SEARCH FOR $t\bar{t}H$, $H \to b\bar{b}$ AT THE CMS EXPERIMENT IN 2016 USING 12.9 fb^{-1} OF pp COLLISION DATA*

Joosep Pata

Institute for Particle Physics and Astrophysics, ETH Zürich
8093 Zürich, Switzerland
pata@phys.ethz.ch

(Received April 24, 2018)

We present an overview of the search for $t\bar{t}H$ in the CMS experiment using up to 12.9 fb^{-1} of data collected during 2016. The analysis is carried out in the $H \to b\bar{b}$ final state with at least one of the top quarks decaying leptonically, resulting in a multi-parton final state with a combinatorial self-background. Discriminators based on machine learning and the direct computation of matrix elements from observed jet and lepton properties are used to distinguish between the $t\bar{t}H$ signal and the $t\bar{t}$ + jets background. Using a combined fit of the multivariate discriminants in several event categories, we find an observed (expected) upper limit of $\mu < 1.5$ (1.7) at the 95% confidence level. We further discuss how this analysis can be extended to the full Run 2 dataset.

DOI:10.5506/APhysPolBSupp.11.249

1. Introduction

After the discovery of the Higgs boson with $m_H = 125$ GeV at the LHC [1, 2], focus of the experimental work in the Higgs sector has shifted to determining the couplings between the Higgs boson and the known Standard Model (SM) fields. In Run 1 of the LHC, the coupling modifiers between the Higgs boson and gauge bosons have been determined relatively precisely, under the assumption of no beyond the Standard Model (BSM) physics [3]. In Run 2, we seek to determine the couplings to fermions, in particular to the top quark, which in the SM is predicted to be of the order of 1 in the Yukawa mechanism. Existing constraints on the coupling modifier for the top quark rely in large parts on the production of Higgs bosons in the gluon fusion (ggF) channel, which is a loop-induced process enhanced due

* Presented at the Final HiggsTools Meeting, Durham, UK, September 11–15, 2017.

to the large contribution from the top-quark loop. In the presence of BSM fields, there may be additional contributions in the loop which obscure the interpretation, therefore, a direct determination of the top-Higgs coupling is desirable.

The $t\bar{t}H$ process, with a cross section of $0.53^{+7.8\%}_{-5.5\%}$ pb [4], presents a channel where the top-Higgs coupling can be accessed at tree-level. The $H \to b\bar{b}$ decay channel with a branching ratio of $\simeq 58\%$ partially alleviates statistical concerns from the very low signal cross section. In this analysis, we focus on the cases where at least one of the top quarks decays leptonically, thus allowing the efficient use of lepton triggers in the experiment. The dominant background in this channel is the QCD production of $t\bar{t}+$jets, which has an inclusive cross section of $832^{+2.3\%}_{-3.5\%}$ (scale)$^{+4.2\%}_{-4.2\%}$ (PDF+α_s)$^{+2.7\%}_{-2.7\%}$ (m_t) pb [5], several orders of magnitude higher than the signal process. In addition, the production of top-quark pairs in association with a bottom-quark pair in the $t\bar{t}+b\bar{b}$ process results in an irreducible background, with 4 bottom quarks in the final state for both the signal and background processes in the leading order (LO) description, in addition to light quarks, charged leptons and neutrinos.

Furthermore, the Higgs invariant mass peak cannot be reconstructed, since the expected width of the peak is several orders of magnitude below the detector resolution and there is a combinatorial self-background arising from multiple bottom quarks in the final state. Therefore, we need to resort to multivariate statistical methods, in particular the use of machine learning techniques, which are complemented with an *ab initio* method of evaluating per-event likelihoods based on the observable event quantities and the underlying matrix elements for the hard interactions in the matrix element method (MEM).

In these proceedings, we present a search for $t\bar{t}H$, $H \to b\bar{b}$ using up to 12.9 fb^{-1} of proton–proton data, depending on the decay channel, collected by the CMS experiment during 2016 [6]. We will briefly describe the analysis method in Section 2, followed by the results in Section 3. We discuss the extension of this analysis to the full 2016 dataset, with a specific attention on the matrix element method, in Section 4.

2. Analysis

In the $t\bar{t}H$, $H \to b\bar{b}$ search, we rely on the identification of charged leptons from the top-quark decay for triggering and reducing the multijet QCD to negligible levels. We expect one (two) electrons or muons in the semileptonic (dileptonic) category with sufficient transverse momentum, with neutrinos giving rise to missing transverse energy (MET) in the detector. In the semileptonic (dileptonic) channel, we expect 6 (4) jets at the LO

description from $t\bar{t}H$, $H \to b\bar{b}$, 4 out of which arise from the hadronisation of bottom quarks. Therefore, we require that at least 3 jets must pass a b discriminator threshold (b tag) that has an acceptance of $\simeq 70\%$ for b jets and a fake rate of $\simeq 1\%$ for light jets. We use the combined secondary vertex (CSVv2) b discriminator algorithm [7]. In order to reduce the contributions from interactions within the same bunch crossing (pileup), we require jets, clustered using the anti-k_t algorithm with $R = 0.4$, to pass a transverse momentum cut of $p_T > 30$ GeV in the semileptonic channel, with the threshold for the subleading jets in the dileptonic channel reduced to $p_T > 20$ GeV. Jets are required to be within tracker acceptance by $|\eta| < 2.4$.

After identifying the jets and leptons, we group the events into mutually exclusive categories based on jet and b tag multiplicities, such that the categories are composed of different fractions of the signal and background processes. The signal process is predicted using MC simulation implemented in `POWHEG` [8] at next-to-leading-order (NLO). The primary background arises from $t\bar{t}+$jets, similarly predicted using `POWHEG`. The production of $t\bar{t}+b\bar{b}$ is only described at LO or parton shower (PS) accuracy, therefore, we account for possible theoretical uncertainties by assigning uncorrelated normalisation uncertainties to the sub-processes of $t\bar{t}+$jets.

The analysis relies on multivariate methods for distinguishing between the signal and background processes. In particular, we use boosted decision trees (BDT) optimised for $t\bar{t}H$ *vs.* $t\bar{t}+$jets (inclusive) discrimination and the MEM, distinguishing between $t\bar{t}H$ *vs.* the irreducible $t\bar{t}+b\bar{b}$ background. As we cannot identify a pure $t\bar{t}+b\bar{b}$ control region, we extract the signal and background processes in a combined fit of the discriminators across all event categories, such that each of the final categories is split into a signal-enriched and background-enriched part using the BDT discriminator, with the MEM being used as a discriminator in these final categories.

2.1. Matrix element method

The matrix element method has been proposed for analyses involving irreducible backgrounds, in particular $t\bar{t}H$, $H \to b\bar{b}$ [9] and used successfully by the LHC experiments already in Run 1 [10]. This method does not require extensive MC simulation and furthermore provides an efficient discriminator in the presence of combinatorial backgrounds. Briefly, the MEM relies on the numerical evaluation of event weights for the signal and background processes of the form of

$$w_{\text{sig,bkg}}(y) \propto \int |\mathcal{M}_{\text{sig,bkg}}(x)|^2 W(x|y)\ \mathrm{d}x \tag{1}$$

with $|\mathcal{M}_{\text{sig,bkg}}(x)|^2$ being the scattering amplitude for the signal (background) hypothesis and $W(x|y)$ the detector transfer function that sum-

marises the evolution of parton-level quantities x to detector-level quantities y, integrated over the full phase space. An optimal discriminator between the signal and background processes can be then constructed as $P_{\mathrm{s/b}} = w_{\mathrm{sig}}/(w_{\mathrm{sig}} + \alpha w_{\mathrm{bkg}})$, with $\alpha \simeq 0.1$ being a normalisation factor[1]. We show the expected performance of the MEM in figure 1, where we see that we can reject the $t\bar{t}+\mathrm{jets}$ background at a level of $\simeq 80\%$ at a signal efficiency of 50%. This is comparable with state-of-the art machine learning methods.

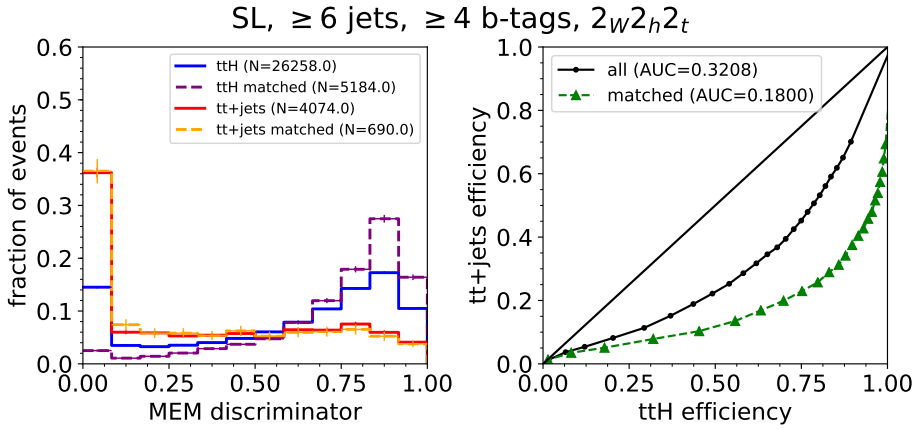

Fig. 1. (Colour on-line) The expected performance of the MEM discriminator in the semileptonic category with at least 6 jets, at least 4 of which must be b tagged. We show the distributions for $t\bar{t}H$ and $t\bar{t}+\mathrm{jets}$ (left) and the $t\bar{t}H$ *vs.* $t\bar{t}+\mathrm{jets}$ efficiency (right), where we compare the expected performance based on all detector-level jets (black) to the performance on jets that were geometrically matched to partons from the hard interaction (grey/green). The latter represents a theoretical upper bound on the performance of the MEM discriminator under ideal detector efficiency. We compare the overall performance in terms of the receiver operating characteristic (ROC) area under curve (AUC), with smaller values corresponding to better signal-to-background discrimination.

We have developed the MEM as applied to $t\bar{t}H$, $H \to b\bar{b}$ further in Run 2 by extending the phase space accessible to the MEM by integrating over unobserved or poorly measured quantities, carried out extensive validation using alternative MC simulation and optimised the method for large-scale use on LHC computing resources. The MEM is evaluated on events that contain at least 4 jets, requiring between 1–10 minutes of integration time per

[1] The exact choice of α only affects the signal-to-background discrimination in the case the distribution is discretised as a histogram, with the choice of $\alpha = 0.1$ corresponding to optimal separation.

event. The computational demand arises from the need to consider between $\mathcal{O}(10)$–$\mathcal{O}(100)$ permutations for the jet-to-parton assignment as well as the difficulty of integrating over jet properties in partially-reconstructed events. Nevertheless, it has been demonstrated to be feasible and to significantly improve the analysis sensitivity in Run 2.

3. Results

We extract the signal strength modifier $\mu = \sigma/\sigma_{\mathrm{SM}}$ from a binned maximum likelihood fit of the MEM discriminator distribution over all the categories. We show an example of the post-fit distributions in figure 2, with the full set of distributions being available in [6]. The best-fit value of μ is $\mu = -0.19^{+0.45}_{-0.44}\,(\mathrm{stat.})^{+0.66}_{-0.68}\,(\mathrm{syst.})$ with a total uncertainty of $^{+0.80}_{-0.81}$. It is compatible within 1.5σ with the SM expectation of $\mu = 1$. This allows us to set observed (expected) exclusion limits on $\mu < 1.5$ (1.7) at a 95% confidence level. The result is dominated by systematic uncertainties, out of which the uncertainty on the modelling of $t\bar{t} + b\bar{b}$ is the most significant. The best fit value and the limits are shown in figure 3.

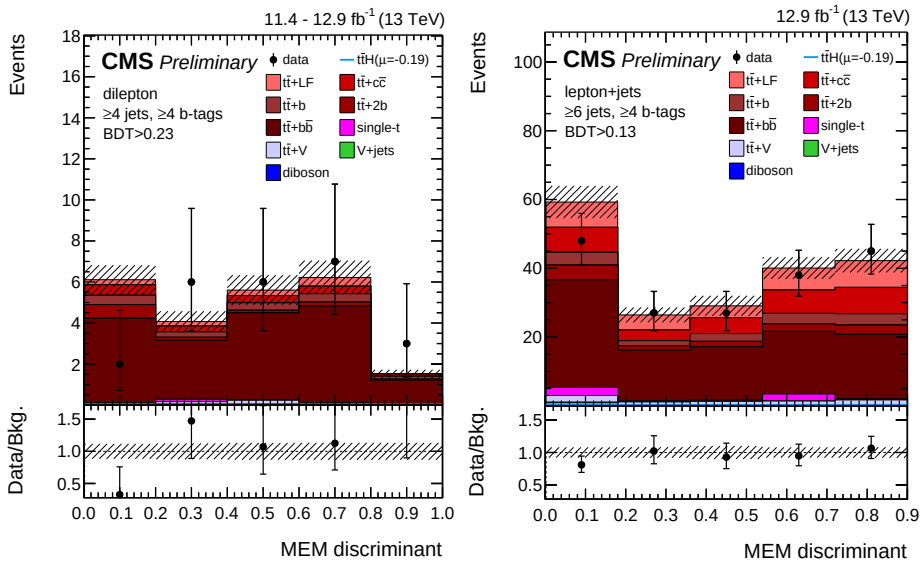

Fig. 2. The post-fit distributions of the MEM discriminant in the dileptonic (left) and semileptonic (right) signal-enriched categories.

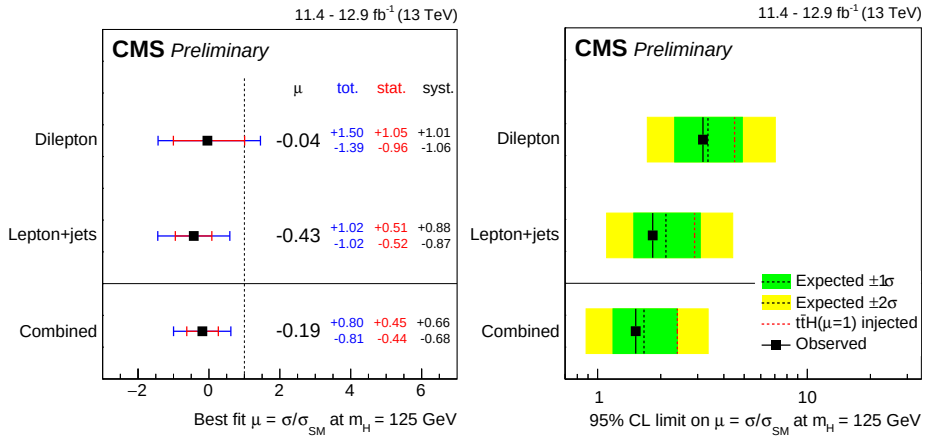

Fig. 3. The best fit value (left) and the upper limits at a 95% confidence level (right) of the signal strength $\mu = \sigma/\sigma_{\mathrm{SM}}$ in the individual dileptonic and semileptonic categories and in the combined fit.

4. Discussion and future work

Improvements to the analysis sensitivity are expected to arise from a better understanding of the systematic uncertainties, as well as including the full 2016 dataset corresponding to an integrated luminosity of 35.9 fb^{-1}. An improved theoretical understanding of the QCD production of $t\bar{t}+b\bar{b}$ is crucial for future precision measurements of the $t\bar{t}H$, $H \to b\bar{b}$ process. Out of the experimental uncertainties, we have focused on an improved description of the uncertainties related to jet energy corrections (JEC), as the presence of multiple jets in the final state makes the analysis relatively sensitive to any variations in jet energies. The $t\bar{t}H$, $H \to b\bar{b}$ search is a contribution to the full set of $t\bar{t}H$ searches in Run 2, which can be combined to determine the cross section of this process. Very recently, as also reported within these proceedings, the ATLAS Collaboration has published a combined $t\bar{t}H$ analysis in the $H \to b\bar{b}$, multilepton and $\gamma\gamma$ channels [11, 12], reporting evidence for this process at an observed (expected) significance level of 4.2σ (3.8σ).

4.1. Uncertainties in the matrix element method

A particular challenge that arises in the use of the MEM in the analysis is the sensitivity estimation of the MEM discriminant to variations in the observed jet momenta. This requires the MEM phase-space integral (Eq. (1)) to be carried out $\mathcal{O}(10^2)$ times per event under very similar conditions in order to account for various sources of uncertainty in the jet energy corrections. To make this computationally feasible, we have introduced an

approximate method based on the functional form of the integral, where detector effects are summarised in the transfer functions. We see that the weight $w_{\mathrm{sig,bkg}}(y)$ depends on the observable quantities y primarily through the transfer functions $W(x|y)$, with changes to the integration ranges having only a secondary effect. Therefore, by promoting the integrand to a vector, such that

$$|\mathcal{M}(x)|^2 W(x|y) \Rightarrow |\mathcal{M}(x)|^2 \begin{pmatrix} W(x|y) \\ W(x|y+\delta y_1) \\ \cdots \\ W(x|y+\delta y_N) \end{pmatrix}, \tag{2}$$

we have been able to carry out the full sensitivity analysis under variations $\delta y_n, n = 1 \ldots N$ using a single numerical integration of a vector-valued quantity. We verify the validity of this method by comparing it to the full variations in figure 4. Using this approximation, we have been able to use the MEM discriminant in the analysis of the full Run 2 dataset without significantly exceeding the computational capabilities on the LHC grid.

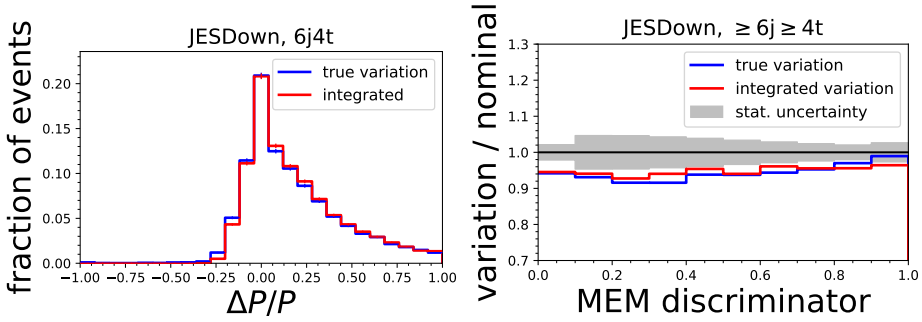

Fig. 4. (Colour on-line) We verify the sensitivity of the MEM under the approximate jet energy variations (grey/red) as compared to the true variation (black/blue) when the integral is recomputed. On the left, we show the relative change in the weight for the signal hypothesis for events with exactly 6 jets, out of which 4 are b tagged. On the right, we show the relative change in the signal-to-background weight ratio used as the final discriminator. We see that the approximate variations reproduce the true change to an acceptable degree. These distributions are derived using $t\bar{t}H$ MC simulation.

5. Summary

We have presented the search for $t\bar{t}H$, $H \to b\bar{b}$ in the final states where at least one of the top quarks decays leptonically using a part of the data collected in 2016. The observed (expected) upper limit on the signal strength

value μ is found to be $\mu < 1.5$ (1.7) at a 95% confidence level. Various multivariate techniques are in use to distinguish between the signal and the irreducible $t\bar{t} + b\bar{b}$ background. We have seen, in particular, that the use of the matrix element method provides significant sensitivity to the analysis. In Run 2, we have extended the applicability of the MEM considerably as applied to $t\bar{t}H$, such that it can be evaluated in final states that are not fully reconstructed. We have carried out extensive validation studies of the method and demonstrated that it is possible to estimate the effect of systematic uncertainties on the MEM through an approximate technique relying on vector integration. We are in the process of searching for the $t\bar{t}H$, $H \to b\bar{b}$ process in the full 2016 dataset and improving the treatment of the experimental systematic uncertainties in this analysis.

This research was supported in part by the Research Executive Agency (REA) of the European Union under the Grant Agreement PITN-GA2012-316704 ("HiggsTools"). The author would like to thank Günther Dissertori and Maren Meinhard for comments on these proceedings and the organisers of the workshop in Durham for an excellent environment for discussion.

REFERENCES

[1] S. Chatrchyan *et al.*, *Phys. Lett. B* **716**, 30 (2012).

[2] G. Aad *et al.*, *Phys. Lett. B* **716**, 1 (2012).

[3] G. Aad *et al.*, *J. High Energy Phys.* **1608**, 045 (2016).

[4] D. de Florian *et al.*, `arXiv:1610.07922 [hep-ph]`.

[5] M. Czakon, A. Mitov, *Comput. Phys. Commun.* **185**, 2930 (2014).

[6] CMS Collaboration, Technical Report CMS-PAS-HIG-16-038, CERN, Geneva, 2016.

[7] CMS Collaboration, Technical Report CMS-PAS-BTV-15-001, CERN, Geneva, 2016.

[8] S. Frixione, P. Nason, C. Oleari, *J. High Energy Phys.* **0711**, 070 (2007).

[9] P. Artoisenet, P. de Aquino, F. Maltoni, O. Mattelaer, *Phys. Rev. Lett.* **111**, 091802 (2013).

[10] V. Khachatryan *et al.*, *Eur. Phys. J. C* **75**, 251 (2015).

[11] ATLAS Collaboration, Technical Report ATLAS-CONF-2017-076, CERN, Geneva, 2017.

[12] ATLAS Collaboration, Technical Report ATLAS-CONF-2017-077, CERN, Geneva, 2017.