Boosted Decision Trees and b-jet Trigger Calibration Studies for $t\bar{t}H(b\bar{b})$ Fully Hadronic Analysis

Bartolini Giovanni

CPPM
University of Aix-Marseille

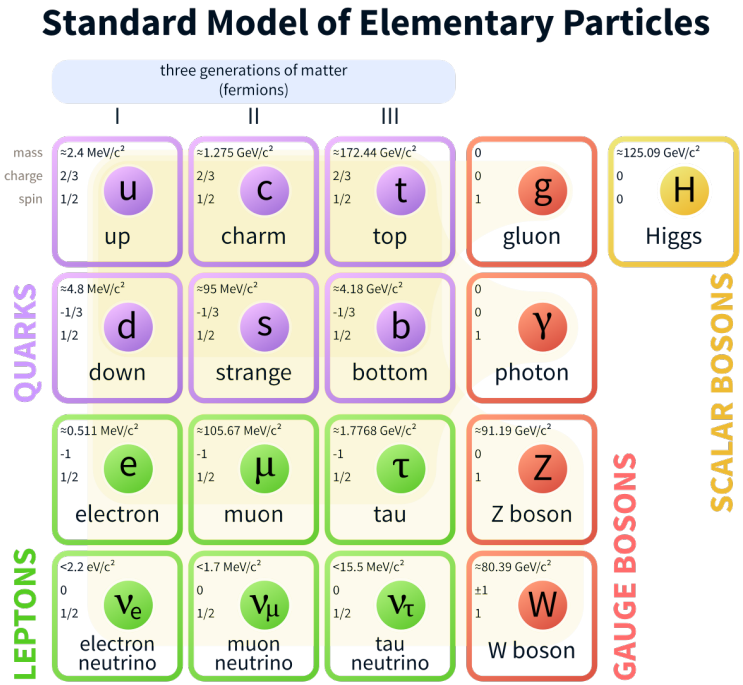
October 18, 2018

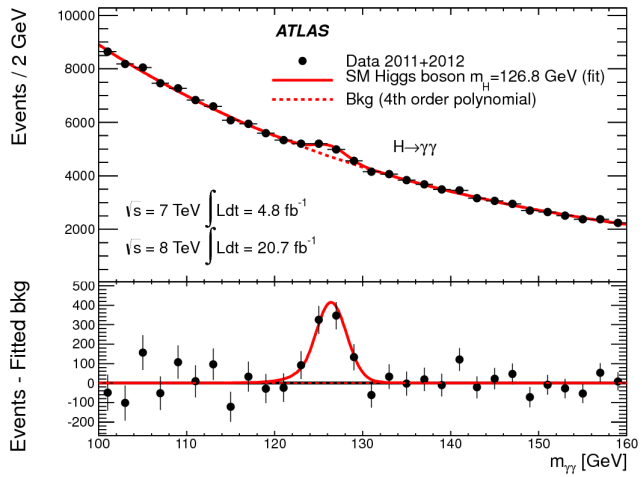
Outline

Introduction

The Standard Model of Particle Phisics

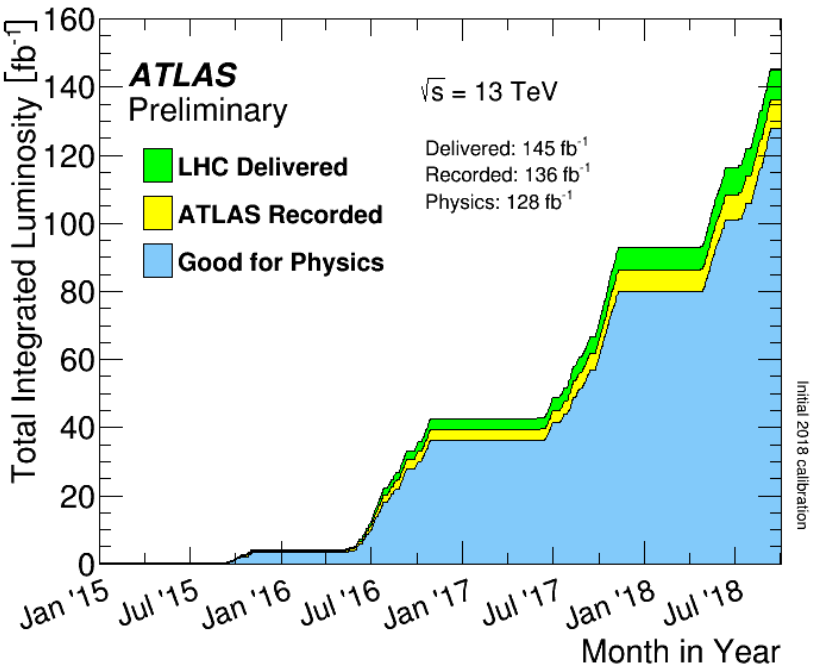
- The **Standard Model** (SM) of Particle Physics is a gauge theory that classifies all known elementary particle and describes 3 of the 4 known interaction forces: Strong, Weak and Electromagnetic
- Discovery of **Higgs** boson in 2012 completed the set of predicted elementary particle
- Very successful and predictive theory, but has still many shortcomings:
  - inclusion of gravity
  - neutrino masses and oscillations
  - matter/anti-matter asymmetry
  - evidence of dark matter existence
- Strong chase for New Physics
  - **direct search** of new particles
  - **indirect evidence** trough deviations in SM predictions

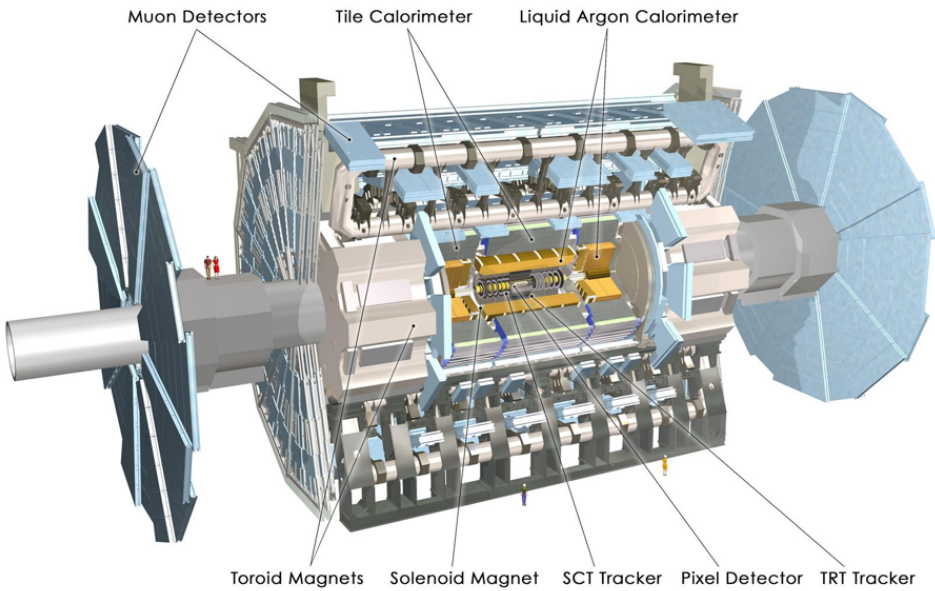




The ATLAS experiment at LHC

- The ATLAS experiment is placed in one of the 4 interaction point of LHC
- Almost at the end of Run 2:
  - already collected $\sim$130 fb^{-1}

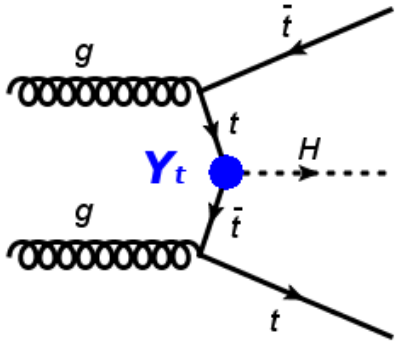


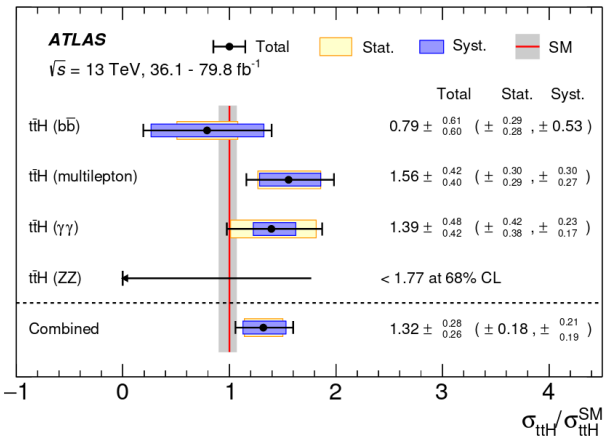


Higgs Boson and Top Yukawa Coupling

- The discovery of the **Higgs** boson in 2012 started an effort on the precise measure of its properties

- The **top** quark is the heaviest elementary particle
  - has the highest Yukawa coupling: Y_t
- Anomalous values for Y_t are predicted by many *Beyond the Standard Model* (BSM) theories
- Associated production ($t\bar{t}H$) only way to directly measure Y_t
  - many accessible final states: $\gamma\gamma$, multi-lepton, lepton+jets, all hadronic
- ATLAS first $t\bar{t}H$ observation recently published using Run 1 and Run 2 data*
  - with observed(expected) significance of 6.3(5.1)



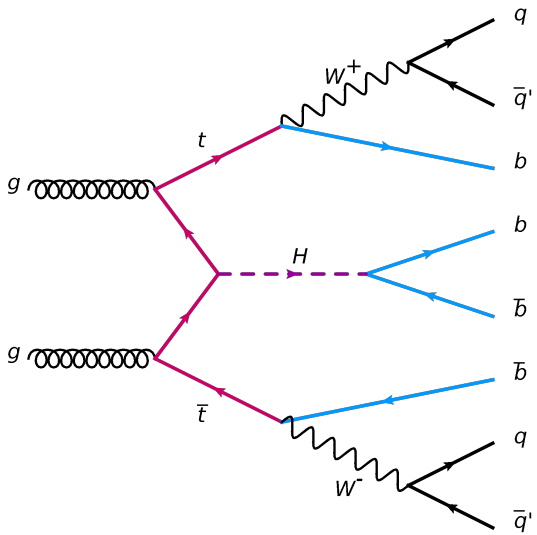


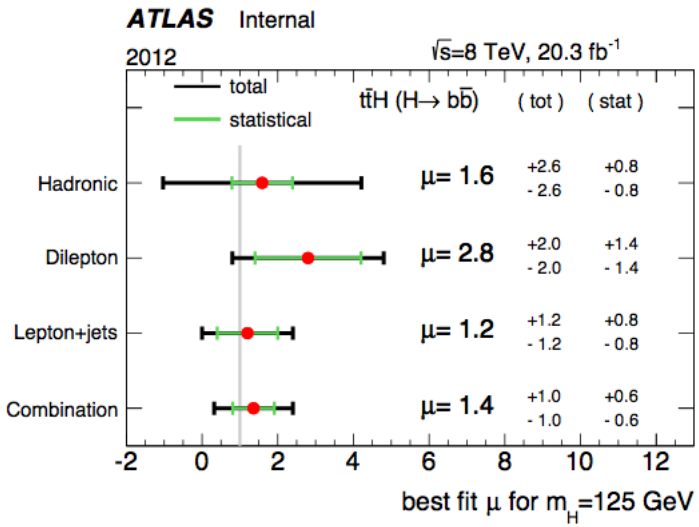
*arXiv:1806.00425 [hep-ex]

Fully Hadronic ttH Analysis Strategy

$t\bar{t}H$ Production in Fully Hadronic Final State

- Most abundant final state
  - $\sim$33% of total $t\bar{t}H$ production
- No neutrino
  - full event reconstruction
- Ideal for differential analysis
  - explore the CP nature of Y_t
- Challenging experimental signature:
  - 8 quarks, 4 b-quarks
  - Large QCD background
  - irreducible $t\bar{t} + b\bar{b}$ background
- First analysis published in Run 1*:
  $\mu_{t\bar{t}H} = 1.6 \pm 2.6$
  - brought 10% improvement on the significance of the $t\bar{t}H(H \to b\bar{b})$ combination

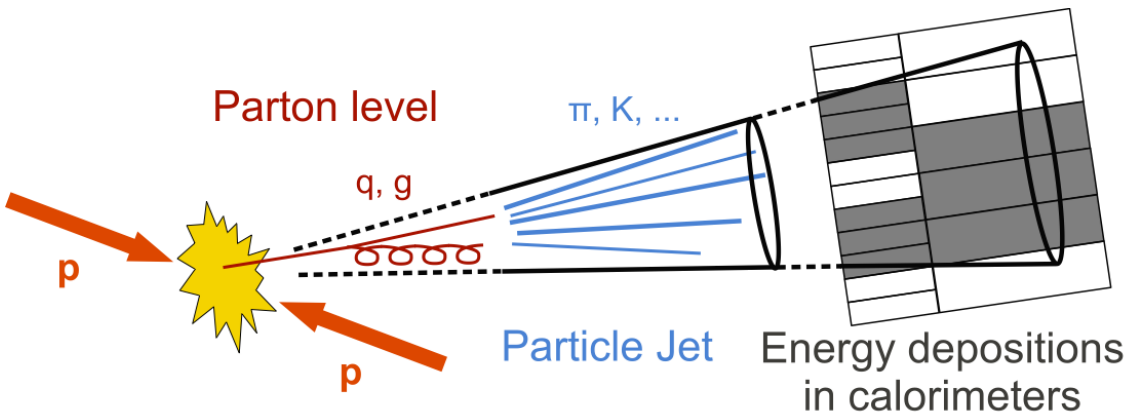


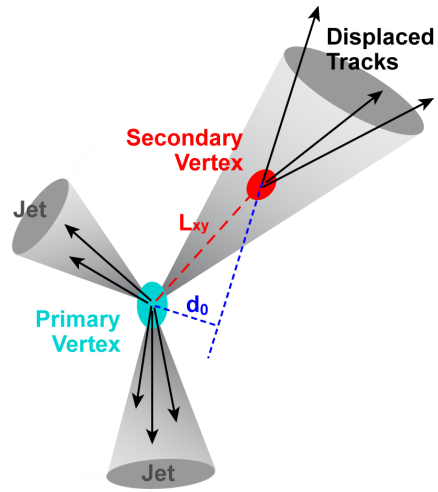


* arXiv:1604.03812 [hep-ex]

Jets as Quark objects

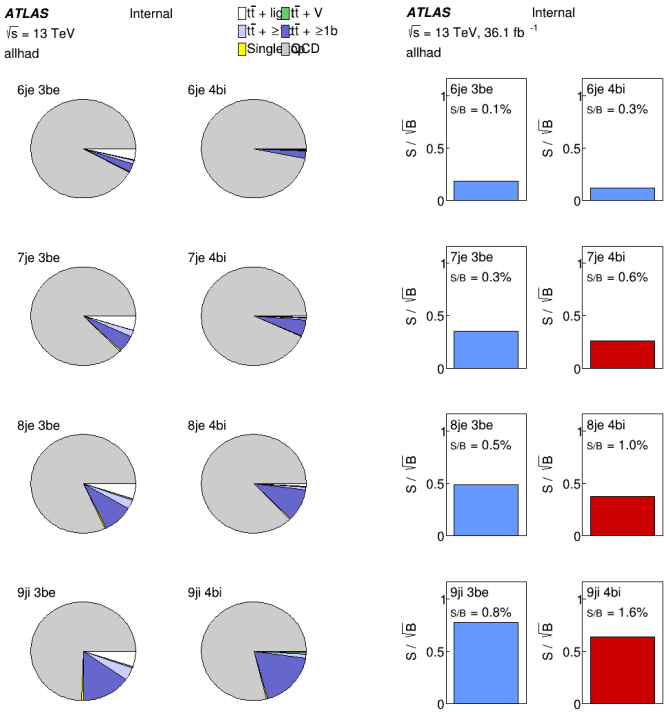
- The nature of the strong interaction does not allow to quarks to be available as free particles
  - How can we see quarks as final state particles?
- Quarks and gluons that are produced from a collision will *hadronize* producing a collimated flow of hadronic particles
  - a specific algorithm is used to reconstruct this flow as a single object, called **JET**
- Jets coming from *b* quarks have particolar properties that allow them to be distinguished from other jets
  - the identification of a jet as coming from a *b* quark is called *b*-**tagging**





Event Selection and Categorization

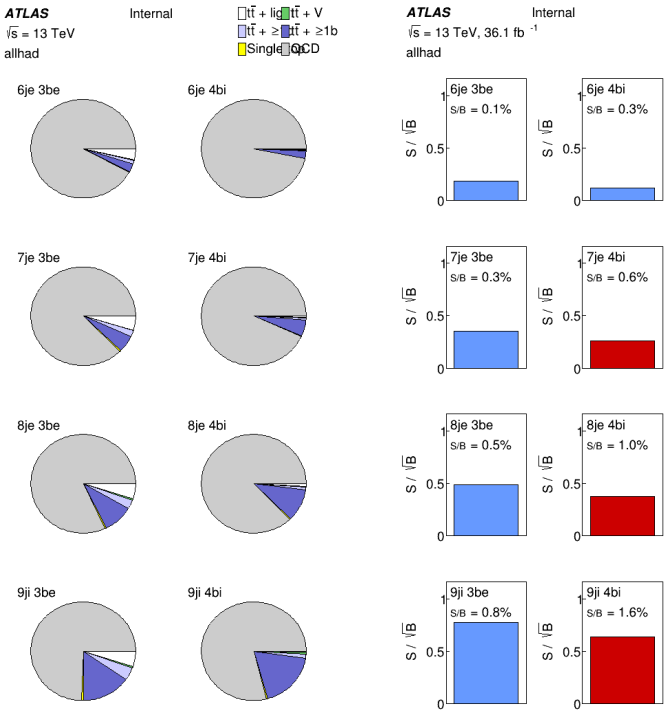
- Multijet trigger
  - $\geq$4 jets with $p_T > 100$ GeV (120 GeV for 2017 data) and $|\eta| < 2.5$
- Lepton veto
  - to avoid overlap with other channels
- Offline selection:
  - same as the trigger + $\geq$ 2 additional jets with $p_T > 25$ GeV and $|\eta| < 2.5$
  - $\geq$ 2 jets *b*-tagged
- Categorization in jet and *b*-jets multiplicity
  - 6, 7, 8 or $\geq$9 jets
  - 2, 3 or $\geq$4 *b*-jets



data 15+16	8j,3b	8j,$\geq$4b	$\geq$9j,3b	$\geq$9j,$\geq$4b
ttH	50.9 $\pm$ 0.3	13.9 $\pm$ 0.2	76.4 $\pm$ 0.4	24.5 $\pm$ 0.2
ttb	1033.7 $\pm$ 25.4	149.3 $\pm$ 9.6	1502.6 $\pm$ 30.8	281.9 $\pm$ 12.8
ttc	360.4 $\pm$ 15.8	10.1 $\pm$ 2.7	437.9 $\pm$ 17.5	17.6 $\pm$ 3.6
ttl	505.0 $\pm$ 19.0	8.0 $\pm$ 2.5	431.3 $\pm$ 17.6	4.3 $\pm$ 1.8
ttV	30.9 $\pm$ 1.0	5.2 $\pm$ 0.4	45.2 $\pm$ 1.5	10.2 $\pm$ 0.7
singletop	63.5 $\pm$ 4.2	6.5 $\pm$ 1.4	57.3 $\pm$ 3.9	6.3 $\pm$ 1.3
QCD	8946.1 $\pm$ 35.6	1203.9 $\pm$ 5.1	7238.9 $\pm$ 34.9	1196.9 $\pm$ 5.9
Total Background	10940 $\pm$ 50	1373 $\pm$ 12	9713 $\pm$ 53	1518 $\pm$ 15

Event Selection and Categorization

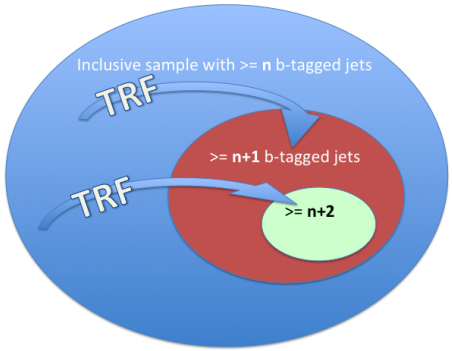
- Multijet trigger
  - $\geq$4 jets with $p_T > 100$ GeV (120 GeV for 2017 data) and $|\eta| < 2.5$
- Lepton veto
  - to avoid overlap with other channels
- Offline selection:
  - same as the trigger + $\geq$ 2 additional jets with $p_T > 25$ GeV and $|\eta| < 2.5$
  - $\geq$ 2 jets *b*-tagged
- Categorization in jet and *b*-jets multiplicity
  - 6, 7, 8 or $\geq$9 jets
  - 2, 3 or $\geq$4 *b*-jets

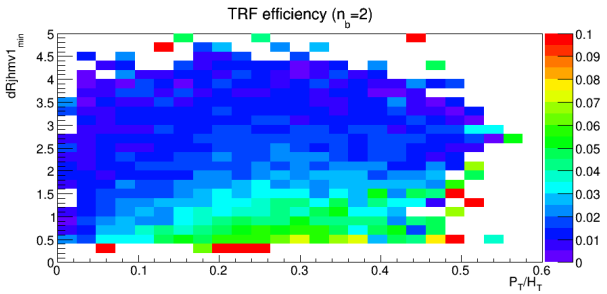


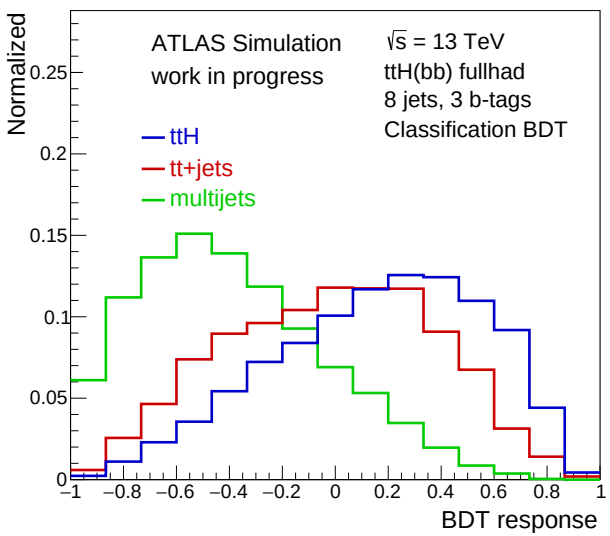
data 17	8j,3b	8j,$\geq$4b	$\geq$9j,3b	$\geq$9j,$\geq$4b
ttH	36.0 $\pm$ 0.4	9.4 $\pm$ 0.2	63.0 $\pm$ 0.5	19.8 $\pm$ 0.3
ttb	594.9 $\pm$ 22.1	75.8 $\pm$ 6.6	1023.2 $\pm$ 29.7	178.8 $\pm$ 11.6
ttc	197.2 $\pm$ 12.9	9.2 $\pm$ 4.1	321.1 $\pm$ 17.3	7.0 $\pm$ 2.0
ttl	278.1 $\pm$ 16.6	5.4 $\pm$ 2.7	278.1 $\pm$ 16.7	4.7 $\pm$ 1.9
ttV	20.6 $\pm$ 1.1	3.6 $\pm$ 0.5	32.6 $\pm$ 1.6	5.8 $\pm$ 0.7
singletop	46.2 $\pm$ 4.2	1.7 $\pm$ 0.6	45.7 $\pm$ 4.1	4.7 $\pm$ 1.0
QCD	5748.7 $\pm$ 28.5	763.6 $\pm$ 4.0	5279.2 $\pm$ 30.2	859.1 $\pm$ 5.1
Total Background	6886 $\pm$ 42	859 $\pm$ 9	6980 $\pm$ 49	1060 $\pm$ 13

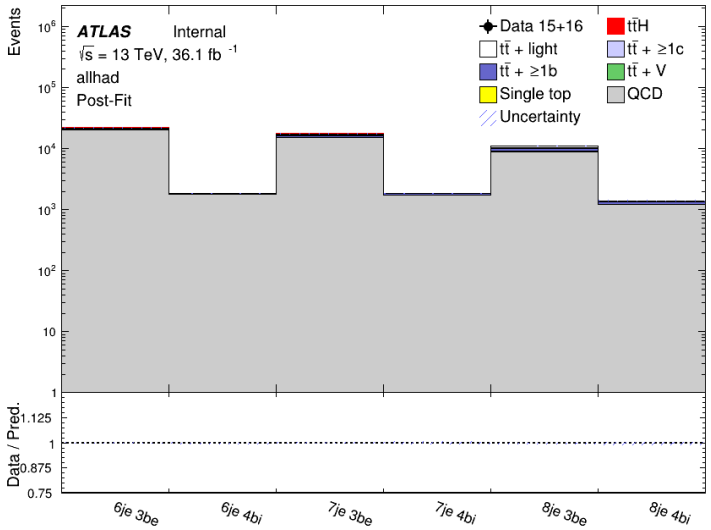
Analysis Strategy

- QCD multijet is main background, but really hard to reproduce in simulation
  - estimation with data-driven method: Tag Rate Function Multijet (TRF_{MJ})
- Signal vs background discrimination
  - with Multivariate Analysis (MVA) method called Boosted Decision Tree (BDT)
- Maximum likelihood fit
  - to extract the final results





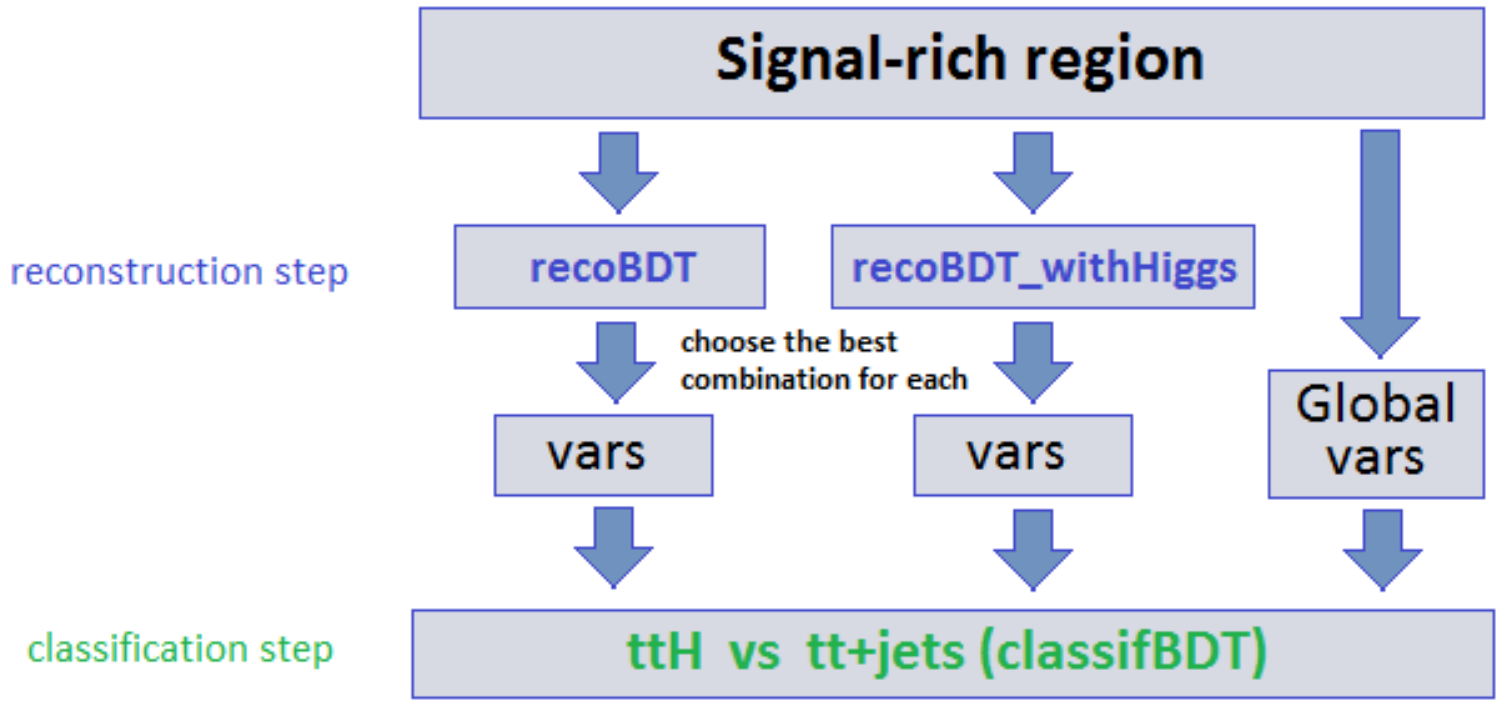




BDT studies

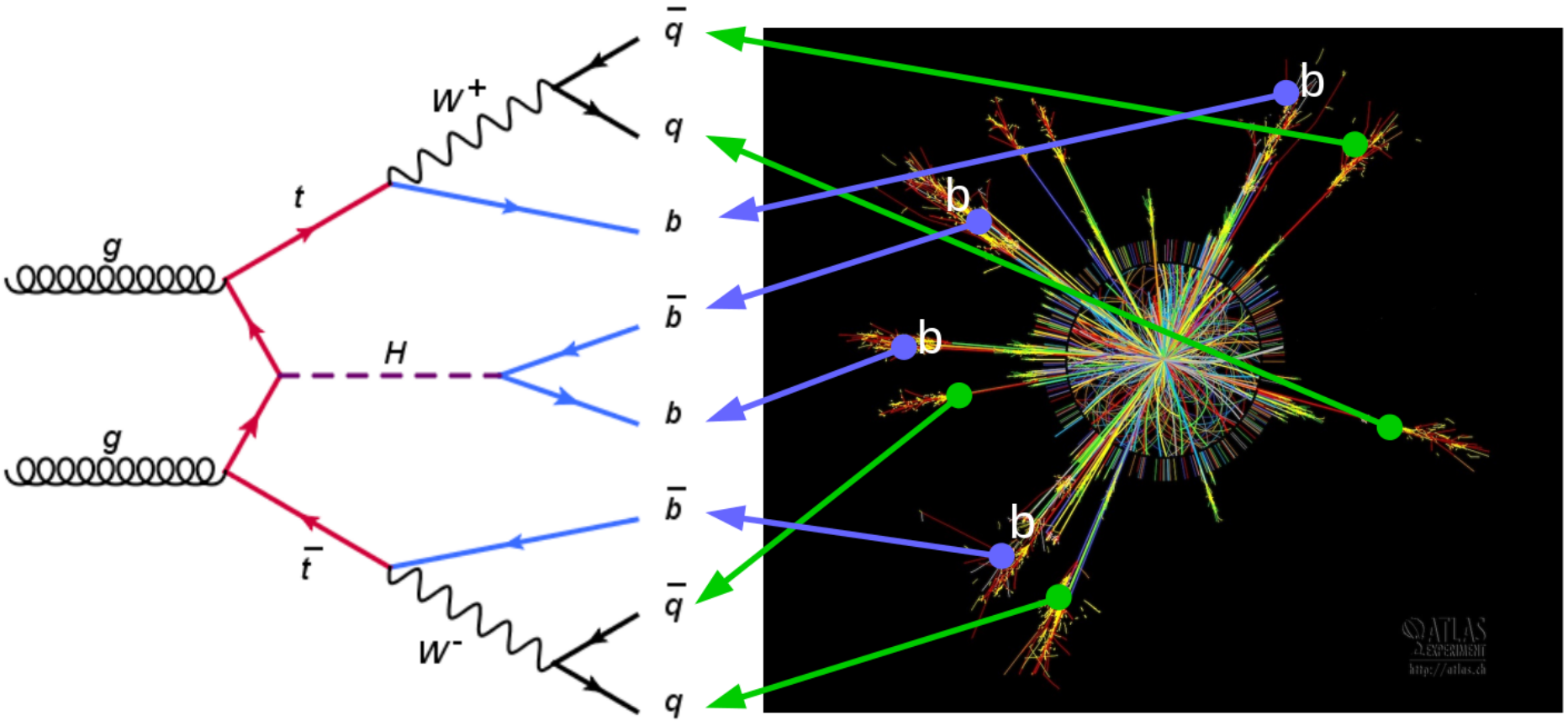
Multivariate Analysis for Fully Hadronic $t\bar{t}H$

Analysis chain



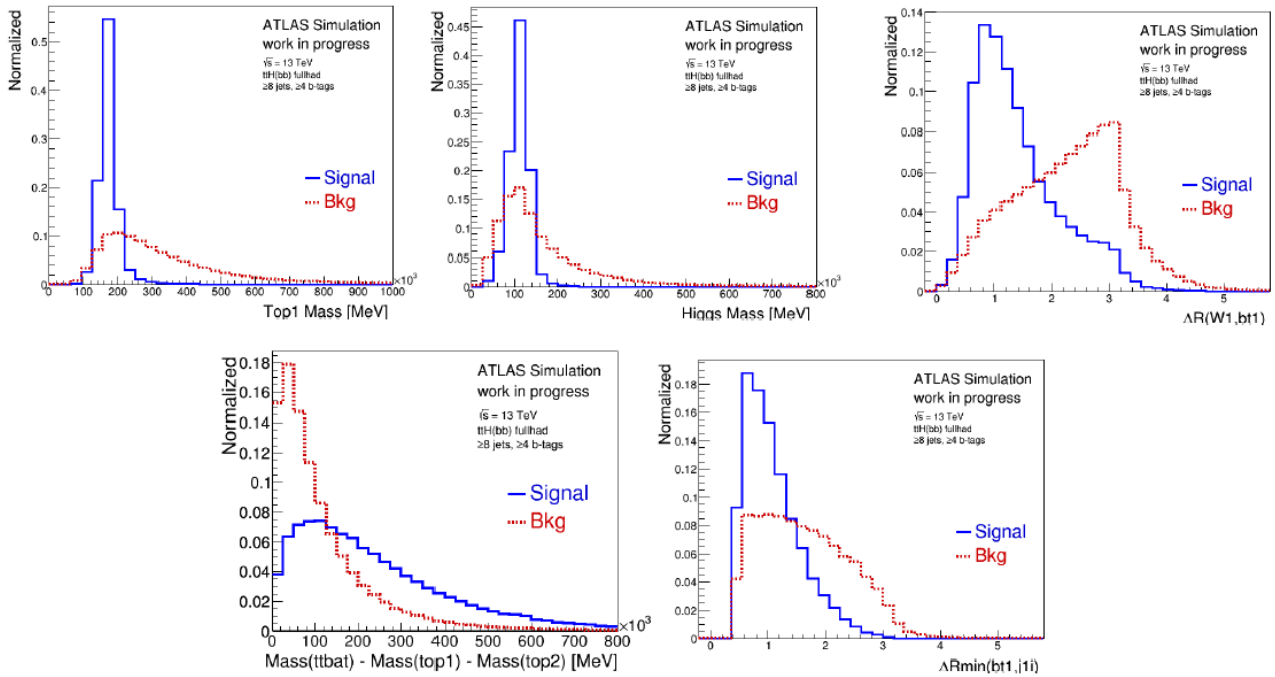
Reconstruction BDT Step

- **Goal**: find the best association between jets reconstructed in the detector and the final state partons
- Large mutliplicites $\rightarrow$ large combinatorics
  - from 36 up to thousands possible ways to reconstruct the $t\bar{t}H$ system

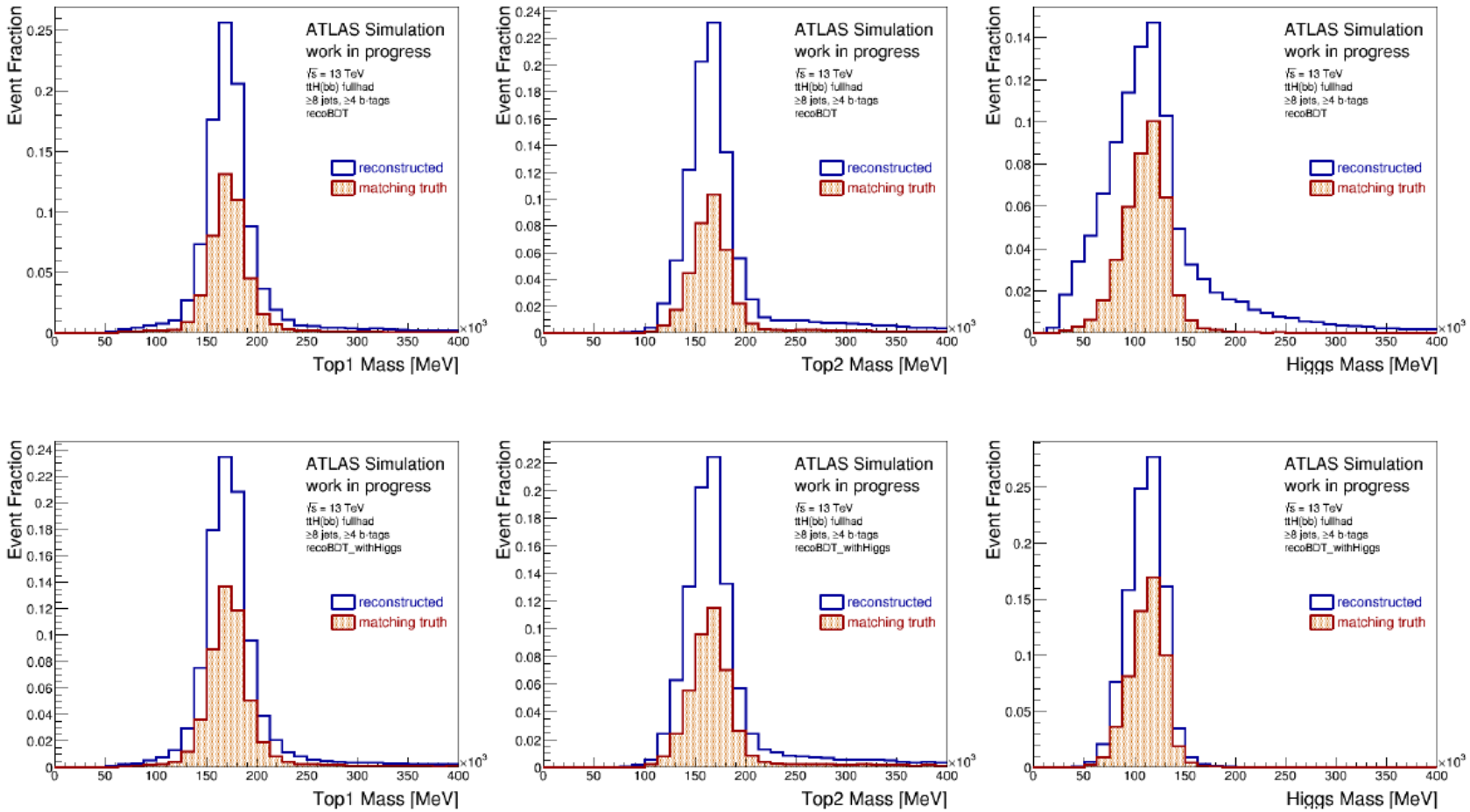


Reconstruction BDT Step

- 2 different BDTs using reconstructed resonances and angular correlations between jets
  - **recoBDT**: tries to reconstruct only $t\bar{t}$ system
    $\rightarrow$ no bias on the Higgs candidate mass
  - **recoBDT_withHiggs**: full $t\bar{t}H$ system reconstruction
    $\rightarrow$ higher reconstruction efficiency
- Trained using $t\bar{t}H$ simulation
  - signal: correct quarks-jets association
  - background: all other possible associations



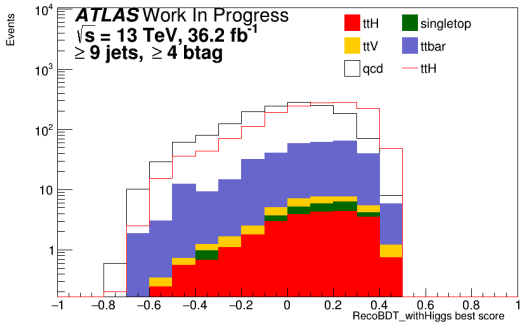
Reconstruction BDT Step

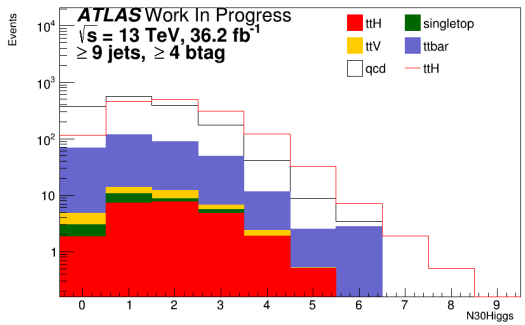


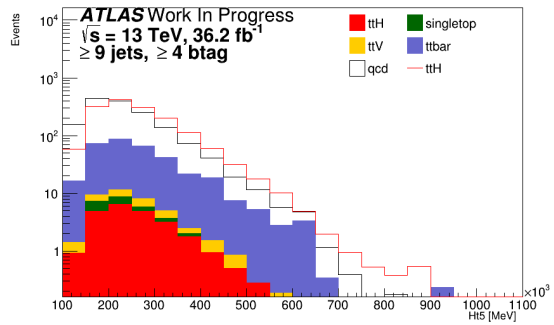
- Higgs boson candidate properly reconstructed 57% and 75% of times and full event properly assigned 41% and 53% respectively in events with full matching available

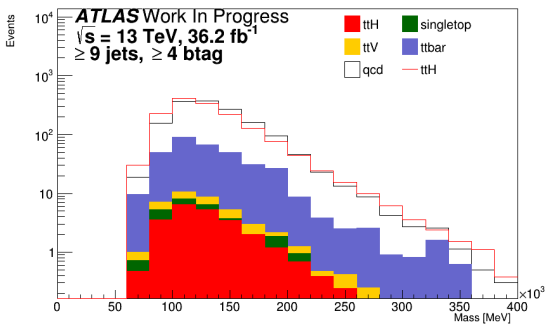
Classification BDT Step

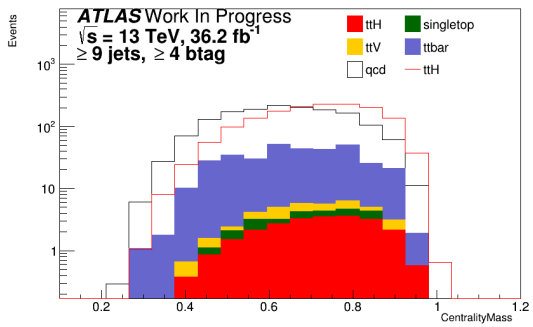
- **Goal**: perform signal vs background discrimination
- Combines reconstruction results from previuos step with global event kinematics
- Optimization against 2 different main background sources:
  - irreducible $t\bar{t} + b\bar{b} \rightarrow$ need more simulation
  - QCD multijet, using data-driven simulation from TRF_{MJ}

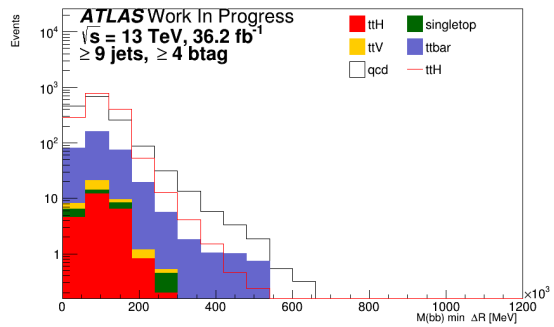






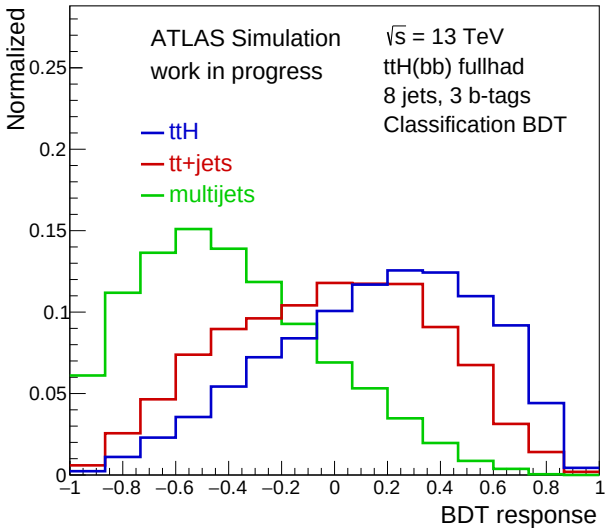


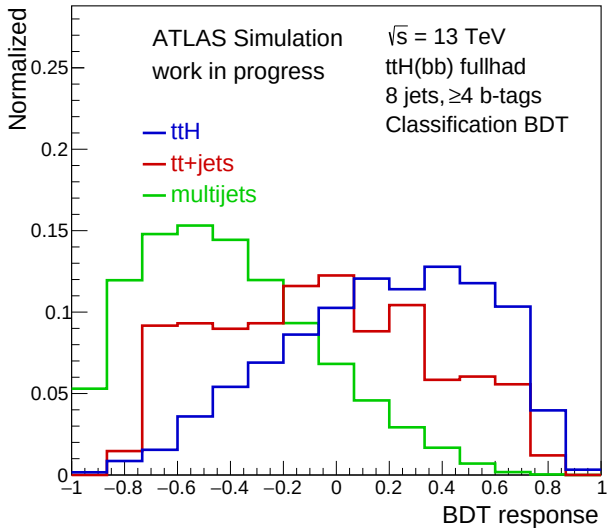




Classification BDT Step

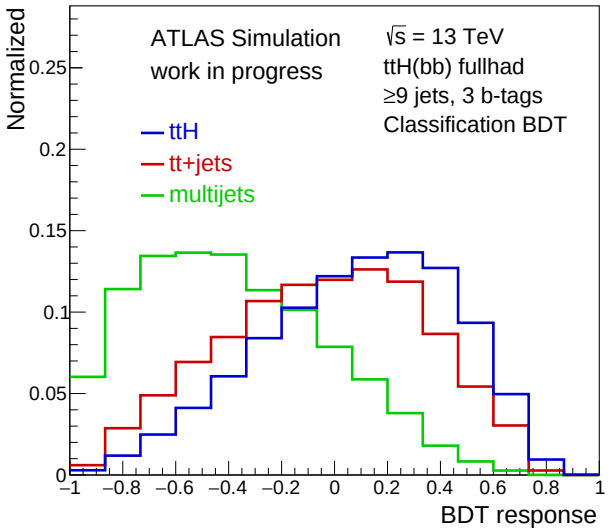
- Variables optimization performed separately in each signal region with a recursive method
  - From a preselected set of variables takes the one with the highest s/b separation
  - Adds recursively the variable that brings the biggest improvement
  - stops when the improvement from the addiction of a new variable is less than 1%

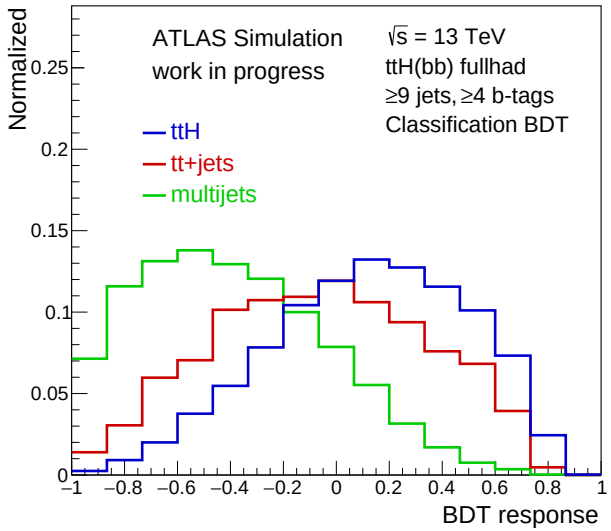




Classification BDT Step

- Variables optimization performed separately in each signal region with a recursive method
  - From a preselected set of variables takes the one with the highest s/b separation
  - Adds recursively the variable that brings the biggest improvement
  - stops when the improvement from the addiction of a new variable is less than 1%

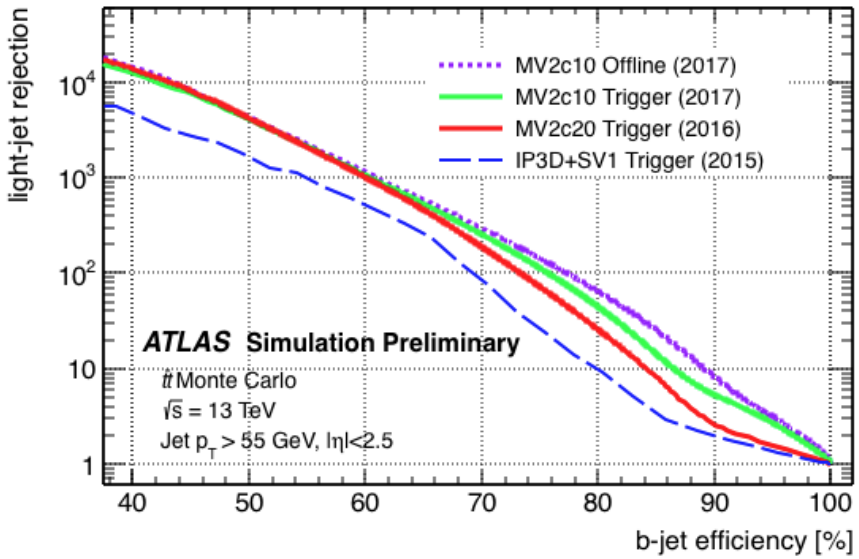




b-jet Trigger Calibration with tt Dilepton Events

b-jet Trigger Calibration

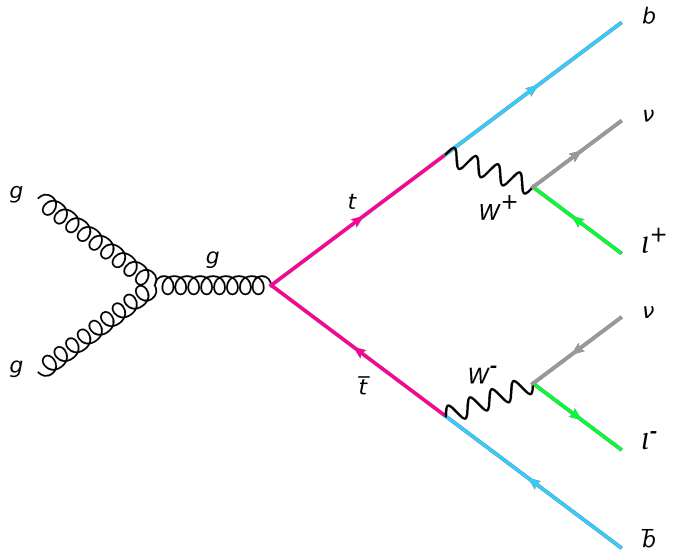
- trigger-level *b*-tagging improved with respect to Run 1, now close in performance to offline algorithms
- Use of *b*-jet trigger results on an increase of signal efficiency of a factor $\sim$3



- Data/MC Scale factors are used to calibrate online algorithms
  - SFs are derived for different jet flavors
- *b*-tagging used both at trigger level and at recostruction level
  - combined online + offline calibration
  - Geometrical association between online and offline jet objects

b-jet Trigger Calibration with $t\bar{t}$ Events

- Event selection:
  - activate one Lepton + *boffperf* trigger
  - exactly 2 leptons with $p_T > 28$ GeV and opposite charge
  - exactly 2 jets with $p_T > 35$ GeV
  - $e\mu$ channel
  - m_{lj} cuts
  - both jets matched



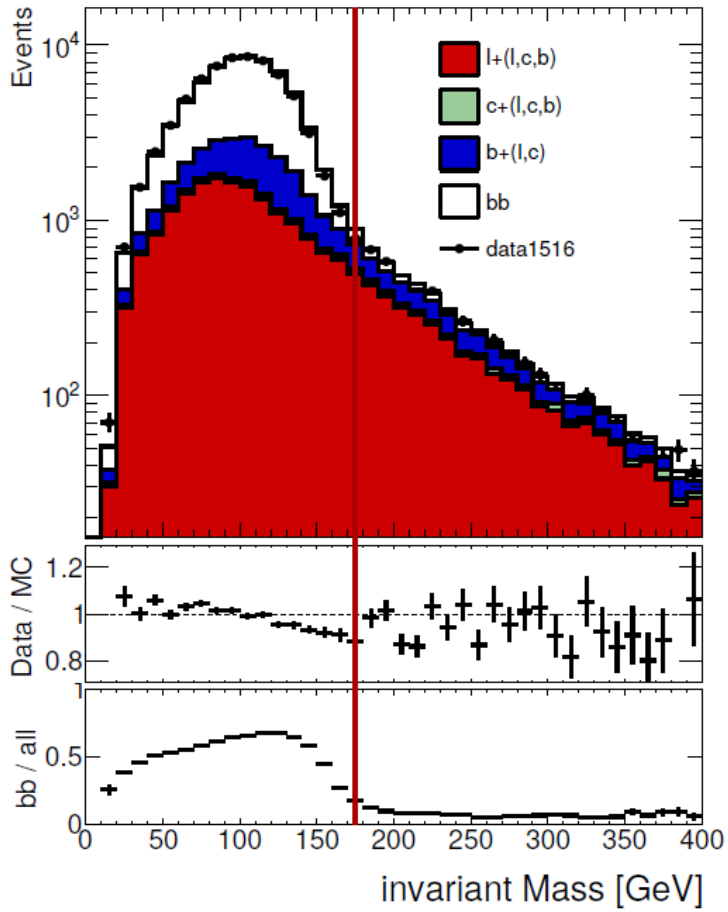
Lepton + boffperf prescaled triggers

b-tagging algorithm will run for each jet without taking decisions

m_{lj} cuts

- Idea: improve bb-purity by finding Jet + lepton combinations which corresponds to the top quarks.
- For *b*-Jets the invariant mass of the combination should be smaller then the top mass.
- The combination we found which seems to be the most promising in reducing background is the one which minimizes the sum of the squared invariant mass of both possible "lj-combinations" in the event.

lj-Combination, invariant Mass



- Veto events with one $m_{\mathrm{lj}} > 175\mathrm{GeV}$ ($\approx$ top-mass), or constrain flavor fractions

b-jet Trigger Calibration with $t\bar{t}$ Events

- b-jet content extracted using likelihood method

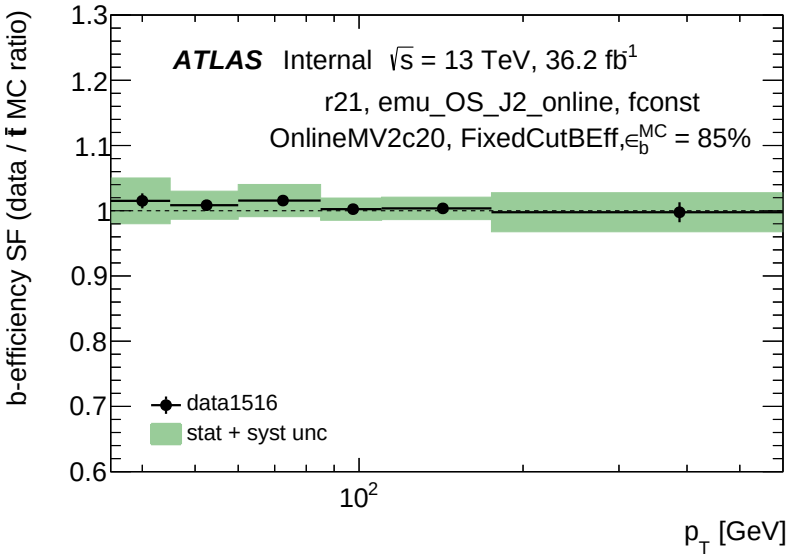
$$\begin{aligned}\mathcal{L}_{\mathrm{E}}(p_{\mathrm{T},1}, p_{\mathrm{T},2}, w_1, w_2|\mathcal{P}_{\mathrm{b}}(w|p_{\mathrm{T}})) =& [f_{\mathrm{bb}}(p_{\mathrm{T},1}, p_{\mathrm{T},2})\mathcal{P}_{\mathrm{b}}(w_1|p_{\mathrm{T},1})\mathcal{P}_{\mathrm{b}}(w_2|p_{\mathrm{T},2}) \\ &+ f_{\mathrm{bl}}(p_{\mathrm{T},1}, p_{\mathrm{T},2})\mathcal{P}_{\mathrm{b}}(w_1|p_{\mathrm{T},1})\mathcal{P}_{\mathrm{l}}(w_2|p_{\mathrm{T},2}) \\ &+ f_{\mathrm{lb}}(p_{\mathrm{T},1}, p_{\mathrm{T},2})\mathcal{P}_{\mathrm{l}}(w_1|p_{\mathrm{T},1})\mathcal{P}_{\mathrm{b}}(w_2|p_{\mathrm{T},2}) \\ &+ f_{\mathrm{ll}}(p_{\mathrm{T},1}, p_{\mathrm{T},2})\mathcal{P}_{\mathrm{l}}(w_1|p_{\mathrm{T},1})\mathcal{P}_{\mathrm{l}}(w_2|p_{\mathrm{T},2})]\end{aligned} \tag{1}$$

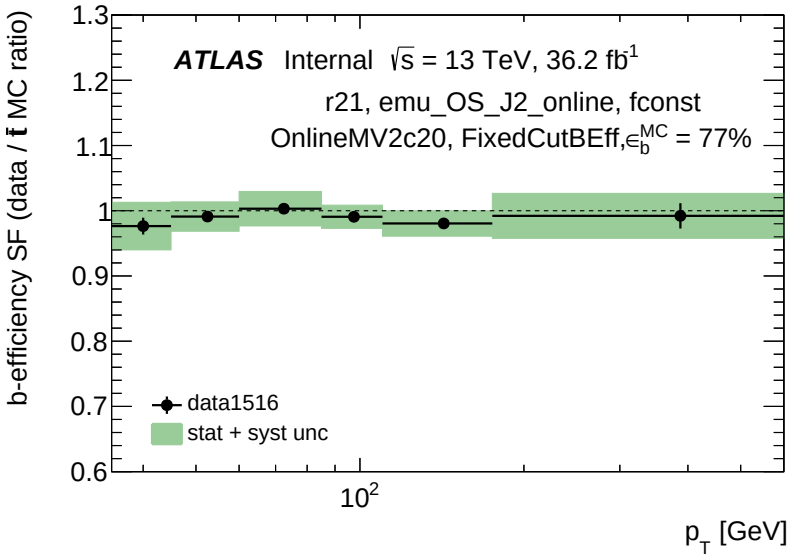
$f_{\mathrm{f}_1,\mathrm{f}_2}(p_{\mathrm{T},1}, p_{\mathrm{T},2})$ = fraction of flavour combination $[\mathrm{f}_1, \mathrm{f}_2]$. (Extracted from Simulation)
$\mathcal{P}_{\mathrm{f}}(w_1|p_{\mathrm{T},1})$ = pdf for a b-tagging weight w of jet with flavour f and a given $pT, 1$.

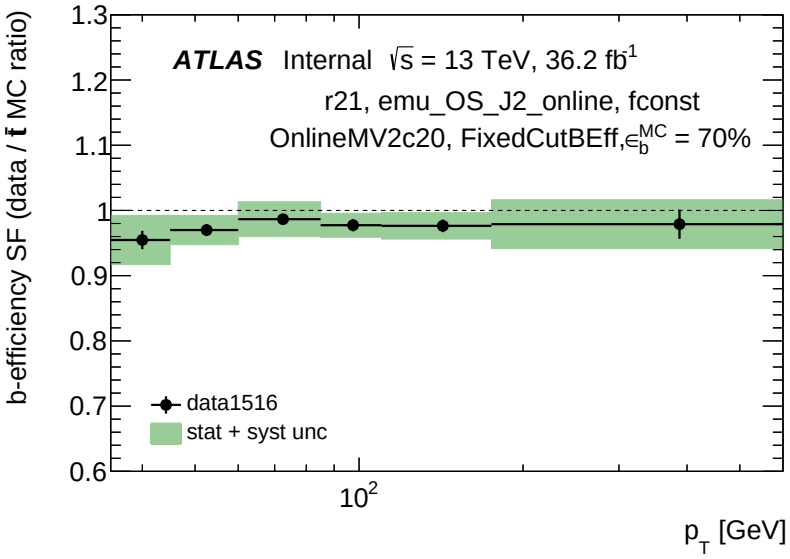
$$\mathcal{L}(\mathcal{P}_{\mathrm{b}}(w|p_{\mathrm{T}})) = \prod_{\mathrm{data}} \mathcal{L}_{\mathrm{E}}(\mathrm{data}|\mathcal{P}_{\mathrm{b}}(w|p_{\mathrm{T}})) \tag{2}$$

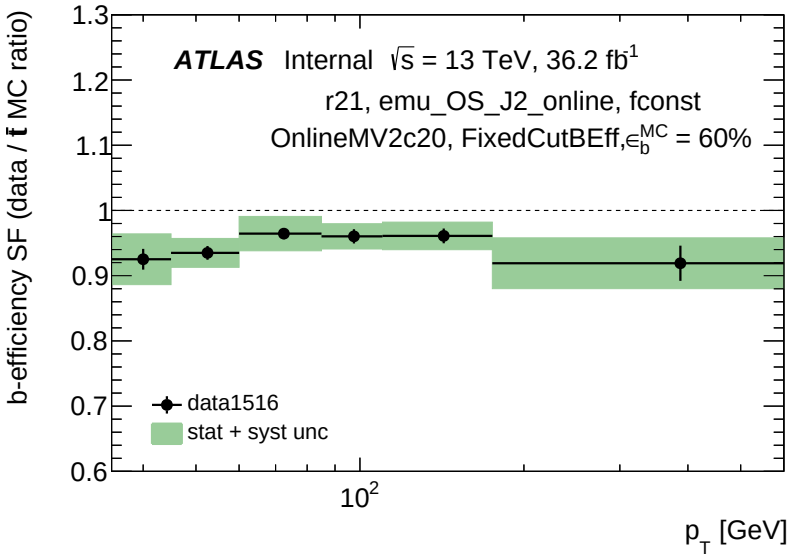
$$\epsilon_{\mathrm{b}}(p_{\mathrm{T}}) = \int_{w_{\mathrm{cut}}}^{\infty} dw' \mathcal{P}_{\mathrm{b}}(w_{\prime}|p_{\mathrm{T}}) \tag{3}$$

Calibration Results: Online only

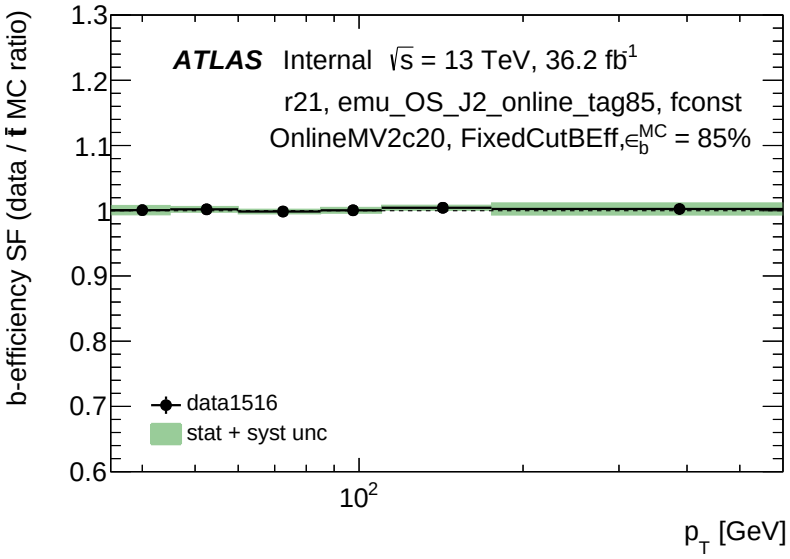


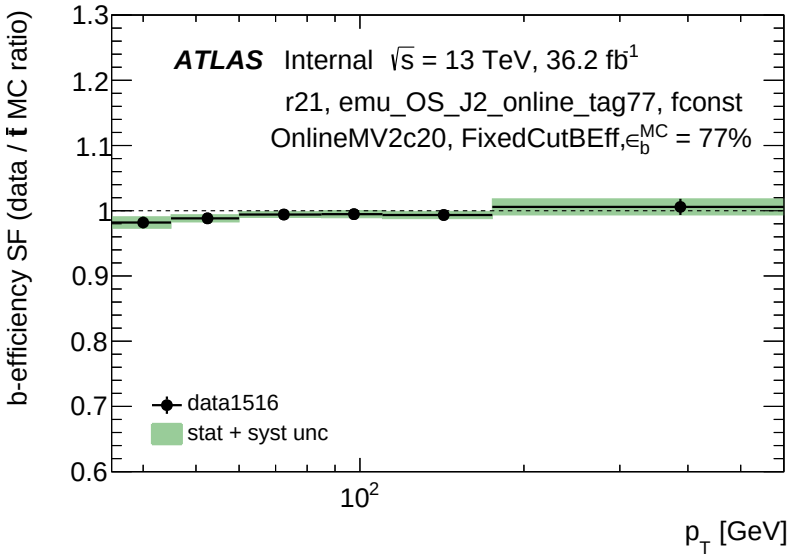


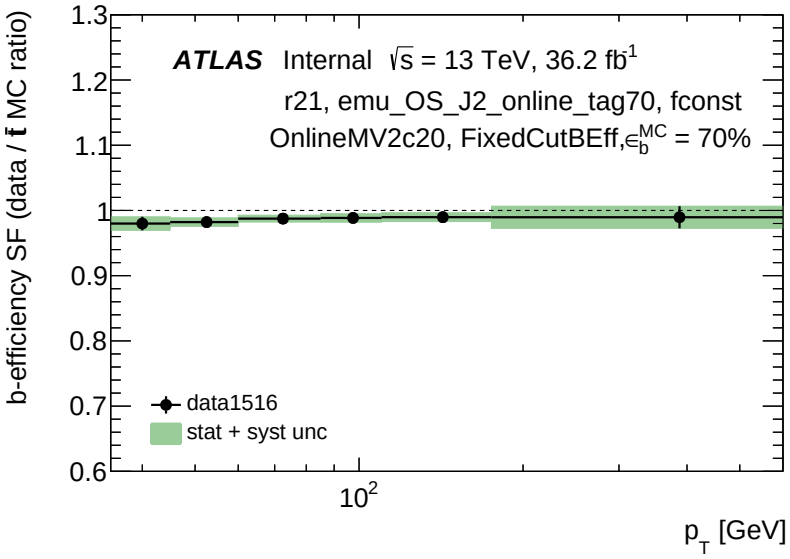


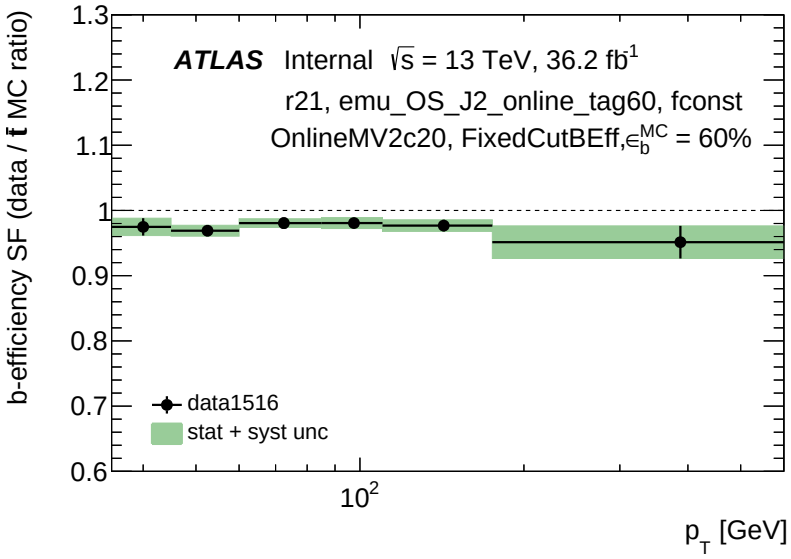


Calibration Results: Combined



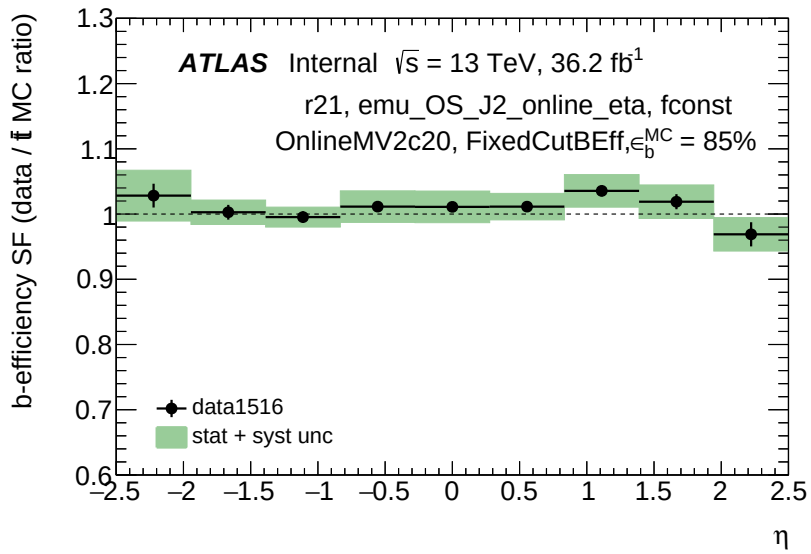


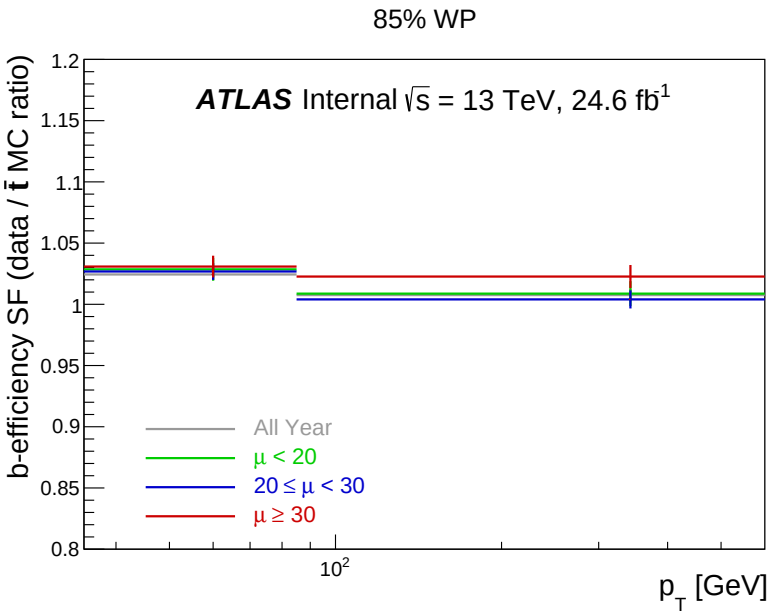


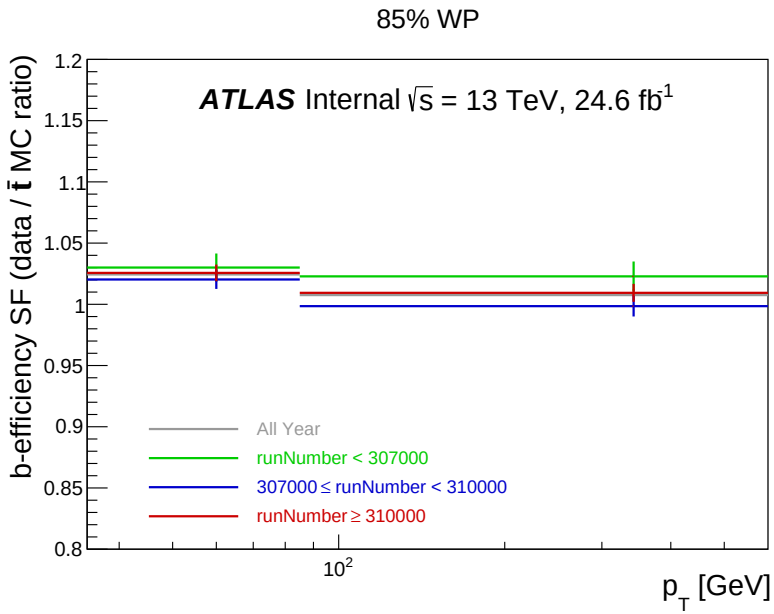


SF Stability Studies

- Studied SF dependency to different possible source of sistematic uncertainties:
  - η, pile-up and data period
- SF seems to be robust to all of them



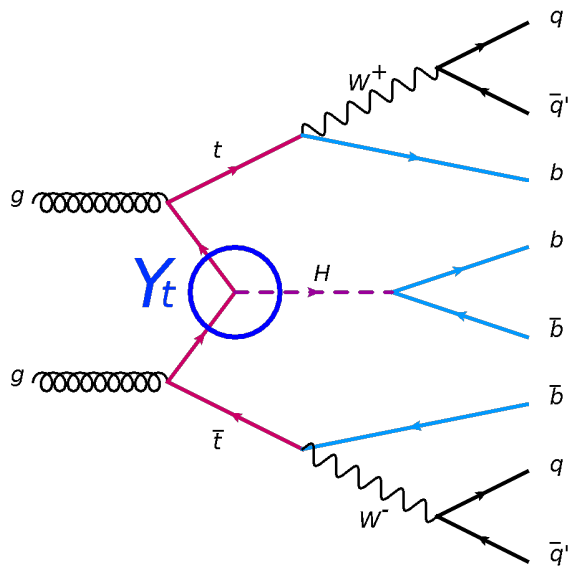




Conclusion

Conclusion

- Fully hadronic $t\bar{t}H$ channel:
  - Large BR, but dominated by QCD multi-jet production
  - Large statistic available and event fully reconstructable



- My contribution to improve the analysis:
  - Calibration of trigger b-tagging efficiency
    - ★ replace multijet trigger with b-jet trigger to increase by a factor $\sim$3 the signal selection efficiency
  - Implementing a 2 steps strategy for MVA based signal/background discrimination
    - ★ **Reconstruction step** to resolve combinatorics
    - ★ **Classification step** with optimization for QCD background discrimination

END

Thanks for the attention!