Search for the Higgs boson decaying to a pair of bottom quarks and produced in association with a pair of top quarks in the LHC Run 2 data with the ATLAS detector using a likelihood technique

Yulia Rodina

CPPM-T-2017-06

Aix-Marseille Université
École Doctorale 352
Faculté des Sciences
Centre de Physique des Particules de Marseille

Universidad Autónoma de Barcelona
Programa de Doctorado en Física
Departamento de Física
Facultad de Ciencias
Instituto de Física de Altas Energías

Thèse présentée pour obtenir le grade universitaire de docteur

Discipline :
PHYSIQUE ET SCIENCES DE LA MATIÈRE

Spécialité :
Physique des Particules et Astroparticules

Yulia Rodina

Search for the Higgs boson decaying to a pair of bottom quarks and produced in association with a pair of top quarks in the LHC Run 2 data with the ATLAS detector using a likelihood technique

Thesis director in AMU: Laurent Vacavant
Thesis director in UAB: Aurelio Juste Rozas
Thesis tutor in UAB: Maria Pilar Casado Lechuga

Soutenue le 23 Novembre 2017 devant le jury composé de:

Henri	Bachacou	Rapporteur
Florencia	Canelli	Rapporteur
Daniel	Bloch	Examinateur
Stefan	Guindon	Examinateur
Aurelio	Juste Rozas	Directeur de thèse
Laurent	Vacavant	Directeur de thèse

Contents

Abstract

Following the discovery of the Higgs boson by the ATLAS and CMS collaborations at the Large Hadron Collider (LHC) in 2012, the attention has been focussed on studying the properties of the newly discovered particle to test the predictions of the Standard Model (SM). An object of particular interest is the top quark Yukawa coupling - the coupling of the Higgs boson to the top quark, which is predicted to be close to unity in the SM and at the same time very sensitive to the possible effects of new physics beyond the SM. The production of the Higgs boson in association with a pair of top quarks, $t\bar{t}H$, is the process that gives direct access to the top quark Yukawa coupling. The decay of the Higgs boson into a pair of b-quarks, $H \to b\bar{b}$, is dominant in the SM for a value of the Higgs boson mass of 125 GeV (its branching ratio is approximately 58%). This decay channel also allows measuring the b-quark Yukawa coupling, the second largest coupling of the Higgs boson to a fermion in the SM.

In this dissertation the search for $t\bar{t}H$ ($H \to b\bar{b}$) in the single-lepton channel, resulting from the semileptonic decay of the $t\bar{t}$ system, is presented. The analysis is based on 36.1 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV recorded with the ATLAS detector in 2015 and 2016. The study is performed using a likelihood-based method that exploits kinematic properties of the selected events to separate the signal from the background, which is dominated by $t\bar{t}$ produced in association with additional jets. This search relies on the high multiplicity of jets originating from b-quarks (b-jets), so identification of these jets (b-tagging) is crucial. A study on the optimisation of b-jet identification algorithms in ATLAS is also presented in this dissertation. The ratio of the measured $t\bar{t}H$ cross-section to the SM expectation is found to be $\mu = 0.84^{+0.64}_{-0.61}$, assuming a Higgs boson mass of 125 GeV. This result is consistent with both the background-only hypothesis and the $t\bar{t}H$ SM prediction.

Résumé

Suite à la découverte du boson de Higgs au Large Hadron Collider (LHC) par les collaborations ATLAS et CMS en 2012, l'attention s'est portée sur l'étude des propriétés de cette nouvelle particule pour tester les prédictions du modèle standard (MS). Un objet d'intérêt particulier est le couplage de Yukawa au quark top - le couplage du boson de Higgs au quark top - qui devrait être proche de l'unité dans le MS et en même temps très sensible aux effets possibles de nouvelle physique au-delà du MS. La production du boson de Higgs en association avec une paire de quarks top, $t\bar{t}H$, est le canal qui donne un accès direct au couplage de Yukawa au quark top.

La désintégration du boson de Higgs en une paire de quarks b, $H \to b\bar{b}$, domine dans le MS pour la valeur de la mass du boson de Higgs de $m_H = 125$ GeV: son rapport de branchement est d'environ 58%. Ce canal de désintégration permet également de mesurer le couplage de Yukawa au quark b - le deuxième plus grand couplage du boson de Higgs à un fermion dans le MS.

Dans cette thèse, la recherche de $t\bar{t}H$ ($H \to b\bar{b}$) dans le canal à un lepton, résultant de la désintégration semi-leptonique du système $t\bar{t}$, est présentée. L'analyse est basée sur 36.1 fb^{-1} de collisions pp à $\sqrt{s} = 13$ TeV enregistrées avec le détecteur ATLAS en 2015 et 2016. L'étude est réalisée avec une méthode de vraisemblance, qui exploite les propriétés cinématiques des événements sélectionnés pour séparer le signal du bruit de fond, qui est dominé par les paires de quarks $t\bar{t}$ produites en association avec des jets supplémentaires. Cette recherche repose sur une grande multiplicité de jets issus de quarks b (jets b). Pour cette raison l'identification de ces jets (b-tagging) est cruciale. Une étude sur l'optimisation des algorithmes d'identification des jets b dans ATLAS est également présentée dans cette dissertation.

Le rapport de la section efficace mesurée de $t\bar{t}H$ à la prédiction de MS est $\mu = 0.84^{+0.64}_{-0.61}$, en supposant une masse du boson de Higgs de 125 GeV. Ce résultat est cohérent avec l'hypothèse de fond seulement ainsi qu'avec la prédiction du MS pour le signal $t\bar{t}H$.

Resumen

Tras el descubrimiento del bosón de Higgs por las colaboraciones ATLAS y CMS en el Gran Colisionador de Hadrones (LHC, por sus siglas en inglés) en 2012, la atención se ha centrado en estudiar las propiedades de la partícula recientemente descubierta para probar las predicciones del Modelo Estándar (SM, por sus siglas en inglés). Un objeto de particular interés es el acoplamiento de Yukawa del quark top - el acoplamiento del bosón de Higgs al quark top, que se prevé que tenga un valor cercano a la unidad en el SM y al mismo tiempo es muy sensible a los posibles efectos de nueva física más allá del SM. La producción del bosón de Higgs en asociación con una pareja de quarks top, $t\bar{t}H$, es el modo que permite medir directamente al acoplamiento de Yukawa del quark top. La desintegración del bosón de Higgs en una pareja de quarks b, $H \to b\bar{b}$, es dominante en el SM para un valor de la masa del bosón de Higgs de $m_H = 125$ GeV (ocurre aproximadamente el 58% de las veces). Este canal de desintegración también permite medir el acoplamiento de Yukawa del quark b - el segundo mayor acoplamiento del bosón de Higgs a un fermión en el SM.

En esta tesis se presenta la búsqueda del proceso $t\bar{t}H$ ($H \to b\bar{b}$) en sucesos con un sólo leptón en el estado final, resultante de la desintegración semileptónica del sistema $t\bar{t}$. El análisis se basa en 36.1 fb^{-1} de datos de colisiones protón-protón a una energía del centro de masas de $\sqrt{s} = 13$ TeV registrados con el detector ATLAS en 2015 y 2016. El estudio se realiza utilizando un método basado en verosimilitud que explora las propiedades cinemáticas de los eventos seleccionados para separar la señal del fondo, que está dominado por $t\bar{t}$ producido en asociación con chorros hadrónicos (jets) adicionales. Esta búsqueda explota la alta multiplicidad de jets originados a partir de quarks b (b-jets), por lo que la identificación de los mismos es crucial. En esta tesis también se presenta un estudio sobre la optimización de los algoritmos de identificación de b-jets en ATLAS. La razón entre la sección eficaz de $t\bar{t}H$ medida y la correspondiente predicción del SM es $\mu = 0.84^{+0.64}_{-0.61}$, asumiendo un bosón de Higgs con una masa de 125 GeV. Este resultado es consistente ambos con la hipótesis de sólo background así como con la predicción del SM incluyendo el proceso $t\bar{t}H$.

Introduction

In ancient times people were searching for answers to fundamental questions such as: "What is the world surrounding us made of?", "What is matter?".

Modern particle physics has found answers to some of these questions, but also has added new ones to the list: "How do elementary particles interact with each other?", "What is common between different physical interactions?", "Why there is more matter than antimatter in the universe?" or "What is the origin of mass?".

A theory that provides a coherent, but not yet fully complete, picture of elementary particles and the interactions among them is the Standard Model (SM). It gives a unified description of three of the four known fundamental forces. Many theoretical predictions of the SM have been verified experimentally with a remarkable accuracy since the 1960s, when the model was established.

One of the fundamental problems raised and solved in the SM is the origin of the mass of the elementary particles. A priori the elementary particles described by the theory are expected to be massless, in contradiction with the observation. Therefore a mechanism that allows particles to acquire their mass was introduced to provide agreement with experimental evidence. This mechanism assumes the existence of a quantum scalar field, whose excitations manifest themselves as a new physical particle called the Higgs boson. The SM predicts some properties of the Higgs boson, but its mass is a free parameter of the theory and can only be obtained from experiment. The search for this particle has been one of the main goals of the Large Hadron Collider (LHC), the world's biggest particle accelerator, built at CERN. The discovery of the Higgs boson in 2012 by the ATLAS and CMS collaborations was a triumph of the SM: the last particle predicted by this theory had finally been found.

One of possible modes for the Higgs boson production at the LHC is the production in association with top-quark pairs ($t\bar{t}H$). This production channel has one of the smallest cross sections at the LHC. At the same time it is of particular physical interest: the coupling of the Higgs boson to top quarks, that can be directly measured in this channel, is an important property of the SM. If the measured value of this parameter is significantly different from unity predicted by the SM, this would be an indication for a new physics beyond the SM. Therefore observing the Higgs boson production in association with top quarks is now one of the most important physics goals of the LHC.

In this dissertation the 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 registered with the ATLAS detector at the LHC in 2015 and 2016, is presented. This analysis is focussed on the semileptonic decay of the $t\bar{t}$ system, resulting in a final state with a single lepton and many jets. This dissertation describes in detail my main contribution, i.e. the development and optimisation of a likelihood-based method to distinguish the signal ($t\bar{t}H$) from the background (dominated by $t\bar{t}$ produced in association with additional jets). Particular kinematic features of both signal and background events are exploited in the method.

Information on the multiplicity of the jets originating from b-quarks (b-jets) is important in the $t\bar{t}H$ ($H \to b\bar{b}$) search, as there are four b-jets in the final state. Therefore the identification of these jets, known as b-tagging, plays a key role. My contribution to the

optimisation of the ATLAS b-tagging algorithms for LHC Run 2 is also presented in this dissertation.

This document is organized as follows. Chapter 1 contains a theoretical overview of the Standard Model and the physics of the Higgs boson at hadron colliders. Chapter 2 introduces the LHC and the ATLAS detector and describes the reconstruction of the various physical objects out of the signals recorded by the detector. Chapter 3 presents the b-tagging algorithms developed in ATLAS and their optimisation for LHC Run 2. My contribution to the optimisation of the algorithms relying on the track impact parameter (IP2D, IP3D) is presented. Chapter 4 is an overview of the $t\bar{t}H$ ($H \to b\bar{b}$) analysis and shows the results obtained, including a detailed description of my main contribution - the likelihood discriminant method.

1 Theoretical background

1.1 The Standard Model

The Standard Model (SM) [1–3] of particle physics describes elementary particles and their interactions via three of the four known fundamental physical forces (gravity is not included). The SM was developed in the 1960s and since then it has been successfully tested in many experiments. The observation of the Higgs boson at the LHC in 2012 represents a triumph of the SM, with its last missing ingredient being discovered.

1.1.1 Elementary particles

According to the SM, there are two types of elementary particles: fermions and bosons. Matter is composed of fermions that interact through the exchange of bosons, which mediate the forces: electromagnetic, strong and weak.

Fermions are classified into quarks and leptons, both categorised in three generations with a mass hierarchy (the mass increasing from lighter particles in the first generation to heavier in the third). Quarks carry an attribute denoted "colour" (red, green, blue) and participate in electromagnetic, weak and strong interactions. Quarks can be observed only in bound states, making composite particles (hadrons). The quarks can be classified into three generations: up (u) and down (d) quarks in the first generation, charm (c) and strange (s) quarks in the second generation and top (t) and bottom (b) quarks in the third generation. Leptons participate in electromagnetic and weak interactions and do not participate in strong interactions, so they do not form bound states. They are also classified into three generations. The charged leptons are the electron (e), muon (μ) and tau-lepton (τ), while neutral leptons are the neutrinos, one associated to each charged lepton generation: ν_e, ν_μ and ν_τ. In addition for each quark and lepton an antiparticle with the same mass, but opposite charge and opposite other quantum numbers, exists. The SM fermions with the values of their mass and charge are presented in table 1.

	Quarks			Leptons		
Generation	Flavour	Mass	Charge (e)	Flavour	Mass	Charge (e)
1	u	2.2 MeV	2/3	e	0.511 MeV	-1
	d	4.7 MeV	-1/3	ν_e	< 2 eV	0
2	c	1.28 GeV	2/3	μ	105.7 MeV	-1
	s	96 MeV	-1/3	ν_μ	< 0.19 MeV	0
3	t	173.1 GeV	2/3	τ	1776.9 MeV	-1
	b	4.18 GeV	-1/3	$\nu\tau$	< 18.2 MeV	0

Table 1: Quarks and leptons with the values of their mass and electric charge. From Ref. [4].

The ordinary matter is built of the first generation particles: u and d, that are constituents of protons and neutrons, and electrons. Particles from the second and third generations can be observed only in cosmic rays and high energy physics experiments.

Gauge bosons are responsible for interactions between particles. Photons (γ) mediate electromagnetic interactions, whereas eight gluons (g) mediate strong interactions. Both of them are massless. The carriers of the weak interaction are extremely massive: two electrically-charged $W^{\pm}$ bosons and neutral Z boson. The properties of the gauge bosons are summarized in table 2.

Boson	Interaction	Mass	Charge (e)
g	Strong	0	0
γ	Electromagnetic	0	0
$W^{\pm}$	Weak	80.39 GeV	± 1
Z		91.19 GeV	0

Table 2: Gauge bosons with the type of interaction they mediate and the values of their mass and electric charge. From Ref. [4].

1.1.2 The Standard Model formalism

The SM is based on a renormalisable relativistic quantum field theory. The gauge symmetry group of the SM is

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

where

- $SU(2)_L \otimes U(1)_Y$ is the symmetry group of the electroweak interaction, according to the unified electroweak (EW) theory developed by Glashow, Salam and Weinberg [1–3]. $U(1)_Y$ is an abelian group that introduces a new conserved quantum number, the hypercharge Y. $SU(2)_L$ is a non-abelian group that describes the weak interaction, with weak isospin $\vec{I}$ as conserved quantity. The electric charge Q is related to the third component of the weak isospin I_3 and the hypercharge Y by the Gell-Mann Nishijima formula:

$$Q = I_3 + \frac{Y}{2}. \tag{2}$$

- $SU(3)_C$ is a non-abelian group that describes the strong interaction. The colour (C) is the conserved charge for this group. The theory of the strong force is described by Quantum Chromodynamics (QCD) [5–9].

The SM Lagrangian can be divided in two terms, one describing the electroweak interaction and another describing the strong interaction:

$$\mathcal{L}_{\mathrm{SM}} = \mathcal{L}_{\mathrm{EW}} + \mathcal{L}_{\mathrm{QCD}}. \tag{3}$$

1.1.3 The electroweak theory

The starting point for constructing the part of SM Lagrangian that describes electromagnetic interactions, is considering two terms: one corresponding to the fermions and another one related to the gauge bosons.

The fermions are represented as Dirac fields composed of left-handed and right-handed components, defined as:

$$\begin{aligned} \psi_L &= \frac{1}{2}(1-\gamma^5)\psi, \\ \psi_R &= \frac{1}{2}(1+\gamma^5)\psi. \end{aligned} \tag{4}$$

The left-handed fermions form weak-isospin doublets:

$$Q_L^i = \begin{pmatrix} u^i \\ d^i \end{pmatrix}_L, \; L_L^i = \begin{pmatrix} \nu^i \\ l^i \end{pmatrix}_L, \tag{5}$$

whereas the right-handed fermions are represented as weak-isospin singlets:

$$u_R^i, d_R^i, l_R^i, \tag{6}$$

where $i = 1, 2, 3$ denotes the generation number. The right-handed fermions do not participate in weak interactions. Therefore, the right-handed neutrinos would not participate in any interaction, and thus they are not considered in the SM. For this reason the SM neutrinos are treated as massless particles.

The term of the Lagrangian describing fermions is given by

$$\mathcal{L}_{\mathrm{fermion}} = \sum_{f=l,q} \bar{f} i\gamma^\mu D_\mu f, \tag{7}$$

where f is the fermion field and the covariant derivative is defined as

$$D_\mu = \partial_\mu - ig\vec{I}\cdot\vec{W}_\mu - ig'\frac{Y}{2}B_\mu. \tag{8}$$

Another term of the Lagrangian describes gauge bosons:

$$\mathcal{L}_{\mathrm{gauge}} = -\frac{1}{4}F^i_{\mu\nu}F^{\mu\nu i} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}, \tag{9}$$

where the field tensors are defined as

$$\begin{aligned} F^i_{\mu\nu} &= \partial_\mu W^i_\nu - \partial_\nu W^i_\mu - g\epsilon_{ijk}W^j_\mu W^k_\nu, \\ B_{\mu\nu} &= \partial_\mu B_\nu - \partial_\nu B_\mu, \end{aligned} \tag{10}$$

g and g' are the gauge couplings of the $SU(2)_L$ and $U(1)_Y$ groups, W^i_μ (where $i = 1, 2, 3$) and B_μ denote the gauge fields of these groups, and ϵ_{ijk} is the totally antisymmetric tensor. The B and W_3 fields mix, giving the photon and the Z boson.

1.1.4 The Brout-Englert-Higgs mechanism

The Lagrangian composed of the two terms described above, $\mathcal{L}_{\text{fermion}}$ and $\mathcal{L}_{\text{gauge}}$, is invariant under local gauge transformations only if assuming that particles are massless. Adding explicit mass terms for gauge bosons or fermions would break the local invariance: gauge symmetry for bosons and chiral symmetry in the case of fermions. But breaking gauge invariance would consequently break the renormalisability of the SM. However, it is known from experiment that the fermions and $W^\pm$ and Z bosons have a mass. This is solved via the Brout-Englert-Higgs mechanism [10–12]. An additional field, called the Higgs field, is introduced, that allows SM particles to acquire masses by interacting with it.

The Higgs field is a weak isospin doublet of one charged and one neutral complex scalar fields:

$$\phi(x) = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}. \tag{11}$$

The Lagrangian of this field is given by

$$\mathcal{L}_{\text{Higgs}} = (D_\mu \phi)^\dagger (D^\mu \phi) - V(\phi), \tag{12}$$

where the first term is kinetic, with the covariant derivative D_μ given by equation 8 and the Higgs potential $V(\phi)$, defined as

$$V(\phi) = \mu^2 \phi^\dagger \phi + \lambda (\phi^\dagger \phi)^2. \tag{13}$$

The first term of $V(\phi)$ can be interpreted as a mass and the second term represents the self-interaction of the field. The minimum of this potential is known as the vacuum expectation value of the Higgs field.

The Higgs potential depends on two parameters, μ and λ. To provide a stable potential minimum, λ is required to be positive. For the μ parameter there are two possibilities: $\mu^2 > 0$ and $\mu^2 < 0$, presented in figure 1. For $\mu^2 > 0$ the minimum of the potential $V(\phi)$ is at $\langle 0|\phi|0\rangle \equiv \phi_0 = 0$, so the $SU(2)_L \otimes U(1)_Y$ symmetry is respected. For the case of $\mu^2 < 0$ the minimum of potential is obtained at a non-zero value of ϕ:

$$\langle 0|\phi^2|0\rangle \equiv \phi_0^2 = -\frac{\mu^2}{2\lambda} = \frac{v^2}{2}. \tag{14}$$

In this case the vacuum state of the field is not invariant under the $SU(2)_L \otimes U(1)_Y$ symmetry. This effect is known as spontaneous symmetry breaking.

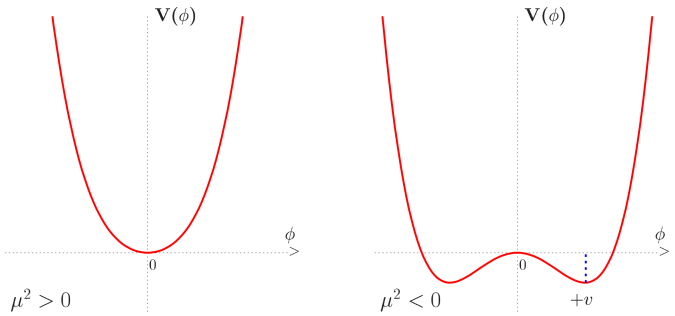


Figure 1: Higgs potential in the case $\mu^2 > 0$ and $\mu^2 < 0$, with $\lambda > 0$ in both cases. From Ref. [13].

To satisfy the requirement that the photon should be massless, the minimum of potential is chosen to be

$$\phi_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}. \tag{15}$$

The expression for the field ϕ can be rewritten as

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}, \tag{16}$$

where $h(x)$ describes small perturbations with respect to the vacuum state. This represents a physical field, associated with the Higgs boson.

Then the part for the Lagrangian corresponding to the Higgs field (see equation 12) develops into an expression with explicit W and Z mass terms:

$$\mathcal{L}_{\text{Higgs}} = (\partial_\mu h)^2 + \frac{1}{4} g^2 W_\mu W^\mu (v+h)^2 + \frac{1}{8} \left(\sqrt{g^2 + g'^2}\right)^2 Z_\mu Z^\mu (v+h)^2 - V(\frac{1}{2}(v+h)^2), \tag{17}$$

The masses of gauge bosons can then be expressed as

$$m_W = \frac{1}{2} g v, \; m_Z = \frac{1}{2} v \sqrt{g^2 + g'^2}. \tag{18}$$

The mass of the Higgs boson is given by

$$m_H = \sqrt{2\lambda} v = \sqrt{-2\mu^2}, \tag{19}$$

although its value is not predicted, as μ is a free parameter of the SM.

Masses of fermions can be generated using the same scalar field ϕ. For each fermion generation a $SU(2)_L \otimes U(1)_Y$ invariant term, known as Yukawa Lagrangian, is introduced:

$$\mathcal{L}_{\text{Yukawa}} = -\lambda_e \bar{L} \phi e_R - \lambda_d \bar{Q} \phi d_R - \lambda_u \bar{Q} \tilde{\phi} u_R + h.c., \tag{20}$$

where λ_e, λ_d and λ_u are Yukawa couplings to fermions, ϕ is the scalar field given by equation 16 and $\tilde{\phi} = i\tau_2\phi^*$, with τ_2 denoting the second Pauli matrix.

The fermion masses are then defined as [13]

$$m_e = \frac{\lambda_e v}{\sqrt{2}},\ m_u = \frac{\lambda_u v}{\sqrt{2}},\ m_d = \frac{\lambda_d v}{\sqrt{2}}. \tag{21}$$

The full expression for the EW component of the SM Lagrangian is

$$\mathcal{L}_{\mathrm{EW}} = \mathcal{L}_{\mathrm{gauge}} + \mathcal{L}_{\mathrm{fermion}} + \mathcal{L}_{\mathrm{Higgs}} + \mathcal{L}_{\mathrm{Yukawa}}, \tag{22}$$

where the different terms are defined by equations 9, 7, 12 and 20.

To summarise, the Brout-Englert-Higgs mechanism allows to obtain masses for the gauge bosons and fermions, while maintaining invariance under the $SU(2)_L \otimes U(1)_Y$ symmetry. The electromagnetic $U(1)_Q$ and the color symmetry $SU(3)_C$ remain unbroken.

1.1.5 The top quark Yukawa coupling

Experimental measurements of Yukawa couplings to fermions represent an important test of the SM. If the Yukawa coupling of a given fermion calculated from its mass is different from the one extracted from direct measurements in data, it would indicate new physics. The top quark Yukawa coupling y_t is especially important to measure, as it is a key parameter of the SM.

Since the SM is a renormalisable theory, new physics can affect evolution of some coupling constants with growing energy. An important parameter is the Higgs boson self-coupling λ. Since the top quark is the fermion most strongly coupled to the Higgs boson, it gives the largest contribution to the Higgs self-coupling. The energy dependence of λ via the renormalisation group evolution at NLO is given by

$$16\pi^2 \frac{d\lambda}{d\ln\mu} = 24\lambda^2 + 12\lambda y_t^2 - 9\lambda(g^2 + \frac{1}{3}g'^2) - 6y_t^4 + \frac{9}{8}g^4 + \frac{3}{8}g'^4 + \frac{3}{4}g^2 g'^2, \tag{23}$$

where μ is the renormalisation scale and g and g' are the SM gauge couplings.

The effective potential of the Higgs field is very sensitive to the value of the top quark Yukawa coupling y_t, as shown in figure 2. Close to a critical value y_t^{crit} a new minimum of the potential appears at large values of the Higgs field. For $y_t > y_t^{\mathrm{crit}}$ the new minimum is deeper than our electroweak vacuum, which means that our vacuum is metastable.

If $y_t > y_t^{\mathrm{crit}}$, the SM is valid up to a certain energy scale, where new physics must appear.

1.1.6 Quantum chromodynamics

Quantum chromodynamics (QCD) is a non-abelian theory based on the gauge group $SU(3)_C$ that describes strong interactions. The conserved quantity under the group symmetry transformations is called colour.

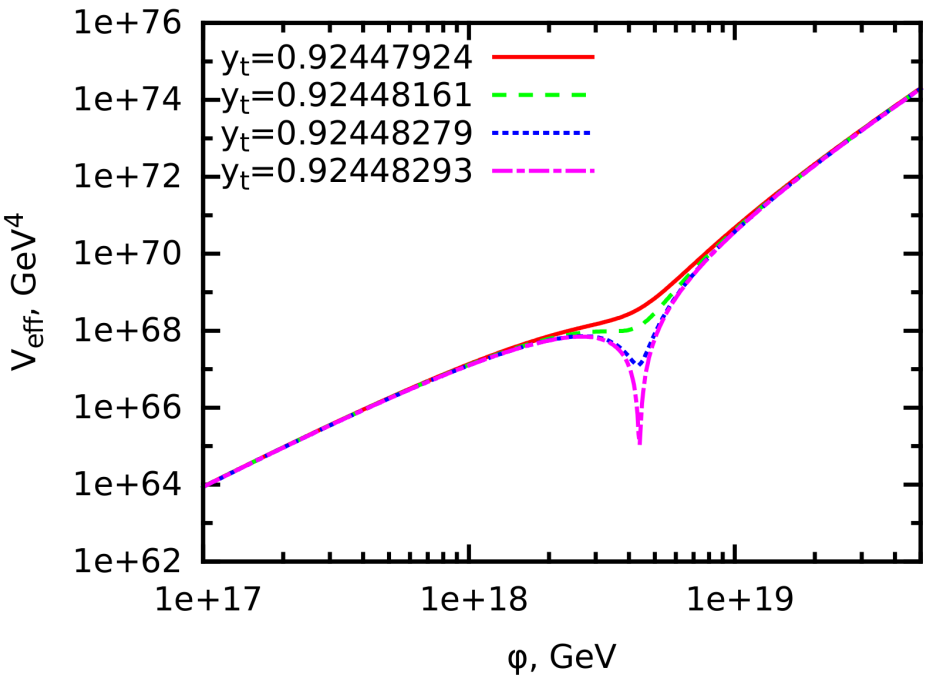


Figure 2: Effective potential of the Higgs field for several values of the top quark Yukawa y_t at renormalisation scale $\mu = 173.2$ GeV as function of the Higgs field. From Ref. [14].

The QCD Lagrangian in the SM is given by

$$\mathcal{L}_{\text{QCD}} = \bar{q} i\gamma^\mu D_\mu q - \frac{1}{4} G^a_{\mu\nu} G^{a\,\mu\nu}, \tag{24}$$

where q are the quark fields and D_μ is the covariant derivative given by

$$D_\mu = \partial_\mu - i g_s T_a G^a_\mu, \tag{25}$$

where g_s is the strong coupling constant, T_a ($a = 1,..8$) are the $SU(3)_C$ generators, and G^a_μ are the gluon fields. $G^a_{\mu\nu}$ is a field tensor, defined as

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu - g_s f_{abc} G^b_\mu G^c_\nu, \tag{26}$$

with f_{abc} being the structure constants of the $SU(3)_C$ group.

Gluons carry colour charge, therefore they can interact with each other. This interaction is described by the last term in equation 26.

The gluon self-interaction has a dramatic effect on the energy dependence of the strong coupling constant, as shown in figure 3. This dependence can be approximated as

$$\alpha_S(Q^2) = \frac{12\pi}{(33 - 2n_f)\log\left(\frac{Q^2}{\Lambda^2_{QCD}}\right)}, \tag{27}$$

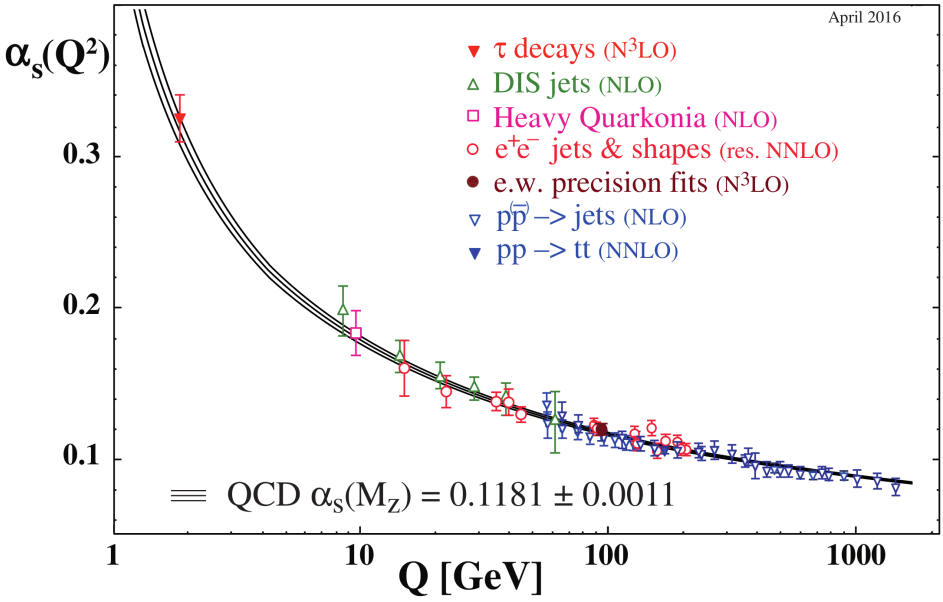


Figure 3: Measurements of α_S as a function of the energy scale Q. From Ref. [4].

where α_S is related to the strong coupling constant g_S as $\alpha_S = g_S^2/4\pi$, Q is the energy scale, n_f is the number of active flavour quarks ($m_q < Q$), Λ_{QCD} is the scale below which the perturbative approximation is no longer valid ($\Lambda_{QCD} \sim 200$ MeV).

With increasing energy (or decreasing distance) α_S decreases. For energies reaching the limit $Q^2 \to \infty$ quarks become free, a phenomenon known as asymptotic freedom. At lower energies (or larger distances) α_S increases and diverges at $Q^2 \to 0$. Due to this, quarks and gluons do not exist as free particles. This feature is referred to as confinement. Quarks produced in high-energy interactions tend to create new bound states with quarks with opposite colour charge from vacuum and produce collimated streams of hadrons known as jets.

1.1.7 Beyond the SM

Despite the fact that the SM is a very successful theory and its predictions are confirmed by the various experiments, there are some physics problems that it is not able to solve:

- The observation of the neutrino oscillations indirectly shows that neutrinos have mass, while they are treated as massless in the SM.
- The matter-antimatter asymmetry in the universe requires a level of violation of the combined symmetries of charge conjugation and parity, known as CP violation, that is significantly larger than that predicted by the SM.
- The SM does not provide an explanation of dark matter, whereas measurements of the rotation curves of galaxies as well as other cosmological studies indicate that it forms a large fraction of the total energy density of the universe.

- The gravitational interaction cannot be described within the quantum field theory, thus a theory unifying all physical forces still does not exist.

These and other open problems in particle physics suggest that the SM is not a complete theory, and thus motivate physicists to search for new phenomena beyond it.

1.2 Search for the Higgs boson at the LHC

The Higgs boson was the last discovered particle predicted by the SM. Since the 1960s, when the SM was developed, and until the discovery in 2012, a broad program of Higgs boson searches was carried out at several colliders: LEP [15], Tevatron [16] and LHC [17]. One of the main goals of the LHC machine was the discovery of the Higgs boson (or proving its absence). After it was discovered, the attention has been focussed on detailed study of its properties.

1.2.1 Production at hadron colliders

There are four main modes for the production of a single Higgs boson at hadron colliders, which are illustrated in figure 4:

- Gluon-gluon fusion (ggH): the Higgs boson is produced via gluon-gluon fusion, mediated by a virtual quark loop, where the main contribution is from the top quark, owing to its large Yukawa coupling. This is the main mechanism of Higgs boson production at the Tevatron and the LHC.

- Vector boson fusion (VBF): two W or Z bosons originating from the initial quarks interact and produce a Higgs boson. This production mode has a special signature that allows to distinguish it from the background: the presence of two light jets in the forward and backward regions of the detector with difference in the pseudorapidity $\Delta\eta \sim$ 3-4 with a maximum transverse momentum of about half of the mass of the vector boson.

- Associated production with a vector boson (VH), or Higgs strahlung: the Higgs boson is produced in association with a W or Z boson, which is typically required to decay leptonically. This was the main production mechanism exploited in the search for a light SM Higgs boson at the Tevatron [18]. In 2017 an evidence for the Higgs boson associated production with a vector boson decaying into b-quarks was announced by the ATLAS [19] and CMS [20] collaborations.

- $t\bar{t}H$ and $b\bar{b}H$: the Higgs boson is produced in association with a top-antitop quark pair ($t\bar{t}$) or a bottom-antibottom quark pair ($b\bar{b}$). The $t\bar{t}H$ process gives direct access in a tree level diagram to the Yukawa coupling to the top quark. The $t\bar{t}H$ process has a smaller cross-section than the three processes described above, but its contribution grows with the energy of pp collisions. The $b\bar{b}H$ process has a comparable cross-section to the $t\bar{t}H$ process, but its signature is similar to ggH, since the associated b-quarks are often produced along the beam direction, thus making this channel very difficult to explore experimentally.

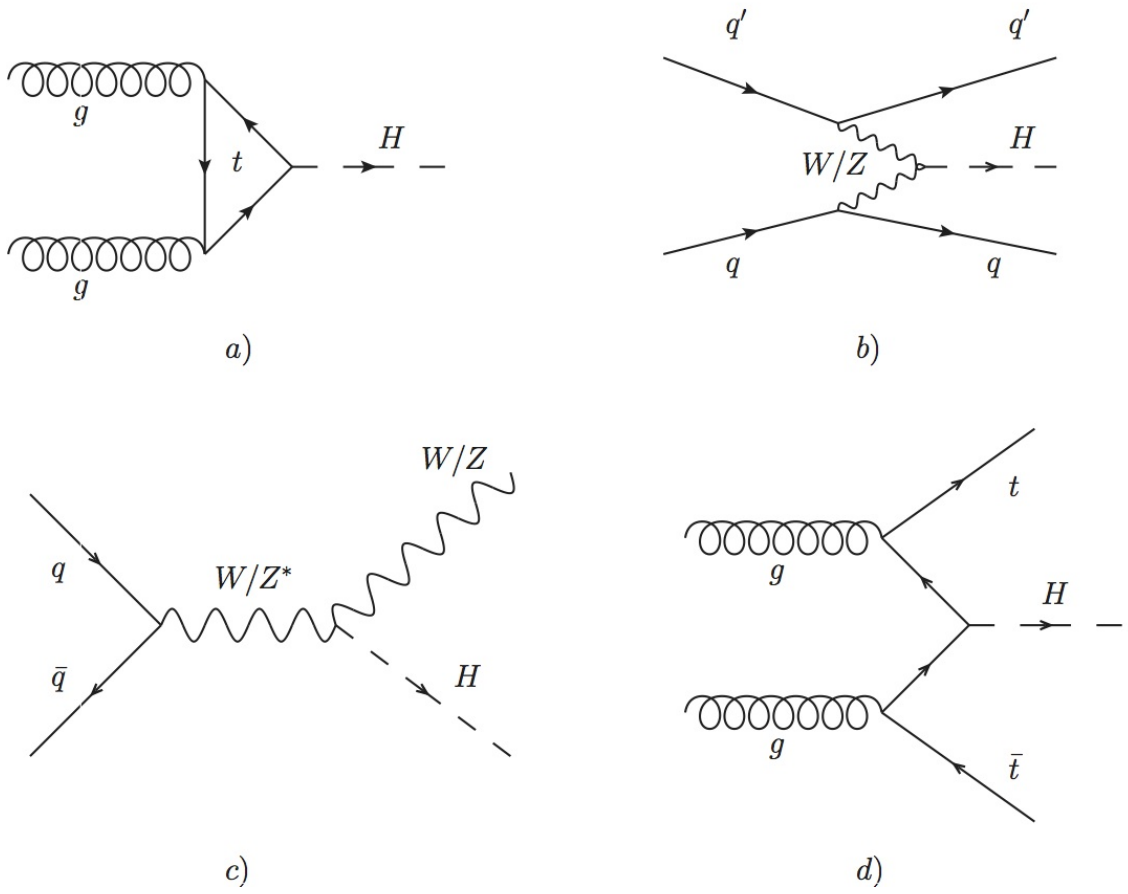


Figure 4: Representative Feynman diagrams for the main production modes for SM Higgs boson production at hadron colliders: (a) gluon-gluon fusion, (b) vector boson fusion, (c) associated production with a vector boson (Higgs strahlung) and (d) $t\bar{t}H$ production.

Production mode	Cross section [pb]
ggH	44.1
VBF	3.78
WH	1.37
ZH	0.88
$t\bar{t}H$	0.507
$b\bar{b}H$	0.488

Table 3: Cross sections of different Higgs-boson production modes predicted by the SM for pp collisions at $\sqrt{s} = 13$ TeV, assuming a value of the Higgs-boson mass of $M_H = 125$ GeV. From Ref. [21].

The SM cross-sections for the different Higgs-boson production modes in pp collisions as a function of the center-of-mass-energy are presented in figure 5. The values of predicted cross-sections for $\sqrt{s} = 13$ TeV are listed in table 3.

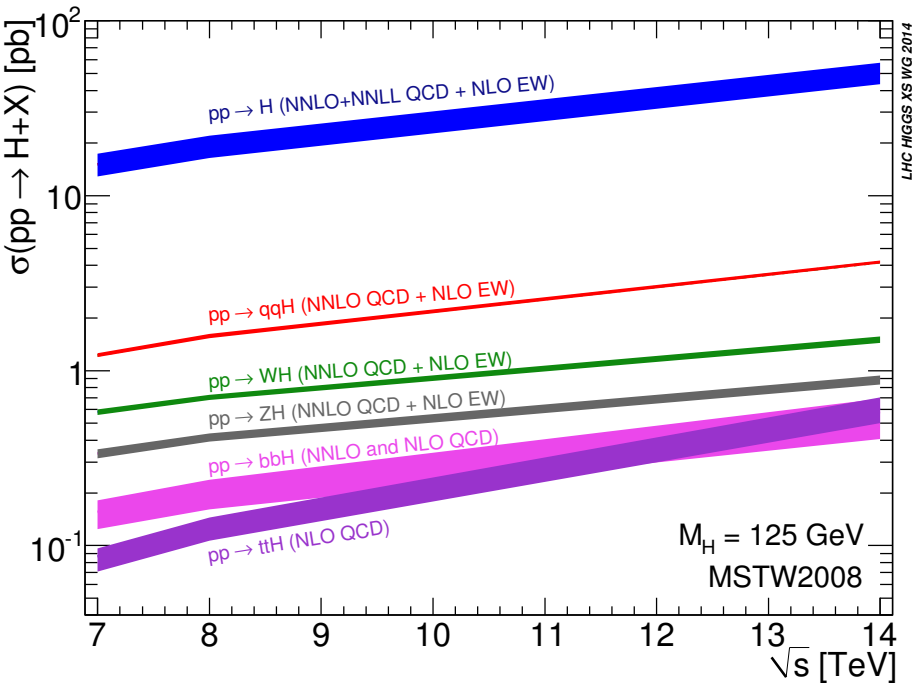


Figure 5: Cross-sections of different Higgs boson production modes in pp collisions as a function of centre-of-mass energy $\sqrt{s}$ assuming a Higgs-boson mass $M_H = 125$ GeV. From Ref. [21].

1.2.2 Decay modes

The Higgs boson decays preferentially to the heaviest particles kinematically accessible. Loop-induced decays to photons and gluons are also possible. Figure 6 shows the different decay branching ratios for the SM Higgs boson as a function of its mass. The branching ratios for the SM Higgs boson assuming $M_H = 125$ GeV are presented in table 4. At this mass, the largest branching ratio corresponds to the $H \to b\bar{b}$ channel.

1.2.3 Discovery

In July 2012 the ATLAS and CMS collaborations announced the discovery of a new particle, compatible with the Higgs boson [23, 24]. The search was performed exploiting the $\gamma\gamma$, $ZZ^* \to 4\ell$ ($\ell = e, \mu$) and W^+W^- channels and using data collected at centre-of-mass energies of $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV. The most significant excesses were observed in the $\gamma\gamma$ and $ZZ^* \to 4\ell$ channels (see figure 7). The mass of the particle was measured to be 126.0 ± 0.4 (stat) $\pm$ 0.4 (syst) GeV by ATLAS and 125.3 ± 0.4 (stat) $\pm$ 0.5 (syst) GeV by CMS. The significance of the observation was 5.9σ for ATLAS and 5.8σ for CMS.

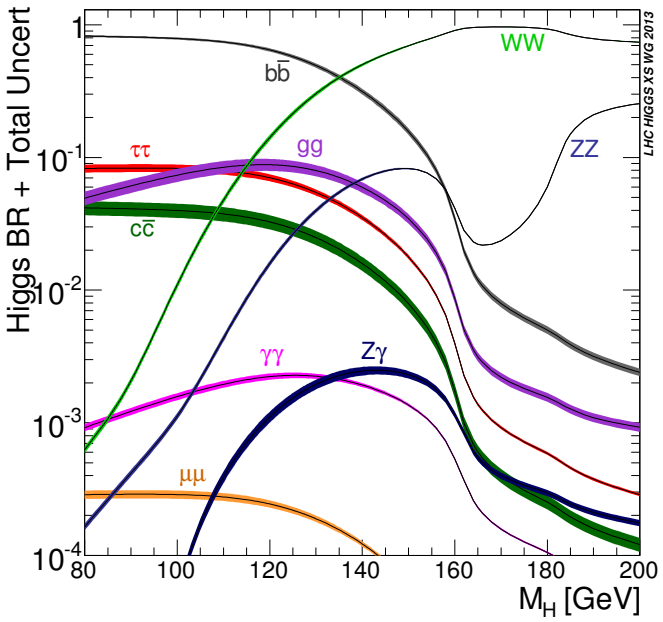
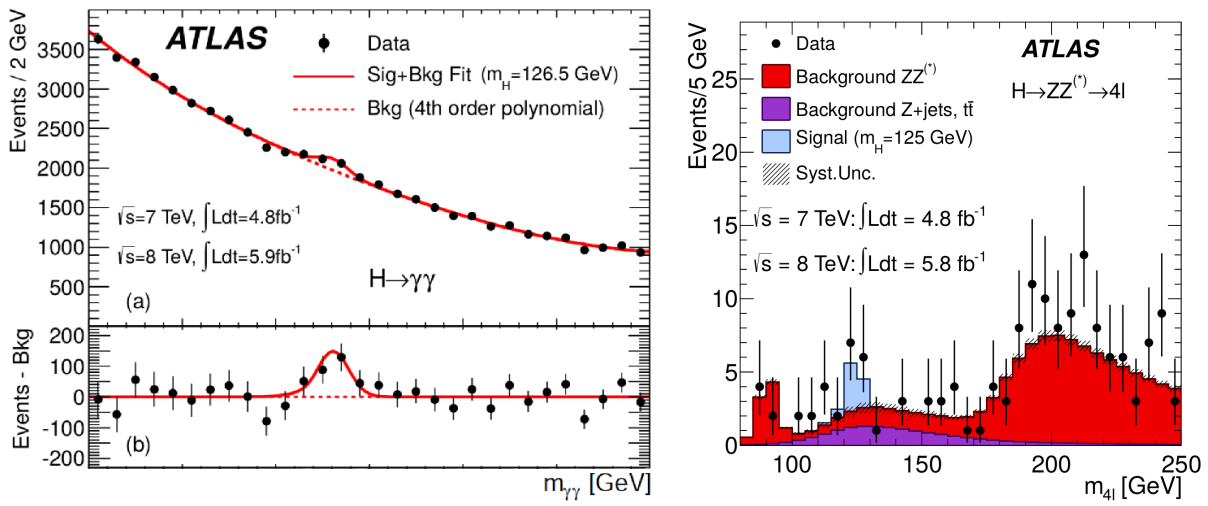


Figure 6: Branching ratios for the different Higgs-boson decay modes as a function of Higgs-boson mass M_H. From Ref. [22].

Decay mode	Branching ratio [%]
$H \to b\bar{b}$	57.7
$H \to WW$	21.6
$H \to gg$	8.55
$H \to \tau\bar{\tau}$	6.37
$H \to c\bar{c}$	2.67
$H \to ZZ$	2.6
$H \to \gamma\gamma$	0.229
$H \to Z\gamma$	0.155
$H \to s\bar{s}$	0.044
$H \to \mu\bar{\mu}$	0.022

Table 4: Branching ratios for the different Higgs-boson decay modes assuming $M_H = 125$ GeV. From Ref. [21].



(a) (b)

Figure 7: Reconstructed mass distribution in the (a) $H \to \gamma\gamma$ and (b) $H \to ZZ^* \to 4\ell$ searches by the ATLAS Collaboration. A peak at a mass value around 125 GeV is observed. From Ref. [23].

1.2.4 Further study of the Higgs boson properties

After the discovery the properties of the new particle were extensively studied to test its compatibility with the SM Higgs boson. Measurements of its couplings to the SM particles were found to be consistent to the SM prediction, as shown in figure 8. In addition the spin and parity of the particle were examined and were found to be consistent with the 0^+ hypothesis, predicted by the SM [25]. The value of the Higgs boson mass measured by ATLAS collaboration using 36 fb^{-1} of 2015 and 2016 data from the LHC at $\sqrt{s} = 13$ TeV is 124.98 $\pm$ 0.28 GeV [26].

The Higgs boson was observed in ggH and VBF production modes with the combined measurements performed by the ATLAS and CMS collaborations in LHC Run 1 at $\sqrt{s} = 7$ TeV and 8 TeV. But the VH and $t\bar{t}H$ processes were not yet discovered, therefore observation of the Higgs boson produced in these two channels are among the priorities of the LHC Run 2. The result of the combined ATLAS and CMS measurement of the signal strength of the $t\bar{t}H$ production with 5 fb^{-1} at $\sqrt{s} = 7$ TeV and 20 fb^{-1} at $\sqrt{s} = 8$ TeV pp collision data is $\mu = 2.3^{+0.7}_{-0.6}$. The observed significance of the signal strength measurement with respect to background-only hypothesis is 4.4 σ, while 2.0 σ is expected [17].

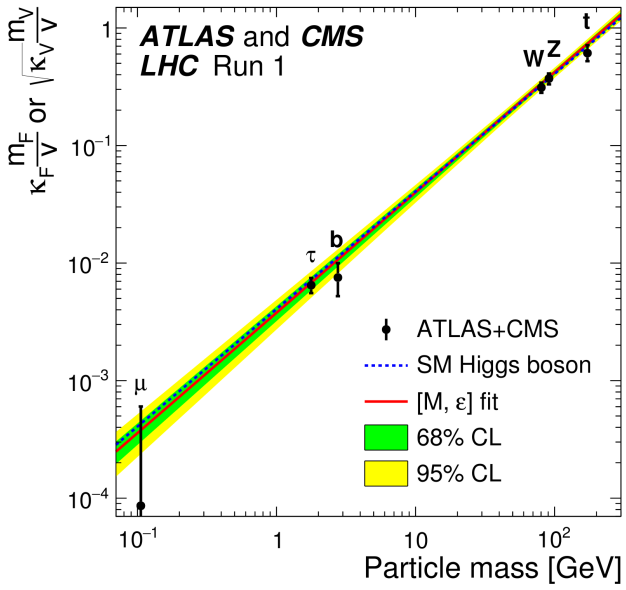


Figure 8: The combined ATLAS and CMS result for scaled couplings of the Higgs boson to W and Z bosons and fermions as a function of the particle mass, measured at $\sqrt{s} = 7$ and 8 TeV, assuming a SM Higgs boson with a mass of 125.09 GeV. The dashed blue line line indicates the prediction of the SM. From Ref. [17].

2 The ATLAS experiment

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is a circular particle accelerator located at the European Organization for Nuclear Research (CERN) situated at the French-Swiss border near Geneva. It accelerates and collides head-on beams of charged particles: protons and heavy ions. It is placed in a 27 km circumference tunnel underground at a depth ranging from 50 m to 170 m that was used previously for the Large Electron-Positron collider (LEP).

The aim of the LHC is to test the predictions of the SM and to search for new phenomena, and this defined its technical parameters. The LHC was designed to collide protons at a centre-of-mass energy of $\sqrt{s} = 14$ TeV. This allows studying SM processes that occur at the high energy scale, as well as to probe various scenarios beyond the SM. The design of high instantaneous luminosity (up to 10^{34} cm^{-2}s^{-1}) produces large data sets, which are crucial to study rare processes, such as Higgs production and to perform high-precision measurements of SM parameters.

The choice of proton-proton collisions was made to minimize the loss of energy due to synchrotron radiation when accelerating charged particles in a curved trajectory: for an electron collider this loss would be prohibitive at such high energies. A proton-antiproton machine would not allow to reach the physics goals because of significantly lower rate of antiproton production, and the consequently lower instantaneous luminosity.

2.1.1 Accelerator complex

The protons are obtained via ionisation of hydrogen gas with a strong electric field. Before protons reach the LHC they go through several acceleration steps to reach the final energy. The CERN accelerator chain is shown in figure 9. In a first step, protons are accelerated to an energy of 50 MeV at the Linear Accelerator 2 (LINAC2). After that three circular accelerators Booster, Proton Synchrotron (PS) and Super Proton Synchrotron (SPS) increase the energy of the proton beams to 450 GeV. Then the beams go inside the LHC main ring where they reach the final energy of 7 TeV (currently 6.5 TeV). The protons are accelerated with electromagnetic field in 16 superconducting radiofrequency (RF) cavities.

There are mainly two types of magnets at LHC: bending dipoles and focusing quadrupoles, together they form the desired beam trajectory. There are in total 1232 dipole magnets in the LHC supplemented by various multipole components to correct imperfections of the magnetic field at the extremities. The dipole magnets of the LHC ring provides a magnetic field of 8.3 T which is needed to keep a beam with an energy of 7 TeV in the desired circular trajectory. The magnets, made of superconductor material, are kept at a temperature of 1.9 K.

As the two proton beams should be accelerated in opposite directions, there are two beam pipes in a dipole element, and two opposite-sign magnetic fields are created around them. Figure 10 shows the cross-section of the LHC dipole element and the beam pipes.

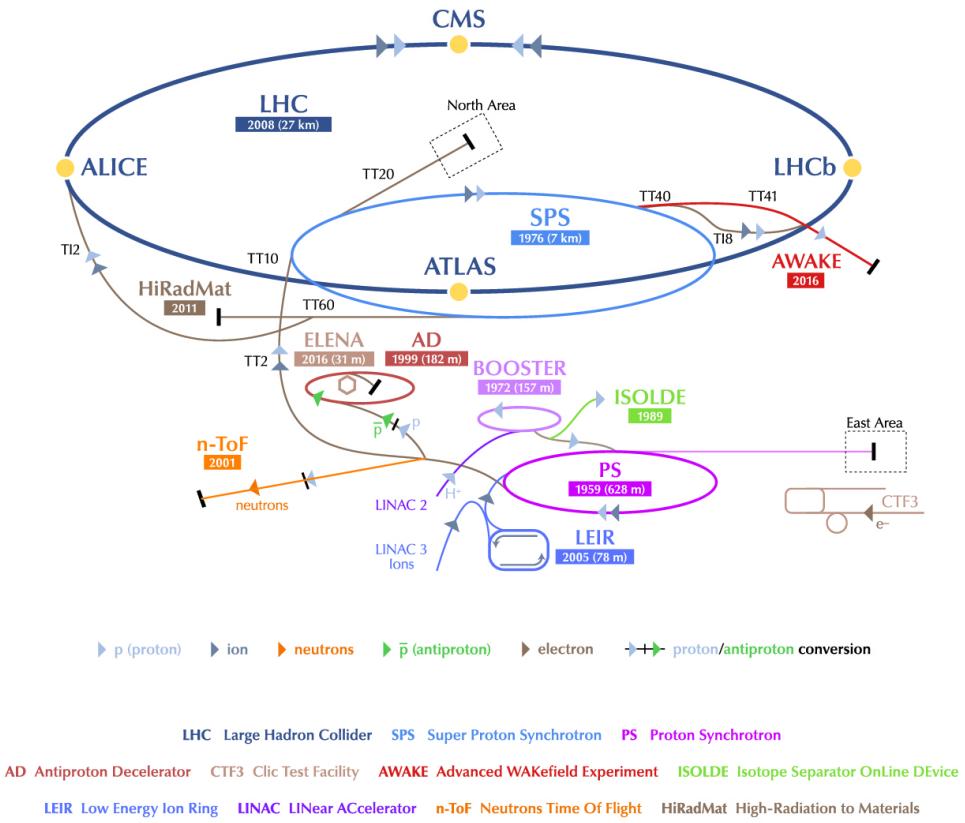


Figure 9: The CERN accelerator complex. From Ref. [27].

The design parameters of the LHC are listed in table 5.

2.1.2 Experiments at the LHC

There are four interactions points where the LHC beams cross, with detectors deployed in these regions to register the collisions:

- ATLAS (A Toroidal LHC Apparatus) [27] is a multipurpose detector, designed to explore various physics processes. This is the largest detector at the LHC (25 m in height and 44 in length) with a weight of 7000 tonnes.

- CMS (Compact Muon Solenoid) [29] is a multipurpose detector with a similar physics program as the ATLAS experiment. With much smaller size than ATLAS (14.6 m in height and 21.6 in length), it weights 12500 tonnes. ATLAS and CMS are placed at opposite intersection points of the LHC ring. Both the detector design and data analysis performed at the two experiments are totally independent, allowing to cross-check the final results.

- ALICE (A Large Ion Collider Experiment) [30] is a detector to study the physics of quarks and gluons behaviour at extreme energy densities (quark-gluon plasma) in heavy ions collisions.

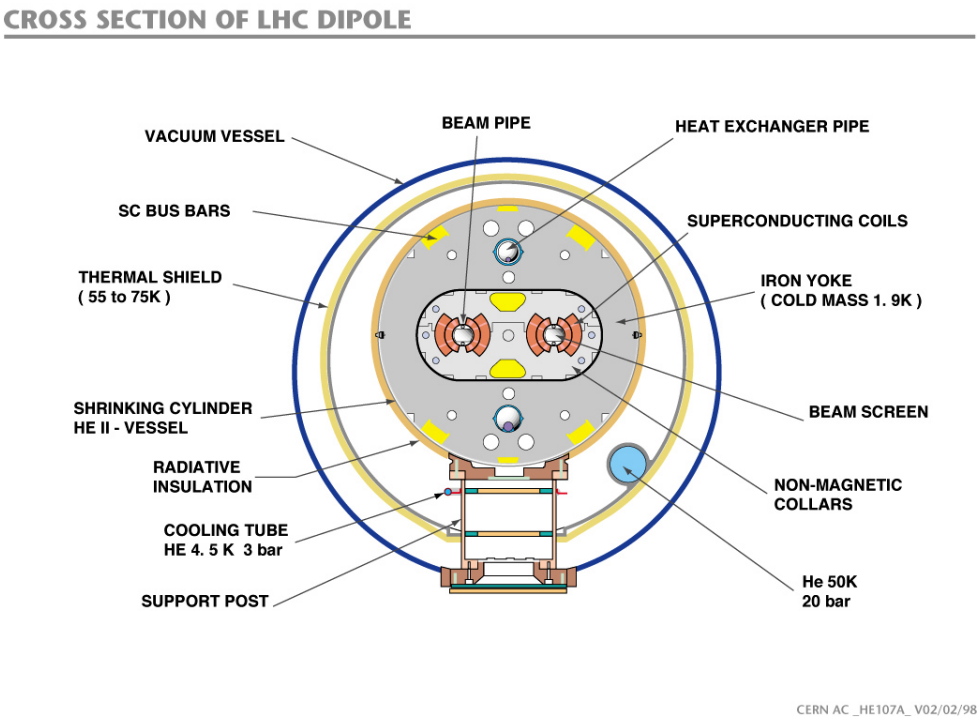


Figure 10: Cross-section of an LHC superconducting dipole element. From Ref. [27].

Beam energy	7 TeV
Injected beam energy	0.45 TeV
Peak luminosity	10^{34} cm^{-2}s^{-1}
Particles per bunch	1.1×10^{11}
Number of bunches	2808
Bunch spacing	24.95 ns
Vertical beam size	18 μm
Horizontal beam size	71 μm
Beam crossing angle	285 μrad
Beam lifetime	13.9 h
Beam energy loss per turn	7 keV
Number of dipole magnets	1232
Max dipole field	8.3 T
Main dipole operation temperature	1.9 K

Table 5: Design parameters of the LHC. From Ref. [28].

- LHCb (Large Hadron Collider beauty) [31] is designed to explore properties of b-hadrons, in particular CP violation in B-meson decay.

In addition, several smaller detectors are installed, such as LHCf (Large Hadron Collider forward) to observe particles at the angles close to the beam direction, TOTEM (TOTal Elastic and diffractive cross section Measurement) that is designed to measure the total proton-proton cross-section and study the proton structure, and MOEDAL (Monopole and Exotics Detector at the LHC) that searches for the magnetic monopole.

2.1.3 Proton-proton collisions

A pp collision at the LHC is a superposition of interactions between the proton constituents: three valence quarks (uud), as well as gluons and $q\bar{q}$ pairs ("sea quarks"), that are produced due to quantum fluctuations. These processes occur in two regimes: high energy (i.e. short distance) interactions, that can be described by perturbative QCD, and low energy (i.e. long distance) non-perturbative effects of the proton structure.

The *factorisation theorem* [32] shows that these two types of processes can be factorised in the calculation of the cross section for a given process X produced in a pp collision:

$$\sigma_{(pp\to X)} = \sum_{a,b} \int dx_a dx_b f_a(x_a, \mu_F^2) f_b(x_b, \mu_F^2) \hat{\sigma}_{ab\to X}(x_a p_a, x_b p_b, \mu_F^2, Q^2), \tag{28}$$

where the sum runs over the possible initial parton types, $\hat{\sigma}_{ab\to X}$ is the perturbative cross-section for partons a and b that is calculated at a fixed order in perturbation theory, Q^2 is the *hard scale* of the process, usually defined as the invariant mass M^2 of the final state of the process. The *factorisation scale* μ_F determines the limit between the high- and low-energy regimes. The *parton distribution functions (PDFs)* $f_i(x_i, \mu_F^2)$ describe the probability for a parton of type i to carry a fraction x_i of the proton momentum.

2.1.3.1 Luminosity

The luminosity describes the rate of collisions produced by a collider. The instantaneous luminosity $\mathcal{L}$ is defined as:

$$\mathcal{L} = \frac{n_1 n_2 n_b f_{\text{rev}} F}{4\pi\sigma_1\sigma_2}, \tag{29}$$

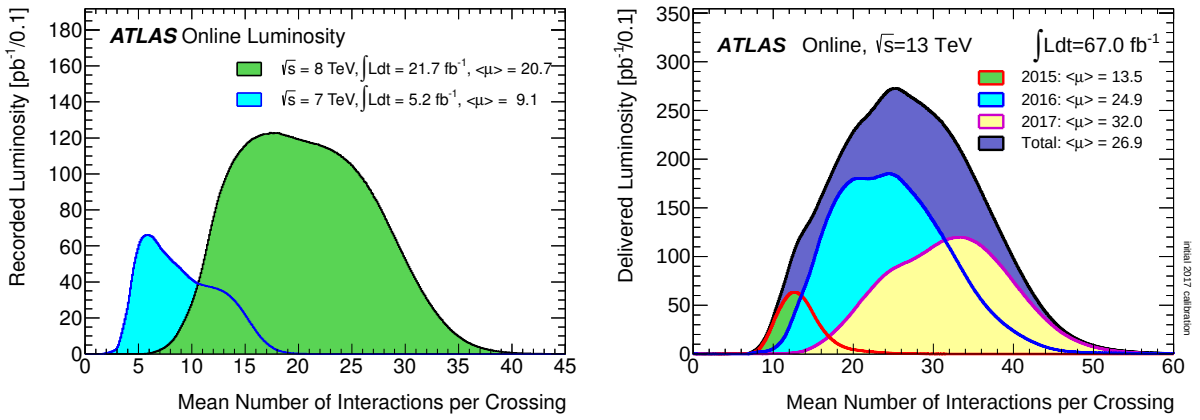
where n_1 and n_2 are numbers of protons per bunch in the two beams, n_b is the number of bunches per beam, f_{rev} is the beam revolution frequency, F is a factor that represents the crossing angle of the two beams, and σ_1 and σ_2 denote the transverse beam dispersions.

The integrated over time luminosity L is defined as

$$L = \int \mathcal{L} dt. \tag{30}$$

and the number of events produced during a time period with a certain reaction is given by

$$N = L\sigma = \int \mathcal{L}\sigma dt, \tag{31}$$



(a) (b)

Figure 11: Luminosity-weighted distribution of the mean number of interactions per bunch crossing for (a) Run 1 and (b) Run 2 *pp* collision data. Full dataset delivered to ATLAS during stable beams is considered. From Refs. [33] and [34].

where σ is the cross-section of this reaction.

2.1.3.2 Pile-up

The events from different *pp* interactions that are registered in the detector simultaneously, on top of the hard-scatter interaction of interest, are referred to as *pile-up*. They are classified into the *in-time* and *out-of-time pile-up* interactions.

The in-time pile-up interactions are those taking place between different protons in the same bunch crossing. The pile-up activity is quantified by the mean of the Poisson distribution of number of interactions per bunch crossing μ:

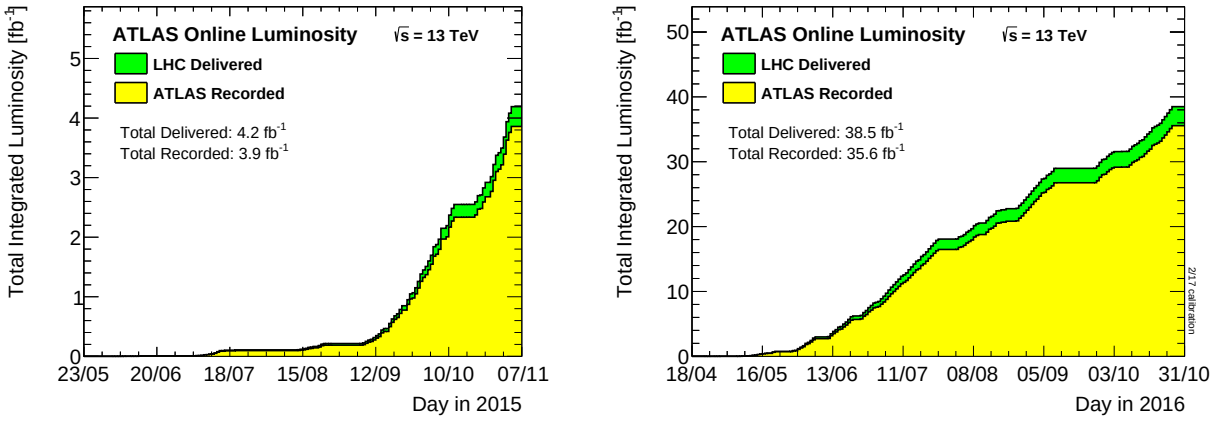
$$\mu = \frac{\mathcal{L}_{\text{bunch}} \times \sigma_{\text{inel}}}{f_{\text{rev}}}, \tag{32}$$

where $\mathcal{L}_{\text{bunch}}$ is the per bunch instantaneous luminosity, σ_{inel} is the inelastic cross section (80 mb for 13 TeV), and f_{rev} is the LHC beam revolution frequency.

The mean number of interactions averaged over all bunch crossings for a considered dataset is denoted as $< \mu >$. Distributions of μ for the 2015, 2016 and 2017 *pp* collision data are presented in figure 11.

The out-of-time pile-up interactions are those coming from events prior or posterior to the analysed one. This overlay of interactions of different bunch crossings happens because of the limited read-out time of some subsystems of the detector.

With higher luminosities it becomes more and more challenging to separate the hard scattering process of interest from the pile-up activity.



(a) (b)

Figure 12: Integrated luminosity versus time delivered to (green) and recorded by ATLAS (yellow) during stable beams for pp collisions at 13 TeV centre-of-mass energy in (a) 2015 and (b) 2016. From Ref. [34].

2.1.4 Experimental data

The first period of data taking at the LHC in 2009-2013 is referred to as Run 1. The first *pp* collisions took place in September 2008, but an electrical fault causing damage to several superconducting magnets led to more than one year of repairs. The machine started again in November 2009, and in 2010 the first collisions at $\sqrt{s} = 7$ TeV were obtained. The *pp* collisions data collected in 2009-2010 were used for detector commissioning, as well as for the first physics studies. Most of the data used in physics analysis was collected in 2011-2012. The ATLAS detector recorded 4.57 fb^{-1} of high-quality data at $\sqrt{s} = 7$ TeV in 2011 and 20.3 fb^{-1} at $\sqrt{s} = 8$ TeV in 2012.

The long maintenance period, from February 2013 to June 2015, allowed to consolidate the LHC and its detectors and prepare for the higher energies and luminosity. New electrical insulation systems and pressure relief devices were installed, the magnet system was upgraded (four quadrupole and 15 dipole magnets were replaced), additional electrical resistance measurements and leak tightness tests were performed to ensure robustness of the system under the new conditions.

Finally, the *pp* collisions restarted in June 2015 at $\sqrt{s} = 13$ TeV. The second LHC data taking period (2015-2018) is referred to as Run 2. The total luminosity delivered and recorded by ATLAS in 2015 and 2016 years is shown in figure 12. As of 4 October 2017, the luminosity recorded by ATLAS in 2017 is 27.9 fb^{-1}.

The search presented in this thesis uses 36.1 fb^{-1} of data collected at $\sqrt{s} = 13$ TeV in 2015 and 2016.

2.2 The ATLAS detector

ATLAS [27] is one of the two multi-purpose experiments that is designed to test the SM predictions and make precise measurements of the SM parameters, as well as to search for new phenomena at the energy frontier being probed in pp collisions at the LHC.

The ATLAS detector is located in an underground cavern in one of the interaction points of the LHC. It is a cylindrical forward-backward symmetric detector, that provides almost full spatial coverage around the pp interaction point. As illustrated in figure 13, the ATLAS detector is composed of several subsystems:

- The inner detector, surrounded by a thin superconducting solenoid. The aim of this detector is finding the tracks of charged particles and measuring their momenta and other parameters using the curvature of their trajectories in the magnetic field.
- The electromagnetic and hadronic calorimeters that allow to measure the energy of particles (e, γ, π, K) that they deposit through their destructive interaction with the detector material.
- The muon spectrometer, designed to detect muons. Three large superconducting toroids form a magnetic field for muon momenta measurements.

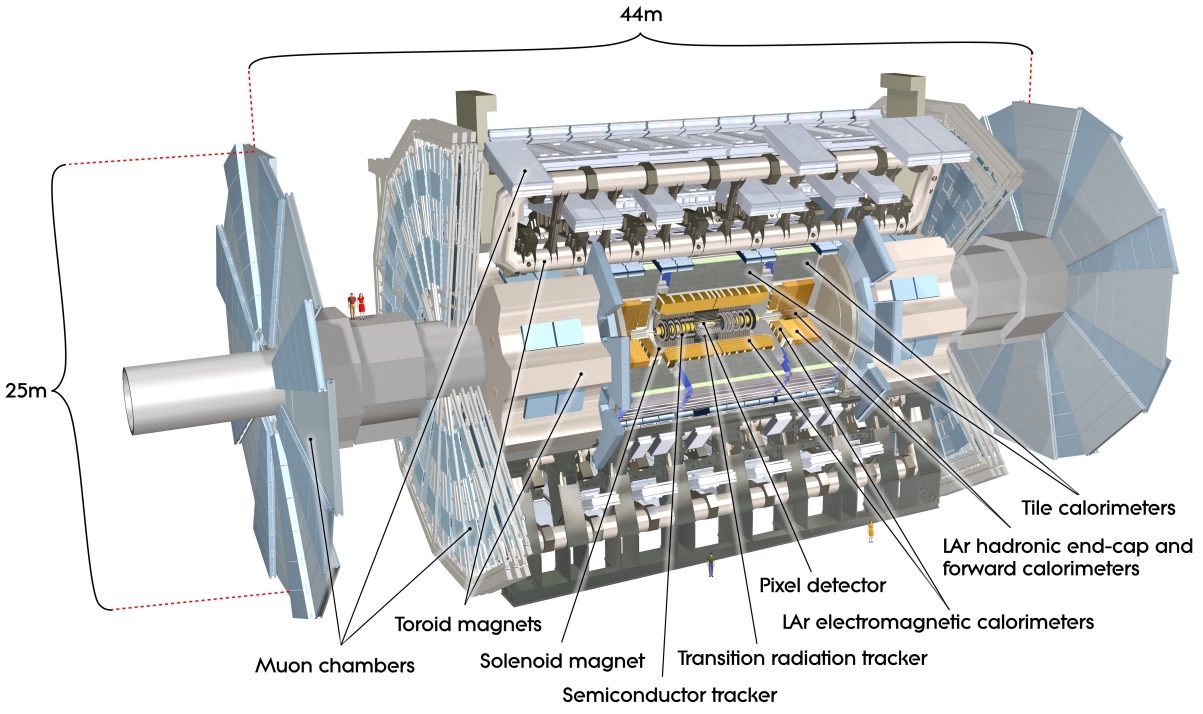


Figure 13: Cut-away view of the ATLAS detector, showing the different subdetectors and the magnet systems. From Ref. [27].

2.2.1 Coordinate system

A right-handed coordinate system with the origin in the interaction point is used in ATLAS. The z-axis is defined by the beam direction, the x-axis points towards the centre of the LHC ring and the y-axis points upwards.

The asimuthal angle ϕ is defined with respect to x-axis in the x-y plane in a range $-\pi < \phi < \pi$, and θ is the polar angle with respect to the z-axis ($0 < \theta < \pi$).

The z-component of initial partons momentum is unknown, therefore choice of variables that are boost-invariant along the z-axis is preferable.

The transverse momentum p_T and the transverse energy E_T are defined in the x-y plane, so that they are also boost-invariant along the z-axis:

$$E_T = E \sin\theta, \;\; p_T = p \sin\theta. \tag{33}$$

The rapidity variable, which is used for massive objects such as jets, is defined as

$$y = \frac{1}{2} \ln \frac{(E + p_z)}{(E - p_z)}. \tag{34}$$

For those particles which mass can be neglected, the rapidity becomes equal to the pseudorapidity η:

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right). \tag{35}$$

Differences in rapidity Δy and pseudorapidity $\Delta\eta$ for massless particles are boost-invariant along the z-axis, unlike $\Delta\theta$.

The distance ΔR in the pseudorapidity-azimuthal angle space is defined as

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}. \tag{36}$$

2.2.2 Magnet system

A magnetic field bends the trajectories of the charged particles, which allows to calculate the particle momentum using the measured track curvature.

The ATLAS magnet system was designed to provide a field mostly orthogonal to the particle trajectory. It consists of four large superconducting magnets: the Central Solenoid which provides the magnetic field to the inner detector, and three toroids, that generate the magnetic field for the muon spectrometer.

The central solenoid surrounds the inner detector and produces a 2 T magnetic field, parallel to the beam axis. It is placed in the same cryostat as the calorimeter.

The magnetic field for the muon spectrometer is generated by three toroids: one large around the barrel part of the calorimeter system and two smaller in the end-caps of the detector. These magnets produce field with a strength of approximately 0.5 T and 1 T for the muon detector in the central and the end-cap regions. The toroid configuration was chosen to achieve the desired magnetic field over a large detector volume with a relatively

small amount of material. Figure 14 illustrates a scheme of the ATLAS magnet system. In figure 15 a photograph of the barrel toroid after installation is shown.

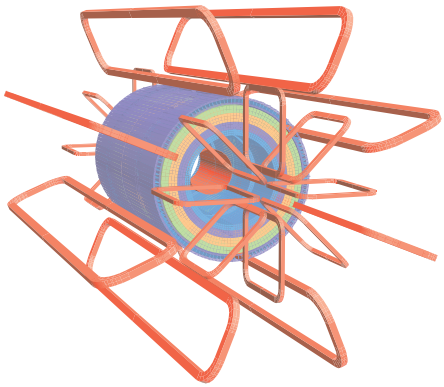


Figure 14: Scheme of the ATLAS magnet system (in orange) and tile calorimeter steel. Windings of barrel and endcaps toroid coils are visible, as well as the solenoid, that lies inside the calorimeter volume. From Ref. [27].

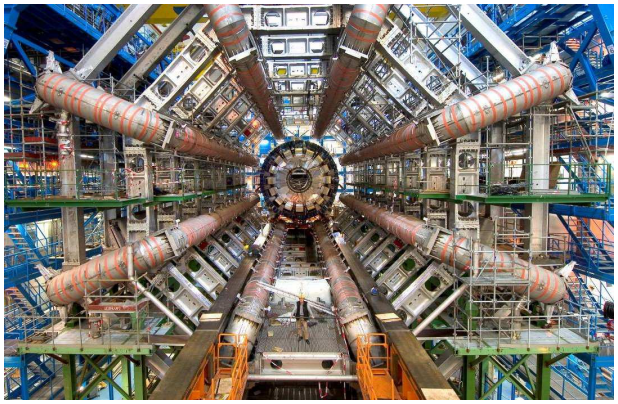


Figure 15: Barrel toroid after installation in the underground cavern. From Ref. [27].

2.2.3 Inner detector

The purpose of the ATLAS inner detector, or tracker, is to perform tracking measurements. With LHC high luminosities the track density in the detector is very high. Thus to provide track resolution as high as required for physics analyses a very high detector granularity is needed, and this demand and the radiation hardness requirement determined the design of the inner-detector system.

There are three components of the inner detector. The pixel detector, located in the innermost part of the detector, performs precise 3D track measurements with high-granularity silicon sensors. The silicon microstrip, or semiconductor tracker (SCT), uses small-angle stereo strips to obtain 2D spatial track measurements. Finally, the transition radiation tracker (TRT) provides 1D track measurements at a larger radius via straw tubes filled with gas. The layout of the ATLAS inner detector is presented in figure 16. The inner detector measures tracks within $|\eta| < 2.5$.

Each of the inner detector subsystems plays an important role in tracking. The pixel detector with the highest granularity provides the most accurate measurements, while the SCT improves the track spatial resolution due to its stereo strips rotated at a small angle, that allow to measure both ϕ and $z(R)$ coordinates in the barrel (endcaps), thus their combination gives very robust pattern recognition and high precision in R, ϕ and z coordinates. Multiple less precise measurements by TRT at larger radii improve the momentum resolution.

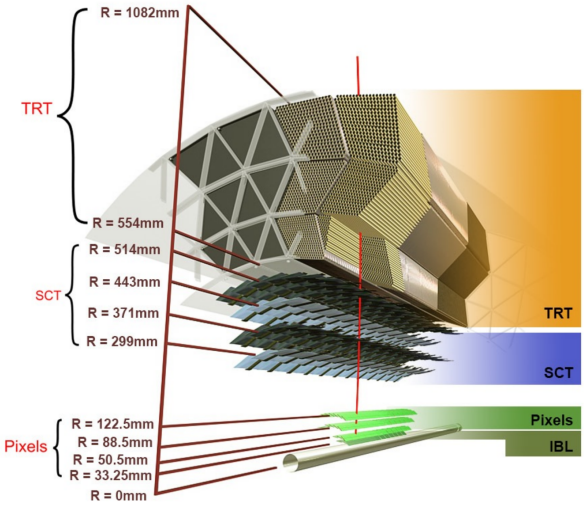


Figure 16: Scheme of the ATLAS inner detector. From Ref. [35].

2.2.3.1 Pixel detector

The pixel detector is the innermost part of the ATLAS tracking system. Being located close to the interaction point, the pixel detector plays an important role for the identification of particles with lifetime, such as b- and c- hadrons and τ-leptons.

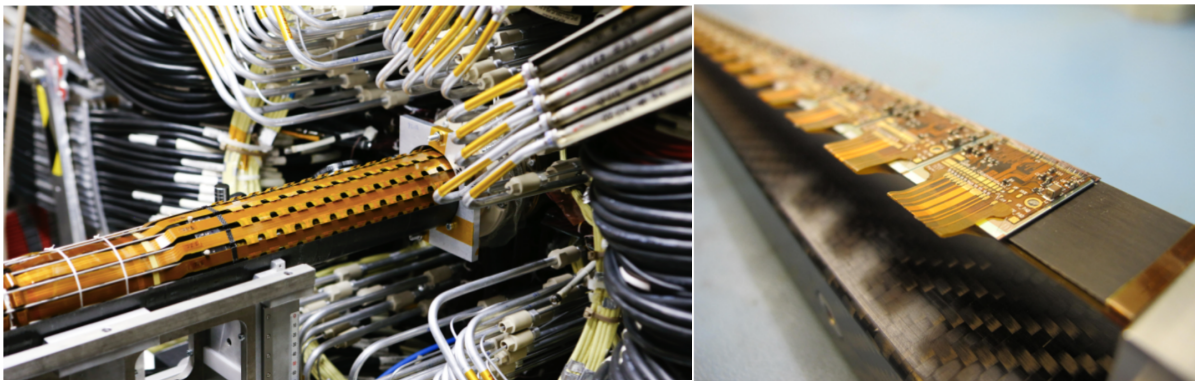
For Run 1 the pixel detector consisted of three cylindrical layers and 6 disks. The barrel parts of the three subdetectors surround the beam axis at a radius of $r = 50.5$, 88.5, and 122.5 mm. In the end-cap regions three disc layers perpendicular to the beam axis are installed on each side of the detector.

There are 1744 detector modules, each composed of 47232 pixels. The size of a pixel is 50×400 μm. Each sensor is a silicon wafer with 16 front-end chips for readout. In total the pixel detector has 80 million readout channels.

The Insertable B-layer

The major ATLAS inner detector upgrade for Run 2 was the addition of the Insertable B-Layer (IBL) [36], a fourth innermost pixel layer that was inserted inside the existing pixel detector at a radius of ≈ 3.3 cm from the beamline. Figure 17 shows the IBL prior to the insertion and an IBL stave with mounted detector modules. As a result, the average number of pixel measurements on a single track was increased from three to four. This improves the tracking robustness with respect to pile-up and possible pixel module failures.

The main reason for the IBL installation was the possible decrease of the former innermost layer efficiency due to radiation damage. With high luminosities the detector components are affected by radiation. The sensors are more sensitive to the impact of neutral particles, whereas the electronic components are mostly damaged by charged particles. At the time of the ATLAS inner detector design there was no technology that would allow to construct a small-radius layer that is sufficiently robust against radiation for seven years of operation. The lifetime of the former innermost layer is limited to



(a) (b)

Figure 17: (a) The IBL before insertion inside the pixel detector. (b) An IBL stave with detector modules mounted on carbon fibre support structures. From Ref. [37].

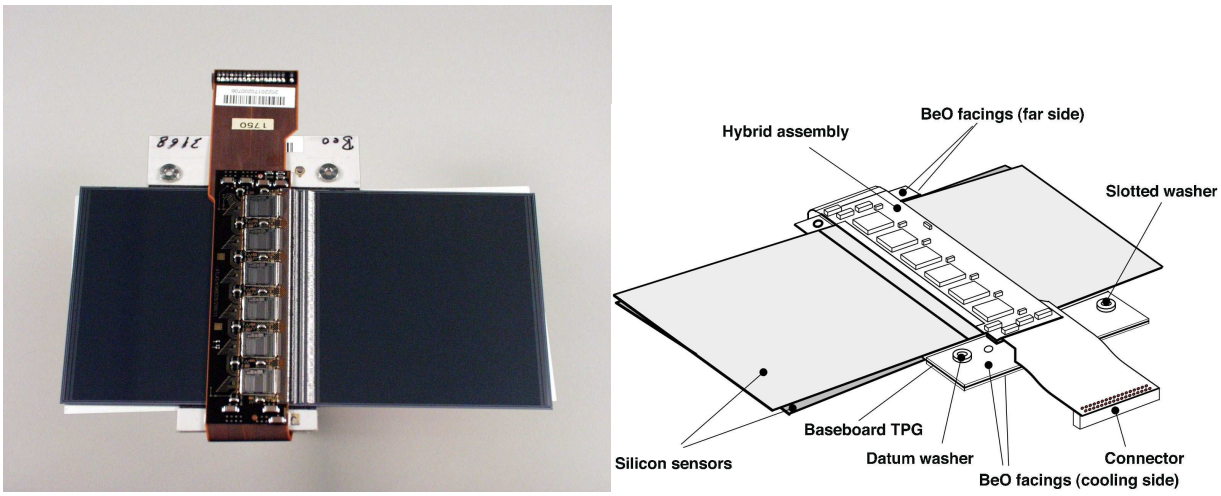
300 fb^{-1}, corresponding to a total ionising dose (TID) of 100 Mrad. The more advanced technologies used for the IBL provide increased resistance against higher ionising dose. It was successfully tested to withstand a TID of up to 250 Mrad.

Apart from radiation damage, the higher pile-up worsens the tracking performance, since the significant increase in the number of tracks makes it difficult to find and reconstruct them. The pixel detector used during Run 1 was designed for a peak luminosity of $\mathcal{L} \approx 10^{34}\text{cm}^{-2}\text{s}^{-1}$, while during Run 2 the peak luminosity is often $\mathcal{L} \approx 2 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$. In particular, with higher pile-up the detector readout efficiency is decreased. This affects the innermost detector layer more than other layers and thus has huge impact on the b-tagging performance. The novel readout techniques of the IBL design allow to maintain tracking and b-tagging performance despite the increased pile-up. The pixel detector with the IBL is expected to efficiently perform track pattern recognition for peak luminosities as high as $\mathcal{L} \approx 3 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$ [38].

The presence of the fourth layer is also important for the areas of the pixel detector with so called *dead modules*. Irreparable failures in the pixel detector inevitably appear over time and impact tracking and vertexing performance. Until the end of Run 1 around 10% of the former innermost layer modules were not operational. During the 2013-2014 downtime most of them were repaired, but $\sim 1\%$ were not working at the start of Run 2. With the IBL three pixel hits are still provided in the segments where former innermost layer modules are not operational, so tracks still can be reconstructed effectively.

Another advantage of the IBL is its higher granularity, with pixels of size 50 μm $\times$ 250 μm instead of 50 μm $\times$ 400 μm for the former innermost pixel layer. It allows to simplify the track finding and improves the precision of the track measurements.

The location of the IBL close to the beamline plays important role for secondary vertex finding and track impact parameter measurements, necessary for b-tagging.



(a) (b)

Figure 18: Components of a barrel SCT module. From Ref. [27].

2.2.3.2 Semiconductor Tracker

The Semiconductor Tracker (SCT) is based on sensors with microstrips. It consists of four cylindrical layers parallel to the beam axis and nine end-cap disks with radially oriented strips. There are 2112 detector modules in the barrel layers and 988 in the endcap disks on each side. The barrel layers are located at a distance of 299-514 mm from the beam line.

Figure 18 shows the components of a barrel module. The four sensors, two on the top and two on the bottom side of the module, are rotated by a stereo angle of ± 20 mrad, which allows to improve the spatial resolution of the detector. An SCT sensor has a pitch of 80 μm and consists of a wafer and 768 active microstrips of 12.8 cm length. The wafer is a n-type semiconductor and the strips are p-type semiconductors. The total number of readout channels in the SCT is approximately 6.3 millions. The intrinsic accuracies of SCT are 17 μm (R-ϕ) and 580 μm (z) in the barrel and 17 μm (R-ϕ) and 580 μm (R) in the end-caps.

2.2.3.3 Transition radiation tracker

The Transition radiation tracker (TRT) is designed for two main goals. The first is to perform track measurements at large radii (554 mm $< R <$ 1082 mm in the barrel and 617 mm $< R <$ 1106 mm in the end-cap), that play an important role for the momentum measurement. The second goal is to detect the transition radiation, which is important for electron identification.

The TRT consists of straw tubes filled with gas, embedded in a matrix of polypropylene fibres, which is used as transition radiation material. The transition radiation (photon emission) rate is inversely proportional to the mass of the particle. Therefore, among the particles that leave a track in the detector, electrons produce the largest amount of photons, so measuring the transition radiation allows to identify them. For electrons with

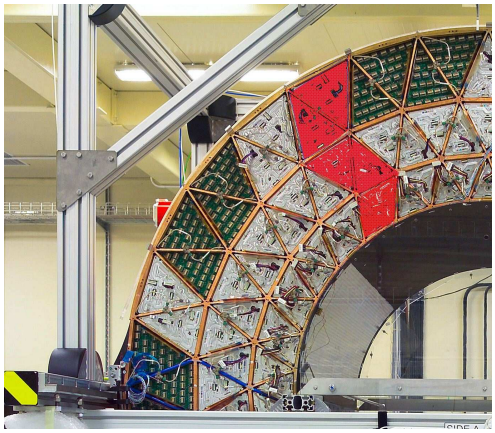


Figure 19: Photograph of a quadrant of the TRT barrel during the integration of the modules at CERN. The shapes of one outer, one middle and one inner TRT module are highlighted. From Ref. [27].

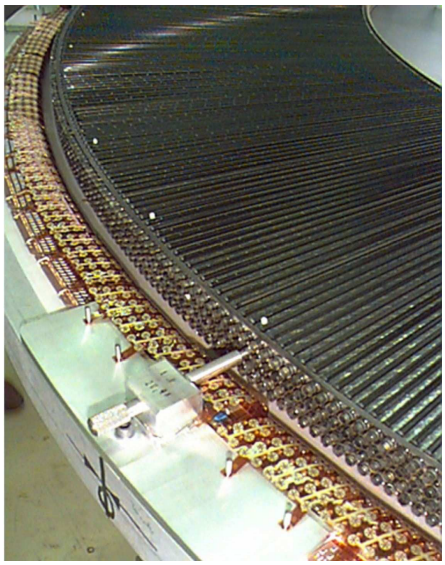


Figure 20: Photograph of a four-plane TRT end-cap wheel during asssembly. From Ref. [27].

energies above 2 GeV from seven to ten high-threshold hits from transition radiation are expected.

The TRT contains up to 73 layers of straws in the barrel and 160 straw planes in the end-cap. Originally the straws were filled with a gas mixture of 70% Xe, 27% CO_2 and 3% O_2. Starting from 2012, gas leaks were observed in some modules of the detector due to cracks in the pipes. In the damaged parts the Xe-based gas was replaced by an Ar-based gas mixture [39]. Figure 19 shows a quadrant of the TRT barrel and figure 20 presents a four-plane TRT end-cap wheel during assembly.

A track with $p_T > 0.5$ GeV and $|\eta| < 2$ traverses a minimum of 36 straws (except in the barrel-end-cap transition region of $0.8 < |\eta| < 1.0$, where a track crosses a minimum of 22 straws). The TRT provides only R-ϕ measurement, with a precision of 130 μm.

2.2.4 Calorimeters

The ATLAS calorimetry system measures the energy of electromagnetically and hadronically interacting particles. The structure of the ATLAS calorimetry system is presented in figure 21. It consists of several detectors providing full ϕ-symmetry and pseudorapidity coverage of $|\eta| < 4.9$. The innermost calorimeters are the electromagnetic barrel calorimeter (EMB), the electromagnetic end-cap calorimeter (EMEC), the hadronic end-cap calorimeter (HEC) and the forward calorimeter (FCal). For these calorimeters liquid argon (LAr) is used as the active detector material. The outer calorimeter part is the Tile barrel, which consists of scintillator tiles and steel as absorber medium. LAr barrel and EMEC are electromagnetic calorimeters, designed for the measurement of electromagnetic showers that electrons and photons produce through the Bremsstrahlung process and pair

production. HEC, FCal and Tile are hadronic calorimeters, used to measure the energy of hadronic showers (jets).

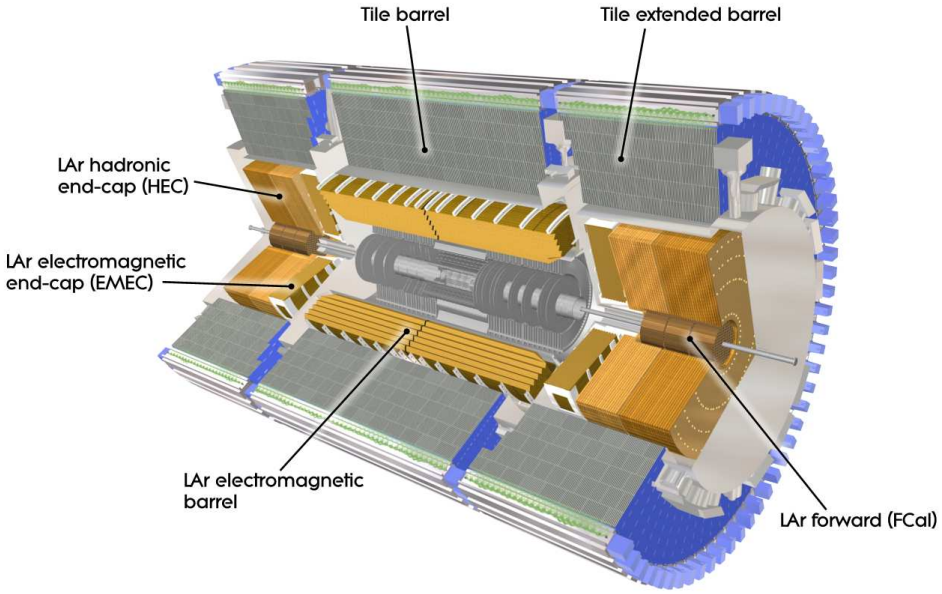


Figure 21: Cut-away view of the ATLAS calorimeter system. From Ref. [27].

2.2.4.1 Electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) is based on high-granularity liquid argon (LAr) technology, that allows to make high-precision measurements of electrons and photons energy. It consists of the barrel section, covering the pseudorapidity region $|\eta| < 1.475$ and two end-caps in the range $1.375 < |\eta| < 3.2$. Each of the three components is housed in its own cryostat. The liquid argon was chosen as the active detector material due to its stability of response over time and resistance to radiation.

Both barrel and end-cap of ECAL consist of accordion-shaped kapton electrodes and lead absorber plates immersed in liquid argon. The accordion geometry has been chosen to provide a full coverage in ϕ without gaps, and a fast extraction of the signal at the back and at the front of the electrodes. A photograph of a barrel ECAL module is shown in figure 22.

The detector consist of three sampling layers: the strip layer, the middle and the back. The electronic readouts are organised such that the volume of detector is divided into virtual cells representing elementary energy deposits. An electromagnetic shower is registered in the detector as a cluster of such cells and its energy is measured by summing the energies deposited in each cell. A segment of ECAL scheme is presented in figure 23.

The ECAL energy resolution as function of energy is given by

$$\frac{\sigma(E)}{E} = \frac{a}{\sqrt{E}} \oplus b, \tag{37}$$

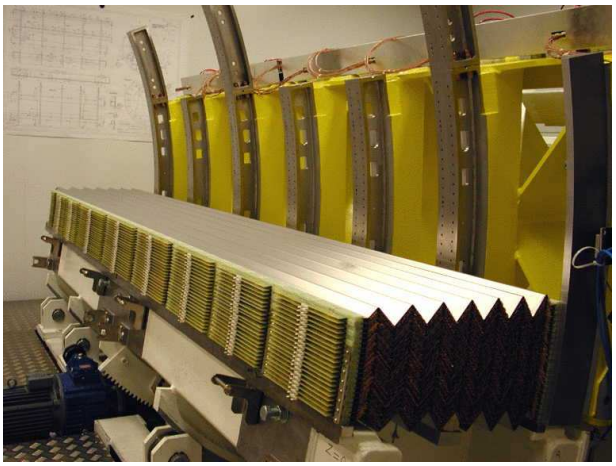


Figure 22: Photograph of a partly stacked barrel electromagnetic LAr module. From Ref. [27].

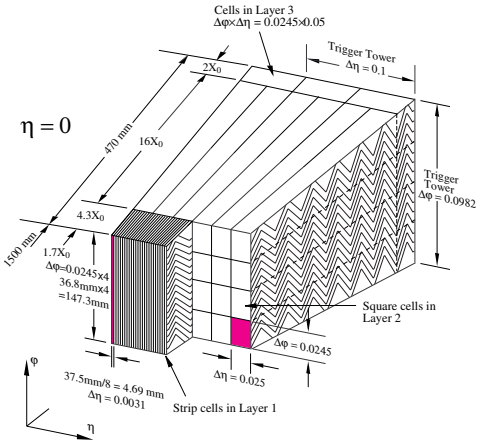


Figure 23: Sketch of an electromagnetic LAr barrel module. From Ref. [27].

where for the barrel part $a \sim 9 - 10\%$ is the stochastic term, $b \sim 0.2\%$ is the constant term (both values obtained from measurement) and E is expressed in GeV.

2.2.4.2 Hadronic calorimeters

Hadronic calorimeters are designed to measure the energy of hadronic showers (jets). The ATLAS hadronic calorimeter system consists of the Tile hadronic calorimeter, the LAr hadronic end-cap calorimeter and the LAr forward calorimeter.

Tile calorimeter

The Tile calorimeter is a sampling calorimeter that consists of iron plates as absorber material and plastic scintillator tiles as active material. It is composed of a central barrel, covering the region $|\eta| < 1.0$, and two extended barrels, covering the range $0.8 < |\eta| < 1.7$.

Scintillation light produced in the tiles goes through wavelength shifting fibres to photomultiplier tubes (PMTs), where the resulting electronic signal is measured. Figure 24 shows the integration of the mechanical assembly and the optical readout of the tile calorimeter.

LAr hadronic end-cap calorimeter

The hadronic end-cap calorimeter (HEC) uses liquid argon as active material and copper as absorber. It is used to measure the energy of particles in the range $1.5 < |\eta| < 3.2$. The HEC consists of two end-cap sections, located behind the electromagnetic calorimeter end-caps in the same cryostat. Each of the two hadronic end-cap sections is composed of four layers: two independent wheels, additionally divided into two segment in depth. The wheels that are located closer to the interaction point are built of 25 mm

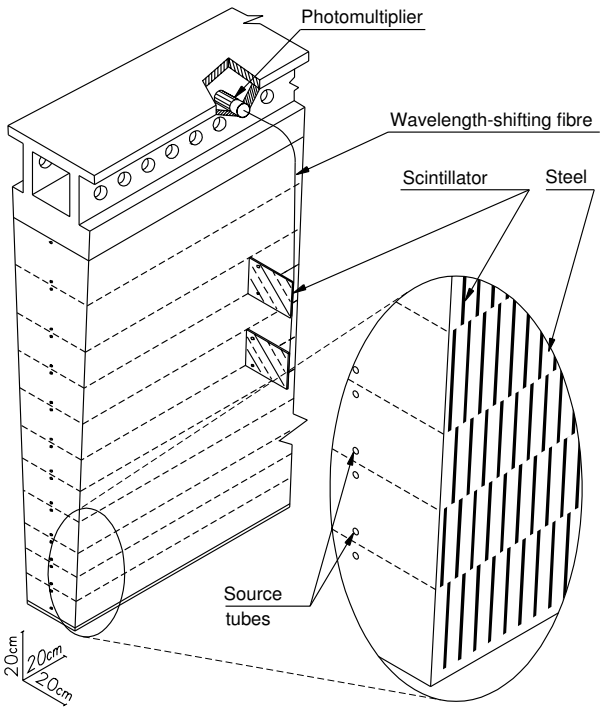


Figure 24: Scheme of the mechanical assembly and the optical readout of the Tile calorimeter. From Ref. [27].

copper plates, the outer wheels use 50 mm copper plates. The 8.5 mm gaps between the plates are filled with liquid argon.

LAr forward calorimeter

The forward calorimeter (FCal) detects particles in the very forward region in the range $3.1 < |\eta| < 4.9$. It consists of two sections (one per end-cap), located inside EMEC and HEC. Each FCal section consists of three modules: the first module is made of copper and is designed for electromagnetic measurements, while the second and third modules are made of tungsten and measure the energy of hadronic showers.

2.2.5 Muon spectrometer

Muons are the only particles detectable by ATLAS that can traverse the calorimeters. They are measured by the muon spectrometer (MS), the outermost section of the detector. This detector is designed for the identification of muons, the reconstruction of their tracks and precision measurement of their momenta. The presence of a muon with certain characteristics can be a sign of an interesting physics process; therefore information from the MS is used by the ATLAS trigger.

The MS is located in the magnetic field generated by the toroid magnets: the barrel toroid in the range of $|\eta| < 1.4$ and two end-cap toroids in the range of $1.6 < |\eta| < 2.7$.

In the range in between, referred to as transition region, $1.4 < |\eta| < 1.6$, tracks are bent by both barrel and end-cap toroids.

The MS consists of four subdetectors: Monitored Drift Tubes (MDT), Cathode Strip Chambers (CSC), Resistive Plate Chambers (RPC), and Thin Gap Chambers (TGC). Those are muon chambers of two types: the first two (MDT in the barrel and CSC in the end-caps) are designed for track precision measurement, whereas the last two (RPC in the barrel and TGC in the end-caps) provide the trigger information. A scheme of the muon system is presented in figure 25.

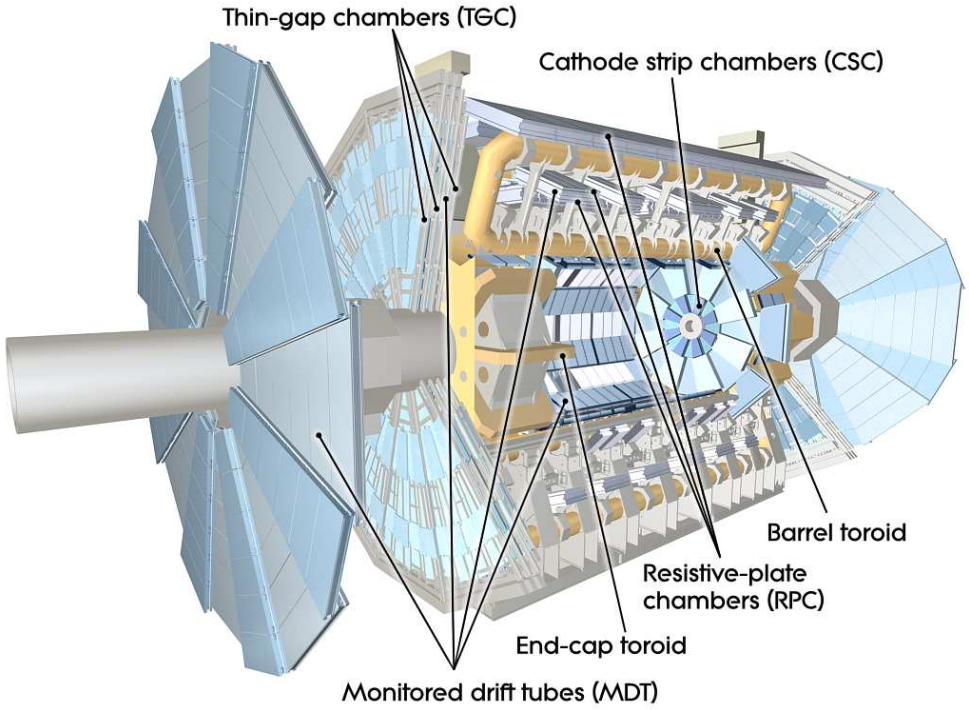


Figure 25: Cut-away of the ATLAS muon system. From Ref. [27].

Tracking chambers

The MDT subdetector is composed of three to eight layers of drift tubes with gas mixture of Ar (97%) and CO_2 (3%). It covers the range $|\eta| < 2.7$. The MDT chambers feature very good spatial resolution of 80 μm per tube and 35 μm per chamber, but their counting rate is limited to 150 Hz/cm^2. Therefore, in the innermost tracking layer in the forward region ($2 < |\eta| < 2.7$) CSCs are used that provide higher rate capability - up to 1000 Hz/cm^2, and better time resolution - their maximum drift time for signal collection is 40 ns compared to 700 ns for the MDTs.

The CSCs are proportional chambers using the same gas mixture as the MDT with multiple anode wires oriented in the radial direction and cathodes segmented into strips in the orthogonal direction. This allows to reconstruct muons tracks in (η, ϕ) space. The whole CSC system is composed of two disks with eight chambers per disk. Each chamber

consists of four planes, which allows to perform four independent measurements in η and ϕ along each track. The spatial resolution of the CSCs is 40 μm in the bending η-plane and 5 mm in the non-bending ϕ-plane.

Triggering chambers

The triggering chambers of the MS are designed to provide to the trigger system fast and coarse information on muon tracking. In addition, the goal of these chambers is to perform track measurements in the non-bending ϕ-plane additionally to those performed by tracking chambers.

The RPCs are placed in the barrel and cover the range $|\eta| < 1.05$. They consist of two parallel electrode plates (no wires are used), with a 2 mm gap between filled with a gas mixture of $C_2H_2F_4$/Iso-C_4H_{10}/SF_6. An electric field of 4.9 kV/mm is formed between the plates that causes the formation of avalanches along the ionising tracks. The RPCs allow a good timing resolution of 1.5 ns.

The TGCs are placed in the end-cap wheels, covering the region $1.05 < |\eta| < 2.4$. They are based on a multiwire proportional chamber technique with the copper wires oriented in the radial direction and carbon strips oriented in the ϕ direction. Each chamber is filled with a gas mixture of 55% CO_2 and 45% n-C_5H_{12}. The TGCs feature a wire-to-cathode distance smaller than the wire-to-wire distance (1.4 mm compared to 1.8 mm), which allows very fast signal collection in order to achieve a time resolution of 4 ns.

2.2.6 Trigger system

The high luminosity of the LHC requires fast and effective selection of events that are interesting for physics. This is performed by the three-level ATLAS trigger system. The scheme of ATLAS trigger and data-acquisition (DAQ) system is presented in figure 26.

The Level-1 trigger (L1) is based on hardware: the selection of events is performed by logical electronics. It decreases event rate from 40 MHz (the bunch-crossing rate) to $\sim$ 100 kHz. The L1 trigger uses information from coarse-granularity calorimeter and muon spectrometer. The decision is taken based on the E_T or p_T threshold and the multiplicity of physical objects registered in detectors: electrons, muons, photons, jets, hadronic tau and E_T^{miss}.

The Level-2 (L2) and Level-3 (L3, event filter) triggers are based on software. In Run 1 the L2 trigger reduced the rate of events down to 2-3 kHz and then the L3 was making the final decision, decreasing the rate of events down to 300-400 Hz. In Run 2 these two triggers are merged into a High Level Trigger System (HLT) farm that uses multivariate analysis techniques. The new approach and various optimisations of the trigger system allowed to simplify and speed up the selection process, which is crucial for high luminosities at Run 2. Using the full event information from different detectors the HLT reduces the event rate down to about 600 Hz to 1.5 kHz. A single-electron trigger with a p_{T} threshold of 24 GeV has a peak event rate of 18 kHz at L1 and $\sim$ 140 Hz for the HLT.

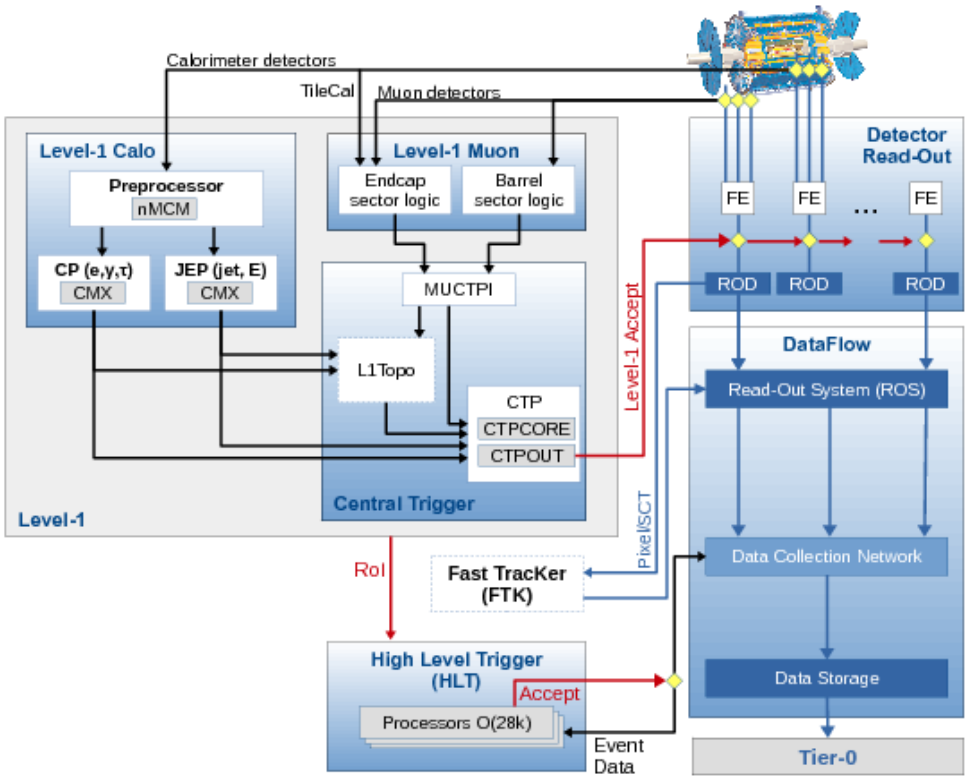


Figure 26: The scheme of ATLAS trigger and data-acquisition (DAQ) system. From Ref. [40].

2.3 Event simulation

The simulation of various physical processes that is used for ATLAS analyses is performed based on theoretical predictions via the so-called Monte Carlo (MC) method, based on the generation of pseudo-random numbers to sample variables governed by complex probability density functions. A physics analysis is based on the comparison of theory with experimental data, therefore it is very important to make a good choice of MC modelling for the particular processes considered in this analysis.

The full scheme of ATLAS simulation software is summarised in figure 27.

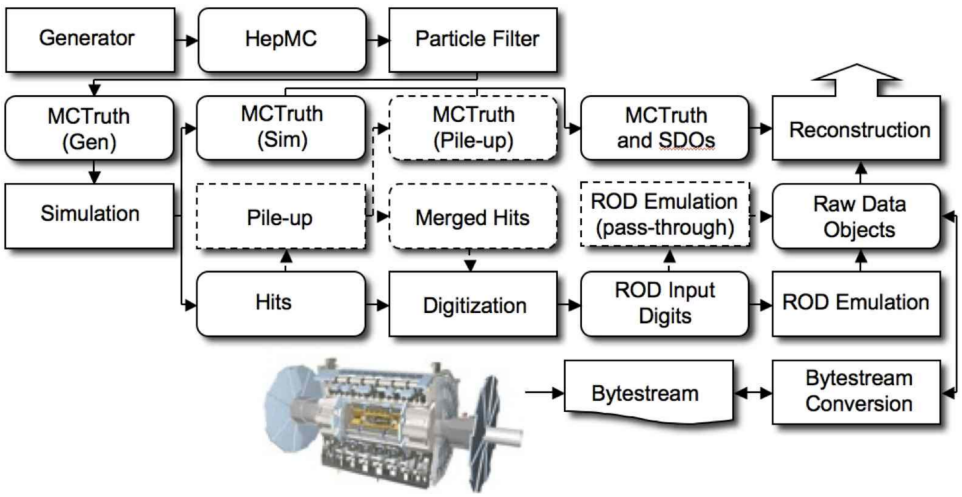


Figure 27: The flow of the ATLAS simulation software from event generators (top left) to reconstruction (top right). Algorithms are placed in square-cornered boxes and persistent data objects are placed in rounded boxes. The optional pile-up portion of the chain, used only when events are overlaid, is dashed. Generators are used to produce data in HepMC format. Monte Carlo truth information is saved in addition to energy depositions in the detector (hits). This truth information is merged into Simulated Data Objects (SDOs) at the digitisation step. Read Out Driver (ROD) electronics are also simulated during the digitisation. From Ref. [41].

The final step of the chain is the reconstruction process, that builds physics objects such as tracks, and calorimeter energy clusters. This procedure is applied in the same way to the data and to simulated events, using digitised information on the hits (real hits in the case of data and simulated hits in the case of MC events) in subdetector systems.

2.3.1 Event generation

The simulation of pp collisions needs the description of physics processes at very different energy scales. According to the factorisation theorem, several subprocesses of a physical event that happen at different energy scales can be separated. Therefore the simulation of a physical process can be performed in several steps: hard scattering, parton

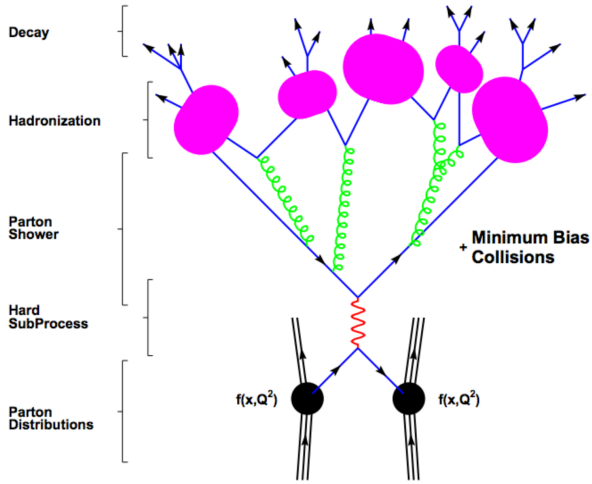


Figure 28: The basic structure of a pp collision event simulation. From Ref. [42].

shower and hadronisation. A visualisation of the different steps of the event generation is shown in figure 28.

The momenta of the hard-scatter partons inside the colliding protons are obtained from the PDFs, that determine the probabilities to find a parton of a certain type carrying a certain fraction of the proton momentum. The partonic content of the proton cannot be described via perturbative QCD, and thus PDFs are obtained from fits to experimental data from deep-inelastic scattering experiments and hadron colliders.

The information on the kinematics and flavour of partons is then used for the evaluation of the cross section for the hard-scatter process in fixed order perturbation theory, known as matrix element (ME) calculation. This part of the simulation in ATLAS is performed by so-called *parton-level generators* (or *matrix element generators*). The parton-level generators used in the analysis presented in this thesis are MadGraph5_aMC@NLO [43], Powheg-Box [44] and Sherpa [45].

The second step in the event generation is the modelling of the parton shower process: partons radiate gluons, which can then further split into other gluons or quark-antiquark pairs. A parton shower generator provides a higher order correction to the matrix element calculation due to this radiation: it simulates the emission of quarks and gluons from the partons in the final or initial state.

The third step is the hadronisation or fragmentation, when the partons in the shower, which have reached non-perturbative energy scales, form hadrons that further decay. The hadronisation process is simulated using phenomenological models.

For simulation of the parton showering and hadronisation for the presented analysis Pythia [46, 47] and Sherpa [48] are used.

The pile-up and the *underlying events* (interactions of the remaining proton constituents that did not participate in the hard scattering) are low energy scale processes, and are simulated using phenomenological models with parameters derived from experimental data.

To summarise, the output of MC generators is a set of four-vectors of the particles

at the step of reaching the detector: after decay and hadronisation of most of the unstable particles (except long-lived particles that decay in the detector material), but before interaction with the detector material.

2.3.2 Detector simulation

The final step after generating events is the simulation of the geometry of the detector, the interaction of particles with the detector material and the detector response.

The so-called full detector simulation is performed via Geant4 [49], that reproduces the interaction between the particles and detector material, resulting in hits in the detector. After that, hit digitisation is performed and the detector response is simulated.

Full simulation consumes much computational power, which implies that it is impossible to obtain the required MC statistics for many physics studies. The alternative fast simulation approach can be used instead in order to save computing resources. The fast simulation that is used in various ATLAS physics analyses is called ATLFAST-II. The reduction of simulation time is achieved by simplifying the detector description used for simulation, mostly for the calorimeters. ATLFAST-II provides the same output as the full simulation, and is used in physics analyses in those cases with not enough MC statistics with full simulation [41].

2.4 Event reconstruction

The events that pass trigger selection are further processed to better identify the detected particles and reconstruct their properties. Various algorithms are used to reconstruct physics objects, such as photons, electrons, muons, taus and jets, using tracks and energy clusters reconstructed from signals in the detector.

2.4.1 Tracks

Charged particles are bent in the solenoid magnetic field of the inner detector, obtaining a curvature inversely proportional to their momenta.

A track in the magnetic field is a helix described by five parameters (d_0, z_0, ϕ, θ, q/p_{T}), as shown in figure 29:

- The signed transverse impact parameter d_0 is the closest distance between the track and the beam axis in the transverse plane.
- The signed longitudinal impact parameter z_0 is the z coordinate of the track at the point of closest approach.
- The azimuthal angle ϕ is the angle between the transverse momentum vector and the x-axis in the transverse plane, $0 < \phi < \pi$.
- The polar angle θ is the angle between the momentum vector $\vec{p}$ and the z-axis in the R-z plane, $0 < \theta < \pi$.

- The electric charge over the transverse momentum q/p_{T} is obtained using the track curvature radius R_{curv}. The dependence of q/p_{T} on R_{curv} is given by

$$q/p_{\mathrm{T}} = (0.3BR_{\mathrm{curv}})^{-1}, \tag{38}$$

  where B is the magnetic field. B, q, p_{T}, and R_{curv} are measured in units of [T], [e], [GeV], and [m], respectively.

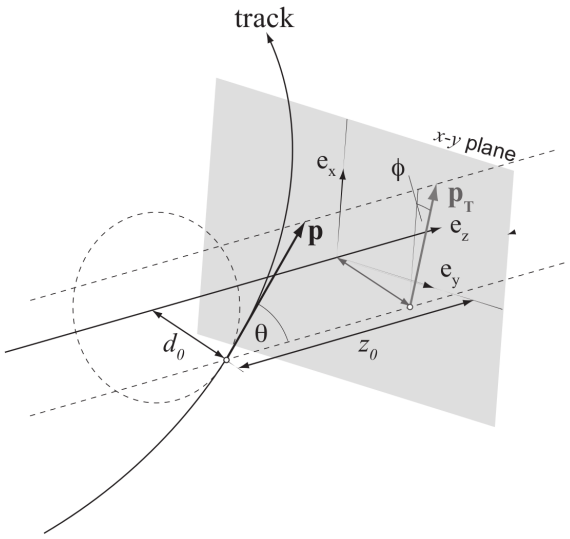
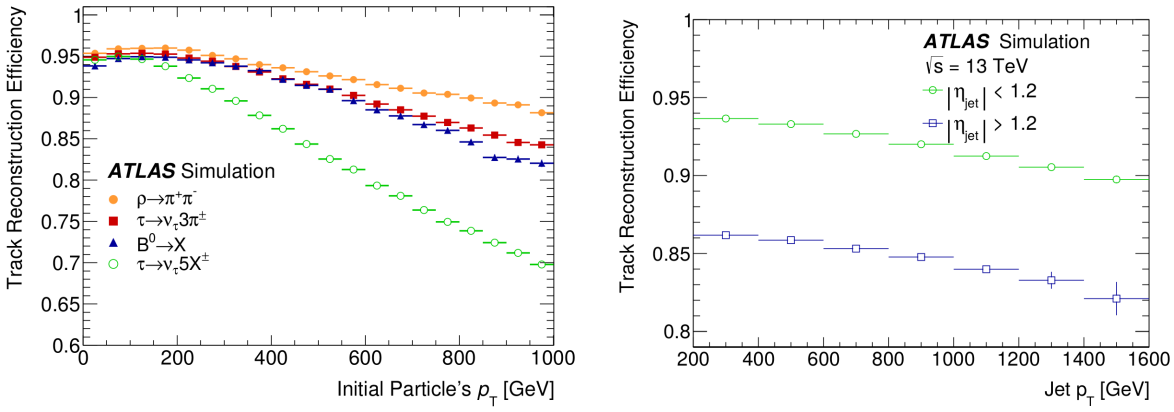


Figure 29: Representation of track parameters. From Ref. [50]

Track reconstruction is performed in two steps: track finding, or pattern recognition, and track fitting.

There are several track finding algorithms used in ATLAS. The *inside-out* pattern recognition algorithm first builds track seeds in the pixel and SCT detectors and then extends the track candidate to the TRT. Most of the tracks used in physics analyses are found with this algorithm. The *outside-in*, or *back-tracking* algorithm starts with segments finding in the TRT, which are then extrapolated back to the pixel and SCT detectors. Within the track finding procedure a combinatorial Kalman filter [51] is used to build track candidates from the chosen seeds. At this stage multiple track candidates per seed are considered if more than one compatible track extrapolation exists on the same layer.

When track candidates are built, an ambiguity solving procedure is applied, which identifies the best track candidates taking into account several quality parameters with the help of a neural network (NN). Track candidates are required to satisfy minimum quality criteria, such as $p_{\mathrm{T}} > 400$ MeV, $|\eta| < 2.5$, at least 7 total hits in total in the pixel detector and SCT (out of 12 expected), $|d_0^{\mathrm{BL}}| < 2.0$ mm, $|z_0^{\mathrm{BL}}| < 3.0$ mm (d_0^{BL} and z_0^{BL} are the transverse and longitudinal impact parameters calculated with respect to the measured beam-line position). Additional requirements are made on the number of *shared*



(a) (b)

Figure 30: Track reconstruction efficiency in the simulation for (a) collimated charged particles in various decay channels as a function of the initial particle p_{T}, and (b) tracks within a jet as a function of jet p_{T} for dijet events. From Ref. [52].

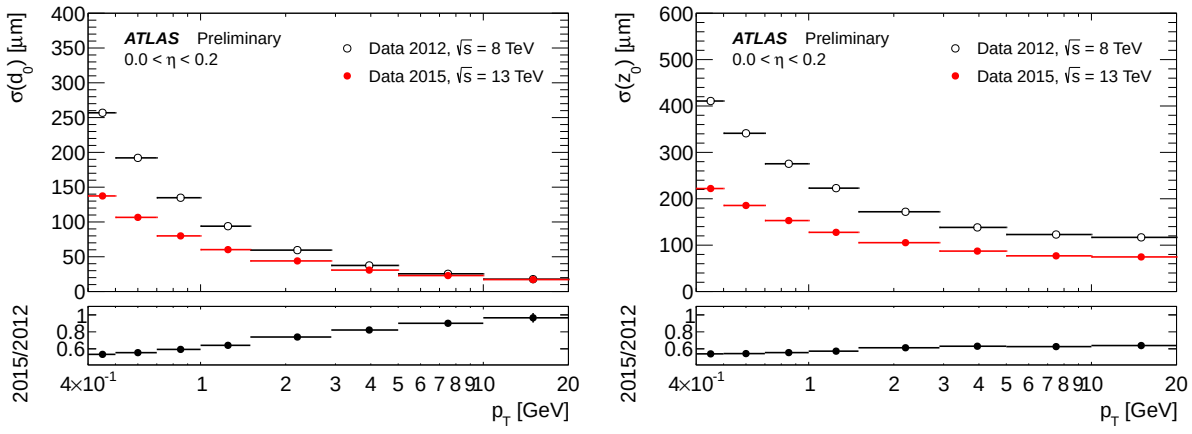
hits (clusters which are shared among several tracks) and *holes* (point of the reconstructed track trajectory with a sensitive detector element that does not contain a matching cluster) [52]. For track candidates that pass the ambiguity solving procedure a high-resolution fit is performed, which allows to measure the track parameters. The track reconstruction efficiency depends on the physical process, as well as on the kinematics of the considered charged particles. The reconstruction of charged particles inside jets is challenging due to the high track densities. Figure 30 shows the track reconstruction efficiencies in the case of collimated charged particles produced in different decays modes and in the case of tracks inside jets.

The most important detector upgrade in Run 2 affecting the performance of tracking is the installation of the IBL. The Run 2 to Run 1 data comparison showed that the IBL significantly improves the track impact parameter resolution: by a factor of two for both transverse and longitudinal components in the case of low-p_{T} tracks ($p_{\mathrm{T}} < 1$ GeV). For a typical track with $p_{\mathrm{T}} = 2$ GeV, the transverse (longitudinal) impact parameter resolution is currently ~ 30 μm (~ 80 μm) [53]. Figure 31 shows the Run 2 and Run 1 impact parameter resolution as function of track p_{T}.

2.4.2 Vertices

Extrapolating the tracks allows to determine the position of the point of the initial interaction of protons known as the primary vertex (PV). But due to the high number of protons per bunch crossing, several vertices can be reconstructed in the same event.

Vertex reconstruction consists of two steps: vertex finding and vertex fitting. First of all, the vertex finding algorithm selects vertex seeds by looking at the local maximum in the distribution of z_0 of the tracks. The second step is the vertex fitting algorithm, that specifies the position and uncertainty of the primary vertex. It takes as input the seed position and the tracks around it and performs a χ^2 fit. Those tracks that are not



(a) (b)

Figure 31: Unfolded (a) transverse and (b) longitudinal impact parameter resolution measured from data in 2015, $\sqrt{s} = 13$ TeV, with the inner detector including the IBL, as a function of p_{T}, in the region $0 < \eta < 0.2$ compared to that measured from data in 2012, $\sqrt{s} = 8$ TeV, without the IBL. From Ref. [53].

.

likely to originate from the vertex are not rejected, but downgraded. The fit procedure is repeated several times, and outlying tracks are progressively downgraded at each next iteration. Tracks that are incompatible with the vertex by more than 7σ are used to seed a new vertex [54].

The vertex with highest sum of the squared p_{T} of its tracks is considered as corresponding to the hardest process in the events; other vertices are assumed to be pile-up interactions.

The vertex reconstruction efficiency is defined as the ratio between events with a reconstructed vertex and events with at least two reconstructed tracks. The vertex reconstruction efficiency measured in low-μ 2015 data is shown in figure 32. The corresponding measured vertex position resolution is shown in figure 33.

It is not always easy the identify the hard-scatter vertex, considering the rest as pile-up interactions. The vertex is referred to as *matched* if tracks identified as originating from the same generated interaction contribute at least 70% of the total weight of tracks fitted to the reconstructed vertex. A *merged* vertex is formed by tracks coming from two or more different interactions. In this case when there is no single interaction that contributes more than 70% of track weight to the vertex. Finally, a *split* vertex describes the case when the generated interaction with the largest contribution to the vertex is also the largest contributor to one or more additional vertices. The vertex position resolution degrades for events with merged vertices [56].

Vertices displaced from the beam collision region are referred to as secondary vertices. The secondary vertex reconstruction algorithm that is used for the identification of b-jets is described in section 3.

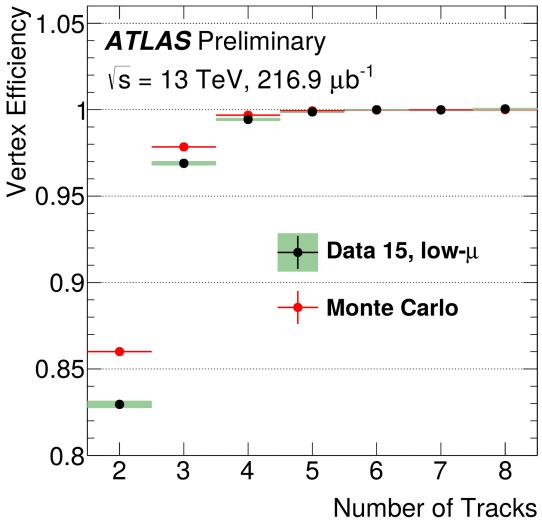
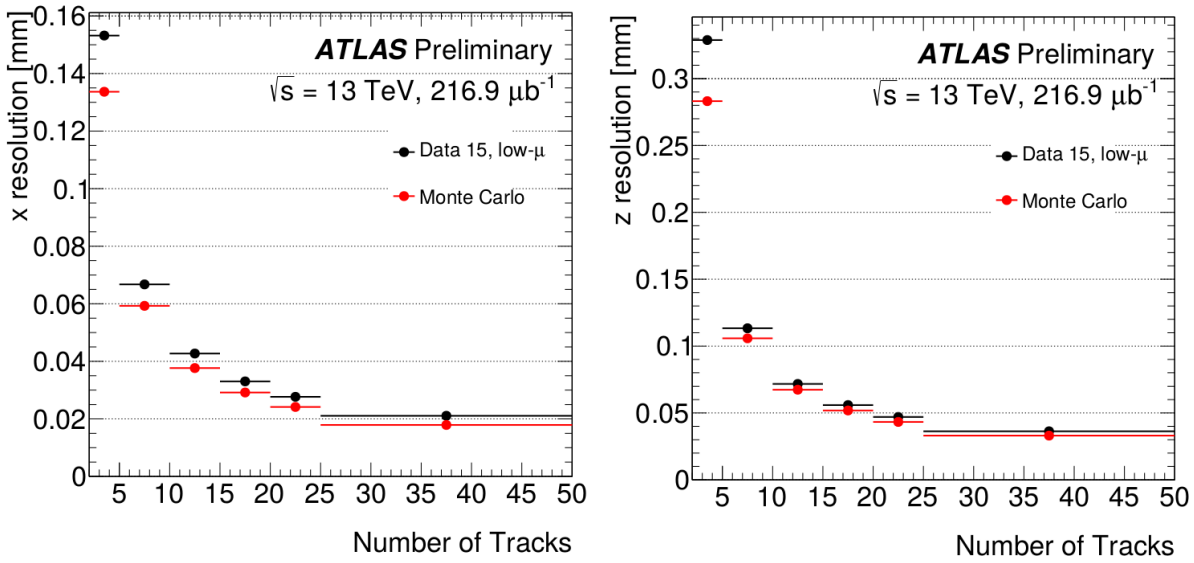


Figure 32: Vertex reconstruction efficiency as a function of the number of tracks in low-μ 2015 data. From Ref. [55].



(a) (b)

Figure 33: Vertex (a) x- and (b) z-resolution as a function of the number of tracks in low-μ 2015 data. From Ref. [55].

2.4.3 Electrons

Electron identification is challenging, as hadronic jets and non-prompt electrons (converted photons) can mimic their signatures in the detector.

Electron reconstruction

Electrons are reconstructed combining information from the ID and the EM calorimeter. An electron candidate is built using information on clusters in the EM calorimeter that are matched to tracks from the ID. Electron candidates that are not matched to a track are removed and considered to be photons.

The electron reconstruction procedure consists of several steps, which are discussed below.

- Cluster reconstruction is performed via the so-called *sliding window* algorithm. The first step of this algorithm is known as tower building. The calorimeter space is split in squares 0.025×0.025 in the $\eta - \phi$ plane, and for each of them the energy deposits in all EM calorimeter layers are summed. After that a rectangular window of 0.025×0.025 scans across the elements of towers in order to find seeds with total cluster transverse energy above 2.5 GeV. Then the clusters are reconstructed around the seeds using a clustering algorithm.

- For track reconstruction the pattern recognition algorithm searches for a track seed, consisting of three hits in different layers of the silicon detectors, with $p_{\mathrm{T}} > 1$ GeV that can be successfully extended to a full track of > 7 hits, that is matched to a given EM cluster region of interest. (A region of interest is usually defined as a cone-size of $\Delta R = 0.3$ around the seed cluster barycentre). This procedure is applied one or two times per track. In a first step, the search is performed with the standard ATLAS pattern recognition that uses the pion hypothesis for energy loss due to interactions with the detector material. If no tracks that are consistent with the pion hypothesis are found, the same procedure, but considering the electron hypothesis instead, is performed. The next step is a χ^2 fit of the track candidates (with either pion or electron hypothesis according to the one used in the pattern recognition step). If the fit with the pion hypothesis fails for an electron candidate, the second attempt is done with the electron hypothesis. This approach of using two hypotheses allows making the electron track reconstruction complementary to the main track reconstruction procedure, without rerunning the algorithm for all electron track candidates. At the same time, the electron performance is improved due to the usage of the electron-based algorithm for those tracks that cannot be reconstructed via the procedure based on the pion hypothesis.

- Electron track fit. The electron track candidates that have ≥ 4 hits are matched to EM clusters in the calorimeter in the η-ϕ plane under loose requirements, additionally taking into account energy loss due to bremsstrahlung. The fit is performed for those track candidates that pass the matching criteria. Next, the matching procedure is repeated with the refit tracks under stricter criteria.

For Run 2 electron track candidates are required to be compatible with the hard-scatter primary vertex, in order to reduce the background from conversions and products of long-lived particle decays, as well as pile-up interactions. Therefore additional requirements on track impact parameters are made: $d_0/\sigma_{d_0} < 5$, and $|\Delta z_0 \sin\theta| < 0.5$ mm, where d_0 and z_0 are transverse and longitudinal impact parameters, θ is the polar angle of the track (see definitions in section 2.4.1) and σ_{d_0} represents uncertainty on d_0. The quantity d_0/σ_{d_0} is referred to as d_0 significance.

Electron identification

Electron identification algorithms are developed to reject misidentified electrons from hadronic jets as well as non-prompt electrons mostly originating from photon conversions and heavy flavour hadron decays.

The baseline algorithm for electron identification in Run 2 is based on the likelihood approach. It considers several properties of electron candidates using signal and background probability density functions. The input variables for this algorithm are quantities related to the electron cluster and track measurements including calorimeter shower shapes, information from the TRT, track-cluster matching related quantities, track properties and variables measuring bremsstrahlung effects. In Run 2 the number of IBL hits is also used as input, as it helps to discriminate between electrons and converted photons. The full list of variables can be found in Ref. [57].

The signal and background probabilities for an electron candidate are evaluated and combined in the final likelihood discriminant:

$$d_{\mathcal{L}} = \frac{\mathcal{L}_S}{\mathcal{L}_S + \mathcal{L}_B}, \tag{39}$$

where signal and backround probabilities are computed as

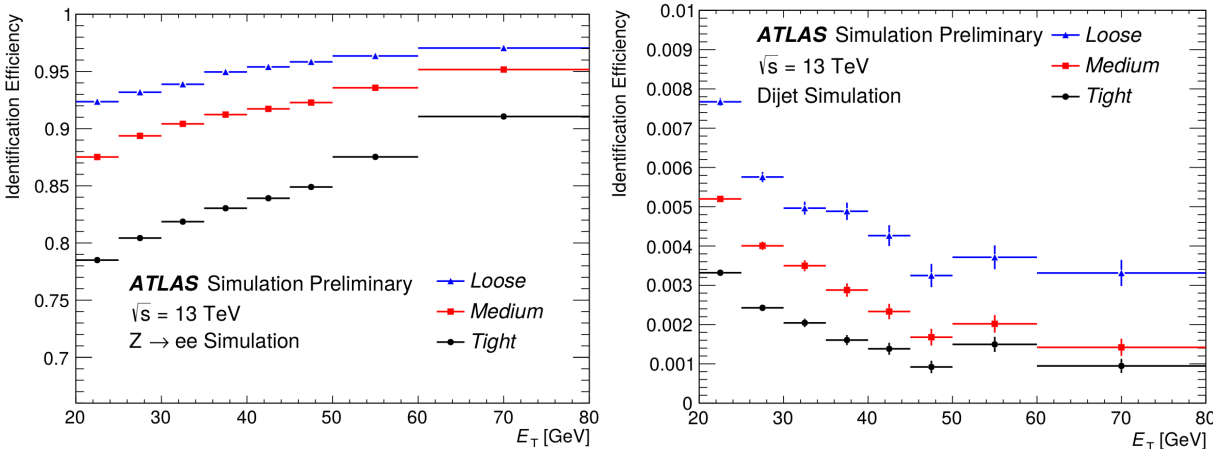
$$\mathcal{L}_{S(B)}(\vec{x}) = \prod_{i=1}^{n} P_{s(b),i}(x_i). \tag{40}$$

In the last equation $\vec{x}$ is the vector of discriminating variable values and $P_{s(b),i}(x_i)$ is the value of the signal (background) probability density function of the i^{th} discriminating variable.

Based on the distribution of the final discriminant $d_{\mathcal{L}}$, three working points are defined, in order of increasing background rejection: Loose, Medium, and Tight. The electron identification efficiency and efficiency of background identification (hadrons identified as electrons) as functions of E_T for these three working points are presented in figure 34.

Electron isolation

To further separate prompt electrons originating from heavy resonance (W, Z, H) decays from the background (converted photons produced in hadron decays, electrons from heavy flavour hadron decays and light hadrons misidentified as electrons) they may



(a) (b)

Figure 34: Efficiency of (a) prompt electron identification obtained from $Z \to ee$ decays and (b) efficiency to identify hadrons as electrons from simulated dijet samples. The efficiencies are obtained using MC simulations, and are measured with respect to reconstructed electrons. From Ref. [57].

be required to be isolated from other activity in the detector. There are two isolation criteria: track-based and calorimeter-based.

The track-based isolation exploits the $p_{\mathrm{T}}^{\mathrm{varcone0.2}}$ variable, which is defined as the sum of transverse momenta of tracks that satisfy certain quality requirements and point to the hard-scatter PV, within a cone $\Delta R = 0.2$ around the electron track. Electron tracks as well as tracks from the converted bremsstrahlung photons are not considered in the calculation.

The calorimeter requirement is based on a variable called calorimetric isolation energy, $E_{\mathrm{T}}^{\mathrm{cone0.2}}$, that is defined as the sum of transverse energies of reconstructed calorimeter clusters with positive energy within a cone of $\Delta R = 0.2$ around the cluster formed by the considered electron. The contributions within a window of size $\Delta\eta \times \Delta\phi = 0.125 \times 0.175$ around the electron cluster barycentre are subtracted [57]. An additional correction for pile-up and the underlying event activity is applied.

2.4.4 Muons

For the reconstruction of muons, measurements from the MS are combined with those from the ID. Information from the EM calorimeter is additionally used.

Muon reconstruction in the MS

The first step of the muon reconstruction in the MS is the search for hit patterns in each of the muon chambers and building segments out of them. The muon track candidates are obtained by fitting together hits from segments in different detector layers. The minimum requirement to build a track is two matching segments, except in the barrel-

endcap transition region, where one segment is allowed to be used. Then hits associated with a track candidate are fitted with a χ^2 fit.

Combined reconstruction

The track measurement in the MS is combined with information from the ID and the EM calorimeter. Five types of reconstructed muons are used in ATLAS, depending on which subdetectors are used in the reconstruction:

- Standalone muons are those reconstructed only in the MS.
- Combined muons use information from both MS and ID.
- Calorimeter-tagged muons are muons detected by the calorimeter and ID, without using information from the MS. They are used for the region $|\eta < 0.1|$, that is uncovered by the MS.
- Segment-tagged muons correspond to the case when the track from the ID is matched to a segment of a track in the MS.
- Extrapolated muons are built using the tracks reconstructed using hits in MS detectors only, but with additional requirement on compatibility with originating from the interaction point.

Combined muons is the best quality muon type: they have the highest fake muons rejection and the best momentum resolution.

Muon identification

After muons are reconstructed, the next step is to distinguish prompt muons from those originating from decays of charged hadrons, mostly from pion and kaon decays. For combined muons several discriminating variables are used:

- q/p significance, defined as the absolute value of the difference between the ratio of the charge q and momentum p of the muons measured in the ID and MS, divided by the sum in quadrature of the corresponding uncertainties;
- ρ', defined as the absolute value of the difference between the p_{T} in the ID and MS divided by the p_{T} of the combined track;
- normalised χ^2 of the combined track fit.

Based on distributions of the above discriminating variables and additional requirements on the number of hits in the ID and MS, four muon selections are defined: Loose, Medium, Tight and High-p_{T}. The muon reconstruction efficiencies for these operating points are evaluated using a $t\bar{t}$ MC sample. Table 6 summarises the muon identification efficiencies of signal (muon candidates from W-boson decays) and background (muon candidates from light-hadron decays) in different p_{T} ranges [58].

	$4 < p_{\mathrm{T}} < 20$ GeV		$20 < p_{\mathrm{T}} < 100$ GeV	
Selection	$\epsilon_{\mu}^{MC}[\%]$	$\epsilon_{\mathrm{Hadrons}}^{MC}[\%]$	$\epsilon_{\mu}^{MC}[\%]$	$\epsilon_{\mathrm{Hadrons}}^{MC}[\%]$
Loose	96.7	0.53	98.1	0.76
Medium	95.5	0.38	96.1	0.17
Tight	89.9	0.19	91.8	0.11
High-p_{T}	78.1	0.26	80.4	0.13

Table 6: Efficiencies of identifying prompt muons from W-boson decays (signal) and in-flight decays of hadrons misidentified as prompt muons (background) computed using a $t\bar{t}$ MC sample. The results are shown for the four muons selections (Loose, Medium, Tight and High-p_{T}) for low and high p_{T} muon candidates with $|\eta| < 2.5$. From Ref. [58].

Muon isolation

Muons may be required to be isolated from other detector activity in order to effectively distinguish those produced in heavy resonance decays from the background from heavy-flavour decays inside jets. As for electrons, track and calorimeter variables can be used as isolation criteria.

The track-based isolation variable is $p_{\mathrm{T}}^{\mathrm{varcone30}}$, defined as the sum of the transverse momenta of the tracks with $p_{\mathrm{T}} > 1$ GeV in a cone of size $\Delta R = \min(10\,\mathrm{GeV}/p_{\mathrm{T}}^{\mu}, 0.3)$ around the muon track, which is not considered in the calculation. The p_{T}-dependence of the cone size allows to improve the performance for muons from decays of high-p_{T} particles.

The calorimeter isolation exploits a variable calorimetric isolation energy, $E_{\mathrm{T}}^{\mathrm{topocone20}}$, defined as the sum of the transverse energy in calorimeter clusters in a cone of size $\Delta R = 0.2$ around the muon. Contribution from the energy deposit of the muon itself is subtracted. An additional correction for pile-up effects is also applied [58].

2.4.5 Jets

Quarks and gluons produced in pp collisions hadronise and create collimated bunches of particles, known as *hadronic jets*. They are experimentally observed as clusters of energy deposits in the calorimeter system that can be associated with charged particle tracks in the inner detector.

Jets that originate from b-quark (b-jets) can be discriminated from other types of jets using special properties of b-hadrons. Identification of b-jets, or b-tagging, is important for many physical analyses and in particular for that presented in this dissertation. It is described in detail in section 3.

Jet reconstruction

Reconstruction of a jet allows to measure the momentum of the initial parton as each of final state particles carries some fraction of it.

First of all, the topological clustering algorithm reconstructs the energy deposit clusters in the calorimeter (known as *topo-clusters*). The algorithm finds a seed cell with a signal-to-noise ratio above the threshold $|S/N| \geq 4$. The noise can be of electronic or pile-up origin. Then the cells surrounding the seed are iteratively attached to the seed if they satisfy the requirement $|S/N| \geq 2$. Finally, the cells with $|S/N| \geq 0$ lying on the perimeter of the resulting cluster are also included. If there is more than one energy local maximum in a cluster, it can be split into several sub-clusters. This helps to separate deposits that are close, but originate from different particles.

After clusters are built, the jet-finding anti-k_T algorithm [59] reconstructs jets from topological clusters. This procedure is characterised by a parameter R that sets the approximate size of jets. An example of jets clusters from a MC simulated parton-level event reconstructed with the anti-k_T algorithm with $R = 1$ is shown in figure 35. The analysis presented in this dissertation uses anti-k_T jets with $R = 0.4$.

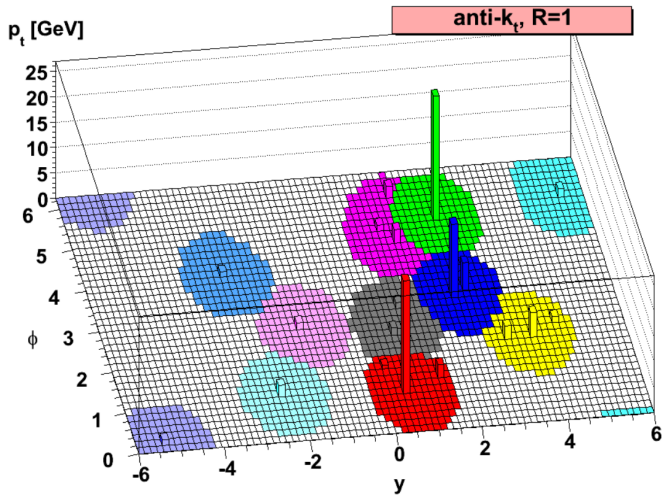


Figure 35: Clusters from a MC simulated parton-level event reconstructed with the anti-k_T algorithm. From Ref. [59].

Jet calibration

After jets are reconstructed, they need to be calibrated to match the predictions of so-called *truth jets* (jets that are reconstructed not from calorimeter energy deposits, but from truth stable particles in MC samples). The full chain of the jet calibration procedure in Run 2 is illustrated in figure 36.

Calorimeter clusters are initially calibrated to the energy scale of electromagnetic showers. After the jet four-momentum is reconstructed from the clusters, corrections are applied to take into accounts various effects.

First of all, the *origin correction* sets the jet direction so that it points to the hard-scatter PV rather than the center of the detector, at the same time preserving the measured jet energy. This correction improves the η resolution of jets.

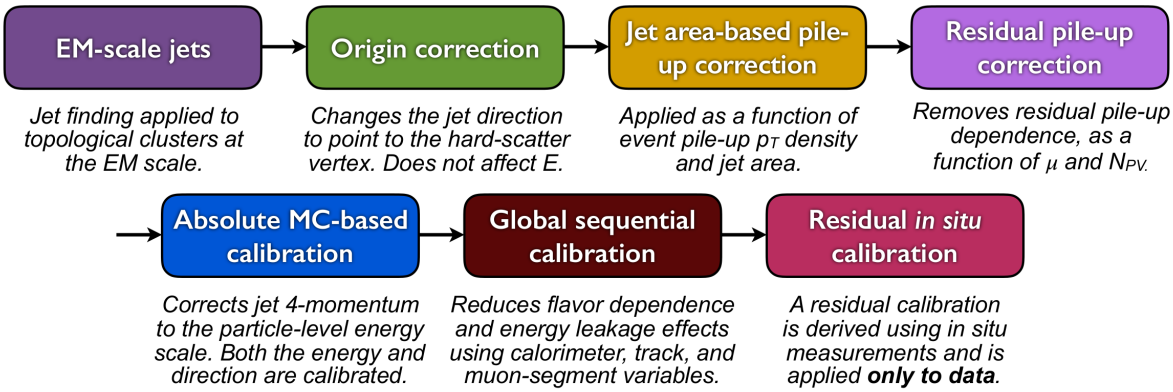


Figure 36: Jet calibration sequence. From Ref. [60].

The next step is the *pile-up correction* that scales the jet energy taking into account the in-time and out-of-time pile-up. It consists of two components: the *area-based correction* and the *residual pile-up correction*. The area-based method calculates the pile-up contribution per event from the median p_{T} density of jets in the η-ϕ plane and then subtracts it to the p_{T} of each jet according to its area. This calculation is imperfect due to the fact that it is evaluated in the central, lower-occupancy region of the calorimeter, so it does not effectively describe the pile-up in the forward region or in the higher-occupancy core of high-p_{T} jets, and some jet p_{T} dependence on the pile-up remains after the correction. Therefore, an additional MC-based residual correction as a function of number of PV and μ is applied.

After pile-up subtraction, MC samples without pile-up added are using to derive the *absolute jet energy scale (JES) calibration*, which corrects the jet four-momentum to the particle-level energy scale (i.e. that of truth jets).

The following step is the *global sequential calibration*, which takes into account remaining differences in the distribution of energy within the jets. The particle composition and shower shape of a jet depends of the type of particle it originates from. In particular, jets initiated by quarks typically contain hadrons that carry a large fraction of jet p_{T}, and thus go further through the calorimeter, enhancing the longitudinal component of the jet. In contrast, gluon-originated jets usually consist of softer (in p_{T}) particles, therefore featuring a wider transverse profile. To take into account these effects, a sequence of jet four-momentum corrections is performed. These corrections are applied to five jet observables (variables from tracker, calorimeters and MS) to match those of truth jets, while conserving the overall energy scale.

Finally, so-called *in-situ* techniques are used to remove the remaining differences between data and MC. The principle of these methods is to compare the p_{T} of a jet with well-measured physics objects (photons and Z-bosons). This calibration is applied only to data [60].

Jet vertex tagger

The discrimination of jets originating from the hard-scattering process from those originating from pile-up activity is challenging given the high luminosities in Run 2.

In Run 1 the rejection of pile-up jets was performed by applying a *jet vertex fraction (JVF)* requirement, defined as the minimum fraction of tracks within a jet that originate from the hard-scatter PV.

For Run 2 a new technique called jet vertex tagger (JVT) is applied to identify pile-up-originated jets. This is a likelihood-based method that uses two variables: a modified version of the JVF variable and the ratio of sum of p_{T} of the tracks in the jet originating from the hard-scatter PV to the full jet p_{T}.

The pile-up events are rejected by applying a $JVT > 0.59$ requirement, that provides a 92% efficiency to select hard-scatter jets. Since the contamination of pile-up is significant for low-p_{T} jets, the requirement is applied only to jets with $p_{\mathrm{T}} < 60$ GeV and $|\eta| < 2.4$.

As shown in figure 37, at a fixed efficiency of 90% the performance of the JVF-based selection depends on the number of vertices, but is stable for the JVT-based selection [61].

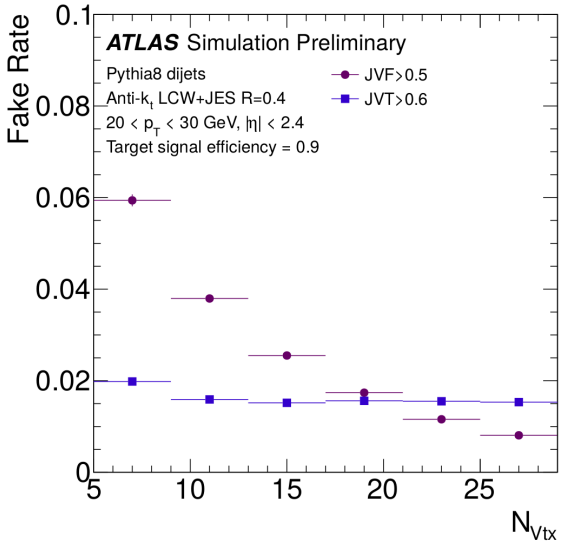


Figure 37: Pile-up jet fake rate at fixed hard scatter jet efficiency 90% as function of number of vertices for JVF and JVT. From Ref. [61].

2.4.6 Missing transverse energy

Neutrinos are not measured by the ATLAS detector. Their presence can be determined by an imbalance in the measured transverse momentum of the detected particles, referred to as missing transverse energy (MET or E_T^{miss}). Theories beyond the SM suggest the existence of additional weakly-interacting particles, so they can also contribute to the missing energy. Therefore, the $E_{\mathrm{T}}^{\mathrm{miss}}$ measurement is of great importance in many new phenomena searches.

The reconstruction of $E_{\mathrm{T}}^{\mathrm{miss}}$ takes into account energy deposits in the calorimeters and muons reconstructed in the MS. It is obtained from the negative vector sum of the momenta of all reconstructed and calibrated physics objects. Their contributions are considered in a specific order: electrons, photons, hadronically decaying τ-leptons, jets and then muons. Soft energy contributions that are not associated with any of these objects (mostly from underlying events and soft radiation) are also considered [62]. If the combined muon momentum is used, the muon energy loss in the calorimeters is subtracted in the calculation in order to avoid double counting [63].

The x and y components of $E_{\mathrm{T}}^{\mathrm{miss}}$ are defined as

$$E_{x(y)}^{\mathrm{miss}} = E_{x(y)}^{\mathrm{miss,e}} + E_{x(y)}^{\mathrm{miss},\gamma} + E_{x(y)}^{\mathrm{miss},\tau} + E_{x(y)}^{\mathrm{miss,jets}} + E_{x(y)}^{\mathrm{miss,SoftTerm}} + E_{x(y)}^{\mathrm{miss},\mu}. \tag{41}$$

The magnitude of $E_{\mathrm{T}}^{\mathrm{miss}}$ is then obtained as

$$E_{\mathrm{T}}^{\mathrm{miss}} = \sqrt{(E_x^{\mathrm{miss}})^2 + (E_y^{\mathrm{miss}})^2}. \tag{42}$$

In Run 1 the soft term $E_{x(y)}^{\mathrm{miss,SoftTerm}}$ was calculated using soft energy deposits in the calorimeter not associated with any of the reconstructed hard objects. In Run 2 the soft term is calculated from the momenta of tracks from the ID. This method is more robust against pile-up interaction contamination than the calorimeter-based approach, since tracks from pile-up vertices can be effectively discriminated [64].

The resolution in the $E_{\mathrm{T}}^{\mathrm{miss}}$ measurement in $Z \to \mu\mu$ events achieved for Run 2 is presented in figure 38. The $Z \to \mu\mu$ process is a well suited channel to evaluate the $E_{\mathrm{T}}^{\mathrm{miss}}$ as its value is expected to be zero, with the additional advantages of a small background and the possibility to precisely measure the kinematics of the Z boson.

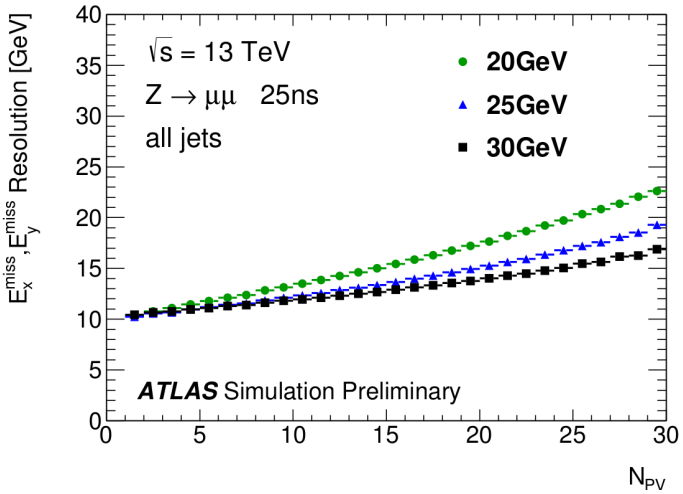


Figure 38: $E_{\mathrm{T}}^{\mathrm{miss}}$ resolution in $Z \to \mu\mu$ MC events for different values of the jet p_{T} threshold as a function of the number of PV. The E_x^{miss} and E_y^{miss} were found to have similar performance, so the two distributions were combined. From Ref. [64].

3 Identification of *b*-jets

Identifying jets originating from b-quarks, a capability known as b-tagging, is important as many physics analyses performed by the ATLAS experiment, such as SM measurements (top quark physics and Higgs physics) and searches beyond the SM, involve b-quarks in the final state. Physics processes with b-quarks in the final state are of particular interest since the b-quarks are the heaviest quarks in the SM that form hadrons. In the $t\bar{t}H$ ($H \to b\bar{b}$) many jets are expected in the final state, and four of them originate from b-quarks, so it is important to identify them.

The b-tagging performance in the ATLAS experiment was improved in Run 2 thanks to the insertion of the IBL and algorithmic enhancements in both tracking and b-tagging [65]. One of the important developments is a new track categorisation that takes advantage of the IBL addition. In this chapter an overview of b-tagging in ATLAS in Run 2 is presented. It also includes a detailed discussion on my main contribution to the optimisation of the impact-parameter-based b-tagging algorithms, via the development of a new track categorisation.

3.1 Properties of *b*-hadrons

When a b-quark is produced, it hadronises and forms a b-hadron ($B^{\pm}$, B^0, etc), that subsequently decays. A jet formed by the particles produced in the fragmentation of a b-quark and the following decay of the corresponding b-hadron is referred to as a *b-jet*.

Important properties of b-hadrons are their relatively long lifetime ($\sim$ 1.6 ps) and high mass ($\sim$ 5 GeV). A b-hadron can travel several millimetres through the detector before decaying. Consequently, the vertex of the b-hadron decay, referred to as *secondary vertex* (SV), is significantly displaced with respect to the *primary vertex* (PV) (see figure 3).

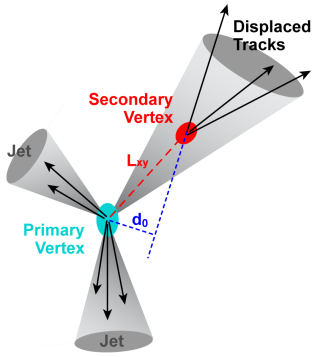


Figure 39: Sketch of a b-hadron decay, showing the most relevant variables for its identification.

The high mass of the b-hadron provides an angular difference between the direction of the initial b-hadron propagation and its decay products. All these features allow to distinguish b-jets from other jets.

A b-hadron dominantly decays to a c-hadron ($D^{\pm}$, D^{0}, etc), which also has a significant lifetime, so that the position of its decay (*tertiary vertex*) is displaced with respect to both secondary and primary vertices.

The *c-jets* have similar features to b-jets, but they are more difficult to distinguish due to relatively lower c-hadron mass and lifetime. The jets originated from u, d, s quarks and gluons are referred to as *light jets* in the following.

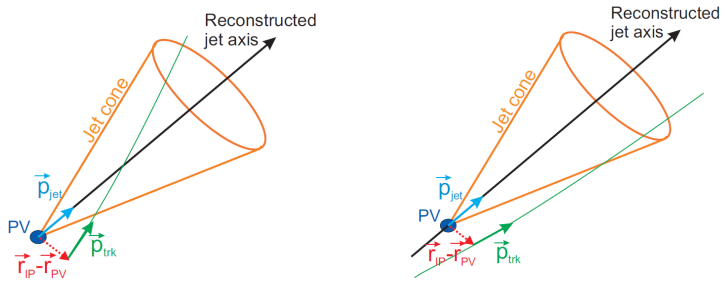
3.2 Key *b*-tagging ingredients

3.2.1 Impact parameter

The position of a displaced track is described by two parameters:

- d_0, the transverse impact parameter, which is the distance of the closest approach of the track to the PV in the transverse plane. It is schematically shown in figure 39.
- z_0, the longitudinal coordinate of the track at the point of closest approach to the PV. The longitudinal impact parameter is defined as $z_0 \sin\theta$.

The sign of the impact parameter is defined in a different way from the one used for the track description presented in section 2.4.1. It is positive (negative) if track crosses the jet axis in front of (behind) the primary vertex with respect to the jet direction, as shown in figure 40.



(a) (b)

Figure 40: Definition of the sign of the impact parameter d_0. When track (or its extrapolation) crosses the jet axis in front of the PV (a) - sign is positive, and behind PV (b) - negative. From Ref. [66].

To give more weight to well-measured tracks, the variable that is taken into account in b-tagging is the impact parameter significance:

$$S_{d_0} = d_0/\sigma(d_0), \quad S_{z_0} = z_0/\sigma(z_0), \tag{43}$$

where $\sigma(d_0)$ and $\sigma(z_0)$ denote the uncertainties on the d_0 and z_0 measurements.

3.2.2 Vertices

Vertex reconstruction plays a key role in b-tagging. First of all, the correct choice of the PV and the precise measurement of its position is crucial. Then the SV needs to be efficiently reconstructed. One of the essential SV parameters used for b-tagging is the distance between the PV and SV in the transverse plane, referred to as the *decay length* L_{xy} (see figure 39).

The performance of vertex reconstruction can be evaluated by defining the reconstruction *rate*:

$$R_{\text{vertex}} = \frac{N_{\text{events}}^{\text{vertex}}}{N_{\text{events}}^{\text{total}}}, \tag{44}$$

where $N_{\text{events}}^{\text{vertex}}$ is the number of events with a vertex successfully reconstructed and $N_{\text{events}}^{\text{total}}$ is the total number of events.

The SV and JetFitter algorithms, described in sections 3.4.1 and 3.4.2, are the main ATLAS SV reconstruction algorithms.

3.2.3 Track quality criteria

To achieve better separation between b-jets and light jets, tracks are required to satisfy certain quality criteria. A minimum track p_{T} requirement is used to reject low-momentum tracks that are not well reconstructed because of multiple scattering, as well as those originating from the pile-up activity. A selection based on $|d_0|$ and $|z_0 \sin\theta|$ is used to exclude tracks from long-lived particles such as K_s, Λ etc. These particles degrade the identification of b-jets as they also form SVs with tracks displaced with respect to the PV, but their lifetime is longer than that from a b-hadron; therefore, they can be rejected with an impact parameter requirement.

Other selection criteria are necessary to reject poorly reconstructed tracks. Those are tracks with hits missing in some detector layers, as well as tracks with ambiguities in pattern recognition. Several characteristics that are used to evaluate the quality of a track are defined below.

- *Shared hits* denote detector clusters that are shared among more than one track. They degrade track reconstruction as their resulting position is shifted with respect to the points where the particles actually crossed the detector layer.

- *Split hits* are the clusters shared among more than one track that have been identified with help of a neural network (NN) as originating from different particles and have therefore been split into sub-clusters. The split and shared categories are exclusive: hits that are identified as split by the NN belong to the split category and not to the shared category.

- *Expected* hits and *holes* are determined using the track extrapolation. A hit in a pixel detector layer is expected when the curve obtained by a helix extrapolation of the hits in other pixel and SCT layers crosses a working module of the given layer.

A hole is a missing hit in a given layer, while it is expected from interpolation of hits in other layers including the hits in the two closest surrounding layers. Holes are not defined in the first layer of the pixel detector, i.e. the IBL, and the last SCT layer. Unactive sensors, so-called *dead modules*, and insensitive detector regions, such as edge areas on the silicon sensors, are not considered as holes and hits are not expected there.

The list of track quality requirements for b-tagging purposes in Run 2 is presented in table 7. The selection criteria vary for different b-tagging algorithms. For instance, the requirements made on track p_{T} and $|d_0|$ and $|z_0 \sin\theta|$ are tigher for the impact-parameter-based algorithms (IP2D/IP3D) than for SV and JetFitter. The impact-parameter based studies performed by the IP2D/IP3D algorithms require very high track quality. On the other hand, a relatively looser track selection allow SV and JetFitter to reconstruct more vertices. These algorithms are described in section 3.4.

Parameter	Impact parameter based	SV	JetFitter
p_{T} [GeV]	> 1.0	> 0.7	> 0.5
$\lvert d_0 \rvert$ [mm]	< 1.0	< 5.0	< 7.0
$\lvert z_0 \sin\theta \rvert$ [mm]	< 1.5	< 25	< 10
Number of IBL hits	≥ 1	≥ 0	≥ 0
Number of pixel hits	≥ 2 (≥ 1)	≥ 1	≥ 1
Number of SCT hits	≥ 0	≥ 4	≥ 4
Number of pixel/SCT hits	≥ 7	≥ 7	≥ 7
Number of shared hits	≤ 1	≤ 1	≤ 1
Number of pixel holes	≤ 1	≤ 1	≤ 1
Number of pixel/SCT holes	≤ 2	≤ 2	≤ 2

Table 7: Track selection criteria for the main b-tagging algorithms, described in section 3.4. The number of pixel hits was required to be ≥ 2 for the studies presented in this chapter. The requirement was afterwards changed to ≥ 1 as it was found that it does not degrade the performance.

3.2.4 Track-to-jet association

To exploit track properties for b-jets identification, it is necessary to perform an association of tracks to jets. A track is matched to a jet if the ΔR distance between the track and the jet axis is below a certain value ΔR_{max} that is given by

$$\Delta R_{\mathrm{max}} = 0.239 + e^{-1.22 - 1.64 \times 10^{-5} p_{\mathrm{T}}}, \tag{45}$$

with p_{T} of the jet in MeV. The p_{T} dependence of the threshold value helps to reduce the contamination of tracks from quark fragmentation that typically have larger ΔR with

respect to the jet direction. For higher-p_{T} jets the fragmentation track component is larger, but at the same time the signal tracks (originating from b-hadron decay) are more collimated towards the jet axis. This allows to reduce the fragmentation track component by applying a tighter ΔR requirement.

3.2.5 Jet truth labelling

To estimate the b-tagging performance, it is necessary to know the truth flavour of particles that jets originate from. To accomplish this, a procedure referred to as *jet labelling* is performed. Jets are matched to truth particles from the MC simulation with $p_{\mathrm{T}} > 5$ GeV if the spatial distance between them is $\Delta R < 0.3$. The matching is done exclusively. If a b-hadron is found within the cone of $\Delta R < 0.3$ the jet is labelled as a b-jet. If no b-hadron is found, the matching algorithm searches for c-hadrons, and then, if a c-hadron is also not found, for τ-leptons. If none of these particles are found, the jet is labelled as a light jet.

3.2.6 Efficiency and rejection rate

The *b-tagging efficiency* is defined as the ratio of the number of truth b-jets that are tagged to the total number of truth b-jets:

$$\epsilon_b = \frac{N_{b,\mathrm{truth}}^{\mathrm{tagged}}}{N_{b,\mathrm{truth}}^{\mathrm{total}}}. \tag{46}$$

The *rejection rate* of light jets is defined as the inverse efficiency of mis-tagging light jets as b-jets (referred to as *mis-tag rate*):

$$\mathcal{R} = \frac{1}{\epsilon_{\mathrm{light}}}. \tag{47}$$

The goal of a b-tagging algorithm is to tag as many b-jets as possible while rejecting as many light jets as possible.

3.2.7 *b*-tagging calibration

The b-tagging efficiencies vary between the data and simulation. To take this into account, the efficiencies obtained from simulation are calibrated to match the efficiencies measured in data. In ATLAS there are several methods to measure the b- and c- tagging efficiency and mis-tag rate in data [67]. The samples selected for calibration are those enriched in jets with one dominant flavour. For the b-tagging calibration those are $t\bar{t}$ samples with one or two leptons in the final state or samples of jets that contain a muon (events with muons are enriched in b-jets that originate from semileptonic decays of b-hadrons). For the analysis presented in this dissertation $t\bar{t}$ samples with one or two leptons in the final state are used.

The c-tagging calibration uses events with a high probability to contain c-jets, in particular, those originating from decays of D mesons or W bosons. The analysis presented in this dissertation uses c-jets from W boson decays in $t\bar{t}$ events.

Finally, the light jet mis-tag rate is measured in an inclusive jet sample using the so-called *negative tag* method. This method evaluates the rate of light jet misidentification caused only by detector resolution effects (and not due to contamination of the tracks from particles with long lifetime, described in section 3.2.3). To do that, jets containing SV and tracks with impact parameters consistent with a negative lifetime are used. Tracks that originate from the PV are expected to have a symmetric distribution of their impact parameter significance. Therefore, selecting tracks with negative impact parameter and decay length and inverting their sign allows to evaluate the mis-tag rate caused by finite detector resolution.

3.3 Sample and event selection

The studies presented in this chapter are performed using $t\bar{t}$ MC events corresponding to 13 TeV pp collisions simulated with Powheg-Box + Pythia 6 and the CT10 [68] parton distribution functions. EvtGen [69] is used to model the decays of b and c-hadrons. Only $t\bar{t}$ events with at least one lepton from a W boson decay are considered.

Jets are reconstructed as described in section 2.4.5 with the anti-k_{T} algorithm with a radius parameter R=0.4. Only jets with $p_{\mathrm{T}} > 20$ GeV and $|\eta| < 2.5$ are considered.

To reject jets that originate from pile-up activity the JVT variable described in section 2.4.5 is used. Jets with $p_{\mathrm{T}} < 50$ GeV and $|\eta| < 2.4$ are rejected if they have a JVT output of less than 0.641. This corresponds to an expected efficiency of about 92% for jets from the hard-scatter and a 2% efficiency for pile-up jets. The JVT selection is close to 100% efficient for b-tagged b- and c-jets.

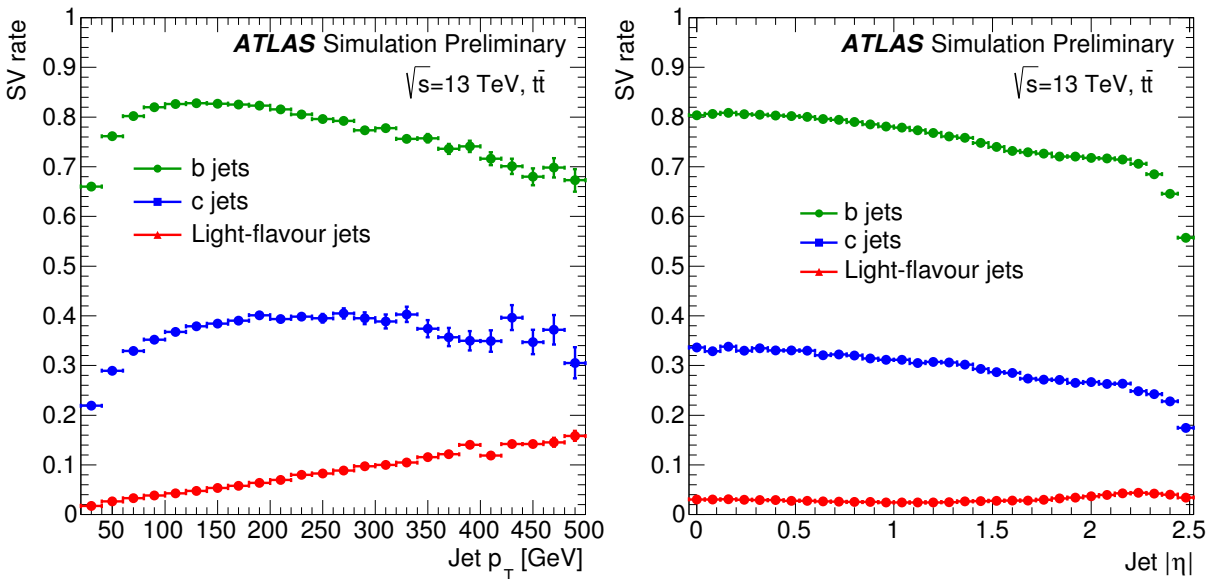
3.4 *b*-tagging algorithms in ATLAS

For Run 2 the b-tagging algorithms were revisited and optimised. The basic Run 2 ATLAS b-tagging algorithms are:

- Secondary vertex finding (SV), reconstructing an inclusive displaced secondary vertex within the jet.
- Decay chain multi-vertex fit (JetFitter), attempting to reconstruct the full b-hadron decay chain.
- Impact parameter-based (IP2D, IP3D), making use of the fact that tracks from b-hadron decays are not pointing to the primary vertex.
- A multivariate algorithm (MV2) that combines several observables from other algorithms together with kinematic variables and provides the final discriminant between the different jet flavours.

3.4.1 Secondary vertex finding

The secondary vertex algorithm attempts to reconstruct the inclusive SV formed by the decay products of the b-hadron, including those from the subsequent c-hadron decay.



(a) (b)

Figure 41: Secondary vertex reconstruction rate as a function of jet (a) p_T and (b) η for b-, c- and light jets in $t\bar{t}$ MC events. From Ref. [65].

Firstly it searches for all two-track pairs that form a good vertex, using tracks displaced from the primary vertex. Then the algorithm removes those tracks that are compatible with decays of long-lived particles (K_s, Λ etc) or interactions with the detector material. After this selection the algorithm fits an inclusive secondary vertex. Several properties of this vertex are useful to discriminate b-jets, such as its mass, number of tracks, distance to the PV, or the energy fraction of tracks associated with the SV with respect to all tracks in the jet.

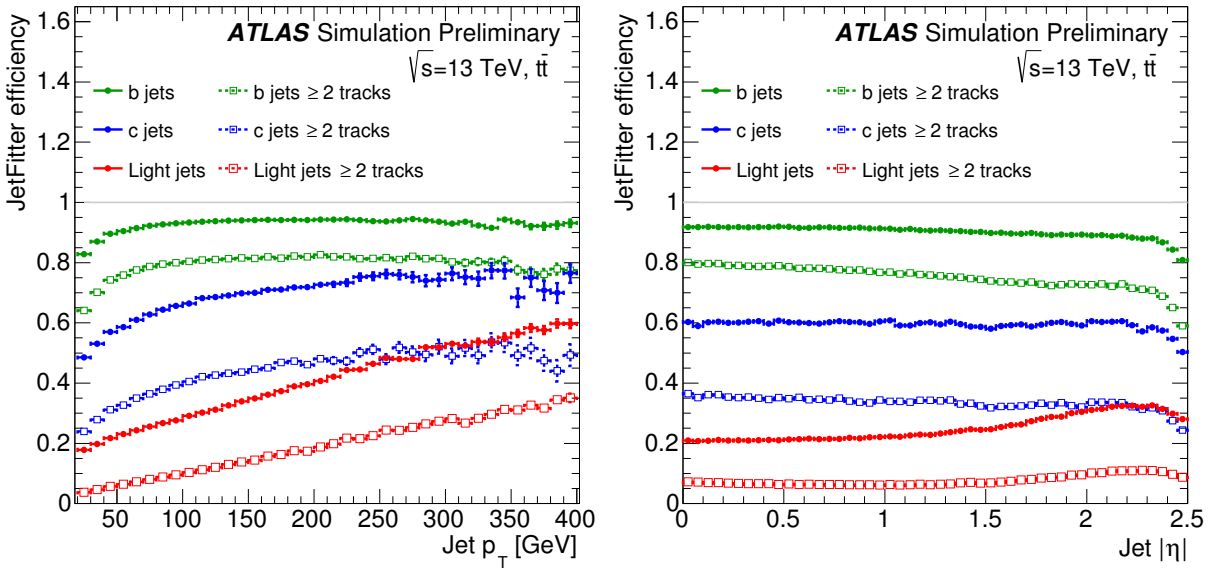
Figure 41 shows the SV vertex reconstruction rate as function of jet p_{T} and η for b-, c- and light-flavour jets.

3.4.2 Multi-vertex fit

The JetFitter algorithm attempts to reconstruct the full PV to b- to c-hadron cascade decay topology. The approach used in the algorithm is based on the assumption that the PV and both b- and c-hadron decay vertices are placed along one line, approximating the b-hadron path. A Kalman filter [51] is used to find such a common line and the positions of vertices on it.

One of the advantages of the JetFitter algorithm is that it allows to separate b- and c-hadron vertices even if only one track is attached to each of them. Reconstruction of the full b-hadron decay chain allows to improve the separation against light jets. Finally, the separation between b-jets and c-jets is also improved, as JetFitter is able to distinguish the b-jet topology with two displaced vertices against the c-jet with one vertex.

Figure 42 shows the JetFitter efficiency to reconstruct a vertex with at least one or two



(a) (b)

Figure 42: JetFitter vertex reconstruction rate as function of jet (a) p_T and (b) η for b-, c- and light jets in $t\bar{t}$ MC events. From Ref. [65].

tracks as function of jet p_T and η for different jet flavours. The efficiency to have at least a single-track vertex is higher than the efficiency to have a vertex with ≥ 2 tracks. However, the higher efficiency of reconstructed vertices in b-jets comes at the cost of higher rate also in light-jets.

3.4.3 Impact-parameter-based algorithms

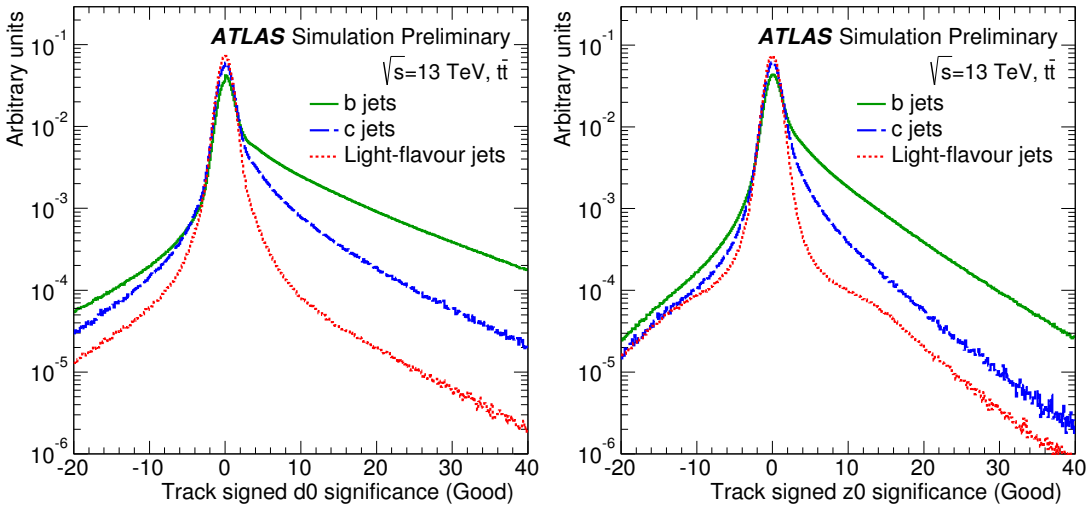
The impact-parameter-based algorithms IP2D, IP3D use a log-likelihood approach to evaluate the probability of a jet to originate from a b-quark, taking advantage from features of the tracks associated with this jet. In particular, they exploit the d_0 and z_0 parameters, that describe the position of the tracks with respect to the PV, as well as their uncertainties.

The IP2D algorithm uses as input S_{d_0}, while IP3D builds a 2D likelihood of both S_{d_0} and S_{z_0}, thus taking into account their correlation.

The IP2D algorithm defines the probability of a given track to belong to a b-jet as

$$p_b(S_{d_0}) = \frac{\mathcal{P}_b(S_{d_0})}{\mathcal{P}_b(S_{d_0}) + \mathcal{P}_c(S_{d_0}) + \mathcal{P}_{\text{light}}(S_{d_0})}, \tag{48}$$

where $\mathcal{P}_b(S_{d_0})$, $\mathcal{P}_c(S_{d_0})$ and $\mathcal{P}_{\text{light}}(S_{d_0})$ are evaluated using the S_{d_0} pdfs derived from the MC simulation. In the same way the track probability to belong to a light jet $p_{\text{light}}(S_{d_0})$ is calculated.



(a) (b)

Figure 43: Pdfs of (a) d_0 and (b) z_0 significance of tracks from b-, c- and light jets in $t\bar{t}$ MC events for the "Good" track category. From Ref. [65].

The b-tagging weight of a track is defined as

$$w_{\text{track}} = \frac{p_b}{p_{\text{light}}}, \tag{49}$$

and the weight of a jet is computed by summing the logarithms of the weights of all tracks that are associated with this jet:

$$w_{\text{jet}} = \sum_{\text{tracks}} \log w_{\text{track}}. \tag{50}$$

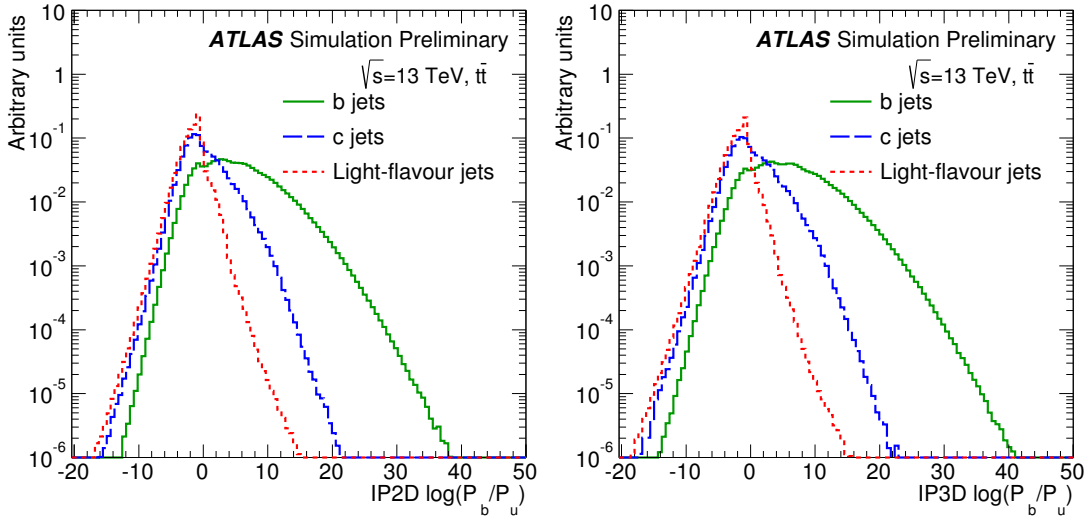
IP2D and IP3D algorithms use different pdfs for different track categories, depending on the quality of the tracks, which is defined using information on hits in the different silicon layers of the inner detector. Figure 43 shows S_{d_0} and S_{z_0} pdfs for the best quality track category for tracks from b-, c- and light jets. Figure 44 shows the final output weights of the IP2D and IP3D algorithms for b-, c- and light jets.

3.4.4 Multivariate algorithm

The discriminating variables from several b-tagging techniques are combined via the MV2 algorithm based on a boosted decision tree (BDT). The kinematic properties (p_{T} and η) of the jets are included in the training to take advantage of correlations with the other input variables.

The default algorithm used for the analysis presented in this dissertation, MV2c10, is a version of the MV2 algorithm, which is trained using b-jets as signal and a mixture of light-flavour jets and c-jets as background (the fraction of c-jets in the background is set to 10% of the amount of light-jets). The previous default version of MV2 was MV2c20 (with 20% of c-jets with respect to the amount of light-jets).

MV2 is an upgrade of the main Run 1 b-tagging algorithm MV1, which combined the outputs of the various b-tagging algorithms using a neural network approach. The MV2



(a) (b)

Figure 44: The log likelihood ratio for the (a) IP2D and (b) IP3D algorithm for b-, c- and light jets in $t\bar{t}$ MC events. From Ref. [65].

algorithm provides better performance and easier retraining and software maintenance. Figure 45 shows the MV2c10 output distribution for b-, c- and light jets. The b-jets are expected to have higher MV2 weight. Applying a tighter (closer to 1) MV2 output requirement provides a lower b-tagging efficiency, but a higher light-jet rejection, than in the case of a looser MV2 requirement. A detailed description of the MV2 algorithm can be found in Ref. [70].

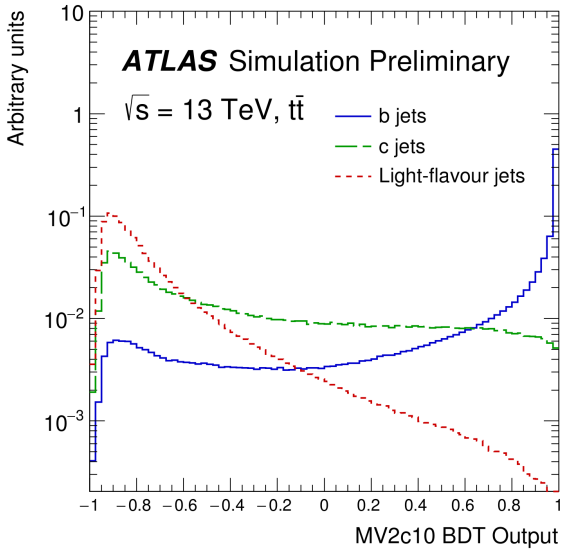


Figure 45: The MV2c10 output distribution for b-, c- and light jets in $t\bar{t}$ MC events. From Ref. [71].

3.5 Impact-parameter-based algorithms optimisation

3.5.1 Track categorisation

Many tracks are well reconstructed and therefore have a good resolution on the impact parameter ($\sigma(d_0) \sim 50\mu$m for a 2 GeV track). But there are also tracks of worse quality: those with a missing hit in one of the pixel detector layers or with ambiguities in the pattern recognition. Rejecting all low-quality tracks would increase light-jet rejection, but at a cost of significant decrease in b-tagging efficiency. To make an efficient use of low-quality tracks, it is necessary to divide them into categories and treat each category differently (with dedicated pdf used in the likelihood calculation).

The Run 2 track categorisation was improved with respect to Run 1 by making use of several new tracking variables (including those related to the presence of the IBL). To determine the properties of the b-hadron decay (position of the PV and SV, track impact parameters), precise measurements, especially close to the interaction point, are necessary. Therefore the two innermost layers of the pixel detector play a key role for b-tagging. Throughout this chapter the following notation is used: L0 is the IBL, the innermost pixel detector layer for Run 2; L1 refers to the next to the innermost layer for Run 2 (which was the innermost layer for Run 1). The variables used for Run 2 categorisation are:

- The number of hits in L0 and L1. Tracks with missing hit(s) are considered to be worse reconstructed, so they should be treated separately from the "better" tracks.
- Information on whether a hit in L0 and L1 is expected or not (see definition in section 3.2.3). The case when there is no hit in a layer, while it is expected, indicates that the track is poorly reconstructed. These tracks are very different from the "better" tracks with no hit in a layer where it is not expected (those are mostly tracks crossing a region outside the detector coverage or a dead module), therefore they must be separated into a different category. This is a new tracking variable, not used in Run 1.
- The number of shared hits in L0, L1 as well as in other pixel detector layers and the SCT.
- The number of split hits in L0, L1 and other pixel detector layers.

To study the impact parameter resolution effects, only tracks that originate from light jets were considered, as they are expected to have a symmetric impact parameter distribution. Tracks were divided into 14 exclusive categories based on the variables above. Their d_0 and z_0 significance distributions were studied.

1. No hits in the first two layers (L0 and L1), while hits are expected in both. The transverse impact parameter distribution has a double peak structure (see figure 46). The tracks in this category originate mostly from photon conversions, with contamination of those from decay of of K_s, λ, and interactions with the detector material.

These tracks do not have hits in the first layers as the particles they originate from decay beyond these layers (while the hit might still be "expected" if the interpolated curve crosses working modules of these layers). The impact of each of these different track types is described in section 3.5.2.

2. No hits in the first two layers, while a hit is expected in L0 and not expected in L1.
3. No hits in the first two layers, while a hit is not expected in L0 and expected in L1.
4. No hits in the first two layers and not expected in both. These tracks mostly lie outside the detector coverage. These tracks are better reconstructed than those in categories 1, 2 and 3 where some of the hits are expected, but missing, and the impact parameter resolution is significantly better.
5. No hit in L0, while expected, and a hit is present in L1.
6. No hit in L0 and not expected, and a hit is present in L1. The resolution of the impact parameter resolution is better, than for category 5.
7. No hit in L1, while expected, and a hit is present in L0.
8. No hit in L1 and not expected, and a hit is present in L0. This category has a larger fraction of tracks with respect to category 6 because of the existing dead modules in L1.
9. Shared hits in both L0 and L1.
10. Shared hits in L0, or in L1, or in other pixel layers.
11. Two or more shared hits in the SCT.
12. Split hits in both L0 and L1.
13. Split hits in L0 or L1 or other pixel layers.
14. "Good" tracks: tracks not in any of the above categories.

A finer track grading of 18 categories was additionally considered, splitting category 10 described above into three categories ("Shared hits in L0 (and not in L1)", "Shared hits in L1 (and not in L0)" and "Shared hits in other pixel layers") and category 13 into three categories ("Split hits in L0 (and not in L1)", "Split hits in L1 (and not in L0)" and "Split hits in other pixel layers"). It was shown that configurations with 14 or 18 categories have similar performance of the IP3D algorithm, and the 14 categories grading was chosen as default. The percentage of tracks from b-, c- and light flavour jets in each of the Run 2 b-tagging track categories is presented in table 8.

The distributions of d_0 significance were approximated by a double gaussian function:

$$f(x) = \frac{a_1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu)^2}{2\sigma_1^2}} + \frac{a_2}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu)^2}{2\sigma_2^2}}, \tag{51}$$

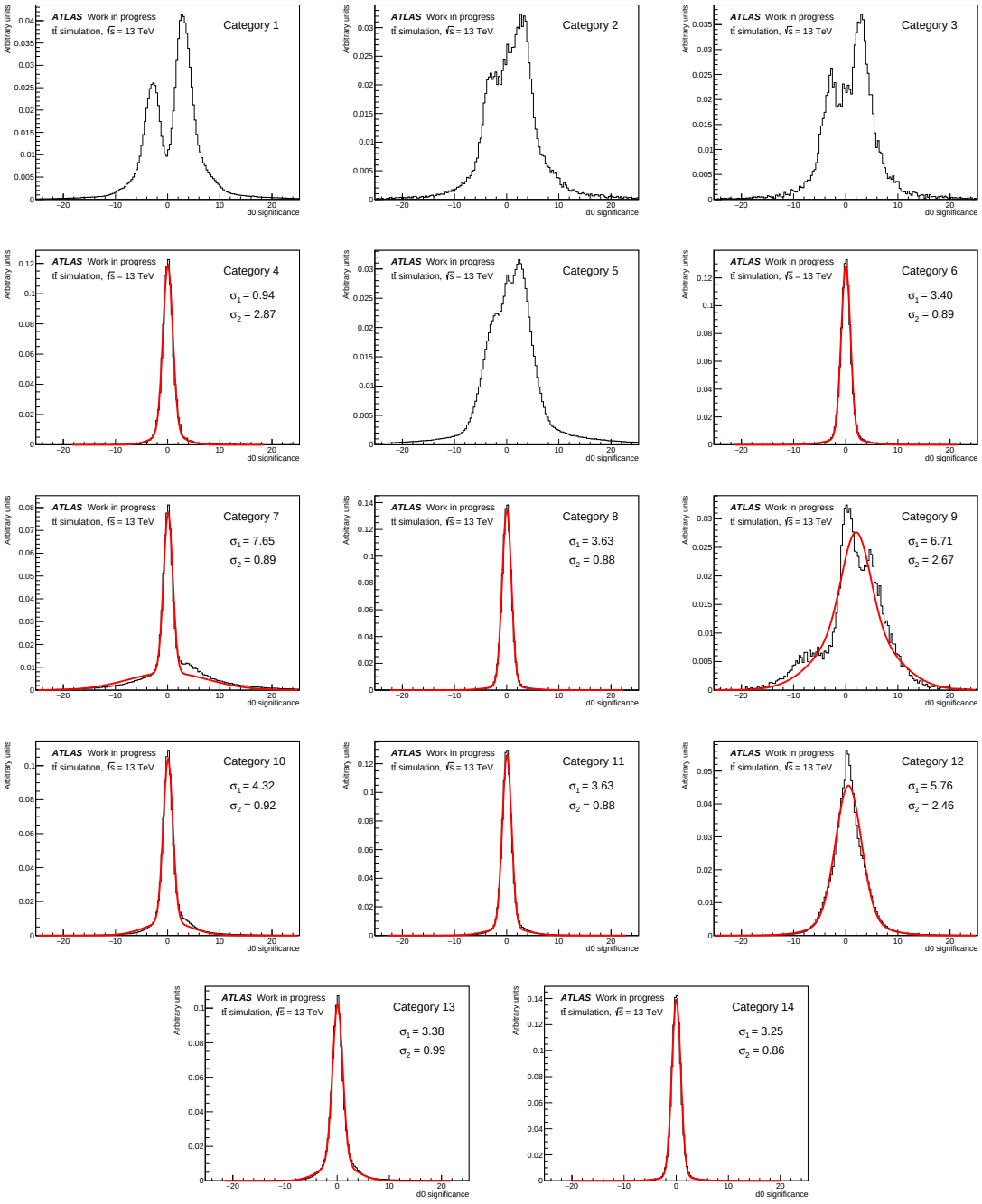


Figure 46: The S_{d_0} distributions for the 14 b-tagging track categories and the result of the fit to a double gaussian function, given by equation 51. No attempt to fit categories 1, 2, 3 and 5 is made. The two peaks of distributions in these categories are formed by tracks originating from photon conversions.

#	Category	light jets	b-jets	c-jets
1	No hits in first two layers; expected hit in both L0 and L1	1.6%	1.5%	1.6%
2	No hits in first two layers; exp. hit in L0 and no exp. hit in L1	0.1%	0.1%	0.1%
3	No hits in first two layers; no exp. hit in L0 and exp. hit in L1	0.03%	0.03%	0.03%
4	No hits in first two layers; no exp. hit in L0 and L1	0.02%	0.03%	0.03%
5	No hit in L0; exp. hit in L0	2.1%	2.4%	2.3%
6	No hit in L0; no exp. hit in L0	0.9%	0.9%	0.9%
7	No hit in L1; exp. hit in L1	0.5%	0.5%	0.5%
8	No hit in L1; no exp. hit in L1	2.3%	2.4%	2.4%
9	Shared hit in both L0 and L1	0.04%	0.01%	0.01%
10	Shared hits in other pixel layers	1.8%	2.1%	1.6%
11	Two or more shared SCT hits	2.2%	2.4%	2.2%
12	Split hits in both L0 and L1	0.8%	1.2%	1.1%
13	Split hits in other pixel layers	1.1%	2.1%	1.6%
14	Good: a track not in any of the above categories	86.6%	84.3%	85.5%

Table 8: Run 2 IP2D and IP3D track categories and corresponding fractions of tracks from b-, c- and light jets in each category in $t\bar{t}$ MC events.

where μ is the mean of the distribution, σ_1 and σ_2 are the standard deviations of the wider and the core gaussian peak, respectively, and a_1 and a_2 are arbitrary normalisation factors.

Categories 1, 2, 3 and 5 are mostly formed by tracks originating from photon conversions (see section 3.5.2) that give a double-peak structure for the S_{d_0} distribution, and therefore is not fitted by a double gaussian function. For the remaining categories the standard deviation values of the peaks σ_1 and σ_2 in the gaussian approximation of their S_{d_0} distributions were studied. The tails in the distributions of categories 7, 10 and 13 (resulting in large σ_1 values) are due to large fractions of tracks originating from photon conversions and significant contamination of "fake" tracks (see section 3.5.2). The shape of category 9 is mainly formed by the very large "fake" component, while photon conversions also play a role, resulting in a visible double peak. Categories 9 and 12 are composed of tracks with severe pattern-recognition ambiguities (with either shared or split hits in both L0 and L1), that have significantly worse resolution (see large σ_2 values).

In similar way the S_{z_0} distributions were analysed. The final set of categories was defined based on both S_{d_0}, which is the main discriminating variable of the IP2D/IP3D algorithms, and S_{z_0} distributions. The S_{d_0} distributions for the different categories and the result of the fit to a double gauss function are shown in figure 46.

The impact of each of the 14 categories was evaluated by retraining the IP3D algorithm with one category at a time removed (if a track belongs to this category, it is rejected and its contribution is not taken into account in the IP3D weight calculation). The performance degrades when removing most of the categories, but does not change when removing some low-statistics categories: 2, 3, 4, 7 and 9. The performance of versions of the IP3D

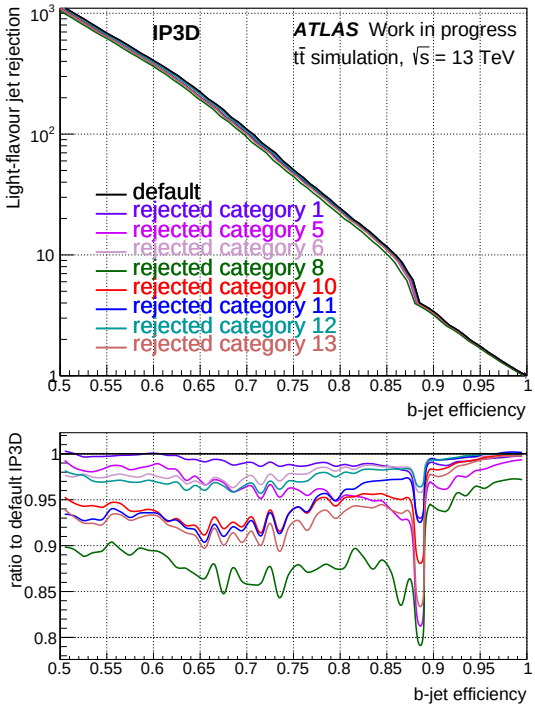


Figure 47: Performance of the versions of IP3D algorithm with each track category at a time removed from the calculation compared with the default IP3D in simulated $t\bar{t}$ events: light jet rejection vs b-jet efficiency. Rejecting categories 2, 3, 4, 7 and 9 has no impact on performance, and the corresponding curves are not shown.

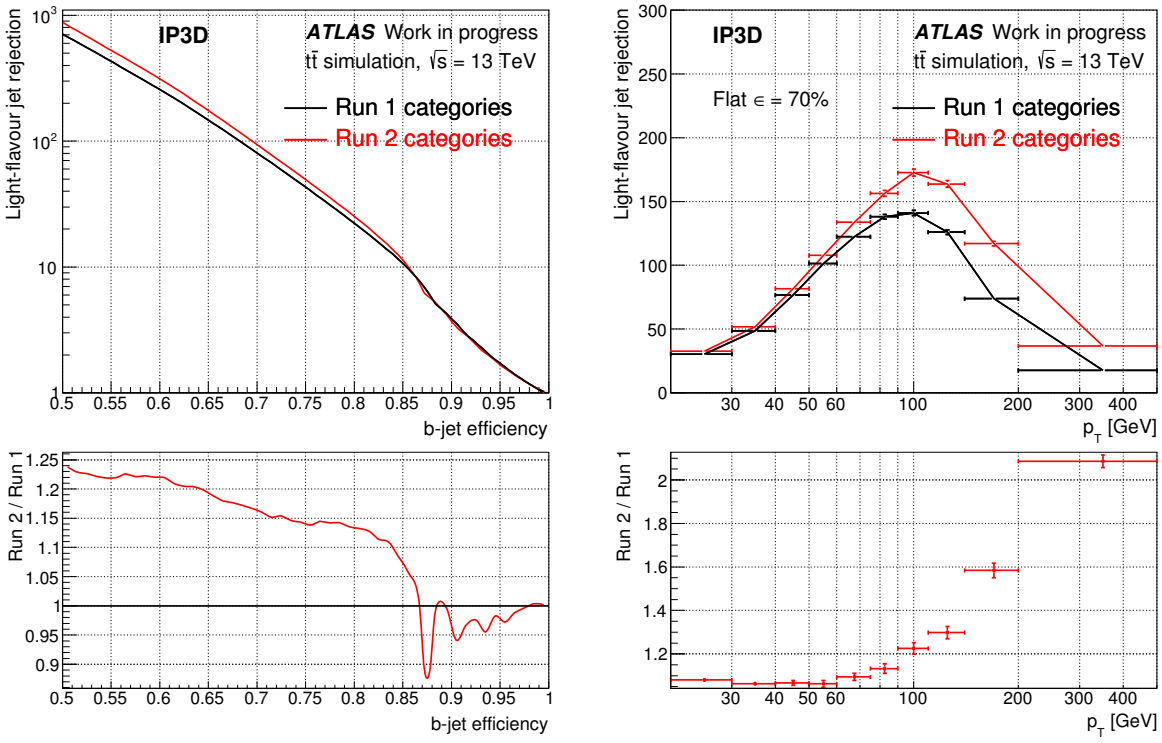
algorithm with different categories removed with respect to the default algorithm are presented in figure 47. No increase in performance when removing a category is observed; therefore all these categories are kept in the Run 2 configuration.

The comparison of the IP3D algorithm performance in Run 2 with the new tracking categories with respect to the old ones used in Run 1 is presented in figure 48. The new IP3D configuration results in $\sim 15\%$ higher light-jet rejection for a b-tagging efficiency of 70%.

3.5.2 Track selection

Another aspect of the impact-parameter-based algorithms optimization regards the track selection. Most tracks associated with light jets originate from the PV, and thus have an impact parameter close to zero, which allows to distinguish them from tracks from b-jets. These tracks are referred to as *primary* tracks. However, in both light and b-jets there is contamination from so-called *secondary* tracks or "bad" tracks, that originate from long-lived particles, such as K_s, Λ, interactions with the detector material, and photon conversions ($\gamma \to e^+e^-$ decay). These tracks have a large impact parameter, so they can degrade the b-jets identification.

To reduce the negative effect of "bad" tracks one needs to identify and reject them.



(a) (b)

Figure 48: Performance of the IP3D algorithm with Run 1 categories and the new Run 2 categories in $t\bar{t}$ MC events: (a) light-jet rejection vs b-jet efficiency and (b) light-jet rejection as a function of jet p_{T} for a fixed b-jet efficiency of 70% in each bin.

This is partially achieved by applying a requirement on $|d_0|$ and $|z_0 \sin\theta|$ when selecting the tracks, as described in section 3.2.3. Additionally, the SV algorithm identifies whether a track is likely to be coming from a long-lived particle, as explained in section 3.4.1.

A MC study was performed to evaluate the contamination of "bad" tracks of different truth origin for each of track categories described in section 3.5.1, the expected gain for b-tagging from removing the "bad" tracks, and the efficiency of SV "bad" track identification procedure.

The following definitions are used in the categorisation of "bad" tracks:

- "fakes" are the tracks with a low matching probability with a MC particle: those with less than 75% of hits associated from the same MC particle (based on Geant4 information).

- "pile-up tracks" are the tracks with no association link with a MC particle; most of these tracks originate from pile-up interactions rather than from hard scattering process; therefore, the information about the MC particles they originate from is not stored.

The truth origin of tracks associated with light jets in $t\bar{t}$ MC events for different b-tagging categories was studied. In the largest "good" track category 97% of the tracks are primary, while there are categories with significantly different composition. For instance, categories 1, 2, 3 (with missing hit in both L0 and L1, while at least one of them is expected) and 5 (with missing hit in L0 while expected) the majority of tracks originate from photon conversions (81% for category 1, 56% for category 2, 66% for category 3 and 57% for category 5). The contamination of photon conversions is also significant for categories 7 (22%) and 12 (18%). Categories 1, 2, 3 and 5 have the largest contamination of tracks originating from decay of K_s (2-3%), λ (2-3%) and other particles including interaction with detector material (3-4%). In some categories the fraction of "fake" tracks is significant: this contamination is especially large (33%) for category 9.

In order to evaluate the expected impact on b-tagging performance due to different types of "bad" tracks a simulation study was performed. The contaminations of "bad" tracks of different types were rejected from the MC samples when calculating the IP3D likelihood. The performance of the different IP3D algorithms, the default one and ones with each type of "bad" tracks removed, is compared in figure 49.

The most damaging for b-tagging are tracks coming from K_s decays: if those particles are removed, the rejection of light jets when the b-jet efficiency is 70% is increased by $\sim 23\%$ with respect to the default algorithm (while this difference is even larger for lower efficiencies). The tracks originating from Λ decays have a less significant effect: the relative gain in light-jet rejection of removing those particles from the IP3D calculation is $\sim 5\%$. Tracks originating from photon conversions do not have any negative impact, but rather provide a slight improvement (1 to 2% at 70% b-jet efficiency). This can be explained by the fact that some of the electrons originating from photon conversions have transverse energy close to the initial photon's, so the transverse impact parameter is not large enough to be harmful for b-tagging. The tracks from the rest of the long-lived particles, including interactions with the detector material also have a significant impact: if removed, the light-jet rejection increases by $\sim 15\%$ at 70% b-jet efficiency.

Applying the SV "bad" track removal procedure to select tracks for the IP3D algorithm allows to increase by $\sim 10\%$ the light-jet rejection at 70% b-jet efficiency, as shown in figure 50. It was enabled in the IP2D and IP3D algorithms in 2016.

3.6 *b*-tagging performance in Run 2

The overall b-tagging performance in Run 2 was improved due to the addition of the IBL, updates in track reconstruction [35, 72] and b-tagging algorithms optimisation.

Figure 51 (a-b) shows a comparison of the performance of the Run 2 b-tagging algorithm MV2c20 and the equivalent Run 1 b-tagging algorithm MV1c. Figure 51 (c) shows a comparison of the performance of the Run 1 default b-tagging algorithm MV1 with the Run 2 equivalent MV2c00 (i.e. trained only with a light-jet sample, without contamination of c-jets). The improvement at low and medium p_{T} is mostly due to the addition of the IBL, while the improvement at high p_{T} is due to algorithm improvements. At 70% efficiency the light-jet rejection in Run 2 is improved by a factor of 4 compared to Run 1. This corresponds to a $\sim 10\%$ gain in b-jet efficiency at a constant light-jet rejection.

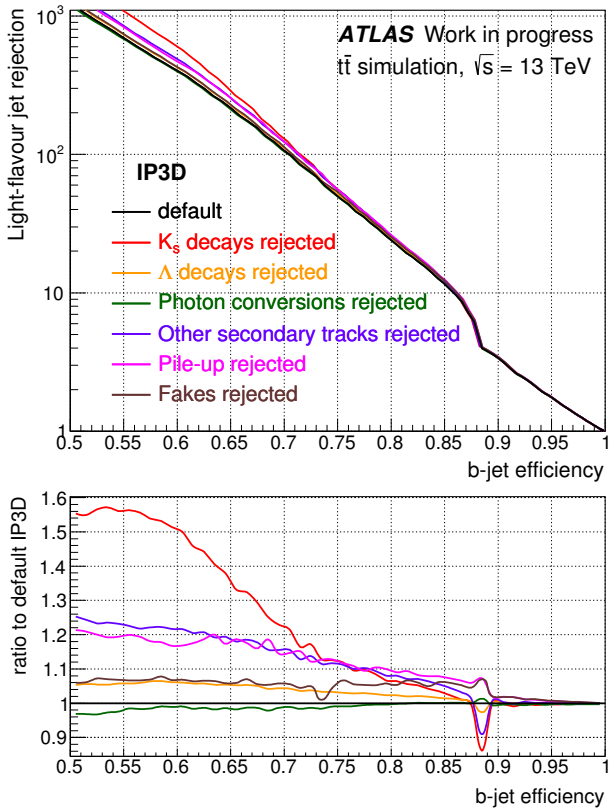
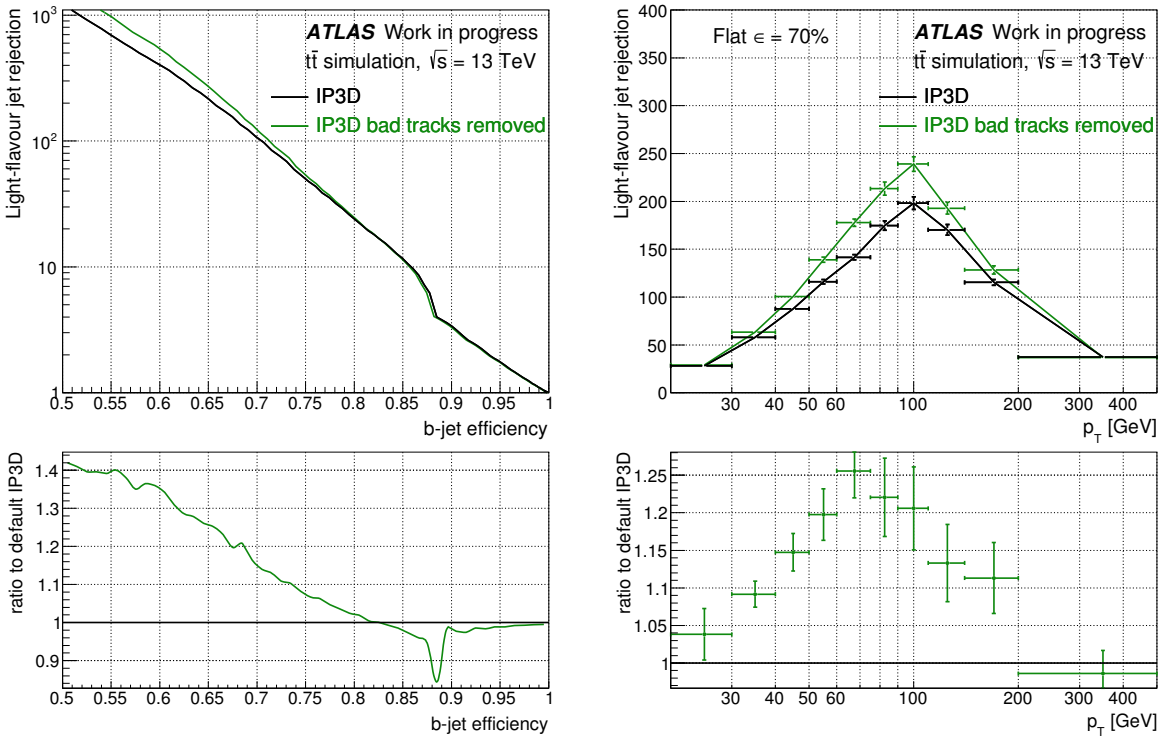


Figure 49: Effect of different types of "bad" tracks on the IP3D performance. Configurations of the IP3D algorithm with each type of "bad" track rejected are compared to the default algorithm: light jet rejection vs b-jet efficiency. Origin of rejected tracks is determined from truth MC particles.

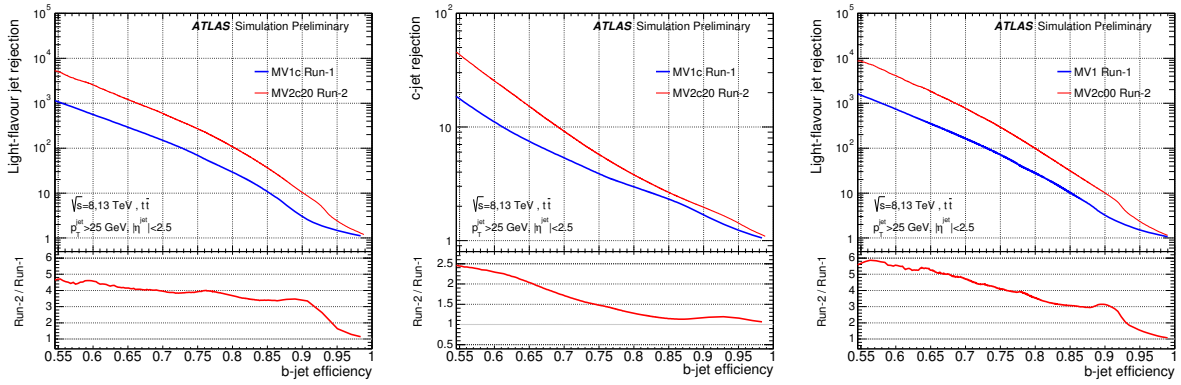
To validate the MC performance, the results in simulation were compared to data. The study was performed using pp collision data collected by ATLAS at the centre-of-mass energy of 13 TeV with 50 ns bunch-spacing on a high purity b-jet sample of $e+\mu$ di-leptonic $t\bar{t}$ candidate events. Only jets with $p_{\mathrm{T}} > 20$ GeV and $|\eta| < 2.5$ are considered [73].

Input observables as well as the output of the multivariate algorithm MV2c20 have been studied. Figure 52 shows the log-likelihood ratio of the IP3D algorithm and the output distribution of the MV2c20 algorithm. In all plots the data are shown by the points and the simulation by the filled area, divided into b (red), c (light green) and light-flavour (blue) components. The dark green shaded area represents the total systematic and statistical uncertainty on the simulation and the error on the points corresponds to the statistical uncertainty on the data. In general, good agreement between the data and the simulation is observed.



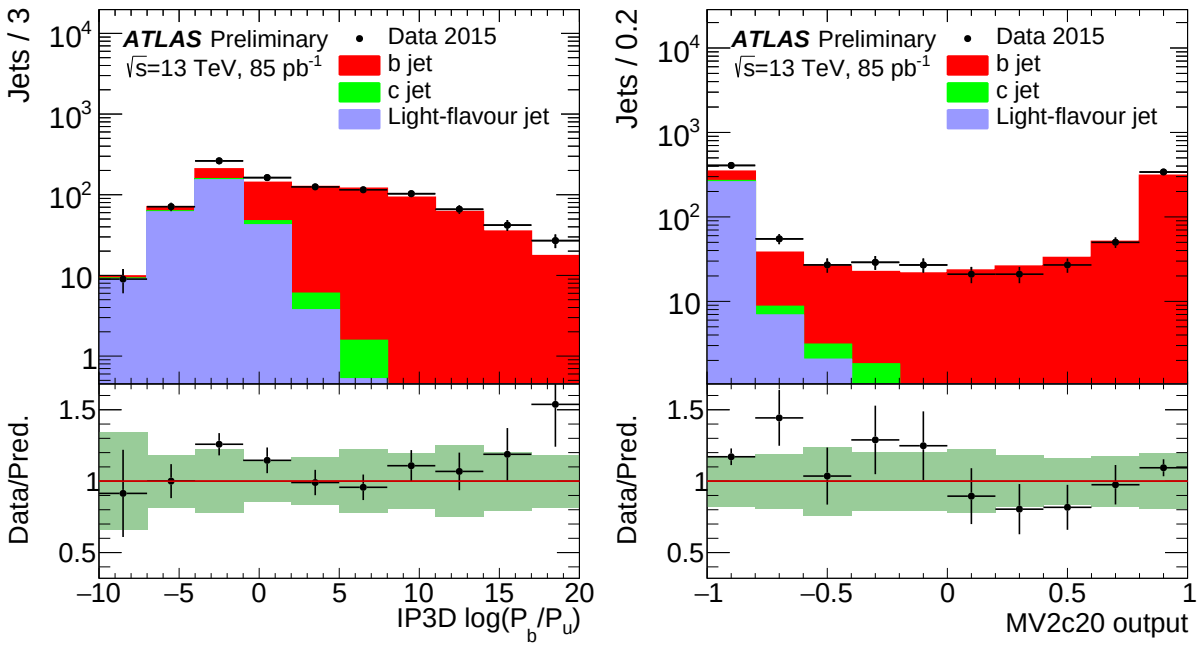
(a) (b)

Figure 50: Effect of the SV "bad" tracks removal procedure on the IP3D performance: (a) light-jet rejection vs b-jet efficiency and (b) light-jet rejection as a function of jet p_{T} for a fixed b-jet efficiency of 70% in each bin.



(a) (b) (c)

Figure 51: Comparison of Run 2 and Run 1 b-tagging algorithms in simulated $t\bar{t}$ events: default Run 2 algorithm MV2c20 and the equivalent Run 1 b-tagging algorithm MV1c: (a) light jet rejection vs b-jet efficiency and (b) c-jet rejection vs b-jet efficiency; (c) default Run 1 b-tagging algorithm MV1 and the Run 2 equivalent MV2c00. From Ref. [65].



(a) (b)

Figure 52: Output distribution of (a) the IP3D algorithm and (b) the final MV2c20 algorithm for jets selected from the $t\bar{t}$-dominated $e + \mu$ sample From Ref. [73].

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

In this chapter the search for the Higgs boson in the $t\bar{t}H$ ($H \to b\bar{b}$) channel, using 36.1 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV recorded with the ATLAS detector at the LHC in 2015 and 2016, is presented. The search is focussed on the semileptonic decay of the $t\bar{t}$ system. My main contribution to the analysis is the development and optimisation of a likelihood-based technique to distinguish the signal from the background.

4.1 Introduction

The Higgs boson, discovered in Run 1 of the LHC, was observed in several production modes, but not in the channel of associated production with top quarks ($t\bar{t}H$). The observation of the Higgs boson production in this channel is one of the most important goals of the LHC Run 2.

The decay mode of the Higgs boson into a pair of b-quarks, $H \to b\bar{b}$, is dominant in the SM for the value of the Higgs boson mass of 125 GeV. However, it is challenging to be observed experimentally, compared to the channels with photons or leptonically-decaying W and Z bosons in the final state, due to a large background. Apart from that, this decay channel is particularly interesting as it allows measuring the Yukawa coupling to the b-quark, which is the second largest coupling of the Higgs boson to a fermion in the SM.

Three different channels of the $t\bar{t}H$ ($H \to b\bar{b}$) process are explored in ATLAS:

- The fully-hadronic channel $t\bar{t}H \to (q\bar{q}b)(q\bar{q}b)(b\bar{b})$. Although this channel has the most favourable branching ratio, it has the challenges of the large QCD background, which is difficult to model accurately, as well as a large jet combinatorial background in the signal itself.

- The dilepton channel $t\bar{t}H \to (\ell\nu b)(\ell\nu b)(b\bar{b})$. The two neutrinos from the leptonic W boson decays both contribute to the $E_{\mathrm{T}}^{\mathrm{miss}}$, therefore making the reconstruction of the event kinematics less precise.

- The single-lepton channel $t\bar{t}H \to (\ell\nu b)(q\bar{q}b)(b\bar{b})$. This is the most sensitive analysis channel of the three. The reconstruction of the event kinematics is easier than for the dilepton channel, as there is only one neutrino in the final state and its kinematics can be determined from the $E_{\mathrm{T}}^{\mathrm{miss}}$. The requirement of a lepton in the final allows to reduce the QCD background.

The term *lepton* here and in the following refers to either an electron or a muon. Electrons and muons from decays of τ-leptons are also included.

The single-lepton channel is subdivided into *resolved* and *boosted* regions. The boosted search targets the final state with the Higgs boson and top quarks produced with a high boost such that their decay products are reconstructed as a single large-radius (large-R) jet. The resolved search covers all other possible final state topologies.

This chapter describes the full analysis chain in the resolved single-lepton channel. The final result from the combination with the dilepton channel and the boosted channel is also presented. Finally, the combination of all ATLAS searches for $t\bar{t}H$ production is presented.

4.2 Object selection

The physics objects considered in this analysis are electrons, muons, $E_{\mathrm{T}}^{\mathrm{miss}}$ and jets, described in section 2.4. Additional requirements are discussed below.

Electrons are selected with the *Tight* likelihood identification criteria (see section 2.4.3). Muons are required to satisfy *Medium* quality requirements (see section 2.4.4). An additional requirement of $p_{\mathrm{T}} > 27$ GeV is applied to both electrons and muons. In addition, both leptons are required to satisfy *Loose* isolation selection.

Jets are reconstructed with the anti-k_{T} algorithm (see section 2.4.5) with a radius parameter $R = 0.4$. They are required to have $p_{\mathrm{T}} > 25$ GeV and $|\eta| < 2.5$. Additional quality criteria are applied to reject jets originating from non-collision source or detector noise: events containing at least one jet failing such quality criteria are removed.

To reject pile-up jets, jets with low transverse momentum ($p_{\mathrm{T}} < 60$ GeV) in the central detector region ($|\eta < 2.4|$) are required to have JVT > 0.59.

A procedure known as overlap-removal is applied to avoid considering a given detector measurement as two different physics objects. The electron overlap-removal is performed to avoid double-counting jets as electrons. First of all, the closest jet whose axis is within $\Delta R < 0.2$ of a selected electron is removed. Then electrons that are lying within $\Delta R < 0.4$ of the remaining jets are removed. Muons are removed if they are within $\Delta R < 0.4$ to the closest jet. In the case this jet has three or less associated tracks, the muon is kept and the jet is removed (this maintains reasonable efficiency for high-energy muons with significant energy loss in the calorimeter). The muon overlap-removal helps to reduce the background that originates from decays of b and c-quarks inside jets.

The identification for b-jets plays a key role in this analysis, as 6 jets are expected in the final state, and 4 of them are b-jets. For this purpose the main b-tagging algorithm MV2c10 (described in detail in section 3) is used. There are four threshold values of the MV2c10 weight (MV2c10 discriminant output) that are defined as *loose*, *medium*, *tight* and *very-tight* working points. For these points the efficiency of b-jet identification is expected to be 85%, 77%, 70% and 60%, respectively. A jet is considered to be b-tagged at a given working point if its MV2c10 weight passes the corresponding threshold.

There are two approaches of using the b-tagging information in a physics analysis:

- *Cumulative* b-tagging employs an overall fixed working point (*loose*, *medium*, *tight* or *very-tight*). The selection of b-jets is performed throughout the analysis considering this working point. This is the approach followed in the $t\bar{t}H(H \to b\bar{b})$ searches performed in Run 1 [74], as well as in the previous Run 2 search.

- *Pseudo-continuous* b-tagging considers several bins of the b-tagging algorithm weight (defined at the same *loose*, *medium*, *tight* or *very-tight* working points), as shown in figure 53. For example, if a jet has an MV2c10 weight that falls into the bin

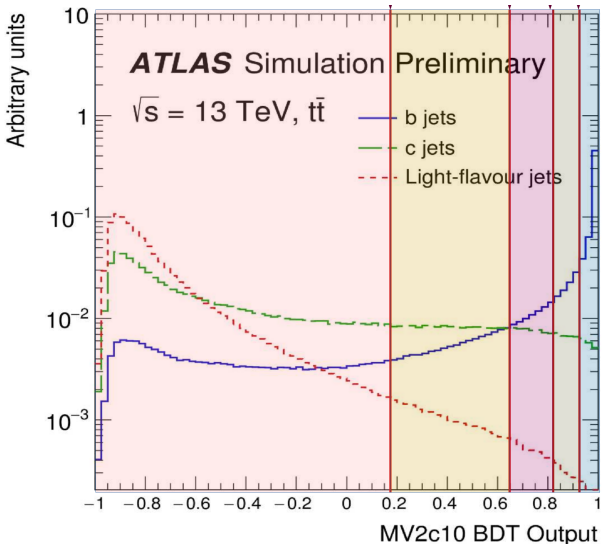


Figure 53: The five bins of the jet MV2c10 weight distribution considered: tagged at *very-tight* (blue), *tight* (green), *medium* (magenta) and *loose* (yellow) working points, and not tagged (orange). From Ref. [71].

between the *medium* and *tight* thresholds it is considered to be "tagged at the *medium* working point". This jet is less likely to be originating from b-quark than, for instance, a jet "tagged at the *tight* working point". Thus, jets are divided into five grades (tagged at *very-tight*, *tight*, *medium* or *loose* working points, and not tagged), and this information is used in the categorisation of events. This is the approach adopted in the analysis that is described in this dissertation.

The $E_{\mathrm{T}}^{\mathrm{miss}}$ is reconstructed as described in section 2.4.6 and is used without additional requirements.

4.3 Signal and background modelling

To estimate the contributions from the signal and most backgrounds, the MC simulation is used (see section 2.3). The generated samples are normalised to the theoretical cross-section. The details on the different steps of the generation of the samples used in this analysis are summarised in table 9.

4.3.1 Signal

The matrix element (ME) calculation for the $t\bar{t}H$ process is performed with the MadGraph5_aMC@NLO (further referred to as MG5_aMC@NLO) generator [43]. The PDF set NNPDF3.0NLO [75] is used with factorisation μ_{F} and renormalisation μ_{R} scales set to $\mu_{\mathrm{F}} = \mu_{\mathrm{R}} = H_{\mathrm{T}}/2$, where H_{T} denotes the scalar sum of transverse masses $\sqrt{p_{\mathrm{T}}^2 + m^2}$ of all final state particles. The mass of the Higgs boson is set to 125 GeV and all Higgs boson decay modes are considered. The parton shower simulation was performed with

Sample	Generator	PDF	Shower	Normalisation
$t\bar{t}H$	MG5_aMC	NNPDF3.0NLO	Pythia 8.2	(N)NLO
$t\bar{t}$	Powheg-Box	CTEQ6L1	Pythia 8.2	NNLO+NNLL
W + jets	Sherpa	CT10	Sherpa 2.2.1	NNLO
Z + jets	Sherpa	CT10	Sherpa 2.2.1	NNLO
Single top (s-channel, Wt)	Powheg-Box	CT10	Pythia 6.428	aNNLO
Single top (t-channel)	Powheg-Box	CT10f4	Pythia 6.428	aNNLO
$t\bar{t}V$	MG5_aMC	NNPDF3.0NLO	Pythia 8.2	NLO
Diboson	Sherpa	CT10	Sherpa 2.1.1	NLO

Table 9: Details on the event generation for the signal and background samples used in this analysis.

Pythia 8.210 [46] using the A14 tune [76] for the underlying events modelling. The $t\bar{t}H$ cross section and the Higgs boson decay branching fractions are taken from NLO QCD and NLO QCD + EW theoretical calculations from Ref. [22]. An alternative sample interfaced to Herwig++ is used to estimate the uncertainty on signal modelling.

4.3.2 $t\bar{t}$ + jets background

The production of $t\bar{t}$ in association with jets ($t\bar{t}$+jets) is the dominant background in this analysis. Powheg-Boxv2 [44] with the NNPDF3.0NLO PDF set is used to model the inclusive $t\bar{t}$ sample. The hdamp parameter, that regulates the p_{T} of the first additional emission beyond the Born configuration, is set to 1.5 times the top quark mass ($m_t = 172.5$ GeV). Parton shower and hadronisation are modelled by Pythia 8.2 [46] with the A14 tune [76]. The sample is generated separately for $t\bar{t}$ hadronic and non-hadronic (i.e. including dileptonic and semileptonic) decay modes. To reach sufficient statistics in the high b-jet multiplicity regions that are crucial for this analysis, each of these samples is additionally generated with filters that require additional b-jets (those not originating from top quarks decay). To simulate bottom and charm hadron decays, the EvtGen v1.2.0 package [69] is used. The sample is normalised to the inclusive $t\bar{t}$ cross section of 832^{+46}_{-52} pb, calculated with top++2.0 [77] at next-to-next-to-leading order (NNLO) in QCD and including resummation of next-to-next-to-leading logarithmic (NNLL) soft-gluon terms.

The $t\bar{t}$ + jets events are divided into three categories with respect to the flavour of additional jets: $t\bar{t}+ \geq 1b$, $t\bar{t}+ \geq 1c$ and $t\bar{t}+$ light. To perform this categorisation, a spatial matching of b- and c-hadrons with particle jets is performed. Particle jets are reconstructed from all stable truth particles, except muons and neutrinos, with the anti-k_{T} algorithm with a radius parameter R = 0.4, and are additionally required to have $p_{\mathrm{T}} > 15$ GeV and $|\eta| < 2.5$. The matching allows to classify events in three exclusive categories:

- First of all, if a particle jet is located within a cone of $\Delta R < 0.3$ to a b-hadron with $p_{\mathrm{T}} > 5$ GeV not originating from a top quark decay, the event is considered as

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

- If not, then another matching procedure is performed and if there is a particle jet that is matched to a c-hadron not originating from a W boson decay, the event is defined as $t\bar{t}+ \geq 1c$.
- The events that did not satisfy any of these two matching criteria are labelled as $t\bar{t}+$ light.

A more refined classification can be considered for $t\bar{t} + \geq 1b$ events, which are of particular importance for this analysis:

- $t\bar{t} + b\bar{b}$ - two particle jets matched to a b-hadron each (excluding those from the top quarks decay),
- $t\bar{t} + b$ - a single particle jet matched to a single b-hadron,
- $t\bar{t} + B$ - a single particle jet matched to two b-hadrons, and
- $t\bar{t}+ \geq 3b$ - more than two particle jets matched to b-hadrons.

The prediction of an NLO $t\bar{t}+b\bar{b}$ sample that is generated with Sherpa+OpenLoops (in the following referred to as SherpaOL) [45], [78] is more accurate than that from the default Powheg-Box+Pythia 8, where additional b-jets are modeled with the parton shower. Thus, to improve the Powheg-Box+Pythia 8 description, the normalisation of each of the $t\bar{t} + b\bar{b}$, $t\bar{t} + b$, $t\bar{t} + B$ and $t\bar{t}+ \geq 3b$ categories at the particle level are scaled to match the predictions of the SherpaOL sample. This sample is generated with Sherpa version 2.1 and the CT10 [68, 79] four-flavour (4F) scheme PDF set. The renormalisation scale for this sample is set to the CMMPS [80] value, $\mu_{\text{CMMPS}} = \prod_{i=t,\bar{t},b,\bar{b}} E_{T,i}^{1/4}$. The factorisation scale is set to $H_{\text{T}}/2 = \frac{1}{2}\sum_i E_{T,i}$, where the sum runs over the hard-scatter partons. The resummation scale, which sets an upper bound for the hardness of the parton shower emissions, is also set to $\mu_Q = H_{\text{T}}/2$.

Alternative MC samples are used to assess uncertainties on the modelling of the different $t\bar{t}$+jets background components, as discussed in section 4.8.2.

4.3.3 Other simulated backgrounds

Other simulated backgrounds considered in this analysis are W/Z and diboson production in association with jets, ($t\bar{t}V$, $V = W, Z$ vector boson), and single top quark production (s-channel, t-channel and Wt production). The samples for W/Z+jets and diboson production in association with jets are generated using Sherpa 2.2.1.

In the case of W/Z+jets samples, the matrix element is calculated for up to two partons at NLO and four partons at leading order (LO) using Comix [81] and OpenLoops [78] and merged afterwards with the Sherpa parton shower [48] using the ME+PS@NLO prescription according to Ref. [82]. The CT10 PDF set is used. The resulting W/Z+jets events are then normalised to the NNLO cross section prediction [83].

For diboson+jets a similar approach is used, but considering zero (for WW and WZ) or one (for ZZ) additional partons at NLO and up to three additional partons at LO. The samples are normalised to the NLO cross sections.

The Wt and s-channel single top quark processes are generated with Powheg-Box 2.0 using the CT10 PDF set. To prevent overlap between $t\bar{t}$ and Wt final states, the diagram removal procedure was applied [84]. The t-channel single top-quark samples were produced with the Powheg-Box v1 generator, based on the NLO matrix element and the CT10f4 PDF set. The parton shower for all single top quark samples are simulated with Pythia 6.428 with the Perugia 2012 underlying-event tune. Bottom and charm decays were modelled with EvtGen v1.2.0. The t- and s-channel samples are normalised to the NNLO theoretical cross sections. The Wt, t- and s-channel single top quark samples are normalised to the approximate NNLO theoretical cross-sections [85], [86], [87].

In the case of $t\bar{t}V$ samples, the matrix element calculation is performed with the MG5_aMC@NLO interfaced to Pythia 8 [46] with the NNPDF3.0NLO PDF and the A14 with an underlying event tune [76]. For uncertainties on MC generator for $t\bar{t}V$ alternative samples were used. For these samples the matrix element was calculated in LO with up to two additional partons using MadGraph5 and interfaced to Pythia 8.

Single top quark produced in association with W boson and Higgs boson (tWH) samples are produced with MG5_aMC@NLO interfaced to Herwig++ [88] with the CTEQ6L1 PDF set. Samples of single top quark produced with Higgs boson and additional jets ($tHjb$) were generated with Madgraph 5 interfaced to Pythia 8, using the CT10 PDF set. Alternative samples for the case of $tHjb$ are interfaced to Herwig++ with the CTEQ6L1 PDF set.

4.3.4 Misidentified-lepton background

The contribution from events with misidentified leptons, or *fake* leptons, is a small, but non-negligible background in this analysis. Such misidentified leptons consist of jets or photons misidentified as electrons and non-prompt electrons and muons from semileptonic decays of b and c-quarks. The simulation of these processes is challenging, and thus the misidentified lepton background contribution is estimated from data. A technique known as the Matrix Method is used [89].

The method uses the difference in the efficiency of the lepton identification in the case of prompt and fake leptons. In order to take this difference into account, events are categorised into those that satisfy the lepton selection and isolation criteria, denoted as *tight selection*, and those events that satisfy less strict selection and isolation requirements (*loose selection*). The contamination of fakes is higher for the loose selection. The composition of real and fake leptons in the two selections is given by

$$N^{\text{loose}} = N^{\text{loose}}_{\text{real}} + N^{\text{loose}}_{\text{fake}}, \tag{52}$$

$$N^{\text{tight}} = N^{\text{tight}}_{\text{real}} + N^{\text{tight}}_{\text{fake}} = \epsilon_{\text{real}} N^{\text{loose}}_{\text{real}} + \epsilon_{\text{fake}} N^{\text{tight}}_{\text{fake}}, \tag{53}$$

where $\epsilon_{\text{real}}/\epsilon_{\text{fake}}$ are the fractions of real/fake leptons in the loose selection that also satisfy

the tight selection requirement. The desired number of the fake leptons in the tight selection can be found solving the above system of two equations:

$$N_{\text{fake}}^{\text{tight}} = \frac{\epsilon_{\text{fake}}}{\epsilon_{\text{real}} - \epsilon_{\text{fake}}}(\epsilon_{\text{real}} N^{\text{loose}} - N^{\text{tight}}). \tag{54}$$

The real efficiencies ϵ_{real} are obtained from $Z \to ee$ and $Z \to \mu\mu$ events. The fake efficiencies ϵ_{fake} are obtained from data samples dominated by non-prompt and fake leptons.

The fake background is estimated by applying to data an event weight

$$w_i = \frac{\epsilon_{\text{fake}}}{\epsilon_{\text{real}} - \epsilon_{\text{fake}}}(\epsilon_{\text{real}} - \delta_i), \tag{55}$$

where $\delta_i = 1$ if the loose event i passes the tight event selection and $\delta_i = 0$ otherwise. The fake and real efficiencies are parametrised as functions of lepton kinematics and the b-jet multiplicity. The total fake-lepton background yield is then given by the sum of w_i over all events [90].

4.4 Event selection

Events are selected with single-electron and single-muon triggers with different p_{T} thresholds, which are combined in a logical "OR" in order to obtain higher efficiency. Those triggers with lower p_{T} thresholds have additional lepton isolation requirements. For the 2015 and 2016 datasets different triggers are used due to the change in data-taking conditions, which was necessary to provide the same event selection rate at the higher instantaneous luminosities in 2016. In particular, the p_{T} threshold for single-electron (single-muon) triggers was increased from 24 GeV (20 GeV) in 2015 to 26 GeV in 2016. All triggers that were used for the 2015 and 2016 data in this analysis are listed in table 10.

Events are required to have one lepton and at least five jets satisfying the selection criteria described in section 4.2. Additional requirements are made based on b-tagging information. At least two jets are required to be b-tagged at the *very tight* working point or at least three jets are required to be b-tagged at the *medium* working point.

To make possible an eventual combination of results, events selected by other $t\bar{t}H$ searches are removed from the selection. These include events with two reconstructed leptons selected by the dilepton $t\bar{t}H$ ($H \to b\bar{b}$) analysis, events with at least two hadronic τ-leptons selected by the $t\bar{t}H$ multilepton search, and single-lepton events selected by the boosted $t\bar{t}H$ ($H \to b\bar{b}$) search.

Events selected for the boosted $t\bar{t}H(H \to b\bar{b})$ analysis channel are removed. Large-R jets are used to identify *boosted* (high-p_{T}) hadronically-decaying top quark and Higgs boson candidates. These are obtained via re-clustering of $R = 0.4$ jets using the anti-k_{T} algorithm with a radius parameter $R = 1.0$. A boosted Higgs boson candidate is required to be a large-R jet with $p_{\text{T}} > 200$ GeV and consist of at least two $R = 0.4$ jets among which at least two are b-tagged. A boosted top quark candidate is required to have $p_{\text{T}} > 250$ GeV and contain exactly one b-tagged jet and at least one additional not b-tagged jet. All b-tagging requirements are at the *loose* working point. Events with at least one boosted Higgs boson candidate, at least one boosted top quark candidate and at

Type	Name	p_{T} threshold, [GeV]	Isolation requirement
2015 data			
electron	HLT_e24_lhmedium_L1EM20VH	24	yes
	HLT_e60_lhmedium	60	no
	HLT_e120_lhloose	120	no
muon	HLT_mu20_iloose_L1MU15	20	yes
	HLT_mu50	50	no
2016 data			
electron	HLT_e26_lhtight_nod0_ivarloose	26	yes
	HLT_e60_lhmedium_nod0	60	no
	HLT_e140_lhloose_nod0	140	no
muon	HLT_mu26_ivarmedium	26	yes
	HLT_mu50	50	no

Table 10: Single-lepton triggers with different p_{T} thresholds used within a given data-taking year. Triggers used for each of the two datasets are combined in a logical "OR".

least one additional jet b-tagged with the *loose* working point are selected by the boosted channel. The remaining events belong to the resolved channel.

Finally, events with more than one reconstructed hadronic τ-leptons are removed to avoid overlap with the search for $t\bar{t}H$ with two τ in the final state.

4.5 Event categorisation

After selection, the data sample is dominated by the background from $t\bar{t}$ production. Preselected events are categorised into exclusive regions ("analysis regions") based on their jet multiplicity and b-tagging characteristics in order to take advantage of the higher jet and b-jet multiplicities of the $t\bar{t}H$ signal process.

Regions that provide high sensitivity to the signal (*signal regions* or SR) are those with the highest signal-to-background ratio (S/B), and signal statistical significance ($S/\sqrt{B}$), where S and B denote the number of expected signal events and the number of expected background events, respectively.

The remaining regions are referred to as *background regions* or *control regions* (CR). The control regions do not provide separation between the signal and the background, but they are used in the fit, together with the signal regions, to improve the background prediction and reduce the impact of its associated systematic uncertainties.

The regions are defined based on their background composition, taking into account

≥ 6 jets		5 jets	
Region	Definition	Region	Definition
Signal regions			
$\mathrm{SR}_1^{\geq 6\mathrm{j}}$	$> 60\%\ t\bar{t}+ \geq 2b$	$\mathrm{SR}_1^{5\mathrm{j}}$	$> 60\%\ t\bar{t}+ \geq 2b$
$\mathrm{SR}_2^{\geq 6\mathrm{j}}$	$> 45\%\ t\bar{t}+ \geq 2b$	$\mathrm{SR}_2^{5\mathrm{j}}$	$> 20\%\ t\bar{t}+ \geq 2b$
$\mathrm{SR}_3^{\geq 6\mathrm{j}}$	$> 30\%\ t\bar{t}+ \geq 2b$		
Background regions			
$\mathrm{CR}_{t\bar{t}+1b}^{\geq 6\mathrm{j}}$	$> 30\%\ t\bar{t}+1b$	$\mathrm{CR}_{t\bar{t}+1b}^{5\mathrm{j}}$	$> 20\%\ t\bar{t}+1b$
$\mathrm{CR}_{t\bar{t}+\geq 1c}^{\geq 6\mathrm{j}}$	$> 30\%\ t\bar{t}+ \geq 1c$	$\mathrm{CR}_{t\bar{t}+\geq 1c}^{5\mathrm{j}}$	$> 20\%\ t\bar{t}+ \geq 1c$
$\mathrm{CR}_{t\bar{t}+\mathrm{light}}^{\geq 6\mathrm{j}}$	Rest	$\mathrm{CR}_{t\bar{t}+\mathrm{light}}^{5\mathrm{j}}$	Rest

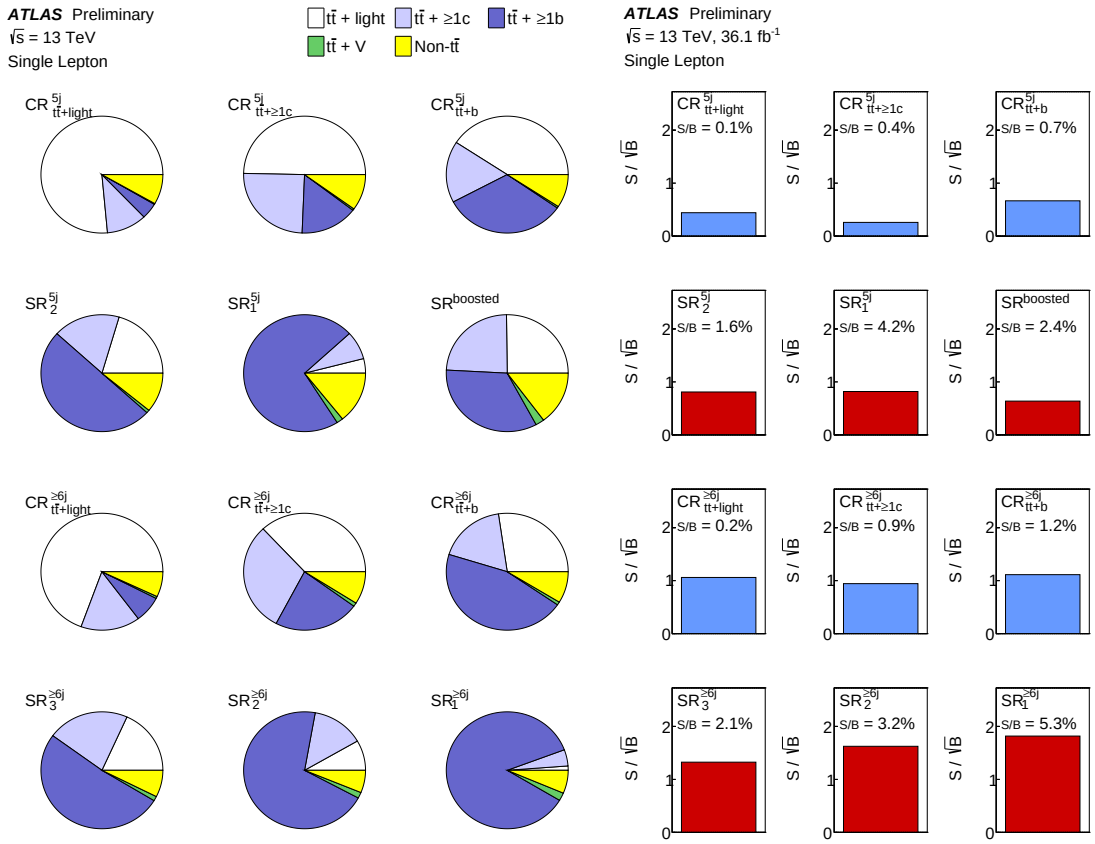
Table 11: Analysis regions in the resolved channel and their definitions with respect to the background composition: requirements on the minimum amount of the background of a given type: $t\bar{t}+ \geq 2b$, $t\bar{t}+1b$ and $t\bar{t}+ \geq 1c$. The rest of events (not satisfying any of these requirements) belong to the $t\bar{t}$ + light enriched regions ($\mathrm{CR}_{t\bar{t}+\mathrm{light}}^{5\mathrm{j}}$and $\mathrm{CR}_{t\bar{t}+\mathrm{light}}^{\geq 6\mathrm{j}}$).

information on jet multiplicity and b-tagging. Five jet grades are defined using the pseudo-continuous b-tagging approach described in section 4.2. Then, for both 5 jet and ≥ 6 jet cases, events are divided into categories based on the grade of the four jets with the highest b-tagging weight in the events. These refined categories are afterwards merged, as shown in figure 55, according to the relative amount of different sample components (see section 4.3.2): $t\bar{t}H$ and $t\bar{t}+ \geq 2$ b-jets, $t\bar{t}+1$ b-jet, $t\bar{t}+ \geq 1$ c-jets and $t\bar{t}+$ light jets. The criteria on the background composition in the resolved channel for each of the regions are presented in table 11.

The signal regions are referred to as $\mathrm{SR}_1^{\geq 6\mathrm{j}}$, $\mathrm{SR}_2^{\geq 6\mathrm{j}}$, $\mathrm{SR}_3^{\geq 6\mathrm{j}}$, $\mathrm{SR}_1^{5\mathrm{j}}$ and $\mathrm{SR}_2^{5\mathrm{j}}$ and, finally, $\mathrm{SR}^{\mathrm{boosted}}$ (formed by events that satisfy the *boosted* single-lepton requirement). The purest signal regions, i.e. those with the highest S/B ratio, are $\mathrm{SR}_1^{\geq 6\mathrm{j}}$ and $\mathrm{SR}_1^{5\mathrm{j}}$, which are selected requiring at least four jets tagged at the *very tight* working point.

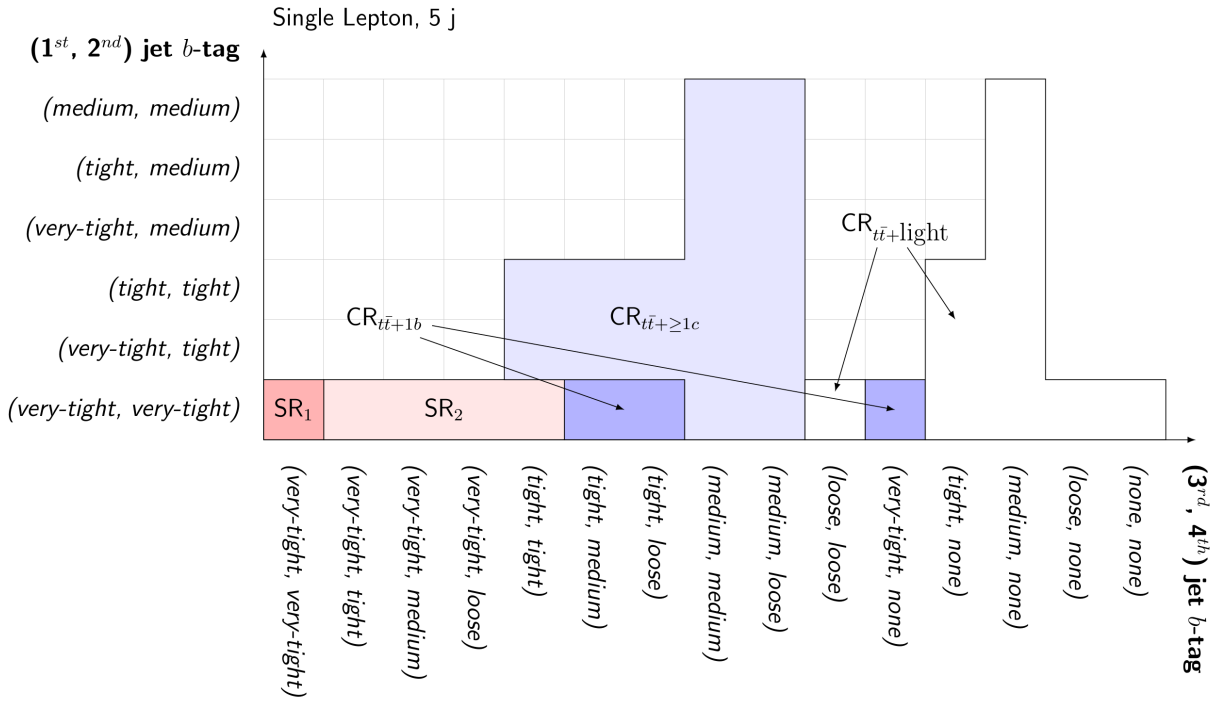
The dominant background in the signal regions is $t\bar{t}+ \geq 2$ b-jets. The control regions are $\mathrm{CR}_{t\bar{t}+1b}^{\geq 6\mathrm{j}}$, $\mathrm{CR}_{t\bar{t}+\geq 1c}^{\geq 6\mathrm{j}}$, $\mathrm{CR}_{t\bar{t}+\mathrm{light}}^{\geq 6\mathrm{j}}$, $\mathrm{CR}_{t\bar{t}+1b}^{5\mathrm{j}}$, $\mathrm{CR}_{t\bar{t}+\geq 1c}^{5\mathrm{j}}$ and $\mathrm{CR}_{t\bar{t}+\mathrm{light}}^{5\mathrm{j}}$, named to indicate their leading background component.

Figure 54 shows the background composition, S/B and $S/\sqrt{B}$ for all defined single-lepton analysis regions.

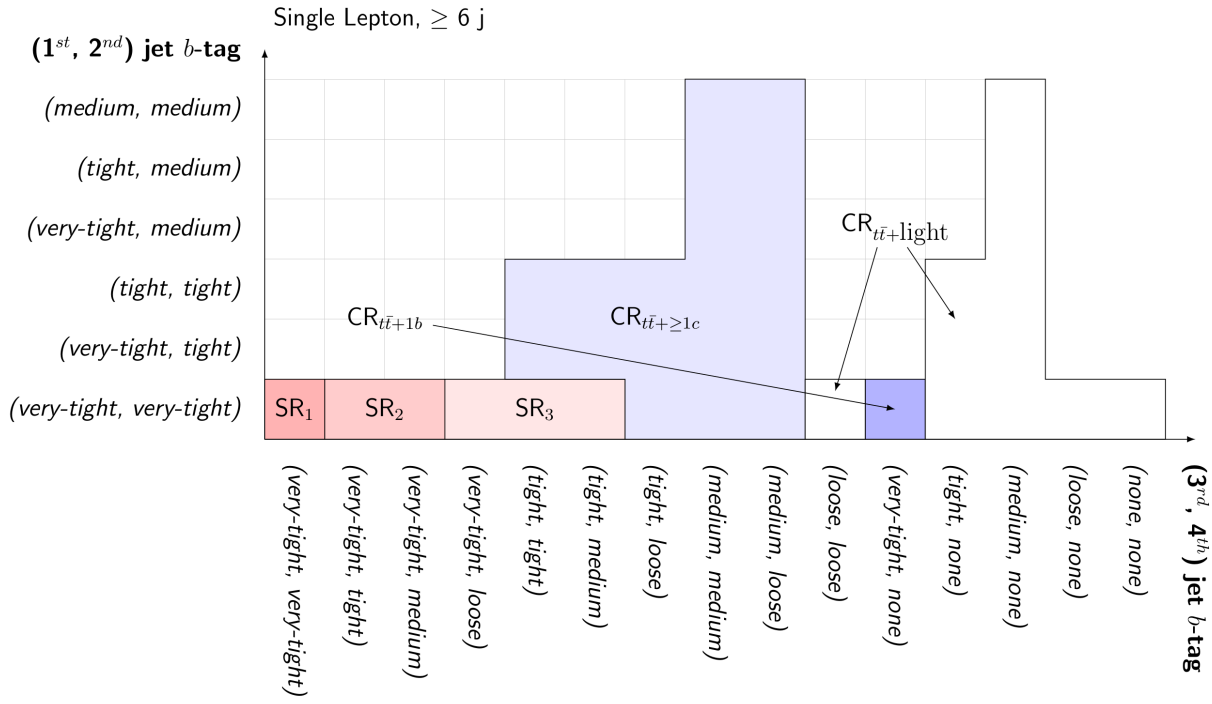


(a) (b)

Figure 54: The analysis regions: (a) background composition and (b) signal-to-background ratio (S/B) and signal statistical significance ($S/\sqrt{B}$). From Ref. [91].



(a)



(b)

Figure 55: Definition of regions based on b-tagging information in the case of (a) 5 jets and (b) ≥ 6 jets. The final categorisation into signal and background regions is shown in colors. From Ref. [91].

4.6 Signal-to-background discrimination

The small expected signal yield compared to the irreducible $t\bar{t} + \geq 1b$ background (see figure 54) makes this analysis particularly challenging. Thus, the use of effective discriminating techniques is crucial. To provide better discrimination of signal from background several methods are developed. These include three different methods that exploit the presence of a $H \to b\bar{b}$ resonance, whose outputs are afterwards combined, along with other variables, into a final multivariate discriminant.

The reconstruction boosted decision trees (BDT) method attempts to reconstruct the $t\bar{t}H$ signal topology, in particular, the kinematics of the Higgs boson. It reconstructs the $t\bar{t}H$ system by finding the best combination of assignments between the jets and the partons, with a multivariate approach.

The matrix element method evaluates the likelihood probabilities under the signal $t\bar{t}H$ and the main $t\bar{t} + b\bar{b}$ background hypotheses by computing the normalized differential cross section at the reconstruction level from the matrix elements of these processes. This method is applied only in the signal region with the highest signal-to-background ratio ($\mathrm{SR}_1^{\geq 6\mathrm{j}}$).

The likelihood discriminant method also computes the signal and background likelihoods, but using probability density functions (pdfs) derived from the MC simulation rather than from the matrix element calculation. It exploits kinematic information from all reconstructed final-state objects, and tests the events under both the signal $t\bar{t}H$ and the main $t\bar{t}$ + jets background hypotheses, considering all possible assignments between the reconstructed jets and the final-state partons. The development and optimisation of this novel method is the main contribution of this thesis. The detailed description of the technique is presented in section 4.7.

Finally, the classification BDT takes as input the information provided by the three methods described above and combines it together with other kinematic variables, as well as the information on the b-tagging weights of the selected jets, using a multivariate approach. The output of the classification BDT is the final discriminating variable used for the fit procedure in all signal regions.

4.6.1 Reconstruction BDT

This BDT technique employs the TMVA package [92]. The first step of the reconstruction BDT training process with the MC events is the matching of the reconstructed jets to the partons originating from the top quark and the Higgs boson decays. A jet is matched to a parton if their separation is $\Delta R < 0.3$. The combination of jets is considered to be correct if all six partons (b-quarks from top quark decay, b-quarks from the Higgs boson decay and light and charm quarks from the hadronic W-boson decay) are matched with jets, or if all except one quark from the W-boson decay are matched. This combination is treated as the signal in the BDT training. Other "wrong" jet assignment combinations are considered as the background.

At this point the Higgs boson, the top quarks and the W boson are reconstructed using kinematic properties for a given combination. The distributions of the invariant masses

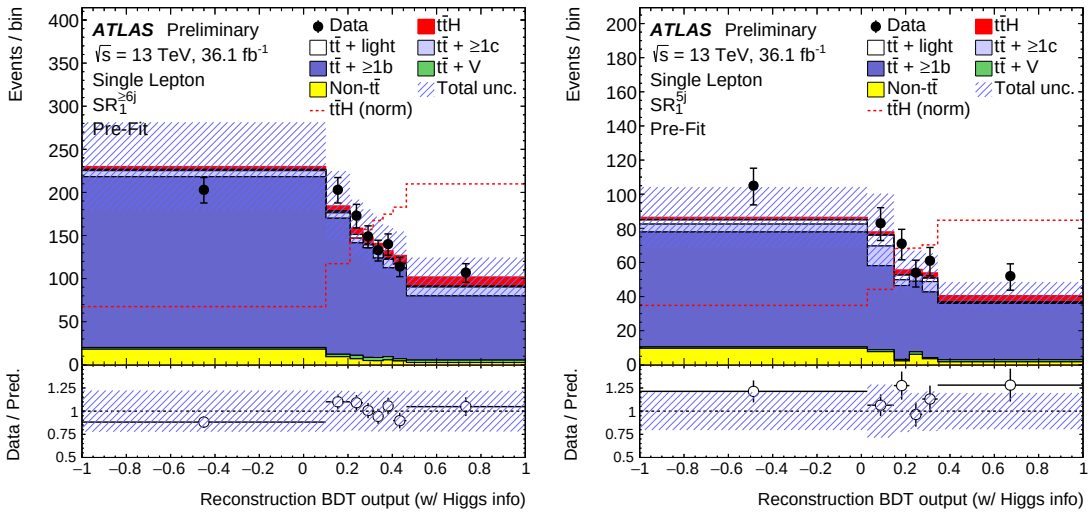


Figure 56: Distribution of the output of the reconstruction BDT with the Higgs boson kinematics in the $\mathrm{SR}_1^{\geq 6\mathrm{j}}$ and $\mathrm{SR}_1^{5\mathrm{j}}$ regions. From Ref. [91].

as well as the ΔR-separation between these objects are obtained for both signal (correct jet assignment combinations for $t\bar{t}H$ system) and background (wrong combinations).

After training is done, the BDT is evaluated for each jet assignment combination for the $t\bar{t}H$ and $t\bar{t}$ + jets MC samples. The b-tagging information is used to select only the combinations with the jets assigned to partons with corresponding flavour (b-tagged jets are assigned to b-quarks, while non-b-tagged jets are assigned to light and charm quarks from the hadronic W boson decay). The combination with the highest BDT output is chosen for the final state reconstruction.

The most powerful discriminating variables between the $t\bar{t}H$ and $t\bar{t}$ + jets are those related to the kinematic properties of the Higgs boson, so a BDT version making use of these properties provides a good performance. But when attempting to reconstruct the $t\bar{t}$ + jets events, this configuration makes the background peak under the signal in bb invariant mass distribution. Therefore two versions of the reconstruction BDT are used. The first uses in training the reconstructed jets that correspond to the six quarks of the $t\bar{t}H$ system. It considers the Higgs boson related variables. The second takes into account only the variables corresponding to the top quarks and the W boson. In this case only the jets matched to the four quarks from the $t\bar{t}$ system are considered in the training.

The output distributions of the reconstruction BDT with Higgs kinematics in the two main SRs are presented in figure 56.

The reconstruction BDT with (without) the Higgs kinematics information correctly reconstructs the Higgs boson in the $\mathrm{SR}_1^{\geq 6\mathrm{j}}$ in 49% (33%) of the events. A detailed description of the method can be found in Ref. [93].

4.6.2 Matrix element method

The principle of the matrix element method (MEM) is to evaluate the likelihood of an event to originate from either the signal ($t\bar{t}H$) or the background ($t\bar{t}+b\bar{b}$), based on the matrix element for these two processes. The MEM was used for the Run 1 $t\bar{t}H$ ($H \to b\bar{b}$) searches by both ATLAS [74] and CMS [94].

For each event two likelihoods are calculated under the signal and background hypotheses:

$$L_{S/B} = \Sigma \int \frac{f_1(x_1, Q^2) f_2(x_2, Q^2)}{|\vec{q_1}||\vec{q_2}|} \left|M_{S/B}(\mathbf{Y})\right|^2 T(\mathbf{X}, \mathbf{Y}) \mathrm{d}\mathbf{\Phi}_n(\mathbf{Y}), \tag{56}$$

where the sum runs over the possible initial partonic configurations, and all possible jet-parton assignments. f_1 and f_2 are the PDFs for two initial state partons carrying fractions x_1 and x_2 of the proton momentum in a collision at energy Q. The LO matrix element $M_{S/B}$ is calculated for a phase space configuration $\mathbf{Y}$ at the parton level for either the signal or background processes. The connection between the parton-level phase space ($\mathbf{Y}$) and the reconstructed in the detector objects ($\mathbf{X}$) is provided by transfer functions $T(\mathbf{X}, \mathbf{Y})$, which describe the probabilities of the reconstructed objects to originate from this partonic configuration. The phase space factor $\mathrm{d}\mathbf{\Phi}_n(\mathbf{Y})$ allows to take into account the unknown parameters, in particular, the neutrino longitudinal momentum.

MG5_aMC@NLO is used for the leading-order matrix element evaluation. The pdfs are modelled with the CT10 set interfaced with the LHAPDF package [95].

The final discriminating variable is then given by

$$MEM_{D_1} = \log_{10} L_S - \log_{10} L_B. \tag{57}$$

The distribution of the MEM_{D_1} variable in the $\mathrm{SR}_1^{\geq 6\mathrm{j}}$ region is shown in figure 57.

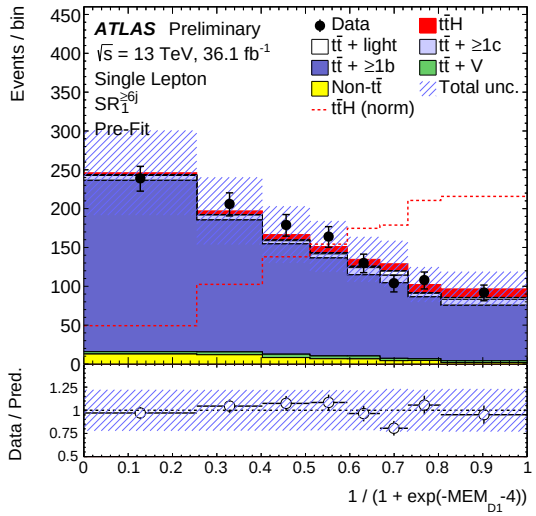


Figure 57: Distribution of the MEM_{D_1} discriminant in the $\mathrm{SR}_1^{\geq 6\mathrm{j}}$ region. From Ref. [91].

4.6.3 Classification BDT

The outputs of the reconstruction BDT, likelihood discriminant method and matrix element method are combined together with general kinematic variables, such as the invariant masses of pairs of reconstructed jets and leptons and masses between them, as well as pseudo-continuous b-tagging information into the classification BDT. The full list of variables used as input for the classification BDT is presented in table 12. Like the reconstruction BDT, it is also based on the TMVA package [92]. The distributions of the classification BDT in all signal regions are presented in figures 58 and 59.

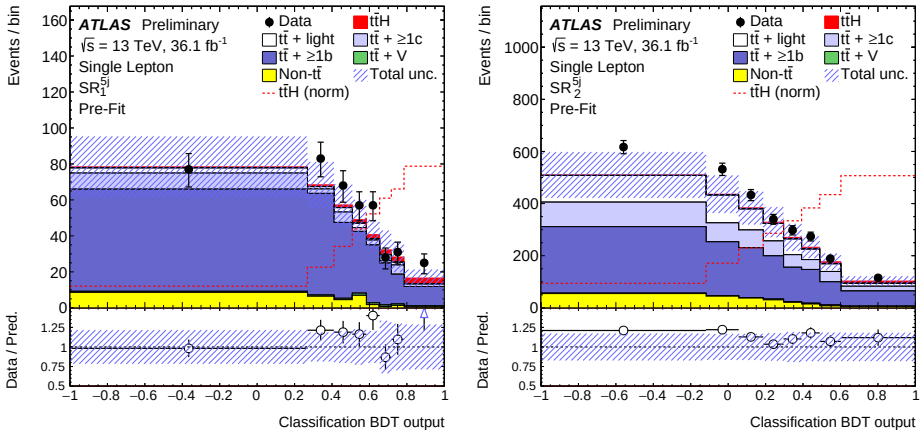


Figure 58: Distributions of the classification BDT in the signal regions with 5 jets.

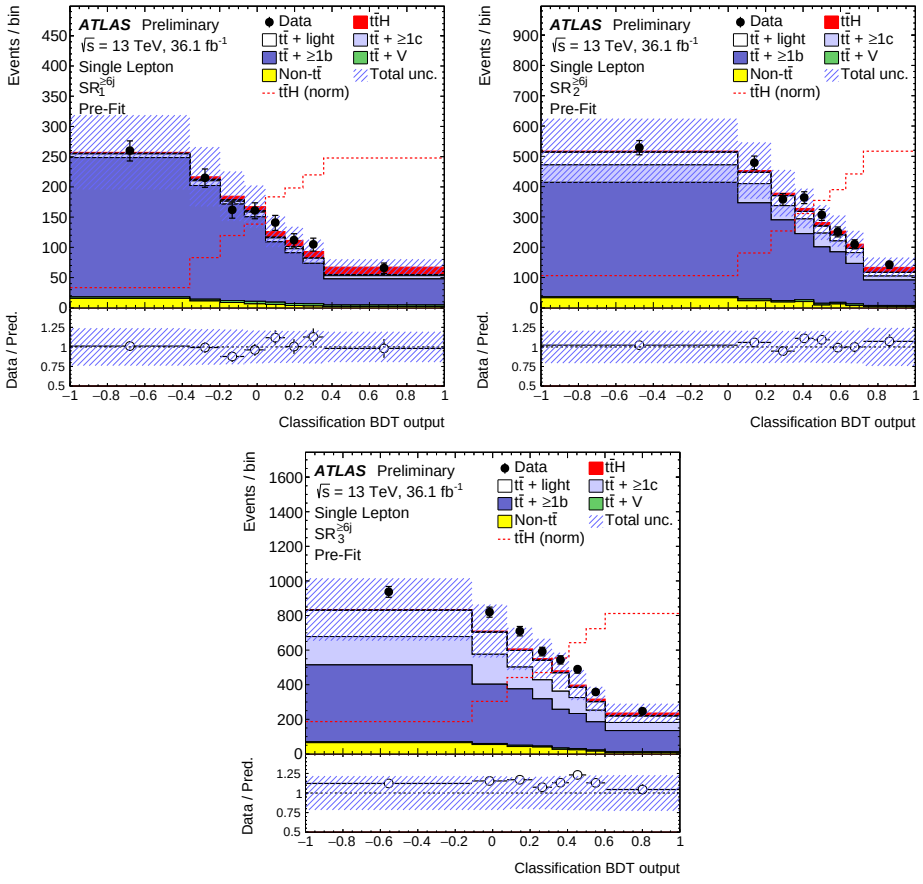


Figure 59: Distributions of the classification BDT in the signal regions with ≥6 jets. From Ref. [91].

Variable	Definition	Region	
		$\geq$ 6j	5j
General kinematic variables			
$\Delta R_{bb}^{\text{avg}}$	Average ΔR for all b-tagged jet pairs	✓	✓
$\Delta R_{bb}^{\text{max } p_{\text{T}}}$	ΔR between the two b-tagged jets with the largest vector sum p_{T}	✓	–
$\Delta\eta_{jj}^{\text{max } \Delta\eta}$	Maximum $\Delta\eta$ between any two jets	✓	✓
$m_{bb}^{\text{min } \Delta R}$	Mass of the combination of the two b-tagged jets with the smallest ΔR	✓	–
$m_{jj}^{\text{min } \Delta R}$	Mass of the combination of any two jets with the smallest ΔR	–	✓
$\text{N}_{30}^{\text{Higgs}}$	Number of b-jet pairs with invariant mass within 30 GeV of the Higgs boson mass	✓	✓
$H_{\text{T}}^{\text{had}}$	Scalar sum of jet p_{T}	–	✓
$\Delta R_{\text{lep}-bb}^{\text{min } \Delta R}$	ΔR between the lepton and the combination of the two b-tagged jets with the smallest ΔR	–	✓
Aplanarity	$1.5\lambda_2$, where λ_2 is the second eigenvalue of the momentum tensor built with all jets	✓	✓
$H1$	Second Fox–Wolfram moment computed using all jets and the lepton	✓	✓
Variables from reconstruction BDT output			
BDT	BDT output	✓*	✓*
m_H	Higgs boson mass	✓	✓
$m_{H,b_{\text{lep top}}}$	Mass of Higgs boson and b-jet from leptonic top	✓	–
$\Delta R_{\text{Higgs bb}}$	ΔR between b-jets from the Higgs boson	✓	✓
$\Delta R_{H,t\bar{t}}$	ΔR between Higgs boson and $t\bar{t}$ system	✓*	✓*
$\Delta R_{H,\text{lep top}}$	ΔR between Higgs boson and leptonic top	✓	–
$\Delta R_{H,b_{\text{had top}}}$	ΔR between Higgs boson and b-jet from hadronic top	–	✓*
Variable from Likelihood calculation			
LHD	Likelihood discriminant	✓	✓
Variable from Matrix Method calculation			
MEM_{D1}	Matrix Method	✓	–
Variables from b-tagging			
w_b^H	Sum of binned b-tagging weights of jets from best Higgs candidate	✓	✓
B_{j^3}	3rd jet binned b-tagging weight (sorted by weight)	✓	✓
B_{j^4}	4th jet binned b-tagging weight (sorted by weight)	✓	✓
B_{j^5}	5th jet binned b-tagging weight (sorted by weight)	✓	✓

Table 12: Classification BDT input variables in 6 jets and 5 jets signal regions. Variables from the reconstruction BDT labeled with * are from the BDT using Higgs boson information, others are from the reconstruction BDT without Higgs boson information. The MEM_{D1} variable is only used in the signal region with the highest signal-to-background ratio ($\geq$ 6 jets, $\geq$ 4 are b-tagged at 60% WP), while b-tagging weights are not used in this region).

4.7 Likelihood discriminant

The method for discriminating the signal from the background presented in this section is based on a so-called combinatorial likelihood approach. Probabilities of a given event under the signal $P^{\text{sig}}(\mathbf{x})$ or background $P^{\text{bkg}}(\mathbf{x})$ hypotheses (see figure 60) are computed with the aid of MC-based probability density functions (pdfs). The pdfs are functions of the four-momentum vectors $\mathbf{x}$ of reconstructed objects in this event: the jets, the lepton and the neutrino. The final discriminating variable is defined as

$$D = \frac{P^{\text{sig}}(\mathbf{x})}{P^{\text{sig}}(\mathbf{x}) + P^{\text{bkg}}(\mathbf{x})}. \tag{58}$$

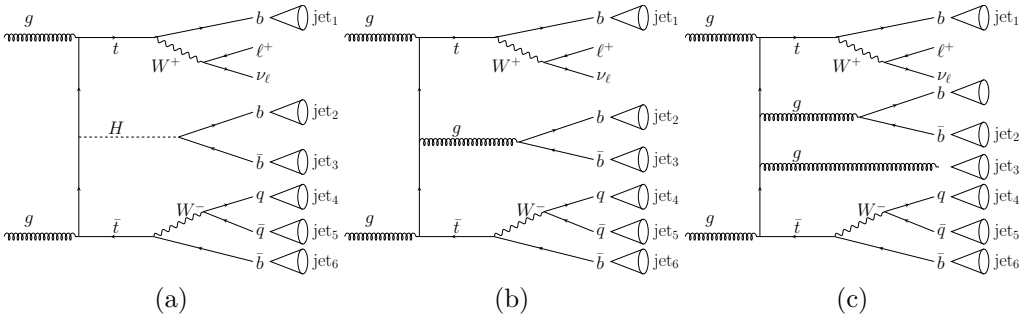


Figure 60: Representative Feynman diagrams illustrating the partonic configurations and parton-jet assignments for the (a) the signal $t\bar{t}H$ and (b,c) the background $t\bar{t}$+jets hypotheses.

Various invariant mass resonances provide useful information to separate the signal from the background: the Higgs boson mass for the signal hypothesis, and the masses of the leptonic top quark, the hadronic top quark and the hadronic W boson for both signal and background hypotheses. The pdfs of these invariant masses are the most significant ones used in this method, and they are presented in sections 4.7.1 and 4.7.2. Other pdfs exploited in the method correspond to additional mass variables, described in section 4.7.3, and several angular variables, described in section 4.7.4.

The signal probability P^{sig} is defined as the product of the probabilities of the invariant masses in this event (see figure 60 (a)): the leptonic top quark mass $M_{t_l}(l, \nu, b_l)$, the hadronic top quark mass $M_{t_h}(q_1, q_2, b_h)$, the hadronic W boson mass $M_{W_h}(q_1, q_2)$ and the Higgs boson mass $M_H(b_1, b_2)$. The background probability P^{bkg} is defined in a similar fashion, but using the pdf for the invariant mass of additional jets b_1 and b_2 instead of the Higgs boson mass.

The distributions of the invariant masses are obtained from simulated signal and background events using the four-momentum vectors of the reconstructed lepton and jets, and the $E_{\text{T}}^{\text{miss}}$. The partonic origin of jets is identified by applying a so-called truth-matching procedure: a jet is defined to be matched to a quark if this quark is within a cone $\Delta R < 0.3$ to the jet. The histograms filled with these mass distributions are normalised to unit area

and used as pdfs in the calculation of the signal and background probabilities. A smoothing procedure is applied in the high mass range of the pdfs, where the MC statistics is limited.

4.7.1 Signal probability

Higgs boson invariant mass

A very important feature that can be exploited for discriminating the signal from the background is the presence of the Higgs boson mass resonance. The Higgs boson invariant mass $M_H(b_1, b_2)$ pdf is built from the jets that are matched to two b-quarks from the Higgs boson decay (b_1 and b_2) in signal MC events (see figure 61).

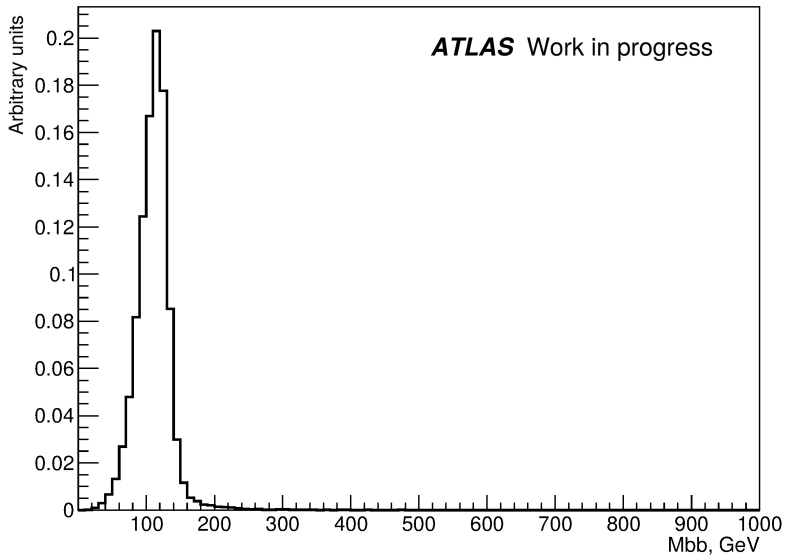


Figure 61: Pdf for the Higgs boson invariant mass in $t\bar{t}H$ MC events with ≥ 6 jets and ≥ 4 b-tagged jets (at the *tight* working point).

Leptonic top quark invariant mass

The invariant mass of the leptonic top quark, $M_{t_l}(l, \nu, b_l)$, is reconstructed using the four-momenta of the jet matched to the b_l, the lepton and the neutrino. The longitudinal component of the neutrino momentum is not known, but it can be calculated using the constraint from the measured value of the W boson mass, $M_W = 80.4$ GeV, which provides a quadratic equation with one unknown (p_{z_ν}):

$$M_W^2 = (p_l + p_\nu)^2. \tag{59}$$

When the discriminant of this quadratic equation is positive ($\Delta > 0$), there are two solutions:

$$p_{z_\nu}^{\pm} = \frac{p_{z_l}\beta \pm \sqrt{\Delta}}{2(E_l^2 - p_{z_l}^2)}, \tag{60}$$

where

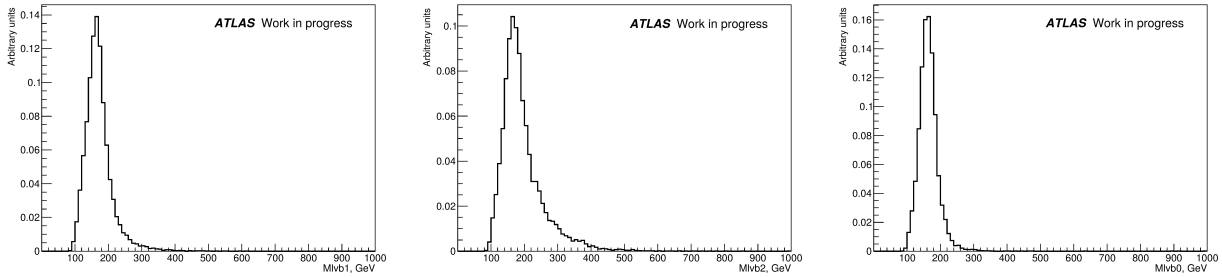
$$\beta = M_W^2 - M_l^2 + 2p_{x_l}p_{x_\nu} + 2p_{y_l}p_{y_\nu}, \tag{61}$$

$$\Delta = E_l^2(\beta^2 + (2p_{z_l}p_{T_\nu})^2 - (2E_l p_{T_\nu})^2). \tag{62}$$

In the case of two neutrino solutions, they are ordered with respect to $|p_{z_\nu}|$, so that $|p_{z_\nu 1}| < |p_{z_\nu 2}|$. It was determined that in $\sim 65\%$ of the signal events $p_{z_\nu 1}$ is closer to the truth neutrino p_z than $p_{z_\nu 2}$. Two separate pdfs for the leptonic top quark invariant mass are built, each corresponding to each neutrino solution ($p_{z_\nu 1}$ or solution 1, and $p_{z_\nu 2}$ or solution 2). The $P^{\text{sig}}(M_{t_l})$ is then constructed using both of them, via calculating two probabilities and summing them with different weights: 0.65 for solution 1 and 0.35 for solution 2.

Due to the finite resolution on the $E_{\text{T}}^{\text{miss}}$ measurement the quadratic equation 59 does not have a real solution in $\sim 35\%$ of the signal events. In this case the solution for p_{z_ν} is approximated: the $E_{\text{T}}^{\text{miss}}$ is varied until $\Delta = 0$ and one neutrino solution p_{z_ν} is obtained.

The pdfs for the different neutrino solutions discussed are presented in figure 62.

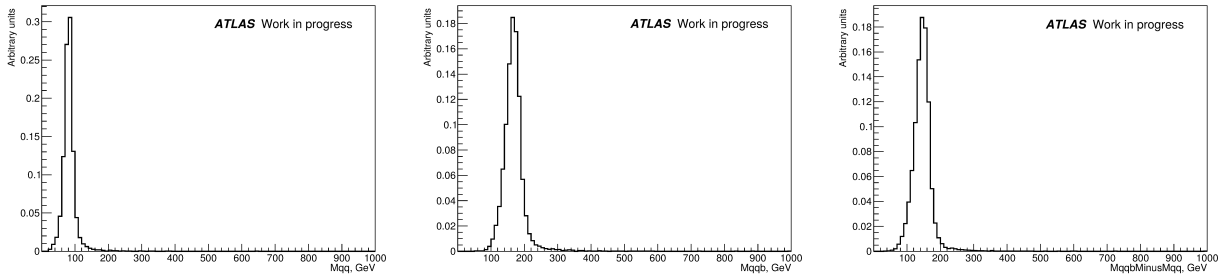


(a) (b) (c)

Figure 62: Pdfs for the leptonic top quark mass M_{t_l} in the case of two neutrino solutions, (a) solution 1 and (b) solution 2, and in (c) no real solution case, in the $t\bar{t}H$ MC events with ≥ 6 jets and ≥ 4 b-tagged jets (at the *tight* working point).

Hadronic W boson and hadronic top quark invariant masses

The pdfs for the invariant masses of the hadronic W boson $M_{W_h}(q_1, q_2)$ and the hadronic top quark $M_{t_h}(q_1, q_2, b_h)$ are built in a similar way based on information from the jet truth-matching. However, these two invariant masses are correlated, which would render suboptimal the construction of the signal and background probabilities as a product of one-dimensional pdfs. Therefore in the final probability calculation, instead of the hadronic top quark mass, M_{t_h} the difference between the hadronic top quark and the



(a) (b) (c)

Figure 63: Pdfs for (a) hadronic W boson mass M_{W_h}, (b) hadronic top quark mass M_{t_h} and (c) $M_{t_h} - M_{W_h}$ in $t\bar{t}H$ MC events with ≥ 6 jets and ≥ 4 b-tagged jets (at the *tight* working point).

hadronic W boson masses $M_{t_h} - M_{W_h}$ is used. These three distributions are presented in figure 63. Figure 64 shows that $M_{t_h} - M_{W_h}$ is less correlated with M_{W_h} than M_{t_h}.

Sum over all jet permutations and *b*-tagging weights

Considering the above kinematic variables, the expression for the signal probability is given by:

$$P^{\text{sig}}_{\text{kin}} = P^{\text{sig}}(M_H)P^{\text{sig}}(M_{t_l})P^{\text{sig}}(M_{t_h} - M_{W_h})P^{\text{sig}}(M_{W_h}). \tag{63}$$

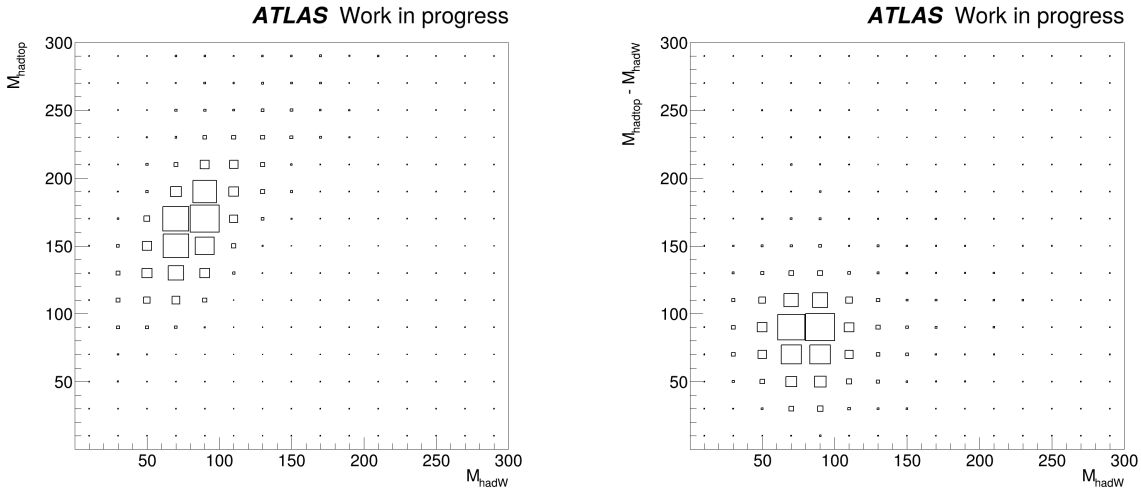
However, as the partonic origin of the jets is not known, the signal probability must be calculated summing over all possible jet permutations N_p in the event. The b-tagging information is then used to give different weights to permutations. The expression for the signal probability becomes

$$P^{\text{sig}} = \frac{\sum_{k=1}^{N_p} P^{\text{sig}}_{\text{kin}} P^{\text{sig}}_{\text{btag}}}{\sum_{k=1}^{N_p} P^{\text{sig}}_{\text{btag}}}, \tag{64}$$

where $P^{\text{sig}}_{\text{kin}}$ is given by equation 63 and the b-tagging term $P^{\text{sig}}_{\text{btag}}$ is defined as

$$P^{\text{sig}}_{\text{btag}} = P_b(\text{jet}_1)P_b(\text{jet}_2)P_b(\text{jet}_3)P_l(\text{jet}_4)(f_l P_l(\text{jet}_5) + f_c P_c(\text{jet}_5))P_b(\text{jet}_6). \tag{65}$$

In this expression $\text{jet}_i (i = 1, ..6)$ denotes a reconstructed jet, and $P_f(\text{jet}_i)$ represent the probability that jet_i originates from a parton of flavour f. These probabilities are computed using the jet MV2c10 b-tagging weight $w(\text{jet}_i)$. The estimated b-tagging efficiencies for jets tagged at different working points and not tagged are used to evaluate the probabilities. If a jet_i has a weight $w(\text{jet}_i)$ between the threshold values for the two working points WP_1 and WP_2: $w_{WP_1} < w(\text{jet}_i) < w_{WP_2}$, then $P_f(\text{jet}_i) = \epsilon_f^{WP_1} - \epsilon_f^{WP_2}$, where $\epsilon_f^{WP_1}$, $\epsilon_f^{WP_2}$ are the b-, c- and light efficiencies at *loose*, *medium*, *tight* and *very-tight* working



(a) (b)

Figure 64: Mass of the hadronic top quark mass vs the hadronic W boson mass (a) and difference between the hadronic top quark and W boson masses vs W boson mass (b) in $t\bar{t}H$ MC events with ≥ 6 jets and ≥ 4 b-tagged jets (at the *tight* working point).

points and the boundary values (0% and 100%). The weighted sum $f_l P_l(\text{jet}_5) + f_c P_c(\text{jet}_5)$ is used to take into account two possible decay channels of W boson: to u,d and to c,s. The weights f_l and f_c are the truth fractions of two W boson decay modes in the MC events.

4.7.2 Background probability

The dominant background in the main signal region (≥ 6 jets, ≥ 4 b-jets) is $t\bar{t}+ \geq 1$ additional b-jets. Therefore, two background hypotheses are considered:

(A) $t\bar{t}+ \geq 2$ additional b-jets, which makes $\sim 80\%$ events (see figure 60 (b)), and

(B) $t\bar{t} + 1$ additional b-jet, which makes the $\sim 20\%$ events (see figure 60 (c)).

These fractions are obtained in $t\bar{t}$ MC events with ≥ 6 jets and ≥ 4 jets b-tagged at the *tight* working point.

For a given event, the background probability P^{bkg} is calculated in a similar way as P^{sig}, i. e. exploiting the invariant masses of the resonances: the leptonic top quark, the hadronic top quark and the hadronic W boson. In order to keep P^{bkg} in the same dimensionality as P^{sig}, the pdf for the invariant mass of two additional b-jets, $M_{b_1 b_2}$, is used in the same way as the Higgs boson invariant mass pdf in the calculation of P^{sig}.

For hypothesis (A) the pdf is constructed using the invariant mass of two highest-p_{T} additional (i.e. not matched to $t\bar{t}$ decay products) b-labelled jets, for hypothesis (B) the single additional b-labelled jet and the highest-p_{T} additional jet not b-labelled. Two pdfs are then used in the P^{bkg} calculation with different weights, corresponding to the fractions

of the (A) and (B) topologies in background events: in events with ≥ 6 jets and ≥ 4 jets b-tagged at the *tight* working point the weights are $f_A = 0.8$ and $f_B = 0.2$. These two pdfs are displayed in figure 65, where they are compared to that for $t\bar{t}H$ signal events. Most of the discrimination between signal and background comes from this variable.

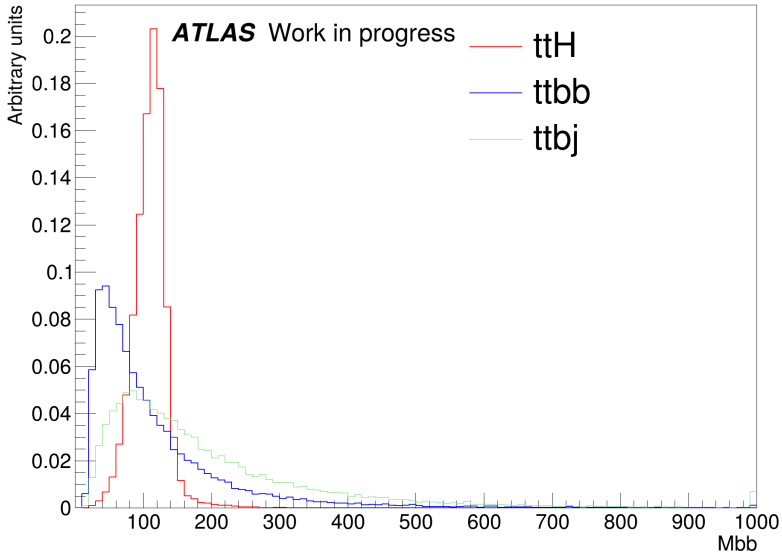


Figure 65: Pdfs for the Higgs boson invariant mass (red) in the $t\bar{t}H$ MC events and two extra jets invariant mass in $t\bar{t}+ \geq 2$ additional b-jets (blue) and $t\bar{t} + 1$ additional b-jet (green) cases in $t\bar{t}$ MC events with ≥ 6 jets and ≥ 4 b-tagged jets (at the *tight* working point).

The background probability is defined in a similar way as the signal probability, but with additional sum over the two hypotheses (A, B):

$$P^{\text{bkg}} = \frac{\sum_{k=1}^{N_p} \sum_{j=\text{A,B}} f_j P_{\text{kin}}^{\text{bkg,j}} P_{\text{btag}}^{\text{bkg,j}}}{\sum_{k=1}^{N_p} \sum_{j=\text{A,B}} f_j P_{\text{btag}}^{\text{bkg,j}}}, \tag{66}$$

where the kinematic terms for the two hypotheses are given by

$$P_{\text{kin}}^{\text{bkg,A}} = P^{\text{bkg}}(M_{b_1 b_2}) P^{\text{bkg}}(M_{t_l}) P^{\text{bkg}}(M_{t_h} - M_{W_h}) P^{\text{bkg}}(M_{W_h}), \tag{67}$$

$$P_{\text{kin}}^{\text{bkg,B}} = P^{\text{bkg}}(M_{b,j}) P^{\text{bkg}}(M_{t_l}) P^{\text{bkg}}(M_{t_h} - M_{W_h}) P^{\text{bkg}}(M_{W_h}). \tag{68}$$

and the b-tagging terms are defined as

$$P_{\text{btag}}^{\text{bkg,A}} = P_b(\text{jet}_1) P_b(\text{jet}_2) P_b(\text{jet}_3) P_l(\text{jet}_4) (f_l P_l(\text{jet}_5) + f_c P_c(\text{jet}_5)) P_b(\text{jet}_6), \tag{69}$$

$$P_{\text{btag}}^{\text{bkg,B}} = P_b(\text{jet}_1)P_b(\text{jet}_2)(f_l P_l(\text{jet}_3) + f_c P_c(\text{jet}_3))P_l(\text{jet}_4)(f_l P_l(\text{jet}_5)+ \\ + f_c P_c(\text{jet}_5))P_b(\text{jet}_6). \quad (70)$$

The probabilities $P_f(\text{jet}_i)$ and weights f_l and f_c are computed in the same way as those in equation 65.

The two hypotheses A and B are combined by summing the corresponding probabilities with weights f_j ($j = \text{A}, \text{B}$), which are the truth fractions for the two hypotheses in the MC events.

4.7.3 Additional invariant mass variables

The kinematic probabilities terms $P_{\text{kin}}^{\text{sig}}$ and $P_{\text{kin}}^{\text{bkg}}$ as defined in equations 63, 67 and 68 exploit the invariant mass distributions of different resonances in the event. However, there are additional invariant mass terms that can be used to improve the separation between the signal and the background: the invariant mass of the $t\bar{t}$ system, $M_{t\bar{t}}$, and the invariant mass of the $t\bar{t} + b\bar{b}$ system, $M_{t_h t_l b_1 b_2}$.

These two invariant masses depend on the neutrino solution p_{z_ν} in the same way as the leptonic top quark invariant mass (see section 4.7.1). Thus, the pdfs are derived separately for the different neutrino solutions. The illustrative pdfs shown in this section correspond to the "solution 1" case.

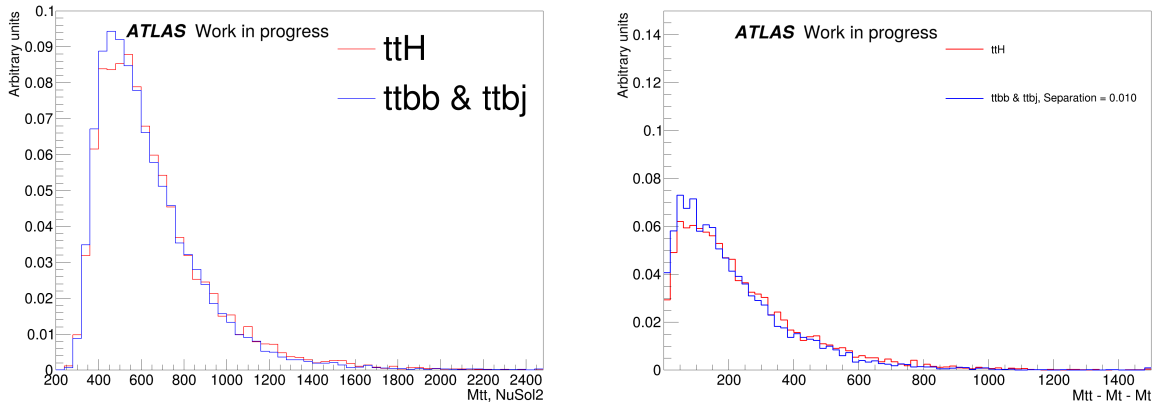
The invariant mass of the $t\bar{t}$ system $M_{t\bar{t}}$ is correlated with the invariant masses of the top quarks M_{t_l} and M_{t_h}. Therefore, the mass difference $M_{t\bar{t}} - M_{t_l} - M_{t_h}$ is used instead. The distributions for $M_{t\bar{t}}$ and $M_{t\bar{t}} - M_{t_l} - M_{t_h}$ are presented in figure 66. Figure 67 shows that $M_{t\bar{t}} - M_{t_l} - M_{t_h}$ has significantly less correlations with both mass variables M_{t_l} and M_{t_h} than $M_{t\bar{t}}$ for the $t\bar{t}H$ signal events. The pdfs corresponding to these variables are derived in the same way for the signal and the background events.

The invariant mass of the $t\bar{t} + b\bar{b}$ system, $M_{t_h t_l b_1 b_2}$, is built from the four-momentum vectors of the leptonic top quark, hadronic top quark and the two b-jets b_1 and b_2, in the case of the signal and background hypothesis (A), and the single b-jet and the highest-p_{T} additional not b-labelled jet j, for background hypothesis (B). As all three sets of pdfs (the signal hypothesis and the two background hypothesis) are built in the same way, for convenience the notation $M_{t_h t_l b_1 b_2}$ will refer to all of them. The correlations of the invariant mass of the $t\bar{t} + b\bar{b}$ system $M_{t_h t_l b_1 b_2}$ with $M_{t_l t_h}$ and $M_{b_1 b_2}$ are reduced by using instead $M_{t_h t_l b_1 b_2} - M_{t_h t_l} - M_{b_1 b_2}$. The pdfs for $M_{t_h t_l b_1 b_2}$ and $M_{t_h t_l b_1 b_2} - M_{t_h t_l} - M_{b_1 b_2}$ are presented in figure 68; the two-dimensional distributions are shown in figure 69.

Therefore, the expression for the kinematic term for the signal probability becomes

$$P_{\text{kin}}^{\text{sig}} = P^{\text{sig}}(M_H)P^{\text{sig}}(M_{t_l})P^{\text{sig}}(M_{t_h} - M_{W_h})P^{\text{sig}}(M_{W_h})\times \\ \times P^{\text{sig}}(M_{t_h t_l} - M_{t_h} - M_{t_l})P^{\text{sig}}(M_{t_h t_l b_1 b_2} - M_{t_l t_h} - M_{b_1 b_2}). \quad (71)$$

and for the background probability



(a) (b)

Figure 66: Pdfs for (a) $M_{t_h t_l}$ and (b) $M_{t_h t_l} - M_{t_l} - M_{t_h}$ in $t\bar{t}H$ (red) and $t\bar{t}$ (blue) MC events with ≥ 6 jets and ≥ 4 b-tagged jets (at the *tight* working point).

$$P_{\text{kin}}^{\text{bkg}} = P^{\text{sig}}(M_{b_1 b_2}) P^{\text{sig}}(M_{t_l}) P^{\text{sig}}(M_{t_h} - M_{W_h}) P^{\text{sig}}(M_{W_h}) \times \\ \times P^{\text{sig}}(M_{t_h t_l} - M_{t_h} - M_{t_l}) P^{\text{sig}}(M_{t_h t_l b_1 b_2} - M_{t_l t_h} - M_{b_1 b_2}), \quad (72)$$

respectively.

4.7.4 Angular variables

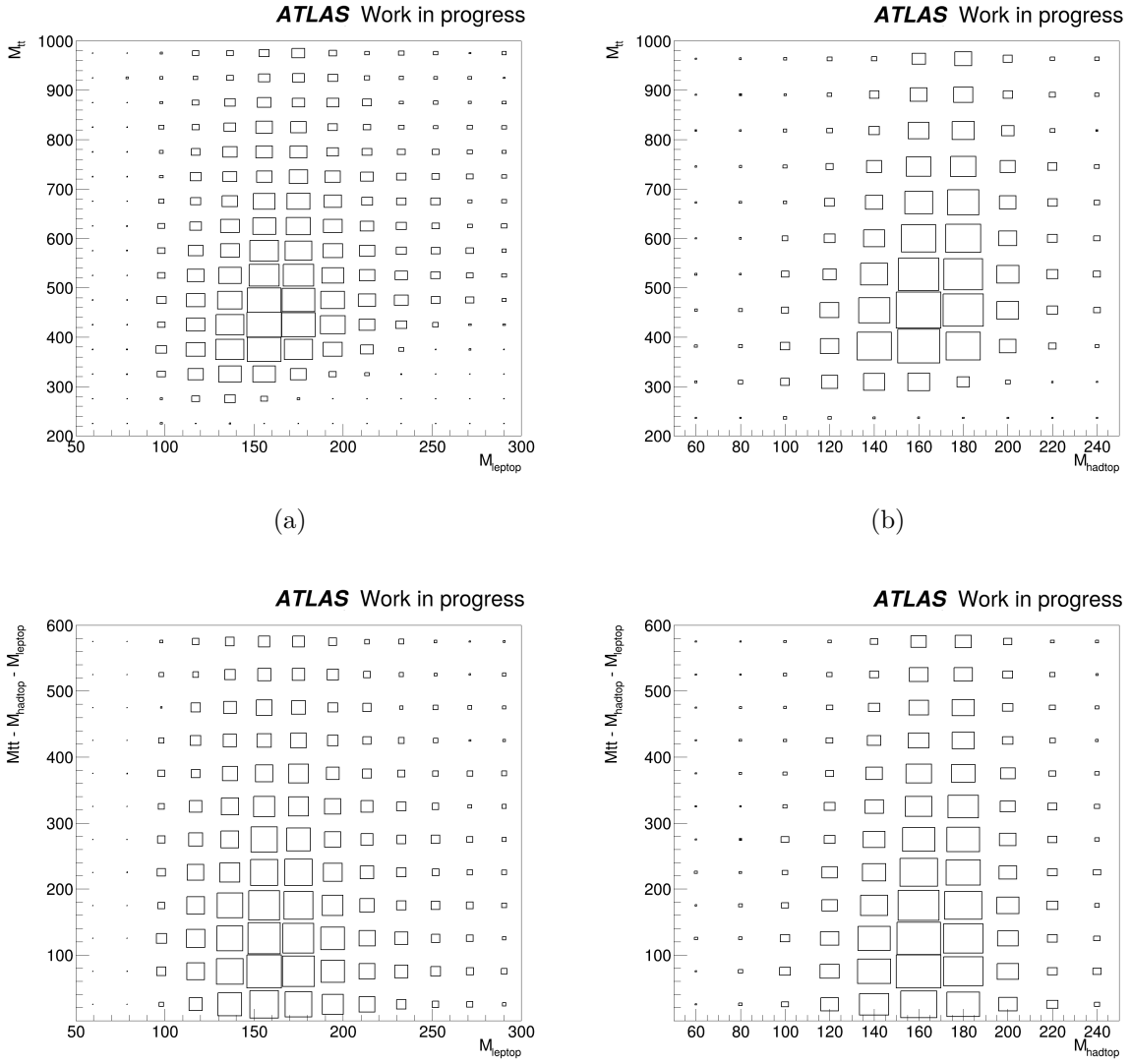
Apart from the invariant masses, additional discrimination power can be gained by exploiting information on the spin of the different resonances, in particular the Higgs boson. The most discriminating variables are:

- $\cos\theta^*_{b,bb}$, where θ^* is the angle between the b_1 direction in the $b_1 b_2$ system rest-frame and the direction of the momentum of the $b_1 b_2$ system in the laboratory frame.
- $\cos\theta^*_{bb,ttbb}$, where θ^* is the angle between the $b_1 b_2$ system in the $t\bar{t} + 2b$ system rest-frame and the direction of the momentum of the $t\bar{t} + 2b$ system in the laboratory frame.

The two definitions apply to both $t\bar{t}H$ and $t\bar{t}+ \geq 2b$ background hypotheses. In the case of the $t\bar{t} + 1b$ hypothesis, the $b_1 b_2$ system is replaced by the bj system. The distributions of these variables are presented in figure 70.

Additionally the following angular variables were tested:

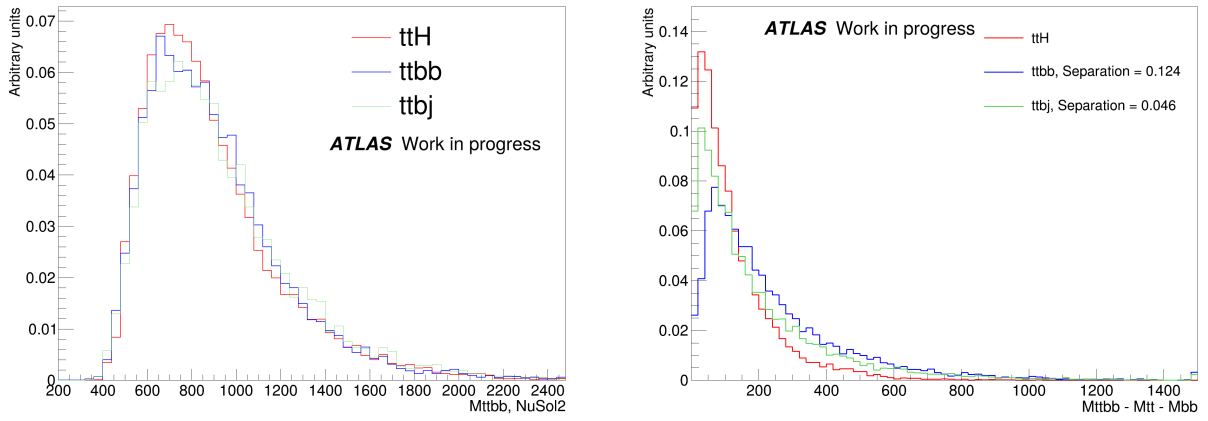
- $\cos\theta^*_{b_l,t_l}$, where θ^* is the angle between the b_l direction in the t_l rest frame and the leptonic top quark direction in the laboratory rest frame,



(a) (b)

(c) (d)

Figure 67: Two-dimensional distributions of $M_{t_h t_l}$ vs (a) M_{t_l} and (b) M_{t_h}, and $M_{t\bar{t}} - M_{t_l} - M_{t_h}$ vs (c) M_{t_l} and (d) M_{t_h} in $t\bar{t}H$ MC events with ≥ 6 jets and ≥ 4 b-tagged jets (at the *tight* working point).



(a) (b)

Figure 68: Pdfs for (a) $M_{t_h t_l b_1 b_2}$ and (b) $M_{t_h t_l b_1 b_2} - M_{t_h t_l} - M_{b_1 b_2}$ in $t\bar{t}H$ (red) and $t\bar{t}$ + jets (blue and green) MC events with ≥ 6 jets, ≥ 4 b-tagged jets (at the *tight* working point).

- $\cos\theta^*_{b_h,t_h}$, where θ^* is the angle between the b_h direction in the t_h rest frame and the t_h direction in the laboratory rest frame, and
- $\cos\theta^*_{q_W,W}$, where θ^* is the angle between the direction of the W boson jet q_W in the hadronic W boson rest frame and the hadronic W boson direction in the laboratory rest frame.

The corresponding pdfs are shown in figure 71.

The impact of these additional variables in the final discriminating power was found to be small, thus only the two angular variables $\cos\theta^*_{b,bb}$ and $\cos\theta^*_{bb,ttbb}$ are used in the calculation.

To include the angular variables, the $P^{\text{sig/bkg}}$ is now defined as

$$P^{\text{sig/bkg}}_{\text{kin}} = P^{\text{sig/bkg}}_{\text{mass}} \times P^{\text{sig/bkg}}_{\text{ang}}, \tag{73}$$

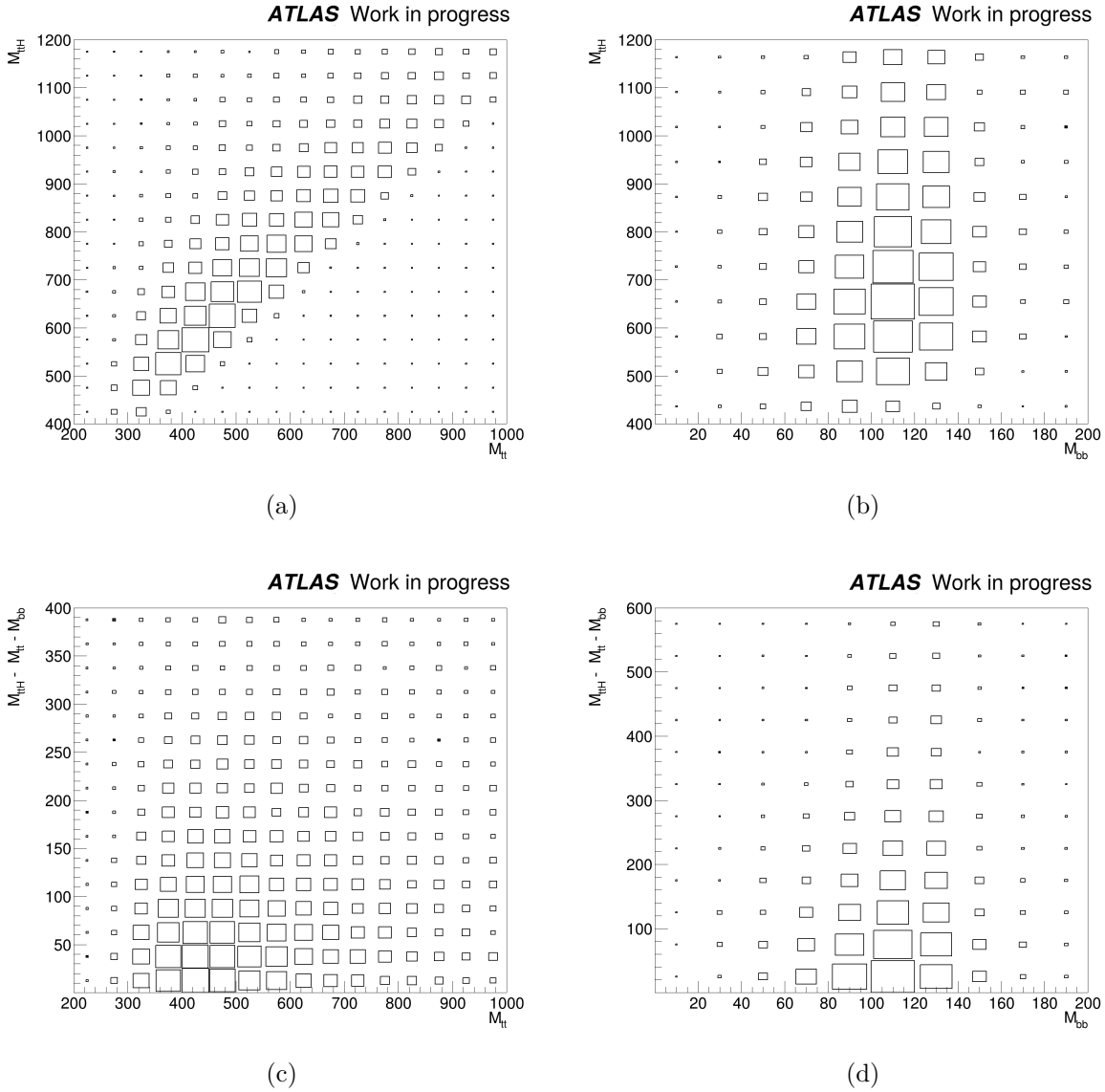
where $P^{\text{sig/bkg}}_{\text{mass}}$ is defined in equations 71 and 72, and $P^{\text{sig/bkg}}_{\text{ang}}$ is

$$P^{\text{sig/bkg}}_{\text{ang}} = P^{\text{sig/bkg}}(\cos\theta^*_{b,bb}) P^{\text{sig/bkg}}(\cos\theta^*_{bb,ttbb}). \tag{74}$$

Figure 72 shows the correlation coefficients for the final set of variables considered, demonstrating they are mostly uncorrelated.

4.7.5 Missing jet hypothesis

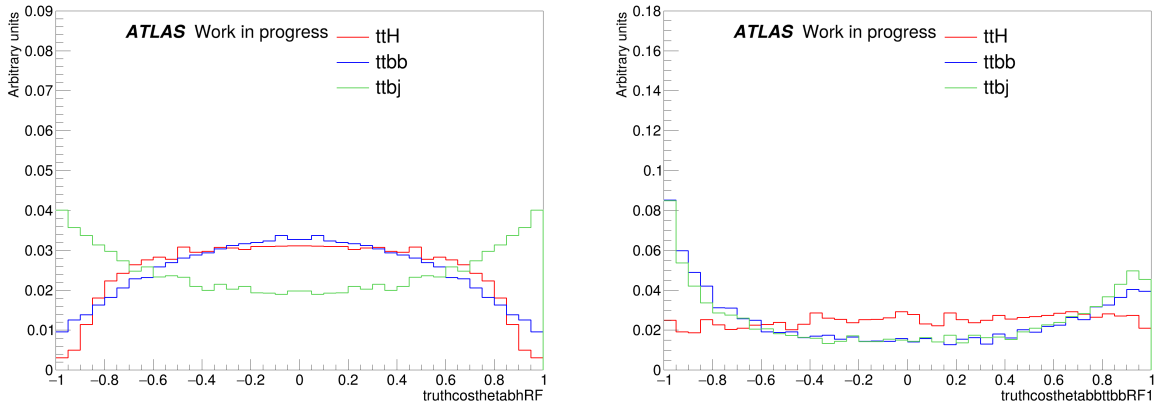
The signal and background probabilities defined in equations 71 - 74 are built under the assumption that the jets originating from the Higgs boson, the top quarks and the W boson are reconstructed and pass the selection criteria. However, often there are jets that fail the p_{T} and η requirements. It was verified that only in $\sim 40\%$ of the signal events



(a) (b)

(c) (d)

Figure 69: Two-dimensional distributions of $M_{t_h t_l b_1 b_2}$ vs $M_{t_h t_l}$ (a) and $M_{b_1 b_2}$ (b), and $M_{t_h t_l b_1 b_2} - M_{t_l} - M_{t_h}$ vs $M_{t_h t_l}$ (c) and $M_{b_1 b_2}$ (d) in $t\bar{t}H$ MC events with ≥ 6 jets, ≥ 4 b-tagged jets (at the *tight* working point).



(a) (b)

Figure 70: Pdfs for the two most important angular variables (a) $\cos\theta^*_{bb,ttbb}$ and (b) $\cos\theta^*_{bb,ttbb}$ in $t\bar{t}H$ (red) and $t\bar{t}$ (blue and green) MC events with ≥ 6 jets, ≥ 4 b-tagged jets (at the *tight* working point).

with ≥ 6 jets, ≥ 4 b-tagged jets (at the *tight* working point) all six partons from the $t\bar{t}H$ decay can be matched to selected jets. In $\sim 36\%$ of the events only 5 partons are matched to jets, while the remaining $\sim 24\%$ of events have < 5 matches (see fractions for other regions in table 13). This means that to describe most of the events correctly, one needs to introduce an additional "missing jet" hypothesis and combine it with the existing "all jets matched" hypothesis using as weights the fraction of events for each of the hypotheses (according to table 13).

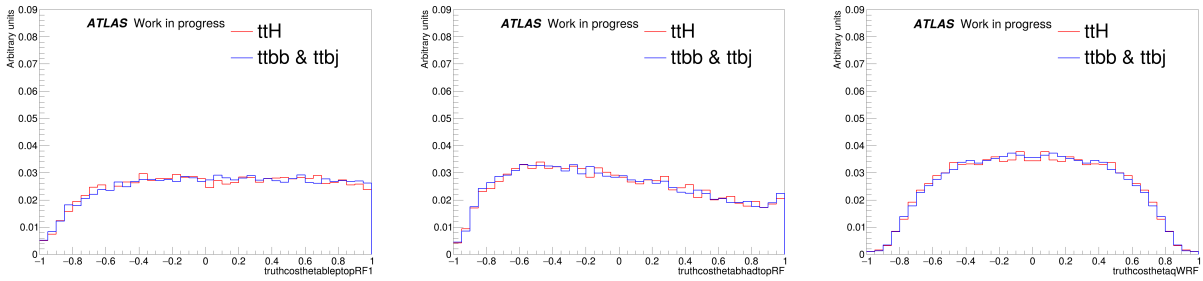
	All 6 jets	One missing jet	Two or more missing jets
≥ 6 jets, ≥ 4 b-jets	40%	36%	24%
5 jets, ≥ 4 b-jets	-	58%	42%
≥ 6 jets, 3 b-jets	24%	36%	40%

Table 13: Fraction of events with given number of partons matched to jets in $t\bar{t}H$ events for different jet and b-jet multiplicities. The number of b-jets is defined at the *tight* b-tagging working point.

≥ 6 jets, ≥ 4 b-jets

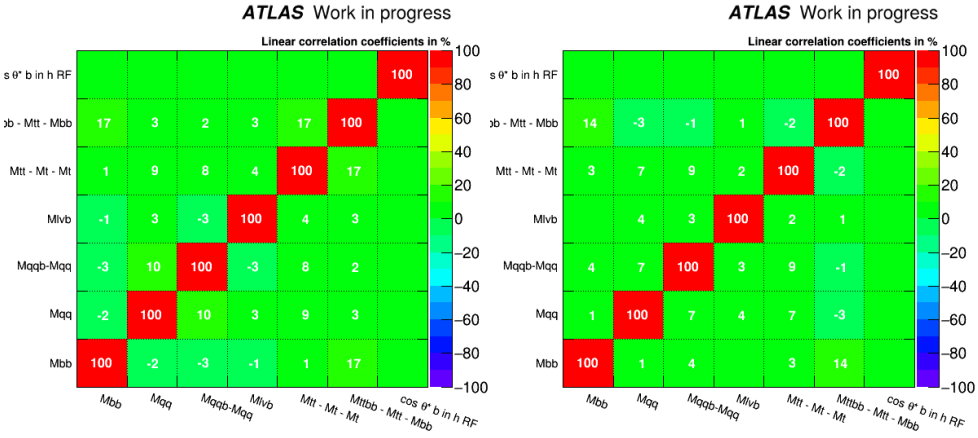
In more than 70% of the $t\bar{t}H$ events with ≥ 6 jets and ≥ 4 b-jets (at the *tight* working point) the missing jet corresponds to the softest parton in the W boson decay, as shown in figure 73 (a). Therefore, this assumption is used for the "missing jet hypothesis".

The probabilities to be the signal and background are defined for this hypothesis in the same way as described by equations 71 - 74, but those probability terms that were calculated using pdfs that contain information from jets from W boson q_1 and q_2 are now



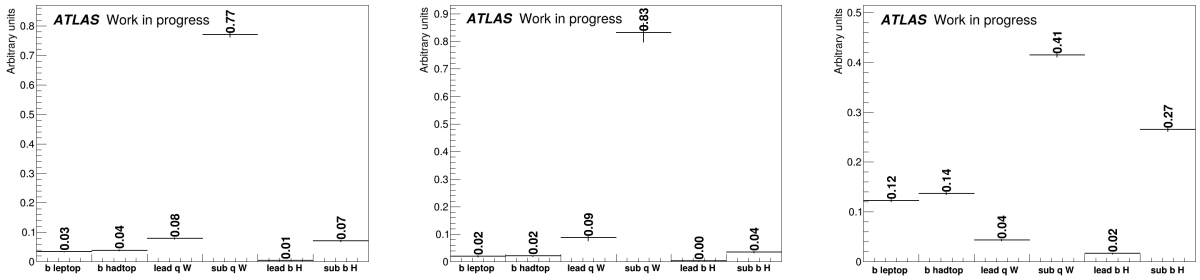
(a) (b) (c)

Figure 71: Pdfs for additional angular variables tested: (a) $\cos\theta^*_{b_l,t_l}$, (b) $\cos\theta^*_{b_h,t_h}$ and (c) $\cos\theta^*_{q_W,W}$ in $t\bar{t}H$ (red) and $t\bar{t}$ (blue) MC events with ≥ 6 jets, ≥ 4 b-tagged jets (at the *tight* working point).



(a) (b)

Figure 72: Correlation matrix for the variables used in the final discriminant calculation: (a) signal and (b) background.



(a) (b) (c)

Figure 73: Distributions of the partonic origin of the missing jet in the $t\bar{t}H$ events with (a) ≥ 6 jets, ≥ 4 b-jets, (b) 5 jets, ≥ 4 b-jets and (c) ≥ 6 jets, 3 b-jets. The number of b-jets is evaluated at the *tight* working point.

All jets hypothesis	Missing jet hypothesis
$M_H(b_1, b_2)$	$M_H(b_1, b_2)$
$M_{t_l}(l, \nu, b_l)$	$M_{t_l}(l, \nu, b_l)$
$M_{W_h}(q_1, q_2)$	$M_{W'_h}(q_W, q_X)$
$[M_{t_h} - M_{W_h}](q_1, q_2, b_h)$	$[M_{t'_h} - M_{W'_h}](q_W, q_X, b_h)$
$[M_{t_h t_l} - M_{t_h} - M_{t_l}](l, \nu, b_l, q_1, q_2, b_h)$	$[M_{t'_h t_l} - M_{t'_h} - M_{t_l}](l, \nu, b_l, q_W, q_X, b_h)$
$[M_{t_h t_l b_1 b_2} - M_{t_l t_h} - M_H](b_1, b_2, l, \nu, b_l, q_1, q_2, b_h)$	$[M_{t'_h t_l b_1 b_2} - M_{t_l t'_h} - M_H](b_1, b_2, l, \nu, b_l, q_W, q_X, b_h)$
$cos\theta^*_{b,bb}(b_1, b_2)$	$cos\theta^*_{b,bb}(b_1, b_2)$
$cos\theta^*_{bb,ttbb}(b_1, b_2, l, \nu, b_l, q_1, q_2, b_h)$	$cos\theta^*_{bb,tt'bb}(b_1, b_2, l, \nu, b_l, q_W, q_X, b_h)$

Table 14: Summary of the pdfs used in the signal probability calculation for events with ≥ 6 jets under the "all jets" and "missing jet" hypotheses.

replaced by pdfs built with one jet from the W boson q_W and the highest-p_{T} non-matched jet q_X. For example, the hadronic W boson invariant mass $M_{W_h}(q_1, q_2)$ is replaced by the invariant mass $M_{W'_h}(q_W, q_X)$. The full list of variables used in the signal probability calculation for the two hypotheses is presented in table 14 (the background probability is calculated accordingly).

5 jets, ≥ 4 *b*-jets

Introducing the "missing jet" hypothesis allows to calculate likelihood discriminant for events with 5 jets. The calculation is performed in a similar way as for the ≥ 6 jets, ≥ 4 b-jets region, but only considering the "missing jet" hypothesis, which is built under the same assumption that most of the times the missing jet originates from the hadronic W boson (see figure 73(b)). The full set of variables used for the signal probability calculation for 5 jets, ≥ 4 b-jets events is presented in table 15.

Missing jet hypothesis
$M_H(b_1, b_2)$
$M_{t_l}(l, \nu, b_l)$
$M_{t''_h}(q_W, b_h)$
$[M_{t''_h t_l} - M_{t''_h} - M_{t_l}](l, \nu, b_l, q_W, b_h)$
$[M_{t''_h t_l b_1 b_2} - M_{t_l t''_h} - M_H](b_1, b_2, l, \nu, b_l, q_W, b_h)$
$cos\theta^*_{b,bb}(b_1, b_2)$
$cos\theta^*_{bb,tt''bb}(b_1, b_2, l, \nu, b_l, q_W, b_h)$

Table 15: Summary of the pdfs used in the signal probability calculation for events with 5 jets (only the "missing jet" hypothesis is considered).

$\geq$ 6 jets, 3 *b*-jets

In this region there are several hypotheses of which jet is missing (see figure 73(c)). As the hypothesis of a missing jet from the W boson is still the leading one, it was used exactly in the same way as for the $\geq$ 6 jets, $\geq$ 4 b-jets (with the variables summarised in table 14). Introducing new hypotheses in this region can be considered as a possible future optimisation.

4.7.6 Final discriminant and performance

The likelihood discriminant was calculated as defined by equation 82, including all refinements discussed previously for the construction of the signal and background probabilities.

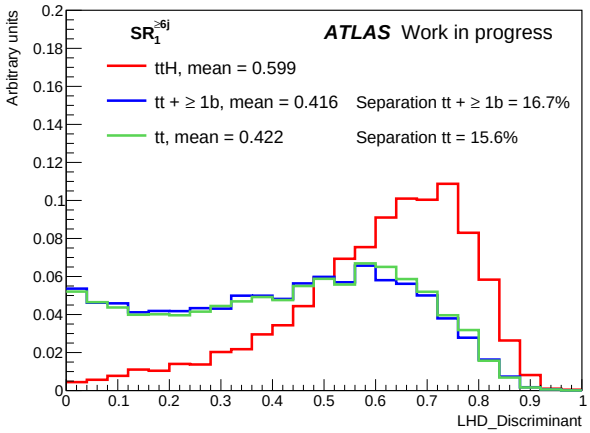
The performance of a discriminating variable is evaluated using the signal-to-background separation power defined as

$$S = \frac{1}{2} \sum_{i}^{N_{bins}} \frac{(N_i^S - N_i^B)^2}{N_i^S + N_i^B}. \tag{75}$$

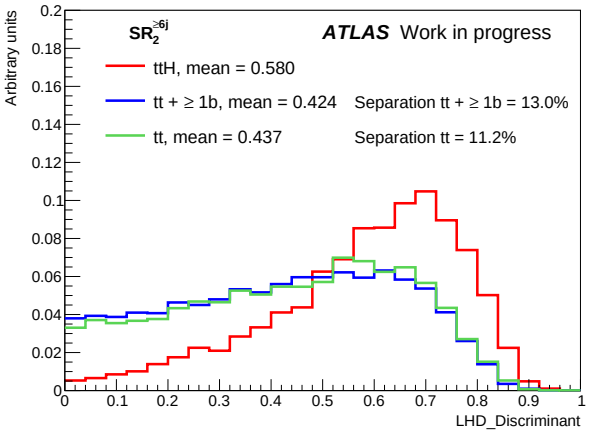
where N_i^S and N_i^B denote the number of signal and background events in bin i of the distribution respectively.

The final likelihood discriminant (LHD) variable becomes an input for the classification BDT together with the reconstruction BDT, the MEM discriminant and other kinematic variables in all signal regions. In the $\mathrm{SR}_1^{\geq 6\mathrm{j}}$, $\mathrm{SR}_2^{\geq 6\mathrm{j}}$ and $\mathrm{SR}_3^{\geq 6\mathrm{j}}$ regions it is calculated under the $\geq$ 6 jets, $\geq$ 4 b-jets hypothesis as described in section 4.7.5, while in the $\mathrm{SR}_1^{5\mathrm{j}}$ and $\mathrm{SR}_2^{5\mathrm{j}}$ regions it is calculated under the $\geq$ 5 jets, $\geq$ 4 b-jets hypothesis presented in section 4.7.5. The distributions of the final discriminant are shown in figure 74.

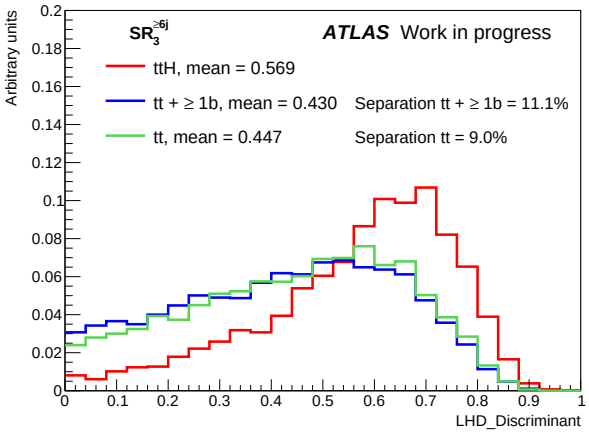
In order to test the agreement between data and simulation, the LHD variable is computed in background regions. In particular, it is computed in the $\mathrm{CR}^{\geq 6\mathrm{j}}_{t\bar{t}+\geq 1c}$ region under the $\geq$ 6 jets, $\geq$ 4 b-jets hypothesis and in the $\mathrm{CR}^{\geq 6\mathrm{j}}_{t\bar{t}+1b}$ region under the $\geq$ 6 jets, 3 b-jets hypothesis (as described in section 4.7.5). In the case of the $\mathrm{CR}^{\geq 6\mathrm{j}}_{t\bar{t}+\mathrm{light}}$ region, the LHD variable is evaluated under the $\geq$ 6 jets, $\geq$ 4 b-jets hypothesis in the case of two jets b-tagged at the *very tight* and two at the *loose* working points, and under $\geq$ 6 jets, 3



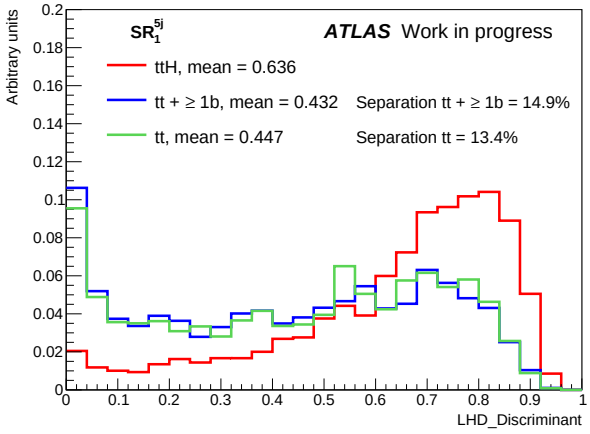
(a)



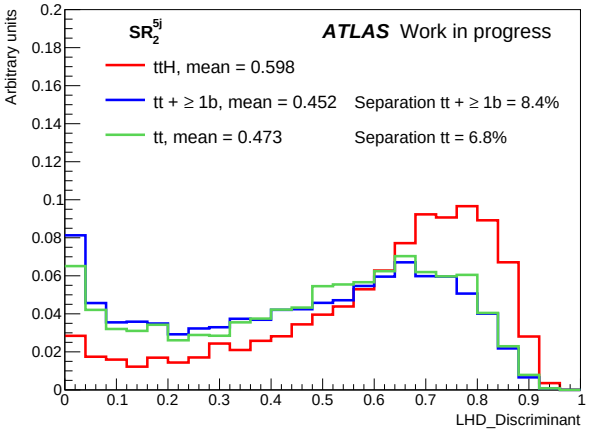
(b)



(c)



(d)



(e)

Figure 74: Final likelihood discriminant distribution for the $t\bar{t}H$ signal (red) and background $tt+ \geq 1$ b-jet (blue) and $t\bar{t}$+jets inclusive (green) MC events in (a) $\mathrm{SR}_1^{\geq 6\mathrm{j}}$, (b) $\mathrm{SR}_2^{\geq 6\mathrm{j}}$, (c) $\mathrm{SR}_3^{\geq 6\mathrm{j}}$, (d) $\mathrm{SR}_1^{5\mathrm{j}}$ and (e) $\mathrm{SR}_2^{5\mathrm{j}}$ regions. The separation power defined by equation 75 is shown.

Variable	Separation [%]
LHD	14.5
MEM_{D1}	12.0
Reconstruction BDT output	9.2
$\Delta R_{bb}^{\mathrm{avg}}$	7.5
m_H	5.3
$\Delta R_{bb}^{\mathrm{max}\ p_{\mathrm{T}}}$	5.3
$\mathrm{N}_{30}^{\mathrm{Higgs}}$	5.1
$\Delta\eta_{jj}^{\mathrm{max}\ \Delta\eta}$	5.0
$\Delta R_{H,t\bar{t}}$	4.8
$m_{bb}^{\mathrm{min}\ \Delta R}$	4.7
$\Delta R_{H,\mathrm{lep\ top}}$	3.1
Aplanarity	3.0
$\Delta R_{\mathrm{Higgs\ bb}}$	2.6
$m_{H,b_{\mathrm{lep\ top}}}$	2.5
$H1$	1.3

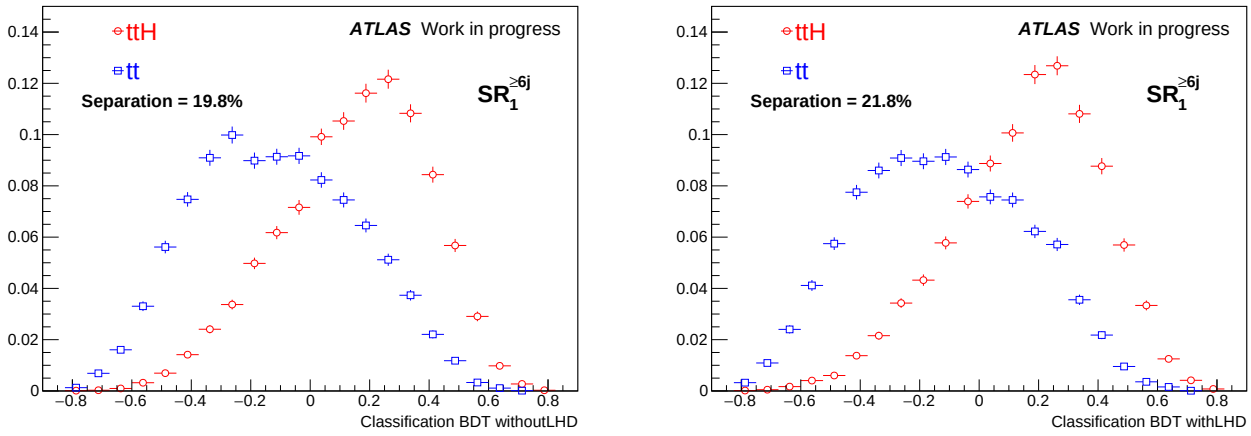
Table 16: Values of separation power for classification BDT input variables in $\mathrm{SR}_1^{\geq 6\mathrm{j}}$.

b-jets hypothesis in the case of two jets b-tagged at the *very tight* and one at the *tight* or *medium* working points (and the remaining jets in the event not tagged). For the rest of the events in the $\mathrm{CR}_{t\bar{t}+\mathrm{light}}^{\geq 6\mathrm{j}}$ region the LHD variable is not calculated. In the $\mathrm{CR}_{t\bar{t}+1b}^{5\mathrm{j}}$ and $\mathrm{CR}_{t\bar{t}+\geq 1c}^{5\mathrm{j}}$ regions the LHD variable is calculated under the ≥ 5 jets, ≥ 4 b-jets hypothesis. Finally, in the $\mathrm{CR}_{t\bar{t}+\mathrm{light}}^{5\mathrm{j}}$ region the calculation is performed only in the case of two jets b-tagged at the *very tight* and two at the *loose* working points, under the ≥ 5 jets, ≥ 4 b-jets hypothesis.

The LHD variable was found to be the single most discriminating variable in the analysis. Table 16 presents the variables that are used as input to the classification BDT and the corresponding values of the separation power in the $\mathrm{SR}_1^{\geq 6\mathrm{j}}$ region.

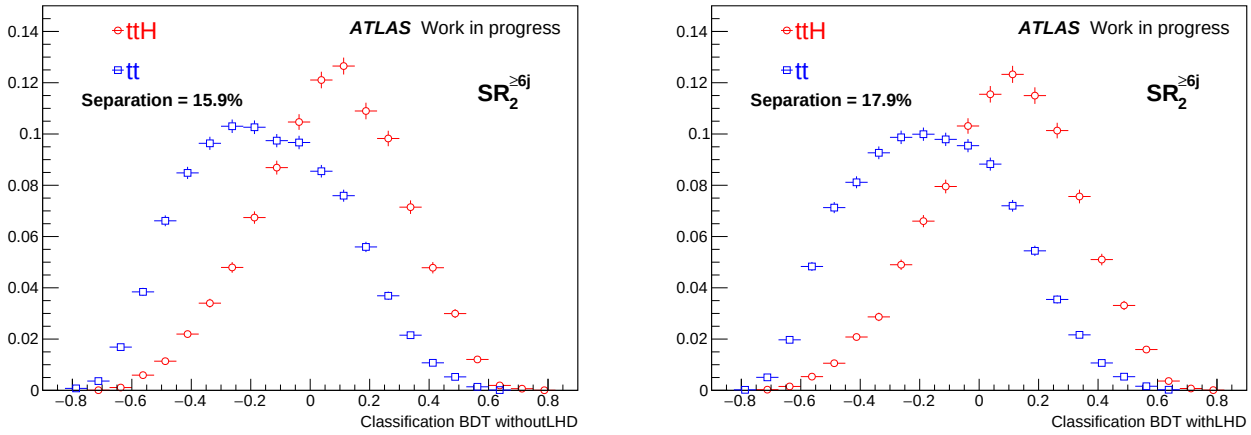
The classification BDT distributions without and with LHD as input variable in the $\mathrm{SR}_1^{\geq 6\mathrm{j}}$ and $\mathrm{SR}_2^{\geq 6\mathrm{j}}$ are shown in figures 75 and 76. The corresponding values of separation power and relative gain due to the addition of the LHD variable is summarised in table 17. The ROC curves of the classification BDT without and with the LHD variable for the same regions are shown in figure 77.

The distributions of the LHD output for data and MC prediction before the fit in signal and background regions are shown in figures 103-100.



(a) (b)

Figure 75: Distributions of the classification BDT (a) without and (b) with the LHD variable for $t\bar{t}H$ signal and $t\bar{t}$ + jets background in the $\mathrm{SR}_1^{\geq 6j}$ region.

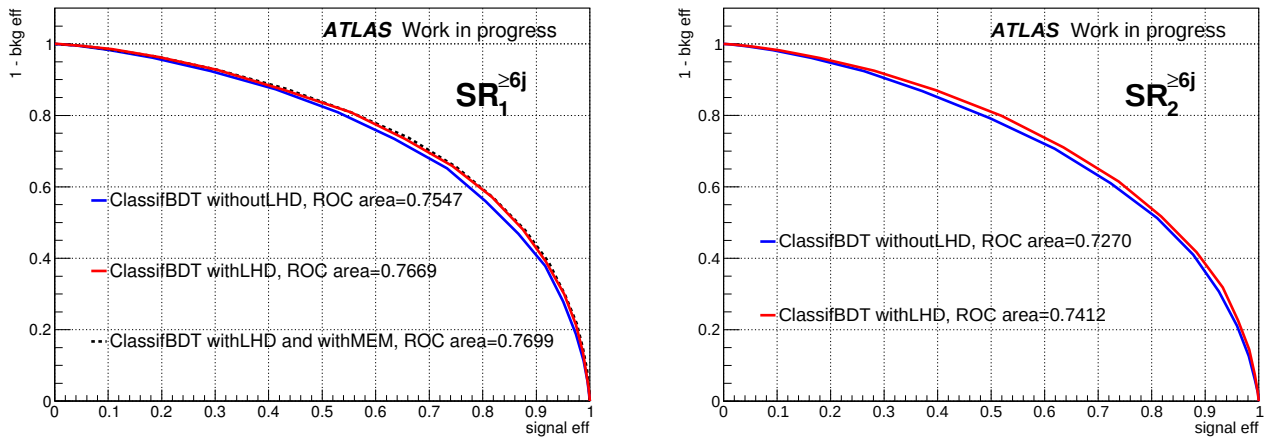


(a) (b)

Figure 76: Distributions of the classification BDT (a) without and (b) with the LHD variable for $t\bar{t}H$ signal and $t\bar{t}$ + jets background in the $\mathrm{SR}_2^{\geq 6j}$ region.

Region	Separation without LHD [%]	Separation with LHD [%]	Relative gain [%]
$SR_1^{\geq 6j}$	19.8 (20.6)	21.8 (22.3)	9.9 (7.8)
$SR_2^{\geq 6j}$	15.9	17.9	12.4

Table 17: Values of the separation power of the classification BDT distributions without and with the LHD variable and relative gain in separation for the $t\bar{t}H$ signal and the $t\bar{t}$ + jets background in the $SR_1^{\geq 6j}$ and $SR_2^{\geq 6j}$. For the $SR_1^{\geq 6j}$ values in brackets correspond to the version of classification BDT with MEM_{D1} as input.



(a) (b)

Figure 77: ROC curves for the classification BDTs without and with the LHD variable in the (a) $SR_1^{\geq 6j}$ region and (b) $SR_2^{\geq 6j}$ regions. The classification BDT with the MEM variable is also shown in $SR_1^{\geq 6j}$ region.

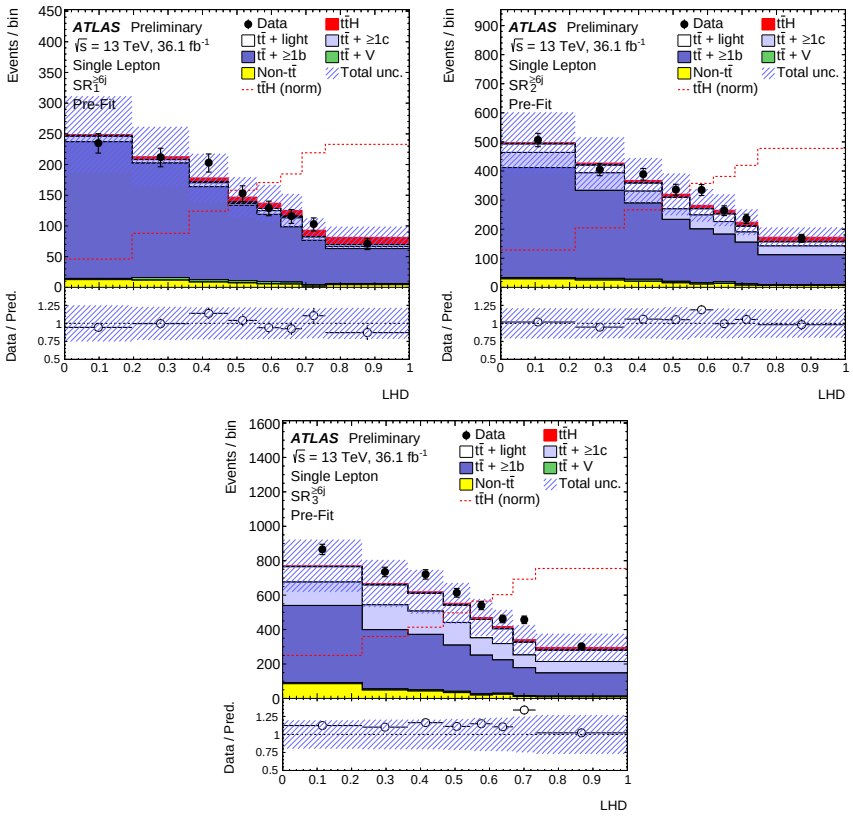


Figure 78: Distributions of the LHD output in the $\mathrm{SR}_1^{\geq 6\mathrm{j}}$, $\mathrm{SR}_2^{\geq 6\mathrm{j}}$ and $\mathrm{SR}_3^{\geq 6\mathrm{j}}$. The signal and background predictions are before the fit to data.

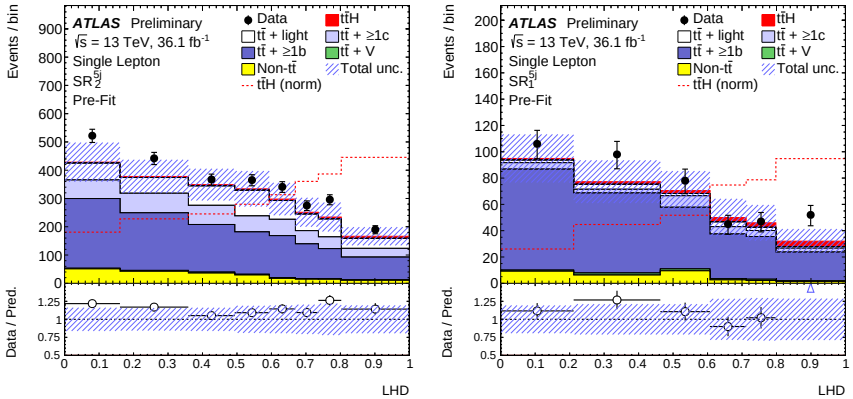


Figure 79: Distributions of the LHD output in the $\mathrm{SR}_1^{5\mathrm{j}}$ and $\mathrm{SR}_2^{5\mathrm{j}}$. The signal and background predictions are before the fit to data.

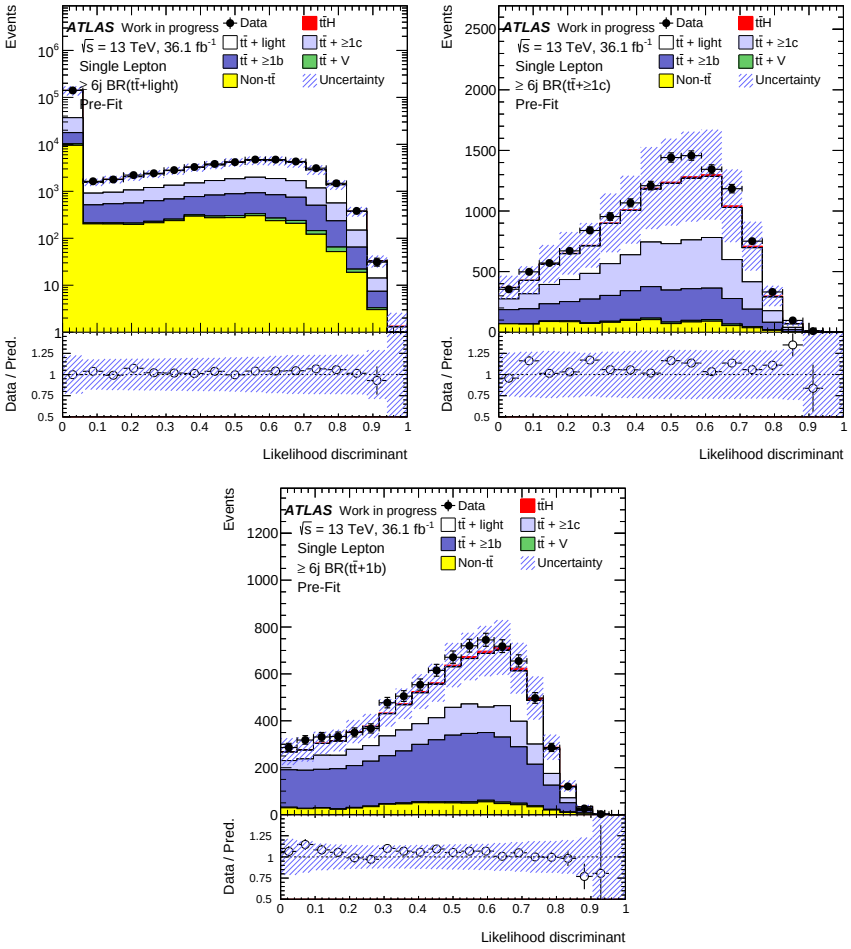


Figure 80: Distributions of the LHD output in the $\mathrm{CR}^{\geq 6\mathrm{j}}_{t\bar{t}+\mathrm{light}}$, $\mathrm{CR}^{\geq 6\mathrm{j}}_{t\bar{t}+\geq 1c}$ and $\mathrm{CR}^{\geq 6\mathrm{j}}_{t\bar{t}+1b}$. The signal and background predictions are before the fit to data.

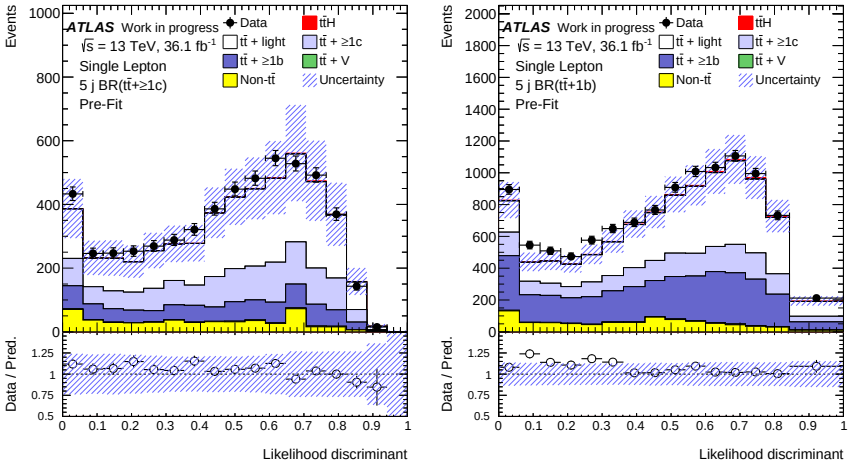


Figure 81: Distributions of the LHD output in the $\mathrm{CR}^{5\mathrm{j}}_{t\bar{t}+\geq 1c}$ and $\mathrm{CR}^{5\mathrm{j}}_{t\bar{t}+1b}$ regions. The signal and background predictions are before the fit to data.

4.8 Systematic uncertainties

Systematic uncertainties affect the normalisation and shape of the signal and background distributions in each of the analysis regions considered, so they are taken into account as nuisance parameters in the fit procedure, as discussed in section 4.9. Systematic uncertainties can be classified into experimental and modelling uncertainties. The main sources of uncertainties in this analysis are those related to the modelling of the $t\bar{t} + \geq 1b$ background. The full list of systematic uncertainties considered is shown in table 18.

4.8.1 Experimental uncertainties

Luminosity

The systematic uncertainty on the 2015+2016 integrated luminosity is 2.1%. The estimation was done with a preliminary calibration of the luminosity scale using the x-y beam-separation scans performed in August 2015 and May 2016. The detailed description of this estimation method can be found in [96]. The systematic uncertainty on the luminosity affects the normalisation of all MC samples.

Leptons

Uncertainties related to leptons are originating from the trigger, reconstruction, identification, isolation, as well as the lepton momentum scale and resolution. The reconstruction, identification and isolation of electrons and muons, as well as the efficiency of the trigger used to record the events, are slightly different between the data and simulation. This is taken into account by so-called scale factors (SF), which are used as weights applied to the MC events. Other uncertainties are related to the difference of the lepton momentum scale and resolution in data and in MC. The corrections that are used to adjust these discrepancies are derived from samples of $Z \to \ell^+\ell^-$, $J/\psi \to \ell^+\ell^-$ and $W \to e\nu$ events. The lepton-related uncertainties have a very small effect for this analysis.

Jets

The uncertainties associated with jets are related to the jet energy scale (JES), jet energy resolution (JER) and the efficiency to pass the JVT selection.

The uncertainty on the jet energy scale is estimated using information from the test-beam data, collision data and simulation as described in Ref. [97]. It consists of 21 components, corresponding to different uncertainty sources: difference in *in-situ* techniques of JES calibration (statistical, modelling, detector and mixed), corrections on pile-up mismodelling, flavour of jets (due to the fact that response of the calorimeter is different to jets originated from quarks or gluons), and high-p_{T} jets measurement. The JES uncertainty is about 5.5% for jets with $p_{\mathrm{T}} = 25$ GeV and decreases for higher jet p_{T}. For central jets with p_{T} in the range of 100 GeV – 1.5 TeV it is below 1.5%. This is one of the main systematic uncertainties related to reconstructed objects.

Systematic uncertainty	Type	Components
Luminosity	N	1
Reconstructed Objects		
Electron trigger+reco+ID+isolation	SN	4
Electron energy scale+resolution	SN	2
Muon trigger+reco+ID+isolation	SN	10
Muon momentum scale+resolution+saggita	SN	5
Taus detector, insitu and model	SN	3
Pileup modelling	SN	1
Jet vertex tagger	SN	1
Jet energy scale	SN	21
Jet energy resolution	SN	2
Missing transverse energy scale+resolution	SN	3
b-tagging efficiency	SN	30
c-mistag rate	SN	15
Light-mistag rate	SN	80
Mistag extrapolation $c \to \tau$	SN	1
Background and Signal Model		
$t\bar{t}$ cross section	N	1
$t\bar{t}+ \geq 1c$: normalisation	N (free floating)	1
$t\bar{t}+ \leq 2b$: normalisation	N (free floating)	1
$t\bar{t}+ \geq 3b$: normalisation	N	1
$t\bar{t}+\geq 1b$: NLO Shape	SN	9
$t\bar{t}+\geq 1c$: NLO Shape	SN	1
$t\bar{t}+\geq 1b$: 4F vs 5F Shape	S	1
$t\bar{t}$ modelling: residual Radiation	SN	3
$t\bar{t}$ modelling: residual NLO generator	SN	3
$t\bar{t}$ modelling: residual parton shower+hadronisation	SN	3
W+jets normalisation	N	3
Z+jets normalisation	N	3
Single top cross section	N	1
Single top model	SN	2
Diboson normalisation	N	1
Fakes normalization	SN	6
$t\bar{t}V$ cross section	N	4
$t\bar{t}V$ modelling	SN	2
tZ cross section	N	2
tWZ cross section	N	1
$t\bar{t}WW$ cross section	N	2
4-tops cross section	N	1
$tHjb$ cross section	N	3
WtH cross section	N	2
$t\bar{t}H$ cross section	N	2
$t\bar{t}H$ branching ratios	N	3
$t\bar{t}H$ modelling	SN	1

Table 18: The list of systematic uncertainties. N - the uncertainty considered to be affecting normalisation only, SN - both normalisation and shape of distributions are affected. Some of the uncertainties are split into several components for a more accurate treatment.

The JER uncertainties was measured in the Run 1 data and simulated dijet events. They were found to agree within 10% [98]. Additional uncertainties were obtained from the extrapolation from Run 1 to Run 2 conditions [97].

Missing transverse energy

The $E_{\mathrm{T}}^{\mathrm{miss}}$ uncertainties are propagated from those related to leptons and jet energy scales and resolutions. Additional uncertainties related to the resolution and scale of the soft term of $E_{\mathrm{T}}^{\mathrm{miss}}$ are considered. These uncertainties have a very small impact on the analysis.

Flavour tagging

The efficiencies of the b- and c-jet identification and light jet mis-tag rates obtained from simulation are corrected by applying SFs to match the efficiencies measured in data (see section 3.2.7). These SF depend on jet p_{T} in the case of b- and c-jets and on jet p_{T} and η for light-jets. The efficiencies are derived for the four b-tagging working points and then combined, and the corresponding uncertainty components are taken into account. The uncertainties corresponding to these measurements are factorised into independent sources (those corresponding to different working points, and several bins in p_{T} and η): 30 for b-jets, 15 for c-jets, and 80 for light jets.

4.8.2 Modelling uncertainties

Signal modelling

An uncertainty of $+10\%/-13\%$ on the $t\bar{t}H$ signal cross-section is applied. This includes the contributions from scale and PDF uncertainties, considered to be uncorrelated [99]-[100]. Uncertainties on the Higgs boson branching ratios are considered; for $H \rightarrow b\bar{b}$ it is 2.2% [101]. An uncertainty on the choice of parton shower and hadronisation model is obtained from the difference between MG5_aMC@NLO interfaced to either Pythia 8 (nominal model) or Herwig++.

tt+jets modelling

The modelling of $t\bar{t}$ + jets events is the main source of systematic uncertainties in this analysis. The full list of corresponding uncertainties is presented in table 19.

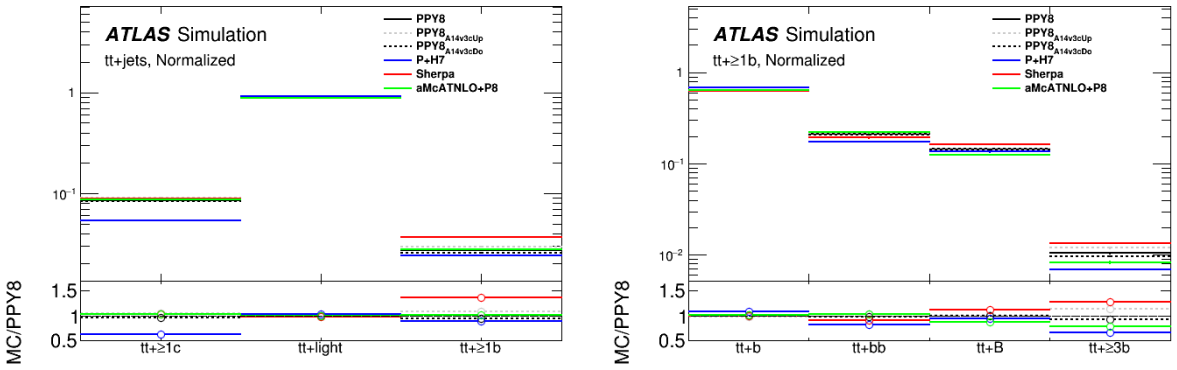
For the inclusive $t\bar{t}$ production cross-section at NNLO+NNLL an uncertainty of $\pm 6\%$ is applied according to [77]. There is no prior uncertainty on the normalisation of $t\bar{t}$ + $\geq 1b$ and $t\bar{t}$ + $\geq 1c$, those parameters are let to float freely in the fit. The uncertainty associated with the choice of NLO generator is estimated comparing the nominal Powheg-Box+Pythia 8 sample with a sample generated with Sherpa (5F). The uncertainty associated with the choice of parton shower and hadronisation models is evaluated by comparing the prediction from the Powheg-Box generator interfaced to either to Pythia 8 or Herwig 7. An uncertainty on modelling of initial and final state radiation (ISR and FSR) is obtained with two Powheg-Box+Pythia 8 samples with different

Systematic source	How evaluated	$t\bar{t}$ categories
$t\bar{t}$ cross-section	Up or down by 6%	All, corr.
NLO generator	Powheg-Box+Pythia 8 vs. Sherpa 5F	All, uncorr.
ISR / FSR	Variations of μ_{R}, μ_{F}, h_{damp} and A14 parameters	All, uncorr.
PS & hadronisation	Powheg-Box+Pythia 8 vs. Powheg-Box+Herwig 7	All, uncorr.
$t\bar{t} + \geq 1b$ renorm. scale	Up or down a by factor of two	$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 1b$ resumm. scale	Vary μ_{Q} from $H_{\mathrm{T}}/2$ to μ_{CMMPS}	$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 1b$ global scales	Set μ_{Q}, μ_{R}, and μ_{F} to μ_{CMMPS}	$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 1b$ shower recoil	Alternative model scheme	$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 1b$ PDF	CT10 vs. MSTW or NNPDF	$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 1b$ FSR	ISR / FSR variation samples vs. nominal	$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 3b$ normalisation	Up or down by 50%	$t\bar{t} + \geq 3b$
$t\bar{t} + \geq 1b$ 4F vs 5F	Powheg-Box+Pythia 8 vs. SherpaOL	$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 1b$ MPI	Up or down by 50%	$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 1c$ ME calculation	MG5_aMC@NLO +Herwig++ inclusive vs. ME prediction	$t\bar{t} + \geq 1c$

Table 19: The systematic uncertainties on the $t\bar{t}$ + jets modelling. For the $t\bar{t} + \geq 1b$ background, the inclusive $t\bar{t}$ sample is reweighted to the NLO $t\bar{t} + \geq 1b$ prediction.

values of h_{damp} and A14 eigentune parameters. All these uncertainties, except that on the inclusive $t\bar{t}$ cross-section, are considered to be uncorrelated among $t\bar{t} + \geq 1b$, $t\bar{t} + \geq 1c$ and $t\bar{t}$ + light. For the $t\bar{t} + \geq 1b$ process, an additional uncertainty on the choice in two alternative schemes of $\geq 1b$ production, known as four-flavour (4F) and five-flavour (5F) schemes, is considered. It is obtained from the difference between the prediction of the nominal Powheg-Box+Pythia 8 sample (5F) and the SherpaOL (4F). For each of the above alternative samples the fractions of $t\bar{t} + \geq 1b$ subcategories are reweighted to match the prediction of SherpaOL in the same way as the nominal sample, as described in section 4.3.2. Additionally, uncertainties on the SherpaOL prediction for $t\bar{t} + \geq 1b$ at NLO are estimated by applying variations to the renormalisation, factorisation and resummation scales in SherpaOL. To take into account uncertainty on the choice of PDF, two different sets are considered: NNPDF (nominal) and MSTW [102] (alternative). Another uncertainty is obtained with an alternative shower recoil scheme. Additionally, a 50% uncertainty is associated to the events not included in the original NLO calculation but coming from multiple parton interactions (MPI). To take into account significant difference in the $t\bar{t} + \geq 3b$ component between Powheg-Box+Pythia 8 and SherpaOL, an additional 50% uncertainty on the normalisation of $t\bar{t} + \geq 3b$ events is considered.

Another uncertainty is applied to take into account the difference between $t\bar{t} + \geq 1c$ calculated in the matrix element with the default approach of using charm jets produced in the parton shower. This uncertainty is derived by comparing the nominal $t\bar{t}$ + jets sample with $t\bar{t} + c\bar{c}$ NLO matrix element calculation with MadGraph5_aMC@NLO interfaced to Herwig++.



(a) (b)

Figure 82: Relative fractions of different $t\bar{t}$ + jets components (a) and $t\bar{t}+ \geq 1b$ (b) sub-components for different generators: nominal POWHEG-BOX+PYTHIA 8 (black line) and systematic samples: POWHEG-BOX+PYTHIA 8 sample with more (gray dotted line) and less (black dotted line) radiation, POWHEG-BOX+HERWIG 7 (blue), SHERPA 5 FS (red) and MG5_aMC@NLO + PYTHIA 8 (green). The distributions are obtained using particle jets with $p_{\mathrm{T}} > 15$ GeV.

The fractions of various $t\bar{t}$ + jets background components (a) and $t\bar{t}+ \geq 1b$ (b) sub-components for different generators are presented in figure 82.

Other backgrounds modelling

An uncertainty of 40% is considered for the W + jets cross-section, and an additional 30% uncertainty for the W + heavy flavour jets events. These uncertainties are estimated by varying the SHERPA scales and matching parameters. A 35% uncertainty is applied for Z + jets normalisation. It also takes into account variations of the SHERPA parameters, as well as the uncertainty on the correction factor of ~ 1.3 for the heavy flavour component, that is derived from data.

For the cross-section of single top production a theoretical uncertainty of $^{+5\%}_{-4\%}$ is considered [85, 87]. As for the $t\bar{t}$ background, an uncertainty associated with initial and final-state radiation is used. An additional uncertainty takes into account the interference between the $t\bar{t}$ and Wt processes at NLO [84]. It is derived by comparing the default diagram removal scheme with so-called diagram subtraction scheme.

For the diboson background, a 50% normalisation uncertainty on cross-section and additional jet production is considered [103]. An uncertainty on the $t\bar{t}V$ NLO cross-section of 15% is used [104]. An additional uncertainty on $t\bar{t}V$ associated with the choice of generator, parton shower and hadronisation model is considered. It is obtained from the comparison of the nominal sample with alternative generated with SHERPA. For the $t\bar{t}t\bar{t}$ background a normalisation uncertainty of 50% is considered.

Misidentified-lepton background

For the data-driven non-prompt lepton background estimation an uncertainty of 50% is used. It is considered uncorrelated across the 5 jet and ≥ 6 jet regions as well as between the electron and muon channels.

4.9 Statistical analysis

The ratio of the measured signal to the Standard Model prediction, or signal strength, $\mu = \sigma/\sigma_{SM}$ is obtained with a fitting procedure based on the RooStats framework [105]. The statistical method used in this analysis is based on a binned maximum likelihood function $\mathcal{L}(\mu, \theta)$, where θ is the set of nuisance parameters (NP), corresponding to the considered systematic uncertainties. This function is a 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:

$$\mathcal{L}(\mu, \theta) = \prod_{j} \prod_{i=bin} \frac{(\mu s_i(j) + b_i(j))^{N_i^{(j)}}}{N_i^{(j)}!} e^{-\mu s_i(j) - b_i(j)} \prod_{\theta} func(\theta|0,1), \tag{76}$$

where $func$ is given by a Gaussian or log-normal pdfs, the value $\theta = 0$ corresponds to the nominal value of the prediction, $\theta = \pm 1$ correspond to ± 1 standard deviation of given systematic uncertainty. $N_i^{(j)}$ is the number of observed events in the i-th bin of the j-th signal region, $s_i(j)$ and $b_i(j)$ are the expected numbers of signal and background events, that are expressed as a function of θ.

The test statistics is defined as a profile likelihood ratio

$$q_\mu = -2 \ln(\mathcal{L}(\mu, \hat{\hat{\theta}}_\mu)/\mathcal{L}(\hat{\mu}, \hat{\theta})), \tag{77}$$

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

The test statistics is used to determine the compatibility of the data measurement with the background-only hypothesis (μ=0) and predict the upper limit on μ using the confidence level (CL_S) method [106, 107].

A test of a hypothesized value of μ is a measure of discrepancy between the data and hypothesis, with higher values of q_μ corresponding to increasing disagreement. The disagreement is quantified using the p-value

$$p_\mu = \int_{q_\mu,\mathrm{obs}}^{\infty} f(q_\mu|\mu) dq_\mu, \tag{78}$$

where $q_{\mu,\mathrm{obs}}$ is the statistic value observed in data, $f(q_\mu|\mu)$ is the pdf of q_μ assuming a signal strength of μ.

The compatibility of the result with the signal-plus-background hypothesis is then given by

$$p_{s+b} = f(q \geq q_{\rm obs}|1) = \int_{q_\mu,{\rm obs}}^{\infty} f(q_\mu|1)dq_\mu, \tag{79}$$

while that with the background-only hypothesis is quantified by

$$p_b = f(q \geq q_{\rm obs}|0) = \int_{-\infty}^{q_\mu,{\rm obs}} f(q_\mu|0)dq_\mu, \tag{80}$$

Example distributions of the test statistics under the signal-plus-background and background-only hypotheses and corresponding p-values are presented in figure 83.

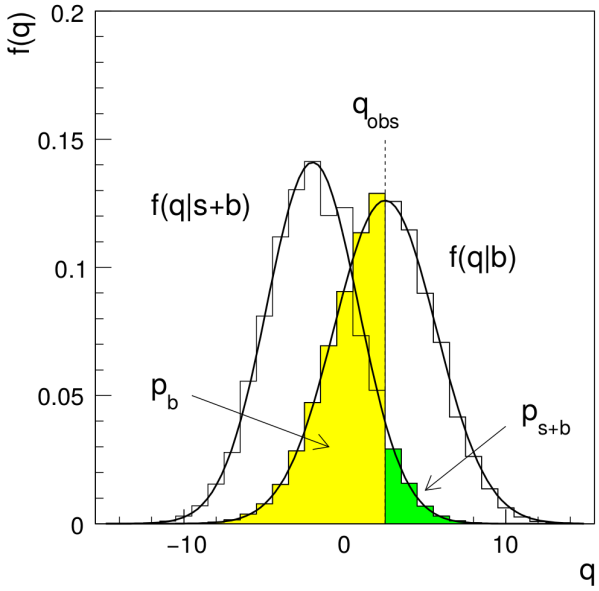


Figure 83: The distribution of the statistics $q_\mu = -2\ln(\mathcal{L}_{s+b}/\mathcal{L}_b)$ under the signal-plus-background ($\mu = 1$) and background-only ($\mu = 0$) hypotheses. The p-values for both hypotheses are also shown with respect to the observed value of the statistic $q_{\rm obs}$ [106].

The confidence level for the signal hypothesis is then defined as

$$CL_s = \frac{CL_{s+b}}{CL_b} = \frac{p_{s+b}}{1-p_b}. \tag{81}$$

Values of μ for which CLs <0.05 are excluded at the 95% CL.

4.10 Results

The final result is obtained by combining the single-lepton and the dilepton channels. A simultaneous binned maximum-likelihood fit is performed in all analysis regions. In all signal regions the classification BDT output is used in the fit to obtain the maximum sensitivity. In the single-lepton background regions $\mathrm{CR}^{\geq 6\mathrm{j}}_{t\bar{t}+\geq 1c}$ and $\mathrm{CR}^{5\mathrm{j}}_{t\bar{t}+\geq 1c}$ the scalar sum of the p_{T} of the jets, $H_{\mathrm{T}}^{\mathrm{had}}$, is used as the discriminating variable. In the remaining single-lepton and all dilepton background regions a single bin is used. The $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ normalisation factors are kept as free-floating parameters in the fit.

The results of the fit to data in the single-lepton and dilepton channels as well as their combination are presented in this section. For the two channels the signal strength values are obtained by performing a single combined fit with all regions included, keeping all the NPs and normalisation factors correlated across channels.

A good agreement of the data with the prediction in the single-lepton regions before and after the combined fit is observed, as presented in figure 84. The post-fit uncertainties are reduced due to constraints and NP correlations, which are measured by the fit.

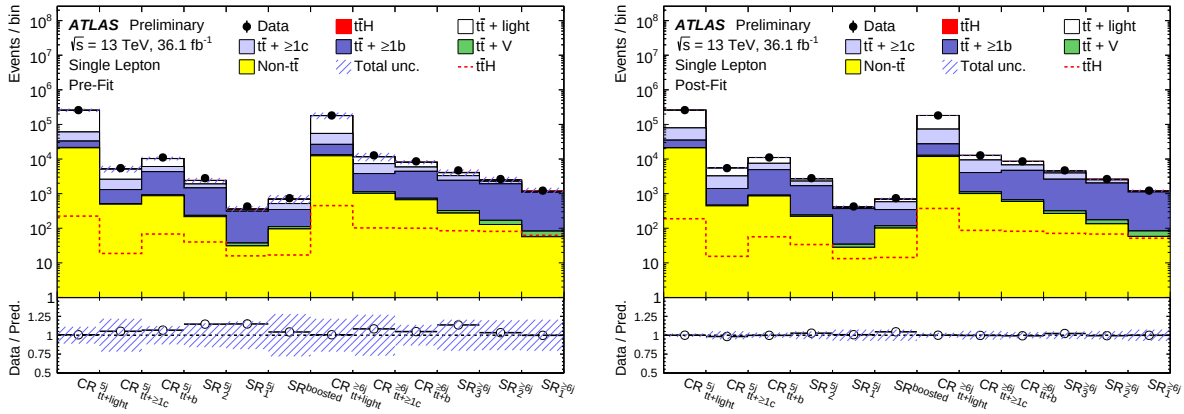


Figure 84: Single-lepton yield summary before and after the combined fit to data. From Ref. [91].

The observed signal strength values with their statistical and systematic components are given in figure 85. The statistical uncertainty is estimated performing a fit with fixing all the NPs, except the free-floating $k(t\bar{t}+\geq 1b)$, $k(t\bar{t}+\geq 1c)$ and μ, to their post-fit values. The systematic uncertainty is then calculated via subtraction in quadrature of the statistical uncertainty from the total uncertainty. The contribution of the systematic uncertainty is larger than the statistical for both single channel signal strengths as well as the combined signal strength.

The combined best-fit signal strength value is

$$\mu = 0.84 \pm 0.29\ (\mathrm{stat.})\ ^{+0.57}_{-0.54}\ (\mathrm{syst.}) = 0.84^{+0.64}_{-0.61}.$$

When a different μ is fitted separately in each channel in the combined fit, the best-fit values are $0.95^{+0.65}_{-0.62}$ in the single-lepton channel and $-0.24^{+1.02}_{-1.05}$ in the dilepton channel.

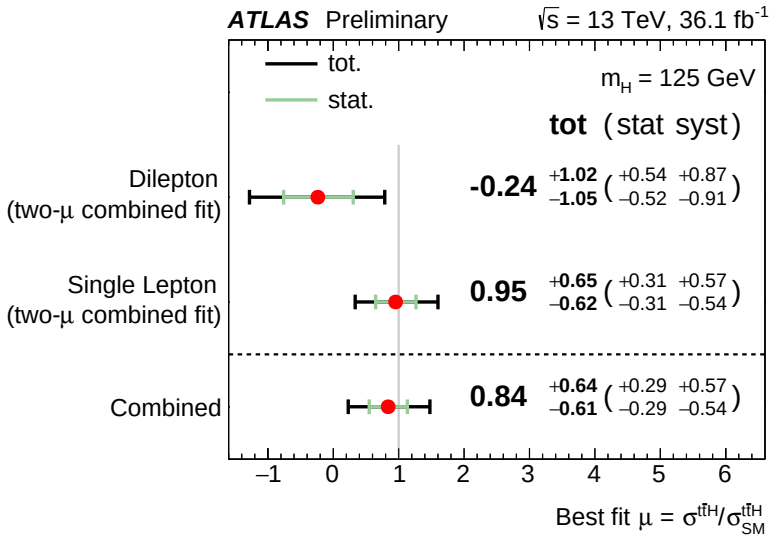


Figure 85: The signal strength for the single-lepton and dilepton channels and their combination. From Ref. [91].

The observed significance of μ measurement with respect to the background-only hypothesis is 1.4 standard deviations (σ), for an expected significance of 1.6 σ. A signal strength larger than 2.0 can be excluded at the 95% confidence level, as shown in figure 86.

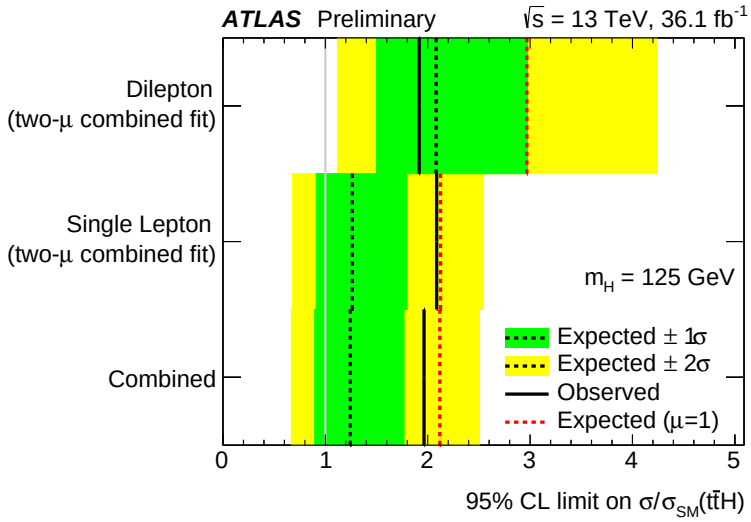


Figure 86: Upper limits on signal strength at 95% confidence level in dilepton channel, single-lepton channel and their combination. From Ref. [91].

The best-fit values of the $t\bar{t} + \geq 1b$ and $t\bar{t} + \geq 1c$ normalisation factors are $k(t\bar{t}+ \geq 1b) = 1.24 \pm 0.10$ and $k(t\bar{t}+ \geq 1c) = 1.63 \pm 0.23$, respectively.

The contributions of the different systematic uncertainties on the signal strength obtained in the combined fit are presented in table 20. The dominant uncertainty in this analysis originates from the modelling of the $t\bar{t} + \geq 1b$ background. The next significant sources of uncertainties, in order of their impact, are the background simulation statistics, the uncertainty on b-tagging efficiency, the jet energy scale and resolution and the signal modelling.

Uncertainty source	$\Delta\mu$	
$t\bar{t} + \geq 1b$ modelling	$+0.46$	-0.46
Background model statistics	$+0.29$	-0.31
Jet flavour tagging	$+0.16$	-0.16
Jet energy scale and resolution	$+0.14$	-0.14
$t\bar{t}H$ modelling	$+0.22$	-0.05
$t\bar{t} + \geq 1c$ modelling	$+0.09$	-0.11
Jet-vertex association, pileup modelling	$+0.03$	-0.05
Other background modelling	$+0.08$	-0.08
$t\bar{t}$+light modelling	$+0.06$	-0.03
Luminosity	$+0.03$	-0.02
Light lepton (e, μ) ID, isolation, trigger	$+0.03$	-0.04
Total systematic uncertainty	$+0.57$	-0.54
$t\bar{t} + \geq 1b$ normalisation	$+0.09$	-0.10
$t\bar{t} + \geq 1c$ normalisation	$+0.02$	-0.03
Statistical uncertainty	$+0.29$	-0.29
Total uncertainty	$+0.64$	-0.61

Table 20: Summary of the effects of the uncertainties on the measured signal strength μ. The background model statistics refers to the statistical uncertainties from the limited number of simulated events and from the data-driven determination of the non-prompt and fake lepton background component in the single-lepton channel. The normalisation factors $k(t\bar{t}+ \geq 1b)$ and $k(t\bar{t}+ \geq 1c)$ are included in the statistical component. From Ref. [91].

The 20 NPs with highest impact on the uncertainty on the signal strength are shown in figure 87 for the combined fit. The four most important NPs are related to the theoretical uncertainties on $t\bar{t} + \geq 1b$ modelling. The fit to data decreases these uncertainties. This is reflected by the post-fit impact which is up to three times smaller than the pre-fit impact

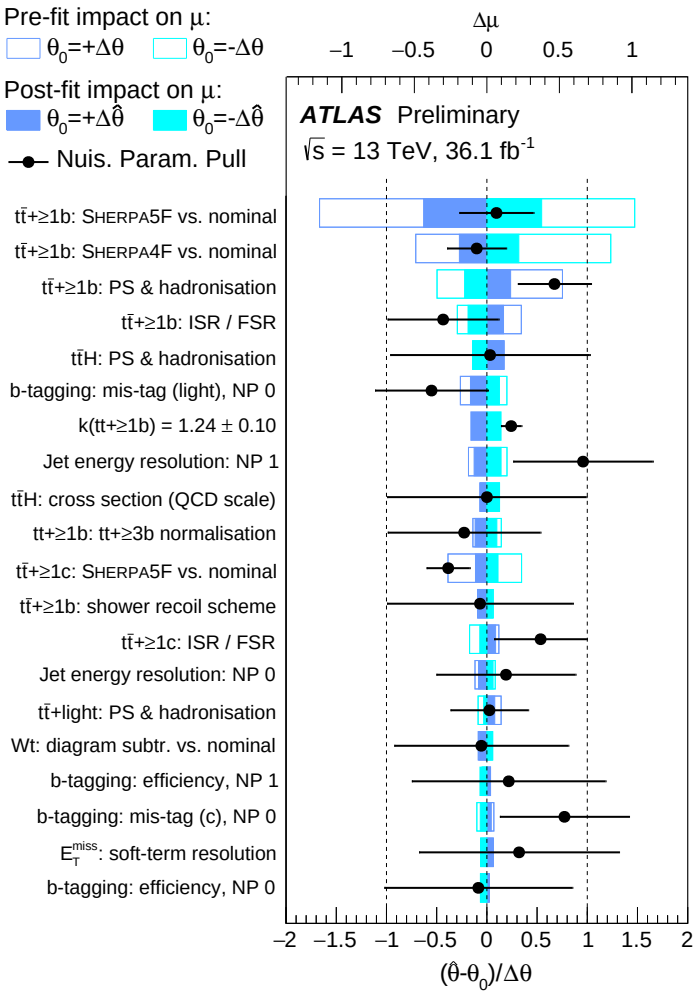


Figure 87: Ranking of the NPs with respect to their impact on the measured signal strength μ. The empty blue rectangles indicate the pre-fit impact on μ, the filled blue ones - post-fit impact on μ. The black points show the fitted values and uncertainties of the NPs. From Ref. [91].

on the uncertainty on the signal strength. The $t\bar{t} + \geq 1b$ normalisation uncertainty also has a high impact.

Further details on the observed and expected pulls and constraints for the NPs can be found in appendices A and B, respectively.

The distributions of the discriminating variables used in the fit in the single-lepton control and signal regions before and after the combined fit to the data are shown in figures 88-90. Because of the adjustment of the NPs the agreement between data and prediction is improved in all distributions after the fit. The corresponding comparison for the LHD discriminant in all considered regions can be found in appendix C.

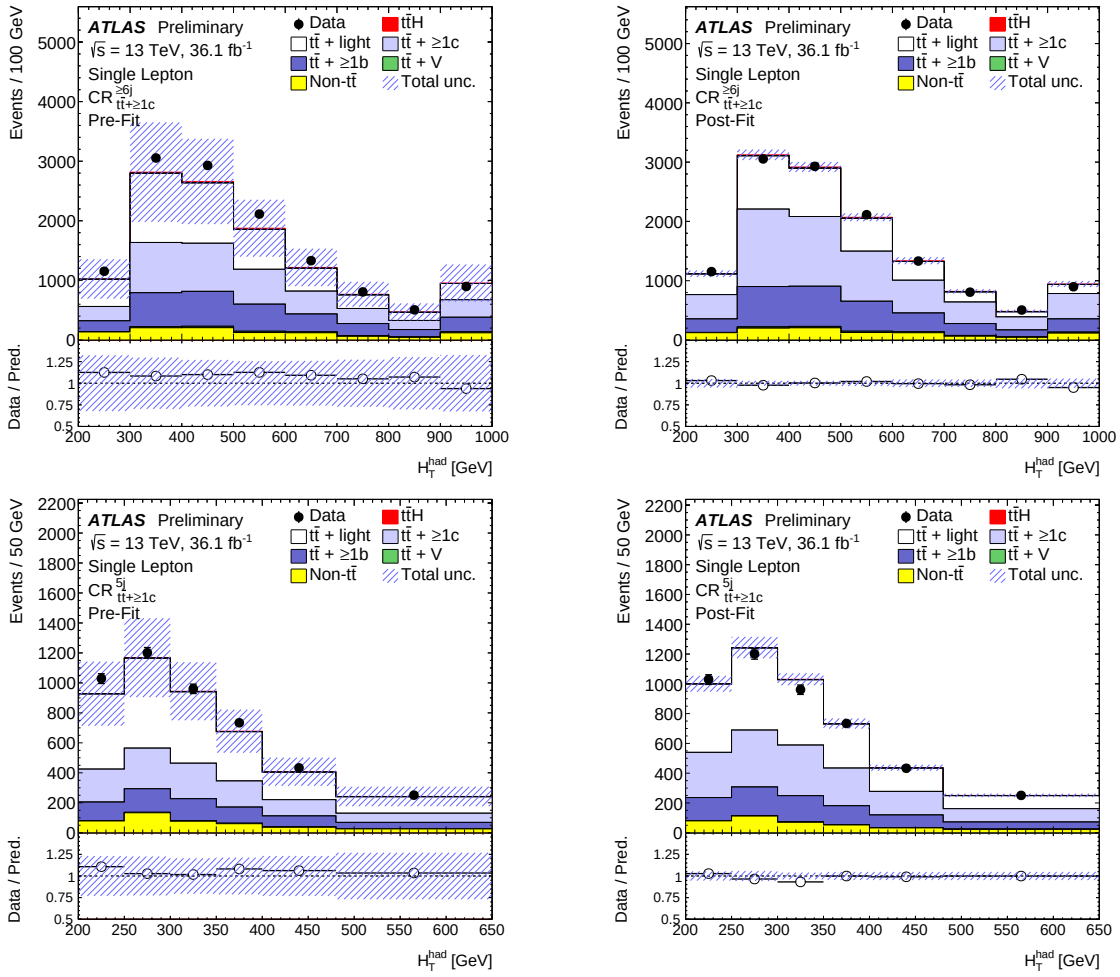


Figure 88: $H_{\mathrm{T}}^{\mathrm{had}}$ distribution in the single-lepton $\mathrm{CR}^{\geq 6\mathrm{j}}_{t\bar{t}+\geq 1c}$ and $\mathrm{CR}^{5\mathrm{j}}_{t\bar{t}+\geq 1c}$ regions (left) before and (right) after the fit to data. The pre-fit plots do not include an uncertainty on the $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$ normalisation. From Ref. [91].

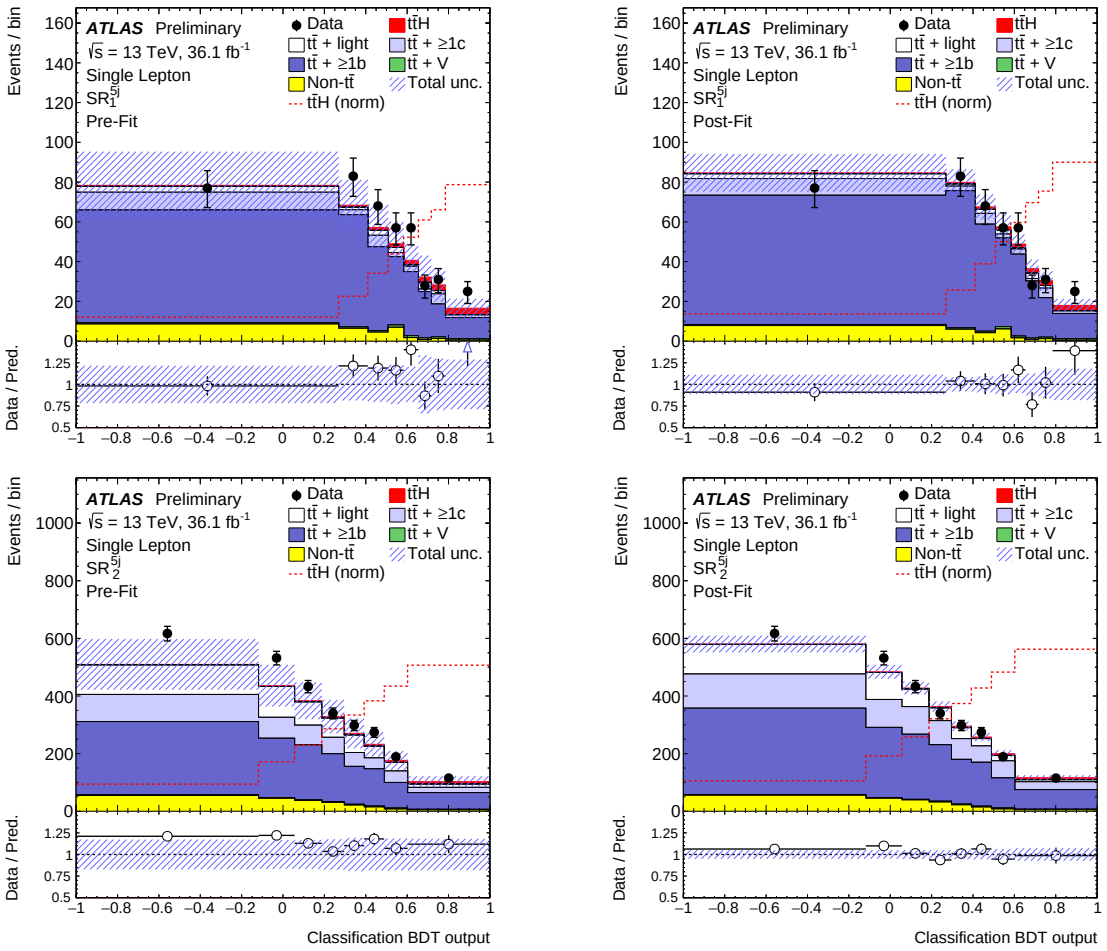


Figure 89: Classification BDT distribution in the single-lepton $\mathrm{SR}_1^{5\mathrm{j}}$ and $\mathrm{SR}_2^{5\mathrm{j}}$ regions (left) before and (right) after the fit to data. The $t\bar{t}H$ signal yield (solid) is normalised to the SM cross-section before the fit and to the fitted μ after the fit. The dashed line shows the $t\bar{t}H$ signal distribution normalised to the total yield. The pre-fit plots do not include an uncertainty on the $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$ normalisation. From Ref. [91].

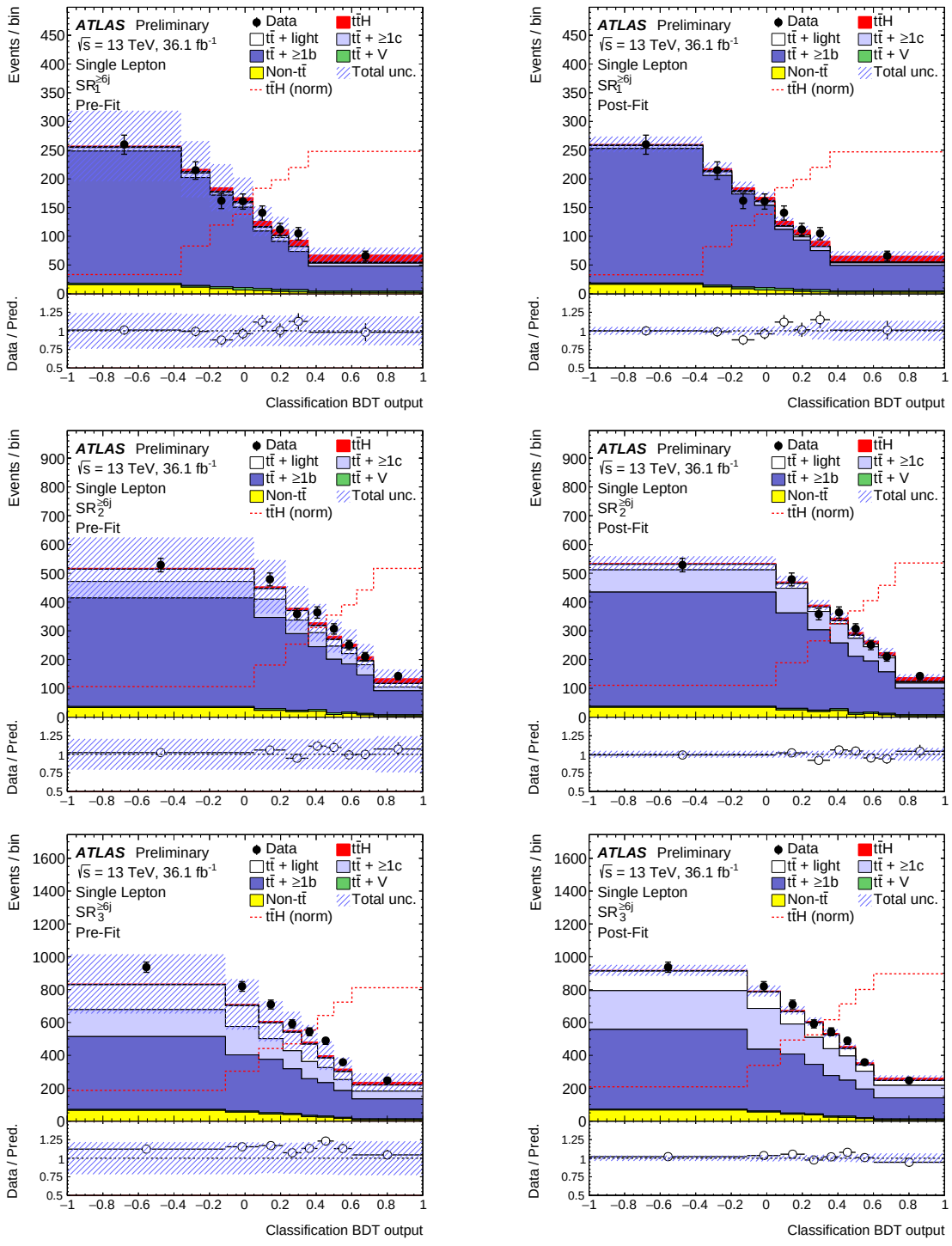


Figure 90: Classification BDT distribution in the single-lepton $\mathrm{SR}_1^{\geq 6\mathrm{j}}$, $\mathrm{SR}_2^{\geq 6\mathrm{j}}$ and $\mathrm{SR}_3^{\geq 6\mathrm{j}}$ regions (left) before and (right) after the fit to data. The $t\bar{t}H$ signal yield (solid) is normalised to the SM cross-section before the fit and to the fitted μ after the fit. The dashed line shows the $t\bar{t}H$ signal distribution normalised to the total yield. The pre-fit plots do not include an uncertainty on the $t\bar{t}+\geq 1b$ or $t\bar{t}+\geq 1c$ normalisation. From Ref. [91].

4.11 Combination of ATLAS $t\bar{t}H$ searches

The results from the $t\bar{t}H(H \to b\bar{b})$ search was combined with those obtained in other channels:

- $t\bar{t}H$ ML [108], a search targeting the final states with three or more leptons, or two same-sign charged light leptons,
- $t\bar{t}H$ $(H \to \gamma\gamma)$ [109], and
- $t\bar{t}H$ $(H \to ZZ^* \to 4\ell)$ [110].

All these analyses use the same 36.1 fb^{-1} of pp collision data registered with the ATLAS detector in 2015 and 2016. The $t\bar{t}H$ production is modelled using the same MC generators, assuming the Higgs boson mass of 125 GeV.

The best-fit $t\bar{t}H$ signal strength value obtained in the combination of four searches is

$$\mu = 1.17 \pm 0.19 \text{ (stat.)} \, ^{+0.28}_{-0.25} \text{ (syst.)}.$$

The signal strength values obtained in each analysis, and the result of the combination is shown in figure 91.

The observed significance with respect to the background-only hypothesis is 4.2σ, while expected significance is 3.8σ. This represents evidence for $t\bar{t}H$ production.

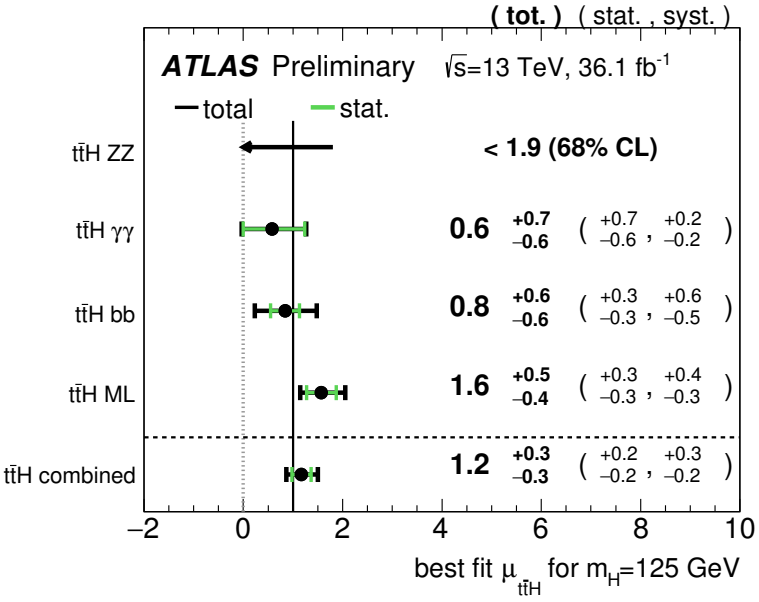


Figure 91: The best-fit signal strength values for individual analyses and their combination. From Ref. [108].

Conclusions

A search for the associated production of the Higgs boson with a pair of top quarks, $t\bar{t}H$ ($H \to b\bar{b}$), with a single lepton in the final state is presented in this thesis. The search was performed using 36.1 fb^{-1} of pp collision data at a centre-of-mass energy of 13 TeV recorded with the ATLAS detector at the LHC in 2015 and 2016. Measuring the $t\bar{t}H$ cross-section is very important, since it gives a direct access to the measurement of the top quark Yukawa coupling. A significant deviation of this parameter from the SM prediction would indicate physics beyond the SM.

The major difficulty of this measurement arises from the fact that the main background, $t\bar{t}$ with additional b-jets, has the same signature as the signal. To separate the signal from the background several analysis techniques are employed. The main contribution by the author to this analysis, presented in this dissertation, is the development of the likelihood discriminant (LHD) method, that exploits specific kinematic properties of $t\bar{t}H$ ($H \to b\bar{b}$) and $t\bar{t}$+jets events to distinguish them. The method is used in combination with other discriminating variables via multivariate techniques, in order to achieve improved discrimination between the signal and the background. The LHD was for the first time applied in this analysis and was found to be the most discriminating single variable, providing an improvement in the separation power of $\sim 10\%$.

The ratio of the measured $t\bar{t}H$ cross-section to the SM expectation obtained in a combination with the dilepton channel is $\mu = 0.84^{+0.64}_{-0.61}$, assuming a Higgs boson mass of 125 GeV. This result is consistent with both the background-only hypothesis and the $t\bar{t}H$ SM prediction. A value of μ higher than 2.0 is excluded at the 95% confidence level. The combination of ATLAS searches for $t\bar{t}H$ production yields a measured signal strength of $\mu = 1.17^{+0.33}_{-0.30}$. The result has an observed significance of 4.2σ, which represents evidence for $t\bar{t}H$ production.

The identification of the jets originating from b-quark fragmentation, or b-tagging, plays a key role in this search. Work on the optimisation of the b-tagging impact-parameter-based algorithms (IP2D, IP3D) for LHC Run 2 is also presented in this dissertation. The major contribution made by the author is the development of a new classification of tracks that considers several new tracking variables, in particular, taking advantage of the installation of the IBL - a new pixel detector layer. The expected improvement in light jet rejection at 70% b-jet efficiency of the IP3D algorithm due to the new track categorisation is $\sim 15\%$, while the overall improvement of the optimisations described is $\sim 27\%$.

Résumé

Dans les temps anciens, les gens cherchaient des réponses à des questions fondamentales telles que: "De quoi le monde qui nous entoure est-il fait?", "Qu'est-ce que la matière?".

La physique des particules moderne a apporté quelques réponses à certaines de ces questions, mais en a aussi ajouté de nouvelles à la liste: "Comment les particules élémentaires interagissent?", "Qu'est-ce qui est commun entre les différentes interactions physiques?", "Pourquoi y a-t-il plus de matière que d'antimatière dans l'univers?" ou "Quelle est l'origine de la masse?".

Une théorie qui fournit une image des particules élémentaires cohérente, mais pas complète, est le modèle standard (MS). Il donne une description unifiée de trois des quatre forces fondamentales, à l'exception de la gravité. De nombreuses prédictions théoriques du MS ont été vérifiées expérimentalement avec une remarquable précision depuis les années 1960, lorsque le modèle a été établi.

L'un des problèmes fondamentaux soulevés et résolus dans le MS est l'origine de la masse des particules élémentaires. A priori, on s'attend à ce que les particules élémentaires décrites par la théorie soient sans masse, en contradiction avec l'observation. Par conséquent, un mécanisme qui permet aux particules d'acquérir leur masse a été introduit pour fournir un accord avec les preuves expérimentales. Ce mécanisme suppose l'existence d'un champ scalaire dont les excitations se manifestent comme une nouvelle particule physique appelée le boson de Higgs. Le MS prédit certaines propriétés du boson de Higgs, mais sa masse est un paramètre libre de la théorie et ne peut être obtenue qu'à partir de l'expérience. La recherche de cette particule a été l'un des principaux objectifs du Large Hadron Collider (LHC), le plus grand accélérateur de particules au monde construit au CERN. La découverte du boson de Higgs en 2012 par les collaborations ATLAS et CMS fut un triomphe du MS: la dernière particule prédite par cette théorie avait finalement été trouvée.

L'un des modes possibles pour la production du boson de Higgs au LHC est la production en association avec des paires de quarks top ($t\bar{t}H$). Ce canal de production possède l'une des plus petites sections efficaces de production. En même temps, il présente un intérêt physique particulier: le couplage du boson de Higgs aux quarks top, qui peut être directement mesuré dans ce canal, est une propriété importante du MS. Si la valeur mesurée de ce paramètre est significativement différente de l'unité prédite par le MS, ce serait une indication pour une nouvelle physique au-delà du MS. Par conséquent, l'observation de la production du boson de Higgs en association avec les quarks top est maintenant l'un des objectifs les plus importants du LHC.

Une recherche de la production du boson de Higgs en association avec une paire de quarks top, $t\bar{t}H$ ($H \to b\bar{b}$) dans le canal à un lepton est présentée dans cette thèse. La recherche a été effectuée en utilisant 36.1 fb^{-1} de données de collisions pp à une énergie de 13 TeV dans le centre de masse, enregistrées avec le détecteur ATLAS au LHC en 2015 et 2016.

Cette recherche repose sur la grande multiplicité des jets issus de quarks bottom (b) dans l'état final, pour cette raison il est crucial d'identifier ces jets.

Identification de *b*-jets

L'identification des jets issus de quarks b (b-tagging) est importante car de nombreuses analyses avec de tels quarks dans l'état final sont effectuées par l'expérience ATLAS, comme les mesures dans le secteur du MS (la physique du quark top et du boson de Higgs) et des recherches au-delà du MS. Les processus physiques avec des quarks b dans l'état final sont d'un intérêt particulier puisque ce sont les quarks les plus lourds du SM formant des hadrons.

Quand un quark b est produit, il s'hadronise et forme un hadron b ($B^{\pm}$, B^0, etc.), qui se désintègre par la suite. Un jet formé par les particules produites dans la fragmentation d'un quark b et la désintégration suivante des hadrons b est appelé *jet b*.

Les propriétés importantes de hadrons b sont leur durée de vie relativement longue ($\sim$ 1.6 ps) et leur grande masse ($\sim$ 5 GeV). Un hadron b peut voler sur plusieurs mm à travers le détecteur avant de se désintégrer. De ce fait, le vertex de la désintégration du hadron b, appelé *vertex secondaire* (SV), est significativement déplacé par rapport au *vertex primaire* (PV). En même temps, la grande masse du hadron b fournit une différence angulaire entre la direction de la propagation initiale de hadron b et ses produits de désintégration. Toutes ces paramètres permettent de distinguer les jets b des autres jets.

Il y a plusieurs types d'algorithmes de b-tagging dans ATLAS. La performance de b-tagging pour Run 2 a été améliorée grâce à l'insertion de l'IBL et des développements des algorithmes de reconstruction des traces et de b-tagging.

Les algorithmes basés sur les paramètres d'impact (IP2D, IP3D) utilisent le fait que les traces de la désintégration de hadron b ne pointent pas sur le vertex primaire. Ils utilisent les paramètres d'impact transversal et longitudinal. Le paramètre d'impact transversal d_0 est la distance de l'approche de la trace au PV dans le plan transverse. Le paramètre d'impact longitudinal est défini comme $z_0 \sin\theta$, où z_0 est la coordonnée longitudinale de la trace au point d'approche au PV.

L'algorithme de recherche de vertex secondaire (SV) reconstruit le SV inclusif formé par les produits de désintégration du hadron b, y compris ceux de la désintégration hadron c subséquente. Premièrement, il recherche toutes les paires de deux traces qui forment un vertex, en utilisant des traces déplacées du PV. Ensuite, l'algorithme supprime les traces compatibles avec les désintégrations de particules à vie longue (K_s, Λ, etc) ou l'interaction avec le matériau du détecteur. Après cette sélection, l'algorithme reconsruit un vertex secondaire inclusif. Plusieurs propriétés de ce vertex sont utiles pour identifier les jets b, telles que sa masse, le nombre de traces, la distance au PV, la fraction d'énergie des traces au vertex par rapport à toutes les traces dans le jet.

L'algorithme JetFitter tente de reconstruire la topologie de désintégration en cascade complètement, du PV au hadron b, et ensuite au hadron c. L'approche utilisée dans l'algorithme est basée sur l'hypothèse que le vertex primaire et les deux vertex de désintégration des hadrons b et c sont placés le long d'une ligne, approchant la trajectoire du hadron b.

Enfin, les observables discriminantes de plusieurs techniques de b-tagging sont combinées dans un algorithme basé sur un arbre de décision boosté (BDT) et appelé MV2. Les propriétés cinématiques (p_{T} et η) des jets sont incluses dans l'entraînement pour

utiliser des corrélations avec les autres variables d'entrée. MV2 est une mise à niveau de l'algorithme MV1, qui combine les sorties des différents algorithmes de b-tagging en utilisant lui une approche par réseau de neurones. L'algorithme MV2 offre de meilleures performances et facilite l'entraînement et la maintenance logicielle.

Ma contribution principale présentée dans cette thèse est l'optimisation des algorithmes de b-tagging basés sur les paramètres d'impact, en particulier le développement d'une nouvelle catégorisation des traces qui profite de l'ajout de IBL.

Certaines traces sont bien reconstruites et ont donc une meilleure résolution en paramètre d'impact. Mais il y a aussi des traces de moindre qualité: celles qui ont un hit manquant dans l'une des couches du détecteur de pixels ou avec des ambiguïtés dans la reconnaissance des formes. Rejeter toutes les traces de faible qualité réduirait significativement l'efficacité de b-tagging, mais pour en faire un usage efficace, il est nécessaire de diviser les traces en catégories et de traiter chaque catégorie séparément.

La catégorisation des traces au Run 2 a été améliorée par rapport à Run 1, en utilisant plusieurs nouvelles variables de tracking (y compris celles liées à la présence de l'IBL). En particulier, les traces avec un (des) hit(s) manquant(s) sont considérées comme étant moins bien reconstruites que les autres, elles doivent donc être traitées séparément des "meilleures" traces. Les informations sur les hits dans les deux couches du détecteur de pixel les plus internes sont particulièrement importantes pour déterminer les propriétés de la désintégration du hadron b (position du PV et du SV, paramètres d'impact des traces), des mesures précises, particulièrement près du point d'interaction, sont nécessaires. Les traces ont été divisées en 14 catégories exclusives en fonction de leurs distributions des significances d_0/σ_{d_0} et z_0/σ_{z_0}. La nouvelle catégorisation a permis d'améliorer le performance de l'algorithme IP3D au Run 2 par rapport au Run 1 de $\sim 15\%$.

Une autre partie de l'optimisation des algorithmes basée sur les paramètres d'impact concerne la sélection des traces. La plupart des traces associées aux jets de quarks légers proviennent du PV, et ont donc un paramètre d'impact proche de zéro, qui permet de les distinguer des traces de jets b. Cependant, à la fois dans les jets de quarks légers et jets b, il existe une contamination des traces dites *secondaires* ou "mauvaises", qui proviennent de particules de longue durée de vie, telles que K_s, Λ, interactions avec le matériel du détecteur et conversions de photons ($\gamma \to e^+e^-$). Ces traces ont un grand paramètre d'impact, elles peuvent donc être dommageables pour l'identification des jets b.

Pour réduire l'effet négatif des traces "mauvaises", il faut les identifier et les rejeter. Pour ce faire, l'algorithme SV identifie si un traces est susceptible de provenir d'une particule à longue durée de vie. Cette procédure a été activée dans les algorithmes IP2D et IP3D en 2016. L'application de la procédure de suppression de traces "mauvaises" SV pour sélectionner des traces permet d'augmenter le performance de l'algorithme IP3D de $\sim 10\%$.

Recherche de $t\bar{t}H$ ($H \to b\bar{b}$)

Le boson de Higgs, découvert au Run 1 du LHC, a été observé dans plusieurs modes de production, mais pas dans le canal de production en association avec une paire de quarks top. La recherche de la production du boson de Higgs dans ce canal est l'un des objectifs

les plus importants du Run 2 du LHC. La mesure de la section efficace de $t\bar{t}H$ serait un test essentiel du MS, car elle permettrait de déterminer la valeur du couplage de Yukawa au quark top, un paramètre important du MS.

Le mode de désintégration du boson de Higgs en une paire de quarks b, $H \to b\bar{b}$, est dominant dans le MS. Cependant, il est difficile à observer expérimentalement, par rapport aux canaux avec des photons ou des électrons dans l'état final, en raison d'un grand bruit de fond multijet. En dehors de cela, ce canal de désintégration est particulièrement intéressant car il permet de mesurer le couplage de Yukawa au quark b, qui est le deuxième plus grand couplage du boson de Higgs à un fermion dans le MS.

Trois canaux différents du processus $t\bar{t}H$ ($H \to b\bar{b}$) sont explorés dans ATLAS:

- Le canal entièrement hadronique $t\bar{t}H \to (q\bar{q}b)(q\bar{q}b)(b\bar{b})$. Il s'agit d'une analyse complexe en raison de l'importance de la production multijet, difficile à modéliser avec précision, ainsi que du grand bruit de fond combinatoire.

- Le canal dileptonique $t\bar{t}H \to (\ell\nu b)(\ell\nu b)(b\bar{b})$. Deux neutrinos contribuent tous deux à la $E_{\mathrm{T}}^{\mathrm{miss}}$, ce qui rend difficile la reconstruction de la topologie de l'événement.

- Le canal à un lepton $t\bar{t}H \to (\ell\nu b)(q\bar{q}b)(b\bar{b})$. C'est le canal d'analyse le plus sensible parmi les trois. La reconstruction de la topologie de l'événement est plus facile que pour le canal de dileptonique, car il n'y a qu'un seul neutrino dans l'état final et sa cinématique peut être déterminée à partir de la $E_{\mathrm{T}}^{\mathrm{miss}}$. Le terme *lepton* se réfère ici à un électron ou un muon.

La recherche présentée dans cette thèse est basée sur des données de collisions pp à $\sqrt{s} = 13$ TeV enregistrées par le détecteur ATLAS en 2015 et 2016. La luminosité intégrée correspondante est de 36.1 fb^{-1}. Pour estimer les contributions du signal et de la plupart des bruits de fond, une simulation Monte Carlo (MC) a été effectuée. Les événements sont sélectionnés avec des triggers à un seul électron ou à un seul muon. Les événements doivent avoir un lepton et au moins cinq jets. Des exigences supplémentaires sont faites en fonction des informations de b-tagging.

Après la présélection, les événements sont dominés par le bruit de fond de la production $t\bar{t}$. Les événements présélectionnés sont catégorisés en régions exclusives (régions de signal et de fond) en fonction de leur multiplicité de jets et de leurs caractéristiques de b-tagging pour profiter de la multiplicité plus élevée des jets et des jets-b pour le signal $t\bar{t}H$.

Le faible nombre d'événements de signal par rapport à ceux de bruits de fond irréductibles rend cette analyse difficile. Ainsi, des techniques de discrimination efficaces sont cruciales. Pour mieux distinguer le signal du bruit de fond, plusieurs méthodes ont été développées. Celles-ci comprennent trois techniques différentes qui exploitent la présence d'une résonance $H \to b\bar{b}$. Leurs sorties sont combinées avec d'autres variables dans un discriminant multivarié final, qui est utilisé pour le fit procédure dans toutes les régions de signal. La reconstruction avec des arbres de décision boostés (BDT) tente de reconstruire la topologie du signal $t\bar{t}H$, en particulier la cinématique du boson de Higgs. Elle reconstruit le système $t\bar{t}H$ en trouvant le meilleur appariemment entre les jets observés et les partons.

La méthode des éléments de matrice (MEM) évalue les probabilités de vraisemblance sous les hypothèses du signal $t\bar{t}H$ et du bruit de fond $t\bar{t}+b\bar{b}$ en calculant la section efficace différentielle normalisée au niveau de la reconstruction à partir des éléments de la matrice pour ces processus. Cette méthode est appliquée uniquement dans la région du signal avec le rapport signal sur bruit le plus élevé.

La méthode discriminante de vraisemblance (LHD) calcule également les vraisemblances de signal et de fond, mais en utilisant des fonctions de densité de probabilité (pdfs) dérivées de la simulation MC plutôt que du calcul d'éléments de matrice. Elle exploite les informations cinématiques de tous les objets de l'état final reconstruits, et teste les événements sous les hypothèses du signal $t\bar{t}H$ et du bruit de fond $t\bar{t}$ + jets, en considérant toutes les affectations possibles entre les jets reconstruits et les partons. Le développement et l'optimisation de cette nouvelle méthode est la contribution principale de cette thèse.

Les probabilités d'un événement donné sous les hypothèses du signal $P^{\text{sig}}(\mathbf{x})$ ou du bruit de fond $P^{\text{bkg}}(\mathbf{x})$ sont calculées à l'aide de fonctions de densité de probabilité basées sur de la simulation MC (pdfs). Les pdfs sont des fonctions des quadri-vecteurs $\mathbf{x}$ d'objets reconstruits dans cet événement: les jets, le lepton et le neutrino. La variable discriminante finale est définie comme

$$D = \frac{P^{\text{sig}}}{P^{\text{sig}} + P^{\text{bkg}}}. \tag{82}$$

Différentes résonances des masses invariantes fournissent des informations utiles pour séparer le signal du bruit de fond. Ce sont: la masse du boson de Higgs pour l'hypothèse du signal, et les masses du quark top leptonique, du quark top hadronique et du boson W hadronique pour les hypothèses de signal et de fond. Les pdfs de ces masses invariantes sont les variables les plus significatives utilisées dans cette méthode. Les autres pdfs des masses invariantes exploitées dans la méthode sont la masse invariante du système $t\bar{t}$ et la masse invariante du système $t\bar{t}+b\bar{b}$. Outre les masses invariantes, un pouvoir de discrimination supplémentaire est fourni en exploitant les informations sur le spin des différentes particules, en particulier du boson de Higgs. Comme l'origine partonique des jets n'est pas connue, la probabilité de signal doit être calculée en sommant toutes les permutations de jet possibles dans l'événement. Les informations de b-tagging sont ensuite utilisées pour donner des poids différents aux permutations.

La sortie du discriminateur de vraisemblance finale LHD devient une entrée pour la classification BDT avec le BDT de reconstruction, le discriminant MEM et d'autres variables cinématiques dans toutes les régions de signal. Le LHD a été appliqué pour la première fois dans cette analyse et s'est avéré être la variable unique la plus discriminante, fournissant une amélioration de la puissance de séparation de $\sim 10\%$.

Le rapport de la section efficace $t\bar{t}H$ mesurée à la prédiction du MS, obtenue dans une combinaison des canaux à un lepton et deux leptons, est $\mu = 0.84^{+0.64}_{-0.61}$, en supposant une masse du boson de Higgs de 125 GeV. Ce résultat est cohérent à la fois avec l'hypothèse de fond seulement et avec la prédiction $t\bar{t}H$ SM. Une valeur de μ supérieure à 2.0 est exclue à un niveau de confiance de 95%. Le résultat de la combinaison des recherches de production de $t\bar{t}H$ par ATLAS a fourni $\mu = 1.17^{+0.33}_{-0.30}$. La signification observée est 4.2σ, ce qui représente une évidence de la production de $t\bar{t}H$.

References

[1] S. Weinberg, *A model of leptons*, Phys. Rev. Lett. **19** (1967) 1264.

[2] A. Salam, *Weak and electromagnetic interactions*, Proc. of the 8th Nobel Symposium (1969) 367.

[3] S. L. Glashow, *Partial symmetries of weak interactions*, Nucl. Phys. **22** (1961) 579.

[4] Particle Data Group Collaboration, C. Patrignani et al., *Review of Particle Physics*, Chin. Phys. C **40** (2016) 100001, `http://pdg.lbl.gov`.

[5] M. Gell-Mann, *A Schematic Model of Baryons and Mesons*, Phys. Lett. **8** (1964) 214.

[6] H. Fritzsch, M. Gell-Mann, and H. Leutwyler, *Advantages of the Color Octet Gluon Picture*, Phys. Lett. B **47** (1973) 365.

[7] S. Weinberg, *Nonabelian Gauge Theories of the Strong Interactions*, Phys. Rev. Lett. **31** (1973) 494.

[8] D. J. Gross and F. Wilczek, *Ultraviolet Behavior of Nonabelian Gauge Theories*, Phys. Rev. Lett. **30** (1973) 1343.

[9] G. Altarelli, *Partons in Quantum Chromodynamics*, Phys. Rept. **81** (1982) 1.

[10] F. Englert and R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, Phys. Rev. Lett. **13** (1964) 321.

[11] P. W. Higgs, *Broken Symmetries, Massless Particles and Gauge Fields*, Phys. Lett. **12** (1964) 132.

[12] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, Phys. Rev. Lett. **13** (1964) 508.

[13] 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]`.

[14] F. Bezrukov and M. Shaposhnikov, *Why should we care about the top quark Yukawa coupling?*, J. Exp. Theor. Phys. **120** (2015) 335–343, `arXiv:1411.1923 [hep-ph]`, [Zh. Eksp. Teor. Fiz.147,389(2015)].

[15] OPAL, DELPHI, LEP Working Group for Higgs boson searches, ALEPH, L3, *Search for the standard model Higgs boson at LEP*, Phys. Lett. **B565** (2003) 61–75, `arXiv:hep-ex/0306033 [hep-ex]`.

[16] CDF and D0 Collaborations, *Higgs Boson Studies at the Tevatron*, Phys. Rev. **D88** (2013) 052014, `arXiv:1303.6346 [hep-ex]`.

[17] ATLAS and CMS Collaborations, *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*, JHEP **08** (2016) 045, arXiv:1606.02266 [hep-ex].

[18] CDF and D0 Collaborations, *Evidence for a particle produced in association with weak bosons and decaying to a bottom-antibottom quark pair in Higgs boson searches at the Tevatron*, Phys. Rev. Lett. **109** (2012) 071804, arXiv:1207.6436 [hep-ex].

[19] ATLAS Collaboration, *Evidence for the $H \to b\bar{b}$ decay with the ATLAS detector*, arXiv:1708.03299 [hep-ex].

[20] CMS Collaboration, *Evidence for the decay of the Higgs Boson to Bottom Quarks*, CMS-PAS-HIG-16-044 (2017), https://cds.cern.ch/record/2278170.

[21] LHC Higgs Cross Section Working Group https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWGCrossSectionsFigures.

[22] LHC Higgs Cross Section Working Group Collaboration, J. R. Andersen et al., *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties*, arXiv:1307.1347 [hep-ph].

[23] 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. B **716** (2012) 1–29, arXiv:1207.7214 [hep-ex].

[24] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. B **716** (2012) 30–61, arXiv:1207.7235 [hep-ex].

[25] ATLAS Collaboration, *Measurements of the properties of the Higgs-like boson in the four lepton decay channel with the ATLAS detector using 25 fb^{-1} of proton-proton collision data*, ATLAS-CONF-2013-013 (2013), https://cds.cern.ch/record/1523699.

[26] ATLAS Collaboration, *Measurement of the Higgs boson mass in the $H \to ZZ^* \to 4\ell$ and $H \to \gamma\gamma$ channels with $\sqrt{s}$=13 TeV pp collisions using the ATLAS detector*, ATLAS-CONF-2017-046 (2017), https://cds.cern.ch/record/2273853.

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

[28] O. S. Bruning, P. Collier, P. Lebrun, S. Myers, R. Ostojic, J. Poole, and P. Proudlock, *LHC Design Report Vol.1: The LHC Main Ring*, CERN-2004-003 (2004).

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

[30] ALICE Collaboration, K. Aamodt et al., *The ALICE experiment at the CERN LHC*, JINST **3** (2008) S08002.

[31] LHCb Collaboration, A. A. Alves, Jr. et al., *The LHCb Detector at the LHC*, JINST **3** (2008) S08005.

[32] J. C. Collins, D. E. Soper, and G. F. Sterman, *Factorization of Hard Processes in QCD*, Adv. Ser. Direct. High Energy Phys. **5** (1989) 1–91, arXiv:hep-ph/0409313 [hep-ph].

[33] ATLAS Collaboration, *Luminosity Results for Run-1*, https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResults.

[34] ATLAS Collaboration, *Luminosity Results for Run-2*, https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2.

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

[36] ATLAS Collaboration, M. Capeans et al., *ATLAS Insertable B-Layer Technical Design Report*, CERN-LHCC-2010-013, ATLAS-TDR-19 (2010), https://cds.cern.ch/record/129163.

[37] A. La Rosa, *The ATLAS Insertable B-Layer: from construction to operation*, JINST **11** (2016) C12036, arXiv:1610.01994 [physics.ins-det].

[38] A. Miucci, *The ATLAS Insertable B-Layer project*, JINST **9** (2014) C02018.

[39] E. Celebi et al., *Test beam studies of the TRD prototype filled with different gas mixtures based on Xe, Kr, and Ar*, J. Phys. Conf. Ser. **798** (2017) 012172, arXiv:1612.02623 [physics.ins-det].

[40] Y. Nakahama, *The ATLAS Trigger System: Ready for Run-2*, J. Phys. Conf. Ser. **664** (2015) 082037.

[41] ATLAS Collaboration, *The ATLAS Simulation Infrastructure*, Eur. Phys. J. C **70** (2010) 823, arXiv:1005.4568 [physics.ins-det].

[42] M. Dobbs and J. B. Hansen, *The HepMC C++ Monte Carlo event record for High Energy Physics*, Comput. Phys. Commun. **134** (2001) 41.

[43] J. Alwall et al., *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].

[44] 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].

[45] T. Gleisberg, S. Höche, F. Krauss, M. Schönherr, S. Schumann, et al., *Event generation with SHERPA 1.1*, JHEP **0902** (2009) 007, arXiv:0811.4622 [hep-ph].

[46] 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].

[47] T. Sjöstrand, S. Mrenna, and P. Z. Skands, *PYTHIA 6.4 Physics and Manual*, JHEP **0605** (2006) 026, arXiv:hep-ph/0603175.

[48] S. Schumann and F. Krauss, *A Parton shower algorithm based on Catani-Seymour dipole factorisation*, JHEP **0803** (2008) 038, arXiv:0709.1027 [hep-ph].

[49] GEANT4 Collaboration, *GEANT4: A Simulation toolkit*, Nucl. Instrum. Meth. A **506** (2003) 250.

[50] ATLAS Collaboration, *Updates of the ATLAS Tracking Event Data Model (Release 13)*, ATL-SOFT-PUB-2007-003 (2007), https://cds.cern.ch/record/1038095.

[51] R. Frühwirth, *Application of Kalman filtering to track and vertex fitting*, Nucl. Instrum. Meth. **A262** (1987) 444.

[52] ATLAS Collaboration, *Performance of the ATLAS Track Reconstruction Algorithms in Dense Environments in LHC Run 2*, CERN-EP-2017-045 (2017), arXiv:1704.07983 [hep-ex].

[53] ATLAS Collaboration, *Impact Parameter Resolution*, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/IDTR-2015-007/.

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

[55] ATLAS Collaboration, *Vertex reconstruction performance of the ATLAS detector at $\sqrt{s}$ = 13 TeV*, ATL-PHYS-PUB-2015-026 (2015), http://cdsweb.cern.ch/record/2037717.

[56] K. Grimm, *Primary Vertex Reconstruction at the ATLAS Experiment*, ATL-SOFT-PROC-2017-051 (2017), https://cds.cern.ch/record/2253428.

[57] ATLAS collaboration, *Electron efficiency measurements with the ATLAS detector using the 2015 LHC proton-proton collision data*, ATLAS-CONF-2016-024 (2016), https://cds.cern.ch/record/2157687.

[58] ATLAS Collaboration, *Muon reconstruction performance of the ATLAS detector in proton-proton collision data at $\sqrt{s}$ =13 TeV*, Eur. Phys. J. C **76** (2016) 292, arXiv:1603.05598 [hep-ex].

[59] M. Cacciari, G. Salam, and G. Soyez, *The anti-k_t jet clustering algorithm*, JHEP **04** (2008) 063, arXiv:0802.1189 [hep-ph].

[60] ATLAS Collaboration, *Jet energy scale measurements and their systematic uncertainties in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, arXiv:1703.09665 [hep-ex].

[61] ATLAS Collaboration, *Tagging and suppression of pileup jets*, ATLAS-CONF-2014-018, https://cds.cern.ch/record/1700870.

[62] ATLAS Collaboration, *Performance of Missing Transverse Momentum Reconstruction in ATLAS studied in Proton-Proton Collisions recorded in 2012 at 8 TeV*, ATLAS-CONF-2013-082 (2013), https://cds.cern.ch/record/1570993.

[63] ATLAS Collaboration, *Performance of Missing Transverse Momentum Reconstruction in Proton-Proton Collisions at 7 TeV with ATLAS*, Eur. Phys. J. C **72** (2012) 1844, arXiv:1108.5602 [hep-ex].

[64] ATLAS Collaboration, *Expected performance of missing transverse momentum reconstruction for the ATLAS detector at $\sqrt{s} = 13$ TeV*, ATL-PHYS-PUB-2015-023 (2015), https://cds.cern.ch/record/2037700.

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

[66] G. Piacquadio, *Identification of b-jets and investigation of the discovery potential of a Higgs boson in the $WH \to l\nu b\bar{b}$ channel with the ATLAS experiment*, 2010. https://cds.cern.ch/record/1243771.

[67] ATLAS Collaboration, *Performance of b-Jet Identification in the ATLAS Experiment*, JINST **11** (2016) P04008, arXiv:1512.01094 [hep-ex].

[68] J. Gao, M. Guzzi, J. Huston, H.-L. Lai, Z. Li, et al., *CT10 next-to-next-to-leading order global analysis of QCD*, Phys. Rev. D **89** (2014) 033009, arXiv:1302.6246 [hep-ph].

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

[70] K. Mochizuki, *Search for the Higgs boson in the $WH \to \ell\nu b\bar{b}$ channel with the ATLAS detector - Development of high performance b-jet identification algorithms.* PhD thesis, Marseille, CPPM, 2015-10-27. http://inspirehep.net/record/1503539/files/CERN-THESIS-2015-367.pdf.

[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] ATLAS Collaboration, *The Optimization of ATLAS Track Reconstruction in Dense Environments*, https://cds.cern.ch/record/2002609.

[73] ATLAS Collaboration, *Commissioning of the ATLAS b-tagging algorithms using $t\bar{t}$ events in early Run-2 data*, https://cds.cern.ch/record/2047871.

[74] 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** (2015) 349, arXiv:1503.05066 [hep-ex].

[75] NNPDF Collaboration, R. D. Ball et al., *Parton distributions for the LHC Run II*, JHEP **04** (2015) 040, arXiv:1410.8849 [hep-ph].

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

[77] 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].

[78] F. Cascioli, P. Maierhöfer, and S. Pozzorini, *Scattering Amplitudes with Open Loops*, Phys. Rev. Lett. **108** (2012) 111601, arXiv:1111.5206 [hep-ph].

[79] M. Guzzi, P. Nadolsky, E. Berger, H.-L. Lai, F. Olness, and C. P. Yuan, *CT10 parton distributions and other developments in the global QCD analysis*, arXiv:1101.0561 [hep-ph].

[80] 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. B **734** (2014) 210–214, arXiv:1309.5912 [hep-ph].

[81] T. Gleisberg and S. Höche, *Comix, a new matrix element generator*, JHEP **0812** (2008) 039, arXiv:0808.3674 [hep-ph].

[82] 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].

[83] *Measurement of W and Z Boson Production Cross Sections in pp Collisions at root s = 13 TeV in the ATLAS Detector*, ATLAS-CONF-2015-039 (2015), https://cds.cern.ch/record/2045487.

[84] 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].

[85] N. Kidonakis, *Two-loop soft anomalous dimensions for single top quark associated production with a W- or H-*, Phys. Rev. D **82** (2010) 054018, arXiv:1005.4451.

[86] N. Kidonakis, *NNLL resummation for s-channel single top quark production*, Phys. Rev. D **81** (2010) 054028, arXiv:1001.5034.

[87] N. Kidonakis, *Next-to-next-to-leading-order collinear and soft gluon corrections for t-channel single top quark production*, Phys. Rev. D **83** (2011) 091503, arXiv:1103.2792.

[88] M. Bahr et al., *Herwig++ Physics and Manual*, Eur. Phys. J. **C58** (2008) 639–707, arXiv:0803.0883 [hep-ph].

[89] 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), https://cdsweb.cern.ch/record/1951336.

[90] ATLAS Collaboration, *Estimation of fake lepton background for top analyses using the Matrix Method with the 2015 dataset at $\sqrt{s}$ =13 TeV with AnalysisTop-2.4.13*, ATL-COM-PHYS-2016-198 (2016), https://cds.cern.ch/record/2135116.

[91] ATLAS Collaboration, *Search for the Standard Model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector* , https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2017-076/.

[92] A. Hocker et al., *TMVA - Toolkit for Multivariate Data Analysis*, PoS **ACAT** (2007) 040, arXiv:physics/0703039 [PHYSICS].

[93] R. E. Ticse Torres, *Search for the Higgs boson in the $ttH(H \to bb)$ channel and the identification of jets containing two B hadrons with the ATLAS experiment (PhD thesis)*, CERN-THESIS-2016-123, http://cds.cern.ch/record/2226037/.

[94] CMS Collaboration, V. Khachatryan et al., *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. C **75** (2015) 251, arXiv:1502.02485 [hep-ex].

[95] M. R. Whalley, D. Bourilkov, and R. C. Group, *The Les Houches accord PDFs (LHAPDF) and LHAGLUE*, in *HERA and the LHC: A Workshop on the implications of HERA for LHC physics. Proceedings, Part B*. 2005. arXiv:hep-ph/0508110 [hep-ph].

[96] ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s}$ = 8 TeV using the ATLAS detector at the LHC*, Eur. Phys. J. C **76** (2016) 653, arXiv:1608.03953 [hep-ex].

[97] 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.

[98] ATLAS Collaboration, *Jet energy measurement and its systematic uncertainty in proton-proton collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector*, Eur. Phys. J. C **75** (2015) 17, arXiv:1406.0076 [hep-ex].

[99] J. N. Ng and P. Zakarauskas, *QCD-parton calculation of conjoined production of Higgs bosons and heavy flavors in p anti-p collisions*, Phys. Rev. D **29** (1984) 876.

[100] S. Dawson, L. H. Orr, L. Reina, and D. Wackeroth, *Associated top quark Higgs boson production the LHC*, Phys. Rev. D **67** (2003) 071503, arXiv:hep-ph/0211438.

[101] LHC Higgs Cross Section Working Group Collaboration, D. de Florian et al., *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*, arXiv:1610.07922 [hep-ph].

[102] 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].

[103] ATLAS Collaboration, *Multi-Boson Simulation for 13 TeV ATLAS Analyses*, ATL-PHYS-PUB-2016-002 (2016), https://cds.cern.ch/record/2119986.

[104] J. M. Campbell and R. K. Ellis, *$t\bar{t}W^{+-}$ production and decay at NLO*, JHEP **1207** (2012) 052, arXiv:1204.5678 [hep-ph].

[105] L. Moneta, K. Belasco, K. S. Cranmer, S. Kreiss, A. Lazzaro, D. Piparo, G. Schott, W. Verkerke, and M. Wolf, *The RooStats Project*, PoS **ACAT2010** (2010) 057, arXiv:1009.1003 [physics.data-an].

[106] 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].

[107] A. L. Read, *Presentation of search results: The CL(s) technique*, J. Phys. G **28** (2002) 2693.

[108] ATLAS Collaboration, *Evidence for the associated production of the Higgs boson and a top quark pair with the ATLAS detector*, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2017-077/.

[109] ATLAS Collaboration, *Measurements of Higgs boson properties in the diphoton decay channel with 36.1 fb^{-1} pp collision data at the center-of-mass energy of 13 TeV with the ATLAS detector*, https://cds.cern.ch/record/2273852.

[110] ATLAS Collaboration, *Measurement of the Higgs boson coupling properties in the $H \to ZZ^* \to 4\ell$ decay channel at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, `https://cds.cern.ch/record/2273849`.

A Observed results from the fit to data

The results of fit to data under the signal plus background (S+B) hypothesis for single-lepton channel, dilepton channel and their combination are presented. Figures 92 and 93 show the NPs corresponding to theoretical and instrumental systematic uncertainties. The normalisation factors for $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ components and the ratio of the measured $t\bar{t}H$ cross-section to the SM prediction μ are shown in figure 94.

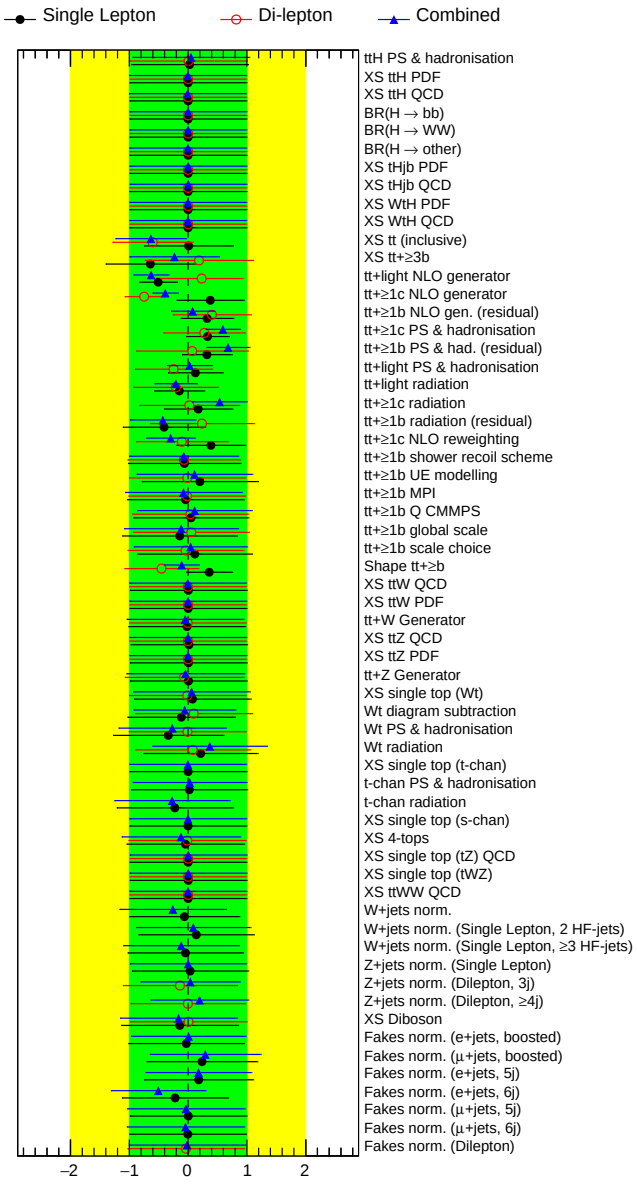


Figure 92: Nuisance parameter corresponding to theoretical systematic uncertainties for BDT-based fits to data.

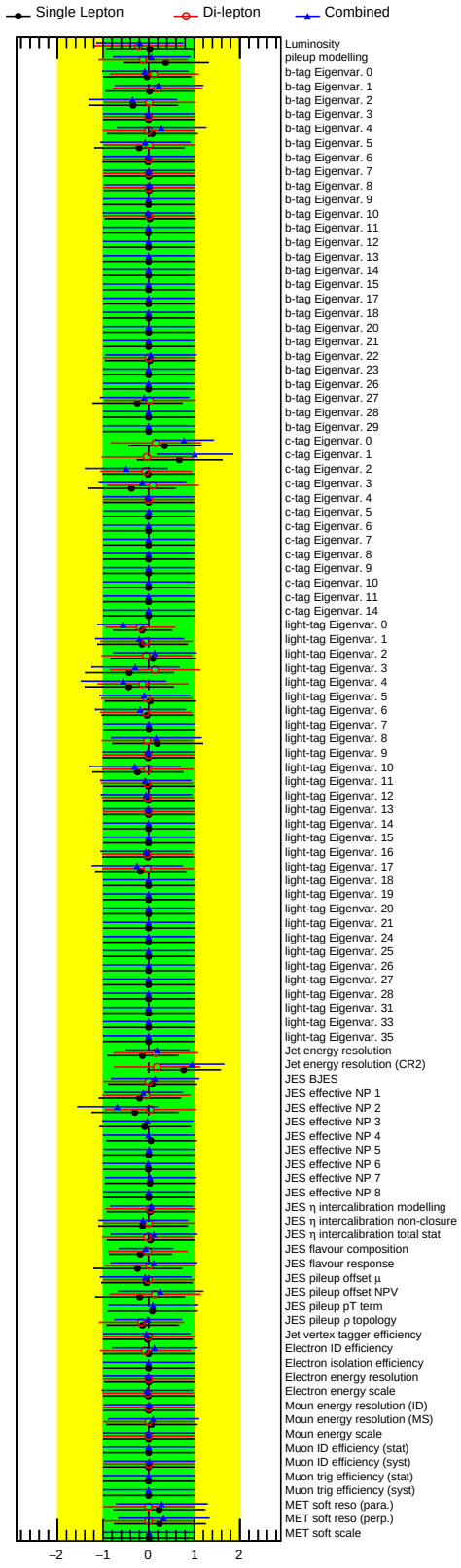


Figure 93: Nuisance parameter corresponding to instrumental systematic uncertainties for BDT-based fits to data.

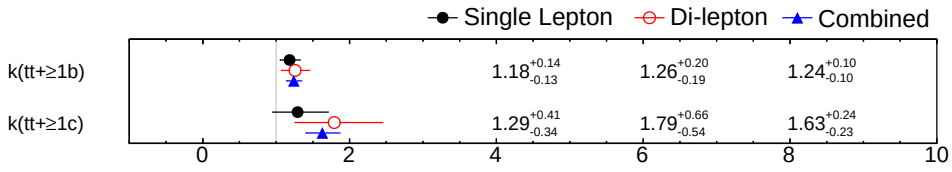


Figure 94: Normalisation factors for $t\bar{t}+ \geq 1b$ and $t\bar{t}+ \geq 1c$ components and the signal strength μ for BDT-based fits to data.

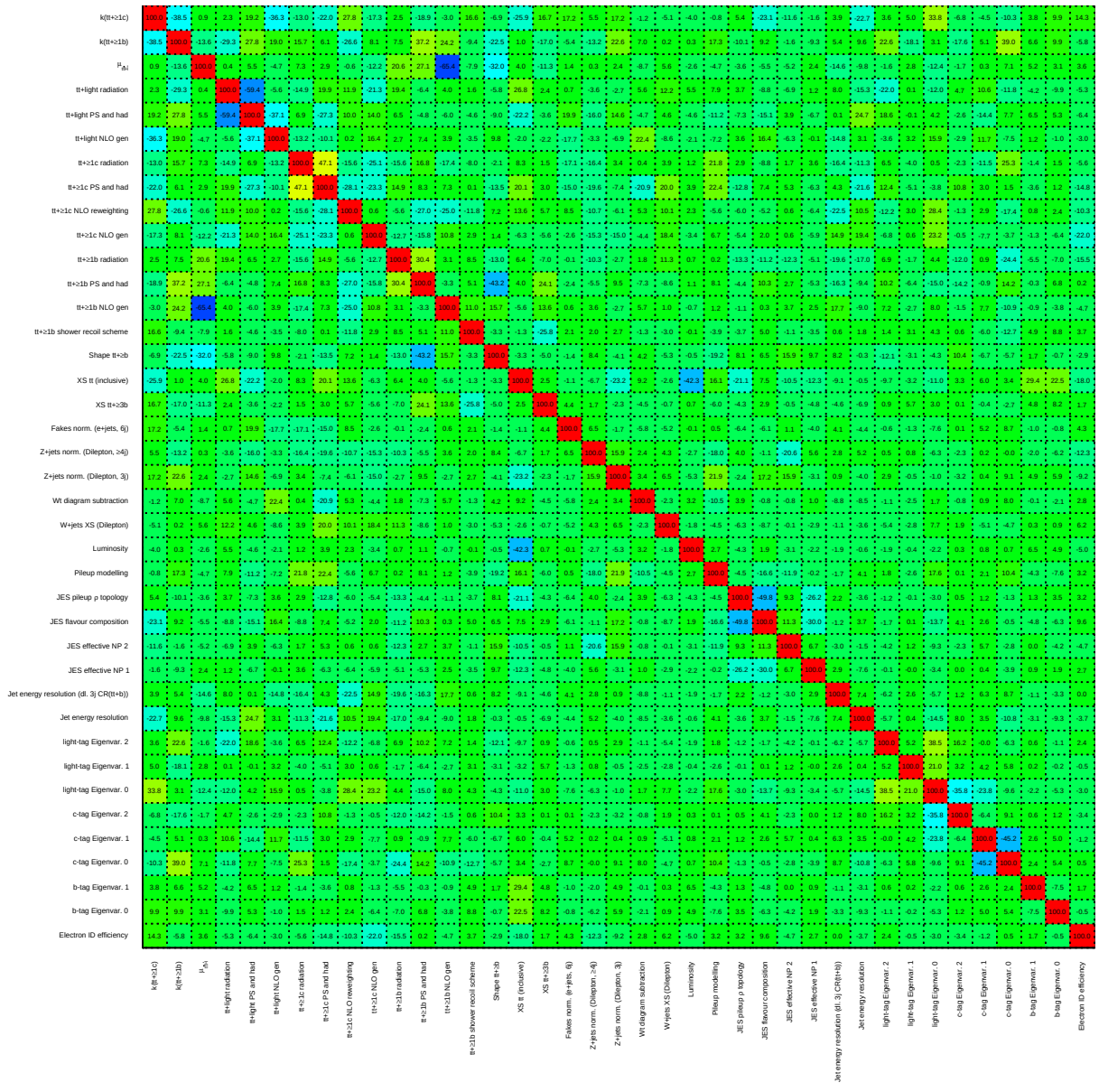


Figure 95: Correlation matrix corresponding to the fits to data under the S+B hypothesis.

B Expected results from the fit to the Asimov dataset

The expected results obtained by performing fits to the Asimov data set under the signal plus background (S+B) hypothesis for single-lepton channel, dilepton channel and their combination are presented. The expected constraints on NPs corresponding to theoretical and instrumental systematic uncertainties resulting from the fits are shown in figures 96 and 97. The corresponding expected correlation matrix for the fitted NPs can be found in figure 99. The expected uncertainties on the normalisation factors for $t\bar{t} + \geq 1b$ and $t\bar{t} + \geq 1c$ and the signal strength are shown in figure 98.

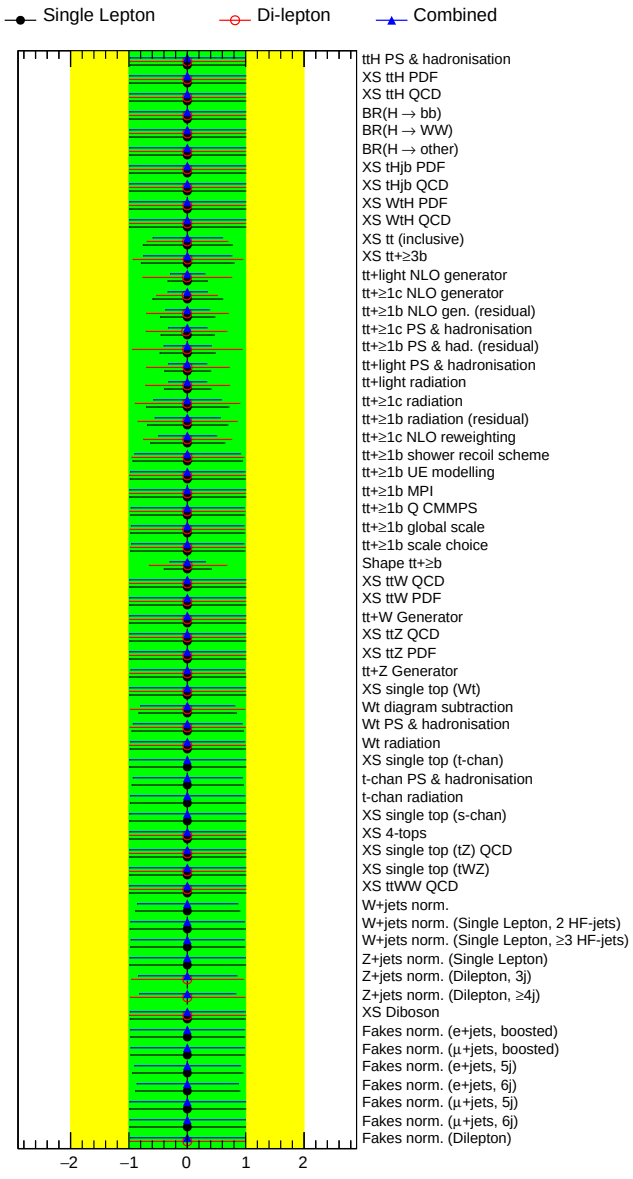


Figure 96: Nuisance parameter corresponding to theoretical systematic uncertainties corresponding to the fit to the Asimov dataset under the S+B hypothesis.

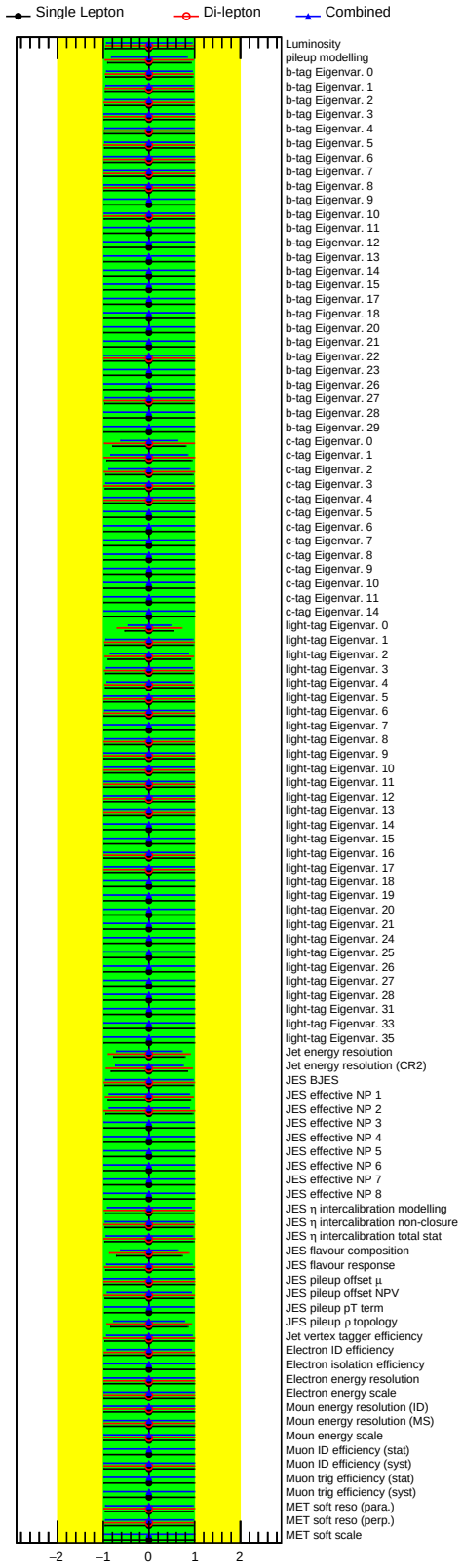


Figure 97: Nuisance parameter corresponding to instrumental systematic uncertainties corresponding to the fit to the Asimov dataset under the S+B hypothesis.

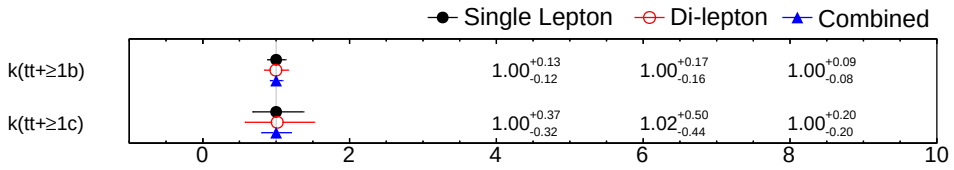


Figure 98: Expected uncertainties on the normalisation factors for $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ components and the signal strength corresponding to the fit to the Asimov dataset under the S+B hypothesis.

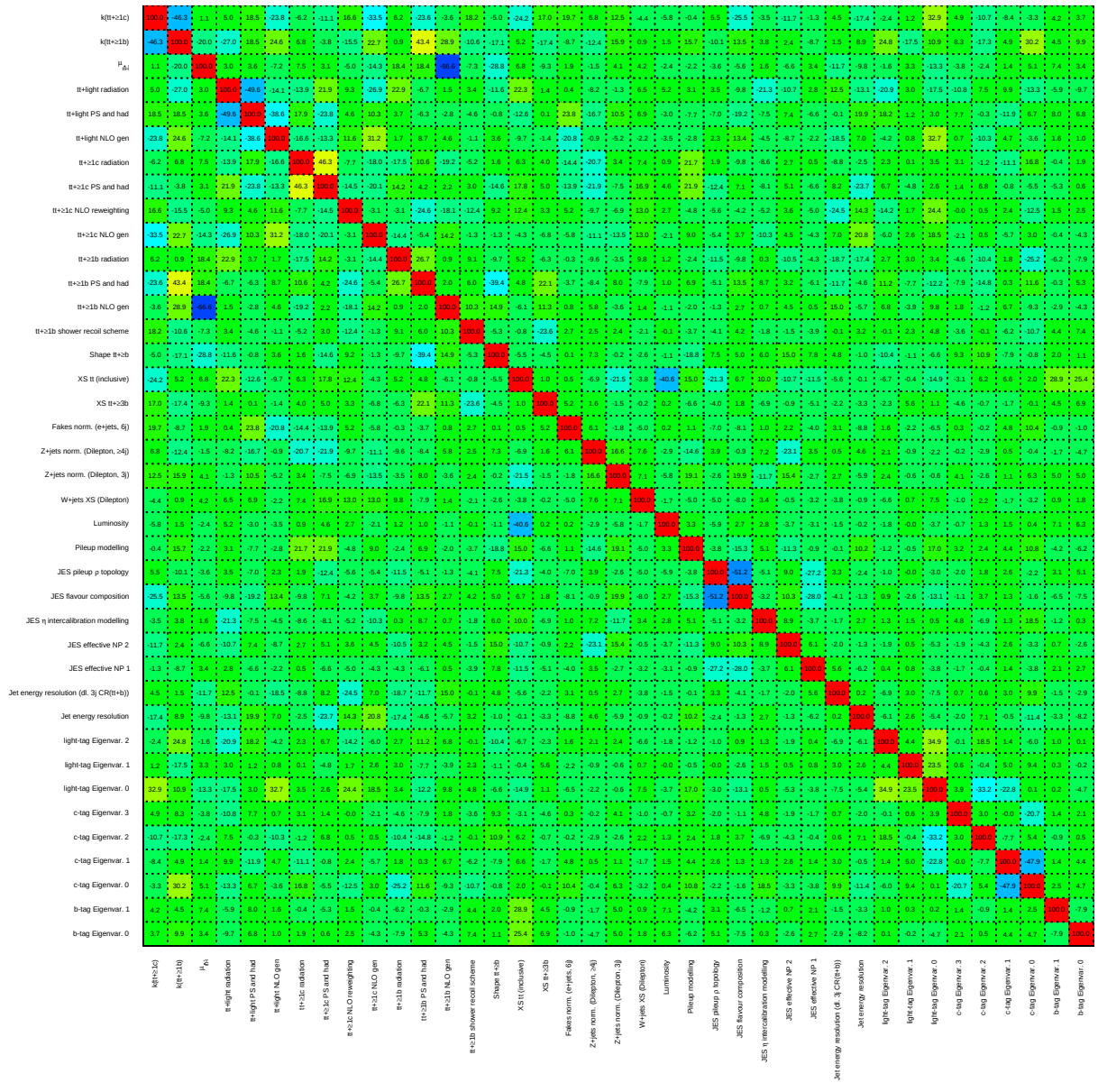


Figure 99: Correlation matrix corresponding to the fit to the Asimov dataset under the S+B hypothesis.

C LHD distributions before and after the fit

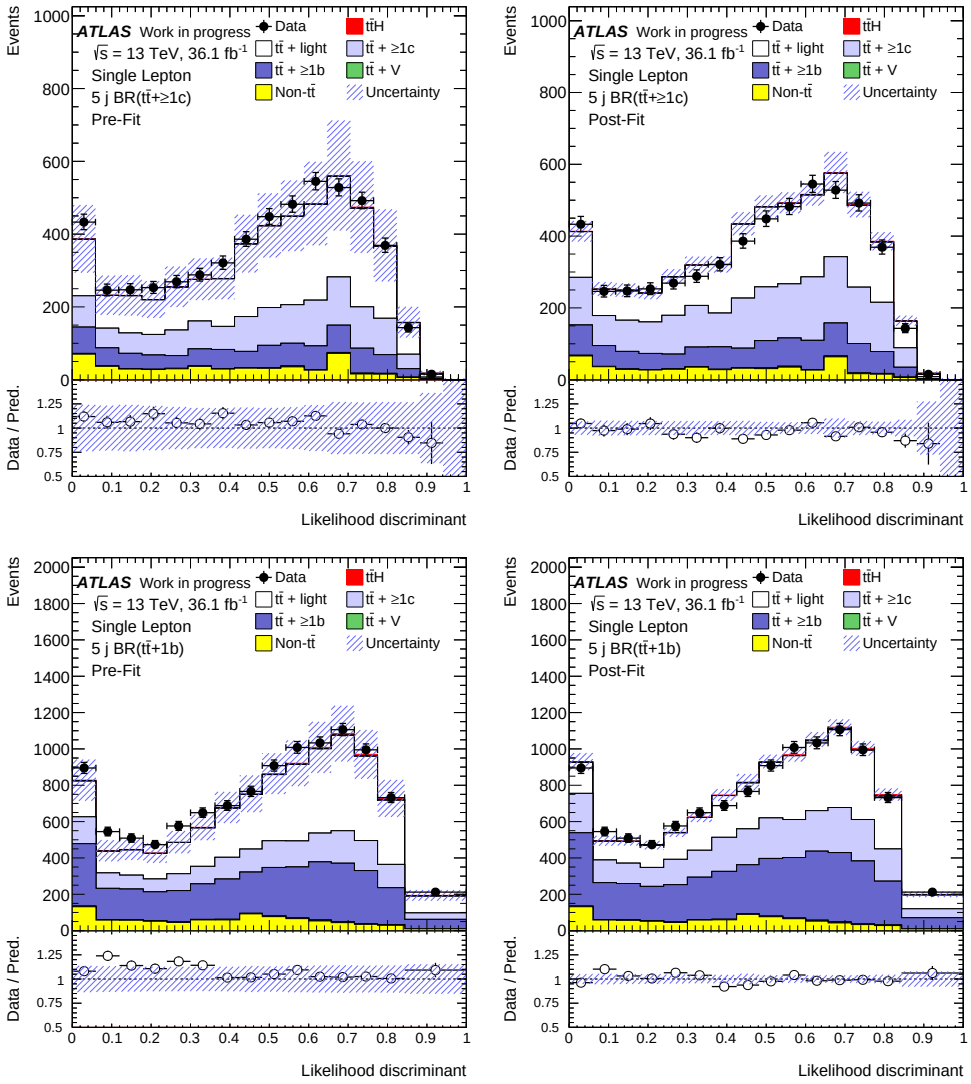


Figure 100: Distributions of the LHD output in the $\mathrm{CR}^{5\mathrm{j}}_{t\bar{t}+\geq 1c}$ and $\mathrm{CR}^{5\mathrm{j}}_{t\bar{t}+1b}$ regions (left) before and (right) after the fit to data.

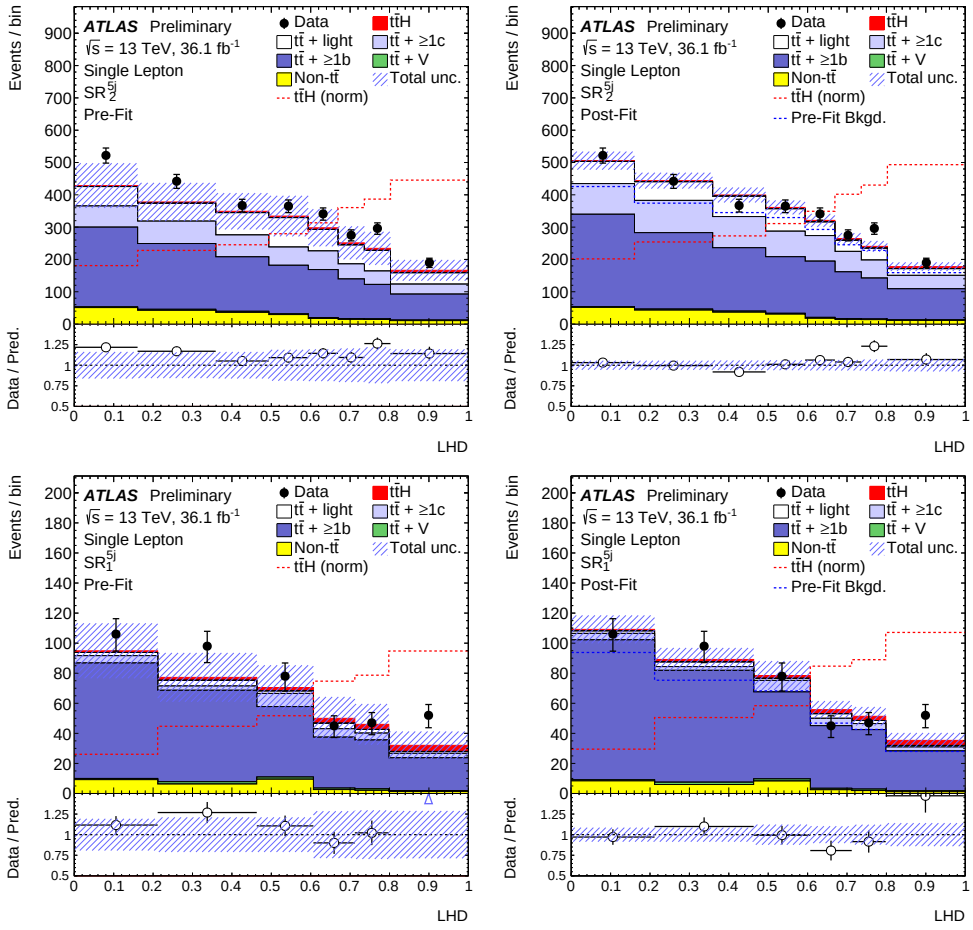


Figure 101: Distributions of the LHD output in the SR_1^{5j} and SR_2^{5j} regions (left) before and (right) after the fit to data.

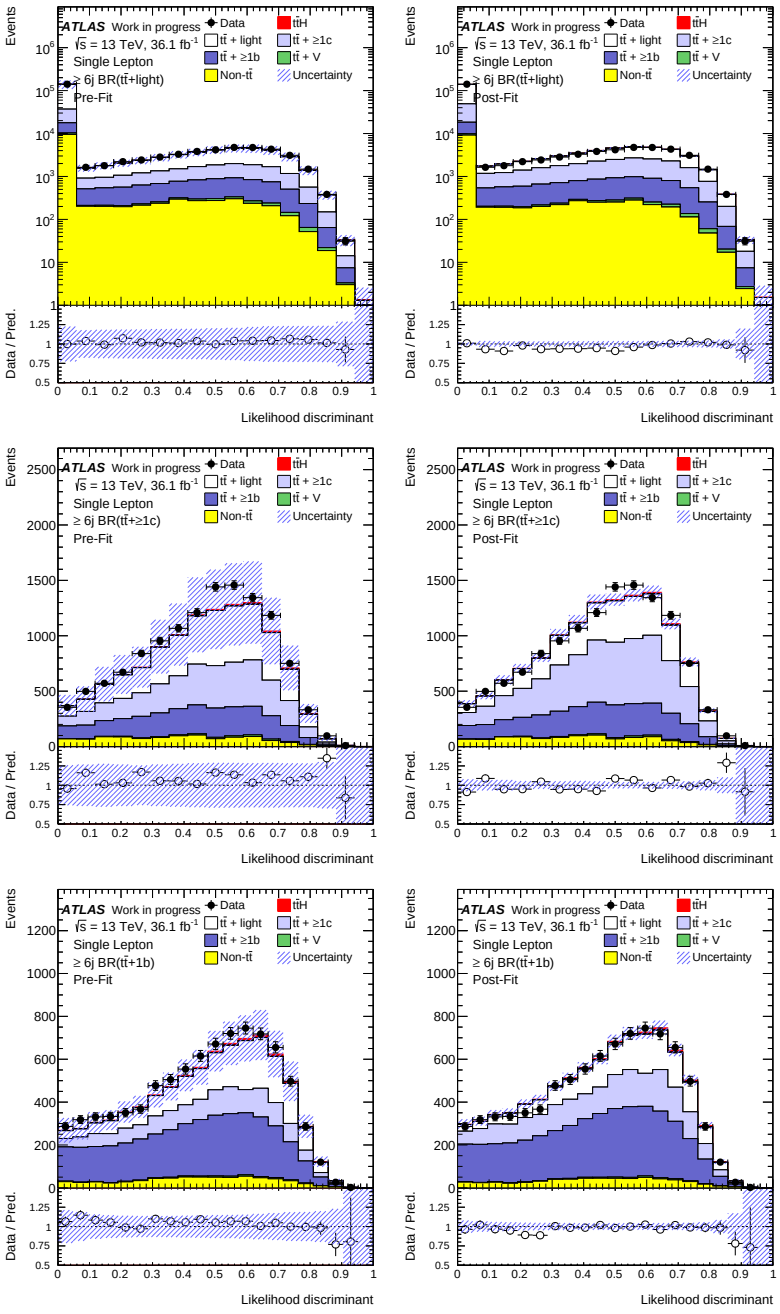


Figure 102: Distributions of the LHD output in the $\mathrm{CR}^{\geq 6\mathrm{j}}_{t\bar{t}+\mathrm{light}}$, $\mathrm{CR}^{\geq 6\mathrm{j}}_{t\bar{t}+\geq 1c}$ and $\mathrm{CR}^{\geq 6\mathrm{j}}_{t\bar{t}+1b}$ regions (left) before and (right) after the fit to data.

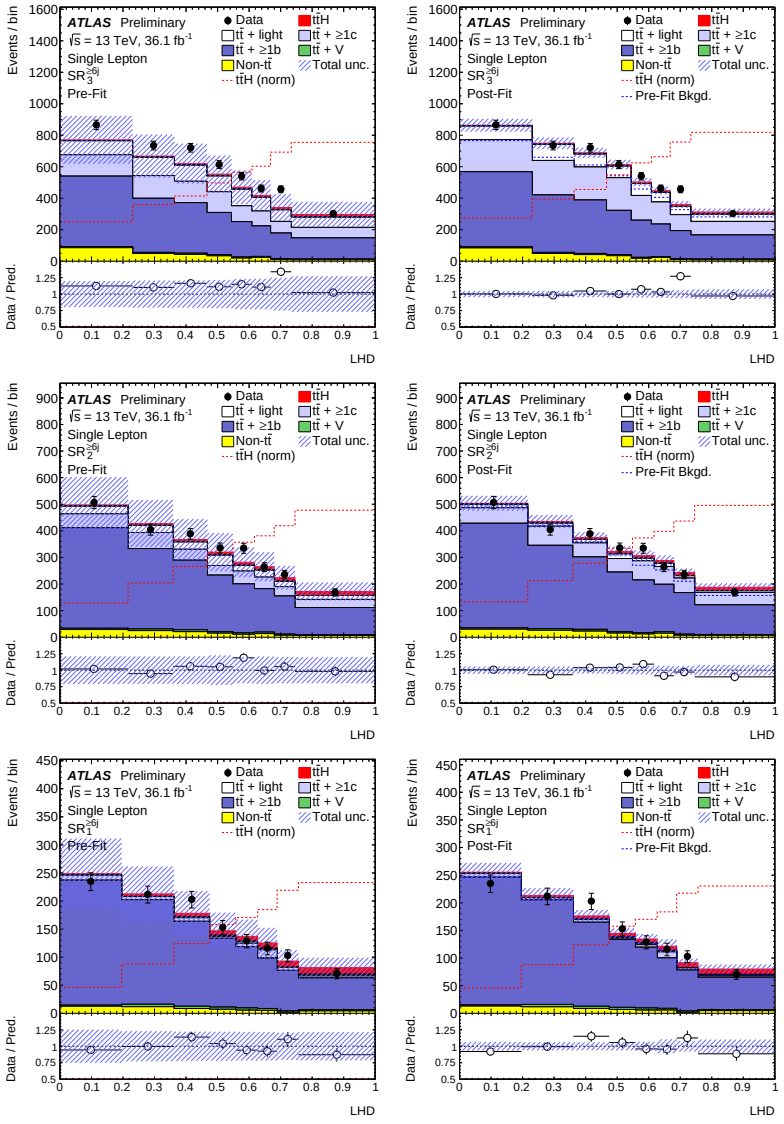


Figure 103: Distributions of the LHD output in the $\mathrm{SR}_3^{\geq 6\mathrm{j}}$, $\mathrm{SR}_2^{\geq 6\mathrm{j}}$ and $\mathrm{SR}_1^{\geq 6\mathrm{j}}$ regions (left) before and (right) after the fit to data.