CMS Physics Analysis Summary

Contact: cms-pog-conveners-jetmet@cern.ch 2019/07/24

Machine learning-based identification of highly Lorentz-boosted hadronically decaying particles at the CMS experiment

The CMS Collaboration

Abstract

In this note, machine learning (ML) based techniques are presented to identify and classify hadronic decays of highly Lorentz-boosted W/Z/H bosons and top quarks, to be used by the CMS Collaboration. The techniques presented include the Energy Correlation Functions tagger, the Boosted Event Shape Tagger, the ImageTop tagger, and the DeepAK8 tagger. Techniques without ML have also been evaluated and are included for comparison. An alternative approach for jet clustering and identification, the Heavy Resonance Tagger with Variable-R, has been also studied. The identification performance is studied in simulated events and directly compared among algorithms. The algorithms are also validated using 35.9 fb^{-1} of proton-proton events collected at $\sqrt{s} = 13$ TeV, and systematic uncertainties are assessed. The new techniques studied in this note provide significant performance improvements over non-ML techniques, reducing the background rate by up to a factor of $\sim$10 for the same signal efficiency.

1 Introduction

At the Large Hadron Collider (LHC) [1] at CERN, efficiently classifying hadronic decays of heavy standard model (SM) particles that are reconstructed within a single jet can provide a powerful handle for improving the sensitivity in searches for physics beyond the standard model (BSM) and in measurements of SM parameters. The understanding of jet substructure and highly Lorentz-boosted W/Z/H bosons and top (t) quark jets has advanced dramatically in recent years, both experimentally [2] and theoretically [3]. For a particle with a Lorentz boost of γ, the angular separation between its decay products scales like $\theta \sim 2/\gamma$. Understanding the radiation pattern of these jets and their substructure is a topic of high theoretical and experimental interest.

In this note, we present studies evaluating and comparing the performances of a suite of algorithms designed to distinguish hadronically decaying massive SM particles with large Lorentz boosts, namely W/Z/H bosons and t quarks, from other jets originating from light-flavor quarks (u/d/s/c/b) or gluons (g), using the CMS detector [4] at the CERN LHC. We refer to such jets as "boosted W/Z/H/t jets," or "W/Z/H/t-tagged jets".

The theoretical and experimental understanding of jet substructure has gained significant precision in recent years. The CMS Collaboration has many relevant measurements of jet substructure and boosted jets, including measurements of the cross section of highly Lorentz-boosted t quarks [5], jet mass in $\mathrm{t\bar{t}}$ [6], dijet [7, 8], and light flavor-enriched [7] samples, and substructure observables in jets of different light-quark flavors [9] in resolved $\mathrm{t\bar{t}}$ events. Similar measurements by the ATLAS Collaboration can be found in Refs. [10–14]. Overall, the systematic effects of jet substructure and boosted jets are well-understood, and after correcting for detector effects, the results are generally consistent with theoretical expectations.

The maturity of these tools comes with significant advantages. These measurements show reasonable (albeit not perfect) agreement between data and simulation and hence give confidence in the ability for this note to use simulation samples to develop advanced techniques based on machine learning (ML). Residual differences between data and simulation will be accounted for by means of scale factors.

The ML-based approaches can be tailored to suit the needs of individual analyses. Overall there are two broad categorizations. First, there are analyses that have background estimates primarily relying on shape comparisons or simulation. Oftentimes, these analyses must have as much signal efficiency as can be attained for a fixed background rejection. Second, there are analyses that rely on sideband extrapolations for background estimates. These analyses require predictable smooth transitions from control regions to signal regions, usually manifesting as simple dependencies on kinematics (ordinarily, p_{T}). A characteristic example is the use of jet mass sidebands for the background estimation. In this case, removing such dependencies is collectively referred to as "mass decorrelation", as described in Ref. [15]. This note will provide tools for both of these scenarios, informed by a strong program of previous study [16–20] to garner confidence in these advanced techniques.

A brief description of the CMS detector is presented in Section 2. The Monte Carlo (MC) simulated events used for the results are discussed in Section 3, and details of the CMS event reconstruction and the event selections used for the studies are summarized in Sections 4 and 5, respectively. Section 6 presents an overview of the methods currently used in CMS for heavy resonance identification, and describes a set of novel algorithms that utilize ML methods and observables for this task. For the former, the discussion builds on the work documented in [16–20]. Section 7 details the studies performed to understand the complementarity between the

algorithms using MC simulated events. The performance of the algorithms is validated in data samples collected in proton-proton (pp) collisions at $\sqrt{s} = 13\,\text{TeV}$ by the CMS experiment at the LHC in 2016, and corresponding to an integrated luminosity of $35.9\,\text{fb}^{-1}$. The results, along with the effect of systematic uncertainties in their performance, are discussed in Section 8.

2 The CMS detector

The central feature of the CMS apparatus is a superconducting solenoid of 6 m internal diameter, providing a magnetic field of 3.8 T. Within the superconducting solenoid volume are a silicon pixel and strip tracker, a lead tungsten crystal electromagnetic calorimeter (ECAL), and a brass and scintillator hadron calorimeter (HCAL), each composed of a barrel and two endcap sections. Forward calorimeters extend the pseudorapidity (η) coverage provided by the barrel and endcap detectors [21]. Muons are measured in gas-ionization chambers embedded in the steel flux-return yoke outside the solenoid.

In the barrel section of the ECAL, an energy resolution of about 1% is achieved for unconverted or late-converting photons in the tens of GeV energy range. The remaining barrel photons have a resolution of about 1.3% up to a pseudorapidity of $|\eta| = 1$, rising to about 2.5% at $|\eta| = 1.4$. In the endcaps, the resolution of unconverted or late-converting photons is about 2.5%, while the remaining endcap photons have a resolution between 3 and 4% [22].

In the region $|\eta| < 1.74$, the HCAL cells have widths of 0.087 in pseudorapidity and 0.087 in azimuth (ϕ). In the η-ϕ plane, and for $|\eta| < 1.48$, the HCAL cells map on to 5×5 ECAL crystals arrays to form calorimeter towers projecting radially outwards from close to the nominal interaction point. At larger values of $|\eta|$, the size of the towers increases and the matching ECAL arrays contain fewer crystals.

Muons are measured in the pseudorapidity range $|\eta| < 2.4$, with detection planes made using three technologies: drift tubes, cathode strip chambers, and resistive plate chambers. Matching muons to tracks measured in the silicon tracker results in a relative transverse momentum resolution for muons with $20 < p_\text{T} < 100$ GeV of 1.3–2.0% in the barrel and better than 6% in the endcaps, The p_T resolution in the barrel is better than 10% for muons with p_T up to 1 TeV [23].

The silicon tracker measures charged particles within the pseudorapidity range $|\eta| < 2.5$. It consists of 1440 silicon pixel and 15 148 silicon strip detector modules and is located in the 3.8 T field of the superconducting solenoid. Isolated particles of $p_\text{T} = 100$ GeV emitted at $|\eta| < 1.4$ have track resolutions of 2.8% in p_T and 10 (30) μm in the transverse (longitudinal) impact parameter [24].

Events of interest are selected using a two-tiered trigger system [25]. The first level (L1), composed of custom hardware processors, uses information from the calorimeters and muon detectors to select events at a rate of around 100 kHz within a time interval of less than 4 μs. The second level, known as the high-level trigger (HLT), consists of a farm of processors running a version of the full event reconstruction software optimized for fast processing, and reduces the event rate to around 1 kHz before data storage.

A more detailed description of the CMS detector, together with a definition of the coordinate system used and the relevant kinematic variables, can be found in Ref. [4].

3 Simulated events

Simulated pp collision events are generated at a center-of-mass energy of 13 TeV using various generators described below. They are used for the design and the performance studies of the heavy resonance identification algorithms, and to compare to their performance in data. The signal samples, enriched in one or more of boosted W/Z/H/t jets, are obtained from the simulation of BSM processes. The t and W jet signal samples are obtained from heavy spin-1 Z′ resonances decaying to either a pair of t quarks ($t\bar{t}$) or a pair of W bosons, respectively. These resonances are narrow, having intrinsic widths equal to 1% of the resonance mass. The Z and Higgs jet signal samples are obtained from the decay of spin-2 Graviton resonances to a pair of Z or Higgs bosons, respectively, following the narrow width assumption. The $Z^{'}$ and Graviton samples are simulated with MADGRAPH [26] and interfaced with PYTHIA 8.212 [27, 28] for the hadronization. Signal events are generated for different Z′ and Graviton mass scenarios, allowing for signal jets over a wide range of p_T. The background sample is represented by jets produced via the strong interaction of quantum chromodynamics (QCD), referred to as "QCD multijet" processes. The QCD multijet events are generated using PYTHIA 8.212 in exclusive p_T bins of the leading quark or gluon using the NNPDF2.3LO [29] parton distribution function (PDF) set.

A set of MC samples is needed for the study of the performance of the tagging algorithms in data. The $t\bar{t}$ process is generated with the next-to-leading-order (NLO) generator POWHEG v2.0 [30–32] interfaced to PYTHIA for the showering. Simulated events originating from W+jets, Z+jets and γ+jets, are generated using MADGRAPH5_aMC@NLO 2.3.3 [33] at leading order (LO) accuracy using the LO NNPDF3.0 [54] PDF set. The WZ, ZZ, $t\bar{t}W$, $t\bar{t}Z$, and $t\bar{t}\gamma$ processes are generated using MADGRAPH5_aMC@NLO at NLO accuracy, the single t quark process in the tW channel and the WW process are generated at NLO accuracy with POWHEG v2.0, all using the NLO NNPDF3.0 PDF set. In all of the aforementioned cases, parton showering and hadronization is simulated in PYTHIA 8.212. Double counting of partons generated using PYTHIA with those using MADGRAPH5_aMC@NLO is eliminated using the MLM [34] and the FXFX [33] matching schemes, for the LO and NLO samples, respectively.

The systematic uncertainties associated with the performance of the taggers are evaluated using simulated events produced with alternative generation settings. For the $t\bar{t}$ process, an additional sample is generated using POWHEG v2.0 interfaced with HERWIG++ v2.7.1 [35, 36] to assess systematic uncertainties related to the modeling of the parton showering and hadronization. Additional QCD multijet samples are generated using MADGRAPH5_aMC@NLO 2.3.3, interfaced with PYTHIA 8.212 to test the modeling of the hard scatter in background jets, or generated solely with HERWIG++ providing an alternative description of the background jets.

The most precise cross section calculations are used to normalize the SM simulated samples. In most cases, this is next-to-next-to-leading order (NNLO) accuracy in the inclusive cross section. Finally, the p_T spectrum of top quarks in $t\bar{t}$ events is reweighted (referred to as "top quark p_T reweighting") to account for effects due to missing higher-order corrections in MC simulation, according to the results presented in Ref. [37]. The simulation of the QCD multijet and γ+jets processes is based on LO calculations. To account for missing higher order corrections, the simulated QCD multijet events and the γ+jets events are reweighted such that the p_T distribution of the leading jet in simulation matches data. In both cases, contributions from other processes are subtracted from data using the predicted cross sections before extracting the weights.

A full GEANT 4-based model [38] is used to simulate the response of the CMS detector to SM background samples. Event reconstruction is treated in the same manner for MC simulation

as for data. A nominal distribution of multiple pp collisions in the same or neighboring bunch crossings (referred to as "pileup") is used to overlay the simulated events. The events are then reweighted to match the pileup profile observed in the collected data. For the data used in this note, there were an average of 23 interactions per bunch crossing.

4 Event reconstruction and physics objects

Events are reconstructed using the CMS particle-flow (PF) event algorithm [39], which aims to reconstruct and identify each individual particle with an optimized combination of information from the various elements of the detector. Particles are identified as charged hadrons, neutral hadrons, photons, electrons, or muons, and constitute the mutually exclusive list of PF candidates in the event. The PF candidates are then used to build higher level objects such as jets. Events are required to have at least one reconstructed vertex. In the case of multiple events with multiple reconstructed vertices, the one with the largest value of summed physics object p_{T}^2 is taken to be the primary pp interaction vertex. The physics objects are those returned by a jet-finding algorithm [40, 41] applied to the tracks associated with the vertex, and the associated $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$.

Photons are reconstructed from energy depositions in the ECAL using identification algorithms that utilize a collection of variables related to the spatial distribution of shower energy in the supercluster (a group of 5x5 ECAL crystals), the photon isolation, and the fraction of the energy deposited in the HCAL behind the supercluster relative to the energy observed in the supercluster [22, 42]. The requirements imposed on these variables ensure an efficiency of 80% in selecting prompt photons. Photon candidates are required to be reconstructed with $p_{\mathrm{T}} > 200$ GeV and $|\eta| < 2.5$. Simulation-to-data correction factors are used to correct photon identification performance in MC.

Electrons are reconstructed by combining information from the inner tracker with energy depositions in the ECAL [42]. Muons are reconstructed by combining tracks in the inner tracker and in the muon system [23]. Tracks associated with electrons or muons are required to originate from the PV, and a set of quality criteria is imposed to assure efficient identification [23, 42]. To suppress misidentification of charged hadrons as leptons, we require electrons and muons to be isolated from jet activity within a p_{T}-dependent cone size defined by a radius R_{rel} in the η-ϕ plane, where ϕ is the azimuthal angle in radians, and $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$. The relative isolation, I_{rel}, is defined as the scalar sum of the p_{T} of the PF candidates within the cone divided by the lepton p_{T}. Charged PF candidates not originating from the PV, as well as PF candidates identified as electrons or muons, are not considered in the sum.

The isolation sum I_{rel} is corrected for contributions of neutral particles originated from pileup interactions using an area-based estimate [43] of pileup energy deposition in the cone. The requirements imposed to the electron and muon candidates lead to an average efficiency of 70% and 95%, respectively. In addition the electron and muon candidates are required to have $p_{\mathrm{T}} > 40$ GeV and be within the tracker acceptance of $|\eta| < 2.5$. The electron and muon identification performance in simulation is corrected to match the performance in data.

The primary jet collection in this note is produced by clustering PF candidates using the anti-k_{T} algorithm [40] with a distance parameter of $R = 0.8$ with the FASTJET 3.1 software package [40, 41]. This jet collection will be referred to as "AK8 jets". A collection of jets produced using the Cambridge-Aachen (CA) [44, 45] clustering algorithm with $R = 1.5$ is also used in this note. This jet collection will be referred to as "CA15 jets". In both jet collections, the mitigation of the effect of pileup, relies on the "PileUp Per Particle Identification (PUPPI)" [46] method, which

uses local shape information around each particle in the event, event pileup properties, and tracking information together, to mitigate the effect of pileup on jet observables. PUPPI thus operates at the PF candidate level, before any jet clustering is performed. A local variable α is computed which contrasts the collinear structure of QCD with the low-p_{T} diffuse radiation arising from pileup interactions. This α variable is used to calculate a weight correlated with the probability that an individual PF candidate originates from a pileup collision. These per-PF candidate weights are used to rescale the four-momenta of each PF candidate to correct for pileup. The resulting PF candidate list is used as input to the clustering algorithm. A detailed description of the PUPPI implementation in CMS can in found in Ref. [47]. No additional pileup corrections are applied to jets clustered from these weighted inputs. Corrections are applied to the jet energy scale to compensate for nonuniform detector response [48]. Jets are required to have $p_{\mathrm{T}} > 200$ GeV and $|\eta| < 2.4$.

A collection of smaller R jets, which is distinct from the collection of jets discussed earlier, is used to define the event samples for the validation of the algorithms. These jets are reconstructed with the anti-k_{T} algorithm with $R = 0.4$, and will be referred to as "AK4 jets". To reduce the effect of pileup collisions, charged PF candidates identified as originating from pileup vertices are removed before the jet clustering, based on the method known as "charged-hadron subtraction" [48]. An event-by-event jet area-based correction [48] is applied to the jet four-momenta to remove the remaining energy from pileup vertices. As with the AK8 and CA15 jets described above, additional corrections to the jet energy scale are applied to compensate for nonuniform detector response. The AK4 jets are required to have $p_{\mathrm{T}} > 30$ GeV and be contained within the tracker volume of $|\eta| < 2.4$.

Jets originating from the hadronization of bottom (b) quarks are identified, or "tagged", using the combined secondary vertex (CSVv2) b tagging algorithm [49]. The working point used provides an efficiency for the b tagging of jets originating from b quarks that varies from 60 to 75% depending on p_{T}, whereas the misidentification rate for light quarks or gluons is $\sim 1\%$, and $\sim 15\%$ for charm quarks.

For the studies presented in this note, the simulated signal jets (AK8 or CA15 jets) are identified as boosted W/Z/H/t jets when the ΔR between the reconstructed jet and the closest truth particle (W/Z/H boson or t quark) before the decay, denoted as $\Delta R(\text{jet}, \text{truth-particle})$, is less than 0.6 for both jet collections. This definition allows for a consistent comparison of the performance of the algorithms using collections of jets clustered with different R. The fraction of AK8 jets with $\Delta R(\text{AK8}, \text{truth-particle}) < 0.6$ as a function of the p_{T} of the truth particle, for jets initiated from the decay of a W boson (left) or t quark (right), is shown in Fig. 1. This "matching" efficiency of W bosons (t quarks) reaches a plateau of nearly 100% for $p_{\mathrm{T}} \gtrsim 200$ (400) GeV. The corresponding efficiency curve for CA15 jets with $\Delta R(\text{CA15}, \text{truth-particle}) < 0.6$ is superimposed on the plots, showing consistent results with AK8 jets. A similar efficiency is obtained when a relaxed selection of $\Delta R(\text{CA15}, \text{truth-particle}) < 1.2$ is applied on CA15 jets. This justifies the use of the same $\Delta R(\text{jet}, \text{truth-particle})$ reconstruction criteria for both jet collections.

Additional criteria are applied on the simulated jets for the evaluation of the performance in data and the calibration of the algorithms. The partonic decay products (b, q_1, q_2 for t quarks, or q_1, q_2 for W, Z or Higgs bosons) are required to be fully contained in the AK8 (CA15) jet, satisfying $\Delta R(\text{AK8}, \mathrm{q}_i) < 0.6$ ($\Delta R(\text{CA15}, \mathrm{q}_i) < 1.2$). The above requirements are based on studies carried out in [17]. The "merging" efficiency as a function of the p_{T} of the truth particle (i.e. the efficiency for the decay products of the t quark or W boson to be fully contained in a single jet based on the above requirements), is superimposed on Fig. 1. For W bosons (t quarks) with $p_{\mathrm{T}} \gtrsim 200$ (650) GeV, at least 50% of the AK8 jets fully contain the W (t) decay products.

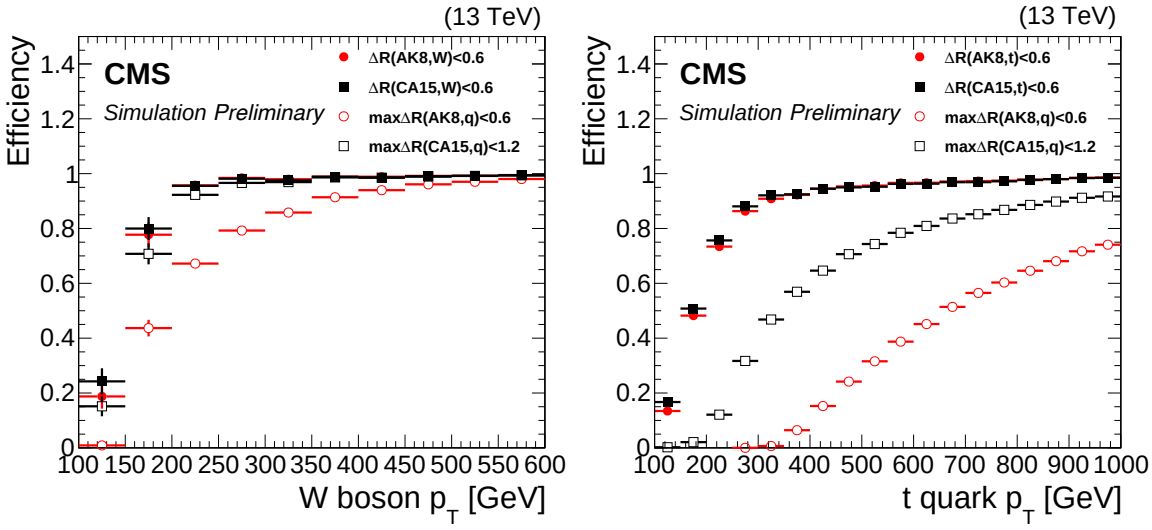

Figure 1: Matching efficiency as a function of the p_{T} of the truth particle; hadronically decaying W bosons (left) and t quarks (right). This efficiency is defined as the fraction of the truth particles (t quarks or W bosons) that are within $\Delta R < 0.6$ with an AK8 or CA15 jet with $p_{\mathrm{T}} > 200$ GeV and $|\eta| < 2.4$. Superimposed is the merging efficiency as a function of the truth particle p_{T} when all decay products are within $\Delta R(\mathrm{AK8}, \mathrm{q}_i) < 0.6$ ($\Delta R(\mathrm{CA15}, \mathrm{q}_i) < 1.2$) with an AK8 (CA15) jet.

In the case of CA15 jets, similar efficiency is achieved for W bosons (t quarks) with $p_{\mathrm{T}} \gtrsim 150$ (350) GeV.

In the case of background jets, partons (u, d, s, c, b, and gluon) from the hard scattering are required to be contained in the jet cone.

Lastly, the $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$, is defined as the negative of the vectorial sum of the $\vec{p}_{\mathrm{T}}$ of all PF candidates in the event [50]. Its magnitude is denoted as $p_{\mathrm{T}}^{\mathrm{miss}}$. The jet energy scale corrections applied to the jets are propagated to $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$.

5 Event selection

Several samples are utilized to validate the performance of the tagging algorithms in data. The single-μ signal sample is used to calibrate the t quark and W boson identification performance in a sample enriched in hadronically decaying t quarks. The di-jet sample, dominated by light-flavor quarks and gluons, enables the study of the identification probability of background jets (misidentification rate) in a wide range of p_{T}. The misidentification rate depends on the flavor of the parton that initiated the jet. Therefore, in addition to the di-jet sample, the single-γ background sample is further utilized. The di-jet and single-γ samples differ in the light-quark and gluon fractions. The former has a larger fraction of gluon jets than the latter.

Systematic effects are quantified using these samples for this analysis, motivated by dominant uncertainties in measurements corrected for detector effects.

5.1 The single-μ signal sample

The single-μ signal sample has been recorded using a single-muon trigger that selects events online based on the p_{T} of the muon. Candidate events are required to have exactly one muon

with $p_{\mathrm{T}} > 55\,\mathrm{GeV}$, satisfying the identification criteria defined in Section 4, except for the requirement related to the isolation of leptons I_{rel}. In the boosted semi-leptonic $\mathrm{t\bar{t}}$ regime, the lepton from the W boson decay often overlaps with the b jet from the t quark decay, leading to large values of I_{rel}. Therefore, a custom isolation criterion is applied by requiring a minimal distance between the muon and the nearest AK4 jet, $\Delta R(\mu, \mathrm{AK4}) > 0.4$, or the perpendicular component of the muon p_{T} with respect to the nearest AK4 jet, $p_{\mathrm{T,rel}} > 25\,\mathrm{GeV}$. This has been extensively used in measurements [5] and searches [51–54] involving boosted t quarks in the semi-leptonic $\mathrm{t\bar{t}}$ sample.

The AK4 jets used in this selection are clustered from PF candidates after removing muons with $p_{\mathrm{T}} > 55\,\mathrm{GeV}$. The custom isolation requirement results in an up to 40% increase in the statistical power of the sample. To suppress the contribution from QCD multijet processes we require $p_{\mathrm{T}}^{\mathrm{miss}} > 50$ GeV. To enhance the sample purity in $\mathrm{t\bar{t}}$ events, we require the presence of two or more AK4 jets, at least one of which is reconstructed as a b jet. In addition, to probe boosted topologies we require the p_{T} of the leptonically decaying W, defined as $\vec{p}_{\mathrm{T}}(W) = \vec{p}_{\mathrm{T}}(\mu) + \vec{p}_{\mathrm{T}}^{\mathrm{miss}}$, and the scalar sum p_{T} of the AK4 jets, denoted as H_{T}, to be greater than 250 GeV. The t/W candidate is the highest p_{T} AK8 or CA15 jet in the event with $p_{\mathrm{T}} > 200\,\mathrm{GeV}$, satisfying the criteria discussed in Section 4. To further improve the purity, we require the azimuthal angle $\Delta\phi$ between the AK8 or CA15 jet and the muon to be greater than 2 radians. The purity of the sample in semi-leptonic $\mathrm{t\bar{t}}$ events is $\sim 70\%$. Other contributions arise from QCD multijet ($\sim 15\%$) and W+jets ($\sim 10\%$) processes.

5.2 The di-jet background sample

The di-jet background sample has been recorded using a trigger that requires H_{T}, where H_{T} is defined as the scalar sum of the p_{T} of the AK4 jets in the event. Events with $H_{\mathrm{T}} > 1000$ GeV are selected to ensure 100% trigger efficiency. Events are required to have at least one AK8 or CA15 jet meeting the requirements presented in Section 4, and the absence of electrons or muons, leading to a sample dominated by jets from the QCD multijet process, which are backgrounds to the algorithms presented here.

5.3 The single-γ background sample

The single-γ background sample has been collected using an isolated single-photon trigger. Events with a photon with $p_{\mathrm{T}} > 200\,\mathrm{GeV}$ are selected to ensure 100% trigger efficiency. The photon is further required to satisfy the criteria presented in Section 4. In addition to the photon, the single-γ sample is required to have at least one AK8 or CA15 jets and no electrons or muons. The sample consists of $\sim 80\%$ γ+jets events, whereas smaller contribution from QCD multijet events is $\sim 15\%$.

6 Overview of the algorithms

This section presents recently developed CMS heavy object tagging methods. However, to understand the historical developments and their limitations, we first present tagging algorithms that do not rely on ML-based methods, which rely on selections on a set of jet substructure observables ("cut-based" approaches). In order to better explore the complementarity between the jet substructure variables, alternative tagging algorithms were developed using multivariate methods. Lastly, to exploit the full potential of the CMS detector and event reconstruction, methods based on Deep Neural Networks (DNN) are explored using either high level inputs (e.g. jet substructure observables), or lower level inputs, such as PF candidates and secondary vertices. For a wider overview of the most recent developments in ML-based tagging see for

example Ref.[55]. Finally, dedicated versions of the algorithms are developed that are only loosely correlated with the jet mass. A detailed discussion of each algorithm is presented in this Section and a summary of all t quark, and W, Z or Higgs boson identification algorithms is presented in Table 1.

Table 1: Summary of the CMS algorithms for the identification of hadronically decaying t quarks and W, Z and Higgs bosons. The column "p_T (jet)" indicates the jet p_T threshold to be used in each algorithm.

Algorithm	p_T (jet) [GeV]	t quark	W boson	Z boson	Higgs boson	decay modes
$m_{SD} + \tau_{32}$	400	✓				
$m_{SD} + \tau_{32} + b$	400	✓				
$m_{SD} + \tau_{21}$	200		✓	✓		
HOTVR	200	✓				
$N_3 - BDT$ (CA15)	200	✓				
$m_{SD} + N_2$	200		✓	✓	✓	
BEST	500	✓	✓	✓	✓	
ImageTop	600	✓				
DeepAK8	200	✓	✓	✓	✓	✓
Jet mass decorrelated algorithms						
$m_{SD} + N_2^{DDT}$	200		✓	✓	✓	
double–b	300			✓	✓	
ImageTop–MD	600	✓				
DeepAK8–MD	200	✓	✓	✓	✓	✓

6.1 Jet grooming and substructure variable-based algorithms

Historically, the boosted t quark and W/Z/H boson tagging methods used by the CMS Collaboration are based on a combination of selection criteria on the jet mass and and the energy distribution inside the jet [16–20].

The jet mass is one of the most powerful observables to discriminate t quark and W/Z/H boson jets from background jets (i.e. jets stemming from the hadronization of light quarks or gluons). QCD will cause a radiative shower of quarks and gluons, which will be collimated within a jet. The probability for a gluon to be radiated from a propagating quark or gluon is inversely proportional to the angle and energy of the radiated gluon, hence will tend to appear close to the direction of the original quark or gluon. These radiated gluons tend to be soft, resulting in a characteristic "Sudakov" peak structure. This is explained in detail in Ref. [8]. Contributions from initial state radiation, the underlying event, and pileup also contribute strongly to the jet mass, especially at larger values of R. As such, the jet mass from QCD scales as the product of the jet p_T and R.

Methods have been developed to remove soft or uncorrelated radiation from jets, called "grooming" methods. These methods strongly reduce the "Sudakov" peak structure in the jet mass distribution. Removing the soft and uncorrelated radiation results in a much weaker dependence of the jet mass on its p_T.

The t quark and W/Z/H bosons have an intrinsic mass, and the jet substructure tends to be dominated by electroweak splittings at larger angles than QCD. This can be exploited to separate such jets from jets arising from heavy SM particles.

The grooming method used most often in CMS is the "modified mass drop tagger" algorithm (mMDT) [56], which is a special case of the "soft drop" (SD) method [57]. This algorithm

systematically removes the soft and collinear radiation from the jet in a manner that can be theoretically calculated [58, 59] (see comparisons to data in Ref. [8]).

The first step in the SD algorithm is the reclustering of the jet constituents with the CA algorithm, and then the identification of two "subjets" within the main jet by reversing the CA clustering history. The jet is considered as the final jet if the two subjets meet the SD condition:

$$\frac{\min(p_{T1}, p_{T2})}{p_{T1} + p_{T2}} > z_{\text{cut}} \left(\frac{\Delta R_{12}}{R_0} \right)^{\beta}, \tag{1}$$

where p_{T1} (p_{T2}) is the p_{T} of the leading (sub-leading) subjet and ΔR_{12} their angular separation. The parameters z_{cut} and β define what the algorithm considers "soft" and "collinear," respectively. The values used in CMS are $z_{\text{cut}} = 0.1$ and $\beta = 0$ (making this identical to the mMDT algorithm, although for notation we still denote this as SD). If the SD condition is not met, the lower-p_{T} subject is removed and the same procedure is followed until Eq. 1 is satisfied or no further declustering can be performed.

The two subjets returned by the SD algorithm are used to calculate the jet mass. Figure 2 shows the distribution of the AK8 jet mass after applying the SD algorithm (m_{SD}) in signal and background jets in simulation. The jet mass has been measured in data at CMS for t [6] and QCD jets [7, 8] in previous papers.

The m_{SD} in background jets peaks below 20 GeV due to the suppression of the Sudakov peak, whereas the m_{SD} for signal jets peaks around the mass of the heavy SM particle (t quark, or W/Z/H bosons). Similar conclusions also hold for CA15 jets. Based on these observations, we define three regions in m_{SD}. The "W/Z mass" region with $65 < m_{\text{SD}} < 105$ GeV, the "H mass" region with $90 < m_{\text{SD}} < 140$ GeV, and the "t mass" region with $105 < m_{\text{SD}} < 210$ GeV. These definitions will be used throughout the studies in this note unless stated otherwise.

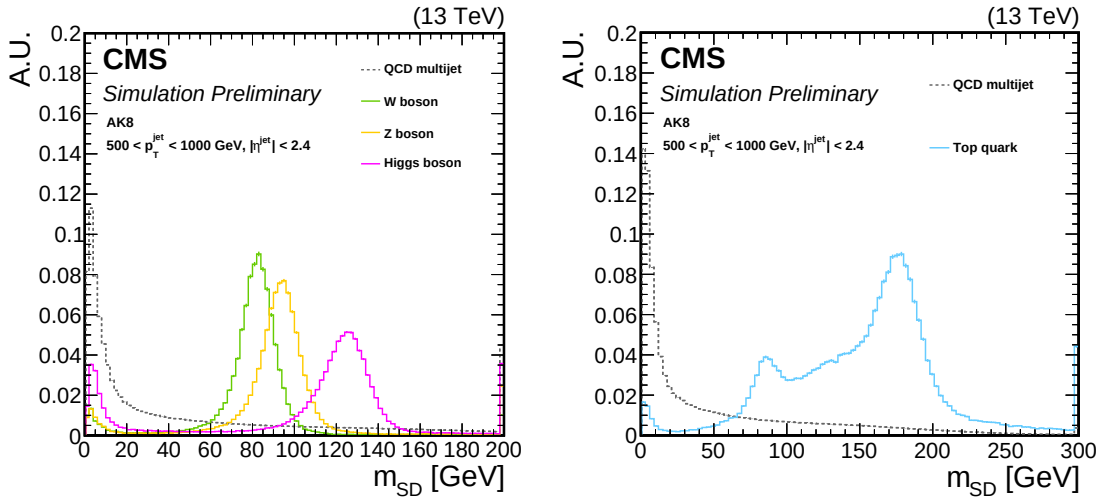

Figure 2: Comparison of the m_{SD} shape in signal and background AK8 jets in simulation. The fiducial selection on the jets is displayed on the plots. Signal jets are defined as jets arising from hadronic decays of W/Z/H bosons (left) or t quarks (right), whereas background jets are obtained from the QCD multijet sample.

An additional handle to separate signal from background events is to exploit the energy distribution inside the jet. Jets resulting from the hadronic decays of a heavy particle to N separate quarks or gluons are expected to have N subjets. For two-body decays like W/Z/H, there are

two subjets, while for t quarks, there are three. In contrast, jets arising from the hadronization of light quarks or gluons are expected to only have one or two such regions (in the case of gluon splitting). The N-subjettiness variables [60, 61],

$$\tau_N = \frac{1}{d_0} \sum_i p_{\mathrm{T},i} \min \left[\Delta R_{1,i}, \Delta R_{2,i}, \ldots, \Delta R_{N,i} \right], \tag{2}$$

provide a measure of the number of subjets that can be found inside the jet. The index i refers to the jet constituents, while the ΔR terms represent the spatial distance between a given jet constituent and the subjets. The quantity d_0 is a normalization constant. The centers of hard radiation are found by performing the exclusive k_T algorithm [62, 63] on the jet constituents before the application of any grooming techniques. The values of the τ_N variables are typically small if the jet is compatible with having N or more subjets. However, a more discriminating observable is the ratio of different τ_N variables. To this end, the ratio τ_{32} is used for t quark identification, whereas the ratio τ_{21} is used for W/Z/H boson identification. The distribution τ_{21} and τ_{32} for signal and background AK8 jets is shown in Fig. 3. Measured values of these distributions at CMS can also be found for light-flavor jets in Ref. [9]. Typical operating regions for τ_{32} (τ_{21}) are 0.44–0.89 (0.35–0.65), which correspond to a misidentification rate after the m_{SD} selection of 0.1–10% (0.1–10%), respectively.

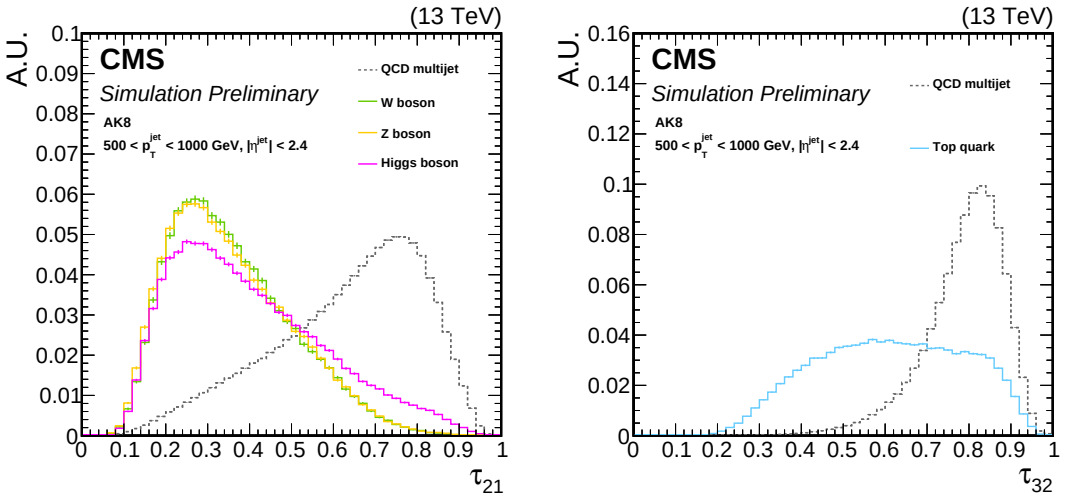

Figure 3: Comparison of the τ_{21} (left) and τ_{32} (right) shape in signal and background AK8 jets. The fiducial selection on the jets is displayed on the plots. As signal jets we consider jets stemming from hadronic decays of W, Z or H bosons (left) or t quarks (right), whereas background jets are obtained from the QCD multijet sample.

The baseline W and Z boson, collectively referred to as V boson, tagging algorithm, based on selections on m_{SD} and τ_{21} will be referred to as "$m_{\mathrm{SD}} + \tau_{21}$" in this note. V tagging with this method is used frequently in current analyses (see for example Ref.[64–67]) starting at approximately 200 GeV in p_{T}.

For t quark tagging we studied a tagger based on m_{SD} and τ_{32}, which will be referred to as "$m_{\mathrm{SD}} + \tau_{32}$". An additional improvement in the performance of the t quark identification is achieved by applying the CSVv2 b tagging algorithm discussed in Section 4 on the subjets returned by the SD algorithm. In the studies presented in this note we require at least one of the two subjets to pass the loose working point of the CSVv2 algorithm, corresponding to b quark identification efficiency $\sim$ 85%, with a misidentification rate for light quarks and gluons

$\sim$ 10%, and $\sim$ 60% for c quarks. This version of the baseline t quark tagging algorithm will be referred to as "$m_{\mathrm{SD}} + \tau_{32} + \mathrm{b}$". Top tagging with this method is used extensively in physics analyses (see for example Ref. [54, 66–68]) tagging high p_{T} tops, which start to merge into the AK8 cone at around 350 GeV and are fully efficient at around 600 GeV. For applications below this mass range, analyses can profit from the larger (or variable) R clustering algorithms discussed in the following sections.

6.2 Heavy Resonance Tagger with Variable R

The Heavy Resonance Tagger with Variable R (HOTVR) [69] is a new cut-based algorithm for the identification of boosted jets. It introduces a new jet clustering technique with a variable R and removal of soft contributions during the clustering. The clustering is similar to other standard sequential clustering algorithms like the CA algorithm, where particles are sequentially added. However, instead of a fixed R, HOTVR uses a p_{T}-dependent R (R_{HOTVR}), defined as:

$$R_{\mathrm{HOTVR}} = \begin{cases} R_{\mathrm{min},} \text{ for } \rho/p_{\mathrm{T}} < R_{\mathrm{min}} \\ R_{\mathrm{max},} \text{ for } \rho/p_{\mathrm{T}} > R_{\mathrm{max}} \\ \rho/p_{\mathrm{T}}, \text{ elsewhere} \end{cases} . \tag{3}$$

The values of ρ correspond to the typical scale of the event (O(100) GeV). In the case of $\rho \to 0$ the algorithm is identical to the CA algorithm for $R = R_{\mathrm{min}}$, whereas for $\rho \to \infty$ is identical to the CA algorithm for $R = R_{\mathrm{max}}$. Higher values of ρ result in larger jet sizes. The parameters R_{min} and R_{max} are introduced for robustness of the algorithm with respect to experimental effects.

Inspired by [69], at each clustering step the invariant mass, m_{ij}, between two subjets, "i" and "j", entering the jet clustering, is calculated. If m_{ij} is greater than a mass threshold, μ, the following condition is verified:

$$\theta m_{ij} > \max(m_i, m_j), \tag{4}$$

where m_i and m_j are the masses of the two subjets, and θ is a parameter that determines the strength of the condition and ranges between 0 and 1. If the condition in Eq. 4 is not fulfilled the subjet with the lowest mass is discarded, otherwise depending on the relative p_{T} difference of the subjets they are either combined into a single subjet or the softer one is discarded. The algorithm continues until no other subjet is found. The detailed description of the HOTVR algorithm is presented in [69]. Table 2 lists the values of the HOTVR parameters used in CMS. In the CMS implementation, HOTVR jets are clustered using PUPPI corrected PF candidates.

Table 2: Summary of the HOTVR parameters. The p_{Tsub} is the minimum p_{T} threshold of each subjet.

R_{min}	R_{max}	ρ [GeV]	θ	p_{Tsub} [GeV]	μ [GeV]
0.1	1.5	600	0.7	30	30

The HOTVR clustering algorithm is currently explored in CMS for the t quark identification. The jets returned by HOTVR (i.e. "HOTVR jets") are required to have mass consistent with m_{t}, namely $140 < m_{\mathrm{HOTVR}} < 220$ GeV, and at least three subjets, $N_{\mathrm{sub,\,HOTVR}} \geq 3$, the minimum pairwise mass of which should be $m_{\mathrm{sub,\,min}} > 50$ GeV. In addition, the p_{T} of the hardest subjet

should be less than 80% of the p_T of the HOTVR jet. Lastly, to further improve the discrimination, $\tau_{32} < 0.56$ is required. The shape comparison of the main variables of the HOTVR algorithm for signal and background, for different parton p_T ranges, is shown on Fig. 4.

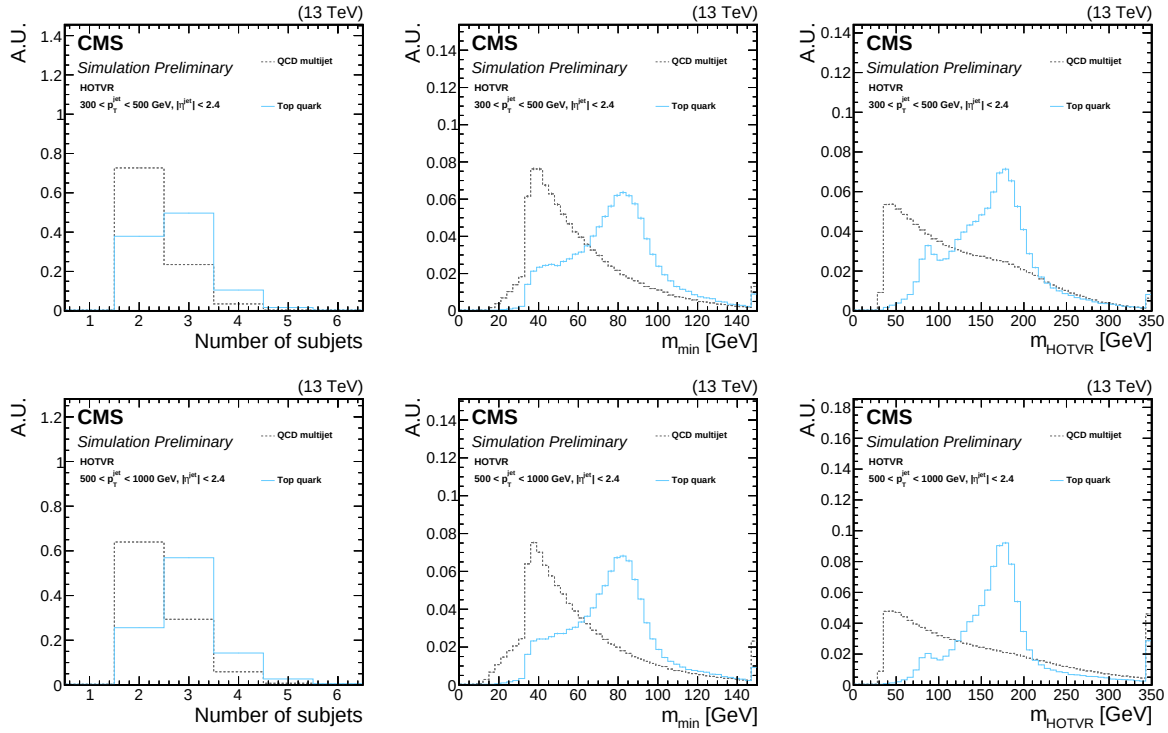

Figure 4: Shape comparison of the main variables of the HOTVR algorithm for signal and background jets, in two different regions of the parton p_T as displayed on the plots.

6.3 Energy correlation functions

A new set of N-prong identification algorithms, the generalized energy correlation functions (ECF) [70], are now used by the CMS Collaboration. The ECFs explore the energy distribution inside a jet by aiming to quantify the number of centers of hard radiation using an axis-free approach, differing form the axis-dependent definition used by N-subjettiness, which reduces the dependence of the observable on the jet p_T. This allows the exploration of complementary information between the two techniques.

For a jet containing N_C particles, an ECF is defined as:

$$ {}_{o}e^{\beta}_{N} = \sum_{1\leq i_1<i_2<\cdots<i_N\leq N_C} \left[\prod_{1\leq k\leq N} \frac{p_T^{i_k}}{p_T^{J}}\right] \prod_{m=1}^{o} \min_{i_j<i_k\in\{i_1,i_2,\cdots,i_N\}}^{(m)} \left\{\Delta R^{\beta}_{i_j,i_k}\right\} , \tag{5} $$

where $1 \leq i_1 < i_2 < \cdots < i_N \leq N_C$ range over the jet constituents. The symbols $p_T^{i_k}$ and p_T^J are the p_T of the constituent i_k and the p_T of the jet, respectively. The notation $\min^{(m)}$ refers to the m^{th} smallest element, and $\Delta R_{i_j,i_k}$ is the distance between constituents i_j and i_k. The parameters N and o must be positive integers, whereas β must be positive. For a concrete example, we calculate the ECF corresponding to $o = 2, N = 3, \beta = 1$. This ECF tests the compatibility of a jet with three centers of hard radiation, but only considering the two smallest angles ($o = 2$):

$$ {}_2e_3^1 = \sum_{1\leq a<b<c\leq M} \frac{p_\mathrm{T}^a}{p_\mathrm{T}^J}\frac{p_\mathrm{T}^b}{p_\mathrm{T}^J}\frac{p_\mathrm{T}^c}{p_\mathrm{T}^J} \min\{\Delta R_{ab}\Delta R_{ac}, \Delta R_{ab}\Delta R_{bc}, \Delta R_{bc}\Delta R_{ac}\}. \tag{6} $$

Moreover, there is the possibility to select subsets of the jet that contain large energy fractions and pairwise opening angles only if the size of the subset is less than or equal to the number of the centers of radiation in the jet. In general, a jet with N centers of radiation has $e_N \gg e_M$, for $M > N$.

6.3.1 ECF for 3-prong decay identification

The ratios of type $(N = 4)/(N = 3)$ can identify the hadronic decays of 3-body decays like t quarks. The paper [70] proposes the specific ratio N_3 for this purpose:

$$ N_3^{(\beta)} = \frac{{}_2e_4^\beta}{({}_1e_3^\beta)^2}. \tag{7} $$

Since a jet contains $N_C \sim \mathcal{O}(p_\mathrm{T}/\mathrm{GeV})$ constituents, and the sum has $\binom{N_C}{N}$ terms, it is prohibitively expensive to compute $e(N = 4)$ on high-momentum jets. For example, about 10–15% of CA15 jets with $p_\mathrm{T} \sim 500\,\mathrm{GeV}$ have more than 100 particles. However, we find that functions are dominated by the hardest particles, and therefore limiting to the 100 hardest particles makes the calculation tractable without significant performance degradation.

In our reconstruction, the ECF ratios are calculated on jets after the SD grooming is applied, which improves the stability of the ECF as a function of mass and p_T. An example of the ECF ratios is shown in the left plot of Fig. 5 for t quark and QCD jets in simulation. The ECF ratios are measured in data in Ref. [9]. While the N_3 is designed to have comparable performance with τ_{32}, its dependence on p_T is reduced.

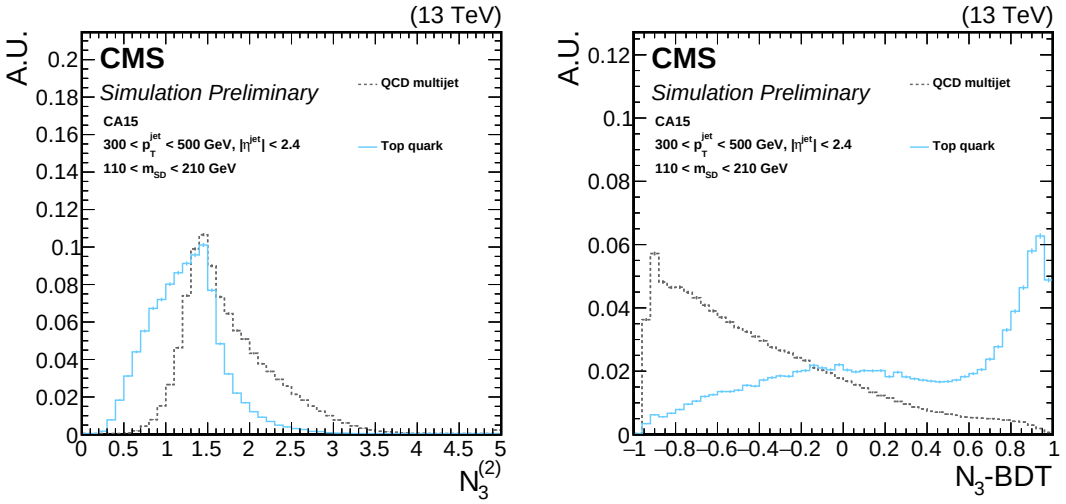

Figure 5: Comparison of the distribution of $N_3^{(2)}$ (left) and the $N_3 - \mathrm{BDT}\,(\mathrm{CA15})$ discriminant (right) in t quarks jets (signal) and jet from QCD multijet processes (background).

To this end, a set of ECF is chosen based on the improvement in the performance of the t tagging algorithm, while in parallel maintaining small dependence on $p_\mathrm{T}(\mathrm{jet})$. Despite the fact that the terms of the ECF are dimensionless, the angular component of the ECF function will

be modified based on the boost of the jet. Therefore, scale invariant ECF ratios are constructed by only considering those ratios that satisfy:

$$\frac{{}_a e_N^{\alpha}}{({}_b e_M^{\beta})^x}, \text{ where } M \leq N \text{ and } x = \frac{a\alpha}{b\beta}. \tag{8}$$

Only ratios that are not highly correlated among themselves are considered for the t quark tagging algorithm. ECFs ratios that are not well-described by simulation are discarded. The following 11 ECF ratios are finally selected:

$$\begin{gathered}\frac{{}_1 e_2^{(2)}}{\left({}_1 e_2^{(1)}\right)^2}, \frac{{}_1 e_3^{(4)}}{{}_2 e_3^{(2)}}, \frac{{}_3 e_3^{(1)}}{\left({}_1 e_3^{(4)}\right)^{3/4}}, \frac{{}_3 e_3^{(1)}}{\left({}_2 e_3^{(2)}\right)^{3/4}}, \frac{{}_3 e_3^{(2)}}{\left({}_3 e_3^{(4)}\right)^{1/2}}, \\ \frac{{}_1 e_4^{(4)}}{\left({}_1 e_3^{(2)}\right)^2}, \frac{{}_1 e_4^{(2)}}{\left({}_1 e_3^{(1)}\right)^2}, \frac{{}_2 e_4^{(1/2)}}{\left({}_1 e_3^{(1/2)}\right)^2}, \frac{{}_2 e_4^{(1)}}{\left({}_1 e_3^{(1)}\right)^2}, \frac{{}_2 e_4^{(1)}}{\left({}_2 e_3^{(1/2)}\right)^2}, \frac{{}_2 e_4^{(2)}}{\left({}_1 e_3^{(2)}\right)^2}.\end{gathered} \tag{9}$$

In addition to the ECF, two jet substructure observables are employed to further distinguish t quark jets from light quarks or gluons. The first observable is τ_{32} calculated after applying the SD method on the CA15 jets, defined as τ_{32}^{SD} and the second is the f_{rec} variable of the HEPTopTagger algorithm [71–73], which quantifies the difference between the reconstructed W boson and t quark masses and their expected values, and is defined as:

$$f_{\text{rec}} = \min_{i,j} \left| \frac{m_{ij}/m_{123}}{m_{\text{W}}/m_{\text{t}}} - 1 \right|, \tag{10}$$

where i, j range over the three chosen subjets, m_{ij} is the mass of subjets i and j, and m_{123} is the mass of all three subjets.

The ECF-based t quark tagger, referred to as "$N_3 - \text{BDT}\ (\text{CA15})$", is based on a Boosted Decision Tree (BDT) [74] with the 11 ECF ratios, the τ_{32}^{SD}, and the f_{rec} as inputs. The $N_3 - \text{BDT}\ (\text{CA15})$ was trained using jets with $110 < m_{\text{SD}} < 210\,\text{GeV}$. To avoid possible bias in the identification performance due to differences in the p_{T} spectrum of the signal (t quarks) and background (light quarks or gluons) jets, their contributions are reweighted such that they have a flat distribution in $p_{\text{T}}(\text{jet})$.

Figure 5 on the right shows a comparison of the $N_3 - \text{BDT}\ (\text{CA15})$ discriminant distribution between signal and background jets. The final $N_3 - \text{BDT}\ (\text{CA15})$ algorithm also requires at least one of the two subjets returned by the SD method to be identified as a b jet by the CSVv2 algorithm using the loose working point. The ECF BDT tagger is used for top jet identification in the context of dark matter production in association with a single top in the $p_{\text{T}} > 250\,\text{GeV}$ range [75].

6.3.2 ECF for 2-prong decay identification

Similarly to the identification of 3-prong decays, ECF are explored for the identification of 2-prong decays like W/Z/H bosons. In this case, the signal jets have a stronger 2-point correlation than a 3-point correlation and the discriminant variable N_2^1 can be used to separate jets originating from W/Z/H bosons. The N_2 variable is constructed via the ratio:

$$N_2^1 = \frac{{}_2e_3^1}{({}_1e_2^1)^2}, \tag{11}$$

and presents similar performance to N-subjetiness ratio τ_{21}, with the advantage that it is more stable as a function of the jet mass and p_T. This method will be referred to as "$m_{SD} + N_2$" in this note.

A decorrelation procedure is further applied to avoid distorting the jet mass distribution when a selection based on N_2 is made. We design a transformation from N_2 to N_2^{DDT}, where DDT stands for "designed decorrelated tagger" [15]. The transformation is defined as a function of the dimensionless scaling variable $\rho = \ln(m_{SD}^2/p_T^2)$ and the jet p_T:

$$N_2^{DDT}(\rho, p_T) = N_2(\rho, p_T) - N_2^{(X\%)}(\rho, p_T)\,, \tag{12}$$

where $N_2^{(X\%)}$ is the X percentile of the N_2 distribution in simulated QCD events. This ensures that the selection $N_2^{DDT} < 0$ yields a constant QCD background efficiency of $X\%$ across the mass and p_T range considered with no loss in performance. The value $X = 5$ is used throughout this note, following the choice in [76]. The distributions of the N_2 and N_2^{DDT} in signal and background jets are shown in Fig. 6. Signal jets populate smaller values, whereas background jets have larger values. The N_2 DDT is used for V tagging with p_T in excess of 500 GeV in the search for light dijet resonances [76].

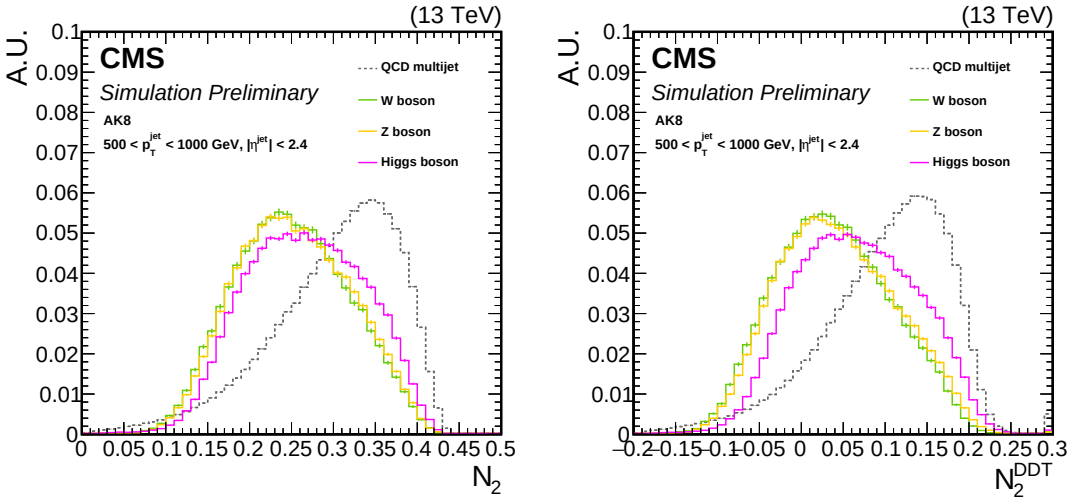

Figure 6: Distributions of the $m_{SD} + N_2$ (left) and $m_{SD} + N_2^{DDT}$ (right) in signal and background jets.

The $m_{SD} + N_2^{DDT}$ observable has been used and validated in several analyses, including [76] and [77].

6.4 The double-b tagger

The standard b-tagging tools, as the CSVv2 discussed in Section 4, can be applied to the subjets returned by the SD algorithm on AK8 jets. Characteristic examples are the $m_{SD} + \tau_{32} + b$ and $N_3 - BDT$ (CA15) algorithms. However these tools can become limited in certain topologies, for example when the two subjets become very collimated. The "double-b" tagger was developed to specifically target Higgs decays to pairs of b quarks in the boosted regime [78]. While it utilizes many of the variables used in the standard CSVv2 b tagging algorithm, it also

employs variables related to the track properties, such as the track impact parameter and the impact parameter significance, the position of the secondary vertices, and information from the two-secondary-vertex system, among others listed in Ref. [78]. These variables are then used as inputs to a BDT. A key feature of the double-b algorithm is that it is designed to minimize the dependence of the BDT discriminant on the jet mass and p_T, making it suitable for other topologies such as boosted Z decays to bottom quarks [77].

The performance of the double-b tagger in simulation is detailed in [78] using Higgs jets as signal, and for background, single-b jets, double-b jets from gluon splitting to a pair of b quarks, and light quark or gluon jets. The $H \to b\bar{b}$ identification efficiency is $\sim 25\%$ ($\sim 70\%$) for $\sim 1\%$ ($\sim 10\%$) misidentification rate [78].

The double-b tagger performance in data is studied in [78] using data events in a recent inclusive search for the Higgs boson in the $b\bar{b}$ decay mode [77]. In that analysis, the Z boson was observed for the first time in the single-jet topology and $b\bar{b}$ decay mode, with a rate consistent within uncertainties with the SM expectation, validating the double-b tagging algorithm for the Higgs and future new physics searches.

The double-b tagger will serve as a reference for the performance of the new methods explored in CMS.

6.5 Boosted Event Shape Tagger

The boosted event shape tagger (BEST) [79] is a multi-classification algorithm designed to discriminate hadronic decays of high-p_T t quarks, and W/Z/H bosons from jets arising from b quarks, light flavor quarks, and gluons. The original algorithm was demonstrated using generator-level particles and efficiently separated jets originating from W/Z/H bosons, t quarks, and b jets. The algorithm has been extended and deployed for use in the CMS experiment, adding an additional category to discriminate jets from light flavor quarks and gluons.

The BEST algorithm obtains discrimination on a jet-by-jet basis by transforming the entire set of jet constituents four times, each with a different boost vector. The boost vectors are obtained by assuming the jet originated from one of the heavy objects under consideration (W/Z/H/t). The jet momentum is held constant while the mass of the jet is adjusted to the theoretical value of the corresponding particle. This results in four distributions of constituents that can be used to discriminate between particle origins. If a jet did originate from the corresponding heavy object hypothesis, its jet constituents will, in general, be more isotropic in the rest frame of that particle. By examining the differences between heavy object hypotheses, discrimination is obtained between the categories of interest (W/Z/H/t/b/other).

In total 59 quantities are used to train a neural network (NN) and classify the AK8 jets. The list of variables is seen in Table 3. For each boost transformation, we calculate the following observables: Fox-Wolfram moments [80], the aplanarity, sphericity, and isotropy quantities based on the eigenvalues of the sphericity tensor as defined in Ref. [81], as well as the jet thrust [82]. Additionally, in each boost hypothesis, AK4 subjets are clustered from the boosted constituents and used to compute pairwise subjet masses for the leading three subjets, as well as the combined mass of the leading four subjets m_{1234}. These AK4 subjets are also used to compute the longitudinal asymmetry A_L, defined as the ratio of the sum of the longitudinal components of the AK4 subjet momenta to the sum of the total AK4 subjet momenta. In addition to these quantities evaluated for each set of boosted jet constituents, the m_{SD}, rapidity, charge, τ_{32}, τ_{21}, and the CSVv2 discriminant for each subjet are used as additional inputs.

The NN is trained with the `scikit-learn` package [83] using the `MLPClassifier` module.

BEST Training Quantities		
Jet Charge	Fox-Wolfram Moment H_1/H_0 (t,W,Z,H)	m_{12} (t,W,Z,H)
Jet η	Fox-Wolfram Moment H_2/H_0 (t,W,Z,H)	m_{23} (t,W,Z,H)
Jet τ_{21}	Fox-Wolfram Moment H_3/H_0 (t,W,Z,H)	m_{13} (t,W,Z,H)
Jet τ_{32}	Fox-Wolfram Moment H_4/H_0 (t,W,Z,H)	m_{1234} (t,W,Z,H)
Jet soft-drop mass	Sphericity (t,W,Z,H)	A_L (t,W,Z,H)
Subjet 1 CSV Value	Aplanarity (t,W,Z,H)	
Subjet 2 CSV Value	Isotropy (t,W,Z,H)	
Maximum Subjet CSV Value	Thrust (t,W,Z,H)	

Table 3: List of input quantities used for the training and evaluation of the BEST algorithm on AK8 jets.

The network architecture is fully-connected and consists of 3 hidden layers with 40 nodes in each layer using a rectified linear unit (ReLU) [84] activation function. The six output nodes correspond to the 6 particle species of interest. We use 500,000 jets in total to train the network, split evenly between the 6 training samples. The training is performed using the ADAM [85] optimizer to minimize the cross-entropy loss, with a constant learning rate of 0.001. BEST W/Z/H/t/b/other multi-classification is currently used for tagging high p_{T} jets in the search for VLQ pair production [67].

6.6 Identification using particle flow candidates: ImageTop

Recent studies, e.g. in Ref. [86], have shown that jet identification algorithms deploying ML methods directly on the jet constituents, yield significantly improved performance compared to traditional algorithms.

To this end, the "ImageTop" t quark identification algorithm is developed. The ImageTop algorithm follows closely the network framework described in Ref. [86], which is an optimization based on the DeepTop framework described in Ref. [87]. This tagging approach uses standard image recognition techniques based on two dimensional Convolutional Neural Networks (CNN) to discriminate boosted t jets from QCD jets. This is performed by pixelizing the jet energy deposits and colorizing based on relevant detector information. Before pixelization, the centroid of the jet is shifted so that it is at the origin and then a rotation is performed such that the major principal axis is vertical. The image is then flipped along both the horizontal and vertical axes such that the maximum intensity is in the lower-left quadrant. After this, the image intensity is normalized and the image is pixelized using 37×37 pixels with a total $\Delta\eta = \Delta\phi = 3.2$, with colors split into neutral p_{T}, track p_{T}, number of muons, and number of tracks. The network architecture uses a layer of 128 feature maps with a 4×4 kernel followed by a second convolutional layer of 64 feature maps. Then a max-pooling layer with a 2×2 reduction factor is used, followed by two more consecutive convolutional layers with 64 features maps followed by another max-pooling layer. A zero-padding in each convolutional layer is used to correct for image-border effects. In the last pooling layer, the 64 maps are flattened into a single one that is passed into a set of three fully connected dense layers of 64, 256 and 256 neurons each. The training uses the ADADELTA optimizer [88] with a learning rate of 0.3, a minibatch size of 128, and the binary cross entropy loss function.

The tagger is modified to use as inputs the PF candidates comprising the AK8 jets, with the colors being the p_{T} of the PF candidates for the full greyscale image, and a separate color for each PF candidate flavor, namely charged hadrons, neutral hadrons, photons, electrons, and muons. The characteristic flavor of the t quark decay is included by applying the DeepFlavour [89] b tagging algorithm to the SD subjets of the AK8 jet. The subjet b-tagging outputs include the

probability of the jet to originate from the following six sources: b quark, $\mathrm{b\bar{b}}$ pair, leptonic b decays, c quark, light quark or gluon. These output probabilities calculated for both subjets along with m_{SD}, are used as inputs (13 in total) into a 64-neuron dense layer and merged with the previous flattened CNN layer and finally input into three fully connected layers of 256 neurons each. The factorization of the b flavor discrimination is important for the versatility of the network, allowing for the flavor identification to be easily removed or validated in parallel, which can be necessary for the validation of objects with no SM analogue. The diagram of the CMS application of this NN can be seen in Fig. 8. The training of the ImageTop is performed using a TitanXP GPU donated by the NVIDIA Corporation.

In order to sustain the ImageTop performance over a wide range of $p_{\mathrm{T}}(\mathrm{jet})$ (the training is performed for jets in the $p_{\mathrm{T}} > 600\,\mathrm{GeV}$ region), the image is adaptively zoomed based on $p_{\mathrm{T}}(\mathrm{jet})$ in order to account for the increased collimation of the t quark decay products at high Lorentz boosts and maintain a static pixel size. The functional form of the zoom is extracted from the average ΔR of the three generator hadronic t quark decay products, and jet energy deposits are corrected to make this constant on average as evaluated from a fit using the inverse jet p_{T} functional form $f(p_{\mathrm{T}}) = p_0 + p_1/p_{\mathrm{T}}$ with $p_0 = 0.066$ and $p_1 = 264$.

As input to the training, a jet p_{T} bias is further reduced by ensuring that the input p_{T} distributions for signal and background jets are similarly shaped by probabilistically removing QCD events based on the ratio of t and QCD p_{T} distributions. A training is also performed by additionally constraining the m_{SD} in a similar manner in order to reduce the mass correlation of the tagger. Since the inputs are relatively simple and do not exhibit secondary mass correlation, this passive approach for decorrelating the ImageTop network is sufficient to remove the mass bias in the fiducial training region ($p_{\mathrm{T}} > 600\,\mathrm{GeV}$ and $|\eta| < 2.4$). This method of mass decorrelation also leads to a factorized sensitivity, where the sensitivity of the full ImageTop network in the top mass region is closely approximated by the sensitivity of the mass-decorrelated version after including a mass selection. This is considered evidence that the decorrelation procedure is nearly optimally sensitive, in contrast to an active decorrelation approach, which could have an impact on the training performance through penalty terms in the loss function. This version of ImageTop will be referred to as "ImageTop–MD".

6.7 Identification using particles flow candidates: DeepAK8

An alternative approach to exploit particle-level information directly with customized ML methods is the "DeepAK8" algorithm, a multiclass classifier for the identification of hadronically decaying particles, with five main categories, W/Z/H/t/other. To increase the versatility of the algorithm, the main classes are further subdivided to the minor categories, corresponding to the decay modes of each particle (e.g. $Z \to \mathrm{b\bar{b}}$, $Z \to \mathrm{c\bar{c}}$ and $Z \to q\bar{q}$).

In the DeepAK8 algorithm, two lists of inputs are defined for each jet. The first list ("particle" list) consists of up to 100 jet constituent particles, sorted by decreasing p_{T}. Typically less than 5% of the jets have more than 100 reconstructed particles, therefore restricting to the 100 hardest particles results in negligible loss of performance. Measured properties of each particle, such as the p_{T}, the energy deposit, the charge, the angular separation between the particle and the jet axis or the subjet axes, etc., are included to help the algorithm extract features related to the substructure of a jet. For charged particles, additional information measured by the tracking detector is also included, such as the displacement and quality of the tracks, etc. These inputs are particularly useful for the algorithm to extract features related to the presence of heavy flavor (b or c) quarks. In total, 42 variables are included for each particle in the "particle" list. A secondary vertex (SV) list consists of up to 7 SVs, each with 15 features, such as the SV kine-

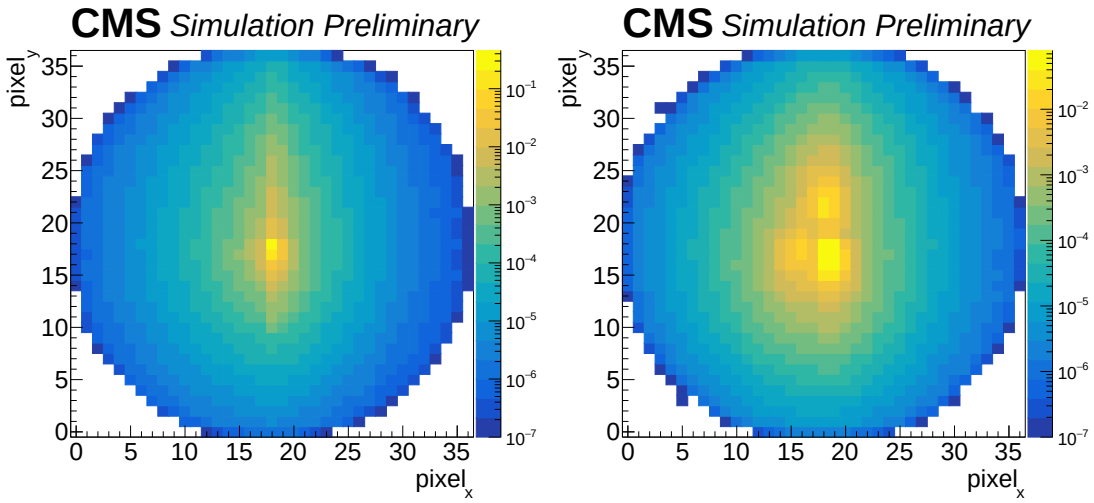

Figure 7: The pixelized greyscale images used in the ImageTop network for QCD (left) and top (right). The x and y axes are the pixel number, and roughly scale with ΔR. The z axis is the intensity of the greyscale image in the given pixel, (particle flow candidate p_T) and has been normalized to unity. This figure shows an ensemble of overlaid images after the image post processing, where we can see clear differences between the top and QCD energy deposition patterns.

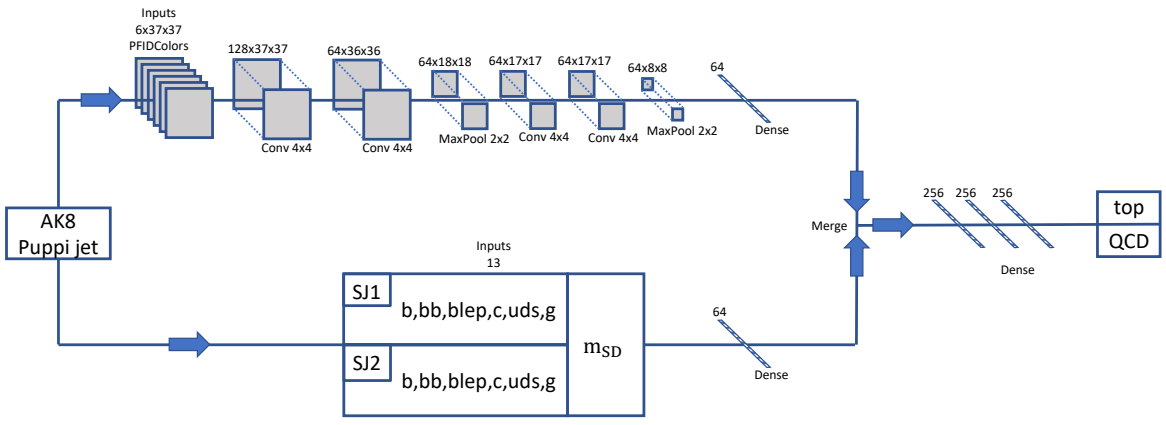

Figure 8: The ImageTop network architecture. The NN inputs are the 37x37 pixelized PF candidate p_T map, which is split into colors based on the PF candidate flavor, as well as the DeepFlavour subjet b tags applied to both subjets. The pixelized images are sent through a two dimensional CNN, and the subjet b tags are inputs into a dense layer. The NN are merged before being input into three dense layers and finally the two node output which is used as the top tagging discriminator.

matics, the displacement, and quality criteria. The SV list provides additional contributions to extracting features related to the heavy flavor content of the jet.

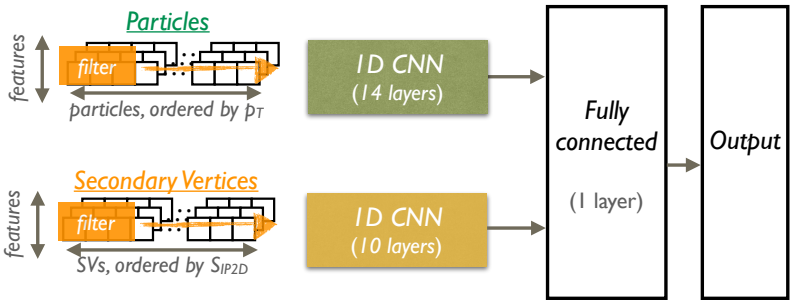

Figure 9: The network architecture of DeepAK8.

A significant challenge posed by the direct use of particle-level information is a substantial increase in the number of inputs. Meanwhile, the correlations between these inputs are of vital importance, thus an algorithm that can both process the inputs efficiently and exploit the correlations effectively is required. A customized DNN architecture is thus developed in DeepAK8 to fulfill this requirement. As illustrated in Figure 10, the architecture consists of two steps. In the first step, two one-dimensional CNNs are applied to the particle list and the SV list in parallel to transform the inputs and extract useful features. Then, in the second step, the outputs of these CNNs are combined and processed by a simple fully-connected network to perform the jet classification. The CNN structure in the first step is based on the ResNet model [90], but adapted from two-dimensional images to one-dimensional particle lists. The CNN for the particle list has 14 layers and the one for the SV list has 10 layers. A convolution window of length 3 is used, and the number of output channels in each convolutional layer ranges between 32 to 128. The ResNet architecture allows for an efficient training of deep CNNs, thus leading to a better exploitation of the correlations between the large inputs and improving the performance. The CNNs in the first step already contain strong discriminatory ability, so the fully-connected network in the second step consists of only one layer with 512 units, followed by a ReLU activation function and a Dropout [91] layer of 20% drop rate. The neural network is implemented using the MXNet package [92] and trained with the ADAM optimizer to minimize the cross-entropy loss. The initial learning rate is set to 0.001 and then reduced by a factor of 10 at the 10th and 20th epochs to improve convergence. The training is stopped after 35 epochs. A sample of 50 million jets is used, of which 80% are used for training and 20% are used for development and validation. Jets from different signal and background samples are reweighted to yield flat distributions in p_{T} to avoid any potential bias in the training process. The DeepAK8 algorithm is designed for jets with $p_{\mathrm{T}} > 200\,\mathrm{GeV}$ and typical operating regions which correspond to a misidentification rate great than 0.1%.

6.7.1 A mass-decorrelated version of DeepAK8

As it will be discussed in Section 7, background jets selected by the DeepAK8 algorithm exhibit a modified mass distribution similar to that of the signal. This is because the mass of a jet is one of the most discriminating variables, and although it is not directly used as an input to the algorithm, the CNNs are able to extract features that are correlated to the mass to improve the discrimination power. However, such modification of the mass distribution may be undesirable (as described in Ref. [15]) if the mass variable itself is in use for separating signal and background processes. Thus, an alternative DeepAK8 algorithm, "DeepAK8-MD", is developed to be largely decorrelated with the mass of a jet while preserving the discrimination

power as much as possible using an adversarial training approach [93]. Jets from different signal and background samples are also reweighted to yield flat distributions in both p_{T} and m_{SD} to aid the training.

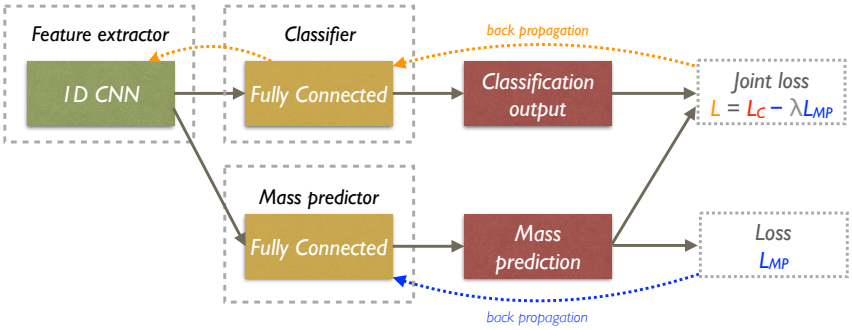

Figure 10: The network architecture of DeepAK8-MD.

The architecture of DeepAK8-MD is shown in Figure 10. Compared to the nominal version of DeepAK8, a mass prediction network is added with the goal of predicting the mass of a jet from the features extracted by the CNNs. When properly trained, the mass prediction network becomes a good indicator of how strongly the features extracted by the CNNs are correlated with the mass of a jet, as the stronger the correlation is, the more accurate the mass prediction will be. With the introduction of the mass prediction network, the training target of the algorithm can be modified to include the accuracy of the mass prediction as a penalty, therefore preventing the CNNs from extracting features that are correlated with the mass. In this way, the final prediction of the algorithm also becomes largely independent of the mass. As the features extracted by the CNNs evolve during the training process, the mass prediction network itself needs to be updated regularly to adapt to the changes of its inputs and remain as an effective indicator of mass correlation. Forcing the algorithm to be decorrelated with the jet mass, inevitably leads to a loss of discrimination power, and the resulted algorithm is a balance between performance and mass-independence. As the training of DeepAK8-MD is carried out only on jets with $30 < m_{\mathrm{SD}} < 250\,\mathrm{GeV}$, jets with m_{SD} outside this range should be removed when using DeepAK8-MD.

7 Performance in simulation

As presented in Section 6, a variety of algorithms has been developed by the CMS Collaboration to identify the hadronic decays of W/Z/H/t jets. To gain an initial understanding of the tagging performance and the complementarity between the different approaches, the algorithms were studied in simulated events. The performance of the algorithms is evaluated using the signal and background efficiency, ϵ_S and ϵ_B, respectively, as a figure of merit. The ϵ_S and ϵ_B are defined as:

$$\epsilon_S = \frac{N_S^{\text{tagged}}}{N_S^{\text{total}}} \quad \text{and} \quad \epsilon_B = \frac{N_B^{\text{tagged}}}{N_B^{\text{total}}}, \tag{13}$$

where $N_{S(B)}^{\text{tagged}}$ is the number of signal (background) jets satisfying the identification criteria of each algorithm, and $N_{S(B)}^{\text{total}}$ is the total number of generated particles considered to be signal (background). Hadronically decaying W/Z/H bosons or t quarks are considered to be signal, while quarks (excluding t quarks) and gluons from the QCD multijet process are considered to be background.

First, for each algorithm, the ϵ_B as a function of ϵ_S is evaluated in terms of a receiver operating characteristic (ROC) curve. Figures 11-14 summarize the ROC curves of all algorithms for the identification of t quarks and W/Z/H bosons, respectively. The comparisons are performed at low and high values of the truth particle p_{T}. The fiducial selection criteria applied to the truth particles are displayed on the plots. For the cut-based algorithms, namely $m_{\text{SD}} + \tau_{32}$, $m_{\text{SD}} + \tau_{32} + \text{b}$, $m_{\text{SD}} + \tau_{21}$, $m_{\text{SD}} + N_2$, and $m_{\text{SD}} + N_2^{\text{DDT}}$, all selections except the selection on τ_{32}, τ_{21}, or N_2, are applied as described in Sections 6.1 and 6.3.2.

In t-tagging, the addition of the subjet b tagging in the $m_{\text{SD}} + \tau_{32}$ algorithm reduces the misidentification probability for t quarks by up to $\sim 50\%$ depending on the p_{T}. The performance of the HOTVR algorithm lies between $m_{\text{SD}} + \tau_{32}$ and $m_{\text{SD}} + \tau_{32} + \text{b}$, whereas the $N_3 - \text{BDT}\ (\text{CA15})$ algorithm shows improved performance compared to the aforementioned algorithms, particularly in the low p_{T} range. The improved performance stems from the usage of the ECFs, which provide complementary information to τ_{32}. Particularly in the low-p_{T} region, the gain is mainly due to the use of larger-cone jets (i.e. jets clustered with $R = 1.5$). The BEST algorithm targets the high-p_{T} regime and shows similar performance to the ECF algorithm in this regime. The best discrimination is achieved with algorithms based on lower level information, namely the ImageTop and DeepAK8 algorithms. ImageTop and DeepAK8-MD yield comparable performance in the low and high p_{T} regions, whereas the optimal performance in terms of ROC curves is achieved with the nominal version of DeepAK8 over the entire p_{T} region.

Various arguments contribute to the significantly improved performance of ImageTop and DeepAK8 with respect to the other algorithms. First, the usage of lower-level variables as inputs to the network allows better exploitation of the high granularity of the CMS detector. The architectures of these algorithms allows for usage of quark-gluon discrimination information. Moreover, information about the jet flavor content is extracted, which is particularly important for t quark and Z/H identification. The flavor identification in boosted jets is very challenging since the decay products overlap and traditional b tagging algorithms are significantly less performant. The usage of the flavor of the PF candidates, and the secondary vertices in the case of DeepAK8, allows for a more precise description of the flavor content inside the jet.

Similar conclusions hold for the identification of hadronically decaying W and Z bosons. The BEST, DeepAK8, and DeepAK8-MD algorithms show improved performance compared to the simpler $m_{\text{SD}} + \tau_{21}$ algorithm. The gain in terms of misidentification rate can be as large as

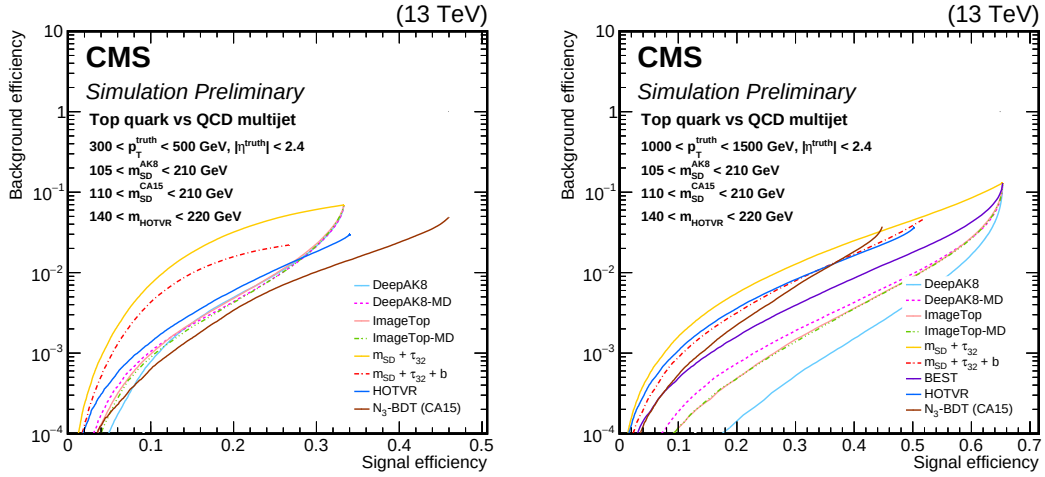

Figure 11: Performance comparison of the hadronically decaying t quark identification algorithms in terms of receiver operating characteristic (ROC) curves in two regions based on the p_T of the truth particle; Left: $300 < p_T < 500$ GeV, and Right: $1000 < p_T < 1500$ GeV. Additional fiducial selection criteria applied to the jets are displayed on the plots.

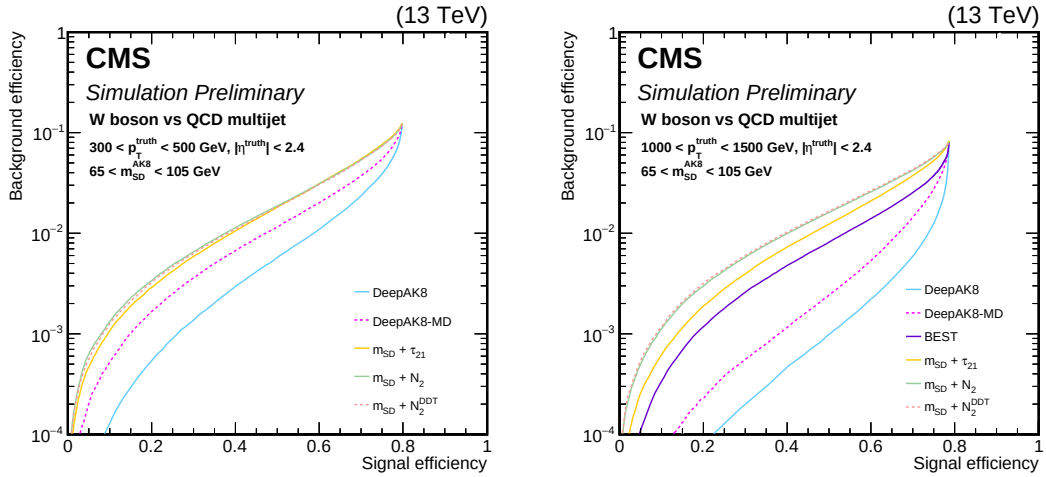

Figure 12: Performance comparison of the hadronically decaying W boson identification algorithms in terms of receiver operating characteristic (ROC) curves in two regions based on the p_T of the truth particle; Left: $300 < p_T < 500$ GeV, and Right: $1000 < p_T < 1500$ GeV. Additional fiducial selection criteria applied to the jets are displayed on the plots.

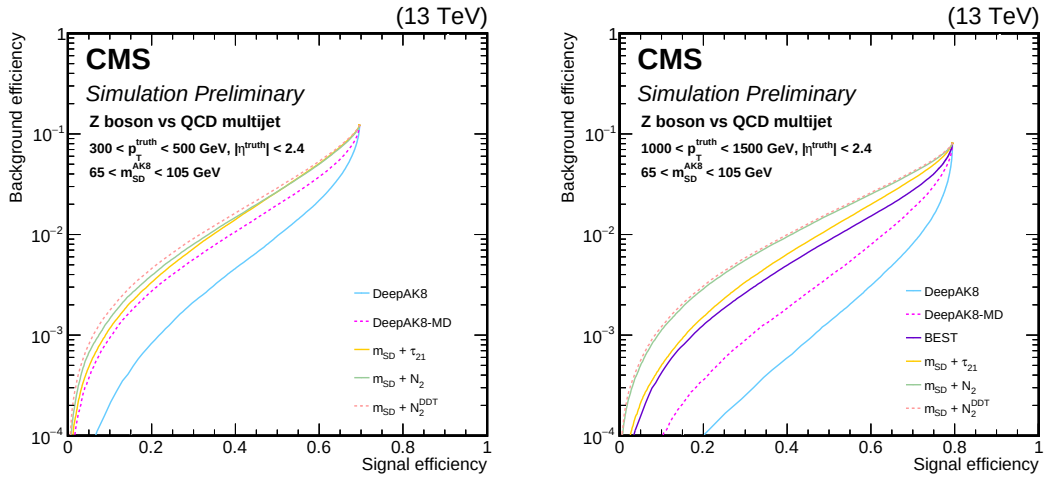

Figure 13: Performance comparison of the hadronically decaying Z boson identification algorithms in terms of receiver operating characteristic (ROC) curves in two regions based on the p_T of the truth particle; Left: $300 < p_T < 500$ GeV, and Right: $1000 < p_T < 1500$ GeV. Additional fiducial selection criteria applied to the jets are displayed on the plots.

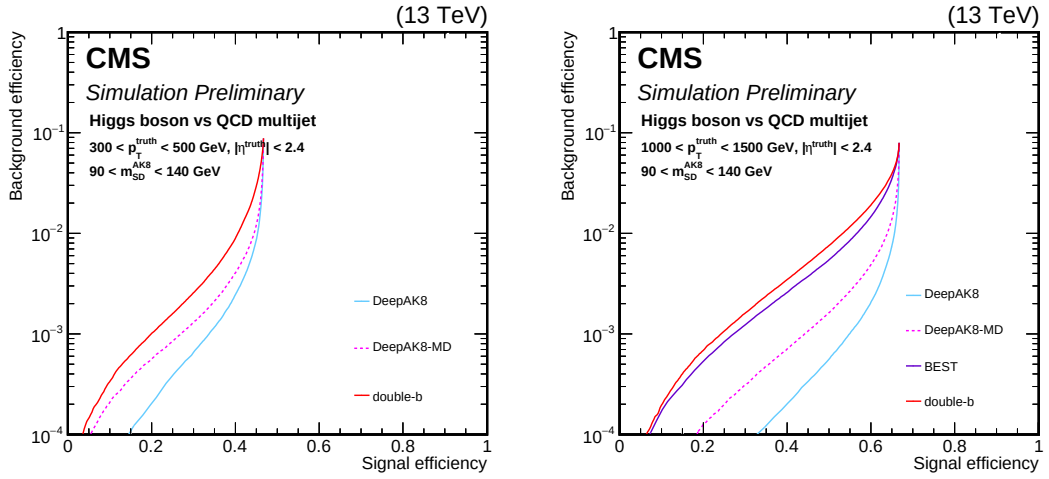

Figure 14: Performance comparison of the hadronically decaying H boson identification algorithms in terms of receiver operating characteristic (ROC) curves in two regions based on the p_T of the truth particle; Left: $300 < p_T < 500$ GeV, and Right: $1000 < p_T < 1500$ GeV. The H boson is forced to decay in a pair of b quarks. Additional fiducial selection criteria applied to the jets are displayed on the plots.

an order of magnitude in the case of DeepAK8. The smaller relative gain of DeepAK8 over BEST between the W or Z bosons, and the t quark identification is mainly explained by the fact that for the first two, flavor information is not as critical as for the latter. The $m_{\mathrm{SD}} + N_2$ and $m_{\mathrm{SD}} + N_2^{\mathrm{DDT}}$ show weaker performance compared to the $m_{\mathrm{SD}} + \tau_{21}$ algorithm.

Lastly, the double-b, BEST, DeepAK8, and DeepAK8-MD algorithms can be used to identify hadronic decays of the H boson. In Fig. 14 the H is forced to decay to a pair of b quarks. The double-b algorithm lies between BEST and DeepAK8. The gain for DeepAK8 is expected for similar arguments as for the t quark identification.

To gain a deeper understanding of the DeepAK8 performance, alternative versions of DeepAK8 were trained using a subset of the input features. Three sets of input features were studied and compared. The "Particle (kinematics)" set consists of only the kinematic information of the PF candidates, e.g., the four momenta, the distances to the jet and subjet axes. This set serves as a baseline to evaluate the performance using only substructure of the jets. The "Particle (w/o Flavour)" set includes additional experimental information of each PF candidate, such as the electric charge, particle identification and track quality information. Compared to the nominal DeepAK8 algorithm, input features that contribute to the identification of heavy-flavor quarks, such as the displacement of the tracks, the association of tracks to the reconstructed vertices, as well as the SV features, are not included in the "Particle (w/o Flavour)" set. The performances of the three versions of DeepAK8 are compared in Fig. 15 for top and Z identification. In both cases, the addition of experimental information brings sizable improvement in performance. While the additional features contributing to heavy flavor identification leads to no improvement for the identification of Z bosons decaying to a pair of light quarks, a significant improvement is observed for Z decaying to a pair of b quarks, as well as t quarks, showing the strong complementarity between heavy flavor identification and jet substructure for heavy resonance identification where heavy flavor quarks are involved in the decay.

7.1 Robustness of tagging algorithms

In addition to the performance of the algorithms in pure discrimination, an important ingredient is their robustness to changes in jet kinematics and data-taking conditions. To quantify this, we study the ϵ_S and ϵ_B of the algorithms as a function of the p_{T} of the truth particle and the number of reconstructed vertices ($\mathrm{N_{vtx}}$). For the sake of these studies, a common working point is defined, corresponding to $\epsilon_S = 30\%(50\%)$ for t quark (W, Z, and H boson) with $500 < p_{\mathrm{T}}(\text{truth particle}) < 600\,\mathrm{GeV}$. Working points used in CMS analyses are typically optimized to achieve the best sensitivity for the targeted signal processes, therefore vary from analysis to analysis. For example, CMS employs a top tagging working point at approximately 40% signal efficiency in the search for BSM $\mathrm{t\bar{t}}$ production [54], a W tagging working point at approximately 20% signal efficiency in the search for BSM diboson production [64], and a H tagging working point at approximately 30% signal efficiency in the search for di-Higgs production [94].

The distributions of the ϵ_S and ϵ_B as a function of the p_{T} of the truth particle for the different particle identification scenarios are displayed in Figs. 16 and 17, respectively. In the low-p_{T} range for the t-tagging case, the ϵ_S for the algorithms using AK8 jets increases rapidly until $p_{\mathrm{T}} \gtrsim 600\,\mathrm{GeV}$, where a sufficient fraction of jets contain all the t decay products. As expected, the $N_3 - \mathrm{BDT}$ (CA15) and HOTVR algorithms have a stable ϵ_S as a function of the truth particle p_{T}. Similar behavior is observed for the t quark misidentification rate.

In the case of the W and Z boson tagging, the ϵ_S for the $m_{\mathrm{SD}} + \tau_{21}$ algorithm decreases as a function of $p_{\mathrm{T}}(\text{truth particle})$, whereas for the BEST, DeepAK8, and DeepAK8-MD algorithms

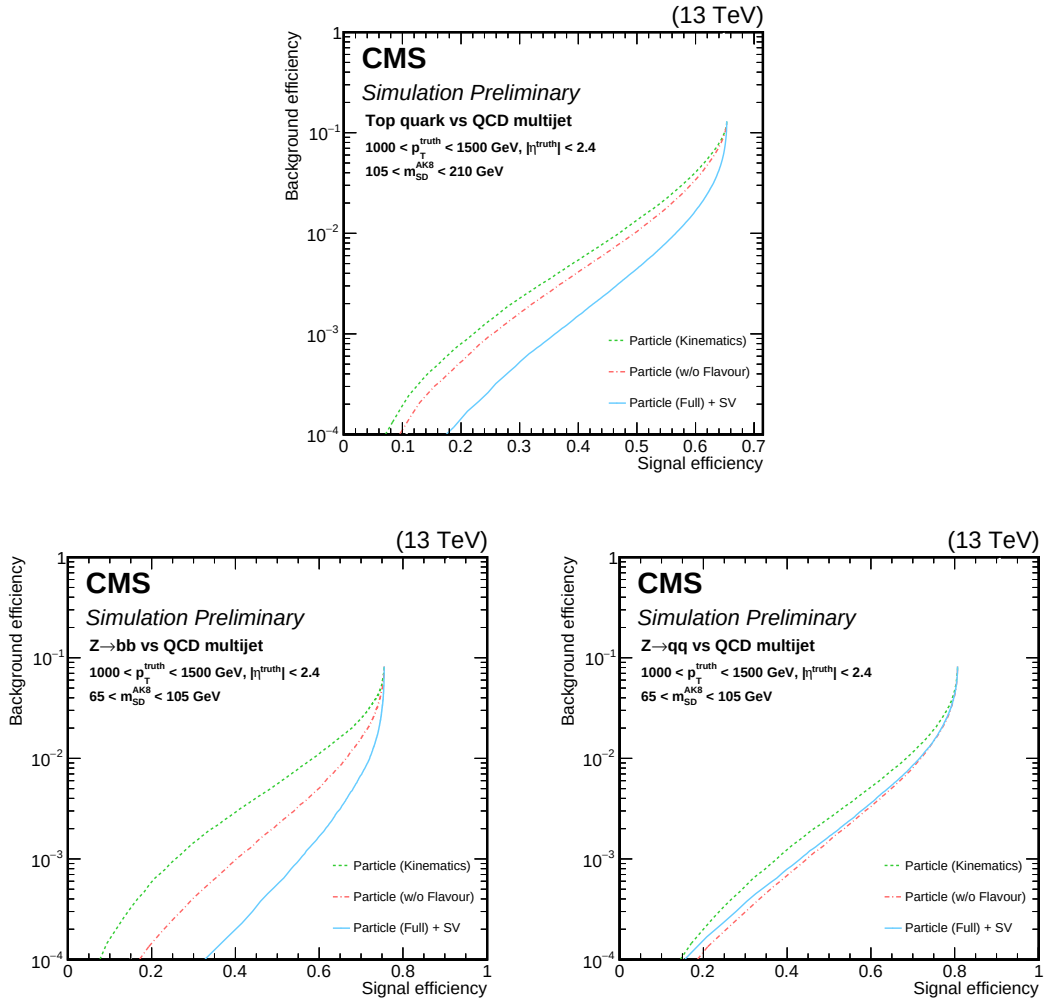

Figure 15: Alternative versions of DeepAK8 trained using a subset of the input features. The details about each version are discussed in the text. The performances of the three versions of DeepAK8 are compared for t quark (upper) and Z (lower) identification. For the latter, the left plot corresponds Z bosons decaying to a pair of b quarks, and the right to a pair of light quarks.

exhibit improvements in ϵ_S as a function of p_T(truth particle). The drop in ϵ_S for $m_{SD} + \tau_{21}$ is a result of the correlation that $m_{SD} + \tau_{21}$ has with the jet p_T, leading to a shift in the jet mass distribution to higher values. The $m_{SD} + N_2$ shows similar behavior to BEST and DeepAK8, while the ϵ_S in the case of $m_{SD} + N_2^{DDT}$ is stable as a function of p_T(truth particle). In contrast to N-subjetiness, the ECF observable uses an axis-free approach, which is more efficient in the case of highly collimated decay products.

The misidentification rate has a non trivial behavior for most algorithms. In the case of DeepAK8 and DeepAK8-MD the ϵ_B decreases with p_T(truth particle), which is mainly a result of the use of low level features as inputs to the algorithm. For $m_{SD} + N_2$, the ϵ_B increases with p_T(truth particle), whereas for $m_{SD} + N_2^{DDT}$, is, by design, significantly more stable. In the case of $m_{SD} + \tau_{21}$, the decrease of ϵ_B as function of p_T(truth particle) is mainly due to the strong shift of the m_{SD} shape of the background jets to larger values due the selection on τ_{21}. This will be discussed in more detail in Section 7.2. Finally, for BEST the ϵ_B decreases up to p_T(truth particle) $\sim$ 1000 GeV, and then increase again. This is a feature of the training of the BEST algorithm, stemming from imbalance in the relative fraction of jets between the low and high p_T regimes.

In the case of H tagging, BEST and the DeepAK8 algorithms have stable ϵ_S for p_T(truth particle) $\gtrsim$ 600 GeV, whereas for double-b the ϵ_S starts to decrease around this p_T regime. There are two main arguments for this behavior. First, double-b exploits axis-dependent observables, similar to τ_{21}, which are less efficient at high p_T where the decay products become highly collimated. Second, the selection on the tracks used to construct the variables used for the training of double-b, discussed in Section 6.4, is suboptimal in the very high-p_T regime. The ϵ_B for both double-b and DeepAK8 decreases as a function of p_T(truth particle), whereas for BEST shows a modest increase for p_T(truth particle) $\gtrsim$ 1000 GeV, for the same arguments as in the W and Z case.

The dependence of the algorithms on N_{vtx} is also examined using simulated events. Figure 18 (19) displays the distribution of ϵ_S (ϵ_B) as a function of N_{vtx} for truth particles with 500 $<$ p_T(truth particle) $<$ 1000 GeV, operating at a working point with $\epsilon_S = 30\%$ ($\epsilon_S = 50\%$) for t quark (W, Z, and H boson) identification as defined above. The algorithms make use of jets that exploit PUPPI for pileup mitigation, which results to a roughly constant ϵ_S and ϵ_B across the different pileup scenarios.

7.2 Correlation with jet mass

Finally, a set of studies was performed to understand the correlation of the algorithms with the jet mass. This is an essential step in order to benefit from the theoretical progress made in jet substructure [3], which can result in reduced systematic uncertainties in analyses [15]. The jet mass is one of the most discriminating variables, and many analyses require a smoothly falling background jet mass spectrum under a signal peak (for instance, in Ref. [95]). Figure 20 displays the shape of the m_{SD} distribution for jets obtained from the QCD multijet sample, inclusively and after applying a selection on each algorithm. The working point chosen corresponds to $\epsilon_S = 30\%$ ($\epsilon_S = 50\%$) for t quark (W, Z, and H boson). The results are displayed for one p_T region of the truth particle distribution, but the conclusions hold for other p_T regions as well. By design, the BEST and the nominal version of the DeepAK8 algorithms feature significant sculpting of the background jet mass shape. For analyses that do not explicitly use the jet mass distribution for signal extraction, this is not problematic.

To quantify the level of mass sculpting we use the Jensen-Shannon divergence [96] (JSD), which is a symmetrized version of the Kullback-Leibler divergence [97] (KLD), and provides a metric

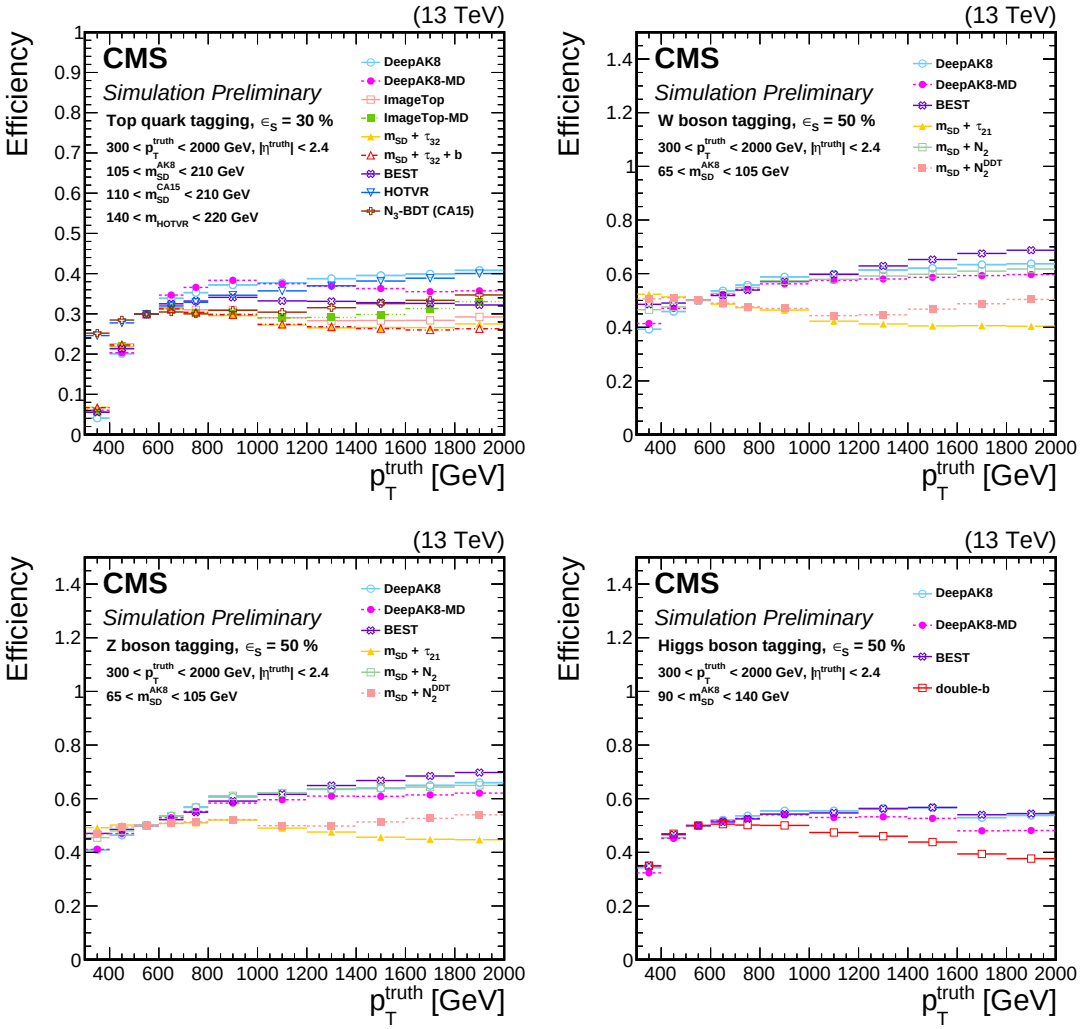

Figure 16: The distribution of ϵ_S as a function of the p_T of the truth particle for a working point corresponding to $\epsilon_S = 30\%(50\%)$ for t quark (W, Z, and H boson) identification. Upper left: t quark, upper right: W boson, lower left: Z boson, lower right: H boson. The error bars represent the statistical uncertainty in each specific bin, due to the limited number of simulated events. Additional fiducial selection criteria applied to the jets are displayed on the plots.

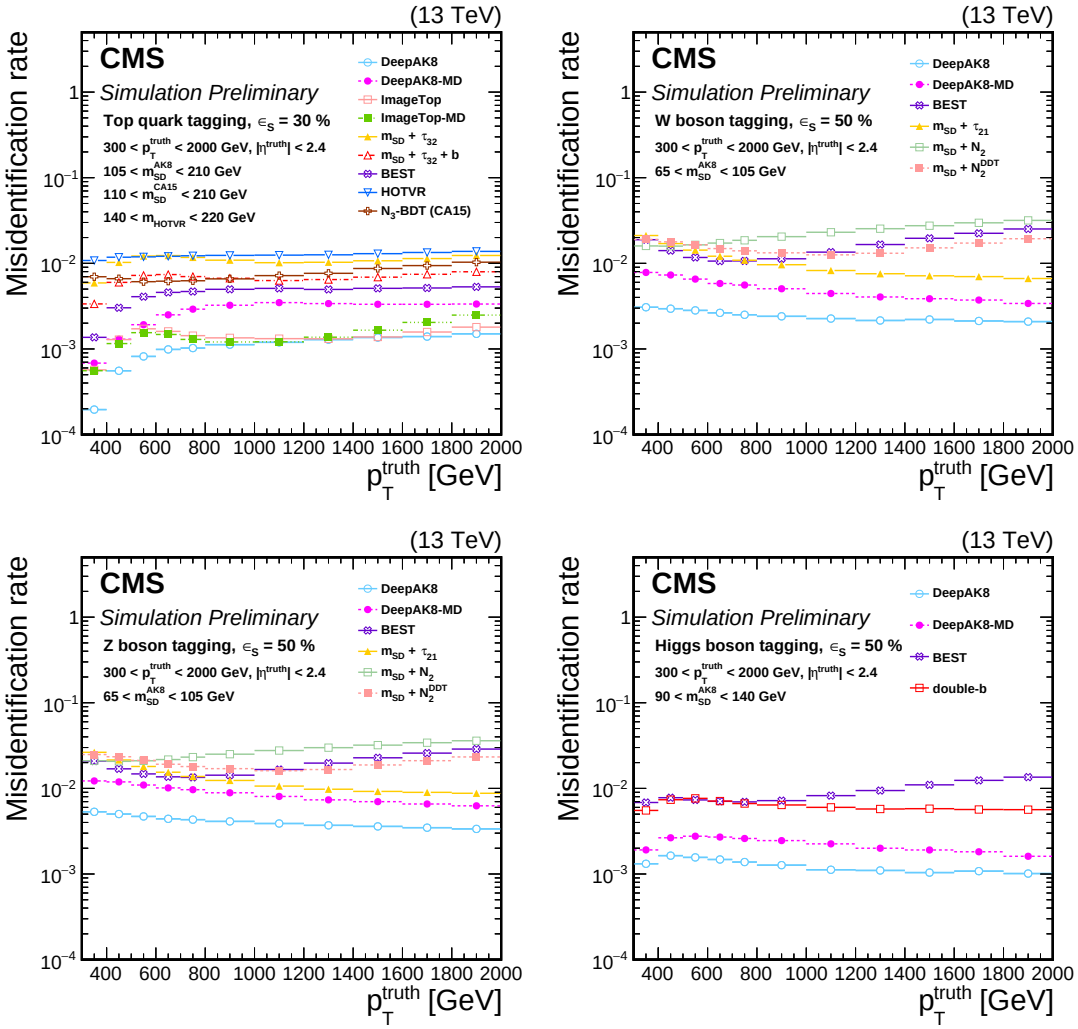

Figure 17: The distribution of ϵ_B as a function of the p_T of the truth particle for a working point corresponding to $\epsilon_S = 30\%(50\%)$ for t quark (W, Z, and H boson) identification. Upper left: t quark, upper right: W boson, lower left: Z boson, lower right: H boson. The error bars represent the statistical uncertainty in each specific bin, due to the limited number of simulated events. Additional fiducial selection criteria applied to the jets are displayed on the plots.

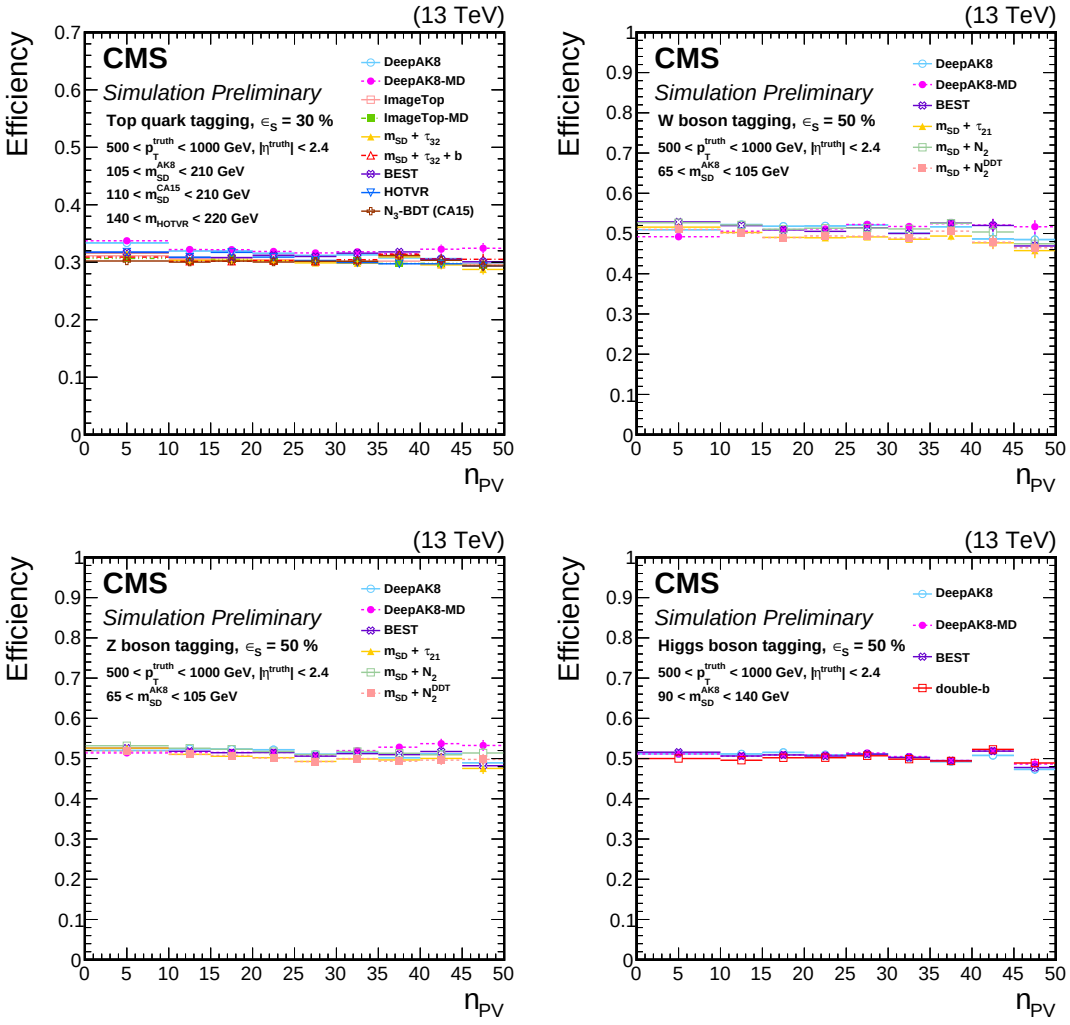

Figure 18: The ϵ_S as a function of N_{vtx} for truth particles with $500 < p_T(\text{truth particle}) < 1000$ GeV at a working point corresponding to $\epsilon_S = 30\%(50\%)$ for t quark (W, Z, and H boson) identification. Upper left: t quark, upper right: W boson, lower left: Z boson, lower right: H boson. The error bars represent the statistical uncertainty in each specific bin, due to a limited number of simulated events. Additional fiducial selection criteria applied to the jets are displayed on the plots.

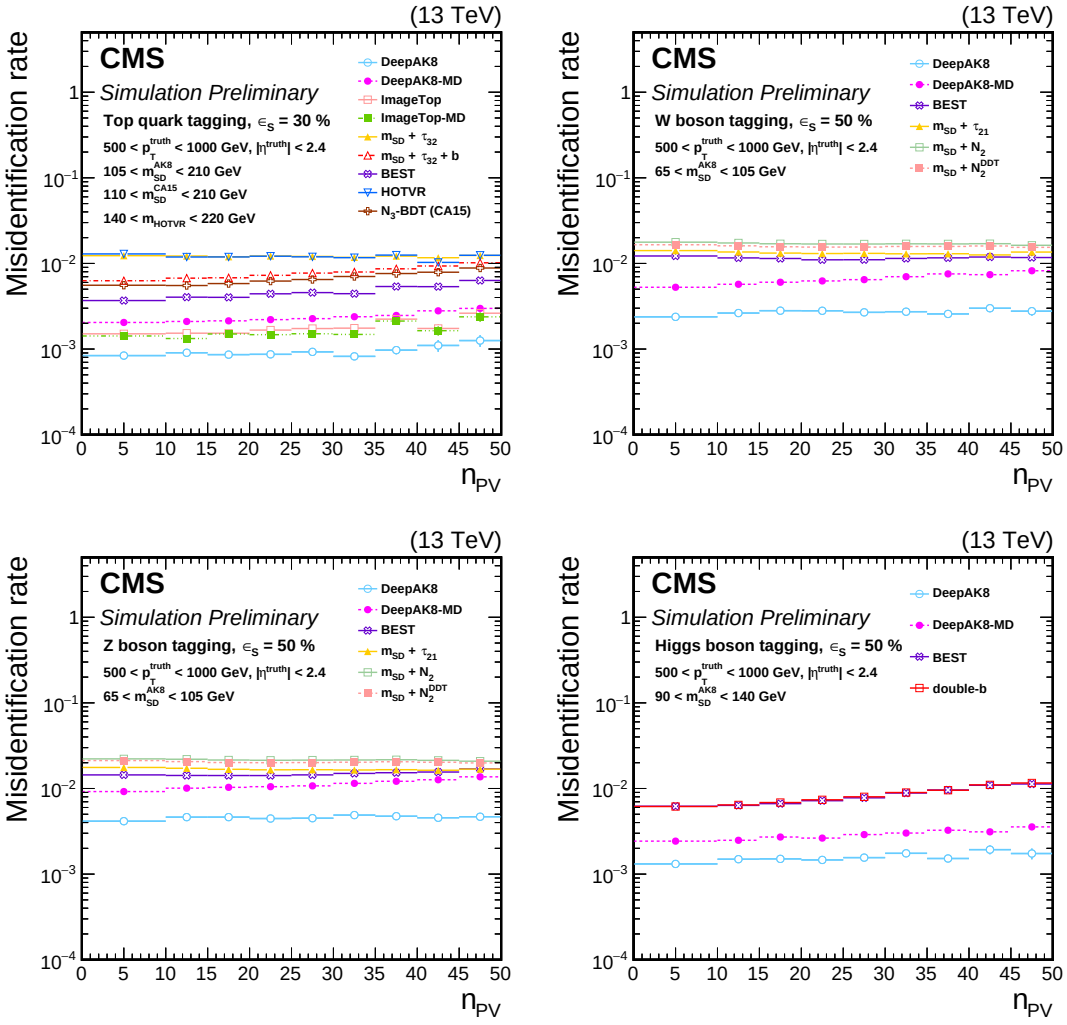

Figure 19: The ϵ_B as a function of N_{vtx} for truth particles with $500 < p_T(\text{truth particle}) < 1000$ GeV at a working point corresponding to $\epsilon_S = 30\%(50\%)$ for t quark (W, Z, and H boson) identification. Upper left: t quark, upper right: W boson, lower left: Z boson, lower right: H boson. The error bars represent the statistical uncertainty in each specific bin, due to the limited number of simulated events. Additional fiducial selection criteria applied to the jets are displayed on the plots.

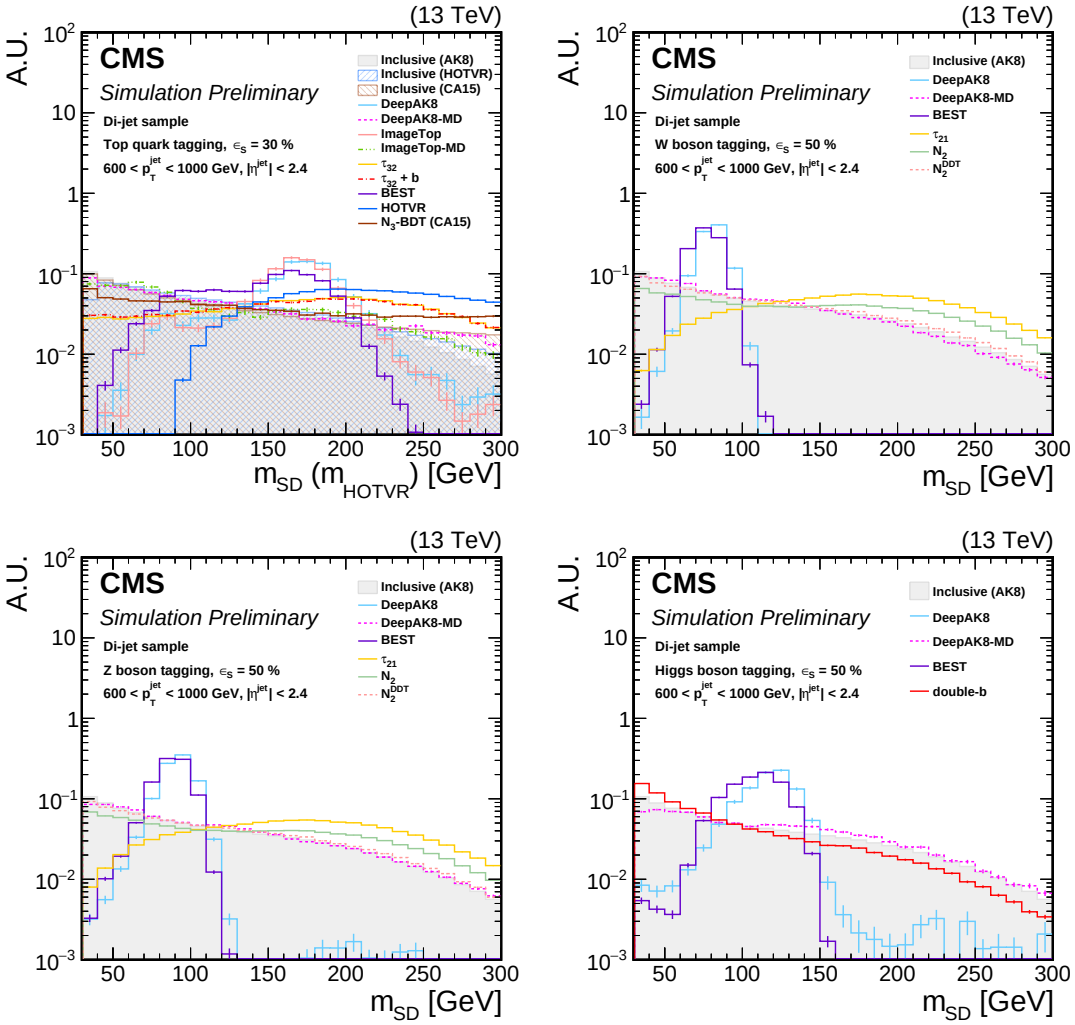

Figure 20: The shape of the softdrop mass distribution for background jets with 600 < p_T(jet) < 1000 GeV, inclusively and after selection by each algorithm. The working point chosen corresponds to $\epsilon_S = 30\%$ ($\epsilon_S = 50\%$) for t quarks (W, Z, and H bosons). Upper left: t quark, upper right: W boson, lower left: Z boson, lower right: H boson. The error bars represent the statistical uncertainty in each specific bin, due to the limited number of simulated events. Additional fiducial selection criteria applied to the jets are displayed on the plots.

for the similarity of the shape between distributions. The KLD is defined as:

$$\mathrm{KLD}(P||Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)}\,, \tag{14}$$

where $P(i)$ and $Q(i)$ are the normalized mass distributions of the background jets that fail and pass a selection on a given algorithm, respectively. The index "i" runs over the bins of the distributions. The jet mass distributions lay between 30 and 300 GeV with a bin size of 10 GeV.

The JSD metric is defined as:

$$\mathrm{JSD}(P||Q) = \frac{1}{2}(\mathrm{KLD}(P||M)) + \mathrm{KLD}(Q||M)), \text{ where } M = \frac{P+Q}{2}\,. \tag{15}$$

Lower values of JSD indicate larger similarity between the mass distributions of jets passing and failing a selection on a given algorithm.

The JSD values for successively tighter selections (expressed in terms of ϵ_B) on the various t- and W-tagging algorithms are shown in Fig 21. The best decorrelation for the t-tagging case is achieved with the DeepAK8-MD algorithm, which exploits an adversarial network to reduce the correlation of the tagging score with the jet mass. For W-tagging, $m_{\mathrm{SD}} + N_2^{\mathrm{DDT}}$ and DeepAK8-MD achieve similar level of mass decorrelation. As expected, tighter selection on the tagging score results in an increase of the mass sculpting. A similar behavior is observed for all algorithms.

The robustness of the mass decorrelation techniques was further studied as a function of the jet p_{T} and as a function of $\mathrm{N_{vtx}}$. These studies are carried out for a working point corresponding to $\epsilon_S = 50\%$ and $\epsilon_S = 30\%$ for t- and W-tagging, respectively. Figure 22 shows the JSD values as a function of the jet p_{T} for jets from QCD multijet events. The majority of the algorithms show modest dependence on jet p_{T}, except for ImageTop-MD, where the mass dependence increases rapidly when $p_{\mathrm{T}} \lesssim 600$ GeV as the training was performed only for jets with $p_{\mathrm{T}} > 600$ GeV. The DeepAK8-MD and $m_{\mathrm{SD}} + N_2^{\mathrm{DDT}}$ for W-tagging also show modestly increased mass dependence in the p_{T} range of $p_{\mathrm{T}} \gtrsim 1200$ and $p_{\mathrm{T}} \gtrsim 1600$, respectively. The dependence of the mass mitigation techniques on $\mathrm{N_{vtx}}$ was also studied and was found to be small across the different $\mathrm{N_{vtx}}$ regions.

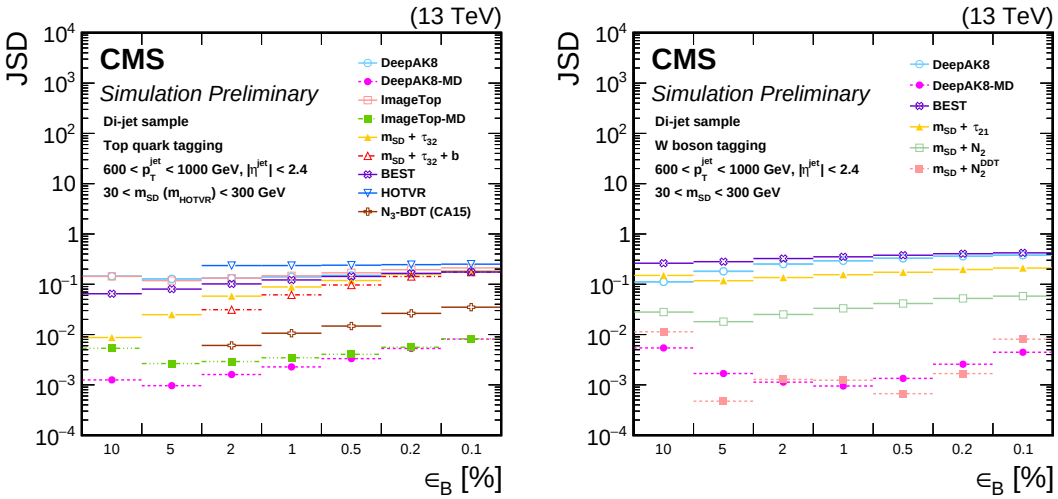

Figure 21: The JSD as a function of successively tighter selections (expressed in terms of ϵ_B) for the various t- (left) and W (right) tagging algorithms. Lower values of JSD indicate larger similarity of the M_{SD} in QCD multijet events passing and failing the selection on the tagging algorithm. Additional fiducial selection criteria applied to the jets are displayed on the plots.

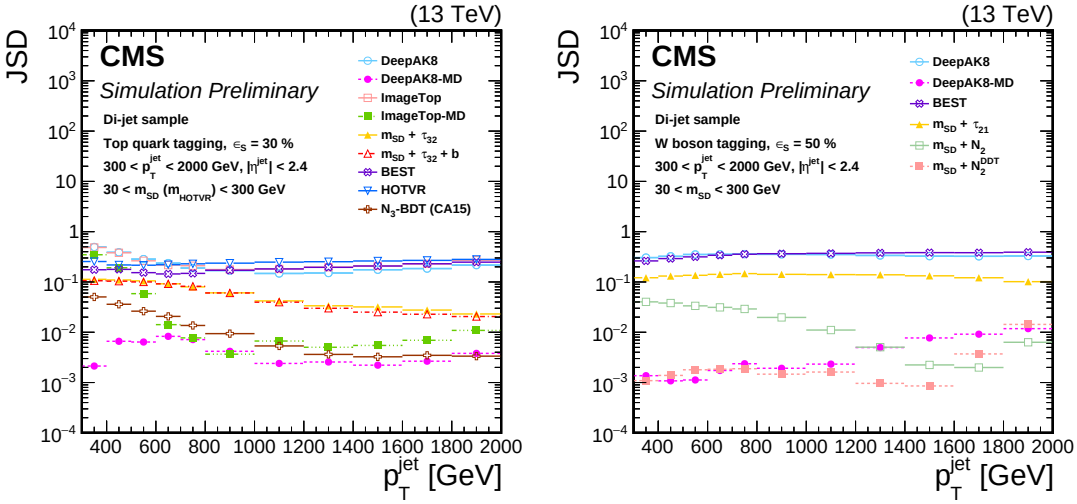

Figure 22: The JSD as a function of the jet p_T for the various t (left) and W (right) tagging algorithms. Lower values of JSD indicate larger similarity of the M_{SD} in QCD multijet events passing and failing the selection on the tagging algorithm. Additional fiducial selection criteria applied to the jets are displayed on the plots.

8 Performance in data and systematic uncertainties

In this section, the validation of the performance of the algorithms in data is presented. The validation is performed in two steps. In the first step, we focus on studying the overall modeling of key variables in simulation and their agreement with data, as well as the dependence on the simulation details. The second step is to use these results to extract corrections to the simulation so their performance matches that in data. Differences in the performance between data and simulation are taken into account by means of scale factors (SF) extracted by comparing the efficiencies in data and simulation. To account for effects not captured in the SF, multiple sources of systematic uncertainties are considered. The data and simulated samples used for these studies are described in Section 5.

In this note we focus on the calibration of the t quark and W/Z boson tagging algorithms. The calibration of Z and H boson tagging algorithms where Z and H decay to a pair of bottom or charm quarks requires alternative methods that go beyond the scope of this note. In a nutshell, since with the current luminosity is challenging to obtain a pure Z or H sample, the calibration of such taggers relies on the use of proxy jets. Data-to-simulation correction factors are extracted based on these proxy jets, which are then applied on signal jets. Therefore, the proxy jets should be selected such to have similar characteristics to the signal jets. To this end, jets arising from gluon splitting to $\mathrm{b\bar{b}}$ or $\mathrm{c\bar{c}}$ are used as proxy jets from a sample dominated by QCD multijet events. Such approaches have been followed in Refs. [77, 78, 98], and this is a point to expand in a future publication.

8.1 Systematic uncertainties

A number of sources of systematic uncertainties can affect the modeling of the performance of the algorithms in data by the simulation. These include systematic uncertainties in the parton showering model, renormalization and factorization scale, PDF, jet energy scale and resolution, $p_{\mathrm{T}}^{\mathrm{miss}}$ unclustered energy, trigger and lepton identification, pileup, and luminosity, as well as statistical uncertainties in both simulation and data.

Parton shower uncertainties for signal jets are evaluated using samples with the same event generator but different choice for the modeling of the parton showering. For background jets, a sample produced using alternative generator for both the hard-scatter and the parton shower is utilized. Details about the samples used can be found in Section 3. Changes in renormalization (μ_R) and factorization (μ_F) scales are estimated by varying μ_R and μ_F separately by a factor of two relative to the choices of these scale values used in the sample generation. The uncertainty related to the choice of PDF is obtained from the standard deviation in 100 variations of the NNPDF3.0 PDF set [29]. The jet energy scale and resolution are changed within their $p_{\mathrm{T}}-$ and $\eta-$ dependent uncertainties based on the studies presented in [48]. Their effect is also propagated to $p_{\mathrm{T}}^{\mathrm{miss}}$. The effect of the uncertainty in the measurement of the unclustered energy (i.e. contribution of PF candidates not associated to any of the physics objects) is evaluated based on the momentum resolution of each PF candidate, which depends on the type of the candidate [50]. Uncertainties on the measurement of the trigger efficiency and on the energy scale and resolution of the leptons are propagated in the SF extraction. The uncertainty in the pileup reweighting procedure is determined by varying the minimum bias cross section used to produce the pileup profile away from the measured central value of 69.2 mb [99, 100] by $\pm$5%. The limited size of the simulated samples and the size of the data control samples are also considered.

The uncertainties described above contribute in different ways to the modeling of the jet kinematics and the extraction of the SF. For example, the trigger and lepton identification uncer-

tainties are a few percent, and do not include uncertainties on the kinematic distribution. The identification of leptons, especially muons, is nearly fully efficient, and the trigger selection is chosen in order to ensure full efficiency in the regime of interest. The jet energy scale and resolution uncertainties are similar, where shape components are included, and are between 1–5% for the high-p_{T} jets studied here. Uncertainties related to pileup and the luminosity measurement have an effect smaller than $\sim 3\%$.

As many of the algorithms detailed in this note use jet substructure and jet constituent information, either directly, or as input to multivariate techniques, the uncertainties in the choice of parton shower is significant. Different parton showers directly affect the number, momentum, and distribution of jet constituents, influencing the observables used as inputs to the multivariate techniques, and eventually propagating to the outputs of those algorithms. The magnitude of this source of systematic uncertainty lies in the range of 10–30%. The uncertainty in the value of the renormalization and factorization scales chosen for event generation also has a sizable impact (5–15%), as this changes the amount of radiation that can enter into a reconstructed jet. In total, these dominant components contribute a total combined uncertainty in the range of 10–50%, depending on the specific jet kinematics of interest.

Nevertheless, these uncertainties partially cancel in the SF measurement, as will be discussed in Section 8.4.

8.2 The t quark and W boson identification performance in data

The single-μ event selection discussed in Section 5.1 provides a sample dominated by $\mathrm{t\bar{t}}(1\ell)$ events. This selection has a high fraction of events with leptonically decaying W bosons that decay from one of the t quarks, providing a sample with a high purity of $\mathrm{t\bar{t}}$ events, whereas the other t quark (which decays hadronically) provides boosted t quarks and W bosons, which are used for the validation of the algorithms.

In order to study possible dependence of the tagging efficiency on the parton showering scheme, we consider two alternative simulated $\mathrm{t\bar{t}}$ samples. As discussed in Section 3, both samples are generated with the same generator (i.e. POWHEG), but one uses PYTHIA for the modeling of the parton showering, whereas the other HERWIG++. The total SM expectation from simulation using the latter $\mathrm{t\bar{t}}$ sample will be referred to as "SM (Herwig)" in what follows. It will be seen that the choice of the parton showering generator has a small impact on the overall agreement between data and simulation in signal jets.

To account for the differences in the design of the algorithms, the large-R jets discussed in Section 5.1 are either AK8, CA15, or HOTVR jets. For the sake of brevity we will focus mainly on results using AK8 jets, unless stated otherwise. Nevertheless, the conclusions from the validation in data are similar between the three jet collections.

The data-to-simulation comparison of fundamental jet substructure variables, such as m_{SD}, the $p_{\mathrm{T}}(\mathrm{jet})$, the N-subjetiness ratios, τ_{32} and τ_{21}, and the N_2 and N_2^{DDT}, are shown in Fig. 23. These have all been measured by the CMS Collaboration as noted above. A second set of comparisons is related to the main observables of the HOTVR algorithm. Figure 24 displays the distributions of m_{HOTVR}, $m_{\mathrm{min,HOTVR}}$ and $N_{\mathrm{sub,HOTVR}}$ in data and simulation. The next set of comparisons includes tagging algorithms that are based on high level jet substructure observables and explore ML techniques to improve performance, namely the BEST and the N_3 − BDT (CA15) algorithms. Figure 25 shows the t quark and W boson identification probabilities of BEST and the t-tagging discriminant for the N_3 − BDT (CA15), in data and simulation. The last set of comparisons is related to the ImageTop and the DeepAK8 algorithms, which both explore lower-level

observables. Figure 26 displays the distributions of the t quark identification probability for the two versions of ImageTop, and the t quark and W boson identification probabilities for DeepAK8 algorithms.

As the selection applied to events shown in Figs. 23-26 results in a sample with a small purity of fully merged t quarks, we also study the same distributions after applying a stricter requirement on the jet momenta: $p_T > 500\,\text{GeV}$. This selection results in a sample consisting of a higher fraction of fully merged t quark jets, relative to the boosted W boson jet component. Figures 27-30 show the same distributions for this high-p_T selection.

To account for effects related to differences in the overall normalization between data and simulation, the total background yield is normalized to the observed number of data events. The systematic uncertainties discussed in Section 8.1 are also considered and are shown via the shaded blue band in the figures. Overall, the shapes in data are compatible with the expectation from simulation within uncertainties for all the algorithms.

8.3 Misidentification probability in data

The misidentification probability of the algorithms is studied in the di-jet and single-γ data samples. The two samples differ in the relative fraction of light quarks and gluons in the final state. In order to study the dependence of the misidentification probability on the choice of the event generator and the parton showering scheme, we consider two different simulated samples to model the QCD multijet background. The nominal sample uses MADGRAPH for the event generation and PYTHIA 8 (P8) for the parton showering and hadronization, whereas the alternative sample uses HERWIG++ for event generation and the modeling of the parton showering. More information on the generation details on these samples are discussed in Section 3. The total SM estimated using the HERWIG++ QCD multijet sample will be referred to as "SM (Herwig)". Similarly to Section 8.2, we will focus on results using jets with $R = 0.8$, unless stated otherwise. To account for possible differences in the p_T distribution of the QCD multijet and γ+jet simulated events, the total background yield is reweighted to match the p_T distribution in data, following the procedure discussed in Section 3.

The distribution of m_{SD}, $p_T(\text{jet})$, the N-subjetiness ratios τ_{32} and τ_{21}, and the N_2 and N_2^{DDT}, in the di-jet sample are displayed in Fig. 31. For this event selection, the shape of m_{SD} and the N-subjetiness ratios are described well by simulation, whereas there is disagreement between data and simulation for high values of N_2 and N_2^{DDT}. A better description of the data, particularly for N_2^{DDT}, is achieved with the HERWIG++ QCD multijet sample, which hints that the disagreement is related to the description of the parton shower. For the other observables we observe similar level of agreement between the two generators.

The same set of variables is presented in Fig. 32 for the sample. From previous measurements [8], the m_{SD} is observed to agree very well with simulation except at low masses. The modeling of the N-subjetiness and N_2 ratios is poorer in the single-γ sample.

Figures 33 and 34 show the distribution of the main observables of the HOTVR algorithm, namely m_{HOTVR}, $m_{min,HOTVR}$ and $N_{sub,HOTVR}$, in data and simulation, in the di-jet and single-γ samples, respectively. In both samples m_{HOTVR} and $m_{min,HOTVR}$ shows good agreement between data and simulation. The $N_{sub,HOTVR}$ distribution in data is softer compared to simulation. Similar conclusions hold using HERWIG++ to simulate the QCD multijet events. The difference is more pronounced in the single-γ sample. The $N_{sub,HOTVR}$ is particularly sensitive to the precise modeling of the parton showering.

The distribution of the t quark and W boson identification probabilities for BEST and the top

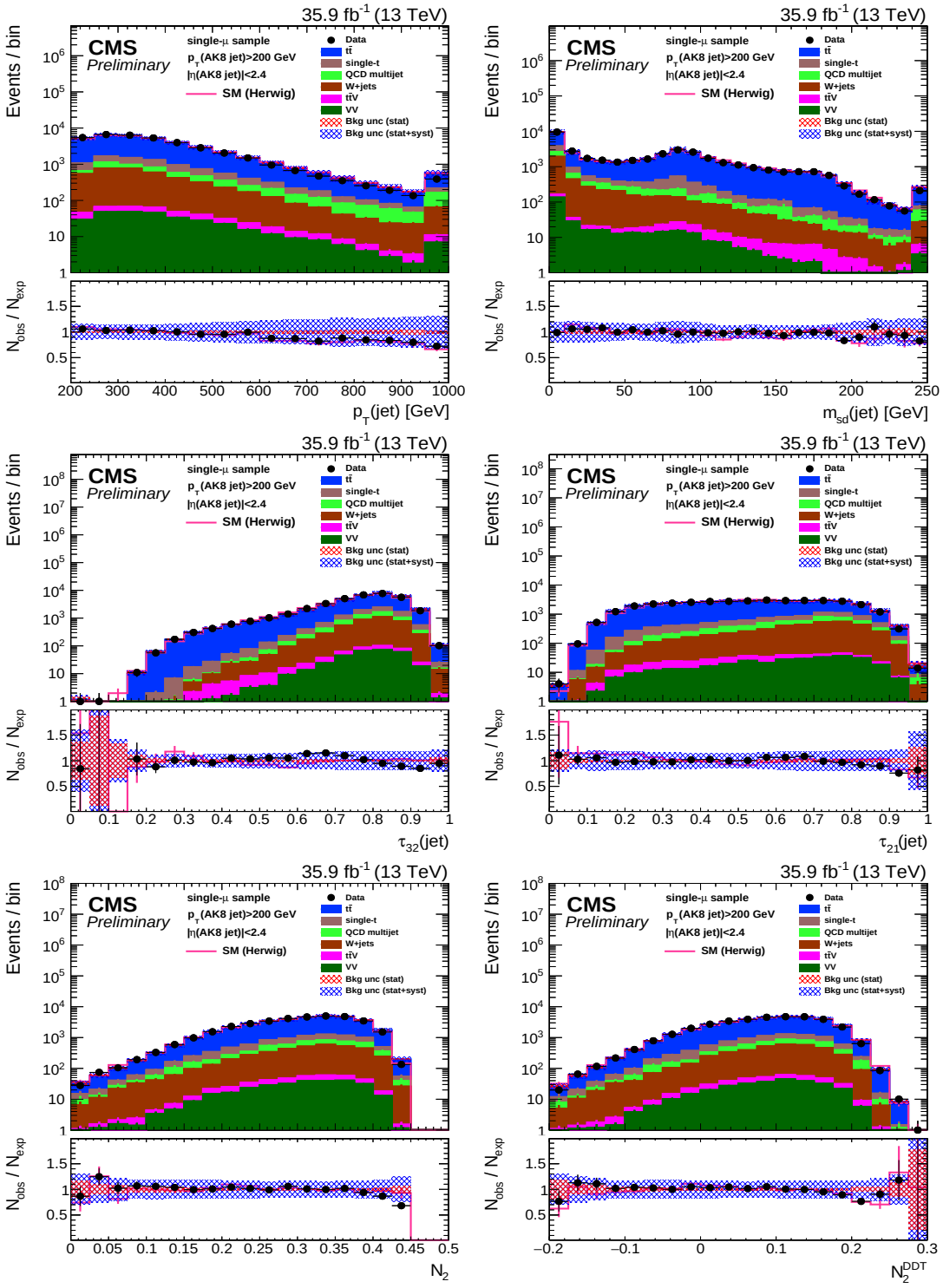

Figure 23: Distribution of the the jet p_{T} (upper-left), jet mass, m_{SD} (upper-right), the N-subjetiness ratios, τ_{32} (middle-left) and τ_{21} (middle-right), and the N_2 (lower-left) and N_2^{DDT} (lower-right) in data and simulation in the single-μ signal sample. The pink solid line corresponds to the simulation distribution obtained using the alternative $\mathrm{t\bar{t}}$ sample. The background event yield is normalized to the total observed data yield. The lower panel shows the data to simulation ratio. The shaded blue (red) band corresponds to the total uncertainty (statistical uncertainty of the simulated samples), the pink line to the data to simulation ratio using the alternative $\mathrm{t\bar{t}}$ sample, and the vertical lines correspond to the statistical uncertainty of the data. The distributions are weighted according to the top p_{T} reweighting procedure described in the text.

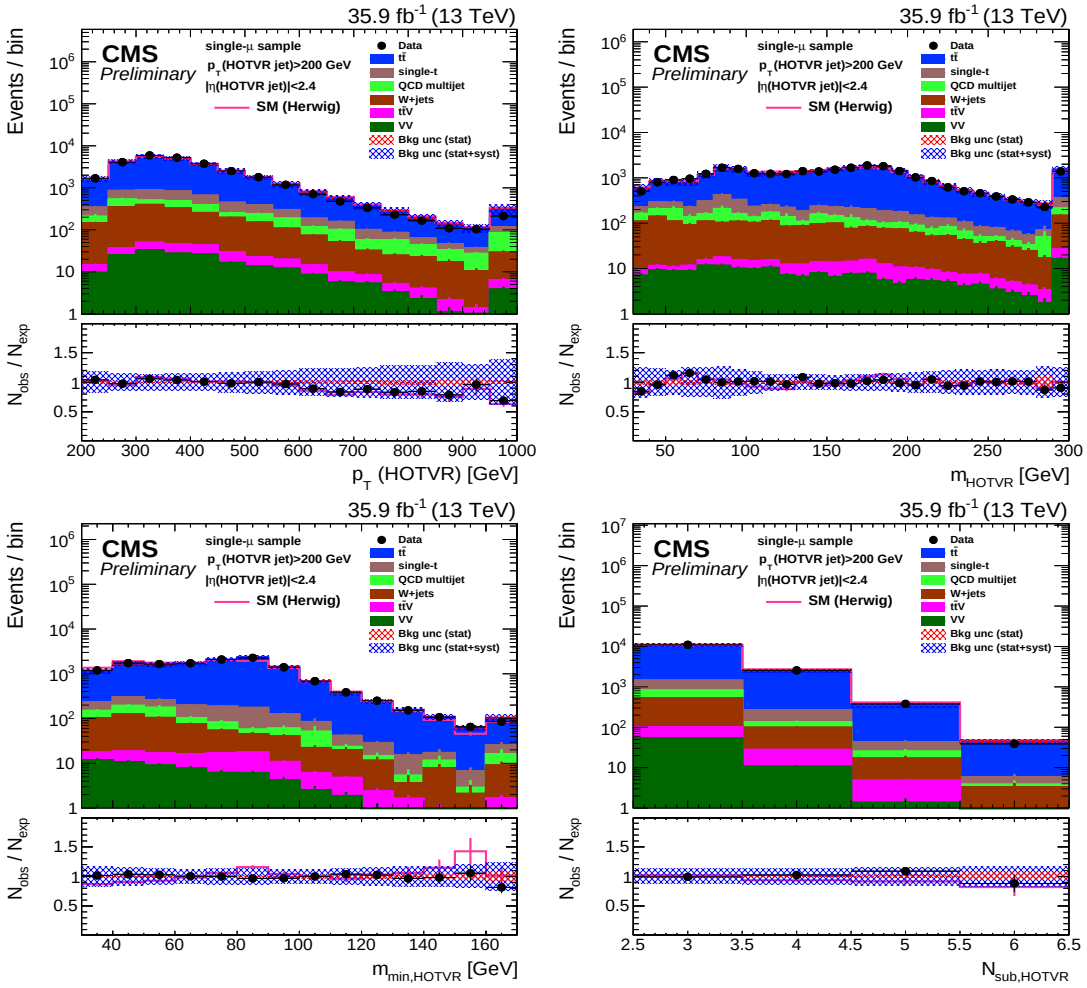

Figure 24: Distribution of the main observables of the HOTVR algorithm, p_T(HOTVR jet) (upper-left), m_{HOTVR} (upper-right), $m_{\text{min,HOTVR}}$ (lower-left), and $N_{\text{sub,HOTVR}}$ (lower-right) in data and simulation in the single-μ signal sample. The pink solid line corresponds to the simulation distribution obtained using the alternative $\mathrm{t\bar{t}}$ sample. The background event yield is normalized to the total observed data yield. The lower panel shows the data to simulation ratio. The shaded blue (red) band corresponds to the total uncertainty (statistical uncertainty of the simulated samples), the pink line to the data to simulation ratio using the alternative $\mathrm{t\bar{t}}$ sample, and the vertical lines correspond to the statistical uncertainty of the data. The distributions are weighted according to the top p_{T} reweighting procedure described in the text.

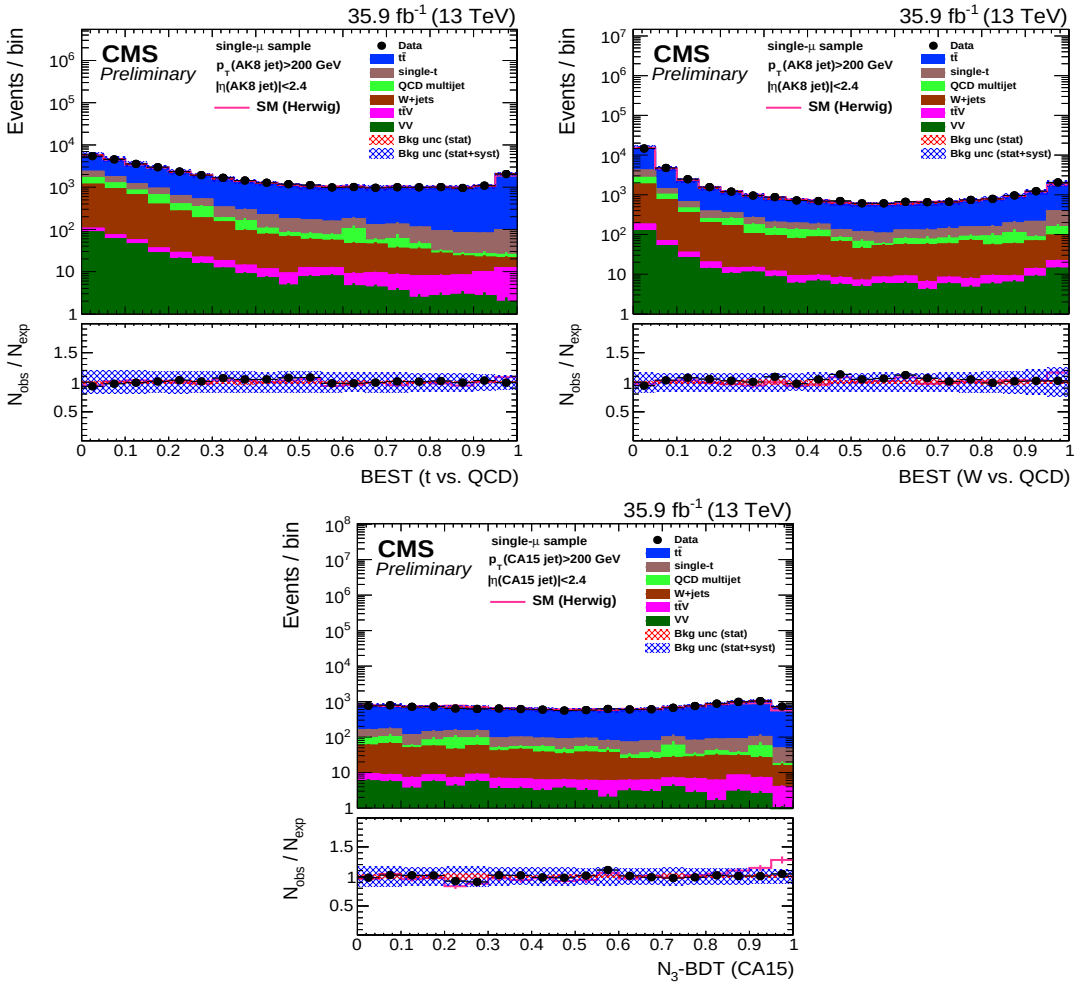

Figure 25: Distribution of the t quark (upper-left) and W boson (upper-right) identification probabilities for the BEST algorithm, and the $N_3 - \mathrm{BDT}\ (\mathrm{CA15})$ discriminant, in data and simulation in the single-μ signal sample. The pink solid line corresponds to the simulation distribution obtained using the alternative $\mathrm{t\bar{t}}$ sample. The background event yield is normalized to the total observed data yield. The lower panel shows the data to simulation ratio. The shaded blue (red) band corresponds to the total uncertainty (statistical uncertainty of the simulated samples), the pink line to the data to simulation ratio using the alternative $\mathrm{t\bar{t}}$ sample, and the vertical lines correspond to the statistical uncertainty of the data. The distributions are weighted according to the top p_{T} reweighting procedure described in the text.

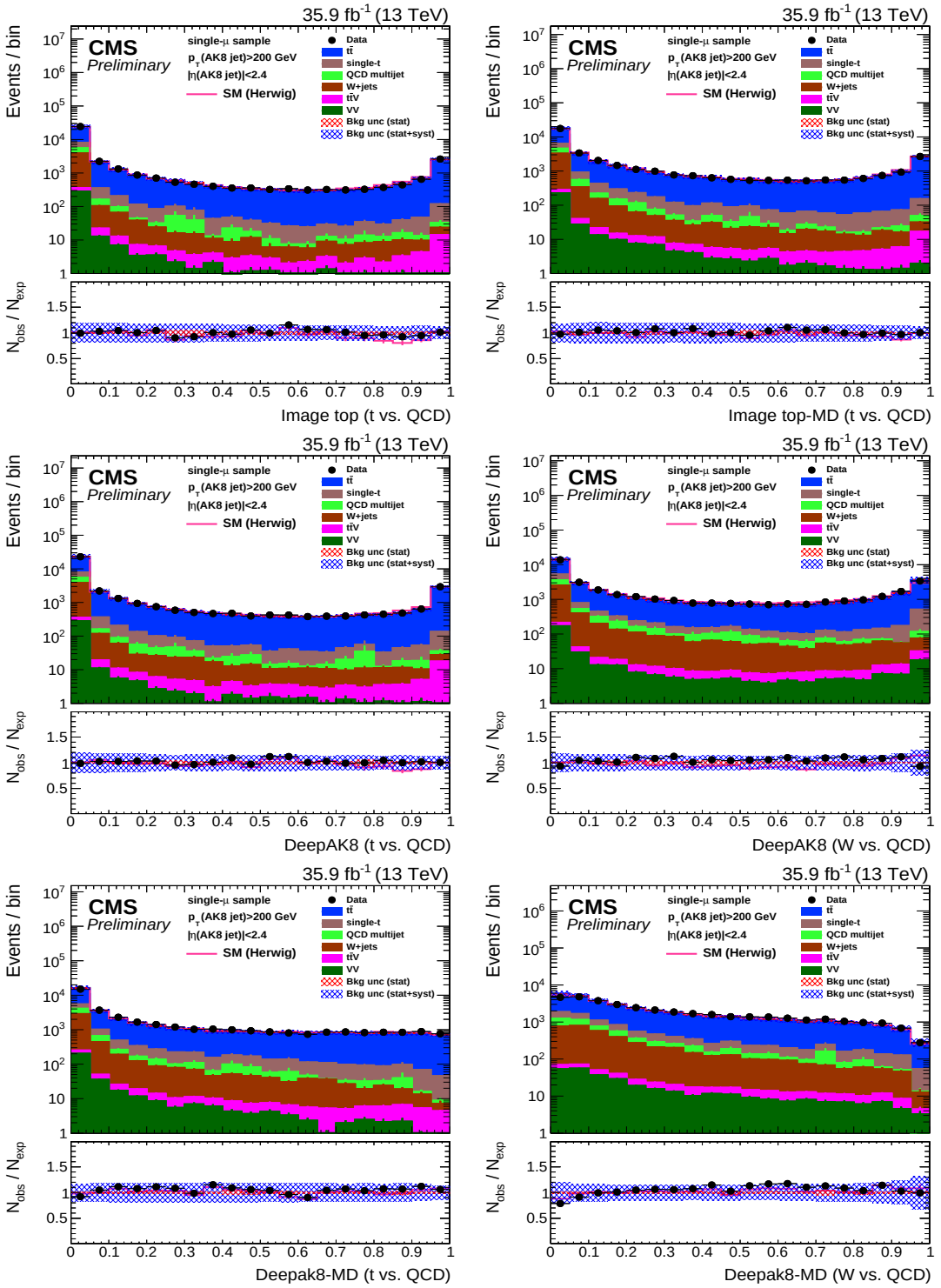

Figure 26: Distribution of the ImageTop (upper-left) and ImageTop-MD (upper-right) discriminant in data and simulation in the single-μ sample. The plots in the middle row show the t quark (left) and W boson (right) identification probabilities in data and simulation for the DeepAK8 algorithm. The corresponding plots for DeepAK8-MD are displayed in the lower row. The pink solid line corresponds to the simulation distribution obtained using the alternative $t\bar{t}$ sample. The background event yield is normalized to the total observed data yield. The lower panel shows the data to simulation ratio. The shaded blue (red) band corresponds to the total uncertainty (statistical uncertainty of the simulated samples), the pink line to the data to simulation ratio using the alternative $t\bar{t}$ sample, and the vertical lines correspond to the statistical uncertainty of the data. The distributions are weighted according to the top p_T reweighting procedure described in the text.

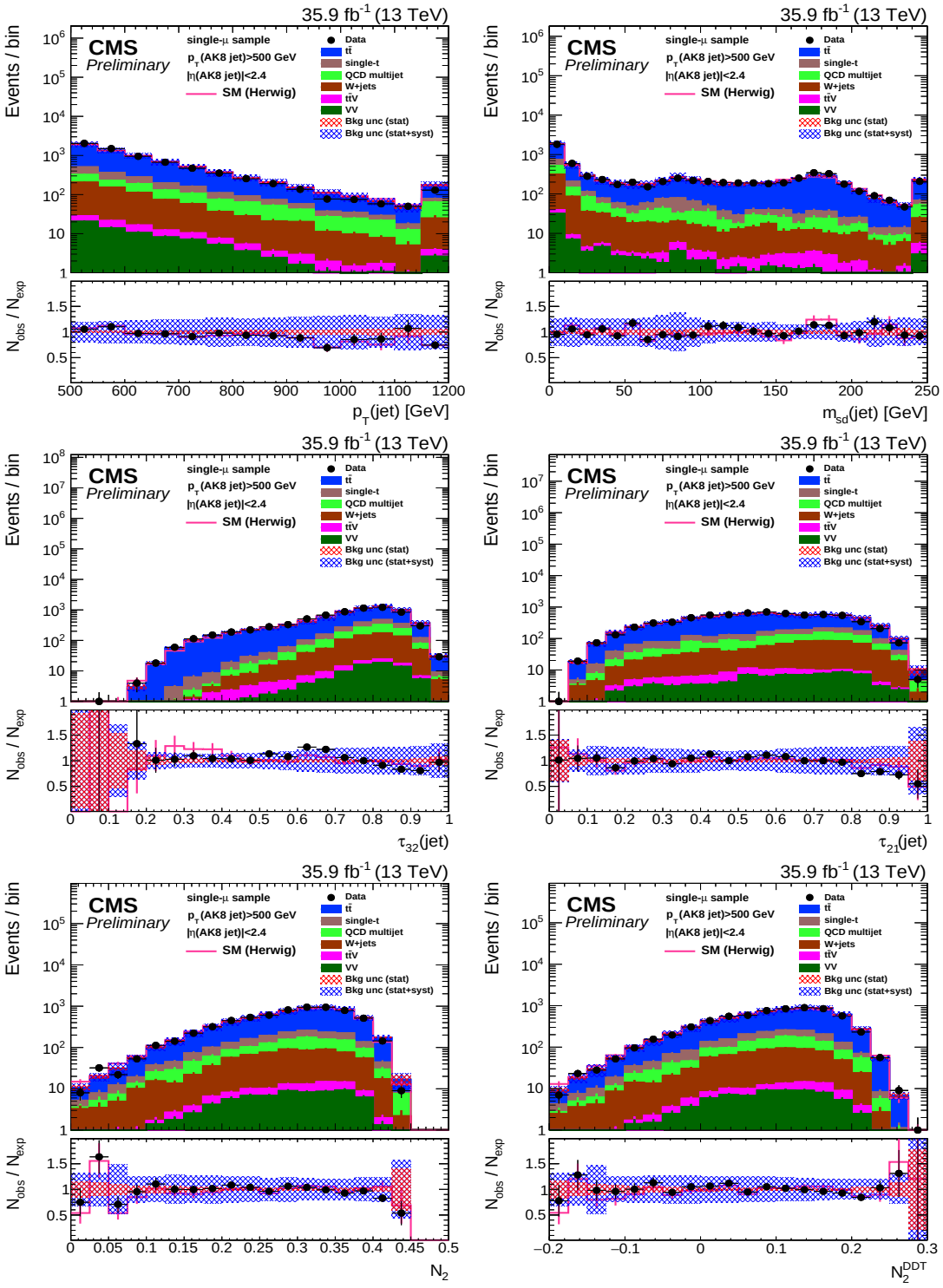

Figure 27: Distribution of the jet p_T (upper-left), the jet mass, m_{SD} (upper-right), the N-subjetiness ratios, τ_{32} (middle-left) and τ_{21} (middle-right), and the N_2 (lower-left) and N_2^{DDT} (lower-right) in data and simulation in the single-μ signal sample, after applying a jet momentum cut $p_T > 500$ GeV. The pink solid line corresponds to the simulation distribution obtained using the alternative $t\bar{t}$ sample. The background event yield is normalized to the total observed data yield. The lower panel shows the data to simulation ratio. The shaded blue (red) band corresponds to the total uncertainty (statistical uncertainty of the simulated samples), the pink line to the data to simulation ratio using the alternative $t\bar{t}$ sample, and the vertical lines correspond to the statistical uncertainty of the data. The distributions are weighted according to the top p_T reweighting procedure described in the text.

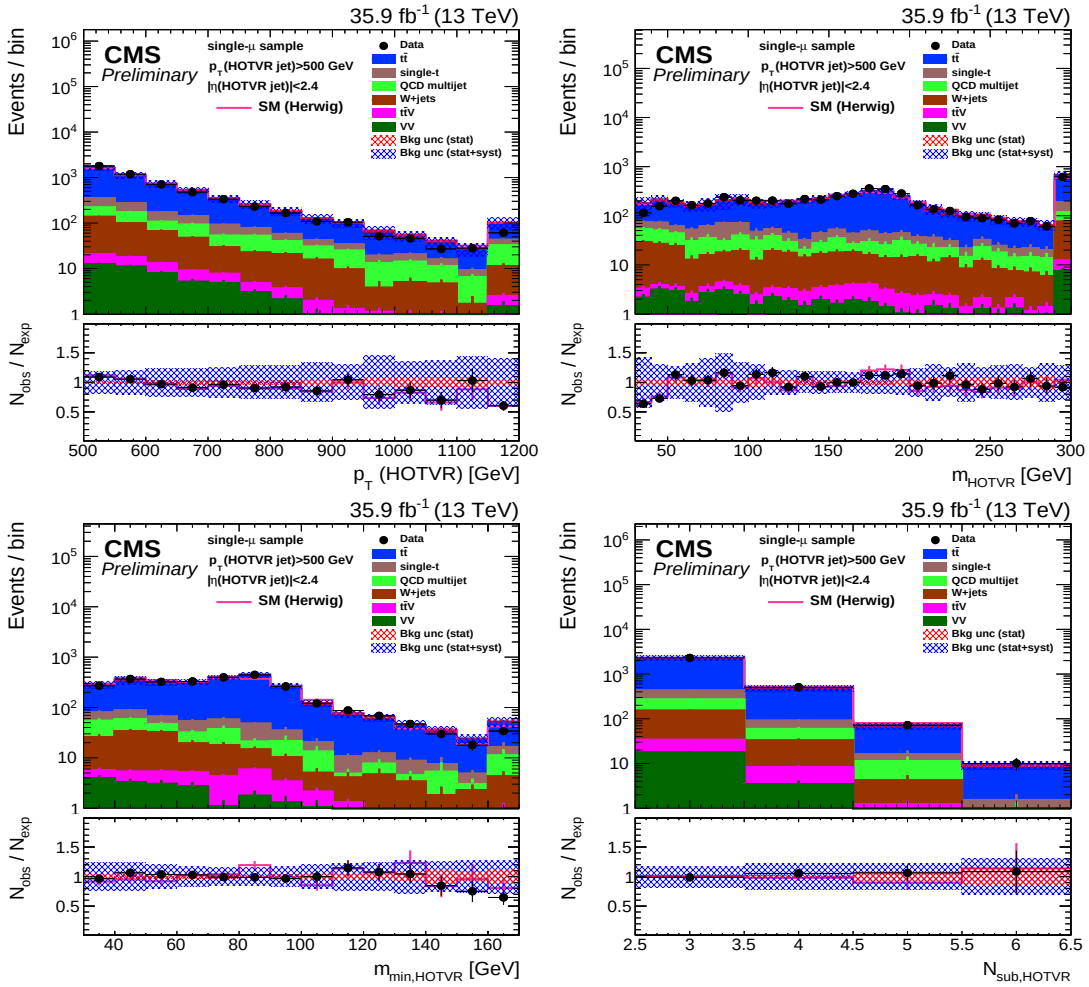

Figure 28: Distribution of the main observables of the HOTVR algorithm, p_T(HOTVR jet) (upper-left), m_{HOTVR} (upper-right), $m_{\text{min,HOTVR}}$ (lower-left) and $N_{\text{sub,HOTVR}}$ (lower-right) in data and simulation in the single-μ signal sample, after applying a jet momentum cut $p_{\text{T}} > 500\,\text{GeV}$. The pink solid line corresponds to the simulation distribution obtained using the alternative $\text{t}\bar{\text{t}}$ sample. The background event yield is normalized to the total observed data yield. The lower panel shows the data to simulation ratio. The shaded blue (red) band corresponds to the total uncertainty (statistical uncertainty of the simulated samples), the pink line to the data to simulation ratio using the alternative $\text{t}\bar{\text{t}}$ sample, and the vertical lines correspond to the statistical uncertainty of the data. The distributions are weighted according to the top p_{T} reweighting procedure described in the text.

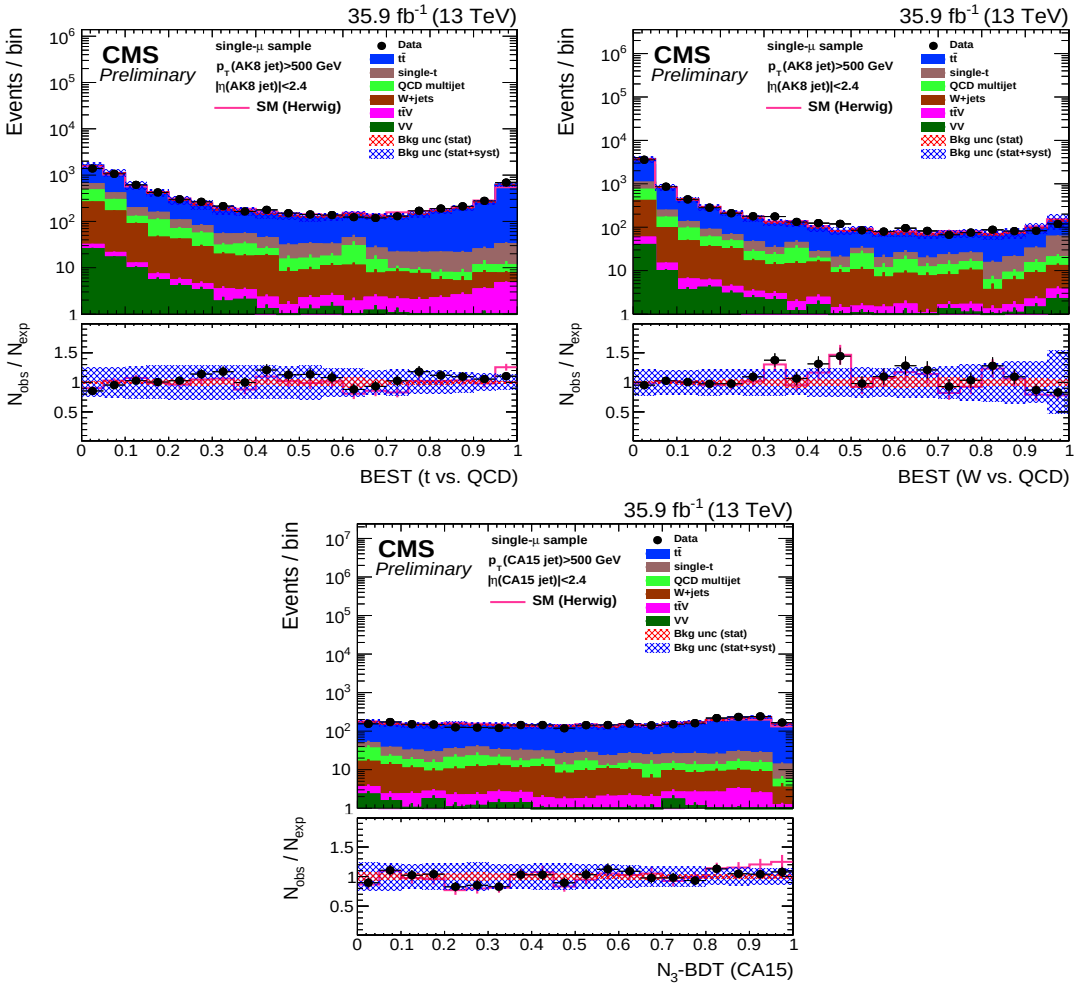

Figure 29: Distribution of the t quark (upper-left) and W boson (upper-right) identification probabilities for the BEST algorithm, and the $N_3 - \mathrm{BDT}\ (\mathrm{CA15})$ discriminant, in data and simulation in the single-μ signal sample, after applying a jet momentum cut $p_\mathrm{T} > 500$ GeV. The pink solid line corresponds to the simulation distribution obtained using the alternative $\mathrm{t\bar{t}}$ sample. The background event yield is normalized to the total observed data yield. The lower panel shows the data to simulation ratio. The shaded blue (red) band corresponds to the total uncertainty (statistical uncertainty of the simulated samples), the pink line to the data to simulation ratio using the alternative $\mathrm{t\bar{t}}$ sample, and the vertical lines correspond to the statistical uncertainty of the data. The distributions are weighted according to the top p_T reweighting procedure described in the text.

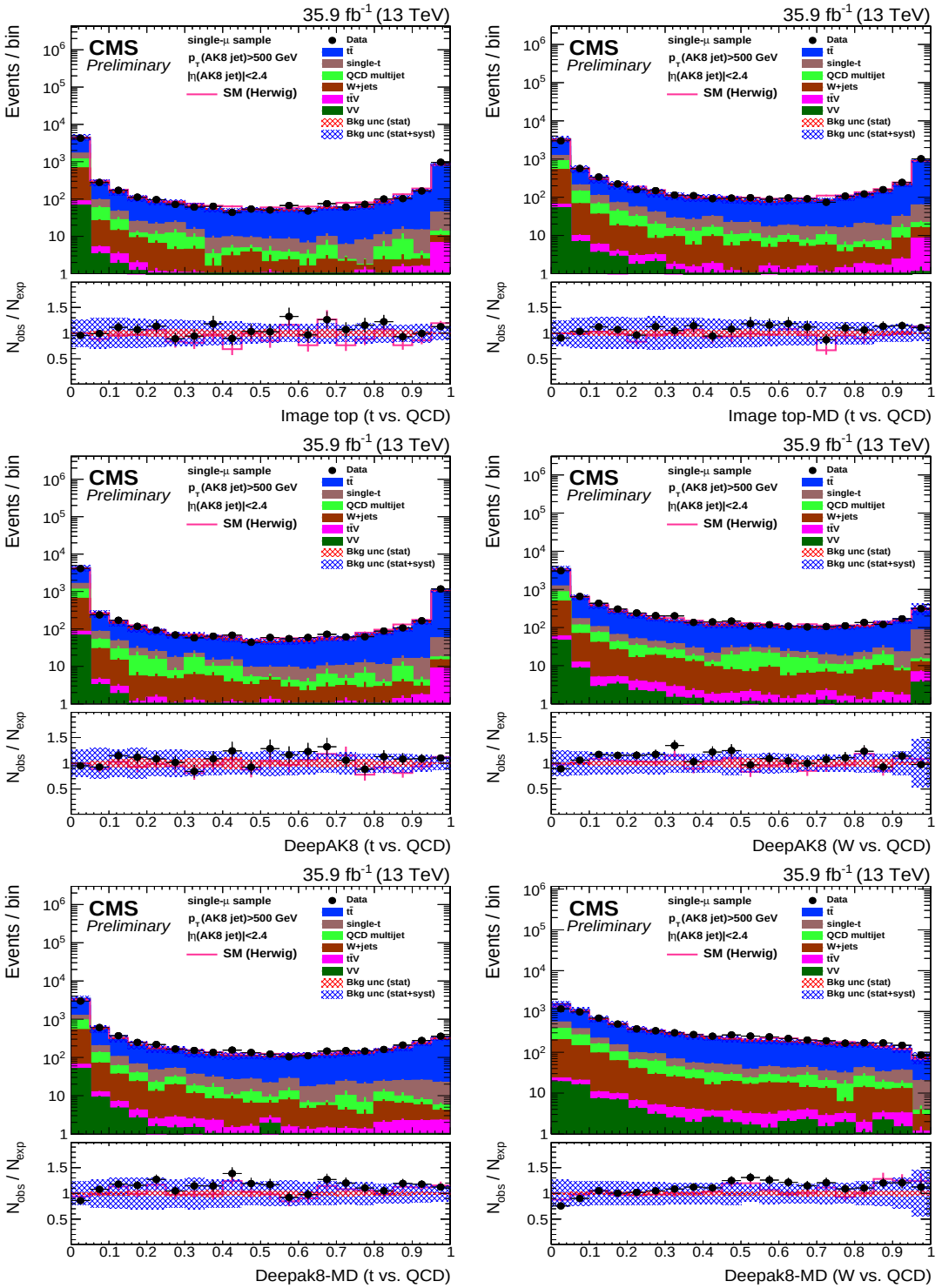

Figure 30: Distribution of the ImageTop (upper left) and ImageTop-MD (upper-right) discriminant in data and simulation in the single-μ sample. The plots in the middle row show the t quark (left) and W boson (right) identification probabilities in data and simulation for the DeepAK8 algorithm, after applying a jet momentum cut $p_T > 500$ GeV. The corresponding plots for DeepAK8-MD are displayed in the lower row. The pink solid line corresponds to the simulation distribution obtained using the alternative $t\bar{t}$ sample. The background event yield is normalized to the total observed data yield. The lower panel shows the data to simulation ratio. The shaded blue (red) band corresponds to the total uncertainty (statistical uncertainty of the simulated samples), the pink line to the data to simulation ratio using the alternative $t\bar{t}$ sample, and the vertical lines correspond to the statistical uncertainty of the data. The distributions are weighted according to the top p_T reweighting procedure described in the text.

tagging discriminant for the $N_3 - \text{BDT}$ (CA15) algorithm in the di-jet sample are presented in Fig. 35, while the equivalent plots for the single-γ selection are shown in Fig 36. In both samples the agreement between data and simulation is found to be reasonable. Some tension is observed in the very high values ($\gtrsim 0.95$) for the t quark identification probability of the BEST algorithm in the single-γ sample. The disagreement is observed in the region of the t quark probability greater than 0.95, which is significantly tighter than the recommended operating points. Some disagreement is observed between the nominal QCD multijet simulated sample and the alternative for large values of the W boson probability of the BEST algorithm, with the nominal sample showing better agreement with the data.

The distributions of the ImageTop and DeepAK8 discriminants are shown in Figs 37 and 38 for jets in the di-jet and single-γ samples, respectively. The agreement between data and simulation in the single-γ is overall better than in the di-jet sample. Moreover, the discrepancy on the shape is mainly observed in the very low values of the discriminant and more enhanced in the t-tagging case. The di-jet sample is dominated by jets initiated by gluons, especially at low values of the discriminant. In addition, ImageTop and DeepAK8 are very sensitive to mismodeling of quarks or gluons in the simulation, so exhibit more sample dependence. To this end, QCD multijet events simulated using HERWIG++ show generally better agreement with the data.

8.4 Corrections to simulation

The measurement of the t quark and W boson tagging efficiency in data is performed in the single$-\mu$ sample using a "tag & probe" method. The muon, in combination with the b tagged jet, is used as the "tag". In the opposite hemisphere of the event, the jet is considered as the "probe jet".

The total SM sample is decomposed into three categories based on the spatial separation of the partons from the t quark decay with respect to the AK8 jet, following the discussion in Section 4. The "Merged t quark" category includes cases were the three partons and the jet have $\Delta R < 0.6$. In the "Merged W boson" category are cases where only the two partons from the W boson decay are within $\Delta R < 0.6$ of the jet, and the b quark from the top decay is outside the jet cone. Any other scenario falls in the "Unmerged" category. In the cases of the HOTVR and $N_3 - \text{BDT}$ (CA15) algorithms the matching requirement is adjusted from 0.6 to 1.2.

The m_{jet} distributions in simulation of each one of the three categories are used to derive templates to fit the m_{jet} distribution in data. For a given working point, the fit is done simultaneously for both the "passing" and "failing" events, for all three categories. The fit is performed in the range from 50 GeV to 250 GeV with a bin width of 10 GeV. The sources of systematic uncertainties discussed in Section 8.1 are considered and are treated as nuisance parameters in the fit. After calculating the efficiencies in data and simulation, the SF is determined as:

$$SF = \frac{\epsilon_{\text{Data}}}{\epsilon_{\text{Simulation}}}. \tag{16}$$

The SFs are extracted differentially in jet p_{T}. For the case of t quark identification the following exclusive jet p_{T} regions are considered: $300 - 400$, $400 - 480$, $480 - 600$, and $600 - 1200$ GeV. In order to increase the purity of "Merged W boson" candidates, we consider regions with lower jet p_{T} : $200 - 300$, $300 - 400$, $400 - 550$, and $550 - 800$ GeV. The effects of the systematic sources discussed in Section 8.1 are propagated to uncertainties in the SF. An example of m_{jet} distributions for data and simulation in the passing and failing categories for $400 < p_{\text{T}} < 480$ GeV after performing the maximum likelihood fit are displayed in Fig 39.

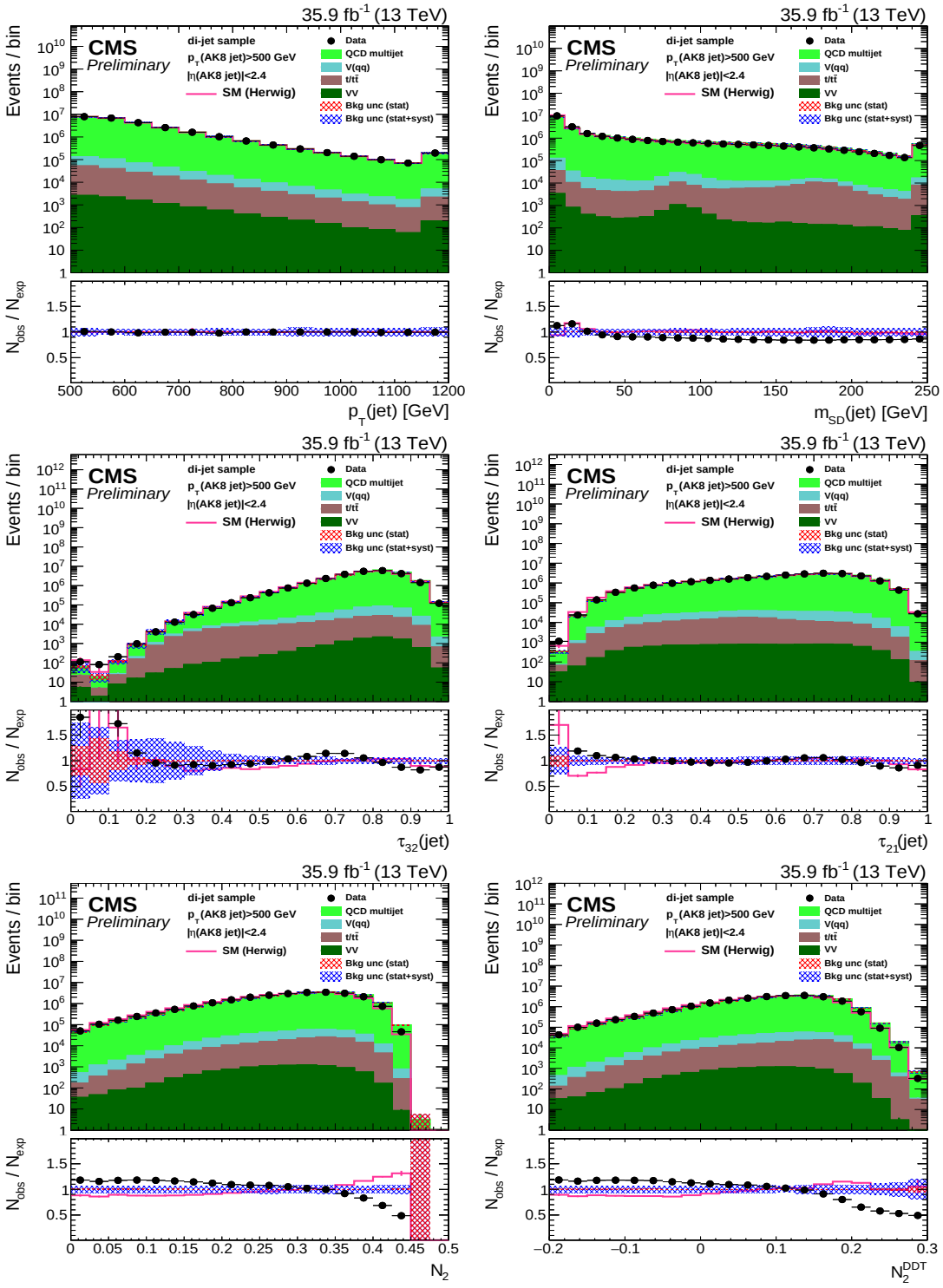

Figure 31: Distribution of the jet p_T (upper-left), the jet mass, m_{SD} (upper-right), the N-subjetiness ratios, τ_{32} (middle-left) and τ_{21} (middle-right), and the N_2 (lower-left) and N_2^{DDT} (lower-right) in data and simulation in the di-jet sample. The pink solid line corresponds to the simulation distribution obtained using the alternative QCD multijet sample. The background event yield is normalized to the total observed data yield. The lower panel shows the data to simulation ratio. The shaded blue (red) band corresponds to the total uncertainty (statistical uncertainty of the simulated samples), the pink line to the data to simulation ratio using the alternative QCD multijet sample, and the vertical lines correspond to the statistical uncertainty of the data. The distributions are weighted so that the jet p_T distribution of the simulation matches the data.

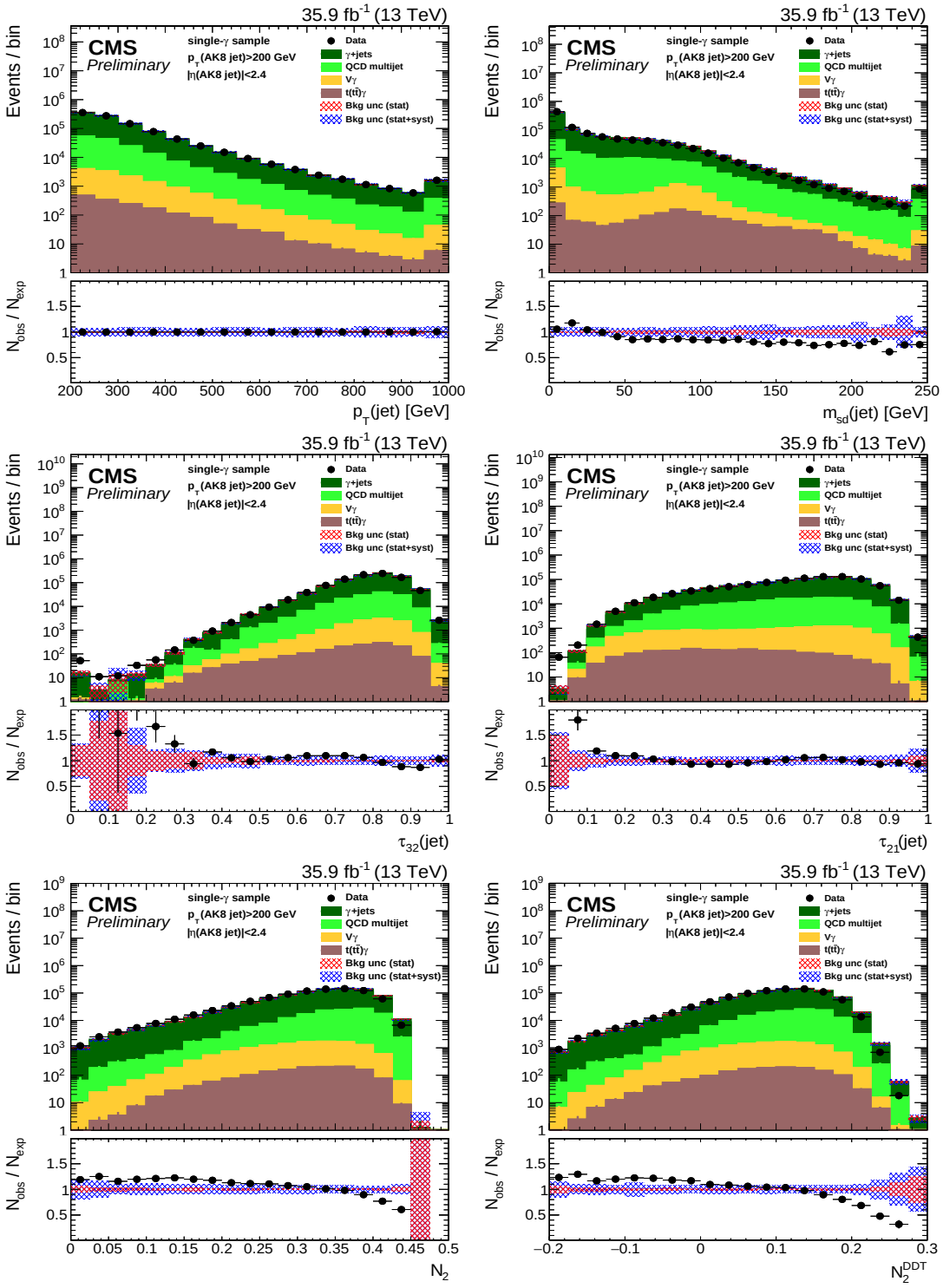

Figure 32: Distribution of the jet p_T (upper-left), the jet mass, m_{SD} (upper-right), the N-subjetiness ratios, τ_{32} (middle-left) and τ_{21} (middle-right), and the N_2 (lower-left) and N_2^{DDT} (lower-right) in data and simulation in the single-γ sample. The background event yield is normalized to the total observed data yield. The lower panel shows the data to simulation ratio. The shaded blue (red) band corresponds to the total uncertainty (statistical uncertainty of the simulated samples), and the vertical lines correspond to the statistical uncertainty of the data. The distributions are weighted so that the jet p_T distribution of the simulation matches the data.

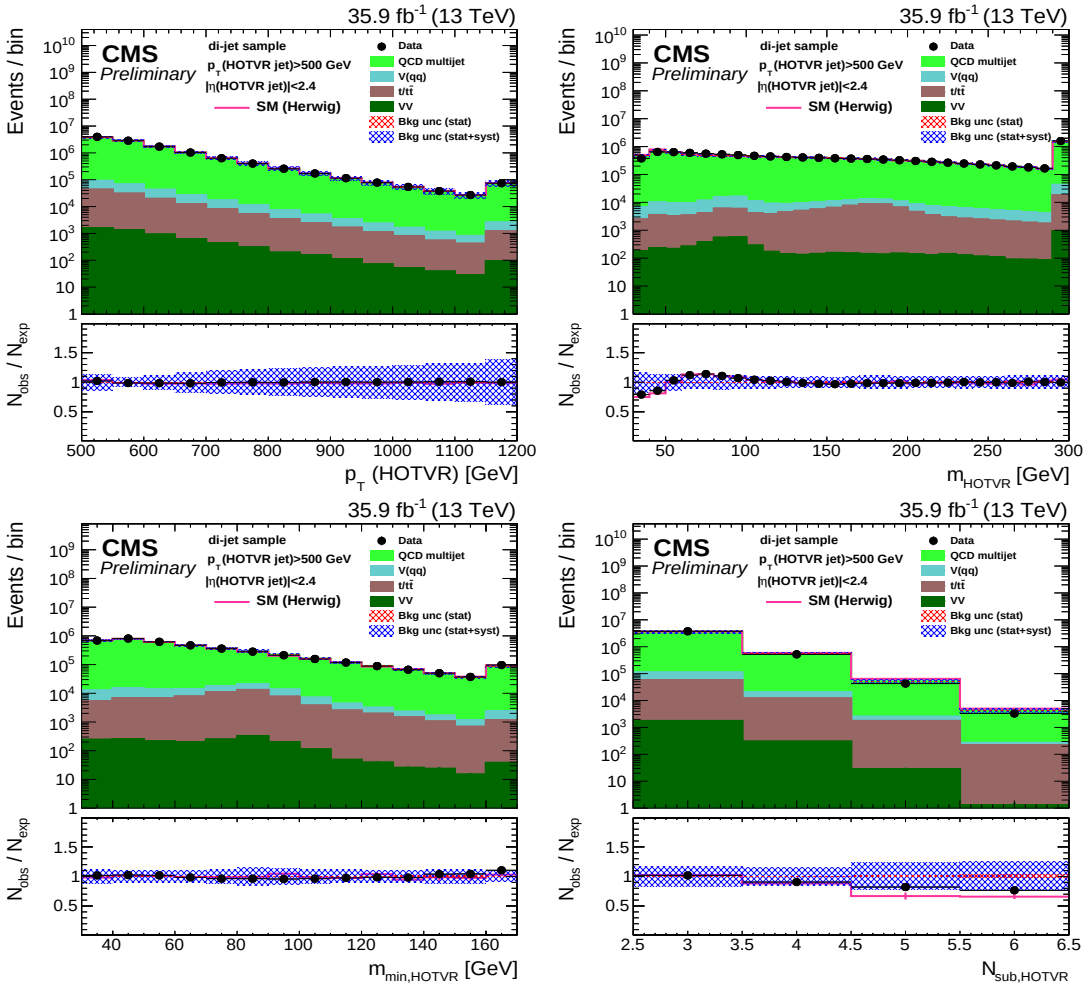

Figure 33: Distribution of the main observables of the HOTVR algorithm, p_T(HOTVR jet) (upper-left), m_{HOTVR} (upper-right), $m_{\text{min,HOTVR}}$ (lower-left) and $N_{\text{sub,HOTVR}}$ (lower-right) in data and simulation in the di-jet sample. The pink solid line corresponds to the simulation distribution obtained using the alternative QCD multijet sample. The background event yield is normalized to the total observed data yield. The lower panel shows the data to simulation ratio. The shaded blue (red) band corresponds to the total uncertainty (statistical uncertainty of the simulated samples), the pink line to the data to simulation ratio using the alternative QCD multijet sample, and the vertical lines correspond to the statistical uncertainty of the data. The distributions are weighted so that the jet p_{T} distribution of the simulation matches the data.

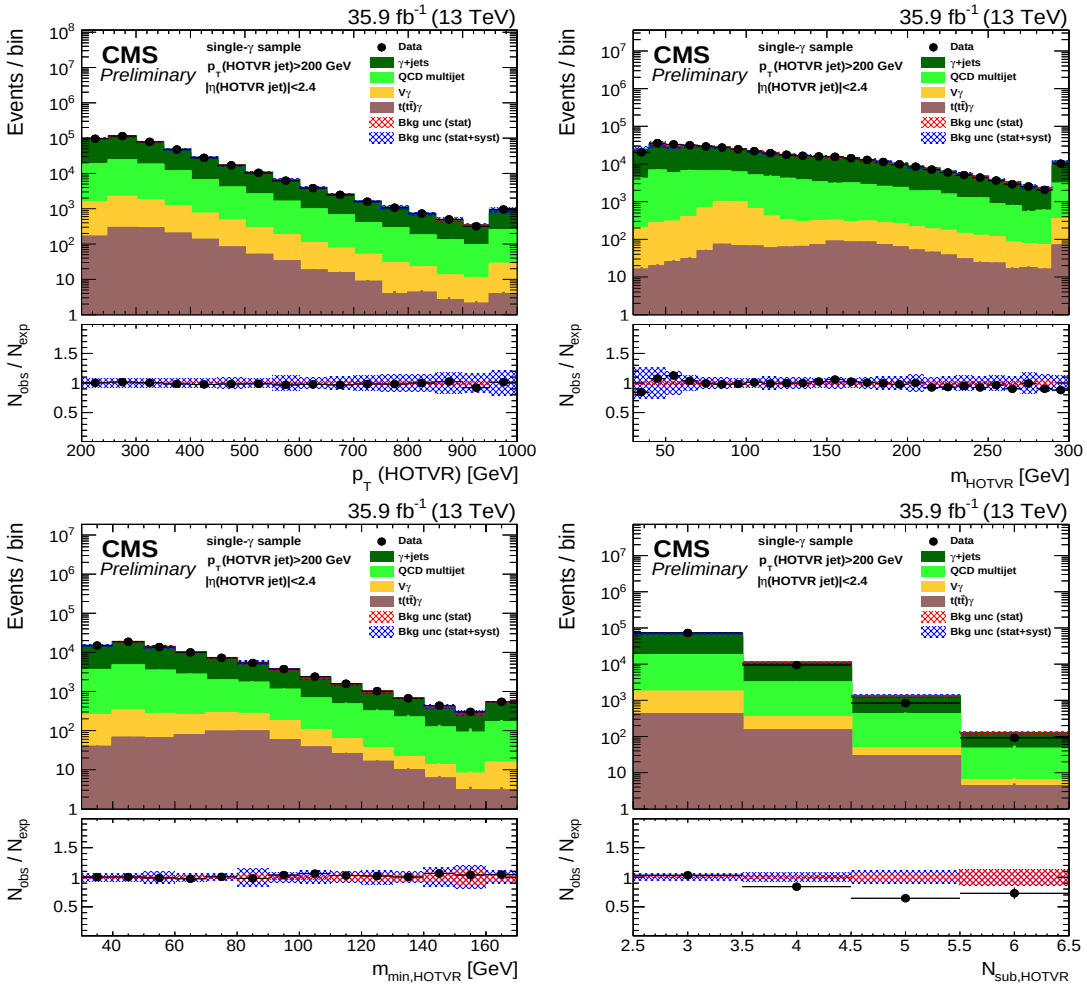

Figure 34: Distribution of the main observables of the HOTVR algorithm, p_T(HOTVR jet) (upper-left), m_{HOTVR} (upper-right), $m_{\text{min,HOTVR}}$ (lower-left) and $N_{\text{sub,HOTVR}}$ (lower-right) in data and simulation in the single-γ sample. The background event yield is normalized to the total observed data yield. The lower panel shows the data to simulation ratio. The shaded blue (red) band corresponds to the total uncertainty (statistical uncertainty of the simulated samples), and the vertical lines correspond to the statistical uncertainty of the data. The distributions are weighted so that the jet p_{T} distribution of the simulation matches the data.

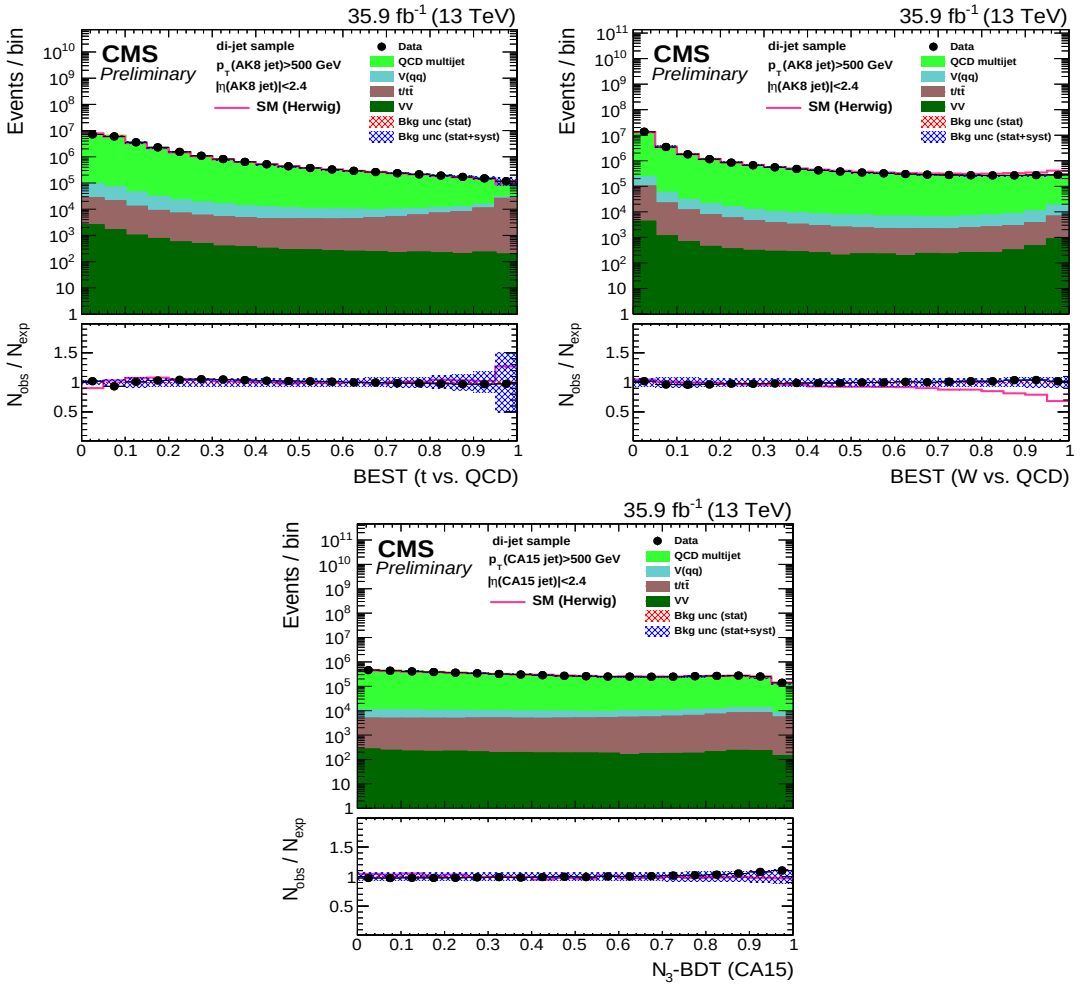

Figure 35: Distribution of the t quark (upper left) and W boson (upper-right) identification probabilities for the BEST algorithm, and the $N_3 - \mathrm{BDT}\ (\mathrm{CA15})$ discriminant, in data and simulation in the di-jet sample. The background event yield is normalized to the total observed data yield. The pink solid line corresponds to the simulation distribution obtained using the alternative QCD multijet sample. The background event yield is normalized to the total observed data yield. The lower panel shows the data to simulation ratio. The shaded blue (red) band corresponds to the total uncertainty (statistical uncertainty of the simulated samples), the pink line to the data to simulation ratio using the alternative QCD multijet sample, and the vertical lines correspond to the statistical uncertainty of the data. The distributions are weighted so that the jet p_{T} distribution of the simulation matches the data.

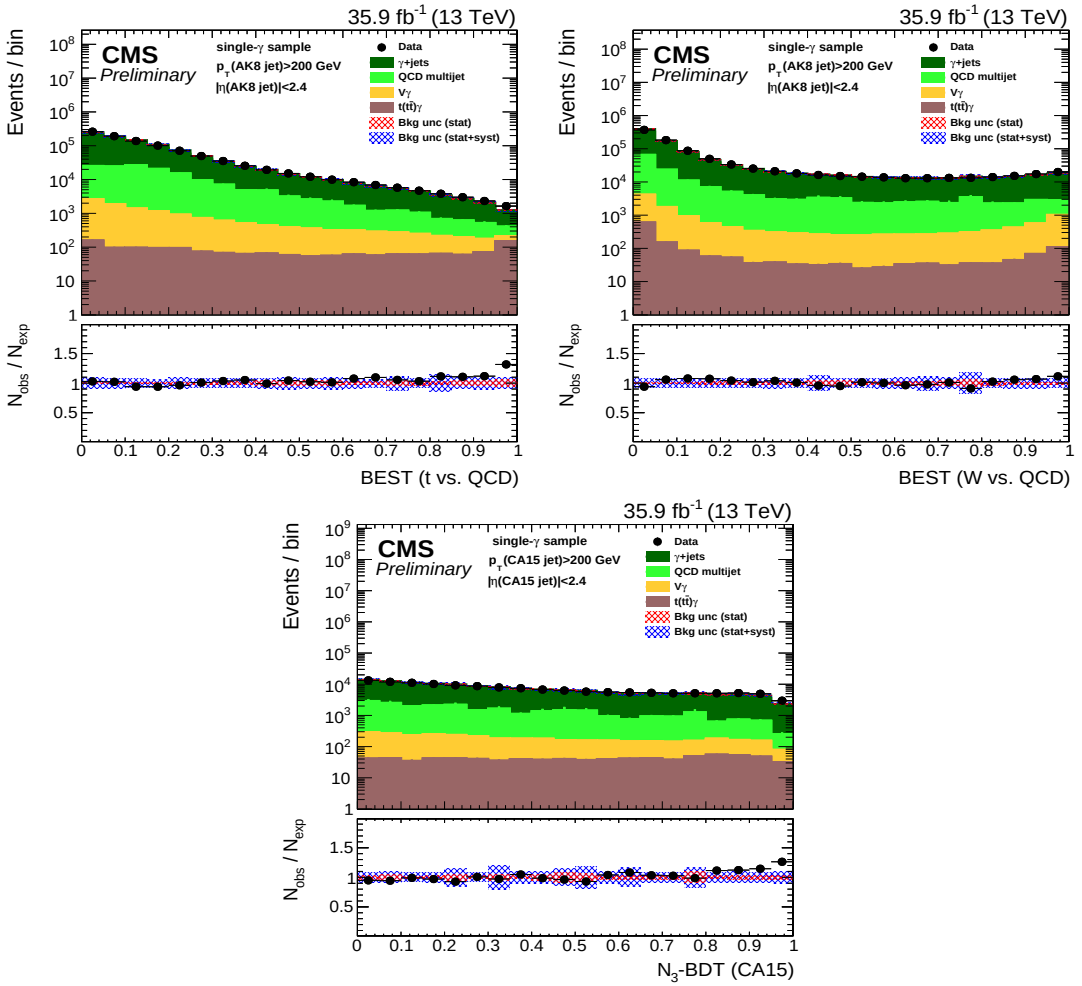

Figure 36: Distribution of the t quark (upper left) and W boson (upper-right) identification probabilities for the BEST algorithm, and the N_3 − BDT (CA15) discriminant, in data and simulation in the single-γ sample. The background event yield is normalized to the total observed data yield. The background event yield is normalized to the total observed data yield. The lower panel shows the data to simulation ratio. The shaded blue (red) band corresponds to the total uncertainty (statistical uncertainty of the simulated samples), and the vertical lines correspond to the statistical uncertainty of the data. The distributions are weighted so that the jet p_T distribution of the simulation matches the data.

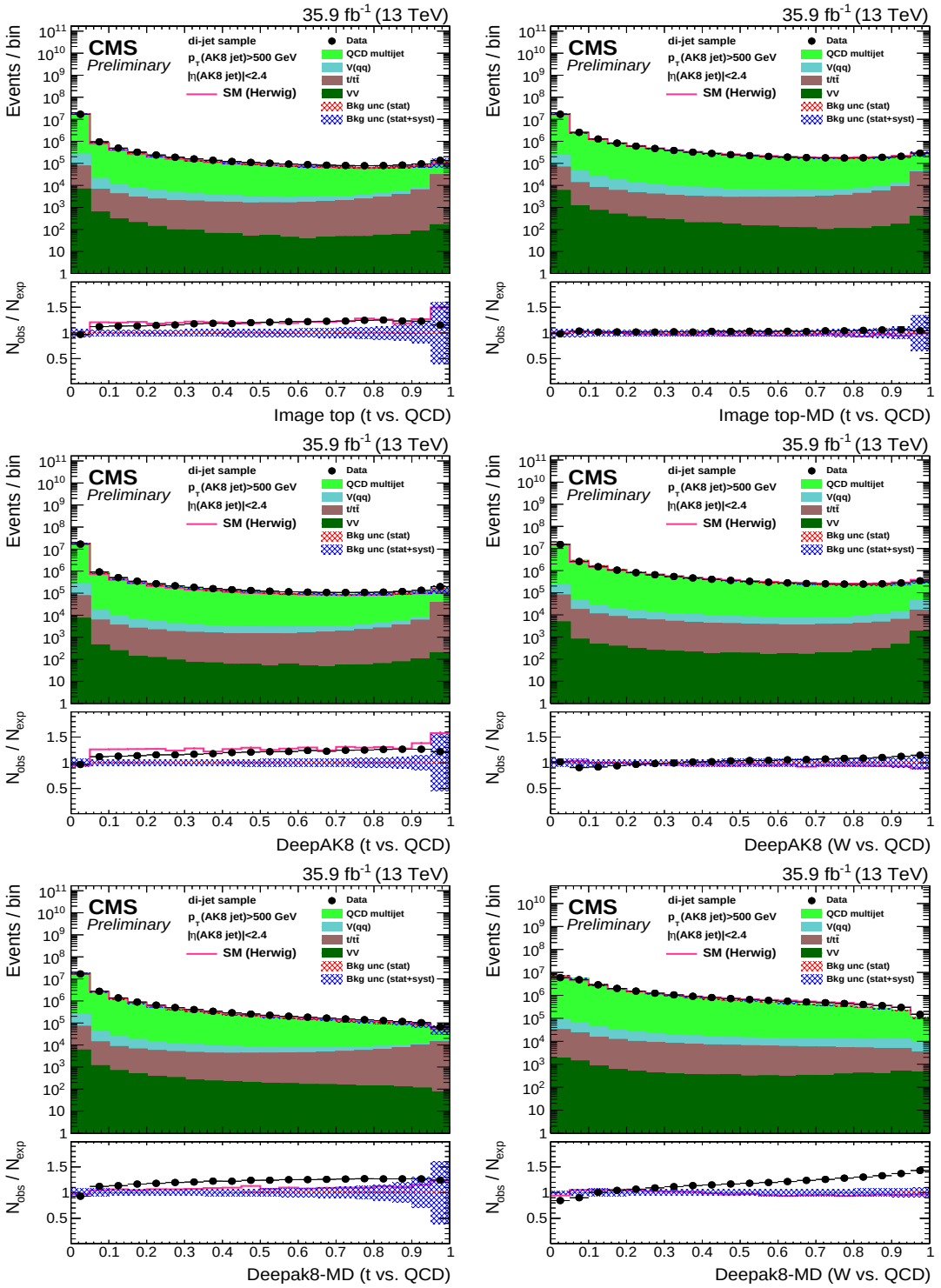

Figure 37: Distribution of the ImageTop (upper-left) and ImageTop-MD (upper-right) discriminant in data and simulation in the di-jet sample. The plots in the middle row show the t quark (left) and W boson (right) identification probabilities in data and simulation for the DeepAK8 algorithm. The corresponding plots for DeepAK8-MD are displayed in the lower row. The pink solid line corresponds to the simulation distribution obtained using the alternative QCD multijet sample. The background event yield is normalized to the total observed data yield. The lower panel shows the data to simulation ratio. The shaded blue (red) band corresponds to the total uncertainty (statistical uncertainty of the simulated samples), the pink line to the data to simulation ratio using the alternative QCD multijet sample, and the vertical lines correspond to the statistical uncertainty of the data. The distributions are weighted so that the jet p_{T} distribution of the simulation matches the data.

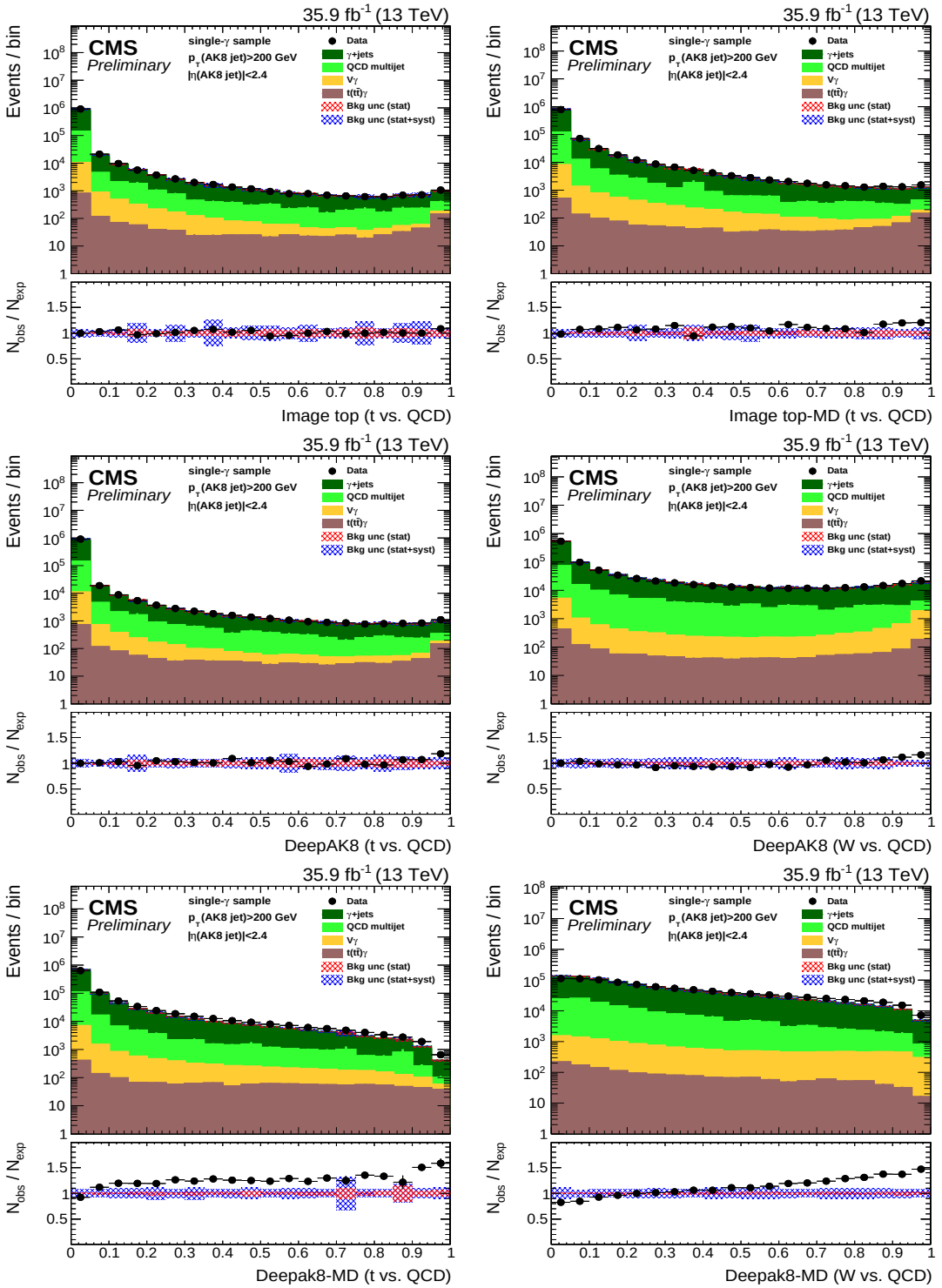

Figure 38: Distribution of the ImageTop (upper-left) and ImageTop-MD (upper-right) discriminant in data and simulation in the single-γ sample. The plots in the middle row show the t quark (left) and W boson (right) identification probabilities in data and simulation for the DeepAK8 algorithm. The corresponding plots for DeepAK8-MD are displayed in the lower row. The background event yield is normalized to the total observed data yield. The lower panel shows the data to simulation ratio. The shaded blue (red) band corresponds to the total uncertainty (statistical uncertainty of the simulated samples), and the vertical lines correspond to the statistical uncertainty of the data. The distributions are weighted so that the jet p_{T} distribution of the simulation matches the data.

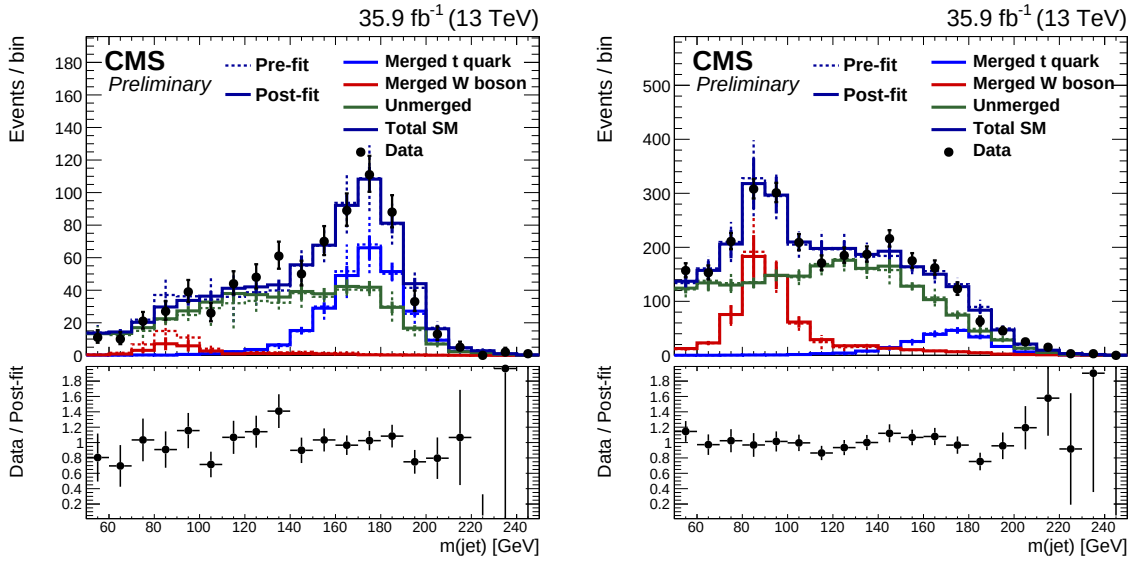

Figure 39: The m_{jet} distributions for data and simulation in the passing (left) and failing (right) categories for $400 < p_{\text{T}}(\text{jet}) < 480$ GeV. The solid lines correspond to the contribution of each category after performing the maximum likelihood fit as described in the text. The dashed lines are the expectation from simulation before the fit. The lower panel shows the data to simulation ratio.

The SF measured for each of the t quark and W boson identification algorithms are summarized in Figs. 40 and 41, respectively. The SFs are typically consistent with unity, within uncertainties. The largest SF is measured for the identification of t quarks using DeepAK8-MD. The statistical and parton shower uncertainties dominate the SF measurement. Another observation is that algorithms designed to avoid strong dependence on the mass, like the DeepAK8-MD, have typically smaller uncertainties compared to the other algorithms. Another point is that the effect of the systematic uncertainties is more pronounced on algorithms that utilize a larger set of observables to increase discrimination power. These algorithms (i.e. BEST, ImageTop, and DeepAK8) are more sensitive to the simulation details. Both points are more evident in the W case, due to the larger sample size of the "Merged W boson" category compared to the "Merged t quark" category, which allows the exploration of finer details of the modeling of data by simulation.

The misidentification rate as a function of the p_{T} of the jet is displayed in Figs 42 and 43 for the t and W tagging algorithms. In order to study the dependence of the misidentification probability on the hard-scatter generator, and on the modeling of the parton showering, we use an additional simulation sample for the QCD multijet background, which uses HERWIG++ for both the hard scattering generation and the parton showering. In some cases, the misidentification probabilities show an important dependence (up to $\sim$ 25%) on the simulation details, particularly for the ImageTop and DeepAK8 algorithms. The main source of this dependence is the description of the gluon content; these are the only algorithms that have access to quark-gluon separation to improve performance. Differences in the quark/gluon content can have large effects on the uncertainties.

Moreover, the misidentification probability is studied in the single-γ sample. Overall the performance in data and simulation in this sample is in better agreement than in the di-jet sample. This can be attributed to the fact that the single-γ sample has a larger fraction of light-quarks, which are better modeled in simulation [18].

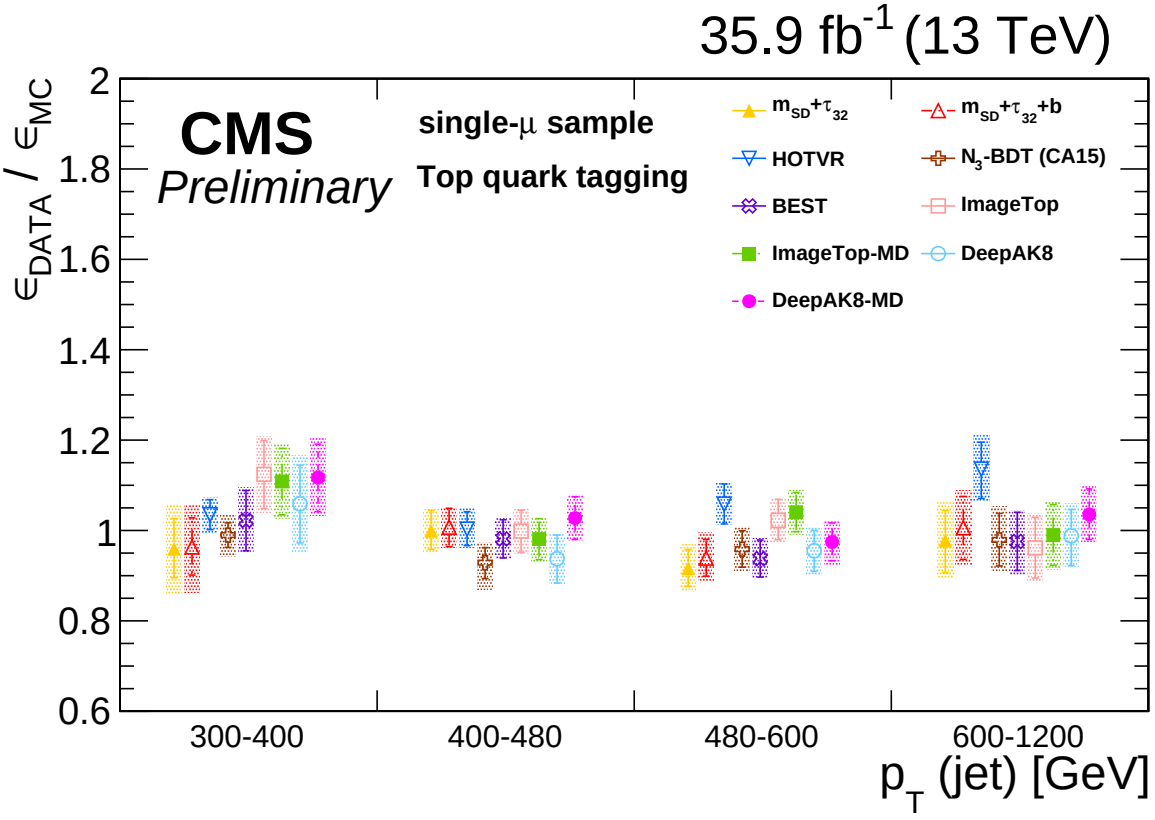

Figure 40: Summary of the SFs measured for each of the t quark identification algorithms. The markers correspond to the SF value, the error bars to the statistical uncertainty on the SF measurement, and the band is the total uncertainty (statistical + systematic).

9 Summary

A review of the heavy object tagging methods recently developed in CMS has been presented. Tagging algorithms based on theory inspired higher-level observables, which were studied in LHC Run1, serve as a reference. New tagging approaches, such as the ECF tagger and the BEST algorithm, utilize multivariate methods (i.e., boosted decision trees or deep neural networks) on higher-level observables and result in enhanced performance. A novel set of tagging algorithms, ImageTop and DeepAK8, are developed based on candidate level information, allowing to explore more of the CMS potential. Lower-level information is processed using advanced machine learning methods. This approach results in significant performance improvement which in some cases leads to $\sim O(10)$ gain in background rejection for the same signal efficiency. Moreover, the BEST and DeepAK8 algorithms are developed to provide multi-class tagging capabilities, which can potentially enable new measurements and search approaches. Finally, dedicated versions of the algorithms which are only loosely correlated with the jet mass are developed.

The performance of these new techniques has been directly compared in simulation in a jet transverse momentum range from 200 to 2000 GeV. The techniques have also been validated in collision data events, with scale factors extracted including systematic uncertainties.

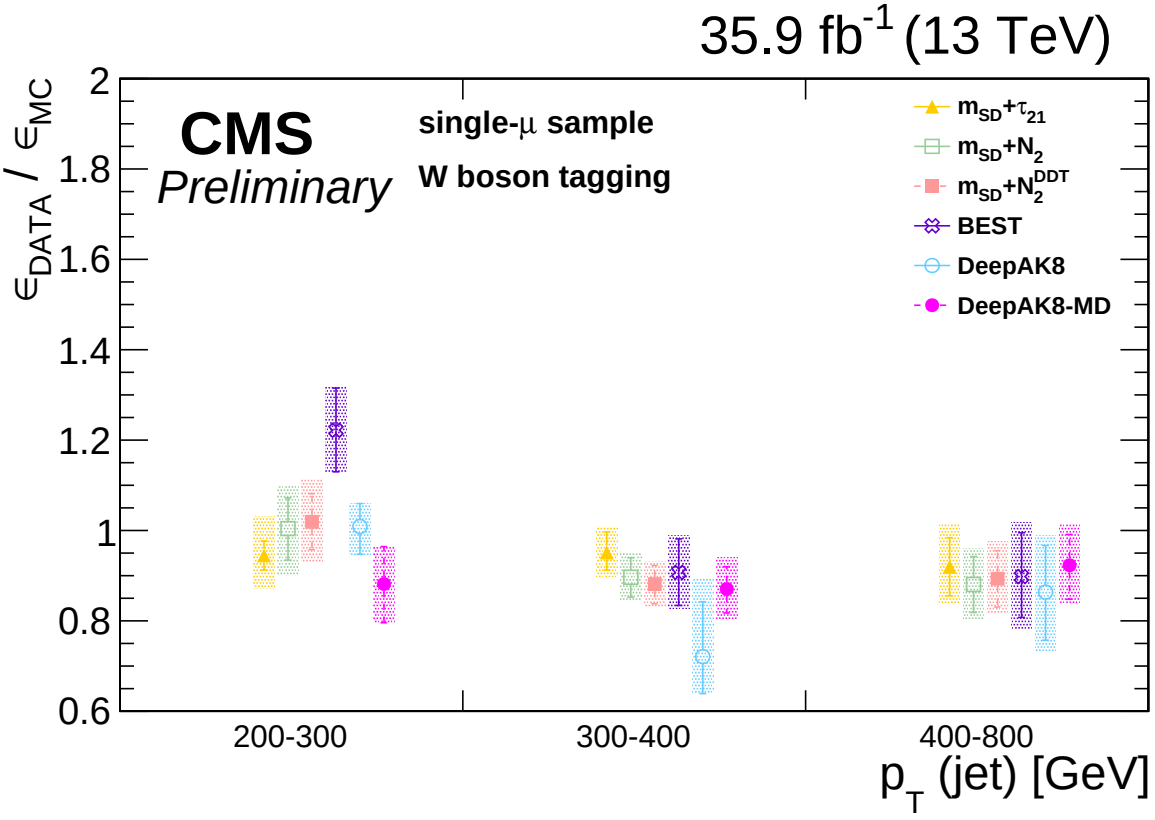

Figure 41: Summary of the SF measured for each of the W boson identification algorithms. The markers correspond to the SF value, the error bars to the statistical uncertainty on the SF measurement, and the band is the total uncertainty (statistical + systematic).

References

[1] L. Evans and P. Bryant (editors), "LHC Machine", *JINST* **3** (2008) S08001, doi:10.1088/1748-0221/3/08/S08001.

[2] L. Asquith et al., "Jet Substructure at the Large Hadron Collider : Experimental Review", arXiv:1803.06991.

[3] A. J. Larkoski, I. Moult, and B. Nachman, "Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning", arXiv:1709.04464.

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

[5] CMS Collaboration, "Measurement of the integrated and differential $t\bar{t}$ production cross sections for high-p_t top quarks in pp collisions at $\sqrt{s} = 8$ TeV", *Phys. Rev. D* **94** (2016) 072002, doi:10.1103/PhysRevD.94.072002, arXiv:1605.00116.

[6] CMS Collaboration, "Measurement of the jet mass in highly boosted $t\bar{t}$ events from pp collisions at $\sqrt{s} = 8$ TeV", *Eur. Phys. J. C* **77** (2017) 467, doi:10.1140/epjc/s10052-017-5030-3, arXiv:1703.06330.

[7] CMS Collaboration, "Studies of Jet Mass in Dijet and W/Z + Jet Events", *JHEP* **05** (2013) 090, doi:10.1007/JHEP05(2013)090, arXiv:1303.4811.

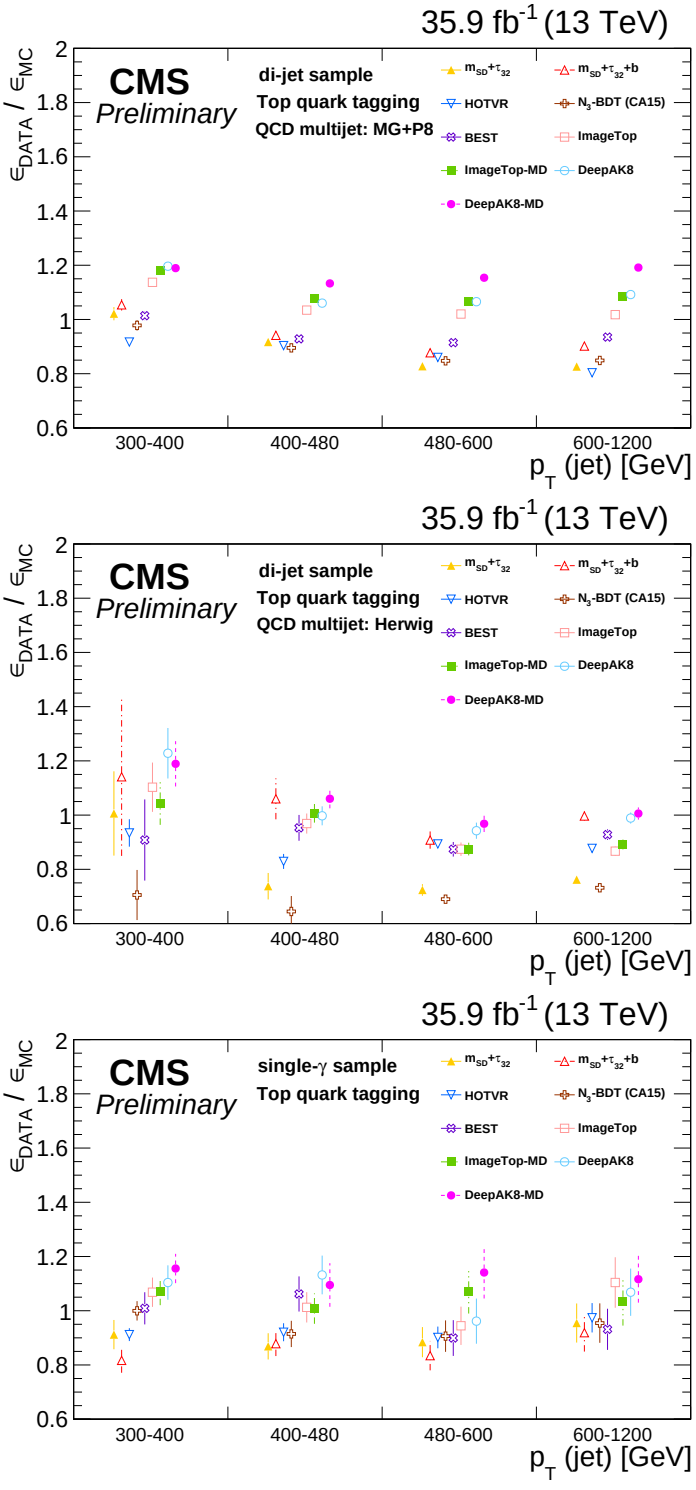

Figure 42: The ratio of the misidentification rate of t quarks in data and simulation in the di-jet (upper and middle rows) and the single-γ (lower row) samples. The QCD multijet process is simulated using MADGRAPH for the hard process and PYTHIA for parton showering (upper) and HERWIG++ for both (middle).

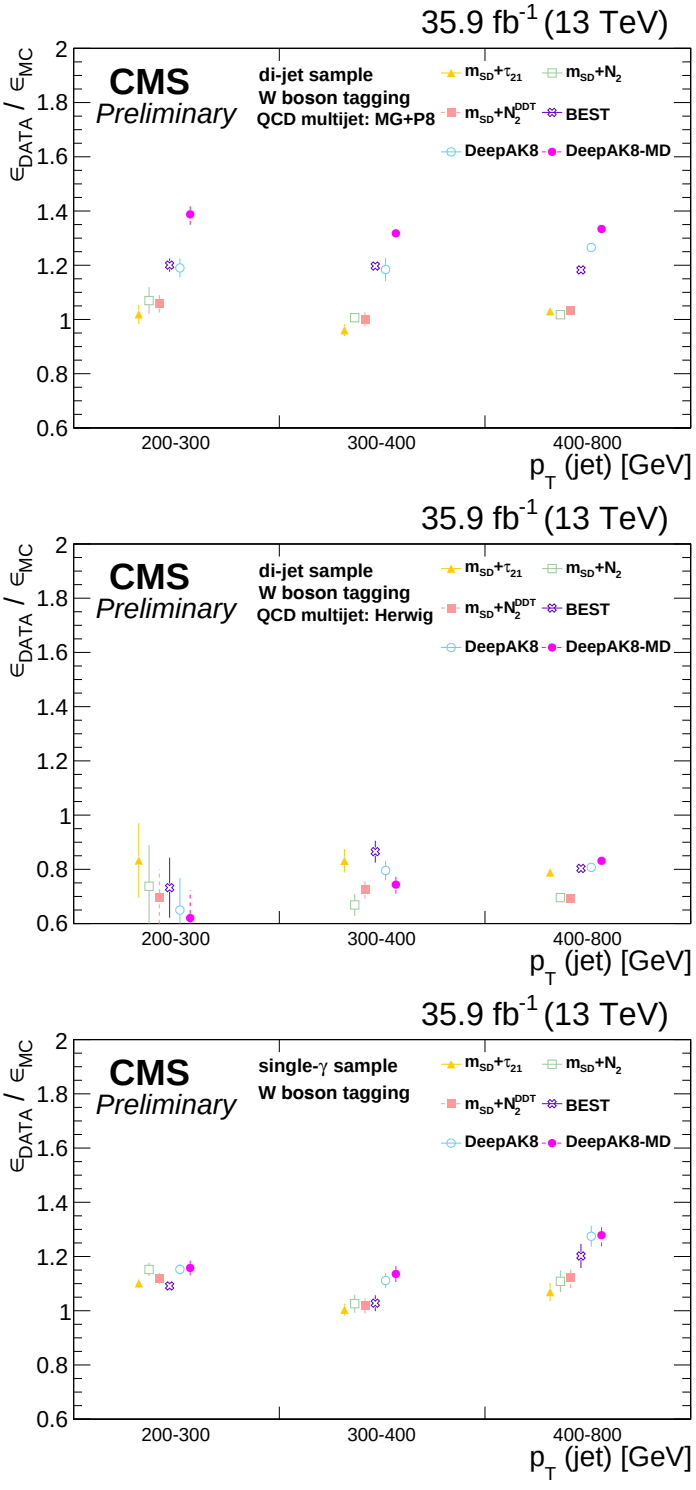

Figure 43: The ratio of the misidentification rate of W bosons in data and simulation in the di-jet (upper and middle rows) and the single-γ (lower row) samples. The QCD multijet process is simulated using MADGRAPH for the hard process and PYTHIA for parton showering (upper) and HERWIG++ for both (middle).

[8] CMS Collaboration, "Measurements of the differential jet cross section as a function of the jet mass in dijet events from proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **11** (2018) 113, doi:10.1007/JHEP11(2018)113, arXiv:1807.05974.

[9] CMS Collaboration, "Measurement of jet substructure observables in $\mathrm{t\bar{t}}$ events from proton-proton collisions at $\sqrt{s} = 13$TeV", *Phys. Rev. D* **98** (2018) 092014, doi:10.1103/PhysRevD.98.092014, arXiv:1808.07340.

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

[11] ATLAS Collaboration, "Measurement of the Soft-Drop Jet Mass in pp Collisions at $\sqrt{s} = 13$ TeV with the ATLAS Detector", *Phys. Rev. Lett.* **121** (2018) 092001, doi:10.1103/PhysRevLett.121.092001, arXiv:1711.08341.

[12] ATLAS Collaboration, "Measurement of the cross-section of high transverse momentum vector bosons reconstructed as single jets and studies of jet substructure in pp collisions at $\sqrt{s}$ = 7 TeV with the ATLAS detector", *New J. Phys.* **16** (2014) 113013, doi:10.1088/1367-2630/16/11/113013, arXiv:1407.0800.

[13] ATLAS Collaboration, "Measurements of $t\bar{t}$ differential cross-sections of highly boosted top quarks decaying to all-hadronic final states in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector", *Phys. Rev. D* **98** (2018) 012003, doi:10.1103/PhysRevD.98.012003, arXiv:1801.02052.

[14] ATLAS Collaboration, "Measurement of jet-substructure observables in top quark, W boson and light jet production in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", *Submitted to: JHEP* (2019) arXiv:1903.02942.

[15] J. Dolen et al., "Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure", *JHEP* **05** (2016) 156, doi:10.1007/JHEP05(2016)156, arXiv:1603.00027.

[16] CMS Collaboration, "Boosted top jet tagging at cms", CMS Physics Analysis Summary CMS-PAS-JME-13-007, 2014.

[17] CMS Collaboration, "Top tagging with new approaches", CMS Physics Analysis Summary CMS-PAS-JME-15-002, 2016.

[18] CMS Collaboration, "Jet algorithms performance in 13 tev data", CMS Physics Analysis Summary CMS-PAS-JME-16-003, 2017.

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

[20] CMS Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV", *JINST* **13** (2018) P05011, doi:10.1088/1748-0221/13/05/P05011, arXiv:1712.07158.

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

[22] CMS Collaboration, "Performance of photon reconstruction and identification with the CMS detector in proton-proton collisions at $\sqrt{s} = 8\,\text{TeV}$", *JINST* **10** (2015) P08010, doi:10.1088/1748-0221/10/08/P08010, arXiv:1502.02702.

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

[24] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", *JINST* **9** (2014) P10009, doi:10.1088/1748-0221/9/10/P10009, arXiv:1405.6569.

[25] CMS Collaboration, "The CMS trigger system", *JINST* **12** (2017) P01020, doi:10.1088/1748-0221/12/01/P01020, arXiv:1609.02366.

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

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

[28] P. Skands, S. Carrazza, and J. Rojo, "Tuning PYTHIA 8.1: the Monash 2013 Tune", *Eur. Phys. J. C* **74** (2014) 3024, doi:10.1140/epjc/s10052-014-3024-y, arXiv:1404.5630.

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

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

[31] S. Frixione, P. Nason, and G. Ridolfi, "A positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction", *JHEP* **09** (2007) 126, doi:10.1088/1126-6708/2007/09/126, arXiv:0707.3088.

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

[33] R. Frederix and S. Frixione, "Merging meets matching in MC@NLO", *JHEP* **12** (2012) 061, doi:10.1007/JHEP12(2012)061, arXiv:1209.6215.

[34] J. Alwall et al., "Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions", *Eur. Phys. J. C* **53** (2008) 473, doi:10.1140/epjc/s10052-007-0490-5, arXiv:0706.2569.

[35] M. Bahr et al., "Herwig++ Physics and Manual", *Eur. Phys. J.* **C58** (2008) 639–707, doi:10.1140/epjc/s10052-008-0798-9, arXiv:0803.0883.

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

[37] CMS Collaboration, "Measurement of differential cross sections for top quark pair production using the lepton + jets final state in proton-proton collisions at 13 tev", *Phys. Rev. D* **95** (May, 2017) 092001, doi:10.1103/PhysRevD.95.092001.

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

[39] CMS Collaboration, "Particle-flow reconstruction and global event description with the CMS detector", *JINST* **12** (2017) P10003, doi:10.1088/1748-0221/12/10/P10003, arXiv:1706.04965.

[40] 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.

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

[42] CMS Collaboration, "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s}$ = 8 TeV", *JINST* **10** (2015), no. 06, P06005, doi:10.1088/1748-0221/10/06/P06005, arXiv:1502.02701.

[43] CMS Collaboration, "Pileup removal algorithms", CMS Physics Analysis Summary CMS-PAS-JME-14-001, 2014.

[44] 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.

[45] M. Wobisch and T. Wengler, "Hadronization corrections to jet cross-sections in deep inelastic scattering", in *Proceedings of the Workshop on Monte Carlo Generators for HERA Physics, Hamburg, Germany*, p. 270. 1998. arXiv:hep-ph/9907280.

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

[47] CMS Collaboration, "Pile up mitigation at CMS in 13 TeV data", CMS Physics Analysis Summary CMS-PAS-JME-18-001, 2019.

[48] CMS Collaboration, "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV", *JINST* **12** (2017) P02014, doi:10.1088/1748-0221/12/02/P02014, arXiv:1607.03663.

[49] CMS Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV", *JINST* **13** (2018) P05011, doi:10.1088/1748-0221/13/05/P05011, arXiv:1712.07158.

[50] CMS Collaboration, "Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector", *Submitted to: JINST* (2019) arXiv:1903.06078.

[51] CMS Collaboration, "Searches for new physics using the $t\bar{t}$ invariant mass distribution in pp collisions at $\sqrt{s}$=8 TeV", *Phys. Rev. Lett.* **111** (2013) 211804, doi:10.1103/PhysRevLett.111.211804, 10.1103/PhysRevLett.112.119903, arXiv:1309.2030. [Erratum: Phys. Rev. Lett.112,no.11,119903(2014)].

[52] CMS Collaboration, "Search for resonant $t\bar{t}$ production in proton-proton collisions at $\sqrt{s} = 8$ TeV", *Phys. Rev. D* **93** (2016) 012001, doi:10.1103/PhysRevD.93.012001, arXiv:1506.03062.

[53] CMS Collaboration, "Search for $\mathrm{t\bar{t}}$ resonances in highly boosted lepton+jets and fully hadronic final states in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **07** (2017) 001, doi:10.1007/JHEP07(2017)001, arXiv:1704.03366.

[54] CMS Collaboration, "Search for resonant $\mathrm{t\bar{t}}$ production in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **04** (2019) 031, doi:10.1007/JHEP04(2019)031, arXiv:1810.05905.

[55] A. Butter et al., "The Machine Learning Landscape of Top Taggers", arXiv:1902.09914.

[56] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, "Towards an understanding of jet substructure", *JHEP* **09** (2013) 029, doi:10.1007/JHEP09(2013)029, arXiv:1307.0007.

[57] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, "Soft drop", *JHEP* **05** (2014) 146, doi:10.1007/JHEP05(2014)146, arXiv:1402.2657.

[58] C. Frye, A. J. Larkoski, M. D. Schwartz, and K. Yan, "Factorization for groomed jet substructure beyond the next-to-leading logarithm", *JHEP* **07** (2016) 064, doi:10.1007/JHEP07(2016)064, arXiv:1603.09338.

[59] S. Marzani, L. Schunk, and G. Soyez, "A study of jet mass distributions with grooming", *JHEP* **07** (2017) 132, doi:10.1007/JHEP07(2017)132, arXiv:1704.02210.

[60] 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.

[61] 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.

[62] S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber, "Longitudinally invariant K_t clustering algorithms for hadron hadron collisions", *Nucl. Phys. B* **406** (1993) 187–224, doi:10.1016/0550-3213(93)90166-M.

[63] S. D. Ellis and D. E. Soper, "Successive combination jet algorithm for hadron collisions", *Phys. Rev. D* **48** (1993) 3160–3166, doi:10.1103/PhysRevD.48.3160, arXiv:hep-ph/9305266.

[64] CMS Collaboration, "A multi-dimensional search for new heavy resonances decaying to boosted WW, WZ, or ZZ boson pairs in the dijet final state at 13 TeV", arXiv:1906.05977.

[65] CMS Collaboration, "Search for heavy resonances that decay into a vector boson and a Higgs boson in hadronic final states at $\sqrt{s} = 13$ TeV", *Eur. Phys. J. C* **77** (2017) 636, doi:10.1140/epjc/s10052-017-5192-z, arXiv:1707.01303.

[66] CMS Collaboration, "Search for vector-like T and B quark pairs in final states with leptons at $\sqrt{s} = 13$ TeV", *JHEP* **08** (2018) 177, doi:10.1007/JHEP08(2018)177, arXiv:1805.04758.

[67] CMS Collaboration, "Search for pair production of vector-like quarks in the fully hadronic final state", arXiv:1906.11903.

[68] CMS Collaboration, "Search for a W′ boson decaying to a vector-like quark and a top or bottom quark in the all-jets final state", *JHEP* **03** (2019) 127, doi:10.1007/JHEP03(2019)127, arXiv:1811.07010.

[69] T. Lapsien, R. Kogler, and J. Haller, "A new tagger for hadronically decaying heavy particles at the lhc", *The European Physical Journal C* **76** (2016), no. 11, 600, doi:10.1140/epjc/s10052-016-4443-8.

[70] I. Moult, L. Necib, and J. Thaler, "New angles on energy correlation functions", *Journal of High Energy Physics* **2016** (2016), no. 12, 153, doi:10.1007/JHEP12(2016)153.

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

[72] T. Plehn, M. Spannowsky, M. Takeuchi, and D. Zerwas, "Stop Reconstruction with Tagged Tops", *JHEP* **10** (2010) 078, doi:10.1007/JHEP10(2010)078, arXiv:1006.2833.

[73] G. Kasieczka et al., "Resonance Searches with an Updated Top Tagger", *JHEP* **06** (2015) 203, doi:10.1007/JHEP06(2015)203, arXiv:1503.05921.

[74] H. Voss, A. Höcker, J. Stelzer, and F. Tegenfeldt, "TMVA, the toolkit for multivariate data analysis with ROOT", in *XIth International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT)*, p. 40. 2007. arXiv:physics/0703039.

[75] CMS Collaboration, "Search for dark matter in events with energetic, hadronically decaying top quarks and missing transverse momentum at $\sqrt{s} = 13$ TeV", *JHEP* **06** (2018) 027, doi:10.1007/JHEP06(2018)027, arXiv:1801.08427.

[76] CMS Collaboration, "Search for low mass vector resonances decaying into quark-antiquark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **01** (2018) 097, doi:10.1007/JHEP01(2018)097, arXiv:1710.00159.

[77] CMS Collaboration, "Inclusive search for a highly boosted Higgs boson decaying to a bottom quark-antiquark pair", *Phys. Rev. Lett.* **120** (2018) 071802, doi:10.1103/PhysRevLett.120.071802, arXiv:1709.05543.

[78] CMS Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV", *JINST* **13** (2018) P05011, doi:10.1088/1748-0221/13/05/P05011, arXiv:1712.07158.

[79] J. S. Conway, R. Bhaskar, R. D. Erbacher, and J. Pilot, "Identification of high-momentum top quarks, higgs bosons, and w and z bosons using boosted event shapes", *Phys. Rev. D* **94** (Nov, 2016) 094027, doi:10.1103/PhysRevD.94.094027.

[80] G. C. Fox and S. Wolfram, "Observables for the analysis of event shapes in e^+e^- annihilation and other processes", *Phys. Rev. Lett.* **41** (1978) 1581, doi:10.1103/PhysRevLett.41.1581.

[81] J. D. Bjorken and S. J. Brodsky, "Statistical model for electron-positron annihilation into hadrons", *Phys. Rev. D* **1** (1970) 1416, doi:10.1103/PhysRevD.1.1416.

[82] E. Farhi, "Quantum chromodynamics test for jets", *Phys. Rev. Lett.* **39** (1977) 1587, doi:10.1103/PhysRevLett.39.1587.

[83] F. Pedregosa et al., "Scikit-learn: Machine learning in Python", *J. Mach. Learn. Res.* **12** (2011) 2825–2830, arXiv:1201.0490.

[84] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines", in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 807–814. Omnipress, USA, 2010.

[85] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", 2014. http://arxiv.org/abs/1412.6980.

[86] S. Macaluso and D. Shih, "Pulling Out All the Tops with Computer Vision and Deep Learning", *JHEP* **10** (2018) 121, doi:10.1007/JHEP10(2018)121, arXiv:1803.00107.

[87] G. Kasieczka, T. Plehn, M. Russell, and T. Schell, "Deep-learning Top Taggers or The End of QCD?", *JHEP* **05** (2017) 006, doi:10.1007/JHEP05(2017)006, arXiv:1701.08784.

[88] M. D. Zeiler, "ADADELTA: an adaptive learning rate method", *CoRR* **abs/1212.5701** (2012) arXiv:1212.5701.

[89] CMS Collaboration, "Performance of b tagging algorithms in proton-proton collisions at 13 TeV with phase 1 CMS detector", CMS Detector Performance Note CMS-DP-2018-033, 2018.

[90] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", *CoRR* **abs/1512.03385** (2015) arXiv:1512.03385.

[91] N. Srivastava et al., "Dropout: A simple way to prevent neural networks from overfitting", *Journal of Machine Learning Research* **15** (2014) 1929.

[92] T. Chen et al., "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems", *CoRR* **abs/1512.01274** (2015) arXiv:1512.01274.

[93] G. Louppe, M. Kagan, and K. Cranmer, "Learning to pivot with adversarial networks", in *Advances in Neural Information Processing Systems 30*, I. Guyon et al., eds., p. 981. Curran Associates, Inc., 2017.

[94] CMS Collaboration, "Search for production of Higgs boson pairs in the four b quark final state using large-area jets in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **01** (2019) 040, doi:10.1007/JHEP01(2019)040, arXiv:1808.01473.

[95] CMS Collaboration, "Search for low-mass resonances decaying into bottom quark-antiquark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV", *Phys. Rev. D* **99** (2019) 012005, doi:10.1103/PhysRevD.99.012005, arXiv:1810.11822.

[96] J. Lin, "Lin jh.. divergence measures based on the shannon entropy. ieee trans inform theory 37: 145-151", *IEEE Transactions on Information Theory* **37** (01, 1991) 145, doi:10.1109/18.61115.

[97] S. Kullback and R. A. Leibler, "On information and sufficiency", *Ann. Math. Statist.* **22** (1951), no. 1, 79–86.

[98] CMS Collaboration, "Search for the standard model higgs boson decaying to charm quarks", CMS Physics Analysis Summary CMS-PAS-HIG-18-031, 2019.

[99] CMS Collaboration, "Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 13$ TeV", *JHEP* **07** (2018) 161, doi:10.1007/JHEP07(2018)161, arXiv:1802.02613.

[100] ATLAS Collaboration, "Measurement of the Inelastic Proton-Proton Cross Section at $\sqrt{s} = 13$ TeV with the ATLAS Detector at the LHC", *Phys. Rev. Lett.* **117** (2016), no. 18, 182002, doi:10.1103/PhysRevLett.117.182002, arXiv:1606.02625.