$HH \rightarrow bb\tau\tau$ ATLAS and CMS overview: comparison of reconstruction and background estimation methods

Francesco Brivio

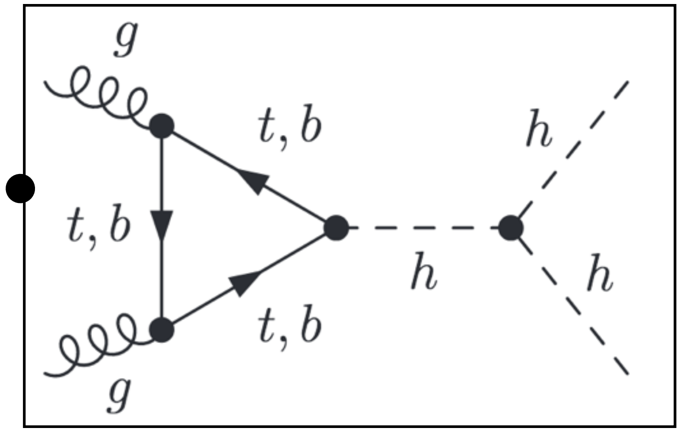
on behalf of the CMS and ATLAS Collaborations

ATLAS EXPERIMENT · CMS · INFN Istituto Nazionale di Fisica Nucleare · UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

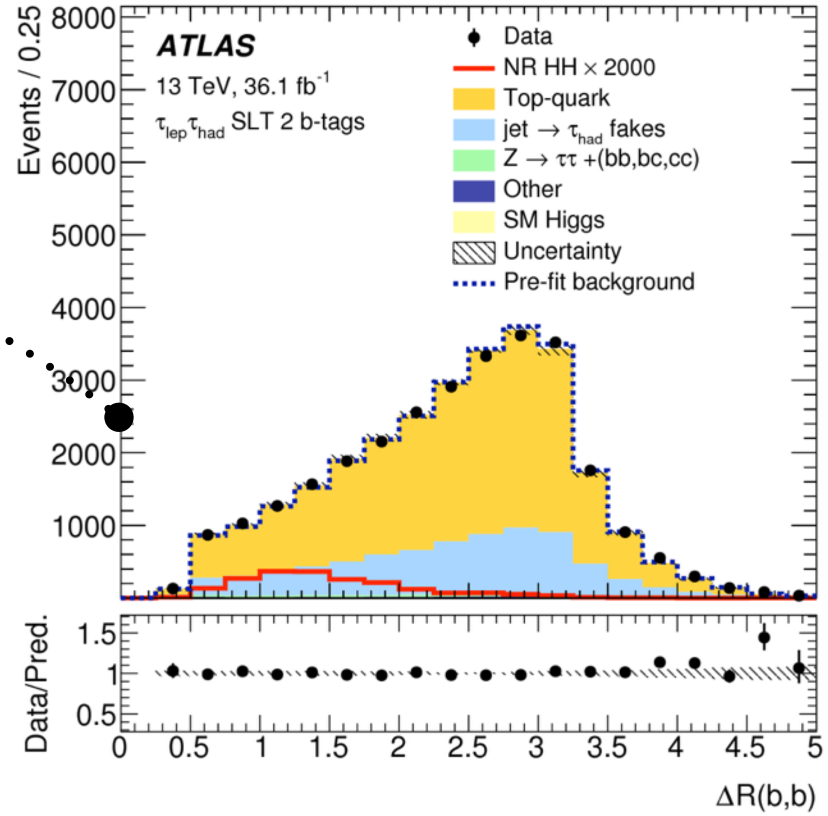
Double Higgs Production at Colliders Workshop

4-9th September 2018
Fermilab, USA

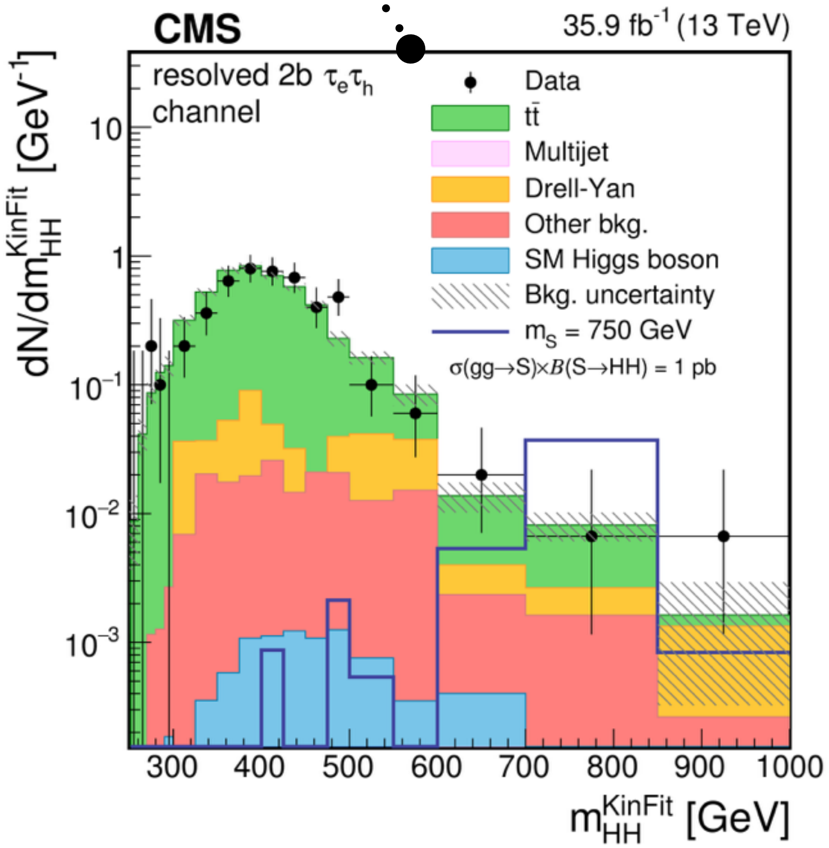
Outline



- **Introduction**
- **Object Reconstruction**



- **Background Estimation**

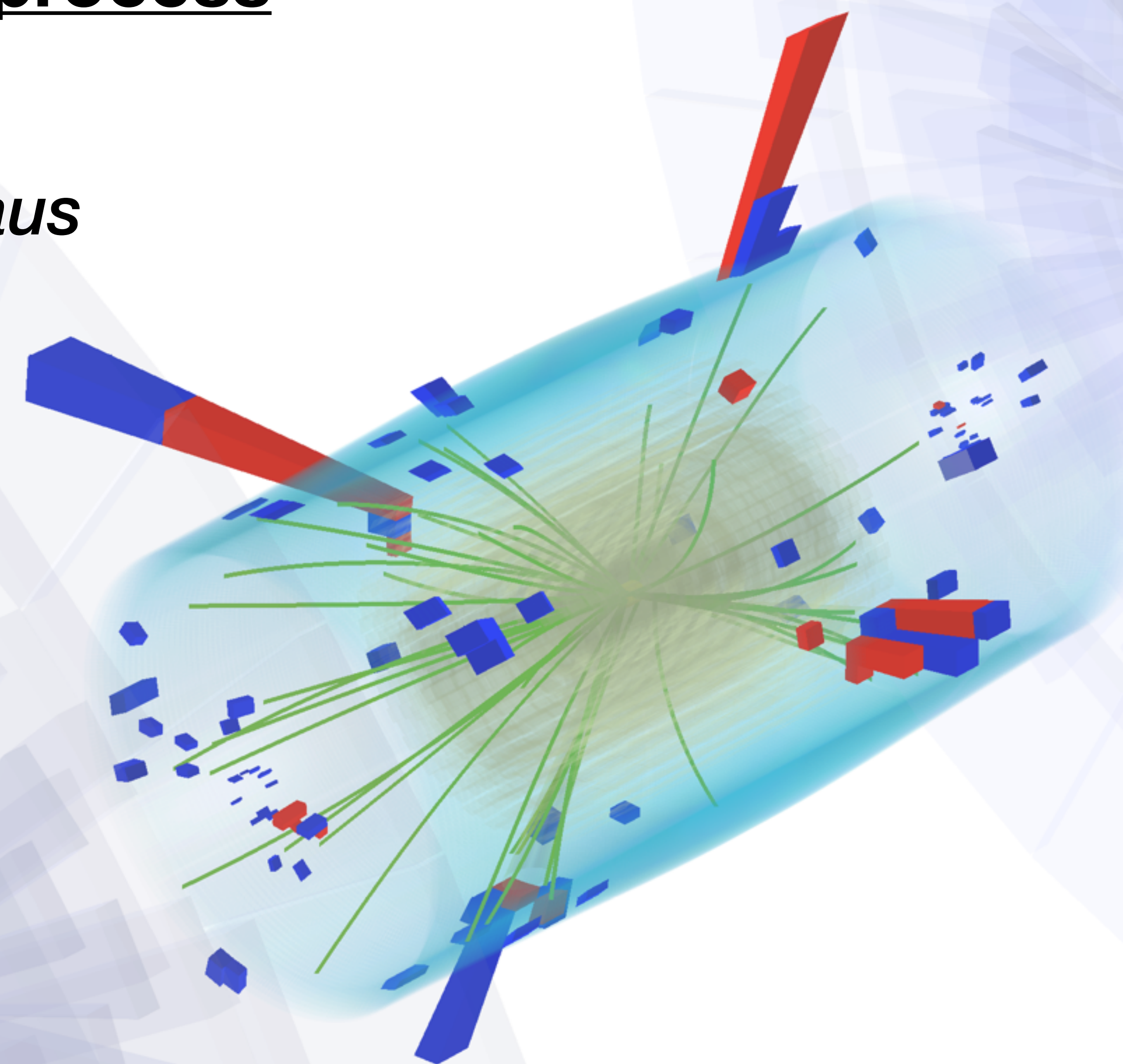


- **Conclusion**

Introduction

HH→bb$\tau\tau$ is a very complex decay process

- 4 objects in final state
- Two jets + hadronically decaying *taus*
- b-tagging needed
- All families of leptons (e, μ, τ)
- MET



Need to control the backgrounds

- Many processes giving the same final state
- Main backgrounds: $t\bar{t}$, DY+jets, QCD
- τ_h fakes

Object Reconstruction

Low Level Objects - Leptons

CMS strategy

- **Electrons**
  - Standard ParticleFlow algorithm: MVA technique (tracker + ECAL information)
  - Only isolated leptons considered
  - Efficiency: 80% (tight WP)
- **Muons**
  - Standard **ParticleFlow** algorithm
  - Only isolated leptons considered
  - Efficiency: ~90%
- **Hadronic Taus**
  - Hadron Plus Strips algorithm seeded by jets (1 or 3 tracks + calo deposits)
  - MVA discriminant for ID and ISO Medium WP: efficiency 60% and misidentification rate 0.1-1%

ATLAS strategy

- **Electrons**
  - Likelihood technique + track-hit requirements
  - Only isolated leptons considered
  - Eff/mis-ID rate: 87-95% / 0.15-0.5%
- **Muons**
  - Tracker+Muon systems information
  - Only isolated leptons considered
  - Eff/mis-ID rate: 96.1% / 0.17%
- **Hadronic Taus**
  - Seeded by jets (1 or 3 tracks associated)
  - Tau Vertex association algorithm
  - BDT against *quark/gluon-* jets Medium WP: 60/55% (ID only) misidentification rate 1%

Low Level Objects - Jets

CMS strategy

- **AK4 Jets**

  - *Anti-k_T* algorithm with distance parameter of 0.4

- **AK8 Jets**

  - *Anti-k_T* algorithm with distance parameter of 0.8

- **b-tagging**

  - Standard CMS b-tagging

  - Combined Secondary Vertex (CSV) algorithm to discriminate from *light quarks/gluon*-jets

  - WP: efficiency 60% and misidentification rate 1%

ATLAS strategy

- **AK4 Jets**

  - *Anti-k_T* algorithm with distance parameter of 0.4

- **AK8 Jets**

  - Not used -

- **b-tagging**

  - Standard ATLAS b-tagging

  - MV2c10 algorithm to discriminate from *light quarks* jets (trained also against *c*-hadrons)

  - WP: efficiency 70% and rejection factors 12/381/55 for c-jets/light-jets/taus

***More on b-tagging in Luca Mastrolorenzo's talk**

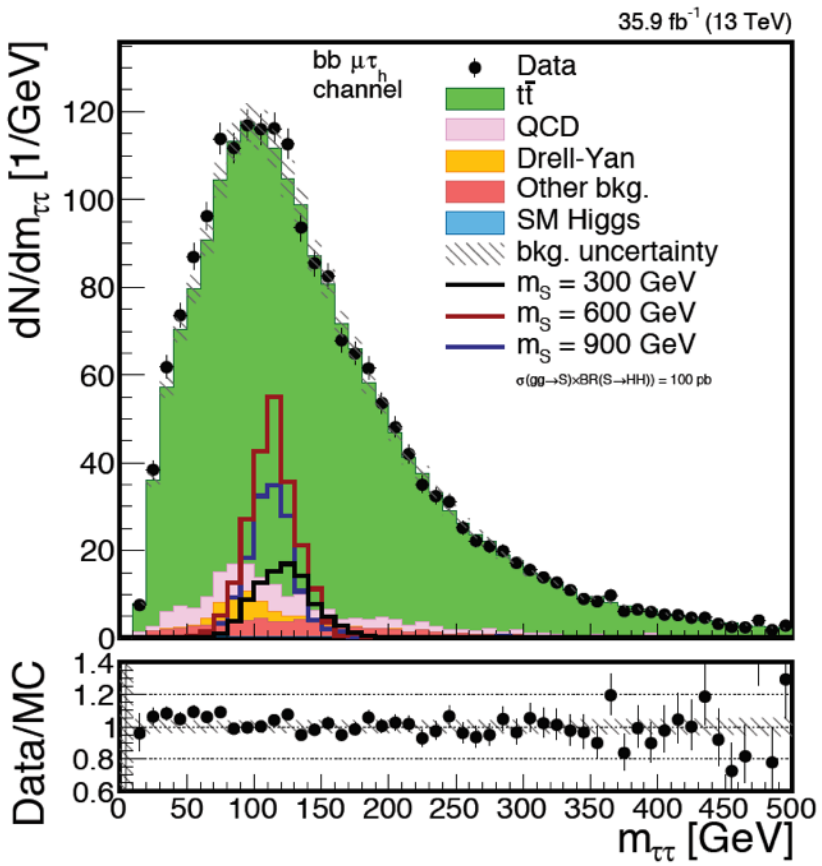
High Level Objects - $\tau\tau$ Invariant Mass

SVfit (2014 J. Phys.: Conf. Ser. 513 022035)

MMC (Nucl. Instrum. Meth. A 654 (2011) 481)

CMS strategy

- **SVfit**
  - Dynamic likelihood technique
  - Visible *taus* + MET

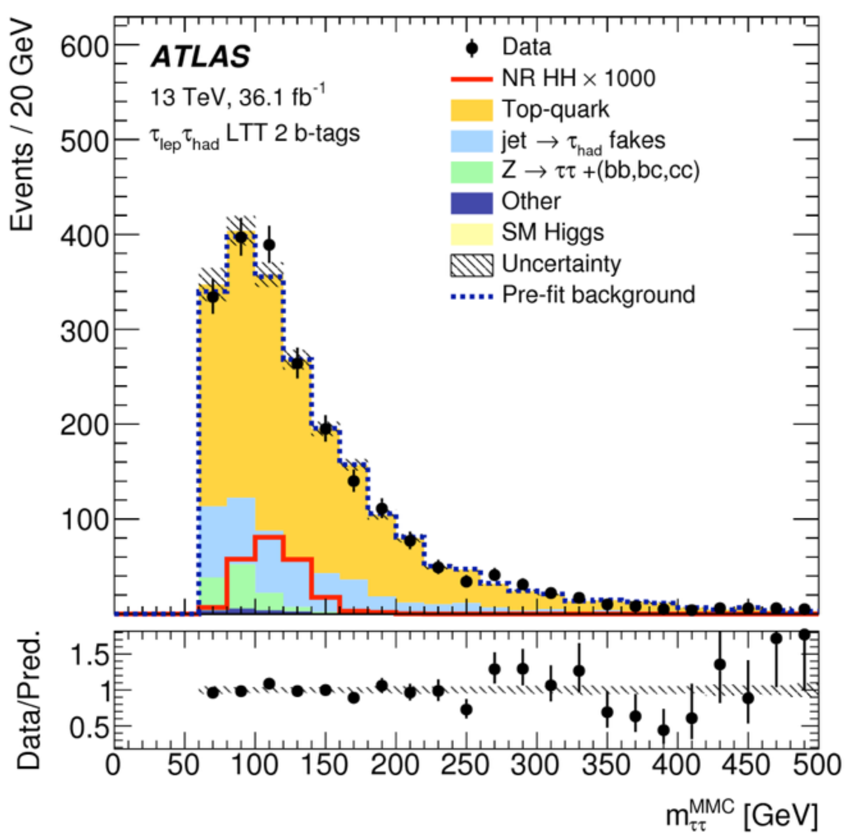


- **Elliptic cut on the invariant masses**

$$\frac{(m_{\tau\tau} - 116\,\mathrm{GeV})^2}{(35\,\mathrm{GeV})^2} + \frac{(m_{\mathrm{bb}} - 111\,\mathrm{GeV})^2}{(45\,\mathrm{GeV})^2} < 1$$

ATLAS strategy

- **Missing Mass Calculator (MMC)**
  - Dynamic likelihood technique
  - Visible *taus* + MET



- **Cut on the invariant mass**

$$m^{MMC} > 60\ \mathrm{GeV}$$

High Level Objects - $\tau\tau$ Invariant Mass

SVfit (2014 J. Phys.: Conf. Ser. 513 022035)

MMC (Nucl. Instrum. Meth. A 654 (2011) 481)

- **Similar starting point:**
  - Tau decay kinematics under-constrained from the observables
    - Event-by-Event Likelihood approach
    - Include information on the tau decay to find the bes estimate

CMS

$$P(M_{\tau\tau}) = \int \delta\left(M_{\tau\tau} - M_{\tau\tau}(\vec{y},\vec{a})\right)\, p(\vec{x}|\vec{y},\vec{a})\, d\vec{a}.$$

- Where: $\vec{x} = obs.\ MET$
  $\vec{y} = obs.\ \tau_{vis}$
  $\vec{a} = unknown\ \tau\ kinem.$
- Scan in steps of $\delta m_{\tau\tau}$ for all possible configurations compatible with measured observables
- Best estimate from the maximization of the likelihood

ATLAS

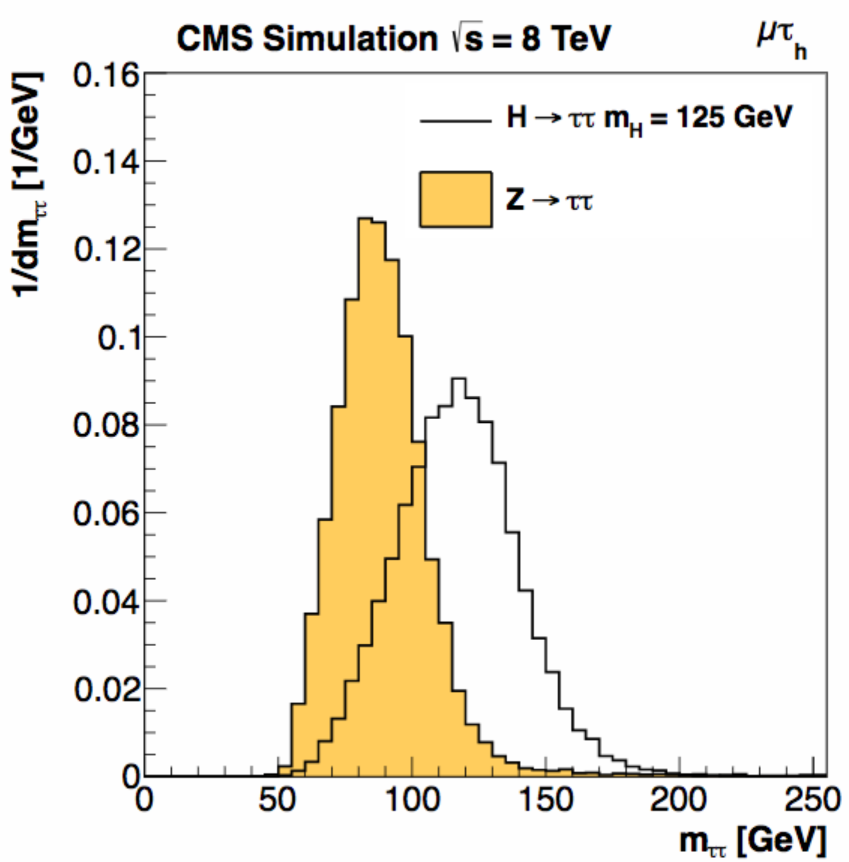
$$\mathcal{L} = -\log\left(\mathcal{P}(\Delta R_1, p_{\tau 1}) \times \mathcal{P}(\Delta R_2, p_{\tau 2}) \times \mathcal{P}(\Delta \not{E}_{\mathrm{T}_x}) \times \mathcal{P}(\Delta \not{E}_{\mathrm{T}_y})\right),$$

- Where: $\mathcal{P}(\Delta \not{E}_T) = MET\ resolution\ pdf$
  $\mathcal{P}(\Delta R, p_T) = \tau\ decay\ kinem.\ pdf$
- Scan all points of the 3D hyper-plane (ϕ_{mis1}, ϕ_{mis2}, m_{mis}) to build a $m_{\tau\tau}$ distributions
- Most probable value of the distribution used as best estimate

High Level Objects - $\tau\tau$ Invariant Mass

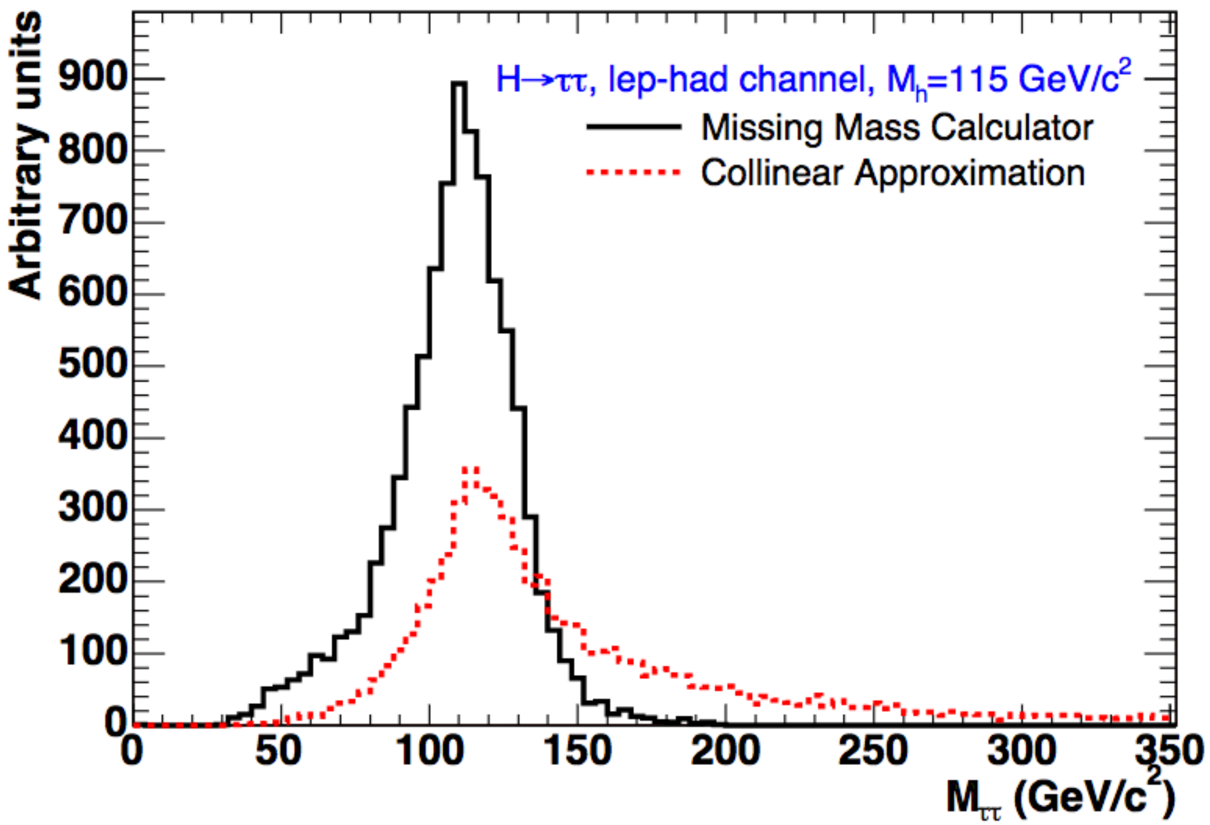
- **Expected performances from the original MMC and SVfit papers**

CMS



- More symmetrical shape for the Higgs distribution
- Wider distribution with tails on both sides

ATLAS

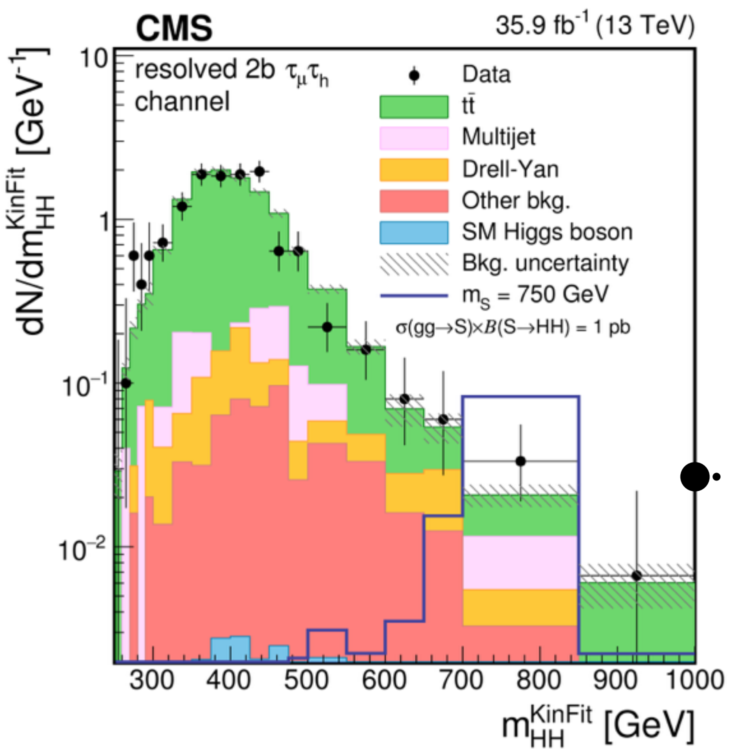


- Very asymmetrical shape for the Higgs distribution
- Longer tail on the left side which is cut out by requiring m^{MMC}>60 GeV

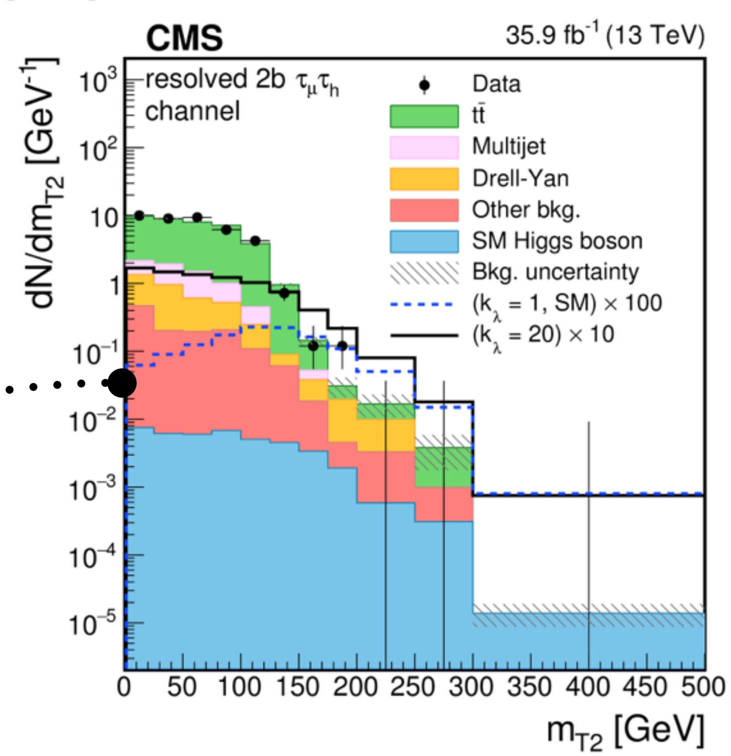
The only way to fully appreciate the differences is to compare the performances directly on the same set of events

High Level Objects - Discriminating Variables

CMS strategy



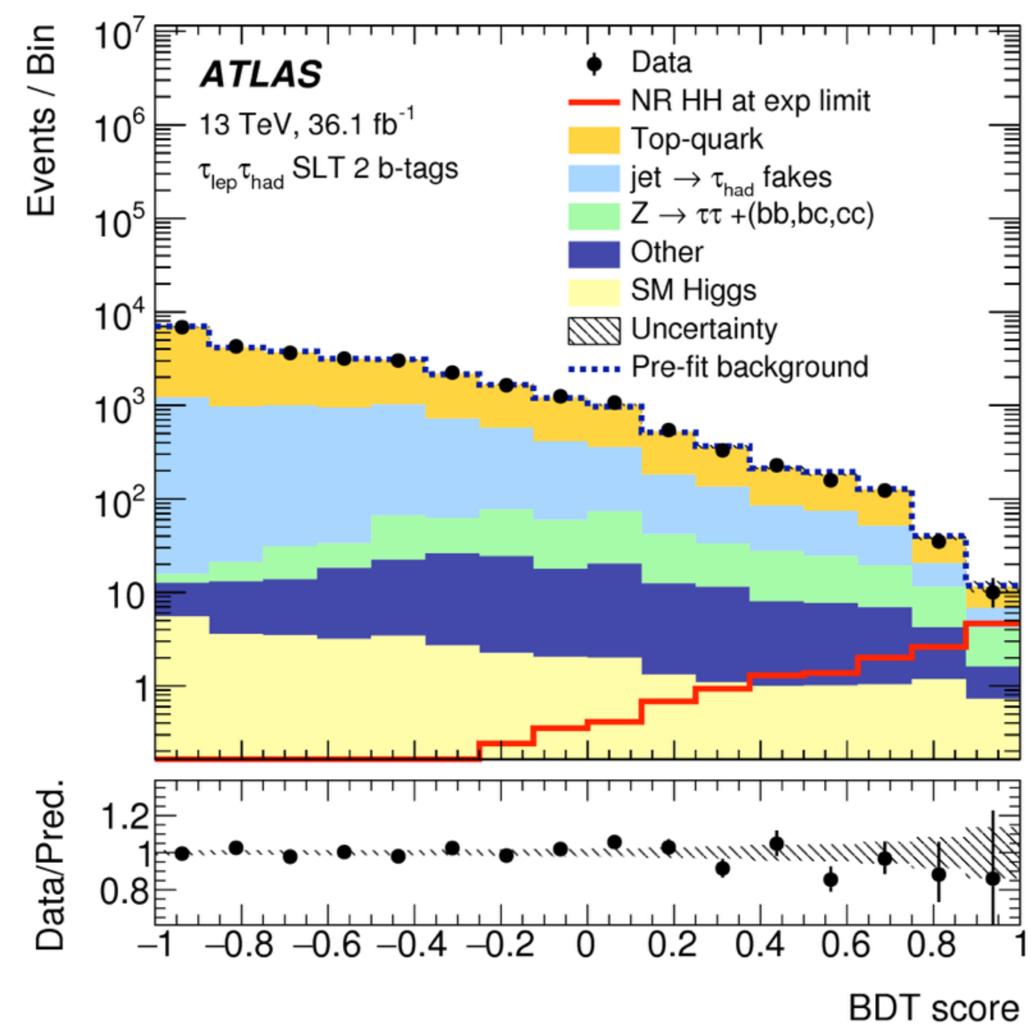
Resonant: m_{HH}^{KinFit}



Non- Resonant: MT2

ATLAS strategy

BDT Score



*More in Arnaud's talk

Objects Reconstruction Conclusion

- **Overall very similar reconstruction methods**
  - For all objects involved (leptons, jets…)
  - With obvious differences due to different detectors
  - Slight differences in the working points chosen
    (efficiencies and mistag rate of tau identification and b-tagging)

- **Main difference in the reconstruction of the $\tau\tau$ pair mass**
  - SVfit vs Missing Mass Calculator
  - Would be nice to directly compare the performances of the 2 algorithms
    - with respect to one another
    - with respect to the visible $\tau\tau$ mass

Background Estimation

Background Estimation - MC simulation

- Backgrounds estimated using MC simulation only

CMS		
Background	**Generator**	**XS precision**
$t\bar{t}$	POWHEG 2.0	NNLO
Single Top	POWHEG 2.0	NNLO
Single Higgs	MADGRAPH5	NNLO
W + jets	MADGRAPH5	NNLO
Di-boson	MADGRAPH5	NLO
DY+jets*	MADGRAPH5	NNLO

ATLAS		
Background	**Generator**	**XS precision**
$\bar{t}t$*	POWHEG-BOX	NNLO+NNLL
Single Top	POWHEG-BOX	NLO
ttH **ZH**	MADGRAPH POWHEG	NLO NLO+NLL
W + jets	SHERPA	NNLO
Di-boson	SHERPA	NLO
DY+jets*	SHERPA	NNLO

* corrected with data-driven methods

Background Estimation - DY background

- **Drell-Yan is one of the most prominent backgrounds, especially when associated with the production of 2 jets**

CMS strategy

- **From MC simulation**

  Corrected with SF from control regions

- **Control regions**

  $Z \to \mu\mu$ + 2 jets

  - events with selection similar to SR (mass and isolation)
  - binned in number of b-jets (0,1,2)

- **SF extraction**

  - Simultaneous fit on the 3 regions
  - Fit on $m_{\mu\mu}$
  - $m_{\mu\mu}$ between 60 and 120 GeV

	Z + light jets	Z + one b jet	Z + two b jets
Scale factors	1.1412 ± 0.0017	1.187 ± 0.015	1.170 ± 0.029

ATLAS strategy

- **From MC simulation**

  Corrected with SF from control regions

- **Normalization**

  From $Z \to \mu\mu$ + heavy-flavor CR

  - events with selection similar to SR
  - $m_{\mu\mu}$ between 81 and 101 GeV
  - m_{bb} below 80 GeV or above 140 GeV to remove SM ZH

- **Normalization extraction**

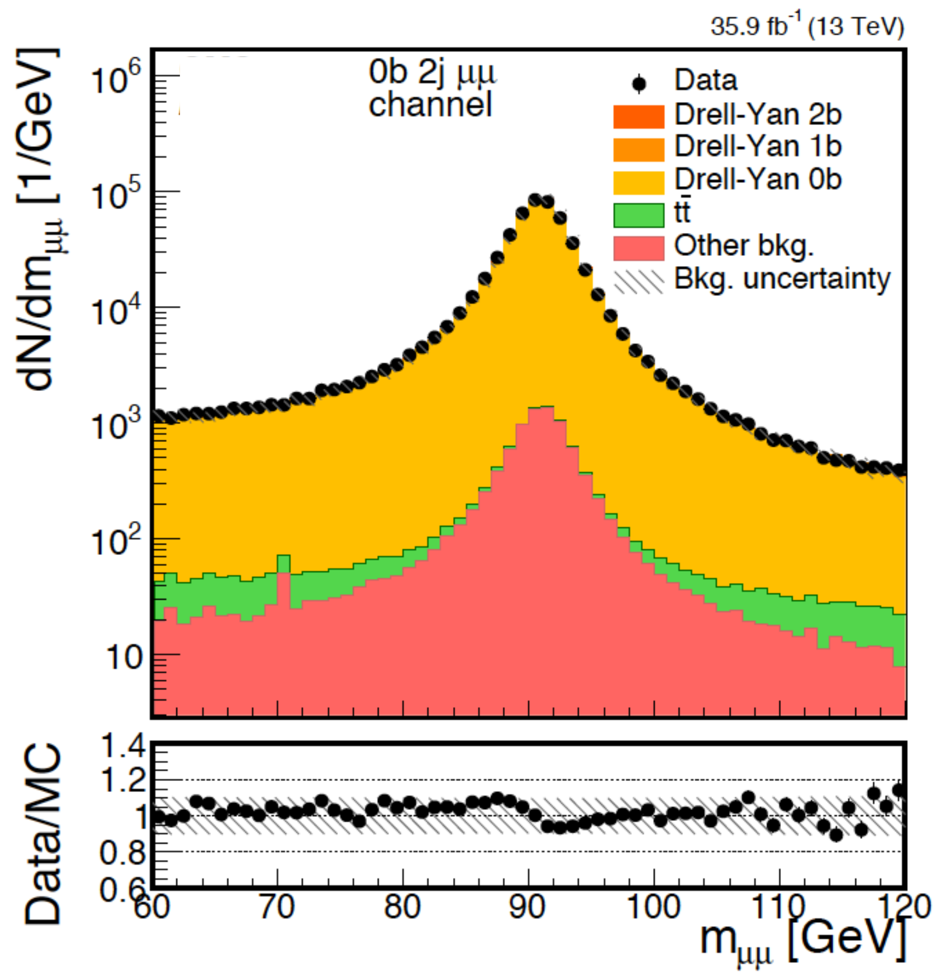
  Including this CR yield in the final fit

Normalization factor
1.34 ± 0.16

Background Estimation - DY background

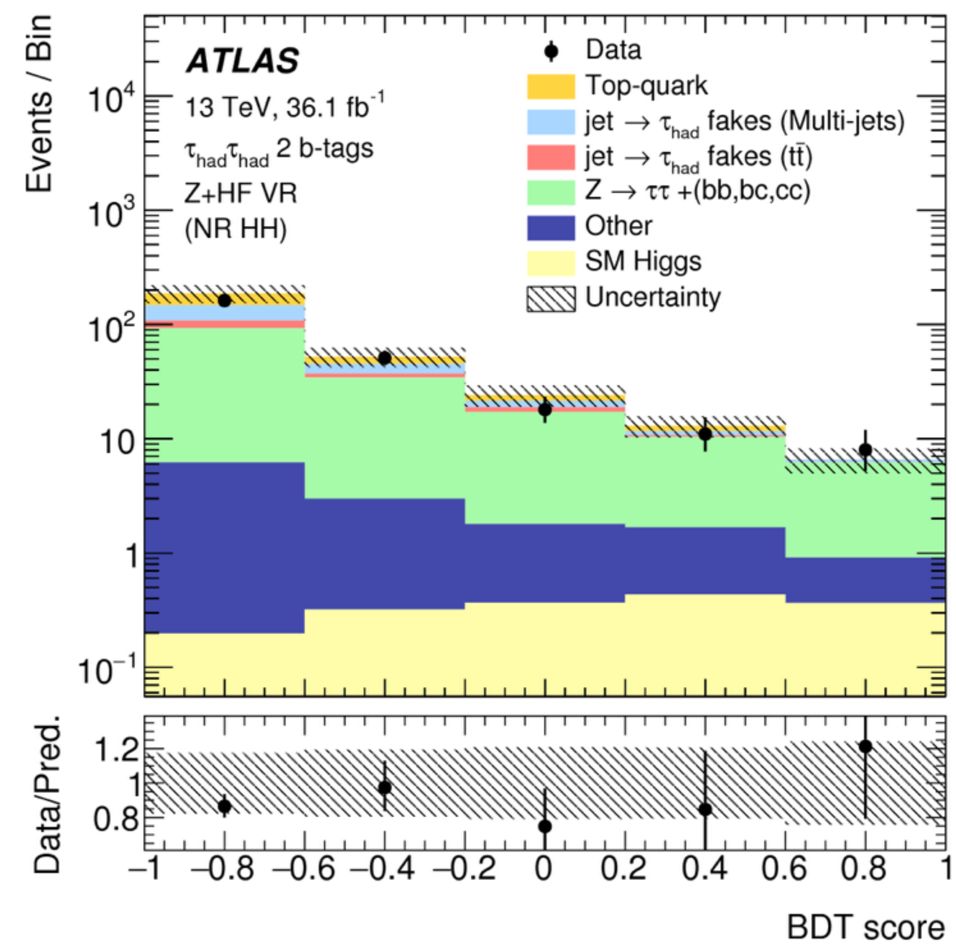
- Validation/Control Regions for the DY+ jets estimation

CMS strategy



- **Z→$\mu\mu$** + 2 jets, 0 b-tag control region
- After the application of the SF

ATLAS strategy

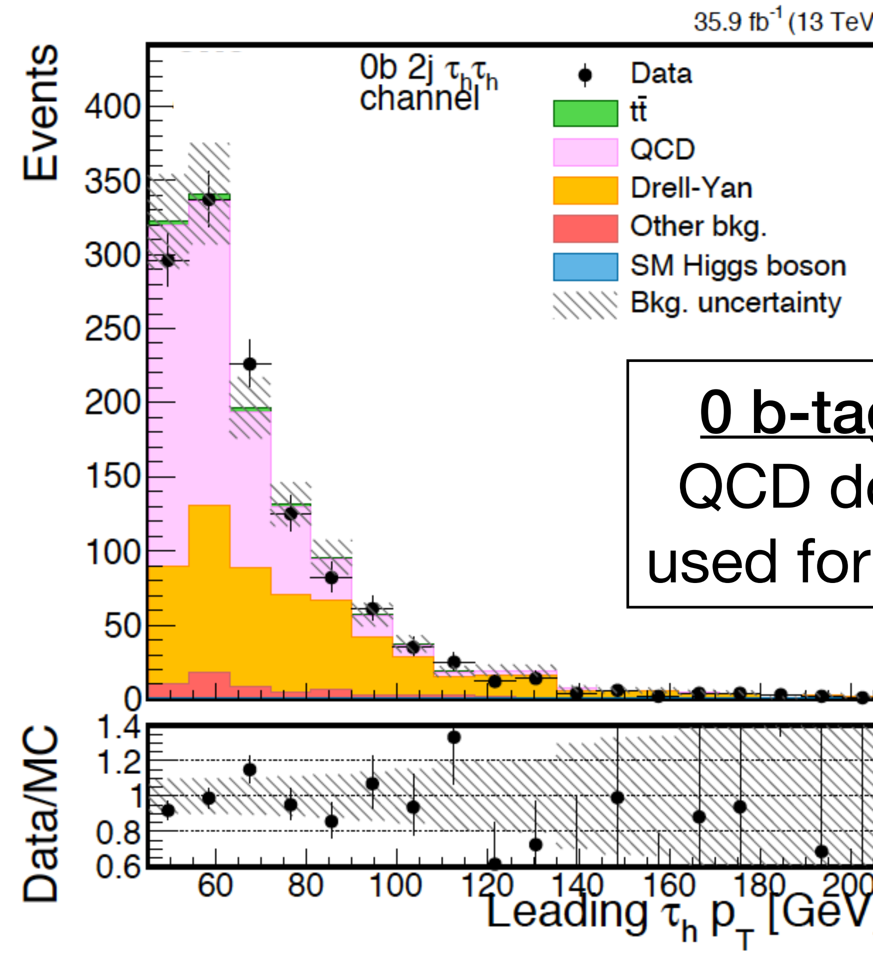


- Z+HeavyFlavor validation region

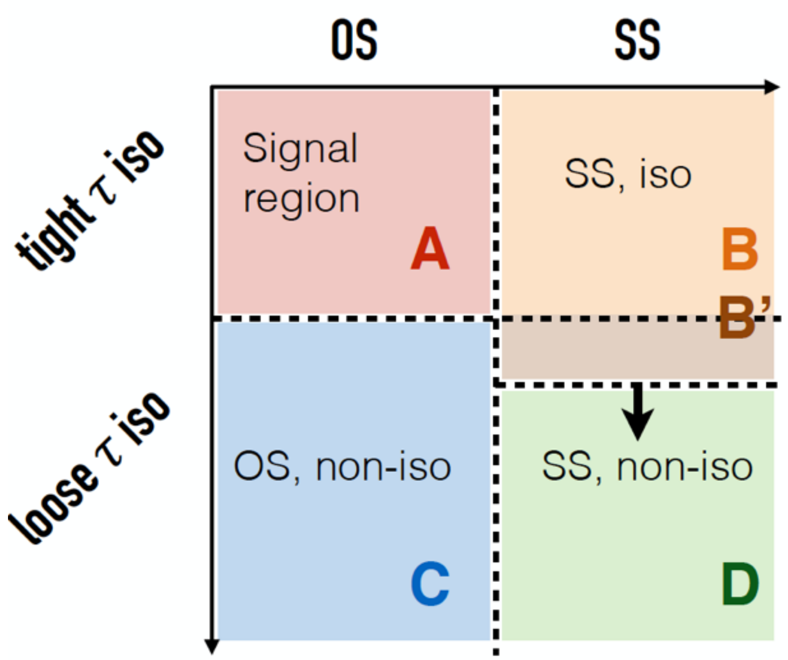
Background Estimation - Multijet background

CMS strategy

- **ABCD method**
  - Both yield and shape from jet enriched regions in data
  - Other background contributions subtracted from MC simulation
  - No parametrization in pT, eta, #tracks



0 b-tag region
QCD dominated
used for validation



Isolation inverted only on the sub-leading τ lepton

shape: B'
yield: B * C/D

Shape
From a SS relaxed
τ lepton isolation region
(due to statistics reasons)

Yield
From SS region
corrected by ratio OS/SS in
jet enriched region

Background Estimation - Fake *taus*

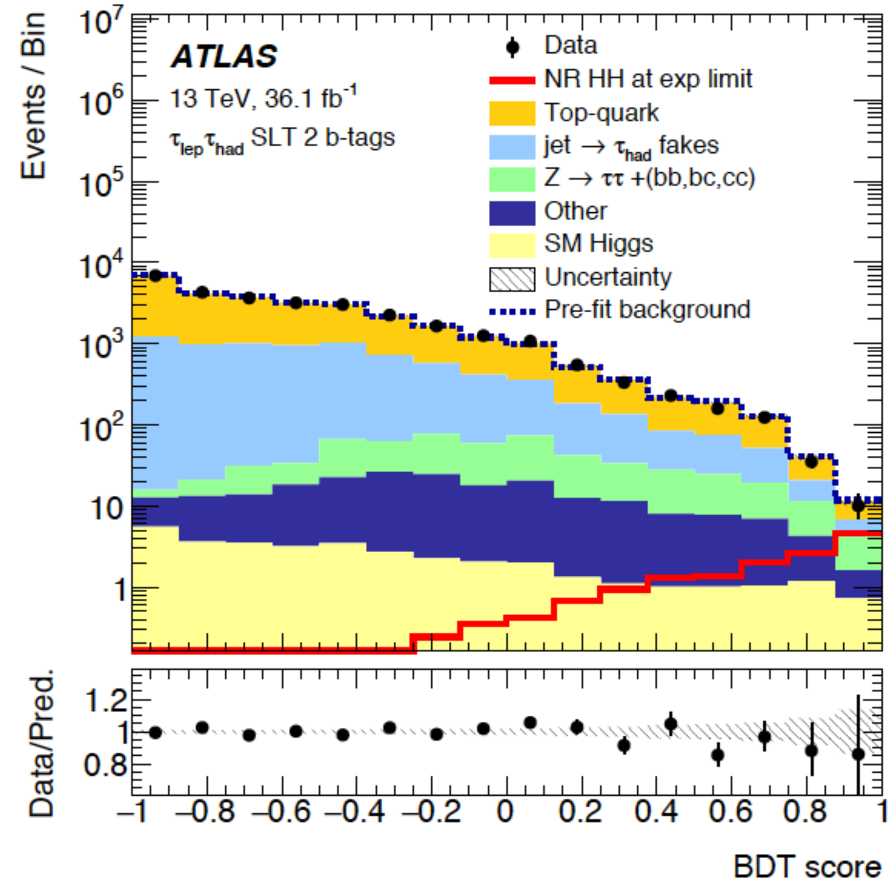
ATLAS strategy

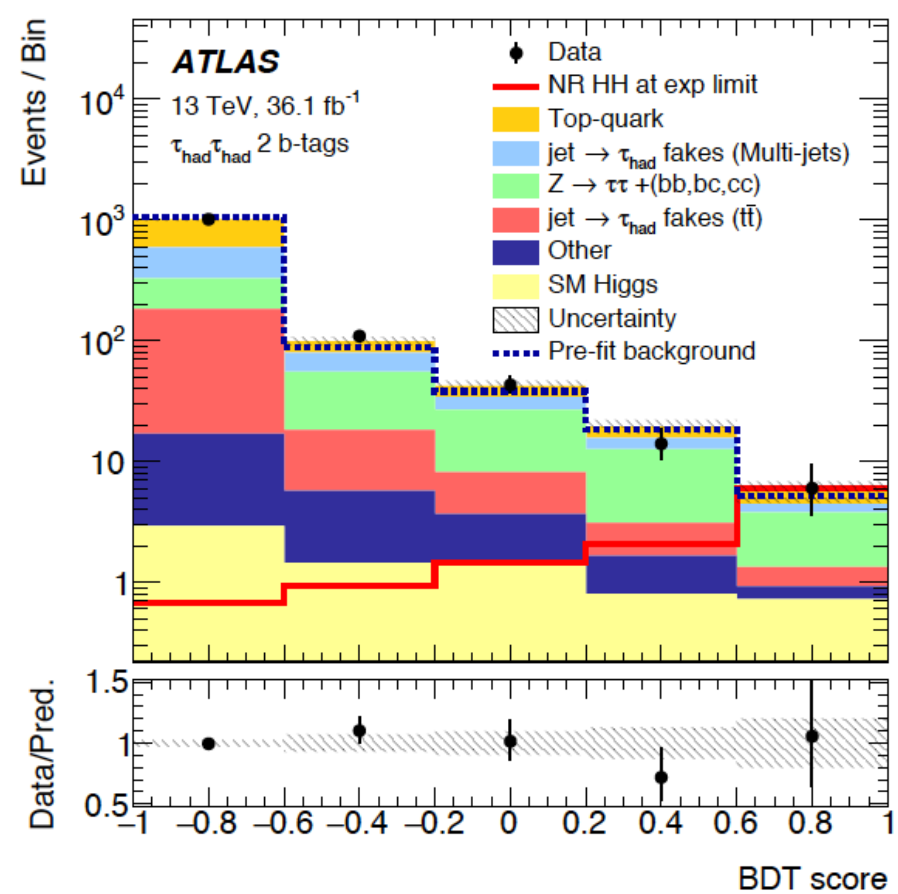
- **Fake *taus* estimated using FakeRates and FakeFactor methods**

  - From fake *taus* enriched sample as similar as possible to the SR:
    - tau identification < “medium” WP
    - “very loose” requirement on the BDT score
  - **Templates** from this enriched region after subtraction of real *taus* (from MC simulation)
  - **FakeFactors** to scale the templates
    - Extracted from Control Regions
    - Binned in p_T and number of associated tracks

$$FF = \frac{\#\text{jets pass tauID}}{\#\text{jets fail tauID}}$$

$$FR = \frac{\#\text{jets pass tauID}}{\#\text{all } \tau \text{ candidates}}$$





Background Estimation - Fake *taus*

ATLAS strategy

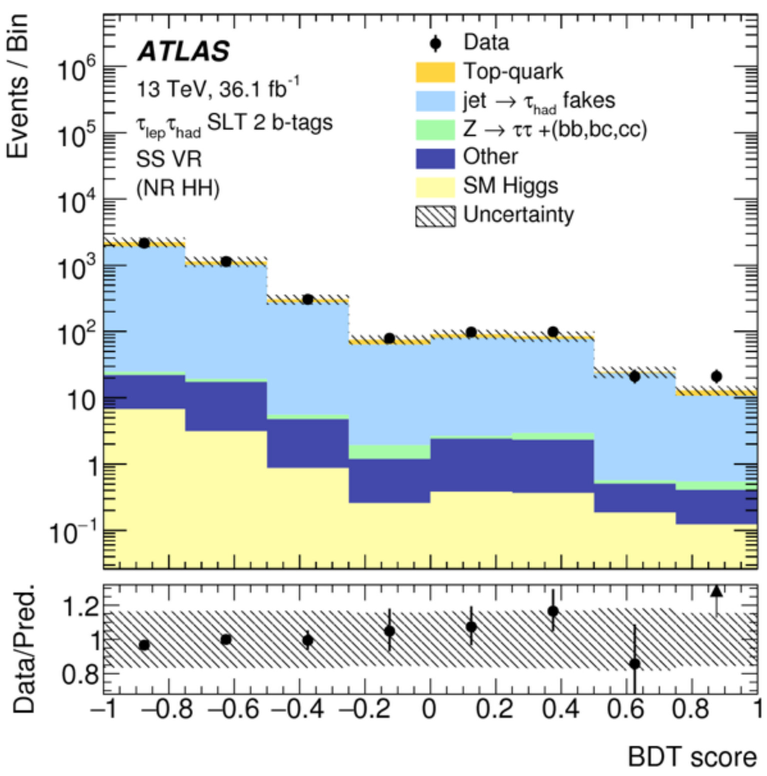
- **Semileptonic case**
  - Full estimation using FF derived for multijet, $t\bar{t}$ and W+jets
  - FF combined to account for different *quark/gluon* fractions

$$\mathrm{FF(comb)} = \mathrm{FF(QCD)} \times r_{\mathrm{QCD}} + \mathrm{FF}(t\bar{t}/W + \mathrm{jets}) \times (1 - r_{\mathrm{QCD}})$$

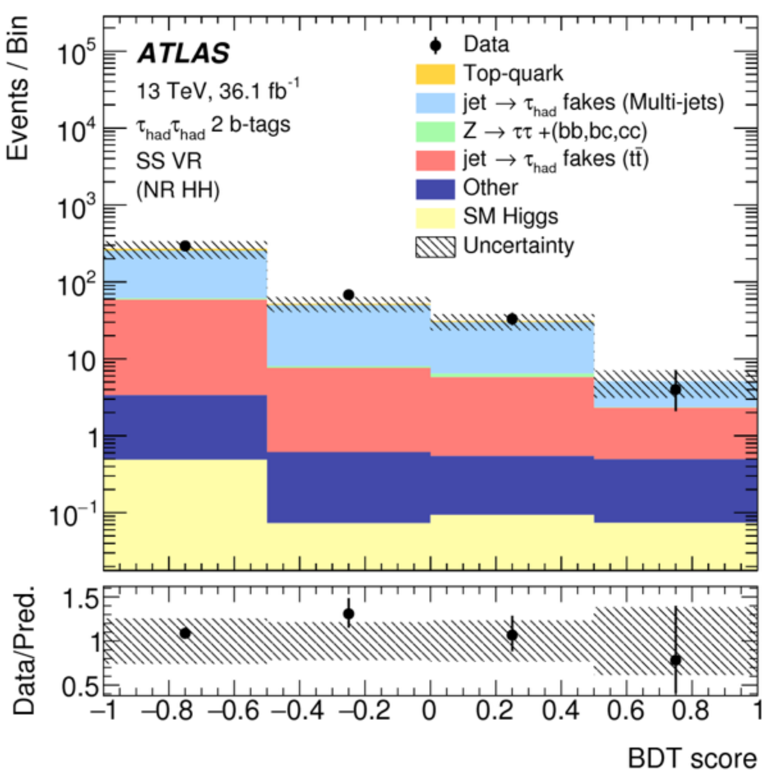
  - Validated in Same Sign region

- **Fully hadronic case**
  - Multijet background: same as semileptonic case
    - templates from 1 b-tag SS control region
    - FF from 2 b-tag SS control region
  - $t\bar{t}$ Background (≥1 fake tau from W decay):
    - from MC simulation
    - corrected by FR measured in semileptonic control region region (parametrized as function of *eta* number of *prongs*)

Semilep. VR



Hadronic VR



Background Estimation Conclusion

- **Backgrounds from MC simulation only**
  - Same backgrounds for CMS and ATLAS
  - Same generator for the “top-quark backgrounds” (POWHEG)
  - Mixture of generators for the others (MADGRAPH, SHERPA)
  - Similar precision on the cross sections

- **Similar approach for the DY+jets background**
  - Simulation corrected with scale factors from $Z \rightarrow \mu\mu$ events

- **Main difference in the estimation of multijet/fake-τ_h backgrounds**
  - **CMS:** ABCD for multijet background only
  - **ATLAS:** FakeFactor for treatment of fake-τ_h from multijet and $t\bar{t}$

  - A comparison between the two methods is not as easy in this case…
    - Different approach but good data/MC agreement for both experiments

Systematic Uncertainties

Systematic Uncertainties Comparison

- Similar systematic uncertainty sources considered in both experiments
- Shown here:
  - **CMS:** systematic uncertainties affecting the normalization of different processes.
  - **ATLAS:** impact on the simulated non-resonant signal strength, *i.e.* the expected yield assuming σ x *BR* equal to the expected exclusion limit (14.8 times SM prediction)

CMS

PLB paper link

Systematic uncertainty	Value
Luminosity	2.5%
Lepton trigger and reconstruction	2–6%
τ energy scale	3–10%
Jet energy scale	2–4%
b tag efficiency	2–6%
Background cross section	1–10%
$Z/\gamma^* \to \ell\ell$ SF uncertainty	0.1–2.5%
Multijet normalization	5–30%
Scale unc.	+4.3%/−6.0%
Theory unc.	5.9%

ATLAS

arXiv link

Experimental Uncertainties	
Luminosity	± 2.4
Pileup reweighting	± 1.7
τ_{had}	± 16
Fake-τ estimation	± 8.4
b-tagging	± 8.3
Jets and $E_{\mathrm{T}}^{\mathrm{miss}}$	± 3.3
Electron and muon	± 0.5
Theoretical and Modeling Uncertainties	
Top	± 17
Signal	± 9.3
$Z \to \tau\tau$	± 6.8
SM Higgs	± 2.9
Other backgrounds	± 0.3

Summary
and
Conclusions

Summary and Conclusion

- **Overall similar strategies about:**
  - Objects reconstructions
  - Background estimation

- **Two main exceptions:**
  1. Reconstructions of the $\boldsymbol{\tau\tau}$ pair mass:

     two different algorithm (**SVfit** & **MMC**), but with the same "*philosophy*":
     combine visible *taus* and MET with information about the tau decay kinematics

  2. Estimation of the multijet/fake-τ_{h} background:

     Different "*philosophy*" behind the estimation methods
     Different treatment of the fake-τ_{h} contribution

- **Suggested improvements:**
  - **ATLAS & CMS:** Explicitly compare SVfit and MMC
  - **CMS:** Explore the effect of jet—>fake-τ_{h} as function of (p_{T}, eta…) and from different background sources (*e.g.* $t\bar{t}$)

Backup

Documentation

- **CMS documentation**
  1. PLB paper link
  2. Auxiliary material

- **ATLAS documentation**
  1. arXiv link
  2. Auxiliary material

Systematic Uncertainties Comparison

CMS

Systematic uncertainty	Value	Processes
Luminosity	2.5%	all but multijet, $Z/\gamma^* \to \ell\ell$
Lepton trigger and reconstruction	2–6%	all but multijet
τ energy scale	3–10%	all but multijet
Jet energy scale	2–4%	all but multijet
b tag efficiency	2–6%	all but multijet
Background cross section	1–10%	all but multijet, $Z/\gamma^* \to \ell\ell$
$Z/\gamma^* \to \ell\ell$ SF uncertainty	0.1–2.5%	$Z/\gamma^* \to \ell\ell$
Multijet normalization	5–30%	multijet
Scale unc.	+4.3%/−6.0%	signals
Theory unc.	5.9%	signals

ATLAS

Source	Uncertainty (%)
Total	± 54
Data statistics	± 44
Simulation statistics	± 16
Experimental Uncertainties	
Luminosity	± 2.4
Pileup reweighting	± 1.7
τ_{had}	± 16
Fake-τ estimation	± 8.4
b-tagging	± 8.3
Jets and E_T^{miss}	± 3.3
Electron and muon	± 0.5
Theoretical and Modeling Uncertainties	
Top	± 17
Signal	± 9.3
$Z \to \tau\tau$	± 6.8
SM Higgs	± 2.9
Other backgrounds	± 0.3