Alma Mater Studiorum · Università di Bologna

DOTTORATO DI RICERCA IN FISICA
Ciclo XXXIII

Settore concorsuale: 02/A1
Settore scientifico discplinare: FIS/01

Search for $t\bar{t}H(b\bar{b})$ events at 13 TeV in the fully hadronic final state targeting large-radius jets and a resolved decay of the Higgs boson with the CMS experiment

Presentata da:
Fabio Iemmi

Coordinatore dottorato:
Prof. Michele Cicoli

Supervisore:
Prof. Andrea Castro

Esame finale anno 2021

[...] εἰ γὰρ ἠδύνατο ἕκαστον τῶν ὀργάνων κελευσθὲν ἢ προαισθανόμενον ἀποτελεῖν τὸ αὑτοῦ ἔργον, <καὶ> ὥσπερ τὰ Δαιδάλου φασὶν ἢ τοὺς τοῦ Ἡφαίστου τρίποδας, οὓς φησιν ὁ ποιητὴς αὐτομάτους θεῖον δύεσθαι ἀγῶνα, οὕτως αἱ κερκίδες ἐκέρκιζον αὐταὶ καὶ τὰ πλῆκτρα ἐκιθάριζεν, οὐδὲν ἂν ἔδει οὔτε τοῖς ἀρχιτέκτοσιν ὑπηρετῶν οὔτε τοῖς δεσπόταις δούλων.

[...] *For if every instrument could accomplish its own work, obeying or anticipating the will of others, like the statues of Daedalus, or the tripods of Hephaestus, which, says the poet, "of their own accord entered the assembly of the Gods"; if, in like manner, the shuttle would weave and the plectrum touch the lyre without a hand to guide them, chief workmen would not want servants, nor masters slaves.*

Arist. *Pol.* I.4, 1253b33-39

Abstract

A search for $t\bar{t}H(b\bar{b})$ events in the fully hadronic final state is performed using data collected by the CMS experiment in proton-proton collisions at the center-of-mass energy of 13 TeV corresponding to an integrated luminosity of 35.9 fb^{-1}. The analysis is based on the study of events in the boosted topology, *i.e.*, events in which the decay products of at least one particle have a high Lorentz boost and are thus reconstructed in the detector as a single, large-radius jet. In this search, the Higgs boson is required to decay to a pair of well-separated jets. The observed (expected) upper limit at 95% confidence level on the signal strength parameter $\mu_{t\bar{t}H}$, defined as the ratio of the measured $t\bar{t}H$ production cross section to the one expected from a standard model Higgs boson, is found to be 11.5 (15.8) times the expectation from the standard model. By performing a combination with a complementary search targeting decays of the Higgs boson to a large-radius jet, the observed (expected) upper limit at 95% confidence level on the signal strength parameter is found to be 7.1 (9.0) times the expectation from the standard model.

Contents

Chapter 1

Introduction

The Higgs boson (H) plays a fundamental role in the Standard Model of the electroweak and strong interactions (SM), being the particle associated with the field that gives mass to all the elementary particles. The discovery of the Higgs boson in 2012 is among the most outstanding results obtained by the ATLAS and CMS Collaborations [1, 2] and, since then, many analyses have been developed in order to measure the Higgs boson mass and properties as precisely as possible [3, 4, 5, 6]. Up to date, no deviations from the SM predictions have been found.

At the CERN Large Hadron Collider (LHC), the SM Higgs boson, *i.e.*, a neutral, spinless, CP-even particle of measured mass of 125.38 ± 0.14 GeV [7], can mainly be created in four different processes or production modes: gluon-gluon fusion (ggH), vector boson fusion (VBF), Higgs-strahlung (VH) and top-antitop-Higgs ($t\bar{t}H$) associated production. In the ggH production mode, a gluon-induced fermionic loop, mostly dominated by top quarks, gives rise to a Higgs boson with no other associated particles. This is the most probable process, with a theoretical cross section at 13 TeV $\sigma_{ggH} = 48.6 \pm 2.4$ pb [8]. In the VBF process, two initial-state quarks emit a couple of virtual vector bosons, which annihilate to give rise to a Higgs boson. This process shows the characteristic signature of two jets in the final state, which can be used to target this particular production mode, and has a theoretical cross section at 13 TeV $\sigma_{VBF} = 3.78 \pm 0.08$ pb [8]. In the VH process, a couple of quarks annihilates to produce a virtual vector boson, which then irradiates a Higgs boson, resulting in a theoretical cross section at 13 TeV $\sigma_{WH,\,ZH} = 1.37 \pm 0.03, 0.88 \pm 0.04$ pb [8]. Finally, in the $t\bar{t}H$ production mode, which will be the main subject of this work, two gluons split in a top quark-antiquark ($t\bar{t}$) pair each, with one of such pairs annihilating to a Higgs boson, resulting in a final state with a Higgs boson, a top quark and a top antiquark. This process is the rarest, having a theoretical cross section at 13 TeV $\sigma_{t\bar{t}H} = 0.50^{+0.05}_{-0.07}$ pb [8]. The four, main production mechanisms are depicted in Fig. 1.1, where the leading order Feynman diagrams for such

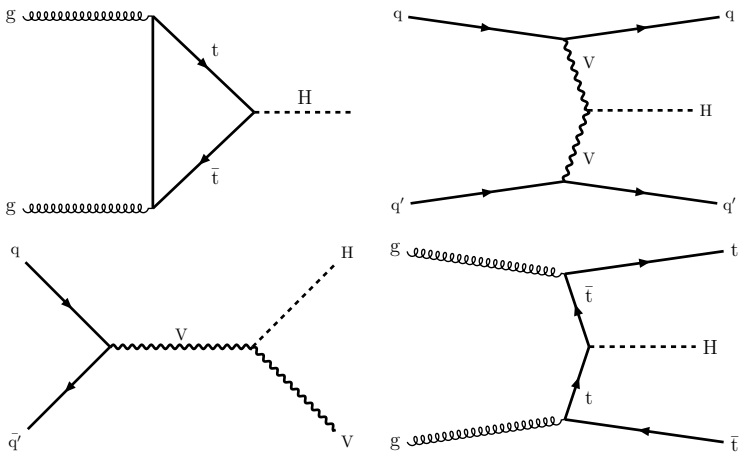

Figure 1.1: Leading order Feynman diagrams for the main production modes of the SM Higgs boson at the LHC: ggH (upper left), VBF (upper right), VH (lower left) and $t\bar{t}H$ (lower right).

processes are reported.

Once produced, the SM Higgs boson can decay in several different ways. As it will be described in detail in the following, the Higgs boson coupling with SM particles is in general proportional to the mass of the particles. If we take into account some limitations due to the kinematics, which imply that the decays to a $t\bar{t}$ pair and to a pair of real vector bosons are forbidden, the dominant decay channel for the Higgs boson is predicted to be the one to a pair of bottom quarks, $H \to b\bar{b}$, with a branching ratio $BR(H \to b\bar{b}) = (58.4 \pm 1.9)\%$ [8], which has been observed by the ATLAS and CMS Collaborations [9, 10]. Even though it provides a high event yield, searches in this decay channel are challenging, given the high level of background which tends to mask the $b\bar{b}$ decay.

The second most dominant decay mode is predicted to be the decay to a couple of W bosons, $H \to WW^*$, where, due to the kinematic limitations mentioned above, one of the W bosons is produced off-shell. This channel has a branching ratio $BR(H \to WW^*) = (21.4 \pm 0.8)\%$ [8], and has been observed by the ATLAS and CMS Collaborations [11, 12] in the case of a decay to a lepton-neutrino pair for both the vector bosons. The $H \to WW^*$ channel benefits from a relatively high branching fraction but has the drawback of the presence of neutrinos in the final state, which make the full reconstruction of the final state impossible.

The subdominant decay channels are the $H \to \tau\tau$ and $H \to ZZ^*$ channels, in which the Higgs boson decays to a pair of τ leptons and Z bosons respectively. The search of the former process is relevant to investigate the coupling of the Higgs boson with third generation fermions and is challenging in the

fact that many final states are possible, which cannot be fully reconstructed. The search of the latter process is very important since, in the case of a decay of the ZZ^* pair to four leptons (sometimes referred to as the "golden channel"), it gives a very clean, fully reconstructable signature with low background. Both such decay modes have been observed by the ATLAS and CMS Collaborations [13, 14, 15, 16].

Another important decay channel is the $H \to \gamma\gamma$, in which the Higgs boson decays to a pair of photons. Being the photon a massless particle, the $H \to \gamma\gamma$ decay must proceed through a loop-level process, dominated by a top quark loop. As in the case of the ZZ^* decay mode, also the $\gamma\gamma$ decay channel offers a clean, fully reconstructable final state, nevertheless having a lower branching fraction. This decay channel has been observed by the ATLAS and CMS Collaborations [17, 18].

Finally, even rarer decay modes are predicted by the SM, such as the decay to a Z boson and a photon, $H \to Z\gamma$ and the decay to a pair of muons, $H \to \mu\mu$. The former is interesting as it proceeds through loop processes which could involve new physics, resulting in a modified rate with respect to the SM one. The latter decay mode is important as it would be the fist observation of the coupling of the Higgs boson to a second generation fermion. None of these rare decay modes have been observed so far, due to the very small branching fractions. However, the CMS Collaboration has recently reported the first evidence for the $H \to \mu\mu$ decay with a significance of 3.0 standard deviations [19].

The main results obtained by the ATLAS and CMS Collaborations in the aforementioned production and decay modes are summarized in Table 1.1.

The discovery of the top quark dates back to 1995, when the two experiments CDF and D0, operating at the Tevatron, announced independently [20, 21] the observation of a particle compatible with a new quark. The top quark plays a very important role in the SM due to its very large mass $m_t = 173.0 \pm 0.4$ GeV [8], so that the precise knowledge of its properties is critical for the general understanding of the theory. For example, it turns out from SM calculations that the mass of the W boson can be computed as:

$$m_W = \left(\frac{\pi\alpha}{G_F\sqrt{2}}\right)^{\frac{1}{2}} \frac{1}{\sin\theta_W\sqrt{1-\Delta r}},$$

where α is the fine-structure constant, G_F is the Fermi constant, θ_W is the Weinberg angle and Δr is a function related to the radiative corrections of the W propagator, see Fig. 1.2, of the form $\Delta r \sim f(m_t, \log m_H)$. Since the parameters α, G_F and θ_W are measured with very high precision, accurate measurements of m_t (and m_W) led, in the past, to constraints on the value of the Higgs boson mass m_H [22] and, now that the Higgs boson has been discovered, can be used to check the overall consistency of the SM.

Decay mode	ggH	VBF	VH	$t\bar{t}H$
$b\bar{b}$ (A)	–	-3.9 ± 2.8	0.9 ± 0.27	0.83 ± 0.63
$b\bar{b}$ (C)	2.3 ± 1.66	2.8 ± 1.5	1.2 ± 0.4	0.82 ± 0.43
WW^* (A)	1.21 ± 0.22	0.62 ± 0.36	3.2 ± 4.3	1.50 ± 0.61
WW^* (C)	1.38 ± 0.23	0.29 ± 0.48	3.27 ± 1.84	1.97 ± 0.67
$\tau\tau$ (A)	1.14 ± 0.44	0.98 ± 0.46	2.3 ± 1.6	1.36 ± 1.11
$\tau\tau$ (C)	1.2 ± 0.5	1.11 ± 0.34	-0.33 ± 1.02	0.28 ± 1.02
$ZZ^*(4\ell)$ (A)	1.04 ± 0.17	2.8 ± 0.95	0.9 ± 1.0	< 1.8 68% CL
$ZZ^*(4\ell)$ (C)	1.20 ± 0.22	0.05 ± 0.04	0.0 ± 1.5	< 1.3 68% CL
$\gamma\gamma$ (A)	0.81 ± 0.18	2.0 ± 0.6	0.7 ± 0.8	1.4 ± 0.4
$\gamma\gamma$ (C)	1.10 ± 0.19	0.8 ± 0.6	2.4 ± 1.1	2.3 ± 0.8
$Z\gamma$ (A)	$< 6.6(5.2)$	Incl.	–	–
$Z\gamma$ (C)	$< 3.9(2.9)$	Incl.	Incl.	–
$\mu\mu$ (A)	$< 3.0(3.1)$	Incl.	–	–
$\mu\mu$ (C)	$< 2.6(1.9)$	–	–	–

Table 1.1: Summary of the ATLAS (A) and CMS (C) measurements and upper limits for different production modes and decay channels of the Higgs boson. For observed channels, the measured cross section times branching ratio, normalized to the expectation of the SM, is reported. For channels which have not been observed yet, observed (expected) upper limits at 95% CL on the same quantity are reported. In the case of rare decay modes, results are reported as limits and secondary production mechanisms which are included in the analyses are labeled as "Incl.".

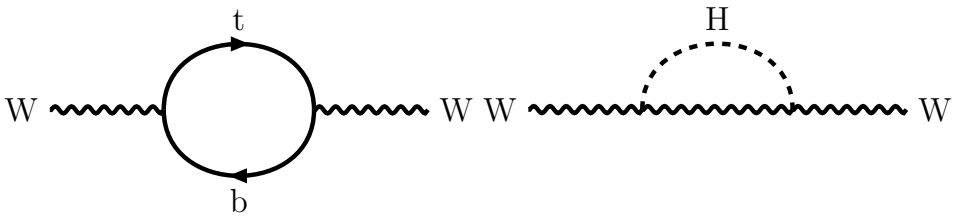

Figure 1.2: Radiative corrections to the W boson propagator. Top-bottom quarks loop (left) and Higgs boson loop (right).

Once produced, a t ($\bar{\text{t}}$) quark decays in a time comparable to 10^{-24} s to a W^+b ($\text{W}^-\bar{\text{b}}$) pair, different decays being highly subdominant. Such a short decay time is due to the large mass of the top quark and the related wide phase space available for the decay, and does not allow the top quark to form bounded states. Thus, three decay channels are available for a $\text{t}\bar{\text{t}}$ pair: a leptonic decay channel, in which both the W bosons arising from the top quarks decay to a lepton-neutrino pair, with a branching ratio of about 5%; a semileptonic decay channel, in which one W boson decays to a lepton-neutrino pair and the other decays to a pair of light quarks, with a branching ratio of about 30%; and finally an all-hadronic decay channel, in which both the W bosons decay to a pair of light quark, with a branching ratio of about 46%.

Searches for the associated production of a Higgs boson and a $\text{t}\bar{\text{t}}$ pair are of crucial importance in the context of the SM. First of all, they complete the measurements of the couplings of the Higgs boson with third generation fermions. In the SM framework, the Higgs boson is predicted to couple to fermions with a Yukawa-like interaction, the coupling being proportional to the fermion mass. Since the top quark is the heaviest fermion, it shows the greatest coupling to the Higgs boson. However, given the fact that the mass of the top quark exceeds the mass of the Higgs boson, the direct decay $\text{H} \rightarrow \text{t}\bar{\text{t}}$ is kinematically forbidden. Indirect measurements of the top-Higgs coupling can be performed in loop-level processes, such as the Higgs boson production through gluon-gluon fusion reported in Fig. 1.1 or the Higgs boson decay to photons; however, a direct measurement of the top-Higgs coupling exploiting tree-level processes, such as in the case of the $\text{t}\bar{\text{t}}\text{H}$ associated production, is crucial in order to exclude beyond standard model contributions which could enter the loops unnoticed. Both the ATLAS and CMS Collaborations have reported the observation of the $\text{t}\bar{\text{t}}\text{H}$ process with a significance above five standard deviations [23, 24].

Also, the Higgs boson and top quark masses play a fundamental role in determining the nature of the stability of the electroweak vacuum. A crucial question is if it is possible to extend the spontaneous symmetry breaking mechanism (see Chapter 2) up to high energy scales while keeping the minimum of the scalar potential that breaks the electroweak symmetry stable. Rather intriguingly, current measurements of the Higgs and top masses lead to a metastable configuration for the electroweak vacuum [25, 26], that is, it may exists a lower energy vacuum state available to which the electroweak vacuum can decay into [27]. This has very important consequences at the cosmological level, implying, among other things, a finite age for the universe [28]. A deeper understanding of the top-Higgs Yukawa coupling, which is related to the top mass, may lead to improvements in the comprehension of this scenario.

Given the above description of the decay modes of the Higgs boson and of the top quark, it appears evident that many different decay channels

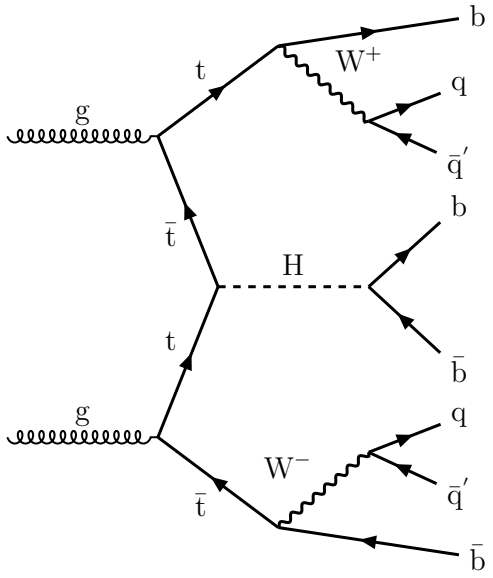

Figure 1.3: Leading order Feynman diagram for the $t\bar{t}H$ production and the subsequent decay in the FH channel. The final state is nominally composed by eight partons, four of them which are bottom quarks.

are possible for a $t\bar{t}H$ triplet. The analysis described in this works targets the so called fully hadronic (FH) final state, in which the Higgs boson decays to a $b\bar{b}$ pair and the $t\bar{t}$ pair decays in the all-hadronic final state. As a result, the final state is composed by at least eight partons (more are possible due to initial/final state radiation), four of them which are bottom quarks. Combining the branching ratios of the Higgs boson and top quark decay modes, a total branching ratio for the FH $t\bar{t}H$ final state $\mathrm{BR}(t\bar{t}H \rightarrow \mathrm{FH}) = 0.58 \times 0.46 \approx 0.27$ is found, making the FH one the most probable among all the possible final states.

Even though it shows the highest branching fraction, the FH final state turns out to be very challenging, having many particles that are difficult to identify, such as the b quarks, and suffering from large contamination from the main background at hadron colliders, namely the QCD multijet production. Moreover, also the production of $t\bar{t}$ pairs with additional jets is found to be an overwhelming background for this kind of search. Given the presence of many jets, measurements in the FH channel involve larger uncertainties (coming from jet energy corrections and b tagging) than in the leptonic channels, but offer the unique possibility to fully reconstruct the $t\bar{t}H$ system, since no missing energy/momentum is present. The leading order Feynman diagram for the $t\bar{t}H$ production and subsequent decay in the FH channel is shown in Fig. 1.3

The standard approach adopted in the CMS Collaboration to search for

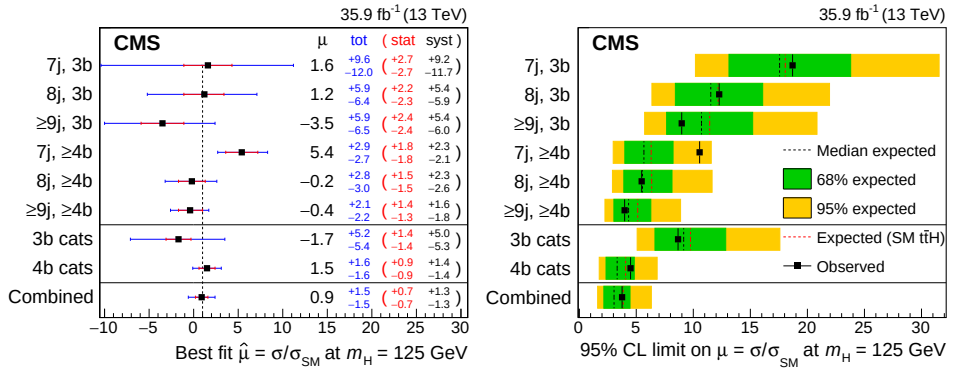

Figure 1.4: Best fit values in the signal strength modifiers $\hat{\mu}$, and their 68% CL intervals as split into the statistical and systematic components (left), and median expected and observed 95% CL upper limits on μ (right). The expected limits are displyed with their 68% and 95% CL intervals, as well as with the expectation for an injected SM signal of $\mu = 1$. (Figures and caption from [29]).

$t\bar{t}H$ events in the FH final state [29] is to look for events consisting of eight, well separated jets (the so called resolved topology), four of them which are supposed to be identified as coming from the hadronization of bottom quarks. Due to the final state with high jet multiplicity, sophisticated methods such as quark-gluon discrimination techniques [30, 31] and the matrix element method [32, 33] must be exploited to discriminate quark-induced jets from gluon-induced jets and to separate the $t\bar{t}H$ signal from the background. The results of this search are reported in terms of the so called signal strength parameter $\mu_{t\bar{t}H}$, which is defined as the ratio of the measured $t\bar{t}H$ production cross section to the one expected from a SM Higgs boson with a mass equal to 125 GeV. From a combined fit of the signal and background template shapes to the data in all the event categories, observed and exepcted upper limits $\mu_{t\bar{t}H} < 3.8$ and < 3.1 have been obtained at 95% CL, corresponding to a best fit value $\hat{\mu}_{t\bar{t}H} = 0.9 \pm 0.7$ (stat) ± 1.3 (syst). These results are summarized in Fig. 1.4.

However, in the latest runs of the LHC at a centre-of-mass energy $\sqrt{s} =$ 13 TeV, very high-p_T top quarks and Higgs bosons can be produced. If their Lorentz boost is sufficiently high (this typically happens for top quarks with $p_T > 350$ GeV and Higgs bosons with $p_T > 300$ GeV), their decay products are found to be very collimated and merge into a single, wide jet, which is usually referred to as a boosted jet. The properties of the decaying particles can then be studied by looking at the substructure of such wide jets, which encodes the information about the decay products. Thus, the search presented in this work exploits a novel approach by focusing on $t\bar{t}H$ events

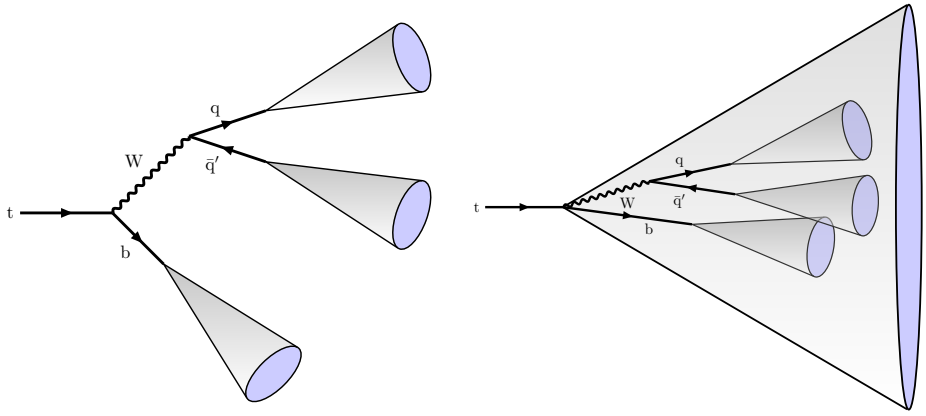

Figure 1.5: Different fully hadronic decay topologies for a top quark: resolved decay (left), in which three, well separated jets are reconstructed in the detector, and boosted decay (right) in which, due to the high Lorentz boost of the particle, the decay products are highly collimated and are reconstructed as a single, large-radius jet.

in the boosted, FH final state, namely the FH final state in which at least one boosted jet is present. In particular, it targets resolved decays of the Higgs boson (resolved-Higgs channel, RHC), *i.e.*, decays to a well separated pair of jets identified to be the result of the hadronization of bottom quarks, thus the boosted jets being supposed to come from the decay of top quarks. The RHC is one of the two channels forming a wider search of $\mathrm{t\bar{t}H}$ events in the boosted, FH final state. The complementary channel, the boosted-Higgs channel (BHC) targets, instead, decays of the Higgs boson to a boosted jet. As it will be explained in the following chapters, the RHC presented in this work has been constructed to be orthogonal to the BHC, in such a way that the two channels can be safely combined to achieve a better sensitivity. In the end of this work, the results of such a combination will be shown. Both the searches in these mutually exclusive channels are performed using proton-proton (pp) collisions collected by the CMS experiment during the 2016 data-taking period and corresponding to an integrated luminosity of 35.9 fb^{-1}.

Analyses can evidently benefit from the novel approach involving boosted jets: since the decay products of one or more particles are collected in a single, wide jet, the combinatorics of the problem is sensibly reduced, and the properties of the particles, such as masses or transverse momenta, can be inferred by directly looking at such jets, as can be seen in the pictorial representation of a boosted decay reported in Fig. 1.5. Also, the exploitation of decays in the boosted regime makes possible to more than double the spectrum of differential measurements as a function of the particles p_{T}, reaching the TeV frontier [34] (however, such measurements are still not

possible for rare processes such as the $t\bar{t}H$ production, as more integrated luminosity is needed).

On the other hand, as far as the $t\bar{t}H$ searches are concerned, such advantages are achieved at the price of a significantly lower signal acceptance, as the resolved decay topology is found to be still favored at $\sqrt{s} = 13$ TeV. However, since in the coming years the number of multiple pp collisions within the same bunch crossing is expected to increase up to an average of 60 during Run3 and of 140 in the context of the High-Luminosity LHC project [35], low-p_{T} jets are expected to become more difficult to trigger, and searches will benefit for the inclusion of boosted jets in the final state.

Chapter 2

The Standard Model of the electroweak and strong interactions

This chapter is devoted to the formal exposition of the theory describing the interaction among elementary particles. Being the Higgs boson and its interactions the subject of this thesis, a special focus on the spontaneous simmetry breaking mechanism is provided. The material presented in the following is a personal elaboration of the matter discussed in [36] and in the notes for the Quantum Field Theories courses held by Prof. Roberto Soldati at the Alma Mater Studiorum University of Bologna, whom I would like to thank for the support in the computation of the full electroweak Lagrangian.

2.1 Gauge invariance in classical electrodynamics

As an introductory step, let us review the well known results concerning the gauge arbitrariness in classical electrodynamics.

The Gauss' theorem for the magnetic field $\mathbf{B}$ (second Maxwell's equation)

$$\nabla \cdot \mathbf{B} = 0, \tag{2.1}$$

suggests the possibility of writing the magnetic field as the curl of some vector potential $\mathbf{A}$, *i.e.*

$$\mathbf{B} = \nabla \times \mathbf{A}. \tag{2.2}$$

In fact, the well known identity $\nabla \cdot (\nabla \times \mathbf{A}) = 0$, which is valid for every $\mathbf{A}$, guarantees that the $\mathbf{B}$ field is divergenceless.

In a similar fashion, the Faraday-Neumann-Lenz equation for the electric

field **E** (Maxwell's third equation)

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{2.3}$$

can be rewritten, making use of Eq. 2.2, as

$$\nabla \times (\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t}) = 0, \tag{2.4}$$

which suggests the possibility of writing $\mathbf{E} + \partial \mathbf{A}/\partial t$ as the gradient of some scalar potential V, *i.e.*

$$\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} = -\nabla V. \tag{2.5}$$

In fact, the well known identity $\nabla \times (\nabla V) = 0$, which is valid for every V, guarantees that Eq. 2.4 is satisfied. Let us now note that, if we add an arbitrary gradient to the vector potential and an arbitrary time derivative to the scalar potential, namely if we perform the transformations

$$\begin{aligned} &\mathbf{A} \to \mathbf{A} + \nabla \Lambda \\ &V \to V - \partial \Lambda / \partial t, \end{aligned} \tag{2.6}$$

the electric and magnetic fields given by Eqs. 2.2 and 2.5 are unchanged.

All this can be summarized by adopting a covariant notation. First we introduce the antisymmetric electromagnetic field-strength tensor

$$F^{\mu\nu} = -F^{\nu\mu} = \partial^\nu A^\mu - \partial^\mu A^\nu = \begin{pmatrix} 0 & E_1 & E_2 & E_3 \\ -E_1 & 0 & B_3 & -B_2 \\ -E_2 & -B_3 & 0 & B_1 \\ -E_3 & B_2 & -B_1 & 0 \end{pmatrix}, \tag{2.7}$$

which is build up starting from the four-vector potential

$$A^\mu = (V; \mathbf{A}), \tag{2.8}$$

and which is unchanged by the gauge transformation

$$A^\mu \to A^\mu - \partial^\mu \Lambda, \tag{2.9}$$

where $\Lambda(x)$ is an arbitrary function of the coordinates. The fact that many different four-vector potentials lead to the same electric and magnetic fields, and thus describe the very same physics, is known as the gauge invariance of classical electrodynamics.

The introduction of the dual field-strength tensor

$$ {}^*F^{\mu\nu} = -\frac{1}{2}\varepsilon^{\mu\nu\rho\sigma}F_{\rho\sigma} = \begin{pmatrix} 0 & B_1 & B_2 & B_3 \\ -B_1 & 0 & -E_3 & E_2 \\ -B_2 & E_3 & 0 & -E_1 \\ -B_3 & -E_2 & E_1 & 0 \end{pmatrix}, \tag{2.10} $$

where we used the convention that the Levi-Civita's asymmetric symbol $\epsilon^{\alpha\beta\gamma\delta}$ is equal to ∓ 1 for even or odd permutations of $\{0, 1, 2, 3\}$ and that $\epsilon^{\alpha\beta\gamma\delta} = -\epsilon_{\alpha\beta\gamma\delta}$, leads to a very elegant and compact expression for the Maxwell's equations 2.1 and 2.3. In fact, they can be written as

$$ \partial_\mu \, {}^*F^{\mu\nu} = 0. \tag{2.11} $$

The remaining Maxwell's equations, namely

$$ \begin{aligned} \nabla \cdot \mathbf{E} &= \rho \\ \nabla \times \mathbf{B} &= \mathbf{J} + \frac{\partial \mathbf{E}}{\partial t}, \end{aligned} \tag{2.12} $$

are obtained, in covariant notation, as

$$ \partial_\mu F^{\mu\nu} = -J^\nu, \tag{2.13} $$

where we have introduced the electromagnetic four-current

$$ J^\nu = (\rho; \mathbf{J}). \tag{2.14} $$

As a final remark, let us point out two important and immediate consequences of Eq. 2.13: first, given the completely antysimmetric nature of the electromagnetic field-strength tensor, the electromagnetic current is conserved, that is

$$ \partial_\nu J^\nu = -\partial_\nu \partial_\mu F^{\mu\nu} = 0. \tag{2.15} $$

Also, by expanding them and in the Lorentz gauge $\partial_\mu A^\mu = 0$, these equations become

$$ \Box A^\nu = 0, \tag{2.16} $$

where $\Box \equiv \partial_\mu \partial^\mu$ is the D'Alembert operator, which means that each component of the four-vector potential, which can be identified with the photon, satisfies a massless Klein–Gordon equation. In such a way, a connection is made between gauge invariance, current conservation and massless vector fields.

Let us now explore in more detail this subject by switching to quantum field theories.

2.2 Phase invariance in quantum field theories

Let us consider the Lagrangian for a complex scalar field, namely

$$\mathcal{L} = \partial_\mu \phi \partial^\mu \phi^* - m^2 \phi \phi^*. \tag{2.17}$$

If we make use of the well known Euler–Lagrange equations for the field $u(x)$

$$\partial_\mu \frac{\delta \mathcal{L}}{\delta \partial_\mu u(x)} - \frac{\delta \mathcal{L}}{\delta u(x)} = 0, \tag{2.18}$$

and we compute them for ϕ and ϕ^*, we readily obtain the Klein–Gordon equations

$$(\Box + m^2)\phi(x) = 0, \quad (\Box + m^2)\phi^*(x) = 0. \tag{2.19}$$

If we now perform a global (namely, coordinate-independent) phase transformation of these fields, *i.e.*,

$$\phi(x) \to e^{iq\alpha}\phi(x), \quad \phi^*(x) \to e^{-iq\alpha}\phi^*(x) \tag{2.20}$$

and we make use of the well known results by Emmy Noether, we can eventually identify a conserved Noether current of the form

$$\begin{aligned} j^\mu &= -iq\left[\frac{\delta \mathcal{L}}{\delta\left(\partial_\mu \phi\right)}\phi - \frac{\delta \mathcal{L}}{\delta\left(\partial_\mu \phi^*\right)}\phi^*\right] \\ &= iq\left[\phi^* \partial^\mu \phi - \left(\partial^\mu \phi^*\right)\phi\right] \equiv iq\phi^* \overset{\leftrightarrow}{\partial^\mu} \phi, \end{aligned} \tag{2.21}$$

which satisfies

$$\partial_\mu j^\mu = 0. \tag{2.22}$$

With the identification of q as the electric charge, we can interpret Eq. 2.21 as the electromagnetic current of the charged scalar field. The connection between global phase invariance and current conservation is guaranteed by the Noether's theorem [37], which states that to every continuous transformation of coordinates and fields which makes the variation of the action equal to zero, there always corresponds a constant of motion, *i.e.* a combination of the fields and their derivatives which is constant in time.

Let us now make a step further and consider the consequences of imposing a local (namely, coordinate-dependent) phase rotation to the scalar fields, that is

$$\phi(x) \to e^{iq\alpha(x)}\phi(x), \quad \phi^*(x) \to e^{-iq\alpha(x)}\phi^*(x). \tag{2.23}$$

While the terms of the Lagrangian depending only on the fields are left unchanged by the action of Eqs. 2.23, the transformation of gradient terms involves more than a simple phase change, that is

$$\partial_\mu \phi(x) \to e^{iq\alpha(x)} \left[\partial_\mu \phi(x) + iq \left(\partial_\mu \alpha(x)\right) \phi(x)\right], \tag{2.24}$$

with the presence of the gradient term of the local function $\alpha(x)$ actually spoiling the local phase space invariance. We will now show that, eventually, the local phase space invariance can be recovered by modifying the derivative operator and introducing the electromagnetic field A^μ. In fact, let us first replace the gradient ∂_μ in the Lagrangian with the gauge-covariant derivative defined by

$$\mathcal{D}_\mu = \partial_\mu + iqA_\mu(x). \tag{2.25}$$

We also impose that, under the phase rotations 2.23, the vector field A_μ transforms as

$$A_\mu(x) \to A_\mu(x) - \partial_\mu \alpha(x), \tag{2.26}$$

which does not change the physics at all by virtue of the gauge invariance of electromagnetism, as explained in Section 2.1. Under these conditions, an object such as $\mathcal{D}_\mu \phi$ will simply transform with the same phase rotation as the fields, that is

$$\mathcal{D}_\mu \phi \to e^{iq\alpha(x)} \mathcal{D}_\mu \phi, \tag{2.27}$$

and local gauge invariance is preserved.

To point out the consequences of the introduction of the covariant derivative, let us have a look at the Dirac Lagrangian for a free particle of spin 1/2, *i.e.*

$$\mathcal{L}_{\text{free}} = \overline{\psi} \left(i\gamma^\mu \partial_\mu - m\right) \psi \tag{2.28}$$

and replace the gradient operator with the covariant derivative. We get

$$\begin{aligned} \mathcal{L} &= \overline{\psi} \left(i\gamma^\mu \mathcal{D}_\mu - m\right) \psi \\ &= \overline{\psi} \left(i\gamma^\mu \partial_\mu - m\right) \psi - qA_\mu \overline{\psi} \gamma^\mu \psi \\ &= \mathcal{L}_{\text{free}} - J^\mu A_\mu, \end{aligned} \tag{2.29}$$

where we have obtained an interaction term in the Lagrangian between the spinor and vector fields, which is $-qA_\mu \overline{\psi} \gamma^\mu \psi$, containing the conserved electromagnetic current in the familiar form

$$J^\mu = q\bar{\psi} \gamma^\mu \psi. \tag{2.30}$$

Note that this form of the current is exactly the one that can be derived from the global gauge invariance using the Noether's theorem. Also, it is easy to show that the Lagrangian of Eq. 2.29 is invariant under a local phase rotation of the fields ψ and $\overline{\psi}$ and under the transformation of Eq. 2.26 for the scalar field A_μ.

We have thus shown that invariance under a local phase rotation can be achieved at the price of introducing an electromagnetic interaction. This shows the possibility of using local gauge invariance as a dynamical principle.

As a final remark, we shall obtain the complete Lagrangian for QED by simply adding a kinetic energy term for the vector field to Eq. 2.29, which is the well known, manifestly gauge invariant term $-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}$:

$$\mathcal{L}_{\text{QED}} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \mathcal{L}_{\text{free}} - J^{\mu}A_{\mu}. \tag{2.31}$$

Notice that a mass term for the photon would have the form $\frac{1}{2}m^2A_{\mu}A^{\mu}$, which is manifestly violating the gauge invariance. Thus, the requirement of a gauge invariant Lagrangian leads to the existence of a massless photon.

2.3 Non-abelian gauge theories

We have just shown that the Lagrangian for the QED reported in Eq. 2.31 represents a gauge invariant theory, which means that is unchanged by the action of the transformations

$$\begin{aligned} A_{\mu}(x) &\to A_{\mu}(x) - \partial_{\mu}\alpha(x) \\ \psi(x) &\to e^{iq\alpha(x)}\psi(x). \end{aligned} \tag{2.32}$$

Non-abelian gauge theories consist in a generalization of the concepts developed for QED, which involves a non-abelian $SU(N)$ transformation group instead of the abelian $U(1)$ group of phase rotations. The $SU(N)$ group effectively acts on N-dimensional objects of the kind

$$\psi(x) = \begin{pmatrix} \psi_1(x) \\ \psi_2(x) \\ \vdots \\ \psi_N(x) \end{pmatrix}, \tag{2.33}$$

where each of the $\psi_i(x)$ is a four-component Dirac spinor. The action of the symmetry group on the N-dimensional spinors is a generalization of the abelian case and is now given by

$$\psi(x) \to U_{\omega}(x)\psi(x), \tag{2.34}$$

where

$$U_{\omega} = \exp\{ig\tau_F^a\omega^a(x)\} \tag{2.35}$$

is a generic element of the group $SU(N)$ in its exponential representation. Here, the group index a runs from 1 to (N^2-1) and identifies the (N^2-1) infinitesimal generators of the symmetry group, τ_F^a, while the $\omega^a(x)$ are

generic functions of the coordinates. Summation over repeated group indices is understood in this formula and in the following. Finally, g plays the role of a coupling constant. Let us recall that the non-abelian nature of the $SU(N)$ symmetry group is given by the commutation rules that exist among the infinitesimal generators, *i.e.*

$$\left[\tau_F^a, \tau_F^b\right] = i f^{abc} \tau_F^c, \tag{2.36}$$

where f^{abc} are completely anti-symmetrical, real quantites called the structure constants of the group.

In order to proceed with the generalization of QED, we first need a generalized covariant derivative. To this goal we define

$$D_\mu \psi(x) = \left(\mathbb{1}_{N\times N}\partial_\mu + ig\tau_F^a A_\mu^a(x)\right)\psi(x) \equiv \left(\partial_\mu + igA_\mu(x)\right)\psi(x), \tag{2.37}$$

where $\mathbb{1}_{N\times N}$ is the N-dimentional identity matrix and we have introduced (N^2-1), non-abelian vector fields $A_\mu^a(x)$. Note that, now, the quantity denoted with $A_\mu(x)$ is actually a matrix. To familiarize with this generalization, we shall compute the form of $A_\mu(x)$ in the case of the symmetry group $SU(2)$ acting on a two-dimensional space. For such a group and representation, the infinitesimal generators are the well known Pauli matrices

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}; \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}; \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}; \tag{2.38}$$

in such a way that we obtain

$$\begin{aligned} A_\mu \equiv \tau_F^a A_\mu^a(x) = &\begin{pmatrix} 0 & A_\mu^1(x) \\ A_\mu^1(x) & 0 \end{pmatrix} + \begin{pmatrix} 0 & -iA_\mu^2(x) \\ iA_\mu^2(x) & 0 \end{pmatrix} + \\ &+ \begin{pmatrix} A_\mu^3(x) & 0 \\ 0 & -A_\mu^3(x) \end{pmatrix} = \\ &= \begin{pmatrix} A_\mu^3(x) & A_\mu^1(x) - iA_\mu^2(x) \\ A_\mu^1(x) + iA_\mu^2(x) & -A_\mu^3(x) \end{pmatrix}, \end{aligned} \tag{2.39}$$

so that the matrix nature of A_μ has been made explicit.

As a second step, let us generalize the transformation for the vector potential in the following way:

$$A_\mu(x) \to U_\omega(x) A_\mu(x) U_\omega^{-1}(x) + \frac{i}{g}\left[\partial_\mu U_\omega(x)\right] U_\omega^{-1}(x). \tag{2.40}$$

First, we note that this transformation reduces to the previously described transformation given by Eq. 2.26 in the case of a $U(1)$ phase rotation $e^{ig\alpha(x)}$. Also, most importantly, it is easily shown, using Eqs. 2.34 and 2.40, that objects such as $D_\mu\psi(x)$ transform in an homogeneous way, in fact:

$$\begin{aligned}
D_\mu\psi(x) \to &\left[\partial_\mu + ig\left(U_\omega(x)A_\mu(x)U_\omega^{-1}(x) + \frac{i}{g}(\partial_\mu U_\omega(x))U_\omega^{-1}(x)\right)\right]U_\omega(x)\psi(x) = \\
&= (\partial_\mu U_\omega(x))\psi(x) + U_\omega(x)\partial_\mu\psi(x) + igU_\omega(x)A_\mu(x)U_\omega^{-1}(x)U_\omega(x)\psi(x) + \\
&- (\partial_\mu U_\omega(x))U_\omega^{-1}(x)U_\omega(x)\psi(x) = \\
&= (\partial_\mu U_\omega(x))\psi(x) + U_\omega(x)\partial_\mu\psi(x) + igU_\omega(x)A_\mu(x)\psi(x) + \\
&- (\partial_\mu U_\omega(x))\psi(x) = \\
&= U_\omega(x)\left[\partial_\mu + igA_\mu(x)\right]\psi(x),
\end{aligned}$$

which means that

$$D_\mu\psi(x) \to U_\omega(x)D_\mu\psi(x). \tag{2.41}$$

As a final step towards our generalization of the QED framework, we need an expression for the Maxwell's field-strength tensor $F_{\mu\nu}$. First, we observe that, in the abelian case, the field-strength tensor can be written as

$$F_{\mu\nu} = -\frac{i}{q}[\mathcal{D}_\nu, \mathcal{D}_\mu], \tag{2.42}$$

with $\mathcal{D}_\nu$ given by Eq. 2.25, where the commutator term between the vector fields vanishes in the case of a $U(1)$ theory. Inspired by this, we can compute the analogous commutator $[D_\nu, D_\mu]$ for the non-abelian case. This reads:

$$\begin{aligned}
[D_\nu, D_\mu] &= [\partial_\nu + igA_\nu, \partial_\mu + igA_\mu] = \\
&= (\partial_\nu + ig\tau_F^a A_\nu^a)(\partial_\mu + ig\tau_F^b A_\mu^b) - (\partial_\mu + ig\tau_F^b A_\mu^b)(\partial_\nu + ig\tau_F^a A_\nu^a) = \\
&= \partial_\nu\partial_\mu + ig\tau_F^b\partial_\nu A_\mu^b + ig\tau_F^b A_\mu^b\partial_\nu + ig\tau_F^a A_\nu^a\partial_\mu - g^2\tau_F^a A_\nu^a\tau_F^b A_\mu^b + \\
&- \partial_\mu\partial_\nu - ig\tau_F^a\partial_\mu A_\nu^a - ig\tau_F^a A_\nu^a\partial_\mu - ig\tau_F^b A_\mu^b\partial_\nu + g^2\tau_F^b A_\mu^b\tau_F^a A_\nu^a = \\
&= ig(\tau_F^b\partial_\nu A_\mu^b - \tau_F^a\partial_\mu A_\nu^a) - g^2[\tau_F^a, \tau_F^b]A_\nu^a A_\mu^b = \\
&= ig(\partial_\nu A_\mu - \partial_\mu A_\nu) - ig^2 f^{abc}\tau_F^c A_\nu^a A_\mu^b,
\end{aligned}$$

where in the last equality we have made use of Eq. 2.36. Note that the equation above is similar in form to the previously defined abelian field-strength tensor, since it contains the term $(\partial_\nu A_\mu - \partial_\mu A_\nu)$, but it also has an additional term depending on the structure constant of the symmetry group. Thus, after multiplying by $(-i)$ and dividing by g, we shall identify the quantity

$$
\begin{aligned}
-\frac{i}{g}\left[D_\nu, D_\mu\right] &= (\partial_\nu A^c_\mu - \partial_\mu A^c_\nu - g f^{abc} A^a_\nu A^b_\mu)\tau^c_F = \\
&= (\partial_\nu A^c_\mu - \partial_\mu A^c_\nu + g f^{abc} A^a_\mu A^b_\nu)\tau^c_F \equiv \\
&\equiv F^c_{\mu\nu}\tau^c_F \equiv F_{\mu\nu}
\end{aligned}
\tag{2.43}
$$

with the generalized Maxwell's field-strength tensor, where we have actually introduced a family of (N^2-1) Maxwell's field-strength tensors. Let us note that, since the structure constants of a $SU(N)$ are antisymmetric in the indices a, b, c, the resulting quantity $F^c_{\mu\nu}\tau^c_F$ is antisymmetric in μ, ν as it was the case for the Maxwell's tensor of QED, Eq. 2.7. In a similar fashion to what it was done for the generalized covariant derivative, it is possible to explicitly compute how the generalized Maxwell's tensor transforms under a $SU(N)$ rotation. Similar (but longer) calculations lead to

$$
F_{\mu\nu} \to U_\omega(x) F_{\mu\nu} U^{-1}_\omega,
\tag{2.44}
$$

that is, the generalized field-strength tensor transforms in an homogeneous way.

With all the generalized objects in hand, we can now write the Lagrangian for a non-abelian theory, which was historically first obtained by C. Yang and R. Mills [38] in order to describe the isospin conservation. According to the general requirements of a local and gauge invariant Lagrangian, we are led to the form

$$
\mathcal{L}_{\text{Y-M}} = -\frac{1}{2}\text{tr}\left[F_{\mu\nu}F^{\mu\nu}\right] + \overline{\psi}(i\gamma^\mu D_\mu - M)\psi.
\tag{2.45}
$$

In fact, the spinor part of the Lagrangian is manifestly invariant given the transformation properties of $\overline{\psi}$ and $D_\mu\psi$, while the vector part can be shown to be gauge invariant by making use of the cyclic property of the trace:

$$
\begin{aligned}
-\frac{1}{2}\text{tr}\left[F_{\mu\nu}F^{\mu\nu}\right] &\to -\frac{1}{2}\text{tr}\left[U_\omega(x)F_{\mu\nu}U^{-1}_\omega(x)U_\omega(x)F^{\mu\nu}U^{-1}_\omega(x)\right] = \\
&= -\frac{1}{2}\text{tr}\left[U_\omega(x)F_{\mu\nu}(x)F^{\mu\nu}U^{-1}_\omega(x)\right] = \\
&= -\frac{1}{2}\text{tr}\left[U^{-1}_\omega(x)U_\omega(x)F_{\mu\nu}F^{\mu\nu}\right] = \\
&= -\frac{1}{2}\text{tr}\left[F_{\mu\nu}F^{\mu\nu}\right].
\end{aligned}
\tag{2.46}
$$

By making use of Eqs. 2.37 and 2.43 in 2.45, and after some proper manipulation, it can be shown that the Yang-Mills Lagrangian can be split into a free quadratic part and a cubic and fourth-order interaction Lagrangian, *i.e.*

$$\begin{aligned}
\mathcal{L}_{\mathrm{YM}} &= \mathcal{L}_{\mathrm{free}} + \mathcal{L}_{\mathrm{int}} \\
\mathcal{L}_{\mathrm{free}} &= -\frac{1}{2} g^{\mu\rho} g^{\nu\sigma} \left(\partial_\mu A^a_\nu - \partial_\nu A^a_\mu \right) \partial_\rho A^a_\sigma + \\
&\quad + \bar{\psi}(i\gamma^\mu \partial_\mu - M)\psi \\
\mathcal{L}_{\mathrm{int}} &= g\bar{\psi}\gamma^\mu \tau^a_F \psi A^a_\mu + \\
&\quad - g f^{abc} g^{\mu\rho} g^{\nu\sigma} A^b_\rho A^c_\sigma \partial_\mu A^a_\nu + \\
&\quad - \frac{1}{4} g^2 f^{abc} f^{ade} g^{\mu\rho} g^{\nu\sigma} A^b_\mu A^c_\nu A^d_\rho A^e_\sigma,
\end{aligned} \tag{2.47}$$

where tensorial indices of the vector fields have been raised and lowered by making use of the metric tensor $g_{\alpha\beta}$ for the ease of reading. In the interaction Lagrangian we clearly see the presence of a non-abelian electromagnetic current $\bar{\psi}(x)\gamma^\mu \tau^a_F \psi(x) A^a_\mu(x)$, as a generalization of Eq. 2.30, plus interaction terms involving the vector fields. These addictional terms arise from the non-abelian nature of the $SU(N)$ group and can be interpreted as interaction terms between the bosons of the theory.

As we shall point out in the following, non-abelian gauge theories are the cornerstones of the Standard Model of the electroweak and strong interactions. According to this model, the gauge symmetry group is the unitary group $SU(3) \otimes SU(2) \otimes U(1)$ of dimension $8 + 3 + 1 = 12$. Thus, the fermion multiplets which undergo strong interactions will posses three degrees of freedom named colors (red, green and blue) in eight possible combinations, associated to the $SU(3)$ symmetry governing the strong interactions, while the $SU(2)$ spinor multiplets do exhibit two weak isospin degrees of freedom named flavors, that will be attached to the weak interactions together with the $U(1)$ hypercharge that will be related to the electromagnetic interactions.

2.4 The Brout–Englert–Higgs mechanism

In the past sections we have fully explored the connection which exists between symmetries of the Lagrangian and conserved quantities, and we found that the requirement of local gauge invariance can be used as a dynamical principle to construct interacting theories. Although some degree of mathematical elegance has been achieved, such theories are still unsatisfactory, because the gauge invariance principle led us to interactions which are mediated by massless vector bosons (see, *e.g.*, Eqs. 2.31 and 2.45), which is in contrast with the experimental evidence concerning weak interactions.

In this section, we will focus on the difference existing between exact and approximate symmetries of the Lagrangian. We shall see that there are cases in which the Lagrangian is symmetric with respect to some transformation, while the physical vacuum state, namely the state which corresponds to the minimum potential energy, is not. In this case, the symmetry of the

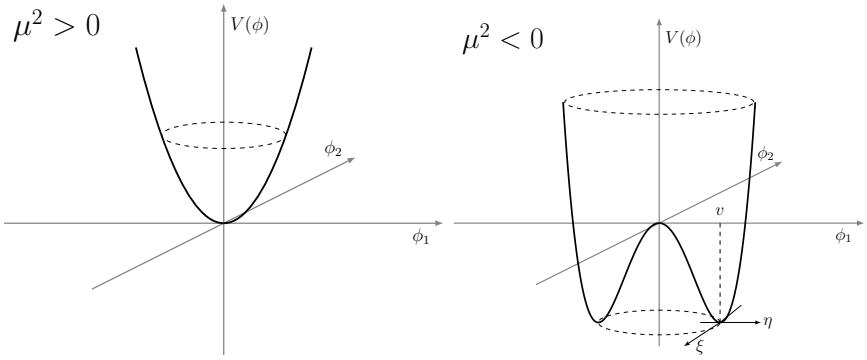

Figure 2.1: Effective potential for the Lagrangian of two scalar fields ϕ_1 and ϕ_2. If the parameter μ^2 is positive (left), the minimum is non degenerate and we recover the usual Lagrangian; if the parameter μ^2 is negative (right), the potential assumes the so called "Mexican hat" shape, with a degenerate set of minima lying on a circumference of radius v. In this latter case, the vacuum in not invariant under the action of the $SO(2)$ symmetry group and spontaneous symmetry breaking occurs.

Lagrangian is said to be spontaneously broken and some very interesting effects come to light.

2.4.1 Spontaneus symmetry breaking of a global symmetry

Let us first describe the spontaneous symmetry breaking of a continuous, global symmetry by inspecting a model, first introduced by J. Goldstone [39], which is based on the Lagrangian of two scalar fields ϕ_1 and ϕ_2,

$$\mathcal{L} = \frac{1}{2}\left(\partial_\mu\phi_1\partial^\mu\phi_1 + \partial_\mu\phi_2\partial^\mu\phi_2\right) - V(\phi_1^2 + \phi_2^2), \tag{2.48}$$

written as a kinetic term plus an effective potential V. Such Lagrangian is invariant under the action of the $SO(2)$ group of rotations in the euclidean plane

$$\boldsymbol{\phi} \equiv \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} \rightarrow \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}, \tag{2.49}$$

and, in order to investigate the nature of the vacuum state, we define the effective potential as

$$V(\phi^2) = \frac{1}{2}\mu^2\phi^2 + \frac{1}{4}|\lambda|(\phi^2)^2, \tag{2.50}$$

where $\phi^2 = \phi_1^2 + \phi_2^2$. We can now distinguish two cases, depending on the sign of μ^2.

First, we shall examine the case $\mu^2 > 0$. A positive value for the parameter μ^2 corresponds to the ordinary case of an exact symmetry. The potential energy has a single global minimum, as we can see in Fig. 2.1, and the value of the fields corresponding to the minimum is

$$\langle \boldsymbol{\phi} \rangle = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{2.51}$$

which is invariant under the $SO(2)$ rotation. Thus, approximating the Lagrangian for small oscillations, we get

$$\mathcal{L}_{\text{S.O.}} = \frac{1}{2}(\partial_\mu \phi_1 \partial^\mu \phi_1 - \mu^2 \phi_1^2) + \frac{1}{2}(\partial_\mu \phi_2 \partial^\mu \phi_2 - \mu^2 \phi_2^2), \tag{2.52}$$

which is in every way equivalent to the Lagrangian in Eq. 2.17.

We shall now have a look at the $\mu^2 < 0$ case. This choice of the parameter value indeed leads to a spontaneous symmetry breaking. In fact, the absolute minimum of the potential, as we can see from Fig 2.1, is now degenerate, being achieved for all the points of the (ϕ_1, ϕ_2) plane laying on the circumference of equation

$$\phi_1^2 + \phi_2^2 = \frac{-\mu^2}{|\lambda|} \equiv v^2. \tag{2.53}$$

We can now choose a point belonging to this circumference as the vacuum state, for example

$$\langle \boldsymbol{\phi} \rangle = \begin{pmatrix} v \\ 0 \end{pmatrix}, \tag{2.54}$$

which, note, is not invariant under the $SO(2)$ rotation, and expand about the vacuum configuration by a suitable change of coordinates,

$$\boldsymbol{\phi} - \langle \boldsymbol{\phi} \rangle \equiv \begin{pmatrix} \eta \\ \xi \end{pmatrix}, \tag{2.55}$$

which means

$$\begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} = \begin{pmatrix} \eta + v \\ \xi \end{pmatrix}. \tag{2.56}$$

Trivial calculation yield to the following Lagrangian for small oscillations:

$$\mathcal{L}_{\text{S.O.}} = \frac{1}{2}(\partial_\mu \eta \partial^\mu \eta + 2\mu^2 \eta^2) + \frac{1}{2}(\partial_\mu \xi \partial^\mu \xi), \tag{2.57}$$

where higher order terms in the fields and irrelevant constants have been neglected. We see that two particles appeared in the spectrum. The η particle, which may be thought to be associated with radial oscillations,

has a squared mass of $-2\mu^2 > 0$, and thus behaves as any usual scalar we encountered so far. The ξ particle, which may be thought to be associated with angular oscillations, is instead massless. The mass of the η particle can be interpreted as the result of the restoring force of the potential against radial oscillations, while the zero mass of the ξ particle can be interpreted as a consequence of the $SO(2)$ invariance of the Lagrangian, meaning that there is no restoring force against angular oscillations. The appearance of massless, spin-zero particles is called the Goldstone phenomenon and such particles are usually referred to as Goldstone bosons. In the general case, one Goldstone boson will show up for each broken generator of the considered symmetry group. Even though, from the point of view of unobserved massless particles, the Goldstone phenomenon seems to double the trouble, adding massless, scalar particles to the massless, vector particles that arose from the gauge theories, we will see in the following that, if the broken symmetry is a local symmetry, an astonishing result is obtained. In fact, the interplay between the Goldstone bosons and the massless bosons coming from the gauge theory will endow the gauge bosons with a mass and remove the Goldstone bosons from the spectrum. This is known as the Brout–Englert–Higgs mechanism [40, 41].

2.4.2 Spontaneus symmetry breaking of a local symmetry

Let us now consider a locally invariant Lagrangian which gives rise to spontaneous symmetry breaking. A simple example can be the so called abelian Higgs model, which is just the locally gauge-invariant extension of the Goldstone model, namely a $U(1)$-invariant theory describing the electrodynamics of a couple of scalar fields. The Lagrangian is

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + (\mathcal{D}_\mu\phi)^*(\mathcal{D}^\mu\phi) - \mu^2\phi\phi^* - |\lambda|(\phi^*\phi)^2, \tag{2.58}$$

where

$$\phi = \frac{\phi_1 + i\phi_2}{\sqrt{2}} \tag{2.59}$$

is a complex scalar field and the usual definitions given in Eqs. 2.7 and 2.25 hold. This Lagrangian is invariant under both global $U(1)$ rotations

$$\phi \to e^{i\theta}\phi \tag{2.60}$$

and local gauge transformations

$$\begin{aligned} &\phi \to e^{iq\alpha(x)}\phi \\ &A_\mu \to A_\mu - \partial_\mu\alpha(x). \end{aligned} \tag{2.61}$$

As we discussed before, two cases, depending on the value of μ^2, are possible.

If $\mu^2 > 0$, the effective potential has a single minimum at $\phi = 0$ and the symmetry of the Lagrangian is an exact symmetry. The spectrum of particles is just the ordinary spectrum of scalar QED, with a massless photon A_μ and two charged scalar particles, ϕ and ϕ^*, sharing a common mass μ.

On the other hand, the case $\mu^2 < 0$ corresponds to a spontaneous symmetry breaking. The effective potential has a degenerate minimum consisting of the points in the plane $(\Re(\phi), \Im(\phi))$ belonging to the circumference of equation

$$|\phi|^2 = \frac{-\mu^2}{2\lambda} \equiv \frac{v^2}{2}. \tag{2.62}$$

To explore the particle spectrum, we shall choose a minimum, such as

$$\langle\phi\rangle = \frac{v}{\sqrt{2}} \tag{2.63}$$

and then shift the ϕ field by $\phi - \langle\phi\rangle$ to expand the Lagrangian in term of displacements from the vacuum. As we shall see, the most useful parametrization to achieve this task is

$$\phi = \frac{\phi_1 + i\phi_2}{\sqrt{2}} = \frac{1}{\sqrt{2}} e^{i\frac{\xi}{v}}(v + \eta) \simeq \frac{\eta + v + i\xi}{\sqrt{2}}, \tag{2.64}$$

where, in the last equality, we have shown that this parametrization is nothing but the one given by Eq. 2.56 that we used for the Goldstone model, where quadratic terms in the fields have been neglected. If we insert the parametrization for ϕ in the Lagrangian given by Eq. 2.58, we get:

$$\begin{aligned}\mathcal{L} = &\frac{1}{2}(\partial_\mu\eta\partial^\mu\eta + 2\mu^2\eta^2) + \frac{1}{2}\left(\frac{v+\eta}{v}\right)^2\left(\partial_\mu\xi\partial^\mu\xi + q^2v^2A_\mu A^\mu + 2qvA_\mu\partial^\mu\xi\right) + \\ &+ \frac{\mu^2}{v}\eta^3 + \frac{\mu^2}{4v^2}\eta^4.\end{aligned} \tag{2.65}$$

As we already know from the discussion of the Goldstone phenomenon, a Goldstone boson ξ has appeared in the spectrum. We also see a new feature, namely, the field A_μ has acquired a mass term, but is also weirdly coupled with the Goldstone boson in the term $2qvA_\mu\partial^\mu\xi$. However, this can be fixed by a proper choice of the gauge. In fact, we note that the terms in the Lagrangian involving A_μ and ξ can conveniently be rewritten as

$$q^2v^2\left(A_\mu + \frac{1}{qv}\partial_\mu\xi\right)\left(A^\mu + \frac{1}{qv}\partial^\mu\xi\right), \tag{2.66}$$

which evidently inspires the gauge transformation

$$A'_\mu = A_\mu + \frac{1}{qv}\partial_\mu\xi, \tag{2.67}$$

which corresponds to the phase rotation of the scalar field

$$\phi' = e^{-i\frac{\xi}{v}}\phi. \tag{2.68}$$

Note that the gauge transformation for the vector field A_μ is not changing the kinematic term $-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}$ in Eq. 2.58. Thus, we obtain the crucial result:

$$\begin{aligned}\mathcal{L} = &-\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}(\partial_\mu\eta\partial^\mu\eta + 2\mu^2\eta^2) + \\ &+ \frac{q^2v^2}{2}A'_\mu A'^\mu + \frac{q^2}{2}A'_\mu A'^\mu\eta^2 + q^2vA'_\mu A'^\mu\eta + \frac{\mu^2}{v}\eta^3 + \frac{\mu^2}{4v^2}\eta^4.\end{aligned} \tag{2.69}$$

In this gauge the spectrum of particles contains:

- a massive scalar field η with mass $\sqrt{-2\mu^2}$;
- a massive vector field A'_μ with mass qv;
- no ξ field.

By a wise choice of the gauge, which is usually called the unitary gauge or U-gauge, the unwanted Goldstone boson disappeared, was "gauged away". However, it was not simply cancelled, but it was rather absorbed by the scalar field which acquired an additional degree of freedom. In fact, before spontaneous symmetry breaking, the theory had two scalar fields ϕ and ϕ^* plus the two helicity states of the massless field A_μ, for a total of four degrees of freedom. After spontaneous symmetry breaking, we are left with one scalar particle η plus the three helicity states of the massive field A'_μ, leading again to four degrees of freedom. The η field appearing in the Lagrangian is known as the Higgs boson.

The result we just obtained is truly remarkable, since it suggests the possibility of constructing gauge theories (possibly non-abelian gauge theories) in which the interactions are mediated by massive bosons, in agreement with the experimental evidence. This will be briefly explored in the following Subsection.

2.4.3 Spontaneous symmetry breaking of a non-abelian symmetry

In order to explore the consequences of spontaneous symmetry breaking in non-abelian theories, we shall use, as example of a non-abelian symmetry, the $SU(2)$ group in its three-dimensional representation, which can act on scalar fields triplets of the form

$$\boldsymbol{\phi} = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}. \tag{2.70}$$

The generators of the three-dimensional representation of $SU(2)$ are

$$T_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}; \quad T_2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}; \quad T_3 = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}; \tag{2.71}$$

and satisfy the algebra

$$[T^i, T^j] = i\epsilon_{ijk}T^k \tag{2.72}$$

as a special case of Eq. 2.36, with the structure constants being the elements of the Levi-Civita antysimmetric tensor. Having in hand the definitions for the non-abelian covariant derivative and Maxwell's tensor given by Eqs. 2.37 and 2.43, where now $a \in \{1, 2, 3\}$, we can immediately write the Lagrangian as

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}(D_\mu\boldsymbol{\phi})^\dagger(D^\mu\boldsymbol{\phi}) - V(\boldsymbol{\phi}\cdot\boldsymbol{\phi}), \tag{2.73}$$

where a generalized effective potential of the usual form has been introduced.

As we discussed in the previous examples, we have a case in which a unique minimum in the effective potential is present, which is achieved when $\boldsymbol{\phi} = 0$ and corresponds to an ordinary Yang-Mills theory describing three massive scalar particles with mass μ and three massless gauge bosons A^a_μ. Thus, the number of degrees of freedom of such a theory is $3 \times 1 + 3 \times 2 = 9$. We also have the spontaneously broken case, in which the vacuum state can be choosen to be

$$\langle\boldsymbol{\phi}\rangle = \begin{pmatrix} 0 \\ 0 \\ v \end{pmatrix}. \tag{2.74}$$

In order to know how many of the three generators of $SU(2)$ are broken, let us recall that the vacuum is left invariant by a generator T_i if, using the exponential representation given by Eq. 2.35,

$$\exp\{igT_i\omega_i(x)\}\langle\boldsymbol{\phi}\rangle = \langle\boldsymbol{\phi}\rangle, \tag{2.75}$$

which for an infinitesimal transformation becomes

$$(1 + igT_i\omega_i(x))\langle\boldsymbol{\phi}\rangle = \langle\boldsymbol{\phi}\rangle, \tag{2.76}$$

in such a way that the condition for T_i for leaving the vacuum invariant is simply

$$T_i\langle\boldsymbol{\phi}\rangle = 0. \tag{2.77}$$

Using Eq. 2.77 it is immediate to verify that T_1 and T_2 are broken, while T_3 is not. Thus, we can act in a similar fashion to what we did in Eq. 2.64 and expand the Lagrangian about the minimum using

$$\boldsymbol{\phi} = \exp\left\{\frac{i}{v}(\xi_1 T_1 + \xi_2 T_2)\right\}\begin{pmatrix} 0 \\ 0 \\ (v+\eta)/\sqrt{2} \end{pmatrix} \equiv U_\xi^{-1}(x)\begin{pmatrix} 0 \\ 0 \\ (v+\eta)/\sqrt{2} \end{pmatrix}, \tag{2.78}$$

where we introduced the two Goldstone bosons ξ_1 and ξ_2 corresponding to the two broken generators of $SU(2)$. Following analogous steps to the ones of the abelian theory, we move to the U-gauge by letting

$$A'_\mu = U_\xi(x) A_\mu U_\xi^{-1}(x) + \frac{i}{g}\left(\partial_\mu U_\xi(x)\right) U_\xi^{-1}(x). \tag{2.79}$$

Inserting Eqs. 2.78 and 2.79 in the Lagrangian, we obtain:

$$\begin{aligned}\mathcal{L} = &-\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}(\partial_\mu\eta\partial^\mu\eta + 2\mu^2\eta^2) + \\ &+ \frac{g^2(v+\eta)^2}{2}(A^1_\mu A^{1\mu} + A^2_\mu A^{2\mu}) + \frac{\mu^2}{v}\eta^3 + \frac{\mu^2}{4v^2}\eta^4,\end{aligned} \tag{2.80}$$

from which we see that:

- the η field has become a massive Higgs scalar with mass $\sqrt{-2\mu^2}$;
- the gauge bosons A^1_μ and A^2_μ, corresponding to the broken generators T_1 and T_2, acquired a common mass gv;
- the gauge boson A^3_μ, corresponding to the unbroken generator T_3, remains massless;
- the Goldstone bosons ξ_1 and ξ_2 are gauged away;
- after the spontaneous symmetry breaking, the number of degrees of freedom is still $1 \times 1 + 2 \times 3 + 1 \times 2 = 9$.

Having all these powerful tools in our hands, we are now ready to introduce the basic elements that give rise to the Standard Model of the electroweak and strong interactions.

2.5 Electroweak interactions of leptons

In order to describe the Standard Model of particle physics, let us begin by designating the spectrum of the fundamental leptons which enter the theory. We start by introducing the leptons belonging to the first generation, namely

the electron and its neutrino, which form a left-handed weak isospin doublet and are described by the two, four-dimensional Dirac bispinors e and ν, *i.e.*

$$\mathbf{L} \equiv \begin{pmatrix} \nu_e \\ e \end{pmatrix}_{\mathrm{L}} \tag{2.81}$$

where the left-handed states, obtained by making use of the well known chirality projectors, are

$$\begin{aligned} \nu_{\mathrm{L}} &= \frac{1}{2}\left(\mathbb{1} - \gamma_5\right)\nu \\ e_{\mathrm{L}} &= \frac{1}{2}\left(\mathbb{1} - \gamma_5\right)e. \end{aligned} \tag{2.82}$$

Neutrinos are known to have a very small mass [42], nevertheless it is convenient to assume them to be exactly massless, in which case the right-handed state

$$\nu_{\mathrm{R}} = \frac{1}{2}\left(\mathbb{1} + \gamma_5\right)\nu = 0 \tag{2.83}$$

does not exist. Thus, we can identify the only right-handed fermion with the right-handed electron,

$$\mathbf{R} \equiv e_{\mathrm{R}} = \frac{1}{2}\left(\mathbb{1} + \gamma_5\right)e, \tag{2.84}$$

which is a weak-isospin singlet. Note that the different nature of $\mathbf{L}$ (a weak isospin doublet) and $\mathbf{R}$ (a weak isospin singlet) implies different transformations for the two, that is:

$$\begin{aligned} &\mathbf{L} \to \exp\left\{ig\sigma^a\omega^a(x)\right\}\mathbf{L} \\ &\mathbf{R} \to \mathbf{R}, \end{aligned} \tag{2.85}$$

where σ^a are the infinitesimal generators of the two-dimensional representation of $SU(2)$, namely the well known Pauli matrices of Eq. 2.38. We see that the weak isospin doublet is transformed in the usual non-abelian way given by Eq. 2.34, while the weak isospin singlet is left unchanged.

In order to include electromagnetism in the framework, let us first introduce a weak hypercharge Y. Assuming the validity of the Gell-Mann–Nishijima relation for the electric charge, which relates the weak-isospin projection I_3 and the weak hypercharge Y,

$$Q = I_3 + \frac{1}{2}Y, \tag{2.86}$$

we can derive values for the left and right weak hypercharges. In fact, ν_{L} and e_{L} belong to a weak isospin doublet, and thus we have $I = 1/2$, $I_3^{\nu_{\mathrm{L}}} = 1/2$ and $I_3^{e_{\mathrm{L}}} = -1/2$, which, using Eq. 2.86, leads to a value of weak hypercharge equal to -1 for both the left-handed electron and neutrino; on the other

hand, the right-handed electron is a weak isospin singlet, from which it follows that $I = 0$, $I_3^{e_\mathrm{R}} = 0$ and thus the weak hypercharge is equal to -2. In summary, the hypercharge operator acts as:

$$\begin{aligned} Y\mathbf{L} &= -1\mathbf{L} \\ Y\mathbf{R} &= -2\mathbf{R}. \end{aligned} \tag{2.87}$$

Now, we take the group of transformations generated by I and Y to be the gauge group $SU(2)_L \otimes U(1)_Y$ of a non-abelian gauge theory. To build up the theory, we first introduce the four gauge bosons

$$\begin{aligned} b_\mu^a(x) \quad & \text{for } SU(2)_L, \\ A_\mu(x) \quad & \text{for } U(1)_Y. \end{aligned} \tag{2.88}$$

where $a \in \{1, 2, 3\}$. Then, we write the Lagrangian of the theory as

$$\mathcal{L} = \mathcal{L}_\text{gauge} + \mathcal{L}_\text{leptons} \tag{2.89}$$

where the first term involves the gauge bosons, while the second involves the leptons. The first term can be written in the usual form

$$\begin{aligned} \mathcal{L}_\text{gauge} &= -\frac{1}{2}\mathrm{tr}\left[F_{\mu\nu}F^{\mu\nu}\right] - \frac{1}{4}f_{\mu\nu}f^{\mu\nu} = \\ &= -\frac{1}{4}F^a_{\mu\nu}F^{a\mu\nu} - \frac{1}{4}f_{\mu\nu}f^{\mu\nu} \end{aligned} \tag{2.90}$$

where we used the relation $\mathrm{tr}[\sigma^a, \sigma^b] = \frac{1}{2}\delta^{ab}$ and the Maxwell's field-strength tensors, as we have seen previously, are

$$\begin{aligned} F^l_{\mu\nu} = \partial_\nu b^l_\mu - \partial_\mu b^l_\nu + g\varepsilon_{jkl}b^j_\mu b^k_\nu \quad & \text{for } SU(2)_L, \\ f_{\mu\nu} = \partial_\nu A_\mu - \partial_\mu A_\nu \quad & \text{for } U(1)_Y. \end{aligned} \tag{2.91}$$

The form of the second term is driven by the experimental evidence. In fact, the charged vector bosons of the weak interaction are found to couple only with the left-handed components of the spinor fields, contrary to the photon that couples with both left-handed and right-handed spinors. Eventually, we shall write the second term as:

$$\mathcal{L}_\text{leptons} = \overline{\mathbf{R}}i\gamma^\mu\left(\partial_\mu + \frac{ig'}{2}A_\mu Y\right)\mathbf{R} + \bar{\mathbf{L}}i\gamma^\mu\left(\partial_\mu + \frac{ig'}{2}A_\mu Y + \frac{ig}{2}\sigma^a b^a_\mu\right)\mathbf{L}. \tag{2.92}$$

The coupling constant of the $SU(2)$ group is called g and the coupling constant of the $U(1)$ group is called g', with some $1/2$ factors added in the Lagrangian to simplify calculations later on.

Even though it was inspired by the experimental evidence, the Lagrangian we just wrote in unsatisfactory for two main reasons: first, it describes

four massless gauge bosons $b^1_\mu, b^2_\mu, b^3_\mu, A_\mu$, in contrast with the observation of only one massless gauge boson, namely the photon; second, note that the different transformations for $\mathbf{L}$ and $\mathbf{R}$ forbid to include a mass term for the electron, which would result to be non-invariant under a $SU(2)$ transformation. However, what seems to be a troublesome bottleneck can be avoided by exploiting the spontaneous symmetry breaking of the theory. In order to see this explicitly, let us introduce in the theory a complex doublet of scalar fields

$$\boldsymbol{\phi} = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}, \tag{2.93}$$

which transforms as an $SU(2)$ doublet and thus has a weak hypercharge $T_{\boldsymbol{\phi}} = +1$, given by the Gell-Mann–Nishijima relation. We then add to the Lagrangian the term

$$\mathcal{L}_{\text{scalar}} = (D_\mu \boldsymbol{\phi})^\dagger (D_\mu \boldsymbol{\phi}) - \mu^2 (\boldsymbol{\phi}^\dagger \boldsymbol{\phi}) - |\lambda| (\boldsymbol{\phi}^\dagger \boldsymbol{\phi})^2, \tag{2.94}$$

where the covariant derivative is

$$D_\mu = \partial_\mu + \frac{ig'}{2} A_\mu Y + \frac{ig}{2} \sigma^a b^a_\mu, \tag{2.95}$$

as in Eq. 2.92. We also add an interaction term, which is taken to be the most general Yukawa term involving scalars and fermions which is invariant under the action of $SU(2)_L \otimes U(1)_Y$, namely

$$\mathcal{L}_{\text{Yukawa}} = -G_e \left[\overline{\mathbf{R}} (\boldsymbol{\phi}^\dagger \mathbf{L}) + (\overline{\mathbf{L}} \boldsymbol{\phi}) \mathbf{R} \right], \tag{2.96}$$

where G_e is a coupling constant. Let us now compute the consequences of spontaneous symmetry breaking. We assume $\mu^2 < 0$ and choose as the vacuum state the value of the scalar field

$$\langle \boldsymbol{\phi} \rangle = \begin{pmatrix} 0 \\ \frac{v}{\sqrt{2}} \end{pmatrix}, \tag{2.97}$$

which yields to very promising results. In fact, we readily see that all the generators of $SU(2)_L \otimes U(1)_Y$ are individually broken:

$$\begin{aligned}
\sigma^1 \langle \boldsymbol{\phi} \rangle &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ \frac{v}{\sqrt{2}} \end{pmatrix} \neq 0 \\
\sigma^2 \langle \boldsymbol{\phi} \rangle &= \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \begin{pmatrix} 0 \\ \frac{v}{\sqrt{2}} \end{pmatrix} \neq 0 \\
\sigma^3 \langle \boldsymbol{\phi} \rangle &= \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 0 \\ \frac{v}{\sqrt{2}} \end{pmatrix} \neq 0 \\
Y \langle \boldsymbol{\phi} \rangle &= +1 \langle \boldsymbol{\phi} \rangle \neq 0,
\end{aligned} \tag{2.98}$$

while the linear combination corresponding to the electric charge is not:

$$Q\langle\boldsymbol{\phi}\rangle = \frac{1}{2}(\sigma^3 + Y)\langle\boldsymbol{\phi}\rangle = 0. \tag{2.99}$$

This way, we will hopefully obtain that the photon will remain massless, while the three remaining bosons will acquire a mass. As usual, we expand the Lagrangian about the minimum of the effective potential by writing

$$\boldsymbol{\phi} = \exp\left\{\frac{i\xi^a\sigma^a}{2v}\right\}\begin{pmatrix} 0 \\ (v+\eta)/\sqrt{2}\end{pmatrix} \equiv U_\xi^{-1}(x)\begin{pmatrix} 0 \\ (v+\eta)/\sqrt{2}\end{pmatrix} \tag{2.100}$$

and transform to the U-gauge:

$$\sigma^a b_\mu^a \equiv b_\mu \to b'_\mu = U_\xi(x) b_\mu U_\xi^{-1}(x) + \frac{i}{g}\left[\partial_\mu U_\xi(x)\right] U_\xi^{-1}(x), \tag{2.101}$$

$$A_\mu \to A_\mu, \tag{2.102}$$

$$\mathbf{L} \to \mathbf{L}' = U_\xi(x)\mathbf{L}, \tag{2.103}$$

$$\mathbf{R} \to \mathbf{R}. \tag{2.104}$$

We can now express the Lagrangian in terms of the U-gauge fields and investigate the consequences of spontaneous symmetry breaking.

Let us start with a closer look to the Yukawa term. We insert Eqs. 2.100 and 2.103 in Eq. 2.96 and obtain:

$$\begin{aligned}\mathcal{L}_{\text{Yukawa}} = -G_e\Big\{\overline{\mathbf{R}}\left[\begin{pmatrix} 0 & \frac{v+\eta}{\sqrt{2}}\end{pmatrix} U_\xi(x)U_\xi^{-1}(x)\mathbf{L}'\right] + \\ + \left[\left(U_\xi^{-1}(x)\mathbf{L}'\right)^\dagger \gamma_0 U_\xi^{-1}(x)\begin{pmatrix} 0 \\ \frac{v+\eta}{\sqrt{2}}\end{pmatrix}\right]\mathbf{R}\Big\},\end{aligned} \tag{2.105}$$

which, after some trivial calculations becomes

$$\begin{aligned}\mathcal{L}_{\text{Yukawa}} &= -\frac{G_e(v+\eta)}{\sqrt{2}}\left\{\overline{e_R}e_L + \overline{e_L}e_R\right\} = -\frac{G_e(v+\eta)}{\sqrt{2}}\bar{e}e = \\ &= -\frac{G_e v}{\sqrt{2}}\bar{e}e - \frac{G_e}{\sqrt{2}}\bar{e}e\eta,\end{aligned} \tag{2.106}$$

in such a way that the electron has clearly acquired a mass

$$m_e = \frac{G_e v}{\sqrt{2}}, \tag{2.107}$$

and an interaction term between electrons and the Higgs boson has shown up, where the coupling is found to be proportional to the fermion mass.

As a second step, we shall inspect the scalar term. By making use of Eqs. 2.100 and 2.101, $D_\mu\phi$ assumes the form:

$$D_\mu\phi = U_\xi^{-1}(x)\left[\begin{pmatrix} 0 \\ \frac{\partial_\mu\eta}{\sqrt{2}} \end{pmatrix} + \frac{ig'}{2}A_\mu \begin{pmatrix} 0 \\ \frac{v+\eta}{\sqrt{2}} \end{pmatrix} + igb'_\mu \begin{pmatrix} 0 \\ \frac{v+\eta}{\sqrt{2}} \end{pmatrix}\right], \quad (2.108)$$

which shows that the Goldstone-like scalar fields $\xi^a(x)$ can be factorized into a unitary matrix and eventually decoupled from the other fields thanks to the U-gauge. The form of the adjoint term $(D_\mu\phi)^\dagger$ and of the terms $-\mu^2(\boldsymbol{\phi}^\dagger\boldsymbol{\phi})$ and $-|\lambda|(\boldsymbol{\phi}^\dagger\boldsymbol{\phi})^2$ is easily computed, so that we obtain:

$$\begin{aligned}\mathcal{L}_{\text{scalar}} = &\frac{1}{2}\left(\partial_\mu\eta\partial^\mu\eta + 2\mu^2\eta^2\right) + \\ &+ \frac{(v+\eta)^2}{8}\left[g^2(b^1_\mu b^{1\mu} + b^2_\mu b^{2\mu}) + \left(g'A_\mu - gb^3_\mu\right)\left(g'A^\mu - gb^{3\mu}\right)\right] + \\ &+ \frac{\mu^2}{v}\eta^3 + \frac{\mu^2}{4v^2}\eta^4,\end{aligned} \quad (2.109)$$

where we omitted primes to avoid notational clutter. We see that the η field has acquired a mass $m_{\text{H}} = \sqrt{-2\mu^2}$. Also, if we define the charged gauge fields to be

$$W^\pm_\mu = \frac{b^1_\mu \mp ib^2_\mu}{\sqrt{2}}, \quad (2.110)$$

we can rearrange the term proportional to b^1_μ and b^2_μ as:

$$\begin{aligned}\frac{g^2v^2}{8}(b^1_\mu b^{1\mu} + b^2_\mu b^{2\mu}) &= \frac{g^2v^2}{16}(b^1_\mu b^{1\mu} + b^2_\mu b^{2\mu} + b^1_\mu b^{1\mu} + b^2_\mu b^{2\mu}) = \\ &= \frac{g^2v^2}{16}\left[2(W^+_\mu)^\dagger W^\mu_+ + 2(W^-_\mu)^\dagger W^\mu_-\right] = \\ &= \frac{g^2v^2}{8}\left[(W^+_\mu)^\dagger W^\mu_+ + (W^-_\mu)^\dagger W^\mu_-\right],\end{aligned} \quad (2.111)$$

which is a mass term for the charged gauge bosons corresponding to the mass

$$m_{W^\pm} = \frac{gv}{2}. \quad (2.112)$$

Also, defining the combinations

$$Z_\mu = \frac{-g'A_\mu + gb^3_\mu}{\sqrt{g^2+g'^2}} \quad (2.113)$$

$$\mathcal{A}_\mu = \frac{gA_\mu + g'b^3_\mu}{\sqrt{g^2+g'^2}} \quad (2.114)$$

we can rearrange the term proportional to A_μ and b^3_μ as:

$$\frac{v^2}{8}\left(g' A_\mu - g b^3_\mu\right)\left(g' A^\mu - g b^{3\mu}\right) = \frac{v^2(g^2+g'^2)}{8} Z_\mu Z^\mu, \tag{2.115}$$

which is a mass term for the neutral gauge boson corresponding to the mass

$$m_Z = \frac{v\sqrt{g^2+g'^2}}{2}. \tag{2.116}$$

Note that the neutral gauge boson $\mathcal{A}_\mu$, which is interpreted to be the photon, is not present in the scalar term and thus is massless, as it should be. Note also the presence of cubic and quartic self-interaction terms for the Higgs boson and of interaction terms between the gauge bosons and the Higgs boson.

The mixing between the A_μ and b^3_μ vector fields that gives rise to the neutral gauge bosons can be conveniently parametrized by means of a mixing angle, which was first introduced by S. Glashow [43], although it is often and curiously referred to as the Weinberg angle. That is:

$$\begin{aligned} Z_\mu &= -A_\mu \sin\theta_W + b^3_\mu \cos\theta_W \\ \mathcal{A}_\mu &= A_\mu \cos\theta_W + b^3_\mu \sin\theta_W, \end{aligned} \tag{2.117}$$

with

$$\sin\theta_W = \frac{g'}{\sqrt{g^2+g'^2}}, \quad \cos\theta_W = \frac{g}{\sqrt{g^2+g'^2}}, \quad \tan\theta_W = \frac{g'}{g}. \tag{2.118}$$

The use of the Weinberg angle makes it easier to interpret the term of the Lagrangian involving the leptons given by Eq. 2.92. By making use of Eqs. 2.104 and 2.103 we readily see that the part involving $\mathbf{R}$ is untouched by the change to the U-gauge, while the covariant derivative involving $\mathbf{L}$ assumes the form:

$$D_\mu \mathbf{L}' = U^{-1}_\xi(x)\left[\partial_\mu \mathbf{L}' - \frac{ig'}{2} A_\mu \mathbf{L}' + \frac{ig}{2} b'_\mu \mathbf{L}'\right], \tag{2.119}$$

which once again shows that the Goldstone-like scalar fields $\xi^a(x)$ can be factorized into a unitary matrix and decoupled from the other fields thanks to the U-gauge. After some manipulation, the leptonic term in the Lagrangian becomes

$$\begin{aligned}\mathcal{L}_{\text{leptons}} = {} & \overline{\nu_L} i\gamma^\mu \partial_\mu \nu_L + \bar{e} i\gamma^\mu \partial_\mu e + \\ & - \frac{g}{\sqrt{2}} \left(W^+_\mu \overline{\nu_L} \gamma^\mu e_L + W^-_\mu \overline{e_L} \gamma^\mu \nu_L \right) - \frac{\sqrt{g^2 + g'^2}}{2} Z_\mu \overline{\nu_L} \gamma^\mu \nu_L + \\ & + \frac{Z_\mu}{\sqrt{g^2 + g'^2}} \left(-g'^2 \overline{e_R} \gamma^\mu e_R + \frac{g^2 - g'^2}{2} \overline{e_L} \gamma^\mu e_L \right) + \frac{g g'}{\sqrt{g^2 + g'^2}} \mathcal{A}_\mu \bar{e} \gamma^\mu e,\end{aligned} \tag{2.120}$$

where now are clearly present: two kinetic terms for electrons and neutrinos respectively; interaction terms between the charged bosons and the left-handed leptons; interaction terms between the Z boson and both left- and right-handed leptons; an interaction term between the photon and left- and right-handed electrons. The terms contracted with the gauge bosons can be interpreted as weak and electromagnetic currents $J^\mu_\pm$, J^μ_0, J^μ, provided that we identify the electric charge of the electron with

$$e = \frac{g g'}{\sqrt{g^2 + g'^2}}. \tag{2.121}$$

Note that, under this crucial identification, the coupling constants g and g', belonging to $SU(2)_L$ and $U(1)_Y$ respectively, are both found to be proportional to the electric charge *via* the Weinberg angle, that is:

$$\begin{aligned} g &= \frac{e}{\sin\theta_W} \\ g' &= \frac{e}{\cos\theta_W}, \end{aligned} \tag{2.122}$$

which shows that the different intensities of the electromagnetic and weak interactions are now related by a single parameter and are thus now unified in a single interaction, called the electroweak interaction.

Finally, we shall compute the form of the term involving the gauge bosons. This will give rise to kinetic terms for the bosons of the theory, as well as interaction terms between bosons, as already pointed out in Eq. 2.47. In fact, if we define the abelian-like field-strength tensors for the gauge bosons as

$$\begin{aligned} \Gamma_{\mu\nu} &= \partial_\nu \mathcal{A}_\mu - \partial_\mu \mathcal{A}_\nu \\ W^\pm_{\mu\nu} &= \partial_\nu W^\pm_\mu - \partial_\mu W^\pm_\nu \\ Z_{\mu\nu} &= \partial_\nu Z_\mu - \partial_\mu Z_\nu, \end{aligned} \tag{2.123}$$

we obtain, after some degree of manipulation, the following expression:

$$\begin{aligned}\mathcal{L}_{\text{gauge}} &= -\frac{1}{4}\Gamma_{\mu\nu}\Gamma^{\mu\nu} - \frac{1}{2}W^{+}_{\mu\nu}W^{\mu\nu}_{-} - \frac{1}{4}Z_{\mu\nu}Z^{\mu\nu} + \\ &- i\left[W^{\nu}_{-}W^{+}_{\mu\nu} - W^{\nu}_{+}W^{-}_{\mu\nu}\right]\left[e\mathcal{A}^{\mu} + g\cos\theta_W Z^{\mu}\right] \\ &- i\left[e\Gamma^{\mu\nu} + g\cos\theta_W Z^{\mu\nu}\right]W^{-}_{\mu}W^{+}_{\nu} \\ &- W^{+}_{\mu}W^{\mu}_{-}\left[e\mathcal{A} + g\cos\theta_W Z\right]^2 \\ &+ W^{+}_{\mu}W^{-}_{\nu}\left[e\mathcal{A}^{\mu} + g\cos\theta_W Z^{\mu}\right]\left[e\mathcal{A}^{\nu} + g\cos\theta_W Z^{\nu}(x)\right] \\ &- \frac{1}{2}g^2\left[W^{+}_{\mu}W^{\mu}_{-}\right]^2 + \frac{1}{2}g^2\left[W^{+}_{\mu}W^{-}_{\nu}\right]\left[W^{\mu}_{+}W^{\nu}_{-}\right]\end{aligned} \tag{2.124}$$

where the complex nature of the interaction between gauge bosons in a non-abelian theory has been made explicit. In fact, we see that the spontaneous symmetry breaking produces triple and quartic gauge interactions involving gauge bosons, which effectively arise from many different combinations of contracted spacetime indices.

Finally, in order for the electroweak theory of leptons to be complete, we have to include the second and third generation leptons. This extension is actually pretty straightforward, since one can simply add new left-handed weak-isospin doublets

$$\mathbf{L}_{\mu} \equiv \begin{pmatrix} \nu_{\mu} \\ \mu \end{pmatrix}_{\text{L}} \qquad \mathbf{L}_{\tau} \equiv \begin{pmatrix} \nu_{\tau} \\ \tau \end{pmatrix}_{\text{L}}, \tag{2.125}$$

and right-handed singlets

$$\mathbf{R}_{\mu} \equiv \mu_{\text{R}} \qquad \mathbf{R}_{\tau} \equiv \tau_{\text{R}}, \tag{2.126}$$

in such a way that the "Yukawa-like" part of the Lagrangian given by Eq. 2.96 can readily be generalized as:

$$\mathcal{L}_{\text{Yukawa}} = \sum_{\iota=e,\mu,\tau} -G_{\iota}\left[\overline{\mathbf{R}}_{\iota}(\boldsymbol{\phi}^{\dagger}\mathbf{L}_{\iota}) + (\bar{\mathbf{L}}_{\iota}\boldsymbol{\phi})\mathbf{R}_{\iota})\right], \tag{2.127}$$

namely by simply adding Yukawa terms for the second and third generation fermions. In a similar fashion to what it was done for the electron, the spontaneous symmetry breaking mechanism applied to this term will also produce a mass term for the muon and the tau leptons, as well as interaction terms between these leptons and the Higgs boson. Similarly, the term of the Lagrangian involving the leptons given by Eq. 2.92 shall be generalized as:

$$\begin{aligned}\mathcal{L}_{\text{leptons}} = \sum_{\iota=e,\mu,\tau}\Bigg[&\overline{\mathbf{R}}_{\iota}i\gamma^{\mu}\left(\partial_{\mu} + \frac{ig'}{2}A_{\mu}Y\right)\mathbf{R}_{\iota} + \\ &+ \bar{\mathbf{L}}_{\iota}i\gamma^{\mu}\left(\partial_{\mu} + \frac{ig'}{2}A_{\mu}Y + \frac{ig}{2}\sigma^{a}b^{a}_{\mu}\right)\mathbf{L}_{\iota}\Bigg],\end{aligned} \tag{2.128}$$

which, after spontaneous symmetry breaking, gives rise to kinetic terms for the second and third generation fermions and interaction terms between the gauge bosons and the second and third generation fermions.

2.6 Electroweak interactions of quarks

The extension of the previously developed model to quarks is to some degree straightforward, even though it must be carried out with some care since the leptonic and hadronic sectors show a couple of important differences. First of all, we shall remark the experimental evidence which shows that all the quarks have a non-zero mass. Secondly, let us recall the fact that individual quark quantum numbers are not conserved by the weak interaction (*e.g.*, in the strangeness-violating decay $K^+ \to \pi^+\pi^0$).

To extend the model to account for the non-vanishing quark masses, we must introduce right-handed singlets for both members of each family. That is, an additional right-handed singlet will be present when quarks are taken into account, together with the left-handed doublet and right-handed singlet we have encountered so far. This can be written as:

$$\begin{aligned} \boldsymbol{q}_{L,\iota} &= \frac{1}{2}(\mathbb{1} - \gamma_5) \begin{pmatrix} u_\iota \\ d_\iota \end{pmatrix} \\ \boldsymbol{u}_{R,\iota} &= \frac{1}{2}(\mathbb{1} + \gamma_5) u_\iota \\ \boldsymbol{d}_{R,\iota} &= \frac{1}{2}(\mathbb{1} + \gamma_5) d_\iota, \end{aligned} \tag{2.129}$$

where the index $\iota = 1, 2, 3$ runs over the three families of quarks:

$$u_\iota = \{u, c, t\} \qquad d_\iota = \{d, s, b\}. \tag{2.130}$$

An additional color index should be added to the quark fields, which are sensitive to the strong interaction, but it will understood in the following for the sake of clarity. The left-handed doublet and the right-handed singlets that we just introduced transform in the usual way under the action of $SU(2)$, as given by Eq. 2.85. Also, given the well-known fractional electric charges of the quarks, the hypercharge of the quark doublet and singlets are readily found to be:

$$Y(\boldsymbol{q}_{L,\iota}) = \frac{1}{3} \qquad Y(\boldsymbol{u}_{R,\iota}) = \frac{4}{3} \qquad Y(\boldsymbol{d}_{R,\iota}) = -\frac{2}{3}. \tag{2.131}$$

The presence of an additional, right-handed singlet implies that, for a given generation, two distinct Yukawa terms can be built from the left-handed doublet and right-handed singlets. Such Yukawa terms would be of the form:

$$\begin{aligned} &- y_d \left[\overline{\boldsymbol{d}}_{R,\iota} (\boldsymbol{\phi}^\dagger \boldsymbol{q}_{L,\iota}) + (\overline{\boldsymbol{q}}_{L,\iota} \boldsymbol{\phi}) \boldsymbol{d}_{R,\iota} \right] + \\ &- y_u \left[\overline{\boldsymbol{u}}_{R,\iota} (\widetilde{\boldsymbol{\phi}}^\dagger \boldsymbol{q}_{L,\iota}) + (\overline{\boldsymbol{q}}_{L,\iota} \widetilde{\boldsymbol{\phi}}) \boldsymbol{u}_{R,\iota} \right], \end{aligned} \tag{2.132}$$

where an additional Higgs doublet $\widetilde{\boldsymbol{\phi}} = i\sigma_2 \boldsymbol{\phi}^*$ with hypercharge $Y(\widetilde{\boldsymbol{\phi}}) = -1$ has been introduced to construct the second Yukawa term, in an indentical fashion to what it was done for the leptonic sector. As a naive approach, one could be now tempted to build the Lagrangian for the quark fields by simply adding three terms such as Eq. 2.132 (one for each family) and also by introducing three terms involving covariant derivatives of the fields such as it was done in Eq. 2.92 (again, one for each family). Unfortunately, such model is experimentally inconsistent. In fact, it turns out to be incompatible with the observation made by N. Cabibbo [44] that the strange quark is actually mixed with the down quark forming a weak eigenstate that differs from the mass eigenstates. Indeed, incorporating the Cabibbo model in the naive description of quarks would imply, for example, the presence of flavor-changing neutral currents (FCNC), an hypothesis that has been strongly rejected by the experimental evidence.

Thus, the correct form for the Yukawa term involving the quark fields is found to be slightly more complicated, even though still quite familiar:

$$\begin{aligned} \mathcal{L}_{\text{Yukawa}} = -\sum_{\iota=1}^{3}\sum_{\jmath=1}^{3} \Big(& \overline{\boldsymbol{d}}_{R,\jmath} Y^{d*}_{\jmath\iota} (\boldsymbol{\phi}^\dagger \boldsymbol{q}_{L,\iota}) + (\overline{\boldsymbol{q}}_{L,\iota} \boldsymbol{\phi}) Y^{d}_{\iota\jmath} \boldsymbol{d}_{R,\jmath} + \\ & + \overline{\boldsymbol{u}}_{R,\jmath} Y^{u*}_{\jmath\iota} (\widetilde{\boldsymbol{\phi}}^\dagger \boldsymbol{q}_{L,\iota}) + (\overline{\boldsymbol{q}}_{L,\iota} \widetilde{\boldsymbol{\phi}}) Y^{u}_{\iota\jmath} \boldsymbol{u}_{R,\jmath} \Big), \end{aligned} \tag{2.133}$$

which is nothing more than a generalization of the naive Yukawa term, where now the weak eigenstates for the quarks are mixed through two complex matrices of elements $Y^u_{\iota\jmath}$ and $Y^d_{\iota\jmath}$. This term can be readily expressed in the U-gauge, by means of identical calculations to the ones of the leptonic sector, to give:

$$\begin{aligned} \mathcal{L}_{\text{Yukawa}} = -\sum_{\iota=1}^{3}\sum_{\jmath=1}^{3} \frac{v+\eta}{\sqrt{2}} \Big[& \overline{\boldsymbol{d}}_{R,\jmath} Y^{d*}_{\jmath\iota} \boldsymbol{d}_{L,\iota} + \overline{\boldsymbol{d}}_{L,\iota} Y^{d}_{\iota\jmath} \boldsymbol{d}_{R,\jmath} + \\ & \overline{\boldsymbol{u}}_{R,\jmath} Y^{u*}_{\jmath\iota} \boldsymbol{u}_{L,\iota} + \overline{\boldsymbol{u}}_{L,\iota} Y^{u}_{\iota\jmath} \boldsymbol{u}_{R,\jmath} \Big], \end{aligned} \tag{2.134}$$

where we see that the mass term coming from the spontaneously-broken Yukawa term is actually *non-diagonal in flavor*, suggesting that the weak eigenstates are not actually mass eigenstates, according to the experimental evidence. In fact, we can recover the usual form for the mass terms by noting that, in a similar fashion to what it is done with real, symmetric matrices, a generic complex matrix like $Y^u_{\iota\jmath}$ (or $Y^d_{\iota\jmath}$) can always be diagonalized by means

of two unitary matrices. This means that we can always introduce 4 unitary matrices $V_L^u, V_R^u, V_L^d, V_R^d$ in such a way that:

$$\begin{aligned} \frac{v}{\sqrt{2}} V_L^{d\dagger} Y^d V_R^d &= \mathrm{diag}(m_d, m_s, m_b) \\ \frac{v}{\sqrt{2}} V_L^{u\dagger} Y^u V_R^u &= \mathrm{diag}(m_u, m_c, m_t). \end{aligned} \tag{2.135}$$

Thus, if we define the weak eigenstates $\boldsymbol{u}_{L,\iota}, \boldsymbol{u}_{R,\iota}, \boldsymbol{d}_{L,\iota}, \boldsymbol{d}_{R,\iota}$ to be a mixing of the mass eigenstates $\boldsymbol{U}_{L,\iota}, \boldsymbol{U}_{R,\iota}, \boldsymbol{D}_{L,\iota}, \boldsymbol{D}_{R,\iota}$:

$$\begin{aligned} \boldsymbol{u}_L &= V_L^u \boldsymbol{U}_L \\ \boldsymbol{u}_R &= V_R^u \boldsymbol{U}_R \\ \boldsymbol{d}_L &= V_L^d \boldsymbol{D}_L \\ \boldsymbol{d}_R &= V_R^d \boldsymbol{D}_R, \end{aligned} \tag{2.136}$$

we can readily insert Eqs. 2.136 into Eq. 2.134 to obtain a mass term of the form

$$- \sum_{\iota=d,s,b} m_\iota \overline{\boldsymbol{D}}_\iota \boldsymbol{D}_\iota - \sum_{\jmath=u,c,t} m_\jmath \overline{\boldsymbol{U}}_\jmath \boldsymbol{U}_\jmath, \tag{2.137}$$

which appears to be an usual Dirac mass term for the spinor fields describing the mass eigenstates of the quarks.

We shall conclude this section with a last but crucial remark. The mixing between weak eigenstates has an important consequence on the charged current for the quarks. In fact, we can add a $\mathcal{L}_{\mathrm{quarks}}$ term in the Lagrangian, with exactly the same approach that we used for leptons (see Eq. 2.128), that is:

$$\begin{aligned} \mathcal{L}_{\mathrm{quarks}} = \sum_{\iota=1}^{3} \Bigg[& \overline{\boldsymbol{u}_{R,\iota}} i\gamma^\mu \left(\partial_\mu + \frac{ig'}{2} A_\mu Y \right) \boldsymbol{u}_{R,\iota} + \\ & + \overline{\boldsymbol{d}_{R,\iota}} i\gamma^\mu \left(\partial_\mu + \frac{ig'}{2} A_\mu Y \right) \boldsymbol{d}_{R,\iota} + \\ & + \overline{\boldsymbol{q}_{L,\iota}} i\gamma^\mu \left(\partial_\mu + \frac{ig'}{2} A_\mu Y + \frac{ig}{2} \sigma^a b_\mu^a \right) \boldsymbol{q}_{L,\iota} \Bigg], \end{aligned} \tag{2.138}$$

and transform this term to the U-gauge. If we do so, we obtain the kinetic terms for the quark mass eigenstates and we also see that the charged current for quarks becomes:

$$J_+^\mu = (J_-^\mu)^\dagger = -\frac{g}{\sqrt{2}} \sum_{\iota=1}^{3} \overline{\boldsymbol{U}_{L,\iota}} \gamma^\mu \left(V_L^{u\dagger} V_L^d \right)_{\iota\jmath} \boldsymbol{D}_{L,\jmath}, \tag{2.139}$$

where the quark mass eigenstates appear to be connected by the action of the famous Cabibbo–Kobayashi–Maskawa (CKM) matrix [45] $V_L^{u\dagger} V_L^d \equiv V$ which

describes the non-conservation of the flavor quantum number in the weak charged interactions. On the contrary, the neutral currents associated with the photon and the Z boson are found not to change the flavor of quarks, as they read:

$$\begin{aligned} J^{\mu} &= e\sum_{\iota=1}^{3}\left[-\frac{2}{3}\overline{\boldsymbol{u}}_{\iota}\gamma^{\mu}\boldsymbol{u}_{\iota}+\frac{1}{3}\overline{\boldsymbol{d}}_{\iota}\gamma^{\mu}\boldsymbol{d}_{\iota}\right] \\ J_0^{\mu} &= g\sec\theta_W\sum_{\iota=1}^{3}\Big\{-\frac{1}{2}\overline{\boldsymbol{u_L}}_{,\iota}\gamma^{\mu}\boldsymbol{u}_{L,\iota}+\frac{1}{2}\overline{\boldsymbol{d}}_{L,\iota}\gamma^{\mu}\boldsymbol{d}_{L,\iota}+ \\ &\quad+\sin^2\theta_W\left[\frac{2}{3}\overline{\boldsymbol{u}}_{\iota}\gamma^{\mu}\boldsymbol{u}_{\iota}-\frac{1}{3}\overline{\boldsymbol{d}}_{\iota}\gamma^{\mu}\boldsymbol{d}_{\iota}\right]\Big\}, \end{aligned} \tag{2.140}$$

in such a way that the absence of FCNC is granted.

2.7 Electroweak Lagrangian

We have now all the ingredients needed in order to write the complete Lagrangian for the electroweak interaction, which is internally consistent and coherent with the experimental evidence. It is simply a matter of rearranging the terms we have inspected so far to obtain the rather intriguing form:

$$\begin{aligned} \mathcal{L}_{\rm EW} &= -\frac{1}{4}\Gamma_{\mu\nu}\Gamma^{\mu\nu}-\frac{1}{2}W^{+}_{\mu\nu}W^{\mu\nu}_{-}-\frac{1}{4}Z_{\mu\nu}Z^{\mu\nu}+ \\ &\quad-i\left[W^{\nu}_{-}W^{+}_{\mu\nu}-W^{\nu}_{+}W^{-}_{\mu\nu}\right]\left[e\mathcal{A}^{\mu}+g\cos\theta_W Z^{\mu}\right]+ \\ &\quad-i\left[e\Gamma^{\mu\nu}+g\cos\theta_W Z^{\mu\nu}\right]W^{-}_{\mu}W^{+}_{\nu}+ \\ &\quad-W^{+}_{\mu}W^{\mu}_{-}\left[e\mathcal{A}+g\cos\theta_W Z\right]^2+ \\ &\quad+W^{+}_{\mu}W^{-}_{\nu}\left[e\mathcal{A}^{\mu}+g\cos\theta_W Z^{\mu}\right]\left[e\mathcal{A}^{\nu}+g\cos\theta_W Z^{\nu}(x)\right]+ \\ &\quad-\frac{1}{2}g^2\left[W^{+}_{\mu}W^{\mu}_{-}\right]^2+\frac{1}{2}g^2\left[W^{+}_{\mu}W^{-}_{\nu}\right]\left[W^{\mu}_{+}W^{\nu}_{-}\right]+ \\ &\quad+\frac{1}{2}\left(\partial_{\mu}\eta\partial^{\mu}\eta-m^2_{\rm H}\eta^2\right)+\frac{\mu^2}{v}\eta^3+\frac{\mu^2}{4v^2}\eta^4+ \\ &\quad+\frac{1}{2}\left[m^2_W(W^{+}_{\mu})^{\dagger}W^{\mu}_{+}+m^2_W(W^{-}_{\mu})^{\dagger}W^{\mu}_{-}+m^2_Z Z_{\mu}Z^{\mu}\right]\left(1+\frac{\eta}{v}\right)^2+ \\ &\quad+i\sum_{\iota=1}^{3}\left[\bar{\nu}_{L,\iota}\not{\partial}\nu_{L,\iota}+\bar{\ell}_{\iota}\not{\partial}\ell_{\iota}+\overline{\boldsymbol{U}}_{\iota}\not{\partial}\boldsymbol{U}_{\iota}+\overline{\boldsymbol{D}}_{\iota}\not{\partial}\boldsymbol{D}_{\iota}\right]+ \\ &\quad-\sum_{\iota=1}^{3}\left[m^{\ell}_{\iota}\bar{\ell}_{\iota}\ell_{\iota}+m^{u}_{\iota}\overline{\boldsymbol{U}}_{\iota}\boldsymbol{U}_{\iota}+m^{d}_{\iota}\overline{\boldsymbol{D}}_{\iota}\boldsymbol{D}_{\iota}\right]\left(1+\frac{\eta}{v}\right)+ \\ &\quad+\mathcal{A}_{\mu}J^{\mu}+Z_{\mu}J^{\mu}_0+W^{+}_{\mu}J^{\mu}_{+}+W^{-}_{\mu}J^{\mu}_{-}, \end{aligned} \tag{2.141}$$

where the Feynman slash notation $\not{\partial} \equiv \gamma^\mu \partial_\mu$ has been introduced and the currents are the sum of the leptonic and hadronic parts, namely:

$$
\begin{aligned}
J^\mu &= e \sum_{\iota=1}^{3} \left[\bar{\ell}_\iota \gamma^\mu \ell_\iota - \frac{2}{3} \overline{\boldsymbol{U}}_\iota \gamma^\mu \boldsymbol{U}_\iota + \frac{1}{3} \overline{\boldsymbol{D}}_\iota \gamma^\mu \boldsymbol{D}_\iota \right] \\
J_0^\mu &= \frac{1}{2} g \sec\theta_W \sum_{\iota=1}^{3} \Big\{ -\bar{\nu}_{L,\iota} \gamma^\mu \nu_{L,\iota} + \overline{\ell_{L,\iota}} \gamma^\mu \ell_{L,\iota} - \overline{\boldsymbol{U}_{L,\iota}} \gamma^\mu \boldsymbol{U}_{L,\iota} + \overline{\boldsymbol{D}_{L,\iota}} \gamma^\mu \boldsymbol{D}_{L,\iota} + \\
&\quad + \sin^2\theta_W \left[-2\bar{\ell}_\iota \gamma^\mu \ell_\iota + \frac{4}{3} \overline{\boldsymbol{U}}_\iota \gamma^\mu \boldsymbol{U}_\iota - \frac{2}{3} \overline{\boldsymbol{D}}_\iota \gamma^\mu \boldsymbol{D}_\iota \right] \Big\} \\
J_+^\mu &= -\frac{g}{\sqrt{2}} \sum_{\iota=1}^{3} \left[\overline{\nu_{L,\iota}} \gamma^\mu \ell_{L,\iota} + \overline{\boldsymbol{U}_{L,\iota}} \gamma^\mu V_{\iota\jmath} \boldsymbol{D}_{L,\jmath} \right] = \left(J_-^\mu \right)^\dagger .
\end{aligned}
\tag{2.142}
$$

2.8 Strong interactions of quarks

After having deeply investigated the connection between local gauge invariance and electroweak interactions, we shall now spend a few words concerning the strong interactions. The formalism of the theory will not be developed here, but we shall point out the main differences between the strong and electroweak interactions and give a qualitative description of the theory describing the strong interaction among quarks, namely the quantum chromodynamics (QCD).

Several piece of evidence point in the direction of the quarks being actually triplets, rather than singlets. For example, the magnitude of the cross section for electron-positron annihilation into hadrons ($e^+e^- \to q\bar{q}$) is experimentally found to be three times higher than the theory prediction where quarks are considered to be singlets. Thus, the quarks have been assigned a quantum number, called color, which can take the three values red, blue and green (R, B, G). The leptons, on the other side, are found not to interact *via* the strong interaction. This distinction suggests the possibility that color plays the role of a charge for the strong interactions.

An important difference between the strong and electroweak interactions in due to the non-observation of free quarks, which suggests a theory in which long-range forces mediated by massless gague bosons are present. In this respect, the spontaneous symmetry breaking of the theory giving mass to the gauge bosons is not needed in the case of QCD. Even though this simplifies things with respect to the electroweak interaction, the picture is dramatically complicated by the intensity of the strong interaction, which does not guarantee the possibility of using perturbation theory to compute observable quantities. However, in a general quantum field theory, the effective value of the coupling constant is not actually constant, but it

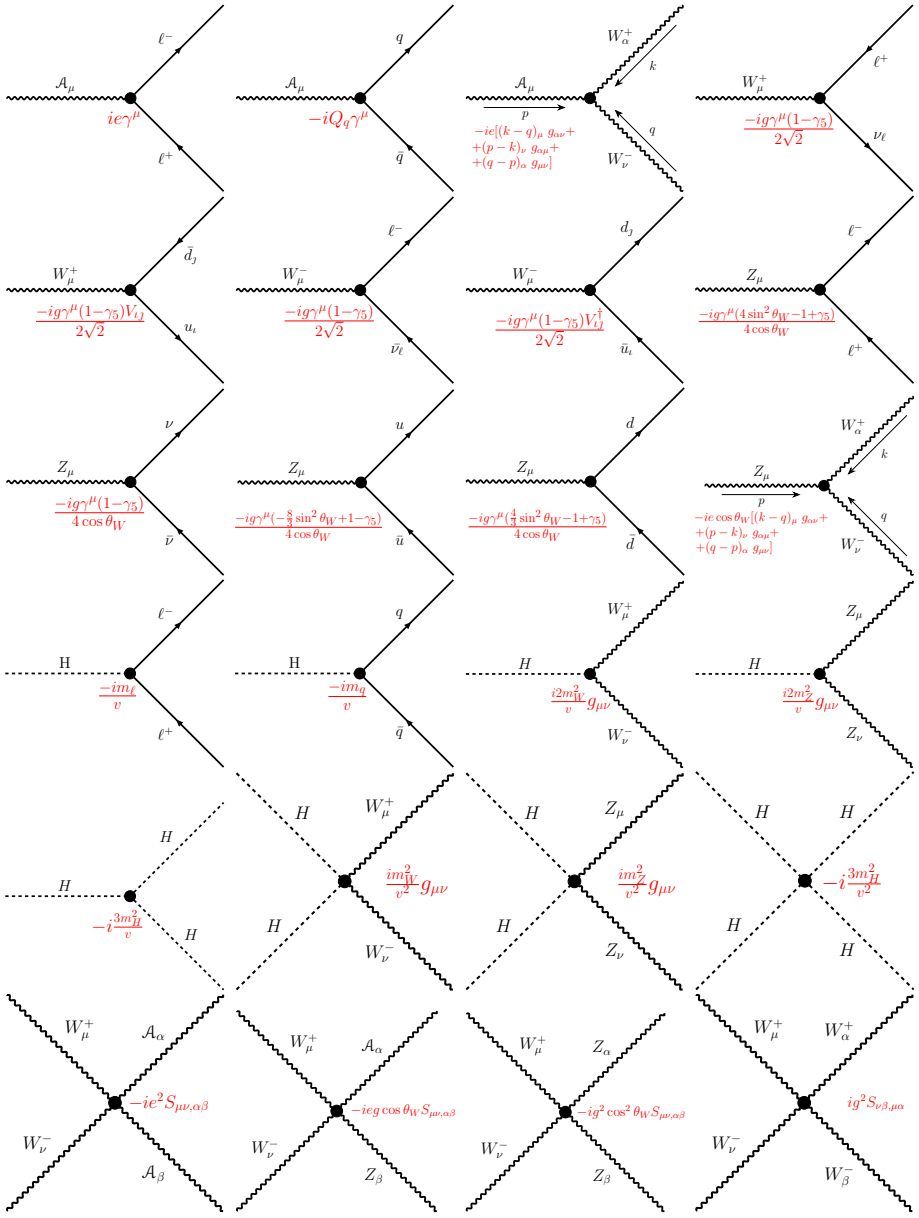

Figure 2.2: Complete set of the electroweak vertices and corresponding Feynman rules that arise from the spontaneously broken Lagrangian of Eq. 2.141. The shortcut $S_{\mu\nu,\alpha\beta} \equiv 2g_{\mu\nu}g_{\alpha\beta} - g_{\alpha\mu}g_{\beta\nu} - g_{\alpha\nu}g_{\beta\mu}$ has been used.

depends on the momentum scale because of renormalization effects. A peculiar property of non-abelian gauge theories, called asymptotic freedom, is that the effective coupling decreases at high momentum scales. This raises the possibility that, in the high momentum regime, perturbative methods can be reliable for QCD.

In analogy with the electroweak theory, and taking into account the evidence of the quark being color triplets, the QCD can be constructed to be a non-abelian theory based on the $SU(3)$ symmetry group. The group will act on 3-dimensional Dirac spinors of the form

$$\psi_\iota = \begin{pmatrix} q_{\iota,\mathrm{R}} \\ q_{\iota,\mathrm{B}} \\ q_{\iota,\mathrm{G}} \end{pmatrix}, \tag{2.143}$$

where the index ι runs on the 6 quark flavors. The gauge bosons associated with the theory are called gluons and, given the dimensionality of the group, can appear in 8 different color combinations. A simplified Lagrangian of the theory can evidently be written as

$$\mathcal{L} = -\frac{1}{2}\mathrm{tr}\left[G_{\mu\nu}G^{\mu\nu}\right] + \sum_\iota \overline{\psi}_\iota \left(i\gamma^\mu D_\mu - m_\iota\right)\psi_\iota, \tag{2.144}$$

where the covariant derivative and the generalized Maxwell's field-strength tensor have the familiar form:

$$\begin{aligned} D_\mu &= \partial_\mu + \frac{ig}{2}\lambda^a b_\mu^a \\ G_{\mu\nu} &= -\frac{i}{g}\left[D_\nu, D_\mu\right], \end{aligned} \tag{2.145}$$

and the 8 matrices λ^a are know as the Gell–Mann matrices. As a final remark to close this brief section, we shall have a closer look at the interaction term arising from the QCD Lagrangian. For a given quark flavor, this term reads

$$\mathcal{L}_{\mathrm{int}} = -\frac{g}{2}b_\mu^a \overline{\psi}_\alpha \gamma^\mu \lambda^a_{\alpha\beta}\psi_\beta, \tag{2.146}$$

which leads at once to the Feynman rule for the quark-gluon vertex $-\frac{ig}{2}\gamma^\mu\lambda^a_{\alpha\beta}$ and to the 8 color combinations of the gluons. In fact, one can explicitly compute the interaction term to see that the gluons can carry the color combinations

$$\begin{array}{ll} (\mathrm{R}\overline{\mathrm{B}} + \mathrm{B}\overline{\mathrm{R}})/\sqrt{2} & i(\mathrm{R}\overline{\mathrm{B}} - \mathrm{B}\overline{\mathrm{R}})/\sqrt{2} \\ (\mathrm{R}\overline{\mathrm{G}} + \mathrm{G}\overline{\mathrm{R}})/\sqrt{2} & i(\mathrm{R}\overline{\mathrm{G}} - \mathrm{G}\overline{\mathrm{R}})/\sqrt{2} \\ (\mathrm{B}\overline{\mathrm{G}} + \mathrm{G}\overline{\mathrm{B}})/\sqrt{2} & i(\mathrm{B}\overline{\mathrm{G}} - \mathrm{G}\overline{\mathrm{B}})/\sqrt{2} \\ (\mathrm{R}\overline{\mathrm{R}} - \mathrm{B}\overline{\mathrm{B}})/\sqrt{2} & (\mathrm{R}\overline{\mathrm{R}} + \mathrm{B}\overline{\mathrm{B}} - 2\mathrm{G}\overline{\mathrm{G}})/\sqrt{6}. \end{array} \tag{2.147}$$

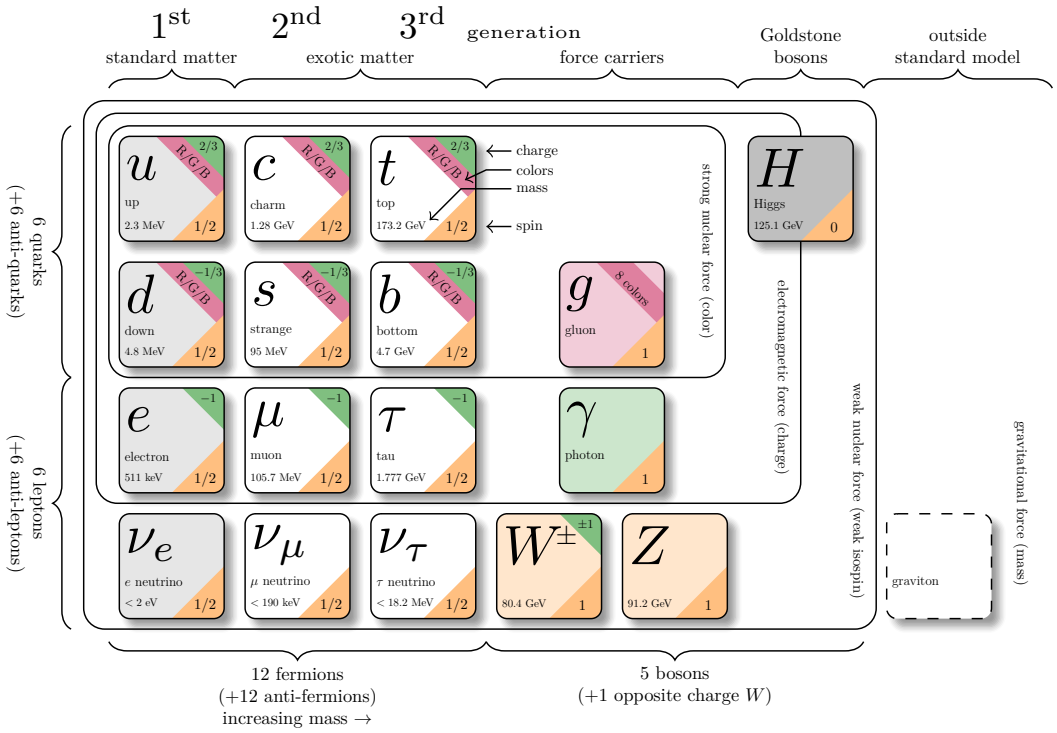

Figure 2.3: Schematic view of the particles belonging to the Standard Model of elctroweak and strong interactions.

2.9 Standard Model summary

As a summary and a partial integration of the matter developed in the previous sections, in Fig. 2.3 we report a schematic view of all the particles entering the SM framework, the symbols commonly used to label them, their masses, their electric and color charges and their spins. Particles are classified as bosons (fermions) based on their integer (half–integer) spin, and are grouped based on the fundamental interactions they are subject to. Fermions are further divided in quarks and leptons and split into the well known three generations. Right beside the SM particles, a blank space is left empty, as up to now no coherent way to include the gravitational interaction in the SM has been found. The related carrier of the interaction, the graviton, is yet unobserved.

Chapter 3

Experimental apparatus: the LHC machine and the CMS experiment

This chapter is devoted to the description of the experimental apparatus that made possible the data analysis presented in this work. In the following, we shall describe the Large Hadron Collider, the particle accelerator which provided the proton-proton collisions, and the Compact Muon Solenoid experiment, the detector that recorded the collisions and provided the data used in this work.

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [46] is a double-ring, proton-proton, superconducting particle accelerator operating at CERN (Geneva, Switzerland). The machine is placed in a 27 km-long, pre-existing, underground tunnel which hosted, from 1989 to 2000, the Large Electron-Positron collider (LEP).

The approval of the LHC project was given by the CERN Council in December 1994. In those early times, the plan was to first build a machine capable of developing a center-of-mass energy of 10 TeV, to be upgraded in a second stage up to 14 TeV. However, intense negotiations held with non-member states in 1995–1996 secured substantial contributions to the project and in December 1996 the CERN Council approved the construction of the 14 TeV machine in a single stage.

The decision to build the LHC at CERN was highly influenced by the cost saving coming from the reuse the LEP tunnel, its facilities and its injection chain. The LEP tunnel is composed by eight arcs where the protons, bent by magnetic fields, are free to circulate and eight straight sections, which are used for various tasks, such as beam injection/dump, particle acceleration or beam collisions. A schematic view of the LHC is shown in Fig. 3.1.

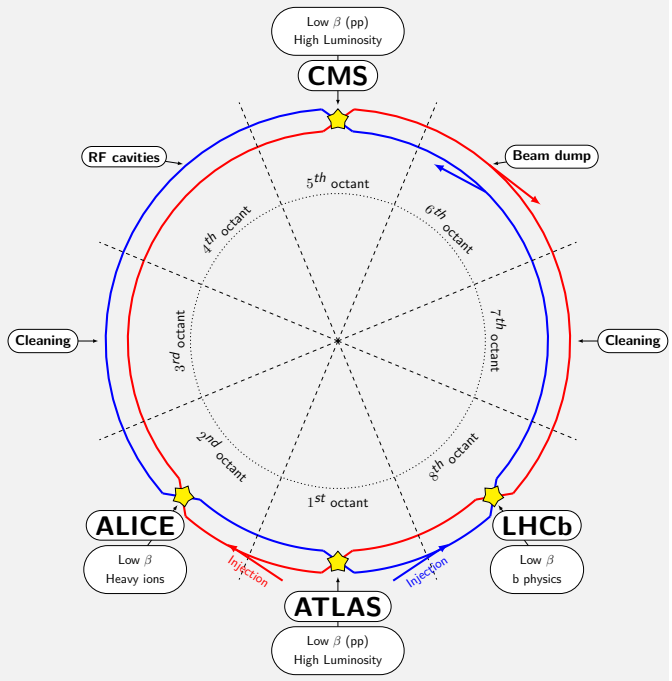

Figure 3.1: Schematic representation of the Large Hadron Collider. The four main experiments are depicted, together with other facilities such as the radiofrequency cavities, the beam cleaning sections, the beam dump section and the injection section.

The LHC is the last stage of a complex acceleration system, and a process of pre-acceleration is needed before the beams are injected in the machine. As a first step, protons are obtained by the ionization of hydrogen atoms; then the Linear accelerator 2 (Linac2) transfers to the particles the first amount of energy, accelerating them up to an energy of 50 MeV. Starting from Run3, the Linac2 will be replaced by the Linac4, a new linear accelerator capable to accelerate negative hydrogen ions (H^-, consisting of hydrogen atoms with an additional electron) up to 160 MeV. The Linac4 will be one of the key elements in the project to increase the luminosity of the LHC. In this new scenario, the ions will be stripped of their two electrons during the injection in the second stage of the acceleration system.

In the following step, the protons enter the superimposition of four synchrotron rings called the Proton Synchrotron Booster (PSB) and are accelerated up to 1.4 GeV. Then, they enter the Proton Synchrotron (PS) and the Super Proton Synchrotron (SPS) to reach an energy of 450 GeV. Finally, they are injected in the LHC and accelerated for the last time to the nominal energy. A scheme of this acceleration system is presented in Fig. 3.2.

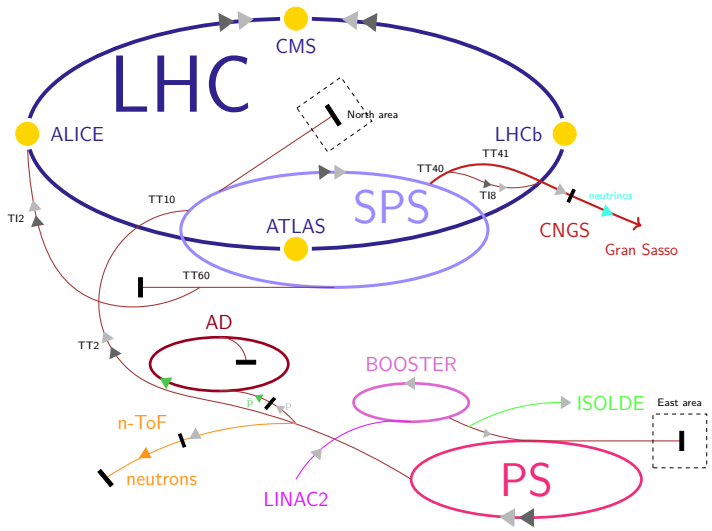

Figure 3.2: The acceleration system at CERN. The protons travel from the Linac2 through the PSB, the PS and the SPS to finally reach the LHC, where they undergo the final step of the accelerating process and they eventually collide at the four interaction points where the experiments are placed.

3.1.1 Performance goals and limitations

One of the main goals of the LHC is to produce rare events, possibly coming from physics beyond the SM. To achieve this, collisions must take place at high rate $R = \sigma \cdot \mathcal{L}$, where σ is the cross section of the process of interest and $\mathcal{L}$ is the instantaneous luminosity of the machine. Thus, a high instantaneous luminosity is crucial for the success of the experimental program. The instantaneous luminosity depends exclusively on the machine design and can be written as

$$\mathcal{L} = \frac{n^2 N_b \gamma \nu}{4\pi \epsilon_n \beta^*} F,$$

where n stands for the particle content per bunch; N_b is the number of bunches per beam, γ is the relativistic factor, ν is the revolution frequency, ϵ_n is the transverse beam emittance, β^* is the beta function at the collision point and F is the the so-called geometric luminosity reduction factor, which accounts for the crossing angle at the interaction point. The maximum particle density per bunch is limited by beam-beam interactions that particles experience when the bunches of the two beams collide. This effect can be measured by the linear tune shift

$$\xi = \frac{N_b r_p}{4\pi \epsilon_n},$$

where $r_p = e^2/4\pi\epsilon_0 m_p c^2$ is the classical radius of the proton. Given the experience gained in previous hadron colliders, the total linear tune shift summed over all the interaction points (IPs) should not exceed the value $\xi = 0.015$. This means that, having the LHC three proton experiments requiring head-on collisions, the linear tune shift at each IP must satisfy the requirement $\xi < 0.005$.

The LHC hosts two, high-luminosity experiments: A Toroidal LHC ApparatuS (ATLAS) [47] and Compact Muon Solenoid (CMS) [48], both designed for a peak, nominal, instantaneous luminosity of $10^{34}\text{cm}^{-2}\text{s}^{-1}$ (reached for the first time on June 26, 2016 and exceeded during 2017 with a peak value of $2.06 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$); there is then a third, low-luminosity experiment, LHCb [49], designed for studies on the physics of the bottom quark (designed for a luminosity of $2\times10^{29}\text{cm}^{-2}\text{s}^{-1}$) and a fourth experiment, A Large Ion Collider Experiment (ALICE) [50], studying heavy-ions collisions (designed for a luminosity of $10^{27}\text{cm}^{-2}\text{s}^{-1}$).

The LHC instantaneous luminosity is not constant during data-taking, but it rather decreases due to effects such as the degradation of intensities and emittances of the beams, the main reason for this being the beam loss caused by collisions. It is found that the instantaneous luminosity decreases with a decay factor

$$\tau_{\text{decay}} = \frac{N_0}{\mathcal{L}_0 \sigma_{\text{tot}} k},$$

where N_0 is the initial beam intensity, $\mathcal{L}_0$ the initial luminosity, σ_{tot} is the total cross section ($\sigma_{\text{tot}} \approx 10^{25}\text{cm}^2$ at 14 TeV) and k is the number of IPs, following a decay law of the form

$$\mathcal{L}(t) = \frac{\mathcal{L}_0}{(1 + t/\tau_{\text{decay}})^2}.$$

To quantify the amount of data collected during a run, the instantaneous luminosity must be integrated over time to get the integrated luminosity:

$$L = \int \mathcal{L}(t)\ \mathrm{d}t.$$

When the instantaneous luminosity becomes too low, the beams must be dumped and the accelerator refilled. Taking into account the filling time of the injection chain, the minimum time required for ramping the LHC up to the nominal 14 TeV energy and possible beam aborts, the total turnaround time, defined as the time needed to establish stable beams conditions after a beam dump has occurred, is estimated to be around 70 minutes, as a theoretical minimum. In reality, up to seven hours can be needed, as on average only one fill out of six leads to a successful fill at full energy.

3.1.2 Magnets

The LHC relies on superconducting magnets that are at the edge of the present technology. In particular, they exploit the technology based on niobium-titanium (Nb-Ti) Rutherford cables, cooled by superfluid helium to a temperature of 1.9 K. With such a design, a magnetic field of 8.33 T can be achieved. Colliding two counter-rotating proton beams requires opposite magnetic dipole fields in both rings. The LHC is therefore designed with separate magnet fields and vacuum chambers in the main arcs. The LEP tunnel has a diameter of 3.7 m, which made almost impossible to install two completely separated proton beams with independent magnet system. Thus, the twin-bore magnet design [51] was adopted, in which the two proton rings are housed in a common cryostat, with the magnetic flux circulating in the opposite sense through the two channels.

The LHC hosts 1232 main dipole magnets or cryodipoles. The core of a cryodipole is the dipole cold mass, which contains all the components cooled by superfluid helium. It is provided with two apertures for the bore tubes in which the protons circulate. It has an overall length of about 16.5 m, a diameter of 570 mm and a mass of about 27.5 t. The mass is also curved in the horizontal plane with an angle of 5.1 mrad in order to correctly match the particle trajectories. The manufacturing of these magnets has been very challenging, as the successful operation of the LHC requires for them to have practically identical characteristics, with relative variations in the field strength that should not exceed $\sim 10^{-4}$. Such a high level of reproducibility required a close control on the building parameters, such as the coil diameter and length and the length ratio between the magnetic and non-magnetic parts of the yoke.

In correspondence of the eight intersections along the LHC ring (four in the experiment areas, two for the beam cleaning, one for the radiofrequency cavities and one for the beam dumping, see Fig. 3.1), the main quadrupoles are placed. These magnets accomplish many tasks, such as collimating the beams in proximity of the experiments in order to maximize the probability of pp collisions. A set of multipole correcting magnets are coupled with the quadrupoles, giving rise to 10 different combinations of magnets used for a fine tuning of the magnetic fields.

3.1.3 Radiofrequency cavities

To store and accelerate the beams that are injected in the LHC, radiofrequency cavities (RFs) are used. Such RFs are split in two independent systems, one for each beam, and work in a superconductive regime. In order for an RF cavity to accelerate particles, an RF power generator supplies an electromagnetic field. An RF cavity is designed in such a way that electromagnetic waves become resonant and build up inside the cavity. Protons passing through

the cavity feel the overall electromagnetic force of the resulting field and are accelerated along the rings.

The field inside an RF cavity is oscillating in time and must be synchronized with the revolution frequency of the beam. In order for this to happen, the ratio between the frequency of the RF voltage, ν_{RF}, and the revolution frequency, ν_{rev}, must be a constant integer, called harmonic number h. The frequency at which the LHC RF operate is 400.8 MHz, the radius of the LHC machine is approximately 4245 m and the particles circulate at approximately the speed of light, so that we can write

$$h = \frac{\nu_{\mathrm{RF}}}{\nu_{\mathrm{rev}}} = \frac{\nu_{\mathrm{RF}} 2\pi R}{c} \approx 35640.$$

This means that, in principle, there are 35640 spots along the LHC ring, called buckets, in which the particles can be stored. However, not all the buckets are filled with particles, as a certain number of buckets in a row (the so called abort gap) is always left empty to give time to the magnets to perform a beam dump, if needed. The number of actually occupied buckets in the LHC is 2808.

Since the field inside an RF oscillates, a particle with exactly the right energy will experience zero accelerating force when the LHC is at full energy. Such a proton is called a synchronous particle. On the other hand, protons with slightly higher energies than the synchronous particle will travel on a longer orbit, thus arriving later and being decelerated, so that they stay close to the energy of the ideal particle. An analogous reasoning can be made for particles with slightly lower energies than the synchronous particle, which in turn arrive early and are thus accelerated. Suppose that a particle has a higher energy than the synchronous particle. Then, after a given number of turns in which the particle is decelerated, it will reach the energy of the synchronous particle, but still being late on the last turn. This way, it will be decelerated under the optimal energy and it will experience an accelerating force in the following turns. In a similar fashion, the particle will then increase its energy and exceed again the the optimal energy. This oscillating pattern takes the name of synchrotron oscillations and leads to the splitting of the protons into bunches. This is visually shown in Fig. 3.3.

3.1.4 Vacuum system

The LHC exploits the properties of ultra-high vacuum in several different ways. The use of ultra-high vacuum has two main goals: first, to insulate the cryogenically-cooled parts of the machine, and second, to avoid collisions between the beams and gas molecules, thus ensuring a high beam lifetime. The system is split into three independent parts: one for insulating the cryomagnets, one for insulating the helium distribution line and one for the beam pipe. All the three vacuum systems are then subdivided into

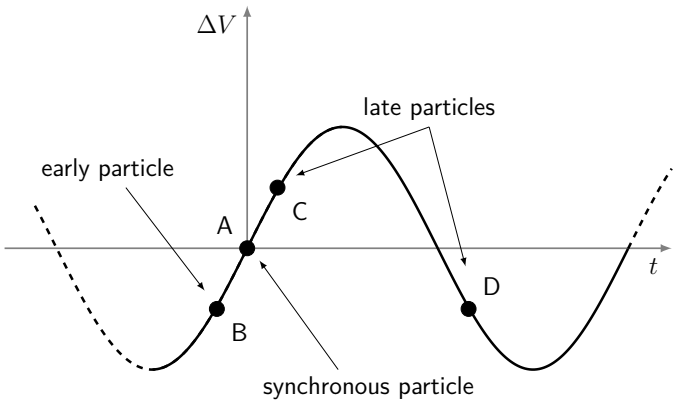

Figure 3.3: Time dependence of the potential difference inside a radiofrequency. A synchronous proton (A) will feel zero accelerating force; a proton with a slightly lower energy (B) will arrive early, thus experiencing a negative potential difference and being accelerated to approach the optimal energy; a proton with a slightly higher energy (C) will arrive late, thus experiencing a positive potential difference and being decelerated to approach the optimal energy; if a proton is even more energetic and is very late, it arrives during the second half period of the oscillation (D), thus experiencing a negative potential difference, being accelerated and separated from the other particles, forming a second bunch.

manageable sectors by the use of vacuum barriers in the insulating part and vacuum valves for the beam pipe.

The design of the beam vacuum system is driven by the dynamic phenomena that could spoil the beam insulation, such as synchrotron radiation, which can hit and heat the beam pipe, and electron cloud, mainly caused by the ionization of residual gas and synchrotron emission. In order to intercept these sources of heat, a beam screen has been installed. The system is actually split in two independent chambers, one for each beam pipe, which merge into a single one in the collision regions.

The insulation vacuum systems includes the magnet cryostats and the helium distribution line. Vacuum barriers are installed in such a way that the two systems are kept separated and warming of individual machine cells is possible.

3.1.5 Beam injection/dump

Beam injection is performed in sectors 2 and 8. In both insertions, the beams approach the machine from outside and below the machine plane. The beams are first directed by dipole magnets into the injection line; then, a set of 5 septum magnets deflect the beam in the horizontal plane and finally four, fast-pulsed kicker magnets deflect the beam vertically. To protect the

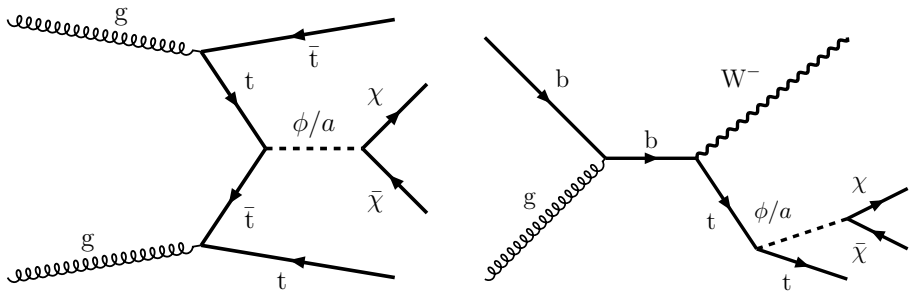

Figure 3.4: Examples of leading-order Feynman diagrams for the associated production of WIMPs (χ) and top quarks in the so called minimal flavor violation scenario. In this framework, the interaction between DM and SM particles is mediated by a scalar particle (ϕ or a) with a Yukawa-like coupling identical in structure to that of the SM.

LHC from malfunctioning of the kickers, an injection beam stopper is placed downstream. The design of the kickers is of particular interest, as they must provide an intense magnetic field in a fast pulse. A total integrated field of 1.2 Tm and pulses with rise/fall time of the order of tens of ns are needed. To achieve this, each magnet is powered by a separate pulse-forming network.

The dedicate beam dumping system is located in sector 6. The system is made by extraction kicker magnets, septum magnets and dilution kicker magnets. To protect the machine, the system is able to fast-extract the beam in a single beam turn and direct it towards to an external absorber. The absorbing material must be positioned sufficiently away to allow for beam dilution in order not to overheat the absorbing material. Given the destructive potential of the LHC beams, they cannot be dumped in a single point but must rather be diluited in a "e" shape, in such a way that the energy is deposited on a wider surface. The system is not only able to dump ordinary beams, but it can also deal with beams outside of the normal parameters, to prevent incidents due to equipment failures or abnormal optic settings in the rings.

3.1.6 Scientific goals

The LHC machine, the most powerful particle accelerator ever built, has been designed to accomplish several scientific goals. A first, important task, was to confirm or discard the CDF and D0 results [52] which indicated the evidence for a particle consistent with the Higgs boson in the mass range $115 < m_{\mathrm{H}} < 140$ GeV with significance of 2.9 standard deviations. This search ended successfully on July 4^{th}, 2012 when the ATLAS and CMS Collaborations announced independently the discovery of a new particle compatible with the SM Higgs Boson.

However, the experimental program of the machine goes far beyond the Higgs boson discovery. For example, the LHC experiments could possibly be able to determine the nature of dark matter (DM). Firstly postulated in the 30s [53], the existence of DM is supported by indirect evidence coming from astrophisical observations, such as galactic rotation velocity dispersion curves [54], and could account for roughly 25% of the matter of the universe. At the LHC, proton-proton collision could produce in a direct way DM candidates, such as weakly interacting massive particles (WIMPs). An example of processes that directly couple the SM and DM sectors is given in Fig. 3.4.

Moreover, the LHC could possibly observe the production of supersymmetric particles. The theory of supersymmetry (SUSY) [55, 56], if confirmed by the experimental evidence, could shed light on aspects of grand unified theories [57, 58], give an answer to the hierarchy problem [59] and solve the puzzle of the radiative corrections to the Higgs boson mass [60, 61]. Through the detailed study of the properties of b-flavored hadrons, the experiments could be able to enhance our knowledge on the sources of CP violation [62]. Finally, exploiting data coming from ion-ion collisions, we may get an insight into the properties of the exotic state of matter called quark-gluon plasma [63].

3.2 The Compact Muon Solenoid detector

The CMS experiment [64] is a multi-purpose particle detector operating at the CERN LHC. It is housed in sector 5 of the LHC ring, in the Point 5 LHC area, about 100 m underground, near the French village of Cessy. In its basic design, CMS is a 21.6 m long, 14.6 m high cylinder made of concentric layers, each layer corresponding to a different subdetector or piece of equipment.

Being coupled with the most powerful particle accelerator ever, CMS has to accomplish many challenging tasks. At design luminosity, approximately one billion proton-proton collisions per second take place in the heart of the detector, and thus a very effective online selection process must take place to reduce the rate to around 100 events per second in order for data acquisition and storage to be possible. The very short time of 25 ns between proton bunches makes mandatory to have ultra-fast acquisition systems, in such a way that collisions belonging to different bunches do not overlap during data taking. The high flux of radiation coming from the interaction point hits the front end electronic devices of the detector, which thus must be produced with a radiation-hard design.

The coordinate system adopted by CMS has the origin centered at the nominal collision point inside the experiment, with the y-axis pointing vertically upward and the x-axis pointing radially inward toward the center of the LHC. Thus, the z-axis points along the beam direction toward the

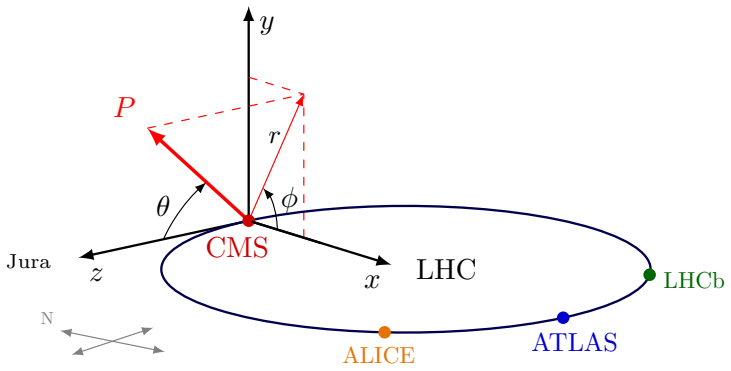

Figure 3.5: Coordinate system adopted by the CMS experiment.

Jura mountains from LHC Point 5. Given the evident symmetry of the detector, a cylindrical coordinate system is often used. The azimuthal angle ϕ is measured from the x-axis in the x-y plane and the radial coordinate in this plane is denoted by r. The polar angle θ is measured from the z-axis. A visual representation of the coordinate system adopted by CMS is reported in Fig. 3.5. The widely-used pseudorapidity variable is defined as $\eta = -\ln[\tan(\theta/2)]$. Also, the momentum transverse to the beam direction, denoted by p_{T}, is thus computed from the x and y components of the momentum vector.

In the following subsections, we shall provide a description of all the relevant CMS subdetectors, starting from the inner layers and going outwards.

3.2.1 Tracking system

The CMS tracking system has been designed to efficiently and precisely measure the trajectories of charged particles emerging from the interaction point, as well as to reconstruct secondary vertices. It is the closest subdetector to the beam pipe and has a length of 5.8 m and a diameter of 2.5 m. Given the high number of multiple proton-proton collisions within the same bunch crossing and the 25 ns bunch spacing, a detector technology featuring high granularity and fast response is required, in such a way that the trajectories can be identified reliably and attributed to the correct bunch crossing. Also, being so close to the interaction point, the intense particle flux can severely damage the tracking system; thus, a radiation-hard technology is required.

All these challenging quality requirements have been addressed using a silicon-based detector technology. The readout chips of the CMS tracking system were found to be severely damaged by radiation at the end of 2016; thus, during an extended end-of-year shutdown during winter 2016/2017, the system has been fully replaced.

With respect to the previous design, the new tracker features an additional

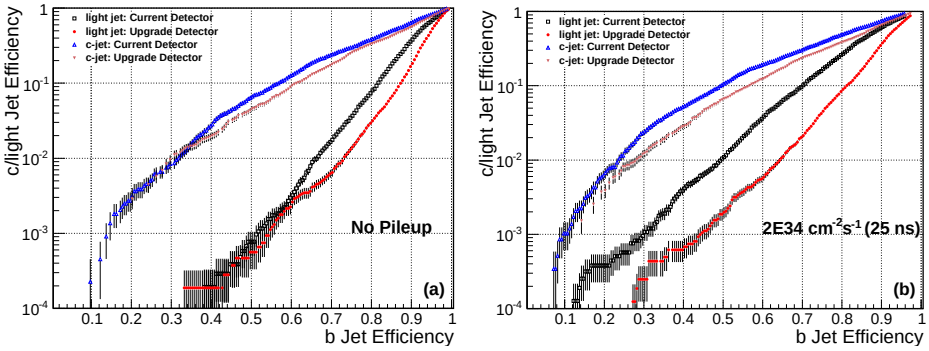

Figure 3.6: Performance of the combined secondary vertex b tagging algorithm (see Section 4.2) for jets with $p_{\mathrm{T}} > 30$ GeV in a $\mathrm{t\bar{t}}$ sample with zero pileup (left), and an average pileup of 50 (right). The plots are taken from the CMS TDR for the pixel detector upgrade [65], which was published in 2012 and described the upgrade plans for the tracking system; thus, points labeled as "Current Detector" actually refer to the previously-installed system, while "Upgrade Detector" refers to the detector now present inside CMS. The results show that the use of the new detector outperforms the old one in terms of b tagging efficiency.

fourth layer of pixel modules in the barrel and a third disk per side in the endcap region. The number of channels has almost doubled, reaching 124 million, giving a four-hit coverage in the whole tracking region up to $|\eta| < 2.5$. To improve the vertex resolution and b tagging efficiency, the radius of the innermost layer is now smaller, corresponding to 29 mm. This implied that a new beam pipe was also needed, which was installed during Long Shutdown 1 in 2013/2014. As an example of a performance which is improved with the new design, in Fig. 3.6 we report the gain in performance of an algorithm used to identify jets produced by the hadronization of bottom quarks.

Moreover, the material budget of the new tracker has been significantly reduced, firstly by switching from a C_6F_{14} cooling system to a two-phase CO_2 cooling system, which allows lower coolant mass and smaller pipes, and secondly by moving the electronic boards and connections out of the tracking volume.

The basic design of pixels basically remained unchanged. A module consists of 285 μm thick silicon sensor comprising 66540 pixels, each with a size of 100 μm $\times$ 150 μm. A technique based on charge sharing between neighboring pixels is used to get a single hit resolution of 5–7 μm.

3.2.2 Electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) is a hermetic, homogeneous calorimeter made of 61200 lead tungstate ($PbWO_4$) scintillating crystals mounted in the central barrel part, closed by 7324 crystals in each of the two endcaps. Also, a preshower detector is placed in front of the endcap crystals.

The high density (8.28 g/cm^3), short radiation length (0.89 cm) and small Molière radius (2.2 cm) result in a compact, high-granularity calorimeter. Avalanche photodiodes are used in the barrel as photodetectors, while vacuum phototriodes are used in the endcaps. The use of high-density crystals gives ECAL a fine granularity, a fast response and good radiation hardness. The shape of the crystals is a truncated pyramid and they have a scintillation decay time which is the same order of magnitude of the LHC bunch spacing, with about 80% of the light being emitted in 25 ns.

The barrel part of the ECAL (EB) covers a pseudorapidity range of $|\eta| < 1.48$, and the cross-section of the crystals in this region corresponds to 22 mm $\times$ 22 mm at the front face of the crystal, and to 26 mm $\times$ 26 mm at the rear face. The crystal length is 230 mm, corresponding to 25.8 radiation lenghts. The centers of the front faces of the crystals are at a radius 1.29 m.

The endcap part of the ECAL (EE) covers the range $1.48 < |\eta| < 3.0$. The longitudinal distance between the interaction point and the endcap envelope is 3.15 m, which takes into account the estimated shift toward the interaction point by about 2 cm when the magnetic field is switched on. The crystal geometry in this region is identical to the one of the barrel region.

The number of scintillation photons emitted by the crystals and the performance of the photodiodes are temperature dependent, both variations being negative with the increase of temperature. The temperature of the system must therefore be kept as constant as possible, requiring a cooling system capable of extracting the heat from crystals, photodiodes and readout electronics. The nominal operating temperatore of the ECAL is 18 °C. The cooling system uses water to stabilize the detector.

The preshower detector is designed to identify pions in the endcap region, to help the identification of electrons against minimum ionizing particles and to improve the position determination of electrons and photons. It is basically a sampling calorimeter made of two layers: first, lead radiators initiate the shower induced by incoming photons or electrons, and then silicon strip sensors placed after each radiator measure the energy deposit. The preshower detector covers the region $1.65 < |\eta| < 2.6$.

3.2.3 Hadron calorimeter

The hadron calorimeter (HCAL) is designed to measure the properties of hadron jets and, indirectly, of neutrinos or exotic particles resulting in missing transverse energy. It is composed by two half barrels (HB) covering

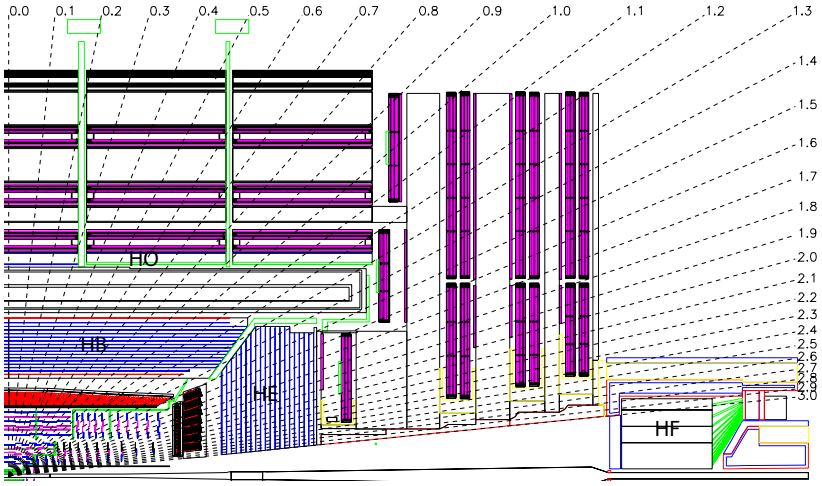

Figure 3.7: Longitudinal view of the CMS detector, showing the locations of the hadron barrel (HB), endcap (HE), outer (HO) and forward (HF) calorimeters.

a pseudorapidity range of $|\eta| < 1.3$, completed by two endcaps (HE) which cover the pseudorapidity range of $1.3 < |\eta| < 3.0$. Beyond $|\eta| = 3.0$, a forward hadron calorimeter (HF) placed about 11 m after the interaction point extends the pseudorapidity coverage to $|\eta| = 5.2$, making the CMS HCAL an almost hermetic calorimeter. Since the space for the barrel is limited by the ECAL and the magnet coil, a complementary outer hadron calorimeter (HO) is placed outside of the magnet. Figure 3.7 shows where each of these components are placed inside of the CMS detector.

The HB calorimeter is a sampling calorimeter composed by 36 identical azimuthal wedges which form the two half-barrels, each of them divided in 4 sectors. The absorber is made of alternating layers of stainless steel or brass, while the active material is a plastic scintillator which exhibits long term stability and a good radiation hardness. The scintillator is used in the tile plus wavelength-shifting fibre design to collect the light produced in the scintillating material. The light collected by the fibres is transferred to photodiodes which enhance the signal and make possible its usage.

The forward region covered by HE calorimeter is expected to contain about 34% of the particles produced in the final state of a collision. Given the high luminosity delivered by the LHC, the HE must handle rates up to the MHz and must have an excellent radiation tolerance. Also, since it is inserted at the ends of a 4-T magnet, the absorber must be made of a non-magnetic material and show good mechanical properties. These requirements led to the choice of trapezoidal-shape plastic scintillators and cartridge brass as the absorbing material. Plus, the absorber design is optimized to minimize the

cracks between HB and HE, since the single-particle energy resolution would be anyway spoiled by multiple interactions within the same bunch crossing, fragmentation effects and so on.

In the central pseudorapidity region, the combined effect of the EB and HB calorimeters is not enough to provide sufficient confinement of the hadronic showers. To ensure an adequate sampling depth in the $|\eta| < 1.3$ region, the hadron calorimeter is extended outside of the magnet with the HO calorimeter. This subdetector uses the magnet as an additional absorber and is able to identify late showers. The presence of the HO prevents shower leakage and leads to important improvements on the measurement of the missing transverse energy.

The HF calorimeter undergoes very intense particle fluxes. After ten years of operation of the LHC, the absorbed dose received by the HF in the direction $|\eta| = 5$ has been estimated to be ≈ 10 MGy. This hostile environment presents a considerable challenge to calorimetry, and the design of the HF was mainly driven by the necessity to survive in these harsh conditions. The calorimeter consists of an absorber composed by steel, grooved plates. Quartz fibers are inserted in such grooves. The fibers are divided in two sets, the first one running over the full depth of the absorber, the second one starting at a depth of 22 cm from the front of the detector. By reading out these two sets separately, the detector is able to distinguish showers initiated by photons and electrons, which deposit most of the energy in the first 22 cm of the detector, from showers initiated by hadrons which on average produce equals signals in two regions. The HF uses a Čerenkov-based technology. The signal is generated when particles above the threshold generate Čerenkov light inside the fibers, which is then read by photomultipliers.

3.2.4 Superconducting magnet

The superconducting magnet of the CMS experiment is designed to reach a field of 4 T. The joint requirements of such an intense field, keeping at the same time a contained size, presented severe technological challenges. However, the choice of an intense magnetic field is crucial, as lower intensities would correspond to higher Level-1 trigger rates, to a degradation in the particles momentum resolution, and to more problematic calibrations of the ECAL.

The solenoid has a diameter of 6 m and is 12.5 m long, making it the biggest superconducting magnet ever built, and stores an energy of 2.6 GJ at full current. In order to achieve such intense current, the magnet must operate in a superconducting regime at a temperature of 4.6 K. The distinctive features of the cold mass are the winding, which consists in 4 layers made from Nb-Ti conductor, the high ratio between stored energy and cold mass, which is 11.6 kJ/kg, and the thinness of the coil, with a radial extension Δr

which is small compared to the radius of the solenoid, $\Delta r/r \approx 0.1$.

The flux of the magnetic field is returned by a 10000-t iron yoke made of 5 wheels and 2 endcaps, with the central wheel housing the coil and its cryostat. The yoke is designed in a 12-sided structure that also houses the muon chambers. The movement of such a heavy setup is granted by the combined use of heavy-duty air pads and grease pads. This makes possible to easily and quickly open the detector for maintenance tasks.

3.2.5 Muon system

An efficient detection of muons is a powerful tool to recognize signatures of interesting events in a QCD-dominated environment such as the LHC collisions. A decay to four muons is the golden decay channel for the Higgs boson, and exotic theories such as SUSY also predict final states containing muons. Therefore, as it is suggested by the name of the experiment, a great attention has been payed in the design of the muon system in order to achieve robust and precise muon measurements.

The muon system is devoted to three functions: muon identification, momentum measuring and triggering. It is constructed by making use of three different kinds of gaseous detectors and it is composed by a barrel region plus two endcap regions.

In the barrel region, where the magnetic field is uniform and mostly contained in the return yoke, drift chambers with rectangular drift cells are used. These barrel drift tube (DT) chambers cover the pseudorapidity range $|\eta| < 1.2$. Each of the five wheels composing the barrel is divided in 12 sectors, each of them containing 4 DT chambers. Each DT chamber is made by 3 (or 2) superlayers (SLs), each made of 4 layers composed by 60–90 rectangular cells. This complex system, shown schematically in Fig. 3.8, results in more than 172000 sensitive wires. The wires in inner and outer SLs are parallel to the beam line and provide a track measurement in the r–ϕ plane, while the wires in the middle SL are orthogonal to the beam line and measure the z position along the beam.

The endcap region of the muons system is equipped with cathode strip chambers (CSCs). The CSCs are trapezoidal, multiwire proportional chambers comprised of 6 anode wire planes interleaved among 7 cathode panels, covering the pseudorapidity range $0.9 < |\eta| < 2.4$. The wires run in the azimuthal direction and measure the radial coordinate of a track. Also, the ϕ coordinate is obtained by interpolating charges induced on the strips. The CSCs are a very versatile detector, since can operate at high rates, in non-uniform magnetic fields and they do not require high-precision measurements of gas temperature and pressure.

Finally, both the barrel and endcap regions are equipped with resistive plate chambers (RPCs). The RPCs are gaseous parallel-plate detectors that achive adequate spatial resolution with a time resolution comparable to that

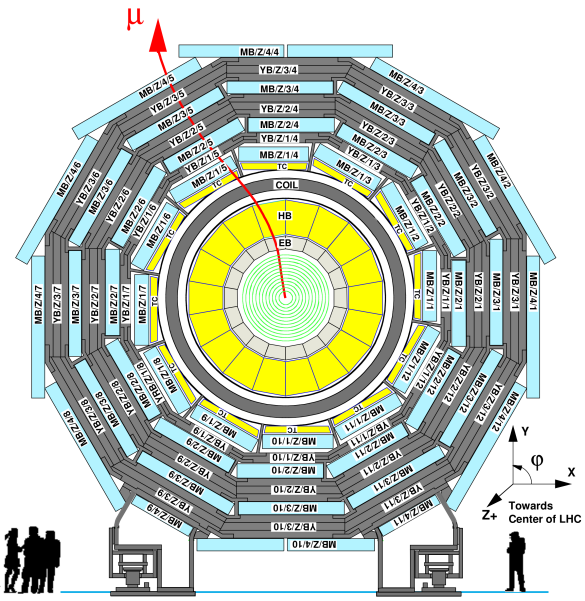

Figure 3.8: Layout of the barrel muon DT chambers. The segmentation in 12 sectors, each composed of 4 chambers (drawn in light blue) is shown.

of scintillators. An RPC is capable of tagging the time of a ionizing signal in a time which is shorter than the 25 ns bunch spacing of the LHC. The RPC basic design consists of two gaps operated in avalanche mode with common pick-up read-out strips in between. The total signal collected by the RPC is the sum of the signals produced in the two gaps, thus allowing the operation at a lower gas gain, with an increase detector efficiency with respect to a single-gap design.

3.2.6 Trigger

The LHC provides a very high rate of proton-proton collisions: with a nominal bunch spacing of 25 ns, it corresponds to 40 MHz. Since no storage capacity and computing power are available to save and analyze such a big amount of data, a drastic reduction in rate must be achieved. This is done by the trigger system, which is the starting point of the event selection. The reduction in the rate of events is obtained in two steps: the Level-1 (L1) trigger and the High Level Trigger (HLT).

The L1 trigger is hardware-based and consists of custom-designed, largely programmable electronics. To achieve a good level of flexibility, the L1 hardware is implemented in FPGA technology, when possible. The system uses coarse information from the calorimeters and the muon system, while holding the high-resolution data in pipelined memories in the front-end electronics.

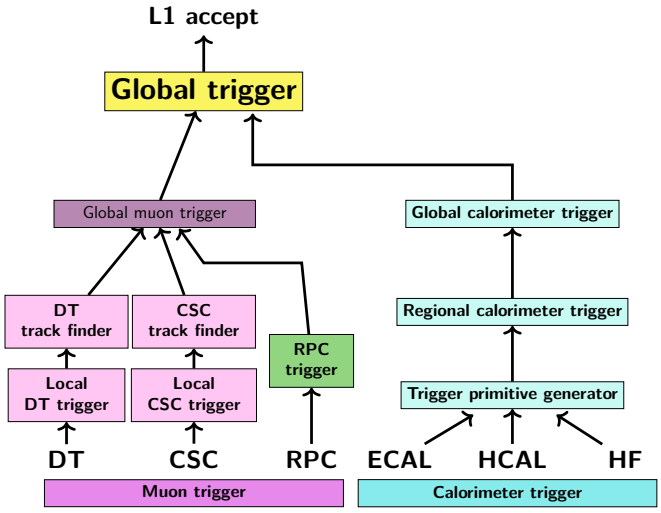

Figure 3.9: Layout of the Level-1 trigger.

The L1 trigger is composed by local, regional and global components. At the bottom end, the local triggers, also known as the trigger primitive generators, are based on calorimetric deposits and hits in the muon system. Then the regional triggers combine this information using pattern logic to determine and rank trigger objects such as electrons or muons, based on energy, momentum and quality. The ranking reflects the level of confidence attributed to the L1 parameter measurements, based on detailed knowledge of the detectors and trigger electronics and on the amount of information available. Finally, the global calorimeter and muon triggers determine the highest-rank calorimeter and muon objects and pass them to the global trigger, which is the top-level entity of the L1 hierarchy. This last element takes the decision whether to accept the event or reject it. This hierarchic structure of the L1 trigger is shown in Fig. 3.9. The rate reduction capability of the L1 trigger is such that the rate of collisions is reduced from the nominal 40 MHz developed by the LHC to 100 kHz after the L1 selection has taken place.

The HLT is a software system implemented in a farm of about 1000 commercial processors. The HLT algorithms are very numerous, quickly evolve in time and are related to the needs of specific fields of research. For this reason, they cannot be investigated in full detail here. A description of the HLT path used in the analysis described in this work will be given in Section 4.3. However, in general, the HLT has access to the complete readout data and can therefore perform complex calculations, similar to those that are made during offline analyses, if required for specially interesting events. The HLT is able to reduce the rate of collisions from the 100 kHz of the L1 trigger 100 Hz.

Chapter 4

Data analysis

This chapter describes in detail the main steps of the data analysis procedure. It is organized as follows: in Section 4.1, the data and Monte Carlo samples used to perform the analysis are described, while in Section 4.2, the physical objects entering the analysis are listed. After that, in Section 4.3, the signal trigger path is presented, and in Section 4.4 all the steps forming the event selection are described and justified. Subsequently, in Section 4.5, several consistency checks concerning the selection procedure are shown. In Sections 4.6 and 4.7, the estimation of the main backgrounds of the analysis is described. Finally, in Sections 4.8 and 4.9, the template shapes used for the extraction of the signal strength parameter and the systematic uncertainties affecting the search are discussed.

4.1 Data and Monte Carlo samples

4.1.1 Data

The present analysis is based on pp collision events recorded during the 2016 data-taking period. All the different eras of data-taking, *i.e.*, from B to H, are used, taking into account only good runs and luminosity sections which are present in the golden JSON. As far as era B is concerned, only the version 2 dataset is used, since the version 1 dataset does not contain events in the golden JSON. As far as era H is concerned, both version 2 and version 3 dataset are used, as they contain no overlapping events. This results in a total integrated luminosity of 35.922 fb^{-1}. Data events collected by the CMS detector are first of all processed through the HLT. From there, several HLT paths are designated to live inside specific primary datasets. Events collected by the trigger paths included in the JetHT primary dataset are used in this analysis. The different data-taking eras contributing to this analysis, together with the corresponding run ranges and collected integrated luminosities, are listed in Table 4.1.

Sample	Run range	Luminosity (pb^{-1})
Run2016B_ver2	273150-275376	5750
Run2016C	275656-276283	2573
Run2016D	276315-276811	4242
Run2016E	276947-277420	4025
Run2016F	277932-278808	3105
Run2016G	278820-280385	7576
Run2016H_ver2	281613-284035	8435
Run2016H_ver3	284036-284044	216

Table 4.1: Data samples used in the analysis, along with the corresponding run ranges and integrated luminosity.

4.1.2 Monte Carlo

The Monte Carlo (MC) samples used in this analysis include the simulation of the signal process, as well as the simulation of all the backgrounds relevant to the search.

The samples of $t\bar{t}H(b\bar{b})$ events, in which the Higgs boson decays to a $b\bar{b}$ pair, and its complementary $t\bar{t}H(nob\bar{b})$ in which the Higgs boson decays to any other particles, have been simulated with the MadGraph5_aMC@NLO generator [66] using the next-to-leading-order (NLO) merging scheme proposed by Frederix-Frixione [67] interfaced with the pythia8 [68] parton shower.

The dominant background, namely the QCD multijet production, has been simulated with MadGraph, interfaced with pythia8. It has been divided into six subsamples, each one corresponding to a given range in H_T, defined as the scalar sum of the transverse momenta of all the jets in the event. The main background coming from the $t\bar{t}$ associated production has been simulated with the powheg generator [69], interfaced with pythia8. In addition to this nominal sample, two additional $t\bar{t}$ samples have been used, which have a cut on the invariant mass of the $t\bar{t}$ system at parton level ($m_{t\bar{t}}$). These samples are obtained through powheg and pythia8 as well. The usage of such samples will be discussed in the following. With the goal of estimating the impact of the uncertainties affecting the MC parameters of the $t\bar{t}$ simulation, additional samples are used. The only difference with respect to the nominal sample is the value of the MC parameters, which are shifted according to their uncertainties. The irreducible background of $t\bar{t}Z$ events has been simulated using MadGraph5_aMC@NLO, once again interfaced with pythia8.

Concerning the subdominant backgrounds, several sources are taken into account: the production of a single top quark (or antiquark) in the t-channel and in the tW channel, as well as the production of a pair of W bosons

decaying to four quarks, have been simulated with POWHEG. Drell-Yan (DY) events in which the virtual photon or Z boson decays to a pair of quarks, as well as the production of a W boson decaying to quarks plus additional jets, have been simulated with MADGRAPH. Finally, the $t\bar{t}W$ production, in which the W boson decays to quarks, and the production of a pair of Z bosons decaying to four quarks, have been simulated with MADGRAPH5_aMC@NLO. All the aforementioned generators have been coupled to PYTHIA8 for the simulation of the parton shower. A comprehensive list of all the MC samples used in this analysis, together with the corresponding generator, theoretical cross section and number of generated events, is reported in Table 4.2.

Next-to-leading-order predictions obtained with the use of the MADGRAPH5_aMC@NLO generator lead to weighted MC events. This means that, when filling the distribution of a given observable, events do not actually count as one, but rather count as their event weight, and the integral of the resulting distribution will not be equal to the number of times the histogram has been filled. For such kind of events, a comparison has been made between weighted and unweighted shapes, the latter being the shapes obtained by simply filling the distributions with events counting as one, *i.e.*, neglecting the event weight. As a result, the weighted and unweighted shapes have generally been found to be compatible within the statistical uncertainties, with the unweighted shapes showing smaller error bars. Therefore, in the final fit performed in this analysis, for all the weighted samples, unweighted shapes are used, which are normalized to the weighted yield (*i.e.*, to the yield of weighted shapes). An analogous argument holds for $t\bar{t}$ events, where the shapes obtained from the samples with the $m_{t\bar{t}}$ cut are found to be more populated and to be compatible, within the uncertainties, with the shapes obtained from the nominal file. Thus, $t\bar{t}$ shapes included in the final fit are obtained from the $m_{t\bar{t}}$ samples and normalized to the yield obtained from the nominal file.

Monte Carlo samples are generated with distributions of the number of pileup (PU) interactions, *i.e.*, the number of multiple proton-proton collisions within the same bunch crossing, which are meant to roughly cover, though not exactly match, the actual conditions of a given data-taking period. The distribution of the number of reconstructed primary vertices (PVs) is sensitive to details in the PV reconstruction, to differences in the description of the underlying event in data and MC, and potentially to effects coming from the offline selection criteria. Thus, the PU distribution in MC must be reweighted with a proper pileup scale factor (SF) to match the corresponding distribution in data. In order to account for all the previously described effects, instead of reweighting the MC by the number of reconstructed vertices, we choose instead to reweight by the number of pileup true interactions, as stored in the simulations. The corresponding distribution in data is obtained starting from the instantaneous luminosity distribution and the total proton-proton inelastic cross section, which is taken to be 69.4 mb. Finally, the SF is

Sample	Generator	Events ($\times 10^6$)	σ (pb)
$t\bar{t}H(b\bar{b})$	MadGraph5_aMC@NLO	9.8912	0.2934
$t\bar{t}H(nob\bar{b})$	MadGraph5_aMC@NLO	10.045	0.2151
QCD ($300 < H_T < 500$ GeV)	MadGraph	54.537	3.477×10^5
QCD ($500 < H_T < 700$ GeV)	MadGraph	62.271	3.21×10^4
QCD ($700 < H_T < 1000$ GeV)	MadGraph	45.412	6831
QCD ($1000 < H_T < 1500$ GeV)	MadGraph	15.127	1207
QCD ($1500 < H_T < 2000$ GeV)	MadGraph	11.827	119.9
QCD ($2000 < H_T < \infty$ GeV)	MadGraph	6.039	25.24
$t\bar{t}$	powheg	77.0811	832
$t\bar{t}$ ($700 < m_{t\bar{t}} < 1000$)	powheg	38.4226	69.64
$t\bar{t}$ ($1000 < m_{t\bar{t}} < \infty$)	powheg	24.563	16.74
$t\bar{t}$ (tune up)	powheg	58.954	832
$t\bar{t}$ (tune down)	powheg	29.984	832
$t\bar{t}$ (isr up)	powheg	156.179	832
$t\bar{t}$ (isr down)	powheg	149.763	832
$t\bar{t}$ (fsr up)	powheg	152.618	832
$t\bar{t}$ (fsr down)	powheg	155.992	832
$t\bar{t}$ (hdamp up)	powheg	58.859	832
$t\bar{t}$ (hdamp down)	powheg	57.764	832
$t\bar{t}Z$	MadGraph5_aMC@NLO	0.749	0.5297
Single-top (t-channel)	powheg	67.241	136.02
Single-antitop (t-channel)	powheg	38.811	80.95
Single-top (tW)	powheg	6.953	35.6
Single-antitop (tW)	powheg	6.933	35.6
Drell-Yan	MadGraph	12.055	1460
W+jets	MadGraph	22.402	3539
WW	powheg	1.998	51.723
ZZ	MadGraph5_aMC@NLO	30.454	22.29
$t\bar{t}W$	MadGraph5_aMC@NLO	0.833	0.4062

Table 4.2: Monte Carlo samples used in the analysis. The kind of process, the generator used to simulate it, the number of simulated events and the theoretical cross section are reported.

obtained as the ratio between the pileup distributions in data and in MC.

4.2 Object reconstruction

4.2.1 Jets

Given the targeted fully hadronic final state, jets are the most used and important objects of the analysis. Since they are the result of the hadronization of quarks, jets are effectively build up starting from large collections of objects that are reconstructed in the detector. Indeed jets can first be categorized based on the type of input objects and three different categories of reconstructed jets exist in CMS:

1. Calorimetric jets (CaloJets): such jets are clustered starting from calorimetric-only information;

2. Jet-plus-track jets (JPTJets): such jets are clustered starting from calorimetric energy depositions, where an algorithm corrects the energy

of a jet using the momentum of charged particles measured in the tracker;

3. Particle-flow jets (PFJets): such jets are clustered starting from particles (often referred to as "candidates") which have been reconstructed by the particle-flow (PF) algorithm [70].

The analysis presented in this work uses jets clustered from PF candidates, *i.e.*, PFJets. In the PF event reconstruction all the stable particles, *i.e.*, muons, electrons, photons and charged and neutral hadrons, are reconstructed exploiting information coming from all the CMS subdetectors, resulting in an optimal determination of energies, directions and types. The building blocks of the PF event reconstruction are charged-particles tracks, calorimeter clusters and muon tracks. In order for these object to be delivered with a high efficiency and low fake rates even in high-density environments such as the LHC collisions, advanced tracking and clustering algorithms are used.

As far as tracks are concerned, an iterative-tracking strategy is adopted, where in the first step the tracks are reconstructed with tight criteria, which lead to a moderate efficiency but low fake rates; then, in the following iterations, hits unambiguously assigned to the tracks found in the previous iterations are removed, and the track seeding criteria are progressively made looser. This way, the efficiency increases and the fake rate is kept low, due to the low combinatorics of the hits.

Moreover, a specific calorimeter clustering algorithm has been developed for the PF event reconstruction. First, "cluster seeds" are formed from local calorimetric deposits above a given energy threshold; then, "topological clusters" are formed by aggregating cells with at least one side in common with a seed already in the cluster. Finally, a "particle-flow cluster" is formed from each seed and an iterative procedure determines the clusters energies and positions.

A given particle is expected, in general, to give rise to more than one PF object: one charged-particle track, and/or one or more calorimetric clusters, and/or a muon tracks. These elements are connected by a link algorithm to fully reconstruct each particle. The link algorithm is performed for each pair of elements in the event and defines a distance between any two linked elements to quantify the quality of the link. The algorithm then produces "blocks" of elements linked directly or indirectly. Given the high granularity of the CMS detectors, blocks typically contain only up to three elements, and constitute simple inputs for the particle reconstruction and identification algorithm which is described in full detail in [70].

A quantitative illustration of the performances of the PF algorithm can be found in Fig. 4.1. Using simulations, reconstructed CaloJets and PFJets are matched to the closest jet among the generated jets (GenJet), *i.e.*, jets clustered from all the stable particles, except for neutrinos. To establish a match, a distance in the η-ϕ plane of less than 0.1 is required.

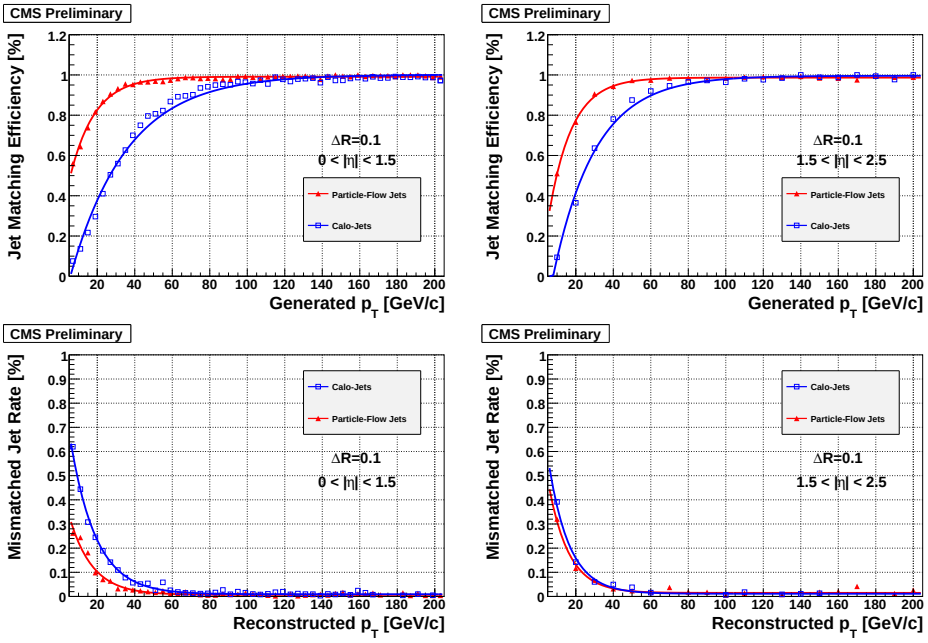

Figure 4.1: Matching efficiency between reconstructed and generated jets in the barrel region (upper left) and endcap region (upper right); mismatch rate between between reconstructed and generated jets in the barrel region (lower left) and endcap region (lower right).

The matching jet efficiency, *i.e.*, the fraction of GenJets that give rise to a matched reconstructed jet, and the mismatched jet rate, *i.e.*, the fraction of reconstructed jets that do not have a matched generated jet, are plotted, as a function of the jet p_{T}, in the barrel ($0 < |\eta| < 1.5$) and the end-cap ($1.5 < |\eta| < 2.5$) regions. As a result of the usage of the PF algorithm, jet matching efficiencies are found to be significantly higher than the ones obtained by the simpler CaloJets, while at the same time the mismatch rate is lower, especially in the barrel region.

For each given type of object which serves as an input to the jet clustering procedure, different possible ways of performing such clustering exist. Jet reconstruction algorithms are among the most useful tools used in particle physics in order to analyze data from hadronic collisions. Theoretical QCD calculations provide results in terms of quarks and gluons in the final state which, once produced in the detectors, undergo hadronization and give rise to the characteristic signature of QCD jets. A quantitative mapping between hadrons in the form of jets and final state quarks and gluons is then required in order to correctly interpret the theoretical results and compare them with the experimental evidence. This mapping is provided by jet reconstruction algorithms. A wide variety of algorithms has been developed over time, differing from one another by the internal parameters used to reconstruct a

jet. However, a good algorithm should show some fundamental properties: it should be fast and easy to implement during data analysis and it should show the key feature called infra-red and collinear (IRC) safety, meaning that collinear gluon splitting and soft radiation should not change the set of jets in the event. All the jet reconstruction algorithms widely used at the present time in CMS are IRC safe and belong to the class of the so called sequential recombination jet algorithms. They are based on some kind of definition of how likely two partons arise from a QCD splitting, and they proceed sequentially to construct a jet evaluating which partons are close in this measure. Thus, jets can further be categorized based on the clustering algorithm:

1. k_T algorithm (KT) jets [71];
2. Anti-k_T algorithm (AK) jets [72];
3. Cambridge-Aachen (CA) jets [73].

The jets used in this analysis are reconstructed with the anti-k_T algorithm. First, the following two quantities are defined:

$$\begin{aligned} d_{ij} &= \min(p_{\mathrm{T},i}^{-2}, p_{\mathrm{T},j}^{-2})\frac{\Delta R_{ij}^2}{R^2} \qquad \Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2 \\ d_{iB} &= p_{\mathrm{T},i}^{-2}, \end{aligned} \tag{4.1}$$

where d_{ij} is the "distance" between particle i and particle j, d_{iB} is the distance between particle i and the beam, R is the so called distance parameter, y_i and ϕ_i are the rapidity [1] and azimuthal angle of particle i. Then the procedure starts and all the distances d_{ij} and d_{iB} are evaluated from the list of all the final-state particles. The minimum distance is found and if it is a d_{ij}, particles i and j, are recombined together and the evaluation of distances is repeated; if, instead, the minimum distance is found to be d_{iB}, particle i is declared to be a jet, is removed from the final state and the procedure starts again. The algorithm stops when no particles remain.

Depending on the distance parameter used to cluster the jets, AK jets can further be categorized based on the jet size. The present analysis uses AK jets clustered with distance parameters of 0.8 and 0.4, which are referred to as AK8 and AK4 jets respectively. AK8 jets are well suited to collect boosted decays of particles, whereas AK4 jets are used to collect resolved decays. Since there is a non negligible probability for an AK4 jet to be reconstructed as an AK8 jet (and vice versa), the geometrical matching of AK4 and AK8 jets in the $\eta - \phi$ space is checked and any AK4 jet showing a $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} < 0.4$ from any accepted AK8 candidate is removed

[1] The rapidity is defined as $y = 0.5 \times \ln[(E + p_z)/(E - p_z)]$

from the AK4 collection. This way, the two jet collections are disjoint, meaning that no object can be ambiguously counted twice.

The clustered jets, especially the large-radius ones, can suffer from contamination of particles coming from PU. This contribution can be reduced by using PU mitigation algorithms. In this analysis, the charged-hadron subtraction (CHS) method [74] is used for both the AK4 and AK8 collections. In this method, the interaction vertices are ordered by the quadratic sum of the transverse momenta of their tracks, $\sum p_{\mathrm{T}}^2$, and the vertex showing the highest $\sum p_{\mathrm{T}}^2$ is considered to be the primary vertex, while the others are considered to be PU vertices. Charged hadrons coming from PU are removed from the list of PF candidates used to cluster the jets. In addition, the jet p_{T} is corrected to take into account the PU contribution coming from photons, neutral hadrons and particles out of the tracker acceptance.

As it has been already pointed out, an important feature of analyses in the boosted regime is that, for particles having enough Lorentz boost, all the decay products can be collected in a single, large-radius jet, so that the properties of the decaying particles can be inferred directly from the substructure of such jets. A key feature of boosted decays is the pattern of energy deposits inside a jet. In fact, decays coming, for example, from top quarks decays, will mostly produce AK8 jets which contain three energy deposits, while AK8 jets coming from QCD multijet production will more likely contain a smaller number of such deposits. This suggests the possibility of using the number of prongs inside an AK8 jet to discriminate between ordinary QCD jets and jets coming from Higgs boson or top quark decays. Thus, an important variable to study the substructure of AK8 jets is the n-subjettiness variable τ_i [75], defined as:

$$\tau_i = \frac{1}{\sum_k p_{\mathrm{T},k} R} \sum_k p_{\mathrm{T},k} \min\left(\Delta R_{1k}, \Delta R_{2k}, ...\Delta R_{ik}\right), \tag{4.2}$$

where the index k enumerates the constituents of the input jet, $p_{\mathrm{T},k}$ is the p_{T} of the k-th constituent, R is the distance parameter of the original jet, and ΔR_{ik} is the angular distance in the $\eta - \phi$ space between the i-th subjet and the k-th constituent. The n-subjettiness variable τ_i measures the compatibility of a jet with the hypothesis of containing i prongs. In the case of exactly i prongs, the value of τ_i tends to zero. So, n-subjettiness ratios are found to be powerful variables to discriminate between QCD jets and multi-prong decays. Figure 4.2 shows the ratios τ_3/τ_2 and τ_3/τ_1 for simulated QCD and $\mathrm{t\bar{t}}$ events, where some general selection criteria have been asked to select boosted-, all-jets-like events. We see that both variables are found to have a considerable discriminating power to reject the QCD background.

In order to compute the invariant mass of AK8 jets, the soft drop declustering algorithm [76] is used. This method is able to perform jet grooming by removing wide-angle, soft radiation from a jet thus mitigating

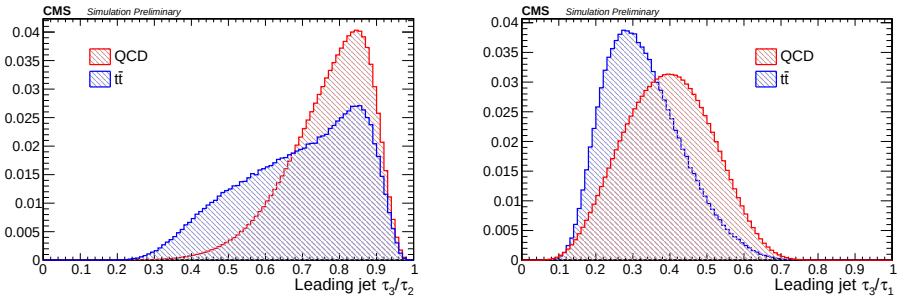

Figure 4.2: n-subjettiness ratios for QCD and $\mathrm{t\bar{t}}$ events. Both τ_3/τ_2 and τ_3/τ_1 are found to have a non-negligible discriminating power between QCD and $\mathrm{t\bar{t}}$ jets.

effects from initial state radiation (ISR), underlying event (UE) and pileup. In the soft drop procedure, the clustering algorithm that reconstructed a jet j is reverted step by step, decomposing j into two subjets, j_1 and j_2, at each iteration. If the subjets fulfill the soft drop condition

$$\frac{\min(p_{\mathrm{T}_1}, p_{\mathrm{T}_2})}{p_{\mathrm{T}_1} + p_{\mathrm{T}_2}} > z_{\mathrm{cut}} \left(\frac{\Delta R_{12}}{R} \right)^{\beta}, \tag{4.3}$$

then j is declared to be the final jet; otherwise j is redefined to be the leading subjet and the procedure is iterated. The degree of grooming is defined by the two parameters of the algorithm, z_{cut} and β, controlling the strength of the fractional p_{T} selection and the suppression of collinear radiation, respectively. In this analysis, the default CMS values $z_{\mathrm{cut}} = 0.1$ and $\beta = 0.0$ are used. Note that, given the nature of the soft drop condition, the action of the algorithm always results in exactly two subjets.

Jets that are likely to come from the hadronization of bottom quarks are identified with the combined secondary vertex version 2 (CSVv2) b tagging algorithm [77]. This algorithm exploits the long mean life of b-flavored hadrons present in jets originating from the hadronization of b quarks (b jets) and is able to combine in an optimal way the information coming from the tracking system, such as the number of tracks, their pseudorapidity and impact parameters, to identify secondary vertices generated by the decay of such b-flavored hadrons. A pictorial view of a b quark-induced jet is shown in Fig. 4.3. The algorithm provides a continuous discriminator output, which states how likely a jet is indeed a b jet. Cuts placed on the value of such discriminator define several working points for the algorithm, corresponding to different b tagging efficiencies and mistag probabilities (*i.e.*, the probability of tagging a light-flavored jet as a b jet); in the present analysis, the medium working point is chosen. As far as large-radius jets are concerned, we consider AK8 jets to be b tagged if we are able to find at least one b tagged subjet inside them.

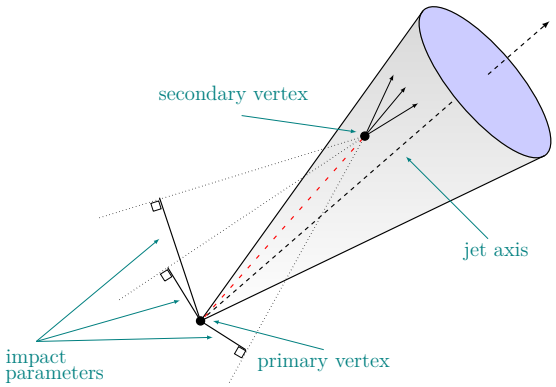

Figure 4.3: Pictorial representation of a b jet. Due to its long mean life, a b-flavored hadron leaves the primary vertex and travels some distance before decaying, thus creating a displaced secondary vertex. The impact parameters of the b hadron track are shown.

In general, the b tagging efficiency and the mistag probability are found to be different when computed in data and in simulation. Thus, scale factors are needed in order to correct the simulation to match the data. In this analysis, the number of b tagged jets in the event is used to categorize signal events; also, the multivariate methods that have been developed to tag the candidates use CSVv2 scores as input variables. Thus, scale factors describing the discrepancies in the b tagging algorithm performance between data and simulations must affect both the MC expected yields and the overall shape of the discriminators. This is achieved with a tag-and-probe approach through the IterativeFit method [78]. This approach relies on samples of events with two high-p_{T}, charged leptons and exactly two AK4 jets. Requests made on the lepton pair, event variables and one of the jets, called tag jet, are exploited to select samples either enriched in dilepton $\mathrm{t\bar{t}}$ events, used to derive a heavy-flavor SF, or Z+jets events, used to derive a light-flavor SF. No requirements are made on the second jet, called probe jet, which is used to extract the SF by comparing its CSVv2 distribution in data to that predicted by MC. These SFs are determined separately for exclusive bins of CSVv2 score, p_{T} and η of the jet. In the absence of a data-driven calibration sample for charmed jets, the scale factor for such jets is set to 1.0. Since the present analysis uses both AK4 and AK8 jets, we apply the previously described SF to AK4 jets and to the subjets which are found inside AK8 jets. That is, the total scale factor for the event is the product of all the scale factors of all the AK4 jets and subjets:

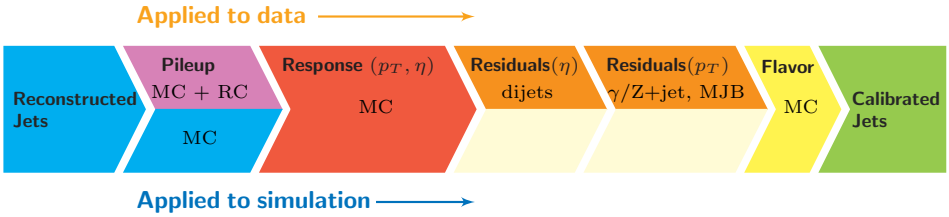

Figure 4.4: Consecutive stages of JEC, for data and MC simulation. All corrections marked with MC are derived from simulation studies, RC stands for random cone, and MJB refers to the analysis of multijet events. (Figure and caption from [79]).

$$\mathrm{SF}_{\mathrm{total}} = \prod_{i}^{N_{\mathrm{jets}}} \mathrm{SF}_{\mathrm{jet}_i} = \prod_{j}^{N_{\mathrm{AK4jets}}} \mathrm{SF}_{\mathrm{AK4jet}_j} \times \prod_{k}^{N_{\mathrm{subjets}}} \mathrm{SF}_{\mathrm{subjet}_k}. \tag{4.4}$$

Like all the experimentally-reconstructed objects, jets need to be calibrated to have the correct energy scale. The purpose of jet energy corrections (JEC) is to relate, on average, the energy of jets measured in the detector to the energy of true particle jets, *i.e.* jets originating from the clustering of all stable particles coming from the fragmentation of partons. In fact, effects of detection efficiencies, measurements resolution and systematic biases can lead to non negligible differences between true jets and jets reconstructed after the interaction of particles with the detector. Thus, corrections are applied to the jets 4-momenta in order to calibrate the jet energy scale (JES). Also, the simulated jet energy resolution (JER), namely the spread in the jet response, defined as the ratio between the p_{T} of reconstructed and GenJets, is found to be better than the one measured in data using methods such as the p_{T}-asymmetry method in dijet events. Thus, a smearing of the jets p_{T} in the simulation must be performed in order for the MC events to match the data.

The CMS Collaborations adopts a factorized approach to the problem of JEC [79], where different levels of correction are used. Each level of correction is basically a scaling of the jet four-momentum, which may depend on different jet-related quantities, such as the p_{T}, η, flavor, etc. The different levels of correction are applied in a precise sequence, which is visually shown in Fig. 4.4, with the output of each step being the input of the following one.

In the first level (L1) of corrections, PU offset corrections are applied in order to remove the energy coming from PU particles. Such corrections are determined from the simulation of a sample of QCD dijet events processed with and without pileup overlay.

In the second level (L2L3), corrections to the simulated jet response are determined by making use of QCD dijet events, by comparing the

reconstructed p_{T} to the particle-level one. The corrections are derived as a function of the jet p_{T} and η.

In the third step (L2L3 residuals), the remaining small differences in the jet response between data and MC are accounted for. Such corrections are obtained as a function of the jet p_{T} and η by making use of dijet, Z/γ plus jets and multijet events.

Finally, in the fourth step (flavor corrections), a correction related to the difference in the jet flavor which is present in the samples used to compute L2L3 residual corrections is applied. This takes into account the fact that QCD dijet and multijet events are enriched in gluon-initiated jets, while Z/γ plus jets events are enriched in quark-initiated jets.

A visualization of the effect of JEC is presented in Fig. 4.5. The average value of the jet response at various stages of the JEC procedure is shown for simulated QCD multijet events measured at central rapidities ($|\eta| < 1.3$) and in bins of GenJet transverse momentum $p_{\mathrm{T,\ ptcl}}$. Distributions corresponding to different average numbers of PU interactions, indicated by μ, are shown separately, to display the dependence of the response on the level of PU. It is evident that, without any correction, the responses diverge as a function of the number of PU collisions, especially at low jet p_{T}, where the fraction of pileup particles inside a jet is expected to be higher; when PU corrections are applied, the divergence is removed, even though the response is still somewhat far from one, especially at low jet p_{T}; after all the correction are applied, the response is fairly equal to one over all the jet p_{T} spectrum and for all the levels of pileup.

4.2.2 Leptons

Even though no leptons are present in the final state targeted by this analysis, isolated leptons are used for vetoing purposes. PF leptons (electrons and muons) are reconstructed using the default collections present in the CMS software. Electrons are identified with the so called tight electron ID, a set of simple and robust cuts on ECAL variables. Muons are identified with the so called medium muon ID, a set of identification criteria designed to be highly efficient for prompt muons and for muons from heavy quark decays and, at the same time, to keep the rejection of fake muons fairly high. To select isolated leptons, the mini-isolation variable $I^{\mathrm{rel}}_{\mathrm{mini}}$, originally suggested in [80], is used, which is defined as the sum of the transverse energy of charged hadrons, neutral hadrons and photons contained in a variable-size cone around the lepton direction, divided by the lepton p_{T}. The cone size scales with the lepton p_{T} as

$$\Delta R(p_{\mathrm{T},\ell}) = \frac{10\ \mathrm{GeV}}{\min\Big[\max\left[p_{\mathrm{T},\ell}, 50\ \mathrm{GeV}\right], 200\ \mathrm{GeV}\Big]}. \tag{4.5}$$

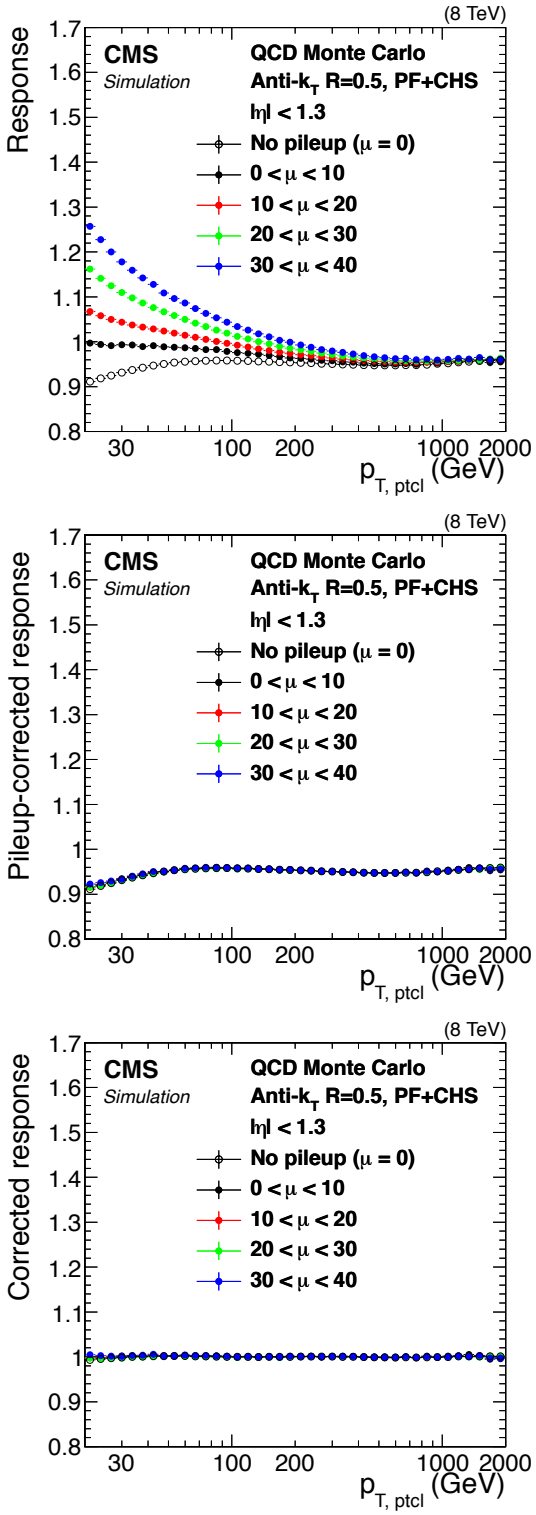

Figure 4.5: Average value of the jet response, as a function of the GenJet transverse momentum $p_{\mathrm{T,\ ptcl}}$ and for different levels of pileup, at different stages of the JEC procedure: before any corrections (upper panel), after the pileup offset correction (middle panel) and after all the corrections (lower panel).

Both electrons and muons are requested to have a mini-isolation $I^{\text{rel}}_{\text{mini}} < 0.1$. Finally, in order to guarantee the lepton isolation, which could be spoiled by the misidentification of leptons as jets, the lepton collection is cleaned from the jet collections by removing each reconstructed lepton showing $\Delta R < 0.4$ with either an AK8 or AK4 jet.

4.3 Signal trigger

The choice of a proper signal trigger is the first, crucial step towards a successful analysis. The chosen trigger should guarantee a good background rejection and, at the same time, an adequate signal efficiency. In order to find the best trigger among all the unprescaled triggers available in the JetHT menu, we use simulated events belonging to the $t\bar{t}H(b\bar{b})$ signal and to the two dominant backgrounds, namely the QCD multijet production and the $t\bar{t}$ associated production. A pretty general selection is applied to such events, in order to pick boosted-, fully hadronic-like events: we request the presence of at least one AK8 jet, with the leading one having $p_{\text{T}} > 300$ GeV, and we veto the presence of leptons. A useful quantity to perform this study is found to be the selection efficiency given the request of trigger j, ϵ_j, which, for every sample, is defined as

$$\epsilon_j = \frac{N_{\text{pass},j}}{N_{\text{gen}}}, \tag{4.6}$$

where $N_{\text{pass},j}$ is the number of simulated events that pass the aforementioned selection requirements and the requirements of trigger j, while N_{gen} is the number of generated events in a given sample. Since we are looking for a trigger with high signal efficiency and, at the same time, low background efficiency, we compute the efficiency given by Eq. 4.6 both for the signal and background samples and we choose the ratio $\epsilon_{j,\text{sig}}/\epsilon_{j,\text{bkg}}$ as a figure of merit for the wanted trigger. Eventually, the trigger showing the highest $\epsilon_{j,\text{sig}}/\epsilon_{j,\text{bkg}}$ ratio is found to be `HLT_AK8PFHT700_TrimR0p1PT0p03Mass50`. At L1 it is seeded by the logical OR of six trigger bits: `L1_HTT240`, `L1_HTT255`, `L1_HTT270`, `L1_HTT280`, `L1_HTT300` and `L1_HTT320`, which are satisfied if the online H_{T} exceeds the specified values (in GeV). At HLT, the decision is performed in three steps: first, a CaloJet filter is applied, requiring for the scalar sum of the momenta of AK8 jets in the event, H_{T}, reconstructed from calorimetric information only, to be greater than 600 GeV. For events passing the CaloJet filter, the full PF algorithm is run and PFJets are reconstructed. As a second step, the event H_{T}, reconstructed from PFJets, is required to be greater than 700 GeV. Finally, the presence of at least one AK8, PF jet with trimmed mass greater than 50 GeV is required.

Once a signal trigger has been chosen, it is important to study its properties, and in particular its efficiency as a function of event variables.

The trigger efficiency could in principle be defined as the ratio between the number of events passing the trigger requirements and some offline criteria and the number of events passing the same offline criteria; however, this definition does not make sense for data, which are always collected with a given trigger, implying that the denominator would not be a meaningful quantity for data events. Thus, it is customary to compute the trigger efficiency using events collected with some reference trigger. This reference should collect events with looser and, if possible, orthogonal criteria. Eventually, the following formula is used to describe the trigger efficiency as a function of some event variable x:

$$\varepsilon(x) = \frac{N_{\text{trig., offl., ref.}}(x)}{N_{\text{offl., ref.}}(x)}, \tag{4.7}$$

where $N_{\text{trig., offl., ref.}}(x)$ is the number of events passing the signal trigger requirements, the offline criteria and the reference trigger requirements, while $N_{\text{offl., ref.}}(x)$ is the number of events passing the offline criteria and the requirements of the reference trigger. As a set of offline selection requirements for this study, we use the requests that were made for the selection efficiency given by Eq. 4.6. The simulated events used here come from QCD and $\mathrm{t\bar{t}}$ production, mixed according to the corresponding theoretical cross sections reported in Table 4.2. The trigger efficiency has been measured with respect to the reference trigger `HLT_AK8PFJet200`, which requires the presence of an AK8 jet with $p_{\mathrm{T}} > 200$ GeV, as a function of the event H_{T}, the leading AK8 jet p_{T} and, for events showing at least two AK8 jets, the second-leading AK8 jet p_{T}. Figure 4.6 shows the outcome of the study. As a result, the reference is found to be a kinematically unbiased reference, since the trigger efficiency completely overlaps with the true (*i.e.*, computed with respect to no reference trigger) MC efficiency. Additionally, the trigger efficiency has been measured in data with respect to the same reference trigger.

Once the unbiased nature of the reference has been verified, a final study has been performed in order to find under which conditions the signal trigger is fully efficient. Given the fully hadronic final state targeted by this analysis, it is possible to find events with a large hadronic activity and low AK8 multiplicity. In this sense, the previously-developed study of the trigger efficiency as a function of H_{T} does not seem to be optimal, while a better choice can be to compute the efficiency as a function of the event S_{T}, namely the scalar sum of the momenta of all the jets present in the event: $S_{\mathrm{T}} = H_{\mathrm{T,AK8}} + H_{\mathrm{T,AK4}}$. Figure 4.7 shows such a final study. Once again, the reference trigger is found to be unbiased; also, by superimposing the S_{T} distribution for the signal events to the trigger efficiency, we see that events with $S_{\mathrm{T}} > 900$ GeV guarantee a high trigger efficiency and, at the same time, constitute a great fraction of the signal. Thus, our analysis will place a cut on the event S_{T} considering only events with $S_{\mathrm{T}} > 900$ GeV. Finally, since a small deviation is found in the trigger efficiency when computed using data

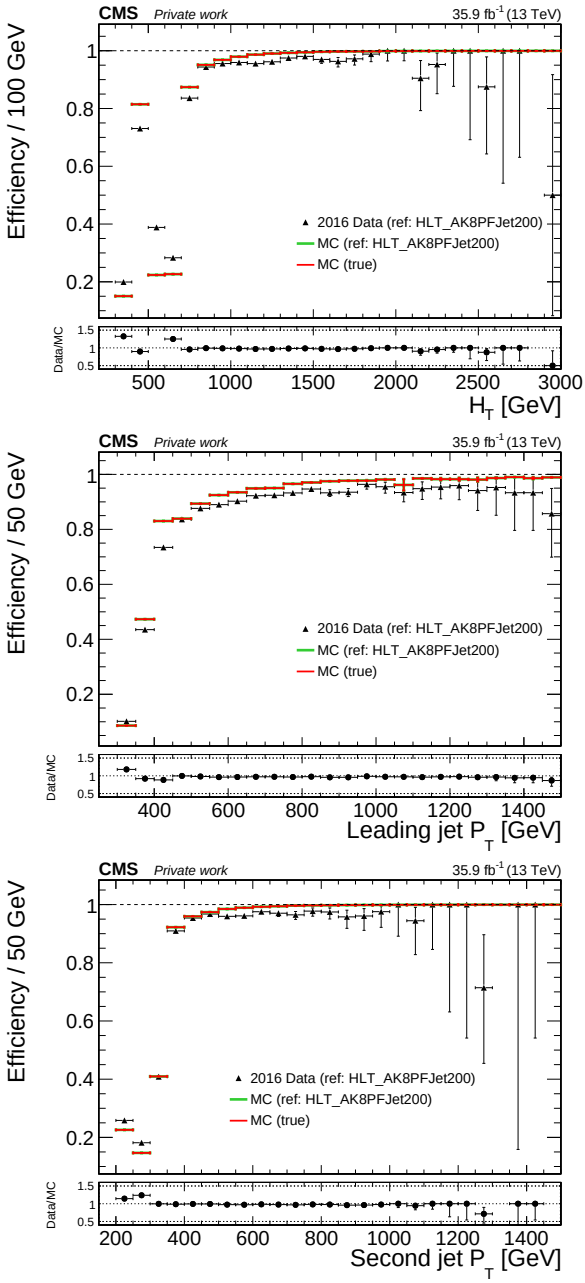

Figure 4.6: Trigger efficiency as a function of the event H_{T} (upper plot), the leading AK8 jet p_{T} (middle plot) and, for events showing at least two AK8 jets, the second-leading AK8 jet p_{T} (lower plot). The efficiency is computed in MC events with respect to the reference trigger `HLT_AK8PFJet200` (green line), in MC events with respect to no reference trigger (red line) and in data (black dots). Since the true MC efficiency completely overlaps with the efficiency computed with respect to the reference, we conclude that the reference is unbiased.

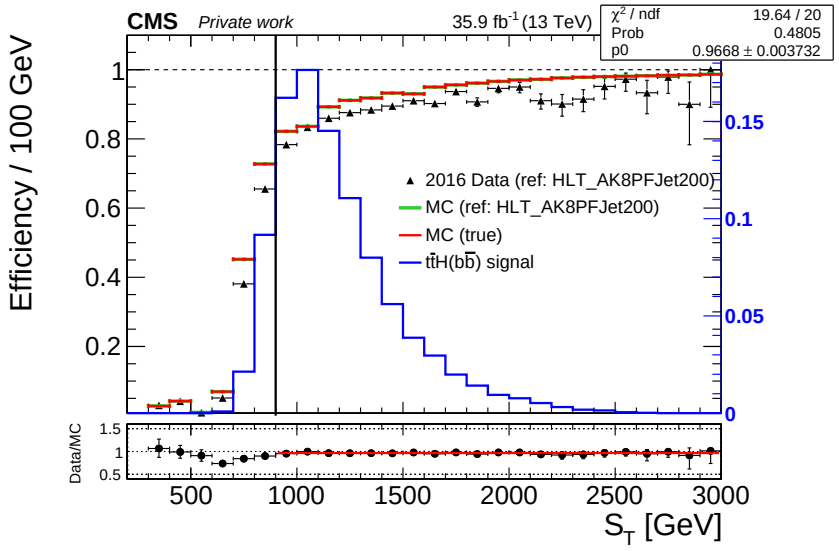

Figure 4.7: Trigger efficiency as a function of the event S_{T}. The efficiency is computed in MC events with respect to the reference trigger `HLT_AK8PFJet200` (green line), in MC events with respect to no reference trigger (red line) and in data (black dots). The reference trigger is found to be unbiased. The S_{T} distribution of signal events (blue line) is shown as well. The solid, vertical line corresponds to the offline $S_{\mathrm{T}} > 900$ GeV cut which makes the trigger fully efficient. The bottom panel shows the ratio between the efficiencies in data and MC, along with the fitting constant line which is used as a scale factor for the simulation to match the data.

and MC, we fit the ratio plot in Fig. 4.7 to obtain a scale factor that can be applied to the simulations to match the distribution of the data. Since the ratio is almost flat, the fitting curve has been chosen to be a constant line.

4.4 Event selection

4.4.1 Baseline selection

As a first step towards a robust and complete event selection targeting boosted, fully hadronic $\mathrm{t\bar{t}H}$ events, a baseline selection is set up. In order to enhance the efficiency on boosted events, we first request the presence of at least one AK8 jet in the event with soft drop mass greater than 50 GeV; also, to ensure that genuine boosted events are selected, we request the p_{T} of the leading AK8 jet to be greater than 300 GeV. Secondly, to select events in the fully hadronic final state, we veto on the presence of leptons. Then, the events are required to fulfill the requirements of the signal trigger described in Section 4.3. Given the studies presented above, the

Requirement	**Purpose**
$N_{\text{AK8}} \geq 1$	Select boosted regime
Leading jet $p_{\text{T}}>$ 300 GeV	Select boosted regime
$N_{\text{leptons}} = 0$	Select fully hadronic final state
Signal trigger	Select boosted, fully hadronic final state
$S_{\text{T}} > 900$ GeV	Have a fully efficient trigger

Table 4.3: Summary of the baseline selection criteria.

events used in this analysis must also have $S_{\text{T}} > 900$ GeV in order to make the signal trigger fully efficient. A summary of the selection criteria forming the baseline selection is reported in Table 4.3, where we also briefly highlight the purpose of each requirement.

4.4.2 Boosted-jets BDTs

In the context of the boosted-Higgs channel, which makes part of the search of $\text{t}\bar{\text{t}}\text{H}$ events using large-radius jets, an important task is to identify AK8 jets coming either from Higgs bosons or top quarks decays and separate them by AK8 jets induced by ordinary QCD processes. This goal is achieved by means of multivariate methods that look at substructure variables of AK8 jets since, as it was already pointed out in Subsection 4.2.1, such variables have been found to be well suited for this task. Events passing the baseline selection are used to create three collections of jets: a first one, composed by AK8 jets that are matched with a Higgs boson in the $\text{t}\bar{\text{t}}\text{H}(\text{b}\bar{\text{b}})$ simulation; a second one, composed by AK8 jets that are matched with a top quark in the $\text{t}\bar{\text{t}}$ simulation; a third one, composed by AK8 jets coming from the QCD simulation. These jet collections are trained against each other to obtain the three BDTs called BDT_HvsQCD, BDT_TvsQCD and BDT_HvsT, which discriminate between Higgs-boson- and QCD-induced, top-quark- and QCD-induced and Higgs-boson- and top-quark-induced jets respectively.

The variables used in all the trainings are the invariant masses of the leading and second-leading subjets, their CSVv2 discriminant distributions and the n-subjettiness variables τ_1, τ_2 and τ_3.

Figure 4.8 shows the scores of the three BDTs for the "signal" and "background" distributions, where clearly the notion of "signal" and "background" changes for each training, *e.g.*, top-quark-induced jets are considered to be signal in the BDT_TvsQCD training and background in the BDT_HvsT training. In all the cases, a good discriminating power is found between signal and background jets.

The aforementioned BDTs are used to identify a boosted Higgs candidate, namely an AK8 jet induced by the boosted $\text{b}\bar{\text{b}}$ decay of a Higgs boson. The boosted Higgs candidate is defined to be the AK8 jet having the highest sum

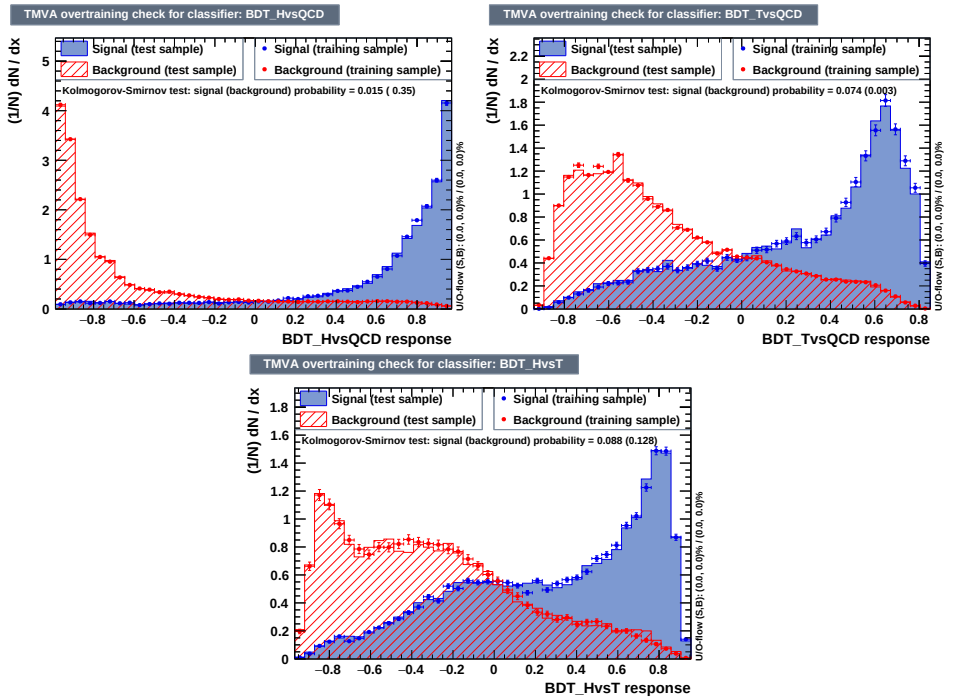

Figure 4.8: Signal and background distributions for the boosted-jets BDTs: BDT_HvsQCD distribution (upper left), BDT_TvsQCD distribution (upper right), BDT_HvsT distribution (lower row).

of the BDT_HvsQCD and BDT_HvsT scores; then, in order to enhance the purity of the candidate, lower bounds are put on each score, requesting the BDT_HvsQCD score to be greater than 0.8 and the BDT_HvsT score to be greater than 0.1. Finally, the selected jet must have a p_{T} greater than 300 GeV and a soft drop mass of at least 70 GeV.

Boosted Higgs candidates are used in the resolved-Higgs channel for vetoing purposes. In fact, since the targeted final state has a resolved decay of the Higgs boson, the logical negation of the aforementioned cuts is preformed and the absence of a boosted Higgs candidate is requested. Since, on the other side, events entering the boosted-Higgs channel must contain a boosted Higgs candidate, thanks to this request the two channels composing the search for $\mathrm{t\bar{t}H}$ events in the boosted, fully hadronic final state are made orthogonal by construction, thus making a combination possible.

4.4.3 Resolved-Higgs BDT

As we pointed out in the previous section, events entering this analysis must not contain a boosted Higgs candidate. Then, a way to identify $\mathrm{t\bar{t}H(b\bar{b})}$ events containing a resolved decay of the Higgs boson must be found. The

naive idea of just selecting events containing exactly two AK4, b tagged jets does not seem to be the optimal choice. In fact, events in which both the Higgs boson and one top quark decay in a resolved topology contain at least three AK4, b tagged jets, and thus are lost if the naive selection is assumed. This raises the problem of how to treat events with more than two AK4, b tagged jets and how to identify, among all the possible combinations, the correct pair of jets corresponding to the decay of the Higgs boson. As it will be discussed in the following, this is dealt with by the use of a resolved-Higgs BDT which makes possible, at the same time, to reject events belonging to the main backgrounds of the analysis and to choose the correct pair of jets arising from the decay of the Higgs boson.

Training

The simulated events used for the training of the BDT come from the $t\bar{t}H(b\bar{b})$ signal sample and from the two dominant backgrounds of the analysis, *i.e.* the QCD multijet production and the $t\bar{t}$ associated production. Both the signal and background events are requested to fulfill the baseline selection criteria described in Subsection 4.4.1 and to show exactly two AK4, b tagged jets. This selection is aimed to select boosted, all-jets -like events where a Higgs boson can be reconstructed and, at the same time, to guarantee an adequate background yield ensuring a sufficiently general training. The $t\bar{t}H(b\bar{b})$ events in which the generated Higgs boson is matched to the system of two AK4, b tagged jets within $\Delta R < 0.3$ are considered to be signal events. On the other hand, unmatched $t\bar{t}H(b\bar{b})$ events are considered to be background events. All QCD and $t\bar{t}$ events passing the aforementioned selection are considered to be background events as well.

Once the training selection is set up, a set of training variables showing high discriminating power between signal and background must be chosen. First, variables concerning the pair of b tagged jets are considered. The CSVv2 b tag scores of both the AK4 jets are used as input variables for the resolved-Higgs BDT. Kinematic variables related to the jet pair, such as the p_{T} of both jets, their mass and their ΔR distance are found to be powerful discriminating variables; nevertheless they are not used in the training as they are directly correlated with the invariant mass of the dijet pair, which will be the observable used to extract the final results of the analysis. This way we prevent the BDT to create artificial biases and peaks in the background distributions; also, the low correlation between the resolved-Higgs BDT score and the invariant mass of the dijet pair will turn out to be useful in the estimation of the QCD background, see Section 4.6. In any case, information on the p_{T} of the event is exploited by including the offline H_{T} in the set of training variables. Also, the number of AK8 and b tagged AK8 jets are included. Moreover, since the $t\bar{t}H$ events contain more final state particles than the dominant backgrounds, we expect the jets in the signal events to

be closer to each other with respect to the background. Thus, the distances between the first and second AK4, b tagged jet and the closest AK8 jet are used as discriminating variables. We also exploit some event variables and substructure variables of the leading AK8 jet. Eventually, the following set of variables is used:

1. b tag score of the first AK4, b tagged jet;
2. b tag score of the second AK4, b tagged jet;
3. number of AK8 jets;
4. number of AK8, b tagged jets;
5. scalar sum of the p_{T} of the AK8 jets;
6. minimum ΔR between the first AK4, b tagged jet and the AK8 jets;
7. minimum ΔR between the second AK4, b tagged jet and the AK8 jets;
8. event aplanarity;
9. event centrality;
10. event sphericity;
11. τ_3/τ_1 of the leading AK8 jet;
12. τ_3/τ_2 of the leading AK8 jet;
13. b tag score of the first subjet inside the leading AK8 jet;
14. b tag score of the second subjet inside the leading AK8 jet.

The centrality of the event is defined as the ratio of the scalar sum of the transverse momenta of all the jets (AK4 and AK8) to the invariant mass of the multijet system, $\sum p_{\mathrm{T}}/\sqrt{\hat{s}}$. Starting from the spatial components of the jets four-momenta, it is also possible to construct the sphericity tensor $M_{ab} = \sum_j p_{ja} p_{jb}$, calculated in the centre-of-mass of the multijet system, where $a, b \in \{x, y, z\}$ and the index j runs over all the jets (AK4 and AK8) of the event. The aplanarity of the event is defined as $3/2\ \mathcal{Q}_1$, where $\mathcal{Q}_1$ is the smallest of the three normalised eigenvalues of the sphericity tensor; in a similar fashion, the sphericity of the event is defined as $3/2\ (\mathcal{Q}_1 + \mathcal{Q}_2)$. The distributions of the training variables are shown, both for signal and background, in Fig. 4.9, while Fig. 4.10 shows the linear correlation coefficients between the variables, both for signal and background.

The resolved-Higgs BDT has been trained using the TMVA package [81]. Many different combinations of BDT type, number of trees in the forest and learning rate have been explored in order to find the most performing

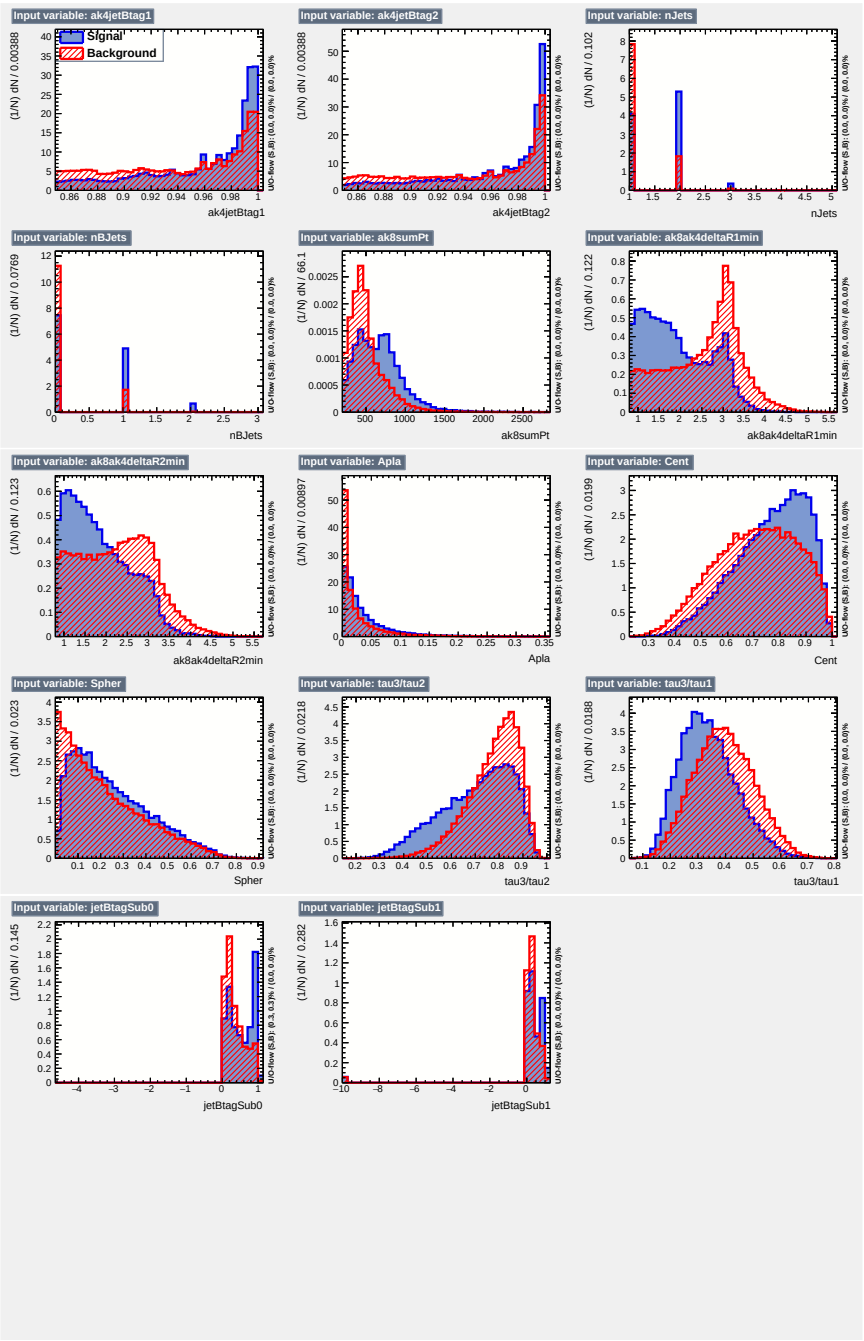

Figure 4.9: Distribution of the 14 training variables for signal and background events.

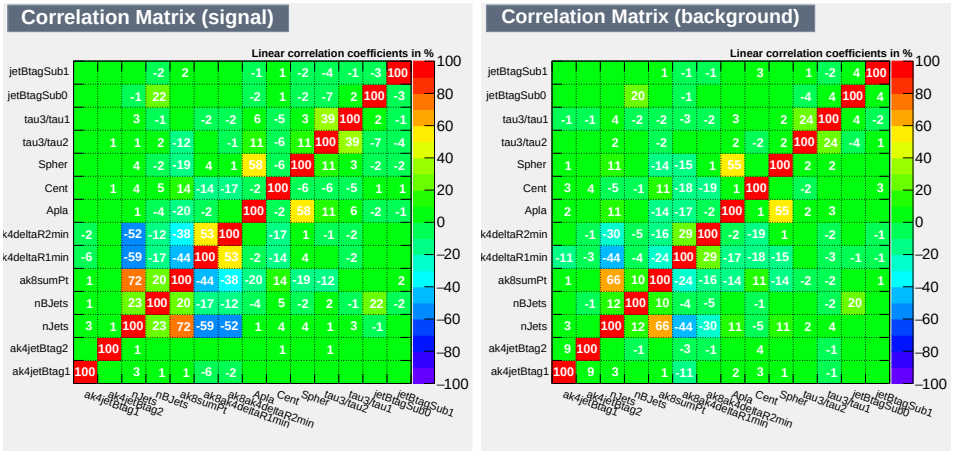

Figure 4.10: Linear correlation coefficients between the variables used to train the resolved-Higgs BDT for signal (left) and background (right) events.

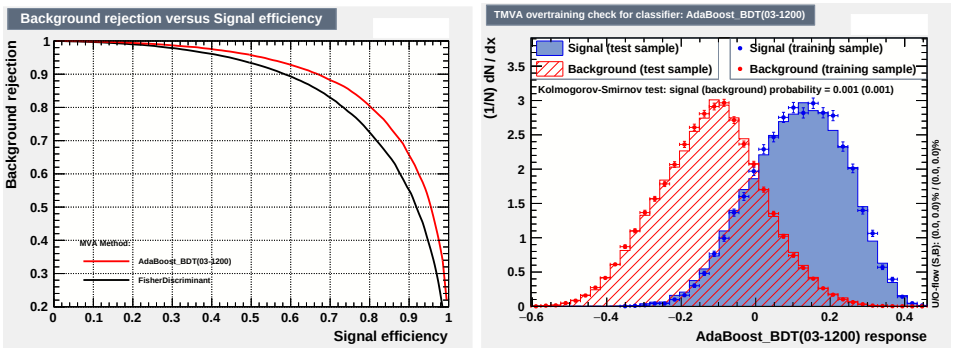

Figure 4.11: Background rejection vs. signal efficiency ROC curve for the resolved-Higgs BDT and the Fisher discriminant (left); signal and background distribution for the resolved-Higgs BDT score (right).

one. In the end, a BDT based on the Adaptive Boost method, composed of 1200 trees and with learning rate of 0.3 has been chosen. A simple Fisher discriminant has been trained as well, in order to check if there is a significant gain in exploiting more complex multivariate methods. Figure 4.11, showing the performance of the two discriminants expressed as a background rejection vs. signal efficiency ROC curve, testifies that the use of a BDT is justified by the enhanced performance. The performance of the BDT is quantified by the area under the ROC curve, which is found to be 0.886. Finally, Fig. 4.11 also shows the resolved-Higgs BDT score for signal and background events, where we see that a good discriminating power between signal and background events is achieved.

4.4.4 Higgs boson and top quark candidates

As a further step in the analysis selection, a way to identify jets corresponding to the Higgs boson and top quark decays is developed. In order to do so, the resolved-Higgs BDT and the BDTs which have been developed in the context of the boosted-Higgs channel are extensively used. All the cuts reported in the following have been optimized in order to maximize the analysis sensitivity and, at the same time, obtain a good agreement ("closure") in the QCD background estimation procedure (see Section 4.6).

Higgs boson candidate

The resolved Higgs boson candidate targeted by this analysis is identified through the resolved-Higgs BDT. First, events with at least two AK4, b tagged jets are considered. All the possible combinations of two, b tagged AK4 jets are taken into account and, for each of them, a resolved-Higgs BDT score is computed. This way, the dijet combinations can be ordered based on how likely they come from the resolved decay of a Higgs boson, with combinations showing larger scores having a greater likelihood to be the result of a Higgs boson decay, as Fig. 4.11 testifies. The resolved Higgs boson candidate is defined to be the combination of two, b tagged AK4 jets showing the highest resolved-Higgs BDT score, having it greater than 0.07 and lying in the invariant mass window $[70, 270]$ GeV. The Higgs booson candidate purity obtained with this selection criteria is evaluated in the $\mathrm{t\bar{t}H(b\bar{b})}$ simulated sample and is defined as the ratio of the number of resolved Higgs boson candidates matched to a Higgs boson to the total number of resolved Higgs boson candidates. This purity is found to be $\approx 33\%$ and is visually shown in Fig. 4.12.

Since the observable used to perform the final fit is the mass of the Higgs boson, events entering the signal selection of this analysis must contain a reconstructed Higgs boson candidate.

Top quark candidate

The top quark candidate is identified by making use of the boosted BDTs. It is defined to be the AK8 jet showing the highest BDT_TvsQ score, having $p_{\mathrm{T}} > 300$ GeV, BDT_TvsQ > 0.2, BDT_HvsT < -0.2 and lying in the soft drop mass (m_{SD}) window $[120, 240]$ GeV. The cut on the BDT_TvsQ score is aimed at rejecting the QCD background, while the cut on the BDT_HvsT score has been added in order to avoid the presence of a secondary Higgs boson peak in the soft drop mass distribution. In fact, since a boosted Higgs boson is required not to be present, actual boosted Higgs bosons that failed the identification may become top quark candidates and contaminate the invariant mass distribution of the top quark candidate. The top quark candidate purity obtained with this selection criteria is evaluated in the

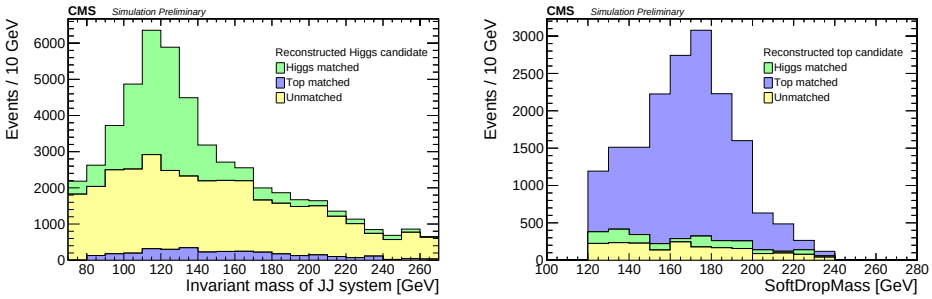

Figure 4.12: Invariant mass distribution of the resolved Higgs boson candidate (left), showing the fraction of resolved Higgs boson candidates which are correctly matched to a Higgs boson (green, corresponding to $\approx 33\%$ of the total), matched to a top quark (blue) or unmatched (yellow); soft drop mass distribution of the top quark candidate (right), showing the fraction of top quark candidates which are correctly matched to a top quark (blue, corresponding to $\approx 83\%$ of the total), matched to a Higgs boson (green) or unmatched (yellow area).

Higgs boson candidate	**Top quark cadidate**
$N_{\text{AK4bjets}} \geq 2$	Highest BDT_TvsQCD score
Highest resBDT score combination	$p_{\text{T}} > 300$ GeV
resBDT score ≥ 0.07	BDT_TvsQCD > 0.2
$70 < m_{\text{jj}} < 270$ GeV	BDT_HvsT < -0.2
	$120 < m_{\text{SD}} < 240$ GeV

Table 4.4: Summary of the selection requirements defining the Higgs boson and top quark candidates.

$\text{t}\bar{\text{t}}\text{H}(\text{b}\bar{\text{b}})$ simulated sample and is defined as the ratio of the number of top quark candidates matched to a top quark to the total number of top quark candidates. This purity is found to be $\approx 83\%$ and is visually shown in Fig. 4.12.

A summary of the requirements defining the Higgs boson and top quark candidates is given in Table 4.4.

4.4.5 Signal categories

To further enhance the analysis sensitivity, events that have reached this stage of the selection are split into four, mutually-exclusive signal categories based on the number of AK8 jets, top quark candidates and AK4 jets. Since the Higgs boson decays in a resolved topology, boosted $\text{t}\bar{\text{t}}\text{H}$ events are expected to contain up to two AK8 jets corresponding to boosted decays of top quarks. Thus, categories with two and one AK8 jets are taken into account. In the

following, a description of each category is given. The counting starts by nine in order to avoid confusion with the categories of the BHC, when the combination between the RHC and BHC will be presented.

Categories with two AK8 jets

- *category* 9: In this category, events in which at least one AK8 jet is tagged as a top quark candidate are selected. Given the resolved decay of the Higgs boson and the boosted decays of the top quarks, no additional AK4 jets are expected and no requests are set concerning AK4 multiplicity. This category shows a purity of reconstructed Higgs boson candidates of about 39% and a purity of reconstructed top quark candidates of about 82%.
- *category* 10: Given that there is a non negligible probability for a high-momentum top quark to fail the identification, in this category we include events with two AK8 jets, with none of them being identified as a top quark candidate. The same reasoning as before concerning the AK4 multiplicity holds in this case. This category shows a purity of reconstructed Higgs boson candidates of about 30%.

Categories with one AK8 jet

- *category* 11: This category contains event with one AK8 jet which is identified as a top quark candidate. The presence of only one AK8 jet implies that the remaining top quark decays in the resolved topology. Thus, a nominal number of five AK4 jets is expected in the event, three of them being b tagged. Since there is a finite probability for a jet to be misidentified with a different object, a conservative request $N_{\text{AK4jets}} \geq 4$ is made. The correct combination of AK4 jets that forms the Higgs candidate is identified by making use of the resolved Higgs BDT as explained in Section 4.4.3. This category shows a purity of reconstructed Higgs boson candidates of about 39% and a purity of reconstructed top quark candidates of about 81%.
- *category* 12: Given that there is a non negligible probability for a high-momentum top quark to fail the identification, in this category we include events with one AK8 jet which fails the top quark candidate identification. The same reasoning as before concerning the AK4 multiplicity holds in this case. This category shows a purity of reconstructed Higgs boson candidates of about 28%.

A summary of the requirements that define each category is given in Table 4.5. As a general remark, we see that, while the top quark candidate purity is

Category	N_{AK8}	Higgs-tagged	Top-tagged	N_{AK4}
9	2	✓	✓	-
10	2	✓	×	-
11	1	✓	✓	≥ 4
12	1	✓	×	≥ 4

Table 4.5: Summary of the selection requirements defining the signal categories.

pretty stable across categories, the Higgs boson candidate purity is found to be higher in categories where top quark candidates are found as well. This is consistent with the resolved-Higgs BDT not being a Higgs tagger, but rather an event-variable-based BDT which is able to identify $t\bar{t}H(b\bar{b})$ events in which the Higgs boson decays in the resolved topology. Thus, events containing top quark candidates are expected to be more likely $t\bar{t}H(b\bar{b})$ events and the Higgs boson candidate purity is increased for such events. In Fig. 4.13 the sample purity of the reconstructed Higgs boson candidates is shown for each signal category. The percentage of real reconstructed Higgs boson candidates in each category varies between 28% in category 12 and 39% in categories 9 and 11. Figure 4.13 also shows the sample purity of reconstructed top quark candidates for categories in which a top quark candidate is identified. The percentage of real top quarks reconstructed as top quark candidates is around 80% for each category. Finally, Fig. 4.14 shows the expected yields in each signal category, corresponding to an integrated luminosity of 35.9 fb^{-1}, as predicted by the simulations. We clearly see that QCD multijet production contributes in a dominant way in all categories. The same figure also shows the background composition of each category, where we note the expected behavior of categories with a top quark candidate being more contaminated by the $t\bar{t}$ background with respect to categories where a top quark candidate is not found.

Eventually, events passing the signal selection of this analysis must fulfill selection requirements that can be summarized as follow:

- pass the baseline selection;
- do not find a boosted Higgs boson candidate;
- consider events with $N_{\text{AK4bjets}} \geq 2$;
- find a resolved Higgs boson candidate;
- split events in signal categories.

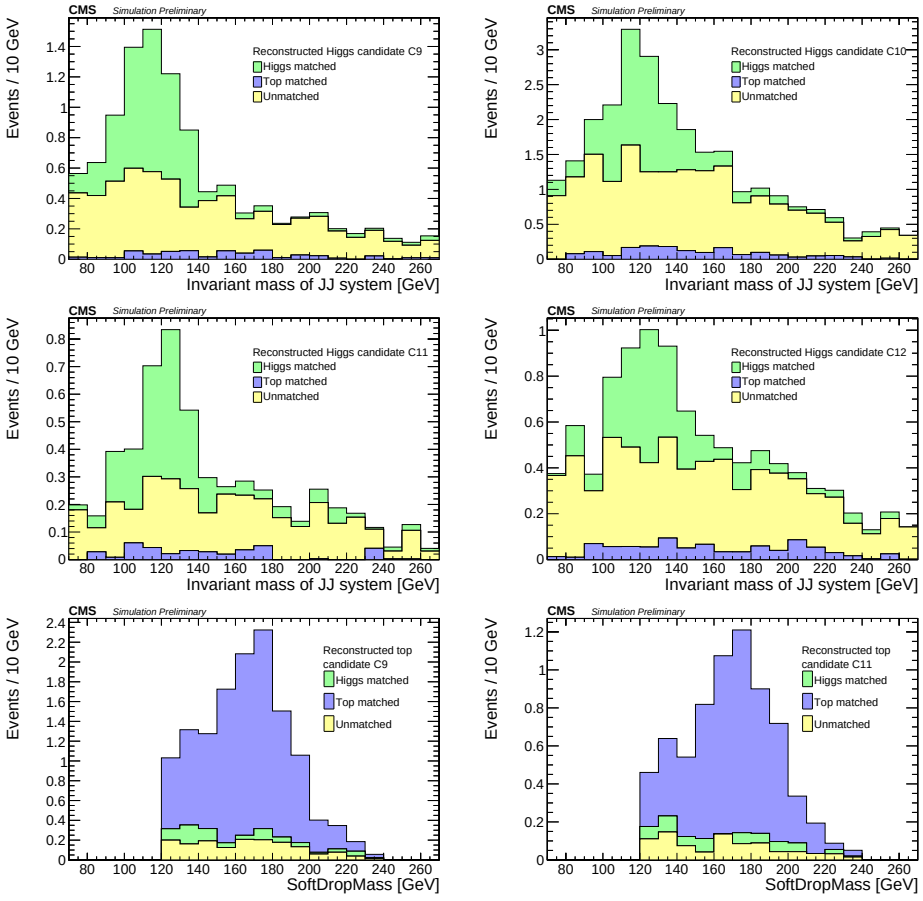

Figure 4.13: Composition of the reconstructed Higgs boson candidate for the signal categories (first and second row); composition of the reconstructed top quark candidate for the signal categories in which a top candidate is found (third row).

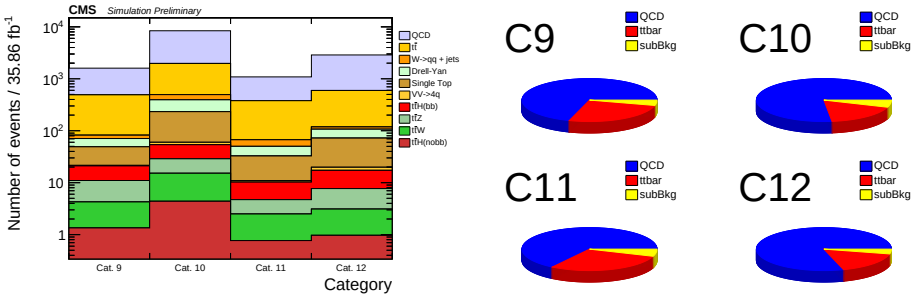

Figure 4.14: Expected yields in each signal category (left) and background composition in each signal category (right).

$\mathcal{B}$	$\mathcal{E}_{\text{train}}$	$\mathcal{E}_{\text{test}}$
0.99	0.276	0.265
0.9	0.677	0.669
0.7	0.882	0.878

Table 4.6: Overtraining check for the resolved-Higgs BDT. The BDT efficiency is measured for different background rejection values in two independent samples.

4.5 Validation

In this section several checks are presented, which are aimed at demonstrating the consistency of the previously developed event selection. First, we perform sanity checks concerning the BDT-based identification of resolved Higgs boson decays; then, data-MC comparisons are extensively used to test the goodness and coherence of the signal selection.

4.5.1 Resolved-Higgs BDT consistency checks

In order to test the overall sanity of the resolved-Higgs BDT, several checks have been performed. First, we performed an overtraining check. The most evident symptom of overtraining is a seemingly increased performance in the classification over the objectively achievable one, if measured on the training sample, and an effective performance decrease when measured with an independent test sample. Thus, the resolved-Higgs BDT efficiency $\mathcal{E}$ has been measured, in correspondence of different values of the background rejection $\mathcal{B}$, for the training and test samples. The results of this checks are reported in Table 4.6. Since only small deviations are found in the efficiency when computed in the two independent samples, we conclude that the resolved-Higgs BDT does not suffer from substantial overtraining.

As a second check regarding the resolved-Higgs BDT performance, we computed the shapes of the observable for the $t\bar{t}H(b\bar{b})$ signal and the QCD main background using the simulations. This check is aimed at showing whether the cut on the resolved-Higgs BDT score introduces unwanted biases and distortions in the distributions. Since we took care of excluding variables which are strongly correlated with the dijet invariant mass from the training, no such biases are found. This is testified by Fig. 4.15 which shows that, in each category, the $t\bar{t}H$ shape peaks around the Higgs boson mass value, while the QCD shape shows a smooth, decreasing behavior.

As a third check, we filled the scatter plot highest resolved-Higgs BDT score vs. corresponding invariant mass, both for the $t\bar{t}H(b\bar{b})$ signal and QCD background, right before splitting events in categories, as shown in Fig. 4.16. As expected, while in both cases the BDT score profile shows a smooth

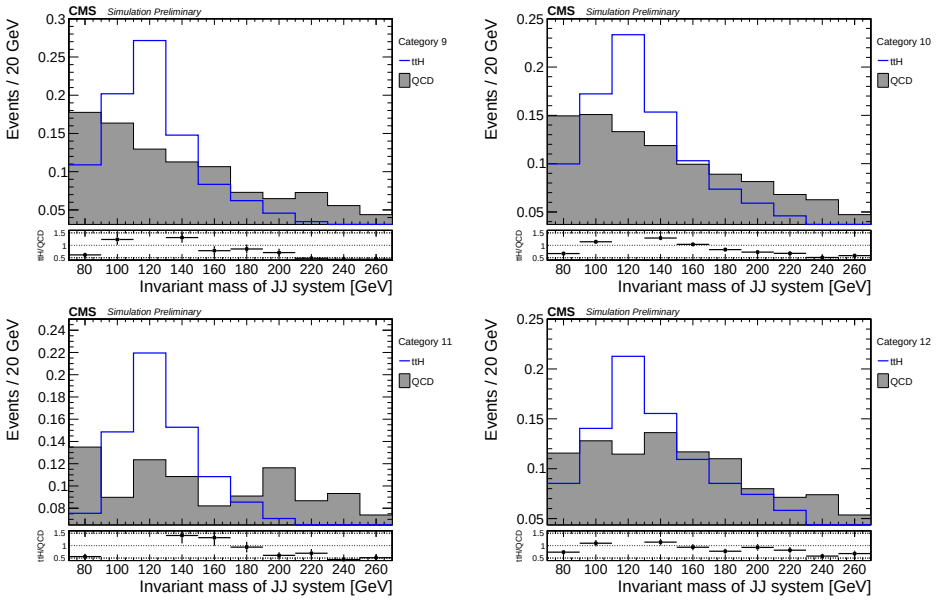

Figure 4.15: Comparisons between simulated $t\bar{t}H$ and QCD observable shapes for category 9 (upper left), category 10 (upper right), category 11 (lower left) and category 12 (lower right). The distributions show the expected behaviors, with $t\bar{t}H$ peaking around the Higgs mass and QCD showing a decreasing pattern.

increasing-decreasing trend, the Higgs boson candidate mass shows a peak around the Higgs boson mass for the $t\bar{t}H(b\bar{b})$ signal and is instead smoothly decreasing for the QCD multijet background.

Finally, in Fig. 4.17 we plot the decay mode of the $t\bar{t}H$ system before and after the cut on the resolved-Higgs BDT score using the $t\bar{t}H(b\bar{b})$ and $t\bar{t}H(nob\bar{b})$ simulated samples. More precisely, we first plot the decay mode for events that pass the baseline selection only; then, we plot the same variable for events in which a resolved Higgs boson candidate is also reconstructed. As a result of the cut, the fraction of events decaying to an all-jets final state becomes significantly higher.

4.5.2 Data vs. Monte Carlo comparisons

In order to demonstrate the overall consistency of the data and of the selection requirements, we present data vs. MC comparisons for various variables of interest. In all cases the $t\bar{t}H$ signal, the $t\bar{t}$ background, the $t\bar{t}Z$ irreducible background and the subdominant backgrounds are normalized to the 2016 integrated luminosity using the corresponding theoretical cross sections reported in Table 4.2. The QCD multijet background is first normalized to the 2016 integrated luminosity, using the cross section of each H_T slice; then

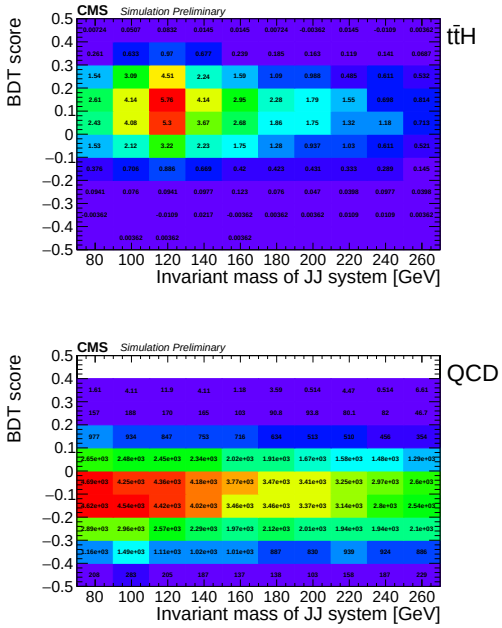

Figure 4.16: Resolved-Higgs BDT score vs. invariant mass of the resolved Higgs candidate for $t\bar{t}H$ signal (up) and QCD multijet background (down).

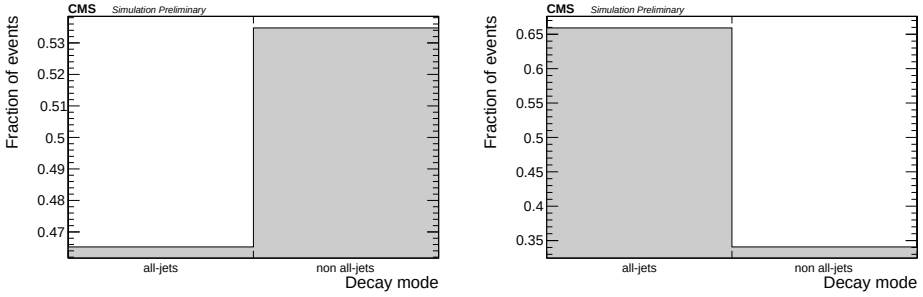

Figure 4.17: Decay modes for the $t\bar{t}H$ system before (left) and after (right) the cut on the resolved-Higgs BDT score. The bin labeled "all-jets" contains the fraction of $t\bar{t}H$ events in which the Higgs boson decays to a $b\bar{b}$ pair and the $t\bar{t}$ pair decays hadronically; the bin labeled "non all-jets" contains all the other decays.

a global k-factor is applied in such a way that the total simulated yield is equal to the number of events in data. We must point out that the QCD simulation is used here with the only purpose of checking the goodness of the signal selection, but it will not enter the final fit as the QCD multijet background will be estimated from data, as it will be described in Section 4.6. The comparisons presented here are performed both before and after events are split into categories.

First, as a further validation of the resolved-Higgs BDT performance, we report the data-MC agreement for the variables used in the training and for the resolved-Higgs BDT score itself. Figures 4.18, 4.19 and 4.20 show such comparisons for events that pass the baseline selection and show a resolved reconstructed Higgs boson candidate. In general, an overall good agreement between data and simulations is found, which justifies the use of the resolved-Higgs BDT on the data. A good agreement is also found in the resolved-Higgs BDT distribution, which is shown prior to any cut on it.

Next, we show data vs. simulations comparisons for some variables of interest after the event categorization. Figure 4.21 shows a set of such variables for category 9. The soft drop mass of the leading jet clearly shows a peak corresponding to the top quark mass; the bump in the mass window $[120, 240]$ GeV corresponds to leading jets which are identified as top quark candidates, while values outside of this range correspond to leading jets which fail the top quark candidate identification. The mass of the leading subjet within the leading AK8 jet clearly shows the W boson resonance, as a result of highly collimated decay products of the W boson that end up being collected in a single subjet. On the other hand, the mass of the second-leading subjet within the leading AK8 jet does not show such behavior. The same set of variables is shown in Figs. 4.22, 4.23 and 4.24 for categories 10, 11 and 12 respectively. In category 10, where no top quark candidates are found, the top quark resonant peak in the leading jet soft drop mass in not present anymore. Similarly, no resonances are found in the mass distributions of the subjets. In category 11, where a top quark candidate is found, we clearly see the top quark and W boson mass peaks in the leading jet soft drop mass and leading subjet mass distributions. In a similar fashion, no resonances are found in category 12.

4.6 QCD background estimation

As we already highlighted in the previous sections, the dominant background to the $t\bar{t}H$ associated production in the fully hadronic decay channel is the QCD multijet production. In fact, there is a non negligible probability for events produced by generic QCD interactions to mimic the $t\bar{t}H(b\bar{b})$ topology and enter the signal categories. Due to the very large cross section of such events in proton-proton collisions, this background becomes by far the most

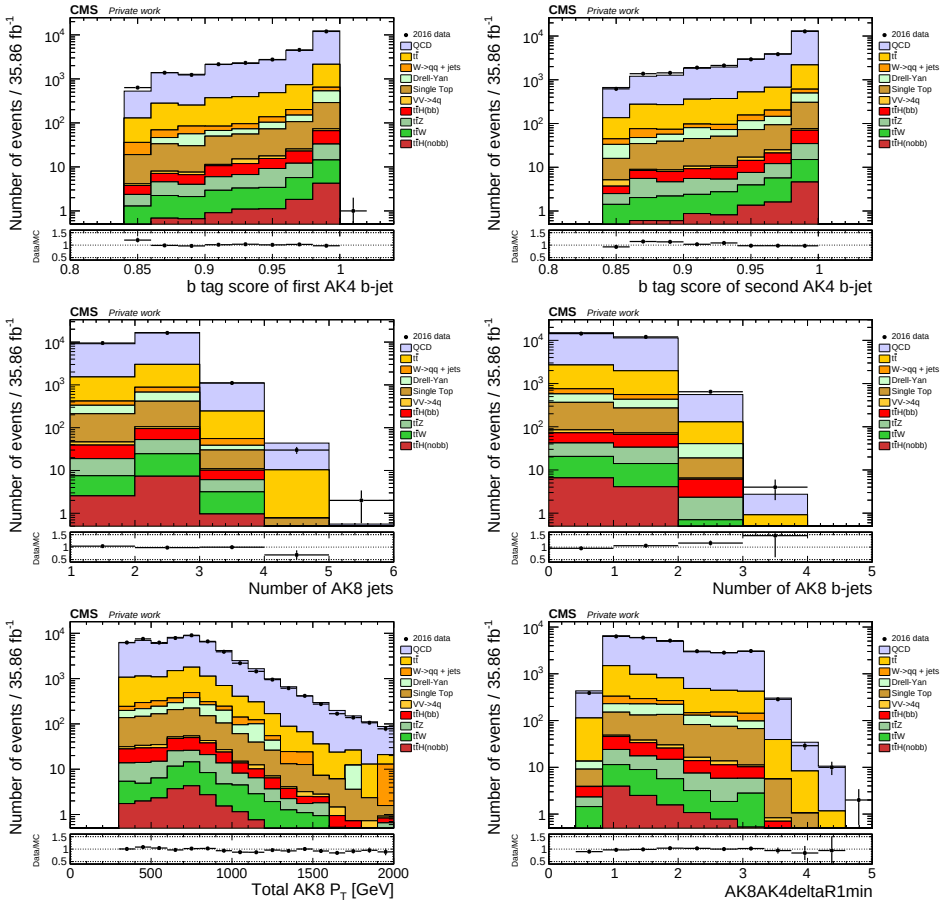

Figure 4.18: Data vs. simulations for the variables used to train the resolved-Higgs BDT. The b tag score of the first AK4 b jet and the b tag score of the second AK4 b jet (upper row), the number of AK8 jets and the number of AK8 b tagged jets (middle row), the scalar sum of the p_{T} of the AK8 jets and the minimum ΔR between the first AK4, b tagged jet and the AK8 jets (lower row) are shown.

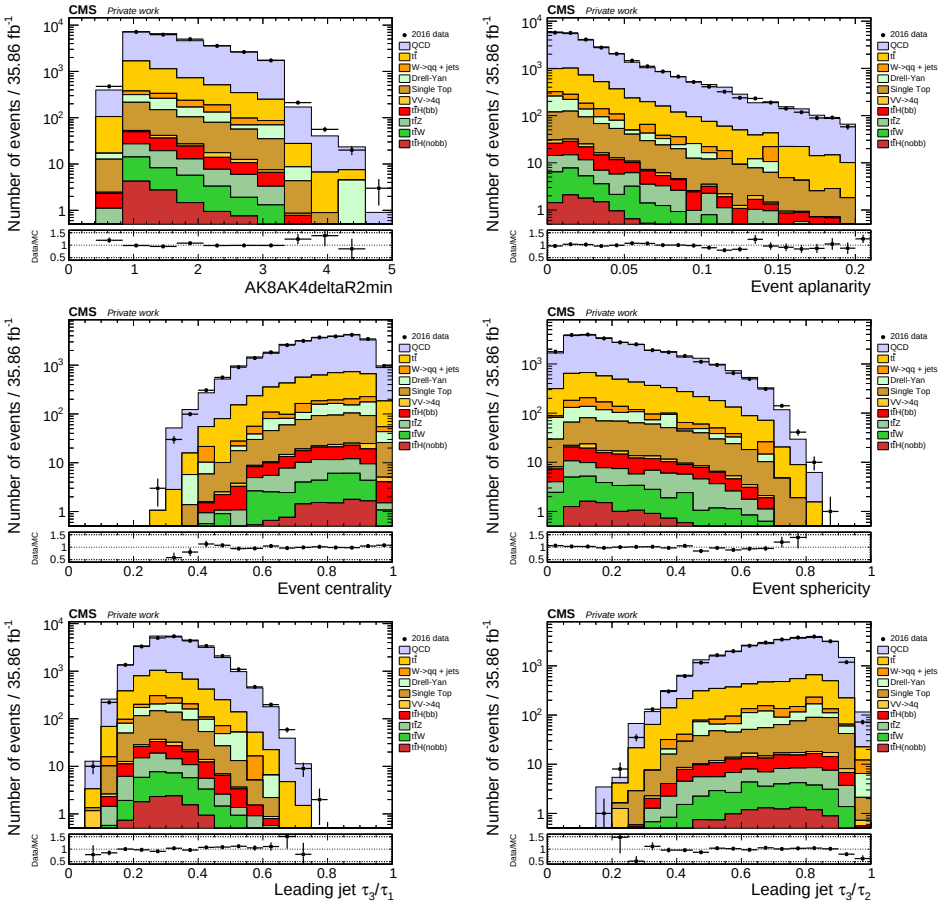

Figure 4.19: Data vs. simulations for the variables used to train the resolved-Higgs BDT. The minimum ΔR between the second AK4, b tagged jet and the AK8 jets and the event aplanarity (upper row), the event centrality and the event sphericity (middle row), the τ_3/τ_1 of the leading AK8 jet and the τ_3/τ_2 of the leading AK8 jet (lower row) are shown.

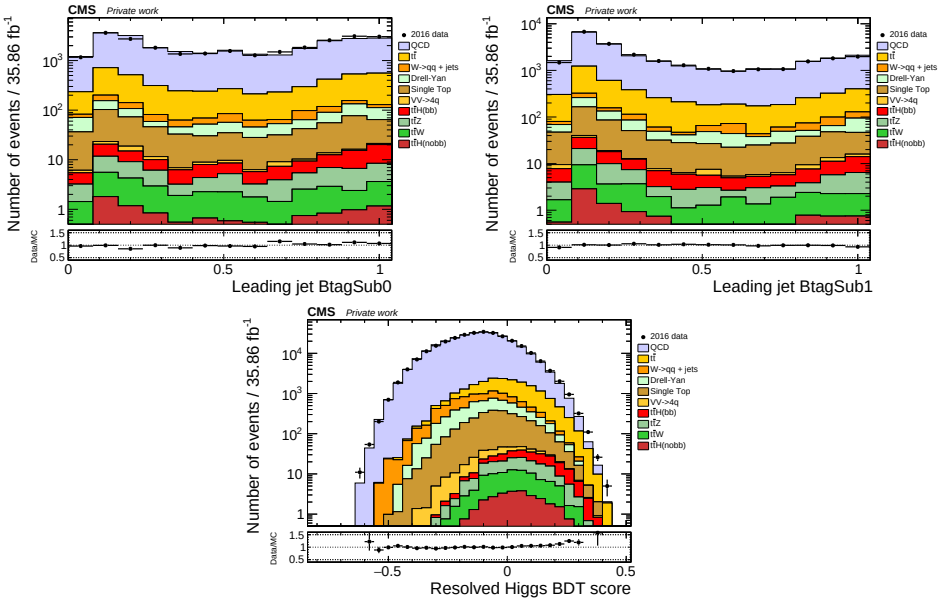

Figure 4.20: Data vs. simulations for the variables used to train the resolved-Higgs BDT and resolved-Higgs BDT score distribution. The b tag score of the first subjet inside the leading AK8 jet and the b tag score of the second subjet inside the leading AK8 jet are shown, as well as the distribution of the resolved-Higgs BDT scores prior to any cut on it.

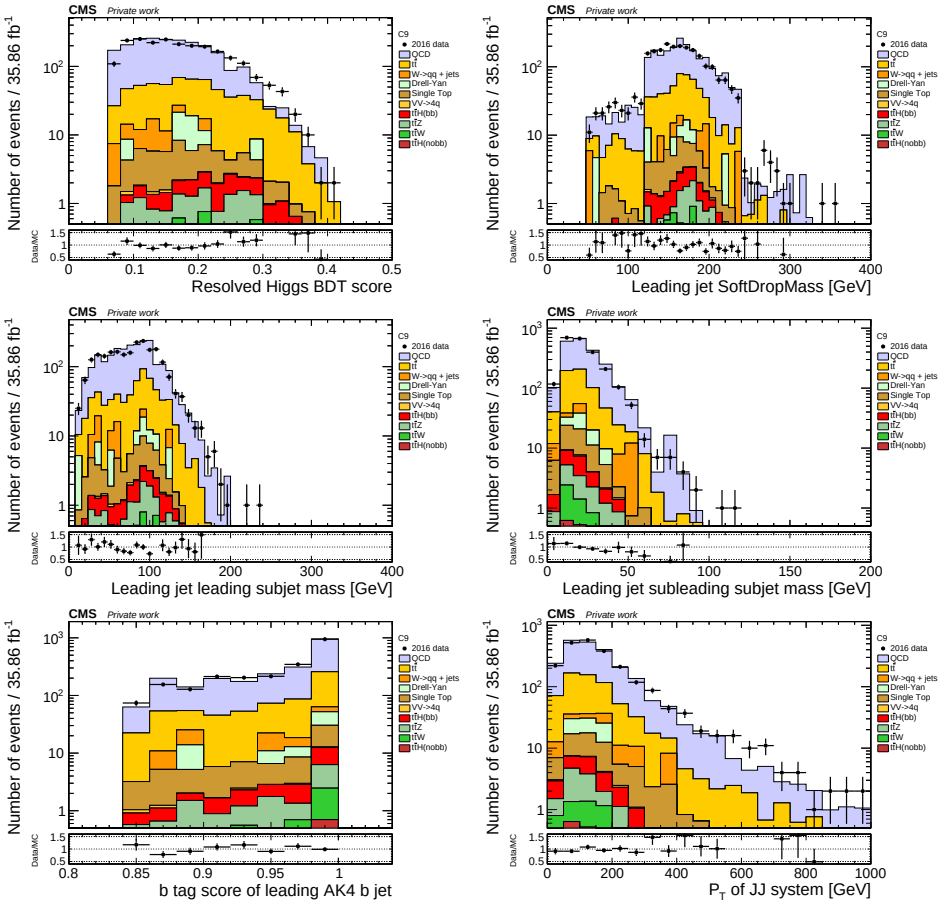

Figure 4.21: Data vs.simulations for some variables of interest in category 9. The resolved-Higgs BDT score and the soft drop mass of the leading AK8 jet (upper row), the mass of the first subjet inside the leading AK8 jet and the mass of the second subjet inside the leading AK8 jet (middle row), the b tag score of the leading AK4, b tagged jet forming the resolved Higgs boson candidate and the p_{T} of the resolved Higgs boson candidate jets (lower row) are shown.

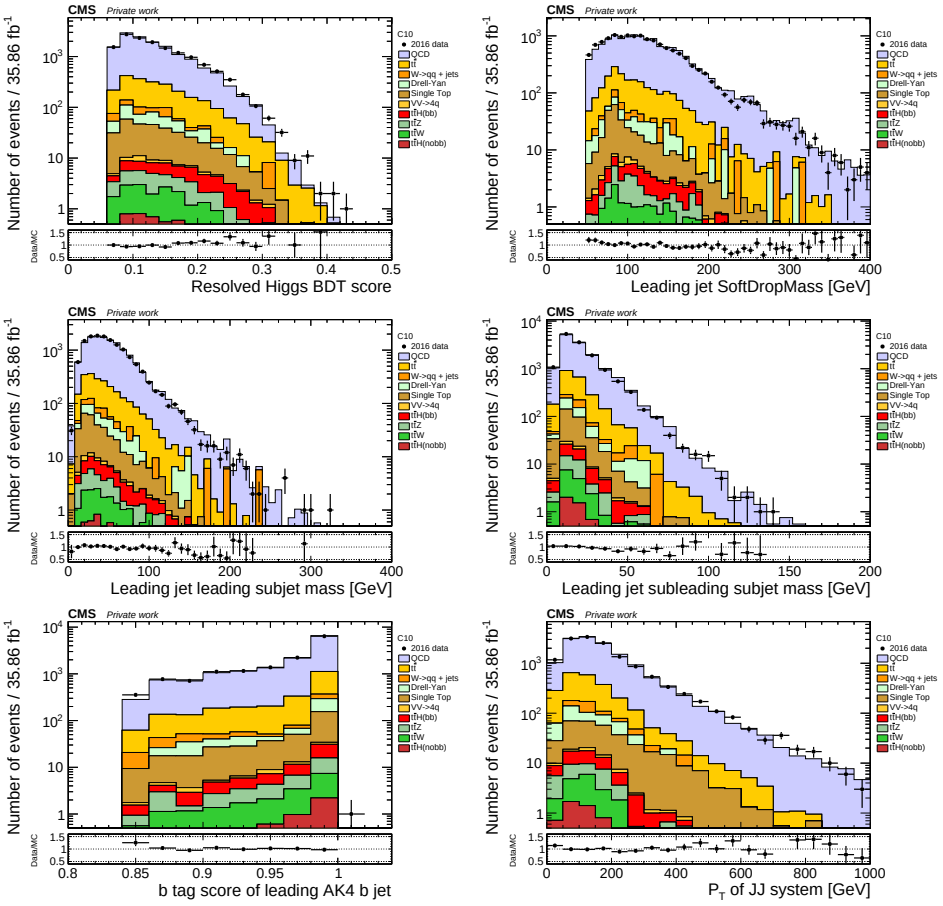

Figure 4.22: Data vs.simulations for some variables of interest in category 10. The resolved-Higgs BDT score and the soft drop mass of the leading AK8 jet (upper row), the mass of the first subjet inside the leading AK8 jet and the mass of the second subjet inside the leading AK8 jet (middle row), the b tag score of the leading AK4, b tagged jet forming the resolved Higgs boson candidate and the p_{T} of the resolved Higgs boson candidate jets (lower row) are shown.

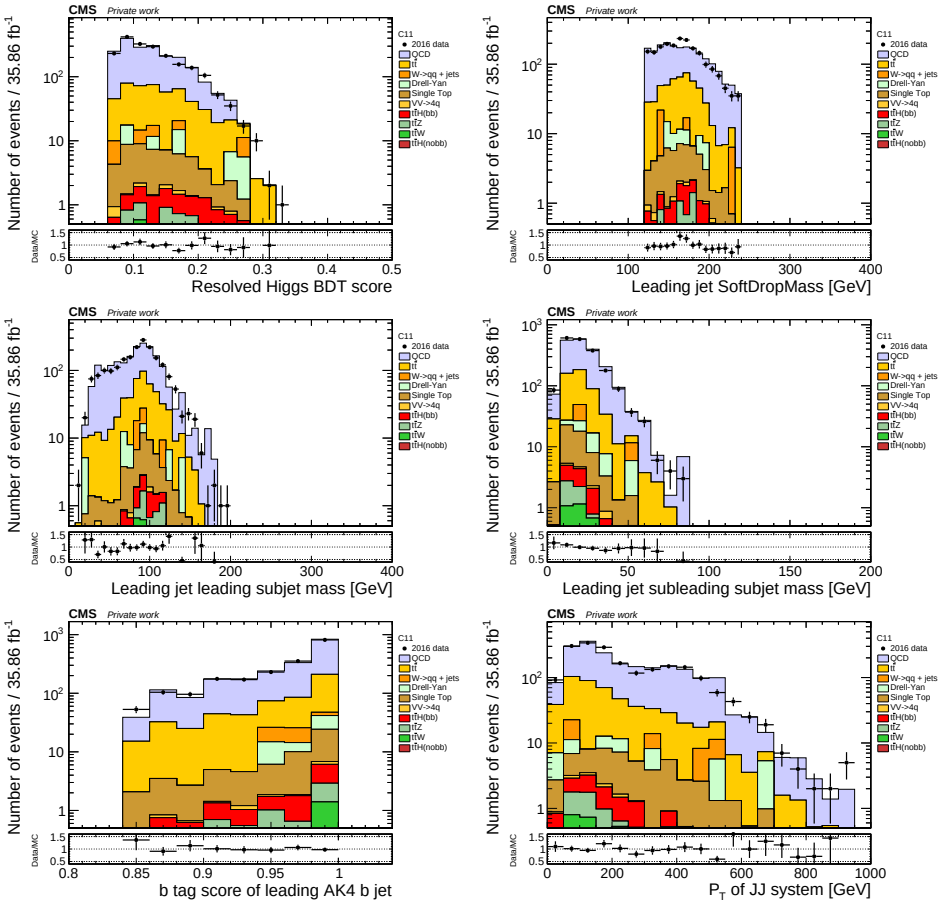

Figure 4.23: Data vs.simulations for some variables of interest in category 11. The resolved-Higgs BDT score and the soft drop mass of the leading AK8 jet (upper row), the mass of the first subjet inside the leading AK8 jet and the mass of the second subjet inside the leading AK8 jet (middle row), the b tag score of the leading AK4, b tagged jet forming the resolved Higgs boson candidate and the p_{T} of the resolved Higgs boson candidate jets (lower row) are shown.

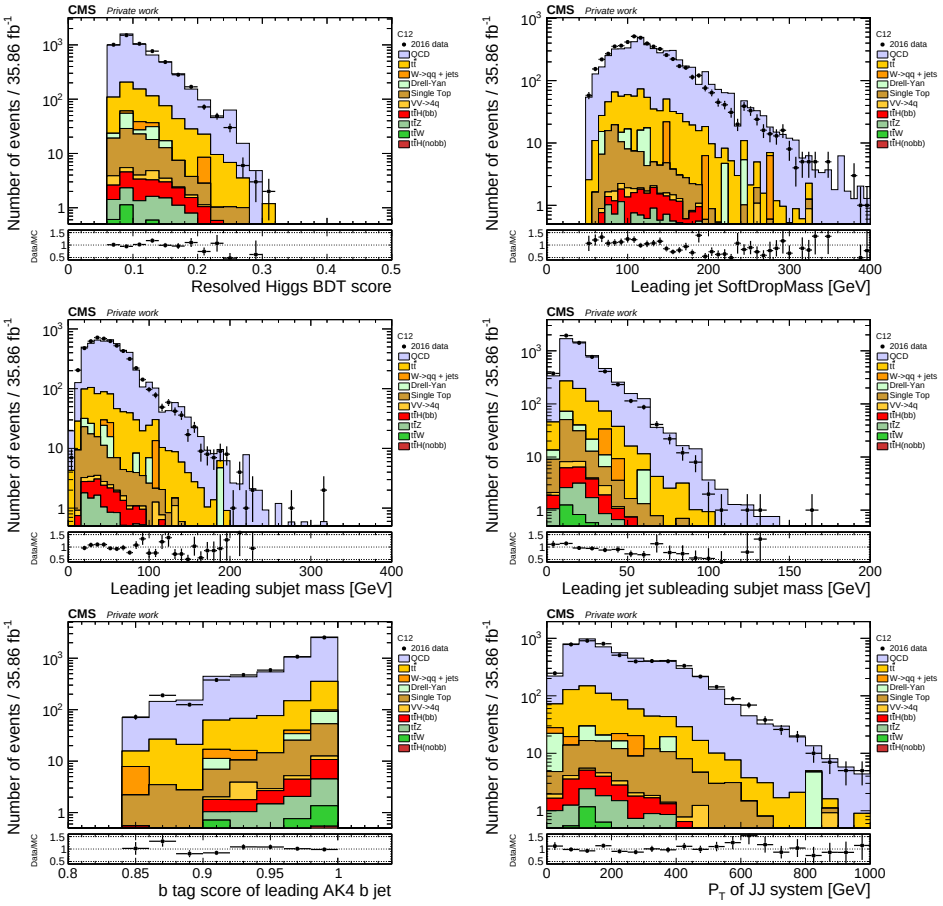

Figure 4.24: Data vs.simulations for some variables of interest in category 12. The resolved-Higgs BDT score and the soft drop mass of the leading AK8 jet (upper row), the mass of the first subjet inside the leading AK8 jet and the mass of the second subjet inside the leading AK8 jet (middle row), the b tag score of the leading AK4, b tagged jet forming the resolved Higgs boson candidate p_T of the resolved Higgs boson candidate jets (lower row) are shown.

abundant one. Unfortunately, MC predictions of QCD processes cannot be safely used. First of all, they suffer from big theoretical uncertainties on the cross sections and on next-to-leading order corrections, which lead to a poor description of the data, especially for events with high jet multiplicity such as the ones targeted by this analysis; second, despite the big number of generated events, the efficiency after the selection is usually very low, resulting in small usable samples. Given these limitations, the yields and shapes of the QCD multijet background are obtained with a data-driven approach, which makes possible to avoid the big theoretical uncertainties related to the simulated events and also guarantees a good population in the template histograms used in the final fit.

4.6.1 QCD shapes - QCD control region

In order to estimate the QCD shapes, we define a control region (CR) enriched in QCD events. This CR should show two important properties: first, it should be as kinematically close as possible to the signal region defined in Section 4.4, in such a way that events falling in this region belong to a phase space sufficiently similar to the one used in the analysis; second, it should be orthogonal to the signal region, so that double counting of events is avoided. In order to find the selection criteria defining a proper CR, we note that the critical request to identify signal events is to find a resolved Higgs boson candidate. This is achieved through the resolved-Higgs BDT, performing a cut on the BDT score distribution. This cut helps in rejecting the QCD background and selecting $t\bar{t}H(b\bar{b})$ events. Reverting this cut, a QCD-enriched sample is obtained, in which the events are expected to show similar kinematic properties to the events in the signal region. Also, since the BDT cut is reverted, orthogonality between the two regions is obviously achieved. Thus, in order to define the CR, events passing the baseline selection and in which a boosted Higgs boson candidate is not found are considered. Then, events with at least two AK4, b tagged jets are considered and the combination showing the highest resolved-Higgs BDT score is selected. If this score is found to be less than 0.07, the event enters the CR. Then, events are split in categories with the same critera that were discussed in Subsection 4.4.5, in such a way that a QCD shape is obtained for each category.

Eventually, events passing the control selection of this analysis must fulfill selection requirements that can be summarized as follow:

- pass the baseline selection;
- do not find a boosted Higgs boson candidate;
- consider events with $N_{\text{AK4bjets}} \geq 2$;
- find a “reverted” resolved Higgs boson candidate;

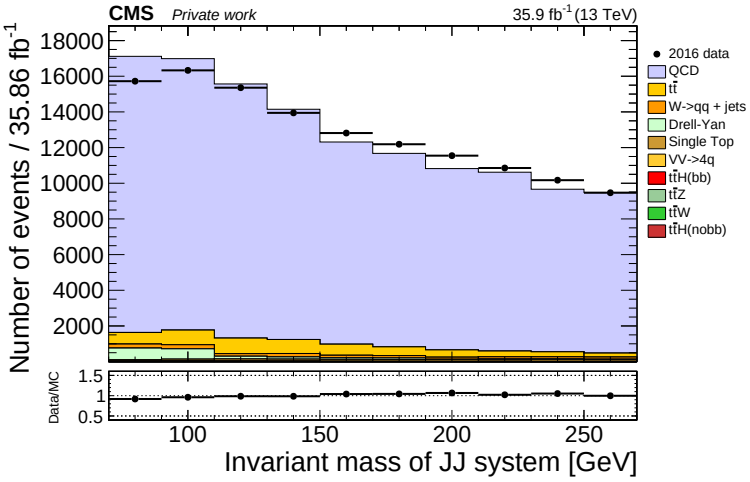

Figure 4.25: Invariant mass of the resolved Higgs boson candidate for events in the QCD CR. The multijet events are found to be 93% of the total number of events.

- split events in categories.

To testify that these cuts effectively lead to a QCD-dominated region, in Fig. 4.25 we report the invariant mass of the resolved Higgs boson candidate for events entering the CR, before the splitting into categories. Indeed, the fraction of multijet events with respect to the total is found to be 93%.

To check that the control selection is indeed kinematically close to the signal selection, a closure test is performed. Using the simulation, we compare the QCD distribution of the invariant mass of the dijet pair with the highest resolved-Higgs BDT score for the signal and control selections. The outcome of such study is shown in Fig. 4.26, where we see that an overall good agreement is found between the signal and control shapes in all the categories. In order to account for residual kinematic differences between the two selections, a transition factor is computed as the ratio between the distributions of the dijet invariant mass in the signal and control regions. The transition factor is then fitted with a smooth curve to obtain a transition function which is used to switch from the control region to the signal region. Eventually, the following formula is used:

$$B^{\text{data}}_{\text{signal}}(m_{\text{jj}}) = \frac{B^{\text{MC}}_{\text{signal}}}{B^{\text{MC}}_{\text{control}}}(m_{\text{jj}}) B^{\text{data}}_{\text{control}}(m_{\text{jj}}), \tag{4.8}$$

where "signal" and "control" refer to the selection with regular and reverted resolved-Higgs BDT cut respectively. The lower panels in Fig. 4.26 show the transition factors and the fitting transition functions. In all the cases, the

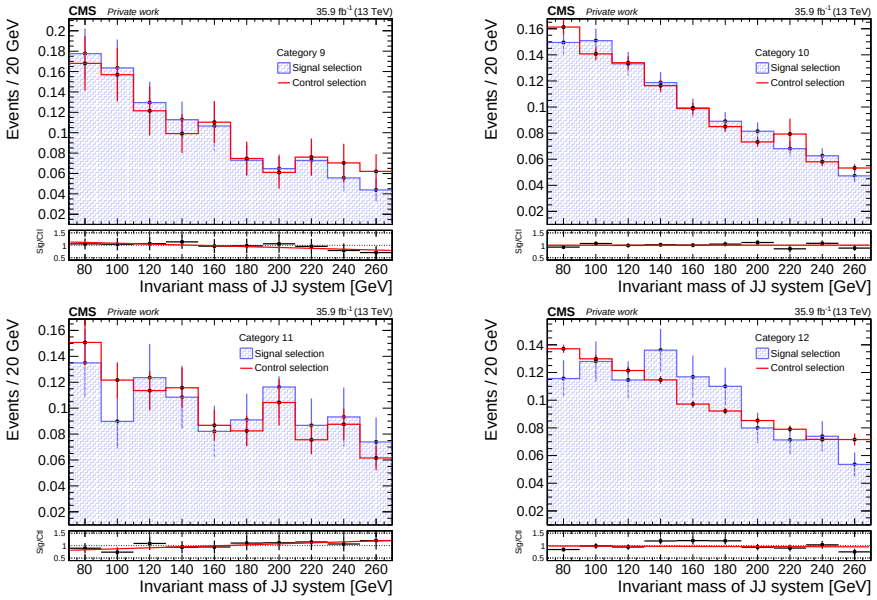

Figure 4.26: Closure test for the estimation of the QCD shapes in category 9 (upper left), category 10 (upper right), category 11 (lower left) and category 12 (lower right). Transition factors, along with the fitting transition functions are shown as well.

transition factor trend is approximately linear, so that the fitting functions have been chosen to be straight lines.

Once the selection criteria are validated by the closure test, the QCD shapes are computed from data events entering the CR. Then, the shapes are corrected with the transition functions as described in Eq. 4.8. To show the goodness of this procedure, the corrected distributions in data are compared with the simulated QCD shapes in the signal region. Such comparisons are shown in Fig. 4.27, where we see that the method results in a good agreement between the two distributions.

4.6.2 QCD yields - ABCD method

The QCD background yields in each signal category are obtained from data with an ABCD-like method. The ABCD method is a widely used procedure which provides a way to estimate the number of background events in the signal region by exploiting the discriminating power, in the phase space of the analysis, of two loosely correlated variables. To apply the method, events passing the baseline selection and having at least two AK4, b tagged jets are used. The two loosely correlated variables are chosen to be the highest resolved-Higgs BDT score among all the possible combinations, and the corresponding invariant mass. The degree of correlation between these two

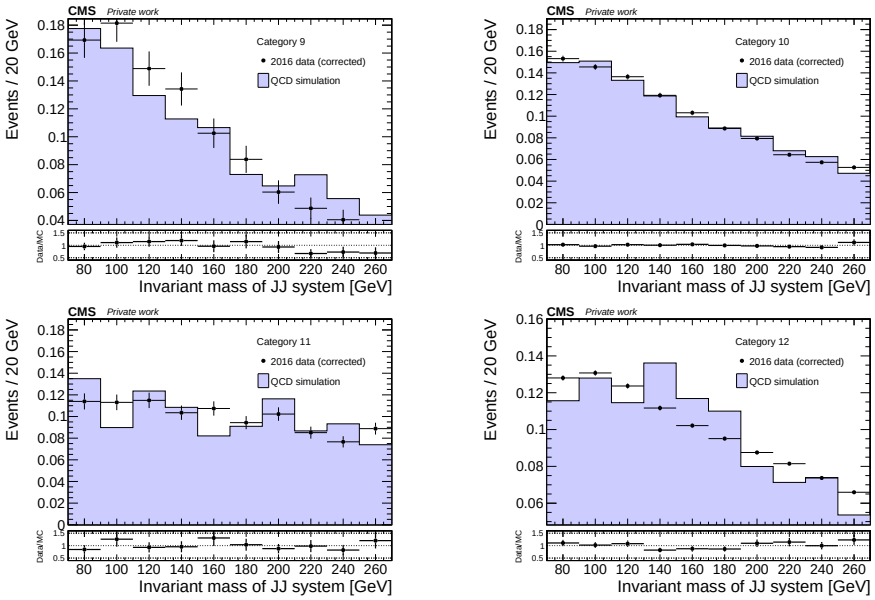

Figure 4.27: Comparison between simulated QCD shapes in the signal region and corrected distributions in data in category 9 (upper left), category 10 (upper right), category 11 (lower left) and category 12 (lower right).

variables has been checked and a correlation coefficient of -0.18 has been found. In the phase space defined by the two variables, four orthogonal regions A, B, C and D are defined, with region A being the region containing signal events, while regions B, C and D mostly contain background events. Region A is identified by a resolved-Higgs BDT score greater than 0.07 and an invariant mass lying in the window $[70, 270]$ GeV, in such a way that events falling in each signal category effectively correspond to subsets of the events falling in region A. If the cuts defining the background-enriched regions are properly chosen, and given that the variables are loosely correlated, the ratio of events N_A/N_B should be equal to the ratio N_C/N_D and the number of signal events can be estimated as

$$N_A = \frac{N_B N_C}{N_D}. \tag{4.9}$$

Since the analysis phase space is pretty complicated, with multiple signal categories, two extended categories corresponding to events with one and two AK8 jets are defined, and for each of them four regions are defined, which are shown in Fig. 4.28.

As a first step towards the estimation of the QCD yields, a closure test is performed using the MC QCD events. In such a check, we count the simulated events falling in region A and compare this number with the result coming from Eq. 4.9 both for the extended categories with two and one AK8

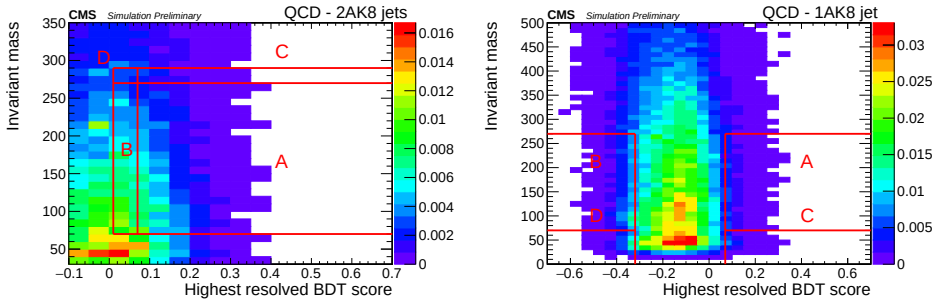

Figure 4.28: Extended categories used in the ABCD method for events with two AK8 jets (left) and one AK8 jet (right).

$N_{\text{AK8 jets}}$	MC prediction	ABCD prediction
2	8015 ± 174	8308 ± 1290
1	4363 ± 128	4316 ± 326

Table 4.7: MC closure test for the ABCD method. The second column reports the number of events falling in region A, while the third column reports the prediction obtained by making use of Eq. 4.9. The errors on the predictions reflect the statistical uncertainties.

jets. The outcome of the counting is reported in Table 4.7. The reported uncertainties are statistical, where we assumed that the yields in regions B, C and D are independent, in such a way that the uncertainty on the ABCD prediction simply comes by linear error propagation. Since both the predictions from ABCD and simple counting are in agreement within the uncertainties, we argue that the method closes and can be applied to the data. The QCD simulation is also used to compute the fractions of events falling in the A regions of the extended categories that also fall in the signal categories of the analysis.

Once the method is validated in the simulation, we apply it to the data. Before using Eq. 4.9, we subtract the contributions of the $\mathrm{t\bar{t}}$ and subdominant backgrounds from regions B, C and D in data using the simulation, in such a way that the outcome numbers will genuinely predict the QCD yields, without contamination. To make sure that the method is consistent in data as well, we compare the ABCD yield with the number of data events falling in region A after the subtraction of $\mathrm{t\bar{t}}$ and subdominant backgrounds. The resulting numbers are found to be compatible within the uncertainties for data events as well. Finally, we use the fractions of events falling in the A regions of the extended categories that also fall in the signal categories, computed from the simulation, to obtain the expected QCD yields in each signal category. The

	$N_{\mathrm{AK8\,jets}}$	Data prediction	ABCD prediction	Fraction	Yield
C9	2	8406 ± 114	8698 ± 942	0.137229	1194
C10				0.794051	6906
C11	1	4934 ± 84	4757 ± 212	0.162647	774
C12				0.52327	2489

Table 4.8: Summary of the results obtained from the application of the ABCD method on data. For each category, we report: the number of events falling in region A, after the subtraction of background contributions (second column), the prediction obtained by making use of Eq. 4.9 after the subtraction of background contributions (third column), the fractions of events falling in the A regions of the extended categories that also fall in the signal categories, obtained from simulated QCD (fourth column) and the yield in each category which is obtained from the multiplication of the two previous columns (fifth column).

aforementioned procedure and its results are summarized in Table 4.8.

4.6.3 Final QCD estimation

Eventually, the estimation of the QCD background in each signal category is obtained by combining the previously described methods: shapes obtained from the data events in the CR are normalized to the expected yields (fifth row of Table 4.8) coming from the ABCD method applied to data. Thus, the dominant background of this analysis is entirely estimated from data, in such a way that the big theoretical uncertainties related to the simulated QCD events are avoided. Nevertheless, the simulation is used to compute the fraction of ABCD events falling in each signal category. This is found to introduce a small uncertainty, as it will be shown in Subsection 4.9.1.

4.7 $\mathrm{t\bar{t}}$ background estimation

The second dominant background of the search is the $\mathrm{t\bar{t}}$ associated production, with additional jets produced together with the top quark-antiquark pair. The all-jets decay of the $\mathrm{t\bar{t}}$ pair is expected to produce six partons in the final state; in the case of gluon emission and/or splitting, additional partons can be produced and the final state can mimic the one targeted by this analysis. Moreover, given the relatively large cross section of the process, considerable yields are expected.

This source of background is estimated entirely using MC events. However, some care must be used when dealing with the simulated $\mathrm{t\bar{t}}$ events, since some degree of mismodelling in the description of the data has notoriously

been found in previous analyses. In the following, we describe the procedure used to deal with such mismodelling and the strategy used to reduce the uncertainty on the modelling of the process.

4.7.1 Loose $t\bar{t}$ validation region

Several analyses [34, 82] previously developed by the CMS Collaboration found the p_T spectrum of top quarks in data to be significantly softer than the one predicted by LO and NLO generators interfaced with parton showers. Since the origin of this suboptimal description is yet unclear, a general procedure has been developed inside the collaboration to reweight the top quark p_T in the simulation to better describe the data. This procedure works for low-p_T top quarks, while no special recommendation holds for highly boosted top quarks. Thus, we derive a corrective factor for the $t\bar{t}$ cross section which is used to account for the p_T and any other possible sources of mismodelling in the simulations.

We start by defining a highly boosted $t\bar{t}$-enriched region, which we will call in the following loose $t\bar{t}$ validation region (loose VR). Events belonging to the loose VR must first pass the requirements of the signal trigger; also, they must contain no isolated leptons in the final state and have at least two AK8 jets. Subsequently, requests are made on the nature of the leading and subleading AK8 jets: both should have a p_T greater than 300 GeV and have at least one b tagged subjet. Finally, to select a leading jet with a tight top quark-like topology, cuts are placed on the boosted BDTs asking for the BDT_TvsQCD score to be greater than 0.7 and for the BDT_HvsT score to be less than 0.1. The selection requirements that form the loose VR selection are summarized in Table 4.9. Since strong requirements on the top quark-like nature of the leading jet have been made, we assume it to be the top quark candidate in the loose VR and we use its soft drop mass to derive the corrective factor. The data vs. simulation plot for the invariant mass of the leading jet is shown in Fig. 4.29. We see that the loose VR is effectively dominated by $t\bar{t}$ events but, without a corrective factor for the $t\bar{t}$ cross section, the data-MC agreement is not optimal.

The dominant background in the loose VR is the QCD multijet production. The shape of such background is treated in an analogous way as it has been done in Section 4.6: a control region in data, enriched in QCD events, is defined and used to estimate the distribution of the soft drop mass of the leading jet for multijet events. Since one of the critical requests to identify boosted jets induced by the decay of top quarks is the presence of b tagged subjets, we first revert this condition asking for events with AK8 jets with no b tagged subjets. Then, in order to enhance the purity in QCD events, we revert the cut on TvsQCD for the leading jet, asking for it to be in the window $[0.5, 0.7]$. This cut also makes the CR orthogonal to the requirements of the loose VR. The selection requirements that form the QCD CR for the

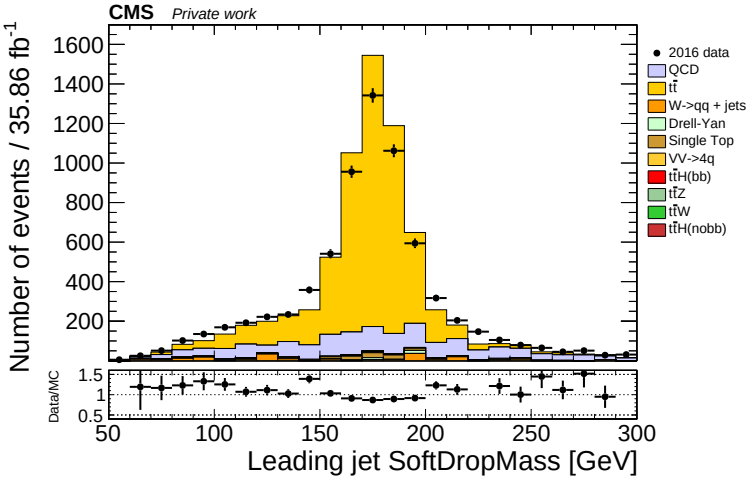

Figure 4.29: Soft drop mass of the leading jet in the loose VR without corrective factor on the $t\bar{t}$ cross section. The data-MC agreement is not optimal.

loose VR selection are summarized in Table 4.9. The data vs. simulation plot for the invariant mass of the leading jet for events in the CR is shown in Fig. 4.30, where we see that the CR is effectively dominated by QCD events. Following the same steps that we developed for the QCD estimation in the signal categories, we perform a closure test in order to verify that the loose VR and the related CR are kinematically close, comparing the shape of the soft drop mass of the leading jet for the loose VR and CR selections. Such test is shown in Fig. 4.30, where we see that a good agreement is achieved between the two distributions. The residual kinematic differences are accounted for by fitting the ratio plot with a straight line, which is used as a transition function to correct the CR distribution in data to match the distribution in the loose VR. This is shown in Fig. 4.30, where we see that a reasonably good agreement is obtained between the corrected QCD shape and the simulated shape.

Having in hand the shapes of the soft drop mass of the leading jet for the $t\bar{t}$ signal (from the simulation) and for the QCD dominant background (from the data), we set up the procedure to correct the $t\bar{t}$ cross section. First we note that the expected $t\bar{t}$ yield obtained using the nominal cross section $\sigma_{t\bar{t}} = 832$ pb, and given the selection efficiency ε and the 2016 luminosity $\mathcal{L}$, is found to be

$$N_{t\bar{t}}^{\text{exp}} = \sigma_{t\bar{t}} \cdot \mathcal{L} \cdot \varepsilon = 5188. \tag{4.10}$$

Then, the idea is to use the $t\bar{t}$ and QCD shapes to fit the soft drop mass distribution of the leading jet in data with a function of the form

$$D(m_{\text{SD}}) = N_{t\bar{t}}^{\text{fit}} \cdot S(m_{\text{SD}}) + N_{\text{QCD}}^{\text{fit}} \cdot B(m_{\text{SD}}), \tag{4.11}$$

Observable	Loose VR	CR
-	Signal trigger	Signal trigger
N_{leptons}	$= 0$	$= 0$
N_{AK8jets}	≥ 2	≥ 2
Leading jet p_{T}	> 300 GeV	> 300 GeV
Second-leading jet p_{T}	> 300 GeV	> 300 GeV
Leading jet $N_{\text{b-subjet}}$	> 0	$= 0$
Second-leading jet $N_{\text{b-subjet}}$	> 0	$= 0$
Leading jet TvsQCD	> 0.7	$[0.5, 0.7]$
Leading jet HvsT	< 0.1	< 0.1

Table 4.9: Summary of the requirements defining the loose $\mathrm{t\bar{t}}$ VR and its CR.

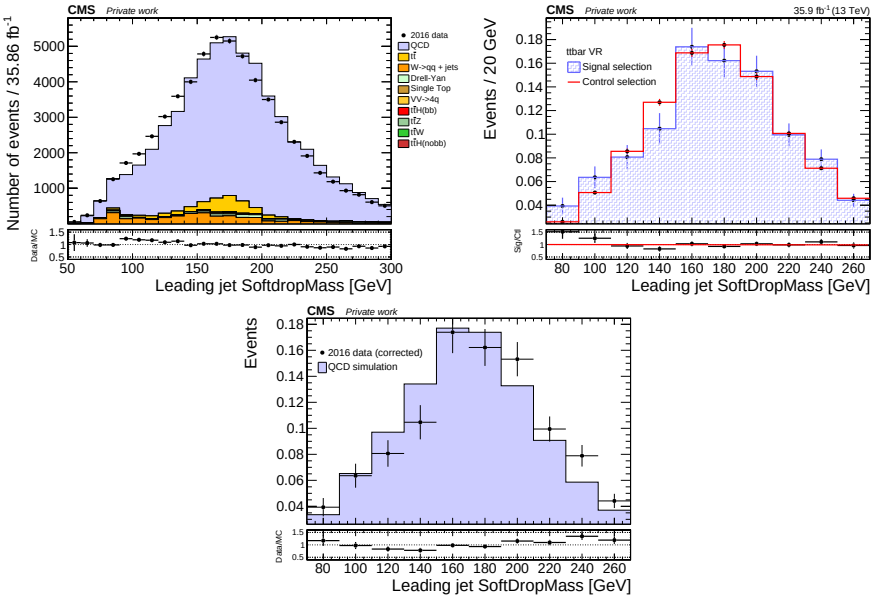

Figure 4.30: Data-MC agreement for events in the CR of the loose $\mathrm{t\bar{t}}$ VR (upper left); closure test for the QCD estimation in the loose $\mathrm{t\bar{t}}$ VR (upper right); comparison between corrected QCD shape in data and MC simulation in the loose $\mathrm{t\bar{t}}$ VR (lower panel).

where $S(m_{\mathrm{SD}})$ and $B(m_{\mathrm{SD}})$ are two probability density functions (pdf) describing the signal and background shapes respectively, while $N^{\mathrm{fit}}_{\mathrm{t\bar{t}}}$ and $N^{\mathrm{fit}}_{\mathrm{QCD}}$ are the free parameters which are fitted from the data. Once the number of $\mathrm{t\bar{t}}$ events actually observed $N^{\mathrm{fit}}_{\mathrm{t\bar{t}}}$ is obtained, we can compare it with the expected number of events coming from Eq. 4.10 and obtain the corrective factor for the cross section.

All the maximum-likelihood fits described in the following have been performed using the RooFit toolkit [83]. As a first step, we fit the soft drop mass distribution of the leading jet for the $\mathrm{t\bar{t}}$ simulated events in order to obtain the pdf $S(m_{\mathrm{SD}})$ describing the signal process. The signal model is chosen to be the sum of a Crystal Ball function and a third-order Bernstein polynomial. As a second step, we fit the soft drop mass distribution of the leading jet for the QCD events in data, corrected with the transition function, in order to obtain the pdf $B(m_{\mathrm{SD}})$ describing the background process. The background model is chosen to be the sum of a Crystal Ball function and a fifth-order Bernstein polynomial. The outcome of this two fits is shown in Fig. 4.31, upper row. Finally, keeping frozen the parameters describing the signal and background models, we use such models to perform a final maximum-likelihood fit to the data, where the data model is given by Eq. 4.11. The outcome of this final fit is shown in Fig. 4.31, lower row. The fitted values for the number of signal and background events are found to be

$$\begin{aligned} N^{\mathrm{fit}}_{\mathrm{t\bar{t}}} &= 4084 \pm 106 \\ N^{\mathrm{fit}}_{\mathrm{QCD}} &= 2836 \pm 97. \end{aligned} \tag{4.12}$$

By comparing the fitted number of $\mathrm{t\bar{t}}$ events with the expected one from simulation with nominal cross section, we obtain a correction factor

$$r = \frac{\sigma^{\mathrm{fit}}_{\mathrm{t\bar{t}}}}{\sigma^{\mathrm{exp}}_{\mathrm{t\bar{t}}}} = \frac{N^{\mathrm{fit}}_{\mathrm{t\bar{t}}}}{N^{\mathrm{exp}}_{\mathrm{t\bar{t}}}} = 0.787. \tag{4.13}$$

This result is found to be in agreement with the results found in [34, 82]. By comparing this value with similar studies performed in the boosted-Higgs decay channel, we assume a correction factor $r = 0.75$ for the $\mathrm{t\bar{t}}$ cross section, which translates in an effective cross section of 0.75×832 pb $= 624$ pb. Whenever the $\mathrm{t\bar{t}}$ simulation is used in this analysis, the effective cross section is implied, unless otherwise stated.

Figure 4.32 reports the data vs. simulation plot for the invariant mass of the leading jet in the loose VR, where the effective cross section is used to normalize the $\mathrm{t\bar{t}}$ shape. We see that the data-MC agreement is enhanced with respect to Fig. 4.29.

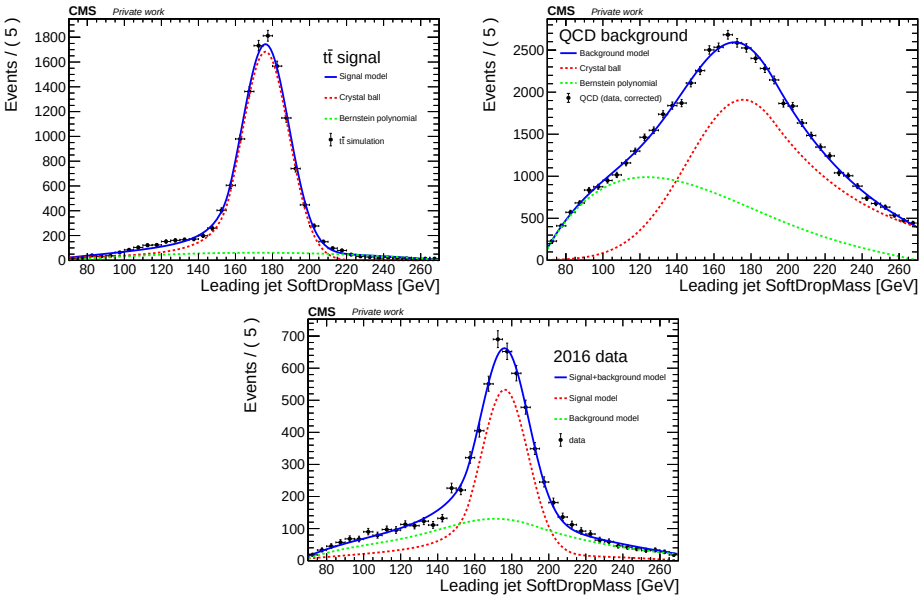

Figure 4.31: Maximum-likelihood fits used to derive the corrective factor for the $t\bar{t}$ cross section.

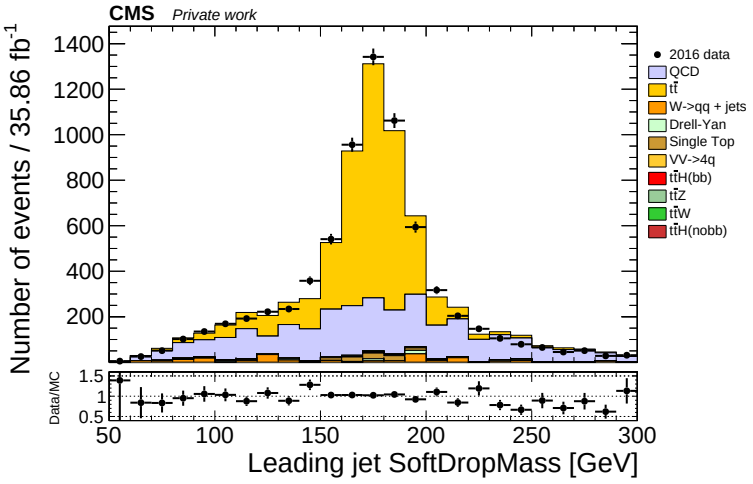

Figure 4.32: Soft drop mass of the leading jet in the loose $t\bar{t}$ VR with corrective factor on the $t\bar{t}$ cross section.

4.7.2 Tight $t\bar{t}$ validation region

A widely used strategy in the CMS Collaboration, aimed at reducing the impact of systematic uncertainties affecting a given background, is to define a background-enriched region and to include it as an additional category into the final simultaneous fit used to extract the signal strength parameter. If the nuisance parameters affecting the background process are assumed to be correlated between the categories, a constraint on such nuisance parameters is expected (see Chapter 5 for a detailed description of the statistical treatment of the data employed in this analysis). Thus, a $t\bar{t}$-enriched region is defined, which will enter the simultaneous fit, with the purpose of constraining the uncertainties related to the $t\bar{t}$ simulation. We will refer to this region as tight VR, as it is inspired by the selection criteria given in Subsection 4.7.1 but assumes somewhat more stringent cuts.

The selection criteria defining the tight VR are identical to the ones listed in Table 4.9, with the exception of the number of AK8 jets, which is required to be exactly equal to two, and with the additional request for the number of AK4, b tagged jets to be exactly zero. This latter cut makes the tight VR orthogonal to the signal categories, making it possible to include it in the simultaneous fit, and is also a natural request since in the highly boosted topology selected by the VR, both the b quarks are expected to be collected inside the two AK8 jets.

In an identical fashion to what it has been done in Subsection 4.7.1, the shape of the QCD multijet production in the tight VR is estimated from a CR in data. The cuts defining the CR are unchanged with respect to the ones introduced previously. The selection requirements that form the tight VR and the corresponding CR are summarized in Table 4.10. Figure 4.33 shows the checks used to validate the choice of the CR, namely the closure test evaluating the agreement between the signal and control selections and the comparison between the corrected shape in data and the signal selection. Also, the soft drop mass of the leading jet in the tight VR is shown both for data and simulated events.

Concerning the estimation of the QCD yield in the tight VR, we once again proceed in an analogous fashion to what it has been done in the previous sections: the ABCD method is used to extract the QCD yield from data. The variables used to set up the procedure are the BDT_TvsQCD score and the number of b tagged AK4 jets, which show a low degree of correlation, since no information regarding AK4 jets is used in the training of the boosted-jets BDTs. Events fulfilling the requirements of Table 4.10, first column, are used, except for the cuts on the TvsQ score and the number of b tagged AK4 jets, which are not applied in order to define the four regions A, B, C and D shown in Fig. 4.34. Events in region A effectively fulfill the requirements of the aforementioned table and thus belong to the tight VR. This means that no fractions of events are needed in this case. Identical

Observable	Tight VR	CR
-	Signal trigger	Signal trigger
N_{leptons}	$= 0$	$= 0$
N_{AK8jets}	$= 2$	$= 2$
$N_{\text{AK4b-jets}}$	$= 0$	$= 0$
Leading jet p_{T}	> 300 GeV	> 300 GeV
Second-leading jet p_{T}	> 300 GeV	> 300 GeV
Leading jet $N_{\text{b-subjet}}$	> 0	$= 0$
Second-leading jet $N_{\text{b-subjet}}$	> 0	$= 0$
Leading jet TvsQCD	> 0.7	$[0.5, 0.7]$
Leading jet HvsT	< 0.1	< 0.1

Table 4.10: Summary of the requirements defining the tight $\mathrm{t\bar{t}}$ VR and its CR.

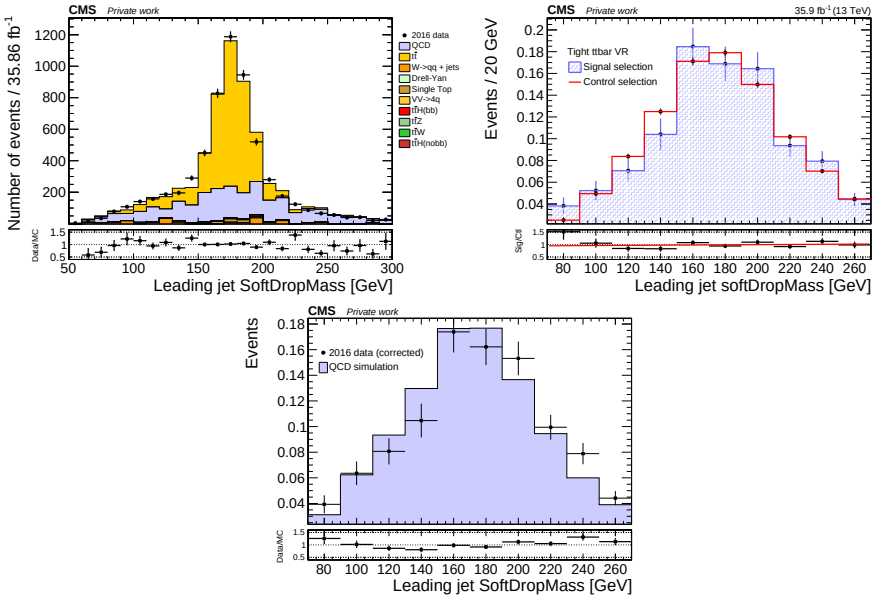

Figure 4.33: Data-MC agreement for events in the tight VR (upper left); closure test for the QCD estimation in the tight VR (upper right); comparison between corrected QCD shape in data and MC simulation in the tight VR (lower panel).

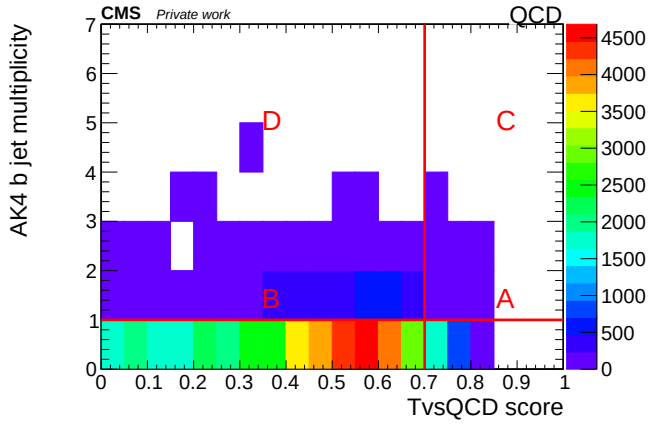

Figure 4.34: ABCD regions for the QCD yield estimation in the tight VR.

steps to the ones developed before lead to an expected QCD yield in the tight VR of 2683 ± 91 events.

For ease of reading, in the following we will simply refer to the tight $\mathrm{t\bar{t}}$ VR as the "$\mathrm{t\bar{t}}$ VR", since this region is the one entering the final fit used to extract the signal. The loose $\mathrm{t\bar{t}}$ VR, on the other hand, is only exploited to extract the correction to the $\mathrm{t\bar{t}}$ cross section and will not be further used.

4.8 Template shapes

The value of the signal strength modifier $\mu_{\mathrm{t\bar{t}H}}$ is obtained by performing a binned maximum-likelihood fit to the data. Such kind of method is called shape analysis, and is described in full detail in Chapter 5. The observable used in the fit is the invariant mass of the resolved Higgs boson candidate in the four signal categories; in addition, the soft drop mass of the leading jet in the $\mathrm{t\bar{t}}$-enriched VR is added to the fitting procedure, in order to constrain the systematic uncertainties related to the $\mathrm{t\bar{t}}$ events, as it will be discussed in Subsection 4.9.2. The estimation of the composition of the data sample is achieved taking into account the following processes:

1. $\mathrm{t\bar{t}H(b\bar{b})}$;
2. $\mathrm{t\bar{t}H(nob\bar{b})}$;
3. QCD;
4. $\mathrm{t\bar{t}}$;
5. $\mathrm{t\bar{t}Z(b\bar{b})}$;
6. subdominant backgrounds.

The $t\bar{t}H(b\bar{b})$ and $t\bar{t}H(nob\bar{b})$ processes are both considered to contribute to the signal and are estimated using the simulations, while QCD and $t\bar{t}$ are the main backgrounds of the analysis. As it has been described in the previous sections, the former is entirely estimated from data, while the latter is estimated from simulation. In addition, minor backgrounds are considered.

The $t\bar{t}Z(b\bar{b})$ process is an irreducible background of this search, since its topological signature in the detector cannot be distinguished from the one of signal events. Also, its theoretical cross section is comparable to that of $t\bar{t}H(b\bar{b})$, and thus it is included in the fit as an independent template. Such background is estimated completely from simulation.

Finally, we consider the subdominant backgrounds. They consist of nine processes: single top quark and top antiquark production in the t-channel and in association with a W boson, Drell-Yan events with the production of a $q\bar{q}$ pair and additional jets, $W \to q\bar{q}$ and additional jets, $WW \to q\bar{q}q\bar{q}$, $ZZ \to q\bar{q}q\bar{q}$, $t\bar{t}W(q\bar{q})$. Such processes show in general a low selection efficiency but can have a large cross section compared to the $t\bar{t}H(b\bar{b})$ one. This can lead to spikes in the distribution due to poorly populated distributions of events with high cross section. In order to partially mitigate this effect, the subdominant backgrounds are added together to form a single template, taking into account the proportions given by the theoretical cross sections. The subdominant backgrounds are all modeled using the simulations. As an example of template shapes entering the maximum-likelihood fit, in Fig. 4.35 we present the distributions of the mass of the resolved Higgs boson candidate in category 9. As a result of the signal selection, a sharp peak around the Higgs boson mass in the $t\bar{t}H(b\bar{b})$ template and, to a less extent, in the $t\bar{t}H(nob\bar{b})$ template, is found. On the other hand, background events show smoother behaviors, in such a way that a good discriminating power between signal and background is found in the shapes used in the fit. The set of all the template shapes entering the maximum-likelihood fit is presented in Appendix A.

4.9 Systematic uncertainties

This section describes in full detail the sources of systematic uncertainty affecting the measurement and the way they are treated. All the sources described in the following are implemented as nuisance parameters entering the likelihood function which is used to compute the signal strength value and the corresponding upper limit. For an exhaustive description of how nuisance parameters enter the likelihood and how they impact on it, we refer to the matter described in Chapter 5 of this work.

In general, the sources of systematic uncertainty can be categorized based on their effect on the template shapes entering the final fit. First, we have rate uncertainties, which are supposed to change the expected yield of a given

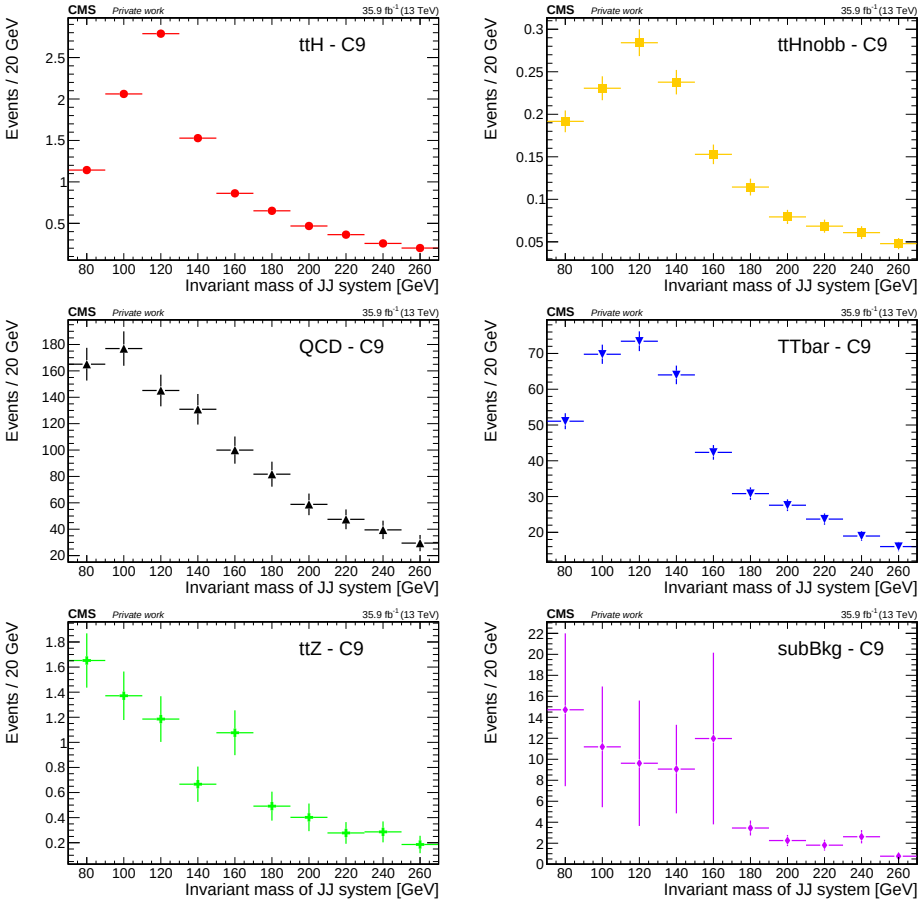

Figure 4.35: Template shapes for events in category 9. Signal processes are shown in the upper row, while middle and lower rows show background processes. Each template is normalized to the expected yield in the 2016 data-taking period.

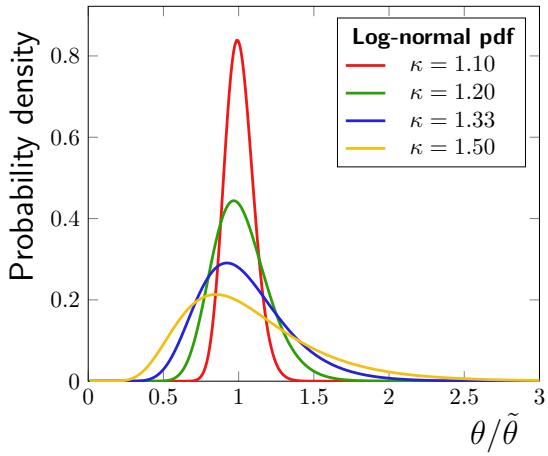

Figure 4.36: Log-normal pdf, as a function of $\theta/\tilde{\theta}$, for different values of the parameter κ. As κ becomes larger, the pdf broadens, reflecting the increased change induced in the observable affected by the lnN uncertainty.

process leaving the shape of the related distribution untouched. In order to parameterize this kind of effect, a nuisance parameter θ is assumed to follow a pdf $\rho(\theta)$ of type log-normal (lnN), which is best suited for uncertainties affecting positively defined observables like yields, cross sections, efficiencies, etc. The variable θ follows a lnN pdf if and only if the variable $N = \ln\theta$ follows a normal distribution with expected value $\tilde{\theta}$. The expression of the lnN distribution reads

$$\rho(\theta) = \frac{1}{\sqrt{2\pi}\ln\kappa}\exp\left\{-\frac{\left[\ln\left(\theta/\tilde{\theta}\right)\right]^2}{2\left(\ln\kappa\right)^2}\right\}\frac{1}{\theta} \tag{4.14}$$

where the width of the log-normal pdf is characterized by κ. The width is used to state by which factor a given observable can change, *e.g.*, $\kappa = 1.10$ implies that the observable can be larger or smaller than the nominal value by a factor 1.10, with both deviations having a chance of 16%. The behavior of the lnN pdf for different values of the parameter κ is shown in Fig. 4.36.

On the other side, we have shape uncertainties, which are supposed to change the actual shape of the observable. These uncertainties are implemented in the simultaneous fit by providing alternative shapes to the framework. If the change in shape also implies a change in yield, this is again treated with a lnN uncertainty.

The sources of uncertainties taken into account in this analysis can be split in two main categories: experimental uncertainties and theoretical uncertainties. The former group can be essentially described as the group of uncertainties which arise from differences that are found between the

simulation and the actual data in the description of the detector performance. These uncertainties can in principle be partially reduced with the increased integrated luminosity collected by the experiments, which leads to a better understanding of the detectors, enhanced calibrations of objects and algorithms, etc. The latter group is instead related to uncertainties in the theoretical parameters describing the physical processes, which cannot be reduced (at least, not directly) by collecting more data, but should rather be constrained by developments at the theory level. In the following, we list all the sources of systematic uncertainty affecting the measurements and, in the coming subsections, we will investigate more deeply some of such sources. A summary of the systematic uncertainties is given at the end of the section, in Table 4.11.

Experimental uncertainties

- Jet energy corrections: several strategies have been developed inside the CMS Collaboration [79, 84] to implement corrective procedures for the jet energy scale and resolution, which have associated uncertainties that translate to systematic uncertainties on the observable of this analysis. Therefore, in MC simulations, jets belonging to the $t\bar{t}H(b\bar{b})$ and $t\bar{t}H(nob\bar{b})$ signal samples and to the $t\bar{t}$ dominant background are shifted (smeared) according to the JES (JER) η-dependent uncertainties and new templates are obtained. Since such uncertainties affect the p_T of jets, which is extensively used in the event selection, a change in yield, as well as a change in shape, is expected. Given the above reasoning, the JES and JEC uncertainties are included in the simultaneous fit as rate+shape uncertainties;
- QCD background prediction: both shapes and yields for the QCD dominant background are estimated from data. Nevertheless, some degree of uncertainty is related to the shapes, since they are extracted from a control region using transition functions resulting from a fitting procedure. Thus, the parameters describing the transition functions have an uncertainty coming from the fit, which translates into an uncertainty on the shape. Also, in the final estimation of the yields, QCD simulation is used in order to compute the fractions of events obtained by the ABCD method that fall in the signal categories. This will introduce an uncertainty on the yields, as it will be discussed in Subsection 4.9.1;
- $t\bar{t}$ background modeling uncertainties: in order to achieve a good modeling of $t\bar{t}$ events, the parameters entering the MC simulation need to be properly tuned. The tuning of such parameters is achieved with some degree of uncertainty, which reflects on the

shapes and rates of $t\bar{t}$ events. This subject is described in more detail in Subsection 4.9.2;

- $t\bar{t}Z$ and subdominant backgrounds modeling: a systematic uncertainty of type lnN is assigned to the rates of minor backgrounds, based on how well the simulations are known to predict the various processes. Since these rate uncertainties are supposed to cover a wide range of possible mismodellings and the correlation among them is unclear, we assume them to be uncorrelated between categories. A 20% uncertainty (*i.e.*, $\kappa = 1.20$) is associated with the $t\bar{t}Z$ production, while a 50% uncertainty (*i.e.*, $\kappa = 1.50$) is associated with the subdominant backgrounds;
- b tagging scale factors: since differences in the performance of the CSVv2 b tagging algorithm in MC events have been found with respect to data events, scale factors must be applied to the simulations to enhance the description of the data. The extraction of such SF is obtained with the method summarized in Subsection 4.2.1, which is affected by several sources of systematic uncertainty. This matter is described in detail in Subsection 4.9.3;
- Trigger scale factor: differences in the description of the signal trigger efficiency in data and MC are taken into account by fitting the ratio plot in in Fig. 4.7 with a constant straight line. The y-intercept resulting from the fit comes with an uncertainty, which in principle translates to an uncertainty on the SF. However, the relative uncertainty on this parameter is very small, in such a way that no meaningful change in rates or shapes is found by shifting up or down the y-intercept within its uncertainty. For this reason, the trigger scale factor uncertainty is neglected and eventually does not contribute to the experimental uncertainties;
- Pileup scale factor: the distribution of the number of PU interactions in the simulation shows discrepancies with respect to the one obtained from actual data events. Therefore, the PU distribution in MC events needs to be corrected to match the data. This is done by taking the ratio between the PU distributions in data and simulation and by applying the resulting weight distribution on a per-bin basis to each event. In order to assess the uncertainty on this procedure, which is described in more detail in Subsection 4.1.2, the total inelastic proton-proton cross section is shifted up and down by its uncertainty, which is found to be 4.6%, and new distributions for the number of PU interactions in data are obtained, which translate in upwards and downwards shifted pileup SFs. The results of this procedure are presented in Fig. 4.37. The shifted SFs are then applied to the simulations to obtain upwards and downwards shifted shapes. This uncertainty is assumed to

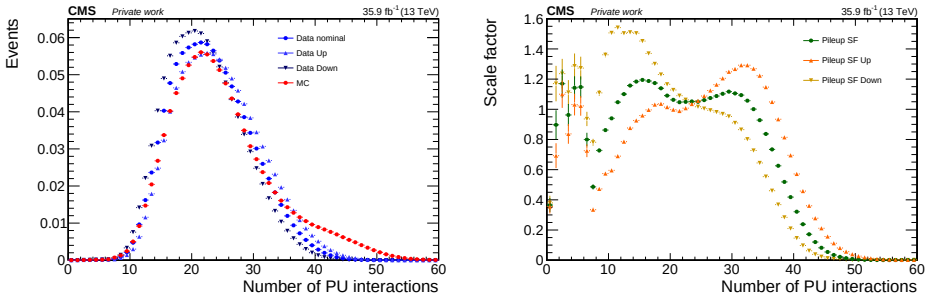

Figure 4.37: Distributions of the number of pileup interactions in data and MC (left) and pileup scale factors (right). The upwards (downwards) shifted distributions in data are obtained by shifting up (down) the nominal, total inelastic proton-proton cross section by 4.6%. The ratios between the data distributions and the MC distribution result in the scale factors.

impact both the rate and the shape of processes and is considered to be correlated between processes and categories;

– Luminosity uncertainty: since the overall uncertainty on the value of the integrated luminosity for the 2016 data-taking period has been estimated to be 2.5% [85], we assign a rate uncertainty of type lnN corresponding to this value ($\kappa = 1.025$) to all the processes, except for QCD, which is estimated directly from data. This uncertainty is assumed to be correlated between processes and categories;

– Finite size of the MC samples: the simultaneous fit used in this analysis to extract the value of the signal strength involves the estimation of the composition of the data sample based on MC events. In a shape analysis (see Section 5.1), data are binned, so that a few MC events can be found in some regions of the observable, leading to high fluctuations. In order to incorporate the finite statistics of simulated samples, the Barlow–Beeston method [86] is used. This method assigns a single nuisance parameter to scale the sum of process yields in a given bin, instead of requiring separate parameters, one for each process, thus reducing the number of parameters required in the maximum-likelihood fit.

Theoretical uncertainties

– PDFs: parton distribution functions (PDFs) describe the internal structure of protons in terms of their quark and gluon constituents and of the momentum fraction carried by each of them. A precise knowlege of PDFs is crucial ingredient of precision measurements at the LHC. PDFs are described in terms of several parameters

that should be extracted by looking at the data; for example one parametrization among many possible can be

$$xf(x) = a_0 x^{a_1}(1-x)^{a_2} \exp\left\{a_3 x + a_4 x^2 + a_5\sqrt{x} + a_6 x^{-a_7}\right\}$$

where $f(x)$ is the PDF, x is the fraction of the total momentum carried by the parton and $\vec{a}$ is a set of parameter describing the PDF. To account for uncertainties on the set of parameters, the approach used by the NNPDF collaboration is used. In this method, called MC replicas method [87], a set of $N_{\text{rep}} = 100$ replica copies of the vector of parameters $\vec{a}$ is given, and the observable used in this analysis is then computed repeating its determination N_{rep} times, each time using a different parameter replica. Then, the root mean square (RMS) of these 100 histograms with respect to the nominal is computed on a per-bin basis. The up and down shifted shapes used in the final fit are finally computed as nominal+RMS and nominal−RMS for each bin, respectively. This uncertainty is assumed to affect the $t\bar{t}H(b\bar{b})$ and $t\bar{t}H(nob\bar{b})$ signal and the $t\bar{t}$ dominant background as a rate+shape uncertainty, which is taken to be correlated between processes and categories;

– Renormalization and factorization scales: when computing cross sections for processes arising from pp collisions, the amplitude of such processes may often be affected by ultraviolet and infrared divergences. Such divergences are cured by introducing in the calculations the renormalization scale μ_{R} and the factorization scale μ_{F} respectively. Evidently, μ_{R} and μ_{F} are fictitious parameters that are introduced *ad hoc*, and physical observables should ideally not depend on them. The community of theoretical physicists has developed ways to treat this problem and to assess the theoretical uncertainty arising from this procedure. In the simulation, weights are applied to the events corresponding to different μ_{R} and μ_{F} choices. Once again, the RMS of the resulting histograms with respect to the nominal is used to obtain up and down shifted shapes as nominal+RMS and nominal−RMS for each bin. This uncertainty is assumed to affect the $t\bar{t}H(b\bar{b})$ and $t\bar{t}H(nob\bar{b})$ signal and the $t\bar{t}$ dominant background as a rate+shape uncertainty, which is taken to be correlated between processes and categories;

– Strong coupling: weights are applied to the simulated events which reflect the theoretical uncertainty on the strong coupling α_S. The RMS method is used to obtain the shifted shapes. This uncertainty is assumed to affect the $t\bar{t}H(b\bar{b})$ and $t\bar{t}H(nob\bar{b})$ signal and to the $t\bar{t}$ dominant background as a rate+shape uncertainty, which is taken to be correlated between processes and categories.

4.9.1 Uncertainties on the QCD prediction

The shapes for the dominant QCD background are obtained, both for the signal categories and for the VR, from control samples in data. Transition functions, which are taken to be straight lines, are then applied in order to translate these shapes into the signal regions. Such transition functions are obtained by fits to the ratios in Figs. 4.26 and 4.33, each one giving as a result two parameters, the y-intercept q and slope m of the straight line. These two parameters come with an uncertainty, which translates into an uncertainty on the shape of the QCD observable. In general, the covariance matrix $\mathcal{V}$ of the fit is not diagonal, meaning that there is some degree of correlation between the two parameters. In order to properly take this into account, the following procedure is used. Since the covariance matrix is a real, symmetric matrix, it can always be diagonalized by means of an orthogonal transformation $\mathcal{O}$. Starting from the "real" parameters, described by the vector $\mathbf{p}^T = (q, m)$, we first transform them to some auxiliary parameters $\tilde{\mathbf{p}}^T = (\tilde{q}, \tilde{m})$:

$$\tilde{\mathbf{p}} = \mathcal{O}\mathbf{p}, \tag{4.15}$$

where the orthogonal matrix $\mathcal{O}$ has the eigenvectors of $\mathcal{V}$ as columns. In this auxiliary space, the covariance matrix $\mathcal{D}$ of the auxiliary parameters is diagonal, is computed as

$$\mathcal{D} = \mathcal{O}^{-1}\mathcal{V}\mathcal{O} = \mathcal{O}^T\mathcal{V}\mathcal{O} \tag{4.16}$$

and has the eigenvalues of $\mathcal{V}$ as diagonal elements, which can be interpreted as the variances of the parameters in the auxiliary space:

$$\mathcal{D} = \begin{bmatrix} \tilde{\sigma}_{\tilde{q}}^2 & 0 \\ 0 & \tilde{\sigma}_{\tilde{m}}^2 \end{bmatrix}. \tag{4.17}$$

Since in this auxiliary space the two parameters are fully decorrelated, they can be shifted up and down freely and independently. Thus, we define the upwards (downwards) shifted transition function in the auxiliary space as the transition function described by $\tilde{\mathbf{p}}_{\text{up}}^T = (\tilde{q} + \tilde{\sigma}_{\tilde{q}}, \tilde{m} + \tilde{\sigma}_{\tilde{m}})$ ($\tilde{\mathbf{p}}_{\text{down}}^T = (\tilde{q} - \tilde{\sigma}_{\tilde{q}}, \tilde{m} - \tilde{\sigma}_{\tilde{m}})$). Finally, the real parameters describing the upwards (downwards) shifted transition function in the real parameter space are obtained by performing the inverse transformation:

$$\mathbf{p}_{\text{up}} = \mathcal{O}^{-1}\tilde{\mathbf{p}}_{\text{up}} \qquad \mathbf{p}_{\text{down}} = \mathcal{O}^{-1}\tilde{\mathbf{p}}_{\text{down}}. \tag{4.18}$$

The transition functions described by $\mathbf{p}_{\text{up}}$ and $\mathbf{p}_{\text{down}}$ are applied to the QCD distributions in the control region to obtain upwards and downwards shifted shapes that are included as a shape uncertainty in the simultaneous fit. Since the parameters describing the transition functions come from five

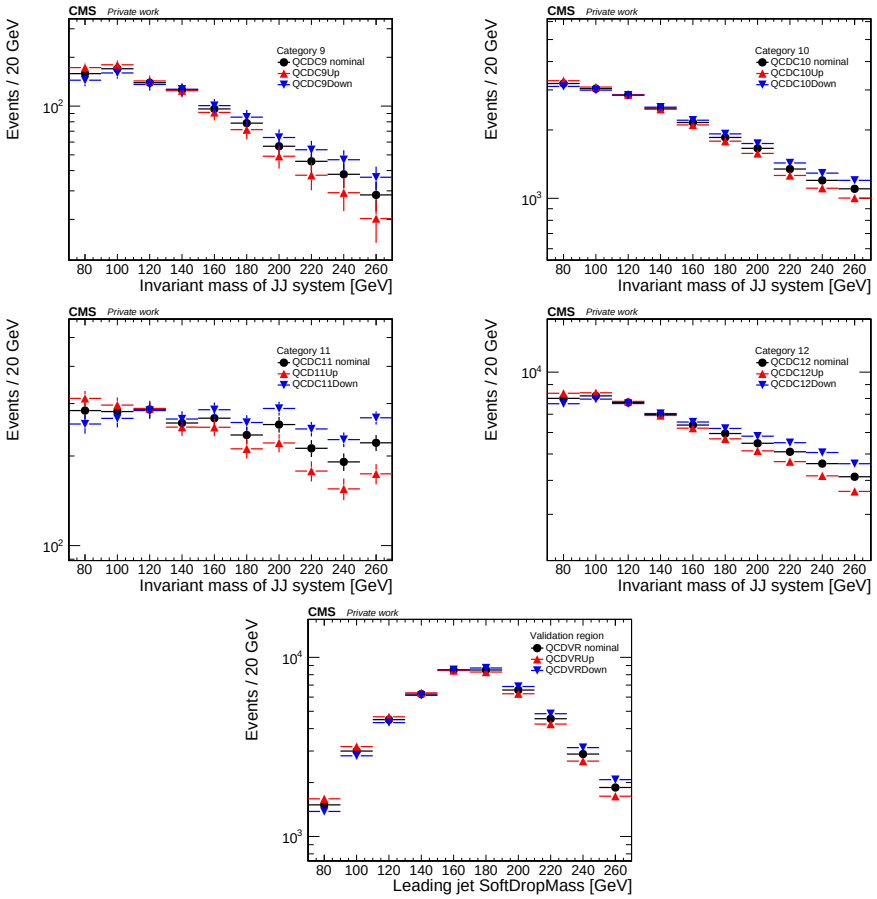

Figure 4.38: Shape uncertainty for the QCD background in category 9 and 10 (upper row), category 11 and 12 (middle row) and in the VR (lower row).

independent fits, the QCD shape uncertainty is assumed to be uncorrelated between categories. Figure 4.38 shows the nominal QCD shapes together with the upwards and downward shifted shapes for each signal category and for the VR. The discrepancy between nominal and shifted shapes is bigger in categories 9 and 11, where the smaller MC population translates to bigger error bars in the ratios of Fig. 4.26 and subsequently to larger uncertainties on the fit parameters.

The QCD yields are estimated from data using the ABCD method. However, MC QCD events are used to determine the fractions of events falling in region A that also fall into signal categories. In order to assess the degree of uncertainty introduced by the use of the simulation, we check the data-MC agreement for the variables that are used to split events in categories, namely the number of AK8 jets, the number of AK4 jets and the number of AK4, b tagged jets. From the result of this check, shown in

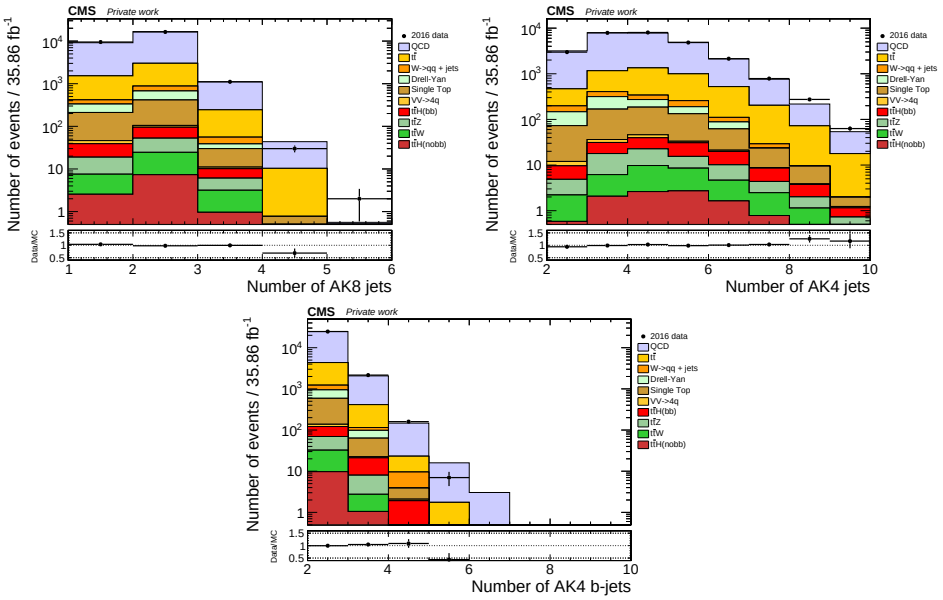

Figure 4.39: Data-MC agreement for the variables used to categorize the events. The AK8 and AK4 multiplicities are shown in the upper row, while the multiplicity of AK4, b tagged jets is shown in the lower row.

Fig. 4.39, we argue that the simulation describes pretty well the variables under investigation and thus, the fraction of events are well modeled. In order to account for the small disagreement between data and MC events, a 10% uncertainty on the QCD yield of each category is introduced. This uncertainty is assumed to be uncorrelated between categories. Also, an additional uncertainty is assumed to affect the QCD yields, coming from the number of events obtained by making use of the ABCD method. Such numbers come with an uncertainty, as we shown in Table 4.8 and Subsection 4.7.2. Thus, a lnN rate uncertainty equal to the relative error on such numbers is assigned, which affects categories with the same AK8 multiplicity in a correlated way. Based on the numbers reported in the previous sections, we assign a 11% uncertainty to categories 9 and 10, a 4% uncertainty to categories 11 and 12, and a 3% uncertainty to the VR.

4.9.2 Uncertainties on the $t\bar{t}$ production tune

An accurate modelling of the $t\bar{t}$ pair production is a crucial aspect to many analyses. This is particulary true at the LHC, where the large integrated luminosity collected by the experiments leads to measurements in which modelling uncertainties dominate over the statistical uncertainties. Thus, a proper tuning of the MC parameters must be performed in order to improve the description of the data. The $t\bar{t}$ simulated samples used in this analysis rely

on the so called CUETP8M2T4 tune, which has been developed to address some mismodellings in the jet multiplicity of previous tunes to correctly describe differential measurements. Several parameters describing the tune are extracted from 8 and 13 TeV data [88] by fitting differential distributions of several quantities of interest. The most relevant parameters of the tune are shifted up and down according to their uncertainties to obtain new MC samples which are subsequently used to compute new $t\bar{t}$ template shapes. The parameters taken into account are:

- h_{damp}, controlling the matrix element to parton shower (PS) matching and regulating the high-p_{T} radiation by damping real emission in POWHEG with a factor $h^2_{\text{damp}}/(p^2_{\text{T}} + h^2_{\text{damp}})$
- ISR, controlling the initial state radiation at PS level
- FSR, controlling the final state radiation at PS level
- Tune, controlling global parameters of the tuning

The uncertainties on the $t\bar{t}$ tune are expected to change both the expected rates and shapes. However, we keep the two effects separated in the following way: first, we scale the shifted shapes to the nominal yields, in such a way that the yield-changing effect is factored out. This shape uncertainty is assumed to be correlated between categories. As an example, Fig. 4.40 shows the comparisons between nominal and shifted template shapes for category 10. Then, we introduce a rate uncertainty of type lnN affecting the yield in each category by 20%, which is intended to cover the changes in rate coming for the four sources taken into account, and we assume this uncertainty to be correlated between categories in order to constrain it by fitting simultaneously the signal categories with the VR.

4.9.3 Uncertainties on the b tagging scale factors

The method used to extract the b tagging scale factors relies on a tag-and-probe approach. Dijet events are selected and, applying a selection on event variables and on the properties of one of such jets, samples either enriched in $t\bar{t}$ events or in Z+jets are obtained. The $t\bar{t}$ events, given the top quark decay to a bottom quark, are used to extract a heavy-flavor (HF) scale factor, while events in which a Z boson is produced with additional light jets are used to extract a light-flavor (LF) scale factor. In order to account for light- (heavy-) flavor contamination in the selected heavy (light) flavor samples, simulated events are used. Eventually, the scale factors are obtained as the ratio of CSVv2 distributions in data and MC events, where the non-relevant contamination in data is subtracted using the simulation:

$$SF(\text{CSV}, p_{\text{T}}, \eta) = \frac{\text{Data} - \text{MC}_A}{\text{MC}_B}, \tag{4.19}$$

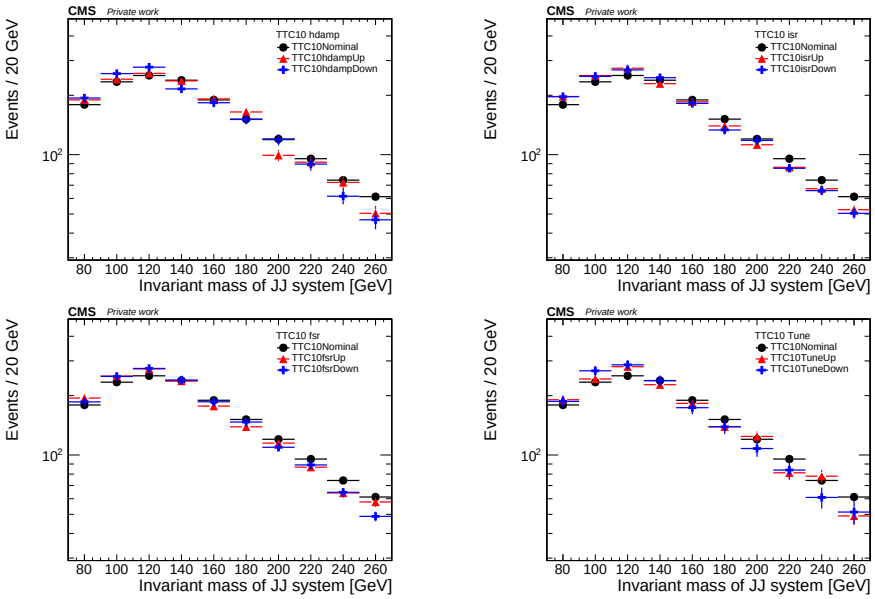

Figure 4.40: Shape uncertainties for the $t\bar{t}$ background in category 10. Changes in shape related to h_{damp} (upper left), ISR (upper right), FSR (lower left) and tune (lower right) are show, along with the nominal shape. Shifted shapes are normalized to the nominal yields, in such a way that differences in the distributions reflect changes in shapes only.

where A, B is the HF or LF component. As Eq. 4.19 testifies, such scale factors are obtained for exclusive bins in CSVv2 score, p_{T} and η of the jet.

As a first source of systematic uncertainty, JES is considered. Instead of nominal MC samples, JES shifted samples are used to recompute the SF. Shifts in the JES change the p_{T} of jets, which could cause events to migrate out of the selected samples or to migrate to different p_{T} bins. The b tagging uncertainty related to the JES is considered to be fully correlated with the "overall" JES described above, which means that, when the JES for the jet kinematics is shifted up or down by one standard deviation, the SF is shifted in the same direction.

As a second source of systematic uncertainty, the purity of the samples used to extract the SF is considered. In both HF and LF calculations, MC is used for the purpose of subtracting the nonrelevant part. Thus, the uncertainty on the MC predictions used in Eq. 4.19 and their propagation to SF must be taken into account. This leads effectively to two sources of systematic uncertainty, one for the HF and one for the LF scale factors.

As a third source of systematic uncertainty, the size of the samples used to extract the SF must be taken into account. To address this problem, two sources of uncertainty are introduced for each flavor, the first assessing for statistical fluctuations that would tend to tilt the SF distribution, while the

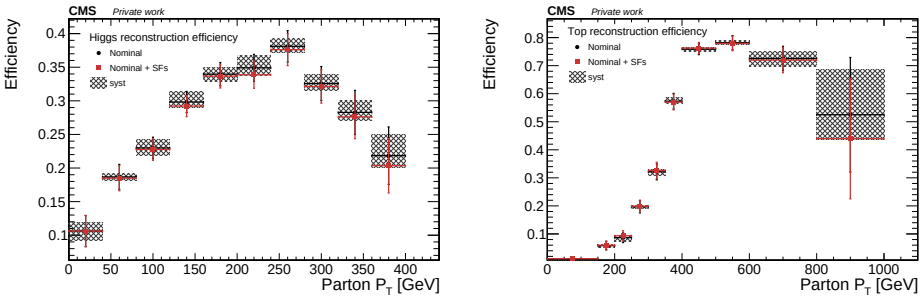

Figure 4.41: Higgs boson (left) and top quark (right) reconstruction efficiencies, before (black line) and after (red line) the application of b tagging scale factors, as a function of the generated p_{T}. The shaded regions corresponds to the nine independent sources of b tagging uncertainty, added together in quadrature.

second accounting for fluctuations that increase or decrease the SF value in the center of the CSVv2 distribution.

Charm-flavored jets are treated separately. For such jets, the scale factor is set to 1.0 and twice the relative uncertainty of b-flavored jets is used. Following the same approach as before, two sources of uncertainty related to the statistical fluctuations are introduced.

Eventually, nine independent sources of systematic uncertainty arise from the extraction of b tagging scale factors. Each of them is assumed to be correlated between different processes and categories. All the processes involved in this analysis are affected by such uncertainties, except for QCD, which is estimated from data and thus has no SF applied. Even though the aforementioned procedure is not supposed to induce any migration of events from one b tag multiplicity bin to another, as each event gets a weight based on a per jet basis, it is nevertheless possible for the global scale factor to be different from one. This is why the sources of uncertainty related to the b tagging procedure are assumed to change both the expected rates and shapes. Since the identification of jets coming from the hadronization of bottom quarks is one of the most important steps in the signal selection of this analysis, studies have been performed to assess the impact of the b tagging scale factors and of the related systematic uncertainties on the Higgs boson and top quark candidates.

First, we consider the Higgs boson and top quark reconstruction efficiencies as a function of the generated p_{T}. Such quantities are defined as the ratio between the number of events in which a generated Higgs boson or top quark is matched with a reconstructed Higgs boson or top quark candidate and the total number of Higgs bosons or top quarks. To perform these studies, we use events passing the signal selection just before the splitting into signal categories.

The reconstruction efficiencies are reported in Fig. 4.41, in which we show that no substantial changes are found in both the distributions when SF are applied. The contributions of the nine independent sources of systematic uncertainty are summed in quadrature and reported as shaded, gray regions. Note the decrease in the Higgs boson reconstruction efficiency at around 300 GeV in the parton p_{T}, which is indeed generally considered to be the threshold between the resolved and boosted decays of Higgs bosons. When boosted decays become dominant, the reconstruction efficiency as a $\mathrm{b\bar{b}}$ system becomes steadily worse.

Second, we study the impact of b tagging SF and related uncertainties on the Higgs boson candidates. This check has been performed both before and after splitting events in categories. Figure 4.42 shows the Higgs boson candidate for events passing the signal selection just before the splitting into signal categories. No considerable changes or distortions are found to be induced by the application of scale factors. The same conclusions can be drawn regarding the Higgs boson candidates for events belonging to the signal categories, which are shown in Fig. 4.42 as well. Given the overall consistency between the distributions before and after the application of b tagging scale factors, we conclude that scaled shapes and yields can be safely used.

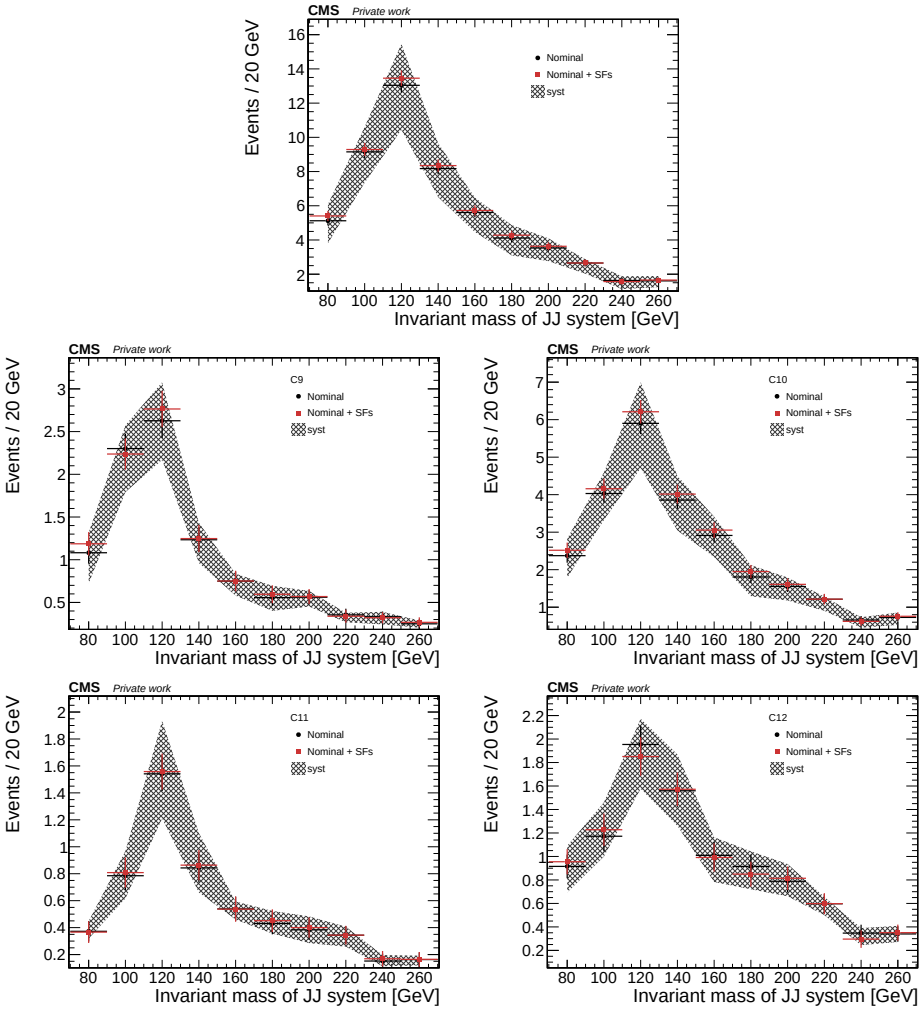

Figure 4.42: Reconstructed Higgs boson candidate before splitting the events into signal categories (upper row) and for events entering the signal categories (middle and lower rows), before (black) and after (red) the application of b tagging scale factors. The shaded regions corresponds to the nine independent sources of b tagging uncertainty, added together in quadrature. The simulated $t\bar{t}H$ events are scaled to the luminosity of the data, in such a way that differences in the distributions reflect both the change in yields and shapes.

Source	Type	Correlation
JES	rate lnN + shape	✓ ✓
JER	rate lnN + shape	✓ ✓
QCD shape	shape	×
QCD fractions	rate lnN	×
QCD ext. yields	rate lnN	✓
$t\bar{t}$ norm	rate lnN	×
h_{damp}	rate lnN + shape	✓
ISR	rate lnN + shape	✓
FSR	rate lnN + shape	✓
Tune	rate lnN + shape	✓
$t\bar{t}Z$ norm	rate lnN	×
subBkg norm	rate lnN	×
b tagging JES	rate lnN + shape	✓ ✓
b tagging HF purity	rate lnN + shape	✓ ✓
b tagging LF purity	rate lnN + shape	✓ ✓
b tagging HF stats1	rate lnN + shape	✓ ✓
b tagging HF stats2	rate lnN + shape	✓ ✓
b tagging LF stats1	rate lnN + shape	✓ ✓
b tagging LF stats2	rate lnN + shape	✓ ✓
b tagging CF stats1	rate lnN + shape	✓ ✓
b tagging CF stats2	rate lnN + shape	✓ ✓
Pileup	rate lnN + shape	✓ ✓
Luminosity	rate lnN	✓ ✓
MC statistics	Barlow–Beeston	×
PDF	rate lnN + shape	✓ ✓
α_S	rate lnN + shape	✓ ✓
μ_F, μ_R	rate lnN + shape	✓ ✓

Table 4.11: Summary of the systematic uncertainties affecting the measurement. All the sources of uncertainty are listed, together with their type and their kind of correlation. A cross symbol (×) means that the uncertainty is uncorrelated between processes and categories, a single checkmark (✓) means that the uncertainty is correlated between different categories only, while a double checkmark (✓ ✓) indicates that the uncertainty is correlated between categories and processes.

Chapter 5

Statistical methods for the search of new phenomena in particle physics

This chapter is devoted to the description of the statistical techniques used in the context of the search of new phenomena in particle physics. The procedure to correctly claim for a discovery and to set upper limits on parameters of interest will be discussed. The material presented in the following is a personal elaboration of the matter discussed in [89] and [90].

5.1 Statistical formalism of a search

The search for new phenomena in particle physics is most usually done in the context of a frequentist statistical test. The goal of a statistical test is to give a quantitative description of how well the observed data agree with some given hypothesis. The hypothesis under consideration is traditionally called the null hypothesis, H_0, and a statement about the validity of H_0 usually involves the comparison with some alternative hypotheses, denoted by H_1, H_2, ..., H_n. If the goal of the test is to discover a signal associated to a new process, one takes the null hypothesis as the one describing only known processes, which can be referred to as the background. This is to be tested against an alternative hypothesis which includes both the new signal and the background. In order for a new signal to be found, the null hypothesis must be rejected with some degree of confidence. On the other hand, if the goal of the test is to set limits on a given parameter of interest (POI), namely to exclude some regions of the parameter space, the signal+background hypothesis plays the role of the null hypothesis, which is tested against the background-only hypothesis. Also in this case, regions in the parameter space are excluded if the null hypothesis is rejected with some degree of confidence.

To summarize the outcome of such searches, it is customary to quantify the level of agreement between the observed data and a given hypothesis H by computing the p-value p of the observation, *i.e.*, the probability, under the assumption of H, to observe data of equal or greater incompatibility with the expectation from H. In particle physics p-values are often converted to equivalent significances, Z. Considering a variable following a standard Gaussian distribution, Z is defined in such a way that when this variable is found Z standard deviations above its mean value, it has an upper tail probability equal to p. Eventually, the following formula can be written,

$$Z = \Phi^{-1}(1 - p), \tag{5.1}$$

where Φ^{-1} is the quantile of the standard Gaussian. As a simple example, suppose that one wants to check whether a coin has been fixed or not. Then, one can take the null hypothesis H_0 to be "the coin is fair" and check if this hypothesis can be rejected. For a fair coin, the probabilities of head and tail are both equal to $1/2$, so that one can simply take the number of heads n_h as a statistical test, which is supposed to follow a binomial distribution with probability $P = 1/2$, namely:

$$\mathcal{B}(n_h; N) = \frac{N!}{n_h!(N - n_h)!}\left(\frac{1}{2}\right)^{n_h}\left(\frac{1}{2}\right)^{(N-n_h)}, \tag{5.2}$$

where N is the number of times the coin has been tossed. Suppose that $N = 20$ tosses are made and $n_h = 17$ heads are observed. Since the expected vale for n_h is $E[n_h] = NP = 10$, then the probability, assuming that the coin is fair, to observe an outcome with equal or greater incompatibility (the p-value) is the sum of the binomial probabilities for $n_h = 0, 1, 2, 3, 17, 18, 19, 20$. Using Eq. 5.2 one gets $p = 0.0026$, which making use of Eq. 5.1 translates to a significance of $Z = 2.79$. One should now decide if the level of disagreement which has been found is sufficient to reject the null hypothesis.

In the context of particle physics searches, there is a general agreement to regard the rejection of background hypotheses with a significance of at least $Z = 5$, corresponding to a p-value of 2.87×10^{-7}, as an appropriate level to claim a discovery. On the other hand, for purposes of setting limits, a threshold p-value of 0.05 (corresponding to the so called 95% confidence level) is usually adopted, which corresponds to $Z = 1.64$.

In particle physics, the problem of assessing discovery or exclusion is usually treated with a frequentist statistical test using a likelihood ratio as a test statistic. The likelihood ratio will, in general, be a function of some POI of the signal process (such as rates, cross sections, signal strengths) and will also contain a set of nuisance parameters which are not known *a priori* but rather must be fitted from the data. The impact of systematic uncertainties on experimental measurements is generally treated assuming that each source of systematic uncertainty enters the likelihood ratio as a nuisance parameter.

To illustrate the use of the profile likelihood ratio, consider a so called shape analysis, namely an experiment where for each selected event one measures the values of some kinematic quantities which can be stored in one or more histograms. Suppose that, for each event in the signal sample, the variable x is measured and the corresponding values are used to construct an histogram $\mathbf{n} = (n_1, n_2, \ldots, n_N)$ composed by N bins. The expectation values of each n_i can be written as

$$E[n_i] = \mu s_i + b_i, \tag{5.3}$$

where μ is the so called signal strength parameter, with $\mu = 0$ corresponding to the background-only hypothesis and $\mu = 1$ corresponding to the nominal signal hypothesis. The mean number of entries in the ith bin from signal and background, s_i and b_i, can be written as

$$\begin{aligned} s_i &= s_{\text{tot}} \int_{\text{bin } i} f_s\left(x; \boldsymbol{\theta}_s\right) dx, \\ b_i &= b_{\text{tot}} \int_{\text{bin } i} f_b\left(x; \boldsymbol{\theta}_b\right) dx. \end{aligned} \tag{5.4}$$

The functions $f_s\left(x; \boldsymbol{\theta}_s\right)$ and $f_b\left(x; \boldsymbol{\theta}_b\right)$ are the probability density functions of the variable x for signal and background events, the integrals in Eq. 5.4 thus representing the probabilities for an event to be found in bin i. Note that from Eq. 5.4 follows that the expected values for the number of entries in each bin depend on the nuisance parameters $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_b$, meaning that systematic uncertainties can smear these expected values, as they are supposed to do. While the quantity b_{tot} can be regarded as a nuisance parameters as well, note that instead the quantity s_{tot} is not an adjustable parameter but is rather fixed to the value predicted by the nominal signal model. In the following we will denote as $\boldsymbol{\theta} = (\boldsymbol{\theta}_s, \boldsymbol{\theta}_b, b_{\text{tot}})$ the full set of nuisance parameters.

In addition to the measured histogram $\mathbf{n}$, one may also want to make further subsidiary measurements in order to constrain the nuisance parameters. For example, many data analyses select a control sample which is mostly populated by background events and, from them, an histogram of some kinematic variable is constructed. This then gives a set of values $\mathbf{m} = (m_1, m_2, \ldots, m_M)$ for the number of entries in each of the M bins of the histogram. In a similar fashion to Eq. 5.3 the expectation value for each m_i can be written as

$$E[m_i] = u_i(\boldsymbol{\theta}), \tag{5.5}$$

where the u_i are calculable quantities that depend on the nuisance parameters.

The likelihood function of observing exactly the histograms $\mathbf{n}$ and $\mathbf{m}$ is the product of Poisson probabilities for all bins:

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}. \tag{5.6}$$

To test some value for μ, one can consider the likelihood ratio

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}, \tag{5.7}$$

where $\hat{\hat{\boldsymbol{\theta}}}$ in the numerator denotes the value of $\boldsymbol{\theta}$ that maximizes the likelihood for the μ being tested, namely it is the conditional maximum-likelihood (ML) estimator for $\boldsymbol{\theta}$ (and is thus a function of μ). The denominator is instead the maximized likelihood function, namely $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$ are their ML estimators. Notice once again the presence of nuisance parameters which smear and broaden the profile likelihood as a function of μ, relative to what one would have if they were fixed. This reflects the loss of information about μ due to the impact of systematic uncertainties. From the definition of $\lambda(\mu)$ one easily sees that $0 \leq \lambda(\mu) \leq 1$, with values near to 1 implying good agreement between the data and the tested value of μ.

5.1.1 Test statistic $t_\mu = -2 \ln \lambda(\mu)$

A convenient and indeed equivalent statistic to $\lambda(\mu)$ is obtained by computing

$$t_\mu = -2 \ln \lambda(\mu). \tag{5.8}$$

Given the properties of $\lambda(\mu)$, one easily sees that $0 \leq t_\mu \leq +\infty$, with greater values of t_μ corresponding to greater incompatibility between the data and the tested μ. It is now possible to define a test for μ by using the statistic t_μ itself as a measure of the discrepancy between the data and the hypothesis. To quantify the level of disagreement, a p-value is needed, which can be computed as

$$p_\mu = \int_{t_{\mu,\,\mathrm{obs}}}^{\infty} f\left(t_\mu | \mu\right) dt_\mu, \tag{5.9}$$

where $t_{\mu,\,\mathrm{obs}}$ is the value of the test statistic t_μ observed from the data, and $f\left(t_\mu | \mu\right)$ is the pdf of t_μ under the assumption of data being distributed according to a signal strength μ. In Section 5.2 we will point out that is perfectly fine and possible to compute the pdf of t_μ under the assumption of data being distributed with respect to a different value $\mu' \neq \mu$, $f\left(t_\mu | \mu'\right)$; however, in order to compute a p-value, one needs the pdf of the test statistic under the assumption of data being distributed with the very same value of μ that is being tested.

5.1.2 Test statistic $\tilde{t}_\mu$ for processes with $\mu \geq 0$

In many particle physics searches, one can safely assume that a signal process can only enhance the rate of measured events beyond the expectation coming from background alone, which means that the signal strength is bounded in the region $\mu \geq 0$ [1]. For such a signal model, if data is found such that the observed signal strength $\hat{\mu}$ is below zero, then the best level of agreement between the data and any allowed value of μ is obtained for the lowest possible μ, namely $\mu = 0$. Thus, a modified likelihood ratio $\tilde{\lambda}(\mu)$ can be defined:

$$\tilde{\lambda}(\mu) = \begin{cases} \frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu},\hat{\boldsymbol{\theta}})} & \hat{\mu} \geq 0, \\ \frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0,\hat{\hat{\boldsymbol{\theta}}}(0))} & \hat{\mu} < 0. \end{cases} \tag{5.10}$$

In a similar fashion with what has been done in Eq. 5.8, the variable $\tilde{\lambda}(\mu)$ can be used to obtain the corresponding test statistic, which we denote with $\tilde{t}_\mu$, that is

$$\tilde{t}_\mu = -2\ln\tilde{\lambda}(\mu) = \begin{cases} -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu},\hat{\boldsymbol{\theta}})} & \hat{\mu} \geq 0, \\ -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0,\hat{\hat{\boldsymbol{\theta}}}(0))} & \hat{\mu} < 0. \end{cases} \tag{5.11}$$

As it was done in Eq. 5.9 for the test statistic t_μ, it is possible to assess the level of disagreement between the data and the signal strength being tested by computing the p-value.

5.1.3 Test statistic q_0 for the discovery of a positive signal

An important special case of the test statistic $\tilde{t}_\mu$ is the one where the value $\mu = 0$ is tested, since rejecting the background hypothesis $\mu = 0$ leads to the discovery of a new signal process. For this special case, the notation $q_0 \equiv \tilde{t}_0$ is adopted. The definition given by Eq. 5.11, evaluated for $\mu = 0$, greatly simplifies and leads to

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \geq 0, \\ 0 & \hat{\mu} < 0, \end{cases} \tag{5.12}$$

where $\lambda(\mu)$ is the likelihood ratio as defined in Eq. 5.7. Notice that, if data fluctuate such that the observed signal strength is found below zero, the maximum agreement between the data and the background-only hypothesis is set, $q_0 = 0$. This means that, even though a negative value for the observed signal strength indicates some discrepancy between the data and

[1] Note, however, that there are important exceptions, such as searches for neutrino oscillations, where neutrinos of a given flavor can oscillate and "disappear".

the background-only hypothesis, this does not come from an excess in signal events, but most likely comes from some kind of systematic error. If, instead, the data yield increases above the expected background, increasingly large values for q_0 are obtained, corresponding to an increasing level of incompatibility between the data and the $\mu = 0$ hypothesis. As usual, to quantify this level of incompatibility, a p-value has to be computed in the same manner as done with t_μ, namely

$$p_0 = \int_{q_{0,\text{obs}}}^{\infty} f\left(q_0|0\right) dq_0, \tag{5.13}$$

where $f\left(q_0|0\right)$ is the pdf of the test statistic q_0 under the assumption of the background only hypothesis. A useful approximation of this pdf will be given in Section 5.2.

5.1.4 Test statistic q_μ for upper limits setting

In order to establish an upper limit on the signal strength parameter, it is possible to define the test statistics

$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \leq \mu, \\ 0 & \hat{\mu} > \mu, \end{cases} \tag{5.14}$$

where $\lambda(\mu)$ is the likelihood ratio as defined in Eq. 5.7. Notice that, if data fluctuate such that the observed signal strength is found above the tested signal strength, the maximum agreement between the data and the tested hypothesis is set, $q_\mu = 0$. This means that, in the context of upper limits setting, one does not regard values greater and greater than the tested value for μ as representing less and less compatibility with μ. It is also important to note that the test statistic q_0 defined previously is not a special case of q_μ as they have different and, in some sense, opposite definitions: q_0 is zero when data fluctuate downwards ($\hat{\mu} < 0$), while q_μ is zero if data fluctuate upwards ($\hat{\mu} > \mu$). With this caveat in mind, in the following sections we may often refer to q_μ, meaning either q_0 or q_μ depending on the context.

As usual, to assess the level of agreement between the data and the tested value of μ, a p-value has to be computed,

$$p_\mu = \int_{q_{\mu,\text{obs}}}^{\infty} f\left(q_\mu|\mu\right) dq_\mu, \tag{5.15}$$

where $f\left(q_\mu|\mu\right)$ is the pdf of q_μ assuming the hypothesis μ. A useful approximation of this pdf will be given in Section 5.2.

For the case in which a positive signal is considered ($\mu > 0$), the variable $\tilde{\lambda}(\mu)$ can be used in Eq. 5.14 to obtain the corresponding test statistic $\tilde{q}(\mu)$. However, the difference between tests based on q_μ and $\tilde{q}(\mu)$ are usually found to be negligible, while the use of q_μ leads to important algebraic simplifications and will thus be assumed in the following.

5.2 Approximate formulae for sampling distributions

In order to find the p-value of a hypothesis using, *e.g.*, Eqs. 5.13 or 5.15, the sampling distributions for the corresponding test statistics are needed. In the case of discovery, since the value being tested is $\mu = 0$, we need the sampling distribution for the test statistic q_0 under the hypothesis of data being distributed according to $\mu = 0$, $f(q_0|0)$. In the case of the setting of upper limits, when a nonzero value of μ is being tested, we need the sampling distribution of the test statistic q_μ under the assumption of data being distributed according to the very same value of μ, $f(q_\mu|\mu)$. For the sake of clarity, let us recall that, in this notation, the subscript of q refers to the value of the signal strength being tested, while the second argument in $f(q_\mu|\mu)$ refers to the value of the signal strength hypothesized in the distribution of the data.

In the context of particle physics experiments, it may also be useful to report the significance and upper limit that one would obtain for a variety of signal hypothesis, the so called expected (or, more precisely, median) significance and upper limit. In order to do so, we also need the distribution $f(q_\mu|\mu')$ with $\mu' \neq \mu$ which describes how the test statistic is distributed if the data correspond to a strength parameter different from the one being tested. Note that the sampling distributions needed in order to compute the p-values can always be obtained as a special case of $f(q_\mu|\mu')$ by setting μ' equal to the signal strength being tested.

5.2.1 Wald approximation for the distribution of the profile likelihood ratio

Let us consider a test of the strength parameter μ, which can either be zero in the case of discovery or nonzero in the case of upper limit setting. Also, suppose that data are distributed according to a strength parameter μ'. A. Wald showed [91] that, in the case of a single parameter of interest and in the large sample limit, the likelihood ratio is Gaussian distributed and can be written as

$$-2\ln\lambda(\mu) = \frac{(\mu-\hat{\mu})^2}{\sigma^2} + \mathcal{O}(1/\sqrt{N}), \tag{5.16}$$

where $\hat{\mu}$ follows a Gaussian distribution with mean μ' and standard deviation σ, and N represents the data sample size. The standard deviation σ of $\hat{\mu}$ can be obtained from the covariance matrix of the estimators of all parameters (including POI), $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$. In the large sample limit, the bias of ML estimators tends to zero and the Rao-Cramer-Frechet (RCF) bound holds, in such a way that the covariance matrix can be estimated by computing

$$V_{ij}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right], \tag{5.17}$$

where the expectation value assumes a strength parameter μ'.

5.2.2 Approximate distribution for q_0

Assuming the validity of the Wald approximation of Eq. 5.16 and starting from Eq. 5.12 one easily gets

$$q_0 = \begin{cases} \hat{\mu}/\sigma^2 & \hat{\mu} \geq 0, \\ 0 & \hat{\mu} < 0, \end{cases} \tag{5.18}$$

where $\hat{\mu}$ follows a Gaussian distribution with mean μ' and standard deviation σ. From this, it can be shown that the pdf of q_0 takes the form

$$f(q_0|\mu') = \left(1 - \Phi\left(\frac{\mu'}{\sigma}\right)\right)\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}\exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right]. \tag{5.19}$$

For the important case of $\mu' = 0$, which serves as a baseline to compute p-values, the equation above reduces to

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}e^{-q_0/2}. \tag{5.20}$$

From Equation 5.19 it can be shown that the corresponding cumulative function is

$$F(q_0|\mu') = \Phi\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right), \tag{5.21}$$

which in the special case $\mu' = 0$ simplifies to

$$F(q_0|0) = \Phi(\sqrt{q_0}). \tag{5.22}$$

Given that the p-value of the $\mu = 0$ hypothesis is

$$p_0 = 1 - F(q_0|0), \tag{5.23}$$

therefore the significance, using Eq. 5.1, simply becomes

$$Z_0 = \Phi^{-1}(1 - p_0) = \sqrt{q_0}. \tag{5.24}$$

5.2.3 Approximate distribution for q_μ

Assuming the validity of the Wald approximation of Eq. 5.16 and starting from Eq. 5.14 one easily gets

$$q_\mu = \begin{cases} \frac{(\mu-\hat{\mu})^2}{\sigma^2} & \hat{\mu} < \mu, \\ 0 & \hat{\mu} > \mu, \end{cases} \tag{5.25}$$

where again $\hat{\mu}$ follows a Gaussian distribution with mean μ' and standard deviation σ. The pdf $f(q_\mu|\mu')$ is found to be

$$f(q_\mu|\mu') = \Phi\left(\frac{\mu'-\mu}{\sigma}\right)\delta(q_\mu) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_\mu}}\exp\left[-\frac{1}{2}\left(\sqrt{q_\mu} - \frac{\mu-\mu'}{\sigma}\right)^2\right], \tag{5.26}$$

so that for the special case $\mu = \mu'$ we get

$$f(q_\mu|\mu) = \frac{1}{2}\delta(q_\mu) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}e^{-q_\mu/2}. \tag{5.27}$$

The cumulative distribution is

$$F(q_\mu|\mu') = \Phi\left(\sqrt{q_\mu} - \frac{\mu-\mu'}{\sigma}\right), \tag{5.28}$$

thus leading to a special case $\mu' = \mu$ that has the same form as what was found for q_0, namely,

$$F(q_\mu|\mu) = \Phi(\sqrt{q_\mu}). \tag{5.29}$$

The p-value of the μ being tested is given by

$$p_\mu = 1 - F(q_\mu|\mu) = 1 - \Phi(\sqrt{q_\mu}), \tag{5.30}$$

and the corresponding significance is

$$Z_\mu = \Phi^{-1}(1-p_\mu) = \sqrt{q_\mu}. \tag{5.31}$$

If p_μ is found below a give threshold α, which is usually taken to be $\alpha = 0.05$, then the tested value of μ is said to be excluded at a confidence level (CL) of $1-\alpha$. The observed upper limit on the signal strength parameter is the smallest tested μ with $p_\mu < \alpha$. An explicit formula for the upper limit can be obtained by simply setting $p_\mu = \alpha$ and solving for μ:

$$
\begin{aligned}
&p_\mu = \alpha \\
&1 - \Phi(\sqrt{q_\mu}) = \alpha \\
&1 - \Phi\left(\frac{\mu - \hat{\mu}}{\sigma}\right) = \alpha \\
&\Phi^{-1}(1-\alpha) = \frac{\mu - \hat{\mu}}{\sigma}
\end{aligned}
$$

thus leading to

$$\mu_{\text{up}} = \hat{\mu} + \sigma\Phi^{-1}(1-\alpha). \tag{5.32}$$

As an important point, it should be noted that some of the formulae given above require the standard deviation σ of $\hat{\mu}$. We already gave a method to estimate the variance using the RCF bound (see Eq. 5.17); however, in many situations, it turns out to be impractical to compute the RCF bound analytically, since it requires the expectation value of the second derivatives of the log-likelihood function (*i.e.*, an integration over the variable x). In the following, we will show a way to overcome this bottleneck which relies on the use of a special, artificial data set that we call the "Asimov data set".

5.2.4 The Asimov data set and the variance of $\hat{\mu}$

First, let us give a definition of the Asimov data set and show how it is constructed in some interesting cases.

Asimov data set. *An Asimov data set is a data set constructed in such a way that, when one uses it to evaluate all the ML estimators of all the parameters, one gets their true values.*

As a first important example of Asimov data set, let us suppose we are performing a simple counting experiment: we measure some number of events n with $E[n] = \mu s + b$, where the expected signal and background yields s and b are assumed to be known with negligible uncertainty. The observed number of events n is taken to be a Poisson variable and so the likelihood can be written as

$$L(\mu) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)}.$$

In this simple case, the only parameter entering the likelihood is the signal strength μ and the Asimov data set is the expected signal+background yield $\mu s + b$. Let us demonstrate this statement by computing the ML estimator for μ. The ML estimator for μ is found by setting the partial derivative $\partial L/\partial\mu = 0$. Straightforward calculations lead to

$$\hat{\mu} = \frac{n - b}{s}. \tag{5.33}$$

If we impose the Asimov condition $\hat{\mu} = \mu$ and solve for n we obtain:

$$n_A = \mu s + b. \tag{5.34}$$

As a second example, let us move to a slightly more complex case, namely a shape analysis which is described by the likelihood reported in Eq. 5.6. Let $\theta_0 \equiv \mu$, in such a way that the full set of parameters (the POI μ and the nuisance parameters) can be written as $\boldsymbol{\theta}$. In order to maximize this likelihood let us compute the partial derivatives of the log-likelihood with respect to the parameters and set them equal to zero:

$$\begin{aligned}
\ln L &= \ln\left(\prod_{i=1}^{N} \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-(\mu s_i + b_i)} \prod_{j=1}^{M} \frac{u_j^{m_j}}{m_j!} e^{-u_j}\right) = \\
&= \sum_{i=1}^{N} \ln\left(\frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-(\mu s_i + b_i)}\right) + \sum_{j=1}^{M} \ln\left(\frac{u_j^{m_j}}{m_j!} e^{-u_j}\right) = \\
&= \sum_{i=1}^{N} \Big[n_i \ln \underbrace{(\mu s_i + b_i)}_{\nu_i} - \ln n_i! - \underbrace{(\mu s_i + b_i)}_{\nu_i} \Big] + \\
&+ \sum_{j=1}^{M} \Big[m_j \ln u_j - \ln m_j! - u_j \Big].
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ln L}{\partial \theta_k} &= \sum_{i=1}^{N} \Big[n_i \frac{\partial}{\partial \theta_k} \ln \nu_i - 0 - \frac{\partial \nu_i}{\partial \theta_k} \Big] + \sum_{j=1}^{M} \Big[m_j \frac{\partial}{\partial \theta_k} \ln u_j - 0 - \frac{\partial u_j}{\partial \theta_k} \Big] = \\
&= \sum_{i=1}^{N} \Big(\frac{n_i}{\nu_i} - 1 \Big) \frac{\partial \nu_i}{\partial \theta_k} + \sum_{j=1}^{M} \Big(\frac{m_j}{u_j} - 1 \Big) \frac{\partial u_j}{\partial \theta_k} = 0.
\end{aligned} \tag{5.35}$$

Notice that now we cannot proceed as we did previously, because we are now dealing with more than one parameter. It is not correct to perform the partial derivative of the likelihood with respect to θ_0 and find the $\boldsymbol{n}_0$ and $\boldsymbol{m}_0$ that give $\hat{\theta}_0 = E[\theta_0]$, then perform the partial derivative of the likelihood with respect to θ_1 and find the $\boldsymbol{n}_1$ and $\boldsymbol{m}_1$ that give $\hat{\theta}_1 = E[\theta_1]$ etc., because the Asimov data set is a single data set $\boldsymbol{n}_A$, $\boldsymbol{m}_A$ that, when used to evaluate all the ML estimators, gives the true values for the parameters. We can nevertheless note that condition (5.35) holds for every k (namely, for all the parameters) if $n_i = \nu_i$ and $m_j = u_j$. This means that the Asimov data set in the case of a shape analysis is the set of N values $n_{i,A}$ and M values $m_{i,A}$ fulfilling:

$$\begin{aligned}
n_{i,A} &= E[n_i] = \nu_i = \mu s_i + b_i, \\
m_{j,A} &= E[m_j] = u_j,
\end{aligned} \tag{5.36}$$

namely an N-bins histogram made of the expected signal+background yields for each of the observable bins and a M-bins histogram made of the expected background yields for each of the subsidiary measurement bins.

As previously stated, we can use the Asimov data set to find an estimation for the variance of $\hat{\mu}$, σ. As a first step, let us exploit the Asimov data set to compute an "Asimov likelihood" L_A and the corresponding profile likelihood λ_A:

$$\lambda_A(\mu) = \frac{L_A(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L_A(\hat{\mu}, \hat{\boldsymbol{\theta}})} = \frac{L_A(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L_A(\mu', \boldsymbol{\theta})}, \tag{5.37}$$

where the last equality follows by the very same definition of Asimov data set. As we already mentioned, a standard way to find σ is to use the prescription of the RCF bound, namely to estimate the matrix of second derivatives of the log-likelihood to obtain the inverse of the covariance matrix, then to invert it and finally to extract the element V_{00} corresponding to the variance σ of $\hat{\mu}$. The second derivative of the log-likelihood is

$$\begin{aligned} \frac{\partial^2 \ln L}{\partial\theta_k \partial\theta_\ell} = &\sum_{i=1}^{N} \left[\left(\frac{n_i}{\nu_i} - 1 \right) \frac{\partial^2 \nu_i}{\partial\theta_k \partial\theta_\ell} - \frac{\partial\nu_i}{\partial\theta_k} \frac{\partial\nu_i}{\partial\theta_\ell} \frac{n_i}{\nu_i^2} \right] + \\ &+ \sum_{j=1}^{M} \left[\left(\frac{m_j}{u_j} - 1 \right) \frac{\partial^2 u_j}{\partial\theta_k \partial\theta_\ell} - \frac{\partial u_j}{\partial\theta_k} \frac{\partial u_j}{\partial\theta_\ell} \frac{m_j}{u_j^2} \right]. \end{aligned} \tag{5.38}$$

The equation above shows that the second derivative of the log-likelihood depends linearly on the data values n_i and m_j, which means that its expectation value can be found by simply evaluating it with the expectation values of the data, which, given Eq. 5.36, are simply the Asimov data. Therefore, the inverse of the covariance matrix can be obtained as

$$V_{k\ell}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_k \partial\theta_\ell} \right] = -\frac{\partial^2 \ln L_A}{\partial\theta_k \partial\theta_\ell} = \sum_{i=1}^{N} \frac{\partial\nu_i}{\partial\theta_k} \frac{\partial\nu_i}{\partial\theta_\ell} \frac{1}{\nu_i} + \sum_{j=1}^{M} \frac{\partial u_j}{\partial\theta_k} \frac{\partial u_j}{\partial\theta_\ell} \frac{1}{u_j}. \tag{5.39}$$

In practical cases, one can evaluate the derivatives of $\ln L_A$ numerically, use them to find the inverse of the covariance matrix, invert it and extract the variance of $\hat{\mu}$.

5.3 Median discovery and exclusion significances

In order to state the sensitivity of a given experiment, one is not only interested in the significance obtained from the observed data set, but also in the expected (more precisely, median) significance with which one would be able to reject different values of the tested μ. Specifically, in the

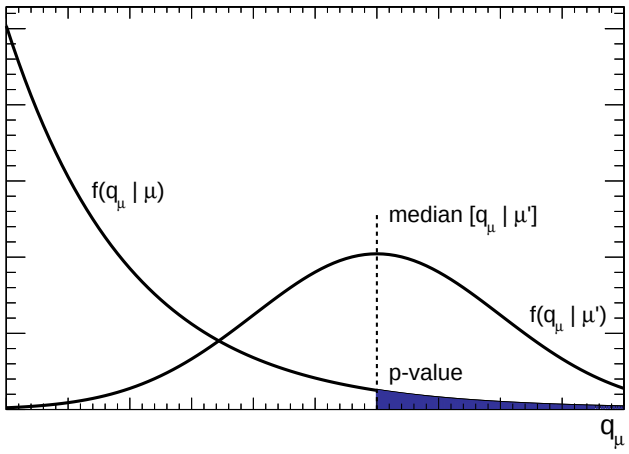

Figure 5.1: Pictorial representation of the p-value corresponding to the median q_μ assuming an alternative hypothesis μ'.

case of discovery one may be interested in the median significance, under the assumption of the nominal signal hypothesis ($\mu' = 1$) with which the background-only hypothesis ($\mu = 0$) can be rejected. On the other hand, in the case of exclusion, one may be interested in the median significance, assuming the background-only hypothesis ($\mu' = 0$), with which a nonzero value of μ (usually $\mu = 1$) can be rejected. The median significance of an experiment is illustrated in Fi. 5.1, where the pdfs for q_μ are shown for both an hypothesized value of μ and μ'. Note that the distribution $f(q_\mu|\mu')$ is shifted towards higher values of q_μ, as it should be, since some level of disagreement between the data and the tested value of the signal strength is expected when data are distributed with respect to a different signal strength value. The sensitivity of an experiment can be characterized by the p-value corresponding to the median q_μ assuming the alternative hypothesis μ', namely the shaded region in Fig. 5.1.

In the following we will describe how to obtain simple expressions for the experimental sensitivity (*i.e.*, the median discovery or exclusion significance) by making use of the previously defined Asimov data sample.

5.3.1 Expected significance for a discovery

As a starting point in order to obtain the expected significance in the case of a discovery, let us note that the observed significance reported in Eq. 5.24 is a monotonic function of the test statistic q_0. Thus, the expected significance is obtained by simply evaluating the function for the median value of q_0 for a hypothesized value μ'. Since the Wald approximation is assumed, q_0 follows approximately a Gaussian distribution, for which the median can be

approximated with the expectation value, which in turn can be approximated by its Asimov value, *i.e.*:

$$\text{med}[Z_0|\mu'] = \sqrt{q_{0,A}}. \tag{5.40}$$

The Asimov data set must obviously be chosen in such a way that the evaluation of $\hat{\mu}$ using this data set gives μ' as a result. Given the form of q_0 in the Wald approximation that was found in Eq. 5.18, we can readily obtain the following expression for the expected significance in the case of discovery:

$$\text{med}[Z_0|\mu'] = \sqrt{q_{0,A}} = \frac{\mu'}{\sigma}. \tag{5.41}$$

Notice that the $\hat{\mu} < 0$ part of Eq. 5.18 is not taken into account since using the Asimov data set we obtain $\hat{\mu} = \mu'$ and the hypothesized μ' value is greater than zero.

By exploiting the important properties of the Asimov data set, we were able to find the median significance, assuming some strength parameter μ', for rejecting the background-only hypothesis. However, it is also useful to know by how much the expected significance would vary, given the expected fluctuations in data. To achieve this, it is convenient to compute error bars for the median significance corresponding to $\pm N\sigma$ variations of $\hat{\mu}$. Let us go through the two cases separately:

- *Positive variation* $+N\sigma$. Combining the results of Eqs. 5.18 and 5.24, we can write the following expression for the observed significance:

$$Z_0(\hat{\mu} + N\sigma) = \begin{cases} \dfrac{\hat{\mu} + N\sigma}{\sigma} & \hat{\mu} + N\sigma \geq 0, \\ \\ 0 & \hat{\mu} + N\sigma < 0. \end{cases}$$

  If we evaluate this using the Asimov data set to find the median significance we readily obtain:

$$Z_0(\mu' + N\sigma) = \begin{cases} \dfrac{\mu'}{\sigma} + N & \mu' + N\sigma \geq 0, \\ \\ 0 & \mu' + N\sigma < 0, \end{cases}$$

  where the only possible case is actually the first one, since $\mu' \geq 0$. Thus, the median significance for a $+N\sigma$ variation is found to be

$$Z_0(\mu' + N\sigma) = \frac{\mu'}{\sigma} + N = \text{med}[Z_0|\mu'] + N. \tag{5.42}$$

- *Negative variation* $-N\sigma$. In this case, the significance takes the form

$$Z_0(\hat{\mu} - N\sigma) = \begin{cases} \dfrac{\hat{\mu} - N\sigma}{\sigma} & \hat{\mu} - N\sigma \geq 0, \\ \\ 0 & \hat{\mu} - N\sigma < 0. \end{cases}$$

  Following the same reasoning as before and using the Asimov data set, we get:

$$Z_0(\mu' - N\sigma) = \begin{cases} \text{med}[Z_0|\mu'] - N & \mu' - N\sigma \geq 0, \\ \\ 0 & \mu' - N\sigma < 0. \end{cases}$$

  Notice that now $\mu' - N\sigma$ has not a definite sign, so both of the cases are possible. Moreover, $\text{med}[Z_0|\mu'] - N$ may be negative and a negative significance does not have a physical meaning. For this reason we take the higher of the possible values for the expected significance as the median significance corresponding to a $-N\sigma$ variation:

$$Z_0(\mu' - N\sigma) = \max\big[\text{med}[Z_0|\mu'] - N,\ 0\big]. \tag{5.43}$$

We can summarize the results obtained for the median discovery significance and its error bands as follows:

$$\begin{cases} \text{med}[Z_0|\mu'] = \sqrt{q_{0,\text{A}}} \\ \\ Z_0(\mu' + N\sigma) = \text{med}[Z_0|\mu'] + N \\ \\ Z_0(\mu' - N\sigma) = \max\big[\text{med}[Z_0|\mu'] - N,\ 0\big] \end{cases} \tag{5.44}$$

5.3.2 Expected significance for exclusion

As a starting point in order to obtain the expected significance in the case of exclusion, let us note that the observed significance reported in Eq. 5.31 is a monotonic function of the test statistic q_μ. Thus, we can repeat the very same reasoning of the previous subsection and find the expected significance by evaluating Eq. 5.31 with the Asimov data set, obtaining:

$$\text{med}[Z_\mu|\mu'] = \sqrt{q_{\mu,A}} = \frac{\mu - \mu'}{\sigma}. \tag{5.45}$$

where in the last equality we used the asymptotic expression for q_μ, Eq. 5.25 and the Asimov data set must be chosen in such a way that the evaluation of $\hat{\mu}$ using this data set gives μ' as a result. In order to find the expected

upper limit, let us start from the approximate form of the observed upper limit given by Eq. 5.32 and find the median value by evaluating it with an Asimov data set corresponding to μ':

$$\text{med}[\mu_{\text{up}}|\mu'] = \mu' + \sigma\Phi^{-1}(1-\alpha). \tag{5.46}$$

In a similar fashion, one can compute the observed upper limit for $\hat{\mu} \pm N\sigma$ and evaluate it using the Asimov data set, obtaining the error bands for the median upper limit:

$$\text{band}_{N\sigma} = \mu' + \sigma[\Phi^{-1}(1-\alpha) \pm N]. \tag{5.47}$$

Note, as highlighted at the beginning of Section 5.3, that the usual procedure to state expected exclusion significances and expected upper limits uses $\mu' = 0$. This means that one usually computes the limit that one would obtain if the most probable (more precisely: median) value of the test statistic q_μ in the background only hypothesis was measured.

5.3.3 Combination of multiple channels

In most of the cases, data analyses split the phase space of the search in many channels, often referred to as categories, in order to enhance the sensitivity of the measurement. For each channel i there is a likelihood function $L_i(\mu, \boldsymbol{\theta}_i)$, where $\boldsymbol{\theta}_i$ stands for the set of nuisance parameters for the ith channel, some of which may be shared between channels. The signal strength parameter is assumed to be the same for all channels. If the channels are statistically independent (which can be easily arranged), then the full likelihood can be written as the product of the likelihoods over all the channels,

$$L(\mu, \boldsymbol{\theta}) = \prod_i L_i\left(\mu, \boldsymbol{\theta}_i\right), \tag{5.48}$$

where $\boldsymbol{\theta}$ stands for the full set of nuisance parameters. Thus, the profile likelihood ratio $\lambda(\mu)$ becomes:

$$\lambda(\mu) = \frac{\prod_i L_i(\mu, \hat{\hat{\boldsymbol{\theta}}}_i)}{\prod_i L_i(\hat{\mu}, \hat{\boldsymbol{\theta}}_i)}. \tag{5.49}$$

Because the Asimov data set has no statistical fluctuations, one has $\hat{\mu} = \mu'$ for each channel in such a way that, when using an Asimov data set corresponding to a signal strength parameter μ', one finds that the global Asimov likelihood ratio factorizes in the product of the Asimov likelihood ratios over all channel, that is:

$$\lambda_A(\mu) = \frac{\prod_i L_i(\mu, \hat{\hat{\boldsymbol{\theta}}}_i)}{\prod_i L_i\left(\mu', \boldsymbol{\theta}\right)} = \prod_i \lambda_{A,i}(\mu). \tag{5.50}$$

This greatly simplify the task of computing Eqs. 5.40 and 5.45 in the case of multiple channels, since the global likelihood can be obtained as the product of separate likelihoods for each channel. Nevertheless, in order to find observed significances or upper limits, one needs to construct the full likelihood function, which cannot be factorized, and perform a ML fit to construct the likelihood ratio.

5.4 Goodness of the asymptotic formulae

In this final section, let us show with a simple example the goodness of the asymptotic formulae given above. We consider a shape analysis in which one is looking for a peak in some invariant mass distribution. The background is assumed to follow a Rayleigh distribution, while the signal is modeled with a Gaussian with known mean and width. To simplify things, no subsidiary measurement is performed. We assume the likelihood function to have the form

$$L(\mu,\theta) = \prod_{i=1}^{N} \frac{(\mu s_i + \theta f_{\mathrm{b},i})^{n_i}}{n_i!} e^{-(\mu s_i + \theta f_{\mathrm{b},i})}, \tag{5.51}$$

where the mean signal yield in each bin, s_i, and the probability to find a background event in bin i, $f_{\mathrm{b},i}$, are assumed to be known, while the total expected number of background events, θ, is a nuisance parameter. For a given data set $\mathbf{n} = (n_1, n_2, \ldots, n_N)$ we can now evaluate the likelihood and from this determine any of the previously discussed test statistics. If we focus on the test statistic q_μ used to set an upper limit on μ and compute the distribution $f(q_\mu|\mu')$ a first time by generating Monte Carlo toys and a second time using the asymptotic formula in Eq. 5.26, we obtain the results that are shown in Fig. 5.2, where the distributions $f(q_\mu|0)$ (red) and $f(q_\mu|\mu)$ (blue) are plotted for both the MC toys, as histograms, and for the approximated formulae, as functions. We see pretty clearly that the asymptotic formulae agree well with the MC predictions.

If one has only the MC histograms, the expected upper limit at $1-\alpha$ on the signal strength parameter can be found by testing different values of μ and finding the one for which the p-value, assuming that the median q_μ in the background-only hypothesis is measured, is equal to α (see Fig. 5.2). As a final remark, we will show that having the asymptotic formulae in hand, this procedure reduces to the formula that was given in Eq. 5.46 evaluated for $\mu' = 0$. Indeed, following the given procedure we must find a μ_{up} such that $p_{\mu_{\mathrm{up}}} = \alpha$. This means

$$p_{\mu_{\mathrm{up}}} = \int_{\mathrm{med}[q_{\mu_{\mathrm{up}}}|0]}^{+\infty} f(q_{\mu_{\mathrm{up}}}|\mu_{\mathrm{up}})\ \mathrm{d}q_{\mu_{\mathrm{up}}} = \alpha$$

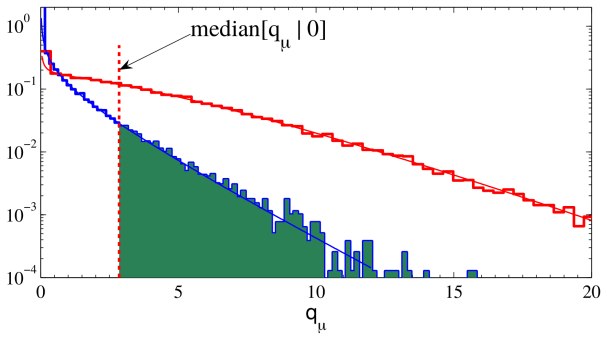

Figure 5.2: The distributions $f(q_\mu|0)$ (red) and $f(q_\mu|\mu)$ (blue) for both MC toys and asymptotic formulae. The shaded, green area under the curve $f(q_\mu|\mu)$ is the median p-value under the assumption of the background-only hypothesis. The value of the tested μ for which $p_\mu = \alpha$ is said to be the expected upper limit at $1-\alpha$ CL.

which can also be written as

$$p_{\mu_{\text{up}}} = 1 - F(q_{\mu_{\text{up}}}|\mu_{\text{up}}) = \alpha. \tag{5.52}$$

Equation 5.52 can be written, making use of the expression in Eq. 5.29 for $F(q_\mu|\mu)$ given by the Wald approximation, as

$$\begin{aligned} p_{\mu_{\text{up}}} &= 1 - \Phi\left(\sqrt{q_{\mu_{\text{up}}}}\right) = \\ &= 1 - \Phi\left(\sqrt{\text{med}[q_{\mu_{\text{up}}}|0]}\right) = \alpha, \end{aligned} \tag{5.53}$$

which means

$$\sqrt{\text{med}[q_{\mu_{\text{up}}}|0]} = \Phi^{-1}(1-\alpha). \tag{5.54}$$

Recall that median$[q_{\mu_{\text{up}}}|0]$ is the specific $\tilde{q}_{\mu_{\text{up}}}$ value that gives $F(\tilde{q}_{\mu_{\text{up}}}|0) = \frac{1}{2}$. Using Eq. 5.28 for $F(q_\mu|\mu')$ in the case $\mu' = 0$ we get

$$\begin{aligned} &\Phi\left(\sqrt{\tilde{q}_{\mu_{\text{up}}}} - \frac{\mu_{\text{up}}}{\sigma}\right) = \frac{1}{2} \\ &\sqrt{\tilde{q}_{\mu_{\text{up}}}} - \frac{\mu_{\text{up}}}{\sigma} = \Phi^{-1}\left(\frac{1}{2}\right) \\ &\sqrt{\tilde{q}_{\mu_{\text{up}}}} - \frac{\mu_{\text{up}}}{\sigma} = 0 \\ &\sqrt{\tilde{q}_{\mu_{\text{up}}}} \equiv \sqrt{\text{med}[q_{\mu_{\text{up}}}|0]} = \frac{\mu_{\text{up}}}{\sigma}. \end{aligned} \tag{5.55}$$

So, in the end, putting Eq. 5.55 into Eq. 5.54, we get

$$\begin{aligned} \frac{\mu_{\text{up}}}{\sigma} &= \Phi^{-1}(1-\alpha) \\ \mu_{\text{up}} &= \sigma\Phi^{-1}(1-\alpha), \end{aligned} \tag{5.56}$$

which corresponds to Eq. 5.46 when $\mu' = 0$.

Chapter 6

Results

6.1 Upper limit in the CL_s prescription

In order to compute the upper limit on the signal strength parameter $\mu_{t\bar{t}H}$, the test statistic introduced in Eq. 5.14 is used. As we discussed in Subsection 5.2.3, a possible way to compute the upper limit at $(1-\alpha)$ CL is to find the smallest value of the tested μ that gives $p_\mu < \alpha$, where p_μ is given by Eq. 5.30. However, a more refined procedure, which is discussed in full detail in [92], can be implemented by defining, together with p_μ, the p-value $1-p_b$ associated with the observation for the background-only hypothesis, *i.e.*,

$$1 - p_b = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f\left(q_\mu|0\right) dq_\mu. \tag{6.1}$$

Both the p-values are shown schematically in Fig. 6.1. Then, we can introduce the quantity $\mathrm{CL}_s(\mu)$ by taking the ratio of these two probabilities,

$$\mathrm{CL}_s(\mu) = \frac{p_\mu}{1-p_b}, \tag{6.2}$$

and find the smallest value of the tested μ for which $\mathrm{CL}_s < \alpha$.

This way, one can get small values of $\mathrm{CL}_s(\mu)$ (and thus, eventually, exclude the tested μ) for either small values of p_μ, *i.e.*, in the case of great incompatibility between the signal+background hypothesis and the observed data, or in the case of large values for $1-p_b$, *i.e.*, in the case of great compatibility between the background-only hypothesis and the observed data. Also, as $1-p_b$ is always less than or equal to unity, the exclusion criterion based on $\mathrm{CL}_s(\mu)$ is more conservative than the usual criterion $p_\mu < \alpha$, and the actual coverage probability of the corresponding intervals will exceed the nominal confidence level $1-\alpha$. In this work, we report upper limits obtained with the CL_s prescription.

Using the asymptotic formulae given by Eqs. 5.28, 5.30 and 5.45, it is possible to obtain a rather simple form for CL_s, that is

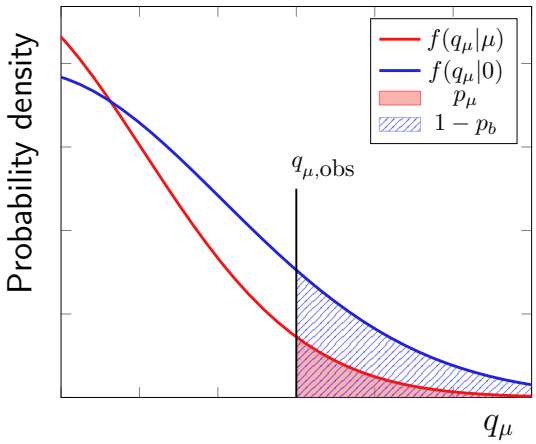

Figure 6.1: p-values p_μ (red area) and $1-p_b$ (blue, dashed area) used in the CL_s prescription for the computation of upper limits.

$$\text{CL}_s = \frac{1-\Phi\left(\sqrt{q_\mu}\right)}{\Phi\left(\sqrt{q_{\mu,A}}-\sqrt{q_\mu}\right)}, \tag{6.3}$$

which can be used to impose $\text{CL}_s = \alpha$ to find the observed and median upper limits. Using Eq. 5.25 and doing similar calculations to the ones performed in Subsection 5.2.3 one finds that the formula for the observed upper limit reads

$$\mu_{\text{up}} = \hat{\mu} + \sigma\Phi^{-1}\left[1-\alpha\Phi\left(\frac{\hat{\mu}}{\sigma}\right)\right]. \tag{6.4}$$

Also, with identical steps to the ones developed in Subsection 5.3.2, one finds that the formulae for the median upper limit and the related error bands are

$$\text{med}[\mu_{\text{up}}|0] = \sigma\Phi^{-1}\left(1-\frac{1}{2}\alpha\right) \tag{6.5}$$

and

$$\text{band}_{N\sigma} = \sigma\left\{\Phi^{-1}\left[1-\alpha\Phi\left(\pm N\right)\right]\pm N\right\} \tag{6.6}$$

respectively.

6.2 Blinded results

The CMS Collaboration performed, and still performs, many searches of new processes by looking for signals amidst the background of already-discovered

processes. If the indication for a new signal emerges, more data should be collected in order to make the measurement statistically more significant and to confirm or reject the existence of the sought-after signal.

However, a well-known tendency of the human beings, present at conscious or unconscious level, is to influence and optimize analyses based on what has been seen previously. Historically, several ways have been developed to prevent this from happening. In this work, the method of "keeping the signal box closed" is adopted, *i.e.*, the analysis is developed and optimized without looking at the data distributions in the signal region. More precisely, we do not have access to the distribution of the invariant mass of the resolved Higgs candidate for the events entering the signal categories.

The metric used to optimize the performances of the blinded analysis is the median (expected) upper limit at 95% CL on the signal strength parameter $\mu_{\mathrm{t\bar{t}H}}$, computed in the CL_s prescription. Even though the observed signal strength cannot be extracted at the level of the blinded analysis, since no data in the signal region is available to perform the ML fit, it is nevertheless convenient and recommended, before reporting the final value of the expected upper limit, to perform sanity checks for the fitting procedure by making use of an Asimov dataset. We perform such checks using an Asimov dataset corresponding to the signal+background (S+B) hypothesis. The CMS Higgs Physics Analysis Group suggests some useful quantities to be monitored in order to assess the consistency of the procedure. These quantities are:

- The fitted value of the signal strength. By the very definition of an Asimov dataset corresponding to the S+B hypothesis, a consistent fit should return the fitted value $\mu_{\mathrm{t\bar{t}H},\,A} = 1$;

- The pulls on the nuisance parameters. The pull on a given NP is defined as the difference between the post-fit and pre-fit values θ and θ_0 of the NP, normalized to the prefit uncertainty $\Delta\theta_0$, *i.e.*,

$$\frac{\theta - \theta_0}{\Delta\theta_0}. \tag{6.7}$$

  Since we are fitting an Asimov dataset, the expected behavior for all the parameters is not to be pulled, *i.e.*, the pulls are all expected to be zero. Also, a related useful information is the ratio between the post-fit and pre-fit uncertainties on the NP, with values smaller than 1 implying that the NP is constrained by the fitting procedure;

- The impacts of the nuisance parameters on the signal strength $\mu_{\mathrm{t\bar{t}H}}$. The impact of a NP θ on the signal strength is defined as the shift $\Delta\mu_{\mathrm{t\bar{t}H}}$ that is induced by a shift of $\pm\Delta\theta$ on the NP, where $\Delta\theta$ is the post-fit uncertainty, while the remaining NPs are profiled as usual. The impact reflects the level of correlation between the NP and the signal

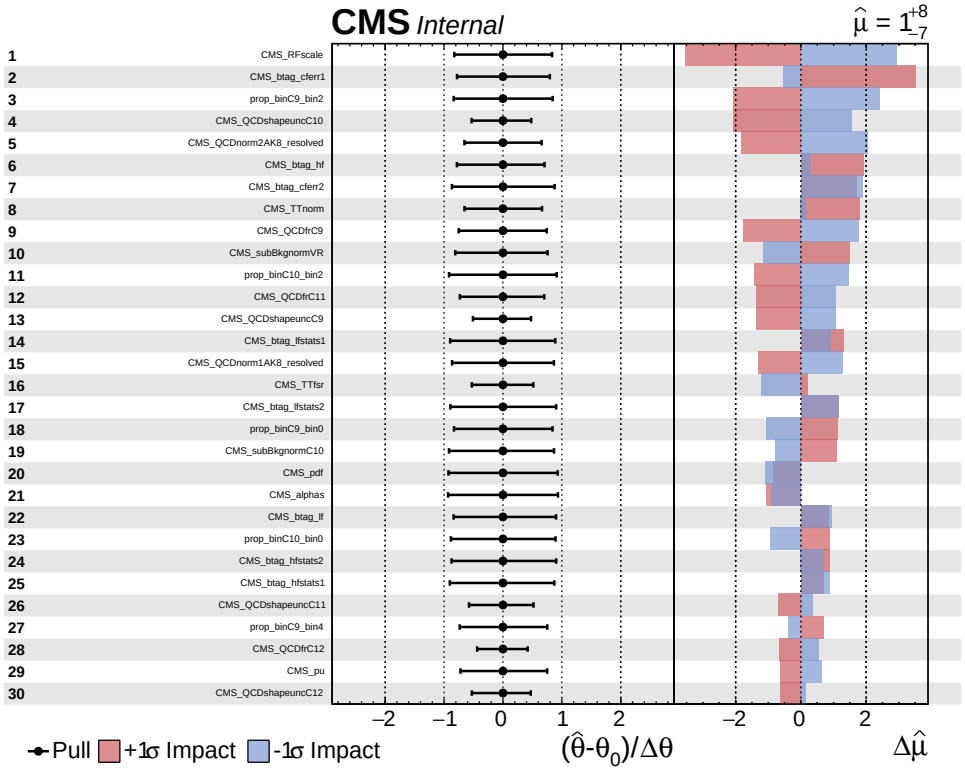

Figure 6.2: Pulls and impacts for the 30 nuisance parameters that affect the most the signal strength $\mu_{t\bar{t}H}$ in the RHC, as the result of the fit to the S+B Asimov dataset.

strength and is useful to understand which NPs have the largest effect on the signal strength uncertainty.

All such information is summarized in the pulls and impacts plot which is shown in Fig. 6.2. In the upper right corner, the fitted value of $\mu_{t\bar{t}H}$ is reported. The left column shows the names of the 30 nuisance parameters impacting the most on the signal strength parameter $\mu_{t\bar{t}H}$, sorted in decreasing order of the impact. The central column shows the pull of each NP, along with the error bars corresponding to the ratio of post-fit and pre-fit uncertainties. Finally, the third column shows the impact of each NP on the signal strength, with red bands standing for a positive shift of the NP and blue bands standing for a negative shift of the NP. Based on the orientation of such bars, one can deduce if a give NP is correlated or anticorrelated with $\mu_{t\bar{t}H}$.

As a first remark, we shall note that the fitted value for the signal strength is 1, as expected from a fit to a S+B Asimov dataset.

Secondly, among the nuisance parameters having the biggest impact on the signal strength parameter, we find the renormalization and factorization

scales, the b tagging uncertainty related to c-flavored jets, the MC statistics of the second bin of category 9, the shape uncertainty for the QCD background in category 10 and the uncertainty on the QCD normalization of categories with two AK8 jets. As expected, the prediction of the dominant background plays a major role in determining the uncertainty on $\mu_{\mathrm{t\bar{t}H}}$, even though by estimating its contribution with a data-driven method we are able to obtain lower impacts with respect to different scenarios in which, for example, the normalization was obtained from the MC prediction and the related uncertainty was freely floating.

Finally, we see that all the pulls are equal to zero, as expected, and that some nuisance parameters are somewhat constrained. No indication of artificial overconstrain is found in any of the NP. In particular, we note that the NP related to the $\mathrm{t\bar{t}}$ normalization is found to be constrained, which is the desired behavior achieved by including in the fit the $\mathrm{t\bar{t}}$ VR. The pulls and impacts for the full set of nuisance parameters entering the analysis can be found in Appendix B.1.

Once the fitting procedure has been validated, we proceed in computing the expected upper limit at 95% CL on the signal strength parameter. By performing a fit to the four signal categories and to the $\mathrm{t\bar{t}}$ VR simultaneously, the expected upper limit at 95% CL on $\mu_{\mathrm{t\bar{t}H}}$ is found to be 15.9 times the expectation from the SM, *i.e.*,

$$\mathrm{med}[\mu_{\mathrm{t\bar{t}H,\,up}}|0] = 15.9. \tag{6.8}$$

6.3 Unblinded results

Once the diagnostic of the fitting procedure has been carried out, we proceed in the extraction of the signal strength value. A comparison between the pre-fit template shapes and the observed data is reported in Fig. 6.3 for the signal categories and for the $\mathrm{t\bar{t}}$-enriched VR.

By performing a simultaneous ML fit to the data sample in the signal categories and in the $\mathrm{t\bar{t}}$ VR, the observed signal strength is found to be

$$\hat{\mu}_{\mathrm{t\bar{t}H}} = -7^{+8}_{-7}, \tag{6.9}$$

a value which is compatible with the SM expectation.

Since the fitted value is also compatible with the background-only hypothesis, we additionally set an exclusion limit at 95% CL using the asymptotic approximation of the CL_s prescription. Combining all the signal categories and the $\mathrm{t\bar{t}}$ VR, the observed and expected upper limits on $\mu_{\mathrm{t\bar{t}H}}$ are found to be $\mu_{\mathrm{t\bar{t}H}} < 11.5$ and < 15.8, respectively. The observed and expected upper limits in each signal category, together with the observed and expected upper limit which result from the combination, are shown in Fig. 6.4.

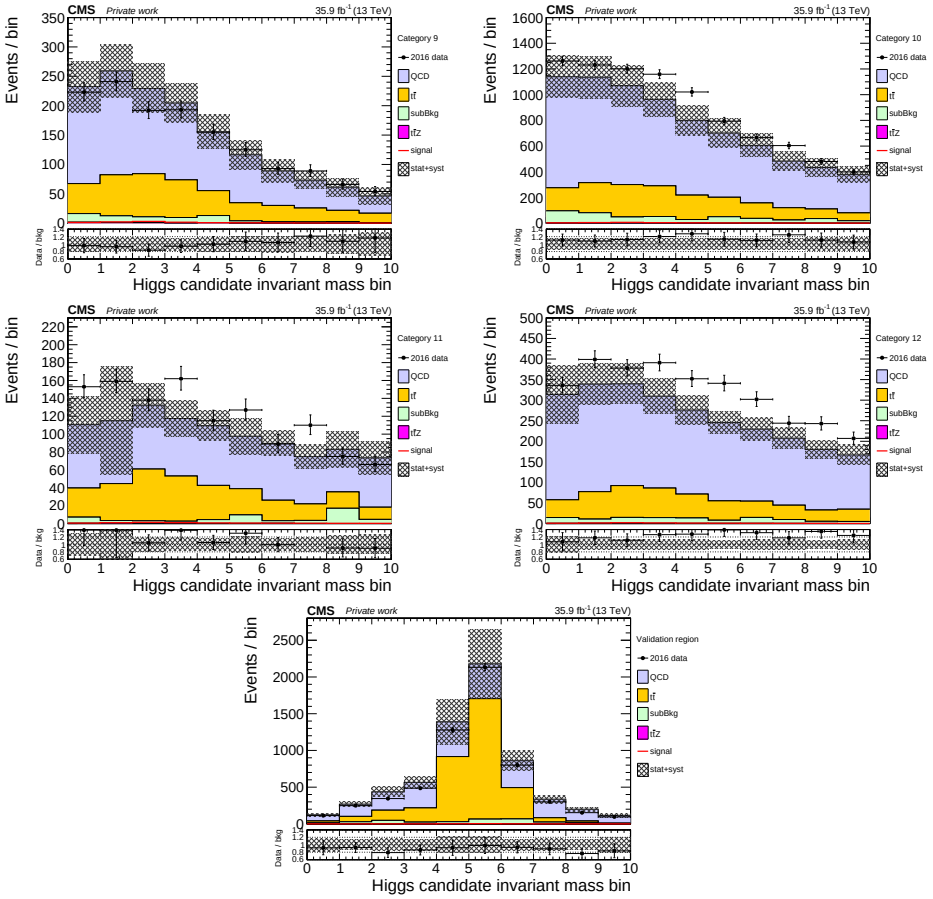

Figure 6.3: Comparison between the pre-fit template shapes and the observed data for the signal categories and for the $t\bar{t}$ VR. The sum of all the background processes is shown as a stacked plot, while the signal distributions are superimposed. The observed data are shown as black dots. The pre-fit uncertainties on the total background are reported as shaded regions. The ratio between the observed data and the total background is shown in the bottom panels.

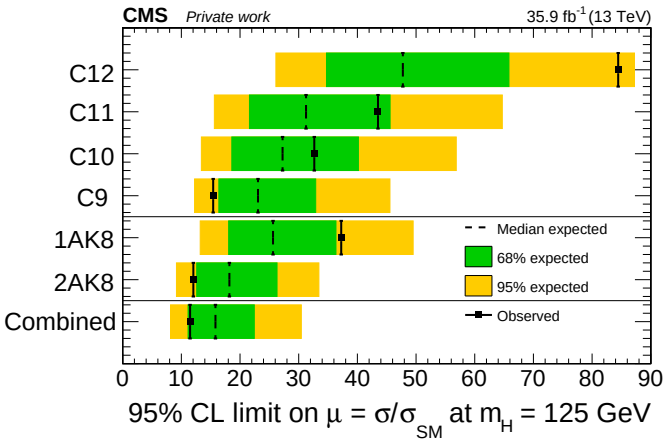

Figure 6.4: Observed and expected upper limits at 95% CL in each signal category, together with the observed and expected upper limit which result from the combination. Also, the observed and expected upper limits which result from the combination of categories with 2 and 1 AK8 jets are reported. The expected limits are displayed with the 68% and 95% CL intervals. In all the cases, the fit has been performed simultaneously with the VR in order to constrain the uncertainties related to the $t\bar{t}$ process.

In order to check the sanity of the fitting procedure also in the case of a fit to the observed data sample, we plot the pulls of the NPs and their impacts on the signal strength parameter. The result of such a check is shown in Fig. 6.5.

In the ideal case in which the pre-fit prediction perfectly describes the observed data, the NPs should not be pulled. As this is not, obviously, the real-life case, we rather expect the pulls to spread in a roughly Gaussian way around the value of zero. Also, considering the definition of the pulls given by Eq. 6.7, we expect the standard deviation of this distribution to be 1, *i.e.*, 95% of the NPs should lie in the pull mass window $[-2, 2]$. By having a look at Fig. 6.5, we see that this is indeed the case, with an important exception which we shall describe below.

The NP called "CMS_subBkgnormC12", which is related to the uncertainty on the normalization of the subdominant backgrounds in category 12, is strongly pulled, with a fitted value of 4.16. This can be explained by checking the pre-fit prediction of the data sample given in Fig. 6.3. We shall note that, in category 12, our pre-fit prediction underestimates the data yield, and thus, the fit accommodates this by strongly increasing the normalization of the subdominant backgrounds.

Regarding the constraints on the nuisance parameters, we shall give below an explanation for the NPs which are found to impact the most on

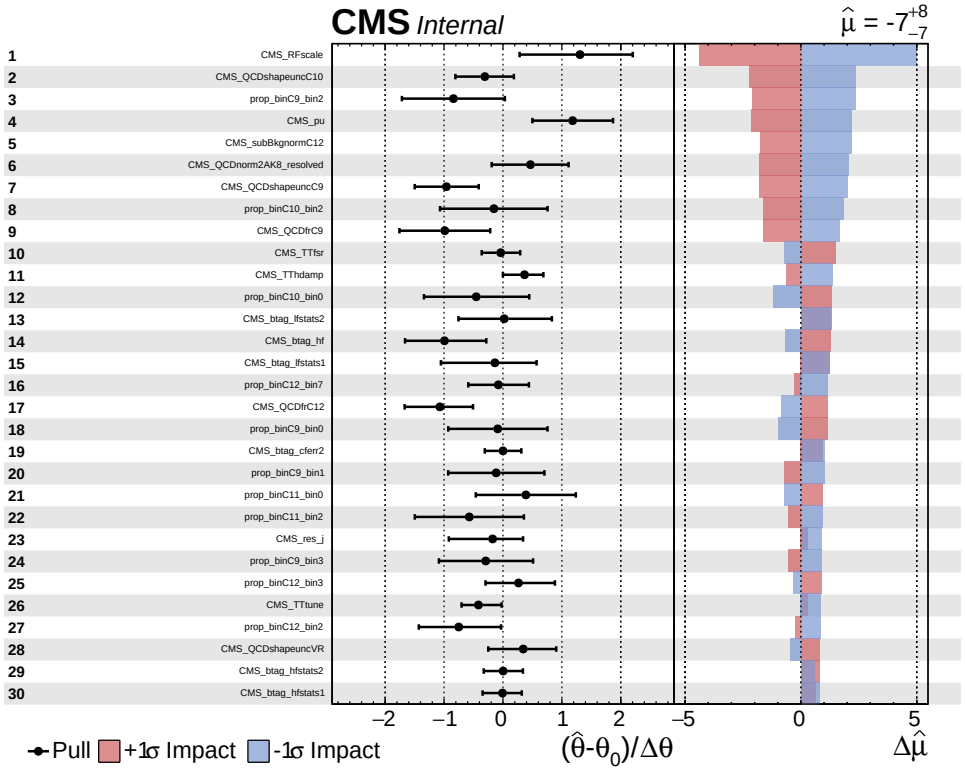

Figure 6.5: Pulls and impacts for the 30 nuisance parameters that affect the most the signal strength $\mu_{t\bar{t}H}$ in the RHC, as the result of the fit to the observed data sample.

the signal strength. As far as the NPs related to the uncertainties on the shape of the QCD background are concerned ("CMS_QCDshapeuncC9" and "CMS_QCDshapeuncC10"), several methods have been tested in order to estimate this source of uncertainty and, eventually, the most conservative one has been chosen. Thus, our choice is probably over-conservative and the fit is constraining this uncertainty to its effective magnitude. The NP which deals with the uncertainty on the profile of the PU distribution ("CMS_pu") is assumed to be correlated between categories and processes, thus leading to a moderate constraint. Similarly, the NP which treats the uncertainty on the QCD yield for categories with 2AK8 jets ("CMS_QCDnorm2AK8_resolved") is assumed to be correlated between categories, thus leading to a moderate constraint. Finally, the NP which deals with the uncertainty on the final state radiation in the $t\bar{t}$ process ("CMS_TTfsr") is assumed to be correlated between categories, and the inclusion of the $t\bar{t}$-enriched VR in the fit leads to a significant constraint. The pulls and impacts for the full set of nuisance parameters entering the analysis can be found in Appendix B.2.

The final distributions for the invariant mass of the Higgs boson candidate,

resulting after the combined fit to the data, are shown in Fig. 6.6.

6.4 Combination of the RHC and BHC

As it was pointed out in Chapter 1, the RHC described in this work, where the Higgs boson decays to a couple of well-separated, b tagged jets, is part of a wider $t\bar{t}H$ search which also involves a second channel, the BHC, in which the Higgs boson decays to a large-radius jet with peculiar substructure. We described in Subsection 4.4.2 that the two channel have been designed to be mutually exclusive, and thus, a combination can be easily performed in order to enhance the overall analysis sensitivity.

The BHC consists of nine, mutually exclusive signal categories targeting $t\bar{t}H$ events in which the decay of the Higgs boson is collected using an AK8 jet. The soft drop mass of the AK8 jet which is identified to be the Higgs candidate is the observable used in this channel to perform the final fit. In a similar fashion to what has been discussed in Subsection 4.4.5, the signal categories belonging to the BHC are determined based on the AK8 and AK4 multiplicities, as well as the number of top quark candidates. The same $t\bar{t}$ validation region that has been described in Subsection 4.7.2 is used to obtain a constraint on the uncertainties related to the $t\bar{t}$ process.

Also, whenever it is possible, the systematic uncertainties are treated in the same way in the two channels, in order to guarantee a simple and coherent combination procedure. The same sanity checks concerning the blinded results that we showed in this work have been performed in the BHC as well.

By performing a simultaneous ML fit to the data sample in the signal categories and in the $t\bar{t}$ VR, the observed signal strength in the BHC is found to be $\hat{\mu}_{t\bar{t}H} = -1.5^{+5.3}_{-5.4}$, a value which is compatible with the SM expectation. Since the fitted value is also compatible with the background-only hypothesis, additionally, an exclusion limit at 95% CL using the asymptotic approximation of the CL_s prescription is set. The combination of all the signal categories and the $t\bar{t}$ VR leads to observed and expected upper limits on $\mu_{t\bar{t}H}$ which are found to be $\mu_{t\bar{t}H} < 9.4$ and < 10.4, respectively.

By combining the RHC and BHC, the final $t\bar{t}H$ analysis results in 13, mutually exclusive signal categories which are simultaneously fitted together with the $t\bar{t}$ VR. As a result of the fit, the observed value for the signal strength parameter is found to be

$$\hat{\mu}^{\text{comb}}_{t\bar{t}H} = -3^{+4}_{-5}, \tag{6.10}$$

a value which is consistent with the SM expectation. Since the fitted value is also compatible with the background-only hypothesis, we additionally set an an exclusion limit at 95% CL using the asymptotic approximation of the CL_s prescription. Combining all the signal categories and the $t\bar{t}$ VR, the observed

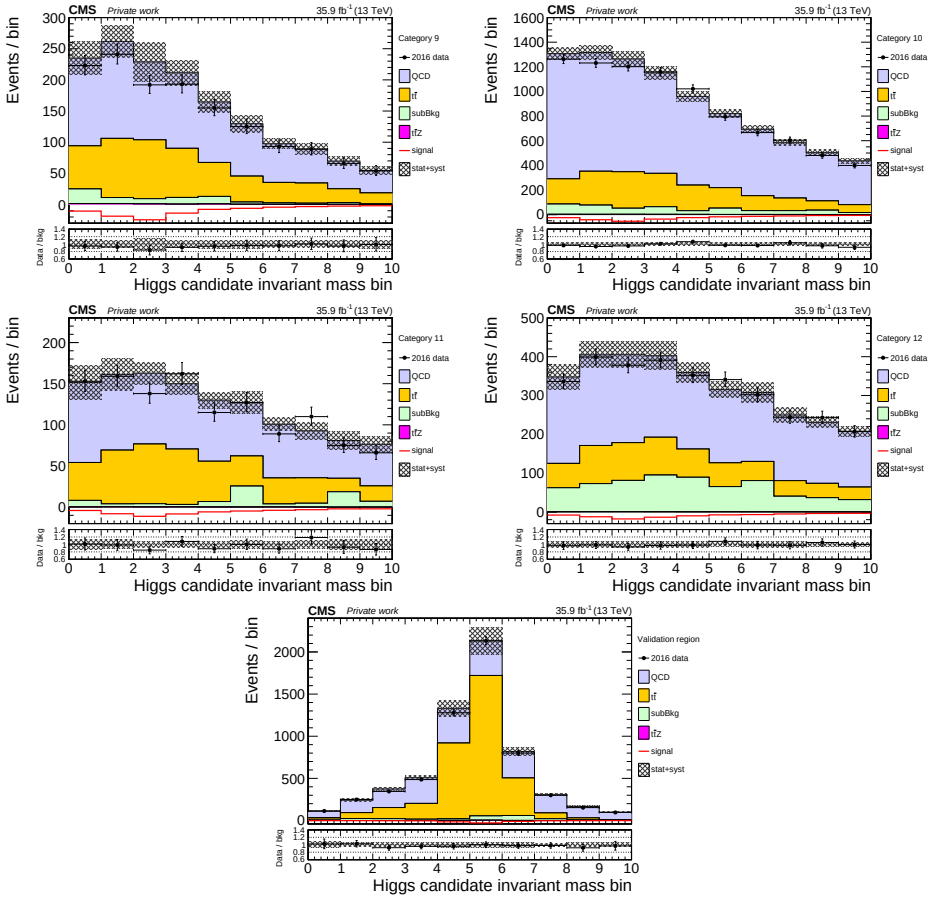

Figure 6.6: Comparison between the post-fit distributions of the invariant mass of the Higgs boson candidate and the observed data for the signal categories and for the $t\bar{t}$ VR. The sum of all the background processes is shown as a stacked plot, while the signal distributions are superimposed. The observed data are shown as black dots. The pre-fit uncertainties on the total background are reported as shaded regions. The ratio between the observed data and the total background is shown in the bottom panels.

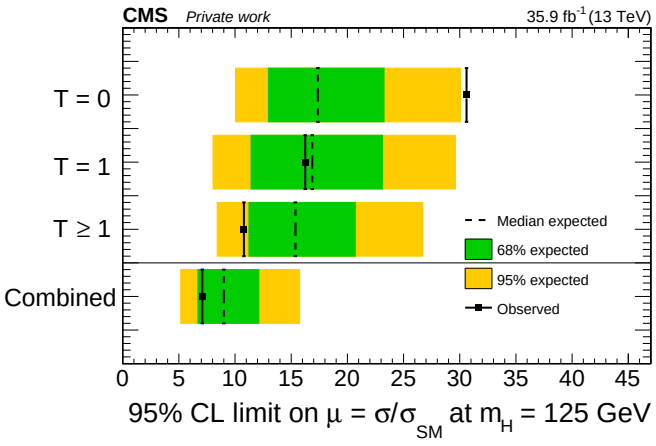

Figure 6.7: Observed and expected upper limits at 95% CL resulting from the combination of the RHC and the BHC, together with the observed and expected upper limits which result from the combination of categories with at least one top candidate, exactly one top candidate and no top candidates. As expected, the presence of top candidates makes the categories more sensitive. The expected limits are displayed with the 68% and 95% CL intervals. In all the cases, the fit has been performed simultaneously with the VR in order to constrain the uncertainties related to the $t\bar{t}$ process.

and expected upper limits on $\mu_{t\bar{t}H}$ are found to be $\mu_{t\bar{t}H} < 7.1$ and < 9.0, respectively. These combined limits are shown in Fig. 6.7, where we also show the expected and observed upper limits obtained by the combination of categories with at least one top candidate, exactly one top candidate and no top candidates.

In Fig. 6.8 we show the pulls of the NPs and their impacts on the signal strength parameter as the result of the combination of the RHC and the BHC. As expected, the NPs impacting the most on the signal strength are the ones which are related to the normalization of the dominant background of the search. The pulls and impacts for the full set of nuisance parameters entering the combined analysis can be found in Appendix B.3.

6.5 Conclusions and prospects

In this work, we developed a new approach to the searches for fully hadronic $t\bar{t}H$ events, which can be considered to be complementary to the searches that are currently performed by the CMS Collaboration. In fact, we explored a different (or at least, partially different) phase space by looking for event containing at least one large-radius jet. This strategy makes possible to reduce the large number of combinatorial permutations of jets in the event,

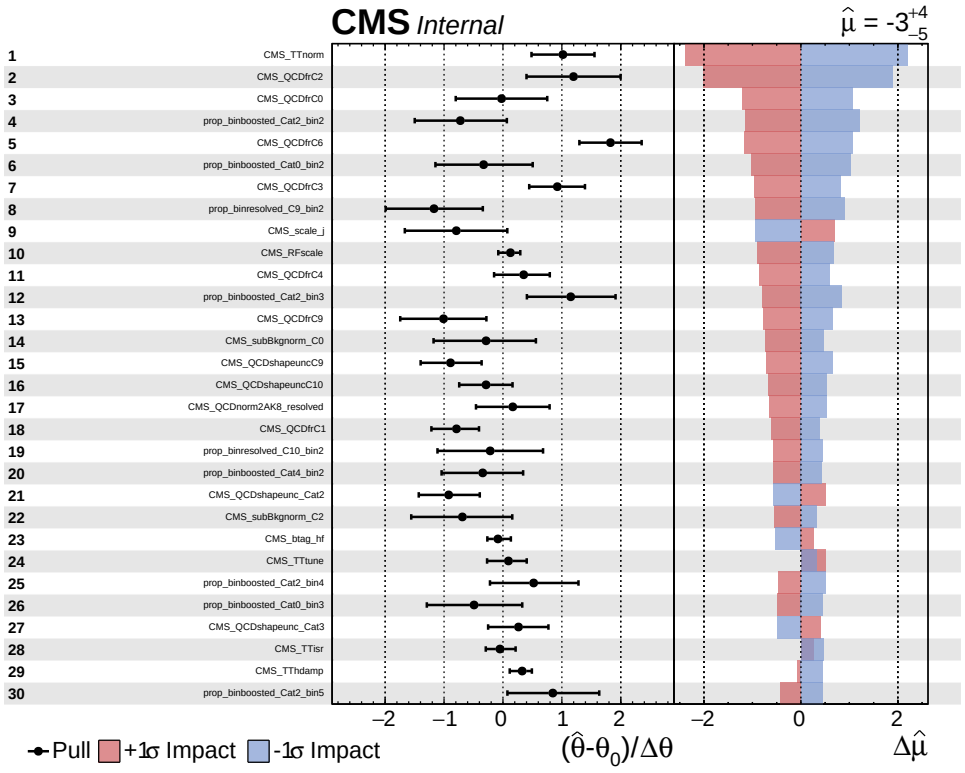

Figure 6.8: Pulls and impacts for the 30 nuisance parameters that affect the most the signal strength $\mu_{t\bar{t}H}$ in the combination of the RHC and the BHC, as the result of the fit to the observed data sample.

which is expected to show at least eight small-radius jets, and to infer the properties of the decaying particles by looking at the substructure of large-radius jets. Properties such as the number of energy prongs inside a jet, or the kinematic variables of subjets created by the soft drop algorithm, can be successfully used to separate the $t\bar{t}H$ signal from the overwhelming QCD multijet background. The analysis is split in two orthogonal channels, the resolved Higgs channel and the boosted Higgs channel, with the former being the focus of this work.

As a first step, we have developed a coherent trigger strategy by finding the best performing trigger in terms of the ratio between the trigger efficiency in data and background events. Also, a study of the trigger efficiency as a function of several variables of interest has been performed in order to find selection cuts that make the signal trigger fully efficient.

A baseline selection has been developed, aimed at selecting $t\bar{t}H$ events in the FH final state and in the boosted topology. The key features of this baseline selection are the absence of isolated leptons in the final state, which selects FH decays of the $t\bar{t}H$ triplet, the presence of at least one AK8 jet,

which defines boosted topologies, and the request for the scalar sum of the transverse momenta of all the jets in the event to be greater than 900 GeV, which makes the signal trigger fully efficient.

On top of the baseline selection, we have developed multivariate methods to identify Higgs boson and top quark candidates. Boosted-jets BDT are used to tag boosted Higgs boson and top quark candidates, which are reconstructed as AK8 jets with peculiar substructure properties, while a resolved-Higgs BDT has been developed to identify $t\bar{t}H$ events in which a Higgs boson decays to a pair of well-resolved, b tagged AK4 jets. The invariant mass of the pair which has been identified as the resolved Higgs candidate has been chosen as the target observable in the resolved Higgs channel. To make the RHC and BHC mutually exclusive channels, the absence or presence of a boosted Higgs candidate has been requested, respectively.

Events have been then split into four, mutually exclusive signal categories in order to enhance the sensitivity of the analysis. The split has been made based on the AK8 and AK4 multiplicities, as well as on the presence or absence of top quark candidates.

As a crucial step of the analysis, the main backgrounds of the search have been estimated. In order to avoid the large theoretical uncertainties involved in the simulation of the QCD multijet processes, this background has been estimated completely from data. Concerning the shapes of the invariant mass of the resolved Higgs candidate, they have been estimated by setting up a dedicated, QCD-enriched control region. Data distributions in this region have been corrected with transition functions to match the expected distributions in the signal region. Concerning the expected yields, they have been estimated using the ABCD method. In order to account for possible mismodellings in the $t\bar{t}$ simulation, a corrective factor for the $t\bar{t}$ cross section in the boosted regime has been derived. Also, a $t\bar{t}$-enriched validation region has been implemented, which is used to constrain the systematic uncertainties related to the $t\bar{t}$ process.

Six kind of processes have been considered in this analysis and translated to template shapes. First, both the $t\bar{t}H(b\bar{b})$ and $t\bar{t}H(nob\bar{b})$ processes have been considered to contribute to the signal; then, we have taken into account the QCD multijet dominant background, the $t\bar{t}$ background, the $t\bar{t}Z$ irreducible background and the subdominant backgrounds (coming from processes such as Drell–Yan, single top quark production, diboson production, etc.), which are added together to form a single template shape.

The effect of systematic uncertainties on both the shapes and normalizations of the template shapes has been fully taken into account. Several sources of uncertainty required a customized treatment, such as the uncertainties related to the shape of the QCD templates, or the uncertainties on the QCD yield coming from the ABCD method. Alternatively, in every case in which a commonly agreed procedure exists, the standard methods recommended by the CMS Collaboration have been used to estimate the

impact of the systematic uncertainties.

A simultaneous, maximum-likelihood fit to the data collected by the CMS experiment during the 2016 data-taking period, and corresponding to an integrated luminosity of 35.9 fb^{-1}, has been performed, including the four signal categories and the $\text{t}\bar{\text{t}}$ validation region, in order to extract the signal strength parameter $\mu_{\text{t}\bar{\text{t}}\text{H}}$, defined as the ratio of the measured $\text{t}\bar{\text{t}}\text{H}$ production cross section to the one expected from SM calculations. As a result, the fitted value for the signal strength parameter has been found to be $\hat{\mu}_{\text{t}\bar{\text{t}}\text{H}} = -7^{+8}_{-7}$, a value which is consistent with the expectation from the standard model, but also with the background-only hypothesis. For this reason, we have also set an observed (expected) upper limit on the signal strength parameter, which is found to be 11.5 (15.8) times the expectation from the standard model.

The combination of the RHC and the BHC resulted in the simultaneous fit of 13 signal categories, together with the $\text{t}\bar{\text{t}}$ validation region. As a result, the fitted value for the signal strength parameter has been found to be $\hat{\mu}_{\text{t}\bar{\text{t}}\text{H}} = -3^{+4}_{-5}$, a value which is consistent with the expectation from the standard model, but also with the background-only hypothesis. For this reason, we have also set an observed (expected) upper limit on the signal strength parameter, which is found to be 7.1 (9.0) times the expectation from the standard model. By performing the combination, the uncertainty on the signal strength value is reduced with respect to the values found in the individual channels, and the upper limit is found to be lower. This results in the first search for $\text{t}\bar{\text{t}}\text{H}$ events in the FH final state using large-radius jets performed by the CMS Collaboration.

In the future, several sources of improvement are expected for this analysis. First, we shall mention the obvious reduction in statistical uncertainties that would come from the analysis of the 2017 and 2018 datasets, which would increase the integrated luminosity by almost a factor of four. A simple way to get an estimation of the gain that would result from the processing of the full Run2 data is to scale the expected yields of all the template shapes entering the ML fit to the integrated luminosity collected during Run2 (140 fb^{-1}) and perform a new fit to the corresponding Asimov dataset. By doing so, the expected (blinded) upper limit on the signal strength, corresponding to the full Run2 integrated luminosity, is found to be

$$\text{med}[\mu_{\text{t}\bar{\text{t}}\text{H},\,\text{up}}|0]_{\text{Run2}} = 12.3, \tag{6.11}$$

to be compared with the value of 15.9 found for the 2016 dataset only (see Eq. 6.8). The exploitation of the full Run2 dataset leads, in the RHC, to a roughly 25% lower expected limit on the signal strength. An analogous test can be made for the combination of RHC and BHC, which results in an expected (blinded) upper limit on the signal strength of 6.9, to be compared with the corresponding value obtained using the 2016 dataset only, which

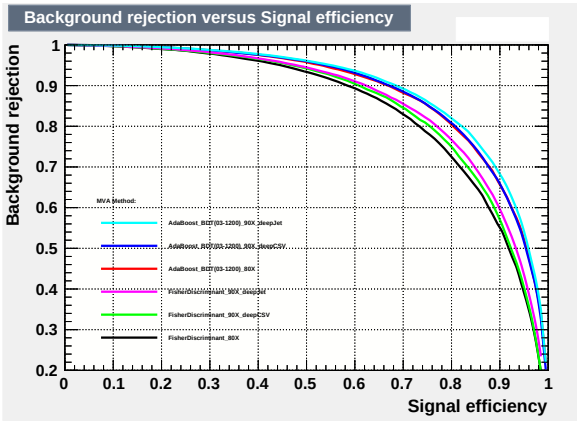

Figure 6.9: Performance of the resolved-Higgs BDT, expressed as a background rejection vs. signal efficiency ROC curve, when trained with CSVv2, DeepCSV and DeepJet discriminants as input variables. Fisher discriminants have been trained as well, as a reference.

is found to be 8.5. The exploitation of the full Run2 dataset leads, for the combination of RHC and BHC, to a roughly 20% lower expected limit on the signal strength.

Moreover, the analysis has been developed using robust and well-understood methods, and improvements may arise from the exploitation of more advanced algorithms. For example, b tagging techniques based on deep neural networks could be exploited, such as DeepCSV [93] or DeepJet [94]. Figure 6.9 shows the results, expressed in terms of the ROC curve, of the retraining of the resolved-Higgs BDT, where the b tagging discriminators that are used as input variables come from the standard CSVv2 algorithm, from DeepCSV and from DeepJet. As a result, we see that the performance of the BDT is comparable with the standard approach when using DeepCSV and slightly increased when using DeepJet. Also, even more advanced b tagging algorithms, based on graph neural networks, are currently under development and are expected to become available in the future.

As a second example of advanced techniques that could improve the analysis performance, we shall mention the novel jet tagging algorithms that have been developed inside the CMS Collaboration. For example, DeepAK8 [95] is a multi-class identification algorithm, based on AK8 jets, which is able to identify the hadronic decays of top quarks, Higgs bosons and vector bosons. Also in this case, intense work is ongoing inside the CMS Collaboration in order to develop even more advanced techniques based on graph neural networks, such as ParticleNet [96].

Finally, in the future runs of the LHC, the PU conditions are expected to change. While we faced an average number of PU interaction per bunch crossing $\langle \text{nPU} \rangle \approx 30$ during Run2, this value is expected to increase up to

$\langle \text{nPU} \rangle \approx 60$ during Run3 and up to $\langle \text{nPU} \rangle \approx 140$ during the High-Luminosity LHC project. This will result in a very crowded environment, which will require more advanced PU mitigation techniques and which will make it more difficult to trigger AK4 jets, that could be masked by the overwhelming PU. In this context, boosted topologies are expected to become more important as they will enhance the chance of selecting $t\bar{t}H$ events at trigger level.

6.6 Acknowledgments

I would like to spend my last words to express my deepest gratitude to all the people that made this work possible: to Andrea Castro, who first introduced me to data analysis and who always supervised, helped and supported me during all my years at the University of Bologna; to Konstantinos Kousouris, the first person I met at CERN, who has been a great teacher during past projects and who gave his precious supervision to this work; to Garyfallia Paspalaki, for being such a good workmate and friend.

To express how much I am grateful to my parents for their love and constant support, not only throughout my Ph.D. years, but throughout my whole life, words tend to be inadequate.

Appendices

Appendix A

Template shapes

A.1 Template shapes for category 9

In Fig. A.1, we list the template shapes entering the final fit for events belonging to category 9.

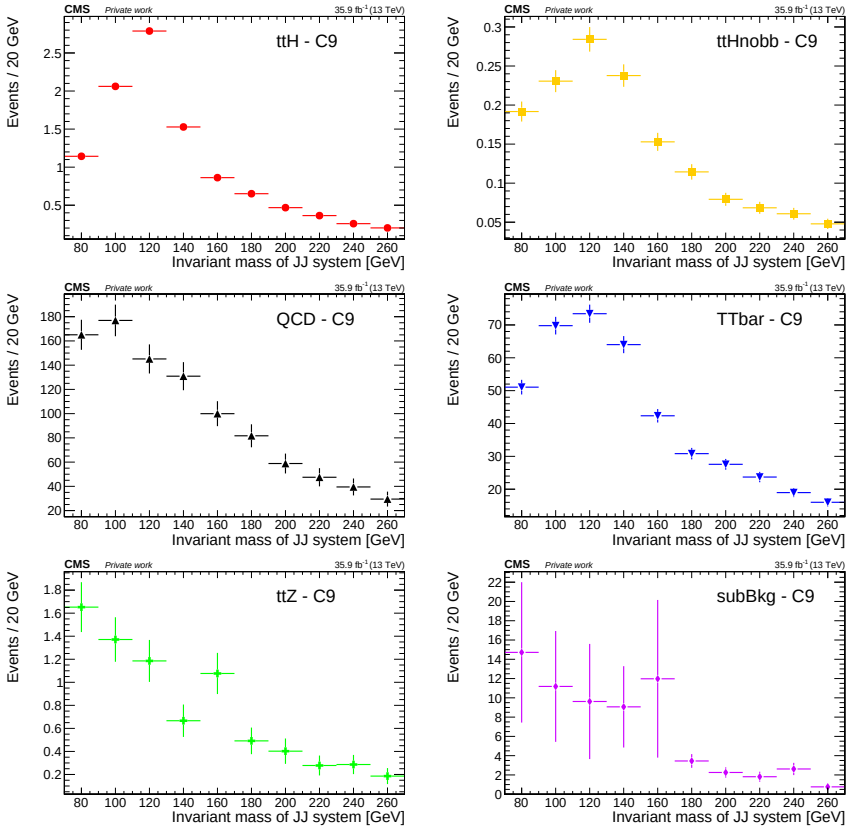

Figure A.1: Template shapes for events in category 9. Each template is normalized to the expected yield in the 2016 data-taking period.

A.2 Template shapes for category 10

In Fig. A.2, we list the template shapes entering the final fit for events belonging to category 10.

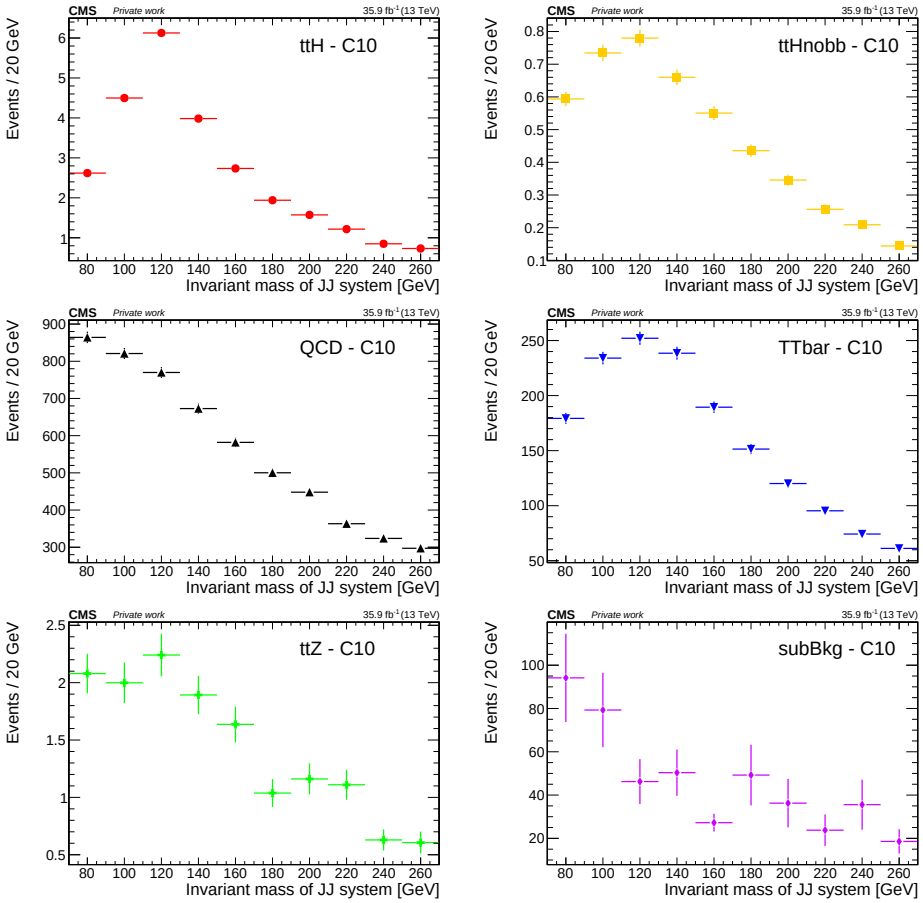

Figure A.2: Template shapes for events in category 10. Each template is normalized to the expected yield in the 2016 data-taking period.

A.3 Template shapes for category 11

In Fig. A.3, we list the template shapes entering the final fit for events belonging to category 11.

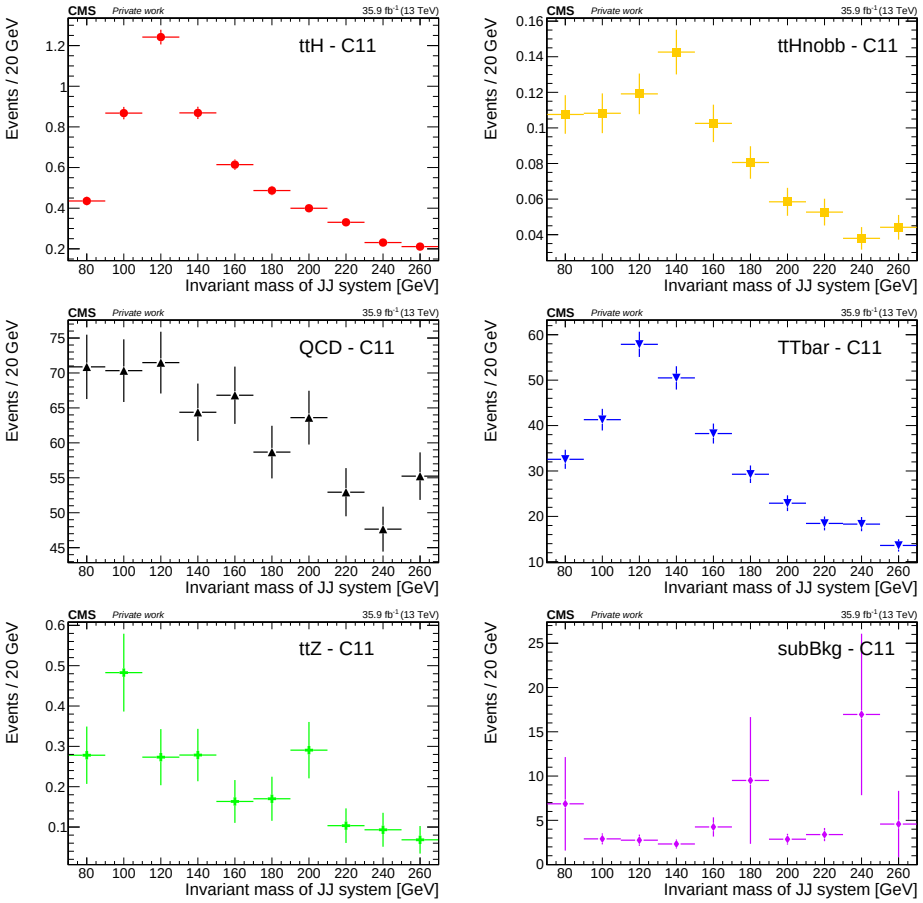

Figure A.3: Template shapes for events in category 11. Each template is normalized to the expected yield in the 2016 data-taking period.

A.4 Template shapes for category 12

In Fig. A.4, we list the template shapes entering the final fit for events belonging to category 12.

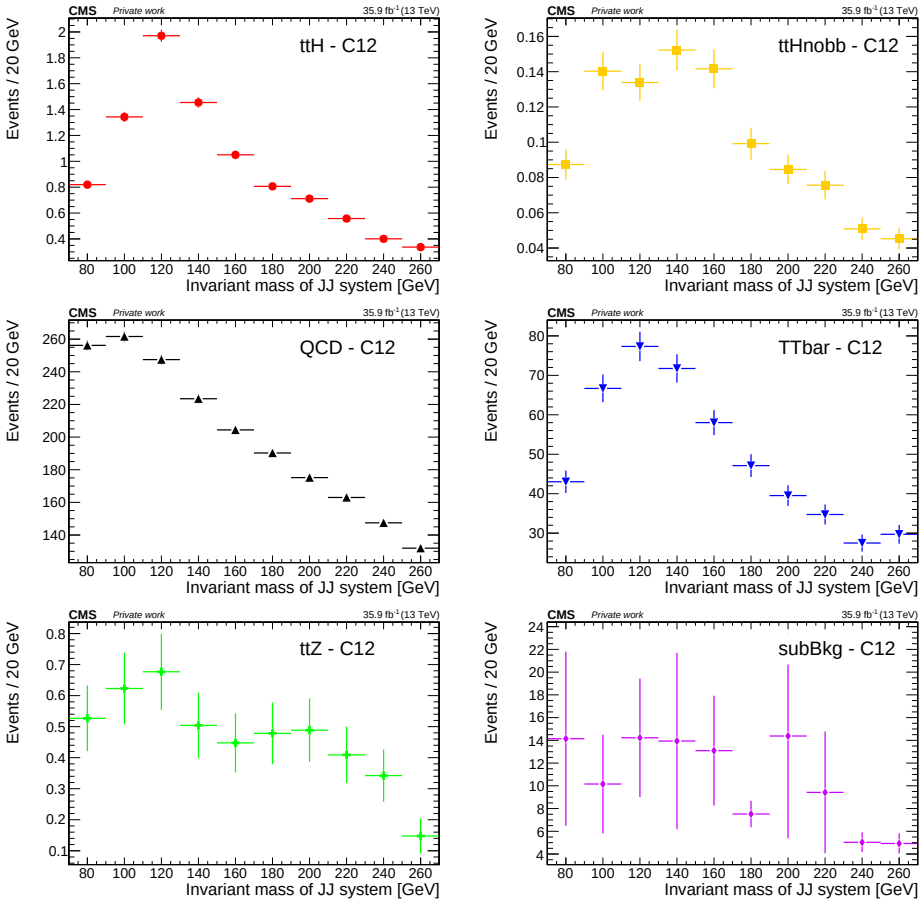

Figure A.4: Template shapes for events in category 12. Each template is normalized to the expected yield in the 2016 data-taking period.

A.5 Template shapes for the $t\bar{t}$ VR

In Fig. A.5, we list the template shapes entering the final fit for events belonging to the $t\bar{t}$ VR.

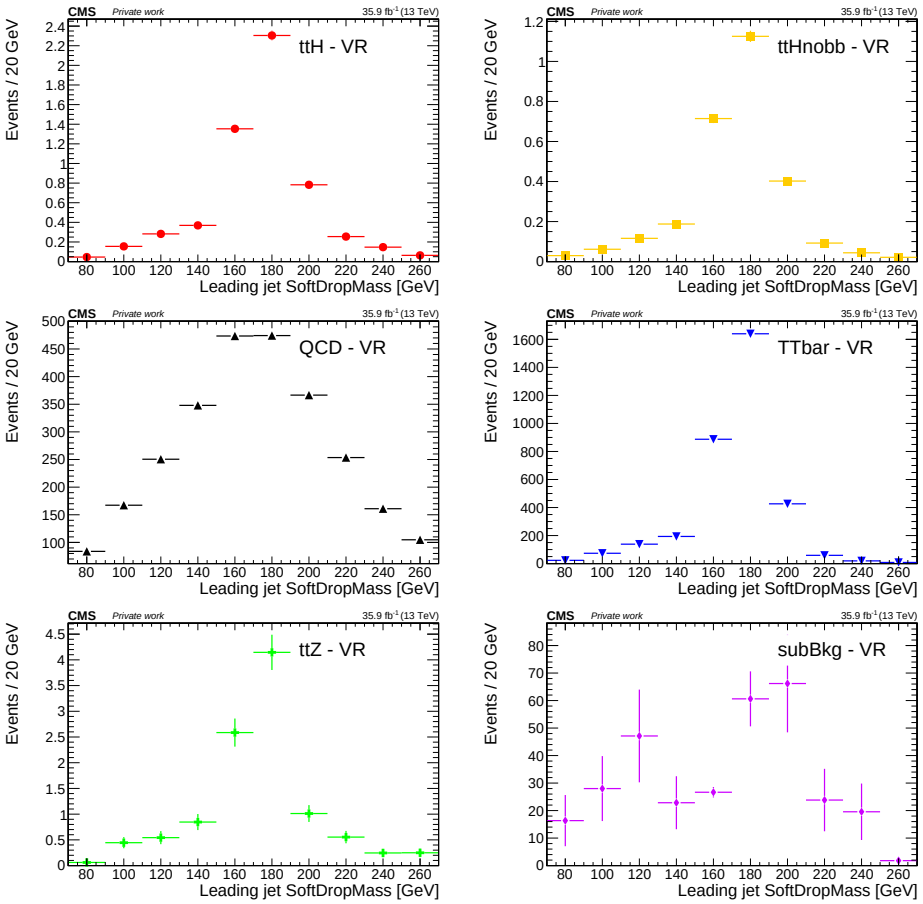

Figure A.5: Template shapes for events in the $t\bar{t}$ VR. Each template is normalized to the expected yield in the 2016 data-taking period.

Appendix B

Pulls and impacts plots

B.1 Blinded results

In Figs. B.1 and B.2 we report the full list of the nuisance parameters entering the fit to the Asimov dataset in the S+B hypothesis, with the corresponding pulls and impacts on the signal strength parameter.

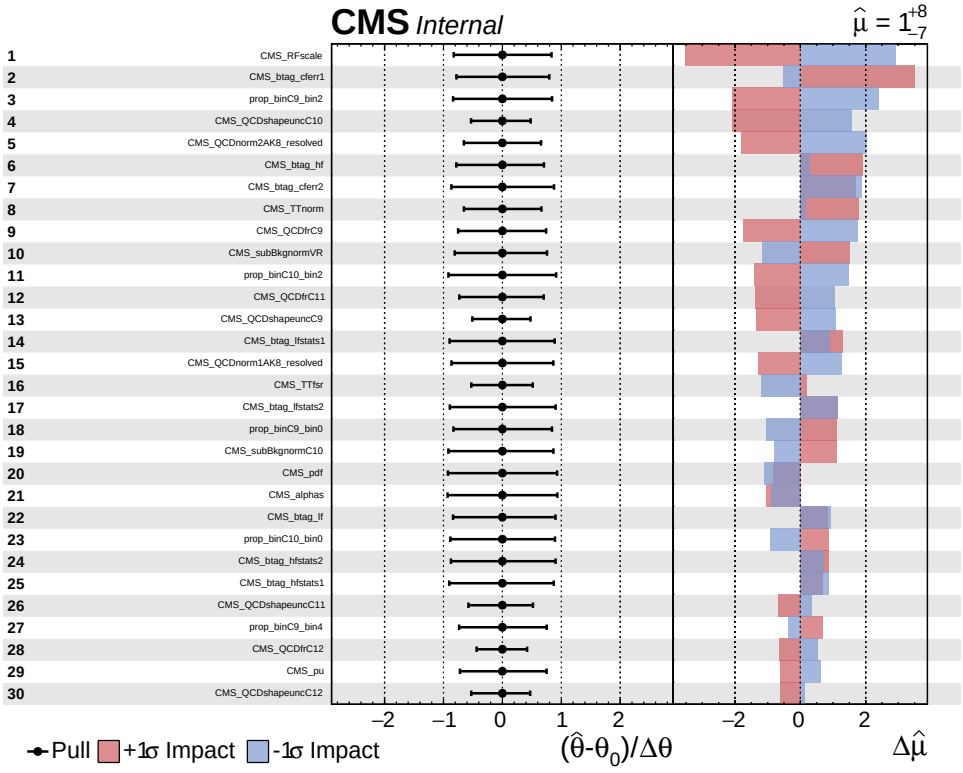

Figure B.1: Pulls and impacts for nuisance parameters in the RHC, as the result of the fit to the S+B Asimov dataset.

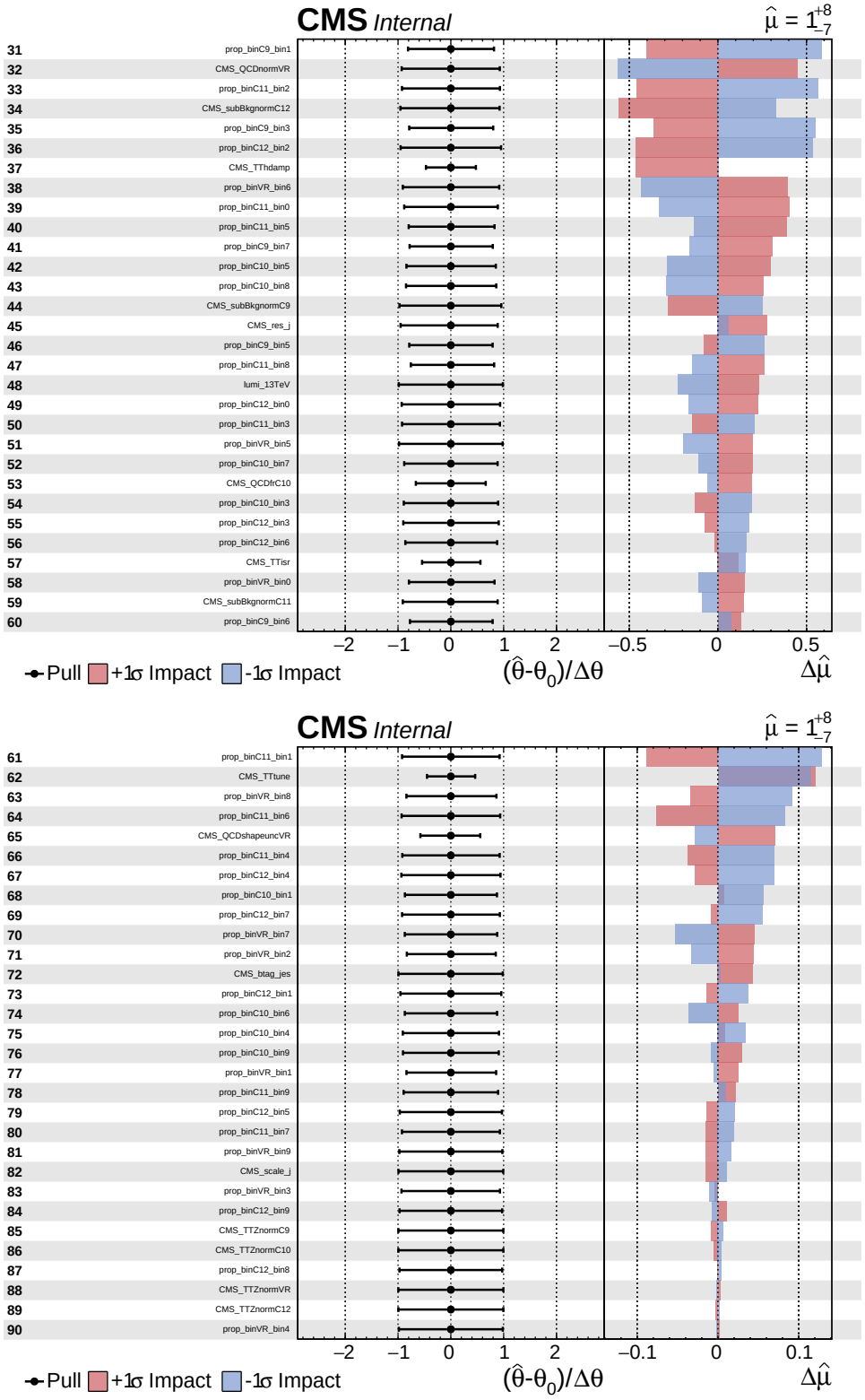

Figure B.2: Pulls and impacts for nuisance parameters in the RHC as the result of the fit to the S+B Asimov dataset.

B.2 Unblinded results

In Figs. B.3 and B.4 we report the full list of the nuisance parameters entering the fit to the observed data sample, with the corresponding pulls and impacts on the signal strength parameter.

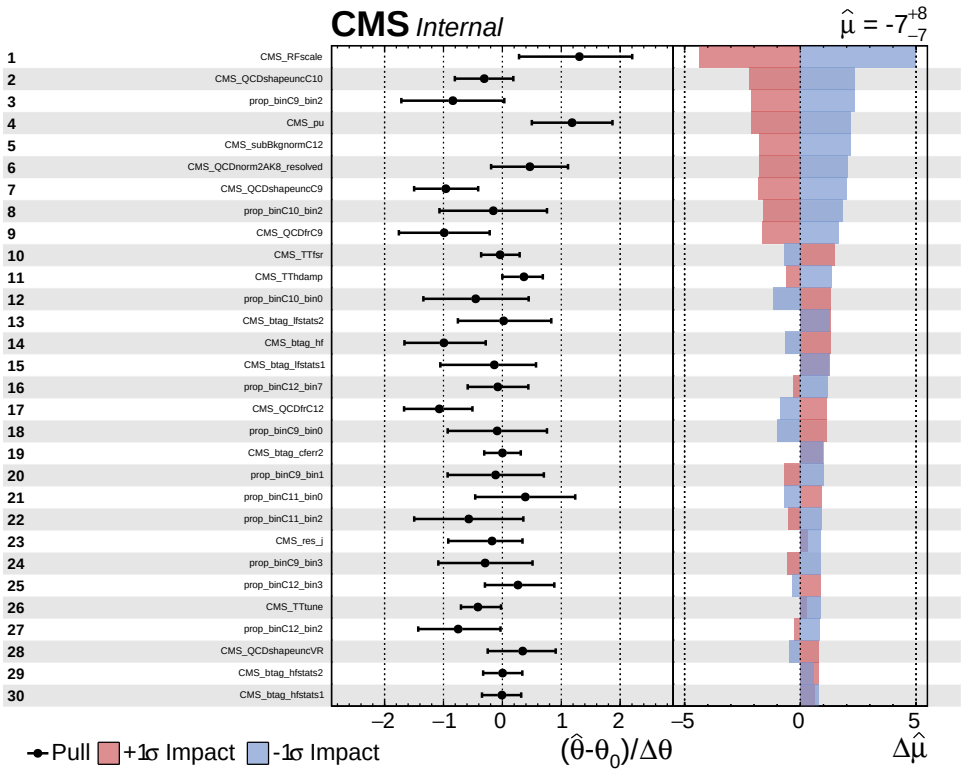

Figure B.3: Pulls and impacts for nuisance parameters in the RHC, as the result of the fit to the observed data sample.

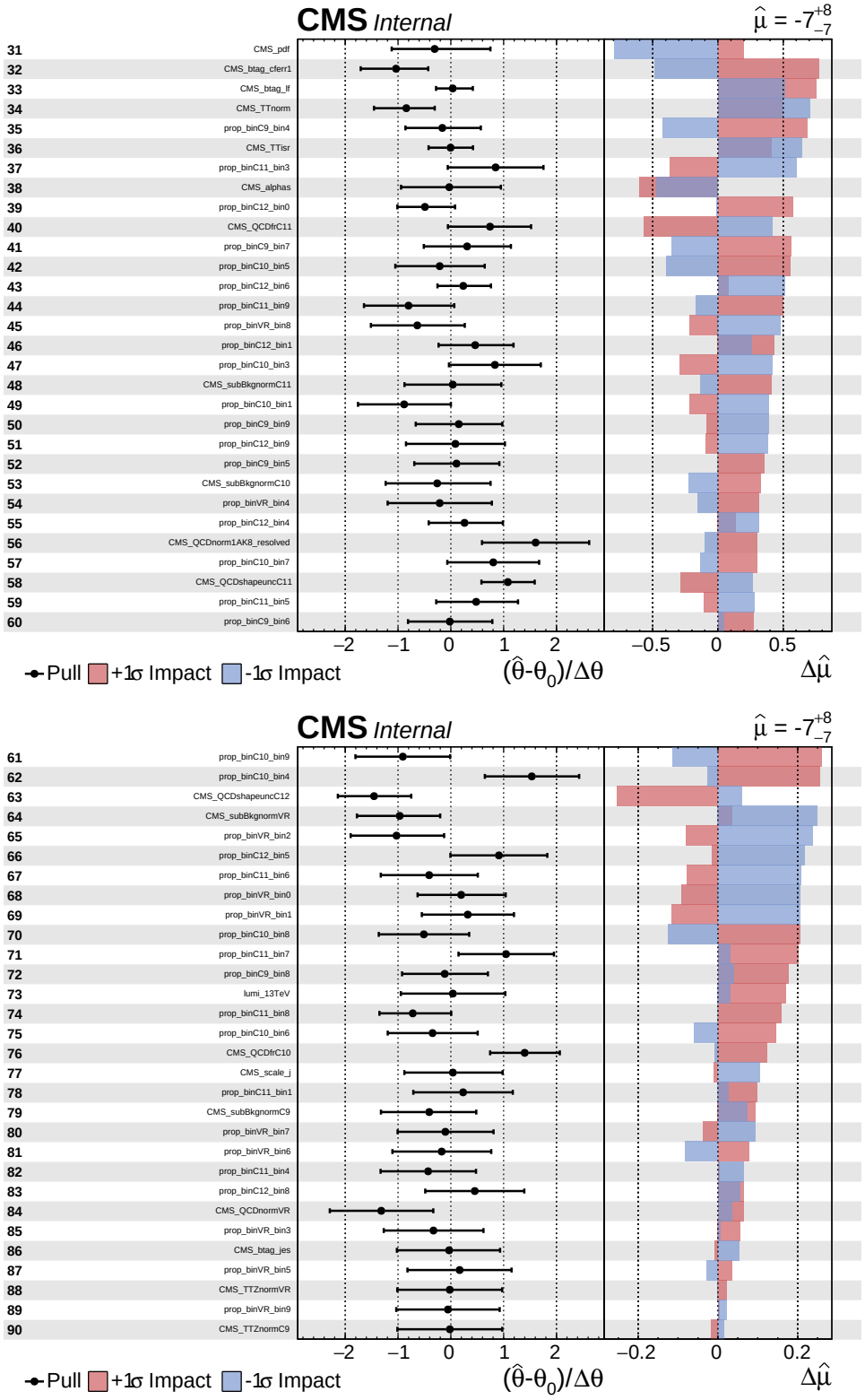

Figure B.4: Pulls and impacts for nuisance parameters in the RHC as the result of the fit to the observed data sample.

B.3 Combination

In Figs. B.5, B.6 and B.7 we report the full list of the nuisance parameters entering the fit to the observed data sample for the combination of the RHC and the BHC, with the corresponding pulls and impacts on the signal strength parameter.

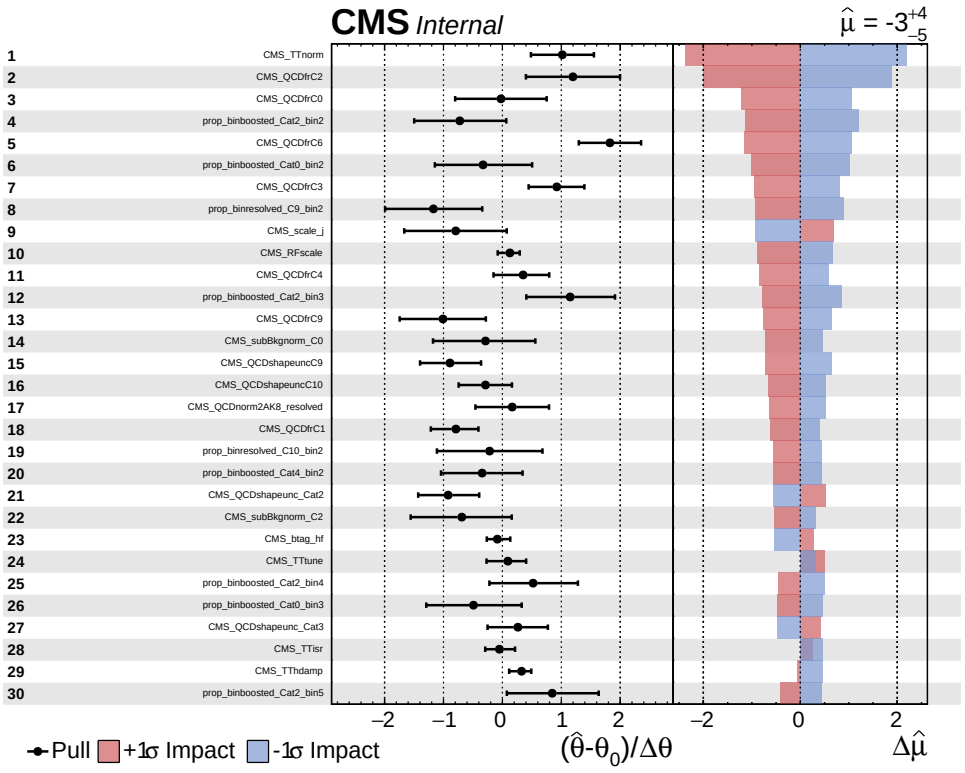

Figure B.5: Pulls and impacts for nuisance parameters in the combination of the RHC and the BHC as the result of the fit to the observed data sample.

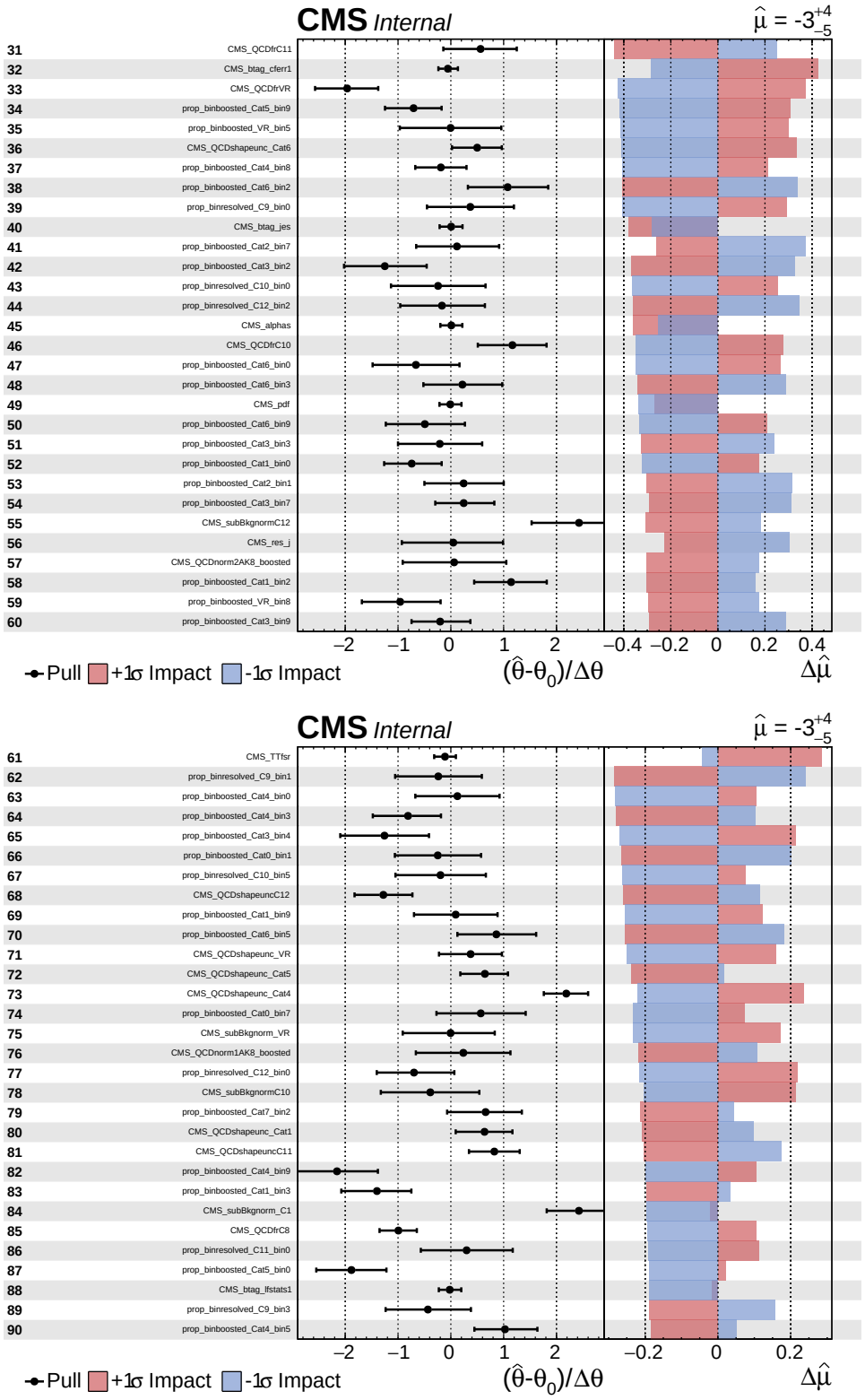

Figure B.6: Pulls and impacts for nuisance parameters in the combination of the RHC and the BHC as the result of the fit to the observed data sample.

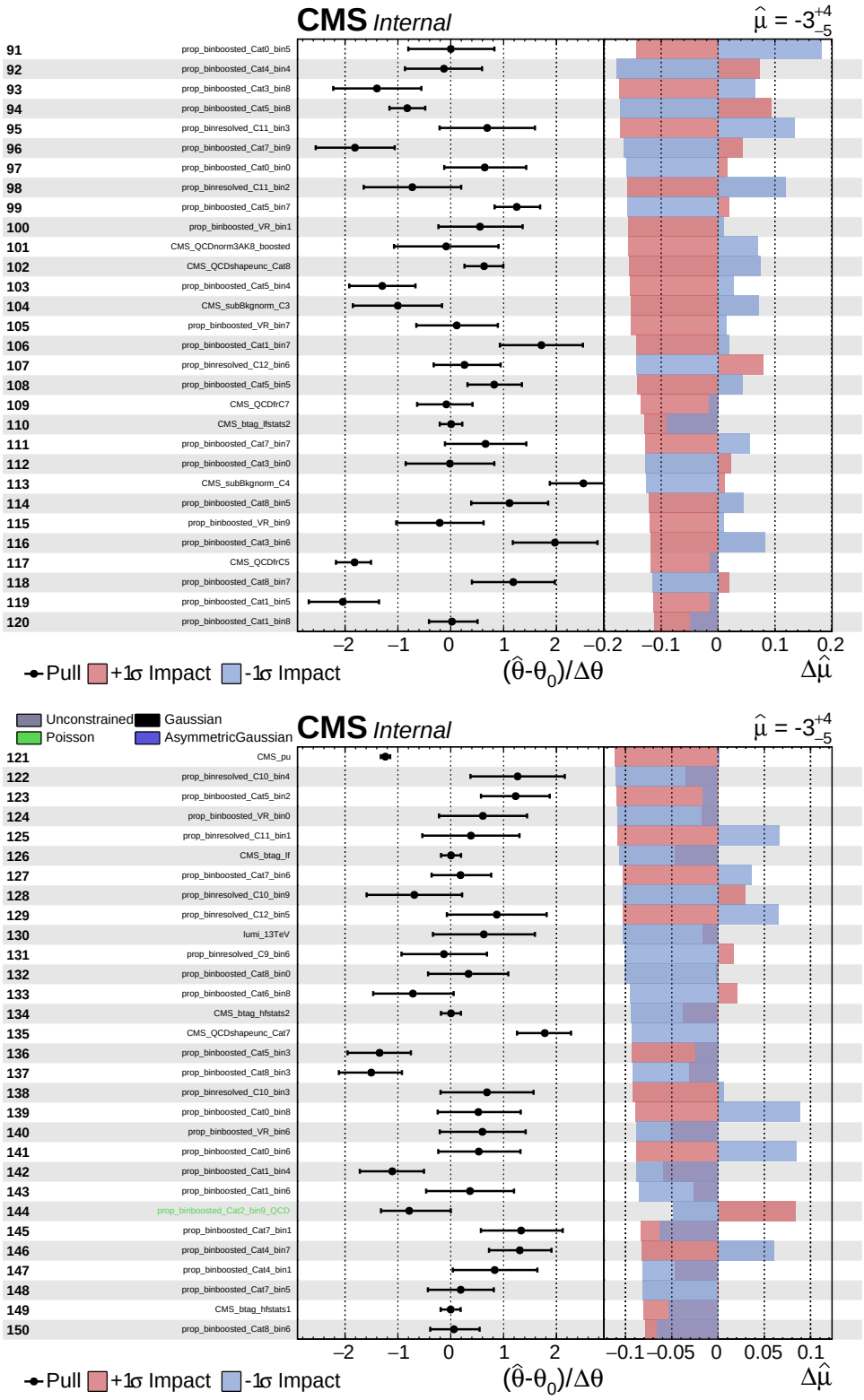

Figure B.7: Pulls and impacts for nuisance parameters in the combination of the RHC and the BHC as the result of the fit to the observed data sample.

Bibliography

[1] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC.* Phys. Lett. B **716** (2012) 1.

[2] 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.

[3] ATLAS Collaboration, *Evidence for the spin-0 nature of the Higgs boson using ATLAS data.* Phys. Lett. B **726** (2013) 120.

[4] 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.

[5] CMS Collaboration, *Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV.* Eur. Phys. J. C **75** (2015) 212.

[6] 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.

[7] CMS Collaboration, *A measurement of the Higgs boson mass in the diphoton decay channel.* Phys. Lett. B **805** (2020) 135425.

[8] M. Tanabashi et al., *Review of particle physics.* Phys. Rev. D **98** (2018) 030001.

[9] ATLAS Collaboration, *Observation of* $H \to b\bar{b}$ *decays and* VH *production with the ATLAS detector.* Phys. Lett. B **786** (2018) 59.

[10] CMS Collaboration, *Observation of Higgs boson decay to bottom quarks.* Phys. Rev. Lett. **121** (2018) 121801.

[11] ATLAS Collaboration, *Measurements of gluon-gluon fusion and vector-boson fusion Higgs boson production cross-sections in the $H \to WW^* \to e\nu\mu\nu$ decay channel in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector.* Phys. Lett. B **789** (2019) 508.

[12] CMS Collaboration, *Measurements of properties of the Higgs boson decaying to a W boson pair in pp collisions at $\sqrt{s} = 13$ TeV.* Phys. Lett. B **791** (2019) 96.

[13] 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.

[14] CMS Collaboration, *Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector.* Phys. Lett. B **779** (2018) 283.

[15] ATLAS Collaboration, *Measurement of the Higgs boson coupling properties in the $H \to ZZ^* \to 4\ell$ decay channel at $\sqrt{s} = 13$ TeV with the ATLAS detector.* JHEP **03** (2018) 095.

[16] CMS Collaboration, *Measurements of properties of the Higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s} = 13$ TeV.* JHEP **11** (2017) 047.

[17] CMS 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.

[18] CMS Collaboration, *Observation of the diphoton decay of the Higgs boson and measurement of its properties.* Eur. Phys. J. C **74** (2014) 3076.

[19] CMS Collaboration, *Evidence for Higgs boson decay to a pair of muons.* JHEP **01** (2021) 148.

[20] CDF Collaboration, *Observation of top quark production in $\bar{p}p$ collisions.* Phys. Rev. Lett. **74** (1995) 2626.

[21] D0 Collaboration, *Observation of the top quark.* Phys. Rev. Lett. **74** (1995) 2632.

[22] A. Olchevski and M. Winter, *High precision tests of the standard model and determination of the top quark and higgs boson masses.* Comptes Rendus Physique - C R PHYS **3** (2002) 1183.

[23] 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.

[24] CMS Collaboration, *Observation of* $\mathrm{t\bar{t}H}$ *production.* Phys. Rev. Lett. **120** (2018) 231801.

[25] S. Alekhin, A. Djouadi, and S. Moch, *The top quark and Higgs boson masses and the stability of the electroweak vacuum.* Phys. Lett. B **716** (2012) 214.

[26] G. Degrassi, S. Di Vita, J. Elias-Miro, J. R. Espinosa, G. F. Giudice, G. Isidori, and A. Strumia, *Higgs mass and vacuum stability in the Standard Model at NNLO.* JHEP **08** (2012) 098.

[27] D. Buttazzo, G. Degrassi, P. P. Giardino, G. F. Giudice, F. Sala, A. Salvio, and A. Strumia, *Investigating the near-criticality of the Higgs boson.* JHEP **12** (2013) 089.

[28] T. Markkanen, A. Rajantie, and S. Stopyra, *Cosmological Aspects of Higgs Vacuum Metastability.* Front. Astron. Space Sci. **5** (2018) 40.

[29] CMS Collaboration, *Search for* $\mathrm{t\bar{t}}H$ *production in the all-jet final state in proton-proton collisions at* $\sqrt{s} = 13$ *TeV.* JHEP **06** (2018) 101.

[30] CMS Collaboration, *Performance of quark/gluon discrimination in 8 TeV pp data*, CMS Physics Analysis Summary CMS-PAS-JME-13-002, 2013.

[31] CMS Collaboration, *Performance of quark/gluon discrimination in 13 TeV data*, CMS Detector Performance Note CMS-DP-2016-070, 2016.

[32] D0 Collaboration, *A precision measurement of the mass of the top quark.* Nature **429** (2004) 638.

[33] CMS Collaboration, *Search for a Standard Model Higgs Boson Produced in Association with a Top-Quark Pair and Decaying to Bottom Quarks Using a Matrix Element Method.* Eur. Phys. J. C **75** (2015) 251.

[34] CMS Collaboration, *Measurement of the* $\mathrm{t\bar{t}}$ *production cross section at 13 TeV in the all-jets final state*, CMS Physics Analysis Summary CMS-PAS-TOP-16-013, 2016.

[35] G. Apollinari, I. Béjar Alonso, O. Brüning, P. Fessia, M. Lamont, L. Rossi, and L. Tavian, *High-Luminosity Large Hadron Collider (HL-LHC): Technical Design Report V. 0.1.* CERN Yellow Reports: Monographs **4/2017** (2017).

[36] C. Quigg, *Gauge Theories of the Strong, Weak, and Electromagnetic Interactions: Second Edition*, Princeton University Press, USA, 9 2013.

[37] E. Noether, *Invariant Variation Problems.* Gott. Nachr. **1918** (1918) 235.

[38] C.-N. Yang and R. L. Mills, *Conservation of Isotopic Spin and Isotopic Gauge Invariance.* Phys. Rev. **96** (1954) 191.

[39] J. Goldstone, *Field Theories with Superconductor Solutions.* Nuovo Cim. **19** (1961) 154.

[40] F. Englert and R. Brout, *Broken symmetry and the mass of gauge vector mesons.* Phys. Rev. Lett. **13** (1964) 321.

[41] P. W. Higgs, *Broken symmetries and the masses of gauge bosons.* Phys. Rev. Lett. **13** (1964) 508.

[42] M. Shiozawa, *Evidence for neutrino oscillations in atmospheric neutrino observations.* Nucl. Instrum. Meth. A **433** (1999) 307.

[43] S. Glashow, *Partial Symmetries of Weak Interactions.* Nucl. Phys. **22** (1961) 579.

[44] N. Cabibbo, *Unitary Symmetry and Leptonic Decays.* Phys. Rev. Lett. **10** (1963) 531.

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

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

[47] W. Armstrong et al., *ATLAS: Technical proposal for a general-purpose pp experiment at the Large Hadron Collider at CERN.* CERN-LHCC-94-43 (1994).

[48] CMS Collaboration, *CMS, the Compact Muon Solenoid: Technical proposal.* CERN-LHCC-94-38, CERN-LHCC-P-1 (1994).

[49] S. Amato et al., *LHCb technical proposal: A Large Hadron Collider Beauty Experiment for Precision Measurements of CP Violation and Rare Decays.* CERN-LHCC-98-04, CERN-LHCC-P-4 (1998).

[50] ALICE Collaboration, *ALICE: Technical proposal for a large ion collider experiment at the CERN LHC.* CERN-LHCC-95-71, CERN-LHCC-P-3 (1995).

[51] J. Blewett, *200-GeV Intersecting Storage Accelerators.* eConf **C710920** (1971) 501.

[52] CDF and D0 Collaborations, *Updated Combination of CDF and D0 Searches for Standard Model Higgs Boson Production with up to 10.0 fb^{-1} of Data*, in Summer 2012 Conferences, 7 2012.

[53] F. Zwicky, *On the Masses of Nebulae and of Clusters of Nebulae.* Astrophys. J. **86** (1937) 217.

[54] M. Persic, P. Salucci, and F. Stel, *The Universal rotation curve of spiral galaxies: 1. The Dark matter connection.* Mon. Not. Roy. Astron. Soc. **281** (1996) 27.

[55] R. Barbieri, S. Ferrara, and C. A. Savoy, *Gauge Models with Spontaneously Broken Local Supersymmetry.* Phys. Lett. B **119** (1982) 343.

[56] J. Wess and B. Zumino, *Supergauge Transformations in Four-Dimensions.* Nucl. Phys. B **70** (1974) 39.

[57] J. C. Pati and A. Salam, *Unified Lepton-Hadron Symmetry and a Gauge Theory of the Basic Interactions.* Phys. Rev. D **8** (1973) 1240.

[58] H. Georgi and S. Glashow, *Unity of All Elementary Particle Forces.* Phys. Rev. Lett. **32** (1974) 438.

[59] R. Barbieri and G. F. Giudice, *Upper bounds on supersymmetric particle masses.* Nucl. Phys. B **306** (1987) 63.

[60] S. Weinberg, *Implications of Dynamical Symmetry Breaking.* Phys. Rev. **D13** (1976) 974.

[61] L. Susskind, *Dynamics of Spontaneous Symmetry Breaking in the Weinberg-Salam Theory.* Phys. Rev. D **20** (1979) 2619.

[62] A. Sakharov, *Violation of CP Invariance, C asymmetry, and baryon asymmetry of the universe.* Sov. Phys. Usp. **34** (1991) 392.

[63] E. V. Shuryak, *Theory of Hadronic Plasma.* Sov. Phys. JETP **47** (1978) 212.

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

[65] A. Dominguez, D. Abbaneo, K. Arndt, N. Bacchetta, A. Ball, E. Bartz, W. Bertl, G. M. Bilei, G. Bolla, H. W. K. Cheung, M. Chertok, S. Costa, N. Demaria, D. D. Vazquez, K. Ecklund, W. Erdmann, K. Gill, G. Hall, K. Harder, F. Hartmann, R. Horisberger, W. Johns, H. C. Kaestli, K. Klein, D. Kotlinski, S. Kwan, M. Pesaresi, H. Postema, T. Rohe, C. Schäfer, A. Starodumov, S. Streuli, A. Tricomi, P. Tropea, J. Troska, F. Vasey, and W. Zeuner, *CMS Technical Design Report for the Pixel Detector Upgrade*, Tech. Rep. CERN-LHCC-2012-016. CMS-TDR-11, Sep 2012.

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

[67] R. Frederix and S. Frixione, *Merging meets matching in MC@NLO.* JHEP **12** (2012) 061.

[68] T. Sjöstrand, S. Mrenna, and P. Z. Skands, *A Brief Introduction to PYTHIA 8.1.* Comput. Phys. Commun. **178** (2008) 852.

[69] P. Nason, *A New method for combining NLO QCD with shower Monte Carlo algorithms.* JHEP **11** (2004) 040.

[70] CMS Collaboration, *Particle–flow event reconstruction in CMS and performance for jets, taus, and MET*, CMS Physics Analysis Summary CMS-PAS-PFT-09-001, 2009.

[71] S. Catani, Y. L. Dokshitzer, M. Seymour, and B. Webber, *Longitudinally invariant K_t clustering algorithms for hadron hadron collisions.* Nucl. Phys. B **406** (1993) 187.

[72] M. Cacciari, G. P. Salam, and G. Soyez, *The anti-k_t jet clustering algorithm.* JHEP **04** (2008) 063.

[73] CMS Collaboration, *A Cambridge-Aachen (C-A) based Jet Algorithm for boosted top-jet tagging*, CMS Physics Analysis Summary CMS-PAS-JME-09-001, 2009.

[74] CMS Collaboration, *Particle-flow reconstruction and global event description with the CMS detector.* JINST **12** (2017) P10003.

[75] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness.* JHEP **03** (2011) 015.

[76] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, *Soft Drop.* JHEP **05** (2014) 146.

[77] CMS Collaboration, *Identification of b-Quark Jets with the CMS Experiment.* JINST **8** (2013) P04013.

[78] N. Bartosik et al., *Calibration of the combined secondary vertex b-tagging discriminant using dileptonic ttbar and drell-yan events*, CMS Physics Analysis Note CMS-AN-13-130, 2013.

[79] CMS Collaboration, *Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV.* JINST **12** (2017) P02014.

[80] K. Rehermann and B. Tweedie, *Efficient Identification of Boosted Semileptonic Top Quarks at the LHC*. JHEP **03** (2011) 059.

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

[82] CMS Collaboration, *Measurement of differential* $\mathrm{t\bar{t}}$ *production cross sections using top quarks at large transverse momenta in pp collisions at* $\sqrt{s} = 13$ *TeV*. Submitted to Phys. Rev. D (2020).

[83] W. Verkerke and D. P. Kirkby, *The RooFit toolkit for data modeling*. eConf **C0303241** (2003) MOLT007.

[84] CMS Collaboration, *Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS*. JINST **6** (2011) P11002.

[85] CMS Collaboration, *CMS Luminosity Measurements for the 2016 Data Taking Period*, CMS Physics Analysis Summary CMS-PAS-LUM-17-001, 2017.

[86] R. J. Barlow and C. Beeston, *Fitting using finite Monte Carlo samples*. Comput. Phys. Commun. **77** (1993) 219.

[87] S. Forte, *Parton distributions at the dawn of the LHC*. Acta Phys. Polon. B **41** (2010) 2859.

[88] CMS Collaboration, *Investigations of the impact of the parton shower tuning in Pythia 8 in the modelling of* $\mathrm{t\bar{t}}$ *at* $\sqrt{s} = 8$ *and 13 TeV*, CMS Physics Analysis Summary CMS-PAS-TOP-16-021, 2016.

[89] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*. Eur. Phys. J. C **71** (2011) 1554. [Erratum: Eur.Phys.J.C 73, 2501 (2013)].

[90] G. Cowan, *Statistical data analysis*, Oxford University Press, 1998.

[91] A. Wald, *Tests of statistical hypotheses concerning several parameters when the number of observations is large*. Transactions of the American Mathematical Society **54** (1943) 426.

[92] ATLAS and CMS Collaborations, *Procedure for the LHC Higgs boson search combination in Summer 2011*, Tech. Rep. CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11, CERN, Geneva, Aug 2011.

[93] CMS Collaboration, *Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV*. JINST **13** (2018) P05011.

[94] CMS Collaboration, *Performance of the DeepJet b tagging algorithm using 41.9/fb of data from proton-proton collisions at 13 TeV with Phase 1 CMS detector*, CMS Detector Performance Note CMS-DP-2016-070, 2016.

[95] CMS Collaboration, *Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques.* JINST **15** (2020) P06005.

[96] H. Qu and L. Gouskos, *Jet tagging via particle clouds.* Phys. Rev. D **101** (2020) 056019.

List of Figures

List of Tables