$t\bar{t}H$ $3\ell + \tau$ Run 2 analysis overview

David A. DeMarco, Tina Ojeda, Pierre Savard

April 13, 2017

Contents

1 Object selection

2 Event selection & signal region optimization

Table 1: MC yields in $3\ell + \tau_{\mathrm{had}}$ signal region

$3\ell + \tau_{\mathbf{had}}$ **signal region**	ttH	ttZ	Rare prompt	Nonprompt	ttbar	Sum bkg
Loose leptons	124.63 ± 0.80	530.60 ± 1.96	8407.69 ± 158.88	65597.72 ± 766.32	24430.22 ± 90.19	98966.23 ± 787.80
Tight leptons	55.15 ± 0.54	321.57 ± 1.37	3054.69 ± 35.20	10148.02 ± 155.36	484.23 ± 12.51	14008.51 ± 159.79
Trigger-match	54.85 ± 0.53	321.02 ± 1.37	3021.01 ± 35.06	10086.83 ± 153.61	479.41 ± 12.46	13908.27 ± 158.05
Z-veto	46.81 ± 0.49	80.71 ± 0.71	1758.00 ± 23.77	2698.82 ± 115.74	421.55 ± 11.79	4959.07 ± 118.74
Low-mass veto	46.27 ± 0.49	76.56 ± 0.69	1599.28 ± 22.52	2456.71 ± 111.58	412.77 ± 11.68	4545.32 ± 114.43
NTau==1	2.52 ± 0.09	3.46 ± 1.27	11.46 ± 1.88	8.30 ± 2.46	1.97 ± 0.76	25.19 ± 3.43
Total charge==0	2.33 ± 0.09	3.16 ± 0.15	10.39 ± 1.79	5.72 ± 1.75	1.12 ± 0.63	20.39 ± 2.59
NJet $\geq$2	2.04 ± 0.08	2.77 ± 0.14	2.75 ± 0.70	1.27 ± 0.37	0.38 ± 0.27	7.18 ± 0.85
NBjet $\geq$1	1.68 ± 0.08	2.26 ± 0.13	0.44 ± 0.07	0.30 ± 0.04	0.21 ± 0.21	3.21 ± 0.26
Truth-match tau	1.41 ± 0.06	1.53 ± 0.11	0.31 ± 0.06	0.12 ± 0.03	0.00 ± 0.00	1.96 ± 0.12

3 Fake τ_{had} estimate

3.1 Fake factor method

A fake factor method is used to estimate the background contribution from processes with a fake hadronic tau. The fake τ_{had} estimate is extrapolated from a region that is enriched in the relevant backgrounds, primarily $t\bar{t}$, using a fake factor derived in a high-stats $2\ell OS + \tau_{had}$ region. The fake factor represents the probability of a τ_{had} that passes a loosened selection to pass the tight selection. It is parametrized in bins of p_T of the τ_{had} and is computed as the ratio of events with a good τ_{had} and those with an anti-τ_{had}.

$$FF(p_T) = \frac{N_{\tau}(p_T)}{N_{\not\tau}(p_T)} \tag{1}$$

Table 2: Fake factor derivation regions

Good τ_{had} (numerator) region	Anti-τ_{had} (denominator) region
2ℓ(OS) + τ_{had} == 2 jets, >= 1 b-jets τ_{had} (Medium τ_{had} ID)	2ℓ(OS) + $\not\tau_{had}$ == 2 jets, >= 1 b-jets $\not\tau_{had}$ (Very Loose τ_{had} ID and not Medium)

The 2-jet selection above ensures orthogonality with the $2\ell(\mathrm{OS})+\tau_{had}$ channel signal region and can be used for extrapolation because the fake factor is flat with respect to jet multiplicity (ADD FIGURE AND REFERENCE). The p_T-dependent fake factor is applied to a $3\ell + \not\tau_{had}$ sideband extrapolation region which has an identical selection to the signal region but with an anti-τ_{had}.

In Table 4, the MC yields are listed for the fake factor derivation and extrapolation regions and for the $3\ell + \tau_{had}$ signal region. Background processes listed as prompt are those which may yield a $3\ell + \tau_{had}$ final state without an object faking a light lepton or τ_{had} while non-prompt processes

Table 3: Fake estimate extrapolation region

Extrapolation sideband region
$3\ell + \not\tau_{had}$ $>=$ 2 jets, $>=$ 1 b-jets $\not\tau_{had}$ (Very Loose τ_{had} ID and not Medium)

are those which require a faked object. The full list of processes, samples and their categoration can be found in Table 10.

Table 4: Nominal Monte Carlo yields for fake estimate regions and for the $3\ell + \tau_{had}$ signal region ($t\bar{t}$ from Powheg+Pythia8 non-all hadronic low stats). Entries labeled "true" require the τ_{had} to be truth-matched.

Process	Numerator	Denominator	Extrapolation	SR MC
$t\bar{t}H$	2.467 ± 0.080	4.399 ± 0.108	2.278 ± 0.110	1.675 ± 0.078
Prompt bkg	4.084 ± 0.273	9.589 ± 0.608	3.716 ± 0.706	2.700 ± 0.147
Prompt bkg. (true)	2.906 ± 0.234	1.643 ± 0.338	0.723 ± 0.074	1.842 ± 0.122
Non-prompt bkg.	920.07 ± 21.20	8647.67 ± 59.08	6.250 ± 1.200	0.510 ± 0.213
Non-prompt bkg (true)	34.59 ± 3.07	199.71 ± 8.53	0.051 ± 0.011	0.117 ± 0.028

3.2 MC estimate results and closure tests

To check the closure of this procedure, the extrapolation is performed and the integrated value compared with the signal region yield in MC. Closure is calculated as:

$$\frac{\text{(MC yield - Fake estimate)}}{\text{Fake estimate}} \times 100\% \tag{2}$$

3.2.1 Estimate w. background subtraction

The fake factor is calculated for fake τ_{had} by subtracting all signal $t\bar{t}H$ events and background events which are matched to a true τ_{had} from the total Monte Carlo (or data) yield:

$$FF(p_T)_{\text{MC (data)}} = \frac{N_\tau(p_T)^{\text{All MC (data)}} - N_\tau(p_T)^{\text{Truth-matched MC}} - N_\tau(p_T)^{t\bar{t}H\text{MC}}}{N_{\not\tau}(p_T)^{\text{All MC (data)}} - N_{\not\tau}(p_T)^{\text{Truth-matched MC}} - N_{\not\tau}(p_T)^{t\bar{t}H\text{MC}}} \tag{3}$$

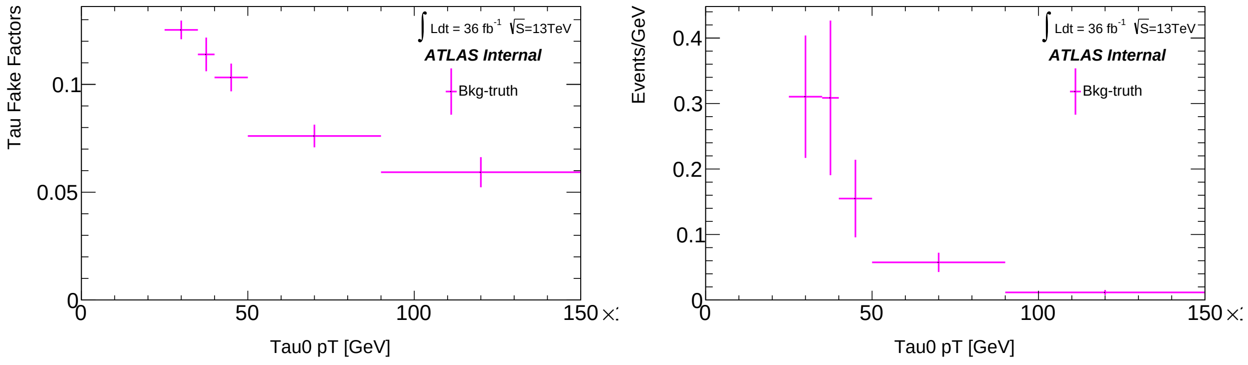
The estimate from fake taus in the signal region is then computed by applying these fake factors to the $3\ell + \not\tau$ extrapolation region:

$$N_\tau(p_T)^{\text{fakes}} = FF(p_T)_{\text{MC (data)}} \cdot \left[N_{\not\tau}(p_T)^{\text{All MC (data)}} - N_{\not\tau}(p_T)^{\text{Truth-matched MC}} - N_{\not\tau}(p_T)^{t\bar{t}H\text{MC}}\right] \tag{4}$$

Results are shown below using four different Powheg+Pythia8 $t\bar{t}$ samples. The PP8 sample listed (DSID 410501) is the baseline, and is compared with a dilepton-filtered sample (DSID 410503) and two higher-stats productions. Figures are produced using the non-all hadronic high-stats sample.

Table 5: Extrapolation and closure test, (All MC - Truth-matched MC - $t\bar{t}H$)

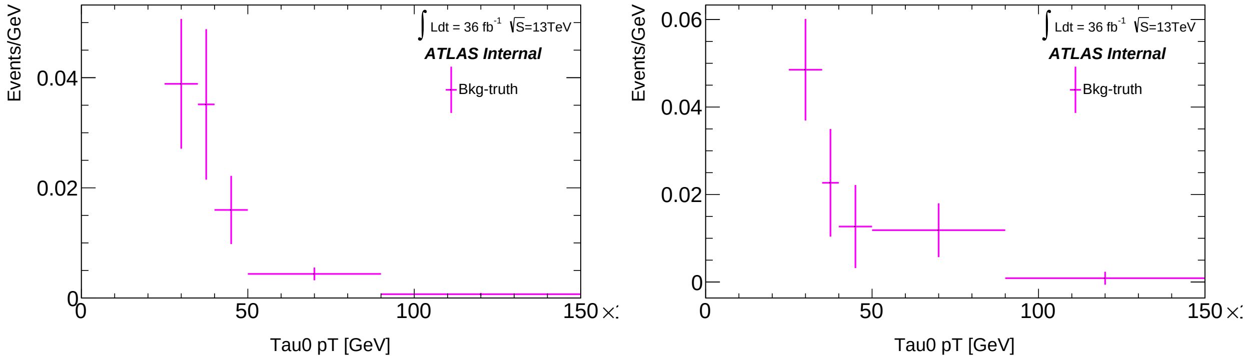
$t\bar{t}$ sample	Integrated fake estimate	SR MC	Closure
PP8 non-all hadronic	0.869 ± 0.171	1.446 ± 0.462	$66 \pm 62\%$
PP8 non-all hadronic, high stats	0.941 ± 0.157	1.252 ± 0.308	$33 \pm 40\%$
PP8 dilepton	0.733 ± 0.136	1.273 ± 0.323	$74 \pm 55\%$
PP8 dilepton, high stats	0.806 ± 0.125	1.447 ± 0.327	$80 \pm 49\%$



(a) p_T-parametrized fake factors

(b) Extrapolation sideband region p_T spectrum

Figure 1: Fake factors and extrapolation sideband region, (All MC - Truth-matched MC - $t\bar{t}H$)



(a) Fake estimate in p_T bins

(b) MC SR yield in p_T bins

Figure 2: p_T spectra for fake estimate and MC yield, (All MC - Truth-matched MC - $t\bar{t}H$)

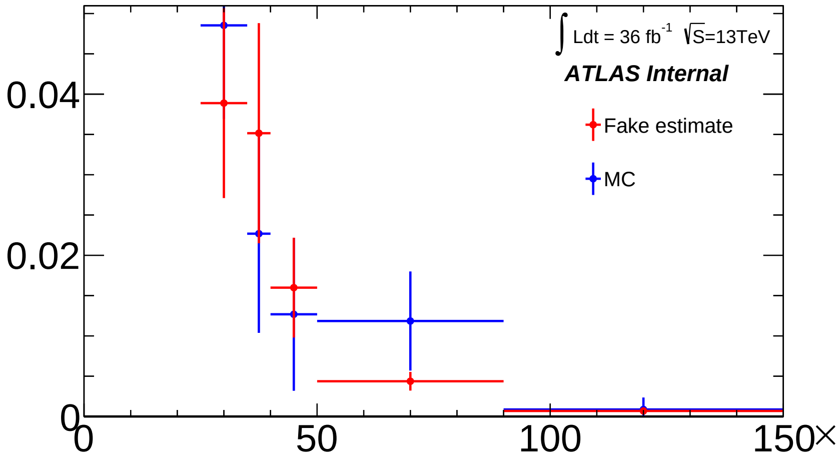
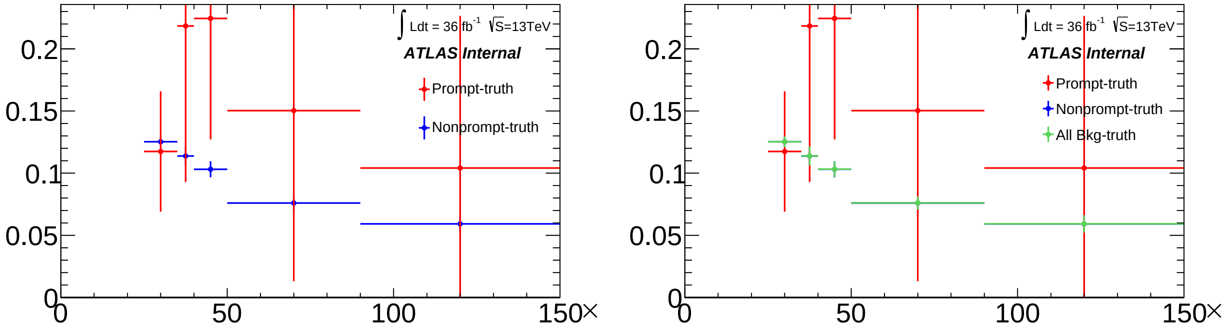


Figure 3: Overlaid fake estimate and MC prediction (All MC - Truth-matched MC - $t\bar{t}H$)



(a) Prompt (red) and non-prompt (blue)

(b) Prompt, non-prompt and combined (green)

Figure 4: Comparison of fake factors derived from prompt MC, non-prompt MC and from the full combined (prompt + non-prompt - truth) method

A Appendix: Additional studies

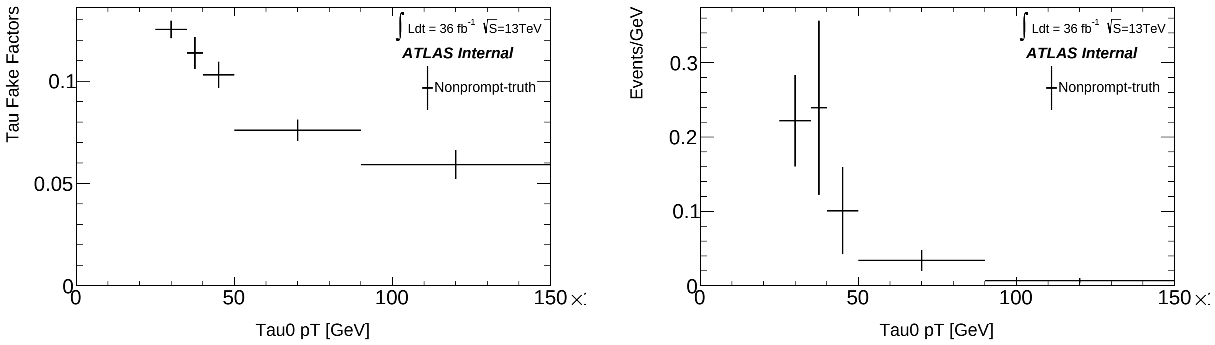
A.1 Estimate with non-prompt MC - truth-matched

$$FF(p_T)_{MC} = \frac{N_{\tau}(p_T)^{\text{Non-prompt MC}} - N_{\tau}(p_T)^{\text{Truth-matched non-prompt MC}}}{N_{\not\tau}(p_T)^{\text{Non-prompt MC}} - N_{\not\tau}(p_T)^{\text{Truth-matched non-prompt MC}}} \tag{5}$$

Table 6: Extrapolation and closure test, (Non-prompt MC - truth-matched)

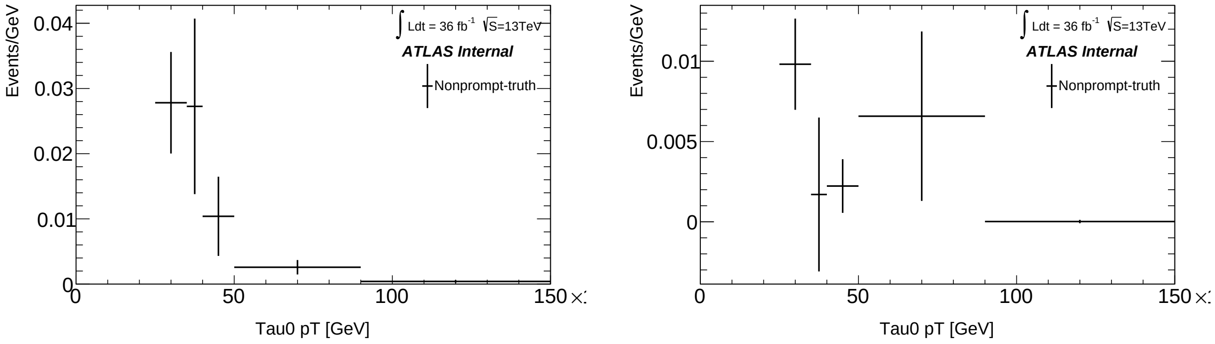
$t\bar{t}$ sample	Integrated fake estimate	SR MC	Raw, total ($t\bar{t}$)	Closure
PP8 non-all hadronic	0.574 ± 0.143	0.587 ± 0.406	138 (1)	$2 \pm 75\%$
PP8 non-all hadronic, high stats	0.646 ± 0.128	0.393 ± 0.215	138 (1)	$-39 \pm 35\%$
PP8 dilepton	0.442 ± 0.104	0.414 ± 0.236	138 (1)	$-6 \pm 58\%$
PP8 dilepton, high stats	0.519 ± 0.088	0.588 ± 0.242	140 (3)	$13 \pm 51\%$

Figures are produced using the non-all hadronic high-stats sample.



(a) p_T-parametrized fake factors (b) Extrapolation sideband region p_T spectrum

Figure 5: Fake factors and extrapolation sideband region, (Non-prompt MC - truth)



(a) Fake estimate in p_T bins (b) MC SR yield in p_T bins

Figure 6: p_T spectra for fake estimate and MC yield, (Non-prompt MC - truth)

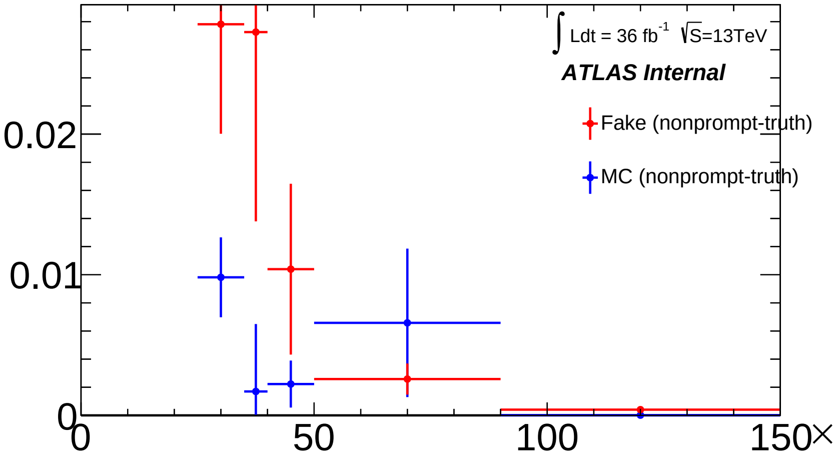


Figure 7: Overlaid fake estimate and MC prediction (Non-prompt MC - truth)

Table 7: Raw yield comparison for signal region comparing low-stats $t\bar{t}$ non-all hadronic and dilepton low-stats samples

	ttH	Prompt bkg	Nonprompt bkg (no ttbar)	ttbar non-all hadronic	ttbar dilepton
Input	7196834.000 ± 2682.692	5923407.000 ± 2433.805	155015314.000 ± 12450.515	11452200.000 ± 3384.110	6562556.000 ± 2561.749
CutBlind	7196834.000 ± 2682.692	5923407.000 ± 2433.805	155015314.000 ± 12450.515	11452200.000 ± 3384.110	6562556.000 ± 2561.749
Cleaning	7196834.000 ± 2682.692	5923407.000 ± 2433.805	155015314.000 ± 12450.515	11452200.000 ± 3384.110	6562556.000 ± 2561.749
CutSLORDLTrigger	5124841.000 ± 2263.811	4621965.000 ± 2149.876	120080497.000 ± 10958.125	7944424.000 ± 2818.585	5279307.000 ± 2297.674
Cut3Leptons	247338.000 ± 497.331	770943.000 ± 878.034	1235939.000 ± 1111.728	51759.000 ± 227.506	80321.000 ± 283.410
CutLepTightMVA	103371.000 ± 321.514	466468.000 ± 682.985	574917.000 ± 758.233	996.000 ± 31.559	1639.000 ± 40.485
CutTrigFlat	102924.000 ± 320.818	465119.000 ± 681.996	574130.000 ± 757.714	987.000 ± 31.417	1624.000 ± 40.299
CutZCandVeto	88789.000 ± 297.975	121906.000 ± 349.150	105100.000 ± 324.191	869.000 ± 29.479	1409.000 ± 37.537
CutLowMass12	87981.000 ± 296.616	115654.000 ± 340.079	102436.000 ± 320.056	849.000 ± 29.138	1390.000 ± 37.283
CutNTau	5888.000 ± 76.733	4825.000 ± 69.462	896.000 ± 29.933	4.000 ± 2.000	4.000 ± 2.000
Cut3LepCharge	5519.000 ± 74.290	4080.000 ± 63.875	613.000 ± 24.759	4.000 ± 2.000	3.000 ± 1.732
CutNJet	4991.000 ± 70.647	3004.000 ± 54.809	356.000 ± 18.868	2.000 ± 1.414	1.000 ± 1.000
Cut1BTag	4113.000 ± 64.133	2173.000 ± 46.615	237.000 ± 15.395	1.000 ± 1.000	1.000 ± 1.000
CutTauTruthMatch	3627.000 ± 60.225	1354.000 ± 36.797	100.000 ± 10.000	0.000 ± 0.000	0.000 ± 0.000

Table 8: Raw yield comparison for signal region comparing low-stats $t\bar{t}$ non-all hadronic and dilepton high-stats samples

	ttH	Prompt bkg	Nonprompt bkg (no ttbar)	ttbar non-all hadronic	ttbar dilepton
Input	7196834.000 ± 2682.692	5923407.000 ± 2433.805	155015314.000 ± 12450.515	22070812.000 ± 4697.958	12822814.000 ± 3580.896
CutBlind	7196834.000 ± 2682.692	5923407.000 ± 2433.805	155015314.000 ± 12450.515	22070812.000 ± 4697.958	12822814.000 ± 3580.896
Cleaning	7196834.000 ± 2682.692	5923407.000 ± 2433.805	155015314.000 ± 12450.515	22070812.000 ± 4697.958	12822814.000 ± 3580.896
CutSLORDLTrigger	5124841.000 ± 2263.811	4621965.000 ± 2149.876	120080497.000 ± 10958.125	15309277.000 ± 3912.707	10316107.000 ± 3211.870
Cut3Leptons	247338.000 ± 497.331	770943.000 ± 878.034	1235939.000 ± 1111.728	99709.000 ± 315.767	156885.000 ± 396.087
CutLepTightMVA	103371.000 ± 321.514	466468.000 ± 682.985	574917.000 ± 758.233	1985.000 ± 44.553	3263.000 ± 57.123
CutTrigFlat	102924.000 ± 320.818	465119.000 ± 681.996	574130.000 ± 757.714	1963.000 ± 44.306	3235.000 ± 56.877
CutZCandVeto	88789.000 ± 297.975	121906.000 ± 349.150	105100.000 ± 324.191	1718.000 ± 41.449	2819.000 ± 53.094
CutLowMass12	87981.000 ± 296.616	115654.000 ± 340.079	102436.000 ± 320.056	1682.000 ± 41.012	2781.000 ± 52.735
CutNTau	5888.000 ± 76.733	4825.000 ± 69.462	896.000 ± 29.933	11.000 ± 3.317	15.000 ± 3.873
Cut3LepCharge	5519.000 ± 74.290	4080.000 ± 63.875	613.000 ± 24.759	7.000 ± 2.646	10.000 ± 3.162
CutNJet	4991.000 ± 70.647	3004.000 ± 54.809	356.000 ± 18.868	2.000 ± 1.414	5.000 ± 2.236
Cut1BTag	4113.000 ± 64.133	2173.000 ± 46.615	237.000 ± 15.395	1.000 ± 1.000	3.000 ± 1.732
CutTauTruthMatch	3627.000 ± 60.225	1354.000 ± 36.797	100.000 ± 10.000	0.000 ± 0.000	0.000 ± 0.000

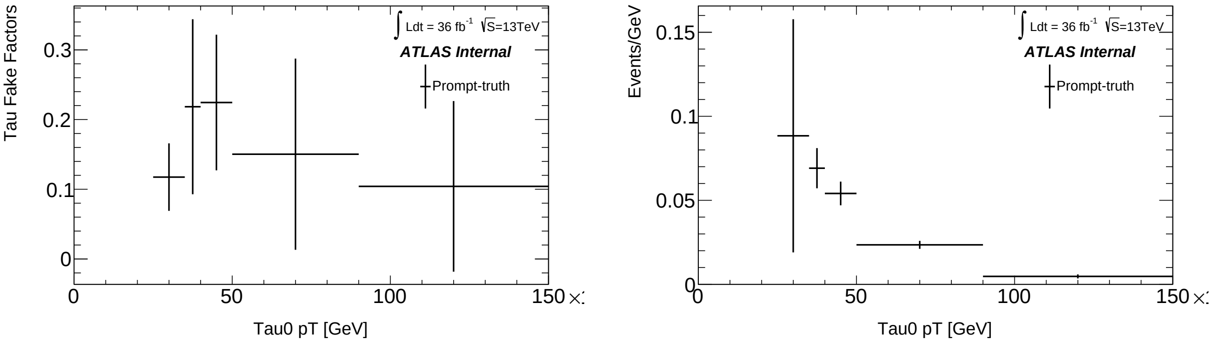
A.2 Estimate with (Prompt MC - truth-matched MC)

$$FF(p_T)_{MC} = \frac{N_{\tau}(p_T)^{\text{Prompt MC}} - N_{\tau}(p_T)^{\text{Truth-matched prompt MC}}}{N_{\not\tau}(p_T)^{\text{Prompt MC}} - N_{\not\tau}(p_T)^{\text{Truth-matched prompt MC}}} \tag{6}$$

Table 9: Extrapolation and closure test, (Prompt MC - truth-matched)

Integrated fake estimate	SR MC	Closure
0.471 ± 0.178	0.859 ± 0.191	$82 \pm 80\%$

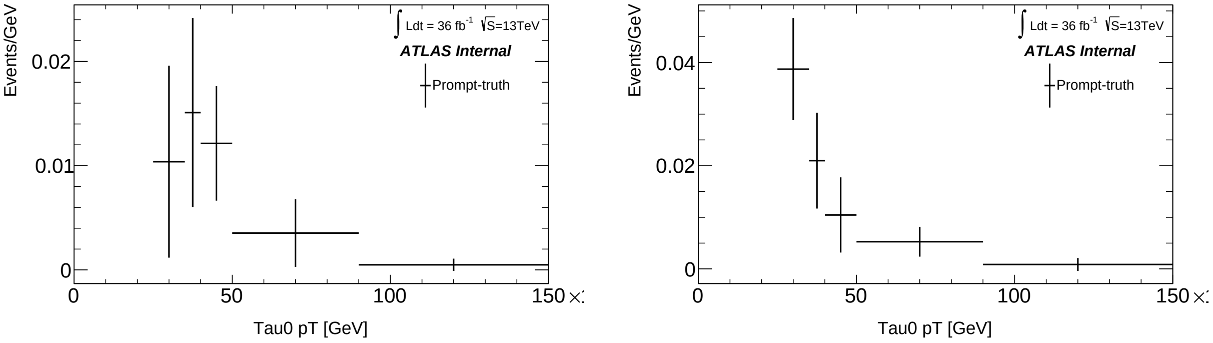
As is the case with the prompt MC without truth-match subtraction, most of the non-closure comes from the low-p_T bin. Omitting this bin, the fake estimate (0.368 ± 0.152) and MC prediction (0.472 ± 0.163) close to within $28 \pm 70\%$.



(a) p_T-parametrized fake factors

(b) Extrapolation sideband region p_T spectrum

Figure 8: Fake factors and extrapolation sideband region, (Prompt MC - truth)



(a) Fake estimate in p_T bins

(b) MC SR yield in p_T bins

Figure 9: p_T spectra for fake estimate and MC yield, (Prompt MC - truth)

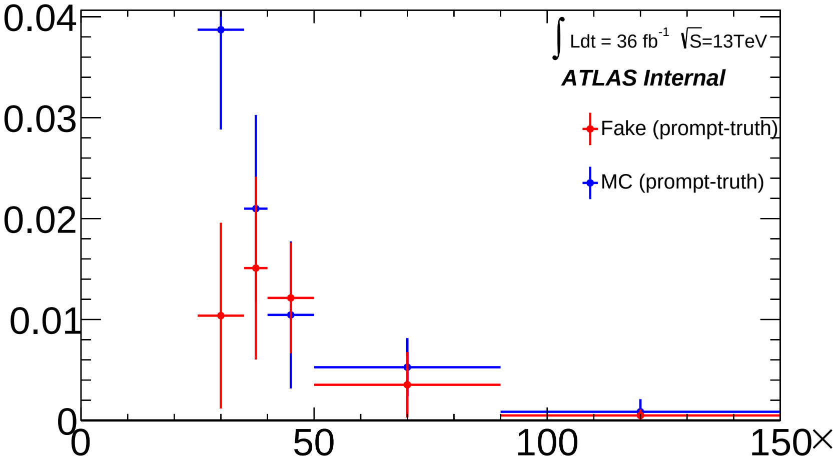
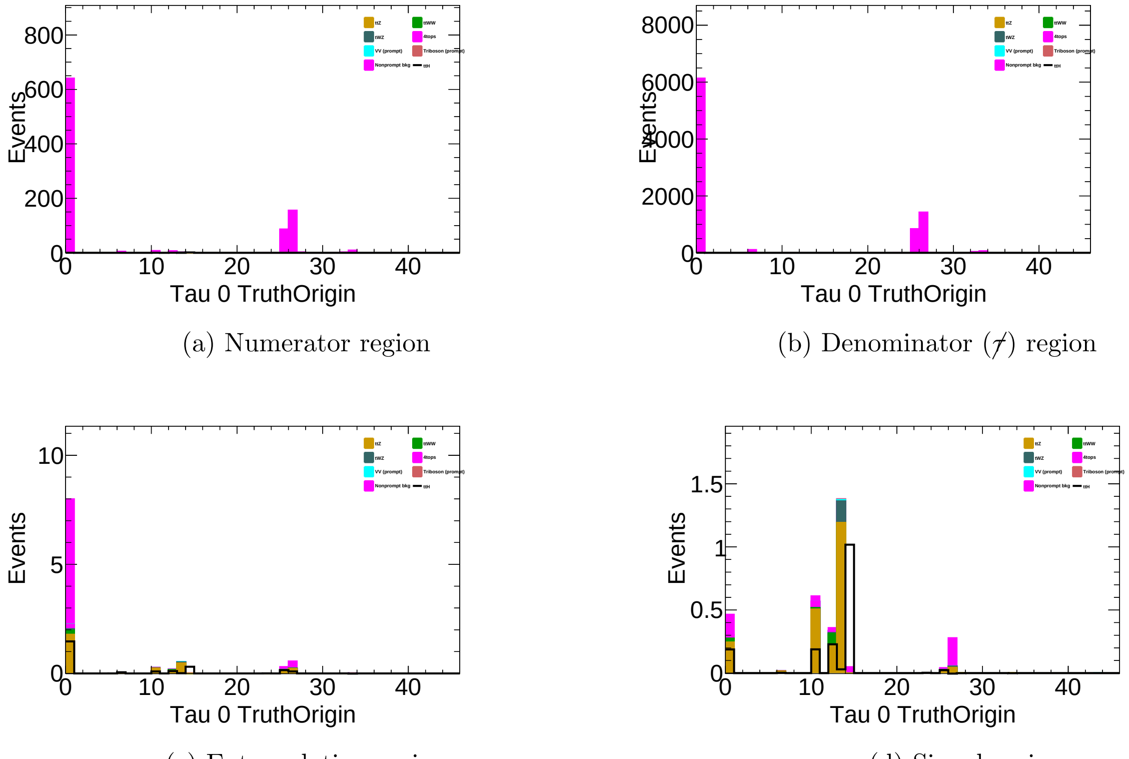


Figure 10: Overlaid fake estimate and MC prediction (Prompt MC - truth)

A.3 τ_{had} Truth origin studies



(a) Numerator region

(b) Denominator ($\not{\tau}$) region

(c) Extrapolation region

(d) Signal region

Figure 11: Truth origin of τ_{had} in fake-estimate control regions and signal region

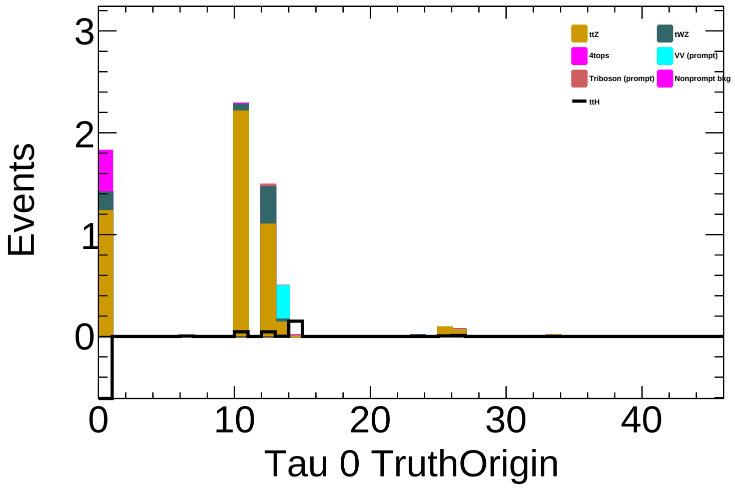


Figure 12: Truth origin of τ_{had} in $t\bar{t}Z$ control region (signal region selection + selecting events with 1 pair of OS-SF leptons within the Z-mass window

Table 10: Monte Carlo samples & categorization

ttH	343365 343366 343367
Prompt bkg	ttZ, ttWW, tWZ, 4top, VH, diboson (4l) and triboson (4l): 342284 342285 361063 361072 361073 361621 361623 361625 361626 410080 410081 410156 410157 410215 410218 410219 410220
Nonprompt bkg	tZ, ttW, 3top, single top, Z+jets, W+jets, diboson (non-4l), triboson (non-4l), tHbj, ttbar: 304014 341998 342001 342004 343267 343270 343273 361064 361065 361066 361067 361068 361069 361070 361071 361077 361091 361092 361093 361094 361095 361096 361097 361620 361622 361624 361627 364100 364101 364102 364103 364104 364105 364106 364107 364108 364109 364110 364111 364112 364113 364114 364115 364116 364117 364118 364119 364120 364121 364122 364123 364124 364125 364126 364127 364128 364129 364130 364131 364132 364133 364134 364135 364136 364137 364138 364139 364140 364141 364156 364157 364158 364159 364160 364161 364162 364163 364164 364165 364166 364167 364168 364169 364170 364171 364172 364173 364174 364175 364176 364177 364178 364179 364180 364181 364182 364183 364184 364185 364186 364187 364188 364189 364190 364191 364192 364193 364194 364195 364196 364197 364198 364199 364200 364201 364202 364203 364204 364205 364206 364207 364208 364209 364210 364211 364212 364213 364214 364215 410011 410012 410015 410016 410025 410026 410049 410155 410501

A.4 Estimate with non-prompt MC

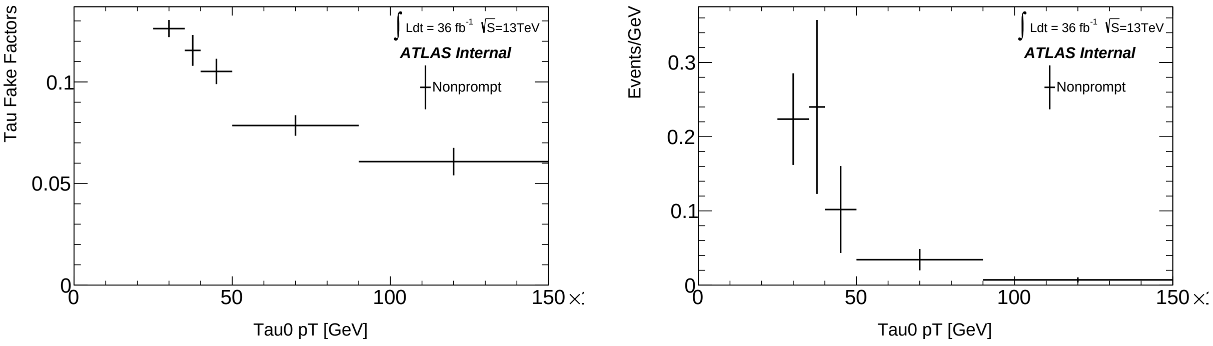
Here there is no subtraction performed and the estimate and closure test is done on the non-prompt MC only.

$$FF(p_T)_{MC} = \frac{N_{\tau}(p_T)^{\text{Non-prompt MC}}}{N_{\not\tau}(p_T)^{\text{Non-prompt MC}}} \tag{7}$$

Table 11: Extrapolation and closure test, (Non-prompt MC)

$t\bar{t}$ sample	Integrated fake estimate	SR MC	Raw	Closure
PP8 non-all hadronic	0.590 ± 0.146	0.704 ± 0.405	238	$19 \pm 75\%$
PP8 non-all hadronic, high stats	0.661 ± 0.130	0.510 ± 0.213	238	$-23 \pm 36\%$
PP8 dilepton	0.454 ± 0.105	0.532 ± 0.234	238	$17 \pm 58\%$
PP8 dilepton, high stats	0.530 ± 0.089	0.706 ± 0.241	240	$33 \pm 50\%$

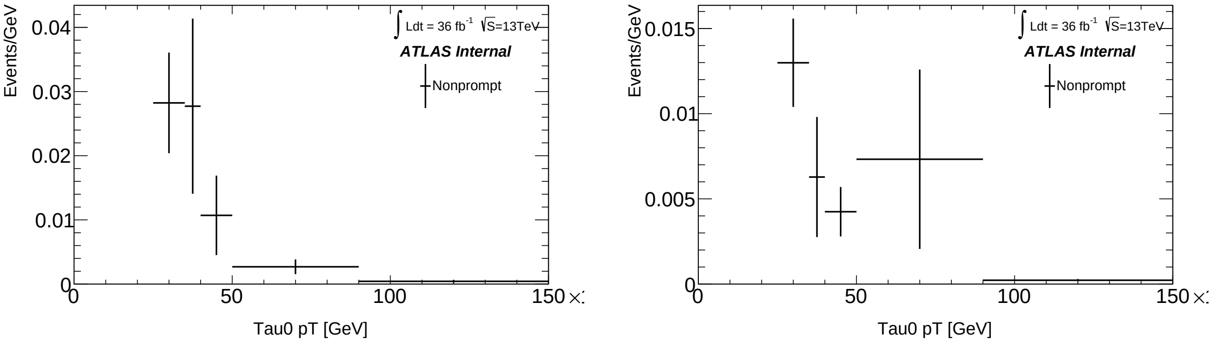
Figures are produced using the non-all hadronic high-stats sample.



(a) p_T-parametrized fake factors

(b) Extrapolation sideband region p_T spectrum

Figure 13: Fake factors and extrapolation sideband region, (Non-prompt MC)



(a) Fake estimate in p_T bins

(b) MC SR yield in p_T bins

Figure 14: p_T spectra for fake estimate and MC yield, (Non-prompt MC)

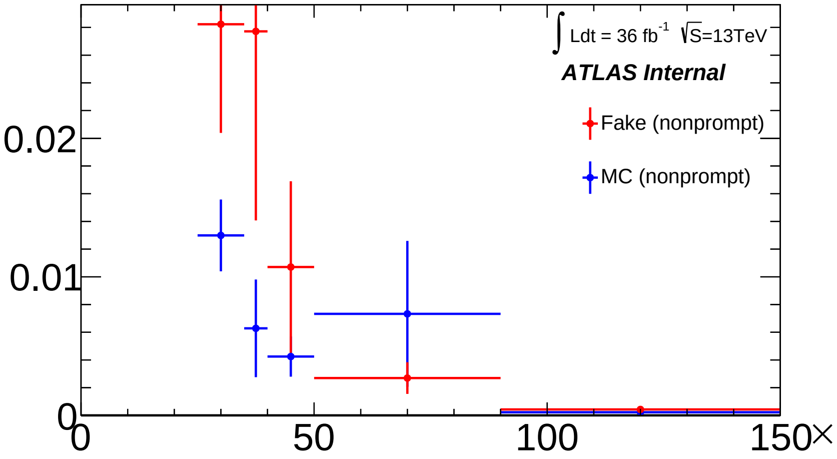


Figure 15: Overlaid fake estimate and MC prediction (Non-prompt MC)

A.5 Estimate with prompt MC

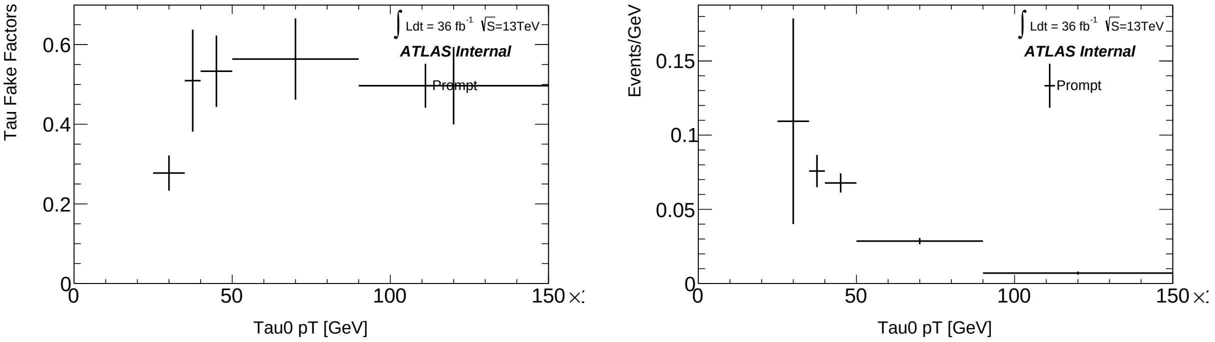
Here there is no subtraction performed and the estimate and closure test is done on the prompt MC only.

$$FF(p_T)_{MC} = \frac{N_\tau(p_T)^{\text{Prompt MC}}}{N_{\not\tau}(p_T)^{\text{Prompt MC}}} \tag{8}$$

Table 12: Extrapolation and closure test, (Prompt MC)

Integrated fake estimate	SR MC	Closure
1.71 ± 0.26	2.70 ± 0.15	$58 \pm 25\%$

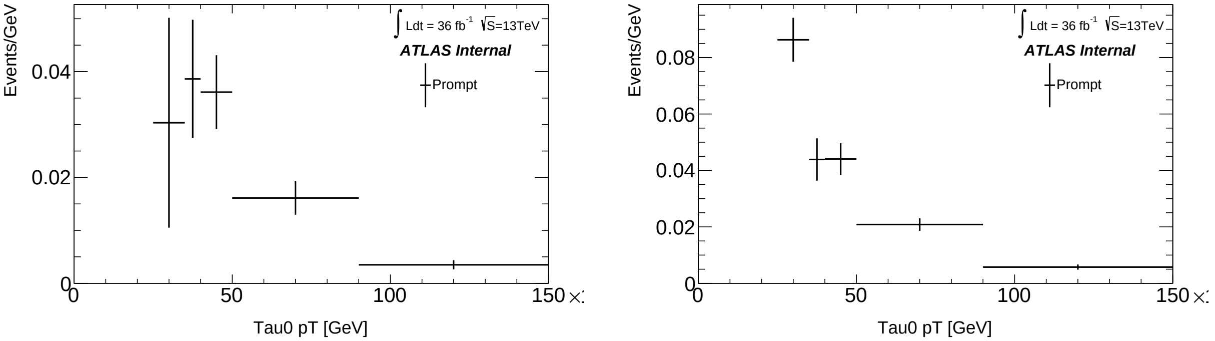
A large difference between the fake estimate procedure and the MC prediction is seen at low p_T. If the lowest p_T bin is omitted, the closure between the fake estimate (1.41 ± 0.16) and the MC (1.84 ± 0.12) improves to $30 \pm 18\%$.



(a) p_T-parametrized fake factors

(b) Extrapolation sideband region p_T spectrum

Figure 16: Fake factors and extrapolation sideband region, (Prompt MC)



(a) Fake estimate in p_T bins

(b) MC SR yield in p_T bins

Figure 17: p_T spectra for fake estimate and MC yield, (Prompt MC)

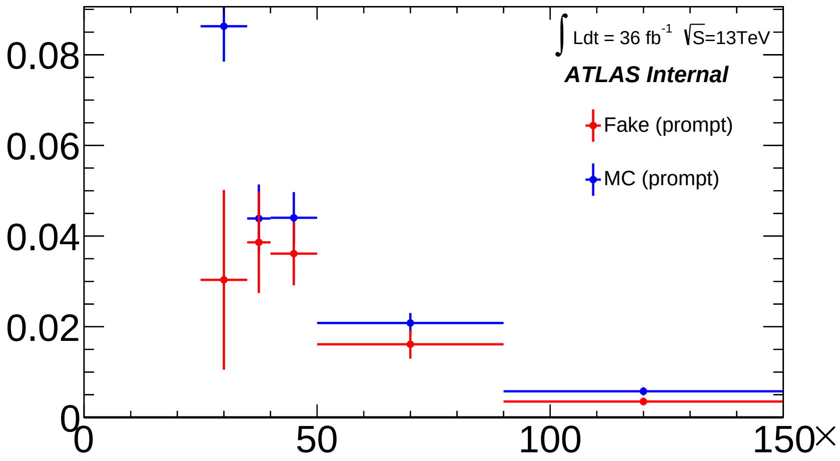


Figure 18: Overlaid fake estimate and MC prediction (Prompt MC)

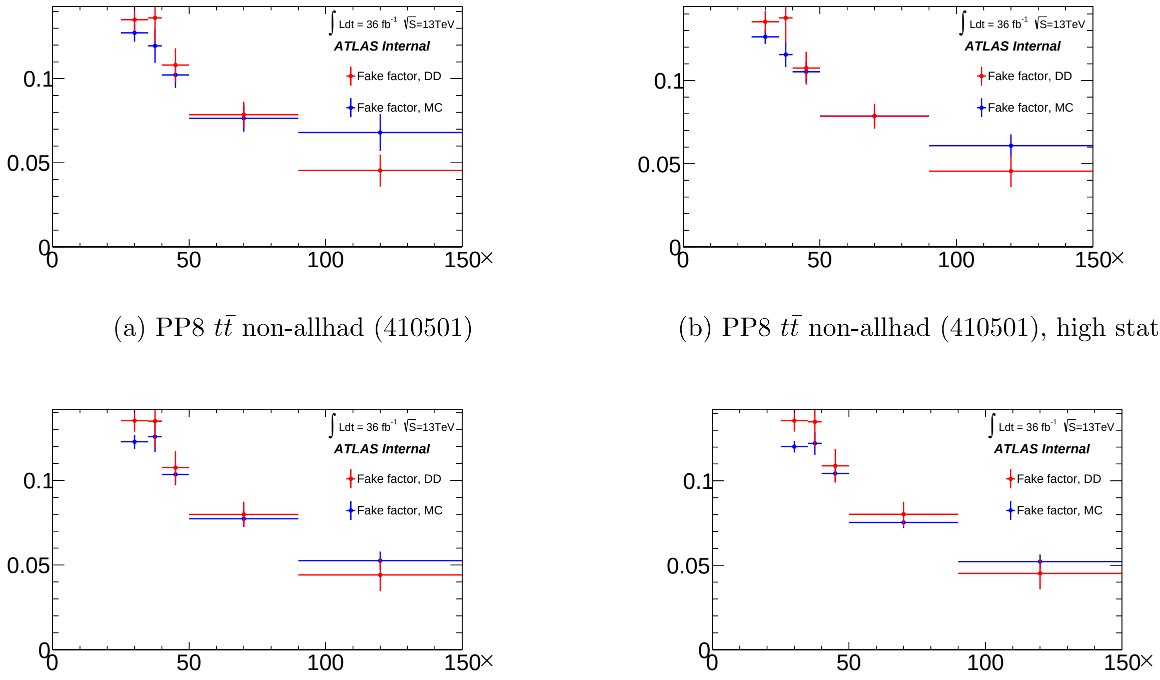
A.6 Data/MC comparison for $t\bar{t}$ samples in $2\ell OS + \tau$

Table 13: Data & MC in $2\ell OS + \tau$ with different $t\bar{t}$ samples

$t\bar{t}$ sample	$t\bar{t}H$	Prompt	Nonprompt	Data	Data/MC
PP8	2.467 ± 0.080	4.084 ± 0.273	924.376 ± 28.516	951	1.029
PP8 (high stats)	"	"	920.071 ± 21.203	"	1.034
PP8 (dilep)	"	"	896.863 ± 20.615	"	1.060
PP8 (dilep, high stats)	"	"	882.809 ± 16.779	"	1.077

Table 14: Data & MC in $2\ell OS + \not\tau$ with different $t\bar{t}$ samples

$t\bar{t}$ sample	$t\bar{t}H$	Prompt	Nonprompt	Data	Data/MC
PP8	4.399 ± 0.108	9.589 ± 0.608	8664.639 ± 76.760	8271	0.954
PP8 (high stats)	"	"	8647.666 ± 59.076	"	0.956
PP8 (dilep)	"	"	8580.240 ± 62.630	"	0.964
PP8 (dilep, high stats)	"	"	8590.145 ± 49.257	"	0.963



(a) PP8 $t\bar{t}$ non-allhad (410501)

(b) PP8 $t\bar{t}$ non-allhad (410501), high stats

(c) PP8 $t\bar{t}$ dilepton (410503)

(d) PP8 $t\bar{t}$ dilepton (410503), high stats

Figure 19: MC fake factors compared with data

A.7 Data-driven estimate

B Appendix: Ongoing to-do list

- Check results with new higher-stats $t\bar{t}$ samples, dilep-filtered samples and $t\bar{t}\gamma$ overlap samples
- Look into possible problem with calculation of weights/errors
- Get fit framework running (see David Hohn) w. systematics ntuples
- DONE: Validate data yields between vector- and flat-branch ntuple versions