ttH to leptons: Tau channels and non-prompt backgrounds estimation

Xuan Yang (LPSC)

Top-LHC France meeting

24th Apr. 2019

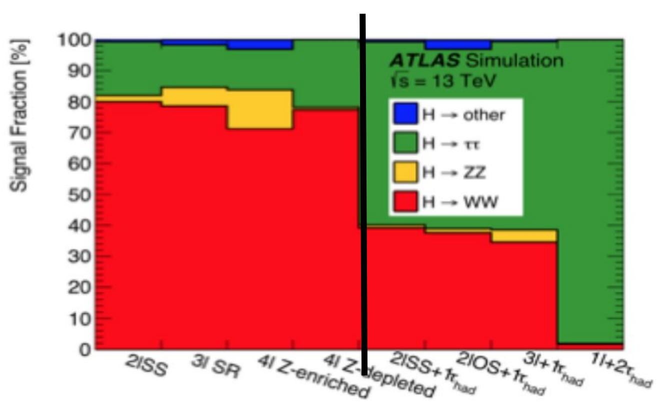
Outline

- Overview of ttH multileptons
- Overview of ATLAS tau channels
- Fake tau estimation
- Fake lepton estimation
  - Fake factor
  - Matrix method
- Statatistical results of tau channels
- Future

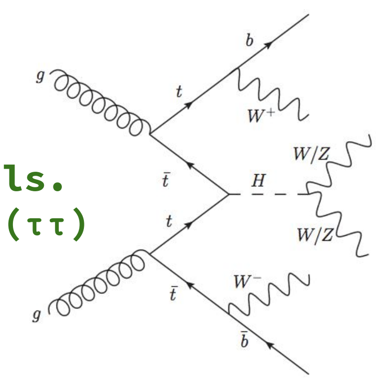
ttH

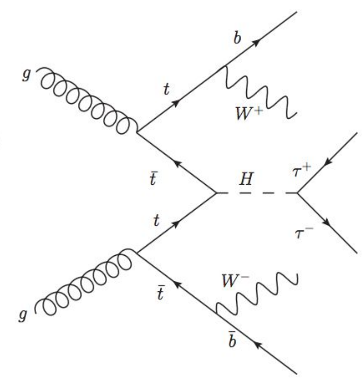
- Direct measurement of the Yukawa coupling between top quark and Higgs boson at tree level
- ttH prodution xs at 13 TeV is ~1 % of total Higgs production xs -> Very small!
- ttH-multilepton mainly targets at decay modes with >= 1 leptons (WW, ZZ, ττ)
- split by number of τhad: 0τ-channels, τ-channels
  - In the following, “tau” is always referred to as hadronic tau

0-tau channels. mainly H(WW)

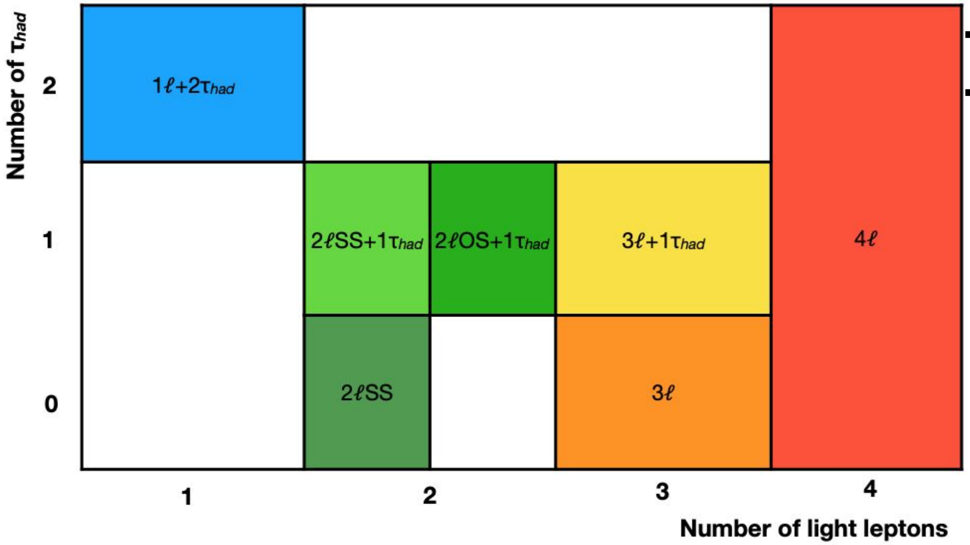


tau channels. also H(ττ)





ttH multilepton



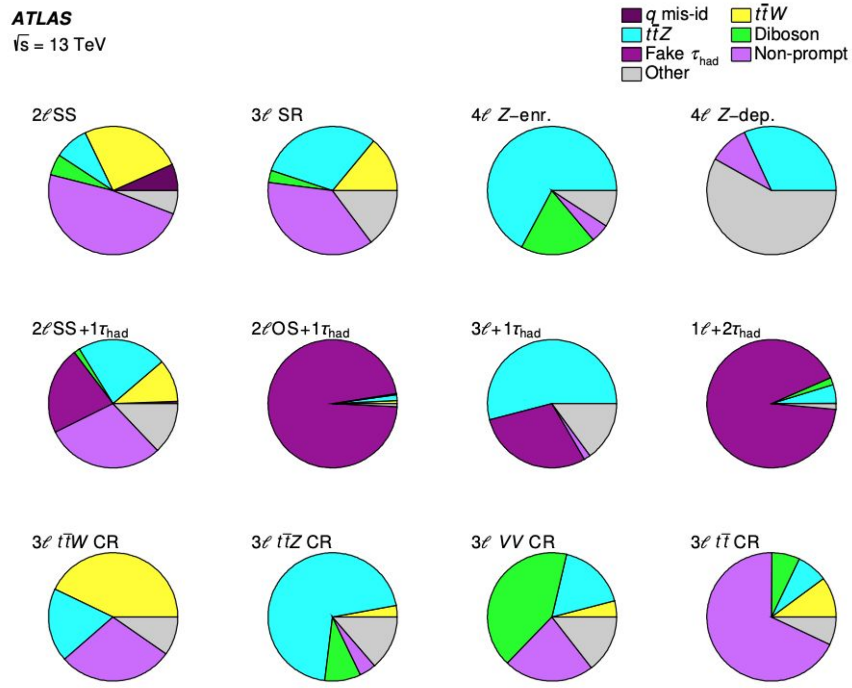
What we have in multilepton:

- ➔ 2/3/4 leptons (e/μ)
- ➔ 1 or 2 hadronic tau (τ_had)
- ➔ several jets (usually >=4)
- ➔ in which some are b-tagged (usually >=1)

The selections among different sub-channels are orthogonal at object level to avoid overlaps. Each channel has their own requirement on the objects (i.e. tight lepton or loose lepton) to get maximum statistics and sensitivity. Detailed selections can be found in backup slides.

Backgrounds

- Reducible backgrounds: Can be suppressed by optimizing event selection, improving estimate method, etc
  - Non-prompt leptons
  - Fake tau
  - Charge mis-id (charge flip)
  - Diboson
  - …
- Irreducible background: Can really mimic the signal topology, which are not possible to reduce by simple cuts
  - ttW
  - ttZ



In most of the channels, non-prompt and fake tau are the main backgrounds. It is essential to get a good estimation of these backgrounds.

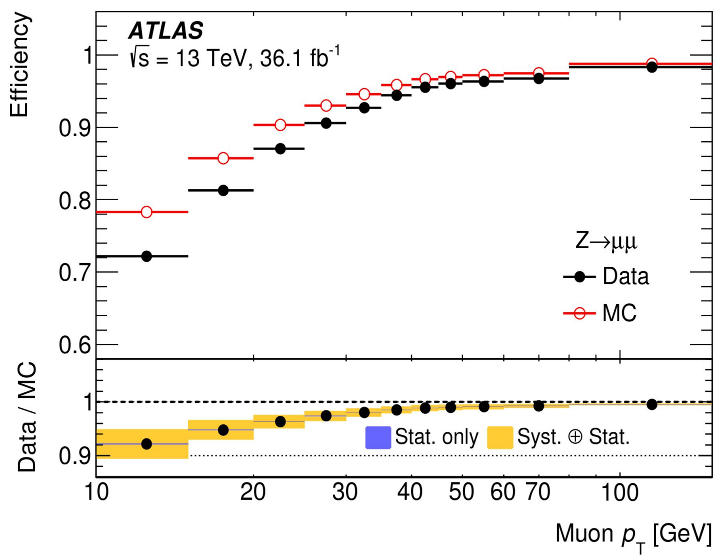
Leptons!

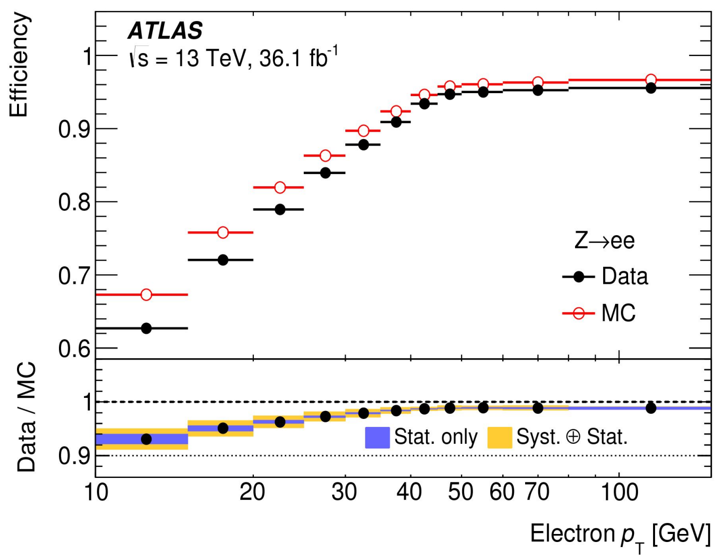
- The most important background: non-prompt leptons from semi-leptonic b decay
- Implement a new variable to reject non-prompt leptons -> PromptLeptonIso(PLI)

PLI is a BDT trained with:

- ❑ lepton and overlapping track jets properties
- ❑ lepton track/calorimeter isolation variables

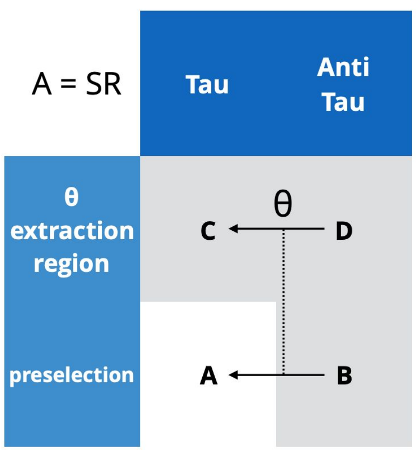
Scale factors (ratio of efficiency in data and in MC) to be used, measured from Z(ll) events. Maximum 0.95 at low pT.

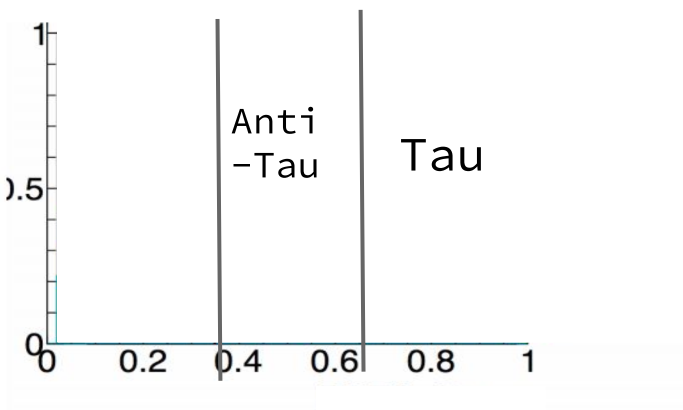




Fake tau estimation

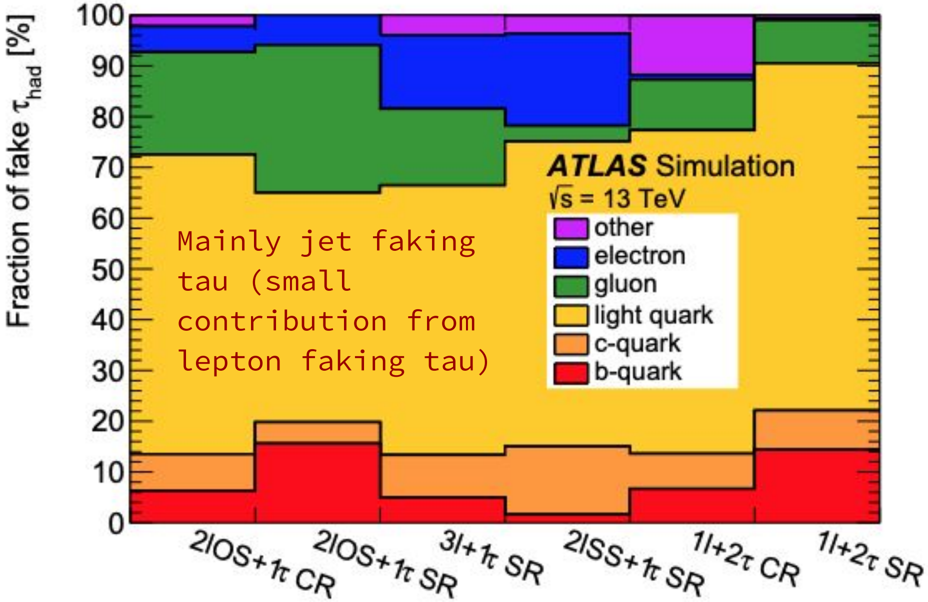
- Fake tau is estimated from 2lOS+1tau control region, where is dominated by fake tau contribution.
    - Apply 2los1tau selection but requires at least 3 jets and veto b-tagged jets
- Fake factor method (ABCD method)
    - Fake rates parametrised in jet pT
    - Derived in C/D and applied to B to estimate A
- Fake tau in 1l2tau channel is measured from a 1l2tau CR with SS tau pair (OS in SR)
    - Jets have identical chance to be reconstructed as positve or negative charged tau
    - The estimation is taken from the SS data with small corrections from simulation samples (truth tau contribution)



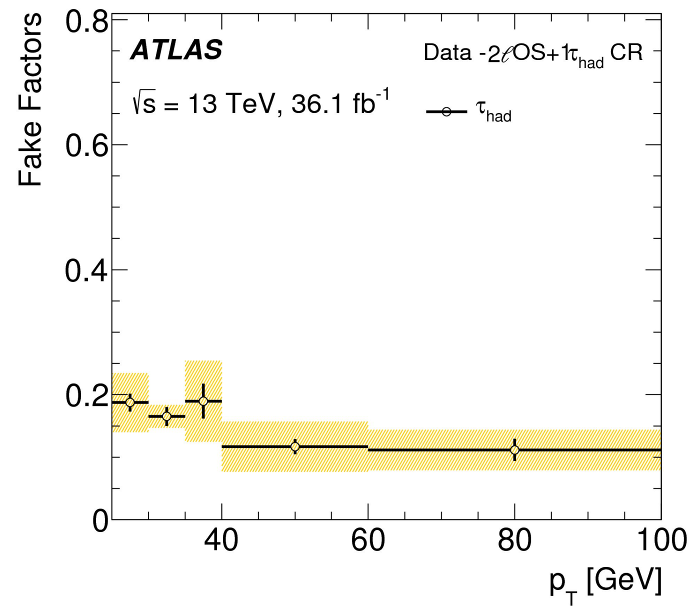


Cutting a BDT meant to identify hadronically decaying taus.

Fake tau estimation



(b) Fake τ_{had} composition

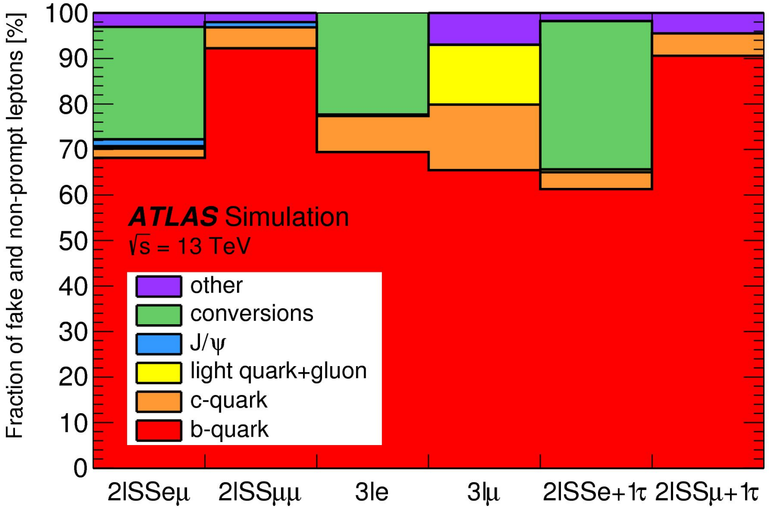


Fake tau composition is similar across channels, which allow us to just scale the factors measured in 2los1tau. A scale factor of 1.36 derived from 2los1tau CR (DD/Data) is applied to Monte carlos to get correct estimate in 3l1tau and 2lss1tau regions.

Fake lepton estimation

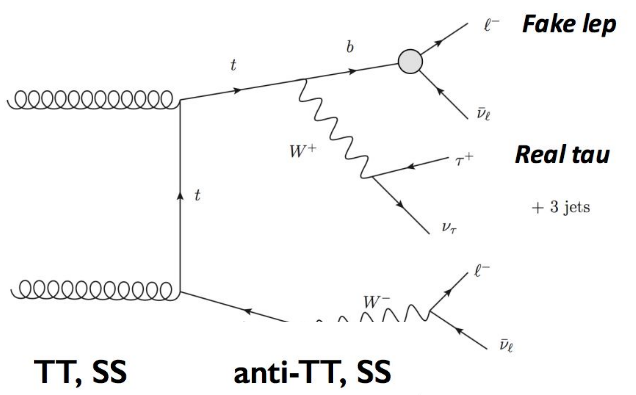
- Non-prompt lepton is the most important background in many channels.
- Crucial to the analysis to get a good estimates
- Extremely difficult to estimate from MC
- We developed two data-driven methods:
    - Fake factor: used in 2lss1tau
    - Matrix method: used in 2lss, 3l

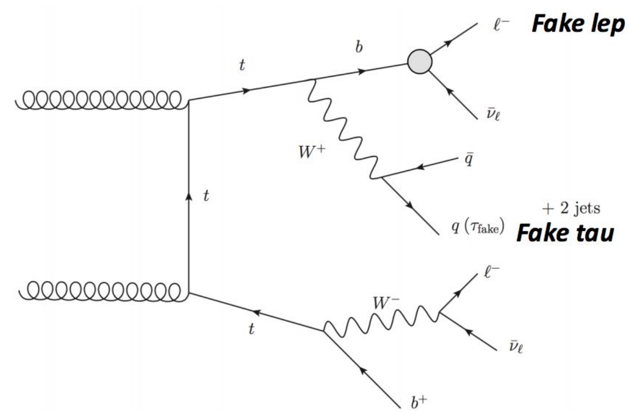
- Charge mis-identification contribution is estimated from Z->ee events with a pure DD method.

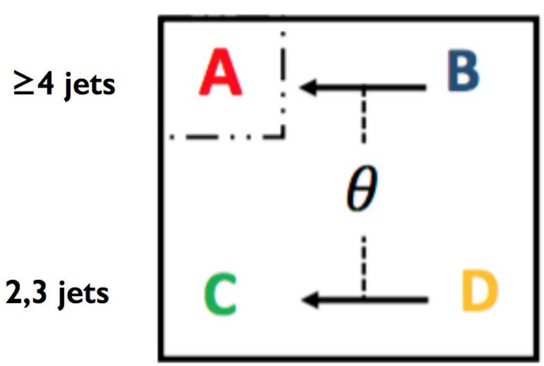


Fake factor

- ABCD method, same as for fake tau estimation.
- Most of the fakes are originated from ttbar
- fake factor is parameterised as a function of pT







C and D also include 0 tau events to increase stats.

anti-Tight = !tight, invert all the tight lepton selections.

Matrix method

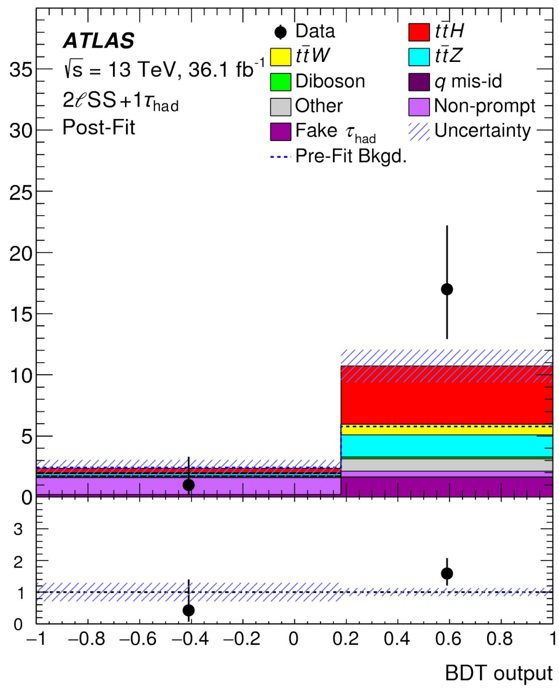
- Idea is similar: estimate fake lepton from looser region

$$\begin{pmatrix} N^{TT} \\ N^{T\not{T}} \\ N^{\not{T}T} \\ N^{\not{T}\not{T}} \end{pmatrix} = \begin{pmatrix} \varepsilon_{r,1}\varepsilon_{r,2} & \varepsilon_{r,1}\varepsilon_{f,2} & \varepsilon_{f,1}\varepsilon_{r,2} & \varepsilon_{f,1}\varepsilon_{f,2} \\ \varepsilon_{r,1}\not{\varepsilon}_{r,2} & \varepsilon_{r,1}\not{\varepsilon}_{f,2} & \varepsilon_{f,1}\not{\varepsilon}_{r,2} & \varepsilon_{f,1}\not{\varepsilon}_{f,2} \\ \not{\varepsilon}_{r,1}\varepsilon_{r,2} & \not{\varepsilon}_{r,1}\varepsilon_{f,2} & \not{\varepsilon}_{f,1}\varepsilon_{r,2} & \not{\varepsilon}_{f,1}\varepsilon_{f,2} \\ \not{\varepsilon}_{r,1}\not{\varepsilon}_{r,2} & \not{\varepsilon}_{r,1}\not{\varepsilon}_{f,2} & \not{\varepsilon}_{f,1}\not{\varepsilon}_{r,2} & \not{\varepsilon}_{f,1}\not{\varepsilon}_{f,2} \end{pmatrix} \begin{pmatrix} N^{rr} \\ N^{rf} \\ N^{fr} \\ N^{ff} \end{pmatrix}$$

Elements in the matrix is the efficiency of a real/fake lepton to pass tight/anti-tight selection, which are measured in data in a dedicated CR

The fake composition can be different between regions, we implement α-function to account for the difference since different composition has different fake rate. The fake rate measured in CR is reweighted by the α to get the fake rate in SR

2lSS+1tau

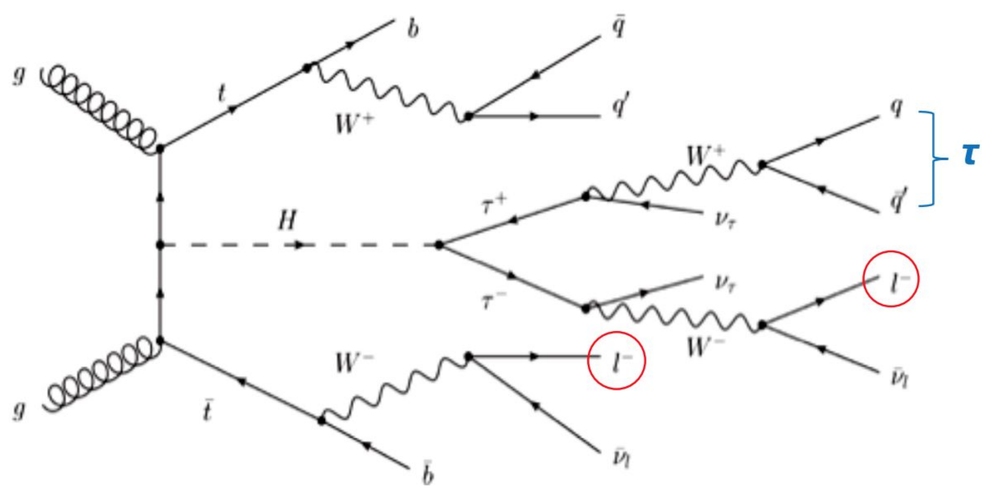


Post-fit distribution

BDT discriminant

- Main background:
  - ttbar where additional light jet fakes a tau
  - Non-prompt light leptons
- Main systematic:
  - Fake factor non-closure
  - JES
- Sensitivity:
  - Expected: 1.1 σ
  - Observed: 3.4 σ

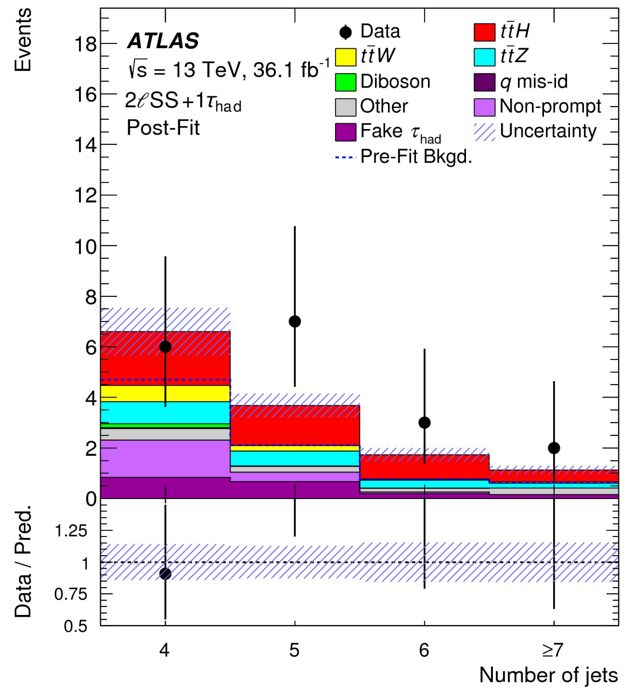
Fitted signal strength $3.5\ ^{+1.5}_{-1.2}\ (\text{stat.})\ ^{+0.9}_{-0.5}\ (\text{syst.})$



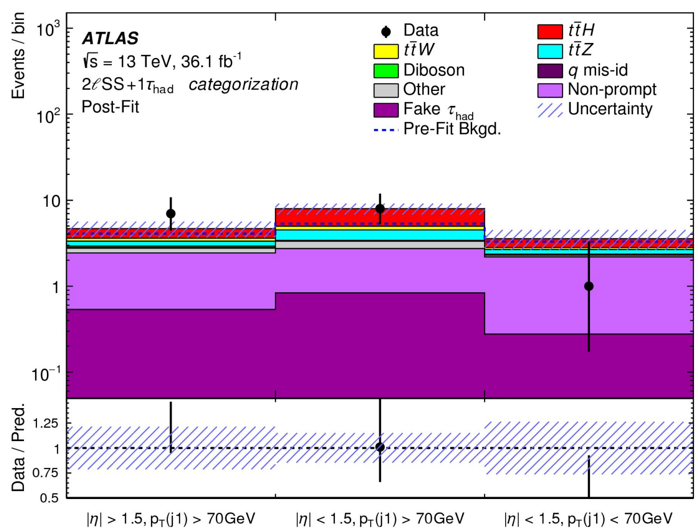
2lSS+1tau

The most sensitive variables in the BDT training are:

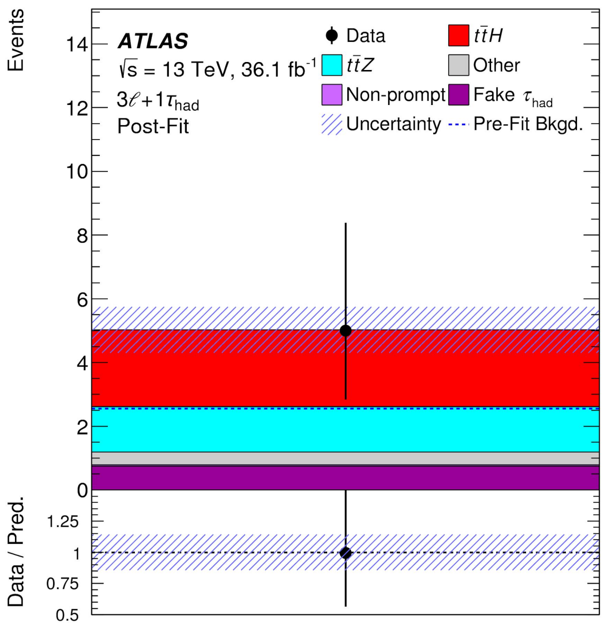
- HTjets: Scalar sum of jets pt
- Jet multiplicity



Beside the BDT approach, we also developed an alternative cut&count analysis, categorisation. It give similar sensitivity to the BDT analysis. Details are in the backup slides.



3l+1tau

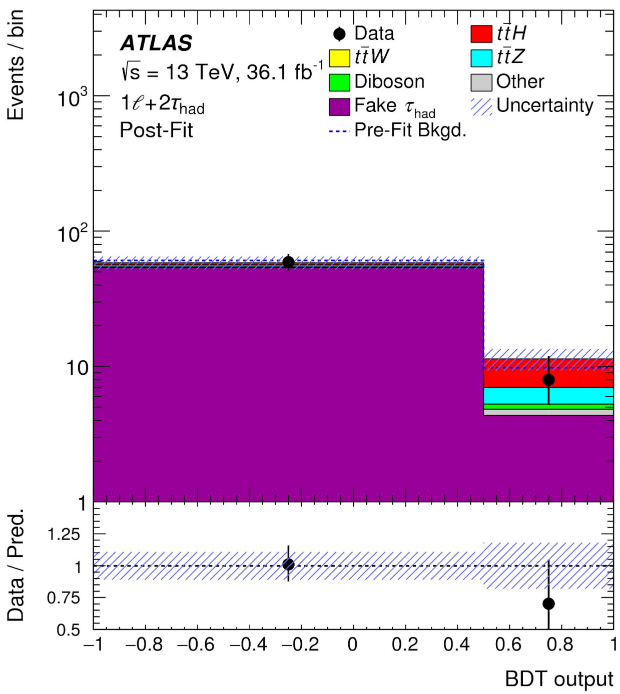


Post-fit distribution

One bin channel

- Main background:
  - ttbar where additional light jet fakes a tau
  - ttZ
- Main systematic:
  - Tau scale factor
  - theoretical uncertainties
  - Fake tau related
  - Statistical uncertainty (γ)
- Sensitivity:
  - Expected: 0.9 σ
  - Observed: 1.3 σ

1l+2tau

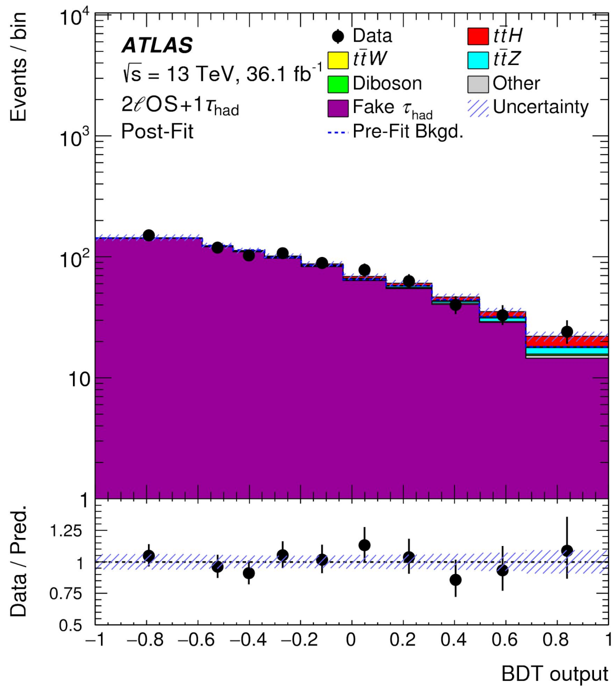


Post-fit distribution

BDT discriminant

- Main background:
  - ttbar where additional light jet fakes a tau
- Leading discriminant:
  - the minimum ΔR between the selected jets in the event
  - the scalar sum of jet pT
  - the di-tau invariant mass
- Main systematic:
  - Fake tau related
- Sensitivity:
  - Expected: 0.6 σ
  - Observed: - (the fitted signal strength is negative)

2lOS+1tau

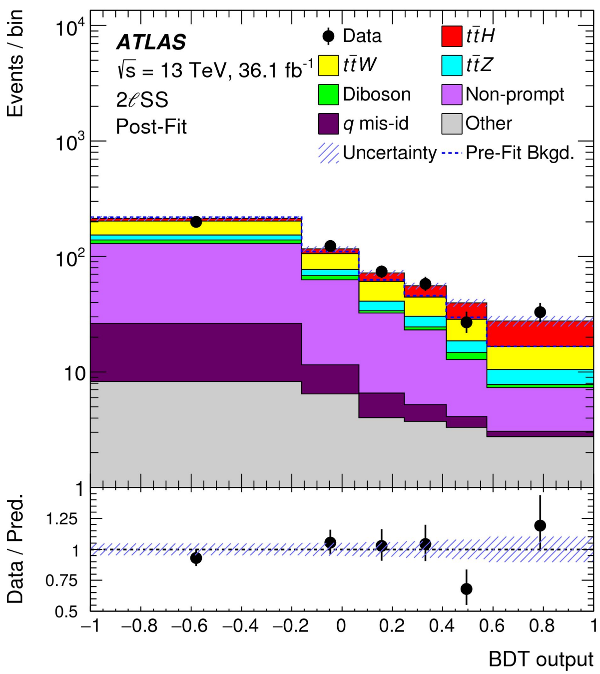


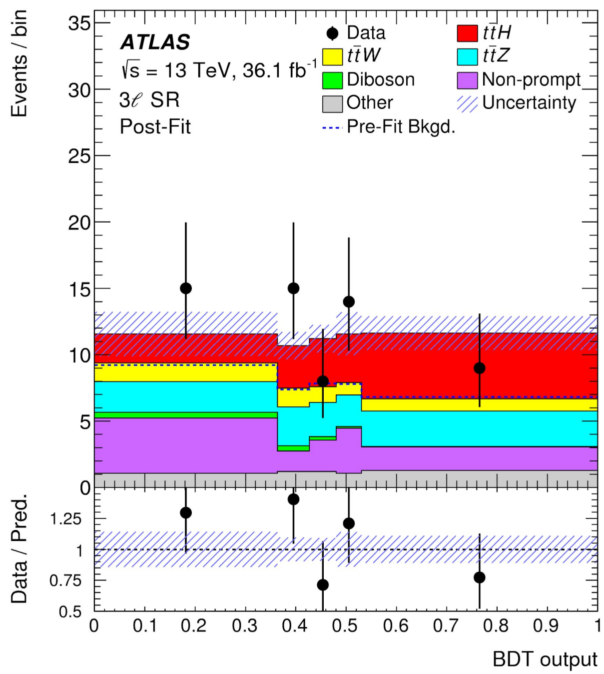
Post-fit distribution

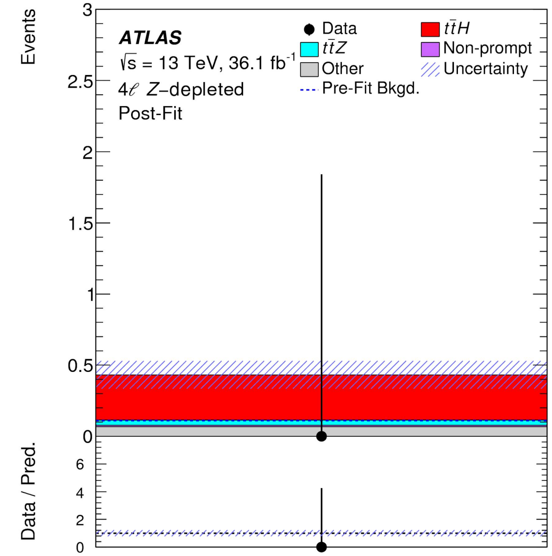
BDT discriminant

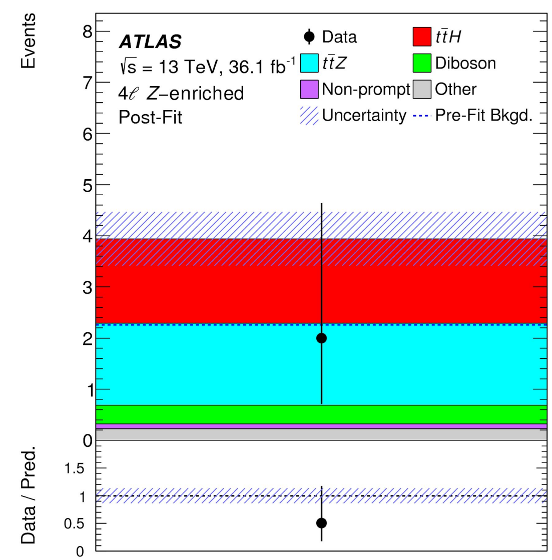
- Main background:
  - ttbar where additional light jet fakes a tau
- Leading discriminant:
  - invariant mass of light lepton pair
  - transverse momentum of hadronic tau
  - smallest ΔR distance between a lepton and a jet
- Main systematic:
  - Fake tau related
- Sensitivity:
  - Expected: 0.5 σ
  - Observed: 0.9 σ

Non-tau channels







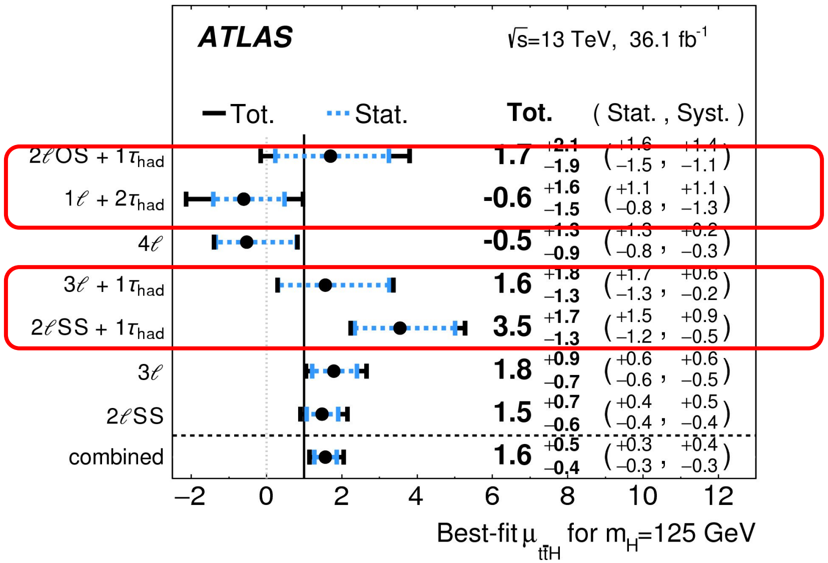


2lss0tau: combination of BDT (vs. ttbar) and BDT (vs. ttV)

3l0tau: multi-class BDT

4l: split into Z-enriched and z-depleted

Statistical results



Analysis strategy:

- 2lss1tau: BDT (ttH-vs-allbkg)
- 3l1tau: event counting
- 1l2tau: BDT (ttH-vs-top)
- 2los1tau: BDT (ttH-vs-ttbar)

2lss1tau is a bit off (3.5σ), but still compatible. Overall compatibility (comparing to combined result) for all channels is 34%

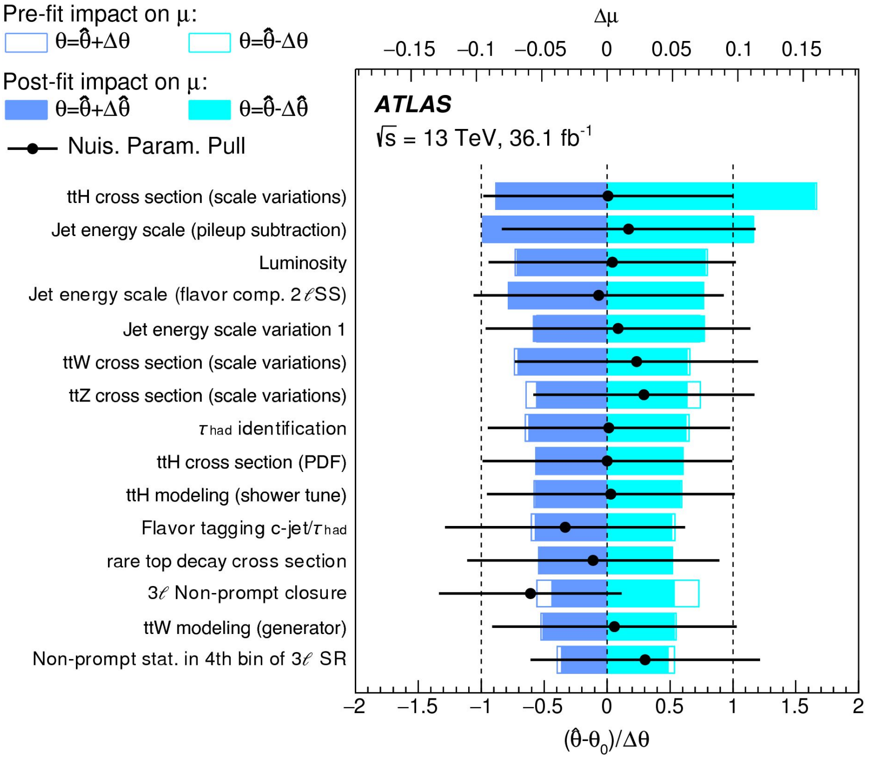
Observed 4.1σ for 0tau and tau channels combined fit. Compatible with SM prediction

Channel	Best-fit μ Observed	Best-fit μ Expected	Significance Observed	Significance Expected
$2\ell OS+1\tau_{had}$	$1.7\ ^{+1.6}_{-1.5}$ (stat.) $^{+1.4}_{-1.1}$ (syst.)	$1.0\ ^{+1.5}_{-1.4}$ (stat.) $^{+1.2}_{-1.1}$ (syst.)	0.9σ	0.5σ
$1\ell+2\tau_{had}$	$-0.6\ ^{+1.1}_{-0.8}$ (stat.) $^{+1.1}_{-1.3}$ (syst.)	$1.0\ ^{+1.1}_{-0.9}$ (stat.) $^{+1.2}_{-1.1}$ (syst.)	–	0.6σ
4ℓ	$-0.5\ ^{+1.3}_{-0.8}$ (stat.) $^{+0.2}_{-0.3}$ (syst.)	$1.0\ ^{+1.7}_{-1.2}$ (stat.) $^{+0.4}_{-0.2}$ (syst.)	–	0.8σ
$3\ell+1\tau_{had}$	$1.6\ ^{+1.7}_{-1.3}$ (stat.) $^{+0.6}_{-0.2}$ (syst.)	$1.0\ ^{+1.5}_{-1.1}$ (stat.) $^{+0.4}_{-0.2}$ (syst.)	1.3σ	0.9σ
$2\ell SS+1\tau_{had}$	$3.5\ ^{+1.5}_{-1.2}$ (stat.) $^{+0.9}_{-0.5}$ (syst.)	$1.0\ ^{+1.1}_{-0.8}$ (stat.) $^{+0.5}_{-0.3}$ (syst.)	3.4σ	1.1σ
3ℓ	$1.8\ ^{+0.6}_{-0.6}$ (stat.) $^{+0.6}_{-0.5}$ (syst.)	$1.0\ ^{+0.6}_{-0.5}$ (stat.) $^{+0.5}_{-0.4}$ (syst.)	2.4σ	1.5σ
$2\ell SS$	$1.5\ ^{+0.4}_{-0.4}$ (stat.) $^{+0.5}_{-0.4}$ (syst.)	$1.0\ ^{+0.4}_{-0.4}$ (stat.) $^{+0.4}_{-0.4}$ (syst.)	2.7σ	1.9σ
Combined	$1.6\ ^{+0.3}_{-0.3}$ (stat.) $^{+0.4}_{-0.3}$ (syst.)	$1.0\ ^{+0.3}_{-0.3}$ (stat.) $^{+0.3}_{-0.3}$ (syst.)	4.1σ	2.8σ

Systematics

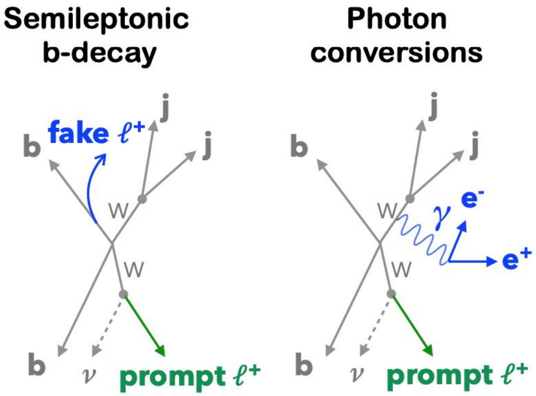
- Largest impact on μ:
  - Signal modelling
  - JES and JER
  - non-prompt leptons
- No significant constrain or pulls of nuisance parameters

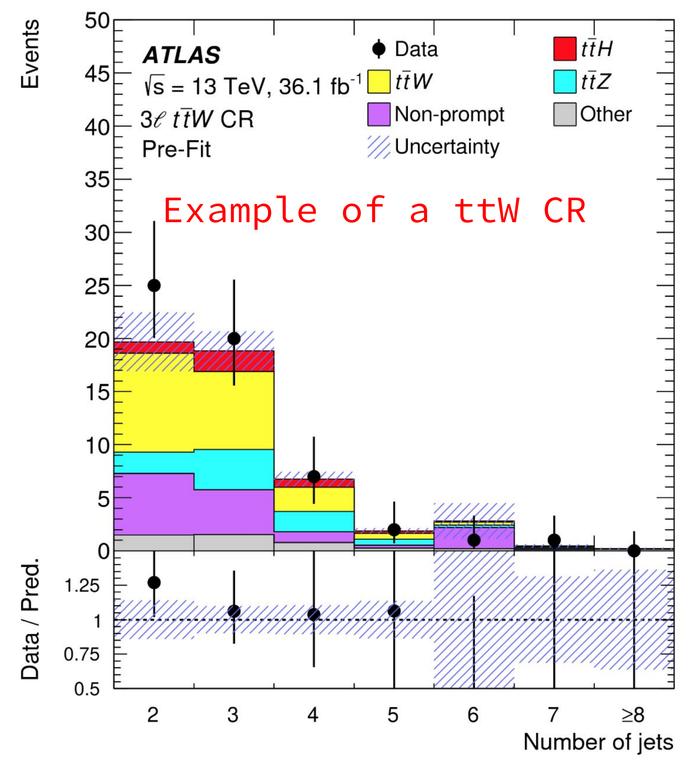
Uncertainty Source	$\Delta\mu$	
$t\bar{t}H$ modelling (cross section)	+0.20	−0.09
Jet energy scale and resolution	+0.18	−0.15
Non-prompt light-lepton estimates	+0.15	−0.13
Jet flavour tagging and τ_{had} identification	+0.11	−0.09
$t\bar{t}W$ modelling	+0.10	−0.09
$t\bar{t}Z$ modelling	+0.08	−0.07
Other background modelling	+0.08	−0.07
Luminosity	+0.08	−0.06
$t\bar{t}H$ modelling (acceptance)	+0.08	−0.04
Fake τ_{had} estimates	+0.07	−0.07
Other experimental uncertainties	+0.05	−0.04
Simulation statistics	+0.04	−0.04
Charge misassignment	+0.01	−0.01
Total systematic uncertainty	+0.39	−0.30



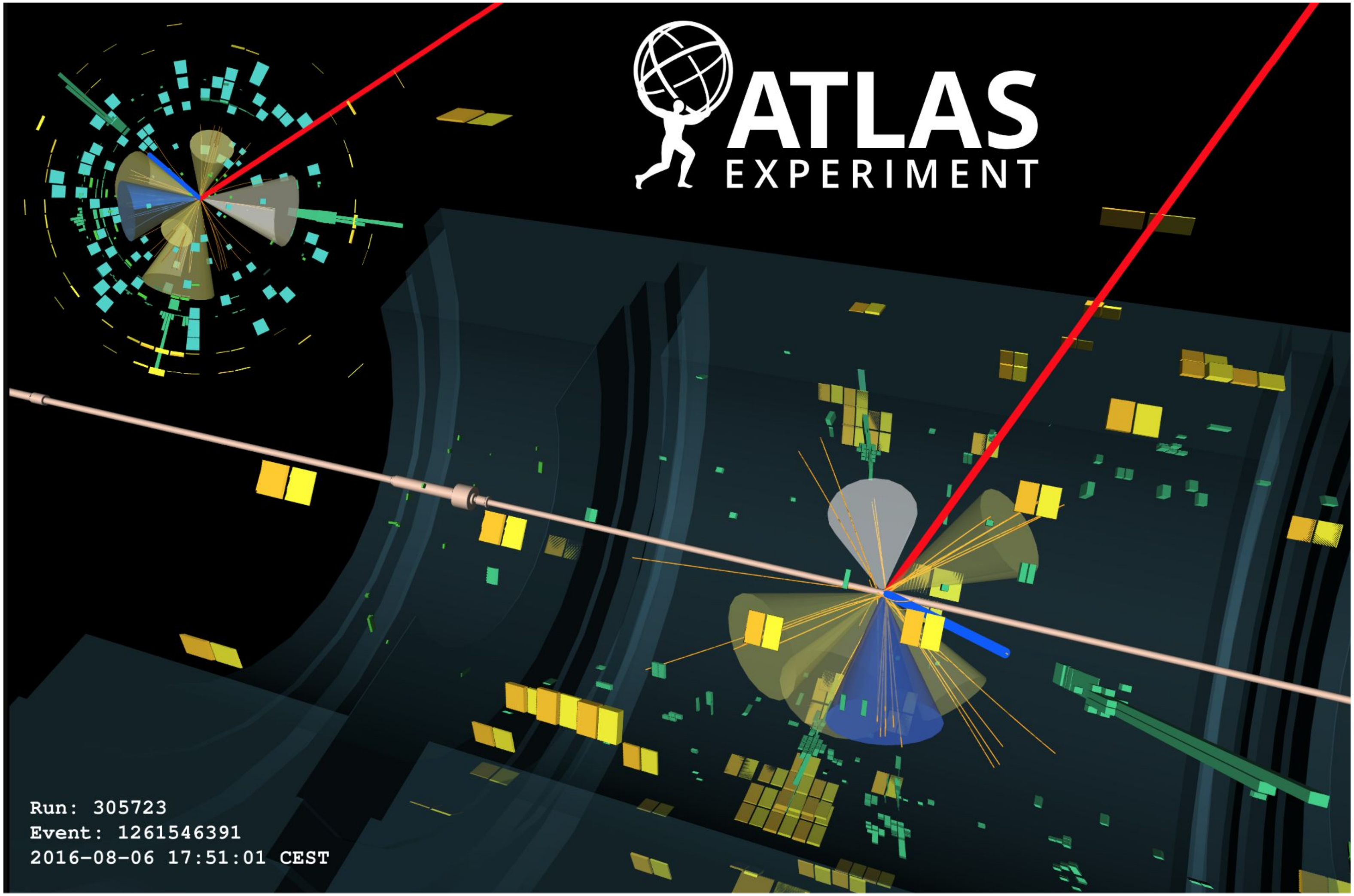
Future

- Working on updates with more statistics with a further improved analysis in terms of backgrounds understanding
- The tighter we cut on heavy flavour backgrounds, the more there is a non negligible presence of conversions which forces us to reinforce non-prompt estimate
- The tighter we cut and reduce non-prompt background, the more ttZ and ttW backgrounds are becoming dominant. In the future we have increasing interests towards building ttW regions that can be used as a control regions
- Exploring new possiblilities:
    - New methods: Machine learning further than BDT (xgboost, DNN...), MEM





Event display of 2lss1tau candidate

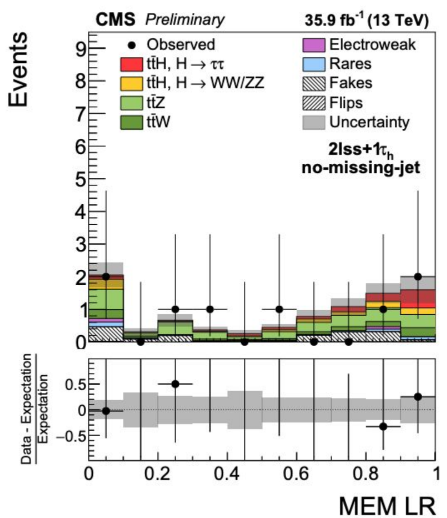


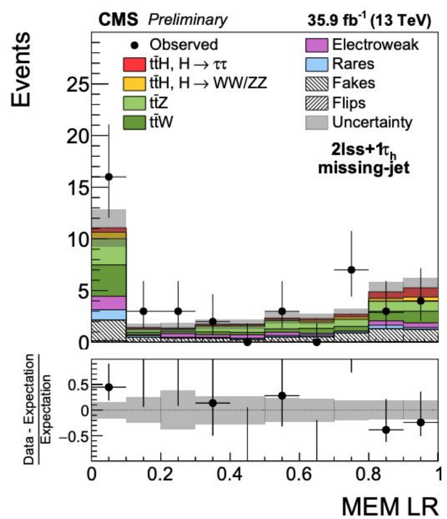
BACKUP

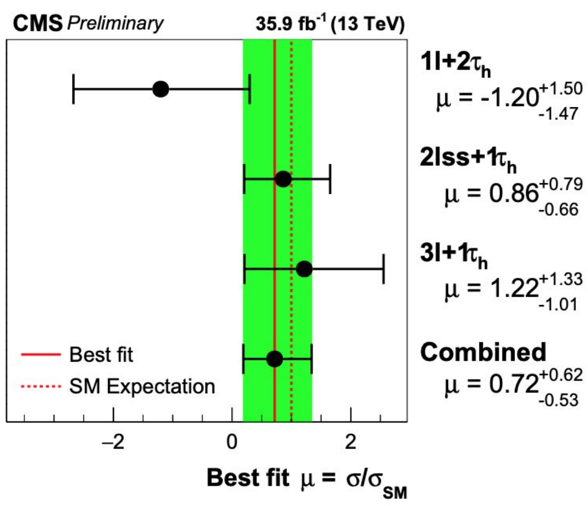
CMS tau channels

More to come in Cristina's talk!

- 2lss1tau, 1l2tau, 3l1tau (same as ATLAS, except 2los1tau)
  - 2los1tau will become an useful tau CR in next round in ATLAS
- BDT is trained for 1l2tau and 3l1tau
- Matrix element method (MEM) is implemented in 2lss1tau for event reconstruction
- Events selected in 2lss1tau SR are further categorised:
  - "No missing jet category": full event reconstruction
  - "missing jet category": one of the jets originated from W boson is missing

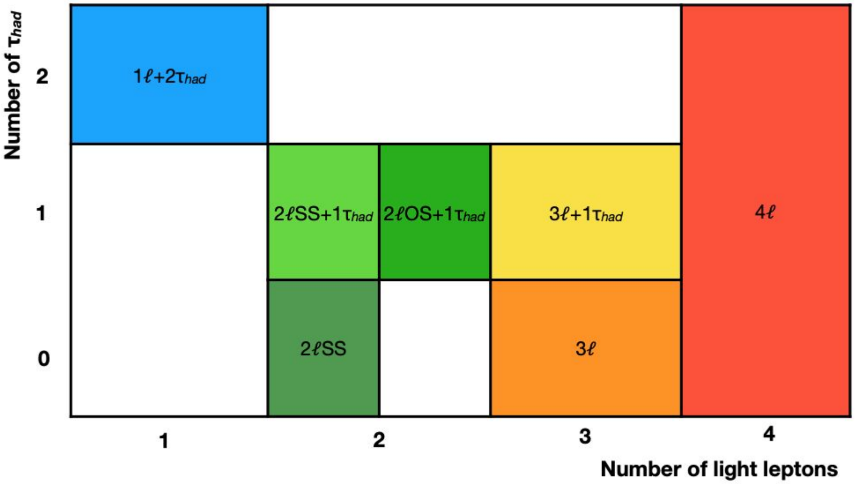






ttH multilepton

Orthogonal among all channels at object level to avoid overlaps.



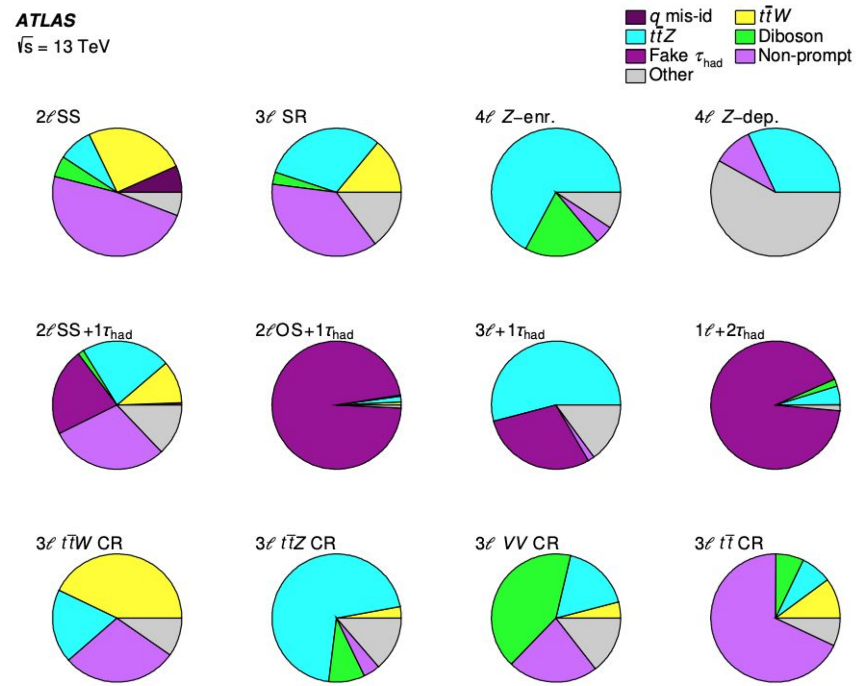
	e					μ			
	L	L†	L*	T	T*	L	L†	L*	T
Isolation	No	Yes				No	Yes		
Non-prompt lepton BDT	No		Yes			No		Yes	
Identification	Loose			Tight		Loose			
Charge mis-assignment veto	No				Yes	N/A			
Tranverse impact parameter significance $\lvert d_0 \rvert / \sigma_{d_0}$	< 5					< 3			
Longitudinal impact parameter $\lvert z_0 \sin\theta \rvert$	< 0.5 mm								

Lepton definitions

	Non-tau channels			Tau channels			
	2ℓSS	3ℓ	4ℓ	1ℓ+$2\tau_{\text{had}}$	2ℓSS+$1\tau_{\text{had}}$	2ℓOS+$1\tau_{\text{had}}$	3ℓ+$1\tau_{\text{had}}$
Light lepton	2T*	1L*, 2T*	2L, 2T	1T	2T*	2L†	1L†, 2T
τ_{had}	0M	0M	–	1T, 1M	1M	1M	1M
$N_{\text{jets}}, N_{b-\text{jets}}$	$\geq 4, = 1,2$	$\geq 2, \geq 1$	$\geq 2, \geq 1$	$\geq 3, \geq 1$	$\geq 4, \geq 1$	$\geq 3, \geq 1$	$\geq 2, \geq 1$

Event selection

Channel	Selection criteria
Common	$N_{\text{jets}} \geq 2$ and $N_{b\text{-jets}} \geq 1$
2ℓSS	Two very tight light leptons with $p_T > 20$ GeV Same-charge light leptons Zero medium τ_{had} candidates $N_{\text{jets}} \geq 4$ and $N_{b\text{-jets}} < 3$
3ℓ	Three light leptons with $p_T > 10$ GeV; sum of light-lepton charges ± 1 Two same-charge leptons must be very tight and have $p_T > 15$ GeV The opposite-charge lepton must be loose, isolated and pass the non-prompt BDT Zero medium τ_{had} candidates $m(\ell^+\ell^-) > 12$ GeV and $\lvert m(\ell^+\ell^-) - 91.2\text{ GeV}\rvert > 10$ GeV for all SFOC pairs $\lvert m(3\ell) - 91.2\text{ GeV}\rvert > 10$ GeV
4ℓ	Four light leptons; sum of light-lepton charges 0 Third and fourth leading leptons must be tight $m(\ell^+\ell^-) > 12$ GeV and $\lvert m(\ell^+\ell^-) - 91.2\text{ GeV}\rvert > 10$ GeV for all SFOC pairs $\lvert m(4\ell) - 125\text{ GeV}\rvert > 5$ GeV Split 2 categories: Z-depleted (0 SFOC pairs) and Z-enriched (2 or 4 SFOC pairs)
$1\ell+2\tau_{\text{had}}$	One tight light lepton with $p_T > 27$ GeV Two medium τ_{had} candidates of opposite charge, at least one being tight $N_{\text{jets}} \geq 3$
$2\ell\text{SS}+1\tau_{\text{had}}$	Two very tight light leptons with $p_T > 15$ GeV Same-charge light leptons One medium τ_{had} candidate, with charge opposite to that of the light leptons $N_{\text{jets}} \geq 4$ $\lvert m(ee) - 91.2\text{ GeV}\rvert > 10$ GeV for ee events
$2\ell\text{OS}+1\tau_{\text{had}}$	Two loose and isolated light leptons with $p_T > 25, 15$ GeV One medium τ_{had} candidate Opposite-charge light leptons One medium τ_{had} candidate $m(\ell^+\ell^-) > 12$ GeV and $\lvert m(\ell^+\ell^-) - 91.2\text{ GeV}\rvert > 10$ GeV for the SFOC pair $N_{\text{jets}} \geq 3$
$3\ell+1\tau_{\text{had}}$	3ℓ selection, except: One medium τ_{had} candidate, with charge opposite to the total charge of the light leptons The two same-charge light leptons must be tight and have $p_T > 10$ GeV The opposite-charge light lepton must be loose and isolated



In most of the channel, non-prompt and fake tau are the main backgrounds. It is essential to get a good estimation of these backgrounds.

Charge Mis-ID

- The charge of electrons can be mis-identified due to:
  - Wrong assignment of EM cluster
  - Slightly curved track that induces measurement error
- Estimated by a pure data-driven method

$N^{os} = (1 - 2\varepsilon + 2\varepsilon^2)N$ opposite-sign events,

$N^{ss} = 2\varepsilon(1 - \varepsilon)N$ same-sign events,

→

$N^{ss} = \frac{\varepsilon_i + \varepsilon_j - 2\varepsilon_i\varepsilon_j}{1 - \varepsilon_i - \varepsilon_j + 2\varepsilon_i\varepsilon_j} N^{os}$ for the ee channel,

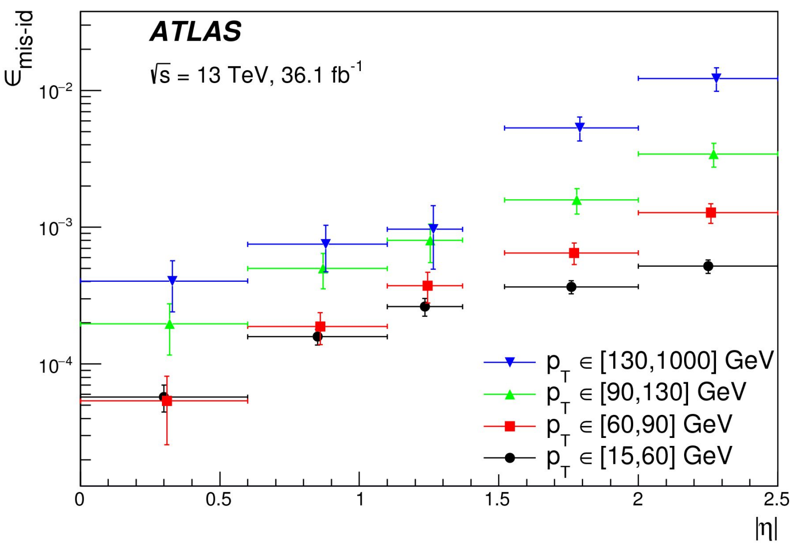
$N^{ss} = \frac{\varepsilon}{1-\varepsilon} N^{os}$ for the $e\mu$ channel,

ε is the rate of a single electron to be mis-identified.
N is the total number of true OS events

ε is measured in Z->ee events, by a likelihood method.

$$N_{ss}^{ij} = N^{ij}(\varepsilon_i + \varepsilon_j - 2\varepsilon_i\varepsilon_j).$$

Number of SS pair in Zee peak.



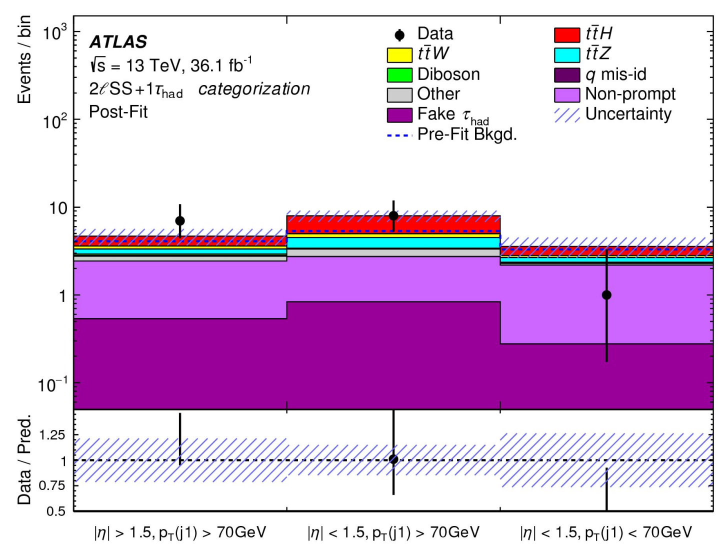
2lss+1tau: alternative

- Beside BDT, a cut-based analysis, categorisation, is also performed as a cross-check and alternative approach
  - Since it's a statistical limited channel, cut-based analysis could be more valid
- Categories set by cutting on two variables based on SR

- Cat 1 = maxeta2leadlepton > 1.5 && jet1pT> 70 GeV.
- Cat 2 = maxeta2leadlepton < 1.5 && jet1pT> 70 GeV.
- Cat 3 = maxeta2leadlepton < 1.5 && jet1pT< 70 GeV.

maxeta2leadlepton: maximum pseudo-rapidity of two leading leptons

Observed significance: 2.3 σ



Bkg composition