Tau isolation study for CMS run2

Marcel Stefko
CERN, Geneva, Switzerland

Abstract
Final report on study of hadronic tau fake rate behavior for $H \to \tau\tau$ analysis of CMS run2 data. Differences between same-sign and opposite-sign events were studied, and Monte Carlo simulation was compared with experimental data. This study has been carried out by summer student Marcel Stefko under supervision of Yuta Takahashi and Jan Steggemann of PH-CMG-PS group.

1 Introduction

Hadronic tau (τ_h) identification plays an important part in $H \to \tau\tau$ analysis. It is important to accurately evaluate the rate at which jets are misidentified as τ_h to accurately estimate our backgrounds from processes such as W+jets, QCD multijet production, and TTbar production in the semi-leptonic decay channel.

2 Same-sign / opposite-sign event study

2.1 Motivation

In $H \to \tau\tau$ analysis, is is often assumed that fake rate (see definition in Eq. (2)) behaves in the same way in same-sign (SS) and opposite sign (OS) regions.[1] This assumption is then used to measure certain properties of the fake rate in SS region, where almost no real τ are expected,[2] since all events with real τ are OS due to conservation of charge (see Figure 1).

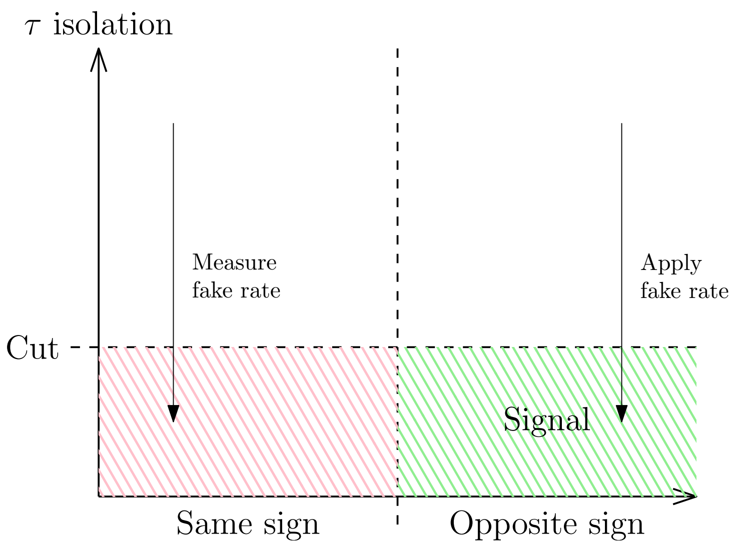


Fig. 1: In analysis, often certain properties of the fake rate are measured in same-sign region and then applied in opposite-sign region.

The first part of this study regards the investigation, whether this assumption is correct.

[1] Same-sign region is the set of events in Monte Carlo simulation or real detector data, where the two reconstructed leptons have charge of same sign (either both are positive, or both negative). In opposite-sign events, these leptons have opposite charge.

[2] More precisely, there are only few SM processes with fairly low cross sections that give rise to two same-sign leptons, such as WZ or ZZ production, but these are about two orders of magnitude suppressed in our final selection.

2.2 Dataset & isolation variable

For this study spring15 MC simulations of $t\bar{t}$, W+jets, QCD and DY processes are investigated. In these samples, one muon (μ / mu) and one reconstructed hadronic tau candidate were selected (τ / tau). On this dataset, cuts excluding all leptons and corresponding to loose Higgs-to-taus selection in the mu-tau channel were applied.[3]

An isolation criterium is a condition (for example a value of a certain variable), which separates events of certain kind (i.e. real τ) from the rest of the dataset. Important properties of the isolation criterium are its signal efficiency[4] and fake rate[5]. For this study, the isolation criterium to separate events with real tau is chosen as

tau_byCombinedIsolationDeltaBetaCorrRaw3Hits < 2.[6]

"The absolute [tau] isolation is defined as

$$I \equiv \sum_{\text{charged}} p_{\mathrm{T}} + \max\left(0, \sum_{\text{neutral}} p_{\mathrm{T}} + \sum_{\gamma} p_{\mathrm{T}} - \frac{1}{2} \sum_{\text{charged, PU}} p_{\mathrm{T}}\right), \tag{1}$$

where $\sum_{\text{charged}} p_{\mathrm{T}}$ T is the scalar sum of the transverse momenta of the charged hadrons, electrons, and muons originating from the primary vertex and located in a cone of size $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} = 0.4$ centred on the lepton direction. The sums $\sum_{\text{neutral}} p_{\mathrm{T}}$ and $\sum_{\gamma} p_{\mathrm{T}}$ represent the same quantity for neutral hadrons and photons, respectively. In the case of τ_h, the particles used in the reconstruction of the τ_h are excluded from the sums. The contribution of pileup photons and neutral hadrons is estimated from the scalar sum of the transverse momenta of charged hadrons from pileup vertices in the cone, $\sum_{\text{charged, PU}} p_{\mathrm{T}}$. This sum is multiplied by a factor of $1/2$ which corresponds approximately to the ratio of neutral to charged hadron production in the hadronization process of inelastic proton-proton collisions, as estimated from simulation."[7]

Fake rate of this variable is defined as

$$\text{Fake rate} = \frac{N(\text{isolation criterium} = \text{True AND particle is Tau} = \text{False})}{N(\text{particle is Tau} = \text{False})}. \tag{2}$$

2.3 $t\bar{t}$ study

We investigate the tau p_{T} dependence of fake rate in OS and SS regions using a $t\bar{t}$ sample - Figure 2. We can see that fake rate does behave *differently* in these regions, so one has to be careful when applying measurements from one region to the other one.

One theory to explain this difference is, that these two regions have different compositions of particles (especially quarks), and individual particles have different fake rates. This can explain the difference, and in Figs. 3a,3b and 4, we see that these differences do exist.

If we multiply each fraction of the particle (Figs. 3a, 3b) with its respective fake rate (Fig. 4), and sum over all particles for SS and OS region, we should obtain the original dependence (Fig. 2). However, while the resulting calculated dependence (Fig. 5) does show a difference between OS/SS fake rates, the

[3](tau_gen_pdgId $\neq$ 11-15) && !veto_otherlepton && !veto_thirdlepton && !veto_dilepton && (tau_againstMuon3 > 1) && (tau_againstElectronMVA5 > 2)

[4]What fraction of targeted events pass the isolation criterium.

[5]What fraction of non-targeted events pass the isolation criterium (eq. 2).

[6]Results were also verified by repeating the study with isolation citerium **tau_byIsolationMVAnewDMwLT >= 3**. This verification led to same conclusions as the study.

[7]Source: Evidence for the 125 GeV Higgs boson decaying to a pair of tau leptons. **arXiv:1401.5041 [hep-ex]**

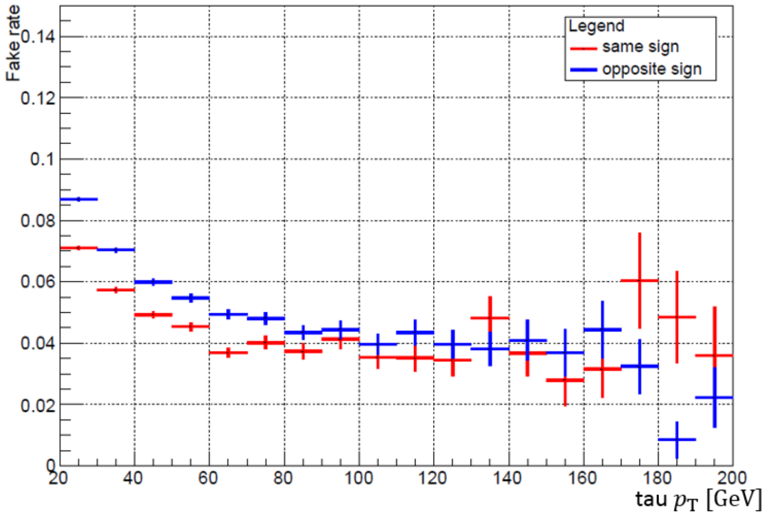
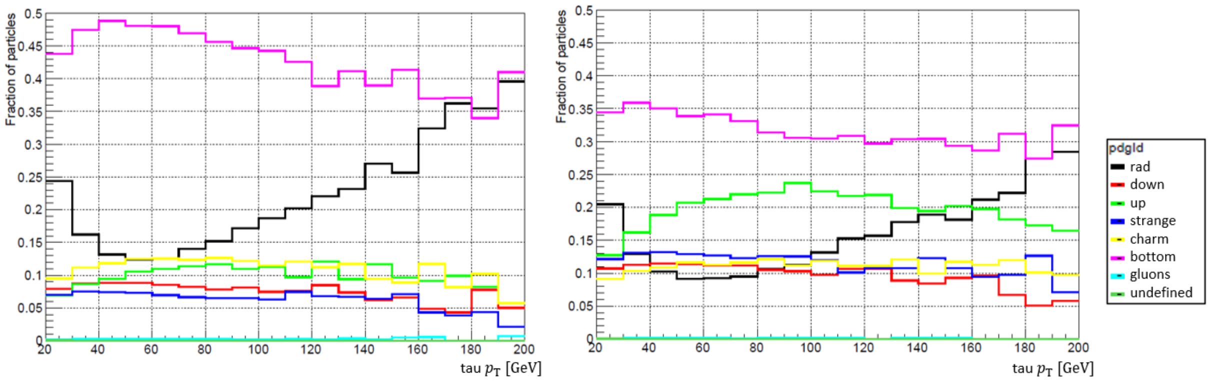


Fig. 2: Comparison of fake rate between same-sign and opposite-sign events in $t\bar{t}$ sample, with dependence on tau p_T.



(a) Same-sign.

(b) Opposite-sign.

Fig. 3: Comparison of particle composition of same-sign and opposite-sign regions of $t\bar{t}$ sample. There are non-negligible differences in composition. For example, the same-sign sample has a noticeably higher fraction of b-quarks, opposite sign has a higher fraction of u-quarks.

difference is not as significant as in real data. This can be explained by the fact, that individual parton fake rate also varies between SS and OS regions. Fig. 6 shows, that SS events, with the exception of b-quarks, have lower fake rates than OS events.

The proposed hypothesis to explain this phenomenon is following: Assume that the muon is correctly reconstructed as μ^-. Then the fake tau (u or $\bar{d}$) has positive charge (the W^+ decays into 2 positively-charged quarks - Fig. 7), and if it is identified as τ^-, it had to undergo a process with a larger energy deposit, which is less likely to pass the isolation criterium. So a fake tau with same charge as the mu (and thus opposite charge from its origin particle) produces lower fake rate. The same argument applies also for the muon being μ^+. Examples of processes where this "sign flip" occurs are depicted in Fig. 8a and 8b, and respective fake rates of these two kinds of processes are compared in Fig. 9.

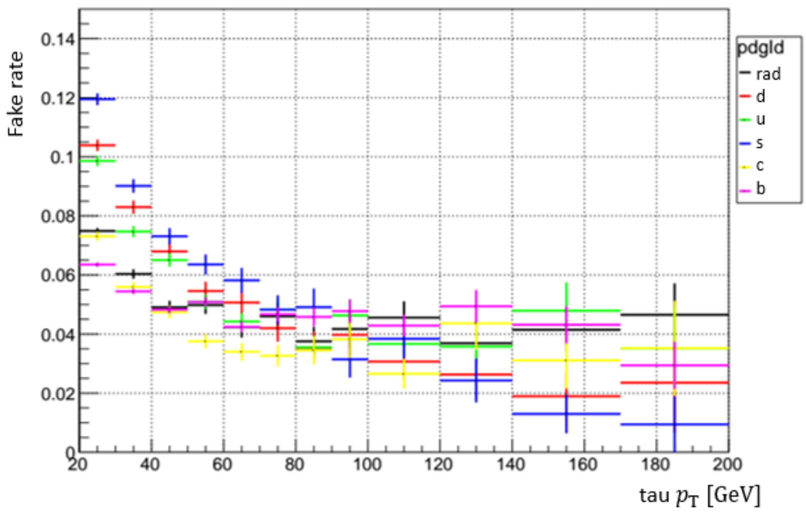


Fig. 4: Fake rate of individual quarks, with dependence on tau p_T.

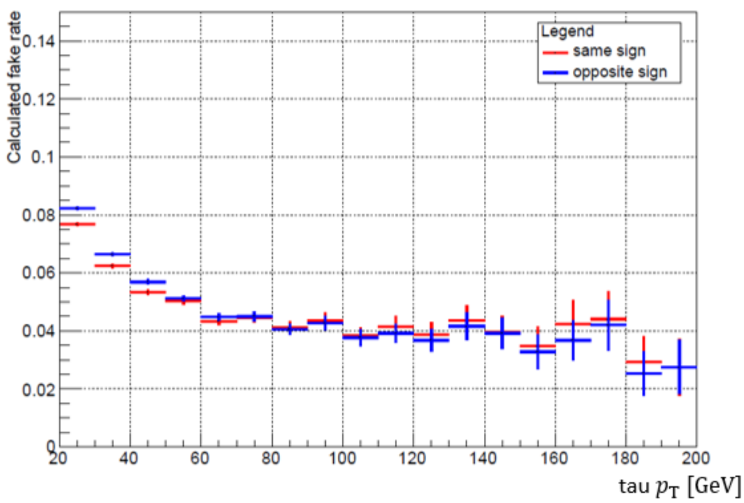


Fig. 5: Calculated difference between SS and OS fakerates, with dependence on tau p_T. The calculated difference is not as pronounced as the real difference (Fig. 2). This suggests that different particle compositions are not the only cause of the difference.

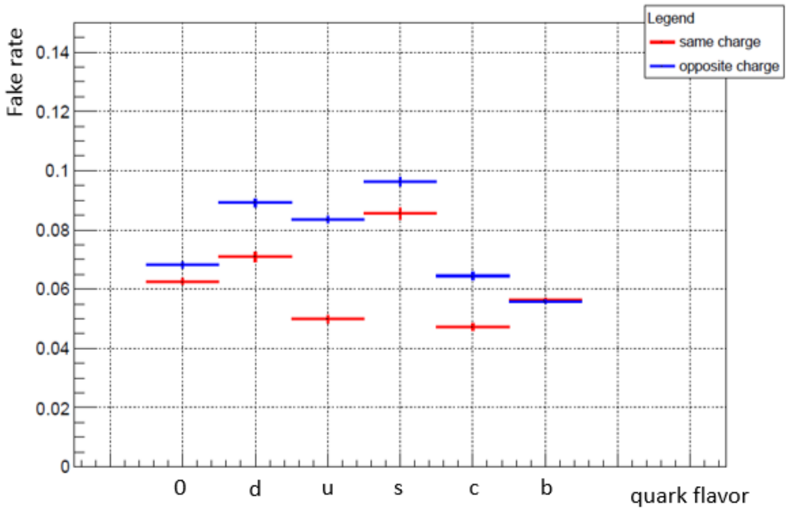


Fig. 6: Fake rates of different quarks in SS/OS regions of TTbar sample.

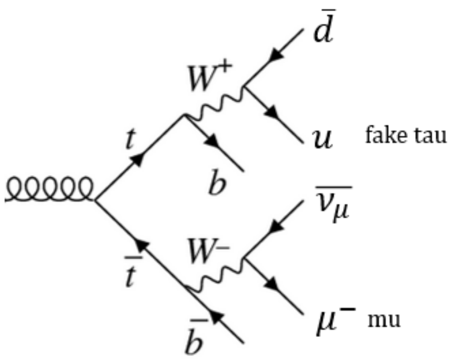
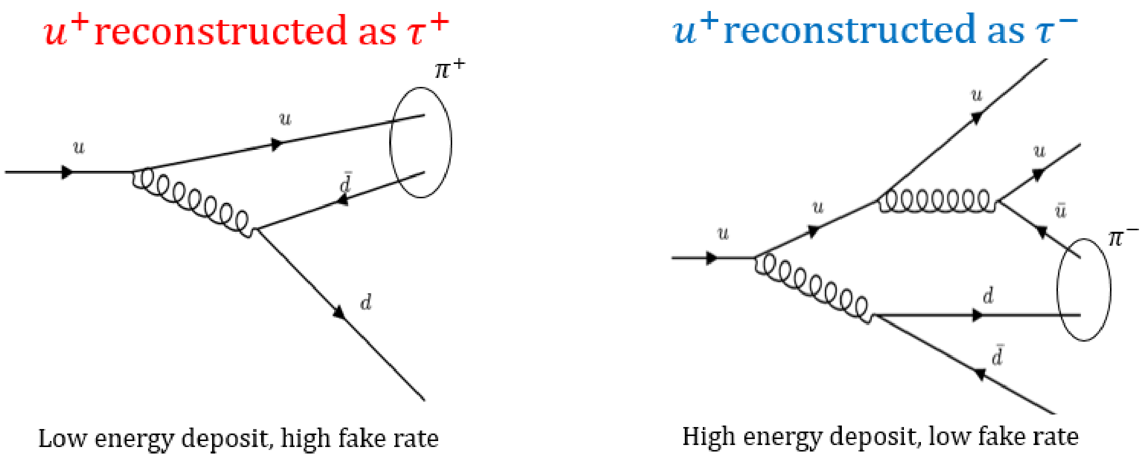


Fig. 7: Example of a $t\bar{t}$ process. If the muon is correctly identified as μ^-, then the fake tau (in this case the u-quark) has to be positively-charged due to conservation of charge.



(a) Process with no sign flip.

(b) Process with sign flip.

Fig. 8: Comparison of two u-quark hadronisation processes. A process, in which the u turns into a π^+ (no sign flip), has a lower energy deposit, than a process where sign flip occurs.

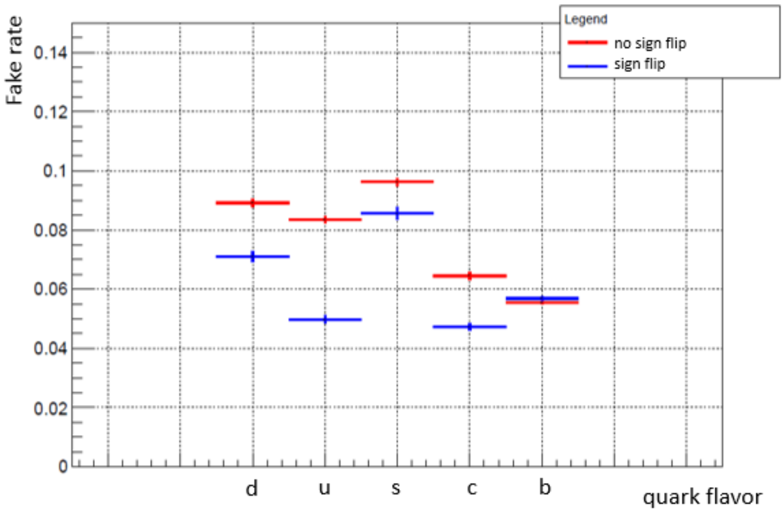


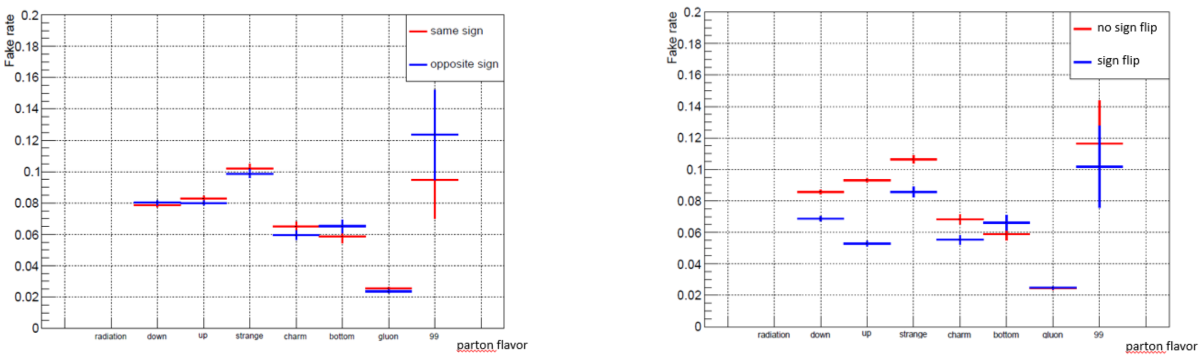
Fig. 9: Comparison of fake rates between sign-flip and no sign-flip events in TTbar sample for different quark flavors. The no sign-flip events have higher fake rate, which is in agreement with our hypothesis (except for b-quarks, for which the explanation is not yet known).

2.4 Other samples

By repeating the study for other samples, we can verify if this phenomenon is found also in other events than $t\bar{t}$. In Table 1, ratios between SS and OS region fake rates using two different isolation criteria are displayed. While the data shows that for some samples, there is no difference between SS and OS fake rates, the difference between fake rates in sign-flip and no-sign-flip events is still present in all samples (an example is shown in Figs. 10a and 10b for the Drell-Yan sample).

Table 1: Ratio of SS/OS fake rates for different samples and isolation variables.

	deltaBeta	byIsolationMVA
TTbar	1.23 ± 0.01	1.28 ± 0.01
W+Jets	1.43 ± 0.02	1.89 ± 0.04
QCD	0.96 ± 0.07	1.04 ± 0.10
Drell-Yan	0.99 ± 0.02	1.01 ± 0.03



(a) SS/OS

(b) Sign flip/no sign flip

Fig. 10: Comparison of fake rates of SS/OS regions and sign-flip/no-sign-flip regions. In the SS/OS comparison, there is no significant difference, but in sign-flip/no-sign-flip regions, the difference is present in accordance with the hypothesis.

Also, it is worth pointing out, that fake rate is strongly dependent on parton flavor (Fig. 11), and the ratio of amount of sign-flip/no-sign-flip particles is constant for each parton flavor and process-invariant (Fig. 12).

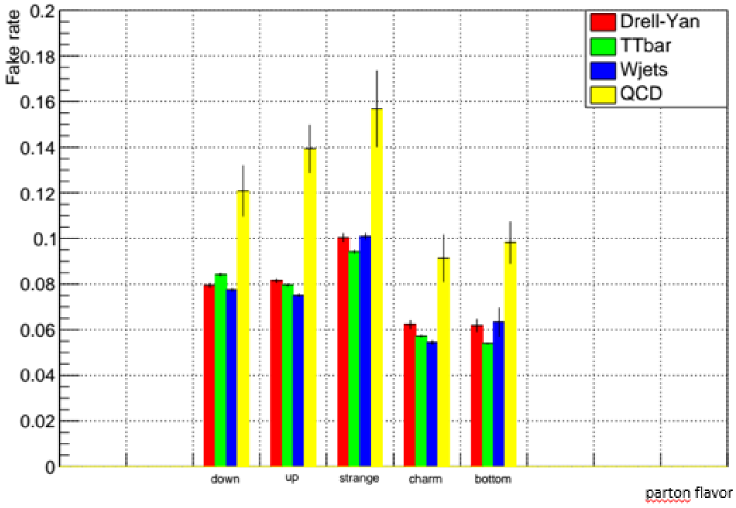


Fig. 11: Fake rate for each parton flavor in different processes.

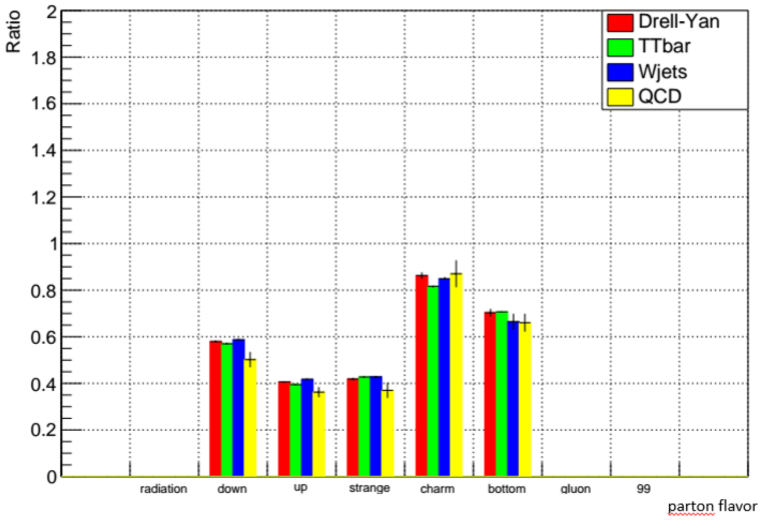


Fig. 12: Ratio of particles with/without sign flip for each parton flavor and process.

3 Verification of Monte Carlo simulation accuracy

In this study, Monte Carlo simulations aree compared with real CMS data from 2015 (40 pb^{-1}, centre-of-mass energy 13 TeV). In the sample, enriched regions of $t\bar{t}$[8], W+jets[9] and QCD[10] samples are defined, and the MC fakerate is compared with data fakerate. Fake rate in this study is defined as

$$\text{Fake rate}(X = \text{MC,Data}) = \frac{X_{\text{iso}} - \text{MC}_{\text{iso}}^{\text{real }\tau}}{X_{\text{non-iso}} - \text{MC}_{\text{non-iso}}^{\text{real }\tau}}. \tag{3}$$

Resulting fake rates are marked in Table 2. Within uncertainties, simulation matches the experiment.

[8] n_bjets > 0
[9] (n_bjets==0) && (mt>70 && mt<100)
[10] (mu_pt<50) && (met_pt<40) && (mt<60) && SameSign

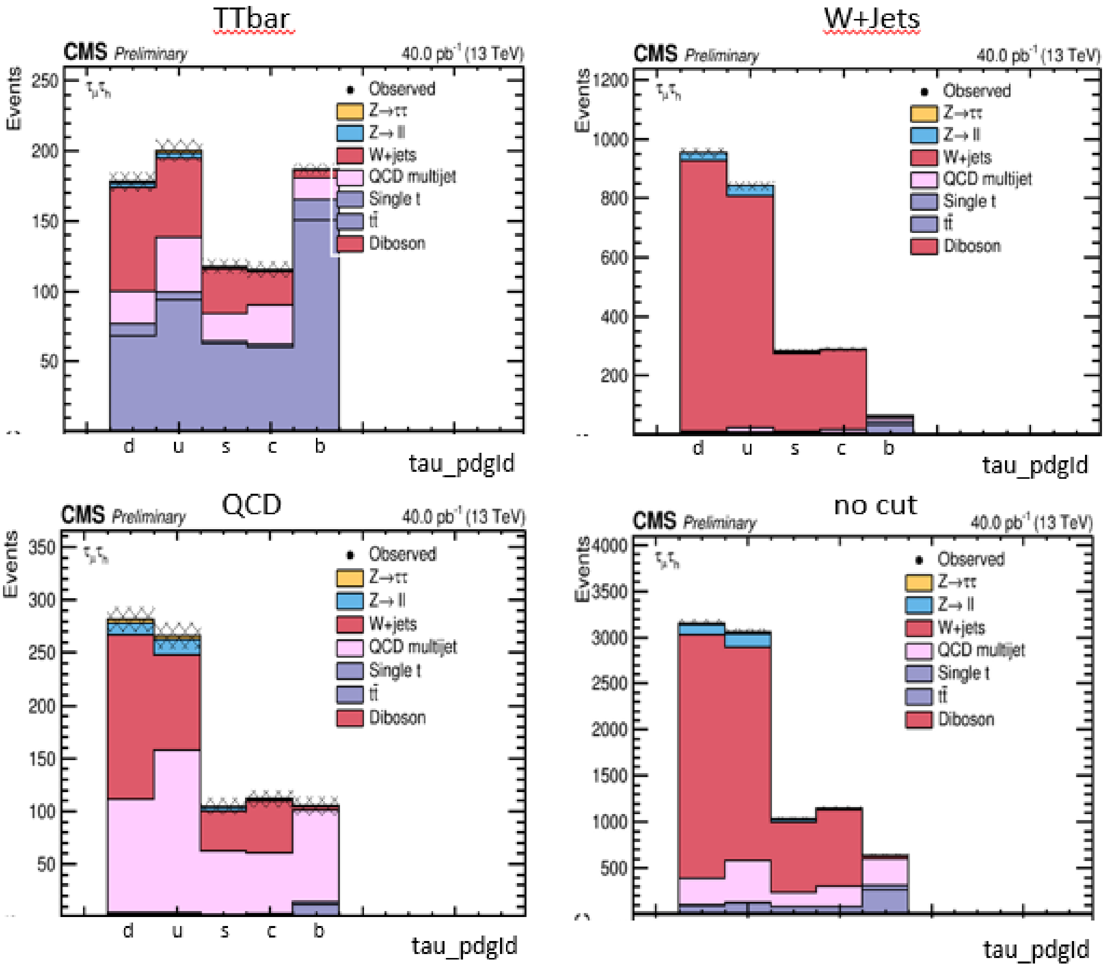


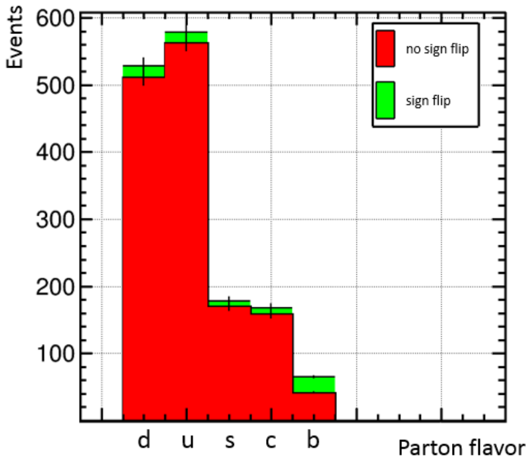
Fig. 13: Particle composition of enriched regions for different processes. Note that this composition varies between regions. This could be used in the future to define regions enriched in individual partons, to verify that fake rate for individual partons does vary, as proposed in Section 2, Fig. 11.

Table 2: Fake rates of data and MC samples in enriched regions. Data and simulation agree within uncertainties.

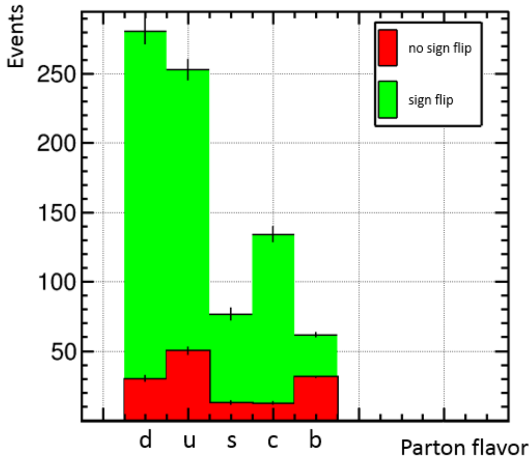
	TTbar	W+jets	QCD
Data fake rate [%]	7.29 ± 0.81	6.40 ± 0.37	7.06 ± 0.59
MC fake rate [%]	5.82 ± 0.31	6.28 ± 0.20	6.31 ± 0.46
Ratio Data/MC	1.25 ± 0.22	1.05 ± 0.12	1.12 ± 0.14

3.1 Sign flip - fake rate relation

To verify the relation between sign flip and fake rate outlined in Section 2, two regions are defined: OShighMT[11] and SShighMT[12], which are rich in no-sign-flip and sign-flip events respectively - see Fig. 14. Fake rates are evaluated in Table 3. The results support our proposed hypothesis, that sign-flip events have lower fake rate than no-sign-flip events.



(a) OShighMT



(b) SShighMT

Fig. 14: Composition of OS and SS high mt regions. We can see that OS region is dominated by non-sign-flip events, and SS region by sign-flip events, which allows us to study their fake rates - see Table 3.

Table 3: Fake rates of data and MC samples in regions enriched in non-sign-flip and sign-flip events. Data and simulation support the hypothesis proposed in Section 2.3

	OShighMT	SShighMT
Data fake rate [%]	7.45 ± 0.56	4.47 ± 0.49
MC fake rate [%]	7.29 ± 0.27	4.10 ± 0.24
Ratio Data/MC	1.02 ± 0.11	1.09 ± 0.18

4 Conclusion

From the results of this study, several conclusions can be made:

– Fake rate does not behave in same way in same/opposite sign events âĂŞ distinguishing between these events can help us better estimate the fake rate.
– Tau candidates with ”flipped sign” have lower fake rates due to larger energy deposits, which make them easier to identify and exclude - this is also supported by available data.
– Probability of ”sign flip” of tau candidate is dependent on quark type, but process-invariant.
– Within uncertainties, data and MC fake rates agree in all defined enriched regions.
– With more data, tighter selections regions can be defined, either to achieve better sample purity, or to try to define regions enriched in individual quark flavors.

[11] mt>80 && OS
[12] mt>80 && SS