PROCEEDINGS OF SCIENCE

PoS(EPS-HEP2017)698

ATL-PHYS-PROC-2017-214
16 October 2017

BDTs in the Search for $t\bar{t}H$ Production with Higgs Decays to $b\bar{b}$ at ATLAS

Paul Glaysher* On behalf of the ATLAS Collaboration
DESY
E-mail: paul.glaysher@cern.ch

An observation of Higgs boson production in association with a top quark pair ($t\bar{t}H$) is a key outstanding measurement of the LHC programme. Boosted Decision Trees are vital for the search of $t\bar{t}H$, playing a role in event reconstruction and discriminating signal from a large background. The use of BDTs in the ATLAS search for $t\bar{t}H$, $H \to b\bar{b}$ is presented for the preliminary results on the $\sqrt{s} = 13$ TeV, 13.2 fb^{-1} dataset.

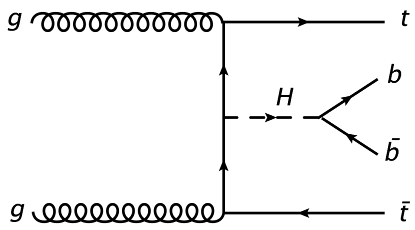
The European Physical Society Conference on High Energy Physics
5-12 July, 2017
Venice

*Speaker.

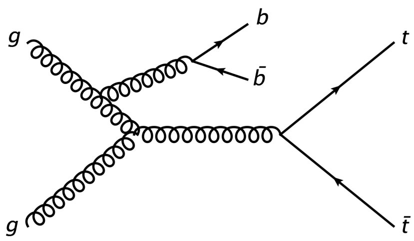
1. Introduction

The discovery of the Higgs boson in 2012 [1, 2] was a major breakthrough for the field of particle physics. The initial discovery made use of the dominant production mode of the Higgs boson at the LHC, via gluon-gluon fusion. Under study is a rarer production mode where the Higgs boson is produced in association with top quarks ($t\bar{t}H$). Measureing $t\bar{t}H$ production will provide a direct probe of the top Yukawa coupling, crucial to validating the Standard Model nature of the Higgs boson. However, the $t\bar{t}H$ cross section is predicted to be only 1% of total Higgs cross section [3]. This analysis at the ATLAS detector [4] therefore makes use of the largest Higgs branching ratio to b quarks of 58%.

The search for $t\bar{t}H$, $H \to b\bar{b}$, has two main challenges that are addressed by the use of multivariate techniques. Firstly, the $t\bar{t}H$ final state provides high combinatorics of b-jet pairs from which a Higgs boson candidate must be determined. This is aided by a Reconstruction Boosted Decision Tree (BDT). Secondly, the $t\bar{t}H$ signal is distinguished from the large $t\bar{t}$ + b-jets background, illustrated in Figure 1, by a Classification BDT.



(a)



(b)

Figure 1: Feynman diagrams of (a) the $t\bar{t}H$ signal and (b) the $t\bar{t}$ + b-jets background.

To reduce background, associated $t\bar{t}$ decays to one or two charged leptons in the final state are considered. Here only the one lepton case is presented from the preliminary 13 TeV ATLAS analysis [5], where lepton refers to either an electron or muon. The analysis is divided into regions based on number of jets and number of b-tagged jets to form signal and background enriched regions. The two stage multivariate approach is applied to the signal regions. Lastly, the ttH signal strength is determined from a profile likelihood fit over all signal and background regions.

2. Boosted Decision Trees

For both event reconstruction and classification, Boosted Decision Tree (BDT) algorithms are used. These represent shallow learners and are thus robust against over-training. A BDT is a decision tree ensemble built by incrementally targeting previously misclassified training instances. An implementation of Adaptive Boosting as part of the TMVA package [6] was used.

BDTs are trained on signal and background Monte Carlo simulated data. The hyper-parameters and choice of input variables are optimised, using the area under the ROC curve as a measure of performance. To mitigate over-training a two-fold cross validation is performed. Additionally, all

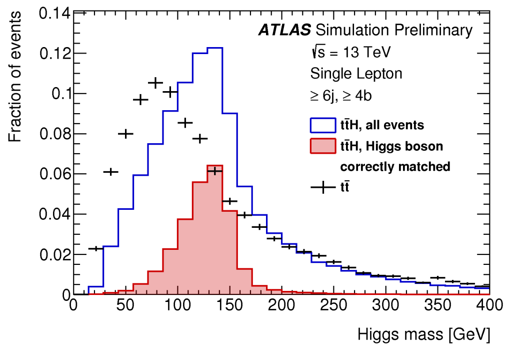


Figure 2: Normalised reconstructed invariant mass of Higgs candidate from ttH MC and fraction of events correctly matched, compared to the $t\bar{t}$ background [5].

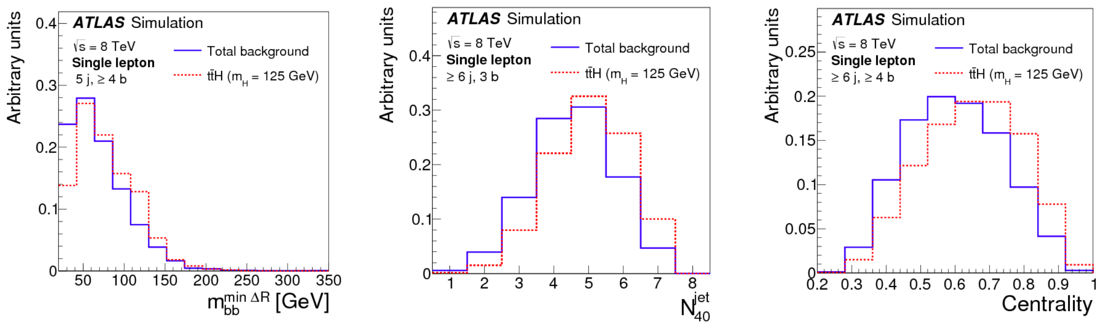
training variables must be well modelled in MC compared to data. Independent BDTs are trained and optimised for each signal region.

3. Reconstruction BDT

The Reconstruction BDT matches reconstructed jets to final-state partons from either Higgs or top decays in $t\bar{t}$H MC data, training correct against incorrect matching. The BDT score thus represent the probability of correct assignment of the b-jets from Higgs boson decay. The most probable Higgs boson candidate is therefore established despite a large combinatorial ambiguity and kinematic variables such as its mass can be computed, as in Figure 2.
The Reconstruction BDT is trained on up to 22 kinematic variables depending on the region. The jets from the Higgs boson are correctly matched in about 30% of the cases. The best possible efficiency is achieved by including information on the Higgs boson to the training, such as the Higgs mass. This however biases the identification of gluon→bb to be more Higgs-like. For variable reconstruction other than M_{bb} and $\Delta R(b,b)$ the prior Higgs information is used to enhance the matching efficiency. In this case the maximum efficiency to fully match all object in an event is 12%. The output of the Reconstruction BDT are used as training inputs for the Classification BDT.

4. Classification BDT

The Classification BDT is trained to classify events as more signal or background-like. Training inputs are variables from the Reconstruction BDT in addition to further quantities. Each individual input variable shows only small kinematic differences between signal and background, as shown in Figure 3, but combined in the BDT provide better separation of signal from the background. An example of a BDT discriminant distribution is given in Figure 4, showing signal and background vs. data. The overall signal to background (S/B) ratio is 5% in this region and it can be seen that with the BDT a sub-selection can be made with higher signal purity. The signal rich last bin has a S/B $\approx$ 20%. The distributions of the classification BDT scores serve as the final discriminant in the fit of MC to data.



(a) (b) (c)

Figure 3: Examples of training varibales for the Classification BDT [7], showing normalised distributions for signal and background of (a) the invariant mass of the two b-jets with the smallest opening angle, (b) the number of jets with a p_T of at least 40 GeV and (c) the sum of p_T divided by the total energy of all jets and the lepton.

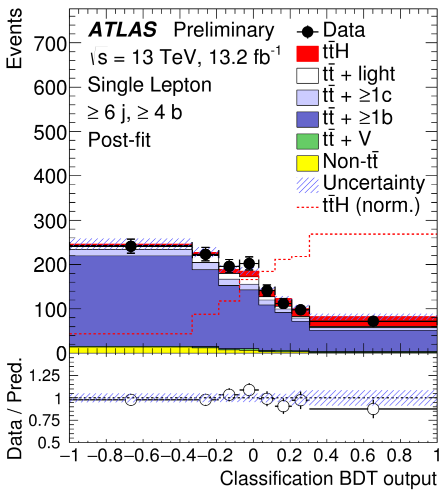


Figure 4: The Classification BDT score for the most sensitive region defined by ≥ 6 jets and ≥ 4 b-tags [5]. The data to MC agreement is shown after the MC is scaled by the fit.

5. Results and Conclusion

The use of multivariate techniques for event reconstruction and classification are crucial to the ATLAS $t\bar{t}$H, H $\rightarrow$ bb search. The shown two stage BDT represents the core of this analysis. Following this procedure, a $t\bar{t}$H, H $\rightarrow$ b$\bar{b}$, signal strength of $\mu_{ttH} = 1.6 \pm 1.1$ was determined with the 13 TeV 13.2 fb^{-1} dataset for the single lepton channel, where μ is the ratio of the measured cross section over the Standard Model prediction. This preliminary measurement is therefore both compatible with the SM expectation and the background-only hypothesis. A $\mu_{ttH} > 2.2$ was excluded at 95% CL. The sensitivity is limited by $t\bar{t}$ + b-jets modelling uncertainty.

References

[1] Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, ATLAS Collaboration, Phys. Lett. B 716 (2012) 1

[2] Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC, CMS Collaboration, Phys. Lett. B 716 (2012) 30

[3] Handbook of LHC Higgs Cross Sections: 3. Higgs Properties, S Heinemeyer et al., arXiv: 1307.1347 [hep-ph]

[4] ATLAS Collaboration, 2008 JINST 3 S08003

[5] Search for the Standard Model Higgs boson produced in association with top quarks and decaying into a bb pair in p p collisions at sqrts = 13 TeV with the ATLAS detector, ATLAS Collaboration, ATLAS-CONF-2016-080

[6] TMVA - Toolkit for Multivariate Data Analysis, A. Hoecker et al., PoS ACAT:040,2007

[7] Search for the Standard Model Higgs boson produced in association with top quarks and decaying into bb in pp collisions at sqrts = 8 TeV with the ATLAS detector, ATLAS Collaboration, Eur. Phys. J. C (2015) 75: 349

PoS(EPS-HEP2017)698