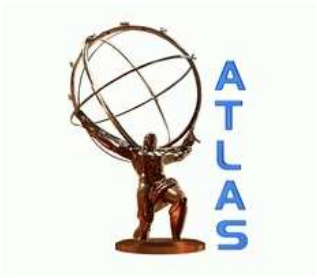
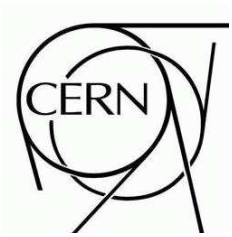


ATLAS NOTE



April 28, 2009

Search for $t\bar{t}H(H \to b\bar{b})$

The ATLAS Collaboration[1)]

This note is part of CERN-OPEN-2008-020. This version of the note should not be cited: all citations should be to CERN-OPEN-2008-020.

ATL-PHYS-PUB-2009-057
10 May 2009

Abstract

For a light Higgs boson, with $m_H \leq 135$ GeV, the largest decay mode is $H \to b\bar{b}$. Events where the Higgs boson is produced in association with a $t\bar{t}$ pair manifest a distinct signature due to the presence of two W bosons and four b quarks. Topological and kinematical quantities are used to reconstruct the $t\bar{t}$ system. The identification of an additional $b\bar{b}$ pair from the Higgs boson decay is used to further reduce the background.

In this analysis we focus on the sensitivity to a light Standard Model Higgs boson with the ATLAS detector in the channel $t\bar{t}H(H \to b\bar{b})$ using the semi-leptonic final state with 30 fb^{-1} of integrated luminosity. The relevant backgrounds to the channel are investigated and the impact of their associated systematic uncertainties is explored.

[1)]This note prepared by: G. Aad, L. Asquith, S. Boeser, C. Bernius, C. Collins-Tooth, S. d'Auria, S. Dean, L. Feligioni, R. Goncalo, N. Konstantinidis, W.J. Murray, A. Rozanov, L. Vacavant.

1 Signal

At the LHC $t\bar{t}H$ production is dominated (90%) by gluon fusion, as illustrated in Fig. 1. The remaining 10% arises from quark-antiquark interactions. For a Higgs boson mass between 115 GeV and 130 GeV the production cross-section times branching ratio to $b\bar{b}$ varies between roughly 0.4 and 0.2 pb at leading order. The top quarks decay almost exclusively to bW, and therefore the various final states can be classified according to the decays of the W bosons.

The all-hadronic channel is the one with the highest branching fraction, with a value of 43%. Unfortunately, the large QCD multijet cross-section does not allow easy triggering with jets. Only tight requirements on the jet p_T and on the jet multiplicity could lead to reasonable rates in the first level of the trigger, but these requirements come at the expense of signal efficiency. This is being studied for this final state together with the use of b-tagging at the second level of the trigger to reduce the jet p_T threshold.

The fully-leptonic final state analysis is probably the least feasible, despite presenting a simpler signature to trigger on, given the presence of two isolated leptons. The branching fraction (5%) is low and the two neutrinos prevent the reconstruction of the top quarks.

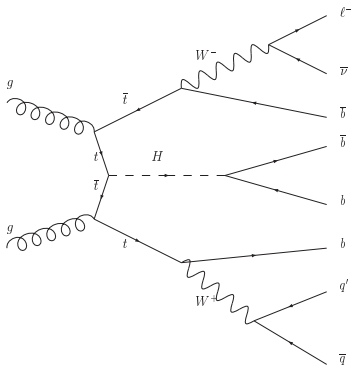


Figure 1: One of the Feynman diagrams for $t\bar{t}H$ production in the semi-leptonic final state.

The semi-leptonic final state is a good compromise with a branching fraction of about 28% excluding tau leptons. The experimental signature consists of one energetic isolated lepton, a high jet multiplicity with multiple b-tags, and missing transverse energy from the escaping neutrino, as shown in Fig. 1. The trigger relies on the presence of the high-p_T lepton. The non-b-tagged jets can be used for reconstructing the hadronically decaying W boson, decreasing the possible combinatorial permutations.

2 Physics backgrounds

The production of $t\bar{t}$ events is the main background for the $t\bar{t}H$ process. Given the high jet multiplicity in the signal process (≥ 6 jets), only $t\bar{t}$ events produced together with at least two extra jets contribute to the preselected data sample. Since most of these extra jets come from the hadronisation of light quarks, this contribution is greatly reduced by asking for four jets to be identified as b-jets.

The irreducible background comes from $t\bar{t}b\bar{b}$ production. This can proceed via QCD or electroweak, (EW), interactions with a total cross-section of the order of 9 pb. Some of the Feynman diagrams involved in the two production mechanisms are shown in Fig. 2 and Fig. 3. While the QCD production cross-section is ten times larger than the EW production, the latter is also important. The two b-jets not coming from the $t\bar{t}$ decay have large momenta and also have a total invariant mass which is typically close to the Z boson mass, and can therefore contaminate the signal region.

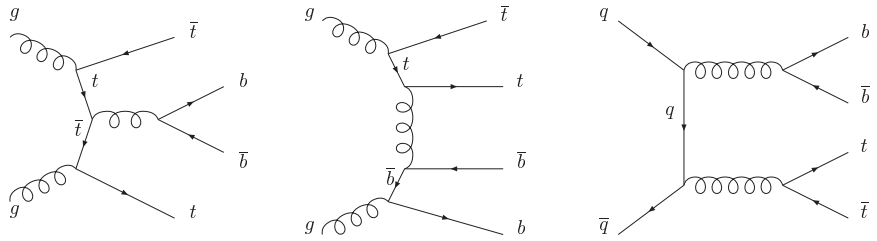


Figure 2: Example of Feynman diagrams for the $t\bar{t}b\bar{b}$ QCD production.

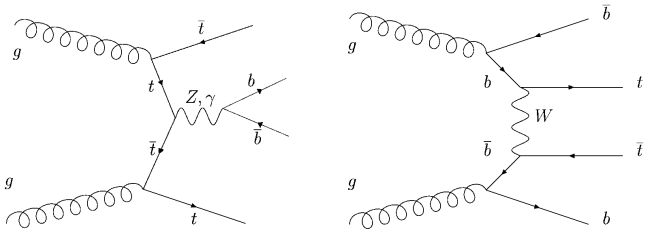


Figure 3: Example of Feynman diagrams for the $t\bar{t}b\bar{b}$ EW production.

The $t\bar{t}c\bar{c}$ background cross-section is 60% higher than $t\bar{t}b\bar{b}$ [1], so it could also play an important role. However, it is found upon investigation that due to the c-jet rejection factor, the $t\bar{t}c\bar{c}$ background plays a negligible part in comparison with the $t\bar{t}b\bar{b}$ background. No dedicated sample is therefore simulated for this study. Some $t\bar{t}c\bar{c}$ events are however present in the inclusive $t\bar{t}$ +jets sample used.

Several other backgrounds, such as W+jets, tW production and QCD multijet production, could also have a non-negligible impact on the analysis. Even though the W plus two jets inclusive cross-section is about 1200 pb per lepton flavor [2], it has been shown [3] and confirmed in this analysis that the contribution can be reduced to a negligible level if the four b-tags requirement is applied. This is also true for the less abundant tW background, which has a cross-section of 9.5 pb [4]. Even when four b-jets are requested in the event, contamination via QCD $b\bar{b}b\bar{b}$ production, which has a cross-section of a few hundred nb [5], is still possible. The reconstruction of the $t\bar{t}$ system allows a certain degree of safety against non-top background. None of these samples are presented in what follows.

3 Monte Carlo samples and cross-sections

This study uses the leading order cross-sections for the signal and $t\bar{t}b\bar{b}$ samples. No calculation has yet been performed for $t\bar{t}b\bar{b}$ at next-to-leading order, (NLO). This study also uses a $t\bar{t}$ background, simulated at next-to-leading order. This is the only NLO Monte Carlo sample used, and the only $t\bar{t}$ large sample available in ATLAS at the time this analysis was performed. No K-factors are applied to the samples simulated at leading order in the significance estimates.

The signal sample is generated for a Higgs boson mass of $m_H = 120$ GeV with PYTHIA [6] 6.403. The exact generated process was $pp \to t\bar{t}H\mathrm{X} \to \ell\nu bq\bar{q}'b\bar{b}b\mathrm{X}$, with $\ell =$ e or μ. The factorization and renormalization scales used are identical and are listed in Table 1. The signal and the $t\bar{t}b\bar{b}$ events are generated with a lepton filter requiring at least one electron or one muon with pseudorapidity $|\eta| < 2.7$ and transverse momentum above 10 GeV. The leading order production cross-section used is $\sigma(t\bar{t}H) =$ 537 fb [7]. The branching ratios $H \to b\bar{b}$ of 67.5% at 120 GeV [7], $W \to \ell\nu$ of 10.66% [8], and $W \to$ hadrons of 67.6% [8] is applied. Finally the lepton filter efficiency of $\varepsilon = 0.953$ is also applied. The resulting cross-section is 100 fb.

For both $t\bar{t}b\bar{b}$ QCD and EW samples, the exact process generated is $\mathrm{gg} \to t\bar{t}b\bar{b}\mathrm{X} \to \ell\nu bq\bar{q}'b\bar{b}b\mathrm{X}$,

with $\ell = \mathrm{e}$ or μ. Both processes can be initiated by a $q\bar{q}$ pair, but only the dominant gluon fusion is simulated, with the cross-sections being increased to allow for the $q\bar{q}$ pair production. For the $t\bar{t}b\bar{b}$ QCD sample, AcerMC 3.4 [9] is used and interfaced to PYTHIA 6.403 for the simulation of the initial and final state radiation, hadronisation and decay. The $t\bar{t}b\bar{b}$ EW sample is generated using AcerMC 3.3 and PYTHIA 6.403. The leading order $t\bar{t}b\bar{b}$ QCD cross-section is $\sigma(pp \to t\bar{t}b\bar{b}) = 8.2(gg)(+0.5(q\bar{q}))$ pb and the lepton filter efficiency is $\varepsilon = 0.946$. For the $t\bar{t}b\bar{b}$ EW sample, the leading order cross-section is $\sigma(pp \to t\bar{t}b\bar{b}) = 0.90(gg)(+0.04(q\bar{q}))$ pb and the lepton filter efficiency is $\varepsilon = 0.943$.

The reducible $t\bar{t}$ background events are generated with the MC@NLO [10] program, interfaced to HERWIG [11] and Jimmy [12]. The events in this sample correspond to the processes $pp \to t\bar{t} \to (\ell\nu, q\bar{q}')b\ell\nu b$ with $\ell = \mathrm{e}, \mu, \tau$. The generator versions used are MC@NLO 3.1 and HERWIG 6.510. For the inclusive $t\bar{t}$ cross-section we use the NLO+NLL calculation of $\sigma(pp \to t\bar{t}) = 833$ pb. The $t\bar{t}$ sample is also produced using a filter requiring one electron or one muon with pseudorapidity $|\eta| < 2.7$ and transverse momentum above 14 GeV. The $t\bar{t}$ filter also applies requirements on the jets in the generated events which are reconstructed using a seeded fixed-cone algorithm with a cone size of $\Delta R = 0.4$ [13], by requiring at least:

- six jets with $p_T > 14$ GeV and $|\eta| < 5.2$
- four jets with $p_T > 14$ GeV and $|\eta| < 2.7$

The efficiency of this generator filter on inclusive $t\bar{t}$ events is 0.146.

For the $t\bar{t}$ sample, about 10% of events are $t\bar{t}b\bar{b}$ and are removed following the overlap treatment explained in Ref. [14] together with their associated cross-section.

Table 1 summarizes the cross-sections, calculated using the Monte Carlo generators, of the different processes considered for this analysis, together with the corresponding numbers of generated events and the equivalent integrated luminosity. All branching fractions and filter efficiencies are included.

Table 1: Summary of the different samples used for the analysis. The cross-sections are taken from the generators and include all branching fractions and filter efficiencies. The fourth column shows the equivalent integrated luminosity, taking into account all corrections (see text). For the scale calculations, $m_H = 120$ GeV and $m_t = 175$ GeV are used. $\max({p_T}^2_t, {p_T}^2_{\bar{t}})$ corresponds to the higher of the two values of ${p_T}^2$ when both the top and anti-top quarks are considered.

Process	σ (fb)	Events	L (fb^{-1})	Fact. & Renorm Scale	PDF set
$t\bar{t}H$ (LO)	100	92750	931	$Q^2 = m_t^2 + \max({p_T}^2_t, {p_T}^2_{\bar{t}})$	CTEQ6L1
$t\bar{t}b\bar{b}$ QCD (LO)	2371	98350	42	$Q = m_H/2 + m_t = 235$ GeV	CTEQ6L1
$t\bar{t}b\bar{b}$ EW (LO)	255	24750	97	$Q = m_H/2 + m_t = 235$ GeV	CTEQ6L1
$t\bar{t}$ filtered (NLO)	109487	710321	6.5	$Q^2 = m_t^2 + \frac{1}{2}({p_T}^2_t + {p_T}^2_{\bar{t}})$	CTEQ6M

4 Analysis overview

The analysis consists of an initial preselection requirement which is applied to the events to ensure that the fundamental physics objects associated with $t\bar{t}H$ are reconstructed. Following preselection, three different analysis techniques are implemented in order to reconstruct the top quark pairs and the Higgs boson through the identification of their decay products.

The identification and association of decay products is directly related to the quality of the reconstructed Higgs boson signal. It mainly suffers from the misassociation of the four b-tagged jets to the original partons. For this reason, the initial cut-based approach is complemented by two multivariate algorithms, called the pairing likelihood and constrained mass fit.

5 Preselection

At the preselection level we require that the event passes the trigger requirement to identify at least one high-p_T lepton (muon or electron) coming from the decay of one of the W bosons. We then require that the event reconstruction identifies exactly one isolated high-p_T lepton (muon or electron). Vetoing the presence of a second isolated lepton is intended to remove additional sources of background. After the lepton requirements are met, we require at least six calorimeter jets, of which at least four must be loosely b-tagged jets from the decay of the top quarks and the Higgs boson.

5.1 Trigger requirements

The presence of one high-p_T lepton, together with missing transverse momentum, is a distinct signature of W boson production. These leptons can generally be used to trigger on W production with high efficiency. A logical OR of the single isolated electron (`e22i`) [15], high-p_T electron (`e55`) [15] and single muon (`mu20`) [16] triggers is used. The inclusion of the `e55` trigger is found to improve the efficiency for high-p_T electrons where the `e22i` trigger efficiency was reduced due to the isolation requirement. Missing energy triggers were not available at the time of writing, but could be used in future analysis.

The trigger efficiency is approximately 82% for the semileptonic top decays for those events which would otherwise pass the offline analysis. This is included consistently in the following sections.

5.2 Reconstructed high p_T lepton selection

In this Section we explain the selection criteria used for reconstructed electrons and muons produced in the semi-leptonic decay of the $t\bar{t}$ system. As previously mentioned, exactly one high-p_T isolated electron or muon must be reconstructed for the event to pass the preselection.

To be considered for the analysis, reconstructed electrons must have transverse momentum $p_T >$ 25 GeV and pseudorapidity $|\eta| < 2.5$. Further calorimeter-based cuts are applied to the loose electron [17] definition. An isolation cut is also applied to the candidate electrons in the form of an upper limit of 0.15 on the ratio of the p_T of the additional tracks inside a cone of size 0.2 in ΔR, ($\sqrt{\Delta\eta^2 + \Delta\phi^2}$), around the electron track to the electron p_T.

Muon candidates are reconstructed using a combination of the Inner Detector and Muon Spectrometer [18]. They must pass the acceptance cuts $p_T > 20$ GeV and $|\eta| < 2.5$. In order to remove poorly reconstructed muons, cuts are applied to the muon track fit quality and its transverse impact parameter, which helps to discriminate against muons generated by the decay of long-lived mesons.

An isolation cut of 0.30 on the ratio between the transverse energy deposited inside a cone of size 0.2 in ΔR around the muon track and the muon p_T is applied.

5.3 Jets

To reconstruct the energy of the partons produced in the original collision, calorimeter jets are reconstructed using a seeded fixed-cone algorithm with a cone size of $\Delta R = 0.4$ [13]. Cuts on $p_T > 20$ GeV and $|\eta| < 5.0$ are initially applied. Only events with at least 6 jets are kept for the analysis. All electrons reconstructed as jets are identified and removed from the jet collection according to the electron overlap removal procedure described in Section 5.3.1. Reconstructed muons which are not isolated are combined into jets where applicable (see Section 5.3.3), and only after this step are the jet energies calibrated for residual effects. The jet multiplicity is shown in Fig. 4, calculated after electron overlap removal and the jet η and p_T cuts.

In the following analyses the concept of 'correct' jets is important. This is defined by finding the closest reconstructed jet to each parton, after final state radiation. This match is in ΔR space and must

be closer than 0.4. A W or H boson is correctly matched if both the jets being used are associated to the partons from its decay, while for a top quark the matching refers to the b quark jet only. Normally 'correct' is applied in this note only to one quark or boson at a time.

5.3.1 Treatment of overlaps between jets and electrons

Since most electrons are also reconstructed by the jet algorithm it becomes necessary to identify them in the jet collection in order to avoid double counting. The criteria for the jet-electron overlap removal is the following: each jet matching a well-reconstructed electron (i.e. fulfilling the cuts defined in Section 5.2) within a ΔR of 0.2, and for which the ratio of the electron to the jet transverse momenta is greater than 0.75, is discarded from the jet collection. About 4% of the jets in the signal sample are removed by this selection, 99% of them being actual true electrons.

5.3.2 *b*-tagging

b-jets are identified using the IP3D+SV1 tagger [19], which exploits both the impact parameter of tracks and the properties of an inclusive secondary vertex, using a likelihood approach which leads to a single discriminating variable: the *b*-tagging weight. In order to allow for a projected decrease in light jet rejection of approximately 30%, the *b*-tagging weight for this study is increased by 0.9 for central jets ($|\eta| < 2.5$) having no associated ($\Delta R < 0.3$) heavy quark or lepton (b,c,τ) in the Monte Carlo simulation history. The *b*-tag weight spectrum for *b*-, *c*-, and light jets in the signal sample is shown in Fig. 4. A cut on the weight defines which jets will be eventually identified as *b*-jets in the analysis. The rejection of *c*- and light jets versus the *b*-jet efficiency, obtained by varying the weight cut, is also shown in Fig. 5. The different samples exhibit very similar behaviour. The rejection for "purified" jets is shown in the bottom row of Fig. 5. Purified jets are those which have no heavy flavor (b, c, τ) quark or lepton within 0.8 in ΔR.

In the preselection, a loose set of criteria are initially used to define a sample of jets as a first step to identifying b jets for the analyses. The requirements are that the jet is in the central region of the detector $|\eta| < 2.5$, and has *b*-tag weight ≥ 0. If there are fewer than four of these jets, the event is discarded.

The cut-based analysis and the pairing likelihood (Sections 8 and 9) require that there are at least four *b*-jets having *b*-tag weight ≥ 5.5. The *b*-tag weight ≥ 0 working point implies a *b*-tagging efficiency of about 85% and a rejection of light (*c*-) jets of about 8.6 (2.4), whereas the working point at *b*-tag weight ≥ 5.5 implies a *b*-tagging efficiency of about 65% and a rejection of light (*c*-) jets of about 60 (6).

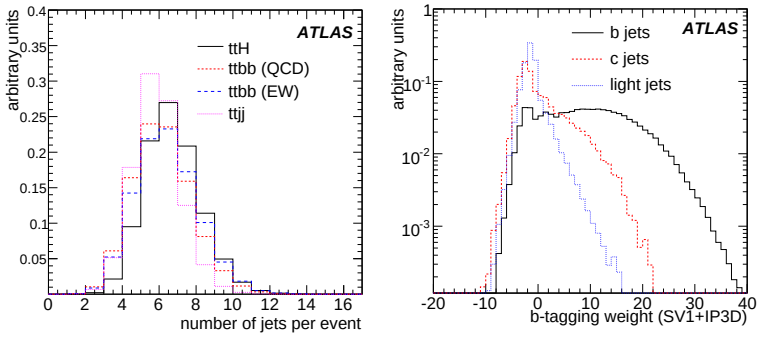


Figure 4: Left: Multiplicity of jets, inside p_T and η acceptance. Right: Distribution of *b*-tagging weight for *b*-, *c*- and light jets in $t\bar{t}H$ events, using the IP3D+SV1 tagger.

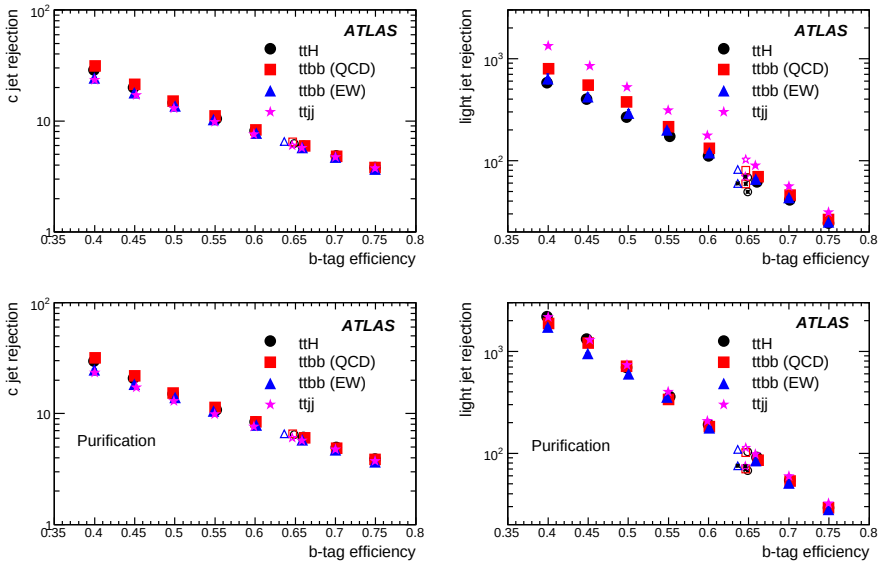


Figure 5: Rejection of light and *c*-jets versus *b*-tagging efficiency. Red open markers indicate the working point (*b*-tag weight ≥ 5.5) used for the cut-based and paring likelihood analyses, open markers with a central dot (right plots) represent the performance when the 30% performance degradation is applied. The lower plots show results for purified jets, where no heavier quark existed within a wide cone of $\Delta R < 0.8$.

5.3.3 Treatment of low p_T muons

About 20% of the time a *B*-meson decay cascade gives rise to a muon. With a four *b*-jet signature in this channel, these muons, also called *soft muons* are present in almost every event. In order to improve the estimate of the momentum of the original *b* quark, these muons must be used to correct the jet four-momenta by adding the muon four-momentum to a jet. Two different algorithms (high and low-p_T [20]) identify the muons, which are required to be within $\Delta R < 0.4$ of the jet axis. Among the candidates from the high-p_T algorithm which are separated from the selected hard lepton by $\Delta R > 0.1$, only the one with the best track quality is considered for addition. All the neighbouring candidates from the low-p_T algorithm are considered for addition, provided they fulfil $p_T > 4$ GeV and $p_T < 100$ GeV, $|\eta| < 2.5$ and chi-squared per degree of freedom $\chi^2/n < 30$ for the combined fit. In addition, a loose anti-isolation cut is applied, requiring that the energy reconstructed in the calorimeter within a cone of size 0.2 in ΔR around the muon track divided by the muon p_T is higher than 0.1. Table 2 shows that adding low p_T muons to jets improves both the mean jet p_T and resolution. Fig. 6 shows that there is also an improvement in the Higgs boson mass and resolution, when the correct jet combination is chosen.

5.3.4 Calibration

A Monte Carlo based jet correction has been derived to take into account residual calibrations, *e.g.* out-of-cone effects and neutrinos. The parametrization was derived from full simulation, so that the jet four-momentum is corrected by a flavor dependent rescaling factor which scales all components of the four-momentum. Table 2 shows that for *b*-jets, the residual calibration brings the jet and associated parton p_T into agreement from an offset of 5.4 to -0.5 GeV. Each of the analyses uses the light jet correction for those jets which are assigned to the W boson and corrects the other four jets as *b* quark originated. The impact of the calibration on the Higgs boson mass peak where the correct jet combination is chosen is shown in Fig. 6. Both the mass peak location and the resolution are improved.

Table 2: The true parton p_T minus the measured jet p_T with and without adding muons and making the out-of-cone correction. The quoted values are the results of a Gaussian fit in the region $\pm$20 GeV.

Treatment	Value	No Calibration	Added muons	Calibrated	Both corrections
All b jets	Mean, GeV	5.7$\pm$0.05	5.4$\pm$0.05	-0.1$\pm$0.04	-0.5$\pm$0.04
	Sigma, GeV	10.0$\pm$0.05	9.8$\pm$0.05	10.2$\pm$0.05	10.0$\pm$0.04
b jets with muons	Mean, GeV	26$\pm$2.6	7.6$\pm$0.2	11.4$\pm$0.5	2.3$\pm$0.2
	Sigma, GeV	18.2$\pm$1.1	11.8$\pm$0.2	13.6$\pm$0.4	12.0$\pm$0.2
All light jets	Mean, GeV	2.4$\pm$0.04	2.3$\pm$0.04	-1.4$\pm$0.04	-1.5$\pm$0.04
	Sigma, GeV	8.4$\pm$0.04	8.4$\pm$0.04	8.9$\pm$0.04	8.3$\pm$0.04
Light jets with muons	Mean, GeV	10.5$\pm$0.9	2.4$\pm$0.4	7.3$\pm$0.7	-1.0$\pm$0.4
	Sigma, GeV	11.3$\pm$0.8	10.9$\pm$0.5	11.5$\pm$0.6	10.7$\pm$0.4

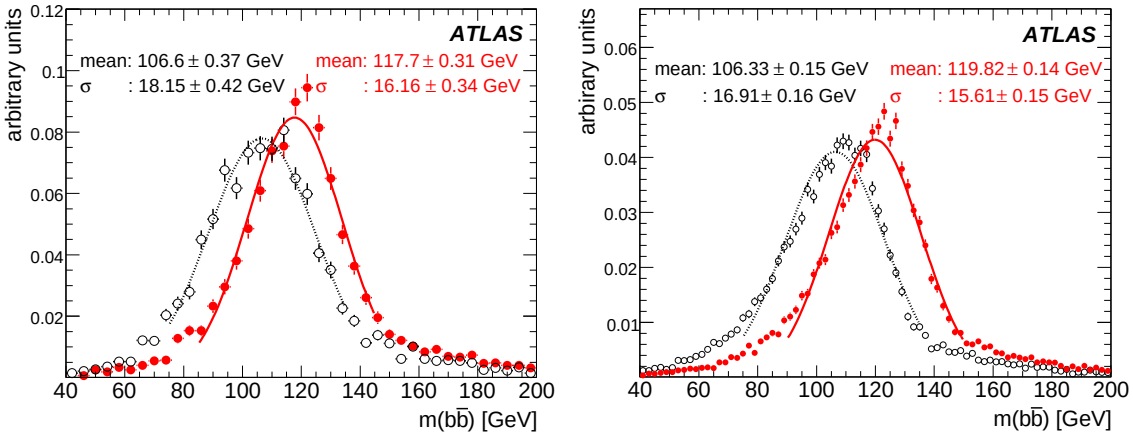


Figure 6: Effect of adding a reconstructed low p_T muon to at least one of the b jets (left), and the effect of jet calibration (right) shown for the Higgs boson mass where the correct jet combination is chosen. Red solid (black open) markers and solid (dotted) line show the mass distribution and fitted values for jets after (before) correction. The effect of the accidental wrong muon matches can be seen in the shape distortion. All distributions are normalized to unity.

5.4 Results of preselection on signal and background

The effect of the event preselection on signal and background samples is illustrated in Table 3. The efficiency for the four b-tag cut is different in the signal and in the irreducible background because the two additional b-jets have different p_T and $|\eta|$ spectra. The preselection, at a level of four loose b-tags, removes a large fraction of the signal, but also reduces the backgrounds to a level where they can be handled more easily; for example 98% of the $t\bar{t}$X background is removed. It also ensures that the selections are a subset of the requirements placed at generator level. The selected sample has approximately a 0.6% signal component, with a little more than 8% of the background being irreducible $t\bar{t}b\bar{b}$. Further tightening the b-tagging requirement to four jets with weights of at least 5.5 reduces the samples with four b quarks by a further factor of four while removing 90% of the remaining $t\bar{t}$X background.

Table 3: Cross-sections after each preselection cut for signal and background. The last row shows the further effect of tightening the b-tag requirements to the level of the final selection in the cut based and pairing likelihood analyses. In the last column the contribution of $t\bar{t}b\bar{b}$ has been removed. The errors are statistical only.

Preselection cut	$t\bar{t}H$(fb)	$t\bar{t}b\bar{b}$(EW) (fb)	$t\bar{t}b\bar{b}$(QCD) (fb)	$t\bar{t}$X (fb)
lepton	$57. \pm 0.2$	141 ± 1.0	1356 ± 6	63710 ± 99
+ ≥ 6 jets	36 ± 0.2	77 ± 0.9	665 ± 4	26214 ± 64
+ ≥ 4 loose b-tags	16.2 ± 0.2	23 ± 0.7	198 ± 3	2589 ± 25
+ ≥ 4 tight b-tags	3.8 ± 0.06	4.2 ± 0.2	30 ± 0.8	51 ± 2

6 Reconstruction of the hadronically decaying W bosons

Reconstructing the W boson four-momenta is necessary to reconstruct the top quarks. The reconstruction of the hadronically decaying W boson is done in different ways by the three analyses.

For the cut-based and pairing likelihood analyses (see Sections 8 and 9), the highest four b-tagged jets (with b-jet weight ≥ 5.5) are excluded from the hadronically decaying W reconstruction, however all other jets are paired to form W candidates. Figure 7 shows the mass distribution and multiplicity of all W candidates for the cut-based analysis. Only candidates within 25 GeV of the true W mass are kept. Even with these cuts the hadronically decaying W candidate multiplicity is still very high. All jets used to form these W candidates are calibrated with the light-jet calibration. The likelihood analyses do not require an explicit cut upon this mass as the definition of the likelihood imposes it automatically, but the constrained fit likelihood (Section 10) imposes a requirement that the mass be between 30 and 150 GeV to reduce the number of combinations to be evaluated.

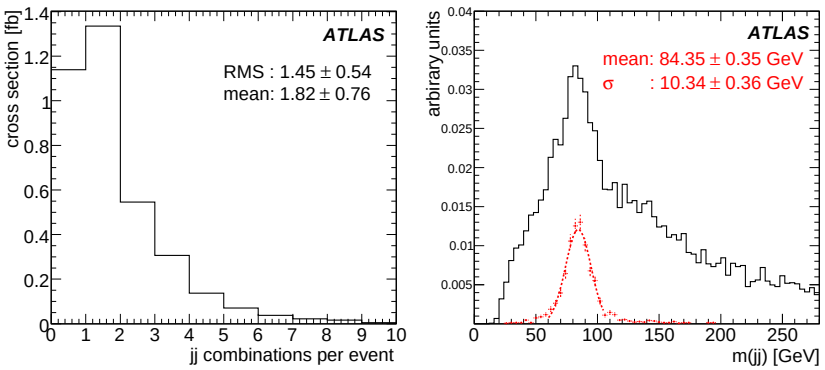


Figure 7: Cross-sections of hadronically decaying W combinations per event giving candidates within 25 GeV of the true W mass in the signal sample. (left). Right: Invariant mass spectrum for hadronically decaying W candidates normalized to unity. The dotted line shows combinations where the jets from the W are correctly matched.

7 Reconstruction of the leptonically decaying W bosons

When reconstructing the leptonically decaying W, we use the lepton four-momentum as measured in the detector. The neutrino, transverse momentum can be inferred by measuring the imbalance of the transverse energy in the event. This measured quantity is referred to as missing transverse energy.

7.1 Neutrino p_z estimation

Once the missing transverse energy is identified with $p_{T\nu}$, the invariant mass of the sum of the lepton and neutrino four-momenta can be constrained to the W boson mass [21]. Because of the limited measurement resolution on the transverse missing energy, for a significant fraction of events the quadratic constraint equation does not have a real solution. In this case the "$\Delta = 0$ approximation" can be made by dropping the imaginary part of those solutions with complex roots. Another method (the "collinear approximation") assumes that the W boson decay products are produced preferentially in the same direction (due to the large top quark mass boosting the W boson). For the collinear approximation, one can assume that $p_{z\ell} = p_{z\nu}$.

Considering $t\bar{t}H$ events where one lepton is reconstructed, 72% of the time $p_{z\nu}$ has real solutions. In this case, both solutions are carried forward into the analyses, and the best performing final state solution is used. For these events the $p_{z\nu}$ resolution is 19.5 GeV. In the other 28% of cases where there are no real $p_{z\nu}$ solutions, the $\Delta = 0$ approximation is used and the $p_{z\nu}$ resolution is 40 GeV. This performs better than the collinear approximation where the $p_{z\nu}$ resolution is 54 GeV. The quality of the W-boson reconstruction can be seen in Fig. 8. For the events where there is no real solution for $p_{z\nu}$, the direction and the mass of the W boson is better represented by the $\Delta = 0$ approximation.

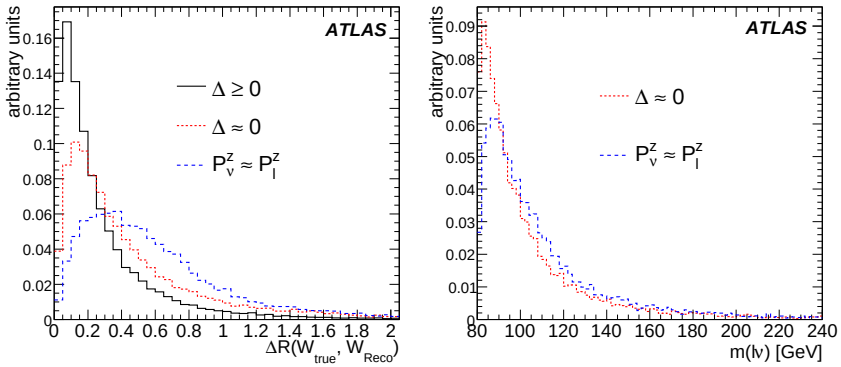


Figure 8: Distributions of ΔR between the true and the reconstructed W boson (left) and of the reconstructed leptonically decaying W mass (right) for events where a solution for p_z is found (solid black line) and events where an approximation is used (dotted red for $\Delta = 0$, dashed blue for the collinear approximation). All distributions are normalized to unity.

Since the mass constraint is lost when using the $\Delta = 0$ approximation (the same would be true for the collinear approximation) there is an actual cut for the reconstructed W boson mass. The cut-based and pairing likelihood analyses only consider W candidates having a mass less than 140 GeV, while the constrained fit analysis does not make an explicit requirement on this but the poor χ^2 will remove extreme cases.

8 Cut-based analysis

This section describes the algorithms used to reconstruct the $t\bar{t}$ system and the Higgs boson. The b-jets are associated with the leptonically and hadronically decaying W boson candidates, to build a list of top quark candidates. The combination of b-jets resulting in the best reconstruction for the top quark candidates is taken as the final choice. It is important to note that the b-jets themselves satisfy the cut on b-tag weight ≥ 5.5 and that if there are more than four of these jets, then the four with the highest weight are treated as b-jets. Events where there is no combination giving a satisfactory top quark mass reconstruction are discarded. The two remaining b-jets are used to form the Higgs boson candidate.

8.1 Top-antitop quark system and combinatorial background

In each event, top quarks are reconstructed by pairing two *b*-jets with the *W* boson candidates in the way which minimizes the χ^2 expressed as:

$$\chi^2 = \left(\frac{m_{jjb} - m_{top}}{\sigma_{m_{jjb}}}\right)^2 + \left(\frac{m_{lvb} - m_{top}}{\sigma_{m_{lvb}}}\right)^2, \quad (1)$$

where $\sigma_{m_{jjb}}$ and $\sigma_{m_{lvb}}$ are the reconstructed mass resolutions estimated in simulated signal events and are 13 and 19 GeV respectively. Only combinations fulfilling $|m_{jjb} - m_{top}| < 25$ GeV and $|m_{lvb} - m_{top}| < 25$ GeV are considered for the χ^2 calculation. The top quark mass distributions for the chosen combination in the signal sample is shown in Fig. 9. The two remaining *b*-tagged jets are used to form the Higgs boson candidates. The mass distribution for all Higgs boson candidates in the signal sample is shown in Fig. 10, and in the same plot the signal and physics background cross-sections are adjacent. Here the difficulty of the analysis is clearly shown, requiring dedicated studies to measure the background normalization and its shape in data.

As a final cut, to discriminate against $t\bar{t}$ events where no Higgs boson is produced, only events in a mass window of 30 GeV from the nominal Higgs boson mass are used for the final estimation of the cut-based analysis significance.

The effect of the final selection for the cut-based analysis on signal and background samples is shown in Table 4. The selections have reduced the signal by a factor of sixteen from the preselection, but the signal to background has increased from 0.006 to 0.11. The irreducible $t\bar{t}b\bar{b}$ background is 46% of the total.

Table 4: Accepted cross-section after each successive mass-window selection cut for signal and background in the cut-based analysis. In the last column the contribution of $t\bar{t}b\bar{b}$ is removed. Errors are statistical only.

cut	$t\bar{t}H$(fb)	$t\bar{t}b\bar{b}$(EW) (fb)	$t\bar{t}b\bar{b}$(QCD) (fb)	$t\bar{t}$X (fb)
$W_{\text{had}} + W_{\text{lep}}$	2.49 ± 0.05	2.9 ± 0.2	18.2 ± 0.7	22.5 ± 1.9
+ $t\bar{t}$+Higgs	2.04 ± 0.05	2.2 ± 0.2	14.7 ± 0.6	14.3 ± 1.5
+ Higgs boson mass window	1.00 ± 0.03	0.52 ± 0.07	3.6 ± 0.3	4.9 ± 0.9

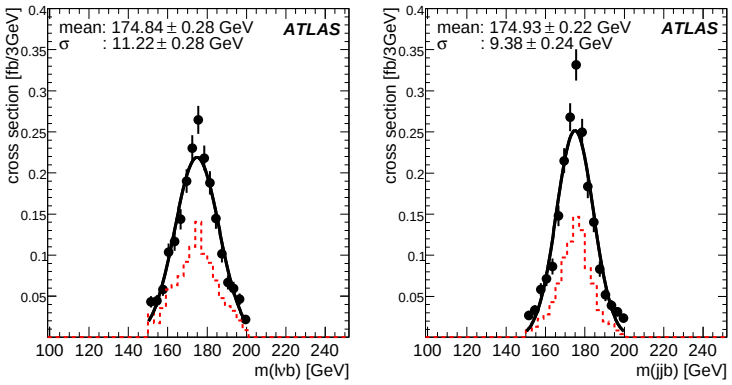


Figure 9: Reconstructed invariant mass spectrum for selected leptonic (left) and hadronic (right) top quark candidates in the signal sample. The dotted red line indicates the candidates formed by assigning the correct *b*-jet to the top quark being considered. All distributions are given in cross-section.

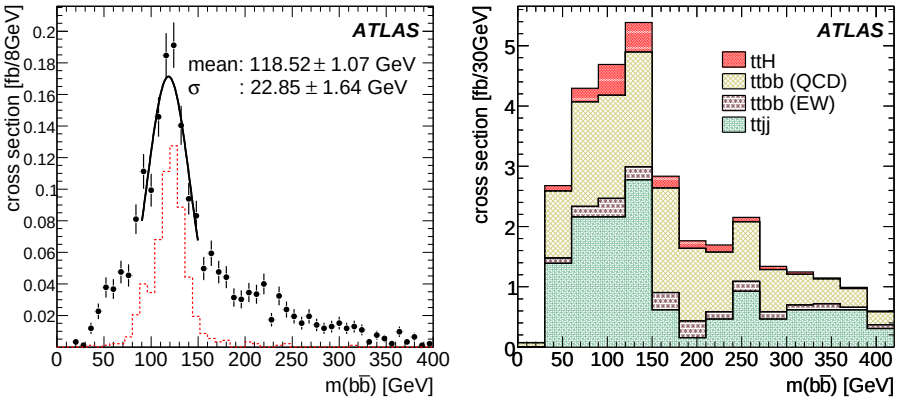


Figure 10: Left: Reconstructed invariant mass spectrum for Higgs boson candidates in the signal sample. The dotted red line indicates the candidates formed by assigning the correct *b*-jets. Right: Reconstructed invariant mass spectrum for signal and backgrounds after the cut-based selection. All distributions are given in cross-section.

9 Pairing likelihood analysis

In the previous Section we used a cut-based approach to identify the top quark decay products. A straightforward improvement to such an approach is to use several discriminating topological distributions combined together in order to build a pairing likelihood. As a first step the analysis considers only top quark properties as likelihood templates. Even though Higgs boson properties could help associating *b*-jets, if used, those could lead to bias in the background distributions. A correct combination is obtained when the objects used for the reconstructed variables match the Monte Carlo partons, regardless of whether other objects are correctly associated. On the other hand all the wrongly reconstructed objects are used to form the wrong combination templates. The variables used, shown in Fig. 11, are:

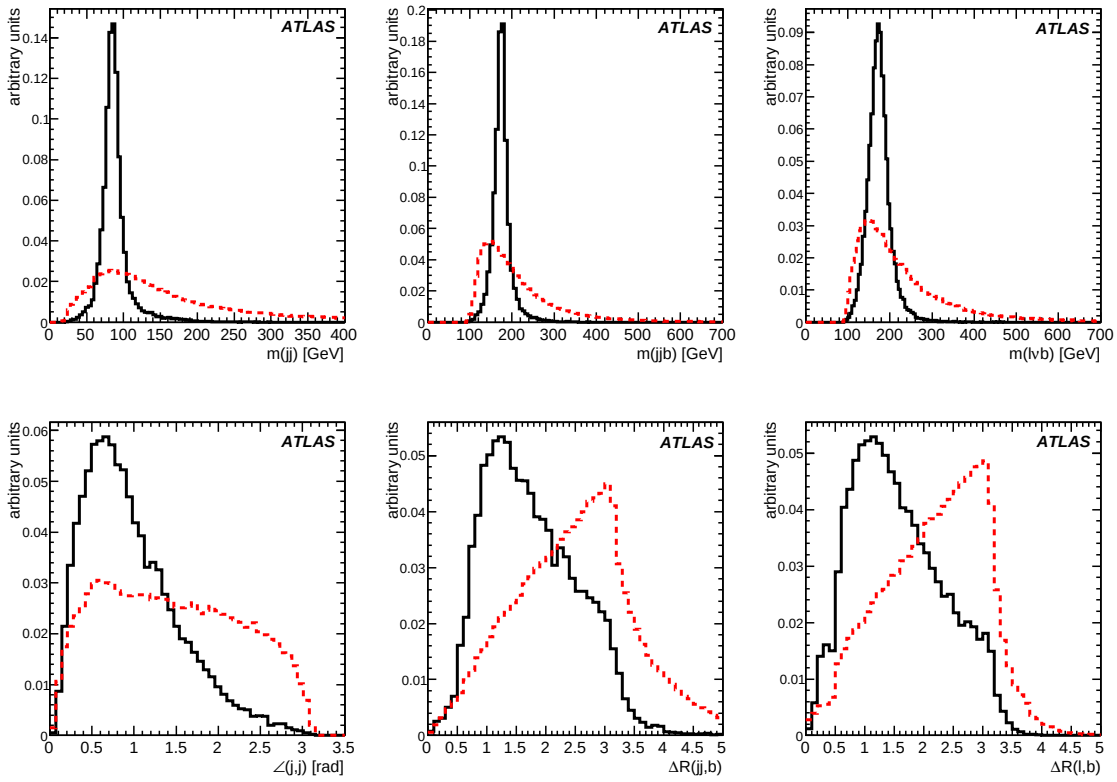


Figure 11: Pairing likelihood templates for top quark topological distributions, derived from the $t\bar{t}H$ signal sample. Solid lines represent the correct combination while the dotted lines show the combinatorial background in the signal itself. See text for a description of the variables.

- m_{jj}: The invariant mass of the light jets from the hadronic W decay.
- m_{jjb}: The invariant mass of the hadronic top decay products.
- m_{lvb}: The invariant mass of the leptonic top decay products.
- $\angle(j,j)$: The angle between the light jets from the hadronic W decay.
- $\Delta R(jj,b)$: The distance in R between the hadronic W and b jet from the hadronic top decay.
- $\Delta R(l,b)$: The distance in R between the lepton and the b jet from the leptonic top decay.

The output of the pairing likelihood for the correct and wrong b-jet combinations is shown in Fig. 12. As shown in this plot, even though the correct distributions are peaked at 1, the wrong combinations still have a large probability of being selected. The only combination used is the one which maximizes the likelihood output. In order to avoid the presence of a large combinatorial contribution a cut on the likelihood output of 0.9 is used to select well-reconstructed events. After this cut, b-jets are associated to reconstruct the Higgs boson, as shown in Fig. 12, and the two top quarks, shown in Fig. 13. The invariant mass distribution for the selected Higgs boson candidates for signal and backgrounds is shown in Fig. 14.

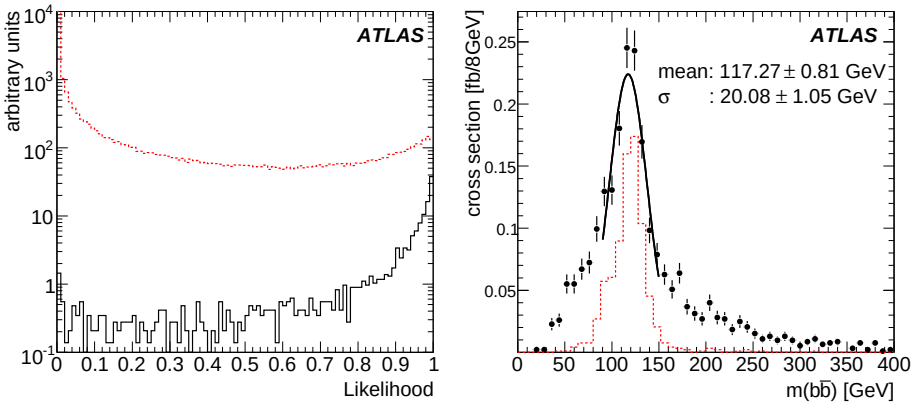


Figure 12: Left hand side: combinatorial likelihood output for $t\bar{t}H$ events. Black solid (red dotted) histogram indicates the correct (wrong) combinations. Right hand side: invariant mass for the Higgs boson candidates reconstructed using the maximum likelihood configuration, after applying a cut on the likelihood. Dotted histogram indicates the correct combinations. The differential cross-section is shown in fb.

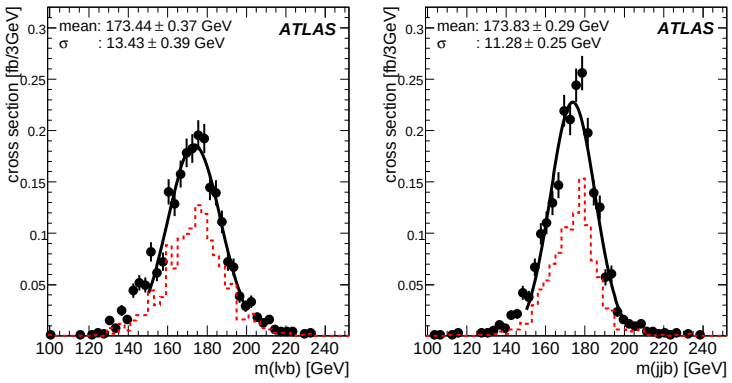


Figure 13: On the left (right) hand side is shown the leptonic (hadronic) top quark candidates reconstructed invariant mass using the maximum likelihood configuration, after applying a cut on the likelihood output. The dotted histogram indicates the correct b quark jet for the top quark being considered. The differential cross-section is shown in fb.

A final cut on the reconstructed Higgs boson mass, requiring it to be within 30 GeV of the Higgs boson nominal mass is applied. The event yield for the whole analysis using the pairing likelihood is shown in Table 5. This analysis reduces the signal by a factor thirteen, and produces a sample which has a signal to background ratio of 0.1. The irreducible $t\bar{t}b\bar{b}$ background increases to 45% of the total.

Table 5: Cross-sections after each selection cut for signal and backgrounds for the pairing likelihood analysis. In the last column the contribution of $t\bar{t}b\bar{b}$ has been removed. Errors are statistical only.

applied cuts	$t\bar{t}H$(fb)	$t\bar{t}b\bar{b}$(EW) (fb)	$t\bar{t}b\bar{b}$(QCD) (fb)	$t\bar{t}$X (fb)
Leptonic W	3.6 ± 0.06	4.1 ± 0.2	29 ± 0.8	48 ± 2.7
+ Best likelihood > 0.9	2.3 ± 0.05	2.5 ± 0.2	16 ± 0.6	19 ± 1.7
+ Higgs boson mass window	1.2 ± 0.04	0.68 ± 0.08	4.6 ± 0.3	6.5 ± 1.0

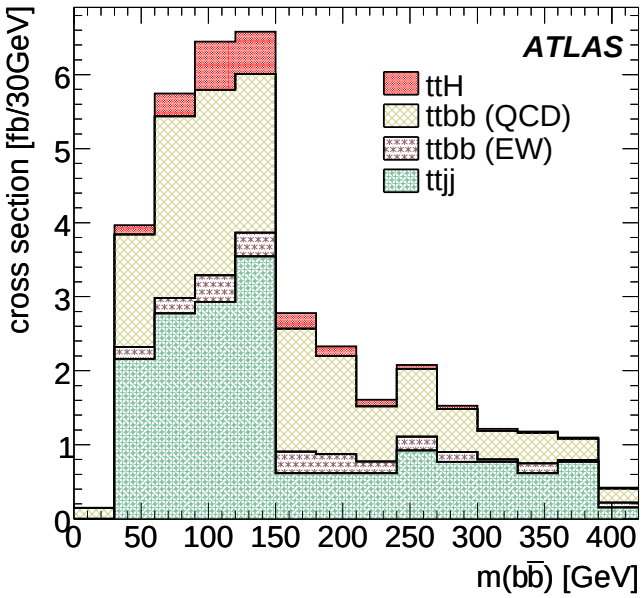


Figure 14: Reconstructed invariant mass spectrum for Higgs boson candidates for signal and backgrounds after pairing likelihood selection. The differential cross-section is shown in fb.

10 Constrained fit analysis

An alternative analysis uses a mass-constrained fit to the measured missing energy and jet and lepton four-momenta to help with the jet combinatorics. There are six quarks produced from the top quark and Higgs boson decays, and these are matched to the reconstructed jets. The χ^2 from this fit is used in a likelihood technique together with kinematic variables, *b*-tagging and jet charge. Then, all jet combinations passing loose criteria are tested and the one with the best likelihood is chosen. The analysis starts from the preselection as described in Section 5. The signal events are then separated from background in a second likelihood step.

10.1 Mass-Constrained fit technique

The fit varies a scale factor for the four-momenta of the jets and the z component of the neutrino momentum. Adjustments to the jet momenta and masses through the scale parameters, f^i, produce accompanying changes in the missing energy and hence in the parameters used in the reconstruction of the leptonically decaying W as the transverse components of its neutrino are taken to be the missing energy. The longitudinal component of the momentum of decaying W boson's neutrino $p_{z\nu}$ is the last fit parameter. The parameters are constrained by the estimated jet errors and by the masses of the top quarks and W bosons. These later are included as approximate Gaussian χ^2 contributions calculated using the masses inferred from the current parameters as indicated in the following equation:

$$\chi^2 = \sum_{i=1}^{6} \left(\frac{f^i_{jet} - 1}{\sigma^i_{jet} / P^{i,initial}_{jet}} \right)^2 + \frac{(m^{lep}_W - 80.425)^2}{\sigma^2_W} + \frac{(m^{lep}_t - 175)^2}{\sigma^2_t} \tag{2}$$

where the W and top quark widths σ_W and σ_t are 2.1 and 1.5 GeV respectively.

To simplify the fit, the hadronic top quark and W are forced to be exactly on mass shell. The scale factor of the higher p_T jet from the W is externally varied, while the other two scale factors are calculated to give the correct masses. The momenta of all six jets are varied, but these three are linked. There are therefore five free parameters and not seven. This implies that these two particles are fixed to their nominal masses, while the leptonic top quark and W are given widths.

The calculation of the momentum of the neutrino from the leptonic W decay normally has two solutions, as discussed in Section 7. Both of these are used as starting points for the fit, to ensure that it does not find only one local minimum, and the fit with the larger χ^2 is discarded. If the W neutrino solutions had complex roots then the real part of these is used as an initial value.

The jet momenta are calibrated as discussed in Section 5.3.4, and the following errors are used in the fit:

$$\sigma_{P_{light}} / P_{light} = 0.988/\sqrt{p_T} \oplus 0.035 \tag{3}$$

$$\sigma_{P_b} / P_b = 0.888/\sqrt{p_T} \oplus 0.125 \tag{4}$$

where the momenta are measured in GeV. This form comes from comparing reconstructed jet p_T with simulated initial quark p_T; in other words it includes not only detector effects but also fragmentation. Both light and *b*-jet momentum errors are treated as Gaussian distributed; for *b* quark jets in particular this is not a good description as the frequent presence of a neutrino gives tails to the measured energy response.

The fit adjusts the momenta of all the jets, including those from the Higgs boson, but the fitted Higgs boson mass is not used in the analysis, as it offers no improvement.

10.2 *b*-tag information

In order to use the *b*-tagger output, as described in Section 5.3.2, for a likelihood, the distributions of weights expected for *b*-jets and non *b*-jets in this environment are required. This is done using jets taken from the signal simulation, where only jets with a parent quark within 0.4 in ΔR and no second quark within 0.6 in ΔR are used. The ratio of smoothed weight distributions with *b*-jet over light flavored jets (u,d,s) is shown in Fig. 15.

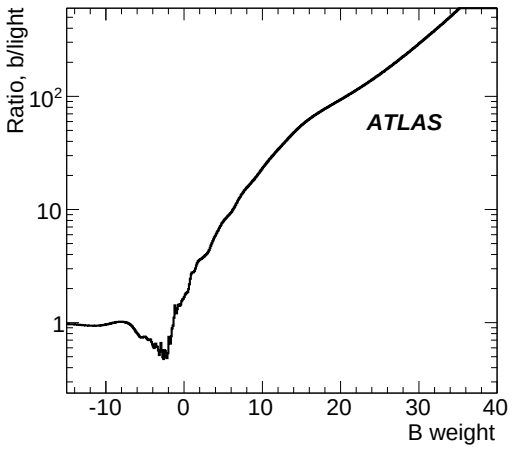


Figure 15: The *b*-tagging likelihood ratio extracted from signal simulation as the ratio of *b*-weight distributions for *b*-jets and light jets. The degree of smoothing reflects the statistical precision at each point.

In the analysis, four jets are taken to be from *b* quarks. For each jet we compute $\mathscr{L}_b^i$, the ratio of $b/light$ for jet *i* as shown in Fig. 15. In selecting combinations the sum $\Sigma_i \log_{10} \mathscr{L}_b^i$ is taken over the four jets. There is no requirement made on the jets from the *W* boson.

10.3 Jet charge

The assignment of jets to quarks can benefit from the jet charge measurement as we know the expected charges of the quarks involved. The jet charge is the momentum weighted sum of the charged particle charges within the jet, and it shows some correlation with the initial quark charge. The $\bar{b}(b)$ quark has only a charge of (-)1/3, and furthermore, after hadronisation there are oscillations which reduce the sensitivity, but there is some information.

The analysis requires exactly one high p_T lepton which has charge Q_l, and therefore we know the expected charge of both of the *b* quarks associated to the top quarks via the relationship $sign(Q_l) = sign(Q_{t_{lept}}) = -1 \times sign(Q_{t_{had}})$, and the sum of the charges of the jets from the hadronically decaying *W* boson $Q_{W_{had}} = -1 \times Q_l$. The measured values of these are then compared with the expectations using a likelihood. The sum of the Higgs boson jet charges is much less sensitive because the expected value is zero, but it is also used.

The jet charge plots in Fig. 16 are calculated using jets from the signal sample. The *W* plots are made using only true light jets which were tagged as light jets, and the *b* plots only from tagged *b*-jets associated with a *b* quark. This is to ensure that the jet charge is independent of *b*-tag information. The W wrong combinations distribution has spikes at integral values which generally involve at least one jet outside the tracker acceptance contributing a charge of zero. These are less frequent in the correct combinations which tend to be central. The other distributions reflect *b*-tagged jets which must therefore have charged

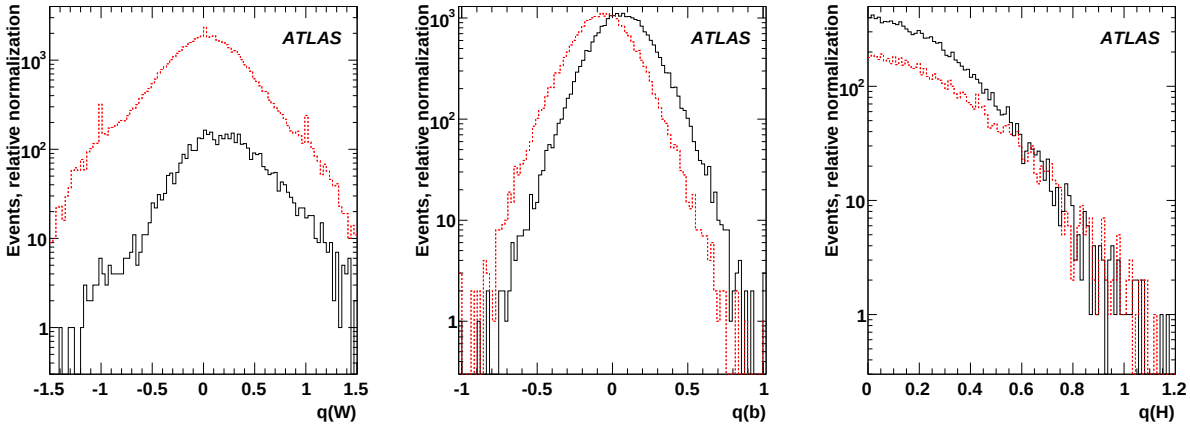


Figure 16: Left: The jet charge of W boson candidates, based on the sum of the jet-charges of the two jets signed by the high-p_T lepton. Center: Jet charges of individual b-jets from top quarks; signed so that those where quark charge agreeing with the lepton charge from W decay are positive. Right: The magnitude of the sum of the jet charges of the jets assigned to the Higgs boson. Wrong b-jet combinations have a somewhat flatter distribution. Correct combinations are solid (black) and wrong combinations are dashed (red).

particle tracks. It is assumed that the jet charge information can be calibrated from the plentiful top quark pair events. The normalizations are arbitrary, as they offset every combination equally. The jet-charge is used as $\mathscr{L}_{jet-charge} = \mathscr{L}_{q(b)}^{hadronic\ top} \times \mathscr{L}_{q(b)}^{leptonic\ top} \times \mathscr{L}_{q(W)} \times \mathscr{L}_{q(H)}$.

10.4 Likelihood analysis for jet assignment

All possible assignments of jets to quarks are evaluated in turn. Those combinations which fail a loose quality requirement are discarded. This quality requirement is that:

- Both of the jets assigned to the Higgs boson must have a b-likelihood greater than zero. From Fig. 15 that corresponds to a cut of about -2 on the b-tagging variable.
- Selections on the mass of the W and top quark which decay to jets. The W mass calculated from the jets without fitting must lie between 30 and 150 GeV, and the top quark mass between 100 and 250 GeV. Note that the jet energy correction factor applied depends upon whether or not the jet is considered to originate from a b quark in this hypothesis.
- Total b-likelihood greater than 8. This is the sum of the log-likelihoods of the four jets which are assigned to b-jets; this roughly corresponds to the mean b-weight of the jets being 4 or greater.

In the preselected signal sample there are a mean of 5811 combinations to be tested per event, but the above quality requirement reduces this to 233; a considerable saving in time. 90% of the preselected signal events have at least one combination passing the above requirements.

Events which pass these selections are processed by the $t\bar{t}H$ fitting code. Correct or wrong combinations are then used to define a likelihood ratio.

The elements of that likelihood ratio are as follows:

- The $\log_{10}$ of the χ^2 from the fit.

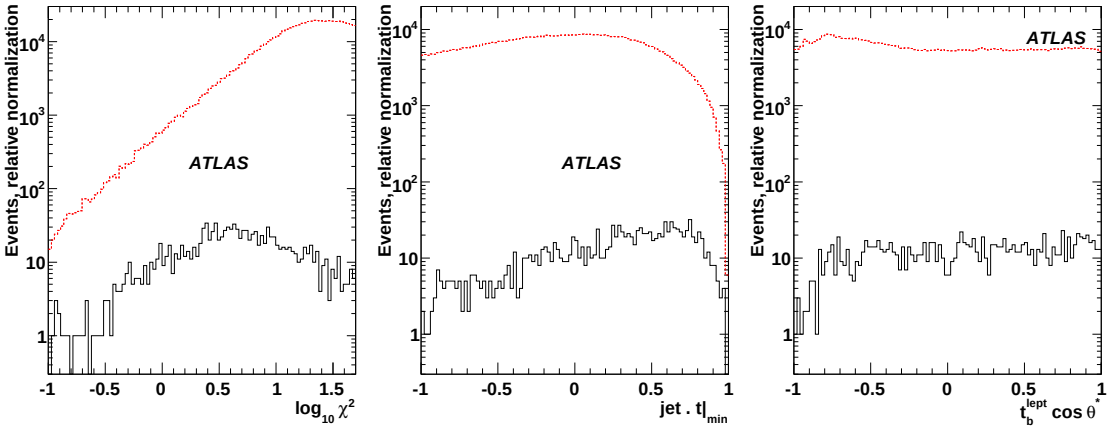


Figure 17: Left: The χ^2 of the fit. Center: jet $\cdot$ top$|_{\rm min}$. The cosine of the minimum angle between top quarks and their daughter jets. Correct combinations have more collimated tops. Right: The $t_b^{\rm lept}\cos\theta^*$. The decay angle in the rest frame of the leptonically decaying top quark of the b-jet relative to the top quark direction in the lab frame. Correct combinations are solid (black) and wrong combinations are dashed (red).

- jet $\cdot$ top$|_{\rm min}$: The minimum cosine of the angle between the top quarks and any of their four jets in the $t\bar{t}H$ center of mass frame.
- $t_b^{\rm lept}\cos\theta^*$: The decay angle in the rest frame of the leptonically decaying top quark of the b-jet relative to the top quark direction in the lab frame.
- $t_b^{\rm lept}\Delta$R: the distance in ΔR space between the lepton and the b quark assigned to the same top quark.
- $|\eta|^{\rm max}$: The maximum $|\eta|$ of the considered jets. Jets from the $t\bar{t}H$ system tend to be more central than those from the underlying event.
- $\cos\theta^*_{\rm H-jet}$: Measured in the Higgs boson rest frame, this is the cosine of the angle between the higher p_T of the two jets from the Higgs boson and the boost applied to shift from the lab frame to the Higgs boson rest frame.
- m_t: The hadronic top quark mass before the fit is performed.

The variables are displayed in Figs. 17 and 18 for correct and wrong combinations. In this case 'correct' implies that all six quarks are correctly assigned. Fig. 19 shows how each variable would perform if used individually to separate correct and incorrect pairings. For each variable combinations are selected by the likelihood ratio found using that variable alone. The Fig. shows the fraction of wrong pairings which would be accepted as a function of the fraction of correct ones. The fit χ^2 is the most powerful single variable over much of Fig. 19, but the masses of the hadronic top quark and W work well at high efficiency while jet $\cdot$ top$|_{\rm min}$ is also rather powerful. Clearly the variables have correlations, and these are taken into account by evaluating the likelihood in a 3D space defined by the three variables under study. This explicitly includes the correlations, but is limited by the simulation statistics required to populate the space. The combination of χ^2, jet $\cdot$ top$|_{\rm min}$ and $t_b^{\rm lept}\cos\theta^*$ was adopted as the most powerful set of three variables found.

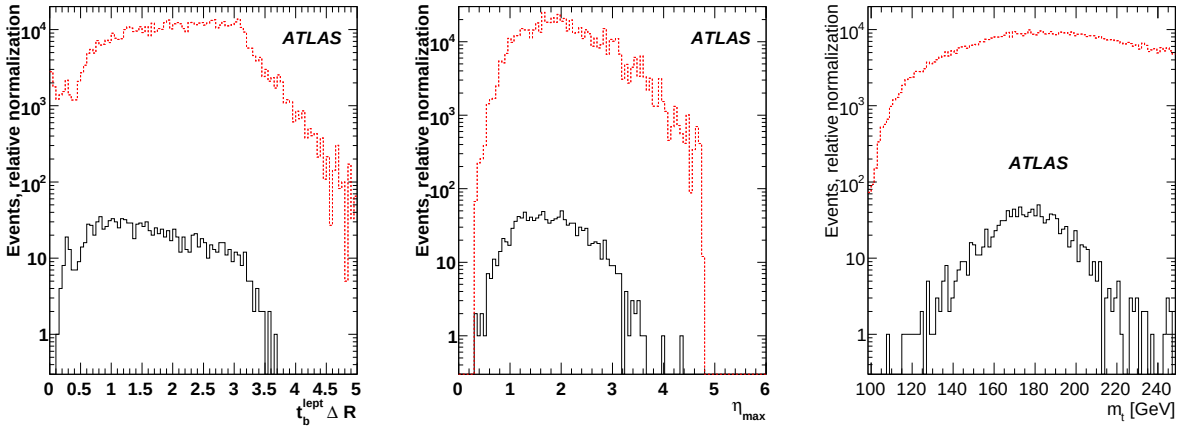


Figure 18: Left: The distribution of distances, ΔR, between the lepton and the b quark from the leptonic top quark. Center: The maximum $|\eta|$ of any jet in the combination being tested. Correct combinations are more central. Right: The reconstructed mass of the hadronically decaying top quark, before any fit is performed. Correct combinations are solid (black) and wrong combinations are dashed (red).

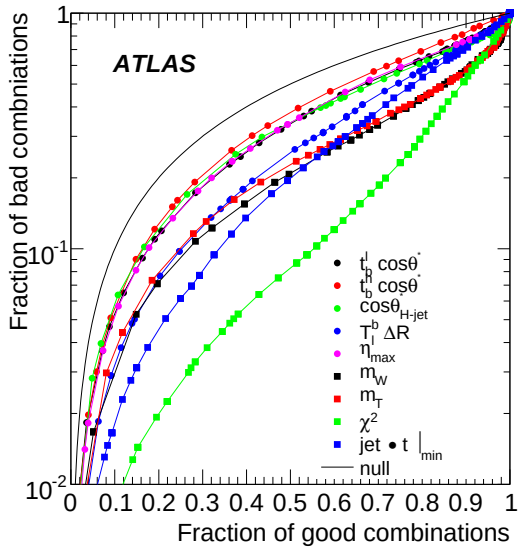


Figure 19: The performance of a range of possible variables if they are used individually to find the correct quark-jet pairing in a $t\bar{t}H$ event. For a given efficiency for selecting the correct pairing (x axis), what fraction of the incorrect pairings will also be chosen (y axis). The line labelled null shows the effect of selecting combinations at random.

The remaining variables were tested to see whether their addition as uncorrelated likelihood contributions made a significant improvement, and those that did were included. The final likelihood is defined as follows:

$$\mathscr{L}_{pairing} = \log \mathscr{L}^{3D}_{\log\chi^2,\mathrm{jet}\cdot\mathrm{top}|_{\min},\mathrm{t_b^{lept}}\cos\theta^*} + \log \mathscr{L}_{\mathrm{t_b^{lept}}\Delta\mathrm{R}} + \log \mathscr{L}_{|\eta|^{\max}} + \log \mathscr{L}_{\cos\theta^*_{\mathrm{H-jet}}} + \log \mathscr{L}_{m_t} + \log \mathscr{L}_{b-tag} + \log \mathscr{L}_{jet-charge} \quad (5)$$

The combination which produces the largest likelihood for each event is adopted. The quality of the chosen combination is examined in Section 11.

10.5 Signal and background separation

The separation of signal from background is again done using the likelihood technique. There are two rather different backgrounds considered: the $t\bar{t}j\bar{j}$ component for which b-tagging is the primary tool and the 'irreducible' $t\bar{t}b\bar{b}$ background which differs from the signal only in kinematic ways, which can be exploited to give some separation. The variables used to separate signal and background are:

- $\mathscr{L}_{pairing}$: From combinatorics.
- $\Sigma_i \log_{10} \mathscr{L}^i_b$: Sum of the log-likelihoods of the four jets used as b's in the combination chosen.
- Σ^H_{b-tag}: The sum of the b-tagging weight of the two jets from the Higgs boson.
- $\Delta\eta(\mathrm{H},\mathrm{top})^{\min}$: The difference in η between the Higgs boson and the closer top quark.
- $\cos(tH)^{max}$: The higher of the two angles between top quarks and Higgs boson in the center of mass of the $t\bar{t}H$ system.
- H_p in C.o.M: The Higgs boson momentum in the center of mass frame of the $t\bar{t}H$ system.
- $\cos\theta^*_{\mathrm{H-jet}}$: As defined in Section 10.4.

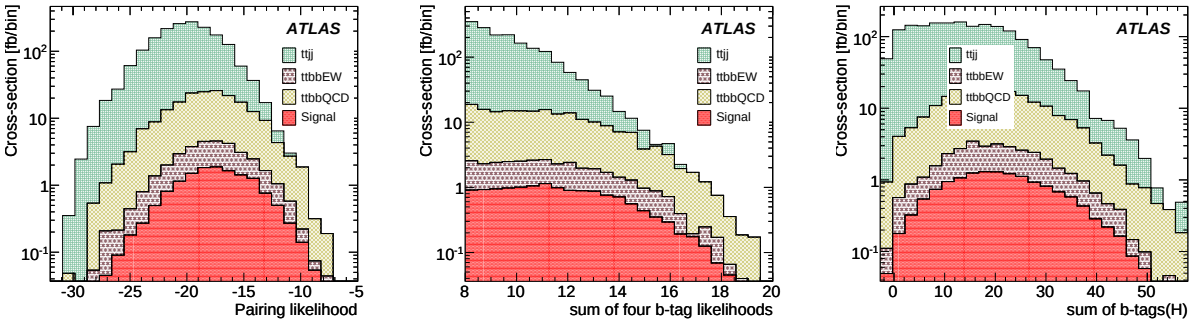


Figure 20: Left: Distributions for signal and backgrounds of the likelihood used to find the combinatorics solution. Center: The sum of the b-tag likelihoods of the four jets used as b's in the chosen combination. There was cut at 8 in the combinatorics preselection. Right: The sum of the b-tags of the jets associated to the Higgs boson. This variable removes $t\bar{t}j\bar{j}$ more than $t\bar{t}b\bar{b}$. Histograms are filled for every event where there is a successful fit. In all histograms, the sum of the individual histograms is shown. They are stacked to indicate relative contributions.

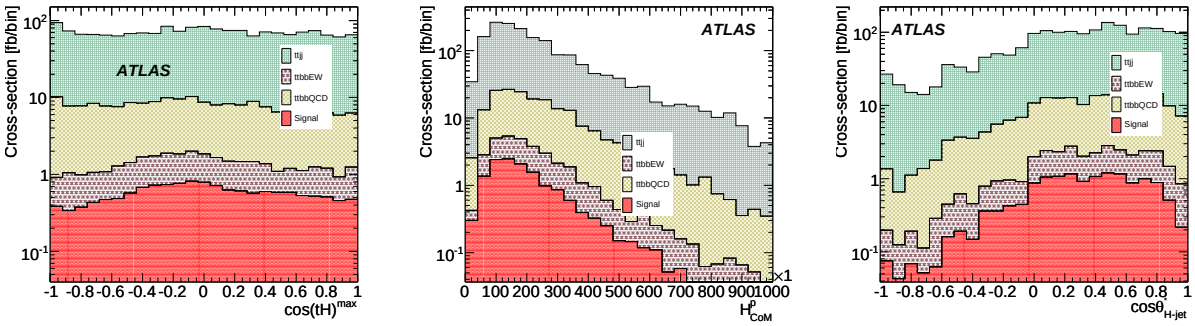


Figure 21: Left: The maximum cosine of the angle between either top quark and the Higgs boson when boosted into the $t\bar{t}H$ center of mass frame. Center: The momentum of the Higgs boson candidate in the center of mass frame. Note that the true Higgs bosons have a lower momentum than the background. Right: The $\cos\theta^*$ of the higher p_T of the jets from the Higgs boson. This tends to be more central for signal events. In all histograms, the sum of the individual histograms is shown. They are stacked to indicate relative contributions.

Figures 20 and 21 show the data used to produce the likelihood ratio for most of the variables. The backgrounds fall into two basic classes - the $t\bar{t}j\bar{j}$ and the $t\bar{t}b\bar{b}$, and to deal with these two 3D likelihoods are defined. The first includes the three quantities which contain b-tagging information: $\mathscr{L}_{pairing}, \Sigma_i \log_{10}\mathscr{L}_b^i, \Sigma_{b-tag}^H$. These are all powerful but are highly correlated and therefore benefit from a correct treatment of those correlations. The second is $\cos(tH)^{max}, H_p, \cos\theta^*_{\mathrm{H-jet}}$, which carries discrimination based on event kinematics. It too has important correlations. These likelihoods are combined as if independent, with one further likelihood, derived from $\Delta\eta(\mathrm{H,top})^{\mathrm{min}}$, added as well. All the likelihoods have been smoothed so that the expected fluctuations are below 10%, and there is therefore little over-training, as separate test and training samples are maintained.

The distributions used to define the likelihood are constructed using all events for which a constrained fit was made, and all the components are normalized to the cross-sections at that stage. The signal separation likelihood is:

$$\mathscr{L}_{s/b} = 1/3\left(\log\mathscr{L}^{3D}_{\mathscr{L}_{pairing},\Sigma_i\log_{10}\mathscr{L}_b^i,\Sigma_{b-tag}^h} + \log\mathscr{L}_{\Delta\eta(\mathrm{H,top})^{\mathrm{min}}} + \log\mathscr{L}^{3D}_{\cos(tH)^{max},H_p,\cos\theta^*_{H-jet}}\right) \quad (6)$$

The factor of 3 makes this an average, rather than a sum, and is there purely for convenience. Note that there is nothing in this definition to prefer correctly paired signal events.

Figure 22 shows the distribution of the final likelihood. It is generally dominated by $t\bar{t}j\bar{j}$ events, but at the largest likelihood values the $t\bar{t}b\bar{b}$ and signal events are more prevalent. It can be seen that any $t\bar{t}H$ analysis will be selecting a tail of the signal, and controlling this will be important. The signal to background ratio, within the mass window, rises to about 25%, and any rise above that is in a region affected by lack of simulation statistics.

The final choice of working point will depend upon the details of the systematic error evaluation. The tighter the selection on the likelihood the higher the signal to background ratio but the smaller the samples in data and simulation; the latter is an important consideration.

The maximum significance which might be expected in a measurement (evaluated as $s/\sqrt{b}$) ignoring all systematic uncertainties, is obtained by cutting at a log likelihood of -4.44. This would yield a significance of 2.78σ, but at low purity. Reducing the mass range by requiring that the candidate has a mass within the range 90 to 150 GeV does not appear to improve the results when systematic errors are not considered.

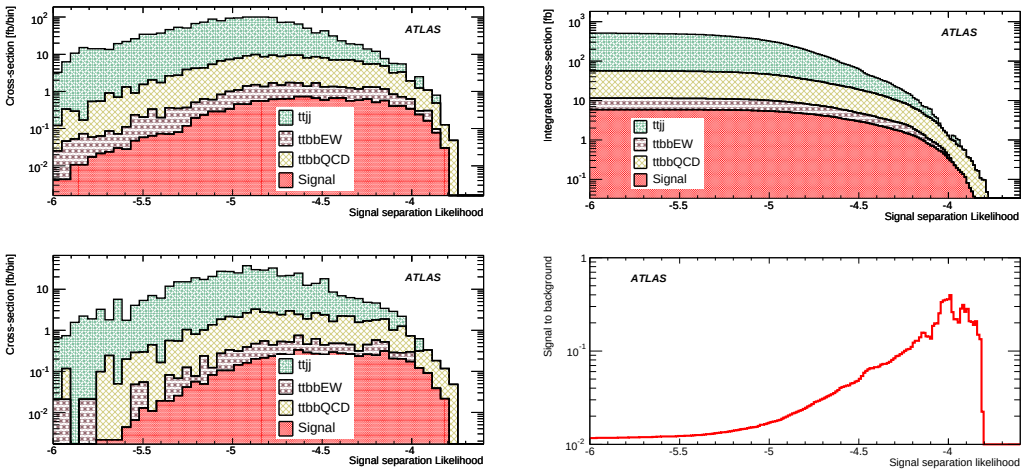


Figure 22: Left: The total signal separation likelihood. The top figure shows all events while the bottom shows only those within a Higgs boson candidate mass window of 90 to 150 GeV. Right: The integrated version of the lower left plot, so the total event rates passing any cut can be seen. The bottom half of this plot is the signal to background ratio implied.

If an arbitrary ten per cent error on the background level is assumed then the significance for this cut, evaluated as $s/\sqrt{b+(\delta b)^2}$, decreases to below 0.5σ, while the highest significance is around 1.8σ for a cut at -4.05. This is shown numerically in Table 6, where the expected event rates are shown for three different cut values. No final choice is really possible without complete evaluation of systematic errors, but -4.2 with the mass window cut applied does seem to be a plausible working point. The statistical significance for this selection is 2.18σ. At this point the signal is reduced by a factor of twelve from the preselection, but signal to background ratio has become 0.125 ± 0.01. The irreducible $t\bar{t}b\bar{b}$ background has increased to 50% of the total.

Table 6: The accepted cross-sections for signal and the main backgrounds at various stages of the analysis. The $t\bar{t}j\bar{j}$ cross-section suffers from limited statistics.

Selection	$t\bar{t}H$(fb)	$t\bar{t}b\bar{b}$(EW) (fb)	$t\bar{t}b\bar{b}$(QCD) (fb)	$t\bar{t}$X (fb)
Initial Sample	100	255	2371	109487
Pass preselection	16	23	198	2589
Fit quality requirements	14	20	165	1584
$\mathscr{L}_{s/b} >$ -4.40	4.9	5.1	35	58
$\mathscr{L}_{s/b} >$ -4.20	2.5	2.3	13.9	11.9
$\mathscr{L}_{s/b} >$ -4.10	1.4	0.96	7.11	4.5
Mass window 90 to 150 GeV.				
$\mathscr{L}_{s/b} >$ -4.40	2.3±0.07	1.4±0.17	10.8±0.7	22±3.1
$\mathscr{L}_{s/b} >$ -4.20	1.3±0.05	0.62±0.12	4.6±0.5	5.3±1.5
$\mathscr{L}_{s/b} >$ -4.10	0.71±0.04	0.23±0.07	2.5±0.35	2.2±1.0

The distribution of the masses of the candidates can be seen in Fig. 23, at a cut of $\mathscr{L}_{s/b} > -4.2$. The right hand side of Fig. 23 shows details of the mass distribution for signal only.

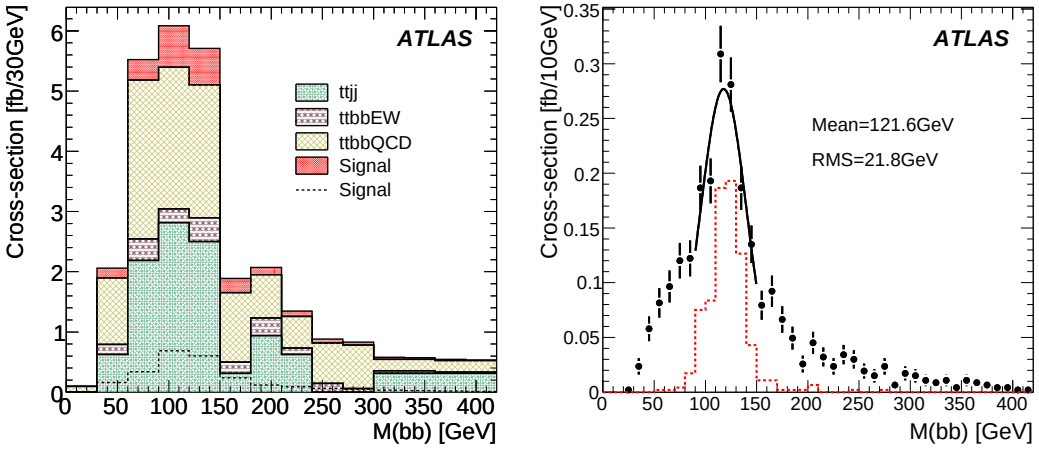


Figure 23: The mass distribution after cutting at -4.2 in $\mathscr{L}_{s/b}$. Left: All samples, showing the contributions stacked. The signal distribution is also shown separately at the bottom. Right: Signal only, the dashed (red) line equals events where the correct jets from the Higgs boson are selected.

11 Comparison between the three analysis techniques

The performance of the cut-based, pairing likelihood and the constrained fit analyses in terms of purity versus selection efficiency can be seen in Fig. 24. For the likelihood analyses, the different working points are obtained by varying the final cut on the likelihood discriminant. In the case of the cut-based analysis, the same variation is achieved by loosening or tightening the mass-window cuts on the hadronically decaying W and the reconstructed top quarks. For this section the efficiency is defined as the selection efficiency relative to the total events simulated. The purity is defined in terms of the correctness of the assignment of b-jets used to reconstruct the final objects. For instance, one has a *pure* hadronic top quark when the b-jet matches the true b parton from the top quark decay, regardless of whether the same happens for the hadronic W boson decay products.

Fig. 24 clearly shows the increase of performance when using more information (likelihood) than just the mass of the reconstructed particles (cut-based).

The chosen working points are indicated with solid markers on Fig. 24. Those points have not been optimized in terms of statistical significance, because of a lack of statistics for the $t\bar{t}$X background and because due consideration of the systematic errors should also influence the decision. However, the significance does not change much with the choice of the cut on the pairing likelihood output since this likelihood is not designed to discriminate signal from physics background events.

The ability of the three analyses to correctly identify objects in the event is compared in Table 7. The likelihood-based assignments perform noticeably better than the cut-based analysis. The signal efficiency and statistical significance are also improved.

12 Background Shapes

The success of this analysis relies on the accurate knowledge of the background level and shape. Monte Carlo predictions are affected by large systematic uncertainties, as the background rejection depends critically also on the jet flavor composition. For this reason it is mandatory to develop methods to measure background directly from real data.

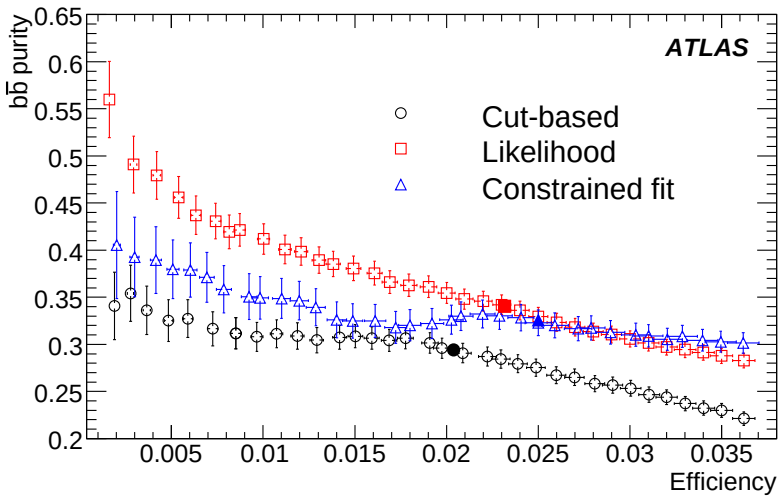


Figure 24: Comparison of the purity of reconstructed $b\bar{b}$ invariant mass before the final mass window cut versus selection efficiency for the cut-based, the pairing likelihood and the constrained fit analysis. The solid markers show the selected working points.

Table 7: A comparison of quality criteria for the three analyses at their working points. The mass window of 90 to 150 GeV is only applied for the last two rows.

	Cut Based	Pairing likelihood	Constrained fit
b jet from Hadronic top correct	44.4±1.1%	49.2±1.1%	51.0±1.5%
b jet from Leptonic top correct	50.5±1.2%	57.4±1.1%	56.2±1.5%
Higgs boson jets correctly chosen	29.4±1.0%	34.0±1.0%	32.0±1.4%
Four b quarks correct	23.3±1.0%	27.5±1.0%	27.1±1.3%
Higgs boson mass peak resolution, GeV	22.8±1.6	20.1±1.1	22.3±2.1
Signal Efficiency	2.04±0.05%	2.32±0.05%	2.49±0.07%
Signal to background	0.110±0.014	0.103±0.014	0.123±0.019
$s/\sqrt{b}$, 30fb^{-1}	1.82	1.95	2.18

One important result of the present study is that the Higgs boson candidate mass spectrum depends weakly upon the b-tagging working point. This is shown in Fig. 25, which reports the difference in $b\bar{b}$ invariant mass shape for the $t\bar{t}b\bar{b}$ and $t\bar{t}$+jets processes after applying the pairing likelihood analysis with the loose and tight b-tagging requirement as defined in Section 5.3.2

The complete determination of the background shape from data depends crucially on the relative contributions of the $t\bar{t}b\bar{b}$ and $t\bar{t}$+jets distributions, which in turn depends on the strength of the b-tagging cut applied. The $b\bar{b}$ invariant mass can be studied for a b-tag requirement, the "medium b-tag", between the loose and tight, such that the possible presence of signal can be still neglected. We choose a medium working point corresponding to a b-tagging weight cut of 3, such that the ratio of the contribution of $t\bar{t}b\bar{b}$ with respect to $t\bar{t}$+jets goes from 11% to 30%, with a signal contamination of less than 3%.

One strategy contemplated is to use the $t\bar{t}b\bar{b}/t\bar{t}$+jets fraction coming from the Monte Carlo prediction and the total number of events from the data to normalize the Monte Carlo at the loose working point where the signal level is less than 1%. Using the Monte Carlo jet flavor composition and the ratio of the b-, c- and light jet efficiencies ($\varepsilon^{medium}_{b,\,c,\,light}(p_T,\eta)/\varepsilon^{loose}_{b,\,c,\,light}(p_T,\eta)$) at the loose and medium working

points, it is then possible to predict the shape and normalization at the medium working point.

The data reduction when moving the *b*-tag quality from "loose" to "medium" is explained by the ratio of the *b*-tag efficiencies at these two working points applied to the $t\bar{t}b\bar{b}$ and $t\bar{t}$+jets data. With a 50 pb^{-1} data sample, the *b*-tagging efficiency of *b*- and *c*-jets will be known with an accuracy of 5% [19], while the rejection of light quark jets will be measured with a 10% uncertainty. We expect a significantly more accurate knowledge of the *b*-tagging performance with a data sample of approximately 30 fb^{-1}. This will allow the measurement of the background level of $t\bar{t}b\bar{b}$ and $t\bar{t}$+jets as a function of the $b\bar{b}$ invariant mass for the loose and medium *b*-tagging working points. These measurements can be used to verify and tune with data the background prediction given by the Monte Carlo simulation, which will be used to extrapolate the event yeld expectation of known processes when the *b*-tag quality is moved from "medium" to "tight". This extrapolation can be monitored, and eventually further corrected, by looking at the comparison with the measured data outside the mass window, where the number of signal events expected is small (about 4%). If necessary, this procedure can be extended by asking for three *b*-tagged jets to further constrain the background composition and its shape and absolute normalization, to achieve the 5% systematic uncertainty necessary for the analysis of this processes.

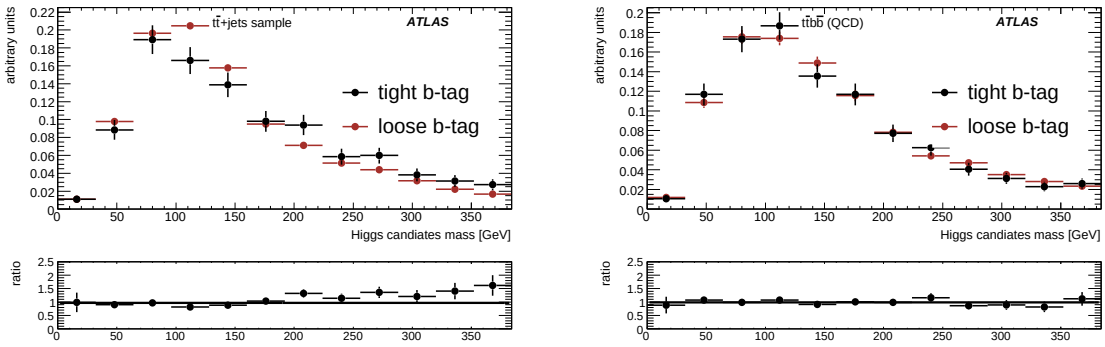


Figure 25: Ratio of the invariant mass spectrum for Higgs boson candidates after combinatorial likelihood analysis and using a loose and tight cut on the *b*-tag weight. Left hand side: $t\bar{t}$+jets, right hand side: $t\bar{t}b\bar{b}$. The signal region shows very consistent behaviour.

13 Systematic uncertainties

The evaluation of systematic uncertainties, especially in the background level, is of vital importance in this analysis. Unfortunately it has not yet been brought to a satisfactory level and a robust method to infer background shapes and normalization from data, vital for this channel, still needs to be developed. Following the estimation of systematic uncertainties due to the standard detector effects, Table 8 shows the various contributions for all three analyses. It is noticeable how important the jet uncertainties are for both signal and background. Indeed the knowledge of the jet energy and of the *b*-tagging performance have a crucial impact on the kinematic quantities used for the reconstruction of the $t\bar{t}$ system and for the correct identification of the *b*-jets used for the analysis. Large fluctuations on the background estimations arise due to the lack of statistics for the $t\bar{t}$X sample, giving rise to a relative statistical error up to 20%.

While the theoretical uncertainties for the signal and background normalization are quite large, their impact can be reduced by making direct measurements. This is certainly the case for the $t\bar{t}$ cross-section, where the theoretical uncertainties associated with the NLO+NLL calculation are around 12% [14] while with only 100 pb^{-1} of data, a direct measurement of the cross-section for the semileptonic final state using *b*-tagging could be performed with a much smaller error [22]. The $t\bar{t}b\bar{b}$ background is only calculated at LO, the cross-section calculation has a strong scale dependence, a factor 4 when changing

from $Q^2_{QCD} = \hat{s}$ to $Q^2_{QCD} =< p^2_T >$ [9]. Even though the signal cross-sections used for this work are LO, NLO calculations are already available with a theoretical uncertainty including errors coming from parton distribution functions of the order of 15-20% [23] (compared to the 100-200% uncertainty of the LO cross-section).

Table 8: Effect of the various systematic uncertainties on the signal and background efficiencies.

Source		Cut-based		Likelihood		Constrained fit	
		signal	background	signal	background	signal	background
Electron	energy scale	± 0.5%	± 2%	± 0.3%	± 3%	± 1%	± 3%
	resolution	± 0.5%	± 0.6%	± 0%	± 1%	± 0.2%	± 4%
	efficiency	± 0.2%	± 2%	± 0.2%	± 1%	± 0.5%	± 0.2%
Muon	energy scale	± 0.7%	± 3%	± 0.6%	± 0.2%	± 0.4 %	± 4%
	resolution	± 0.8%	± 0.6%	± 0.3%	± 0.4%	± 1%	± 3%
	efficiency	± 0.3%	± 0.1%	± 0.8%	± 0.1%	± 0.4%	± 0.1%
Jet	energy scale	± 9%	± 5%	± 9%	± 14%	± 9%	± 8%
	resolution	± 0.3%	± 7%	± 1%	± 5.5%	± 5%	± 14%
	b-tag	± 16%	± 20 %	± 18%	± 20%	± 16%	± 20%
	b mis-tag	± 0.8%	± 5%	± 1.1%	± 3%	± 3%	± 10%
summed in quadrature		± 18%	± 22%	± 20%	± 25%	± 19%	± 28%

13.1 Effect of pile-up on signal

The portion of the semi-leptonic $t\bar{t}H$ signal sample used here is simulated a second time, but with the anticipated effects of pile-up and cavern background included. It is important to stress that the same generated events are used as input for both pile-up and non pile-up samples. The pile-up actually applied to the events is that expected for running at instantaneous luminosity $\mathscr{L}$ of $10^{33} cm^{-1} s^{-1}$.

The effect of pile-up on the preselection of events (applicable to all three analyses) is shown in Table 9, and as can be seen, the effect of the trigger requirement is the most significant.

Table 9: The effect of pile-up on the samples at successive stages of preselection with relative efficiencies.

	$t\bar{t}H$ σ (fb)	
Quantity \ Sample	No pile-up	pile-up
Starting Sample Generated	100	100
Pass Trigger (`e22i`, `e55`, `mu20`)	65 (65%)	62 (62%)
One high-p_T Lepton	56 (87%)	53 (86%)
$\geq$ 6 jets ($p_T >$ 20 GeV, $\|\eta\| < 5$)	36 (64%)	34 (64%)
$\geq$ 4 central *b*-jet candidates, ($\|\eta\| < 2.5$ & *b*-jet weight > 0)	16 (45%)	15 (44%)
Preselected	**16 (45%)**	**15 (44%)**

The distribution of the number of high p_T leptons in the events before preselection is shown in Fig. 26 for the pile-up and non pile-up samples. They are very similar, suggesting that the electron and muon reconstruction are not significantly affected by pile-up; however the trigger requirement reduces the number of events with pile-up available to the rest of the preselection.

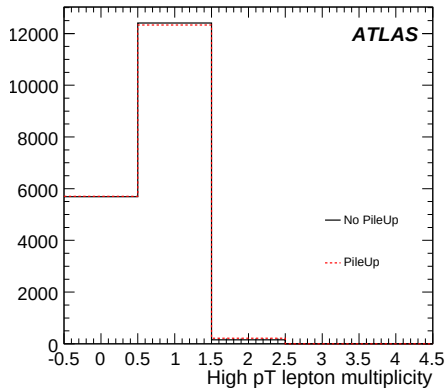


Figure 26: High p_T lepton multiplicity ($e^{\pm}, \mu^{\pm}$) for pile-up and non pile-up samples (before preselection).

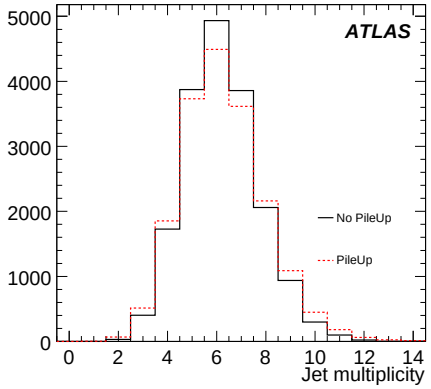


Figure 27: Jet multiplicity (before preselection) for pile-up and non pile-up samples. The cuts $p_T < 20$ GeV and $|\eta| < 5$ are applied to the individual jets.

Figure 27 shows the jet multiplicity before preselection. It can be seen that the number of events having exactly 6 jets is reduced by approximately 10%, and the number of events having more than 7 jets is increased. The net effect will be an increase in the combinatorial background, though the extra jets will typically have a low p_T.

For the cut-based and pairing likelihood analyses, candidate b-jets are designated as those jets lying in the central region of the detector ($|\eta| < 2.5$), with $p_T > 20$ GeV and b-jet weight > 5.5, however, if there are more than four of these then the jets with the highest b-jet weights are used.

The number of b-jets in the events both with and without pile-up after the other preselection cuts are applied can be seen in Fig. 28, where there is a reduction in the number of events having the requisite four b-jets.

The extent of the reduction in events at the various stages of the cut-based analysis is shown in Table 10. The most pronounced difference comes from the reduction in the number of b-jets, and the net effect of pile-up is a $\sim$12% reduction in the number of events where it is possible to reconstruct a Higgs particle, as shown in Fig. 29. This effect will also manifest itself in the backgrounds, since they all have either two or four b-quarks. The most interesting background pile-up study would now be with the $t\bar{t}j\bar{j}$ sample, since this relies on mis-tagging of light jets to create a physics background, however this is beyond the scope of this study which only examines the signal.

It should be noted that in the course of this study, the b-jet efficiency and light-jet rejections were studied, however no visible differences were observed. This can be explained by the fact that even a 1% drop in efficiency from 50% to 49% causes almost an 8% drop in events having four b-jets.

14 Significance estimates

The number of remaining events in the Higgs boson mass window (30 GeV around the nominal Higgs boson mass) have been used to compute a crude estimate of the statistical significance for this channel with 30 fb^{-1}. For such a channel in which the signal and backgrounds are very alike, this naive estimate is not the most relevant figure of merit, but it is still useful to compare analyses. For the cut-based analysis, a significance of 1.8 is achieved with signal to background ratio of approximately 0.11. It is worth noting that the addition of the low p_T muons to jets and the residual jet calibrations performed in Sections 5.3.3

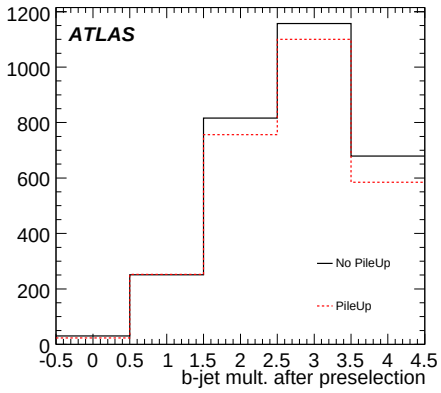


Figure 28: b jet multiplicity (after preselection) for pile-up and non pile-up samples. The cuts $p_T < 20$ GeV, $|\eta| < 2.5$ and $b$jet weight > 5.5 are applied to the individual jets.

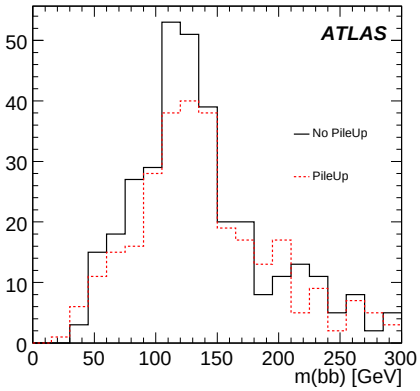


Figure 29: Reconstructed Higgs boson mass peak (m_{bb}) for pile-up and non pile-up samples.

Table 10: The extent of the reduction in events for pile-up and non pile-up samples for the cut-based analysis with relative efficiencies in parentheses. The harshest reduction comes from the four b-jet requirement.

	$t\bar{t}H$ σ (fb)	
Quantity \ Sample	No pile-up	pile-up
Preselected events	16.0	14.8
≥ 4 b-jets (b-jet weight > 5.5)	3.7 (23.1%)	3.2 (21.5%)
Had & Lep W inside mass-window	2.5 (66.4%)	2.1 (66.8%)
t, $\bar{t}$-quarks rec. in mass-window	2.0 (81.6%)	1.8 (83.1%)

and 5.3.4 improved the cut-based analysis significance by 0.3. With the pairing likelihood approach the significance is 1.95 for a signal to background ratio of 0.1. Finally the constrained fit likelihood gives 2.2 (1.7) for a signal over background value of 0.12 (0.14), obtained with a cut on $\mathscr{L}_{s/b}$ of -4.2 (-4.1). Figure 30 shows the total significance $S/\sqrt{B+(\Delta B)^2}$ as a function of the systematic error on the background (ΔB) for the different analyses. As is shown in the Fig., only a background uncertainty level below 10% allows exploitation of the statistical power of the mass constrained fit analysis with respect to the cut-based analysis, and even less for the case the pairing likelihood. Even for a robust analysis such as the cut-based approach, the large systematic uncertainties estimated in Table 8 provide a clear indication that a data driven background estimation is necessary.

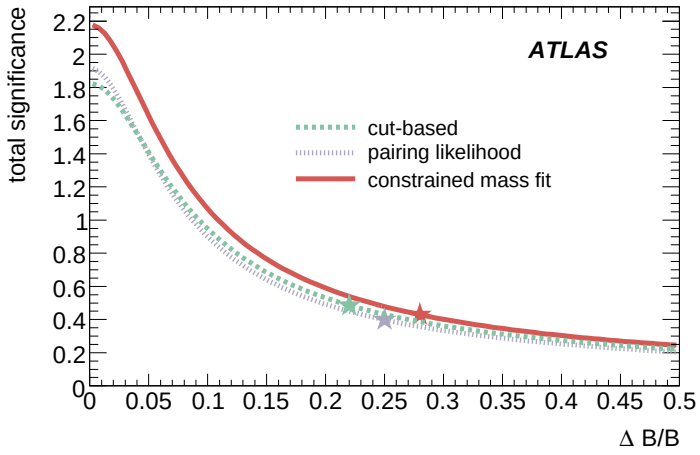


Figure 30: Comparison of the total significance as function of systematic uncertainties (ΔB), for the cut-based, the pairing likelihood and the constrained fit analysis. Markers indicate the significance corresponding to the background uncertainty estimated in Table 8.

15 Conclusion

We performed a baseline sensitivity study for the detection of a Standard Model Higgs boson decaying to $b\bar{b}$ when produced together with a $t\bar{t}$ pair. After the definition of a common preselection, three different techniques are used, all aimed at the reconstruction of the $t\bar{t}$ system. The first one is based on the reconstruction of the top quark and W candidate masses (cut-based analysis). The second one (pairing likelihood analysis) uses a more complete description of the kinematic properties of the $t\bar{t}$ system to build a likelihood discriminant and isolate the jets coming from the Higgs boson decay. The third approach (constrained fit) uses the known masses and jet errors as constraints to produce a combinatoric likelihood, and a second likelihood to separate signal from background. While the cut-based analysis is certainly the most stable one, relying only on the reconstructed invariant masses of the top quark candidates, it also performs worse with respect to the other two likelihood based analyses. On the other hand, these likelihood based analyses can be used successfully only after all kinematical variables are well understood together with their correlations. Although beyond the scope of this work, the use of more advanced multivariate techniques is foreseen to reduce both the combinatorial and physics background.

The statistical significance obtained for the three approaches was 1.82 for the cut-based, 1.95 for the pairing likelihood and 2.18 for the constrained mass fit at signal-to-background ratios of 0.11, 0.10 and 0.12 respectively. All the analyses suffer drastic reduction in significance as the overall systematic

uncertainty increases. The most important individual uncertainties are those for the jet energy scale and *b*-tagging efficiency.

From this study emerges the necessity of a strong *b*-tagging algorithm which is important not only to suppress the $t\bar{t}$+jets physics background but also to help reduce the combinatorial background by improving the hadronically decaying *W* reconstruction. It is also clear that the combinatorial background, responsible for the dilution of the Higgs boson mass peak, needs to be further reduced, possibly using multivariate techniques, in order to improve the statistical significance of the channel. Improvements in the mass peak resolution would also enhance the ability of a shape analysis from two perspectives; firstly it would be easier to select a signal-depleted region for any shape fits, and secondly the mass peak itself would become more pronounced.

The results presented in this work can be compared with a previous ATLAS study [3] performed using fast simulation with a parametrized *b*-tagging efficiency which had a higher performance than the one used here and also used PYTHIA in order to simulate the $t\bar{t}$+X background. It resulted in a significance of 1.9 and 2.6 respectively for the cut-based and likelihood analyses. The results presented in this note can also be compared with a recent CMS study [24] reporting a significance of 1.8 for the electron channel and 1.6 for the muon channel, in both cases for an integrated luminosity of 60 fb^{-1}. While a detailed comparison between the two experiments is not attempted in this work, it is noteworthy that the jet energy resolution quoted the CMS paper could be a key factor in explaining the improved sensitivity seen by ATLAS for this channel.

The measurement of the background normalization from data is vital for this channel. Subsequent studies must be performed in this regard. Further methods of extracting shape information from data must also be developed, in particular, the extraction of the signal in the presence of a *quasi-signal-like* background as is exhibited in the invariant mass plots at the ends of the analyses. The shape information and any estimate of the significance obtained from it could be used in conjunction with the counting experiment information to improve the overall significance.

References

[1] M.L. Mangano et al., *JHEP* 0307 (2003) 001 (and updates).

[2] J. Campbell, R. Ellis and D. Rainwater, *Phys. Rev.* **D68** (2003) 094021.

[3] J. Cammin and M. Schumacher, ATL-PHYS-2003-024.

[4] ATLAS Collaboration, Prospect for Single Top Quark Cross-Section Measurements, this volume.

[5] S. Tsuno et al., hep-ph/0204222v2.

[6] T. Sjostrand, S. Mrenna and P. Skands, *JHEP* 0605 (2006) 26.

[7] ATLAS Collaboration, Introduction on Higgs Boson Searches, this volume.

[8] W.-M. Yao et al., Journal of Physics, G 33, 1 (2006).

[9] B. Kersevan and E. Richter-Was, *Comp. Phys. Comm.*149 (2003) 142 (and updates).

[10] S. Frixione and B.R. Webber, *JHEP* 0206 (2002) 029 (and updates).

[11] G. Corcella et al., HERWIG 6.5 *JHEP* 0101 (2001) 010.

[12] J. Butterworth et al., http://projects.hepforge.org/jimmy/.

[13] ATLAS Collaboration, Jet Reconstruction Performance, this volume.

[14] ATLAS Collaboration, Cross-Sections, Monte Carlo Simulations and Systematic Uncertainties, this volume.

[15] ATLAS Collaboration, Physics Performance Studies and Strategy of the Electron and Photon Trigger Selection, this volume.

[16] ATLAS Collaboration, Performance of the Muon Trigger Slice with Simulated Data, this volume.

[17] ATLAS Collaboration, Reconstruction and Identification of Electrons, this volume.

[18] ATLAS Collaboration, Muon Reconstruction and Identification: Studies with Simulated Monte Carlo Samples, this volume.

[19] S. Corréard et al., *b*-tagging performance, ATL-PHYS-2004-006.

[20] ATLAS Collaboration, Soft Muon *b*-Tagging, this volume.

[21] ATLAS Collaboration, Top Quark Mass Measurements, this volume.

[22] ATLAS Collaboration, Determination of the Top Quark Pair Production Cross-Section, this volume.

[23] S. Dawson et al., *Phys. Rev.* **D68** (2003) 034022.

[24] CMS Collaboration, CMS Physics TDR, CERN-LHCC-2006-21 (2006).