Measurement of the $t\bar{t}H$ production cross-section with a collimated $H \to b\bar{b}$ decay in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector

Dissertation submitted
for the award of the title

"Doctor of Natural Sciences"

to the Faculty of Physics, Mathematics, and Computer Science
of Johannes Gutenberg University Mainz

Eftychia Tzovara

Born in Athens, Greece

JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

Mainz, December 2022

Date of oral examination: 30.06.2023

1^{st} reviewer: Prof. Dr. Lucia Masetti

2^{nd} reviewer: Prof. Dr. Dmitry Budker

Summary

The predictions of the Standard Model (SM) of particle physics have been probed with remarkable accuracy, so far. The Large Hadron Collider (LHC) at CERN has significantly contributed to this quest. A remarkable achievement of the ATLAS and CMS experiments at the LHC was the discovery of the Higgs boson in 2012, the last missing piece of the SM. With the increasing amount of proton-proton collisions delivered by the LHC, more precise measurements of the Higgs boson are now possible, while rare processes are accessible as well.

A property of the Higgs boson that is of particular importance is its coupling to the top quark, which is expected to be the strongest in the SM due to the high mass of the top quark. Therefore, its precise measurement is a stringent test of the SM. A direct measurement of the top-quark Yukawa coupling can be assessed through the Higgs-boson production in association with a pair of top quarks ($t\bar{t}H$). This thesis presents the measurement of the $t\bar{t}H$ process with a subsequent Higgs-boson decay to a pair of b-quarks ($H \to b\bar{b}$), the decay mode with the largest branching ratio. The measurement is performed with data collected by the ATLAS detector, corresponding to an integrated luminosity of 139 fb^{-1} at a center-of-mass energy of 13 TeV.

Events with one or two charged leptons from the $t\bar{t}$ decay in the final state are considered to the measurement. The main challenge of the $t\bar{t}H(H \to b\bar{b})$ channel emerges from the large SM backgrounds from the production of top-quark pairs with additional jets ($t\bar{t}$+jets). Also the many jets coming from b-hadrons (b-jets) in the final state cause combinatorial ambiguities. Thus, the identification of such jets is decisive in order to determine the signal and reject many background processes. The $t\bar{t}H$ events are split into exclusive analysis regions, based on the number of leptons, jets, and jets tagged as b-jets, providing regions enhanced in signal, or in the main background components. Specifically in the single-lepton channel, a boosted category is defined by selecting events in which the Higgs boson and possibly also the hadronically decaying top quark are produced with high transverse momentum (p_T), with their decay products being collimated in large-radius jets. The single-lepton boosted channel targets events with Higgs-boson candidate $p_T \geq 300$ GeV and is the main scope of this thesis.

To identify the reconstructed objects with the underlying particles and to maximise the discrimination of the $t\bar{t}H$ signal from the overwhelming $t\bar{t}$+jets background events in the signal-enriched regions, machine-learning algorithms are employed. The background is dominated by a $t\bar{t}$ process with an additional gluon in the final state which further splits into a pair of b-quarks ($t\bar{t}+b\bar{b}$). Besides, a large number of heavy-flavour jets in the final state is not well modelled, thus many systematic uncertainties have to be considered, decreasing the sensitivity of the measurement. All the defined analysis regions are analysed together in a combined profile likelihood fit to test for the presence of signal. The fit simultaneously determines the event yields for the signal and the most important background component, while constraining the overall background model within the assigned systematic uncertainties.

Eventually, the ratio of the measured $t\bar{t}H$ cross section to the SM expectation in the inclusive cross-section measurement is found to be $0.35^{+0.36}_{-0.34}$, corresponding to an observed (expected) significance of 1.0 (2.7) standard deviations. A $t\bar{t}H$ signal strength larger than the SM prediction is excluded at 95% confidence level. The measurement uncertainty is dominated by systematic uncertainties, mainly regarding the theoretical knowledge of the $t\bar{t}+\geq 1b$ background process. Finally, to further test the SM, the cross-section is measured differentially as a function of the generator-level Higgs-boson p_T, taking advantage of the reconstruction of the Higgs-boson kinematics.

Contents

Chapter 1

Introduction

During the last century, a worthwhile progress has been made in the field of high energy physics. The predictions of the Standard Model (SM) of particle physics, the most successful theory that describes the building blocks of matter and their interactions, have been probed with remarkable accuracy. The Large Hadron Collider (LHC) [108], the world's largest and highest-energy particle collider placed at CERN, has significantly contributed to this quest. A remarkable achievement of the ATLAS [116] and CMS [117] experiments at the LHC was the discovery of the Higgs boson in 2012 [10, 11], the last missing piece of SM. This discovery confirms the mechanism that generates massive vector bosons and fermion masses through the Yukawa coupling [5–9]. Various properties of the Higgs boson have been determined so far, confirming its compatibility with the SM, but there are still many that are yet to be measured. Any potential deviation from the SM may give an insight into physics beyond the SM [69–73]. With the increasing amount of proton-proton collisions delivered by the LHC, more precise measurements of the Higgs boson are possible, while rare processes are accessible as well.

Considering that the Yukawa coupling increases proportionally to the fermion masses, the coupling of the top quark, the heaviest particle in the SM, to the Higgs boson is expected to be the strongest. Therefore, its precise measurement is a stringent test of the SM. However, the Higgs boson decay into a pair of top quarks on mass shell is strongly suppressed. The first attempts to extract this Yukawa coupling was through the gluon fusion [12] and the Higgs-boson decay to a pair of photons [89], which provide a clear signature and found to be in agreement with the SM. Nevertheless, both processes involve a top-quark loop, providing only an indirect evidence of that coupling, since not all the contributors to the loop can be known.

A direct measurement of the top-quark Yukawa coupling can be assessed through the Higgs-boson production in association with a pair of top quarks ($t\bar{t}H$). Although it contributes only around 1% of the total Higgs boson production cross section at the LHC [106], the top quarks in the final state offer a distinctive signature in the detector, allowing access to many Higgs boson decay modes. This thesis presents the measurement of the $t\bar{t}H$ process with a subsequent Higgs-boson decay to a pair of b-quarks ($H \to b\bar{b}$), the largest decay mode with a branching ratio of about 58% [106]. Although the $t\bar{t}H$ [102] and $H \to b\bar{b}$ [104] processes have been observed independently, the combined $t\bar{t}H(H \to b\bar{b})$ process has not been observed, yet. The measurement is performed using data collected by the ATLAS detector, corresponding to an integrated luminosity of 139 fb^{-1} at a center-of-mass energy of 13 TeV [100].

For this measurement, events with one or two charged leptons from the $t\bar{t}$ decay in the final state are considered. The $t\bar{t}H(H \to b\bar{b})$ channel suffers from the large SM backgrounds from the production of top-quark pairs with additional jets ($t\bar{t}$+jets). Due to the many jets

originating from b-hadrons (b-jets) in the final state, combinatorial ambiguities arise. Thus, the identification of such jets is decisive in order to determine the signal and reject many background processes. Although the analysis targets the $H \to b\bar{b}$ decay, all the decay modes may contribute to the signal. The $t\bar{t}H$ events are split into exclusive analysis regions, based on the number of leptons, jets, and jets tagged as b-jets, providing regions enhanced in signal, or in the dominant background components. Especially in the single-lepton channel, a boosted category is designed to select events in which the Higgs boson and possibly also the hadronically decaying top quark are produced with high transverse momentum (p_T), so that their decay products are collimated in large-radius jets. The single-lepton boosted channel, which targets events with Higgs-boson candidate $p_T \geq 300$ GeV, is the main subject of this thesis.

Machine-learning algorithms are used to identify the reconstructed objects with the underlying particles and to maximise the separation between the $t\bar{t}H$ signal events and the overwhelming $t\bar{t}$+jets background, classifying events in the signal-enriched regions. The background is dominated by a $t\bar{t}$ process with an additional gluon in the final state which further splits into a pair of b-quarks ($t\bar{t}+b\bar{b}$). However, a large number of heavy-flavour jets in the final state is not well modelled, thus many systematic uncertainties have to be accounted for, decreasing the sensitivity of the measurement. The signal-enriched regions are analysed together with the signal-depleted ones in a combined profile likelihood fit to test for the presence of signal. The fit simultaneously determines the event yields for the signal and the most important background component, while constraining the overall background model within the assigned systematic uncertainties. An inclusive cross-section measurement is performed and a differential one, which could be sensitive to effects beyond the SM. In the latter, the cross-section is measured as a function of the generator-level Higgs boson p_T in the simplified template cross-sections formalism, exploiting the possibility to reconstruct the Higgs boson kinematics.

The thesis starts with a general description of the theory behind the SM, the importance of the Higgs mechanism in the SM as well as the coupling of fermions to the Higgs field, described in Chapter 2. It also details the Higgs-boson and top-quark production and decay modes. Then, the final-state products of the complex $t\bar{t}H(H \to b\bar{b})$ process are measured by multiple subsystems within the ATLAS detector at the LHC, which are outlined in Chapter 3. Afterwards, Chapter 4 summarises the Monte Carlo (MC) simulations that serve as the theoretical predictions of the signal and background processes that take place at the LHC.

Furthermore, the particle reconstruction and identification of the measured physics objects are detailed in Chapter 5. Then, the selection criteria applied to events and physics objects, describing the event categorisation in the different analysis regions, are discussed in Chapter 6. Also, the multivariate analysis techniques exploited to reconstruct the $t\bar{t}H$ signal events and to separate them from the dominant $t\bar{t}$+jets background, explicitly in the single-lepton boosted region, are outlined. In Chapter 7, the various systematic uncertainties that affect the measurement, arising from the physics object reconstruction and the modelling of the physics processes, are discussed.

The statistical analysis employed to extract the $t\bar{t}H$ signal cross-section is introduced in Chapter 8 and the final fit model is summarised. Also, the fit results for both the inclusive and differential cross-section measurements, after combining all the analysis regions, as well as their interpretation are presented. Finally, Chapter 9 contains a summary and the conclusions of the study presented in this thesis. Also, possibilities of further tools and techniques that could improve the $t\bar{t}H(H \to b\bar{b})$ measurement, and especially optimise the performance of the single-lepton boosted region, are discussed.

Chapter 2

The Standard Model of Particle Physics

The 20^{th} century was a milestone in the evolution of physics. It marked the beginning of a new era, referred to as modern physics, which changed our interpretation of the universe and its content. Already at the dawn of that century, Einstein proposed the special theory of relativity [1], demonstrating the relation between space and time. Soon after it was followed by the general theory of relativity [1] encompassing gravity, as well as describing the history of our universe. A few years later, quantum mechanics [2] was formulated as the theory of matter and light at the atomic scale.

Although special relativity and quantum mechanics could successfully interpret many experimental results, they failed in explaining interactions involving the relativistic creation and annihilation of particles. These processes are of particular interest in high-energy scattering experiments. A fully relativistic quantum theory required the development of quantum field theory (QFT) [15]. The fundamental objects of this theory are the quantum fields describing elementary particles, defined at all points in space-time.

The first complete QFT, quantum electrodynamics (QED) [18] was formulated about in the middle of 20^{th} century, providing a fully relativistic quantum description of the electromagnetic interaction. QED describes all phenomena involving electrically charged particles interacting with the electromagnetic field. About a decade later, an attempt to unify the weak and electromagnetic interactions into a single QFT, the so-called electroweak theory (EWT) [19–21], was made. It was finalised when the spontaneous symmetry breaking mechanism was incorporated, through which originally massless gauge bosons and fermions could acquire mass. A few years later, another QFT was formulated, quantum chromodynamics (QCD) [17], in order to explain phenomena involving the strong interaction.

These theoretical breakthroughs, EWT and QCD, constitute the Standard Model (SM) of particle physics, which was finalised in the mid-1970s along with the experimental confirmation of the existence of quarks. The SM summarises the known fundamental structures of matter and forces (electromagnetic, weak, and strong interactions) in the universe except gravity. The SM predictions have been experimentally verified in the subsequent decades. Especially, the discovery of the top quark (1995) [23], and the Higgs boson (2012) [5–11] (important component of the spontaneous symmetry breaking mechanism) have added further credence to the SM. However, the SM cannot be considered a complete theory, since it cannot describe all phenomena observed in nature, such as gravity.

2.1 The fundamental particles and interactions

The SM of particle physics represents our current understanding of the fundamental particles and the interactions with each other. The particle content of the SM and the basic attributes of particles are depicted in Fig. 2.1. The elementary particles are divided into two categories, the fundamental fermions and fundamental bosons, according to their quantum statistics.

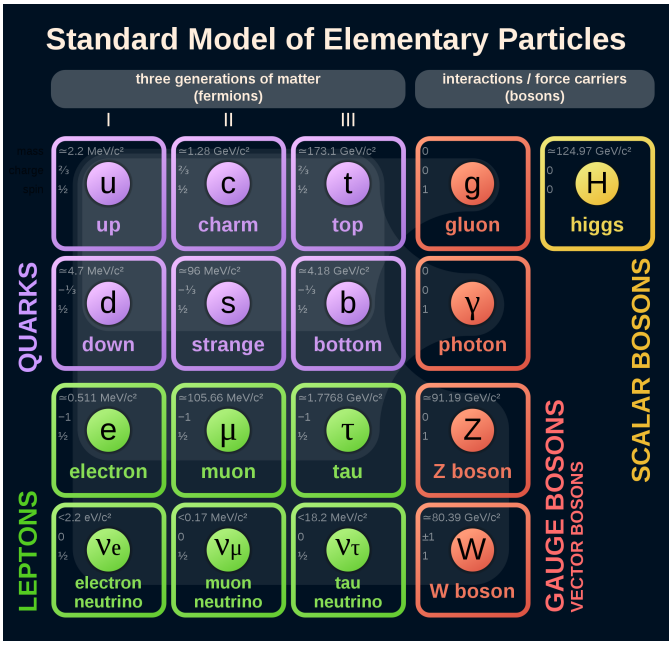


Fig. 2.1: Elementary particles of the Standard Model of particle physics [3]. They consist of three generations of quarks and leptons, as well as five force-carrying bosons. For each particle the mass, charge and spin is given.

Fermions obey the *Fermi–Dirac statistics* and according to the *spin–statistics theorem*, they have half-integer spin. Moreover, fermions respect the *Pauli exclusion principle*, as a result, more than one fermion cannot be in the same quantum state at the same time. The SM includes 12 fundamental fermions of spin $1/2$. They are classified into two families, quarks and leptons, according to their quantum numbers, as well as, the interactions they undergo. They are further divided into three generations with the same quantum numbers, apart from their masses. Particles of the 2^{nd} and 3^{rd} generation are unstable and decay into 1^{st} generation particles. Therefore, stable matter in our universe consists solely of 1^{st} generation particles, while the other generations can only be observed in high-energy collisions, such as in particle accelerators. Each fermion has a corresponding antiparticle. Each antiparticle has the same mass, mean lifetime, and spin as its respective fermion, while its charges have equal magnitude but opposite sign.

In the SM, there are six flavours of quarks (and the corresponding anti-quarks). They are further categorised as up-type quarks (up (u), charm (c), top (t)) carrying an electric charge of $+2/3$e; and down-type quarks (down (d), strange (s), bottom (b)) with electric charge $-1/3$e. Eventually, each generation contains an up- and a down-type quark. In addition, quarks have a unique attribute called colour charge, which comes in three different types, arbitrarily labeled as red, blue, and green. Each of them is complemented by the corresponding anti-colour.

Furthermore, there are six flavours of leptons (and the corresponding anti-leptons, as

well). There are also two types of leptons. The ones carry an integer electric charge of -1e (electron (e), muon (μ), tau (τ)) while, the others are electrically neutral (electron- (ν_e), muon- (ν_μ), tau-neutrino (ν_τ)). Each generation contains a charged lepton and the corresponding neutrino.

In relativistic QFT, the behaviour of the free spin-$1/2$ fermions is described by the *Dirac equation* $(i\gamma^\mu\partial_\mu - m)\psi = 0$[1] [41], expressed in natural units (i.e. $\hbar = c = 1$). The wavefunction ψ, which represents a fermionic field, is the product of a plane wave and a four-component column vector in the spin space, the so-called *Dirac spinor*. There are four independent solutions of the Dirac equation, two with positive energy $\psi = u(p)e^{-ip\cdot x}$ and two with negative energy $\psi = \upsilon(p)e^{ip\cdot x}$, while the Dirac spinors depend on the four-momentum. The positive-energy solutions can be interpreted as the up- and down-spin states of a fermion, while the negative-energy solutions correspond to an anti-fermion.

For relativistic massless particles, the projection of their spin $\vec{S}$ on the direction of their motion $\hat{p} = \frac{\vec{p}}{|\vec{p}|}$ is a conserved quantity. This quantity is called *helicity* and its conservation law follows from the solution of the Dirac equation. A particle with spin pointing into the direction of its motion possesses a positive helicity and it is called *right-handed*, or left-handed if opposite. However, for massive particles, helicity is not a Lorentz invariant quantity. Therefore, *chirality* is defined, as the related Lorentz invariant quantity, through the *chiral projection operators* $\frac{1}{2}(1 \pm \gamma^5)$, where γ^5 is the Dirac matrix $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3 = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix}$. Any fermion field can thus be projected into its left- or right-handed component by acting with the projection operators

$$\psi_L = \frac{1}{2}(1 - \gamma^5)\psi \qquad\qquad \psi_R = \frac{1}{2}(1 + \gamma^5)\psi. \tag{2.1}$$

Experimental results show that all produced and observed neutrinos have left-handed helicities, and all anti-neutrinos have right-handed helicities. Hence, these are the only chiralities included in the SM of particle interactions.

In contrast with fermions, bosons follow Bose–Einstein statistics and according to the spin–statistics theorem, they have integer spin. Additionally, they do not follow the Pauli exclusion principle, namely, more than one boson can simultaneously be in the same quantum configuration. The SM describes various fundamental bosons; twelve gauge bosons (vector bosons) which have spin 1 (photon (γ), $W\pm$, Z, 8 gluons (g)) and a scalar boson with spin 0 (Higgs (H)). Gluons exist in eight different states and carry a combination of colour and anti-colour charge. The elementary vector bosons mediate interactions among fermions, acting as force carriers, whereas the Higgs boson is responsible for the intrinsic mass of particles. In fact, the elementary fermions and gauge bosons are initially considered to be massless. However, the SM posits that all of them, except for the photon and gluons, acquire their masses when interacting with the Higgs field, through the so-called *Higgs mechanism* [6, 7, 9] (Sec. 2.2.5).

Moreover, among fermions, neutrinos are considered to remain massless within the SM, which is still in agreement with current direct experimental observations of the mass. However, recent indirect experiments observing neutrino oscillations[2] require neutrinos with non-

[1] The coefficient ∂_μ is the four-gradient $\partial_\mu = (\frac{\partial}{\partial t}, \vec{\nabla})$, while γ^μ represents the 4×4 Dirac γ-matrices $\gamma^\mu = (\gamma^0, \vec{\gamma}) = (\gamma^0, \gamma^1, \gamma^2, \gamma^3)$.

[2] Neutrino oscillation is a quantum mechanical phenomenon which describes transformations of the neutrino's lepton flavour. It was experimentally discovered by the Super-Kamiokande Observatory and the Sudbury Neutrino Observatories.

vanishing masses. In addition, there are theories that treat neutrinos as Majorana[3] particles, which are investigated in neutrino-less double beta-decay experiments. Such theories, in combination with the seesaw mechanism, may explain the existence of neutrino masses and why they are many orders of magnitude smaller than the charged fermions, which would constitute extensions of the SM. After all, since neutrino masses are supposed to be very small, neutrinos are treated as massless Dirac fields in the context of this thesis.

The elementary particles interact with each other through the four fundamental forces in nature. The strong interaction is the strongest force among them, followed by the electromagnetic force which is about two orders of magnitude weaker. Subsequently, the weak interaction is several orders of magnitude ($10^{-7} - 10^{-6}$) less strong than the strong force. Finally, the gravitational force, although it is the most familiar interaction, it is not yet included in the SM. However, it is by far the weakest and its impact is assumed to be negligible in the interactions among elementary particles.

The electromagnetic force, is a physical interaction that occurs between quarks or electrically charged leptons. It is mediated by neutral, massless and non self-interacting photons. Electromagnetism is the only interaction with infinite range in the SM, given that the photon has no mass, and it is described by QED. It is responsible for electromagnetic radiation such as light, as well as for atomic structure, due to the electromagnetic attraction between atomic nuclei and their orbital electrons.

The weak force is responsible for the radioactive decay of atoms. It acts on all quarks and leptons and it is mediated by $W^{\pm}$ or Z bosons. Especially neutrinos, that do not carry any electric charge, interact only weakly, resulting in a very low probability to detect them in nowadays detectors. Therefore, their momentum is determined by measuring the other particles involved in the interaction and applying conservation laws. The weak interaction is very short-ranged, due to the massive mediating bosons. It has the unique feature of changing quark or lepton flavours. The $W^{\pm}$ bosons are electrically charged, therefore they mediate interactions that change the particle flavour and charge (*charged current interactions*). Also, there are weak interactions in which the flavour of the particle is not changed (*neutral current interactions*) and are mediated by the neutral Z bosons. In the SM, the weak force is understood in terms of the EWT.

Colour charged particles, i.e. quarks, interact also through the strong force, mediated by massless gluons. Since gluons themselves carry colour charge (combination of a colour and an anti-colour), they can also interact with each other via the strong force (*gluon self-interaction*). Even though the gluons have no mass, the strong interaction is short-ranged, acting mainly at distances comparable to the diameter of a nucleon. The strong interaction makes up ordinary matter, since it confines quarks into colour-neutral composite particles called hadrons[4], such as the proton and neutron in atomic nuclei. It is described by QCD within the SM.

2.2 Theories describing the fundamental interactions

The SM is a relativistic QFT, implying that its fundamental objects are interpreted as quantum fields that pervade space-time. QFT provides the mathematical framework for the SM based on the Lagrangian formalism, which controls the kinematics and dynamics of a system. Inter-

[3]Majorana fermions do not have an anti-particle partner, but represent their own anti-particles.

[4]*Hadrons* consist of quarks held together by the strong force. They are categorised into mesons (quark, anti-quark pairs) such as pions and kaons, or baryons (3 quarks) like the proton and neutron.

actions among particles are described by interaction terms in the Lagrangian involving their corresponding quantum fields. Every field theory of particle physics is based on certain symmetries observed in nature. According to Noether's theorem [26], every differentiable symmetry of the action of a physical system has an associated conserved quantity. Hence, any formalism describing the nature of fundamental particles must possess such a symmetry, giving the same result at any point in space-time. Finally, the construction of SM Lagrangian proceeds by first postulating the symmetries of the system, and then by expressing the renormalisable Lagrangian from its field content, being invariant under these symmetries.

A field theory in which the Lagrangian is invariant under local transformations is called *gauge theory*. These gauge (or phase) transformations, form the *symmetry* (or *gauge*) *group* of the theory. The fundamental interactions are represented by the symmetry groups in the SM. A corresponding field (usually a vector field) necessarily arises for each group generator, called the *gauge field*. Gauge fields, possessing the properties of causality and locality, are included in the Lagrangian to ensure its invariance under the local group transformations, also called *gauge invariance*. When such a theory is quantised, the quanta of the gauge fields are the *gauge* (or *vector*) *bosons*, that carry the fundamental interactions.

If the symmetry group of a theory is commutative, then this is called *abelian* gauge theory. In 1954, C.N. Yang and R.L. Mills extended the concept of abelian gauge theories of local symmetry groups, to non-abelian gauge theories, also known as Yang–Mills theories [25], based on more complicated local symmetry groups. The quanta of the Yang–Mills field must be massless in order to maintain gauge invariance. Also, the gauge bosons acting as force carriers in a Yang-Mills theory carry charge, as a result they can interact with themselves and radiate further carrier particles.

2.2.1 Quantum Electrodynamics: the theory of electromagnetic interaction

A representative example of a gauge invariant theory is QED. Its development began in the 1920s with the description of radiation and matter interaction by many remarkable scientists. However, a major theoretical obstacle soon emerged with the appearance of various infinities in higher than first order perturbative calculations[5]. The problem was solved with the invention of the renormalisation[6] procedure [124]. Eventually, the complete QED theory was formulated in the late 40's by R.P. Feynman [27, 28], S.I. Tomonaga [29, 30], J. Schwinger [31, 32], and F. Dyson [33, 34]. QED is an abelian gauge theory, which aims at describing the electromagnetic interaction between charged fermions and the massless neutral photon. The group $U(1)_{em}$[7] is thus defined as the gauge group of electromagnetism. The electric charge is a conserved property of the electromagnetic interactions and serves as the generator of the $U(1)_{em}$ symmetry group. The gauge field, which mediates the interaction between the electrically

[5] In particle physics, the mathematical description of interaction processes has no exact solutions, though it is described by an infinite number of additional terms. The first order perturbation theory describes the most basic procedure, while higher orders include additional radiation or loop processes. For the simulation of these processes mostly the leading order (LO) and next-to-leading order (NLO), and sometimes also next-to-next-to-leading order (NNLO) calculations are used.

[6] Renormalisation is a computational procedure, in QFT, for treating infinities which arise in calculated quantities (such as mass and charge) at low or high energy. The main idea is to replace the calculated values of these quantities with their finite measured values. The renormalisation procedure can be applied to arbitrary order in perturbation theory.

[7] The *unitary group* $U(1)$ contains 1×1 unitary matrices, i.e. satisfying $U^\dagger U = 1$, of the form $U = e^{i\alpha(x)}$. $U^\dagger$ is the Hermitian conjugate of matrix U.

charged fermions, is the electromagnetic four-vector potential A_μ, with the photon being the gauge boson.

The kinematics of free fermions, described as spin-1/2 Dirac fields, satisfy the so-called *Dirac Lagrangian*

$$\mathcal{L} = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi. \tag{2.2}$$

In order to obtain the Lagrangian that describes the electromagnetic interaction, the $U(1)_{em}$ local gauge invariance, namely invariance under the local gauge transformation $\psi \rightarrow \psi^{'} = e^{i\alpha(x)}\psi$, should be imposed. This is achieved by replacing the partial derivative by a covariant derivative, $\partial_\mu \rightarrow D_\mu = \partial_\mu - ig_e A_\mu$ and introducing the gauge field A_μ, which is regarded as the physical photon field and couples to the Dirac particle. The field A_μ possesses very specific transformation properties, $A_\mu \rightarrow A^{'}_\mu = A_\mu + \frac{1}{g_e}\partial_\mu \alpha(x)$, in order to ensure the invariance of the Lagrangian. Consequently, the QED Lagrangian consists mainly of three terms; the free motion of fermion fields described by the Dirac equation, the interaction term between the gauge and the fermion field, as well as the free motion of photons described by Maxwell's equations,

$$\begin{aligned} \mathcal{L}_{QED} &= i\bar{\psi}\gamma^\mu D_\mu \psi - m\bar{\psi}\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} \\ &= \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi + g_e \bar{\psi}\gamma^\mu A_\mu \psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} \end{aligned} \tag{2.3}$$

where ψ is the four-vector Dirac spinor for a free spin-$\frac{1}{2}$ particle with mass m. In addition, $\bar{\psi} = \psi^\dagger \gamma^0$ is the adjoint spinor, while γ^μ represents the four Dirac γ-matrices. Also, $g_e \propto eQ$ represents the coupling constant of $U(1)_{em}$ gauge symmetry, where Q is the charge operator corresponding to each fermion and e is the fundamental charge (i.e. the charge of the positron). Lastly, $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the electromagnetic field tensor, which is invariant under the A_μ transformation.

In QED, a mass term, like $\frac{1}{2}m^2 A_\mu A^\mu$, for the gauge boson is prohibited, since it breaks gauge invariance, thus it is not included in the Lagrangian 2.3. As a result, the gauge particle of this theory must be massless, a notion that corresponds to the physical observation of the massless photon, and the gauge field should have infinite range. This theory is one of the best achievements in physics with several predictions confirmed experimentally up to very high precision, such as the anomalous magnetic moment of the electron and the Lamb shift of the energy levels of hydrogen. QED, due to its success, has served as the template to model the weak and the strong interactions.

2.2.2 The theory of weak interaction

The first theory of the weak interaction was developed by E. Fermi [38] in analogy to QED in 1933. He suggested that the $\beta-$decay of neutrons [37] could be explained by a four-fermion interaction, involving a contact force with no range between two vector currents ($V-V$ interaction). The coupling strength of this interaction is described by the *Fermi constant* G_F. This theory couldn't explain some features of the $\beta-$decay, though. Until then, it was believed that all fundamental interactions are invariant under space inversion, namely they conserve parity[8]. In 1956, T.D. Lee and C.N. Yang suggested that the weak interaction could instead violate parity [39]. Eventually, this theory was confirmed by experiments conducted

[8]The *Parity* (*P*) *symmetry* states that the physics laws of a system are invariant under a reflection in space, i.e. a left-right interchange.

by C.S. Wu [40] about a year later. After this discovery, the CP symmetry[9] was proposed to compensate for the parity violation. However until today, various experiments have shown that this symmetry is violated during certain types of weak decay. Finally, the weak interaction is the only fundamental force that violates P- (maximally) and, more rarely, CP-symmetry.

To accommodate parity violation, Fermi's theory making use of only vector currents needed to be extended to include some axial-vector component. In 1958 E.C.G Sudarshan and R.E. Marshak [44, 45] on one side, R. Feynman and M. Gell-Mann [42, 43] on the other, developed an effective field theory based on vector and axial-vector currents ($V-A$ interaction). According to this theory, the weak interaction couples only to left-handed fermions (and right-handed anti-fermions), although in QFTs, such as QED and QCD, left- and right-handed fermions are treated equally. Fermi's theory as well as its extension do not involve any propagators and are thus non-renormalisable. Eventually, the SM completes the description of the weak interaction by introducing massive vector fields, the $W^{\pm}$ the neutral Z boson bosons, as propagators. The coupling strengths of the two theories are related through

$$\frac{G_F}{\sqrt{2}} = \frac{g^2}{8M_W^2} \tag{2.4}$$

where g is the weak coupling constant and M_W is the mass of the charged bosons that mediate this interaction.

Another unique characteristic of the weak force is that it is capable of changing quark or lepton flavours (only within the same generation) through the charged current interaction. Especially in the quark sector, quark transitions in the weak decays are observed predominantly within a generation but also, to a lesser degree, from one generation to another. The mixing among the weak interaction eigenstates of the down-type[10] quarks (d', s', b') and the corresponding mass eigenstates (d, s, b) is characterised by the known Cabibbo-Kobayashi-Maskawa (CKM) matrix [46, 47]. The probability for a transition from a quark q to a quark q' is proportional to the squared magnitude of the matrix element, $|V_{qq'}|^2$.

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = V_{CKM} \begin{pmatrix} d \\ s \\ b \end{pmatrix}, \qquad V_{CKM} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \tag{2.5}$$

$$|V_{CKM}| = \begin{pmatrix} 0.97401 \pm 0.00011 & 0.22650 \pm 0.00048 & 0.00361^{+0.00011}_{-0.00009} \\ 0.22636 \pm 0.00048 & 0.97320 \pm 0.00011 & 0.04053^{+0.00083}_{-0.00061} \\ 0.00854^{+0.00023}_{-0.00016} & 0.03978^{+0.00082}_{-0.00060} & 0.999172^{+0.000024}_{-0.000035} \end{pmatrix} \tag{2.6}$$

The CKM matrix is a unitary matrix and contains four independent parameters; three real angles, that control the mixing among each generation pair, and an imaginary phase, which is responsible for CP-violating phenomena. Various experiments [48] have been conducted for the best determination of the magnitudes of its matrix elements, as depicted in eq. 2.6. The diagonal elements of the CKM matrix describe transitions within a quark generation and

[9] The *Charge Parity* (*CP*) *symmetry* is a combination of the charge conjugation (C) symmetry and the P symmetry. It states that the laws of physics should remain the same if a particle is interchanged with its antiparticle (*C symmetry*) while its spatial coordinates are inverted (*P symmetry*).

[10] The choice of down-type quarks in the definition is a convention, and does not represent a physically preferred asymmetry between up-type and down-type quarks.

deviate from unity by only a few percent. On the contrary, the coupling between different generations is denoted by its off-diagonal elements, being up to two orders of magnitude smaller than the diagonal ones. However, these terms explain the W boson coupling to quarks belonging to two different generations.

In the leptonic sector, the masslessness of the neutrinos implies that there is no mixing of the different lepton generations as there is for quarks. Because in this case, the neutrino eigenstates of the weak interaction are considered to be the same as the corresponding mass eigenstates. However, the existence of neutrino oscillations indicates that neutrinos have non-vanishing masses and that the neutrino eigenstates of the weak interaction are a superposition of the mass eigenstates, in a similar way to d', s' and b' being a superposition of the strong interaction eigenstates d, s and b. The mixing of the neutrinos of different flavours is large, opposite to the weak mixing of the quarks of the different families.

All particles have a property called *weak isospin*, which serves as an additive quantum number that characterises the behaviour of a particle under the weak interaction. It plays the same role as the electric and colour charge in the electromagnetic and strong interaction, respectively. Weak isospin is described by two quantum numbers; the total isospin I and its third component I_3. The latter corresponds to the eigenvalues of the isospin projection on the z-axis for which the flavour states are eigenstates.

Fermions with negative chirality, i.e. left-handed fermions, carry a total weak isospin of $I = \frac{1}{2}$. Each generation of left-handed quarks and leptons forms a doublet of fermions with $I_3 = \pm\frac{1}{2}$, which can transform into each other by emitting (or absorbing) a $W^{\pm}$ boson. By convention, the I_3 ascribed to the electrically charged fermions has the same sign as their electric charge Q. Lacking any distinguishing electric charge, neutrinos are assigned the I_3 opposite to their corresponding charged lepton. Additionally, the electric charges of the two fermions in a doublet always differ by one unit. On the contrary, right-handed fermions have $I = 0$ and can only form singlets with $I_3 = 0$. They do not undergo charged-current weak interactions, but they do all[11] interact with the Z boson. All in all, each fermion has its corresponding anti-fermion with reversed chirality as well as the sign of I_3 and Q (whenever they are non-zero, otherwise they remain zero). As discussed in Sec. 2.1, only left-handed neutrinos and right-handed anti-neutrinos are observed in the SM.

The weak interaction is described by the ${SU(2)_L}$[12] gauge group. The weak isospin operators $\vec{I} = \frac{\tau_i}{2}$, where τ_i ($i = 1, 2, 3$) are the spin Pauli matrices, are the generators[13] of the group. The basis for this representation is conventionally chosen to be the eigenvectors of I_3. The subscript L on the symmetry group denotes that the weak interaction couples only to left-handed fermions.

Unlike charged-current weak interactions, the neutral-current interactions are observed to not have a pure $V - A$ structure, since they couple to right-handed fermions (left-handed anti-fermions) as well, albeit with smaller strength. Therefore, the weak neutral current does not respect the $SU(2)_L$ symmetry. In parallel, one can recall that the electromagnetic interaction is also a neutral current interaction with right- as well as left-handed components. The problem with the soft chiral asymmetry in the neutral-current weak interactions is resolved by the unification with the electromagnetic interaction, outlined in the following.

[11] As already mentioned, right-handed neutrinos are not considered part of the SM. If they actually exist, they would not interact weakly, electromagnetically, or strongly, due to the lack of hypercharge, electric and color charge. They would interact gravitationally though, due to their mass.

[12] The *special unitary group* $SU(2)$ contains 2×2 unitary matrices, transforming as $U(\alpha_i) = e^{-i\alpha_i \tau_i/2}$.

[13] In general, a special unitary group $SU(N)$ has $N^2 - 1$ generators.

2.2.3 Unifying electromagnetic & weak interactions under electroweak theory

Eventually, the SM of particle physics unifies the weak and electromagnetic interactions into a single theoretical system, in which they appear as different manifestations of one fundamental "electroweak" interaction. Building on the parallel with isospin, we are led to consider a weak analog of hypercharge of the strong interaction, the so-called *weak hypercharge*. It is a quantum number relating the electric charge Q and the third component of weak isospin I_3 by the Gell-Mann-Nishijima formula [49, 50],

$$Q = I_3 + \frac{Y}{2}. \tag{2.7}$$

Just as the electric charge Q generates the symmetry group $U(1)_{em}$ of the electromagnetic interaction, so the weak hypercharge Y generates a symmetry group $U(1)_Y$. Thus, these two interactions have been unified under an enlarged symmetry group $SU(2)_L \times U(1)_Y$.

Since we have a product of symmetry groups in the electroweak theory, the generator Y must commute with the generators $\hat{I}$. As a consequence, all the members of an isospin multiplet must have the same value of the hypercharge. This is why $U(1)$ cannot be identified with $U(1)_{em}$ and weak hypercharge had to be introduced. Thus, the leptons in a weak isospin doublet have $Y = -1$, while the quarks $Y = +\frac{1}{3}$. Also, the weak isospin lepton singlets posses $Y = -2$, while up-type and down-type quark singlets have $Y = +\frac{4}{3}$ and $Y = -\frac{2}{3}$, respectively. The weak hypercharge for an anti-fermion is the opposite of that of the corresponding fermion, because Q and I_3 reverse sign under charge conjugation. A summary of the quantum numbers of all fundamental fermions, denoting their coupling to the unified interactions is given in Table 2.1, including their electric charge (electromagnetic interaction) and weak isospin (weak interaction), splitting their ground states into singlets or doublets, as well as the hypercharge.

The SM of electroweak interactions is a Yang-Mills theory, invariant under the weak isospin and hypercharge transformations of the $SU(2)_L \times U(1)_Y$ gauge group. Being non-abelian, the theory introduces massless gauge bosons as mediators in order to maintain gauge invariance. This theory was initially proposed by S. Glashow in 1961 [19], but it was extended, incorporating the spontaneous symmetry breaking, to accommodate the massive vector bosons ($W^{\pm}, Z$), by S. Weinberg (1967) [22] and M.A. Salam (1968) [21]. Later in 1971, G. Hooft [124] proved that non-abelian gauge theories, such as the electroweak theory are renormalisable. Finally, the electroweak theory was completed with its extension from leptons to quarks by S. Glashow, J. Iliopoulos, and L. Maiani [51].

The $SU(2)_L \times U(1)_Y$ invariance of the electroweak sector can be mapped to the $U(1)_{em}$ invariance and the weak sector. Just as the QED Lagrangian resulted from imposing $U(1)_{em}$ local gauge invariance, so the electroweak Lagrangian is obtained by requiring local gauge invariance under the $SU(2)_L \times U(1)_Y$ symmetry group, i.e. under the phase transformations $\psi \to \psi' = e^{i\alpha_i(x)\tau_i/2}\psi$ and $\psi \to \psi' = e^{i\alpha(x)}\psi$, respectively. This is achieved by introducing the covariant derivative $D_\mu = \partial_\mu + ig\frac{1}{2}\vec{\tau}\cdot\vec{W}_\mu + ig'\frac{1}{2}YB_\mu$. The gauge fields of the symmetry groups are represented by the three $SU(2)_L$ gauge bosons $\vec{W}_\mu$ (W^1_μ, W^2_μ, W^3_μ), which couple to the weak isospin, and the $U(1)_Y$ gauge boson B_μ, that couples to weak hypercharge. These fields transform as $B_\mu \to B'_\mu = B_\mu - \frac{1}{g'}\partial_\mu\alpha(x)$ and $W^i_\mu \to W^{i'}_\mu = W^i_\mu - \frac{1}{g}\partial_\mu\alpha_i(x) - \epsilon^{ijk}a_j(x)W^k_\nu$, where g and g' are the coupling constants of the gauge groups $SU(2)_L$ and $U(1)_Y$ respectively,

assuring the invariance of the Lagrangian. Thus, the electroweak Lagrangian is

$$\begin{aligned}\mathcal{L}_{EW} =& \sum_{\chi_L} i\bar{\chi}_L\gamma^\mu D_\mu\chi_L + \sum_{\psi_R} i\bar{\psi}_R\gamma^\mu D_\mu\psi_R - \frac{1}{4}W^a_{\mu\nu}W^{\mu\nu}_a - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} \\ =& \sum_{\chi_L} \bar{\chi}_L\gamma^\mu(i\partial_\mu - g\frac{1}{2}\vec{\tau}\cdot\vec{W}_\mu - g'\frac{1}{2}YB_\mu)\chi_L + \sum_{\psi_R}\bar{\psi}_R\gamma^\mu(i\partial_\mu - g'\frac{Y}{2}B_\mu)\psi_R \\ &- \frac{1}{4}W^a_{\mu\nu}W^{\mu\nu}_a - \frac{1}{4}B_{\mu\nu}B^{\mu\nu},\end{aligned} \tag{2.8}$$

The first two terms of the Lagrangian describe the coupling of fermions to the gauge bosons. The fermion fields ψ decompose in their chirality states, described in eq. 2.1. As already discussed, in the electroweak theory left-handed fermions are arranged in weak isospin doublets, denoted by $\chi_L(= L^i_L, Q^i_L)$, while the right-handed ones are the weak isospin singlets $\psi_R(= l^i_R, u^i_R, d^i_R)$. The last two terms of the Lagrangian are the kinetic energy and self-coupling of the of the $\vec{W}_\mu$ fields as well as the kinetic energy of the B_μ field. The field tensors of the $SU(2)_L$ and $U(1)_Y$ gauge groups are represented as $W^i_{\mu\nu} = \partial_\mu W^i_\nu - \partial_\nu W^i_\mu - g\epsilon^{ijk}W^j_\mu W^k_\nu$, where $i = 1, 2, 3$ and ϵ^{ijk} is the anti-symmetric Levi-Civita symbol, and $B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$, respectively.

Symbol	Multiplets			I	I_3	Y	Q
			Quarks				
Q^i_L	$\begin{pmatrix}u\\d\end{pmatrix}_L$	$\begin{pmatrix}c\\s\end{pmatrix}_L$	$\begin{pmatrix}t\\b\end{pmatrix}_L$	$\frac{1}{2}$	$+\frac{1}{2}$ $-\frac{1}{2}$	$+\frac{1}{3}$	$+\frac{2}{3}$ $-\frac{1}{3}$
u^i_R	u_R	c_R	t_R	0	0	$+\frac{4}{3}$	$+\frac{2}{3}$
d^i_R	d_R	s_R	b_R	0	0	$-\frac{2}{3}$	$-\frac{1}{3}$
			Leptons				
L^i_L	$\begin{pmatrix}\nu_e\\e\end{pmatrix}_L$	$\begin{pmatrix}\nu_\mu\\\mu\end{pmatrix}_L$	$\begin{pmatrix}\nu_\tau\\\tau\end{pmatrix}_L$	$\frac{1}{2}$	$+\frac{1}{2}$ $-\frac{1}{2}$	-1	0 -1
l^i_R	e_R	μ_R	τ_R	0	0	-2	-1

Table 2.1: Quantum numbers of the fundamental fermions of the SM: the weak isospin I and its third component I_3, the weak hypercharge Y and the electric charge Q. L denotes left-handed fermions forming weak isospin doublets Q^i_L for quarks and L^i_L for leptons. R denotes right-handed fermions forming weak isospin singlets u^i_R and d^i_R for up-type and down-type quarks respectively, as well as for charged leptons l^i_R.

The electroweak Lagrangian describes massless gauge bosons and massless fermions. Mass terms, such as $\frac{1}{2}M^2B_\mu B^\mu$ and $-m_f\bar{\psi}\psi$ are not gauge invariant, hence they cannot be included. The requirement of a massless gauge boson is familiar from QED, as explained in Sec. 2.2.1.

A fermion mass term, decomposed in chirality states, would be

$$\begin{aligned} -m_f\bar{\psi}\psi &= -m_f(\bar{\psi}_R + \bar{\psi}_L)(\psi_R + \psi_L) \\ &= -m_f(\bar{\psi}_R\psi_L + \bar{\psi}_L\psi_R), \qquad since\ \ \bar{\psi}_R\psi_R = \bar{\psi}_L\psi_L = 0. \end{aligned} \tag{2.9}$$

Since ψ_L, which is a member of an isospin doublet, and the singlet ψ_R transform differently under $SU(2)_L \times U(1)_Y$ rotations

$$\begin{aligned} \chi_L \rightarrow \chi_L^{'} &= e^{ia(x)I + \beta(x)Y}\chi_L \\ \psi_R \rightarrow \psi_R^{'} &= e^{i\beta(x)Y}\psi_R, \end{aligned} \tag{2.10}$$

a mass term like eq. 2.9 is not gauge invariant and as a result it would manifestly break the gauge symmetry. However, this contradicts the experiments that confirmed the existence of massive fermions and electroweak mediators [52–54], with the latter having masses of $m_{W^\pm} = 80.4$ GeV and $m_Z = 91.2$ GeV [101]. To generate the particle masses in a gauge invariant way, the Higgs mechanism is exploited. That is, by spontaneously breaking the gauge symmetry, gauge bosons and fermions acquire their masses through the interaction with the Higgs field, known as the Yukawa Interaction. As it is shown in Sec. 2.2.5, the spontaneous symmetry breaking makes the neutral bosons W_μ^3 and B_μ mix into the two physical neutral bosons, photon (eq. 2.34) and Z boson (eq. 2.33).

2.2.4 Quantum Chromodynamics: the theory of strong interaction

In contrast to EWT, the SM of particle physics includes the strong interaction in a standalone theory, QCD. The current theoretical picture of the strong interaction finds its origin in the model that identified the spectrum of strongly interacting particles in terms of their elementary constituents, the quarks, proposed by M. Gell-Mann [57,59] and G. Zweig [58] in 1963. About a year later, O.W. Greenberg [55], M.Y. Han and Y. Nambu [56] suggested that quarks carry an additional, unobserved quantum number, the colour charge. Then, H. Fritzsch, M. Gell-Mann, and H. Leutwyler discovered in 1973 [60] that certain phenomena involving the strong interaction could be explained by a non-abelian gauge theory. Eventually, QCD is formulated as a Yang-Mills theory [25], described by $SU(3)_C$[14] gauge group of phase transformations on the quark colour fields, with gluons being its quanta.

In order to obtain the Lagrangian that describes the strong interaction, the invariance under local colour gauge transformations $\psi \rightarrow \psi^{'} = e^{i\alpha(x)T_a}\psi$ of the $SU(3)_C$ group should be imposed. This is achieved by replacing the partial derivative in the Lagrangian 2.2 by a covariant derivative, $D_\mu = \partial_\mu + ig_sT_aG_\mu^a$, where g_s denotes the coupling constant of $SU(3)_C$ gauge symmetry. In this way, eight gauge bosons G_μ^a are introduced, associated to the $SU(3)_C$ generators, which correspond to the eight gluon fields and couple to the quark colour fields. The fields G_μ^a transform as $G_\mu^a \rightarrow G_\mu^a - \frac{1}{g_s}\partial_\mu\alpha_a - f_{abc}\alpha_bG_\mu^c$, in order to ensure the invariance of the Lagrangian. In this transformation, f_{abc} are the real and antisymmetric structure constants of the group, which follow the commutation rule $[T_a, T_b] = if_{abc}T_c$. Consequently, the QCD Lagrangian describes the free motion of quarks described by the Dirac equation, the free

[14]The *special unitary group* $SU(3)$ contains 3×3 unitary matrices, transforming as $U(\alpha_a) = e^{i\alpha_aT_a}$, where α_a are the local phase parameters, and $T_a = \frac{\lambda_a}{2}$ with $a = 1, ..., 8$ are matrices with the generators λ_a being the 3×3 Gell-Mann matrices [67] (linearly independent and traceless). The matrices T_a are hermitian, meaning that the group parameters α_a are real.

propagation of gluons, the quark-gluon interaction, and the self-interaction between gluons,

$$\begin{aligned}\mathcal{L}_{QCD} &= i\bar{q}\gamma^{\mu}D_{\mu}q - m\bar{q}q - \frac{1}{4}G^{a}_{\mu\nu}G^{\mu\nu}_{a} \\ &= \bar{q}(i\gamma^{\mu}\partial_{\mu} - m)q + g_s(\bar{q}\gamma^{\mu}T_a q)G^{a}_{\mu} - \frac{1}{4}G^{a}_{\mu\nu}G^{\mu\nu}_{a}\end{aligned} \tag{2.11}$$

where q is the four-vector Dirac spinor for a free quark with mass m. Additionally, $\bar{q} = q^{\dagger}\gamma^{0}$ is the adjoint spinor, while γ^{μ} represents the four Dirac γ-matrices. Lastly, $G^{a}_{\mu\nu} = \partial_{\mu}G^{a}_{\nu} - \partial_{\nu}G^{a}_{\mu} - g_s f_{abc}G^{b}_{\mu}G^{c}_{\nu}$, $a = 1, 2, ..., 8$ is the field strength tensor for gluons [60], which is invariant under the G^{a}_{μ} transformations. Analogously to QED, mass terms for the gauge bosons would destroy the local gauge invariance of the Lagrangian 2.11. As a result, the gluons are massless, which corresponds to the physical observation.

Unlike the electromagnetic field quantum, the neutral photon, gluons themselves carry colour charge which allows them to directly interact with each other, thus they are described by a non-abelian gauge group. As a consequence of the gluon self-interaction, QCD theory is governed by two special properties, asymptotic freedom and colour confinement.

In contrast to the electromagnetic interaction, which becomes stronger at short distances, the strength of the strong interaction decreases asymptotically as the energy scale[15] of the interaction increases ($Q \to \infty$), while the distance between quarks decreases. Therefore, in the limit of extremely high-energy interactions, quarks and gluons are considered as "free" non-confined particles, interacting with a Coulomb-like force. This property is called *asymptotic freedom* and it was proved for non-abelian gauge theories by D. Gross, F. Wilczek, and H. D. Politzer, in 1973 [62, 63].

On the other side of the spectrum, as the distance between quarks increases and the energy diminishes ($Q \to 0$), the coupling strength becomes stronger. As discussed earlier, the electric field between electrically charged particles decreases rapidly as these particles are separated. In contrast, when a quark gets separated from other quarks, the energy in the strong colour field between them is enough so as to emit QCD radiation in the form of gluons, which subsequently split into many gluons or quark/anti-quark pairs. Hence, colour charged particles cannot be found in isolation, but they are constantly bound into colour-neutral states, the hadrons. This property is called *colour confinement* [61] and explains the short range character of the strong interaction, despite of the massless gauge bosons.

In QFT, the mathematical description of the various physical processes is based on perturbative calculations. This a priori produces infinities for finite order perturbation expansion, which are removed through the renormalisation procedure, as introduced in Sec. 2.2.1. Thus, the coupling constant describing the interaction between two particles is an effective constant which depends on the energy scale (Q^2) and in fact on the renormalisation scale μ_R (defined in Sec. 4.1.1). Due to this dependence, it is called *running coupling constant* [64, 65]. In the electromagnetic interaction the dependence of the electromagnetic coupling $a_{em} \equiv e^2/4\pi$ increases with energy, albeit is very weak. In the strong interaction, however, the dependence is very strong, since gluons carry colour themselves, and therefore can also couple to other gluons. Figure 2.2 shows the strong coupling constant $\alpha_s \equiv g_s^2/4\pi$ and its dependence on the energy scale *Q* of the interaction. The strong and electromagnetic coupling will eventually become comparable at energies of order 10^{15} - 10^{17} GeV [68, 101].

[15] It is the positive quantity for the square of the four-momentum transfer, $Q^2 = -q^2$, where q is the four-momentum transferred by the exchanged particle in an interaction.

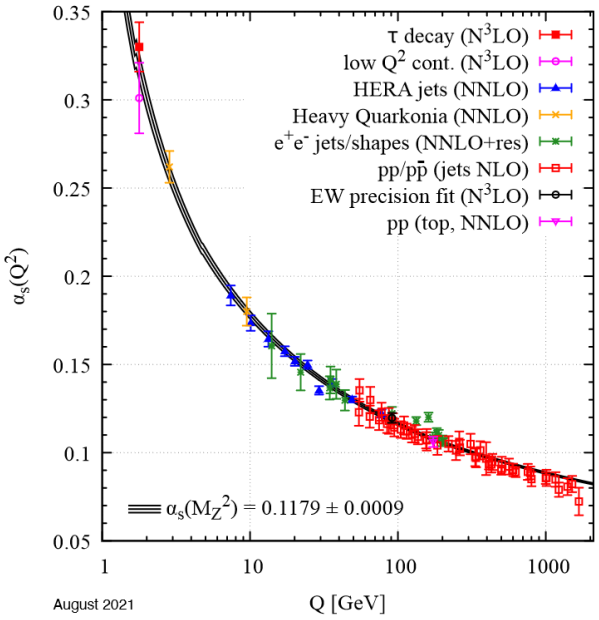


Fig. 2.2: Summary of measurements of the QCD running coupling constant α_s as a function of the energy scale Q [101]. The respective degree of QCD perturbation theory used in the extraction of α_s is indicated in brackets. Also, the world average of α_s measured at the energy scale of the *Z*-boson mass is illustrated.

A first-order perturbation calculation in QCD coupling is expressed by [14, 66]

$$\alpha_s(Q^2) = \frac{\alpha_s(\mu_R^2)}{1 + \beta_0 \alpha_s(\mu_R^2) \log(\frac{Q^2}{\mu_R^2})} \quad \text{with} \quad \beta_0 = \frac{11N_C - 2n_f}{12\pi}, \tag{2.12}$$

where $N_C = 3$ is the number of colours, n_f is the number of quark flavours involved. Since a heavy virtual quark–antiquark pair has a very short lifetime and range, it can be resolved only at very high Q^2. Hence, n_f depends on Q^2, as it refers to quarks with mass$^2 < Q^2$, taking the values $n_f \approx 3 - 6$, and resulting in $\beta_0 > 0$. From eq. 2.12 it is apparent that, at sufficiently low Q^2, the effective coupling becomes large. The Q^2 scale at which this happens is denoted by Λ^2_{QCD}, and eq. 2.12 can be written as

$$\alpha_s(Q^2) = \frac{1}{\beta_0 \log(\frac{Q^2}{\Lambda_{QCD}})}. \tag{2.13}$$

So, the two special properties of QCD, are associated to the running of the strong coupling constant. More precisely, for $Q^2 \gg \Lambda^2_{QCD}$, the effective coupling, $\alpha_s(Q^2)$, is small, vanishing asymptotically, and a perturbative description in terms of quarks and gluons interacting weakly makes sense with the higher order terms have a smaller impact. For $Q^2 \lesssim \Lambda^2_{QCD}$, the interquark coupling increases strongly, so that quarks and gluons arrange themselves into strongly bound states, while the perturbative QCD breaks down as $\alpha_s(Q^2) \to 1$. Thus, Λ_{QCD} can be considered as the boundary between quasi-free quarks and gluons and the bounded hadrons. The value of Λ_{QCD} is a free parameter determined from experiment and its order of magnitude is of a typical hadronic mass, $\Lambda_{QCD} \approx 300$ MeV. Instead of Λ_{QCD}, it is more accurate to quote the value of $\alpha_s(Q^2)$ at a given scale, typically the mass of the *Z* boson, which is illustrated in Fig. 2.2.

In collider experiments, high energetic partons[16] produced in the final state cannot be directly observed in the detector, since they cannot exist freely due to colour confinement. Nonetheless, they form hadrons, which may further decay into a large number of final-state particles, traveling in roughly the same direction as the initial partons. Finally, highly energetic quarks and gluons, are manifested in the detector as bundles of collimated hadrons, called *jets*.

2.2.5 Spontaneous Symmetry Breaking: the Higgs mechanism

The SM is successful in describing the electromagnetic, weak and strong interactions for elementary particles. Nevertheless, no mass term for fermions is allowed in the formalism described up to now, and the electroweak theory is based on four massless gauge bosons. This is in contradiction with the observed masses of the fermions and of the three electroweak mediators ($W^{\pm}, Z$). However, such mass terms in the Lagrangian would violate the $SU(2)_L \times U(1)_Y$ gauge invariance. To solve this inconsistency between the SM theory and the experimental measurements [101], the so-called *Higgs Mechanism* is incorporated to the electroweak theory by S. Weinberg [22] and M.A. Salam [21] in about 1967, which then obtains its modern form.

In fact, the Higgs mechanism, postulated by R. Brout, F.B. Englert [6], P.W. Higgs [5, 7], G.S. Guralnik, C.R. Hagen, and T.W.B. Kibble in 1964 [8], proposes a spontaneous breaking of the local gauge invariant $SU(2)_L \times U(1)_Y$ symmetry in a Yang-Mills theory. The elementary particles acquire their mass through their interaction with the Higgs field. Especially, the massless gauge bosons of the electroweak $SU(2)_L \times U(1)_Y$ mix after the spontaneous symmetry breaking, producing the three massive weak bosons ($W^{\pm}, Z$), while keeping the photon massless. Last but not least, it predicts a scalar particle, the *Higgs boson*, whose mass is a free parameter of the theory.

To attain the spontaneous symmetry breaking an additional weak isospin doublet of complex scalar fields, known as the *Higgs field* ϕ, is introduced

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}. \tag{2.14}$$

Firstly, this field should be an $SU(2)_L \times U(1)_Y$ multiplet, so that the Lagrangian remains gauge invariant. Additionally, in order to generate masses for the three $SU(2)_L$ and the one $U(1)_Y$ gauge bosons, the relevant symmetries should be broken. As a consequence, the Higgs field should be specifically a doublet with $I = \frac{1}{2}$ and $Y = 1$. The electric charges of the upper and lower component of the doublet are chosen to ensure that the $Y = 1$. The same Higgs doublet (summarised in Table 2.2) is sufficient to generate fermion masses as well.

The Higgs field, being a scalar field, follows the Klein-Gordon equation of motion. Its gauge invariant Lagrangian consists of a kinetic term and a potential

$$\mathcal{L}_{Higgs} = (D^\mu \phi)^\dagger (D_\mu \phi) - V(\phi) \tag{2.15}$$

where $D_\mu = \partial_\mu + ig\frac{1}{2}\vec{\tau} \cdot \vec{W}_\mu + ig'\frac{Y}{2}B_\mu$ is the covariant derivative associated to the $SU(2)_L \times U(1)_Y$ symmetry, introduced in Sec. 2.2.3. Renormalisability and $SU(2)_L \times U(1)_Y$ invariance require the Higgs potential to be of the form

$$V(\phi) = \mu^2 \phi^\dagger \phi + \lambda(\phi^\dagger \phi)^2. \tag{2.16}$$

[16] The components of hadrons, i.e. quarks and gluons, are collectively referred to as *partons*.

Symbol	Doublet	I	I_3	Y	Q
			Higgs		
ϕ	$\begin{pmatrix}\phi^+\\ \phi^0\end{pmatrix}$	$\frac{1}{2}$	$+\frac{1}{2}$ $-\frac{1}{2}$	1	1 0

Table 2.2: Quantum numbers of the Higgs boson: the weak isospin I and its third component I_3, the weak hypercharge Y and the electric charge Q.

The first term in eq. 2.16 can be associated with the mass of the field, while the second term stands for the self-interaction of the field. The unitarity requires that the free parameters μ^2 and λ are real.

To determine the ground state (*vacuum*) of the system, the minimum of the potential needs to be found. The extrema of the potential 2.16 are the following

$$\phi_0 = 0 \quad or/and \quad \phi_0 = \pm\frac{1}{\sqrt{2}}\sqrt{-\frac{\mu^2}{\lambda}} \tag{2.17}$$

depending on the values of the parameters λ and μ^2. In order to obtain a finite value for the minima of the potential (vacuum stability) the condition $\lambda > 0$ is imposed, otherwise it becomes unphysical. Furthermore, the parameter of the mass μ can be chosen freely; for $\mu^2 > 0$ the potential assumes a unique minimum at $\phi_0 = 0$, leading to a symmetric ground state under $SU(2) \times U(1)$. The Lagrangian describes a system of four scalar particles ϕ_i, each of mass μ, interacting with four massless gauge bosons ($\vec{W}_\mu$, B_μ). But there is already the analogous electroweak Lagrangian for fermion fields in eq. 2.8, so this case is not interesting at this point. On the contrary, for $\mu^2 < 0$, the shape of the potential is modified as illustrated in Fig. 2.3. In this case, the extremum $\phi_0 = 0$ does not correspond to the minimum energy state and the minima are $\phi_0 = \pm\frac{1}{\sqrt{2}}\sqrt{-\frac{\mu^2}{\lambda}} = \pm\frac{1}{\sqrt{2}}\upsilon$. The minima of the potential ϕ_0 can be identified with the *vacuum expectation value* (v.e.v.) υ of the Higgs field, which is defined as the absolute value of the field at the minimum of the potential.

At first sight, the first term of the potential 2.16 looks like it describes a particle ϕ with an imaginary mass μ. Taking a closer look at the potential in Fig. 2.3, it seems pointless to investigate the particle spectrum using the Lagrangian 2.15. A perturbation series in ϕ would not converge, because it is an expansion around an unstable point $\phi = 0$. For this reason, the field ϕ needs to be studied in the region around its vacuum in a perturbative approach.

There is an infinite number of vacua that satisfy $\phi_0^2 = \frac{1}{2}(\phi_1^2 + \phi_2^2 + \phi_3^2 + \phi_4^2) = -\frac{\mu^2}{2\lambda}$, which follow the symmetry of the Lagrangian. Nevertheless, a specific vacuum should be chosen, in order to perform the perturbative expansion. Thus, without loss of generality, the direction of the vacuum is chosen at $\phi_1 = \phi_2 = \phi_4 = 0$, $\phi_3 = \sqrt{-\frac{\mu^2}{\lambda}}$. Eventually, the vacuum and its expectation value are

$$\phi_0 = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\ \upsilon\end{pmatrix}, \qquad \upsilon = \sqrt{-\frac{\mu^2}{\lambda}}. \tag{2.18}$$

The fact that the neutral component of the Higgs field acquires the vacuum expectation value, ensures the conservation of electric charge. Any choice of the physical vacuum state sponta-

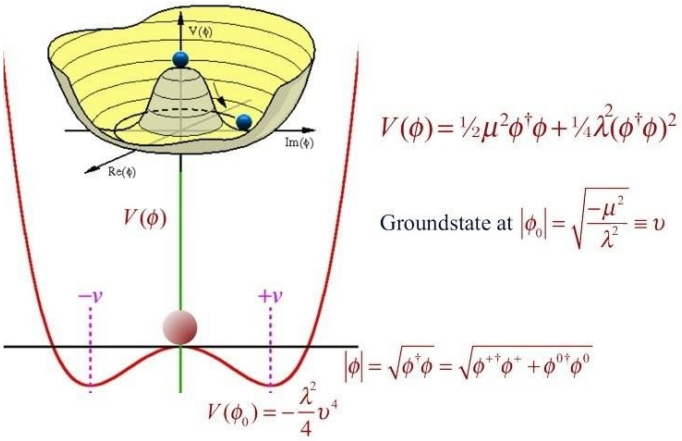


Fig. 2.3: The shape of the Higgs potential $V(\phi)$. The vacuum state is randomly chosen from an infinite number of choices, leading to spontaneous symmetry breaking.

neously breaks the symmetry of the Lagrangian. As a result, this state is not invariant under the $SU(2)_L$ and the $U(1)_Y$ symmetry groups. However it is invariant under $U(1)_{em}$ transformations, since it conserves the electric charge.

When a continuous symmetry of a physical system is broken, massless scalars, referred to as *Goldstone bosons*, appear, according to the *Goldstone theorem* [4]. Especially, when a local gauge symmetry is spontaneously broken, a gauge field can absorb a massless scalar as a longitudinal polarisation component, and as a consequence acquire mass. In addition, if the vacuum ϕ_0 is invariant under some subgroups of the original symmetry $SU(2) \times U(1)$, any gauge bosons associated with that subgroup will remain massless. Therefore, since the chosen vacuum is invariant under $U(1)_{em}$ transformations, the photon remains massless.

In order to determine the particle spectrum, the behaviour of the Lagrangian is studied under small fluctuations around the vacuum, by adding a real field $h(x)$. Provided that the Lagrangian 2.15 is invariant under $SU(2)_L \times U(1)_Y$ transformations, the local gauge invariance should be exploited for a general perturbation. This is achieved by adding the term $e^{i[\vec{\tau}\cdot\vec{\theta}(x)+\xi(x)]/v}$ to the field, where $\vec{\theta}(x)$, $\xi(x)$ denote four real fields (Goldstone bosons). More precisely, the three fields $\theta_1(x), \theta_2(x), \theta_3(x)$ account for the three broken generators of the $SU(2)_L$ symmetry, while $\xi(x)$ stands for the broken generator of the $U(1)_Y$ symmetry. Nevertheless, due to local gauge invariance of the Lagrangian, this phase term is suppressed for infinitesimal transformations, so the field under the expansion is

$$\phi(x) = \frac{e^{i[\vec{\tau}\cdot\vec{\theta}(x)+\xi(x)]/v}}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix} \longrightarrow \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}. \tag{2.19}$$

Eventually, the Lagrangian of the Higgs field becomes

$$\mathcal{L}_{Higgs} = (D^\mu\phi)^\dagger(D_\mu\phi) - \frac{1}{2}(-2\mu^2)h^2(x) - \lambda v h^3(x) - \frac{1}{4}\lambda h^4(x) - \frac{1}{4}\mu^2 v^2. \tag{2.20}$$

Any term involving the $\theta-$ and $\xi-$fields do not appear in the Lagrangian. According to the Goldstone theorem, these additional degrees of freedom (Goldstone bosons) are absorbed by the gauge fields associated to the broken symmetries as longitudinal polarisation components. This is how the masses of the gauge fields $W^1_\mu, W^2_\mu, W^3_\mu, B_\mu$ are generated. In summary, this Lagrangian describes four massive gauge fields and a massive scalar boson $h(x)$, the Higgs

boson. The second term in eq. 2.20 corresponds to the tree-level mass term of scalar field $h(x)$, so the Higgs-boson mass is given by

$$m_h = \sqrt{-2\mu^2} = \sqrt{2\lambda\upsilon^2}. \tag{2.21}$$

Finally, the third and fourth terms in eq. 2.20 describe the Higgs boson self-interactions (three-Higgs-boson and four-Higgs-boson vertex, respectively), while the fifth term is a constant term. The coupling strength of the Higgs-boson self-interactions is expressed as

$$g_{HHH} = \lambda\upsilon = \frac{m_h^2}{2\upsilon} \tag{2.22}$$

$$g_{HHHH} = \frac{\lambda}{4} = \frac{m_h^2}{8\upsilon^2}. \tag{2.23}$$

The Lagrangian 2.20 is just 2.15 written in terms of the ground state of the field, describing exactly the same physical system. Although the Lagrangian retains its original symmetry in ϕ, the perturbations around the minimum are not symmetric in $h(x)$. However, eq. 2.20 is the suitable formulation to generate the physical states of the system. Expressing the field in terms of its vacuum with the addition of the real scalar field $h(x)$, results in the spontaneous breaking of the symmetry in the Lagrangian. This way of allowing the symmetry to hide itself, through which the masses of the relevant scalar and gauge bosons are revealed, is known as the *Higgs mechanism*, according to which the symmetry $SU(2)_L \times U(1)_Y$ breaks down to $U(1)_{em}$.

The masses of the gauge bosons are acquired by their coupling to the Higgs field, with the gauge couplings g of $SU(2)_L$ and g' of $U(1)_Y$. Analysing further the Lagrangian 2.20, the $(D^\mu\phi)^\dagger(D_\mu\phi)$ terms give rise to the masses of the gauge bosons ($\propto \upsilon^2$) as well as the interaction of the gauge bosons with the Higgs boson ($\propto \upsilon h(x)$, $\propto h^2(x)$)

$$\begin{aligned} D_\mu\phi &= [\partial_\mu + ig\frac{1}{2}\vec{\tau}\cdot\vec{W}_\mu + ig'\frac{Y}{2}B_\mu]\frac{1}{\sqrt{2}}\begin{pmatrix}0\\ \upsilon + h(x)\end{pmatrix} \\ (D^\mu\phi)^\dagger &= [\partial^\mu - ig\frac{1}{2}\vec{\tau}\cdot\vec{W}^\mu - ig'\frac{Y}{2}B^\mu]\frac{1}{\sqrt{2}}\begin{pmatrix}0 & \upsilon + h(x)\end{pmatrix}. \end{aligned} \tag{2.24}$$

In order to extract the vector boson masses, the perturbation term $h(x)$ can be ignored for simplicity, so the interesting terms from the $(D^\mu\phi)^\dagger(D_\mu\phi)$ are analysed as follows

$$\begin{aligned} &[-ig\frac{1}{2}\vec{\tau}\cdot\vec{W}^\mu - ig'\frac{Y}{2}B^\mu]\frac{1}{\sqrt{2}}\begin{pmatrix}0 & \upsilon\end{pmatrix}[ig\frac{1}{2}\vec{\tau}\cdot\vec{W}_\mu + ig'\frac{Y}{2}B_\mu]\frac{1}{\sqrt{2}}\begin{pmatrix}0\\ \upsilon\end{pmatrix} \\ =&\frac{-i}{2\sqrt{2}}[g(\tau_1W^{1\mu} + \tau_2W^{2\mu} + \tau_3W^{3\mu}) + g'YB^\mu]\begin{pmatrix}0 & \upsilon\end{pmatrix}\cdot\frac{i}{2\sqrt{2}}[g(\tau_1W_\mu^1 + \tau_2W_\mu^2 + \tau_3W_\mu^3) + g'YB_\mu]\begin{pmatrix}0\\ \upsilon\end{pmatrix} \\ =&\frac{1}{8}\begin{pmatrix}0 & \upsilon\end{pmatrix}\begin{pmatrix}gW^{3\mu} + g'YB^\mu & g(W^{1\mu} - iW^{2\mu})\\ g(W^{1\mu} + iW^{2\mu}) & -gW^{3\mu} + g'YB^\mu\end{pmatrix}\begin{pmatrix}gW_\mu^3 + g'YB_\mu & g(W_\mu^1 - iW_\mu^2)\\ g(W_\mu^1 + iW_\mu^2) & -gW_\mu^3 + g'YB_\mu\end{pmatrix}\begin{pmatrix}0\\ \upsilon\end{pmatrix} \\ =&\frac{\upsilon^2}{8}\begin{pmatrix}g(W^{1\mu} + iW^{2\mu}) & -gW^{3\mu} + g'YB^\mu\end{pmatrix}\begin{pmatrix}g(W_\mu^1 - iW_\mu^2)\\ -gW_\mu^3 + g'YB_\mu\end{pmatrix} \\ =&\frac{\upsilon^2g^2}{8}[(W_\mu^1)^2 + (W_\mu^2)^2] + \frac{\upsilon^2}{8}(g'YB_\mu - gW_\mu^3)(g'YB^\mu - gW^{3\mu}) \\ =&\frac{\upsilon^2g^2}{8}[(W_\mu^1)^2 + (W_\mu^2)^2] + \frac{\upsilon^2}{8}\begin{pmatrix}W^{3\mu} & B^\mu\end{pmatrix}\begin{pmatrix}g^2 & -gg'\\ -gg' & g'^2\end{pmatrix}\begin{pmatrix}W_\mu^3\\ B_\mu\end{pmatrix}, \end{aligned} \tag{2.25}$$

where the mixed terms between the W_μ^3 and B^μ fields are written in a matrix notation.

According to the electroweak theory, the gauge fields $W_\mu^1, W_\mu^2, W_\mu^3, B_\mu$ are combined in order to form the electroweak bosons. More precisely, the W_μ^1 and W_μ^2 fields mix forming the massive charged W^+ and W^- bosons, according to the relation

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2), \tag{2.26}$$

while the massive neutral Z boson and the massless photon are described as orthogonal linear combinations of the neutral W_μ^3 and B_μ fields. The mass eigenstates for the neutral bosons can be acquired by diagonalising the 2×2 matrix in eq. 2.25, and expressed in terms of W_μ^3 and B_μ. The two eigenvalues and the corresponding eigenvectors of this matrix are

$$\begin{aligned} k &= 0 && \frac{1}{\sqrt{g^2+g'^2}}\begin{pmatrix} g' \\ g \end{pmatrix} \\ k &= g^2+g'^2 && \frac{1}{\sqrt{g^2+g'^2}}\begin{pmatrix} g \\ -g' \end{pmatrix} \end{aligned} \tag{2.27}$$

Finally, using eq. 2.26, 2.27 and giving to hypercharge the value $Y = 1$, as it was chosen for the vacuum, the relevant part of the Lagrangian 2.25 results in

$$\begin{aligned} &[-ig\frac{1}{2}\vec{\tau}\cdot\vec{W}^\mu - ig'\frac{Y}{2}B^\mu]\frac{1}{\sqrt{2}}\begin{pmatrix} 0 & \upsilon \end{pmatrix}[ig\frac{1}{2}\vec{\tau}\cdot\vec{W}_\mu + ig'\frac{Y}{2}B_\mu]\frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ \upsilon \end{pmatrix} \\ =&(\frac{1}{2}\upsilon g)^2 W_\mu^+ W^{-\mu} + \frac{1}{2}\frac{\upsilon^2}{4}\left[(g^2+g'^2)\left(\frac{gW_\mu^3 - g'B_\mu}{\sqrt{g^2+g'^2}}\right)^2 + 0\cdot\left(\frac{g'W_\mu^3 + gB_\mu}{\sqrt{g^2+g'^2}}\right)^2\right]. \end{aligned} \tag{2.28}$$

Comparing these terms with general mass terms expected for the physical gauge bosons $M_{W^\pm}^2 W_\mu^{\pm 2} + \frac{1}{2}M_Z^2 Z_\mu^2 + \frac{1}{2}M_A^2 A_\mu^2$ [14], it turns out that the tree-level mass for the charged bosons $W^\pm$ is

$$M_{W^\pm} = \frac{1}{2}\upsilon g, \tag{2.29}$$

while the tree-level masses and the fields of the Z boson and the photon are defined as

$$Z_\mu = \frac{gW_\mu^3 - g'B_\mu}{\sqrt{g^2+g'^2}} \qquad \text{with } M_Z = \frac{1}{2}\upsilon\sqrt{g^2+g'^2} \tag{2.30}$$

$$A_\mu = \frac{g'W_\mu^3 + gB_\mu}{\sqrt{g^2+g'^2}} \qquad \text{with } M_A = 0. \tag{2.31}$$

As expected, the photon remains massless, since the $U(1)_{em}$ symmetry leaves the vacuum invariant. However, this is not a prediction, given that it was required to keep the vacuum neutral when constructing the model.

In terms of the electroweak unification, and since the photon couples to charged fermions, the corresponding coupling constants are connected through

$$g_e = g\sin\theta_W = g'\cos\theta_W, \tag{2.32}$$

where θ_W is the *electroweak mixing* or *Weinberg angle*. This mass eigenstates of the physical neutral bosons can be also expressed as a rotation of the neutral fields W_μ^3 and B_μ through the

electroweak mixing angle

$$Z_\mu = -B_\mu \sin\theta_W + W^3_\mu \cos\theta_W \tag{2.33}$$

$$A_\mu = B_\mu \cos\theta_W + W^3_\mu \sin\theta_W. \tag{2.34}$$

Even though there is no absolute prediction for the mass of the $W\pm$ and Z bosons within the SM, there is a clear prediction on the ratio between the two masses. From eq. 2.29 and 2.30 and exploiting the relation in eq. 2.32, the ratio is

$$\frac{M_W}{M_Z} = \cos\theta_W. \tag{2.35}$$

At the time this theory was formulated, the three massive vector bosons had not been observed. Eventually, they were experimentally confirmed at the Super Proton-Antiproton Synchrotron at CERN in 1983 [52–54]. Moreover, it is worth mentioning that the mass term of the $W^\pm$ boson fixes the v.e.v. of the Higgs potential, although the SM does not fix the value of M_W. The Higgs boson v.e.v. can be extracted from the relations in eq. 2.29 and 2.4 and the precisely measured Fermi constant $G_F = 1.166 \cdot 10^{-5}$ GeV^{-2}

$$\upsilon = \frac{1}{\sqrt{\sqrt{2} G_F}} \approx 246 \text{ GeV}. \tag{2.36}$$

Although the Higgs boson v.e.v. is known, λ is a free parameter therefore, the mass of the Higgs boson, given in eq. 2.21, is not predicted in the SM. The confirmation of the existence of the weak vector bosons gave great weight to the Higgs mechanism, leaving one remaining SM particle to be observed experimentally, the Higgs boson.

Now, returning to the Higgs field Lagrangian (eq. 2.20) and considering also the perturbation term in the Higgs-boson vacuum expectation value (eq. 2.19), the covariant derivative (starting from eq. 2.24 and combining eq. 2.28, 2.30, and 2.31) results in

$$\begin{aligned}(D^\mu\phi)^\dagger (D_\mu \phi) = \; & \underbrace{\frac{1}{2}\partial_\mu h(x) \partial^\mu h(x)}_{\text{Higgs-boson kinematics}} + \underbrace{(\frac{1}{2}\upsilon g)^2 W^+_\mu W^{-\mu}}_{W\text{-boson mass}} + \underbrace{\frac{1}{2}\frac{\upsilon^2}{4}(g^2 + g'^2) Z_\mu Z^\mu}_{Z\text{-boson mass}} \\ & + \underbrace{\frac{1}{8}(2\upsilon h(x) + h^2(x))(2g^2 W^+_\mu W^{-\mu} + (g^2 + g'^2) Z_\mu Z^\mu)}_{\text{Higgs-boson + } W/Z\text{-boson interactions}}.\end{aligned} \tag{2.37}$$

Eventually, also the interactions between gauge bosons and the Higgs boson are introduced, apart from the vector boson mass terms that have been already discussed. The coupling strength of the interactions between gauge bosons and Higgs boson is proportional to the squared mass of the vector boson, m_V

$$g_{HVV} = \frac{2m_V^2}{\upsilon} \tag{2.38}$$

$$g_{HHVV} = \frac{2m_V^2}{\upsilon^2}. \tag{2.39}$$

As already pointed out, the Higgs mechanism is used to accommodate massive gauge bosons in the SM, while keeping the local gauge invariance. In addition to this, fermions also acquire their mass through the spontaneous breaking of the $SU(2)_L \times U(1)_Y$ gauge symmetry. This is achieved by introducing a Yukawa term in the Lagrangian that describes the coupling

of the Higgs to the fermion fields. Furthermore, such a term should be a singlet under $SU(2)_L$ and $U(1)_Y$ so as to remain gauge invariant. For this purpose, the same complex Higgs doublet, which generates $W^{\pm}$ and Z boson masses, can also be used to give mass to fermions.

According to the above, the gauge invariant Yukawa Lagrangian is

$$L_{Yukawa} = y_f[\bar{\chi_L}\phi\psi_R + \chi_L\bar{\phi}\bar{\psi_R}] \tag{2.40}$$

where y_f denotes the coupling of the Higgs boson to a fermion f, known as *Yukawa coupling constant*, and ϕ refers to the Higgs vacuum expansion described in eq. 2.19. However, this vacuum gives mass only to the down-type fermions, i.e. only to the lower components of the isospin doublet. In order to get the mass terms for the up-type fermions, a new complex Higgs doublet, constructed from ϕ as its complex conjugate $\tilde{\phi}(x) = -i\tau_2\phi^*$ with opposite hypercharge $Y = -1$, is exploited

$$\tilde{\phi} = \begin{pmatrix} -\bar{\phi}^0 \\ \phi^- \end{pmatrix} \underset{expansion}{\overset{vacuum}{\longrightarrow}} \frac{1}{\sqrt{2}} \begin{pmatrix} \upsilon + h(x) \\ 0 \end{pmatrix} \tag{2.41}$$

that is also gauge invariant under $SU(2)_L \times U(1)_Y$. Inserting also the new Higgs doublet and writing out the terms of eq. 2.40 in the weak interaction eigenstates basis results in

$$L_{Yukawa} = y_l^{ij}\overline{L'_{Li}}\phi l'_{Rj} + y_d^{ij}\overline{Q'_{Li}}\phi d'_{Rj} + y_u^{ij}\overline{Q'_{Li}}\tilde{\phi}u'_{Rj} + h.c. \tag{2.42}$$

where i, j stands for the three fermion generations and $h.c.$ refers to the hermitian conjugate terms. The Yukawa couplings y_f^{ij} are in fact 3×3 matrices that connect the flavour eigenstates between different generations. In order to express the weak eigenstates as mass eigenstates, i.e. states with proper mass terms, the matrices y_f^{ij} should be diagonalised using unitary transformations. In the quark sector, the rotation to the mass eigenstate basis is done through the CKM matrix, providing a mixing among the quark flavours, as explained in Sec. 2.2.2. By contrast, in the leptonic sector this transformation has no effect due to the absence of right-handed neutrinos, which for this reason cannot acquire mass through the Yukawa coupling. Eventually, the interaction of the Higgs boson with the fermion mass eigenstates are flavour diagonal, thus the Higgs does not mediate flavour changing interactions. Finally, the complete Yukawa Lagrangian is

$$L_{Yukawa} = y_l\overline{L_{Li}}\phi l_{Ri} + y_d\overline{Q_{Li}}\phi d_{Ri} + y_u\overline{Q_{Li}}\tilde{\phi}u_{Ri} + h.c. \tag{2.43}$$

The Yukawa Lagrangian, after substituting the Higgs doublets from eq. 2.19 and eq. 2.41, contains interaction terms ($\propto \frac{y_f}{\sqrt{2}}h(x)$), which couple the Higgs fields to the fermions, as well as mass terms ($\propto \frac{y_f\upsilon}{\sqrt{2}}$) for each fermion. Comparing the latter terms to the mass terms of the Dirac Lagrangian eq. 2.2, it arises that the Yukawa coupling is proportional to the tree-level mass of the fermion

$$g_{Hf\bar{f}} = \frac{y_f}{\sqrt{2}} = \frac{m_f}{\upsilon} \tag{2.44}$$

As a result, the interaction terms are proportional to the mass of the corresponding fermion. However, the fermion masses are not predicted in the SM, since y_f is a free parameter. This relation demonstrates that the Higgs boson couples more strongly to more massive particles. Given that the top quark is the heaviest particle in the SM, the top-quark Yukawa coupling is the largest Higgs coupling to fermions with a value of $y_t = \sqrt{2}\frac{m_{top}}{\upsilon} \approx 1$. This coupling is considered particularly interesting, since it could be sensitive to effects of physics beyond the SM (BSM) [69–73].

2.2.6 Complete Standard Model Lagrangian

The complete SM of elementary particles, which includes the quantum field gauge theories of QCD and the unified EWT, successfully describes all fundamental interactions except gravity. According to the above sections, the SM is a Yang-Mills theory and it is based on local gauge invariance under the $SU(3)_C \times SU(2)_L \times U(1)_Y$ gauge group. These symmetries dictate the internal generators of the SM, which are related to the gauge bosons that mediate the corresponding interactions, i.e. the eight gluons for $SU(3)_C$, $W^{\pm}, Z$ bosons and the photon for $SU(2)_L \times U(1)_Y$. What is more, it is a renormalisable theory, ensuring that all physical observables to be finite.

Especially, the $SU(2)_L \times U(1)_Y$ symmetry in the SM requires the electroweak mediators as well as the fundamental fermions to be massless. Whereas, experimental measurements have shown that so the fermions as the three electroweak gauge bosons $W^{\pm}, Z$ are massive, explicit mass terms in the Lagrangian would violate the gauge invariance. The Brout-Englert-Higgs mechanism introduces a spontaneous breaking of the local gauge $SU(2)_L \times U(1)_Y$ symmetry that resolves the inconsistency among the SM theory and the measurements. As a consequence, the elementary particles acquire their masses through their interactions with the Higgs field. The strength of the interaction of the particle determines its acquired mass which is proportional to the Higgs field. Last but not least, this mechanism predicts the scalar Higgs boson, whose mass is a free parameter of the theory.

Combining eq. 2.8, 2.11, 2.15 and 2.43, the complete SM Lagrangian describing the electromagnetic, weak, and strong interactions can be expressed as

$$
\begin{aligned}
\mathcal{L}_{SM} =& \mathcal{L}_{Gauge} + \mathcal{L}_{Fermions} + \mathcal{L}_{Higgs} + \mathcal{L}_{Yukawa} \\
=& \underbrace{-\frac{1}{4} W^a_{\mu\nu} W_a^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G^a_{\mu\nu} G_a^{\mu\nu}}_{W^{\pm},\, Z,\, \gamma,\, g \text{ kinetic energies and their self-interactions}} + \underbrace{\sum_{\chi_L} \bar{\chi}_L \gamma^\mu (i\partial_\mu - g\frac{1}{2}\vec{\tau}\cdot\vec{W}_\mu - g'\frac{1}{2} Y B_\mu)\chi_L}_{\text{lepton kinetic energies and interactions with } W^{\pm},\, Z,\, \gamma} \\
&+ \underbrace{\sum_{\psi_R} \bar{\psi}_R \gamma^\mu (i\partial_\mu - g'\frac{Y}{2} B_\mu)\psi_R + \sum_{quarks} g_s (\bar{q}\gamma^\mu T_a q) G^a_\mu}_{\text{quark kinetic energies and their interactions with } W^{\pm},\, Z,\, \gamma,\, g} \\
&+ \underbrace{|(i\partial_\mu - g\frac{1}{2}\vec{\tau}\cdot\vec{W}_\mu - g'\frac{Y}{2} B_\mu)\phi|^2 - V(\phi)}_{W^{\pm},\, Z,\, \gamma \text{ and Higgs masses and couplings}} + \underbrace{y_l \overline{L_{Li}}\phi l_{Ri} + y_d \overline{Q_{Li}}\phi d_{Ri} + y_u \overline{Q_{Li}}\tilde{\phi} u_{Ri} + h.c.}_{\text{lepton and quark masses and coupling to Higgs}}.
\end{aligned}
\tag{2.45}
$$

The first group of terms of the analytic form correspond to the kinetic energies of the massless gauge bosons $W^{\pm}$, Z, γ, g and their self-interactions (except for the photon). The next two groups describe the kinetic energies of the massless elementary fermions and their interactions with the respective gauge bosons. The fourth group denotes how the Higgs and the electroweak gauge bosons acquire their mass, while the photon and the gluons remain massless, as well as the coupling of each of the gauge bosons to the Higgs. Finally, the last group of terms represents the fermion masses through their coupling to the Higgs boson.

So far, all SM predictions have been met with remarkable experimental confirmation. The Higgs boson, being determinant to the mechanism of spontaneous symmetry breaking, was finally detected in 2012 at CERN [10, 11], marking the complete verification of the existence of all the SM constituents.

2.3 Feynman diagrams

In principle, for a field theory formulated in terms of a Lagrangian density of quantum fields, the time evolution of arbitrary initial states can be calculated. The interaction between particles is described by the interaction terms in the Lagrangian. However, while the time evolution can be solved exactly in a free theory, a theory containing interactions requires approximate calculations. For scattering reactions these calculations are usually performed in perturbation theory. The basic idea is to start with time-independent states of the free theory, that describe the incoming and outgoing particles and include the interaction as a small perturbation. The time evolution of the free-theory states due to the perturbation can be expressed in terms of the interaction terms in the Lagrangian. The quantitative formulation of elementary particle dynamics amounts to the calculation of decay rates and scattering cross sections (outlined in App. A.1).

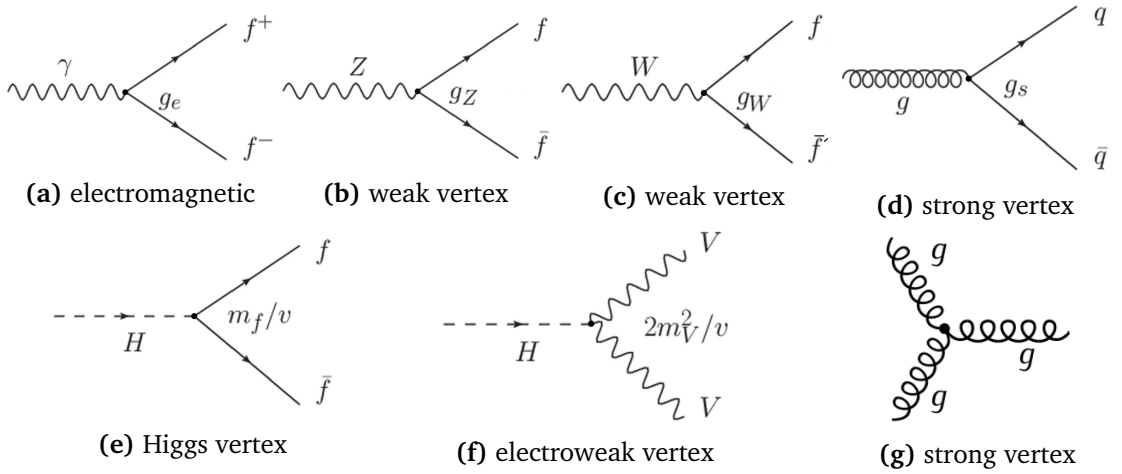
In scattering experiments, the states before and after a scattering (or decay) process, are characterised by the four-momenta of the colliding and scattered (or decaying and produced) particles. All the physics that depends on the dynamics of the process is contained in the Lorentz invariant quantum mechanical *matrix element* $\mathcal{M}$. It can be perturbatively calculated from the interaction part of the Lagrangian with the help of the Dyson series.

The calculation of the matrix element is a major challenge, since its expression can not be defined exactly. The best that can be obtained is a formal expression for $\mathcal{M}$ as a perturbation series in the strength of the interaction, and the evaluation of the first few terms in this series. Nevertheless, there is an elegant procedure to organise and visualise the perturbation series, the *Feynman diagrams* [35, 36]. In short, a diagram is drawn for a specific process, and then it is used to write the mathematical form of the quantum mechanical amplitude for that process to occur.

Feynman diagrams display the flow of particles during a scattering process. The lowest-order term in the perturbation series can be represented by a diagram, called *tree-level* or *leading-order diagram* (LO). Such diagrams are made up of three types of components. There are external lines which depict the incoming and outgoing particles. The points at which three or more particles meet are called vertices, and correspond to the type of interaction that takes place during a scattering process. There are also internal lines between vertices representing the propagation of virtual particles, which serve as force carriers. It is conventional to use straight lines for fermions, helical for gluons, wavy for vector bosons, and dashed for the Higgs boson.

Some of the interaction vertices of the SM are shown in Fig. 2.4. Figures 2.4a - 2.4e corresponds to interactions among fermions and bosons (gauge or scalar). Depending on the time direction, they can represent the creation of a fermion-antifermion pair from a boson (left to right), the annihilation of fermion and anti-fermion into a boson (right to left), the interaction of a fermion with a boson (top to bottom), or the interaction of an anti-fermion with a boson (bottom to top). Additionally, Figs. 2.4f - 2.4g depict the interaction between the Higgs and the heavy gauge bosons, and the gluon self interactions, respectively. They can be interpreted analogous to the fermion-interactions, depending on their orientation. More interactions with up to four participating fields are possible, too.

Each diagram can be translated directly into a contribution to the transition amplitude, following the *Feynman rules* [35]. According to these rules, a short algebraic factor is assigned to each of the aforementioned components of a diagram. The product of these factors gives



(a) electromagnetic **(b)** weak vertex **(c)** weak vertex **(d)** strong vertex

(e) Higgs vertex **(f)** electroweak vertex **(g)** strong vertex

Fig. 2.4: Examples of interaction vertices in the SM: interaction of a) charged fermions *f* and a photon γ, b) same-flavour fermions with the *Z* boson, c) different-flavour fermions with the *W* boson, d) quarks *q* with a gluon *g*, e) fermions with the Higgs boson, f) heavy vector bosons *V* with the Higgs, and g) the gluon self-interaction.

the value of the corresponding term in the perturbation series. For every vertex in particular, a factor proportional to the coupling strength of the participating fields is defined, $g \propto \sqrt{\alpha}$, depending on the type of interaction that is represented by the diagram. An LO diagram normally contains two vertices, hence $\mathcal{M} \propto \sqrt{\alpha} \cdot \sqrt{\alpha} = \alpha$ (or, equivalently $\mathcal{M} \propto g \cdot g = g^2$).

Higher-order diagrams include greater number of vertices and higher powers of α in the matrix element, as well. Then, cross sections and other observables can be expressed as a series in the coupling constant α; the higher the order the more precise is the calculated observable. As long as α is sufficiently small, higher order corrections are small and the perturbation series can be stopped after a few terms. For most of the processes, calculations at next-to-leading order (NLO) or next-to-next-to-leading order (NNLO) are available. Since the strong coupling constant is significantly larger than the coupling constant of the electromagnetic interaction, electromagnetic corrections are of smaller impact at hadron colliders for most processes.

2.4 The Higgs boson

The importance of the Higgs mechanism is evident from Sec. 2.2.5, since it assures the credibility of the SM. High luminosity particle accelerators, reaching center of mass energies up the TeV scale (LEP [74], Tevatron [75], and LHC [108]), provided the environment for intensive testing of the SM. Despite the huge effort, the Higgs boson was the last fundamental component of the SM that remained undiscovered until the beginning of the 21^{st} century. The discovery of a particle compatible with the SM Higgs boson was announced on 4 July 2012 by the ATLAS [10] and CMS [11] experiments at the Large Hadron Collider (LHC), with a mass of about 125 GeV [12], being an important milestone in the history of physics. This discovery confirms the success of the proposed theory about the existence of an associated Higgs field that describes electroweak symmetry breaking as a mechanism to generate massive vector bosons, in addition to fermion masses through Yukawa coupling. Afterwards, an intensive work has taken place so as to measure its production and decay rates and compare with the predictions of the SM, in order to determine the properties of this newly discovered particle.

2.4.1 Production mechanisms of the Higgs boson

The different production mechanisms of a Higgs boson with a mass of 125 GeV in proton-proton (pp) collisions, taking place at the LHC, are described here in decreasing order of production cross-section. Also, a summary of these production cross-sections is depicted in Fig. 2.5 as a function of the centre-of-mass energy $\sqrt{s}$.

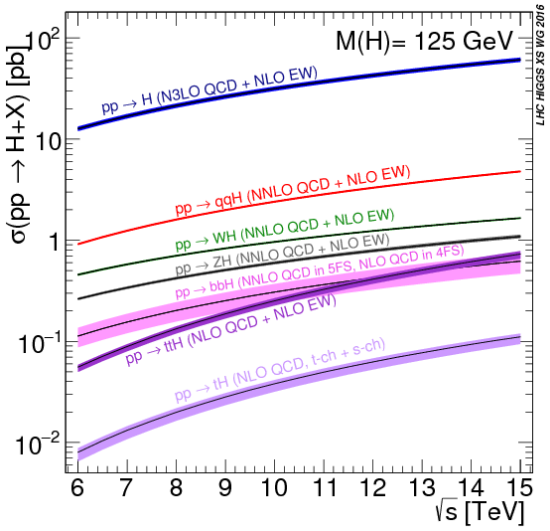


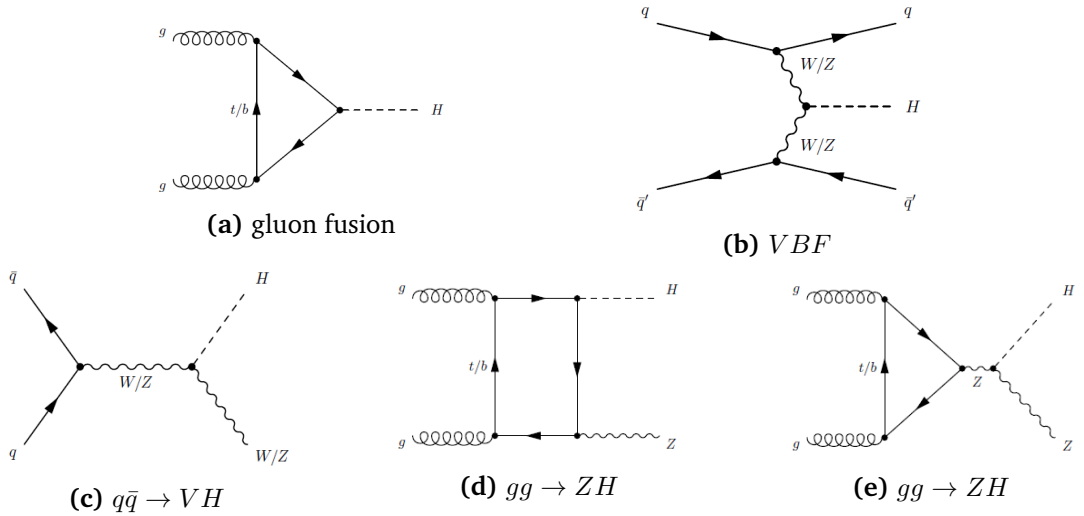
Fig. 2.5: The production cross-sections of the SM Higgs boson with mass 125 GeV, as a function of the center-of-mass energy, $\sqrt{s}$, for pp collisions [106]. The theoretical uncertainties are indicated as bands.

For all the centre-of-mass energies, the dominant SM Higgs production mechanism at the LHC is the *gluon fusion* ($gg \to H$) process, due to the overwhelming presence of gluons in high-energy pp collisions (also explained in Sec. 4.2.1). Especially at $\sqrt{s} = 13$ TeV, this process has a cross-section of $\sigma_{ggH} = 48.58^{+2.22}_{-3.27}$ (theory) ± 1.56 (PDF+α_s) pb [106]. However, the Higgs boson cannot couple directly to gluons, since they are massless. Instead, two merging gluons create a quark loop resulting in the creation of a Higgs boson, as illustrated in the LO diagram in Fig. 2.6a. This production is mainly mediated by a virtual top- or bottom-quark loop, because the matrix element is proportional to the squared Yukawa coupling and subsequently to the squared mass of the corresponding quark, the lighter quark loops are highly suppressed.

The second leading production process, that occurs about an order of magnitude less often than gluon fusion, is the *vector boson fusion* (VBF or $q\bar{q} \to H$) with a cross-section of $\sigma_{VBF} = 3.781^{+0.016}_{-0.012}$ (scale) ± 0.079 (PDF+α_s) pb [106] at $\sqrt{s} = 13$ TeV. In this process, vector bosons V ($W^{\pm}$ or Z), which are radiated from two scattering quarks, merge and create a Higgs boson. In the Feynman diagram in Fig. 2.6b, the presence of a vertex connecting the bosons to the Higgs without being in a loop, is referred to as direct coupling. In VBF the incoming quarks undergo a large momentum transfer, resulting in energetic jets (defined in Sec. 2.2.4) in the forward direction, allowing a direct measurement of the Higgs coupling to vector bosons, with respect to other bosonic decays of the Higgs.

The third most frequent production mode is the *associated production of the Higgs boson with vector bosons* (or *Higgs-strahlung*) (VH or $q\bar{q}, gg \to VH$). For instance, the production associated with a W boson has a cross-section of $\sigma_{WH} = 1.373^{+0.007}_{-0.010}$ (scale) ± 0.026 (PDF+α_s)

pb [106] at $\sqrt{s} = 13$ TeV. LO Feynman diagrams for $q\bar{q}$ and gg initiated process are shown in Figs. 2.6c and 2.6d-2.6e, respectively.



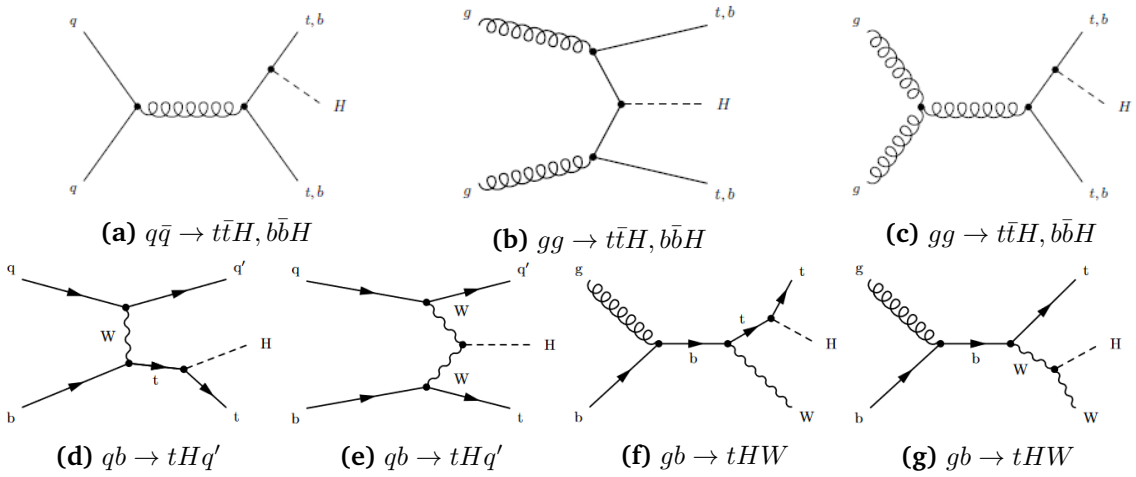
(a) gluon fusion **(b)** VBF **(c)** $q\bar{q} \to VH$ **(d)** $gg \to ZH$ **(e)** $gg \to ZH$

Fig. 2.6: LO Feynman diagrams contributing to the SM Higgs-boson production in pp collisions through bosonic processes.

Then, the Higgs boson production *in association with a heavy quark pair* ($q\bar{q}, gg \to t\bar{t}H$ and $b\bar{b}H$) follows, which is suppressed by two orders of magnitude compared to gluon fusion. It is remarkable that, while the $b\bar{b}H$ cross section is much higher than that of $t\bar{t}H$ at lower center-of-mass energies, the latter increases rapidly, more than any other production mechanism, with the advancing centre-of-mass energy, as shown in Fig. 2.5. Eventually, both processes have a comparable cross-section at $\sqrt{s} = 13$ TeV, which precisely for the $t\bar{t}H$ process is $\sigma_{t\bar{t}H} = 0.507^{+0.029}_{-0.047}$ (scale) ± 0.018 (PDF+α_s) pb [106]. Figures 2.7a-2.7c include Feynman diagrams for $q\bar{q}$ and gg initiated $t\bar{t}H$ and $b\bar{b}H$ processes (analogously to the $t\bar{t}$ production in Sec. 2.5.1), that involve direct coupling of the Higgs boson to the top or bottom quark, respectively.

The rarest considered process is the production of the Higgs boson *in association with a single top quark* (tH), either in the t-channel ($qb \to tHq$) or in the tW-channel ($gb \to tHW$), as depicted in Figs. 2.7d-2.7e and 2.7f-2.7g respectively. The cross-section for the corresponding process at $\sqrt{s} = 13$ TeV is $\sigma_{tHq} = 74.3^{+4.8}_{-11.1}$ (scale) ± 2.7 (PDF+α_s) fb [106] and $\sigma_{tHW} = 15.2^{+0.7}_{-1.0}$ (scale) ± 1.0 (PDF+α_s) fb [106]. The ($q\bar{q} \to tHb$) s-channel production is negligible due to its low cross-section of 2.9 fb [106]. In both modes the Higgs boson is mainly radiated from the top quark, but it can be also radiated from the W-boson propagator. This leads to two Feynman diagrams with the same final state for each of the production modes, which thus cannot be distinguished. Therefore, the coupling to top quarks cannot be directly accessed. Moreover, the interference of these two diagrams leads to a sensitivity of the relative sign between the Higgs-boson coupling to top quarks and to vector bosons, which is positive causing a destructive interference in SM. This effect causes the tH production cross section to be so small. In BSM theories the aforementioned sign can be negative though, resulting in constructive interference, that could significantly enhance the production cross section [76].

Among the Higgs boson couplings to fermions, the top-quark Yukawa coupling is of particular interest. It is not only the largest, but also remarkably close to unity. For the measurement of the top-quark Yukawa coupling, $t\bar{t}H$ is the preferred production process since it has higher cross-section compared to the tH process.



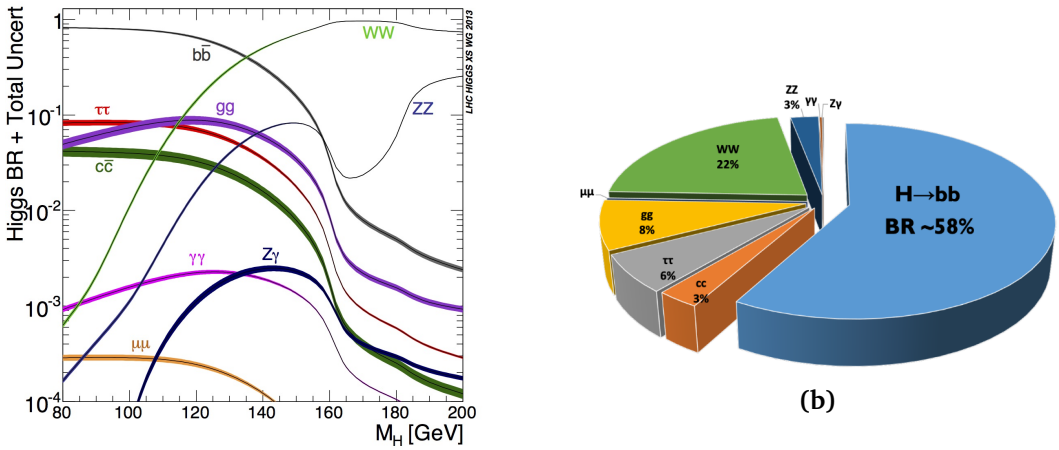
(a) $q\bar{q} \to t\bar{t}H, b\bar{b}H$ **(b)** $gg \to t\bar{t}H, b\bar{b}H$ **(c)** $gg \to t\bar{t}H, b\bar{b}H$

(d) $qb \to tHq'$ **(e)** $qb \to tHq'$ **(f)** $gb \to tHW$ **(g)** $gb \to tHW$

Fig. 2.7: LO Feynman diagrams for the SM Higgs-boson production in pp collisions associated with a pair of top or bottom quarks (upper), or a single top quark (bottom). In the latter, the Higgs boson couples to the top quark d),f) or to the W boson e),g).

2.4.2 Higgs-boson decays

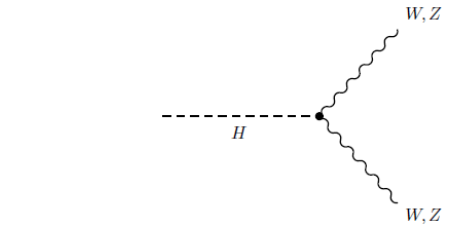
The SM Higgs boson has a lifetime of 10^{-22} s, thus it can be only indirectly observed from its decay products. In general, the branching ratios of the Higgs boson decay modes are dependent on the Higgs mass, as it is depicted in Fig. 2.8a.



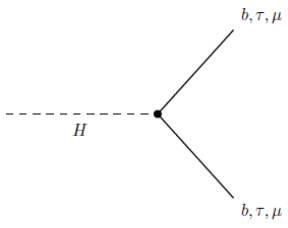
(a) **(b)**

Fig. 2.8: The predicted branching ratios of the various decay modes for the SM Higgs boson a) as a function of its mass [106] and b) at the observed value of $m_H = 125$ GeV. In figure a) the theoretical uncertainties are indicated as bands.

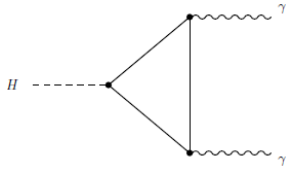
In principle, the Higgs boson can decay into any pair of massive SM particles. Given that the Higgs coupling to particles is proportional to their mass (see eq. 2.44 and eq. 2.39), heavier daughter particles, that are kinematically allowed, are favoured from the Higgs boson decays. Consequently, the Higgs boson decays, through tree-level processes, mostly to pairs of massive electroweak gauge bosons ($W^{\pm}, Z$) (Fig. 2.9a) and into pairs of heavy quarks and leptons (b, τ) (Fig. 2.9b), but also to lighter fermions (c, μ) with smaller rates. Its decay to a top-quark



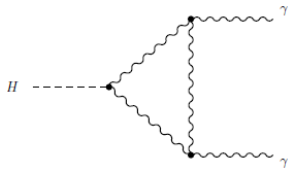
(a) $H \to W^+W^-, ZZ^*$



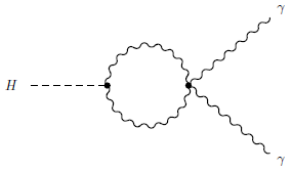
(b) $H \to b\bar{b}, \tau^+\tau^-, c\bar{c}, \mu^+\mu^-$



(c) $H \to \gamma\gamma$ with t, b loop



(d) $H \to \gamma\gamma$ with $W^\pm$ loop



(e) $H \to \gamma\gamma$ with $W^\pm$ loop

Fig. 2.9: LO Feynman diagrams for SM Higgs-boson decays via tree-level (upper) or virtual loop (bottom) processes.

pair is suppressed, since the top quark is much heavier than the Higgs boson, thus none of the quarks can be produced on-shell. As already mentioned, the Higgs boson cannot couple to massless particles. However, decays into a pair of massless gauge bosons (g, γ) can be realised, induced by heavy particle virtual loops as shown in Figs. 2.9c-2.9e. Such processes result in an indirect coupling of the massless vector bosons to the Higgs boson.

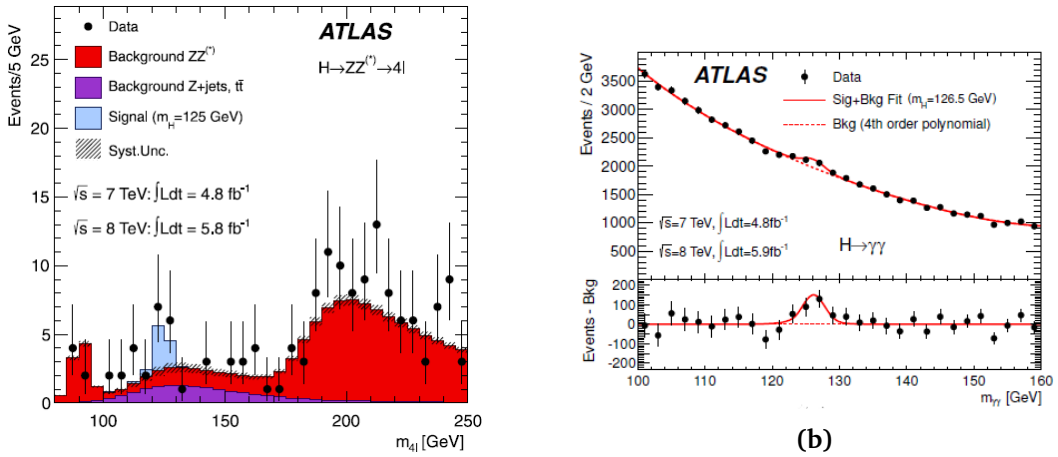
Finally, Fig. 2.8b illustrates the Higgs decay branching ratios at the measured mass of 125 GeV. Among all Higgs-boson decays, the $H \to b\bar{b}$ mode is the dominant one with a branching ratio of about 58%. Although both the $W^\pm$ and Z bosons have much larger masses, their production is suppressed since one of them has to be produced off-shell.

2.4.3 Discovery and properties of the Higgs boson

In March 2010, the LHC started to produce pp collisions at the never achieved before centre-of-mass energy of $\sqrt{s} = 7$ TeV with a total integrated luminosity (defined in Sec. 3.1.1) $\mathcal{L} = 5.5$ fb^{-1} until 2011, while during 2012 it reached $\sqrt{s} = 8$ TeV and $\mathcal{L} = 22.8$ fb^{-1}. During this period, referred to as LHC *Run 1*, Higgs boson searches were performed in all its decay modes. Nevertheless, the initial searches were focused on the bosonic decay modes of the Higgs boson, as they provide better signal sensitivity compared to fermionic final states.

More precisely, the $H \to \gamma\gamma$ decay channel is very sensitive, since it produces a very clear signature with two isolated photons of high transverse momentum. Also, the $H \to ZZ^*$ decay although it has a small branching ratio ($BR \sim 2.6\%$), it gives a clear signature when requiring a decay into charged leptons. Figure 2.10 illustrates the comparison between data recorded by the ATLAS detector as well as the SM predictions for the $H \to ZZ^* \to 4l$ (2.10a) and the $H \to \gamma\gamma$ (2.10b) channels [10]. These two famous bumps at ~ 125 GeV indicate the presence of the new boson compatible with the SM. In parallel, searches in the same channels conducted by the CMS collaboration were equally successful [11]. These two independent observations of the same Higgs boson, with a 5σ significance, demonstrate the validity of each single discovery.

After the discovery of the Higgs boson, a quest to study it in as many production and decay



(a) **(b)**

Fig. 2.10: The distribution of the a) four-lepton invariant mass, m_{4l}, and b) di-photon invariant mass, $m_{\gamma\gamma}$, for the selected candidates in $H \to ZZ^* \to 4l$ and $H \to \gamma\gamma$ events, respectively [10]. They are compared to the background expectation, for the combination of the $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV data.

modes as possible started, in order to determine experimentally all the properties of this new boson. By the end of Run 1, the ggH production as well as the $H \to ZZ^*$, $H \to W^+W^-$ and $H \to \gamma\gamma$ decay modes had been observed by both the ATLAS and CMS collaborations [12]. The combination of the analyses performed within the ATLAS and CMS collaborations in 2016 [13] improved our knowledge of the Higgs boson production and decay modes. Furthermore, it allowed the observation of the VBF production mode [79] and the first fermionic decay $H \to \tau^+\tau^-$ [77, 78]. Afterwards, the $t\bar{t}H$ production mode was observed, in the middle of 2018 [102, 103]. Soon afterwards, the observation of the dominant Higgs-boson decay mode $H \to b\bar{b}$ took place through the VH production mode [104, 105]. This production mode allows to easily access to the $H \to b\bar{b}$ process, since it benefits from the leptonic decays of the additional vector bosons to eliminate the multi-jet background (defined in Sec. 4.5.7).

According to the SM, the Higgs boson is a neutral particle whose mass is a free parameter that needs to be determined experimentally. Measuring precisely the mass of the Higgs boson is necessary, in order to determine its branching ratios and the cross-section of its production modes at the LHC. This measurement was done in the context of the $H \to ZZ^* \to 4l$ and $H \to \gamma\gamma$ decays where, as depicted in Fig. 2.10, the Higgs-boson mass peak is narrow and gives a high experimental resolution of a few GeV. Also, a combined measurement of the Higgs-boson mass in the ATLAS and CMS collaborations with the full Run 1 dataset was performed, resulting in the measured Higgs-boson mass $m_H = 125.09 \pm 0.21$ (stat.) ± 0.11 (syst.) GeV [12].

In the SM, the Higgs boson is introduced as a spin-0 and CP-even particle ($J^P = 0^+$), but other models can generate other types of Higgs bosons. To discriminate among these representations, precise measurements of the Higgs-boson spin and parity were conducted, based on the kinematic properties of the $H \to \gamma\gamma$, $H \to ZZ^* \to 4l$ and $H \to W^+W^- \to l\nu l\nu$ decays, which differ depending on J^P. Eventually, the spin-1 and 2 hypotheses were rejected at confidence levels higher than 99.7% and 99.9%, respectively, using the "8 TeV ATLAS data" [80], while similar studies were performed by the CMS collaboration [81]. These results are an evidence of the spin-0 nature of the Higgs boson and a preference for the even parity, which

are compatible with the SM prediction.

2.5 The top quark

The third generation of quarks was postulated by M. Kobayashi and T. Maskawa in 1973 [46], to explain the observed CP-violation in Kaon decays. A few years later the bottom quark was discovered, but it took many more years for the top quark to be discovered. Due to its large mass, compared to the other fermions in the SM, it can only be produced in high energy processes. Finally, the discovery of the top quark was announced by the CDF [24] and D$\emptyset$ [23] experiments at Tevatron in 1995, which was the first particle accelerator reaching energies capable of producing anti-/top-quark pairs.

The top quark is the heaviest particle in the SM with a mass of 172.76 ± 0.30 GeV [101]. Due to this, other unique properties arise, such as the very short lifetime of about 5.0×10^{-25} s, and consequently the large value of its decay width ($1.42^{+0.19}_{-0.15}$ GeV) [101]. This is about a twentieth of the timescale for strong interactions, connoting that the top quark decays before any hadronisation effect can take place. As a result, the top quark cannot form hadrons, as all other quarks do, thus it is the only quark that has been directly observed from its decay products. This also gives physicists the unique opportunity to detect its properties from its decay products undiluted by non-perturbative effects.

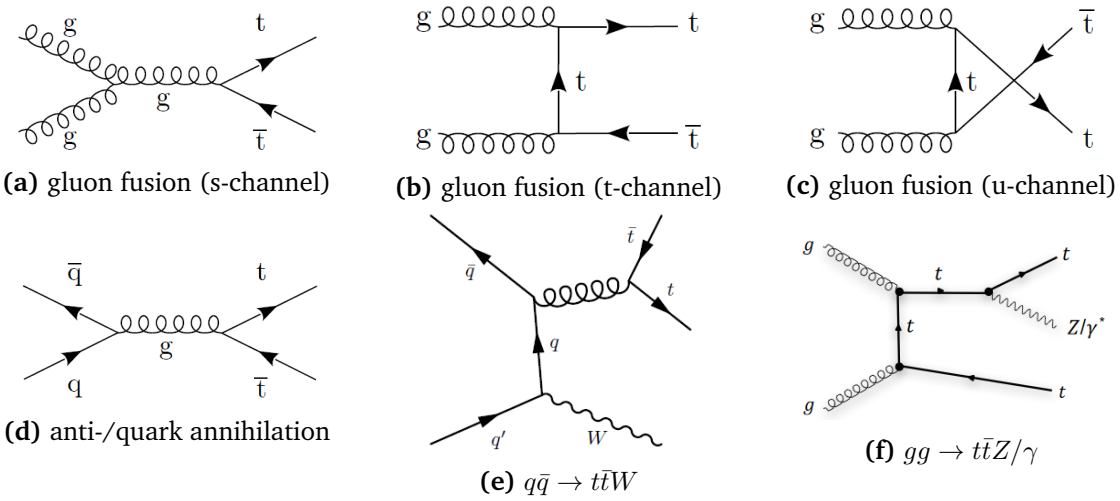
An important consequence of the top quark being so massive is the strong coupling to the Higgs boson y_t, which is very close to 1, as already discussed in the end of Sec. 2.2.3. This might be a coincidence but could also have a deeper reason, while any experimental deviations could be a hint for new physics BSM [69–73]. In the following, the top-quark pair production and decay modes are outlined, since they are a main constituent of the analysis presented in this thesis. In particular, the former constitute the overwhelming background of the analysis. The top-quark pair decays contribute to the background as well, but they are also part of the signal process, since the Higgs boson is produced in association with a top-quark pair ($t\bar{t}H$).

2.5.1 Top-quark production

The creation of a top quark requires large amounts of energy, due to its large mass. Nowadays, the LHC is the only accelerator that generates beams of sufficient energy so as to produce real top quarks, at a center-of-mass energy of $\sqrt{s} = 13$ TeV.

The dominant mechanism to produce top quarks in hadron collider experiments is the anti-/top-quark pair production ($t\bar{t}$), dominated by the strong interaction. At LO perturbation theory anti-/top-quark pairs are produced mainly through the *gluon fusion* ($gg \to t\bar{t}$) and *anti-/quark annihilation* ($q\bar{q} \to t\bar{t}$) processes, as denoted by the Feynman diagrams in Figs. 2.11a-2.11d. Including NLO corrections also allows for quark-gluon initial states. The dominant $t\bar{t}$ production mode at the Tevatron ($p\bar{p}$ collider at 1.96 TeV) was the $q\bar{q}$ annihilation ($\sim 85\%$ of the $t\bar{t}$ cross-section) in which the collisions happen mainly between the valence quarks from the proton and the anti-proton. By contrast, at the LHC (at 13 TeV) about 90% of the $t\bar{t}$ pairs are produced via gluon fusion (see also Sec. 4.2.1), while most of the rest are produced through $q\bar{q}$ annihilation. However, there are also production processes associated with vector bosons ($q\bar{q} \to t\bar{t}W$, $gg \to t\bar{t}Z/\gamma$), as depicted in Figs. 2.11e-2.11f, but they are much more rare.

Figure 2.12 illustrates both the theoretical prediction and the measurements of the $t\bar{t}$



(a) gluon fusion (s-channel) **(b)** gluon fusion (t-channel) **(c)** gluon fusion (u-channel)

(d) anti-/quark annihilation **(e)** $q\bar{q} \to t\bar{t}W$ **(f)** $gg \to t\bar{t}Z/\gamma$

Fig. 2.11: LO Feynman diagrams of $t\bar{t}$ pair production through strong interaction.

production cross-section as a function of the center-of-mass energy $\sqrt{s}$. The theoretical computation is made at NNLO in perturbative QCD. For a top quark mass $m_{top} = 172.5$ GeV, the cross-section is predicted to be $\sigma_{t\bar{t}}(8\text{ TeV}) = 252.89^{+6.39}_{-8.64}$ (scale) ± 11.67 (PDF+α_s) pb, and $\sigma_{t\bar{t}}(13\text{ TeV}) = 831.76^{+19.77}_{-29.20}$ (scale) ± 35.06 (PDF+α_s) pb [101], in which the uncertainties come from variations of the renormalisation and factorisation QCD scales, the parton distribution functions and the strong coupling constant. Also, the measured cross section for the $t\bar{t}$ pair production in association with a W or Z boson is $\sigma_{t\bar{t}W}(13\text{ TeV}) = 0.87 \pm 0.13$ (stat.) ± 0.14 (syst.) pb and $\sigma_{t\bar{t}Z}(13\text{ TeV}) = 0.99 \pm 0.05$ (stat.) ± 0.08 (syst.) pb [101].

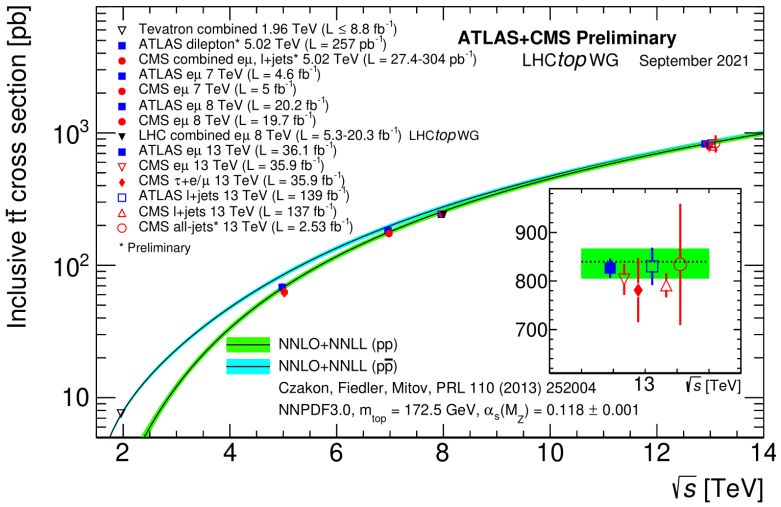


Fig. 2.12: Summary of LHC and Tevatron measurements of the $t\bar{t}$ pair production cross-section as a function of $\sqrt{s}$ [82] compared to the NNLO QCD calculation complemented with NNLL resummation [262]. The theory band includes the uncertainties due to renormalisation and factorisation scales, parton density functions and the strong coupling. The measurements and the theory calculation are quoted at $m_{top} = 172.5$ GeV.

A distinctly different mechanism to produce top quarks is the single top quark process via the weak interaction, with much smaller cross-section though. Single top production proceeds

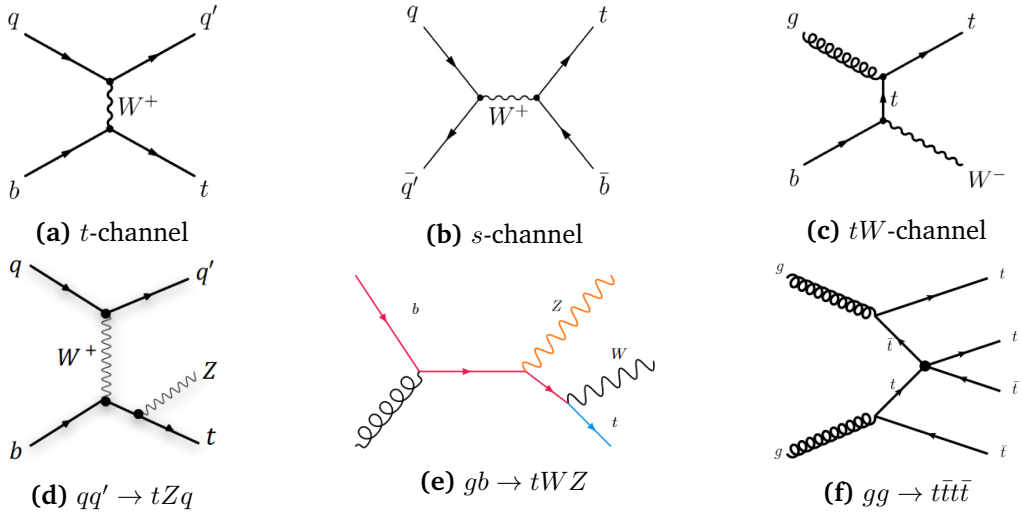


Fig. 2.13: LO Feynman diagrams of a)-d) single top quark production via weak and e) strong interaction, and f) four top-quarks production process.

through several separate sub-processes at the LHC, as illustrated in Figs. 2.13a-2.13e. In the following, the considered predicted cross-sections refer to $\sqrt{s} = 13$ TeV and $m_{top} = 172.5$ GeV. Among the single-top processes, the most probable is the t-channel ($qb \to q't$), with a cross-section of $\sigma_{tq}^{t+\bar{t}} = 216.99^{+6.62}_{-4.64}$ (scale) $\pm$ 6.16 (PDF+α_s) pb [101], in which a bottom quark (mostly arising from a gluon decay in a $b\bar{b}$ pair) transforms to a top quark by exchanging a virtual W boson. There is the s-channel process ($q\bar{q} \to t\bar{b}$) too, where the intermediate virtual W boson decays into a top and anti-bottom quarks, but it is quite rare having $\sigma_{tb}^{t+\bar{t}} = 10.32^{+0.29}_{-0.24}$ (scale) $\pm$ 0.27 (PDF+α_s) pb [101]. Moreover, a single top quark can be produced in association with a real W boson, referred to as tW-channel ($bg \to W^-t$) with $\sigma_{tW}^{t+\bar{t}} = 71.7 \pm 1.8$ (scale) $\pm$ 3.4 (PDF+α_s) pb [101], requiring an initial-state bottom quark. A relevant subprocess is also the single top-quark production in association with a W and a Z boson ($gb \to tWZ$). Additionally, there is the extremely rare single top-quark production in association with a Z boson ($q\bar{q} \to tZq$, Fig. 2.13d), with the measured cross section being $\sigma_{tZq} = 97 \pm 13$ (stat.) $\pm$7 (syst.) fb [101].

Finally, another, overly rare, process to produce top quarks is the four top-quarks production $t\bar{t}t\bar{t}$ (Fig. 2.13f), which is yet to be observed. Its measured cross section is $\sigma_{t\bar{t}t\bar{t}} = 24 \pm 4$ (stat.) $^{+5}_{-4}$ (syst.) fb [101].

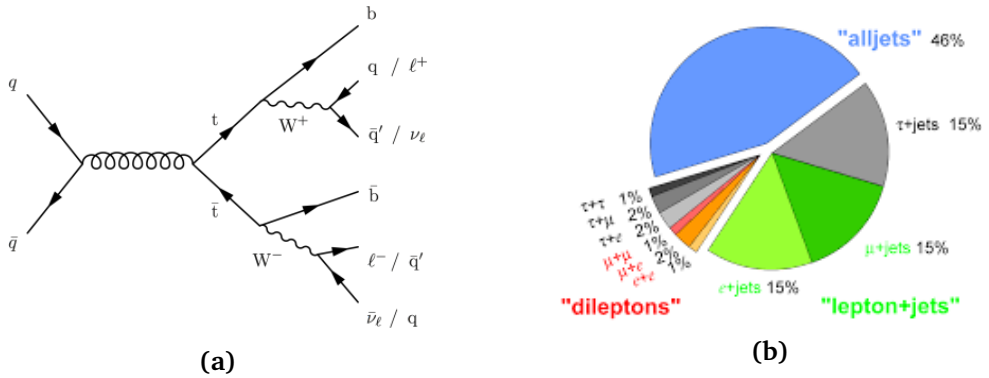
2.5.2 Top-quark decay

As already discussed, the top quark is extremely short-lived because of its enormous mass, thus it does not form hadrons before decaying. Although, it interacts through both the strong and electroweak forces, it can decay only through the weak force. It decays almost exclusively into a W boson and a bottom quark, and more rarely to a W boson and a strange or a down quark. The decay rates for the different quark flavours are proportional to the square of the CKM matrix elements $|V_{tq}|$ (see eq. 2.6), where q can be any down-type quark (d, s, b). More precisely, the probability for a decay to a bottom quark $|V_{tb}| = 0.999$ is almost 100%, with a large decay width $\Gamma(t \to Wb) = 1.35$ GeV (for $m_{top} = 173.3$. GeV) [101].

Furthermore, while the b-quark hadronises forming a parton shower, the W boson decays

further through a hadronic or a leptonic process. In the hadronic decay channel, the W boson decays into a pair of light quarks $q\bar{q}'$ ($u\bar{d}$ or $c\bar{s}$, $BR \sim 68\%$). The decay of a W boson to a top quark and another down-type quark is not possible, since the mass of the latter is larger than the mass of the former. In the leptonic decay channel, the W boson decays into a charged lepton l and the corresponding neutrino ν_l ($W \to l\nu_l$, $BR \sim 32\%$). The analysis presented in this thesis focuses on the $t\bar{t}H$ production for the signal process, and as a result on the decays of the $t\bar{t}$ system. The final state of a $t\bar{t}$ decay is determined upon the number and flavour of the decay products of the two W bosons present in the event. The different $t\bar{t}$ signatures are depicted in the Feynman diagram 2.14a and their branching fractions in Fig. 2.14b.

The most probable channel is the *all-hadronic* or *all-jets* ($t\bar{t} \to W^+bW^-\bar{b} \to bq\bar{q}'\bar{b}q''\bar{q}'''$), which corresponds to a branching ratio of $BR \sim 46\%$. The W boson from each top-quark decays hadronically into an anti-/quark pair, which then hadronise and each forms a parton shower. So in the final state, six jets are expected, two of which are b-jets (jets originating from b-hadrons) resulting from the top-quark decay.



(a) **(b)**

Fig. 2.14: a) LO Feynman diagram of $t\bar{t}$ decay channels. b) Pie chart illustrating the branching ratios of a $t\bar{t}$ pair.

Then, the *lepton+jets* or *single-lepton* channel ($t\bar{t} \to W^+bW^-\bar{b} \to bq\bar{q}'\bar{b}l^-\bar{\nu}_l$ or $bl^+\nu_l\bar{b}q''\bar{q}'''$) is almost equally probable as the all-hadronic one, corresponding to $BR \sim 44\%$. Here, one of the two W bosons decays hadronically into an anti-/quark pair, forming a parton shower, while the other decays leptonically to a charged lepton and its neutrino. This final state contains four jets, two of them being b-jets, the charged lepton and the neutrino resulting in missing transverse energy (defined in Sec. 5.5).

The least probable is the *di-leptonic* or *dilepton* channel ($t\bar{t} \to W^+bW^-\bar{b} \to bl^+\nu_l\bar{b}l'^-\bar{\nu}_{l'}$), with $BR \sim 10\%$. The W boson of each top decay decays leptonically into a charged lepton and its neutrino. The final state involves two b-jets, two opposite sign charged leptons, and two neutrinos resulting in missing transverse energy.

The quarks in the above final states evolve into jets of hadrons. In addition to the quarks resulting from the top-quark decays, extra QCD radiation (quarks and gluons) from the coloured particles in the event can lead to extra jets. Moreover, although l in the above processes refers to electrons, muons, or taus, the following analysis distinguishes the e- and μ- from the τ-channel, which is more difficult to reconstruct. More precisely, the τ lepton decays leptonically ($\tau \to l\nu_l\nu_\tau$, $l = e, \mu$) or hadronically ($\tau \to hadrons\ \nu_\tau$) and it is usually treated separately. The leptonic decay of τ leptons results in the same signature as described above and is experimentally included into the dilepton and single-lepton channels. Therefore

in what follows, l refers to e or μ originating either directly from the W-boson or through a τ-lepton decay, unless otherwise stated.

2.6 Direct measurement of the top-quark Yukawa coupling

It has been already highlighted that the coupling of the Higgs boson to the top quark, namely the top-quark Yukawa coupling, y_t, is much stronger than that of the other quarks in the SM, due to its large measured mass. In addition, y_t is expected to be close to unity, thus it is argued to be a quantity that might give insight into the scale of new physics. However, direct access to the top-quark Yukawa coupling is not possible in measurements of the Higgs boson decays, as the top quark is too heavy to allow the Higgs boson to decay into a pair of top quarks.

When performing Higgs-boson measurements, the main challenge is to distinguish the Higgs-boson signal process from other processes with similar experimental signatures, referred to as background. Different Higgs-boson decay modes have different background compositions and experimental challenges. Therefore, y_t was, at first, experimentally accessible by measuring the cross-section of the gluon fusion (ggH) production process or the $H \to \gamma\gamma$ decay [13], as they provide a clear signature. Nevertheless, the Higgs-boson coupling to the top quark arises indirectly from a top-quark loop. This case requires the assumption that no BSM particle couples to the Higgs boson, contributing with additional induced loops, in order to measure y_t. If no assumption is made about the particle content of such loop contributions, they may also contain non-SM particles, which could compensate deviations in the top-quark Yukawa coupling. For this reason, a direct measurement of y_t is needed.

As discussed in Sec. 2.4.1, the most favourable process that allows to probe directly the Yukawa coupling y_t is the production of the SM Higgs boson in association with a top-quark pair ($t\bar{t}H$). This coupling is of substantial importance to assess the SM behaviour of the observed Higgs boson. Furthermore, a comparison of the direct measurement of this coupling to its indirect measurement through ggH allows to characterise the content of the loop in ggH and reveal potential BSM contributions [69–73]. Although this production mode contributes only around 1% of the total Higgs boson production cross-section at the LHC [106], the top quarks in the final state offer a distinctive signature, providing access to many Higgs boson decay modes.

The $t\bar{t}H$ production mode is split into three main analyses depending on the Higgs boson decay mode, i.e. $H \to b\bar{b}$, $H \to$ multi-leptons ($H \to WW$, ZZ, $\tau\tau$) and $H \to \gamma\gamma$. Currently, the most sensitive channel is $t\bar{t}H(H \to \gamma\gamma)$ [93, 94], where the two photons in the final state allow to measure the invariant mass of the Higgs boson with a good resolution. This analysis is limited by low statistics, though. Among the multi-lepton channels [90–92], the $t\bar{t}H(H \to ZZ \to 4l)$ channel is the most sensitive, but it has a low branching ratio and is currently statistically limited. The biggest disadvantage of the remaining multi-lepton channels is that not all decay products are detected, making it difficult to reconstruct the event kinematics. Then, the decay into two b-quarks ($H \to b\bar{b}$) is the most probable, with a SM $BR \sim 58\%$, and thus with the highest statistics. Furthermore, in the $H \to b\bar{b}$ decay mode the reconstruction of the Higgs boson kinematics is possible, which allows to extract additional information about the structure of the interaction between the top-quark and the Higgs-boson [83–86]. Additionally, this decay mode is sensitive to the b-quark's Yukawa coupling, the second largest in the SM. Although the branching ratio of the $H \to b\bar{b}$ decay mode is large, its measurement is challenging as there are background processes with the same final-state particles, namely a

$t\bar{t}$ process with additional two b-quarks. Also, it is more difficult to identify b-jets compared to photons and the easily detectable charged leptons. Last but not least, the modelling of events with additional heavy-flavour quarks (b- or c-quarks) in the final state generally has large systematic uncertainties, which decrease the sensitivity of the measurement.

The analysis presented in this thesis aims at selecting events with a Higgs boson produced in association with a pair of top quarks, which subsequently decays into a pair of b-quarks, $t\bar{t}H(H \to b\bar{b})$. Such events are further separated into three channels based on the decay of the top-quark pair, i.e. the all-hadronic, the dilepton and the single-lepton channel, as described in Sec. 2.5.2. Although the all-hadronic channel has the highest branching ratio, the absence of any lepton and the presence of many jets makes it difficult to separate from the large background stemming from multi-jet processes (where a jet is misidentified as electron or muon - defined in Sec. 4.5.7), and assign the decay products to the two top-quark decays. Such background process is difficult to model with the generated samples, thus data-driven methods are used to estimate it. The resulting model still has significant uncertainties, though. This channel is the subject of a standalone analysis [96] and will not be presented in this thesis. Then, the dilepton channel provides the cleanest topology with a very high separation from multi-jet background. However this channel suffers from a low branching ratio of 10%. What is more, it involves two neutrinos, which have to be reconstructed by finding a reasonable way to split the missing transverse energy (Sec. 5.5) information. Last but not least, the single-lepton channel provides a compromise between high branching ratio and relatively clean topology with reduced multi-jet background and combinatorial ambiguities, when trying to match the final-state jets to their original particles. In particular, it offers a branching ratio of 44% close to the all-hadronic channel ($BR \sim 46\%$), while it includes one lepton allowing to extract the signal from the multi-jet background. One neutrino is involved as well, which allows for a cleaner event reconstruction of missing transverse energy.

Concluding from the above, the presented analysis considers the single-lepton and dilepton channels. They are analysed separately, though, and combined in the final fit. Despite the lower statistics, the single-lepton and dilepton channels are preferred over the all-hadronic channel due to the significant lower background arising from multi-jet processes, and the ability to select events with at least one lepton. Representative[17] Feynman diagrams for the $t\bar{t}H(H \to b\bar{b})$ signal are depicted in Fig. 2.15a and 2.15b for the two channels, respectively. With many final-state particles, one of the main experimental challenges is the low efficiency to reconstruct and identify all of them. Another difficulty is the large combinatorial ambiguities when trying to assign the many jets containing b-hadrons in the final state to the decay products of the Higgs boson and top quarks, which makes it hard to reconstruct the latter. Moreover, the $t\bar{t}H$ signal cross-section is clearly smaller than that for the $t\bar{t}$ production process. As a result, the $t\bar{t} + jets$ processes constitute the overwhelming background of this analysis, making it particularly challenging, especially when these jets originate from b- or c-quarks. The additional *b*- or *c*-quarks can arise from QCD radiation or loop-induced QCD processes.

Among the different $t\bar{t} + jets$ processes, the $t\bar{t}$ process with additional b-jets in the final state, coming from a splitting of a gluon emission ($t\bar{t} + b\bar{b}$), is the dominant background of the analysis, illustrated in Fig. 2.15c. In the phase space of the $t\bar{t}H(H \to b\bar{b})$ signal, with high number of jets and jets identified as b-jets, b-tagging (described in Sec. 5.3) plays a determining role in the analysis. Furthermore, this background has large theoretical uncertainties and is poorly constrained by existing data measurements. The final sensitivity of the analysis

[17] These diagrams show the $t\bar{t}H$ production via gluon fusion, but also via $q\bar{q}$ annihilation is possible (Sec. 2.4.1).

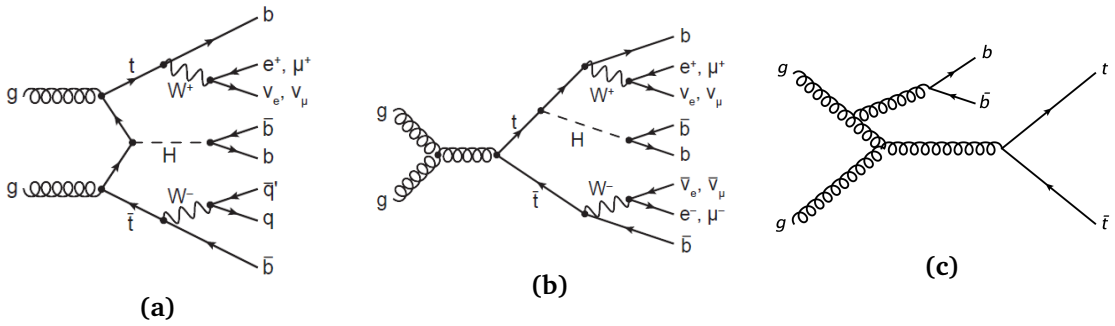


Fig. 2.15: Exemplary tree-level Feynman diagrams for the production of a Higgs boson in association with a top-quark pair (here produced via gluon fusion), $t\bar{t}H$, including the subsequent decays of the top quark-antiquark pair in a) the single-lepton channel (t-channel) and b) the dilepton channel (s-channel), as well as the subsequent decay of the Higgs boson into a bottom quark-antiquark pair ($H \to b\bar{b}$), and for (c) the main $t\bar{t} + b\bar{b}$ background process.

is driven by the modelling of the $t\bar{t} + jets$ background and the most important sources of uncertainties on the signal strength are systematic uncertainties mainly on the $t\bar{t} + b\bar{b}$ process. Even though being challenging the $t\bar{t}H(H \to b\bar{b})$ channel is expected to give high sensitivity to the Higgs boson coupling to third generation quarks, since it involves (at leading order) only the couplings of the Higgs boson to top or bottom quarks. The Higgs boson coupling to the top quark can then be constrained in the combination with the other decay modes.

The ATLAS Collaboration searched for $t\bar{t}H$ production with Higgs-boson decays to $b\bar{b}$ using $t\bar{t}$ decays with at least one lepton in the final state, during Run 1 and Run 2 (defined in Sec. 3.3). The CMS collaboration, also, searched for the same process during these periods. In both experiments, the $t\bar{t}H(H \to bb)$ signal production is parametrised by the ratio of the measured cross-section to the predicted one by the SM, referred to as *signal strength* $\mu_{t\bar{t}H}$.

The first $t\bar{t}H(H \to bb)$ search conducted by the ATLAS collaboration with data collected during Run 1, corresponding to an integrated luminosity of $\mathcal{L} = 20.3$ fb^{-1} at $\sqrt{s} = 8$ TeV, using $t\bar{t}$ decays with at least one lepton [95] (shown in Fig. 2.16a) or no leptons [96]. Then, a combined signal strength $\mu_{t\bar{t}H} = 1.4 \pm 1.0$ was measured [96]. Also, a corresponding search was performed by the CMS Collaboration at $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV resulting in $\mu_{t\bar{t}H} = 0.7 \pm 1.9$ [87]. Both results are compatible with the SM prediction within the uncertainties. Afterwards, these results were combined with each other, and with results from Higgs-boson decay to vector bosons [87], to τ-leptons [88] or to photons [89], resulting in $\mu_{t\bar{t}H} = 2.3^{+0.7}_{-0.6}$ with an observed (expected) significance of 4.4σ (2.0σ) for $t\bar{t}H$ production [13]. Eventually, a strong evidence of the $t\bar{t}H$ production was found.

The $t\bar{t}H(H \to b\bar{b})$ search resumed by the ATLAS Collaboration also during Run 2, with data collected in 2015 and 2016, corresponding to an integrated luminosity $\mathcal{L} = 36.1$ fb^{-1} at $\sqrt{s} = 13$ TeV. During Run 2, the analysis is benefited from the higher centre-of-mass energy reached, since the SM cross-section of the $t\bar{t}H$ production process increases faster than that of the dominant $t\bar{t}$+jets background, resulting in an improved signal-to-background ratio. The analysis followed the Run 1 strategy with the main improvement being the addition of a multivariate analysis technique for the reconstruction of the $t\bar{t}H(H \to b\bar{b})$ final state. Also, the background and fit models are revisited to improve the analysis sensitivity, in particular by constraining the $t\bar{t} + jets$ background. A combined signal strength of $0.84^{+0.64}_{-0.61}$ was measured, illustrated in Fig. 2.16b, with an observed (expected) significance of 1.4σ (1.6σ) standard

deviations [97]. This result was further combined with the analyses of Higgs boson decays into massive vector bosons, τ-leptons, or photons as well as with the results from Run 1. Eventually, the observed (expected) significance was 6.3σ (5.1σ), leading to the observation of the $t\bar{t}H$ production mode [102]. The CMS Collaboration searched for the same processes using $\mathcal{L} = 35.9$ fb^{-1} of data collected at $\sqrt{s} = 13$ in 2016, and measured $\mu_{t\bar{t}H} = 0.72 \pm 0.45$ [98,99]. This result, combined with the result from Run 1, also contributed to the observation of the $t\bar{t}H$ production mode [103].

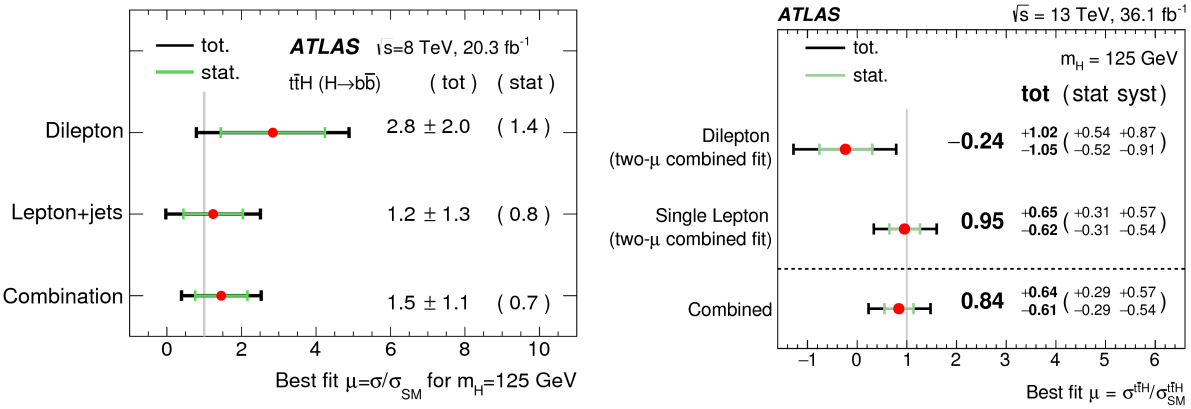


Fig. 2.16: Observed values of the $t\bar{t}H(H \to b\bar{b})$ signal strength and their uncertainties for the individual channels and their combination obtained using a) $\mathcal{L} = 20.3$ fb^{-1} data at $\sqrt{s} = 8$ TeV [95] and b) $\mathcal{L} = 36.1$ fb^{-1} data at $\sqrt{s} = 13$ TeV [97], with the ATLAS experiment.

In this thesis, the measurement of the $t\bar{t}H$ production cross-section in the $H \to b\bar{b}$ decay channel using the full LHC Run 2 data collected by the ATLAS detector, corresponding to an integrated luminosity of $\mathcal{L} = 139$ fb^{-1} at $\sqrt{s} = 13$ TeV [100], is presented. Events with either one or two leptons are analysed separately in exclusive single-lepton or dilepton categories defined according to the number of leptons, the number of jets and the number of jets identified as originating from b-hadrons (b-jets). In the single-lepton channel, a specific category, referred to as "boosted" in the following (defined in Sec. 6.4.1), is designed to select events in which the Higgs boson and possibly also the hadronically decaying top quark are produced with a high transverse momentum (p_T) relative to their rest mass. As a result, their decay products are collimated in large-radius jets, with a radius parameter equals to 1 (see Sec. 5.2.5). The boosted channel is also potentially more sensitive to deviations from the SM [107] at the LHC due to the high centre-of-mass energy.

Chapter 3

The Large Hadron Collider and the ATLAS Detector

In order to test the validity of a theory, its predictions should be compared to experimental observations. The main goal of experiments is either to measure already observed phenomena with higher precision looking for potential deviations from the theory prediction, or to search for new phenomena either already predicted by a theory or totally unexpected. In the field of high energy physics, the most common are collider, fixed-target, or astroparticle experiments.

Particle accelerators, reaching unprecedented energies and interaction rates, provide a suitable environment for intensive testing of the Standard Model (SM) of particle physics. They use electromagnetic fields to propel charged particles, such as protons, up to velocities close to the speed of light and contain them in collimated beams. Colliding these beams at specific interaction points, allows to study the complex processes in particle physics and simulate the conditions right after the Big Bang[1]. These collisions produce massive particles, such as the Higgs boson or the top quark. By measuring their properties, scientists can get a better understanding of matter and of the origins of the Universe. However, these massive particles cannot be directly observed, since they almost immediately decay (or transform) into lighter particles, which in turn also decay. Eventually, the particles emerging from the successive decays are tracked and identified by detectors.

The Large Hadron Collider (LHC) [108] is the largest and most powerful accelerator in the world. The study presented in this thesis has been performed using data collected with the ATLAS detector [116], placed at one of the collision points at the LHC.

3.1 The Large Hadron Collider at CERN

The LHC [108] is a synchrotron accelerator constructed to produce a high rate of proton-proton (pp) and lead ions collisions at the TeV scale. It is located at CERN (Conseil Européen pour la Recherche Nucléaire) [109], which is an international organisation for particle physics research established near Geneva. The LHC is suited in a circular tunnel with a circumference of approximately 27 km, lying between 45 m and 170 m under the French-Swiss boarders so as to minimise the background from the cosmic radiation. It is also built at a inclination of 1.4% towards lac Léman in Geneva, in order to minimise excavation costs as well as to line up with

[1]The Big Bang is the widely accepted theory for creation of our universe. It states that all matter in the universe evolved from an extremely dense and hot singularity in an explosive event about 13.7 billion years ago.

the other tunnels in the CERN accelerator complex. Though both protons and heavy ions are collided at the LHC, the following discussion mostly focuses on the former, as this dissertation reports on a measurement based on pp collision data. The LHC is designed to boost protons up to an energy of 7 TeV, generating collisions at a centre-of-mass energy[2] of $\sqrt{s} = 14$ TeV.

3.1.1 Luminosity and pileup

As discussed in Sec. A.1.1, a scattering process is described in terms of a cross section σ, which is thought of as the effective area the incoming particles have to hit in order to initiate the interaction and create new particles. Then, the rate of events $\dot{N}$ in collisions depends on the process-specific cross section, representing the underlying physics, and the particle flux, described by the *instantaneous luminosity* L,

$$\dot{N} = L\sigma. \tag{3.1}$$

The instantaneous luminosity is, in fact, a measure of the number of collisions that take place in a detector per cm^2 and per second, depending on the beam and accelerator properties, as follows [111, 112]

$$L = \frac{N_{b,1} N_{b,2} n_b f_{rev}}{4\pi\sigma_x\sigma_y} F. \tag{3.2}$$

Considering pp collisions at the LHC, $N_{b,1}$ and $N_{b,2}$ are the number of protons per bunch in each of the two beams ($N_{b,1} \simeq N_{b,2} \simeq 1.15 \cdot 10^{11}$) and n_b is the number of bunches injected at the LHC per revolution, since each particle in a bunch might collide with anyone from the bunch approaching head on. Also, f_{rev} is the machine revolution frequency which is approximately $f_{rev} \sim \frac{c}{27\text{km}} \simeq 11$ kHz, while F is the geometric luminosity reduction factor that serves as a small correction factor to account for the crossing angle between beams at the interaction point. Finally, LHC is a circular collider and a Gaussian-shaped effective beam is assumed with area $4\pi\sigma_x\sigma_y$, where σ_x and σ_y stand for the horizontal (x-scan) and vertical (y-scan) Gaussian widths of the colliding beams.

Eventually, the design instantaneous luminosity of the LHC in pp collisions is $L = 10^{34}$ cm^{-2}s^{-1} [108], which means that detectors in the LHC might produce 10^{34} collisions per second and per cm^2. Nonetheless, due to the excellent performance and various improvements to the machine this value was surpassed, reaching a peak luminosity of $1.9 \cdot 10^{34}$ cm^{-2}s^{-1} during 2018 data taking [111]. It should also be noted that the luminosity provided by the LHC is not constant. Provided that the bunches collide repeatedly, the number of protons in each bunch decreases and consequently also the luminosity (eq. 3.2). Additionally, the beam parameters can change during the year.

The total amount of data recorded over a certain period is called *integrated luminosity*. It is defined by integrating the instantaneous luminosity $\mathcal{L} = \int L dt$ and is often measured in units of inverse femtobarns (fb), where 1 fb = 10^{-39} cm^2. As a result, the produced events by the LHC in a given time interval, which are the number of collisions occurred, can be retrieved by integrating over the data-taking period in eq. 3.1

$$N = \sigma \int L dt = \sigma\mathcal{L}. \tag{3.3}$$

[2]The total available energy for producing new particles in collision experiments, in the centre-of-mass frame, is called *centre-of-mass energy*, $\sqrt{s}$, and it is invariant in any frame. When two beams of identical particles with equal momentum collide head on, the energy of the collision is $\sqrt{s} = 2 \cdot E_{beam}$.

Larger luminosity signifies larger statistics to analyse, but it also comes with difficulties. The LHC does not collide individual protons but bunches (see Sec. 3.1.3), and given the high density of the beam bunches and the high frequency of collisions, more than one collision per bunch crossing may occur simultaneously. This effect, resulting in the overlap of the electronic signals from multiple interactions, is called *pileup* [114]. The actual number of interactions is a random Poisson variable. Thus, the pileup, μ, i.e. the mean number of interactions per crossing, corresponds to the mean of the Poisson distribution of the number of interactions per crossing calculated for each bunch, and is determined through [111, 113]

$$\mu = \frac{L_{bunch}\sigma_{inel}}{f_{rev}}, \tag{3.4}$$

where L_{bunch} is the instantaneous luminosity per bunch, σ_{inel} is the inelastic cross section ($\sigma_{inel} = 80$ mb for $\sqrt{s} = 13$ TeV pp collisions), and f_{rev} is the LHC revolution frequency. The original design value of pileup averaged over all colliding bunch pairs for the LHC is $\langle\mu\rangle = 19$.

In general, there is at most one hard-scattering process (interaction of highest energy) per bunch crossing, producing an event interesting for a physics analysis. The other interacting protons usually result in a soft (low-energy) scattering. Only hard-scatter events are considered as a signal, while the additional interactions in the same bunch crossing are considered as background (pileup interactions) in most analyses, which play a significant role in the reconstruction of physics objects. From eq. 3.4, it follows that larger luminosity results in larger number of protons interacting per bunch crossing. This leads to larger noise and background in the detector, and to more challenging identification of particles originating from the hard-scattering process (interaction of highest energy) and reconstruction of objects.

Moreover, due to the large number of protons within a bunch, more than one pp collisions can occur within a bunch crossing, so the additional pp collisions that occur in the same bunch crossing are referred to as *in-time pileup*. Higher $N_{b,1(2)}$ produces more interactions within a given bunch crossing, i.e. results in higher luminosity as seen in eq. 3.2, meaning higher in-time pileup. On the contrary, the additional pp collisions occurring in bunch-crossings just before and after the collision of interest, which cannot be resolved fast enough by the detector, are called *out-of-time pileup*. Large n_b reduces the bunch spacing, causing interactions from different bunch crossings to overlap, increasing the out-of-time pileup.

3.1.2 A proton-proton collider

Nowadays, the LHC, as a circular hadron collider, is suitable for reaching the highest possible energy frontiers at the TeV scale, making it particularly sensitive to potential discoveries. By contrast, e^+e^- circular colliders, such was the LEP [74], suffer from a large loss of energy due to synchrotron radiation and as a result cannot achieve such high energies. Furthermore, pp collisions primarily take place at the LHC (heavy ions collisions are typically performed for one month a year) in order to achieve high luminosity, namely large amount of data. On the contrary, proton-antiproton ($p\bar{p}$) collisions, which used to take place at Tevatron [75], cannot offer a large amount of data due to the difficulty to produce antiprotons.

Nevertheless, pp collisions also come with some difficulties. The hard scatter (interaction of interest) occurs between constituents of the protons, namely quarks (q) and gluons (g). At the LHC gg-initiated processes are favoured with respect to $q\bar{q}$- or qg-initiated processes due to the parton dynamics inside protons. Partons carry only a fraction of the proton energy following the parton distribution function (see Fig. 4.3). However, this effect involves non

perturbative QCD phenomena, requiring input from other experiments which come with their uncertainties. Moreover, on top of the hard scatter, the remaining partons in the protons can interact generating an underlying event (see Sec. 4.2.5), which is not well described by the existing models. Another challenge in pp collisions is the overwhelming production of gluons and quarks (observed as multi-jet events) due to the large QCD coupling. These events are an important source of background events for many analyses of the LHC physics programme, thus need to be suppressed.

Figure 3.1 shows the production cross-section of several of the main SM processes as a function of the pp center-of-mass energy. It also illustrates that pp collisions products are dominated by multi-jet events as mentioned earlier. Considering also the relatively large cross section of top-quark production modes, the LHC allows for and favours precise measurements of the top-quark properties. In addition the operating energy of the LHC rises the Higgs boson production rate to an accessible value, making the discovery of the Higgs boson and its properties possible. Nevertheless, the Higgs production modes are orders of magnitudes lower than many other SM processes, therefore sophisticated techniques are necessary to extract Higgs-boson signals in this analyses. As already introduced, for the $t\bar{t}H$ production mode, the main background originates from $t\bar{t}$ processes, whose cross section is more than two orders of magnitudes larger.

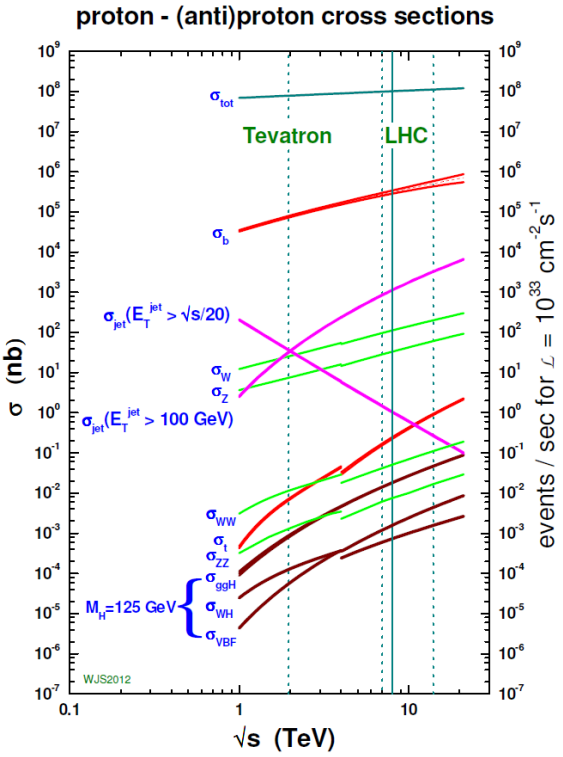


Fig. 3.1: Expected cross sections for a few typical SM processes in $pp(p\bar{p})$ collisions as a function of the center-of-mass energy [115].

3.1.3 The LHC setup

As depicted in Fig. 3.2, the LHC is the final part of the large CERN accelerator complex. Protons are fed to the LHC starting from a small bottle of hydrogen gas, whose atoms are

ionised within an electric field. The resulting protons follow a pre-acceleration procedure, to successively increase their speed. At first, they are sent through the linear accelerator LINAC2[3] (80 m long), where they are accelerated up to 50 MeV. These low energy protons are subsequently injected into the proton-synchrotron BOOSTER (circular accelerator of 157 m circumference) and then enter the Proton Synchrotron (PS) (circular accelerator of 628 m circumference), where they are accelerated to 1.4 GeV and 25 GeV, respectively. The last pre-acceleration stage is the Super Proton Synchrotron (SPS) (circular accelerator of 7 km circumference), boosting the protons, as a single beam, to energies of up to 450 GeV. The proton beam from the SPS is finally injected into the LHC, where it splits into two beams traveling in opposite directions through separate vacuum tubes, to be further accelerated to their final energies.

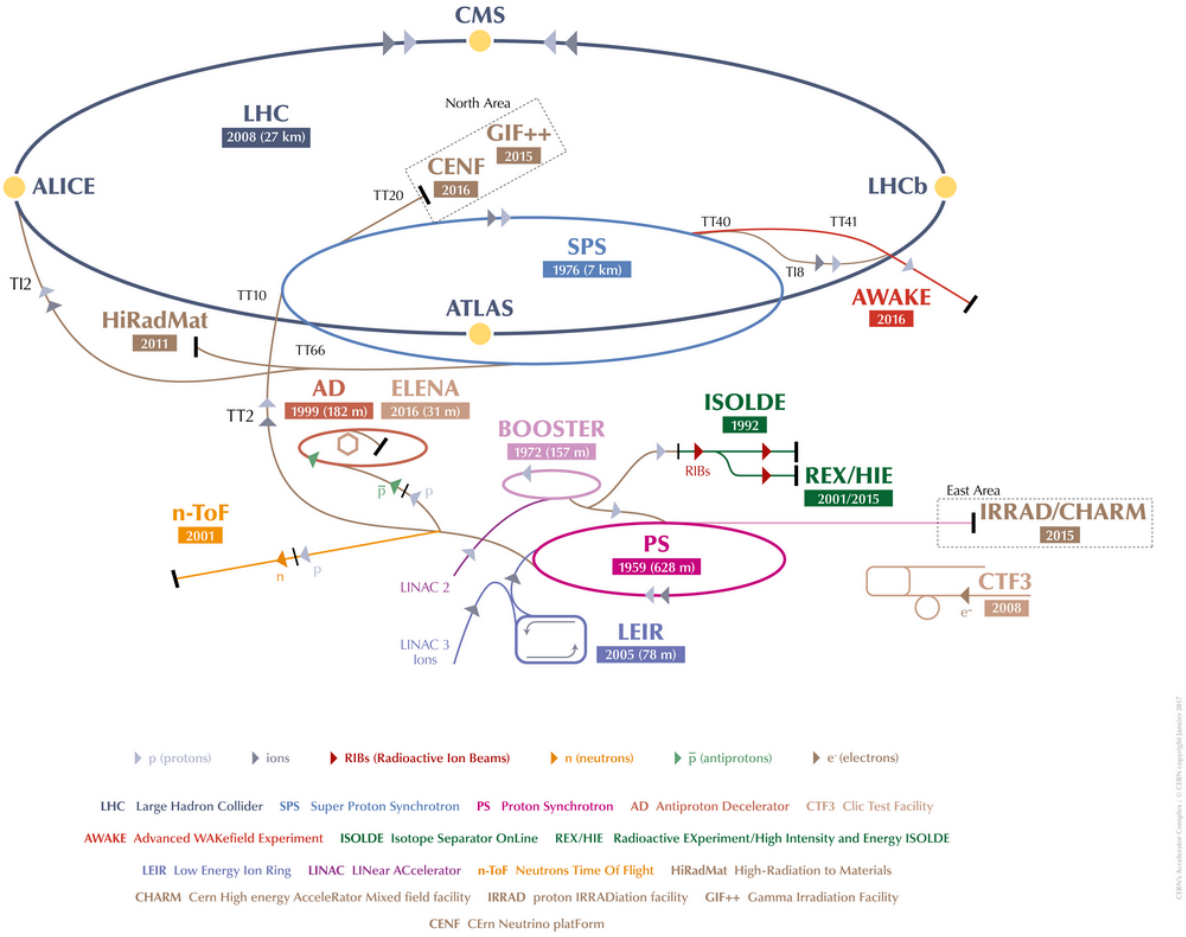


Fig. 3.2: Overview of the CERN accelerator complex, showing the pre-accelerators providing proton beams to the LHC for the four largest experiments [110].

Inside the LHC (with circumference of 27 km) the protons from each beam are accelerated up to 6.5 TeV energy, leading to a center-of-mass energy $\sqrt{s} = 13$ TeV during Run 2 [4] (see Sec. 3.3). Protons at the LHC are accelerated by eight radiofrequency (RF) cavities per beam, each delivering a voltage of 2 MV and oscillating at 400 MHz. In this way, the particle beam is

[3]For the third LHC data taking period (*Run 3*), started in mid-2022 and is scheduled to last until the end of 2025, the new LINAC4 accelerator is used.

[4]For the LHC Run 3, a centre-of-mass energy of $\sqrt{s} = 13.6$ TeV has been achieved.

sorted into bunches of around $1.15 \cdot 10^{11}$ protons[5]. Between each consecutive bunch there are 7.5 m of space and considering bunches moving around the LHC ring at almost the speed of light, the resulting bunch spacing, i.e. the time interval between pp bunch crossings, is 25 ns. During collisions the effective number of bunches in the LHC is 2808.

The LHC relies on superconducting electromagnets to keep particle beams on course around the accelerator and reach energies in TeV scale. They produce a magnetic field of about 8 T, operating at 1.9 K, colder than the 2.7 K of outer space, to achieve resistance-free electrical conduction. In particular, 1232 dipole magnets (15 m in length) bend the beams, and 392 quadrupole magnets (each 5–7 m long) focus the beams. They are used to squeeze the beams at each interaction point, narrowing the effective size of the beams to maximise the number of colliding particles. Just prior to collision, the RF cavities squeeze further the 2808 proton bunches to increase the chances of collisions, ensuring high luminosity at the collision points. The LHC proton beams are declared stable, once they are aligned, squeezed, focused and eventually steered to collide head-on, verifying that the collision mechanism is ready to take data that are good for physics studies. Finally, the collisions start and continue until the beam luminosity has decreased by roughly 50%, possibly up to 10 hours. A new fill starts once the bunches have lost a significant amount of their protons, impacting the data collection rate.

The beams inside the LHC collide at four locations around the accelerator ring, corresponding to the positions of four main experiments which are equipped with different detectors, illustrated in Fig. 3.2. The ATLAS (A Toroidal LHC Apparatus) [116] and the CMS (Compact Muon Solenoid) [117] are general-purpose detectors pursuing a wide range of physics, comprising SM precision measurements as well as searches for BSM phenomena, such as supersymmetry, exotic particles or Dark Matter searches. However, they use a different technology with the major differences being about their muon and tracking systems. ATLAS and CMS, as two separate experiments with different detectors, are independent but also complementary. Each of them can provide a confirmation of particle discovery by the other experiment and results can be combined for enhanced precision. LHCb (LHC beauty) [118] is dedicated to heavy-flavour physics and the search for BSM effects, especially investigating CP-violating processes, via precise measurement of hadrons containing *beauty*- and *charm*-quarks. ALICE (A Large Ion Collider Experiment) [119] focuses on QCD measurements for strongly interacting matter as well as quark-gluon plasma description at large energy densities and high temperature in heavy-ion collisions. As already mentioned, this thesis uses data collected by the ATLAS detector, which is briefly described in the following.

3.2 The ATLAS detector

As already introduced, the ATLAS detector [116] is one of the two general-purpose detectors at the LHC covering a wide range of particle physics. It measures the properties of particles produced in high-energy pp collisions, as well as heavy-ion collisions. The high luminosity and centre-of-mass energy at the LHC allow for searches and precision measurements of diverse SM processes, offering also a good opportunity for the discovery of potential BSM physics in the TeV energy regime. In order to observe these rare events, the main challenge is to distinguish them from other processes and particles produced during pp collisions.

The detector is situated in the Interaction Point 1 of the LHC ring, 100 m underneath the ground. It is a hermetic detector of nearly 4π coverage in solid angle around the central

[5]During Run 3 almost $1.8 \cdot 10^{11}$ protons per bunch are reached

interaction point, which is necessary for reconstructing the energy flow in an event. The ATLAS detector is the largest volume particle detector ever constructed. It weighs approximately 7000 tons and has a cylindrical profile, 25 m in diameter and 44 m in length. It consists of sub-detectors, magnets, and supporting infrastructure inserted as concentric cylinders around the interaction point where the proton and ion beams of the LHC collide. These subsystems are designed to reconstruct the products of the collisions.

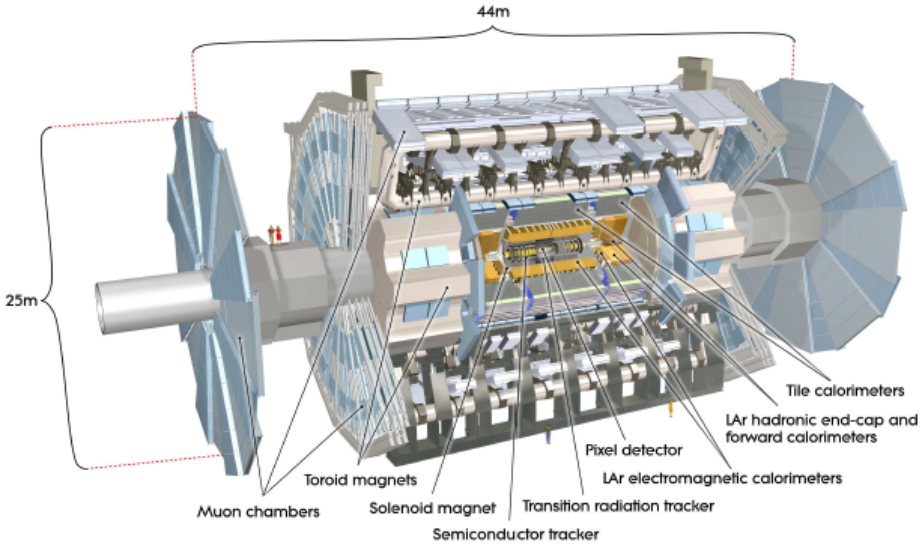


Fig. 3.3: The ATLAS detector and its subsystems [116].

Figure 3.3 shows an overview of the ATLAS detector and its subsystems. The innermost part of the detector is the Inner Detector (ID), composed of three tracking sub-detectors which are used to reconstruct tracks of charged particles and their interaction vertices. The ID contains a silicon pixel detector surrounded by a semiconductor microstrip detector and a straw-tube tracker that can detect electron transition radiation. Additionally, it is enclosed by a superconducting solenoid magnet, which provides axial magnetic fields of up to 2 T and bends the tracks of charged particles to allow for precise momentum measurement of their transverse momentum and charge. Around the ID, the calorimeter system is placed, where charged and neutral particles exiting the ID are absorbed and measured. Closest to the ID the electromagnetic calorimeter is positioned, which primarily measures photons and electrons. Then, hadrons are mainly measured in the hadronic calorimeter, which stops all the remaining particles apart from muons and neutrinos. The latter pass through the whole detector undetected, given their low interaction rate. Finally, the muon spectrometer surrounds the ATLAS calorimeters and measures the position and energy of charged muon tracks. It consists of three large superconducting air-core magnets, and a multi-component tracking system to detect muons traversing the detector. They provide a 4 T toroidal magnetic field to allow for muon momentum measurements.

3.2.1 Detector geometry and coordinate system

The ATLAS detector has a cylindrical geometry and its various components are generally divided into two main categories based on their geometry. Firstly, it is composed of a barrel, which has cylindrical structure and covers the central region of the detector around the in-

teraction point of the two beams. Further from the interaction point there are two end-caps, which have a planar circular geometry perpendicular to the beam-pipe and cover the forward and backward region of the detector at the end of the central barrel.

The ATLAS detector uses a right-handed coordinate system, illustrated in Fig. 3.4, with its origin being at the centre of the detector where the collisions take place. In Cartesian coordinates, the direction of the beam pipe defines the z-axis and the $x-y$ plane is defined transverse to it. The positive x-axis is defined as pointing from the interaction point towards the centre of the LHC ring and the positive y-axis is defined as pointing upwards. Moreover, in the polar coordinate system, the polar angle θ is defined with respect to the beam axis, while the azimuthal angle ϕ is computed in the $x-y$ plane around the beam axis.

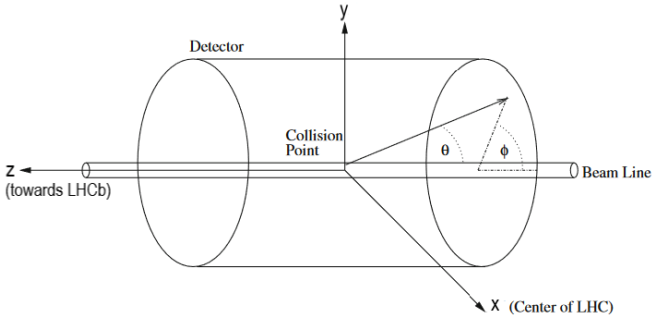


Fig. 3.4: Coordinate system of the ATLAS detector.

In high energy physics, particles move with a velocity close to the speed of light, so relativistic effects have to be considered. Using the polar angle θ is not always practical, due to large dependence of the particles produced in collisions on the angle, where most of them are produced closer to the beam-line. For this purpose, other quantities are defined and used instead, exploiting the fact that the incident velocities of the particles taking part in a collision are along the beam axis. The rapidity, y, is a Lorentz invariant expression that describes the motion of a relativistic massive particle (with four-momentum $p = (E, p_x, p_y, p_z)$) relative to the beam axis. According to its definition in eq. 3.5, when a very high-energy particle is moving transverse to the beam axis, the rapidity tends $y \to 0$. Whereas, when the particle is moving close to the beam axis in either direction, it results in $y \to \pm\infty$. However, it can be hard to measure y for highly relativistic particles ($E \gg mc^2$), since both the energy and the total momentum are needed.

$$y = \frac{1}{2\ln \frac{E+p_z}{E-p_z}} \qquad \overset{E \gg mc^2}{\longrightarrow} \qquad \eta = -\ln \tan\left(\frac{\theta}{2}\right) \tag{3.5}$$

Therefore, the pseudorapidity, η, is defined as the measure of the polar angle against the beam line (eq. 3.5), being a low mass approximation of the rapidity but much easier to measure for highly energetic particles. In fact, for highly relativistic particles ($E \gg mc^2$), which are approximately treated as being massless, it holds that $\eta \approx y$. By definition, the pseudorapidity is positive in the forward region (defined by the positive z-axis from the interaction point) and negative in the backward region, while small $|\eta|$ values refer to the central region of the detector. In case of a particle transverse to the beam axis the pseudorapidity yields in $\eta(\theta = 90°) = 0$, while it diverges to ∞ for particles close to the beam axis ($\theta = 0°$ or 180°).

Also, it is worth mentioning that particles exceeding the detector region of $|\eta| > 4.9$ can not be considered for analyses, because the calorimeters do not cover areas exceeding this

range (described in Sec. 3.2.3). For the purpose of this analysis, physics objects reconstructed within the coverage of the ID are considered, namely within the region of $|\eta| < 2.5$.

Another important quantity is the angular separation between two particles i and j emerging from the interaction point. The distance ΔR in the $\phi - \eta$ plane between these two objects can be defined from the polar angle ϕ and the pseudorapidity η

$$\Delta R = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2} = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}. \tag{3.6}$$

In some cases, the angular separation is also defined using the rapidity instead of the pseudorapidity. This quantity is used for isolation criteria of physics objects (Sec. 6.3.2), or for the clustering of energy deposits to form jet objects (Sec. 5.2.1) when defined in the $\phi - y$ plane.

Finally, the *transverse momentum* p_T is defined as the momentum in the $x - y$ plane, i.e. the plane transverse to the beam axis

$$\vec{p_T} = \begin{pmatrix} p_x \\ p_y \end{pmatrix} \qquad \text{and} \qquad p_T = \sqrt{p_x^2 + p_y^2}, \tag{3.7}$$

and is of particular importance in the ATLAS measurements. Since protons are composite particles and only part of them reacts in pp collisions, the longitudinal component of the momentum of the initial partons is unknown. However, given that protons at the LHC collide head on in the longitudinal direction, the transverse momentum of the initial partons is known to be zero in the lab system at the time of the collision. Also, the solenoidal magnetic field of the ID allows to measure the p_T of charged particles directly from the curvature of the track.

3.2.2 Inner Detector

The Inner Detector (ID) [116] is a series of tracking detectors and its main purpose is to reconstruct tracks of charged particles produced in the collisions at the LHC. Thousands of particles are produced in every collision and since the ID is situated close to the interaction point, it should have a good granularity and a great resistance to radiation damage. It is around 6 meters long and 2 meters high, covering a pseudorapidity range up to 2.5, and it is composed of three different subsystems. The Pixel detector is set close to the beam pipe. Beyond this layer the Semi-Conductor Tracker (SCT) is placed and then there is the Transition Radiation Tracker (TRT) surrounded by a superconducting solenoid magnet, which provides a uniform 2 T axial magnetic field. The superconducting magnet made of NbTi is cooled via liquid helium to a temperature of 1.8 K. Each sub-detector is split into cylindrical concentric barrel modules covering the central region and disk-shaped end-cap modules covering the forward/backward regions. This structure offers high-precision measurements with fine detector granularity achieving better momentum and vertex resolution. An overview of the arrangement of the subsystems in the central detector and end-cap region is shown in Fig. 3.5.

The pixel detector is particularly important for the track reconstruction, the primary vertex reconstruction as well as for secondary vertex finding (Sec. 5.1). It consists of three concentrically arranged pixel module (barrel) layers around the beam pipe in the central detector region. In addition to them, there are three discs in each of the two end-caps of the detector, limited to a coverage of $|\eta| < 2.5$. The modules consist of silicon semiconductor sensors and readout electronics providing a high granularity crucial for the spatial resolution, which is necessary for track reconstruction in the high pileup environment in pp collisions at the LHC. Also, between Run 1 and Run 2, an extension of the original system was placed closer to the

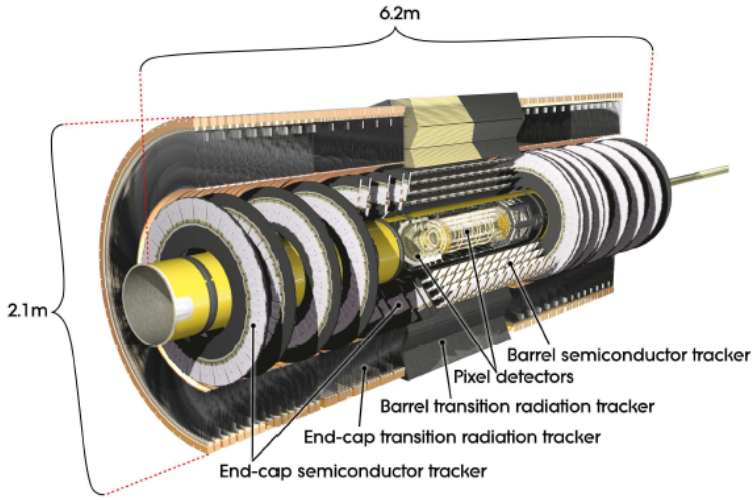


Fig. 3.5: Cut-away view of the ATLAS inner detector [116]

beam pipe, the Insertable B-layer (IBL) [121], as a fourth layer of pixel modules only in the barrel region. Its addition improved the detector granularity and helped to mitigate negative effects of the radiation damage in the other layers of the Pixel detector. It also provided more precise vertex measurements and a better performance of the identification of jets originating from b-quarks (Sec. 5.3).

The Semi-Conductor Tracker (SCT) is located beyond the pixel detector and also consists of silicon semiconductor sensors. Instead of pixels, the SCT sensors are segmented into strips. The SCT is arranged in four cylindrical double-layers of silicon microstrip detectors in the barrel region. Additionally, there are nine discs in the end-caps on each side of the detector, covering a pseudorapidity region of $|\eta| < 2.5$. In general, the semiconductor-based detectors in ATLAS operate at a temperature between -10 °C and -5 °C to suppress different types of electronic noise and radiation damage. The SCT enhances the momentum resolution of tracks by providing higher radius hits.

Surrounding both the silicon-based detectors the Transition Radiation Tracker (TRT) is situated in radial direction, which combines tracking and identification of particles. It is divided into a barrel and end-caps and it covers a region only up to $|\eta| < 2.0$. The TRT consists of densely packed straw tubes filled with a mixture of gases (70% Xe, 27% CO_2, 3% O_2). A particle passing through the mixture ionises the gas, generating a current. The region around the tubes is filled with a material that enhances the electron transition radiation. Highly relativistic particles traversing these materials with different dielectric constants emit transition radiation photons resulting in high-energy depositions in the straws. These hits can be well distinguished from low energy track ionisation hits allowing for electron identification apart from the tracking information. Nevertheless, due to leakage of the gas mixture, some tubes were refilled but with an Argon- instead of Xenon-dominated mixture, due to budget constraints. As a result, the electron-discriminating properties of the TRT got suppressed.

Combining the hit information of charged particles along their trajectory through the subsystems of the ID provides high precision track reconstruction. Also, the magnetic field provided by the solenoid magnet, deflects and bends the trajectory of the charged particles in the ID, via the Lorentz force depending on their momentum and charge. This allows for the determination of their momentum and charge by measuring the curvature of their tracks.

3.2.3 Electromagnetic and Hadronic Calorimeters

Surrounding the ID and the solenoid, the calorimeter system [116] of the ATLAS detector is placed. The ATLAS calorimeter system is divided into two different parts, the electromagnetic (EMCal) and hadronic (HCal) calorimeters, each optimised for different particle types, and an overview is depicted in Fig. 3.6. The calorimeter system covers a large range of $|\eta| < 4.9$.

In contrast to the tracking detectors, which simply measure points along the particle trajectory with as little material as possible, calorimeters stop the particles that exit the ID in a dense material. The energy deposited by a particle and by the products of its interaction with the material is collected and measured. This allows for the determination of the position and energy of the original particle. Furthermore, they are optimised to measure the shower[6] properties to allow for particle identification. The calorimeters are able to detect not only charged but also neutral particles produced in the pp collisions at the LHC. Nonetheless, only the neutrinos and muons have an interaction rate small enough to pass the calorimeters unaffected, though muons still deposit a small amount of energy.

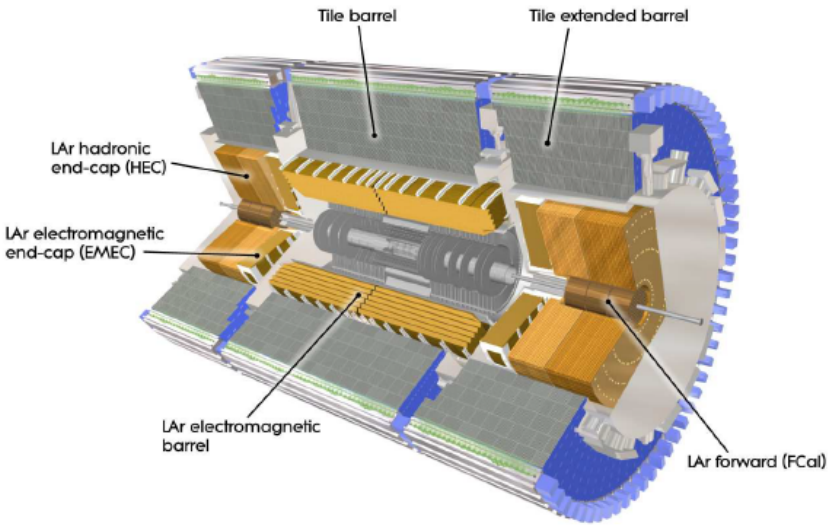


Fig. 3.6: Cut-away view of the ATLAS calorimrters [116]

The calorimeters used in ATLAS are composed of alternating layers of dense passive material and an active medium. The passive material is composed of heavy absorber material that interacts with charged and neutral particles. The active medium is a sampling material composed of plastic scintillators or liquid argon that reacts in the presence of charged particles and measures their deposited energy. Each calorimeter unit then consists of numerous layers of absorbers and samplers. Concluding from the above, particles passing through the passive material induce particle showers which in turn induce signals in the active medium through ionisation or scintillation that are proportional to the total released energy.

Electromagnetic calorimeter

The EMCal is located just beyond the solenoid magnet and is closer to the beam pipe than the HCal, thus having higher granularity. It is designed to measure energies of light particles

[6] Showers are cascades of secondary particles with lower momenta, which are formed when a highly energetic particle interacts with dense material.

which interact mainly through the electromagnetic interaction, such as electrons and photons. The absorber of this calorimeter consists of multiple layers of lead having Liquid Argon (LAr) as active sampling material, providing a homogeneous response. Electrodes then collect the charge generated in the LAr by the particle shower. This allows for the determination of the energy of the showering particle. The EMCal is divided into a central barrel, endcap calorimeter (EMEC) in the backward region and a forward calorimeter (FCal) in the forward region, and overall covers a range of $|\eta| < 3.2$. This coverage is assured by a large number of cells with a high granularity in the $\eta - \phi$ plane with a low size of 0.025×0.025. The detector has an accordion geometry, which provides a full ϕ coverage.

In general, electrons and photons produce additional electrons and photons when interacting with material, resulting in the characteristic electromagnetic showers. In particular, electrons emit a photon usually via bremsstrahlung, while photons split into an electron-positron pair (photon conversion), which can in turn undergo bremsstrahlung. The distance in which an electron looses $1/e$ of its energy due to bremsstrahlung is called *radiation length*, X_0. It is a practical unit to measure absorption properties of an EMCal and strongly depends on the type of material the electron traverses. The thickness of the EMCal is over $22X_0$ in the barrel and over $24X_0$ in the end-caps, large enough to contain most of the electromagnetic showers.

Hadronic calorimeter

The HCal is installed around the EMCal and is optimised to measure energies of all the remaining particles (except for muons and neutrinos) within its volume, namely neutral and charged hadrons. Hadrons lose energy mainly through strong and weak nuclear interactions with the absorber, though charged particles also lose energy by ionising the detector material. This produces also photons and electrons, resulting eventually in electromagnetic showers. The hadronisation (defined in Sec. 4.2) of quarks and gluons due to the strong interaction produces hadronic parton showers. The distance over which hadrons lose on average $1/e$ of their energy due to nuclear interactions is referred to as *nuclear interaction length*, λ. It provides a good description of the absorption properties of the detector.

The HCal is divided into two main parts, the Tile hadronic Calorimeter (TileCal) and the liquid-argon Calorimeter (LAr), which is further split into the Hadronic End-cap Calorimeter (HEC), and the Forward Calorimeter (FCal). The TileCal forms an outer cylindrical envelope around the other calorimeters with combined coverage of the barrel and end-cap of $|\eta| < 1.7$. It has cells with granularity of 0.1×0.1 in the $\eta - \phi$ plane, significantly larger than that of the EMCal. It uses steel as the absorber and plastic scintillators for the sampling, both forming several layers (tiles). Overall they form a volume with an average length of 7.4λ, covering most of the particle shower. The light in the scintillators is collected and measured using wave-shifters and photomultipliers located at the edges of the tiles. Then, the HEC is the first extension of the hadronic calorimeter to higher $|\eta|$, covering a region of $1.5 < |\eta| < 3.2$, slightly overlapping with the TileCal. The absorber is made of copper interlaced with gaps filled with LAr. The granularity of the HEC in the $\eta - \phi$ plane is the same as that of the TileCal for low pseudorapidity and slightly larger (0.2×0.2) for the forward parts of the detector. Finally, the highest values of $|\eta|$ are covered by the FCal, covering a region of $3.1 < |\eta| < 4.9$. It has by far the lowest granularity compared to the other hadronic calorimeters with much larger cells. Due to higher radiation levels in the forward/backward regions of the detector, the HEC uses copper or tungsten and liquid argon as passive and active material, respectively. Overall the FCal has a depth of 10λ.

3.2.4 Muon Spectrometer

The outermost part of the ATLAS detector, surrounding the calorimeters, is the large Muon Spectrometer (MS) [116], which is responsible for the detection and precise measurement of muons. This is achieved by a system of tracking detectors embedded in a magnetic field in ϕ-direction (typically perpendicular to the muon trajectory), which is induced by three air-core toroidal magnets (one in the barrel and the other two in the end-caps). Whereas other charged particles deposit all or most of their energy in the calorimeters, muons traverse these detector components almost without losing any energy. Therefore, the MS contributes to the identification of muons and reconstruction of their momenta measuring their curvature in the toroidal magnetic field. The MS consists of four detector systems grouped into high-precision muon tracking and trigger chambers. An overview of the muon system is depicted in Fig. 3.7.

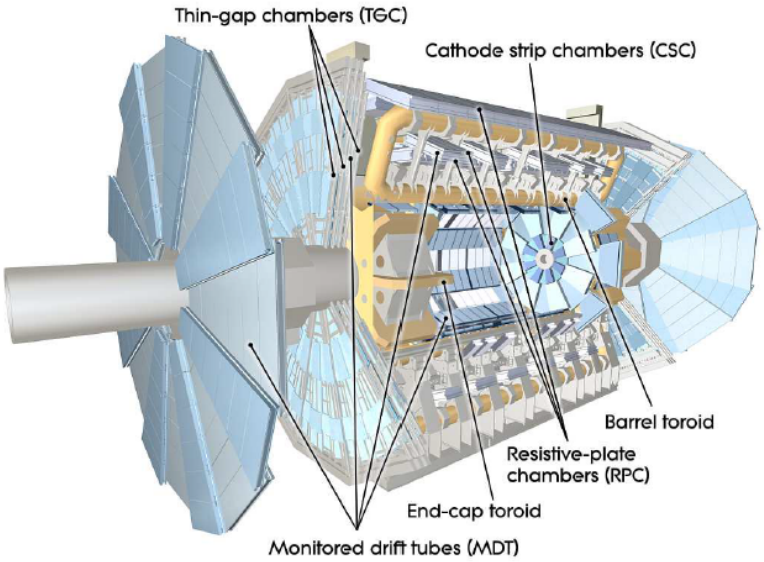


Fig. 3.7: Cut-away view of the ATLAS muon system [116]

The Monitored Drift Tube (MDT) chambers constitute the largest part of the MS and are installed in the barrel and end-cap regions, oriented in the ϕ direction, covering $|\eta| < 2.7$. They are aluminium drift tubes filled with an Ar/CO_2 (93/7%) gas mixture and a wire at their centre. Typically, each chamber contains 3-8 layers of drift tubes offering a spatial resolution of 35 μm. Their main purpose is to provide precision tracking information in the central region.

However, the expected particle multiplicity at high-$|\eta|$ is too high for the MDTs to handle. As a result, Cathode Strip Chambers (CSCs) are installed in the forward/backward regions, which are proportional multi-wire chambers with cathodes divided into strips, each side perpendicular to the other. This provides a good spacial resolution of 40 μm. The CSCs have an end-cap geometry with eight segments covering the whole ϕ range and $2.0 < |\eta| < 2.5$, while the wires of each chamber are perpendicular to the beam pipe. Each chamber has four layers, hence provides four independent measurements of the track trajectory.

The muon trigger chambers, covering the whole ϕ range and $|\eta| < 2.4$, are designed to provide a fast (15-25 ns) read out about the presence of muons to the trigger system. They also provide additional measured coordinates for track reconstruction. In the barrel region of $|\eta| < 1.05$, three layers of Resistive Plate Chambers (RPCs) are placed. The RPCs are gaseous detectors built of two parallel plates with high resistivity under high voltage. The gap between

them is filled with a gas mixture mostly of $C_2H_2F_4$. Particles traversing the detector produce avalanches in the volume of the detector and the charge is then collected on the plates. The plates are segmented into strips to allow for a position measurement with a spatial resolution of 10 mm. The end-caps ($1.05 < |\eta| < 2.4$) are covered by the Thin Gap Chambers (TGCs). They are proporional multi-wire chambers filled with a gas mixture of 55% CO_2 and 45% n-C_5H_{12}. The radial coordinate is determined by the wires and the azimuthal by radial strips. Apart from the trigger information, the TGCs provide ϕ information with a resolution of 5 mm.

Last but not least, the magnet system is of major importance since it provides momenta and charge measurements. A vast toroidal magnet system is embedded in the MS, comprising one barrel toroid and two end-cap toroids with eight coils each. The toroidal magnets deliver an inhomogeneous magnetic field of roughly 0.5 T and 1 T in the central and end-cap regions, respectively. The coils are made up of a mixture of aluminum, copper, niobium, and titanium and are cooled with liquid helium to 4.5 K. The muon p_T resolution of the MS is limited though, by the non-uniformity of the magnetic field.

3.2.5 Trigger and Data Acquisition

The high luminosity of the LHC produces numerous collisions per second resulting in a vast amount of data, corresponding to an expected output rate of 40 TB/s for ATLAS, which would be practically impossible to be recorded. Furthermore, most of these events are irrelevant to the main physics goals of the LHC. In order to handle the high event rates and reduce the amount of data to be recorded without losing important information, a series of requirements, called *triggers*, is placed on individual events determining whether a given event is recorded. The ATLAS trigger system [122] during Run 2 consists of a hardware-based Level-1 (L1) trigger and a software-based High-level trigger (HLT).

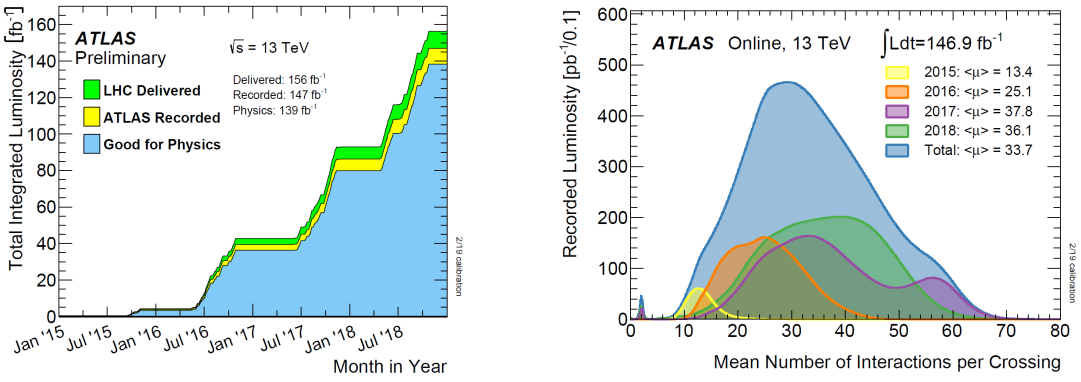
The L1 trigger performs the initial event selection based on information from the muon triggers (RPCs and TGCs) and the calorimeters to identify high p_T electrons, muons, photons, jets and high missing transverse energy (defined in Ch. 5). The measured signal is processed by the Central Trigger Processor (CTP), which then decides whether the event is collected or not. It also applies a preventive dead time to avoid overlapping read-out, or to keep the front-end buffer from overflowing. The L1 trigger has a very fast latency of 2.5 μs and reduces the event rate from 40 MHz to 100 kHz. Eventually, it identifies regions of interest (RoIs) in η and ϕ and passes this information to the HLT.

Then, the L1 trigger is followed by the HLT. The latter is fully software-based and uses the full detector information within the RoIs to reduce the event rate down to approximately 1 kHz, with a latency of 200 ms. In addition, a basic reconstruction is performed, including a reconstruction of tracks, charged particles, identification of jets from B-hadron decays, or a first rough computation of the missing transverse energy. This is only a crude fast reconstruction though, while more precise algorithms are used in the offline step discussed in Ch. 5. Finally, the data is transferred to a computing centre for further processing and storage.

The selection on objects is often accompanied by requirements on the quality of the reconstruction, which together with the kinematic requirements determine the average event rate. In case the event rates still too high, a prescale factor N can be applied, where only one out of N events is recorded. This allows for recording of events with a looser selection, but most of them are discarded randomly. In addition to the triggers, other requirements are set on each event to assure the quality of the events for data analyses. For instance, the detector has to be fully functional, while there should usually be a reconstructed primary vertex (Sec. 5.1.2).

3.3 LHC and ATLAS Data Taking

The analysis presented in this thesis utilises data from pp collisions $\sqrt{s} = 13$ TeV provided by the LHC and collected by the ATLAS detector during the second LHC data-taking period, started in the beginning of 2015 and completed in the end of 2018, referred to as *Run 2*. The evolution of the total integrated luminosity of the LHC Run 2 over the years is shown in Fig. 3.8a. In total, the amount of data delivered by the LHC during Run 2 corresponds to $\mathcal{L} = 156$ fb^{-1} [7], of which 147 fb^{-1} were actually recorded[8] by the detector. Due to various limiting factors, such as the availability and the performance of the detector, the actual luminosity usable for physics analyses is lower. Consequently, the total $\mathcal{L}$ collected with all subsystems of the ATLAS detector operational amounts to 139 fb^{-1} with an uncertainty of 1.7% [111]. In particular, the dataset for each individual year of Run 2 corresponds to $\mathcal{L}$ of 3.2 ± 0.1 fb^{-1} in 2015, 32.9 ± 0.7 fb^{-1} in 2016, 44.3 ± 1.0 fb^{-1} in 2017, and 58.5 ± 1.2 fb^{-1} in 2018 [111].



(a) **(b)**

Fig. 3.8: a) Cumulative luminosity versus, time delivered to ATLAS (green), recorded by ATLAS (yellow), and certified to be good quality data (blue) during stable beams for pp collisions at $\sqrt{s} = 13$ TeV during 2015-2018 period [120]. b) Luminosity-weighted distribution of the mean number of interactions per crossing for the 2015-2018 pp collision data at $\sqrt{s} = 13$ TeV, considering all data recorded by ATLAS during stable beams [120]. The recorded integrated luminosity and the average pileup $\langle\mu\rangle$ for each year and for all years combined are displayed.

The distributions of the mean number of interactions per bunch crossing, μ, for each year during Run 2 and for all the years combined are depicted in Fig. 3.8b. The pileup distributions differ among the years. Since the instantaneous luminosity (corresponding to the slope of the distribution in Fig. 3.8a) was constantly rising, more interactions per bunch crossing occurred. Exceptionally, some special low-μ runs performed in 2017 and 2018, demonstrated by the small peak at $\mu = 1$ in the figure. The average pileup over all bunch crossings over each year increased during the first years, during the ramp up of the collider. After all, the average pileup over all bunch crossings and over all years is $\langle\mu\rangle = 33.7$, higher than the designed one reported earlier. The increased pileup makes it more difficult to reconstruct the hard interactions of interest in the midst of many other concurrent pp collisions. This would affect the object reconstruction, if not identified and treated accordingly.

[7] The *delivered luminosity* is the luminosity delivered from the start of stable beams until the LHC requests ATLAS to put the detector in a safe standby mode to allow a beam dump or beam studies.

[8] The *recorded luminosity* reflects the DAQ inefficiency and the inefficiency of the "warm start", i.e. when "stable beams" is declared until the tracking detectors ramp up to high-voltage and the pixel system turns on preamplifiers.

Chapter 4

Particle Interactions and Simulation

In order to validate the SM, or its possible extensions, the theoretical predictions are compared to the observed data, collected by the detectors. Therefore, the simulation of the physics processes and the interaction of particles with the detector is of particular importance in high energy physics, so as to model the expected contributions from different background or signal sources. Additionally, simulated events are exploited to optimise the sensitivity of a process, such as the $t\bar{t}H(H \to b\bar{b})$ process studied in this thesis.

Computer programs known as *Monte Carlo* (MC) *event generators* are able to simulate events from defined physics processes. Pseudo-random numbers are generated to simulate particle collision events reproducing on average the predicted probability distributions, based on phase-space integrations of matrix element calculations. Furthermore, MC techniques are used to simulate the interaction of particles with the detector materials and the read-out of the detector. Apart from modelling the signal and background processes, MC techniques also provide theoretical uncertainties using the most up to date theoretical knowledge.

The simulation of *pp* collisions requires the description of physics processes including a wide range of energy scales. At high-energy scales, deep-inelastic scattering between partons occurs, calculated in perturbative QCD. In contrast, at low energies, the evolution of partons into stable hadrons takes place, which cannot be calculated perturbatively. Thus, it is determinant to factorise the different energy scales involved in the process. The simulation of the hard interaction can be computed up to a fixed order in perturbation theory, while the description of the softer scales can be done with phenomenological models. The first step of the simulation of *pp* collisions is the event generation, which is further divided into several steps due to the complexity of physics processes. Afterwards, the simulation of the detector response follows.

4.1 Treating high-order divergences

The matrix element of a scattering process is computed as a series in perturbation theory, as introduced in Sec. 2.3. So far, only tree-level processes have been discussed, which are described by the first-order term in perturbation series. All such processes receive higher-order contributions though, known as *radiative corrections*, from diagrams that contain additional particle loops, i.e. intermediate states of virtual particles. In each of these diagrams, there is one "virtual" particle whose momentum is not determined by the four-momentum conservation at the vertices. Since perturbation theory requires to sum over all possible intermediate states, the integration over all possible values of this momentum is needed. However, the loop-

momentum integrals turn out to be infinite.

The evaluation of diagrams containing loops, which correspond to higher-order terms, involve integrals over the four-momenta of intermediate (virtual) particles. These integrals are often divergent in the loop-momentum $k \to \infty$ or *ultraviolet region*. These divergences can be regularised and absorbed by a redefinition of physical quantities of a theory (such as masses, coupling constants, as well as fields), a procedure called *renormalisation* [15, 124].

After ultraviolet renormalisation, higher-order perturbative QCD contributions still contain divergences in the *infrared region*, when the loop-momentum $k \to 0$. They arise either from real emissions of soft or collinear partons, or from soft or collinear configurations of momenta in virtual loops. Observables, like cross sections, can finally obtain finite expression by factorising into an infrared safe component, describing short-distance interactions at very high-energy scales, and a non-perturbative infrared singular component, describing long-distance physics at low-energy scales. This procedure accounts for the *factorisation theorem* [125].

In the following sections we will see that, an arbitrary renormalisation scale μ_R and a factorisation scale μ_F are introduced to scale the finite set of parameters in the QCD theory to counteract divergent contributions. To avoid unnaturally large logarithms reappearing in the perturbation series, it is sensible to choose μ_F and μ_R values of the order of the typical momentum scales of the hard-scattering process, while $\mu_F = \mu_R$ is also often assumed.

4.1.1 Renormalisation

As a first step in the renormalisation procedure, a cut off is imposed on the loop momentum at some large but finite momentum M, in order to regularise the divergent integrals. However, such a theory cannot describe physics at asymptotically high energies ($M \to \infty$). In order to eliminate the dependence of physical quantities on this arbitrary parameter M, but also to remove the infinities of a theory, renormalisation requires the introduction of a reference scale. This is an arbitrary and unphysical scale called *renormalisation scale*, μ_R, on which the physical quantities depend. By subtracting the expression of a physical quantity as a function of the renormalisation scale from that of the physical scale, the final expression is finite and independent of the cut off M. In this case, the theory is said to be *renormalisable*. Moreover, different choices of μ_R lead to different expansions of an observable, i.e. to different *renormalisation schemes*. In general, physical quantities must be independent of the arbitrary scale μ_R. Nevertheless, the truncated perturbation series introduces a scale dependence of the approximate result. To reduce the influence of higher-order corrections, for a scattering process the physical quantities are evaluated at the energy scale of the process.

4.1.2 The factorisation theorem

In *pp* collisions, interactions among the components of protons occur, the partons. The primary scattering of the partons, known as the *hard* (*scattering*) *process*, happens at very high-energy scales. On the contrary, many low-energy-scale processes take place inside the proton. A *factorisation scale* μ_F is employed to allow for the separation of low-energy dynamics within the proton from the high-energy dynamics of the hard process.

The highly-energetic hard scattering $pp \to X$ is illustrated in Fig. 4.1. The incoming partons of type (a, b), that constitute the proton, are essentially free, carrying a momentum fraction x_a,x_b of the proton momentum p_a,p_b, respectively. Each of the colliding partons is

described by a *parton density function* (*PDF*) (defined in Sec. 4.2.1). According to the *factorisation theorem* [125, 126], the cross section for this process can then be calculated as a convolution of the PDFs $f_a(x_a, \mu_F^2)$,$f_b(x_b, \mu_F^2)$ at the energy scale μ_F^2 of the interaction, and the cross section of the hard interaction of free partons $\hat{\sigma}_{ab\to X}$

$$\sigma_{pp\to X} = \sum_{a,b} \int dx_a dx_b f_a(x_a, \mu_F^2) f_b(x_b, \mu_F^2) \hat{\sigma}_{ab\to X}(x_a p_a, x_b p_b, \mu_R^2, \mu_F^2). \tag{4.1}$$

In accordance with the process, collinear QCD splittings up to different μ_F scales are absorbed in the PDFs. Depending on μ_F, the proton momentum can, thus, contain large contributions from quarks and gluons in addition to the three valence quarks. Their contribution (x_a, x_b) in turn depends on the mass of the produced particle with respect to the centre-of-mass energy of the colliding protons, as shown in Sec. 4.2.1.

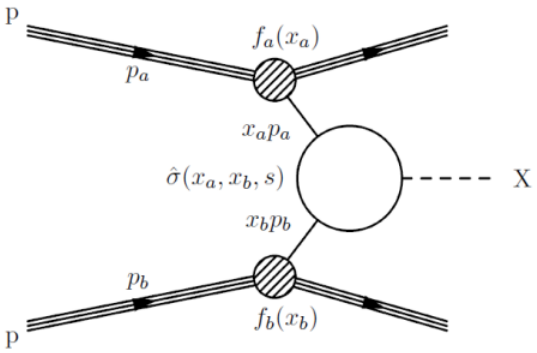


Fig. 4.1: Illustration of a generic hard scattering process. The partons, emerged from the colliding *pp* pair, carry a momentum fraction with respect to the proton energy, described by a PDF. The scattering of the partons is computed perturbatively. Thus, the kinematic properties of the final state object *X* can be predicted.

However, the partonic cross sections show collinear divergences connected to long-distance soft interactions. These divergences are factored out and absorbed into a redefinition of the PDFs introducing a factorisation scale μ_F, that separates the perturbative and non-perturbative effects. In the kinematic region above μ_F, the emissions are treated as part of the short-distance hard scattering and are described by perturbation theory. Conversely, interactions with energy below μ_F are included in the PDFs, describing the non-perturbative long-distance soft physics. Perturbative calculations are then used for a given process, such as the production and subsequent decay of top quark pairs in *pp* collisions. According to eq. 4.1, the sum of these processes is the total hard-scatter cross section.

In order to reduce the impact of higher-order corrections, the factorisation scale is often set to the same value at which α_s is evaluated, namely to the renormalisation scale. As a result, the energy scale Q of the investigated process is determined. For the simulation of the analysis presented in this thesis, μ_F and μ_R are set to the mass or the transverse momentum of the final state system, and are presented in Sec. 4.5 for each process.

4.2 Event generation and simulated samples

A *pp* collision is a complicated physics process because of the composite internal structure of hadrons, as well as of the sub-processes arising from the interactions of the final-state particles

with the detector and with each other. They consist of several components, that describe the physics from very short-distance scales, up to the typical scale of hadron formation and decay. Since QCD is weakly interacting at short distances, the components of simulation dealing with short-distance physics, such as the hard interaction, are computed up to a fixed order in perturbation theory. At larger distances, all soft hadronic phenomena, like the hadronisation, cannot be computed perturbatively, thus follow QCD phenomenological models.

All these physics processes are simulated by MC software, producing samples according to both theoretical and phenomenological models. The MC event generation of the simulation of pp collisions, is disaggregated in several steps, depicted in Fig. 4.2, which are outlined below. In the end, all the information from the event generation is stored in the MC history and the particles it contains are referred to as *true particles*.

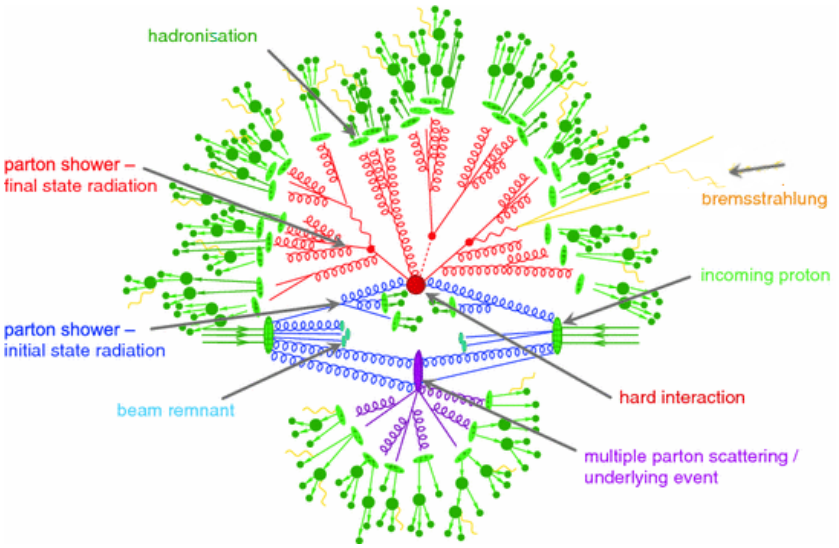


Fig. 4.2: Representation of a hadron-hadron collision event containing all the factorised stages in the MC event generation chain [127].

In Fig. 4.2, the two incoming protons of the pp collision are represented by big green blobs, and the momenta of the three partons are depicted by the continuous green lines. As already discussed, the extraction of the partons from the protons is a non-perturbative process and thus cannot be computed explicitly with the SM. Therefore, the modelling of the partons inside the proton can be separated from the actual interaction. During a collision, the partons transfer a fraction of the proton's momentum, described by the PDFs.

In the first step, the scattering probability of the hard scattering process (red blob), such as $pp \rightarrow t\bar{t}H$, is calculated through the evaluation of the matrix element at a fixed order in perturbation theory, given the high-energy scale of the interaction. The outgoing partons (particles going out of the red blob) are randomly distributed in the available phase space. Since the partons involved in the collision are colour charged they will emit gluons, apart from photons, due to bremsstrahlung. The gluons in turn radiate further gluons or split into quark/anti-quark pairs, forming parton showers. The initial (blue lines) and final state (red lines) parton showers are simulated by appropriate parton shower algorithms.

The radiation process continues until the partons reach an energy scale of Q $\approx$ 1 GeV. At this stage hadronisation takes place, and partons recombine into collimated bunches of colourless hadrons (light green blobs). These hadrons further decay into the final state parti-

cles (dark green blobs) that interact with the detector, leaving energy deposits referred to as jets (discussed in Sec. 5.2). There is also soft photon radiation (yellow wavy lines) coming from the hadron decays. Phenomenological models are used to describe the hadronisation process and the decay of hadrons. Finally, other final state partons, produced from secondary interactions between other partons of the protons (purple blob) involving smaller momentum transfers, and remnants (light blue blobs) are considered, forming the underlying event.

4.2.1 PDFs and DGLAP Equations

As already introduced, during *pp* collisions at the LHC, interactions among the components of protons occur. A proton consists of three *valence quarks*, namely two up and one down quark (*uud*). Additionally, there are gluons, mediating the strong force between the valence quarks, keeping them in a bound state. These gluons can form virtual quark-antiquark pairs, called *sea quarks*. Partons constituting hadrons behave as asymptotically free particles at high energy, where a perturbative description is applied.

All partons within a proton carry a part of its momentum. The momentum distribution of the partons inside the proton follows a probability distribution called *parton density function* (*PDF*) [14]. Thus, the PDFs are a measure for the probability of observing a certain parton within the proton, carrying a fraction *x* of the whole proton momentum at a certain energy scale (Q^2). Because of the inherent non-perturbative nature of partons which cannot be observed as free particles, PDFs cannot be derived from perturbative QCD calculations. Instead, they are measured from several hadron colliders and deep inelastic scattering experiments such as H1 and ZEUS at the electron-proton HERA collider [128, 129].

The PDFs depend on the energy of the proton. However, they can be measured at a certain energy scale Q^2 and extrapolated to the energy regime of interest. The energy dependence of the PDFs is described by the *Dokshitzer–Gribov–Lipatov–Altarelli–Parisi* (*DGLAP*) *evolution equations* [130–132] as

$$\begin{aligned}\frac{\partial q_i(x,Q^2)}{\partial \log Q^2} &= \frac{\alpha_s(Q^2)}{2\pi}\int_x^1 \frac{dz}{z}\left\{P_{q_iq_j}(z,\alpha_s(Q^2))q_j(\frac{x}{z},Q^2) + P_{q_ig}(z,\alpha_s(Q^2))g(\frac{x}{z},Q^2)\right\}\\ \frac{\partial g(x,Q^2)}{\partial \log Q^2} &= \frac{\alpha_s(Q^2)}{2\pi}\int_x^1 \frac{dz}{z}\left\{P_{gq_j}(z,\alpha_s(Q^2))q_j(\frac{x}{z},Q^2) + P_{gg}(z,\alpha_s(Q^2))g(\frac{x}{z},Q^2)\right\}\end{aligned} \tag{4.2}$$

where $q_i(x,Q^2)$ and $g(x,Q^2)$ is the quark and gluon PDF, respectively. Also, $P_{ab}(z,a_s(Q^2))$ are the *Altarelli-Parisi splitting functions* [132, 133] that can be expanded in powers of the running coupling.

PDFs are a key ingredient in the MC simulation of *pp* collisions at the LHC. Various collaborations constantly work to improve the PDF fits using the most recent data. Experimental data from the LHC can also be included in these fits. The PDF sets that are commonly used at the LHC and in the analysis presented in this thesis are NNPDF [135] and CTEQ [134]. Since the PDF groups use slightly different assumptions for the DGLAP equation, different groups are used to estimate the theoretical uncertainty. An example of the proton NNPDF3.1 PDF set is shown in Fig. 4.3. It includes previous LHC among many other datasets and is calculated at NNLO accuracy for two different factorisation scales.

According to both examples in Fig. 4.3, the proton momentum contains large contributions from quarks and gluons in addition to the three valence quarks. The valence quarks dominate at large values of x, with the contribution of the u-quark being twice as large as of the d-quark. The sea partons increase towards low values of x, with the gluon contribution being

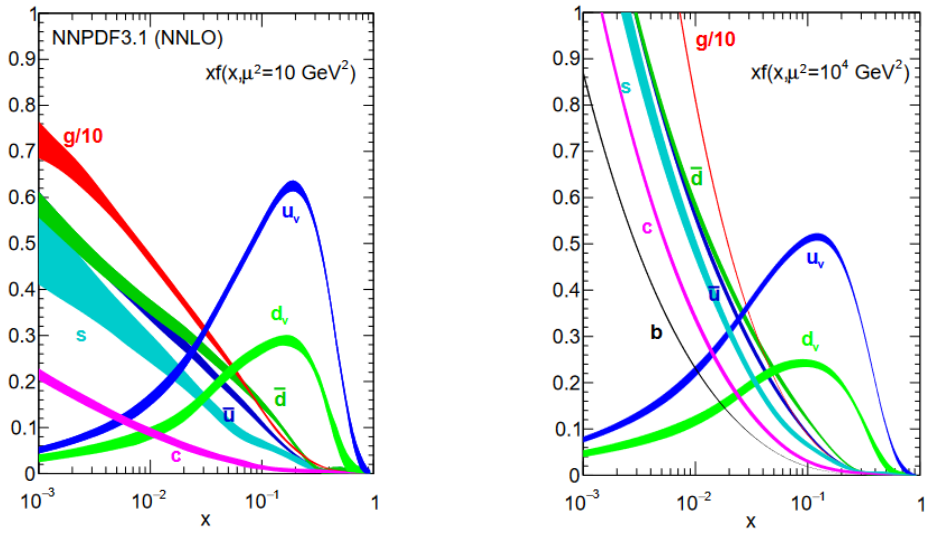


Fig. 4.3: Example of NNLO PDFs for valence and sea quarks, as well as gluons, as a function of the fractional parton momentum x at a factorisation scale of a) 10 GeV^2 and b) 10^4 GeV^2. The results are from NNPDF3.1 set using also data from the LHC [135].

an order of magnitude higher than that of quarks. At the high energy scale (Fig. 4.3b), the fraction of gluons, anti-quarks, and heavier quark flavours that carry a significant fraction of the proton momentum increases with respect to the low energy scale (Fig. 4.3a) because higher-energetic QCD splittings are included in the PDF. The contribution of the partons to the proton momentum depends on the energy scale at which the process takes place, like the centre-of-mass energy of the colliding protons, as well as on the mass of the produced particle.

Due to the composite structure of the proton, multiple production modes of a process or a particle are possible. For instance, there are various production mechanisms for the Higgs boson (described in Sec. 2.4.1). The energy scale $Q^2 = 10^4$ GeV in Fig. 4.3b corresponds to the typical momentum transfer of the Higgs boson production at the LHC. Then, at 13 TeV, the most likely values of x of the incoming partons are around 10^{-2} (assuming symmetric collisions). So, the Higgs boson is dominantly produced through gluon fusion, due to the overwhelming presence of gluons at these energies, while the valence quarks play a very minor role. Generally, the order of the PDF calculation should be equivalent to the order of the matrix elements used in the hard process part of the MC calculation. Analogously, although the $t\bar{t}$ as well as the $t\bar{t}H$ process can be produced via gluon fusion or $q\bar{q}$ annihilation (see Figs. 2.11a-2.11d and 2.7a-2.7c, respectively), the gluon fusion is favoured at the LHC.

4.2.2 Matrix Element

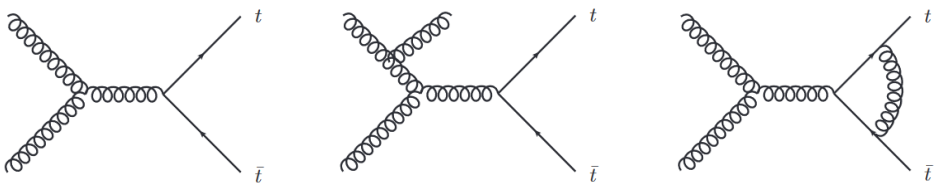
The hard process is characterised by a large invariant mass or large momentum transfer. As it happens at high energy scales, it can be calculated in perturbative QCD. The simulation of the hard scattering depends on the calculation of the matrix element (ME), that describes the transition from an initial to a final state. As discussed in Sec. 2.3, the matrix element accounts for the underlying mathematical description of Feynman diagrams.

In order to compute the ME of a certain process, a perturbation series corresponding to

all possible Feynman diagrams, representing the process, has to be evaluated. Thanks to the factorisation introduced earlier, the description of the hard scattering can exclude the protons and be modelled only as an interaction of the component partons. Also, particles sensitive to the strong interaction tend to emit additional particles, either through gluon emissions, or gluons splitting into $q\bar{q}$ pairs. As a result, the total inclusive cross section for producing any final state (X) from a parton collision is given to all orders in perturbation theory from (starting from eq. A.9)

$$\hat{\sigma}_{ab\to X} \sim \sum_{k=0}^{\infty} \int_{n+k} \delta^4(p_a + p_b - \sum_{f=1}^{n+k} p_f) \left| \sum_{l=0}^{\infty} \mathcal{M}^{(l)}_{n+k}(p_a, p_b \to p_f) \right|^2 \prod_{f=1}^{n+k} \frac{d^3\vec{p}_f}{2E_f}, \tag{4.3}$$

where n is the number of particles in final-state X, k denotes the number of additional "real emissions", l represents the number of "virtual correction" (loops). Also, $\mathcal{M}^{(l)}_{n+k}$ is the matrix element corresponding to the sum of the Feynman diagrams with l loops and $n+k$ final-state particles. Figure 4.4 shows an example of three Feynman diagrams for a $t\bar{t}$ final state at tree level ($k = 0$, $l = 0$), including a first emission ($k = 1$, $l = 0$), or a virtual correction ($k = 0$, $l = 1$).



(a) **(b)** **(c)**

Fig. 4.4: Example of Feynman diagrams of a $t\bar{t}$ production a) at leading order, b) for a first real emission, and c) for a first virtual correction.

As the number of Feynman diagrams is infinite, an exact calculation of eq. 4.3 is not possible analytically. Nevertheless, it uses a perturbative approach and different order of precision can be used to achieve a good approximation, mainly depending on the complexity of the final state. A leading order calculation is determined by $k = l = 0$, so that the higher order Feynman diagrams with additional emissions and loops are not included. Analogously, the full NLO calculation includes diagrams with $k = 1$, as well as interference with loop diagrams $l = 1$.

The hard scatter is the only step which purely relies on the theoretical predictions. In order to generate events, the matrix element of a process is usually calculated at a fixed order of the relevant coupling constant in perturbation theory, mainly at LO and NLO. The aforementioned calculations can be performed at NLO accuracy with various MC generators, each solving differently eq. 4.3 with respect to real emissions and loop corrections. POWHEG BOX [151–153] is a NLO parton-level event generator computing matrix element in perturbative QCD using the Powheg method [150]. MADGRAPH5_AMC@NLO [156–158] is a MC generator known for the automated computation of the matrix element at LO and NLO. The NLO calculation depends on the MC@NLO method [155], while the LO one is provided by MADGRAPH5. One of the main differences is that the former method generates events only with positive weights, while the latter allows also the possibility of negative weights. SHERPA [162] is a NLO/LO multi-purpose MC generator used for many final states. It can be interfaced with additional libraries to compute loop amplitudes. SHERPA interfaced with the OPENLOOPS library [164]

is used to model the $t\bar{t} + b\bar{b}$ process at NLO which constitutes the largest background for the analysis presented in this thesis.

4.2.3 Parton Shower

As already introduced, since the hard process involves large momentum transfers, the partons involved in it are violently accelerated, which, being coloured particles, emit QCD radiation in the form of gluons. Also, given that gluons themselves carry colour charge, they can emit further radiation or split into quark/anti-quark pairs, leading to cascades of partons called *parton showers* (PS). This is a perturbative process, which is alternatively called *fragmentation*. When the gluons are radiated off partons in the initial state before the hard scatter (see Fig. 4.2), it is referred to as *initial-state radiation* (ISR), while when they are radiated off final-state particles, it is called *final-state radiation* (FSR).

The parton shower is included in the MC simulation to approximately account for the higher-order corrections to the hard process, emulating a complete final state. In particular, the parton shower generators simulate the successive emission of quarks and gluons from the partons in the initial and final states. Given that the subsequent radiations are of lower energy compared to the couplings considered in the matrix element, the coupling constant becomes large. Therefore, a perturbative approach is insufficient to model the parton shower. Instead, an approximation scheme is used, described by the Altarelli-Parisi splitting functions [133, 135] and Sudakov form factor [101].

The parton shower corrections to the hard-process cross section are predicted by considering the dominant contributions at each order in perturbation theory. These contributions are associated with the collinear (small-angle) parton splitting in the direction of the parent parton (*collinear splitting*), or the soft (low energy) gluon emission (*infrared radiation*). The possible processes for QCD emission or splitting that can occur are $g \to gq$, $g \to gg$, $g \to q\bar{q}$. Assuming the *n*-parton differential cross section before splitting $d\sigma_n$, then after the splitting, for $n+1$ particles, it becomes

$$d\sigma_{n+1} \approx d\sigma_n dP_{ji}(z) \approx d\sigma_n \frac{\alpha_s}{2\pi} \frac{dq^2}{q^2} dz P_{ji}(z) \tag{4.4}$$

at LO, where $P_{ji}(z)$ is the $i \to j$ *splitting function*, which describes the distribution of the fraction of momentum *z* of the parton *i* carried by the parton *j*. Additionally, q^2 is the evolution variable of the parton shower, that can denote the squared virtual mass of the partons in the shower, also called *virtuality*. Except for the virtuality, the opening angle between the split partons (θ) can be also used as the evolution variable, by applying $\frac{dq^2}{q^2} = \frac{d\theta^2}{\theta^2}$. The simulation algorithm develops the shower from each parton involved in the hard process, by applying eq. 4.4 iteratively. The upper limit on the initial virtuality is set by some momentum transfer scale *Q* of the hard process, $q^2 < Q^2$, and the shower is terminated when the virtualities have fallen to the *hadronisation scale*, $q^2 = Q_0^2 \approx 1$ GeV^2 .

Apart from the collinear real parton emissions, the parton shower approximation takes also into account virtual (loop) effects of the same order in perturbation theory. They are included in the probability of a parton *i* evolving from an initial scale q_1^2 to a lower scale q_2^2 without splitting, i.e. without emitting QCD radiation above a certain scale, which is given by the *Sudakov form factor* [101]

$$\Delta_i(q_1^2, q_2^2) = e^{-\sum_i \int_{q_2^2}^{q_1^2} \int_{z_{min}}^{z_{max}} dP_i(z,q^2)}. \tag{4.5}$$

In that way, additional emissions at increasingly lower energies can be added to the final state of the hard process. Similarly, initial-state radiation [136] is added to the event, with the additional complication that the way initial-state splittings are accounted for in the PDF has to be considered.

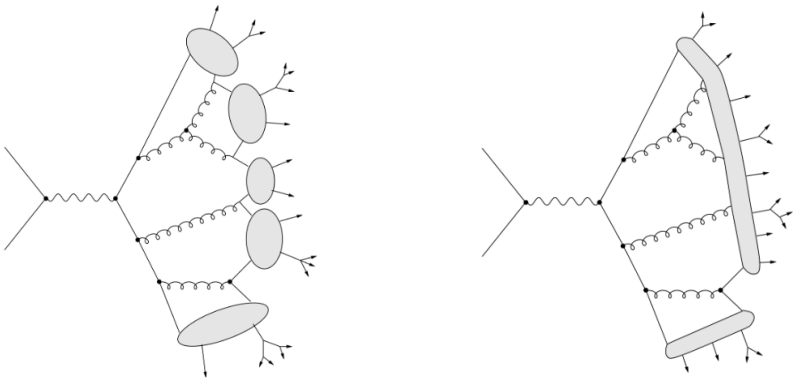
The showering relies on theoretical predictions tuned to data. Various multi-purpose MC generators are used in this thesis for the parton shower simulation. PYTHIA [159] is an event generator, that uses PS with emissions ordered in virtuality or in transverse momentum. HERWIG [160, 161] is another MC generator, which uses PS with emissions ordered in opening angle that includes colour-coherence effects with special description of radiation from heavy particle. Both PS generators are interfaced with the ME generators. Last but not least, the SHERPA generator contains its own parton shower algorithm, based on the Catani-Seymour dipole factorisation formalism [162, 163], thus it does not need interfacing.

The parton shower algorithms are based on a combination of the collinear and soft contributions and are thus inaccurate for hard and large-angle emissions, which are generated as part of the ME. In that sense, the PS algorithms are complementary to the simulation of particles from ME calculation. Besides, the diagram with an NLO emission of a particle is the same as an LO ME with additional emissions from the PS. So, there is an overlap between emission from the ME and the PS though, which has to be removed to avoid double counting. Therefore, the phase space covered by the ME calculation, and the space covered by the PS evolution needs to be separated. This is achieved by the so-called *ME-PS matching algorithms* [137]. The ME-PS algorithms define a transverse momentum and angular cutoff above which additional radiation is simulated by the ME calculation, thereby avoiding soft and collinear divergences. Afterwards, these events are showered and clustered with a jet-clustering algorithm with an angular resolution in the order of the angular cutoff. An event is only used if no jets in addition to the ones simulated in the ME, and above the p_T threshold, are generated in the shower. This ensures that only soft and collinear radiation is simulated in the PS, below the cutoff. In terms of this analysis, the Catani-Krauss-Kuhn-Webber (CKKW) [165] matching scheme is used, which is based on the k_T clustering algorithm [166].

4.2.4 Hadronisation

As already discussed in Sec. 2.2.4, the coupling α_s increases rapidly at low energies and large distances. In particular, at the energy scale of the order of Λ_{QCD} or smaller the QCD confining effects become important, while perturbation theory becomes invalid. Colour confinement (Sec. 2.2.4) forbids the colour-charged particles, produced in the hard scattering process and in the parton shower, to propagate freely. Instead, individual partons bind into colourless baryons and mesons, i.e. hadrons. Thus, this process is called *hadronisation* and the energy scale that it starts to happen, $Q_0^2 \approx 1$ GeV^2, is referred to as *hadronisation scale*.

As the partons evolve and radiate, the value of the parton shower evolution scale Q^2 decreases bringing the parton virtuality below the hadronisation scale. At this point, the parton shower is terminated, while the hadronisation process starts to evolve, leading to the formation of the observed final-state hadrons. These hadrons might be excited and also decay into many lower-energy states. The hadronisation process is not amenable to the currently available non-perturbative techniques for calculation, therefore event generators have to rely on phenomenological models based on general features of QCD. Hadronisation is typically simulated through either the cluster hadronisation model [140, 141] or the Lund string fragmentation model [138, 139].



(a) **(b)**

Fig. 4.5: Illustration of the a) cluster hadronisation model where individual colour-singlets are considered individually, and the b) Lund string fragmentation model.

The *cluster hadronisation model* relies on the concept of preconfinement [142]. After the parton cascade, the preconfinement starts with the non-perturbative splitting of final-state gluons into colour-singlet $q\bar{q}$ pairs, as illustrated in Fig. 4.5a. Colour-singlet combinations are then grouped into colourless clusters of partons, with a mass of a few GeV, evaluated as excited hadron resonances. Especially, clusters with a mass below ~ 3 GeV decay into hadrons through a two-body decay. Unstable hadrons further decay, until the final state consists of stable[1] particles. HERWIG event generator models the hadronisation based on the cluster model. Finally, SHERPA generator has its own hadronisation model based on the cluster model.

In the *Lund string fragmentation model*, as illustrated in Fig. 4.5b, the confinement between partons, induced by the colour force, is represented by a gluonic string of ~ 1 fm. For a $q\bar{q}$ pair, as the quarks move apart, the string is stretched and its potential energy grows because of confinement. When the energy becomes of the order of hadron masses, it becomes energetically favourable for the string to break and create a new $q\bar{q}$ pair. Likewise, the two string segments begin to stretch and eventually break again, and it proceeds iteratively until all the energy has been converted into $q\bar{q}$ pairs connected by short string segments and only hadrons with on-shell mass remain. Moreover, radiated gluons are considered as kinks along the string, carrying momentum. PYTHIA MC generator models the hadronisation based on the Lund string model.

EVTGEN [145], is a MC event generator that runs after the parton shower and hadronisation, which are simulated by the above mentioned PYTHIA or HERWIG generators. EVTGEN simulates the decays of heavy flavour particles, such as B-mesons. In particular, it includes detailed models for semi-leptonic or CP-violating decays, while it produces accurate results for angular distributions in sequential decays, including all correlations.

4.2.5 Underlying Event

The hard interaction occurs between two partons of the incoming protons. The remaining quarks and gluons from the colliding hadrons are also subject to secondary soft interactions, referred to as *multi-parton interactions* (MPI) [143], generating further multiple distinct scat-

[1] As *stable* (also called *long-lived*) is considered a particle that does not decay before reaching the detector. Typically in hadron colliders, this is a particle with a decay length $c\tau \geq 1$ cm [101].

ters. The products of MPI are colour-charged as well, thus they also undergo hadronisation. MPI together with other non-interacting beam remnants create additional activity, known as the *underlying event* (UE). These effects are usually not of direct interest for an analysis. Nevertheless, they leave traces in the detector and have to be considered in the simulation.

Due to the low energy scale of the UE processes, their modelling relies on phenomenological models with free parameters, and require extensive tuning based on experimental data [144]. PYTHIA and HERWIG MC generators use multiple-interaction models for the simulation of the underlying event. Finally, SHERPA uses a multiple-interaction model based on that of PYTHIA.

4.3 ATLAS detector simulation

The output of the MC generators is a list of four-vectors of all stable particles produced in the event, after hadronisation and decay of the intermediate unstable particles. This output is used so as to study the physics processes at the so-called stable *particle level*. In order to compare it with the recorded data, the MC has to be analysed after the reconstruction in the detector, i.e. at the *reconstruction level*. Therefore, it is important to account for the detector acceptance defined by the geometry and the resolution of the different sub-detectors, which affects all data collected by ATLAS. For this purpose, a detailed simulation of the effect of the interactions between the particles and the detector materials is required. These effects, accounting for the finite efficiency and resolution of the various technologies and utilised in recording the particle collisions, are applied to the final state particles output from the MC generators.

The detector simulation software, based on the GEometry ANd Tracking (GEANT4) framework [147], models the interaction of particles and their decay products with the ATLAS detector and all its components. For the simulation, energy deposits are converted into simulated electronic signals taking into account the detector geometry and the response of the readout electronics. However, the detector simulation is a computationally complex process, being highly CPU intensive, and is often the dominant contribution to the total generation time of samples. Especially, the development of a particle shower in the calorimeter system of the ATLAS detector requires the largest amount of computing resources to be simulated. It can last for several minutes per event for typically several millions of events per sample. This method is referred to as *full simulation* [146].

The large amount of time needed for the simulation motivates the use of an alternative, faster, and less refined simulation algorithm, called *fast simulation* (AFII) [148, 149]. Each of the three sub-detectors (discussed in Sec. 3.2) simulation times can be reduced. Nevertheless, the main contribution to the simulation time comes from the shower of particles in the calorimeters. Thus, the fast simulation imposes a parametrised description of the particle shower shapes in the calorimeters. As a result, the CPU time required to process the events is considerably reduced, at the order of a magnitude or more, at the cost of a poorer description of the calorimeter response. This effect can be mitigated via dedicated calibration of fast simulation MC samples, resulting in good agreement between data and MC at the level of the high-level physics objects (discussed in Ch. 5).

In general, the full simulation provides a higher precision and is favoured as the main simulation method used in the production of MC samples of the analyses. By contrast, the fast simulation method allows to produce multiple alternative samples, that are compared to the

nominal samples in optimisation studies, or assess theoretical systematic uncertainties.

4.4 Monte Carlo corrections

Despite all the efforts and expertise on the simulation, differences between data and MC simulations could arise. These differences are not always reducible to detector effects, but they originate from the approximated description of a process to be studied. In case of discrepancies that are common between different experiments, the simulation can be corrected empirically through appropriate weights, that take into account this variation and decrease or eliminate the differences on the process.

In order to form an accurate description of the detector effects including reconstruction and identification of physics objects, the simulated MC event samples are compared to data and corrected with multiplicative scale factors (SFs), defined as

$$SF = \frac{\epsilon_{data}}{\epsilon_{MC}}, \tag{4.6}$$

where ϵ_{data} and ϵ_{MC} are measured in dedicated data calibration samples and in the equivalent MC simulation, respectively. For example, the energy scale and resolution of the different physics objects in the simulated MC events are corrected to match the corresponding data measurements.

Another necessary correction so that the MC events match the data is the normalisation of the recorded luminosity (defined in Sec. 3.1.1). Furthermore, pileup effects (see Sec. 3.1.1) need to be considered. Hence, the simulated events are weighted to match the expected number of interactions per bunch crossing, μ, in real data-taking conditions. Lastly, MC samples are corrected to reproduce the best known theoretical cross section, usually at NLO or NNLO, even when they are produced with a lower order MC generator.

4.5 Signal and background modelling

The analysis described in this thesis studies the associated production of the Higgs boson with a $t\bar{t}$ pair ($t\bar{t}H$), with the Higgs boson decaying in $b\bar{b}$ pair and the $t\bar{t}$ system decaying semi-leptonically (lepton + jets, see Fig. 2.15a). The background sources affecting the $t\bar{t}H$ channel emerge from processes with a final signature that resembles the $t\bar{t}H$ one. It may depend on the similarity of the decay products, or on the not negligible possibility of objects misidentification. The main physics process that contributes to the background composition of this analysis is the $t\bar{t}$ + jets. In addition to the main background, there are smaller contributions from the associated production of a vector boson and a $t\bar{t}$ pair ($t\bar{t}+V$, $V = W, Z$), as well as non-$t\bar{t}$ processes. The latter can be the production of a single top, followed by the production of a vector boson in association with jets (W/Z+jets), diboson (WW, WZ, ZZ) production, tH production, or other rare top-quark processes.

The $t\bar{t}$ + jets is the dominant process among all the background processes since it has the largest cross section, which is significantly higher than that of signal. Therefore, it constitutes the overwhelming background of this analysis, so its precise estimation is crucial. Especially for the $t\bar{t}$ + jets background, a heavy flavour classification, which categorises jets based on the particle they originate from, is also defined. The additional quarks that the jets stem from can arise from QCD radiation or loop-induced QCD processes. The final contribution

of the various background sources to the total background events of the analysis depend on the event selection criteria that define the analysis regions (discussed in Sec. 6.4). After the event categorisation, it becomes apparent that the $t\bar{t} + b\bar{b}$ process (Fig. 2.15c) is the dominant background component. It is a $t\bar{t}$ process with two additional *b*-quarks coming e.g. from an emission of a gluon which further splits into two *b*-quarks. The $t\bar{t} + b\bar{b}$ background remains almost irreducible with respect to the $t\bar{t}H(H \to b\bar{b})$ process though, since both processes have four b-quarks in the final state. Therefore, the modelling of this background is a challenge for the analysis.

The $t\bar{t}H(H \to b\bar{b})$ analysis, as many other studies on the Higgs boson, start from final-state signatures which include b-quarks. Thus, the accurate theoretical prediction of fixed-order perturbative calculations of high-energy processes, which involve the production of b-quarks, is required. Depending on the way the *b*-quarks in the initial state are treated, the proton can either be modelled in the four-flavour scheme (4FS) or in the five-flavour scheme (5FS) [123].

The 4FS assumes the proton to contain quarks of the four lightest flavours (up, down, strange, and charm). Since *b*-quarks are significantly heavier than the proton, in this case they are only created in pairs by gluon splittings. Consequently, there is no *b*-flavour PDF describing the proton, thus no *b*-quark contributes to ME in the initial state. So, b-quarks decouple from the QCD perturbative evolution and thus also from the running α_s, where the number of quark flavours is set to $n_f = 4$ in eq. 2.12. However, it can only be generated as a massive final state, impacting calculations at lower energy scales. In this approach, the gluon splitting is described as part of the hard-scattering process, allowing the ME computation to cover the full $t\bar{t} + b\bar{b}$ phase space hence, the kinematic properties of the *b*-quark are described more accurately.

On the contrary, at high energy scales the mass effects are negligible. This case is described by the 5FS, in which the initial-state *b*-quarks are considered as partons inside the proton. Therefore they are treated as massless, as the other light quarks, comprising a b-quark PDF and $n_f = 5$ is set to the running α_s. In this case, the gluon splitting is regarded as part of the PS, while this description does not account for the full phase space, so as to avoid divergences from soft and collinear emissions. Although the 5FS does not account properly for the mass of the *b*-quark, it results in easier calculations since there is one less final-state quark.

Simulated event samples, obtained with MC event generators, are used in the analysis in order to construct the nominal model of the $t\bar{t}H$ signal and the background processes. They are also used to calculate detector acceptance, as well as to train the Boosted Decision Trees (outlined in Sec. 6.5.3). Furthermore, in order to estimate the systematic modelling uncertainties, either variations of the nominal model are produced by tuning the various parameters of the nominal samples, or alternative samples are generated. All simulated samples, i.e. the nominal samples used for the baseline modelling of this analysis, as well as the alternative ones used to estimate systematic uncertainties, are outlined below (also listed in Table A.1 in Appendix). The choice of the generators and the various settings for the generation of the nominal samples is made, in principle, based on what describes better each physics process, according to the latest measurements from the corresponding analyses. Also, the agreement of the simulated events of a sample with the measured data play a role in the final choice, as well as the available statistics of a sample.

4.5.1 Common treatment in MC samples generation

As discussed in Sec. 4.3, the MC samples are produced using either the full ATLAS detector simulation based on GEANT4, or the fast simulation where the full simulation of the calorimeter

response is replaced by a detailed parametrisation of the shower shapes. For the observables used in the analysis, the two simulations are found to give similar modelling.

In order to simulate the effects of pileup (see Sec. 3.1.1), additional interactions are generated using PYTHIA 8 [159] with a set of tuned parameters, the A3 tune [167], and overlaid onto the simulated hard-scattering event. Simulated events are reweighted to match the pileup conditions observed in the full Run 2 dataset, with an average pileup over all bunch crossings over all years during Run 2 $\langle\mu\rangle \simeq 34$ (see Sec. 3.3). All simulated events are processed through the same reconstruction algorithms and analysis chain as the data.

The precision of the ME generators is NLO in QCD for most samples. However, some samples are normalised to higher precision in QCD (NNLO) or with electroweak (EW) corrections. For all samples generated using MADGRAPH5_AMC@NLO at NLO in QCD for the ME, the shower starting scale has the functional form $\mu_q = \frac{H_T}{2}$ [171], where H_T is defined as the scalar sum of the p_T of all outgoing partons.

Furthermore, in all samples where the PS, hadronisation, and MPI are generated with either PYTHIA 8 or HERWIG 7 [161], the decays of *b*- and *c*-hadrons are simulated using the EVTGEN 1.6 program [167]. For PYTHIA 8, the A14 set of tuned parameters [168] is used for the simulation of UE with the NNPDF2.3LO PDF set [169]. Analogously, for HERWIG 7, the H7UE tune [161] is used with the MMHT2014LO PDF set [170].

Then, for all samples generated using MADGRAPH5_AMC@NLO for the ME, and in the *t*-channel single-top POWHEG samples, top quarks, *Z*- and *W*-bosons are decayed at LO using MADSPIN [172, 173] to preserve all spin correlations. In all samples with top quarks, the top-quark mass is set to $m_t = 172.5$ GeV. The mass of the *b*-quarks from the top-quark decays is set to $m_b = 4.95$ GeV (4.75 GeV) when the decay is modelled by POWHEGBOX or MADSPIN (SHERPA [162]).

Also, for all samples with a Higgs boson in the final state, its decay is done by the PS generator including all decay modes, using the recommended branching ratios [106, 174]. The Higgs-boson mass is set to $m_H = 125.0$ GeV, with a *b*-quark mass set at $m_b = 4.80$ GeV (4.50 GeV) for samples using PYTHIA 8 (HERWIG 7).

4.5.2 Signal model

In $t\bar{t}H$ signal events the production and decays (Fig. 2.15a) are modelled in the 5FS using the POWHEG BOX generator [151–153], which provides ME at NLO accuracy in the strong coupling constant α_s. The NNPDF3.0NLO [169] PDF set is used for the ME calculation, and the functional form of the renormalisation and factorisation scales are both set to $\mu_R = \mu_F = \sqrt[3]{m_T(t) \cdot m_T(\bar{t}) \cdot m_T(H)}$ [2]. The generated events are then interfaced to the PYTHIA 8 PS and hadronisation model. The h_{damp} parameter[3] is set to $\frac{3}{4}(m_t + m_{\bar{t}} + m_H)$=352.5 GeV. The $t\bar{t}H$ sample is normalised to the cross-section, $\sigma_{t\bar{t}H} = 507^{+35}_{-50}$ fb, determined at NLO accuracy in QCD including also NLO EW corrections [106] for a Higgs-boson mass of 125 GeV. In the $t\bar{t}H$ sample, all Higgs-boson decay modes (Sec. 2.4.2) are included in proportion to their branching ratios, hence containing only a very small number of $H \to \gamma\gamma$ and $H \to ZZ \to 4l$ events.

Moreover, the impact of the PS and hadronisation model is evaluated by comparing the nominal generator setup with a sample produced with the same ME generator and PDF settings

[2] The transverse mass of a generated particle is defined as $m_T = \sqrt{m^2 + p_T^2}$ where m represents its invariant mass, and p_T its transverse momentum.

[3] The h_{damp} parameter controls the p_T of the first additional gluon emission beyond the LO Feynman diagram in PS and therefore, regulates the high-p_T emission against which the $t\bar{t}H$ system recoils.

but the generated events are showered with HERWIG 7. Finally, in order to assess the uncertainty due to the choice of the NLO matching scheme (described in Sec. 4.2.3), the nominal sample is compared to an alternative sample, for which the calculation of the hard-scattering is done with the MADGRAPH5_AMC@NLO generator [156–158] with the NNPDF3.0NLO PDF set. The renormalisation and factorisation scale choice is the same as for the POWHEG setup. The generated events are showered with PYTHIA 8 as in the nominal sample.

4.5.3 $t\bar{t}$ + heavy flavour jets classification

Events arising from the $t\bar{t}$ pair production in association with additional jets ($t\bar{t}$ + jets) are distinguished into three non-overlapping categories according to the flavour of additional jets (not originating from the $t\bar{t}$ decay chain) in the event [95]. The classification is performed before reconstruction, using true- (generator-) level particle jets from the MC simulation. These are reconstructed from stable particles, as described in Sec. 5.2.1, using the anti-k_t algorithm with a radius parameter $R = 0.4$, and are required to have $p_T > 15$ GeV and $|\eta| < 2.5$. Then, hadrons are matched to particle jets, if they are within a distance of $\Delta R < 0.4$ from the jet axis, excluding jets produced by the top-quark or *W*-boson decays. Jets matched to at least one *b*- or *c*-hadron, the leading of which having $p_T > 5$ GeV, are labelled as *b*- or *c*-jets, respectively. Hence, $t\bar{t}$ events are classified, based on the flavour of the additional jets, as $t\bar{t}+\geq 1b$ if at least one *b*-jet is identified. In case no additional *b*-jet but at least one *c*-jet is identified, the event is labelled as $t\bar{t}+\geq 1c$. Both kind of events are collectively referred to as $t\bar{t}$ + heavy-flavour (HF) jets. Events not containing any HF jets, aside from those from top-quark or *W*-boson decays, are categorised as $t\bar{t}$ + light. With this classification, the $t\bar{t}+b\bar{b}$ (events with exactly two additional *b*-jets, each matched to at least one *b*-hadron) process falls under the $t\bar{t}+\geq 1b$ category. Finally, where necessary, the $t\bar{t}+\geq 1b$ events are further separated into $t\bar{t}+1b$ (exactly one jet is matched to at least one *b*-hadron) and $t\bar{t}+\geq 2b$ (all remaining events).

4.5.4 $t\bar{t}$ + jets background model

As already highlighted, the $t\bar{t}$ + jets production process is the dominant background to the $t\bar{t}H$ signal, with the $t\bar{t}$ pair decaying leptonically or semi-leptonically (Fig. 2.14a). The large phase space covered by this analysis requires a $t\bar{t}$ simulation that describes correctly the different topologies, especially the emission of additional jets and the heavy-flavour fraction. Not only the normalisation, but also the kinematics of the full final state have to be correctly modelled since several kinematic variables are used to build the final discriminants. Several MC samples, with different perturbative accuracy in the ME generator and with different choices for the PS and hadronisation processes, are used in this analysis to model the main $t\bar{t}$ + jets background. These samples, described below, are used for the nominal $t\bar{t}$ + jets model, for the modelling systematic uncertainties described in Sec. 7.1.2, or to study potential biases due to particular model components.

To accurately model the irreducible $t\bar{t}+\geq 1b$ background, a sample simulating the ME of the $t\bar{t}+b\bar{b}$ process (Fig. 2.15c) at NLO QCD accuracy in the 4FS is produced, using the POWHEG BOX RES generator [154] and OPENLOOPS [164] with the NNPDF3.0NLO nf4 PDF set. As introduced earlier, in the 4FS a $b\bar{b}$ pair in addition to the $t\bar{t}$ pair is generated only in the ME level, with the two additional *b*-quarks being massive. It is then interfaced to PYTHIA 8 for the PS and hadronisation. The scales are set to $\mu_F = \frac{1}{2}\sum_{i=t,\bar{t},b,\bar{b},j} m_T(i)$ (j stands for extra partons) and $\mu_R = \sqrt[4]{m_T(t)\cdot m_T(\bar{t})\cdot m_T(b)\cdot m_T(\bar{b})}$, while the parameter $h_{damp} =$

$\frac{1}{2}\sum_{i=t,\bar{t},b,\bar{b},j} m_T(i)$. The mass of the two *b*-quarks produced in the ME in association with the two top quarks is set to the same value as the mass of the *b*-quarks from the top-quark decays. The POWHEGBOXRES+PYTHIA8 $t\bar{t}b\bar{b}$ (4FS) sample is the nominal $t\bar{t} + \geq 1b$ prediction and its normalisation is given by the $t\bar{t} + b\bar{b}$ cross section. However, the normalisation of the $t\bar{t} + \geq 1b$ background is let free-floating in the fit process, as described in Sec. 8.2.

Inclusive $t\bar{t}$ + jets events are generated simulating the ME $t\bar{t}$ decay at NLO QCD accuracy in the 5FS, using the POWHEG BOX generator. As already introduced, according to the 5FS the additional *b*-quarks originate exclusively from the PS and are considered to be massless. The PS and hadronisation are modelled by PYTHIA 8, with the same settings as for the POWHEG-BOX+PYTHIA8 signal samples. Here, it holds $h_{damp} = \frac{3}{2}m_t$ [175], while the functional form of the scales is set to $\mu_R = \mu_F = m_T(t)$ [4]. The sample is normalised using the predicted $t\bar{t}$ cross section, $\sigma_{t\bar{t}} = 832^{+46}_{-51}$ pb, calculated at NNLO in perturbative QCD [262–265]. So, this POWHEGBOX+PYTHIA8 $t\bar{t}$ (5FS) sample is used as nominal only for the $t\bar{t}$ + $\geq 1c$ and $t\bar{t}$ + light predictions, since the 4FS sample provides more accurate prediction for the $t\bar{t} + b\bar{b}$ process.

However, for the determination of the modelling uncertainties no alternative 4FS MC samples with sufficient statistics are available. In the specific case of the $t\bar{t} + \geq 1b$ process, comparing the 5FS variations to the nominal $t\bar{t}b\bar{b}$ (4FS) sample would lead to a double counting of the difference between the $t\bar{t} + b\bar{b}$ coming from the ME and the PS. Therefore, the $t\bar{t}$ 5FS sample is used also for the $t\bar{t}$ + $\geq 1b$ prediction when assigning the uncertainties. Eventually, the impact of the PS and hadronisation model is evaluated by comparing the POWHEGBOX+PYTHIA8 (5FS) nominal to a POWHEGBOX+HERWIG7 (5FS) setup. Furthermore, the uncertainty due to the choice of the matching scheme (see Sec. 4.2.3) is assessed by comparing the POWHEG-BOX+PYTHIA8 (5FS) nominal to a MADGRAPH5_AMC@NLO+PYTHIA8 (5FS) setup. These additional samples are also normalised to the inclusive $t\bar{t}$ cross section. The aforementioned comparisons are used to evaluate the uncertainties for $t\bar{t}$ + $\geq 1c$ and $t\bar{t}$ + light, but also for $t\bar{t}$ + $\geq 1b$ background components, as described in Sec. 7.1.2. Using systematic variations derived with the 5FS $t\bar{t}$ sample, though, indicates that the two additional b-quarks are generated in the PS and their production is thus directly affected by these systematics. This implies that these uncertainties are overestimated compared to the case where the two b-quarks are directly part of the ME at NLO.

Finally, to enhance the statistics in the phase-space relevant for this analysis, for all the $t\bar{t}$ (5FS) MC samples described above, dedicated filtered samples are produced, for each of the three decay channels (dilepton, single-lepton, all-hadronic) and for each setup. They require *b*- or *c*-hadrons in addition to those arising from the decays of the top quarks. Hence, one sample is produced with at least two additional *b*-hadrons with $p_T > 15$ GeV. Another sample is produced with at least one additional *b*-hadron with $p_T > 5$ GeV, while failing the previous requirement. A last sample is produced with at least one additional *c*-hadron with $p_T > 15$ GeV, and failing the previous two requirements.

4.5.5 Single-top production background model

The single top-quark production is the second largest contribution to the background, after the $t\bar{t}$ + jets. All the three production processes, *t*-, *s*-, and *tW*-channel (Fig. 2.13), are considered, with the *t*-channel having the highest production cross section among them. The single-top

[4] This scale is calculated in the $t\bar{t}$ rest-frame, hence the p_T values of the top quark and top antiquark are equivalent.

production processes are modelled using the POWHEGBOX generator for the ME calculation at NLO accuracy in QCD. For *s*-channel and *tW* production, events are generated using the 5FS for the ME calculation with the NNPDF3.0NLO PDF set, while the scales are set to the top-quark mass. By contrast, for t-channel production, events are generated in the 4FS with the NNPDF3.0NLO nf4 PDF set, and the functional form of the renormalisation and factorisation scales is set to $m_T(b)$ [176]. In this case, it was found that the 4FS describes better the hardest b-quark not originating from the top-quark decay [176]. Generated events are then showered with PYTHIA 8. These samples are normalised using the theory prediction calculated at NLO in QCD [177, 266].

The impact of the PS and hadronisation model is evaluated by comparing the nominal POWHEGBOX+PYTHIA8 samples with alternative samples produced with the same ME generator setup, but events are showered with HERWIG 7. Moreover, to assess the uncertainty due to the choice of the matching scheme, the nominal samples are compared to samples generated with the MADGRAPH5_AMC@NLO generator, while the showering of the events is the same. The ME calculation is performed at NLO in QCD, in the 4FS (5FS) with the NNPDF3.0NLO nf4 (NNPDF3.0NLO, CT1.0NLO) PDF set, for *t*-channel (*s*-, *tW*-channel) production.

Some of the Feynman graphs that contribute to the *tW* channel can be interpreted as the $t\bar{t}$ pair production at LO, with subsequent decay of the *t* ($\bar{t}$) into a *bW* ($\bar{b}W$) pair. In order to handle the overlap between the $t\bar{t}$ and the *tW* final states, the diagram removal (DR) scheme [269] is employed. Additionally, an alternative sample is generated also for the *tW* production, which applies the diagram subtraction (DS) scheme [178]. Then, this is compared to the nominal POWHEGBOX+PYTHIA8 sample to assess an uncertainty in the modelling of this interference.

4.5.6 Rare top-quark processes background modelling

The production of a $t\bar{t}$ pair, decaying (semi-)leptonically, in association with a vector boson (i.e. $t\bar{t}W$ - Fig. 2.11e, or $t\bar{t}Z$ - Fig. 2.11f), collectively referred to as $t\bar{t}V$, is also considered a background. Such events can be misinterpreted as $t\bar{t}H$ events when the *W/Z* boson decays hadronically ($W^{\pm} \to qq'$, $Z \to q\bar{q}$) and the $t\bar{t}$ system decays semi-leptonically, or when the *W/Z* boson decays leptonically ($W^{\pm} \to l^{\pm}\nu_{l^{\pm}}$, $Z \to l^{+}l^{-}$) and the $t\bar{t}$ system decays hadronically. Both processes are modelled using the MADGRAPH5_AMC@NLO generator, which provides ME at NLO in QCD with the NNPDF3.0NLO PDF set. The scales are set to the default $\mu_R = \mu_F = \frac{1}{2}\sum_i m_T(i)$, where the sum runs over all the particles generated from the ME calculation. Generated events are interfaced with PYTHIA 8. Additional $t\bar{t}V$ samples are produced with the SHERPA 2.2.0 generator [162] at LO accuracy, along with the NNPDF3.0NNLO PDF set and the same scales as the former samples. Also, the MEPS@LO prescription [179, 180] is used, with up to one additional parton for the $t\bar{t}Z$ sample and two additional partons for the others. For the PS generation, the default SHERPA 2.2.0 PS is employed.

For events with four top quarks ($t\bar{t}t\bar{t}$, Fig. 2.13f) the production and decays are modelled using the MADGRAPH5_AMC@NLO generator at NLO in QCD with the NNPDF3.1NLO PDF set. The functional form of the scales is set to $\mu_R = \mu_F = \frac{1}{4}\sum_i m_T(i)$, where the sum runs over all particles generated from the ME calculation [181]. The events are showered with PYTHIA 8.

The SM *tZq* (Fig. 2.13d) events are generated using the MADGRAPH5_AMC@NLO generator in the 4FS at LO in QCD [182], with the CTEQ6L1 [183] PDF set. The renormalisation and factorisation scales were set to $4m_T(b)$ [176], with the *b*-quark coming from the gluon splitting. The generated events are showered with PYTHIA 8.

The *tWZ* (Fig. 2.13e) sample is produced using the MADGRAPH5_AMC@NLO generator in the 5FS at NLO in QCD with the NNPDF3.0NLO PDF set. The top quark decays inclusively, while the Z boson decays to a pair of leptons. The renormalisation and factorisation scales are set to the top-quark mass. The events are showered with PYTHIA 8. The DR scheme is employed to handle the interference between *tWZ* and $t\bar{t}Z$, and is applied to the *tWZ* sample.

4.5.7 Other backgrounds modelling

The associated production of a single top-quark and a Higgs boson, expressed by $tHjb$ and tHW subprocesses (Figs. 2.7d-2.7g), is rare in the SM, hence it has a negligible contribution to the $t\bar{t}H(H \to b\bar{b})$ analysis. However, both subprocesses are valuable to the top-quark Yukawa coupling measurement in the $t\bar{t}H$ combination, as their cross sections are asymmetric with respect to the sign of the top Yukawa coupling. Therefore, they are included as additional background in this analysis. MC samples for the two subprocesses are generated using the MADGRAPH5_AMC@NLO generator at NLO in QCD. The functional form of the scales is set to $\mu_R = \mu_F = \frac{1}{2}\sum_i m_T(i)$, where the sum runs over all the particles generated from the ME calculation. For $tHjb$ (tHW), events are generated in the 4FS (5FS) using the NNPDF3.0NLO nf4 (NNPDF3.0NLO) PDF set, and are showered with PYTHIA 8. Finally, the DR scheme is employed to handle the interference with $t\bar{t}H$ in the tHW sample [184,269]. The other Higgs-boson production modes are negligible in terms of this analysis, thus they are not considered.

The production of *W* or *Z* boson with additional jets (QCD *V*+jets processes, $V = W/Z$) also contributes to the background of this analysis, where each vector boson can subsequently decay either hadronically or leptonically. Both processes are simulated with the SHERPA 2.2.1 generator. The NLO-accurate MEs for up to two partons, and LO accurate MEs for up to four partons are calculated with the OPENLOOPS and Comix [185] libraries. They are matched with the default SHERPA PS by using the MEPS@NLO prescription with the set of tuned parameters developed by the SHERPA authors and based on the NNPDF3.0NNLO set of PDFs. These samples are normalised to the NNLO prediction [186].

Diboson events are simulated with the SHERPA 2.2.1 and 2.2.2 generators. In this set-up, the MEs are calculated up to one (*ZZ*) or zero (*WW*, *WZ*) additional partons at NLO, and up to three additional partons at LO. For semi-leptonically and fully-leptonically decaying diboson event samples, the virtual QCD correction for MEs at NLO accuracy is provided by the OPENLOOPS library. For EW *VVjj* production, the calculation is performed in the G_μ scheme, ensuring an optimal description of pure EW interactions at the EW scale [187–189]. Then, the multiple MEs are matched and merged with the default SHERPA PS using the MEPS@NLO prescription. All diboson samples are generated using the NNPDF3.0NNLO PDF set, along with the dedicated set of tuned PS parameters developed by the SHERPA authors.

Finally, the multi-jet background, consisting of events with several jets, has a negligible contribution in the single-lepton region. Multi-jet events arise from the misidentification of jets or photons as leptons (i.e. fake leptons), or the presence of well identified leptons not coming from the PV (i.e. non-prompt leptons). The latter may originate from semi-leptonic *b*- and *c*-hadron decays, or photon conversions (especially for electrons). However, with the isolation criteria applied at the trigger level (Sec. 6.3.1), as well as the purity-enhancing identification criteria for electrons and muons at event selection (Sec. 6.3.2), the majority of fake and non-prompt leptons in the single-lepton channel are removed. Thus the multi-jet background is not a determinable background source for this measurement.

Chapter 5

Physics Objects Definition and Reconstruction at Detector Level

The several components of the ATLAS detector provide useful information, which is used to reconstruct the paths and energies of leptons and parton showers in the detector. The identification (fig. 5.1) and reconstruction of these particles, produced in the ATLAS detector during the proton-proton interactions, is crucial for a proper reconstruction of the full event.

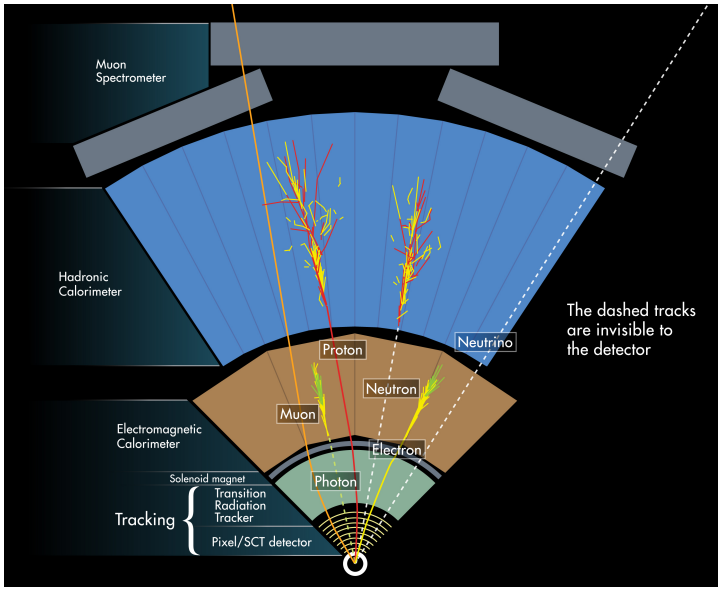


Fig. 5.1: Overview of particle identification in the ATLAS detector [190]. The solid and dashed curves show the tracks of charged and neutral particles, respectively. Arising from the interaction region, the muon passes through the whole detector being tracked by the ID and the MS. Electrons and photons are caught, mainly by the EMCal. Hadrons are detected mainly by the HCal. The charged particles leave a track also in the ID, while the neutral ones do not. Neutrinos escape from the entire detector without leaving any signature.

For this purpose, the first step is to build *low-level objects* representing individual particles. In the ID (see Sec. 3.2.2), tracks are constructed from space point hits, while calorimeter-cell clusters are formed in the calorimeter. Afterwards, various quality criteria are imposed for the tracks that are used in each analysis, in order to reject tracks that do not originate from the

particles produced in the primary collision. Then, *high-level objects*, such as electrons, muons, and jets, are constructed from the tracks and the calorimeter-cell clusters.

Undoubtedly, there is always the possibility that an object of one type is considered as of another type. In order to reject the mis-identified objects, particle identification schemes are constructed from the actual object properties. Furthermore, various corrections, based on shower properties, are applied to calibrate the energy or momentum of the reconstructed objects. All the aforementioned objects are finally used to construct the missing transverse momentum, which is a measure of the momentum carried away by particles that traverse the detector undetected, such as neutrinos.

5.1 Low level objects

5.1.1 Tracks

The charged particles deposit a small fraction of their total energy, referred to as *hit*, in the Pixel, SCT and TRT (inner) detector (outlined in Sec. 3.2.2) components while traversing the ATLAS detector. The locations traversed by a charged particle are represented by three-dimensional *space-points* in the SCT or in the Pixel detector. Then, the trajectories of the charged particles are reconstructed exploiting the information from the ID components and they form the *tracks*, using a sequence of algorithms [191, 192].

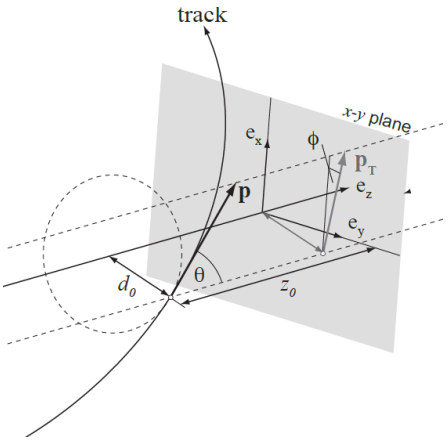


Fig. 5.2: Geometric illustration of the track helix parameters [193].

The magnetic fields from the ATLAS solenoid (Sec. 3.2.2) and toroid (Sec. 3.2.4) magnets bend the trajectories of these particles forming helices. In the ATLAS coordinate system, these helices are characterised by five parameters exploiting the full geometry and kinematics of the incoming particles

$$(d_0, z_0, \phi, \theta, q/|\vec{p}_T|), \tag{5.1}$$

where d_0 (*transverse impact parameter*) is the track's signed distance of closest approach to the z-axis (the sign gives the direction of rotation) and z_0 (*longitudinal impact parameter*) indicates the coordinate of the track along the z-axis at the point of closest approach. In addition, ϕ is the *azimuthal angle* of the track in the $x-y$ plane at the point of closest approach and θ expresses the *polar angle* of the track to the z-axis. Last but not least, $q/|\vec{p}_T|$ stands for the ratio between the charge of the particle and transverse momentum of the track. This parameter can

be determined by the radius of the curvature *R* in the magnetic field *B* by $q/|\vec{p}_T| = (RB)^{-1}$. The sign of the electric charge can be extracted from the direction of the curvature. Figure 5.2 illustrates a geometric definition of the track parameters. The detector has been designed to provide a p_T resolution for the tracking, in the plane perpendicular to the beam axis, of $\sigma_{p_T}/p_T = 0.05\% p_T$ GeV $\oplus\, 1\%$.

According to the main strategy of track reconstruction, different algorithms process the hit points of a track, in order to reconstruct the outgoing tracks starting from seeds[1] close to the interaction point (*inside-out*) and extrapolate them to the TRT. Nevertheless, some of these initial track seeds may not be found or do not even exist, i.e. tracks originating from secondary vertices of long-lived particles, photon conversions, and material interactions which can be found inside the ID. To take into account such topologies a consecutive algorithm is used, starting with seeds in the TRT and reconstructs the tracks backwards towards the inner parts of the detector (*outside-in*). In order to reduce the time required for the reconstruction and minimise double counting, the outside-in tracking procedure excludes all the TRT hits that have already been assigned to inside-out tracks. Both reconstructed tracks are combined to enhance the track reconstruction efficiency[2] [194].

5.1.2 Vertices

ID tracks are then deployed to construct *vertices*, points at which particles interact and produce divergent tracks. The *primary vertex* (PV) is defined as the point in space where pp interactions have occurred and is reconstructed from at least two associated tracks. The hardest pp interaction in a given event is referred to as the *hard scattering* and is associated with a reconstructed hard-scatter vertex, which is considered the hardest among all reconstructed PVs in the event. Thus, the *hard-scattering PV* is defined as the one with the largest sum of the squared transverse momentum ($\sum p_T^2$) of its associated tracks. However, multiple inelastic pp interactions may occur, reconstructed as a single physics event with many PVs. The PVs other than the hard-scatter one correspond to in-time pile-up interactions (see Sec. 3.1.1). All the other vertices, that are incompatible with the beam collision region, are considered as *secondary vertices* (SV) and may originate from pile-up interactions, multi-parton interactions, or decays of long lived particles such as b-hadrons.

The primary vertices are reconstructed from the combination of reconstructed tracks, using vertex finding and fitting algorithms [196], and are required to lie within the estimated position of the beam spot[3]. In the first step, the reconstructed tracks are associated to the vertex candidates and a seed position for the first vertex is selected (*vertex finding*). Afterwards, the optimal vertex position is obtained from an iterative fit with the seed and the reconstructed tracks as inputs (*vertex fitting*). During each iteration the tracks are weighted according to their compatibility with the corresponding vertex estimate. Once the vertex is found, the tracks that are incompatible with the vertex are removed from it and they are considered as inputs for a new vertex finding iteration. This whole procedure is repeated until no tracks are left in the event, or no vertex can be formed. The position resolution of the obtained vertices depends on the number of tracks used. In fact, the vertexing algorithms typically achieve a resolution

[1] A track *seed* is composed of space-points only from the Pixel detector or the SCT, or from combination of them.

[2] The efficiency is defined as the fraction of particles which are matched to reconstructed tracks passing the quality cuts.

[3] The beam spot is the spatial region around the interaction point, where the collisions take place within the detector.

of 30 μm in the (x, y)-plane for events with high multiplicity of reconstructed tracks, while it is below 20 μm in the z-axis for all track-multiplicities.

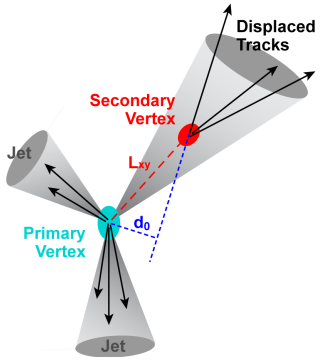


Fig. 5.3: Illustration of a collision where three jets, with their track content, emerge from the PV and one of them creates a SV. Its tracks are characterised by the transverse impact parameter d_0, while the decay lifetime length L_{xy} of the displaced tracks allows to resolve the SV [195].

5.1.3 Clusters

When high energy photons and electrons interact with the detector medium, several secondary particles are produced, which constitute the electromagnetic showers and deposit their energy in the EMCal (Sec 3.2.3) cells. Similarly, high energy hadrons interact with the detector medium developing hadronic showers. They include electromagnetic sub-cascades, though, initiated by neutral pions, which are within the particles of the hadronic showers. Hence, the particles from the hadronic showers deposit their energy in the HCal (Sec 3.2.3) cells, accompanied by hits in the EMCal cells. Different clustering algorithms group the calorimeter cells into clusters and sum the energy deposits within their constituents cells. These energies are then calibrated to account for the energy deposited outside the cluster and in dead material.

The calorimeter clusters can be reconstructed using the *sliding-window* algorithm, which is efficient for electromagnetic shower reconstruction, and are called *towers* [197]. Thus, the calorimeter towers are used for electron and photon identification. In order to build a tower, the $\eta - \phi$ space of EMCal is divided into a grid of $N_\eta \times N_\phi = 250 \times 256$ elements of size $\Delta\eta \times \Delta\phi = 0.025 \times 0.025$ within the $|\eta| < 2.5$ region. Inside each element, the energy of all cells in all layers of the calorimeter is summed into the tower energy. Afterwards, a window of fixed size ($N_\eta \times N_\phi = 5 \times 5$) is moved across each element of the tower grid (in $\Delta\eta$-$\Delta\phi$ steps). The position of the window is adjusted so that its contained energy is a local maximum. Finally, all cells within a rectangle of a certain size, centered around the position of the window, are assigned to EM clusters. Clusters of different sizes are built depending on the hypothesised particle and the location of the cluster in the calorimeter. An optimal cluster should contain most of the energy deposited by the particle, so as to limit the lateral shower leakage contribution to the energy resolution. At the same time, it should include as little noise as possible, by not including cells without a physical signal. A calorimeter tower has an energy equal to the energy sum of all included calorimeter cells. The formed Lorentz four-momentum has zero mass.

Besides the sliding-window algorithm, there is the topological clustering algorithm which groups noise-suppressed clusters of topologically connected calorimeter cells, the so-called

topoclusters [198]. This algorithm is more efficient for the jet (Sec. 5.2) and missing transverse energy (Sec. 5.5) reconstruction. The topological clustering starts from calorimeter cells with energy exceeding four times the expected noise, called *seeds*. The expected noise is defined by the standard deviations of the electronic and pile-up noise[4], summed in quadrature. Cells adjacent to the seed (in three dimensions) and with energy exceeding two times the noise are iteratively added to the cluster. Afterwards, all direct neighbouring cells are added to the perimeter of each cluster, irrespective of their energy. The lack of energy threshold at the perimeter ensures that tails of hadronic showers are not discarded, while the high thresholds for seeds and neighbours effectively suppress noise. Finally, a splitting step is included in the algorithm in order to optimise the separation of showers from distinct close-by particles, which form local energy maxima[5]. All cells in a topocluster are searched for local energy maxima, which are then used as seeds for a new iteration of topological clustering, splitting the original cluster into more topoclusters, each with a variable number of associated cells. A topocluster is interpreted as a massless pseudo-particle with energy equal to the sum of its constituent cell energies.

The energy of a cluster is calibrated at the electromagnetic (EM) energy scale, which correctly reconstructs the energy deposited in the calorimeter by particles produced in the EM showers. An additional calibration using the local cluster weighting (LCW) scheme [198] can be applied, especially to topoclusters, to compensate for the lower calorimeter response to the hadronic components of the shower. The LCW method classifies the topoclusters at the correct particle-level energy scale, based on the energy density and the longitudinal shower depth. Energy corrections are applied to each cluster to account for energy losses due to noise threshold effects and in the inactive material (non-instrumented regions) within the detector, the out-of-cluster energy depositions, as well as for the non-compensating response of the calorimeters.

5.2 Jets

As already established in the context of QCD (Sec. 2.2.4), partons emerging from high-energy particle collisions, being colour-charged particles, cannot be directly observed in the detector. Instead, after the hard scattering process, additional particles are produced through fragmentation and hadronisation until a stable colourless final state is achieved. The experimental signature of the resulting showers of collimated hadrons is the jets, whose reconstruction is of particular importance for the analysis. The jet reconstruction aims at producing objects which preserve the original kinematic characteristics (energy and momentum) of the parent partons, in order to infer their properties from the corresponding jets. Then, the jets have to be calibrated and can be further analysed to discover their substructure, as well as, identify the flavour of the initial partons.

[4]The pile-up noise comes from extra interactions that can either be overlaid in the same beam crossing or occur during crossings close-in-time with that of the primary interaction.

[5]A local energy maximum is defined as a clustered cell with energy > 500 MeV. Also, this cell is required to have at least four neighbouring cells and none of them having energy larger than that.

5.2.1 Jet reconstruction

Jets are reconstructed using clustering algorithms, which attempt to reduce the complexity of the multi-hadron final states using simpler four-vector objects, which represent the energy and direction of the initial hard-scattering partons. The total jet four-momentum is therefore defined as the sum of the four-momenta sum of all its constituents.

Jets may be defined in various ways depending on the type of objects and algorithms used to construct them. Experimentally, the final-state particles are observed as tracks in the tracker systems or as clusters of energy deposits in the finely segmented calorimeter cells. Therefore, the input objects to the jet algorithm can be reconstructed charged-particle tracks in the ID, originating from the primary hard scattering vertex, (*track jets*), or three-dimensional topoclusters with positive energy (*calorimeter jets*). Except for these types of jets reconstructed at detector level, jets can be also defined at particle level. Particle-level[6] jets are formed from stable (see Sec. 4.2.4) interacting[7] final-state particles emerging from fragmentation processes in MC simulations (*true jets*), excluding particles from pile-up interactions.

A jet finding algorithm [208] should be theoretically well defined at all orders of perturbation theory, as well as safe against infrared and collinear radiation (defined in Sec. 4.2.3) (*IRC safe*) so that the number and properties of reconstructed jets remain unbiased. In particular, collinear safety ensures that the jet formation is independent of the number of particles within the hadronic shower; the jet remains unchanged if a particle is replaced by two collinear particles with a total energy equal to the original. Infrared safety requires the jet clustering to be driven by the hardest energy deposits, namely a hard jet is unaffected by the addition of a soft particle from the initial parton.

In addition, a well-defined jet algorithm should be invariant to boosts along the beam direction and insensitive to non-perturbative effects like hadronisation, or underlying events due to additional hadron-hadron collisions per bunch crossing. Lastly, the boundaries of a jet should be well-defined even in the case of overlapping jets. The algorithm will assign the shared objects to one of the overlapping jets, depending on the energy and the distance between the object and the four-vectors of the jets.

The jet finding algorithm combines objects (particles, tracks, or topoclusters), which are then aggregated into individual jets, mapping the momenta of the final state particles into the momenta of the resulting jets. It also contains a resolution parameter, R, which controls the extension of the jet, denoting the characteristic conical shape of parton showers. Objects that are likely to have resulted from the same initial parton are grouped together starting from the highest p_T of a four-vector object *i*, which sets the initial direction. Then, all objects *j* within a radius R around *i* satisfying specific requirements are grouped together, summing up their four-momenta. The distance between the two four-vectors in the $y - \phi$ space is defined as

$$\Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2 < R^2, \tag{5.2}$$

where $y_{i(j)}$ is the rapidity and $\phi_{i(j)}$ is the azimuthal angle of particle $i(j)$. A new four-vector is assigned to this group of objects and nearby objects, within a radius R, are recalculated at its center in the same way. This iterative procedure continues until no further objects match the conditions and the grouped objects eventually form a four-vector jet.

[6]Particle- (or generator-) level refers to a state where all final state particles of an event after the PS and the hadronisation are defined, but before the decays and propagation through the detector.

[7]A true particle is considered to be interacting if it deposits most of its energy in the calorimeters; thus, muons and neutrinos are considered to be non-interacting particles.

In this analysis, a *sequential recombination algorithm* [208] is exploited for the jet reconstruction, which relies on the pair-wise combination of objects into final jets, attempting to "undo" the showering of partons. A distance between pairs of particles is defined, according to which successive recombinations of pairs of closest particles are performed, until all resulting objects are too far apart. Moreover, this algorithm takes advantage of the minimal energy-weighted geometrical distance between the particles that are combined. This ensures that the distance between two soft, back-to-back particles is larger than that between a soft and a hard particle that is nearby in angle. The distance metric d_{ij} between two jet-candidates *i* and *j*, as well as the distance d_{iB} between each four-vector and the LHC beam are defined (using the transverse energy/momentum $E_{T,i(j)} = |\vec{p}_{T,i(j)}|$ of an object *i*(*j*)) as

$$d_{ij} = min\left(p_{T,i}^{2m}, p_{T,j}^{2m}\right)\frac{\Delta R_{ij}^2}{R}, \tag{5.3}$$

$$d_{iB} = p_{T,i}^{2m}. \tag{5.4}$$

The variable ΔR_{ij}^2 denotes the radial distance between two jet-candidates, as defined in eq. (5.2), while *R* is the radius of a jet determining its size. The jet clustering begins by combining the two four-vectors with the smallest distance d_{ij}. If $d_{ij} > d_{iB}$ for all four-vector combinations, then the jet-candidate *i* is classified as a final jet and it is not taken into account in further calculations. Otherwise, the two jet-candidates are merged into a new four-vector. Then, the distances d_{ij} are recalculated and the sequential recombination procedure continues until all inputs have been clustered into jets. Finally, the parameter m is an integer that determines how the ordering of the cluster sequence behaves and designates the different algorithm types.

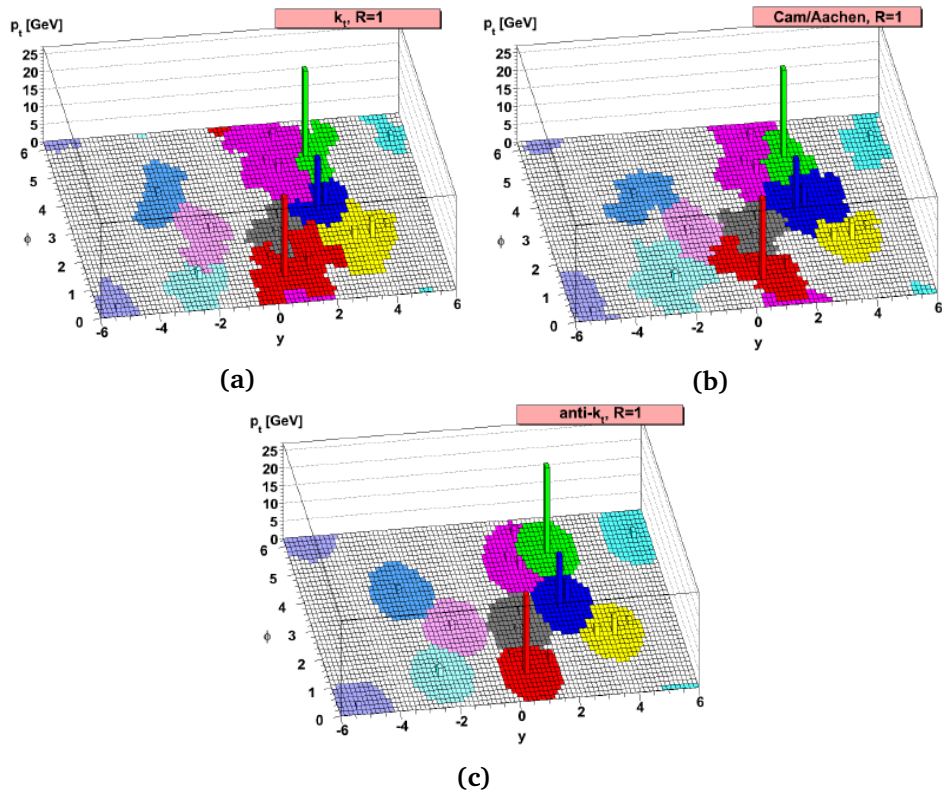


Fig. 5.4: Illustration of topoclusters, in a parton-level event with soft particles, grouped into jets with the a) k_t, b) Cambridge/Aachen and c) anti-k_t algorithms [212].

A positive m value clusters the constituents from lowest (softest) to highest in p_T (hardest), and the case $m = 1$, commonly known as the k_t algorithm [209], allows for a reverse parton shower. This leads to non-conical shapes and the highest p_T clusters are not always the center of the reconstructed jet, as depicted in Fig. 5.4a). When $m = 0$, the clustering procedure is independent of the energy of the inputs, referred to as the Cambridge/Aachen (C/A) algorithm [210]. This results in combining particles based only on their relative distance and the jets grow around the hardest topoclusters ending up also with somewhat irregular shape, as illustrated in Fig. 5.4b). On the other hand, negative m values refer to a clustering process combining the hardest jet-candidates first, and $m = -1$ corresponds to the anti-k_t algorithm [212]. This is the most ubiquitous algorithm, since it creates conical shaped hard jets, centered around their highest-p_T constituents, while only the softer jets have more complex shapes (e.g. the pair of jets near $\phi = 5$ and $y = 2$ in Fig. 5.4c). Also, the energy entries of overlapping jets are properly assigned to the hardest close-by jet.

In the study presented in this thesis, the calorimeter jets are used for the definition of the analysis, which are reconstructed with the anti-k_t jet algorithm with a radius parameter of $R = 0.4$, and referred to as *small-R jets*. Afterwards, the calorimeter jets need to be calibrated, in order to restore the jet energy scale to that of jets reconstructed from stable simulated particles. Thus, true jets are employed as a reference for jet calibration purposes using MC simulation. Reconstructed calorimeter jets are geometrically matched to true jets using the angular distance requirement $\Delta R < 0.3$, for the various generator-level studies, included in Sec. 6.4.1. Furthermore, also track jets are used in the jet calibration procedure. Given that only tracks originating from the hardest PV in the collision are used in the jet finding, track jets are insensitive to the pile-up activity providing a rather stable kinematic reference for matching with calorimeter jets. Both true and track jets are reconstructed with the same configurations as calorimeter jets.

5.2.2 Jet calibration

The reconstructed calorimeter jets are calibrated to correct their energy scale (JES), in order to correspond to that of the true jets reconstructed at the particle level. The calorimeter jets are reconstructed from topoclusters calibrated either at the EM scale (*EMTopo jets*) or at the LCW scale (*LCTopo jets*). Concentrating on EMTopo jets for the purpose of this analysis, the jet calibration procedure consists of several consecutive stages derived from a combination of MC-based methods and in situ techniques, based on 13 TeV data, as illustrated in Fig. 5.5. Each calibration stage corrects the reconstructed jet four-momentum unless otherwise stated, scaling the jet p_T, energy, and mass. The MC-based calibrations account for features of the detector, the jet fragmentation and reconstruction algorithm, as well as the busy data-taking environment resulting from pile-up interactions. The in situ techniques are used to measure the difference in the jet response between data and MC simulation, with residual corrections applied to jets in data only.

A jet produced in the hard-scattering interaction is expected to originate from the PV. However, the calorimeter jets are reconstructed using the geometrical centre of the ATLAS detector as a reference to calculate their direction and constituents. Therefore, a correction to account for the position of the PV in each event, called the *origin correction* [225], is applied to every topocluster. The jet four-momentum is recalculated so that the direction of each topocluster points to the PV rather than the detector centre, while the jet energy is unaffected. This correction improves the angular resolution of jets, as measured from the difference between

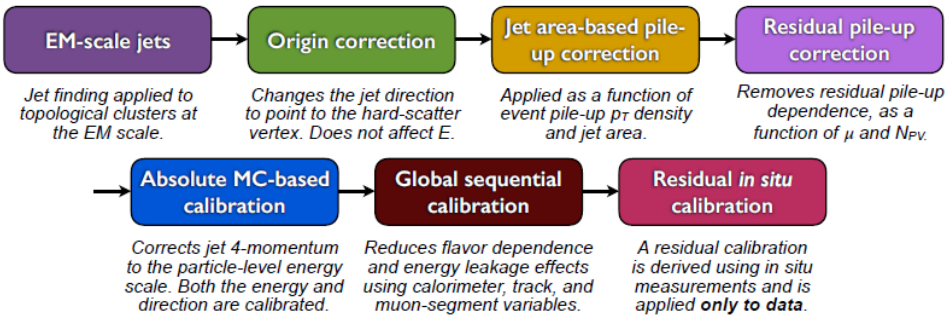


Fig. 5.5: Jet energy scale calibration stages, each one applied to the four-momentum of the jet, except for the origin correction [225].

reconstructed and true jets in MC simulation.

Afterwards, the *pile-up correction* [225], [211] removes the excess energy due to additional pp interactions within the same (in-time) or nearby (out-of-time pile-up) bunch crossings. It consists of two components: a *correction based on the jet area and* p_T *density* and a *residual correction*. The area-based p_T density method subtracts the per-event pile-up contribution from the p_T of each jet according to its area. The pile-up contribution is estimated from the median p_T density $\rho = \left\langle \frac{p_T}{A} \right\rangle$ of jets in the $\eta - \phi$ plane. The jet area A is a measure of the susceptibility of an individual jet to pile-up and it is determined by the relative number of ghost particles[8] associated with a jet after clustering. However, after this correction some dependence of the jet p_T on pile-up remains, therefore an additional correction is applied. The residual correction derived from MC simulation is defined as the difference between the reconstructed and true jet p_T. It is applied as a function of the expected number of additional pp collisions per bunch crossing (μ) and the number of reconstructed primary vertices in the event (N_{PV}), sensitive to in-time and out-of-time pileup (defined in Sec. 3.1.1), respectively. The jet p_T, after both pile-up corrections, is given by

$$p_T^{corr} = p_T^{reco} - \rho \cdot A - \alpha \cdot (N_{PV} - 1) - \beta \cdot \mu, \tag{5.5}$$

where p_T^{reco} refers to the EM-scale p_T of the reconstructed jet before any pileup correction is applied. The coefficients α and β correspond to the residual p_T dependence on N_{PV} and μ respectively and are derived from linear fits in bins of true jet p_T and $|\eta|$.

Moreover, the *absolute jet energy scale* and η *calibrations* [225] correct the reconstructed jet four-momentum to the particle-level energy scale, accounting for non-compensating[9] calorimeter response, energy losses in inactive material, out-of-cone effects and biases in the jet η reconstruction. The latter are mainly caused by the transition between different calorimeter technologies and sudden changes in calorimeter granularity. These calibrations are derived from dijet MC events and are applied after the *origin* and *pile-up corrections*. The *JES* calibration corrects the reconstructed jet energy to the true jet energy, using isolated[10] reconstructed jets that are geometrically matched to true jets within $\Delta R = 0.3$, and it is defined as the inverse

[8] Ghost particles [213] are simulated particles of infinitesimal momentum that are added uniformly in solid angle to the event before jet reconstruction.

[9] Non-compensation is an intrinsic property of the calorimeters, according to which the calorimeter response to the electromagnetic components of a hadronic shower is different (in fact smaller for the ATLAS HCal) from the response to the non-electromagnetic (purely hadronic) ones.

[10] An isolated jet has no other (only one) calorimeter (true) jet of $p_T > 7$ GeV within $\Delta R = 0.6$ (1.0).

of the average energy response[11]. Also, gaps and transitions between calorimeter subdetectors result in a lower energy response due to absorbed or undetected particles. An additional correction is applied to compensate for the bias in the reconstructed jet η direction, defined as the difference between the reconstructed and true η_{det}. The η calibration alters only the jet p_T and η and not the full four-momentum. The absolute JES and η calibrations are also derived for fast simulation samples using the same methods as in those with the full detector simulation.

Following the previous jet calibrations, residual dependencies of the JES on longitudinal and transverse features of the jet are observed. The calorimeter response and the jet reconstruction are sensitive to fluctuations in the flavour and energy distribution of the constituent particles. In addition, the jet average particle composition and shower shape vary between jets initiated mainly by quarks (including hadrons with higher jet p_T fraction, penetrating further into the calorimeter) and gluons (containing more particles of softer p_T, leading to a lower calorimeter response). The *global sequential calibration* (*GSC*) [225] is a series of independent multiplicative corrections applied to the jet four-momentum, as a function of $|\eta_{det}|$ and p_T^{true} (with a stronger dependence) by inverting the reconstructed jet response, using reconstructed jets geometrically matched to true jets. The GSC exploits the topology of energy deposits in the calorimeter, tracking information, as well as information related to the activity in the muon chamber behind jets. The purpose of each correction is to reduce the effects from the aforementioned fluctuations, while improving the jet resolution[12] without changing the average jet energy. Also, in order to reduce the non-Gaussian tails in the jet response distribution, caused by high-p_T jets not fully contained in the calorimeter (associated with muon track segments), a correction as a function of the jet energy is applied, referred to as punch-through correction.

The MC-based calibrations, mentioned so far, correct the EM-scale jets for MC mis-modelling. However, there are also differences in the jet response between data and MC simulation. They arise from the imperfect description of the detector material and simulation of the physics processes involved (hard scatter and underlying event), the pileup and the jet formation, as well as the EM and hadronic interactions with the detector. Therefore, the *residual in situ calibration* [225] contains corrections derived from real data, that are applied to the jet four-momentum of data only, accounting for these remaining differences. The jet response is defined as the average ratio of the jet p_T to a well-measured reference object p_T, in bins of the reference object p_T. It is equivalent to the calorimeter response to jets, while it is sensitive to the presence of additional radiative jets or the loss of energy outside the jet cone. In order to mitigate these secondary effects, the ratio of the jet responses, measured separately in data and MC simulation, is applied as the final correction, being a reliable measure of the JES difference between data and MC. More precisely, the η-intercalibration corrects the energy scale of forward ($0.8 < |\eta_{det}| < 4.5$) jets to match those of central ($|\eta_{det}| < 0.8$) jets in dijet events. The Z/γ+jet balance uses a well-calibrated photon or Z boson to correct the p_T response of the hadronic jet recoil in the central region. The multijet (topologies with three or more jets) balance calibrates central ($|\eta_{det}| < 1.2$), high-p_T jets recoiling against a collection of well-calibrated lower-p_T jets. The Z/γ+jet and multijet calibrations are statistically combined into a single calibration covering the full kinematic range.

[11] The average energy response (i.e. average calorimeter response to jets) is defined as the mean of a Gaussian fit to the core of the E^{reco}/E^{true} distribution for jets, binned in E^{true} and η_{det} (pointing from the geometric center of the detector).

[12] The jet resolution is given by the standard deviation of a Gaussian fit to the jet p_T response distribution, where the latter is defined as p_T^{reco}/p_T^{true}.

5.2.3 Jet Energy Scale and Resolution uncertainties

A large number of systematic uncertainties arise from the JES calibrations [214], derived mainly from the in situ measurements, pile-up effects, and flavour dependence, providing an accurate understanding of the overall JES uncertainty. In addition, correlations between the JES uncertainties for two jets at different η and p_T exist. However, many physics analyses would be hampered by the implementation and evaluation of them all having no benefit from the rigorous conservation of all correlations. Instead, a single JES uncertainty, by adding in quadrature all the independent components, would lead to an unrealistic assumption of full correlation between the JES uncertainties for any values of η and p_T. As a result, a reduced set of systematic uncertainties is preferred preserving as much as possible the correlations across jet p_T and η.

The majority of the JES uncertainty components stem from the Z/γ+jet (through the decay channels $Z \rightarrow e^+e^-$ and $\mu^+\mu^-$) and multijet in-situ calibrations. They account for assumptions in the event topology, MC simulation, sample statistics, and propagated uncertainties of the electron, muon, and photon energy scales. Since they are functions only of p_T, their behaviour can be easily represented by a smaller number of orthogonal terms. The category reduction scheme [225] combines these p_T-dependent uncertainty components in separate categories based on their source (*detector* description, *statistics* and method, physics *modelling*, or *mixed* detector and modelling), resulting in fifteen new effective nuisance parameters. In addition, a high-p_T "single-particle" uncertainty term is derived from studies of the response to individual hadrons and it is applied to jets with $p_T > 2$ TeV, where the multijet balance analysis has no statistical power. Finally, five uncertainties are associated with the η-intercalibration technique, accounting for potential physics mismodelling, statistical uncertainty, and the non-closure[13] of the method.

The remaining systematic uncertainties are related to the MC-based calibrations. Four pileup uncertainty terms are included to account for potential MC mismodelling of N_{PV}, μ, and ρ topology, as well as the residual p_T dependence of N_{PV} and μ. These uncertainties are derived from either data or MC simulation studies, or their difference. Furthermore, when calibrating AFII (defined in Sec. 4.3) MC samples, an uncertainty to account for the small non-closure in the absolute JES and η calibrations is introduced only for AFII samples. It accounts for the difference in the jet response between the fast and full detector simulated samples, due to the approximate treatment of the hadronic showers in the forward calorimeters. In addition, a punch-through uncertainty is considered for the mis-modelling of the GSC correction to jets which pass through the calorimeter and into the muon system. It is derived from the difference in the jet response between data and MC simulation, as a function of the muon detector activity. Finally, two jet-flavour composition uncertainties are derived from simulation to reflect the differences in the calorimeter response to quark- and gluon-initiated jets. An additional uncertainty is applied only to b-initiated jets, in order to cover the difference in response between jets from $light$- and heavy-flavour quarks.

Even if JES is perfectly calibrated, the precise energy of a jet can not still be measured due to noise, stochastic jet-by-jet of fluctuations in the calorimeter response, and detector calibration effects. The calorimeter response to jets is expected to be distributed approximately like a Gaussian distribution with a width σ, referred to as the jet energy resolution (JER) [214].

[13] After the jets in the nominal jet MC simulation sample are calibrated, the jet energy and p_T response still show slight deviations from unity at low p_T. This so-called *non-closure* refers to a failed consistency test when the calibration is applied to the same sample from which it is derived.

The analyses employed to measure the JER are essentially the same as for the jet calibration, but the observable of interest is not the mean of the jet energy response but is its standard deviation. For the central rapidity region, the JER is measured with good precision using Z/γ+jet simulated events, as well as in situ techniques. In the forward η region and for high p_T, dijet events provide the most precise determination of the JER, while for very low p_T jets there is a significant contribution from pile-up particles and electronic noise. The dependence of the relative JER on the jet p_T may be parametrised using a functional form, with three independent contributions

$$\frac{\sigma(p_T)}{p_T} = \frac{N}{p_T} \oplus \frac{S}{\sqrt{p_T}} \oplus C, \tag{5.6}$$

where the noise (N) term stands for the effect of electronic noise to the signal measured by the detector electronics, as well as due to pileup. Statistical fluctuations in the amount of energy deposited are captured by the stochastic (S) term. The constant (C) term corresponds to fluctuations that are a constant fraction of the jet p_T, such as energy depositions in passive material, the starting point of the hadron showers, and non-uniformities of response across the calorimeter.

In order to measure the JER, jet momentum must be measured precisely. This implies that the jets must either recoil against a reference object whose momentum can be measured precisely, or be balanced against one another in a well-defined dijet or Z/γ+jet system. The JER measurements based on the latter approach are statistically combined using a chi-squared minimisation of the function in eq. 5.6. The uncertainties in each term are evaluated in the same way they were in the JES determination. A set of JER systematic uncertainties arise from this fit parametrisation and stability. What is more, the energy resolution for jets in the MC simulation is very close to the resolution observed in data. The uncertainty on the JER measurement in data is propagated as an uncertainty in the response in MC simulation. The observed difference in response between the varied and the nominal results is defined as the systematic uncertainty due to JER. The non-closure of the method is largely due to the differences in topocluster formation sensitivity to pile-up and electronic noise in the presence versus absence of hard-scatter particles.

5.2.4 Jet Vertex Tagger

Several pileup jets remain above the p_T threshold used in this analysis, even after the pileup subtraction during the jet calibration procedure, mainly due to localised fluctuations in pile-up activity which are not fully corrected by the ρ term in eq. 5.5. Information from the tracks matched to each jet may be used to further reject jets not originating from the hard-scattering interaction. For the identification of pile-up jets a track-based tagging discriminant, called the *Jet Vertex Tagger* (*JVT*) [211], has been developed. The JVT combines the information from two variables, the *corrected jet vertex fraction* (*corrJVF*) and R_{p_T}, into a multivariate analysis.

The variable corrJVF identifies the primary vertex from which the jet originated, accounting also for a pile-up (N_{PV}) dependence. It is defined as

$$corrJVF = \frac{\sum_m p_{T,m}^{track}(PV_0)}{\sum_l p_{T,l}^{track}(PV_0) + \frac{\sum_{n\geq 1}\sum_l p_{T,l}^{track}(PV_n)}{k \cdot n_{track}^{PU}}}, \tag{5.7}$$

where $\sum_m p_{T,m}^{track}(PV_0)$, $\sum_{n\geq 1}\sum_l p_{T,l}^{track}(PV_n) = p_T^{PU}$ denote the scalar sum of the p_T of the tracks associated with the jet, originating from the hard-scattering PV (PV_0) or from any of the

pile-up interactions, respectively. The term $k \cdot n_{track}^{PU}$ corrects for the linear increase of $< p_T^{PU} >$ with the total number of pile-up tracks per event (n_{track}^{PU}), and k is approximately their slope. The resulting discrimination between hard-scatter and pile-up jets is insensitive to the choice of k. The corrJVF is expected to be close to 1 for hard-scattering jets and close to 0 for pile-up jets, since they are not originating from the PV.

The R_{p_T} is the ratio of the scalar sum of the p_T of the associated tracks originating from the hard-scattering vertex, to the fully calibrated jet p_T after pile-up subtraction

$$R_{p_T} = \frac{\sum_i p_{T,i}^{track}(PV_0)}{p_T^{jet}}. \tag{5.8}$$

It peaks at 0 and steeply falls for pile-up jets, since tracks from the hard-scattering vertex rarely contribute. However, for hard-scattering jets, the R_{p_T} has the meaning of a charged p_T fraction and it is a broad distribution with a mean value and a spread larger than those for pile-up jets.

A cut on JVT can suppress spurious calorimeter jets resulting from local fluctuations in pile-up activity, as well as real QCD jets originating from single pile-up interactions, resulting in improved stability of the reconstructed jet multiplicity against pile-up. However, the differences in fragmentation and showering between gluon- and light-quark-initiated jets affect the corrJVF and R_{p_T} distributions and thus the performance of the JVT pile-up jet suppression. Jets initiated by light quarks have on average a lower number of associated hard-scattering tracks, but a slightly higher jet energy response. Both effects lead to an increased number of jets with no associated tracks from the hard-scattering PV with respect to the gluon-initiated jets. The hard-scattering jet efficiency and the corresponding scale factors, for such a cut on JVT, are derived from $Z(\rightarrow \mu^+\mu^-)$+jets events in data and simulation with tag-and-probe techniques [211]. A good agreement is observed between data and simulation, while the slight difference is within the statistical uncertainty. The systematic uncertainty associated with the JVT requirement is determined by accounting for the differences in efficiency observed between different MC generators.

5.2.5 Reclustered (large-*R*) jets

Especially for the boosted topology (explained in Sec. 6.2), another type of jets is also exploited, the so-called *reclustered* (RC) *jets* [215]. The small-R jets, described in Sec. 5.2.1, are used as input constituents to the anti-k_T algorithm, which reclusters these with a radius parameter of $R = 1.0$, resulting in a collection of fully calibrated large radius jets. This allows for direct propagation of the systematic uncertainties associated with the input small-*R* jets, which are already calibrated, thus no further calibration or uncertainties are needed. Also, the *b*-tagging (explained in the following) associated with reclustered large-*R* jets can be done directly on the constituent subjets, so no ΔR matching is performed at this point. What is more, the combination with resolved category is straightforward, since the same objects are used. Finally, in order to remove the impact of pile-up on the jets, *reclustered jet trimming* [216] is applied. According to this, a threshold f_{cut} is applied removing any small-*R* jet with p_T^{sub} being a subjet of a large-*R* jet with p_T^{large}, if $\frac{p_T^{sub}}{p_T^{large}} < f_{cut} = 0.1$. Nevertheless, this has minimal effect, since the constituent small-*R* jets have already had pile-up suppression techniques applied through the JVT requirement.

5.3 *b*-tagging

The identification of jets containing hadrons from the fragmentation of *b*-quarks (*b-jets*), is referred to as *b-tagging*. It is of major importance for this analysis, since the final state contains many *b*-quarks (see Fig. 2.15). The aim is to identify *b*-jets against the large jet-background originating from *c*-hadrons but no *b*-hadron (*c-jets*), or originating from *light*-flavour partons (*u*, *d*, *s*-quarks, or gluons) but no *b-*/*c*-hadron (*light-jets*). For this purpose, the long lifetime, high mass and high decay multiplicity of *b*-hadrons as well as the properties of the *b*-quark fragmentation are exploited.

The *b*-hadrons have sufficient lifetime, $\tau \sim 1.5$ ps ($c\tau \sim 450\mu$m), hence they travel some distance in the detector before decaying, resulting in a significant decay length ($l = \beta\gamma c\tau$) of several millimeters. This leads to topologies with at least one secondary vertex (SV), corresponding to the decay vertex of a *b*-hadron displaced from the hard-scattering collision point (PV). The distance of the secondary from the primary vertex corresponds to the decay length of the initiating particle. Also, the PV comprises the reference point with respect to which the impact parameters (Sec. 5.1.1) and vertex displacements are expressed. The transverse impact parameter, d_0, is the signed distance of closest approach of the track to the PV point in the *x*-*y* plane. The longitudinal impact parameter, z_0, is the difference between the *z* coordinates of the PV position and of the track at the point of closest approach. A positive sign is assigned to the impact parameters if the track intersects the jet axis in the transverse plane in front of the PV, and a negative one otherwise. Displaced charged-particle tracks, originating from *b*-hadron decays, tend to have large impact parameters which can be distinguished from tracks stemming from the PV. Figure 5.3 illustrates a SV with tracks displaced from the PV, and the large impact parameters.

Another significant property of *b*-hadrons that can be exploited is their high mass ($\sim$ 5 GeV), which is at least two times higher than the mass of *c*- and *light*-hadrons and than anything they decay into. Thus, their decay products tend to have higher transverse momentum. Additionally, thanks to their large mass, the *b*-hadrons produce a large number of charged particles, resulting in wider *b*-jets with higher track multiplicities than *c*- and *light*-jets. Finally, *b*- and *c*-hadrons can decay to electrons and muons. This signature can be employed to identify *b*- and *c*-jets.

Eventually for the *b*-jet identification at detector level, ID tracks are associated to reconstructed jets based on the angular separation ΔR between the track and the jet axis directions. Given that the decay products from higher-p_T *b*-hadrons are more collimated, the ΔR requirement varies as a function of jet p_T, being wider for low-p_T jets and narrower for high-p_T jets. If more than one jet fulfills the matching criteria, the closest jet is preferred.

Analogously, jet flavour labels are also attributed to the jets in the simulation. In this case, true jets are labelled according to their flavour by spatially matching the jet with generator-level hadrons. In particular, jets are labelled as *b-jets*, if they are matched to at least one weakly decaying *b*-hadron having $p_T > 5$ GeV within a cone of size $\Delta R = 0.3$ around the jet axis. If no *b*-hadrons are found, then *c*-hadrons are searched for, and the jets matched to them accordingly are labelled as *c-jets*. If neither *b*- nor *c*- hadrons are found, the jets are labelled as *light-jets*.

5.3.1 *b*-tagging algorithms

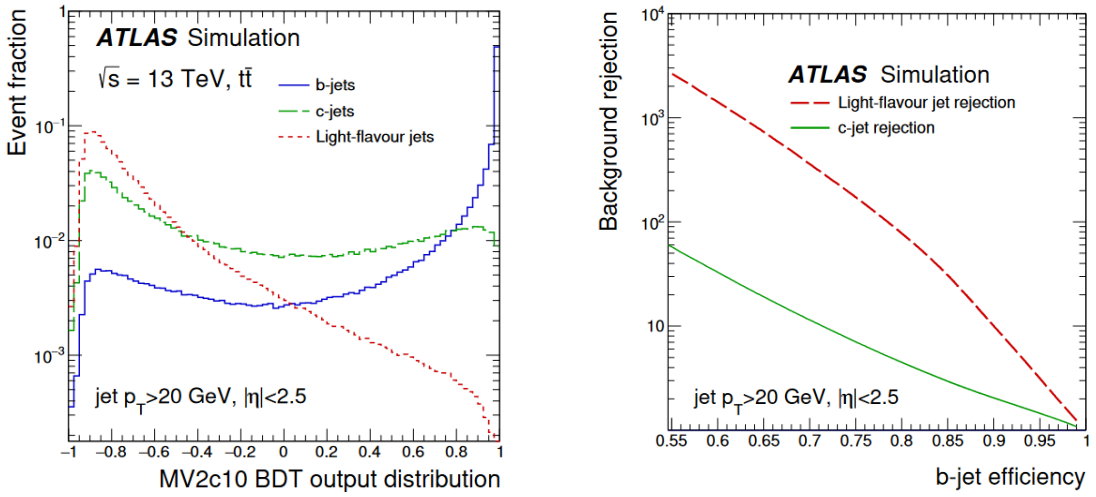
In order to identify the jets that originate from *b*-quarks, various algorithms have been developed by the ATLAS Collaboration, referred to as *b-tagging algorithms* [217]. Firstly, the *low-level algorithms* reconstruct the characteristic features of the *b*-jets via two complementary approaches. The one uses the individual properties of charged-particle tracks, with $p_T > 0.5$ GeV, associated with a hadronic jet. It is an inclusive approach that exploits the large impact parameters of the tracks originating from the *b*-hadron decay products. The other approach combines the tracks to explicitly reconstruct displaced vertices. Finally, in order to maximise the *b*-tagging performance, the results of the low-level *b*-tagging algorithms are combined into *high-level algorithms* using multivariate classifiers. The performance of a *b*-tagging algorithm is characterised by the probability of tagging a *b*-jet (*b-jet tagging efficiency*, ε_b) and the probability of mistakenly identifying a *c*-jet or a *light*-jet as a *b*-jet (*c-jet mistag-rate*, ε_c, or *light-jet mistag-rate*, ε_l).

Starting with the low-level algorithms, the *impact parameter-based* algorithm (IP3D) exploits the signed impact parameter significances ($S \equiv d_0/\sigma_{d_0}, z_0/\sigma_{z_0}$) of the tracks associated with a jet in a two-dimensional template to account for their correlation. Probability density functions (pdfs) for these significances of the tracks associated with *b*-, *c*- and *light*-jets are derived from MC simulation, and they are then combined in three log-likelihood ratio (LLR) discriminants. To further discriminate *b*- and *light*-jets, a three dimensional vertex formed by the *b*-hadron decay product, including the products of the possible subsequent *c*-hadron decay, can be sought by the *secondary*- or *multi-vertex finding algorithms* (SV1 or JetFitter). The former algorithm aims to explicitly reconstruct an inclusive displaced SV within the jet. The latter algorithm exploits the topological structure of weak *b*- and *c*-hadron decays inside the jet and tries to reconstruct two vertices for the full *b*-hadron decay chain. The properties of the reconstructed vertices, such as the displaced vertex invariant mass and track multiplicity, the fraction of the sum of the tracks energies in the vertex to the sum of the energies of all tracks in the jet, the significance of the total distance between the primary and displaced vertices, are then combined in multivariate techniques. For the SV1 algorithm, the LLR discriminant is used, based on pdfs for the *b*-, *c*- and *light*-jet hypotheses. Also, the JetFitter algorithm uses a Kalman filter [218] in order to find a common line on which the primary, bottom and charm vertices lie, approximating the b-hadron decay length and vertex positions.

The outputs of the aforementioned low-level tagging algorithms are combined using a multi-variate analysis (MVA), resulting in a high-level tagger, the MV2 [219], [220]. A boosted decision tree (BDT) (outlined in Sec. 6.5.3) is trained in order to enhance the discrimination of *b*-jets against *c*- or *light*-jets. The training is performed assigning *b*-jets as signal and a mixture of *light*- and *c*-jets as background, which specifies several variations of the MV2 algorithm. Particularly the MV2c10 variation, that is employed for the $t\bar{t}H(H \rightarrow b\bar{b})$ analysis, uses a background sample consisting of 93% *light*-jets and 7% *c*-jets.

The MV2c10 *b*-tagging algorithm assigns to each jet a b-tagging output, ranging from -1 to 1. The MV2c10 BDT output distribution for b-jets, c-jets and $light$-jets in a $t\bar{t}$ sample is illustrated in Fig. 5.6a. For *light*-jets it peaks towards -1 and for *b*-jets towards +1, while for *c*-jets the MV2c10 values tend to lie between the two. The rejection rates for $light$-jets and c-jets are defined as the inverse of the efficiency for tagging a $light$-jet or a c-jet as a b-jet, respectively. Figure 5.6b shows the corresponding $light$-jet and c-jet rejection rates as a function of the b-jet tagging efficiency.

Different working points (WPs) are defined according to the desired *b*-jet tagging effi-



(a) **(b)**

Fig. 5.6: (a) The MV2c10 output distribution for b-jets (solid line), c-jets (dashed line) and $light$-flavour jets (dotted line) in simulated $t\bar{t}$ events [221]. (b) The $light$-flavour jet (dashed line) and c-jet rejection factors (solid line) as a function of the b-jet tagging efficiency of the MV2c10 b-tagging algorithm [221].

ciency, evaluated on simulated $t\bar{t}$ events, by requiring a single-cut value on the b-tagging discriminant distribution. This distribution is also divided into five *pseudo-continuous b-tagging* (PCB) bins, delimited by the cut values used to define the b-tagging efficiency WPs, and bounded by the trivial 100% and 0% values. The available working points for the MV2c10 algorithm with the benchmark performance values are listed in Table 5.1.

PCB	WP [%]	BDT cut value	b-jet efficiency (ε_b) [%]	c-jet mistag rate (ε_c) [%]	$light$-jet mistag rate (ε_l) [%]
1	100	-1	100	100	100
2	85	0.1758	85	32	2.9
3	77	0.6459	77	16	0.77
4	70	0.8244	70	8.3	0.26
5	60	0.9349	60	2.9	0.065

Table 5.1: Working points for the MV2c10 b-tagging algorithm, including cut values for the efficiency and mistag rates, extracted from $t\bar{t}$ events with $p_T^{jet} > 20$ GeV [224].

5.3.2 *b*-tagging calibration

MC simulated samples, including the various quark flavours, are used to evaluate the b-tagging performance. However, additional calibration is often needed to account for differences between data and simulation, originating for instance from an imperfect description of the detector response or from physics modelling effects. The efficiencies of the b-tagging algorithm, derived from MC simulation, are calibrated as a function of the jet p_T and, if relevant $|\eta|$, in

order to match those in data. For the calibrations data samples enriched in *b*-, *c*-, and *light*-jets respectively, are used. Relevant scale factors $SF = \frac{\varepsilon_{data}}{\varepsilon_{MC}}$ (ε_{data} is the efficiency measured in data, while ε_{MC} denotes the efficiency predicted by the simulation) are extracted, correcting for the mismodeling in the input variables used from the *b*-tagging algorithm. The resulting *b*-, *c*-, and *light*-jet SFs are applied event-by-event by multiplying together the per-jet SFs.

The calibration of the *b*-jet efficiency is derived from $t\bar{t}$ events in the the dilepton topology (requiring two opposite-charged leptons in the final state), exploiting the very pure sample of *b*-jets arising from the decays of the top quarks [221]. A combinatorial likelihood approach is used to simultaneously extract the jet flavour composition of the sample and the *b*-jet tagging efficiencies of each WP, in a p_T range from 20 to 600 GeV.

Then, for the calibration of the *c*-jet mistag rate a sample with $t\bar{t}$ events is used in the single-lepton topology, exploiting the *c*-jets from the hadronically decaying *W*-bosons [222]. However, two real *b*-jets and a *light*-jet are expected along with the *c*-jet in the single lepton $t\bar{t}$ events final state. For this reason, a kinematic likelihood fitter (KLFitter) [223] is used to reduce the combinatorial background arising from the improper assignment of *b*-jets or additional jets in the event as decay products of the hadronic *W*-boson.

Finally, the calibration for the mistag rate of *light*-jets is calculated using the negative-tag method in high statistics data samples of Z+jets events [224]. This method relies on the assumption that *light*-jets are mistagged as *b*-jets mainly because of the finite resolution of the reconstructed ID track trajectories and impact parameters. According to the method, some of the discriminating variables of the *b*-tagging algorithm are reversed, while the mistag rate is calculated by applying the same tagging criteria, taking into account the effects of the finite detector resolution. Due to the differences in the track resolutions in the central and more forward regions of the tracking system, the efficiency and SFs of the *light*-jets are calculated for two η regions separately.

5.4 Leptons and photons

The reconstruction of photons is discussed here, just because they participate in the reconstruction of the missing transverse energy (Sec. 5.5). There is no special selection cut applied to photons in the analysis presented in this thesis.

5.4.1 Electrons and photons

Electrons can lose a significant amount of their energy due to bremsstrahlung[14] when interacting with the atomic electrons of the material they traverse. Then, the radiated photon may further split into e^+e^- pair (*photon conversion*) when interacting with the detector material. These interactions can occur inside the ID volume (even in the beam pipe), generating multiple tracks there. As electrons and photons have very similar signatures in the EMCal, their reconstruction algorithms [199], [201] proceed in parallel. Figure 5.7 illustrates the elements of the detector that enter into the reconstruction and identification of an electron and a photon.

The reconstruction of electrons (positrons) and photons takes place in the central region of the ATLAS detector, $|\eta| < 2.47$, starting from clusters of energy deposits in the EMCal. This is partially done to reduce noise arising from cells from $|\eta| > 2.5$, which have coarser granularity,

[14] *Bremsstrahlung* is the electromagnetic radiation produced by the deceleration of a charged particle when deflected by another charged particle (typically an electron by an atomic nucleus).

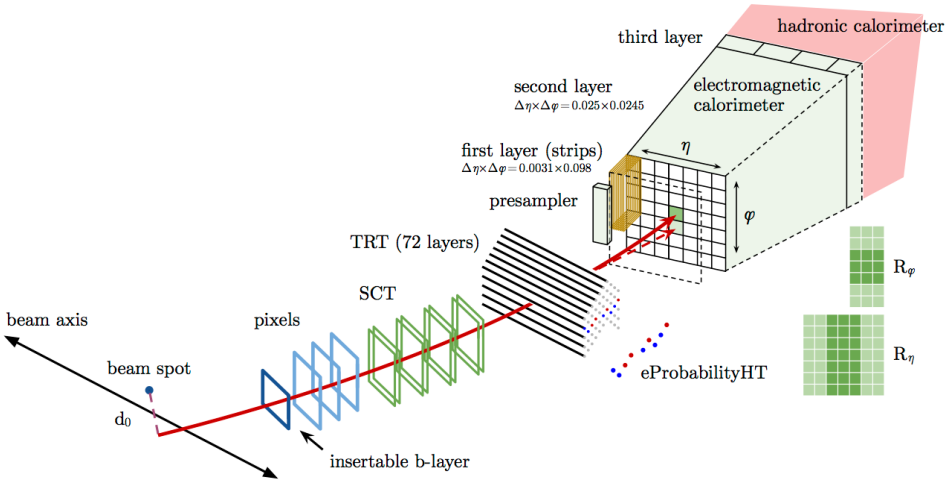


Fig. 5.7: Schematic illustration of an electron and a photon through the detector. The red trajectory shows the hypothetical path of the electron, which first traverses the tracking system (pixel detectors, silicon-strip detectors, and lastly TRT) and then enters the EMCal. The dashed red trajectory indicates the path of the photon produced by the interaction of the electron with the material in the tracking system [202].

but also due to a bad description of material in the ID, which leads to bad simulation of the electron response. However, the transition region ($1.37 < |\eta| < 1.52$) between the barrel and endcaps of the calorimeter is excluded, due to extra material and the difficulty of matching and calibrating energy from cells of different geometries and orientation. The first step is the search of cluster seeds with $\eta \times \phi = 3 \times 5$ cells and a size of $\Delta\eta \times \Delta\phi = 0.025 \times 0.025$, corresponding to the granularity of the EMCal middle layer. A sliding window algorithm scans the full EMCal acceptance and saves cluster seeds if their energy[15] is higher than the detector noise background ($E_T > 2.5$ GeV). Then, the clusters matched to a reconstructed ID track, consistent with originating from an electron produced in the beam interaction region (coming from a PV), are classified as electrons. On the other hand, clusters without matching tracks are classified as unconverted photons, while converted photons are defined as clusters matched to a track (or a two-track vertex) - coming from a SV, consistent with originating from a photon conversion in the material of the ID. In the last stage of the reconstruction, the clusters are enlarged[16], specifically in the ϕ direction, in order to capture the full electron energy including the lost energy from bremsstrahlung.

However, not all objects, built by the electron reconstruction algorithms, are signal electrons originating from the hard interaction. Background objects include hadronic jets, as well as non-signal charged particles which have similar properties (e.g. charged pions) and could be misidentified as electrons. Furthermore, background electrons can occur from photon conversions into e^+e^- pairs, leaving tracks and energy deposits in the detector, that are often very difficult to distinguish from signal electrons. Meanwhile, the identification of prompt photons in hadronic collisions is also challenging, since the overwhelming majority of final state photons originate from neutral hadron decays or from radiative decays of other particles. In order

[15]The energy collected in the first, second, and third calorimeter layers, as well as in the presampler (only for $|\eta| < 1.8$ - the region where the presampler is located), is summed to form the energy of the cluster seeds.

[16]Optimal cluster size for electrons and converted(unconverted) photons: $\eta \times \phi = 3 \times 7(3 \times 5)$ cells in the EMCal barrel and $\eta \times \phi = 5 \times 5$ EMCal end-cap

to properly identify the signal-like electrons and photons with high efficiency, while rejecting as much of these backgrounds as possible, further criteria are required.

Electron and photon identification [200], [201] in ATLAS is based on quantities related to the electron and photon clusters, describing the shape and properties of the associated electromagnetic showers. In case of electrons, information from the TRT, track-cluster matching related quantities and track properties are also considered. In addition, the electron identification is performed by imposing sequential requirements on these discriminating quantities, known as *cut-based identification*, or a multivariate analysis technique using the signal and background probability density functions (pdfs) of the discriminating variables, referred to as *likelihood-based (LH) identification*[17]. The LH method provides higher background rejection for a given signal efficiency compared to the cut-based identification.

Three identification operating points are provided for the electron identification in Run 2, based on the identification efficiency of an electron at $E_T = 40$ GeV. They are referred to as *Loose* (93%), *Medium* (88%), and *Tight* (80%), starting from the highest efficiency. They are defined in such a way that each operating point uses the same variables to define the LH discriminant but a different cut value is applied on this discriminant. Therefore, the samples selected by the three operating points are subsets of one another: electrons selected by *Tight* are all selected by *Medium* and those selected by *Medium* are also selected by *Loose*. The identification operating points are optimised in several bins of η and E_T. On the contrary, the photon identification can be performed by applying a set of cuts on the aforementioned discriminating quantities. Two working points are provided for the photon identification, *loose* and *tight*, which are optimised in bins of η. Also, the *tight* working point is separately optimised to differentiate between converted and unconverted photons.

In order to discriminate the prompt[18] from non-signal (non-prompt[19] or fake[20]) electrons, that typically have a large activity in the proximity of the electron, and further suppress this background, electron isolation criteria [203] are imposed. There are two kinds of isolation requirements, depending on calorimeter energy deposit or track p_T densities, that rely on the relation of reconstructed electrons to other objects. The calorimeter-based isolation, is calculated via the sum of the energy of the calorimeter clusters in a cone of radius $\Delta R = 0.2$ around the candidate electron, removing cells that contain the electron energy deposits. Lower values are not practically accessible due to the granularity of the EMCal. In addition, energy leakage and pile-up corrections are applied as well. The track-based isolation is estimated via the variable $p_T^{\text{varcone}0.2}$. It is estimated as the sum of the transverse momentum of the tracks within a cone of $\Delta R = \min\left(\frac{10\text{ GeV}}{p_T^e\,[\text{GeV}]}, 0.2\right)$ centered around the electron track originating from the PV of the hard collision, excluding tracks coming from the electron and its radiation. The

[17] Based on the signal and background pdfs $P_{s(b),i}(x_i)$ (of the i^{th} variable evaluated at x_i) - which are treated as uncorrelated, an overall probability is calculated for the object to be signal $\mathcal{L}_{\mathcal{S}}(\vec{x})$ or background $\mathcal{L}_{\mathcal{B}}(\vec{x})$. These probabilities are then combined into a discriminant d_L:

$$d_{\mathcal{L}} = \frac{\mathcal{L}_{\mathcal{S}}}{\mathcal{L}_{\mathcal{S}} + \mathcal{L}_{\mathcal{B}}}, \qquad \mathcal{L}_{\mathcal{S}(\mathcal{B})}(\vec{x}) = \prod_i^n P_{s(b),i}(x_i) \tag{5.9}$$

where $\vec{x}$ is the vector of discriminating variable values.

[18] Real electrons (muons) produced from heavy short-lived particle decays (like *W/Z*-bosons, top quarks and tau leptons), or originating from the hard interaction, are called *prompt* electrons (muons).

[19] Real electrons (muons) that do not originate from the hard scatter, but are produced in subsequent heavy hadron decays or, especially in case of electrons, in photon conversions, are called *non-prompt*.

[20] Signals being selected as electrons (muons) but without a real electron (muon) being present, coming mostly from pions and kaons, are called *fake* electrons (muons).

selected area used in the isolation selection can be flexible compared to the calorimeter due to the higher granularity of the ID. Additional track quality and kinematic requirements in order to suppress tracks from pile-up are also considered. Several isolation working points are available, based on the isolation efficiency. Nevertheless, in the context of this analysis only the *Gradient* isolation is important, which has a p_T^e dependent requirement on the efficiency while being uniform in η, giving an efficiency of 90% (99%) at $p_T = 25$ (60) GeV. Analogous isolation criteria [203], based on the energy deposited in the calorimeters, in a cone around the photon candidate, can be used to further suppress the main background from neutral hadrons decaying into two photons.

The reconstruction, identification, and isolation efficiencies of electrons (photons) are measured with the data-driven tag-and-probe method [203]. This method selects, from well-known resonances such as $Z \to e^+e^-$ and $J/\psi \to e^+e^-$ ($Z \to l^+l^-\gamma$), unbiased samples of electrons (probes) by using tighter selection requirements on the second object (tags) produced from the heavy particle's decay. The events are selected on the basis of the electron–positron invariant mass. The efficiency of a given requirement can then be determined by applying it to the probe sample after accounting for residual background contamination. However, the simulation of the detector is not perfect and differences arise in electron response between data and MC. In order to compensate for differences in electron reconstruction, identification, and isolation algorithms performance in data and MC, the ratio of their efficiencies measured in each of these algorithms are used to derive the corresponding SFs. The latter are subsequently applied as an event weight to MC events, so that they agree with data. Finally, the energy scale and resolution of electrons (photons) are calibrated using $Z \to e^+e^-$ and $J/\psi \to e^+e^-$ ($Z \to l^+l^-\gamma$) decays [203]. The energy resolution of the electron (photon) is particularly optimised using a multivariate regression algorithm based on the properties of the shower development in the EMCal.

5.4.2 Muons

Like any other charged particle, muons are reconstructed from tracks in the ID, according to the standard ID tracking algorithms (Sec. 5.1.1). Then, since they have a low interaction rate with the material, they traverse the calorimeter system typically without significant energy loss. However, they leave tracks also in layers furthest from the center of the detector, in the muon system. The muon reconstruction [204] in the MS (Sec.3.2.4) is performed independently from that in the ID. In the MS, it starts by searching for MDT hits inside each muon chamber, which form segments after a straight-line fit is applied. Also, a combinatorial search is performed to associate tracks to CSC segments. Muon track candidates are then built by fitting together the hits from segments in the different layers. These segments are selected using criteria based on hit multiplicity and fit quality. The MS muon tracks are then obtained by applying a χ^2 fit to the hits. Eventually, the muon tracks are formed according to various algorithms which combine hit and track information provided by the ID, MS, and calorimeters.

There are four types of reconstructed muons depending on which subdetectors are used in the reconstruction. The *Combined* (CB) *muons* are reconstructed by independent tracks in the ID and MS, and a combined track is formed with a global refit that uses the hits from both subdetectors. These muons are reconstructed in the MS (ID) and then extrapolated inward (outward) and matched to an ID (MS) tracks. Both approaches are used as complementary. The *Segment-tagged* (ST) *muons* are reconstructed from ID tracks, extrapolated to typically one track segment in the MDT or CSC and are mostly used with muons crossing only the first

MS layers[21]. In addition, the *Calorimeter-tagged* (CT) *muons* are reconstructed from ID tracks that are matched to an energy deposit in the calorimeter, compatible with a minimum-ionising particle. Although this type has the lowest purity, it recovers acceptance in the region where the MS is only partially instrumented[22]. Lastly, the *Extrapolated* (ME) *muons* are reconstructed only from MS tracks originating from the interaction point, taking into account the energy loss of the muons in the calorimeters, extending the acceptance for muon reconstruction into the region $2.5 < |\eta| < 2.7$ which is not covered by the ID. Only the CB muons are used in the $t\bar{t}H(H \to b\bar{b})$ analysis.

Muon identification [204] is performed in order to discriminate prompt muons from background events, containing particles that escape the inner parts of the ATLAS detector, mainly from pion and kaon decays. Four identification working points (WPs) are defined using basic muon quality cuts: *Loose*, *Medium*, *Tight*, and *High-*p_T with a muon identification efficiency of approximately 98%, 96%, 92%, and 80%, respectively. For the *Medium* WP, CB muons are further required to have at least 3 hits (1 hit and at most 1 hole[23]) in at least two MDT layers for $|\eta| > 0.1$ ($|\eta| < 0.1$). Additionally, a loose requirement on the ID and MS momentum compatibility is applied, in order to suppress the contamination due to hadrons misidentified as muons. This selection minimises the systematic uncertainties associated with muon reconstruction and calibration. The *Loose* identification criteria are designed to maximise the reconstruction efficiency while providing good-quality muon tracks. The *Tight* WP maximises the muon purity but loses in efficiency. Only CB muons satisfying the *Medium* criteria with enhanced track quality cuts are identified as *Tight*. Finally, the *High-*p_T WP targets muons with $p_T > 100$ GeV in order to maximise the momentum resolution. CB muons passing the *Medium* criteria and having at least three hits in three MS stations are selected. Eventually, the *Medium* WP is used in this analysis.

Muons originating from the decay of heavy particles, such as *W-/Z-*, or Higgs bosons, are often produced isolated from other particles; unlike muons from heavy-flavour hadron semileptonic decays, which are embedded in jets. In order to further reduce the contamination from non-prompt or fake muons, additional muon isolation criteria [204] on a low activity of particles around the lepton are applied, for which seven working points are provided. Similarly to electrons, the isolation requirements depend on a combination of track-based and calorimeter-based variables. The calorimeter-based isolation parameter is defined as the sum of the energies of topoclusters (defined in Sec. 5.1.3) in a cone of $\Delta R = 0.2$ around the muon, subtracting the energy deposit of the muon as well as pile-up contributions. In contrast to electrons, since muons are minimum-ionising particles, they do not contribute as much to the calorimeter deposition. The track-based isolation parameter, $p_T^{\text{varcone0.3}}$, is estimated by the scalar sum of the transverse momenta of the tracks surrounding the muon within a cone of flexible size $\Delta R = \min\left(\frac{10\text{ GeV}}{p_T^{\mu}[\text{GeV}]}, 0.3\right)$, excluding the muon track. This p_T dependence improves the performance at high p_T while keeping a reasonable cone size of 0.3 at low p_T. Both parameters are divided by the p_T^{μ} of the muon. For the scope of this analysis, the *FixedCutTightTrackOnly* isolation operating point is employed, which relies only on the track component of the isolation satisfying $p_T^{\text{varcone0.3}}/p_T^{\mu} < 0.06$.

Similarly to the electrons, the muon performance is studied in $Z \to \mu^+\mu^-$ and $J/\psi \to$

[21] These muons have either low p_T, or they fall in MS regions of reduced acceptance.

[22] The MS $|\eta| < 0.1$ region is partially instrumented in order to allow for cabling and services to the calorimeters and ID.

[23] A hole is defined as an active sensor traversed by the track but containing no hits.

$\mu^+\mu^-$ decays, where a pure sample of muons can be studied using the tag-and-probe method [204]. Differences in the efficiency between data and simulation are then corrected for all stages of the muon reconstruction, identification and isolation, applied as event weights to MC. The muon momentum scale and resolution are calibrated also using events from the above processes. Correction factors, as a function of the muon momentum in various η regions, are derived by fitting the dimuon resonance peak and applied to the simulated muon momentum to match data.

5.4.3 Tauons

Tauons (tau-/τ-leptons) are the heaviest leptons with a mass of 1.777 GeV, a lifetime of 2.9×10^{-13} s and a decay length of 87 μm [101]. They result from the leptonic *W*-boson decay, and they quickly decay either leptonically or hadronically as mentioned Sec. 2.5.2 (in 35% or 65% of the cases, respectively), before reaching any detector layer. Thus, they can only be identified via their decay products. Tauons are discussed here because a veto is applied to the hadronically decaying tau leptons in order to maintain orthogonality with other $t\bar{t}H$ channels. Otherwise, they are not further used in the analysis.

Ideally leptonically decaying tau leptons would be identified as an electron or a muon associated to a track not pointing towards the PV and with missing energy. However, due to their short decay length, it is difficult to distinguish tau leptons decaying to electrons or muons from prompt electrons and muons. This is also because the only other particle produced in their decay, the neutrino, is not detected. Therefore, tau identification focuses on reconstructing hadronically decaying tau leptons (τ_{had}).

The overwhelming majority (>90%) of the hadronic tau decays yield in one (*1-prong*) or three (*3-prong*) charged hadrons (mainly pions and rarely kaons), up to two neutral pions and a tau neutrino. The neutrino passes through the detector undetected, hence the neutral and charged hadrons make up the visible decay products of the tau lepton. The main background to hadronic tau-lepton decays is from jets of energetic hadrons produced via the fragmentation of quarks and gluons. Other important backgrounds are electrons (also muons), which can mimic the signature of tau-lepton decays with one charged hadron.

The hadronically decaying τ-leptons are reconstructed [205] using jets and their associated tracks - within the core region ($\Delta R < 0.2$ from the initial jet-axis). Since tau leptons decay via the weak interaction, they are expected to give narrower jets and low track multiplicities compared to gluons or quarks. Therefore, discriminating variables based on the narrow shower shape, the distinct number of charged particle tracks, the displaced tau-lepton decay vertex, and the kinematic information from tracks and jets, are employed for the hadronic τ-identification. These variables are combined in a multivariate algorithm that employs Boosted Decision Trees and likelihood methods. The output distributions of these techniques are used to discriminate the tau leptons from the QCD-jets and electrons. Three working points, labeled as *Tight*, *Medium* and *Loose*, are provided, corresponding to different τ-identification efficiency[24] values.

Given that the hadronic decays of τ-leptons can be reconstructed as jets, which can mimic *b*-jets, it is useful to discriminate the ones from the others. Thus, jets containing hadronically decaying τ-leptons and no *b*- or *c*-hadrons are labelled as τ-jets from the *b*-tagging algorithm.

[24]The identification efficiency is defined as the fraction of 1-prong (3-prong) hadronic tau decays reconstructed as 1-track (3-track) hadronic tau candidates, which also pass the BDT selection criteria.

5.5 Missing Transverse Energy

The energy (momentum) after a collision is equal to the energy (momentum) before the collision, according to the energy-momentum conservation law. However in hadron collisions, such as pp collisions, only part of the initial protons reacts, while the rest of them fly along the beam pipe undetected. As already discussed in Sec. 4.1.2, it is impossible to know the fraction of the centre-of-mass energy contained in the colliding partons. Furthermore, these undetected particles also carry away some energy which cannot be measured, because the longitudinal component of their momenta is unknown. As a result, the energy-momentum conservation law cannot be used in the total system.

In pp collisions at the LHC, the proton beams collide head on in the longitudinal direction thus, the initial partons have no momentum in the plane transverse to the beam axis before scattering. Therefore, the conservation of momentum in the transverse plane implies that the transverse momenta ($\vec{p}_T$) of the collision products would sum to zero, if all particles in the final state were detected. The sum of the transverse momenta of the visible particles, is known as *missing transverse momentum* ($\vec{p}_T^{miss}$). A deviation of this sum from zero may indicate weakly interacting particles, which traverse the detector without being detected, since they deposit hardly any energy in the detector material. Within the SM these are the neutrinos, but there are also prospects for new particles suggested in theories beyond the Standard Model (BSM). Moreover, the p_T^{miss} measurement is affected by interacting SM particles which are poorly reconstructed thus, it is an important measure of the overall event reconstruction performance.

The missing transverse momentum is reconstructed as the negative vector sum of the transverse momenta of all detected particles in the final state, $\vec{p}_T^{miss} = -\sum_i \vec{p}_T^{i}$. The magnitude of the missing transverse momentum vector is the *missing transverse energy* E_T^{miss} and the azimuthal angle ϕ^{miss} is its direction in the transverse plane [206]. The measurement of E_T^{miss} strongly depends on the energy scale and resolution of the selected reconstructed and fully calibrated physics objects (electrons, muons, photons, hadronically decaying τ-leptons, and jets) which constitute the *hard term* of the E_T^{miss}. In addition, the E_T^{miss} *soft term* [207] is added to account for soft radiation. It contains momentum contributions from reconstructed ID tracks, which are associated with the hard-scattering PV, that are not attributed to any of the physics objects included in the hard term.

The reconstruction of E_T^{miss} is challenging because it involves all detector subsystems and requires the most complete and unambiguous representation of the hard interaction of interest by calorimeter and tracking signals. This representation is obscured by limitations introduced by the detector acceptance and by signals and signal remnants from additional *pp* interactions occurring in the same, previous and subsequent LHC bunch crossings (pile-up) which overlap with the hard-scattering process. Both the hard and soft terms are affected by pile-up, thus various techniques have been developed in order to suppress the pile-up effects [206]. In general, the E_T^{miss} reconstruction is based on the combined information from the energy deposits in the calorimeters and the ID tracks associated to the hard-scattering vertex. The former is more vulnerable to pile-up effects, but provides information about the neutral particles. The latter is insensitive to neutral particles, since they do not leave tracks in the ID, but offers greater resilience under conditions of increased pile-up.

The missing transverse energy is discussed in this thesis because the leptonic final states (from the W-boson decay) in the $t\bar{t}H(H \rightarrow b\bar{b})$ analysis contain neutrinos, which are not detected and only appear as E_T^{miss}.

Chapter 6

Analysis Strategy in the $t\bar{t}H(H \to b\bar{b})$ Single-Lepton Boosted Channel

As already introduced, a measurement of the $t\bar{t}H$ production cross-section in the $H \to b\bar{b}$ decay channel, with the full LHC Run 2 dataset, is presented in this dissertation. The specific signature, that is studied throughout this analysis, has been introduced in Sec. 2.6. Events with either one or two leptons are analysed separately in exclusive single-lepton or dilepton categories defined according to the number of leptons, the number of jets and the number of jets identified as originating from *b*-hadrons (*b-jets*). Consequently, *b*-tagging is crucial for this analysis. Particularly in the single-lepton channel, a specific category, referred to as "boosted" in the following, is designed to select events in which the Higgs boson and possibly also the hadronically decaying top quark are produced with high transverse momentum relative to their rest mass. As a result, their decay products are collimated in large-*R* jets. Due to the fairly high branching ratio, the single-lepton boosted channel provides sufficient statistics also in higher p_T regimes. Nevertheless, no boosted category can be defined in the dilepton channel because the expected number of events in the high-p_T regime is small.

The highly complex final state of the $t\bar{t}H(H \to b\bar{b})$ topology (fig. 2.15) with many jets and the overwhelming $t\bar{t}$+jets background, poses great challenges. Especially, the irreducible $t\bar{t}+b\bar{b}$ background, coming from a splitting of a gluon emission (fig. 2.15c), has similar kinematics to the $t\bar{t}H(H \to b\bar{b})$ signal and a cross-section about one or two orders of magnitude larger than that of the $t\bar{t}H$ production depending on the analysis phase space [290]. In addition, it is quite difficult to model due to the high number of jets in the final state and the presence of heavy-flavour particles with significantly different masses. Thus, it is poorly constrained by data measurements and has large theory uncertainties which limit the analysis.

At first, events are selected creating a phase space with enhanced signal contribution. Then, the events are split into signal-depleted and signal-enriched categories in each channel. In addition, machine-learning algorithms are exploited to improve the event reconstruction, by matching jets to the final state partons from the top-quark and Higgs-boson decays. Various topological and kinematic discriminating variables are defined based on the reconstructed Higgs-boson and top-quark candidates. Particularly in the signal-enriched regions, some of these variables are combined with other multivariate analysis techniques to better classify events between signal and background. The output distributions of these multivariate algorithms are used as the main discriminant to extract the signal (detailed in Ch. 8). Furthermore, making use of the possibility to reconstruct the Higgs-boson kinematics in the $H \to b\bar{b}$ mode,

the cross-section is measured as a function of the Higgs boson "true" transverse momentum $\hat{p}_T^H$ in the simplified template cross-sections formalism (STXS). This $\hat{p}_T^H$ is the p_T of the true Higgs-boson object before it decays as obtained from Monte Carlo simulation. An STXS measurement allows for and benefits from the combination of measurements in all decay channels.

In this chapter, the $t\bar{t}H(H \to b\bar{b})$ analysis overview is presented with a focus on the single-lepton boosted channel, which is the main subject of this study, giving insights in its motivation, challenges and strategy. Nevertheless, a brief overview of the single-lepton and dilepton resolved regions is also given for completeness, since all the analysis regions are combined in the signal extraction fit (Ch. 8).

6.1 Simplified Template Cross-Section

The STXS formalism [106, 232] has been developed and adopted by the LHC experiments in order to provide a consistent basis for finely-grained measurements for individual Higgs production modes in various kinematic regions. The primary goals of the STXS framework are to maximise the sensitivity of the measurements, while simultaneously eliminating the dependence on the theoretical uncertainties that are directly folded into the measurements.

With the STXS formalism, differential measurements of the physicsl cross-sections are performed in mutually exclusive kinematic phase-space regions ("STXS bins") to avoid introducing statistical correlations among the different bins. The number of STXS bins is kept minimal to avoid losing experimental sensitivity, while the bins are common in each individual analysis. In parallel, the STXS allows for the use of advanced analysis techniques such as event categorisation or multivariate techniques optimised to achieve maximal sensitivity. As a consequence, a common framework is provided for a subsequent global combination of all measurements in the different decay channels as well as between the ATLAS and CMS experiments. Eventually, tests of the SM in the kinematics of the different Higgs-boson production modes with an improved sensitivity are allowed when combining all decay channels.

Furthermore, in order to reduce the dependence on theory predictions and uncertainties that are folded into the measurements, extrapolations of the measurement from a certain phase-space region to the full phase space should be avoided, especially when they carry sizeable theoretical uncertainties. To avoid such extrapolations, the STXS bins are preferably defined by quantities that are directly measured by the experiments, such as the true $\hat{p}_T^H$.

After the discovery of the most prominent Higgs-boson decay channels, the Run 2 statistics allow to perform differential cross-section measurements also in the $t\bar{t}H$ production channel. In parallel, the $H \to b\bar{b}$ decay mode allows to probe the differential cross sections due to its large decay rate. Thus, even the high Higgs-boson p_T regime can be assessed, in which e.g. the $t\bar{t}H(H \to \gamma\gamma)$ decay mode is lacking statistics [94]. Additionally, the possibility to reconstruct the Higgs-boson kinematics in the $t\bar{t}H(H \to b\bar{b})$ analysis is exploited for an improved definition of the STXS bins. Consequently, a differential cross-section measurement is performed in the $t\bar{t}H$ production channel as a function of the true $\hat{p}_T^H$ with the STXS formalism. According to the recommendations for the STXS bins in other Higgs-boson production channels [232], the following STXS bins are used in this analysis: $\hat{p}_T^H \in$ [0,120), [120,200), [200,300), [300,450), [450,∞) GeV. For each STXS bin, a separate $t\bar{t}H$ signal template is defined which is the signal MC prediction in the targeted kinematic region at true level. Finally, such a measurement in the $t\bar{t}H$ channel allows to access the CP structure of the Higgs boson [234] and to probe anomalous Higgs self-couplings [235] with increased sensitivity.

6.2 Boosted topology

With the increase in the center-of-mass energy at $\sqrt{s} = 13$ TeV during Run 2, which far exceeds the masses of the known SM particles, the LHC has been exploring a completely new physics regime. At such energies, more particles than ever before, including heavy particles such as *W*-, *Z*-, Higgs bosons and top quarks, are often produced with high transverse momentum that implies large Lorentz boost for their decay products. For this reason, such particles are called *boosted particles* and their decay products are collimated to the momentum direction of the boosted parent particle in the detector rest frame.

Traditionally, the decay products of heavy particles are reconstructed individually. In particular, top quarks, decaying through $t \to Wb \to qqb$, and Higgs bosons, decaying via $H \to b\bar{b}$, would be typically reconstructed as three or two well-separated and approximately-conic sprays of mostly hadronic particles, the jets (Sec. 5.2). This is referred to as the *resolved topology*. However, these decay products become collimated at high p_T, eventually causing the jets to overlap and become unresolved. An alternative topology, referred to as the *boosted regime*, utilises this collimation by reconstructing high p_T heavy particles as a single larger radius jet, as demonstrated in Fig. 6.1. It depicts the jet configuration of a low-p_T Higgs boson decaying to *b*-hadrons, represented by two small-*R* jets according to the resolved regime (on the left). While p_T increases, the angular separation of these jets is reduced (in the middle), resulting in a configuration where the two *b*-quarks appear in a single large-*R* jet (on the right).

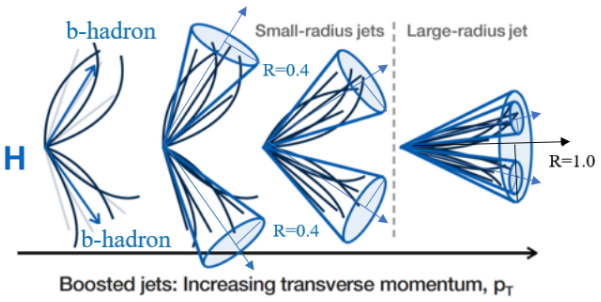


Fig. 6.1: Graphical representation of the transition between the resolved and the boosted regimes for the Higgs boson decay, while the Higgs p_T increases.

In general, the angular separation, ΔR, of the decay products of a particle, with mass *m* and transverse momentum p_T, is approximately estimated by

$$\Delta R \simeq \frac{2m}{p_T}, \qquad \text{where} \quad \Delta R = \sqrt{(\Delta y)^2 + (\Delta \phi)^2}, \tag{6.1}$$

where y is the rapidity and ϕ is the azimuthal angle around the beam axis (see Sec. 3.2.1). According to eq. 6.1, a jet of radius $R = 1.0$ can typically fully contain a Higgs boson with $p_T \geq 250$ GeV, or a top quark with $p_T \geq 350$ GeV. This is indeed evident in Fig. 6.2, that shows the true angular separation between the two *b*-quarks of a Higgs boson, $\Delta R(b, \bar{b})$, as a function of the true Higgs $\hat{p}_T^H$ in simulated $t\bar{t}H$ events. Besides, the ability to resolve the individual Higgs-boson decay products using standard narrow-cone (small-*R*) jets degrades further for $\hat{p}_T^H \geq 300$ GeV, since the two *b*-quarks tend to have a separation $\Delta R(b, \bar{b}) \leq 0.8$, where the two small-*R* jets start to overlap.

Eventually, the traditional reconstruction algorithms, according to which individual decay products are described by small-R jets, loose significantly their efficiency at high p_T, due to

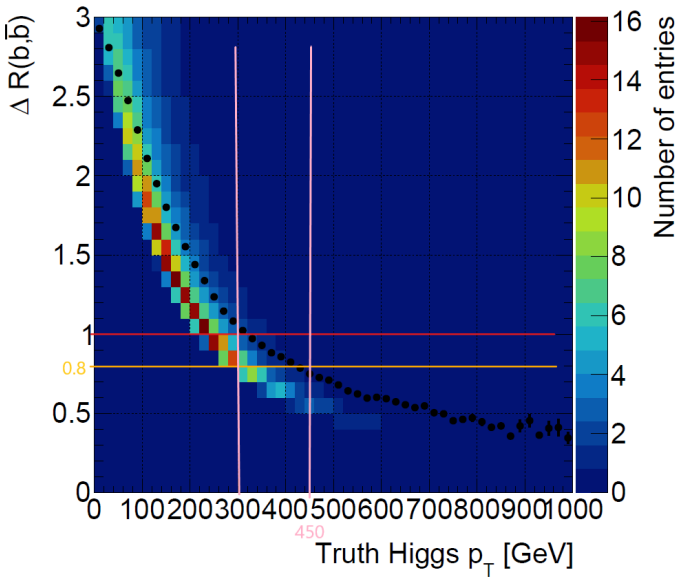


Fig. 6.2: True angular separation between the two b-quarks from a Higgs boson decay, $\Delta R(b, \bar{b})$, as a function of the true Higgs $\hat{p}_T^H$ in simulated $t\bar{t}H$ events. The dots on the plot represent the median of each column.

the overlap of the jets coming from a hadronic decay of a parent particle. As a consequence, at high p_T, the decay products of a hadronically decaying object merge into a single, energetic and large-radius jet. Especially for this analysis, the RC jets are exploited (introduced in Sec. 5.2.5) with a characteristic substructure different from those initiated by a single parton. The advantage of the boosted topology is that it reduces the combinatoric ambiguity of which jet originated from which decay product. In addition, the selection efficiency increases relative to the resolved regime as p_T increases, until the angular separation $\Delta R(b, \bar{b})$ gets close to 0.4, allowing to explore higher energies of the $t\bar{t}H(H \to b\bar{b})$ final state.

6.3 Reconstructed object and event selection

The events for the $t\bar{t}H(H \to b\bar{b})$ analysis are extracted from the full Run 2 dataset, selected by specific triggers. The main physics objects considered in this analysis are electrons, muons, jets and b-jets. The reconstruction, identification, isolation, and calibration definitions of these objects are described in Chapter 5. Various selection requirements are applied on the events at the detector level, in order to define the single-lepton and dilepton regions aiming at the $t\bar{t}H$ signal and rejecting all the relevant background processes.

6.3.1 Dataset and trigger requirements

The events for this analysis are selected from *pp* collision data recorded by the ATLAS detector between 2015 and 2018 at a centre-of-mass energy $\sqrt{s} = 13$ TeV, corresponding to an integrated luminosity of 139 fb^{-1}. Only events for which the LHC beams were in stable-collision mode and all relevant ATLAS subsystems were operational are considered [236]. Throughout Run 2, the primary luminosity measurements were obtained using the LUCID-2 detector [230]. Table 6.1 summarises the integrated luminosity for partial datasets together with their relative

uncertainties. The total integrated luminosity for the full Run-2 dataset is 139 fb^{-1} with an uncertainty of 1.7% [237], which is lower than the uncertainty of any of the partial datasets, due to correlations.

Years	$\int Ldt$ [fb^{-1}]	Uncertainty [%]
2015-2016	36.2	2.1
2017	44.3	2.4
2015-2017	80.5	2.0
2018	58.5	2.0
2015-2018	139.0	1.7

Table 6.1: Integrated luminosity for the full LHC Run 2 and for partial datasets and their relative uncertainties [237]. The table indicates the integrated luminosity for single-lepton triggers used in the analysis.

The data collected by the ATLAS experiment are recorded using specific triggers (see Sec. 3.2.5). In this analysis, at least one electron or one muon is expected from the single- or di-leptonic top-quark decay. Therefore, the analysis is based on events where the detector read-out is triggered by the presence of at least one electron or one muon, referred to as *single-lepton triggers*. Each trigger is defined by a p_T threshold of the lepton and by identification and isolation requirements.

In terms of the analysis, triggers with different lepton p_T thresholds are combined in a logical OR in order to maximise the overall efficiency. In particular, the events are required to fire triggers with a low lepton p_T threshold and an isolation requirement on the candidate lepton. Nevertheless, such triggers result in inefficiency at high p_T, which can be recovered by triggers with a higher threshold but a looser identification criterion and without any isolation requirement. The isolation requirement, at low-p_T threshold, is applied in order to keep the trigger rate under control and to reduce the high trigger rate of leptons produced in hadron decays. It also reduces the amount of fake and non-prompt leptons (defined in Sec. 5.4.1), i.e. the multijet background, so that it becomes negligible in this analysis. In contrast, at high-p_T threshold this background is not significant, thus the isolation requirement can be dropped, so as to increase the trigger efficiency. Furthermore, the trigger p_T requirement subsequently impacts the event selection at detector level. Therefore, the lepton p_T requirement in the analysis (mentioned in Sec. 6.3.2) is typically ~ 1 GeV higher than the trigger requirement, in order to remove a region where the efficiency of the trigger is not well measured, resulting in large uncertainties.

In particular, events were recorded using the lowest unprescaled[1] single-lepton triggers, meaning that the lowest p_T threshold at trigger level was used, which for muons is 20 (26) GeV [238], while for electrons it is 24 (26) GeV [239] in 2015 (2016-2018) data taking. The exact selection on the lepton candidates is differentiated between the 2015 and 2016-2018 data sets to take into account the lower instantaneous luminosity which allows for lower thresholds. Table 6.2 summarises the triggers used in the analysis for the 2015 and 2016-2018

[1]Unprescaled refers to triggers which have no prescales applied. Prescales are factors that allow the experiment to either disable triggers completely or to set the fraction of events accepted by them.

data taking periods.

Period	Object	p_T [GeV]	identification	isolation
2015	electron	24	medium	-
		60	medium	-
		120	loose	-
	muon	20	loose	-
		50	-	-
2016 - 2018	electron	26	tight	loose
		60	medium	-
		140	loose	-
	muon	26	medium	medium
		50	-	-

Table 6.2: Single-electron and single-muon trigger menus used in the dilepton and single-lepton channels, depending on the year of data-taking. The selected object, its p_T threshold, as well as the identification and isolation working points are shown.

6.3.2 Object and event selection at detector level

First and foremost, the events are required to have at least one primary vertex (PV) (defined in Sec. 5.1.2) associated with two or more tracks with $p_T > 0.5$ GeV [196], in order to enhance the resolution on the vertex spatial position. In case more than one vertex is found, the hard-scattering PV is separated from pile-up vertices by selecting the one with the highest sum of squared transverse momenta of associated tracks.

Electrons are reconstructed from clusters of energy deposits in the EMCal associated with tracks reconstructed in the ID, as described in Sec. 5.4.1, and are required to have $p_T > 10$ GeV and be in the cetnral calorimeter region $|\eta| < 2.47$. However, candidates in the calorimeter barrel-endcap transition region ($1.37 < |\eta| < 1.52$) are excluded. Additionally, electrons must satisfy the *Medium* likelihood identification criterion, based on a likelihood discriminant combining observables related to the shower shape in the calorimeter and to the track matching the electromagnetic cluster. Finally, electron tracks should match the PV of the event, thus the longitudinal impact parameter is required to satisfy $|z_0 sin(\theta)| < 0.5$ mm, while the transverse impact parameter significance is $|d_0/\sigma(d_0)| < 5$.

Muons are reconstructed from either track segments or full tracks in the MS which are matched to tracks in the ID corresponding to the *combined muon* type, as described in Sec. 5.4.2. For the reconstruction, the *Loose* identification criterion is used. Tracks are then re-fitted using information from both detector systems. Muons are required to have $p_T > 10$ GeV as electrons, but a slightly larger η acceptance $|\eta| < 2.5$. Muon tracks must also match the PV of the event, which is ensured by the requirements for the longitudinal and transverse impact parameters, $|z_0 sin(\theta)| < 0.5$ mm and $|d_0/\sigma(d_0)| < 3$, respectively. The impact parameter requirements in muons, as well as in electrons, are optimised to reduce the amount of fake and non-prompt leptons, which are considered backgrounds in lepton identification.

As already pointed out, events are required to have exactly one lepton in the single-lepton channel, while in the dilepton channel exactly two leptons with opposite electric charge are needed. In both channels, at least one reconstructed lepton is required to have $p_T > 27$ GeV, so as to ensure full trigger efficiency, and match a lepton with the corresponding flavour reconstructed by the trigger algorithm within $\Delta R < 0.15$. In the dilepton channel, events are categorised into ee, $\mu\mu$ and $e\mu$ samples. The sub-leading lepton must have $p_T > 15$ GeV in the *ee* channel, or $p_T > 10$ GeV in the $e\mu$ or $\mu\mu$ channels. Additionally, in the *ee* and $\mu\mu$ channels, the invariant mass of the two leptons must be $m_{ll} > 15$ GeV in events with more than two *b*-jets, to suppress contributions from the decay of hadronic resonances, such as the J/ψ and Υ, into a same-flavour lepton pair. A further cut on m_{ll} is applied in the *ee* and $\mu\mu$ categories to reject events close to the *Z*-boson mass, $|m_{ll} - m_Z| > 8$ GeV $\Rightarrow m_{ll} \notin [83, 99]$ GeV, reducing the contribution from *Z*+jets events.

Events which fail the dilepton channel requirements and contain exactly one lepton with $p_T > 27$ GeV are considered for the single-lepton channel. Moreover, to improve the purity in dilepton and single-lepton events, the leading-p_T leptons are further required to satisfy additional identification and isolation criteria. In particular, electrons (muons) must pass the Tight (Medium) identification criterion and the Gradient (FixedCutTightTrackOnly) isolation criteria (outlined in Sec. 5.4.1 and 5.4.2). This is to enhance the selection of prompt leptons, increasing the background rejection of fake leptons.

Jets are reconstructed from three-dimensional noise-suppressed topoclusters of calo- rimeter energy depositions calibrated at the EM-scale (presented in Sec. 5.1.3). For the jet reconstruction the anti-k_t jet algorithm is used, implemented in the FastJet package [228] with a radius parameter of $R = 0.4$, resulting in the EMTopo small-*R* jets (detailed in Sec. 5.2.1). The reconstructed jets are then calibrated to the particle-level energy scale with a series of simulation-based corrections and in situ techniques, as described in Sec. 5.2.2. Also, the average energy contribution from pile-up is subtracted according to the jet area. After the energy calibration, jets are required to satisfy the kinematic requirements $p_T > 25$ GeV, since low p_T jets are possibly not detected due to poor detector resolution or pile-up effects, and $|\eta| < 2.5$ due to the geometry of the detector. Moreover, quality criteria are imposed to identify jets arising from non-collision sources or calorimeter noise, and any event containing such a jet is removed [226]. In order to further reduce the effect of pile-up, the JVT (described in section 5.2.4), which matches the calorimeter-based jets to tracks with $p_T > 0.5$ GeV, is employed. In particular, to identify jets originating from the hard-scattering PV, the requirement JVT > 0.59 is applied to low p_T jets ($p_T < 60$ GeV) in the central region $|\eta| < 2.4$ of the detector [227], since the contribution of pile-up jets at high p_T is negligible.

Furthermore, the selected small-*R* jets are reclustered using the anti-k_t algorithm with $R = 1.0$, resulting in the fully calibrated RC (large-*R*) jets (Sec. 5.2.5). Such large-*R* jets are employed to describe events in the boosted topology in order to identify the *boosted Higgs boson candidates* and possibly also the *boosted hadronic top-quark candidates*. Additionally, RC jets ensure orthogonality between the resolved and boosted lepton+jets channel. These jets are required to have a reconstructed invariant mass higher than 50 GeV, $p_T > 200$ GeV and at least two small-*R* constituent jets. Such RC jets are used as input to a deep neural network (DNN) to identify high-p_T (boosted) top-quark and Higgs boson candidates decaying into collimated hadronic final states (explained in Sec. 6.5.2).

Since the final state of the $t\bar{t}H(H \to b\bar{b})$ process contains a large number of *b* quarks, *b*-tagging and *b*-jets identification play a determining role in this analysis. Jets are identified

as originating from the hadronisation of a *b*-quark using the high-level *b*-tagging algorithm MV2c10 (described in Sec. 5.3.1), which combines information from the impact parameters of displaced tracks as well as topological properties of secondary and tertiary decay vertices reconstructed within the jet. A selection requirement on the MV2c10 discriminant at a single-cut WP with the desired *b*-jet efficiency is applied, corresponding to a specific PCB score (cf. Sec. 5.3.1). The correction factors to data are retrieved for the *b*-jet efficiency, *c*-jet and *light*-jet mistag rates separately (see Sec. 5.3.2). The analysis considers the 77% and 85% *b*-tagging efficiency WPs for the boosted region and 60% and 70% WPs for resolved regions. This allows to define different analysis regions, with different amount of signal and background events.

In the single-lepton channel, events with at least four small-*R* jets, at least three of which being *b*-tagged at the 85% efficiency WP are selected and classified in the *boosted region*. In addition, an RC jet with $p_T \geq 300$ GeV and invariant mass in the range 100–140 GeV, containing exactly two constituent jets *b*-tagged at the 85% WP, is required. This RC jet is, then, flagged as a boosted Higgs-boson candidate if it has a probability $P(H) \geq 0.6$ of originating from a Higgs boson, as estimated by the DNN (explained in Sec. 6.5.2). Moreover, events with at least five small-R jets, at least three of which being *b*-tagged using the 70% efficiency WP and which do not fall in the boosted category, are selected and categorised in the *resolved region*. Lastly, in the dilepton channel, events are required to have at least three small-R jets, at least three of which must be *b*-tagged using the 70% efficiency WP. The exact requirements on the number of jets and *b*-tagged jets implemented to define the analysis regions in the different channels are described in the next section.

Hadronically decaying τ leptons (τ_{had}) (Sec. 5.4.3) are distinguished from jets using their track multiplicity and a multivariate discriminant based on the track collimation, further jet substructure and kinematic information. These τ_{had} candidates are required to have $p_T > 25$ GeV, $|\eta| < 2.5$ and to pass the *Medium* τ-identification working point. To maintain orthogonality with other $t\bar{t}H$ channels, such as the multi-lepton channel [90], events are vetoed if they contain two or more (one or more) τ_{had} candidates in the single-lepton (dilepton) channel.

As discussed in Chapter 5, the energy deposits in the calorimeter are used to reconstruct electrons and jets, consequently a single detector response can be assigned to more than one lepton or jet. To prevent the double-counting of objects, an *overlap removal* procedure is adopted and applied to the reconstructed leptons and small-*R* jets (before the reclustering). First, the closest jet within $\Delta R = \sqrt{(\Delta y)^2 + (\Delta\phi)^2} = 0.2$ of a selected electron is removed, reducing the number of jets reconstructed from electron energy depositions in the calorimeter. Then, if the nearest jet surviving that selection is found within $\Delta R = 0.4$ of the electron or muon, the lepton is discarded. This reduces the background arising from non-prompt electrons or muons from heavy flavour decays inside the jets. However, for muons, if this jet has fewer than three associated tracks, the jet is removed instead. Such a jet often comes from energy depositions of a high p_T muon, hence an inefficiency for high-energy muons undergoing significant energy loss in the calorimeter is prevented. Lastly, a τ_{had} candidate is rejected if it is separated by $\Delta R < 0.2$ from any selected electron or muon.

The missing transverse momentum (with magnitude E_T^{miss} - defined in Sec. 5.5) is reconstructed as the negative vector sum of the p_T of all the selected electrons, muons, τ_{had} and jets described above (as well as photons), with an extra soft term to account for energy in the event which is not associated with any of these. This extra term is built from additional tracks associated with the PV, to make it resilient to pile-up contamination. The missing transverse momentum is not used for the event selection but enters the event reconstruction to describe

the neutrino from the leptonic decay of the *W*-boson. It is also included in the inputs to the multivariate discriminants that are built in the most sensitive analysis categories.

6.4 Analysis region definition

In order to target the $t\bar{t}H(H \to b\bar{b})$ final state, events are classified into mutually exclusive regions among the different channels of the analysis (the single-lepton resolved, the single-lepton boosted, and the dilepton channel). After the event selection described above, most of the SM processes are rejected, leaving out mainly background from $t\bar{t}$+jets events. Especially, the actual contribution of the different $t\bar{t}$+jets categories depends on the specific selection in each analysis region. As discussed in Sec. 4.2.3 and according to the PDF distributions in Fig. 4.3, the production of *c*- and *b*-quarks in a parton shower is suppressed due to their mass. Providing that *light*-jets include *u*, *d*, *s* or gluon induced jets, the $t\bar{t}$+light component dominates, unless a strict selection on the number of *b*-jets is introduced. As a result, at this stage the samples are dominated by the $t\bar{t}$+light component in the single-lepton boosted channel, while the irreducible $t\bar{t}+\geq 1b$ prevails in single-lepton and dilepton resolved channels.

In addition to the dominant $t\bar{t}$+jets background, there are small contributions from the associated production of a vector boson and a $t\bar{t}$ pair ($t\bar{t}+V$; $V = W, Z$) as well as non-$t\bar{t}$ events. The latter originate from the production of a single top, followed by the production of a *W*- or *Z*-boson in association with jets (*W/Z*+jets), the diboson production (*WW*, *WZ*, *ZZ*), as well as other rare top-quark sources. In the following, backgrounds from non-$t\bar{t}$ processes are grouped together in the figures as "Other" (depicted in yellow), unless stated otherwise, since their single contributions are very small. Only the *tH* process is represented separately in pink, which also has a minimal contribution. All the aforementioned background processes and their modelling are explained thoroughly in Sec. 4.5.

The non-overlapping analysis regions are primarily defined by the number of leptons, in order to separate between single-lepton and dilepton channels, as already pointed out in Sec. 6.3.2. In addition, selection criteria on the number of jets as well as *b*-tagged jets at different *b*-tagging efficiencies are applied, so as to profit from the higher jet and *b*-jet multiplicities of the $t\bar{t}H$ signal process with respect to the $t\bar{t}$ background. Specifically, the different *b*-tagging efficiencies are exploited to define regions with different contributions of the $t\bar{t}$+jets background components. Additionally, a requirement on the number of boosted Higgs boson candidates is implemented in order to target the boosted regime of the $t\bar{t}H(H \to b\bar{b})$ signature.

To gain control over the various background components, events are classified into categories, which are often referred to as "regions" of the analysis phase space. Regions are interpreted in terms of relations between "*S*", the expected number of events of the SM Higgs boson signal with $m_H = 125$ GeV, and "*B*", the expected number of background events according to MC simulation. The categorisation of events is based on the *signal-to-background ratio* S/B and the approximation of the *statistical significance of signal* $S/\sqrt{B}$. Regions where the signal model predicts a significant excess of events over the predicted background level, namely with high S/B and $S/\sqrt{B}$, are referred to as *signal-enriched* or just *signal regions* (SR), and they provide most of the sensitivity to the signal. Furthermore, to estimate background processes contaminating the signal regions, analysis regions depleted in signal are defined, called *control regions* (CR).

6.4.1 Single-lepton boosted region definition

Based on the final state of the $t\bar{t}H(H \to b\bar{b})$ signal, four *b*-quarks (two from the top decays and two from the Higgs boson) are contained, two additional quarks from a decay of the hadronic *W*-boson and a single charged lepton. The six quarks in the final state produce jets, four of which can be tagged as *b*-jets. In general, individual jets are represented by small-*R* jets. This holds for the resolved topology, where the decay products are represented by small-*R* jets, while in the boosted regime (Sec. 6.2) they are collimated in large-*R* jets. In the single-lepton boosted region, the Higgs boson candidate (and possibly also the hadronic top-quark candidate) is required to be boosted. As a result, at least four small-*R* jets are expected in the final state, since the boosted candidates are represented by large-*R* (RC) jets, as illustrated in Fig. 6.3. In addition, the signal signature contains four *b*-jets (two from the top-quark decays and two from the Higgs boson). Therefore, events in the single-lepton boosted SR are required to have at least four small-R jets, as well as a specific number of *b*-tagged jets. Furthermore, among the dominant background sources, the $t\bar{t} + b\bar{b}$ production process results in the same final-state signature as the signal, containing also four small-R jets, thus the discrimination between signal and background events becomes challenging.

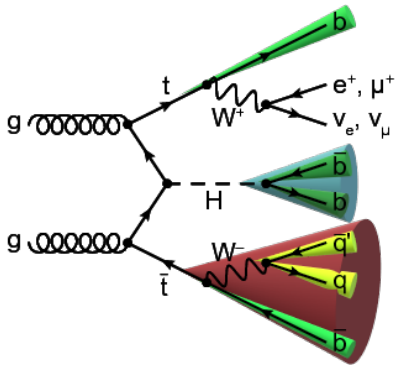
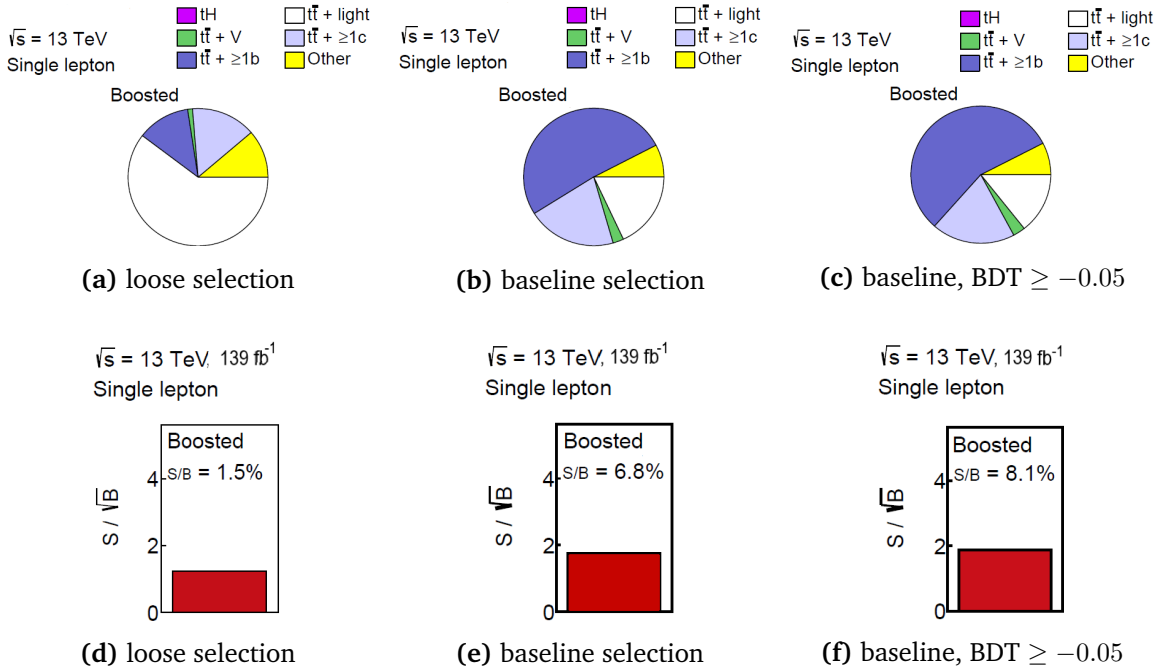


Fig. 6.3: Tree-level Feynman diagram summarising the boosted topology in the single-lepton channel of the $t\bar{t}H(H \to b\bar{b})$ analysis. The Higgs-boson (blue cone) and the hadronic top-quark (red cone) candidates are represented by large-*R* jets with their small-*R* constituent jets (green and yellow cones). Also, the leptonic top-quark jet is also depicted by a small-*R* jet (green cone).

As described in Sec. 6.3.2, a loose cut, requiring at least three *b*-tagged jets at the 85% efficiency WP, was defined in the first place. These loose requirements, instead of requiring more *b*-tagged jets or a tighter *b*-tagging WP, are to account for the limited efficiency of the *b*-tagging algorithm to identify real *b*-jets. According to this event selection, the single-lepton boosted region is dominated by the $t\bar{t}$+light background component, as expected. The background composition of the boosted region is depicted in Fig. 6.4a. Besides, this selection results in low purity ($S/B = 1.5\%$) and low statistical significance of the signal ($\sim 1.2\sigma$) (Fig. 6.4d).

In order to further constrain the $t\bar{t}$+jets background, tighter *b*-tagging requirements are employed to define the boosted SR. In particular, exactly two small-*R* *b*-tagged jets at the 85% efficiency WP are assigned to the boosted Higgs-boson candidate constituents. Then, if more than one boosted Higgs-boson candidates are identified, the one with the invariant mass closest to the Higgs-boson is selected. Additionally, at least two small-*R* jets being *b*-tagged at the 77% WP, which do not belong to the boosted Higgs-boson candidate, are required. As depicted

in Fig. 6.4b, this selection constrains drastically the $t\bar{t}$+light background making the $t\bar{t}+\geq 1b$ component dominant. Nonetheless, the $t\bar{t}$+light background still has a sizeable contribution, however applying even tighter b-tagging requirements would reduce the statistical significance of the region. This selection also confines a little the non-$t\bar{t}$ background contributions. Simultaneously, it results in higher purity ($S/B = 6.8\%$) and a statistical significance of $\sim 1.7\sigma$, as shown Fig. 6.4e. Eventually, this tighter selection is chosen as the "baseline" event selection to define the single-lepton boosted signal region. On the contrary, the looser boosted selection is used for the BDT training in the single-lepton boosted region (detailed in Sec. 6.5.3), since it provides higher statistics.



(a) loose selection **(b)** baseline selection **(c)** baseline, BDT ≥ -0.05

(d) loose selection **(e)** baseline selection **(f)** baseline, BDT ≥ -0.05

Fig. 6.4: Comparison of the expected signal and background composition in the single-lepton boosted channel for the different selections. Figures (a)-(c) show the fractional contributions to the total background and (d)-(f) the ratios S/B and $S/\sqrt{B}$

Moreover, only the single-lepton boosted region with a BDT score (see Sec. 6.5.3) of at least -0.05 is included in the fit for the statistical analysis (discussed in Ch. 8), to further reject $t\bar{t}$+jets background while losing almost no signal. This is motivated by the classification BDT output distribution for signal and background, depicted e.g. in Fig. 6.19a. In this region, mostly the $t\bar{t}$+light background component is further constrained, as shown in Fig. 6.4c. This increases even more the purity to 8.1% and the statistical significance to almost 2σ (Fig. 6.4f). So in the following, this region is referred to as the single-lepton boosted signal region, $SR_{boosted}$. In addition, the $SR_{boosted}$ is further split into two reconstructed p_T^H regions: $300-450$ GeV and ≥ 450 GeV, to allow for the extraction of the differential cross-section measurement, sensitive to new physics effects. These p_T^H ranges are the same as used to define STXS bins with true $\hat{p}_T^H$ (cf. Sec. 6.1), and are chosen to minimise the correlation among signal strengths in different STXS bins. Finally, in both single-lepton boosted SRs the shape and normalisation of the classification BDT output distribution, is used as the final discriminant for the fit (see Ch. 8). The full selection criteria of the $SR_{boosted}$ are summarised in Table 6.3.

The signal event yields are calculated using the "inclusive" $t\bar{t}H$ MC sample (see Sec. 4.5.2),

Region	$SR_{boosted}$	
reco p_T^H [GeV]	[300,450)	[450,∞)
#leptons	$== 1$	
#small-*R* jets	≥ 4	
#boosted Higgs-boson candidates	≥ 1	
RC jet reco p_T [GeV]	≥ 300	
RC jet mass [GeV]	[100-140)	
#small-R subjets *b*-tagged at 85%	$== 2$	
additional #*b*-tags 77%	$\geq 2^{\dagger}$	
Fit input	classification BDT ≥ -0.05	

Table 6.3: Definition of the single-lepton boosted analysis region ($SR_{boosted}$). The *b*-tagged jets flagged with † are extra *b*-jets not part of the boosted Higgs-boson candidate. The $SR_{boosted}$ is further split in reconstructed p_T^H similarly as the STXS binning. The last row specifies the type of input to the signal extraction fit.

in which all Higgs-boson decay modes predicted by the SM (Sec. 2.4.2) are considered. Nevertheless, the analysis is optimised specifically for the $H \to b\bar{b}$ decay. Therefore, after the event selection and region definition, only small fractions of the other decay modes are still present. Especially in the single-lepton boosted signal region, the $H \to b\bar{b}$ events account for 96% of the $t\bar{t}H$ selected events, while the $H \to WW$ and all the other decay modes correspond to the 1.4% and 2.6%, respectively.

Reconstruction of the top-quark candidates

The main experimental challenge for the semi-leptonic $t\bar{t}H$ channel is the low efficiency to reconstruct and identify all final-state particles. In particular, the assignment of the many jets in the final state to their original particles becomes a combinatorial problem, complicating the reconstruction of the Higgs boson and top-quark candidates. Moreover, the large SM background from the production of $t\bar{t}$+jets processes, which have a much larger production cross-section than the $t\bar{t}H$ signal, especially when the associated jets stem from *b*- or *c*-quarks, enhances the combinatorial ambiguity. However, a proper reconstruction of the Higgs boson and top-quark candidates would increase the separation power of their kinematic variables, that would in turn benefit the classification BDT (see Sec. 6.5.3) to better classify events between signal and background.

Therefore, a full event reconstruction, based on various kinematic requirements, is performed in the single-lepton boosted region. This aims to correctly assign the reconstructed jets to the final state partons from top-quark and Higgs boson decays, and to suppress background from wrong combinations. For this reason, *W*-boson, top-quark and Higgs-boson candidates are built from reconstructed jets, missing transverse energy and a lepton. Jets are assigned to the quarks from the $t\bar{t}H(H \to b\bar{b})$ decay and combinations including jets and *b*-jets are used to reconstruct the objects, not only of the Higgs boson that has already been discussed, but also of the top quark.

In the following, the top-quark candidate containing the hadronically (leptonically) decaying W-boson (see Sec. 2.5.2) is referred to as the *hadronic* (*leptonic*) *top-quark candidate*. Specifically the hadronic top-quark candidate is reconstructed either in a boosted way, considering its decay products as part of a large-R jet, or as resolved one using small-R jets. The leptonic top-quark candidate is reconstructed exploiting a small-R jet, or no jet at all. While the boosted Higgs-boson candidate is used as a requirement in the event selection (described in Sec. 6.3), the top-quark candidates are only reconstructed in order to use their properties as input for the classification BDT.

The leptonically decaying *W*-boson candidate is needed so as to reconstruct the leptonic top-quark candidate and it is obtained from the lepton and the neutrino four-momenta p_l and p_ν, respectively. The latter is built from the missing transverse momentum (E_T^{miss}), but its longitudinal component, $p_{\nu,z}$ (the z-direction is defined along the beam pipe), is not measurable. However, it can be inferred by assuming that the lepton and the neutrino (i.e. the E_T^{miss}) are originating solely from the *W*-boson decay, hence by solving the equation $m_W^2 = (p_l + p_\nu)^2$, where m_W represents the nominal *W*-boson mass. If there are two solutions of this quadratic equation, both are considered. Then, the invariant mass of the small-R jet (considered for the leptonic top-quark candidate - described below), the neutrino, and the lepton is computed for both and the one that gives the closest mass to the top-quark mass is chosen. If no real solutions exist, the discriminant of the quadratic equation is set to zero, giving a unique solution.

After the boosted Higgs-boson candidate has been found, additional RC jets, which are required to have $p_T \geq 300$ GeV and a DNN probability $P(t) \geq 0.3$ of originating from a real top quark, are identified as boosted top-quark candidates (defined in Sec. 6.5.2). These RC jets are taken to be different from the one assigned to the boosted Higgs-boson candidate. In case more than one boosted top-quark candidates are identified, the one with the invariant mass closest to the top-quark mass is selected. Then, the leptonic top-quark candidate reconstruction follows, after both the Higgs-boson and the hadronic top-quark candidates have been identified. In particular, a small-*R* jet is searched, requiring the invariant mass of this jet, the lepton and the neutrino to be in the range of $[130, 200)$ GeV. This small-*R* jet should not be a constituent of the two RC jets assigned to the Higgs-boson and hadronic top-quark candidates. If more than one small-*R* jet fulfilling these requirements is found, the one that together with the neutrino and the lepton give an invariant mass closest to the top quark, is taken. On the contrary, if no such non-overlapping jet is found, then the leptonic top-quark is defined as the sum of lepton and neutrino, i.e. the W boson alone is used instead of the top quark. In the latter case, the PCB (pseudo-continuous b-tagging) score for the small-*R* jet from the leptonic top-quark candidate is set to zero.

If the hadronic top-quark candidate has not been found within the large-R jets, small-*R* jets not overlapping with the Higgs-boson candidate are taken into account. In this case, the hadronic top-quark candidate is not considered to be in the boosted regime. The invariant mass of the reconstructed hadronic (m_{hadTop}) and leptonic (m_{lepTop}) top-quark are evaluated for all combinations of a certain number of small-*R* jets simultaneously. Thus, the invariant mass of the jets assigned to the hadronic top-quark candidate has to be in the range of [70,195] GeV, while the invariant mass of the jet, lepton and neutrino $m_{lepTop} \in [130, 200]$ GeV. If there is at least one combination, the one with minimum value of $|m_{hadTop} - 172.5| + |m_{lepTop} - 172.5|$ is chosen. In case there is no non-overlapping combination, the hadronic top-quark is reconstructed from the three highest p_T jets (not overlapping with the Higgs-boson candidate), while the leptonic top-quark is reconstructed from lepton and neutrino. Also in this case, the

PCB score for the small-*R* jet from the leptonic top-quark candidate is set to 0.

Finally, the kinematic variables of the Higgs boson, hadronic and leptonic top-quark candidates, such as the transverse momentum p_T^H of the Higgs-boson candidate that is used to split the $SR_{boosted}$, are computed according to the above reconstruction. Also, other variables like angular separations between the candidates and b-tagging discriminants of the candidates or of combinations among them are computed. Then, some of them are used as inputs to the classification BDT (Sec. 6.5.3), that is then employed to separate signal from background, in the two boosted SRs.

Performance of the reconstruction in the single-lepton boosted region

Summarising from the above, the Higgs-boson candidate is always reconstructed with a RC jet, while the reconstruction of the top-quark candidates is more complicated. In order to better understand the performance of this complex reconstruction procedure, Table 6.4 shows how often the various reconstruction cases happen in the inclusive $t\bar{t}H$ signal and $t\bar{t}$+jets background samples, in the boosted SR. The leptonic top-quark candidate is mostly reconstructed with a small-*R* jet, while the cases in which no jet is assigned are more rare. Then, the hadronic top-quark candidate is reconstructed either with RC or small-*R* jets, with the latter being the dominant one.

	$t\bar{t}H$			$t\bar{t}$		
Reco objects	RC jets	small-R jets	no jet	RC jets	small-R jets	no jet
hadronic top-quark	20.4%	79.6%	0%	13.3%	86.7%	0%
leptonic top-quark	0%	90.4%	9.6%	0%	90.3%	9.7%

Table 6.4: Probabilities of the different reconstruction procedures for each reconstructed top-quark candidate, in the inclusive $t\bar{t}H$ and $t\bar{t}$ samples. Events falling in the boosted SR are considered in these fractions.

Figure 6.5 shows the migration matrix for the Higgs boson candidate between the true $\hat{p}_T^H$ and the reconstructed p_T^H in the $SR_{boosted}$, considering $t\bar{t}H$ signal events with $H \to b\bar{b}$ and semi-leptonic $t\bar{t}$ decays only. In fact, the p_T of a true particle is smeared at detector level, due to the detector response, thus it can take any possible value after the reconstruction. The specific reconstruction in the boosted region seems to sufficiently recover the correspondence between the true and the reconstructed p_T^H, since the majority of the signal events lie in the diagonal of the matrix.

In order to further assess the performance of the reconstruction procedure, the true-reco matching probabilities for the reconstructed objects, considered in the single-lepton boosted region (Higgs, hadronic top and leptonic top), are shown in Tables 6.5 and 6.6 for the signal and dominant background samples. The reconstructed small-R jets, or the RC jets contain subjets that, are geometrically matched to the true jets under the angular distance requirement $\Delta R < 0.3$, as already introduced in Sec. 5.2.1. Thus, the following generator-level ("true") objects have been defined for this study. *True Higgs-boson* means a matching of two *b*-quarks from the Higgs-boson decay. *True hadronic top-quark* means a matching of the *b*-quark from the hadronic top-quark and two *b*-quarks from the *W*-boson decays. Also, *true semi-hadronic top-quark* means a matching of the *b*-quark from the hadronic top-quark and a light-quark from

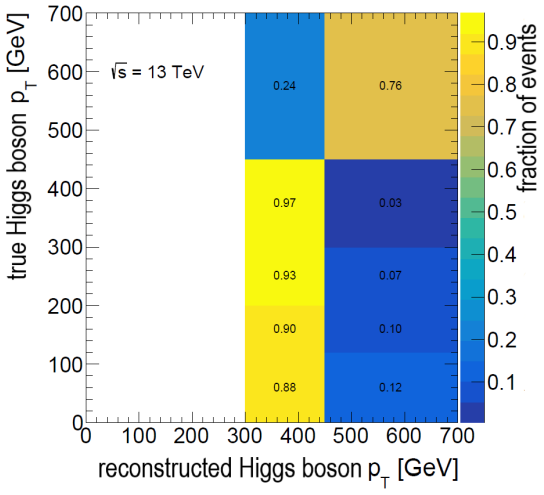


Fig. 6.5: True $\hat{p}_T^H$ vs. reconstructed p_T^H migration matrix showing the fraction of $t\bar{t}H$ signal events in each bin, considering only $H \to b\bar{b}$ and semi-leptonic $t\bar{t}$ decays, in the $SR_{boosted}$.

the *W*-boson, or two light-quarks from the *W*-boson. Lastly, *true leptonic top-quark* means a matching of the *b*-quark from the leptonic top-quark. In this way, one can verify which of the true partons that fall within the large-*R* jets, or just assigned to small-*R* jets (for the leptonic top-quark or possibly also for the hadronic top-quark reconstruction), according to the boosted reconstruction, actually originate from a top quark or a Higgs boson.

Samples	$t\bar{t}H$		
True-reco objects	inclusive	$\hat{p}_T^H \in [300, 450)$ GeV	$\hat{p}_T^H \in [450, \infty)$ GeV
Higgs boson	71%	90%	93%
hadronic top-quark	17%	19%	24%
semi-hadronic top-quark	20%	25%	24%
leptonic top-quark	58%	62%	65%

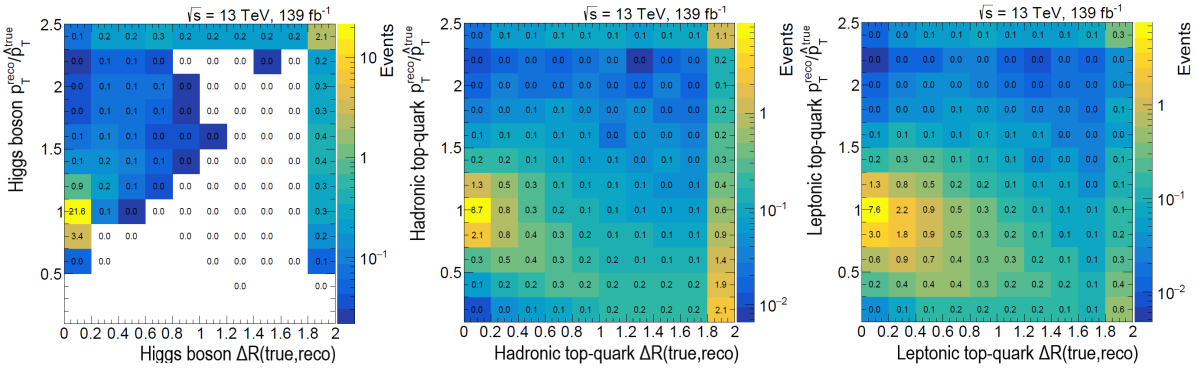
Table 6.5: True-reco matching probabilities for each reconstructed object candidate considered in this analysis, in the $t\bar{t}H$ sample, for the inclusive case and the $\hat{p}_T^H$ splitting required by the analysis strategy. Only events falling in the boosted SR are considered in these fractions.

Samples	$t\bar{t}$	$t\bar{t} + \geq 1c$	$t\bar{t} + \geq 1b$	$t\bar{t}$ + light
True-reco objects				
hadronic top-quark	8%	6%	9%	5%
semi-hadronic top-quark	11%	10%	12%	6%
leptonic top-quark	52%	51%	52%	53%

Table 6.6: True-reco matching probabilities for reconstructed top-quark candidates considered in this analysis, in the inclusive $t\bar{t}$+jets sample and in its heavy-flavour sub-components. Only events falling in the boosted SR are considered in these fractions.

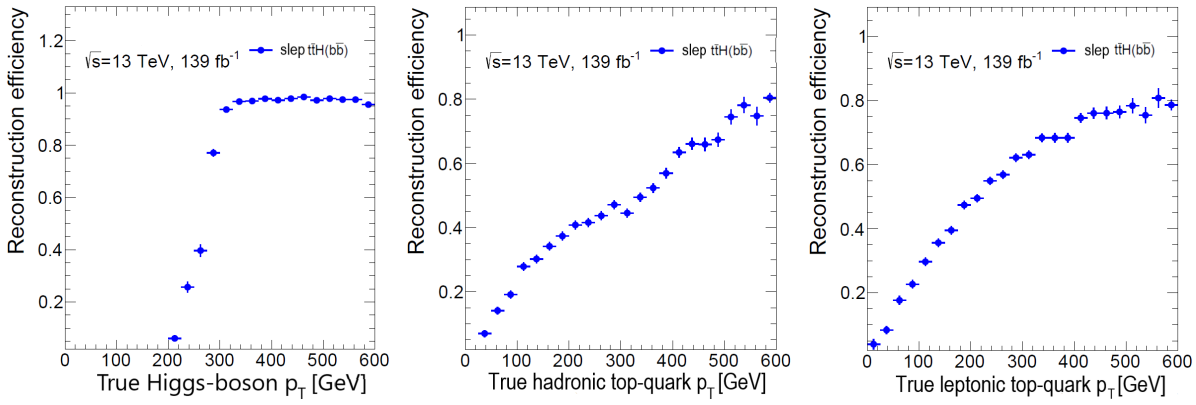
When applying the reconstruction described earlier, the Higgs-boson candidate is found to be well reconstructed in 91% of the cases in the phase space that the boosted region targets, with true $\hat{p}_T^H \geq 300$ GeV (combining the last two columns in Table 6.5). This is an unprecedented purity for this analysis. Besides, in this phase space, the hadronic top-quark candidate (considering also the semi-hadronic top-quark candidate) is well reconstructed in 47% of the cases, while the leptonic top-quark candidate in 63%. As expected, the true-reco matching probabilities of the reconstructed objects are much larger in the high-$\hat{p}_T^H$ signal samples than in the inclusive one, since the boosted region is optimised for this specific regime. Then, no true Higgs-boson is expected in the $t\bar{t}$+jets background (Table 6.6), while the reconstruction purity for the top-quark candidates is smaller than in the signal. This is not surprising, given the significant probability that quarks from the top-quark decay are used to reconstruct the Higgs-boson candidate.

Furthermore, the efficiency of the aforementioned reconstruction procedure is examined, as another figure of merit. Figures 6.6a and 6.6b (6.6c) illustrate the region where most of the boosted signal events fall, in terms of $p_T^{reco}/\hat{p}_T^{true}$ and ΔR(true,reco) of the Higgs-boson and hadronic (leptonic) top-quark candidates, respectively, especially in the semi-leptonic $t\bar{t}H(H \to b\bar{b})$ sample. From these matrices it is apparent that most of the boosted signal events fall in a region characterised by $|p_T^{reco}/\hat{p}_T^{true} - 1| < 0.5$ and ΔR(true,reco) < 0.4 for all candidates. Particularly for the Higgs-boson candidate, the reconstructed and true p_T match very well in most of the signal events, while also the angular separation of the reconstructed with respect to the true Higgs-boson is the smallest possible. Then, Figs. 6.7a and 6.7b (6.7c) show the reconstruction efficiency of the Higgs-boson and hadronic (leptonic) top-quark candidates as a function of their respective true $\hat{p}_T$ in this particular region, where most of the boosted signal events are found. Obviously, the Higgs-boson candidate reconstruction efficiency rises rapidly with the increasing true Higgs-boson p_T and already at 300 GeV reaches an efficiency of $\sim 95\%$, which is then kept almost constant at 95%-100% for the rest of the p_T spectrum. This shows that the Higgs-boson candidate is correctly reconstructed most of the times in the relevant p_T range. Then, for the top-quark candidates, the reconstruction efficiency gradually rises with the increasing true p_T of the considered object, and an efficiency of almost 80% is reached for both in the very high-p_T region.



(a) **(b)** **(c)**

Fig. 6.6: Number of signal events in terms of $p_T^{reco}/\hat{p}_T^{true}$ and $\Delta R(\text{true}, \text{reco})$ for (a) Higgs-boson (b) hadronic top-quark and (c) leptonic top candidates, in the semi-leptonic $t\bar{t}H(H \to b\bar{b})$ sample. In both cases, most of the events fall within the $|p_T^{reco}/\hat{p}_T^{true} - 1| < 0.5$ and $\Delta R(\text{true}, \text{reco}) < 0.4$ region.



(a) **(b)** **(c)**

Fig. 6.7: Reconstruction efficiency as a function of the true $\hat{p}_T$ of the candidate in the region characterised by $|p_T^{reco}/\hat{p}_T^{true} - 1| < 0.5$ and $\Delta R(\text{true}, \text{reco}) < 0.4$ for (a) Higgs-boson (b) hadronic top and (c) leptonic top-quark candidates, in the semi-leptonic $t\bar{t}H(H \to b\bar{b})$ sample.

According to the observed reconstruction efficiencies and purity, there are still cases in which the reconstructed objects, mostly the hadronic top but also the leptonic top-quark candidate and rarely the Higgs-boson candidate, are not reconstructed properly. Because of mis-reconstructions, the different candidates may be constructed by wrong combinations of jets, which do not originate from the respective true objects. Firstly, due to the limited efficiency of the *b*-tagging algorithm a real *b*-jet cannot be always identified. In fact, according to the *b*-tagging requirements of the boosted selection, a real *b*-jet is identified only in 85% of the times for each *b*-jet of the Higgs-boson candidate, and in 77% of the times for each additional *b*-jet. Then, each *b*-tagging working point has a corresponding mistag rate for *c*- and light-jets (see Table 5.1), respectively. In addition, among the hadronic *W*-boson decays $W \to c\bar{s}$ is one of the two most probable ones, with a very high decay rate (it is proportional to $|V_{cs}|^2$ - see Table 2.6). It is quite probable then, that the *c*-quark is mis-tagged as a *b*-jet, which can subsequently be mis-reconstructed as part of one of the candidates. Another possibility is that none of the aforementioned partons are contained in the reconstructed jets. For instance, the Higgs-boson candidate may have been reconstructed with *b*-quarks originating from a gluon radiation and not from a true Higgs boson. Such *b*-quarks may also constitute any of the top-quark candidates. Especially, the low hadronic top-quark reconstruction efficiency could also result from the fact that jets originating from quarks from the *W* boson, are possibly not detected due to their low p_T. It is also hard to reconstruct them, since the jets in the analysis are required to have $p_T > 25$ GeV (see Sec. 6.3.2). Finally, there are cases in which the leptonic top-quark candidate is reconstructed only from a lepton and a neutrino, thus no true jet can be matched with it.

6.4.2 Summary of the $t\bar{t}H(H \to b\bar{b})$ analysis regions

Considering that all the analysis regions are combined for the signal extraction fit, also the single-lepton and dilepton resolved region definitions are briefly described here for completeness. As already introduced, orthogonality between the regions is ensured by the number of leptons for the dilepton and single-lepton regions, and by the number of boosted Higgs candidates reconstructed for the single-lepton boosted and resolved regions. Additional re-

quirements on the number of jets and *b*-tagged jets, using the 60% or 70% efficiency WPs, are employed to define exclusive regions in each resolved channel.

On condition that the $t\bar{t}$ + HF-jets backgrounds (defined in Sec. 4.5.3) are particularly hard to model, apart from the signal regions also control regions are defined. Control regions are specifically designed to have a high purity of one type of background, then the dominant backgrounds can be controlled by comparison to the data samples. They provide stringent constraints on the normalisation and shape of the backgrounds and on systematic uncertainties, thus improving the background prediction in the signal-rich regions. As it is demonstrated in Ch. 8, the resolved regions have high statistics and are dominated by systematic uncertainties, while the single-lepton boosted region is mostly dominated by the statistical uncertainty due to low statistics. Therefore, control regions are only defined in the resolved regions, where they are mostly needed. Finally, the control regions are combined statistically with the signal regions in the final fit, detailed in Ch. 8, to constrain the uncertainties on these backgrounds and maximise the overall sensitivity.

Similarly to the boosted region, also these analysis regions are further split in reconstructed p_T^H bins, according to the STXS true $\hat{p}_T^H$ bins, to allow for the extraction of multiple signal parameters. However, the control regions, that are defined in the resolved channels, are kept inclusive in reconstructed p_T^H to keep increased statistics, so that they maintain the constraints on the background composition.

Events in the single-lepton resolved channel are classified in the control or signal regions according to whether the number of jets is exactly five or at least six, requiring also at least four of them to be *b*-tagged using the 70% efficiency WP. Apparently, tighter b-tagging requirements are used with respect to the boosted region, since the resolved region has much higher statistics. This results in smaller contributions of the $t\bar{t}+\geq 1c$ and $t\bar{t}$+light background components, as well as of the non-$t\bar{t}$ background processes. Furthermore, events in the control region are further subdivided, having at least four or less than four *b*-tagged jets using the 60% efficiency WP, denoted by $CR^{5j}_{\geq 4b\,\text{hi}}$ and $CR^{5j}_{\geq 4b\,\text{lo}}$, respectively. Finally, events in the signal region ($SR^{\geq 6j}_{\geq 4b}$) are further split in five reconstructed p_T^H bins: $0-120$ GeV, $120-200$ GeV, $200-300$ GeV, $300-450$ GeV, and ≥ 450 GeV.

The dilepton analysis regions are defined in an analogous way. Specifically, events in the signal region ($SR^{\geq 4j}_{\geq 4b}$) are required to have at least four jets, at least four of which are *b*-tagged at the 70% efficiency WP. In addition, these events are split in reconstructed p_T^H bins, but with the two highest p_T^H bins merged together because only a small number of events are expected there. So the four p_T^H bins are $0-120$ GeV, $120-200$ GeV, $200-300$ GeV, and ≥ 300 GeV. Then, events in the $CR^{\geq 4j}_{3b\,\text{hi}}$ ($CR^{3j}_{3b\,\text{hi}}$) control region are required to have have at least four (exactly three) jets, exactly three of which are *b*-tagged using the 60% efficiency WP. Lastly, events in the $CR^{\geq 4j}_{3b\,\text{lo}}$ control region are required to have have at least four jets, exactly three of which are *b*-tagged using the 70% efficiency WP, but with less than three of them being *b*-tagged using the 60% efficiency WP. As already introduced, the boosted topology is not studied in the dilepton channel due to a much smaller expected yield in the high-p_T regime.

Tables 6.7 and 6.8 summarise the definitions of the regions into which the selected events are classified in the single-lepton and dilepton resolved channels, respectively. Also in the resolved categories, the Higgs boson candidates are reconstructed using boosted decision trees (BDT) referred to as "reconstruction BDTs", aiming at associating the reconstructed jets to the final state partons. Kinematic variables of these Higgs boson candidates as well as angular separations and *b*-tagging variables are computed according to the respective reconstruction

Region		$SR^{\geq 6j}_{\geq 4b}$					$CR^{5j}_{\geq 4b\,\text{hi}}$	$CR^{5j}_{\geq 4b\,\text{lo}}$
reco p_T^H [GeV]		[0,120)	[120,200)	[200,300)	[300,450)	[450,∞)	inclusive	
#leptons		$== 1$					$== 1$	
#small-*R* jets		≥ 6					$== 5$	
#*b*-tags	70%	≥ 4					≥ 4	
	60%	-					≥ 4	< 4

Table 6.7: Definition of the single-lepton resolved analysis regions, split according to the number of jets, and b-tagged jets using different working points. The $SR^{\geq 6j}_{\geq 4b}$ is further split in reconstructed p_T^H similarly as the STXS binning.

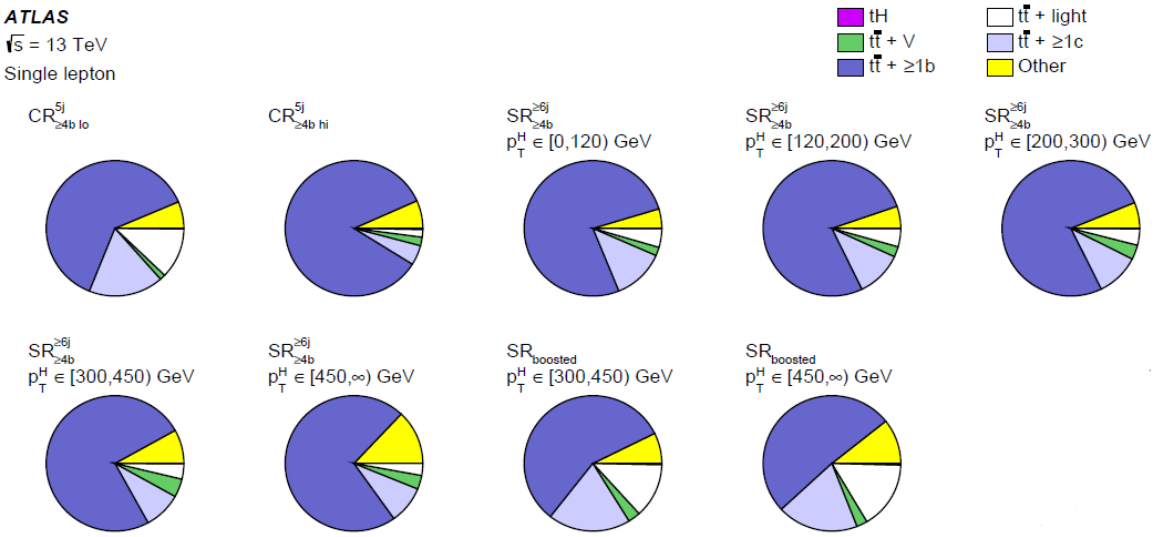
Region		$SR^{\geq 4j}_{\geq 4b}$				$CR^{\geq 4j}_{3b\,\text{hi}}$	$CR^{\geq 4j}_{3b\,\text{lo}}$	$CR^{3j}_{3b\,\text{hi}}$
reco p_T^H [GeV]		[0,120)	[120,200)	[200,300)	[300,∞)	inclusive		
#leptons		$== 2$				$== 2$		
#small-*R* jets		≥ 4				≥ 4		$== 3$
#*b*-tags	70%	≥ 4				$== 3$		
	60%	-				$== 3$	< 3	$== 3$

Table 6.8: Definition of the dilepton resolved analysis regions, split according to the number of jets, and b-tagged jets using different working points. The $SR^{\geq 4j}_{\geq 4b}$ is further split in reconstructed p_T^H similarly as the STXS binning.

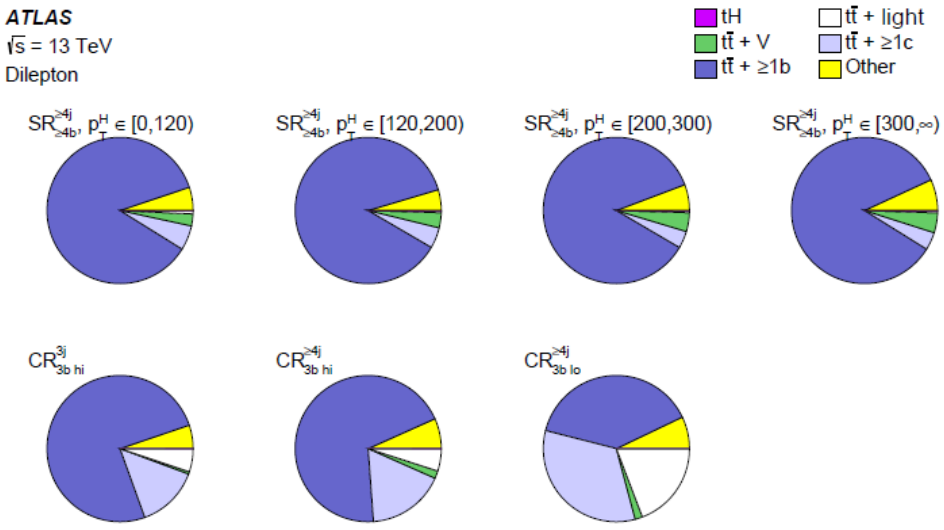
procedure. They are then used as inputs to the "classification BDTs", that are then employed to separate signal from background, in each of the SRs.

In summary, the selected events in the analysis are classified into 16 regions: eleven SRs (dilepton $SR^{\geq 4j}_{\geq 4b}$, single-lepton $SR^{\geq 6j}_{\geq 4b}$ and $SR_{boosted}$, split according to the reconstructed p_T^H into four, five and two regions, respectively), and five CRs. Figure 6.8 shows the expected background composition, in the different analysis regions. The $t\bar{t}$+jets production dominates by far in all the regions and only smaller fractions are coming from $t\bar{t}+V$ or from non-$t\bar{t}$ processes (labelled as "Other"). In addition, different fractions of the $t\bar{t}$+jets components among the regions with different *b*-tagging requirements can be observed. The largest fraction of $t\bar{t}$+jets events consists of the $t\bar{t}+\geq 1b$ background in all SRs which is of particular importance, since it can have an identical final state to the $t\bar{t}H(H \to b\bar{b})$ signal. The $t\bar{t}+\geq 1c$ process follows and then comes the $t\bar{t}$+light. The latter is really negligible in the dilepton SR, which is relatively pure in $t\bar{t}+\geq 1b$ background, since the $t\bar{t}$ systems decays only leptonically. However, it is difficult to define a very pure $t\bar{t}$ + HF-jets region in the single-lepton channel because of the *c*-quark from the hadronic *W*-boson decays (and more generally because of the extra jets). Within the single-lepton channel, the $SR_{boosted}$ has larger $t\bar{t}$+light contribution compared to the $SR^{\geq 6j}_{\geq 4b}$, given the looser *b*-tagging requirements applied in the boosted region. The size of the $t\bar{t}$+light contribution also depends on the mis-identification rate of the *b*-tagging algorithm for the corresponding efficiency working points (see Table 5.1). The CRs in the resolved

channels have different ratios of $t\bar{t} + \geq 1b$ to $t\bar{t} + \geq 1c$ or $t\bar{t}$+light events. In particular, regions labelled with "hi", referring to a higher b-tagging probability, are enriched in $t\bar{t}+\geq 1b$. Whereas, in regions labelled with "lo" the proportion of $t\bar{t} + \geq 1c$, but also $t\bar{t}$+light, events is increased. The different proportions of $t\bar{t}$+jets components in the CRs allow the signal extraction fit to better constrain the relative fractions of these processes in the signal regions.



(a) single-lepton (resolved and boosted) channels



(b) dilepton channel

Fig. 6.8: Fractional contributions to the total background in the $t\bar{t}H(H \to b\bar{b})$ analysis regions in the (a) single-lepton (resolved and boosted) and (b) dilepton channels.

Figure 6.9 shows the signal purity (S/B) and statistical significance ($S/\sqrt{B}$) in the different regions, where S and B are the number of signal and background events, respectively. Among all the SRs, the $SR_{boosted}$ achieves the largest purity of 8.1%, possibly due to exploitation of the boosted topology in the reconstruction procedure of the different objects. On the other hand, the largest significance is observed in the $SR^{\geq 6j}_{\geq 4b}$, which is expected given the much

higher statistics this region has. Finally, the CRs are not expected to have high contribution in signal by construction.

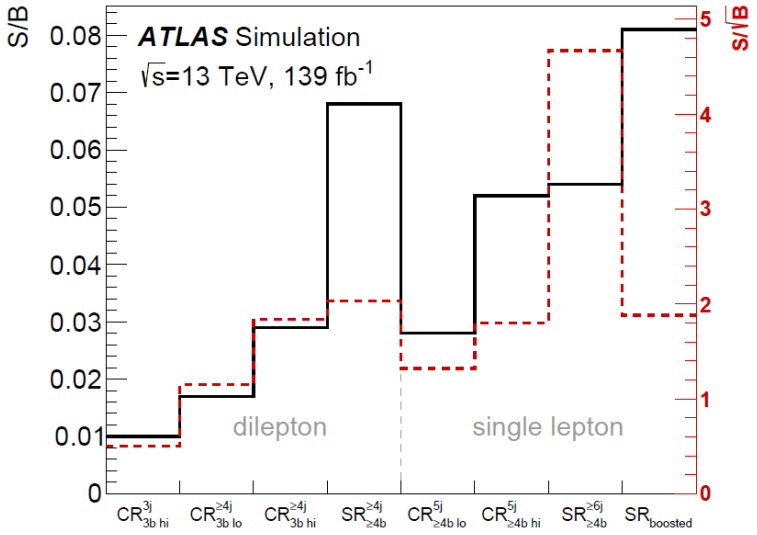


Fig. 6.9: The ratios S/B (black solid line, referring to the vertical axis on the left) and $S/\sqrt{B}$ (red dashed line, referring to the vertical axis on the right) for each category in the inclusive analysis in the dilepton channel (left) and single-lepton channels (right), where S (B) is the number of selected signal (background) events predicted by the simulation and normalised to a luminosity of 139 fb^{-1} [100].

6.4.3 Overlap between single-lepton boosted and resolved regions

As already presented, two categories are defined in the single-lepton channel, based on the resolved or the boosted topology and targeting at events with low and high p_T^H, respectively. In Sec. 6.3.2, it is mentioned that events that do not fall in the boosted category, are selected and categorised in the resolved region. In fact, this is achieved after applying a dedicated veto, because in principle it can occur that some of the selected $t\bar{t}H$ events in the resolved region also pass the boosted selection requirements. Table 6.9 summarises the number of $t\bar{t}H$ events (considering all the $t\bar{t}$ and Higgs-boson decays) that fall in the boosted signal region ($SR_{boosted}$) and those that overlap with the resolved signal ($SR^{\geq 6j}_{\geq 4b}$) and control regions ($CR^{\geq 5j}_{\geq 4b\,\text{hi}}$, $CR^{\geq 5j}_{\geq 4b\,\text{lo}}$) before applying that veto. These are provided both for the inclusive cross-section measurement and the differential cross-section measurement as a function of the true $\hat{p}_T^H$ required by the STXS formalism.

Most of the signal events that overlap between the resolved and the boosted region are in the $\hat{p}_T^H \in [300, 450)$ and $[450, \infty)$ GeV bins, where the boosted category has the highest impact in the analysis. In order to avoid double-counting these events and to preserve orthogonality to the single-lepton channel, two veto strategies have been studied. With the so-called "boosted veto" the overlap events are removed from the resolved category, giving the priority to the boosted one. While, when the "resolved veto" is applied the overlap events are removed from the boosted category, giving the priority to the resolved one. Both veto strategies have been tested in Asimov and data background-only fits, both for the inclusive and the STXS cross-section measurements (reported in App. A.3). According to the fit results, a decreased sensitivity is observed when removing the overlap events from the boosted region. This is

	Inclusive	$\hat{p}_T^H \in [0, 300)$ GeV	$\hat{p}_T^H \in [300, 450)$ GeV	$\hat{p}_T^H \in [450, \infty)$ GeV
$SR_{boosted}$	43.6	11.1	23.9	8.6
$CR^{\geq 5j}_{\geq 4b\,\text{hi}}$	2.5 (5.7%)	0.6 (5.4%)	1.4 (5.9%)	0.4 (4.7%)
$CR^{\geq 5j}_{\geq 4b\,\text{lo}}$	2.8 (6.4%)	0.7 (6.3%)	1.5 (6.3%)	0.6 (7.0%)
$SR^{\geq 6j}_{\geq 4b}$	9.0 (20.6%)	2.1 (18.9%)	5.1 (21.3%)	1.8 (20.9%)

Table 6.9: Absolute numbers of $t\bar{t}H$ signal events that fall only in the boosted category (upper row) and that fall in boosted and a given resolved region (lower rows). Also the corresponding percentage of the overlap events in each category with respect to the boosted events is provided. The $t\bar{t}H$ events are counted inclusively and split into STXS true $\hat{p}_T^H$ bins.

foreseeable, considering that the booted region has very few events compared to the resolved one and the overlap events are a considerable amount of the boosted events, as shown in Table 6.9. Also, the majority of the overlap events lie in the high-p_T region, where the boosted category is more sensitive. In light of the above, it has been decided to give priority to the boosted category over the resolved one and remove ("veto") these events from the resolved category.

6.5 Multivariate Analysis methods

The $t\bar{t}H$ production covers about 1% of the total Higgs production cross-section and, even if the branching ratio of the decay chosen for the Higgs boson ($H \to b\bar{b}$) is the highest possible (almost 60%), the expected signal is much smaller with respect to the background. Besides, the final state of the chosen channel is extremely complex, since the signal has similar final states compared to the dominating background arising from $t\bar{t} + \geq 1b$ events. As a result, the reconstruction of the Higgs boson mass peak is challenging due to the large combinatorial background arising from the presence of four *b*-jets in the final state. In addition, it is difficult to select jets initiated from *b*-quarks against the background of *c*-, *light*-quark, and gluon initiated jets. The conventional approach of applying independent cuts on individual observables is not sufficient to isolate the $t\bar{t}H(H \to b\bar{b})$ signal from the $t\bar{t}$ + jets background process. Thus, more sophisticated techniques are required in order to improve the event reconstruction and classification and to better distinguish the signal from the combinatorial background.

In recent years, it has become increasingly common in high energy physics to utilise machine learning (ML) techniques [244], with the multivariate analysis (MVA) approach having a widespread use. Particularly in this analysis, the absence of a single variable exhibiting a clear separation power among signal and background events makes the use of MVA techniques necessary, in order to enhance the discrimination between signal and background events. The MVA techniques take advantage from the different amount of discrimination between signal and background of various observables and their correlations, allowing for the automated determination of the optimal set of requirements, or the exploitation of nonlinear functions of the different observables. Eventually, various classification problems, such as the identification of the most likely origin of a *pp* event or the source of a jet, can be resolved. In terms of this analysis, two of the most popular MVA algorithms are exploited. In particular, the Deep Neural Networks (DNN) [246, 247] are used to identify the reconstructed objects with the

underlying particles. Lastly, the Boosted Decision Trees (BDT) [240] are also deployed to reconstruct physics topologies as well as to better distinguish the signal from the combinatorial background.

6.5.1 Common aspects

Several common concepts are found among the MVA algorithms used in this analysis. Such algorithms are not optimised just for a specific task but are flexible enough to adapt to different problems by tuning (training) their parameters. The *training* (or *learning*) process of these algorithms takes as input sets of events, characterised by specific features of the object to be classified (or identified), in order to define a function (*classifier*). This is used later during the classification (or identification) process, to discriminate the events between the different classes used in the training. In order to train such algorithms simulated events are exploited, especially in high energy particle physics, since a true class label may be assigned to the events according to the underlying physical processes having been generated.

In the context of this analysis, the algorithms that are employed follow the so-called *supervised learning*, which requires the set of training events to be fully associated with true labels. Consequently, the algorithms identify patterns in data based on these labels. The classifiers resulting from the training step are divided into linear and non-linear. The *linear classifiers* categorise a set of events into a discrete class by applying cuts on a variable or a linear combination of its variables. On the contrary, the *non-linear classifiers* are exploited in the cases where the classes of events are not separated well enough. In this case, a single cut on a variable depends simultaneously on all the other variable cuts, not necessarily in a linear way. BDT and DNN, that are used in the analysis, are non-linear classifiers.

The evaluation of the performance of any MVA, called *testing*, should be performed on a dataset statistically independent of that used in the training, since the MC events utilised for the training are also required to be used in the analysis itself. Therefore, the given dataset is split into two equal sized subsets, the training and testing datasets, and two separate MVAs are trained with the one half and tested with the other. Eventually, data simply uses one of the two MVAs choosing randomly at each event. In the end, the classification (or identification) process takes place, assigning objects or events to one of the possible discrete classes (e.g. signal and background) by the classifier found during the training process.

Any algorithm which requires training faces the risk of *overtraining*, whereby the performance on the data used to train the algorithm exceeds the performance of an independent dataset. In this case, the algorithm relies on features, e.g. statistical fluctuations, of the particular sample used to train the classifier rather than on general features of the kind of events to be selected. Overtraining may occur due to limited training statistics, or a non-representative sample. It can also appear when a machine learning problem has too few degrees of freedom, because too many model parameters of an algorithm are adjusted to too few data points. Overtraining leads to an ostensible increase in the classification performance over the objectively achievable one, if measured on the training sample, and to an effective performance decrease when measured on an independent test sample.

A convenient way to detect overtraining and to measure its impact is to compare the performance results of the discriminant distributions between the training and test samples. A possible inconsistency of the distributions between the training and test samples is a sign of overtraining. The sensitivity to overtraining depends on the MVA method, thus various method-specific solutions exist to counteract overtraining. Eventually in this analysis, over-

training can be mitigated by carefully choosing the hyperparameters. Hyperparameters are parameters of the MVA algorithms themselves, which affect the learning of the algorithm. They can be manually specified and optimised according to some fashion, depending on the performance of the algorithm.

6.5.2 Deep Neural Network (DNN)

Neural networks [245], coined by the loose analogy to the function of neurons in a brain, are composed of artificial neurons connected via weights to each other, forming a network. A neural network consists of an input layer, (at least) a hidden layer, and an output layer (or just a node). In case there are multiple hidden layers, the network is called Deep Neural Network (DNN). The hidden layers store information and make evaluations regarding the significance of an input to the output, and they make associations between the importance of combinations of inputs as well.

In essence, a neural network can be regarded as a mapping of a set of features (input variables) $\vec{x}$ to a classification (set of target labels) $\vec{y}$ with the function $f : \vec{x} \to \vec{y}$. For a DNN specifically, such a function defines a series of transformations, which map the input $\vec{x}$ onto hidden states h_i, until the final transformation maps these hidden states onto the output $\vec{y}$. Mathematically, these transformations are expressed as

$$\vec{h}_i = f_i(W_i\vec{h}_{i-1} + \vec{b}_i) \tag{6.2}$$

where f_i is a mathematical function used to transform the inputs called *activation function*, and a particular h_i is the i^{th} transformation of the information in $\vec{x}$, called the *embedding*. The first embedding is simply the input vector $\vec{h}_1 \equiv \vec{x}$, while the final embedding is the output of the network. The elements of the matrix W are referred to as *weights* and those of vector $\vec{b}$ as *biases*. The general structure of these transformations, such as the dimensionality of each W and the choice of activation function is referred to as the network architecture, which, together with the training parameters constitute the *hyperparameters* of the network.

A typical DNN structure is illustrated in Fig. 6.10. Each layer consists of many nodes; the input layer has one node per input feature, while the *output layer* can have as many outputs as desired (just one for the case of binary classification, or one per class for multi-class output), while each *hidden layer* in between can have an arbitrary number. The connections between each of these nodes are characterised by a set of weights and a bias. In addition, since each node in a given layer is connected to each node in the subsequent layer, this network is referred to as *fully connected*. Furthermore, each element of the input $\vec{x}$ is multiplied by the associated weight of each connection to the first hidden layer. All of the connections to each node in the hidden layer are summed and the corresponding bias is added. Afterwards, this sum is passed through the activation function, to calculate the node score. This node score is passed forward identically through the subsequent hidden layers until reaching the output layer, where the node scores are finally the actual network output. Therefore, this structure is called *feedforward* neural network. The term hidden layer comes from the fact that the node scores for these layers are hidden from the user.

Training a (deep) neural network entails the estimation of the weights and biases for all neurons that lead to the best possible predictions. No prior intuition of the problem is necessary up front when these weights and biases are initialised. The training procedure starts with random weights and biases, and applies the network to the training set. In order

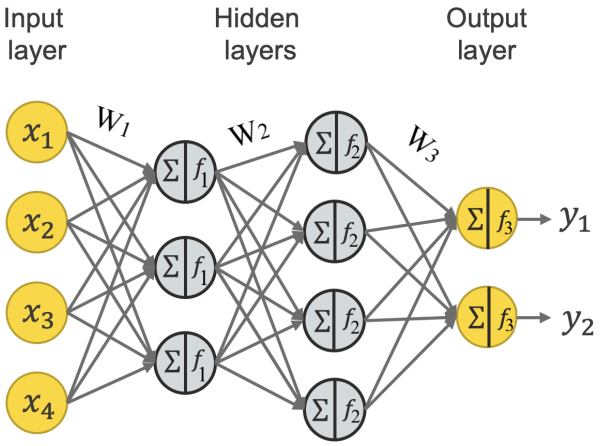


Fig. 6.10: Model of a fully connected, feedforward Deep Neural Network containing two hidden layers [248].

to evaluate the model at making predictions, the so-called *loss function* is employed, which measures the error between the predicted value and the expected outcome. The gradient of the loss function indicates how the parameters should be updated to make the model more accurate. Eventually, the weights in the network are updated using a type of gradient descent, which is an optimisation algorithm utilised to minimise the loss function. An efficient way to calculate the gradient is called *back propagation* [249], according to which the network output is compared to the true value of some training data, and the weights are then updated to better approximate the true value next time. Afterwards, the network with the new weights is again applied to the training set and the procedure is repeated until we have a set of good weights, which describe data as precisely as possible.

The size of the minimisation steps of the gradient descent is a tunable parameter, hence a hyperparameter, called *learning rate*. A very high learning rate can prevent convergence because the optimisation may overshoot the minimum. In contrast, a too low learning rate, although more precise, slows down the optimisation and the optimiser might get stuck in a local minimum. The training of a DNN is performed in batches, namely the training events are divided into equally sized segments. The weights of the DNN are updated after every batch. The number of events per batch is the *batch-size*. It is important that every batch is an adequate representation of the full dataset, typically realised by shuffling the full dataset before slicing the sample. A full iteration of the network over the whole set of events is called *epoch*. As a result, the batch-size and the number of epochs, together with the number of hidden layers and nodes per hidden layers are the most important hyperparameters which determine the training process.

Besides the training performance, an important feature of a DNN model is to be generalisable and to not depend on fluctuations in the training data, i.e. avoid overfitting and overtraining. Similarly to a BDT, a DNN can be sensitive to overtraining, thus optimal choices of the various hyperparameters is of determining importance, in order to avoid this effect. *Overfitting* occurs when a model corresponds too closely to a particular set of data, and may therefore fail to fit to additional data or predict future observations reliably. In order to prevent overfittig, the capacity of the model needs to be sometimes limited producing a simpler and more robust model. This is achieved by using stochastic regularisation methods, such as Dropout [250]. The Dropout method randomly drops a certain percentage of node connec-

tions to neighbouring layers avoiding complex neuron co-adaptions.

DNN training in $t\bar{t}H(H \to b\bar{b})$ analysis

The single-lepton boosted region is originally defined using requirements on RC jets as well as b-tagging of subjets and additional small-*R* jets, as detailed in Sec. 6.4.1. Besides, the boosted region definition can be further improved by utilising a high performance jet tagger, which exploits the primary difference between signal and background owing to the presence of a Higgs boson. Because of the use of RC jets in this analysis, the existing jet taggers, that were used to tag the Higgs boson [251] or the top quark [252] by the time that this analysis was being developed, were not suitable. As a result, a custom-made RC jet tagger has been developed by the Glasgow group using a DNN [253] and is exploited to improve the reconstruction of the Higgs-boson and top-quark final objects.

A multi-class DNN is trained to identify the most likely parent particle of the RC jets, while effectively discriminating between jets produced by top quarks and Higgs bosons, as well as those produced by additional QCD activity. The fact that each of these jet classes have some similar but distinguishable characteristics is deployed; top quarks are the most massive and contain a *b*-jet, while Higgs bosons are slightly less massive and contain two *b*-jets. The third category of jets is expected to be less massive, and contain mostly light quarks and/or gluons. This DNN tagger results in a multi-class output corresponding to classification as a Higgs-boson, top-quark, or QCD jet.

The network is built and trained with the KERAS software package [254], using the TensorFlow backend [255]. The performance of the DNN tagger is dependent on the architecture and training hyperparameters that have been set to optimise the three-class sigmoid output, in order to have a probability for each category. The network hyperparameters optimised for this tagger are listed in Table 6.10. Any hyperparameters not listed there are left to KERAS and TensorFlow defaults. The network architecture is chosen to be 3 layers of 100 nodes, having been chosen to balance performance and training time. Similarly, the number of epochs was selected to be 50 to prevent overtraining. Furthermore, the hidden layers each use the *ReLU* activation function. The choice of each hidden layer to have the same number of nodes and the same activation was chosen for simplicity and not optimised. The final layer is a 3 node softmax[2] output, which constrains the output nodes to sum to unity. This corresponds to probabilities for the three possible labels: *P(H)*, *P(t)* and *P(QCD)* for *Higgs-boson*, *top-quark* and *QCD jets* respectively, with $P(H) + P(t) + P(QCD) = 1$. In this way, a jet can be tagged by assigning it to the category which has the highest output score.

The Higgs-boson and top-quark categories have been obtained by matching the RC jets to a hadronically decaying Higgs boson or top quark at generator level. As a consequence, RC jets whose subjets match to two *b*-quarks, within a cone of $\Delta R = 0.4$, are labelled as *Higgs-boson jets*. Additionally, RC jets whose constituents are matched to one *b*-quark and at least one *W*-boson decay product are labelled as *top-quark jets*. Finally, all the other RC jets are labeled as *QCD jets*. The DNN is trained jet by jet, using the nominal $t\bar{t}H$ POWHEGBOX+PYTHIA8 sample. Events are selected requiring at least one isolated lepton with $p_T > 27$ GeV, and at least four small-*R* jets, at least two of which are *b*-tagged with 85% efficiency WP. Also, two RC jets are required, each with $p_T > 200$ GeV, $m > 50$ GeV, and at least two constituents. The selected

[2]Softmax is a function that assigns probabilities to each target class over all possible target classes in a multi-class problem. It is implemented through a neural network layer just before the output layer. The Softmax layer must have the same number of nodes as the output layer.

Hyperparameter	Value
Number of Layers	3
Nodes per Layer	100
Activation per Layer	ReLu
Learning Rate	0.01
Epochs	50
Decay Rate	10^{-6}
Momentum	0.2

Table 6.10: The hyperparameters used in the DNN training. Any hyperparameters not listed in this table are left to KERAS [254] and TensorFlow [255] defaults.

jets are split into two orthogonal samples to be used for the training and the testing processes, independently.

For the DNN training, 17 variables, related to the RC jet and its constituents, were used as input and are summarised in Table 6.11. These variables were selected from a larger set of variables based on the separation power for each signal and background hypothesis, defined for a binned distribution as

$$< S^2 >= \frac{1}{2}\sum_{i=1}^{N}\frac{(s_i - b_i)^2}{s_i + b_i} \tag{6.3}$$

where *N* is the number of bins in the distribution, s_i is the signal yield in bin *i*, and b_i is the background yield in bin *i*, after the total signal and background yields have been normalised to unity. As the primary difference between signal and background is in the Higgs boson, discrimination of Higgs-boson jets was prioritised for the final selection of the input variables. The separation power of the selected input variables is depicted in Fig. 6.11 and their correlations, in the $t\bar{t}H$ signal MC, are shown in Fig. 6.12. The invariant mass of the *b*-tagged subjets is the most separating variable for the discrimination of Higgs-boson from QCD jets, and simultaneously highly ranked for Higgs-boson vs top-quark jet discrimination. The most difficult discrimination in the network is for top-quark vs QCD jets, where the highest ranked variable is the RC jet mass. The p_T of the sub-leading subjet discriminates well between top-quark and QCD jets, since QCD jets typically have only a single hard prong.

A discriminant function *P* has been built for each category and an optimised working point is defined to obtain a boosted-object tagger with a specific signal efficiency. This results in probabilities for the three possible labels: P(H), P(t) and P(QCD) for Higgs-boson, top-quark and QCD jets respectively, as described above. These output distributions from the DNN are illustrated in Fig. 6.13, showing good separation for each category of jets. P(H), P(t), or P(QCD) values towards 1 signify an increasing probability (up to 100%) that a reconstructed object originates from the true Higgs-boson, top-quark or QCD jet, respectively. On the contrary, values towards 0 indicate that a reconstructed jet has a decreasing probability (up to 0%) to come from the true Higgs-boson, top-quark or QCD jet, accordingly.

Furthermore, the probability to mis-reconstruct the objects with this technique is quantified by the confusion matrix shown in Fig. 6.14. It shows both the true label (*y*-axis) and the predicted label (*x*-axis) for each jet, while the tagging prediction per true category has

Variable	Definition
$m_{\text{RC jet}}$	Mass of reclustered jet
$\sqrt{d_{12}}$	First splitting scale
$\sqrt{d_{23}}$	Second splitting scale
Q_W	Minimum invariant mass of constituent pairs
$n_{\text{constituents}}$	Number of constituents in the RC jet
$p_T^{\text{const}_1}$	Leading p_T of constituent jets
$p_T^{\text{const}_2}$	Sub-leading p_T of constituent jets
B_{const_1}	b-tagging discriminant of leading p_T constituent jet
B_{const_2}	b-tagging discriminant of sub-leading p_T constituent jet
$\Delta R_{\text{const}_1,\text{const}_2}$	ΔR between leading and sub-leading p_T constituent jets
$m_{b\text{-jets}}$	Invariant mass of all b-tagged constituent jets
$m_{\text{light-jets}}$	Invariant mass of all untagged constituent jets
$B_{\text{const}}^{\text{min}}$	Minimum b-tagging discriminant of constituent jets
$B_{\text{const}}^{\text{max}}$	Maximum b-tagging discriminant of constituent jets
$\Delta R_{\text{const}}^{\text{max}}$	Maximum ΔR between two constituents
$\Delta R_{\text{const}}^{\text{min}}$	Minimum ΔR between two constituents
B_{rest}	b-tagging discriminant (Sec. 5.3.1) of all constituents except leading and sub-leading p_T jets

Table 6.11: List of variables included in the DNN training in the single-lepton boosted channel. The jet substructure variables $\sqrt{d_{12}}$, $\sqrt{d_{23}}$ [256] and Q_W are calculated using the constituents information of the RC jets.

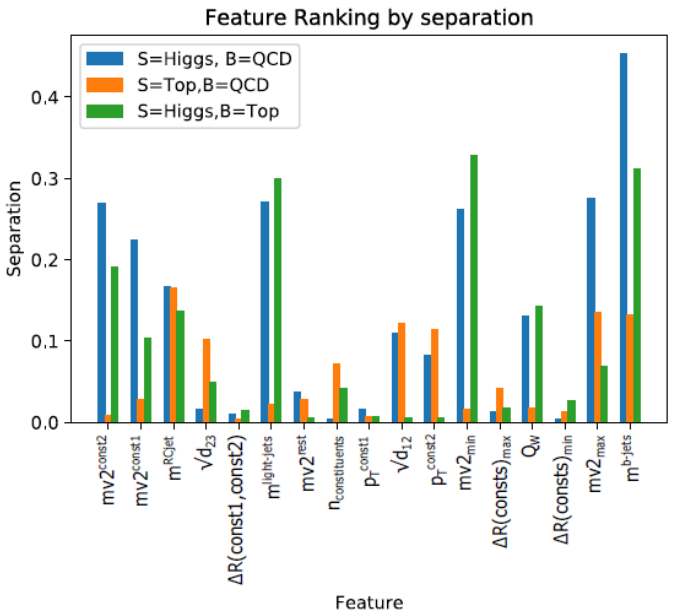


Fig. 6.11: Separation power (eq. 6.3) for input variables used in DNN training, between Higgs-boson/QCD (blue), top-quark/QCD (orange) and Higgs-boson/top-quark (green) jets [253].

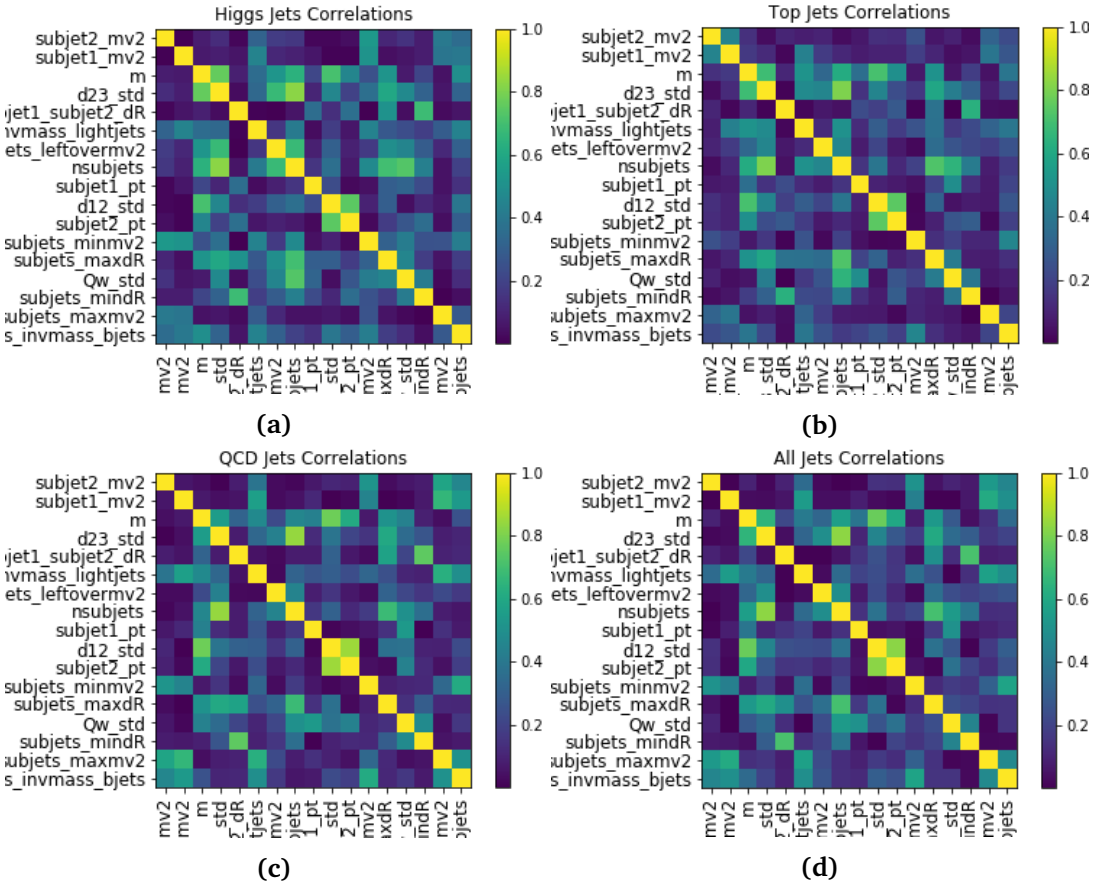


Fig. 6.12: Correlation matrices among the DNN input variables, in $t\bar{t}H$ signal MC, for (a) the Higgs-boson jets, (b) the top-quark jets, (c) the QCD jets and (d) all the jets [253].

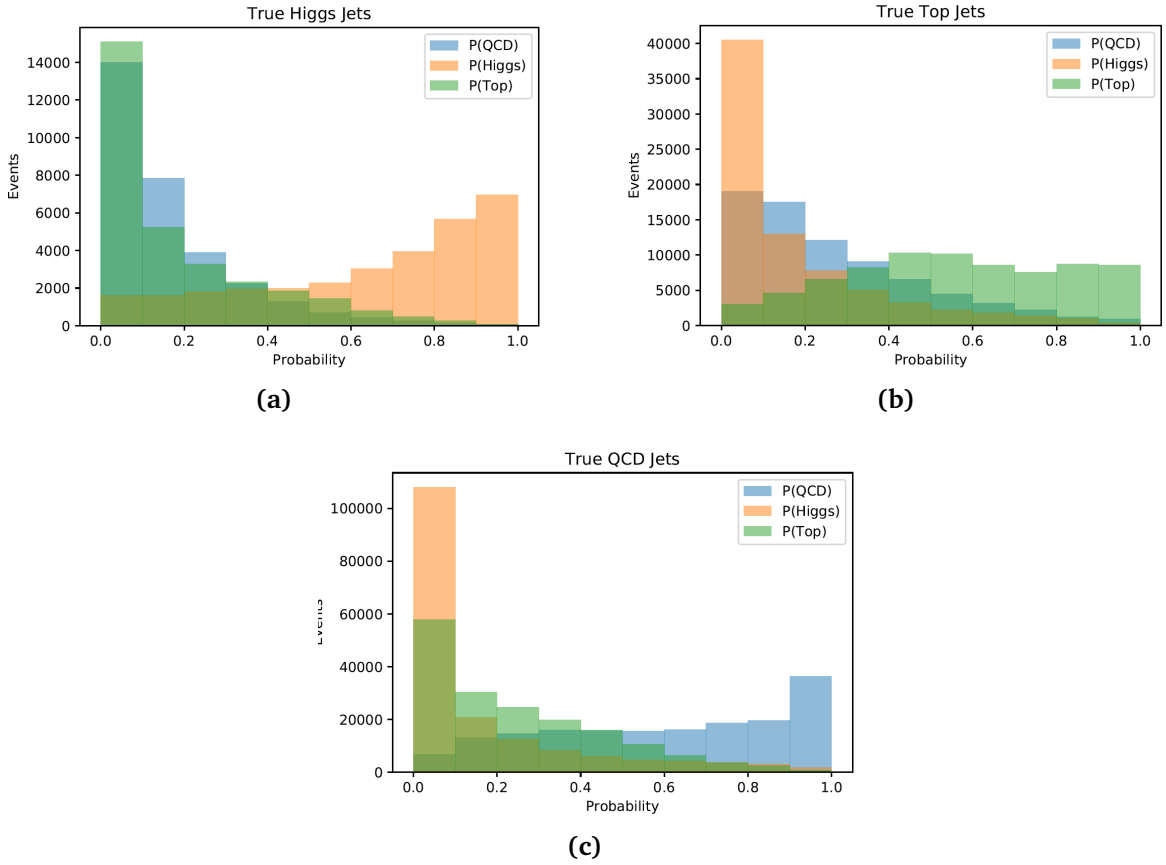


Fig. 6.13: DNN output distributions: the probabilities P(H), P(t) and P(QCD), for the true-reco matched (a) Higgs-boson jets, (b) top-quark jets, and (c) QCD jets. In each case, a high discrimination power is shown [253].

been normalised by row. As a result, greater diagonality corresponds to higher tagging performance. The confusion matrix demonstrates that the Higgs jets are correctly identified 76% of the time, whereas the top-quark jets are correctly identified 67% of the time.

Eventually, in order to improve the single-lepton boosted region definition, a cut on the DNN output distributions P(H) and P(t) is applied. Figures 6.15 depict the full distributions P(H) (Figs. 6.15a-6.15b) and P(t) (Figs. 6.15c-6.15d) for the Higgs-boson and hadronic top-quark candidates respectively, according to the boosted region definition (Sec. 6.4.1) without having applied the cuts on P(H) and P(t). Considering that the DNN tagger is based on RC jets, it can be applied only to the hadronic top-quark candidates reconstructed within large-*R* jets.

The P(H) distribution for the Higgs-boson candidate (Fig. 6.15a) shows some separation between signal and background events. In particular, most of the signal events are concentrated to values $P(H) \to 1$, indicating that the reconstructed Higgs-boson candidate is most of the times identified as originating from a true Higgs-boson jet with high probability. By contrast, one would ideally expect the background events to lie at $P(H) \to 0$. However, the $t\bar{t}+ \geq 1b$ events also increase with increasing P(H) values, while the rest of the background channels have a roughly flat distribution. Background events at $P(H) \to 1$ denote that there are objects wrongly reconstructed as the Higgs-boson candidate, which is identified as being actually a true Higgs-boson jet with high probability. Besides, the P(H) distribution for the

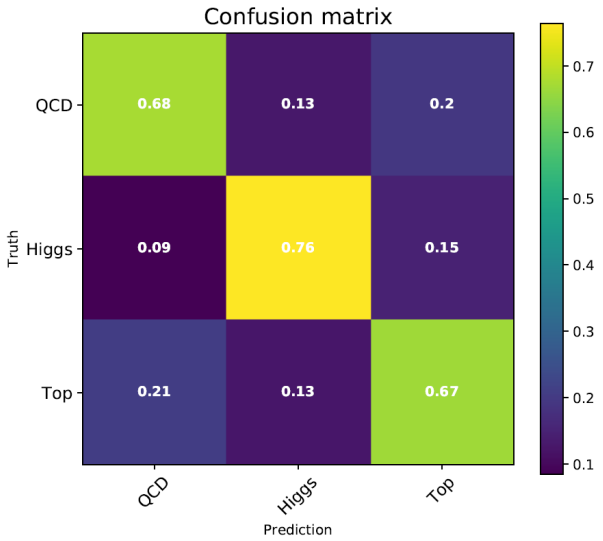


Fig. 6.14: Confusion matrix for the DNN outputs. The *x*-axis shows the prediction according to the DNN output, while the *y*-axis shows the true classification [253].

hadronic top-quark candidate (Fig. 6.15b) illustrates almost no separation between signal and background events. Both of them increase at lower P(H) values, creating a distinct peak at $P(H) \sim 0$. Hence, in most of the events the reconstructed hadronic top-quark candidate has a very small probability to originate from a true Higgs-boson jet.

Then, the P(t) distribution for the hadronic top-quark candidate (Fig. 6.15d) demonstrates a small separation between signal and background events. However, the signal events are almost evenly distributed along the P(t) values, signifying that roughly in the same amount of events the reconstructed hadronic top-quark candidate originates from a true hadronic top-quark jet with any probability. On the other hand, the background events roughly diminish with increasing P(t). So, in most events there are objects wrongly reconstructed as the hadronic top-quark candidate, which is actually identified as a true hadronic top-quark jet with smaller probabilities. Also, the P(t) distribution for the Higgs-boson candidate (Fig. 6.15c) indicates almost no separation between signal and background events, with both kind of events lying in lower P(t) values. Thus, in most events the reconstructed Higgs-boson candidate has small probability to come from a true hadronic top-quark jet.

From these plots one can observe at which P(H) or P(t) values most of the signal or background events are concentrated for the corresponding reconstructed object. Then, a cut value is specified by considering to reject a region enriched in background and depleted in signal events. In particular, the requirement $P(H) \geq 0.6$ is included in the event selection as part of the Higgs-boson candidate reconstruction, and $P(T) \geq 0.3$ is required for the reconstruction of the hadronic top-quark candidate with a large-*R* jet. Eventually, these cuts do not result in a significant increase of the reconstruction efficiency for the Higgs or the hadronic top-quark candidate (only 1%-2% improvement). Nevertheless, they lead to a significant reduction of the background events, while losing only a small amount of signal events.

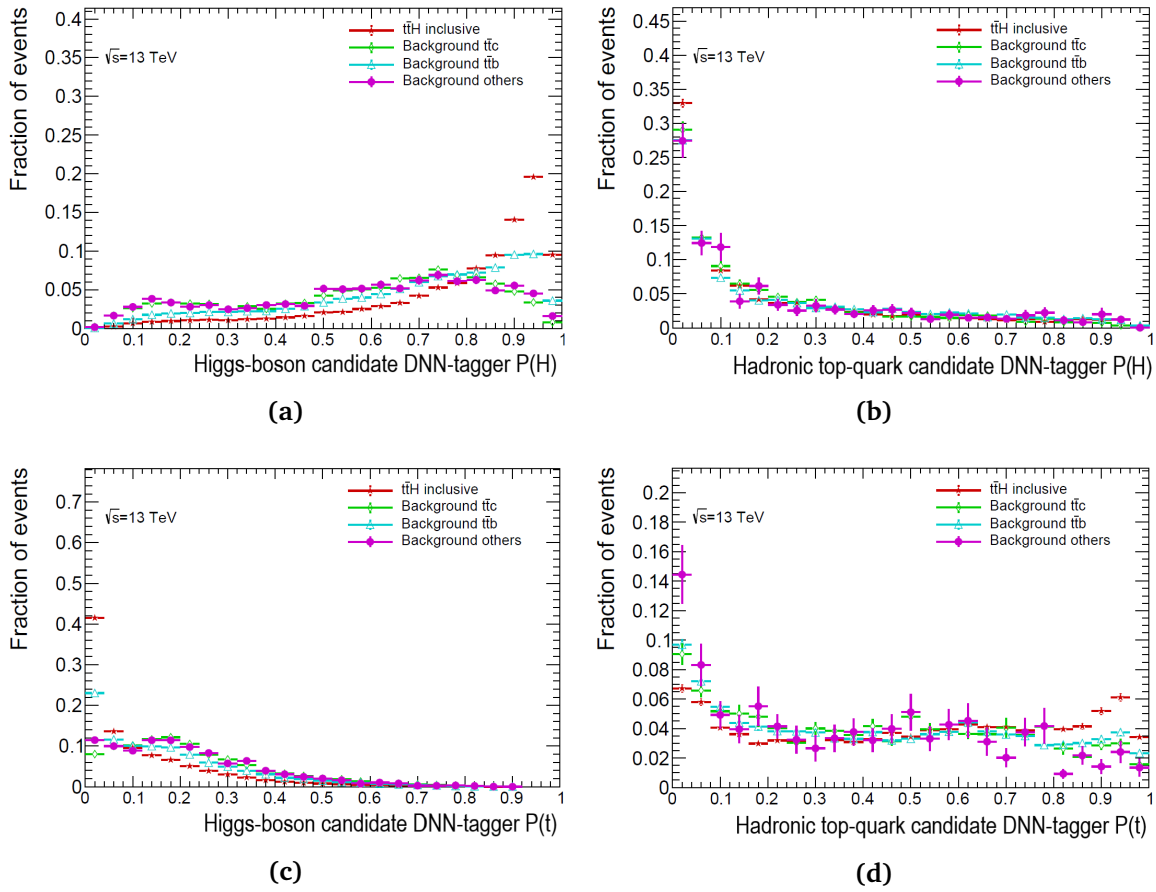


Fig. 6.15: DNN output distributions (a)-(b) P(H) and (c)-(d) P(t) for the Higgs-boson and the hadronic top-quark candidate, respectively. The $t\bar{t}H$ signal process, considering all the $t\bar{t}$ and Higgs boson decays, is depicted inclusive in true $\hat{p}_T^H$ and in the $\hat{p}_T^H$ bins $[300, 450)$ and $[450, \infty)$ GeV. The $t\bar{t}+ \geq 1b$ and $t\bar{t}+ \geq 1c$ background components are illustrated individually, while the $t\bar{t}$+light is grouped together with all the other background processes considered in the analysis, under "Others", since they have significantly smaller contribution. The distribution for each process is normalised to its total number of events.

6.5.3 Boosted Decision Trees (BDT)

A Boosted Decision Tree (BDT) is a structure of decision trees using the boosting technique, which offers a way for classification of events. The decision trees were formalised and developed by Breiman [241] in the context of pattern recognition and data mining, in order to extend a simple cut-based analysis into an MVA. This is achieved by continuing to analyse events that fail a particular criterion until they satisfy a terminating condition, so that they are classified in the best possible way.

A Decision Tree (DT) is a binary tree structured classifier with branches connected via nodes, as illustrated in fig. 6.16. The training of a DT starts from a root node that contains all the events, where an initial splitting criterion for the full training sample is determined, resulting in two subsets of training events. Each subset goes through the same algorithm iteratively, following a sequence of binary splits. At each node the events are separated into the corresponding classification category, i.e. signal or background. For each split, a cut on the discriminating input variable that gives the highest separation between signal and background

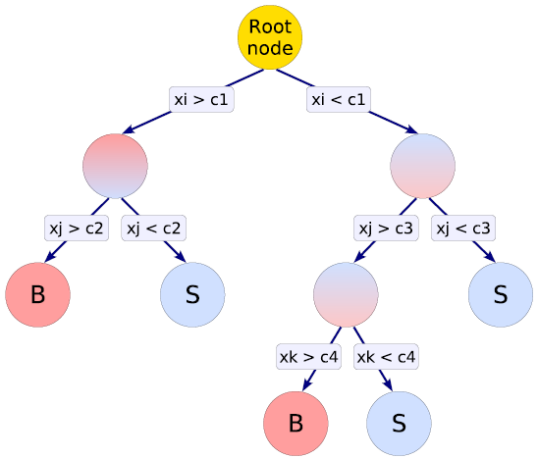


Fig. 6.16: Schematic representation of a decision tree [242]. In each node (circles) the events are divided according to a sequence of binary splits, based on whether they are more signal-like or background-like. For each split a specific discriminating variable (xi, xj, xk) is used, until some stopping condition (defined by the splitting thresholds $c1, c2, c3, c4$) is reached at each branching.

at this node is applied. Thus, the same input variable may be used at several nodes, while others might not be used at all. Also, the cut values are optimised by scanning over the variable range with a specified granularity (*nCuts*). The events are further divided until a stop criterion is fulfilled; such can be: the maximum depth of the decision tree allowed before further splitting is stopped (*MaxDepth*), the minimum percentage of training events in the final node (*MinNodeSize*), so as to avoid statistical fluctuations, or just reaching the desired separation. In this way, the phase space is split into many regions, eventually classified as signal or background, depending on the majority of training events that end up in the final node, also called *leaf* node.

In principle, the splitting could continue until each leaf node contains only signal or only background events, which could suggest that perfect discrimination is achievable. However, such a DT would be strongly overtrained. As already introduced, this means that the DT describes statistical fluctuations in the data set used for the training, leading to a performance that will not be reproducible on independent training data. To avoid overtraining a DT must be pruned, namely the most statistically insignificant nodes must be removed after the full decision tree has been built. Nevertheless, another way to drastically limit the tree depth of decision trees, far stronger than any pruning algorithm would do afterwards, and thus constrain overtraining, is to apply a boosting method already to the decision trees.

Single decision trees are considered a weak MVA method, given their limited separation power. Exploiting the *boosting* method, various decision trees are combined together, forming a "forest". Boosting stabilises the response of the decision trees with respect to fluctuations in the training sample and is a way to considerably enhance the classification performance with respect to a single DT. With this technique, the trees are iteratively created with an MVA algorithm from the same training ensemble by reweighting the wrongly classified events, and are after all combined into a single classifier which is given by an average of the individual decision trees. In terms of the $t\bar{t}H(b\bar{b})$ analysis, the Adaptive Boost (AdaBoost) [240] algorithm is used. This algorithm starts with the original event weights when training the first decision

tree. Then, the subsequent tree is trained using a modified event sample, in which the weights of the previously misclassified events are multiplied by a common "boost" weight. The effect of every boost step on the events' weights can be optimised via the parameter *AdaBoostBeta* for a specific model.

A BDT is still sensitive to residual overtraining, though, due to the precise manner in which the splittings are determined. In order to detect overtraining, the training sample of MC simulated data is split into two statistically independent samples based on the event number. In general, the performance on the training samples should not significantly outperform that on the test sample. Then, the BDT trained on even events is applied on odd events, and vice versa, referred to as *cross training*. Cross training profits from the full available statistics by evaluating the events in the one sample with the BDT trained on the other sample and the other way around. Finally, to counteract any remaining overtraining the various hyperparameters should be effectively optimised for a specific model.

After processing all decision trees, events are classified as signal or background depending on the weighted average of the individual tree classifications. This weighted average is the final *BDT score* which is the likelihood of an event to be a signal or background. The output values of the BDT lie between -1 and 1, and by convention signal (background) events accumulate at large (small) BDT output values.

Finally, in order to evaluate the performance of a trained BDT, *Receiver Operating Characteristic* (ROC) curves are used. These curves illustrate the background rejection versus the signal efficiency caused by a variation of the threshold on the BDT score. The *signal efficiency* (ε_S) is defined as the proportion of signal events above a particular threshold on the BDT score to all signal levels. Whereas, the *background rejection* is defined as "1 - background efficiency" (1 - ε_B), which is the proportion of rejected background by the same threshold. Better BDT performance means higher background rejection at similar signal efficiency, resulting in a more convex ROC curve. As a result, the best BDT performance can be identified by the largest AUC (Area under curve), which ranges from 0.5 to 1.

BDT training in $t\bar{t}H(H \to b\bar{b})$ analysis

As already introduced, the final state of the $t\bar{t}H(H \to b\bar{b})$ process is composed of many jets stemming from the Higgs-boson and top-quark decay products, as well as from additional radiation. Therefore, the boosted decision trees are exploited in the analysis to identify Higgs-boson and top-quark candidate objects (*reconstruction BDT*) as well as to enhance the discrimination between the $t\bar{t}H$ signal and the background events (*classification BDT*). Dedicated reconstruction BDTs are constructed for the single-lepton and dilepton resolved channels. These examine each possible combination of small-*R* jet assignments to the decay products of the hadronic top quark, leptonic top quark, and Higgs boson, and output a score indicating the combination most likely to be correctly assigned. Analogously, a classification BDT is trained independently in each of these channels as well as in the single-lepton boosted channel. The reconstruction and classification BDTs for the single-lepton resolved and dilepton channels are out of the scope of this thesis, hence are not described here. More details can be found in the reference [100].

The studies presented in this thesis are based on BDTs which are trained using the Toolkit for Multivariate Analysis (TMVA) [242] implemented in the ROOT data analysis framework [243]. The training parameters of the classification BDT in the single-lepton boosted channel, optimised to maximise the performance of the trained BDT and minimise its overtraining,

are listed in Table 6.12. In particular, this BDT is constructed from 1300 individual trees in the forest, a fact that offers a good separation between signal and background, while at the same time it prevents the BDT from adapting to statistical fluctuations in the training samples. Additionally, the maximal tree depth allowed is 3 nodes and the minimum percentage of training events in a leaf node is set to 2%. The number of grid points in the input variables' range used in finding the optimal cut in each node splitting is set to 16, which is a good compromise between computing time and step size. Last but not least, the boosting method employed for this BDT is the AdaBoost, which performs best on weak classifiers, such as the small individual decision trees with a tree depth of 3 which are used here. Although such small trees have very little discrimination power by themselves, they are much less prone to overtraining compared to simple decision trees and as an ensemble outperform them by far. The training performance is further enhanced by forcing a "slow learning" and allowing a larger number of boost steps, thus the learning rate of the AdaBoost algorithm is set to 0.3.

Hyperparameter	Value
BoostType	AdaBoost
AdaBoostBeta	0.3
NTrees	1300
MaxDepth	3
nCuts	16
MinNodeSize	2%

Table 6.12: TMVA BDT hyperparameters used for the classification BDT in the single-lepton boosted region.

The single-lepton boosted classification BDT has been trained on events selected with the loose selection, as described in Sec. 6.3.2, and not with the baseline that is used to define the boosted SR (Sec. 6.4.1). The choice of this looser selection has been primarily motivated by the higher statistics that it provides, thus being less prone to overtraining. After all, it is proven that it gives better overall performance (highlighted later in Fig. 6.20).

For the training, the nominal MC16 simulated samples for the release 21 analysis, so for the signal as for the various background processes (described in Sec. 4.5), have been used. Additionally, a true $\hat{p}_T^H \geq 300$ GeV is required for the $t\bar{t}H$ signal sample. This requirement helps to distinguish the $t\bar{t}H$ signal shapes between the $\hat{p}_T^H < 300$ GeV and $\hat{p}_T^H \geq 300$ GeV bins. All background processes are included in the training, since backgrounds other than $t\bar{t}$ + jets are expected to have a non-negligible contribution. After the BDT has been trained on MC simulated events, it is applied to data assuming that the same separation power holds. This is justified if all the used variables in the training and their correlations are well modelled in the simulation compared to data.

The classification BDT in the single-lepton boosted channel is built by combining several input variables that exploit the different kinematics of signal and background events, as well as information from *b*-tagging and the DNN training. A long list of prospective input variables has been examined in order to converge on an optimal set of variables. The modelling of these variables, their signal-to-background separation power, given by eq. 6.3, as well as the

correlations among them, are some of the features that play a determining role in the final selection of the variables that will serve as inputs to the BDT. Furthermore, a ranking of the BDT input variables is provided by the TMVA, evaluating the overall importance of each variable. It is estimated by counting how often the variables are used to split DT nodes, and by weighting each split occurrence by the separation gain achieved as well as by the number of events in the splitting node.

After all, the 21 variables listed in Table 6.13 are selected as inputs to the classification BDT in the boosted channel, since they maximise its performance. General kinematic variables, such as invariant masses, transverse momenta, pseudorapidities, angular separations of pairs of reconstructed jets, and the pseudo-continuous *b*-tagging discriminants of selected jets are included. In particular, $\eta^{\text{lep}}_{\text{Higgs}}$ is defined as $sign \times \eta_{\text{Higgs}}$, where η_{Higgs} is the pseudorapidity of the Higgs-boson candidate and *sign* is equal to 1 if the pseudorapidity of the lepton is $\eta_{\text{lep}} \geq 0$, otherwise *sign* is equal to -1. The $\eta^{\text{lep}}_{\text{had top}}$ variable is defined analogously. The variable $\eta^{\text{lep}}_{\text{Higgs}}$ ($\eta^{\text{lep}}_{\text{had top}}$) is constructed having as a reference frame the lepton in order to artificially increase the statistics in a specific η_{Higgs} ($\eta_{\text{had top}}$) region. This can be implemented since the symmetry of the variables η_{lep} and η_{Higgs} ($\eta_{\text{had top}}$) (defined in the laboratory frame) with respect to the transverse plane at $z = 0$ is taken into account (i.e. the signal or background events which are mirrored through the $z = 0$ plane happen with the same probability). Moreover, the output of the DNN training, precisely the Higgs probability P(H), is used as input variable to the classification BDT. Also, the hadronic and leptonic top reconstruction, described in Sec. 6.4.1, allow to use their properties to improve the final discrimination power of the BDT.

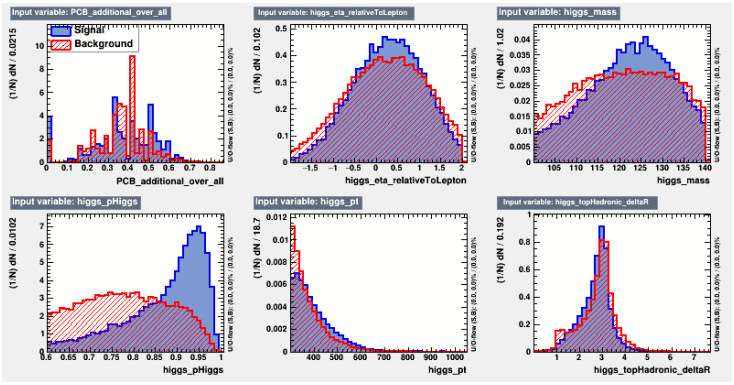
An overview of the separation between $t\bar{t}H$ signal and the backgrounds of the input variables to the classification BDT in the single-lepton boosted channel is illustrated in Fig. 6.17. Each individual variable shows either larger or smaller discrimination between signal and background. The ranking of the input variables based on the size of their separation power, as well as their overall importance during training, are reported in Table 6.14. The most important variables entering the single-lepton boosted BDT, based on their separation power, include the DNN P(H) output for the Higgs boson candidate, the sum of b-tagging discriminants of small-R jets from Higgs, hadronic top and leptonic top candidates, the hadronic top candidate's invariant massthe fraction of the sum of b-tagging discriminants due to all jets not associated with the Higgs or hadronic top candidates, the small-R jet multiplicity. Also, most of the highest ranked variables have the largest overall importance as well. Plots with data/MC comparison of the five highest ranked BDT training input variables are shown and discussed later in Sec. 8.2.

The linear correlations among the input variables for both the signal and the backgrounds are shown in Fig. 6.18. Evidently, there are only a few variables highly (anti-)correlated, however most of them have different amount of correlation in the signal compared to the background samples. Hence, this is still a kind of discrimination power that can contribute to the total BDT performance.

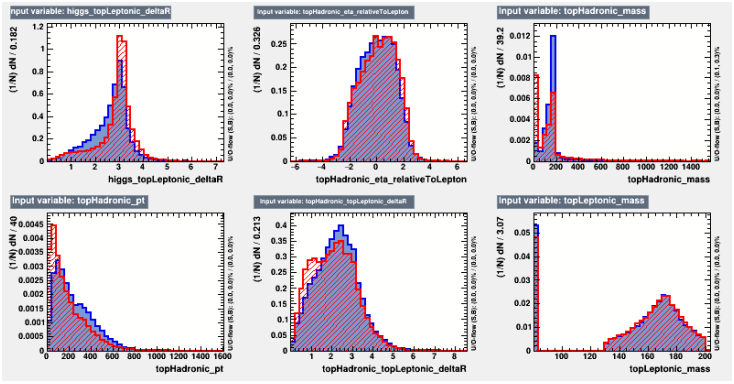
In order to remove a potential bias from overtraining, the classification BDT is trained on events with even numbers and applied to events with odd numbers and vice versa. Any remaining overtraining has no impact on the Data/MC agreement of the BDT fitted distribution, but may yield to a sensitivity loss. The overtraining is checked by comparing the BDT response for the training and the testing samples, as shown in Fig. 6.19a and Fig. 6.19b for the even and odd trainings, respectively. Moreover, a direct comparison between responses with only the training on even events and only the training on odd events has been performed on the full

Variable	Definition
m_{bb}^{Higgs}	Higgs candidate mass
p_T^H	Higgs candidate transverse momentum
$\eta_{\text{Higgs}}^{\text{lep}}$	η of the Higgs candidate relative to the lepton
$P(H)$	DNN Higgs probability for the Higgs candidate
$m_{\text{had top}}$	Hadronic top candidate mass
$p_T^{\text{had top}}$	Hadronic top candidate transverse momentum
$\eta_{\text{had top}}^{\text{lep}}$	η of the hadronic top candidate relative to the lepton
$B_{\text{had top}}^{i}$	i^{th} largest jet b-tagging discriminant associated to the hadronic top candidate
$m_{\text{lep top}}$	Leptonic top candidate mass
$p_T^{\text{lep top}}$	Leptonic top candidate transverse momentum
$B_{\text{lep top}}$	b-tagging discriminant of the jet associated to the leptonic top candidate
n_{jets}	Small-R jets multiplicity
$\Delta R_{H,\text{had top}}$	ΔR between the Higgs and the hadronic top candidates
$\Delta R_{H,\text{lep top}}$	ΔR between the Higgs and the leptonic top candidates
$\Delta R_{\text{had top,lep top}}$	ΔR between the hadronic top and the leptonic top candidates
$p_T^{t\bar{t}H}$	$t\bar{t}H$ system transverse momentum
$p_T^{t\bar{t}}$	$t\bar{t}$ system transverse momentum
$w_{b\text{-tag}}^{\text{sum}}$	Sum of b-tagging discriminants (Sec. 5.3.1) of jets from Higgs, hadronic and leptonic top candidates
$w_{b\text{-tag}}^{\text{add jet}}$	Fraction of the sum of b-tagging discriminants of all jets not associated to Higgs or hadronic top candidates

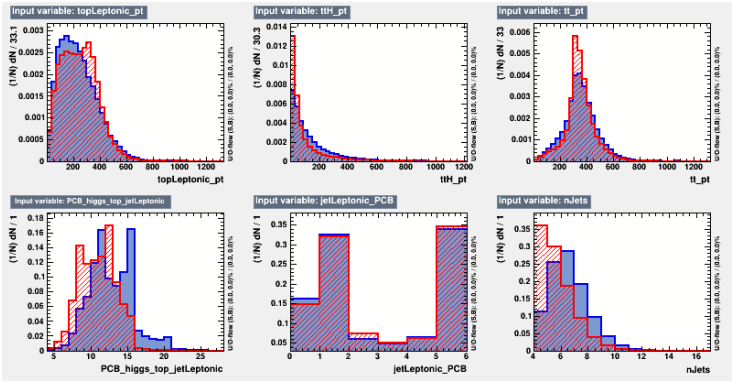
Table 6.13: Input variables to the classification BDT training in the single-lepton boosted channel. For variables depending on b-tagged jets, jets are sorted by their pseudo-continuous b-tag score, and by their p_T when they have the same pseudo-continuous b-tag score. The i index runs from zero to two.



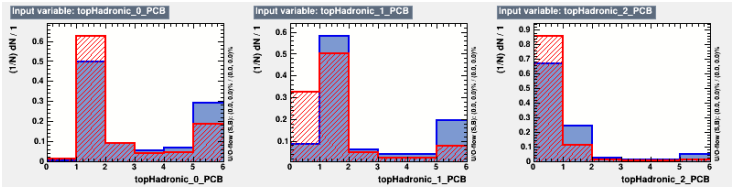
(a)



(b)



(c)

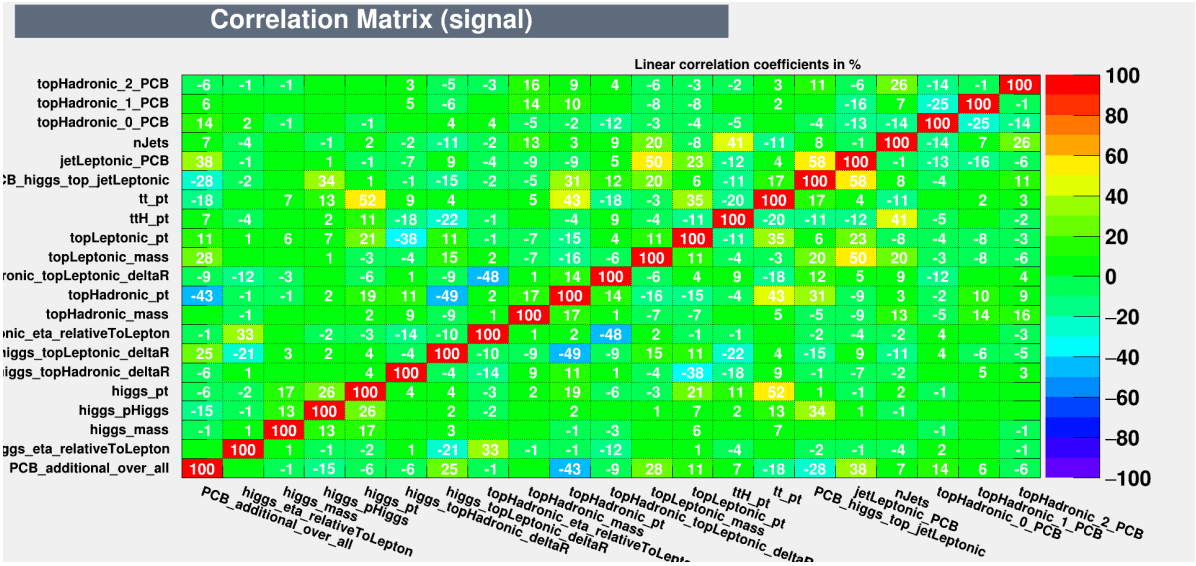


(d)

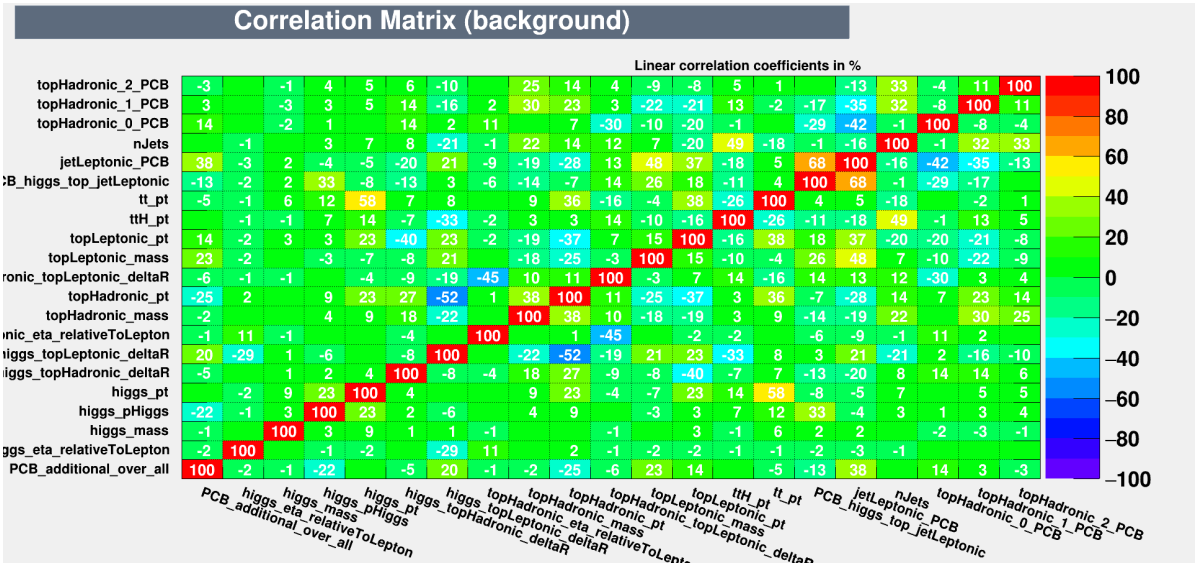
Fig. 6.17: Signal (blue) and background (red) distributions of the input variables for the single-lepton boosted classification BDT. Their separation power is also depicted.

Rank	Variable	Separation power	Importance
1	$P(H)$	0.2018	0.0793
2	$w_{b\text{-tag}}^{\text{sum}}$	0.1329	0.0630
3	$m_{\text{had top}}$	0.1325	0.0521
4	$w_{b\text{-tag}}^{\text{add jet}}$	0.1259	0.0552
5	n_{jets}	0.1164	0.0492
6	$B_{\text{had top}}^{1}$	0.1035	0.0462
7	$p_T^{\text{had top}}$	0.0533	0.0591
8	$\Delta R_{H,\text{lep top}}$	0.0529	0.0461
9	$p_T^{t\bar{t}H}$	0.0518	0.0469
10	$B_{\text{had top}}^{2}$	0.0502	0.0321
11	$B_{\text{had top}}^{0}$	0.0285	0.0395
12	p_T^{H}	0.0270	0.0479
13	$\Delta R_{H,\text{had top}}$	0.0263	0.0462
14	$p_T^{t\bar{t}}$	0.0167	0.0361
15	m_{bb}^{Higgs}	0.0144	0.0468
16	$\eta_{\text{lep}}^{\text{Higgs}}$	0.0118	0.0417
17	$\Delta R_{\text{had top,lep top}}$	0.0099	0.0524
18	$p_T^{\text{lep top}}$	0.0085	0.0442
19	$\eta_{\text{had top}}^{\text{lep}}$	0.0042	0.0431
20	$B_{\text{lep top}}$	0.0012	0.0362
21	$m_{\text{lep top}}$	0.0008	0.0368

Table 6.14: TMVA variable ranking for the single-lepton boosted classification BDT based on the separation power of the variables. The amount of their separation power (3^{rd} column) as well as of their importance (last column) are also included here.



(a) signal sample



(b) background samples

Fig. 6.18: Correlation matrices of input variables of the single-lepton boosted classification BDT in the (a) signal sample and (b) all background samples.

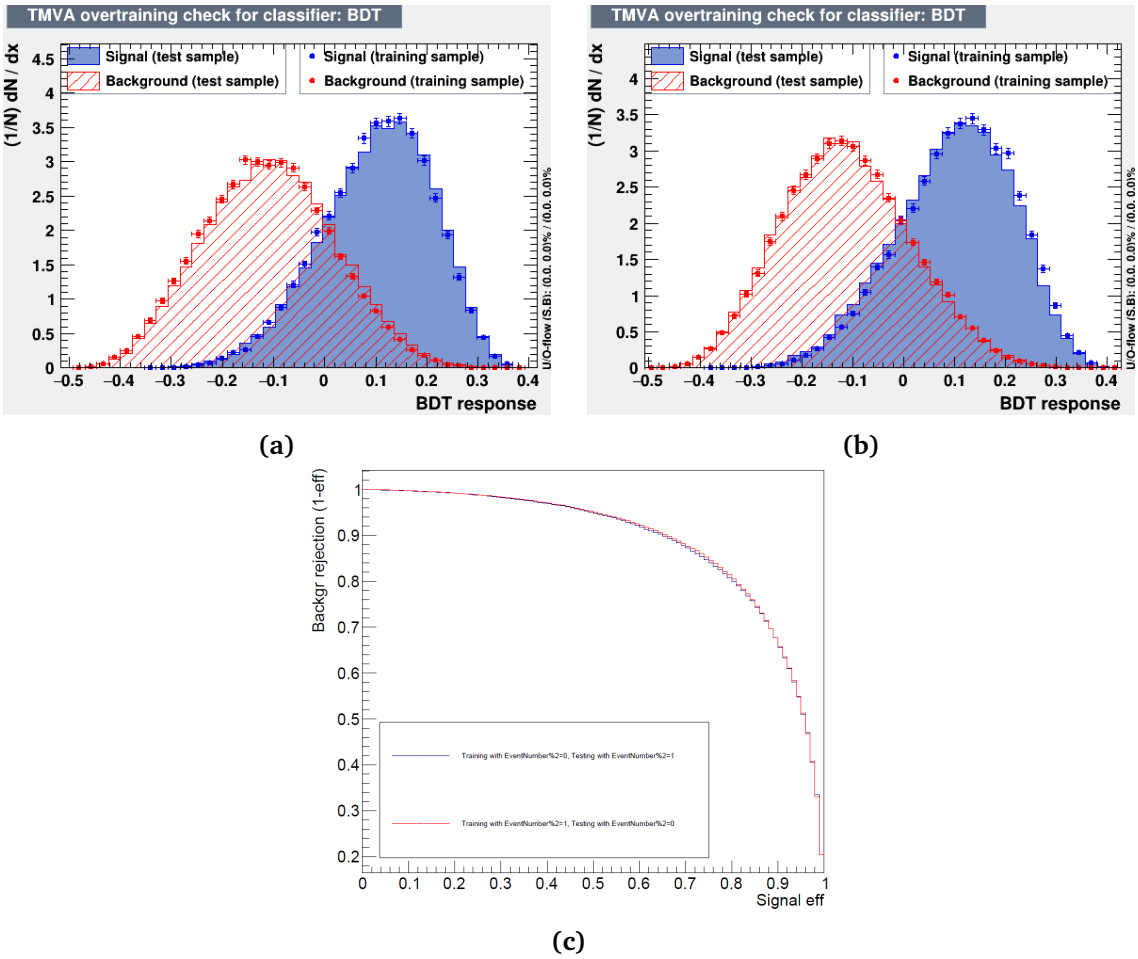
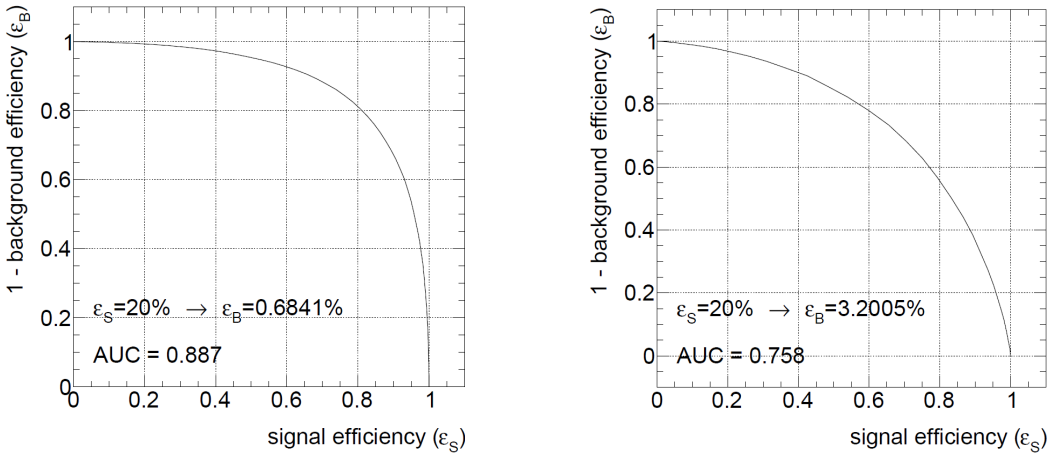


Fig. 6.19: Comparison of single-lepton boosted classification BDT response for training (dots) and testing (histogram) samples for signal (blue) and background (red) for the training on (a) even and (b) odd events. (c) Superimposed BDT ROC curves with only the training on even events (blue) or odd events (red).

MC16 samples and shown in Fig. 6.19c. The similarity between testing and training is found to be sufficiently large, hence no explicit sign of overtraining is observed.

After all, the performance of the training in the single-lepton boosted channel, is depicted by the ROC curve in Fig. 6.20a. The more convex the ROC curve is, and hence the closer the AUC is to 1, the better the performance of the trained BDT is. For completeness, the ROC curve for the BDT trained on events selected with the baseline boosted selection is included in Fig. 6.20b. Comparing the two ROC curves, it is evident that the former is more convex, giving higher background rejection (1 - ε_B) at similar signal efficiency ε_S, resulting in a larger AUC. As a consequence, the training on the looser selection demonstrates by far the better performance.



(a) loose boosted selection

(b) baseline boosted selection

Fig. 6.20: Comparison of classification BDT ROC curves for different selections in the single-lepton boosted channel.

Eventually, the aforementioned training input variables are combined in the BDT discriminant, that results in the best separation power between signal and background events compared to the input variables, reaching a separation of almost 45% (from eq. 6.3). Therefore, it is used as the final discriminant in the profile likelihood fit to data (discussed in Sec. 8) in the single-lepton boosted SR. In Figs. 6.19a-6.19b, it is apparent that the low bins ($BDT < -0.2$) of the classification BDT output have almost no signal events, whereas the high BDT bins ($BDT > 0.1$) are particularly enriched with signal events. The BDT helps to increase the overall signal to background (S/B) ratio from 6.8% (Fig. 6.4e) in the whole boosted SR to $\sim 45\%$ in the region with $BDT \geq 0.2$.

Chapter 7

Systematic Uncertainties

Various sources of systematic uncertainty affect the $t\bar{t}H(H \to b\bar{b})$ measurement, impacting the categorisation of events as well as the shape and normalisation of the final discriminants used in the signal extraction fit (detailed in Ch. 8). They are related either to the reconstruction of the physics object or to the modelling of the physics processes. All of them are described in the following.

The systematic uncertainties are evaluated by varying the discriminant distributions using the MC samples (introduced in Sec. 4.5), which may suffer from statistical fluctuations. The choice of the binning of the discriminant distributions is crucial so as to reduce the fluctuations in the predicted number of events. In order to further mitigate the impact of the statistical fluctuations, symmetrisation and smoothing algorithms are applied to the varied distributions.

7.1 Sources of systematic uncertainty

The systematic uncertainties are classified into two main categories. One includes the *experimental* (*instrumental*) *uncertainties*, which originate from an imperfect knowledge of the detector response, that entails an uncertainty on the parameters used in the identification and reconstruction of the various physics objects and their calibrations. The other category constitutes the *theoretical modelling uncertainties*, which arise from an approximate theoretical modelling of the signal and background processes.

The systematic uncertainties can affect the amount of signal and background estimated in the different regions, i.e. the overall normalisation (N) of a process, or also the shape (SN) of the final discriminants distributions, as indicated in Table 7.1. In particular, all sources of experimental uncertainty considered, with the exception of the one on the luminosity, affect both the normalisation and shape of the distributions in all the simulated event samples. Moreover, the modelling uncertainties affect both the normalisation and shape of the distributions, with the exception of cross-section and normalisation uncertainties which only affect the overall yield of the considered sample. Nevertheless, the normalisation uncertainties modify the relative fractions of the different samples leading to a shape uncertainty in the distribution of the final discriminant for the total prediction in the different analysis regions.

A single independent nuisance parameter (defined in Sec. 8.1) is assigned to each source of systematic uncertainty. Especially, most of the experimental uncertainties, are decomposed into several independent sources. Each individual source, then, has a correlated effect across all the channels, analysis regions, as well as signal and background samples. On the con-

trary, the modelling uncertainties affect only a single sample. Also, they are broken down into several components depending on the signal and background processes, as well as targeting specific physics effects in the event generation, such as scale variations or changing the hadronisation model. Thus, they are uncorrelated between different samples, while they are still correlated across channels and analysis regions with some exceptions though, explained in Sec. 7.1.2. In total, 216 nuisance parameters, corresponding to the systematic components and the free-floating $t\bar{t} + \geq 1b$ normalisation factor, are included in this analysis, presented in Table 7.1.

Last but not least, the statistical uncertainty arising from the limited statistics of the MC simulated samples, is also considered a systematic uncertainty. It is included in the likelihood in the form of additional nuisance parameters (see Sec. 8.1), one for each bin of the discriminant distributions considered in the analysis.

7.1.1 Experimental uncertainties

The $t\bar{t}H(H \to b\bar{b})$ measurement is based on the reconstructed objects, jets and leptons. The identification efficiencies are derived from simulation and are corrected with scale factors to match the data (see Sec. 4.4). In particular, the correction related to SFs applied on efficiencies for triggering, reconstructing, and identifying objects, is applied by modifying the event weight, whereas the correction related to the energy scales and resolutions is applied by smearing or re-scaling the energies of the objects. The uncertainties on these corrections have to be considered. They constitute the experimental systematic uncertainties and describe the performance of the detector, as well as of the reconstruction and calibration procedures of the physics objects. The experimental uncertainties affect all processes determined by MC simulation, as a result they are correlated across all samples. In general, the experimental uncertainties have a rather low impact on the final fit (shown in Ch. 8). Only the uncertainties associated to jets and *b*-tagging have a larger effect.

Luminosity and pileup

First of all, providing that the properties of the colliding bunches are not perfectly understood, an uncertainty of 1.7% on the integrated luminosity is considered for the full Run-2 dataset [229]. It is obtained using the LUCID-2 detector [230] for the primary luminosity measurement. Also, an uncertainty associated with the modelling of pile-up in the simulation is included to cover the difference between the predicted and measured inelastic cross-section values [231].

Leptons

Systematic uncertainties associated with leptons arise from the trigger, reconstruction, identification, and isolation, as well as the lepton momentum scale and resolution (outlined in Sec. 5.4.1 and 5.4.2). They amount to 22 in total, but have only a small impact on the final result. Efficiency SFs for the reconstruction, identification, and isolation efficiency of electrons and muons, as well as the efficiency of the trigger used to record the events, are measured using tag-and-probe techniques on $Z \to l^+l^-$ and $J/\psi \to l^+l^-$ data and simulated samples. They are applied to the simulation to correct for differences with respect to data. The effect of these SFs as well as of their uncertainties are propagated as corrections to the MC event weight,

Systematic uncertainty	Type	Comp.
Experimental uncertainties		
Luminosity	N	1
Pileup modelling	SN	1
Physics Objects		
Electrons	SN	7
Muons	SN	15
Jet energy scale	SN	31
Jet energy resolution	SN	9
Jet vertex tagger	SN	1
$E_{\mathrm{T}}^{\mathrm{miss}}$	SN	3
b**-tagging**		
Efficiency	SN	45
Mis-tag rate (c)	SN	20
Mis-tag rate (light)	SN	20
Signal and background modelling		
Signal		
$t\bar{t}H$ cross-section	N	2
H branching fractions	N	3
$t\bar{t}H$ modelling	SN	4
$t\bar{t}$ **Background**		
$t\bar{t}$ cross-section	N	1
$t\bar{t} + \geq 1c$ normalisation	N	1
$t\bar{t} + \geq 1b$ normalisation	N (free floating)	1
$t\bar{t}$ + light modelling	SN	4
$t\bar{t} + \geq 1c$ modelling	SN	4
$t\bar{t} + \geq 1b$ modelling	SN	17
Other Backgrounds		
$t\bar{t}W$ cross-section	N	2
$t\bar{t}Z$ cross-section	N	2
$t\bar{t}W$ modelling	SN	1
$t\bar{t}Z$ modelling	SN	1
Single top cross-section	N	3
Single top modelling	SN	7
W+jets normalisation	N	3
Z+jets normalisation	N	3
Diboson normalisation	N	1
$4t$ cross-section	N	1
Small backgrounds cross-sections	N	3

Table 7.1: Overview of all systematic uncertainties considered in the analysis. The expression "SN" means that both the shape and normalisation are taken into account for the uncertainty, whereas "N" stands for the normalisation effects only, for all processes and channels affected. The number of components in which each systematic uncertainty is split is indicated in the column labelled as "Comp.". "Small backgrounds" refers to the *tZq*, *tWZ*, *tHjb*, and *tWH* processes.

considering 4 (10) independent components for electrons (muons). Additional sources of uncertainty, originating from the corrections applied to adjust the lepton momentum scale and resolution in the simulation to match those in data, are derived using events of the aforementioned processes. To evaluate the effect of momentum scale uncertainties, the event selection is redone with the lepton energy or momentum varied by $\pm 1\sigma$. For the momentum resolution uncertainties the event selection is redone by smearing the lepton energy or momentum. In total, 3 (5) independent components are considered for electrons (muons).

Jets and heavy-flavour tagging

Moreover, systematic uncertainties associated with jets arise from the efficiency of pile-up rejection by the JVT, from the Jet Energy Scale (JES) and Resolution (JER) correcting the jet four-momentum, and from b-tagging. A total of 126 uncertainties associated to jets are considered, being the dominant ones among the experimental uncertainties. Finally, the small-*R* jet constituent uncertainties related to JES, JER, and JVT, are propagated to RC jets.

A total of 31 independent sources of systematic uncertainty on the JES and 9 on the JER are taken into account. The JES uncertainties (detailed in Sec. 5.2.3) are derived by combining information from test-beam data, LHC collision data and simulation. Additional uncertainties are included, such as those related to the jet flavour, assuming a conservative uncertainty of $\pm 50\%$ on the quark/gluon fraction for all MC samples. Furthermore, pileup corrections and uncertainties from jet kinematics (η dependence, high-p_T jets) are considered, as well as detector simulation differences (GEANT4 vs fast simulation - see Ch. 4.3). The JER, defined as a function of jet p_T and rapidity, as well as its uncertainties are measured in Run 2 data and MC using dijet events (detailed in Sec. 5.2.3). The combined uncertainties are propagated by smearing the jet p_T in simulation. Although the uncertainties are not large, varying between 1% and 5% per jet (depending on the jet p_T), their effect is enhanced by the large number of jets considered in the final state.

One more uncertainty is considered, corresponding to the efficiency to identify and remove jets from pile-up. As described in Sec. 5.2.4, SFs are applied to correct for discrepancies between data and MC for the JVT efficiency, estimated using $Z \to \mu^+\mu^-$ events with tag-and-probe techniques. The effect of these SFs as well as of their uncertainties are propagated as a correction to the MC event weight.

On condition that this analysis relies heavily on *b*-tagging, it also comprises a source of systematic uncertainties. As described in Sec. 5.3.2, *b*-tagging efficiencies in simulated samples are corrected to match efficiencies in data. The efficiency to correctly tag *b*-jets, is measured using dileptonic $t\bar{t}$ events, exploiting the very pure sample of *b*-jets arising from the decays of the top quarks. For *c*-jets mistag rates, single-lepton $t\bar{t}$ events are used, exploiting the *c*-jets from the hadronically decaying W-bosons. The mistag rates for *light*-jets are measured using the negative-tag method applied to Z+jets events. Then, the *b*-tagging calibrations provide uncertainties depending on the different WPs and the jet p_T, which amount to 2%–10% for tagging *b*-jets, and to 10%–25% or 15%–50% for mistagging *c*- or *light*-jets, respectively. For the calibration of the four WPs used in this analysis, a large number of uncertainty components emerge. A principal component analysis is performed, yielding 45, 20, and 20 uncorrelated sources of uncertainties (eigen-variations) for *b*-, *c*-, and *light*-jets, respectively. Each component is characterised by a number and they are all ordered based on the size of their impact in the phase-space they were derived from, so to an extent the order depends on the analysed phase-space.

Furthermore, due to the use of multiple working points of the MV2c10 tagger, the analysis makes use of the pseudo-continuous (PC) calibration - calibration of the MV2c10 output in 5 bins with boundaries corresponding to the efficiency WPs (0%, 60%, 70%, 77%, 85%, 100%). Unlike the cumulative WP calibration, the PC calibration does not have any high-p_T extrapolation uncertainties applied to the jets with p_T outside the calibration range. The *b*-tagging calibration limits for the corresponding true jet flavour are listed in Table 7.2. The sensitivity of the analysis on the missing high-p_T extrapolation uncertainties have been studied for events with at least one jet of the given flavour, above the relevant calibration limit. A test was performed by including the available uncertainties from cumulative WPs (details in App. A.4.1), but it had no significant effect on the expected sensitivity of the analysis. Another study was carried out by removing all events with at least one jet outside the calibration range of either flavour (see App. A.4.2). This study is considered very conservative, given the fact that there is a very high number of jets in the events of the analysis, and having for example one jet outside the calibration range should not have a huge effect in the analysis. In this case, only a small effect on the expected sensitivity was observed. However, this is expected since there is a change on the shape of the fit input distributions, which can then cause a difference in the effect of the modelling systematics. Finally, it was decided to not add any uncertainties to account for b-tagging at high-p_T.

true jet flavour	*b*-tagging calibration limit: p_T [GeV]
b	600
c	250
light	300

Table 7.2: The *b*-tagging calibration limits for each true jet flavour.

Missing Transverse Energy

As defined in Sec. 5.5, the missing transverse momentum is calculated from the reconstructed physics objects and a soft term. Therefore, all aforementioned uncertainties on energy scales or resolutions of the reconstructed objects are propagated to its calculation. Three additional independent uncertainties associated with the scale and resolution of the soft term are also included, to account for disagreement between data and MC for the p_T balance between the hard and soft components. They refer to an offset along the hard component p_T axis, as well as the resolution along and perpendicular to this axis. Considering that the missing transverse momentum is not used in event selection but only in event reconstruction, the associated uncertainties have a minimal impact on the analysis.

7.1.2 Theoretical modelling uncertainties

The modelling uncertainties arise from the approximate theoretical modelling of the signal and background processes through the MC simulation. In principle, these sources of systematic uncertainty are estimated by varying the various parameters of the nominal MC generator and by comparing alternative generators to the nominal one. By contrast to experimental uncertainties, the modelling uncertainties affect a single sample, hence they are not correlated

across all background and signal processes. Besides, they are correlated across the analysis regions, apart from specific cases discussed below.

Uncertainties from the variations of the nominal model parameters are determined based on their physics interpretation. Specifically, a systematic uncertainty related to varying the amount of the initial state radiation (ISR) originates from two sources. One is the simultaneous variation of the factorisation μ_F and renormalisation μ_R scales of the matrix element (described in Sec. 4.2.2). Among other things, they affect properties of the additional gluon emission included in the matrix element. The other component comes from the variation of the parameter for the QCD emission of the ISR in the parton shower, α_s^{ISR}. In contrast, an uncertainty related to the amount of the final state radiation (FSR) is estimated by varying the parameter α_s^{FSR} for the QCD emission of the FSR in the parton shower. The variation of the matrix element is only a part of the ISR and not of the FSR, because the gluon emission from a top quark is suppressed due to its large mass, thus it is not a part of the matrix element in POWHEGBOX. Consequently, the impact of the ISR variation is larger than that of the FSR one. The effect of the systematic uncertainties resulting from parameter variations in the analysis is provided as event weights of the nominal samples, hence they are statistically correlated to the nominal sample.

To evaluate modelling uncertainties, apart from the variations of parameters in the nominal model, it is useful to consider alternative models as well. In this case, the systematic variations are derived from the comparison of two different MC generator setups and the uncertainty is extracted from their difference. They target one modeling component at a time, in order to minimise correlations among the different MC models. Such variations are called *two-point systematics* and two such uncertainties are considered in this analysis. In particular, to assess the uncertainty arising from changing the NLO matching procedure, the nominal sample is compared to an alternative setup which uses a different matrix element generator. Lastly, to estimate the uncertainty related to the choice of PS and hadronisation model, the nominal sample is compared to an alternative setup which uses a different parton shower generator. The difference in the parton shower between the two samples is mainly in the choice of the ordering variable of the shower (Sec. 4.2.3). Additionally, the hadronisation relies on a different approach (string vs cluster model) of the different generators (Sec. 4.2.4).

Signal modelling uncertainties

The nominal $t\bar{t}H$ signal sample is POWHEGBOX+PYTHIA8, as specified in Sec. 4.5.2 with its parameters. Various systematic uncertainties related to the modelling of the $t\bar{t}H$ process are considered in the analysis. In order to quantify the impact of ISR and estimate its uncertainty, a simultaneous variation of the scales μ_R and μ_F in the ME, by a factor 0.5 (2.0) for increased (decreased) parton radiation is considered. Additionally, the QCD emission parameter α_s^{ISR} in the PS [257] of the PYTHIA8 A14 tune is varied and set to 0.140 (0.115) rather than the nominal value 0.127. Analogously, the uncertainty on the FSR is evaluated by varying only the parameter α_s^{FSR} in the PS, as introduced above, which is set to 0.1423 (0.1147) instead of the nominal 0.127. The ISR and FSR uncertainties affect the acceptance and shape of the distribution of several kinematic variables. Furthermore, the nominal POWHEGBOX+PYTHIA8 sample is compared with the POWHEGBOX+HERWIG7 sample to assess an uncertainty related to the choice of PS and hadronisation model, as well as with the MADGRAPH5_AMC@NLO+PYTHIA8 sample to assess the uncertainty arising from changing the NLO matching procedure.

In addition to the modelling systematic variations, theoretical uncertainties on the predicted SM $t\bar{t}H$ signal cross-section are estimated, with a particular focus on the impact on STXS $\hat{p}_T^H$ bins (defined in Sec. 6.1). An uncertainty of $\pm 3.6\%$ from varying the PDF and α_s in the fixed-order calculation is applied [106, 259–261]. The effect of PDF variations in the ME on the acceptance and shape of the distributions considered in this analysis, for the signal process, is found to be negligible. Uncertainties in the Higgs boson branching fractions are considered as well, and amount to 2.2% for the $b\bar{b}$ decay mode [106].

Moreover, uncertainties due to missing higher-order terms in the perturbative QCD calculations, affecting the total cross-section and event migration between STXS p_T^H bins, are evaluated. This is achieved by varying the scales μ_R and μ_F independently by a factor 0.5 (2.0), as well as evaluating the ISR and FSR uncertainties. The largest effect was found to originate from the ISR uncertainty, corresponding to a 9.2% variation of the total cross-section, leading to an uncertainty of 10%–17% in STXS bin migrations, retrieved using the Stewart–Tackmann procedure [258]. All signal uncertainties are correlated across STXS bins, except for bin migration uncertainties.

Background modelling uncertainties

Since the $t\bar{t}$ + jets events represent by far the largest source of background in the analysis, a wide range of uncertainties is considered for them. They are summarised in Table 7.3, where they are distinguished, similarly to the $t\bar{t}$ + jets classification, in the subcategories: $t\bar{t}+\geq 1b$, $t\bar{t}+\geq 1c$ and $t\bar{t}$ + light, since each process is affected by different types of uncertainties. In particular, $t\bar{t}$ + light profits from relatively precise measurements in data. Also, $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ can have similar or different diagrams depending on the precision of the ME and the flavour scheme used for the PDF, while the different masses of the *c*- and *b*-quarks contribute to additional differences between these two processes. Therefore, all uncertainties associated with the $t\bar{t}$ + jets background modelling are assigned independent nuisance parameters for the $t\bar{t}+\geq 1b$, $t\bar{t}+\geq 1c$ and $t\bar{t}$ + light processes, which are uncorrelated among each other. Nonetheless, the uncertainty of each category is correlated across the STXS $\hat{p}_T^H$ bins, with some exceptions that are explained below.

An uncertainty of 6% is considered for the inclusive $t\bar{t}$ production cross-section predicted at NNLO, including effects from varying the factorisation and renormalisation scales, the PDFs, α_s, as well as the top-quark mass [262–265]. This uncertainty is applied only to $t\bar{t}$ + light samples, since this component is dominant in $t\bar{t}$ production in the full phase-space. An uncertainty of 100% in the normalisation of $t\bar{t}+\geq 1c$ events is applied, motivated by the fitted value of this normalisation in the previous analysis [97], where it was a free-floating parameter. This is the consequence of tightening the event selection criteria, reducing significantly the contribution of the $t\bar{t}+\geq 1c$ background component. On the contrary, the normalisation of $t\bar{t}+\geq 1b$ is kept to float freely in the signal extraction fit, designated as $k(t\bar{t}+\geq 1b)$. Similarly to the signal process, the effect of PDF uncertainties was found to be negligible also for the $t\bar{t}$ + jets background process.

As introduced in the beginning of this section, systematic uncertainties in the acceptance and shape of the distributions are extracted from comparisons between the nominal prediction and different MC samples or settings. However, such comparisons would change the fraction of $t\bar{t}+\geq 1b$ events in the phase-space selected by the analysis, which is a poorly modelled property. This is the reason why the normalisation of the $t\bar{t}+\geq 1b$ background component is left free-floating in the fit, thus the normalisation effects of the $t\bar{t}+\geq 1b$ systematic variations

Uncertainty source	Description		Components
$t\bar{t}$ cross-section	$\pm 6\%$		$t\bar{t}$ + light
$t\bar{t} + \geq 1b$ normalisation	Free-floating		$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 1c$ normalisation	$\pm 100\%$		$t\bar{t} + \geq 1c$
NLO matching	MADGRAPH5_aMC@NLO + PYTHIA8 vs POWHEGBOX + PYTHIA8		All
PS & hadronisation	POWHEGBOX + HERWIG7 vs POWHEGBOX + PYTHIA8		All
ISR	Varying α_s^{ISR} (PS), μ_r & μ_f (ME)	in POWHEGBOXRES + PYTHIA8	$t\bar{t} + \geq 1b$
		in POWHEGBOX + PYTHIA8	$t\bar{t} + \geq 1c$, $t\bar{t}$ + light
FSR	Varying α_s^{FSR} (PS)	in POWHEGBOXRES + PYTHIA8	$t\bar{t} + \geq 1b$
		in POWHEGBOX + PYTHIA8	$t\bar{t} + \geq 1c$, $t\bar{t}$ + light
$t\bar{t} + \geq 1b$ fractions	POWHEGBOX + HERWIG7 vs POWHEGBOX + PYTHIA8		$t\bar{t} + 1b$, $t\bar{t} + \geq 2b$
$p_T^{b\bar{b}}$ shape	Shape mismodelling measured from data		$t\bar{t} + \geq 1b$

Table 7.3: Overview of the sources of systematic uncertainty for $t\bar{t}$+jets modelling. The systematic uncertainties listed in the first section of the table are normalisation and cross-section uncertainties. The uncertainties in the second section are evaluated in such a way as to have no impact on the normalisation of the three $t\bar{t} + \geq 1b$, $t\bar{t} + \geq 1c$, and $t\bar{t}$+light components in the phase-space selected in this analysis. The third section lists uncertainties specifically assigned to the $t\bar{t} + \geq 1b$ mis-modelling effects. The last column of the table indicates the $t\bar{t}$+jets components to which a systematic uncertainty is assigned. All systematic uncertainty sources are treated as uncorrelated across the three components.

do not affect the results of the fit. Nonetheless, if their normalisation effect is large, these systematics will be highly correlated to the $k(t\bar{t} + \geq 1b)$ factor, making it difficult to disentangle their effects in the fit. To avoid this correlation, the fraction of $t\bar{t} + \geq 1b$ events in the selected phase-space in all alternative samples is reweighted to match the fraction in the nominal sample. This allows the normalisation of $t\bar{t} + \geq 1b$ to be driven solely by the free-floating parameter in the fit to data. Eventually, the $t\bar{t} + \geq 1b$ systematic variations are renormalised to not change the normalisation inclusively across all analysis regions, though they can have a normalisation effect when used in only one of the channels and in individual regions. This applies to all $t\bar{t} + \geq 1b$ systematics considered in the analysis.

Moreover, the uncertainties associated to the modelling of $t\bar{t} + \geq 1b$, $t\bar{t} + \geq 1c$, and $t\bar{t}$ + light by the respective nominal prediction (defined in Sec. 4.5.4) result in 25 independent sources (17 for $t\bar{t} + \geq 1b$ and 4 for each of the others). The definition of these systematic uncertainties are motivated by the need to distinguish as much as possible different effects in the modelling, while comparing, for each component different MC setups with the same process generated in the ME, computed at the same order in pQCD, and with sufficient statistics to avoid introducing unphysical shapes due to large statistical fluctuations.

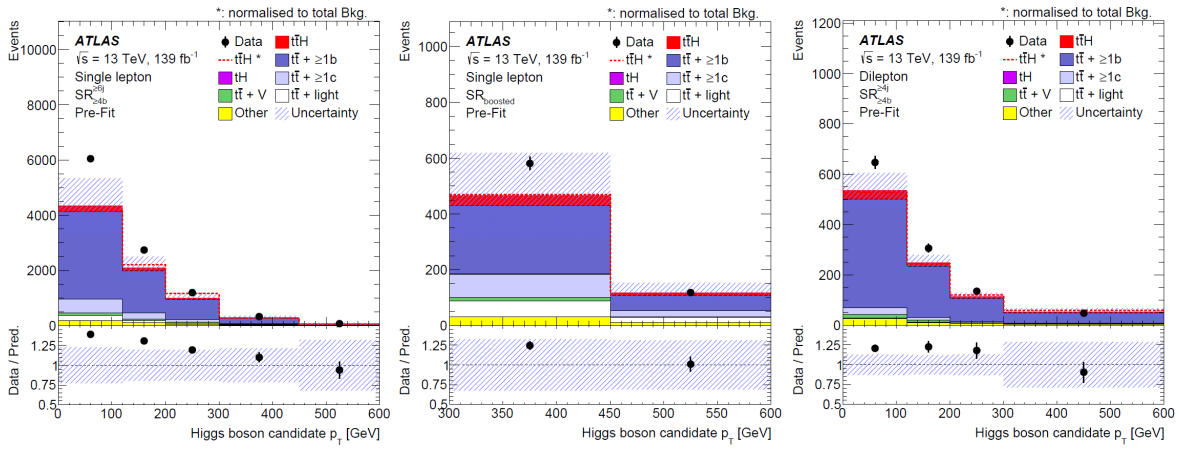
In fact, the systematic uncertainties related to varying the amount of ISR, the amount of FSR, the PS and hadronisation model, and the NLO matching procedure are estimated in the same way as for $t\bar{t}H$, comparing the nominal prediction with alternative samples. So, the variations for ISR and FSR systematic uncertainties are estimated using weights in the ME and PS of the respective nominal samples (POWHEGBOXRES+PYTHIA8 $t\bar{t}b\bar{b}$ (4FS) for $t\bar{t} + \geq 1b$ and POWHEGBOX+PYTHIA8 $t\bar{t}$ (5FS) for $t\bar{t} + \geq 1c$ and $t\bar{t}$ + light components) in a similar way as for the signal sample described above. For the determination of the two-point systematics, the $t\bar{t}$ 5FS sample is used also for the $t\bar{t} + \geq 1b$ component, as already explained in Sec.

4.5.4. Thus, the relative difference between the nominal POWHEGBOX+PYTHIA8 $t\bar{t}$ (5FS) and POWHEGBOX+HERWIG7 $t\bar{t}$ (5FS) samples is used to estimate the PS and hadronisation model uncertainty, also for the $t\bar{t}+\geq 1b$ component. Then, the relative difference is applied as a systematic variation to the nominal $t\bar{t}b\bar{b}$ (4FS) sample. Analogously, the relative difference between the nominal and MADGRAPH5_AMC@NLO+PYTHIA8 $t\bar{t}$ (5FS) samples is used to evaluate the NLO matching uncertainty. These uncertainties do not aim to cover differences between the 4FS and 5FS modelling, since it was found that for the nominal sample the 4FS represents data better than 5FS. Thus, no dedicated uncertainty accounting for this difference is used in this analysis.

Specific consideration is given to the correlation of the two-point modelling uncertainties across the different p_T^H bins, in order to provide the fit with enough flexibility to cover background mismodelling without biasing the signal extraction. The $t\bar{t}+\geq 1b$ NLO matching uncertainty shows a dependency on p_T^H, thus it is decorrelated across p_T^H bins in the SRs. Moreover, the NLO matching as well as the PS and hadronisation uncertainty are further decorrelated between the single-lepton and dilepton channels, particularly for the $t\bar{t}+\geq 1b$ process. This is done in order to avoid transferring constraints from the single-lepton resolved to the dilepton channel, since the latter is less sensitive to the high-p_T^H regime and also less additional radiation is produced there.

Furthermore, the reconstructed p_T^H is not well modelled in the different channels of the analysis, as shown in Fig. 7.1. Especially the pre-fit p_T^H distribution in the resolved lepton+jets channel (Fig. 7.1a) shows a clear slope in the data over MC prediction ratio. The effect is smaller in the single-lepton boosted and dilepton channels. Only the signal regions are considered, since they are split into p_T^H bins, whereas the control regions are inclusive in p_T^H. As a result, an additional uncertainty is evaluated for the $t\bar{t}+\geq 1b$ sample, taking into account only shape effects of this process, so as to cover the mismodelling observed in this distribution. After removing the overall normalisation difference by scaling the $t\bar{t}+\geq 1b$ background in the single-lepton $SR^{\geq 6j}_{\geq 4b}$ (dilepton $SR^{\geq 4j}_{\geq 4b}$), a weight is computed in each reconstructed p_T^H bin of the single-lepton $SR^{\geq 6j}_{\geq 4b}$ (dilepton $SR^{\geq 4j}_{\geq 4b}$). These weights correct the predicted $t\bar{t}+\geq 1b$ contribution so that the data and background model yields agree in each p_T^H bin. Then, the derived weights are not applied to the nominal sample, instead they define the one standard deviation ($+1\sigma$) variation of an additional uncertainty in the $t\bar{t}+\geq 1b$ sample in each reconstructed p_T^H bin. The weights derived from the single-lepton resolved channel are also applied in the boosted channel. Finally, this uncertainty enters the signal extraction fit as a single nuisance parameter (p_T^{bb} shape), correlated across all channels. It is constructed such that a pull of $+1\sigma$ corresponds to fully reweighting the $t\bar{t}+\geq 1b$ sample, and effectively correcting the reconstructed p_T^H spectrum.

Another issue that we had to cope with in this analysis is the fact that the predicted fraction of the $t\bar{t}+\geq 1b$ subcomponents (defined in Sec. 4.5.3) vary among the different MC generators, as shown in Table 7.4. To account for the variations in the $t\bar{t}+\geq 1b$ subcomponent fractions, an additional nuisance parameter is assigned to cover the largest discrepancy between two models for the fraction of $t\bar{t}+1b$ and $t\bar{t}+\geq 2b$. The largest difference is found between POWHEGBOXRES+PYTHIA8 $t\bar{t}b\bar{b}$ (5FS) and POWHEGBOX+HERWIG7 $t\bar{t}$. The 1σ variation of this nuisance parameter corresponds to reducing the amount of $t\bar{t}+\geq 2b$ by 13%, while increasing the amount of $t\bar{t}+1b$ by 22%. This uncertainty is correlated across all analysis regions, and impacts each region differently due to the varying compositions of $t\bar{t}+\geq 1b$. Lastly, as mentioned above, the NLO matching, and the PS and hadronisation uncertainties are



(a) **(b)** **(c)**

Fig. 7.1: Pre-fit distributions of the reconstructed Higgs boson candidate p_T^H for the (a) single-lepton resolved $SR_{\geq 4b}^{\geq 6j}$, (b) single-lepton boosted $SR_{boosted}$, and (c) dilepton $SR_{\geq 4b}^{\geq 4j}$ signal regions [100]. The $t\bar{t}H$ signal yield (solid red) is normalised to the Standard Model expectation. The dashed line shows the $t\bar{t}H$ signal distribution normalised to the total background prediction. The uncertainty band includes the statistical and systematic uncertainties, except for the uncertainty in the $k(t\bar{t}+\geq 1b)$ normalisation factor which is not defined pre-fit. The last bin includes the overflow.

derived in such a way as to have no impact on the normalisation of the $t\bar{t}+1b$ and $t\bar{t}+\geq 2b$ sub-components, in order to avoid double-counting with this additional $t\bar{t}+\geq 1b$ fractions systematic.

MC sample	yield			fraction	
	$t\bar{t}+\geq 2b$	$t\bar{t}+1b$	$t\bar{t}+\geq 1b$	$t\bar{t}+\geq 2b$	$t\bar{t}+1b$
PowhegBox+Pythia8 $t\bar{t}b\bar{b}$ (4FS)	9806.3	5750.5	15556.8	0.630	0.370
PowhegBox+Pythia8 $t\bar{t}$ (5FS)	9206.6	6361.0	15567.6	0.591	0.409
MadGraph5_aMC@NLO+Pythia8 $t\bar{t}$ (5FS)	9243.0	6225.5	15468.5	0.598	0.402
PowhegBox+Herwig7 $t\bar{t}$ (5FS)	6280.8	5206.7	11487.5	0.547	0.453

Table 7.4: Expected yields for $t\bar{t}+\geq 1b$ and its subcomponents $t\bar{t}+1b$ and $t\bar{t}+\geq 2b$ in the single-lepton and dilepton channels for the different $t\bar{t}+\geq 1b$ MC models considered in this analysis. The relative $t\bar{t}+1b$ and $t\bar{t}+\geq 2b$ fractions are also shown.

Non-$t\bar{t}$ simulated background processes such as single top, $t\bar{t}W/t\bar{t}Z$, W/Z+jets, diboson, and others represent a minor fraction of the total background. Consequently, a less refined treatment of the uncertainties associated with these small backgrounds is adopted, since they have a subordinate effect on the sensitivity of the analysis.

An uncertainty of 5% is considered for the cross-sections of each of the three single-top production modes [266–268]. Uncertainties associated with the PS and hadronisation model as well as the NLO matching scheme are evaluated by comparing, for each process, the nominal PowhegBox+Pythia8 sample with a sample produced using PowhegBox+Herwig7 and MadGraph5_aMC@NLO+Pythia8, respectively (see Sec. 4.5.5). In addition, an uncertainty

associated with the interference between *tW*, and $t\bar{t}$ production at NLO [269] is assessed by comparing the nominal POWHEGBOX+PYTHIA8 sample, produced using the diagram removal scheme, with an alternative sample produced with the same generator but using the diagram subtraction scheme.

Furthermore, the theoretical uncertainty on the $t\bar{t}V$ NLO cross-section prediction is 15% [270], split into PDF and scale uncertainties as for $t\bar{t}H$. An additional $t\bar{t}V$ modelling uncertainty, related to the choice of PS and hadronisation model and NLO matching scheme is assessed by comparing the nominal MADGRAPH5_AMC@NLO+PYTHIA8 samples with alternative ones generated with SHERPA (see Sec. 4.5.6). Also, a 50% normalisation uncertainty is assumed for the $t\bar{t}t\bar{t}$ background, covering effects from varying the renormalisation μ_R and factorisation μ_F scales, the PDFs and α_s [271].

The small backgrounds from *tZq* and *tWZ* are each assigned cross-section uncertainties. In particular, for *tZq* two uncertainties are used, 7.9% accounting for renormalisation and factorisation scale variations and 0.9% accounting for PDFs, whereas for *tWZ* a single uncertainty of 50% is used [272]. Aslo, uncertainties in the associated production of a single top-quark and a Higgs boson include μ_R and μ_F scale variations as well as PDF uncertainties: they amount to +6.5/-14.9% (+6.5/-6.7%) and $\pm 3.7\%$ ($\pm 6.3\%$) for *tHjb* (*tWH*), respectively [106].

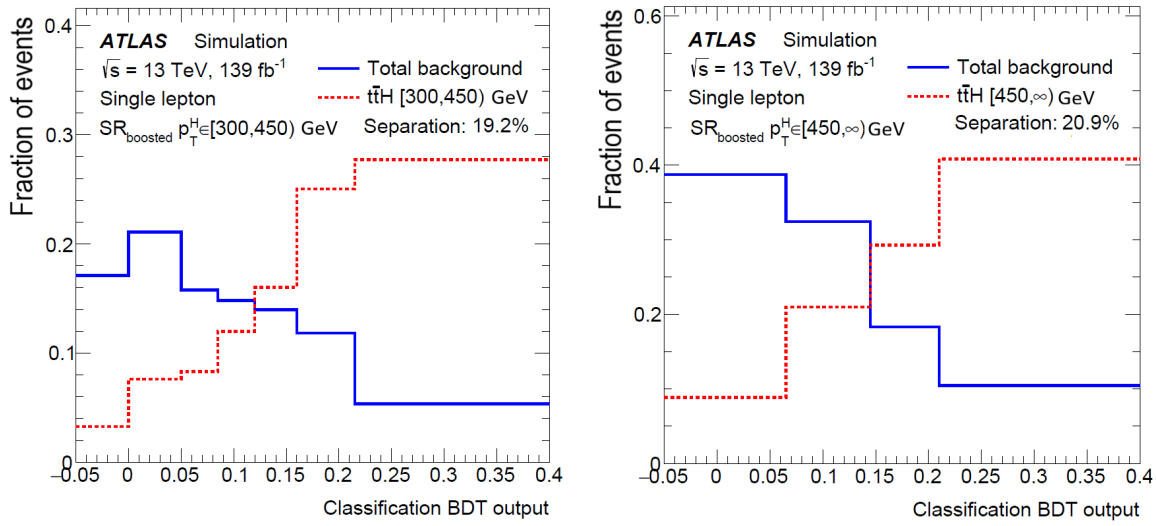
Moreover, an uncertainty of 40% is evaluated for the *W*+jets cross section, with an additional 30% normalisation uncertainty used for *W*+heavy-flavour jets, taken as uncorrelated between events with at least two heavy-flavour jets. These uncertainties are based on variations of the μ_R and μ_F scales and of the Sherpa matching parameters (cf. Sec. 4.5.7). Additionally, an uncertainty of 35% is applied to the *Z*+jets normalisation, uncorrelated across jet bins. This accounts for both the variations of the scales and SHERPA matching parameters as well as the uncertainty in the extraction from data of the correction factor for the heavy-flavour component. Finally, a 50% normalisation uncertainty in the diboson background is considered, which includes uncertainties in the inclusive cross-section and additional jet production [272].

7.2 Fit input preparation for the statistical analysis

In order to develop the statistical model (described in Ch. 8), Monte Carlo discriminant distributions (e.g. the classification BDT output in the single-lepton boosted region) are used to construct templates for the $t\bar{t}H$ signal and each of the backgrounds, which are then used as inputs to the profile likelihood fit. An important step in the development of the fit model is the choice of the binning of these templates. Its definition is crucial since it affects the sensitivity and the impact of the statistical fluctuations. Another crucial point is the practical implementation of the systematic uncertainties on the analysis. A symmetrisation procedure as well as smoothing techniques are employed to mitigate the impact of statistical fluctuations in the MC samples used to estimate the systematic uncertainties.

7.2.1 Binning

The binning of the discriminant distributions is of particular importance for the construction of the histograms for the likelihood fit. Fewer bins, hence larger bins, means larger statistics in each bin, but the trade-off is usually a lower control over the backgrounds when the systematic uncertainties are large, leading to a lower sensitivity. On the contrary, splitting these distributions in more bins would give more hints about the performance of the fit model, revealing



(a) **(b)**

Fig. 7.2: Classification BDT response in the boosted signal region ($SR_{boosted}$) in the $p_T^H \in$ (a) $[300, 450)$ GeV and (b) $[450, \infty)$ GeV regions for the $t\bar{t}H$ signal (red) with $\hat{p}_T^H \in [300, 450)$ and $[450, \infty)$ GeV, respectively, and background (blue) samples, showing the binning in each region. Only the MC simulation is shown in this distribution and the events in each bin are normalised to the total expected signal and background events, respectively. Also, the relative separation power (eq. 6.3) between signal and background events is written.

in advance any potential shape or mismodelling between the data and the simulated events. Especially in the signal regions, where the discriminant is the classification BDT output, the binning in the higher BDT values plays a determining role in the final sensitivity, since most of the signal events lie there.

The binning of the fit input distribution has been optimised independently for each analysis region in each channel. The ultimate goal is to maximise the analysis sensitivity, while keeping enough MC statistics in each bin to avoid fluctuations in the predicted yields of the nominal as well as of the systematics model. Eventually, in the single-lepton boosted channel seven (four) bins in the $SR_{boosted}$ with $p_T^H \in [300, 450)$ ($[450, \infty)$) GeV, depicted in Fig. 7.2, give the best compromise between signal sensitivity and MC statistics, according to fits with an Asimov data-set (explained in Sec. 8.3.1). Less bins have been chosen in the high p_T^H region to reduce fluctuations, due to the much lower statistics.

For completeness, Table 7.5 summarises the discriminating variables used in the fit for each analysis region (presented in Sec. 8.2) as well as the number of bins used to define the template histograms of these observables. The concept of the "blinded bins" is explained in Sec. 8.3.2. In the analysis regions where only the event yield is used as input to the fit, there is no shape to be considered, thus only one bin is defined. The usage of only one bin in these categories allows to reduce the pulls on the fit to data, due to the assumption of fully correlated systematic variations among all regions, and the tensions between the different categories. On the other hand, this induces a loss in sensitivity due to the lower constraints on the nuisance parameters. These parameters are mainly associated with the $t\bar{t}+\geq 1c$ modelling, since the event yield is fitted mostly in the dilepton CRs which are used to control the $t\bar{t}+\geq 1c$ background.

Region	reco p_T^H [GeV]	Fitted observable	# bins (# blinded bins)
$SR^{\geq 4j}_{\geq 4b}$	$[0, 120)$	Classification BDT	4 (2)
	$[120, 200)$	Classification BDT	3 (2)
	$[200, 300)$	Classification BDT	2 (1)
	$[300, \infty)$	Classification BDT	2 (1)
$CR^{\geq 4j}_{3b \text{ hi}}$		Event yield	1 (0)
$CR^{\geq 4j}_{3b \text{ lo}}$		Event yield	1 (0)
$CR^{3j}_{3b \text{ hi}}$		Event yield	1 (0)
$SR^{\geq 6j}_{\geq 4b}$	$[0, 120)$	Classification BDT	4 (2)
	$[120, 200)$	Classification BDT	4 (2)
	$[200, 300)$	Classification BDT	4 (2)
	$[300, 450)$	Classification BDT	4 (2)
	$[450, \infty)$	Event yield	1 (0)
$CR^{5j}_{\geq 4b \text{ hi}}$		$\Delta R^a_{bb} vg$	6 (2)
$CR^{5j}_{\geq 4b \text{ lo}}$		$\Delta R^a_{bb} vg$	6 (0)
$SR_{boosted}$	$[300, 450)$	Classification BDT	7 (3)
	$[450, \infty)$	Classification BDT	4 (2)

Table 7.5: Overview of the discriminating variables used in the fit for each analysis region, split in the respective p_T^H bins, and of the number of bins of the fitted distributions.

7.2.2 Symmetrisation

Each systematic uncertainty is evaluated by varying the corresponding discriminant distribution by one standard deviation (σ) and reweighting accordingly all the events. It can either increase or decrease the yield, denoted by an "up" or a "down" label, arbitrarily. This leads to one or two shifted distributions for the discriminant variable representing the $\pm 1\sigma$ variations with respect to the nominal distribution. However, in many cases the direction of the systematic changes for different values of an observable. In order to further reduce the impact of potential statistical fluctuations on the calculation of systematic uncertainties, which may then propagate to the fit, a symmetrisation procedure is applied.

Systematic uncertainties for which the 1σ variation is available only in one direction, by convention the up variation, are called *one-sided systematics*. This is mainly the case for systematic uncertainties arising from comparing two MC samples, namely the two-point systematics so for the signal as for the background processes. For one-sided uncertainties the symmetrisation provides the down variation as the symmetric of the up variation around the nominal prediction.

In contrast, uncertainties with both the up and down variations provided, are called *two-sided systematics*. Even if the systematic uncertainty is expected to be symmetric, the discriminant distribution may not necessarily reflect that because of statistical fluctuations. Therefore, although both variations are provided, a new variation is calculated as the mean difference between the up and down variations and it is used to re-define the up variation. Afterwards, the one-sided symmetrisation is applied accordingly to define the down variation. This re-

duces the impact of statistical fluctuations without changing the underlying systematic. This symmetrisation method is applied to the experimental uncertainties as well as to the signal and background modelling uncertainties related to the ISR/FSR.

7.2.3 Smoothing

While constructing the systematics model, it can happen that some uncertainties are derived from the comparison of several MC simulations with a limited amount of generated events. Such distributions of the uncertainties can be used in the fit, however, some statistical fluctuations of the systematic variations could possibly match fluctuations in the data, which may have a large impact on the fit result. A way to confine this effect is to introduce statistical uncertainties on the systematics. Then, the model would become much more complicated, while the derivation of such statistical uncertainties is not straightforward for systematic uncertainties which are correlated to the nominal sample.

In order to cope with the statistical fluctuations in the alternative MC samples, that define the systematic variations, smoothing algorithms are applied to the templates used prior to the fit. In fact, the smoothing procedure can significantly reduce the impact of the statistical fluctuations, but it will not eradicate it. Besides, the smoothing of systematic uncertainties is essential in order to reduce the CPU time for fitting, as well as to avoid the problems with the convergence of the minimisation. The impact of the smoothing applied to the templates used to derive systematics, was studied thoroughly in terms of the $t\bar{t}H(H \to b\bar{b})$ analysis and it was found that it does not bias the result [273].

Several smoothing methods [274] are used in the analysis, which preserve the effect on normalisation due to the systematic uncertainty. They are based on merging of bins and subsequent smoothing of the shape, whereas the binning of the nominal distribution remains unchanged. The first method is called PARABOLIC according to which the bins with a large statistical uncertainty ($> 5\%$) are merged with their left bin (starting with the last bin). The whole distribution is smoothed by a running average, keeping the first and the last bins unchanged. The other method is called MAXVARIATION, and it is based on the merging of close bins with a large statistical uncertainty until the statistical uncertainty of the combined bins is lower than a fixed value. Compared to the PARABOLIC method, the MAXVARIATION gives smaller importance to the first and last bins of the distribution, leading to undesirable results since the smoothing algorithm does not capture the shape properly.

According to the above, the MAXVARIATION fails to capture the first and the last bins of the large systematic variations. Since the analysis uses a classification discriminant for the fitting, the last bin is the most sensitive one to the signal and its bad modelling would have a significant impact on the result of the analysis. Especially the major modelling systematic uncertainties on the signal as well as on the main $t\bar{t}$+jets background have much larger shape variations, and as a result a larger impact on the result. Therefore, it has been found that only the PARABOLIC algorithm is able to perform well enough in capturing their shape, without underestimating these systematic variations significantly.

On the contrary, it was observed that the PARABOLIC smoothing introduces a shape for the small systematic variations due to a fluctuation in one bin. Given the large statistical uncertainties of the underlying distribution, avoiding artificial shapes and instead introducing a flat shape, produced by the MAXVARIATION algorithm, seems as a more reasonable approximation. As a consequence, the latter is exploited for the systematic uncertainties of which the difference between bins is comparable to their statistical uncertainties. These are all the in-

strumental detector systematics as well as the modelling systematic uncertainties on the small backgrounds (all backgrounds except for the $t\bar{t}$+jets components). Although this algorithm works better for samples with lower statistics, it still does not fully capture the shape of a distribution. The smoothed distributions follow the shape of the original distribution within the uncertainties, though. Since these systematics do not have such a large impact on the analysis result, this is a reasonable compromise.

7.2.4 Shape of major systematic uncertainties

As described earlier, the various sources of systematic uncertainties on expected signal and background contributions originate from the modelling of the signal and the different background processes as well as from the various objects reconstructed and identified by the detector. In particular, the modelling of the $t\bar{t} + \geq 1b$ process plays a decisive role being the main limiting factor of the analysis result. Indicatively, the variations of the $t\bar{t} + \geq 1b$ modelling uncertainties on the final discriminant used in the signal extraction fit in the single-lepton boosted region are presented in the following. Examples of variations on the signal modelling or experimental uncertainties, that have much smaller impact on the measurement, demonstrating comparable or smaller shape and/or normalisation effects to those described in the following, can be found in App. A.5.

The ISR uncertainty on the $t\bar{t} + \geq 1b$ prediction is illustrated Figs. 7.3a-7.3b. An asymmetry is discernible between the up and down components of the original distributions, before the application of the smoothing and symmetrisation. The effect of a potential symmetrisation and/or smoothing was studied in the full single-lepton channel, where the statistics is high enough so as to not deem these techniques necessary. Eventually, their effect on the fit result was found to be negligible [273], thus they are both applied to reduce the impact of statistical fluctuations on the systematic variations, especially in the dilepton channel where the statistical fluctuations are higher. Apart from this, the shape of the systematic seems to be relatively flat. As already described, the ISR systematic represents variations in the production of additional jets in both the matrix element and the parton shower. This implies that it could significantly affect the jet multiplicity, which is indeed the case and is highlighted in Sec. 8.4.1.

On the contrary, the $t\bar{t} + \geq 1b$ FSR uncertainty, depicted in Figs. 7.3c-7.3d, has large statistical fluctuations in most bins of the classification BDT distribution. This is due to large variations of the weights used to derive this uncertainty. The impact of these fluctuations on the result of the fit was tested using Monte Carlo toys and was found to be negligible [273]. Also, the FSR as the ISR systematic shows a significant dependence on the BDT variable in the low p_T^H region, while they are almost flat in the high p_T^H region.

Then, the two-point systematics for the $t\bar{t} + \geq 1b$ background are shown in Fig. 7.4. The NLO generator matching systematic (Figs. 7.4a-7.4b) develops a large shape effect with difference up to 20% in the high p_T^H region between the first and the last bin. It will be shown later that this systematic is strongly correlated to the signal normalisation, making it a limiting factor of the analysis result. Additionally, the parton shower and hadronisation (PS & had.) systematic, displayed in Figs. 7.4c-7.4d, is relatively flat in the BDT distribution. In both systematics, the shape of the distribution is well captured from the smoothing technique. Also, they both have quite a significant normalisation effect in the high p_T^H region.

Finally, the $t\bar{t} + \geq 1b$ p_T^{bb} shape and fraction uncertainties, are displayed in Fig. 7.5. The shape of the distributions is well captured from the smoothing technique in both systematics. Both develop a shape effect with difference up to 15% in both p_T^H regions between the first

and the last bin, while the p_T^{bb} uncertainty shape has also a normalisation effect, especially in the high p_T^H region.

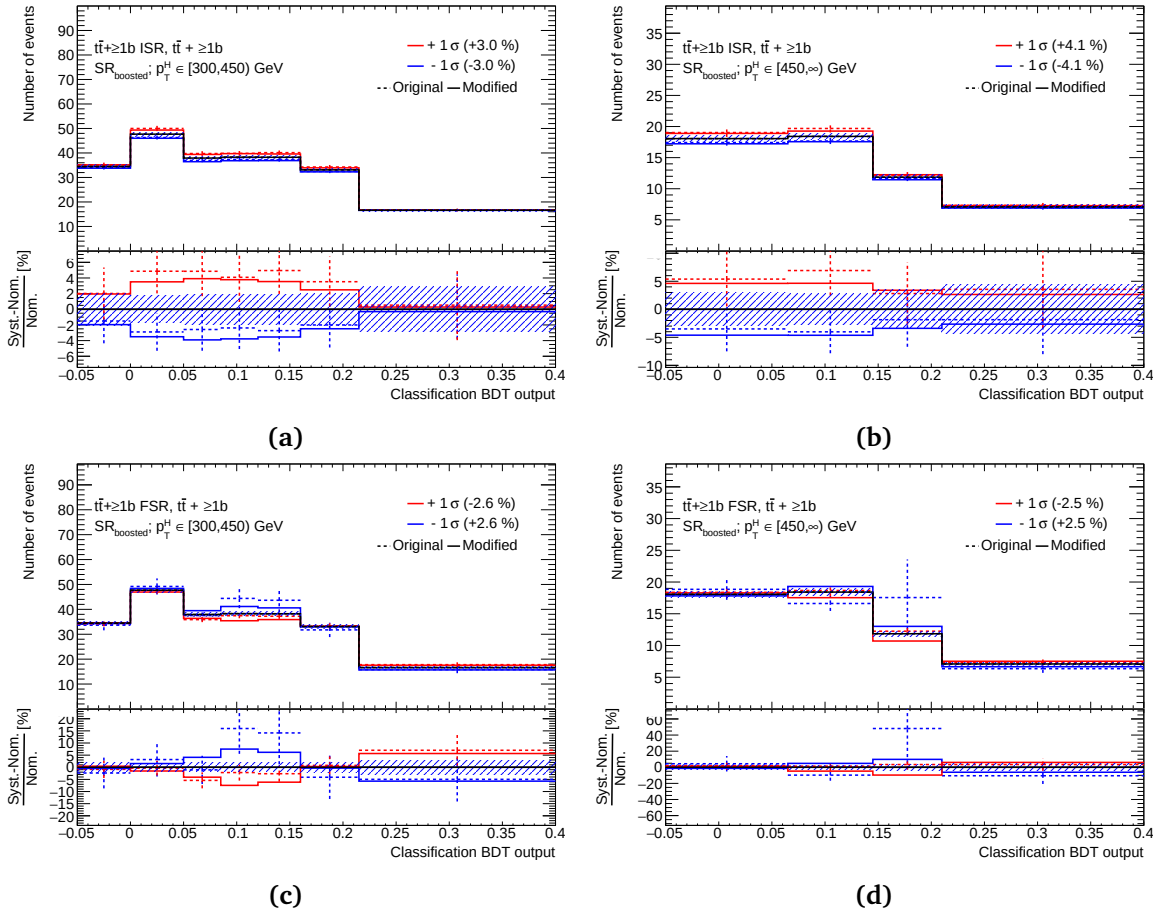


Fig. 7.3: Comparison of the nominal $t\bar{t} + \geq 1b$ prediction (black) with the one standard deviation up (red) and down (blue) variations induced by the (a),(b) $t\bar{t} + \geq 1b$ ISR and (c),(d) $t\bar{t} + \geq 1b$ FSR uncertainties for the classification BDT distribution in the single-lepton boosted signal region ($SR_{boosted}$) in the $[300, 450)$ GeV (left) and $[450, \infty)$ GeV (right) p_T^H bins. *Original* (dashed line) refers to the raw input distribution, while *modified* (solid line) is the distribution after symmetrisation and smoothing.

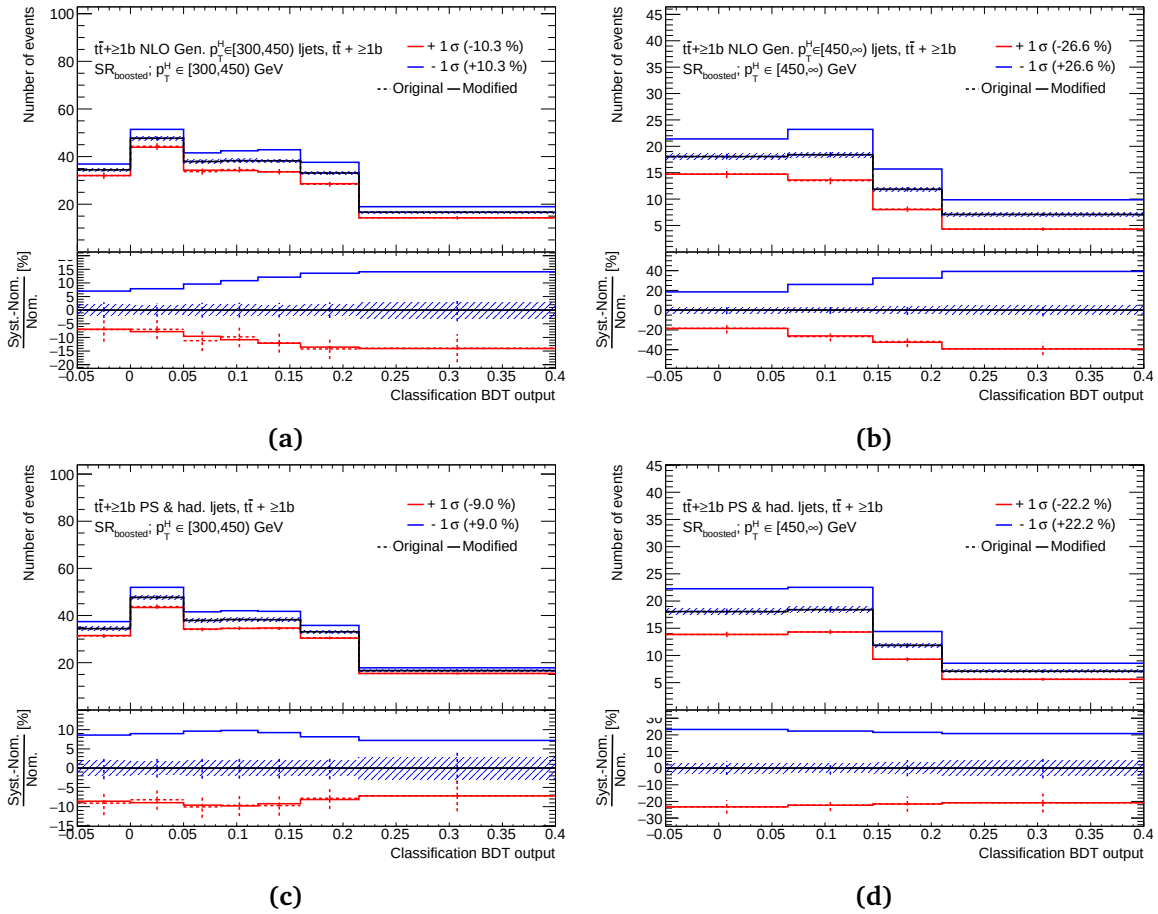


Fig. 7.4: Comparison of the nominal $t\bar{t} + \geq 1b$ prediction (black) with the one standard deviation up (red) and down (blue) variations induced by the (a),(b) $t\bar{t} + \geq 1b$ NLO matching (c),(d) $t\bar{t} + \geq 1b$ PS & hadronisation uncertainties for the classification BDT distribution in the single-lepton boosted signal region ($SR_{boosted}$) in the $[300, 450)$ GeV (left) and $[450, \infty)$ GeV (right) p_T^H bins. *Original* (dashed line) refers to the raw input distribution, while *modified* (solid line) is the distribution after symmetrisation and smoothing.

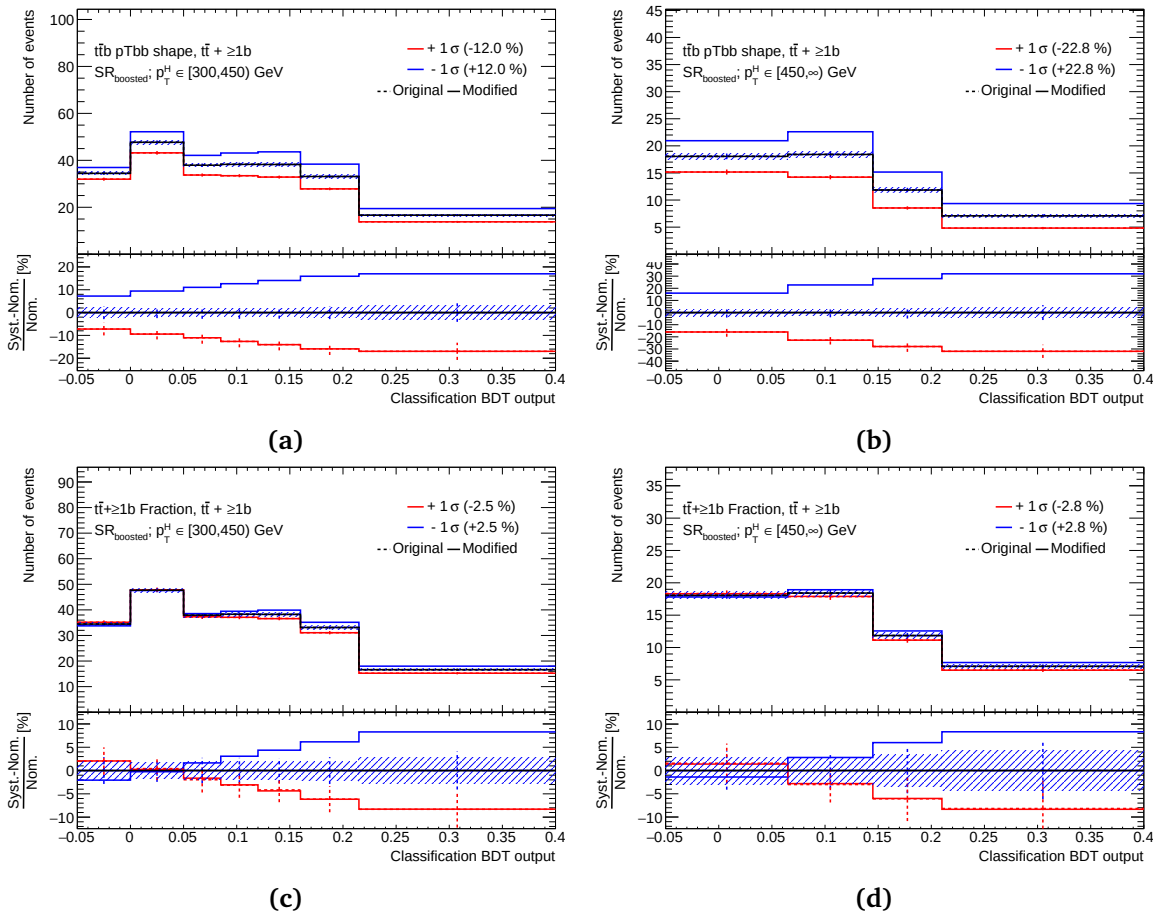


Fig. 7.5: Comparison of the nominal $t\bar{t} + \geq 1b$ prediction (black) with the one standard deviation up (red) and down (blue) variations induced by the (a),(b) $t\bar{t} + \geq 1b$ p_T^{bb} shape and (c),(d) $t\bar{t} + \geq 1b$ fraction uncertainties for the classification BDT distribution in the single-lepton boosted signal region ($SR_{boosted}$) in the $[300, 450)$ GeV (left) and $[450, \infty)$ GeV (right) p_T^H bins. *Original* (dashed line) refers to the raw input distribution, while *modified* (solid line) is the distribution after symmetrisation and smoothing.

Chapter 8

Statistical Analysis and Results

In order to extract the $t\bar{t}H$ signal cross-section from the data, a complex statistical analysis is needed, given that the signal cross-section is very small with respect to the background one. The signal-enriched regions are analysed together with the signal-depleted ones in a template profile likelihood fit, combining the single-lepton resolved and boosted as well as dilepton channels. For each discriminant variable used as input to the fit, template distributions for the signal and each of the backgrounds, built from the MC prediction, are compared to data. All systematic uncertainties (discussed in Sec. 7.1) are taken into account in the fit. This signal extraction fit simultaneously determines the event yields for the signal and the most important background component, while constraining the normalisation and shape of the differential distributions of the backgrounds within the assigned systematic uncertainties.

Distinctive fits for the inclusive and differential cross-section measurements are performed (the latter with the STXS formalism), both using the same strategy. The only difference is that in the latter, the signal is split into five templates according to the true Higgs transverse momentum, $\hat{p}_T^H$, and for every signal template a separate signal normalisation factor is obtained. The performance of the analysis and the validation of the fit model are evaluated from a fit to an Asimov dataset [275]. Additionally, in order to optimise the fit model without biasing the $t\bar{t}H$ signal sensitivity towards a certain result, a background-only fit is performed. In this fit, the data in the bins of the discriminant distributions sensitive to signal events are kept blinded and are removed from the fit. Finally, together with the final fit to the full data, the signal significance and the upper limit of the $t\bar{t}H$ production cross-section are determined.

This $t\bar{t}H(H \rightarrow b\bar{b})$ measurement [100] is performed with increased data statistics compared to the previous measurement [97], which was performed on a subset of the Run 2 dataset. This benefits mostly the high-p_T region, since in the low-p_T one the precision of the analysis was already limited by systematic uncertainties. Nonetheless, more MC simulated events have been also produced, leading to increased statistics of the simulated samples for better estimation of the nominal background, as well as for an improved assessment of its systematic uncertainties. Then, the improved reconstruction algorithms and detector calibrations, in turn enhance the performance of the b-tagging algorithm. This allows for the use of tighter selection criteria, rejecting events in poorly modelled regions of phase space. Moreover, due to the better understanding of the detector and of the reconstruction, smaller instrumental uncertainties are expected. Their improved performance implies that more effort has been committed to the investigation of the background modelling. In addition, an enhanced model is adopted for the $t\bar{t}$+jets background with a new nominal generator for the $t\bar{t} + b\bar{b}$ process, which contains the two *b*-quarks directly in the matrix elements (described in Sec. 4.5.4).

Finally, the analysis benefits from the improved boosted region definition which leads to an increased sensitivity, especially in the high-p_T regime.

8.1 Profile likelihood fit

In general, in order to measure a specific physics process, the amount of the measured signal events needs to be extracted from the data. Thus, the *signal strength*, μ, is measured, defined as the ratio between the measured cross-section and the SM prediction. Measuring the signal strength instead of the cross section of a process is preferred, since a direct comparison with the SM expectation and with other analyses is feasible. For the $t\bar{t}H(H \to b\bar{b})$ analysis, the signal strength of the $t\bar{t}H$ process, $\mu_{t\bar{t}H}$, is expressed by

$$\mu_{t\bar{t}H} = \frac{\sigma_{t\bar{t}H}}{\sigma_{t\bar{t}H}^{SM}}. \tag{8.1}$$

Ideally for a SM Higgs boson, one expects to measure $\mu_{t\bar{t}H} = 1.0$, while a $\mu_{t\bar{t}H}$ value not compatible with 1 within the uncertainties could imply a deviation from the SM indicating new physics. Eventually, any significant positive value of the signal strength denotes the presence of a signal, here the $t\bar{t}H$ signal, whereas $\mu = 0$ means its absence. The latter represents the so-called *background-only hypothesis*. Moreover, the true value of the signal strength cannot be negative, though the measured value can. However, a negative value not compatible with zero within the uncertainties would constitute evidence against the background-only model, or that the amount of the events predicted by the background processes has been overestimated.

Even though the $t\bar{t}H$ production and the $H \to b\bar{b}$ decay channel are already independently discovered [102–105], specifically the $t\bar{t}H(H \to b\bar{b})$ channel has not been observed, yet. For its measurement, a complex fit model is required, following a statistical method [275] that is based on the Neyman-Pearson lemma [276]. According to this, for the discovery of a new signal process, a hypothesis of null signal, i.e. the *null hypothesis* (H_0), is defined, describing only known processes, here designated as background. This is then tested against an *alternative hypotheses* (H_1), which includes both background and the sought-after signal. In order to reject hypothesis H_0 in favour of hypothesis H_1 the most powerful test statistic is the ratio of their likelihoods. Therefore, for the analysis presented here, a template profile likelihood fit is exploited to extract the signal strength, incorporating the predicted yields and uncertainties in every bin of the analysis regions to fit them to data.

For each event in a sample, the discriminant variables are measured and these values are used to construct the relevant histograms. Then, the distributions of the discriminants from each of the analysis channels and regions are combined in the profile likelihood fit to test the presence of a $t\bar{t}H$ signal, assuming a Higgs boson mass of m_H = 125 GeV, and to constrain the backgrounds.

For each bin i of the input distribution of each analysis region r, the number of data events $n_{r,i}$, which is assumed to be Poisson distributed, is compared to the theoretical prediction given

by the expectation value of the bin content

$$E[n_{r,i}(\mu_1,...,\mu_j,k_1,...,k_m,\theta_1,...,\theta_{l_x})] = \sum_{s\in \text{sig}} \mu_s \cdot n^{exp}_{r,i,s}(\theta_1,...,\theta_{l_s}) + \sum_{b\in \text{bkg}} k_b \cdot n^{exp}_{r,i,b}(\theta_1,...,\theta_{l_b}) + \sum_{c\in \text{bkg}} n^{exp}_{r,i,c}(\theta_1,...,\theta_{l_c}), \tag{8.2}$$

where μ_s is the signal strength on the signal template s and j is the number of signal templates. Also, k_b is the normalisation factor (k-factor) on each background b, m is the number of backgrounds, while $c \neq b$ refers to each background to which no normalisation factor is assigned. Lastly, $\theta_1,...,\theta_{l_x}$ is a set of $L_x = 1,..,l_x$ additional parameters that affect the sample x being signal ("sig") or background ("bkg"). The signal strengths (μ_s) are also called *parameters of interest* (POIs). The remaining parameters, θ_{L_x}, are called *nuisance parameters* (NPs), since their value is not considered a generally valid result.

Template distributions are formed from the expected discriminant distributions on which the nuisance parameters are varied. The POIs and k-factors act only on the normalisation of the signal and background template distributions, respectively. The NPs encode the effects of systematic uncertainties on the signal and background expectations. For every systematic variation listed in Table 7.1, there is a NP, θ_{L_x}, that modifies the shape and/or normalisation of the templates, such as in Figs. 7.4a-7.4b, according to the systematic uncertainty it parametrises. Nuisance parameters and normalisation factors are assigned to each template, providing the degrees of freedom that the fit uses to correct the predicted templates and match the data.

The statistical analysis is based on a binned likelihood function which depends on the parameters of interest and the nuisance parameters, $\mathcal{L}(\boldsymbol{\mu},\boldsymbol{k},\boldsymbol{\theta})$. Since the data content in each bin is expected to follow a Poisson probability, the total probability of observing the given data is constructed as a product of Poisson distribution terms over all bins considered in the analysis. This results in the binned likelihood function

$$\mathcal{L}_{\text{Poisson}}(\boldsymbol{\mu},\boldsymbol{k},\boldsymbol{\theta}) = \prod_{r\in \text{regs}} \prod_{i\in \text{bins}} \frac{E[n_{r,i}(\boldsymbol{\mu},\boldsymbol{k},\boldsymbol{\theta})]^{n_{r,i}}}{n_{r,i}!} e^{-E[n_{r,i}(\boldsymbol{\mu},\boldsymbol{k},\boldsymbol{\theta})]}. \tag{8.3}$$

The signal strength parameters and the normalisation factors are regarded as unconstrained, assuming no prior knowledge from theory or subsidiary measurements. Hence, they are referred to as *free-floating* parameters.

On the contrary, the nuisance parameters are implemented in the likelihood function using Gaussian constraints reflecting the prior knowledge of the systematic uncertainty, also called *penalty terms*. As a result, the binned likelihood function is given by the one in eq. 8.3 times a product of Gaussian probabilities for each NP p ($\mathcal{L}^{\text{NP}}_{\text{Gaussian}}$)

$$\mathcal{L}(\boldsymbol{\mu},\boldsymbol{k},\boldsymbol{\theta}) = \mathcal{L}_{\text{Poisson}}(\boldsymbol{\mu},\boldsymbol{k},\boldsymbol{\theta}) \cdot \prod_{p=1}^{l} \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta_p^2}{2}}. \tag{8.4}$$

The value $\theta = 0$ corresponds by construction to the best knowledge of a specific NP (*nominal value*) indicating no correction to the relevant systematic uncertainty, while a deviation by $\theta = \pm 1$ shifts its distribution by $\pm 1\sigma$. After the fit to data, the shift of the nuisance parameters, as well as of the normalisation factors and signal strengths, by an amount of standard deviations

from their nominal values is referred to as a *pull*. Also, the uncertainty on the θ value, which is 1σ before the fit, is usually either not affected or reduced (*constrained*) by the fit to data.

However, the likelihood function in eq. 8.4 does not take into account the statistical uncertainty arising from the limited number of simulated events. So, the statistical uncertainty in the prediction is incorporated [277, 278] in the likelihood as Poisson prior ($\mathcal{L}_{\text{Poisson}}^{\text{MC stat}}$) in the form of additional NPs, $\gamma_{r,i}$, one for each of the considered bins of each analysis region. They have a multiplicative effect on the expected background events in each bin i of each analysis region r, thus a $\gamma_{r,i}$ factor is multiplied in the two background terms in eq. 8.2. Given a background estimation $n_{r,i,bkg}^{exp}$ with a statistical uncertainty $\Delta n_{r,i,bkg}^{exp}$, the relative statistical uncertainty is defined as $\Delta n_{r,i,bkg}^{exp}/n_{r,i,bkg}^{exp}$. This MC estimation corresponds to an auxiliary measurement of $\tau_{r,i} = (n_{r,i,bkg}^{exp}/\Delta n_{r,i,bkg}^{exp})^2$ background events in each bin following a Poisson distribution, where $\tau_{r,i}$ events would fluctuate around the mean $\gamma_{r,i}\tau_{r,i}$, if a new MC samples was generated. Eventually, this Poisson constraint ($\mathcal{L}_{\text{Poisson}}^{\text{MC stat}}$) is multiplied to the likelihood function in eq. 8.4

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{k}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \mathcal{L}_{\text{Poisson}}(\boldsymbol{\mu}, \boldsymbol{k}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \cdot \mathcal{L}_{\text{Gaussian}}^{\text{NP}} \cdot \prod_{r \in \text{regs}} \prod_{i \in \text{bins}} \frac{(\gamma_{r,i}\tau_{r,i})^{\tau_{r,i}}}{\tau_{r,i}!} e^{-(\gamma_{r,i}\tau_{r,i})}. \tag{8.5}$$

The nominal value of these parameters is $\gamma = 1$ by construction, so that their best-fit value corresponds to their value before the fit, while a value of $\gamma = 0$ would scale the MC yield of the corresponding bin to 0.

The best estimate, i.e. the best-fit value of the measurement, for the parameter set $(\boldsymbol{\mu}, \boldsymbol{k}, \boldsymbol{\theta})$ is obtained by maximising the likelihood function, or alternatively by minimising the negative log-likelihood (-log $\mathcal{L}$), with the latter being numerically more stable. The results presented in this thesis are obtained using the minimisation procedure as implemented in the *Minuit2* package [279] of the *RooFit* framework [280, 281], implemented in ROOT, a C++ based framework for data analysis [243]. The template distributions before and after performing the fit are referred to as "pre-fit" and "post-fit", respectively.

In order to test hypothetical values of μ in a particle physics analysis, and as a result to establish a discovery, a significance test is exploited. However, as explained above, the signal and background models also contain nuisance parameters, whose values are not taken as known a priori but rather must be fitted from the data. Eventually, the dependence on these parameters can be approximately confined by the use of the profile likelihood, and the dependence is removed in the limit where the data sample is very large. The *profile likelihood* [101] depends only on the parameters of interest $\boldsymbol{\mu}$ and is defined as

$$\mathcal{L}_P(\boldsymbol{\mu}) = \mathcal{L}(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{k}}}, \hat{\hat{\boldsymbol{\theta}}}), \tag{8.6}$$

where $\hat{\hat{\boldsymbol{\theta}}}$ and $\hat{\hat{\boldsymbol{k}}}$ denote the values of $\boldsymbol{\theta}$ and $\boldsymbol{k}$ that maximise the likelihood for the specified $\boldsymbol{\mu}$, namely they are the conditional maximum-likelihood estimators (MLE) of $\boldsymbol{\theta}$ and $\boldsymbol{k}$ (and thus they are a function of $\boldsymbol{\mu}$). Then, for the Neyman-Pearson lemma the optimal *test statistic* for the significance test is the profile likelihood ratio. This is given by the profile likelihood divided by the unconditional likelihood function at its maximum

$$q_\mu = -2\ln\lambda(\boldsymbol{\mu}) = -2\ln\frac{\mathcal{L}(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{k}}}, \hat{\hat{\boldsymbol{\theta}}})}{\mathcal{L}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{k}}, \hat{\boldsymbol{\theta}})}, \tag{8.7}$$

where $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{k}}, \hat{\boldsymbol{\theta}}$ are the unconditional MLE, i.e. the measured best-fit values of the parameters $\boldsymbol{\mu}, \boldsymbol{k}, \boldsymbol{\theta}$. These correspond to the parameter values which maximise the likelihood function

(with the constraints $0 \leq \hat{\mu} \leq \mu_s$, $\mu_s \in \boldsymbol{\mu}$) [275]. The uncertainty on the best-fit value of the signal strength is extracted by finding the values of μ_s that correspond to varying q_μ by one unit. Overall, the advantage of using the profile likelihood ratio, according to Wilks' theorem [101], is that the asymptotic distribution of $-2\ln\lambda(\boldsymbol{\mu})$ approaches a χ^2 distribution, independently of the nuisance parameters $\boldsymbol{\theta}$, in the limit of a large data sample.

An ATLAS fitting framework, called TRexFitter [282,283], is used to perform this statistical analysis. It builds histograms from the input data and provides them to the tools for statistical analysis. The fit itself is done using the HistFactory package [284], a tool specifically designed for profile likelihood fits in the form of histograms. The HistFactory is built on the RooFit [280] and RooStat [281] packages, which provide the technical implementation of the various statistical tools for the fit.

8.1.1 Compatibility and discovery significance

As it has been highlighted, the ultimate goal of the analysis is the discovery of the $t\bar{t}H(H \to b\bar{b})$ process, since it has not been observed, yet. In order to summarise the outcome of such a search, the level of agreement of the observed data with a given signal hypothesis needs to be quantified. For this purpose, the test statistic, q_μ, is used to assess their compatibility, while it is defined (see eq. 8.7) such that higher values represent increasing incompatibility between the data and the hypothesis. Rejecting effectively the background-only hypothesis, leads to the discovery of a new signal. The compatibility with the background-only hypothesis, corresponding to $\mu_s = 0 \;\forall\; \mu_s \in \boldsymbol{\mu}$, is measured with the relevant test statistic, denoted as q_0.

The level of agreement between the data and a hypothesis is further quantified with the p-value. It is the probability, under the assumption of a signal hypothesis μ (or the background-only $\mu = 0$ hypothesis), of finding data of equal or greater incompatibility with the predictions of the hypothesis than the observed one. For an observed value $q_{\mu,obs}$ ($q_{0,obs}$), the p-value p_μ (p_0) is expressed as

$$p_\mu = \int_{q_{\mu,obs}}^{\infty} f(q_\mu|\mu)dq_\mu \;\overset{\mu=0}{\longrightarrow}\; p_0 = \int_{q_{0,obs}}^{\infty} f(q_0|0)dq_0, \tag{8.8}$$

where $f(q_\mu|\mu)$ ($f(q_0|0)$) denotes the pdf of the test statistic q_μ (q_0) assuming a hypothesis μ ($\mu = 0$), and is depicted in Fig. 8.1a. The hypothesis is regarded as excluded if its p-value is observed below a specified threshold.

In case of a search, a possible excess in data compatible with a certain signal hypothesis, i.e. a potential discovery, is convenient to be assessed by converting the p-value into an equivalent *significance*, Z, as illustrated in Fig. 8.1b. It is defined such that a Gaussian distributed variable q_μ, which is found Z standard deviations (σ) above its mean, has an upper-tail probability equal to the p-value p, and expressed by

$$Z = \Phi^{-1}(1-p), \tag{8.9}$$

where Φ^{-1} is the quantile (inverse of the cumulative distribution) of the standard normal distribution. Especially in the large sample (asymptotic) limit, the significance of a deviation from the background-only hypothesis, assuming a positive measured μ ($\hat{\boldsymbol{\mu}} \geq 0$), is proved to be directly determined from the value of the test statistic (eq. 8.7) at $\boldsymbol{\mu} = 0$ [275]

$$Z = \Phi^{-1}(1-p_0) = \sqrt{q_{0,obs}} = \sqrt{2[\ln\mathcal{L}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{k}}, \hat{\boldsymbol{\theta}}) - \ln\mathcal{L}(0, \hat{\hat{\boldsymbol{k}}}, \hat{\hat{\boldsymbol{\theta}}})]}. \tag{8.10}$$

If the data fluctuate downwards, such that fewer events than even predicted by background processes alone are found, then $\hat{\boldsymbol{\mu}} < 0$ is obtained and a value of $q_{0,obs} = 0$ is chosen. In particle physics, for a signal process, such as the Higgs boson, the rejection of the background-only hypothesis with a significance of at least $Z\sigma = 5\sigma$ is regarded as an appropriate level to constitute a discovery. This corresponds to $p = 2.87 \times 10^{-7}$. For purposes of excluding a signal hypothesis, a threshold p-value of 0.05 (i.e. 95% confidence level) is often considered, which corresponds to $Z = 1.64$.

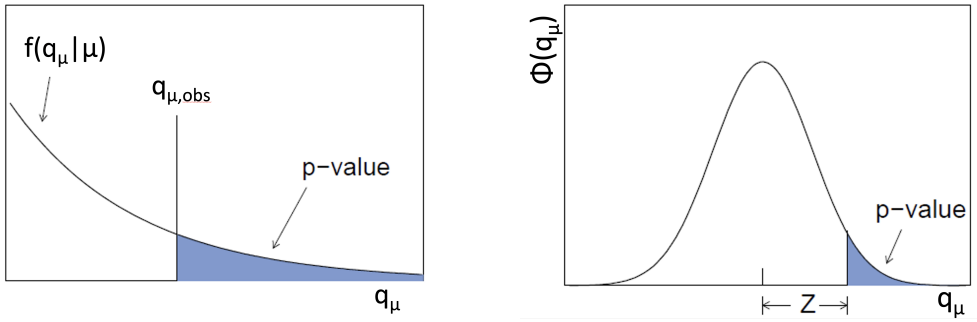


Fig. 8.1: Illustration of a) the relation between the p-value obtained from an observed value of the test statistic, $q_{\mu,obs}$, and b) the standard normal distribution $\Phi(q_\mu)$ showing the relation between the significance of Z standard deviations and the p-value.

To characterise the sensitivity of an experiment to a potential discovery, not only the significance obtained from a single dataset is interesting, but also the *expected* (or *median*) *significance* with which one would be able to reject different values of μ. Specifically, for the case of discovery one would like to know the median, under the assumption of the nominal signal model ($\mu = 1$), with which one would reject the background-only ($\mu = 0$) hypothesis. In this analysis, the estimation of the median significance is achieved by replacing the real data with an Asimov dataset [275] generated under the assumption of $\mu = 1$ (described in Sec. 8.3.1).

8.1.2 Setting upper limits

In the absence of a discovery, namely if the measured signal strength shows no significant excess with respect to the background-only hypothesis ($\mu = 0$), exclusion limits can be set on the production cross-section of the $t\bar{t}H$ process, by performing hypothesis tests based on a frequentist approach. The test statistic, defined in eq. 8.7, is exploited not only to evaluate the validity of the background-only hypothesis, but also to make statistical inferences about μ. Thus, in terms of this analysis, it is used to set upper limits, too, employing the CL_s *method* [285, 286], as implemented in the RooFit package [280].

According to the CL_s method, a test statistic q is used to distinguish between the hypotheses that the data contain signal plus background ($s + b$) or only background (b). These correspond to the distributions $f(q_\mu|s + b)$ and $f(q_\mu|b)$, respectively, as depicted in Fig. 8.2. The compatibility among the observed data (q_{obs}) and a given hypothesis is measured by a p-value. Since the $f(q_\mu|b)$ is here shifted to the right of $f(q_\mu|s + b)$, the p-value of $s + b$ is considered as the probability to find $q \geq q_{obs}$, under assumption of the $s + b$ hypothesis, i.e.

$$p_{s+b} = P(q \geq q_{obs}|s + b) = \int_{q_{obs}}^{\infty} f(q_\mu|s + b)dq_\mu. \tag{8.11}$$

In a similar way, the p-value of the background-only hypothesis is considered to be

$$p_b = P(q \leq q_{obs}|b) = \int_{-\infty}^{q_{obs}} f(q_\mu|b)dq_\mu. \tag{8.12}$$

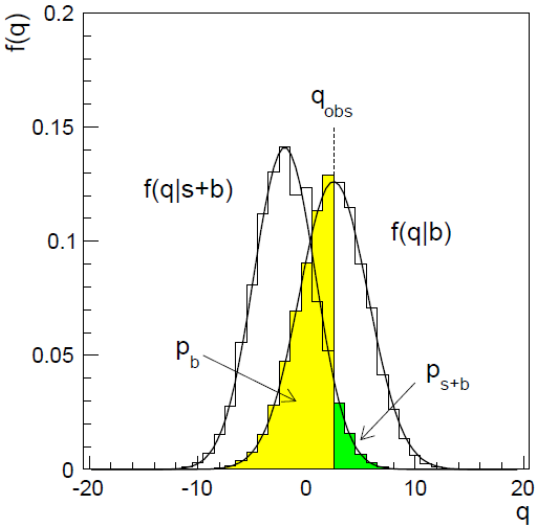


Fig. 8.2: Example of distributions of the test statistics for the background-only (b) and signal-plus-background ($s + b$) hypotheses [275].

In case only the p-value of a hypothesis was considered when carrying out a statistical test, a signal model would be regarded as excluded at a confidence level of 95% if just $p_{s+b} < 0.05$. As a result, hypotheses to which one has little or no sensitivity will be excluded with probability close to 5%. This corresponds to cases where the expected number of signal events is much less than that of background. Therefore, to avoid excluding models to which one has little or no sensitivity, the CLs variable, defined as

$$\text{CL}_s(\mu) = \frac{p_{s+b}}{1 - p_b}, \tag{8.13}$$

is used instead of just the p-value of a hypothesis. Finally, the 95% Confidence Level (CL) upper limit on μ, referred to as $\mu^{95\%CL}$, is the value of μ for which $\text{CL}_s = 0.05$. Hence, a value of μ greater than $\mu^{95\%CL}$ is excluded at 95% CL.

8.2 The fit model

The fit model of the $t\bar{t}H(H \to b\bar{b})$ analysis has been described in Ch. 6 and 7. It consists of the chosen discriminant variables to build the template distributions, as well as the list of systematic uncertainties and their correlations across the defined analysis regions. The statistical matching of the expected distributions to data is done in a template profile likelihood fit, which is performed simultaneously on the signal (SRs) and control (CRs) regions of the single-lepton resolved and boosted as well as dilepton channels. In total, events from sixteen orthogonal analysis regions, summarised in Sec. 6.4.2, are combined in the signal extraction fit, which determines the event yields for the signal and the dominant $t\bar{t}+ \geq 1b$ background. Simultaneously, the normalisation and shape of the differential distributions of the backgrounds are constrained within the assigned systematic uncertainties.

In the SRs across all channels, the distribution of the multivariate discriminant, i.e. the corresponding classification BDT output, is employed for the statistical analysis, in order to maximise the sensitivity to the signal. In all SRs, both the shape and normalisation of the classification BDT distribution are used as input to the fit. The only exception is for the single-lepton resolved region ($SR^{\geq 6j}_{\geq 4b}$) with $p_T^H \in [450, \infty)$ GeV, which has only one bin due to low statistics, thus only the normalisation of the distribution is used.

Additionally, the fit exploits the background dominated regions in order to improve the knowledge of the dominant $t\bar{t}$+jets background, through constraints of the nuisance parameters or their resulting correlations, while constraining the large systematic uncertainties in the $t\bar{t}$+jets modelling. In the single-lepton CRs, the shape and normalisation of the average ΔR distribution for all possible combinations of *b*-tagged jet pairs in an event, ΔR_{bb}^{avg}, is used to help better constrain the background contributions and correct their shape. In the dilepton CRs, which are enriched in $t\bar{t} + \geq 1c$ background events, only the event yield is used to correct the amount of these events predicted from the $t\bar{t}$ sample. This is done to avoid propagating mismodelling effects from the control to the signal regions, but also to prevent arbitrary pulls of the nuisance parameters in order to correct such mismodelling.

The binning of these distributions (discussed in Sec. 7.2.1) is optimised to maximise the analysis sensitivity while keeping the total MC statistical uncertainty in each bin to a level adjusted to reduce fluctuations in the predicted number of events. After all, the bins i of eq. 8.2 refer to the bins of the template distributions, while the regions r correspond to the sixteen analysis regions of the single-lepton and the dilepton channels.

In the statistical model, a nuisance parameter, θ_p, with Gaussian prior constraints, is assigned to each source of systematic uncertainty as part of the likelihood function (eq. 8.4). Also, the MC statistical uncertainties in each bin of the discriminant distributions are taken into account in the likelihood fit through dedicated parameters $\gamma_{r,i}$, as defined in eq. 8.5. Then, the signal strength $\mu_{t\bar{t}H}$, which represents the normalisation of the $t\bar{t}H$ signal, is the parameter of interest in the fit and is allowed to float freely with no prior knowledge applied. For the inclusive cross-section measurement a signal-strength parameter is determined for the whole phase space, whereas for the differential STXS measurement one signal-strength parameter is defined for each of the $\hat{p}_T^H$ bins.

Moreover, the normalisation of the distribution of each background process can be determined from the fit, simultaneously with $\mu_{t\bar{t}H}$, as a free parameter. According to studies performed in the beginning of the analysis, it was shown that the MC simulations underestimate the $t\bar{t} + \geq 1b$ fraction of events with respect to data. Therefore, the normalisation of the $t\bar{t} + \geq 1b$ component, denoted as $k(t\bar{t} + \geq 1b)$ and referred to the k_b factor in eq. 8.2, is a free-floating parameter. Hence, it is applied without any prior uncertainty and is determined by the fit to data. By contrast, none of the other backgrounds has a free-floating normalisation. Instead, their normalisation is controlled through specific nuisance parameters, that reflect the theoretical knowledge of the respective cross sections, and the data themselves.

8.2.1 Pruning

A large amount of systematic uncertainties is considered in the analysis, summarised in Table 7.1. Some of them have a significant impact on the final fit result, such as those related to the $t\bar{t}+ \geq 1b$ modelling, though most of them are negligible. If all of them were included in the likelihood fit, problems would arise with the convergence of the minimisation procedure, while the overall fitting would be very time-consuming. To avoid these issues, the systematic

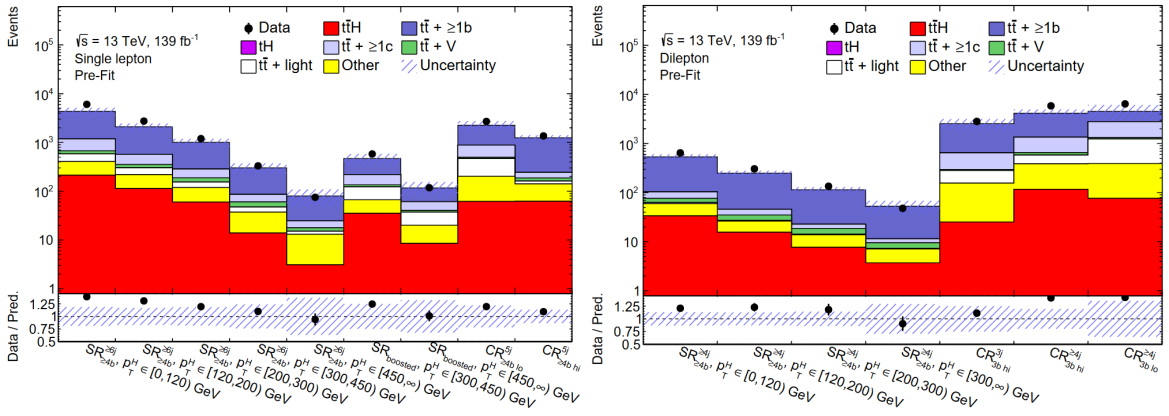
uncertainties that have a negligible effect on the analysis are *pruned*, namely they are removed from the likelihood function, resulting in a simplified fit model.

As already introduced, the systematic variations are split into their normalisation and shape components[1], hence the pruning is applied separately to them. Also, the pruning is done independently for each sample and analysis region. In this analysis, the normalisation of a systematic variation is dropped if the normalisation difference with respect to the nominal distribution is smaller than 0.5%. Similarly, the shape of a systematic is dropped when no bin of the shape component in the given region has a relative difference bigger than 0.5%.

Finally, in order to affirm that the choice of the pruning threshold does not have a significant effect on the fit result, a lower cut-off value of 0.1% for both the shape and the normalisation components was tested. Only inappreciable differences were found on the constraints or pulls of some remaining nuisance parameters, while the uncertainty on μ is almost unchanged.

8.2.2 Pre-fit modelling of fitted and kinematic distributions

With the final set of systematic variations, all components of the fit model are available for the statistical analysis. The quality of the signal and background modelling, detailed in Sec. 4.5, can be assessed by comparing the simulation with the measured data in the various analysis regions, before performing any fit to data (*pre-fit*). For the pre-fit case, the uncertainty band in the following plots includes all the statistical and systematic uncertainties. The uncertainty on the $t\bar{t} + \geq 1b$ background normalisation is not included, though, since it is a free-floating parameter and will be determined from the fit to data.



(a) **(b)**

Fig. 8.3: Comparison of the predicted and observed event yields in each of the signal and control regions, in the (a) single-lepton and (b) dilepton channels, before performing the fit to data (pre-fit). The uncertainty band includes the statistical and systematic uncertainties. The uncertainty on the $t\bar{t} + \geq 1b$ background normalisation is not included as it is a free-floating parameter of the fit.

Figure 8.3 shows the predicted number of events before the fit to data compared to the

[1]The normalisation component describes the effect of the systematic variation on the yield of the nominal distribution in a given region. Then the shape component is the systematic variation modified, such that it does not affect the normalisation in the given region.

amount of observed data events in each analysis region in the single-lepton (Fig. 8.3a) and dilepton (Fig. 8.3b) channels. Data overshoot the prediction in several regions, especially in those with a large fraction of $t\bar{t}+\geq 1b$ and possibly also $t\bar{t}+\geq 1c$ background components. This difference between data and prediction is mainly connected to the bad modelling of the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ backgrounds. In fact, in the previous iteration of the analysis the normalisations of these two background components, which were both free-floating parameters, were found to be underestimated by 24% and 63%, accordingly [97]. Nevertheless, in most of these regions the difference is covered by the total of statistical and systematic uncertainties.

Figures 8.4 and 8.5 show the comparison of data and MC prediction for the distributions of the variables which will eventually enter in the fit, as outlined in the beginning of Sec. 8.2, in the single-lepton and dilepton analysis regions, respectively. As already highlighted in Sec. 6.5.3, the classification BDT helps to increase the overall signal to background (S/B) ratio in the single-lepton boosted, as well as in the other analysis regions. Therefore, the low bins of the classification BDT output have very few signal events, whereas the last bins are enriched in signal events. As already noticed in Fig. 6.9, among the inclusive SRs the highest purity is observed in the boosted channel. Then, among the individual p_T^H regions, the largest increase in the purity is also achieved in the two boosted signal regions, starting from an overall $S/B \sim 8\%$ to 30-40% in the last bin of the BDT score, accordingly. According to the pre-fit modelling of the template distributions, the disagreement between data and MC events is mainly noticed in normalisation, mostly encountered in the resolved low-p_T^H SRs and the dilepton CRs, though it is not always covered by the systematic uncertainties. Notably, there is a normalisation difference between the 5-jet CRs and the 6-jet SRs of the single-lepton resolved channel, which implies that the ISR systematic, discussed in Sec. 7.1.2, will play an important role in the fit to compensate for this discrepancy. Apart from the normalisation effects, there is no significant difference in the shape between the data and prediction.

In order to further test the validity of the background modelling, comparisons between the prediction and the observed data have been studied for various kinematic distributions. In the following, the most indicative ones are presented. In Fig. 8.6, the reconstructed Higgs-boson candidate mass distribution is depicted for the three inclusive in p_T^H signal regions. A peak around the Higgs boson mass is discernible, which is expected according to the selection criteria in each analysis region. No significant mismodelling is observed in this distributions, except for a distinct normalisation effect in the single-lepton resolved region. Nevertheless, the discrepancy between data and simulated events lies within the assigned uncertainties.

A significant variable for this analysis is the reconstructed Higgs boson candidate transverse momentum p_T^H, shown in Fig. 7.1, since it is used to split the SRs of the analysis channels. As already discussed in Sec. 7.1.2, there is mainly a clear shape effect in this distribution and a dedicated uncertainty is assessed to cover this mismodelling. Then, another meaningful variable is the number of jets, which is used to define the various analysis regions. The pre-fit distributions of the number of jets in the three inclusive SRs are illustrated in Fig. 8.7. There is an evident mismodelling, so in the shape as in the normalisation of this distribution. The agreement between the data and the prediction becomes increasingly bad with the increasing jet multiplicity. It is still covered by the modelling uncertainties, though.

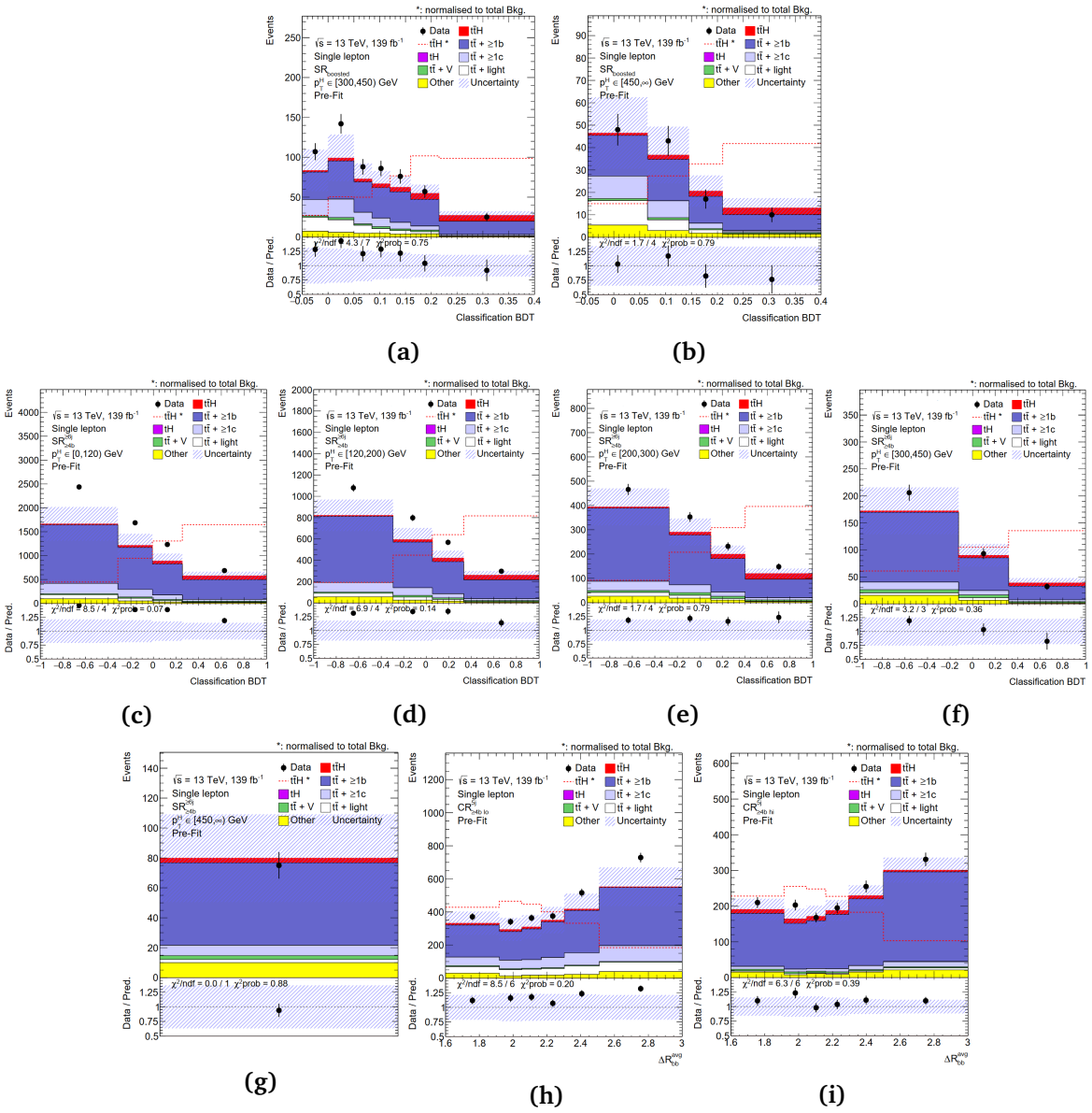


Fig. 8.4: Comparison between data and MC prediction for the BDT and ΔR_{bb}^{avg} discriminants in the SRs and CRs of the single-lepton channel, respectively, before performing the fit to data (pre-fit). The boosted signal region, split in the (a) $300 \leq p_T^H < 450$ GeV (b) $p_T^H \geq 450$ GeV regions, is shown. The resolved signal region, $\text{SR}_{\geq 4b}^{\geq 6j}$, split into (c) $0 \leq p_T^H < 120$ GeV, (d) $120 \leq p_T^H < 200$ GeV, (e) $200 \leq p_T^H < 300$ GeV, (f) $300 \leq p_T^H < 450$ GeV, (g) $p_T^H \geq 450$ GeV (yield only) regions, as well as the control regions (h) $\text{CR}_{\geq 4b\ lo}^{5j}$ and (i) $\text{CR}_{\geq 4b\ hi}^{5j}$ are shown. In the latter, the first (last) bin includes the underflow (overflow). The $t\bar{t}H$ signal yield (solid red) is normalised to the Standard Model expectation. The dashed line shows the $t\bar{t}H$ signal distribution normalised to the total background prediction. The uncertainty band includes the statistical and systematic uncertainties. The uncertainty on the $t\bar{t} + \geq 1b$ background normalisation is not included as it is a free-floating parameter of the fit.

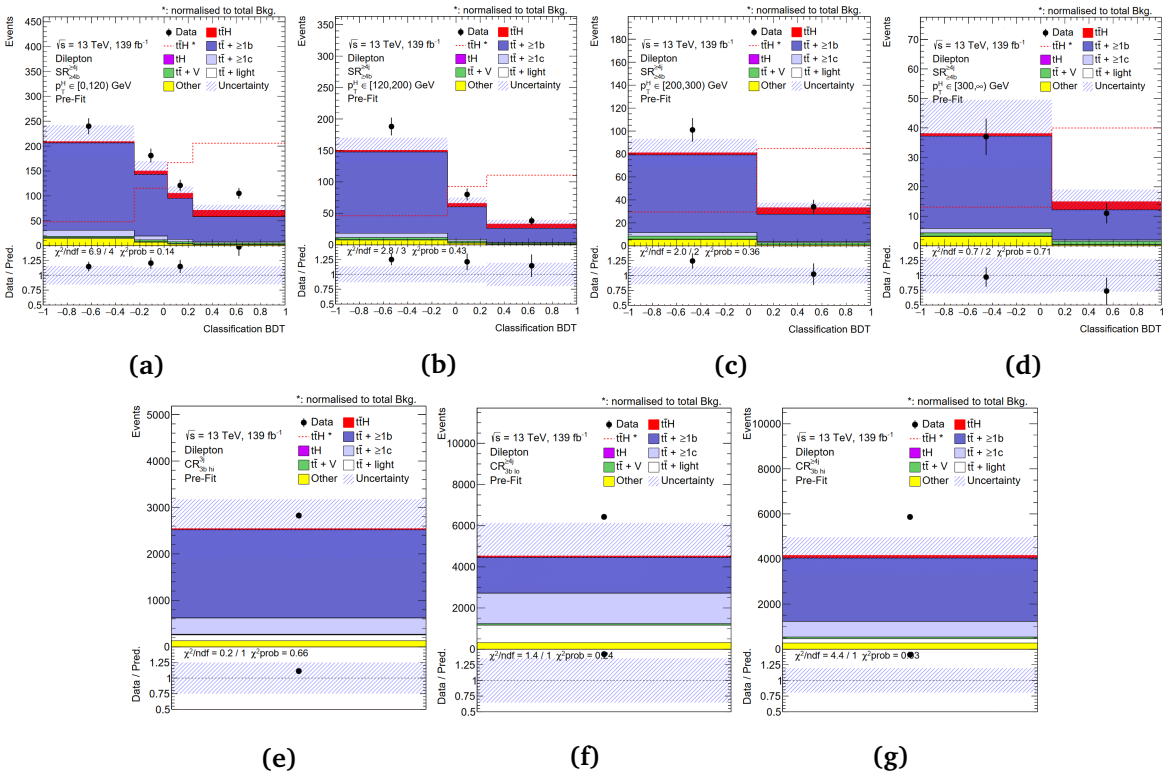
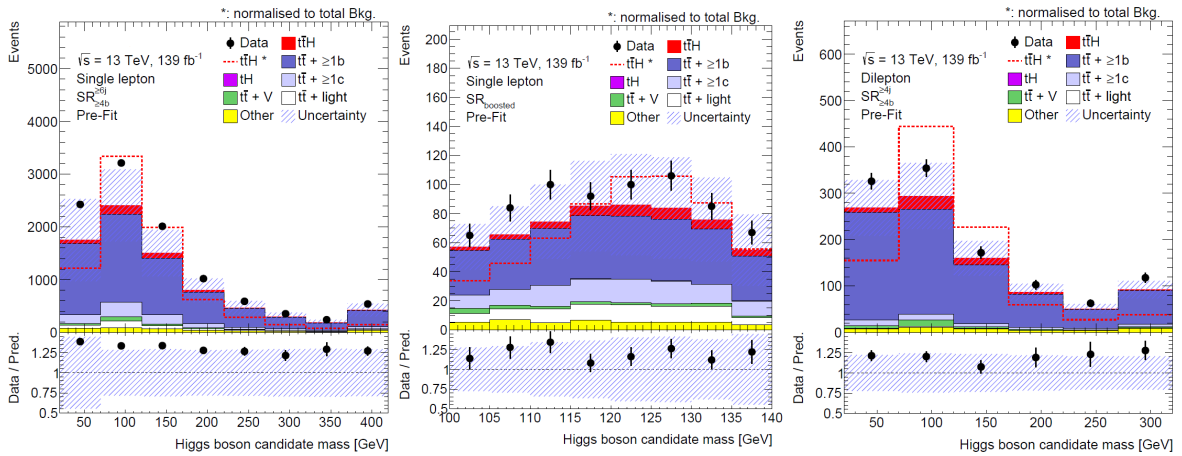
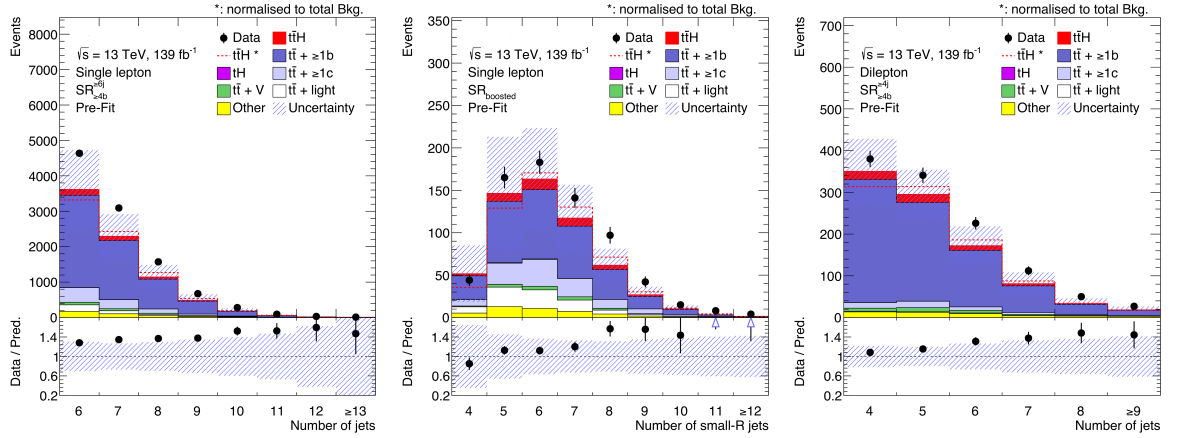


Fig. 8.5: Comparison between data and MC prediction for the BDT discriminant and the event yield in the SRs and CRs of the dilepton channel, respectively, before performing the fit to data (pre-fit). The signal region, $\text{SR}^{\geq 4j}_{\geq 4b}$, split into (a) $0 \leq p_T^H < 120$ GeV, (b) $120 \leq p_T^H < 200$ GeV, (c) $200 \leq p_T^H < 300$ GeV, (d) $p_T^H \geq 300$ GeV regions, as well as the control regions (e) $\text{CR}^{3j}_{\geq 3b\ hi}$, (f) $\text{CR}^{4j}_{\geq 3b\ lo}$, and (g) $\text{CR}^{4j}_{\geq 3b\ hi}$ are shown. The $t\bar{t}H$ signal yield (solid red) is normalised to the Standard Model expectation. The dashed line shows the $t\bar{t}H$ signal distribution normalised to the total background prediction. The uncertainty band includes the statistical and systematic uncertainties. The uncertainty on the $t\bar{t} + \geq 1b$ background normalisation is not included as it is a free-floating parameter of the fit.



(a) (b) (c)

Fig. 8.6: Pre-fit distributions of the reconstructed Higgs boson candidate mass for the (a) single-lepton resolved $SR^{\geq 6j}_{\geq 4b}$, (b) single-lepton boosted $SR_{boosted}$, and (c) dilepton $SR^{\geq 4j}_{\geq 4b}$ signal regions. The $t\bar{t}H$ signal yield (solid red) is normalised to the Standard Model expectation. The dashed line shows the $t\bar{t}H$ signal distribution normalised to the total background prediction. The uncertainty band includes the statistical and systematic uncertainties, except for the uncertainty in the $k(t\bar{t}+\geq 1b)$ normalisation factor which is not defined pre-fit. The first (last) bin includes the underflow (overflow).



(a) (b) (c)

Fig. 8.7: Pre-fit distributions of the number of jets for the (a) single-lepton resolved $SR^{\geq 6j}_{\geq 4b}$, (b) single-lepton boosted $SR_{boosted}$, and (c) dilepton $SR^{\geq 4j}_{\geq 4b}$ signal regions [100]. The $t\bar{t}H$ signal yield (solid red) is normalised to the Standard Model expectation. The dashed line shows the $t\bar{t}H$ signal distribution normalised to the total background prediction. The uncertainty band includes the statistical and systematic uncertainties, except for the uncertainty in the $k(t\bar{t}+\geq 1b)$ normalisation factor which is not defined pre-fit. The last bin includes the overflow.

Lastly, the following two variables, as well as the number of jets, are some of the most highly ranked input variables to the classification BDT training in the single-lepton boosted region (based on Table 6.14). Each distribution is shown for the $300 \leq p_T^H < 450$ GeV (left) and $p_T^H \geq 450$ GeV (right) regions in the boosted channel. Figures 8.8a-8.8b depict the DNN

$P(H)$ output for the Higgs-boson candidate. In Fig. 8.8c-8.8d the hadronic top-quark invariant mass is illustrated, where a distinct peak around the top quark mass is apparent, as expected according to the reconstruction criteria in the $SR_{boosted}$ (see Sec. 6.4.1). There is a reasonable agreement between data and prediction in these distributions. Although the prediction is underestimated compared to data in most bins, they are compatible within the uncertainties.

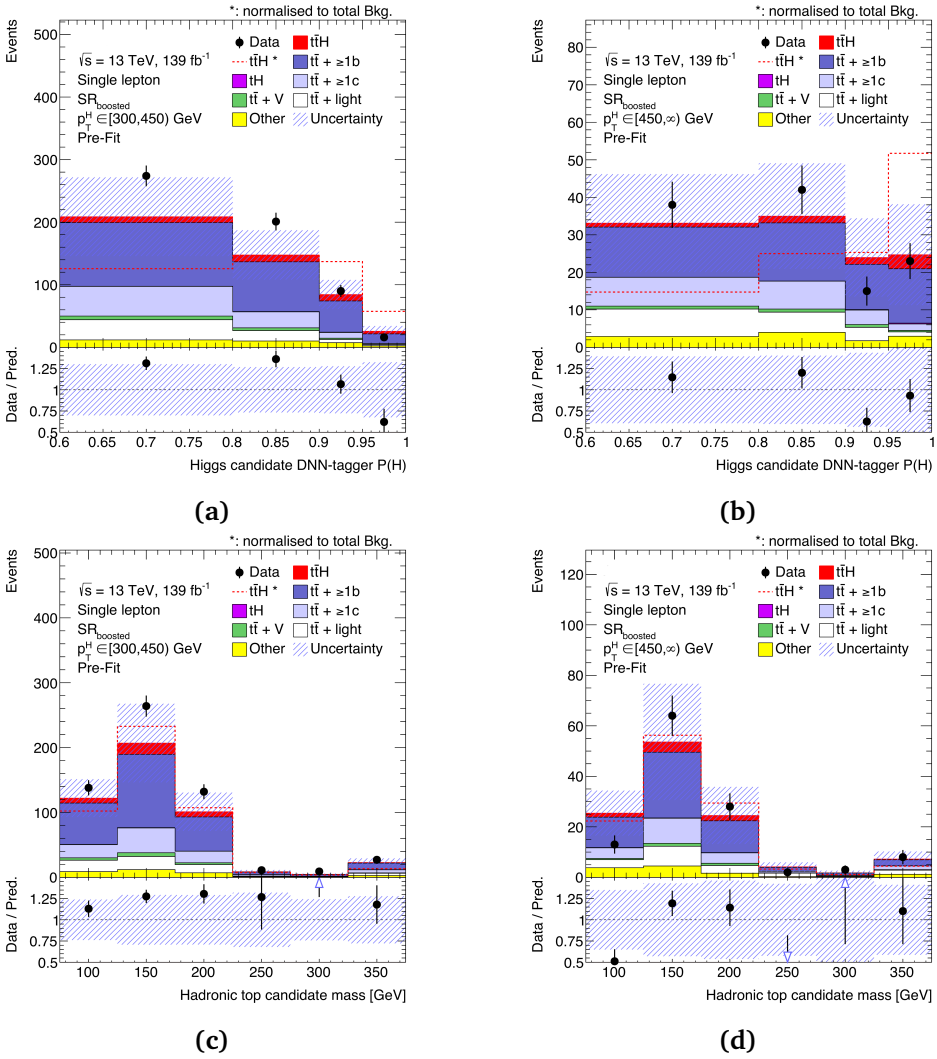


Fig. 8.8: Pre-fit distributions of (a),(b) the DNN $P(H)$ output for the Higgs-boson candidate as well as (c),(d) the hadronic-top invariant mass in the single-lepton boosted $SR_{boosted}$ for the $300 \leq p_T^H < 450$ GeV (left) and $p_T^H \geq 450$ GeV (right) regions, respectively. The $t\bar{t}H$ signal yield (solid red) is normalised to the SM expectation. The dashed line shows the $t\bar{t}H$ signal distribution normalised to the total background prediction. The uncertainty band includes the statistical and systematic uncertainties, except for the uncertainty in the $k(t\bar{t}+\geq 1b)$ normalisation factor which is not defined pre-fit. The first (last) bin includes the underflow (overflow).

8.2.3 Goodness of fit

When a fit procedure is used, it is important to further examine the quality of the fit in the analysis by obtaining the *goodness of fit* status. It is a metric that quantifies how well the fit model

describes the observed data. If the goodness of fit assigns a very small probability, the model should be checked. The goodness-of-fit test is evaluated by using a saturated model [287], including all the input variables to the classification BDTs and to the fit. The saturated model has as many estimated parameters as data points, i.e. one extra free parameter per bin of the fitted distribution. Consequently, such a model can perfectly fit the data, without requiring nuisance parameter pulls that would result in likelihood penalties due to the associated constraint terms. Thus, it can be used to compare to the actual fit model and evaluate its quality. Then, the ratio between the likelihood of the nominal model and of the saturated model follows a χ^2 distribution asymptotically (Wilks theorem), as mentioned in Sec. 8.1. Eventually, the corresponding χ^2 probability [288] is taken as a goodness-of-fit value. To further compare it to the "χ^2 test", the saturated model represents $\chi^2 = 0$, denoting the perfect agreement.

8.3 Expected performance

In general, the physics analyses which search for new processes are initially performed in pseudo-data sets and optimised without looking at the data in regions sensitive to the signal. If the nominal fit to data was performed without excluding the signal events, then optimising the fit model based on its results would lead to a bias, modifying the model to acquire a desired result. Additionally, since the analysis sensitivity is systematically limited, the completeness of the systematics model and the performance of the analysis need to be evaluated after the fit but before looking at the final result in data. Therefore, $t\bar{t}H(H \to b\bar{b})$ is a physics search which needs to be conducted in a blinded way, at first.

The analysis is primarily optimised on simulated signal and background samples. Then, the validation of the fit model and the performance of the analysis are evaluated from a profile likelihood fit to the Asimov dataset [275], instead of the actual data, and a background-only fit with blinded part of the data. Most of the studies and decisions concerning the determination of the analysis strategy were based on the obtained performance of these fits. In particular, the event categorisation and region definition, the construction of the MVAs, the choice of the discriminant distributions, and the optimisation of their binning are some of the fundamental parts of the analysis strategy. Also, studies have been conducted about the systematic uncertainties, understanding their potential constraints and their impact on the signal sensitivity.

In the following, the performance of these fits is presented for the combination of the single-lepton boosted and resolved as well as the dilepton channels. The full combination provides additional information on the fit model and can be used to better control the background. Moreover, both the inclusive and differential cross-section measurements (the latter just called "STXS measurement") are conducted, using the same strategy. The only difference is that in the STXS measurement, the signal template is divided into five templates according to the truth $\hat{p}_T^H$ and for every signal template a separate signal strength μ is considered to the fit. The expected results, based on fits to the Asimov dataset under the signal-plus-background ($S + B$) hypothesis, are presented in Sec. 8.3.1. Then, fits to data under the background-only hypothesis, applying the blinding strategy described in Sec. 8.3.2, are presented in Sec. 8.3.3.

8.3.1 $S + B$ fit to Asimov data

The Asimov dataset [275], in fact a pseudo-data set, is generated for a particular set of model parameters, such that the maximum likelihood estimators of all those parameters correspond

to their true values. It is built as a binned dataset from the prediction of the signal and background nominal model, reflecting the statistics of the real data though. The event count in each bin is set to the expectation value of the respective predicted event yield (the latter following the Poisson distribution) for the chosen model parameters, hence all statistical fluctuations are suppressed. In addition, a Poisson error is assigned to each bin, corresponding to the statistical uncertainty of the data.

Furthermore, the likelihood for an Asimov dataset is constructed as for any other data. Then, the profile likelihood fit performed to the Asimov dataset, by definition, gives the true values for all parameters. As a result, the signal strength the $t\bar{t}+\geq 1b$ background normalisation factor are fixed at 1, which corresponds to the nominal signal and background model. Additionally, no pulls are expected for the nuisance parameters (NPs), namely they should not deviate from their nominal values, thus are fixed at 0 (only the NPs related to the MC statistical uncertainty per bin are fixed at 1), as defined in Sec. 8.1.

As a consequence, uncertainties on the signal strength and the background normalisation can be extracted from the fit to Asimov dataset, based on the prediction of the nominal model. It also provides an estimate of how much the data should be able to constrain the various systematic uncertainties (NPs), while it determines which systematics will have a sizable effect on the actual signal sensitivity. Finally, from this fit the expected (median) significance can be assessed, derived from eq. 8.10 where the measured value $\hat{\mu}$ corresponds to the true value of the nominal model.

Inclusive cross-section measurement

The profile likelihood fit on the Asimov data for the inclusive cross-section measurement, combining all three channels, results in an expected $t\bar{t}H$ signal strength

$$\mu_{t\bar{t}H} = 1.00 \pm 0.18\,(\text{stat.})\,^{+0.30}_{-0.25}\,(\text{syst.}) = 1.00^{+0.36}_{-0.31}, \tag{8.14}$$

corresponding to an expected (median) significance of 3.4σ (standard deviations) with respect to the background-only hypothesis. Also, the expected uncertainty on the free-floating $t\bar{t}+\geq 1b$ background normalisation factor is found to be $k(t\bar{t}+\geq 1b) = 1.00^{+0.07}_{-0.06}$. The contribution of the data statistical uncertainty to the result is evaluated by repeating the fit with all systematic parameters, except for the free-floating parameters $\mu_{t\bar{t}H}$ and $k(t\bar{t}+\geq 1b)$, fixed to their estimated value, which in the case of the Asimov fit is their true value. The $k(t\bar{t}+\geq 1b)$ is basically unaffected by the statistics with an uncertainty of only 1%. By contrast, the effect on the $\mu_{t\bar{t}H}$ is relatively large (18%), but still much smaller than the overall uncertainty, revealing that the analysis is limited by the systematic uncertainties. The total systematic uncertainty is simply derived by quadratically subtracting the statistical uncertainty from the total uncertainty.

Nevertheless, in the fit to the actual data the measured values of the $k(t\bar{t}+\geq 1b)$ factor and the nuisance parameters may deviate from their nominal values and thus modify the sensitivity of the analysis. In order to take into account the effect of the pulls and $t\bar{t}+\geq 1b$ normalisation factor, a more "realistic" significance needs to be assessed. Therefore, a $S+B$ fit is performed to a modified Asimov dataset, which is built using the pulls of a $S+B$ fit to data where the signal strength is set to the SM expectation $\mu = 1$, yielding a "realistic" expected significance of 2.7σ. In the following, this is quoted as the expected significance of the analysis. The expected sensitivity is improved with respect to that from the previous measurement (1.6σ) [97] and the Asimov fit results (eq. 8.14) already foresee a significant reduction of the uncertainties on the measurement (discussed later in Sec. 8.4.1).

Furthermore, a combined fit, where each of the three channels has an individual signal strength, is performed. This would give an estimate of the signal sensitivity in the individual channels, if the background was constrained by all channels. Nonetheless, the fit procedure is identical to the nominal combined fit, hence the nuisance parameters and $k(t\bar{t}+\geq 1b)$ factors are correlated among the channels. Figure 8.9 shows the uncertainties on the expected $\mu_{t\bar{t}H}$ obtained for each channel from a combined fit as well as those from the inclusive-μ combined fit. Apparently, the uncertainties on the inclusive-μ are substantially constrained with respect to those obtained from single-μ fit. The single-lepton resolved channel has the smallest uncertainties and is thus the most sensitive. In accordance with the inclusive-μ fit, the regions are dominated by the systematic uncertainties, except for the single-lepton boosted region which is statistically limited by construction. In addition, the impact of systematic uncertainties in the dilepton region is larger than in the single-lepton resolved region. This is because the latter covers the great majority of the phase space in the analysis and thus has more constraining power on the measurement. Finally, the uncertainties on the $t\bar{t}+\geq 1b$ normalisation factor from this fit are found to be the same as in the inclusive-μ fit.

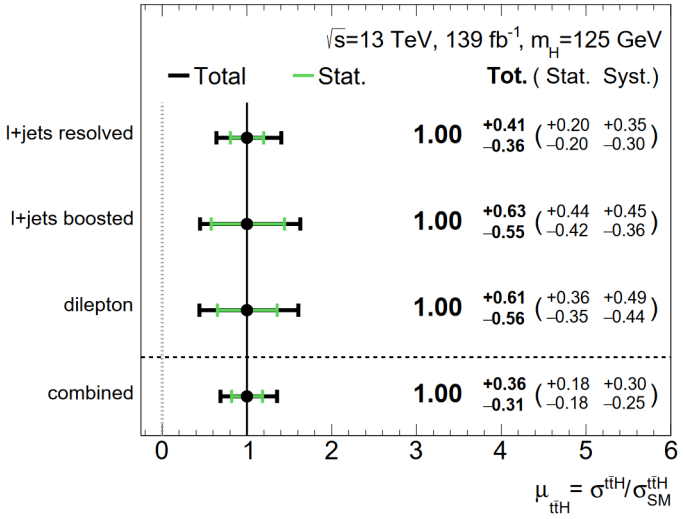


Fig. 8.9: Fitted values of the $t\bar{t}H$ signal-strength parameter from the S+B fit to the Asimov dataset, in the individual channels (single-lepton resolved, single-lepton boosted, and dilepton) and in the inclusive-μ measurement by combining the three channels.

In the previous round of the $t\bar{t}H(H\rightarrow b\bar{b})$ analysis [97], the single-lepton boosted channel had seldom impact on the expected result. In fact, when fitting only this channel individually, the uncertainty on the measured signal strength was about 310% [289], and was therefore not explicitly presented in the publication. In the full Run 2 analysis presented in this thesis though, the performance of this channel has been significantly improved, resulting in an uncertainty on the expected (or also measured) signal strength of roughly 140%. This yields an expected sensitivity comparable to that of the dilepton channel in the single-μ combined fit, already seen in Fig. 8.9. To better quantify the net gain on the sensitivity in the combined inclusive cross-section measurement due to the single-lepton boosted region, an Asimov fit is also performed in which this region is removed. A comparison of the expected uncertainties on $\mu_{t\bar{t}H}$ and $k(t\bar{t}+\geq 1b)$ between the latter configuration and the nominal combined fit is shown in Fig. 8.10. Eventually, the improvement on the expected sensitivity obtained, when including the single-lepton boosted region in the combination, amounts to nearly 4.3% (or 6.3% when

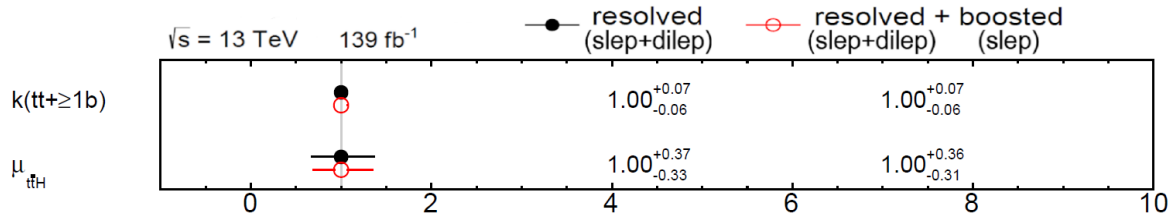


Fig. 8.10: Fitted normalisation factor for the $t\bar{t}+\geq 1b$ background and signal strength from the S+B fit to the Asimov dataset in the inclusive cross-section measurement, combining only the resolved (single-lepton and dilepton) regions (in black) compared to the full combination (including also the single-lepton boosted region) (in red).

considering only the single-lepton regions for the combination). In addition, the uncertainties on the normalisation of the $t\bar{t}+\geq 1b$ background remain unaffected. Although the improvement is rather small, this is expected since, as already explained, the single-lepton resolved region has by far the largest constraining power on the measurement. Nonetheless, this is an advancement with respect to the previous iteration of the analysis, given that the sensitivity of the measurement did not profit from the inclusion of the boosted region then. The gain on the sensitivity from the inclusion of the single-lepton boosted region is more explicit and determinant in the STXS measurement, though.

The sources of the systematic uncertainty are included in the profile likelihood fit through the various nuisance parameters NPs. Although they are included as uncorrelated parameters among each other, the fit creates correlations between complementary nuisance parameters. Large correlations can point to unnecessary degrees of freedom which are covered by other parameters, or to some variables being too correlated to the signal, making the analysis less sensitive. In general, there are only a few large (anti-)correlations among the different parameters and most of them are between the NPs related to the uncertainties arising from the dominant $t\bar{t}+\geq 1b$ background component. They are not considerably large though, and they are more or less expected from a fit model with that many NPs. Nevertheless, the most important correlations are the ones between the nuisance parameters and the signal strength ($\mu_{t\bar{t}H}$). These correlations imply that these systematics have similar features to the signal, thus they would affect the sensitivity of the analysis. There are only a few correlations between the signal strength and mainly $t\bar{t}+\geq 1b$ background NPs, but they are not considerably large. The correlation matrix of the nuisance parameters and the normalisation factors can be found in Appendix (Fig. A.9).

Then, the actual effect of the systematic uncertainties on $\mu_{t\bar{t}H}$ is illustrated in Fig. 8.11. It shows the twenty most important sources of systematic uncertainty, ranked based on the size of their impact on the signal strength after the combined fit to Asimov dataset. Also, it reflects the information from the correlation matrix. In particular, the seven highest-ranked nuisance parameters are also depicted in Fig. A.9 with a post-fit impact on signal strength comparable to the absolute amount of their correlation. Most of these NPs are associated with the $t\bar{t}+\geq 1b$ background modelling, while a couple of them are related to the $t\bar{t}H$ signal modelling. Specifically, the dominant post-fit impact on $\mu_{t\bar{t}H}$ comes from the $t\bar{t}+\geq 1b$ NLO generator matching in the two lowest-p_T^H single-lepton bins. In total, the $t\bar{t}+\geq 1b$ related NPs have smaller post-fit effect on $\mu_{t\bar{t}H}$ compared to what they have pre-fit, since the former depends also on the correlations among the NPs. This difference indicates the constraining power of the fit model particularly on the $t\bar{t}+\geq 1b$ background. Besides the uncertainties from

the $t\bar{t} + \geq 1b$ and $t\bar{t}H$ modelling, a couple of tW background related NPs also show up in the ranking. However, their impact is small compared to that of $t\bar{t} + \geq 1b$ NPs, while it is almost unchanged from their pre-fit impact. In fact, after the first few highly ranked NPs, the effect of the rest systematic variations is relatively similar, though with a slowly decreasing effect. Lastly, no systematic uncertainties related to the performance of the detector (experimental NPs) are listed among them, confirming the low impact of experimental uncertainties on the result of the measurement.

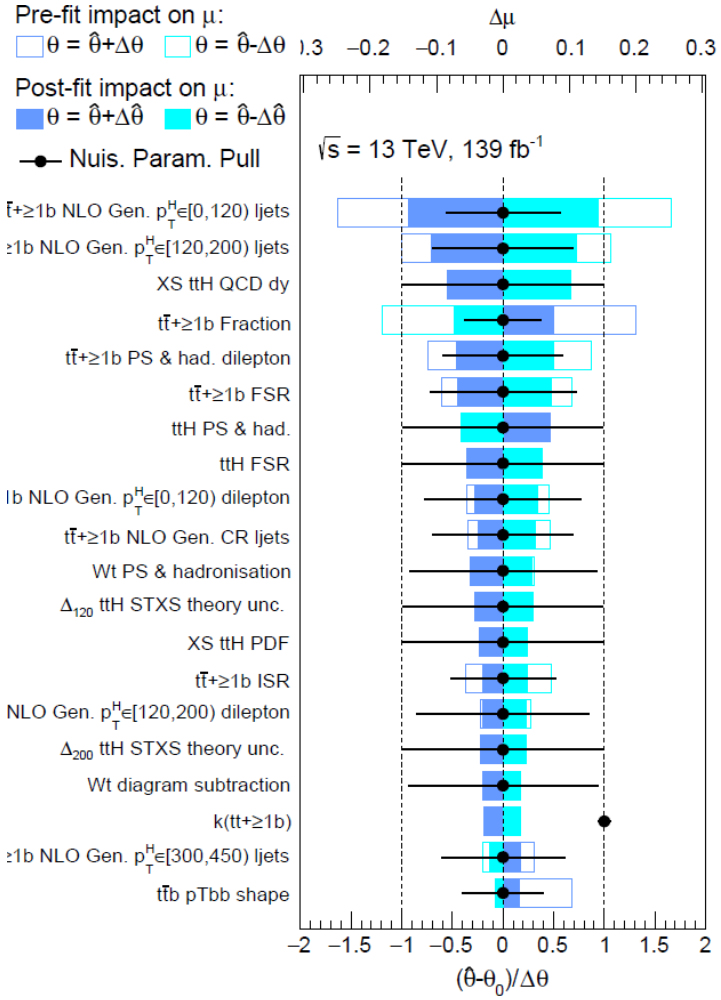


Fig. 8.11: Ranking of the 20 nuisance parameters with the largest post-fit impact on μ in the combined inclusive cross-section fit to the Asimov dataset. Nuisance parameters corresponding to statistical uncertainties in the MC simulated event samples are not included. The empty blue rectangles correspond to the pre-fit impact on μ, while the filled blue ones to the post-fit impact on μ, both referring to the upper scale. The impact of each nuisance parameter, $\Delta\mu$, is computed by comparing the nominal best-fit value of μ with the result of the fit when fixing the considered nuisance parameter to its best-fit value, $\hat{\theta}$, shifted by its pre-fit (post-fit) uncertainties $\pm\Delta\hat{\theta}$ ($\pm\Delta\theta$). The black points show the pulls of the nuisance parameters relative to their nominal values, $\theta_0 = 0$. These pulls and their relative post-fit errors, $\Delta\hat{\theta}/\Delta\theta$, refer to the lower scale. The parameter $k(t\bar{t} + \geq 1b)$ refers to the free-floating normalisation of the $t\bar{t} + \geq 1b$ background, for which the pre-fit impact on μ is not defined, while its nominal value is $\theta_0 = 1$. The "ljets" ("dilep") label refers to the single-lepton (dilepton) channel.

Additionally, Fig. 8.11 shows the constraints of the systematics variations while, as already introduced, no pulls are expected for the NPs from this fit. The modelling $t\bar{t}H$ signal systematics are not constrained at all, implying low sensitivity to the $t\bar{t}H$ modelling within the given precision. On the contrary, most of the NPs related to the $t\bar{t} + \geq 1b$ background exhibit large constraints. This is expected, since the most sensitive analysis regions are dominated by the $t\bar{t} + \geq 1b$ background, hence the model is able to constrain the various $t\bar{t} + \geq 1b$ modelling uncertainties and possibly compensate for any mis-modelling effects. All the fitted nuisance parameters and their constraints can be found in Appendix (Fig. A.10), though almost all the remaining uncertainty sources are not constrained at all.

STXS measurement

Considering that the analysis regions have been designed particularly for the STXS measurement, no big changes with respect to the inclusive cross-section measurement are required in order to perform this fit. As described in Sec. 6.1, the signal template is split into five truth $\hat{p}_T^H$ bins, corresponding to the reconstructed p_T^H bins into which the SRs are split. Also, since each signal template has a dedicated signal strength parameter, the STXS bin migration uncertainties are removed.

The expected signal strengths from the Asimov fit of the STXS measurement are shown in Fig. 8.12. The uncertainties associated to the signal strength parameter are overall fairly large. The total uncertainty is clearly dominated by the systematic uncertainties in the $\hat{p}_T^H \in [0, 120)$ GeV bin, whereas in the $\hat{p}_T^H \in [200, 300)$ GeV and $[300, 450)$ GeV bins the statistical uncertainty is larger. In addition, in the remaining two bins the systematic and the statistical parts of their total uncertainties are comparable. Moreover, the normalisation factor of the $t\bar{t} + \geq 1b$ background is found to be $k(t\bar{t} + \geq 1b) = 1.00 \pm 0.07$, in agreement with the inclusive cross-section fit uncertainties. The statistical uncertainty on these parameters is evaluated as described for the inclusive cross-section fit.

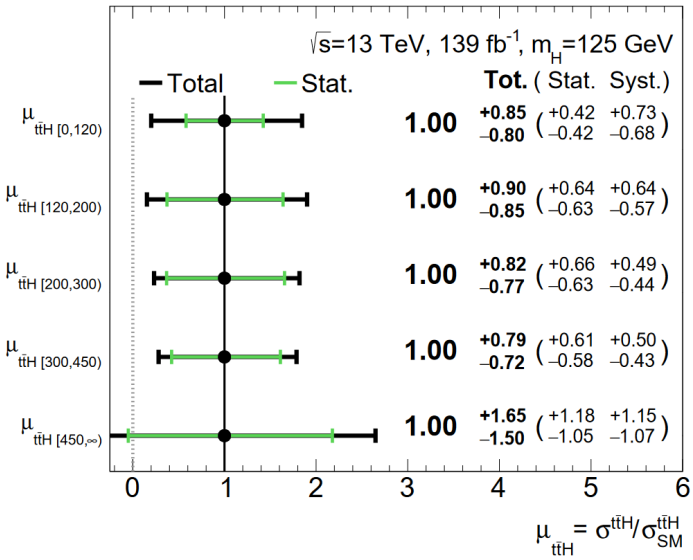


Fig. 8.12: Fitted values of the $t\bar{t}H$ signal-strength parameter from the S+B fit to the Asimov dataset in the individual STXS $\hat{p}_T^H$ bins combining all the analysis channels.

As done for the inclusive cross-section measurement, in order to better quantify the net

gain on the sensitivity in the combined STXS measurement due to the inclusion of single-lepton boosted region, an Asimov fit is also performed in which this region is removed. The comparison of the expected uncertainties on $\mu_{t\bar{t}H}$ of each $\hat{p}_T^H$ bin, as well as on $k(t\bar{t}+\geq 1b)$, between the latter and the nominal fit is shown in Fig. 8.13. So, when including the single-lepton boosted region in the combination, the uncertainties on the $k(t\bar{t}+\geq 1b)$ factor and on $\mu_{t\bar{t}H}$ in the two lowest $\hat{p}_T^H$ bins remain unchanged. However, there is significant improvement on the expected sensitivity which amounts to about 7%, 36%, and 73% in the three higher $\hat{p}_T^H$ bins, respectively. The improvement on the sensitivity from the inclusion of the single-lepton boosted region is remarkable in the two highest $\hat{p}_T^H$ bins. This is expected, since the boosted region targets events with high p_T^H and is thus more sensitive to these two bins of the measurement. Especially in the highest one, the sensitivity of the measurement would not even be measurable without the contribution of the boosted region.

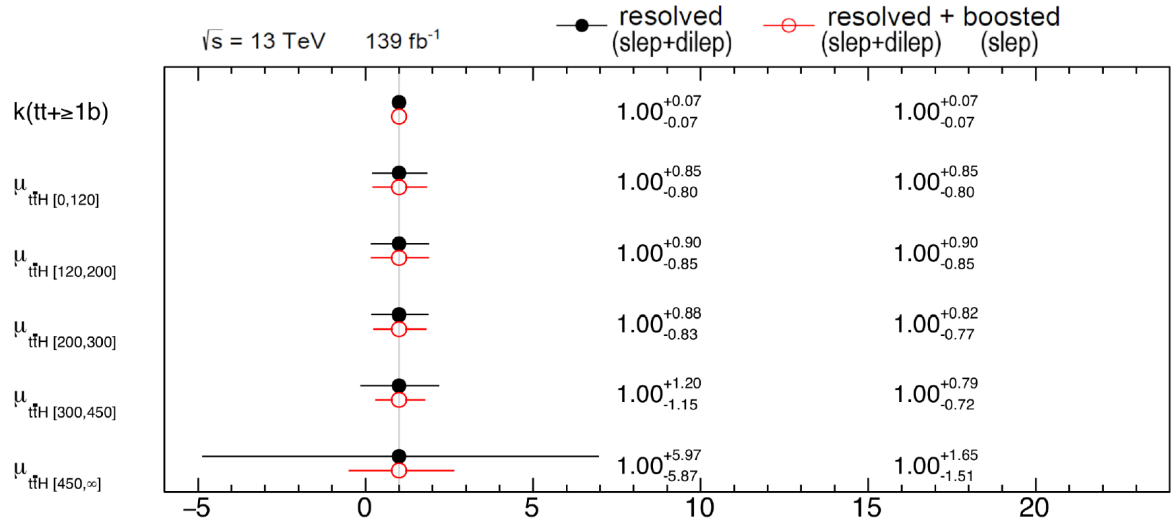


Fig. 8.13: Fitted normalisation factor for the $t\bar{t}+\geq 1b$ background and signal strength from the S+B fit to the Asimov dataset in the individual STXS $\hat{p}_T^H$ bins, combining only the resolved (single-lepton and dilepton) regions (in black) compared to the full combination (including also the single-lepton boosted region) (in red).

The correlation matrix of the nuisance parameters, the $k(t\bar{t}+\geq 1b)$ normalisation factor, and the signal strength parameters, containing the correlation coefficients of a selection of the aforementioned parameters obtained by the combined fit to data, is shown in Fig. 8.14. The correlations are pretty much similar to those from the inclusive cross-section measurement in a comparable amount. The only difference emerges from the fact that now there are multiple signal strength parameters, so the various nuisance parameters are potentially correlated to each one of them separately. As in the inclusive measurement, there are only a few large (anti-)correlations (e.g. $\geq 45\%$) among the different parameters and most of them are between the NPs related to the $t\bar{t}+\geq 1b$ background uncertainties, or specifically to the $t\bar{t}+\geq 1c$ normalisation uncertainty. Also, only a few NPs associated with the experimental uncertainties are present in the matrix, experiencing mostly subtle correlations among each other. A few strong correlations observed between some of the signal strength parameters and some NPs related to the $t\bar{t}+\geq 1b$ background modelling, may imply that these uncertainties could affect the sensitivity in the respective $\hat{p}_T^H$ bins.

Finally, the actual effect of the nuisance parameters is examined separately on each signal strength parameter, as illustrated in Fig. 8.15. Although some instrumental nuisance parameters show up in the ranking, related to the jet energy resolution and to the b-tagging efficiency or mistag rates, the dominant contributions still originate from the $t\bar{t} + \geq 1b$ background modelling. Also, the p_T^{bb} shape uncertainty becomes more dominant in the high STXS bins. This is possibly because the shape effect becomes more prominent in the high $\hat{p}_T^H$ bins by construction. Overall, the constraints of the various NPs are similar to those from the inclusive cross-section measurement. Only the NPs related to the $t\bar{t} + \geq 1b$ NLO matching are slightly less constrained in this fit.

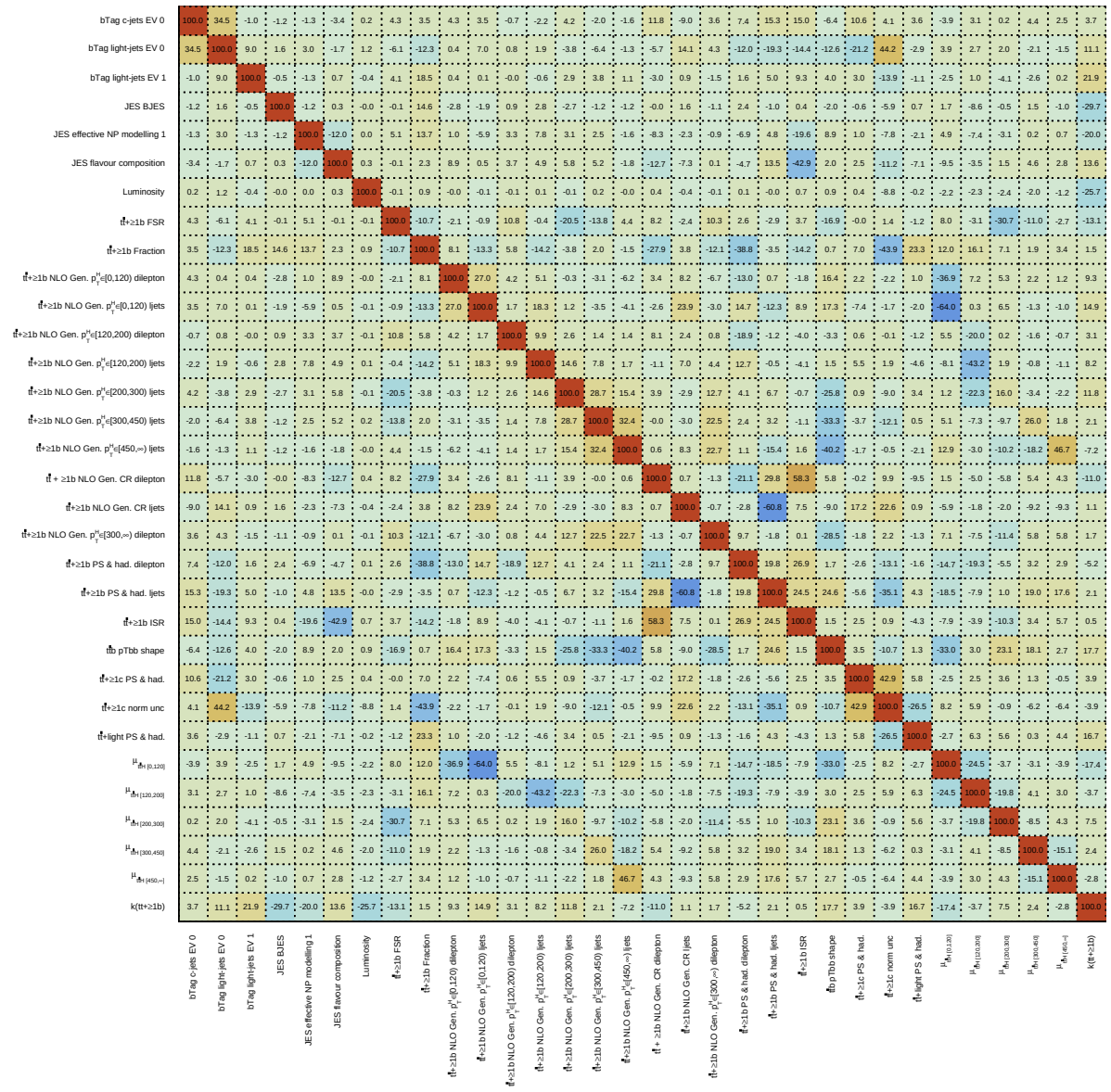


Fig. 8.14: Linear correlation coefficients of the nuisance parameters, normalisation factor, and signal strength, after the combined STXS fit to the Asimov dataset. All values are given in percent. Only nuisance parameters with at least one absolute correlation coefficient above 20% are shown.

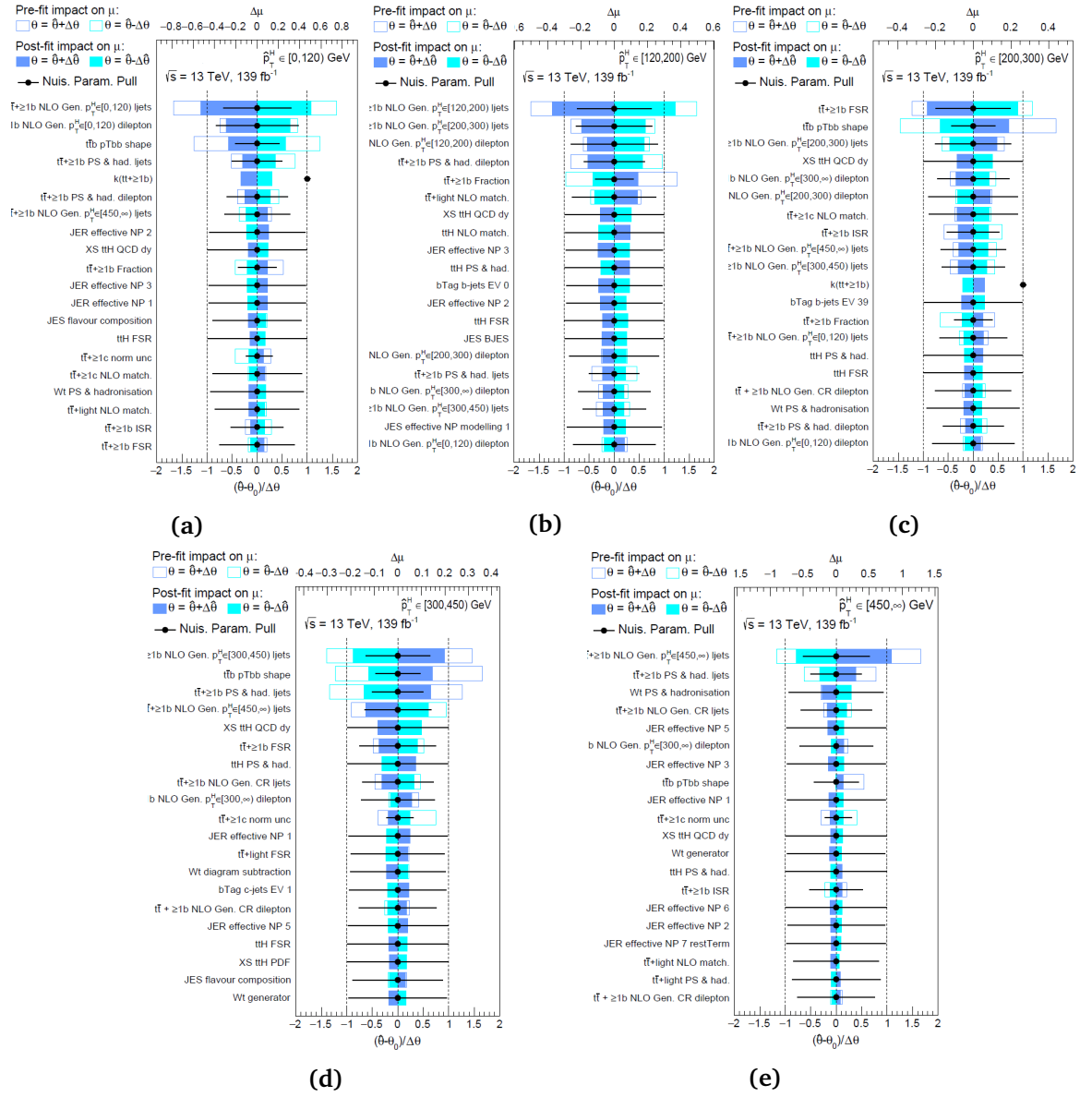


Fig. 8.15: Ranking of the 20 nuisance parameters with the largest post-fit impact on μ in the combined STXS fit to the Asimov dataset, for (a) $0 \leq \hat{p}_T^H < 120$ GeV, (b) $120 \leq \hat{p}_T^H < 200$ GeV, (c) $200 \leq \hat{p}_T^H < 300$ GeV, (d) $300 \leq \hat{p}_T^H < 450$ GeV, and (e) $\hat{p}_T^H \geq 450$ GeV. Nuisance parameters corresponding to statistical uncertainties in the MC simulated event samples are not included. The empty blue rectangles correspond to the pre-fit impact on μ, while the filled blue ones to the post-fit impact on μ, both referring to the upper scale. The impact of each nuisance parameter, $\Delta\mu$, is computed by comparing the nominal best-fit value of μ with the result of the fit when fixing the considered nuisance parameter to its best-fit value, $\hat{\theta}$, shifted by its pre-fit (post-fit) uncertainties $\pm\Delta\hat{\theta}$ ($\pm\Delta\theta$). The black points show the pulls of the nuisance parameters relative to their nominal values, $\theta_0 = 0$. These pulls and their relative post-fit errors, $\Delta\hat{\theta}/\Delta\theta$, refer to the lower scale. The parameter $k(t\bar{t} + \geq 1b)$ refers to the free-floating normalisation of the $t\bar{t} + \geq 1b$ background, for which the pre-fit impact on μ is not defined, while its nominal value is $\theta_0 = 1$. For the experimental uncertainties, that are decomposed into several independent sources, NP X corresponds to the Xth nuisance parameter, ordered by their impact on μ. The "ljets" ("dilep") label refers to the single-lepton (dilepton) channel.

8.3.2 Blinding strategy

As already introduced, in order to avoid biasing the signal sensitivity towards a certain result, the fit model is further optimised without looking at the data in regions sensitive to the signal. According to the blinding procedure, the bins with a large expected contribution of the $t\bar{t}H$ signal are removed from the fit and the data are not displayed in the plots. Specifically for this analysis, the signal over background contribution in every bin considered in the analysis was required to be lower than 5%, at the beginning, providing regions with a low sensitivity to the signal. Later, the blinding threshold was increased to 7.7%, which resulted in revealing one extra bin per region on average. This was done to ascertain that the results of the various verification and optimisation tests were also valid in part of the previously blinded bins, before the fit to the full data. The number of unblinded (revealed) bins (with the 7.7% threshold) of the corresponding discriminant distribution in each analysis region are listed in Table 7.5.

8.3.3 Background-only fit to blinded data

From the fit to the Asimov pseudo-data the expected performance of the fit model is evaluated. However, the Asimov fit naively assumes correctness of the nominal background model, and as such the significance derived from this fit might not reflect the result of the fit to the data in the full phase-space. In addition, given the mis-modelling of the $t\bar{t}$+jets background observed in the analysis, many corrections of the systematics model are required. Therefore, the real sensitivity and performance would differ significantly when fitting the full data.

In order to further validate the background model, a background-only fit is performed only in the unblinded bins. In this case, the profile likelihood follows the background-only (or null-signal) hypothesis which corresponds to signal strength $\mu = 0$. Thus, the $t\bar{t}H$ signal events and its modelling uncertainties are excluded from the likelihood function. Since the signal is omitted, the $k(t\bar{t}+ \geq 1b)$ normalisation is the only free-floating factor. Nonetheless, the background-only fit provides a good approximation of the background modelling and it is used to derive a more realistic estimation of the expected performance of the fit model. Eventually, the value of the $k(t\bar{t} + \geq 1b)$ normalisation factor from this fit is 1.29 ± 0.08, with 1% statistical uncertainty. Its total uncertainty is comparable with that obtained from the Asimov fit. Also, the (anti-)correlations among the different nuisance parameters parameters are along a similar line with those obtained from the Asimov fit.

In contrast to the Asimov fit, the fitted nuisance parameters can be shifted with respect to their nominal value, as depicted in Fig. 8.16. Almost none of the experimental (instrumental) systematics develop any constraints with respect to their prior uncertainties, while only a few of them are slightly pulled. In particular, the JES flavour composition NP and two b-tag related NPs (the first eigen-variation of the b-tagging mistag rate for c-/$light$-jets, "b-tag c-/$light$-jets EV 0"), are the largest pulls among them (but still quite small $< 0.2\sigma$). They are also slightly constrained, as it is the case also in the Asimov fit. Although the b-tagging plays a significant role in this analysis, no notable impact is expected on the sensitivity from these NPs, since the definition of the most sensitive signal regions require at least four b-tagged jets using tight b-tagging operating points. These operating points have very high rejection of $light$-jets as well as of c-jets, as mentioned in Sec. 5.3.1. Furthermore, the modelling systematics of the $t\bar{t}H$ signal process are neither constrained nor pulled, implying low sensitivity to the $t\bar{t}H$ modelling within the given precision. Also, the non-$t\bar{t}$+jets background modelling systematics are not constrained, as in the Asimov fit, though only a couple of them are slightly pulled.

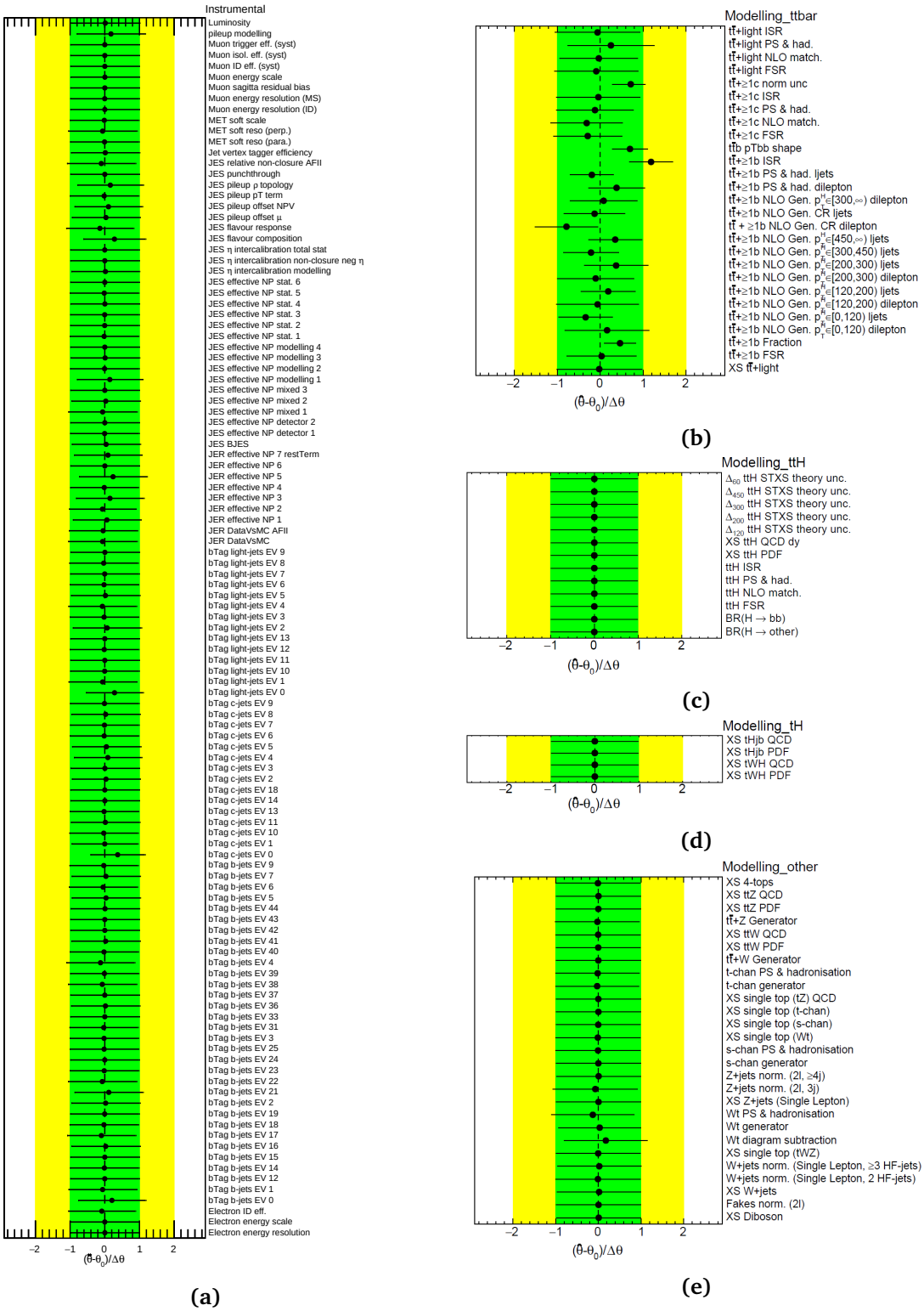


Fig. 8.16: Fitted nuisance parameters of the (a) instrumental, (b) $t\bar{t}$ + jets modelling, (c) $t\bar{t}H$ modelling, (d) tH modelling, and (e) other backgrounds modelling systematic uncertainties from the combined background-only fit to data. The green (yellow) area represents the $\pm 1(2)\sigma$ band on the pre-fit systematic uncertainty. The position of the black points shows the pull of the nuisance parameters, i.e. their best-fit value $\hat{\theta}$ relative to their nominal values, θ_0. The size of the horizontal bars give the constraint of the nuisance parameters, i.e. their post-fit error relative to the pre-fit one, $\Delta\hat{\theta}/\Delta\theta$. Both values are given in units of standard deviation (σ).

Moreover, most of the NPs related to the $t\bar{t}$+jets background are largely constrained, similarly to the Asimov fit, and are also pulled. In particular, those related to the $t\bar{t}$+light and $t\bar{t}+\geq 1c$ background modelling remain almost unconstrained and only a few of them are slightly pulled, given the small contribution of these background components in the analysis regions. The only exception is the $t\bar{t}+\geq 1c$ normalisation uncertainty which is intensely constraint and pulled, probably due to the inclusion of the dilepton CRs which contain a relatively large amount of $t\bar{t}+\geq 1c$ background. On the contrary, most of the $t\bar{t}+\geq 1b$ background systematics are strongly constrained and pulled. This is expected, since the most sensitive analysis regions are dominated by the $t\bar{t}+\geq 1b$ background, hence the model is able to constrain the various $t\bar{t}+\geq 1b$ modelling uncertainties and possibly compensate for any mis-modelling effects. Most of the pulls are still compatible with the nominal value 0 within 1σ, though. There are only a few exceptions where the pulls are large. The largest pull is observed in the $t\bar{t}+\geq 1b$ ISR systematic uncertainty, that mainly affects the distribution of the number of jets in the events for which a mis-modelling is observed pre-fit (Fig. 8.7). The other is on the reconstructed p_T^{bb} shape which is expected from the pre-fit mis-modelling (see Fig. 7.1) and the way this uncertainty is constructed (defined in Sec. 7.1.2). All the major pulls are thoroughly discussed after the fit to data in Sec. 8.4.1.

Although the background-only fit is a way to verify the performance of the fit model, it can still deviate from the complete $S+B$ fit to data. This is because the distributions of the model can be similar in the revealed bins but differ in the blinded bins, leading to different correlations and thus different post-fit values. Also, the complete absence of the signal from the model, even though its contribution in the revealed bins is low, can further bias the fit. For this reason, the background-only and the Asimov fits work as complementary for the optimisation of the analysis.

8.4 $S+B$ fit to data and results

After performing studies on pseudo-data or only part of the actual data, in order to assess the complete fit model, the fit to the full measured data can be carried out. As already introduced, to obtain the final result a profile-likelihood fit to data is performed under the $S+B$ hypothesis, simultaneously on the discriminant distributions of all analysis regions. The $k(t\bar{t}+\geq 1b)$ normalisation factor is determined from the fit together with the signal strength, μ. Also, the fit makes use of the flexibility introduced by the several nuisance parameters, each accounting for a specific physical effect associated with a systematic uncertainty. Eventually, the contributions from $t\bar{t}$+jets, W/Z+jets production, single top, diboson, and $t\bar{t}$V background process are constrained by the uncertainties of the respective theoretical calculations, the uncertainty on the luminosity, and the data themselves. Specifically, corrections from the systematics model arise to compensate for the mismodelling in the $t\bar{t}$+jets background observed in the pre-fit distributions. Both the inclusive cross-section and the STXS measurements are conducted and their fit results are reported in Sec. 8.4.1 and 8.4.2, respectively.

8.4.1 Inclusive cross-section measurement

The combined profile likelihood fit to data of the inclusive cross-section measurement for a Higgs boson mass of $m_{Higgs}=125$ GeV results in a best-fit $t\bar{t}H$ signal strength

$$\mu_{t\bar{t}H}=0.35\pm 0.20\,(\text{stat.})\,^{+0.30}_{-0.28}\,(\text{syst.})=0.35^{+0.36}_{-0.34}, \tag{8.15}$$

corresponding to an observed significance of 1.0σ (standard deviations) with respect to the background-only hypothesis. The significance is lower than the expected value of 2.7σ due to the low measured value of $\mu_{t\bar{t}H}$. The value of $\mu_{t\bar{t}H}$ is low, though still in agreement with the SM expectation of $\mu_{t\bar{t}H} = 1$ within an uncertainty of 1.8σ. The measured $t\bar{t} + \geq 1b$ background normalisation factor is found to be $k(t\bar{t} + \geq 1b) = 1.28 \pm 0.08$, in agreement with the background-only expectation (see Sec. 8.3.3). The statistical uncertainty is obtained by repeating the fit to data after fixing all nuisance parameters to their post-fit values, with the exception of the free normalisation factors in the fit, $k(t\bar{t} + \geq 1b)$ and $\mu_{t\bar{t}H}$. The total systematic uncertainty is obtained by subtracting the statistical variance from the total variance, i.e. $\sigma_{syst} = \sqrt{\sigma_{tot}^2 - \sigma_{stat}^2}$.

The measured inclusive signal strength is smaller than that obtained previously with 36.1 fb^{-1} of data [97] (Fig. 2.16b), but within their uncertainties. Also, the $k(t\bar{t} + \geq 1b)$ normalisation factor is in agreement with the one measured previously, while it is also compatible with the measured value obtained from an independent measurement of the $t\bar{t} + b\bar{b}$ process [290]. Although the expected sensitivity improved from an expected significance of 1.6σ to 2.7σ, the observed significance decreased from 1.4σ to 1.0σ due to the low signal strength. The current, as well as the previous, measurement is dominated by the systematic uncertainties (detailed in Sec. 7.1.2). However, the impact of the systematic uncertainties has been reduced by about a factor of two, as a consequence of a series of improvements. One of the major improvements comes from the enhanced theoretical knowledge in $t\bar{t} + \geq 1b$ modelling. In the previous publication, the $t\bar{t} + \geq 1b$ background was modelled with the 5FS and uncertainties incorporating the differences between the 4FS and 5FS were assigned, which had the second largest impact in the analysis. However, in the current analysis this uncertainty is not assessed, since the $t\bar{t} + \geq 1b$ background is modelled with the 4FS. Moreover, the improvements in the b-tagging calibration and the refined b-tagging scale factors, allowed for a better region definition. Other major improvements emerge from the much larger size of simulated event samples for systematic uncertainty estimation as well as the optimised jet energy scale and resolution measurements. Additionally, the detector reconstruction has changed significantly between the two measurements. Furthermore, the statistical uncertainty is also reduced by 31%, which is a smaller effect than expected though, considering the quadruple increase in statistics. This could be a consequence of the tighter selections applied to the current measurement in order to simplify the analysis. Finally, the inclusion of the single-lepton boosted region contributed in the overall improvement of the measurement, though its contribution is much larger and determining in the STXS measurement, as indicated from the Asimov fit (Sec. 8.3.1).

Provided that the signal strength is measured using data in different phase space regions (channels), it would be interesting to also review the best-fit results for the signal strength in each channel individually. For this purpose, a separate fit in each channel (single-lepton resolved and boosted, and dilepton channel) is done without considering any correlations among the channels. The obtained best-fit results are

$$\begin{aligned} \mu_{t\bar{t}H}^{\text{l+jets resolved}} &= 0.28^{+0.43}_{-0.41} \\ \mu_{t\bar{t}H}^{\text{l+jets boosted}} &= -0.74^{+1.28}_{-1.58} \\ \mu_{t\bar{t}H}^{\text{dilepton}} &= 0.57^{+0.68}_{-0.64} \end{aligned} \tag{8.16}$$

Especially when only fitting the boosted signal region, the measured signal strength is negative, indicating that the MC overestimates the background in the regions where the signal is expected. This can be caused by fluctuations of the MC samples. With a naive approach, the

best-fit μ values differ a lot with each other and with respect to the combined fit, they are all compatible within their uncertainties, though.

A better grounded way to evaluate the compatibility of the best-fit results for the signal strength among channels is by performing a combined fit, where each of the three channels has an individual signal strength. In this way, an estimate of how the signal in the individual channels would perform, if the background was constrained by all channels, is acquired. Nonetheless, the fit procedure is identical to the nominal combined fit, hence the nuisance parameters and $k(t\bar{t}+\geq 1b)$ factors are correlated among the channels. A likelihood ratio test can then be used, comparing the likelihood from the combined fit with the three signal strengths to the likelihood from the combined fit with the single signal strength. The former fit corresponds to the null hypothesis, which states that data in all channels can be described with a single signal strength. If data in all channels prefers very different signal strengths, one can reject the null hypothesis. The negative logarithmic ratio of the likelihoods asymptotically follows the χ^2 distribution (described in Sec. 8.1), with a number of degrees of freedom equal to the difference in degrees of freedom between the two likelihoods, and assuming the null hypothesis is true. In fact, the difference in the number of signal strengths is the number of degrees of freedom to be used in the χ^2 test. In this way, the p-value quantifying compatibility with the null hypothesis is derived. To conclude, the probability of obtaining a discrepancy between the best-fit results of these two fits, equal to or larger than the one observed, corresponds to the probability of obtaining a χ^2 value at least as large.

Figure 8.17 shows the $\mu_{t\bar{t}H}$ value obtained for each channel from a combined fit as well as the signal strength resulting from the inclusive-μ combined fit. Apparently, the single-μ values are compatible with the inclusive-μ obtained from the nominal fit within their uncertainties. The signal strength from the single-lepton resolved has the smallest uncertainty and is thus the most sensitive. The $t\bar{t}+\geq 1b$ normalisation factor from this fit results in $k(t\bar{t}+\geq 1b) = 1.27\pm 0.08$, which is in almost perfect agreement with the value from the inclusive-μ fit. The probability of obtaining a discrepancy among the signal strengths from these two fits equal to or larger than the one observed amounts to 90%. Moreover, the probability of the obtained signal strength from inclusive-μ combined fit being compatible with the SM prediction is 8.5%, estimated by redoing the fit while fixing $\mu_{t\bar{t}H} = 1$.

Looking at the uncertainties on the three $\mu_{t\bar{t}H}$ values in Fig. 8.17, one observes that the impact of systematic uncertainties on the result is larger than that of the statistical uncertainty, as it holds for the inclusive-μ result (eq. 8.15). However, this is not valid for the boosted channel, which by construction is statistically dominated given the low statistics. Also, the impact of systematic uncertainties in the dilepton channel is quite large with respect to the single-lepton resolved channel. This is because the latter has more statistics on data, and thus more constraining power than the former. In fact, before having decorrelated the two-point systematics between the dilepton and single-lepton channels, as described in Sec. 7.1.2, the systematic uncertainties in the dilepton channel were smaller than the current ones. That was because the dilepton channel could benefit from the constraining power of the single-lepton resolved channel thanks to the correlations.

Furthermore, comparing the respective $\mu_{t\bar{t}H}$ values from the individual fits (eq. 8.16) to the single-μ values from the combined fit (Fig. 8.17), one observes that the former are smaller than the latter in all three cases. Though, they are still compatible with each other within their uncertainties. In particular, while the results of the single-lepton resolved and dilepton channels are very similar, the result of the single-lepton boosted channel changes drastically

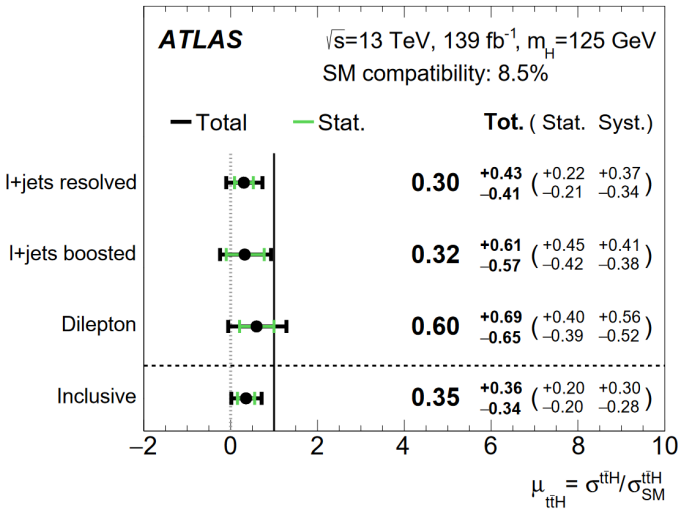


Fig. 8.17: Fitted values of the $t\bar{t}H$ signal-strength parameter from the S+B fit to data in the individual channels (single-lepton resolved, single-lepton boosted, and dilepton) and in the inclusive-mu measurement by combining the three channels [100].

between these two fits. As already discussed, its individual fit gives a negative signal strength, and a more than 100% uncertainty is associated to it with regard to $\mu_{t\bar{t}H} = 1$. This probably ensues from the fact that the boosted region targets the high-p_T^H phase space, thus it can not effectively correct the $t\bar{t}$+jets mismodelling, which mostly occurs in the low-p_T^H region. In parallel, this entails that the boosted channel needs constraints on the $t\bar{t}$+jets background and on the $k(t\bar{t}+\geq 1b)$ factor from the other channels. Overall, the observed differences are caused by correlations of the NPs affecting the different channels, which are not taken into account when fitting the channels separately. Then, the combination of the channels helps to mitigate some of the shortcomings that each channel individually has.

As already pointed out, the NPs are included in the profile likelihood fit as uncorrelated parameters among each other, though the fit creates correlations between complementary NPs. Large correlations can point to unnecessary degrees of freedom which are covered by other parameters, or to some variables being too correlated to the signal, making the analysis less sensitive. The correlation matrix of the nuisance parameters, the $k(t\bar{t}+\geq 1b)$ normalisation factor, and the signal strength, containing the correlation coefficients of a selection of the systematic uncertainty sources obtained by the combined inclusive fit to data, is shown in Fig. 8.18. In general, there are only a few large (anti-)correlations (e.g. $\geq 45\%$) among the different parameters and they are comparable to those observed in the Asimov and background-only fits. The largest correlation is between the $t\bar{t}+\geq 1b$ ISR systematic and the $t\bar{t}+\geq 1b$ NLO Gen. (matching) CR dilepton systematic (65.3%). In parallel, the former systematic source is also highly anti-correlated with the experimental source JES flavour composition (-45.9%). Then, the the $t\bar{t}+\geq 1c$ normalisation uncertainty is strongly correlated with $t\bar{t}+\geq 1c$ PS & hadronisation NP (50.9%), but it is also highly anti-correlated with the $t\bar{t}+\geq 1b$ Fraction (-46.9%). The largest anti-correlation comes from the $t\bar{t}+\geq 1b$ NLO Gen. CR ljets NP with the $t\bar{t}+\geq 1b$ PS & hadronisation (-64.3%).

As already highlighted, the most important correlations are the ones between the nuisance parameters and the signal strength ($\mu_{t\bar{t}H}$), since they affect the sensitivity of the analysis. The most noticeable is the anti-correlation between the signal strength $\mu_{t\bar{t}H}$ with the $t\bar{t}+\geq 1b$ NLO

Gen. parameter in the single-lepton $p_T^H \in [0, 120)$ GeV bin (-47.9%) and the next largest is with the $t\bar{t} + \geq 1b$ NLO Gen. in single-lepton $p_T^H \in [120, 200)$ GeV (-37%). This implies that these systematics have similar features to the signal, and it is thus difficult to separate the signal from the $t\bar{t} + \geq 1b$ background in these regions. Given that the most sensitive analysis regions are dominated by the $t\bar{t} + \geq 1b$ background and considering its observed MC mis-modelling, these strong anti-correlations have dominant impact on the signal strength uncertainty, leading to lower sensitivity. The actual effect of the various systematic uncertainties on $\mu_{t\bar{t}H}$ is illustrated in Fig. 8.22 discussed later.

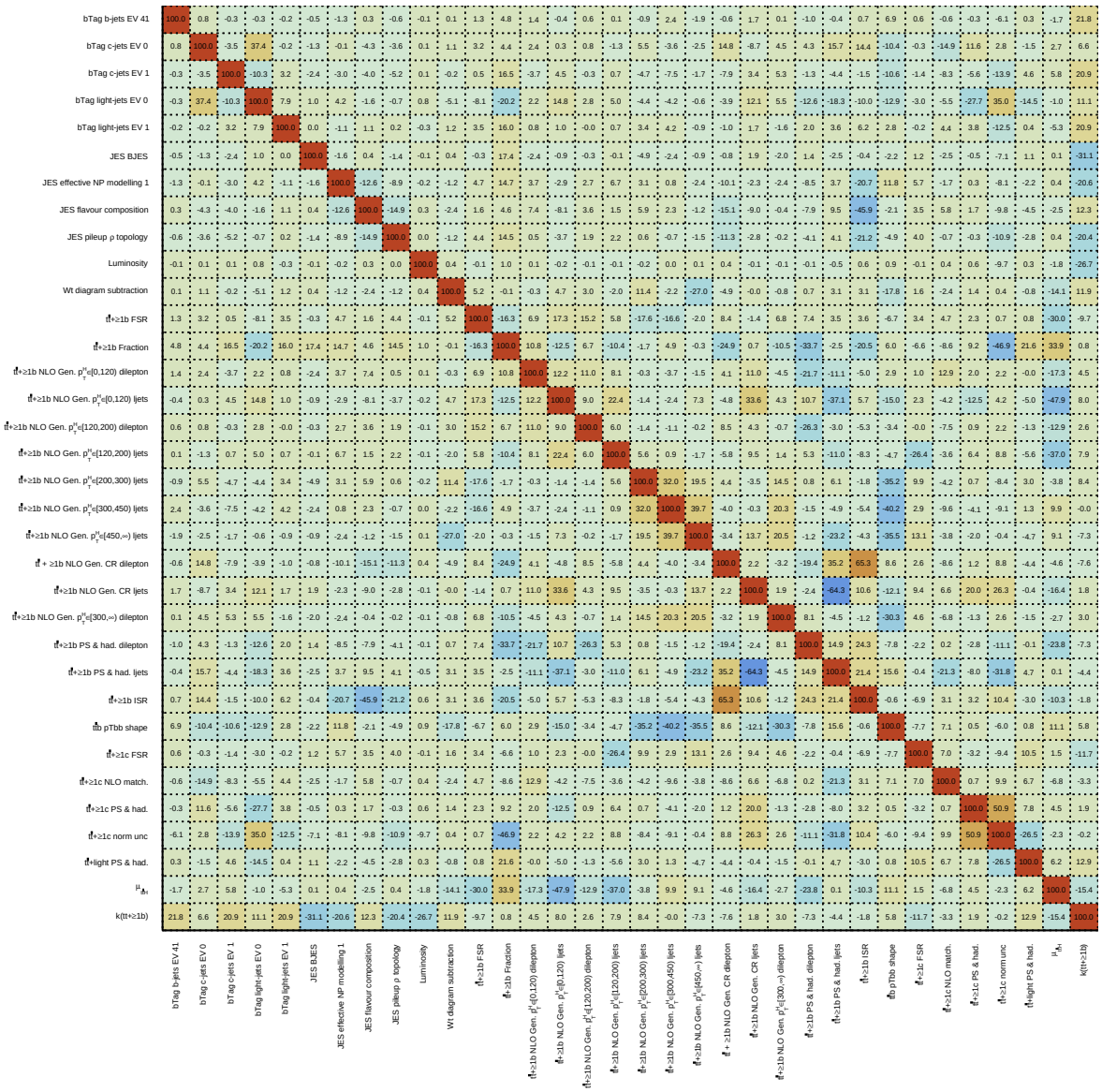


Fig. 8.18: Linear correlation coefficients of the nuisance parameters, normalisation factor, and signal strength, after the combined inclusive-μ fit to data. All values are given in percent. Only nuisance parameters with at least one absolute correlation coefficient above 20% are shown.

Figure 8.19 shows the fitted nuisance parameters in the S+B combined fit on data. Some nuisance parameters are pulled from their nominal values after the fit to data. Only a few of the instrumental (experimental) NPs develop slight pulls and/or subtle constraints, as the NPs

of the non-$t\bar{t}$ background systematics do. By contrast, most of the NPs related to the $t\bar{t}$+jets background exhibit larger pulls and constraints, apart from those related to the $t\bar{t}$+light background modelling which, given its small contribution in the analysis regions, remain almost unconstrained. Besides, most of the pulls are still compatible with the nominal value 0 within 1σ. The major exceptions, where the pulls are large, are discussed below, although it is difficult to explain their source or effect, due to the high number of NPs involved in the analysis and the correlations among them. The modelling $t\bar{t}H$ signal systematics are not constrained at all, indicating low sensitivity to the $t\bar{t}H$ modelling within the given precision. The observed pulls and constraints are comparable to those obtained from the background-only fit.

The largest observed pull on systematic uncertainties is seen in the $t\bar{t}+\geq 1b$ ISR uncertainty and is about 1.2σ. Thorough studies have shown that the pull mostly stems from the renormalisation scale. This pull indicates that the data favours a softer renormalisation scale in the matrix element calculation (Sec. 4.2.2), as also suggested in reference [291], where a lower scale in $t\bar{t}b\bar{b}$ calculations gives better agreement with $t\bar{t}b\bar{b}j$ calculations. In the future, it would be preferable to produce a nominal sample with the renormalisation scale shifted like that, instead of having such a large pull which has a penalty in the likelihood. Extensive studies were performed in order to understand the impact of this pull on the background model and the fit results. In particular, this effect is shown to not affect the shape of the BDT distributions used as input for the fit in each individual region. On the contrary, it significantly corrects the mismodelling of extra radiation in $t\bar{t}+\geq 1b$ events observed in the distribution of the number of jets in the event, by adjusting the amount of additional radiation, which affects the categorisation of events. The pre-fit distribution of the number of jets in the three SRs is already shown in Fig. 8.7, while the post-fit one is in Fig. 8.20. Decorrelating the $t\bar{t}+\geq 1b$ ISR uncertainty in its different components (μ_R, μ_F, a_R^{ISR}), and assigning them independent nuisance parameters, had minimal impact on the fit results. Also, when softer scales are used (scaling μ_R and μ_F by a factor 0.5), only the $t\bar{t}+\geq 1b$ ISR NP changed with respect to the nominal fit. Moreover, decorrelating this uncertainty between the dilepton and single-lepton channels leads to very similar fitted $\mu_{t\bar{t}H}$ values and NP pulls. Finally, the $t\bar{t}+\geq 1b$ ISR pull may cause the pulls in the $t\bar{t}+\geq 1b$ NLO matching uncertainty in the dilepton CRs (-0.7σ) and in the JES flavour composition (0.2σ), on account of their strong (anti-)correlations mentioned earlier.

The second largest pull is observed in the $t\bar{t}+\geq 1b$ NLO matching uncertainty in the dilepton SR in the $0 \leq p_T^H < 120$ GeV bin ($\sim 1\sigma$). Several studies were performed in order to understand the impact of this pull on the background model and on signal strength. Namely, comparisons between the nominal fit and alternative models were made, where the NP related to this uncertainty is fixed to its nominal value, or where this NP is correlated with the corresponding NP from the single-lepton channel, or it is left free-floating. The alternative models show no striking difference in the pulls, while they have a small impact on the $\mu_{t\bar{t}H}$ value. Nevertheless, the observed changes on $\mu_{t\bar{t}H}$ in these tests are within the contribution of the NLO matching uncertainty on μ, which amounts to $[-0.20, +0.21]$ (see Table 8.1). The $t\bar{t}+\geq 1b$ NLO generator matching uncertainty is fully decorrelated across channels and signal regions, in order to limit the size and propagation of constraints from one region to another. Then, an additional study was performed to understand the statistical significance of this pull. It was shown that, due to the large number of nuisance parameters entering into the likelihood fit, there is a non-negligible probability that any one of the $t\bar{t}+\geq 1b$ NLO matching uncertainties could be pulled to a value at least as extreme as that observed from a purely statistical

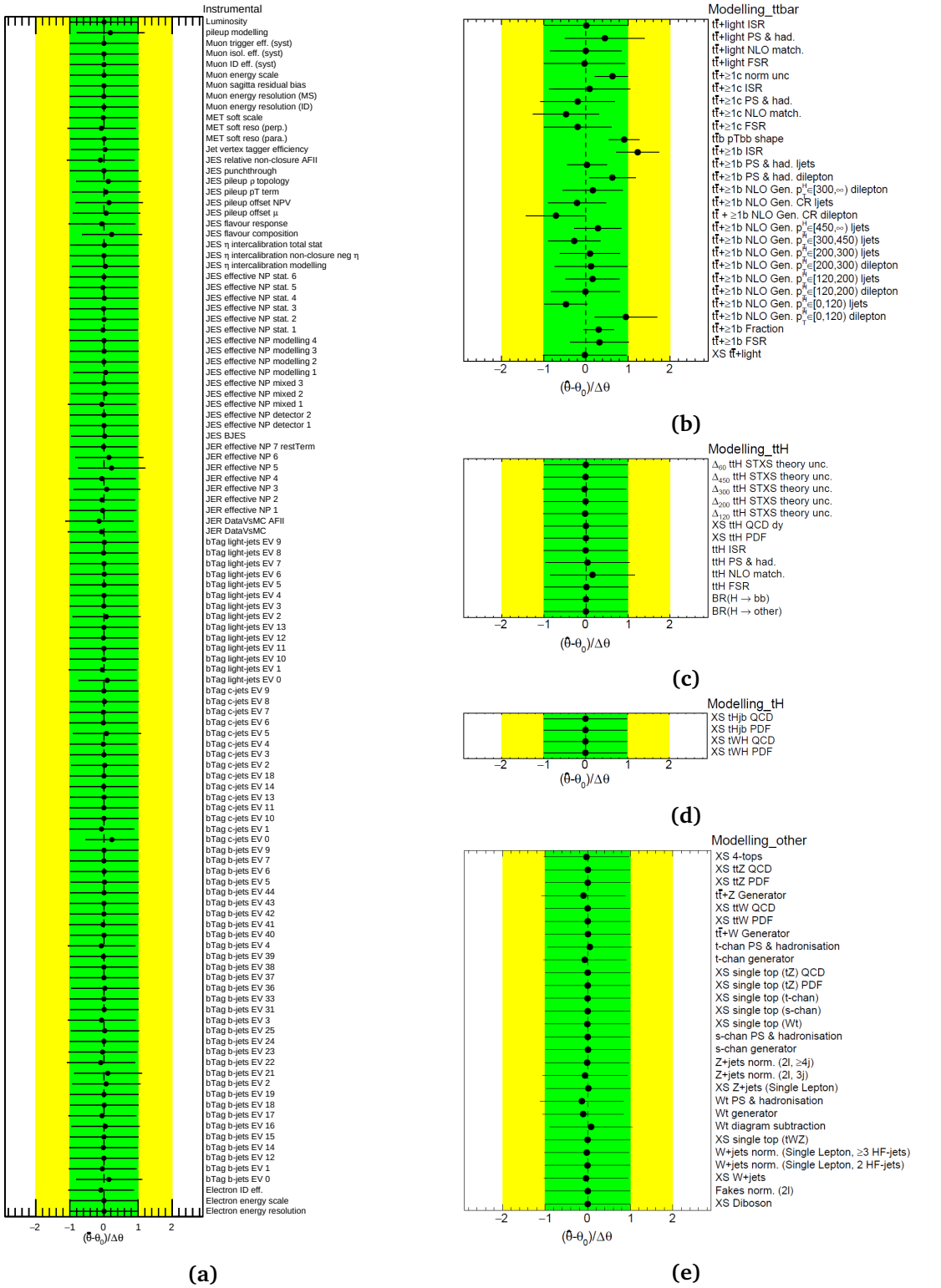


Fig. 8.19: Fitted nuisance parameters of the (a) instrumental, (b) $t\bar{t}$ + jets modelling, (c) $t\bar{t}H$ modelling, (d) tH modelling, and (e) other backgrounds modelling systematic uncertainties from the combined inclusive cross-section fit to data. The green (yellow) area represents the $\pm 1(2)\sigma$ band on the pre-fit systematic uncertainty. The position of the black points shows the pull of the nuisance parameters, i.e. their best-fit value $\hat{\theta}$ relative to their nominal values, θ_0. The size of the horizontal bars give the constraint of the nuisance parameters, i.e. their post-fit error relative to the pre-fit one, $\Delta\hat{\theta}/\Delta\theta$. Both values are given in units of standard deviation (σ).

standpoint. The probability is evaluated to be 17%. Considering also the high goodness-of-fit value in the nominal analysis (reported below), which shows that the fit model has sufficient degrees of freedom to model the data within the assigned uncertainties, it can be concluded that the data could create this pull due to the limited number of expected events.

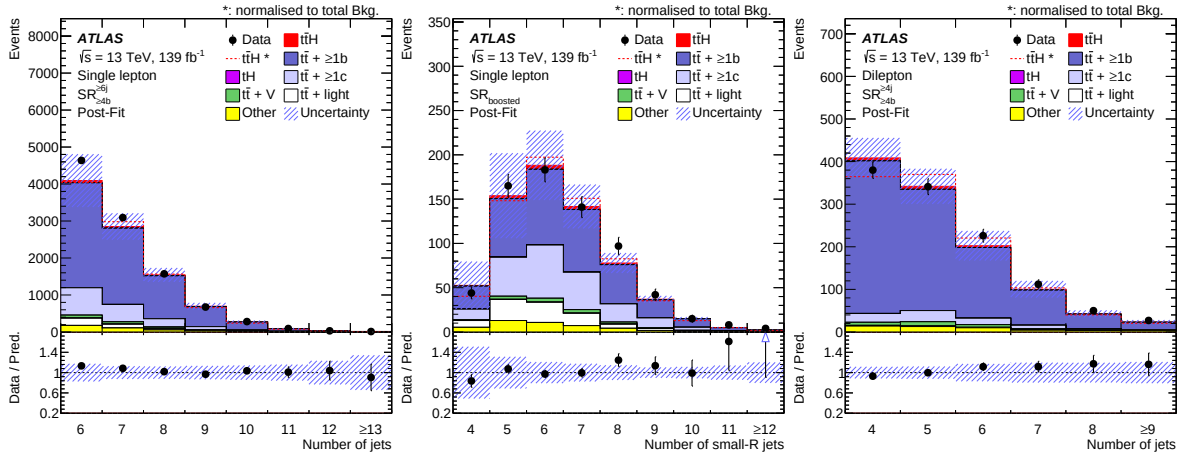
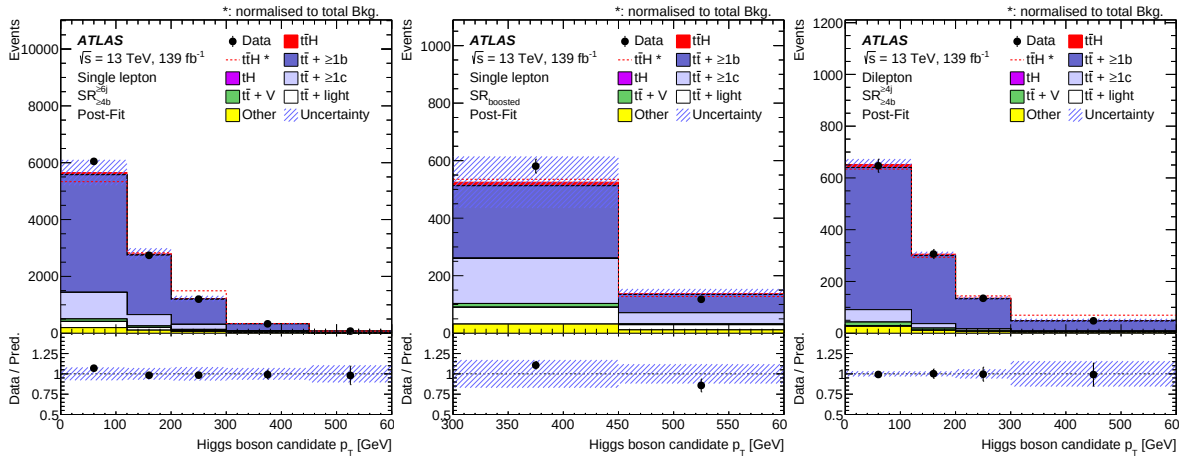


Fig. 8.20: Post-fit distributions of the number of jets for the (a) single-lepton resolved $SR^{\geq 6j}_{\geq 4b}$, (b) single-lepton boosted $SR_{boosted}$, and (c) dilepton $SR^{\geq 4j}_{\geq 4b}$ signal regions [100]. The $t\bar{t}H$ signal yield (solid red) is normalised to the fitted μ value from the inclusive fit. The dashed line shows the $t\bar{t}H$ signal distribution normalised to the total background prediction. The uncertainty band includes all uncertainties as well as their correlations. The last bin includes the overflow.

Another large pull is observed on the reconstructed p_T^{bb} shape uncertainty in the $t\bar{t}+\geq 1b$ background ($\sim 0.9\sigma$), as expected from the pre-fit mismodelling (see Fig. 7.1) and the way this uncertainty is defined. Namely, a $+1\sigma$ variation is derived such that it corrects the reconstructed p_T^H shape, so that it agrees between data and the background model (defined in Sec. 7.1.2), without changing the overall $t\bar{t}+\geq 1b$ background normalisation. The sensitivity of the result to this uncertainty was tested by replacing the data-driven mismodelling with decorrelated free-floating $t\bar{t}+\geq 1b$ normalisation factors across the STXS bins and analysis regions. Nonetheless, no bias was observed on the fitted signal strength. The reconstructed p_T^H distributions display good post-fit agreement between data and simulation, as depicted in Fig. 8.21 (compared to the pre-fit discrepancies shown in Fig. 7.1).

Moreover, despite the large differences between models (see Sec. 7.1.2), the varying $t\bar{t}+\geq 1b$ composition in the CRs and SRs allows the $t\bar{t}+\geq 1b$ subcomponent fraction systematic uncertainty to be constrained. In general, the fit mostly constrains the $t\bar{t}+\geq 1b$ modelling uncertainties as well as the normalisation of the $t\bar{t}+\geq 1c$ background, which is also pulled to 0.6σ. The constraint on the $t\bar{t}+\geq 1c$ normalisation uncertainty comes from the inclusion of the dilepton CRs, which contain a relatively large contribution of the $t\bar{t}+\geq 1c$ background and result in a better estimation of the $t\bar{t}+\geq 1c$ normalisation. In the previous $t\bar{t}H(H\rightarrow b\bar{b})$ publication, the $t\bar{t}+\geq 1c$ normalisation factor was a free-floating parameter in the fit, with a best-fit value of $k(t\bar{t}+\geq 1c) = 1.63 \pm 0.23$ [97]. On the contrary, in the analysis presented here, the $k(t\bar{t}+\geq 1c)$ is not free-floating, which is now reflected in this pull. Eventually, the measured value of the $t\bar{t}+\geq 1c$ normalisation is in agreement with that obtained in the



(a) **(b)** **(c)**

Fig. 8.21: Post-fit distributions of the reconstructed Higgs boson candidate for the (a) single-lepton resolved $SR^{\geq 6j}_{\geq 4b}$, (b) single-lepton boosted $SR_{boosted}$, and (c) dilepton $SR^{\geq 4j}_{\geq 4b}$ signal regions [100]. The $t\bar{t}H$ signal yield (solid red) is normalised to the fitted μ value from the inclusive fit. The dashed line shows the $t\bar{t}H$ signal distribution normalised to the total background prediction. The uncertainty band includes all uncertainties as well as their correlations. The last bin includes the overflow.

previous measurement as well as in the independent $t\bar{t}b\bar{b}$ measurement [290]. Lastly, the pull on the $t\bar{t}+\geq 1c$ normalisation uncertainty corrects for the normalisation discrepancy in the distribution of the number of jets in the event, by increasing the $t\bar{t}+\geq 1c$ contribution. The $k(t\bar{t}+\geq 1b)$ best-fit value also contributes to compensate for the disagreement on the normalisation of this distribution.

As already remarked, the $t\bar{t}H(H\rightarrow b\bar{b})$ measurement is largely dominated by the systematic uncertainties. Specifically, the modelling of the $t\bar{t}b\bar{b}$ background process is the primary source of uncertainties in this measurement. The actual pre-fit and post-fit effect of the nuisance parameters on $\mu_{t\bar{t}H}$ is illustrated in Fig. 8.22. It shows the twenty most important sources of systematic uncertainty, ranked based on the size of their impact on the signal strength after the fit. Additionally, the plot shows the constraints and pulls of the systematics variations, similarly to the plots in Fig. 8.19. The ranking plot reflects the correlation matrix, i.e. the seven highest-ranked nuisance parameters are also depicted in Fig. 8.18 with a post-fit impact on signal strength comparable to the absolute amount of their correlation. These NPs are all associated to the $t\bar{t}+\geq 1b$ background modelling. In total, all the $t\bar{t}+\geq 1b$ related NPs have smaller effect on $\mu_{t\bar{t}H}$ after the fit compared to what they have before, indicating the constraining power of the fit model particularly on the $t\bar{t}+\geq 1b$ background. The post-fit impact is configured also according to the correlations among the NPs, which are not present pre-fit.

Especially, the dominant post-fit impact on $\mu_{t\bar{t}H}$ comes from the $t\bar{t}+\geq 1b$ NLO generator matching in the two lowest-p_T^H bins of the single-lepton region. The fact that they are both quite shifted compared to their nominal value and are also constrained, demonstrates their importance to the background modelling. Indicatively, by shifting the highest ranked systematic by 1σ of its post-fit uncertainty, the $\mu_{t\bar{t}H}$ is shifted by around 15%. In fact, the large variation of these nuisance parameters implies a mismodelling in the matching of the NLO matrix element to the parton shower, that may not be easily interpreted. Given that this systematic

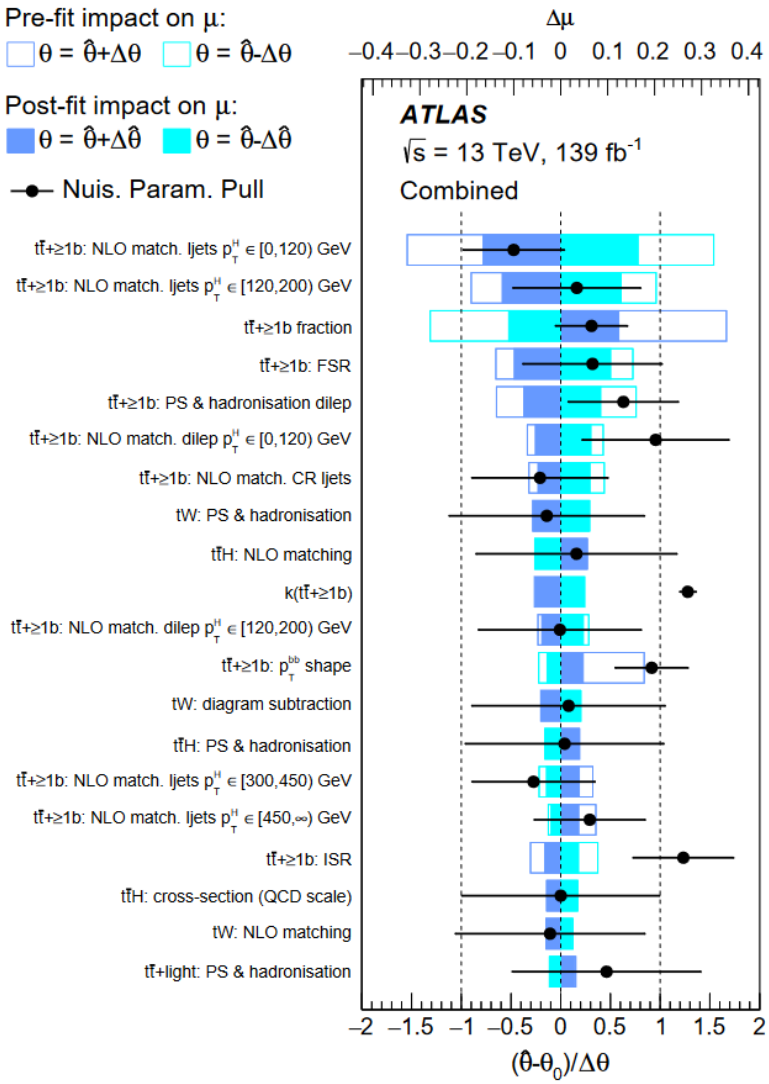


Fig. 8.22: Ranking of the 20 nuisance parameters with the largest post-fit impact on μ in the combined inclusive cross-section fit to data [100]. Nuisance parameters corresponding to statistical uncertainties in the MC simulated event samples are not included. The empty blue rectangles correspond to the pre-fit impact on μ, while the filled blue ones to the post-fit impact on μ, both referring to the upper scale. The impact of each nuisance parameter, $\Delta\mu$, is computed by comparing the nominal best-fit value of μ with the result of the fit when fixing the considered nuisance parameter to its best-fit value, $\hat{\theta}$, shifted by its pre-fit (post-fit) uncertainties $\pm\Delta\hat{\theta}$ ($\pm\Delta\theta$). Usually, the post-fit value of $\Delta\hat{\theta}$ is smaller than the pre-fit $\Delta\theta$ due to constraints that arise from the fit. The black points show the pulls of the nuisance parameters relative to their nominal value $\theta_0 = 0$, that are defined as $\frac{\hat{\theta}-\theta_0}{\Delta\theta}$. These pulls and their relative post-fit errors, $\Delta\hat{\theta}/\Delta\theta$, refer to the lower scale. The parameter $k(t\bar{t} + \geq 1b)$ refers to the free-floating normalisation of the $t\bar{t} + \geq 1b$ background, for which the pre-fit impact on μ is not defined, while its nominal value is $\theta_0 = 1$. The "ljets" ("dilep") label refers to the single-lepton (dilepton) channel.

variation is defined by the comparison of two $t\bar{t}$+jets models (comparing the generators MADGRAPH5_aMC@NLO+PYTHIA8 with POWHEGBOX+PYTHIA8), it could be overestimated and not precisely describe NLO matching of the $t\bar{t}b\bar{b}$ process in the matrix element. Moreover, due to its definition, the negative value of the nuisance parameter does not directly correspond to a specific model. Therefore, producing proper alternative $t\bar{t}b\bar{b}$ models, with the b-quarks included in the matrix element, to assess the NLO matching uncertainty, could be used to also investigate further this effect. However, while preparing this analysis, no alternatives with sufficient statistics were available.

Besides the uncertainties from the $t\bar{t}+\geq 1b$ modelling, also tW background and $t\bar{t}H$ signal modelling related NPs show up in the ranking. However, their post-fit impact is small compared to that of $t\bar{t}+\geq 1b$ NPs, while it is almost unchanged from their pre-fit impact. In fact, after the first few highly ranked NPs, the post-fit effect of the remaining systematic variations is relatively similar, though with a slowly decreasing effect. Most of them are not significantly pulled or constrained, meaning their actual importance to the modelling is not as large. Lastly, no NPs related to experimental systematic uncertainties are listed among them, confirming the low impact of experimental uncertainties on the result of the measurement.

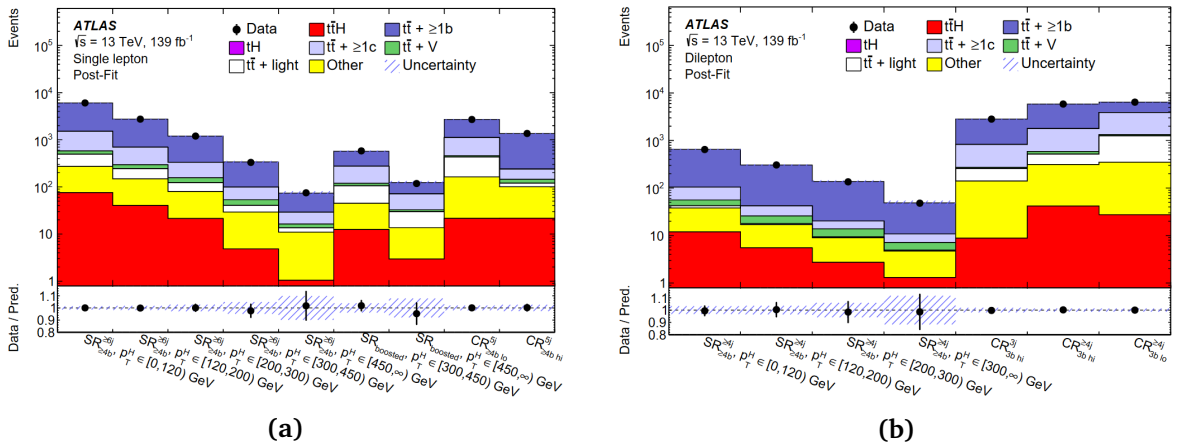
In addition, the impact of systematic uncertainty sources on signal strength is evaluated in groups, according to their contribution to the total uncertainty, and listed in Table 8.1. A consistent picture is drawn, as the dominant impact comes from the $t\bar{t}+\geq 1b$ background modelling, followed by the signal modelling, tW background modelling and flavour-tagging uncertainties. The "background-model statistical uncertainty" refers to the statistical uncertainties on the available MC statistics for the background. In fact, the uncertainty on the background MC statistics is of similar size as the flavour-tagging uncertainties. In the previous round of the analysis [97], the low statistics of the MC background samples had a large impact on the result. In the presented analysis though, the statistics has been significantly increased having a minimal effect on the measurement, which can be further reduced by generating more events. Interestingly, the combined impact of the $t\bar{t}H$ modelling variations has larger impact on the positive variation of $\mu_{t\bar{t}H}$ than on the negative, as it was also observed in the previous measurement [97].

After the combined fit to data is performed (post-fit), the prediction is adjusted according to the fit results. Thus, the agreement between the data and the MC simulated events in the various distributions should be revisited. In the post-fit case, the values of the signal strength and the $k(t\bar{t}+\geq 1b)$ normalisation from the nominal fit results are applied, while their uncertainties are now included in the overall uncertainty. Also, the post-fit uncertainties take into account the correlations of all nuisance parameters and their constraints.

In Fig. 8.23, a summary of the predicted signal and background events compared to the observed yields in all SRs and CRs after the fit to data are depicted. Moreover, the classification BDT distributions in the single-lepton and dilepton SRs, as well as the ΔR_{bb}^{avg} in the single-lepton resolved CRs and the event yield in the dilepton CRs, are presented in Fig. 8.25 and 8.24, respectively. The level of agreement between prediction and data is improved after the fit, due to the parameters of interest and NPs being adjusted by the fit, correcting for the MC deficit in background normalisation observed in several of these regions pre-fit (Sec. 8.2.2). Additionally, the post-fit uncertainty is significantly reduced as a result of the NP constraints and the correlations generated by the fit. Also, it is observed that the uncertainties increase as a function of p_T^H, ranging from 2% to 12%, but are still smaller than pre-fit. Therefore, the precision increases post-fit due to profiling and all distributions are compatible with the data.

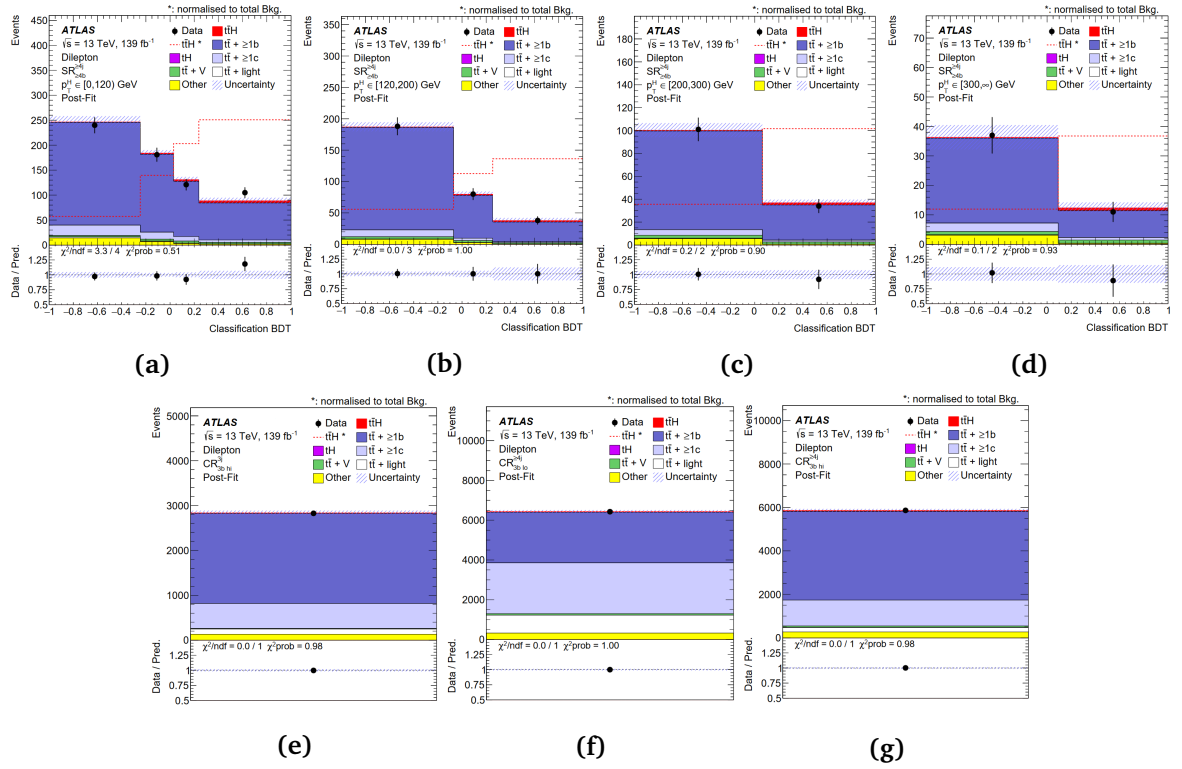
Uncertainty source	$\Delta\mu$	
Process modelling		
$t\bar{t}H$ modelling	$+0.13$	-0.05
$t\bar{t}+\geq 1b$ modelling		
$t\bar{t}+\geq 1b$ NLO matching	$+0.21$	-0.20
$t\bar{t}+\geq 1b$ fractions	$+0.12$	-0.12
$t\bar{t}+\geq 1b$ FSR	$+0.10$	-0.11
$t\bar{t}+\geq 1b$ PS & hadronisation	$+0.09$	-0.08
$t\bar{t}+\geq 1b$ p_T^{bb} shape	$+0.04$	-0.04
$t\bar{t}+\geq 1b$ ISR	$+0.04$	-0.04
$t\bar{t}+\geq 1c$ modelling	$+0.03$	-0.04
$t\bar{t}$ + light modelling	$+0.03$	-0.03
tW modelling	$+0.08$	-0.07
Background-model statistical uncertainty	$+0.04$	-0.05
b-tagging efficiency and mis-tag rates		
b-tagging efficiency	$+0.03$	-0.02
c-mis-tag rates	$+0.03$	-0.03
l-mis-tag rates	$+0.02$	-0.02
Jet energy scale and resolution		
b-jet energy scale	$+0.00$	-0.01
Jet energy scale (flavour)	$+0.01$	-0.01
Jet energy scale (pile-up)	$+0.00$	-0.01
Jet energy scale (remaining)	$+0.01$	-0.01
Jet energy resolution	$+0.02$	-0.02
Luminosity	$+0.01$	-0.00
Other sources	$+0.03$	-0.03
Total systematic uncertainty	$+0.30$	-0.28
$t\bar{t}+\geq 1b$ normalisation	$+0.04$	-0.07
Total statistical uncertainty	$+0.20$	-0.20
Total uncertainty	$+0.36$	-0.34

Table 8.1: Breakdown of the contributions to the uncertainties in μ [100]. The contributions from the different sources of uncertainty are evaluated after the fit. The $\Delta\mu$ values are obtained by repeating the fit after having fixed a certain set of nuisance parameters corresponding to a group of systematic uncertainties, and then evaluating $(\Delta\mu)^2$ by subtracting the resulting squared uncertainty of μ from its squared uncertainty found in the full fit. The same procedure is followed when quoting the effect of the $t\bar{t}+\geq 1b$ normalisation. The total uncertainty is different from the sum in quadrature of the different components due to correlations between nuisance parameters existing in the fit.



(a) (b)

Fig. 8.23: Comparison of the predicted and observed event yields in each of the signal and control regions, in the (a) single-lepton and (b) dilepton channels, after performing the combined fit to data (post-fit) [100]. The uncertainty band includes the statistical and systematic uncertainties as well as their correlations.



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

(e) (f) (g)

Fig. 8.24: Comparison between data and MC prediction for the BDT discriminant and the event yield in the SRs and CRs of the dilepton channel, respectively, after performing the inclusive fit to data (post-fit) [100]. The signal region, $\mathrm{SR}^{\geq 4j}_{\geq 4b}$, split into (a) $0 \leq p_T^H < 120$ GeV, (b) $120 \leq p_T^H < 200$ GeV, (c) $200 \leq p_T^H < 300$ GeV, (d) $p_T^H \geq 300$ GeV regions, as well as the control regions (e) $\mathrm{CR}^{3j}_{\geq 3b\ hi}$, (f) $\mathrm{CR}^{4j}_{\geq 3b\ lo}$, and (g) $\mathrm{CR}^{4j}_{\geq 3b\ hi}$ are shown. The $t\bar{t}H$ signal yield (solid red) is normalised to the fitted μ value from the inclusive fit. The dashed line shows the $t\bar{t}H$ signal distribution normalised to the total background prediction. The uncertainty band includes the statistical and systematic uncertainties as well as their correlations.

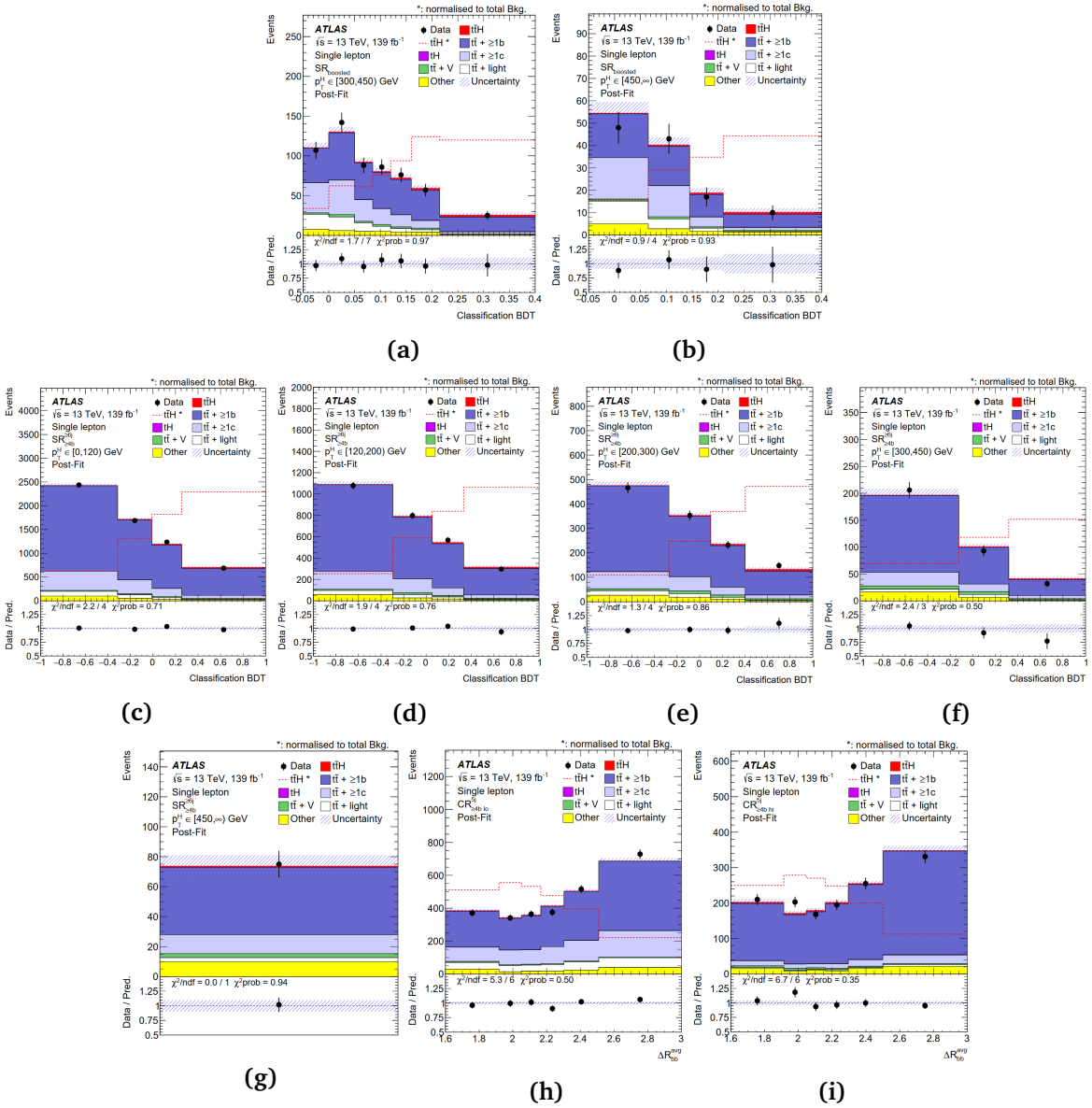


Fig. 8.25: Comparison between data and MC prediction for the BDT and ΔR_{bb}^{avg} discriminants in the SRs and CRs of the single-lepton channel, respectively, after performing the inclusive fit to data (post-fit) [100]. The boosted signal region, split in the (a) $300 \leq p_T^H < 450$ GeV (b) $p_T^H \geq 450$ GeV regions, is shown. The resolved signal region, $\text{SR}^{\geq 6j}_{\geq 4b}$, split into (c) $0 \leq p_T^H < 120$ GeV, (d) $120 \leq p_T^H < 200$ GeV, (e) $200 \leq p_T^H < 300$ GeV, (f) $300 \leq p_T^H < 450$ GeV, (g) $p_T^H \geq 450$ GeV (yield only) regions, as well as the control regions (h) $\text{CR}^{5j}_{\geq 4b\ lo}$ and (i) $\text{CR}^{5j}_{\geq 4b\ hi}$ are shown. In the latter, the first (last) bin includes the underflow (overflow). The $t\bar{t}H$ signal yield (solid red) is normalised to the fitted μ value from the inclusive fit. The dashed line shows the $t\bar{t}H$ signal distribution normalised to the total background prediction. The uncertainty band includes the statistical and systematic uncertainties as well as their correlations.

Furthermore, the background modelling is also reviewed in the various kinematic distributions. As discussed above, two of the largest NP pulls compensate for the mismodelling observed in the number of jets (Fig. 8.20) and reconstructed p_T^H distributions (Fig. 8.21), resulting in a very good data to MC agreement post-fit. Additionally, all the other input variables

to the classification BDTs develop an improved post-fit agreement between data and prediction compared to pre-fit. Indicatively, a few variables used in the single-lepton boosted channel, can be found in App. A.6 (Fig. A.13 and A.12).

To further validate the post-fit modelling, the goodness of fit is evaluated, accounting for all input variables to the classification BDTs and to a fit using the saturated model (outlined in Sec. 8.2.3). The obtained goodness of fit value is 92%, justifying the good post-fit modelling achieved. Another figure of merit is the probability of obtaining a level of agreement worse than observed between the fitted predictions and data, which is assessed by calculating the χ^2 for a given number of degrees of freedom and integrating the cumulative probability distribution to $+\infty$ (summarised in Sec. 8.1.1). Figure 8.26 depicts this data-to-MC-agreement test as a function of the p-value, retrieved from the χ^2 value and the number of degrees of freedom. In order to calculate these p-values, all correlations of the uncertainties are considered. Obviously, the illustrated p-values peak at one and only a few distributions end up in lower values. Typically, one would expect a flat distribution of the p-values, however, given that the analysis is dominated by the systematic uncertainties, the peak at one occurs. Ultimately, in all SRs for all channels combined the mean probability is 60% for the classification BDT training variables and 80% for the classification BDT outputs, denoting an overall good post-fit modelling.

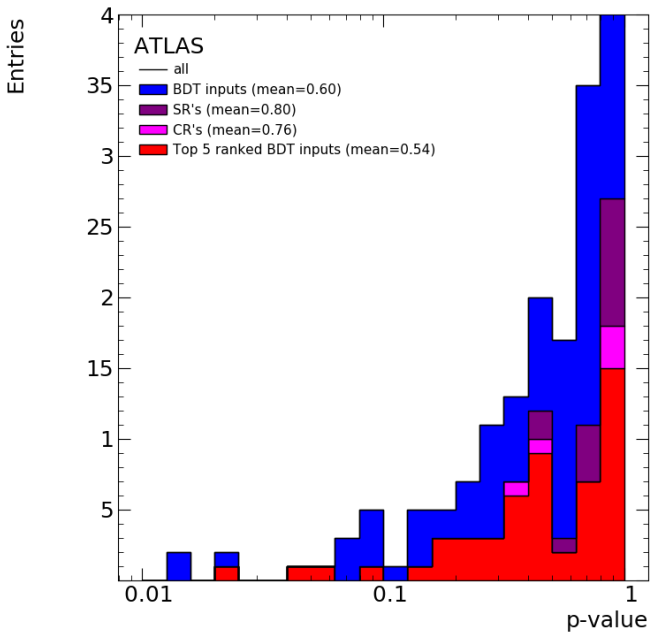


Fig. 8.26: Goodness of fit test as a function of the p-value retrieved from the χ^2 value and the number of degrees of freedom for the combined inclusive-μ fit to data. The p-value is calculated for the classification BDT input distributions in all channels (blue), for the top 5 ranked BDT inputs in all channels (red), as well as for the classification BDT outputs in the signal (purple) and control (pink) regions separately.

Finally, Fig. 8.27 shows the event yield in data compared with the post-fit prediction for all events entering the analysis selection, grouped and ordered by the signal-to-background ratio (S/B) of the corresponding bins entering the fit. The $t\bar{t}H$ signal normalised to the best-fit signal strength (red) and the SM prediction (orange) is depicted. Overall, the data is in good agreement with the nominal fit results. However, while in most bins the data also agrees well with the SM prediction, in the last three bins, which are most sensitive to the signal, the SM

scenario ($\mu_{t\bar{t}H} = 1$) overestimates the event yields.

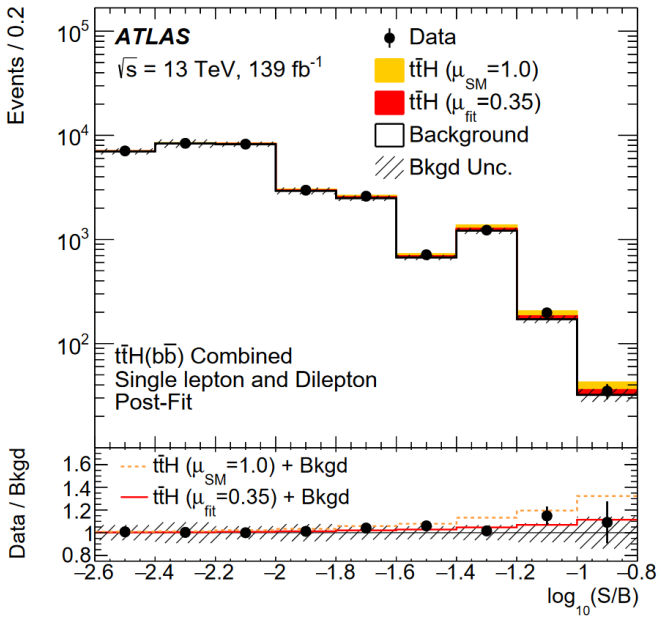


Fig. 8.27: Post-fit yields of signal (S) and total background (B) as a function of $\log(S/B)$ compared with data [100], from the combined inclusive-μ fit. Final-discriminant bins in all single-lepton and dilepton analysis regions are combined into bins of $\log(S/B)$, with the signal normalised to the SM prediction used for the computation of $\log(S/B)$. The signal is then shown normalised to the best-fit value and the SM prediction. The lower frame reports the ratio of data to background, and this is compared with the expected $t\bar{t}H$-signal-plus-background yield divided by the background-only yield for the best-fit signal strength (solid red line) and the SM prediction (dashed orange line).

8.4.2 STXS measurement

As already remarked, no big changes with respect to the inclusive cross-section measurement are required in order to perform the STXS measurement. The signal template is split into five truth $\hat{p}_T^H$ bins, corresponding to the reconstructed p_T^H bins of the SRs, and each template has a dedicated signal strength parameter, thus the STXS bin migration uncertainties are removed. The resulting best-fit values of the signal strength from the STXS measurement and a Higgs boson with $m_{Higgs} = 125$ GeV are shown in Fig. 8.28, together with the one from the inclusive-μ fit discussed above. The uncertainties associated to the signal strength parameter are overall fairly large. All $\mu_{t\bar{t}H}$ values are consistent with each other within their uncertainties as well as with the inclusive measurement. The measurement is dominated by the statistical uncertainty in the $\hat{p}_T^H \in [200, 300)$ GeV and $\hat{p}_T^H \in [300, 450)$ GeV bins, whereas the other bins are dominated by the systematic uncertainties. Apparently, some signal strength parameters are negative, indicating that the MC overestimates the background in the regions where the signal is expected. However, this can be caused just by fluctuations in the MC prediction. Nonetheless, the total signal yield in each bin entering the fit is never zero or negative after the fit, since this effect is compensated by the other signal strength parameters. Moreover, the normalisation factor of the $t\bar{t}+\geq 1b$ background is found to be $k(t\bar{t}+\geq 1b) = 1.28 \pm 0.08$, in perfect agreement with the inclusive-μ fit value. The statistical uncertainty on these parameters is evaluated as described

for the inclusive fit. The probability that the obtained signal strengths are compatible with the SM predictions is 45%, estimated by repeating the fit while fixing to $\mu = 1$ in the five bins.

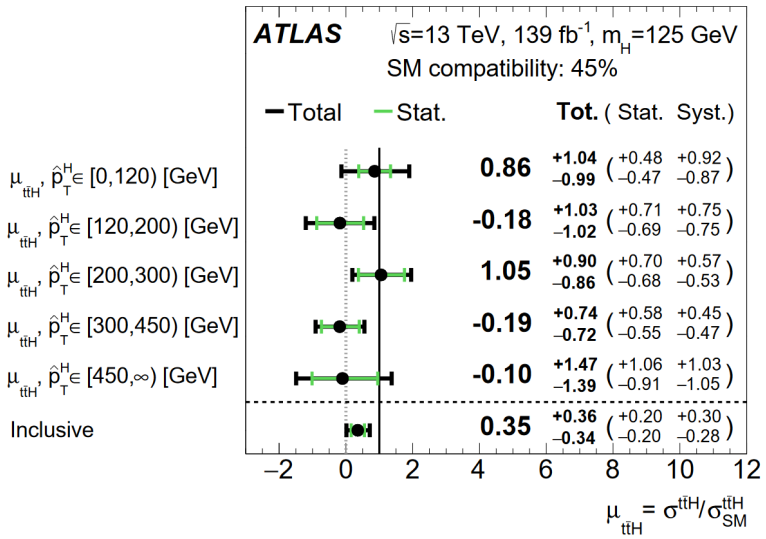


Fig. 8.28: Fitted values of the $t\bar{t}H$ signal-strength parameter from the S+B fit to data in the individual STXS $\hat{p}_T^H$ bins and in the inclusive signal-strength measurement, combining all the analysis channels [100].

The correlation matrix of the nuisance parameters, the $k(t\bar{t} + \geq 1b)$ normalisation factor, and the signal strength parameters, containing the correlation coefficients of a selection of the aforementioned parameters obtained by the combined fit to data, is shown in Fig. 8.29. The correlations are pretty much similar to those from the inclusive measurement in a comparable amount. The only difference emerges from the fact that now there are multiple signal strength parameters, so the various NPs are potentially correlated to each one of them separately. The most noticeable are the anti-correlation between the signal strength $\mu_{t\bar{t}H}, \hat{p}_T^H \in [0, 120)$ GeV with the $t\bar{t}+\geq 1b$ NLO Gen. single-lepton $p_T^H \in [0, 120)$ GeV parameter (-68.7%) as well as with the $t\bar{t} + \geq 1b$ NLO Gen. dilepton $p_T^H \in [0, 120)$ GeV (-43.7). Then, the $\mu_{t\bar{t}H}, \hat{p}_T^H \in [120, 200)$ GeV is highly anti-correlated with the $t\bar{t} + \geq 1b$ NLO Gen. single-lepton $p_T^H \in [120, 200)$ GeV parameter (-49.0%), while the $\mu_{t\bar{t}H}, \hat{p}_T^H \in [450, \infty)$ GeV is strongly correlated with the $t\bar{t} + \geq 1b$ NLO Gen. single-lepton $p_T^H \in [450, \infty)$ GeV parameter (52.5%). These correlations imply that these systematics have similar features to the signal, and it is thus difficult to separate the signal from the $t\bar{t} + \geq 1b$ background in these regions. Given that the most sensitive analysis regions are dominated by the $t\bar{t} + \geq 1b$ background and considering its observed MC modelling, these strong correlations have dominant impact on the signal strength, leading to lower sensitivity mainly in the respective bins.

Overall, the observed pulls and constraints of the fitted NPs, after the $S + B$ combined STXS fit on data, are similar to those from the inclusive measurement, though some of the pulls slightly differ in size. Again, the $t\bar{t} + \geq 1b$ ISR uncertainty has the largest pull followed by the p_T^{bb} shape uncertainty, the $t\bar{t} + \geq 1c$ normalisation uncertainty, and the the $t\bar{t} + \geq 1b$ NLO matching uncertainty in the dilepton SR in the $0 \leq p_T^H < 120$ GeV bin. Only the latter is about 0.2σ less pulled than before, while the others are pulled in relatively the same amount. The figure can be found in Appendix for reference (Fig. A.11).

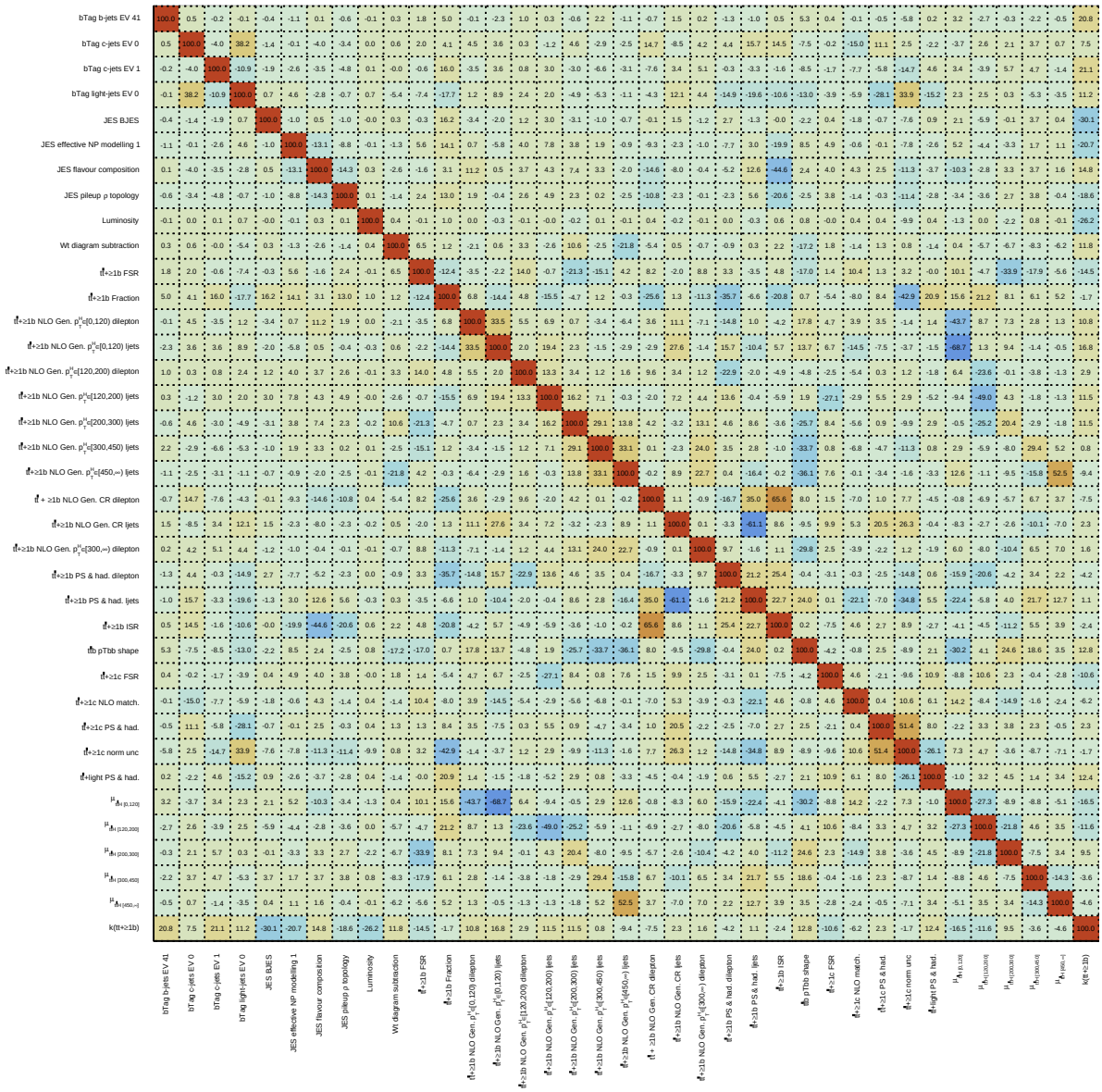


Fig. 8.29: Linear correlation coefficients of the nuisance parameters, normalisation factor, and signal strength, after the combined STXS fit to data. All values are given in percent. Only nuisance parameters with at least one absolute correlation coefficient above 20% are shown.

As already pointed out, also the STXS measurement is mostly dominated by the systematic uncertainties. The actual effect of the various systematic uncertainties is examined separately on each signal strength parameter, as illustrated in Fig. 8.30. Although some instrumental NPs show up in the ranking, related to the jet energy resolution and to the b-taggig efficiency or mistag rates, the dominant contributions still originate from the $t\bar{t}+b\bar{b}$ background modelling. It is also remarkable that the p_T^{bb} shape uncertainty becomes more dominant in the higher STXS bins. This is possibly because the shape effect gets more prominent in the larger p_T^H bins by construction.

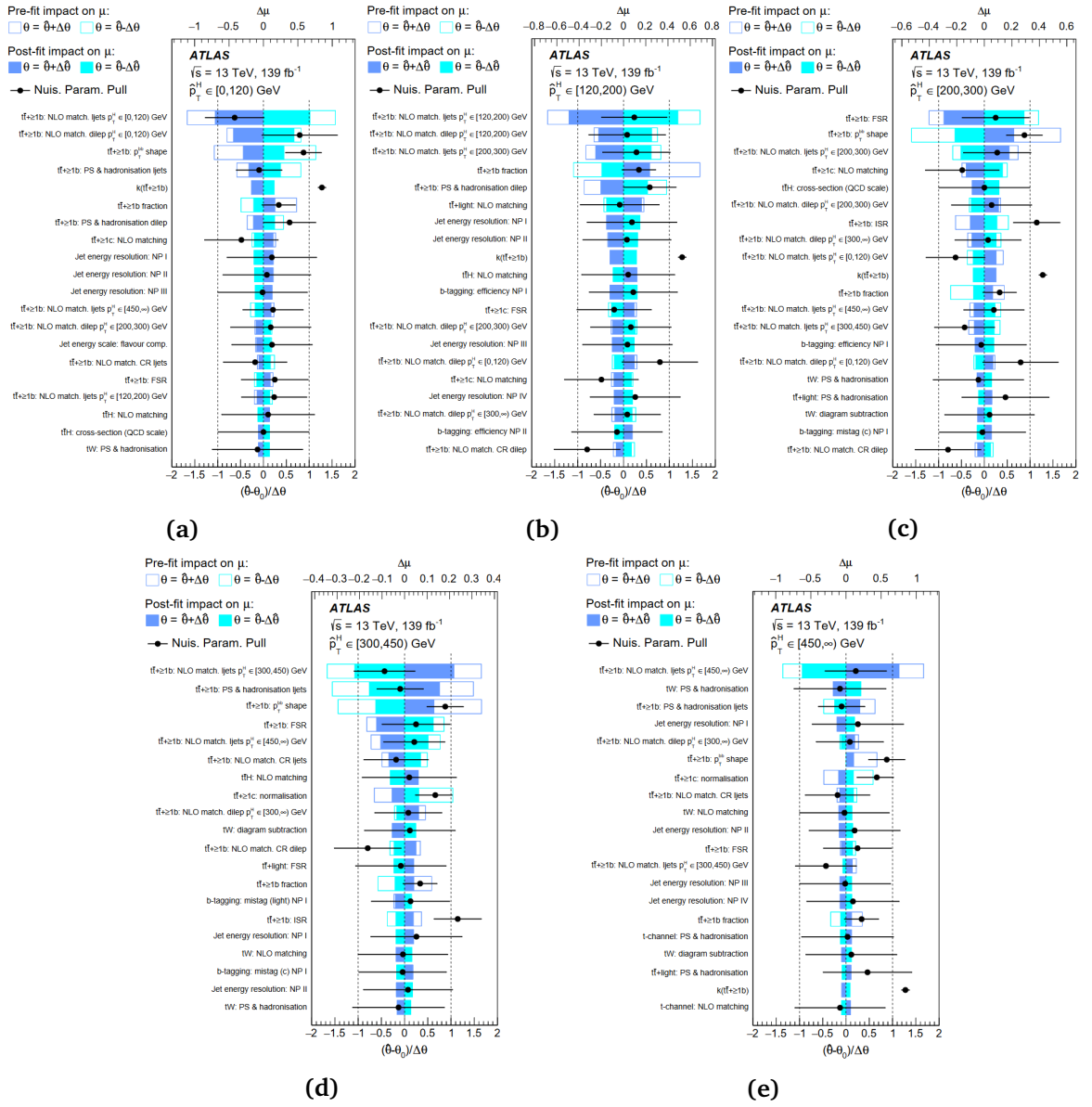


Fig. 8.30: Ranking of the 20 nuisance parameters with the largest post-fit impact on μ in the combined STXS fit to data, for (a) $0 \leq \hat{p}_T^H < 120$ GeV, (b) $120 \leq \hat{p}_T^H < 200$ GeV, (c) $200 \leq \hat{p}_T^H < 300$ GeV, (d) $300 \leq \hat{p}_T^H < 450$ GeV, and (e) $\hat{p}_T^H \geq 450$ GeV [100]. Nuisance parameters corresponding to statistical uncertainties in the MC simulated event samples are not included. The empty blue rectangles correspond to the pre-fit impact on μ, while the filled blue ones to the post-fit impact on μ, both referring to the upper scale. The impact of each nuisance parameter, $\Delta\mu$, is computed by comparing the nominal best-fit value of μ with the result of the fit when fixing the considered nuisance parameter to its best-fit value, $\hat{\theta}$, shifted by its pre-fit (post-fit) uncertainties $\pm\Delta\hat{\theta}$ ($\pm\Delta\theta$). The black points show the pulls of the nuisance parameters relative to their nominal values, $\theta_0 = 0$. These pulls and their relative post-fit errors, $\Delta\hat{\theta}/\Delta\theta$, refer to the lower scale. The parameter $k(t\bar{t} + \geq 1b)$ refers to the free-floating normalisation of the $t\bar{t} + \geq 1b$ background, for which the pre-fit impact on μ is not defined, while its nominal value is $\theta_0 = 1$. For the experimental uncertainties, that are decomposed into several independent sources, NP X corresponds to the Xth nuisance parameter, ordered by their impact on μ. The "ljets" ("dilep") label refers to the single-lepton (dilepton) channel.

After the combined STXS fit to data is performed, the prediction is adjusted according to these fit results. Therefore, the post-fit agreement between the data and the MC simulated events in the various distributions should be revisited. In this case, the fitted values of each signal strength parameter, together with the $k(t\bar{t}+\geq 1b)$ normalisation, are taken into account. Their uncertainties are included in the overall uncertainty, as well. Again, the post-fit uncertainties account for the correlations of all NPs and their constraints. Likewise, the post-fit distributions demonstrate an improved agreement between data and prediction, quite similar as before. The global goodness of fit is 88%, denoting the good post-fit modelling obtained. This is also verified from the χ^2 test as a function of the p-value, depicted in Fig. 8.31, where the mean probability of the classification BDT input variables is 61% and for the classification BDT outputs is 83%, in the SRs of all channels combined. The values of both tests are comparable to those obtained from the inclusive measurement, validating that the post-fit modelling is almost similar between the two measurements.

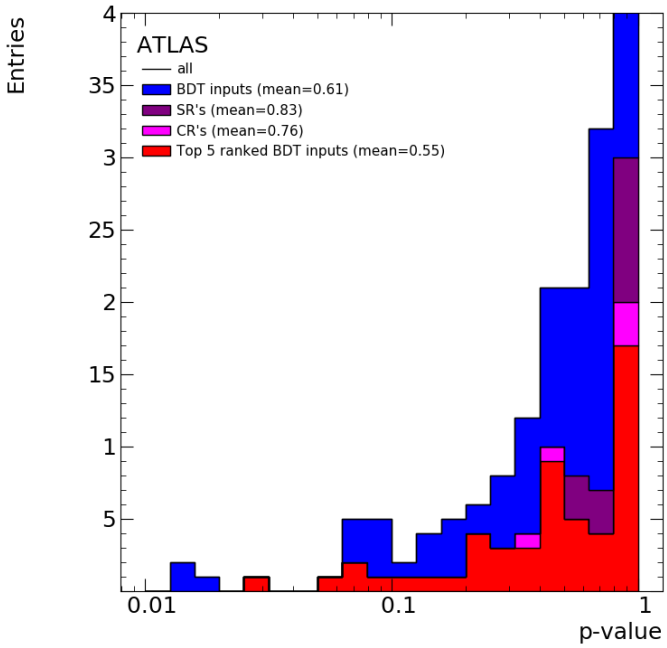


Fig. 8.31: Goodness of fit test as a function of the p-value retrieved from the χ^2 value and the number of degrees of freedom for the combined STXS fit. The p-value is calculated for the BDT input distributions in all channels (blue), in the signal (purple) and control (pink) regions separately, as well as separately for the top 5 ranked BDT inputs in all channels (red).

8.4.3 Setting limits

No significant excess is observed in data compared to the background-only hypothesis. Specifically, an excess of $t\bar{t}H(b\bar{b})$ events over the expected SM background is found with an observed (expected) significance of 1.0 (2.7) standard deviations. Therefore, an upper limit on μ under the background-only hypothesis is determined, using a modified frequentist CLs procedure with the asymptotic method, discussed in Sec. 8.1.2. The expected significance and exclusion limits are calculated using the background estimate after the fit to the data. The observed and expected upper limits in the inclusive measurement as well as in the individual STXS measurement $\hat{p}_T^H$ bins are illustrated in Fig. 8.32 and listed in Table 8.2. Indicatively for the combined

inclusive measurement, an observed upper limit at 95% confidence level (CL) on μ of 1.0 is obtained. This means that a signal strength larger than 1.0 ($\mu_{t\bar{t}H} > 1.0$) is excluded at the 95% confidence level. Simultaneously, in the absence of signal, the expected exclusion would be $\mu_{t\bar{t}H} > 0.68$ at the 95% confidence level.

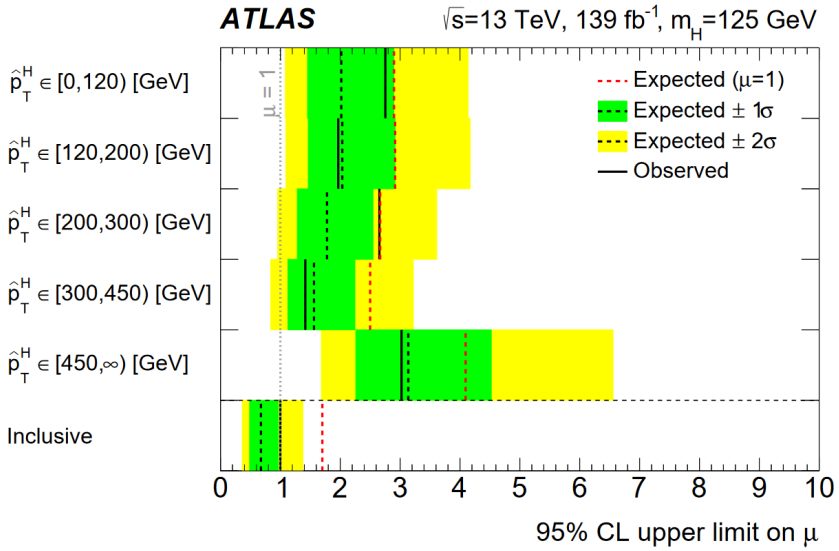


Fig. 8.32: Summary of the 95% confidence level (CL) upper limits on signal-strength in the individual STXS $\hat{p}_T^H$ bins as well as in the inclusive measurement, after the combined fit to data [100]. The observed limits are shown (solid black lines), together with the expected limits both in the background-only hypothesis (dotted black lines) and in the SM hypothesis (dotted red lines). In the case of the expected limits under the background-only hypothesis, 1σ (green) and 2σ (yellow) uncertainty bands are also shown.

$\hat{p}_T^H$ bin [GeV]	95% upper limit		Best-fit μ
	obsreved	expected $\pm 1\sigma$	
$[0, 120)$	2.75	$2.02^{+0.88}_{-0.57}$	$0.86^{+1.04}_{-0.99}$
$[120, 200)$	1.97	$2.03^{+0.88}_{-0.57}$	$-0.18^{+1.03}_{-1.02}$
$[200, 300)$	2.65	$1.78^{+0.77}_{-0.50}$	$1.05^{+0.90}_{-0.86}$
$[300, 450)$	1.42	$1.56^{+0.68}_{-0.44}$	$-0.19^{+0.74}_{-0.72}$
$[450, \infty)$	3.02	$3.14^{+1.39}_{-0.88}$	$-0.10^{+1.47}_{-1.39}$
Inclusive	1.00	$0.68^{+0.29}_{-0.19}$	$0.35^{+0.36}_{-0.34}$

Table 8.2: Best fit value of the signal strength μ and the observed and median expected 95% CL upper limits in the individual STXS $\hat{p}_T^H$ bins as well as in the inclusive measurement for the combined result. Also, the expected limits under the background-only hypothesis and the 1σ uncertainties are quoted.

Chapter 9

Conclusion and Outlook

9.1 Conclusion

The discovery of the SM Higgs boson [10,11] has been a milestone for the LHC and for the particle physics field in general. This discovery confirms the success of the proposed theory about the existence of an associated Higgs field that describes the electroweak symmetry breaking as a mechanism to generate massive vector bosons, in addition to fermion masses through Yukawa couplings. Afterwards, a new era of studies and measurements started in order to understand the properties of this newly discovered particle. In particular, the top-quark Yukawa coupling, which is the strongest Yukawa coupling in the SM, was until recently observed only through indirect measurements [13], since they provide clearer signatures. The recent analysis of the Higgs-boson production associated with a pair of top quarks ($t\bar{t}H$) [102, 103] provided the first insight of this coupling in a combined measurement of several channels, defined by the decay products of the Higgs boson. The result showed an agreement with the SM prediction, although the uncertainty is still quite large, hence a more precise measurement could result in a discrepancy. Moreover, a differential cross-section measurement would probe the CP properties of the coupling, which are not yet established. Both aspects constitute a strong motivation to ameliorate the $t\bar{t}H$ measurement.

In this thesis, a measurement of the $t\bar{t}H$ production cross-section with a subsequent Higgs-boson decay into a pair of b-quarks ($H \to b\bar{b}$) was presented. The results are based on the full Run 2 dataset of pp collision data collected at $\sqrt{s} = 13$ TeV by the ATLAS experiment at the LHC, corresponding to an integrated luminosity of 139 fb^{-1}, and have already been published [100]. Although the $H \to b\bar{b}$ channel has the highest branching ratio, the precision of the measurement in this channel is limited by the low precision of the background modelling. Events with one or two charged leptons in the final state, emerging from the decay of the top-quark pair, are considered. As a result, the final state of the $t\bar{t}H(H \to b\bar{b})$ process contains at least four b-jets, hence the analysis strongly depends on b-tagging. Also, the $t\bar{t}$+jets process, having a cross section a few orders of magnitude larger than the $t\bar{t}H$ production, constitutes the overwhelming background of this analysis, which becomes particularly challenging when the associated jets stem from heavy-flavour quarks. Thus, combinatorial ambiguities arise when trying to assign the various b-jets in the final state to the decay products of the Higgs boson and top quarks, resulting in low reconstruction efficiency of the latter.

Although the analysis targets the $H \to b\bar{b}$ decay, all the decay modes may contribute to the signal. The $t\bar{t}H$ events are split into non-overlapping regions, based on the number of

leptons, of jets, and of b-tagged jets, in order to provide regions enhanced in signal, or in the dominant background components. Especially in the single-lepton channel, a boosted category is designed to select events in which the Higgs boson and possibly also the hadronically decaying top quark are produced with high p_T, so that their decay products are collimated in large-R jets. In fact, the single-lepton boosted channel, which targets events with Higgs-boson candidate p_T greater than 300 GeV, is the main scope of this thesis. By contrast, no boosted category is designed in the dilepton channel, due to the small number of expected events.

Multivariate analysis techniques are employed to identify the reconstructed objects with the underlying particles and to maximise the separation of the signal from the background in the signal-enriched regions. The background is dominated by the $t\bar{t}+b\bar{b}$ component, a $t\bar{t}$ process with additional b-quarks originating from the splitting of a radiated gluon. A large number of heavy-flavour jets in the final state is not well modelled though, and a huge effort has been made to constrain this large background with the latest theoretical predictions. Many systematic uncertainties are accounted for, decreasing the sensitivity of the measurement. Then, the signal-enriched analysis regions are combined with the signal-depleted ones into a profile likelihood fit. The output distributions of these multivariate algorithms are used as the main discriminant to extract the signal strength. The fit simultaneously determines the event yields for the signal process and the most important background component, while constraining the overall background model within the assumed systematic uncertainties.

Eventually, the measured $t\bar{t}H$ signal strength of the inclusive cross-section measurement is found to be $0.35^{+0.36}_{-0.34}$, corresponding to an observed (expected) significance of 1.0 (2.7) standard deviations. This result excludes $t\bar{t}H$ signal strengths larger than the SM prediction at 95% confidence level. The value of the signal strength is lower than expected, though still in agreement with the SM expectation within 1.8σ. The probability of the obtained signal strength being compatible with the SM prediction is 8.5%. The measurement uncertainty is dominated by systematic uncertainties, notably regarding the theoretical knowledge of the $t\bar{t}+\geq 1b$ background process, which still drives the sensitivity, despite the significant improvement relative to the previous measurement [97]. Also, the current analysis reports approximately 1.3 and 1.6 times higher contributions of $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ background components compared to the MC expectation, respectively. These effects are compatible with the previous measurement and with a dedicated analysis of the $t\bar{t}b\bar{b}$ process at $\sqrt{s}=13$ TeV [290].

Although the $t\bar{t}H(H\rightarrow b\bar{b})$ analysis is still limited by systematic uncertainties, their impact has been significantly reduced, by almost a factor of two, compared to the previous analysis [97]. This improvement was mainly achieved by adopting an enhanced model for the $t\bar{t}$+jets background with updated event generator versions, and a higher-order precision prediction for the $t\bar{t}+b\bar{b}$ process, which is now described by the four-flavour scheme. Also, the increased number of simulated events and the improved assessment of uncertainties, while reducing double-counting of certain uncertainty sources, contributed to the optimisation of the systematics model. Moreover, the impact of the experimental uncertainties on the measurement has been appreciably mitigated with respect to the previous analysis, given the improved reconstruction algorithms and detector calibrations, which in turn ameliorate the performance of the b-tagging algorithm. This allows for the use of tighter selection criteria to reject events in poorly modelled regions of phase space, and the definition of analysis regions differential in the Higgs-boson transverse momentum.

Furthermore, the overall statistical uncertainty has been largely confined compared to the previous analysis, considering that now the full Run 2 dataset is analysed. Also, in the previous

measurement, the low statistics of the MC samples had a remarkable impact on the result, whereas now it is found to be almost negligible due to the increased size of the simulated samples.

Also, exploiting the possibility to reconstruct the Higgs boson kinematics in the $H \to b\bar{b}$ mode, the first differential cross-section measurement of the $t\bar{t}H$ signal strength is performed in five bins of the true $\hat{p}_T^H$ in the STXS framework. The signal strength parameters associated to the first two bins are limited by their systematic uncertainties, while the remaining ones are mostly dominated by the statistical uncertainty. In general, the uncertainties are considerably large and the different signal strength parameters from the STXS measurement are in agreement with the measured signal strength in the inclusive cross-section measurement. The probability that the obtained signal strengths are compatible with the SM predictions amounts to 45%.

The contribution of the single-lepton boosted region to the measurement is of substantial importance, especially for the STXS measurement. The improved region definition as well as the employment of a new deep neural network enhanced the performance of the single-lepton boosted channel. In contrast to the previous analysis, there is an evident gain on the sensitivity of the current measurement from the inclusion of the optimised boosted region, whilst small. A significant improvement in the sensitivity of the high $\hat{p}_T^H$ bins is noticed though, since the boosted channel is sensitive in this phase space. In fact, its inclusion is determinant for the differential measurement, otherwise the cross section in the highest $\hat{p}_T^H$ bin would not even be measurable.

9.2 Outlook

Despite the noteworthy advancement in the $t\bar{t}H(H \to b\bar{b})$ measurement presented in this thesis, there are still some caveats that affect the measurement and room for further improvements in the future. At first, the LHC Run 3 data taking period has recently started at an increased centre-of-mass energy (13.6 TeV) and is expected to achieve almost double the luminosity with respect to Run 2 [299]. Thus, much more data are expected to be collected, increasing significantly the statistics for the future analyses.

One of the difficulties in this measurement is the inaccurate modelling of the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ fractions in the $t\bar{t}$+jets process, since their contribution is found to be higher compared to the MC expectation. Another limitation could be mitigated by lowering the renormalisation scale of the nominal $t\bar{t}+b\bar{b}$ sample, as explained in Sec. 8.4.1. However, there is a remaining mismodelling covered by the two point systematic uncertainties, which have a dominant impact on the measurement. Ultimately, the analysis could further benefit from an improved $t\bar{t}+\geq 1b$ background modelling. The dedicated measurements of the $t\bar{t}+b\bar{b}$ process can be an important input for the optimisation the of MC samples. Then, MC generators that properly define alternatives to the NLO matching and to the PS and hadronisation models of the $t\bar{t}+b\bar{b}$ process are required, instead of using extrapolation from the inclusive $t\bar{t}$ sample, as used in this analysis. Also, merging the NLO $t\bar{t}+b\bar{b}$ calculation with the inclusive $t\bar{t}$ production would give the most precise calculation. All in all, better understanding of the background would significantly improve the sensitivity of the $t\bar{t}H(H \to b\bar{b})$ measurement, by further reducing the dominant systematic uncertainties. Apart from this, increasing the amount of generated MC events would further reduce the impact of the statistical uncertainty of the MC predictions on the measurement.

Moreover, a future analysis can take advantage of the most up-to-date reconstruction techniques that now apply in the ATLAS community, in order to improve the background rejection. At first, this means using the particle flow (PFlow) jets [292] instead of the EMTopo jets that were used until now. They combine tracks with clusters in the calorimeter for the jet reconstruction to better distinguish between the charged and neutral particles. The advantage of PFlow jets is their improved energy and angular resolution compared to EMTopo jets, as well as their enhanced reconstruction efficiency and pile-up stability. Then, alongside the PFlow jets, the use of DNN-based DL1r b-tagging algorithm [293] is recommended, instead of the BDT-based MV2c10 tagger [220] used so far. The main advantage of DL1r is its multi-class output, meaning that the network predicts the probabilities for being compatible with the three main flavour classes, i.e. b-, c-, and $light$-flavour jets, for every jet. In principle, this can be also realised with a BDT however, DNNs are more flexible and have more possibilities to customise their structure.

From the increased data and MC statistics, especially in the the high-p_T region, the boosted phase-space would explicitly profit. This would help to improve the performance of the single-lepton boosted channel and increase its impact on the measurement. In addition, with the higher energy available, a higher fraction of events will be more highly energetic, allowing to target even higher-p_T boosted regimes. As already observed (see Fig. 6.2) at very high p_T the two b-quarks from the Higgs boson decay become so collimated that they start to overlap, reaching an angular separation $\Delta R \leq 0.4$. This motivates the construction of a single-lepton ultra-boosted region, targeting at $t\bar{t}H$ events with Higgs-boson $p_T \geq 450$ GeV and with the two b-quarks being reconstructed within a small-R jet [298]. However, the common b-tagging algorithms are not optimised for such a topology. Therefore, in order to avoid these overlaps and to improve b-tagging performance, the $X \rightarrow b\bar{b}$ tagger [296] can be employed. It identifies massive particles decaying to a $b\bar{b}$ pair at high p_T. It is a double b-tagging algorithm based on a neural network which associates variable-radius track (VR-track) subjets [297] to the large-R jet, and then assigns a flavour discriminant to each of the subjets using a flavour-tagging algorithm. Finally, to further enhance the separation between the $t\bar{t}H(H \rightarrow b\bar{b})$ signal and the $t\bar{t} + \geq 1b$ background, other multivariate analysis techniques can be revised.

Chapter A

Appendix

A.1 Particle interactions calculations

A.1.1 Decay and scattering processes

A particle is characterised as "unstable" if there is at least one allowed final state, with multiple other particles, that it can decay into. Particle decays are mediated by one or several fundamental forces. One of the most important characteristics of a particle is its *lifetime*, τ. It depends, on the available decay modes or channels, which are subject to conservation laws for appropriate quantum numbers, coupling strength of the decay process, and kinematic constraints. A decay process is expressed in terms of lifetime, or, equivalently, *decay rate*

$$\Gamma = \frac{1}{\tau}, \tag{A.1}$$

which is a measure of the probability of a specific decay process occurring within a given amount of time in the rest frame of the parent particle.

Hadrons, with a lifetime on the order of 10^{-23} s, are called resonances. They are far too short-lived to be directly observed, so their existence must be inferred from observations on the more stable hadrons to which they decay, via strong interaction. A resonance occurs when the energy of the colliding (parent) particles is sufficient to produce its rest mass. The presence of the resonance is indicated by the peak on the energy distribution of the colliding (or product) particles, which is approximated by the Breit–Wigner formula. The value of the energy range corresponding to the half of the resonance peak is the width $\hbar\Gamma$ of the resonance. Hence, in the system of natural units ($\hbar = c = 1$), Γ is also called *decay width*.

Unstable particles often have multiple decay modes, each with its own associated decay rate, and the *total decay rate/width* is the sum of the rates/widths of the individual modes *i*

$$\Gamma_{total} = \sum_i \Gamma_i. \tag{A.2}$$

In such cases, one is often interested in the probability of a particle to decay by an individual mode. This is called *branching ratio* and it is the fraction of the decay rate of mode *i* to the total decay rate

$$B_i = \frac{\Gamma_i}{\Gamma_{total}}. \tag{A.3}$$

Moreover, a scattering process is characterised by the *cross section*, σ, which is a measure of the probability that two particles interact with each other. The cross section of a process can

be thought of as the effective area within which a specific scattering process occurs. Thus the units of a cross section are the units of an area (cm^2). In contemporary high energy physics experiments, cross sections are typically measured in units of nanobarn (nb) to femtobarn (fb), where a *barn* is defined as 1b = 10^{-24} cm^{-2}. Analogously, in cases that there are multiple decay modes, each one is characterised by an *exclusive cross section*, σ_i. Summing them up one get the *total* (or *inclusive*) *cross section* of the process

$$\sigma_{total} = \sum_i \sigma_i. \tag{A.4}$$

A.1.2 Calculation of widths and cross sections

In principle, for a field theory formulated in terms of a Lagrangian density of quantum fields, the time evolution of arbitrary initial states can be calculated. The interaction between particles is described by the interaction terms in the Lagrangian. However, while the time evolution can be solved exactly in a free theory, a theory containing interactions requires approximate calculations. For scattering reactions these calculations are usually performed in perturbation theory. The basic idea is to start with time-independent states of the free theory, that describe the incoming and outgoing particles and include the interaction as a small perturbation. The time evolution of the free-theory states due to the perturbation can be expressed in terms of the interaction terms in the Lagrangian. The quantitative formulation of elementary particle dynamics amounts to the calculation of decay rates and scattering cross sections.

In scattering experiments, the states before and after a scattering (or decay) process, are characterised by the four-momenta of the colliding and scattered (or decaying and produced) particles. The transition from the initial state $|i\rangle$ to the final state $|f\rangle$ is described by a unitary operator, the so-called *S-matrix*

$$S_{fi} = \delta_{fi} - i(2\pi)^4 \delta^4(p_f - p_i)\mathcal{M}(p_i \to p_f), \tag{A.5}$$

where p_i and p_f are the total momenta of the initial and final state, respectively. However, even if the theory contains interactions, the initial particles have some probability of simply not interacting with other, whereupon the initial and final states are the same. These cases are described by the δ_{fi} term in eq. A.5, as the *S*-matrix becomes the identity operator. Hence, the second term contains the information due to interactions, where the four-dimensional δ-function reflects the energy-momentum conservation between the initial and final state. In particular, all the physics that depends on the dynamics of the process is contained in the Lorentz invariant quantum mechanical *transition amplitude* (or *matrix element*) $\mathcal{M}$. It can be perturbatively calculated from the interaction part of Lagrangian with the help of the Dyson series.

The *transition rate per unit space-time volume*, for a scattering (or decay) process with non-identical initial and final state to occur, is proportional to the square of the transition amplitude. It is determined by Fermi's Golden Rule as

$$\frac{\text{transition rate}}{\text{unit space-time volume}} = (2\pi)^4 \delta^4(p_f - p_i)|\mathcal{M}(p_i \to p_f)|^2 \times \begin{pmatrix} \text{density of} \\ \text{final states} \end{pmatrix}, \tag{A.6}$$

with the Lorentz invariant phase-space factor defined by the number of final states per unit volume, with momenta in element d^3p and normalised to $2E$ particles, as $\frac{d^3\vec{p}}{(2\pi)^3}\frac{1}{2E}$. The quantitative formulation of elementary particle dynamics amounts, in practice, to the calculation

of decay rates and scattering cross sections. Both the decay rate and the cross section express the probability of the the process they describe to take place. They are related to the transition rate by

$$\begin{pmatrix} \text{decay rate} \\ or \\ \text{cross section} \end{pmatrix} = \frac{\text{transition rate/unit space-time volume}}{\text{incident flux}} \tag{A.7}$$

More precisely, for a particle decay with *n* particles in the final state $\alpha \to 1 + 2 + ... + n$, the *incident flux* is defined as $2E_a$ Then, the partial decay rate is given by

$$d\Gamma = \frac{S}{2E_\alpha}(2\pi)^4\delta^4(p_\alpha - \sum_{f=0}^{n} p_f)|\mathcal{M}(p_\alpha \to \{p_f\})|^2 \prod_{f=0}^{n} \frac{d^3\vec{p}_f}{(2\pi)^3}\frac{1}{2E_f}, \tag{A.8}$$

where p_α and E_α are the four-momentum and energy of the initial particle, while p_f (or $\vec{p}_f$) is the four- (or three-) momentum of the *f*th particle. Also, *S* is the product of the statistical factors $\frac{1}{m!}$, for each group of *m* identical particles in the final state. Furthermore, for a general collinear collision between two particles with $\alpha + \beta \to 1 + 2 + ... + n$, the relevant flux is equal to $|\vec{v}_\alpha - \vec{v}_\beta|2E_\alpha 2E_\beta$. In both cases the flux is a Lorentz invariant quantity. After all, the differential cross section is formulated as

$$d\sigma = \frac{S}{4E_\alpha E_\beta|\vec{v}_\alpha - \vec{v}_\beta|}(2\pi)^4\delta^4(p_\alpha + p_\beta - \sum_{f=0}^{n} p_f)|\mathcal{M}(p_\alpha, p_\beta \to \{p_f\})|^2 \prod_{f=0}^{n} \frac{d^3\vec{p}_f}{(2\pi)^3}\frac{1}{2E_f}, \tag{A.9}$$

where E_α, E_β and $|\vec{v}_\alpha - \vec{v}_\beta|$ denote the energy and relative velocity of the colliding particles, respectively.

A.2 Monte Carlo simulated samples

Table A.1 summarises the generator settings for the MC samples used in this analysis. The references of the various generators ad settings can be found in Sec. 4.5.

Process	ME generator	ME PDF	PS	Normalisation
Higgs boson				
$t\bar{t}H$	POWHEG BOX v2	NNPDF3.0NLO	PYTHIA 8.230	NLO+NLO (EW)
	POWHEG BOX v2	NNPDF3.0NLO	HERWIG 7.04	NLO+NLO (EW)
	MADGRAPH5_AMC@NLO 2.6.0	NNPDF3.0NLO	PYTHIA 8.230	NLO+NLO (EW)
$tHjb$	MADGRAPH5_AMC@NLO 2.6.2	NNPDF3.0NLO nf4	PYTHIA 8.230	–
tWH	MADGRAPH5_AMC@NLO 2.6.2 [DR]	NNPDF3.0NLO	PYTHIA 8.235	–
$t\bar{t}+jets$ and single-top				
$t\bar{t}$	POWHEG BOX v2	NNPDF3.0NLO	PYTHIA 8.230	NNLO+NNLL
	POWHEG BOX v2	NNPDF3.0NLO	HERWIG 7.04	NNLO+NNLL
	MADGRAPH5_AMC@NLO 2.6.0	NNPDF3.0NLO	PYTHIA 8.230	NNLO+NNLL
$t\bar{t}+b\bar{b}$	POWHEG BOX RES	NNPDF3.0NLO nf4	PYTHIA 8.230	–
tW	POWHEG BOX v2 [DR]	NNPDF3.0NLO	PYTHIA 8.230	NLO+NNLL
	POWHEG BOX v2 [DS]	NNPDF3.0NLO	PYTHIA 8.230	NLO+NNLL
	POWHEG BOX v2 [DR]	NNPDF3.0NLO	HERWIG 7.04	NLO+NNLL
	MADGRAPH5_AMC@NLO 2.6.2 [DR]	CT10NLO	PYTHIA 8.230	NLO+NNLL
t-channel	POWHEG BOX v2	NNPDF3.0NLO nf4	PYTHIA 8.230	NLO
	POWHEG BOX v2	NNPDF3.0NLO nf4	HERWIG 7.04	NLO
	MADGRAPH5_AMC@NLO 2.6.2	NNPDF3.0NLO nf4	PYTHIA 8.230	NLO
s-channel	POWHEG BOX v2	NNPDF3.0NLO	PYTHIA 8.230	NLO
	POWHEG BOX v2	NNPDF3.0NLO	HERWIG 7.04	NLO
	MADGRAPH5_AMC@NLO 2.6.2	NNPDF3.0NLO	PYTHIA 8.230	NLO
Other				
$W+jets$	SHERPA 2.2.1 (NLO [2j], LO [4j])	NNPDF3.0NNLO	SHERPA	NNLO
$Z+jets$	SHERPA 2.2.1 (NLO [2j], LO [4j])	NNPDF3.0NNLO	SHERPA	NNLO
VV (had.)	SHERPA 2.2.1	NNPDF3.0NNLO	SHERPA	–
VV (lep.)	SHERPA 2.2.2	NNPDF3.0NNLO	SHERPA	–
VV (lep.) + jj	SHERPA 2.2.2 (LO [EW])	NNPDF3.0NNLO	SHERPA	–
$t\bar{t}W$	MADGRAPH5_AMC@NLO 2.3.3	NNPDF3.0NLO	PYTHIA 8.210	NLO+NLO (EW)
	SHERPA 2.0.0 (LO [2j])	NNPDF3.0NNLO	SHERPA	NLO+NLO (EW)
$t\bar{t}\ell\ell$	MADGRAPH5_AMC@NLO 2.3.3	NNPDF3.0NLO	PYTHIA 8.210	NLO+NLO (EW)
	SHERPA 2.0.0 (LO [1j])	NNPDF3.0NNLO	SHERPA	NLO+NLO (EW)
$t\bar{t}Z(qq,\nu\nu)$	MADGRAPH5_AMC@NLO 2.3.3	NNPDF3.0NLO	PYTHIA 8.210	NLO+NLO (EW)
	SHERPA 2.0.0 (LO [2j])	NNPDF3.0NNLO	SHERPA	NLO+NLO (EW)
$t\bar{t}t\bar{t}$	MADGRAPH5_AMC@NLO 2.3.3	NNPDF3.1NLO	PYTHIA 8.230	NLO+NLO (EW)
tZq	MADGRAPH5_AMC@NLO 2.3.3 (LO)	CTEQ6L1	PYTHIA 8.212	–
tWZ	MADGRAPH5_AMC@NLO 2.3.3 [DR]	NNPDF3.0NLO	PYTHIA 8.230	–

Table A.1: Table summarising the generator set-ups for samples used in this analysis. The first row for each sample details the nominal settings used for this process in the analysis. Any additional rows describe samples which are used to evaluate the modelling and performance of the analysis. The precision of the ME generator is NLO in QCD if no additional information is provided in parentheses. The higher-order cross-section used to normalise these samples is listed in the last column and refers to the order of QCD processes. The labels 'lep.' ('had.') means that both bosons decay leptonically (one decays leptonically and one hadronically).

A.3 Overlap removal strategy in single-lepton regions

The overlap between the single-lepton boosted and resolved categories is studied both in the inclusive cross-section measurement and the differential cross-section measurement with the STXS formalism. Two veto strategies regarding the overlap events were examined in Asimov and data background-only fits for both measurements, in order to conclude to the choice of the priority between the two categories in the final analysis regions.

Figure A.1a shows the $k(t\bar{t}+ \geq 1b)$ normalisation factor and the signal strength uncertainties for the Asimov fit of the inclusive cross-section measurement. While no difference is observed in the normalisation factor, a decrease of about 4.7% in the sensitivity is noted when applying the resolved veto. Also, fig. A.1b shows the normalisation factor for the data background-only fit of the same measurement where no difference is observed.

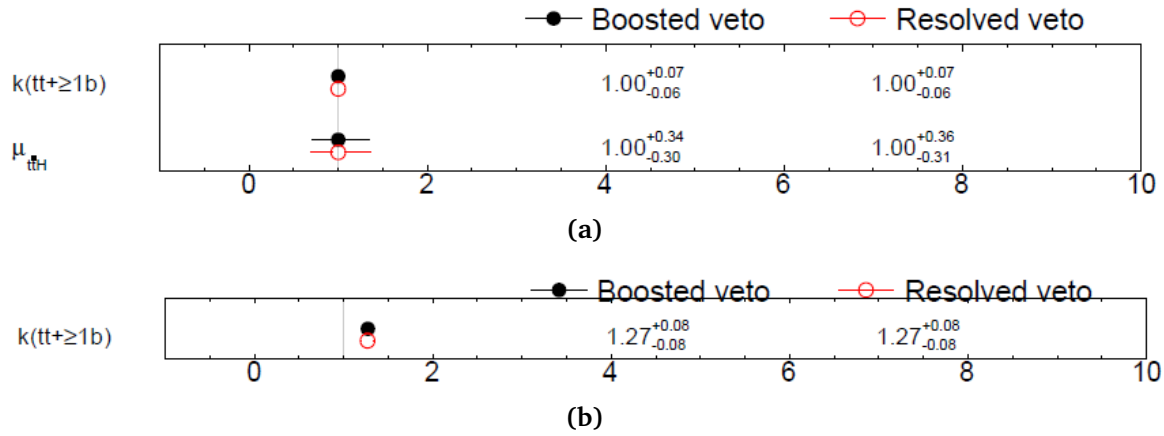


Fig. A.1: Signal strength and/or normalisation factor uncertainties for the (a) Asimov and (b) background-only fit of the inclusive cross-section measurement, comparing the boosted veto (black) and the resolved veto (red).

Then, fig. A.2a shows the normalisation factor and the signal strengths uncertainties for the Asimov fit of the STXS measurement. Again, no difference is observed in the normalisation factor, whereas there is a significant difference in sensitivity when applying the resolved veto, especially n the high p_T^H region as expected. In particular, a decrease of about 13.9% (49.3%) in sensitivity is observed in the p_T^H bin $\in [300, 450)$ ($[450, \infty)$) GeV. This is a strong motivation to give priority to the boosted category, since it has been optimised especially for the high p_T^H phase-space. Finally fig. A.2b shows the normalisation factor for the data background-only fit of the same measurement and no difference is observed.

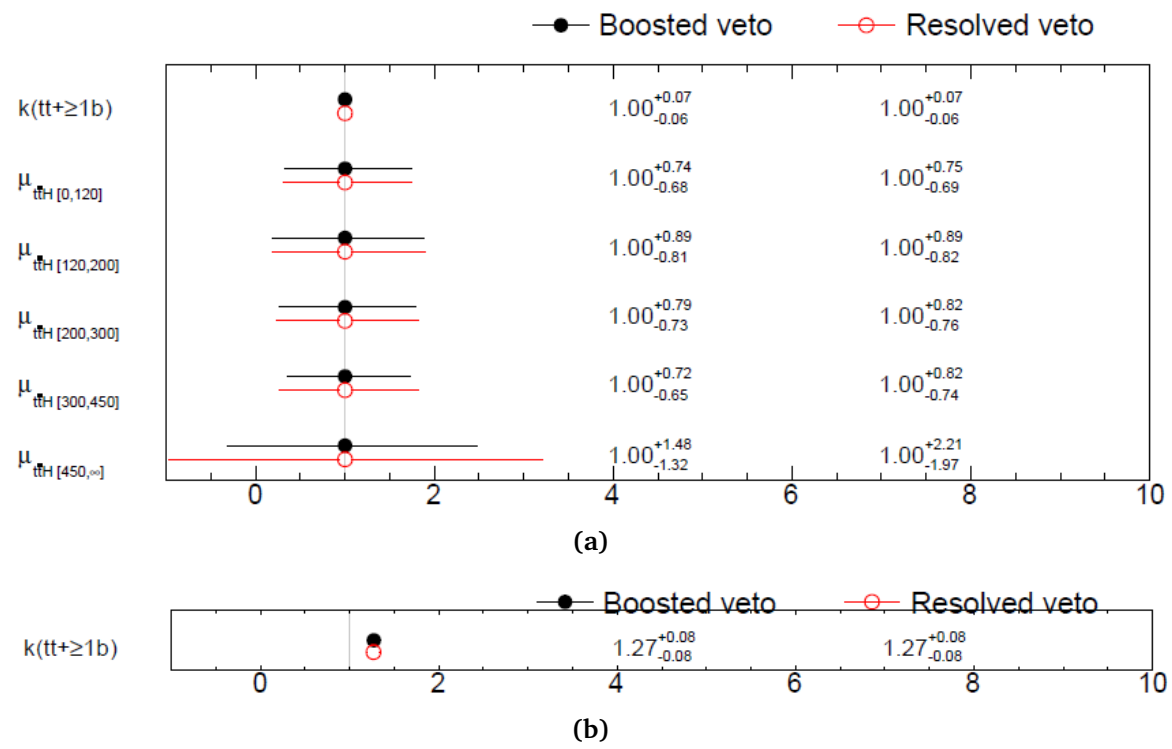


Fig. A.2: Signal strength and/or normalisation factor uncertainties for the (a) Asimov and (b) background-only fit of the STXS measurement, comparing the boosted veto (black) and the resolved veto (red).

A.4 b-tagging extrapolation uncertainties

Due to the use of multiple working points (WPs) of the MV2c10 tagger, the analysis makes use of the pseudo-continuous (PC) calibration - calibration of the MV2c10 output in 5 bins with boundaries corresponding to the efficiency working points (0%, 60%, 70%, 77%, 85%, 100%). Unlike for the cumulative WP calibration, the PC calibration does not have any high-p_T extrapolation uncertainties applied to the jets with p_T outside the calibration range. The b-tagging calibration thresholds for the corresponding true jet flavour are listed in Table 7.2.

In this appendix, the most conservative studies, which ensure that the analysis is not sensitive to the missing high-p_T extrapolation uncertainties, are summarised. The first study (Sec. A.4.1) represents the check of the sensitivity of the analysis when using high-p_T extrapolation uncertainties from the available cumulative working points. The other and very conservative test is done by checking the sensitivity of the analysis when removing all the events with at least one jet outside the calibration range (Sec. A.4.2). All tests are done by making modifications in the single-lepton channel only, while no changes were made in the dilepton channel. This should not affect the overall picture significantly given that there is no dedicated boosted category in the dilepton channel (boosted category mostly exploits the jets outside the calibration range) and that in general, the dilepton channel affects less the final combination in the high-p_T regime.

In the single-lepton channel, the events mainly have very high number of jets (6 or more in the signal-enriched bins), however the analysis had to be modified for each event that contains at least one jet outside the calibration range - given that no calibration scale factors

for individual jets are saved in the ntuples used in terms of the analysis. It was found that the fraction of these events in the boosted region is 18% for the signal (S) and 34% for the background (B) process, while in the resolved channel, 9% either of the signal or the background events have at least one jet outside the calibration range. It was also found that mainly the c-flavoured jets dominate the fraction of jets outside the calibration range, with the following break down:

In the inclusive single-lepton phase-space:

- the fraction of b-jets outside b-tagging calibration range is 0% for signal and background
- the fraction of c-jets outside b-tagging calibration range is S: 4.1% and B: 3.9%
- the fraction of $light$-jets outside b-tagging calibration range is S: 2.9% and B: 2.9%

In the boosted single-lepton phase-space:

- the fraction of b-jets outside b-tagging calibration range is 0% for signal and background
- the fraction of c-jets outside b-tagging calibration range is S: 12% and B: 16%
- the fraction of $light$-jets outside b-tagging calibration range is S: 6% and B: 9.5%

A.4.1 Using extrapolation uncertainty from cumulative *b*-tagging working points

The sensitivity to the missing extrapolation uncertainties was tested by including the available uncertainties from cumulative WPs. For each jet flavour and each tag weight bin, the largest uncertainty out of the four available working points (60%, 70%, 77% and 85%) was taken uncorrelated between tag weight bin and the jet flavour. This results in 15 new nuisance parameters (3 jet flavour times 5 bins of tag weight distribution). Table A.2 shows the values of the uncertainties included in the fit.

tag weight bin	b-jets: $p_T > 600$ GeV	c-jets: $p_T > 250$ GeV	$light$-jets: $p_T > 400$ GeV
1 (100-85 % WP)	0.092	0.015	0.016
2 (85-77 % WP)	0.054	0.037	0.049
3 (77-70 % WP)	0.048	0.043	0.070
4 (70-60 % WP)	0.027	0.054	0.095
5 (60-0 % WP)	0.029	0.079	0.237

Table A.2: The b-tagging calibration thresholds for each true jet flavour.

The inclusion of these new nuisance parameters acting on events with jets outside the calibration range did not show any effect on the expected sensitivity neither for the inclusive signal strength, nor in the STXS fit, as shown in fig. A.3.

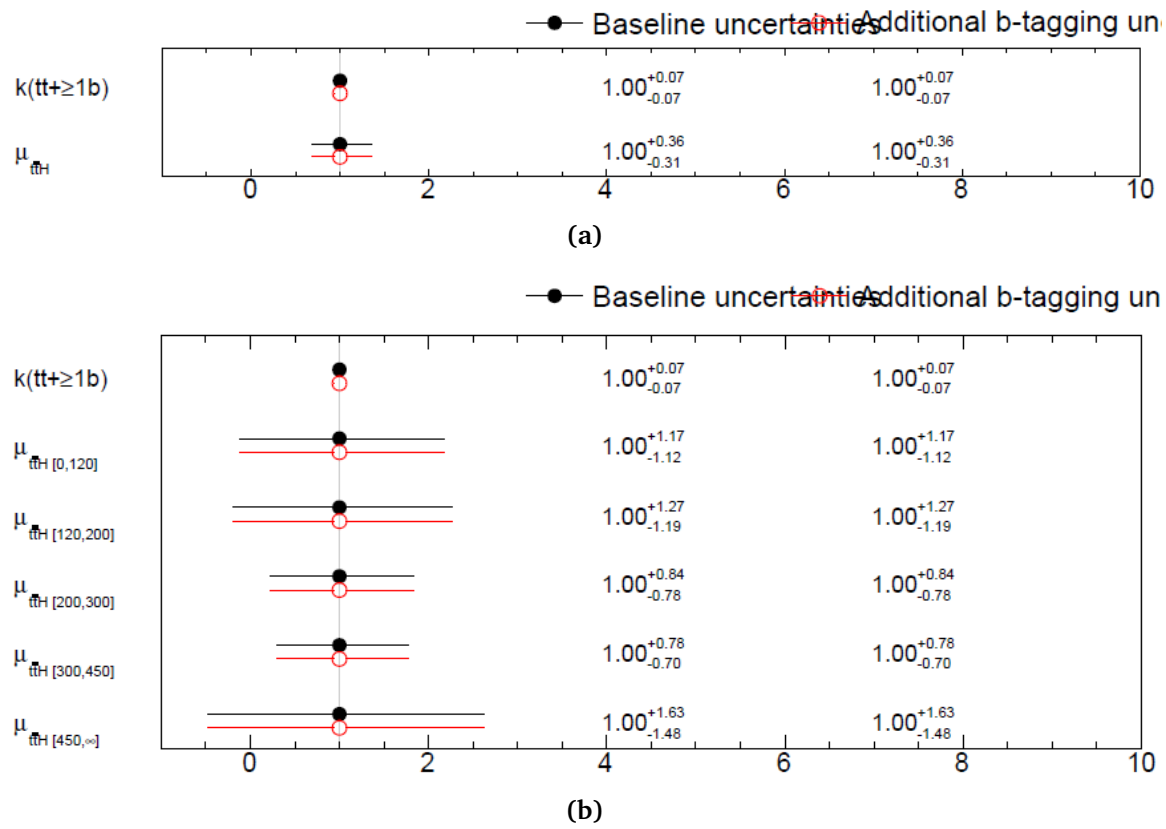


Fig. A.3: Comparison of the expected uncertainty on the signal strength for the (a) inclusive and (b) STXS measurements between the nominal (black) fit and the fit with extra high p_T extrapolation uncertainties (red). The study is done using S+B fit on the Asimov dataset.

A.4.2 Removing events with at least one jet outside the calibration range

Another study was performed to prove that there is no sensitivity on the missing high p_T b-tagging extrapolation uncertainty, in which, all events with at least one jet outside the calibration range of either flavour were removed. This study is considered very conservative given that there is very high number of jets in the events, while having for example one jet outside the calibration range should not have a huge effect in the analysis. The fraction of events removed is documented at the beginning of this Appendix. Figure A.4 shows the comparison of expected uncertainty on the signal strength between nominal fit and the fit with events with jets outside the calibration range removed, both for the inclusive and STXS fits.

For the inclusive measurement, there is no significant change in the uncertainty on the signal strength. Only a small increase of around 5% is observed in both the total and the statistical-only uncertainty. Also, there is no change in expected constraints, while there are changes in ranking of the systematic uncertainties. These are expected though, and are caused by the change in the shape of the fit input distributions, which in turn affect the shape of the modelling uncertainties.

For the STXS measurement, the change in expected uncertainties on the signal strength is slightly larger. In particular, in each STXS bin it is observed:

- [0-120] GeV: increase of 3.5% in the total and 1.0% in the statistical uncertainty

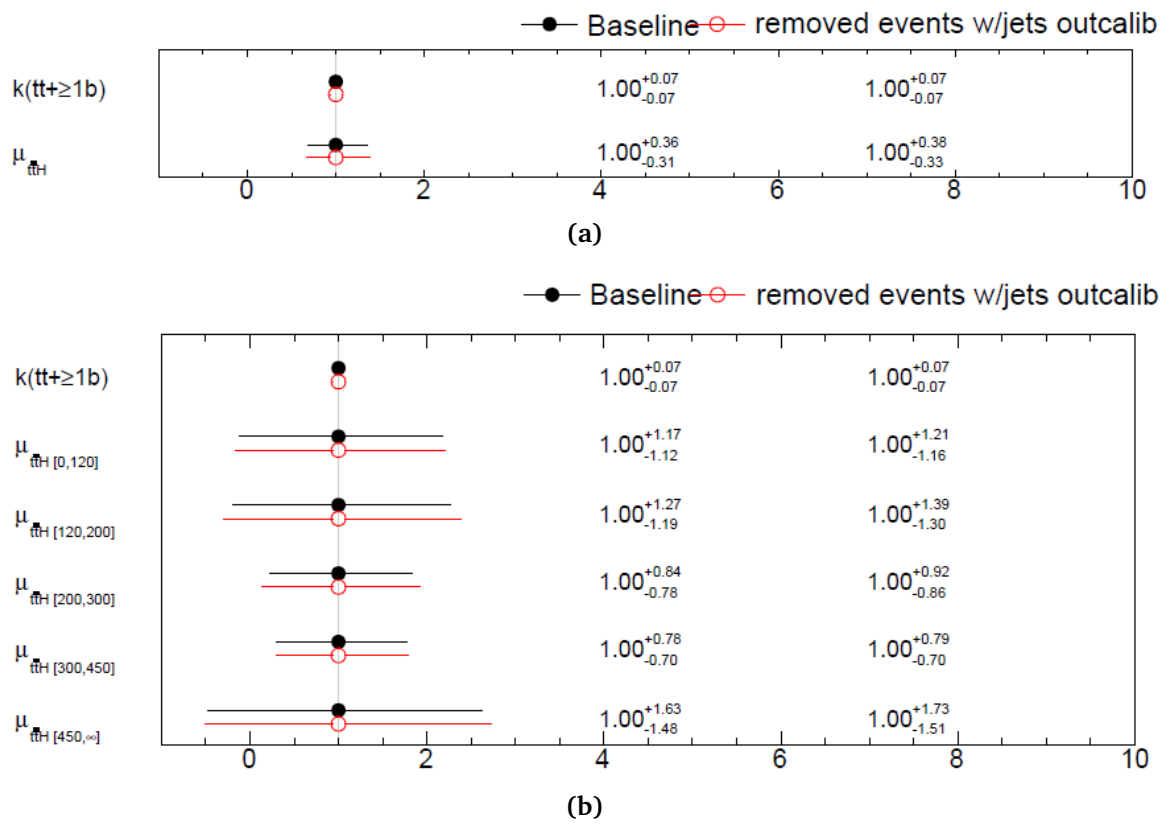


Fig. A.4: Comparison of the expected uncertainty on the signal strength for the (a) inclusive and (b) STXS measurements between the nominal (black) fit and the fit with all events containing at least one jet outside calibration range removed (red). The study is done using S+B fit on the Asimov dataset.

- [120-200] GeV: increase of 9.3% in the total and 1.9% in the statistical uncertainty
- [200-300] GeV: increase of 9.9% in the total and 4.0% in the statistical uncertainty
- [300-450] GeV: increase of 0.1% in the total and 2.5% in the statistical uncertainty
- [450-inf] GeV: increase of 4.2% in the total and 4.1% in the statistical uncertainty

The change in the statistical uncertainty is order of few per-cent, while the largest increase in total uncertainty is observed in bins [120-200] GeV and [200-300] GeV, mainly due to the increase of the effect of the modelling systematic uncertainties. No changes were found in the expected constraints, while the changes in the correlation matrix and ranking of the systematic uncertainties are found to be consistent with the changes in the uncertainty on the signal strengths. These changes, in fact, originate from the change in the shape of the modelling systematic uncertainties, as mentioned earlier.

A.5 Shape of more systematic uncertainties

Examples of $t\bar{t}H$ signal modelling or experimental systematic variations, that have much smaller impact on the measurement, demonstrate comparable or smaller shape and/or normalisation effects to those described in Sec. 7.2.4.

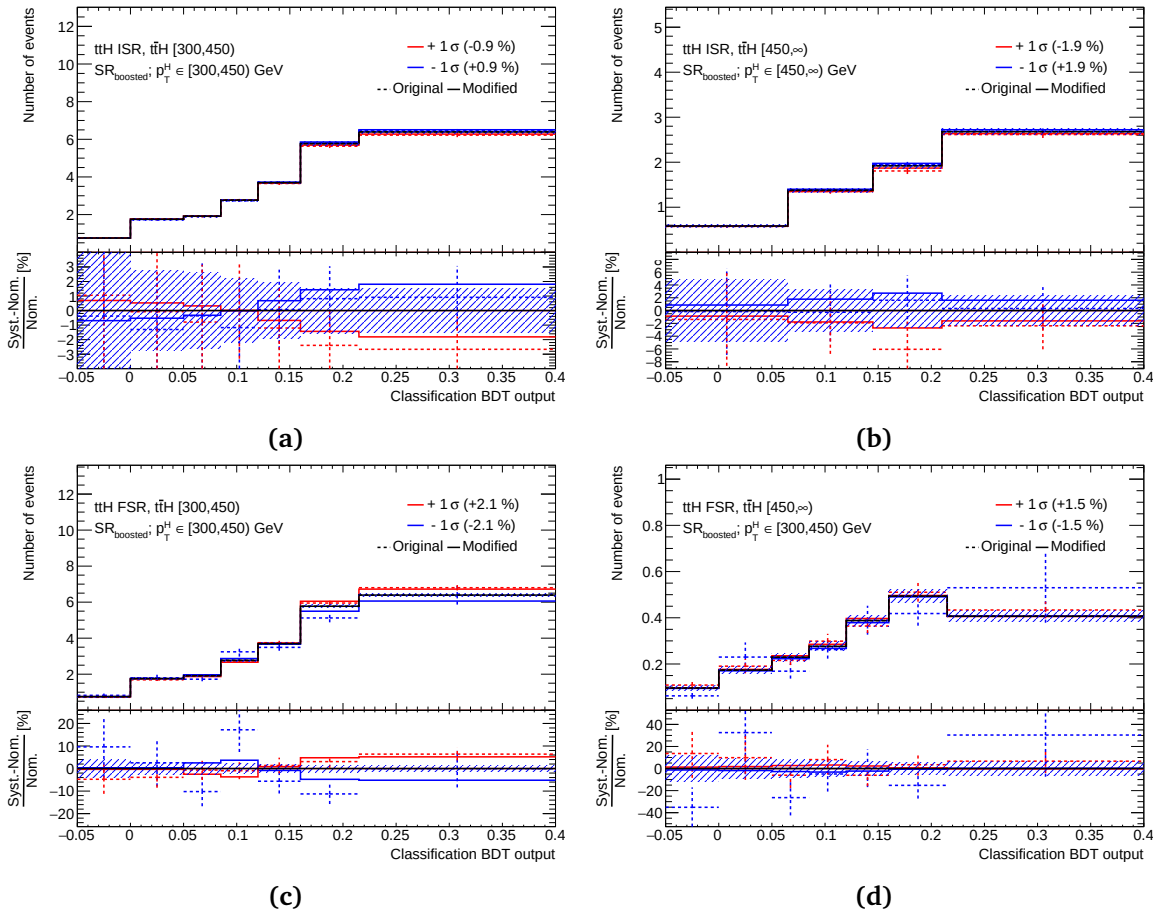


Fig. A.5: Comparison of the nominal prediction (black) with the one standard deviation up (red) and down (blue) variations induced by the (a),(b) $t\bar{t}H$ ISR and (c),(d) $t\bar{t}H$ FSR uncertainties on the $t\bar{t} + \geq 1b$ sample for the classification BDT distribution in the single-lepton boosted signal region ($SR_{boosted}$) in the $[300, 450)$ GeV (left) and $[450, \infty)$ GeV (right) p_T^H bins. *Original* (dashed line) refers to the raw input distribution, while *modified* (solid line) is the distribution after symmetrisation and smoothing.

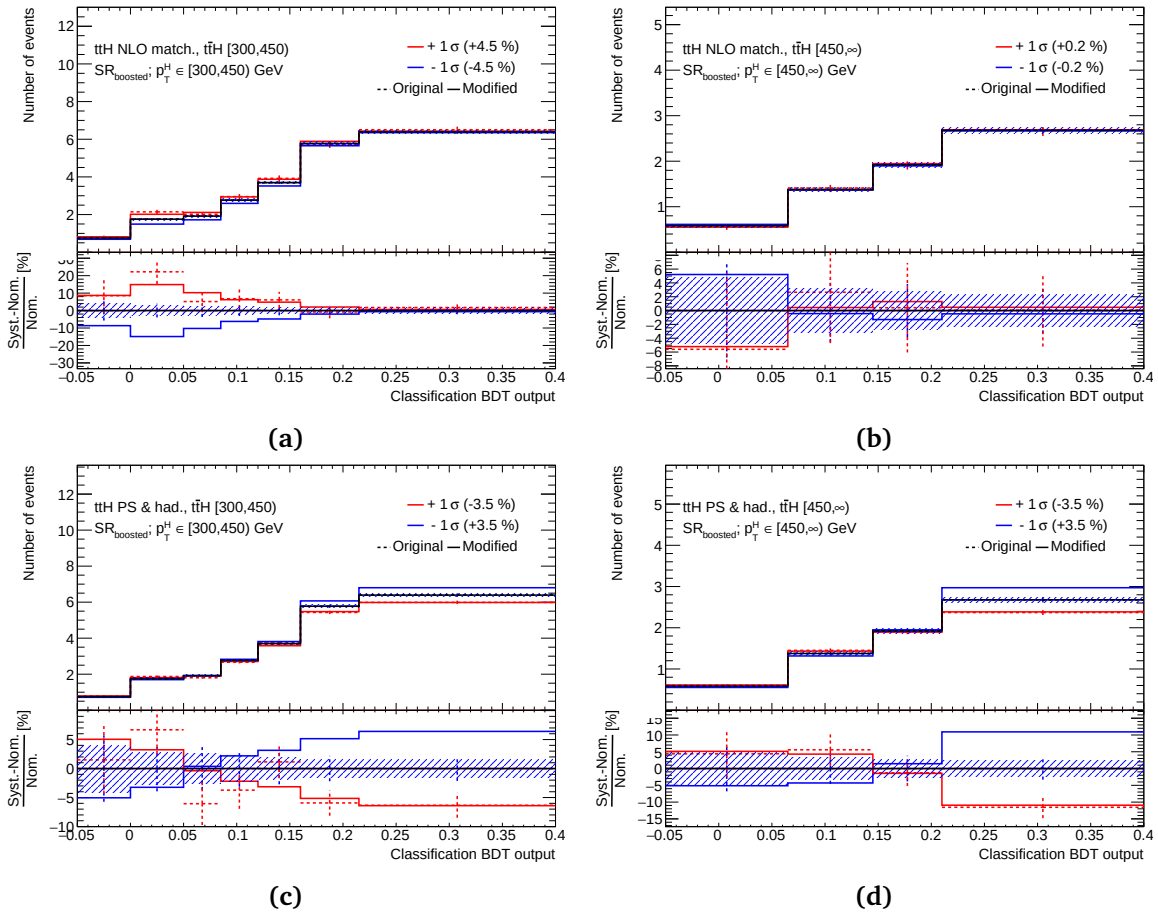


Fig. A.6: Comparison of the nominal prediction (black) with the one standard deviation up (red) and down (blue) variations induced by the (a),(b) $t\bar{t}H$ NLO matching (c),(d) $t\bar{t}H$ PS & hadronisation uncertainties on the $t\bar{t}H$ sample for the classification BDT distribution in the single-lepton boosted signal region ($SR_{boosted}$) in the $[300, 450)$ GeV (left) and $[450, \infty)$ GeV (right) p_T^H bins. *Original* (dashed line) refers to the raw input distribution, while *modified* (solid line) is the distribution after symmetrisation and smoothing.

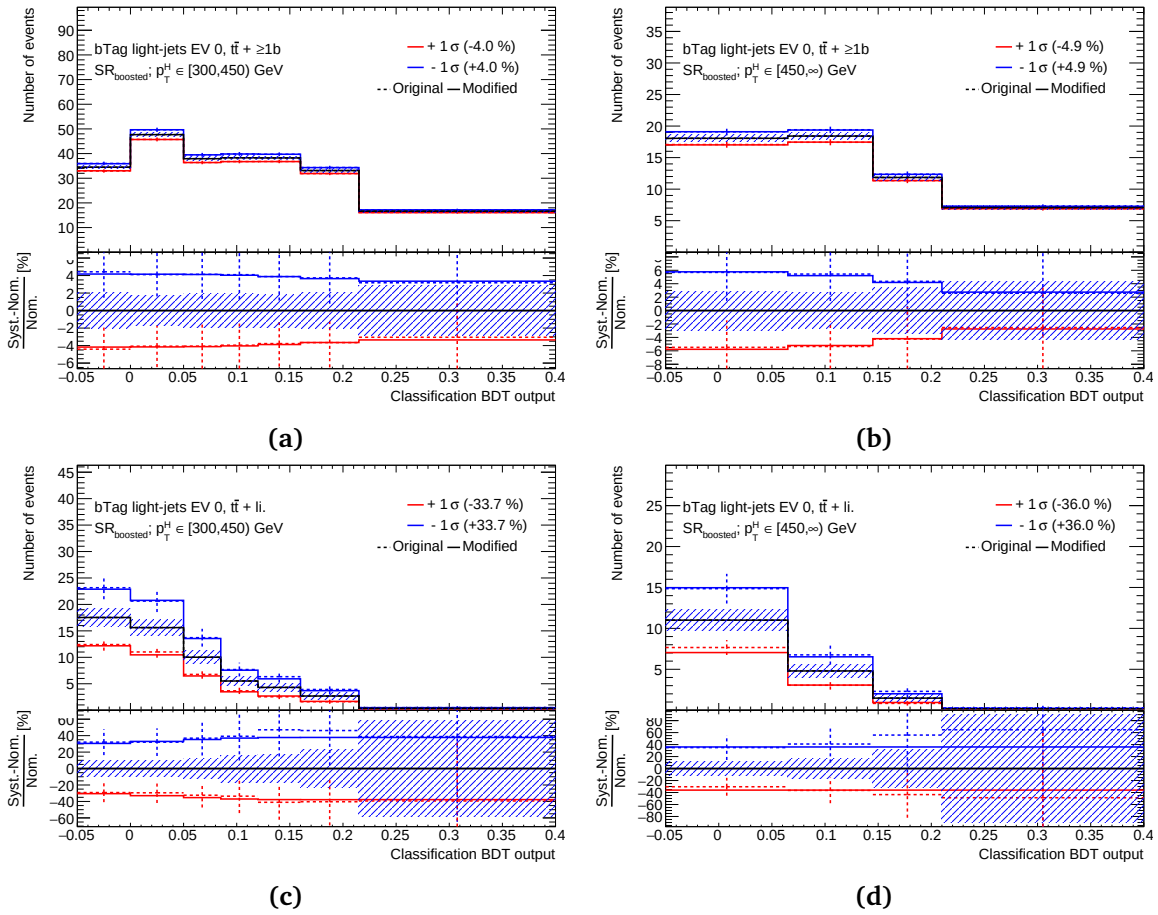


Fig. A.7: Comparison of the nominal prediction (black) with the one standard deviation up (red) and down (blue) variations induced by the "bTag light-jets EV 0" experimental uncertainty on the (a),(b) $t\bar{t}+\geq 1b$ and (c),(d) $t\bar{t}+light$ samples for the classification BDT distribution in the single-lepton boosted signal region ($SR_{boosted}$) in the $[300, 450)$ GeV (left) and $[450, \infty)$ GeV (right) p_T^H bins. *Original* (dashed line) refers to the raw input distribution, while *modified* (solid line) is the distribution after symmetrisation and smoothing.

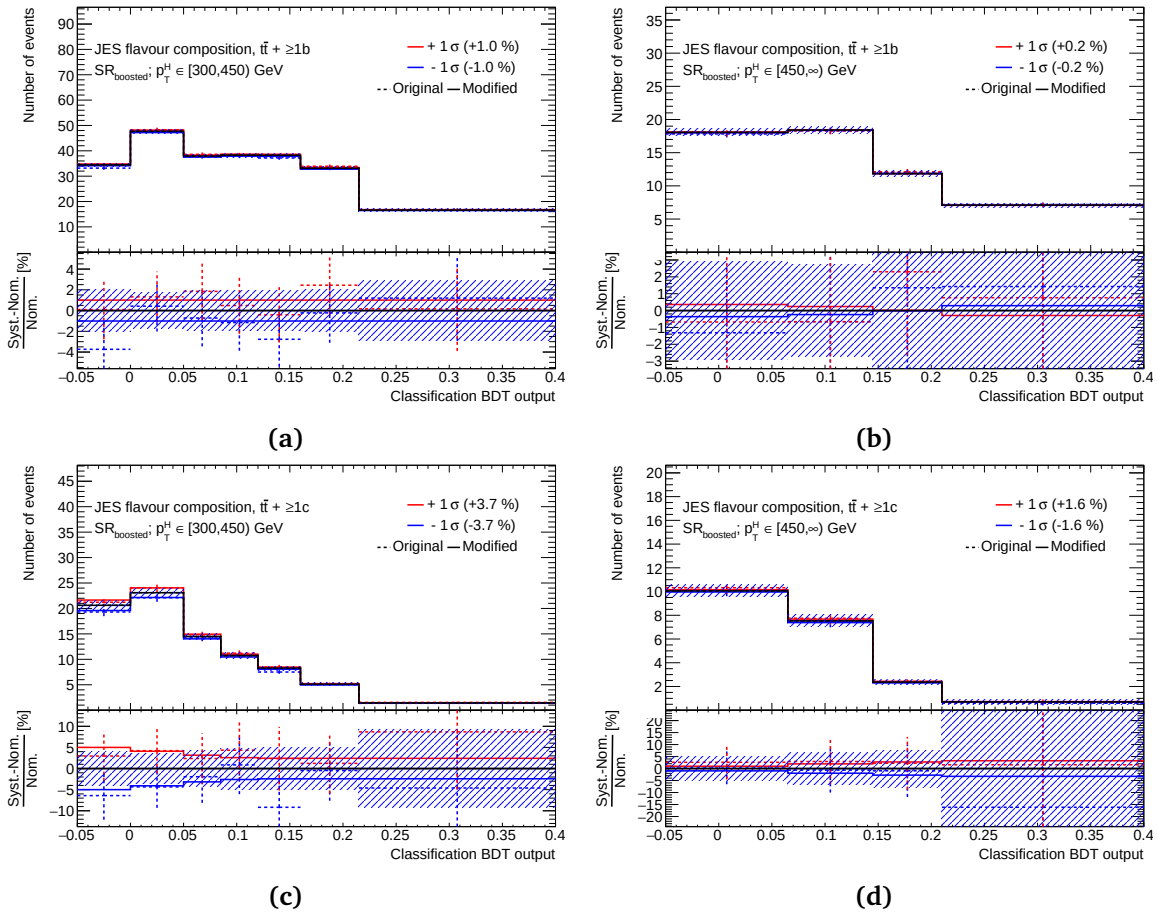


Fig. A.8: Comparison of the nominal prediction (black) with the one standard deviation up (red) and down (blue) variations induced by the "JES flavour composition" experimental uncertainty on the (a),(b) $t\bar{t}+\geq 1b$ and (c),(d) $t\bar{t}+\geq 1c$ samples for the classification BDT distribution in the single-lepton boosted signal region ($SR_{boosted}$) in the $[300, 450)$ GeV (left) and $[450, \infty)$ GeV (right) p_T^H bins. *Original* (dashed line) refers to the raw input distribution, while *modified* (solid line) is the distribution after symmetrisation and smoothing.

A.6 Complementary results

The correlation matrix of the nuisance parameters, the $k(t\bar{t}+\geq 1b)$ normalisation factor, and the signal strength, containing the correlation coefficients of a selection of the systematic uncertainty sources obtained by the combined inclusive fit to Asimov dataset, is shown in fig. A.9. In general, there are only a few large (anti-)correlations (e.g. $\geq 45\%$) among the different parameters and most of them are between the NPs related to the $t\bar{t}+\geq 1b$ background uncertainties, or specifically to the $t\bar{t}+\geq 1c$ normalisation uncertainty. Also, only a few NPs associated with the experimental uncertainties are present in the matrix experiencing mostly subtle correlations among each other. The most considerable correlations of the signal strength are with a few $t\bar{t}+\geq 1b$ background NPs, they are not significantly large though.

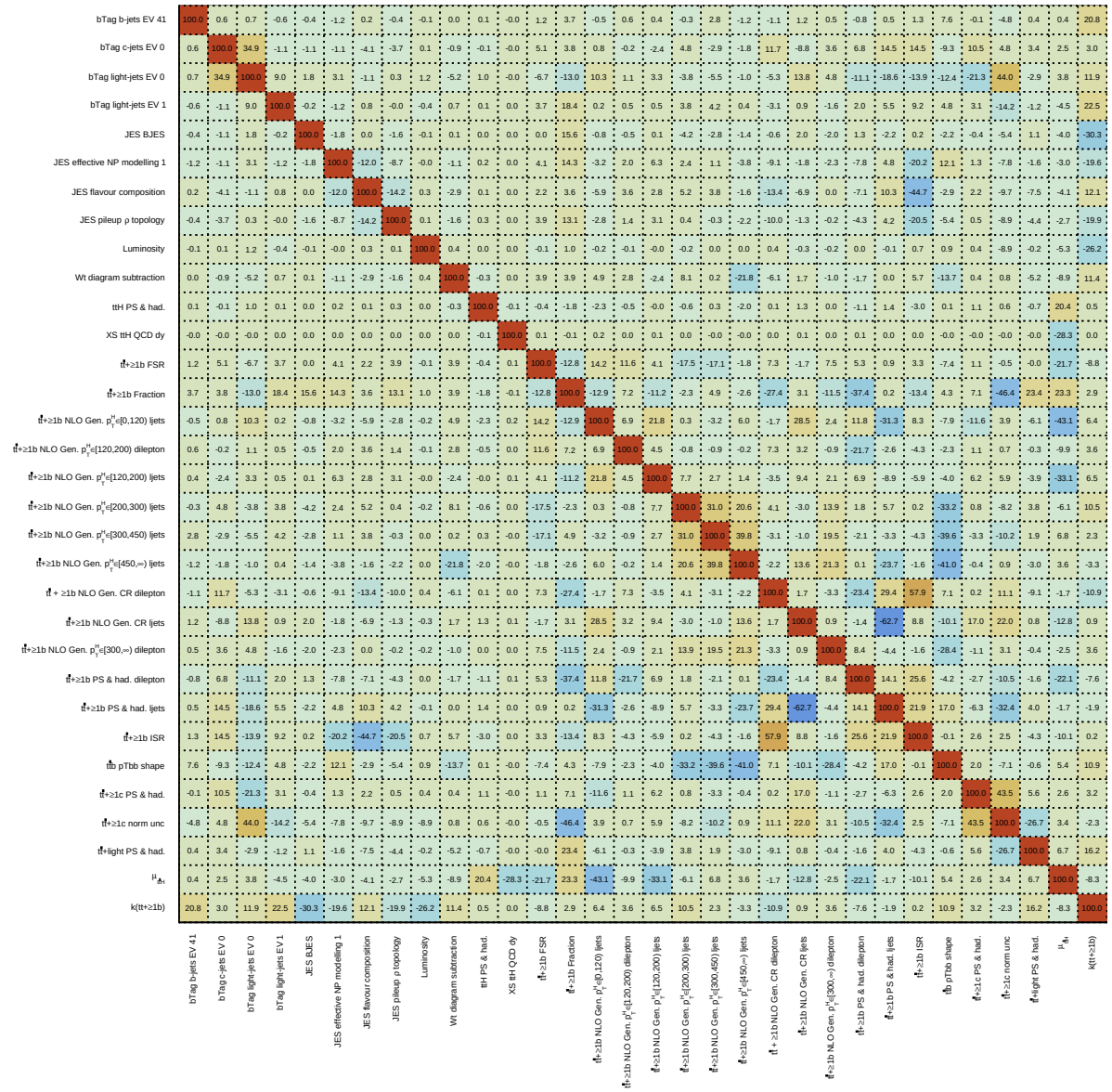


Fig. A.9: Linear correlation coefficients of the nuisance parameters, normalisation factor, and signal strength, after the combined inclusive-μ fit to the Asimov dataset. All values are given in percent. Only nuisance parameters with at least one absolute correlation coefficient above 20% are shown.

Figure A.10 shows the fitted nuisance parameters in the S+B combined fit on Asimov dataset from the inclusive cross-section measurement. As already introduced, no pulls are expected for NPs from this fit. The modelling $t\bar{t}H$ signal systematics are not constrained at all, implying low sensitivity to the $t\bar{t}H$ modelling within the given precision. Also, the NPs of the non-$t\bar{t}$ background systematic uncertainties are not at all constraint. On the contrary, most of the NPs related to the $t\bar{t}$+jets background exhibit large constraints. In fact, these related to the $t\bar{t}$+light and $t\bar{t}+\geq 1c$ background modelling remain almost unconstrained, given the small contribution of these background components in the analysis regions. The only exception is the $t\bar{t}+\geq 1c$ normalisation uncertainty which is intensely constraint, probably due to the inclusion of the dilepton CRs which contain a relatively large amount of $t\bar{t}+\geq 1c$ background.

Then, most of the constraints are associated to the $t\bar{t}+\geq 1b$ background. This is expected, since the most sensitive analysis regions are dominated by the $t\bar{t}+\geq 1b$ background, hence the model is able to constrain the various $t\bar{t}+\geq 1b$ modelling uncertainties and possibly compensate for any mis-modelling effects. Lastly, almost none of the NPs related to the performance of the detector (instrumental NPs) develop any constraints with respect to their prior uncertainties. Only the systematic uncertainties related to the first eigen-variation of the b-tagging mistag rate for $light$-jets (b-tag $light$-jets EV 0) and for c-jets (b-tag c-jets EV 0). No notable impact is expected on the sensitivity from these NPs, though, since the definition of the most sensitive signal regions require at least four b-tagged jets using tight b-tagging operating points. These operating points have very high rejection of $light$-jets as well as of c-jets, as mentioned in Sec. 5.3.1.

Figure A.11 shows the fitted nuisance parameters in the S+B combined fit on data from the STXS measurement. There are no significant differences with respect to that obtain from the inclusive cross-section measurement.

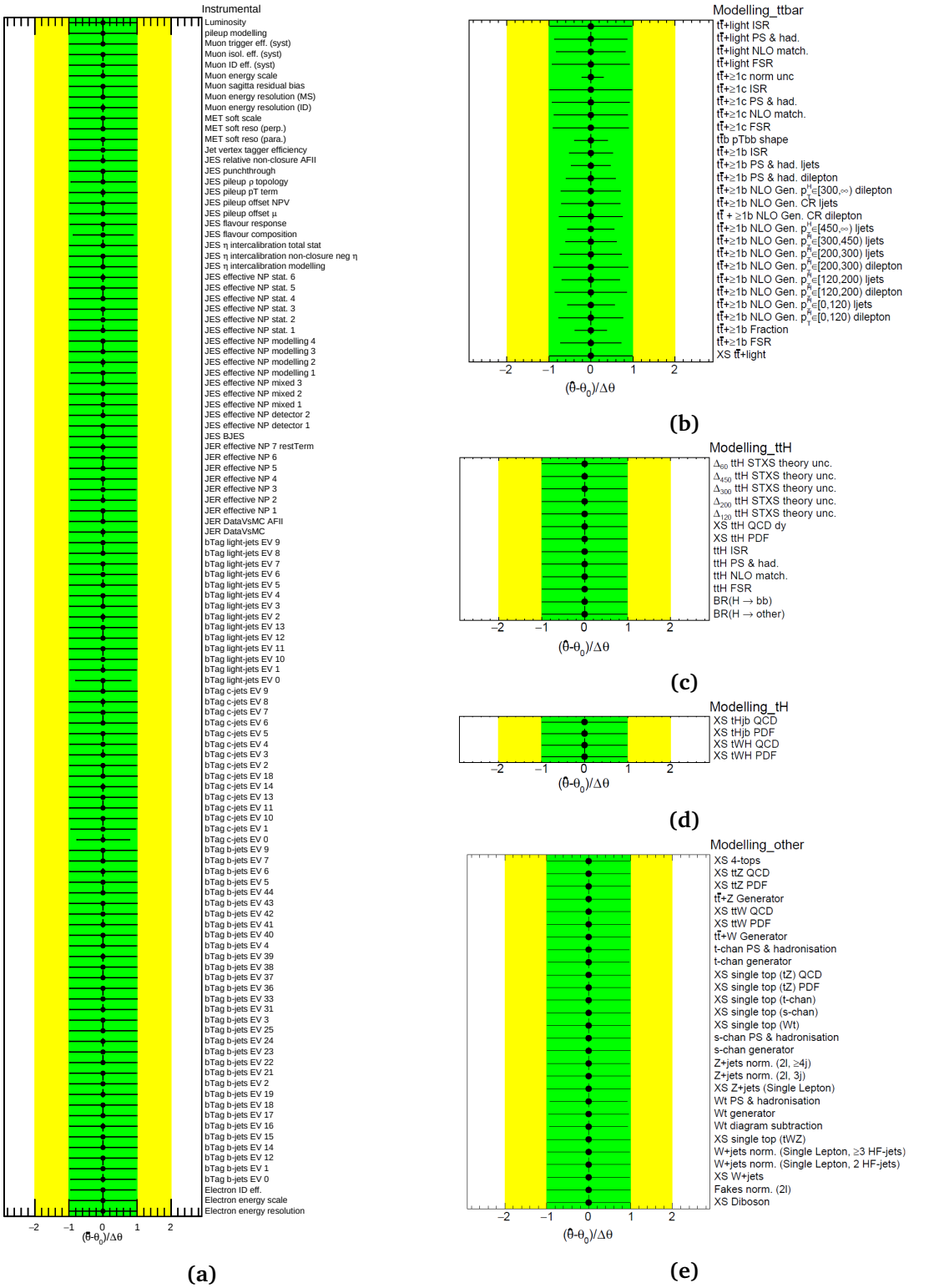


Fig. A.10: Fitted nuisance parameters of the (a) instrumental, (b) $t\bar{t}$ + jets modelling, (c) $t\bar{t}H$ modelling, (d) tH modelling, and (e) other backgrounds modelling systematic uncertainties from the combined inclusive-μ fit to the Asimov dataset. The green (yellow) area represents the $\pm 1(2)\sigma$ band on the pre-fit systematic uncertainty. The position of the black points shows the pull of the nuisance parameters, i.e. their best-fit value $\hat{\theta}$ relative to their nominal values, θ_0. The size of the horizontal bars give the constraint of the nuisance parameters, i.e. their post-fit error relative to the pre-fit one, $\Delta\hat{\theta}/\Delta\theta$. Both values are given in units of standard deviation (σ).

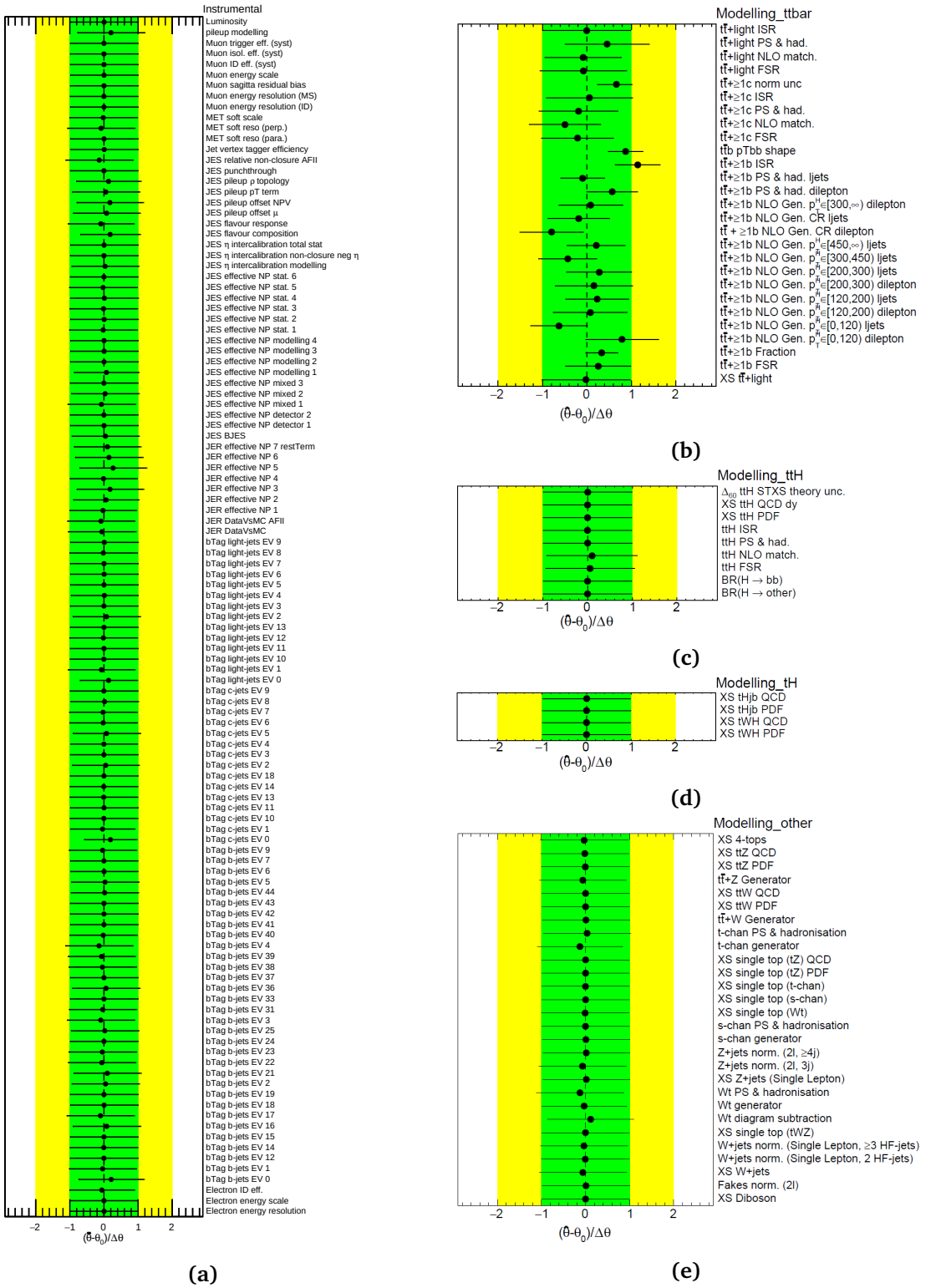


Fig. A.11: Fitted nuisance parameters of the (a) instrumental, (b) $t\bar{t}$ + jets modelling, (c) $t\bar{t}H$ modelling, (d) tH modelling, and (e) other backgrounds modelling systematic uncertainties from the combined STXS fit to data. The green (yellow) area represents the $\pm 1(2)\sigma$ band on the pre-fit systematic uncertainty. The position of the black points shows the pull of the nuisance parameters, i.e. their best-fit value $\hat{\theta}$ relative to their nominal values, θ_0. The size of the horizontal bars give the constraint of the nuisance parameters, i.e. their post-fit error relative to the pre-fit one, $\Delta\hat{\theta}/\Delta\theta$. Both values are given in units of standard deviation (σ).

The post-fit modelling of a few most highly ranked input variables to the classification BDT in the single-lepton boosted channel is illustrated in fig. A.13 and A.12. Overall they show a very good data to MC agreement, although there are still a few bins where the prediction still differs from the measured data. This effect does not point to any trend though, and it is attributed to statistical fluctuations.

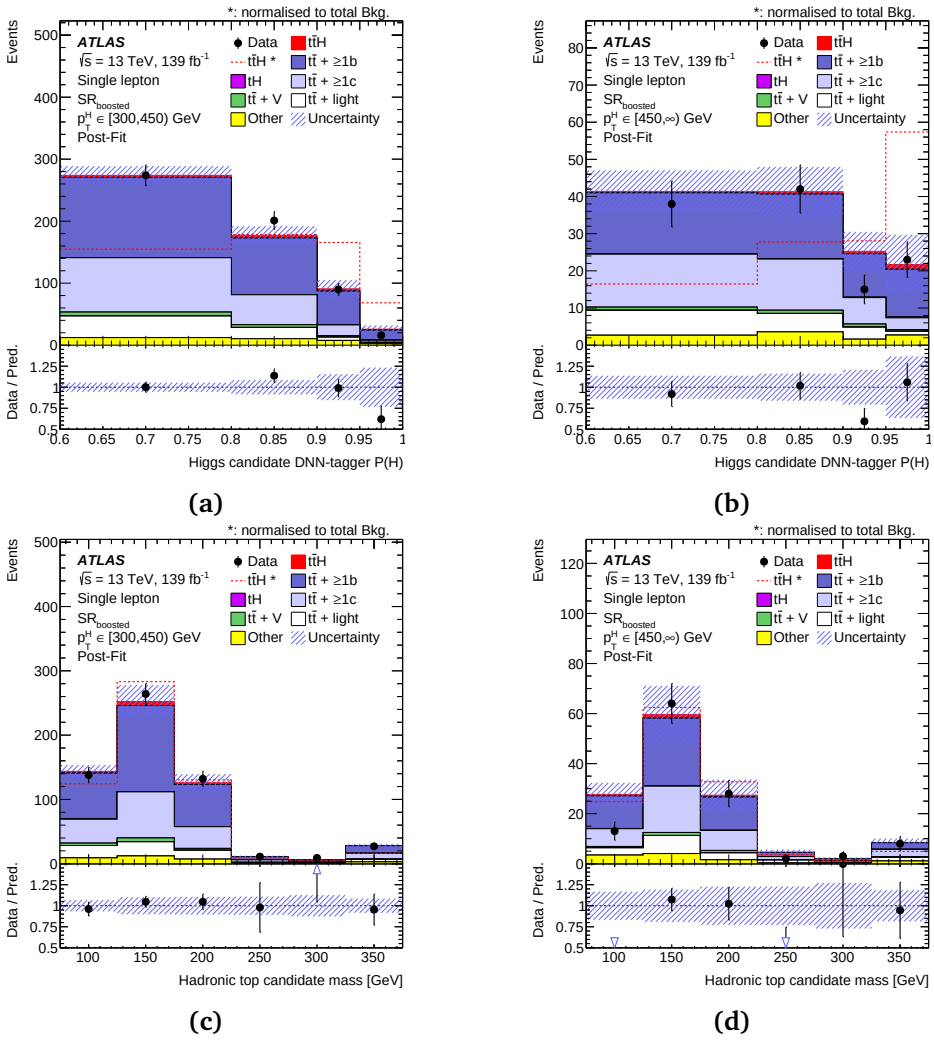
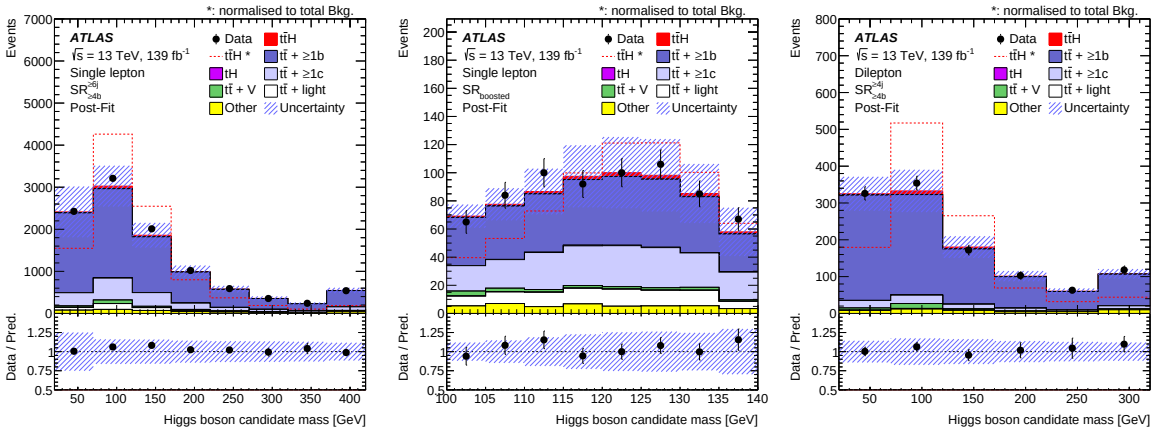


Fig. A.12: Post-fit distributions of (a),(b) the DNN $P(H)$ output for the Higgs-boson candidate as well as (c),(d) the hadronic-top invariant mass in the single-lepton boosted $SR_{boosted}$ for the $300 \leq p_T^H < 450$ GeV (left) and $p_T^H \geq 450$ GeV (right) regions, respectively [100]. The $t\bar{t}H$ signal yield (solid red) is normalised to the fitted μ value from the inclusive fit. The dashed line shows the $t\bar{t}H$ signal distribution normalised to the total background prediction. The uncertainty band includes all uncertainties as well as their correlations. The first (last) bin includes the underflow (overflow).



(a) **(b)** **(c)**

Fig. A.13: Post-fit distributions of the reconstructed Higgs boson candidate mass for the (a) single-lepton resolved $SR^{\geq 6j}_{\geq 4b}$, (b) single-lepton boosted $SR_{boosted}$, and (c) dilepton $SR^{\geq 4j}_{\geq 4b}$ signal regions [100]. The $t\bar{t}H$ signal yield (solid red) is normalised to the fitted μ value from the inclusive fit. The dashed line shows the $t\bar{t}H$ signal distribution normalised to the total background prediction. The uncertainty band includes all uncertainties as well as their correlations. The first (last) bin includes the underflow (overflow).

Cross-section upper limits are also derived in the STXS framework. In this case, the likelihood function is slightly different from the one used to extract signal strengths: the effects of signal scale and PDF uncertainties on the predicted cross-section are not included because, while affecting the signal-strength measurements, they do not affect the cross-section measurements. Scale effects are still present in the statistical model though, via the ISR uncertainty, but with no impact on the overall cross-section. The inclusive cross-section of 507 fb is used to calculate these limits, scaled by the fraction of events in each $\hat{p}_T^H$ bin to establish the fiducial cross-section for each STXS bin. The measured 95% confidence level (CL) cross-section upper limits in each STXS bin are shown in fig. A.14, where the hatched uncertainty bands correspond to the theoretical uncertainty in the fiducial cross-section prediction in each bin.

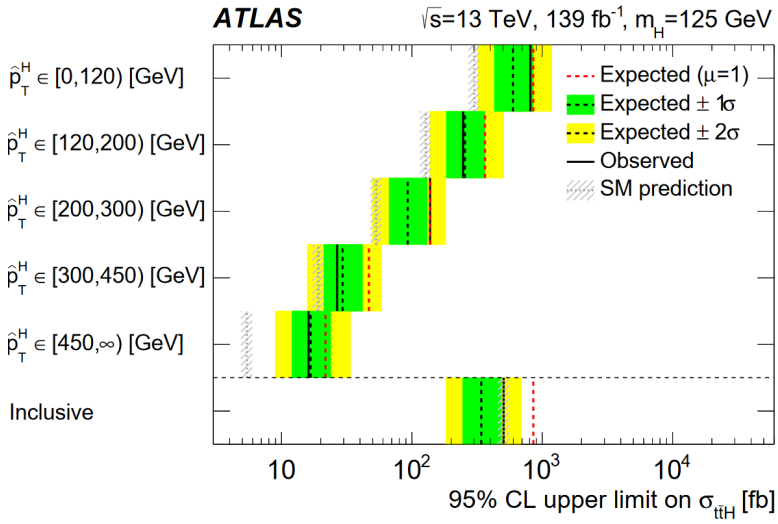


Fig. A.14: Summary of the 95% confidence level (CL) upper limits on cross-section in the individual STXS $\hat{p}_T^H$ bins as well as in the inclusive measurement, after the combined fit to data [100]. The observed limits are shown (solid black lines), together with the expected limits both in the background-only hypothesis (dotted black lines) and in the SM hypothesis (dotted red lines). In the case of the expected limits in the background-only hypothesis, 1σ and 2σ uncertainty bands are also shown. The hatched uncertainty bands correspond to the theory uncertainty in the fiducial cross-section prediction in each bin.

List of Figures

List of Tables

Bibliography

[1] Weinberg S., "Gravitation and cosmology: principles and applications of the general theory of relativity", John Wiley & Sons, Inc. (1972), ISBN: 9780471925675

[2] Griffiths D. J., "Introduction to quantum mechanics", Pearson Education (1995), ISBN: 9781107179868

[3] "Standard Model of Elementary Particles", accessed: 1 February 2022, https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles_dark.svg

[4] J. Goldstone, A. Salam, and S. Weinberg, "Broken Symmetries", Phys. Rev. 127 (1962) 965-970, https://doi.org/10.1103/PhysRev.127.965

[5] P. W. Higgs, "Broken symmetries, massless particles and gauge fields", Phys. Lett. 12 (1964) 132-133, https://doi.org/10.1016/0031-9163(64)91136-9

[6] Englert F. and Brout R., "Broken Symmetry and the Mass of Gauge Vector Mesons", Phys. Rev. Lett. 13 (1964) 321, https://doi.org/10.1103/PhysRevLett.13.321

[7] Higgs P. W., "Broken Symmetries and the Masses of Gauge Bosons", Phys. Rev. Lett. 13 (1964) 508, https://doi.org/10.1103/PhysRevLett.13.508

[8] G. Guralnik, C. Hagen and T. Kibble, "Global Conservation Laws and Massless Particles", Phys. Rev. Lett. 13 (1964) 585, https://doi.org/10.1103/PhysRevLett.13.585

[9] P. W. "Higgs, Spontaneous Symmetry Breakdown without Massless Bosons", Phys. Rev. 145 (1966) 1156-1163, https://doi.org/10.1103/PhysRev.145.1156

[10] The 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, https://doi.org/10.1016/j.physletb.2012.08.020

[11] The 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, https://doi.org/10.1016/j.physletb.2012.08.021

[12] The ATLAS and CMS Collaborations, "Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments, Phys. Rev. Lett. 114 (2015) 191803, https://doi.org/10.1103/PhysRevLett.114.191803

[13] 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, https://doi.org/10.1007/JHEP08(2016)045

[14] Halzen F. and Martin A. D., "QUARKS AND LEPTONS: An Introductory Course In Modern Particle Physics", 978-0-471-88741-6, John Wiley & Sons (1984) USA, https://archive.org/details/QuarksAndLeptonsAnIntroductoryCourseInModernParticlePhysicsHalzenMartin

[15] Peskin M. E. and Schroeder D. V., "An Introduction to quantum field theory", 978-0-201-50397-5, Addison-Wesley (1995) Reading USA, https://doi.org/10.1201/9780429503559

[16] Aitchison I. J. R. and Hey A. J. G., "Gauge Theories in Particle Physics: A Practical Introduction, Volume 1: From Relativistic Quantum Mechanics to QED", 4^{th} Edition, 978-0-429-18538-0, Taylor & Francis (2013), https://doi.org/10.1201/b13717

[17] Ellis R. K., Stirling W. J., and Webber B. R., "QCD and Collider Physics", 978-0-521-54589-1, Cambridge University Press (2003), https://doi.org/10.1017/CBO9780511628788

[18] J. Schwinger, ed., "Selected Papers on Quantum Electrodynamics", Dover Publications (1958), ISBN: 978-0-486-60444-2

[19] S. L. Glashow, "Partial Symmetries of Weak Interactions", Nucl. Phys. 22 (1961) 579-588, https://doi.org/10.1016/0029-5582(61)90469-2

[20] S. Weinberg, "A Model of Leptons", Phys. Rev. Lett. 19 (1967) 1264-1266, https://doi.org/10.1103/PhysRevLett.19.1264

[21] A. Salam and J. C. Ward, "Electromagnetic and Weak Interactions", Physics Letters 13 (1964) p. 168, https://doi.org/10.1016/0031-9163(64)90711-5

[22] S. Weinberg, Non-Abelian Gauge Theories of the Strong Interactions, Phys. Rev. Lett. 31 (1973) 494, https://doi.org/10.1103/PhysRevLett.31.494

[23] The D0 Collaboration, "Observation of the Top Quark", Phys. Rev. Lett. 74 (1995) 2632, https://doi.org/10.1103/PhysRevLett.74.2632

[24] The CDF Collaboration, Observation of top quark production in $p\bar{p}$ collisions, Phys. Rev. Lett. 74 (1995) 2626-2631, https://doi.org/10.1103/PhysRevLett.74.2626

[25] C. N. Yang and R. L. Mills, "Conservation of Isotopic Spin and Isotopic Gauge Invariance", Phys. Rev. 96 (1954) 191, https://doi.org/10.1103/PhysRev.96.191

[26] E. Noether, "Invariant Variation Problems", Gott. Nachr. 1918 (1918) 235–257, https://doi.org/10.1080/00411457108231446

[27] R. P. Feynman, "Space-Time Approach to Non-Relativistic Quantum Mechanics", Rev. Mod. Phys. 20 (1948) 367, https://doi.org/10.1103/RevModPhys.20.367

[28] R. P. Feynman, "Relativistic Cut-Off for Quantum Electrodynamics", Phys. Rev. 74 (1948) 1430, https://doi.org/10.1103/PhysRev.74.1430

[29] S. Kanesawa and S.-I. Tomonaga, "On a Relativistically Invariant Formulation of the Quantum Theory of Wave Fields. V: Case of Interacting Electromagnetic and Meson Fields", Progress of Theoretical Physics 3 (1948) 1, https://doi.org/10.1143/ptp/3.1.1

[30] S.-I. Tomonaga and J. R. Oppenheimer, "On Infinite Field Reactions in Quantum Field Theory", Phys. Rev. 74 (1948) 224, https://doi.org/10.1103/PhysRev.74.224

[31] J. Schwinger, "On Quantum-Electrodynamics and the Magnetic Moment of the Electron", Phys. Rev. 73 (1948) 416, https://doi.org/10.1103/PhysRev.73.416

[32] J. Schwinger, "Quantum Electrodynamics. I. A Covariant Formulation", Phys. Rev. 74 (1948) 1439, https://doi.org/10.1103/PhysRev.74.1439

[33] F. J. Dyson, "The Radiation Theories of Tomonaga, Schwinger, and Feynman", Phys. Rev. 75 (1949) 486, https://doi.org/10.1103/PhysRev.75.486

[34] F. J. Dyson, "The S Matrix in Quantum Electrodynamics", Phys. Rev. 75 (1949) 1736, https://doi.org/10.1103/PhysRev.75.1736

[35] S. Weinberg, "General Theory of Broken Local Symmetries", Phys. Rev. D 7 (1973) p. 1068, https://doi.org/10.1103/PhysRevD.7.1068

[36] Kumericki K., "Feynman Diagrams for Beginners", arXiv:1602.04182 [physics.ed-ph] (2016), https://doi.org/10.48550/arXiv.1602.04182

[37] F. L. Wilson, "Fermi's Theory of Beta Decay", American Journal of Physics 36 (1968) 1150, https://doi.org/10.1119/1.1974382

[38] P. A. M. Dirac, "The Quantum Theory of the Emission and Absorption of Radiation", Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 114 (1927) 243, https://doi.org/10.1098/rspa.1927.0039

[39] T. D. Lee and C. N. Yang, "Question of Parity Conservation in Weak Interactions", Phys. Rev. 104 (1956) 254, https://doi.org/10.1103/PhysRev.104.254

[40] C. S. Wu, E. Ambler, R. W. Hayward, et al., "Experimental Test of Parity Conservation in Beta Decay", Phys. Rev. 105 (1957) 1413, https://doi.org/10.1103/PhysRev.105.1413

[41] P. A. M. Dirac, "The Quantum Theory of the Electron", Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 117 (1928) 610, https://doi.org/10.1098/rspa.1928.0023

[42] R. P. Feynman and M. Gell-Mann, "Theory of the Fermi Interaction", Phys. Rev. 109 (1958) 193, https://doi.org/10.1103/PhysRev.109.193

[43] M. Gell-Mann, "Test of the Nature of the Vector Interaction in β Decay", Phys. Rev. 111 (1958) 362, https://doi.org/10.1103/PhysRev.111.362

[44] S. Okubo, R. E. Marshak, E. C. G. Sudarshan, et al., "Interaction Current in Strangeness-Violating Decays", Phys. Rev. 112 (1958) 665, https://doi.org/10.1103/PhysRev.112.665

[45] E. C. G. Sudarshan and R. E. Marshak, "The nature of the four-fermion interaction", Current Science, 63 2 (1994) 65-75, https://doi.org/10.1142/9789812831408_0046

[46] M. Kobayashi and T. Maskawa, "CP Violation in the Renormalizable Theory of Weak Interaction", Prog. Theor. Phys. 49 (1973) 652-657, https://doi.org/10.1143/PTP.49.652

[47] N. Cabibbo, "Unitary Symmetry and Leptonic Decays", Phys. Rev. Lett. 10 (1963) 531, https://doi.org/10.1103/PhysRevLett.10.531

[48] The CKMfitter group, "CP violation and the CKM matrix: assessing the impact of the asymmetric B factories", The European Physical Journal C - Particles and Fields 41 (2005) 1, http://ckmfitter.in2p3.fr

[49] M. Gell-Mann, "The interpretation of the new particles as displaced charge multiplets", Nuovo Cimento 4 (Suppl 2), (1956) 848 https://doi.org/10.1007/BF02748000

[50] T. Nakano and K. Nishijima, "Charge Independence for V-particles*", Progress of Theoretical Physics 10 (1953) 581, https://doi.org/10.1143/PTP.10.581

[51] Glashow S. L., Iliopoulos J., and Maiani L., "Weak Interactions with Lepton-Hadron Symmetry", Phys. Rev. D 2 (1970) 1285, https://doi.org/10.1103/PhysRevD.2.1285

[52] The UA1 Collaboration, "Experimental observation of isolated large transverse energy electrons with associated missing energy at $sqrts = 540$ GeV", Phys. Lett. B 122 (1983) 103, https://doi.org/10.1016/0370-2693(83)91177-2

[53] The UA1 Collaboration, "Experimental observation of lepton pairs of invariant mass around 95 GeV/c^2 at the CERN SPS collider", Phys. Lett. B 126 (1983) 398, https://doi.org/10.1016/0370-2693(83)90188-0

[54] The UA2 Collaboration, "Evidence for $Z^0 \to e^+e^-$ at the CERN pp collider", Phys. Lett. B 129 (1983) 130, https://doi.org/10.1016/0370-2693(83)90744-X

[55] Greenberg O. W., "Spin and Unitary Spin Independence in a Paraquark Model of Baryons and Mesons", Phys. Rev. Lett. 13 (1964) 598-602, https://doi.org/10.1103/PhysRevLett.13.598

[56] Han M. Y., Nambu Y., "Three-Triplet Model with Double SU(3) Symmetry". Phys. Rev. 139 ((1965) B1006-B1010, https://doi.org/10.1103/PhysRev.139.B1006

[57] M. Gell-Mann, The Eightfold Way: A Theory of strong interaction symmetry", CTSL-20, TID-12608, California Inst. of Tech., Pasadena. Synchrotron Lab. (1961), https://doi.org/10.2172/4008239

[58] Zweig G., "An SU_3 model for strong interaction symmetry and its breaking; Version 1", CERN-TH-401, CERN (1964), http://cds.cern.ch/record/352337

[59] M. Gell-Mann, "Quarks", Elementary Particle Physics: Multiparticle Aspects, ed. by P. Urban, Springer Vienna 9 (1972) 733, https://doi.org/10.1142/9789814618113_0002 , https://doi.org/10.1007/978-3-7091-4034-5_20

[60] Fritzsch H., Gell-Mann M., Leutwyler H., "Advantages of the color octet gluon picture", Physics Letters B 47 4 (1973) 365-368, https://doi.org/10.1016/0370-2693(73)90625-4

[61] Y. Nambu, "The Confinement of Quarks", Sci. Am. 235N5 (1976) 48, https://doi.org/10.1038/scientificamerican1176-48

[62] D. J. Gross and F. Wilczek, "Ultraviolet Behavior of Non-Abelian Gauge Theories", Phys. Rev. Lett. 30 (1973) 1343, https://doi.org/10.1103/PhysRevLett.30.1343

[63] H. D. Politzer, "Reliable Perturbative Results for Strong Interactions?", Phys. Rev. Lett. 30 (1973) 1346-1349, https://doi.org/10.1103/PhysRevLett.30.1346

[64] S. Bethke, "α_s 2002", Nuclear Physics B - Proceedings Supplements 121 (2003) 74, Proceedings of the QCD 02 9th High-Energy Physics International Conference on Quantum ChromoDynamics, https://doi.org/10.1016/S0920-5632(03)01817-6

[65] Zerwas P., "W and Z physics at LEP", Eur. Phys. J. C 34 (2004) 41-49, https://doi.org/10.1140/epjc/s2004-01765-9

[66] Salam G. P., "Elements of QCD for hadron colliders", arXiv:1011.5131 [hep-ph] (2010), https://doi.org/10.48550/arXiv.1011.5131

[67] Gell-Mann M., "Symmetries of Baryons and Mesons", Phys. Rev. 125 (1962) 1067-1084, https://doi.org/10.1103/PhysRev.125.1067

[68] Salam A. an Taylor J., "Unification of Fundamental Forces: The First 1988 Dirac Memorial Lecture", Cambridge University Press (1990), https://doi.org/10.1017/CBO9780511622854

[69] C. Englert et al., "Precision measurements of Higgs couplings: implications for new physics scales", J. Phys. G 41 (2014) 113001, 10.1088/0954-3899/41/11/113001

[70] J. N. Ng and P. Zakarauskas, "QCD-parton calculation of conjoined production of Higgs bosons and heavy flavors in $p\bar{p}$ collision", Phys. Rev. D 29 (1984) 876, https://doi.org/10.1103/PhysRevD.29.876

[71] Z. Kunszt, "Associated production of heavy Higgs boson with top quarks", Nucl. Phys. B 247 (1984) 339, https://doi.org/10.1016/0550-3213(84)90553-4

[72] S. Dawson, L. H. Orr, L. Reina, and D. Wackeroth, "Next-to-leading order QCD corrections to $pp\ t\bar{t}h$ at the CERN Large Hadron Collider", Phys. Rev. D 67 (2003) 071503, https://doi.org/10.1103/PhysRevD.67.071503

[73] W. Beenakker et al., "Higgs Radiation Off Top Quarks at the Tevatron and the LHC", Phys. Rev. Lett. 87 (2001) 201805, https://doi.org/10.1103/PhysRevLett.87.201805

[74] ALEPH Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaboration, The LEP Working Group for Higgs Boson Searches, "Search for the Standard Model Higgs boson at LEP", Physics Letters B 565 (2003) 61 https://doi.org/10.1016/S0370-2693(03)00614-2

[75] The TEVNPH Working Group for the CDF and D0 Collaborations, "Combined CDF and D0 Search for Standard Model Higgs Boson Production with up to 10.0 fb^{-1} of Data, arXiv: 1203.3774 [hep-ex] (2012), https://doi.org/10.48550/arXiv.1203.3774

[76] F. Demartin, F. Maltoni, K. Mawatari, M. Zaro, "Higgs production in association with a single top quark at the LHC", MCnet-15-07, CP3-15-08, arXiv:1504.00611 [hep-ph], https://doi.org/10.48550/arXiv.1504.00611

[77] The ATLAS Collaboration, "Cross-section measurements of the Higgs boson decaying into a pair of τ-leptons in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", Phys. Rev. D 99 (2019) 072001, https://doi.org/10.1103/PhysRevD.99.072001

[78] The CMS Collaboration, "Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector", Phys. Lett. B 779 (2018) 283, https://doi.org/10.1016/j.physletb.2018.02.004

[79] The ATLAS Collaboration, "Observation and measurement of Higgs boson decays to WW^* with the ATLAS detector", Phys. Rev. D 92 (2015) 012006, https://doi.org/10.1103/PhysRevD.92.012006

[80] The ATLAS Collaboration, "Evidence for the spin-0 nature of the Higgs boson using ATLAS data", Phys. Lett. B 726 (2013) 120-144, https://doi.org/10.1016/j.physletb.2013.08.026

[81] The CMS Collaboration, "Study of the Mass and Spin-Parity of the Higgs Boson Candidate Via Its Decays to Z Boson Pairs", Phys. Rev. Lett. 110 (2013) 081803, https://doi.org/10.1103/PhysRevLett.110.081803, Erratum Phys. Rev. Lett. 110 (2013) 189901, https://doi.org/10.1103/PhysRevLett.110.189901

[82] LHCTopWG, "History of LHCTopWG Summary Plots: Top-pair production cross-section as a function of centre-of-mass energy", https://twiki.cern.ch/twiki/bin/view/LHCPhysics/TopPairCrossectionSqrtsHistory

[83] J. Ellis, D. S. Hwang, K. Sakurai and M. Takeuchi, "Disentangling Higgs-top couplings in associated production", JHEP 04 (2014) 004, https://doi.org/10.1007/JHEP04(2014)004

[84] F. Boudjema, D. Guadagnoli, R. M. Godbole and K. A. Mohan, "Laboratory-frame observables for probing the top-Higgs boson interaction", Phys. Rev. D 92 (2015) 015019, https://doi.org/10.1103/PhysRevD.92.015019

[85] M. R. Buckley and D. Gonçalves, "Boosting the Direct CP Measurement of the Higgs-Top Coupling", Phys. Rev. Lett. 116 (2016) 091801, https://doi.org/10.1103/PhysRevLett.116.091801

[86] S. Amor Dos Santos et al., "Probing the CP nature of the Higgs coupling in $t\bar{t}h$ events at the LHC", Phys. Rev. D 96 (2017) 013004, https://doi.org/10.1103/PhysRevD.96.013004

[87] The CMS Collaboration, "Search for the associated production of the Higgs boson with a top-quark pair", JHEP 09 (2014) 087 https://doi.org/10.1007/JHEP09(2014)087

[88] The ATLAS Collaboration, "Search for the associated production of the Higgs boson with a top quark pair in multilepton final states with the ATLAS detector", Phys. Lett. B 749 (2015) 519, https://doi.org/10.1016/j.physletb.2015.07.079

[89] The ATLAS Collaboration, "Search for $H \to \gamma\gamma$ produced in association with top quarks and constraints on the Yukawa coupling between the top quark and the Higgs boson using data taken at 7 TeV and 8 TeV with the ATLAS detector", Phys. Lett. B 740 (2015) 222, https://doi.org/10.1016/j.physletb.2014.11.049

[90] The ATLAS Collaboration, "Evidence for the associated production of the Higgs boson and a top quark pair with the ATLAS detector", Phys. Rev. D 97 (2018) 072003, https://doi.org/10.1103/PhysRevD.97.072003

[91] The ATLAS Collaboration, "Measurement of the Higgs boson coupling properties in the $H \to ZZ^* \to 4l$ decay channel at $\sqrt{s} = 13$ TeV with the ATLAS detector", JHEP 03 (2018) 095, https://doi.org/10.1007/JHEP03(2018)095

[92] The ATLAS Collaboration, "Analysis of $t\bar{t}H$ and $t\bar{t}W$ production in multilepton final states with the ATLAS detector", ATLAS-CONF-2019-045, CERN (2019), http://cds.cern.ch/record/2693930

[93] The ATLAS Collaboration, "Measurements of Higgs boson properties in the diphoton decay channel with 36 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector", Phys. Rev. D 98 (2018) 052005, https://doi.org/10.1103/PhysRevD.98.052005

[94] The ATLAS Collaboration, "Measurement of Higgs boson production in association with a $t\bar{t}$ pair in the diphoton decay channel using 139 fb^{-1} of LHC data collected at $\sqrt{s} = 13$ TeV by the ATLAS experiment", ATLAS-CONF-2019-004, CERN (2019), https://cds.cern.ch/record/2668103

[95] The 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. C 75 (2015) 349, https://doi.org/10.1140/epjc/s10052-015-3543-1

[96] The ATLAS Collaboration, "Search for the Standard Model Higgs boson decaying into $b\bar{b}$ produced in association with top quarks decaying hadronically in *pp* collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector", JHEP 05 (2016) 160, https://doi.org/10.1007/JHEP05(2016)160

[97] The ATLAS Collaboration, "Search for the Standard Model Higgs boson produced in association with top quarks and decaying into $b\bar{b}$ in *pp* collisions at $\sqrt{s} = 13$ TeV with

the ATLAS detector", Phys. Rev. D 97 (2018) 072016, https://doi.org/10.1103/PhysRevD.97.072016

[98] The CMS Collaboration, "Search for $t\bar{t}H$ production in the $H \to b\bar{b}$ decay channel with leptonic $t\bar{t}$ decays in proton-proton collisions at $\sqrt{s} = 13$ TeV", JHEP 03 (2019) 026, https://doi.org/10.1007/JHEP03(2019)026

[99] The CMS Collaboration, "Search for $t\bar{t}H$ production in the all-jet final state in proton-proton collisions at $\sqrt{s} = 13$ TeV", JHEP 06 (2018) 101, https://doi.org/10.1007/JHEP06(2018)101

[100] The ATLAS Collaboration, "Measurement of Higgs boson decay into *b*-quarks in associated production with a top-quark pair in *pp* collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", JHEP 06 (2022) 97 https://doi.org/10.1007/JHEP06(2022)097

[101] R.L. Workman et al. (Particle Data Group), "Review of Particle Physics", Prog. Theor. Exp. Phys. 2022 (2022) 083C01, https://doi.org/10.1093/ptep/ptac097

[102] The ATLAS Collaboration, "Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector", Phys. Lett. B 784 (2018) 173, https://doi.org/10.1016/j.physletb.2018.07.035

[103] The CMS Collaboration, "Observation of $t\bar{t}H$ production", Phys. Rev. Lett. 120 (2018) 231801, https://doi.org/10.1103/PhysRevLett.120.231801

[104] The ATLAS Collaboration, "Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector", Phys. Lett. B 786 (2018) 59, https://doi.org/10.1016/j.physletb.2018.09.013

[105] The CMS Collaboration, "Observation of Higgs boson decay to bottom quarks", Phys. Rev. Lett. 121, 121801 (2018), https://doi.org/10.1103/PhysRevLett.121.121801

[106] Florian D. et al., "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector", CERN-2017-002-M, CERN (2017), https://doi.org/10.23731/CYRM-2017-002

[107] Moretti N., Petrov P., Pozzorini S., and Spannowsky M., "Measuring the signal strength in $t\bar{t}H$ with $H \to b\bar{b}$", Phys. Rev. D 93 (2016) 014019, https://doi.org/10.1103/PhysRevD.93.014019

[108] Bryant P. and Evans L., "LHC Machine", Journal of Instrumentation 3 (2008) S08001, https://dx.doi.org/10.1088/1748-0221/3/08/S08001

[109] "About CERN", accessed: 1 October 2022, https://home.cern/about

[110] Mobs E., "The CERN accelerator complex. Complexe des accélérateures du CERN", (2016), General Photo (accessed 1 October 2022), https://cds.cern.ch/record/2197559

[111] The ATLAS Collaboration, "Luminosity determination in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC", tech. rep. ATLAS-CONF-2019-021, CERN (2019), http://cds.cern.ch/record/2677054

[112] The ATLAS Collaboration, "Improved luminosity determination in pp collisions at $\sqrt{s} =$ 7 TeV using the ATLAS detector at the LHC", Eur. Phys. J. C 73, 2518 (2013), https://doi.org/10.1140/epjc/s10052-013-2518-3

[113] The ATLAS Collaboration, "Luminosity determination in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector at the LHC", Eur. Phys. J. C 76, 653 (2016), https://doi.org/10.1140/epjc/s10052-016-4466-1

[114] Marshall Z. and the ATLAS Collaboration, "Simulation of Pile-up in the ATLAS Experiment", Journal of Physics: Conference Series 513 (2014) 022024 IOP Publishing, doi:10.1088/1742-6596/513/2/022024

[115] Stirling W. J., "Parton Luminosity and Cross-section plots", Imperial College London (accessed 1 October 2022), http://www.hep.ph.ic.ac.uk/~wstirlin/plots/plots.html

[116] The ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider", Journal of Instrumentation 3 (2008) S08003, https://dx.doi.org/10.1088/1748-0221/3/08/S08003

[117] The CMS Collaboration, "The CMS experiment at the CERN LHC", Journal of Instrumentation 3 (2008) S08004, https://dx.doi.org/10.1088/1748-0221/3/08/S08004

[118] The LHCb Collaboration, "The LHCb Detector at the LHC", Journal of Instrumentation 3 (2008) S08005, https://dx.doi.org/10.1088/1748-0221/3/08/S08005

[119] The ALICE Collaboration, "The ALICE experiment at the CERN LHC", Journal of Instrumentation 3 (2008) S08002, https://dx.doi.org/10.1088/1748-0221/3/08/S08002

[120] The ATLAS Collaboration, "Public ATLAS Luminosity Results for Run-2 of the LHC", accessed: 1 October 2022, https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2

[121] B. Abbott et al., "Production and integration of the ATLAS Insertable B-Layer", JINST 13 (2018) T05008, https://doi.org/10.1088/1748-0221/13/05/T05008

[122] The ATLAS Collaboration, "Performance of the ATLAS trigger system in 2015", Eur. Phys. J. C 77 (2017) 317, https://doi.org/10.1140/epjc/s10052-017-4852-3

[123] Maltoni F., Ridolfi G., and Ubiali M., "b-initiated processes at the LHC: a reappraisal", JHEP 07 (2012) 22, https://doi.org/10.1007/JHEP07(2012)022

[124] G. 't Hooft and M. Veltman, "Regularization and renormalization of gauge fields", Nuclear Physics B 44 (1972) 189-213, https://doi.org/10.1016/0550-3213(72)90279-9

[125] Collins J. C., Soper D. E., and Sterman G. F., "Factorization of Hard Processes in QCD", Adv. Ser. Direct. High Energy Phys. 5 (1989) 1-91, https://doi.org/10.1142/9789814503266_0001

[126] Collins J. C. and Soper D. E., "The Theorems of Perturbative QCD", Ann. Rev. of Nucl. and Part. Sc. 37 (1987) 383-409 https://www.annualreviews.org/doi/10.1146/annurev.ns.37.120187.002123

[127] Hoche S., "Introduction to parton-shower event generators", SLAC-PUB 16160, arXiv:1411.4085 [hep-ph] (2014). https://doi.org/10.1142/9789814678766_0005

[128] Placakyte R., "Parton Distribution Functions", Proceedings of 31^{st} International Conference on Physics in Collisions (PIC 2011) Canada, https://doi.org/10.48550/arXiv.1111.5452

[129] Wing M. (on behalf of the ZEUS and H1 Collaborations), "Measurements of deep inelastic scattering at HERA", Proceedings of 32^{nd} International Conference on Physics in Collisions (PIC 2012) Slovakia pp. 93–106 https://doi.org/10.48550/arXiv.1301.7572

[130] Dokshitzer Y. L., "Calculation of the Structure Functions for Deep Inelastic Scattering and e^+e^- Annihilation by Perturbation Theory in Quantum Chromodynamics", Sov. Phys. JETP 46 (1977) 641

[131] Gribov V. and Lipatov L., "e^+e^- annihilation and deep inelastic ep-scattering in perturbation theory", Sov. J. Nucl. Phys. 15 (1972) 675-684, Yad. Fiz 15 (1972) 1218-1237

[132] Altarelli G. and Parisi G., "Asymptotic freedom in parton language", Nuclear Physics B 126 (1977) 298-318 , https://doi.org/10.1016/0550-3213(77)90384-4

[133] Altarelli G., "QCD evolution equations for parton densities", Scholarpedia 4 (1) 7124 (2009) https://doi.org/10.4249/scholarpedia.7124

[134] Lai H.-L., Guzzi M. et al., "New parton distributions for collider physics", Phys. Rev. D82 (2010) 074024, https://doi.org/10.1103/PhysRevD.82.074024

[135] Ball R.D., Bertone V., Carrazza S. et al, "Parton distributions from high-precision collider data", Eur. Phys. J. C 77 (2017) 663, https://doi.org/10.1140/epjc/s10052-017-5199-5

[136] Sjostrand T., "A Model for Initial State Parton Showers", Phys. Lett. B 157 (1985) 321 https://doi.org/10.1016/0370-2693(85)90674-4

[137] S. Hö̈che, F. Krauss, M. Schönherr and F. Siegert, "A critical appraisal of NLO+PS matching methods", JHEP 09 (2012) 049, https://doi.org/10.1007/JHEP09(2012)049

[138] Andersson B., Gustafson G., Ingelman G., and Sjostrand T., "Parton Fragmentation and String Dynamics", Phys. Rept. 97 (1983) 31 https://doi.org/10.1016/0370-1573(83)90080-7

[139] Sjostrand T., "Jet Fragmentation of Nearby Partons", Nucl. Phys. B 248 (1984) 469-502, https://doi.org/10.1016/0550-3213(84)90607-2

[140] Webber B. R., "A QCD Model for Jet Fragmentation Including Soft Gluon Interference", Nucl. Phys. B 238 (1984) 492–528, https://doi.org/10.1016/0550-3213(84)90333-X

[141] Winter J.-C., Krauss F., and Soff G. "A modified cluster-hadronisation model", Eur. Phys. J. C 36, 381–395 (2004), https://doi.org/10.1140/epjc/s2004-01960-8

[142] Amati D. and Veneziano G., "Preconfinement as a Property of Perturbative QCD", Phys. Lett. B 83 (1979) 87, https://doi.org/10.1016/0370-2693(79)90896-7

[143] T. Sj*ö*strand and M. van Zijl, "A multiple-interaction model for the event structure in hadron collisions", Phys. Rev. D 36 (1987) 2019, https://doi.org/10.1103/PhysRevD.36.2019

[144] The ATLAS Collaboration, "Measurement of the underlying event in jet events from 7 TeV proton–proton collisions with the ATLAS detector", Eur. Phys. J. C 74 (2014) 2965, https://doi.org/10.1140/epjc/s10052-014-2965-5

[145] Lange D. J., "The EvtGen particle decay simulation package", Nucl. Instrum. Meth. A 462 (2001) 152-155, https://doi.org/10.1016/S0168-9002(01)00089-4

[146] The ATLAS Collaboration, "The ATLAS Simulation Infrastructure", Eur. Phys. J. C 70 (2010) 823, https://doi.org/10.1140/epjc/s10052-010-1429-9

[147] The GEANT4 Collaboration, S. Agostinelli et al., Geant4 – a simulation toolkit", Nucl. Instrum. Meth. A 506 (2003) 250, https://doi.org/10.1016/S0168-9002(03)01368-8

[148] The ATLAS Collaboration, "Fast Simulation for ATLAS: Atlfast-II and ISF", J. Phys. Conf. Ser. 396 (2012) 022031, https://doi.org/10.1088/1742-6596/396/2/022031

[149] Schaarschmidt J. (on behalf of the ATLAS Collaboration), "The new ATLAS Fast Calorimeter Simulation", J. Phys. Conf. Ser. 898 (2017)042006, https://doi.org/10.1088/1742-6596/898/4/042006

[150] S. Frixione, P. Nason, and C. Oleari, "Matching NLO QCD computations with Parton Shower simulations: the POWHEG method", JHEP 11 (2007) 070, https://doi.org/10.1088/1126-6708/2007/11/070

[151] P. Nason, "A New method for combining NLO QCD with shower Monte Carlo algorithms", JHEP 11 (2004) 040, https://doi.org/10.1088/1126-6708/2004/11/040

[152] Alioli S., Nason P., Oleari C. et al, "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX", J. High Energ. Phys. 43 (2010) 2010, https://doi.org/10.1007/JHEP06(2010)043

[153] B. Hartanto, B. J*ä*ger, L. Reina, and D. Wackeroth, "Higgs boson production in association with top quarks in the POWHEG BOX", Phys. Rev. D 91 (2015) 094003, https://doi.org/10.1103/PhysRevD.91.094003

[154] T. Ježo, J. M. Lindert, N. Moretti and S. Pozzorini, "New NLOPS predictions for $t\bar{t}+b$-jet production at the LHC", Eur. Phys. J. C 78 (2018) 502, https://doi.org/10.1140/epjc/s10052-018-5956-0

[155] S. Frixione and B. R. Webber, "Matching NLO QCD computations and parton shower simulations", JHEP 06 (2002) 029, https://doi.org/10.1088/1126-6708/2002/06/029

[156] Alwall J., Demin P., "MadGraph/MadEvent v4: The New Web Generation", In: JHEP 09 (2007) 028, https://doi.org/10.1088/1126-6708/2007/09/028

[157] R. Frederix and S. Frixione, Merging meets matching in MC@NLO, JHEP 12 (2012) 061, https://doi.org/10.1007/JHEP12(2012)061

[158] Alwall J., Frederix R., Frixione S. et al., "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations", J. High Energ. Phys. 7 (2014) 79 , https://doi.org/10.1007/JHEP07(2014)079

[159] T. Sjöstrand et al., "An introduction to PYTHIA 8.2", Comput. Phys. Commun. 191 (2015) 159, https://doi.org/10.1016/j.cpc.2015.01.024

[160] M. B$\ddot{a}$hr et al., "Herwig++ physics and manual", Eur. Phys. J. C 58 (2008) 639, https://doi.org/10.1140/epjc/s10052-008-0798-9

[161] J. Bellm et al., "Herwig 7.0/Herwig++ 3.0 release note", Eur. Phys. J. C 76 (2016) 196, https://doi.org/10.1140/epjc/s10052-016-4018-8

[162] E. Bothmann et al., "Event Generation with Sherpa 2.2", SciPost Phys. 7 (2019) 034, https://doi.org/10.21468/SciPostPhys.7.3.034

[163] S. Schumann and F. Krauss, "A Parton shower algorithm based on Catani-Seymour dipole factorisation", JHEP 03 (2008) 038 https://doi.org/10.1088/1126-6708/2008/03/038

[164] F. Cascioli, P. Maierhofer, and S. Pozzorini, "Scattering Amplitudes with Open Loops", Phys. Rev. Lett. 108 (2012) 111601 https://doi.org/10.1103/PhysRevLett.108.111601

[165] S. Catani, F. Krauss, R. Kuhn, and B. Webber, "QCD matrix elements + parton showers", JHEP 0111 (2001) 063, https://doi.org/10.1088/1126-6708/2001/11/063

[166] S. Catani, Y. L. Dokshitzer, M. Olsson, G. Turnock, and B. Webber, "New clustering algorithm for multijet cross-sections in e^+e^- annihilation, Phys", Lett. B 269 (1991) 432, https://doi.org/10.1016/0370-2693(91)90196-W

[167] The ATLAS Collaboration, "The Pythia 8 A3 tune description of ATLAS minimum bias and inelastic measurements incorporating the Donnachie-Landshoff diffractive model", ATL-PHYS-PUB-2016-017, CERN (2016) https://cds.cern.ch/record/2206965

[168] The ATLAS Collaboration, "ATLAS Pythia 8 tunes to 7 TeV data", ATL-PHYS-PUB-2014-021, CERN (2014), https://cds.cern.ch/record/1966419

[169] R. D. Ball et al., "Parton distributions for the LHC run II", J. High Energ. Phys. 04 (2015) 040, https://doi.org/10.1007/JHEP04(2015)040

[170] L. Harland-Lang, A. Martin, P. Motylinski and R. Thorne, "Parton distributions in the LHC era: MMHT 2014 PDFs, Eur. Phys. J. C 75 (2015) 204, https://doi.org/10.1140/epjc/s10052-015-3397-6

[171] The ATLAS Collaboration, "Studies on top-quark Monte Carlo modelling with Sherpa and MG5_aMCNLO", ATL-PHYS-PUB-2017-007, CERN (2017), https://cds.cern.ch/record/2261938

[172] S. Frixione, E. Laenen, P. Motylinski, and B. R. Webber, "Angular correlations of lepton pairs from vector boson and top quark decays in Monte Carlo simulations", JHEP 04 (2007) 081, https://doi.org/10.1088/1126-6708/2007/04/081

[173] P. Artoisenet, R. Frederix, O. Mattelaer, and R. Rietkerk, "Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations", 2008 JHEP 03 (2013) 015, https://doi.org/10.1007/JHEP03(2013)015

[174] L. H. C. S. W. Group, "SM Higgs Branching Ratios and Partial-Decay Widths", 2016, https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CERNYellowReportPageBR

[175] The ATLAS Collaboration, "Studies on top-quark Monte Carlo modelling for Top2016", ATL-PHYS-PUB-2016-020, CERN (2016), https://cds.cern.ch/record/2216168

[176] R. Frederix, E. Re, and P. Torrielli, "Single-top t-channel hadroproduction in the four-flavour scheme with POWHEG and aMCNLO", JHEP 09 (2012) 130, https://doi.org/10.1007/JHEP09(2012)130

[177] N. Kidonakis, "Two-loop soft anomalous dimensions for single top quark associated production with a W^- or H^-", Phys. Rev. D 82 (2010) 054018, https://doi.org/10.1103/PhysRevD.82.054018

[178] The ATLAS Collaboration, "Studies on top-quark Monte Carlo modelling for Top2016", ATL-PHYS-PUB-2016-020, CERN (2016), https://cds.cern.ch/record/2216168

[179] S. Höche, F. Krauss, M. Schönherr and F. Siegert, QCD matrix elements + parton showers. The NLO case", JHEP 04 (2013) 027, https://doi.org/10.1007/JHEP04(2013)027

[180] S. Catani, F. Krauss, B. R. Webber and R. Kuhn, "QCD Matrix Elements + Parton Showers", JHEP 11 (2001) 063, https://doi.org/10.1088/1126-6708/2001/11/063

[181] R. Frederix, D. Pagani, and M. Zaro, "Large NLO corrections in $t\bar{t}W^{\pm}$ and $t\bar{t}t\bar{t}$ hadroproduction from supposedly subleading EW contributions", JHEP 02 (2018) 031, https://doi.org/10.1007/JHEP02(2018)031

[182] The ATLAS Collaboration, "Measurement of the production cross-section of a single top quark in association with a Z boson in proton–proton collisions at 13 TeV with the ATLAS detector", Phys. Lett. B 780 (2018) 557 https://doi.org/10.1016/j.physletb.2018.03.023

[183] J. Pumplin et al., "New Generation of Parton Distributions with Uncertainties from Global QCD Analysis", JHEP 07 (2002) 012, https://doi.org/10.1088/1126-6708/2002/07/012

[184] F. Demartin, B. Maier, F. Maltoni, K. Mawatari and M. Zaro, "tWH associated production at the LHC", Eur. Phys. J. C 77 (2017) 34, https://doi.org/10.1140/epjc/s10052-017-4601-7

[185] T. Gleisberg and S. Höche, "Comix, a new matrix element generator", JHEP 12 (2008) 039, https://doi.org/10.1088/1126-6708/2008/12/039

[186] C. Anastasiou, L. J. Dixon, K. Melnikov, and F. Petriello, "High precision QCD at hadron colliders: Electroweak gauge boson rapidity distributions at next-to-next-to leading order", Phys. Rev. D 69 (2004) 094008, https://doi.org/10.1103/PhysRevD.69.094008

[187] A. Denner, S. Dittmaier, M. Roth and L. H. Wieders, "Electroweak corrections to charged-current $e^+e^- \to 4$fermion processes: Technical details and further results", Nucl. Phys. B 724 (2005) 247, https://doi.org/10.1016/j.nuclphysb.2005.06.033, Erratum: Nucl. Phys. B 854 (2012) 504 https://doi.org/10.1016/j.nuclphysb.2011.09.001

[188] S. Dittmaier and M. Huber, "Radiative corrections to the neutral-current Drell-Yan process in the Standard Model and its minimal supersymmetric extension", JHEP 01 (2010) 060, https://doi.org/10.1007/JHEP01(2010)060

[189] J. R. Andersen et al., "Les Houches 2013: Physics at TeV Colliders: Standard Model Working Group Report", arXiv: 1405.1067 [hep-ph] (2014), https://doi.org/10.48550/arXiv.1405.1067

[190] Pequenao J. & Schaffner P., "How ATLAS detects particles: diagram of particle paths in the detector", CERN-EX-1301009, CERN (2013), https://cds.cern.ch/record/1505342

[191] Cornelissen T., Elsing M., Liebig W., Fleischmann S., & Moyse E., "Concepts, design and implementation of the ATLAS new tracking (NEWT)", ATL-SOFT-PUB-2007-007, CERN (2007), https://cds.cern.ch/record/1020106

[192] The ATLAS Collaboration, "The Optimization of ATLAS Track Reconstruction in Dense Environments", ATL-PHYS-PUB-2015-006, CERN (2015), https://cds.cern.ch/record/2002609

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

[194] The ATLAS Collaboration, "Expected Performance of the ATLAS Experiment - Detector, Trigger and Physics", CERN (2008), http://cds.cern.ch/record/1125884

[195] The DØ Collaboration, "Observation of Single Top Quark Production" (2009), https://www-d0.fnal.gov/Run2Physics/top/singletop_observation/singletop_observation_updated.html

[196] The ATLAS Collaboration, "Reconstruction of primary vertices at the ATLAS experiment in Run 1 proton–proton collisions at the LHC", Eur. Phys. J. C 77, 332 (2017), https://doi.org/10.1140/epjc/s10052-017-4887-5

[197] Lampl W., Laplace S., Lelas D., Loch P., Ma H., Menke S., Rajagopalan S., Rousseau D., Snyder S., & Unal G., "Calorimeter Clustering Algorithms: Description and Performance", CERN (2008), https://cds.cern.ch/record/1099735

[198] The ATLAS Collaboration, "Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1", Eur. Phys. J. C 77, 490 (2017), https://doi.org/10.1140/epjc/s10052-017-5004-5

[199] The ATLAS Collaboration, "Electron efficiency measurements with the ATLAS detector using 2012 LHC proton–proton collision data", Eur. Phys. J. C 77, 195 (2017), https://doi.org/10.1140/epjc/s10052-017-4756-2

[200] The ATLAS Collaboration, "Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton–proton collision data at $\sqrt{s} = 13$ TeV", Eur. Phys. J. C 79, 639 (2019), https://doi.org/10.1140/epjc/s10052-019-7140-6

[201] The ATLAS Collaboration, "Measurement of the photon identification efficiencies with the ATLAS detector using LHC Run-1 data", Eur. Phys. J. C 76, 666 (2016), https://doi.org/10.1140/epjc/s10052-016-4507-9

[202] The Sheffield ATLAS group, "ATLAS Electron Reconstruction Activities", https://www.hep.shef.ac.uk/research/atlas/egamma/

[203] The ATLAS Collaboration, "Electron and photon performance measurements with the ATLAS detector using 2015-2017 LHC proton–proton collision data", JINST 14 (2019) P12006, https://doi.org/10.1088/1748-0221/14/12/P12006

[204] The 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, https://doi.org/10.1140/epjc/s10052-016-4120-y

[205] The ATLAS Collaboration, "Reconstruction, Energy Calibration, and Identification of Hadronically Decaying Tau Leptons in the ATLAS Experiment for Run-2 of the LHC", ATL-PHYS-PUB-2015-045, CERN (2015), https://cds.cern.ch/record/2064383

[206] The ATLAS Collaboration, "Performance of missing transverse momentum reconstruction with the ATLAS detector using proton–proton collisions at $\sqrt{s} = 13$ TeV", Eur. Phys. J. C 78, 903 (2018), https://doi.org/10.1140/epjc/s10052-018-6288-9

[207] The ATLAS Collaboration, "Performance of algorithms that reconstruct missing transverse momentum in $\sqrt{s} = 8$ TeV proton–proton collisions in the ATLAS detector", Eur. Phys. J. C 77, 241 (2017), https://doi.org/10.1140/epjc/s10052-017-4780-2

[208] Salam G.P., "Towards jetography", Eur. Phys. J. C 67, 637–686 (2010), https://doi.org/10.1140/epjc/s10052-010-1314-6

[209] Catani S., Dokshitzer Yu.L., Seymour M.H. and Webber B.R., "Longitudinally-invariant $k_\perp$-clustering algorithms for hadron-hadron collisions", Nuclear Physics B 406, 187—224 (1993), https://doi.org/10.1016/0550-3213(93)90166-M

[210] Dokshitzer Yu.L., Leder G.D., Moretti S. and Webber B.R., "Better jet clustering algorithms", JHEP 08 (1997) 001, https://doi.org/10.1088/1126-6708/1997/08/001

[211] The ATLAS Collaboration, "Performance of pile-up mitigation techniques for jets in *pp* collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector", Eur. Phys. J. C 76, 581 (2016), https://doi.org/10.1140/epjc/s10052-016-4395-z

[212] Cacciari M., Salam G.P., and Soyez G., "The anti-k_t jet clustering algorithm", JHEP 04 (2008) 063, https://doi.org/10.1088/1126-6708/2008/04/063

[213] Cacciari M., Salam G. P, and Soyez G., "The catchment area of jets", JHEP 04 (2008) 005, https://doi.org/10.1088/1126-6708/2008/04/005

[214] The ATLAS Collaboration, "Jet energy scale and resolution measured in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", CERN-EP-2020-083, CERN (2020), https://cds.cern.ch/record/2722869

[215] Nachman B., Nef P., Schwartzman A., Swiatlowski M., and Wanotayaroj C., "Jets from jets: re-clustering as a tool for large radius jet reconstruction and grooming at the LHC", JHEP 01 (2015) 75, https://doi.org/10.1007/JHEP02(2015)075

[216] Krohn D., Thaler J., and Wang L. T., "Jet trimming." JHEP 02 (2010) 84, https://doi.org/10.1007/JHEP02(2010)084

[217] The ATLAS Collaboration, "ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13$ TeV", Eur. Phys. J. C 79, 970 (2019), https://doi.org/10.1140/epjc/s10052-019-7450-8

[218] Frühwirth R., "Application of Kalman filtering to track and vertex fitting", Nuclear Instruments and Methods in Physics Research A 262 (1987) 444-450, https://doi.org/10.1016/0168-9002(87)90887-4

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

[220] The ATLAS Collaboration, "Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017-18 LHC run", ATL-PHYS-PUB-2017-013, CERN (2017), https://cds.cern.ch/record/2273281

[221] The ATLAS Collaboration, "Measurements of b-jet tagging efficiency with the ATLAS detector using $t\bar{t}$ events at $\sqrt{s} = 13$ TeV", JHEP 08 (2018) 89, https://doi.org/10.1007/JHEP08(2018)089

[222] The ATLAS Collaboration, "Measurement of b-tagging efficiency of c-jets in $t\bar{t}$ events using a likelihood approach with the ATLAS detector", ATLAS-CONF-2018-001, CERN (2018), https://cds.cern.ch/record/2306649

[223] Erdmann J. et al., "A likelihood-based reconstruction algorithm for top-quark pairs and the KLFitter framework", Nuclear Instruments and Methods in Physics Research A 748 (2014) 18-25, https://doi.org/10.1016/j.nima.2014.02.029

[224] The ATLAS Collaboration, "Calibration of light-flavour b-jet mistagging rates using ATLAS proton-proton collision data at $\sqrt{s} = 13$ TeV", ATLAS-CONF-2018-006, CERN (2018), https://cds.cern.ch/record/2314418

[225] The ATLAS Collaboration, "Jet energy scale measurements and their systematic uncertainties in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", Phys. Rev. D 96, 072002 (2017), https://link.aps.org/doi/10.1103/PhysRevD.96.072002

[226] The ATLAS Collaboration, "Selection of jets produced in 13 TeV proton–proton collisions with the ATLAS detector", ATLAS-CONF-2015-029, CERN (2015), https://cds.cern.ch/record/2037702

[227] ATLAS JetEtmiss group, "Pileup jet recommendations" twiki page, (2021), https://twiki.cern.ch/twiki/bin/view/AtlasProtected/PileupJetRecommendations

[228] Cacciari M., Salam G.P., and Soyez G., "FastJet user manual", Eur. Phys. J. C 72, 1896 (2012), https://doi.org/10.1140/epjc/s10052-012-1896-2

[229] The ATLAS Collaboration, "Luminosity determination in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC", ATLAS-CONF-2019-021, CERN (2019), https://cds.cern.ch/record/2677054

[230] Avoni G. et al., "The new LUCID-2 detector for luminosity measurement and monitoring in ATLAS", JINST 13 (2018) P07017, https://doi.org/10.1088/1748-0221/13/07/p07017

[231] The ATLAS Collaboration, "Measurement of the Inelastic Proton-Proton Cross Section at $\sqrt{s} = 13$ TeV with the ATLAS Detector at the LHC", Phys. Rev. Lett. 117 (2016) 182002, https://link.aps.org/doi/10.1103/PhysRevLett.117.182002

[232] Berger N., Bertella C. et al., "Simplified Template Cross Sections - Stage 1.1", LHCHXSWG-2019-003, arXiv:1906.02754 [hep-ph] (2019), https://doi.org/10.48550/arXiv.1906.02754

[233] Badger S., Bendavid J. et al., "Les Houches 2015: Physics at TeV Colliders Standard Model Working Group Report", arXiv:1605.04692 [hep-ph] (2016), https://doi.org/10.48550/arXiv.1605.04692

[234] Boudjema F., Guadagnoli D., Godbole R. M., and Mohan K. A., "Laboratory-frame observables for probing the top-Higgs boson interaction", Phys. Rev. D 92 (2015) 015019, https://doi.org/10.1103/PhysRevD.92.015019

[235] Maltoni F., Pagani D., Shivaji A. et al, "Trilinear Higgs coupling determination via single-Higgs differential measurements at the LHC", Eur. Phys. J. C 77 887 (2017), https://doi.org/10.1140/epjc/s10052-017-5410-8

[236] The ATLAS Collaboration, "ATLAS data quality operations and performance for 2015–2018 data-taking", JINST 15 (2020) P04003, https://doi.org/10.1088/1748-0221/15/04/P04003

[237] The ATLAS Collaboration, "Luminosity determination in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC", ATLAS-CONF-2019-021, CERN (2019), https://cds.cern.ch/record/2677054

[238] The ATLAS Collaboration, "Performance of the ATLAS muon triggers in Run 2", JINST 15 (2020) P09015, https://doi.org/10.1088/1748-0221/15/09/p09015

[239] The ATLAS Collaboration, "Performance of electron and photon triggers in ATLAS during LHC Run 2", Eur. Phys. J. C 80 (2020) 47, https://doi.org/10.1140/epjc/s10052-019-7500-2

[240] Freund Y. and Schapire R. E., "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting", Journal of Computer and System Sciences 55 (1997) 1 pp. 119-139, https://doi.org/10.1006/jcss.1997.1504

[241] Breiman L., Friedman J., Olshen R., and Stone C., "Classification and Regression Trees", Wadsworth and Brooks, Monterey (1984), https://doi.org/10.1002/cyto.990080516

[242] Hoecker A. et al., "TMVA - Toolkit for Multivariate Data Analysis", 2007, https://doi.org/10.48550/arXiv.physics/0703039

[243] Antcheva I. et al., "ROOT — A C++ framework for petabyte data storage, statistical analysis and visualization", Computer Physics Communications 180 (2009) 12 pp. 2499-2512, https://doi.org/10.1016/j.cpc.2009.08.005

[244] IML WG, "A Living Review of Machine Learning for Particle Physics", accessed: 20 September 2022, https://imlwg.github.io/HEPML-LivingReview/

[245] McCulloch W.S. and Pitts W., "A logical calculus of the ideas immanent in nervous activity", Bulletin of Mathematical Biophysics 5 pp. 115–133 (1943), https://doi.org/10.1007/BF02478259

[246] Goodfellow I., Bengio Y., and Courville A., "Deep learning", MIT Press (2016), http://www.deeplearningbook.org

[247] LeCun Y., Bengio Y., and Hinton G., "Deep learning", Nature 521 pp.436–444 (2015), https://doi.org/10.1038/nature14539

[248] Melcher K., "A Friendly Introduction to [Deep] Neural Networks", KNIME Blog (2021) (accessed 20 May 2022), https://www.knime.com/blog/a-friendly-introduction-to-deep-neural-networks

[249] Rumelhart D., Hinton G., and Williams R., "Learning representations by back-propagating errors", Nature 323 pp. 533–536 (1986), https://doi.org/10.1038/323533a0

[250] Srivastava N., Hinton G., Krizhevsky A., Sutskever I., and Salakhutdinov R., "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", Journal of Machine Learning Research 15 (2014) 56 pp. 1929–1958, http://jmlr.org/papers/v15/srivastava14a.html

[251] ATLAS Collaboration, "Expected Performance of Boosted Higgs ($\rightarrow b\bar{b}$) Boson Identification with the ATLAS Detector at $\sqrt{s} = 13$ TeV", tech. rep. ATL-PHYS-PUB-2015-035, CERN (2015), https://cds.cern.ch/record/2042155

[252] ATLAS Collaboration, "Boosted hadronic top identification at ATLAS for early 13 TeV data", tech. rep. ATL-PHYS-PUB-2015-053, CERN (2015), https://cds.cern.ch/record/2116351

[253] Fenton M., "Boosting to the top: measurements of boosted top quarks and Higgs bosons with the ATLAS detector at the Large Hadron Collider", PhD Thesis (2019), University of Glasgow, http://theses.gla.ac.uk/id/eprint/75176

[254] Chollet F. et al., "Keras", https://github.com/keras-team/keras

[255] Abadi M. et al., "TensorFlow for Keras", https://github.com/tensorflow/tensorflow

[256] ATLAS Collaboration, "Performance of jet substructure techniques for large-*R* jets in proton–proton collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector", JHEP 09 (2013) 076, https://doi.org/10.1007/JHEP09(2013)076

[257] The ATLAS collaboration, "Study of top-quark pair modelling and uncertainties using ATLAS measurements at $\sqrt{s} = 13$ TeV", ATL-PHYS-PUB-2020-023, CERN (2020), https://cds.cern.ch/record/2730443

[258] Stewart I. W., Tackmann F. J., "Theory Uncertainties for Higgs and Other Searches Using Jet Bins", Phys. Rev. D 85 (2012) 034011, https://doi.org/10.1103/PhysRevD.85.034011

[259] Raitio R. and Wada W. W., "Higgs-boson production at large transverse momentum in quantum chromodynamics", Phys. Rev. D 19 (1979) 941, https://doi.org/10.1103/PhysRevD.19.941

[260] Dawson S., Jackson C., Orr L. H., Reina L., and Wackeroth D., "Associated Higgs production with top quarks at the Large Hadron Collider: NLO QCD corrections", Phys. Rev. D 68 (2003) 034022, https://doi.org/10.1103/PhysRevD.68.034022

[261] Frixione S., Hirschi V., Pagani D., Shao H.-S., and Zaro M., "Electroweak and QCD corrections to top-pair hadroproduction in association with heavy bosons", J. High Energ. Phys. 6 (2015) 184, https://doi.org/10.1007/JHEP06(2015)184

[262] Cacciari M., Czakon M., Mangano M., Mitov A., and Nason P., "Top-pair production at hadron colliders with next-to-next-to-leading logarithmic soft-gluon resummation", Phys. Lett. B 710 (2012) 612, https://doi.org/10.1016/j.physletb.2012.03.013

[263] Czakon M. and Mitov A., "NNLO corrections to top-pair production at hadron colliders: the all-fermionic scattering channels", JHEP 12 (2012) 054, https://doi.org/10.1007/JHEP12(2012)054

[264] Czakon M. and Mitov A., "NNLO corrections to top pair production at hadron colliders: the quark-gluon reaction", JHEP 01 (2013) 080, https://doi.org/10.1007/JHEP01(2013)080

[265] Czakon M., Fiedler P., and Mitov A., "Total Top-Quark Pair-Production Cross Section at Hadron Colliders Through $\mathcal{O}(\alpha_s^4)$, Phys. Rev. Lett. 110 (2013) 252004, https://doi.org/10.1103/PhysRevLett.110.252004

[266] Kant P. et al., "HatHor for single top-quark production: Updated predictions and uncertainty estimates for single top-quark production in hadronic collisions", Comput. Phys. Commun. 191 (2015) 74, https://doi.org/10.1016/j.cpc.2015.02.001

[267] Martin A. D., Stirling W. J., Thorne R. S., and Watt G., "Parton distributions for the LHC", Eur. Phys. J. C 63 (2009) 189, https://doi.org/10.1140/epjc/s10052-009-1072-5

[268] Martin A. D., Stirling W. J., Thorne R. S., and Watt G., "Uncertainties on α_s in global PDF analyses and implications for predicted hadronic cross sections", Eur. Phys. J. C 64 (2009) 653, https://doi.org/10.1140/epjc/s10052-009-1164-2

[269] S. Frixione, E. Laenen, P. Motylinski, C. White, and B. R. Webber, "Single-top hadroproduction in association with a W boson", JHEP 07 (2008) 029, https://doi.org/10.1088/1126-6708/2008/07/029

[270] J. M. Campbell and R. K. Ellis, "$t\bar{t}W^{\pm}$ production and decay at NLO", JHEP 07 (2012) 052, https://doi.org/10.1007/JHEP07(2012)052

[271] 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, https://doi.org/10.1007/JHEP07(2014)079

[272] The ATLAS Collaboration, Multi-boson simulation for 13 TeV ATLAS analyses, ATL-PHYS-PUB-2016-002, CERN (2016), https://cds.cern.ch/record/2119986

[273] Nechansky F., "Search for the production of a Higgs boson decaying into a pair of bottom quarks in association with a pair of top quarks at 13 TeV with the ATLAS detector", 2021, https://cds.cern.ch/record/2783832

[274] Data Analysis Techniques for High Energy Particle Physics, "Proceedings of the 1974 CERN School of Computing: God∅ysund, Norway 11 - 24 Aug 1974. 3rd CERN School of Computing", CERN (1974), https://cds.cern.ch/record/186223

[275] Cowan G., Cranmer K., Gross E. et al, "Asymptotic formulae for likelihood-based tests of new physics", Eur. Phys. J. C 71, 1554 (2011), https://doi.org/10.1140/epjc/s10052-011-1554-0

[276] Neyman J. and Pearson E. S., "On the Problem of the Most Efficient Tests of Statistical Hypotheses", Philosophical Transactions of the Royal Society of London, Series A, Containing Papers of a Mathematical or Physical Character 231 (1933), pp. 289-337, ISSN: 02643952, http://www.jstor.org/stable/91247

[277] Cranmer K., Lewis G., Moneta L. et al., "HistFactory: A tool for creating statistical models for use with RooFit and RooStats", Tech. Rep. CERN-OPEN-2012-016 (2012), https://cds.cern.ch/record/1456844

[278] Barlow R. and Beeston C., "Fitting using finite Monte Carlo samples", Comp. Phys. Comm. 77 (1993) 219, https://doi.org/10.1016/0010-4655(93)90005-W

[279] Barlow R. and Beeston C., "Minuit - a system for function minimization and analysis of the parameter errors and correlations", Comp. Phys. Comm. 10 (1975) 343, https://doi.org/10.1016/0010-4655(75)90039-9

[280] Verkerke W. and Kirkby D., "The RooFit toolkit for data modeling", arXiv:physics/0306116 [physics.data-an] (2003), https://doi.org/10.48550/arXiv.physics/0306116arXiv-issuedDOIviaDataCite

[281] Moneta L. et al., "The RooStats project", PoS ACAT2010 (2011) 057, arXiv:1009.1003 [physics.data-an], https://doi.org/10.48550/arXiv.1009.1003

[282] TRexFitter framework twiki page, https://twiki.cern.ch/twiki/bin/view/AtlasProtected/TtHFitter

[283] TRexFitter framework gitlab project, https://gitlab.cern.ch/TRExStats/TRExFitter

[284] Cranmer K., Lewis G. et al., "HistFactory: A tool for creating statistical models for use with RooFit and RooStats", CERN-OPEN-2012-016 (2012), https://cds.cern.ch/record/1456844

[285] Read A. L., "Presentation of search results: the CLs technique", Journal of Physics G: Nuclear and Particle Physics 28 (2002) 2693, https://dx.doi.org/10.1088/0954-3899/28/10/313

[286] Junk T., "Confidence Level Computation for Combining Searches with Small Statistics", Nucl. Instrum. Meth. in Phys. Res. A 434 (1999) 435, https://doi.org/10.1016/S0168-9002(99)00498-2

[287] Cousins R. D., "Generalization of Chisquare Goodness-of-Fit Test for Binned Data using Saturated Models, with Application to Histograms", Dept. of Physics and Astronomy University of California (2013), http://www.physics.ucla.edu/~cousins/stats/cousins_saturated.pdf

[288] Cowan G., "Statistical data analysis", Oxford University Press (1998) USA, ISBN: 9780198501558

[289] Asquith L. et al., "Search for the Standard Model Higgs boson produced in association with top quarks and decaying into $b\bar{b}$ in boosted topologies at $\sqrt{s} = 13$ TeV with the ATLAS detector", ATL-COM-PHYS-2017-396 (2017) CERN, https://cds.cern.ch/record/2260232

[290] The ATLAS collaboration, "Measurements of inclusive and differential fiducial cross-sections of $t\bar{t}$ production with additional heavy-flavour jets in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", High Energ. Phys. 04 (2019) 46, https://doi.org/10.1007/JHEP04(2019)046

[291] Buccioni F., Kallweit S., Pozzorini S. et al., "NLO QCD predictions for $t\bar{t}b\bar{b}$ production in association with a light jet at the LHC", J. High En. Phys. 12 (2019) 15, https://doi.org/10.1007/JHEP12(2019)015

[292] The ATLAS collaboration, "Jet reconstruction and performance using particle flow with the ATLAS Detector", Eur. Phys. J. C 77 466 (2017), https://doi.org/10.1140/epjc/s10052-017-5031-2

[293] The ATLAS Collaboration, "Expected performance of the 2019 ATLAS b-taggers", accessed 1 October 2022, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/FTAG-2019-005/

[294] Zaheer M. and Kottur S., "Deep Sets", arXiv:1703.06114 [cs.LG], https://doi.org/10.48550/arXiv.1703.06114

[295] The ATLAS Collaboration, "Optimisation of large-radius jet reconstruction for the ATLAS detector in 13 TeV proton–proton collisions", Eur. Phys. J. C 81 334 (2021), https://doi.org/10.1140/epjc/s10052-021-09054-3

[296] The ATLAS Collaboration, "Identification of boosted Higgs bosons decaying into b-quark pairs with the ATLAS detector at 13 TeV", Eur. Phys. J. C 79 836 (2019), https://doi.org/10.1140/epjc/s10052-019-7335-x

[297] The ATLAS Collaboration, "Variable Radius, Exclusive-k$_T$, and Center-of-Mass Subjet Reconstruction for Higgs($\rightarrow b\bar{b}$) Tagging in ATLAS", ATL-PHYS-PUB-2017-010 CERN (2017), https://cds.cern.ch/record/2268678

[298] D. Elitez, "Investigation of $t\bar{t}H(b\bar{b})$ events with very high Higgs boson momentum at ATLAS Detector", Bachelor thesis (2021), JGU Mainz

[299] CERN, "LHC Report: The switch was flipped, and the beams were dumped", accessed 10 December 2022, https://home.web.cern.ch/news/news/accelerators/lhc-report-switch-was-flipped-and-beams-were-dumped

Acknowledgements

This work would not have been accomplished without the support of all those who surrounded, both professionally and personally, during the last few years as a PhD student.

First and foremost, I would like to express my gratitude to my supervisor Prof. Dr. Lucia Masetti for her continuous support and encouragement from the very first moment we met. Her guidance in every theoretical and technical issue was invaluable. I appreciate all her insightful discussions as well as her time and effort, especially during the write up of this thesis. Furthermore, I am very grateful for the very constructive collaboration with my colleagues.

Moreover, I would like to acknowledge CERN for the very successful operation of the LHC as well as the ATLAS collaboration for performing this extraordinary experiment providing all the data for this analysis. I would also like to thank the ATLAS $t\bar{t}H(b\bar{b})$ group for the constructive collaboration during the conduction and publication of the analysis. Also, I gratefully acknowledge the computing time granted on the supercomputer Mogon at the Johannes Gutenberg University Mainz (hpc.uni-mainz.de). This PhD project was funded by PRISMA+ and the Mainz Physics Academy.

Last but not least, I owe my deepest gratitude to my family, my dear parents Katerina and Giorgos, and my sisters Marina and Tatiana, for their endless love and continuous support. I would also like to thank all my friends who accompanied me in my life up to now. Above all, I would like to express my thankfulness to Giorgos for his endless love, abiding support, patience, and confidence in me.

Curriculum Vitae

Personal information

Name:	Eftychia Tzovara
Date of birth:	19 May 1992
Nationality:	Greek
Address:	Im Münchfeld 33, 55122, Mainz, Germany
Telephone:	+49 176 34150478
E-mail:	etzovara@uni-mainz.de

Education

***2017 – present*, Doctorate in Physics**
Johannes Gutenberg University Mainz, Germany

***2016 – 2017*, Preliminary Study Fellowship in Physics** (1-year study programme)
Johannes Gutenberg University Mainz, Germany

***2010 – 2015*, Diploma in Physics** (4-year study programme)
National and Kapodistrian University of Athens, Greece

Publications (articles and conference contributions)

Measurement of Higgs boson decay into b-quarks in associated production with a top-quark pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector
The ATLAS Collaboration, JHEP 06 (2022) 97
https://doi.org/10.1007/JHEP06(2022)097

Measurement of the $t\bar{t}H$ production cross-section with $H \to b\bar{b}$ decay in the boosted topology with the ATLAS detector
14^{th} Annual Meeting "Physics at the Terascale", 23-24 November 2021
https://indico.desy.de/event/31325/contributions/112978/

Measurement of the $t\bar{t}H$ production cross-section with a collimated $H \to b\bar{b}$ decay in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector
EPS-HEP Conference 2021, 26-30 July 2021, ATL-COM-PHYS-2021-457
https://cds.cern.ch/record/2775411

Measurement of the $t\bar{t}H$ production cross-section with the $H \to b\bar{b}$ decay channel in the boosted topology
ATLAS-D Meeting 2020, Berlin, 8-11 September 2020
https://indico.cern.ch/event/865390/contributions/3987445/

Optimisation and improvements in the l+jets boosted channel
HTop Workshop 2019, DESY, 15-17 April 2019
https://indico.cern.ch/event/773548/contributions/3367879/