CMS Draft Analysis Note

The content of this note is intended for CMS internal use and distribution only

2015/02/11
Head Id:
Archive Id: Unversioned directory
Archive Date:

Internal Working Note
IISER-Rutgers 2015 Multilepton Analysis Preparation

Christian Contreras-Campaña[1], Emmanuel Contreras-Campaña[1], Sourabh Dube[2], Maximilian Heindl[1], Conan Huang[1], Anshul Kapoor[2], Sunil Somalwar[1], Scott Thomas[1], Peter Thomassen[1], and Matt Walker[1]

[1] Rutgers University
[2] IISER

DRAFT

Abstract

In preparation of the CMS 2015 analysis program, we outline the IISER-Rutgers multilepton analysis approach and list the background methods we are planning to use.

This box is only visible in draft mode. Please make sure the values below make sense.

PDFAuthor: Peter Thomassen
PDFTitle: Internal Note: IISER-Rutgers 2015 Multilepton Analysis Preparation
PDFSubject: CMS SUSY Analyses
PDFKeywords: SUSY

Please also verify that the abstract does not use any user defined symbols

DRAFT

Change log

The change log for this note is kept in the git repository at:
https://gitlab.com/Thomassen/RutgersMultileptonDocs/commits/master/AN-2015multilepton

Contents

1 Introduction . . . 6
2 Data sets and MC Samples . . . 6
2.1 Data sets . . . 6
2.2 Background MC samples . . . 6
2.3 WZ MC details . . . 7
2.4 ttbar MC details . . . 7
3 Analysis Workflow . . . 7
4 Common MC treatment (signal + background) . . . 8
4.1 Pile-up weights . . . 8
4.2 Lepton ID/ISO weights . . . 8
4.3 Trigger weights . . . 9
4.4 $E_{\mathrm{T}}^{\mathrm{miss}}$ resolution correction using Z+j data/MC . . . 9
5 MC background . . . 9
5.1 WZ background . . . 9
5.2 ZZ background . . . 12
5.3 $\mathrm{t\bar{t}}$ background . . . 14
6 Data-driven backgrounds . . . 17
6.1 Light lepton fakes . . . 17
6.2 Fake leptons from asymmetric internal photon conversions (AIC) . . . 18
6.3 Fake leptons from jets . . . 19
6.4 Tau fakes . . . 21
7 Results . . . 24
7.1 3 leptons . . . 24
7.2 4 or more leptons . . . 34
8 Interpretation and Statistical Treatment . . . 44
A To do . . . 45
A.1 Physics . . . 45
A.2 Infrastructure . . . 45
B Miscellaneous Studies . . . 45
B.1 AIC region with DY sample . . . 45
B.2 Z binning . . . 45
B.3 Suitability of tracks as fake proxies . . . 47

DRAFT

Tables

1 Analysis workflow 7

DRAFT

Figures

1 n_{jets} distribution in the WZ-dominated $E_{\text{T}}^{\text{miss}}$ control region, before and after n_{jets} weights . . . 10

2 $p_{\text{T}}(j_{\text{lead}})$ distribution in the WZ-dominated $E_{\text{T}}^{\text{miss}}$ control region (no H_{T} cut). left) without weights based on this variable right) with these weights (all bins above 150 GeV combined into one bin with flat weight). . . . 11

3 n_{jets} distribution in the ZZ-dominated control region left) without weights based on this variable right) with these weights . . . 12

4 $m_{4\ell}$ distribution left) ZZ-dominated control region right) loose cuts (only requiring 4ℓ, no taus) to show that $Z \to 2\ell \to 2\ell\gamma^* \to 4\ell$ is included in the sample . . . 13

5 n_{jets} distribution in $t\bar{t}$-dominated control regions. left) $S_{\text{T}} > 200$ GeV right) $S_{\text{T}} > 300$ GeV . . . 14

6 $E_{\text{T}}^{\text{miss}}$ distribution in $t\bar{t}$-dominated control regions. left) $S_{\text{T}} > 200$ GeV right) $S_{\text{T}} > 300$ GeV . . . 15

7 H_{T} distribution in $t\bar{t}$-dominated control regions. left) $S_{\text{T}} > 200$ GeV right) $S_{\text{T}} > 300$ GeV . . . 15

8 S_{T} distribution in $t\bar{t}$-dominated control region. . . . 16

9 $m_{3\ell}$ distribution in AIC-dominated control region. left) fake muon right) fake electron . . . 18

10 $m_{3\ell}$ distribution in AIC-dominated control region, fine binning. left) fake muon right) fake electron . . . 19

11 Fake rates as a function of the number of prompt non-isolated tracks above 7 GeV (no separation from leptons). Top: Elecron fakes in eee and $e\mu\mu$, bottom: muon fakes in $ee\mu$ and $\mu\mu\mu$. . . . 20

12 p_{T} distributions of the lowest p_{T} lepton. left) fake muon right) fake electron 21

13 $m_{\ell\ell}$ distribution in the dilepton + fake region. left) including events with $m_{\ell\ell\ell}$ on-Z right) events with $m_{\ell\ell\ell}$ on-Z removed from the dilepton-off-Z regions . 21

14 $m_{\ell\ell}$ distribution in the dilepton + fake region. top) e fake bottom) μ fake; left) $Z \to \mu\mu$ right) $Z \to ee$ . . . 22

15 H_{T} distribution in the dilepton fake region (no OSSF pair mass cut) . . . 23

16 Results: H_{T} distribution in L3 region (0 τ_{had}). Left: 0 b-tags, right: b-tagged. Top: 0 OSSF, bottom: OSSF on Z . . . 24

17 Results: H_{T} distribution in L3 region (0 τ_{had}). Left: 0 b-tags, right: b-tagged. Top: 1 OSSF below Z, bottom: OSSF above Z . . . 25

18 Results: $E_{\text{T}}^{\text{miss}}$ distribution in L3DY0 region (0 b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT . . . 26

19 Results: $E_{\text{T}}^{\text{miss}}$ distribution in L3DYl1 region (0 b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT . . . 27

DRAFT

20 Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L3DYz1 region (0 b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT 28

21 Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L3DYh1 region (0 b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT 29

22 Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L3DY0 region (with b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT 30

23 Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L3DYl1 region (with b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT 31

24 Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L3DYz1 region (with b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT 32

25 Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L3DYh1 region (with b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT 33

26 Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DY0 region (0 b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT 34

27 Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DYn1 region (0 b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT 35

28 Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DYn2 region (0 b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT 36

29 Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DYz1 region (0 b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT 37

30 Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DYz2 region (0 b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT 38

31 Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DY0 region (with b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT 39

32 Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DYn1 region (with b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT 40

33 Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DYn2 region (with b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT 41

34 Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DYz1 region (with b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT 42

35 Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DYz2 region (with b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT 43

36 $m_{3\ell}$ distribution in AIC-dominated control region with DYJetsToLL_M-50 sample instead of photon proxies. left) fake muon right) fake electron 46

37 $m_{\ell\ell}$ distribution in in the dilepton fake region (off-Z, trilepton events with $m_{\ell\ell\ell}$ on Z have been vetoed). left) method 1 right) method 2 46

DRAFT

38 Top: m_{et} distribution; bottom: $m_{\mu t}$ distribution. Left: no other leptons present, right: additional lepton present. If a third lepton is present, its flavor is opposite the flavor of the first lepton, and its charge is the same as the track's (rejecting third leptons from Z). **Note:** This is from the dilepton-triggered dataset, i.e. the tracks used here probably were good enough leptons to trigger. 47

DRAFT

1 Introduction

We pursue a broad search in final states with at least three isolated and prompt leptons (e, μ, τ_{had}). We aim at avoiding binning and the application of cuts until the last moment of the analysis, and prefer to categorize events as the last step in the analysis workflow in order to allow for agile adaption to unforeseen analysis needs.

For this reason, our background estimation methods are, as far as possible, independent of a binning scheme. The note at hand describes how these background estimations are achieved.

The most notable backgrounds are:

- $\text{Z} \to \ell\ell$ plus a fake lepton, either from a jet or via AIC
- fully leptonic $\text{t}\bar{\text{t}}$ decays with a fake lepton from a b-jet
- WZ
- ZZ

In addition to these, there are rare backgrounds such as $\text{t}\bar{\text{t}}V$ or VVV.

While the VV backgrounds are generally well modeled by Monte Carlo (MC), some corrections (for example, in the p_{T} spectra) might be necessary. Backgrounds with fake leptons are not as easily modeled by MC and are thus estimated using data-driven techniques or a hybrid method. The purpose of this document is to explain the details of the MC corrections and data-driven methods.

Where comparison to data are made, the data is the full 2012 CMS dataset [1]. Plots contain statistical uncertaintes only (gray hashed band). Inside the gray band, the statistical uncertainty of the underlying MC or fake seed sample is displayed separately (red hashed band). The gray band is the combination of the red band and the Poisson error of the background estimate (added in quadrature).

2 Data sets and MC Samples

2.1 Data sets

We use the following data sets:

- **FIXME:** Add data sets: Name, Trigger, Run, Luminosity

2.2 Background MC samples

We use the following Monte Carlo samples for background determination. The number of events is calculated as 1./treeR->GetWeight() which equals the number of events in the input that was given to EventAnalyzer (this input was Richard's SkimTrees).

DRAFT

Name	Label	xsec [pb]	L [fb^{-1}]	No. events	generation scheme	size (treeR)
WZ	WZJetsTo3LNu	1.22	1653	2016678		380M
ZZ	ZZJetsTo4L	0.179	26809	4804781		980M
TT_SemiL	TTJetsSemiLeptonic	97.97	259	25365231		12G
TT_FullL	TTJetsFullLeptonic	23.08	525	12108679		8.9G
TTWW	TTWWJets	0.002037	106634	217213		22M
TTW	TTWJets	0.23	850	195555		11M
TTZ	TTZJets	0.208	1008	209677		13M
WWW	WWWJets	0.08217	2679	220170		11M
WWZ	WWZJets	0.0633	3504	221805		11M
WZZ	WZZJets	0.019	11549	219428		11M
ZZZ	ZZZNoGstarJets	0.004587	48958	224572		13M
GluGluToHToTauTau		1.2466	776	967566		29M
GluGluToHToWWTo2LAndTau2Nu		0.4437	676	299975		27M
GluGluToHToZZTo4L		0.0053	187758	995117		258M
VBF_HToTauTau		0.0992	10055	997464		39M
VBF_HToWWTo2LAndTau2Nu		0.0282	10617	299401		35M
VBF_HToZZTo4L		0.000423	117910	49876		17M
WH_ZH_TTH_HToTauTau		0.0778	2569	199859		17M
WH_ZH_TTH_HToWW		0.254	788	200197		8.6M

FIXME: Emmanuel knows more about: Generation scheme: jet matching, fullsim, any invariant mass cuts, lepton filters, forced BR's etc

We only read the MC samples up to 100 times the data luminosity, assuming that there is no bias in doing so.

2.3 WZ MC details

FIXME: Maxi Generator level plots showing met, ht, njets and lepton and jet pt's. Also invariant masses that will show offshell generation, eg W*Z and WZ*.

2.4 ttbar MC details

3 Analysis Workflow

The general analysis workflow is as follows:

Table 1: Analysis workflow

	Data	MC backgrounds and signal
1.	Prepare AnalysisTree ntuples. Requires about 0.09 seconds per processed event and about 550 Bytes per saved event.	
1a.	This requires a few variables to be defined: • The Z window is defined to be 91 ± 10 GeV. To define the content of the below- and above-Z bins, we look at the invariant mass of all combinations of two OSSF leptons and decide based on the one closest to the Z, prioritized by 1) in the Z window, 2) below, 3) above. See Appendix B.2. • The W mass (for M_T calculation) is 80.385 GeV.	

Continued on next page

DRAFT

Table 1 – continued

	Data	Backgrounds and signal
1b.	Turn on fake proxies	Turn on fake proxies for backgrounds that include fakes (for subtraction, see Sec. 6.1) **FIXME:** How about signal?
1c.	Global cuts	
1c.I.	Filter cuts	– n/a –
1c.II.	Skip events with an opposite-sign same-flavor (OSSF) pair below 12 GeV **FIXME:** Show justification plot (although things are stable against variable of the cut within reason); don't apply at high H_T/ E_T^{miss} etc.	
2.	Define control regions (don't include in statistical analysis in the last step; more detailed information in other sections)	
2a.	WZ: 3 leptons, 0 b-tags, no τ_{had}, 1 on-Z OSSF pair, 50 GeV $< E_T^{miss} <$ 100 GeV	
2b.	ZZ: 4 leptons, $E_T^{miss} <$ 50 GeV, $H_T <$ 200 GeV, 0 b-tags, no τ_{had}, 2 OSSF pairs (at least one on-Z)	
2c.	$t\bar{t}$ 2ℓ: exactly 2 OSOF leptons ($e^{\pm}\mu^{\mp}$), at least 1 b-tag, no τ_{had}, $S_T >$ 200 GeV	
2d.	Z 3ℓ: 1 OSSF on-Z, 0 b-tag, no τ_{had}, $E_T^{miss} <$ 50 GeV, $H_T <$ 200 GeV, **FIXME:** $E_T^{miss} <$ 30 GeV? M_T?	
3.	– n/a –	Apply weights and corrections (see Sec. 5)
3a.	– n/a –	Assess systematic error of weights and corrections
4.	Measure fake rates	Set up MC subtraction for fake rates
4a.	Assess systematic error	
5.	Define signal regions and determine data/background/signal	
6.	Add signal as a background sample and assess the impact on steps 4 and 5	
7.	Statistical analysis	
7a.	p value calculation, limit setting, ...	

4 Common MC treatment (signal + background)

4.1 Pile-up weights

For each MC sample, the pile-up distribution is compared to the distribution in data, in bins of nvertex. Weights are applied on a per-event basis so that the distribution matches.

FIXME: Matt knows more about this: Pileup subtraction scheme, MC resolution and pileup, nvertex weights, plots showing isolation vs pileup, sensitivity of fake predictions to pileup

4.2 Lepton ID/ISO weights

This is implemented and applied to any MC.

FIXME: Matt knows more about this

DRAFT

4.3 Trigger weights

This is implemented and applied to any MC.

FIXME: Matt knows more about this

4.4 $E_{\mathrm{T}}^{\mathrm{miss}}$ resolution correction using Z+j data/MC

This is applied to any MC.

```
 if(isMC){
    EventVariableSmearMET* MET
        = new EventVariableSmearMET("MET","MET","HT","NRECOVERTICES",2.68,4.14,3.48,2.68,5.10,3.48);
    MET->setSeed(3141592654);
    handler->addEventVariable("MET",MET);
  }else{
    EventVariableSumPT* MET = new EventVariableSumPT("MET","MET");
    handler->addEventVariable("MET",MET);
  }
```

FIXME: Matt knows more about this

5 MC background

FIXME: Say a few words on rare MCs. Higgs backgrounds have a 50 % cross-section uncertainty.

5.1 WZ background

In the following, we will refer to the "WZ control region" (CR) which is defined by 3 leptons, 0 b-tags, no τ_{had}, an on-Z OSSF pair, and either 50 GeV $< E_{\mathrm{T}}^{\mathrm{miss}} <$ 100 GeV ("$E_{\mathrm{T}}^{\mathrm{miss}}$ control region", primary control region) or 50 GeV $< M_{\mathrm{T}} <$ 100 GeV ("M_{T} control region", for cross checks).

We currently do not apply the $H_{\mathrm{T}} < 200$ GeV cut because we find that this poses difficulties to normalization (in low H_{T}) when doing jet p_{T} weights at the same time in a region that has at least two jets (which has large overlap with high H_{T}).

We require all electrons or muons in this sample to come from W, Z, or τ_{had} parent particles.

We use WZ MC with fully leptonic decays and follow the following weighting procedure. Each weight preservers the area.

1. There is a $p_{\mathrm{T}}(\mathrm{Z})$-dependent weight to account for the p_{T} shape from the higher order contributions to boosted Z production: $\exp \frac{-0.005 \cdot p_{\mathrm{T}}(\mathrm{Z})}{\sqrt{m(\mathrm{Z})}}$, ranging between 0.9 and 1.0 in the relevant portion of phase space. It is estimated using MCFM and Pythia from what we understand. **FIXME:** What would be a good variable to show a before/after plot of?

2. In the $E_{\mathrm{T}}^{\mathrm{miss}}$ control region, we find that the n_{jets} distribution does not match, which we correct for by applying event weights: 1.06 for 0 jets, 0.89 for 1 jet, 1.18 for 2 jets, and 1.32 for more than 2 jets. [Fig. 1]

3. We then look at the p_{T} distribution of the leading jet in the $E_{\mathrm{T}}^{\mathrm{miss}}$ CR *with the H_{T} cut removed*. We find that the distribution does not match well and apply weights: 0.75 ($p_{\mathrm{T}}(j_{\mathrm{lead}}) <$ 70 GeV), 1.16 (70..110 GeV), 0.88 (110..150 GeV), 1.26 ($>$ 150 GeV).

DRAFT

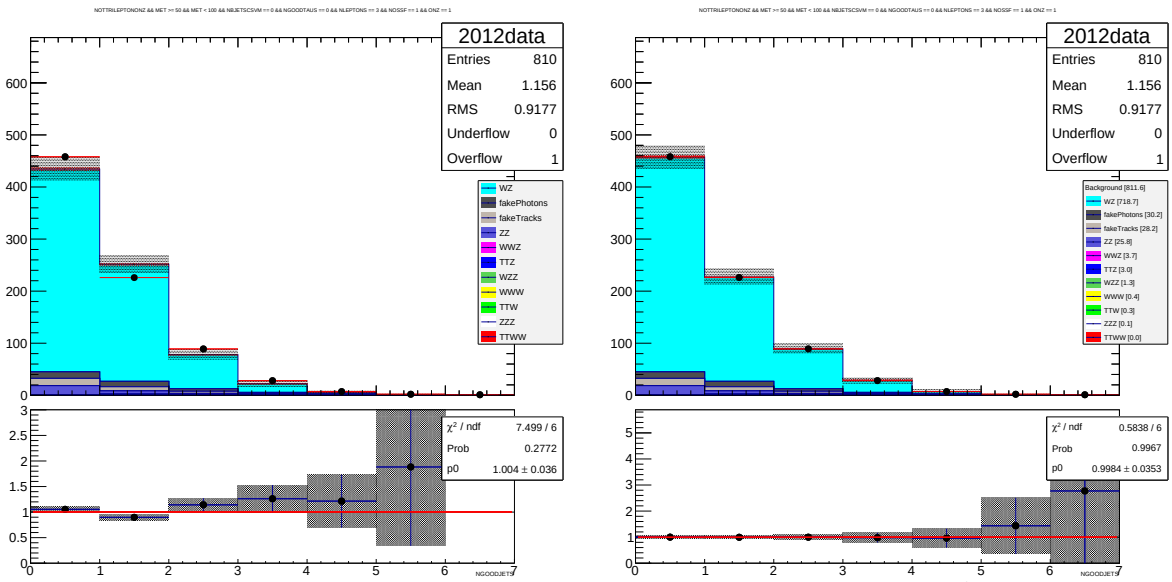

Figure 1: n_{jets} distribution in the WZ-dominated $E_{\text{T}}^{\text{miss}}$ control region, before and after n_{jets} weights

The purpose of this weight is to correct the WZ underprediction in the high-H_{T} region. [Fig. 2]
FIXME: *For Scott:* Are the WZ (and ttbar?) njet weights consistent with the missing higher order terms in the MC generation scheme? (If yes, less syst error.)

4. Finally **FIXME:** or first? Interferes with area vs. theory normalization?, we normalize the cross-section such that the areas of background and data match in the $E_{\text{T}}^{\text{miss}}$ control region (including other backgrounds).

Our cross-section measurement in the $E_{\text{T}}^{\text{miss}}$ control region is 1.192 pb. This includes the branching ratios for cuts applied during MC generation (for example the 3ℓ requirement). In regions where we use this normalization, we apply a 5 % systematic uncertainty based on the normalization statistical uncertainty.

FIXME: Add plots before/after weights, also of H_{T}.

DRAFT

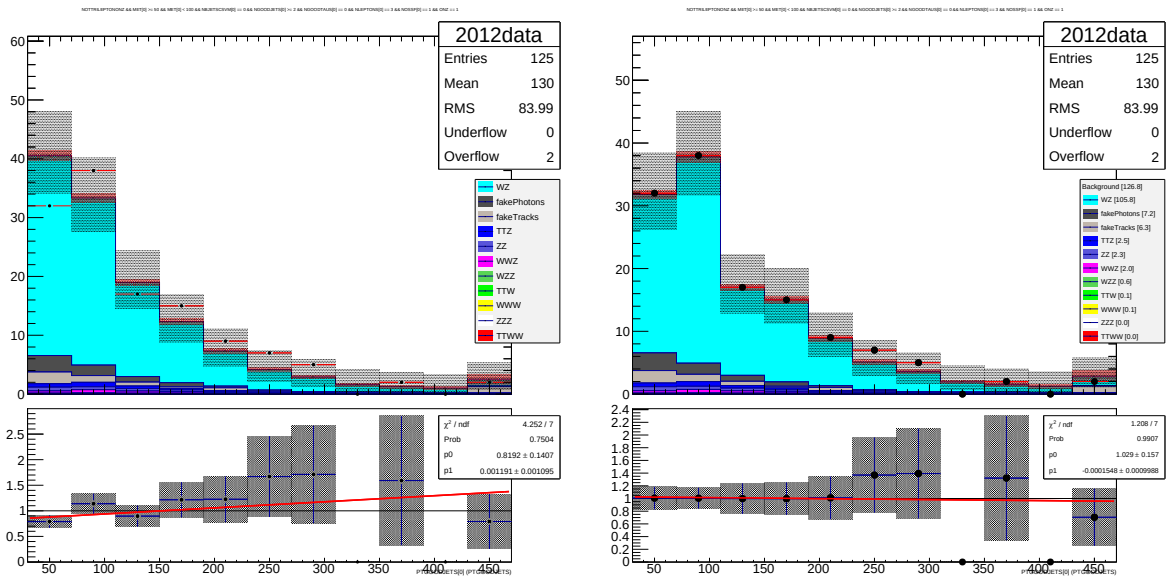

Figure 2: $p_{\mathrm{T}}(j_{\mathrm{lead}})$ distribution in the WZ-dominated $E_{\mathrm{T}}^{\mathrm{miss}}$ control region (no H_{T} cut). left) without weights based on this variable right) with these weights (all bins above 150 GeV combined into one bin with flat weight).

5.2 ZZ background

The "ZZ control region" is defined by 4 leptons, $E_{\mathrm{T}}^{\mathrm{miss}} < 50$ GeV, $H_{\mathrm{T}} < 200$ GeV, 0 b-tags, no τ_{had}, 2 OSSF pairs (at least one on-Z).

We use ZZ MC with fully leptonic decays and apply the following weights:

1. We find that the n_{jets} distribution does not match, which we correct for by applying event weights: 1.084 for 0 jets, 0.75 for 1 jet, 0.333 for 2 jets, and 1.0 for more than 2 jets. [Fig. 3]

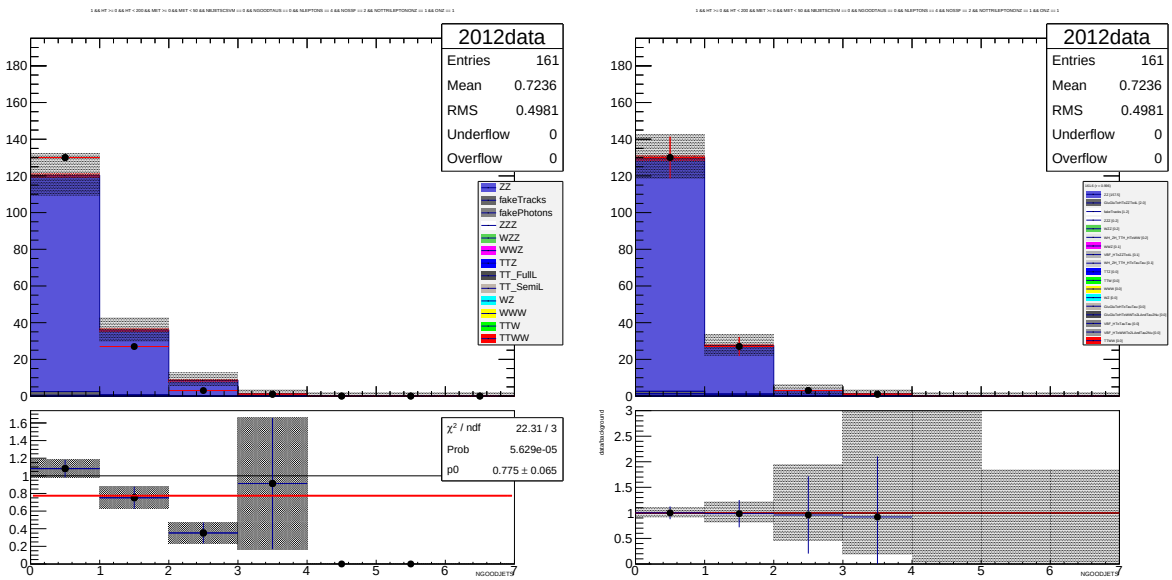

Figure 3: n_{jets} distribution in the ZZ-dominated control region left) without weights based on this variable right) with these weights

Figure 4 shows the 4ℓ mass distribution in the control region, and also for a looser region as a cross-check.

Our cross-section measurement in the ZZ control region is 0.1787 pb. This includes the branching ratios for cuts applied during MC generation.

FIXME: Put mZ2 distribution for mZ1 below Z, on Z, above Z. Need to redo AnalysisTree; use vector capabilites.

DRAFT

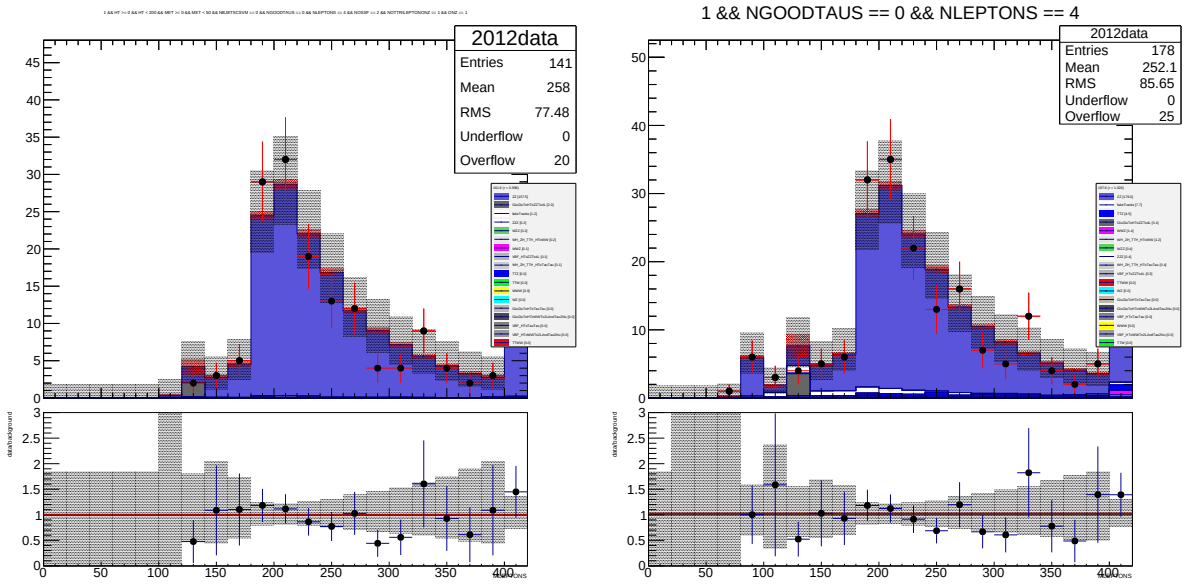

Figure 4: $m_{4\ell}$ distribution left) ZZ-dominated control region right) loose cuts (only requiring 4ℓ, no taus) to show that $Z \to 2\ell \to 2\ell\gamma^* \to 4\ell$ is included in the sample

5.3 $t\bar{t}$ background

We use $t\bar{t}$ MC to predict the number of 3ℓ background events from $t\bar{t}$ decays.

5.3.1 2ℓ studies

5.3.1.1 Dilepton weights The "$t\bar{t}$ 2ℓ control region" is defined by exactly 2 OSOF leptons ($e^{\pm}\mu^{\mp}$), at least 1 b-tag, no τ_{had}, and an S_{T} requirement of 200 GeV or 300 GeV.

We follow the following weighting procedure:

1. We find that the n_{jets} distribution does not match, which we correct for by applying event weights derived in the $S_{\text{T}} > 300$ GeV control region: 1.11 for less than 2 jets, 1.08 for 2 jets, 1.065 for 3 jets, 1.065 for 4 jets, 1.04 for 5 jets, and 1.0 for more than 5 jets. [Fig. 5] The weights differ if derived in the $S_{\text{T}} > 200$ GeV region, especially for 2 jets. Use this for systematic error?

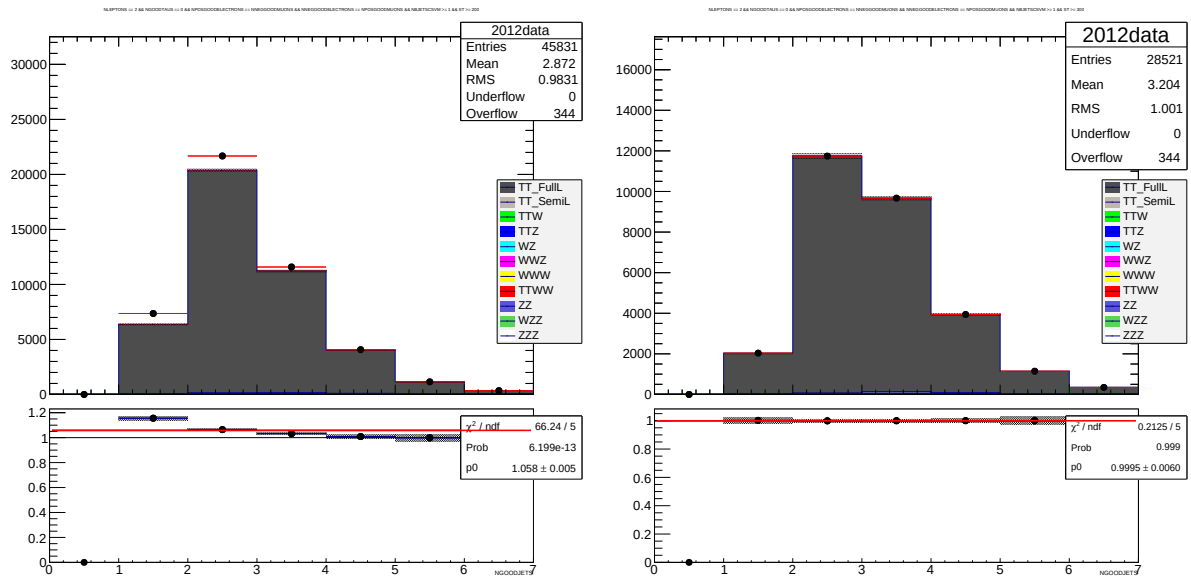

Figure 5: n_{jets} distribution in $t\bar{t}$-dominated control regions. left) $S_{\text{T}} > 200$ GeV right) $S_{\text{T}} > 300$ GeV

FIXME: State our cross-section measurement based on the above

Figures 6–8 show the $E_{\text{T}}^{\text{miss}}$, H_{T}, and S_{T} distributions.

5.3.1.2 MC modeling of isolated fake-leptons We know that the MC does not model the isolation shape of fake-leptons correctly, leading to an underprediction of the number of isolated fake leptons.

To study this (and find a fudge factor), we assume that the semi- and fully leptonic $t\bar{t}$ MC sample are affected by this issue in the same way. To avoid using a signal region, we thus look at a single-lepton-triggered region of semi-leptonic $t\bar{t}$ with an isolated fake lepton. We require a good muon above 30 GeV, at least 3 jets above 40 GeV at least one of which is b-tagged, and $S_{\text{T}} > 300$ GeV. In this region, we normalize the semi-leptonic $t\bar{t}$ MC.

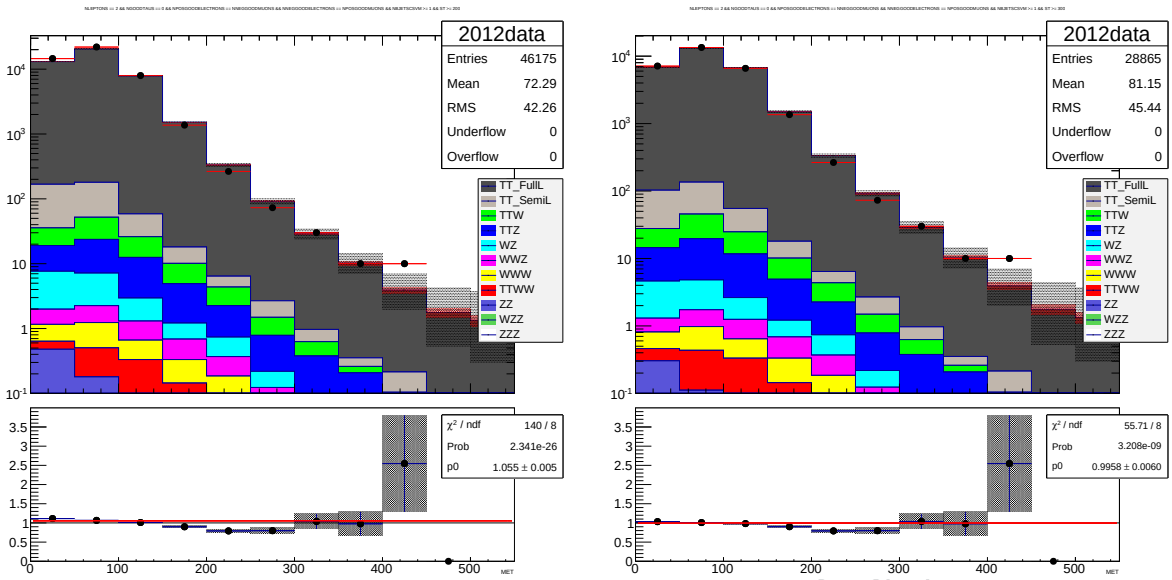

Figure 6: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in $t\bar{t}$-dominated control regions. left) $S_{\mathrm{T}} > 200$ GeV right) $S_{\mathrm{T}} > 300$ GeV

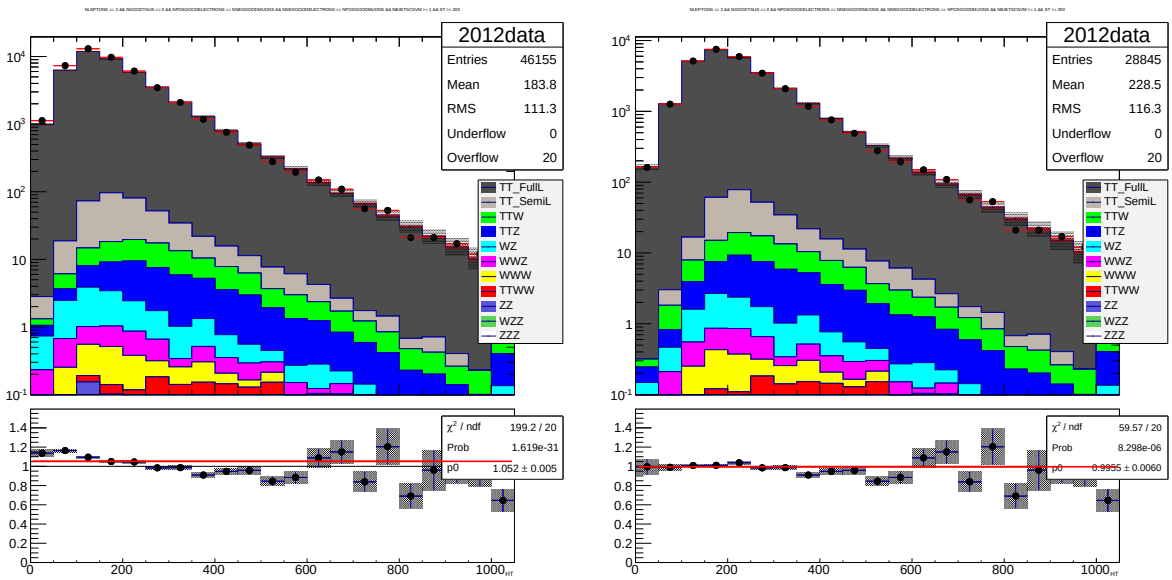

Figure 7: H_{T} distribution in $t\bar{t}$-dominated control regions. left) $S_{\mathrm{T}} > 200$ GeV right) $S_{\mathrm{T}} > 300$ GeV

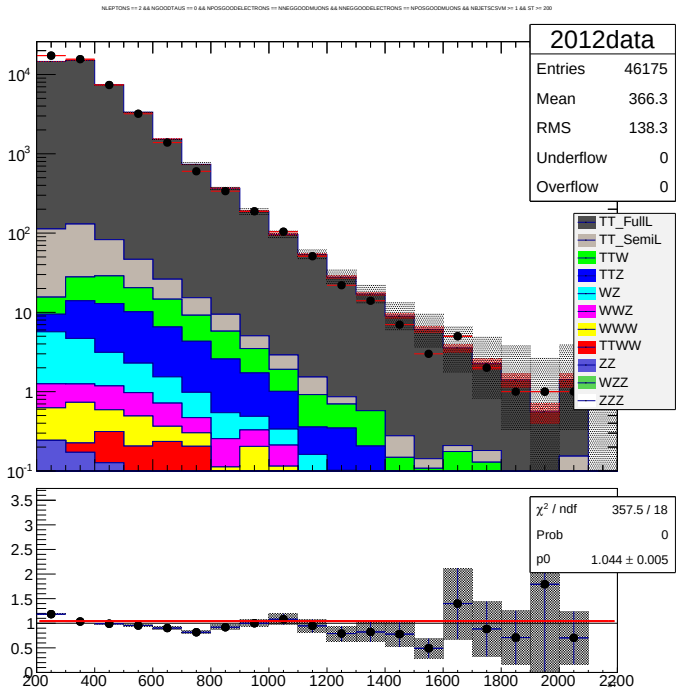

Figure 8: S_T distribution in $t\bar{t}$-dominated control region.

We then require an additional non-prompt muon ($d_{xy} > 0.03$ **FIXME:** unit?) and measure the ratio of data and (semi+full) $t\bar{t}$ MC. This ratio is 1.66 ± 0.32 and applied as a fudge factor for 3ℓ $t\bar{t}$ events in the signal regions.[1]

We cross-check with in an OSOF region (also single-lepton-triggered, thus with a muon above 30 GeV, no special p_T requirement for the electron; dominated by fully-leptonic $t\bar{t}$ decay), where we require only one jet (which needs to be above 40 GeV). We find that the fudge factor in this region is 1.58 ± 0.86, in agreement with the one from the single-lepton region. (The same study in a dilepton-triggered sample yields consistent numbers.)

In regions where we use this correction, we apply a systematic uncertaint of 19.4 % based on the statistical uncertainty of the correction.

FIXME: Add plots. Currently, I don't have them as PDF; look at ttbarFudge_noFake_NGOODJETS.png and ttbarFudge_RELISONONPROMPTMUONS.png in this directory.

5.3.2 3ℓ studies

We apply all weights from the 2ℓ studies, and in addition carry out the following steps:

1. We correct the number of isolated fake leptons by applying the isolation fudge factor as described in the previous section.

2. Some of the 3ℓ background is already predicted by the light lepton fake rate method. To avoid double-counting, we calculate the fake background based on the MC and subtract it from the 3ℓ MC background (see Sec. 6.1). We are assuming here that the track modeling

[1] Our isolation plot looks different from the 2012 one because we're using the "new" (CSV) b-tagging, while Richard's plot used an older algorithm. We determined that W+Jets is not a very significant background, but we don't have a large enough sample to really quantify. Can be covered by systematic, though.

in the 2ℓ region of the $t\bar{t}$ MC is roughly right, and that the light lepton fake rate method is applicable. **FIXME:** Show that (track p_T distribution)
Note: The above isolation fudge factor is also applied to the 2ℓ+track fake events since the tracks are isolated as well, and are thus expected to suffer from the same issue. **FIXME:** Track iso plot

6 Data-driven backgrounds

6.1 Light lepton fakes

To determine the background with fake electrons and muons, we rely on data and use other objects that are emitted further up or on the same level in the decay chain as lepton proxies. We show in a control region that the kinematic properties of these proxies resemble those of the fake leptons. We then generate a fake sample based on the 2ℓ+[proxy object] data where the proxy objects are fully treated as leptons. Further down in the analysis chain, these fake leptons are fully treated as regular leptons (e.g. in computing the dilepton invariant mass for Z-window binning). Proxy objects that can take multiple roles are considered the appropriate number of times (see Sec. 6.3).

The number of 3ℓ events per 2ℓ+[proxy object] event in this fake sample is then evaluated ("fake rate"). With the help of the fake rate, we predict the background in our signal regions, by applying it to the corresponding seed sample which requires one less lepton and a proxy object instead. This is simply done by selecting the signal region from the fake sample.

To compute the fake rate $\frac{N(3\ell)}{N(2\ell+\text{[proxy object]})}$, we subtract contributions from other backgrounds in the numerator and the denominator[2], including other data-driven backgrounds. (This requires an iterative process to converge.) The fake rate thus describes the number of fake leptons as a fraction of the number of 2ℓ+[proxy object] events from all processes that have not been modeled otherwise.

When we apply the fake rate in a signal region, we multiply it by the total number of 2ℓ+[proxy object] events found in the corresponding seed region in data, without any subtractions from the data sample. However, we use MC to obtain the fake contribution for certain backgrounds.[3] In these cases, double-counting needs to be mitigated. Therefore, we take the 2ℓ+[proxy object] component of the background MC sample, apply the same fake rate as for data, and subtract the resulting prediction from the 3ℓ MC prediction (see e.g. Sec. 5.3 for $t\bar{t}$). If the result is negative, we replace it by zero.

FIXME: Currently turned off: Potential mismatch of kinematic properties is corrected for using p_T dependent weights or p_T loss factors (see Sec. 6.2). We also study to what extent the fake rate depends on other properties of the event (for example the jet composition and spectra), and parameterize the fake rate as necessary. The freedom that we find in determining these parameterizations and kinematic weights is used to assess the systematic uncertainty of the background estimate.

[2] In the denominator, we currently only subtract $t\bar{t}$, which turns out to have a negligible contribution. **FIXME:** Think about this

[3] This is especially important for $t\bar{t}$ when a b-tag is not present, since the fake rate is higher in $t\bar{t}$ events, but there is no obvious way to discern these events from non-$t\bar{t}$ events in the seed sample.

DRAFT

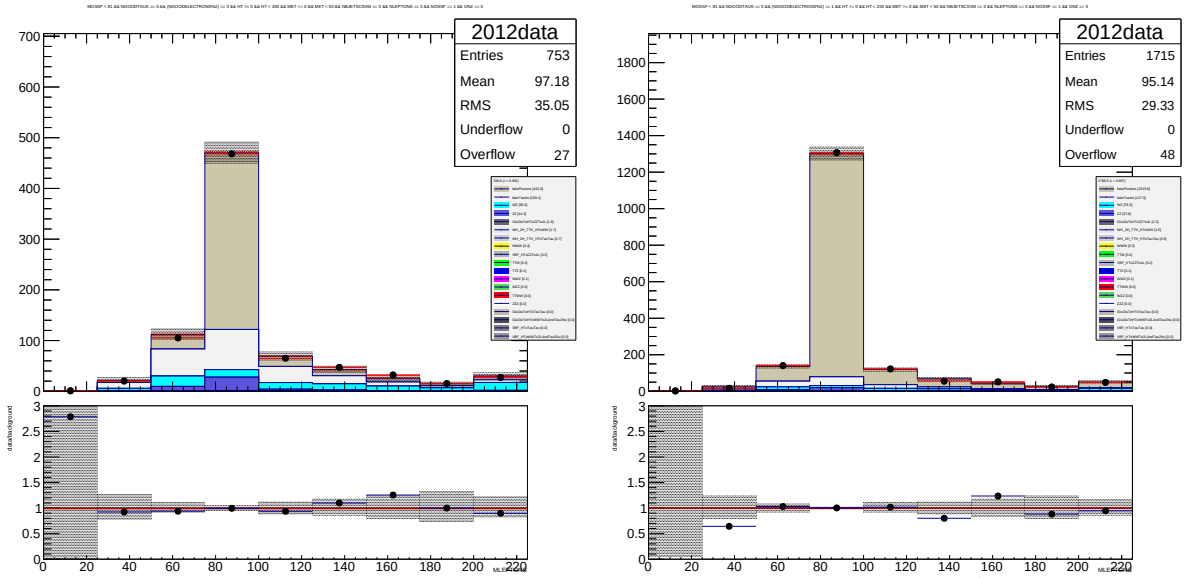

Figure 9: $m_{3\ell}$ distribution in AIC-dominated control region. left) fake muon right) fake electron

6.2 Fake leptons from asymmetric internal photon conversions (AIC)

We look at the number of events that have 3 light leptons (no τ_{had}) including an OSSF pair below Z (i.e. $m_{\ell\ell} < 81$ GeV), no b-tags, $H_{\text{T}} < 200$ GeV, and $E_{\text{T}}^{\text{miss}} < 50$ GeV. This is essentially the Z peak region, except that the dilepton invariant mass is not large enough to fall on the Z peak and a third lepton is present. This region primarily contains events from $\text{Z} \to \ell\ell$ where one of the final state leptons radiates a photon which decays asymmetrically to two additional leptons, one of which carries very low p_{T} and thus escapes the detector undetected.

At the same time, the other additional lepton carries most of the photon p_{T} so that one would expect to recover the Z peak in the trilepton invariant mass (or, equivalently, in the invariant mass of the two primary leptons and the photon, up to a p_{T} loss factor close to unity).

We thus generate a fake sample where photons are treated as leptons by assigning them to the lepton collections. All combinations are taken into account, i.e. dilepton events with a photon enter the fake sample as four events (two possible flavors, two possible charges). Looking in this sample, we find that the $2\ell + \gamma$ mass indeed reproduces the Z peak, as shown in Fig. 9. Note: Here, we are using a Z peak bin with range 75..100 GeV. **FIXME:** See if we can use the regular one The fake rates are

- 0.295 % for muons,
- 0.88 % for electrons.

For photons faking muons, we find better agreement if we apply a loss factor of 0.8 to the photon p_{T} when creating the fake trilepton sample, attributing an average of 20 % of the p_{T} to the lost lepton. If this factor is not applied, the 50–75 GeV bin is not modeled accurately. **FIXME:** Show how it looks without loss factor By analogous reasoning, the photon p_{T} loss factor is also applied in the photon-based estimation of fake muon backgrounds from jets (Sec. 6.3). Note, however, that the width of the background peak is not the same as in data (see finely binned $m_{3\ell}$ distribution in Fig. 10).

For photons faking electrons, we find that the distribution does not match well outside the

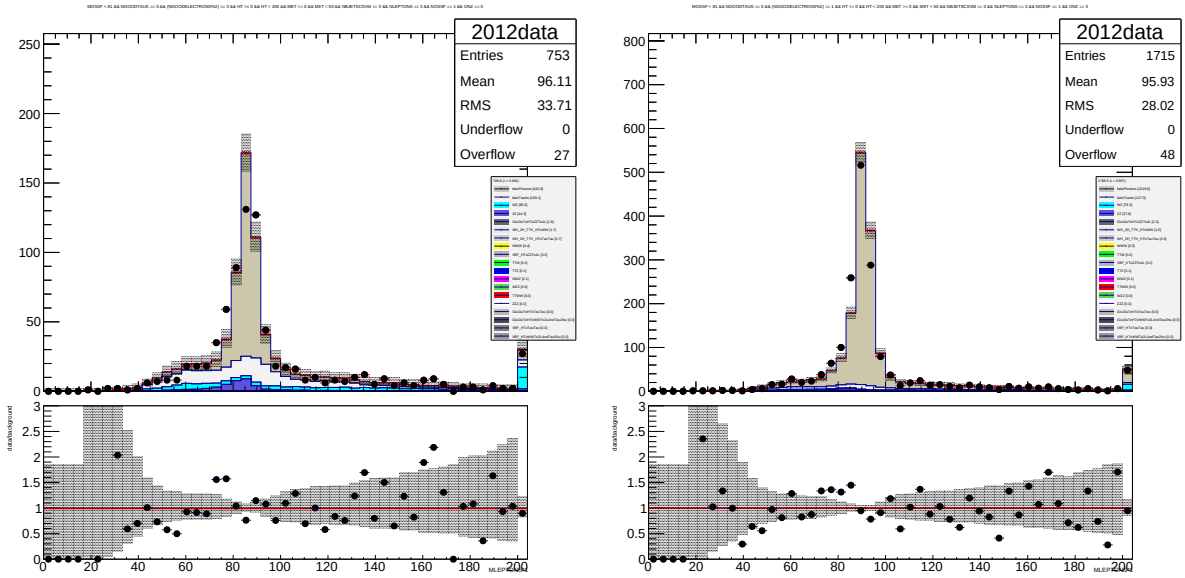

Figure 10: $m_{3\ell}$ distribution in AIC-dominated control region, fine binning. left) fake muon right) fake electron

trilepton Z window unless we apply a scale factor of $\frac{2}{3}$. The plot in Fig. 9 has this factor applied. We find that a loss factor for electrons would not bring a significant improvement.

We currently apply a 15 % systematic uncertainty on the total photon-based background estimate. This is because the photon fake rate depends on the environment flavor within this range (similar to the track-based electron fake rate. **FIXME:** Investigate this further or at least apply uncertainty to electron background only.

6.3 Fake leptons from jets

For fake electrons and muons from jets, our proxies are isolated tracks and photons in the 2ℓ data sample. In addition to the photon fake sample (see Sec. 6.2), we produce a track-based fake 3ℓ background sample by re-assigning tracks to the lepton collections. All combinations are taken into account, i.e. tracks are used to create both a fake-e and a fake-μ event.[4]

We then look at events with 3 light leptons (no τ_{had}) including an OSSF pair on Z (as defined in Section 3), no b-tags, and $E_{\text{T}}^{\text{miss}} < 50$ GeV. This is the prominent Z peak region with an additional lepton, but without an H_{T} requirement. The fake rates are displayed in Fig. 11. Observations and remedies:

- The fake rate does not only depend on the flavor of the fake lepton, but also on whether the Z goes to ee or $\mu\mu$. Therefore, we determine independent fake rates for these regions. For the prediction of $e\mu$+fake, we use an environment-agnostic average of the fake rate that depends on the fake flavor only.
- The fake rate depends on H_{T}. To take care of this, we bin in the number of prompt non-isolated tracks (no separation from leptons required, p_{T} threshold at 7 GeV) and take the linear fit as the fake rate function.
- The muon fake rate goes slightly negative if photon-based fakes are subtracted. We

[4] Multiple fakes in an event are only taken into account if they are from the same proxy type, e.g. two fake leptons from tracks. Hybrid fakes (one from a track, one from a photon) are currently not supported for technical reasons. Given the smallness of the fake rates ($O(10^{-2})$), the contribution from hybrid fakes is negligible.

DRAFT

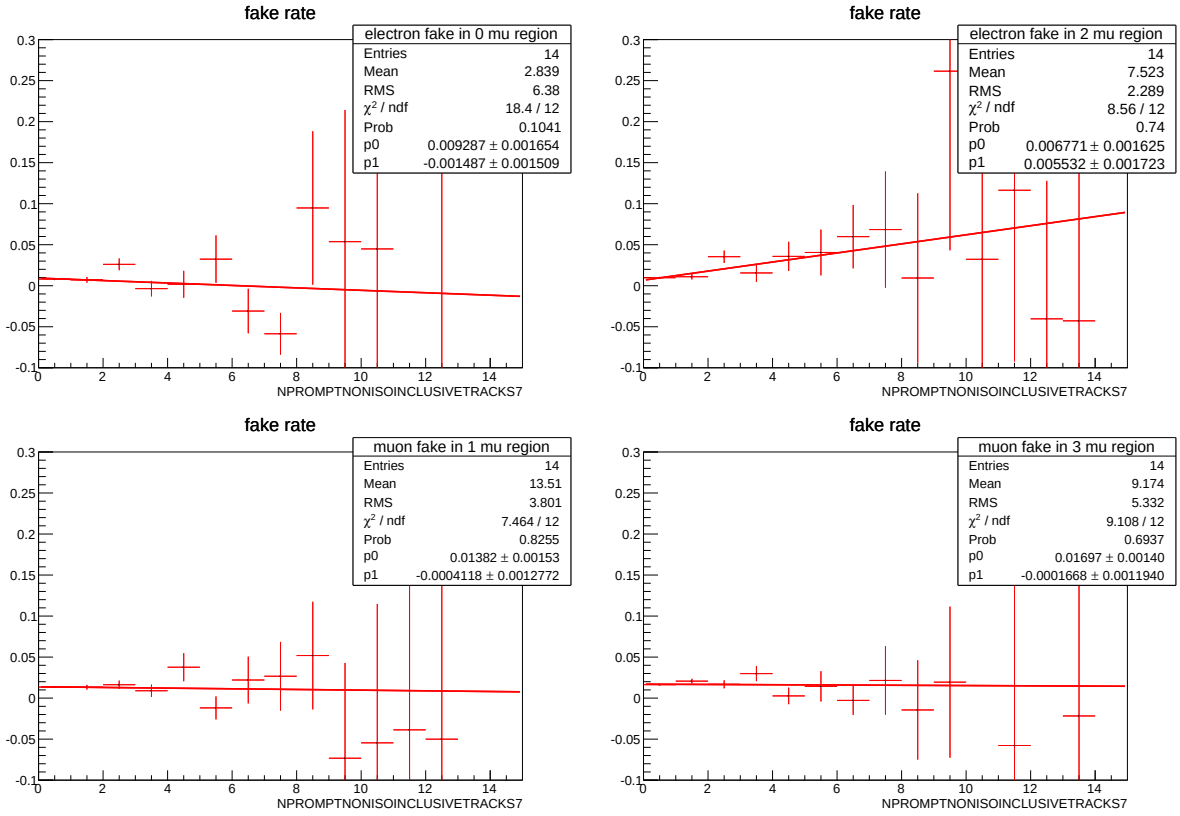

Figure 11: Fake rates as a function of the number of prompt non-isolated tracks above 7 GeV (no separation from leptons). Top: Elecron fakes in eee and $e\mu\mu$, bottom: muon fakes in $ee\mu$ and $\mu\mu\mu$.

thus set the photon fake rate on-Z to 0. This means that electron fakes in Z +jets are modeled by tracks and photons, while muon fakes are modeled by tracks only.

Track- and photon-based fake rates cannot be disentangled and need to be determined iteratively. We find best convergence for the combination if we change the photon-based electron background scale factor from $\frac{2}{3}$ (see Sec. 6.2) to $\frac{1}{2}$. This adjustment is done on-Z only and leads to an attribution of about 60 % of the data-driven electron background to photons, and of roughly 40 % to tracks. The same is true for muons, but only off Z (because the photon fake rate is 0 on Z).

FIXME: Currently turned off: To make the p_{T} distributions match match, we apply weights to the track-based background in bins of the the lowest p_{T} lepton (which is generally the fake – note that the fake proxy objects are treated as leptons here). For the electron part, the weights are between 0.4**FIXME:** Explain that this is so high because of light jets, but leptons come via semileptonic decays from c/s etc. and 2.3 (10..25 GeV), and 4.0 above 25 GeV; for the muons, we only need to scale the first bin (10..15 GeV) by 1.13. [Fig. 12]

Currently, we apply a 17 % systematic uncertainty on the track background estimate, based on the uncertainty of the linear fit. **FIXME:** The uncertainty increases with jet activity – parameterize

Fig. 13 shows the mass distribution of the "best" OS dilepton pair across its full range in the trilepton control region; Fig. 14 distinguishes by flavors. In case of ambiguity, the "best" OS

FIXME: currently turned off

Figure 12: p_{T} distributions of the lowest p_{T} lepton. left) fake muon right) fake electron

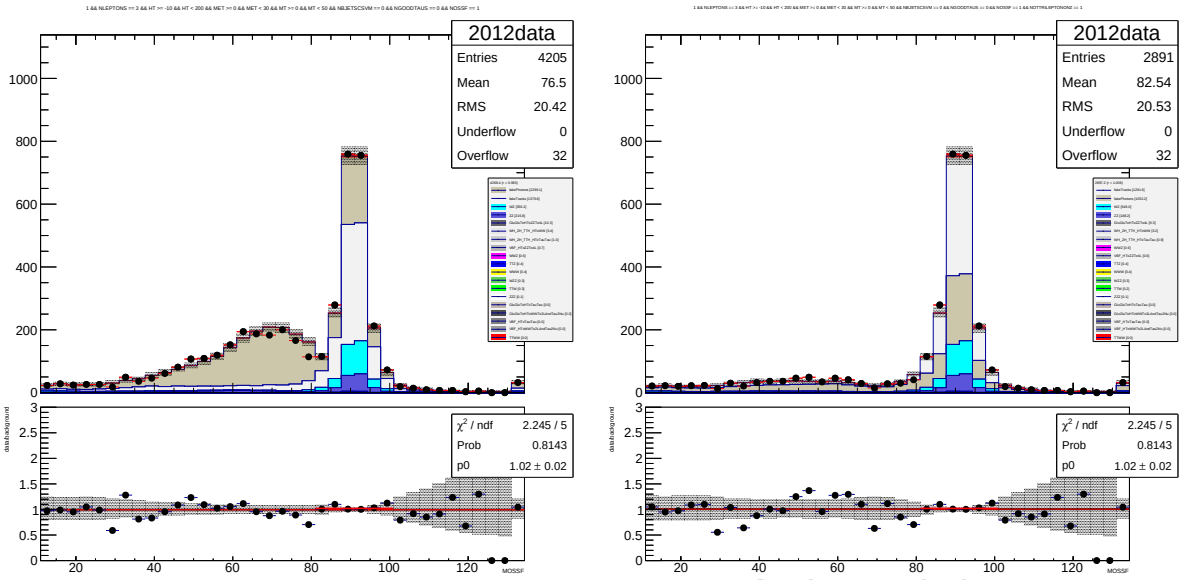

Figure 13: $m_{\ell\ell}$ distribution in the dilepton + fake region. left) including events with $m_{\ell\ell\ell}$ on-Z right) events with $m_{\ell\ell\ell}$ on-Z removed from the dilepton-off-Z regions

dilepton pair is the one whose invariant mass is closest to the Z mass, with the additional condition that pairs above the Z window are not considered if there is a pair below the Z window (thus shifting events from high-Z to low-Z, for a more separative background distribution). For a comparison with another Z-ness binning scheme, see Appendix B.2.

Overall, we get a good prediction across all of H_{T} (see Fig. 15).

6.3.1 b-tagged regions

In b-tagged regions, we scale both the track and the photon fake rate by a factor of 2.4 to make the b-tagged Z control region match. **FIXME:** show plot **FIXME:** This should be done with greater care. Photons/tracks separately? Parameterization?

6.4 Tau fakes

We use tau candidates in the isolation side band (isolation 6 to 15 GeV) as proxies for fake τ_{had}, by creating a fake sample that has these tau candidates in the tau lepton collection. If there are several candidates, one at a time is taken as a fake.

The fake rate is measured on-Z, with $E_{\mathrm{T}}^{\mathrm{miss}} < 50$ GeV and without b-tags (simply Z region with exactly one τ_{had}). For $H_{\mathrm{T}} < 200$ GeV, the fake rate measurement yields 2.2 %, and for high H_{T} it is 1.9 %.

FIXME: Parameterize to take care of jet dependence etc.

DRAFT

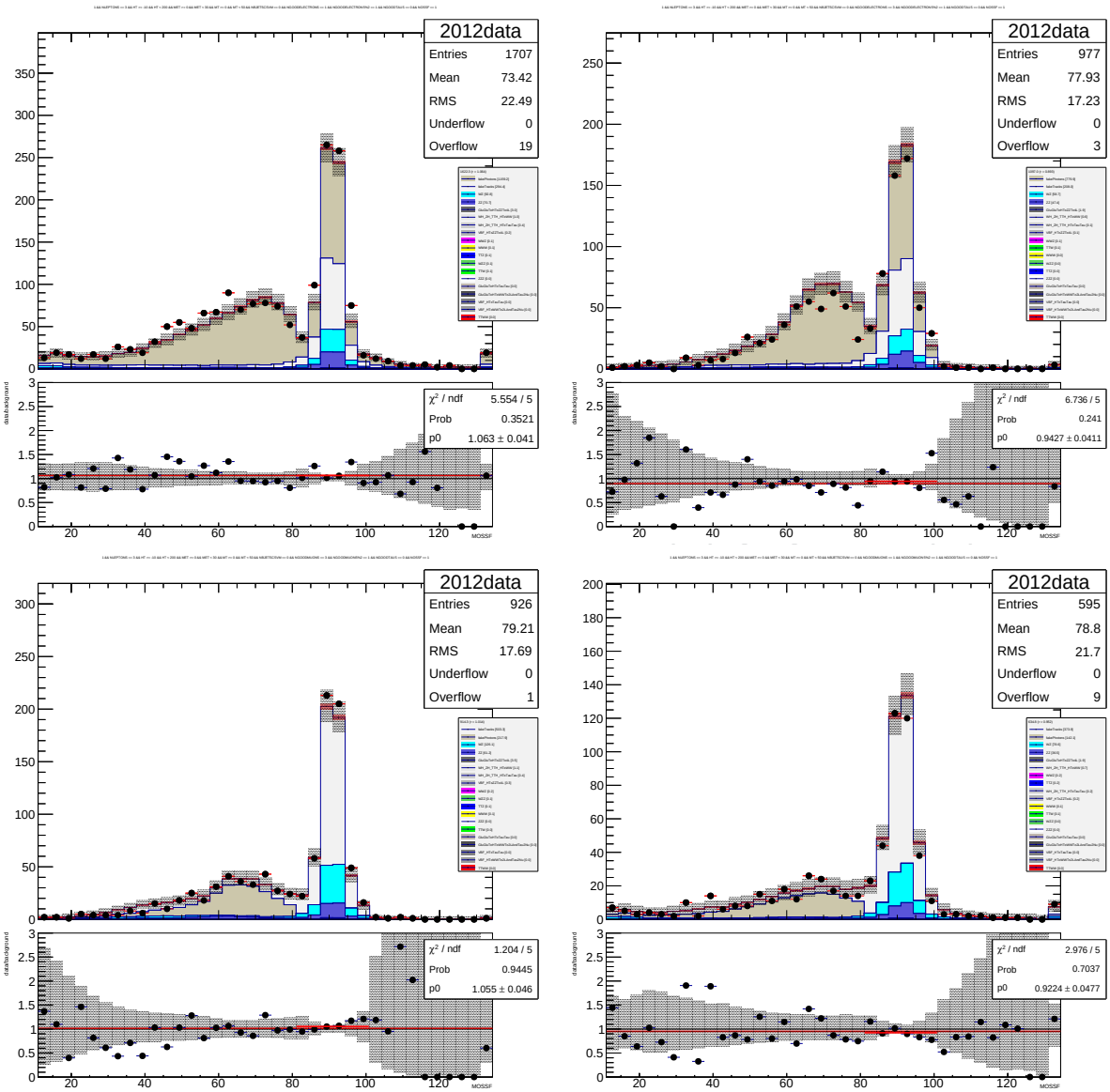

Figure 14: $m_{\ell\ell}$ distribution in the dilepton + fake region. top) e fake bottom) μ fake; left) $Z \to \mu\mu$ right) $Z \to ee$

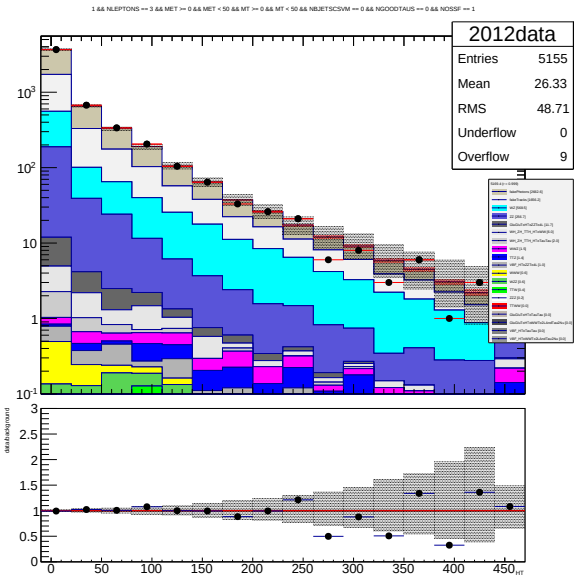

Figure 15: H_T distribution in the dilepton fake region (no OSSF pair mass cut)

7 Results

We present results binned in the number of OSSF pairs, τ_{had}, b-tags, and in H_{T} below and above 200 GeV. If applicable, we distinguish by the mass range of the OSSF pair (in/below/above the Z window). Some of these are control regions are therefore designed to match perfectly.

Note: Legibility improvements are underway. There are also plots available that are integrated over H_{T}, or over H_{T} and the number of b-tags (please e-mail Peter if interested).

7.1 3 leptons

DY0: no OSSF pair; DYl1: one pair, below Z; DYh1: one pair, above Z; DYz1: one pair, on Z

We start with a few summary plots (H_{T} instead of $E_{\text{T}}^{\text{miss}}$ distributions, integrated over $E_{\text{T}}^{\text{miss}}$):

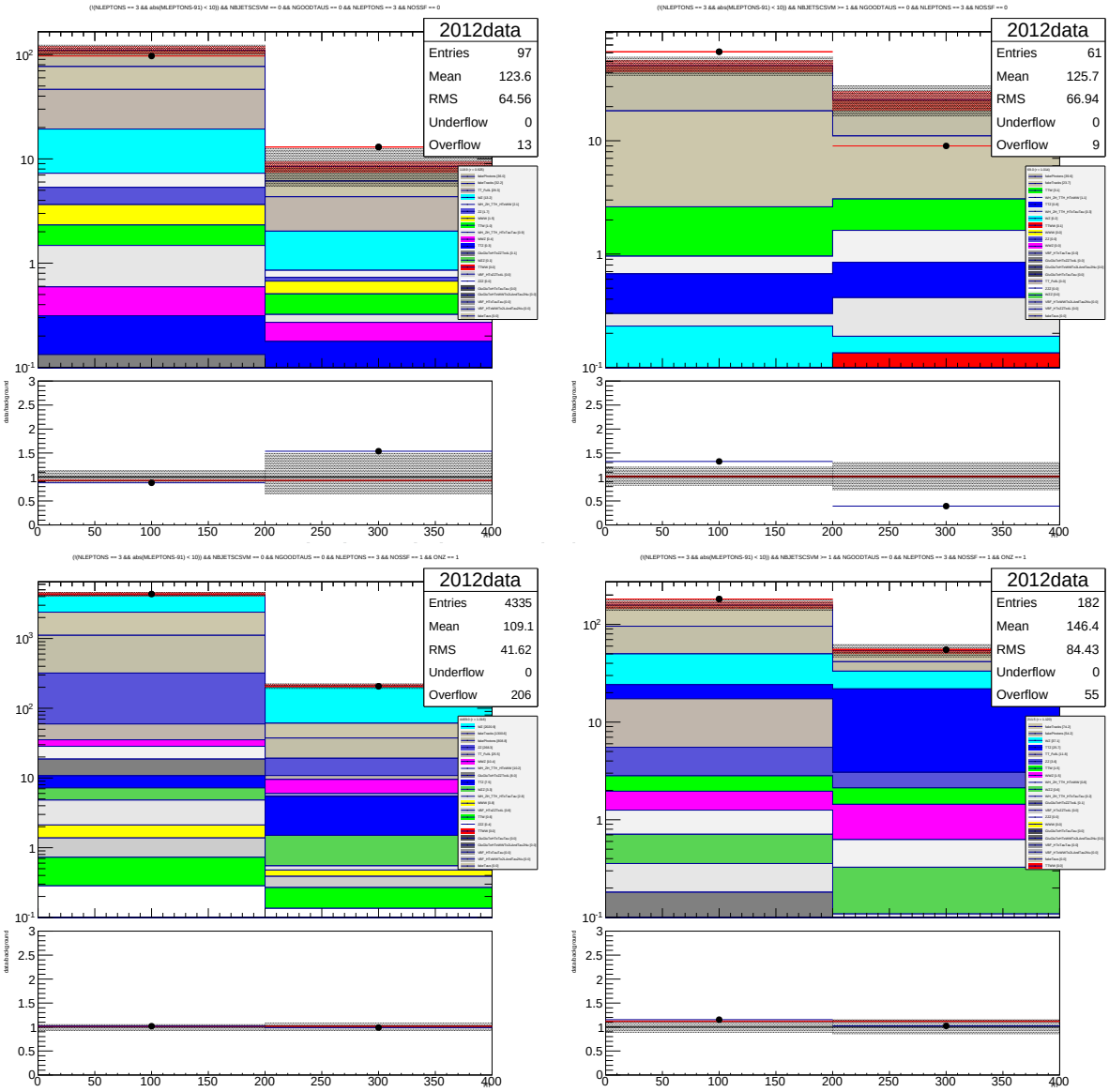

Figure 16: Results: H_{T} distribution in L3 region (0 τ_{had}). Left: 0 b-tags, right: b-tagged. Top: 0 OSSF, bottom: OSSF on Z

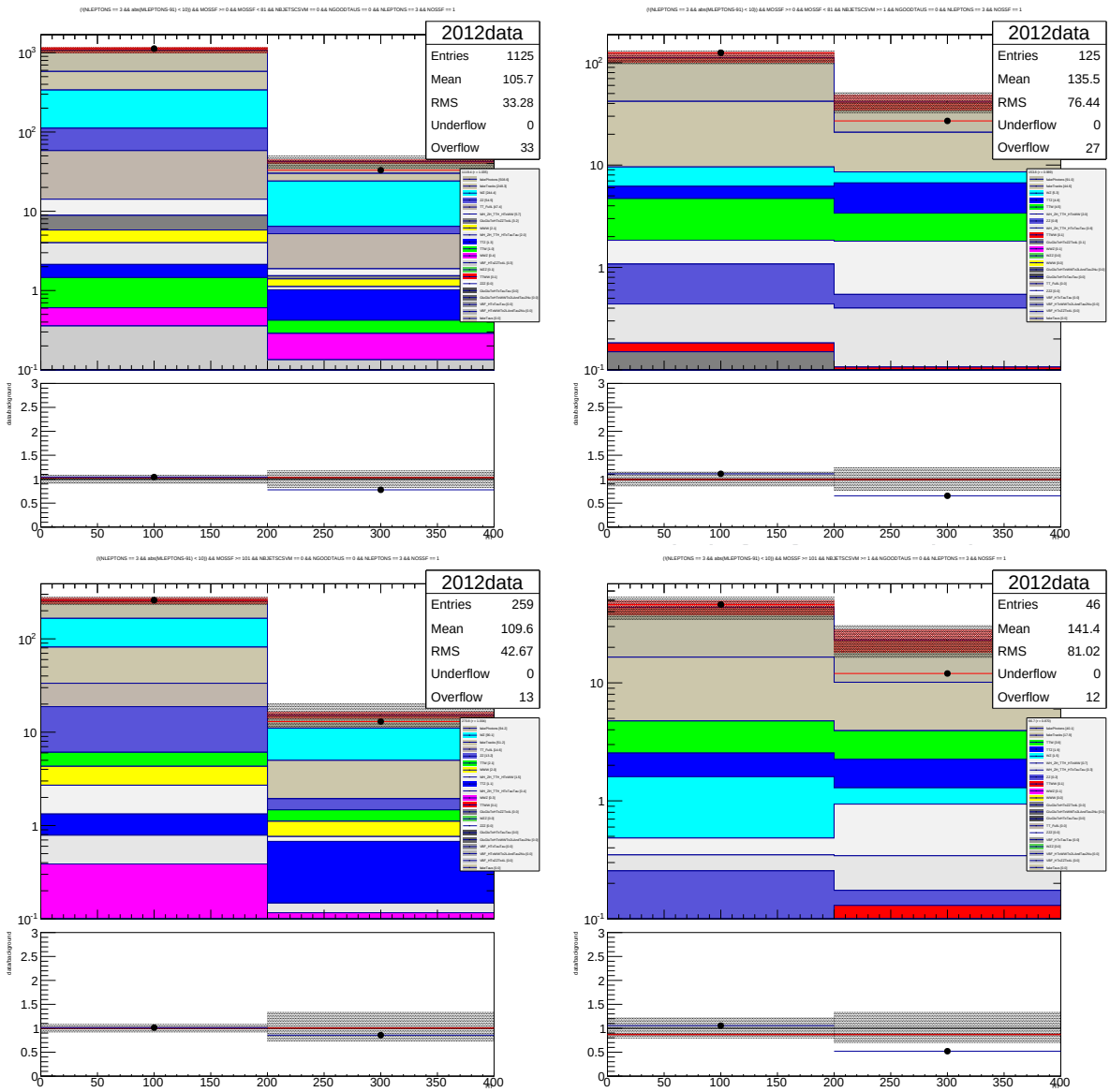

Figure 17: Results: H_T distribution in L3 region (0 τ_{had}). Left: 0 b-tags, right: b-tagged. Top: 1 OSSF below Z, bottom: OSSF above Z

405 Detailed $E_{\mathrm{T}}^{\mathrm{miss}}$ distributions follow:

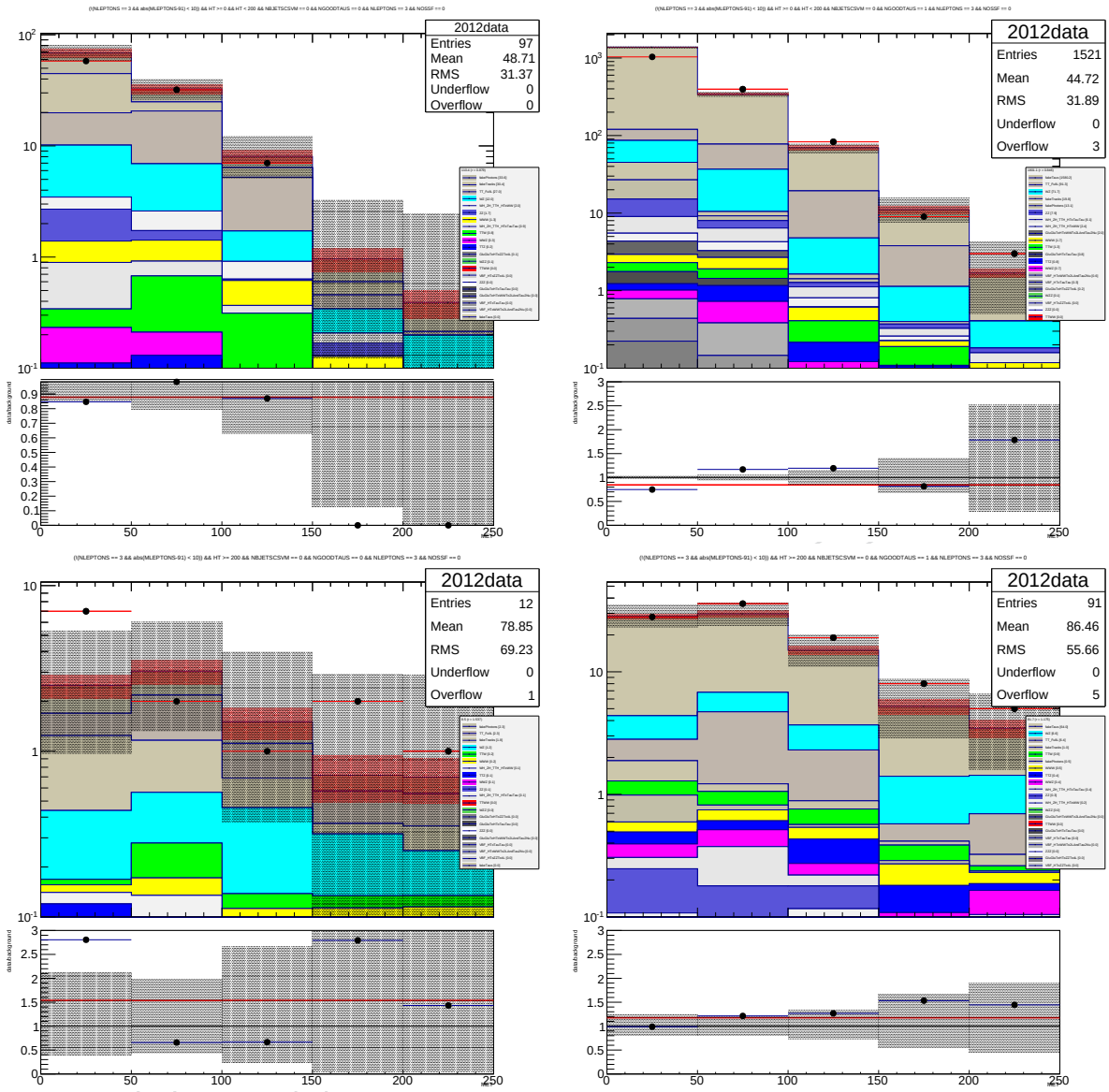

Figure 18: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L3DY0 region (0 b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

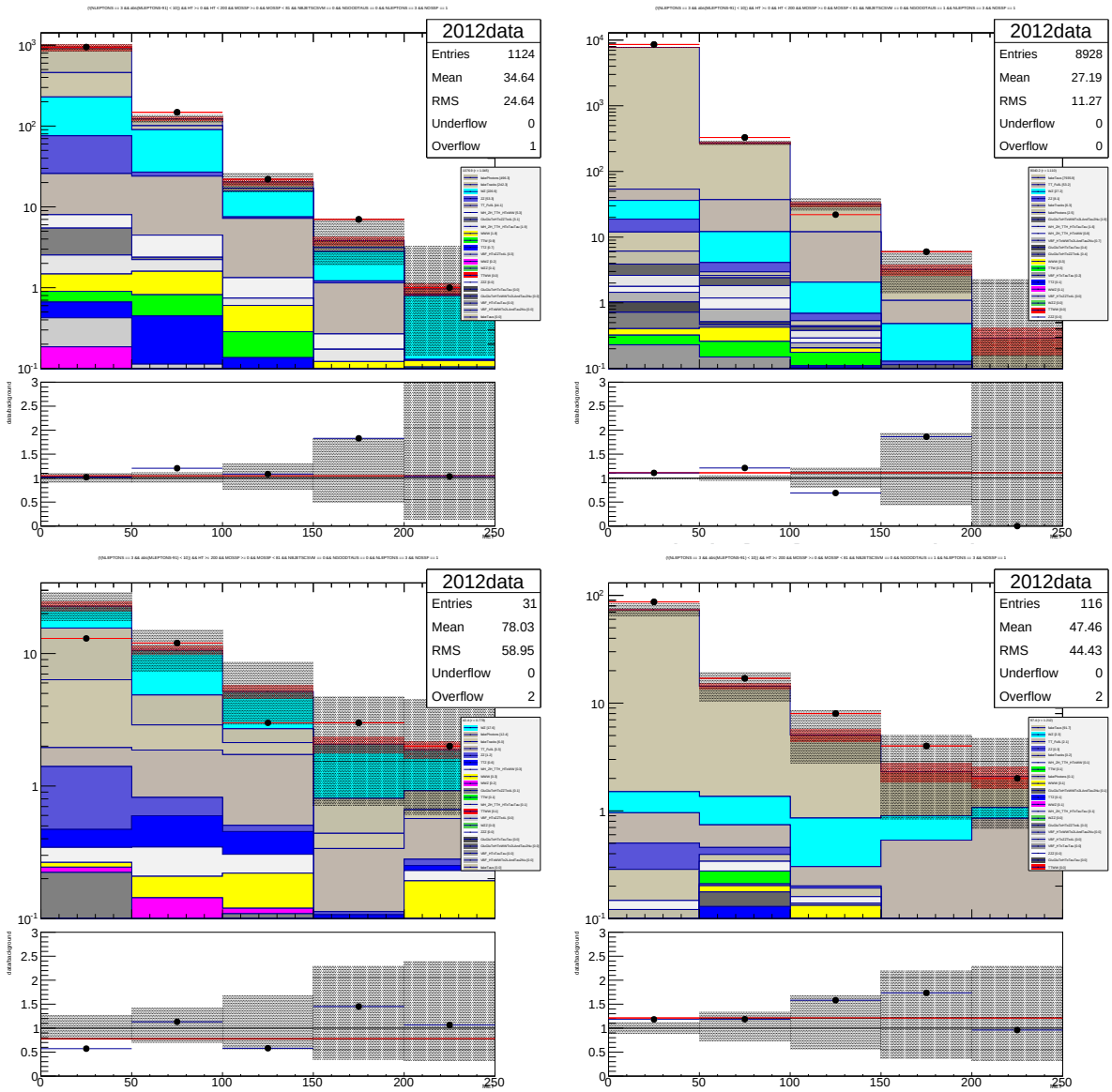

Figure 19: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L3DYl1 region (0 b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

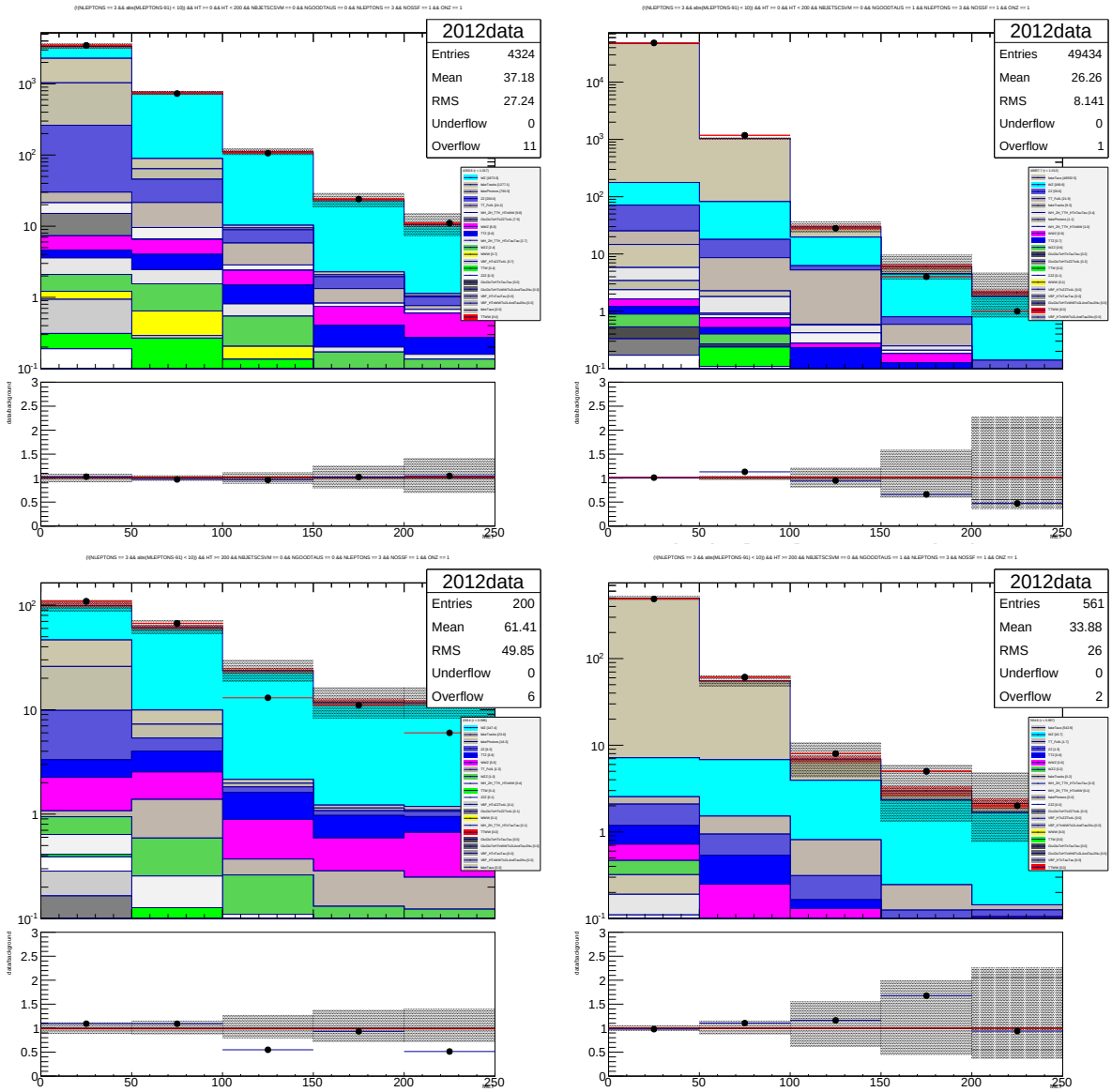

Figure 20: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L3DYz1 region (0 b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

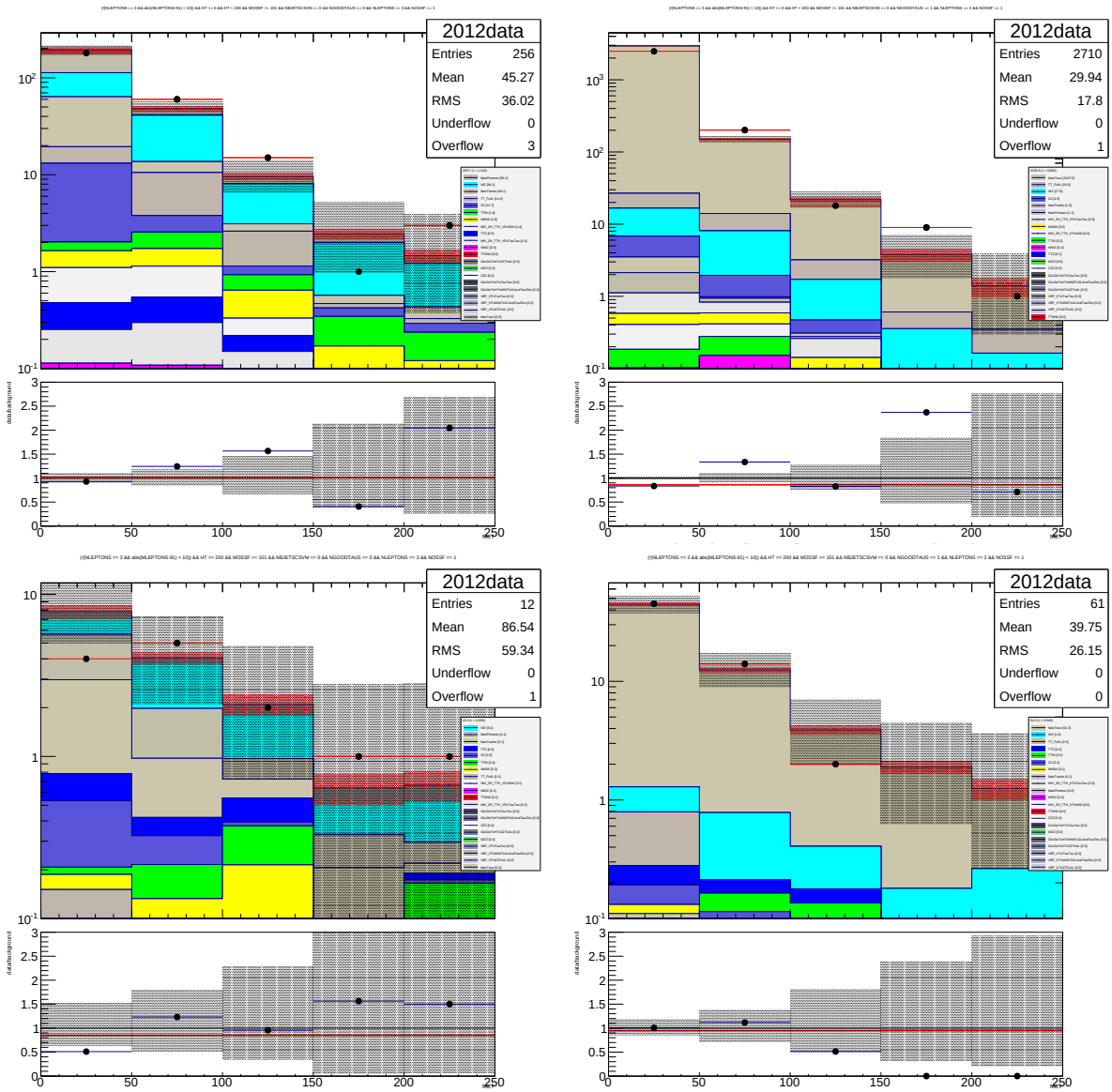

Figure 21: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L3DYh1 region (0 *b*-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

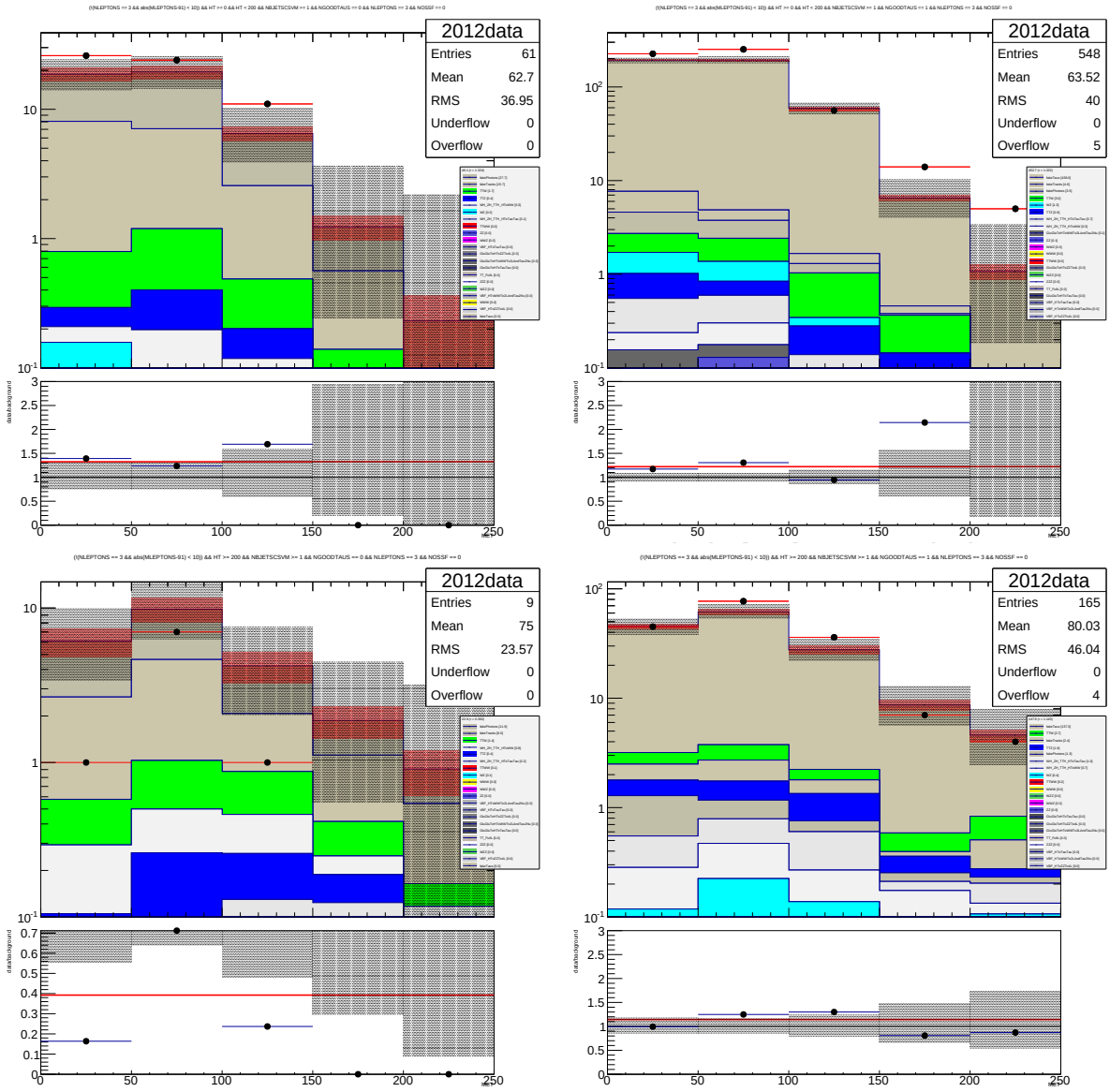

Figure 22: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L3DY0 region (with b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

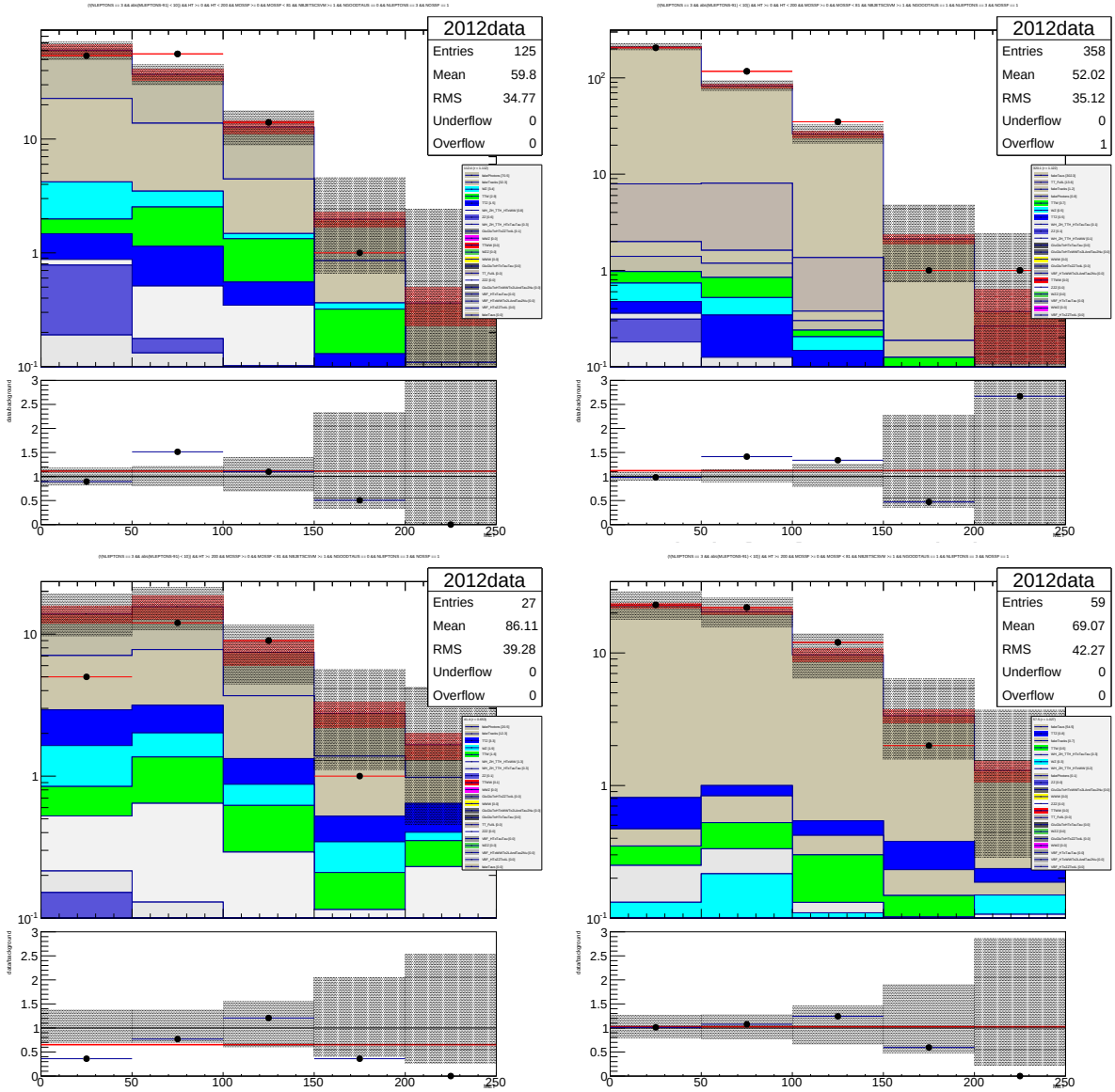

Figure 23: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L3DYl1 region (with b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

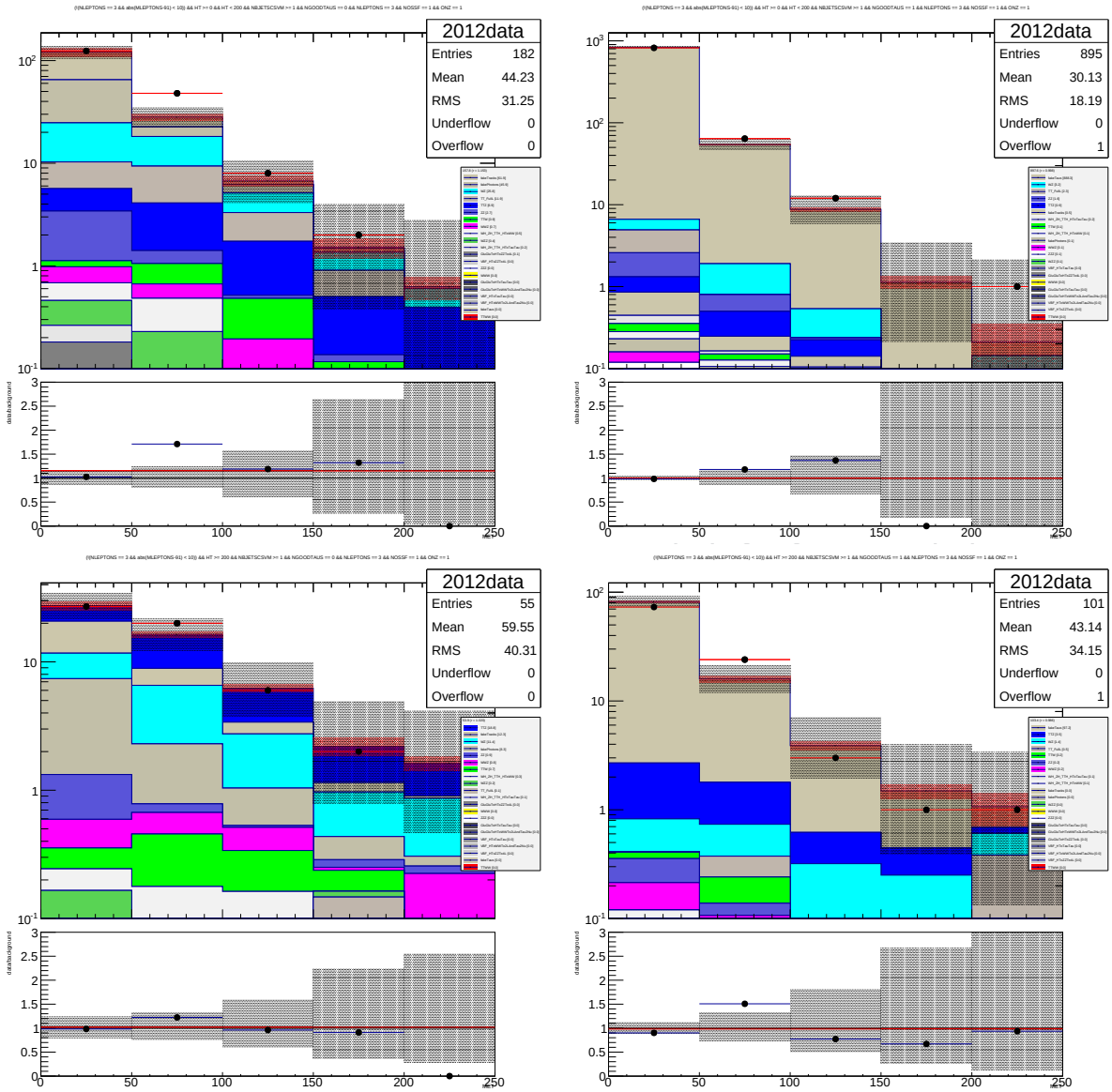

Figure 24: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L3DYz1 region (with b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

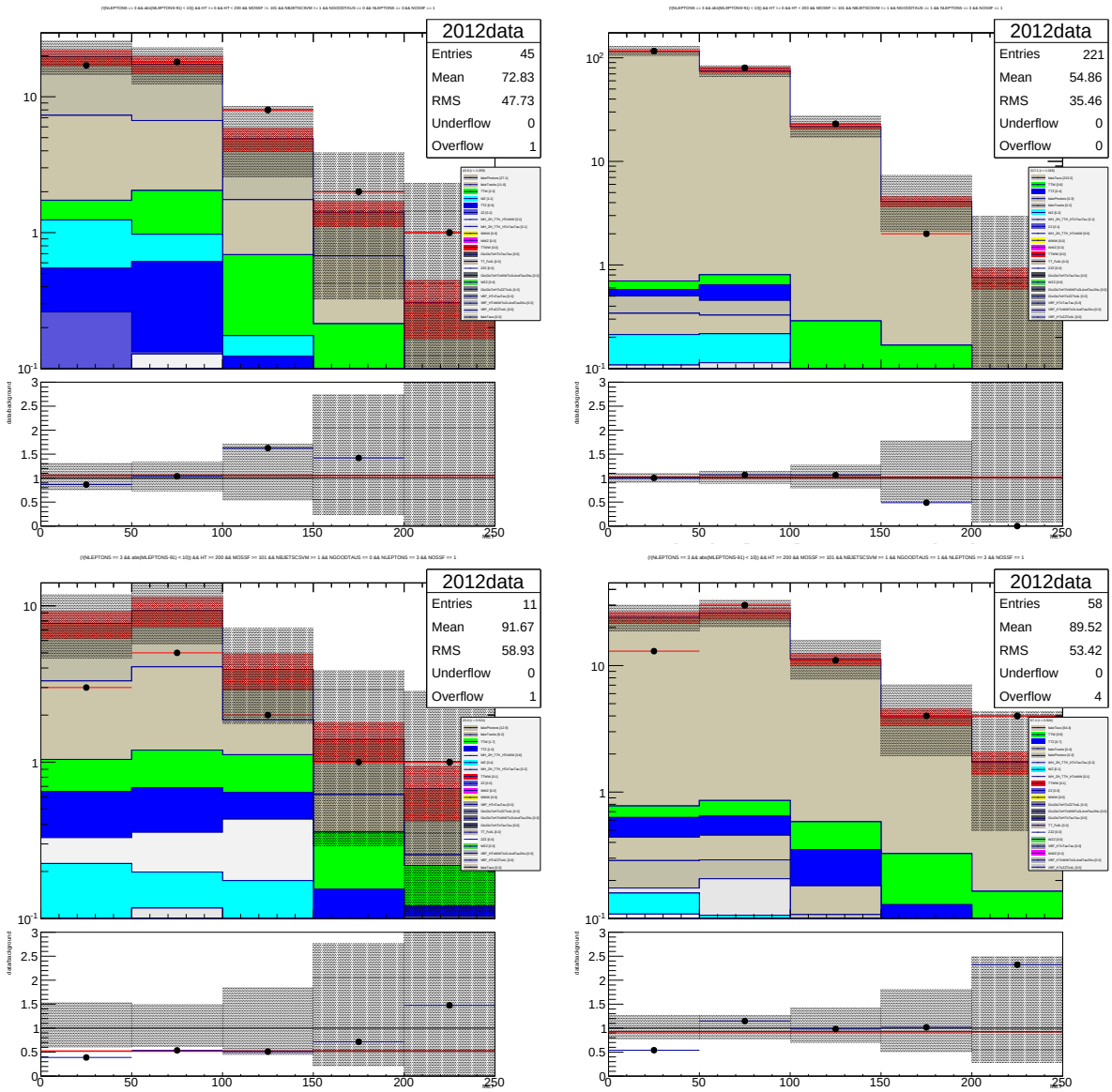

Figure 25: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L3DYh1 region (with *b*-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

7.2 4 or more leptons

DY0: no OSSF pair; DYn1: one pair, but not on Z; DYn2: two pairs, but none on Z; DYz1: one pair, on Z, DYz2: two pairs, at least one on Z

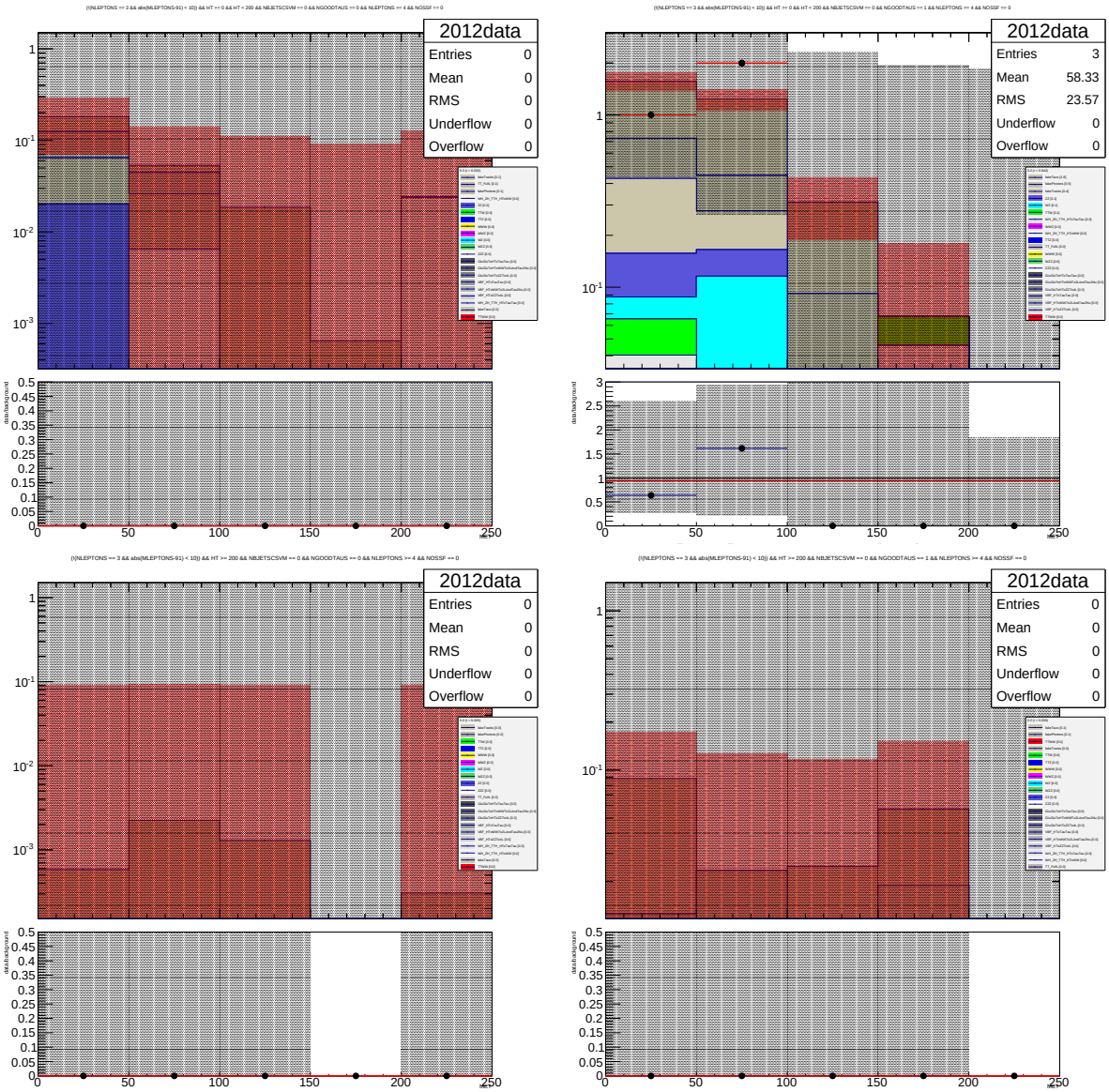

Figure 26: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DY0 region (0 b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

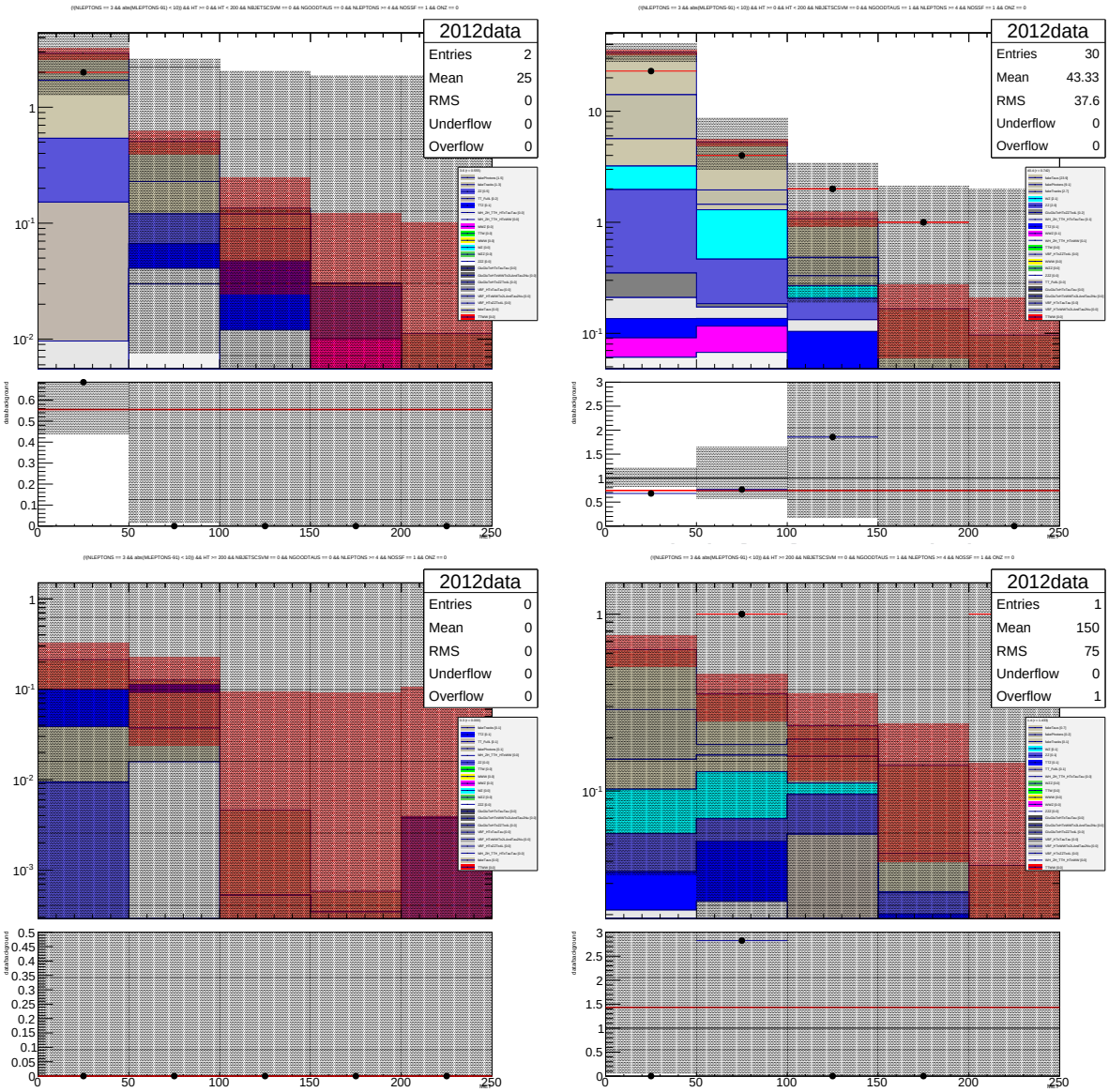

Figure 27: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DYn1 region (0 b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

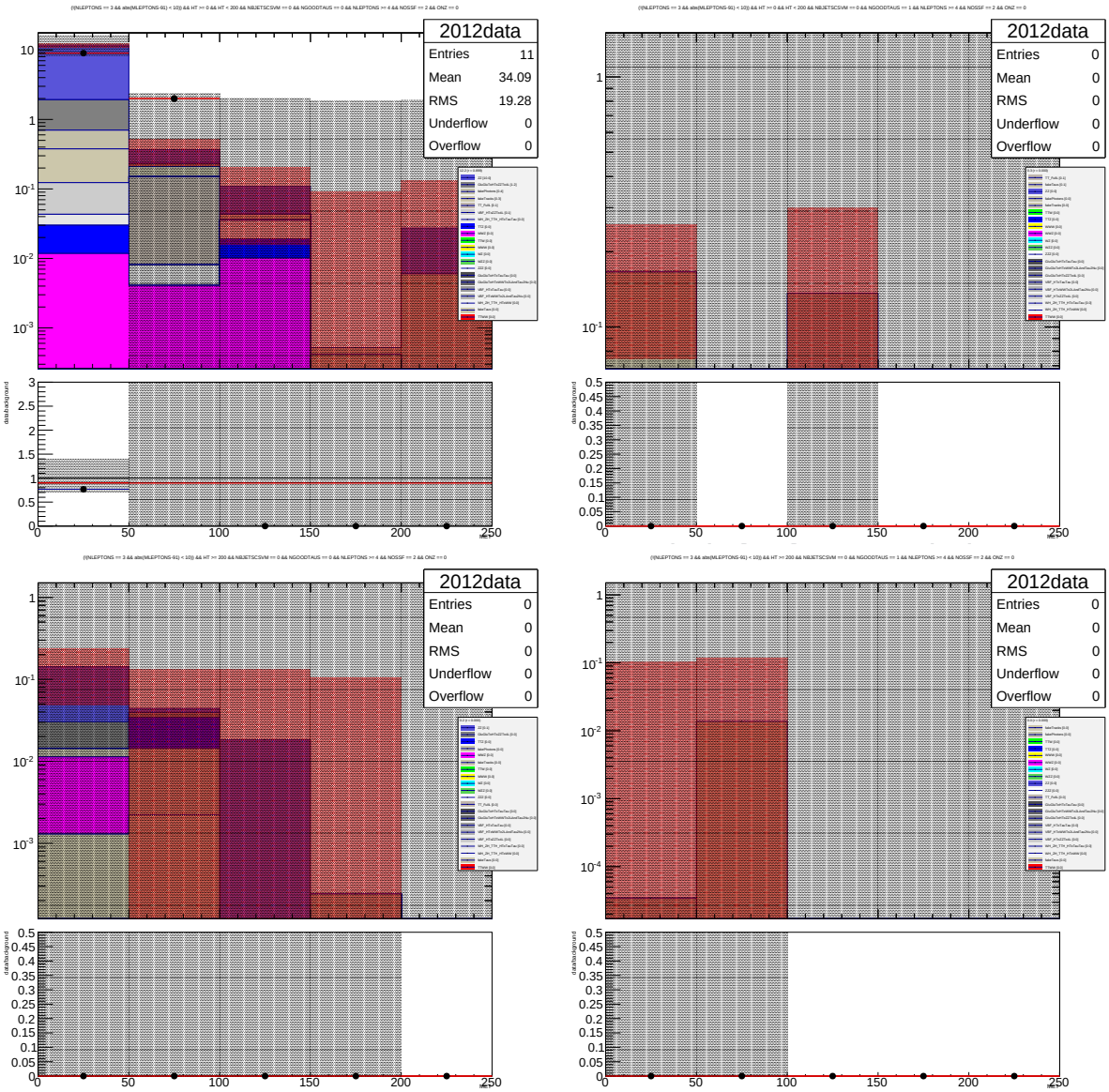

Figure 28: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DYn2 region (0 *b*-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

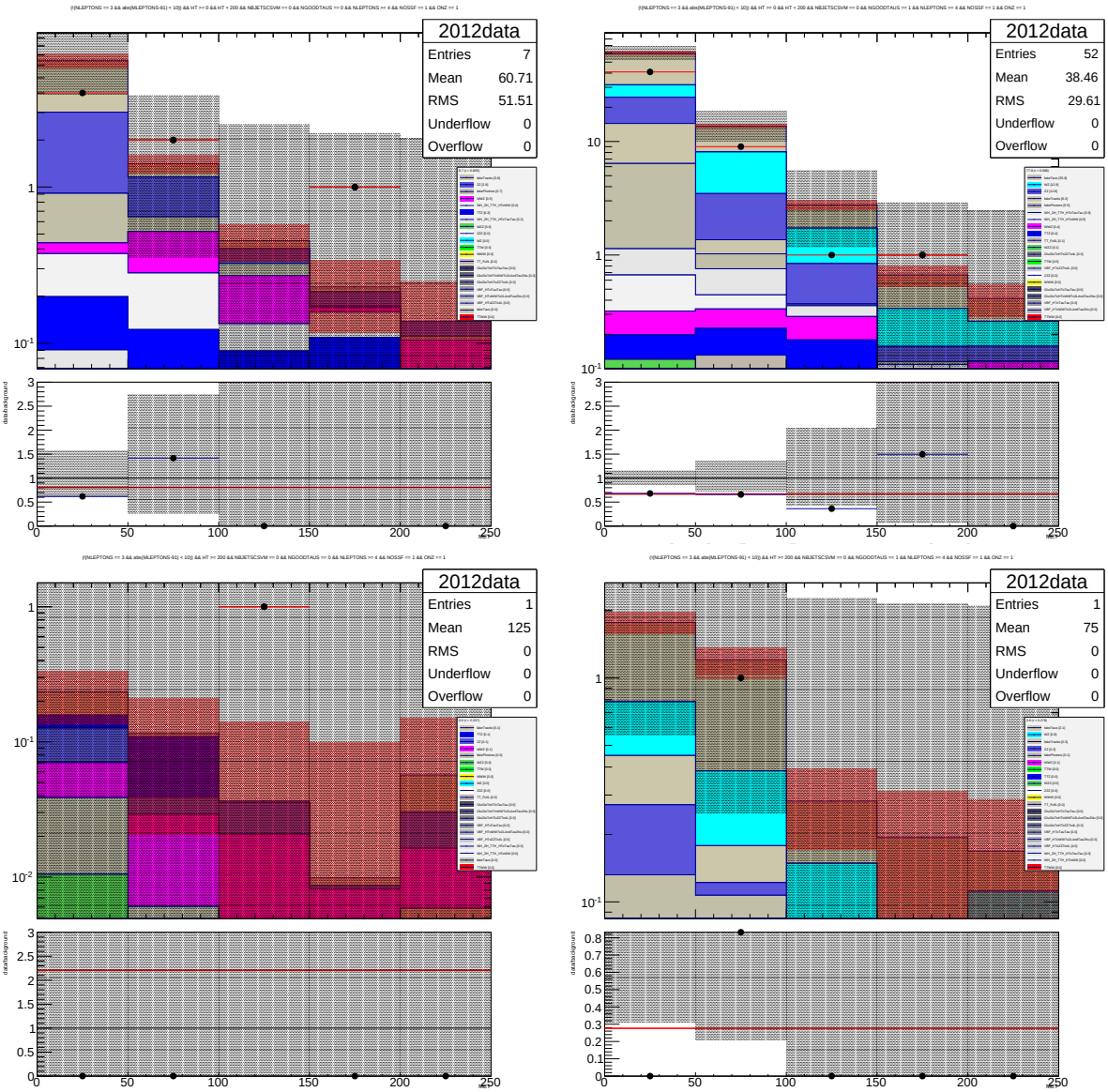

Figure 29: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DYz1 region (0 b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

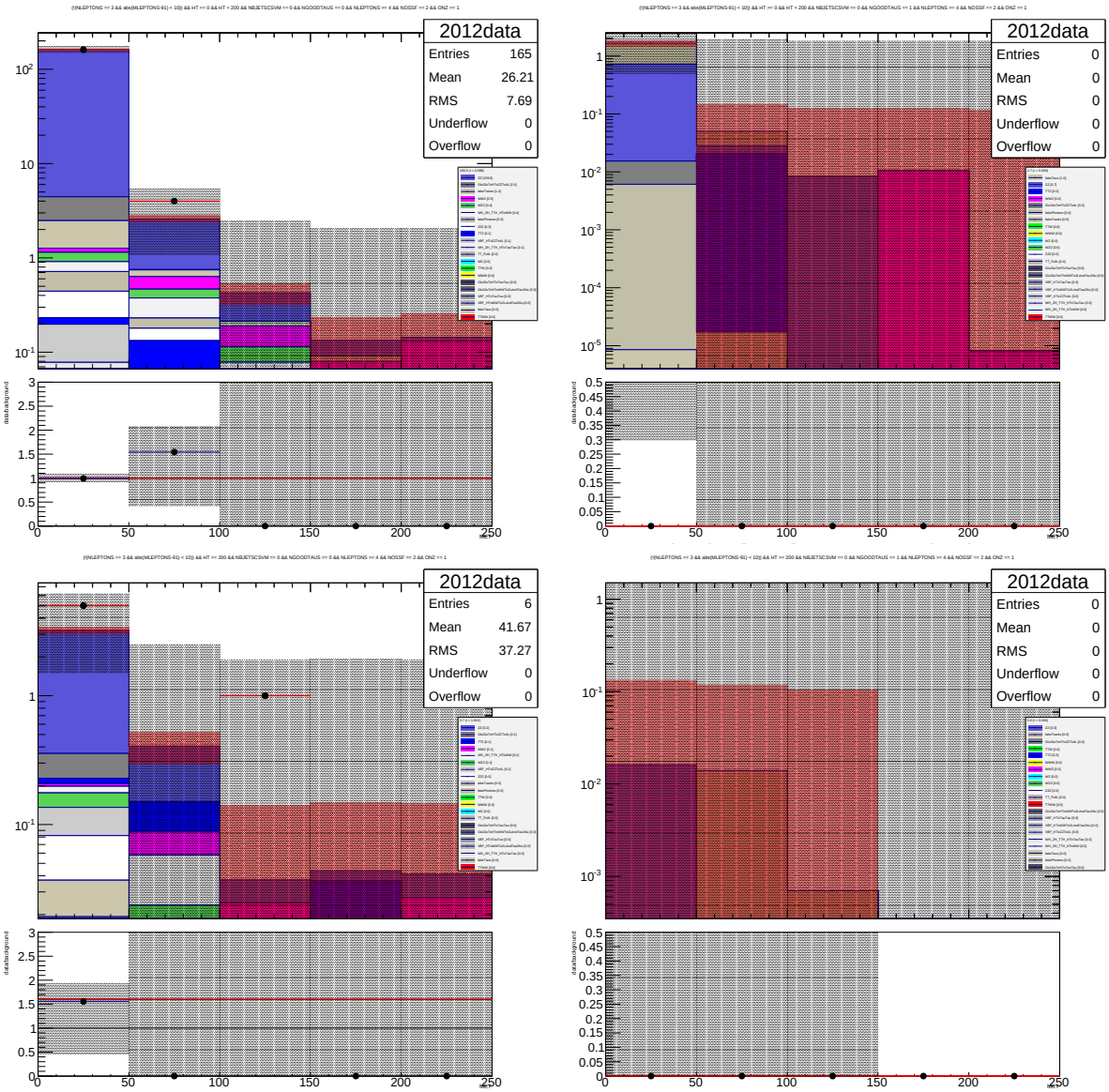

Figure 30: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DYz2 region (0 b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

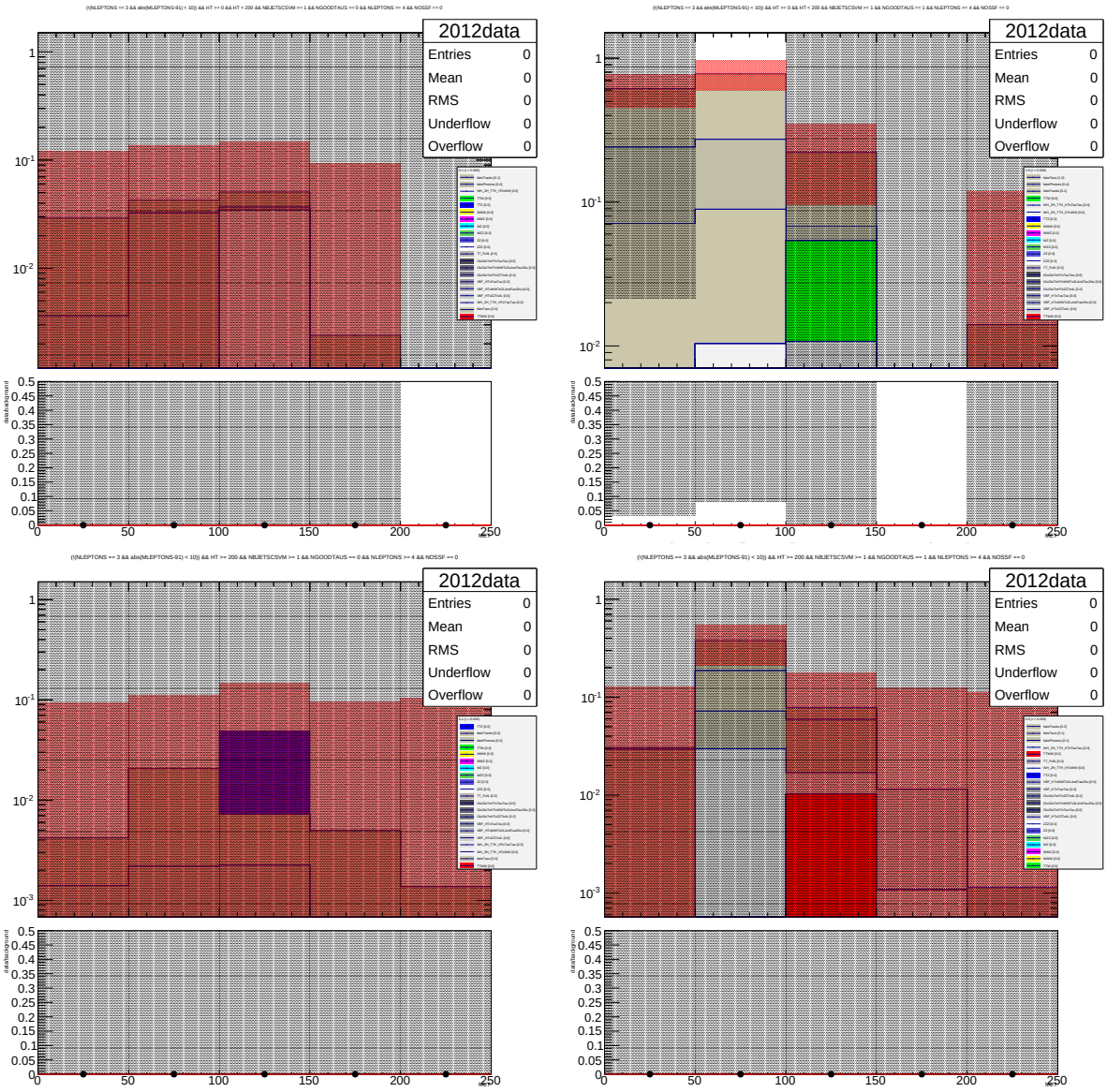

Figure 31: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DY0 region (with b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

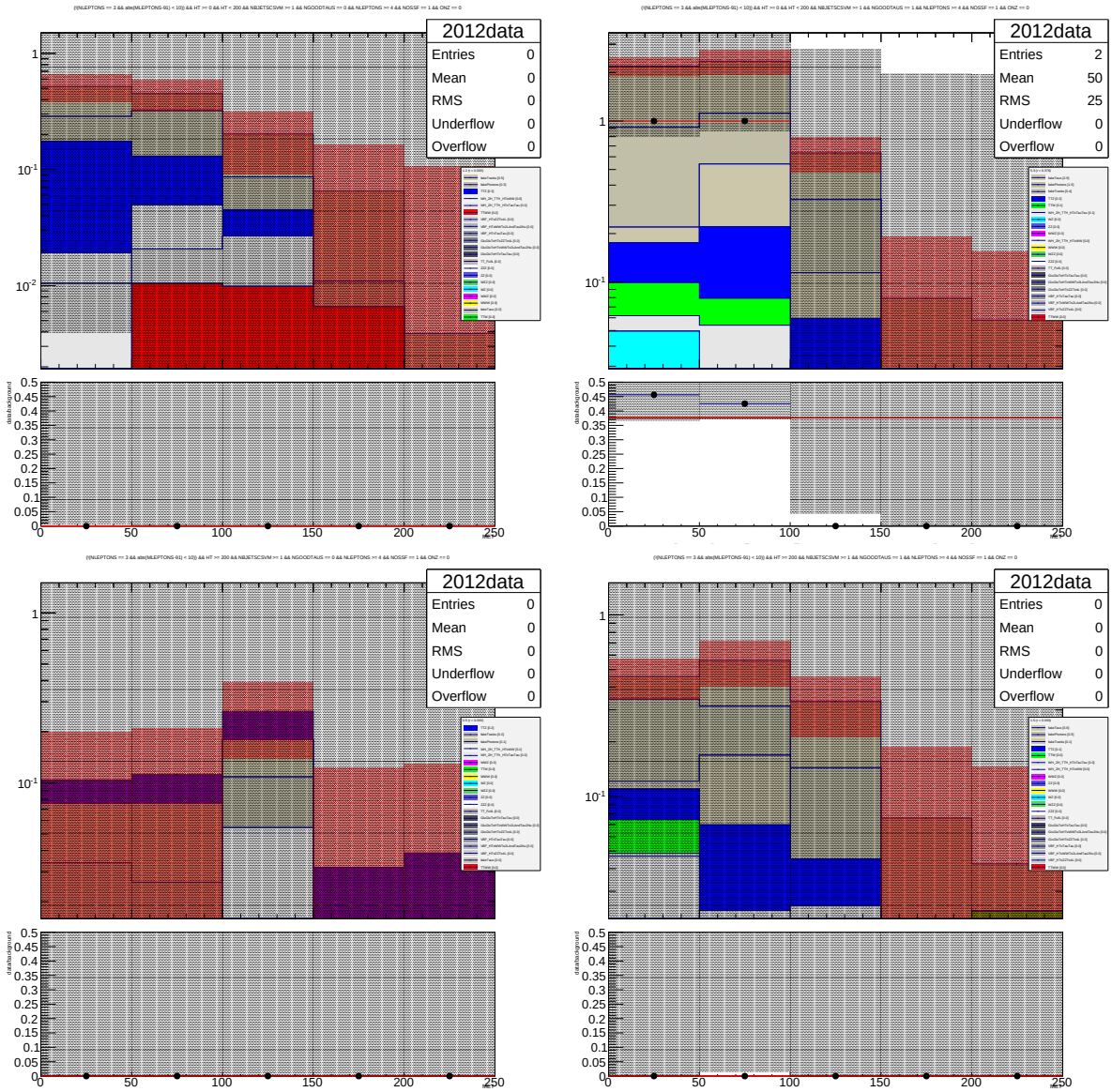

Figure 32: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DYn1 region (with b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

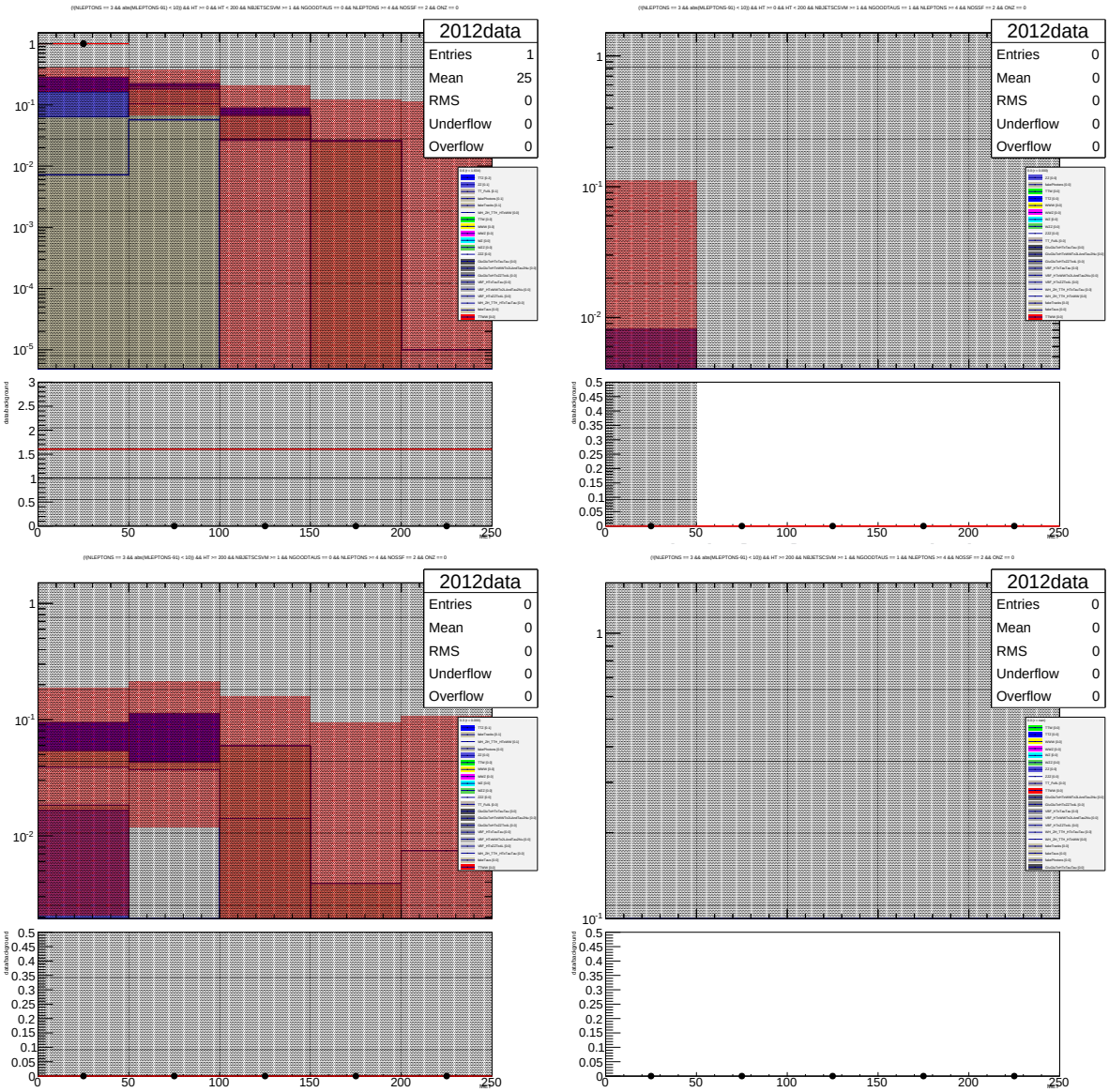

Figure 33: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DYn2 region (with b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

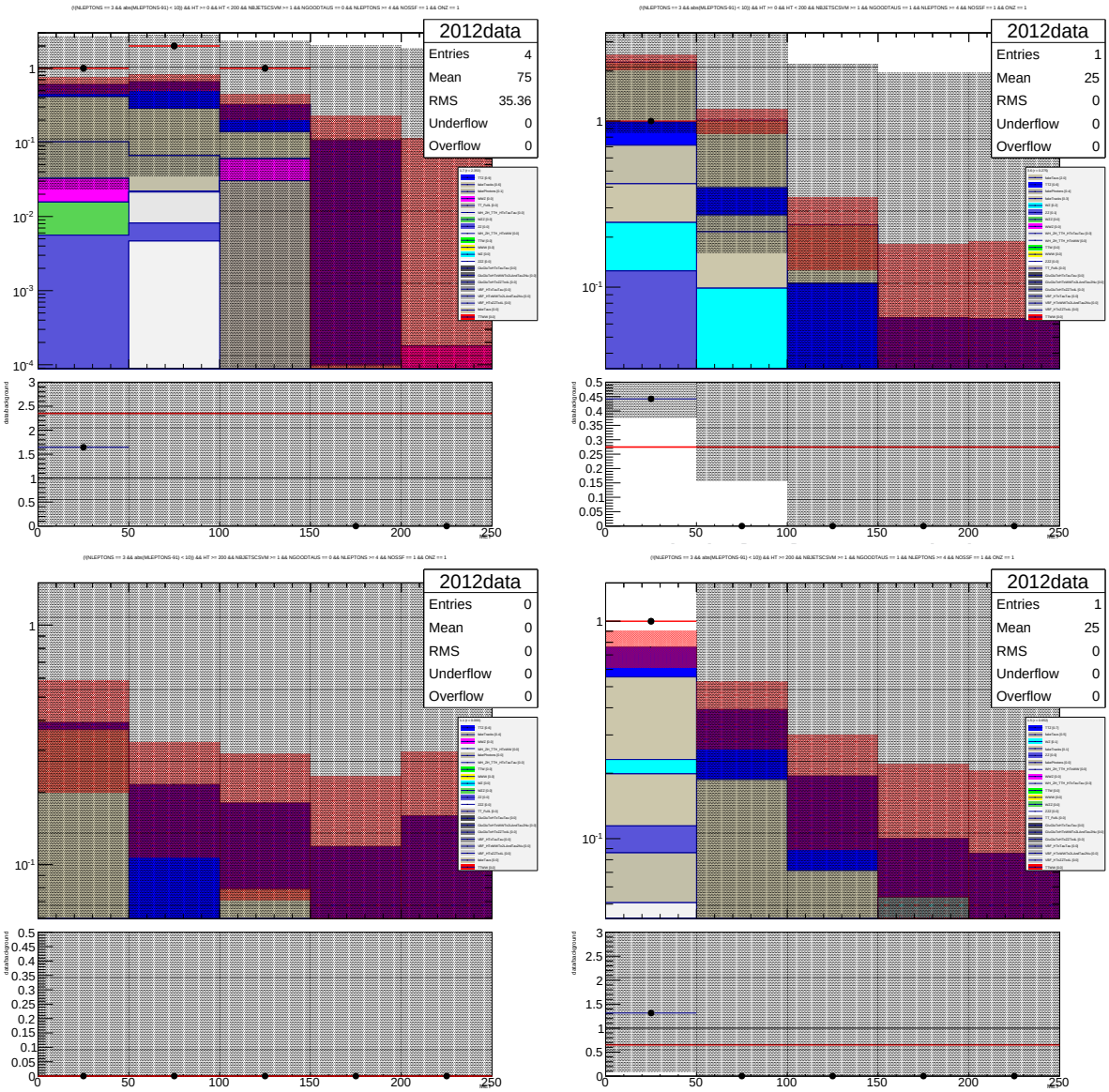

Figure 34: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DYz1 region (with b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

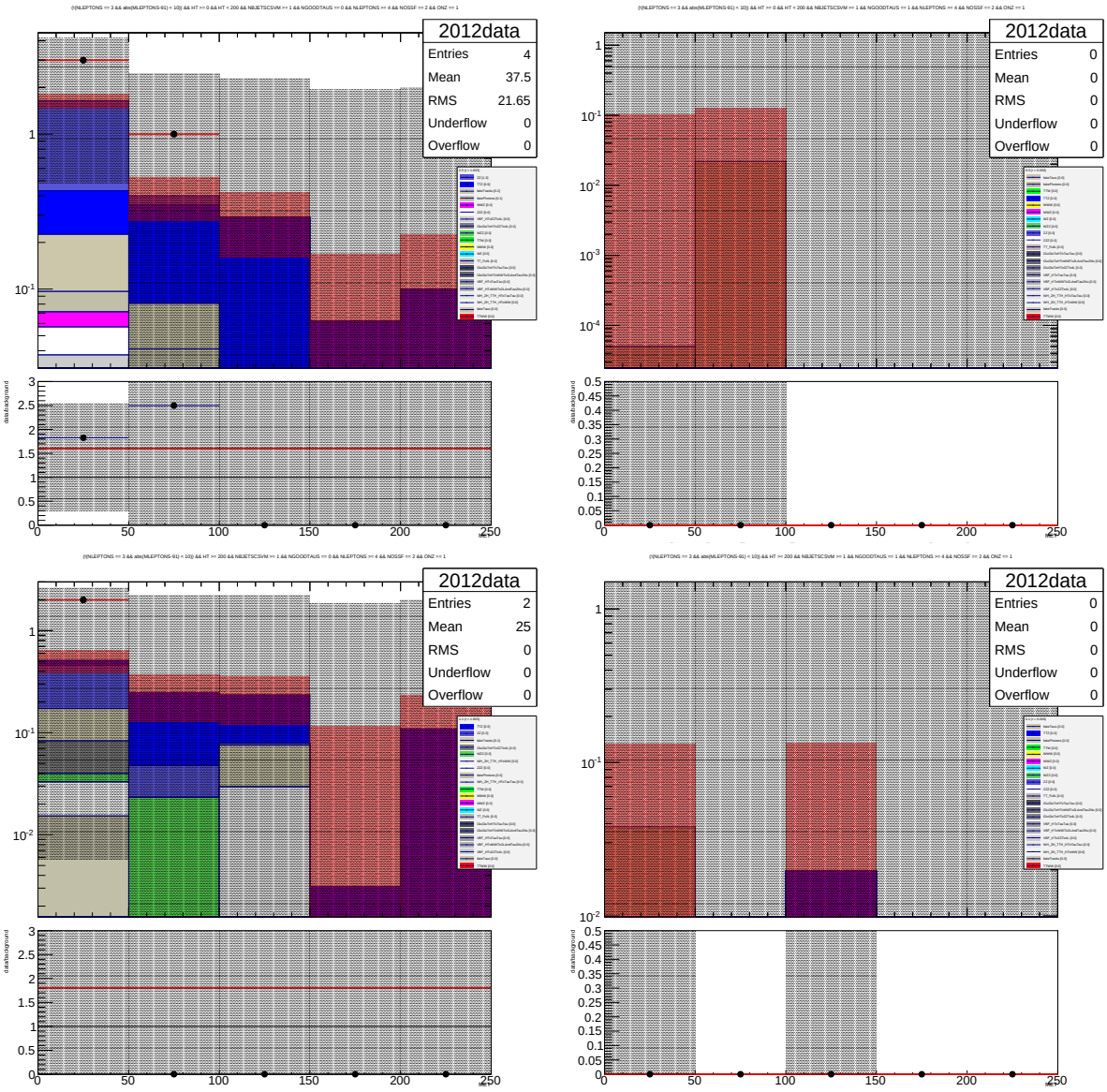

Figure 35: Results: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in L4DYz2 region (with b-tags). Left: 0 taus, right: taus. Top: low HT, bottom: high HT

8 Interpretation and Statistical Treatment

DRAFT

A To do

A loose and incomplete collection of ideas.

A.1 Physics

- Consider narrowing the ZZ control region (require both OSSF pairs to be on-Z, and lower $E_{\mathrm{T}}^{\mathrm{miss}}$?). Shoot for $\sim$20 % error ($\sim$25 events).
- Consider narrowing the WZ control region (apply both $E_{\mathrm{T}}^{\mathrm{miss}}$ and M_{T} cut). We could have a small area in the $E_{\mathrm{T}}^{\mathrm{miss}}$–$M_{\mathrm{T}}$ plane for high statistics, and a larger one for small statistics early in the run.
- AIC study in MC; do the AIC fakes match Z to 3l + soft lepton MC predictions? are we using Z to 4l MC as a rare background MC?
- Systematic errors of background determination

A.2 Infrastructure

- Implement something to avoid using events twice. This would make signal regions exclusive automatically, and remove control region events from signal regions (if things are accessed in the right order).
- Event lists.

B Miscellaneous Studies

B.1 AIC region with DY sample

This is just an attempt to see what happens. The photon background is replaced with the DYJetsToLL_M-50 sample.

Fig. 36: There is a considerable contribution for electrons, but none for muons. This suggests that the 3rd leptons are remnant external conversions. Is that the case? Can we reject them better?

B.2 Z binning

We bin events with an OSSF pair depending on whether the pair invariant mass is in the Z window, or below/above. In case of ambiguity, we need to pick a specific pair. We compare two methods:

1. We take the pair whose invariant mass is closest to the Z mass.
2. We take the pair closest to the Z mass, with the additional condition that pairs above the Z window are not considered if there is a pair below the Z window (thus shifting events from high-Z to low-Z).

Fig. 37 shows that there is little difference to both approaches. We decide to take the second approach to achieve a more separative distribution of background.

DRAFT

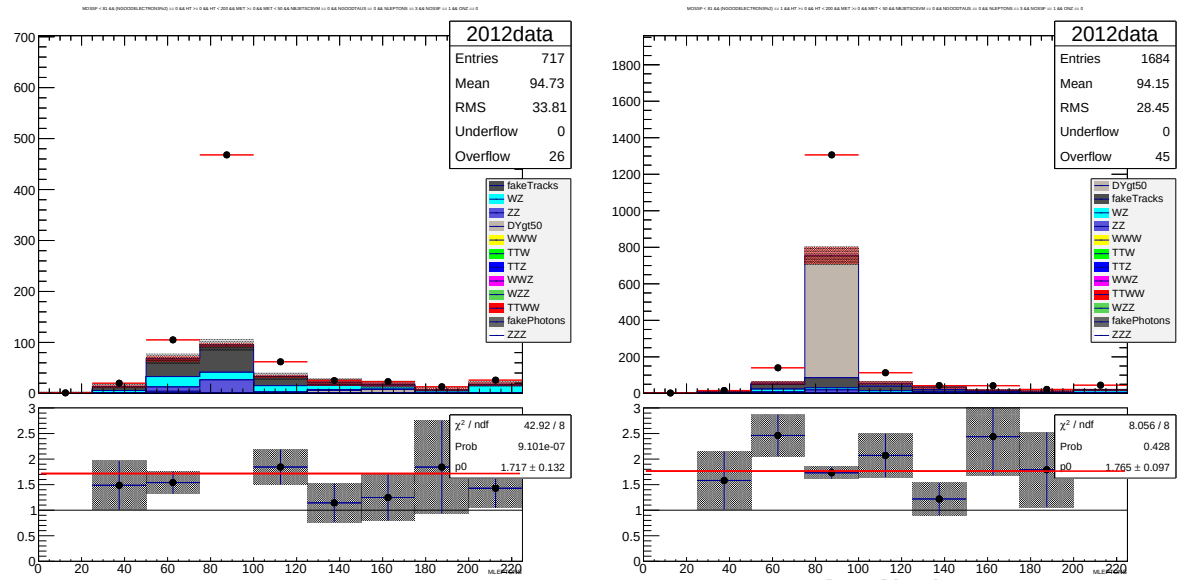

Figure 36: $m_{3\ell}$ distribution in AIC-dominated control region with DYJetsToLL_M-50 sample instead of photon proxies. left) fake muon right) fake electron

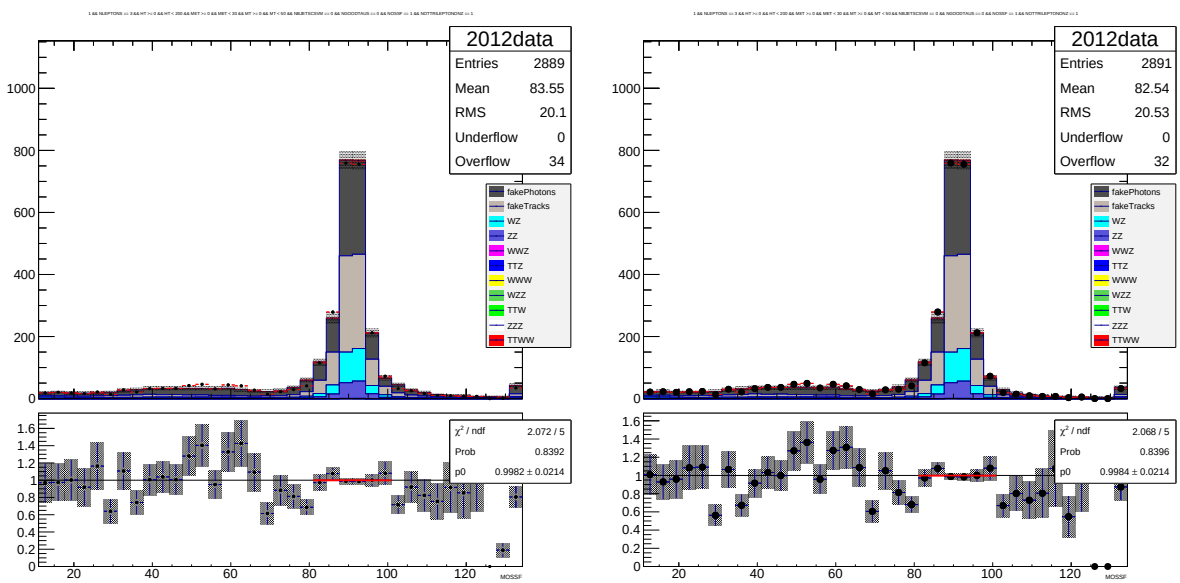

Figure 37: $m_{\ell\ell}$ distribution in in the dilepton fake region (off-Z, trilepton events with $m_{\ell\ell\ell}$ on Z have been vetoed). left) method 1 right) method 2

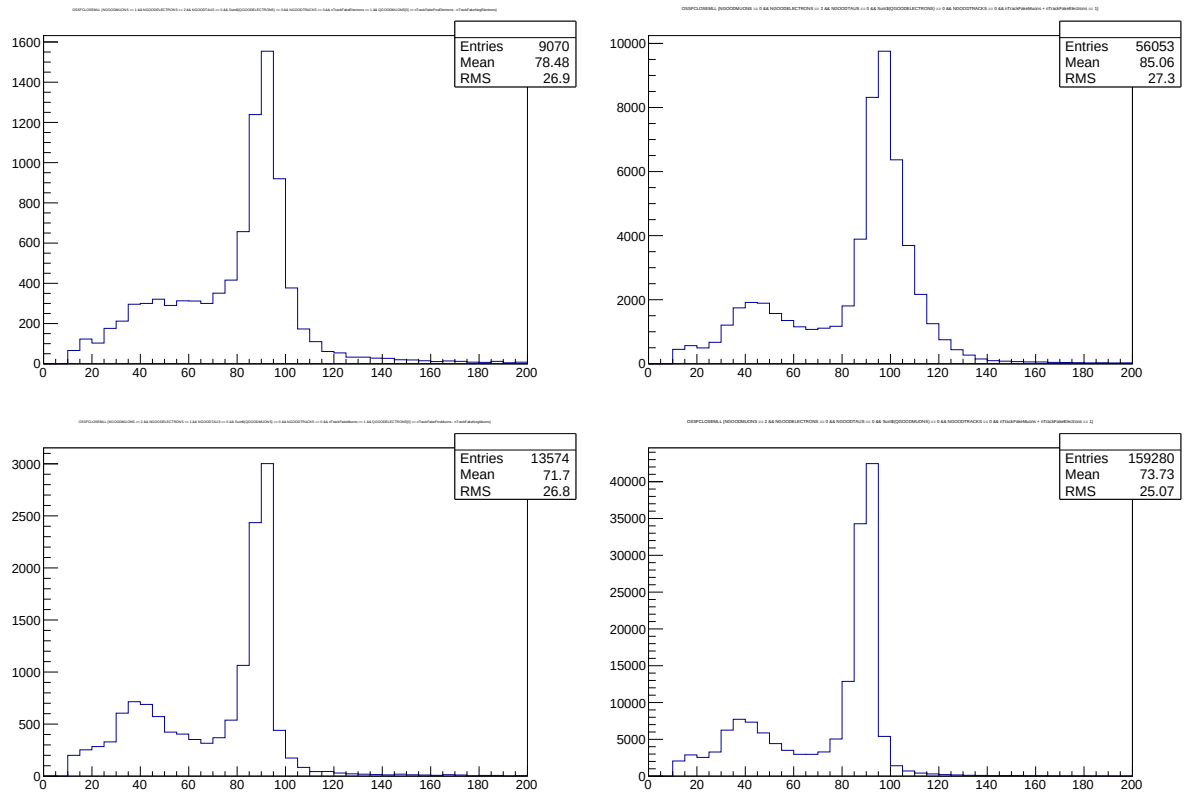

Figure 38: Top: m_{et} distribution; bottom: $m_{\mu t}$ distribution. Left: no other leptons present, right: additional lepton present. If a third lepton is present, its flavor is opposite the flavor of the first lepton, and its charge is the same as the track's (rejecting third leptons from Z).
Note: This is from the dilepton-triggered dataset, i.e. the tracks used here probably were good enough leptons to trigger.

B.3 Suitability of tracks as fake proxies

B.3.1 Low-quality leptons showing up as tracks

Out of curiosity, we look at the invariant mass of an opposite-sign muon+track pair. If the tracks are uncorrelated with the muons, we expect a broad peak at the average invariant mass. If they are correlated, they will be related to their source, for example a Z decay. In fact, we see both (Figure 38).

This suggests that tracks are not only from jets, but also from low quality leptons. So, tracks do model fake leptons from jets, and at the same time also model leptons that were vetoed by quality cuts. However, the general shape of the track background is still expected to be in agreement with the shape of the signal regions. **FIXME:** Does that make sense? Also, understand the left–right difference in Figure 38?

B.3.2 Track p_{T} spectrum

FIXME: TODO

References

[1] CMS Collaboration, "A search for anomalous production of events with three or more leptons using 19.5/fb of sqrt(s)=8 TeV LHC data", CMS Physics Analysis Summary CMS-PAS-SUS-13-002, 2013.

DRAFT