CMS Physics Analysis Summary

Performance of reconstruction and identification of τ leptons in their decays to hadrons and ν_τ in LHC Run-2

The CMS Collaboration

Abstract

The CMS hadrons-plus-strips (HPS) algorithm that was developed to reconstruct tau leptons in their hadronic decays demonstrated a good performance in the LHC Run-1; the algorithm achieves an identification efficiency of $50 - 60\%$ with a probability for quark and gluon jets, electrons, and muons to be misidentified as τ lepton between per cent and per mille levels. In this paper improvements to the HPS algorithm for the LHC Run-2 are described. The performance is evaluated with a sample of proton-proton collisions recorded at a center-of-mass energy of $\sqrt{s} = 13$ TeV in 2015. The data sample corresponds to a total integrated luminosity of 2.3 fb^{-1}.

1 Introduction

Decays to τ leptons constitute an important experimental signature for physics analyses at the LHC: the study of the standard model (SM) Higgs boson that decays into a pair of τ leptons [1, 2], searches for neutral and charged Higgs bosons in the context of the Minimal Supersymmetric SM (MSSM) [3–7], the study of SM processes (Drell-Yan, W boson and a pair production of top quarks) [8–10], and searches for supersymmetry [11–16], leptoquarks [17, 18], W′ [19] bosons and many more [20–25].

With a mass of $m_\tau = 1.777$ GeV [26], the τ lepton is the only lepton heavy enough to decay into hadrons. In about one third of the cases τ's decay leptonically to a muon (τ_μ) or an electron (τ_e) with two neutrinos, and are reconstructed and identified with the usual techniques for muons and electrons [27, 28]. In the remaining cases, τ leptons decay hadronically, to a combination of charged and neutral mesons with a τ neutrino. Hadronically decaying τ's, denoted by τ_h, are reconstructed and identified with the hadrons-plus-strips (HPS) algorithm [29, 30], which was developed for use in the LHC Run-1. The challenge the algorithm has to face is the distinction between genuine τ_h, and quark and gluon jets, which are copiously produced in QCD multijet process and can be misidentified as τ_h. The main handle for reducing these jet $\to \tau_h$ misidentification backgrounds is to utilize the fact that the particles produced in τ_h decays are of lower multiplicity, deposit energy in a narrow region compared to a quark or gluon jet, and are typically isolated with respect to other particles in the event. In some physics analyses, the misidentification of electrons or muons as τ_h candidates may constitute a sizeable background as well.

After the LHC Run-1, the HPS algorithm has been improved in light of 13 TeV data-taking in Run-2. These improvements include

- an improved strip reconstruction algorithm, called dynamic strip reconstruction, that accounts for electromagnetic energy leakage of the τ_h decay;
- an improved multivariate analysis (MVA) based discriminant combining isolation, τ-lifetime and shower shape information to suppress jet $\to \tau_h$ misidentification;
- improvements to the MVA based discriminator to suppress e $\to \tau_h$ misidentification.

After introducing the CMS detector (Section 2), samples used for the performance evaluation (Section 3), and reconstruction of other physics objects (Section 4), Section 5 describes the improved algorithm used in Run-2, with a particular emphasis on the improvements with respect to Run-1 [30]. The event selection in data to evaluate the performance is discussed in Section 6. The performance evaluation of the improved algorithm using data will be presented thereafter: Section 7 describes the tau identification efficiency, Section 8 the τ_h energy scale measurement, Section 9 the charge misidentification probability measurement. Section 10 and 11 summarize the jet $\to \tau_h$ and e $\to \tau_h$ misidentification probability measurements, respectively. The analysis is performed using proton-proton collision data from the LHC Run-2, recorded at $\sqrt{s} = 13$ TeV in 2015 with an integrated luminosity of 2.3 fb^{-1}.

2 The CMS detector

The central feature of the CMS apparatus [31] 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 tungstate crystal electromagnetic calorimeter (ECAL), and a brass and scintillator hadron calorimeter (HCAL), each composed of a barrel and two endcap sections. Muons are measured in gas-ionization detectors embedded in the steel flux-

return yoke outside the solenoid. Extensive forward calorimetry complements the coverage provided by the barrel and endcap detectors.

The CMS tracker is a cylindric detector, constructed from silicon modules, and covers geometric acceptance of $|\eta| \leq 2.5$. Tracks of charged hadrons are reconstructed with an efficiency of typically 80-90%, varying with transverse momentum (p_T) and η [32, 33]. A consequence of the silicon tracker is a significant amount of material in front of the ECAL, mainly due to the mechanical structure, the services, and the cooling system. The minimum number of radiation lengths is $0.4\,X_0$ at $|\eta| \approx 0$, rises to about $2.0\,X_0$ at $|\eta| \approx 1.4$, and decreases to about $1.3\,X_0$ at $|\eta| \approx 2.5$. Photons originating from $\pi^0 \rightarrow \gamma\gamma$ decays therefore have a high probability to convert to e^+e^- pairs within the volume of the tracking detector.

The ECAL is a homogeneous and hermetic calorimeter made of $PbWO_4$ scintillating crystals. It is composed of a central barrel, covering the pseudorapidity region $|\eta| \leq 1.479$, plus two endcaps, covering $1.479 \leq |\eta| \leq 3.0$. The small radiation length ($X_0 = 0.89\,cm$) and small Moliere radius (2.3 cm) of the $PbWO_4$ crystals provide a compact calorimeter with excellent two-shower separation.

3 Data and simulated samples

The τ_h reconstruction and identification is validated using proton-proton collision data delivered by the LHC at a center-of-mass energy of $\sqrt{s} = 13\,TeV$ in 2015. The analyzed dataset corresponds to an integrated luminosity of $2.3\,fb^{-1}$.

The data is compared to the Monte Carlo (MC) simulated events, using samples of $Z/\gamma^* \rightarrow \ell\ell$ ($\ell = e, \mu, \tau$), W+jet, $t\bar{t}$, single top, di-boson (WW, WZ, ZZ) and QCD multi-jet events. The W+jet, $Z/\gamma^* \rightarrow \ell\ell$ samples are generated with MADGRAPH [34], while the single top and $t\bar{t}$ samples are generated with POWHEG [35]. The diboson samples are generated either with MADGRAPH or POWHEG. The QCD multi-jet samples are produced using PYTHIA8 [36] generator. The PYTHIA8 generator, with the tune *CUETP8M1* [37], is used to model parton shower and hadronization processes, as well as tau decays in all samples. The $Z/\gamma^* \rightarrow \ell\ell$, W+jet and $t\bar{t}$ samples are normalized according to cross sections computed at NNLO accuracy [38]. The cross sections for single top and di-boson production are computed at NLO accuracy [39]. The NNPDF3.0 parton distribution functions are used [40].

In addition, simulated signal samples of $H \rightarrow \tau\tau$, $Z' \rightarrow \ell\ell$ (with mass up to 5 TeV), $W' \rightarrow \ell\bar{\nu}_\ell$ (with mass up to 5.8 TeV) and MSSM $H \rightarrow \tau\tau$ (with mass up to 3.2 TeV) are used to optimize the τ_h identification over a wide p_T range. The $H \rightarrow \tau\tau$ sample is produced with POWHEG, while Z' and W' samples with PYTHIA8. For the studies herein the reconstructed τ_h candidates are required to be matched to generated τ_h within a cone of size $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} < 0.3$.

Minimum bias events generated with PYTHIA8 are added to all MC simulated events according to the luminosity profile of the analyzed data. All generated events are passed through a detailed simulation of the CMS apparatus, based on GEANT [41], and are reconstructed using the same CMS event reconstruction software as the data.

4 Object reconstruction

A particle-flow (PF) algorithm [42–45] combines the information from the CMS subdetectors to identify and reconstruct the particles emerging from proton-proton collisions: charged hadrons, neutral hadrons, photons, muons, and electrons. These particles are then used to reconstruct

the missing transverse energy vector, the jets, the τ_h candidates, and to quantify the lepton isolation.

Electrons are reconstructed from clusters of energy deposits in the ECAL and the trajectory in the inner tracker [28]. Electron identification relies on the electromagnetic shower shape and other observables based on tracker and calorimeter information. The selection criteria depend on p_T and $|\eta|$, and on a categorization according to observables that are sensitive to the amount of bremsstrahlung emitted along the trajectory in the tracker.

Muons are reconstructed from both the inner tracker and the muon spectrometer [27]. The PF muons are selected from among the reconstructed muon track candidates by applying minimal requirements on the track components in the muon system and taking into account matching with small energy deposits in the calorimeters.

The isolation of individual electrons or muons is measured relative to their transverse momentum p_T^ℓ, by summing over the transverse momenta of charged hadrons and neutral particles in a cone with $\Delta R < 0.4$ around the lepton direction at the interaction vertex:

$$R_{\text{Iso}}^\ell = \left(\sum p_T^{\text{charged}} + \max \left[0, \sum p_T^{\text{neutral}} + \sum p_T^\gamma - p_T^{\text{PU}}(\ell) \right] \right) / p_T^\ell. \qquad (1)$$

The $\sum p_T^{\text{charged}}$ is the scalar sum of the transverse momenta of charged hadrons originating from the primary vertex. The primary vertex is chosen as the vertex with the highest sum of p_T^2 of its constituent tracks. The $\sum p_T^{\text{neutral}}$ and $\sum p_T^\gamma$ are the scalar sums of the transverse momenta for neutral hadrons and photons, respectively. The neutral contribution to isolation from pileup, $p_T^{\text{PU}}(\ell)$, is estimated using a *jet area* method [46] for the electron. For the muon, the $p_T^{\text{PU}}(\ell)$ contribution is estimated using sum of transverse momenta of charged hadrons not originating from the primary vertex scaled by a factor of 0.5.

Jets are reconstructed using PF objects. The anti-k_T jet clustering algorithm [47] with R = 0.4 is used. The standard jet energy corrections [48] are applied. In order to reject jets formed from pileup collisions, a multivariate-based jet identification algorithm is applied. This algorithm takes advantage of differences in the shape of energy deposits in a jet cone. To identify the jet originated from b-quark, the CSV2 b-tagging algorithm [49] with the medium working point is used.

The missing transverse momentum vector, $\vec{p}_T^{\text{miss}}$, is defined as the projection on the plane perpendicular to the beams of the negative vector sum of the momenta of all reconstructed particles in an event, corrected for pileup contribution. Its magnitude is referred to as E_T^{miss}.

The reconstruction and identification of τ_h will be described separately in Section 5.

5 Tau identification algorithm for Run-2

Hadronic decays of τ leptons are reconstructed and identified with the HPS algorithm [29, 30]. The basic features of the algorithm are identical to those of Run-1 and are briefly summarized in section 5.1. Section 5.2 describes the main modification implemented in Run-2: the dynamic strip reconstruction. Sections 5.3 and 5.4 summarize isolation discriminators against jets and against electrons, respectively, further developed in the context of the LHC Run-2.

5.1 Hadrons plus strips algorithm

The HPS algorithm is seeded by the reconstructed jets. The algorithm looks into the constituents of the jets to reconstruct the neutral pions that are present in most τ_h decays. The high probability for photons originating from $\pi^0 \rightarrow \gamma\gamma$ decays to convert to e^+e^- pairs is accounted for by collecting the photon and electron ($p_T > 0.5\,\mathrm{GeV}$) constituents of the jet into clusters (*strips*). The size of the strips is set to a fixed value of 0.05×0.20 in the $\eta - \phi$ direction in the Run-1 HPS algorithm. Strips containing one or more electron or photon constituents and passing a cut of $p_T > 2.5\,\mathrm{GeV}$ on the transverse momentum sum of electrons plus photons included in the strip are kept as π^0 candidates for further processing. The τ_h candidates are formed by combining the strips with the charged-particle constituents of the jet. Based on the observed number of strips and charged particles, it is assigned to be one of the following *decay modes*:

- a single charged particle without any strips: $h^\pm$;
- combination of one charged particle and one strip: $h^\pm\pi^0$;
- combination of a single charged particle with two strips: $h^\pm\pi^0\pi^0$;
- combination of three charged particles: $h^\pm h^\mp h^\pm$.

5.2 Dynamic strip reconstruction

After Run-1, additional studies were performed in order to optimize the strip size. In practice, there were cases where τ_h decay products contributed to the isolation, such as:

- A charged pion from τ_h decay experiences nuclear interaction with tracker material and produces several secondary particles with low p_T. This ends up with low p_T electrons and photons that go outside strip window. This will affect the isolation of the τ_h, although it is part of the τ_h decay product.
- Photons from $\pi^0 \rightarrow \gamma\gamma$ have a large probability to convert to an e^+e^- pair and, after multiple conversion and bremsstrahlung, electrons and photons may go outside the fixed size window. This will also affect the isolation.

Naïvely, these decay products can be integrated as part of the signal by suitably widening the strip size. On the contrary, if the τ_h has a large p_T the decay product tend to be boosted in the τ_h flight direction. In this case, a smaller strip size than that considered in Run-1 [30] can reduce background contributions in the strip while accounting for all τ_h decay products.

Based on these considerations, the strip reconstruction of the HPS algorithm has been improved for Run-2, and proceeds as follows:

(i) The highest p_T electron or photon (e/γ) not yet included in any strip is used to seed a new strip. The initial position of the strip in η and ϕ is set to the η and ϕ of the seed e/γ.

(ii) The next highest p_T e/γ that is within,

$$\begin{aligned}
\Delta\eta &= f(p_T^\gamma) + f(p_T^{\mathrm{strip}}) \\
\Delta\phi &= g(p_T^\gamma) + g(p_T^{\mathrm{strip}}).
\end{aligned} \tag{2}$$

centered on the strip location is merged into the strip. The functions f and g are determined using a single τ gun MC sample, such that 95% of all electrons and photons, that are due to τ_h decay products, are contained within a strip. The functional form is derived

based on the distance in η and in ϕ between τ_h and e/γ, that are due to tau decay products, as a function of e/γ p_T (Fig. 1). The 95% envelope point in each bin is fitted by the analytical form, a/p_T^b, and following function is obtained (red line in Fig. 1):

$$
\begin{aligned}
f(p_T) &= 0.20 \cdot p_T^{-0.66} \\
g(p_T) &= 0.35 \cdot p_T^{-0.71} .
\end{aligned} \tag{3}
$$

The upper limit of the strip size is set to 0.3 in $\Delta\phi$ and 0.15 in $\Delta\eta$; the lower limit is set to 0.05 for both $\Delta\phi$ and $\Delta\eta$.

Note that the window size also depends on strip p_T, as the strip (after all, it is electrons or photons) can change its position due to particles trajectories changing in the magnetic field. Therefore, the size is determined by considering as a maximum difference between two objects assuming they have different charges and are produced back-to-back in the rest frame. Although this argument can be applied for the ϕ direction only, there are still some cases when the objects can be separated in η direction.

(iii) The strip position is recomputed as a p_T weighted average of all e/γ constituents of the strip:

$$
\begin{aligned}
\eta_\mathrm{strip} &= \frac{1}{p_T^\mathrm{strip}} \cdot \sum p_T^\gamma \cdot \eta_\gamma \\
\phi_\mathrm{strip} &= \frac{1}{p_T^\mathrm{strip}} \cdot \sum p_T^\gamma \cdot \phi_\gamma ,
\end{aligned}
$$

with $p_T^\mathrm{strip} = \sum p_T^\gamma$.

(iv) The construction of the strip ends in case no further e/γ is within the $\eta \times \phi$ window. In this case the clustering proceeds by constructing a new strip, that is seeded by the next highest p_T e/γ.

Finally, mass window cuts [30] are applied to test the compatibility of each hypothesis with the signatures expected for the different hadronic τ decay modes: $0.3 - \Delta m_\tau < m_\tau < 1.3 \cdot \sqrt{p_T/100} + \Delta m_\tau$ for $h^\pm \pi^0$ decay mode, $0.4 - \Delta m_\tau < m_\tau < 1.2 \cdot \sqrt{p_T/100} + \Delta m_\tau$ for $h^\pm \pi^0 \pi^0$ decay mode, and $0.8 - \Delta m_\tau < m_\tau < 1.5 + \Delta m_\tau$ for $h^\pm h^\mp h^\pm$ decay mode. The Δm_τ is the change in the mass of the τ_h object through the addition of the e/γ object of the dynamic strip. It is calculated as,

$$
\Delta m_\tau = \sqrt{\left(\frac{\partial m_\tau}{\partial \eta_\mathrm{strip}} \cdot f(p_T^\mathrm{strip}) \right)^2 + \left(\frac{\partial m_\tau}{\partial \phi_\mathrm{strip}} \cdot g(p_T^\mathrm{strip}) \right)^2} ,
$$

with:

$$
\begin{aligned}
\frac{\partial m_\tau}{\partial \eta_\mathrm{strip}} &= \frac{p_z^\mathrm{strip} \cdot E_\tau - E_\mathrm{strip} \cdot p_z^\tau}{m_\tau} \\
\frac{\partial m_\tau}{\partial \phi_\mathrm{strip}} &= \frac{-\left(p_y^\tau - p_y^\mathrm{strip} \right) \cdot p_x^\mathrm{strip} + \left(p_x^\tau - p_x^\mathrm{strip} \right) \cdot p_y^\mathrm{strip}}{m_\tau} .
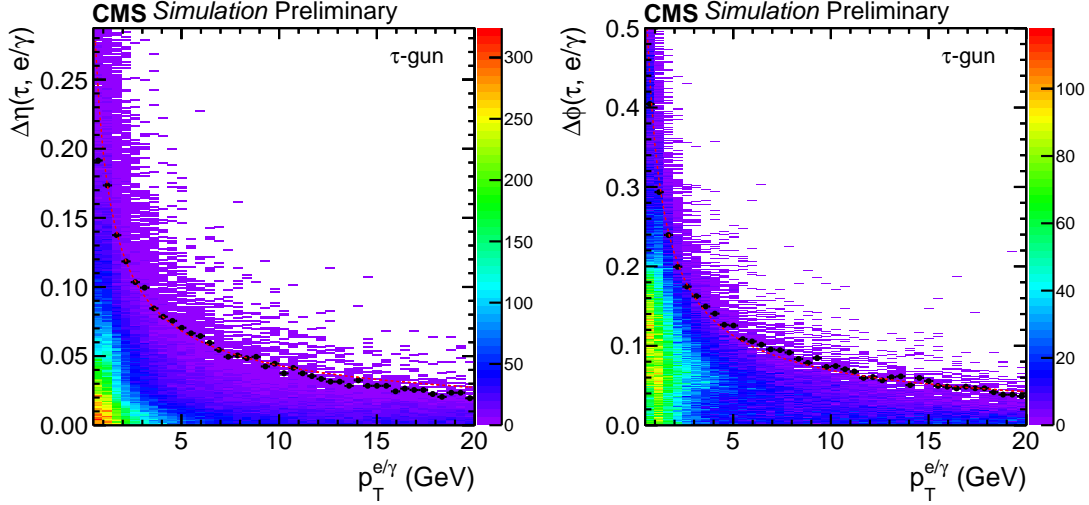\end{aligned}
$$

Figure 1: Distance in η (left) and in ϕ (right) between τ_h and e/γ, that are due to tau decay products, as a function of e/γ p_T. A sample of simulated τ_h decays is used. The size of the window is larger in the ϕ-direction due to magnetic bending. The dotted point shows 95% quantile for the given bin, and the dashed lines represent the fitted functions f and g given by Eq. 3.

5.3 Tau discrimination against jets with dynamic strip reconstruction

Requiring τ_h candidates to pass isolation requirements provides a strong handle for reducing the jet $\rightarrow \tau_h$ misidentification probability. The two types of τ_h isolation discriminators developed in Run-1, namely the isolation-sum and the MVA-based discriminators, have been further optimized and trained for Run-2. A cone with $\Delta R = 0.5$ has been used for evaluation of isolation performance, however, in busy environments (such as Higgs boson production in association with top quarks), it is also useful to use a smaller cone with $\Delta R = 0.3$.

5.3.1 Isolation-sum discriminators

The τ_h candidate isolation is computed by summing the transverse momenta of charged particles of $p_T^{\text{charged}} > 0.5\,\text{GeV}$ and photons of $p_T^{\gamma} > 0.5\,\text{GeV}$ reconstructed with the PF algorithm, within an isolation cone of size $\Delta R = 0.5(0.3)$, centered around the τ_h direction. Charged hadrons and photons that are constituents of the τ_h candidate are excluded from the p_T-sum. The contribution from pileup is suppressed by requiring the tracks associated to charged particles to originate from the production vertex of the τ_h candidate, within $d_z < 0.2\,\text{cm}$. The effect of pileup on photon isolation is estiamted by summing the transverse momenta of charged particles within a cone of size $\Delta R = 0.8$ around the τ_h direction, and with tracks not originating from the τ_h production vertex, $d_z > 0.2\,\text{cm}$. It is then subtracted from photon isolation by multiplying $\Delta\beta$, taking into account the ratio of neutral to charged hadron production in the hadronization process of inelastic proton-proton collisions, as well as the different cone size used for the pileup estimation:

$$I_\tau = \sum p_T^{\text{charged}}(d_Z < 0.2\,\text{cm}) + \max\left(0, \sum p_T^{\gamma} - \Delta\beta \sum p_T^{\text{charged}}(d_Z > 0.2\,\text{cm})\right). \tag{4}$$

In Run-2, $\Delta\beta = 0.2$ is used, obtained by fitting the $\sum p_T^{\text{charged}}(d_z > 0.2\,\text{cm})$ as a function of $\sum p_T^{\gamma}$, in order to make the τ_h identification efficiency insensitive to pileup. During Run-1, an

empirical factor of 0.46 was used to scale the $\Delta\beta$ corrections [30]; it was found to overcorrect the pileup contribution to the isolation.

The Loose, Medium, and Tight working points of the isolation-sum discriminators are defined by requiring I_τ to be less than 2.5, 1.5, and 0.8 GeV, respectively. The thresholds are chosen to keep the τ_h identification efficiency equidistant between the three working points.

In addition to I_τ, the dynamic strip reconstruction provides another handle to further reduce the jet $\to \tau_h$ misidentification probability. A cut on the p_T-sum of e/γ that are included in the strips used to reconstruct the τ_h candidate but are outside of the signal cone, $R_{sig} = 3.0/p_T$ (GeV), is applied as follows:

$$p_T^{\text{strip, outer}} = \sum p_T^{e/\gamma}(\Delta R > R_{\text{sig}}) < 0.10 \cdot p_T^\tau, \tag{5}$$

where the sum extends over all e/γ included in any strip. The upper (lower) limit of R_{sig} is set to be 0.1 (0.05).

The expected performance of the $\Delta\beta$ corrected isolation is shown in Fig. 2, as a comparison between the Run-1 and Run-2 algorithms. To see where the improvement comes from, the Run-1 performance is separately shown for $\Delta\beta = 0.46$, $\Delta\beta = 0.46$ with $p_T^{\text{strip, outer}}$ selection, and retuned $\Delta\beta = 0.2$ with $p_T^{\text{strip, outer}}$ selection. As a signal process, two different MC samples are used; one is H $\to \tau\tau$ and the other is Z$'\to \tau\tau$ (2 TeV). The QCD MC sample is used as a background process with jet p_T up to 100 GeV (for the ROC curve with H $\to \tau\tau$ as a signal) and up to 1000 GeV (for the ROC curve with Z$'\to \tau\tau$ (2 TeV) as a signal), such that its p_T coverage is similar to the corresponding signal process. By comparing the left and right distributions in Fig. 2, the gain of the dynamic strip reconstruction can be seen in high p_T τ_h's, as expected. For the H $\to \tau\tau$ process, the performance of the medium and tight working points slightly improved compared to Run-1. However, in the high efficiency region, the misidentification probability starts to increase faster not only because the τ_h p_T tends to decrease, but also because the strip size increases. As a result, the $p_T^{\text{strip, outer}}$ cut becomes tight.

5.3.2 MVA-based discriminators against jets

As an alternative to the isolation-sum discriminator, the MVA τ_h identification discriminator has been in use since Run-1. It combines the isolation and shape variables with variables sensitive to τ-lifetime information to provide the best possible discrimination of τ_h decays from quark and gluon jets. A Boosted Decision Tree (BDT) is used to achieve a strong reduction in the jet $\to \tau_h$ misidentification probability. The MVA identification method and the variables used as input to the BDT are discussed in [30].

In addition to the variables used in Run-1, a few more variables have been included in Run-2:

- Shape variables: $p_T^{\text{strip, outer}}$ (Eq. 5) and p_T-weighted ΔR, $\Delta\eta$ and $\Delta\phi$ (with respect to the τ_h axis) of photons and electrons in strips inside or outside of signal cone,
- τ-lifetime information: the signed impact parameter of the leading track of the τ_h candidate, and its significance,
- Multiplicity: the total number of photon and electron candidates ($p_T > 0.5$ GeV) in signal and isolation cones.

The BDT is retrained using simulated samples for Run-2. The τ_h candidates are selected with $p_T > 20$ GeV and $|\eta| < 2.3$. Samples of Z/$\gamma^* \to \tau\tau$, W $\to \tau\nu$, H $\to \tau\tau$, Z$' \to \tau\tau$, and W$' \to \tau\nu$ events are used to model the τ_h decays. QCD multi-jet, W+jets, and t$\bar{t}$+jets events
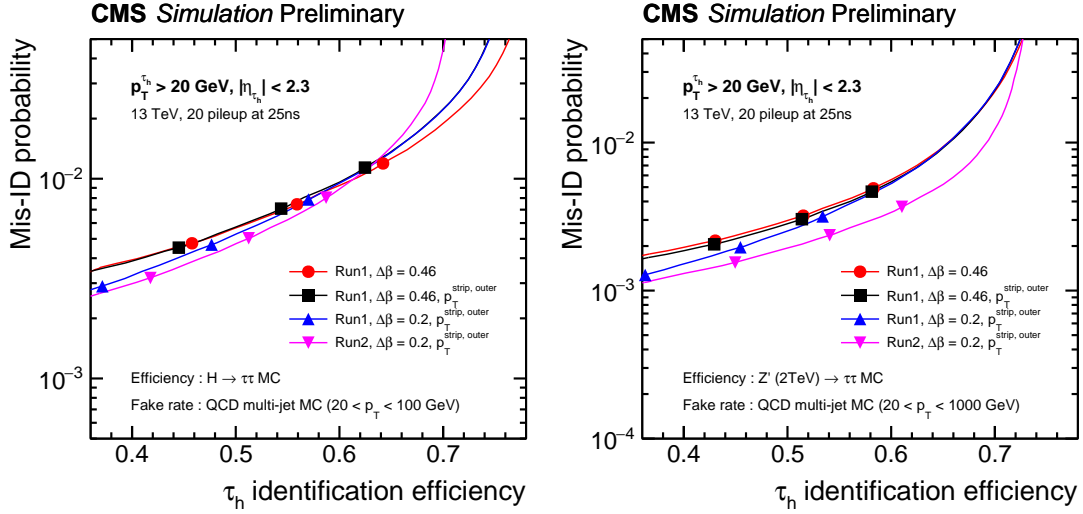
Figure 2: Misidentification probability as a function of τ_h identification efficiency, evaluated using H $\to$ $\tau\tau$ and QCD MC samples (left), and Z$'$ (2 TeV) and QCD MC samples (right). Four different configurations of reconstruction plus isolation method are compared (from top to bottom): Run-1 fixed size strip with $\Delta\beta = 0.46$, Run-1 fixed size strip with $\Delta\beta = 0.46$ and $p_T^{\text{strip, outer}}$ cut, Run-1 fixed size strip with $\Delta\beta = 0.2$ and $p_T^{\text{strip, outer}}$ cut, Run-2 dynamic strip with $\Delta\beta = 0.2$ and $p_T^{\text{strip, outer}}$ cut. The three points on each curve correspond to, from left to right, the Tight, Medium and Loose working point. The misidentification probability is calculated with respect to jets, which pass minimal τ reconstruction requirements.

are used to model quark and gluon jets. The events used for the training are reweighted such that the two-dimensional p_T and η distribution of the τ_h candidates for signal and background are identical, which makes the MVA result independent of event kinematics.

The working points of the MVA isolation discriminator, corresponding to different τ_h identification efficiencies, are defined by cuts on the BDT discriminant. For a given working point, the threshold on the BDT discriminant is adjusted as a function of the transverse momentum of the τ_h candidate to ensure a uniform efficiency over p_T.

The expected jet $\to \tau_h$ misidentification probability as a function of the expected τ_h identification efficiency for both isolation-sum and MVA-based discriminators are shown in Fig. 3, which demonstrates a reduction of misidentification probability by a factor of 2 at a similar efficiency with respect to the cut-based approach. For signal process, two different MC samples are used; one is H $\to \tau\tau$ and the other is Z$' \to \tau\tau$ (2 TeV). The QCD MC sample is used as a background process with jet p_T up to 100 GeV (for the ROC curve with H $\to \tau\tau$ as a signal) and up to 1000 GeV (for the ROC curve with Z$' \to \tau\tau$ (2 TeV) as a signal). Fig. 4 shows τ_h identification efficiency (left) and the misidentification probability (right) as a function of the τ_h p_T and jet p_T, respectively. The efficiency is computed from Z $\to \tau\tau$ events, while the expected jet $\to \tau_h$ misidentification probability is computed from QCD multi-jet events with jet p_T up to 100 GeV.

5.4   Tau discriminators against electrons

Isolated electrons have a high probability to be misidentified as τ_h's that decay to either $h^{\pm}$ or $h^{\pm}\pi^0$. In particular, electrons crossing the tracker material often emit bremsstrahlung photons mimicking π^0s in the decay mode reconstruction. An improved version of the multivariate electron discriminator used during Run-1 [30] has been developed in order to further reduce
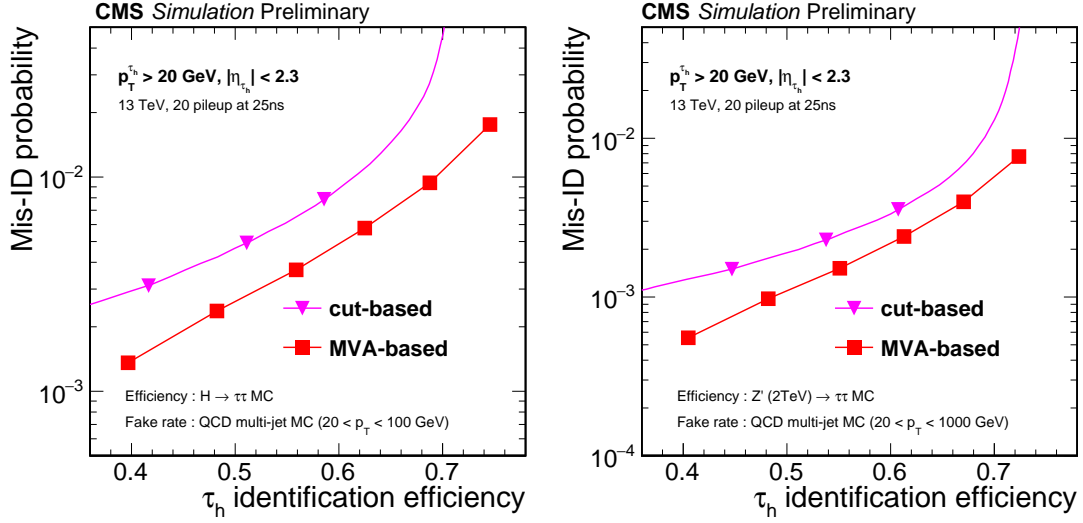
Figure 3: Misidentification probability as a function of τ_h identification efficiency, evaluated using H $\rightarrow \tau\tau$ and QCD MC samples (left), and Z$^{'}$ (2 TeV) and QCD MC samples (right). The MVA-based discriminators are compared to that of the isolation sum discriminators. The points correspond to working points of the discriminators. The three working points of the isolation sum discriminator are Loose, Medium, and Tight working point. The six working points of the MVA-based discriminators are Very Loose, Loose, Medium, Tight, Very Tight, and Very Very Tight working point, respectively. The misidentification probability is calculated with respect to jets, which pass minimal τ reconstruction requirements.
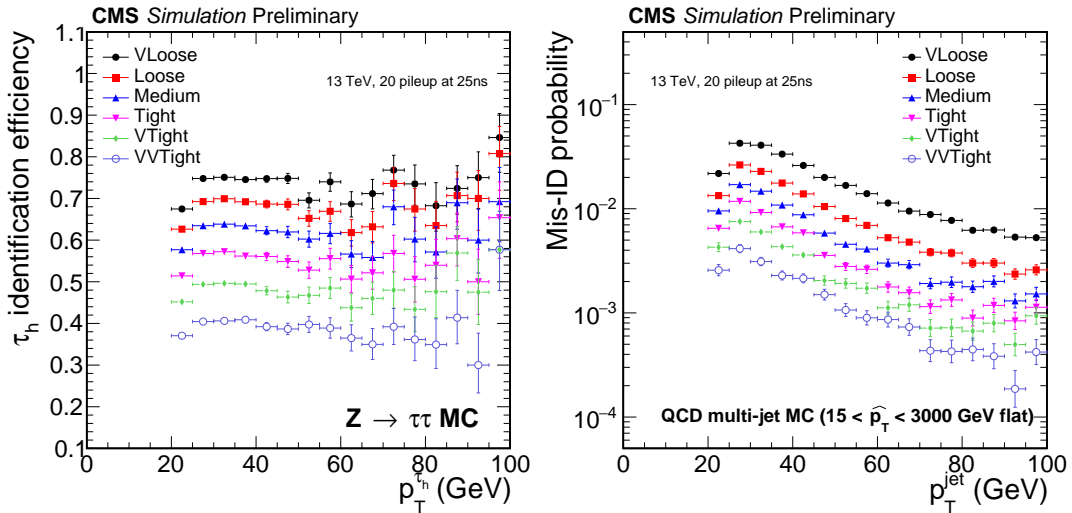

Figure 4: Efficiency of the τ_h identification estimated with simulated $Z/\gamma^* \rightarrow \tau\tau$ events (left) and the misidentification probability estimated with simulated QCD multi-jet events (right) for the Very Loose, Loose, Medium, Tight, Very Tight, and Very Very Tight working points of the MVA based τ_h isolation algorithm. The efficiency is shown as a function of the τ_h transverse momentum while the misidentification probability is shown as a function of the jet transverse momentum.

the e $\rightarrow \tau_h$ misidentification probability while maintaining a high efficiency on real τ_h decays over a wide p_T range. The variables used as input for the BDT are identical to the ones described in Ref. [30], with the addition of the following photon-related variables:

- the number of photons in any of the strips associated with the τ_h candidate;
- the p_T-weighted root-mean-square of the distances in η and ϕ between all photons included in any strip and the leading track of the τ_h candidate;
- the fraction of τ_h energy carried by photons.

These variables are computed separately for photons inside and outside the τ_h candidate signal cone in order to increase their separation power.

The BDT is trained with the simulated samples listed in Section 3, containing the genuine τ_h and the electron. Reconstructed τ_h candidates are considered as signal (background) in the case they are matched to a τ_h decay (electron) at generator level.

Different working points are defined on the BDT output according to the efficiency for real τ_h to pass the discriminator. The expected effect on the efficiency of τ_h reconstruction and on the misidentification probability are presented in Fig. 5.
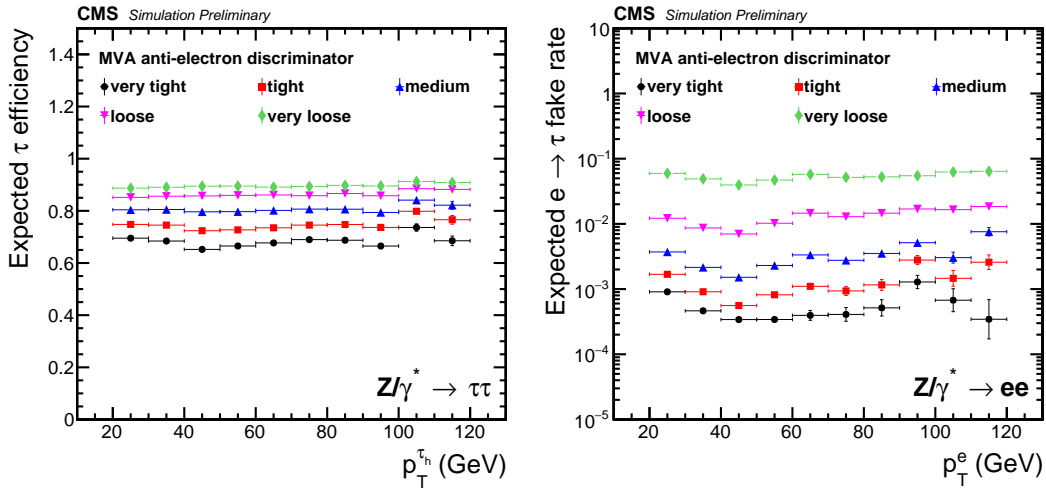

Figure 5: Efficiency of the τ_h identification estimated with simulated $Z/\gamma^* \rightarrow \tau\tau$ events (left) and the e $\rightarrow \tau_h$ misidentification probability estimated with simulated $Z/\gamma^* \rightarrow$ ee events (right) for the Very Loose, Loose, Medium, Tight and Very Tight working points of the MVA based anti-e discrimination algorithm. The efficiency is shown as a function of the τ_h transverse momentum while the misidentification probability is shown as a function of the e transverse momentum. Both efficiency and misidentification probability are calculated for τ_h candidates with a reconstructed decay mode and passing the Loose working point of the isolation sum discriminator.

6   Event samples for the validation with data

The τ_h reconstruction and identification performance is measured in data, with a strategy similar to that described in Ref. [30]. The τ_h identification efficiency, charge misidentification probability, and τ_h energy scale are validated using $Z/\gamma^* \rightarrow \tau_\mu\tau_h$ events. Highly virtual $W^* \rightarrow \tau_h\nu$ events are used to measure the τ_h identification efficiency in high p_T region. Samples of W+jets are used to validate the probability with which quark and gluon jets get misidentified

as hadronic tau decays. The misidentification probabilities for electrons are measured using $Z/\gamma^* \to ee$ events. This section describes the event selection used for these samples and will be referred to by later sections.

6.1 $Z/\gamma^* \to \tau\tau$ **sample**

A high purity sample of $Z/\gamma^* \to \tau_\mu\tau_h$ consists of events with at least one well identified and isolated muon, the *tag*, and one τ_h candidate that passes loose preselection criteria, the *probe*.

The events are required to pass a single muon trigger. Offline, the muon candidate is selected with $p_T > 19\,\text{GeV}$ and $|\eta| < 2.1$, and is required to pass the medium identification working point and have a relative isolation less than 0.1. The τ_h candidate is loosely preselected: it must have $p_T > 20\,\text{GeV}$, $|\eta| < 2.3$, no overlap with any global muon with $p_T > 5\,\text{GeV}$, passes anti-lepton discriminators, and possess at least one track with $p_T > 5\,\text{GeV}$. The two objects are required to be separated by at least $\Delta R = 0.5$ and to carry an opposite-sign electric charge. If several $\mu - \tau_h$ pairs pass this selection, the most isolated objects are selected. The events are rejected if they contain one or more b-tagged jets or an additional isolated lepton with $p_T > 10\,\text{GeV}$. In order to reduce the W+jets background contribution, the transverse mass between the muon and the transverse missing energy, $m_T \equiv \sqrt{2p_T^\mu E_T^{miss}(1 - \cos\Delta\phi)}$, is required to be less than $40\,\text{GeV}$, where $\Delta\phi$ is the difference in azimuthal angle between the muon transverse momentum and $\vec{p}_T^{\text{miss}}$. In addition, the variable P_ζ, orginally developed by the CDF experiment [50], is used in order to benefit from the fact that in $Z/\gamma^* \to \tau\tau$ events the missing energy from the τ decay neutrinos typically forms a small angle with the visible τ decay products. P_ζ is required to be greater than $-25\,\text{GeV}$.

6.2 **Highly virtual** $W^* \to \tau\nu$ **sample**

This selection targets events in which a highly virtual W boson is produced at a small transverse momentum (thus having no accompanying hard hadronic jet) and decays into $W^* \to \tau_h\nu$. The signature of such events consists of a well balanced τ_h and missing transverse momentum.

The events are required to pass a trigger, which only uses calorimeter information (without the muon systems) and require that $E_{T,no\mu}^{miss}$ and $|\vec{H}_{T,no\mu}|$ both are greater than $90\,\text{GeV}$. Offline, events are required to have exactly one τ_h with $p_T > 100\,\text{GeV}$ and $E_T^{miss} > 110\,\text{GeV}$. In order to ensure back-to-back topology between τ_h and E_T^{miss}, $0.7 < p_T(\tau_h)/E_T^{miss} < 1.3$ and $\Delta\phi(\tau_h, p_T^{miss}) > 2.4$ are required. The event is discarded if it has more than one jet with $p_T \geq 20\,\text{GeV}$ and $|\eta| < 4.7$ or loosely identified electrons or muons.

6.3 **W+jet sample**

For this sample, the event is triggered by a single isolated muon trigger with $p_T > 18\,\text{GeV}$. Offline, one well-identified and isolated muon with $p_T > 20\,\text{GeV}$ is required. Events with additional loosely isolated and identified electrons or muons are rejected. In addition, m_T is required to be greater than $60\,\text{GeV}$ to suppress events with real τ_h decays, in particular from Z boson decays. Events with at least one loosely b-tagged jet are removed to reject $t\bar{t}$ events.

6.4 $Z/\gamma^* \to$ **ee**

Finally, a high purity sample of $Z/\gamma^* \to e\tau_h$ events is selected by requiring events with at least one well identified and isolated electron, the *tag*, plus one τ_h candidate that passes loose preselection criteria, the *probe*.

The events are required to pass the single electron trigger. Offline, the electron candidate must have $p_T > 24$ GeV and $|\eta| < 2.1$, pass tight electron identification criteria and have a relative isolation less than 0.1.

The probe τ_h candidate is required to satisfy $p_T > 20$ GeV, $|\eta| < 2.3$, to be separated from the electron by $\Delta R > 0.5$ and to have opposite electric charge with respect to the tag electron.

The W+jets and $t\bar{t}$ backgrounds are reduced by requiring the selected events to pass a cut on the transverse mass of the tag electron plus missing transverse energy of $m_T < 30$ GeV.

7 Measurement of tau identification efficiency

The τ_h identification efficiency in Run-2 is measured using a similar approach as in Run-1 [30]. The goal of the measurement is to derive scaling factors for existing simulations and their uncertainties, which should be used to estimate correspondent systematic uncertainties in physics analyses. The standard candle to compare data and MC simulations is the Z or W production with their decays to τ's. For the low p_T region, two separate measurements are performed; one uses a tag–and–probe method in $Z \to \tau_\mu \tau_h$ events, whereas the second one compares selection efficiencies for $Z/\gamma^* \to \tau_\mu \tau_h$ and $Z/\gamma^* \to \mu\mu$ events. These two measurements can be considered as a cross-check. For the high p_T region, highly virtual $W^* \to \tau_h \nu$ events are used.

7.1 Using the tag–and–probe method in $Z \to \tau_\mu \tau_h$ events

The $Z/\gamma^* \to \tau\tau$ sample is used, as defined in Section 6.1. The events are divided into *pass* and *fail* categories, depending on whether their τ_h candidate respectively passes or fails the isolation discriminator under study. The number of signal events in each category is extracted from a simultaneous maximum likelihood fit of the expected SM contributions to the observed data in the two regions. Two different observables that provide a good discrimination between signal and background are used: m_{vis}, the invariant mass between the muon and τ_h candidates, and $N_{charged}$, the number of charged particles inside the signal and isolation cones of the τ_h candidate.

The signal corresponds to $Z \to \tau_\mu \tau_h$ events in the simulation, where the reconstructed τ_h candidate is matched to the generated one. All other Drell-Yan events are considered as a background in the analysis. The Drell-Yan process contribution, as well as the contributions from diboson, single-top and $t\bar{t}$ events, are estimated from simulation, and normalized to their expected theoretical cross sections. The shape of the W+jets background is taken from simulation, while its normalization is taken from a data sample enriched in W+jets events, defined in the same way as the signal region but without a P_ζ cut and with an inverted selection on the transverse mass ($m_T > 80$ GeV). The QCD multi-jet background is estimated from a control region where the selection is the same as in the signal region, except that the muon and the τ_h candidates are required to have the same electric charge. The QCD multi-jet contribution is estimated in this region by subtracting all other backgrounds, a multiplicative factor 1.06 is used when applying this estimate in the opposite-sign signal region.

The data/MC τ_h identification efficiency scale factor is defined as the parameter of interest in a simultaneous maximum likelihood fit to the passed and failed regions using the two distributions discussed above. The systematics considered in the maximum likelihood fit as nuisance parameters are similar to those used in Ref. [30]: the luminosity uncertainty in 2015 amounts to 2.7%, while 3% uncertainty is associated to the muon identification, isolation and triggering uncertainties in MC simulations. The $t\bar{t}$ cross section is known with 10% precision, and the small diboson and single top processes are attributed a conservative 30% uncertainty, whereas the

Table 1: Data/MC scale factors for the different working points of the isolation-sum and MVA-based discriminator, and for two values of the isolation cone. An uncertainty of 3.9% has been added in quadrature to the uncertainty returned by the fit to account for the tracking efficiency uncertainty. Tag–and–probe method is used to measure the efficiency and its uncertainty.

Isolation discriminant		m_{vis}	$N_{charged}$
	Loose	0.98 ± 0.06	0.98 ± 0.06
Isolation-sum $\Delta R = 0.5$	Medium	0.98 ± 0.06	0.96 ± 0.05
	Tight	0.97 ± 0.06	0.96 ± 0.05
	Loose	0.99 ± 0.06	0.97 ± 0.05
Isolation-sum $\Delta R = 0.3$	Medium	1.00 ± 0.06	0.99 ± 0.05
	Tight	0.99 ± 0.06	0.98 ± 0.05
	Loose	1.00 ± 0.06	1.02 ± 0.06
MVA-based $\Delta R = 0.5$	Medium	1.00 ± 0.06	1.01 ± 0.05
	Tight	1.01 ± 0.06	0.99 ± 0.05
	Very Tight	1.00 ± 0.06	1.00 ± 0.05
	Loose	1.01 ± 0.06	0.99 ± 0.06
MVA-based $\Delta R = 0.3$	Medium	1.01 ± 0.06	0.98 ± 0.06
	Tight	1.02 ± 0.06	0.98 ± 0.05
	Very Tight	1.00 ± 0.06	0.98 ± 0.05

uncertainties on the normalizations of the W+jets and QCD multi-jet background, both data-driven, are equal to 10% and 5% respectively. The τ_h energy scale uncertainty, amounting to 3%, is considered as a shape uncertainty. For the measurement with $N_{charged}$ as an observable, a shape uncertainty is considered on all processes estimated from MC, to account for a 3.9% uncertainty on the track reconstruction efficiency, and an uncertainty of 10% is attributed to the multiplicity of tracks associated to the τ_h candidates that are due to jet $\to \tau_h$ fakes. Bin-by-bin uncertainties, accounting for the limited number of MC events in every bin of the distributions, are considered for all processes.

The results obtained for the different isolation-sum and MVA-based discriminator working points, shown for the standard τ_h identification and using the two different observables are shown in Table 1. An uncertainty of 3.9% has been added in quadrature to the one returned by the fit to take into account the uncertainty on the tracking efficiency. The results with both observables are compatible with each other, and all scale factors are compatible with unity, with an uncertainty of about 6%. The postfit distributions that maximize the likelihood for the Loose working point of the MVA-based isolation are shown in Fig. 6.

7.2 Using the $Z \to \tau_\mu \tau_h / Z \to \mu\mu$ ratio method

The identification efficiency of τ_h can be also measured by comparing the yields of selected $Z/\gamma^* \to \tau_\mu \tau_h$ and $Z/\gamma^* \to \mu\mu$ events. Such a comparison leads to cancellation of some uncertainties (e.g. luminosity and cross section uncertainties).

Events in the $\tau_\mu \tau_h$ region are selected as described in Section 6.1, except that no selection criteria on the p_T of the τ_h track and the P_ζ variable are applied. The selection of events in the $\mu\mu$ region is chosen to be as close as possible to the $\tau_\mu \tau_h$ selection to maximize the correlation of the systematic uncertainties between the two regions. The same single muon trigger is used, and the leading muon is required to pass exactly the same selection criteria as the muon in the $\tau_\mu \tau_h$ region. The selection of the subleading muon is identical to the leading muon selection,
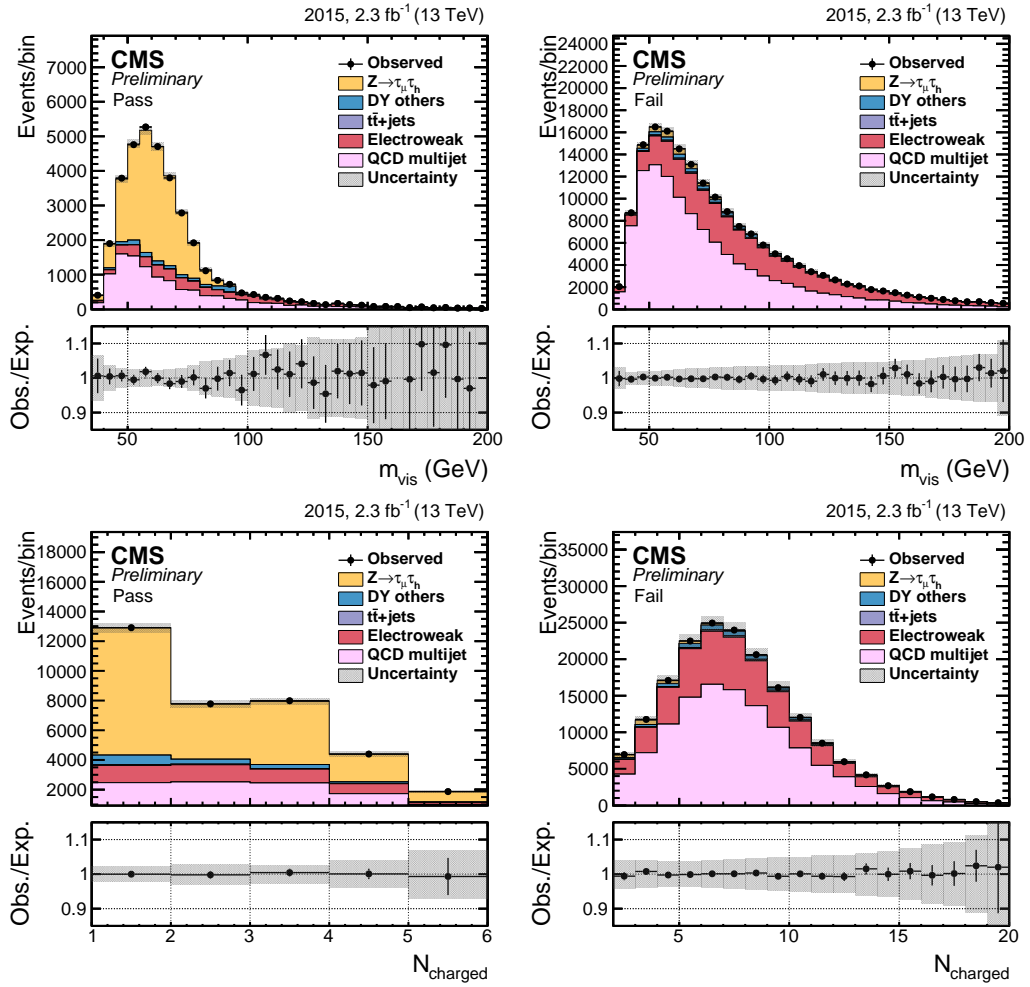
Figure 6: Postfit distributions in the *pass* (left) and *fail* (right) control regions, using m_{vis} (top) or $N_{charged}$ (bottom) as observable, for the Loose working point of the MVA-based isolation.

Table 2: Data/MC scale factors for the different working points of the isolation-sum and MVA-based discriminator, and for two values of the isolation cone. The ratio, $Z \to \tau_\mu \tau_h / Z \to \mu\mu$ is used as a discriminant variable.

Isolation discriminant		Data/MC scale factor
Isolation-sum $\Delta R = 0.5$	Loose	0.92 ± 0.03
	Medium	0.91 ± 0.03
	Tight	0.91 ± 0.03
Isolation-sum $\Delta R = 0.3$	Loose	0.91 ± 0.04
	Medium	0.91 ± 0.03
	Tight	0.92 ± 0.03
MVA-based $\Delta R = 0.5$	Loose	0.94 ± 0.03
	Medium	0.93 ± 0.03
	Tight	0.93 ± 0.03
	Very Tight	0.93 ± 0.03
MVA-based $\Delta R = 0.3$	Loose	0.94 ± 0.03
	Medium	0.94 ± 0.03
	Tight	0.95 ± 0.03
	Very Tight	0.92 ± 0.03

except that $|\eta| < 2.4$ is required. The background estimations are performed in a similar way as described in Section 7.1.

The data/MC τ_h identification efficiency scale factor is considered as the parameter of interest in a simultaneous maximum likelihood fit of the two regions. Similar systematic uncertainties are considered as are described in Section 7.1, with the addition of 5% uncertainty for the muon trigger, identification and isolation efficiencies in the $\mu\mu$ channel. The distributions in both regions after the fit are shown in Fig. 7 for the Loose working point of the isolation-sum discriminator. The obtained data/MC scale factors are summarized in Table 2.

The scale factors measured with this method are largely consistent within uncertainties with those measured using the tag-and-probe method. Although both methods make use of $Z/\gamma^* \to \tau_\mu \tau_h$ events, independent systematic uncertainties are relevant in each case and therefore differing values of the scale factors can be expected.

7.3 Using W* → τν events

The identification efficiency of τ_h's with high transverse momentum ($p_T \geq 100\,\text{GeV}$) is estimated using a data sample with highly virtual W bosons with $m_W \geq 200\,\text{GeV}$. The efficiency is extracted via a maximum likelihood fit to the m_T distribution.

The highly virtual W* → τν sample described in Section 6.2 is used. A control sample of W* → $\mu\nu$ events is defined by selecting events which pass a single-muon trigger and are required to pass a similar event selection as for the signal: events must have an identified and isolated muon with $p_T > 120\,\text{GeV}$, $E_T^{miss} > 100\,\text{GeV}$, $0.7 < p_T(\mu)/E_T^{miss} < 1.3$, and $\Delta\phi(\mu, p_T^{mis}) > 2.4$.

The events selected in the $\tau_h\nu$ final state largely originate from background processes such as QCD multijet, $Z \to \nu\bar{\nu}$+jets, and $W \to \ell\nu$+ jets ($\ell =$ e, μ) events, where a hadronic jet is misidentified as τ_h. These backgrounds are modeled by using a control sample obtained by applying the same selection as used for the signal sample, W* → τν, except that the τ_h candidate is required to be anti-isolated. The events in the control region are then weighted using an ex-
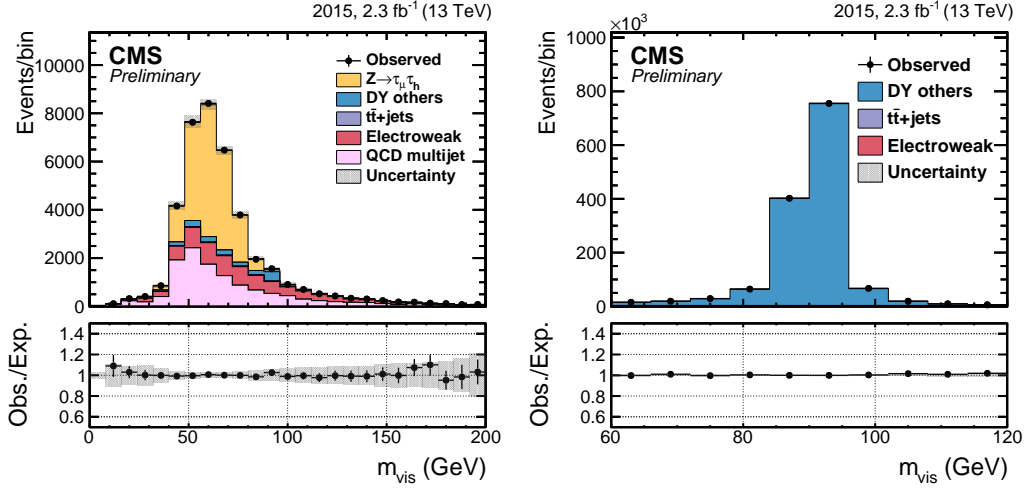
Figure 7: Postfit distributions in the $\mu\tau_h$ (left) and $\mu\mu$ (right) regions, for the Loose working point of the isolation-sum discriminator as derived using the $Z \rightarrow \tau\tau/Z \rightarrow \mu\mu$ ratio method.

trapolation factor $F^{iso} = P(fake\text{-}\tau|iso)/P(fake\text{-}\tau|anti\text{-}iso)$, where $P(fake\text{-}\tau|iso)$ is the probability of a jet to be misidentified as a τ_h and to pass isolation requirements and $P(fake\text{-}\tau|anti\text{-}iso)$ is the probability of a jet to be misidentified as a τ_h but to fail isolation requirements. This extrapolation factor, F^{iso}, is derived from two (quark and gluon) jet dominated control regions in data, and is parameterized as a function of the p_T of the τ_h candidate. Applying the F^{iso} factor to the anti-isolated control region effectively models the shape and estimates the number of events with a jet misidentified as τ_h in the signal region.

The signal is extracted using a maximum likelihood simultaneous fit to the m_T distribution for both the $W^* \rightarrow \tau_h\nu$ and $W^* \rightarrow \mu\nu$ signal and control regions. Such procedure allows to avoid uncertainties related to the yield of the W events in simulation, since it is now defined by the $W^* \rightarrow \mu\nu$ sample. Similar systematic uncertainties are used as are described in section 7.1. The obtained W normalization $\sigma(pp \rightarrow W^* + X|m_{W^*} > 200\,\text{GeV})$, the τ_h identification efficiency scale factor, and the correlation coefficient between the two quantities are shown in Table 3. Fig. 8 shows post-fit m_T distributions in the $W^* \rightarrow \mu\nu$ and the signal region. For the τ_h with a high p_T it is seen that the data/MC scale factor is slightly less than one but still consistent with unity.

Table 3: τ_h identification efficiency scale factor, the nomalization of $\sigma(pp \rightarrow W^* + X|m_{W^*} > 200\,\text{GeV})$, r, and correlation coefficient between the two quantities obtained from the fit. The scale factors are measured for both isolation-sum and MVA-based discriminators.

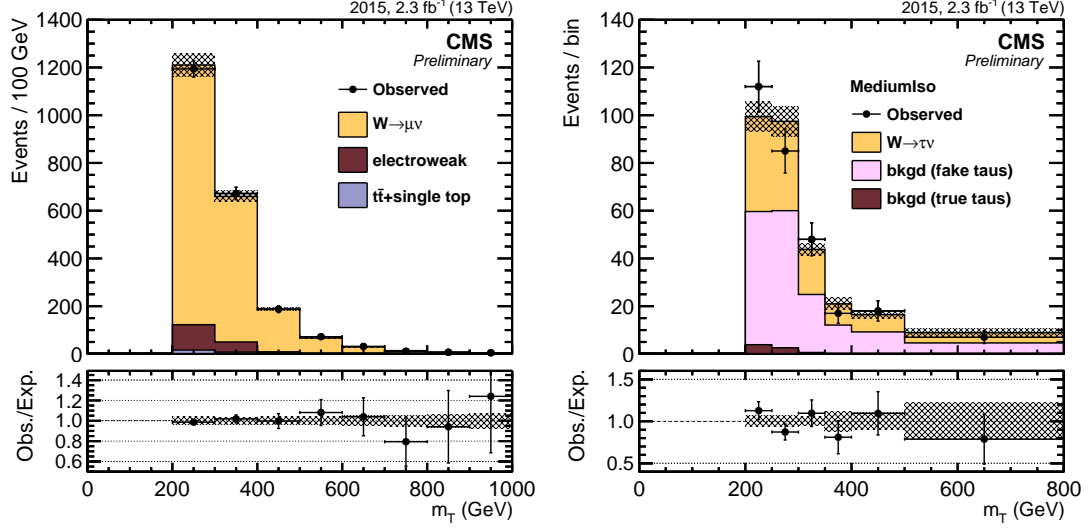| Isolation discriminant		τ id SF	r (W norm.)	correlation
Isolation-sum $\Delta R = 0.5$	Loose	0.94 ± 0.21	1.01 ± 0.05	-0.18
	Medium	0.91 ± 0.19	1.01 ± 0.05	-0.19
	Tight	0.83 ± 0.20	1.01 ± 0.05	-0.17
MVA-based $\Delta R = 0.5$	Loose	0.96 ± 0.17	1.01 ± 0.05	-0.22
	Medium	0.95 ± 0.15	1.01 ± 0.05	-0.25
	Tight	0.94 ± 0.15	1.01 ± 0.05	-0.27
	Very Tight	0.94 ± 0.14	1.01 ± 0.05	-0.27

Figure 8: The transverse mass distribution in the selected sample of $W^* \to \mu\nu$ events after applying maximum likelihood fit (left). The measured transverse mass distribution in the sample of selected $W^* \to \tau_h\nu$ events with Medium working point of isolation-sum discriminator.

8 Measurement of tau energy scale

The energy scale is defined as a ratio of the reconstructed τ_h energy to the true energy of the visible tau decay products at the generator level as defined in the simulation. To check the simulation of the energy scale a sample of $Z \to \tau_\mu\tau_h$ events in data is used. The simulated distributions that are sensitive to the energy scale variations are the reconstructed τ_h mass, m_{τ_h}, and the mass of muon and τ_h system, m_{vis}. These distributions are fit to the data separately for the decay modes $h^\pm$, $h^\pm\pi^0 s$ ($h^\pm\pi^0$ or $h^\pm\pi^0\pi^0$), and $h^\pm h^\mp h^\pm$ to obtain the correction factors between simulation and observed data.

The $Z \to \tau_\mu\tau_h$ events are selected as described in Section 6.1 except that the τ_h candidates are required to pass the Tight working point of the isolation-sum τ_h isolation discriminant in order to further reduce fake τ_h backgrounds. The shape templates for the $Z \to \tau_\mu\tau_h$ signal sample are produced by varying the reconstructed τ_h energy discretely between -6% and $+6\%$ in steps of 0.1%. The m_{vis} and m_{τ_h} templates are recomputed for each step. While the m_{τ_h} shows better sensitivity for the energy scale in $h^\pm\pi^0 s$ and $h^\pm h^\mp h^\pm$ channels, it can not be used in the $h^\pm$ decay, where only the m_{vis} is used. The backgrounds are modeled in the same way as described in Section 7.1, and the templates for background processes are kept unchanged as a function of the τ_h energy scale.

For illustration, the m_{vis} fit templates corresponding to τ_h energy scale shifts of 0, -6, and $+6\%$ are shown in Fig. 9 for an inclusive selection over all decay modes. The data are compared to predictions of MC with different energy scales: $Z/\gamma^\star \to \tau\tau$ and expected background distributions.

A likelihood ratio method is used to extract the τ_h energy scale for each decay mode. Similar systematic uncertainties are used as described in Section 7.1. The obtained results are reported in Table 4.
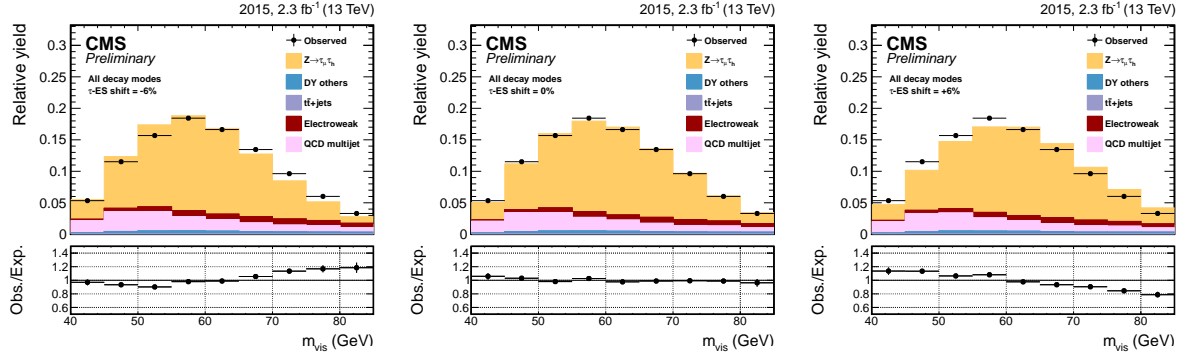
Figure 9: The distributions of m_{vis} of the muon–τ_h system with all τ_h decay modes included. The observed data are compared to predictions with different shift applied to the energy scale: -6% (left), 0% (center) and $+6\%$ (right).

Table 4: Energy scale corrections for τ_h measured in $Z \to \tau\tau$ events for τ_h reconstructed in different decay modes. The inclusive result is obtained by means of an independent fit and hence may be different from the average of τ_h energy scale corrections measured for individual decay modes.

Decay mode	Fit on m_{vis}	Fit on m_{τ_h}
All decay modes	$-1.0^{+1.1}_{-0.6}$	-
$h^{\pm}$	$+1.5^{+1.5}_{-1.8}$	-
$h^{\pm}\pi^0 s$	$-1.5^{+0.8}_{-1.9}$	$-0.5^{+0.7}_{-0.9}$
$h^{\pm}h^{\mp}h^{\pm}$	$-1.0^{+2.3}_{-1.7}$	$+0.0^{+0.2}_{-0.4}$

9   Measurement of tau charge misidentification

Proper identification of the charge of the τ_h is essential as the technique of using opposite-sign (OS) events to define signal region and same-sign (SS) to estimate the background is used in many analyses with τ_h final state.

The $Z/\gamma^* \to \tau_\mu \tau_h$ sample is selected almost identically as defined in Section 6.1, except that the mass of the muon and the τ_h system is required to be less than $100\,\text{GeV}$, and that the τ_h candidate passes the Tight working point of the isolation-sum discriminator. The τ_h charge misidentification probability can be measured with a tag-and-probe method by dividing events into an OS and a SS regions depending on their electric charges. The simultaneous maximum likelihood fit is performed with charge misidentification probability as a parameter of interest, which is anti-correlated between the OS and SS regions.

Except for the QCD multi-jet background, all SM processes are estimated as described in Section 7.1. The QCD multi-jet background cannot be estimated in the SS region with isolated muons and τ_h's, because this region is now used to extract the parameter of interest. Instead, the shape of the QCD multi-jet background is estimated from a SS region where the τ_h candidate satisfies the isolation condition, and where the muon relative isolation is relaxed from 0.1 to 0.5. The limited contribution from other processes is estimated with simulations and subtracted from the observed data. The number of events in this region is about the same as in the signal region. To account for a potential bias caused by the relaxation of the muon isolation, an additional uncertainty is added in the bins with $30 < m_{vis} < 50\,\text{GeV}$ to reflect the disagreement

observed between the shapes in a region with a relaxed muon and anti-isolated τ_h, and isolated muon and anti-isolated τ_h. The normalization of the QCD multi-jet background is determined using the SS region described in Section 7.1, and requiring in addition the visible mass between the muon and τ_h candidates to be greater than 100 GeV.

The systematic uncertainties considered are the same as in Section 7.1, except that an uncertainty of 6% is included to account for the τ_h identification efficiency uncertainty, and that a 10% uncertainty is associated to the estimation of the normalization of the QCD multi-jet background from the high-m_{vis} region.

The expected charge misidentification probability is measured to be $0.22 \pm 0.05\%$, which is compatible with MC expectations. The data/MC scale factor is extracted from a simultaneous maximum likelihood fit and found to be equal to 4^{+5}_{-4}. This extracted value can be considered as an upper limit on the τ_h charge misidentification probability in data, equal to at 68% C.L. For CMS analysis it is recommended simply to use MC for τ_h charge misidentification estimation. The post-fit distributions of m_{vis} in the OS and SS regions are shown in Fig. 10.
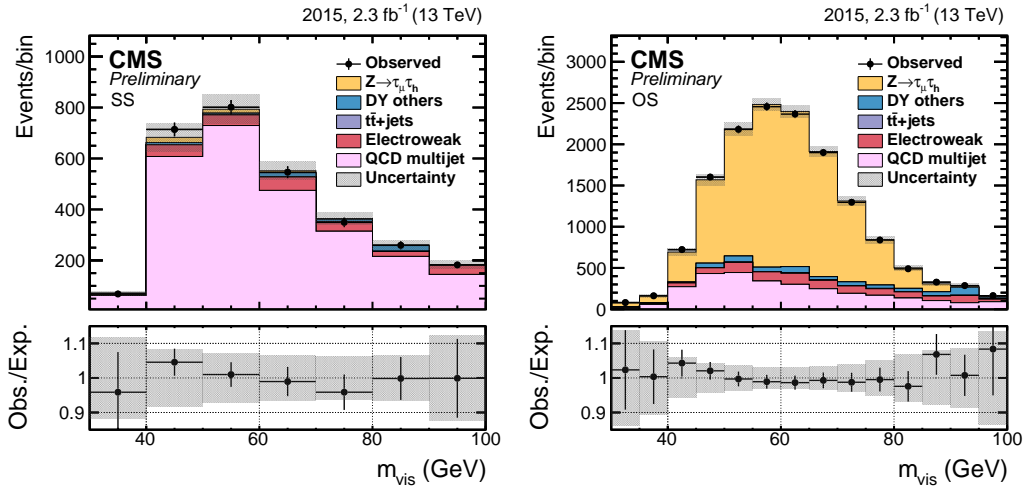

Figure 10: Postfit distributions in the SS (left) and OS (right) regions for the charge misidentification probability measurement.

10 Measurement of jet $\rightarrow \tau_h$ misidentification probability

The probability to misidentify quark and gluon jets as τ_h candidates is measured as a function of jet p_T and η using sample of $W \rightarrow \mu\nu$+jets events selected as described in Section 6.3. In addition to p_T^{jet} and η^{jet}, the misidentification probability also strongly depends on the parton flavor of the considered jet as well as on whether the parton initiating the jet and the reconstructed τ_h have the same or opposite charge. These factors cause a differences of up to a factor of 4 for the misidentification probabilities for c quark and gluon jets, and up to a factor of 2 for jets where the initiating parton has the same or opposite charge with respect to the τ_h candidate. This implies that the numbers for the misidentification probabilities given in this section are representative numbers that apply only to the considered $W \rightarrow \mu\nu$+jets event sample, which has a large fraction of light quark jets and, therefore, a comparably high misidentification probability.

The misidentification probability is computed per jet as follows:

$$P_{misid} = \frac{\text{denominator} \,\&\, p_T^{\tau_h} > 20\,\text{GeV} \,\&\, |\eta^{\tau_h}| < 2.3 \,\&\, \text{decay mode} \,\&\, \tau_h \text{ ID discriminant}}{p_T^{\text{jet}} > 20\,\text{GeV} \,\&\, |\eta^{\text{jet}}| < 2.3} \qquad (6)$$

Note that p_T^{jet} and η^{jet} usually differ from the p_T and η of the τ_h because p_T^{jet} and η^{jet} are computed by summing the momenta of all particle constituents of the jet, while p_T and η of the τ_h are computed only from the charged hadrons and photons included in the decay mode reconstruction of the τ_h candidate. Furthermore, p_T^{jet} is subject to additional jet energy corrections, whereas p_T of the τ_h is not.

In the measurement of the misidentification probability, backgrounds with real τ_h are subtracted based on the expectation from simulated events. The fraction of events with real τ_h in the sample passing the τ_h identification criteria is well below 10% for τ_h with $p_T < 100\,\text{GeV}$, but reaches up to 50% for τ_h with $p_T \sim 300\,\text{GeV}$. Furthermore, backgrounds with prompt muons and electrons giving rise to τ_h candidates are also subtracted based on the expectation from simulated events. To reject events from $Z \rightarrow \mu\mu$ production, a loose anti-muon discriminator is applied to the reconstructed τ_h candidates. This has negligible impact on the inclusive numbers of jets in the denominator of the definition of the misidentification probability, but it largely reduces the number of misidentified τ_h candidates coming from muons in the numerator and thereby reduces the possible impact of the uncertainty in the $\mu \rightarrow \tau_h$ misidentification probability.

The subtraction of backgrounds with real τ_h is subject to an uncertainty of 30%, leading to an uncertainty in the determined jet $\rightarrow \tau_h$ misidentification probability of up to 15%. The jet energy scale also leads to a significant uncertainty, in particular in the lowest considered jet p_T bin due to threshold effects. Additional uncertainties are considered for the probabilities with which electrons are reconstructed as τ_h candidates (100% relative uncertainty) and with which muons are reconstructed as τ_h candidates and pass the loose working point of the anti-muon discriminator (50%). These lead to unceratinties in the measured misidentification probabilities of at most a few percent.

The observed and simulated jet $\rightarrow \tau_h$ misidentification probabilities for the loose working points of the cut-based and MVA based isolation discriminant are shown in Fig.11 as a function of p_T^{jet} and η^{jet}. The observed probabilities are larger than the simulated ones, with the difference exceeding the size of the systematic uncertainties. This difference was not observed in Run-1, whose data and MC difference was up to 20% level [30]. The source of disagreement was traced down to the different spectrum of the charged isolation-sum between data and MC. A significant fraction of this difference can be explained by the use of an updated hadronisation model and tune (CUET) in PYTHIA8 [36] with respect to PYTHIA6 and tune Z2 used in run 1 [51], as studied in simulated events that only differ in the use of the two different generators and tunes.

11   Measurement of e $\rightarrow$ τ_h misidentification probability

The e $\rightarrow \tau_h$ misidentification probability is measured in data via a tag-and-probe method in a $Z/\gamma^* \rightarrow$ ee sample selected as described in Section 6.4.

Depending on whether the probe passes or fails the anti-e discriminator under study, the event enters the *pass* or the *fail* category, respectively. The e $\rightarrow \tau_h$ misidentification probability is then measured using a simultaneous fit to the m_{vis} distribution in the *pass* and the *fail* regions.The
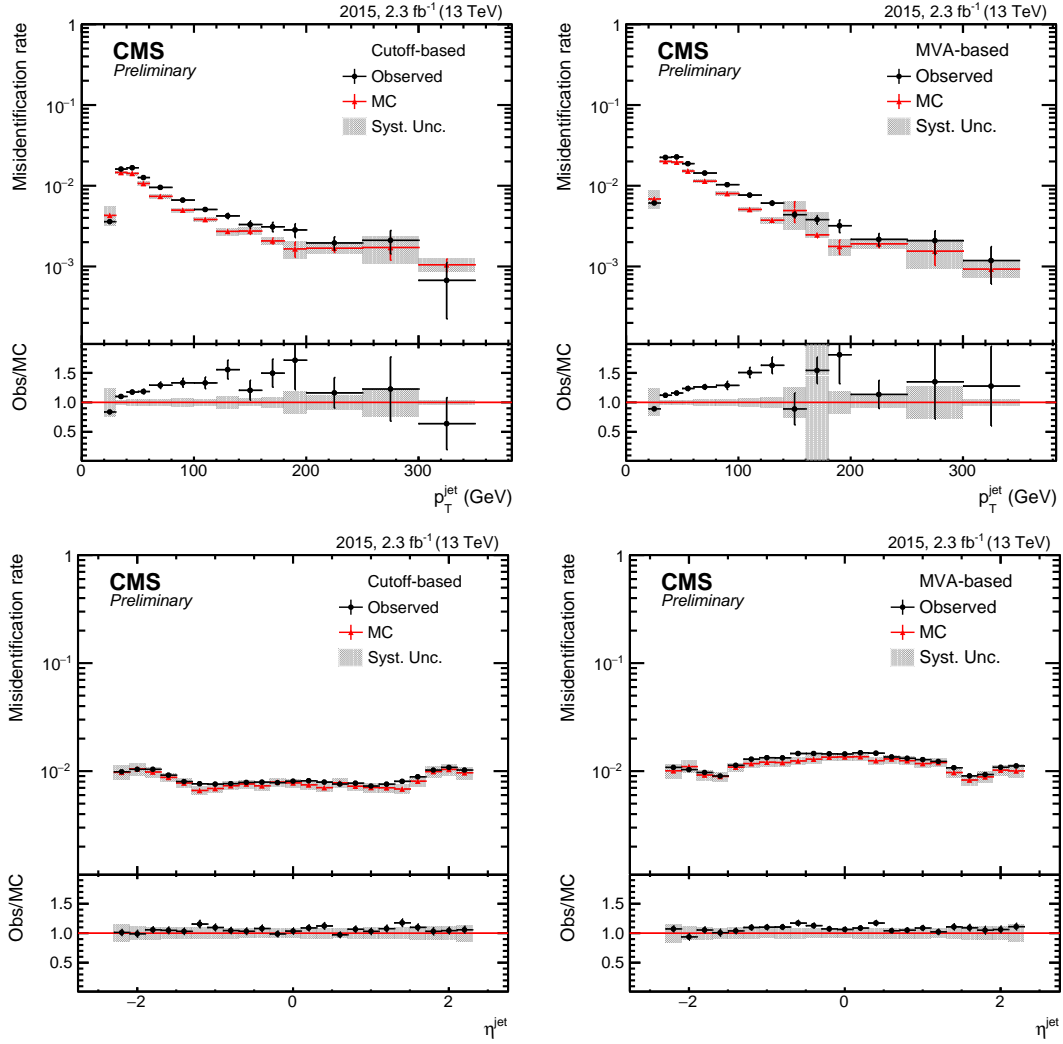
Figure 11: Probabilities for quark and gluon jets in W+jets events to pass the loose working point of either the isolation-sum discriminant (left) or the MVA-based isolation discriminant (right) as a function of p_T^{jet} (top) and η^{jet} (bottom). The misidentification probabilities measured in observed events are compared to the expectation from simulation. The uncertainties in the simulated and observed efficiencies include the statistical uncertainties from the limited numbers of both observed and simulated events, including the background subtraction. The shaded bands include the systematic uncertainties related to the background subtraction and the jet energy scale.

m_{vis} distribution is fitted in the window $60\,\text{GeV} < m_{vis} < 120\,\text{GeV}$ with separate MC shape templates for the $Z/\gamma^* \to ee$ signal and the $Z/\gamma^* \to \tau\tau$, W+jets, $t\bar{t}$, single top-quark, di-boson (WW, WZ, ZZ) and QCD multi-jet backgrounds. The shape templates are derived using the method described in Section 7.1 and simply alternating muon to electron. Similar systematic uncertainties are used as described in Section 7.1. Specfic shape uncertainties to this analysis are the energy scale of the tag electrons (2%) and the energy scale of the probes matched to electrons (5%). The uncertainty on the electron resolution is taken into account as an additional shape systematic, varying the difference between the generated and the reconstructed visible mass of the tag-probe pair upwards and downwards by 20%. Separate fits are used for probes in the barrel ($|\eta| < 1.46$) and in the endcap ($|\eta| > 1.56$).

The post-fit m_{vis} distributions in the *pass* category are shown in Fig. 12 for the Loose, Medium, Tight and Very Tight working points of the anti-e discriminator, in the barrel region. The obtained e $\rightarrow \tau_{h}$ misidentification probabilities are given in Table 5. In the barrel region, the measured misidentification probabilities in data exceed the expected probabilities from simulation. The difference between data and simulation increases for the Tight and Very Tight working points of the discriminator. A similar trend is observed for the probes in the endcap region.

Table 5: Probability for electrons to pass the different working points of the MVA-based anti-e discriminator, splitted in barrel and endcap region. For each working point, the e $\rightarrow \tau_{h}$ misidentification probability is defined as the fraction of probes passing the given discriminator with respect to the total number of probes.

ECAL barrel ($\|\eta\| < 1.46$)			
Working point	Expected	Measured	Scale Factor
Very Loose	$(5.91 \pm 0.62) \cdot 10^{-2}$	$(6.05 \pm 0.63) \cdot 10^{-2}$	$1.02^{+0.05}_{-0.01}$
Loose	$(1.12 \pm 0.12) \cdot 10^{-2}$	$(1.27 \pm 0.14) \cdot 10^{-2}$	$1.14^{+0.03}_{-0.04}$
Medium	$(2.75 \pm 0.31) \cdot 10^{-3}$	$(4.14 \pm 0.46) \cdot 10^{-3}$	$1.50^{+0.12}_{-0.13}$
Tight	$(1.17 \pm 0.13) \cdot 10^{-3}$	$(2.10 \pm 0.24) \cdot 10^{-3}$	1.80 ± 0.23
Very Tight	$(0.67 \pm 0.08) \cdot 10^{-3}$	$(1.27 \pm 0.14) \cdot 10^{-3}$	$1.89^{+0.35}_{-0.32}$
ECAL endcap ($\|\eta\| > 1.56$)			
Working point	Expected	Measured	Scale Factor
Very Loose	$(7.04 \pm 0.74) \cdot 10^{-2}$	$(7.81 \pm 0.81) \cdot 10^{-2}$	$1.11^{+0.05}_{-0.01}$
Loose	$(1.10 \pm 0.12) \cdot 10^{-2}$	$(1.19 \pm 0.13) \cdot 10^{-2}$	1.09 ± 0.05
Medium	$(2.26 \pm 0.26) \cdot 10^{-3}$	$(2.40 \pm 0.27) \cdot 10^{-3}$	$1.06^{+0.17}_{-0.18}$
Tight	$(0.95 \pm 0.11) \cdot 10^{-3}$	$(1.24 \pm 0.14) \cdot 10^{-3}$	$1.30^{+0.32}_{-0.42}$
Very Tight	$(0.54 \pm 0.06) \cdot 10^{-3}$	$(0.91 \pm 0.10) \cdot 10^{-3}$	$1.69^{+0.66}_{-0.68}$

12 Summary

The algorithm used in Run-2 to reconstruct and identify hadronically decaying taus has been described in this note, with a particular emphasis on the changes with respect to Run-1. These changes include among others a dynamical strip reconstruction, and additional variables in the MVA-disriminators against jets and electrons.

The performance has been measured in data collected in 2015 at a center-of-mass energy of $\sqrt{s} = 13$ TeV. The tau identification and reconstruction techniques described are now fully commissioned and ready for use in CMS physics analyses for the remainder of Run-2. The performance in data of the τ_{h} identification efficiency in both low and high p_T regions is similar to that in Monte Carlo simulation, while the performance of the jet $\rightarrow \tau_{h}$ misidentification is found to be moderately different. The energy scale of τ_{h} is measured and its response with respect to the Monte Carlo simulation is found to be close to 1. The reduction in electron $\rightarrow \tau_{h}$ fake probability is seen to perform well in Run-2, and its scale factors have been measured.
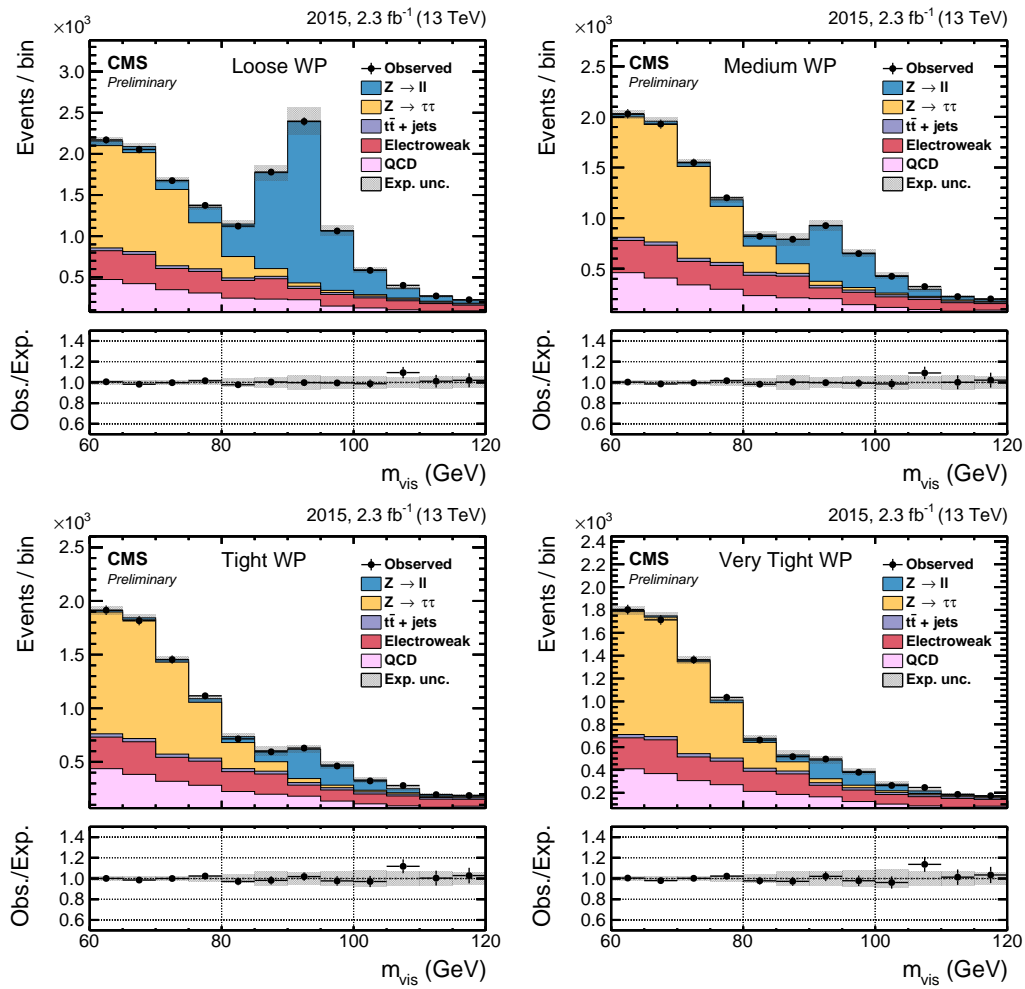
Figure 12: Post-fit plots of the tag and probe mass in the pass category for the Loose (top left plot), Medium (top right), Tight (bottom left) and Very Tight (bottom right) working point of the anti-e discriminator in the barrel region.

References

[1] CMS Collaboration, "Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons", *JHEP* **1405** (2014) 104, `doi:10.1007/JHEP05(2014)104`, `arXiv:1401.5041`.

[2] CMS Collaboration, "Evidence for the direct decay of the 125 GeV Higgs boson to fermions", *Nature Phys.* **10** (2014) 557, `doi:10.1038/nphys3005`, `arXiv:1401.6527`.

[3] CMS Collaboration, "Search for neutral MSSM Higgs bosons decaying to a pair of tau leptons in pp collisions", *JHEP* **10** (2014) 160, `doi:10.1007/JHEP10(2014)160`, `arXiv:1408.3316`.

[4] CMS Collaboration, "Search for additional neutral Higgs bosons decaying to a pair of tau leptons in pp collisions at $\sqrt{s} = 7$ and 8 TeV", Technical Report CMS-PAS-HIG-14-029, CERN, Geneva, 2015.

[5] CMS Collaboration, "Search for a charged Higgs boson in pp collisions at $\sqrt{s} = 8$ TeV", *JHEP* **11** (2015) 018, `doi:10.1007/JHEP11(2015)018`, `arXiv:1508.07774`.

[6] CMS Collaboration, "Updated search for a light charged Higgs boson in top quark decays in pp collisions at $\sqrt{s} = 7$ TeV", *CMS PAS* **HIG-12-052** (2012).

[7] CMS Collaboration, "Search for charged Higgs bosons with the $H^+ \to \tau^+ \nu_\tau$ decay channel in the fully hadronic final state at $\sqrt{s} = 8$ TeV", CMS Physics Analysis Summary CMS-PAS-HIG-14-020, 2014.

[8] CMS Collaboration, "Measurement of the Inclusive Z Cross Section via Decays to Tau Pairs in pp Collisions at $\sqrt{s} = 7$ TeV", *JHEP* **1108** (2011) 117, `doi:10.1007/JHEP08(2011)117`, `arXiv:1104.1617`.

[9] CMS Collaboration, "Measurement of the $W \to \tau\nu$ cross-section in pp Collisions at $\sqrt{s} = 7$ TeV", CMS Physics Analysis Summary CMS-PAS-EWK-11-019, 2011.

[10] CMS Collaboration, "Measurement of the $t\bar{t}$ production cross section in pp collisions at $\sqrt{s} = 8$ TeV in dilepton final states containing one τ lepton", *Phys.Lett.* **B739** (2014) 23, `doi:10.1016/j.physletb.2014.10.032`, `arXiv:1407.6643`.

[11] CMS Collaboration, "Search for Supersymmetry with a single tau, jets and MET", *CMS PAS* **SUS-12-008** (2012).

[12] CMS Collaboration, "Search for charginos and neutralinos produced in vector boson fusion processes in pp collisions at $\sqrt{s} = 8$ TeV", *CMS PAS* **SUS-12-025** (2012).

[13] CMS Collaboration, "Search for RPV supersymmetry with three or more leptons and b–tags", *CMS PAS* **SUS-12-027** (2012).

[14] CMS Collaboration, "A search for anomalous production of events with three or more leptons using 19.5 fb^1 of $\sqrt{s} = 8$ TeV LHC data", *CMS PAS* **SUS-13-002** (2013).

[15] CMS Collaboration, "Search for pair production of third generation leptoquarks and stops that decay to a tau and a b quark", *CMS PAS* **EXO-12-002** (2012).

[16] CMS Collaboration, "Search for high mass resonances decaying into τ^- lepton pairs in pp collisions at $\sqrt{s} = 7$ TeV", *Phys.Lett.* **B716** (2012) 82, `doi:10.1016/j.physletb.2012.07.062`, `arXiv:1206.1725`.

[17] CMS Collaboration, "Search for Third-Generation Scalar Leptoquarks in the tτ Channel in Proton-Proton Collisions at $\sqrt{s}$ = 8 TeV", *JHEP* **07** (2015) 042, `doi:10.1007/JHEP07(2015)042`, `arXiv:1503.09049`.

[18] CMS Collaboration, "Search for pair production of third-generation scalar leptoquarks and top squarks in proton-proton collisions at $\sqrt{s}$=8 TeV", *Phys. Lett.* **B739** (2014) 229, `doi:10.1016/j.physletb.2014.10.063`, `arXiv:1408.0806`.

[19] CMS Collaboration, "Search for W′ decaying to tau lepton and neutrino in proton-proton collisions at $\sqrt{s}$ = 8 TeV", `arXiv:1508.04308`.

[20] CMS Collaboration, "Search for the exotic decay of the Higgs boson to two light pseudoscalar bosons with two taus and two muons in the final state at $\sqrt{s} = 8$ TeV", Technical Report CMS-PAS-HIG-15-011, CERN, Geneva, 2016.

[21] CMS Collaboration, "Searches for a heavy scalar boson H decaying to a pair of 125 GeV Higgs bosons hh or for a heavy pseudoscalar boson A decaying to Zh, in the final states with h $\rightarrow \tau\tau$", `arXiv:1510.01181`.

[22] CMS Collaboration, "Search for a low-mass pseudoscalar Higgs boson produced in association with a b$\bar{\text{b}}$ pair in pp collisions at $\sqrt{s} = 8$ TeV", Technical Report CERN-PH-EP-2015-284. CMS-HIG-14-033-003. arXiv:1511.03610, CERN, Geneva, Nov, 2015. Comments: Submitted to Phys. Lett. B.

[23] CMS Collaboration, "A search for a doubly-charged Higgs boson in *pp* collisions at $\sqrt{s} = 7$ TeV", *Eur.Phys.J.* **C72** (2012) 2189, `doi:10.1140/epjc/s10052-012-2189-5`, `arXiv:1207.2666`.

[24] CMS Collaboration, "Search for lepton-flavour-violating decays of the Higgs boson to etau and emu at sqrt(s)=8 TeV", Technical Report CMS-PAS-HIG-14-040, CERN, Geneva, 2015.

[25] CMS Collaboration, "Search for Lepton-Flavour-Violating Decays of the Higgs Boson", *Phys. Lett.* **B749** (2015) 337, `doi:10.1016/j.physletb.2015.07.053`, `arXiv:1502.07400`.

[26] Particle Data Group, "Review of Particle Physics", *Chin.Phys.* **C38** (2014) 090001, `doi:10.1088/1674-1137/38/9/090001`.

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

[28] 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`.

[29] CMS Collaboration, "Performance of tau-lepton reconstruction and identification in CMS", *JINST* **7** (2012) P01001, `doi:10.1088/1748-0221/7/01/P01001`, `arXiv:1109.6034`.

[30] CMS Collaboration, "Reconstruction and identification of τ lepton decays to hadrons and ν_τ at CMS", *JINST* **11** (2016), no. 01, P01019, `doi:10.1088/1748-0221/11/01/P01019`, `arXiv:1510.07488`.

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

[32] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", *JINST* **9** (2014), no. 10, P10009, `doi:10.1088/1748-0221/9/10/P10009`, `arXiv:1405.6569`.

[33] CMS Collaboration, "CMS Tracking Performance Results from early LHC Operation", *Eur. Phys. J. C* **70** (2010) 1165, `doi:10.1140/epjc/s10052-010-1491-3`, `arXiv:1007.1988`.

[34] F. Maltoni and T. Stelzer, "MadEvent: Automatic event generation with MadGraph", *JHEP* **0302** (2003) 027, `arXiv:0208.0156`.

[35] S. Frixione, P. Nason, and C. Oleari, "Matching NLO QCD computations with Parton Shower simulations: the POWHEG method", *JHEP* **0711** (2007) 070, `arXiv:0709.2092`.

[36] T. Sjöstrand, 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`.

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

[38] K. Melnikov and F. Petriello, "Electroweak gauge boson production at hadron colliders through O(alpha(s)**2)", *Phys.Rev.* **D74** (2006) 114017, `doi:10.1103/PhysRevD.74.114017`, `arXiv:hep-ph/0609070`.

[39] J. M. Campbell, R. K. Ellis, and C. Williams, "Vector boson pair production at the LHC", *JHEP* **1107** (2011) 018, `doi:10.1007/JHEP07(2011)018`, `arXiv:1105.0020`.

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

[41] GEANT4 Collaboration, "GEANT4: A Simulation toolkit", *Nucl.Instrum.Meth.* **A506** (2003) 250, `doi:10.1016/S0168-9002(03)01368-8`.

[42] CMS Collaboration, "Particle–Flow Event Reconstruction in CMS and Performance for Jets, Taus, and $E_{\mathrm{T}}^{\mathrm{miss}}$", CMS Physics Analysis Summary CMS-PAS-PFT-09-001, 2009.

[43] CMS Collaboration, "Commissioning of the Particle-flow Event Reconstruction with the first LHC collisions recorded in the CMS detector", CMS Physics Analysis Summary CMS-PAS-PFT-10-001, 2010.

[44] CMS Collaboration, "Commissioning of the Particle-Flow reconstruction in Minimum-Bias and Jet Events from pp Collisions at 7 TeV", CMS Physics Analysis Summary CMS-PAS-PFT-10-002, 2010.

[45] CMS Collaboration, "Particle-flow commissioning with muons and electrons from J/Psi and W events at 7 TeV", CMS Physics Analysis Summary CMS-PAS-PFT-10-003, 2010.

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

[47] 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`.

[48] CMS Collaboration, "Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS", *JINST* **6** (2011) P11002, `doi:10.1088/1748-0221/6/11/P11002`, `arXiv:1107.4277`.

[49] CMS Collaboration, "Identification of b-quark jets with the CMS experiment", *JINST* **8** (2013) P04013, `doi:10.1088/1748-0221/8/04/P04013`, `arXiv:1211.4462`.

[50] CDF Collaboration, "Search for neutral MSSM Higgs bosons decaying to tau pairs in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV", *Phys.Rev.Lett.* **96** (2006) 011802, `doi:10.1103/PhysRevLett.96.011802`, `arXiv:hep-ex/0508051`.

[51] CMS Collaboration, "Measurement of the Underlying Event Activity at the LHC with $\sqrt{s} = 7$ TeV and Comparison with $\sqrt{s} = 0.9$ TeV", *JHEP* **1109** (2011) 109, `doi:10.1007/JHEP09(2011)109`, `arXiv:1107.0330`.