Summer Program

31st August 2019

Towards defining a new $t\bar{t}H$ category in the had-had $H \to \tau\tau$ analysis

Ewoud Ketele
Supervisors: Ximo Poveda, Tamara Vazquez Schroeder

This summer project was dedicated to develop a new analysis category to explore $t\bar{t}H$ with the Higgs boson decaying as $H \to \tau\tau$. In this new category we use a fully hadronic $t\bar{t}$ decay, with a final state of $0\ell + 2\tau_{\text{had}}$. Optimization of the channel is done considering the full Run-2 dataset of 140fb^{-1}.
Fitting the MVA scores while cutting on the missing mass estimator (MMC) and adding a theoretical uncertainty of 50% to the normalization of Z, $t\bar{t}$ real, and $t\bar{t}$ fakes, gives a signal significance of 1.48σ, trumping the signal significance of the current existing categories.

1 Introduction

This summer project was dedicated to develop a new analysis category to explore $t\bar{t}H$ with the Higgs boson decaying as $H \rightarrow \tau\tau$. None of the current multilepton analysis categories are relying on fully hadronic $t\bar{t}$ decays. Because there is a large correlation between $H \rightarrow WW*$ and $H \rightarrow \tau\tau$ in the multilepton analysis, a pure $H \rightarrow \tau\tau$ category would be really useful. The final state for this new category would be $0\ell + 2\tau_{\mathrm{had}}$. In this new channel, we expect six jets of which two are b-jets. The reconstruction efficiency of b-jets will therefore play an important role in the analysis. The optimization of this channel is done considering the full Run-2 dataset of 140 fb^{-1}.

2 Technical

During this project, we will make use of the $t\bar{t}H$-multilepton ntuple making code. It is modified to save $0\ell + 2\tau_{\mathrm{had}}$ events. We will also work with the same derivations as the $H \rightarrow \tau\tau$ analysis. On the technical side we will be using the multivariate analysis from the TMVA package, in particular the gradient Boosted Decision Tree [1], and for fitting we will be using the TRExFitter package [2]

3 Preselection

3.1 Jet multiplicity

As mentioned before, the topology of our signal events is one with six jets. Sometimes one of the six jets will not be reconstructed properly and we will only find five jets. This is clearly visible in Figure1 where the jet multiplicity of signal and background events is shown. Events with four jets or lower are heavily dominated by the background events. From five jet onwards the signal contribution is no longer negligible.

3.2 B-jet tagging

One of the main points of this summer project is the b-jet tagging. The current b-jet tagging in the ntuples of the $H \rightarrow \tau\tau$ analysis has a working point of 85%. This is reasonable for an analysis which does not heavily rely on the b-jets for identifying signal events. In the new channel, however, we expect 2 b-jets for our signal events. As such, we want to investigate if a 70% working point yields more sensitive results. When we plot the b-jet multiplicity of the 70% WP and compare it with the 85%, we see that a lot of events which were previously tagged with 1 or more b-jets shift towards 0 b-jets. It is clear (see Figure2) that these events with 0 b-jets are mostly background events. The probability of losing 2 real b-jets in a signal event is rather small. This gives us a good handle on the suppression of background events. From here on we will require that every event has at least 1 b-jet at the 70% WP to increase the signal-to-background ratio. To quantify our point, we can calculate the signal-to-background ratio for the b-jet multiplicity after requiring 1 or more b-jets, this yields an increase of $\approx$ 33% compared to b-jet tagging with 85% working point. We can also look at the ditau missing mass estimator (mmc)(see Figure3). This is an estimate for the mass of the Higgs boson given the properties of the 2 τs that were recorded in the event. When we require the 70% WP instead of the 85% WP, we see that the number of Z+jets-events is reduced by a factor $\approx$ 2.5.

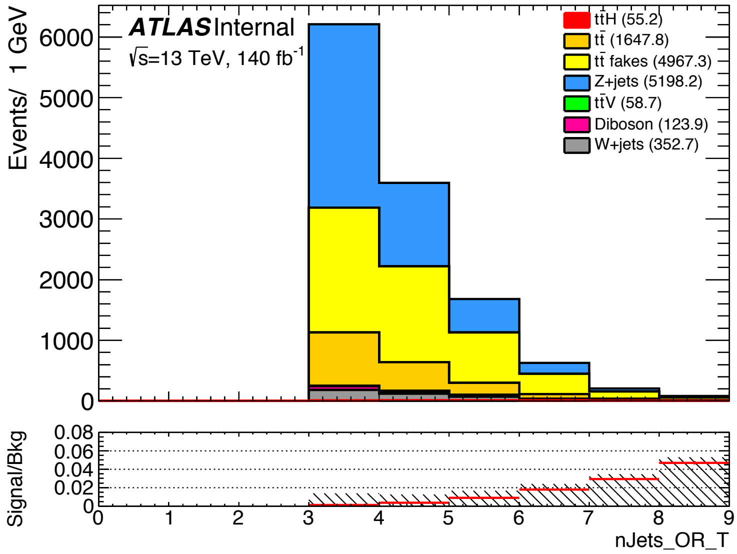

Figure 1: The distribution of jets.

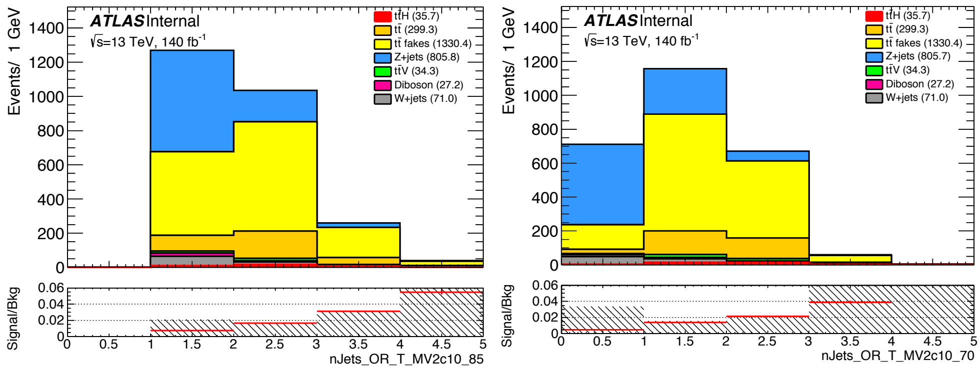

Figure 2: Left: The distribution of b-jets with 85% WP. Right: The distribution of b-jets with 70% WP

The number of $t\bar{t}H$ and $t\bar{t}$ events is reduced by $\approx 10\%$. This is in line with expectations because both $t\bar{t}H$ and $t\bar{t}$ decay to b-jets.

3.3 Preselection proposal

The main take away from this section is that we can reduce the background by requiring 5 or more jets but also by requiring at least 1 b-jet with 70% working point. We propose two channels and one backup

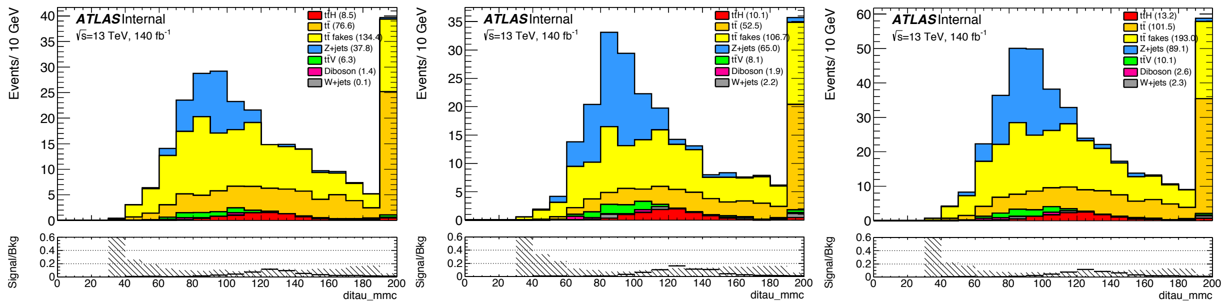

Figure 3: mmc with 70% WP. Left: for Njets=5, Nbjets $\geq$ 2. Middle: for $\geq$ 6, Nbjets $\geq$ 1. Right: for Njets $\geq$ 5, Nbjets $\geq$ 2 or Njets $\geq$ 6, Nbjets $\geq$ 1

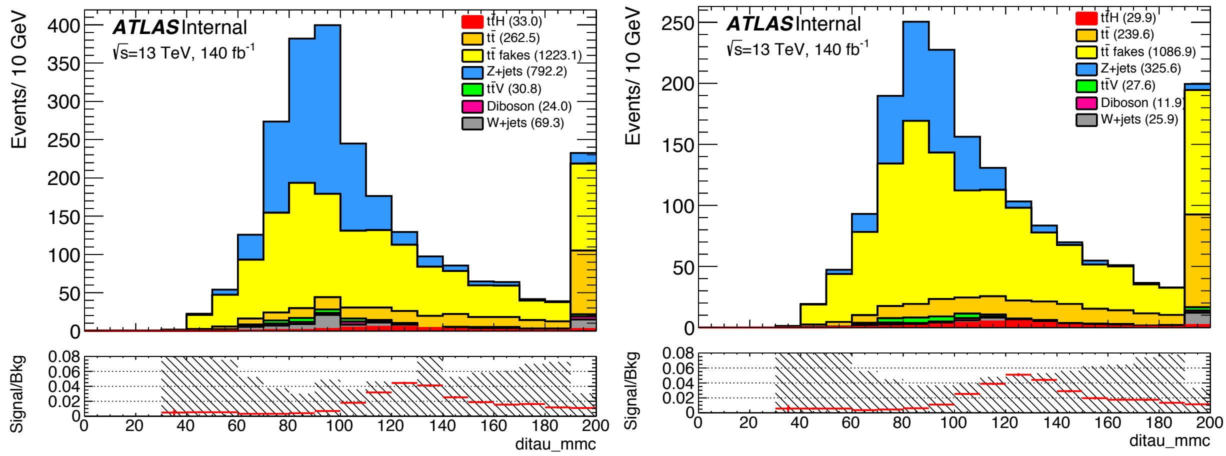

Figure 4: Left: mmc with 85% WP for Njets $\geq$5. Right: mmc with 70% WP for Njets $\geq$5, Nbjets $\geq$1

to investigate further. Either we can have (=5jets, $\geq$2b-jets) and ($\geq$6jets, $\geq$1b-jet) as separate channels, or we combine these two channels into an OR-selection to make up for the lack of Monte Carlo (MC) statistics. The resulting ditau mmc plots for these channels can be seen in Figure4. When we look at the signal to background ratios, we can see that this starts to look promising. We chose to not include the (=5jets, =1b-jet) selection in our channels because the number of Z+jet events is extremely high for this selection with very few signal events populating it.

4 Background and MVA

4.1 Background

Looking at Figure4, It is obvious that the background mainly consists of $t\bar{t}$ and Z+jets. That is why our MVA will also focus on these two. The $t\bar{t}$ events will be split into two contributions: those events producing only real τs and those producing at least one fake τ . The reason behind this decision is that the kinematics of these two types of events are very different (see Figure5). Splitting these contributions allows us to develop dedicated MVAs for each of these backgrounds. Therefore we have three BDTs: signal vs. $t\bar{t}$ with only real τs, signal vs.$t\bar{t}$ with at least one fake τ and signal vs. Z+jets.

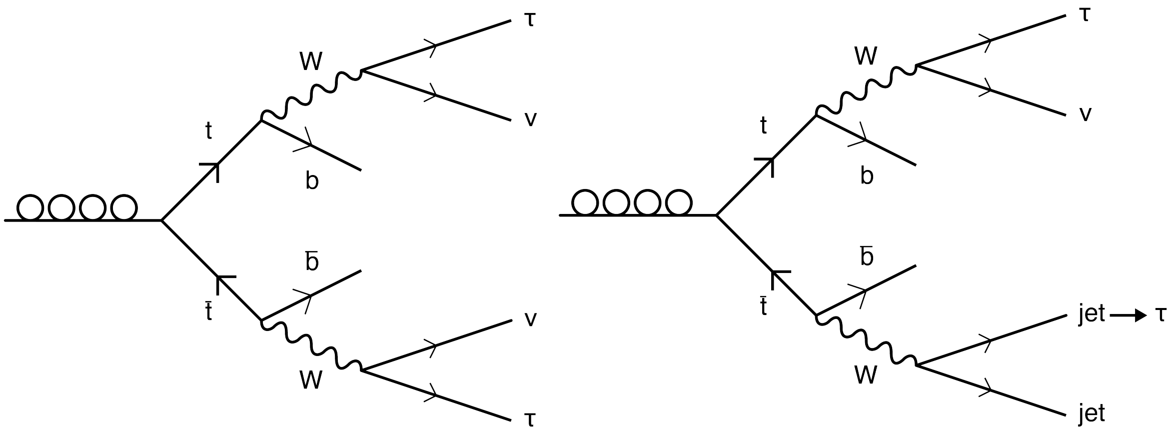

Figure 5: Feynmann diagrams of the $t\bar{t}$ background. Left: the contribution from $t\bar{t}$ with real τs. Right: the contribution from $t\bar{t}$ with at least one fake τ originating from a jet.

Samples	Training events	Testing events	Total
$t\bar{t}H$ vs $t\bar{t}$(Z+jets)	2593(2687)	1397(2687)	3989(3359)
$t\bar{t}$ fake	2934	1580	4514
$t\bar{t}$ real	3771	2514	6285
Z+jets	1878	470	2347

Table 1: The samples used in the BDT trainings.

4.2 MVAs

The particular setup of our MVA models for the different contributions can be found in the appendix. For both the $t\bar{t}$ BDTs we use 65% of the events for training, the other 35% is used for testing. For the Z+jets BDT we use 80% of the events for training while 20% is used for testing. Due to a lack of statistics (see Table1), the $t\bar{t}$ BDTs are trained with requirement of ≥4jets, ≥1b-jet while the Z+jets BDT is trained with a requirement of ≥5jets, ≥1b-jet. We will tighten these requirements again in the evaluation phase.
Before we discuss the variables used in the BDT trainings, it is important to note that we explicitly did not use ditau mmc or any variable correlated to ditau mmc for the training. We want to make our final fits on this variable and introducing this variable in the training phase would lead to a bias in the background mmc distribution.

4.2.1 MVA variables

The full list of variables used throughout the analysis can be seen in Table2. This list consists of five groups of variables. The missing transverse energy will play an important role in the BDT for the real $t\bar{t}$ because this is the only case where neutrinos are produced and thus a significant portion of transverse energy is lost. We are also using variables for reconstruction of the $t\bar{t}$ system, variables for the ditau system and variables

Variable name	
tau_eta_0	η of the leading τ
tau_eta_1	η of the second-leading τ
tau_pt_0	transverse momentum of leading τ
tau_pt_1	transverse momentum of second-leading τ
maxDeltaPhiTauMET	maximum ϕ separation of a τ and the missing transverse energy
minDeltaPhiTauMET	minimum ϕ separation of a τ and the missing transverse energy
ditau_deta	eta separation between leading and second-leading τ
ditau_dR	angular separation between leading and second-leading τ
ditau_pT	transverse momentum of the two τ system
nJets_OR_T_MV2c10_70	number of reconstructed b-jets with a WP of 70%
nJets_OR_T	number of reconstructed jets
lead_jetEta	η of the leading jet
sublead_jetEta	η of the second-leading jet
lead_jetPt	transverse momentum of the leading jet
htjets_1l2tau	sum of transverse energy of all jets (used in ttH multilepton channel)
SumPtBJet	scalar sum of the transverse momentum of all b-jets
jjdrmin_1l2tau	minimum angular separation between two jets for all combinations of jets
DeltaRTauClosestJet	minimum angular separation between jet and a tau
MET_RefFinal_et	missing transverse energy
mWbest	invariant mass of a pair of non-btagged jets closest to the W boson mass
mTopWbest	invariant mass closest to top quark mass of the best W candidate with a b-jet
mWsubbest	invariant mass of a pair of non-btagged jets second closest to the W boson mass

Table 2: The variables used in the BDTs.

that describe jets and τs. In Table3, one can see the variables ranked according to importance as done by the MVA algorithm. Note that the angular separation between the two τs is a variable that is important in every BDT while the $t\bar{t}$ reconstruction parameters only play a role in the $t\bar{t}$ BDTs.

4.2.2 MVA results

The results for the trainings can be seen in Figure6. There is a very good separation between signal and background for the Z+jets contribution and slightly less satisfying results for the $t\bar{t}$ contributions, being more difficult to discriminate signal from the $t\bar{t}$ contributions. The efficiencies for the trainings can be found in Table4.

After the training of the BDTs, we can plot the BDT score for every background contribution. It is clear from Figure6 that for the Z+jets BDT the bulk of the signal resides at high MVA scores. However some of the contributions coming from $t\bar{t}$ also reside at hight Z MVA scores. The picture is less pronounced for the $t\bar{t}$ BDTs.

	Z+jets	$t\bar{t}$ real	$t\bar{t}$ fake
1	htjets_1l2tau	ditau_dR	ditau_dR
2	sublead_jetEta	nJets_OR_T	maxDeltaPhiTauMET
3	ditau_deta	MET_RefFinal_et	tau_pt_0
4	ditau_dR	mWbest	htjets_1l2tau
5	tau_pt_0	sublead_jetEta	nJets_OR_T
6	minDeltaPhiTauMET	maxDeltaPhiTauMET	ditau_deta
7	tau_pt_1	lead_jetEta	lead_jetEta
8	jjdrmin_1l2tau	ditau_pT	mWbest
9	DeltaRTauClosestJet	DeltaRTauClosestJet	DeltaRTauClosestJet
10	maxDeltaPhiTauMET	SumPtBJet	tau_pt_1
11	tau_eta_1	ditau_deta	sublead_jetEta
12	lead_jetEta	jjdrmin_1l2tau	MET_RefFinal_et
13	MET_RefFinal_et	minDeltaPhiTauMET	tau_eta_1
14	SumPtBJet	mTopWbest	jjdrmin_1l2tau
15	nJets_OR_T_MV2c10_70	lead_jetPt	tau_eta_0
16	nJets_OR_T	mWsubbest	SumPtBJet

Table 3: The importance ranking according to the BDT of the variables.

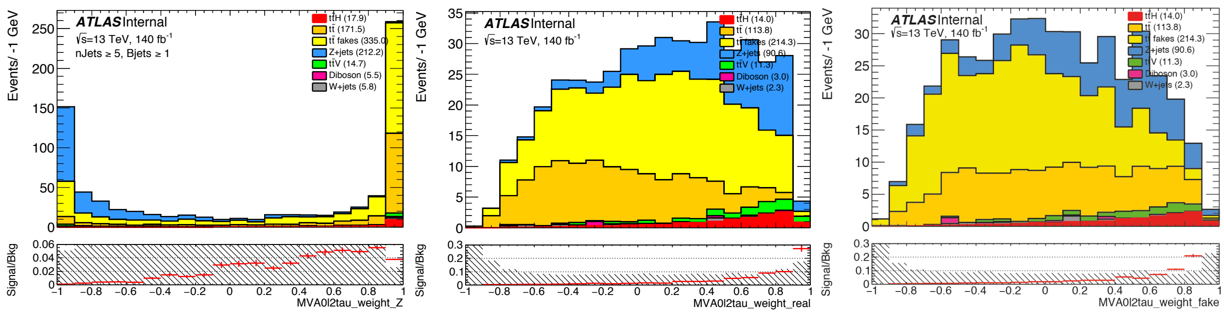

Figure 6: MVA scores. Left: $t\bar{t}H$ versus Z+jets. Middle: $t\bar{t}H$ versus $t\bar{t}$ real. Right: $t\bar{t}H$ versus $t\bar{t}$ fake

	Signal efficiency @ B=0.01 test(training)	Signal efficiency @ B=0.10	Signal efficiency @ B=0.30
$t\bar{t}H$ vs $t\bar{t}$ real	0.203 (0.289)	0.590 (0.626)	0.805 (0.810)
$t\bar{t}H$ vs $t\bar{t}$ fake	0.191 (0.306)	0.549 (0.620)	0.786 (0.831)
$t\bar{t}H$ vs Z+jets	0.228 (0.714)	0.602 (0.868)	0.838 (0.938)

Table 4: The results from the BDT trainings.

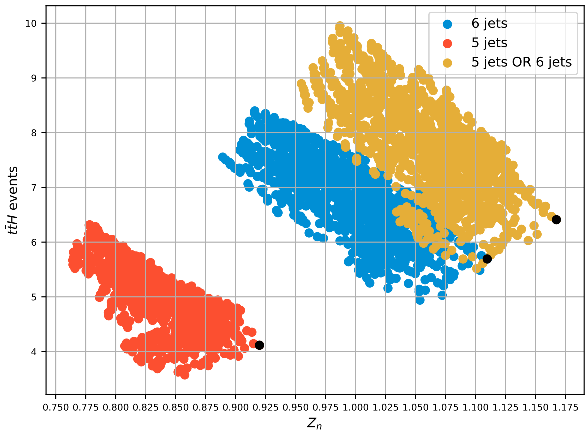

Figure 7: The result from the gridsearch. The OR selection provides the highest significance with cuts at 0.5, 0.4 and 0.3 for respectively the MVA score for $t\bar{t}$ real, $t\bar{t}$ fake and Z+jets. The black dots represent the highest significance-cut for every channel.

5 Analysis strategies

There are two possible roads to take with the results from the BDTs. Either we choose to make cuts on the MVA score of Z+jets, $t\bar{t}$ real and $t\bar{t}$ fake, while fitting on ditau mmc or we choose to apply a cut on the ditau mmc distribution while we fit to the MVA scores. For both options we will be using TRExFitter.

5.1 Cut on MVA - Fit MMC

For a rough idea of the optimal cuts on the MVA, we use a gridsearch in the three dimensional space of all possible combinations of cuts on the three MVAs. The significance of the signal over background is estimated by Formula1. The resulting significance of signal over background versus the number of events can be seen in Figure7. The OR selection clearly gives the highest significance with a result of almost 1.167σ while the six jet and five jet channels give a significance of respectively 1.11σ and 0.92σ.

$$Z_n = BinomialExpZ(sig, bkg, \frac{err_{bkg}}{bkg}) \tag{1}$$

Now that we have this preliminary result, it is possible to do a fully fledged fit using TRExFitter. When using the OR region, we get a resulting signal significance of 1.25σ. That is very close to our earlier calculated result of 1.167σ. When we use the two separate regions, (=5,≥2) and (≥6,≥1), but also use the shape of the missing mass estimator and add a theoretical uncertainty of 50% to the normalization of Z, $t\bar{t}$ real, and $t\bar{t}$ fakes on top, we get a signal significance of 1.08σ (see Figure8).

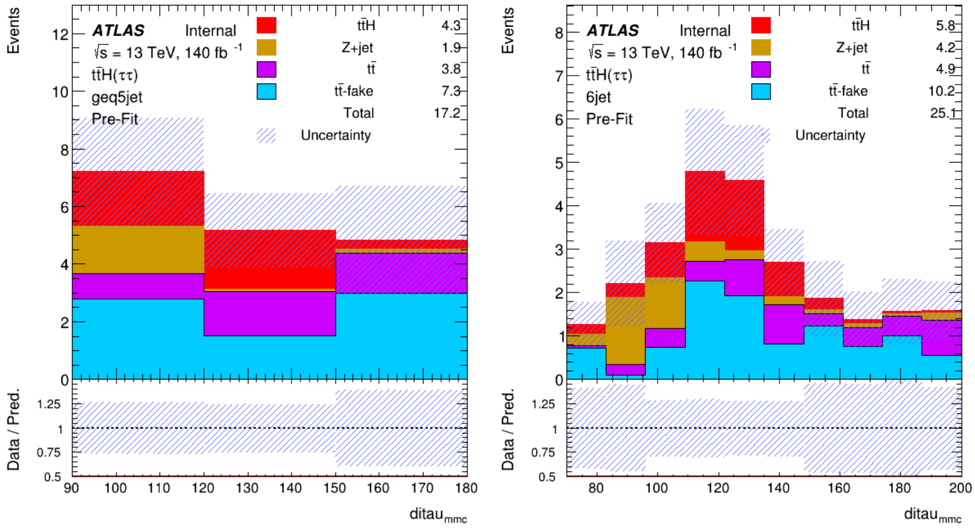

Figure 8: The MMC fit. The resulting significance is 1.85.

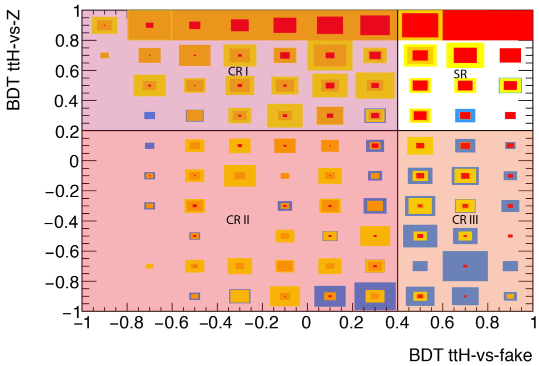

Figure 9: The different regions used in the MVA fit. Red is signal, yellow is $t\bar{t}$ fake, blue is Z+jets, orange is $t\bar{t}$ real.

5.2 Cut on MMC - Fit MVA

Fitting the MVA scores while cutting on the missing mass estimator should yield more or less the same result as calculated in the section above. However, by fitting the MVA scores it is possible to create control regions that can be used to constrain the background. We require the missing mass estimator to be in the range 100≤MMC≤150 GeV and make one signal region and three control regions using Figure9 as guidance. The MVA score of $t\bar{t}H$ versus $t\bar{t}$ real is fitted in the signal region while the MVA score of $t\bar{t}H$ versus $t\bar{t}$ fake is fitted in the control regions. We add, again, a theoretical uncertainty of 50% to the normalization of Z, $t\bar{t}$ real, and $t\bar{t}$ fakes. This procedure gives a signal significance of 1.48σ. Table5 lists the signal significance for every option we discussed.

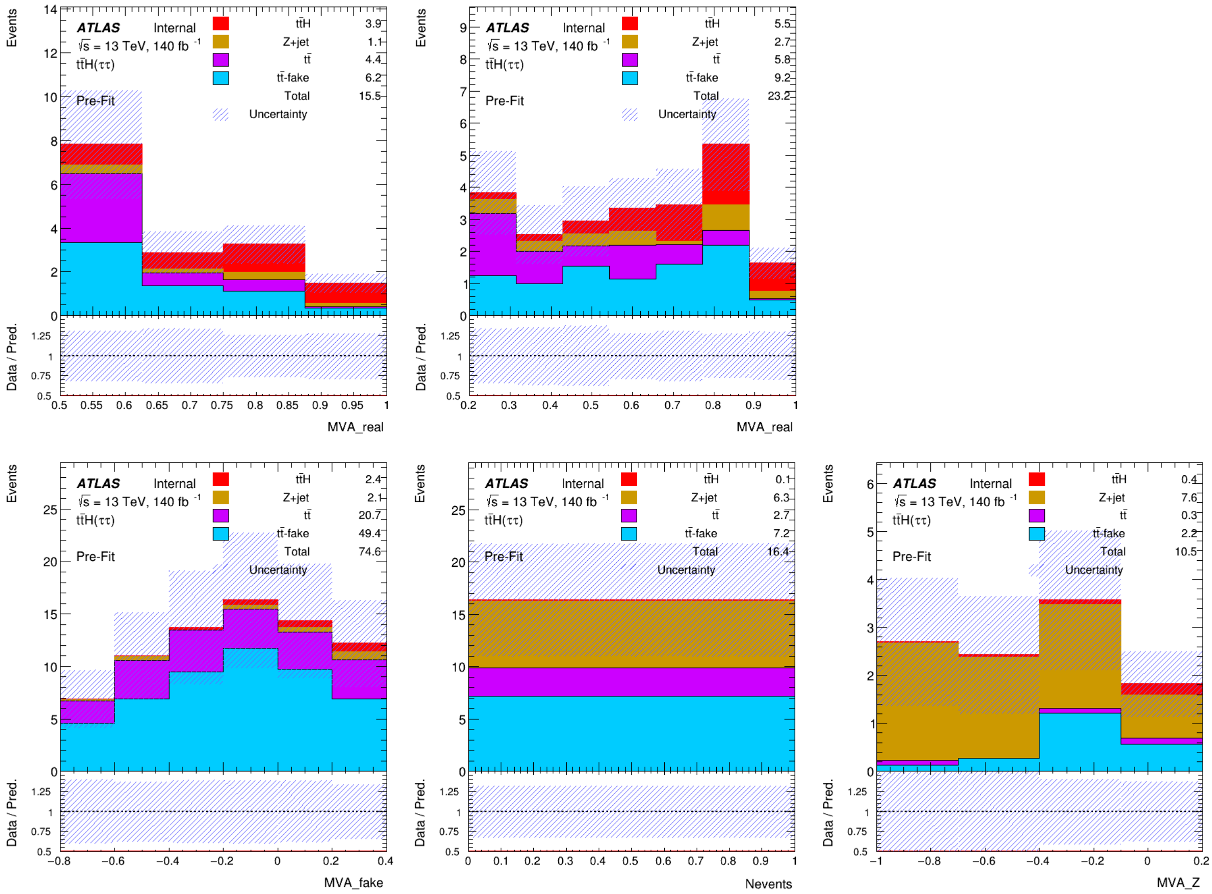

Figure 10: The MVA fit. The resulting significance is 1.48. Upper left is the signal region for (=5,≥2) and upper right is the signal region for (≥6,≥1). Down left is control region 1 where the MVA score of signal vs fake is fitted, down middle is control region 2 and down right is control region 3 where MVA score of signal vs fake is fitted.

	Signal significance syst+stat (stat)
=5j≥2bjet (gridsearch)	(0.92)
≥6j≥1bjet (gridsearch)	(1.11)
OR selection (gridsearch)	(1.167)
OR selection (mmc shape not used)	(1.25)
Separate regions Option 1	1.08
Separate regions Option 2	1.48

Table 5: The signal significance after using formula1 (gridsearch) or after fit.

6 Results

It is clear that the proposed channel is a very powerful one. The new proposed channel will contribute significantly to the sensitivity of the tau channels in $t\bar{t}H$ multilepton. This first optimization study will help in the design of the $0\ell + 2\tau_{\text{had}}$ $t\bar{t}H$ multilepton channel for the full Run-2 publication and the results were shown in the ATLAS $H \to \tau\tau$ plenary meeting in August 2019.

References

[1] *TMVA*, https://twiki.cern.ch/twiki/bin/view/TMVA/WebHome (cit. on p. 2).

[2] *TRExFitter*, https://gitlab.cern.ch/TRExStats/TRExFitter (cit. on p. 2).