Tau reconstruction at CMS with a focus on high p_T taus

Soham Bhattacharya[1]⋆ on behalf of The CMS Collaboration

1 Tata Institute of Fundamental Research, Mumbai

⋆ soham.bhattacharya@cern.ch

*Proceedings for the 15th International Workshop on Tau Lepton Physics,
Amsterdam, The Netherlands, 24-28 September 2018*

Abstract

We present the algorithm and performance of tau reconstruction at the CMS experiment, while highlighting a dedicated reconstruction algorithm that uses calorimeter hits instead of tracks to reconstruct taus with high transverse momentum. Describing the standard Hadrons-Plus-Strips (HPS) algorithm and its dependence on track reconstruction and shower modelling, we present the calorimetric tau (calo-tau) reconstruction that uses minimal track information for high p_T taus. The pros and cons of these algorithms are discussed along with their performance and potential uses. It is found that the calo-tau algorithm outperforms the HPS algorithm in the high efficiency region. This study is work in progress, and is an attempt to tune the reconstruction for high p_T taus. The calo-tau algorithm is not yet an official tau reconstruction algorithm for CMS.

1 Introduction

Searches involving high momenentum τ leptons have gained prominence in proton-proton collision at the LHC, in particular in the context of Beyond Standard Model physics (like heavy resonances decaying to taus). Hence it is important to study τ reconstruction techniques that are more optimized for such high momenta where the standard algorithms may not be as efficient. The τ lepton decays to hadrons with a branching ratio of $\sim$ 65%, as seen in Table 1. The CMS (Compact Muon Solenoid) experiment uses the Hadrons-Plus-Strips (HPS) algorithm to reconstruct these hadronic decay modes. The HPS algorithm relies heavily on track reconstruction. A boosted three-prong τ_h (hadronically decaying τ lepton) may not have all its tracks well resolved, and may appear as a two-pronged object, and hence will not be treated as a τ_h candidate by the algorithm. A tau-reconstruction algorithm that relies primarily on calorimeter deposits only, is expected to be free of these issues, and can be useful at very high p_T ($\sim$ TeV). Moreover, a calorimeter based tau reconstruction algorithm is more robust against possible mismodellings in Monte-Carlo simulations, and can serve as a cross-check of whether potential high p_T τ_h signals are lost in data. In this context, we will first describe the HPS algorithm in Section 2, followed by a calorimeter based algorithm in Section 3. Finally, a comparison between the two algorithms is presented in Section 4.

Table 1: Decay modes of the τ lepton and their branching fractions [1].

Decay mode (DM)	Resonance (mass in GeV)	$\mathcal{B}$ [%]
Leptonic decays		
$\tau^- \to e^- \bar{\nu}_e \nu_\tau$		17.4
$\tau^- \to \mu^- \bar{\nu}_\mu \nu_\tau$		17.8
Total		**35.2**
Hadronic decays		
$\tau^- \to h^- \nu_\tau$		11.5
$\tau^- \to h^- \pi^0 \nu_\tau$	ρ (0.77)	25.9
$\tau^- \to h^- \pi^0 \pi^0 \nu_\tau$	a_1 (1.26)	9.5
$\tau^- \to h^- h^+ h^- \nu_\tau$	a_1 (1.26)	9.8
$\tau^- \to h^- h^+ h^- \pi^0 \nu_\tau$		4.8
Others		3.3
Total		**64.8**
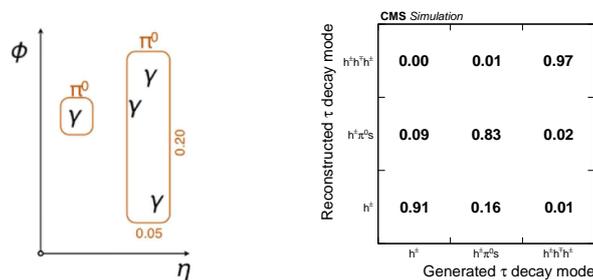
Figure 1: Left: A diagram showing the strip shape and size (Run-1) in the $\eta \times \phi$ plane. Right: Correlation between generated and reconstructed τ_h decay modes for τ_h decays in $Z/\gamma^* \to \tau\tau$ events (Run-1) [2].

2 The Hadrons-Plus-Strips (HPS) algorithm

The Hadrons-Plus-Strips (HPS) algorithm is the default algorithm used for τ_h reconstruction at the CMS experiment. The basic steps of the algorithm are as follows [1]:

- The HPS algorithm is seeded by anti-k_T jets with a distance parameter of 0.4 (AK4 jets).

- The electron and photon constituents in the jet are clustered into "strips" which try to capture the neutral pion decay.

- The strip size was fixed ($\Delta\eta \times \Delta\phi = 0.05 \times 0.2$) in Run-1 (Figure 1, left), and is dynamic (p_T dependent) for Run-2.

- The algorithm forms the following τ_h candidates (corresponding to the τ_h decay modes shown in (Table 1)). The correlation between generated and reconstructed τ_h decay modes is shown in Figure 1 (right).

  - $h^\pm$: A single charged hadron candidate without any strips.
  - $h^\pm\pi^0$: Combination of one charged hadron and one strip.
  - $h^\pm\pi^0\pi^0$: Combination of one charged hadron and two strips.
  - $h^\pm h^\mp h^\pm$: Combination of three charged hadrons without any strips.

2.1 τ_h^{HPS} vs. QCD jet discrimination: Isolation-sum

QCD jets are expected to have higher activity (tracks, calorimeter deposits) in an annular region around the signal cone compared to τ_h jets in which most of the energy is carried

by the charged hadron and the electrons/photons in the signal cone. So one can define an isolation region (cone) around the τ_h axis and place a cut on the energy deposit in that region to discriminate against QCD jets. An isolation cone size of $\Delta R = 0.5$ around the τ_h axis is considered. A smaller cone size (0.3) is also used for busier environments like $t\bar{t}$ events. Then the isolation of a τ_h candidate is computed as [1]:

$$I_\tau = \sum_{d_z < 0.2 \text{ cm}} p_T^{\text{charged}} + \max\left(0, \sum p_T^{e/\gamma} - \Delta\beta \sum_{d_z > 0.2 \text{ cm}} p_T^{\text{charged}}\right). \tag{1}$$

In the above equation, p_T^{charged} and $p_T^{e/\gamma}$ are the transverse momenta of charged hadrons and electrons/photons, respectively. In the first summation, charged hadrons with d_z (longitudinal impact parameter with respect to the primary vertex) greater than 0.2 cm are excluded to reduce tracks from pileup. Charged hadrons and electrons/photons that are part of the τ_h candidate are excluded from the sum. The $\Delta\beta$ term takes care of the contribution from pileup to the photon isolation. The value of $\Delta\beta$ is 0.2 (0.46) for Run-2 (Run-1). Here only the charged hadrons coming from pileup ($d_z > 0.2$ cm) are used. In addition to this, a cut on the p_T-sum of the e/γ that are out of the signal cone but in the strips, helps to reduce the misidentification probability [1].

$$p_T^{\text{strip, outer}} = \sum_{\Delta R > R_{\text{sig}}} p_T^{e/\gamma} < 0.1 \; p_T^{\tau_h},$$
$$R_{\text{sig}} = \frac{3.0 \text{ GeV}}{p_T^{\tau_h} \text{ [GeV]}}, \qquad R_{\text{sig}} \in [0.05, 0.1]. \tag{2}$$

2.2 τ_h^{HPS} vs. QCD jet discrimination: MVA

A BDT (Boosted Decision Tree) is also trained to discriminate τ_h jets from QCD jets. Its relative performance w.r.t. the isolation-based discrimination is shown in Figure 2. The MVA-based discriminator outperforms the isolation-based one at both low ($H \to \tau\tau$ events) and high ($Z'(2 \text{ TeV}) \to \tau\tau$ events) transverse momenta. The efficiencies and misidentification probabilities of the different MVA-based working points as a function of p_T are shown in Figure 3.

3 The calorimetric tau (calo-tau) algorithm

In the previous section we saw that that HPS algorithm relies heavily on track resolution, track momentum measurement, and electron, photon, and charged hadron reconstruction. The calo-tau reconstruction algorithm's robustness lies in its simplicity. The algorithm has been constructed in way such that its dependence on track momentum measurement in minimized, and unlike the HPS-tau algorithm, electron/photon/charged hadron reconstruction plays no role here. The main steps are as follows.

- Seed the algorithm with a calorimeter jet (calo-jet) reconstructed with the anti-k_T algorithm with a distance parameter of 0.4 (AK4 calo-jet).

- The existence of a track with the following condition is required to select the jet as a τ candidate. Note that the p_T measurement of this track does not play a significant role. The track ($p_T > 0.5$ GeV) must be within a cone of $\Delta R < 0.1$ around the jet axis. The track's transverse impact parameter (d_0) must be < 0.1 cm.
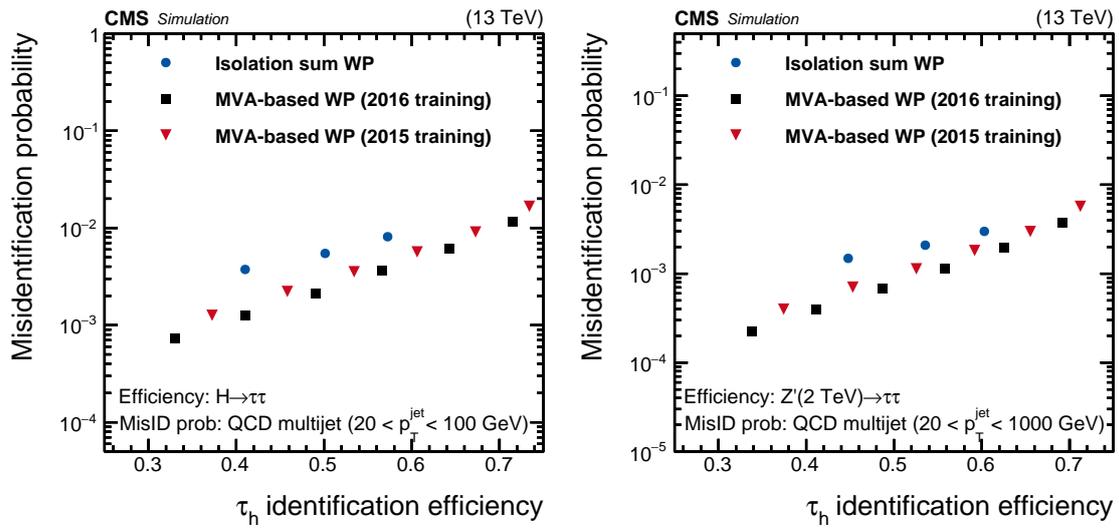
Figure 2: jet $\to \tau_h$ misidentification probability versus the τ_h identification efficiency for the different isolation-based and MVA-based HPS-tau working points. The result in $H \to \tau\tau$ events is shown on the left, and that in $Z'(2\,\text{TeV}) \to \tau\tau$ events, on the right [1].
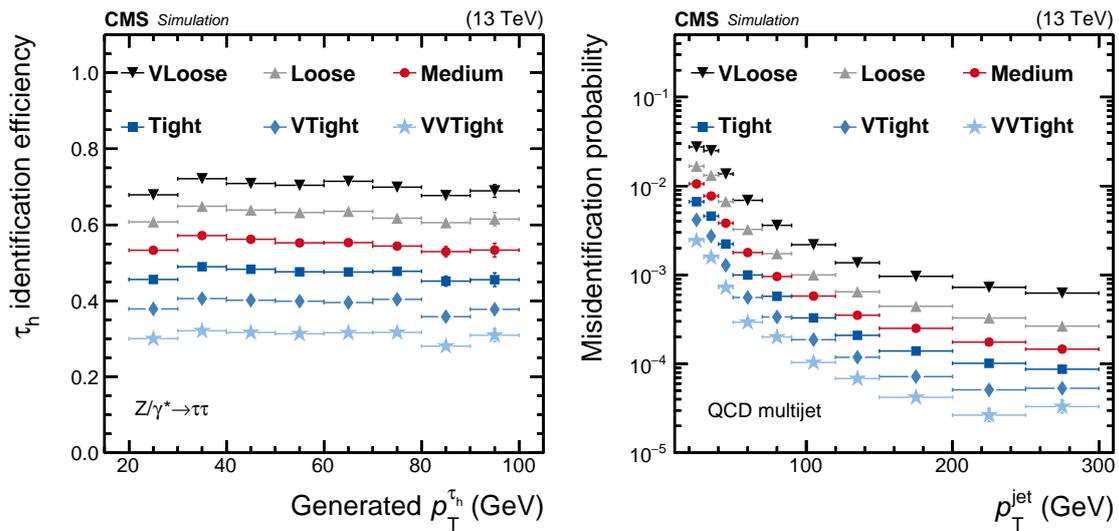
Figure 3: Left: τ_h identification efficiency vs generated τ_h p_T in $Z/\gamma^* \to \tau\tau$ events for the MVA-based HPS-tau working points [1]. Right: Probability of a jet being misidentified as a τ_h in QCD multijet events for the same working points [1].

- Set the 4-momentum of the calo-tau to that of the calo-jet. Note that the track p_T measurement does not play any role here either.

3.1 τ_h^{calo} vs. QCD jet discrimination: Isolation-sum

Similar to HPS-taus, we define the following isolation-sum.

$$I_{\text{iso},\rho}^{\text{comb}} = H_T^{\text{iso-trk}} + \max(0, E_T^{\text{iso-ECAL}} - \rho A_{\text{eff}}) \tag{3}$$

- $H_T^{\text{iso-trk}}$: Scalar sum of the p_T of tracks in the annular region $0.07 < \Delta R < 0.5$ (w.r.t. the

leading signal track) if their longitudinal impact parameter satisfies $\Delta d_z < 1$ cm w.r.t. the leading signal track (to reduce contribution from pileup).

- $E_T^{\text{iso-ECAL}}$: Sum of ECAL (electromagnetic calorimeter) deposits with transverse energy $E_T > 0.5$ GeV within an the annular region $0.15 < \Delta R < 0.5$ w.r.t. the leading signal track.

- ρ is the energy density in the event, and is defined as the median of the calo-jet energies divided by their respective jet-areas.

- A_{eff} is an effective area whose value (0.2) is chosen such that the efficiency is independent of pileup.

- The product ρA_{eff} is the contribution from pileup to the ECAL energy deposits.
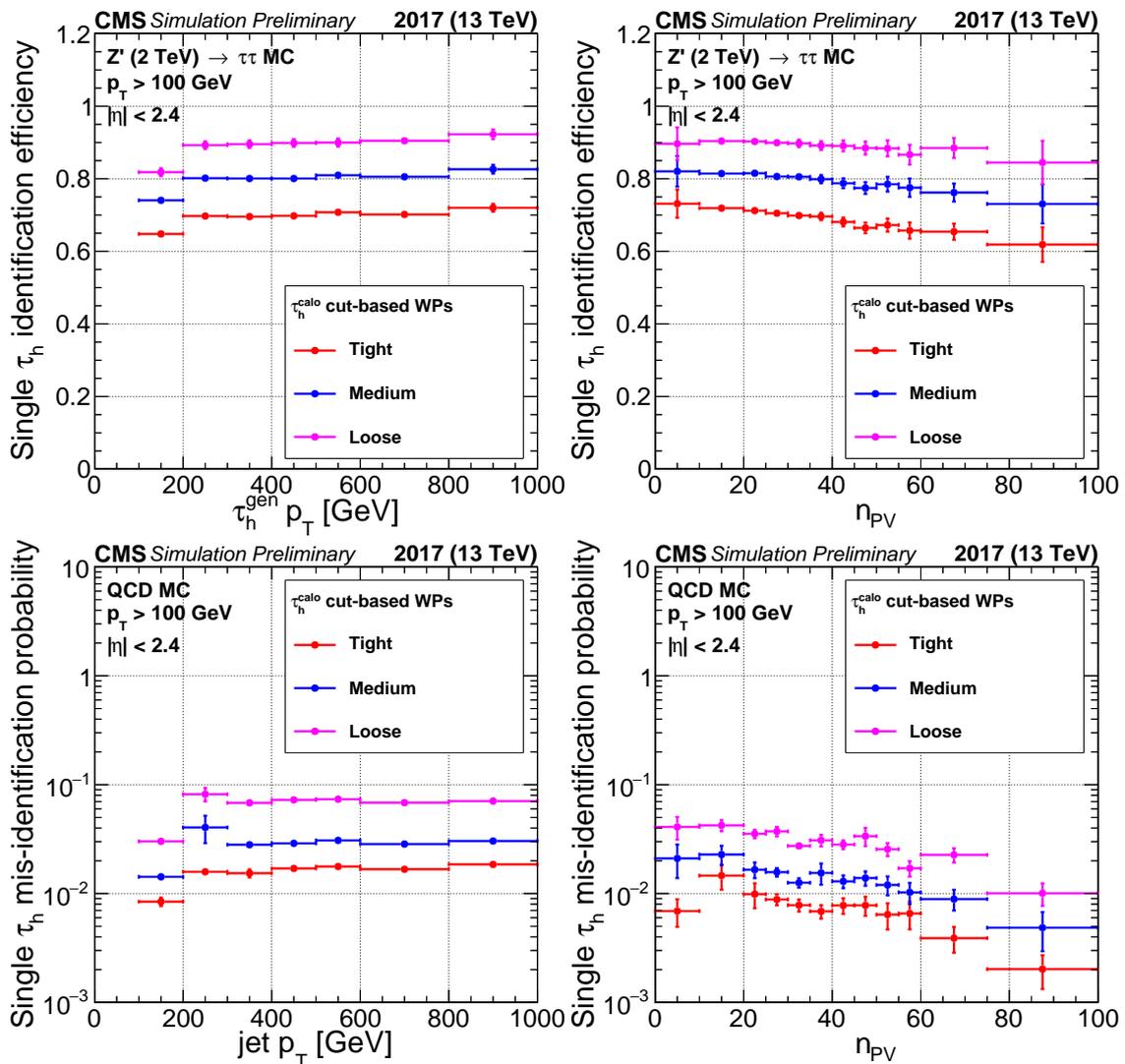
Figure 4: Left: Efficiencies (top) and misidentification probabilities (bottom) of the different isolation-based calo-tau working points as a function of generated τ_h p_T (in $Z'(2 \text{ TeV}) \to \tau\tau$ events) and jet p_T (in QCD events) respectively. Right: Same as left, but as a function of the number of vertices.
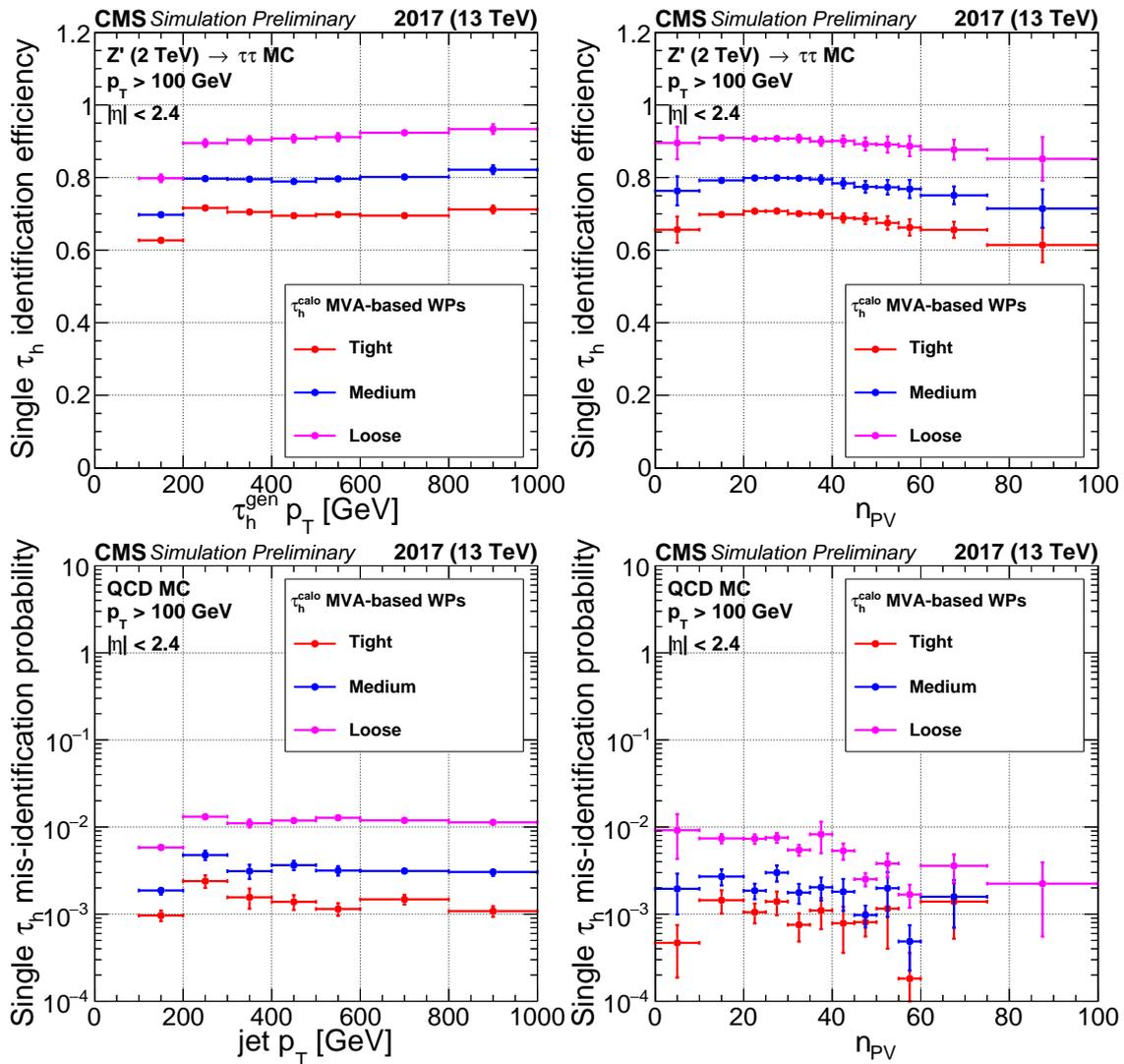
Figure 5: Left: Efficiencies (top) and misidentification probabilities (bottom) of the different MVA-based calo-tau working points as a function of generated τ_h p_T (in $Z'(2\,\mathrm{TeV}) \to \tau\tau$ events) and jet p_T (in QCD events) respectively. Right: Same as left, but as a function of the number of vertices.

The performance of the different working points of the isolation-based discriminant (as a function of p_T and pileup conditions) in $Z'(2\,\mathrm{TeV}) \to \tau\tau$ events is shown in Figure 4. Both the efficiency and misidentification probability are flat across a wide range of p_T. The effect of pileup has also been minimized, as can be seen from flatness of the efficiency and misidentification probability across the number of vertices.

3.2 $\tau_h^{\mathbf{calo}}$ vs. QCD jet discrimination: MVA

A BDT has also been trained to discriminate between genuine τ_h and QCD jets. The most discriminating variables used for the training are:

- $n_{\mathrm{sig\text{-}trk}}$: Number of signal tracks.
  Tracks within $\Delta R < 0.07$ w.r.t. the leading signal track.

- $n_{\mathrm{iso\text{-}trk}}$: Number of isolation tracks.
  Tracks in the annular region $0.07 < \Delta R < 0.5$ w.r.t. the leading signal track.

- m: Invariant mass of the calo-tau.

- E_T^{iso}: Sum of the ECAL energy deposits (transverse component) in the isolation annulus.

- $d_{xy}^{\text{sig-trk}_1}$: The transverse impact parameter of the leading signal track.

- $d_z^{\text{sig-trk}_1}$: The longitudinal impact parameter of the leading isolation track.

- p_T weighted average of ΔR between the τ_h^{calo} and the following:

    - The ECAL energy deposits in the signal cone ($\Delta R < 0.15$ w.r.t. the leading signal track).
    - The ECAL energy deposits in the isolation annulus ($0.15 < \Delta R < 0.5$ w.r.t. the leading signal track).

The number of vertices in the event has also been used by the BDT so that it learns the pileup dependence of the variables and the training is pileup independent. The performance of the different working points of the MVA-based discriminant (as a function of p_T and pileup conditions) in $Z'(2\,\text{TeV}) \rightarrow \tau\tau$ events is shown in Figure 5. Both the efficiency and misidentification probability are flat across a wide range of p_T. The effect of pileup has also been minimized, as can be seen from flatness of the efficiency and misidentification probability across the number of vertices.

4 Comparison between τ_h^{calo} and τ_h^{HPS} performances

The ROC curves of both the isolation-based and the MVA-based discriminators are shown for HPS-taus and calo-taus in Figure 6. The figure also shows some of the standard HPS-tau working points, namely very-loose (VL), loose (L), medium (M), tight (T), and very-tight (VT). Clearly the calo-tau MVA-based discriminant performs better than the isolation-based discriminant, as expected. However, the interesting feature is that the calo-tau algorithm is able to reach higher efficiencies ($> 70\,\%$) than the HPS-tau algorithm and performs better in that region .
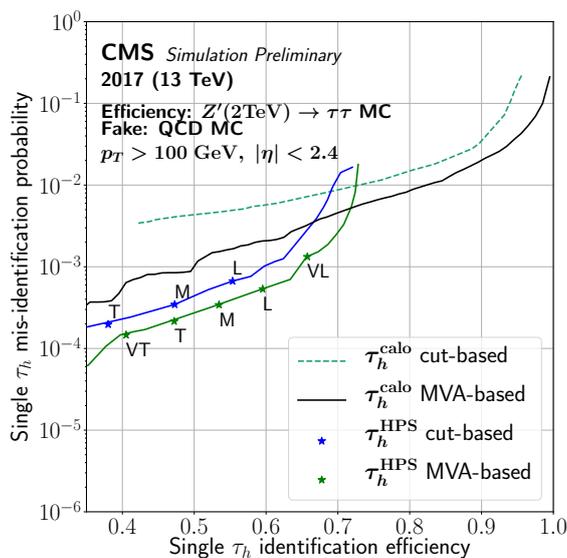
Figure 6: The ROC curves of the calo-tau and HPS-tau algorithms.

5 Conclusion

The excellent performance of the calo-tau algorithm in the high efficiency region can be useful for increasing the signal sensitivity of high momentum tau final state searches which suffer from low event yields in the search region. This can be confirmed after performing a realistic analysis to obtain the sensitivity of a given signal process involving high p_T taus.

References

[1] A. M. Sirunyan et al., *Performance of reconstruction and identification of τ leptons decaying to hadrons and v_τ in pp collisions at $\sqrt{s}$ =13 TeV*, J. Inst. **13**, P10005 (2018), doi:10.1088/1748-0221/13/10/P10005.

[2] [The CMS collaboration], *Reconstruction and identification of τ lepton decays to hadrons and v_τ at CMS*, J. Inst. **11**, P01019 (2016), doi:10.1088/1748-0221/11/01/P01019.