Data/MC comparisons, uncertainties on FR method and miscellanea

Huiling Hua [1] **Fabio Iemmi** [1]
Hongbo Liao [1] Hideki Okawa [2] Yu Zhang [2]

[1]Institute of High Energy Physics (IHEP), Beijing

[2]Fudan University, Shanghai

August 25, 2021

Data/MC comparisons for variables used in BDT

- **Compare data and MC** to see if variables are well modeled by simulation
- **Stacked histogram** with sum of all **MC processes**
  - Signal is added to $t\bar{t}$+X in this histogram
- **Signal** is also reported as a **red, dashed line, scaled** by a multiplicative factor to make it visible
- Apply scale factors that we discussed so far
  - PU
  - Prefiring
  - Trigger
  - b tagging
- Reliable cross sections for single Higgs processes impossible to find, computed them by hand...
- Plots should (more or less) comply with the CMS publication guidelines

1tau1L

Data events: 1633

signal events: 6.68558

ttbar events: 1628.08

QCD events: 2.14882

tt+X events: 75.8976

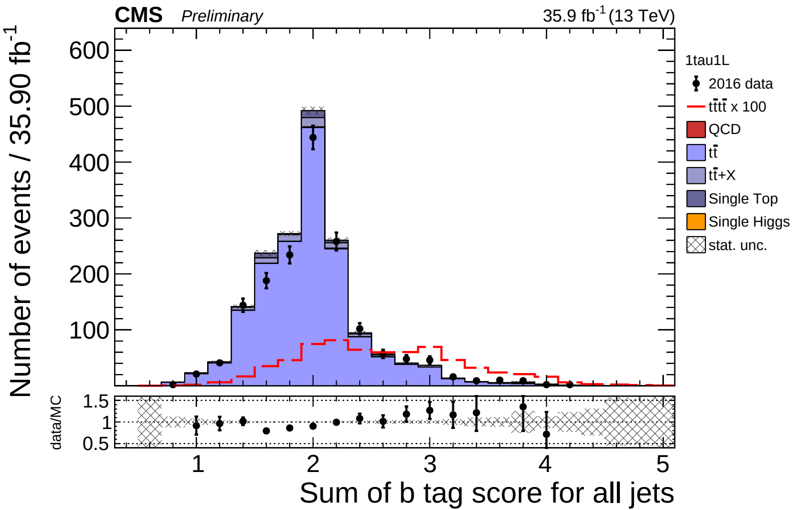
single top events: 32.8338

single Higgs events: 0.0304602

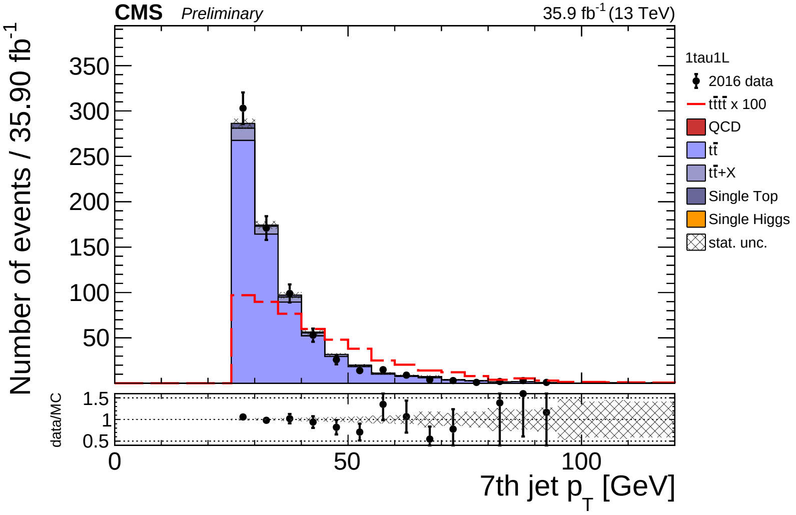
total MC events: 1738.99

data/MC agreement: -6.0947%

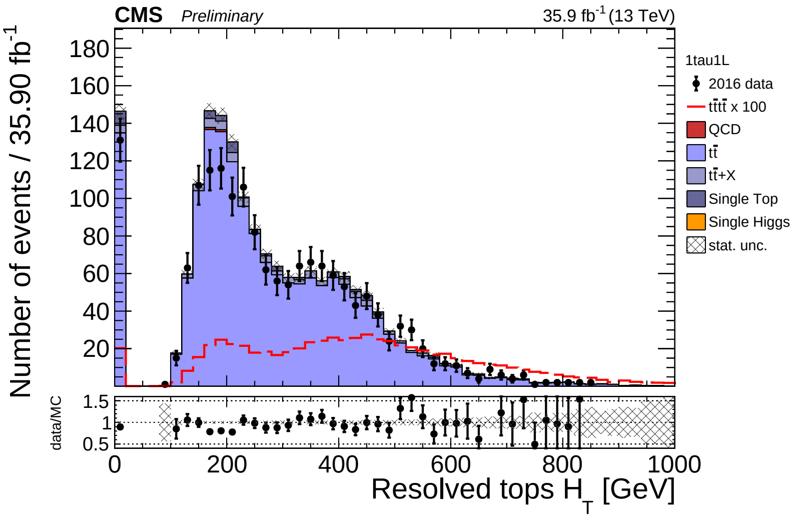
Data/MC agreement



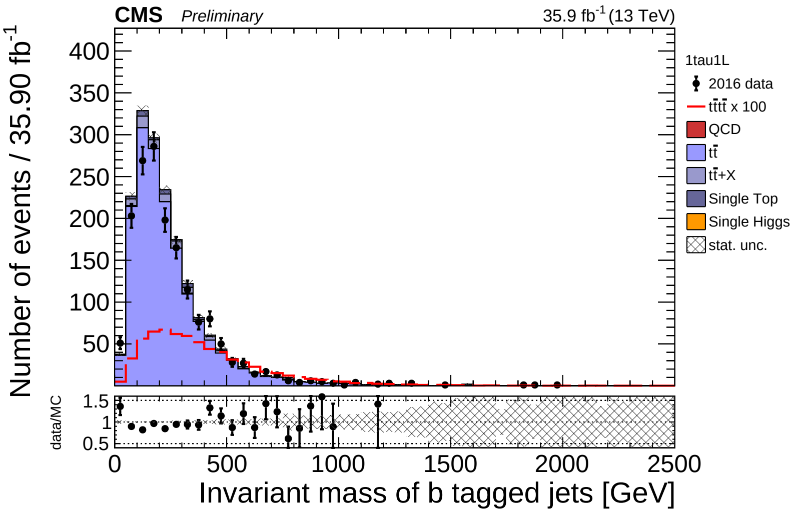
Data/MC agreement



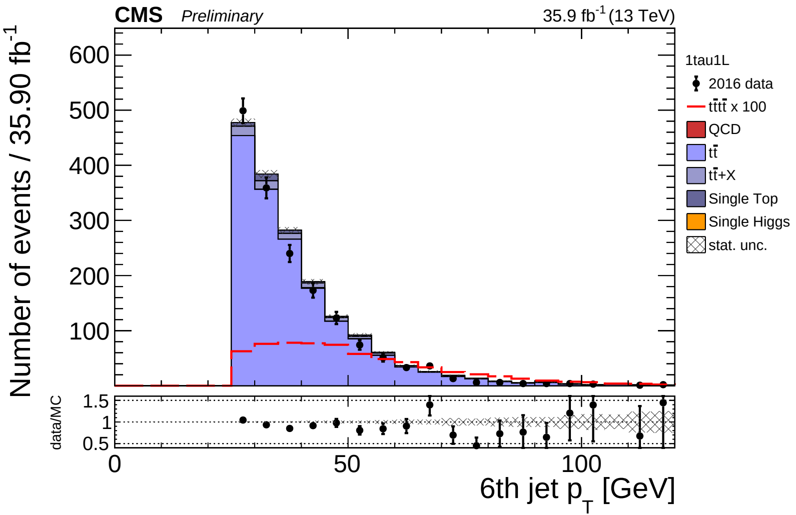
Data/MC agreement



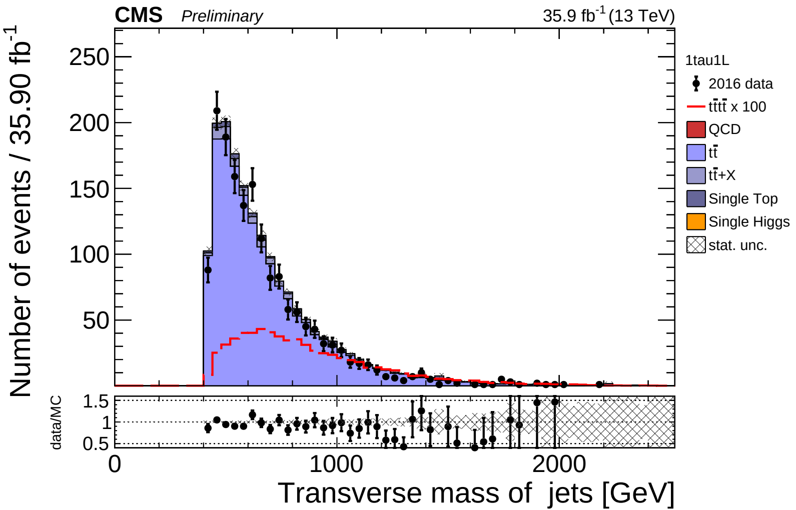
Data/MC agreement



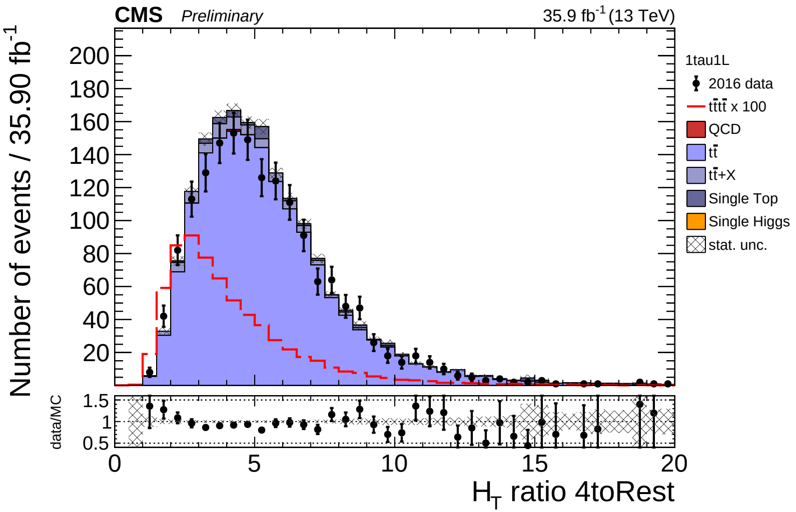
Data/MC agreement



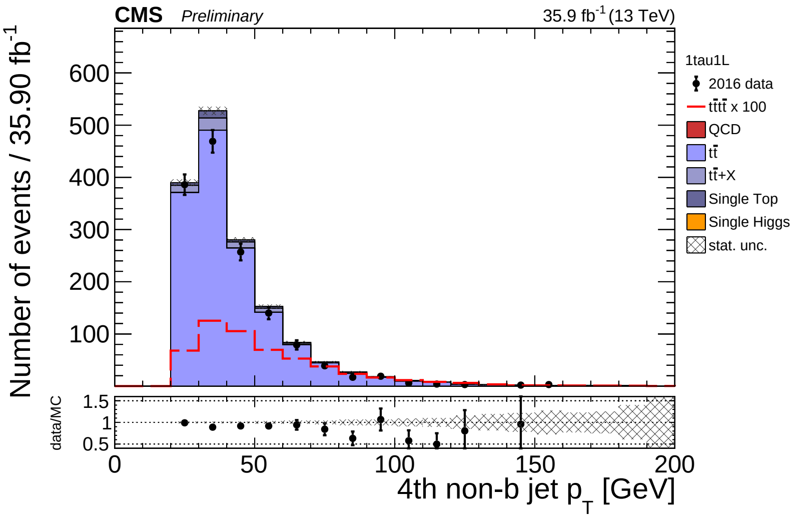
Data/MC agreement



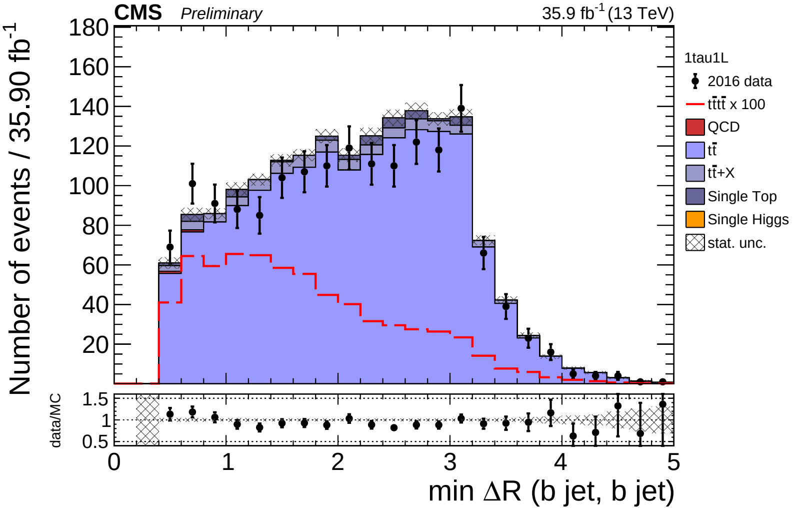
Data/MC agreement



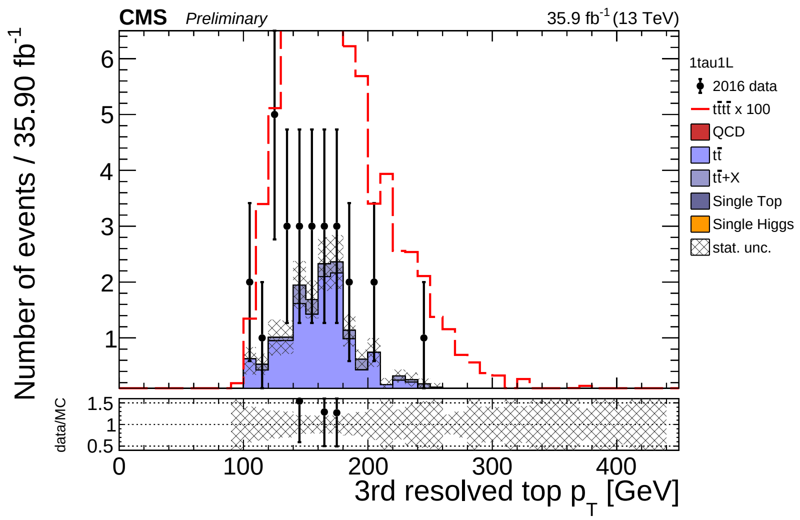
Data/MC agreement



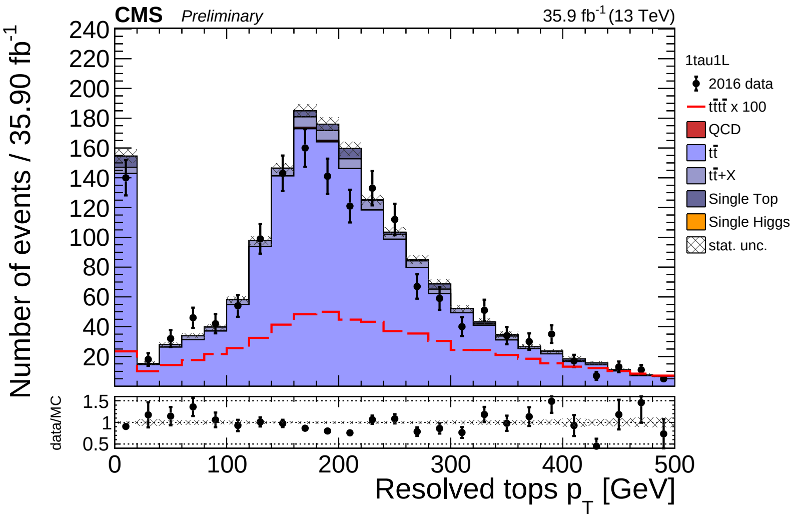
Data/MC agreement



Data/MC agreement



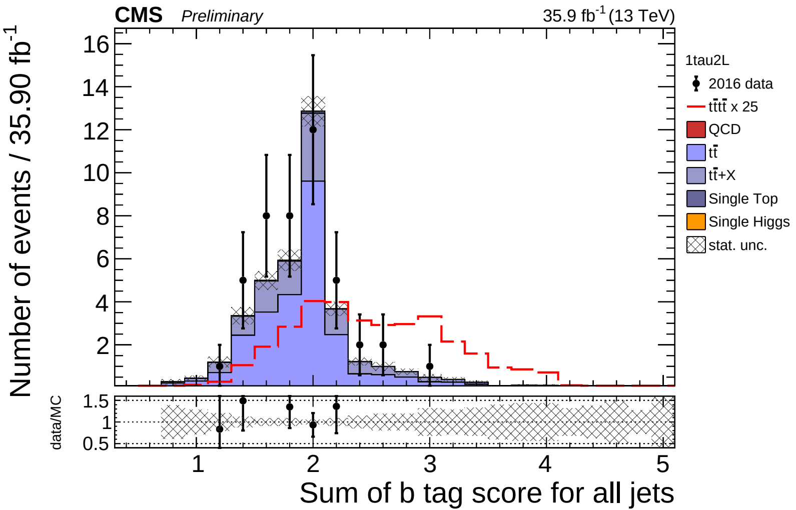
Data/MC agreement



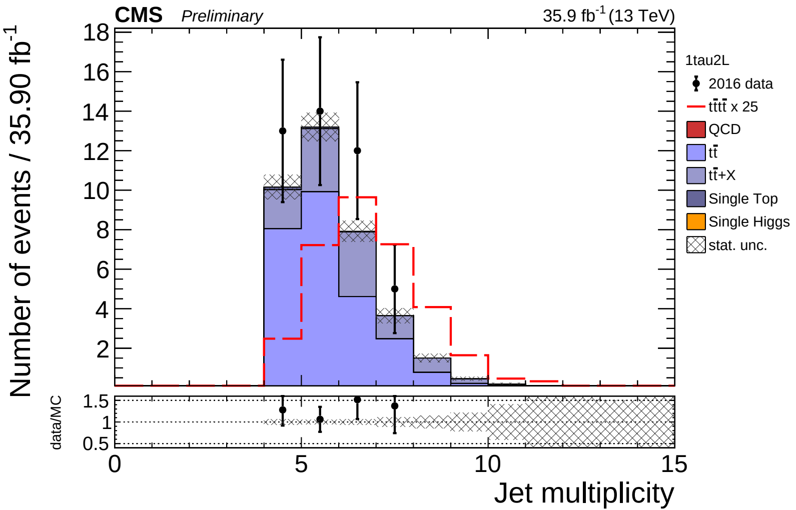
1tau2L

Data events: 44
signal events: 1.32969
ttbar events: 26.2683
QCD events: 0
tt+X events: 10.6613
single top events: 0.23201
single Higgs events: 0.000111213
total MC events: 37.1617
data/MC agreement: 18.4015%

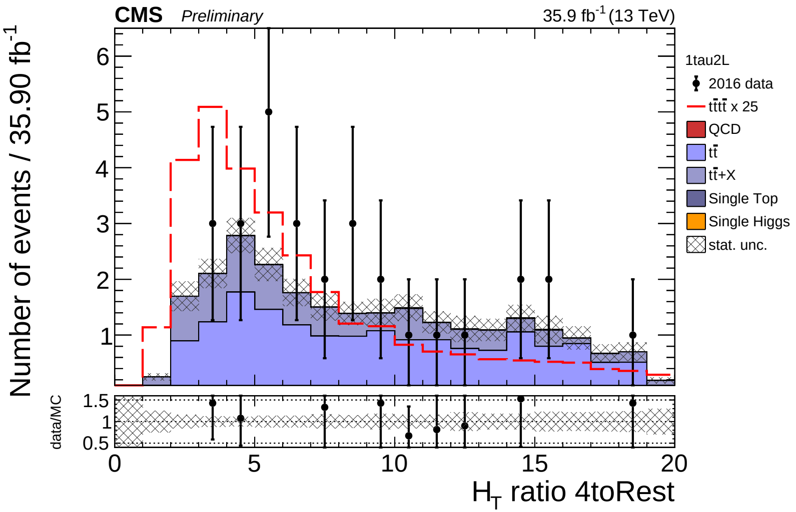
Data/MC agreement



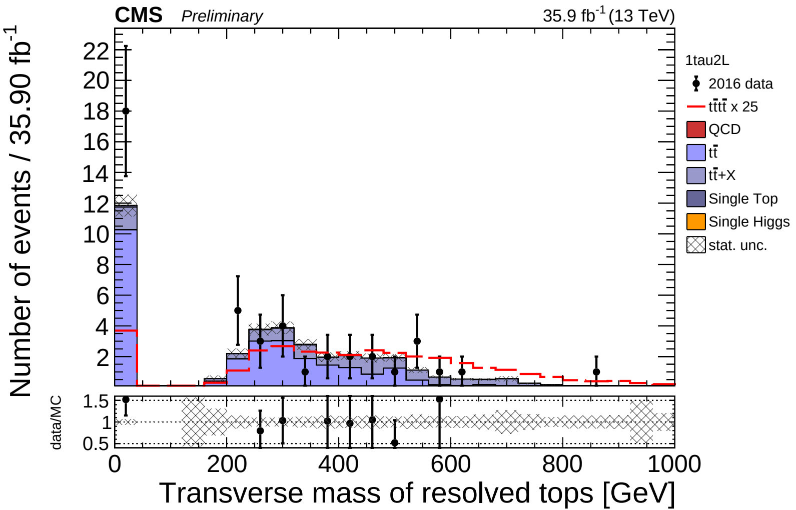
Data/MC agreement



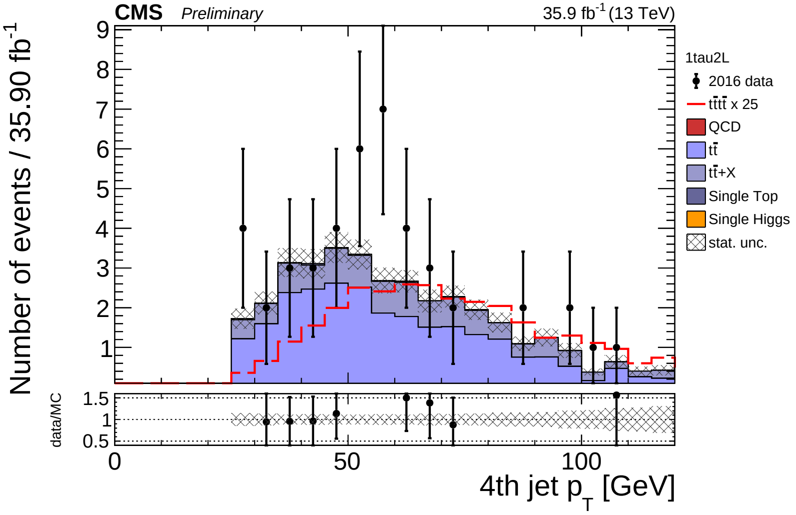
Data/MC agreement



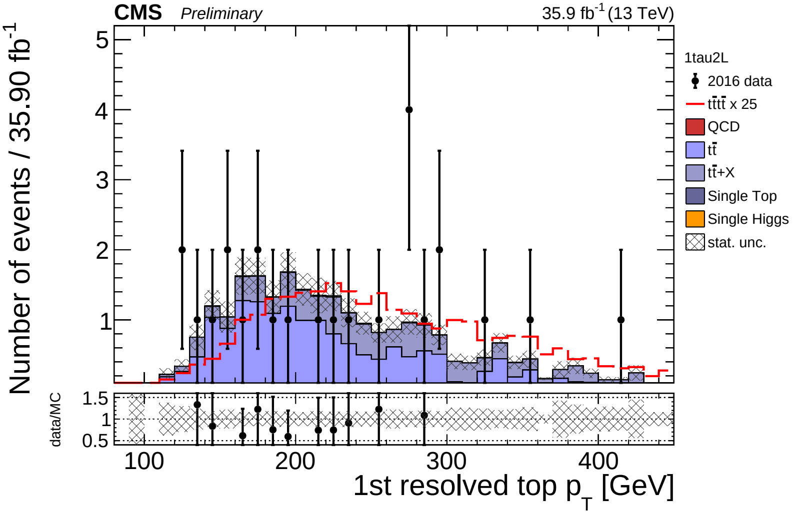
Data/MC agreement



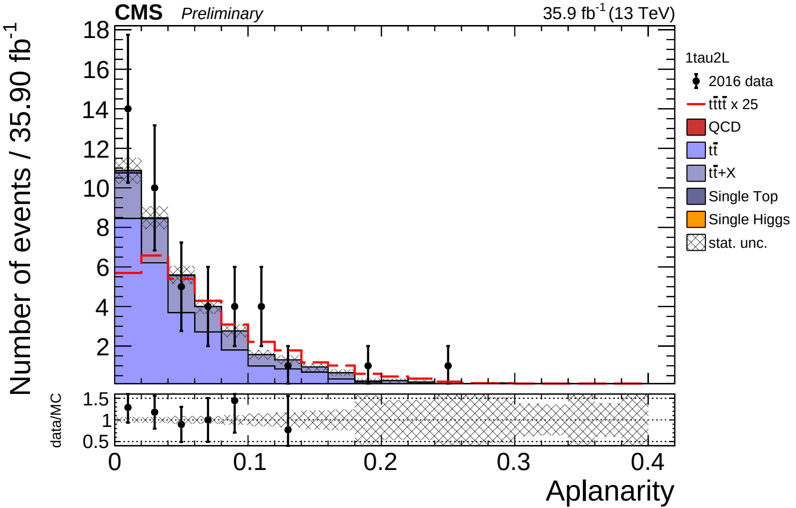
Data/MC agreement



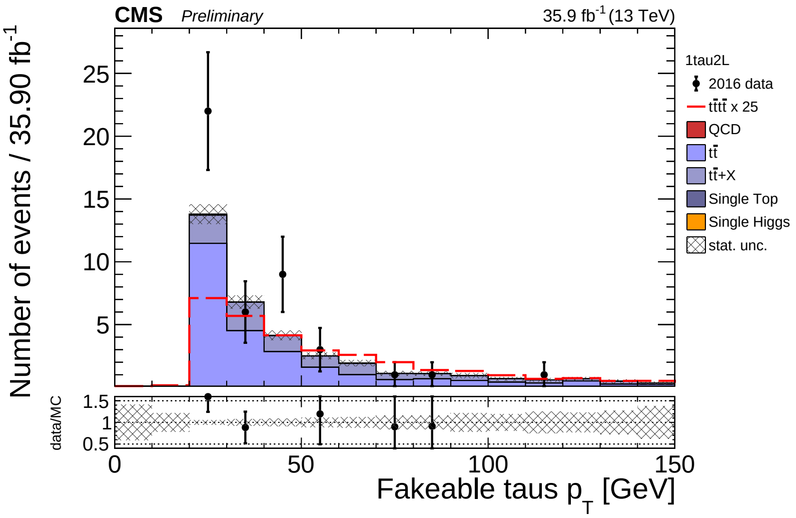
Data/MC agreement



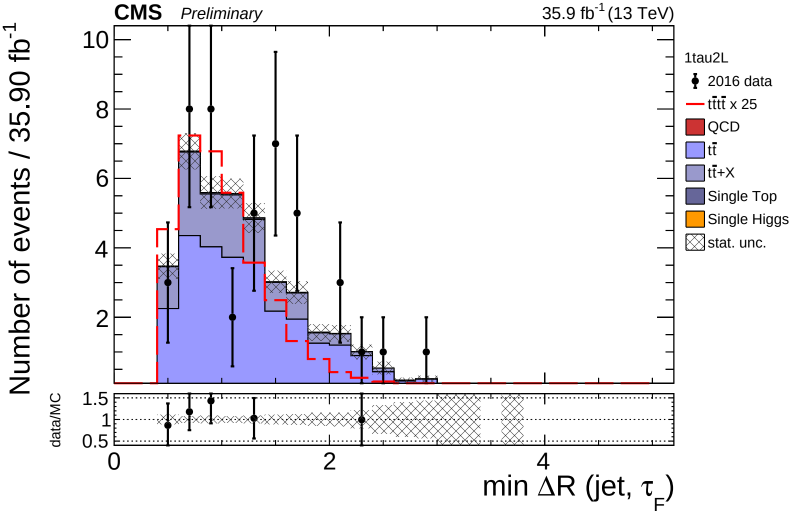
Data/MC agreement



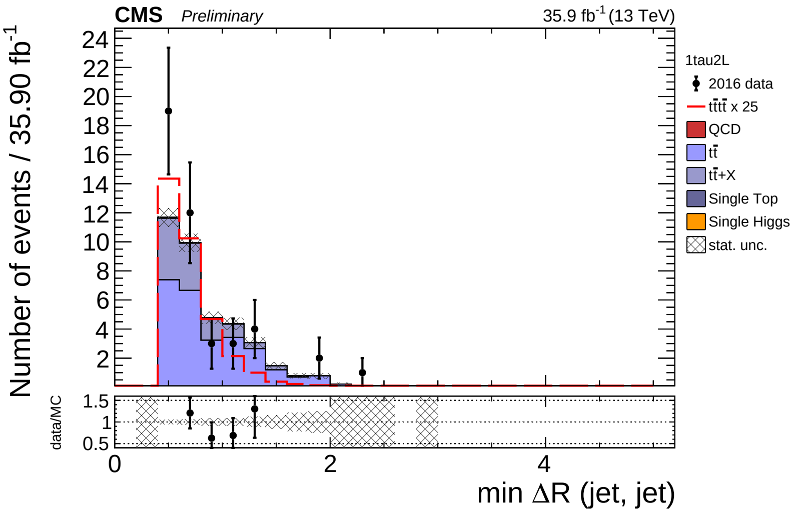
Data/MC agreement



Data/MC agreement



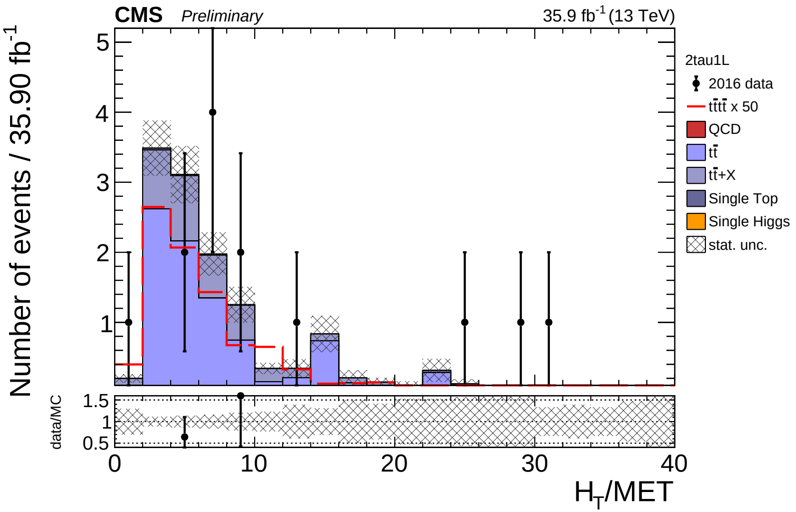
Data/MC agreement



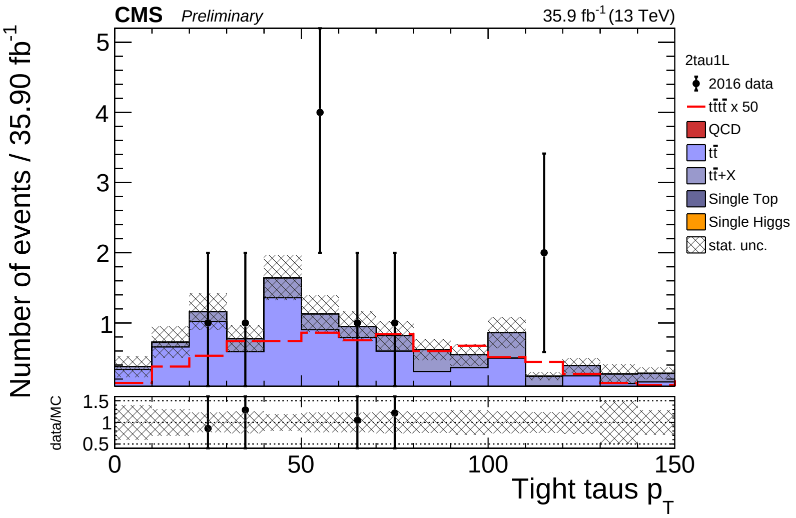
2tau1L

Data events: 13
signal events: 0.180122
ttbar events: 8.93833
QCD events: 0
tt+X events: 3.79847
single top events: 0.07949
single Higgs events: 0
total MC events: 12.8163
data/MC agreement: 1.43343%

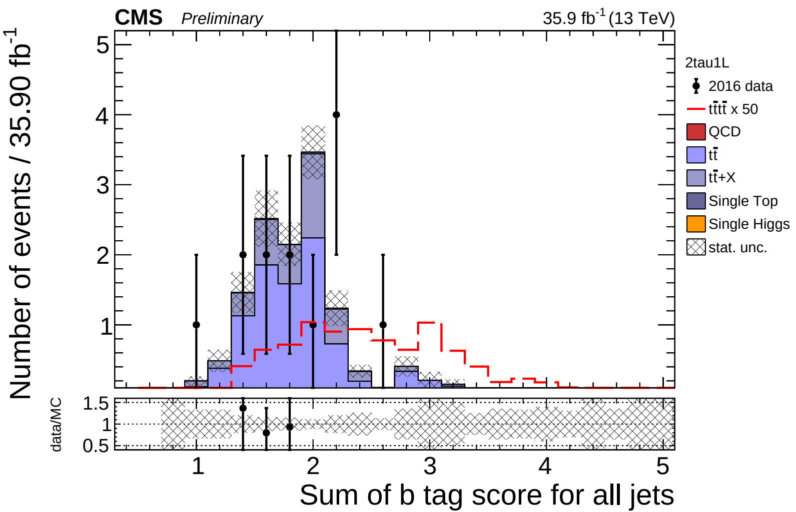
Data/MC agreement



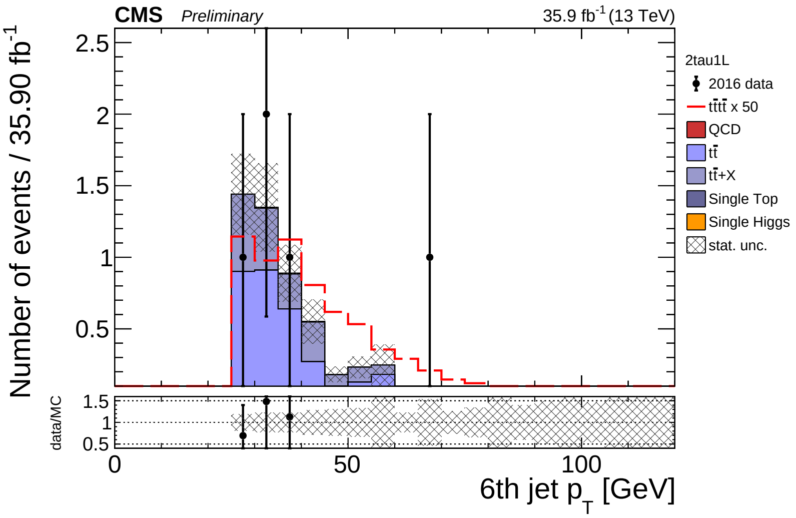
Data/MC agreement



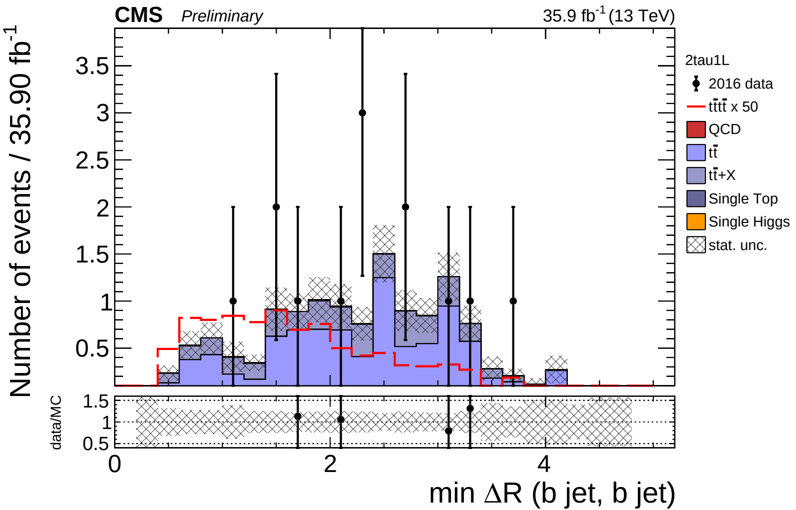
Data/MC agreement



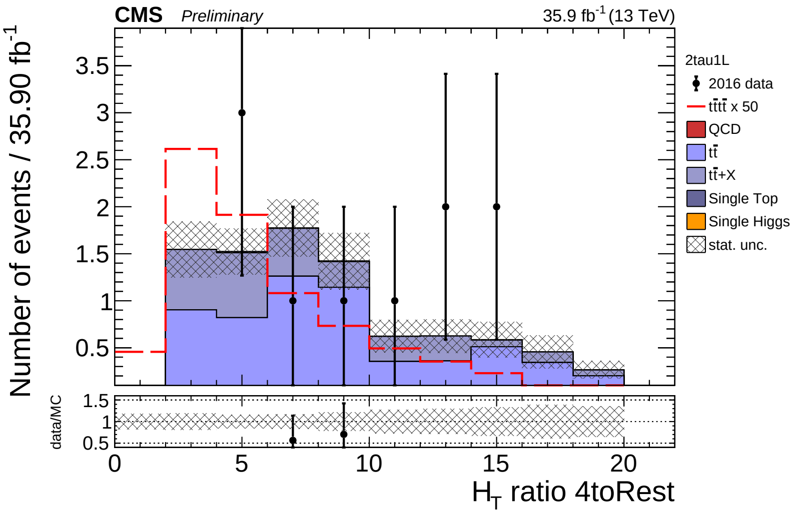
Data/MC agreement



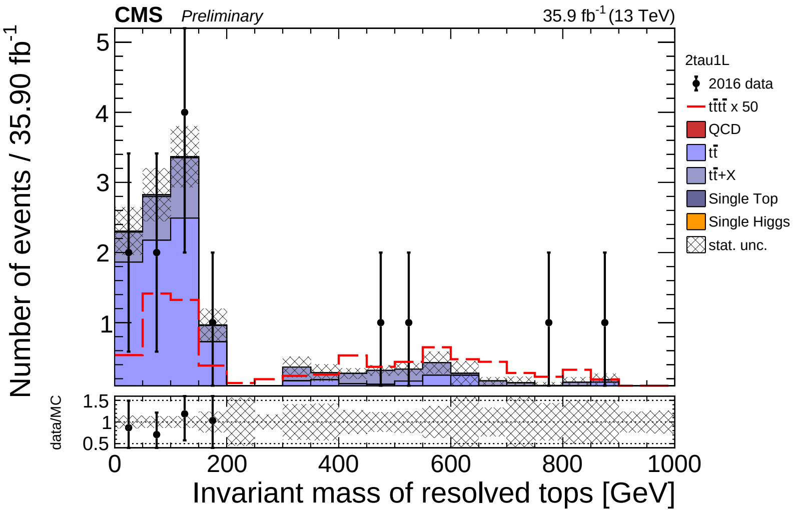
Data/MC agreement



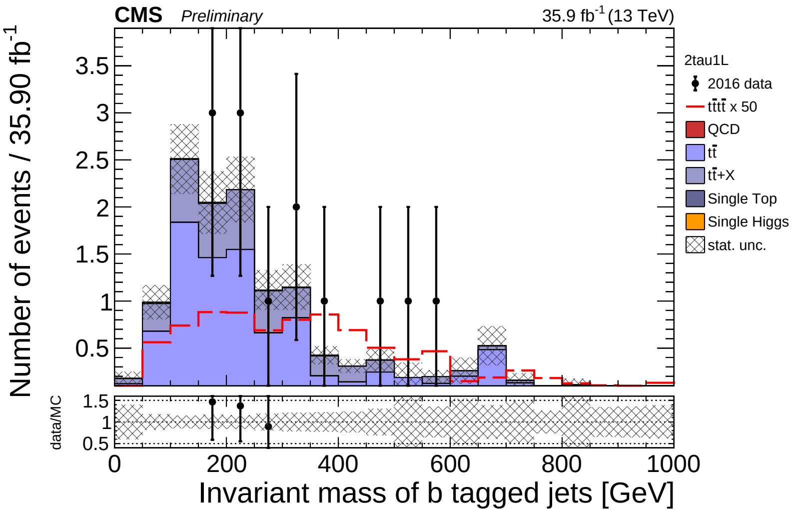
Data/MC agreement



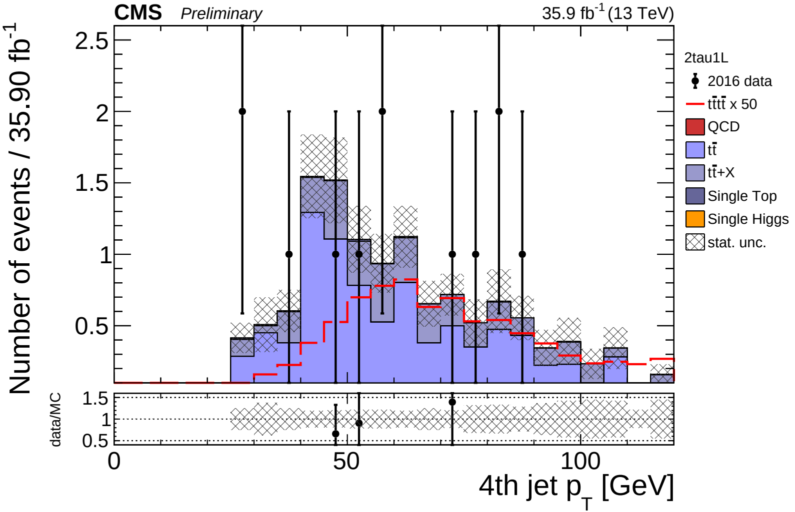
Data/MC agreement



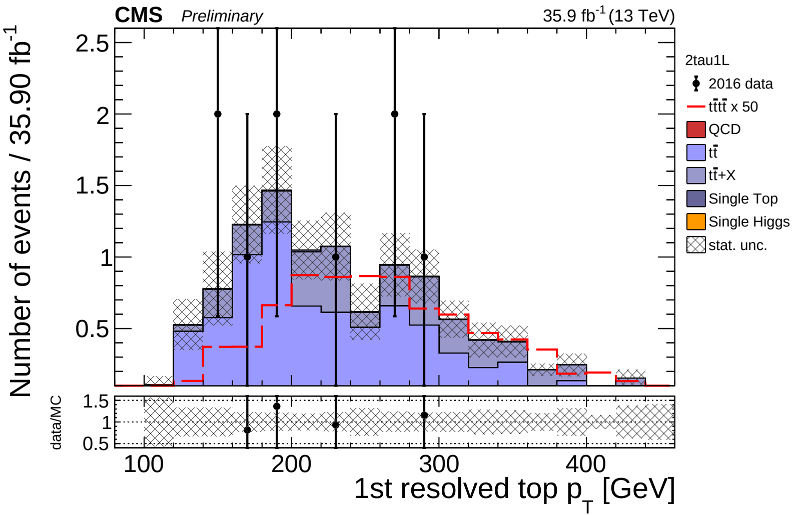
Data/MC agreement



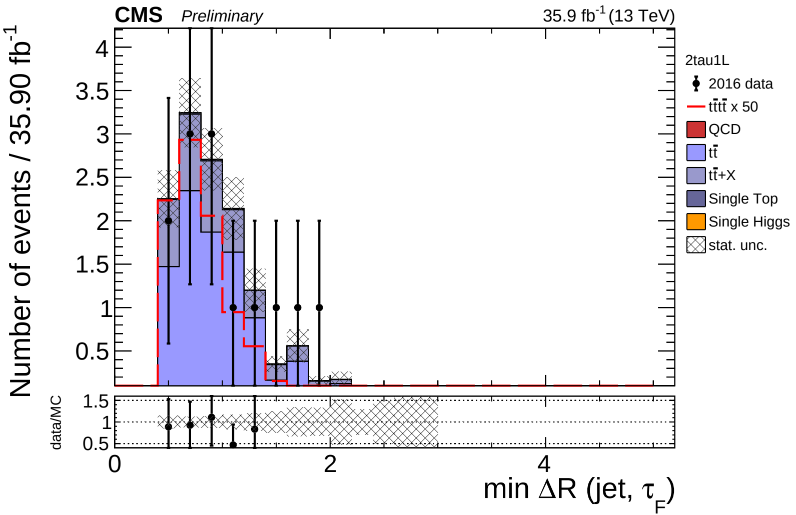
Data/MC agreement



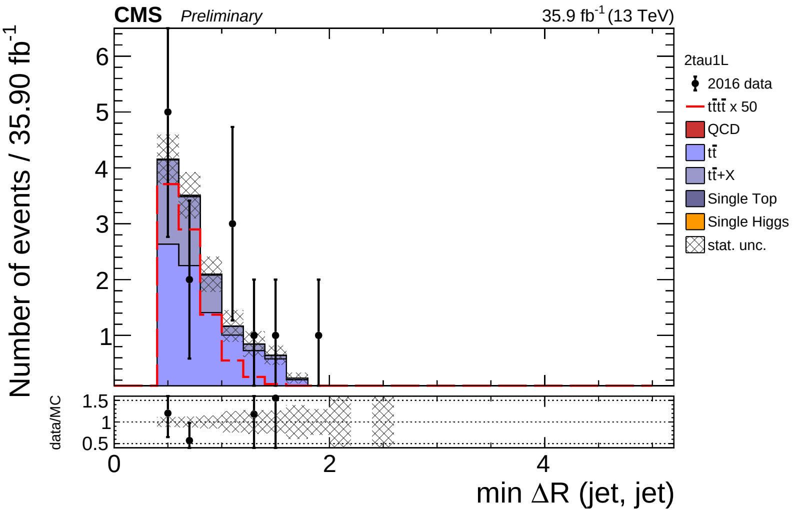
Data/MC agreement



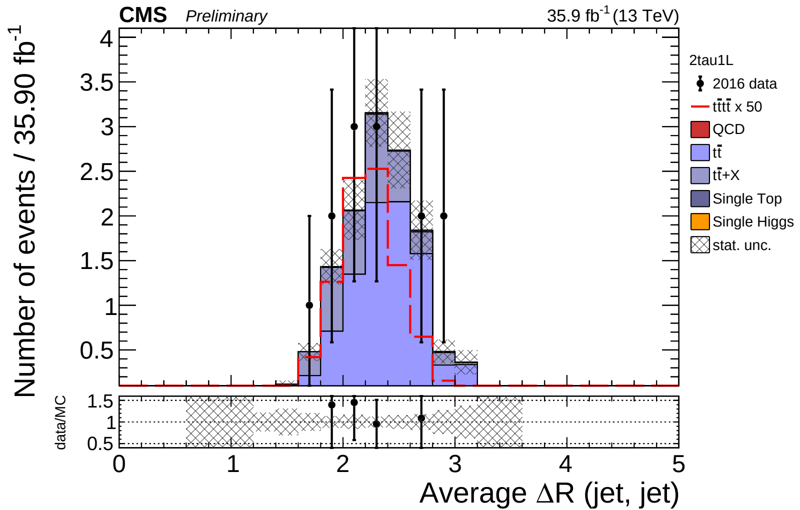
Data/MC agreement



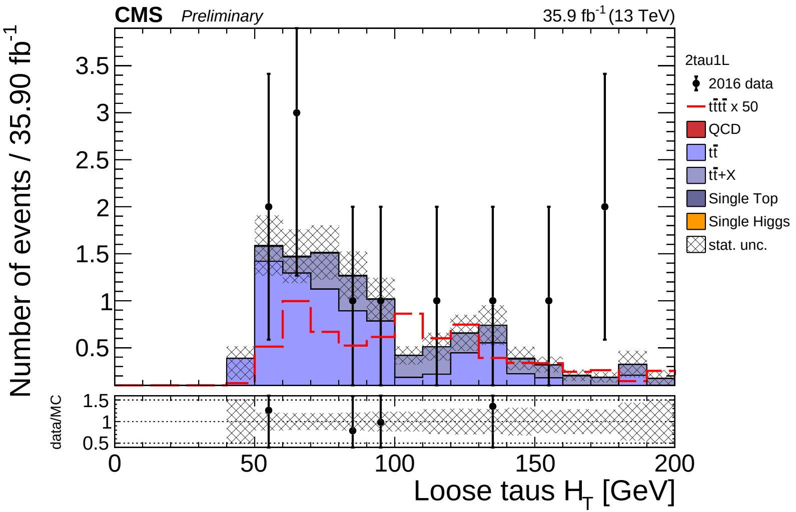
Data/MC agreement



Data/MC agreement



Data/MC agreement



1tau0L

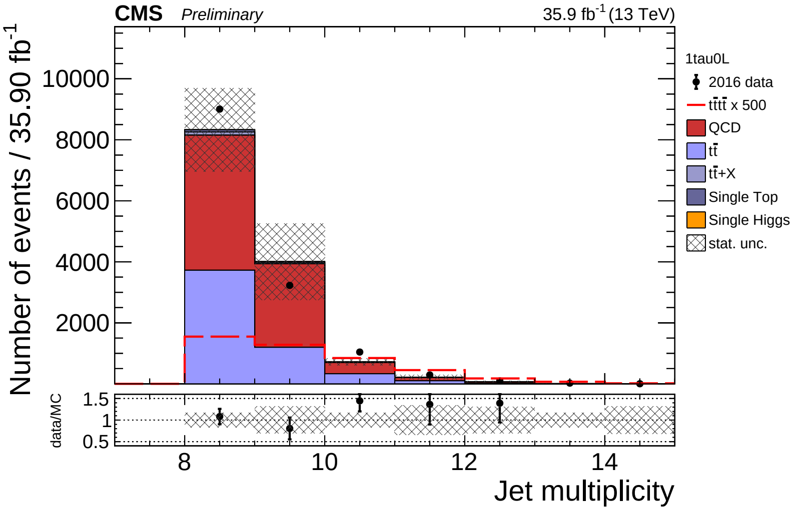
Data events: 13693
signal events: 8.78554
ttbar events: 5389.6
QCD events: 7679
tt+X events: 171.034
single top events: 111.117
single Higgs events: -0.292551
total MC events: 13350.5
data/MC agreement: 2.56573%

Some remarks about 1tau0L

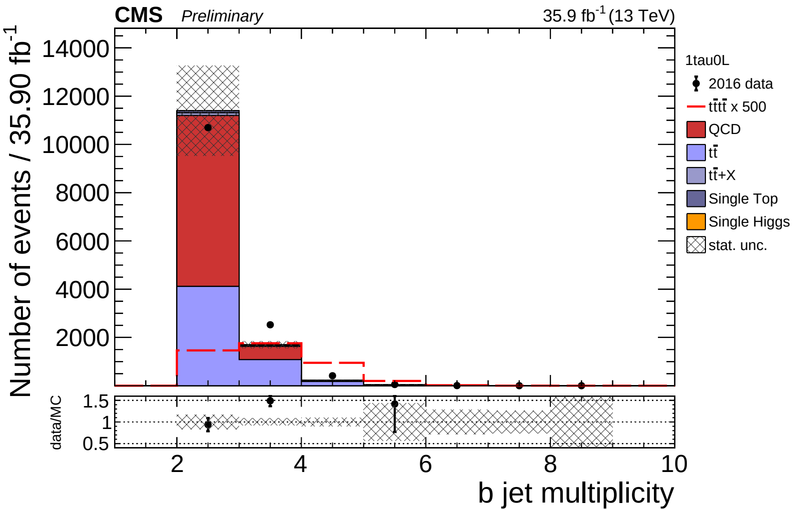
- As we already know, **1tau0L** category is dominated by **QCD background**
- I recently developed a method to **estimate the QCD yield** in 1tau0L completely from data: **FR method**
- In the following, I am **scaling the QCD shape obtained from MC to the FR yield**
- Interestingly, using the FR yield **enhances the data/MC agreement**:

	MC QCD yield	FR QCD yield
data/MC	12.1%	2.6%

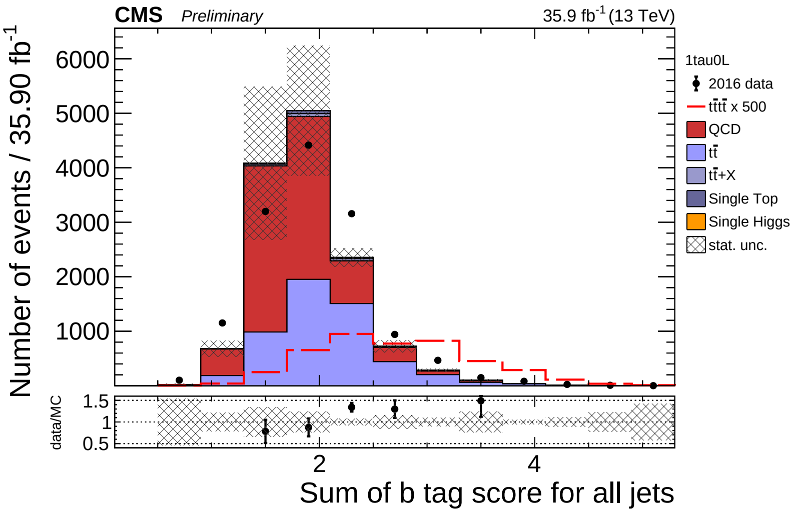
Data/MC agreement



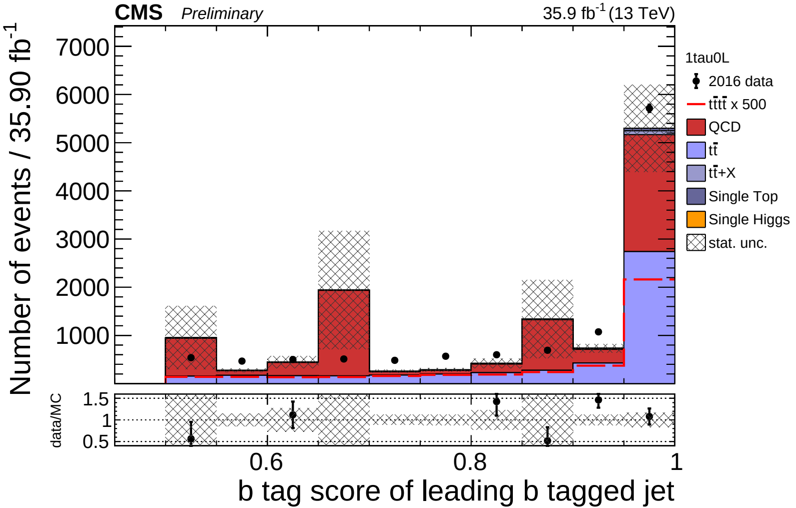
Data/MC agreement



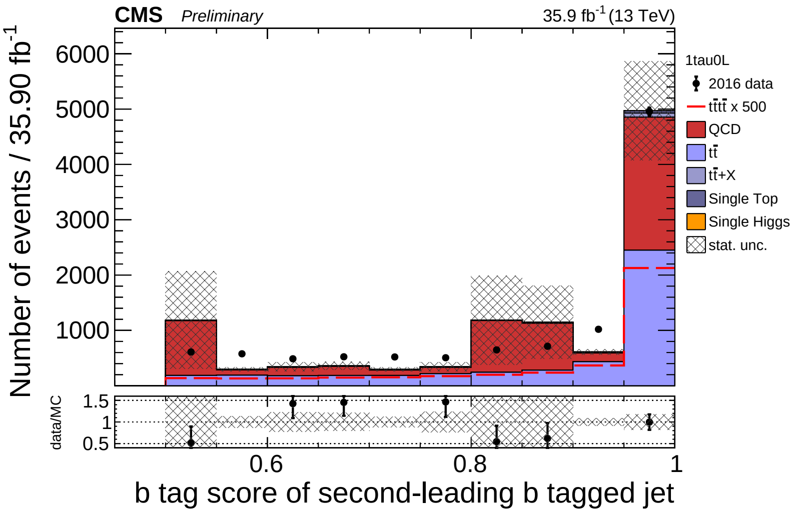
Data/MC agreement



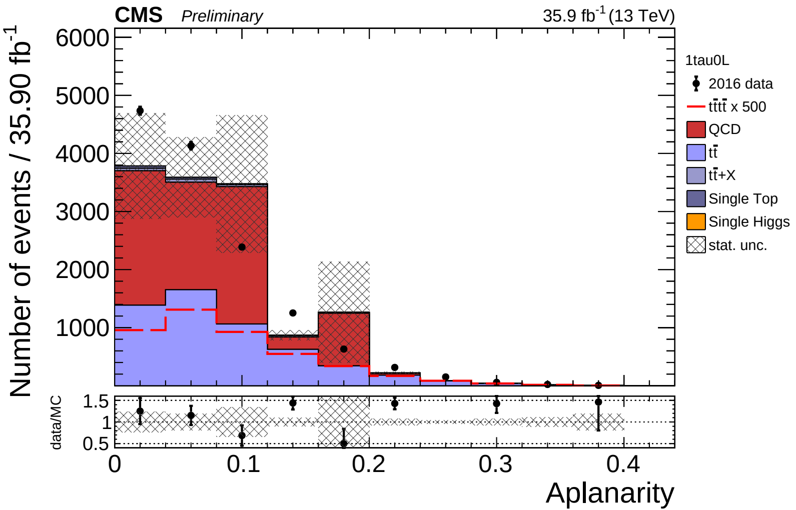
Data/MC agreement



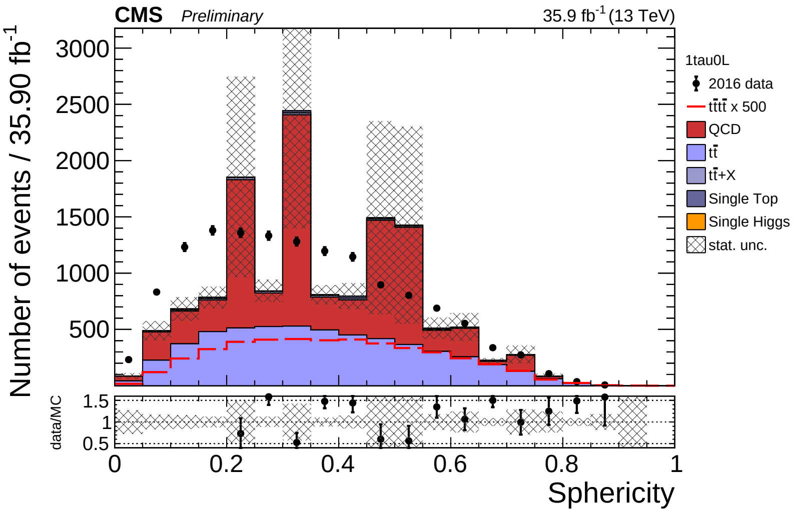
Data/MC agreement



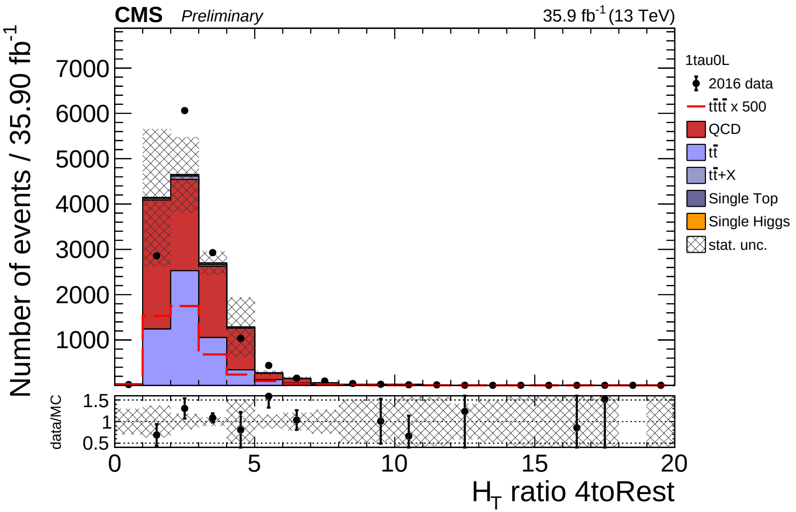
Data/MC agreement



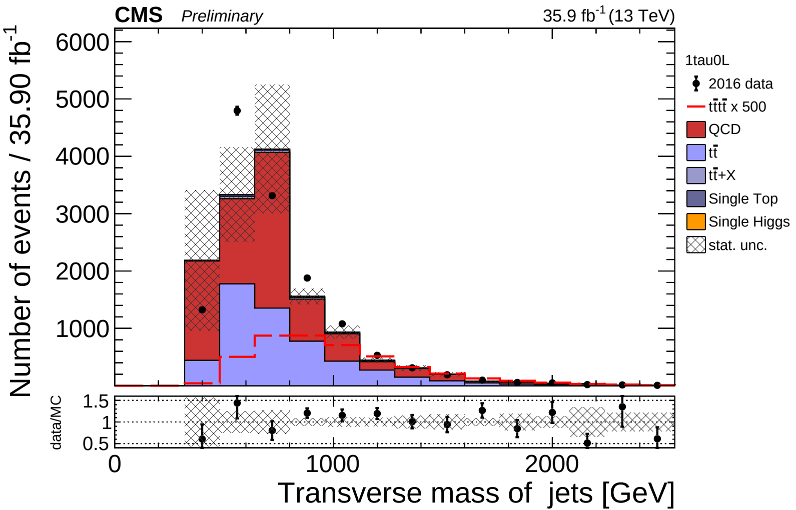
Data/MC agreement



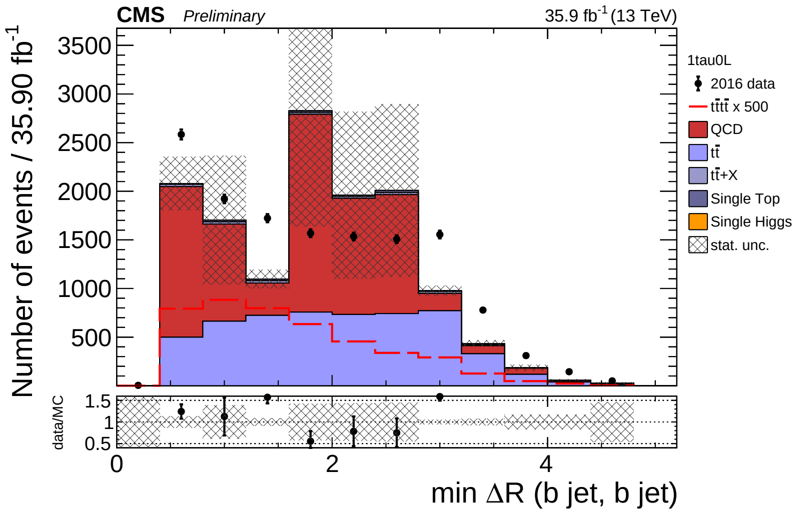
Data/MC agreement



Data/MC agreement



Data/MC agreement



Remarks on data/MC agreement in 1tau0L

- I believe it's nice that the **FR method gives an enhanced agreement in data/MC comparison**
- The simulated QCD shape is giving problems though
  - Spikes caused by a few events with high cross section passing the selection (already observed by Huiling)
- I found out that most of the **spikes are caused by `QCD_HT300to500` sample**
- Could it be worth to increase our HT cut (currently > 400 GeV) to > 500 GeV to rule this sample out from our analysis?
  - Did not try this though...

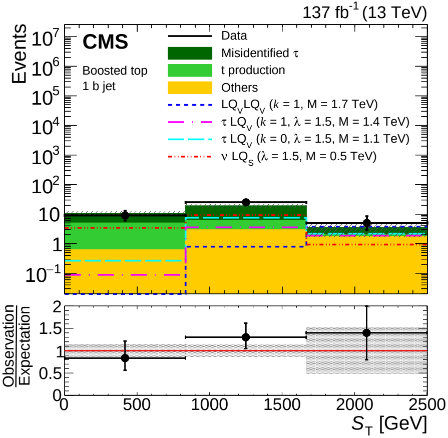
Remarks on BDT variables

- I believe we have a problem with input BDT variables
- Currently, we are **using variables that may be undefined** in a given category
  - For example: 7th jet p_T in 1tau1L
  - We require $N_{jets} \geq 6$ in 1tau1L...
- When a **variable is not defined**, we assign a **ground value of -99**
- This can **introduce fake correlations between variables**
- For example, 7th p_T gets artificially correlated with N_{jets}
- If number of jets is low (6) we assign -99 to 7th p_T so 7th p_T goes lower...
- ...artificial positive correlation between the two
- **We choose variables to use based on their correlation!**
- **Don't think this is safe**

Uncertainties on FR method

Uncertainties on FR method

- I read what they do in EXO-19-015
- Their idea is to perform **validation of the FR method in a region with similar background composition as the signal region**
- Validation is a **data/MC agreement** check on the variable they are going to use in final fit
- I developed the setup for data/MC validation, so tried to do something similar



Definition of the validation region

- As a **reminder**: we **compute fake rates in** the so-called **control region** (CR): same requirements as SR, but no b tagged jets
- I defined the **validation region (VR)** to be both close to CR and SR: same definition of SR but **exactly 1 b tagged jet**
- Orthogonal to both CR and SR
- Being orthogonal to SR, we can look at data here (not blinded)

	N_{τ_h}	N_ℓ	N_{jets}	N_{bjets}
CR	1	0	≥ 8	0
VR	1	0	≥ 8	1
SR	1	0	≥ 8	≥ 2

Definition of the validation region

- The **VR has similar background composition as the SR**: lot's of QCD, non-negligible $t\bar{t}$, some $t\bar{t}$+X

	$t\bar{t}t\bar{t}$	$t\bar{t}$	QCD	$t\bar{t}$+X
CR	0.09	287.46	6051.20	8.17
VR	0.98	2321.43	7792.01	78.91
SR	8.79	5389.60	6539.06	162.25

- It looks fine to perform validation in this region
- **Compute the QCD yield expected by the FR method in the VR**

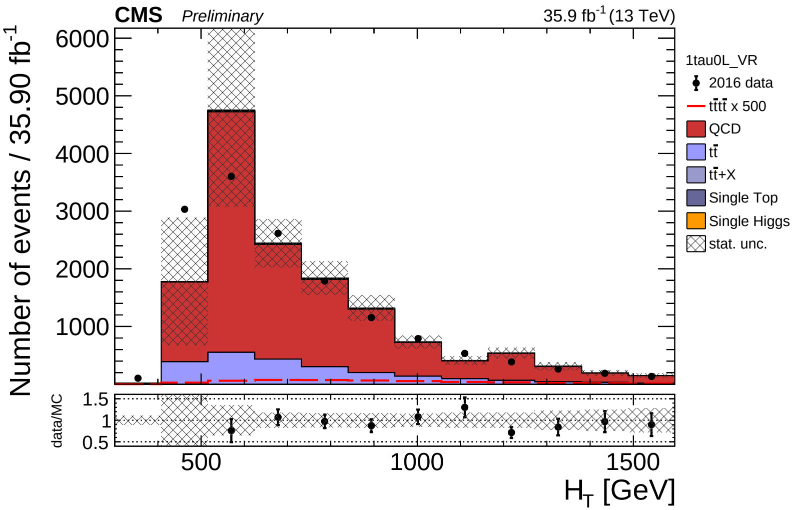
	MC QCD yield	FR QCD yield
exp. yield	7792	12392

Validation of the FR method

- Assumed we are going to fit H_T distribution in this category
  - We don't have a BDT here, at least for now
- Perform **data/MC agreement for H_T** distribution in the VR
- **Scale the MC QCD shape to yield coming from FR method**
- Interestingly, using the FR yield **enhances the data/MC agreement**:

	MC QCD yield	FR QCD yield
data/MC	45%	0.2%

Validation of the FR method



Remarks on validation procedure

- Still not sure which variable we are going to fit, but **this could be the general procedure** to follow
- Based on previous slide agreement, **we should assess the uncertainty on this method**
- I propose to assign **two uncertainties** in the datacard
  - **One** log-normal unc. of $\approx 4\%$ **for the statistical uncertainty** on the yield
  - **One** log-normal unc. of some value **for the above level of agreement**
- MC QCD spikes make it hard to decide the level of agreement
- Binning in EXO-19-015 is pretty coarse, maybe I could do the same (don't like much the idea)
- I could try to get the shape of QCD from data as well
  - We could get way more statistics than the simulation

Miscellanea

Why CWoLa won't work for us

- I read the paper about Classification without labels (CWoLa)
- With **CWoLa**, you can train a classifier entirely from data, which **helps** when you have to deal **with simulation with poor description of the data and low statistics** (as our QCD)
- Unfortunately, it relies on the definition of two data regions with two **conditions that we do not fulfill**:
  1. Your data regions must containt just two processes: signal and background
  2. Your data regions must have different proportions of signal and background
- Concerning 1), we have at least three processes: $t\bar{t}t\bar{t}$, $t\bar{t}$ and QCD
- Concerning 2), $t\bar{t}t\bar{t}$ is very rare, so it's impossible to get very different proportions
- **Sadly, I'm afraid we have to drop this**

Simulated samples

- While looking at the simulated samples I realized that:
  1. Some minor single-Higgs processes are missing (e.g., ggH(ZZ->4l))
  2. **We are using** a mix of top-related processes with **different tunes**
- Concerning 1), shouldn't be a big problem, we can always ntuplize them later
- Concerning 2), it could be a problem when estimating systematic uncertainties
- But we are sooner or later switching to UL, right? There, all the tunes should be the same