



# Data-driven background estimation in CMS

Matti Kortelainen

*Helsinki Institute of Physics*

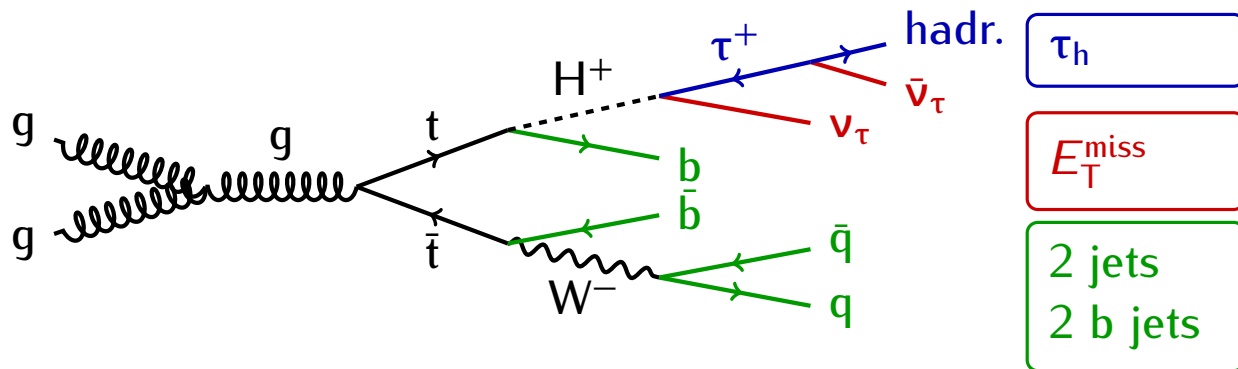
on behalf of the CMS collaboration

4th International Workshop on Prospects for  
Charged Higgs Discovery at Colliders, Uppsala

October 9, 2012

- The CMS analyses consider the following backgrounds
  - JHEP07(2012)143 (arXiv:1205.5736)
- $\tau_h$ +jets final state (Alexandros' talk)
  - QCD multijet events (jet misidentified as  $\tau_h$ )
  - EWK+ $t\bar{t}$  events with genuine  $\tau$  lepton identified as  $\tau_h$
  - EWK+ $t\bar{t}$  events with  $e/\mu$ /jet misidentified as  $\tau_h$
- $e+\tau_h$  and  $\mu+\tau_h$  final states (Pietro's talk)
  - Jet misidentified as  $\tau_h$  ( $W$  + jets,  $t\bar{t}$ )
  - $Z/\gamma^* \rightarrow \tau\tau$ , single top, diboson, and  $t\bar{t}$  with genuine  $\tau$
  - $Z/\gamma^* \rightarrow ee, \mu\mu$ , and  $t\bar{t}$  with  $e/\mu$  misidentified as  $\tau_h$
- $e+\mu$  final state (Pietro's talk)
  - $t\bar{t}$ ,  $Z/\gamma^* \rightarrow \ell\ell$ ,  $W$  + jets, single top, and diboson events
- Backgrounds with blue are measured from data, and are the topic of this talk
  - The remaining backgrounds are estimated using simulation

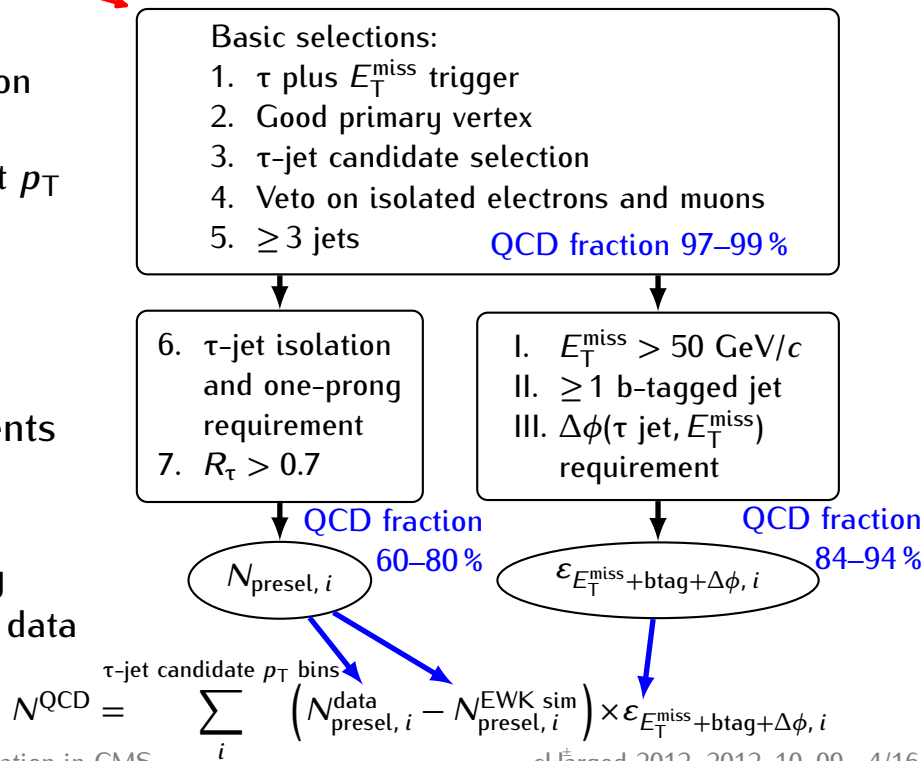
# Reminder: $\tau_h$ +jets final state analysis



1.  $\tau + E_T^{\text{miss}}$  trigger
2. Tight  $\tau_h$  identification,  $p_T > 40 \text{ GeV}/c$ 
  - Tau polarization  $R_\tau > 0.7$
3. Isolated e/ $\mu$  veto,  $p_T > 15 \text{ GeV}/c$
4.  $\geq 3$  hadronic jets  $p_T > 30 \text{ GeV}/c$
5. Missing  $E_T > 50 \text{ GeV}$
6.  $\geq 1$  jet b-tagged
7.  $\Delta\phi(\tau_h, E_T^{\text{miss}}) < 160^\circ$
8. Shape analysis with transverse mass  $m_T(\tau_h, E_T^{\text{miss}})$

## Number of events

- Background from QCD multijet events, where a jet is misidentified as the  $\tau_h$  and no genuine source of  $E_T^{\text{miss}}$
- Shape and normalization of  $m_T$  distribution were measured separately
- Factorized in bins of  $\tau_h$  candidate  $p_T$ , because
  - probability for a quark or a gluon jet to pass isolation and  $R_\tau$  requirements depends on the jet  $p_T$
  - small correlation between  $E_T^{\text{miss}}$  selection and  $\tau_h$  identification is reduced to negligible level
- Selected event samples are dominated by QCD multijet events
  - But contain also impurity from EWK and  $t\bar{t}$  events
  - Amount of them estimated using simulation, and subtracted from data



## Uncertainties and results

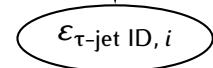
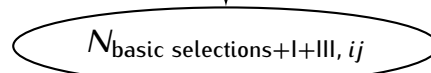
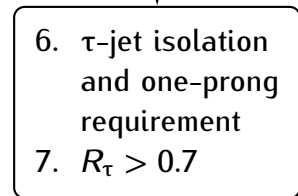
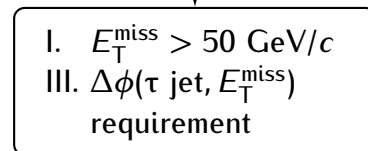
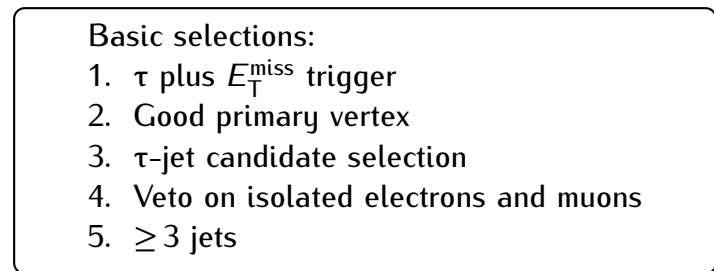
- Dominant uncertainty is the amount of data (6.5 %)
- Systematic uncertainty due to subtraction of EWK+ $t\bar{t}$  events was accounted for by
  - assuming 20 % uncertainty on EWK+ $t\bar{t}$  simulation, and
  - propagating this uncertainty using error propagation
- QCD multijet event yields for three  $\Delta\phi$  selection options:

$\Delta\phi(\tau \text{ jet}, E_T^{\text{miss}})$ option	$N^{\text{QCD}}$
Without $\Delta\phi$ selection	$42 \pm 3 \text{ (stat.)} \pm 2 \text{ (syst.)}$
$\Delta\phi < 160^\circ$	$26 \pm 2 \text{ (stat.)} \pm 1 \text{ (syst.)}$
$\Delta\phi < 130^\circ$	$17.0 \pm 1.2 \text{ (stat.)} \pm 0.6 \text{ (syst.)}$

# QCD multijet background

## Shape of $m_T$ distribution

- Obtain  $m_T$  distributions after  $E_T^{\text{miss}}$  and  $\Delta\phi$  requirements in bins of  $\tau_h p_T$ 
  - B tagging has negligible effect on the shape  $\Rightarrow$  leave out
- Weight distributions by  $\epsilon_{\tau\text{-jet ID}, i}$
- Add up all  $m_T$  distributions
- Summed  $m_T$  distribution normalized to  $N^{\text{QCD}}$  (from previous slide)



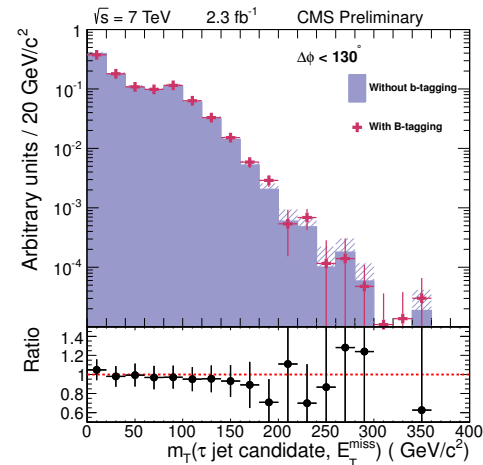
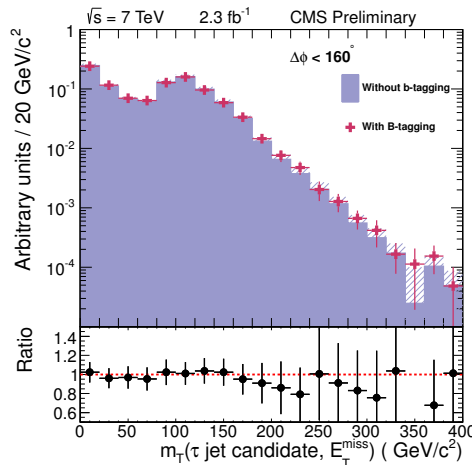
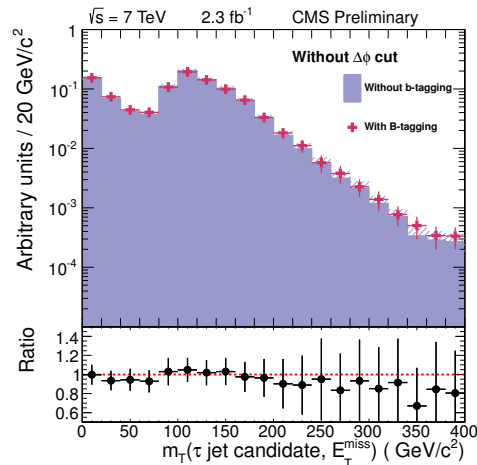
Number of events in  $m_T$  bin  $j$ ,  
 before normalization to  $N^{\text{QCD}}$

$$N(m_T)_j = \sum_i \left( N_{\text{basic selections+I+III}, ij}^{\text{data}} - N_{\text{basic selections+I+III}, ij}^{\text{EWK sim}} \right) \times \epsilon_{\tau\text{-jet ID}, i}$$

$\tau\text{-jet candidate } p_T \text{ bins}$

# QCD multijet background

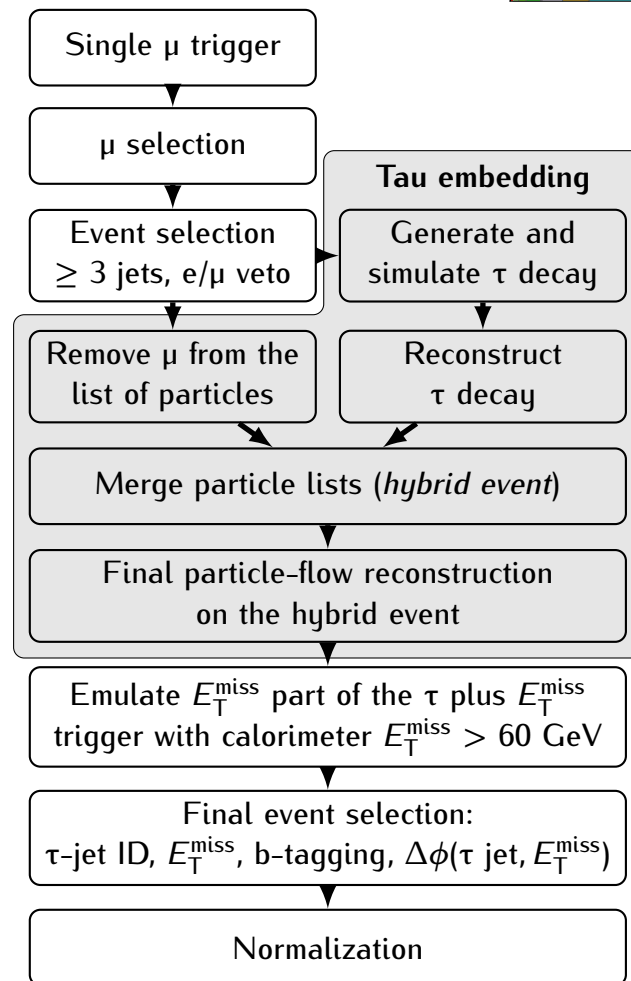
## Shape of $m_T$ distribution: result



- With and without b tagging
  - Distribution shapes agree well
  - ⇒ leaving b tagging out from shape extraction is justified
- Two “bumps”
  - $\tau_h$  energy underestimated/overestimated
  - $m_T \sim 0 \text{ GeV}/c$ :  $\tau_h$  and  $\vec{E}_T^{\text{miss}}$  are collinear
  - $m_T \sim 120 \text{ GeV}/c$  (signal region):  $\tau_h$  and  $\vec{E}_T^{\text{miss}}$  are back-to-back
  - ★ Can be controlled with  $\Delta\phi(\tau \text{ jet}, E_T^{\text{miss}})$  requirement

# EWK+ $t\bar{t}$ genuine $\tau$ background

- Background from SM  $t\bar{t}$ ,  $W + \text{jets}$ ,  $Z/\gamma^*$ , single top, VV events with a genuine  $\tau$  lepton
- Basic idea is to exploit lepton universality  $\mathcal{B}(W \rightarrow \mu) = \mathcal{B}(W \rightarrow \tau)$
- Control sample:  $\mu + \geq 3 \text{ jets}$
- Tau embedding done at particle flow level
  - Tau decay simulated and reconstructed, with tau lepton having same momentum as muon
  - Tau polarization assuming  $W \rightarrow \tau\nu$  decay
- Apply remaining event selections
- Normalization
  - $\tau$  trigger efficiency
  - Muon trigger and ID efficiency
  - Correct for  $W \rightarrow \tau \rightarrow \mu$  events
- Increase statistical precision by repeating embedding 10 times
- Small residual background from ditau events
  - Veto of 2nd  $\mu$  is tighter than veto of 2nd  $\tau_h$
  - Estimated from simulation





# EWK+ $t\bar{t}$ genuine $\tau$ background

## Control sample selection



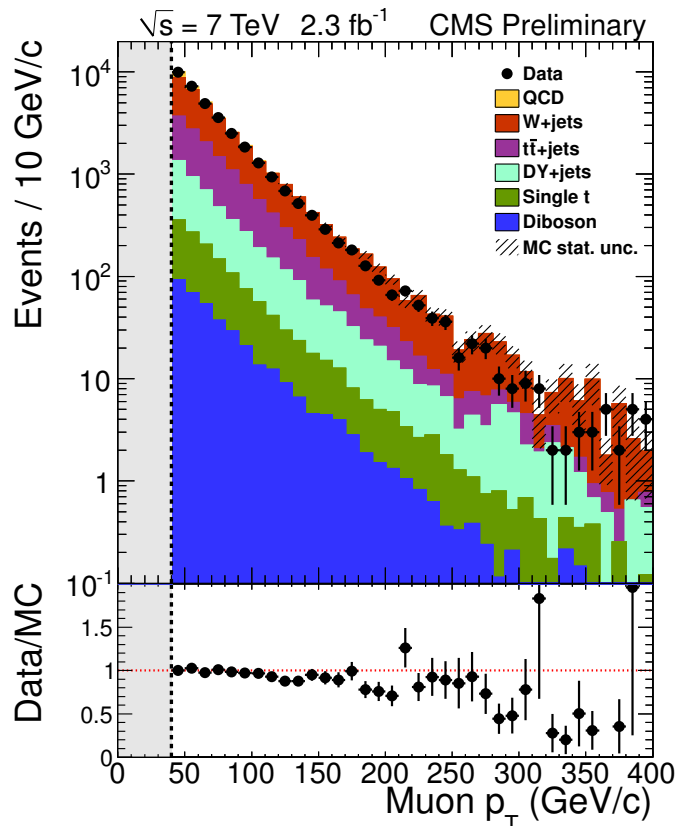
### • Selection

- Single muon trigger
- Require muon with  $p_T > 40 \text{ GeV}/c$ ,  $|\eta| < 2.1$ 
  - ★ Isolation similar to taus, but looser
- Isolated e/other  $\mu$  veto
- Require at least three jets with  $p_T > 30 \text{ GeV}/c$ ,  $|\eta| < 2.4$

### • Reasonable agreement between data and simulation

### • Contamination from QCD multijet events $\sim 6\%$

- After embedding,  $\tau_h$  isolation and  $E_T^{\text{miss}}$  requirement suppress QCD multijet contribution to negligible level



Selected muon  $p_T$  distribution

# EWK+ $t\bar{t}$ genuine $\tau$ background

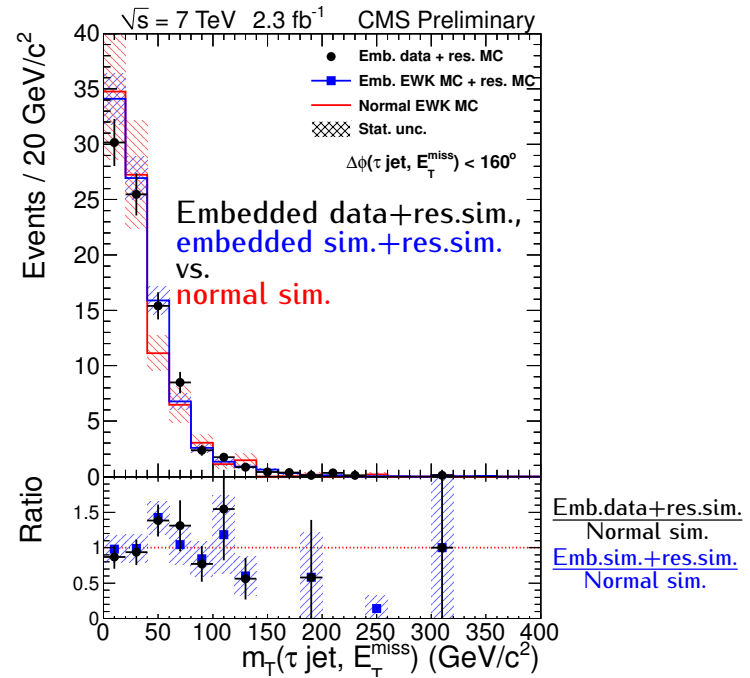
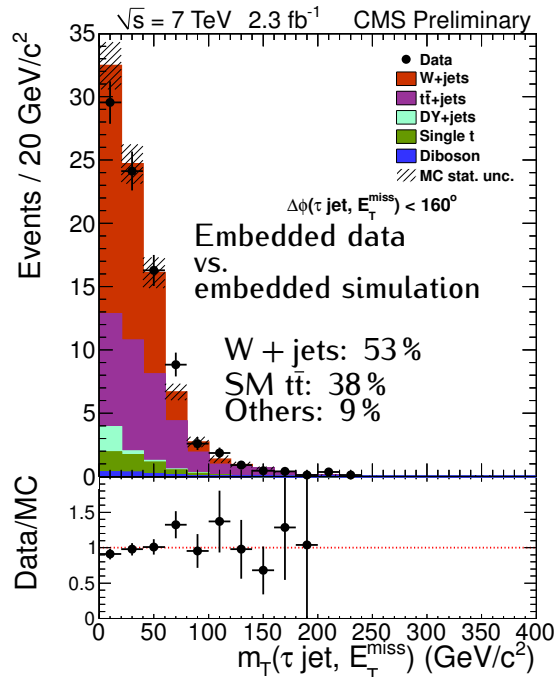
## Validation and uncertainties



- Measurement method was extensively validated by comparing for each selection step
  - Embedded simulation and normal simulation, both without and with accounting for  $\tau$  plus  $E_{\text{T}}^{\text{miss}}$  trigger
  - Embedded data and embedded simulation
  - Both embedded data and embedded simulation plus residual ditau background, and normal simulation
- Dominant uncertainties
  - $\tau$  plus  $E_{\text{T}}^{\text{miss}}$  trigger (11 %)
  - $\tau_{\text{h}}$  energy scale (6.6 %)
  - $\tau_{\text{h}}$  identification (6 %)
  - Statistical uncertainty (3.4 %)

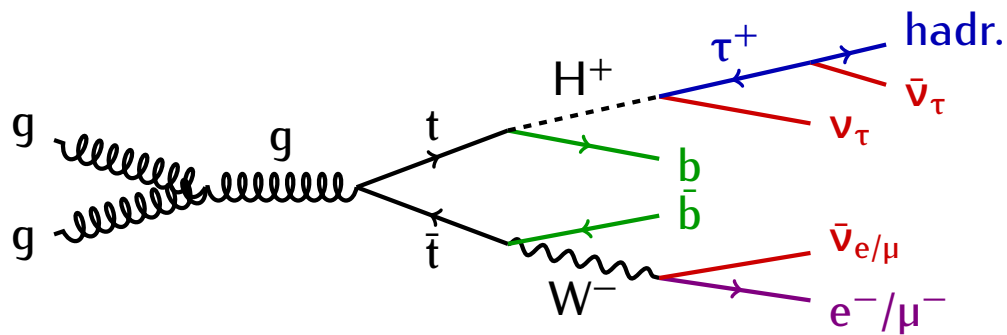
# EWK+ $t\bar{t}$ genuine $\tau$ background

## Results



- Embedded data vs. embedded simulation agree well
- Embedded simulation + residual simulation agree reasonably well with normal simulation
- Result: data:  $78 \pm 3 (\text{stat.}) \pm 11 (\text{syst.})$ ,  
residual ditau from simulation:  $7 \pm 2 (\text{stat.}) \pm 2 (\text{syst.})$

# Reminder: $e+\tau_h$ and $\mu+\tau_h$ final state analyses



$\tau_h$

$E_T^{\text{miss}}$

2 b jets

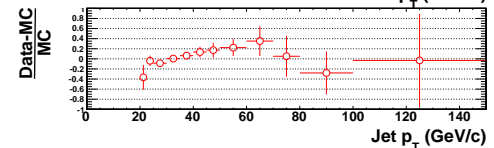
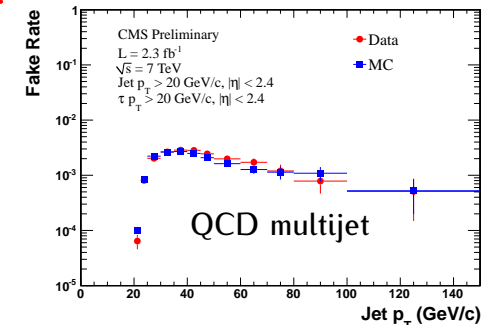
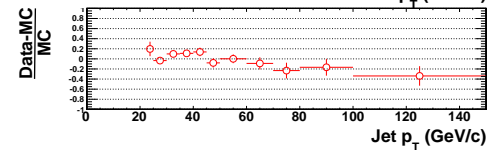
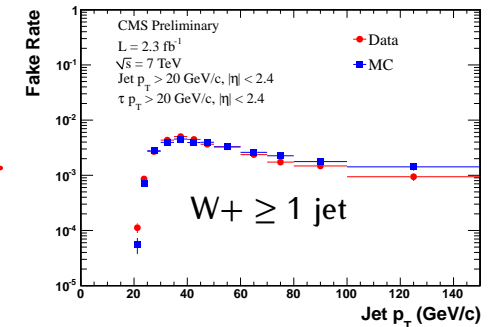
$e/\mu$

1.  $e + 2 \text{ jets} + \text{MHT trigger}$   
Single  $\mu$  trigger
2. Isolated  $e$  with  $p_T > 35 \text{ GeV}/c$ ,  $\eta < 2.5$   
Isolated  $\mu$  with  $p_T > 30 \text{ GeV}/c$ ,  $|\eta| < 2.1$
3.  $\geq 2$  hadronic jets  $p_T > 35(e), 30(\mu) \text{ GeV}/c$
4. Missing  $E_T > 45(e), 40(\mu) \text{ GeV}$
5.  $\geq 1$  jet b-tagged
6.  $\tau_h p_T > 20 \text{ GeV}/c$
7. Opposite-sign (OS) between  $e/\mu$  and  $\tau_h$
8. Counting experiment

# Misidentified $\tau_h$ background measurement

## Misidentification rate

- Background from jets misidentified as  $\tau_h$
- First measure “jet  $\rightarrow \tau$  probability”
- From  $W+ \geq 1$  jet events
  - One isolated  $\mu$  with  $p_T > 20$  GeV/c,  $|\eta| < 2.1$
  - $\geq 1$  jet with  $p_T > 20$  GeV/c,  $|\eta| < 2.4$
  - $m_T(\mu, E_T^{\text{miss}}) > 50$  GeV/c<sup>2</sup>
- From QCD multijet events
  - Single jet trigger ( $p_T > 30$  GeV/c)
  - $\geq 2$  jets with  $p_T > 20$  GeV/c,  $|\eta| < 2.4$
  - All jets except triggering jet used for misidentification rate
    - ★ Except if two jets fire the trigger  $\Rightarrow$  all jets used
- “Jet  $\rightarrow \tau$  probability” parameterized as a function of the jet  $p_T$ ,  $\eta$ , and radius
 
$$(R = \sqrt{\sigma_{\eta\eta}^2 + \sigma_{\phi\phi}^2})$$
  - Using k-Nearest Neighbour (kNN) regression



# Misidentified $\tau_h$ background measurement

## Background estimation

- Select “ $\ell + \geq 3$  jet” events
  - 1 isolated  $e/\mu + E_T^{\text{miss}} + \geq 3$  jets +  $\geq 1$  b-tagged jets
  - Thresholds same as in signal selection
  - Dominated by  $W + \text{jets}$  and  $t\bar{t} \rightarrow \ell + \text{jets}$  events
- Apply to every jet the “jet  $\rightarrow \tau$  probability”
- Subtract a small contribution of genuine  $\tau$  events selected by the requirements above
- Quark and gluon jet composition lies between QCD multijet and  $W + \geq 1$  jet events
  - Take the average of estimates from QCD multijet and  $W + \geq 1$  jet misidentification rates
- Multiply with the efficiency of the opposite-sign requirement ( $\epsilon_{OS}$ )
  - Estimated with simulation, cross-checked with data
- Validated by applying the data-driven method to simulation and comparing with expectation from simulation using generator information


## Results and uncertainties


$e+\tau_h$  final state:

Sample	MC expectation	Estimated from MC	Estimated from data	Residual from MC
QCD multijet	$57.9 \pm 5.1$	54.9	64.1	19.6
W + jets		78.9	86.7	27.4
Average		$66.9 \pm 12.0$	$75.4 \pm 11.3$	$23.5 \pm 3.9$

$\mu+\tau_h$  final state:

Sample	MC expectation	Estimated from MC	Estimated from data	Residual from MC
QCD multijet	$120.1 \pm 8.1$	105.1	113.0	34.4
W + jets		147.3	144.5	44.3
Average		$126.2 \pm 21.1$	$128.8 \pm 15.8$	$39.4 \pm 4.9$

  
 Closure test within uncertainty

  
 Contribution from genuine  $\tau$ 's  
 estimated from simulation  
 Already subtracted  
 from the other numbers

- Final result after multiplication with  $\epsilon_{OS}$

- $e+\tau_h$ :  $54 \pm 6$  (stat.)  $\pm 8$  (syst.)
  - $\mu + \tau_h$ :  $89 \pm 9$  (stat.)  $\pm 11$  (syst.)

- Uncertainties

- Difference in  $\tau_h$  misidentification rates for quark and gluon jets (12%)
    - ★ Use of jet radius decreased the uncertainty from  $\sim 25\%$
  - Number of events for OS efficiency estimate (10%)

- QCD multijet background for  $\tau_h$ +jets final state
  - Normalization and shape of  $m_T$  distribution measured separately
  - Factorization of  $\tau_h$  ID and  $E_T^{\text{miss}} + b\text{-tag} + \Delta\phi(\tau \text{ jet}, E_T^{\text{miss}})$  selections
  - Dominant uncertainties were number of data events, and uncertainties on simulation due to subtraction of EWK+ $t\bar{t}$  events
- EWK+ $t\bar{t}$  genuine  $\tau$  background for  $\tau_h$ +jets final state
  - $\mu + \geq 3$  jets events and tau embedding method
  - Normalization: correct for various efficiencies
  - Dominant uncertainties were  $\tau$  plus  $E_T^{\text{miss}}$  trigger efficiency,  $\tau_h$  energy scale, and  $\tau_h$  identification
- Misidentified  $\tau_h$  for  $e + \tau_h$  and  $\mu + \tau_h$  final states
  - $e/\mu + \geq 3$  jets events and  $\text{jet} \rightarrow \tau_h$  misidentification rate
  - Dominant uncertainties were different  $\tau_h$  misidentification rates for quark and gluon jets, and statistics for OS efficiency estimate