Search for the Higgs boson in the ttH(H->bb) channel and the identification of jets containing two B hadrons with the ATLAS experiment.

Royer Edson Ticse Torres

CPPM-T-2016

AIX-MARSEILLE UNIVERSITE
Faculte Des Sciences de Luminy

Ecole Doctorale 352 : Physique et Sciences de la Matiere
Centre de Physique des Particules de Marseille

Thèse présentée pour obtenir le grade universitaire de docteur

Discipline : Physique et Sciences de la Matiere
Spécialité : Physique des Particules et Astroparticules

Royer Edson TICSE TORRES

Search for the Higgs boson in the $t\bar{t}H(H \to b\bar{b})$ channel and the identification of jets containing two B hadrons with the ATLAS experiment.

Soutenue le 29/09/2016 devant le jury :

Prof. Aurelio JUSTE ROZAS	Rapporteur
Dr. Tim SCANLON	Rapporteur
Dr. Patrice VERDIER	Examinateur
Dr. Eric KAJFASZ	Examinateur, Président du jury
Dr. Arnaud DUPERRIN	Directeur de thèse

Abstract

Keywords: LHC, ATLAS, Higgs boson, $H \to b\bar{b}$, jet flavor tagging, boosted decision trees

In July 2012, CERN announced the discovery of the Higgs boson, the last missing piece of the Standard Model (SM). The Higgs boson has been observed coupling directly to W and Z bosons and tau leptons, and indirectly to top quarks. In order to probe if it is indeed the particle predicted by the SM, or by a theory beyond the SM, direct couplings of the Higgs boson to quarks must also be measured and compared with the SM prediction.

Observing the Higgs boson production in association with a pair of top quarks ($t\bar{t}H$) would allow a direct measurement of the top quark Yukawa coupling and provides an important test of the Higgs mechanism within the SM. This thesis presents a search for the Higgs boson in the $t\bar{t}H(H \to b\bar{b})$ channel using proton-proton collisions at $\sqrt{s}$ = 13 TeV, collected with the ATLAS detector in 2015 and 2016. The Higgs boson decays to two b quarks and top quark pair decays with one lepton, the $t\bar{t}H \to (l\nu b)(jjb)(b\bar{b})$ single lepton channel, are considered.

This document details in particular the contributions made by the author in this search: the full reconstruction of the $t\bar{t}H(H \to b\bar{b})$ single lepton system and the final discrimination between signal and the main background, $t\bar{t}$+jets. A new method was developed to solve the large combinatorial background by fully reconstructing the $t\bar{t}H(H \to b\bar{b})$ final state using a multivariate technique to uniquely associate each reconstructed jets to the initial quarks. A multivariate technique was also used to discriminate between the signal and the main $t\bar{t}$+jets background further increasing the sensitivity of the search compared to the Run 1 analysis. Finally, the first result at $\sqrt{s}$ = 13 TeV is shown. The signal strength (the ratio between the measured and predicted cross sections) is found to be 1.6 $\pm$ 1.1. No significant excess of events above the background expectation is found and an observed (expected) limit of 3.6 (2.2) is set at 95% confidence level.

The $t\bar{t}b\bar{b}$ is one of the main backgrounds for the $t\bar{t}H(b\bar{b})$ search. Recent studies with Monte Carlo events generators show that there is a large fraction of $t\bar{t}b\bar{b}$ events with jets containing two b-hadrons. Standard algorithms for the identification of jets originating from bottom quarks (b-tagging) provide tools to differentiate single b-hadron jets from c- and non-hadron jets but they are not efficient for the identification of jets containing two b-hadrons. A new b-tagging algorithm has been developed to discriminate such jets from single b-hadrons jets. The description of this new b-tagging tool and its performance is presented in this thesis.

Resume

En juillet 2012, le CERN a annoncé la découverte du boson de Higgs qui est la dernière particule manquante du Modèle Standard. Le boson de Higgs observé montre un couplage directe au bosons W, Z et au lepton tau et indirect au quark top. Afin de verifier s'il s'agit bien du boson de Higgs du Modèle Standard ou d'un modèle alternatif, les couplages directs du boson de Higgs au quarks doivent également être mesurés et comparés aux prédictions du Modèle Standard.

La recherche du boson de Higgs produit en association avec une paire de quarks top ($t\bar{t}H$) est le seul moyen pour accéder directement au couplage de Yukawa du boson de Higgs au quark top. Cette measure fournit un test important du mécanisme de Higgs dans le Modèle Standard. Cette thèse présente une recherche du boson de Higgs dans le canal $t\bar{t}H(H \to b\bar{b})$, en utilisant les données de collisions proton-proton à $\sqrt{s} = 13$ TeV, collectées avec le détecteur ATLAS en 2015 et 2016. Le canal considéré est $t\bar{t}H(H \to (l\nu b)(jjb)(b\bar{b})$, le boson de Higgs se désintégrant en deux quarks b et l'un des quarks top se désintégrant avec en lepton.

Ce document détaille en particulier les contributions faites par l'auteur de cette recherche : la reconstruction entière du systeme $t\bar{t}H(H \to b\bar{b})$ en un seul lepton et la discrimination finale entre le signal et le bruit de fond principal $t\bar{t}$+jets. Une nouvelle méthode a été développée pour résoudre le probleme de combinatoire en reconstruisant entièrement l'etat final ttH avec la technique multivariée afin d'associer à chaque jet reconstruit un quark initial unique. Finalement, une technique multivariée a été utilisée pour séparer le signal du bruit de fond principal $t\bar{t}$+jets augmentant ainsi la sensibilité de la recherche par rapport à l'analyse Run 1.

Le $t\bar{t}b\bar{b}$ est l'un des principaux bruit de fonds pour la recherche de $t\bar{t}H(H \to b\bar{b})$). Des études récentes bases sur des événements générés par Monte Carlo montrent qu'il y a une fraction importante d'événements $t\bar{t}b\bar{b}$ avec des jets contenant deux hadrons b. Les algorithmes de b-tagging standard permettent de différencier les jets a un hadron b des jets c et legers, mais ces algorithmes ne sont pas efficaces pour identifier des jets a deux b-hadrons. Un nouvel algorithme de b-tagging a été développé pour séparer ces jets des jets a un hadron b. Une description de ce nouvel outil de b-tagging et de ses performances sont présentées dans cette thèse.

Acknowledgements

First of all, I want to express my best gratitude to Georges Aad. Thank you infinitely for your patience, your invaluable help and your great knowledge that followed me during my years of PhD student. It's my greatest pleasure to work with you and the ATLAS group at CPPM.

My most sincere gratitudes goes next for Arnaud Duperrin, my supervisor. I cannot express enough my thanks for your support, your patience, your kindness. Thank you for accepted me as your PhD student and gave me the chance to do my research about the Higgs boson in the ATLAS experiment. Thank you for believing in me.

A great thanks for my reviewers, Tim Scalon and Aurelio Juste, for the work you have been accomplished for your corrections and suggestions. I also express my thanks to the examiners, Patrice Verdier and Eric Kajfasz.

I also thank to my senior colleagues at CPPM: Yann Coadou and Timothée Theveneaux-Pelzer for gave me precious suggestions on my thesis writing and polished my English word by word.

A big thank you also to all my colleagues and friends from CPPM, especially thanks to Kazuya Mociduki and Sebastien Kahn for help me in many subjects and aspects during my thesis.

I express my gratefulness for the great helps from members of the b-tagging and the $t\bar{t}H(H \rightarrow b\bar{b})$ group in the ATLAS collaboration. Working in such a large experiment gave me the chance to collaborate with scientist around the world.

At last, I would like to take this opportunity to thank to my family for giving me the freedom to follow my own interests.

Contents

List of Figures

List of Tables

Introduction

The Standard Model (SM) of particle physics describes with great precision almost all the observed particle properties and their interactions. The model, developed during the second half of the twentieth century, has been extensively tested with no experimental results that contradict its predictions. The discovery of the vector bosons ($W^{\pm}$ and Z) with the expected properties increased our confidence in the model. The mass of the particles are arbitrary parameters of the model and its origin is described via the Higgs mechanism, an elegant solution responsible for electroweak symmetry breaking. The mechanism generates the mass of all the particles of the model, and also creates an associated particle, the Higgs boson. This key particle of the SM was discovered by the ATLAS and CMS collaboration in July 2012, forty years after its prediction.

Following the discovery of the Higgs boson, further data will allow in-depth investigation of its properties. Due to its large mass, the top quark coupling to the Higgs boson (or top Yukawa coupling) is the largest among the fermions in the SM. Probing the top Yukawa coupling directly requires a process that results in both a Higgs boson and top quarks. The associated production of a Higgs boson with a top quark pair (ttH channel) would allow a direct measurement of the top Yukawa coupling at the LHC. This thesis presents a search for the SM Higgs boson produced in association with top quarks decaying to a b quark pair where one of the top quark pair decays to a lepton, the $t\bar{t}H \rightarrow (l\nu b)(jjb)(b\bar{b})$ single lepton channel is considered.

The analysis is performed with 13.2 fb^{-1} of data recorded with the ATLAS detector in 2015 and part of 2016 at a centre-of-mass energy of 13 TeV. A particular focus is placed on the main contributions made by the author: the full event reconstruction of the $t\bar{t}H(H \rightarrow b\bar{b})$ single lepton system and the final discrimination between the signal and the large $t\bar{t}$+jets background using multivariate techniques. In the $t\bar{t}H(H \rightarrow b\bar{b})$ search, it is very challenging to distinguish the signal from the irreducible and huge background arising from top quark pair production with additional b-quarks ($t\bar{t} + b\bar{b}$). Two steps have been developed in order to increase the signal-to-background separation. In the first part we find the best corresponding match between the observed jets and final-states quarks from the $t\bar{t}H(H \rightarrow b\bar{b})$ system using a multivariate analysis (called MVA reconstruction). This allows variables related to the reconstructed top quarks or the Higgs boson to be defined with a good discrimination power in a natural way (e.g. Higgs mass). In the second step we combine these variables with additional global variables (e.g. average ΔR for all b-tagged jet pairs) in a multivariate technique. Variables from the MVA reconstruction improve the signal-to-background separation by about 16% in the most sensitive region. The novel method to separate signal from background was

successfully implemented in the first ATLAS $t\bar{t}H(H \rightarrow b\bar{b})$ Run 2 result. The best-fit signal strength (the ratio between the measured and predicted cross sections) is found to be 1.6 ± 1.1. No significant excess of events above the background expectation is found and an observed (expected) limit of 3.6 (2.2) is set at 95% confidence level.

The $t\bar{t}+b\bar{b}$ is one of the main backgrounds for the $t\bar{t}H(H \rightarrow b\bar{b})$ search. Recent studies with Monte Carlo events generators show that there is a large fraction of $t\bar{t} + b\bar{b}$ events with jets containing two b-hadrons. The most advanced ATLAS b-tagging algorithms do not provide information on the number of b-hadrons within the jet. This thesis presents a new tagger to identify jets containing two b-hadrons. We use a Multiple Secondary Vertex (MSV) algorithm to reconstruct multiple vertices within jets. Then, a multivariate analysis (Boosted Decision Tree) is used to increase the discrimination power between jets with two b-hadrons and jets containing a single b, c or no hadrons.

This thesis is organised as follows. Chapter 1 gives the basis of the SM of particle physics and introduces the Higgs mechanism and Higgs boson physics. The ATLAS detector is detailed in chapter 2. Chapter 3 presents the identification of jets containing two b-hadrons. The search for the Higgs boson in the $t\bar{t}H(H \rightarrow b\bar{b})$ single lepton channel is described in chapter 4. Finally, chapter 5 summarises the studies in this thesis.

1. Theoretical introduction

In this chapter a short overview of the Standard Model (SM) is presented. After a brief review of the elementary particles and fundamental interactions, we will focus on the electroweak sector of the SM and particularly on the Brout-Englert-Higgs mechanism of the symmetry breaking that generates the mass of particles. This mechanism implies the existence of a new particle, the Higgs boson. The observation of the Higgs boson particle by ATLAS [1] and CMS [2] collaborations in July 2012 experimentally proved the existence of the Higgs field. This was the final missing piece of the Standard Model to be experimentally verified, that had taken 48 years to be experimentally confirmed. The Higgs boson production and decays at hadronic colliders are described in the last section.

1.1. Introduction to Standard Model of particle physics

The Standard Model of particle physics is the theory that describes the elementary constituents of matter and their interactions. With the exception of gravity, it describes the three known fundamental interactions: weak, strong and electromagnetic. The elementary particles described in the SM can be classified into two types: fermions (leptons and quarks) which constitute matter and bosons which act as the mediators of the fundamental forces. They are summarised in figure 1.1.

The Standard Model of electroweak and strong interactions is a quantum field theory based on the gauge group:

$$SU(3)_C \times SU(2)_L \times U(1)_Y, \tag{1.1}$$

where C refer to colour, L to the left-handed weak isospin and Y to hypercharge. The first part of the gauge group $SU(3)_C$ is the non-Abelian symmetry group which describe the strong interaction between quarks. The gluonic gauge fields are coupled to the colour charge as formalised in quantum chromodynamics (QCD) [3, 4]. The second part of the gauge group $SU(2)_L \times U(1)_Y$ represents the unified electroweak interactions known as the Glashow-Salam-Weinberg theory [5–7], which is spontaneously broken via the Brout-Englert-Higgs mechanism [8–11] .

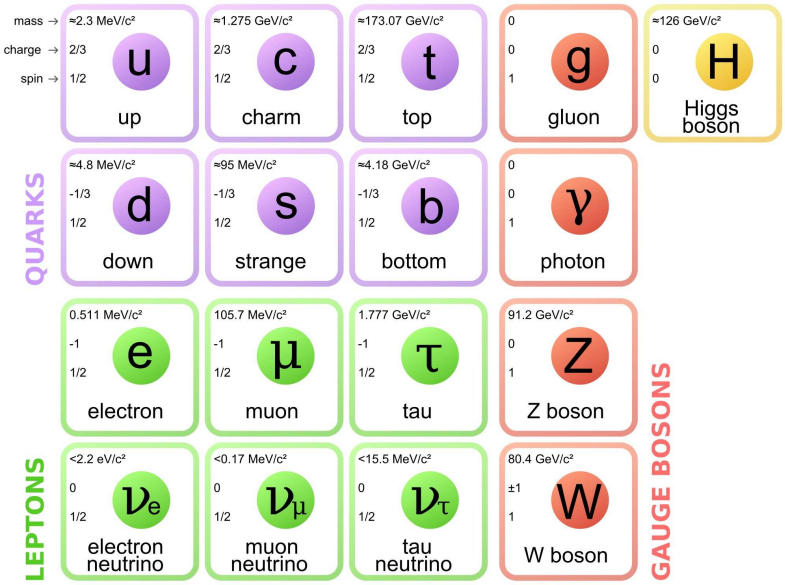


Figure 1.1.: The Standard Model elementary particles. The classification contains some of their characteristics (mass, charge, spin) of the three families of quarks and leptons, intermediate bosons and the Higgs boson.

1.1.1. Elementary particles and fundamental interactions

All the elementary particles of the Standard Model are represented as fundamental fields and can be divided in two groups: matter fields, that is, the three generations of quarks and leptons, carrying spin-1/2 and gauge fields corresponding to the spin-one (or spin-zero for the Higgs) bosons that mediate the interactions.

Leptons and quarks are organised in families, with the left–handed fermions belonging to weak isospin doublets while the right–handed components transform as weak isospin singlets. Table 1.1 lists all leptons and quarks in the SM[a].

The down-type quarks d', s' and b' denote the electroweak eigenstates consisting of a mixture of mass eigenstates via the Cabibbo–Kobayashi–Maskawa (CKM) unitary matrix [12, 13].

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = V_{\mathrm{CKM}} \begin{pmatrix} d \\ s \\ b \end{pmatrix}. \tag{1.2}$$

In the Standard Model the CKM matrix can be described by three angles and one complex phase. Currently the experimental value of each element of the CKM matrix [14] is:

[a] Experiments have shown that neutrinos (ν) are always left-handed. Since right-handed neutrinos do not exist in the Standard Model, the theory predicts that neutrinos can never acquire mass.

Name				
Quarks	Q_L	$\begin{pmatrix} u \\ d' \end{pmatrix}_L$	$\begin{pmatrix} c \\ s' \end{pmatrix}_L$	$\begin{pmatrix} t \\ b' \end{pmatrix}_L$
	q_R	u_R	c_R	t_R
		d_R	s_R	b_R
Leptons	L_L	$\begin{pmatrix} \nu_e \\ e \end{pmatrix}_L$	$\begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_L$	$\begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}_L$
	l_R	e_R	μ_R	τ_R

Table 1.1.: SM fermions. The left-handed (denoted as L) quarks and leptons are in weak isospin doublets and the right-handed (denoted as R) quarks and charged leptons are in weak isospin singlets.

$$\begin{pmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{pmatrix} = \begin{pmatrix} 0.97427 \pm 0.00015 & 0.22534 \pm 0.00065 & 0.00351^{+0.00015}_{-0.00014} \\ 0.22520 \pm 0.00065 & 0.97344 \pm 0.00016 & 0.0412^{+0.0011}_{-0.0005} \\ 0.00867^{+0.00029}_{-0.00031} & 0.0404^{+0.0011}_{-0.0005} & 0.999146^{+0.000021}_{-0.000046} \end{pmatrix}. \tag{1.3}$$

The V_{CKM} terms represent transitions between the different quark generations. For example, the top quark decays almost 100% of the time into a b quark and W boson ($V_{tb} \sim 1$).

The quarks have fractional electric charges: u, c and t quarks have $+\frac{2}{3}e$ (units of elementary charge, e), while the d, s and b quarks have a fractional charge of $-\frac{1}{3}e$. Quarks carry the charge associated with the strong interaction, referred as colour and can be either red, green or blue. Similarly, the leptons are divided into electrically neutral neutrinos and charged leptons with charge -1e. Leptons and quarks have corresponding antiparticles with the same mass, but opposite quantum numbers. While leptons can exist as free particles, quarks are always found in groups of two, three or more[b],and form particles called hadrons (colour-neutral state). There are two well know types of hadrons: the baryons made of three quarks or three antiquarks and mesons composed of a quark-antiquark pair.

The SM contains the gauge bosons corresponding to strong and electroweak interactions, and the Higgs boson, responsible of the masses of the particles. The corresponding gauge fields are shown in table 1.2.

The gauge fields are associated to different representations of the symmetry groups of the SM. In the electroweak sector, the field B_μ corresponds to the generator Y of the $U(1)_Y$ group and the three fields W^i_μ correspond to the generators I^i of the $SU(2)_L$ group. These generators are defined as:

[b] In April 2014 the LHCb collaboration published results of measurements which demonstrated that the $Z(4430)^+$ particle is composed of four quarks ($c\bar{c}d\bar{u}$). And in July 2015 the first observation of two pentaquark particles, i.e. hadrons composed of five quarks, was announced.

Name		
Bosons	Hypercharge	B_μ
	Isospin	W^i_μ $i = 1, 2, 3$
	Colour	G^a_μ $a = 1, \ldots, 8$
	Higgs	h

Table 1.2.: SM gauge fields.

$$I^i = \frac{1}{2}\sigma^i, \tag{1.4}$$

where σ^i are the Pauli matrices. The commutation relation between the generators are given by:

$$[I^i, I^j] = i\epsilon^{ijk} I_k \qquad \text{and} \qquad [Y, Y] = 0, \tag{1.5}$$

where ϵ^{ijk} is the antisymmetric tensor. The four fields are massless in order to conserve the symmetry. However, the symmetry breaking induced by Higgs field changes them. The charged weak bosons $W^\pm$ appear as a linear combination of W^1 and W^2, while the photon (γ) and the neutral weak boson Z are both given by mixture of W^3 and B_μ.

In the strong interaction sector, the spin 1 gluon fields G^a_μ are an octet associated to the eight generator T^a of the $SU(3)_C$ group and which obey the following relations:

$$[T^a, T^b] = if^{abc}T_c \qquad \text{with} \qquad Tr[T^a T^b] = \frac{1}{2}\delta_{ab}, \tag{1.6}$$

where the tensor f^{abc} is the structure constant of the $SU(3)_C$ group.

The bosons are the force carriers that mediate the fundamental interaction. The photon γ mediates the electromagnetic force between electrically charged particles, the $W^\pm$ and Z^0 bosons mediate the weak interaction between particles of different flavors (quarks and lepton) and the gluons mediates the strong interaction between colour charged particles (quarks). Table 1.3 presents a summary of the bosons with their masses and electric charges.

Interaction	Mediator	Electric charge [e]	Mass [GeV]
Strong	8 gluons (g)	0	0
Electromagnetic	photon (γ)	0	0
Weak	charge weak ($W^\pm$)	± 1	80.385 ± 0.015
	neutral weak (Z)	0	91.1876 ± 0.0021

Table 1.3.: Fundamental interactions in the SM. The mediators (bosons), their electric charge and masses are listed as well [14]. The gravitation is not included in the SM.

1.1.2. The Standard Model and the Brout-Englert-Higgs mechanism

In this section, a brief review of the SM formalism is presented. The derivation of the formalism mainly follows the approach of ref [15, 16].

The SM Lagrangian summaries the laws of physics for the three basic interactions, the electromagnetic, the weak and the strong interactions between the leptons and the quarks. Moreover, the specific form of the Higgs interactions generates the mass of the elementary particles. The SM Lagrangian has four contributions:

$$\mathcal{L}_{SM} = \mathcal{L}_{gauge} + \mathcal{L}_{fermions} + \mathcal{L}_{Higgs} + \mathcal{L}_{Yukawa}. \tag{1.7}$$

The first contribution contains the kinematic and self-interactions terms of the various gauge fields. It is formulated with the composition of the $U(1)_Y$ gauge field B_μ, the three $SU(2)_L$ gauge fields W^i_μ and the eight $SU(3)_C$ gauge fields $G^a_{\mu\nu}$:

$$\mathcal{L}_{gauge} = -\frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a - \frac{1}{4}W^a_{\mu\nu}W^{\mu\nu}_a - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}, \tag{1.8}$$

where the gauges field are defined as:

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu + g_3 f^{abc} G^b_\mu G^c_\nu \qquad a \in [1,8], \tag{1.9}$$

$$W^i_{\mu\nu} = \partial_\mu W^i_\nu - \partial_\nu W^i_\mu + g_2 \epsilon^{ijk} W^b_\mu W^c_\nu \qquad i \in [1,3], \tag{1.10}$$

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu, \tag{1.11}$$

where g_2 and g_3 are the weak-isospin and the strong coupling constant of $SU(2)$ and $SU(3)$, respectively.

The second term of the Lagrangian 1.7 describes the dynamics of the fermion-gauge boson coupling:

$$\mathcal{L}_{fermions} = \sum_{\psi_L,\psi_R} \bar{\psi} i\gamma^\mu D_\mu \psi, \tag{1.12}$$

with the sum running over the left- and right- handed field components of the leptons and quarks. The matter fields ψ are coupled to the gauge fields through the covariant derivative D_μ. In case of quarks, the covariant derivative is defined as:

$$D_\mu \psi = (\partial_\mu - i g_3 T_a G^a_\mu - i g_2 I_i W^i_\mu - i g_1 \frac{Y}{2} B_\mu)\psi, \tag{1.13}$$

where the weak-hypercharge coupling constant of $U(1)$ is denoted by g_1.

The electroweak part of the Lagrangian 1.8:

$$\mathcal{L}_{EW} = -\frac{1}{4}W^a_{\mu\nu}W^{\mu\nu}_a - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}, \tag{1.14}$$

can be reformulated by introduction of the transformations:

$$A_\mu = \sin\theta_W W_\mu^3 + \cos\theta_W B_\mu, \tag{1.15}$$

$$Z_\mu = \cos\theta_W W_\mu^3 - \sin\theta_W B_\mu, \tag{1.16}$$

$$W^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp i W_\mu^2), \tag{1.17}$$

where the free parameter θ_W is the Weinberg angle, defined by the ratio of the $SU(2)_L$ and $U(1)_Y$ couplings:

$$\tan\theta_W = g_1/g_2. \tag{1.18}$$

With these transformation the Lagrangian 1.14 changes into:

$$\mathcal{L}_{EW} = -\frac{1}{4}F_{\mu\nu}(x)F^{\mu\nu}(x) - \frac{1}{2}F^\dagger_{W\mu\nu}(x)F_W^{\mu\nu}(x) - \frac{1}{4}F_{Z\mu\nu}(x)F_Z^{\mu\nu}(x), \tag{1.19}$$

where $F_{\mu\nu}$ represent the electromagnetic tensor associated to the photon field A_μ, $F_{W\mu\nu}$ and $F_{Z\mu\nu}$ are tensors related to the electroweak fields $W^\pm$ and Z.

Up to this point, the gauge fields and the fermions fields have been kept massless. The fields $W^\pm$ and Z can acquire a mass if one added the terms $m_W^2 W_\mu^\dagger(x)W^\mu(x) + \frac{1}{2}m_Z^2 Z_\mu(x)Z^\mu(x)$ into 1.19. This incorporation of mass terms for the gauge bosons (and for fermions) leads to a breakdown of the local $SU(2)\times U(1)$ gauge invariance.

Therefore, an essential ingredient of the SM is the scalar potential that is added to the Lagrangian to generate the vector–boson (and fermions) masses without explicit breaking of the $SU(2)\times U(1)$ gauge symmetry. It is made via the Brout-Englert-Higgs mechanism of spontaneous symmetry breaking.

The Brout-Englert-Higgs mechanism

The Higgs field is introduced as a complex scalar doublet:

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}, \tag{1.20}$$

and the Higgs Lagrangian is written as:

$$\mathcal{L}_{Higgs} = (D^\mu\phi)^\dagger(D_\mu\phi) - V(\phi^\dagger\phi), \tag{1.21}$$

with the scalar potential in the form:

$$V(\phi^\dagger\phi) = \mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2, \tag{1.22}$$

For $\mu^2 > 0$ the minimum occurs at $\phi = 0$. That is, the vacuum is empty space and $SU(2)\times U(1)$ gauge symmetry is unbroken at the minimum. For $\mu^2 < 0$ and $\lambda > 0$, one produces the shape of a "mexican hat", as illustrated in figure 1.2. The potential $V(\phi)$ has two critical points: a local maximum at $\phi = 0$ and the nonzero minimum at

$\phi_0 = \sqrt{\frac{-\mu^2}{2\lambda}}$, which breaks the $SU(2) \times U(1)$ symmetry invariance.

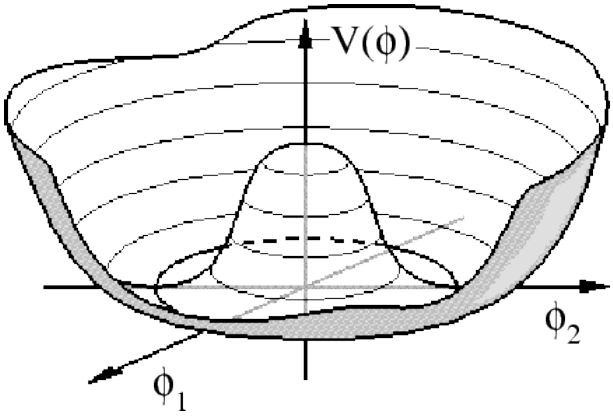


Figure 1.2.: The Higgs potential $V(\phi)$ for $\mu^2 < 0$ and $\lambda > 0$.

The Higgs field can be expanded around this minimum to produce the expression:

$$\phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}, \tag{1.23}$$

where $v = \sqrt{\frac{-\mu^2}{\lambda}}$ is referred to as the vacuum expectation value and the field $h(x)$ describes small perturbations around the ground state. The lowest energy excitation of the Higgs field above its ground state is known as the Higgs boson.

Then, the Higgs Lagrangian can be written in terms of the new field as:

$$\mathcal{L}_{Higgs} = \left| (\partial_\mu - ig_2 I_i W_\mu^i - ig_1 \frac{Y_q}{2} B_\mu) \frac{(v+h)}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right|^2 - \mu^2 \frac{(v+h)^2}{2} - \lambda \frac{(v+h)^4}{4}. \tag{1.24}$$

The first term of 1.24 contains the vector bosons and can be expanded as:

$$\begin{aligned} &= \frac{1}{2} \left| \begin{pmatrix} \partial_\mu - \frac{i}{2}(g_2 W_\mu^3 + g_1 B_\mu) & -\frac{ig_2}{2}(W_\mu^1 - iW_\mu^2), \\ -\frac{ig_2}{2}(W_\mu^1 + iW_\mu^2) & \partial_\mu + \frac{i}{2}(g_2 W_\mu^3 - g_1 B_\mu) \end{pmatrix} \begin{pmatrix} 0 \\ v+h \end{pmatrix} \right|^2 \\ &= \frac{1}{2}(\partial_\mu h)^2 + \frac{1}{8} g_2^2 (v+h)^2 |W_\mu^1 + iW_\mu^2|^2 + \frac{1}{8}(v+h)^2 |g_2 W_\mu^3 - g_1 B_\mu|^2. \end{aligned}$$

In terms of the physical fields A_μ (1.15), Z_μ (1.16) and $W^\pm$ (1.17), it becomes:

$$= \frac{1}{2}\partial_\mu h \partial^\mu h + \frac{g_2^2}{4}(v+h)^2 (W_\mu^+ W^{-\mu} + \frac{g_2^2 + g_1^2}{2g_2^2} Z_\mu Z^\mu). \tag{1.25}$$

The bilinear terms in the fields $W^\pm$, Z, and A are:

$$m_W^2 W_\mu^\dagger W^\mu + \frac{1}{2} m_Z^2 Z_\mu Z^\mu + \frac{1}{2} m_A^2 A_\mu A^\mu.$$

In regards of 1.25, the masses of the W and Z bosons can be written as:

$$m_W = \frac{1}{2} v g_2, \tag{1.26}$$

$$m_Z = \frac{1}{2} v \sqrt{g_2^2 + g_1^2}, \tag{1.27}$$

while the photon remains massless, $m_A = 0$. Therefore, expressing the Higgs field in terms of its ground state via the addition of the real scalar field h induces effective masses for particles propagating through it.

The second term of 1.24 gives rise to terms involving exclusively the scalar field h, namely:

$$-\frac{1}{2}(-2\mu^2)h^2 + \frac{1}{4}\mu 2v^2\left(\frac{4}{v^3}h^3 + \frac{1}{v^4}h^4 - 1\right),$$

where one can see that the Higgs boson mass term is:

$$m_h = \sqrt{-2\mu^2} = \sqrt{2\lambda}v. \tag{1.28}$$

The vacuum expectation value v can be determined via the relation $v = (\sqrt{2}G_F)^{1/2} \approx 246$ GeV. However, the λ parameter is associated purely with the scalar field, and thus cannot be know without knowledge of the scalar field itself. This means that the Higgs mass cannot be predicted from the theory.

Finally, fermions acquire their masses using the same scalar field ϕ. The general $SU(2) \times U(1)$ invariant Yukawa Lagrangian can be introduced as:

$$\mathcal{L}_{Yukawa} = g_f \bar{f} f \phi. \tag{1.29}$$

Replacing the Higgs field by its ground state value $\phi \to v/\sqrt{2}$, it gives the mass term $g_f v/\sqrt{2}\bar{f}f$, where the first constant term is identified as the fermion mass:

$$m_f = g_f \frac{v}{\sqrt{2}}. \tag{1.30}$$

Though the masses of the fermions can be introduced in a consistent way via the Higgs mechanism, the SM does not predict their experimental values. The top quark Yukawa coupling is considered particularly interesting since it is of the order of 1:

$$g_{top} = \sqrt{2}\frac{m_{top}}{v} \approx 1. \tag{1.31}$$

Table 1.4 summarises the intensity of the couplings of the Higgs bosons to fermions (f), the vector gauge bosons ($V \equiv W$ or Z) and to itself. The Higgs couplings are proportional to the particles mass. Therefore, one can establish two general principles: *(i)* the Higgs boson will be produced in association with heavy particles; *(ii)* the Higgs boson will decay into the heaviest particles that are accessible kinematically.

Coupling	Intensity
$Hf\bar{f}$	m_f/v
HVV	$2m_V^2/v$
$HHVV$	$2m_V^2/v^2$
HHH	$3m_H^2/v$
$HHHH$	$3m_H^2/v^2$

Table 1.4.: The Higgs boson couplings to fermions (f), vector gauge bosons ($V \equiv W$ or Z) and the Higgs self-coupling in the SM.

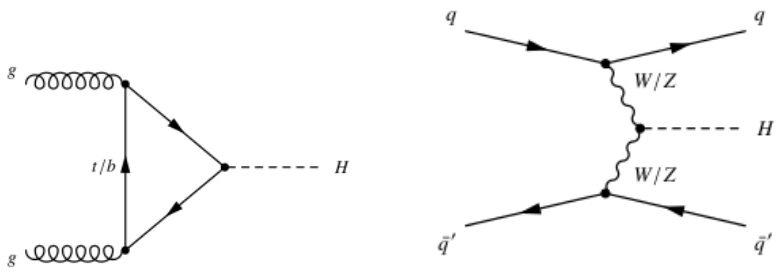
1.2. The Higgs boson

The properties of the SM Higgs boson have been computed by the LHC Higgs Cross Section Working Group [17]. In this section, the most important scattering processes in hadron colliders, particularly at the LHC, and decays of the Higgs boson will be summarised briefly.

1.2.1. Higgs production in hadron colliders

In the SM, the Higgs boson couples preferentially to heavy particles, i.e. mainly to the vector bosons $W^{\pm}$ and Z, and to the top and bottom quarks. Thus, the Higgs boson production at the LHC occurs mainly through the following processes:

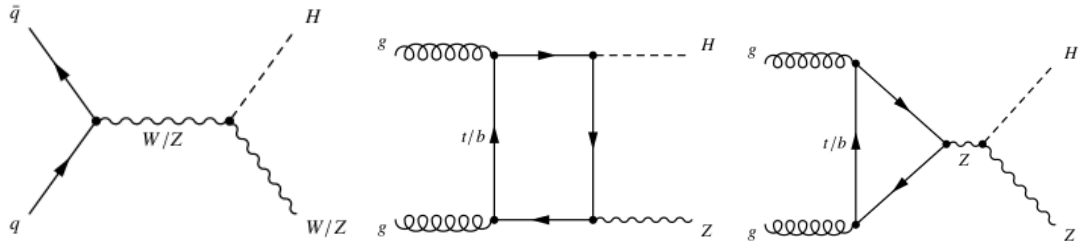
- The dominant process is the gluon fusion (ggF): $pp \to gg \to H$, shown in figure 1.3a, where the loop is dominated by the top quark due to his large mass in comparison to the other quarks.
- The second process in terms of cross section is the production by vector boson fusion (VBF): $pp \to qqVV \to qqH$, shown in figure 1.3b, where the two virtual vector bosons (W or Z) annihilate to create a Higgs boson.



(a) (b)

Figure 1.3.: Leading-order Feynman diagrams for Higgs boson production via the (a) ggF and (b) VBF production processes.

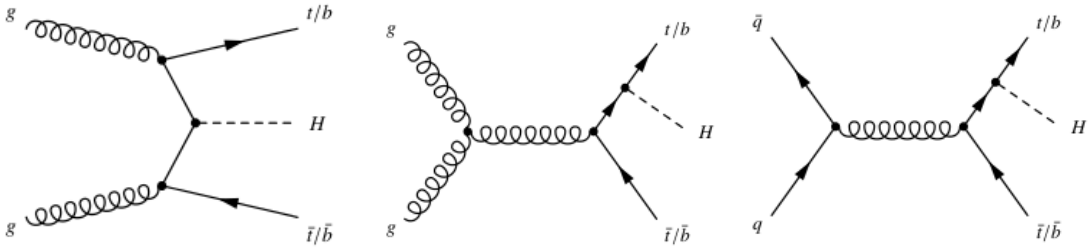
- The two next dominant production modes are the associated production with a vector boson (VH): $pp \to q\bar{q} \to VH$, as shown in figure 1.4a, or $pp \to gg \to ZH$, as shown in figures 1.4b and 1.4c.



(a) (b) (c)

Figure 1.4.: Leading-order Feynman diagrams for Higgs boson production via the (a) qq → VH and (b),(c) gg → ZH production processes.

- Finally we have the associated production with two heavy quarks, dominated by a pair of top quark (ttH): $qq, gg \to ttH$, shown in figure 1.5. The ttH production allows to measure the top Yukawa coupling. The cross section $\sigma(pp \to ttH)$ is directly proportional to the square of this fundamental coupling.

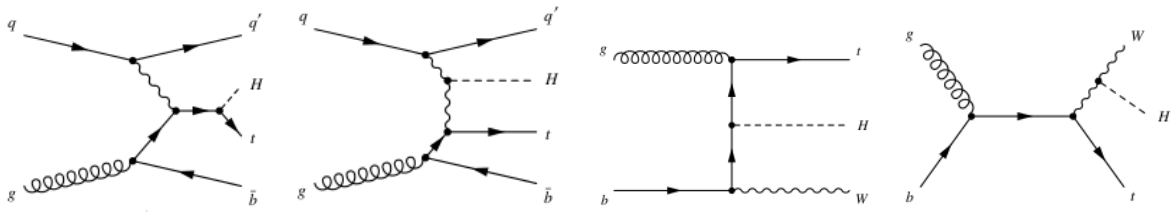


(a) (b) (c)

Figure 1.5.: Leading-order Feynman diagrams of Higgs boson production via the qq/gg → ttH and qq/gg → bbH processes.

- Other less important production processes are the production in association with a single top quark (tH), show in figure 1.6. The tH process is expected to have a negligible contribution in the SM but may become important in some Beyond Standard Model (BSM) scenarios.

The production cross sections at the LHC are quite sizeable so that a large sample of the SM Higgs particles can be produced. The production cross sections at $\sqrt{s} = 13$ TeV are summarised in table 1.5. Experimental difficulties arise from the huge number of background events that come along with the Higgs signal events. This problem can be tackled by triggering on leptonic decays of W, Z and top quarks in the associated productions.



(a) (b) (c) (d)

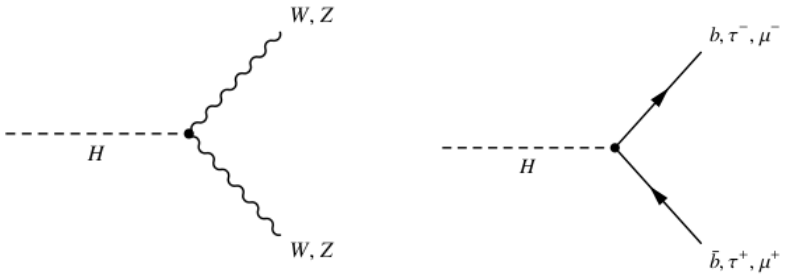
Figure 1.6.: Leading-order Feynman diagrams of the Higgs boson production in association with a single top quark (a),(b) qb → tHq and (c),(d) gb → tHW processes.

Production process	Cross section [pb] $\sqrt{s} = 13$ TeV
ggF	48.6
VBF	3.78
WH	1.37
ZH	0.884
ttH	0.507
tH	0.074

Table 1.5.: The Higgs boson production cross sections at the LHC for various production mechanisms. SM predictions with the Higgs mass of 125 GeV [17].

1.2.2. Higgs decays

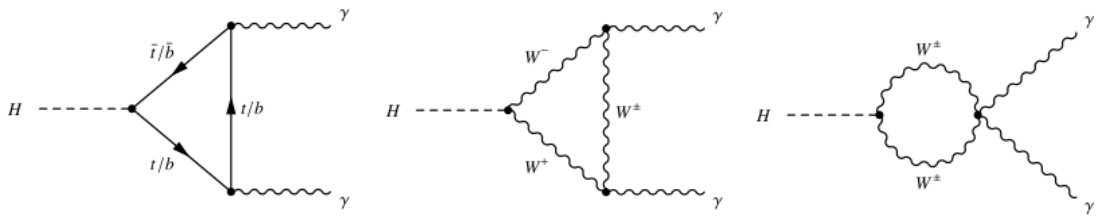
In the SM, the possible decay modes of the Higgs boson are essentially determined by the value of its mass. The Higgs decay width (Γ) is directly related to the coupling factors. The decay widths into massive gauge bosons ($V = W, Z$) or fermions are proportional to the g_{HVV} and $g_{Hf\bar{f}}$ couplings respectively. Thus, the Higgs boson will decay in most of cases to heavy particles such as pairs of electroweak gauge bosons ($W^{\pm}$, Z) and into pairs of quarks and lepton (b, τ, μ), as illustrated in figure 1.7.



(a) (b)

Figure 1.7.: The Higgs boson decays to W and Z bosons (a) and to fermions (b).

The Higgs boson does not couple to massless particles, therefore the decay modes in two photons or two gluons are induced through heavy particle loops, as show in figure 1.8.



(a) (b) (c)

Figure 1.8.: The $H \to \gamma\gamma$ decay mediated by heavy quark loops (a) and W boson (b) ,(c).

The decay branching ratios of the SM Higgs boson are shown in table 1.6.

Decay channel	Branching ratio [%]
$H \to bb$	58.2
$H \to WW$	21.4
$H \to gg$	8.19
$H \to \tau\tau$	6.27
$H \to cc$	2.89
$H \to ZZ$	2.62
$H \to \gamma\gamma$	0.227
$H \to Z\gamma$	0.153
$H \to \mu\mu$	0.022

Table 1.6.: Branching ratios of the Higgs boson with a mass of 125 GeV [17].

1.3. Summary

The Standard Model of elementary particle physics is a very successful theory which provides a successful description of the strong, weak, and electromagnetic interaction between the known elementary particles. The last missing piece of the SM, the Higgs boson, was directly observed by ATLAS and CMS experiments at the LHC in July 2012. The SM Higgs boson was observed in different channels. However, the fermionic decay modes (e.g. $H \to b\bar{b}$) are not yet confirmed as precisely as the bosonic decay modes. Precise measurements of its properties (i.e. mass, spin / CP and couplings) are very important to investigate for possible deviations from the SM. Due to its large mass the top quark coupling to the Higgs boson is the largest among fermions in the SM. Indirect constrains of the top Yukawa coupling were published in Run 1 [18] using a Higgs gluon fusion production and $H \to \gamma\gamma$ decay. The associated production of a Higgs boson with a top quark pair (ttH channel) is the only way for a direct measurement of the top Yukawa coupling at the LHC. Therefore the ttH observation and the top Yukawa coupling measurement will be highlights of Run 2 and one of the main subjects of this thesis.

2. The ATLAS experiment

The ATLAS experiment is a general purpose particle physics experiment at the Large Hadron Collider (LHC) at CERN. It investigates a wide range of physics, from the search for the Higgs boson to new physics in proton collisions at very high energy. The CMS experiment, at the other side of the LHC ring has the same physics programme.

A brief presentation of CERN and its chain of accelerators and experiments are given in section 2.1. An overview of the sub-detectors of the ATLAS detector will be presented in section 2.2. The ATLAS trigger system and data processing are summarised in section 2.3 and section 2.4 respectively. The object reconstruction and identification are described in section 2.5.

2.1. The Large Hadron Collider

2.1.1. CERN

The European Organization for Nuclear Research (CERN) is located at the French-Swiss border near Geneva. The name is derived from the french acronym *Conseil Europeen pour la Recherche Nucleaire* and was founded in 1954 with 12 member states. There are now 21 member states mainly from European countries. CERN employs around 2500 people, scientific and technical staff and 12000 visiting researchers from more than 70 countries working with the CERN facilities. These scientists represent a large community of 120 different nationalities and over 600 universities.

2.1.2. The LHC machine

The LHC [19] project was approved by the CERN Council in December 1994 to replace the Large Electron–Positron (LEP) collider machine. The LHC is a hadron accelerator and collider installed in the existing 27 km long tunnel previously constructed to host the LEP ring. The tunnel lies between 45 m and 170 m below the surface on a plane inclined at 1.4 % .

The aim of the LHC and its experiments is to test the Standard Model or reveal the physics beyond the Standard Model. In order to achieve theses goals, it was decided that the LHC machine would accelerate protons[a] to centre of mass collision energies of 14 TeV.

[a] The LHC also collides lead-ions over one month per year as part of the diverse research programme.

The number of events per second generated in the LHC collisions is given by:

$$N_{\text{event}} = L\sigma_{\text{event}},$$

where σ_{event} is the cross section of the process studied and L the luminosity which depends only on the LHC machine parameters and on the configuration of the magnets in the proximity of the experiments, mainly quadrupoles, which have to focus the beams into the point where the collisions takes place. The luminosity can be written as:

$$L = \frac{N_b^2 n_b f_{\text{rev}} \gamma_r}{4\pi\epsilon_n \beta^*} F,$$

where N_b is the number of particles per bunch, n_b is the number of bunches per beam, f_{rev} is the revolution frequency, γ_r is the relativistic gamma factor, ϵ_n is the normalised transverse beam emittance, β^* is the beta function at the collision point, and F is the geometry luminosity reduction factor due to the crossing angle at the interaction point (IP):

$$F = (1 + (\frac{\theta_c \sigma_z}{2\sigma^*})^2)^{-1/2}$$

θ_c is the full crossing angle at the IP, σ_z is the root mean square (RMS) of the bunch length distribution, and σ^* is the RMS of the transverse beam size at the IP.

The design specifications of the LHC are shown in table 2.1. The LHC running conditions for Run 1 and Run 2 period are summarised in table 2.2.

Beam particle	Protons
Injected beam energy	0.45 TeV
Nominal beam energy	7 TeV
Number of dipole magnets	1232
Max dipole field	8.3 T
Luminosity	10^{34} cm^{-2}s^{-1}
Particles per bunch	1.1×10^{11}
Number of bunches	2808
Bunch spacing	25 ns

Table 2.1.: Design parameters of the LHC.

2.1.2.1. Accelerator and Energy

The CERN accelerator complex is illustrated in Figure 2.1. The first stage in the acceleration is linear, LINAC II strips the electrons from hydrogen atoms to produce protons, which are then linearly accelerated to approximately one third of the speed of light. Then, the protons are injected into the Booster, a small synchrotron. The protons are divided into 4 bunches and circularly accelerated via a pulsing electric field. The Booster

Parameter	Run 1 2012	Run 2 2015	Run 2 2016
Beam energy [TeV]	4	6.5	6.5
Bunch spacing [ns]	50	50 - 25	25
Max. number bunches	1380	2244	2064
Protons per bunch [10^{11}]	1.6	1.15	∼1.2
Beta* [cm]	60	80	40
Peak luminosity [$cm^{-2}s^{-1}$]	8×10^{33}	0.5×10^{34}	$\sim 1 \times 10^{34}$
Collisions per bunch crossing (mean)	21	15	∼25

Table 2.2.: The LHC running conditions during Run 1 and Run 2.

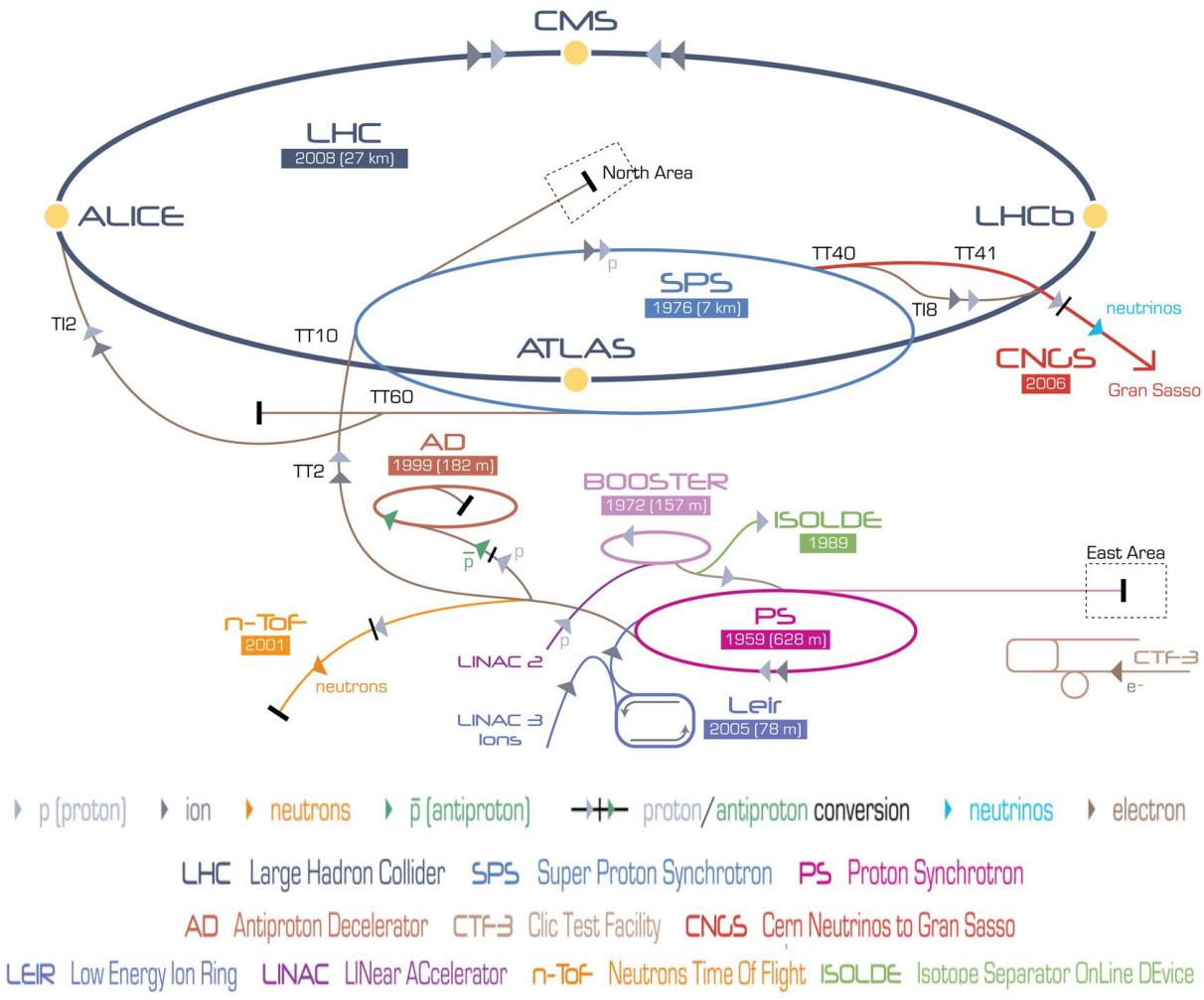


Figure 2.1.: CERN accelerator complex (not to scale), showing the injection system and the component's date of construction.

has 157 m in circumference. It accelerates the protons to $0.916c$ (c denotes the speed of light in vacuum). The third stage of acceleration is the Proton Synchrotron, a circular accelerator with a circumference of 628 m increases the protons to $0.999c$ in 1.2 seconds. The final stage of acceleration before injection into the LHC is the Super Proton Synchrotron with a circumference of 7 Km. At this step the proton beam is separated in two parts to be injected in a counter-rotating configuration in the LHC. The energies reached by the protons at the end of each accelerator are:

- Proton LINear ACcelerator (LINAC): Up to 50 MeV
- Proton Synchrotron Booster (PSB): 1.4 GeV
- Proton Synchrotron (PS): 26 GeV
- Super Proton Synchrotron (SPS): 450 GeV
- LHC: 7 TeV

The maximum beam energy that the LHC can deliver depends strongly of the magnetic field of the dipole magnets needed to keep the particle along the trajectory. The use of superconducting dipoles, shown in Figure 2.2, must supply a magnetic field of 8.3 T which corresponds to a beam energy of 7 TeV. The magnets need to be cooled down to a temperature of 1.9 K.

Almost the same chain of successively energetic accelerators is used to accelerate heavy lead ions Pb^{82} to an energy of 574 TeV which corresponds to a centre of mass energy of 2.76 TeV/nucleon in Pb-Pb collisions.

The accelerator tunnel comprises eight straight sections and eight arcs, as shown in Figure 2.3. The tunnel contains the two rings which produce two counter-rotating particle beams colliding at Points 1, 2, 5 and 8. The four main detectors are built around these points. The beam is accelerated using superconducting Radio-Frequency (RF) cavities, located in the straight section at Point 4, which provide RF energy to the beams and keep the bunches tightly bunched to ensure optimal condition at the collision point. The Points 3 and 7 contain beam collimation systems which shape and clean the beam. The straight section in Point 6 is used as the beam dump, where the beams are removed from the LHC and "dumped" into a graphite target to dissipate the beam's energy.

The arcs are built using a total of 1232 superconducting dipole magnets which keep the beams in the (nearly) circular orbit. Additionally, there are 392 quadrapole magnets, located in the straight sections, which serve to focus the beam.

2.1.2.2. The Experiments on the LHC

The experiments installed on the LHC ring are briefly described below:

ALICE *(A Large Ion Collider Experiment)* [20]: designed with the intention of studying the quark gluon plasma that results from the intense temperatures generated during

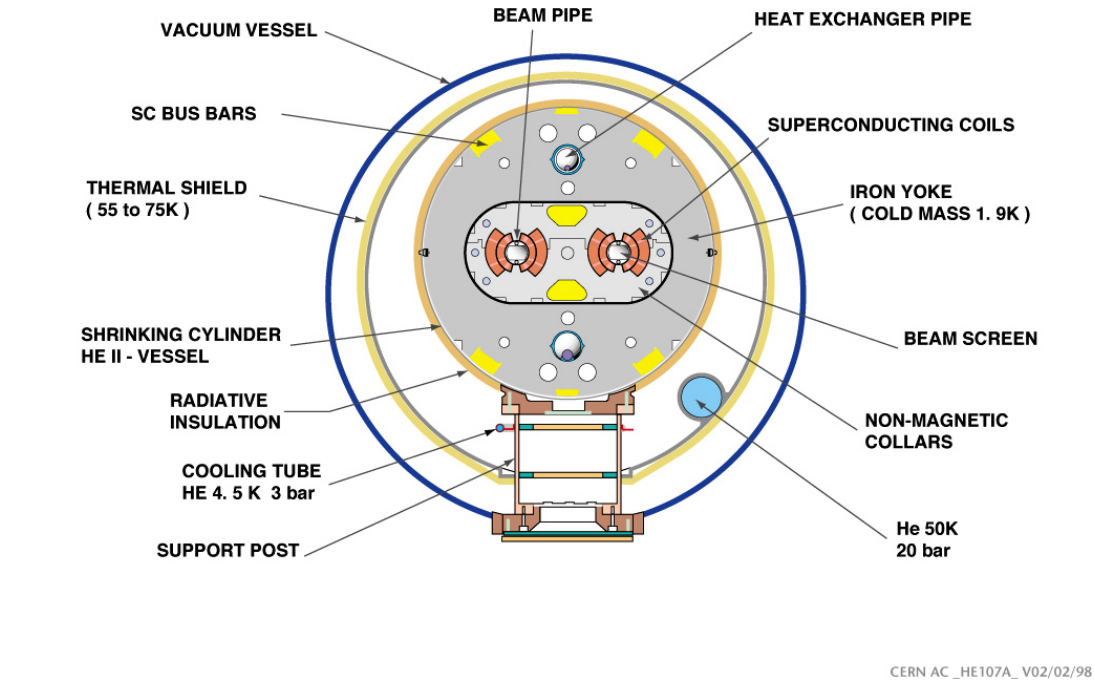


Figure 2.2.: Cross-section of the LHC superconducting dipole.

the heavy ion collisions. Design considerations of ALICE have been made with the ability to cover a large phase space and to detect hadrons, leptons and photons.

ATLAS *(A Toroidal LHC Apparatus)* [21]: the ATLAS detector is the largest detector in operation at LHC. Its design philosophy was to create a detector with the ability to detect the full range of masses allowed for the Higgs boson while retaining the ability to detect the known SM particles such as heavy quarks and gauge bosons.

CMS *(Compact Muon Solenoid)* [22]: The CMS detector has the same research prospect as the ATLAS experiment. CMS is built with a strong superconducting magnetic field of 4 T to collect the maximum energy from the particles. CMS has a very compact design of 12500 tonnes of material.

LHCb *(Large Hadron Collider beauty)* [23]: LHCb is a single-arm spectrometer with a forward angular coverage from approximately $\pm$15 mrad to $\pm$300 mrad in the bending plane. In terms of pseudo-rapidity the acceptance is $1.9 < \eta < 4.9$. The LHCb experiment has as main purpose study the CP violation and the physics of decay in the B-meson system. The geometry is influenced by the fact that both b and $\bar{b}$ hadrons are created in the same forward (or backward) cone. LHCb has excellent particle identification and vertex resolution necessary for the study of rapidly oscillating B mesons.

LHCf *(Large Hadron Collider forward experiment)* [24]: It is the smallest of all the LHC experiments. Its aim is to study the particles generated in the forward region of collisions, to verify hadronic models at very high energy for the understanding of ultra-high energetic cosmic rays. It consists of two small detectors, 140 m on either side of the ATLAS intersection point.

TOTEM [25]: The TOTEM experiment measures the total pp cross-section and study elastic scattering and diffractive dissociation at the LHC. TOTEM also aims to measure the luminosity at the CMS interaction point where it is based. It covers the very forward region in the pseudo-rapidity range.

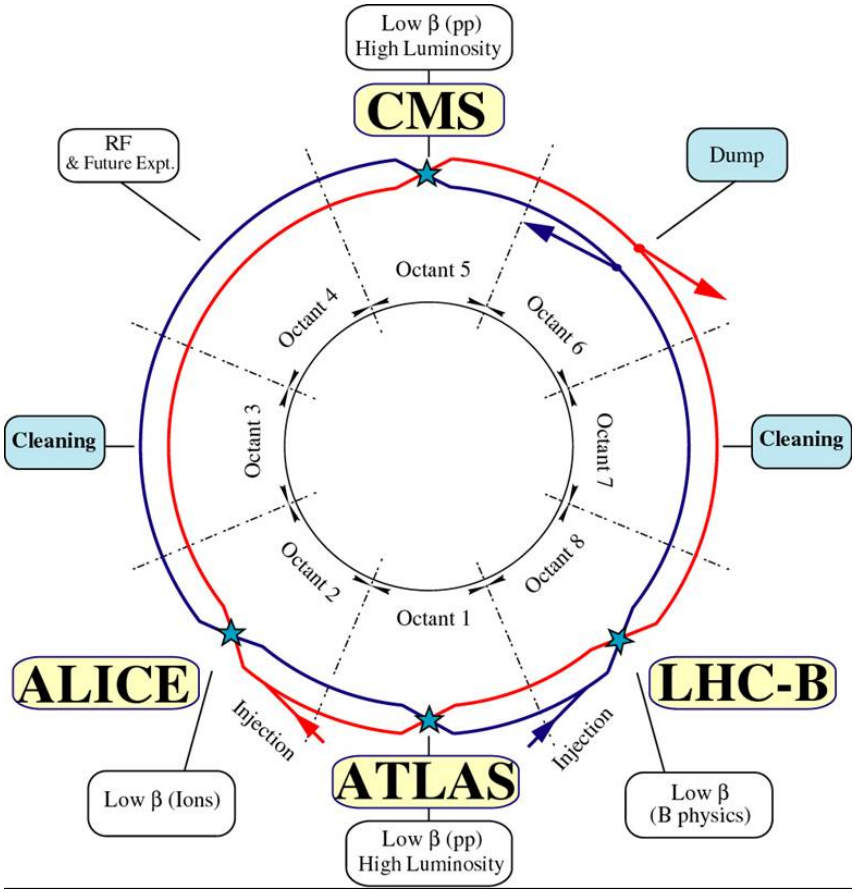


Figure 2.3.: Schematic of the LHC ring showing the four interaction points.

2.2. The ATLAS Detector

The ATLAS detector is a general purpose particle physics experiment. It is designed to achieve the maximum coverage in solid angle around the interaction point. This is realised by several layers of active detector components around the beam axis (barrel) and perpendicular to the beam axis in the forward regions (endcaps). The ATLAS detector consists of four major components, the Inner Detector which measures the momentum of the charged particles, the Calorimeter which measures the energies carried by the particles, the Muon spectrometer which identifies muons and the Magnet system that bends charged particles for momentum measurement. Figure 2.4 show an overview of the ATLAS detector, including all subdetectors and the magnet systems (one solenoid and three air-core toroids). Table 2.3 lists the design performance of the ATLAS detector.

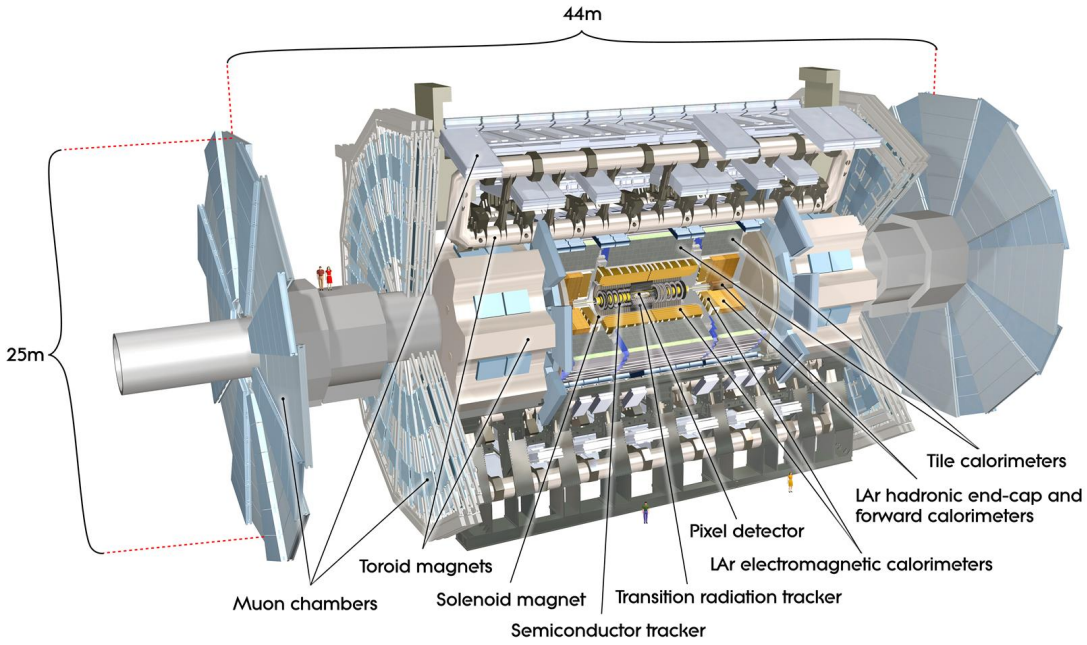


Figure 2.4.: The ATLAS detector and its components.

The ATLAS Coordinate System

The ATLAS Coordinate System is a right-handed system with the x-axis pointing to the centre of the LHC ring, the z-axis following the beam direction and the y-axis going upwards. The azimuthal angle ϕ is defined with respect the beam axis in the x-y plane. ϕ is measured in the range $[-\pi, \pi]$. The polar angle θ is measured from the positive z axis. The pseudorapidity, η, is defined by

$$\eta = -\log(\tan\frac{\theta}{2}),$$

Detector component	Resolution	η coverage measurement
Tracking	$\sigma/p_T = 0.05\%\ p_T \oplus 1\%$	$\lvert\eta\rvert < 2.5$
EM calorimetry	$\sigma/E = 10\%\ \sqrt{E} \oplus 0.7\%$	$\lvert\eta\rvert < 3.2$
Hadronic calorimetry		
- barrel and end-cap	$\sigma/E = 50\%\ \sqrt{E} \oplus 3\%$	$\lvert\eta\rvert < 3.2$
- forward	$\sigma/E = 100\%\ \sqrt{E} \oplus 10\%$	$3.1 < \lvert\eta\rvert < 4.9$
Muon spectrometer	$\sigma/p_T = 10\%$ at $p_T = 1$ TeV	$\lvert\eta\rvert < 2.7$

Table 2.3.: Design performance and coverage of the ATLAS subdetectors [21].

and the distance of objects in the pseudorapidity-azimuthal angle space is defined as:

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$$

2.2.1. The Inner Detector

The ATLAS Inner Detector (ID) provides charged particle tracking with high efficiency over the pseudorapidity range of $|\eta| < 2.5$. The ID consists of three independent but complementary sub-detectors. Figure 2.5 shows a cutaway view of the barrel ID. All the sub-detectors allow precision measurement of charged particle trajectories in an environment of numerous tracks: the Insertable B-Layer (IBL) and the Pixel detector mainly contribute to the accurate measurement of vertices, the silicon microstrip (SCT) measures precisely the particle momentum, and the transition radiation tracker (TRT) enhances the pattern recognition and improve the momentum resolution, with an average of 36 hits per track. The TRT contributes also to electron identification complementary to the calorimeter over a wide range of energies. Figure 2.6 illustrates in more details the sub-detector layers in the barrel and end-cap regions. The ID is immersed in a 2 T axial magnetic field generated by the central solenoid, which extends over a length of 5.3 m with a diameter of 2.5 m.

2.2.1.1. The Pixel detector

The Pixel detector is the innermost part of the ID, originally it was a three layers system. In 2014 a fourth innermost layer, the IBL described in the next section, was installed for Run 2. A Pixel sensor or module is a 16.4 $\times$ 60.8 mm wafer of silicon with 46080 individual channels called pixels of 50 $\times$ 400 micros each. A Pixel module comprises an un-packaged flip-chip assembly of 16 front-end electronics chips bump bonded to a sensor substrate. There are 1744 modules in the pixel detector with a total of more than 80 millions detection units. A cylinder of 1.4 m long and 0.5 m in diameter centred

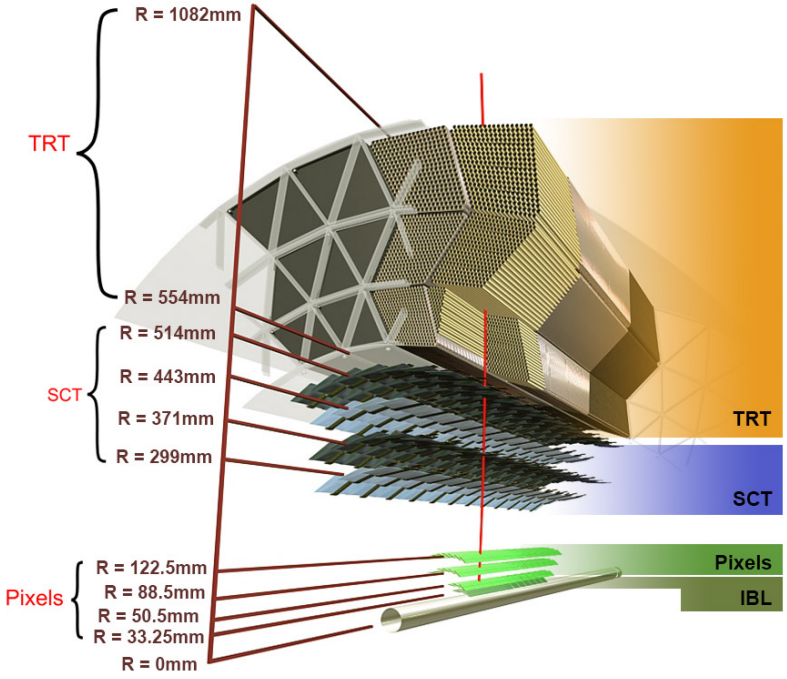


Figure 2.5.: The ATLAS Inner Detector for Run 2.

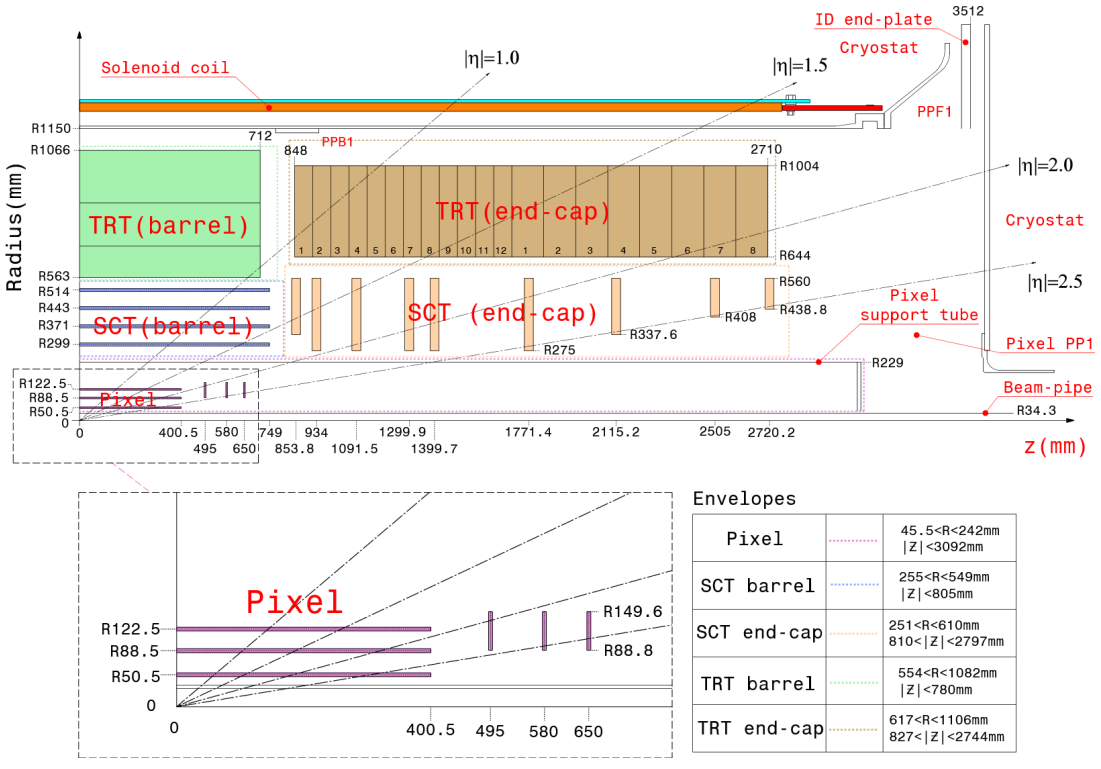


Figure 2.6.: A quarter section of the ATLAS Inner Detector view in r-z plane. The lower part shows a zoom of the pixel region.

on the interaction point supports the active parts. The barrel part of the pixel detector consist of the 3 cylindrical layers with radial positions of 50.5 mm, 88.5 mm and 122.5 mm respectively. There are 22, 38 and 52 staves in each of these layers respectively. Each stave is composed of 13 pixel modules. The staves are mounted with a tilt angle of 20^o to form a layer, this geometry allows overlaps between the modules.

The two pixel end-caps each have three identical disks perpendicular to the beam axis. Each of the disks consist of 8 sectors. Six pixel modules are directly mounted on each sector. The modules are rotated with a tilt angle of 7.5^o to ensure overlap between modules.

The intrinsic measurement accuracies of the pixel detector in the barrel are 10 μm (R-ϕ) and 115 μm (z) and in the disks are 10 μm (R-ϕ) and 115 μm (R). The Pixel detector is designed to measure 4 hits per track in the barrel region and 5 hits per track in the endcaps. The initial Run 1 Pixel detector design allowed to measure only 3 hits per track in the barrel region. The IBL adds an additional hit.

2.2.1.2. The Insertable B-Layer

The Insertable B-Layer (IBL) is the fourth layer added to the Pixel Detector between a new beam pipe and the inner Pixel Detector (B-layer). It consists of 14 tilted staves which are 64 cm long, 2 cm wide and tilted in ϕ by 14^o, equipped with 32 front-end chips per stave and sensors facing the beam pipe over the range of $|\eta| < 2.5$. The inner radius of IBL is 31 mm with an outer radius of 38.2 mm while the sensor are present at an average radius of 33.4 mm. The IBL sensors have 50 $\times$ 250 micron pixels adding an additional 12 million pixels to the pixel system.

The performance of the IBL is critical to the full realisation of the physics capabilities of the ATLAS experiment. The addition of the IBL provides improved precision for vertexing and b-tagging (identification of jets originating from bottom quarks). The improvement in the b-tagging performance due to the addition of the IBL and the algorithmic updates can be found in ref. [26].

2.2.1.3. The SemiConductor Tracker (SCT)

The SCT consist of four cylindrical layers in the barrel region and 9 disks at each end of the barrel (endcap). The barrel SCT consist of 2112 rectangular shape modules. A module is constructed with four rectangular planar p-in-n silicon strip sensors which have a thickness of 285 μm and 768 effective strips with pitch of 80 μm. The four sensors, two of each on the top and bottom side, are rotated with their hybrids by $\pm$ 20 mrad around the geometrical centre of the sensors. They are glued on a 380 μm thick thermal conductive mechanical support. A barrel module with its components is shown in Figure 2.7. The two 768 strip sensors on each side form a 128 mm long unit. The endcap SCT consist of 1976 trapezoidal shape modules placed on 18 endcaps disks, using 4 types of modules which were placed in three rings named as outer, middle and inner on disks. The endcaps modules were constructed in the same manner as the barrel. The strip pitch is varied from 56.9 to 90.4 μm. The intrinsic accuracies per

module in the barrel are 17 μm (R-ϕ) and 580 μm (z), while in the endcap region they are 17 μm (R-ϕ) and 580 μm (z). The total number of readout channels in the SCT is approximately 6.3 millions. The SCT detector is design to measure 8 hits per track in the central region and 9 hits per track in the endcaps.

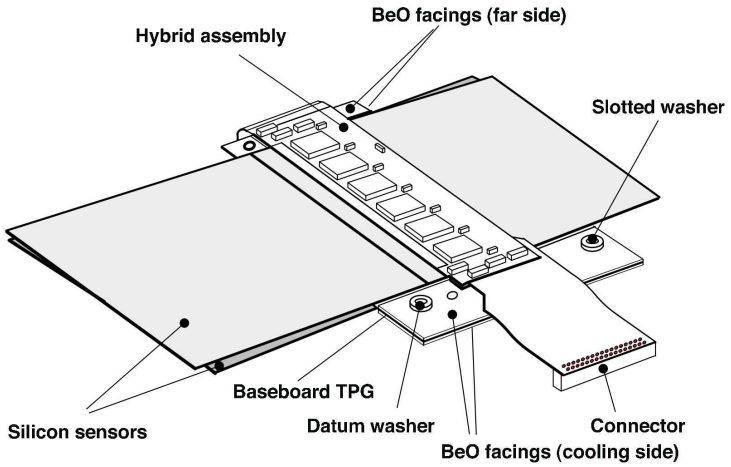


Figure 2.7.: A SCT barrel module. The thermal pyrolytic graphite (TPG) provide a high thermal conductivity path between the coolant and the sensors.

2.2.1.4. The Transition Radiation Tracker (TRT)

The Transition Radiation Tracker (TRT) itself is subdivided into two sections, the TRT barrel ($|\eta| < 1.0$) and the TRT end-caps ($1.0 < |\eta| < 2.0$). The TRT barrel has the sensor layers running parallel to the beam axis, while the sensor layers of the end-cap TRT are radially oriented.

The TRT is based on straws, which in case of the barrel are 144 cm long. They are electrically separated into two halves at $|\eta| = 0$ and arranged in a total of 73 planes. The end-cap straws are 37 cm long, radially arranged in wheels with a total of 160 planes. The straws themselves are polyimide tubes with a diameter of 4 mm. Its wall is made of two 35 μm thick multi-layer films bonded back-to-back to forms the cathode. The straw wall is held at a potential of -1530 V. The anodes are 31 μm diameter gold-plated tungsten wires. They are directly connected to the front-end electronics and kept at ground potential. The straws are operated with a gas mixture of $Xe/CO_2/O_2$(70:27:3). To maintain straw straightness in the barrel, alignment planes made of polyimide with a matrix of holes are positioned each 25 cm along the z-direction of the module.

The TRT operates as a drift chamber: when a charged particle traverses the straw, it ionises the gas, creating about 5-6 primary ionisation clusters per mm of path length. The electron drift towards the wire and they cascade in the strong electric field very close to the wire, thus producing a detectable signal.

As mentioned above the TRT plays a central role for electron identification, cross-checking and complementing the calorimeter. TRT provides substantial discriminating power between electron and pions over the energy range between 1 and 200 GeV.

Typically, the TRT provides 36 hits per track with a precision of about 140 μm in the bending direction.

2.2.2. The calorimeters

The ATLAS calorimeters cover the pseudorapidity range $|\eta| < 4.9$. The design has been guided by the benchmark process of a Higgs boson decaying to two photons, $H \to \gamma\gamma$. For such a physics search the calorimeter must have excellent photon resolution, with uniform photon measurement and good γ/π discrimination across the entire calorimeter. The overview of the ATLAS calorimeter system is illustrated in Figure 2.8. Different technologies are used across different regions in η. Surrounding the inner detector the EM calorimeter is finely segmented for precision measurements of electrons and photons, while the rest of the calorimeter is segmented more coarsely, since it is mainly aimed at reconstructing jets and measuring the missing transverse momentum.

The depth of the calorimeter is important to provide good containment for electromagnetic and hadronic showers and must also limit punch-through into the muon system. The total thickness of the EM calorimeter is $>$ 22 radiation lengths (X_0) in the barrel and $>$ 24 X_0 in the endcaps. The approximately 10 interaction lengths (λ) both in the barrel and in the end-caps are adequate to provide good resolution for high energy jets. The total thickness, including the outer support, is 11 λ at η=0 and has been shown by simulation and measurements to be sufficient to reduce punch-through into the muon system well below the irreducible level of prompt or in-flight decays muons.

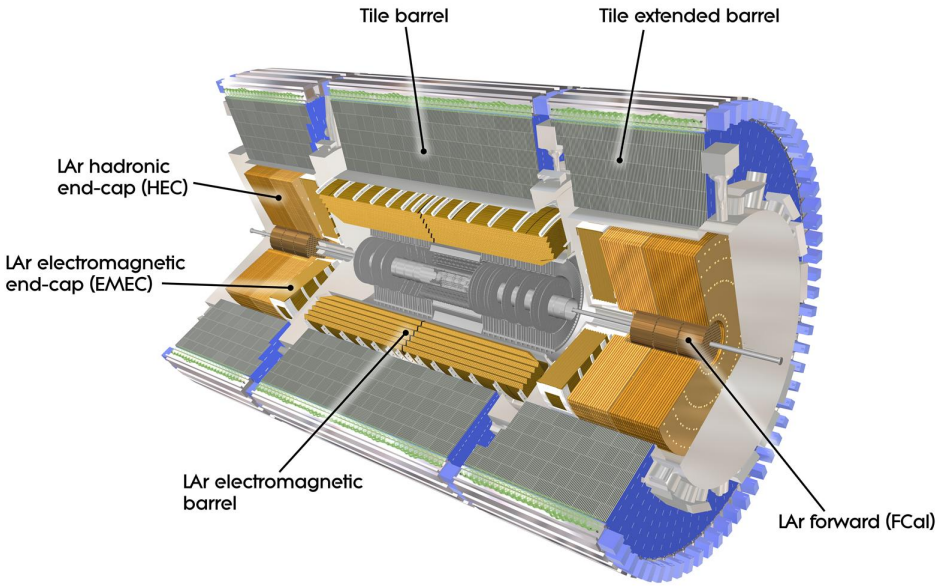


Figure 2.8.: Schematic layout of the ATLAS calorimeters

2.2.2.1. The electromagnetic calorimeter

The Electromagnetic Calorimeter is based on a highly granular liquid-argon technology (LAr). LAr is also used in the end-caps of the Hadron Calorimeter. Both detector elements share the cryostat at the end-cap, which also accommodates a special LAr forward calorimeter. The design is an novel arrangement of the absorber plates and electrodes which are arranged with the "accordion" geometry, with a total of ∼174000 readout channels.

It comprises a barrel section, made of two identical half-barrels, together covering the central pseudorapidity range, $|\eta| < 1.475$ and two endcaps, each covering a region $1.375 < |\eta| < 3.2$. In addition, there is a forward combined electromagnetic/hadronic liquid argon calorimeter at each end, covering the region $3.2 < |\eta| < 4.9$.

In front of the barrel and part of the endcaps, for $|\eta|$<1.8, there is a 10 mm thick presampler to provide an estimation of energy lost in dead material in front of the calorimeter. In the barrel region, the material budget in front of the detector, associated with the solenoid and the tracker, varies from $\sim 2X_0$ at η=0 to 5–6X_0 for η from 1.5 to 1.8. In the endcaps the material budget is $\sim 2.3X_0$.

The liquid argon was chosen as an active medium because it offers an intrinsically linear response which is stable over time and tolerant to high levels of radiation. The accordion geometry provides high granularity and good hermeticity. The readout is at the front and back of the calorimeter, rather than at the sides, which means that adjacent modules can be tightly packed, with full ϕ coverage and no cracks between modules. A section of the barrel calorimeter is shown in Figure 2.9(a).

The energy deposited in the calorimeter is reconstructed by summing the calibrated cell energies in the three sampling layers: the strip layer, the middle and back layers, together with the energy in the presampler for a cluster of cells built around the cell with the largest energy deposit in the middle layer. The layout of the barrel is shown in Figure 2.9(b). The relative energy resolution as a function of energy has been measured for a set of barrel modules in a test beam, with electrons in the energy range 10–245 GeV. The resolution achievable is:

$$\frac{\sigma(E)}{E} = \frac{a}{\sqrt{E(GeV)}} \oplus b,$$

where $a = 10 \pm 0.1\%$ is the stochastic term and $b = 0.17 \pm 0.04\%$ is the constant term. Similar results have been obtained for the endcaps, which satisfy the calorimeter design specification.

2.2.2.2. The Hadronic calorimeter

The ATLAS hadronic calorimeters are divided in: the tile hadronic calorimeter (TileCal), the liquid-argon hadronic end-cap calorimeter (HEC) and the liquid-argon forward calorimeter (FCal), as shown in figure 2.8.

The TileCal is divided into three parts: a barrel, covering the region $|\eta| < 1.0$, and two extended barrels on each side, covering the range $0.8 < |\eta| < 1.7$. Radially, the

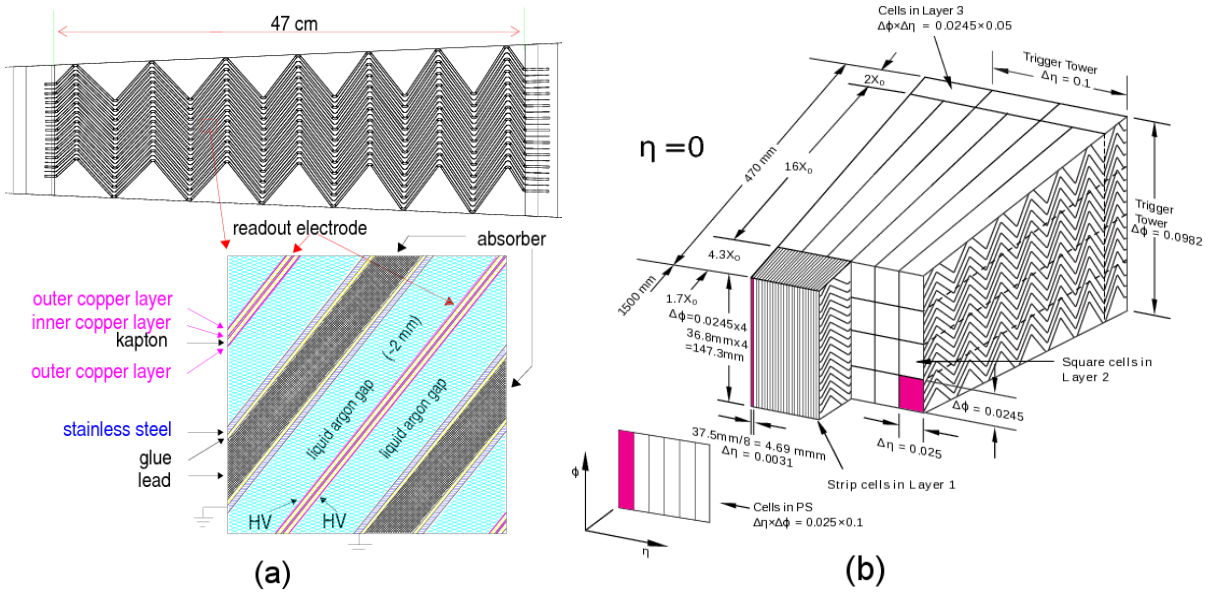


Figure 2.9.: The barrel accordion calorimeter. (a) It is made of succession of lead absorbers and copper electrodes, with gaps of Liquid Argon (LAr) in between. A close-up shows the consecutive layers of absorber. (b) A group of barrel calorimeter cells grouped into readout towers. The fine granularity of the strip towers in Layer 1 improves the γ/π^0 discrimination.

tile calorimeter extends from an inner radius of 2.28 m to an outer radius of 4.25 m. The barrel (and extended barrels) are segmented into 64 azimuthal sections, referred as modules, subtending $\Delta\phi=2\pi/64 \sim 0.1$.

It is a sampling calorimeter using steel as the absorber and scintillator tiles as the active material. The scintillator plates are oriented perpendicularly to the colliding beam axis, and are radially staggered in depth as schematically shown in Figure 2.10. By the grouping of wavelength shifting fibers to specific photo-multipliers (PMTs), modules are segmented in $|\eta|$ and in radial depth. In the direction perpendicular to the beam axis, it is segmented in three layers, approximately 1.5, 4.1 and 1.8 λ for the barrel and 1.5, 2.6 and 3.3 λ for the extended barrels.

The TileCal comprises 4672 readout cells, each equipped with two PMTs that receive light from opposite sides of every tile. The energy response to isolated charged pions of the combined LAr and tile calorimeter tested with test beam is $\frac{\sigma(E)}{E} = \frac{53\%}{\sqrt{E(GeV)}} \oplus 3\%$, close to design specifications.

The HEC, which covers the range $1.5<|\eta|<3.2$, are based on LAr technology. Similar to the EM calorimeter in the barrel region, but copper is used instead of lead as a passive absorber material and with a flat-plate design. The energy resolution to isolated pions is $\frac{\sigma(E)}{E} = \frac{71\%}{\sqrt{E(GeV)}} \oplus 1.5\%$.

The FCal, which covers the range up to $|\eta|=4.9$, use copper as absorber material for the first layer and tungsten for the second and third layer. As a result of test beam data

the energy response to pions is $\frac{\sigma(E)}{E} = \frac{94\%}{\sqrt{E(GeV)}} \oplus 7.5\%$.

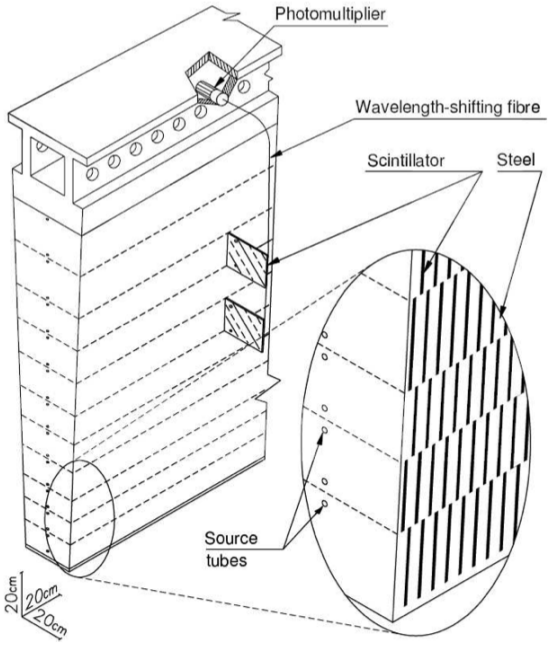


Figure 2.10.: The ATLAS TileCal module. The plastic scintillator tiles are read out from both sides with wavelength shifting fibers into separate PMTs. The staggered absorber/scintillator and the radioactive source tubes are shown on the right.

2.2.3. The muon spectrometer

The muon spectrometer is the outermost part of the ATLAS detector. It is designed to measure muon momentum in the region $|\eta| < 2.7$ and provides trigger information for $|\eta| < 2.4$. The muon momentum is determined by measuring the track curvature in the magnetic field. The magnetic field is provided by three superconducting air-core toroids, one in the barrel ($|\eta| < 1.1$) and one for each endcap ($1.1 < |\eta| < 2.7$), with a field integral between 2 and 8 T.m. The toroids have field lines around the beam axis, which are immersed in the coils thus complicating the instrumentation. A large number of coils is required to keep the field uniform. Each of the three toroids consists of eight coils assembled radially and symmetrically around the beam axis. In Figure 2.11 a schematic view of the ATLAS magnet system is shown.

The layout of the muon spectrometer is shown in Figure 2.12. The muon chambers are categorised into two sets: one for dedicated precision measurement of muon tracks and the second set dedicated for defining a muon trigger. The precision measurements are performed by two different chamber technologies: Monitored Drift Tube chambers (MDTs), covering all the range $|\eta| < 2.7$, except for the innermost layer of the endcap

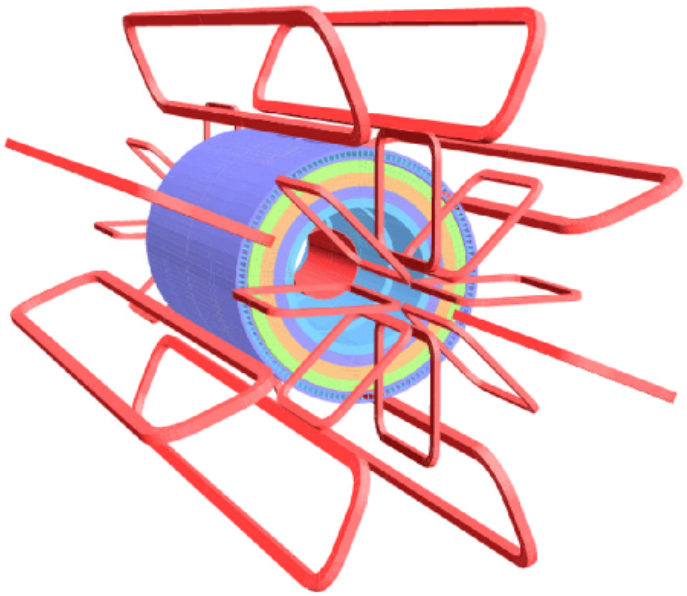


Figure 2.11.: Schematic view of the ATLAS solenoidal (inner cylinder) and toroidal magnets (outer coils).

regions ($2.0 < |\eta| < 2.7$) where Cathode Strip Chambers (CSCs) are installed due to their capability to cope with higher background rates.

The MDT chambers are composed of two multi-layers made of three or four layers of tubes. Each tube is 30 mm in diameter and has a tungsten anode wire of 50 μm diameter. The gas mixture used is 93% Ar and 7% CO_2, the drift velocity is not saturated and the total drift time is about 700 ns. The space resolution of one of the 350k tubes of the MDT is about 80 μm, measured in a test beam. The CSC chambers are multiwire proportional chambers with cathode planes segmented into strips in orthogonal direction. Typical resolution obtained with this scheme is about 50 μm in the R direction and a resolution of 10 mm in the ϕ direction.

The trigger system for muon events is based on the Resistive Plate Chambers (RPC) instrument in the barrel region while Thin Gap Chambers (TGC) are used in the higher background environment of the endcap region. These allow very good timing resolution, 1.5 ns for RCP's and 4 ns for TGC's, appropriate for triggering.

The main parameters of the muon chambers are listed in table 2.4

Muon chamber	Coverage	N^o of chambers	Function
Drift tubes (MDTs)	$\|\eta\| < 2.0$	1170	precision measurement
Cathode Strip Chambers	$2.0 < \|\eta\| < 2.7$	32	precision measurement
Resistive Plate Chambers	$\|\eta\| < 1.05$	1112	Triggering
Thin gap Chambers	$1.05 < \|\eta\| < 2.4$	1578	Triggering

Table 2.4.: Coverage and parameters for the ATLAS muon detectors.

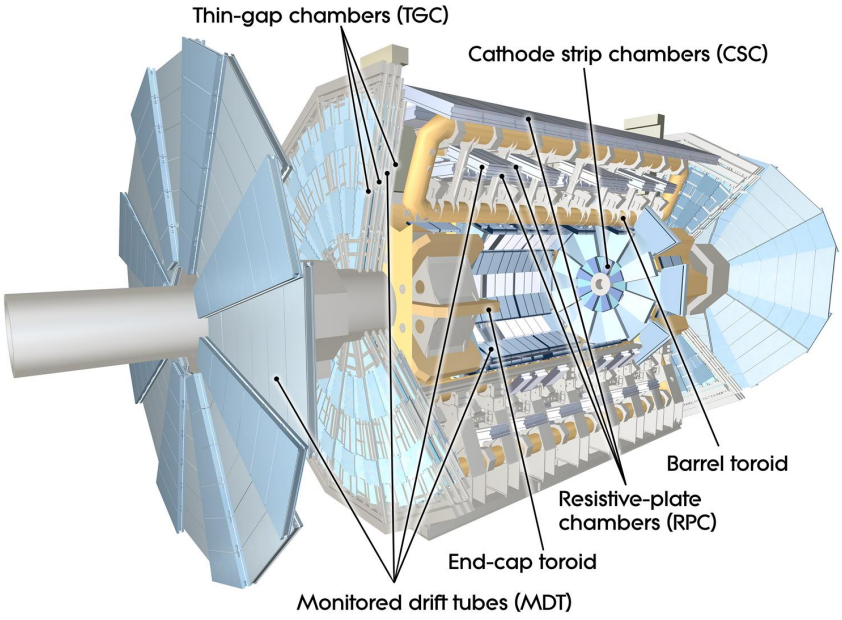


Figure 2.12.: The ATLAS Muon Spectrometer

2.3. The trigger system

The ATLAS trigger system operated very successfully during the Run 1 period [27]. The LHC running conditions for the Run 2 period are challenging for the trigger system. The increase of the beam energy, instantaneous luminosity, and collision frequency implies background rates higher than the Run 1 trigger was designed for. During the Long Shutdown 1 (LS1), there were many important changes and additions to the existing trigger and Data Acquisition (TDAQ) system [28].

The trigger system in Run 2 consists of a hardware Level-1 (L1) and a single software-based High-level trigger (HLT). This new two-stage system reduces the event rate from the bunch-crossing rate of 40 MHz to 100 kHz at L1 and to an average recording rate of 1 kHz at the HLT [29]. Figure 2.13 shows a schematic overview of the ATLAS TDAQ system for Run 2.

The trigger system is configured to use a large set of selection criteria for each event. Each criterion consists of sequential selections in different levels. An event has to satisfy at least one of the triggers in order to be recorded. A proposal of the trigger menu strategy for Run 2 is described in ref. [30].

Level-1

The Level-1 system performs the initial event selection based on information from calorimeters and muon detectors. In Run 2, the Level-1 system consists of the L1 calorimeter trigger system (L1Calo), the L1 muon trigger system (L1Muon), a new L1 topological trigger module (L1Topo) [31] and the Central Trigger Processors (CTP) [32].

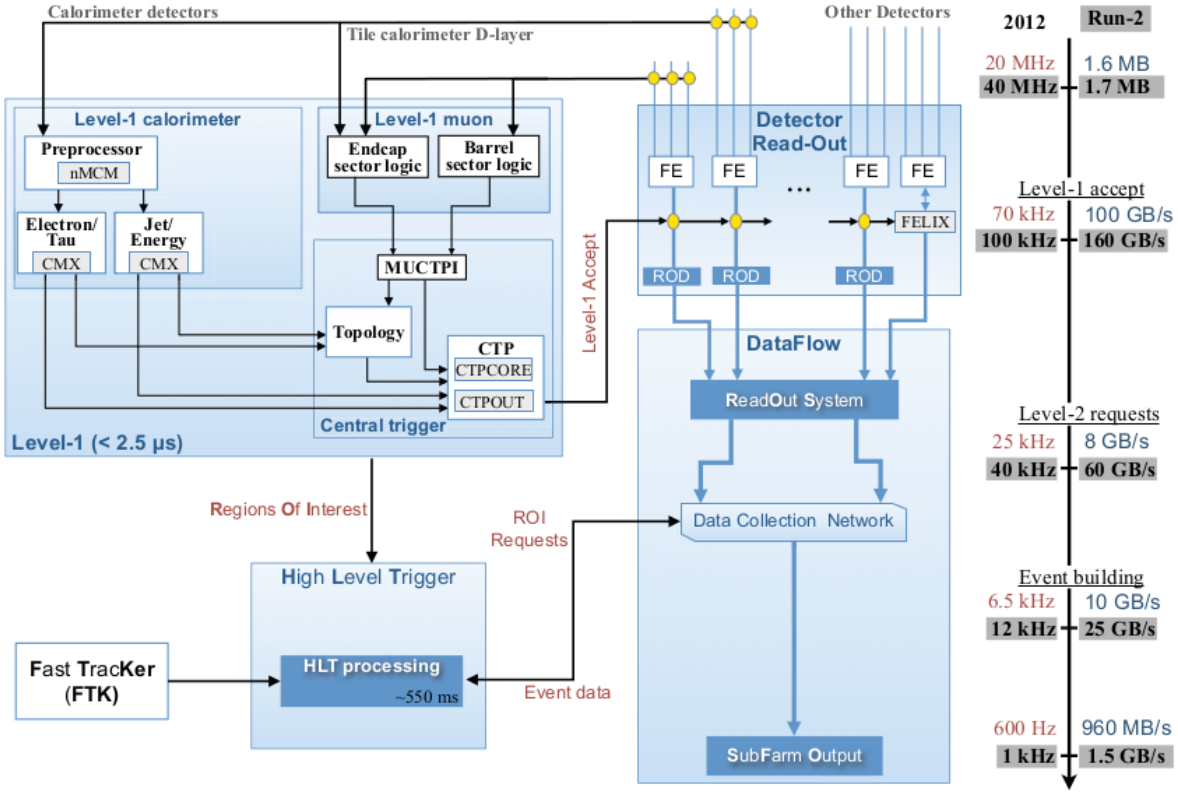


Figure 2.13.: Schematic overview of the Run 2 configuration of the Trigger and DAQ system [29].

The L1Topo system uses detailed information from L1Calo and L1Muon, processed to calculate topological information of the event, such as angles between trigger objects (i.e. electrons, photons, muons, jets, and taus) or the invariant mass of two or more trigger objects. The L1Topo allows a trigger decision to be made using more than just p_T or E_T whose thresholds would be impossible to maintain in Run 2 conditions. The final Level-1 accept decision is made in the CTP and distributes it together with the timing information to the subdetectors via a dedicated network.

Level-1 reduces the 40 MHz bunch-crossing rate to a rate of up to 100 kHz and find regions of interest (RoI) within a latency of 2.5 μs.

High-level trigger

During Run 1 the Level-2 trigger used Level-1 candidates and looked at more detailed physics properties to achieve a further reduction in rate to 2-3 kHz. A third level (event filter) used the full event information, and decided upon storage of the event for offline analysis with a final rate of 300-400 Hz. In Run 2, Level-2 and event filter farms are merged into a unique HLT farm for simplification and dynamic resource sharing. With merged HLT processing nodes, data needs to be requested only once from the read-out system PCs (ROS) hence saving network bandwidth and decreasing the ROS data request rate.

Significant improvements in the algorithms were implemented during the LS1. An initial fast reconstruction helps to reduce the event rate. The final online precision reconstruction is improved and uses offline-like algorithms as much as possible. To improve the performance, multivariate analysis techniques were introduced at the HLT. In particular the upgraded electron and photon trigger system and its performance is described in ref. [33].

Using the information from all subdetectors in a RoI or the full event information the HLT reduces the rate to 600 Hz to 1.5 kHz at peak luminosity within a processing time of 0.2 s on average.

2.4. Data processing

Data from the ATLAS detector selected with the trigger systems are converted into physics objects used in physics analyses through a process called reconstruction. The same reconstruction algorithms are run on both real and simulated data. LS1 has provided an opportunity to re-examine ATLAS reconstruction code. In particular, the object orientation of the Event Data Model (EDM) and the vector algebra library (CLHEP) were identified as a source of heavy CPU consumption that needed to be addressed. Eigen (a computer programming library for matrix and linear algebra operations) was selected as a replacement for CLHEP. About 1000 packages were updated during the EDM migration, not only leading to a significant speedup, but also allowing a significant reduction in the complexity of the code itself.

Processing data taken by the detector requires substantial offline computing which is supplied by Tier-0, Tier-1, and Tier-2 computing centres. The ATLAS prompt data reconstruction is performed at Tier-0. Tier-0 then distributes the data to Tier-1 centres around the world for further processing and analysis using the Event Summary Data (ESD) and Analysis Object Data (AOD) data formats. The large number of Tier-2 centres work in parallel to execute both data analysis and Monte Carlo event generation.

In Run 1, ATLAS followed a "frozen Tier-0" policy to ensure stability and reproducibility. This meant that the AOD files in T0 did not reflect the best available understanding of the running period, so each physics or performance group would first apply all the latest fixes to create a large private dataset for further analysis. In addition, the AODs were not ROOT-readable and a majority of groups created their own large ROOT-readable datasets to use.

In Run 2, the new environment is centred on a new AOD format known as xAOD. The new AOD has a completely redesigned EDM readable by both ROOT and the ATLAS software framework (Athena) [34] allowing full access to all objects. ATLAS Run 2 employs the "staged-Tier-0" policy that allows update of the T0 reconstruction software. The AOD to AOD reprocessing is done to apply improved calibrations and combined reconstruction techniques to the data already reconstructed. The reprocessing is performed using offline software tools, referred to as derivation framework. A standardised derivation framework is run for each analysis group needing a reduced dataset. The physics analysis groups can apply a group-specific reconstruction, skimming (selecting events), slimming (selecting objects), and thinning (dropping information) into their derivation framework. The derivation framework take full advantage of the advanced features of the new xAOD data format. Figure 2.14 show a schematic view of the data analysis model for ATLAS Run 2.

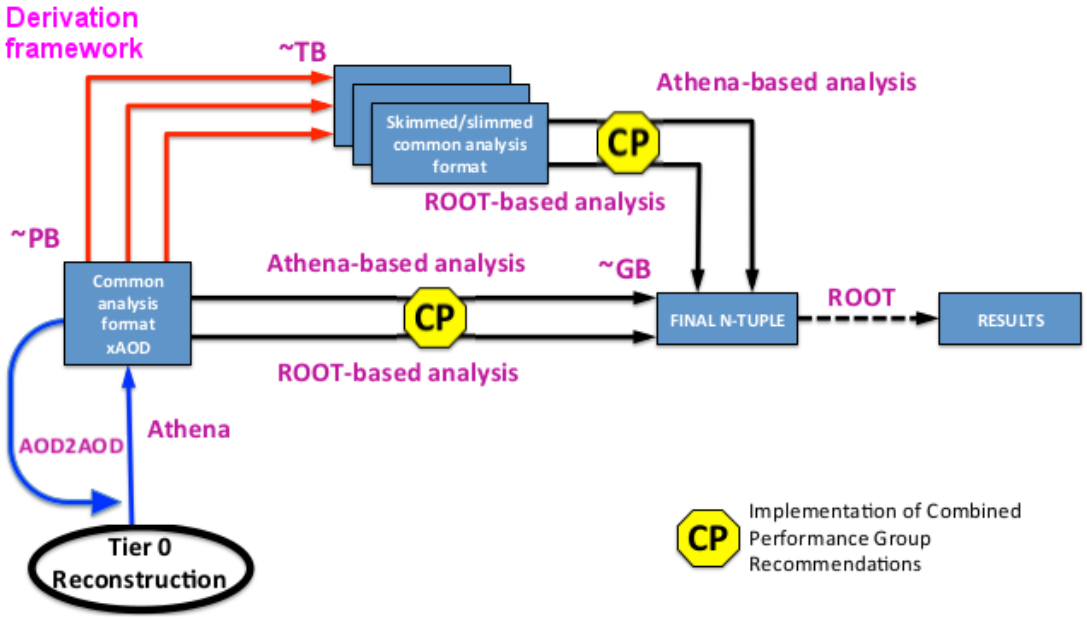


Figure 2.14.: ATLAS Run 2 analysis data flow [35].

As mentioned, the ATLAS reconstruction software was updated to the new EDM required for Run 2. In particular the author participated in the migration of the ATLAS b-tagging software to the new format, mainly the migration of the b-tagging EDM related to secondary vertices. This technical work included:

- Development of a new xAOD class (xAOD::BTagVertex) to be able to store JetFitter vertex information.
- Development of helper functions for easy access to xAOD vertex information which are needed for many taggers.
- Fully redesigned b-tagging interface to secondary vertex algorithms, including the functionalities to store xAOD vertices.

This work was done in several iterations and therefore with some validation work related to the b-tagging code migration to xAOD.

ATLAS reconstruction data flow

The ATLAS online cluster, involving High Level Trigger, produce event data in bytestream format (RAW data). The RAW data are processed in two steps within one job, producing first the ESD and then the derived AOD and Derived Event Summary Data (DESD) in the second step. The RAW data and the reconstruction outputs are exported to ATLAS Grid storage. There is also a small set of monitoring data in specialised file formats (ROOT files) which are produced in data (re)processing for specialised studies (e.g. data quality assessment) [35]. The data re-processing from RAW is performed at Tier-1. Figure 2.15 shows a schematic view of the reconstruction data flow.

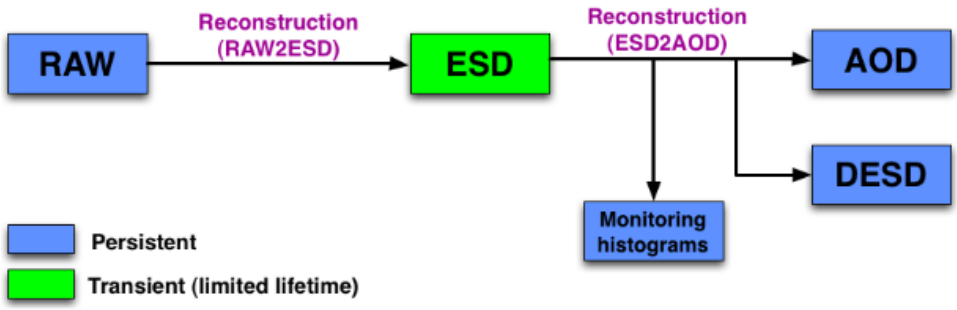


Figure 2.15.: ATLAS reconstruction data flow [35].

ATLAS Monte Carlo simulation flow

In order to simulate the physics processes with the experimental conditions, event generators are used to create a finite number of events. The event generator is carried out step by step through the hard process, parton shower and the hadronisation. Firstly in

the hard process, the main interaction between partons is considered. The cross section of the partonic process is computed explicitly at fixed order in perturbation theory, referred to as the matrix element calculation. Secondly in the parton shower process, the successive emission of photons, quarks and gluons from the partons in the final (or initial) state are generated using QED and QCD. Finally in hadronisation, bunches of particles are generated according to phenomenological models based on general features of QCD. The event generators[b] used in this thesis are PYTHIA 8 [36], ALPGEN [37], POWHEG [38], MadGraph5_aMC@NLO [39] and SHERPA [40]. The first two use leading-order (LO) matrix element while the last three use next-to-leading-order (NLO) matrix element.

The Monte Carlo generator programs produce EVNT files. The EVNT files are then processed to include the detector simulation, producing HITS files. In ATLAS, two approaches of detector simulation are developed: full simulation based on GEANT4 [41], computes the interaction between final state particles and the detector materials, and a less refined simulation, known as Atlfast-II (or AF2) [42].

The modelling of pileup is added in the next processing stage and the detector response (digitisation) is simulated at the same time, producing RDO files. As a separate step, the trigger response simulation is performed again producing RDO files with the simulated trigger information added. The rest of the reconstruction chain is the same as the prompt data reconstruction, producing the physical objects such as tracks, vertices, jets, etc. A schematic view of the Monte Carlo simulation flow is shown in Figure 2.16.

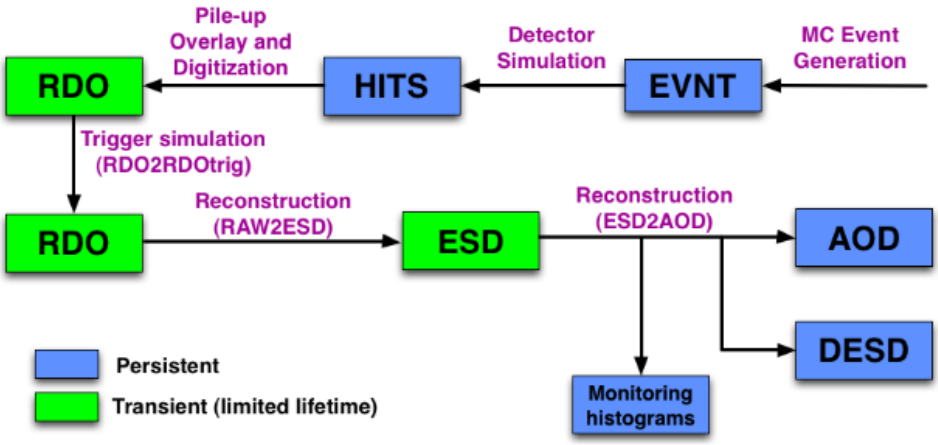


Figure 2.16.: ATLAS Monte Carlo simulation flow [35].

[b] Usually event generators are used in association with parton shower generators. e.g. the simulation of $t\bar{t}H$ is based on next-to-leading order calculations with MadGraph5_aMC@NLO interfaced with Pythia 8 for the modelling of the parton shower.

2.5. Event reconstruction

Physics events that pass the online trigger selection are processed to reconstruct basic quantities like vertices, tracks, and clusters. These quantities are combined to reconstruct and identify the final physical objects used in the analysis, such as electrons, muons, jets, b-jets and missing transverse energy. This section provides a summary of the reconstruction algorithms of the main physics objects used in this thesis.

2.5.1. Charged particle tracks and primary vertex

Charged particle tracks, commonly referred to just as tracks, and vertex finding are a very complex task and indispensable for any physics analysis. The ID track reconstruction consists of several sequences with different strategies. The main sequence is referred to a inside-out track finding. The main steps of the tracking algorithm are the following:

- First, in a pre-processing stage the raw data from the pixel and SCT detectors are converted into clusters, while the TRT raw timing information is calibrated into drift circular curves.
- Then the track finder starts with space points from the pixel layers and the first SCT layer to form track seeds. These are extended throughout the SCT to form track candidates. In this step outliers and fake tracks are rejected by applying quality cuts.
- Finally, the track candidates are extended into the TRT and left-right ambiguities in the association of the tracks to drift circles are solved. A final track fitting is performed with the full information to determine the track parameters.

The inside-out tracking sequence relies on a track seed found in the silicon detector. In the track reconstruction process, some of these initial track seeds may not be found or do even not exist : tracks coming from secondary decay vertices further inside the ID volume (e.g. long-lived particle decays, photon conversions). The sequence outside-in track reconstruction starts with a dedicated segment finding algorithm in the TRT and successive back tracking of the segments into the silicon detector.

To describe the track of a charge particle in a magnetic field, five helix parameters are needed as shown in figure 2.17. Parameters in $x-y$ plane are:

- Q/p_T: the electric charge over the transverse momentum, the relation Q/p_T is determined by the equation : $Q/p_T = (0.3BR_{curv})^{-1}$, where R_{curv} is the curvature radius of the track and B is the magnetic field.
- d_0: signed transverse impact parameter, distance from the beam axis to the point of the closest approach along the track in the transverse plane.
- ϕ: azimuthal angle at the point of the closest approach, defined in $[-\pi, \pi]$.

Parameters in $r - z$ plane:

- z_0: longitudinal impact parameter, defined as the z position of the track at the point of closest approach.
- θ: polar angle at the point of closest approach, defined in $[0, \pi]$.

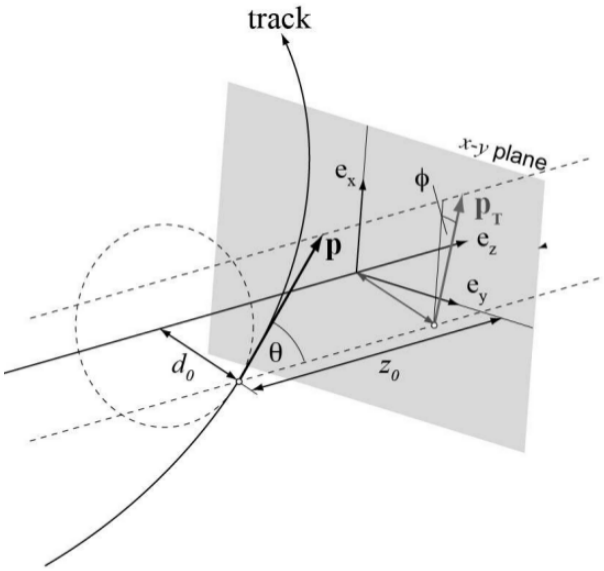


Figure 2.17.: Illustration of helix parameters of a charged track [43].

Track reconstruction performance of the ATLAS inner detector with $\sqrt{s} = 13$ TeV data, including the IBL detector are detailed in ref. [44] and [45].

Primary vertex

Primary vertex reconstruction is performed in two different stages [46]: vertex finding and vertex fitting. The vertex finding algorithms associates tracks into multiple vertex candidates. The vertex fitting algorithms, instead, get the best estimate of the vertex position refitting the associated tracks. A rough outline of the procedure is as follows:

- Vertex seeds are found by looking at the local maximum in the distribution of the z_0 of the tracks. An iterative method is used to find the most likely value.
- Tracks compatible with the seed are grouped together for fitting.
- Vertex fitting is performed by the adaptive fitting algorithm [47] to estimate the position and uncertainty of the vertex.
- Tracks that are not associated to a vertex are down-weighted rather than rejected and then used to repeat the process. Tracks which are incompatible with the vertex by more than 7σ are used to seed a new vertex.

- Finally, the list of vertices are ordered by the sum of the squared p_T of the tracks associated to the corresponding vertex ($\sum_{i=1}^{N_{trk}} p_{T,i}^2$). The vertex with the highest $\sum_{i=1}^{N_{trk}} p_{T,i}^2$ is assumed to be the main vertex of the event corresponding to the hardest proton-proton interaction. The others primary vertices are recognised as pileup vertices.

2.5.2. Jet reconstruction

Jets of particles are produced by the hadronisation of quarks and gluons. Hadrons form a spray of collimated particles that carry the momentum of the original parton. They are key ingredients for many physics measurements and searches for new phenomena

Jets considered in this thesis are reconstructed using the anti-k_T algorithm [48] with a radius parameter of R = 0.4. The input constituents to the jet algorithm are topological calorimeter clusters (topo-clusters) [49].

Topological clusters are groups of calorimeter cells that are designed to follow the shower development of a single particle interacting with the calorimeter. The topo-cluster formation algorithm starts from a seed cell , whose signal-to-noise (S/N) ratio is above a threshold of S/N = 4. Neighboring cells in the three dimensions of the seed (or the cluster being formed) that have a signal-to-noise ratio of at least S/N = 2 are included iteratively. Finally, all calorimeter cells with S/N > 0 in the perimeter to the formed topo-cluster are added. The topo-cluster algorithm effectively suppresses the calorimeter noise.

After the jet is reconstructed based on calibrated clusters, sequences of corrections are applied [50]: pile-up corrections which subtract the pile-up energy from the jet energy, changes of the jet direction due to the primary vertex which could be displaced from the origin of the reference frame, calibration of the jet energy using MC simulation (JES) and the data-to-MC differences are assessed using *in-situ* calibration.

Reconstruction and calibration of jets from the calorimeter are sensitive to pileup effects. Additional jets from softer QCD interactions contributes to the total energy recorded in the calorimeter. In Run 1 pileup jets were effectively removed by a minimal jet vertex fraction (JVF) requirement [51]. The JVF variable is defined as:

$$\mathrm{JVF} = \frac{\sum_{\text{tracks}\in\text{jet}\cap\mathrm{PV}_0} p_T^{\text{track}}}{\sum_{\text{tracks}\in\text{jet}} p_T^{\text{track}}}, \tag{2.1}$$

where the denominator is the scalar sum of p_T of all tracks associated to the jet, and the numerator is the scalar sum of p_T of tracks that are associated with the jet and originate from the hard-scatter vertex.

In Run 2 a multivariate combination of track-based variables called the jet vertex tagger (JVT) was developed [52]. Figure 2.18 (left) shows simulated JVT distributions for jets from hard-scatter vertices and pileup vertices. JVT was developed in such a way that the resulting hard-scatter jet efficiency is stable as a function of number of primary

vertices (N_{vtx}) as is shown in figure 2.18 (right).

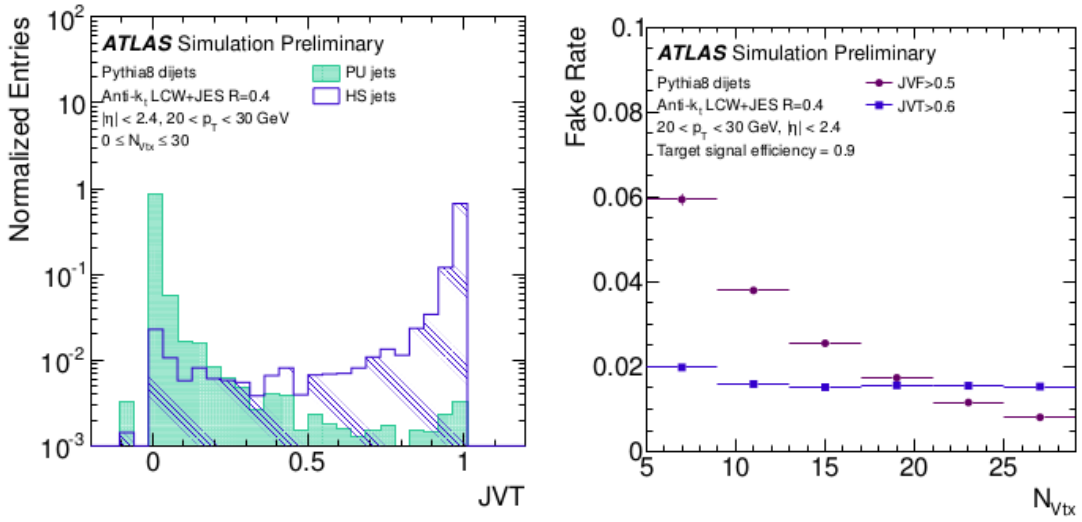


Figure 2.18.: The JVT distribution for pileup and hard-scatter jets with $20 < p_T < 30$ GeV (left). Pileup jet fake rate as a function of N_{vtx} imposing cuts on JVT and JVF for a fixed hard-scatter jet efficiency of 90%(right). [52]

Jets originating from bottom quarks are called b-jets. The procedure to identify such jets is called b-tagging, which is very important for the study of Higgs boson decays to b-quark pairs ($H \rightarrow b\bar{b}$) or in the selection of the top quark ($t \rightarrow bW$) for instance. Details about the b-tagging performance in ATLAS will be described in section 3.2.

2.5.3. Muon reconstruction

Muons are reconstructed using the information from the muon spectrometer, inner detector and the calorimeter. There are four muons categories, which are:

- Standalone muons: the muon trajectory is reconstructed by only the Muon Spectrometer (MS). The standalone muons starts from building track segments in each of the three muon stations. Then these segments are linked together and the track is extrapolated to the beam line. Standalone muons are mainly used to extend the ID acceptance coverage to the range $2.5 < |\eta| < 2.7$.

- Combined muons: a fit combines a standalone muon and an ID track. This category has the highest rejection power for fake muons and the best momentum resolution.

- Segment-tagged muons: if an ID track is matched to a segment of a track in the MS, then it is called a segment-tagged muon. Segment-tagged muons can be used to increase the acceptance in cases where the muon crossed only one layer of a MS chamber.

- Calorimeter-tagged muons: the muon is reconstructed by a combination of the track in the ID and a specific energy deposit in the calorimeter. This type has the lowest purity but recovers acceptance in the uninstrumented regions of the MS.

Overlap between different muon types are resolved before producing a unique collection of muons used in analysis. When two muon types share the same ID tracks, preference is given to combined muons, then to segment-tagged muons and finally to calorimeter-tagged muons. The overlap with standalone muons is resolved by analysing the track hit content and selecting the track with better fit quality and larger number of hits [53].

2.5.4. Electron identification

Electrons are reconstructed using information from the Inner Detector and the electromagnetic calorimeter [54]. The electron reconstruction has two steps: cluster reconstruction and the electron identification. In cluster reconstruction, a candidate electron object is created from clusters in the electromagnetic calorimeter that are matched to a reconstructed track from the ID. The clustering is performed using the *sliding window* [49] algorithm. The algorithm sum cells within a fixed-size rectangular window of 3×5 (in units of the tower size 0.025×0.025 in $\eta \times \phi$ space) adjusting the position of the window in such a way that the total energy deposited is a local maximum.

The electron identification in Run 2 uses a likelihood method to determine whether the reconstructed electron candidates are signal-like objects or background-like objects such as hadronic jets or converted photons [55]. This multivariate technique uses quantities related to the electron cluster and track measurements including calorimeter shower shapes, information from the transition radiation tracker, track-cluster matching related quantities, track properties, and variables measuring bremsstrahlung effects for distinguishing signal from background. The likelihood method use signal and background probability functions (PDFs) of the discriminating variables. Based on these PDFs, a likelihood is created for each hypothesis:

$$\mathcal{L}_s(\vec{x}) = \prod_{i=1}^{n} P_{s,i}(x_i) \tag{2.2}$$

where $\vec{x}$ is the vector of variable values and $P_{s,i}(x_i)$ is the value of the signal probability density function of the i^{th} variable evaluated at x_i. Then a discriminant (d_{Lagr}) is constructed with the signal and background probabilities:

$$d_{\mathcal{L}} = \frac{\mathcal{L}_s}{\mathcal{L}_s + \mathcal{L}_b} \tag{2.3}$$

Three operating points with different levels of purity are typically provided: *loose*, *medium* and *tight*. Each operating point uses the same variables to define the LH discriminant, but the cut value on this discriminant is different for each operating point. The samples selected by these operating points are chosen to be subsets of one another.

Electron efficiency measurements using the 2015 data can be found in reference [56].

2.5.5. Missing transverse energy

The missing energy in the ATLAS detector comes from non-interacting particles such as neutrinos and mismeasured particles. The direction and energy of those particles can be indirectly detected and measured using the momentum conservation in the transverse plane to the beam axis (the initial partons have negligible transverse momenta in comparison to those along the beam axis). This quantity is called missing transverse energy ($E_{\mathrm{T}}^{\mathrm{miss}}$).

$E_{\mathrm{T}}^{\mathrm{miss}}$ is calculated from the combination of all reconstructed and fully calibrated physics objects and from detector signal objects not associated with those objects. The calorimeter cells are associated with reconstructed physics objects in the following order: electrons (e), photons (γ), hadronically decaying tau-leptons (τ), jets and finally muons (μ).

After the corrections and calibrations, the $x(y)$ component of $E_{\mathrm{T}}^{\mathrm{miss}}$ is calculated by:

$$E_{x(y)}^{\mathrm{miss}} = E_{x(y)}^{\mathrm{miss},e} + E_{x(y)}^{\mathrm{miss},\gamma} + E_{x(y)}^{\mathrm{miss},\tau} + E_{x(y)}^{\mathrm{miss,jets}} + E_{x(y)}^{\mathrm{miss},\mu} + E_{x(y)}^{\mathrm{miss,soft}} \tag{2.4}$$

where the soft term ($E^{\mathrm{miss,soft}}$) is reconstructed from the transverse momentum deposited in the detector but not associated with any reconstructed object mentioned above. In Run 2, the soft term is measured by track-based methods to minimise the impact of pileup interactions [57].

Further details of reconstruction and performance of E_T^{miss} in ATLAS can be found in ref. [58].

3. Identification of double b-hadron jets

The ability to identify jets containing two b-hadrons from gluon splitting ($g \to b\bar{b}$) is important to reduce the heavy flavour QCD background to $H \to b\bar{b}$ searches and many new physics searches. This is further exacerbated by the absence of precise theoretical estimates of gluon splitting. A novel approach to identify jets containing two b-hadrons (called bb-jets in the rest of the thesis) has therefore been developed.

This chapter is organised as follows. Section 3.1 presents the motivation for the studies. In section 3.2, the identification of jets originating from b-quarks (b-jets) in ATLAS is summarised. Section 3.3 describes briefly the Multi Secondary Vertex Finder algorithm (MSVF). The simulated samples used for the studies are described in section 3.4. Section 3.5 details the performance of MSVF in b- and bb-jets. The new tagger (MultiSVbb) to identify jets containing two b-hadrons is presented in section 3.6 and section 3.7. Finally, a summary of the results is given in section 3.8.

3.1. Introduction

Bottom quarks are abundantly produced via QCD interactions in pp collisions at the LHC. The leading order (LO) QCD calculation of $b\bar{b}$ production includes a process know as flavour creation (FCR) while at next-to-leading order (NLO), the main mechanism of $b\bar{b}$ production are known as gluon splitting (GSP) and flavour excitation (FEX). FCR includes $b\bar{b}$ production through $q\bar{q}$ annihilation and gluon fusion, plus higher-order corrections of theses processes. In GSP, a gluon splits into a $b\bar{b}$ with a small opening angle between them. In FEX, a $b\bar{b}$ pair from the quark sea of the proton is produced, only one of the these quarks participates in the hard scattering. Examples of Feynman diagrams for the LO and NLO $b\bar{b}$ production are shown in figure 3.1. Most of the Monte Carlo event generators are only capable of describing FCR exactly, NLO effects are included approximately through the parton shower mechanism.

Angular correlations between pairs of b-hadrons have been studied with the CMS detector at a centre-of-mass energy of $\sqrt{s} = 7$ TeV [59]. Such studies allow for a sensitive test of perturbative QCD (pQCD) cross sections at new energy regimes as well as better knowledge of the heavy quark content of the proton. Measurements of the full range of the angular separation between the two b-hadrons demands good angular resolution and requires the ability to resolve small opening angles when the two b-hadrons are

inside a single reconstructed jet. Thus, having an efficient tagging of $b\bar{b}$-jets could lead to a better understanding of the modelling and reduce theoretical uncertainties of such processes.

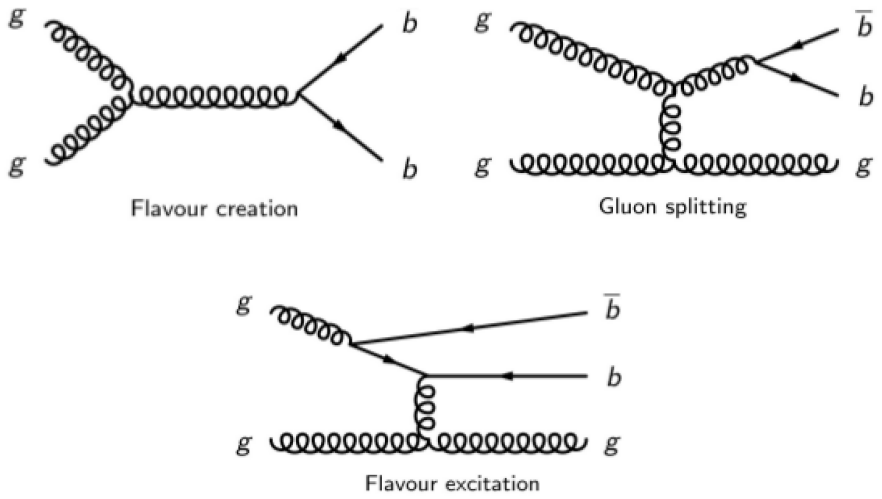
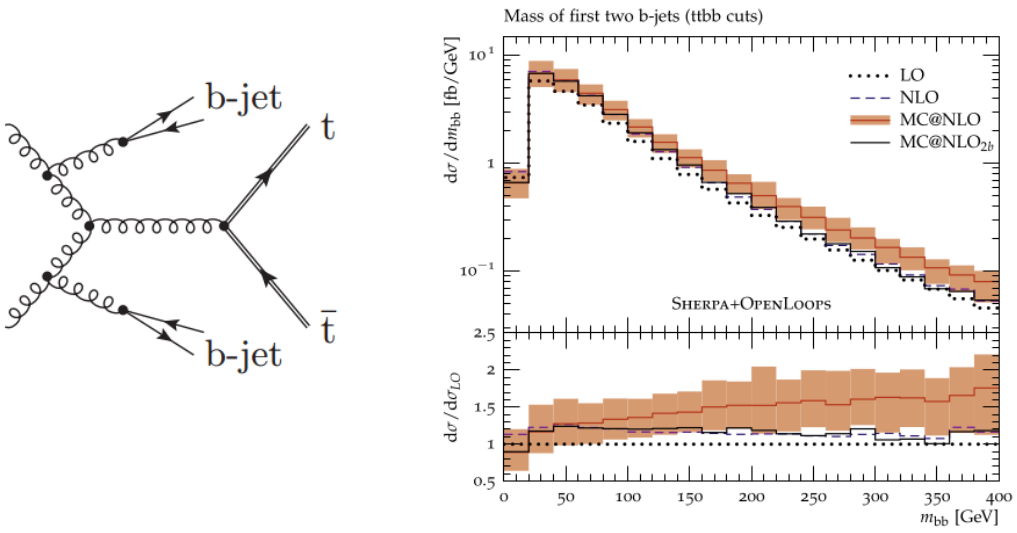


Figure 3.1.: Feynman diagrams that contribute to QCD b-quark production.

In the context of the $t\bar{t}H(H \to b\bar{b})$ analysis, both the signal and the dominant background processes are computed at next-to-leading order in QCD [60]. In order to be applicable to the experimental analyses these calculations need to be matched to parton showers. Modern fixed-order calculations have successfully been embedded in hadron-level simulation based on the MC@NLO [61] method, for the signal $t\bar{t}H(H \to b\bar{b})$ [62] and the dominant irreducible background $t\bar{t}b\bar{b}$ [63]. Moreover, matching massive b-quarks from NLO matrix elements to the parton shower gives access to novel $t\bar{t}$+b-jet production mechanics [64], where b-jets arise from hard gluons via collinear $g \to b\bar{b}$ splittings. In particular, one can describe $t\bar{t}$+2 $b\bar{b}$-jet events where both $b\bar{b}$-jets originate from $g \to b\bar{b}$ splittings, as shown in figure 3.2a. This kind of processes turns out to be a significant background contribution in the $t\bar{t}H(H \to b\bar{b})$ analysis as shown in figure 3.2b. The contribution from double collinear configurations is very relevant in the Higgs region ($m_{b\bar{b}}$ ∼125 GeV). The cross section ratio, MC@NLO/NLO, tends to increase up to 30% in the Higgs-signal region. Thus, a tool for identification of $b\bar{b}$-jets is important to control the dominant background for the $t\bar{t}H(H \to b\bar{b})$ analysis.

Different strategies are being considered in ATLAS to identify jets containing two b-hadrons. One of them, described in ref. [65] relies on the jet substructure techniques. It exploits jet substructure differences between single and merged b-jets combining them in a multivariate analysis. In this thesis, a method that relies on the direct reconstruction of all secondary vertices inside the jet is developed for the first time. The strategy to identify double b-hadrons in jets uses the MSVF algorithm for the direct reconstruction of the two b hadron decays (secondary vertices) and then uses a multivariate technique to increase the discrimination power between jet with two b-hadrons from single b-, c-



(a) (b)

Figure 3.2.: (a) $t\bar{t}b\bar{b}$ production via double collinear $g \to bb$ splitting. (b) Invariant mass of the first two b-jets. The MC@NLO bands display the matching corrections The MC@NLO2b curve is obtained by switching off $g \to bb$ splittings in the parton shower. The MC@NLO/NLO ratio grows with m_{bb} and reaches 25–30% in the Higgs-signal region [64].

and light jets.

3.2. Identification of b-jets in ATLAS

As introduced in section 2.5.2, the procedure to identify jets originating from b-quarks is called b-tagging. The aim of b-tagging in ATLAS [66] is to identify b-jets with high efficiency, while rejecting most of the background from jets originating from fragmentation of light quarks, c-quarks and gluons.

In the hadronisation process, the b-quark forms a b-hadron with $B^{\pm}$, B^0 and B_s being the most likely. b-hadrons are "heavy" particles with masses in the 5-10 GeV range and "long" lived particles typically having a mean lifetime $\tau \approx 10^{-12}s$ and a mean decay length $c\tau \approx 0.45$ mm (a significant flight path length $< l >= \beta\gamma c\tau \sim 1$ mm for a b-hadron momentum of around 10 GeV). Therefore, the identification of b-jets is based on the relatively long decay length of b-hadrons which leads to properties such as a displaced decay secondary vertex and large impact parameter tracks. As depicted in figure 3.3, tracks from b-jets tend to have larger impact parameter than the tracks coming from the primary vertex and the b-hadron decay generate a secondary vertex. The semi-leptonic decay of the b- and c-hadron also supplies useful information to identify b-jets. The presence of leptons is a good signature of the presence of b-hadrons in a jet. However, the small branching ratio (about 20% for each lepton flavour (e,μ)) make it statistically limited.

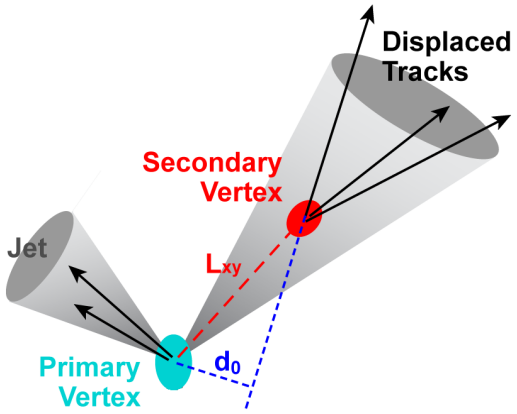


Figure 3.3.: A sketch of the b-jet decay products: displaced tracks and secondary vertex. Large impact parameter (d_0) tracks and a significant flight path length (L_{xy}) of the b-hadron are shown.

3.2.1. b-tagging ingredients

The determination of the position of the primary vertex in each event is important for b-tagging, since it defines the reference point to compute the impact parameters of tracks or for the measurement of displaced vertices. The primary vertex is reconstructed by using the adaptive multi-vertex finding algorithm, as described in section 2.5.1.

3.2.1.1. Association of tracks to jets

The jet flavour identification is performed with tracks in the event that are associated with the jet. Tracks are associated to each jet with a spatial matching based on the ΔR distance between the jet and tracks. Jets at high p_T are more collimated compared to jets at low p_T. This allows the track's distance from the jet axis depending on the jet p_T. Therefore, it is advantageous to have a smaller cone for jets at high p_T, as this reduces the number of events which are not from the b-hadron decay. The cut value in the distance ΔR depends on the jet p_T as expressed by the equation:

$$\Delta R(p_T) = a_0 + e^{a_1 + a_2 \cdot p_T} \tag{3.1}$$

where $a_0 = 0.239$, $a_1 = -1.22$ and $a_2 = -1.64 \cdot 10^{-5}$ [MeV^{-1}]. The value of the coefficients are optimised in order to collect on average 95% of the b-hadron decay products in the associated jet [67]. For a jet p_T of 20 GeV, the ΔR cut is 0.45 while for a jet with a p_T around 150 GeV the p_T cut is 0.26.

All b-tagging algorithms share the same track association except for the Multi Secondary Vertex Finder algorithm (MSVF), described in section 3.3. In order to identify separated secondary vertices in the jet, the ΔR cone size for MSVF is bigger than usual b-tagging algorithms: $\Delta R(p_T) = 0.315 + e^{-0.367 - 1.56.10^{-5} \cdot p_T}$. The ΔR cut is about 0.8 for jet p_T of 20 GeV and about 0.38 for a jet with a p_T around 150 GeV.

3.2.1.2. Track selection

Tracks associated with a jet are subject to specific requirements designed to select well-measured tracks, and to reject poorly reconstructed tracks, tracks from long-lived particles (K^0_S, Λ), material interaction (photon conversions or hadronic interaction) or tracks from pileup interactions. The track selection depends on each specific b-tagging algorithm. For the impact parameter based algorithm, a tight selection is applied. It includes a requirement that the tracks p_T is above 1 GeV, the transverse and longitudinal impact parameters are limited to $|d_0| < 1$ mm and $|z_0 \times \sin\theta|$<1.5 mm, and that there are at least two hits in the pixel detector. For the secondary vertex based algorithms a looser selection is used, relying on the secondary vertex reconstruction to provide additional purity. It requires the p_T of the track to be above 500 or 700 MeV depending on the tagger. Table 3.1 shows the basic quality cuts for the three main b-tagging algorithms used in ATLAS.

Criteria	IP	SSVF/MSVF	JetFitter
p_T [GeV]	>1.0	>0.7	>0.5
$\|d_0\|$ [mm]	<1.0	<5.0	<7.0
$\|z_0 \times \sin\theta\|$ [mm]	<1.5	<25	<10
number of B-Layer hits	≥1	≥0	≥0
number of Pixel hits	≥2	≥1	≥1
number of SCT hits	≥0	≥4	≥4
number of silicon hits	≥7	≥7	≥7
number of shared hits	–	≤1	–

Table 3.1.: Basic track quality selection for three main b-tagging algorithms: Impact Parameter (IP), single secondary vertex finder (SSVF) and JetFitter algorithm. Those algorithms are described in section 3.2.2. The multi secondary vertex finder (MSVF) algorithm uses the same track quality selection as SSVF. It will described in section 3.3.

3.2.2. b-tagging algorithms

Basic b-tagging algorithms based on tracks, secondary vertex and a decay chain fit are described in sections 3.2.2.1-3.2.2.3. They provide the input information for the final multivariate taggers which are described in section 3.2.2.4.

3.2.2.1. Impact parameter based algorithms (IP)

Two algorithms make use of the signed impact parameter significance of the tracks:

- **IP2D tagger**: a likelihood-based tagger using the transverse impact parameter, d_0.

- **IP3D tagger**: a likelihood-based tagger based on the 2D correlation between d_0 and z_0.

The sign of the impact parameter is defined as:

$$\text{sign}(d_0) = \text{sign}(\overrightarrow{d_0} \cdot \overrightarrow{j}_{xy}) \tag{3.2}$$

where $\overrightarrow{d_0}$ is a vector from the primary vertex to the point which defines d_0 on the track, and $\overrightarrow{j}_{xy}$ is a vector of the jet axis on the transverse plane. It is positive if the track intersects the jet axis in front of the primary vertex, and negative if the intersection lies behind the primary vertex.

The likelihood probability for b-, c- and light jet hypotheses are constructed with the track impact parameter significance $S(d_0) = d_0/\sigma_{d_0}$, where σ_{d_0} is the uncertainty on the reconstructed d_0. The track likelihood probability of being a track in the b-jet (p_b^{track}), is given by:

$$p_b^{\text{track}}(d_0/\sigma_{d_0}) = \frac{\mathcal{P}_b(d_0/\sigma_{d_0})}{\mathcal{P}_b(d_0/\sigma_{d_0}) + \mathcal{P}_c(d_0/\sigma_{d_0}) + \mathcal{P}_u(d_0/\sigma_{d_0})} \tag{3.3}$$

where $\mathcal{P}_b$, $\mathcal{P}_c$, and $\mathcal{P}_u$ are the probability density functions (PDFs) of the tracks in b-, c- and light jets. The PDFs are determined using simulated $t\bar{t}$ samples for the different flavour jets.

The individual track probabilities p_b^{track}, p_c^{track} and p_u^{track} are then combined in a single log likelihood ratio discriminant per jet (LLR).

$$\text{LLR}(p_b/p_u) = \sum_{\text{track}} \log\left(\frac{p_b^{\text{track}}}{p_u^{\text{track}}}\right) \tag{3.4}$$

The method allows for the use of different PDFs sets for different track categories. In Run-2, the categorisation of tracks has been significantly refined with a total of 14 categories in order to take advantage of the IBL insertion. The track categories are defined using the quality of the tracks which is based upon the hits from the Inner Detector used in the track reconstruction.

Figure 3.4 shows the LLR_{jet} distributions for the IP3D tagger.

3.2.2.2. Single Secondary Vertex Finding Algorithm

The Single Secondary Vertex Finder (SSVF) algorithm reconstructs only one secondary vertex per jet. Therefore, for a b-jet containing both b- and c-hadron decay vertices, the SSVF merge these vertices into a common single vertex if they are close in space, or it reconstructs the vertex with the largest track multiplicity.

The SSVF algorithm starts from all tracks that are significantly displaced from the primary vertex and form two-track vertex candidates with vertex fit $\chi^2 < 4.5$. The vertices compatible with long-lived particles (V_{0s}: K_s or Λ), photon conversions or hadronic interactions with the detector material are rejected. All tracks from the remaining two-track vertices are combined into a single inclusive vertex. Tracks with the largest con-

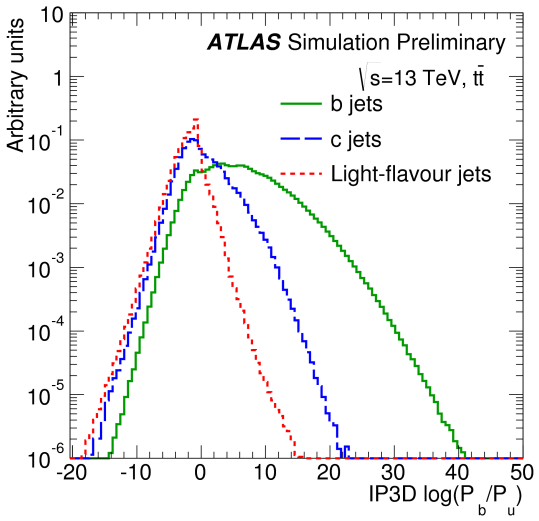


Figure 3.4.: The log likelihood ratio of the IP3D tagger in simulated $t\bar{t}$ events [26].

tribution to the χ^2 of the vertex fit are removed iteratively until a threshold vertex χ^2 (Prob$(\chi^2) > 0.001$) and a vertex invariant mass < 6 GeV are obtained. Figure 3.5 (left) shows the secondary vertex reconstruction efficiency as function of jet p_T for b-, c- and light-flavour jets.

Two algorithms based on the secondary vertex properties, called SV0 and SV1 [68], were developed for ATLAS Run 1. The SV0 tagger is a simple algorithm which takes the decay length significance as its discriminant. The SV1 tagger is a likelihood tagger based on the secondary vertex properties: the number of two-tracks pairs that can form a vertex, the invariant mass of the tracks associated to the secondary vertex, the fraction of the track momentum sum at the secondary vertex to the track momentum sum of the jet and the ΔR distance between the secondary vertex and the jet axis.

3.2.2.3. JetFitter

A different algorithm, JetFitter [69], exploits the topological structure of b- and c-hadron decays inside the jet. It assumes that the b- and c-hadron decay vertices lie on the same line, approximately the b-hadron flight path. A Kalman filter [70] is used to find the common line on which the primary vertex and the bottom and charm vertices lie.

This approach has several advantages, such as increasing the chance to separate b- and c-hadron vertices, even when a single track stemming from the b/c-hadron decay(s) is reconstructed. A single-track vertex can be formed from a track along the b-hadron flight axis which is compatible with the rest of the decay chain. Figure 3.5 (right) shows the efficiency to reconstruct a vertex with at least one or two tracks as function of jet p_T. The efficiency to have at least a single-track vertex is significantly higher than the efficiency to have a vertex with at least two tracks.

The first attempt to combine basic taggers was the JetFitterCombNN tagger. It com-

bined IP3D tagger output with JetFitter variables in a neural network.

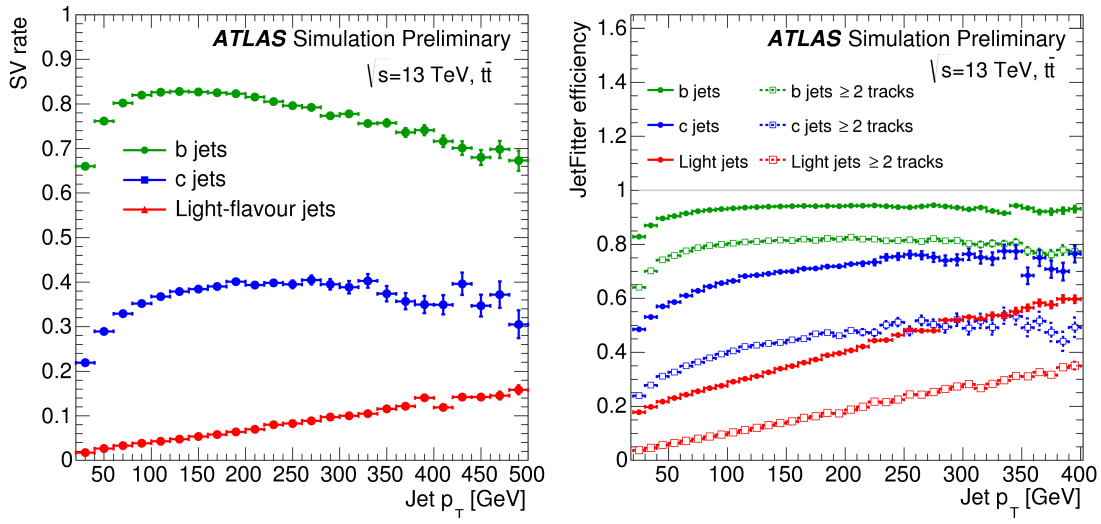


Figure 3.5.: Secondary vertex reconstruction efficiency for SSVF (left) and JetFitter (right) as function of jet p_T [26]

3.2.2.4. Multivariate taggers

Information from the three basic algorithms (IP, SSVF and JetFitter) are combined using multivariate analysis (MVA). The MVA taggers provide the best separation between b- and other flavour jets.

The MV1 tagger was used widely in Run 1 physics analysis. It uses a neural network technique to combine information from intermediate taggers based on likelihood (IP3D, SV0, SV1) and MVA methods (JetFitterCombNN), thus it was not a simple MVA combination.

In Run 2, a new multivariate tagger (MV2) was developed using boosted decision trees (BDT). It combines kinematic information from the jet (p_T, η), the IP2D/IP3D likelihood discriminant, the decay topology and properties of the vertices reconstructed by JetFitter as well as properties of the secondary vertex reconstructed by SSVF. The MV2 taggers used 24 input variables in total. Three variations of the MV2 taggers are provided: MV2c00, MV2c10 and MV2c20, where MV2c00 denotes the MV2 algorithm where no c-jet contribution was present in the training. MV2c10 (MV2c20) denote the MV2 outputs where a 7% (15%) c-jet fractions was present in the background sample (2016 b-tagging configuration) [71].

The MV2c10 output distribution is shown in figure 3.6 (left). A cut value on the MV2 output distribution defines a working point and it is chosen to provide a specific b-jet efficiency on a $t\bar{t}$ sample. Figure 3.6 (right) shows the c-jet rejection as a function of the b-jet efficiency for the different variations of the MV2 tagger. Table 3.2 shows the rejection for c- and light-jet at several b-tagging efficiency for the MV2c10 tagger.

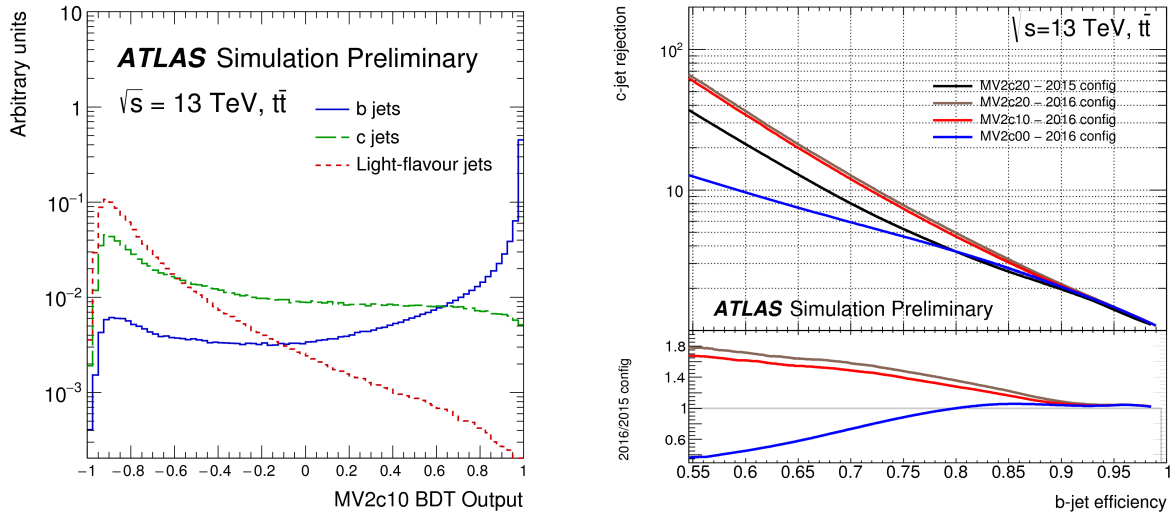


Figure 3.6.: The MV2c10 output for b-, c- and light-jets (left) and c-jet rejection versus b-jet efficiency for different configurations of the MV2 tagger [71].

Cut value	b-jet efficiency [%]	c-jet rejection	light-jet rejection
0.9349	60	34	1538
0.8244	70	12	381
0.6459	77	6	134
0.1758	85	3	33

Table 3.2.: Working points for the MV2c10 tagger, including efficiency and rejections rates on a $t\bar{t}$ sample [71].

3.3. Multi Secondary Vertex Finder algorithm

The Multi Secondary Vertex Finder (MSVF) algorithm finds all the possible vertices inside a jet using tracks associated to the jet. The track association used by the algorithm was described in section 3.2.1.1. The MSVF algorithm uses the same track quality selection and the same procedure to find the list of two-track vertices candidates as the one for the Single Secondary Vertex Finder (SSVF) algorithm described in section 3.2.2.2.

After rejecting 2-track vertices coming from V_{0s}, photon conversion and hadronic interaction with the detector material, the cleaned 2-track vertex set is converted into a graph. Every node in the graph represents a track and an edge connecting the two nodes represents the good 2-track vertex. Then a special graph algorithm, implemented in the BOOST GRAPH library [72], provides all complete subgroups of the original graph (cliques) where all nodes are connected to each other. Thus, the obtained set of cliques is a set of all possible vertices for a given set of tracks. However, the solution is ambiguous, one node (track) can be present in some cliques (vertex candidates). The algorithm requires a final iterative cleaning to arrive to a physical (not mathematical) set of vertices.

The cleaning procedure is [73]:

- If a vertex candidate has a very large χ^2, the track with the largest χ^2 contribution is detached from the given vertex. This track is then combined with another track from the vertex (they are all pair-wise compatible) into a 2-track vertex with minimal χ^2, which is added to the vertex candidate set.
- If two vertices in the vertex set are far from each other but have a common track, this track is detached from the vertex with the largest χ^2 and the vertex position is refitted.
- If two vertices become close to each other – they are merged.

This iterative procedure results in a set of separated physical vertices not having common tracks. Vertices with only one track are allowed. After a good vertex is found its momentum (so the direction) is calculated and used as a pseudo-track to look for crossing with an additional real track. Additional tight quality cuts are applied to the pseudo-track+real track vertex to minimise fake vertices.

Figure 3.7 shows a schematic view of the MSVF algorithm. At the end, the MSVF algorithm provides all the possible vertices inside a jet using tracks associated to the jet. However additional tracks not originating from b- or c-hadron decays can lead to fake vertices inside jets. In addition, instrumental resolution can lead to merged vertices containing tracks from different origins or to split vertices where two or more vertices are reconstructed from tracks belonging to the same spatial point.

Several studies to understand vertexing performance and ambiguous cases (e.g. B/C separation, fakes vertices) in the MSVF algorithm are described in section 3.5.

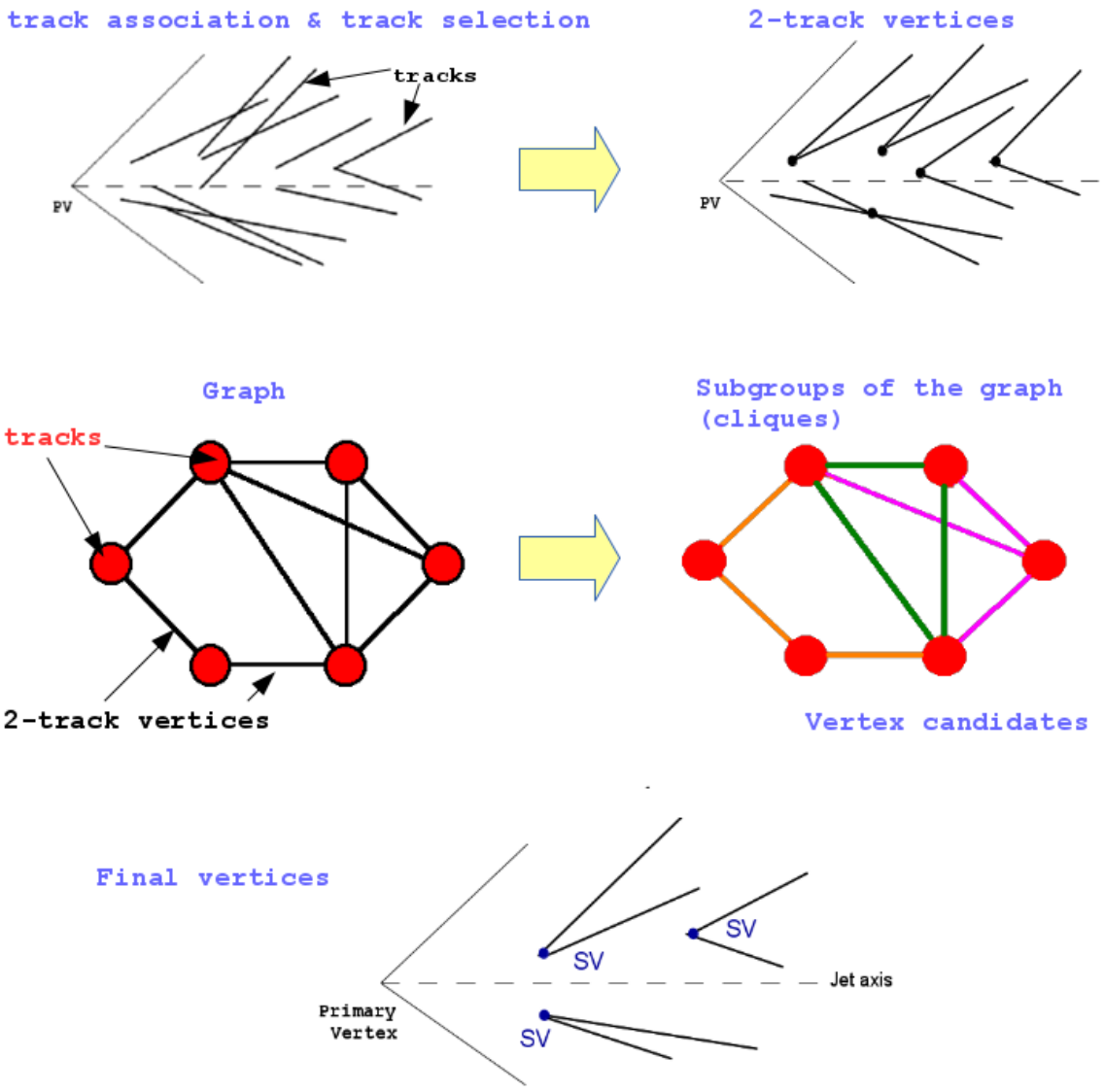


Figure 3.7.: Schematic view of the reconstruction of the secondary vertices by the MSVF algorithm. The MSVF algorithm finds all the possible vertices inside a jet using tracks associated to the jet.

3.4. Simulated samples

Two different samples are used for the performance studies of the MSVF algorithm. A sample of $t\bar{t}$ events is used to study single-b jets. It contains b-jets with high purity since the top quark decays most of the time, to a b-quark and a W boson. The $W + b\bar{b}$ sample is chosen to study the performance of the MSVF algorithm in bb-jets. The relevance of bb-jets in events with a $W+$ two (or more) jets with at least one b-quark is supported by NLO calculations [74], which indicate that the cross section for $W+$ (bb)j (exactly two jets, one of which contains two b-quarks) is almost a factor of two higher than $W + b\bar{b}$ (exactly two jets, both of which contain a b-quark). Figure 3.8 shows two leading order processes that produce W bosons with at least one b-jet.

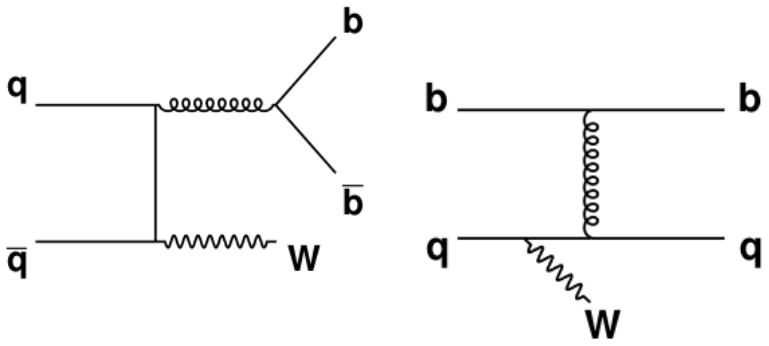


Figure 3.8.: Examples of Feynman diagrams for W production in association with b quarks. In the first process (left), b quarks are produced at small angles by gluon splitting and can be reconstructed as a jet containing two b-quarks.

The samples from proton-proton collision at a centre-of-mass energy of $\sqrt{s} = 8$ TeV were generated with POWHEG [38] plus PYTHIA 6.423 [75] for $t\bar{t}$ events and ALPGEN [37] plus PYTHIA 6.423 for $W + b\bar{b}$ events. The GEANT4 [41] software within the ATLAS simulation framework [34] propagates the generated particles through the ATLAS detector and simulates their interactions with the detector material. The simulated data sample used for the analysis gives an accurate description of the pile-up and detector conditions for the 2012 data-taking period.

The jet algorithm selected for the analysis was the ATLAS default anti-k_t algorithm [48], with a distance parameter R = 0.4, using calorimeter topological clusters [49] as input. Jets are required to have a minimum p_T of 25 GeV and also required to be in a region with full tracking coverage, $|\eta| < 2.5$.

In order to cover a jet p_T range up to 500 GeV, two Monte Carlo samples: $W + b\bar{b}$ and multijet samples, are used for the development of the double b-hadron tagger (called MultiSVbb). QCD multijet Monte Carlo samples were generated with Pythia8 [36] interfaced with EvtGen [76]. The AU2 CT10 tune [77] was used. The samples are divided up into slices depending on the generated jet p_T range. Slices JZ2W (leading truth jet p_T spectrum: 80-200 GeV) and JZ3W (leading truth jet p_T spectrum: 200-500 GeV) have

been used to cover a p_T range of 25-500 GeV. The MultiSVbb tagger will be described in section 3.6.

Using information in the simulation, jets were labelled as bb-jet (b-jet) if they contain two (one) final state b-hadrons with $p_T > 5$ GeV within ΔR(b-hadron,jet) < 0.4. In the same way, jets are labelled as cc-jets (c-jets) if they contain two (one) final state c-hadrons with $p_T > 5$ GeV and ΔR(c-hadron,jet) < 0.4. The remaining jets are labelled as light jets. This label definition is used for the following studies, unless otherwise specified.

3.5. Performance of the Multi Secondary Vertex Finder algorithm

In this section we focus on the understanding of the performance of the MSVF algorithm by investigating the purity of the reconstructed vertices in jets containing single and double b-hadrons.

3.5.1. Vertex Purity Fraction in single-b jets

The MSVF algorithm was tested on the $t\bar{t}$ simulation sample in order to study its performance in single-b jets. This section uses the ATLAS b-tagging Run 1 label definition[a]; jets are labelled as a b-jet if a b-quark with $p_T > 5$ GeV is found in a cone of $\Delta R = 0.3$ around the jet direction. Only b-jets are kept for the next studies.

Figure 3.9 shows the number of reconstructed vertices and the number of tracks in the reconstructed vertex found by the MSVF algorithm in b-jets. About 68% of the b-jets have at least one reconstructed vertex. The fraction of b-jets with two reconstructed vertices ($\sim$20%) is less than the fraction of b-jets with only one reconstructed vertex ($\sim$42%) which indicates that truth vertices from b- and c- hadron decays, in many cases, are merged in one reconstructed vertex. About 41% of the reconstructed vertices have two tracks and around 14% of the reconstructed vertices have only one track, as explained in section 3.3 vertices with only one track are allowed. In order to make studies with truth vertices, a reconstructed vertex is considered only if it has at least two or more associated tracks.

To estimate the performance of the reconstructed vertices the truth B (C) vertex of b-hadron (c-hadron) are defined as points in space where the cascade of charged particles begins, as shown in figure 3.10.

Then a matching procedure is defined as follow:

- A weighted matching probability P_{match} is defined using the ratio of the number of hits which are common to a given track and the corresponding truth particle and

[a] In Run 2, hadrons are used instead of quarks. Thus, jets are labelled as a b-jet (c-jet) if a b-hadron (c-hadron) with $p_T > 5$GeV is found in a cone of $\Delta R = 0.3$ around the jet direction.

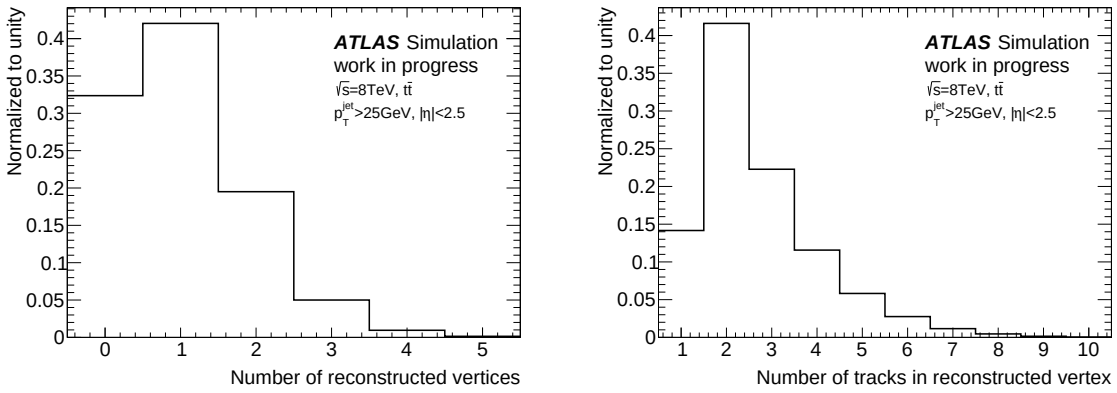


Figure 3.9.: Number of reconstructed vertices (left) and number of tracks in all vertices reconstructed (right) by the MSVF algorithm in b-jets.

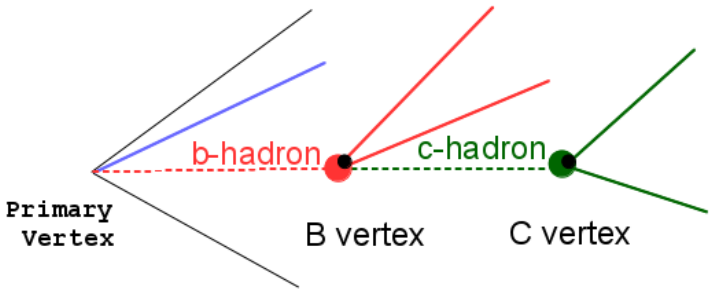


Figure 3.10.: Schematic view of the B and C truth vertices defined as points in space where the cascade of charged particles begins. Only final-state b and c-hadrons are considered ("excited" states can decay into their final state, e.g. $B_s^0 \to D_s^{*+} \to \gamma D_s^+$). Only final state hadrons have a large lifetime and can create a secondary vertex.

the number of hits which form the track:

$$P_{match} = \frac{10 \times N_{Pix}^{common} + 5 \times N_{SCT}^{common} + N_{TRT}^{common}}{10 \times N_{Pix}^{track} + 5 \times N_{SCT}^{track} + N_{TRT}^{track}} \tag{3.5}$$

where N_{Pix}, N_{SCT} and N_{TRT} are the number of hits in the pixel, SCT and TRT detector, respectively. A reconstructed track is considered as matching a given truth particle if $P_{match} > 0.8$ [78].

- An association is performed between reconstructed tracks and charged particles from a truth B or C vertex inside the jet. For tracks corresponding to charged particles, which are not originated from a truth B or C vertex, we define the type X "vertex". These tracks can be generated by charged particles produced in the interaction with the detector material, by vertices V0s (K_s ,Λ) reconstructed inside the jet or by charged particles from truth vertices B or C outside the jet cone.

- A Vertex Purity Fraction (VPF) is defined with respect to B or C or X vertices per reconstructed vertex as:

$$\text{VPF(B/C/X)} = \frac{\text{Number of reconstructed tracks matching particles from B or C or X}}{\text{Total number of reconstructed tracks in the vertex}} \tag{3.6}$$

Figure 3.11 shows the Vertex Purity Fraction with respect to B, C and X vertices for all the reconstructed vertices in single-b jets. A 2D plot of the VPF(B) versus the VPF(C) vertex is shown as well. Around 65% (80%) of the reconstructed vertices have at least one track from B (C) vertices. Around 10% (15%) of the reconstructed vertices are correctly reconstructed with only tracks from B (C) vertices, VPF=1. Around 15% of the reconstructed vertices have at least one track from X vertices. From the 2D plot, around 13% of the reconstructed vertices have VPF(B)=VPF(C)=0.5, they are merged vertices.

The reconstructed vertices are divided according to the value of VPF in 3 inclusive categories: VPF(B)= 1, VPF(B)$\geq$0.5 and VPF(B)$\geq$0.1, similar categorisation is done for VPF(C). And a vertex efficiency per jet is defined as:

$$\text{Eff(B/C)} = \frac{\text{jets with at least one reco vertex with VPF(B/C)}(= 1, \geq 0.5, \geq 0.1)}{\text{jets with at least one true B/C vertex within } \Delta R<0.3} \tag{3.7}$$

The efficiency slightly increases with the jet p_T for category VPF=1 and slightly decrease for VPF$\geq$0.1, as shown in Figures 3.12a, 3.12b. Similar behaviour is observed as a function of the b-hadron p_T, as shown in figure 3.12c. As expected the efficiency increases with the number of charged particles from b-hadron, as shown in figure 3.12d. Figures 3.12e and 3.12f show the efficiency as a function of the transverse distance between B and C vertex. When the distance between B and C vertices is small it is hard

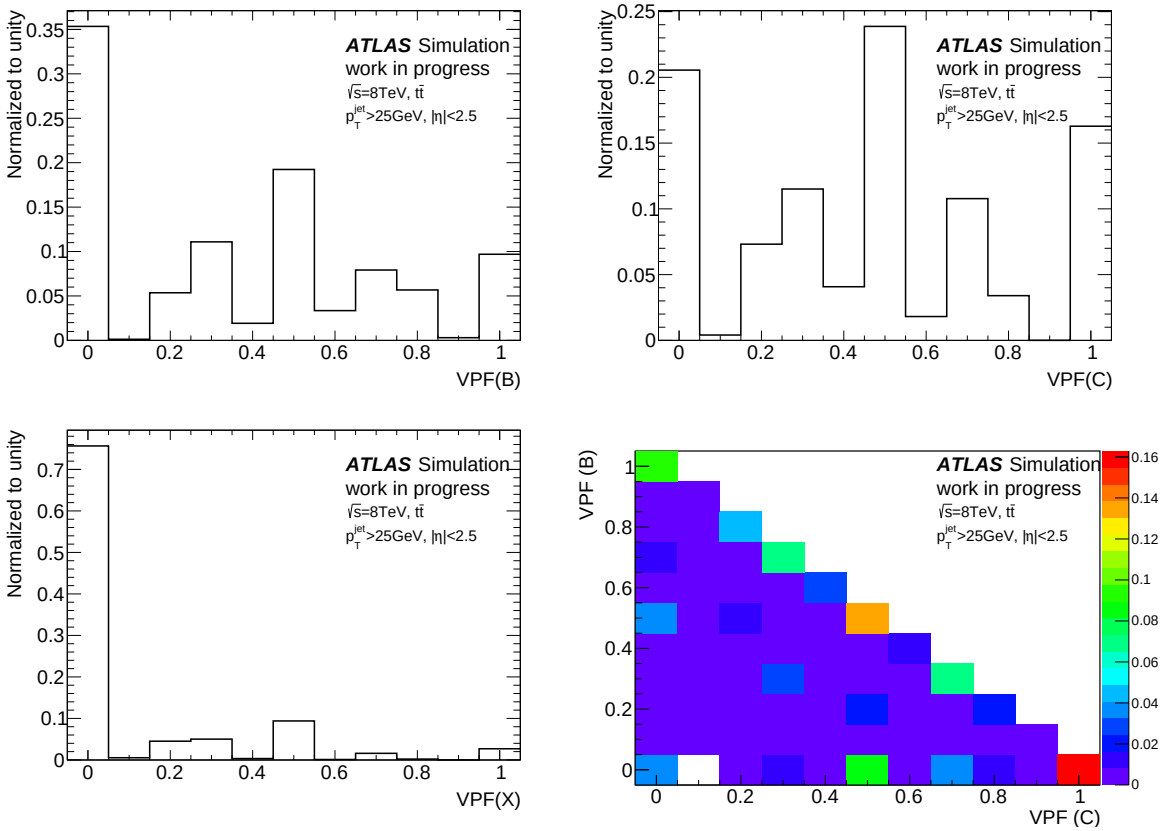


Figure 3.11.: Vertex Purity Fraction in single-b jets

to reconstruct two vertices and hence the efficiency for vertices correctly reconstructed, VPF=1, increases with the distance.

In order to study further the reconstructed vertices one can look at single-b jets with only 2 reconstructed vertices. Figure 3.13 shows Vertex Purity Fraction in single-b jets with exactly 2 reconstructed vertices. Around 63% (63%) of the reconstructed vertices have at least one track from B (C) vertex. For this case, it is more probable to have vertices correctly reconstructed with only B or C tracks. Around 16% (17%) of the vertices are correctly reconstructed, VPF=1. Around 6% of the reconstructed vertices have at least one track from X vertex and 12% of the reconstructed vertices have VPF(C)=VPF(B)=0.5; the merge effect remains.

Figure 3.14 shows the fraction of b-jets with exactly 2 reconstructed vertices for different categories in VPF of the vertices as a function of jet p_T and as a function of the transverse distance between the truth B and C vertices. Three categories are shown: VPF(B)=1 and VPF(C)=1, VPF(B)$\geq$0.5 and VPF(C)$\geq$0.5 and VPF(B)$\geq$0.1 and VPF(C)$\geq$0.1. As expected, the fraction of b-jets for the different categories increases with the jet p_T and the transverse distance between the B and C vertices. However, one can see that the fraction of b-jets with two correctly reconstructed vertices,VPF(B)=1 and VPF(C)=1, is about 3% only.

Fugure 3.15 shows three properties of the reconstructed vertices: the invariant mass of the tracks associated to the secondary vertex (mass), the fraction of the energy of the tracks attached to the vertex to the sum of energies of all tracks associated to the jet (energy fraction) and the number of track in the vertex. Figure 3.15 (left) shows the properties of the reconstructed vertices in jets with one or more vertices while figure 3.15 (right) shows the same properties for jet with exactly 2 reconstructed vertices. Correctly reconstructed vertices (VPF(B)=1) and vertices with VPF(B)$\geq$0.5 tend to have higher mass. This effect is less pronounced in cases with exactly 2 reconstructed vertices since we have a second reconstructed vertex. At VPF(B)=1 or VPF(B)$\geq$0.5, there is low energy fraction (fraction of the vertex energy with respect to the jet energy) for exactly 2 reconstructed vertices in jet. More tracks in reconstructed vertex are observed for the case of one or more reconstructed vertices in the jet.

To be able to study the efficiency of the vertexing reconstruction, different track and vertex categories are defined:

- "B (C) reconstructed tracks": reconstructed tracks coming from b-hadron (c-hadron) inside the jet with $P_{match} > 0.8$ and $p_T \geq 700$ MeV. About 58% of the B and C reconstructed tracks are included in the reconstructed vertices by the MSVF algorithm.

- "Missing B(C) reconstructed tracks": truth charged particles from b-hadron (c-hadron) with $p_T \geq 700$ MeV not associated with a reconstructed track. About 8% of the charged particles from b- and c-hadron are not associated with a reconstructed track.

- "Reconstructable B (C) vertex": truth B (C) vertex with at least two B (C) reconstructed tracks. Figure 3.16 (left) shows the number of truth B vertices versus

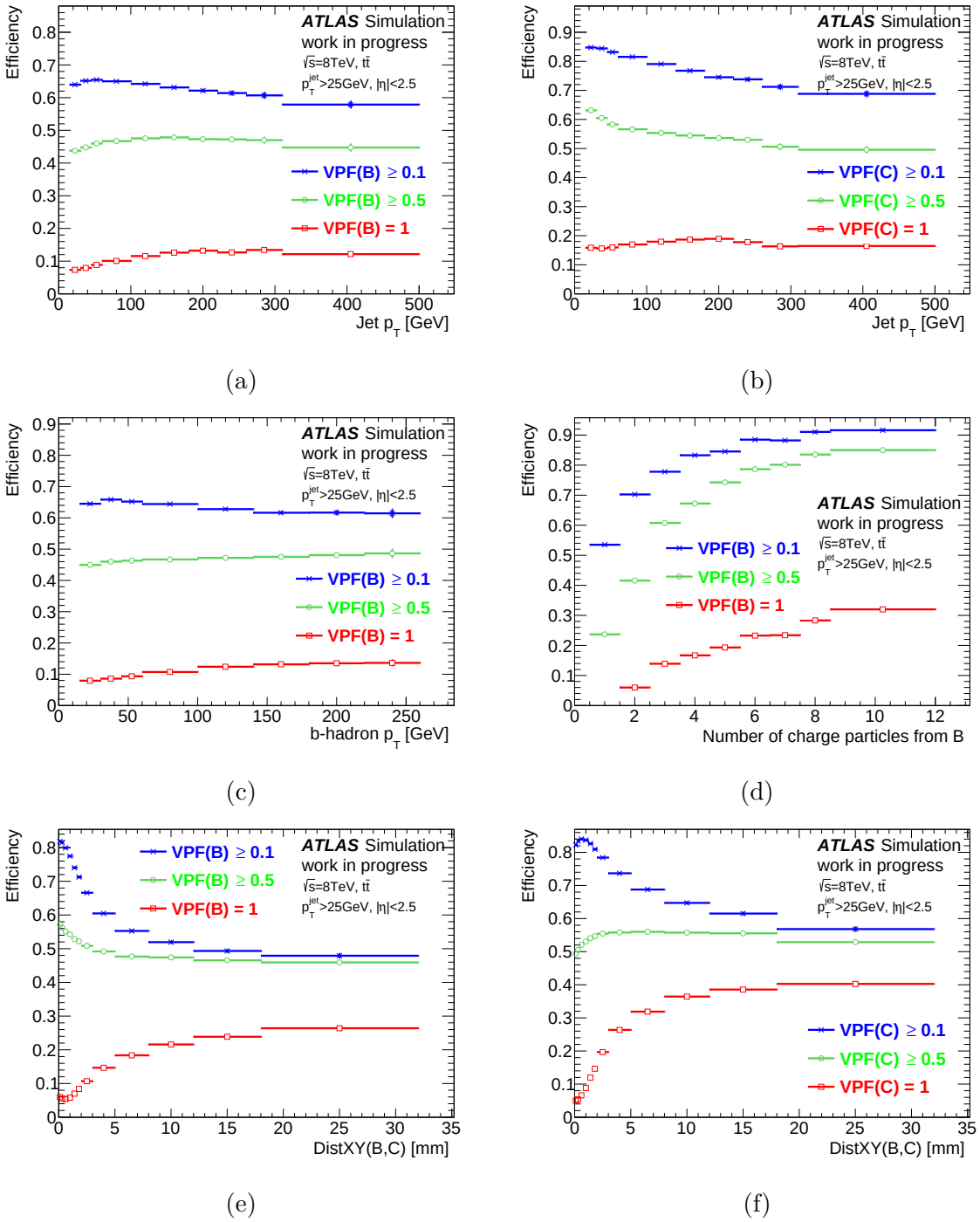


Figure 3.12.: Efficiency in single-b jets. Eff(B) as function of the jet p_T (a), the b-hadron p_T (c), the number of charged particles from b-hadron (d) and the transverse distance between truth B and C vertices (e), and Eff(C) as function of jet p_T (b) and transverse distance between truth B and C vertices (f) are shown.

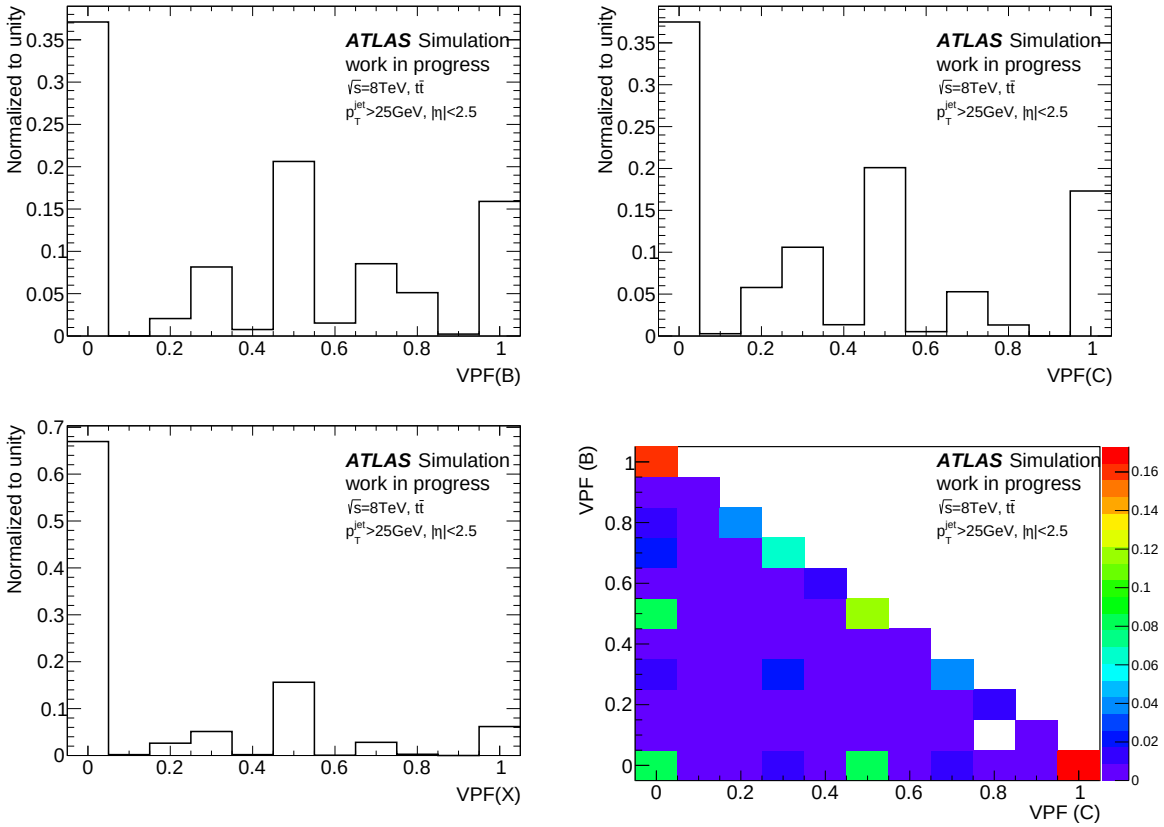


Figure 3.13.: Vertex Purity Fraction in single-b jets with exactly 2 reconstructed vertices.

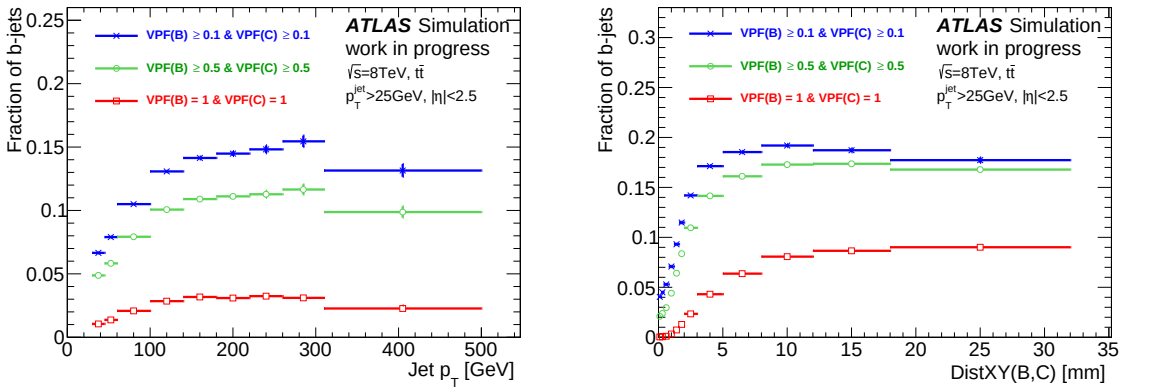


Figure 3.14.: Fraction of single-b jets with exactly two reconstructed vertices in different vertex categories as a function of jet p_T (left) and the transverse distance between the truth B and C vertices (right).

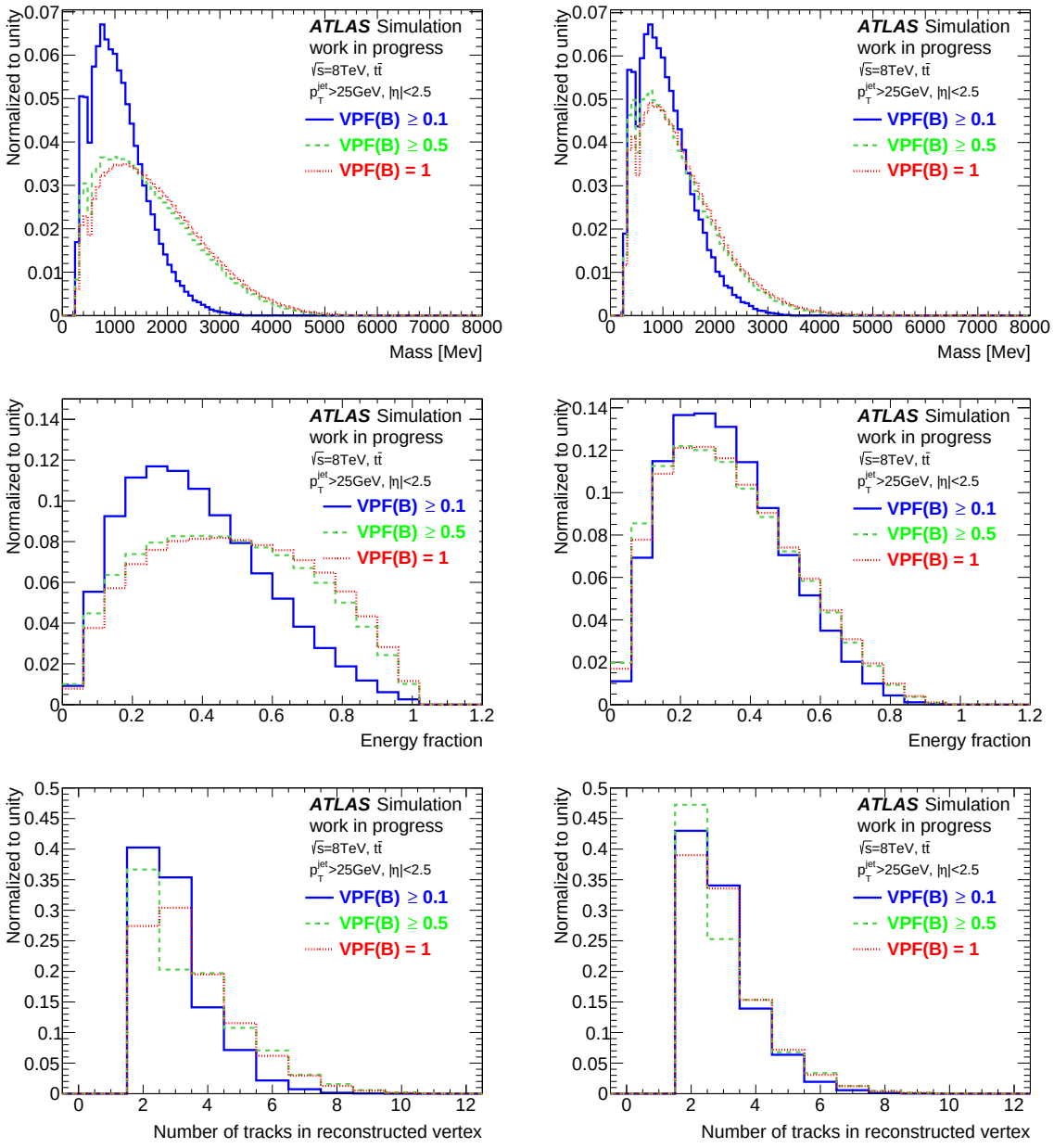


Figure 3.15.: Properties of the reconstructed vertices in b-jet with one or more vertices (left) and exactly 2 reconstructed vertices in b-jet (right).

number of truth C vertices and figure 3.16 (right) shows the number of reconstructable B vertices versus the number of reconstructable C vertices in single-b jets. Even when most of the b-jets have one truth B and one truth C vertices ($\sim$ 79%), the number of b-jets having one reconstructable B and one reconstructable C vertices represent about 28% only.

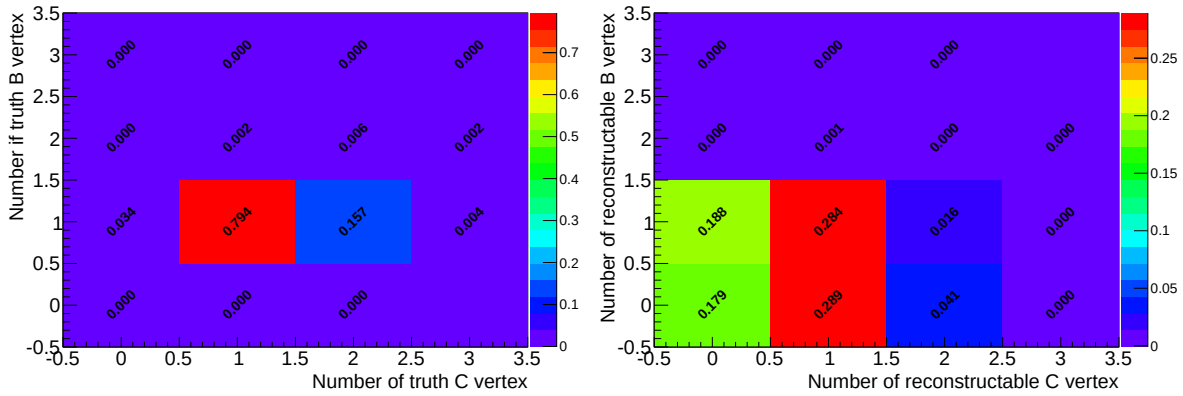


Figure 3.16.: (left) Number of truth B vertex versus number of truth C vertex in single-b jets. (right) Number of reconstructable B vertex versus number of reconstructable C vertex in single-b jets.

Another good criterion to estimate the performance of the MSVF algorithm is to look at the fraction of B and C reconstructed tracks in the reconstructed vertices with respect to the tracks in the vertices (purity) or with respect to the total of B and C reconstructed tracks in jet (efficiency).

$$\text{B/C track purity} = \frac{\text{B and C reconstructed tracks in all vertices in jet}}{\text{Total tracks in all vertices in jet}}$$

$$\text{B/C track efficiency} = \frac{\text{B and C reconstructed tracks in all vertices in jet}}{\text{Total B and C reconstructed tracks in jet}}$$

Figure 3.17 shows the general MSVF purity and efficiency per jet. The efficiency to use B and C tracks in the vertices is found to be 44% and the purity of B and C tracks in the vertices is about 62%.

Finally, figure 3.18 shows the probabilities to have one, at least two and exactly two reconstructed vertices in a single-b jet when this jet has two reconstructable vertices (one B and one C vertex) as a function of the transverse distance between the B and C vertices. As expected, the fraction of single-b jets having two or more vertices increases with the distance between the reconstructable vertices and the fraction with exactly one reconstructed vertex decreases with the distance.

The goal of these studies was estimate the performance of the MSVF algorithm in single-b jets and eventually obtain information sensitive to the presence of multiple secondary vertices in a single-b jets in order to improve the identification of b-jets in

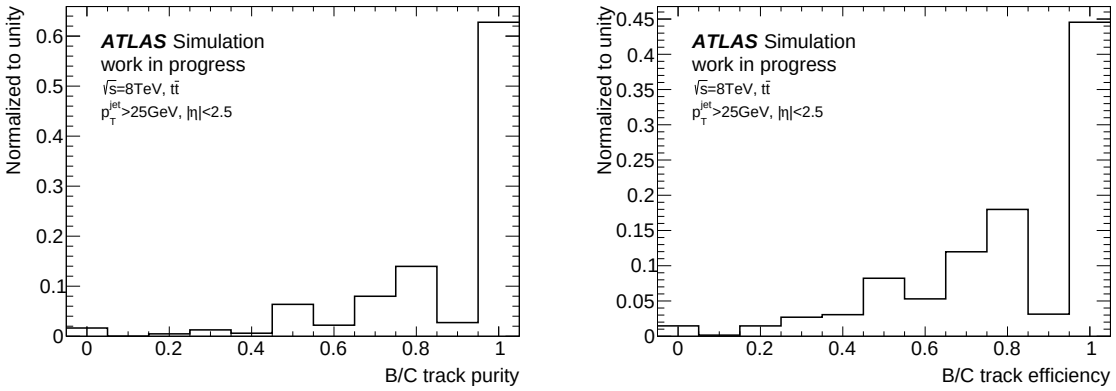


Figure 3.17.: B/C track purity and efficiency per jet

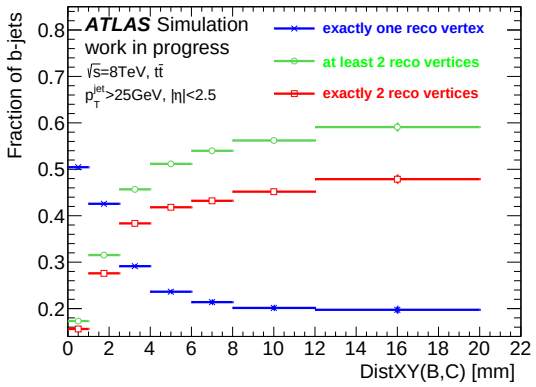


Figure 3.18.: Fraction of single-b jets with two reconstructable B and C vertices as a function of the transverse distance between B and C vertices. Three categories are shown: jets with exactly one reconstructed vertex, jets with at least 2 reconstructed vertices and jets with exactly 2 reconstructed vertices.

ATLAS. As was showed, the MSVF algorithm can not resolve B and C decay vertices efficiently; the fraction of b-jets with two correctly reconstructed B and C vertices is about 3% only. Thus, it is not very suitable for single-b jets. However, the MSVF algorithm can resolve better reconstructable vertices, as shown in figure 3.18. This feature can be exploit in jets containing two b-hadrons (bb-jets). In the next section, studies to characterise the performance of the MSVF algorithm in bb-jets are presented.

3.5.2. Vertex Purity Fraction in bb-jets

As mentioned in section 3.4, a $W + bb$ sample is used to characterise the MSVF performance in bb-jets.

As expected, bb-jets have a higher number of reconstructed vertices in a jet than single b-jets, as shown in figure 3.19 (left). Around 48% of the bb-jets have at least 2 reconstructed vertices. Also many bb-jets have 3 and 4 reconstructed vertices inside the jet which could be produced by the separation of b- and c-hadron vertices. Figure 3.19 (right) shows the number of tracks per reconstructed vertex in bb-jets. Around 43% of the vertices have 2 tracks and 17% have 1-track vertices. Single-track vertices are left out for the following studies.

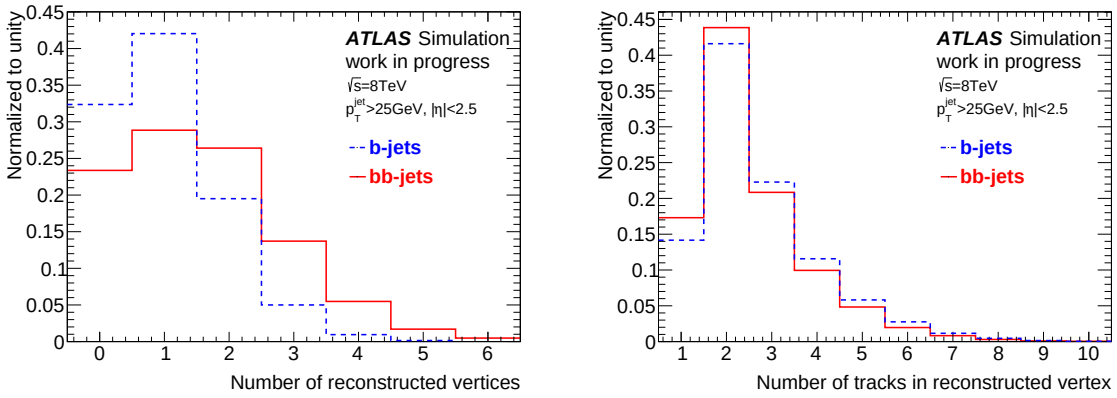


Figure 3.19.: Number of reconstructed vertices (left) and number of tracks per reconstructed vertex (right) by the MSVF algorithm in bb-jets.

In order to quantify the performance of the reconstructed vertices in bb-jets, we define the Truth Secondary Vertex (TSV) position corresponding to the truth B and C vertices, as sketched in figure 3.20. Truth B and C vertices are merged in a unique space point that correspond to the truth B vertex. bb-jet is expected to have two decay cascades of b-hadrons, hence two TSV (called TSV1 and TSV2).

Similar to the VPF definition for single b-jets, a matching procedure is defined as:

- It is required that the bb-jet has at least two reconstructed vertices.
- A reconstructed track is considered as matching a given truth particle if $P_{match} > 0.8$. (see equation 3.5).

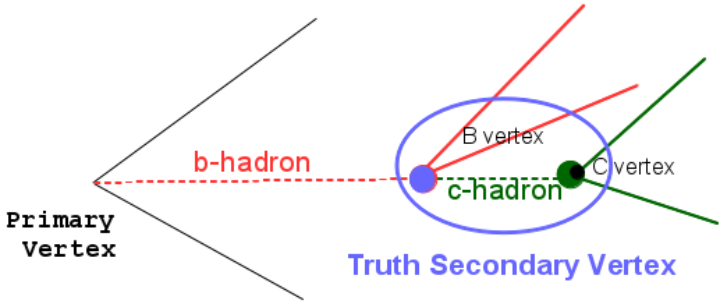


Figure 3.20.: Schematic view of the Truth Secondary Vertex (TSV). Truth B and C vertices are merged in a unique TSV which correspond to the truth B vertex.

- MC truth matching is performed between reconstructed tracks and charged particles to decide if a track comes from a B, C or X vertex.
- A Vertex Purity Fraction (VPF) is defined with respect to the Truth Secondary Vertex per reconstructed vertex as:

$$\text{VPF\_TSV}(1/2) = \frac{\text{Number of B/C reconstructed tracks} \in \text{TSV (1/2)}}{\text{Total number of tracks in the vertex}} \tag{3.8}$$

Figure 3.21 shows the Vertex Purity Fraction with respect to the TSV1 and TSV2. Also a 2D plot of the VPF(TSV1) versus VPF(TSV2) is show. Around 27% of the reconstructed vertices are correctly reconstructed with only tracks from TSV1 or TSV2 (VPF=1). From the 2D plot around 10% of the reconstructed vertices have VPF_TSV1=VPF_TSV2=0.5, the truth secondary vertices cannot be distinguished.

Figure 3.22 shows VPF_TSV1 versus VPF_TSV2 for different transverse distances between the truth secondary vertices. As expected when this distance is small we have more merged vertices. However, the merging effect still exists at large transverse distance between the truth secondary vertices (distxy(TSV1,TSV2)>5 mm).

Figure 3.23 (left) shows the fraction of bb-jets with exactly one reconstructed vertex, exactly 2 reconstructed vertices and at least 2 reconstructed vertices as a function of the transverse distance of the truth secondary vertices. As expected, the fraction of bb-jets with at least 2 reconstructed vertices increases with the transverse distance of the truth secondary vertices and this effect is opposite for bb-jets with only one reconstructed vertex. Figure 3.23 (right) shows the fraction of bb-jets with at least 2 reconstructed vertices as a function of the transverse distance of the truth secondary vertices divided with respect to the bb-jet p_T in three categories: $20 \leq p_T < 60$, $60 \leq p_T < 110$ and $110 \leq p_T < 200$ GeV. From this figure one can see that the fraction of bb-jets with at least two reconstructed vertices increases with the p_T of the bb-jet.

The MSVF algorithm has a good performance in bb-jets. About 48% of the bb-jets have at least two reconstructed vertices. About 27% of the reconstructed vertices are correctly reconstructed with only tracks from TSV. More secondary vertices were found in bb-jets compared to b-jets. Thus, the MSVF algorithm provides good information for

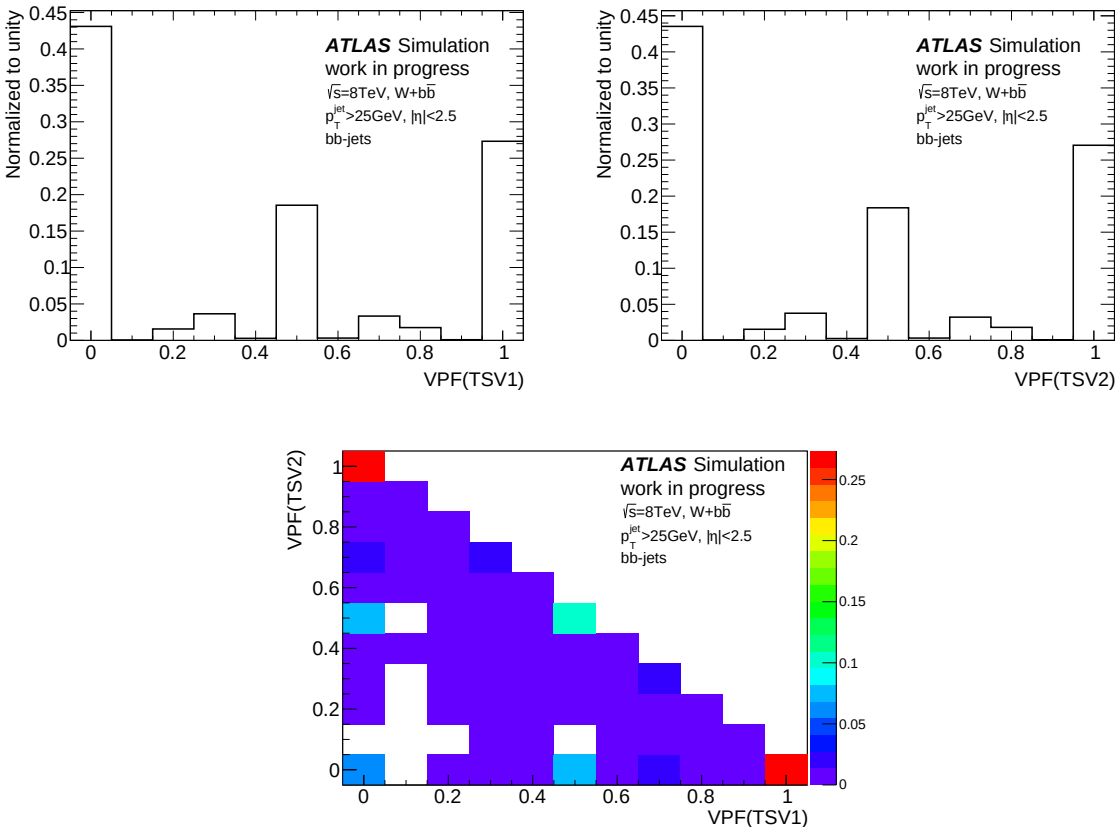


Figure 3.21.: Vertex Purity Fraction with respect to TSV1 and TSV2 in *bb*-jets

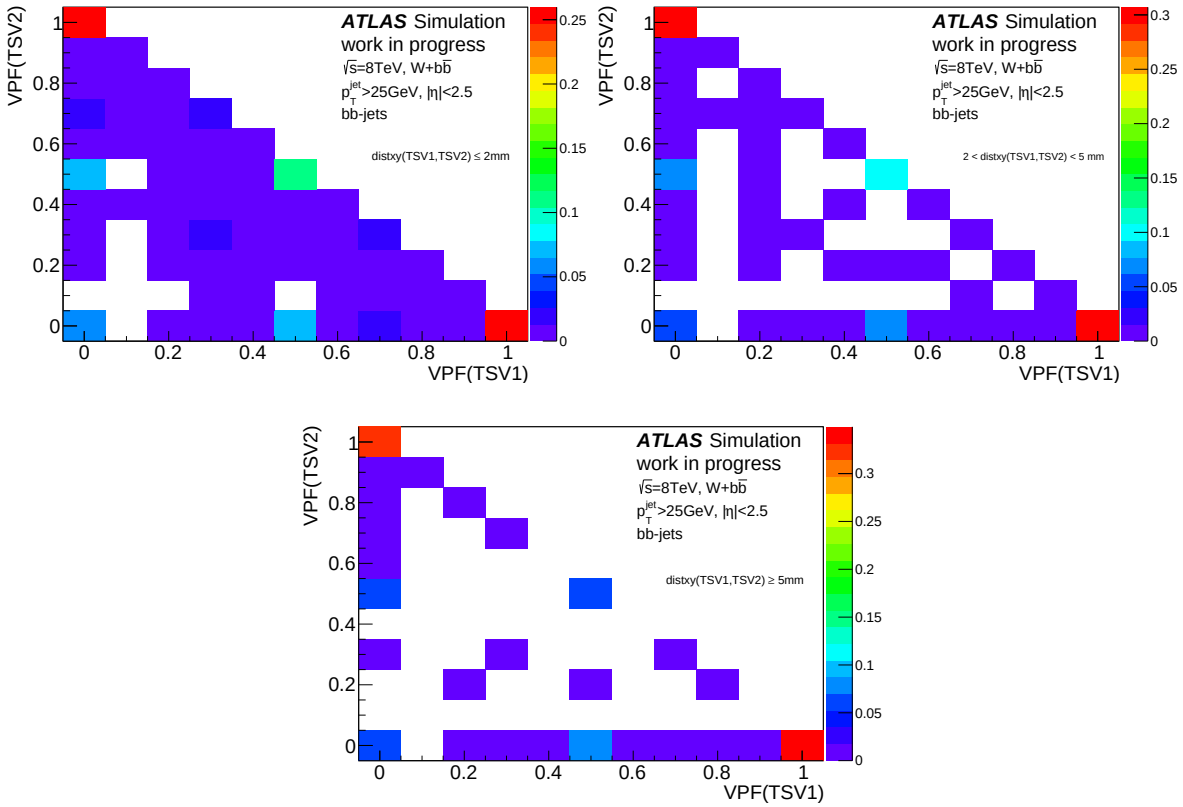


Figure 3.22.: VPF(TSV1) vs VPF(TSV2) for different transverse distances between the two truth secondary vertices

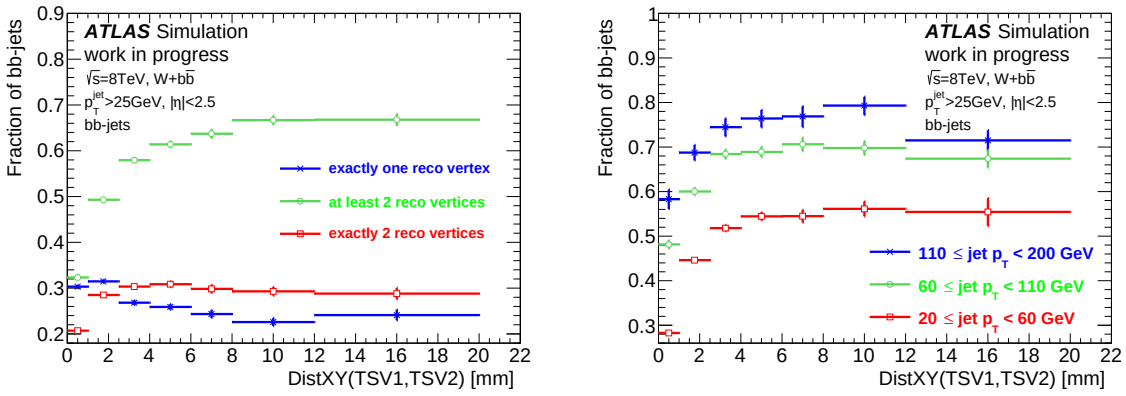


Figure 3.23.: Fraction of *bb*-jet as a function transverse distance between the truth secondary vertices. Fraction of *bb*-jets with exactly one reconstructed vertex, exactly 2 reconstructed vertices and at least 2 reconstructed vertices (left). Fraction of *bb*-jets with at least 2 reconstructed vertices divided in three categories of *bb*-jet p_T (right).

bb-jet identification and could potentially improve the rejection against other jets.

3.6. Development of MultiSVbb taggers

The MSVF algorithm was found suitable for the reconstruction of secondary vertices in bb-jets. In this section, a multivariate analysis is developed to increase the discrimination power between jets with two b hadrons and single-b jets, c-jets and light jets.

The MSVF algorithm is used to reconstruct multiple vertices inside a jet. To further improve the efficiency of tagging bb-jets, a new tagger (MultiSVbb) has been developed. This tagger exploits the properties of the two highest mass reconstructed vertices found by MSVF and defines two sets of variables that use kinematic properties of these vertices and topological variables as listed in table 3.3. To define these variables, jets with at least two reconstructed vertices with two or more tracks in the vertices are required. This cut serves as a first rejection against other jet flavours.

Two versions of the tagger were developed:

- MultiSVbb1: only vertex properties are used as input variables. There are 12 input variables with the complete list outlined in table 3.3.
- MultiSVbb2: includes in addition to the vertex properties, topological variables like the ΔR between the vertices. This version of the tagger has a higher b-jet rejection as shown in the next section. There are 14 input variables as listed in table 3.3.

Boosted decision trees are chosen to maximise the separation between bb-jets and other jets. The Toolkit for Multivariate Data Analysis (TMVA) [79] package is used. Before describing the MultiSVbb tagger, boosted decision trees are briefly explained below.

3.6.1. Boosted decision trees

Boosted decision trees (BDT) are a set of binary structured decision trees using the "boosting" technique. BDT are among the most popular learning techniques used in high energy physics.

3.6.1.1. Decision trees

Decision trees (DT) were developed and formalised by Breiman [80] in the context of data mining and pattern recognition. They consist of extending a simple cut-based analysis into a multivariate technique by continuing to analyse events that fail a particular criterion until they satisfy a terminating condition. A decision tree does not immediately reject events that fail a criterion but instead tries to find others features which may help to classify these events properly [81].

A DT classifies events between signal and background with a sequence of binary splits of the data, as show in figure 3.24. Starting from the root node, the algorithm splits recursively events into two branches using cuts on some discriminating variables x_i, until a stopping condition is satisfied. In each split the best separation variable is used.

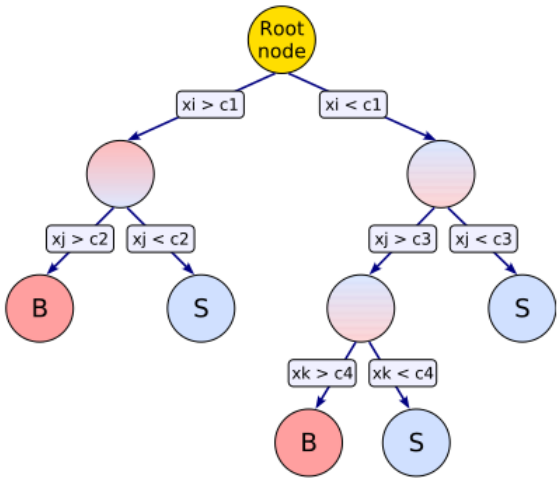


Figure 3.24.: Schematic view of a decision tree. Starting from the root node, a sequence of binary splits using the discriminating variable x_i is applied to the sample [79].

The goal during the training is to find the best split, S^*, between signal and background. It is done after scanning all sets of variables for the events at each node. The S^* is chosen from all splits, as the one that maximises the decrease of impurity:

$$\Delta i(S^*) = \max_{S \in \text{splits}} \Delta i(S), \tag{3.9}$$

where $\Delta i(S)$ is defined as :

$$\Delta i(S) = i - \min[p_{\text{P}} i_{\text{P}}, p_{\text{F}} i_{\text{F}}], \tag{3.10}$$

where p_{P} (p_{F}) is the fraction of events passing (failing) the split S. The most popular impurity definition in decision trees is: $i = 2p \cdot (1-p)$, called the *Gini-index*, where p is the signal purity defined as $s/(s+b)$, in which s (b) is the weighted number of signal (background) events.

The output of the decision tree is defined as the end-node's signal purity. The events with signal-like signature will give values close to 1, while the events with background-like signature will give values close to 0.

The number of parameters of a decision tree is relatively limited. When optimising a cut value on a variable over its full range, the number of intervals to evaluate the cuts is called *nCuts*. Decision trees have a maximal tree depth as condition to satisfy, a tree cannot have more than a certain number of layers. Finally a minimum number of events in each node after splitting is required to ensure statistical significance of the purity.

3.6.1.2. Boosting

Boosting combines many decision trees to form a single strong classifier. Different "boosting" algorithms are available for decision trees. In this thesis the Adaptive Boost (AdaBoost) [82] algorithm is used.

Starting with the original event weights when training the first decision tree (T_k), the subsequent tree is trained using a modified event sample where the weights of previously misclassified events are multiplied by a common boost weight α_k given by

$$\alpha_k = \beta \cdot \ln \frac{1-\epsilon_k}{\epsilon_k} \tag{3.11}$$

where β is a free boosting parameter to adjust boosting strength (1 in the original algorithm), and ϵ_k is the mis-classification (error) rate defined by

$$\epsilon_k = \frac{\sum_{i=1}^{N_k} w_i^k \cdot \text{isMisclassified}_k(i)}{\sum_{i=1}^{N_k} w_i^k} \tag{3.12}$$

where $\text{isMisclassified}_k(i)$ returns 1 when $y_i \cdot (T_k(i) - 0.5) \leq 0$, and 0 otherwise. y_i is a class label equal +1 for signal, -1 for background and $T_k(i)$ is the ith event associated with a weight w_i.

The weights of the entire event sample are renormalised such that the sum of weights remains constant.

The boosted event classification $y_{\text{Boost}}^{\text{Ada}}(\mathbf{x})$ is then given by

$$y_{\text{Boost}}^{\text{Ada}}(\mathbf{x}) = \frac{1}{\sum_i^{N_{\text{tree}}} \alpha_i} \cdot \sum_i^{N_{\text{tree}}} \alpha_i \cdot p_i(\mathbf{x}), \tag{3.13}$$

where $p_i(\mathbf{x})$ is the output of the ith decision tree with $\mathbf{x}$ being the set of input variables.

3.6.2. Multivariate analysis

3.6.2.1. Input variables

The input variables used for training the BDT were chosen in order to maximise the separation between bb-jets and other jets, while avoiding the use of variables not improving the performance significantly. Different kinematic and topological variables from the reconstructed vertices were investigated. The chosen input variables are listed in table 3.3.

The differences between b-jets and bb-jets are expected to arise from the presence of two b-hadrons in bb-jets leading to a higher number of reconstructed displaced vertices. Three groups of variables were defined as follows:

- Properties of vertex with maximum and second maximum mass:

  - Mass of the vertex: invariant mass of the charged particle tracks in the reconstructed vertex, assuming the pion mass for the individual particles.

- Energy Fraction: the energy of the charged particle tracks attached to the vertex divided by the sum of the energies of all charged particle tracks associated to the jet.

- Significance of the decay length of the vertex: the vertex position divided by its error, $\frac{d}{\sigma(d)}$.

- Topological information from the vertices with maximum mass and second maximum mass:

  - Transverse distance between vertices.

  - $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$ between vertices.

  - ΔR between the vertex and the jet axis.

  - Angle between the flight direction from the primary vertex of the two vertices.

- Properties per jet:

  - Transverse momentum of the jet (jet p_T).

  - Total number of reconstructed vertices inside the jet.

  - Sum of the mass of the reconstructed vertices.

  - Total number of tracks in all the reconstructed vertices inside the jet.

  - Difference of number of tracks between the total of number of tracks in all vertices reconstructed by the MSVF algorithm and the total number of tracks in the reconstructed vertex by the SSVF algorithm.

  - Significance of the decay length, the weighted average vertex position divided by its error.

Figures 3.25 to 3.28 show the input variables distributions comparing bb-, b-, cc-, c- and light jets.

3.6.2.2. Training the BDT

An MVA discriminator between bb-jets and different jet flavors was built by training a BDT. The sample of simulated jets is split in two parts: a training and a test sample, with exactly the same kinematic properties. The MVA discriminator is then applied on two independent samples to test for "overtraining". Overtraining occurs when there are too few data events to properly fit the model parameters. If over-trained, the MVA discriminator performance in the training sample and on an independent test sample will differ considerably.

A one-dimensional scan of each of the training parameters was performed to optimise the parameters. The main BDT parameters are:

- Number of trees in the BDT (NTrees) which is set to 250.

Variable	Description	MultiSVbb1	MultiSVbb2
Properties per jet:			
Jet p_T	Transverse momentum of the jet	✓	✓
Nvtx	Total number of reco vertices inside the jet	✓	✓
Total Mass	Sum of the mass of the vertices	✓	✓
Ntrks	Total number of tracks in reco vertices	✓	✓
Diff_ntrk_SSVF	Ntrks - total number of track in reco vertex from SSVF.	✓	✓
NormDist	Significance of the decay length averaged over all vertices	✓	✓
MaxEfrc	Maximum vertex energy fraction	–	✓
Properties of vertex with maximum (vtx1) and second maximum (vtx2) mass:			
Mass_vtx1	Mass of the vertex with maximum mass	✓	–
Mass_vtx2	Mass of the vertex with second maximum mass	✓	–
Efrc_vtx1	Energy fraction of the vtx1	✓	–
Efrc_vtx2	Energy fraction of the vtx2	✓	✓
Dls_vtx1	Significance of the decay length of the vtx1	✓	✓
Dls_vtx2	Significance of the decay length of the vtx2	✓	–
Topological information from the vertices:			
Distxy(vtx1,vtx2)	Transverse distance between vertices	–	✓
ΔR(vtx1,vtx2)	ΔR between vertices	–	✓
ΔR(vtx1, jet)	ΔR between vtx1 and the jet axis	–	✓
ΔR(vtx2, jet)	ΔR between vtx2 and the jet axis	–	✓
Angle(vtx1,vtx2)	Angle between the flight direction from the primary vertex of the two vertices	–	✓

Table 3.3.: List of input variables used in MultiSVbb taggers.

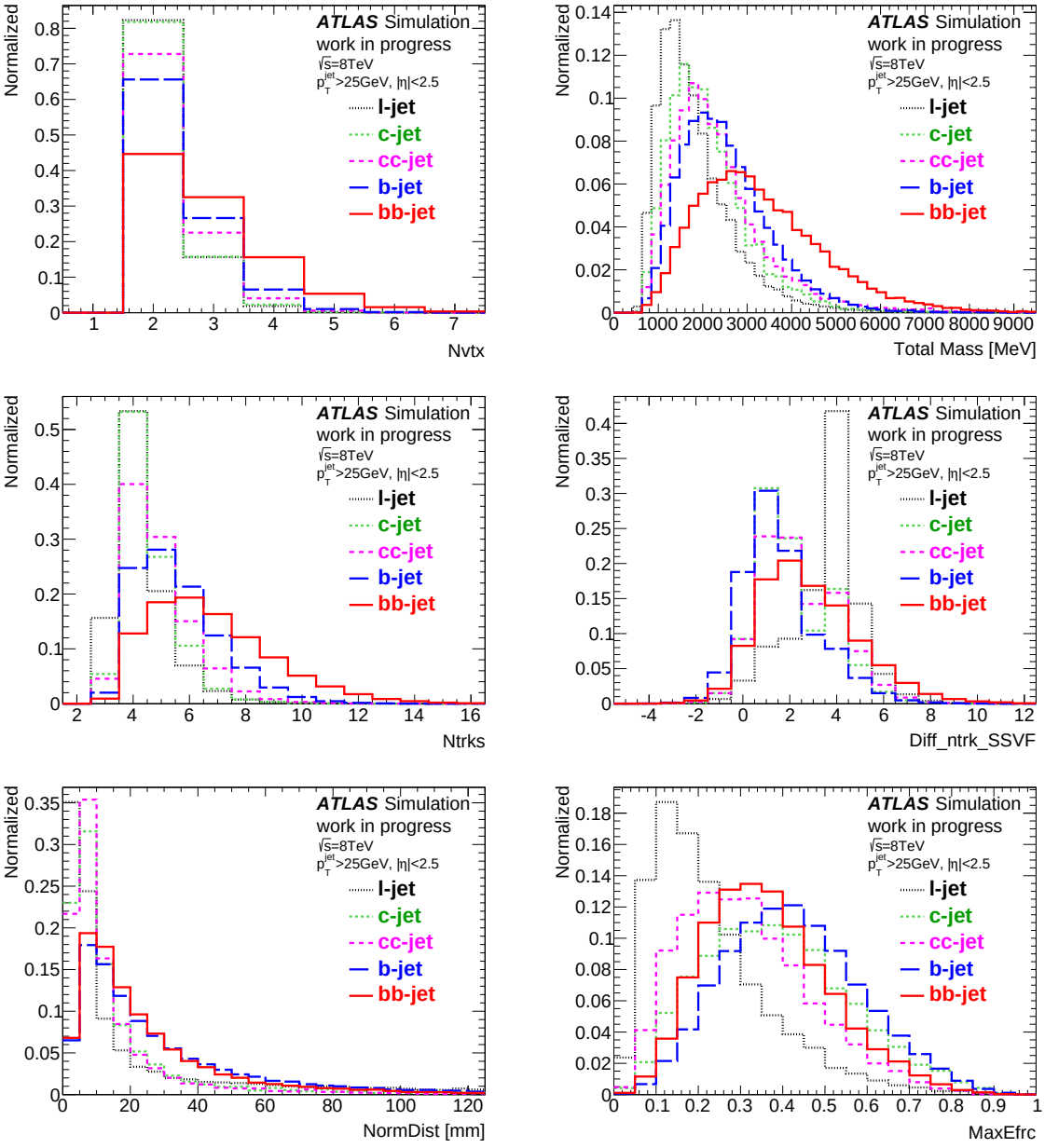


Figure 3.25.: MultiSVbb input variables distributions

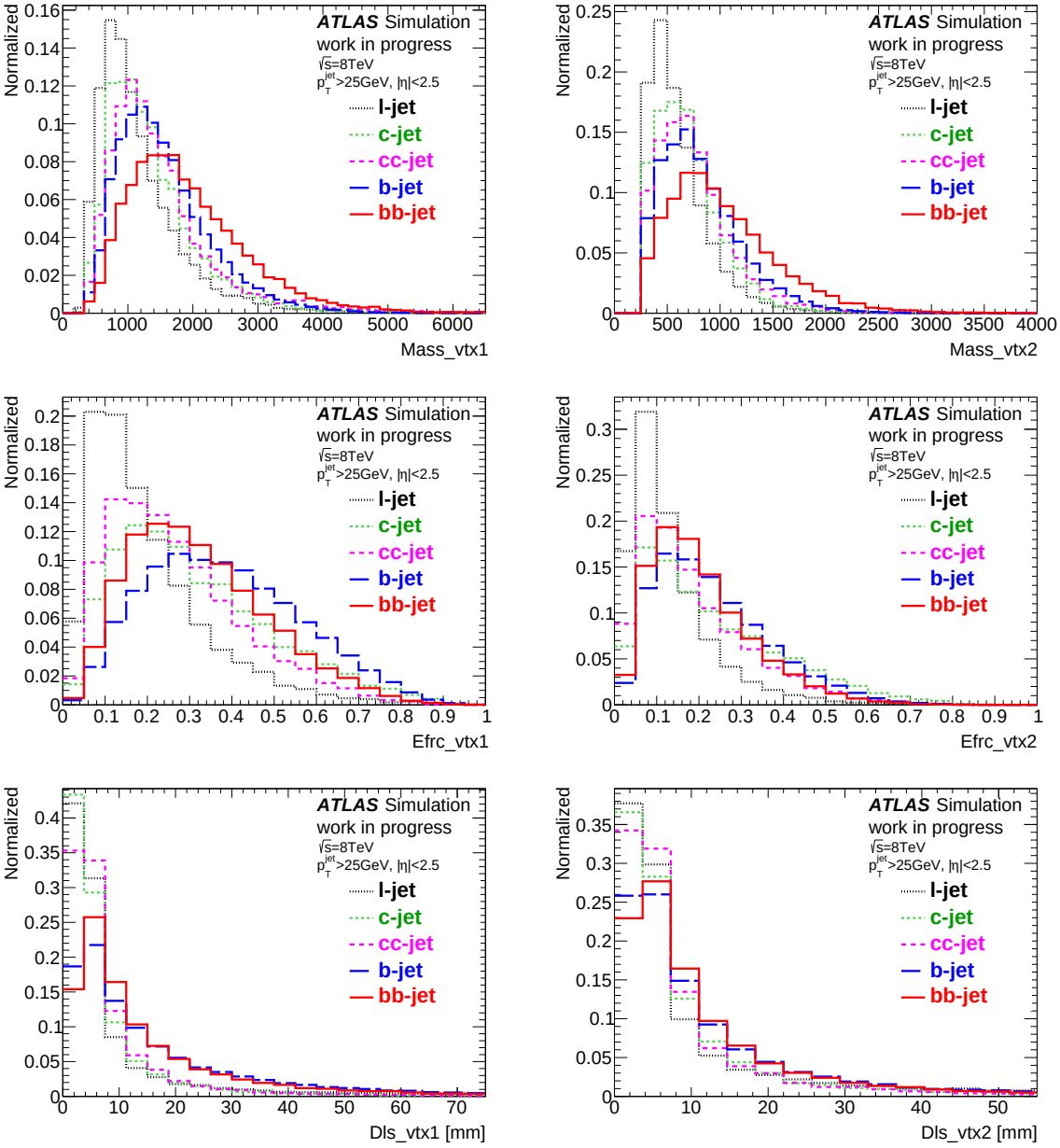


Figure 3.26.: MultiSVbb input variables distributions

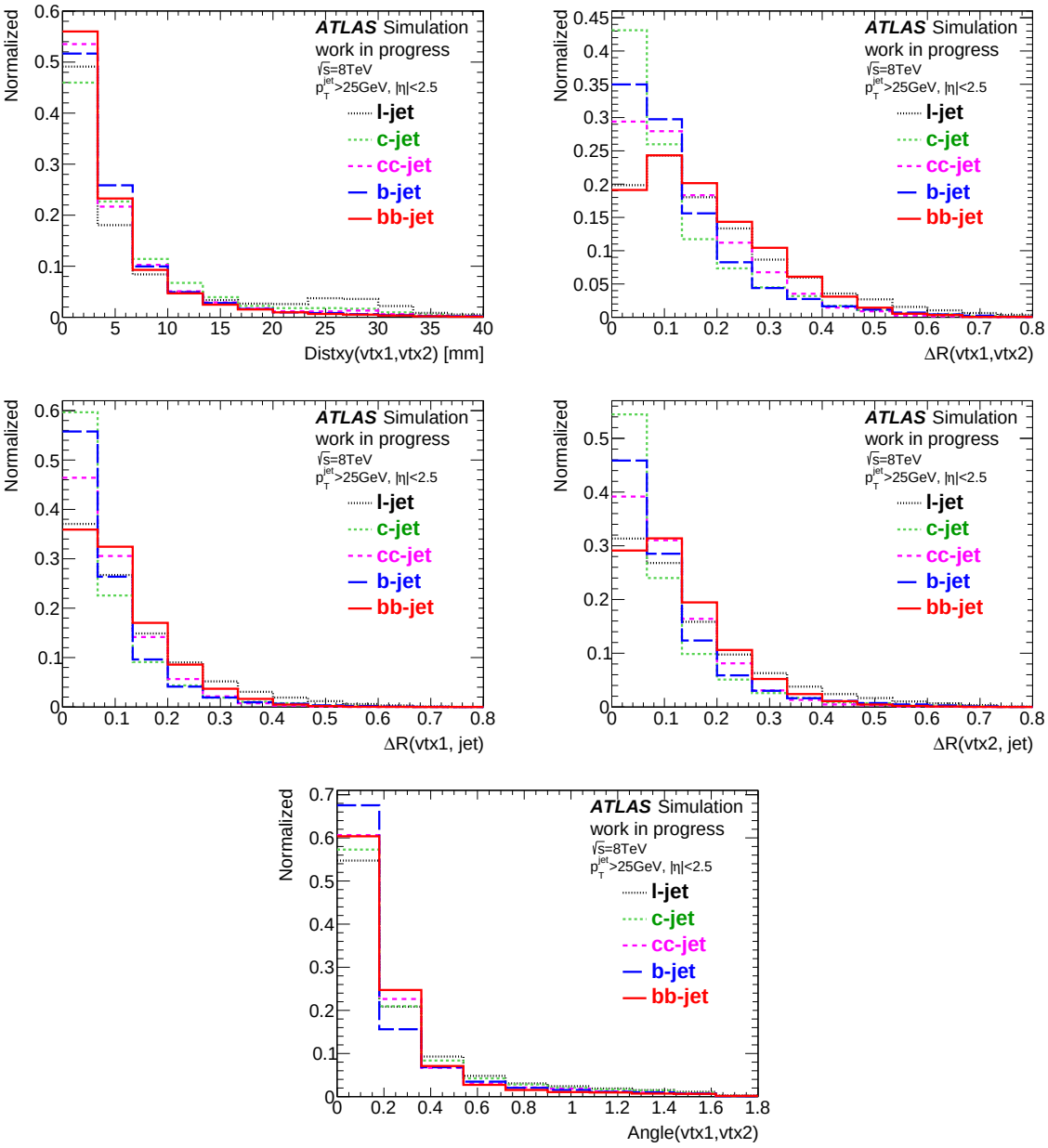


Figure 3.27.: MultiSVbb input variables distributions

- Maximum tree depth (MaxDepth) which is set to 4. This value is a compromise between a proper use of the available information from the input variables while maintaining short trees in order to avoid overtraining.
- Number of cuts which are tested per variable and per node (nCuts). A value of 100 is chosen.
- Minimum number of events in a node (MinNodeSize) which is set to 4% of the total number of events (around 93K (400K) signal (background) events were used in the training).

bb-jets are used as signal and a mixture of jet flavors (b-, c-, light- and cc-jets) for the background. Different background configurations were tested in order to optimise the b-jets rejection while keeping light jets rejection at a good rate. The background composition chosen was 37% b-, 9% c-, 16% cc- and 38% light-jets.

Since the BDT learns that the bb-jet populate in relatively high jet p_T compared to b-jets, a flattening weight was applied to minimise the jet p_T correlation of the BDT output. The jet p_T spectra of all flavors are flattened individually in all the flavors. The jet p_T profiles per flavour of the training sample before flattening are shown in figure 3.28.

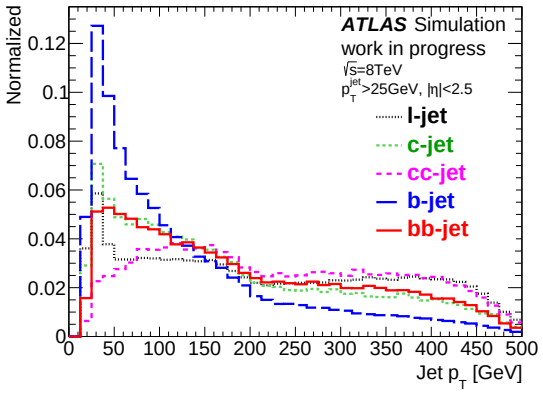


Figure 3.28.: Jet p_T distribution per flavor (normalised to the same area).

3.7. Performance of the MultiSVbb1 and MultiSVbb2 taggers

Figure 3.29 shows the MultiSVbb1 and MultiSVbb2 BDT output for the signal and background components. The MultiSVbb taggers provide better separation between bb-jets and other flavour jets than any individual discriminating variable.

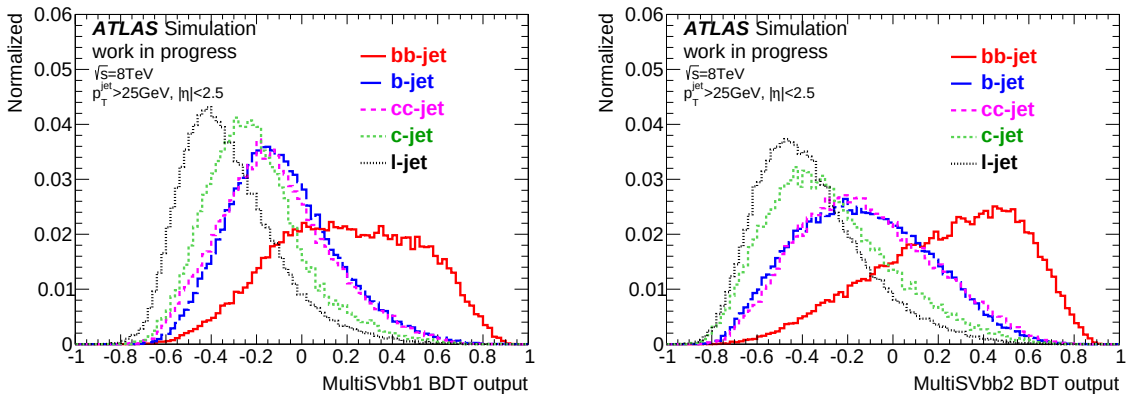


Figure 3.29.: MultiSVbb1 (left) and MultiSVbb2 (right) BDT output for bb-, b-, cc-, c- and light jets.

Figure 3.30 shows the rejection of different jet flavours versus the bb-jet efficiency for the MultiSVbb1 and MultiSVbb2 taggers. The rejection against cc-jets is significantly lower compared to c- or light jets because of the two c-hadrons have a real lifetime and thus similar topology as bb-jets. Figure 3.31 shows the MultiSVbb2 performance with respect to MultiSVbb1. MultiSVbb2 has higher rejection than MultiSVbb1 since it includes topological variables in the training.

A typical working point for the standard b-tagging is 70% b-jet efficiency. The bb-jet efficiency of the MultiSVbb taggers is limited by the vertexing efficiency which is around 50% after requiring at least two reconstructed vertices. This corresponds roughly to the efficiency of tagging two b-jets.

Table 3.4 shows the rejection factor at 35% bb-tagging efficiency for the two taggers, comparing the numbers with the MV1 tagger (default b-tagging algorithm in ATLAS Run 1). The b-jet rejection is about 18 (23) at 35% bb-jet efficiency for the MultiSVbb1 (MultiSVbb2) tagger. The MultiSVbb2 tagger performs 7 times better, in terms of b-jet rejection at the same bb-jet tagging efficiency, compared to MV1 which is not tuned to separate b- and bb-jets.

The dependence of the performance on the jet kinematics is of particular importance. Figure 3.32 shows bb-jet efficiency and b-jet rejection as a function of jet p_T using a global cut at 35% efficiency. The efficiency increases with jet p_T; at high jet p_T we have a relatively better track reconstruction and consequently better secondary vertex reconstruction. The b-jet rejection factor at global 35% bb-jet efficiency goes down as jet p_T increases, the effect is opposite at fixed efficiency as shown in figure 3.33.

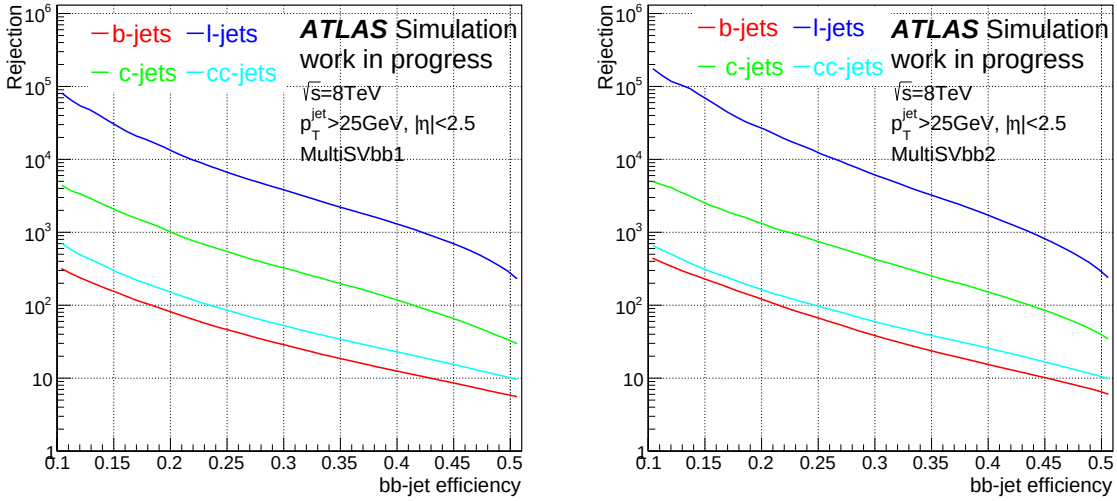


Figure 3.30.: Rejection versus *bb*-jet efficiency for MultiSVbb1 (left) and MultiSVbb2 (right)

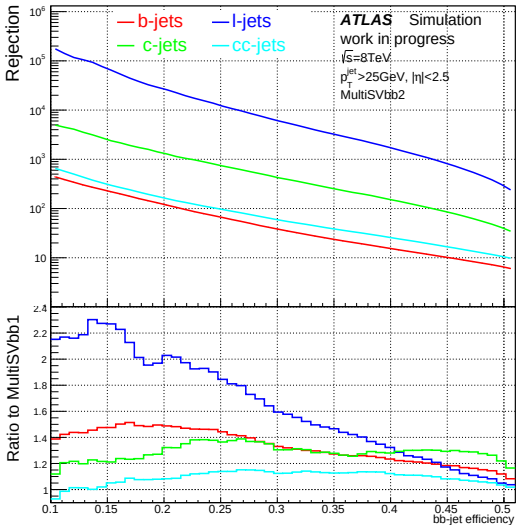


Figure 3.31.: MultiSVbb2 performance with ratio to MultiSVbb1

Rejection	MV1	MultiSVbb1	MultiSVbb2
b-jets	3	18	23
c-jets	40	200	250
l-jets	10000	2400	3200
cc-jets	40	35	38

Table 3.4.: Rejection at 35% of *bb*-jet efficiency.

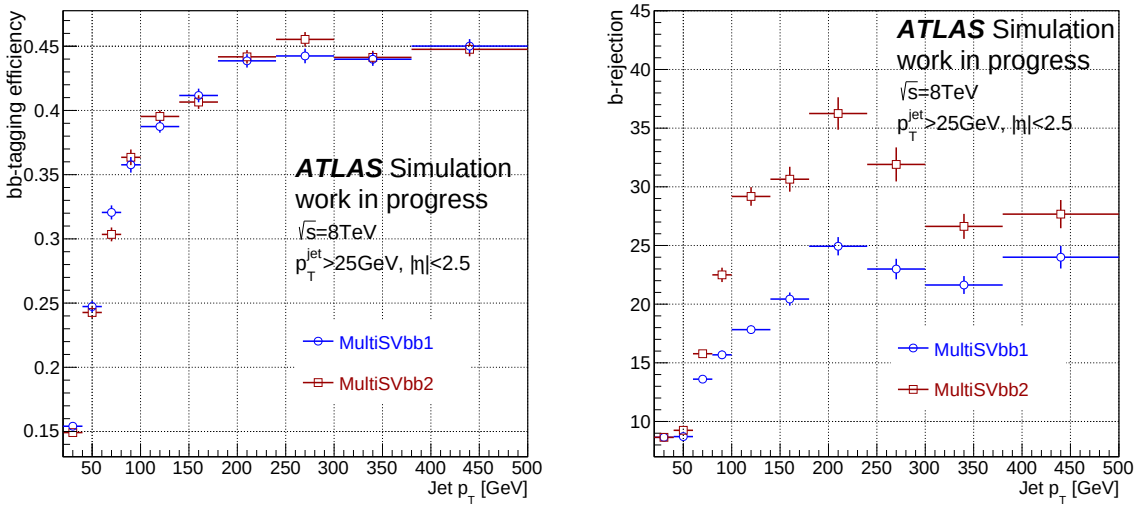


Figure 3.32.: *bb*-jet efficiency as function of p_T and *b*-jet rejection as function of p_T at global 35% *bb*-jet efficiency

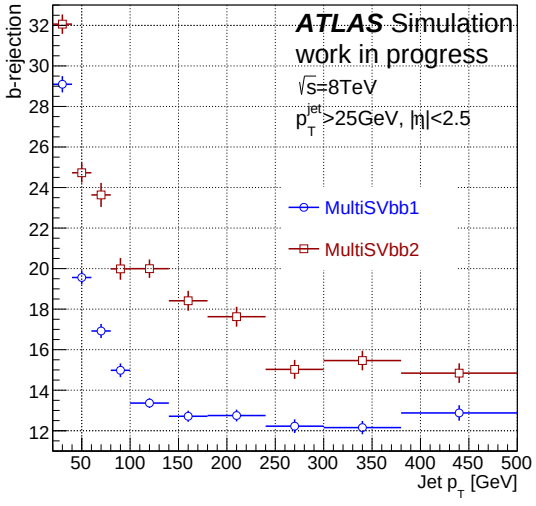


Figure 3.33.: *b*-jet rejection as function of p_T at fixed 35% *bb*-jet efficiency.

3.8. Summary

A new b-tagging tool (MultiSVbb) was developed to identify jets containing two b-hadrons (bb-jets). The method exploits the secondary vertices property differences between bb-jets and b-jets, combining the best discriminant variables with boosted decision trees (BDT). Several variables were investigated and different configurations of BDT were studied. Two configurations are retained (MultiSVbb1 and MultiSVbb2) with the best one (MultiSVbb2) performing 7 times better than the default b-tagging algorithm in ATLAS Run 1 (MV1) which does not separate b-jets from bb-jets. This b-tagging tool can be used to measure the $t\bar{t}$+bb-jets component in the $t\bar{t}$+b-jets processes.

4. Search for the Higgs boson in the single lepton $\mathrm{t\bar{t}H(H \to b\bar{b})}$ channel

This chapter describes the search for the associated production of a Higgs boson with a top quark-antiquark pair in the single lepton channel where the Higgs decays into a bottom quark-antiquark pair, using pp collisions data collected by the ATLAS experiment at centre-of-mass energy of 13 TeV.

A brief overview of the ATLAS+CMS Run 1 analysis in this channel is described in section 4.1. The data and simulated samples used for this analysis are explained in section 4.2. Section 4.3 describes the object selection. Section 4.4 describes the event selection and the splitting into several categories in order to increase the sensitivity of the search. The analysis strategy to separate signal and background events is described in section 4.5. The background estimation methods are introduced in section 4.6. The systematic uncertainties and their impact on the final fit results are discussed in sections 4.7-4.9, followed by conclusions in section 4.10.

4.1. Status of the $\mathrm{t\bar{t}H(H \to b\bar{b})}$ analysis

Events in the $t\bar{t}H(H \to b\bar{b})$ analysis are split into three different channel based on the decay of the top quark pair: the single lepton[a] channel ($t\bar{t}H \to (l\nu b)(q\bar{q}'b)(b\bar{b})$), the dilepton channel ($t\bar{t}H \to (l^-\nu b)(l^+\bar{\nu}b)(b\bar{b})$) and the full hadronic channel ($t\bar{t}H \to (q\bar{q}'b)(q\bar{q}'b)(b\bar{b})$).

Searches for the $t\bar{t}H(H \to b\bar{b})$ in the different channels using Run 1 data at centre-of-mass energy ($\sqrt{s}$) of 7 TeV and 8 TeV were published by the CMS and ATLAS collaboration:

- A search for the associated production of the Higgs boson with a top quark pair using several Higgs decay modes (including $H \to b\bar{b}$) carried out by the CMS collaboration at $\sqrt{s}$ of 7 TeV and 8 TeV can be found in ref. [83]. For the $t\bar{t}H(H \to b\bar{b})$ analysis, the single and dilepton channels were considered and boosted decision trees were used to further improve signal sensitivity.

- A search for $t\bar{t}H(H \to b\bar{b})$ using the matrix element method (MEM) by the CMS collaboration at $\sqrt{s}$ of 8 TeV is described in ref. [84]. Events with one or two

[a] In this chapter, the term "lepton" refers to electron and/or muon. Also taus which decay leptonically.

opposite charged leptons are selected. In order to separate the signal from the larger $t\bar{t}$+jets background, this analysis uses the MEM to assigns a probability density value to each event under the signal or background hypotheses. The results shows an improvement of about 15% in the expected limit compared to those obtained using the same data set and final state as the previous analysis [83].

- A search for $t\bar{t}H(H \to b\bar{b})$ by the ATLAS collaboration at $\sqrt{s}$ of 8 TeV is described in ref. [85], the search uses events containing one or two electrons or muons. A neural network is used to discriminate between signal and background events. In the single lepton channel, variables calculated using the matrix element method are included as inputs to the neural network to improve the discrimination of the irreducible $t\bar{t} + b\bar{b}$ background.

- Finally, a search for the full hadronic $t\bar{t}H(H \to b\bar{b})$ channel by the ATLAS collaboration at $\sqrt{s}$ of 8 TeV was recently published [86]. For this analysis, a data-driven method is used to estimate the dominant multijet background and boosted decision trees are used to discriminate the signal from the background.

The observed signal strengths $\mu = \sigma/\sigma_{SM}$, the ratio of the observed $t\bar{t}H$ production cross section relative to the value expected for a SM Higgs boson, for the individual $t\bar{t}H(H \to b\bar{b})$ channels and for their combination are summarised in figure 4.1. A combined signal strengths μ of 1.4 ± 1.0 is observed by the ATLAS experiment while the CMS experiment has a signal strengths μ of $1.2^{+1.6}_{-1.5}$.

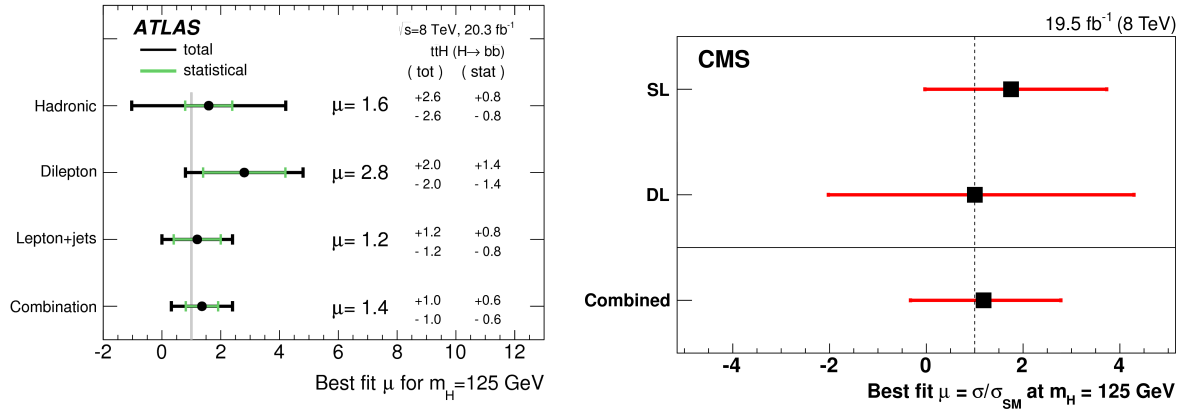


Figure 4.1.: Measurements of the signal strength $\mu = \sigma/\sigma_{SM}$ for $t\bar{t}H(H \to b\bar{b})$ production for the individual channels and for their combination by the ATLAS [86] (left) and CMS [84] (right) experiments.

The observed limits, and those expected with and without assuming SM Higgs boson with $m_H = 125$ GeV, for each channel and their combination are shown in figure 4.2. The best observed (expected) upper limit on $t\bar{t}H(H \to b\bar{b})$ was obtained by the ATLAS collaboration: $\mu < 3.4$ (2.2) at 95% confidence level (CL).

The ATLAS and CMS Run 1 results have been combined, resulting in evidence for the $t\bar{t}H$ production with a measurement (expected) significance of 4.4σ (2.0σ) and a combined signal strengths μ of $2.3^{+0.7}_{-0.6}$ [87].

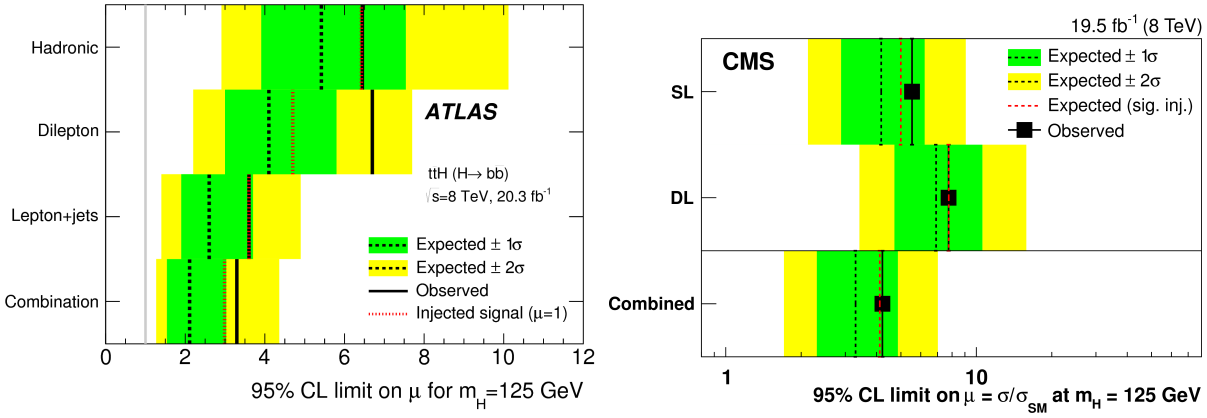


Figure 4.2.: Upper limits on the signal strength μ for the individual $t\bar{t}H(H \to b\bar{b})$ channels as well as for their combination, at 95% CL. The observed limits (solid lines) are compared to the expected (median) limits under the background-only hypothesis (black dashed lines) and under the signal-plus-background hypothesis assuming the SM prediction for $\sigma_{t\bar{t}H}$ (red dotted lines). The surrounding green and yellow bands correspond to the ± 1 standard deviation (s.d.) and ± 2 s.d. ranges, respectively, around the expected limits under the background-only hypothesis

The observation of $t\bar{t}H$ production is one of the major goals of the Higgs boson physics programme for the LHC Run 2. Increasing the centre-of-mass energy to 13 TeV results in a $t\bar{t}H$ production cross section 3.9 times larger than at 8 TeV, while the cross section for the dominant background, $t\bar{t}$ production, is increased by a factor of 3.3 [17].

A search for $t\bar{t}H(H \to b\bar{b})$ with 2.7 fb^{-1} of data recorded with the CMS detector in 2015 at $\sqrt{s} = 13$ TeV has recently been published [88]. This analysis combines the matrix element method with boosted decision trees to separate signal from background events. For the first time, methods for tagging hadronically decaying boosted particles are incorporated in this analysis. The signal strength $\mu = \sigma/\sigma_{SM}$ obtained is $\mu = -2.2 \pm 1.8$, compatible with the SM expectation. An observed (expected) upper limit of $\mu < 2.6$ (3.6) at the 95% CL was set.

4.2. Data and simulation samples

4.2.1. Data

The analysis is performed on pp collisions data recorded at $\sqrt{s} = 13$ TeV with the ATLAS experiment in 2015 and between April and July 2016. The corresponding integrated

luminosities are 3212.96 pb^{-1} and 9994.73 pb^{-1} respectively. Only the periods in which all the sub-detectors were operational are considered.

Triggers

The triggers used for this analysis are single electron triggers and single muon triggers [89]. Table 4.1 summarises the triggers used in the analysis for the 2015 and 2016 data taking periods. The triggers with the lower-p_T threshold include isolation requirements on the candidate lepton, while the triggers with the higher-p_T threshold no lepton isolation is required.

Event filter menu	Online object	p_{T} threshold [GeV]
2015		
`e24_lhmedium_L1EM20VH`	electron	24
`e60_lhmedium`	electron	60
`e120_lhloose`	electron	120
`mu20_iloose_L1MU15`	muon	20
`mu50`	muon	50
2016		
`e24_lhtight_nod0_ivarloose`	electron	24
`e60_lhmedium_nod0`	electron	60
`e140_lhloose_nod0`	electron	140
`mu24_ivarloose_L1MU15`	muon	24
`mu40`	muon	40

Table 4.1.: Single electron and muon triggers used for the analysis.

4.2.2. Simulated samples

Monte-Carlo events have been generated through the ATLAS simulation software chain as explained in section 2.4. The event generators used for the signal and background samples are listed in table 4.2.

4.2.2.1. Signal samples

The $t\bar{t}H$ signal process is modelled using MadGraph5_aMC@NLO [90] (referred to in the following as MG5_aMC) with a NLO matrix element. They are inclusive in Higgs boson decays and are produced with the `NNPDF3.NLO` [91] parton distribution function (PDF) set using factorisation (μ_{F}) and renormalisation (μ_{R}) scales set to $\mu_{\mathrm{F}} = \mu_{\mathrm{R}} = H_T/2$, where H_T is defined as the scalar sum of the transverse masses $\sqrt{p_T^2 + m^2}$ of all final state particles. The Higgs mass is set to 125 GeV. Generated events are interfaced

Process	Generator	Shower	PDF	Tune
		Signal		
$t\bar{t}H$	MG5_aMC	Pythia 8.210	NNPDF3.NLO	A14
		Top-quark		
$t\bar{t}$	Powheg-Box	Pythia 6.428	CT10	Perugia2012
t-channel single top	Powheg-Box	Pythia 6.428	CT10f4	Perugia2012
s-channel single top	Powheg-Box	Pythia 6.428	CT10	Perugia2012
Wt-channel single top	Powheg-Box	Pythia 6.428	CT10	Perugia2012
		V + jets		
W + jets	Sherpa	Sherpa 2.1.1	CT10	Sherpa
Z + jets	Sherpa	Sherpa 2.1.1	CT10	Sherpa
		$t\bar{t}V$		
$t\bar{t}V$	MG5_aMC	Pythia 8.210	NNPDF3.NLO	A14
		Diboson + jets		
WW + jets	Sherpa	Sherpa 2.1.1	CT10	Sherpa
WZ + jets	Sherpa	Sherpa 2.1.1	CT10	Sherpa
ZZ + jets	Sherpa	Sherpa 2.1.1	CT10	Sherpa

Table 4.2.: Processes considered in the analysis and the event generators used for the MC simulation.

with Pythia 8.210 [36] for the parton shower model using the A14 [77] tune for the underlying event (UE tune). The $t\bar{t}H$ cross section and the Higgs boson decay branching fractions are taken from (N)NLO theoretical calculations, collected in ref. [17].

4.2.2.2. $t\bar{t}$+jets background

The $t\bar{t}$+jets sample is generated using the Powheg-Box v2 NLO generator [92–94] with the CT10 PDF set [95]. Parton shower and hadronisation are modelled by Pythia 6.428 [75] with the CTEQ6L1 PDF set [96] and the Perugia2012 (P2012) [97] UE tune. The EvtGen v1.2.0 [76] program is used to simulate the bottom and charm hadron decays. The sample is normalised to the Top++2.0 [98] theoretical cross section of 832^{+46}_{-51} pb, calculated at next-to-next-to-leading order (NNLO) in QCD [99–103].

Alternative $t\bar{t}$ samples are used to derive systematic uncertainties and to reweight the nominal Powheg-Box+Pythia 6 sample. They are described in section 4.7.2.

4.2.2.3. Other backgrounds

Samples of W/Z+jets events, and diboson production in association with jets, are generated using Sherpa 2.1.1. In the W/Z+jets samples, matrix elements are calculated for up to two partons at NLO and four partons at leading order (LO) using the Comix [104] and OpenLoops matrix element generators and merged with the Sherpa parton shower [105] using the ME+PS@NLO prescription [106]. The CT10 PDF set is used. The W/Z + jets events are normalised to the NNLO cross sections [107]. The diboson+jets samples are generated following the same approach but with up to one additional parton at NLO and up to three additional partons at LO. They are normalised to their respective NLO cross sections calculated by the generator.

Samples of Wt and s-channel single top quark backgrounds are generated with Powheg-Box 2.0 using the CT10 PDF set. Overlaps between the $t\bar{t}$ and Wt final states are removed [108]. Electroweak t-channel single top-quark events are generated using the Powheg-Box v1 generator which uses the four-flavour scheme for the NLO matrix elements calculations together with the fixed four-flavour PDF set CT10f4. All single top quark samples are interfaced to Pythia 6.428 with the Perugia 2012 underlying-event tune. The EvtGen v1.2.0 program is used to model properties of the bottom and charm hadron decays. The single top quark t- and s-channel samples are normalised to the approximate NNLO theoretical cross sections [109–111].

Samples of $t\bar{t}V$ events are generated using MG5_aMC with up to two additional partons and interfaced to Pythia 8.210 with NNPDF3.0NLO PDF set and the A14 UE tune.

4.3. Object selection

The main physics objects considered in this analysis are electrons, muons, jets and b-jets. The reconstruction of these objects is described in section 2.5. Below the different requirements of the physics objects are discussed.

Electrons

Electrons must pass a tight likelihood identification criterion (`TightLH`) [54] and further selections on the transverse and longitudinal impact parameters: $|\frac{d_0}{\sigma(d_0)}| < 5$ and $|z_0 \sin\theta| < 0.5$ mm. Electron must have $p_{\mathrm{T}} > 25$ GeV and $|\eta| < 2.5$. To reduce the background from non-prompt electrons (e.g. from decays of hadrons produced in jets), electron candidates are also required to be isolated [112].

Muons

Muons must satisfy "medium" quality [53] and "Gradient" isolation requirements [113]. The absolute value of a muon's d_0 significance must be less than 3, and the value of $|z \sin\theta|$ must be less than 0.5 mm. Muon must have $p_{\mathrm{T}} > 25$ GeV and $|\eta| < 2.5$.

Jets

The reconstructed jets are calibrated to the particle level by the application of a jet energy scale (JES) derived from simulation and *in situ* corrections based on 13 TeV data [114, 115]. After energy calibration jets are required to have $p_{\mathrm{T}} > 25$ GeV and $|\eta| < 2.5$. Quality criteria (also called *jet cleaning*) are imposed to identify jets arising from non-collision sources or detector noise (using the `LooseBad` operating points) and any event containing at least one such jet is removed [116]. To avoid selecting jets from additional collisions within the same bunch crossing, an additional requirement on the tracks associated to the jet [52] is made for low p_{T} ($p_{\mathrm{T}} < 60$ GeV) jets in the central ($|\eta| < 2.4$) region of the detector: such jets must have JVT > 0.59.

b-jets

Jets are identified as originating from the hadronisation of a b quark (b-tagged) via the MV2c10 tagger using the 70% working point. This corresponds to a 70% efficiency to tag a b-jet, with a light-jet rejection factor of 381 and a charm jet rejection factor of 12, as determined for b-tagged jets with $p_{\mathrm{T}} > 20$ GeV and $|\eta| < 2.5$ in simulated $t\bar{t}$ events. Tagging efficiencies in simulation are corrected to match the results of the calibration performed in data [66, 71].

Overlap removal

To avoid double counting of a single detector response, an overlap removal procedure is used. During jet reconstruction, no distinction is made between identified electrons and jet energy deposits. Therefore, if any of the jets lie within ΔR of 0.2 of a selected electron, the single closest jet is discarded in order to avoid double-counting of electrons as jets. After this, electrons which are within ΔR of 0.4 of a remaining jet are removed. Muons are required to be isolated and be separated by $\Delta R > 0.4$ from the nearest selected jet. However, if this jet has fewer than three associated tracks, the muon is

kept and the jet is removed instead to avoid an inefficiency for high-energy muons undergoing significant energy loss in the calorimeter.

4.4. Event selection and categorisation

The event selection is designed to select a sample enriched in $t\bar{t}$ events. Events are required to contain exactly one lepton with a p_{T} above 25 GeV and at least 4 jets. At least two of the jets must be b-tagged. Selected events are then classified based on the number of jets and the number of b-tagged jets.

The regions with a large signal-to-background ratio S/B and $S/\sqrt{B}$ are referred to as "signal-rich" regions, as they provide most of the sensitivity to the signal. The remaining regions are referred to as "signal-depleted" or "control" regions. They are very pure background-only regions and are used to constrain systematic uncertainties, thus improving the background prediction in the signal-rich regions. The regions are analysed separately and combined statistically to maximise the overall sensitivity. In the most sensitive region ($\geq$6 jets, $\geq$4 b-tags), $H \to b\bar{b}$ decays are expected to constitute about 90% of the signal contribution, with the other Higgs boson decay modes also treated as signal.

As a baseline and following the Run 1 analysis, a total of nine independent regions are considered: six signal-depleted regions, (4 jets, 2 b-tags), (4 jets, 3 b-tags), (4 jets, 4 b-tags), (5 jets, 2 b-tags), (5 jets, 3 b-tags), ($\geq$6 jets, 2 b-tags), and three signal-rich regions, (5 jets, $\geq$4 b-tags), ($\geq$6 jets, 3 b-tags) and ($\geq$6 jets, $\geq$4 b-tags). Figure 4.3 shows the $S/\sqrt{B}$ and S/B ratios for the different regions under consideration based on the simulations described in section 4.2. A clear separation between "signal-rich" and "signal-depleted" regions can be noted, though even the regions with the most signal still have a relatively small S/B ratio.

The expected proportions of different backgrounds in each region are shown in figure. 4.4. The $t\bar{t}$ background is divided in different categories, as described in section 4.6.1. The main background contribution in the signal regions corresponds to the $t\bar{t}+\geq 1b$ category.

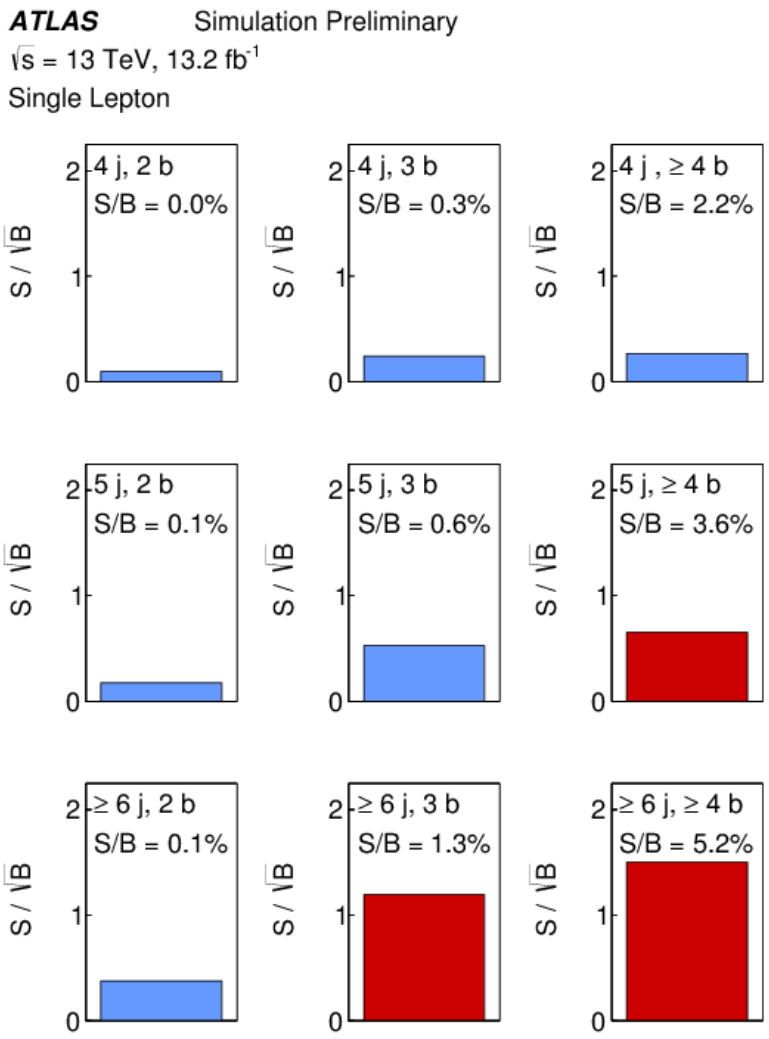


Figure 4.3.: The S/B and $S/\sqrt{B}$ ratio for each of the regions assuming SM cross sections and branching fractions, and $m_H = 125$ GeV. Each row shows the plots for a specific jet multiplicity (4, 5, $\geq$6), and the columns show the b-jet multiplicity (2, 3, $\geq$4). Signal-rich regions are shaded in dark red, while the rest are shown in light blue [117].

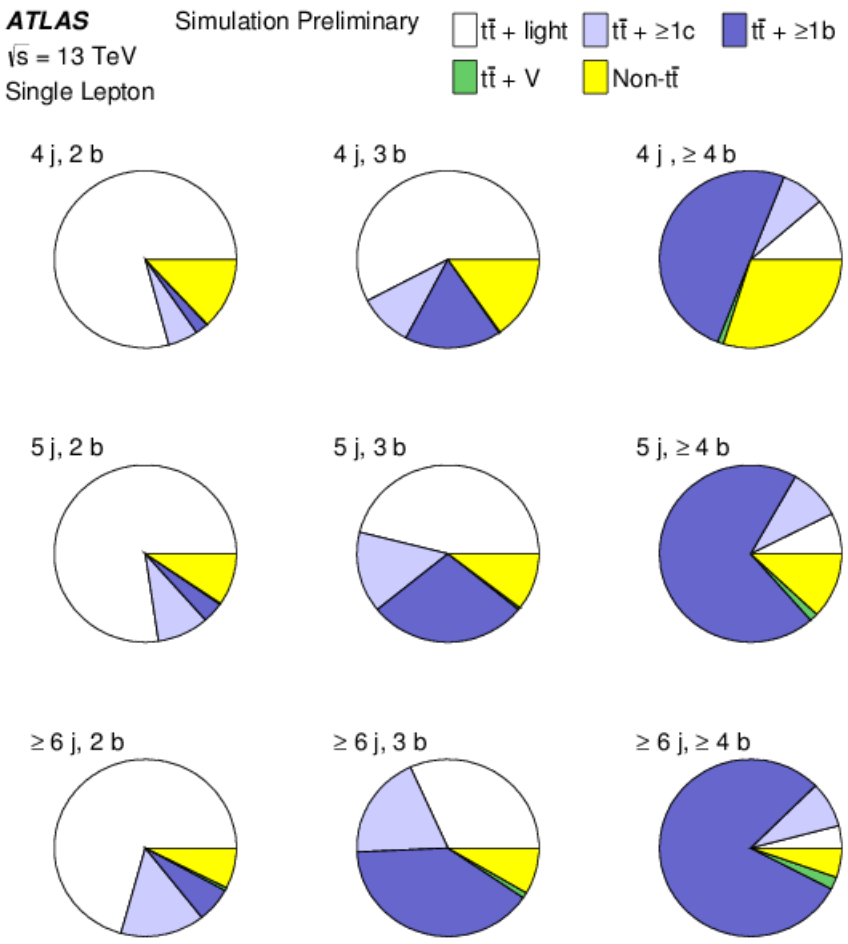


Figure 4.4.: The fractional contributions of the various backgrounds to the total background prediction in each considered region. Each row shows the plots for a specific jet multiplicity (4, 5, ≥6), and the columns show the b-jet multiplicity (2, 3, ≥4) [117].

4.5. Multivariate analysis

The small signal-to-background ratio after the event selection and categorisation mean it is essential to use multivariate techniques (MVA) to further discriminate the signal process from the background.

The Kinematic Likelihood Fitter (KLFitter) algorithm [118] was used in an early analysis performed with the 2011 dataset to reconstructed the entire final state of the $t\bar{t}H(H \rightarrow b\bar{b})$ system [119]. Jets were assigned to the final state partons of the $t\bar{t}$ decay and the remaining b-tagged jets not assigned to the $t\bar{t}$ hypothesis were considered as candidate jets for the Higgs boson decay, with their invariant mass used as the final discriminant in the analysis. The reconstructed Higgs matching efficiency, defined as the subset of events where the two b-tagged jets considered as candidate jets for the Higgs decay are matched to the b-quarks from the Higgs decay, was around 20% and the subset of events where all jets considered in the kinematics fit match the partons from the decays of the top quarks and the Higgs boson was 7.5%. The low matching efficiency is a result of the large jet combinatorics as well as the fact that all products from the $t\bar{t}H(H \rightarrow b\bar{b})$ system might not be present in the selected event because of the requirement on the jet p_T or the η acceptance.

Due to the low efficiency, the kinematic reconstruction of the $t\bar{t}H(H \rightarrow b\bar{b})$ system was not included for the final Run 1 $t\bar{t}H(H \rightarrow b\bar{b})$ analysis instead a multivariate approach was used in the signal-regions. Using several variables as inputs, the MVA produces one output discriminant that exploits the correlations among the input variables. The distribution of the MVA output was used as the final discriminant in the signal-regions in the profile likelihood fit.

One of the main contribution of the work done during this thesis was the development of a new MVA technique to reconstruct the $t\bar{t}H$ system for the Run 2 analysis. These studies make use of boosted decision trees (BDT), as implemented in the TMVA package [79]. In order to increase the signal-to-background separation, it is important to build a correspondence between the reconstructed jets and the final-state quarks of the hard-scattering process. The reconstruction MVA method was optimised to find the best match between the observed jets and the final-state partons from the $t\bar{t}H(H \rightarrow b\bar{b})$ system. In the following these MVAs are referred as "reconstruction BDTs" while the name “classification BDT” is reserved for the final discriminant between the $t\bar{t}H$ signal and the background processes.

Both reconstruction and classification BDTs are separately trained in the three signal-rich regions: (5 jets, $\geq$4 b-tags), ($\geq$6 jets, 3 b-tags) and ($\geq$6 jets, $\geq$4 b-tags).

4.5.1. MVA-based event reconstruction

In each of the signal regions, a full event reconstruction is performed using BDT. For this purpose only the $t\bar{t}H(H \rightarrow b\bar{b})$ simulation sample is used.

4.5.1.1. Truth Matching

For these studies performed, it is necessary to identify jets with the corresponding quarks from the hard scattering process using MC truth matching. This identification is done by requiring a geometric matching based on the spatial distance $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$. A jet is matched to a parton if the ΔR between the jet and the final state parton is less than 0.3. The lepton match is not considered since the lepton is correctly selected in most of the cases.

The fraction of selected events satisfying the different matching requirements are shown in figure 4.5. The bins in the figure are defined as:

- "all": all matched, the six selected jets in the event are matched to the six partons in the final state of the $t\bar{t}H(H \to b\bar{b})$ process.
- "b+1W": four selected jets are matched to the four b-quarks from the top and Higgs decay and one selected jet matches to one quark from the W decay.
- "all b": four selected jets are matched to the four b-quarks from the top and Higgs decay.
- "Higgs": two selected jets are matched to the two b-quarks from Higgs decays.
- "btop": two selected jets are matched to the two b-quarks from $t\bar{t}$ decay.
- "W": two selected jets are matched to the two quarks from hadronic W decay.
- "Hb1": one jet is matched to the leading, in p_T, b-quark from Higgs decay.
- "Hb2": one jet is matched to the sub-leading, in p_T, b-quark from Higgs decay.
- "blt": one jet is matched to the b-quark from leptonic Top decay.
- "bht": one jet is matched to the b-quark from hadronic Top decay.
- "wj1": one jet is matched to the leading, in p_T, quark from hadronic W decay.
- "wj2": one jet is matched to the sub-leading, in p_T, quark from hadronic W decay.

Only 42% of the selected events have all the products from the $t\bar{t}H(H \to b\bar{b})$ decay matching to the reconstructed jets in the most sensitive region ($\geq$ 6jets, $\geq$ 4b-tags). The low fraction can be explained by the fact that not all products from the $t\bar{t}H(H \to b\bar{b})$ decay are present in the event, due to the acceptance of the detector and to the different requirements in the physics objects.

The fraction of events matching the sub-leading quark from W is about 56% in the region with at least 6 selected jets because the sub-leading quark from the hadronic W have a low p_T and hence a large fraction of the jets originated by this quark are removed by requiring jet $p_T > 25$ GeV. This effect is much larger in the region with 5 selected jets. As expected, in the region with $\geq$ 6jets, 3b-tags, the fraction of events matching the b-quarks from the $t\bar{t}H(H \to b\bar{b})$ decay decreases.

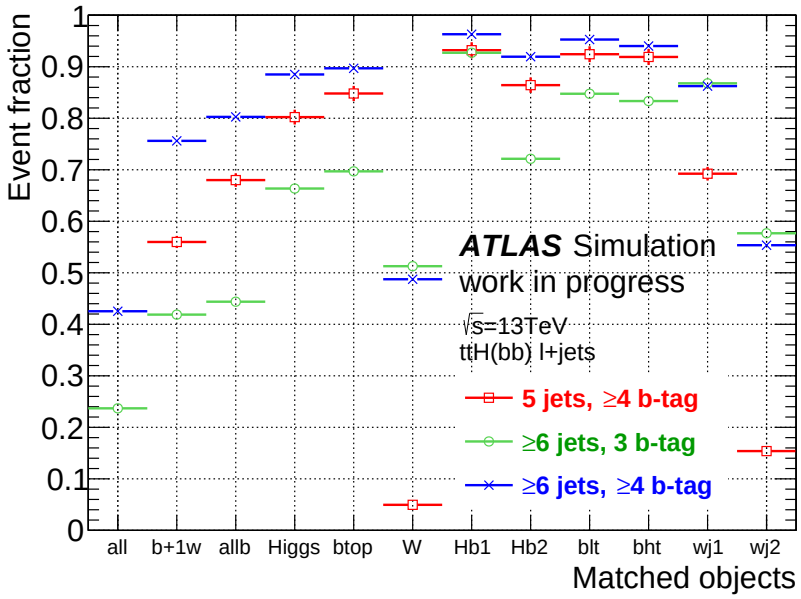


Figure 4.5.: Fraction of selected events satisfying the different matching requirements (see text). These values correspond to the maximum achievable matching efficiency for the reconstruction method assuming perfect identification.

4.5.1.2. Reconstruction algorithm

In each of the three signal-rich categories, the reconstructed jets, the missing transverse energy and one lepton are used to reconstruct the different objects of the $t\bar{t}H(H \to b\bar{b})$ system. Jets are assigned to the quarks from $t\bar{t}H(H \to b\bar{b})$ decay and combinations containing jets and b-tagged jets are used to reconstruct the objects (e.g. the hadronic Top, the Higgs boson, etc.). All jets are considered in the combinations.

Reconstruction of the lepton W boson.

To reconstruct the leptonic W boson, the neutrino momentum is needed. The neutrino transverse momentum can be measured using the imbalance of the transverse energy in the event (missing transverse energy). However, the longitudinal component of the neutrino momentum ($p_{z\nu}$) is not measurable given that the sum of the p_z of the two partons in the hard scattering is not known.

The sum of the lepton and neutrino four momentum is equal to the W four momentum. Hence, constraining the mass of the neutrino-lepton system by the true W boson mass ($M_W = 80.385$ GeV [14]) one can compute the $p_{z\nu}$. It leads to a quadratic equation with two possible solutions:

$$p_{z\nu}^{\pm} = \frac{1}{2}\frac{p_{zl}\beta \pm \sqrt{\Delta}}{E_l^2 - p_{zl}^2}, \tag{4.1}$$

where:

$$\beta = M_W^2 - M_l^2 + 2p_{xl}p_{x\nu} + 2p_{yl}p_{y\nu}, \tag{4.2}$$

$$\Delta = E_l^2(\beta^2 + (2p_{zl}p_{T\nu})^2 - (2E_l p_{T\nu})^2), \tag{4.3}$$

when there is no real solution (around 20% for the $t\bar{t}H$ sample used), the discriminant of the quadratic equation is set to zero (Δ=0). Different approximation were studied in ref. [120] showing that the best $p_{z\nu}$ resolution is obtained with the approximation Δ=0.

With the calculated $p_{z\nu}$, it is possible to reconstruct the leptonic W boson and in cases of two solutions, two different leptonic W bosons are considered.

Reconstruction of the hadronic W boson.

The hadronic W is reconstructed using all combinations of 2 jets that are not considered as b-tagged jets in the selection. If the event contains less than two non-b-tagged jets, a b-tagged jet is then allowed to be used for reconstructing of the hadronic W boson. This happens in a very small fraction of events (less than 1%).

The hadronic W boson is not reconstructed in the region with 5 selected jets since in most of the cases we do not have a jet originating from the sub-leading quark from the W, as shown in figure 4.5.

Reconstruction of the Top quarks and the Higgs boson

The Top quarks are reconstructed by association of one W boson and one b-tagged jet. The remaining b-tagged jets are used to reconstruct the Higgs boson. If an event contains less than 4 b-tagged jets (in the $\geq$6 jets, 3 b-tags region) one and only one non-b-tagged jet is allowed to be used for the reconstruction of one of the Top quarks or the Higgs boson.

In the region with 5 selected jets, the hadronic Top is reconstructed using one b-tagged jet and one non-b-tagged jet. This reconstructed object is referred to as the incomplete hadronic Top.

4.5.1.3. BDT technique for combinatorial solving

The procedure described in section 4.5.1.2 considers all possible permutations of b-tagged and non-b-tagged jets to assign them to the quarks from the $t\bar{t}H(H \to b\bar{b})$ decay. If all jets in a combination are matched to the appropriate quarks, the combination is considered correct.

Boosted Decision Trees (BDT) are used to find the correct combination. Thus, for training the BDT, the correct combination represents the signal. However, due to the small fraction of event with all jets matching the six quarks of the $t\bar{t}H(H \to b\bar{b})$ system, as shown in figure 4.5, combinations with only one of the jets used for the hadronic W reconstruction is not correctly assigned are considered as signal as well. All other

different jets combinations represent the background. These requirements increases the number of entries in the signal category and improves the overall performance.

Natural discriminating variables between the $t\bar{t}H(H \to b\bar{b})$ signal and the dominant $t\bar{t} + b\bar{b}$ background can be defined using the b-quark pair not originating from the top quarks (e.g. the reconstructed Higgs invariant mass). However, the chosen jet combination from the BDT which includes information related to the Higgs boson in the training (recoBDT_withHiggs) biases the distribution of the candidate Higgs mass variable in the $t\bar{t}+b\bar{b}$ background to be closer to the signal, reducing its discriminant power. Therefore, two versions of the reconstruction BDT are used in each signal-enriched regions.

- **Reconstruction BDT (recoBDT)**, targeted to match jets to the four quarks from the decay products of the $t\bar{t}$ system, variables depending only on the top-quark pair system are used in the training of recoBDT.
- **Reconstruction BDT with Higgs related variables (recoBDT_withHiggs)**, it attempts to match the reconstructed jets that correspond to the six quarks of the $t\bar{t}H(H \to b\bar{b})$ system. In addition to variables used in recoBDT it includes variables correlated with the Higgs boson in the training (like the Higgs boson candidate mass).

BDT training

The input variables used for the BDT training are listed in Table 4.3. The variables have been chosen to reveal particular kinematic characteristics of the correct and wrong jet combinations. Figures 4.6 to 4.11 show all input variables for the signal and background both normalised to the same integral.

The total available events are divided into two samples (sample A and sample B) for training and evaluation based on the event number. Cross training is used to profit from the full available statistics in the evaluation step: evaluate events in sample B with the BDT trained on sample A and the opposite. Figure 4.12 show the corresponding receiver operating characteristic (ROC) curves for the the evaluation on sample A, sample B and the combination A+B. This acts as a validation test to see if there is any bias in the response from statistical fluctuations at the moment to split the sample. No large differences have been seen between the two sets in the BDT responses.

The BDT parameters are optimised to get the best possible reconstruction efficiency. An optimal set of parameters for the different reconstruction BDTs is given in Table 4.4.

Variable	Region		
	≥6 jets, ≥4 b-tags	≥6 jets, 3 b-tags	5 jets, ≥4 b-tags
Topological information from $t\bar{t}$:			
Leptonic Top mass	✓	✓	✓
Hadronic Top mass	✓	✓	–
Incomplete hadronic Top mass	–	–	✓
Hadronic W mass	✓	✓	–
Mass of hadW and blepTop	✓	✓	–
Mass of qhadW and blepTop	–	–	✓
Mass of lepW and bhadTop	✓	✓	✓
ΔR(hadW, bhadTop)	✓	✓	–
ΔR(qhadW, bhadTop)	–	–	✓
ΔR(hadW, blepTop)	✓	✓	✓
ΔR(qhadW, blepTop)	–	–	✓
ΔR(lep, blepTop)	✓	✓	✓
ΔR(lep, bhadTop)	✓	✓	✓
ΔR(blepTop, bhadTop)	✓	✓	✓
ΔR(q1hadW, q2hadW)	✓	✓	–
ΔR(bhadTop, q1hadW)	✓	✓	–
ΔR(bhadTop, q2hadW)	✓	✓	–
ΔR^{min}(bhadTop, $q_i$hadW)	✓	✓	–
ΔR^{min}(bhadTop, $q_i$hadW) - ΔR(lep, blepTop)	✓	✓	–
ΔR(bhadTop, qhadW) - ΔR(lep, blepTop)	–	–	✓
Topological information from Higgs :			
Higgs mass	✓	✓	✓
ΔR(b1Higgs, b2Higgs)	✓	✓	✓
ΔR(b1Higgs, lep)	✓	✓	✓
Mass of Higgs and q1hadW	✓	✓	✓
ΔR(b1Higgs, bleptop)	–	✓	✓
ΔR(b1Higgs, bhadtop)	–	✓	✓

Table 4.3.: List of the input variables for the reconstruction BDT in the three signal regions. In the descriptions, hadTop stand for the hadonic Top object, bhadTop correspond to the b-quark from the hadronic Top object, b1Higgs stand for the leading (in p_T) b-quark from the Higgs object and b2Higgs represent the sub-leading (in p_T) b-quark from the Higgs object. The leptonic Top and the hadronic W objects are described using a similar terminology.

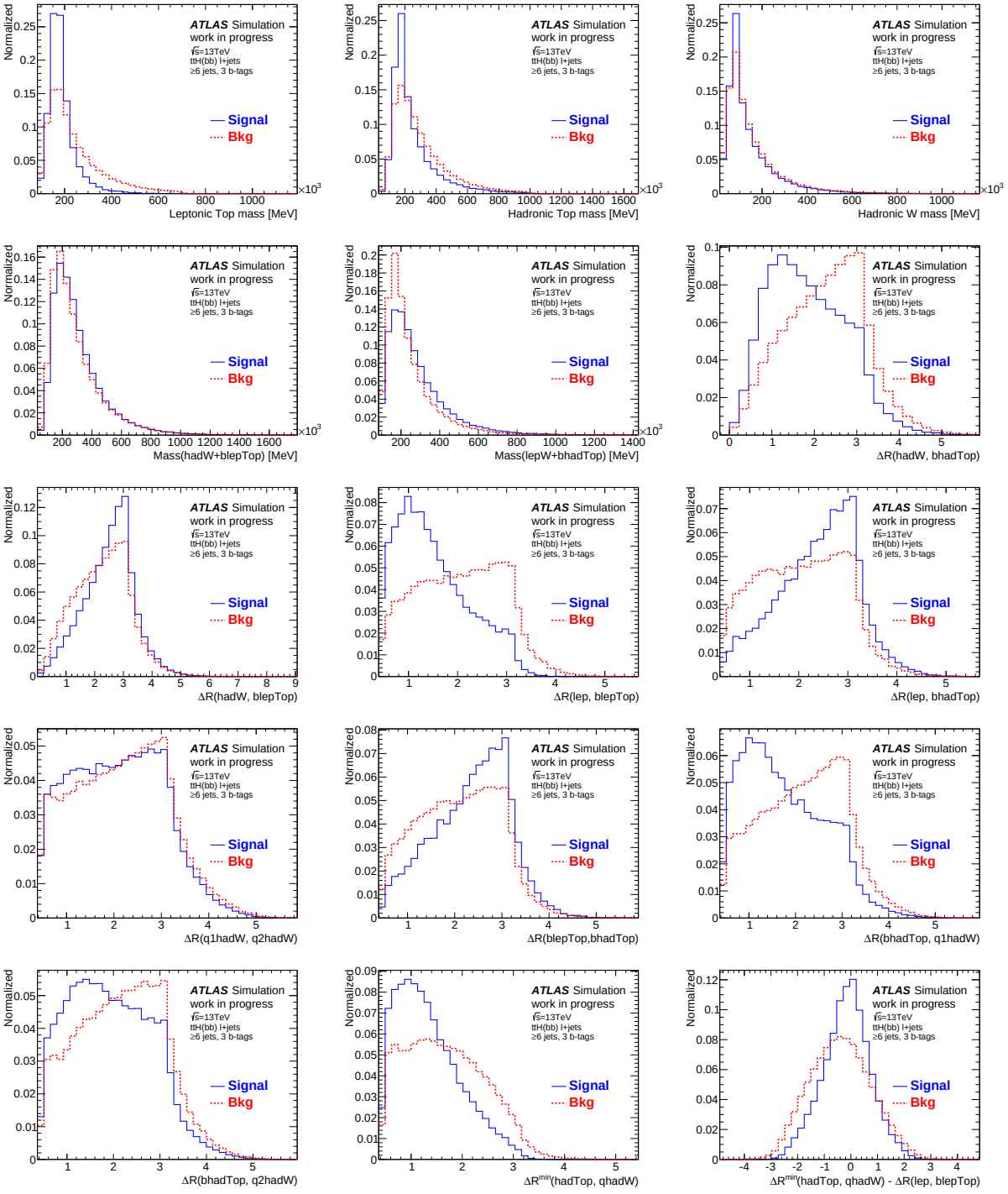


Figure 4.6.: Distributions of the kinematic variables used as inputs for the recoBDT in the ≥6 jets, ≥4 b-tags region. Solid blue lines correspond to the correct combination (signal) while the dashed red lines show the combinatorial background.

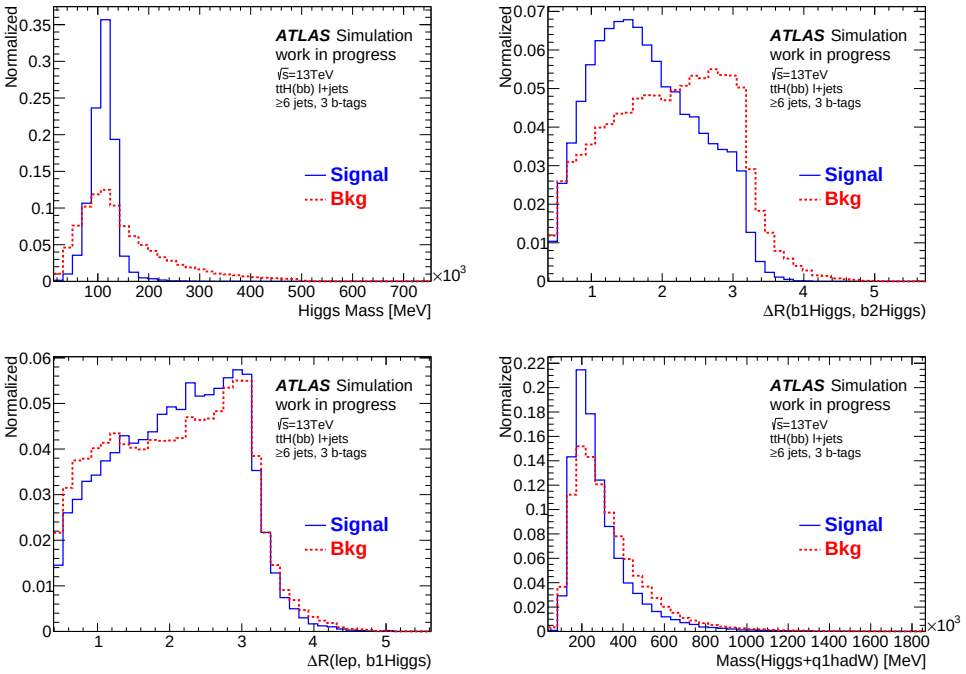


Figure 4.7.: Distributions of Higgs related variables used as inputs for the recoBDT_withHiggs in the ≥6 jets, ≥4 b-tags region. Solid blue lines correspond to the correct combination (signal) while the dashed red lines show the combinatorial background.

TMVA Setting	≥6 jets, ≥4 *b*-tags	≥6 jets, 3 *b*-tags	5 jets, ≥4 *b*-tags
BoostType	AdaBoost	AdaBoost	AdaBoost
AdaBoostBeta	0.15	0.15	0.15
NTrees	400	400	250
MaxDepth	5	5	4
nCuts	80	80	100
MinNodeSize	4%	4%	5%

Table 4.4.: Details of the reconstruction BDT settings in the three signal regions. The parameters are the same in the regions with 6 selected jets. Fewer trees results in an improved stability in the region with less statistics: 5 jets, ≥4 *b*-tags.

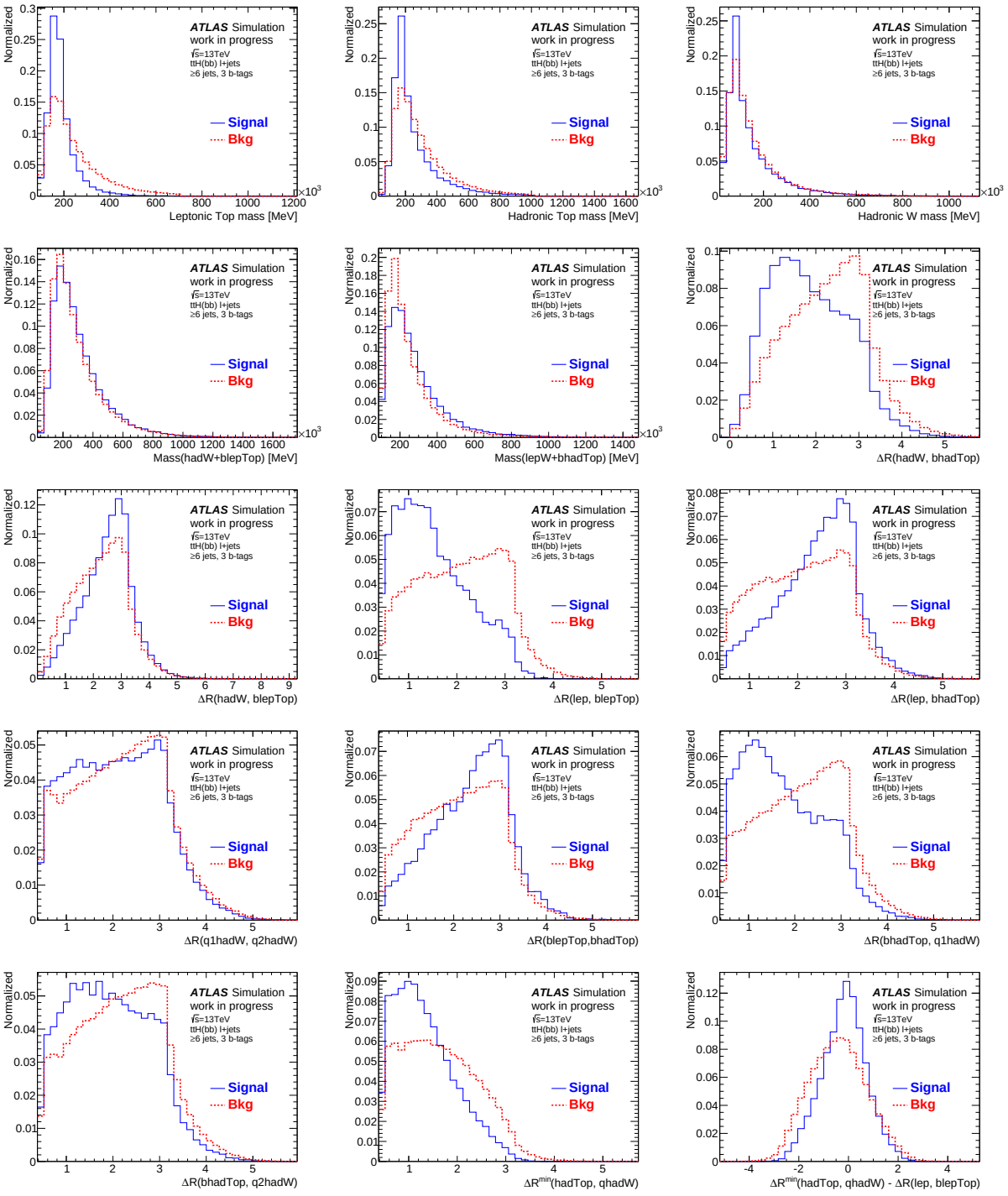


Figure 4.8.: Distributions of the kinematic variables used as inputs for the recoBDT in the ≥6 jets, 3 b-tags region. Solid blue lines correspond to the correct combination (signal) while the dashed red lines show the combinatorial background.

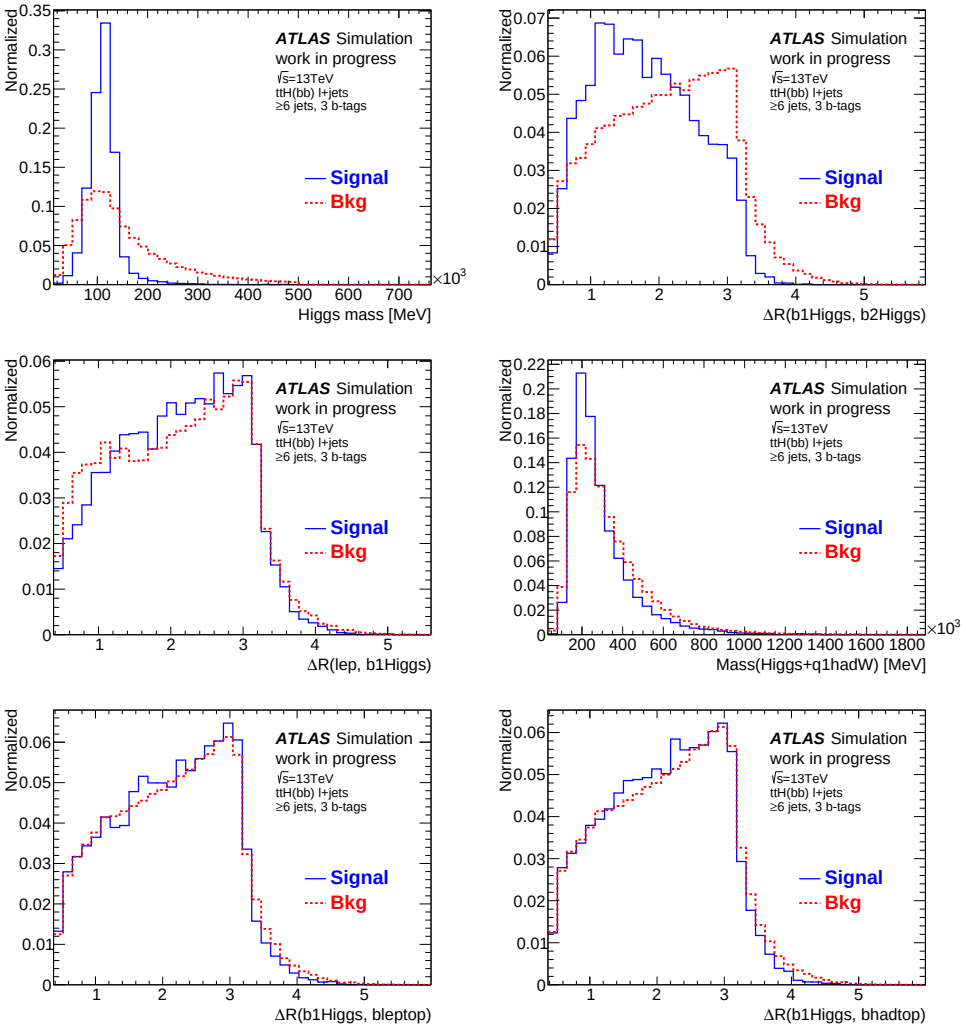


Figure 4.9.: Distributions of Higgs related variables used as inputs for the recoBDT_withHiggs in the≥6 jets, 3 b-tags region. Solid blue lines correspond to the correct combination (signal) while the dashed red lines show the combinatorial background.

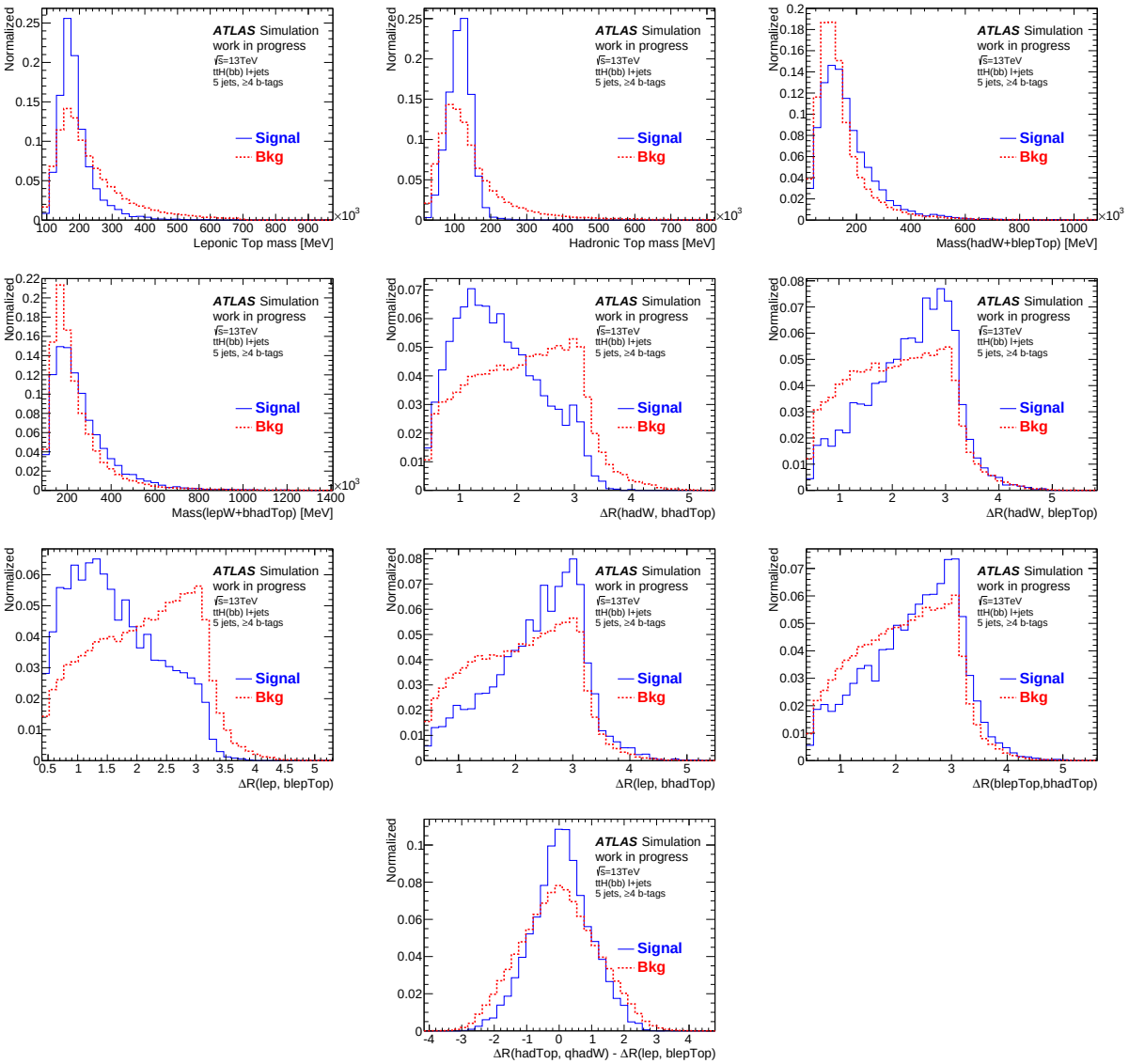


Figure 4.10.: Distributions of the kinematic variables used as inputs for the recoBDT in the 5 jets, ≥4 b-tags region. Solid blue lines correspond to the correct combination (signal) while the dashed red lines show the combinatorial background.

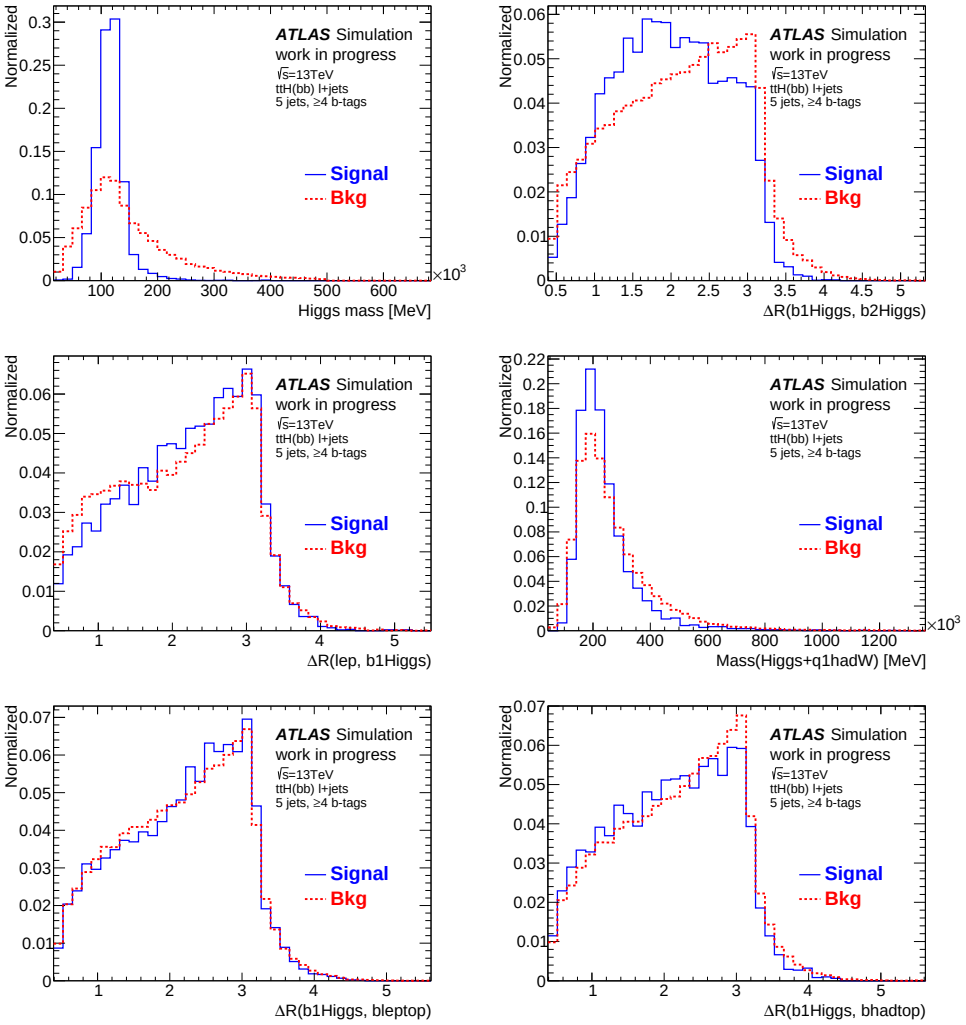


Figure 4.11.: Distributions of Higgs related variables used as inputs for the recoBDT_withHiggs in the 5 jets, ≥4 b-tags region. Solid blue lines correspond to the correct combination (signal) while the dashed red lines show the combinatorial background.

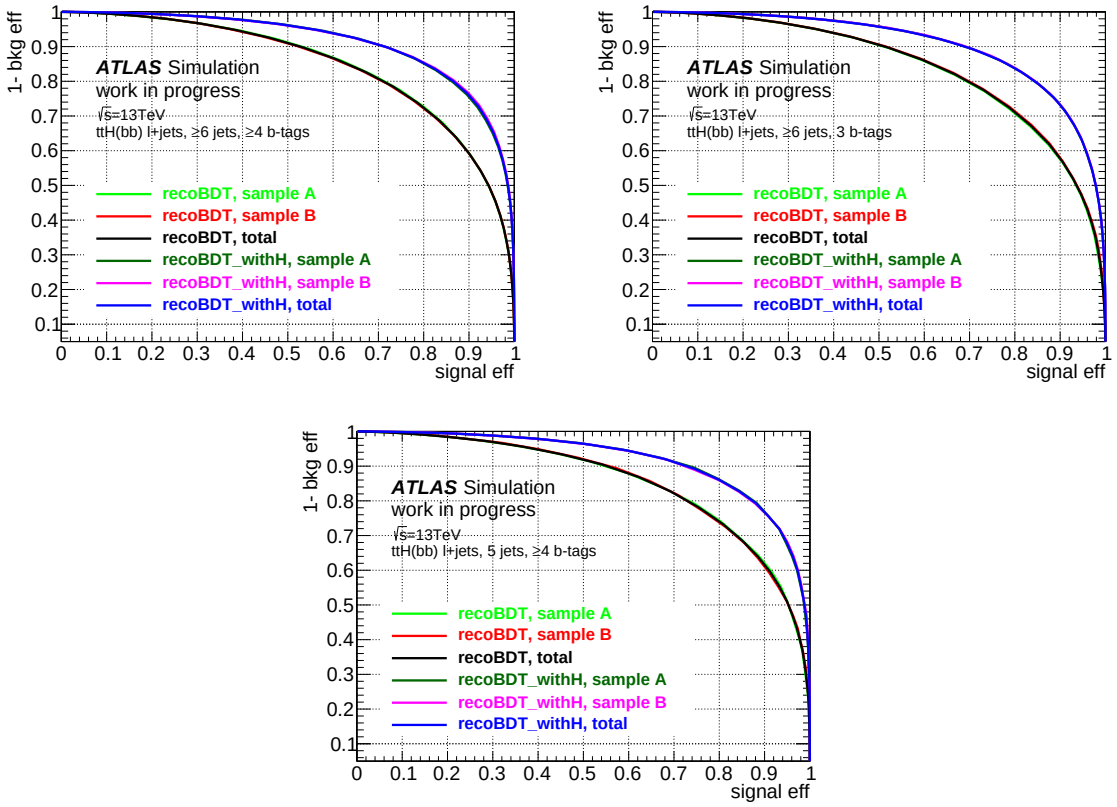


Figure 4.12.: ROC curves for the reconstruction BDT in the signal-rich regions. Results are shown for the evaluation in sample A, sample B and sample (A+B). The A/B curves are not clearly visible as they almost correspond to the same result as A+B.

4.5.1.4. Reconstructed Top and Higgs Masses

To perform the jet assignment in an event, all possible jet combinations are constructed, the trained BDT is evaluated for each jet combination, and the jet combination with the largest BDT output is selected. Since two different reconstruction BDTs are trained, there are two choices for the jet assignment under either the recoBDT or the recoBDT_withHiggs response. Figures 4.13 to 4.15 show the distribution of the invariant mass of the reconstructed Top and the Higgs boson for recoBDT and the recoBDT_withHiggs.

4.5.1.5. Performance

The performance of the MVA-based event reconstruction is quantified with the reconstruction efficiency, defined as the fraction of events for which the chosen combination is the correct one. It is calculated separately for all objects or subset of objects. For instance, the reconstruction efficiency of the Higgs boson is defined as the fraction of events for which the two corresponding jets have been identified correctly (4^{th} column in figure 4.16). A reconstruction Higgs matching efficiency of up to 48% is obtained in the ($\geq$6 jets, $\geq$4 b-tags) region, compared to the maximum achievable matching efficiency of about 89% (as show in figure 4.5), giving a relative Higgs matching efficiency of about 50%. An efficiency of up to 16% to correctly match all jets is achieved in the ($\geq$6 jets, $\geq$4 b-tags) region using Higgs-related variables in the training. The corresponding maximum achievable matching efficiency is about 42%, hence a relative all partons matching efficiency of about 38% is found. The reconstruction efficiencies for the different objects in the three signal regions are shown in figure 4.16.

Appendix A.1 shows the ratio of the reconstruction efficiency to the maximum achievable matching efficiency. The comparison of data and MC prediction for the distributions of the highest reconstruction BDT output per event can be found in appendix A.2

All reconstruction efficiencies obtained with recoBDT are smaller compared to those obtained with recoBDT_withHiggs. This is expected as the latter make use of more information of the Higgs decay topology. In particular, the recoBDT approach does not use information about the b-quarks from the Higgs resulting in a large misidentification rate for the jets assigned to the b-quarks from the Higgs. However, as shown on the bottom of the figures 4.13 to 4.15, recoBDT_withHiggs bias the reconstructed Higgs mass distribution and this effect is similar when it is applied to the $t\bar{t}b\bar{b}$ background, reducing thus its discriminant power.

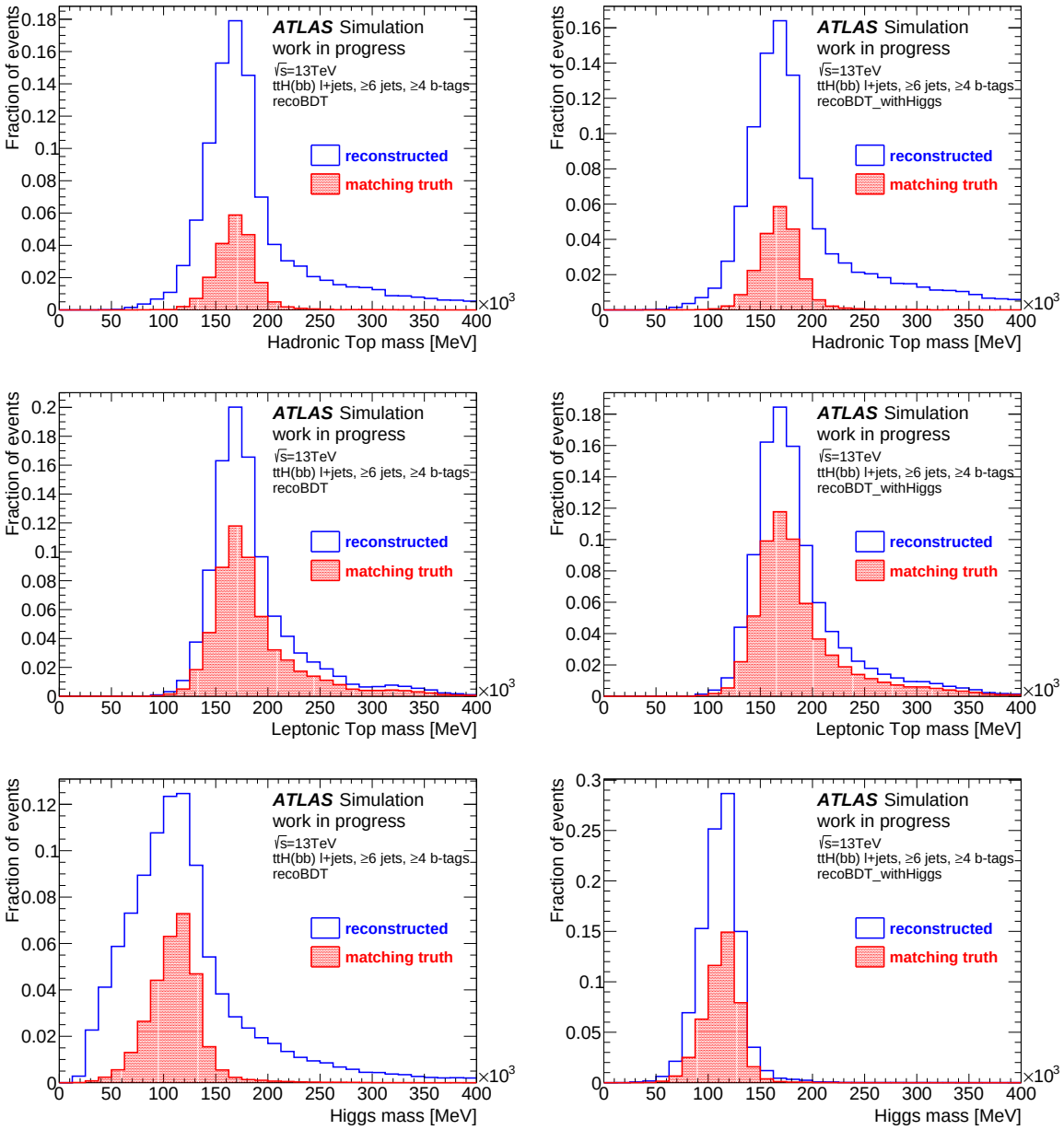


Figure 4.13.: Reconstructed Top and Higgs masses in the ≥6 jets, ≥4 b-tags region. The left (right) distributions show the masses using the jet assignment from recoBDT (recoBDT_withHiggs). Also overlaid are the distributions for the subset of events where the reconstructed objects match the corresponding quarks from Top or Higgs decay.

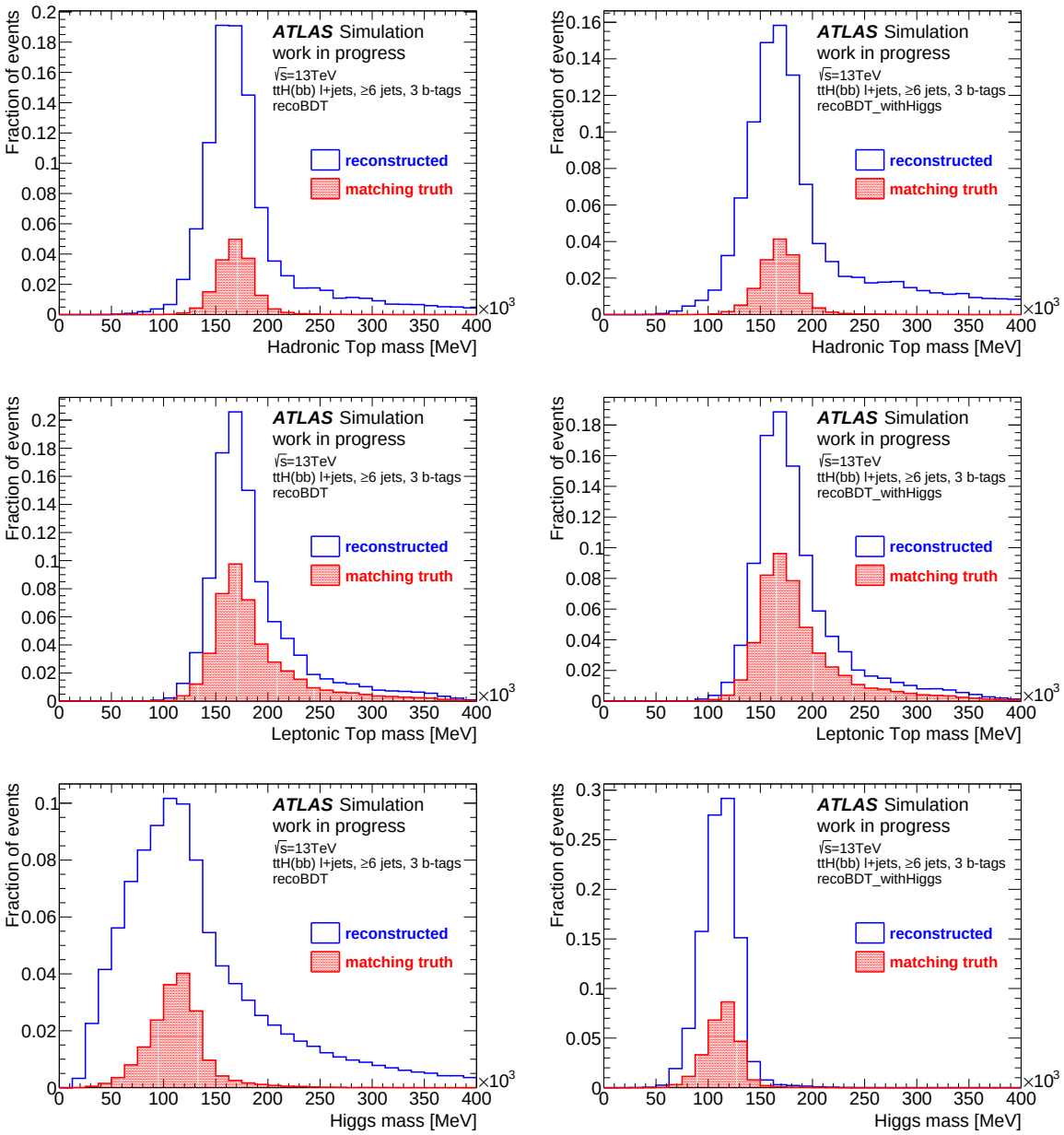


Figure 4.14.: Reconstructed Top and Higgs masses in the ≥6 jets, 3 b-tags region. The left (right) distributions show the masses using the jet assignment from recoBDT (recoBDT_withHiggs). Also overlaid are the distributions for the subset of events where the reconstructed objects match the corresponding quarks from Top or Higgs decay.

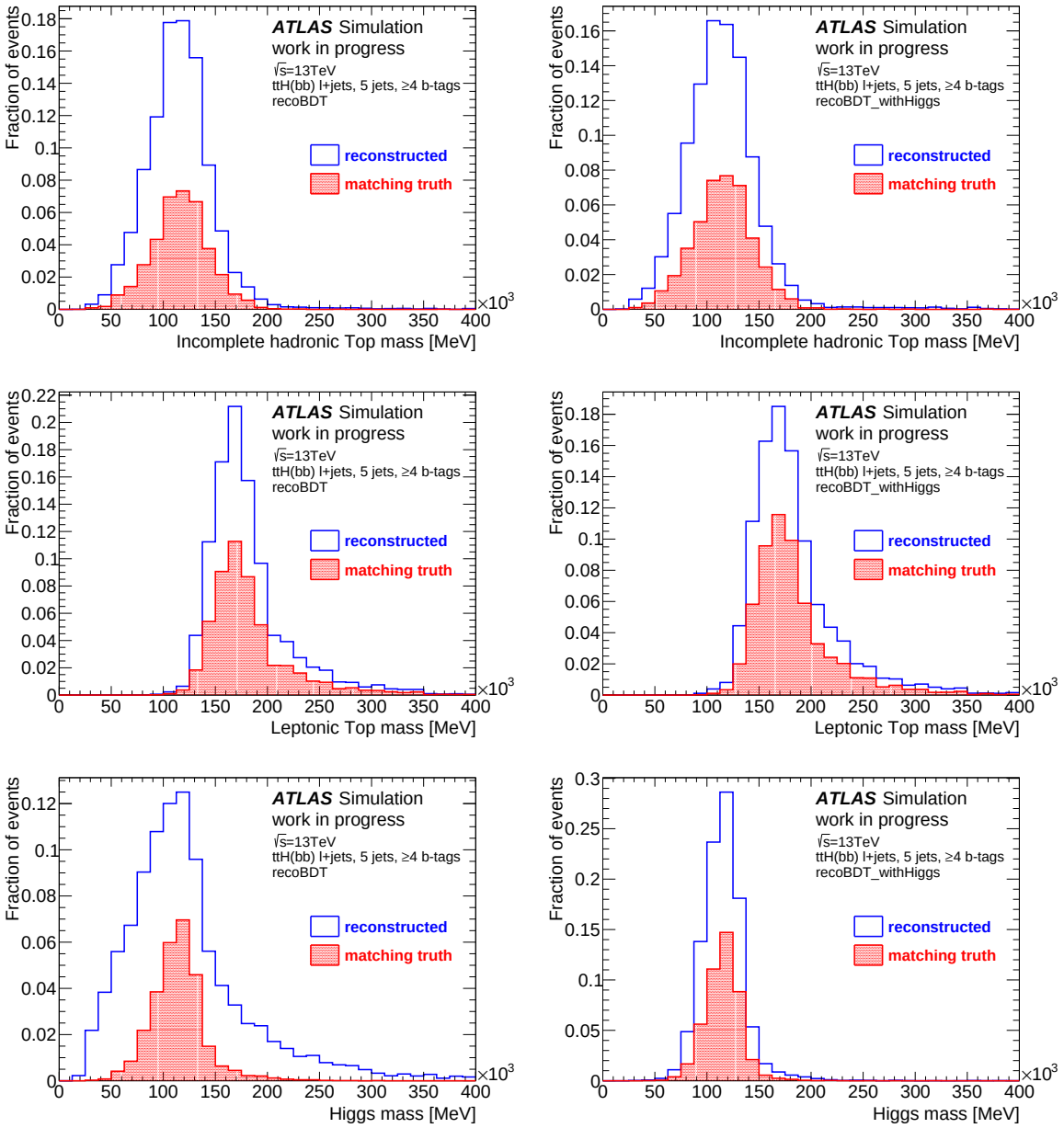


Figure 4.15.: Reconstructed Top and Higgs masses in the 5 jets, ≥4 b-tags region. The left (right) distributions show the masses using the jet assignment from recoBDT (recoBDT_withHiggs). Also overlaid are the distributions for the subset of events where the reconstructed objects match the corresponding quarks from Top or Higgs decay.

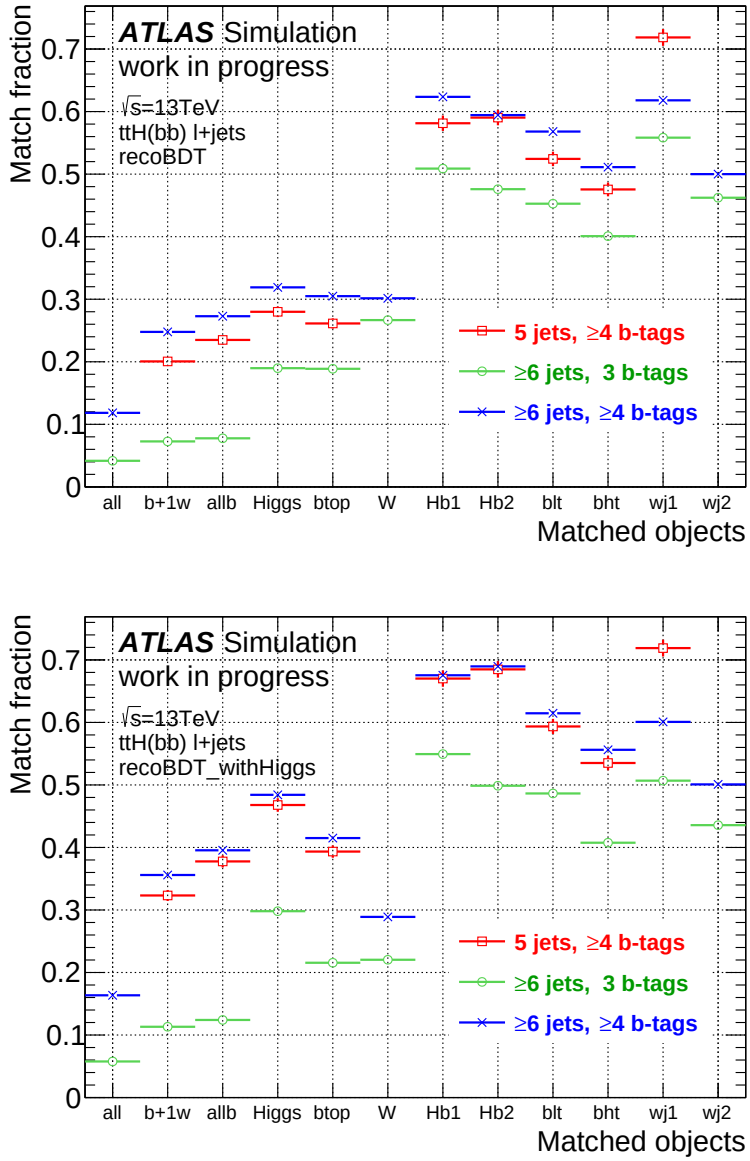


Figure 4.16.: The reconstruction efficiency for all jets correctly assigned (all), jets correctly assigned to the four b-quarks from $t\bar{t}H(H \to b\bar{b})$ system and one jet assigned to one quark from hadronic W (b+1W), jets correctly assigned to the four b-quarks from $t\bar{t}H(H \to b\bar{b})$ system (allb), two jets assigned to the b-quarks from Higgs (Higgs), two jets assigned to the b-quarks from $t\bar{t}$ (btop), two jets assigned to the hadronically decaying W boson (W), one jet assigned to the leading b-quark from Higgs (Hb1), one jet assigned to the sub-leading b-quark from Higgs (Hb2), one jet assigned to the b-quark from leptonic Top (blt), one jet assigned to the b-quark from hadronic Top (bht), one jet assigned to the leading quark from W (wj1) and one jet assigned to the leading quark from W (wj2). (top) Jet combination chosen from RecoBDT, (bottom) jet combination chosen from RecoBDT_withHiggs.

4.5.2. Discrimination between signal and background

The two reconstruction BDTs are both executed for each selected event in the signal region, leading to jet combinations from recoBDT and recoBDT_withHiggs. This allows to construct two sets of variables. In addition, a third group of variables that require no jet assignment are defined. Then, the three sets of variables are used as inputs to a classification BDT that provides the final discrimination between the $t\bar{t}H$ signal and the $t\bar{t}$+jets background as illustrated in the scheme in figure 4.17.

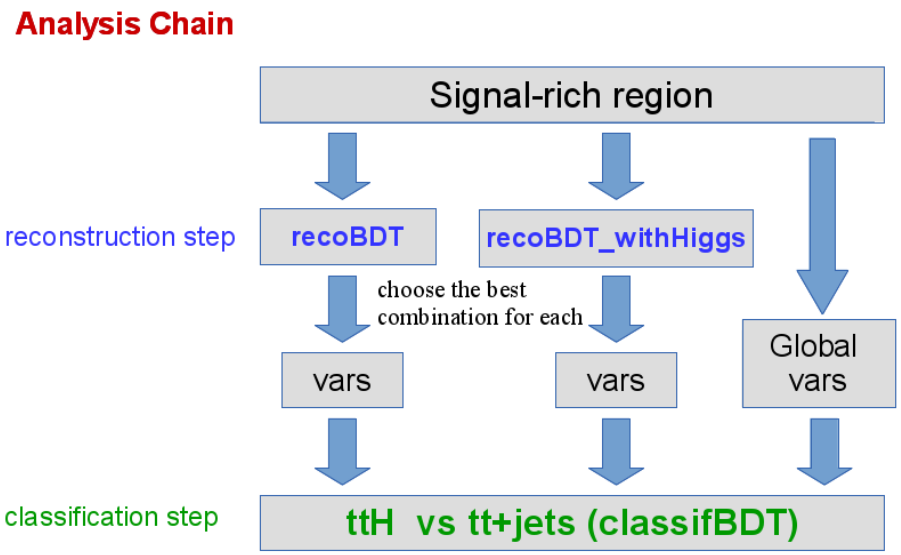


Figure 4.17.: For each signal-rich region, information from the reconstruction BDTs is combined with global kinematic variables in a classification BDT that is used to separate $t\bar{t}H$ from the dominant $t\bar{t}$+jets background.

4.5.2.1. Discriminating variables for the classification BDT

Following as baseline the $t\bar{t}H(H \to b\bar{b})$ Run 1 analysis, several classes of variables were used in the training of the classification BDTs:

- Object kinematic: the transverse momentum of the fifth leading jet ($p_{\mathrm{T}}^{\mathrm{jet5}}$).
- Event kinematic variables: the scalar sum of the transverse momentum of all jets ($H_{\mathrm{T}}^{\mathrm{had}}$) and the number of jets with $p_T \geq 40$ GeV (N_{40}^{jet}),
- Event shape variables: the scalar sum of the p_T divided by the sum of the energy for all jets and the lepton (Centrality), the second Fox-Wolfram moment [121] computing using all jets and the lepton ($H1$) and $1.5\lambda_2$, where λ_2 is the second eigenvalue of the momentum tensor [122] built with all jets (Aplan).
- Object pair properties: the average ΔR for all b-tagged jet pairs ($\Delta R_{\mathrm{bb}}^{\mathrm{avg}}$), the maximum $\Delta\eta$ between any two jets ($\Delta\eta_{\mathrm{jj}}^{\mathrm{max}\ \Delta\eta}$), the mass of the combination of the two

b-tagged jets with the smallest ΔR ($m_{\mathrm{bb}}^{\mathrm{min}\ \Delta R}$), the mass of the combination of a b-tagged jet and any jet with the largest vector sum of p_T ($m_{\mathrm{bj}}^{\mathrm{max}\ p_T}$), ΔR between the two b-tagged jets the largest vector sum of p_T ($\Delta R_{\mathrm{bb}}^{\mathrm{max}\ p_T}$), ΔR between the lepton and the combination of the two b-tagged jets with the smallest ΔR ($\Delta R_{\mathrm{lep-bb}}^{\mathrm{min}\ \Delta R}$), the number of b-tagged jet pairs with invariant mass within 30 GeV of the Higgs boson mass ($\mathrm{N}_{30}^{\mathrm{Higgs}}$), and the mass of the combination of any two jets with the smallest ΔR ($m_{\mathrm{jj}}^{\mathrm{min}\ \Delta R}$).

In addition to the global variables, variables using the information of the MVA-based event reconstruction are used:

- Reconstruction variables using the jets combination from `recoBDT`: the Higgs boson candidate mass (Higgs mass), ΔR between b jets from the Higgs boson candidate (ΔR(b1Higgs, b2Higgs)), the mass of the Higgs boson candidate and the b-jet from the leptonic Top candidate (mass(Higgs+blepTop)) and ΔR between the Higgs boson candidate and the leptonic Top candidate (ΔR(Higgs, lepTop)).

- Reconstruction variables using the jets combination from `recoBDT_withHiggs`: the highest BDT score in the event (highest BDT score), ΔR between the Higgs boson candidate and the $t\bar{t}$ system candidate(ΔR(Higgs, $t\bar{t}$) and ΔR between the Higgs boson candidate and the b-jet from the hadronic Top candidate (ΔR(Higgs, bhadtop)).

An interactive process is used to find an optimal set of variables in each signal-rich region. First the input variables are ranked by their signal-to-background separation power (TMVA separation) defined as:

$$\frac{1}{2}\sum_{i}^{bins}\frac{(N_i^S - N_i^B)^2}{N_i^S + N_i^B}, \tag{4.4}$$

where N_i^S and N_i^B are the entries in each bin of the normalised signal and background histograms, respectively. Then about 30 variables are selected for the starting point due to their discrimination power. One-by-one, variables with no significant improvement of discrimination between signal and background are removed. In the end, only the best 15 variables are selected in each signal region. No significant improvement of discrimination is achieved by selecting more variables. The complete list of variables used in the classification BDTs can be found in table 4.5.

Distributions of signal and background events for the input variables using reconstructed objects are shown in figures 4.18 to 4.20. For each variable the TMVA separation is show as well.

The highest BDT score in the event from `recoBDT_withHiggs` is one of the most important variable in all the signal-rich regions. Since this reconstruction BDT was trained with the full information of the $t\bar{t}H(H \to b\bar{b})$ system the highest BDT score in the event can be interpreted as the probability of the event to be signal. In the most sensitive region ($\geq$6 jets, $\geq$4 b-tags) the three most important variables are: the average ΔR for all b-tagged jet pairs, the highest BDT score from `recoBDT_withHiggs` and centrality.

The comparison between data and simulation for the input variables show a good agreement, as shown in appendix B.1.

Variable	Region		
	≥6 jets, ≥4 b-tags	≥6 jets, 3 b-tags	5 jets, ≥4 b-tags
Global variables:			
Centrality	✓	✓	✓
$\Delta\eta_{\mathrm{jj}}^{\text{max }\Delta\eta}$	✓	✓	✓
$H1$	✓	✓	✓
$p_{\mathrm{T}}^{\mathrm{jet5}}$	✓	✓	✓
$\Delta R_{\mathrm{bb}}^{\mathrm{avg}}$	✓	✓	✓
Aplan	✓	✓	✓
$\mathrm{N}_{30}^{\mathrm{Higgs}}$	✓	–	✓
$m_{\mathrm{bb}}^{\text{min }\Delta R}$	✓	✓	–
$m_{\mathrm{bj}}^{\text{max }p_T}$	–	✓	–
$\Delta R_{\mathrm{bb}}^{\text{max }p_T}$	✓	–	–
$\Delta R_{\mathrm{lep-bb}}^{\text{min }\Delta R}$	–	–	✓
N_{40}^{jet}	–	✓	–
$H_{\mathrm{T}}^{\mathrm{had}}$	–	✓	✓
$m_{\mathrm{jj}}^{\text{min }\Delta R}$	–	–	✓
Variables from `recoBDT`:			
Higgs mass	✓	✓	✓
ΔR(b1Higgs, b2Higgs)	✓	✓	✓
Mass(Higgs+blepTop)	✓	–	–
ΔR(Higgs, lepTop)	✓	–	–
Variables from `recoBDT_withHiggs`:			
Highest BDT score	✓	✓	✓
ΔR(Higgs, $t\bar{t}$)	✓	✓	✓
ΔR(Higgs, bhadtop)	–	✓	✓

Table 4.5.: List of the input variables for the classification BDT. Variables using the information of the MVA-based event reconstruction are defined with the candidate objects from the best jets combination.

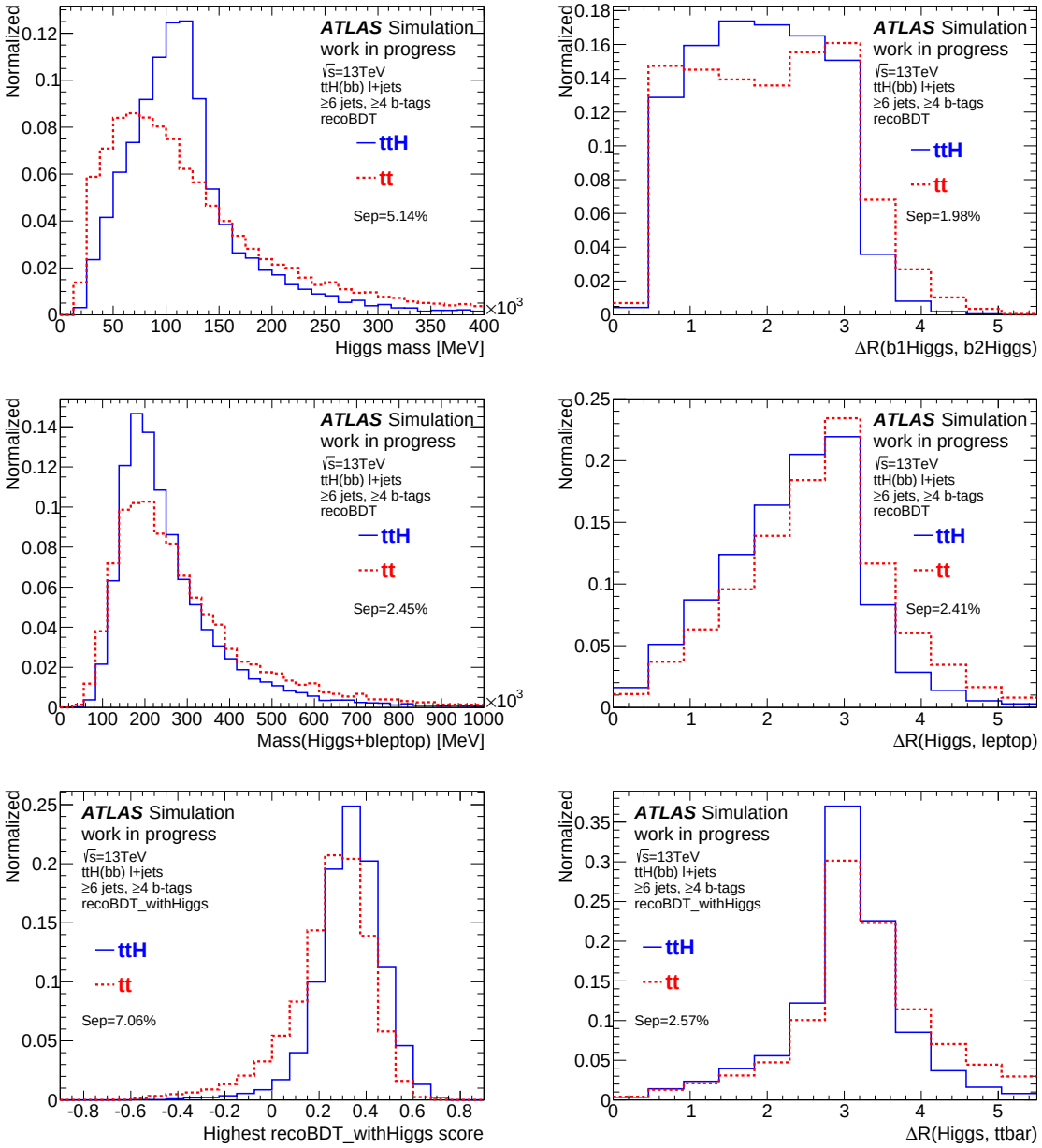


Figure 4.18.: Discriminating variables using reconstructed objects from recoBDT and recoBDT_withHiggs in the region with ≥6 jets, ≥4 b-tags. Each plot shows the normalised distribution for $t\bar{t}H$ signal (solid blue) and the $t\bar{t}$+jets background (dashed red). The TMVA separation (Sep), as defined in equation 4.4, is also shown.

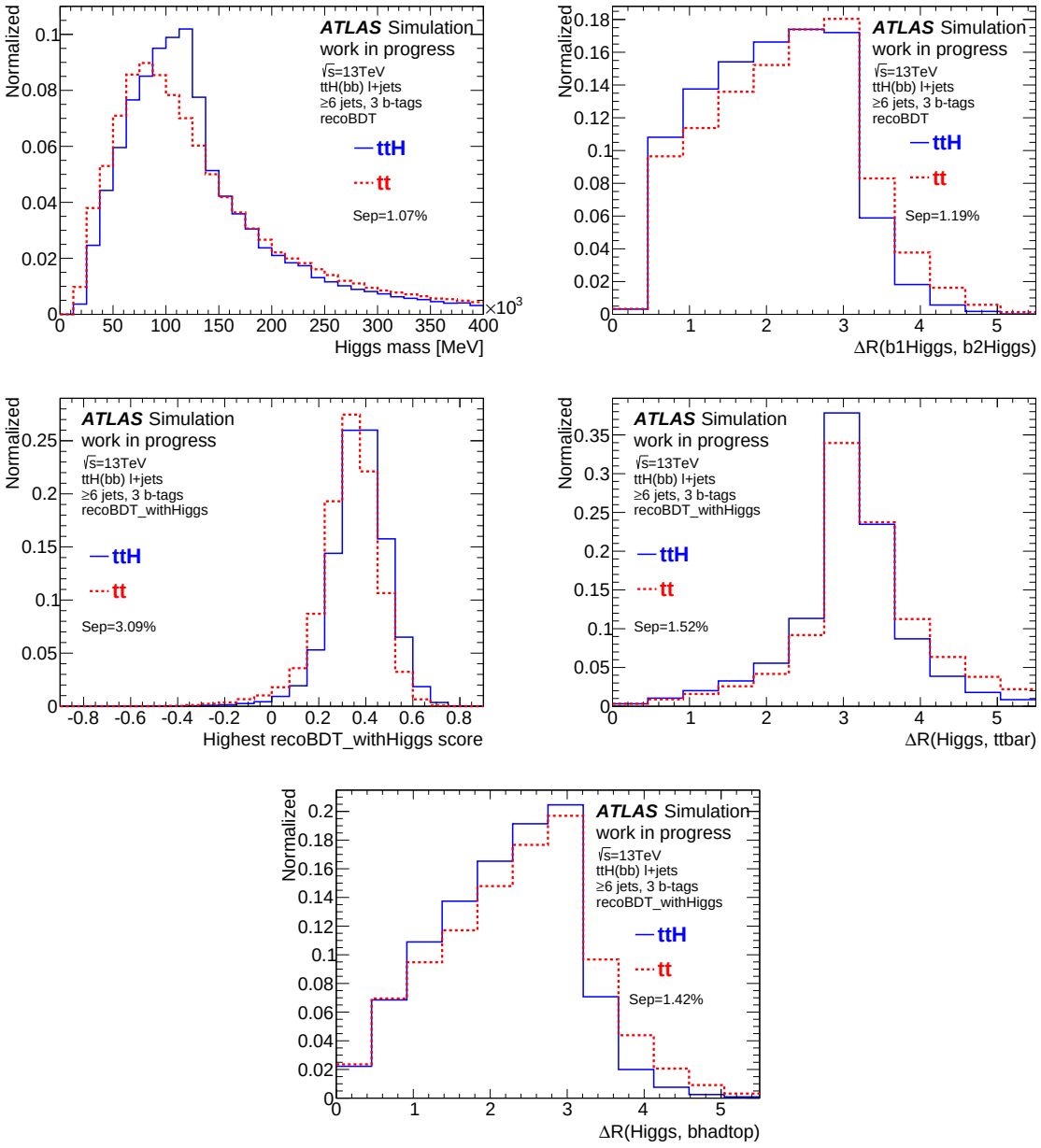


Figure 4.19.: Discriminating variables using reconstructed objects from recoBDT and recoBDT_withHiggs in the region with≥6j, 3b-tags. Each plot shows the normalised distribution for $t\bar{t}H$ signal (solid blue) and the $t\bar{t}$+jets background (dashed red). The TMVA separation (Sep), as defined in equation 4.4, is also shown.

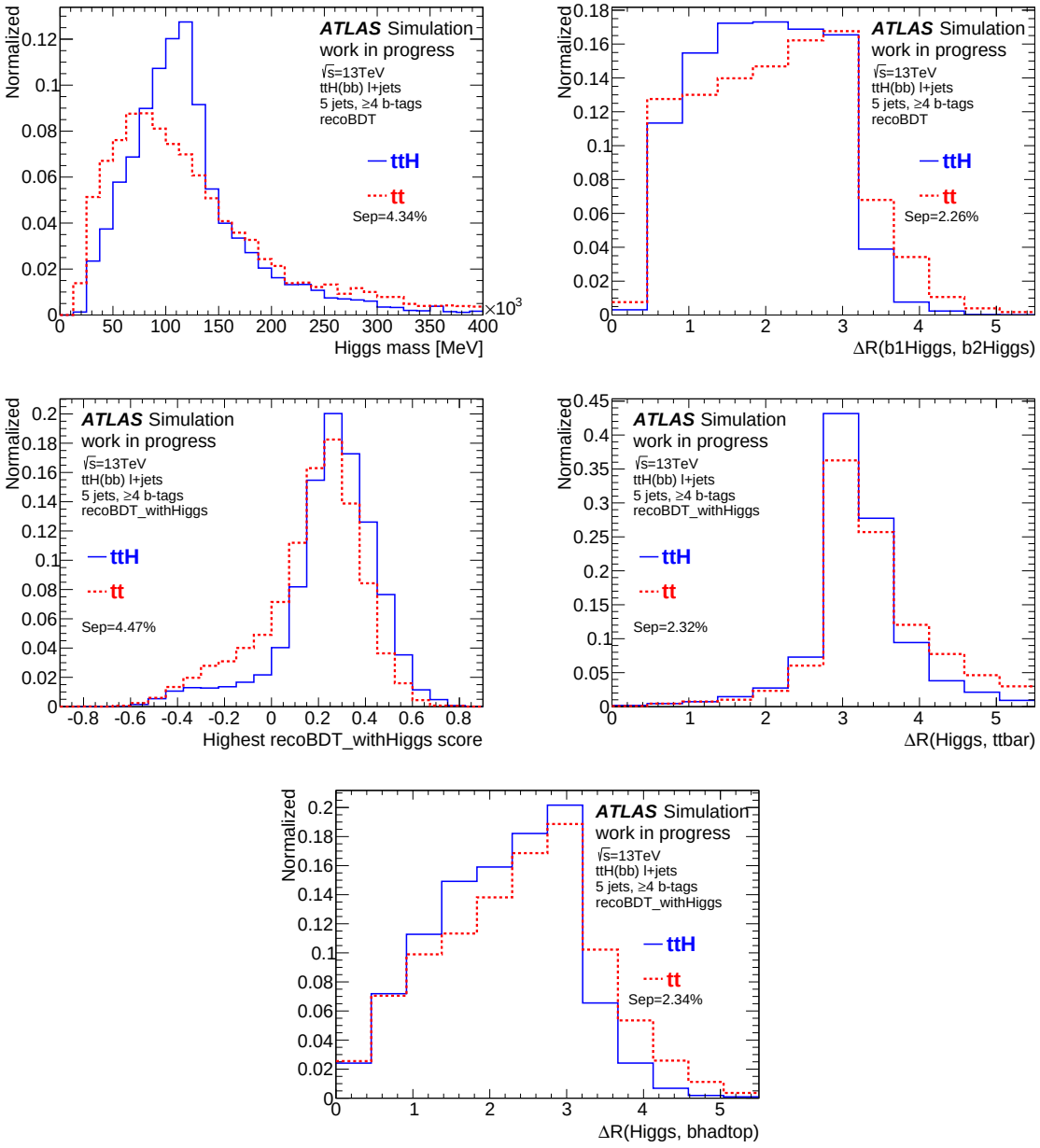


Figure 4.20.: Discriminating variables using reconstructed objects from recoBDT and recoBDT_withHiggs in the region with 5jets, ≥4b-tags. Each plot shows the normalised distribution for $t\bar{t}H$ signal (solid blue) and the $t\bar{t}$+jets background (dashed red). The TMVA separation (Sep), as defined in equation 4.4, is also shown.

4.5.2.2. BDT: setup and training

The classification BDTs are trained with a mixture of the single lepton, dilepton and full hadronic $t\bar{t}H$ samples as the signal. The samples include all Higgs decay modes. For the background $t\bar{t}$ + jets processes are used. Minor backgrounds are not considered for training because of their small impact.

An optimal set of the BDT parameters has been obtained after performing multiple trainings. Table 4.6 lists the chosen BDT parameters for each signal region.

TMVA setting	$\geq$6 jets, $\geq$4 b-tags	$\geq$6 jets, 3 b-tags	5 jets, $\geq$4 b-tags
BoostType	AdaBoost	AdaBoost	AdaBoost
AdaBoostBeta	0.15	0.15	0.15
NTrees	400	400	250
MaxDepth	5	5	4
nCuts	80	80	80
MinNodeSize	4%	4%	5%

Table 4.6.: Details of the classification BDT settings in the three signal regions. The parameters are the same in the regions with 6 selected jets. Fewer trees get better stability in the region with less statistics, 5 selected jets.

Cross training is used to profit from the full available statistics in the evaluation step: evaluate events in sample B with BDT trained on sample A and the opposite. Figure 4.21 shows the cross training validation plots. No large differences have been seen between the two sets.

4.5.2.3. Performance

The performance of the classification BDTs is measured using the TMVA separation defined in eq. 4.4. Table 4.7 shows the separation values for the classification BDTs with and without variables using the information from the MVA-based event reconstruction.

In all regions, the classification BDTs with reconstruction variables have better separation than the classification BDTs with purely global kinematic variables. In particular, the largest improvement with help of the reconstruction is found in the most sensitive region ($\geq$6 jets, $\geq$4 b-tags), an increase of 16.7% in separation is observed.

Figure 4.22 shows the distribution of the classification BDT with reconstruction output for the $t\bar{t}H$ signal and the $t\bar{t}$ + jets background in the signal regions. Also the ROC curves for the classification BDTs with and without reconstruction variables is showed. It can be seen that the ROC curves of the classification BDTs with reconstruction overshoot the ROC curves of the classification BDTs without the inclusion of the additional variables from the reconstruction BDTs.

Figures 4.23 shows the comparison of data and MC prediction for the distributions of the classification BDTs with reconstruction in each of the analysis regions considered. No significant shape disagreement is visible. The discrepancy observed is due to an

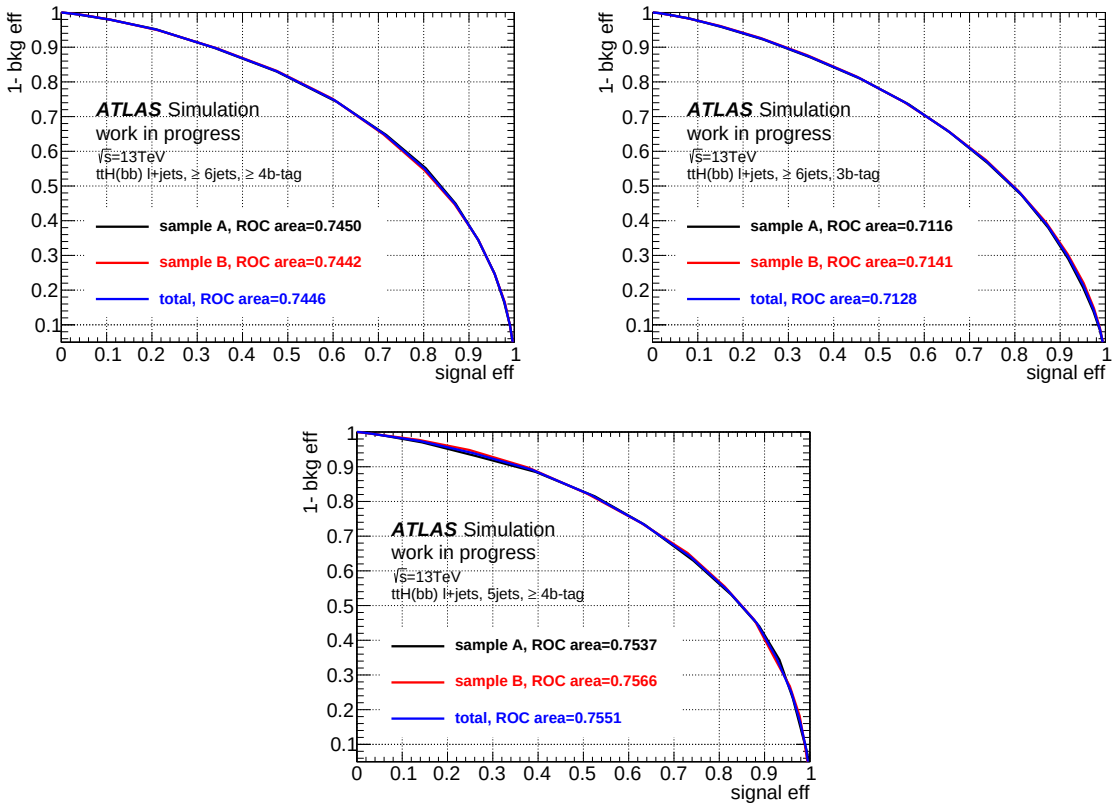


Figure 4.21.: ROC curves for the classification BDTs in the signal-rich regions. Results are shown for the evaluation in sample A, sample B and total (A+B).

TMVA separation (%)	classificationBDT without Reco	classificationBDT with Reco	Gain
≥6 jets, ≥4 b-tags	15.68	18.30	16.7%
5 jets, ≥4 b-tags	18.10	19.88	9.8%
≥6 jets, 3 b-tags	12.99	13.94	7.3%

Table 4.7.: TMVA separation for the classification BDT with and without variables using reconstructed objects in the training.

underestimation of the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ prediction. It will discussed more in section 4.7.2.1. The normalisation uncertainty for these processes are left free floating in the fit to data.

Since classification BDTs with reconstruction show the best separation it is used as the final discriminants in the fit to data. In the following, classification BDT will refer to classification BDT with reconstruction variables, unless otherwise specified.

4.6. Background modelling

This section describes the modelling of the main background components in this analysis.

4.6.1. $t\bar{t}$ + jets background

The $t\bar{t}$+jets sample is generated inclusively, but events are divided into sub-samples based on the flavour of the particle jets that do not originate from the decay of the $t\bar{t}$ system. Particle jets are reconstructed from stable truth particles using the anti-k_t algorithm with parameter R=0.4, in the acceptance region: $p_T > 15$ GeV ,$|\eta| < 2.5$. Events are labelled as $t\bar{t}+\geq 1b$ if at least one particle jet is matched within ΔR < 0.4 to a b-hadron with $p_T > 5$ GeV not originating from the decay of a top quark. Similarly, if at least one particle jet is matched to a c-hadron with $p_T > 5$ GeV, not originating from W boson the event is labelled as $t\bar{t}+\geq 1c$ if it is not already $t\bar{t}+\geq 1b$. Events labelled as either $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$ are referred to as $t\bar{t}$+HF ("heavy flavour") events. The remaining events are labelled as $t\bar{t}$+light events, including those with no additional jets.

The $t\bar{t}$+HF events are further categorised using a finer classification in order to compare different event generators and for the application of systematic uncertainties related to the modelling of $t\bar{t}$+HF. If there are two particle jets matched to an extra b- or c-hadron, the event is referred to as $t\bar{t}+b\bar{b}$ or $t\bar{t}+c\bar{c}$, if there is a single particle jet matched to a single b-hadron or c-hadron, the event is referred referred to as $t\bar{t}+b$ or $t\bar{t}+c$ if there is a single particle jet matched to a b-hadron or c-hadron pair, the event is referred to as $t\bar{t}+B$ or $t\bar{t}+C$ and if there are at least 3 particle jets matched to at least one b-hadron or c-hadron each, the events is referred to as $t\bar{t}+\geq 3b$ or $t\bar{t}+\geq 3c$.

4.6.1.1. $t\bar{t}$+light and $t\bar{t}+\geq 1c$ modelling

From Run 1 $t\bar{t}$ measurements [123], it is known that MC prediction for most generators, particularly Powheg + Pythia6, overpredicts the data at high top quark p_T and $t\bar{t}$ system p_T. Therefore, in order to correct for this effect, $t\bar{t}$+light and $t\bar{t}+\geq 1c$ events are reweighted to match the NNLO calculation for the differential cross-section at 13 TeV [124, 125]. A two-step sequential reweighting is applied in a similar way as the Run 1 $t\bar{t}H(H \rightarrow b\bar{b})$ analysis. First, a reweighting to the NNLO top quark p_T is applied, then the $t\bar{t}$ system p_T is reweighted.

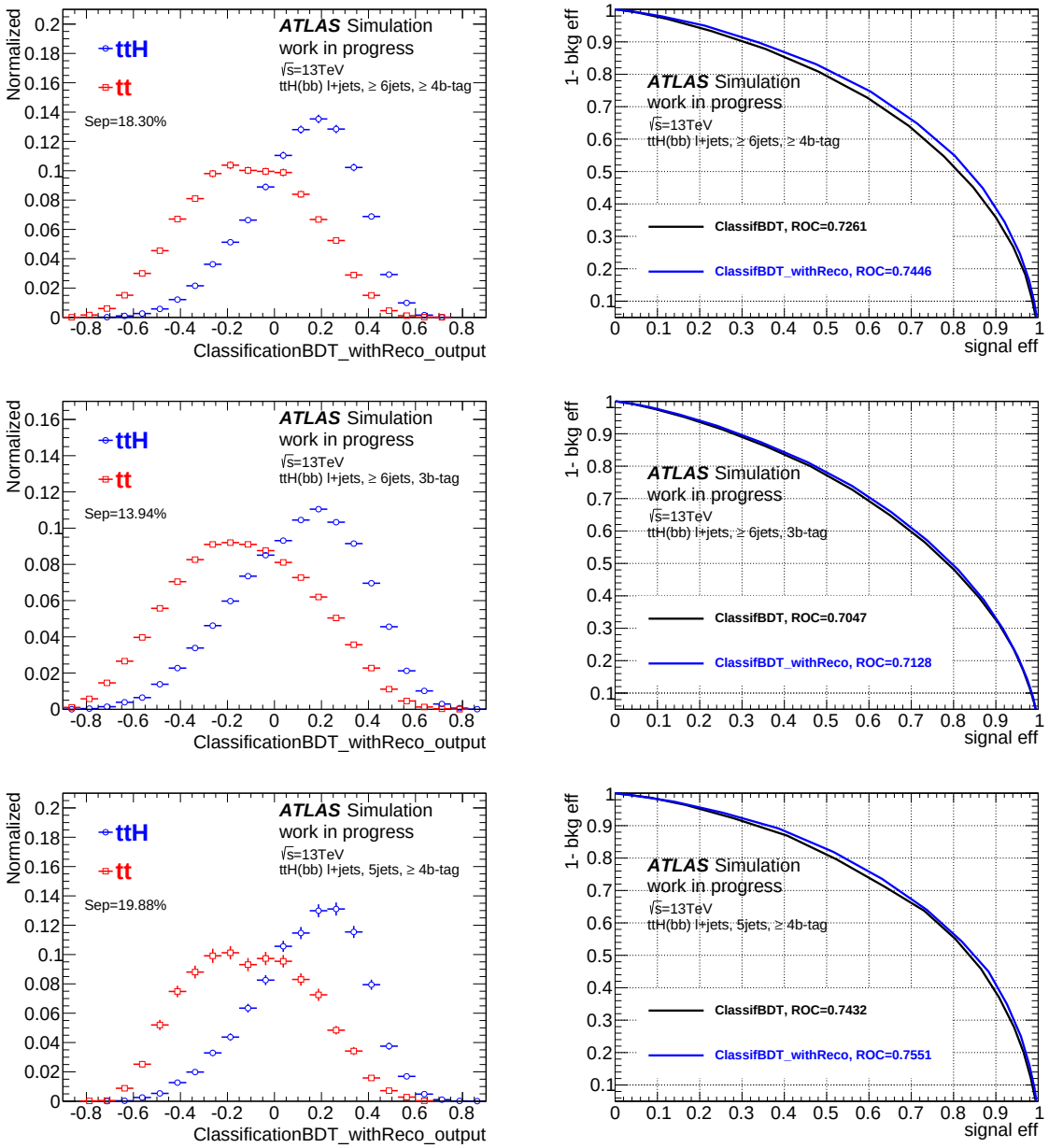


Figure 4.22.: On the left, distributions of the classification BDT with reconstruction output for the $t\bar{t}H$ signal and the $t\bar{t}$ + jets background in the signal regions. On the right, the ROC curves for the classification BDTs with and without reconstruction variables.

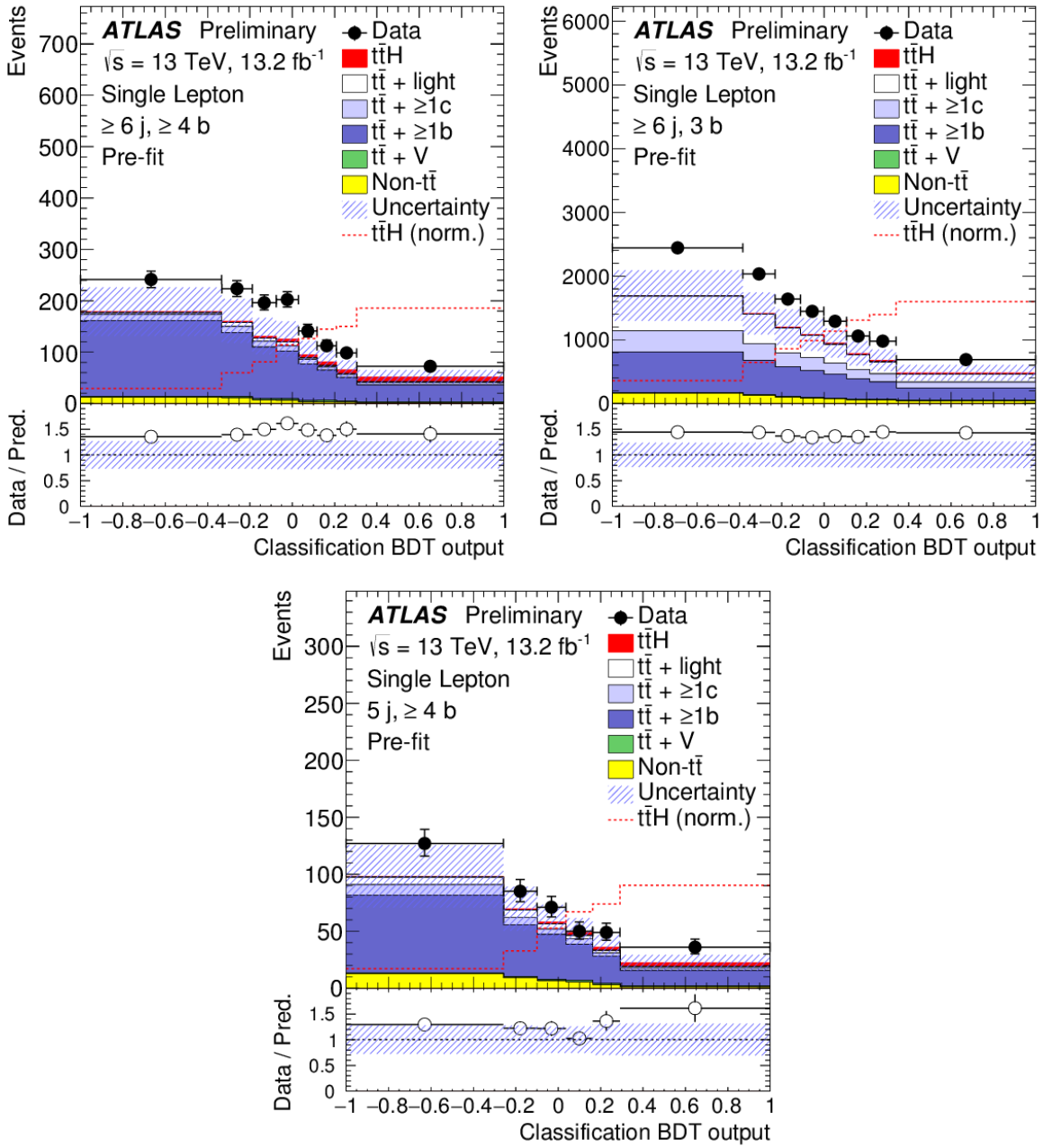


Figure 4.23.: Distributions of the output of the classification BDTs with reconstruction in data and MC for the three signal regions. The uncertainty band contains both statistical uncertainty and systematic uncertainties. Distributions are shown before the fit procedure, uncertainties on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$ [117] are not included.

4.6.1.2. $t\bar{t}+\geq 1b$ modelling

The $t\bar{t}+\geq 1b$ background is reweighted to the NLO prediction based on a $t\bar{t}+b\bar{b}$ sample generated with Sherpa+OpenLoops [126, 127]. This reweighting is performed for different topologies of $t\bar{t}+\geq 1b$ in such a way that the relative normalisation of each of the sub-categories ($t\bar{t}+b\bar{b}$, $t\bar{t}+b$, $t\bar{t}+B$, $t\bar{t}+\geq 3b$) and the relevant kinematic distributions are at NLO accuracy. In each sub-category, a first reweighting is based on the top quark p_T and $t\bar{t}$ system p_T. This is followed in the $t\bar{t}+b$ and $t\bar{t}+B$ sub-categories by a reweighting on the p_T and η of the b-jet; in the $t\bar{t}+b\bar{b}$ and $t\bar{t}+\geq 3b$ sub-categories the reweighting is based on the ΔR and p_T of the di-b-jet system not coming from the top quark decay.

Figure 4.24 shows the cross section of the different $t\bar{t}+\geq 1b$ event categories for the Powheg+Pythia6, Sherpa+OpenLoops samples and MG5_aMC samples.

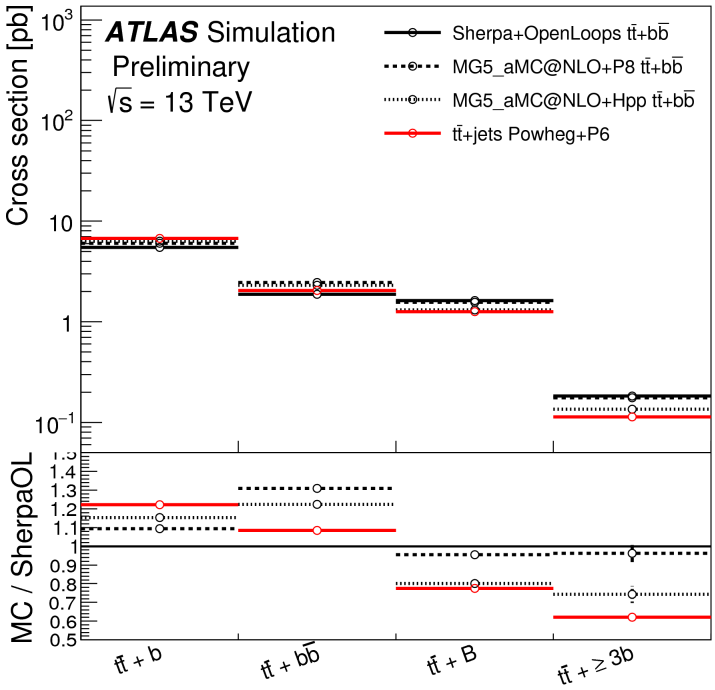


Figure 4.24.: Cross-sections for the different categories of $t\bar{t}+\geq 1b$ events. The inclusive Powheg+Pythia6 sample is compared to Sherpa+OpenLoops and MG5_aMC with different parton showers [117].

4.6.2. Misidentified lepton background

Misidentified leptons produced together with additional jets contribute to the background in the analysis. Although these events have small acceptance rates, production

rates are significantly larger than the processes of interest, resulting in a non-negligible background. The misidentified lepton background contributes via the misidentification of a jet or a photon as an electron ("fake" electron) or the presence of a non-prompt electron (e.g. electrons from heavy-hadron decays) in the electron channel, while in the muon channel the contribution is predominantly due to a non-prompt muon, such as those from b- or c-hadron decays. A data-driven method known as the "matrix method" is used to estimate the expected number of misidentified lepton background in the selected sample [128].

Events are divided into two samples: events with one tight lepton and events with one loose lepton. The former being a subset of the latter. The tight selection applies the same requirements as used in the analysis, as defined in section 4.3. For the loose selection, electrons are those satisfying the medium likelihood-based selection with no requirements on the isolation and muons are those satisfying the tight selection requirements with no requirements on the isolation. The number of events in each sample can then be expressed as a linear combination of the number of events with a real or fake leptons:

$$N^{loose} = N^{loose}_{real} + N^{loose}_{fake}, \tag{4.5}$$

$$N^{tight} = \epsilon_{real} N^{loose}_{real} + \epsilon_{fake} N^{loose}_{fake}, \tag{4.6}$$

where $\epsilon_{real}(\epsilon_{fake})$ is the fraction of real leptons (fake leptons) in the loose selection that also passes the tight selection.

The relative efficiencies ϵ_{real} and ϵ_{fake} depend on the lepton kinematics and the characteristics of the events such as number of jets and b-jets. An event weight is computed from the efficiencies, which are parameterized as a function of several kinematics (e.g. the lepton η and p_T, the ΔR between the lepton and its nearest jet, $\Delta\phi(l, E_T^{miss})$) [129]:

$$w_i = \frac{\epsilon_{fake}}{\epsilon_{real} - \epsilon_{fake}}(\epsilon_{real} - \delta_i), \tag{4.7}$$

where δ_i is equal to 1 if the loose event i passes the tight event selection and 0 otherwise.

The real efficiency ϵ_{real} is derived using the tag-and-probe method from the $Z \to ee$ and $Z \to \mu\mu$ data. The fake efficiency ϵ_{fake} is measured in data samples dominated by non-prompt and fake leptons. Thus, a control region is defined by requiring: $E_T^{miss} + m_T(\text{lepton}, E_T^{miss})$[b] < 60 GeV and $m_T(\text{lepton}, E_T^{miss}) < 20$ GeV. Then, ϵ_{fake} is determined as the ratio between the number of tight and loose events in this region.

4.6.3. Other backgrounds

The W/Z + jets, $t\bar{t}V$, single top (s-channel, t-channel and Wt-channel) and diboson backgrounds are estimated from MC simulations as outlined in section 4.2.2.3.

[b] The transverse mass m_T is defined as $m_T = \sqrt{2p_{\mathrm{T}}^{l} E_{\mathrm{T}}^{\mathrm{miss}}(1 - \cos(\phi^{l} - \phi^{\mathrm{miss}}))}$

4.7. Systematic uncertainties

The sources of systematics are from experimental and modelling uncertainties. Systematic uncertainties can affect the normalisation of the signal and background and/or the shape of the final discriminant distribution. The uncertainties are taken into account via nuisance parameters in the fit procedure, which is described in section 4.8. The sources of systematic uncertainties considered in the analysis are summarised in table 4.8.

4.7.1. Experimental uncertainties

These uncertainties arise from the measurement used to correct the modelling of the physics objects.

4.7.1.1. Luminosity

The uncertainties on the combined 2015+2016 integrated luminosity is 2.9%. The method to derive this uncertainty is similar to the one detailed in ref [130]. This systematic uncertainty affect the overall normalisation of all contributions determined from MC simulations.

In order to correct differences in the pileup distributions between MC simulation and data a pile-up reweighing is applied. An uncertainty is considered on the reweighting of the pileup distribution.

4.7.1.2. Leptons

Uncertainties considered for leptons arises from trigger, reconstruction, identification, isolation, and lepton momentum scale and resolution. These uncertainties generally have a very small impact on the result.

4.7.1.3. Jets

Uncertainties associated to the jet selection arises from the jet energy scale (JES), jet vertex tagger (JVT) and jet energy resolution (JER). The JES uncertainties considered originate from several sources: in-situ calibration techniques (statistical, detector, modelling), pileup dependent corrections and the flavour composition of the jets. In total there are 18 uncertainties associated to the JES. A JER uncertainty has been assessed using the Run 1 uncertainty with an extrapolation from Run 1 to Run 2 conditions [131].

4.7.1.4. Missing transverse momentum

The E_T^{miss} reconstruction is affected by uncertainties associated with leptons and jet energy scales and resolutions which are propagated to E_T^{miss}. Additional uncertainties quantify the resolution and scale of the soft terms. Since E_T^{miss} is not used in selection but only in event reconstruction, its uncertainties typically have a small effect on the analysis

Systematic uncertainty	Type	Components
Luminosity	N	1
Pileup reweighting	SN	1
Reconstructed Objects		
Electron trigger+reco+ID+isolation	SN	4
Electron energy scale+resolution	SN	2
Muon trigger+reco+ID+isolation	SN	6
Muon momentum scale+resolution	SN	3
Jet vertex Tagger	SN	1
Jet energy scale	SN	18
Jet energy resolution	SN	1
Missing transverse momentum	SN	3
b-tagging efficiency	SN	5
c-tagging efficiency	SN	4
Light-jet tagging efficiency	SN	14
High-p_{T} tagging	SN	2
Background Model		
$t\bar{t}$ cross section	N	1
$t\bar{t}$+$b\bar{b}$: NLO Shape	SN	10
$t\bar{t}$+$c\bar{c}$: NLO Shape	SN	1
$t\bar{t}+\geq 1b$ modelling: (residual) Radiation	SN	1
$t\bar{t}+\geq 1b$ modelling: (residual) NLO generator	SN	1
$t\bar{t}+\geq 1b$ modelling: (residual) parton shower+hadronisation	SN	1
$t\bar{t}$+light, $t\bar{t}+\geq 1c$ modelling: Radiation	SN	2
$t\bar{t}$+light, $t\bar{t}+\geq 1c$ modelling: NLO generator	SN	2
$t\bar{t}$+light, $t\bar{t}+\geq 1c$ modelling: parton shower+hadronisation	SN	2
$t\bar{t}$+light, $t\bar{t}+\geq 1c$ NNLO reweighting	SN	4
W+jets normalisation	N	6
Z+jets normalisation	N	1
Single top cross section	N	2
Wt modelling	SN	3
Diboson normalisation	N	1
$t\bar{t}V$ cross section	N	4
Fakes normalisation	N	6
Signal Model		
$t\bar{t}H$ cross section	N	2
$t\bar{t}H$ branching ratios	N	3
$t\bar{t}H$ model	SN	2

Table 4.8.: The list of systematic uncertainties considered in the analysis. An "N" means that the uncertainty is taken as normalisation-only for all processes and channels affected, whereas "SN" means that the uncertainty is taken on both shape and normalisation. Some of the systematic uncertainties are split into several components for a more accurate treatment.

4.7.1.5. Jet flavour tagging

The MV2c10 algorithm is used to distinguish b-jet against the c- and light-jets. The b-tagging related uncertainties are a mixture of statistical, experimental and modelling uncertainties. The b-jet tagging efficiency corrections are derived using $t\bar{t}$ events [132]. The uncertainties depends on p_T and the operating point (and η for light jets [133]) of the b-tagging algorithm. The uncertainties are decomposed to uncorrelated components; five significant uncertainties for b-jets, 4 for c-jets and 14 for light-jets. An additional uncertainty is included for the extrapolation to jets outside the kinematic range of the measurements.

4.7.2. Uncertainties on the background modelling

4.7.2.1. $t\bar{t}$ + jets production

Since $t\bar{t}$+jets represent the largest background in the analysis, a large number of uncertainties have been considered. These include the uncertainty on the theoretical prediction for the inclusive cross section, uncertainties affecting the modelling of $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ production, uncertainties associated with the choice of matrix element generator, the modelling of extra radiation, and the choice of parton shower and hadronisation model. The MC generators used for all $t\bar{t}$ inclusive samples and its setting are listed in table 4.9.

An uncertainty of $\pm 6\%$ is assumed for the inclusive $t\bar{t}$ production cross-section. It includes uncertainties from the PDF, α_s choices and top quark mass. An uncertainty associated with the choice of NLO generator is derived by comparing two alternative predictions, Powheg-Box and MG5_aMC, each of which is showered with Herwig++. Propagating the difference to the default Powheg+Pythia6 prediction (nominal $t\bar{t}$ sample). An uncertainty due to the choice of parton shower and hadronisation model is derived by comparing events produced by Powheg-Box interfaced with Pythia6 or Herwig++. In addition, uncertainties associated with the modelling of initial and final state radiation (ISR/FSR) [134] are obtained by comparing two alternative radiation variation samples of Powheg+Pythia6. All these uncertainties, except the inclusive cross-section, are treated as uncorrelated for the $t\bar{t}+\geq 1b$, $t\bar{t}+\geq 1c$ and $t\bar{t}$+light backgrounds.

All samples are reweighted to NNLO top quark p_T and $t\bar{t}$ p_T prediction before they are used to derive systematic uncertainties for $t\bar{t}$+light and $t\bar{t}+\geq 1c$. An uncertainty on the top quark p_T and $t\bar{t}$ system p_T is derived as the largest difference between the default NNLO prediction and the uncorrected prediction from any of the alternative samples.

Uncertainties associated with the modelling of $t\bar{t}+\geq 1b$ production include those associated with the NLO prediction from Sherpa+OpenLoops. Uncertainties on the NLO prediction are evaluated by varying the renormalisation, factorisation and resummation scales. Also two alternative PDF sets are considered: MTSW [135] and NNPDF [136]. Additional systematic uncertainties are associated to the $t\bar{t}+\geq 1b$ production: an uncertainty on the choice of the generator derived by comparing the prediction from Sherpa+OpenLoops and MG5_aMC+Pythia8 and an uncertainty from the parton

inclusive $t\bar{t}$	PDF	tune
Powheg-Box + Pythia 6.428	CT10	P2012
Powheg-Box + Herwig++2.7.1	CT10	UE-EE5
MG5_aMC + Herwig++2.7.1	CT10	UE-EE5
Powheg-Box + Pythia 6.428	CT10	P2012 radHi
Powheg-Box + Pythia 6.428	CT10	P2012 radLo

Table 4.9.: Summary of the inclusive $t\bar{t}$ samples used to derived systematic uncertainties.

shower and hadronisation model taken from the difference between MG5_aMC showered with Pythia8 or Herwig++. The MC generators used for the $t\bar{t}+b\bar{b}$ samples and its setting are listed in table 4.10. Separate uncertainties are applied to the $t\bar{t}+\geq 1b$ contribution from multi-parton-interactions (MPI) and FSR, which are not included in the Sherpa+OpenLoops prediction; a 50% uncertainty is applied to MPI, while an uncertainty on FSR is estimated from the alternative radiation variation samples. Alternative samples are reweighted to the $t\bar{t}+b\bar{b}$ NLO Sherpa+OpenLoops prior to the evaluation of the uncertainties for $t\bar{t}+\geq 1b$. The remaining differences for $t\bar{t}+\geq 1b$ events are referred to as "residual" uncertainties.

An uncertainty on the $t\bar{t}+\geq 1c$ modelling is obtained from the comparison with a dedicated NLO $t\bar{t}+c\bar{c}$ sample generated with MG5_aMC+Herwig++ [137]. The difference between this sample and an inclusive $t\bar{t}$ sample produced with the same generator is taken as the uncertainty.

The analysis is very sensitive to the $t\bar{t}$+HF modelling. The excess of data observed in the pre-fit plots is compatible with the large uncertainties associated to $t\bar{t}$+HF production. Thus, the normalisation of $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ are allowed to float freely in the fit. No prior normalisation uncertainty is applied.

$t\bar{t}+b\bar{b}$	PDF	tune
Sherpa + OpenLoops	CT10 4F	sherpa
MG5_aMC + Pythia 8.210	NNPDF3.0 4F	A14
MG5_aMC + Herwig++2.7.1	NNPDF3.0 4F	UE-EE5

Table 4.10.: Summary of the $t\bar{t}+b\bar{b}$ samples used to derived systematic uncertainties.

4.7.2.2. Misidentification lepton background modelling

Uncertainties on the data-driven muitijet background arise from the limited sample size in data, particularly at high jet and b-tag multiplicity, as well as from the uncertainty associated with the lepton misidentification rate measurements in different control regions. A combined normalisation uncertainty of 50% is assumed. This uncertainty is taken as uncorrelated across jet and b-tag multiplicity and also uncorrelated between electron and muon channels.

4.7.2.3. Other simulated backgrounds

A conservative normalisation of 30% had been adopted for W+jets events. An additional uncertainty of 30% is applied for events with W+HF jets. A normalisation uncertainty of 45% is applied for Z+jets events. These uncertainties are derived from variations of scales and matching parameters in Sherpa MC sample.

An uncertainty on the cross-sections of the single-top processes of +5%/-4% is used. It is a weighted average of the theoretical uncertainties on t-, Wt- and s-channel production [109–111]. Uncertainties associated with the Wt modelling of initial and final state radiation, parton shower and hadronisation are evaluated in the same manner as for $t\bar{t}$. Additional uncertainties on the interference between Wt and $t\bar{t}$ production at NLO [108] is derived from an alternative sample generated using diagram subtraction technique, as opposed to the nominal diagram removal scheme.

An uncertainties of 50% on the normalisation or diboson production is used, which includes uncertainties on the inclusive cross-section and additional jet production [138].

The theoretical uncertainty on the $t\bar{t}V$ NLO cross-section is 15% [139]. An additional uncertainty associated with the choice of the generator is derived by comparing to the alternative Madgraph+Pythia6 sample.

4.7.3. Uncertainties on the signal modelling

Systematic uncertainties on the modelling of the $t\bar{t}H$ process are estimated from varying the settings or the simulation of showering and hadronisation. A theoretical uncertainty of $^{+10\%}_{-13\%}$ is applied on the cross-section of the $t\bar{t}H$ signal [140–144]. The effect of the QCD scale and PDF set are considered uncorrelated. An additional uncertainty due to the QCD scale choice is estimated by varying the renormalisation and factorisation scales. The uncertainty associated to the showering and hadronisation model is derived by comparing the MG5_aMC samples interfaced to either Pythia8 or Herwig++. Finally, uncertainties on the Higgs boson branching ratios to $b\bar{b}$ ($^{+1.2\%}_{-1.3\%}$), WW ($^{+1.6\%}_{-1.5\%}$) and other final states ($^{+5\%}_{-5\%}$) are considered [145].

4.8. Statistical analysis

The statistical analysis is based on a binned maximum likelihood function $\mathcal{L}(\mu, \theta)$. The likelihood function depends on the signal strength parameter, $\mu = \sigma/\sigma_{\mathrm{SM}}$, and θ, the set of nuisance parameters (NP). The likelihood function is constructed as demonstrated in the following equation, as the product of Poisson-probability terms over the bins of the input distributions including the number of data events and expected signal and background yields, taking into account the effects of the systematic uncertainties:

$$\mathcal{L}(\mu, \theta) = \prod_{j} \prod_{i=\mathrm{bin}} \frac{(\mu s_i(j) + b_i(j))^{N_i(j)}}{N_i(j)!} e^{-\mu s_i(j) - b_i(j)} \prod_{\theta} \mathrm{func}(\theta|0, 1), \tag{4.8}$$

where $N_i(j)$ is the number of observed events in the i^{th} bin in the j^{th} signal region, and $s_i(j)$ and $b_i(j)$ are the expected number of signal and background events, respectively. Since the normalisation factors for $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ are allowed to float freely in the fit, $s_i(j)$ and $b_i(j)$ are also a function of the normalisation factors. The impact of systematic uncertainties on the signal and background expectations is described by the NP. The NP are characterised by a Gaussian or log-normal probability density functions: $\mathrm{func}(\theta|0,1)$, by convention, the value $\theta = 0$ corresponds to the nominal central value of the prediction, while values of ± 1 represent the ± 1 standard deviation of that particular systematic uncertainty. The best estimate for μ is obtained by maximising the likelihood.

The test statistic, q_μ, is defined as the profile likelihood ratio,

$$q_\mu = -2\ln(\mathcal{L}(\mu, \hat{\hat{\theta}}_\mu)/\mathcal{L}(\hat{\mu}, \hat{\theta})), \tag{4.9}$$

where $\hat{\mu}$ and $\hat{\theta}$ are the parameters that maximise the likelihood (with the constraint $0 \leq \hat{\mu} \leq \mu$), and $\hat{\hat{\theta}}_\mu$ are the values of the nuisance parameters that maximise the likelihood function for a given value of μ.

This test statistic is used to measure the compatibility of the observed data with the background-only hypothesis (i.e. for $\mu = 0$), and to make statistical inferences about μ, such as upper limits using the $\mathrm{CL_s}$ method [146, 147].

Figure 4.25 shows the ingredients of the CLs method. For a given value of the test statistic on data (q_{obs}), the distributions of the test statistic for the two hypothesis are shown; the right distribution under the background only hypotheses ($f(q_\mu|b)$) and on the left the signal plus background hypothesis ($f(q_\mu|s+b)$).

Then, the compatibility between the observed data and a given hypothesis is measured by a p-value:

$$p_\mu = P(q_\mu \geq q_\mu^{\mathrm{obs}} \mid \mathrm{signal} + \mathrm{background}) = \int_{q_\mu^{\mathrm{obs}}}^{\infty} f(q_\mu|s+b)dq_\mu \tag{4.10}$$

$$1 - p_b = P(q_\mu \geq q_\mu^{\mathrm{obs}} \mid \mathrm{background} - \mathrm{only}) = \int_{q_\mu^{\mathrm{obs}}}^{\infty} f(q_\mu|b)dq_\mu \tag{4.11}$$

Using these two variables, the CL_s variable is defined as :

$$CL_s(\mu) = \frac{p_\mu}{1 - p_b}. \tag{4.12}$$

To quote the 95% Confidence Level upper limit on μ (denoted as $\mu^{95\%CL}$), the value of μ is adjusted to reach CL_s=0.05. Thus, a value of μ is considered to be excluded at 95% confidence level if $\mu > \mu^{95\%CL}$.

4.9. Results

The signal strength modifier $\mu = \sigma/\sigma_{SM}$ for the $t\bar{t}H$ production cross section is determined in a simultaneous binned maximum-likelihood fit to data in all the regions. In

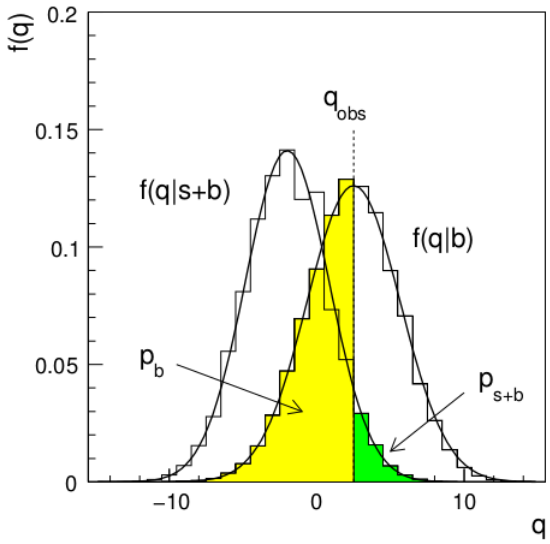


Figure 4.25.: Example of the distribution of the test statistics for background-only ($f(q_\mu|b)$) and signal+background hypothesis ($f(q_\mu|s+b)$), q_{obs} is the value of the test statistic on data [147].

order to improve the sensitivity of the analysis, classification BDTs are used in the template fit for the signal regions. In the other regions, the scalar sum of the jet p_T (H_T^{had}) is used. Table 4.11 summarises the regions and the corresponding variable used in the fit.

Region	2 b-tags	3 b-tags	4 b-tags
4 jets	$H_{\mathrm{T}}^{\mathrm{had}}$	$H_{\mathrm{T}}^{\mathrm{had}}$	$H_{\mathrm{T}}^{\mathrm{had}}$
5 jets	$H_{\mathrm{T}}^{\mathrm{had}}$	$H_{\mathrm{T}}^{\mathrm{had}}$	BDT
$\geq$ 6 jets	$H_{\mathrm{T}}^{\mathrm{had}}$	BDT	BDT

Table 4.11.: Summary of regions and final discriminants used in the fit to data.

The best-fit value of $\mu = \sigma/\sigma_{SM}$ is:

$$\mu = 1.6^{+0.5}_{-0.5}(\text{stat.})^{+1.0}_{-0.9}(\text{syst.}) = 1.6^{+1.1}_{-1.1} \tag{4.13}$$

No significant excess of events above the background expectation is found for the SM Higgs boson with mass of 125 GeV. The expected and observed 95% CL upper limits for the SM Higgs boson production cross-section are shown in table 4.12. A signal 2.2 times larger than the SM Higgs boson is expected to be excluded in the case of no SM Higgs boson. A signal strength larger than 3.6 can be excluded at 95% CL.

The figure 4.26 shows the yields before and after the fit in all regions. The post-fit yields for each process are shown in tables 4.13 to 4.15. The $t\bar{t}+\geq 1b$ contribution

Channel	Expected ($\mu = 1$)	Expected ($\mu = 0$) -2σ	-1σ	median	$+1\sigma$	$+2\sigma$	Observed
$t\bar{t}H$ single lepton	2.9	1.2	1.6	2.2	3.2	4.7	3.6

Table 4.12.: The expected and observed 95% CL upper limits normalised to the SM Higgs boson production for $t\bar{t}H(H \to b\bar{b})$ single lepton search at $m_H = 125$ GeV using 13.2 fb^{-1} at 13 TeV data.

is scaled to $1.24^{+0.23}_{-0.21}$ times the pre-fit value and the $t\bar{t}+\geq 1c$ contribution is scaled by $1.37^{+0.70}_{-0.60}$, consistent with ATLAS Run 1 results [85].

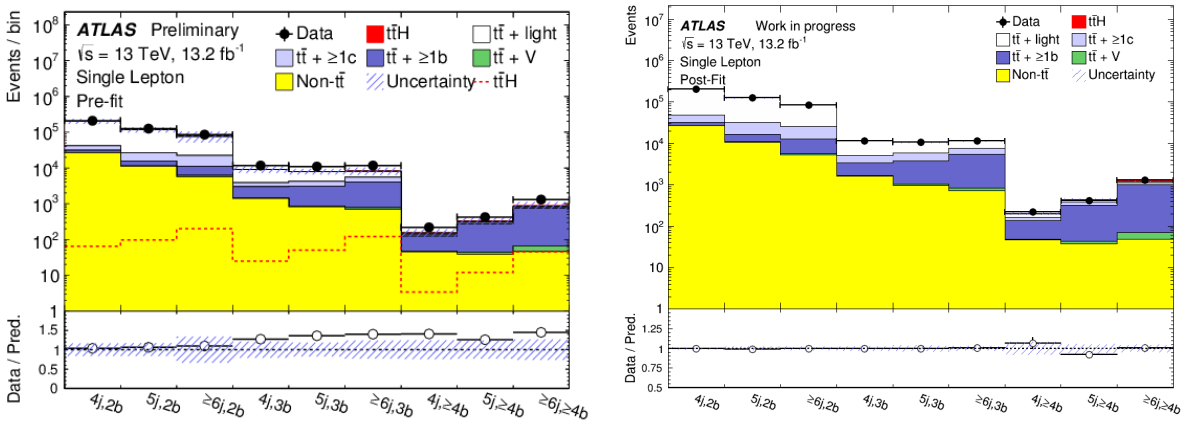


Figure 4.26.: Pre-fit [117] (left) and post-fit (right) yields for each of the nine analysis regions. The pre-fit yields do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

The distribution of H_T^{had} before and after the fit to data and the classification BDTs after the fit are shown in figures 4.27 to 4.29. A good agreement between data and MC simulation is observed. The distributions of all input variables for the classification BDTs pre-fit and post-fit can be found in appendix B.1.

	4 jets, 2 b-tags	4 jets, 3 b-tags	4 jets, 4 b-tags
$t\bar{t}$+light	160000±7030	6580±735	35.6±17.0
$t\bar{t}+\geq 1c$	16000±6560	1620±516	27.3±9.2
$t\bar{t}+\geq 1b$	5230±1240	1740±424	88.7±15.5
$t\bar{t}$+V	216±24	20.4±3.6	1.73±0.35
Single top	10250±1236	469±79	13.1±3.5
W/Z+jets	7900±2346	417±161	2.9±1.3
Diboson	472±230	21.9±12.5	4.6±3.4
fakes	7670±1560	739±239	25.2±25.8
$t\bar{t}H$	98±49	39±25	5.5±3.5
Total	208000±2030	11700±364	205±29
Data	208239	11686	218

Table 4.13.: Yields after the fit in the exclusive four jet region.

	5 jets, 2 b-tags	5 jets, 3 b-tags	5 jets, $\geq$4 b-tags
$t\bar{t}$+light	92000±6560	4710±720	47.1±22.2
$t\bar{t}+\geq 1c$	16300±5890	2170±577	69.4±16.1
$t\bar{t}+\geq 1b$	5470±1060	2680±499	272±36.5
$t\bar{t}$+V	184±30	40.1±6.0	5.46±1.46
Single top	4730±718	340±69	16.1±4.3
W/Z+jets	2872±876	335.4±146.1	3.5±2.8
Diboson	247±124	20.3±11.1	0.48±0.30
fakes	2750±693	286±113	17.3±16.2
$t\bar{t}H$	151±71	80±50	20±12
Total	125000±1770	10700±386	451±31
Data	124688	10755	418

Table 4.14.: Yields after the fit in the exclusive five jet region.

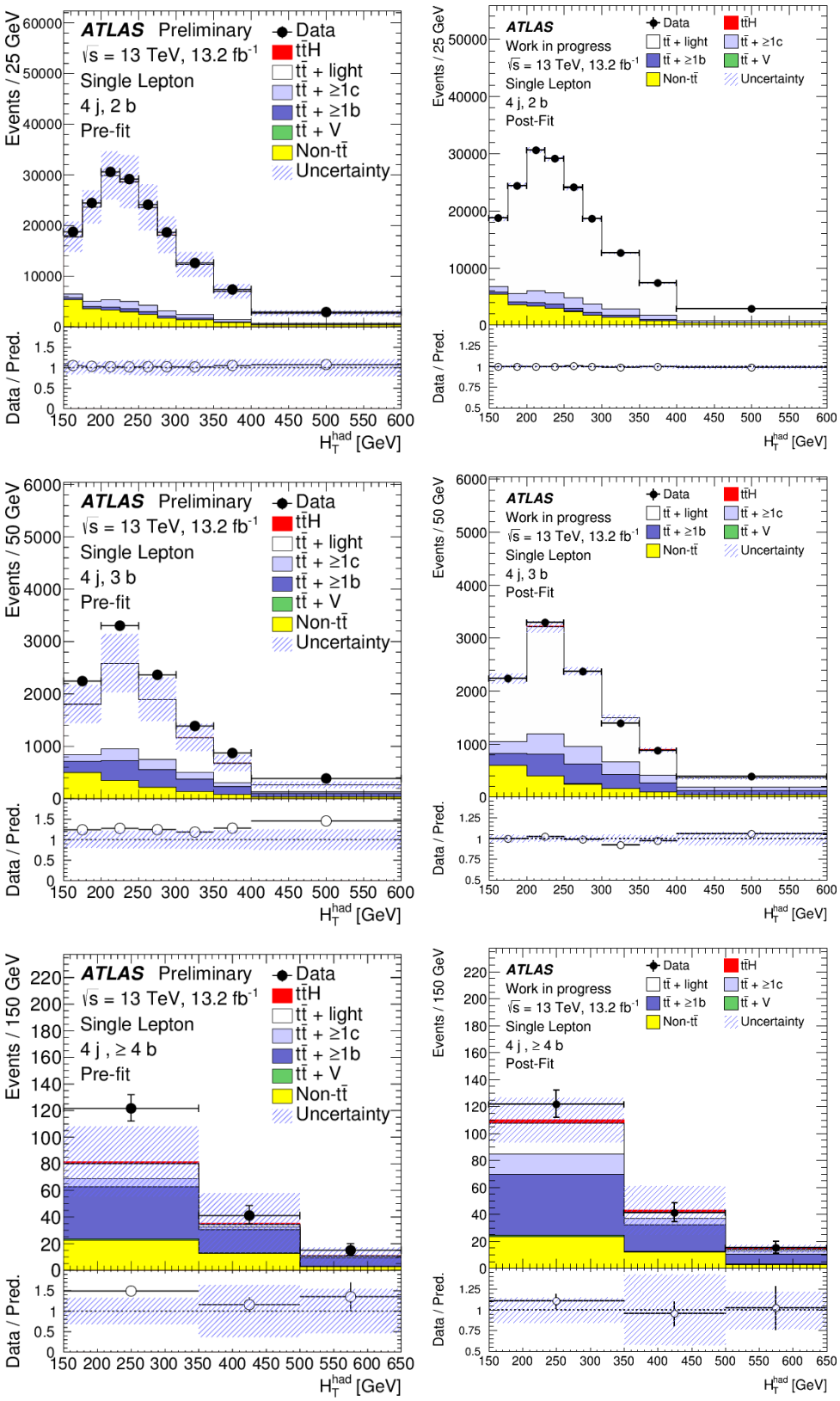


Figure 4.27.: Pre-fit [117] and post-fit plots for the $H_{\mathrm{T}}^{\mathrm{had}}$ variable in the four exclusive jet region. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

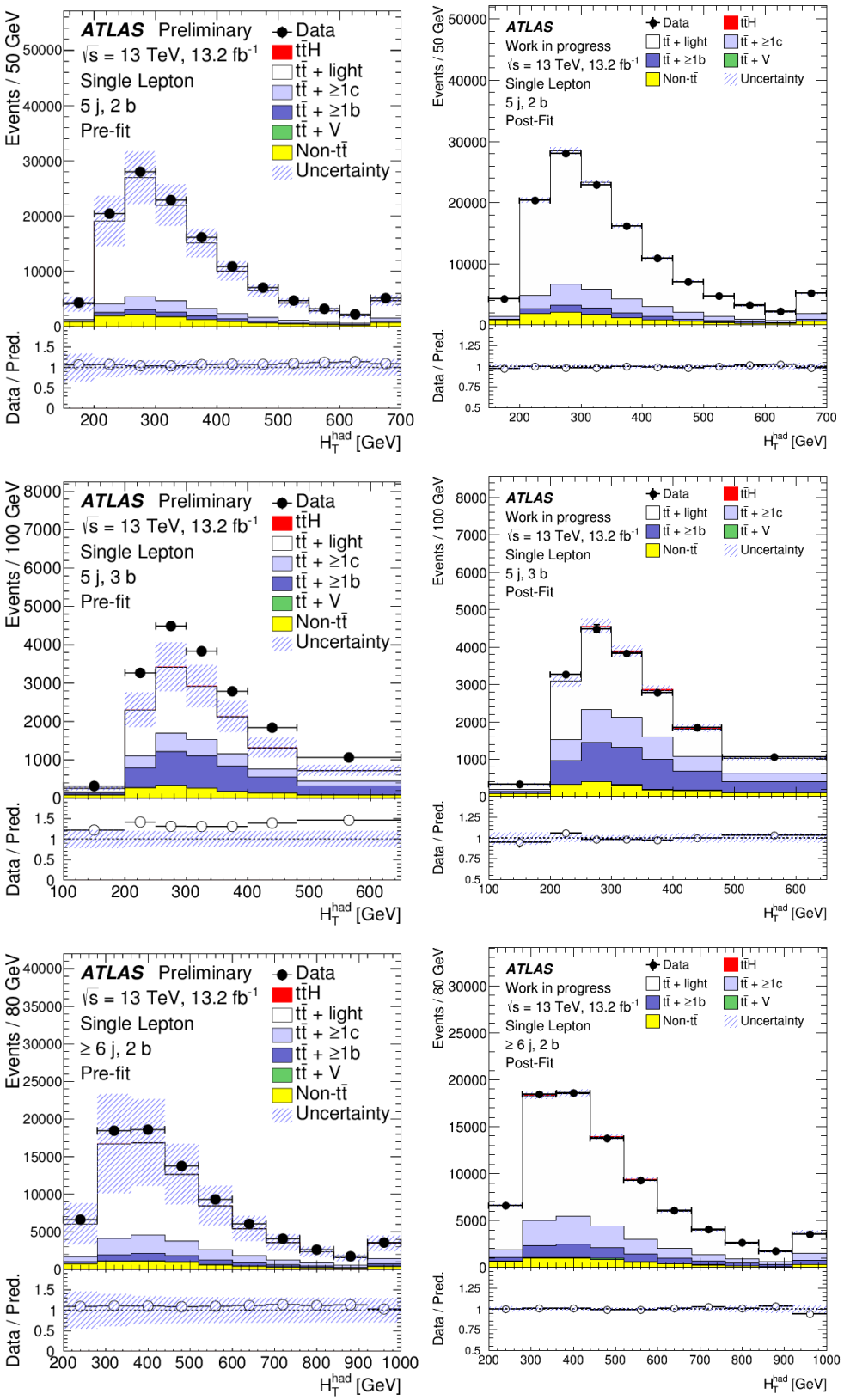


Figure 4.28.: Pre-fit [117] and post-fit plots for the $H_{\mathrm{T}}^{\mathrm{had}}$ variable in the five exclusive jet region. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

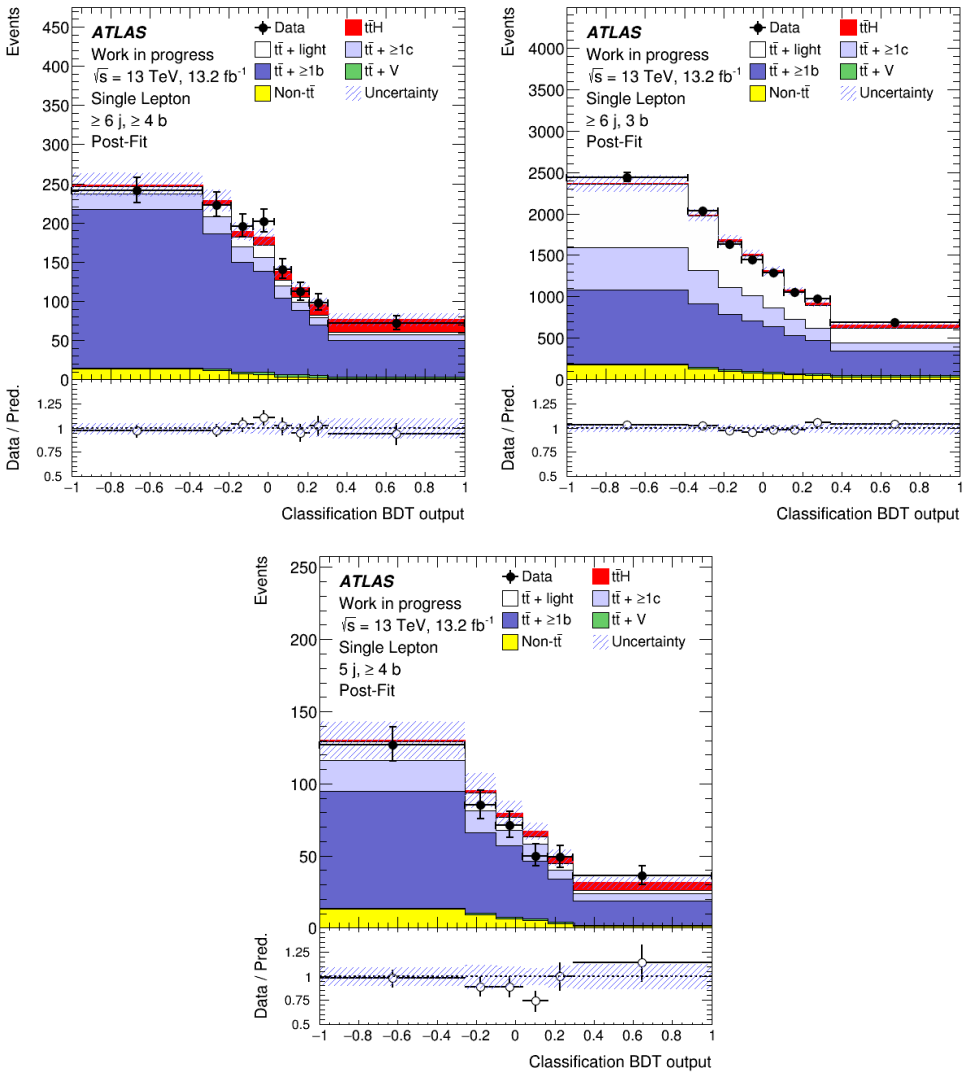


Figure 4.29.: Post-fit plots for the BDT discriminating variable in the signal enriched regions.

	≥6 jets, 2 b-tags	≥6 jets, 3 b-tags	≥6 jets, ≥4 b-tags
$t\bar{t}$+light	58000±5520	3650±684	73.2±36.9
$t\bar{t}+\geq 1c$	13500±6500	2210±777	121±47.9
$t\bar{t}+\geq 1b$	6990±1060	4640±554	935±82.7
$t\bar{t}$+V	504±52	101±11	21.4±3.7
Single top	2460±457	292±72	30.9±11.5
W/Z+jets	1567±481	153±60	12.1±6.2
Diboson	200±99.6	18.4±9.6	2.6±1.5
fakes	1030±315	260±119	1.2±12.3
$t\bar{t}H$	320±136	201±117	79.3±47.1
Total	84700±1220	11500±282	1280±55.7
Data	84556	11561	1285

Table 4.15.: Yields after the fit in the inclusive six jet region.

4.9.1. Combination with the dilepton analysis

A search for the $t\bar{t}H(H \to b\bar{b})$ in the dilepton channel has also been performed by the ATLAS collaboration [117]. It applies a similar analysis strategy; classification of the events according to the number of jets and b-tagged jets and the use of MVA techniques in the signal-rich regions. The MVA procedure is similar to described in this thesis; a MVA-based event reconstruction is implemented, then for the discrimination of the signal and background a BDT is trained using variables from the reconstruction and global kinematic variables. The event selection of the two analysis are designed to be orthogonal hence the combination of both analyses can be performed.

The fitted signal strength for the combined analysis is:

$$\mu = 2.1^{+0.5}_{-0.5}(\text{stat.})^{+0.9}_{-0.7}(\text{syst.}) = 2.1^{+1.0}_{-0.9}, \tag{4.14}$$

which corresponds to an observed significance of 2.4σ where 1.2σ would be expected in the absence of SM $t\bar{t}H(H \to b\bar{b})$ [148].

The results for the single lepton, dilepton and their combination are shown in figure 4.30. A signal 1.9 times larger than the SM prediction is expected to be excluded in absence of the SM $t\bar{t}H(H \to b\bar{b})$. The combination with the dilepton analysis improves the expected limit by 15%. A signal strength larger than 4.0 is excluded by data at 95% CL.

The systematic uncertainties ranked by the post-fit impact on μ are summarised in table 4.16. The source of uncertainties have been grouped into categories. The normalisation factors $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ are included in the statistical component. The sources of systematic uncertainties with the largest impact are those related to the normalisation and modelling of the $t\bar{t}+\geq 1b$, and the jet flavour tagging.

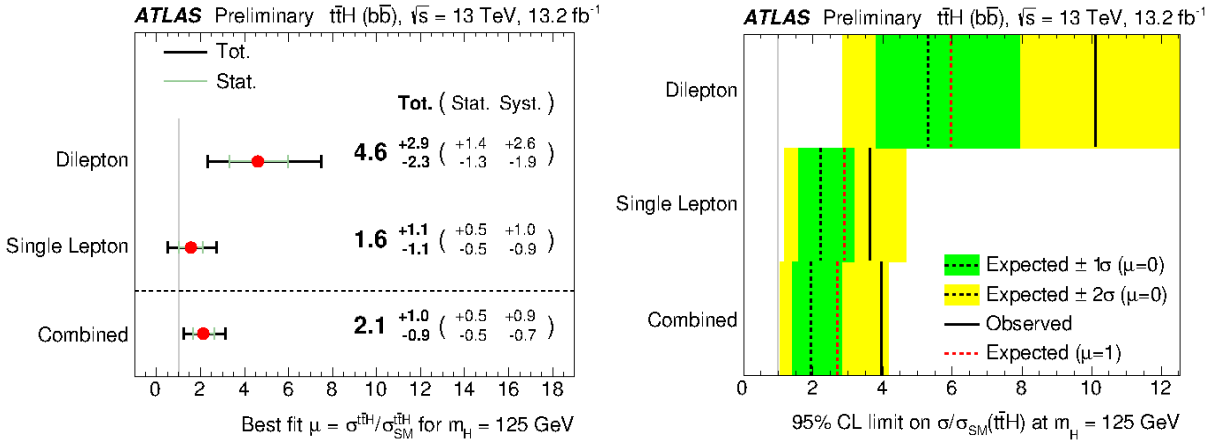


Figure 4.30.: The fitted value of signal strength (left) and the upper limits at 95% CL on cross-section relative to the SM prediction, σ/σ_{SM}, for the individual channels and their combination (right) [117].

Uncertainty Source	$\Delta\mu$	
$t\bar{t}+\geq 1b$ modelling	$+0.53$	-0.53
Jet flavour tagging	$+0.26$	-0.26
$t\bar{t}H$ modelling	$+0.32$	-0.20
Background model statistics	$+0.25$	-0.25
$t\bar{t}+\geq 1c$ modelling	$+0.24$	-0.23
Jet energy scale and resolution	$+0.19$	-0.19
$t\bar{t}$+light modelling	$+0.19$	-0.18
Other background modelling	$+0.18$	-0.18
Jet-vertex association, pileup modelling	$+0.12$	-0.12
Luminosity	$+0.12$	-0.12
$t\bar{t}Z$ modelling	$+0.06$	-0.06
Light lepton (e,μ) ID, isolation, trigger	$+0.05$	-0.05
Total systematic uncertainty	$+0.90$	-0.75
$t\bar{t}+\geq 1b$ normalisation	$+0.34$	-0.34
$t\bar{t}+\geq 1c$ normalisation	$+0.14$	-0.14
Total statistical uncertainty	$+0.49$	-0.49
Total uncertainty	$+1.02$	-0.89

Table 4.16.: Summary of the effect of different sets of systematic uncertainties on the signal strength μ. Since correlations exist between of the different sources of uncertainties, the total systematic uncertainty can be different from the simply combined in quadrature of the individual sources [117].

4.10. Summary

A search for the associated production of the Higgs boson with a top quark pair with the Higgs boson decaying into bottom quarks ($t\bar{t}H(H \to b\bar{b})$) has been performed and presented. The dataset used for this analysis corresponds to an integrated luminosity of 13.2 fb^{-1} from proton-proton collisions at a centre-of-mass energy of 13 TeV, recorded by the ATLAS experiment during 2015 and part of the 2016 data taking period.

The analysis has been carried out in event categories based on the number of jets and b-tagged jets: six signal-depleted regions and three signal-rich regions. In the signal-rich regions a method to reconstruct the $t\bar{t}H(H \to b\bar{b})$ system was implemented. Such a system was not used in the previous Run 1 analysis due to it's complexity and large number of possible combinations of jets. Using a multivariate analysis (Boosted Decision Tree) a good reconstruction efficiency was obtained, with an efficiency of up to 48% to correctly reconstruct the Higgs boson. For the discrimination between the $t\bar{t}H$ signal and the large $t\bar{t}$+jets background, a multivariate analysis (Boosted Decision Tree) was also implemented with variables calculated using the chosen combination from the MVA reconstruction and global kinematic variables. Variables from the MVA reconstruction improve the signal-to-background separation by about 16% in the most sensitive region ($\geq$6 jets, $\geq$4 b-tags).

By performing a fit under the signal-plus-background hypothesis, the best-fit value of μ is found to be 1.6 ± 1.1. The value obtained for μ is compatible with the SM expectation and with background only hypotheses. An observed (expected) 95% confidence level upper limit of 3.6 (2.2) times the SM cross section is obtained.

5. Conclusion

This thesis presented two major studies: the development of a new tagger for the identification of jets containing two b-hadrons (bb-jets) and the search for the Higgs boson in the $t\bar{t}H(H \to b\bar{b})$ single lepton channel.

The ability to identify bb-jets is important to reduce the heavy flavour QCD background in Standard Model (SM) analyses and in new physics searches due to gluon splitting. A method to identify bb- jets based on the reconstruction of secondary vertices inside jets was developed. Properties of the reconstructed vertices, for jets containing at least two reconstructed vertices, are combined in a multivariate analysis (Boosted Decision Tree) to discriminate bb-jets from other jets especially b-jets. The proposed method provides an increase of about 7 times in the separation power between bb-jet and b-jet compared to the default b-tagging algorithm in ATLAS Run 1.

The associated production of a Higgs boson with a top quark pair (ttH channel) is the only way for a direct measurement of the top Yukawa coupling at the LHC. In this thesis a search for the SM Higgs boson produced in association with top quarks, $t\bar{t}H \to (l\nu b)(q\bar{q}'b)(b\bar{b})$ single lepton channel has been presented using 13.2 fb^{-1} of proton-proton collisions at a centre-of-mass energy of 13 TeV recorded by the ATLAS experiment during 2015 and part of the 2016 data taking period.

The full event reconstruction of the $t\bar{t}H(H \to b\bar{b})$ single lepton system is necessary to increase the signal-to-background separation. A MVA-based event reconstruction method was implement in order to find the best corresponding match between the observed jets and the quarks from the decay products of the $t\bar{t}H(H \to b\bar{b})$ system. A MVA approach has also been developed to optimise the separation between the $t\bar{t}H$ signal and the dominant $t\bar{t}$+jets background. This MVA is built using variables from the event reconstruction and global kinematic variables. The signal-to-background separation was improved by about 16% in the most sensitive region ($\geq$6 jets, $\geq$4 b-tags) with help of the MVA reconstruction.

The best-fit value of μ obtained is $1.6^{+1.1(+0.5(\text{stat})+1.0(\text{syst}))}_{-1.1(-0.5(\text{stat})-0.9(\text{syst}))}$. The observed (expected) upper limit on the cross section at 95% confidence level was found to be 3.6 (2.2) times the SM prediction at $m_H = 125$ GeV. The results were found to be consistent either the background-only hypothesis or with the SM prediction.

Bibliography

[1] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. **B716** (2012) 1–29, arXiv:1207.7214 [hep-ex].

[2] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. **B716** (2012) 30–61, arXiv:1207.7235 [hep-ex].

[3] M. Gell-Mann, *A schematic model of baryons and mesons*, Physics Letters **8** no. 3, (1964) 214 – 215. http://www.sciencedirect.com/science/article/pii/S0031916364920013.

[4] G. Altarelli, *Partons in quantum chromodynamics*, Physics Reports **81** no. 1, (1982) 1 – 129. http://www.sciencedirect.com/science/article/pii/0370157382901272.

[5] S. L. Glashow, *Partial-symmetries of weak interactions*, Nuclear Physics **22** no. 4, (1961) 579 – 588. http://www.sciencedirect.com/science/article/pii/0029558261904692.

[6] A. Salam and J. Ward, *Electromagnetic and weak interactions*, Physics Letters **13** no. 2, (1964) 168 – 171. http://www.sciencedirect.com/science/article/pii/0031916364907115.

[7] S. Weinberg, *A Model of Leptons*, Phys. Rev. Lett. **19** (1967) 1264–1266. http://link.aps.org/doi/10.1103/PhysRevLett.19.1264.

[8] F. Englert and R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, Phys. Rev. Lett. **13** (1964) 321–323. http://link.aps.org/doi/10.1103/PhysRevLett.13.321.

[9] P. W. Higgs, *Broken symmetries, massless particles and gauge fields*, Physics Letters **12** no. 2, (1964) 132 – 133. http://www.sciencedirect.com/science/article/pii/0031916364911369.

[10] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, Phys. Rev. Lett. **13** (1964) 508–509. http://link.aps.org/doi/10.1103/PhysRevLett.13.508.

[11] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, *Global Conservation Laws and Massless Particles*, Phys. Rev. Lett. **13** (1964) 585–587. http://link.aps.org/doi/10.1103/PhysRevLett.13.585.

[12] N. Cabibbo, *Unitary Symmetry and Leptonic Decays*, Phys. Rev. Lett. **10** (1963) 531–533. http://link.aps.org/doi/10.1103/PhysRevLett.10.531.

[13] M. Kobayashi and T. Maskawa, *CP Violation in the Renormalizable Theory of Weak Interaction*, Prog. Theor. Phys. **49** (1973) 652–657.

[14] K.A. Olive et al. (Particle Data Group), *Review of Particle Physics*, Chinese Physics C **38** no. 9, (2014) 090001. http://stacks.iop.org/1674-1137/38/i=9/a=090001.

[15] P. Langacker, *Structure of the standard model*, Adv. Ser. Direct. High Energy Phys. **14** (1995) 15–36, arXiv:hep-ph/0304186 [hep-ph].

[16] A. Djouadi, *The Anatomy of electro-weak symmetry breaking. I: The Higgs boson in the standard model*, Phys. Rept. **457** (2008) 1–216, arXiv:hep-ph/0503172 [hep-ph].

[17] LHC Higgs Cross Section Working Group Collaboration, J. R. Andersen et al., *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties: Report of the LHC Higgs Cross Section Working Group*, 2013. https://cds.cern.ch/record/1559921.

[18] ATLAS Collaboration, "Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV." ATLAS-CONF-2015-044, Sep, 2015. http://cds.cern.ch/record/2052552.

[19] L. Evans and P. Bryant, *LHC Machine*, JINST **3** (2008) S08001.

[20] ALICE Collaboration, *The ALICE experiment at the CERN LHC*, JINST **3** (2008) S08002.

[21] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3** (2008) S08003.

[22] CMS Collaboration, *The CMS experiment at the CERN LHC*, JINST **3** (2008) S08004.

[23] LHCb Collaboration, *The LHCb Detector at the LHC*, JINST **3** (2008) S08005.

[24] LHCf Collaboration, *The LHCf detector at the CERN Large Hadron Collider*, JINST **3** (2008) S08006.

[25] T. Collaboration, *The TOTEM experiment at the CERN Large Hadron Collider*, JINST **3** (2008) S08007.

[26] ATLAS Collaboration, "Expected performance of the ATLAS b-tagging algorithms in Run-2." ATL-PHYS-PUB-2015-022, Jul, 2015. https://cds.cern.ch/record/2037697.

[27] ATLAS TDAQ Collaboration, S. Ballestrero, W. Vandelli, and G. Avolio, *ATLAS TDAQ system: current status and performance*, Phys. Procedia **37** (2012) 1819–1826.

[28] ATLAS Collaboration, *Technical Design Report for the Phase-I Upgrade of the ATLAS TDAQ System*, Sep, 2013. https://cds.cern.ch/record/1602235.

[29] Y. Nakahama, *The ATLAS Trigger System: Ready for Run-2*, J. Phys. Conf. Ser. **664** no. 8, (2015) 082037.

[30] T. Hryn'ova and K. Nagano, "Trigger Menu Strategy for Run 2." ATL-COM-DAQ-2014-054, May, 2014. https://cds.cern.ch/record/1703730.

[31] E. Simioni et al., *Upgrade of the ATLAS Level-1 Trigger with event topology information*, J. Phys. Conf. Ser. **664** no. 8, (2015) 082052.

[32] J. Glatzer, *Operation of the Upgraded ATLAS Level-1 Central Trigger System*, J. Phys. Conf. Ser. **664** no. 8, (2015) 082013.

[33] G. Pásztor, *The Upgrade of the ATLAS Electron and Photon Triggers towards LHC Run 2 and their Performance*, in *Meeting of the APS Division of Particles and Fields, Ann Arbor, Michigan, USA, August 4-8*. 2015. arXiv:1511.00334 [hep-ex].

[34] ATLAS Collaboration, *The ATLAS Simulation Infrastructure*, Eur. Phys. J. **C70** (2010) 823–874, arXiv:1005.4568 [physics.ins-det].

[35] I. Bird, P. Buncic, F. Carminati, M. Cattaneo, P. Clarke, I. Fisk, M. Girone, J. Harvey, B. Kersevan, P. Mato, R. Mount, and B. Panzer-Steindel, *Update of the Computing Models of the WLCG and the LHC Experiments*, Apr, 2014. https://cds.cern.ch/record/1695401.

[36] T. Sjöstrand, S. Mrenna, and P. Z. Skands, *A Brief Introduction to PYTHIA 8.1*, Comput. Phys. Commun. **178** (2008) 852, arXiv:0710.3820 [hep-ph].

[37] M. L. Mangano, M. Moretti, F. Piccinini, R. Pittau, and A. D. Polosa, *ALPGEN, a generator for hard multiparton processes in hadronic collisions*, JHEP **07** (2003) 001, arXiv:hep-ph/0206293 [hep-ph].

[38] S. Frixione, P. Nason, and G. Ridolfi, *A Positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction*, JHEP **09** (2007) 126, arXiv:0707.3088 [hep-ph].

[39] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07** (2014) 079, arXiv:1405.0301 [hep-ph].

[40] T. Gleisberg, S. Hoeche, F. Krauss, M. Schonherr, S. Schumann, F. Siegert, and J. Winter, *Event generation with SHERPA 1.1*, JHEP **02** (2009) 007, arXiv:0811.4622 [hep-ph].

[41] GEANT4 Collaboration, S. Agostinelli et al., *GEANT4: A Simulation toolkit*, Nucl. Instrum. Meth. **A506** (2003) 250–303.

[42] W. Lukas, *Fast Simulation for ATLAS: Atlfast-II and ISF*, J. Phys. Conf. Ser. **396** (2012) 022031.

[43] T. G. Cornelissen, N. Van Eldik, M. Elsing, W. Liebig, E. Moyse, N. Piacquadio, K. Prokofiev, A. Salzburger, and A. Wildauer, "Updates of the ATLAS Tracking Event Data Model (Release 13)." ATL-SOFT-PUB-2007-003, Jun, 2007. http://cds.cern.ch/record/1038095.

[44] ATLAS Collaboration, "Track Reconstruction Performance of the ATLAS Inner Detector at $\sqrt{s} = 13$ TeV." ATL-PHYS-PUB-2015-018, Jul, 2015. https://cds.cern.ch/record/2037683.

[45] ATLAS Collaboration, "Measurement of performance of the pixel neural network clustering algorithm of the ATLAS experiment at $\sqrt{s} = 13$ TeV." ATL-PHYS-PUB-2015-044, Sep, 2015. https://cds.cern.ch/record/2054921.

[46] ATLAS Collaboration, "Performance of primary vertex reconstruction in proton-proton collisions at $\sqrt{s} =$7 TeV in the ATLAS experiment." ATLAS-CONF-2010-069, Jul, 2010. https://cds.cern.ch/record/1281344.

[47] W. Waltenberger, R. Frühwirth, and P. Vanlaer, *Adaptive vertex fitting*, Journal of Physics G: Nuclear and Particle Physics **34** no. 12, (2007) N343. http://stacks.iop.org/0954-3899/34/i=12/a=N01.

[48] M. Cacciari, G. P. Salam, and G. Soyez, *The Anti-k(t) jet clustering algorithm*, JHEP **04** (2008) 063, arXiv:0802.1189 [hep-ph].

[49] W. Lampl, S. Laplace, D. Lelas, P. Loch, H. Ma, S. Menke, S. Rajagopalan, D. Rousseau, S. Snyder, and G. Unal, "Calorimeter Clustering Algorithms: Description and Performance." ATL-COM-LARG-2008-003, Apr, 2008. https://cds.cern.ch/record/1099735.

[50] ATLAS Collaboration, *Jet energy measurement with the ATLAS detector in proton-proton collisions at* $\sqrt{s} = 7$ *TeV*, Eur. Phys. J. **C73** no. 3, (2013) 2304, arXiv:1112.6426 [hep-ex].

[51] ATLAS Collaboration, "Pile-up subtraction and suppression for jets in ATLAS." ATLAS-CONF-2013-083, Aug, 2013. https://cds.cern.ch/record/1570994.

[52] ATLAS Collaboration, "Tagging and suppression of pileup jets with the ATLAS detector." ATLAS-CONF-2014-018, May, 2014. https://cds.cern.ch/record/1700870.

[53] ATLAS Collaboration, *Muon reconstruction performance of the ATLAS detector in proton–proton collision data at* $\sqrt{s} = 13$ *TeV*, arXiv:1603.05598 [hep-ex].

[54] ATLAS Collaboration, *Electron reconstruction and identification efficiency measurements with the ATLAS detector using the 2011 LHC proton-proton collision data*, Eur. Phys. J. **C74** no. 7, (2014) 2941, arXiv:1404.2240 [hep-ex].

[55] J. Alison, K. Brendlinger, S. Heim, J. Kroll, and C. M. Lester, "Description and Performance of the Electron Likelihood Tool at ATLAS using 2012 LHC Data." ATL-COM-PHYS-2013-378, Apr, 2013. https://cds.cern.ch/record/1537410.

[56] C. Anastopoulos et al., "Electron efficiency measurements using the 2015 LHC proton-proton collision data." ATLAS-COM-CONF-2016-028, Mar, 2016. https://cds.cern.ch/record/2142831.

[57] ATLAS Collaboration, "Expected performance of missing transverse momentum reconstruction for the ATLAS detector at $\sqrt{s} = 13$ TeV." ATL-PHYS-PUB-2015-023, 2015. http://cdsweb.cern.ch/record/2037700.

[58] ATLAS Collaboration, "Performance of missing transverse momentum reconstruction with the ATLAS detector in the first proton–proton collisions at $\sqrt{s} = 13$ TeV." ATL-PHYS-PUB-2015-027, 2015. http://cdsweb.cern.ch/record/2037904.

[59] CMS Collaboration, *Measurement of* $B\bar{B}$ *Angular Correlations based on Secondary Vertex Reconstruction at* $\sqrt{s} = 7$ *TeV*, JHEP **03** (2011) 136, arXiv:1102.3194 [hep-ex].

[60] S. Dawson, C. Jackson, L. H. Orr, L. Reina, and D. Wackeroth, *Associated Higgs production with top quarks at the large hadron collider: NLO QCD corrections*, Phys. Rev. **D68** (2003) 034022, arXiv:hep-ph/0305087 [hep-ph].

[61] S. Frixione and B. R. Webber, *Matching NLO QCD computations and parton shower simulations*, JHEP **06** (2002) 029, arXiv:hep-ph/0204244 [hep-ph].

[62] R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, R. Pittau, and P. Torrielli, *Scalar and pseudoscalar Higgs production in association with a top–antitop pair*, Phys. Lett. **B701** (2011) 427–433, arXiv:1104.5613 [hep-ph].

[63] F. Cascioli, P. Maierhöfer, N. Moretti, S. Pozzorini, and F. Siegert, *NLO matching for* $t\bar{t}b\bar{b}$ *production with massive* b*-quarks*, Phys. Lett. **B734** (2014) 210–214, arXiv:1309.5912 [hep-ph].

[64] F. Cascioli, P. Maierhöfer, N. Moretti, S. Pozzorini, and F. Siegert, *NLO matching for $t\bar{t}b\bar{b}$ production with massive b-quarks*, Phys. Lett. **B734** (2014) 210–214, arXiv:1309.5912 [hep-ph].

[65] ATLAS Collaboration, "Identification and Tagging of Double b-hadron jets with the ATLAS Detector." ATLAS-CONF-2012-100, 2012. http://cds.cern.ch/record/1462603.

[66] ATLAS Collaboration, *Performance of b-Jet Identification in the ATLAS Experiment*, JINST **11** no. 04, (2016) P04008, arXiv:1512.01094 [hep-ex].

[67] L. Alio, *Search for the Higgs boson produced in association with a Z boson and decaying to a pair of bottom quarks with the ATLAS experiment at LHC*. PhD thesis, Marseille, CPPM, 2014. https://cds.cern.ch/record/2012603. presented 12 Nov 2014.

[68] ATLAS Collaboration, *Measurement of the flavour composition of dijet events in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector*, Eur. Phys. J. **C73** no. 2, (2013) 2301, arXiv:1210.0441 [hep-ex].

[69] G. Piacquadio and C. Weiser, *A new inclusive secondary vertex algorithm for b-jet tagging in ATLAS*, J. Phys. Conf. Ser. **119** (2008) 032032.

[70] R. Fruhwirth, *Application of Kalman filtering to track and vertex fitting*, Nucl. Instrum. Meth. **A262** (1987) 444–450.

[71] ATLAS Collaboration, "Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run." ATL-PHYS-PUB-2016-012, 2016. https://cds.cern.ch/record/2160731.

[72] J. Siek, L.-Q. Lee, and A. Lumsdaine, *The Boost Graph Library: User Guide and Reference Manual*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2002.

[73] M. Cristinziani, M. Ghneimat, and V. Kostyukhin, "Secondary Vertex Finding for Jet Flavour Identification." ATL-COM-PHYS-2016-040, Jan, 2016. https://cds.cern.ch/record/2124747.

[74] J. M. Campbell, R. K. Ellis, F. Maltoni, and S. Willenbrock, *Production of a W boson and two jets with one b-quark tag*, Phys. Rev. **D75** (2007) 054015, arXiv:hep-ph/0611348 [hep-ph].

[75] T. Sjostrand, S. Mrenna, and P. Z. Skands, *PYTHIA 6.4 Physics and Manual*, JHEP **05** (2006) 026, arXiv:hep-ph/0603175 [hep-ph].

[76] D. J. Lange, *The EvtGen particle decay simulation package*, Nucl. Instrum. Meth. **A462** (2001) 152–155.

[77] ATLAS Collaboration, "Summary of ATLAS Pythia 8 tunes." ATL-PHYS-PUB-2012-003, 2012. http://cds.cern.ch/record/1474107.

[78] ATLAS Collaboration, "Performance of the ATLAS Inner Detector Track and Vertex Reconstruction in High Pile-Up LHC Environment." ATLAS-CONF-2012-042, 2012. http://cdsweb.cern.ch/record/1435196.

[79] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss, *TMVA: Toolkit for Multivariate Data Analysis*, PoS **ACAT** (2007) 040, arXiv:physics/0703039.

[80] Breiman, Leo and Friedman, Jerome and Olshen, R. A. and Stone, Charles J., *Classification and regression trees*. Wadsworth, Stanford, 1984.

[81] Y. Coadou, *Boosted Decision Trees and Applications*, EPJ Web of Conferences **55** (2013) .

[82] Y. Freund and R. E. Schapire, *Experiments with a New Boosting Algorithm*, Proceedings of the Thirteenth International Conference on Machine Learning (1996) 148.

[83] CMS Collaboration, *Search for the associated production of the Higgs boson with a top-quark pair*, JHEP **2014** no. 9, (2014) 1–64. http://dx.doi.org/10.1007/JHEP09(2014)087.

[84] CMS Collaboration, *Search for a standard model Higgs boson produced in association with a top-quark pair and decaying to bottom quarks using a matrix element method*, Eur. Phys. J. **C75** no. 6, (2015) 1–28. http://dx.doi.org/10.1140/epjc/s10052-015-3454-1.

[85] ATLAS Collaboration, *Search for the Standard Model Higgs boson produced in association with top quarks and decaying into $b\bar{b}$ in pp collisions at $\sqrt{s}$ = 8 TeV with the ATLAS detector*, Eur. Phys. J. **C75** no. 7, (2015) 1–50. http://dx.doi.org/10.1140/epjc/s10052-015-3543-1.

[86] ATLAS Collaboration, *Search for the Standard Model Higgs boson decaying into $b\bar{b}$ produced in association with top quarks decaying hadronically in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, JHEP **05** (2016) 160, arXiv:1604.03812 [hep-ex].

[87] ATLAS and CMS Collaboration, *Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV*, arXiv:1606.02266 [hep-ex].

[88] CMS Collaboration, *Search for $\mathrm{t\bar{t}H}$ production in the $\mathrm{H} \to \mathrm{b\bar{b}}$ decay channel with $\sqrt{s} = 13$ TeV pp collisions at the CMS experiment*,. https://cds.cern.ch/record/2139578.

[89] ATLAS Collaboration, "2015 start-up trigger menu and initial performance assessment of the ATLAS trigger using Run-2 data." ATL-DAQ-PUB-2016-001, 2016. https://cds.cern.ch/record/2136007/.

[90] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, and T. Stelzer, *MadGraph 5 : Going Beyond*, JHEP **06** (2011) 128, arXiv:1106.0522 [hep-ph].

[91] NNPDF Collaboration, R. D. Ball et al., *Parton distributions for the LHC Run II*, JHEP **04** (2015) 040, arXiv:1410.8849 [hep-ph].

[92] P. Nason, *A New method for combining NLO QCD with shower Monte Carlo algorithms*, JHEP **0411** (2004) 040, arXiv:hep-ph/0409146.

[93] S. Frixione, P. Nason, and C. Oleari, *Matching NLO QCD computations with Parton Shower simulations: the POWHEG method*, JHEP **0711** (2007) 070, arXiv:0709.2092 [hep-ph].

[94] S. Alioli, P. Nason, C. Oleari, and E. Re, *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, JHEP **1006** (2010) 043, arXiv:1002.2581 [hep-ph].

[95] H.-L. Lai, M. Guzzi, J. Huston, Z. Li, P. M. Nadolsky, et al., *New parton distributions for collider physics*, Phys. Rev. D **82** (2010) 074024, arXiv:1007.2241 [hep-ph].

[96] P. M. Nadolsky, H.-L. Lai, Q.-H. Cao, J. Huston, J. Pumplin, et al., *Implications of CTEQ global analysis for collider observables*, Phys. Rev. D **78** (2008) 013004, arXiv:0802.0007 [hep-ph].

[97] P. Z. Skands, *Tuning Monte Carlo Generators: The Perugia Tunes*, Phys. Rev. D **82** (2010) 074018, arXiv:1005.3457 [hep-ph].

[98] M. Czakon and A. Mitov, *Top++: A Program for the Calculation of the Top-Pair Cross-Section at Hadron Colliders*, Comput. Phys. Commun. **185** (2014) 2930, arXiv:1112.5675 [hep-ph].

[99] M. Cacciari, M. Czakon, M. Mangano, A. Mitov, and P. Nason, *Top-pair production at hadron colliders with next-to-next-to-leading logarithmic soft-gluon resummation*, Phys. Lett. **B710** (2012) 612, arXiv:1111.5869 [hep-ph].

[100] P. Bärnreuther, M. Czakon, and A. Mitov, *Percent Level Precision Physics at the Tevatron: First Genuine NNLO QCD Corrections to $q\bar{q} \to t\bar{t}$*, Phys. Rev. Lett. **109** (2012) 132001, arXiv:1204.5201 [hep-ph].

[101] M. Czakon and A. Mitov, *NNLO corrections to top-pair production at hadron colliders: the all-fermionic scattering channels*, JHEP **1212** (2012) 054, arXiv:1207.0236 [hep-ph].

[102] M. Czakon and A. Mitov, *NNLO corrections to top-pair production at hadron colliders: the quark-gluon reaction*, JHEP **1301** (2013) 080, arXiv:1210.6832 [hep-ph].

[103] M. Czakon, P. Fiedler, and A. Mitov, *The total top quark pair production cross-section at hadron colliders through $\mathcal{O}(\alpha_S^4)$*, Phys. Rev. Lett. **110** (2013) 252004, arXiv:1303.6254 [hep-ph].

[104] T. Gleisberg and S. Höche, *Comix, a new matrix element generator*, JHEP **0812** (2008) 039, arXiv:0808.3674 [hep-ph].

[105] S. Schumann and F. Krauss, *A Parton shower algorithm based on Catani-Seymour dipole factorisation*, JHEP **0803** (2008) 038, arXiv:0709.1027 [hep-ph].

[106] S. Höche, F. Krauss, M. Schönherr, and F. Siegert, *QCD matrix elements + parton showers: The NLO case*, JHEP **04** (2013) 027, arXiv:1207.5030 [hep-ph].

[107] J. Butterworth et al., "Single Boson and Diboson Production Cross Sections in pp Collisions at $\sqrt{s}$=7 TeV." ATL-COM-PHYS-2010-695, 2010. https://cds.cern.ch/record/1287902.

[108] S. Frixione, E. Laenen, P. Motylinski, B. R. Webber, and C. D. White, *Single-top hadroproduction in association with a W boson*, JHEP **0807** (2008) 029, arXiv:0805.3067 [hep-ph].

[109] N. Kidonakis, *Two-loop soft anomalous dimensions for single top quark associated production with a W- or H-*, Phys. Rev. **D82** (2010) 054018, arXiv:1005.4451 [hep-ph].

[110] N. Kidonakis, *NNLL resummation for s-channel single top quark production*, Phys. Rev. **D81** (2010) 054028, arXiv:1001.5034 [hep-ph].

[111] N. Kidonakis, *Next-to-next-to-leading-order collinear and soft gluon corrections for t-channel single top quark production*, Phys. Rev. **D83** (2011) 091503, arXiv:1103.2792 [hep-ph].

[112] "Lepton isolation recommendations." https://twiki.cern.ch/twiki/bin/viewauth/AtlasProtected/IsolationSelectionTool.

[113] ATLAS Collaboration, "Muon Combined Performance in Run 2 (25 ns runs)." ATL-COM-MUON-2015-093, Nov, 2015. https://cds.cern.ch/record/2105495.

[114] ATLAS Collaboration, "Jet calibration and systematic uncertainties for jets reconstructed in the ATLAS detector at $\sqrt{s}$ =13 TeV." ATLAS-PHYS-PUB-2015-015, 2015. http://cdsweb.cern.ch/record/2037613.

[115] ATLAS Collaboration, "Monte Carlo Calibration and Combination of In-Situ Measurements of Jet Energy Scale, Jet Energy Resolution and Jet Mass in ATLAS." ATLAS-CONF-2015-037, 2015. http://cdsweb.cern.ch/record/2044941.

[116] ATLAS Collaboration, "Selection of jets produced in 13 TeV proton-proton collisions with the ATLAS detector." ATLAS-CONF-2015-029, 2015. http://cdsweb.cern.ch/record/2037702.

[117] ATLAS Collaboration, "Search for the Standard Model Higgs boson produced in association with top quarks and decaying into $b\bar{b}$ in pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector." ATLAS-CONF-2016-080, 2016. https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2016-080.

[118] ATLAS Collaboration, "Measurement of the Top-Quark Mass using the Template Method in pp Collisions at $\sqrt{s}$ =7 TeV with the ATLAS detector." ATLAS-CONF-2011-033, Mar, 2011. https://cds.cern.ch/record/1337783.

[119] ATLAS Collaboration, "Search for the Standard Model Higgs boson produced in association with top quarks in proton-proton collisions at $\sqrt{s}$ = 7 TeV using the ATLAS detector." ATLAS-CONF-2012-135, Sep, 2012. https://cds.cern.ch/record/1478423.

[120] G. Aad, *Mise en service du détecteur à pixels de l'expérience ATLAS auprès du LHC et étude du canal ttH, H->bb pour la recherche du boson de Higgs* . PhD thesis, Marseille, CPPM, 2009. http://cds.cern.ch/record/1233933. presented 18 Sep 2009.

[121] C. Bernaciak, M. S. A. Buschmann, A. Butter, and T. Plehn, *Fox-Wolfram Moments in Higgs Physics*, Phys. Rev. **D87** (2013) 073014, arXiv:1212.4436 [hep-ph].

[122] V. D. Barger, J. Ohnemus, and R. J. N. Phillips, *Event shape criteria for single lepton top signals*, Phys. Rev. **D48** (1993) 3953–3956, arXiv:hep-ph/9308216 [hep-ph].

[123] ATLAS Collaboration, *Measurements of normalized differential cross sections for $t\bar{t}$ production in pp collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector*, Phys. Rev. **D90** no. 7, (2014) 072004, arXiv:1407.0371 [hep-ex].

[124] M. Czakon, D. Heymes, and A. Mitov, *High-precision differential predictions for top-quark pairs at the LHC*, Phys. Rev. Lett. **116** no. 8, (2016) 082003, arXiv:1511.00549 [hep-ph].

[125] M. Czakon, P. Fiedler, and A. Mitov, *Total Top-Quark Pair-Production Cross Section at Hadron Colliders Through $O(\alpha_S^4)$*, Phys. Rev. Lett. **110** (2013) 252004, arXiv:1303.6254 [hep-ph].

[126] T. Gleisberg, S. Hoeche, F. Krauss, M. Schonherr, S. Schumann, et al., *Event generation with SHERPA 1.1*, JHEP **0902** (2009) 007, arXiv:0811.4622 [hep-ph].

[127] F. Cascioli, P. Maierhofer, and S. Pozzorini, *Scattering Amplitudes with Open Loops*, Phys. Rev. Lett. **108** (2012) 111601, arXiv:1111.5206 [hep-ph].

[128] ATLAS Collaboration, "Estimation of non-prompt and fake lepton backgrounds in final states with top quarks produced in proton–proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS Detector." ATLAS-CONF-2014-058, 2014. http://cdsweb.cern.ch/record/1951336.

[129] F. Derue, "Estimation of fake lepton background for top analyses using the Matrix Method with the 2015 dataset at $\sqrt{s}$= 13 TeV with AnalysisTop-2.3.41." ATL-COM-PHYS-2016-198, 2016. https://cds.cern.ch/record/2135116.

[130] ATLAS Collaboration, *Improved luminosity determination in pp collisions at sqrt(s) = 7 TeV using the ATLAS detector at the LHC*, Eur. Phys. J. **C73** no. 8, (2013) 2518, arXiv:1302.4393 [hep-ex].

[131] ATLAS Collaboration, "Jet Calibration and Systematic Uncertainties for Jets Reconstructed in the ATLAS Detector at $\sqrt{s} = 13$ TeV." ATL-PHYS-PUB-2015-015, 2015. https://cds.cern.ch/record/2037613.

[132] ATLAS Collaboration, "Calibration of b-tagging using dileptonic top pair events in a combinatorial likelihood approach with the ATLAS experiment." ATLAS-CONF-2014-004, 2014. http://cdsweb.cern.ch/record/1664335.

[133] ATLAS Collaboration, "Calibration of the performance of b-tagging for c and light-flavour jets in the 2012 ATLAS data." ATLAS-CONF-2014-046, 2014. http://cdsweb.cern.ch/record/1741020.

[134] ATLAS Collaboration, "Simulation of top-quark production for the ATLAS experiment at $\sqrt{s} = 13$ TeV." ATL-PHYS-PUB-2016-004, 2016. http://cdsweb.cern.ch/record/2120417.

[135] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, *Parton distributions for the LHC*, Eur. Phys. J. **C63** (2009) 189–285, arXiv:0901.0002 [hep-ph].

[136] NNPDF Collaboration, R. D. Ball et al., *Parton distributions for the LHC Run II*, JHEP **04** (2015) 040, arXiv:1410.8849 [hep-ph].

[137] ATLAS Collaboration, "Studies of tt + cc production with MadGraph5_aMC@NLO and Herwig++ for the ATLAS experiment." ATL-PHYS-PUB-2016-011, 2016. http://cdsweb.cern.ch/record/2153876.

[138] J. M. Campbell and R. K. Ellis, *$t\bar{t}W^{+-}$ production and decay at NLO*, JHEP **07** (2012) 052, arXiv:1204.5678 [hep-ph].

[139] ATLAS Collaboration, "Multi-boson simulation for 13 TeV ATLAS analyses." ATL-PHYS-PUB-2016-002, 2016. http://cdsweb.cern.ch/record/2119986.

[140] R. Raitio and W. W. Wada, *Higgs Boson Production at Large Transverse Momentum in QCD*, Phys. Rev. **D19** (1979) 941.

[141] W. Beenakker, S. Dittmaier, M. Krämer, B. Plumper, M. Spira, et al., *NLO QCD corrections to $t\bar{t}H$ production in hadron collisions*, Nucl. Phys. **B653** (2003) 151–203, arXiv:hep-ph/0211352 [hep-ph].

[142] S. Dawson, C. Jackson, L. Orr, L. Reina, and D. Wackeroth, *Associated Higgs production with top quarks at the large hadron collider: NLO QCD corrections*, Phys. Rev. **D68** (2003) 034022, arXiv:hep-ph/0305087 [hep-ph].

[143] Y. Zhang, W.-G. Ma, R.-Y. Zhang, C. Chen, and L. Guo, *QCD NLO and EW NLO corrections to $t\bar{t}H$ production with top quark decays at hadron collider*, Phys. Lett. **B738** (2014) 1–5, arXiv:1407.1110 [hep-ph].

[144] S. Frixione, V. Hirschi, D. Pagani, H.-S. Shao, and M. Zaro, *Electroweak and QCD corrections to top-pair hadroproduction in association with heavy bosons*, JHEP **06** (2015) 184, arXiv:1504.03446 [hep-ph].

[145] A. Djouadi, J. Kalinowski, and M. Spira, *HDECAY: A Program for Higgs boson decays in the standard model and its supersymmetric extension*, Comput. Phys. Commun. **108** (1998) 56–74, arXiv:hep-ph/9704448 [hep-ph].

[146] A. L. Read, *Presentation of search results: The CL(s) technique*, J. Phys. G **28** (2002) 2693.

[147] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. C **71** (2011) 1554, arXiv:1007.1727 [physics.data-an].

[148] ATLAS Collaboration, "Combination of the searches for Higgs boson production in association with top quarks in the $\gamma\gamma$, multilepton, and $b\bar{b}$ decay channels at $\sqrt{s}$=13 TeV with the ATLAS Detector." ATLAS-CONF-2016-068, 2016. https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2016-068.

Appendix

A. Auxiliary materials for the reconstruction BDT

A.1. Relative reconstruction efficiency

The ratio of the reconstruction efficiency to the maximum achievable matching efficiency is shown in figure A.1.

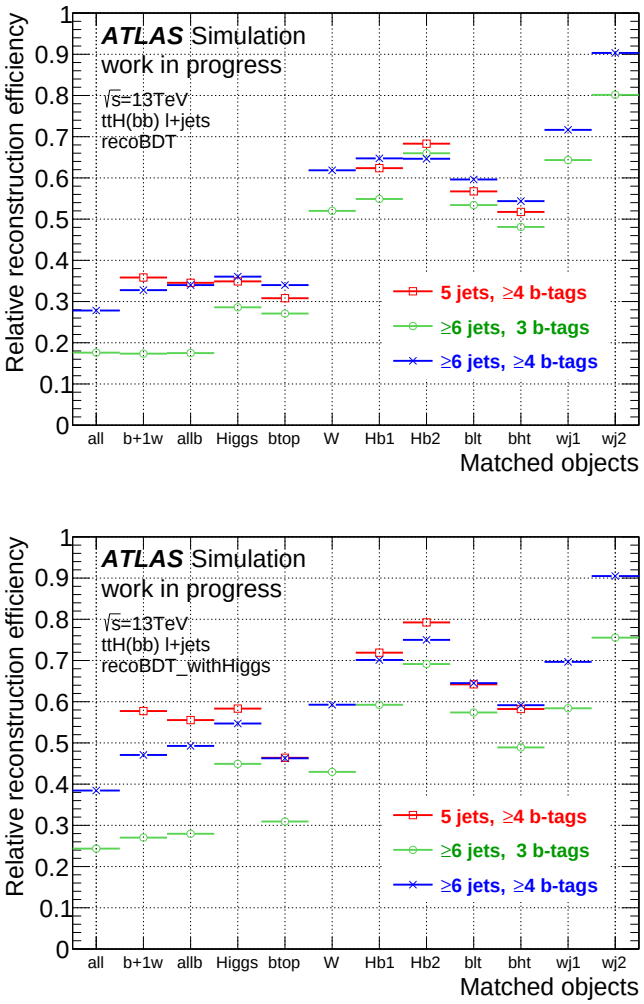


Figure A.1.: Relative reconstruction efficiency for different objects in the signal regions for recoBDT (top) and recoBDT_withHiggs (bottom)

A.2. Pre-fit and post-fit distributions of the reconstruction BDT output

This section shows the distributions of the highest reconstruction BDT output per event before and after the fit to data. The pre-fit plots do not include normalisation factor to $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$. The post-fit plots correspond to the single lepton only fit to data. A good agreement between data and MC simulation is found in the region with more statistics, $\geq$6 jets, 3 b-tags. In 5 jets, $\geq$4 b-tags and $\geq$6 jets, $\geq$4 b-tags statistical fluctuation can be observed.

A.2.1. Region: 5 jets, $\geq$4 b-tags

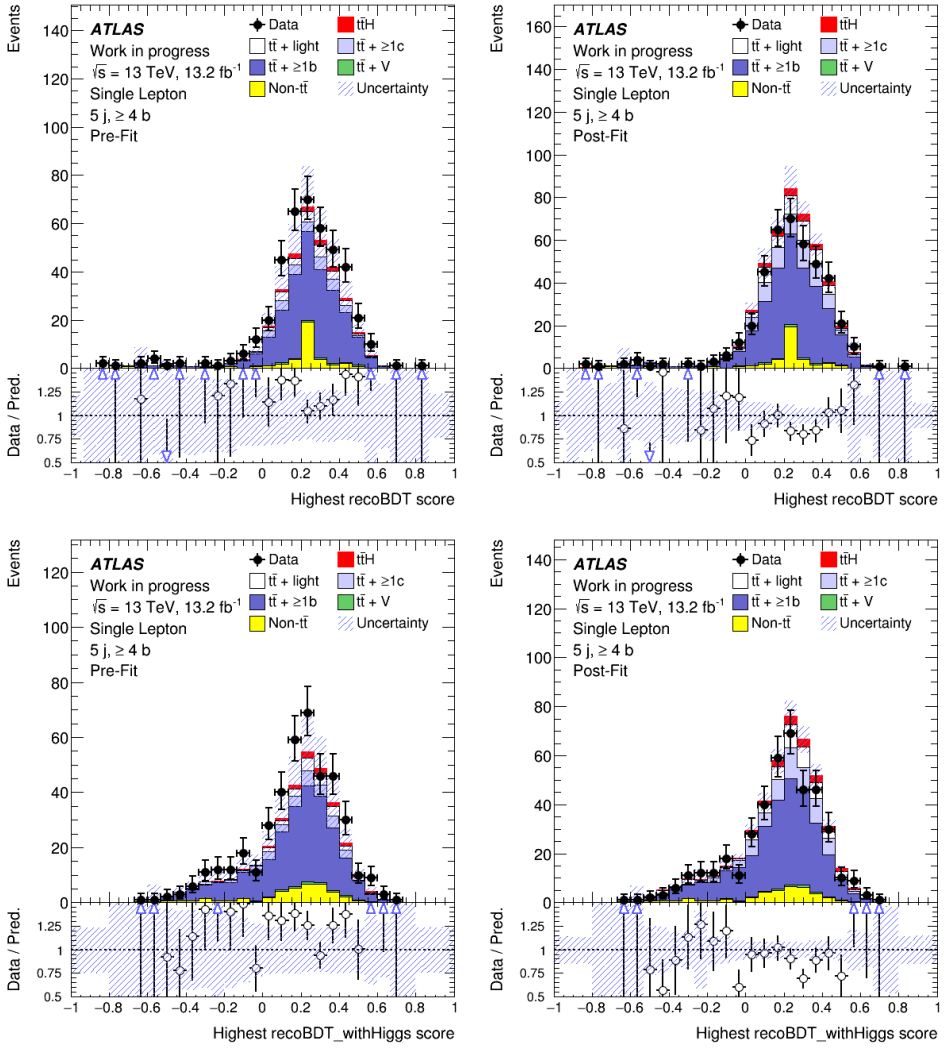


Figure A.2.: Distributions of the highest reconstruction BDT score before and after the fitting procedure in the 5 jets, $\geq$4 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

A.2.2. Region: ≥6 jets, 3 b-tags

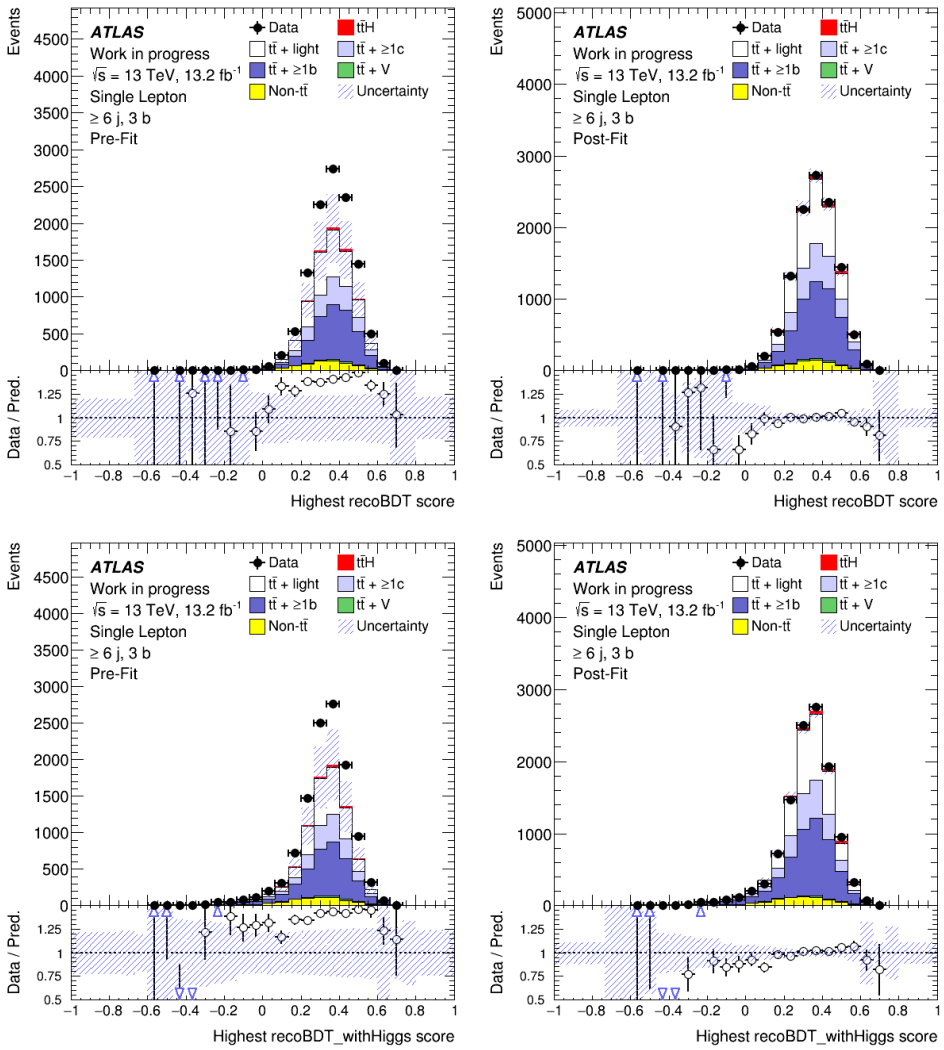


Figure A.3.: Distributions of the highest reconstruction BDT score before and after the fitting procedure in the ≥6 jets, 3 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

A.2.3. Region: ≥6 jets, ≥4 b-tags

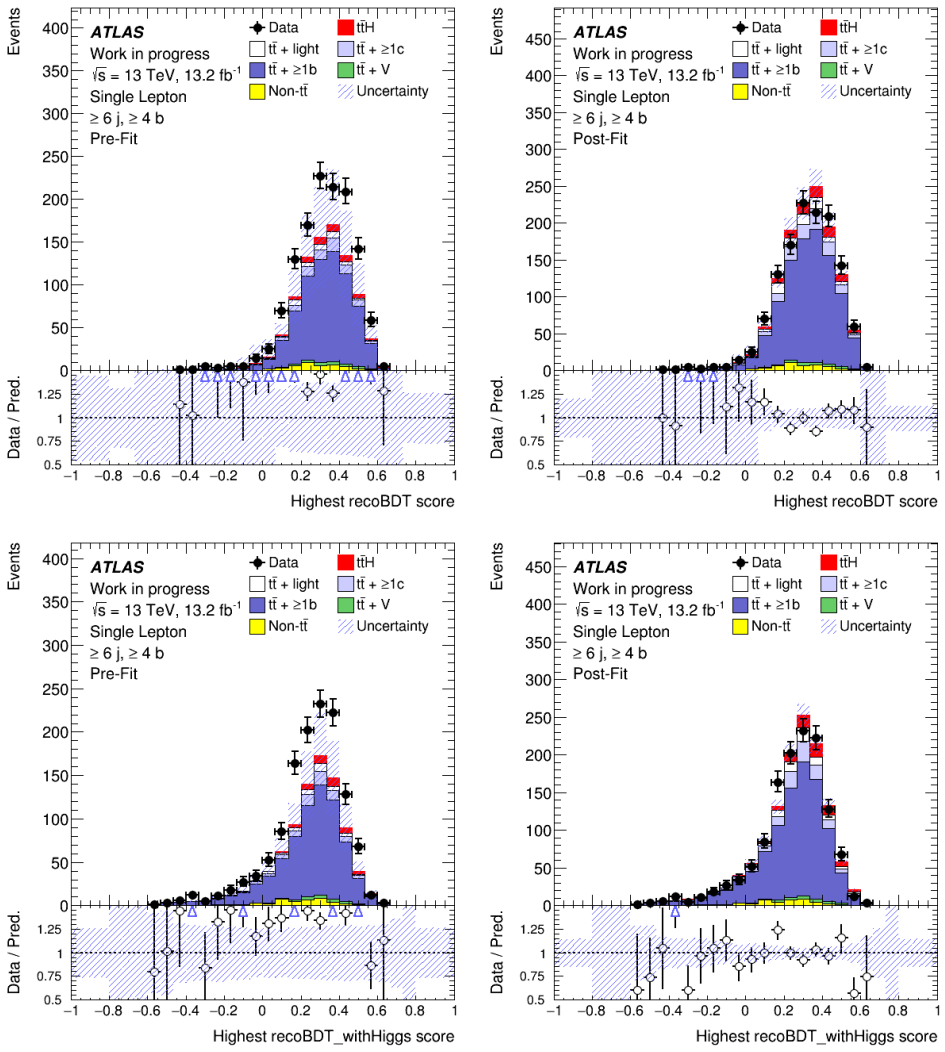


Figure A.4.: Distributions of the highest reconstruction BDT score before and after the fitting procedure in the ≥6 jets, ≥4 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

B. Auxiliary materials for the classification BDT

B.1. Pre-fit and post-fit distributions of the input variables for the classification BDT

This section shows the distributions of the input variables before and after the fit to data. The pre-fit plots do not include normalisation factor to $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$. The post-fit plots correspond to the single lepton only fit to data. A good agreement between data and MC simulation is found in the region with more statistics, $\geq$6 jets, 3 b-tags. In 5 jets, $\geq$4 b-tags and $\geq$6 jets, $\geq$4 b-tags statistical fluctuation can be observed.

B.1.1. Region: 5 jets, $\geq$4 b-tags

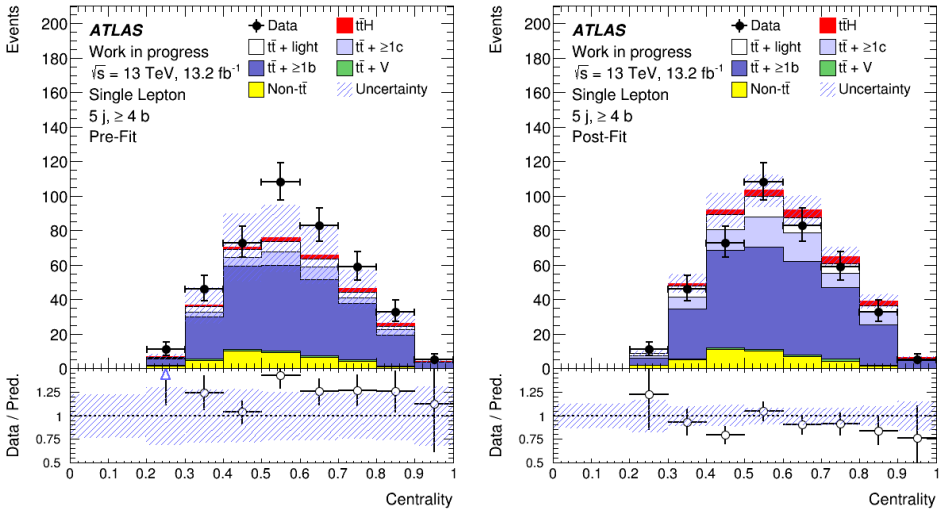


Figure B.1.: Distributions of the discriminating variables before and after the fitting procedure in the 5 jets, $\geq$4 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distribution does not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

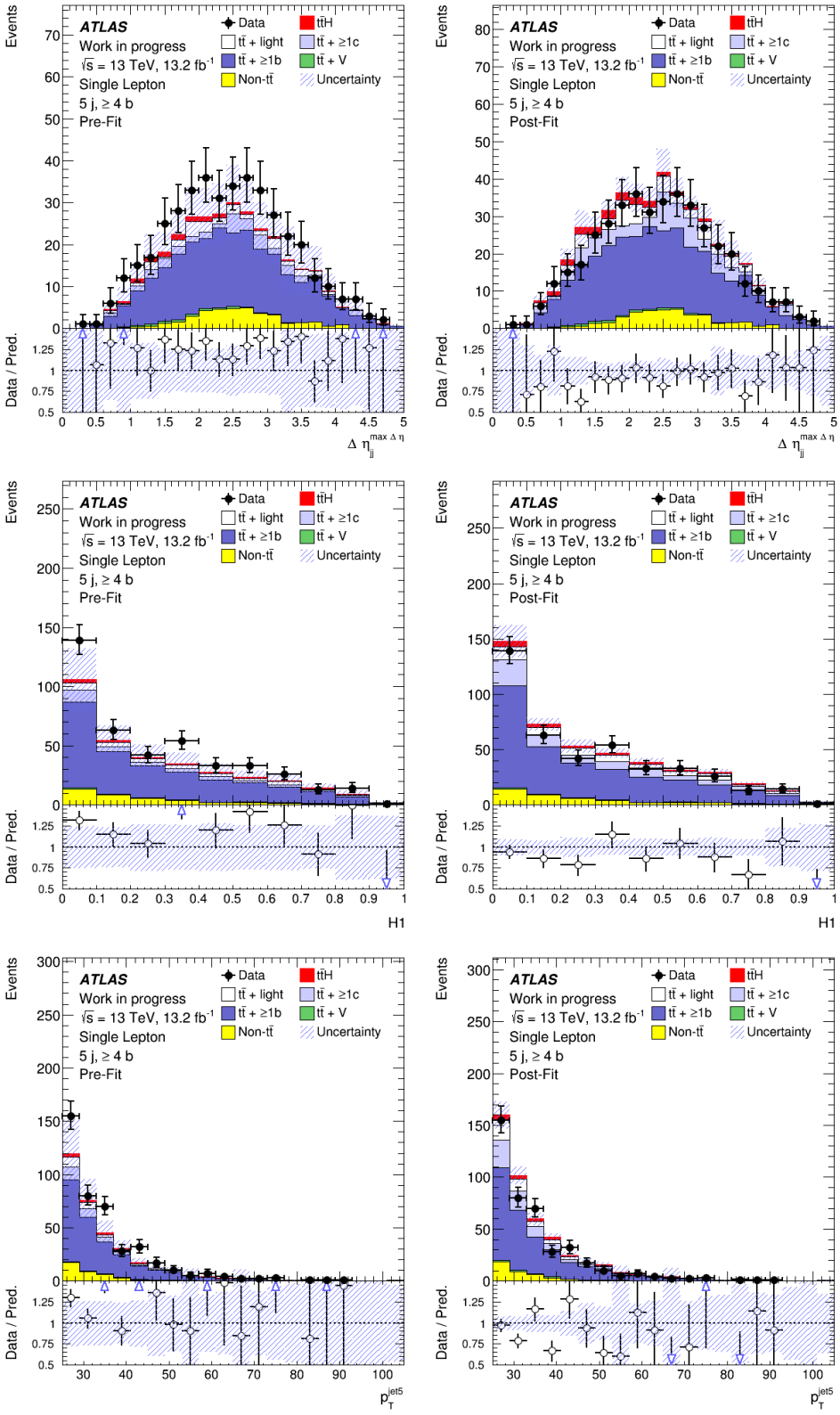


Figure B.2.: Distributions of the discriminating variables before and after the fitting procedure in the 5 jets, $\geq$4 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

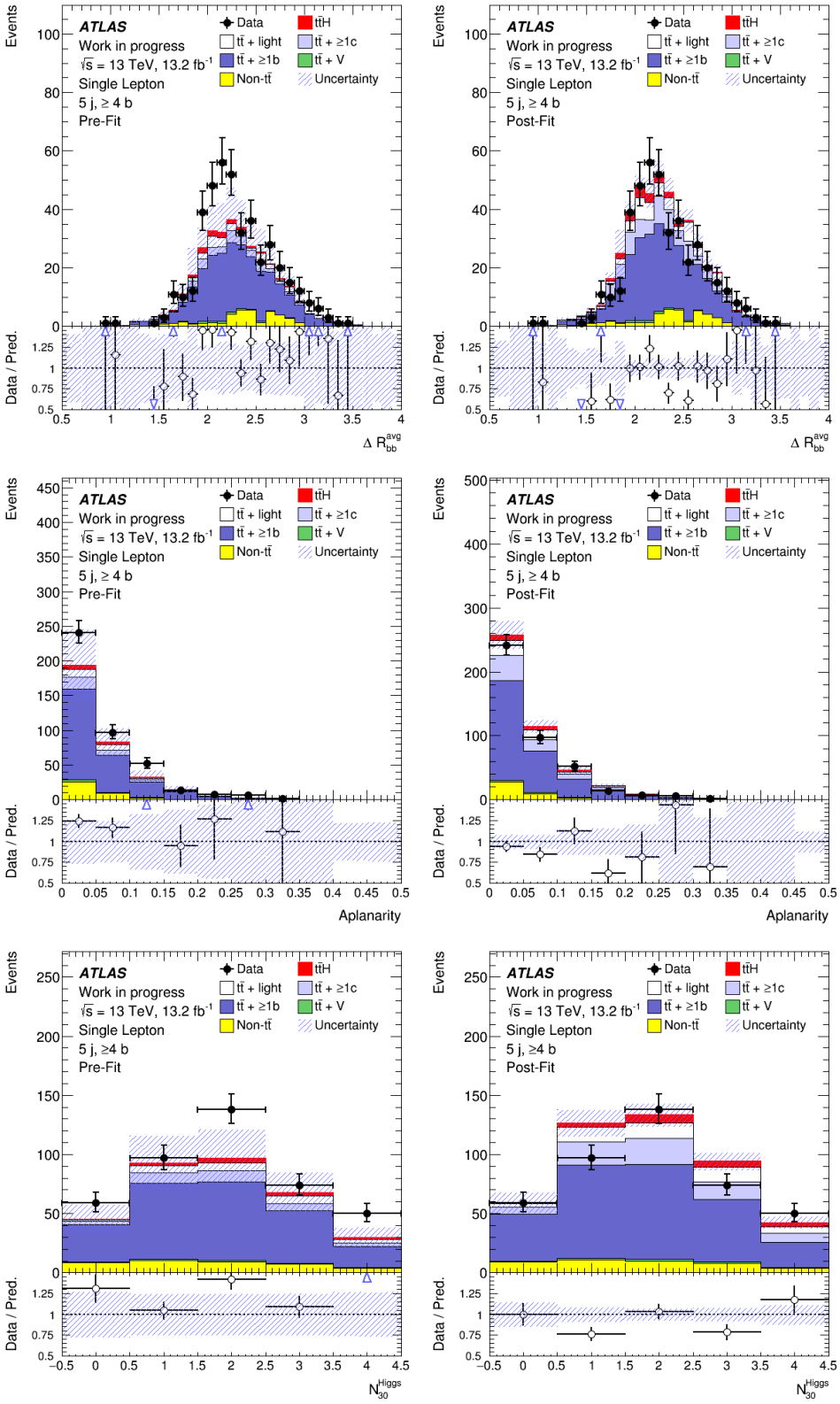


Figure B.3.: Distributions of the discriminating variables before and after the fitting procedure in the 5 jets, $\geq$4 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

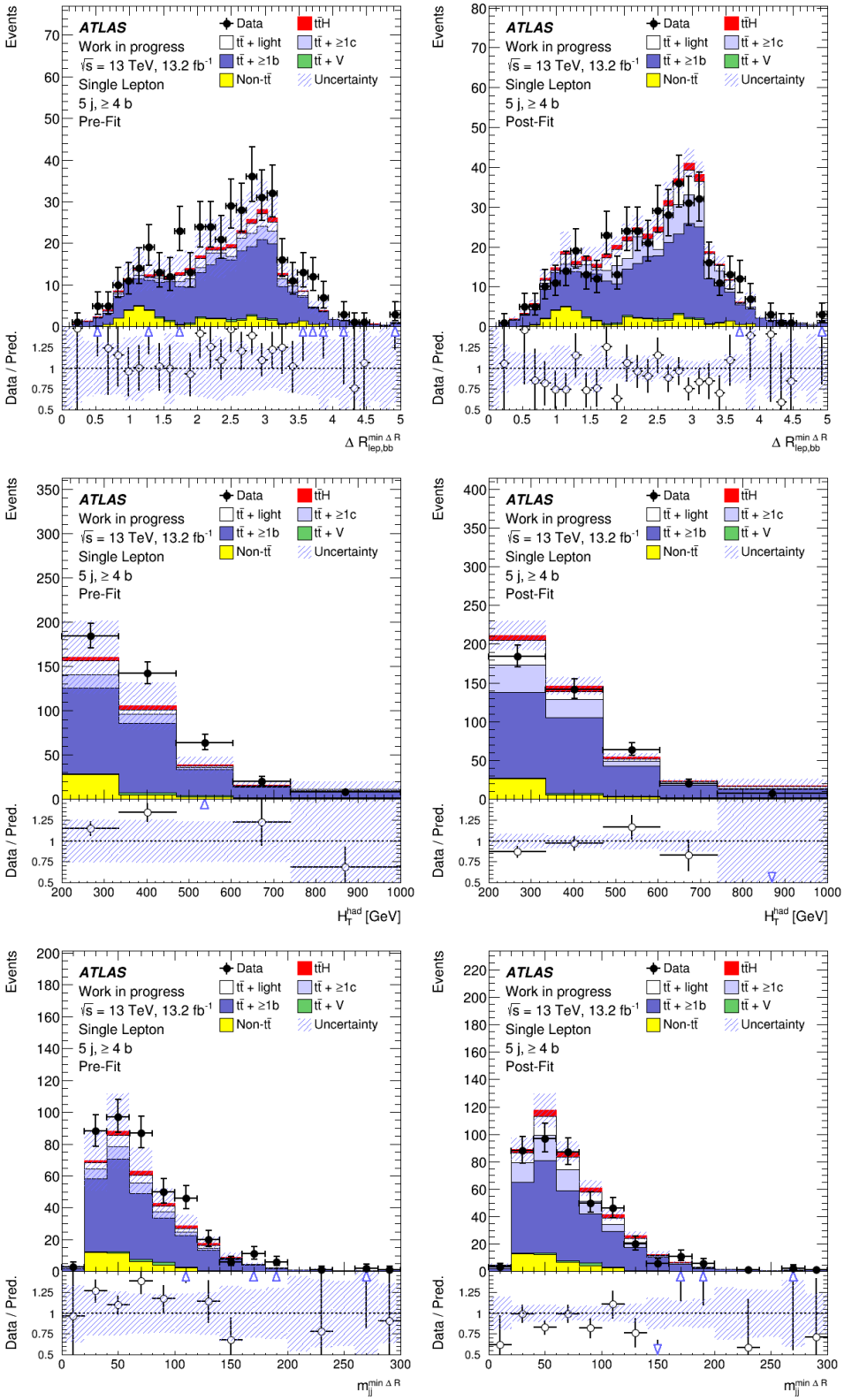


Figure B.4.: Distributions of the discriminating variables before and after the fitting procedure in the 5 jets, ≥4 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

Variables using information from the MVA reconstruction

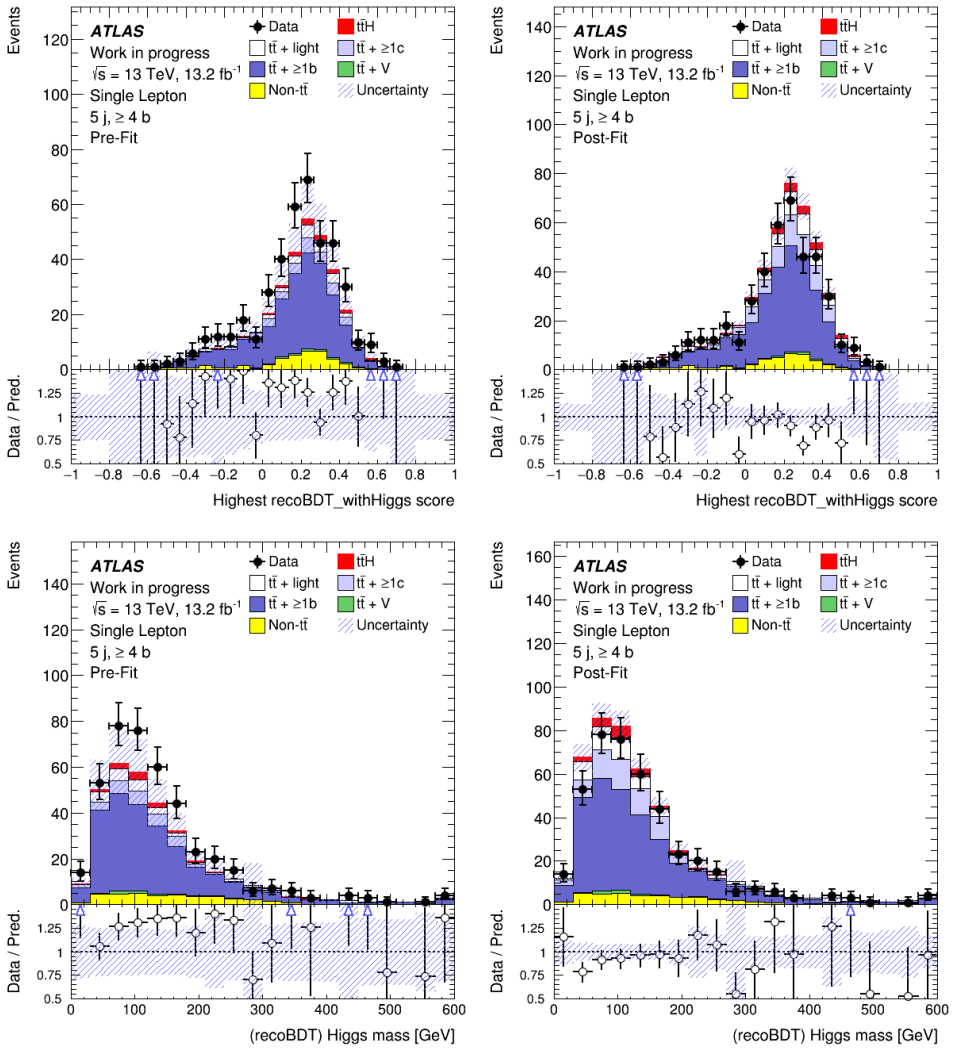


Figure B.5.: Distributions of the discriminating variables before and after the fitting procedure in the 5 jets, $\geq$4 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

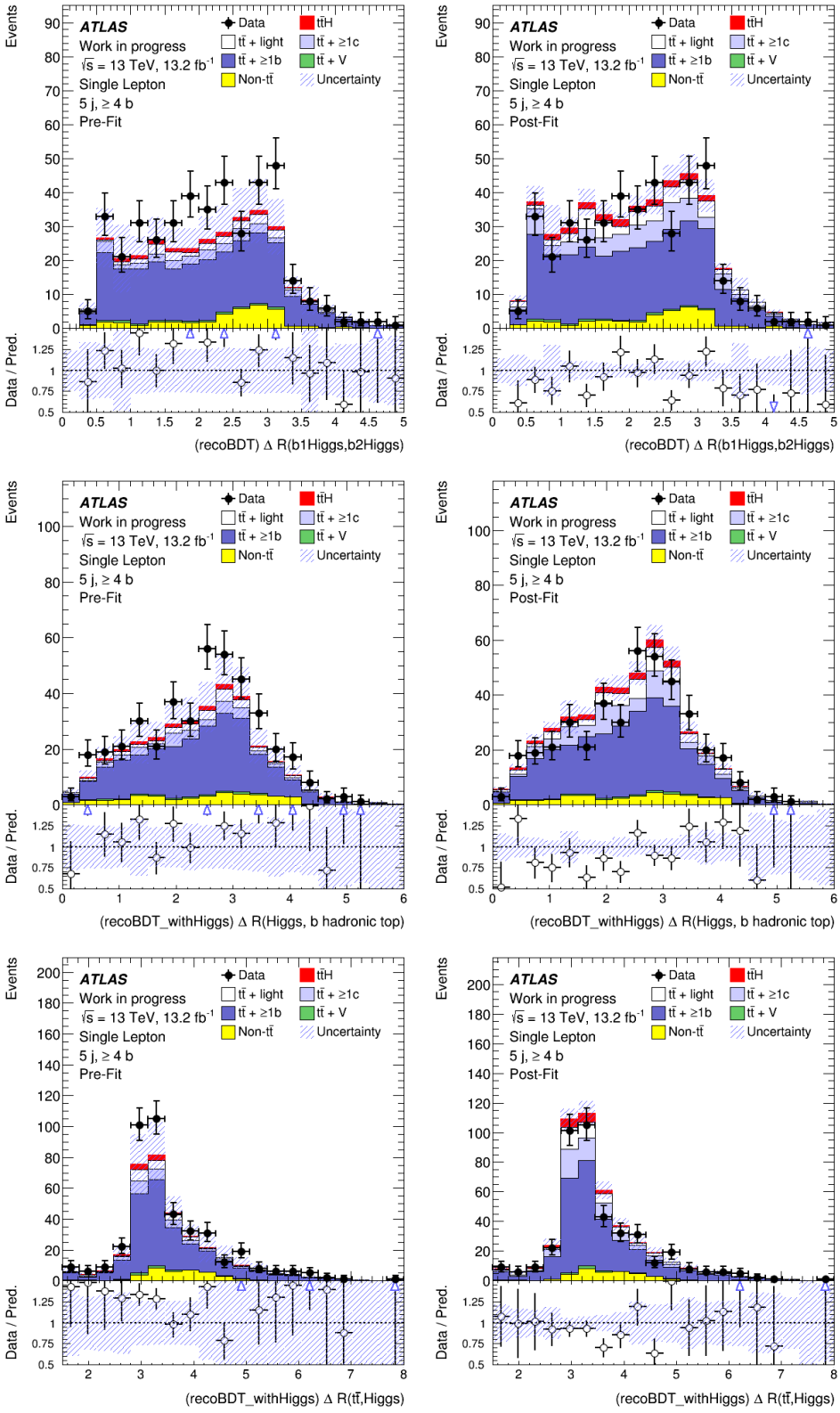


Figure B.6.: Distributions of the discriminating variables before and after the fitting procedure in the 5 jets, $\geq$4 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

B.1.2. Region: $\geq$6 jets, 3 b-tags

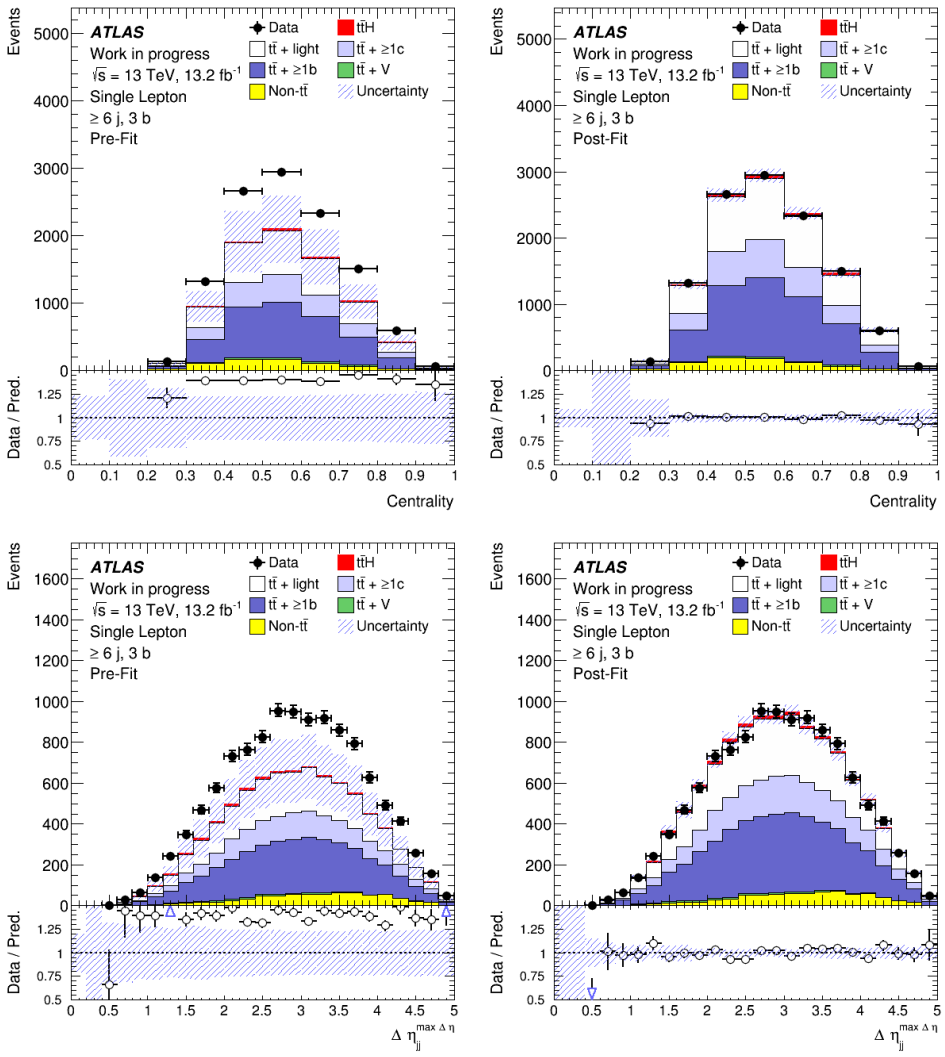


Figure B.7.: Distributions of the discriminating variables before and after the fitting procedure in the $\geq$6 jets, 3 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

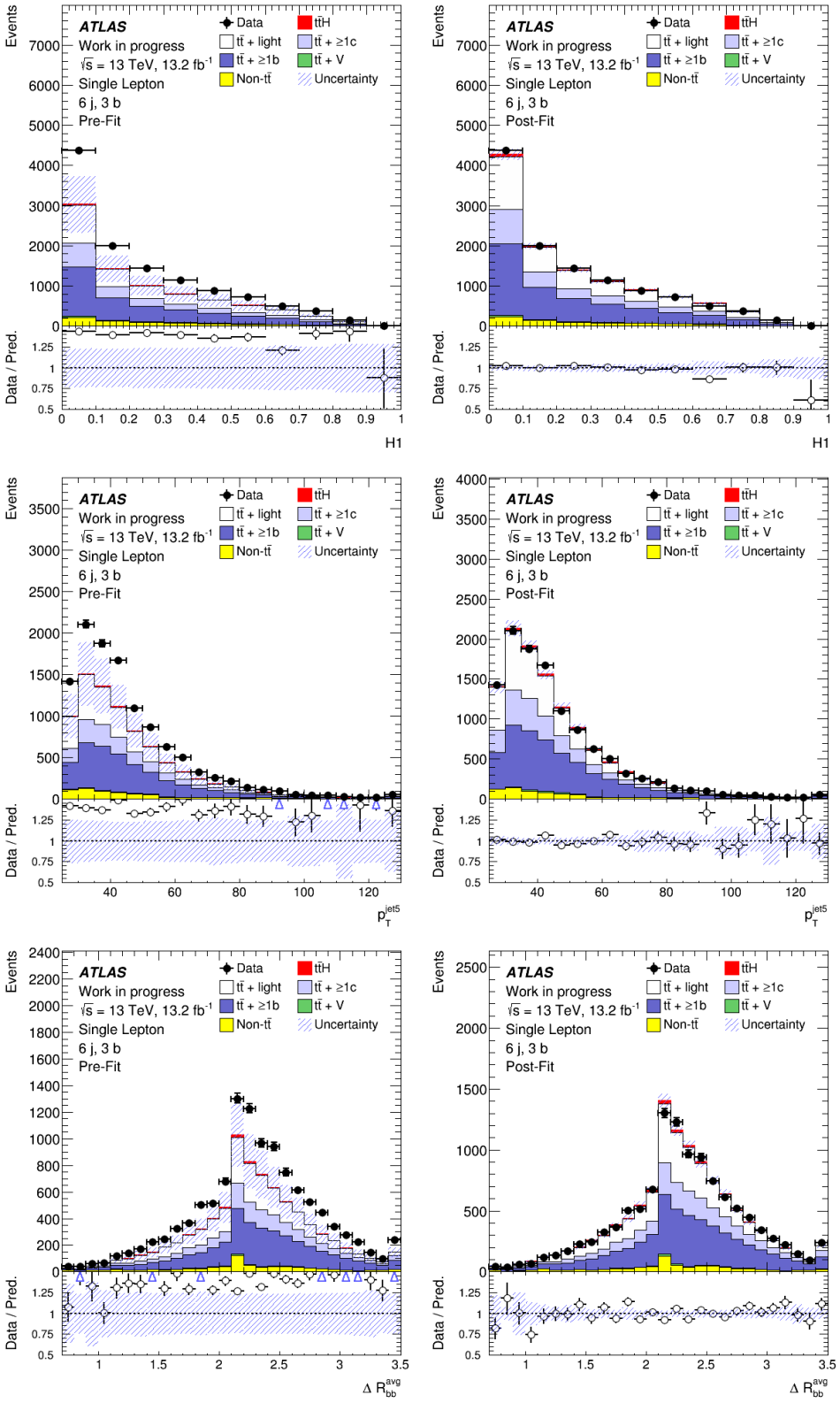


Figure B.8.: Distributions of the discriminating variables before and after the fitting procedure in the ≥6 jets, 3 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

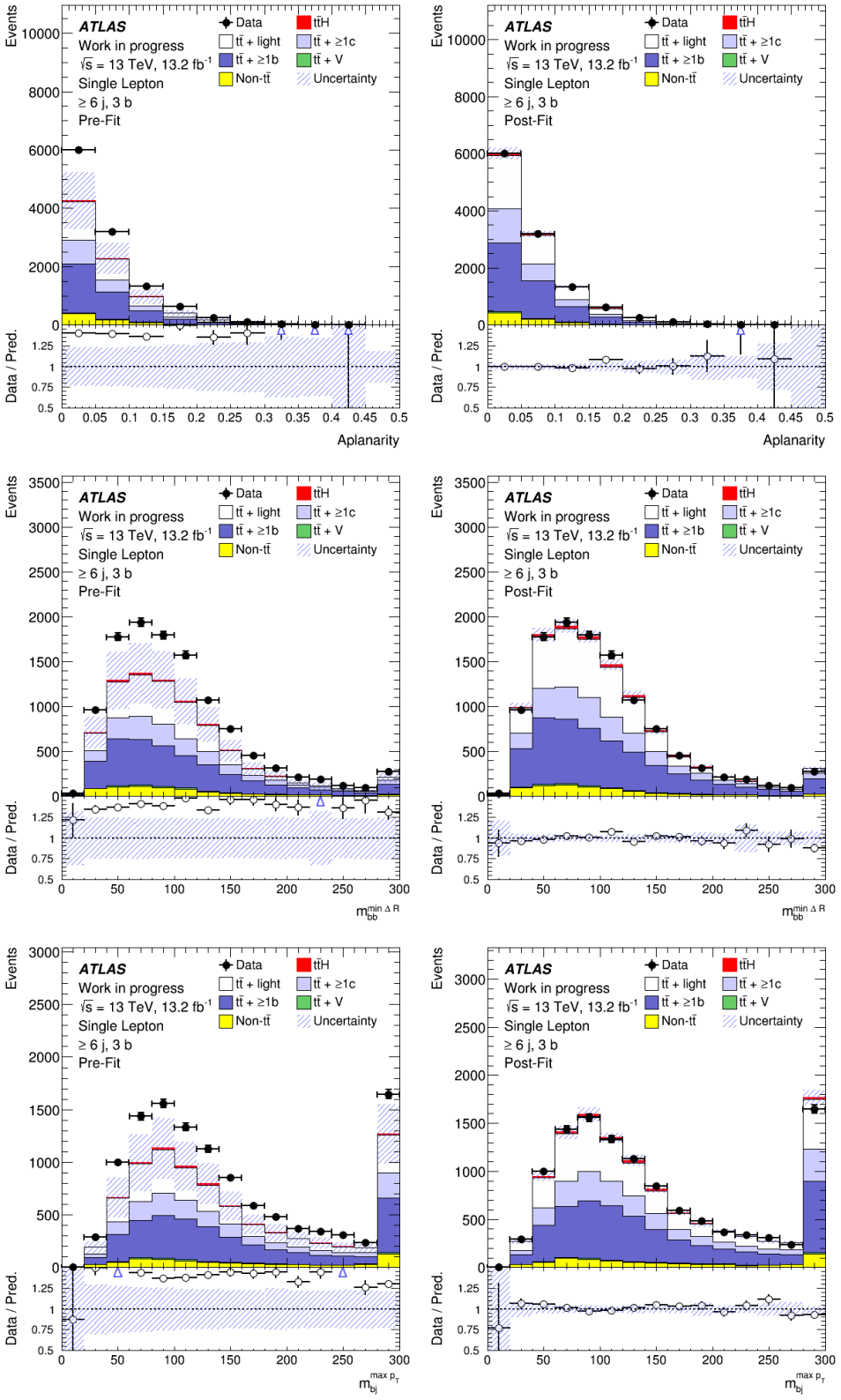


Figure B.9.: Distributions of the discriminating variables before and after the fitting procedure in the ≥6 jets, 3 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

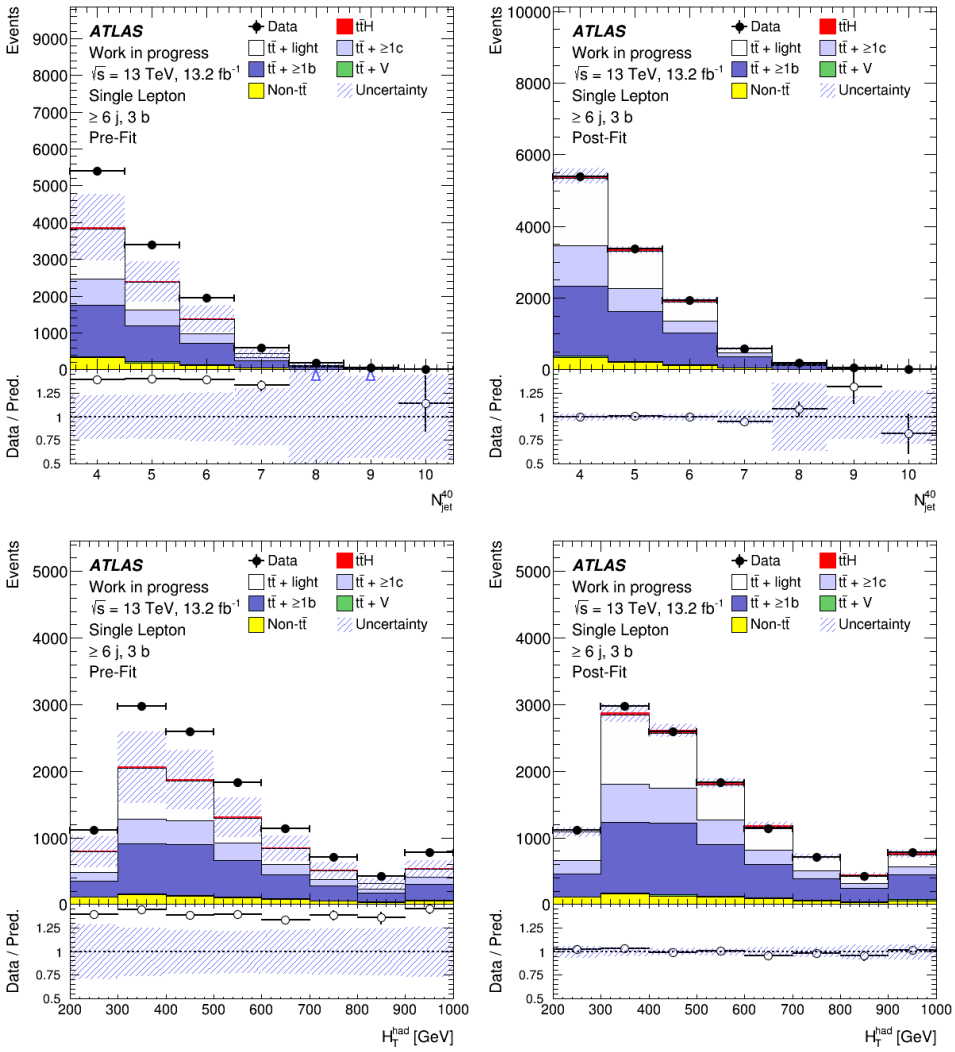


Figure B.10.: Distributions of the discriminating variables before and after the fitting procedure in the ≥6 jets, 3 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

Variables using information from the MVA reconstruction

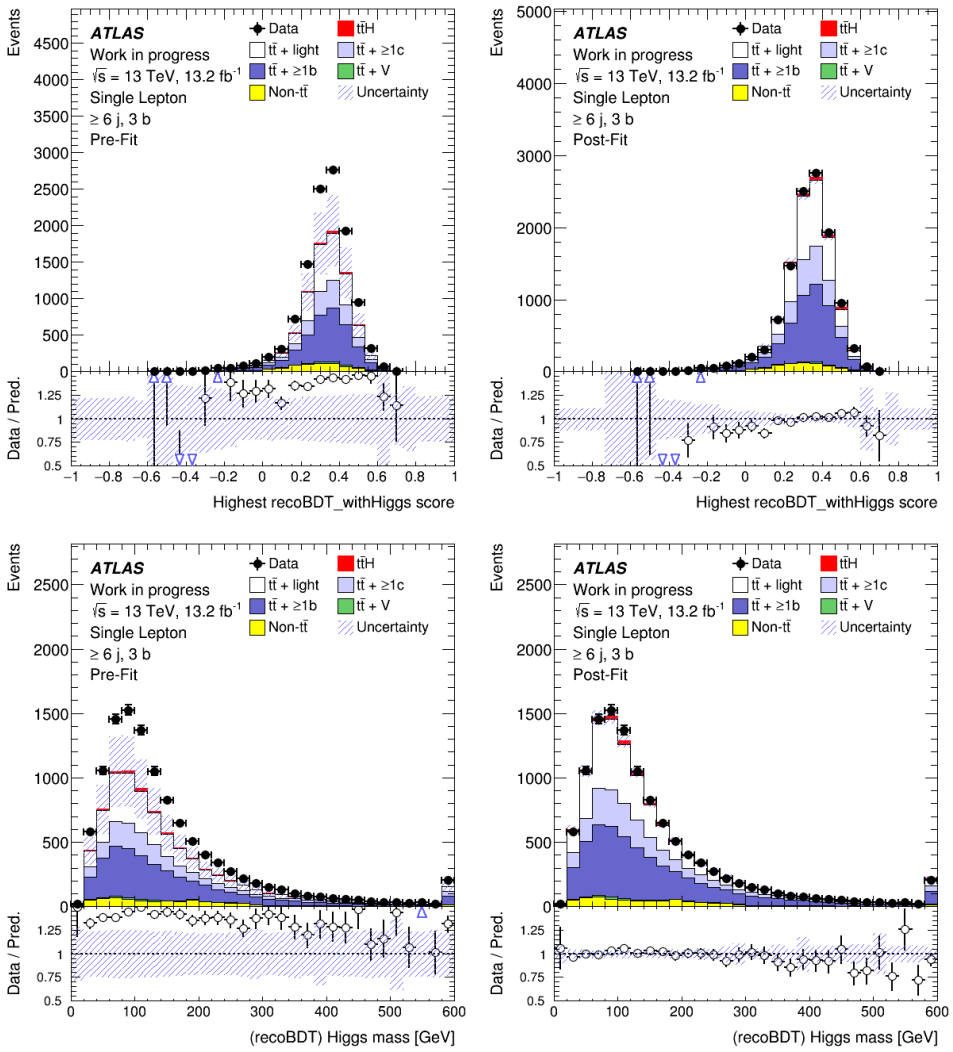


Figure B.11.: Distributions of the discriminating variables before and after the fitting procedure in the ≥6 jets, 3 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

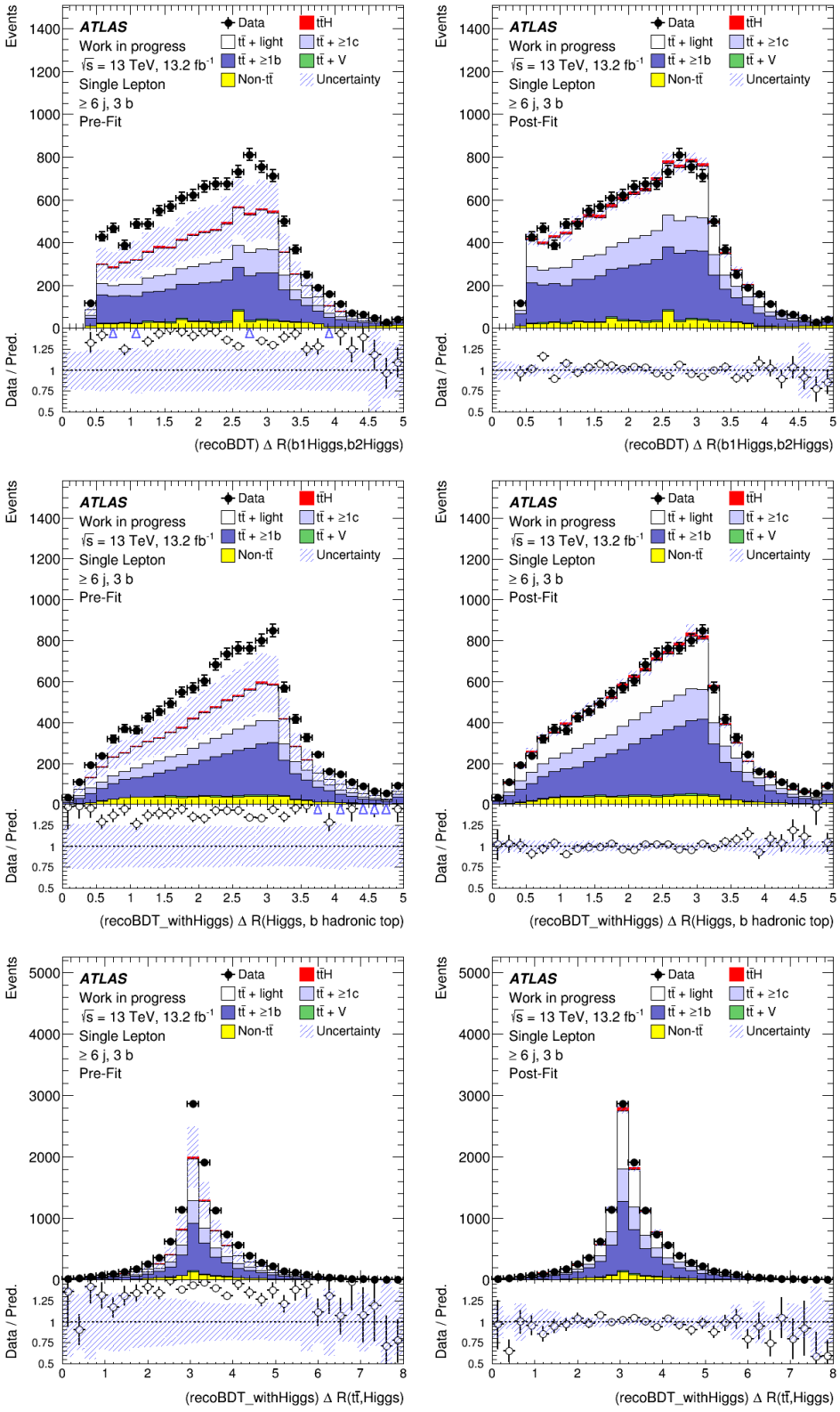


Figure B.12.: Distributions of the discriminating variables before and after the fitting procedure in the ≥6 jets, 3 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

B.1.3. Region: $\geq$6 jets, $\geq$4 b-tags

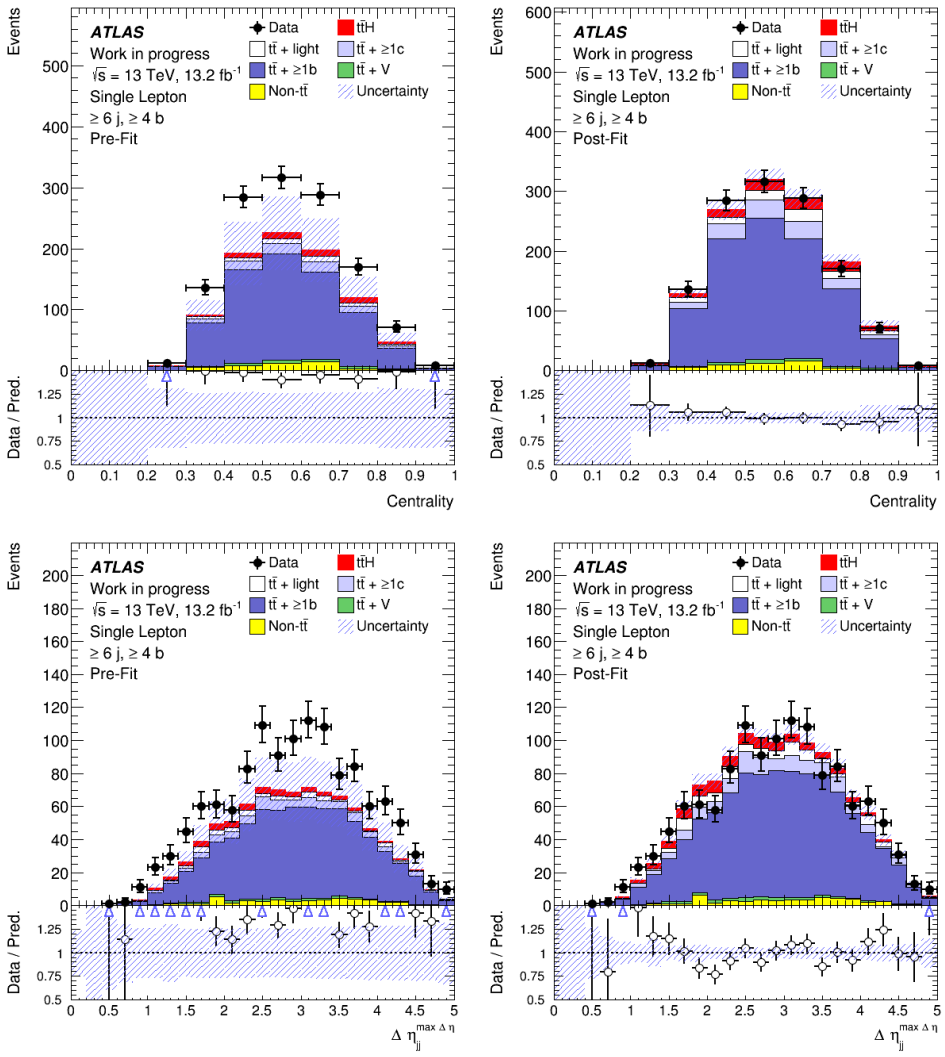


Figure B.13.: Distributions of the discriminating variables before and after the fitting procedure in the $\geq$6 jets, $\geq$4 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

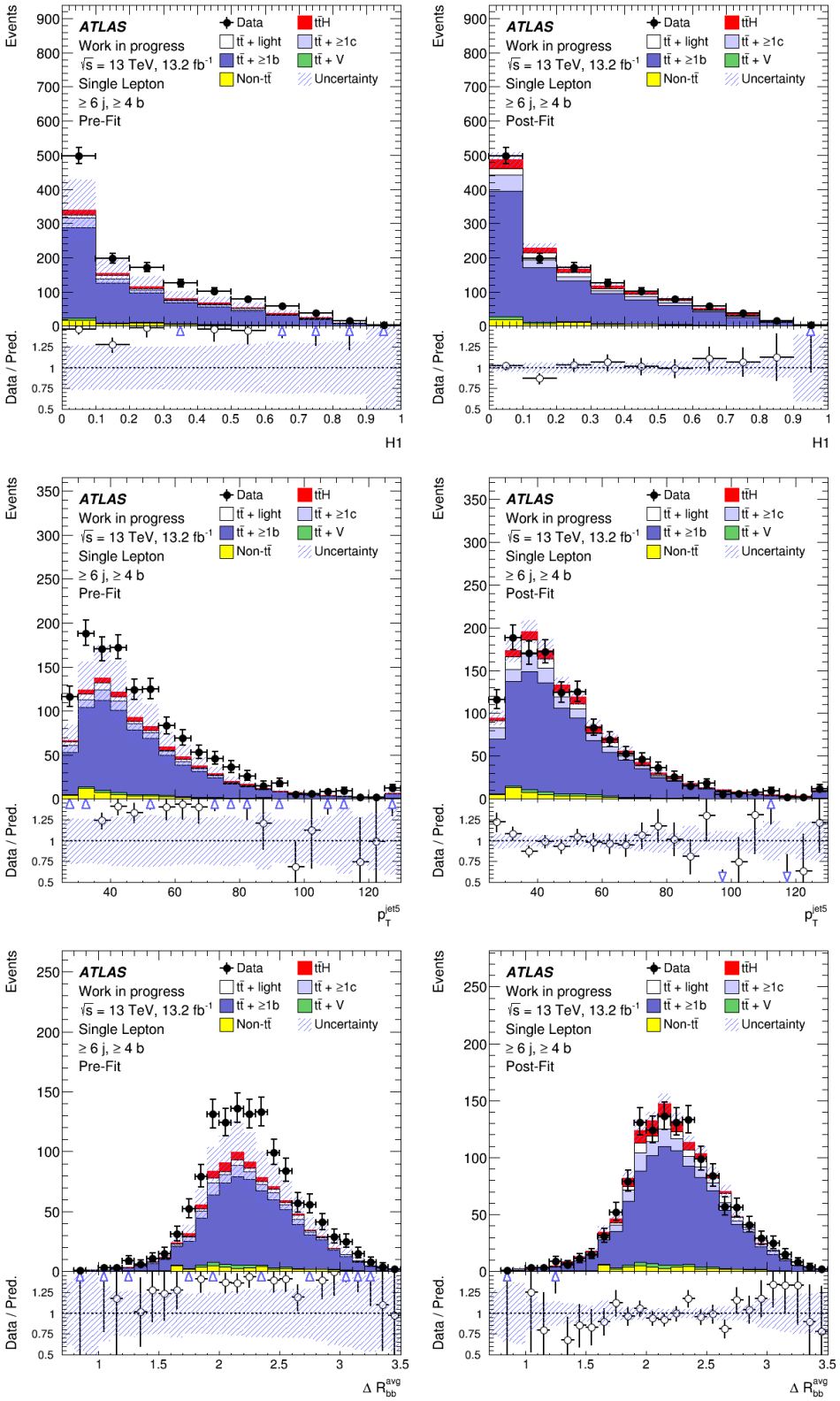


Figure B.14.: Distributions of the discriminating variables before and after the fitting procedure in the ≥6 jets, ≥4 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

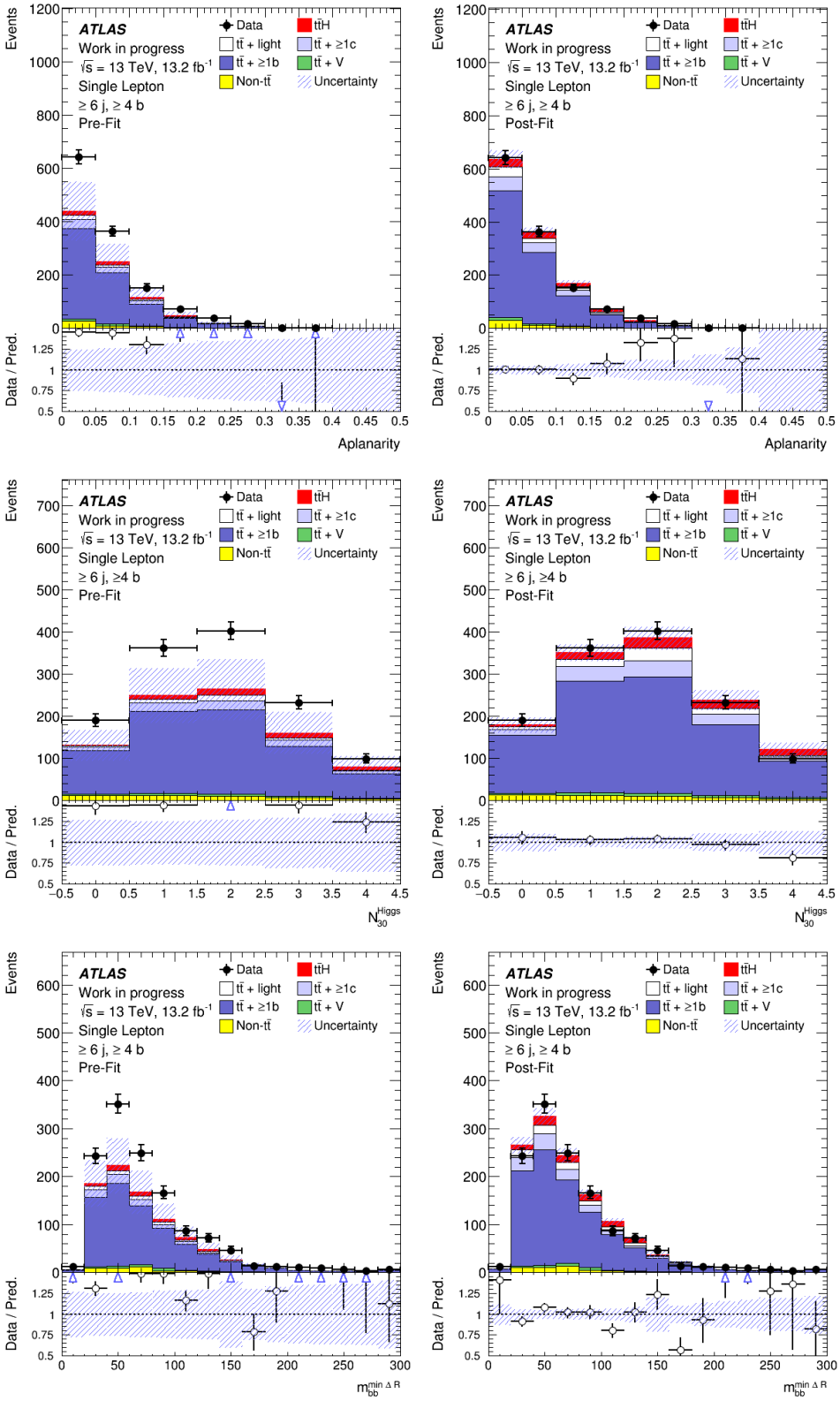


Figure B.15.: Distributions of the discriminating variables before and after the fitting procedure in the ≥6 jets, ≥4 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

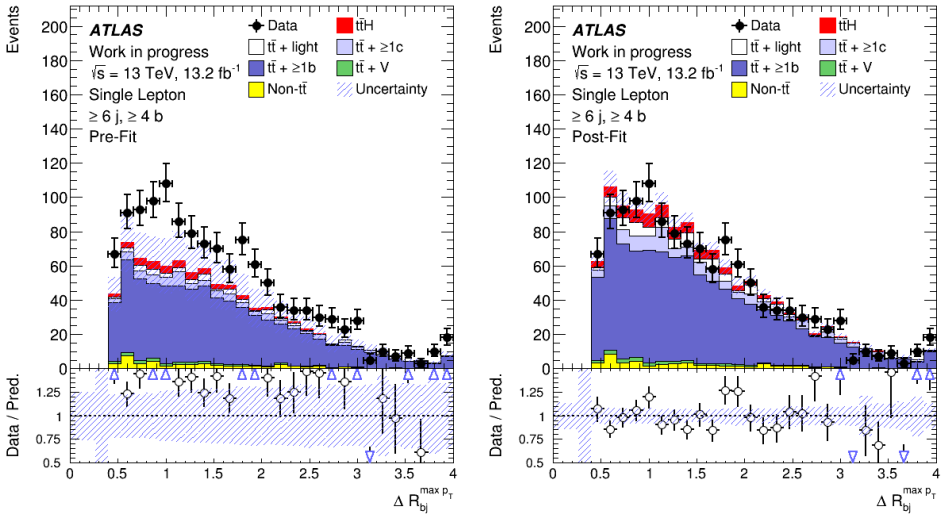


Figure B.16.: Distributions of the discriminating variables before and after the fitting procedure in the $\geq$6 jets, $\geq$4 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

Variables using information from the MVA reconstruction

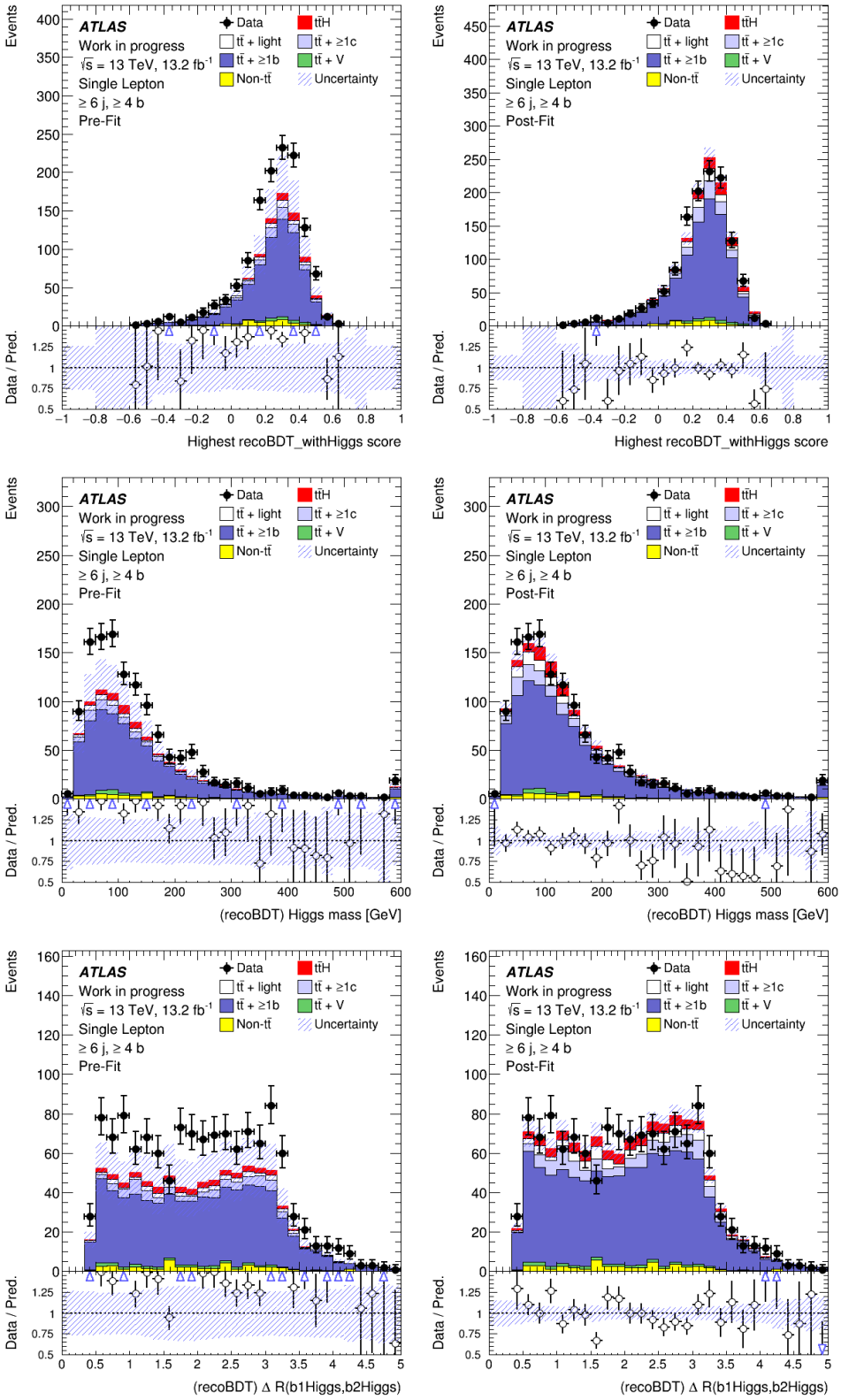


Figure B.17.: Distributions of the discriminating variables before and after the fitting procedure in the ≥6 jets, ≥4 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.

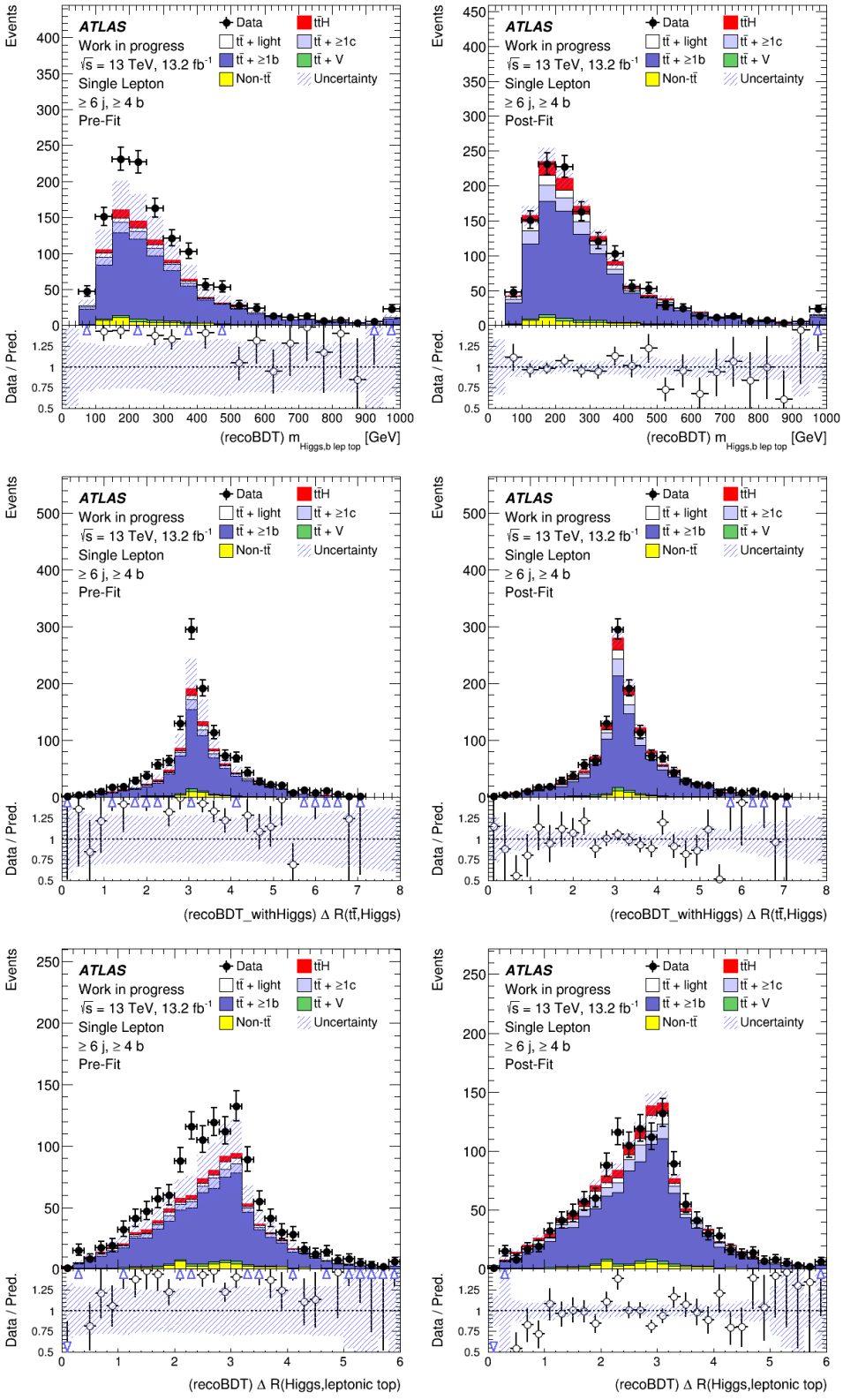


Figure B.18.: Distributions of the discriminating variables before and after the fitting procedure in the ≥6 jets, ≥4 b-tags region. The uncertainty band contains the statistical and systematic contribution. The pre-fit distributions do not include an uncertainty on the normalisation of $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$.