Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Dipartimento di Fisica e Astronomia
Corso di Laurea Magistrale in Fisica

Characterisation of $t\bar{t}H$ production at 13 TeV in the multijet topologies with CMS

Relatore:
Prof. Andrea Castro

Presentata da:
Eric Ballabene

Anno Accademico 2018/2019

II

Abstract

The associated production of a top quark pair with a Higgs boson ($t\bar{t}H$) is very important because, in spite of its small production cross section, it enables the direct measurement of the Yukawa coupling of the Higgs boson to the top quark. The final states of $t\bar{t}H$ events depend on the specific decay of the top quarks and of the Higgs boson, and different topologies are originated, with or without energetic leptons in the final states. The events studied in this work correspond to an all-jet topology, with some of these jets which might be subject to boost given the large transverse momentum involved. It is therefore essential to classify the final states in terms of the number of "resolved" and "boosted" jets, and for each case different triggers are required. Since the background yields (mainly from top quark-antiquark pairs and QCD multijet production) are much larger than the expected signal, special care needs to be used to reduce them. The event selection is based on multivariate analysis algorithms which can distinguish signal from background events, and boosted jets associated to top quarks, to the Higgs bosons or to generic light quarks, enabling the definition of different event categories. The statistical performance of this analysis is characterised by two essential parameters, the upper limits on the signal strength and the signal significance. The expected upper limits on the signal strength are estimated at 95% confidence level, along with the signal significance, for each category and for the combination of all categories.

Sommario

La produzione associata di una coppia di quark top con un bosone di Higgs ($t\bar{t}H$) è molto importante perché, nonostante la sua piccola sezione d'urto di produzione, permette la misura diretta dell'accoppiamento di Yukawa del bosone di Higgs con il quark top. Gli stati finali degli eventi $t\bar{t}H$ dipendono dallo specifico decadimento dei quark top e del bosone di Higgs e differenti topologie possono essere originate, con o senza leptoni energetici negli stati finali. Gli eventi studiati in questo lavoro corrispondono alla topologia *all-jet*, con alcuni di questi jet che possono essere soggetti a *boost* dato l'elevato momento trasverso in gioco. È pertanto essenziale classificare gli stati finali in termini del numero di jet risolti o soggetti a *boost*, e per ciascun caso diversi trigger sono richiesti. Dal momento che il contributo di fondo (principalmente dalla produzione di coppie del quark top e dal fondo QCD *multijet*) è molto più grande del segnale atteso, particolare attenzione deve essere usata per ridurlo. La selezione degli eventi è basata su algoritmi di analisi multivariata che possono distinguere eventi di segnale dal fondo, e jet associati ai quark top, ai bosoni di Higgs o a generici quark leggeri, consentendo la definizione di diverse categorie per gli eventi. Le prestazioni statistiche di questa analisi sono caratterizzate da due parametri essenziali: i limiti superiori sulla *signal strength* e la significanza del segnale. I limiti attesi sulla *signal strength* sono stimati al 95% di livello di confidenza, assieme alla significanza del segnale, per ciascuna categoria e per la combinazione di tutte le categorie.

I would like to express my sincere gratitude to Professor Andrea Castro, for his continuous support throughout the thesis work and helpful advice. I wish to thank my family for the support and encouragement throughout my study.

Contents

Chapter 1

The Standard Model of Particle Physics

The Standard Model (SM) of particle physics is the current description of the fundamental constituents of our universe and interactions between them, developed as a result of a large amount of experimental and theoretical research. The model represents a milestone in the development of the most fundamental theory of matter and outlines the boundaries of the present knowledge of particle physics, beyond which the region of new physics models begin. The aim of the SM has always been represented by providing an unified theoretical description of the three fundamental interactions which are dominant at the particle physics scales, the strong, weak and electromagnetic interactions, the last two being unified in a single Electroweak (EW) interaction.
In this chapter, the general framework of SM is outlined in Section 1.1 and a brief introduction of the elementary particles and their interactions is presented in Sections 1.2 and 1.3. A key concept of the SM is the Spontaneous Symmetry Breaking (SSB) of the EW sector which provides masses to the gauge bosons and matter fermions, reported in Section 1.4. Special attention is reserved to the top quark and Higgs boson, both presented in Section 1.5, and their associated production, presented in Section 1.6.

1.1 General framework

The SM describes the structure of matter as consisting of elementary particles within a spatial scale of $10^{-13} - 10^{-17}$ cm, a scale so small to require a description based on the Quantum Field Theory (QFT) as general framework. The fields and particles are described by the generalized Lagrangian formalism, whose operators are dependent on the space-time point x. The Lagrangian density $\mathcal{L}$ is a functional of the fields $\psi(x)$ and their space-time derivatives $\partial_\mu\psi$, and its exact form is fixed by physical requirements of the local gauge and relativistic invariance, and invariance with respect to groups of internal symmetry. Once the Lagrangian is fixed, the equations of motion are obtained by means of the action principle:

$$\delta S = \delta\Big[\int d^4x \mathcal{L}(\psi, \partial_\mu\psi)\Big] = 0. \tag{1.1}$$

The theory has a gauge symmetry if there is a continuous group of local transformations of the fields (called gauge group) for which the action S remains unmodified. Since each continuous symmetry of $\mathcal{L}$ yields a conserved current and, hence, a conserved charge, the

conservation laws are accounted for by symmetries of the Lagrangian density of the SM under gauge transformations of fields [1, 2, 3, 4].

1.2 Elementary particles

The most fundamental constituents of matter are referred to as elementary particles. They are grouped in two categories according to their spin numbers: fermions, which have half-integer spin, and bosons, which have integer spin.

1.2.1 Elementary fermions

Elementary fermions are further categorized into quarks and leptons.

- There are six quarks that are the constituents of the atomic matter. There are up (u), charm (c), top (t) quarks with electric charge of $+2/3$ and down (d), strange (s), bottom (b) with electric charge $-1/3$. Their charge, mass and spin are reported in Table 1.1 (values are taken from [5]).

Quark	Electric Charge	Mass	Spin
u	$2/3$	$2.16^{+0.49}_{-0.26}$ MeV	$1/2$
d	$-1/3$	$4.67^{+0.48}_{-0.17}$ MeV	$1/2$
c	$2/3$	1.27 ± 0.02 GeV	$1/2$
s	$-1/3$	93^{+11}_{-5} MeV	$1/2$
t	$2/3$	172.9 ± 0.4 GeV	$1/2$
b	$-1/3$	$4.18^{+0.03}_{-0.02}$ GeV	$1/2$

Table 1.1: Relevant physical properties of quarks.

- There are six leptons, three charged and three neutral fermions. Among the charged ones the electron (e) is the well known atomic particle, while the other two are the muon (μ) and the tau (τ) that are heavier counterparts of the electron. The neutral leptons are called neutrinos (ν) and come in three generations ν_e, ν_μ, ν_τ. Their charge, mass and spin are reported in Table 1.2.

Lepton	Electric Charge	Mass	Spin
ν_e	0	< 2.05 eV (95% CL)	$1/2$
e	-1	$0.5109989461 \pm 0.0000000031$ MeV	$1/2$
ν_μ	0	< 0.19 MeV (90% CL)	$1/2$
μ	-1	$105.6583745 \pm 0.0000024$ MeV	$1/2$
ν_τ	0	< 18.2 MeV (95% CL)	$1/2$
τ	-1	1776.86 ± 0.12 MeV	$1/2$

Table 1.2: Relevant physical properties of leptons.

1.2.2 Elementary bosons

Elementary bosons are the force carriers photon (γ), $W^\pm$ and Z bosons, gluons (g) and the mass-giving scalar particle, Higgs (H) boson. Their charge, mass and spin are reported in Table 1.3.

Boson	Electric Charge	Mass	Spin
γ	0	$< 1 \times 10^{-18}$ eV	1
g	0	0	1
W^+	1	80.379 ± 0.012 GeV	1
W^-	-1	80.379 ± 0.012 GeV	1
Z	0	91.1876 ± 0.0021 GeV	1
H	0	125.10 ± 0.14 GeV	0

Table 1.3: Relevant physical properties of bosons.

1.3 Interactions

Once these particles have been introduced, it is interesting to see how they interact with each other, through the electromagnetic, weak and strong forces. The corresponding theoretical parts of the SM are called Quantum Electrodynamics (QED), Quantum Flavordynamics (QFD) and Quantum Chromodynamics (QCD) and drafted in the following.

1.3.1 Electromagnetic interaction

All charged particles interact electromagnetically and the interaction is mediated by the photon as a gauge boson. The photon itself is neutral and does not directly interact with itself. The free Lagrangian $\mathcal{L}_{free}$ of a fermion field ψ, with mass m and charge q, is invariant under U(1) transformations. It can be written as

$$\mathcal{L}_{free} = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi, \tag{1.2}$$

where γ_μ are the Dirac matrices. If we consider a global transformation, the complex fermion field transforms as

$$\psi(x) \to e^{i\alpha}\psi(x), \tag{1.3}$$

where α is a real constant. The same invariance does not hold true under a local U(1) transformation, $\mathrm{U(x)} = e^{i\alpha(x)Q}$, where $\alpha(x)$ is not a constant and it depends on space-time arbitrarily, and Q is the charge operator of the U(1) group. The term that actually breaks the invariance is the derivative of the fermion field, which transforms as

$$\partial_\mu\psi \to e^{i\alpha(x)Q}\partial_\mu\psi + ie^{i\alpha(x)Q}\psi\partial_\mu\alpha. \tag{1.4}$$

In order to have an invariant Lagrangian under this transformation, the derivative must be replaced by the covariant derivative, D_μ, which transforms covariantly like ψ itself and it introduces an addition vector field A_μ to cancel the invariance breaking terms in the above equation

$$D_\mu = \partial_\mu - ieQA_\mu, \tag{1.5}$$

where A_μ transforms:

$$A_\mu \to A_\mu + \frac{1}{e}\partial_\mu\alpha. \tag{1.6}$$

The covariant derivative assures that the free Lagrangian remains invariant under local transformations. This vector field, called the gauge boson A_μ, is the physical photon field. Therefore adding the kinetic energy of the photon field ($F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$), which also needs to be invariant under U(1), leads to the QED Lagrangian

$$\mathcal{L}_{QED} = \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi + eQ\bar{\psi}\gamma^\mu A_\mu\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}. \tag{1.7}$$

A mass term such as $\frac{1}{2}m^2 A_\mu A^\mu$ would not be allowed in the free Lagrangian since it would break the gauge invariance: as a consequence, the gauge field photon is massless.

1.3.2 Strong interaction

All particles with colour charge interact via strong interactions and are mediated by gluons. The strong interactions account for holding the proton and neutron together in an atom, as well for keeping the quarks confined in a hadron. A hadron is a composite particle made of quarks and antiquarks. The gluon itself has a colour charge, which allows for the appearance of self-interactions. None of the coloured particles, quarks and gluons, can be observed as a free particle and they are always confined in colourless states. This phenomenon is called colour confinement. Quarks come in three colours red (r), green (g) and blue (b), antiquarks with anti-colours $\bar{r}$, $\bar{g}$, $\bar{b}$, while gluons carry one unit of colour and one unit of anti-colour. Mesons are colour neutral states formed by quark-antiquark pairs (i.e. $r\bar{r}$) while baryons are groups of three quarks (rgb), antibaryons are groups of three antiquarks ($\bar{r}\bar{g}\bar{b}$). QCD is the theory of the strong interactions between quarks and gluons and describes the $\mathrm{SU(3)_C}$ colour symmetry (where C stands for colour). The free Lagrangian is required to be invariant under the following SU(3) gauge transformation

$$q(x) \to Uq(x) = e^{i\alpha_k(x)T_k} q(x), \tag{1.8}$$

where q is the quark triplet denoting the three colour quark states and U is an arbitrary 3×3 unitary matrix representing the SU(3) transformation. T_k with $k = 1, ..., 8$ are linearly independent traceless matrices and α_k are the group parameters. The local symmetry is restored by introducing the covariant derivative

$$D_\mu = \partial_\mu + ig_s T_k G_\mu^k, \tag{1.9}$$

where g_s is the strong coupling constant, G_μ^k represents the eight gauge fields and transforms in a more complicated way compared to the photon field

$$G_\mu^k \to G_\mu^k - \frac{1}{g}\partial_\mu \alpha_k - f_{klm}\alpha_l G_\mu^m, \tag{1.10}$$

where f_{klm} are the structure constants of the group which is different from the QED case due to the non-abelian structure of the SU(3) group. This leads to the self-interacting gluon terms in the Lagrangian which is also different than the photon field. Adding the gauge invariant kinetic term for each of the gluon fields, the gauge invariant QCD Lagrangian becomes

$$\mathcal{L}_{QCD} = \bar{q}(i\gamma^\mu \partial_\mu - m)q - g(\bar{q}\gamma^\mu T_a q)G_\mu^a - \frac{1}{4}G_{\mu\nu}^a G_a^{\mu\nu}. \tag{1.11}$$

As in the case for U(1) gauge invariance, requiring the Lagrangian to be invariant under colour gauge transformations leads us to 8 self interacting massless gluon fields.

1.3.3 Weak interaction and Electroweak Unification

The weak interactions are mediated by the massive $\mathrm{W^{\pm}}$ and Z vector bosons and they account for the well-known nuclear beta decay. Unlike the electromagnetic and strong interactions, only left-handed fermions and right-handed anti-fermions interact weakly. As a result, the chiral symmetry is broken suggesting that the gauge symmetry of the weak interactions is more complicated compared to the U(1) and SU(3) symmetries. In order to describe the weak interactions of fermions, the electromagnetic and weak interactions

are unified as electroweak interactions. The electroweak interaction is invariant under the $\mathrm{SU(2)_L} \times \mathrm{U(1)_Y}$ weak isospin and hypercharge symmetry, where L stands for left and Y represents the weak hypercharge, defined as $Q = I_3 + \frac{Y}{2}$, with I_3 being the third component of weak isospin. Left-handed fermions are grouped into doublets of weak isospin $I_3 = \pm 1/2$ and right-handed fermions are isospin singlets with $I_3 = 0$. In $\mathrm{SU(2)_L} \times \mathrm{U(1)_Y}$ the left-handed and right-handed fermions transform differently as

$$L \to e^{i\alpha_k(x)\frac{\tau_k}{2} + i\beta(x)\frac{Y}{2}L} \qquad R \to e^{i\beta(x)\frac{Y}{2}R}, \tag{1.12}$$

where $\tau_k/2$ are the generators of weak isospin group $\mathrm{SU(2)_L}$ built from the Pauli matrices τ_k (with $k = 1, 2, 3$), $Y/2$ is the generator of the hypercharge group U(1) and R represents the right-handed fermions. As in U(1) and SU(3) representations, we can introduce the vector fields to ensure the gauge invariance: W_μ^k, with k=1,2,3, is introduced for the $\mathrm{SU(2)_L}$ and a single vector field B_μ for the $\mathrm{U(1)_Y}$. Then the covariant derivative is:

$$D_\mu = \partial_\mu + i\,\frac{g}{2}\,\tau_k\,W_\mu^k + i\,\frac{g'}{2}\,Y\,B_\mu, \tag{1.13}$$

with couplings g and g' for the $\mathrm{SU(2)_L}$ and $\mathrm{U(1)_Y}$ respectively. The $\mathrm{SU(2)_L}$ is a non-abelian group, as SU(3), that the W_k vector fields transform similar to the gauge bosons and the B vector fields. Adding the kinetic terms of the gauge bosons to the free Lagrangian leads us to the electroweak Lagrangian

$$\mathcal{L}_{EWK} = i\bar{L}\gamma^\mu D_\mu L + i\bar{\psi}_R\gamma^\mu D_\mu \psi_R - \frac{1}{4}W_{\mu\nu}^k W^{k,\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}. \tag{1.14}$$

The electroweak Lagrangian is invariant with massless vector bosons, however it is known that the $\mathrm{W^\pm}$ and Z bosons are massive. The masses of the vector bosons need to be added in the theory ensuring the gauge invariance. This happens through the electroweak symmetry breaking mechanism, the so-called the Brout-Englert-Higgs (BEH) mechanism.

1.4 Spontaneous Symmetry Breaking and the BEH mechanism

So far, we have shown that the mass terms of the gauge bosons are not allowed in a gauge invariant theory. As a consequence massive gauge bosons will break the symmetry. In order to allow massive gauge bosons while keeping the Lagrangian invariant under the presented gauge symmetries, we need to introduce the SSB mechanism. The SSB is achieved by adding a scalar field to the Lagrangian, for which the non-zero vacuum expectation values (ground state) break the symmetry. The choice of the field for a SU(2) gauge symmetry is a doublet of complex scalar fields

$$\Phi = \begin{pmatrix} \phi^\dagger \\ \phi \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} \phi^1 + i\phi^2 \\ \phi^3 + i\phi^4 \end{pmatrix}, \tag{1.15}$$

and the SU(2) invariant Lagrangian is

$$\mathcal{L} = (D_\mu\Phi)^\dagger(D^\mu\Phi) - V(\Phi^\dagger\Phi), \tag{1.16}$$

where

$$V(\Phi^\dagger\Phi) = \mu^2\Phi^\dagger\Phi + \lambda(\Phi^\dagger\Phi)^2, \tag{1.17}$$

and the covariant derivative defined in Eq. 1.13. There are two possible forms of this potential depending on the sign of μ^2. If $\mu^2 > 0$, the minimum of the potential can be

set at $< \Phi >= 0$. This represents a system of four scalar particles each with a mass μ interacting with 3 massless gauge bosons W_μ^k and it does not break the symmetry. The most interesting case is $\mu^2 < 0$. The ground state minimum is given by

$$\Phi^\dagger\Phi = \frac{1}{2}(\phi_1^2 + \phi_2^2 + \phi_3^2 + \phi_4^2) = -\frac{\mu^2}{2\lambda}. \tag{1.18}$$

The ground state is associated to a vacuum expectation value $v = \pm\sqrt{\frac{\mu^2}{\lambda}}$. The Lagrangian symmetry is broken by the choice of one of the ground states, it is either $+v$ or $-v$, where the Lagrangian is not symmetric. The field needs to conserve the U(1) symmetry and breaks $\mathrm{SU(2)_L}$. Therefore the field can be fixed to a minimum energy position by choosing $\phi_1 = \phi_2 = \phi_4 = 0$ and $\phi_3^2 = \frac{-\mu^2}{\lambda} = v^2$ and can be parametrised by $h(x)$ which represents the fluctuations of this minimum

$$\Phi = e^{\tau_i\theta^i(x)/v}\frac{1}{\sqrt{2}}\begin{pmatrix}0\\ v + h(x)\end{pmatrix}. \tag{1.19}$$

Here $h(x)$ is the BEH field, $\tau_{1,2,3}$ are the generators of $\mathrm{SU(2)_L}$, and $\theta_{1,2,3}$ are the massless Goldstone bosons. According to the Goldstone theorem, the spontaneously broken symmetry leads to massless scalars as many as the broken generators. The $\mathrm{SU(2)_L}$ symmetry allows to rotate away any dependence on $\theta_i(x)$. Choosing the unitarity gauge $\theta_i(x) = 0$, eliminates the θ_i fields in the Lagrangian so that Goldstone bosons are absorbed by the three gauge bosons that require masses and give the longitudinal components to the massive gauge bosons. The BEH potential of the SSB Lagrangian takes the following form

$$V(\Phi^\dagger\Phi) = \frac{1}{2}\,(2\lambda v^2)\,h(x)^2 + \lambda\, v\, h(x)^3 + \frac{\lambda}{4}\,h(x)^4 - \frac{\lambda}{4}\,v^4. \tag{1.20}$$

The BEH potential has quadratic, cubic and quartic terms of the BEH field. The first term is the mass term of the BEH field

$$m_H = \sqrt{2\lambda}\; v = \sqrt{2}\;|\mu| \tag{1.21}$$

and it depends on the self BEH coupling λ and the v. The cubic and quartic terms correspond to self-interactions of the BEH field, and the last term is a constant. Inserting the new scalar field with the covariant derivative, it becomes

$$(D^\mu\phi)^\dagger(D_\mu\phi) = \frac{1}{2}\,|\partial_\mu h(x)|^2 + \frac{1}{8}\,v^2\,[g^2(W_1^2 + W_2^2) + (gW_3 - g'B_\mu)^2] + \mathcal{O}(h(x)). \tag{1.22}$$

Here the first term is the kinetic term of the BEH field while the last term has the interactions of the BEH field with the gauge boson. We will focus on the second term in the Lagrangian which gives the mass terms of the gauge bosons. We can rewrite this term of the Lagrangian in terms of the known $\mathrm{W^\pm}$, Z and A bosons as

$$\frac{1}{8}\,v^2\,[g^2(W_\mu^+)^2 + g^2(W_\mu^-)^2 + (g^2 + g'^2)\,Z_\mu^2 + 0\cdot A_\mu^2], \tag{1.23}$$

where

$$\begin{aligned}
W_\mu^\pm &= \frac{1}{\sqrt{2}}(W_1 \pm iW_2), && \text{with } M_{W^\pm} = \frac{1}{2}vg\\
Z_\mu &= \frac{1}{\sqrt{g^2 + g'^2}}(gW_3 - g'B_\mu), && \text{with } M_Z = \frac{1}{2}v\sqrt{(g^2 + g'^2)}\\
A_\mu &= \frac{1}{\sqrt{g^2 + g'^2}}(g'W_3 + gB_\mu), && \text{with } M_A = 0
\end{aligned} \tag{1.24}$$

Also, the ratio of the W and Z boson masses is equal to the cosine of the weak mixing angle θ_W

$$M_W/M_Z = g/\sqrt{g^2+g'^2} = \cos(\theta_W). \tag{1.25}$$

The weak mixing angle is a parameter of the SM that rotates the W_3, B_μ vector boson plane producing the Z and A_μ bosons by SSB. Additionally it relates the couplings as

$$e = g\sin(\theta_W) = g'\cos(\theta_W). \tag{1.26}$$

The experimental measurements of M_W, M_Z and θ_W confirm the above relation, which is typically written in the form

$$\sin^2(\theta_W) = 1 - \cos^2(\theta_W) = 1 - \frac{M_W^2}{M_Z^2}. \tag{1.27}$$

The components containing fermion fields can be also expressed in terms of the angle θ_W and the fields $W_\mu^\pm$, Z_μ and A_μ, leading to the neutral-current Lagrangian $\mathcal{L}_{NC}$ and charged-current Lagrangian $\mathcal{L}_{CC}$

$$\mathcal{L}_{NC} = eJ_\mu^A A^\mu + \frac{g}{\cos(\theta_W)} J_\mu^Z Z^\mu, \tag{1.28}$$

$$\mathcal{L}_{CC} = \frac{g}{\sqrt{2}}\,(J_\mu^+ W^{+\,\mu} - J_\mu^- W^{-\,\mu}), \tag{1.29}$$

where the currents J_μ are given by

$$\begin{aligned} J_\mu^A &= Q_f\,\bar\psi\,\gamma_\mu\,\psi, \\ J_\mu^Z &= \frac{1}{2}\,\bar\psi\,\gamma_\mu[(T_f^3 - 2Q_f\sin^2(\theta_W)) - \gamma_5(T_f^3)]\psi, \\ J_\mu^+ &= \frac{1}{2}\,\bar u\,\gamma_\mu(1-\gamma_5)\,d, \end{aligned} \tag{1.30}$$

with u and d representing the up and down-type fermions, while ψ refers to either of them, and Q_f is the electric charge of the fermion.
So far, using the gauge invariance of the theory, we showed how the W and Z bosons gain their mass while the photon remains massless with the addition of the BEH field. But we still need to discuss how fermions acquire their mass. We have shown how the fermion fields transform under $\mathrm{SU(2)_L \times U(1)_Y}$ rotations. The mass terms of the fermions are not allowed since left-handed fermions form an isospin doublet and right-handed fermions form isospin singlets and terms like $m[\bar\psi_L\psi_R + \bar\psi_R\psi_L]$ are not gauge invariant. Therefore a singlet term of $\mathrm{SU(2)_L \times U(1)_Y}$ is needed for an invariant Lagrangian mass term. This can be done by introducing the BEH doublet into the Lagrangian,

$$\mathcal{L}_{fermions} = \lambda_f[\bar\psi_L\phi\psi_R + \bar\psi_R\phi\psi_L], \tag{1.31}$$

for the electron this term becomes

$$-\frac{\lambda_e(v+h)}{\sqrt{2}}[\bar e_L e_R + \bar e_R e_L] = -\frac{\lambda_e v}{\sqrt{2}}\bar e e - \frac{\lambda_e}{\sqrt{2}} h\bar e e \tag{1.32}$$

where e_L, e_R refer to the left- and right-handed electrons. The first term gives the mass term of the electron, $\lambda_e v/\sqrt{2}$ and the second term describes the interactions between the BEH field and the fermions. The λ parameter is very important and describes the Yukawa coupling of a fermion to the BEH field and it is expressed as

$$\lambda_f = \sqrt{2}\,\frac{m_f}{v} \tag{1.33}$$

which is proportional to the mass of the fermion and $v \simeq 246$ GeV. This mass term only gives mass to "down" type of leptons, while keeps the "up" type of leptons, neutrinos, massless. In fact, despite experimental evidence for neutrino oscillations, which implies non-zero neutrino masses, the SM does not predict non-zero masses for neutrinos in a natural way. In order to have mass terms for the up type quarks, an additional term is needed in the Lagrangian. This is done by introducing the charge-conjugate representation of the BEH doublet, which under SU(2) rotations transforms as the original BEH field

$$\tilde{\phi}^C = -i\tau_2\phi^\star = \sqrt{\frac{1}{2}}\begin{pmatrix} v+h \\ 0 \end{pmatrix}, \tag{1.34}$$

where $\tilde{\phi}^C$ is the charge conjugate representation of the BEH doublet. The mass term of the up-type fermion becomes

$$\mathcal{L}_{up} = \lambda_q[\bar{u}_L\tilde{\phi}^C u_R + \bar{u}_R\tilde{\phi}^C u_L] \tag{1.35}$$

where u represents the up-type fermions. This mass term has the same form as the down-type fermions with the corresponding Yukawa couplings. All the mass terms of the SM particles can be expressed in terms of the vacuum expectation value v and the coupling constants: g, g', λ_i where the Yukawa couplings, λ_i, are different for each lepton and quark, and zero for neutrinos in the SM. Finally, we can gather all the ingredients of the SM, $\mathrm{SU(3)} \times \mathrm{SU(2)_L} \times \mathrm{U(1)_Y}$ and summarize all the interaction and mass terms in the Lagrangian

$$\begin{aligned}
\mathcal{L}_{SM} = \quad & \underbrace{-\frac{1}{4}W_{\mu\nu}W^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G_{\mu\nu}G^{\mu\nu}}_{\mathrm{W^\pm,\,Z,\,\gamma\ and\ gluon\ kinetic\ energies\ and\ self-interactions}} \quad + \\
& + \underbrace{\ell_L\gamma^\mu(i\partial_\mu - g\frac{1}{2}\tau^i W^i_\mu - g'\frac{Y}{2}B_\mu)\ell_L + q_L\gamma^\mu(i\partial_\mu - g\frac{1}{2}\tau^i W^i_\mu - g'\frac{Y}{2}B_\mu - g_s T^k G^k_\mu)q_L}_{\mathrm{left-handed\ fermion\ kinetic\ energies\ and\ their\ interactions}} + \\
& + \underbrace{\ell_R\gamma^\mu(i\partial_\mu - g'\frac{Y}{2}B_\mu)\ell_R + q_R\gamma^\mu(i\partial_\mu - g'\frac{Y}{2}B_\mu - g_s T^k G^k_\mu)q_R}_{\mathrm{right-handed\ fermion\ kinetic\ energies\ and\ their\ interactions}} + \\
& + \underbrace{|(i\partial_\mu - g\frac{1}{2}\tau^i W^i_\mu - g'\frac{Y}{2}B_\mu)\phi|^2 - V(\phi^\dagger\phi)}_{\mathrm{W^\pm,Z,\gamma\ and\ BEH\ masses\ and\ coupling}} - \underbrace{(\lambda_f\bar{\ell}_L\phi\ell_R + \lambda_g\bar{q}_L\tilde{\phi}^C q_R + \mathrm{h.c.})}_{\mathrm{fermion\ masses\ and\ couplings\ to\ BEH}},
\end{aligned} \tag{1.36}$$

where ℓ is used for leptons and q for quarks.

1.5 Top quark and Higgs boson physics

The top quark and the Higgs boson are among the most recently discovered SM particles. Due to their large mass and their distinctive properties, they are of special interest to a large fraction of particle-physics analyses performed today. Most of the properties of the top quark and the Higgs boson are well known by now. An overview of them, together with the production and decay modes, is presented in the following. With the discovery of the Higgs boson, the last missing piece of the SM has been found.

1.5.1 Top quark

The top quark is the up-type quark of the third generation of elementary particles, with a mass of approximately 173 GeV. The top quark, together with its antiparticle, the

antitop quark, has been discovered in 1995 by the CDF and D0 experiments at the proton-antiproton collider Tevatron at the Fermilab laboratory [11, 12]. With its high mass, the top quark is particularly interesting for searches for Beyond the Standard Model (BSM) physics and precision measurements of its properties play a significant role at the ongoing physics schedule at the Large Hadron Collider (LHC) at CERN. Moreover, the top quark has other unique and special properties, such as the large value of its width of about 1.35 GeV [5] that causes it to have a very short lifetime of about 5.0×10^{-25} s. This implies that the top quark decays before any hadronization can occur. This allows us to determine spin information transferred to its decay products undiluted by non-perturbative effects.

Top quark production

The top quark can be produced in pairs through the strong interaction or singly through the weak interaction. The top quark-antiquark pair production ($t\bar{t}$) is a pure QCD process and can be initiated in two different ways, either by gluons or by a quark-antiquark pair in the initial state. Both types of $t\bar{t}$ production are illustrated by Feynman diagrams in Fig. 1.1. At the LHC with a centre-of-mass energy of $\sqrt{s} = 13$ TeV, about 90% of the

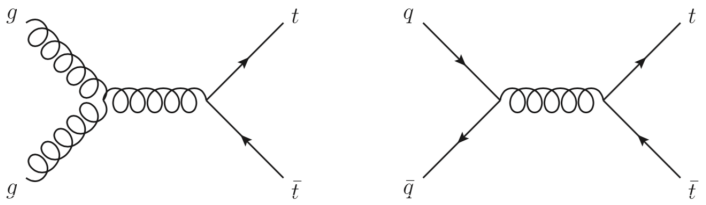

Figure 1.1: Leading order Feynman diagrams for the top quark-antiquark pair production.

top quarks are produced via the gluon-initiated process. Although the top quark pair production requires enough energy to produce two top quarks, it represents the main production mode at the LHC. This fact can be explained by the large coupling constant of the strong interaction.

Single top quarks are produced via the weak interaction. This process includes a vertex of a top quark, a W boson, and a down-type quark. The contribution of different down-type quarks to this vertex is determined by the corresponding Cabibbo-Kobayashi-Maskawa (CKM) matrix element. As the CKM matrix element V_{tb} is close to one and the others negligibly small, the vertex includes a bottom quark in almost all cases. Correspondingly, the single top quark production is well suited for the measurement of the CKM matrix element V_{tb}. The single top quark production is further subdivided into three production modes: the t-channel, the associated production with a W boson (tW) and the s-channel. Feynman diagrams of the single top quark production are given in Fig. 1.2 ordered by their cross section at the LHC. Single top quark production features a cross section that is about five times smaller than top quark pair production at a centre-of-mass energy of $\sqrt{s} = 13$ TeV.

Top quark decay

The top quark decays only through the weak interaction. It decays into a W boson and a bottom quark in almost all of the cases, because of the large CKM matrix element V_{tb}. One distinguishes between the leptonic and the hadronic decay of a top quark, which is characterised by the decay of the W boson. A leptonic decay of a top quark features a

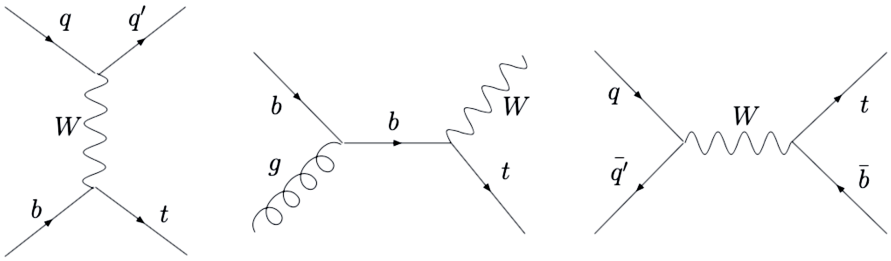

Figure 1.2: Leading order Feynman diagrams for single top quark production. From left to right: t-channel, tW, s-channel.

W-boson decay into a charged lepton and a neutrino. A hadronic decay of a top quark is indicated by a W boson decay into an up-type and a down-type quark and antiquark. A decay of the W boson into a final state featuring a top quark is not possible due to the large mass of the top quark. Accordingly, the hadronic W boson decay produces mainly quarks from the first and the second generation. Taking into account the three different colour charges of quarks, the branching ratio for the hadronic top quark decay occurs twice as often as the leptonic decay. Transferring this categorisation to the decay of a top quark pair provides three different configurations:

- Dileptonic $t\bar{t}$ decay channel: both top quarks decay leptonically. The dileptonic decay channel features a branching ratio of 10.5%.
- Semileptonic $t\bar{t}$ decay channel: one top quark decays leptonically, while the other top quark decays hadronically. The semileptonic decay channel features a branching ratio of 43.8%.
- All-hadronic $t\bar{t}$ decay channel: both top quarks decay hadronically. The all-hadronic decay channel features a branching ratio of 45.7%.

1.5.2 Higgs boson

The Higgs boson is a spin-zero particle resulting from the Higgs mechanism with a mass of approximately 125 GeV. Until its discovery in 2012 by ATLAS and CMS [7, 8], the Higgs boson has been the last missing piece of the SM. The discovery of a new resonance with and the subsequent studies of its properties have provided the first portrait of the BEH mechanism. The Higgs boson mass has been precisely measured and its production and decay rates are found to be consistent, within errors, with the SM predictions. In the following, a brief description of the Higgs boson couplings, production and decay modes is presented, followed by an overview of its discovery.

Higgs boson couplings

The Higgs boson couplings to the fundamental particles are determined by their masses: very weak for light particles, such as light quarks and electrons, but strong for heavy particles such as the W and Z bosons and the top quark. More precisely, the Higgs boson couplings to fermions and gauge bosons, as well as the Higgs boson self coupling, are summarized in the following Lagrangian:

$$\mathcal{L}_H = -g_{Hf\bar{f}}\bar{f}fH + \frac{g_{HHH}}{6}H^3 + \frac{g_{HHHH}}{24}H^4 + \delta_V V_\mu V^\mu\Big(g_{HVV}H + \frac{g_{HHVV}}{2}H^2\Big) \quad (1.37)$$

with

$$g_{Hf\bar{f}} = \frac{m_f}{v}, \quad g_{HVV} = \frac{2m_V^2}{v}, \quad g_{HHVV} = \frac{2m_V^2}{v^2}, \quad g_{HHH} = \frac{3m_H^2}{v}, \quad g_{HHHH} = \frac{3m_H^2}{v^2}. \tag{1.38}$$

It can be seen that the SM Higgs couplings to fundamental fermions are linearly proportional to the fermion masses, whereas the couplings to bosons are proportional to the square of the boson masses. It is also possible for the Higgs boson to self-interact.

Higgs boson production

The Higgs boson can be produced in many ways at the LHC (Fig. 1.3). The main

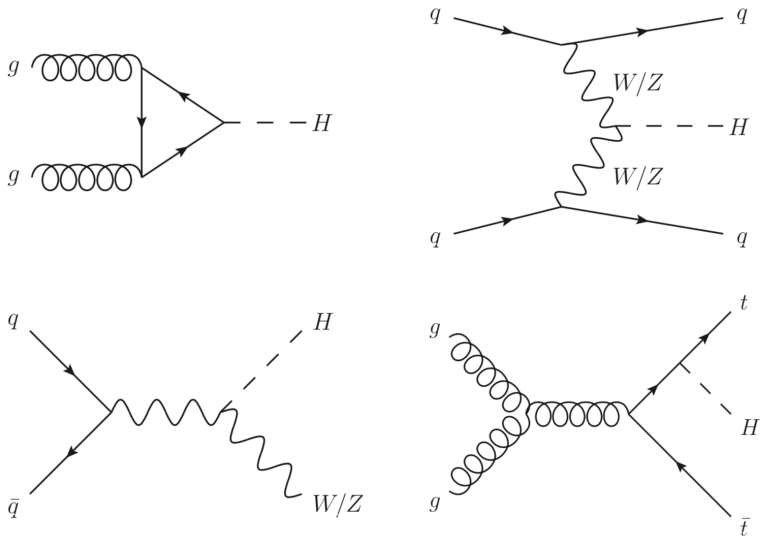

Figure 1.3: Leading order Feynman diagrams for Higgs boson production. Gluon-gluon fusion (upper left), VBF (upper right), VH (lower left), $t\bar{t}H$ (lower right).

production mode at the LHC is the gluon-gluon fusion (ggH), featuring gluons in the initial state. As the Higgs boson does not couple to massless particles, the top quark is produced via intermediately generated particles in this process. Further, the intermediate particles can be only quarks and not gluons which only couple to colour charged particles. The largest contribution is provided by the top quark, due to its large mass and the resulting large coupling to the Higgs boson, as discussed. The main reason for the comparably large cross section is the large number of gluons in a proton-proton collisions with an energy sufficient to enter this process. Gluons carry a large fraction of proton momentum, as it is known from the parton distribution functions which describe the probability distributions for a parton carrying a particular momentum. The second-largest Higgs boson production mode is vector-boson fusion (VBF). This process starts with two quarks in the initial state, which produce virtual vector bosons. The vector bosons in turn produce a Higgs boson. The comparably large cross section can be explained by the large coupling of the Higgs boson to the vector bosons. A special characteristic of this process is the two outgoing quarks. The two quarks form two jets, which are directed in the forward direction of the detector. This special trait simplifies a targeted search for VBF. A further production mode is the associated production of a Higgs boson with a vector boson (VH). This process is also known as Higgs-strahlung, which refers to bremsstrahlung as analogue

process. In the VH process, a vector boson is produced by the annihilation of a quark and an antiquark. The Higgs boson is radiated by the vector boson. VH production is the Higgs boson production mode with the third-largest cross section among all SM Higgs boson production modes. The associated production of a Higgs boson with a top quark pair ($t\bar{t}H$) has the smallest cross section among the four main Higgs boson production modes. In this process, a top quark pair is produced as described in the previous section. The Higgs boson is radiated from one of the top quarks. Even though the coupling of the Higgs boson to the top quark is comparably strong, $t\bar{t}H$ production features a very small cross section. This is mainly due to the enormous amount of energy of about 500 GeV necessary to produce these three massive particles.

Higgs boson decay

The Higgs boson can decay in different channels governed by branching ratios, which are reported in Table 1.4 (values from [5]). The masses of the Higgs boson and of the decay products are the main factors determining the branching ratios. A decay into top quarks, which would be favoured due to the coupling, is not possible as the mass of two top quarks largely exceeds the mass of the Higgs boson. Instead, the largest branching ratio is provided by the Higgs boson decay into two bottom quarks. This decay makes up almost 60% of all Higgs boson decays. However, a search for Higgs bosons decaying into a bottom quark pair at the LHC is challenging due to the large background from QCD processes. The second largest contribution with a branching ratio of about 20% is given by the Higgs boson decay into two W bosons, where one W boson is produced off-shell. In case of the W bosons decaying into leptons, this decay provides a very clean signature. One of the search channels mainly contributing to the Higgs boson discovery in 2012 is based on the Higgs boson decay into two Z bosons. If the Z bosons decay into charged leptons, this decay channel provides a very distinctive signature as there are hardly any backgrounds featuring four charged leptons. Due to the good momentum resolution of charged leptons, a very narrow Higgs boson mass peak can be reconstructed in this search channel. Again, one of the bosons is produced off the mass shell as the invariant mass of two Z bosons exceeds the Higgs boson mass. The second Higgs boson decay mode with a major contribution to the Higgs boson discovery is the decay into two photons. As for the gluons in the Higgs boson production by gluon fusion, the massless photons do not couple to the Higgs boson directly. Instead, this decay proceeds via a loop. Compared to gluon fusion, all electrically charged massive particles may contribute to the loop. Accordingly, a further major contribution is given by the W boson. The Higgs boson decay into two photons also features a very clean final state with a very good Higgs boson mass resolution. However, this decay channel has a very small branching ratio compared to the other Higgs boson decay modes described.

Decay channel	Branching ratio	Rel. uncertainty
$H \to b\bar{b}$	58.4%	$^{+3.2\%}_{-3.3\%}$
$H \to W^+ W^-$	21.4%	$^{+4.3\%}_{-4.2\%}$
$H \to \tau^+\tau^-$	6.3%	$^{+5.7\%}_{-5.7\%}$
$H \to Z Z^*$	2.62%	$^{+4.3\%}_{-4.1\%}$
$H \to \gamma\gamma$	0.23%	$^{+5.0\%}_{-4.9\%}$
$H \to Z\gamma$	0.15%	$^{+9.0\%}_{-8.9\%}$
$H \to \mu\mu$	0.02%	$^{+6.0\%}_{-5.9\%}$

Table 1.4: Decay channels and branching ratios for a SM Higgs boson with $m_H = 125$ GeV.

In the SM, the Higgs boson width is very precisely predicted once the Higgs boson mass is known. For a mass of 125 GeV, the Higgs boson has a very narrow width of 4.2 MeV. The total width is dominated by the fermionic decays at approximately 75%, while the vector boson modes are suppressed and contribute 25% only. Explicitly, the partial widths are given by the relations

$$\Gamma(H \to f\bar{f}) = \frac{G_F m_f^2 m_H N_c}{4\pi\sqrt{2}} \Big(1 - 4m_f^2/m_H^2\Big)^{3/2} \tag{1.39}$$

$$\Gamma(H \to W^+W^-) = \frac{G_F m_H^3 \beta_W}{32\pi\sqrt{2}} \Big(4 - 4a_W + 3a_W^2\Big) \tag{1.40}$$

$$\Gamma(H \to ZZ) = \frac{G_F m_H^3 \beta_Z}{64\pi\sqrt{2}} \Big(4 - 4a_Z + 3a_Z^2\Big) \tag{1.41}$$

where N_c is 3 for quarks and 1 for leptons and where $a_W = 1 - \beta_W^2 = 4m_W^2/m_H^2$ and $a_Z = 1 - \beta_Z^2 = 4m_Z^2/m_H^2$. The decay to two gluons proceeds through quarks loops and the partial width is given by the relation

$$\Gamma(H \to gg) = \frac{\alpha_S^2 G_F m_H^3}{36\pi^3\sqrt{2}} \Big| \sum_q I(m_q^2/m_H^2) \Big|^2 \tag{1.42}$$

where $I(z)$ is complex for $z < 1/4$. For $z < 2 \times 10^{-3}$, $I(z)$ is small so the light quarks contribute negligibly. For $m_H < 2m_t$, $z > 1/4$ and

$$I(z) = 3\Big[2z + 2z(1-4z)\Big(\sin^{-1}\frac{1}{2\sqrt{z}}\Big)^2\Big] \tag{1.43}$$

which has the limit $I(z) \to 1$ as $z \to \infty$.

Higgs boson observation

In 2012, the Higgs boson has been independently observed by the ATLAS and CMS collaborations [7, 8]. The announcement on July 4, 2012, of the observation at the LHC of a narrow resonance with a mass of about 125 GeV was an important landmark in the decades-long direct search [13, 14] for the SM Higgs boson. This was followed by a detailed exploration of properties of the Higgs boson at the different runs of the LHC at $\sqrt{s} = 8$ and 13 TeV. For this discovery, various searches targeting different Higgs boson decay channels have been combined. For a given value of the Higgs boson mass m_H, the sensitivity of a search channel depends on the production cross section of the Higgs boson, its decay branching fraction, reconstructed mass resolution, selection efficiency and the level of background in the final state. For a low-mass Higgs boson (110 GeV $< m_H <$ 150 GeV) where the natural width is only a few MeV, five decay channels play an important role at the LHC. In the $\mathrm{H} \to \gamma\gamma$ and $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ channels, all final state particles can be very precisely measured and the reconstructed m_H resolution is excellent (typically 1−2%). While the $\mathrm{H} \to \mathrm{W^+W^-} \to \ell^+\nu\ell^-\bar{\nu}$ channel has relatively large branching fraction, the m_H resolution is poor (approximately 20%) due to the presence of neutrinos. The $\mathrm{H} \to \mathrm{b\bar{b}}$ and the $\mathrm{H} \to \tau^+\tau^-$ channels suffer from large backgrounds and an intermediate mass resolution of about 10% and 15% respectively. For $m_H > 150$ GeV, the sensitive search channels were $\mathrm{H} \to \mathrm{WW}$ and $\mathrm{H} \to \mathrm{ZZ}$ where the W or Z boson decays into a variety of leptonic and hadronic final states. The candidate events in each Higgs boson decay channel are split into several mutually exclusive categories based on the specific topological, kinematic or other features present in the event. The categorisation of events

increases the sensitivity of the overall analysis and allows a separation of different Higgs boson production processes. In the following, a brief summary of the observation of Higgs boson decay into a $\gamma\gamma$ pair or into a ZZ* is presented.

In the $H \to \gamma\gamma$ channel, a search is performed for a narrow peak over a smoothly falling background in the invariant mass distribution of two high-p_T photons. The background in this channel is conspicuous and stems from prompt diphoton processes for the irreducible backgrounds, and the γ+jet and dijet processes for the reducible backgrounds where one jet fragments typically into a leading π^0. In order to optimise search sensitivity and also to separate the various Higgs production modes, ATLAS and CMS experiments split events into several mutually exclusive categories. Diphoton events containing a high-p_T muon or electron, or missing transverse energy (E_{miss}) consistent with the decay of a W or Z boson are tagged in the VH production category. Diphoton events containing energetic dijets with a large mass and pseudorapidity difference are assigned to the VBF production category, and the remaining events are considered either in the VH category when the two jets are compatible with the hadronic decay of a W or a Z, or in the gluon-fusion production category. While the leptonic VH category is relatively pure, the VBF category has significant contamination from the gluon fusion process. Events which are not picked by any of the above selections are further categorised according to their expected $m_{\gamma\gamma}$ resolution and signal-over-background ratio. Categories with good m_H resolution and larger signal-over-background ratio contribute most to the sensitivity of the search. The $m_{\gamma\gamma}$ distribution after combining all categories is shown for the ATLAS experiment in Fig. 1.4 (left) using Run 2 data. The signal strength $\mu = (\sigma \cdot \mathrm{BR})_{\mathrm{obs}}/(\sigma \cdot \mathrm{BR})_{\mathrm{SM}}$ in the diphoton decay of the Higgs boson is 1.17 ± 0.27 for ATLAS in Run 1 [15] and 0.99 ± 0.14 [16] in Run 2. The signal strengths measured in Run 1 and Run 2 by the CMS collaboration are $0.78^{+0.26}_{-0.23}$ [17] and $1.16^{+0.15}_{-0.14}$ [18] respectively.

In the $H \to ZZ^* \to 4\ell$ channel, a search is performed for a narrow mass peak over a small

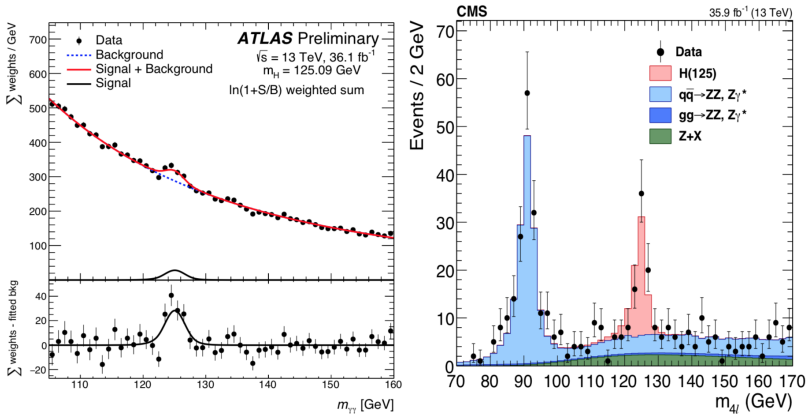

Figure 1.4: (Left) The invariant mass distribution of diphoton candidates, with each event weighted by the signal-over-background ratio in each event category, observed by ATLAS at Run 2. The residuals of the data with respect to the fitted background are displayed in the lower panel. (Right) The $m_{4\ell}$ distribution from CMS Run 2 data.

continuous background dominated by non-resonant ZZ* production from $q\bar{q}$ annihilation and gg fusion processes. The contribution and the shape of this irreducible background are taken from simulation. The subdominant and reducible backgrounds are derived from data. To help distinguish the Higgs signal from the dominant non-resonant ZZ* background, both ATLAS and CMS use a matrix element likelihood approach to construct

a kinematic discriminant built for each 4ℓ event. To further enhance the sensitivity of a signal, various techniques based on the matrix element or the multivariate analysis are used by the experiments. Since the $m_{4\ell}$ resolutions and the reducible background levels are different in the 4μ, 4e and 2e2μ subchannels, they are analysed separately and the results are then combined. The distribution of the reconstructed invariant mass of the four leptons for the CMS experiment is given in Fig. 1.4 (right), showing a clear excess at a mass of approximately $m_H = 125$ GeV. Both experiments also observe a clear peak at $m_{4\ell}$ = 91 GeV from the production of a Z boson on-mass-shell and decaying to four leptons due typically to the emission of an off-shell photon from one of the primary leptons from the Z boson decay. The signal strengths μ for the inclusive $H \to 4\ell$ production measured by the ATLAS and CMS experiments are $1.44^{+0.40}_{-0.33}$ [19] at $m_H = 125.36$ GeV and $0.93^{+0.29}_{-0.25}$ [20] at $m_H = 125.6$ GeV respectively, in Run 1. The signal strengths measured by the ATLAS and CMS experiments in Run 2 are $1.28^{+0.21}_{-0.19}$ [21] and $1.05^{+0.19}_{-0.25}$ [22] respectively, both measurements are made at the combined Run 1 Higgs mass of $m_H = 125.09$ GeV.

1.6 Higgs boson production in association with a top quark pair

The Higgs boson produced in association with a top quark-antiquark pair ($t\bar{t}H$) is a very interesting channel and represents the subject of this thesis. A special characteristic of $t\bar{t}H$ production is to give direct access to the coupling of the Higgs boson to the top quark, the so-called top quark Yukawa coupling. However, the experimental observation of this process is complicated due to its small cross section and an overwhelming amount of background. Substantial indirect evidence of this coupling is provided by the compatibility of observed rates of the Higgs boson produced through gluon fusion involving a top quark loop in the principal discovery decay channels. Direct evidence of this coupling at the LHC is available through $t\bar{t}H$ production which allows a clean measurement of the top quark with the Higgs boson coupling. In fact, according to the SM, the masses of elementary fermions are accounted for by introducing a minimal set of Yukawa interactions, compatible with gauge invariance, between the Higgs and fermion fields. Following the spontaneous breaking of electroweak symmetry, charged fermions of flavour f couple to H with a strength proportional to the mass of those fermions m_f. Measurements of the Higgs boson decay rates to down-type fermions (τ leptons and bottom quarks) agree with the SM predictions within their uncertainties. However, the top quark Yukawa coupling cannot be similarly tested from the measurement of a decay rate since on-shell top quarks are too heavy to be produced in Higgs boson decay. Instead, constraints on the coupling can be obtained through measurement of the $t\bar{t}H$ production process. The Feynman diagram for $t\bar{t}H$ production is represented in Fig. 1.5, with the hadronic decay of the two top quarks and the Higgs boson.

1.6.1 Theoretical motivations for measuring $t\bar{t}H$ production

There are several motivations for studying the $t\bar{t}H$ production:

- The Higgs boson production mode in association with top quarks provides access to a direct measurement of the top quark Yukawa coupling λ_t to the Higgs boson. Precise measurement of the Yukawa couplings of the Higgs boson to fermions λ_f, in general, remains a very important goal of the LHC, with the Yukawa interaction predicted to be the source of fermion masses. Any deviations found between measurements of λ_f and the expected values extracted using the fermion masses

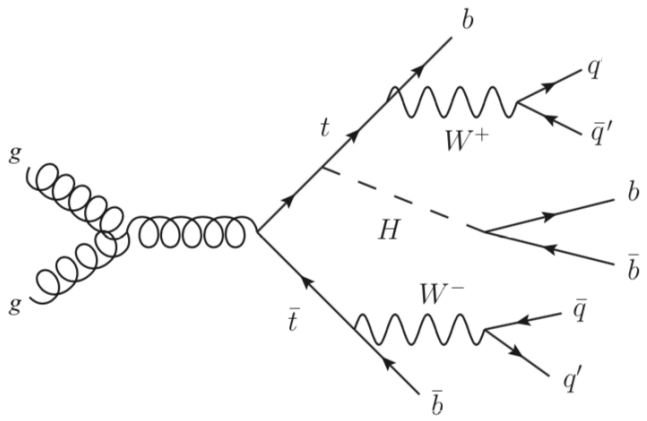

Figure 1.5: Feynman diagram for $t\bar{t}H$ production in the fully hadronic decay.

$m_f = \lambda_f \frac{v}{2}$ would be strong evidence for new physics. The top quark plays a key role, being the heaviest SM particle whose predicted value of $\lambda_t \simeq 1$, with the latest experimental measurement of $1.07^{+0.34}_{-0.43}$ [23] with an upper limit of 1.67 at the 95% confidence level in good agreement with the SM prediction. In comparison to the couplings of the Higgs boson to other fermions, it is almost two orders of magnitude higher than the next largest coupling, λ_b.
Measurements of λ_t can be extracted from processes involving loop effects, such as the gluon-fusion production. However, these channels only provide an indirect measurement where the top quark mediates the interactions in the loops and assumes no BSM effects. Instead, the top-Higgs vertex present in $t\bar{t}H$ production provides a direct measurement of λ_t, significantly reducing the model dependence. A direct measurement of λ_t helps to constrain BSM searches and represents a precision test of the SM consistency. Measurements from direct and indirect searches can also be compared, which would probe the presence of BSM particles mediating loops in indirect processes.

- The top quark Yukawa coupling also provides a window into the scale of new physics. The effective potential of the Higgs field is extremely sensitive to λ_t. Small changes in λ_t can modify the effective potential from a monotonic behaviour which appears as an extra minimum at very large values of the Higgs field [24].
In the absence of BSM signals the only way to address the question of the scale of new physics is to define the energy where the SM becomes theoretically inconsistent or contradicts some observations. Since the SM is a renormalisable quantum field theory, the problems can appear because of the renormalisation evolution of some coupling constants, i.e. when they become large (and the model enters strong coupling at that scale), or additional minima of the effective potential develop changing the vacuum structure. The most dangerous constant turns out to be the Higgs boson self-coupling constant λ with the renormalisation group (RG) evolution at one loop.

$$16\pi \frac{\mathrm{d}\lambda}{\mathrm{d}\ln\mu} = 24\lambda^2 + 12\lambda\lambda_t^2 - 9\lambda(g^2 + \frac{1}{3}g'^2) - 6\lambda_t^4 + \frac{9}{8}g^4 + \frac{3}{8}g'^4 + \frac{3}{4}g^2g'^2 \quad (1.44)$$

The right-hand side depends on the interplay between the positive contributions of the bosons and negative contribution from the top quark. The contribution of

the top quark to the effective potential is very important, as it has the largest Yukawa coupling to the Higgs boson. Moreover, it comes with the minus sign and is responsible for the appearance of the extra minimum of the effective potential at large values of the Higgs field. In general, λ_t should not exceed the critical value λ_{crit}, coinciding with good precision with the requirement of the stability of the electroweak vacuum. To find the numerical value of λ_{crit}, one should compute the effective potential for the Higgs field $V(\phi)$ and determine the parameters at which it has two degenerate minima:

$$V(\phi_{SM}) = V(\phi_1) \quad V'(\phi_{SM}) = V'(\phi_1) = 0 \tag{1.45}$$

The renormalization group effective potential has the form:

$$V(\phi) \propto \lambda(\phi)\phi^4\Big[1 + \mathcal{O}\Big(\frac{\alpha}{4\pi}\log(M_i/M_j)\Big)\Big] \tag{1.46}$$

where α is the common name for the SM coupling constants, and $M_{i,j}$ are the masses of different particles in the background of the Higgs field. For $\lambda_t < \lambda_{\text{crit}} - 1.2 \times 10^{-6}$ the effective potential increases while the Higgs field increases, for $\lambda_t > \lambda_{\text{crit}} - 1.2 \times 10^{-6}$ a new minimum of the effective potential develops at large values of the Higgs field, at $\lambda_t = \lambda_{\text{crit}}$ our electroweak vacuum is degenerate with the new one, while at $\lambda_t > \lambda_{\text{crit}}$ the new minimum is deeper than ours, meaning that our vacuum is metastable. If $\lambda_t > \lambda_{\text{crit}} + 0.04$ the life-time of our vacuum is smaller than the age of the Universe. The case $\lambda_t < \lambda_{\text{crit}} - 1.2 \times 10^{-6}$ is certainly the most cosmologically safe, as our electroweak vacuum is unique. However, if $\lambda_t > \lambda_{\text{crit}} - 1.2 \times 10^{-6}$ the evolution of the Universe should lead the system to our vacuum rather than to the vacuum with large Higgs field (as far as our vacuum is the global minimum). While in the interval $\lambda_t < \lambda_{\text{crit}} - 1.2 \times 10^{-6} < y < \lambda_{\text{crit}}$ our vacuum is deeper than another one, in contrast with the case $y > \lambda_{\text{crit}}$, where it is the other way around.
Variation of the top quark Yukawa coupling in the allowed by experimental and theoretical uncertainties interval changes the place where the scalar self-coupling crosses zero from 10^7 GeV to infinity, without a clear indication of the necessity of new thresholds in particle physics between the Fermi and Planck scales. For the largest allowed top Yukawa coupling the scale μ_{new} is as small as 10^7 GeV, whereas if the uncertainties are pushed in the other direction no new physics would be needed below the Planck mass.

- Also, the $t\bar{t}H$ production has a very important role in the Effective Field Theories (EFT) that study new physics through precise measurements of the production cross section of some processes like the $t\bar{t}H$. In principle, an EFT is a low-energy approximation for a more fundamental theory involving particles of mass scale Λ. In practice, an EFT is based on the construction of an effective Lagrangian $\mathcal{L}_{eff}$ by adding new physics terms to the SM Lagrangian $\mathcal{L}_{SM}$ that have dimension higher than five, respecting the symmetries and conservation laws observed in nature,

$$\mathcal{L}_{eff} = \mathcal{L}_{SM} + \sum_i \frac{c_i^{(6)}}{\Lambda^2} O_i^{(6)} + \mathcal{O}(\Lambda^{-4}) \, . \tag{1.47}$$

The additional terms O_i are the operators constructed from the products of only SM fields weighed by the Wilson coefficients c_i. The greater the dimension of an operator, the more suppressed the corresponding factor, therefore operators of the lowest possible dimension are the most responsible for describing new physics (NP). For this reason, it is common practice to look with 6th order terms, avoiding the

higher dimensions which are suppressed by the increasing $1/\Lambda$ power.
The common EFT analysis strategy is to measure the cross section for a specific physics process and unfold this measurement back to the particle level, then make a comparison with EFT predictions [25]. Deviations from the SM prediction of the cross section are then included in the context of EFT through Wilson coefficients. In this case, for every operator, terms $\mathcal{M}_i$ will be added to the matrix element $\mathcal{M}$ of a process:

$$\mathcal{M} = \mathcal{M}_0 + \sum_i c_i \mathcal{M}_i \tag{1.48}$$

In the simplest case of a single added operator the cross section is then:

$$\sigma_{SM+NP}(c) \propto |\mathcal{M}|^2 = s_0 + s_1 c_i + s_2 c_i^2 \tag{1.49}$$

where $s_0 = \sigma_{SM}$, and s_1 and s_2 parametrise the cross section in terms of Wilson coefficient. The cross section has a quadratic dependence on the Wilson coefficient of the added operator. Notice that the cross section does not necessarily reach its minimal value when $c_i = 0$. While in most cases the cross section is increased by adding an operator, it is possible for the cross section to decrease owing to the partial cancellation with SM terms. According to the EFT, assuming baryon and lepton number conservation, there is a total of fifty-nine independent dimension-six operators [26], thirty-nine of those operators including at least one Higgs field. In the $t\bar{t}H$ production there are two kinds of relevant operators: those with four fermion fields and those with two or no fermion fields [27]. Considering the second kind, three operators can be defined:

$$O_{t\phi} = y_t^3 \Big(\phi^\dagger \phi\Big)(\bar{Q}t)\tilde{\phi}, \tag{1.50}$$

$$O_{\phi G} = y_t^2 \Big(\phi^\dagger \phi\Big) G^A_{\mu\nu} G^{A\mu\nu}, \tag{1.51}$$

$$O_{tG} = y_t g_s \Big(\bar{Q}\sigma^{\mu\nu} T^A t\Big)\tilde{\phi} G^A_{\mu\nu}. \tag{1.52}$$

All three operators contribute to the $t\bar{t}H$ process at the tree level. The first one rescales the top quark Yukawa coupling in the SM, and also gives rise to a new ttHH coupling which contributes to Higgs pair production. The second one is a loop-induced interaction between the gluon and Higgs fields. Even though it does not involve a top-quark field explicitly, it is generally included for consistency because the O_{tG} mixes into this operator, and this operator in addition mixes into $O_{t\phi}$. The third one represents the chromo-dipole moment of the top quark. It modifies the gtt vertex in the SM and produces new four-point vertices, ggtt and gttH, as well as a five-point ggttH vertex. One can obtain the differential distributions at LO and NLO for the $pp \to t\bar{t}H$ process using the MG5_aMC generator [28] framework. As an example, it is reported in Fig. 1.6 the normalised differential cross section distribution as a function of the transverse momentum distributions of the $t\bar{t}$ system. The SM contribution as well as the individual operator contributions, normalised, are displayed, in order to compare the kinematic features from different operators. The magnitudes can be read off from the total cross section tables. In the lower panel the differential K factors are represented for each operator, together with the $\mu_{R,F}$ uncertainties. Both interference and squared contributions are shown. Given the current limits on the coefficients, it is likely that the O_{tG} operator still leads to observable effects on the shape, due to large squared contributions.

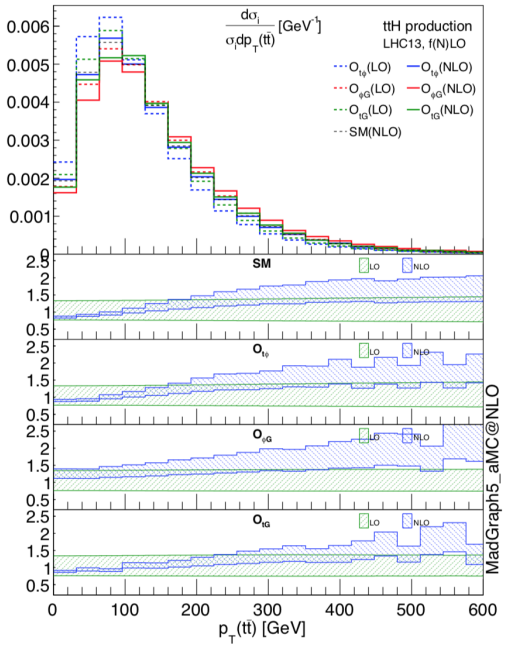

Figure 1.6: Normalised differential cross section distribution as a function of the transverse momentum distributions of the $t\bar{t}$ system

1.6.2 Observation of $t\bar{t}H$ production

Before the $t\bar{t}H$ observation took place, the CMS experiment had already performed several searches for $t\bar{t}H$ production using 7 and 8 TeV collision data from 2011 and 2012, corresponding to $5\,\text{fb}^{-1}$ and $19.5\,\text{fb}^{-1}$, respectively [29, 30]. Searches at a centre-of-mass energy of 13 TeV have been conducted in the W^+W^-/multilepton, ZZ, $\gamma\gamma$ and $\tau\tau$ final states of the Higgs boson with $35.9\,\text{fb}^{-1}$ of data collected in 2016 [31, 32, 33].
The $t\bar{t}H$ production has been observed only recently in 2018 by the ATLAS and CMS Collaborations [34, 35]. This was the result of statistically independent searches for $t\bar{t}H$ decaying in different topologies that were combined together to maximize sensitivity. In the $H \to \gamma\gamma$ channel, $t\bar{t}H$ events are searched for a narrow mass peak in the $m_{\gamma\gamma}$ distribution. The background is estimated from the $m_{\gamma\gamma}$ sidebands. The sensitivity in this channel is mostly limited by the available sample size. The $H \to ZZ^* \to 4\ell$ channel is currently limited by the low yields because of the small branching fraction of the Z decays to leptons. The $H \to b\bar{b}$ channel is intricate because of the large backgrounds, both physical and combinatorial in resolving the $b\bar{b}$ system from the Higgs decay, in events with six jets and four b-tagged jets. Already with the Run 1 dataset, the sensitivity of this analysis is strongly impacted by the systematic uncertainties on the background predictions. In this thesis, special care is reserved for this channel. The channel $H \to \tau^+\tau^-$, where the two τ leptons decay to hadrons, has been also considered. Finally, the W^+W^-, $\tau^+\tau^-$, and ZZ^* final states can be searched for inclusively in multilepton event topologies. The signal over-background-ratio is displayed in Fig. 1.7. The presence of a $t\bar{t}H$ signal is assessed by performing a simultaneous fit to the data from the different decay modes. The test statistic q, defined as the negative of twice the logarithm of the profile likelihood ratio, has been adopted, with the systematic uncertainties incorporated through the use of nuisance parameters treated according to the frequentist paradigm [36]. An excess of events from

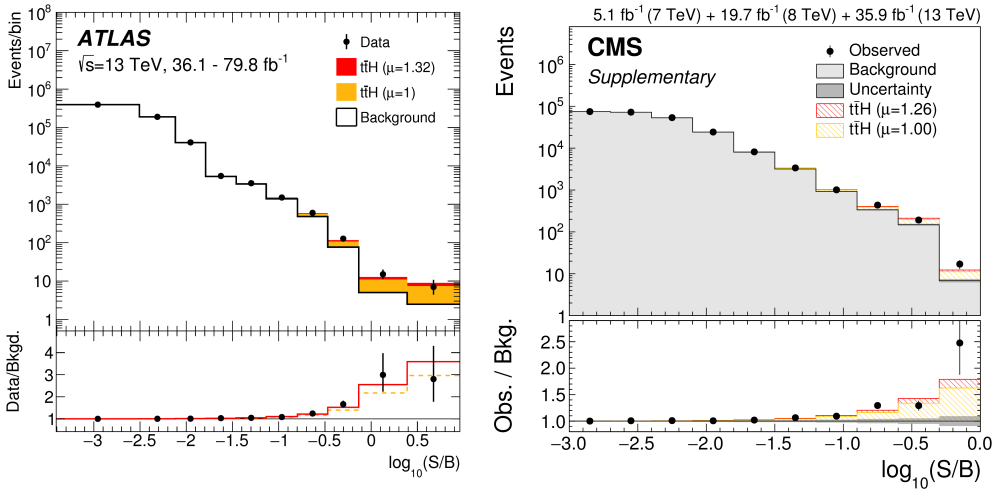

Figure 1.7: Signal-over-background ratio for ATLAS (left) and CMS (right).

the SM for a Higgs boson mass of 125.09 GeV is observed, with an observed (expected) significance of 5.2 (4.2) standard deviations for the CMS collaboration, as can be seen in Fig. 1.8, and an observed (expected) significance of 6.3 (5.1) standard deviations for the ATLAS collaboration. The combined (7+8+13 TeV) best-fit signal strength $\mu_{t\bar{t}H}$,

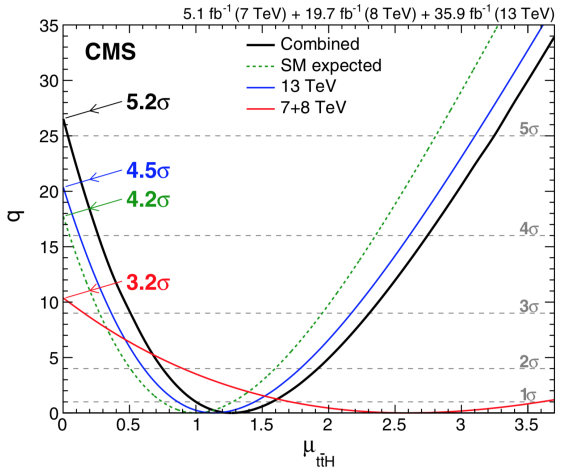

Figure 1.8: Test statistic q as a function of $\mu_{t\bar{t}H}$ for all decay modes at 7 + 8 TeV and at 13 TeV, shown separately and combined. The horizontal dashed lines indicate the p values for the background-only hypothesis obtained from the asymptotic distribution of q, expressed in units of the number of standard deviations.

defined as the observed $t\bar{t}H$ cross section $\sigma_{t\bar{t}H}$ normalized to its the SM prediction $\sigma^{SM}_{t\bar{t}H}$, is $1.32^{+0.28}_{-0.26}(\text{tot})$ for ATLAS and $1.26^{+0.31}_{-0.26}(\text{tot})$ for CMS (see Fig. 1.9).

In addition to comprising the first observation of a new Higgs boson production mechanism, this measurement establishes the tree-level coupling of the Higgs boson to the top quark, and hence to an up-type quark, and is another milestone towards the measurement of the Higgs boson coupling to fermions. Also, the overall agreement observed between

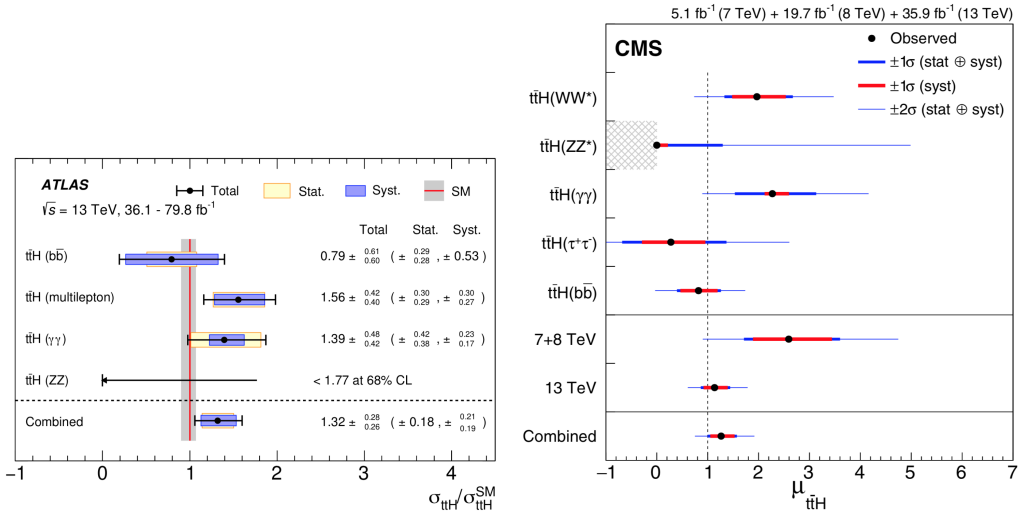

Figure 1.9: Signal strengths for ATLAS (left) and CMS (right).

the SM predictions and data for the rate of Higgs boson production through gluon-gluon fusion decay mode suggested that the Higgs boson coupling to top quarks is SM-like, since the quantum loops in these processes include top quarks. However, non-SM particles in the loops could introduce terms that compensate for, and thus mask, other deviations from the SM. A measurement of the production rate of the tree-level $t\bar{t}H$ process provides clearer evidence for, or against, such new-physics contributions.

1.6.3 Theoretical cross section $t\bar{t}H$ production

The computation for the LO $t\bar{t}H$ cross section is very complicated and must take into consideration all the possible Feynman diagrams, displayed in Fig 1.10. The complete analytical expression for the LO $gg \to t\bar{t}H$ considers the all possibles permutation of exchanging the fermion with the antifermion and the gluons with each other from the Feynman diagram initiated by gluons (b), (c), (d). Following the notation of [37], we begin by denoting the four-momenta of the incoming gluons, top quark, top antiquark and Higgs boson respectively by g_1, g_2, p, $\bar{p}$ and k, and the gluon polarisation four-vectors as ϵ_1 and ϵ_2. The invariant mass squared of the initial gluons is given by $\hat{s} = Q^2 = (g_1 + g_2)^2 = (p + \bar{p} + k)^2$ and the LO scattering amplitudes for the three diagrams shown in (b), (c) and (d), labelled $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$, respectively, are given by:

$$\mathcal{M}_1 = -AX^a_{ik}X^b_{kj}\bar{u}^j(p)\frac{\not{k} + \not{p} + m_t}{2p\cdot k + M_H^2}\not{\epsilon}_2\frac{-\not{\bar{p}} + \not{g}_1 + m_t}{-2g_1\cdot\bar{p}}\not{\epsilon}_1 v^i(\bar{p}) + \begin{Bmatrix} g_1 \leftrightarrow g_2, \epsilon_1 \leftrightarrow \epsilon_2 \\ g_1 \leftrightarrow g_2, \epsilon_1 \leftrightarrow \epsilon_2, p \leftrightarrow \bar{p} \\ p \leftrightarrow \bar{p} \end{Bmatrix} \tag{1.53}$$

$$\mathcal{M}_2 = -AX^a_{ik}X^b_{kj}\bar{u}^j(p)\not{\epsilon}_2\frac{\not{p} - \not{g}_2 + m_t}{-p\cdot g_2}\frac{-\not{\bar{p}} + \not{g}_1 + m_t}{-g_1\cdot\bar{p}}\not{\epsilon}_1 v^i(\bar{p}) + \{g_1 \leftrightarrow g_2, \epsilon_1 \leftrightarrow \epsilon_2\} \tag{1.54}$$

$$\mathcal{M}_3 = iAf^{abc}X^c_{ij}\bar{u}^j(p)\frac{\not{\epsilon}_1\not{\epsilon}_2 Q^\lambda}{\hat{s}} + \Big[2g_1^\nu g^{\lambda\mu} + (g_2 - g_1)^\lambda g^{\mu\nu} - 2g_2^\mu g^{\nu\lambda}\Big]\frac{\not{\bar{p}} + \not{k} - m_t}{2k\cdot\bar{p} + M_H^2}v^i(\bar{p}) + \{p \to \bar{p}\} \tag{1.55}$$

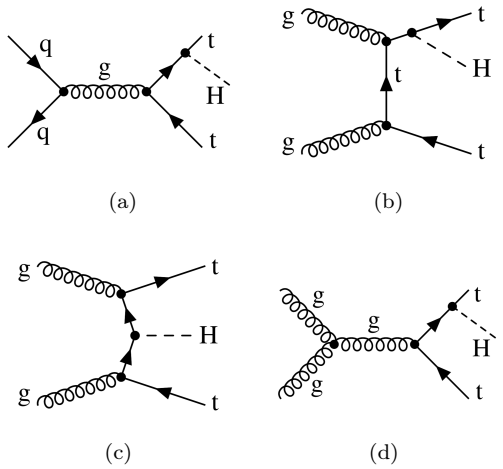

Figure 1.10: Examples of LO Feynman diagrams for $t\bar{t}H$ production: (a) initiated by quarks; (b) initiated by gluons with t-channel exchange and radiation from external lines (c) initiated by gluons with t-channel exchange and radiation from internal lines; (d) initiated by gluons with s-channel exchange and radiation from external lines.

where $A = 4\pi\alpha_S(\sqrt{2}m_t^2 G_F)^{1/2}$ are the coupling factors, and the SU(3) generators X^a and structure constants f^{abc}. The polarisation vectors obey the transversality condition $\epsilon_i \cdot g_i = 0$ and the SU(3) gauge invariance implies $\epsilon_1 \cdot g_2 = \epsilon_2 \cdot g_1$ and the invariance substitutions $\epsilon_i \leftrightarrow g_i$.

The amplitude squared needs to be summed over the colour and spin states of the final quarks, and averaged over the colour and polarisation states of the initial gluons:

$$|\mathcal{M}|^2 = \frac{1}{256} \sum_{\text{spin,col}} |\mathcal{M}_1 + \mathcal{M}_2 + \mathcal{M}_3|^2. \tag{1.56}$$

The trace over the γ matrices and the sum over the indices of the generators and structure function yields:

$$(X^a_{ik}X^b_{kj})^2 = 24, \qquad (f^{abc}X^c_{ij})^2 = 12, \qquad (X^a_{ik}X^b_{kj})(f^{abc}X^c_{ij}) = 0, \tag{1.57}$$

while the average over the gluon polarisation states must be performed in an axial gauge (since the gluons are massless), for example:

$$\sum_{\lambda_i=1}^{2} \epsilon_i^\mu(g_i,\lambda_i)\epsilon_i^\nu(g_i,\lambda_i) = -g^{\mu\nu} + \frac{2}{\hat{s}}(g_1^\mu g_2^\nu + g_1^\nu g_2^\mu) \tag{1.58}$$

The cross section for the core $gg \to t\bar{t}H$ process is then obtained by integrating over the phase space as:

$$\hat{\sigma}_{LO} = \frac{1}{\hat{s}} \frac{\alpha_S^2 G_F m_t^2}{\sqrt{2}\pi^3(2\pi)^9} \int \frac{\mathrm{d}^3p}{2E_t} \frac{\mathrm{d}^3\bar{p}}{2E_{\bar{t}}} \frac{\mathrm{d}^3k}{2E_H} \delta^{(4)}(Q - p - \bar{p} - k)\,|\mathcal{M}|^2 \tag{1.59}$$

The parton level cross section must then be folded with the gluon luminosity to obtain the full cross section for the process $pp \to gg \to t\bar{t}H$:

$$\sigma_{LO} = \int_0^1 \frac{1}{2}\Big[g(x_1,\mu_F)\,g(x_2,\mu_F)\,\hat{\sigma}_{LO}(x_1,x_2,\mu_F) + \big\{x_1 \to x_2\big\}\Big]\mathrm{d}x_1\mathrm{d}x_2 \tag{1.60}$$

Then we should consider the top quark and Higgs boson decays. The scattering amplitude must be multiplied by the decay amplitudes to give:

$$|\mathcal{M}_{\mathrm{gg}\to\mathrm{t\bar{t}H}\to\mathrm{qqb,qqb,b\bar{b}}}|^2 = |\mathcal{M}|^2 \cdot |\mathcal{M}_{\mathrm{t}\to\mathrm{qqb}}|^2 \cdot |\mathcal{M}_{\bar{\mathrm{t}}\to\mathrm{qqb}}|^2 \cdot |\mathcal{M}_{\mathrm{H}\to\mathrm{b\bar{b}}}|^2. \tag{1.61}$$

The top quark and Higgs boson decay amplitudes can be simplified with the narrow width approximation and expressed in terms of the vertex amplitudes:

$$|\mathcal{M}_{\mathrm{t}\to\mathrm{qqb}}|^2 = \frac{\pi}{m_t\Gamma_t}\delta(p^2 - m_t^2)|\mathcal{M}_{\mathrm{q,q,b}}|^2 \tag{1.62}$$

$$|\mathcal{M}_{\mathrm{H}\to\mathrm{b\bar{b}}}|^2 = \frac{\pi}{m_H\Gamma_H}\delta(k^2 - m_H^2)|\mathcal{M}_{\mathrm{b,b}}|^2 \tag{1.63}$$

The phase space must now only include the final state quarks. Denoting the four-momenta of the top quark decay products as q_1, q_1', b_1, those of the top antiquark as q_2, q_2', b_2 and those of the Higgs boson as b, $\bar{b}$, the phase space volume is parameterised as:

$$\mathrm{d}\Phi = \frac{1}{(2\pi)^{24}}\frac{\mathrm{d}\vec{q_1}}{2E_{q1}}\frac{\mathrm{d}\vec{q'}_1}{2E_{q'1}}\frac{\mathrm{d}\vec{b_1}}{2E_{b2}}\frac{\mathrm{d}\vec{q_2}}{2E_{q1}}\frac{\mathrm{d}\vec{q'}_2}{2E_{q'2}}\frac{\mathrm{d}\vec{b_2}}{2E_{b2}}\frac{\mathrm{d}\vec{b}}{2E_b}\frac{\mathrm{d}\vec{\bar{b}}}{2E_{\bar{b}}} \tag{1.64}$$

The cross section for the gluon initiated $\mathrm{t\bar{t}H}$ process in the all-hadronic decay channel is therefore given by:

$$\sigma_{LO}^{\mathrm{gg}\to\mathrm{t\bar{t}H}\to 8\mathrm{q}} = \frac{1}{\hat{s}}\frac{\alpha_S^2 G_F m_t^2}{\sqrt{2}\pi^3}\int \mathrm{d}\Phi\delta^{(4)}(Q - \sum_{i=1}^{8} p_i)|\mathcal{M}_{\mathrm{gg}\to\mathrm{t\bar{t}H}\to\mathrm{qqb,qqb,b\bar{b}}}|^2 \tag{1.65}$$

and the final cross section starting from protons is expressed as:

$$\sigma_{LO}^{\mathrm{pp}\to\mathrm{t\bar{t}H}\to 8\mathrm{q}} = \int_0^1 \frac{1}{2}\Big[g(x_1,\mu_F)g(x_2,\mu_F)\sigma_{LO}^{\mathrm{gg}\to\mathrm{t\bar{t}H}\to 8\mathrm{q}} + \{x_1 \leftrightarrow x_2\}\Big]\mathrm{d}x_1\mathrm{d}x_2 \tag{1.66}$$

Values of the $\mathrm{t\bar{t}H}$ production cross section as a function of the center of mass energy $\sqrt{s}$ are reported in Table 1.5. For our analysis, which is a simulation at $\sqrt{s} = 13$ TeV, the $\mathrm{t\bar{t}H}$ cross section is $0.50^{+9\%}_{-13\%}$ (pb).

$\sqrt{s}$	$\mathrm{t\bar{t}H}$ production cross section (in pb)
1.96	$0.004^{+10\%}_{-10\%}$
7	$0.09^{+8\%}_{-14\%}$
8	$0.13^{+8\%}_{-13\%}$
13	$0.50^{+9\%}_{-13\%}$
14	$0.60^{+9\%}_{-13\%}$

Table 1.5: SM Higgs boson production cross sections for $m_H = 125$ GeV in pp collisions ($\mathrm{p\bar{p}}$ collisions at $\sqrt{s} = 1.96$ TeV for the Tevatron), as a function of $\sqrt{s}$. Values are taken from [5].

1.6.4 The all-hadronic $\mathrm{t\bar{t}H}$ channel

In the all-hadronic $\mathrm{t\bar{t}H}$ decay mode channel, the Higgs boson decays exclusively to $\mathrm{b\bar{b}}$, and each top quark decays to a bottom quark and a W boson, which in turn decays to two quarks. Searches in which the $\mathrm{H}\to\mathrm{b\bar{b}}$ decay mode is selected and the W bosons are allowed to decay into leptons have also been reported by ATLAS [38] and CMS [39].

ATLAS dedicated a search for $t\bar{t}H$ production in the all-hadronic final state at $\sqrt{s} = 8$ TeV, in which the observed and expected upper limits on the signal strength resulted to be 6.4 and 5.4 at 95% CL, and a best fit value for the signal strength of $\hat{\mu} = \sigma/\sigma_{SM} = 1.6 \pm 2.6$ [40]. Six independent analysis regions are considered for the fit used by the ATLAS analysis: two control regions (6j, 3b), (6j, ≥4b) and four signal regions (7j, 3b), (7j, ≥4b), (≥8j, 3b) and (≥8j, ≥4b). In addition, the three regions with exactly two b-tagged jets, (6j, 2b), (7j, 2b) and (≥8j, 2b), are used to predict the multijet contribution to higher b-tagging multiplicity regions, using the tag rate function for multijet events (TRFMJ) method. The categories are analysed separately and combined statistically to maximise the overall sensitivity. The most sensitive regions, (≥8j, 3b) and (≥8j, ≥4b), are expected to contribute more than 50% of the total significance. The combined post-fit event yields for data, total background and signal expectations as a function of $\log_{10}$(S/B) are shown in the left panel of Fig. 1.11. The signal is normalised to the fitted value of the signal strength ($\mu = 1.6$). A signal strength 6.4 times larger than predicted by the SM is also shown in the left panel of Fig. 1.11. The all-hadronic best fit value of $\hat{\mu} = \sigma/\sigma_{SM} = 1.6 \pm 2.6$ has been combined with other $t\bar{t}H$ search channels in which $H \to b\bar{b}$, and the combined result yields a best fit value of $\hat{\mu} = \sigma/\sigma_{SM} = 1.4 \pm 1.0$, as shown in the right panel of Fig. 1.11.

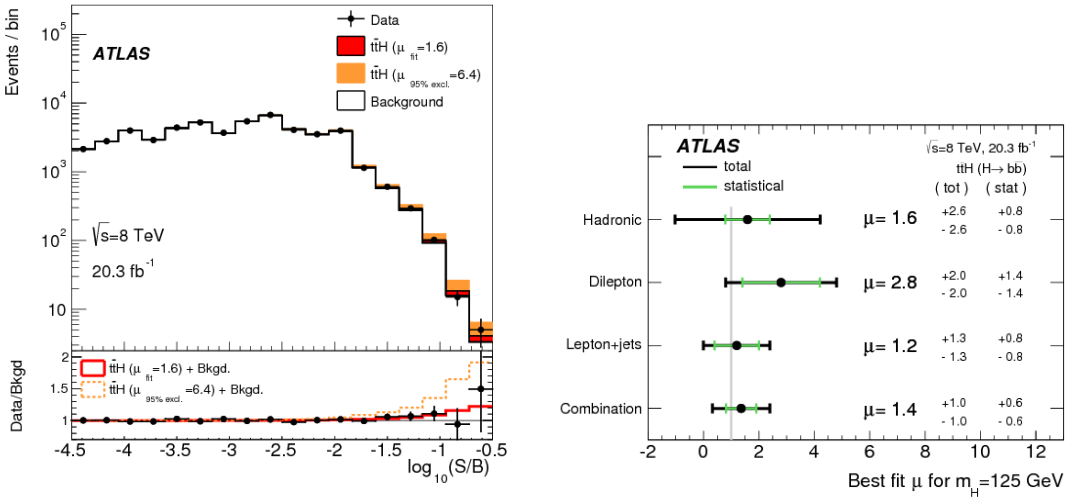

Figure 1.11: (Left) Event yields as a function of $\log_{10}$(S/B) taken from the corresponding BDT discriminant bin. The $t\bar{t}H$ signal is shown both for the best-fit value ($\mu = 1.6$) and for the upper limit at 95% CL ($\mu = 6.4$). (Right) Measurements of the signal strength for the $t\bar{t}H$ production in the $H \to b\bar{b}$ decay mode channels and their combination, assuming $m_H = 125$ GeV. The SM $\mu = 1$ expectation is shown as the grey line.

CMS published a search for $t\bar{t}H$ production in the all-hadronic decay channel at $\sqrt{s} =$ 13 TeV, corresponding to an integrated luminosity of 35.9 fb^{-1} [41]. Events, which are selected to be compatible with the $H \to b\bar{b}$ decay and the all-jet final state of the $t\bar{t}$ pair, are divided into six categories according to their reconstructed jet and b jet multiplicities: (7j, 3b), (7j, ≥ 4b), (8j, 3b), (8j, ≥ 4b), (≥ 9j, 3b), (≥ 9j, ≥ 4b). Events ≥ 7j and ≥ 2b are used to form control regions for the multijet background estimation. The categories are analysed separately and combined statistically to maximise the overall sensitivity.

The combined post-fit event yields for data, total background and signal expectations as a function of log10(S/B) under the signal+background hypothesis are shown in Fig. 1.12. From a combined fit of signal and background templates to the data in all event categories, a best fit value was obtained for the signal strength relative to the SM value of $\hat{\mu} = 0.9 \pm 1.5$, which is compatible with the SM expectation, as shown in the left panel of Fig. 1.13. Observed and expected upper limits for the signal strength are computed separately for each category and combined together. Observed and expected upper limits of 3.8 and 3.1, respectively, are obtained at 95% CL, as shown in the right panel of Fig. 1.13.

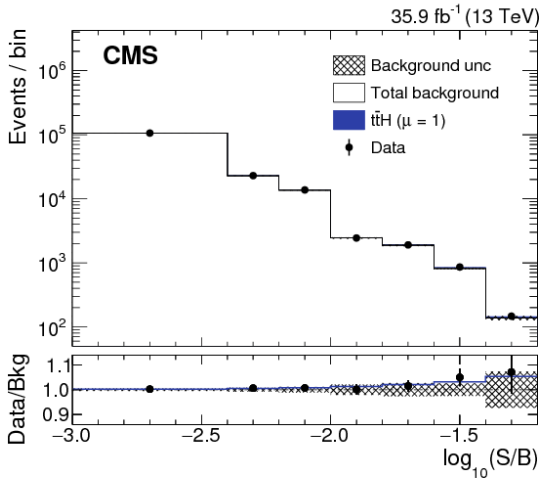

Figure 1.12: Event yields as a function of $\log_{10}$ (S/B) taken from the corresponding MEM discriminant bin. The $t\bar{t}H$ signal is shown both for the signal+background hypothesis $\mu = 1$ at 95% CL.

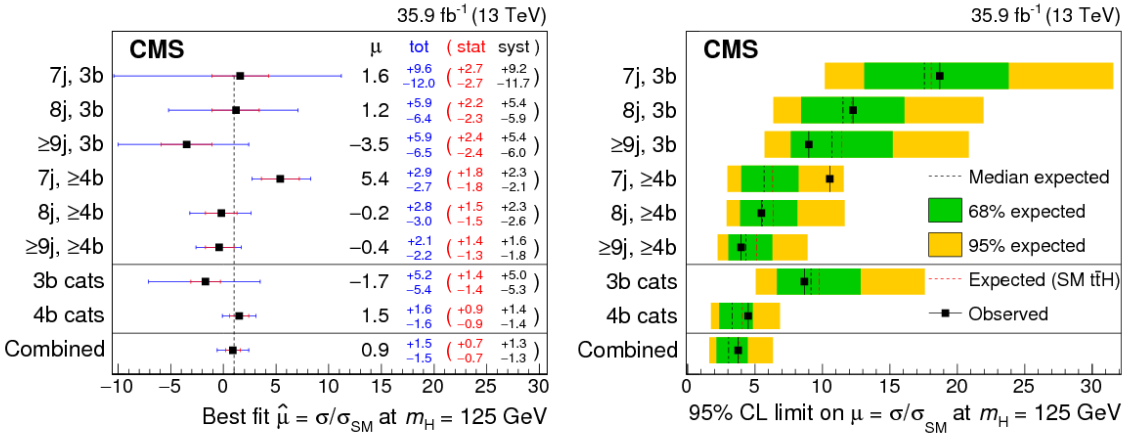

Figure 1.13: (Left) Best fit values of the signal strength, and their 68% CL intervals as split into the statistical and systematic components. (Right) Median expected and observed 95% CL upper limits on μ. The expected limits are displayed with their 68% and 95% CL intervals, as well as with the expectation for an injected SM signal of $\mu = 1$.

1.6.5 Backgrounds for the all-hadronic $t\bar{t}H$ production

There are several SM processes that can produce the same final state as the all-hadronic $t\bar{t}H$ signal, with eight jets including four b-jets. The underlying production mechanisms vary substantially, but in all cases the required number of jets is reached only through radiation. Nevertheless, in high-energy proton-proton collisions, QCD radiation is very common, even up to several consecutive splittings, thus ensuring that the signal rate is overwhelmed by SM background. Furthermore, the presence of four real b-jets is not necessary for background processes as there is a significant probability of one or more light-flavour jets to be incorrectly identified as a b jet in the detector. The SM backgrounds and their main features are described below in order of dominance:

- QCD multijet: the most dominant background is from jets produced through the strong interaction, referred to as QCD multijet events. Such events include multiple gluon radiation and have a large cross section which drops off as the jet and b-jet multiplicity increase and the jet p_T increase. Nevertheless, at eight jets with high p_T the cross section is still substantially above the signal.

- $t\bar{t}$+ jets: The SM $t\bar{t}$ production with additional jets from radiation forms a large and difficult background, as it has a large cross section and involves a final state with very similar kinematic properties to the signal. The process is considered irreducible when the additional jets are b-jets, and is then referred to as $t\bar{t} + b\bar{b}$. If the additional jets are c-jets there is a larger probability of misidentifying them as b-jets, making the process more difficult to distinguish from the signal.

- Single top quark: Single top quark production (single t) constitutes the next most dominant background, although it is considered a minor background. It has a larger cross section than the signal, but since it requires many additional radiated jets, its total contribution in the selected final state is less than the signal. The process can occur through an exchange of a W boson in the t- or s-channel, or in the tW-channel.

- W + jets: W boson production has a much larger cross section than the signal, however to become a background it requires a significant amount of radiation, which effectively reduces its cross section to below that of the signal.

- Z + jets: Z boson production has a lower cross section than W boson production, and at the jet and b-jet multiplicity of the signal, it also has a lower cross section than W + jets.

- $t\bar{t}$ + Z: $t\bar{t}$ production in association with a Z boson has a similar cross section to $t\bar{t}H$ production, however the branching ratio for $Z \to b\bar{b}$ is lower than that for $H \to b\bar{b}$, and therefore it presents a signal-like final state at a lower rate than the signal.

- $t\bar{t}$ + W: $t\bar{t}$ production in association with a W boson also has a similar cross section to $t\bar{t}H$ production, however the W boson cannot decay to two b quarks. Therefore, it makes an even smaller background contribution than $t\bar{t}$ + Z.

- Diboson: The production of two weak vector bosons occurs as WW, WZ or ZZ in decreasing order of cross section. Although the three processes have a cross section one to two orders of magnitude larger than the signal, the number of additional jets required to form a background is large and thus the final contribution is very small.

Chapter 2

The CMS experiment at LHC

CERN is the world largest laboratory dedicated to the exploration of particle physics, originated from a European organisation for nuclear research. It hosts the largest particle accelerator on earth, the Large Hadron Collider (LHC). It is a circular almost 27 km long accelerator where protons are accelerated up to a centre-of-mass energy of $\sqrt{s} = 13$ TeV, the highest energy ever achieved by a man-made accelerator. Four main experiments are located along the LHC, recording and studying the proton-proton collisions or heavy ions collisions: "A Large Ion Collider Experiment" (ALICE), "A Toroidal LHC ApparatuS" (ATLAS), "Compact Muon Solenoid" (CMS), "LHC-beauty" (LHCb). This thesis was carried out at the CMS experiment [42]. CMS is a multipurpose detector and it has an onion-like structure which combines different subdetectors measuring different aspects of the particles arising from the proton-proton collisions. A general overview of the LHC accelerator is provided in Section 2.1 and a description of the CMS detector with its subdetectors in Section 2.2.

2.1 The LHC

The LHC provides the most energetic particle collisions under laboratory conditions. It is a circular synchrotron accelerator, housed in a 26.7 km long tunnel located at 50 to 175 meters below the ground. The LHC is situated beneath the franco-swiss border area in the north-west of Geneva, Switzerland. Before the LHC was built, the same tunnel has accommodated the Large Electron-Positron Collider (LEP), which was shut down in the year 2000. The LHC can be used to collide protons or heavy ions. Beams composed of spatially separated bunches of these particles counter-rotate in two designated beam pipes. The LHC is designed to hold 2808 bunches with each of them containing either about 10^{11} protons or about 10^8 Pb^{82+} ions. The vacuum within the beam pipes prevents interactions of the particles with gas molecules, which could lead to instabilities of the beam. The particles in the LHC are accelerated by 16 superconducting radio-frequency cavities. They are grouped to modules including four cavities each. Within these modules, two cavities are designated for the acceleration of the particles of each beam. The cavities are built from copper coated with niobium on the inside. Using liquid helium, they are cooled down to 4.5 K in order to transfer the niobium to a superconducting state. Within the cavities, electromagnetic oscillations are stimulated at a frequency of 400 MHz. Due to the special shape, only modes longitudinal with respect to the beam direction are stimulated. Particles passing the cavities are accelerated in the oscillating field gradient ranging up to 5 MV/m. Due to the oscillations, the particles are automatically grouped into bunches. A total of 1232 superconducting dipole magnets keep the particles on the

circular path. The coils of the dipole magnets are made of niobium-titanium. They are brought to their superconducting state by cooling them down to 1.9 K with superfluid helium-4. This state allows to operate the dipole magnets with a current of 11850 A for a maximum magnetic field of 8.33 T. More than 8000 additional superconducting magnets with higher multipole orders are installed to focus and stabilise the beam. The beam in one pipe circulates clockwise while the beam in the other pipe circulates anticlockwise. It takes 4 minutes and 20 seconds to fill each LHC ring, and 20 minutes for the protons to reach their maximum energy of 6.5 TeV. Beams circulate for many hours inside the LHC beam pipes under normal operating conditions. Actually, the acceleration of the particles in the LHC represents only the last stage in a sequence of preliminary accelerations. An illustration of the entire acceleration complex is displayed in Fig. 2.1

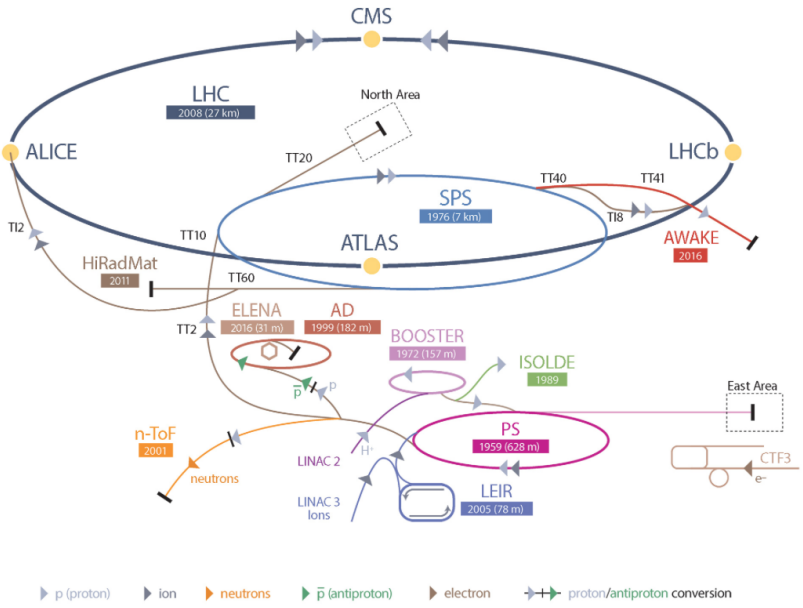

Figure 2.1: Sketch of the CERN particle acceleration complex.

The process of acceleration of protons starts with a simple bottle of hydrogen gas. The atoms of this gas are ionized by an electric field in order to obtain protons (H^+). These are subsequently accelerated to 50 MeV by the linear particle accelerator LINAC2. The accelerated protons are injected into the proton synchrotron BOOSTER, where they reach energies of 1.4 GeV. In a subsequent step, the protons are accelerated to 26 GeV by the Proton Synchrotron (PS). In the last step before the injection into the LHC, the protons are brought to an energy of 450 GeV by the Super Proton Synchrotron (SPS). The accelerator complex also includes the Antiproton Decelerator (AD) and the Online Isotope Mass Separator (ISOLDE) facility, and the Compact Linear Collider test area, as well as the neutron time-of-flight facility (nTOF).

2.2 The CMS experiment

CMS is a multipurpose detector designed to detect a broad range of signatures provided by SM and new physics. It is situated in one of the four spots where the two particle beams are brought to collision, in a cavern built about 100 m beneath the surface. It is a hermetic

detector aiming at detecting as many particles produced in the collisions as possible. Each subdetector system is specialised to measure the properties of different types of particles. The diameter of the detector amounts to 15 m while the length of the detector adds up to 21 m. These dimensions are necessary to ensure proper measurement of the particles properties. Still, compared to the ATLAS experiment, which is about double the size, the CMS indeed is quite compact. This compact build requires the application of very dense materials in order to stop particles before they leave the detector. Accordingly, the total weight of the CMS detector adds up to 14 000 t. A schematic view of the CMS detector is shown in Fig. 2.2.

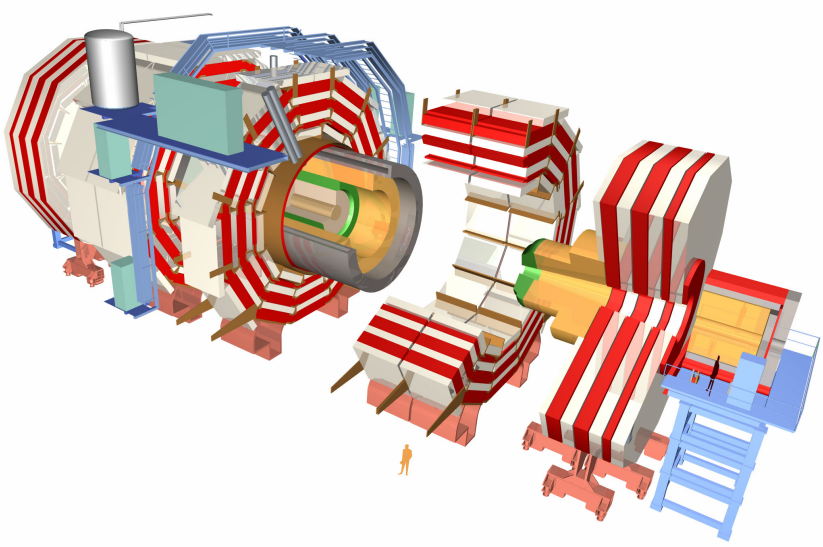

Figure 2.2: Illustration of the CMS detector. The various detector components are the tracker system in beige, the ECAL in green, the HCAL in yellow, the solenoid in grey, the return yoke in red, and the muon system in white.

Starting from the collision point, the innermost subdetector is the tracker system. It consists of different layers of silicon detectors enclosing the interaction point. Each layer allows a precise determination of the position of charged particles passing through it and combining the positions of different layers one can determine the trajectories of the particles. Together with the strong magnetic field provided by the solenoid, the trajectories allow the determination of the momentum and the electric charge of passing particles. Numerous lead-tungstate crystals, which surround the tracking system, form the electromagnetic calorimeter (ECAL). Light electromagnetically interacting particles, like electrons, positrons, and photons, deposit all of their energy within these crystals, allowing an energy measurement. The adjacent hadronic calorimeter (HCAL) consists of alternating layers of absorbers and active material. Hadrons entering this subdetector interact with the absorber material and are expected to be completely stopped within the HCAL. The active medium measures the energy deposited by the initial particles. The HCAL is surrounded by the superconducting solenoid, which provides a strong magnetic field necessary for the determination of the momentum and the electric charge of particles. The return yoke is an iron structure encasing the solenoid. It provides structural support for the detector and guides the magnetic field. These components are interspersed with muon chambers, gas ionization detectors measuring the tracks of passing electrically charged particles. As muons are the only electrically charged particles expected to reach

this part of the detector, signals in the muon system provide good identification criteria for them. The arrangement of the different subsystems in the CMS detector is illustrated in Fig. 2.3.

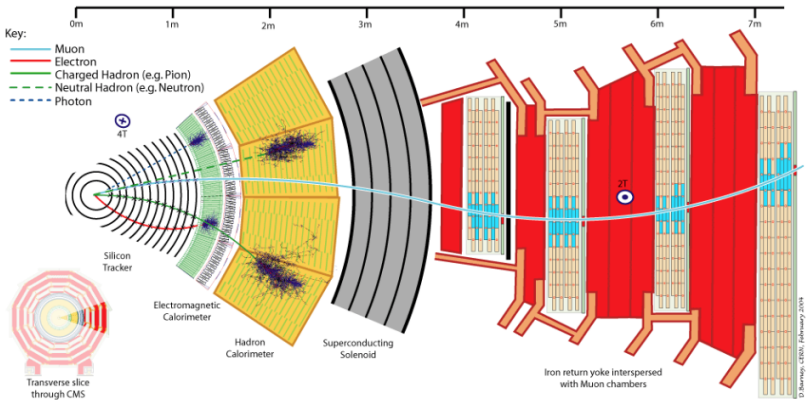

Figure 2.3: Illustration of the CMS subdetectors.

Additionally, this figure shows examples of interactions of different types of particles with the subdetectors. The signals provided by the individual subdetectors are read out by the data acquisition system. However, not all events can be processed and stored, as this would exceed the capabilities of processing and storage resources. Consequently, a large fraction of events lacking interesting features is rejected by a dedicated trigger system. The recorded data is stored and analyzed on a distributed computing infrastructure, the Worldwide LHC Computing Grid.

2.2.1 Coordinates system

At LHC, a specific coordinate system is used to describe the positions and directions of the particles in the detectors. First of all, it is a right-handed coordinate system with its origin at the designated point of collision. The z-axis points in the direction of the counterclockwise rotating beam, which is westwards from the LHC Point 5 to the Jura mountains. The x-axis of the coordinate system points towards the centre of the LHC, whereas the y-axis points vertically upwards. The most common coordinates used for the description of the detector and particles are spherical coordinates. These coordinates include the distance from collision point denoted by r and the two angles ϕ and θ. The azimuthal angle ϕ is located in the x-y plane, which is orthogonal to the beam axis. The polar angle θ is measured with respect to the z-axis. In proton-proton collisions, a large number of interactions with small momentum transfers occur. This causes the regions with low values of the polar angle to be highly populated. In regions with large values of the polar angle, on the other hand, comparably few particles can be found. Accordingly, the distributions of particles are not flat functions of the polar angle. Furthermore, the interacting partons are very likely to feature different momentum fractions of the respective proton. Consequently, their system features a residual longitudinal boost. However, the polar angle is not invariant under a longitudinal boost. A Lorentz invariant variable that additionally provides flat distributions is the rapidity,

$$y = \frac{1}{2} \ln\left(\frac{E + p_z}{E - p_z}\right) \tag{2.1}$$

In this equation, E denotes the energy of the respective particle and p_z is the z-component of the particle momentum. If the mass of a particle is negligible compared to its momentum, the rapidity is identical to the pseudorapidity,

$$\eta = -\ln\left(\tan\left(\frac{\theta}{2}\right)\right) \tag{2.2}$$

which is a direct function of the polar angle θ. In the following, the pseudorapidity is used for the description of positions and directions in the detector instead of the polar angle.

2.2.2 Tracker system

The tracker system is the innermost subdetector of the CMS experiment and performs multiple precise position determinations of the electrically charged particles produced in collision experiments. It is the system closest to the interaction point and as a result it has to record a large throughput of particles produced in the collisions. In order to distinguish individual particles, a very fine granularity is required. For this reason, the CMS tracker system is subdivided into the pixel tracker with a fine granularity and the coarser segmented strip tracker. The pixel tracker represents the inner part of the CMS tracker system while the strip tracker surrounds the pixel tracker [43]. The CMS tracker system covers a pseudorapidity range of $|\eta| < 2.5$. A sketch of the entire system is shown in Fig. 2.4.

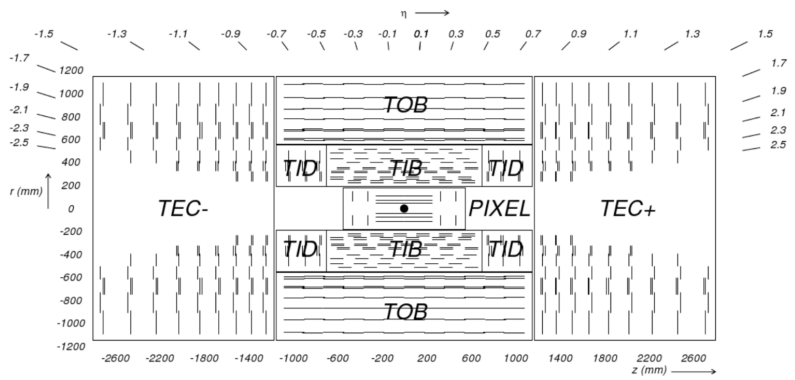

Figure 2.4: Illustration of the CMS tracker.

Additionally, the detector has to be very resistant against radiation damage. For this reason, the CMS tracking system has been developed with an all-silicon configuration. The individual tracker is composed of p-n junctions powered by a high voltage which extends the depletion region over the entire thickness of the module. Electrically charged particles passing the module cause the production of free electrons and holes. For minimum ionizing particles, the number of electrons and holes amount to about 75 per micron thickness. The free charge carriers drift to the pixels or strips implanted into the module, where the signal is read out. The drift trajectory, however, is altered by the magnetic field of the CMS solenoid. This effect is quantified by the Lorentz angle and has to be accounted for in the determination of positions. Silicon detector modules are arranged in 13 to 14 layers depending on the position in the detector. Electrically charged particles passing cause hits in the different layers, which allow the reconstruction of the entire particle trajectory. Such trajectories are crucial for the reconstruction and identification

of particles. Furthermore, tracks serve as input for the reconstruction of vertices and the identification of jets originating from bottom quarks. The trajectories of particles are bent by the magnetic field induced by the CMS solenoid. This effect enables the measurement of the momenta of the particles and their electric charges.

Pixel tracker

The pixel tracker of the CMS experiment is the innermost part of the detector. In the barrel region covering a pseudorapidity range of $|\eta| < 2.2$, pixel modules are arranged in three layers of cylinder barrels. The three barrel layers are positioned at an angular distance from the beam axis of 4.3 cm, 7.3 cm, and 10.4 cm with each of them having a longitudinal length of 53 cm. Two endcaps layers are placed at the z-coordinates of $|z| =$ 34.5 cm and $|z| = 46.5$ cm covering radii from 6 cm to 15 cm. Based on this setup, two to three hits are expected for each electrically charged particle passing the pixel tracker. The close distance to the collision point requires a fine granularity, to distinguish signals caused by different particles. Accordingly, a size of $100\,\mu\text{m} \times 150\,\mu\text{m}$ was chosen for each individual pixel on a module. Each module features a thickness of 285 μm. The pixel tracker consists of about 1 m^2 active detection area. This area is populated with 1400 modules corresponding to 66 million pixels in total. The hit resolution of the pixel tracker amounts to about 10 μm in r-ϕ direction and about 20 μm in r-z direction.

Strip tracker

The strip tracker surrounds the pixel tracker. It is subdivided into four parts. In the barrel region, two of these parts are given by the tracker inner barrel (TIB) and the tracker outer barrel (TOB). The shorter TIB consists of four cylindrical layers of strip modules while the TOB is composed of six cylindrical layers of strip modules. The ten layers of both parts are located at radii ranging from 25 cm to 108 cm. Another part of the strip tracker is the tracker inner disks (TID), which are three disks located at each end of the TIB. Each of these disks consists of three rings of strip modules. The last part of the strip detector is the Tracker EndCaps (TEC) located at each end of the TOB. The TEC consist of nine pairs of disks featuring up to seven rings of modules. The total active detection area of the strip tracker is by far larger than the one of the pixel detector. It adds up to 200 m^2 populated with silicon strip modules. However, the larger distance of the strip tracker to the collision point allows for a granularity that is coarser than the one of the pixel tracker. Accordingly, the strips are larger than the pixels and feature lengths of 9 cm for the inner parts to 21 cm for the outer parts. The pitches between the strips range between 80 μm and 120 μm. The strips themselves are given by p+ doped areas implanted into a n-type bulk with n-type backside. The total number of strips in the strip tracker amounts to about ten million. The spatial resolution of single hits in the CMS strip tracker ranges from 15 μm to 40 μm depending on the pitch between the strips.

2.2.3 Electromagnetic Calorimeter

In order to build up a picture of events occurring in the LHC, CMS must find the energies of all of the emerging particles. The ECAL determines the energies of electrons, positrons, and photons with a hermetic construction which encloses the inner tracking system in a pseudorapidity range of $|\eta| < 3.0$. The ECAL is composed of lead-tungstate crystals (PbWO4) with front cross sections of about 2 cm $\times$ 2 cm and lengths of 23 cm. About 61 000 crystals populate the barrel region, while the endcap region features about 7300 crystals [44].

Electrons, positrons, and photons entering the crystals are expected to deposit their entire energy within the crystals. High-energy photons create electron-positron pairs in interaction with matter, whereas electrons radiate photons via bremsstrahlung. The consecutive repetition of these processes by initial and resulting particles lead to the formation of electromagnetic showers. The large atomic numbers of the elements composing the crystals promote the rate of the mentioned processes, which leads to small shower geometries. Accordingly, the energy of these particles is deposited in a small volume. The radiation length and the Molière radius, which are specific properties of materials, characterise the geometry of electromagnetic showers. The radiation length, which determines the depth of penetration of an electron until its energy has fallen to $1/e$, amounts to 0.89 cm for lead-tungstate. Consequently, the length of an ECAL crystal adds up to 25.8 radiation lengths. The Molière radius determines the transverse extent of the electromagnetic shower. The small value of Molière radius, of approximately 2.2 cm for lead-tungstate, allows for a fine granularity. The lead-tungstate crystals are scintillators. The deposition of energy in the crystal stimulates the emission of photons. However, with the emission of 30 photons per MeV of energy deposited in the crystal, the photon yield is quite low. As a matter of fact, photodetectors with intrinsic amplification are used for the readout of the signal. Additionally, the photodetectors are required to be insensitive to the large magnetic field induced by the CMS solenoid. The photodetectors used are silicon avalanche photodiodes in the barrel region and vacuum phototriodes in the endcap region.
An additional part of the calorimeter system is the preshower (PS) attached prior to the ECAL endcaps. This detector component consists of two layers of lead and silicon strip detectors respectively. The silicon strip detectors feature a much finer granularity than the ECAL. This property allows the distinction between a single highly energetic photon and two spatially close low energetic photons stemming from the decay of a neutral pion. This distinction is crucial for the search of signatures featuring highly energetic photons, where pion decays into photons represent a large background. An important example is the search for a Higgs boson decaying into two photons. The preshower device is only necessary for the endcap regions, where the angles between photons originating from pions are expected to be small.

2.2.4 Hadronic Calorimeter

The CMS hadron calorimeter (HCAL) encloses the ECAL and represents the last subdetector inside the CMS solenoid. Its purpose is to stop strongly interacting particles and measure the energy deposited during this process. The design of the HCAL is chosen to fulfil this purpose, while still fitting in the limited space provided by the solenoid. Accordingly, as much material in terms of interaction lengths as possible is gathered inside the magnet coil. This is accomplished with a sandwich-calorimeter design, which features alternating layers of absorber and active material. The absorber material is brass, which features a small interaction length and is non-magnetic [45]. The brass used for the CMS HCAL was fabricated from over a million brass shell casements from World-War-II provided by the Russian Navy. Hadronic particles passing the absorber material interact with the atomic nuclei, mainly via the strong or the electromagnetic interaction. Secondary particles are detected by the layers of active material. These layers consist of tiles of plastic scintillators emitting ultraviolet light in the interaction with particles. Embedded wavelength-shifting fibres change the ultraviolet light to visible light and direct the photons to multi-channel hybrid photodiodes. The amount of light produced is proportional to the number of particles passing the scintillator. Furthermore, the number of particles produced in the interactions with the material is proportional to the energy of the initial particle.

The structure of the HCAL is subdivided into different parts. The hadron barrel detector (HB) consists of 2304 sandwich-calorimeter towers covering a pseudorapidity range of $|\eta| < 1.4$. Additionally, in the barrel region the hadronic outer detector (HO) can be found. It is made from scintillators located on the outside of the magnet coil. The HO functions as "tail-catcher" measuring the energy of particles leaking out of the HCAL and the solenoid. It covers a pseudorapidity range of $|\eta| < 1.26$ and extends the effective thickness of the HB to over ten interaction lengths. The hadron endcap (HE) is covering a pseudorapidity range of $1.3 < |\eta| < 3.0$. It consists of 2304 sandwich-calorimeter towers. The mentioned parts of the HCAL provide a similar pseudorapidity coverage as the ECAL. However, its granularity 25 times coarser. The last part of the CMS calorimeter system is the hadron forward calorimeter (HF). It covers a pseudorapidity region of $3.0 < |\eta| < 5.0$ and is located 11 m away from the collision point. The HF covers the high pseudorapidity region, which is highly populated by particles originating from collisions with small momentum transfers. Accordingly, a very radiation hard design was chosen. The HF is composed of steel absorbers and active material. The latter is given by quartz fibres embedded into the steel in a grid-like structure parallel to the beam line. Again, incoming particles interact with the atomic nuclei of the absorbers creating secondary particles. Electrically charged particles passing the quartz fibres cause the emission of Cerenkov light. The fibres redirect the produced light to photomultipliers, which extract the signal.

2.2.5 Superconducting Solenoid

The superconducting solenoidal coil positioned after the HCAL produces a uniform axial magnetic field necessary for the determination of the momentum and the charge of particles. Its length adds up to 12.9 m, while its diameter constitutes 5.9 m. Its design is strongly influenced by the fundamental concept of the CMS experiment, which foresees the tracker, the ECAL, and major parts of the HCAL to be located within the magnet coil. At the same time, the magnet is required to provide a field that is large enough to obtain a good resolution in the measurement of the momenta of the particles. The coil of the magnet is manufactured from niobium-titanium, which is coated with aluminium [46]. Liquid helium is used to bring it to its operating temperature of 4.5 K. At this temperature, the niobium-titanium conductors are in a superconducting state allowing a current of 19.5 kA. The current induces a magnetic field with a strength of 3.8 T and an energy of 2.7 GJ stored inside. The field further causes a hoop stress of 64 atm on the structure.

2.2.6 Muon System

The purpose of the muon system is to determine the position of electrically charged particles emerging the hadron calorimeter, especially muons. Muons can penetrate several meters of iron without interacting and without being stopped by any of the CMS calorimeters. Therefore, chambers to detect muons are placed at the very edge of the experiment where they are the only particles likely to register a signal. The measurements are performed in four layers in the barrel and four layers in the endcaps. The muon system is embedded into a return yoke system. As for the tracker system, these measurements can be applied to reconstruct the trajectory of electrically charged particles. In an ideal case, only muons and neutrinos are expected in this region of the detector. Accordingly, the reconstruction of a track in the muon system strongly hints at the occurrence of a muon. The muon system provides 25 000 m^2 of active detection plane [47]. Due to this large surface to be covered, the application of gas-ionization detectors has been chosen. Three

different types of modules are installed, in order to account for the different conditions in the different regions of the detector. The modules are made up of drift tube (DT) chambers, cathode-strip chambers (CSC), and resistive-plate chambers (RPC). In total, 1400 modules are installed in the CMS detector. Fig. 2.5 shows an illustration of the arrangement of the modules in the CMS detector.

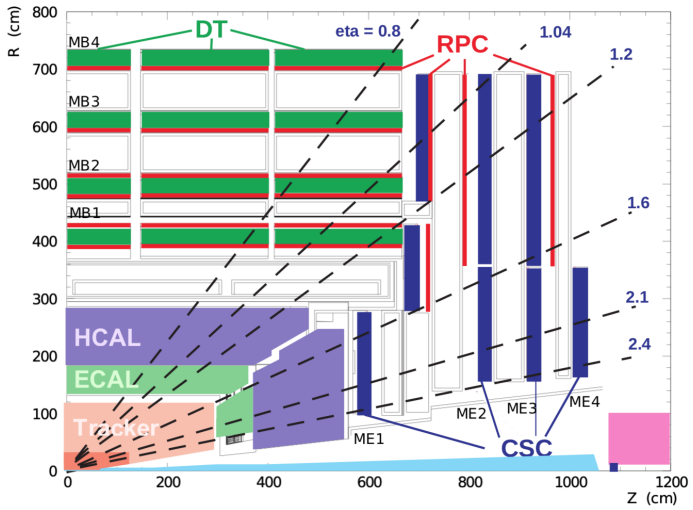

Figure 2.5: Sliced view of a quarter of the CMS detector. The various detector subsystems are highlighted in different colours. The tracker system, the ECAL and the HCAL are displayed in the lower-left corner by the areas coloured in beige, light green, and purple respectively. The subsystems associated to the muon system are illustrated by the dark colours, the DT chambers in dark green, the CSCs in red, and the RPCs in dark blue.

A powerful muon reconstruction software has been developed which reconstructs muons in the stand-alone muon system, using information from all three types of muon detectors, and links the resulting muon tracks with tracks reconstructed in the silicon tracker [48]. The software is designed to work for both offline reconstruction and for online event selection within the CMS High-Level Trigger (HLT). The muon reconstruction algorithm is the Kalman filter which uses muon candidates found by the Level-1 muon trigger as seed for the HLT, including those candidates that did not necessarily lead to a Level-1 trigger accept. These seeds define a region of interest in the muon system, in which local reconstruction is performed. In case of offline reconstruction, a different seed generation has been developed, which performs muon reconstruction in three stages: local pattern recognition, stand-alone reconstruction and global reconstruction.

Drift-Tube Chambers

DT chambers are installed only in the barrel region of the detector and they cover a pseudorapidity of $|\eta| < 1.2$. There, the muon rate, as well as the neutron-induced background and the residual magnetic field, is low. There are a total of 250 DT chambers in CMS that populate the four layers of the muon system. These four layers are located at a distance of about 4.0 m, 4.9 m, 5.9 m, and 7.0 m from the beam axis. The DT muon system is divided into five parts in z-direction and these parts are further subdivided into 12 sectors with respect to the azimuthal angle. Each DT module measures 2 m $\times$ 2.5 m and includes 12 layers of drift tubes. The 12 layers form three layers, of which the middle

one measures the z-direction of passing electrically charged particles whereas the other two layers measure the r-ϕ coordinates associated to the bending plane. Every drift tube consists of a stretched cavity bordered by aluminium, which features a width of 4 cm. The tubes are filled with a gas mixture composed of argon and carbon dioxide. A cathode strip is placed on each side of the tube and an anode wire in the middle. A high voltage is applied, leading to the formation of an electric field, that is shaped by electrodes installed at the top and the bottom of the DT cavity that make the electric field as uniform as possible. When an electrically charged particle passes through the cavity, it ionizes the gas and the resulting electrons drift to the positively charged wire. The field near the wire is so strong that the electrons are able to ionize further gas atoms and the electric signal results from the avalanche multiplication. The DT modules are bordered by one or two resistive-plate chambers depending on the layer. These detectors provide the timing of a particle entering the drift tube modules. Based on this information, the drift time of the electrons can be determined, which allows a position determination more accurate than using only the position of the anode wires. The maximum drift time in each tube is around 400 ns and the spatial resolution of a single hit in the drift tubes is about 200 μm.

Cathode-Strip Chambers

CSCs are installed in the endcaps of the detector covering a pseudorapidity region of $1.2 < |\eta| < 2.4$. In this region, the muon rate as well as the neutron-induced background and the magnetic field is large. There are a total of 468 CSC modules distributed over the layers of the muon system with a trapezoidal shape. Planes of negatively-charged copper cathode strips and planes of positively-charged anode wires alternate with six gas gaps. The anode wires and cathode strips are arranged perpendicular to each other. The gas gaps are filled with a mixture composed of argon, carbon dioxide, and tetrafluoromethane. A high voltage applied to the anode wires induces a strong electric field. Electrically charged particles passing the gap ionize the gas atoms and molecules. In the strong electric field, the electrons produced ionize further gas atoms and molecules, which leads to an avalanche of electric charges registered by the anode wire. The signal on the wire is extremely fast and is therefore applied in the Level-1 trigger system of the CMS experiment. The ionized gas atoms and molecules induce an image charge on the cathode strips. This slower signal is used to quantify the position of the passing electrically charged particle by the technique of the centre of gravity of the measured electric charges. The spatial resolution of a single hit in a CSC module ranges from 50 μm to 240 μm depending on the design, which is slightly different for the different layers of the muon system in the endcap region. The differences mainly concern the number of strips per chamber, the strip width, and the pitch width.

Resistive-Plate Chambers

RPCs are installed in both regions of the detector. There are a total of 480 RPC modules in the barrel region, and a total of 432 RPC modules in the endcaps, all together providing coverage for a pseudorapidity region of $|\eta| < 1.6$. RPCs consist of two negatively-charged cathodes separated with a positively-charged anode and forming two a gas-gaps. Each of the electrodes is covered by the high-resistivity plastic material bakelite. A plane of copper readout strips is sandwiched between the two electrode-gap structures. The gas gaps are filled with a gas mixture mainly composed of tetrafluoroethane and isobutane. Electrically charged particles passing the RPCs ionize the gas molecules. The high voltage applied between the electrodes creates a strong electric field which causes the resulting electrons to ionize further gas molecules. The avalanche of electrons drifts to the positively charged

electrodes which are transparent to the electrons produced allowing them to pass on to the strips and cause the readout signal. Based on the pattern of hits on the strips, a fast estimation of the momentum of the passing particle can be performed. This information is used in the trigger system of the CMS experiment. The RPCs feature a fast response and operate well at high rates, which allows to unambiguously identify the correct bunch crossing. The position resolution is at the order of 1 cm, which is much coarser than the one provided by the DTs and CSCs. The spatial resolution of hits in the RPCs mainly depends on the width of the readout strips.

2.2.7 Data Acquisition & Trigger

The LHC is designed to provide a bunch crossing rate of 40 MHz. One event recorded by the CMS experiment amounts to about 1 MB of zero-suppressed data. The processing and storage of all events would largely exceed the resources provided. The available storage capabilities can store data at O(1) kHz and O(100) MB/s. Accordingly, a huge fraction of the collision events has to be rejected at an early stage. The rejection rate necessary corresponds to a factor of about 10^6. The CMS trigger and data acquisition system achieves such high rejection rates based on a two-staged approach: the Level-1 (L1) trigger and the HLT.

L1 trigger

The front-end electronics situated in the detector receive signals from the various subdetector channels. Part of this information is passed on to the L1 trigger system located in the service cavern, a second cavern next to the one accommodating the CMS detector. A schematic illustration of the L1 trigger is shown in Fig. 2.6. The L1 trigger system selects only events with simple signs of interesting physics [49]. For this purpose, simple objects, so-called trigger-primitive objects, are reconstructed mainly using calorimeter and muon system information. They are processed in several steps before the combined event information is evaluated in the global trigger (GT) and a decision is made whether to accept the event or not.
The L1 calorimeter trigger comprises two stages, a regional calorimeter trigger (RCT) and a global calorimeter trigger (GCT). The RCT receives the transverse energies and quality flags from over 8000 ECAL and HCAL towers, giving trigger coverage over $|\eta| < 5$. The RCT processes this information in parallel and sends as output e/γ candidates and regional E_{T} sums based on 4×4 towers. The GCT sorts the e/γ candidates further, finds jets (classified as central, forward, and tau) using the E_{T} sums, and calculates global quantities such as E_{miss}. It sends as output four e/γ candidates each of two types, isolated and non-isolated, four each of central, tau, and forward jets, and several global quantities.
All three muon detector systems in CMS participate in the L1 muon trigger. The front-end trigger electronics of DT and CSC identifies track segments from the hit information registered in multiple detector planes of a single measurement station. These segments are collected and then transmitted via optical fibres to regional track finders in the electronics service cavern, which then applies pattern recognition algorithms that identifies muon candidates and measure their momenta from the amount they bend in the magnetic field of the flux-return yoke of the solenoid. Information is shared between the DT track finder (DTTF) and CSC track finder (CSCTF) for efficient coverage in the region of overlap between the two systems at $|\eta| \approx 1$. The hits from the RPCs are directly sent from the front-end electronics to pattern comparator trigger (PACT) logic boards that identify muon candidates. The three regional track finders sort the identified muon candidates

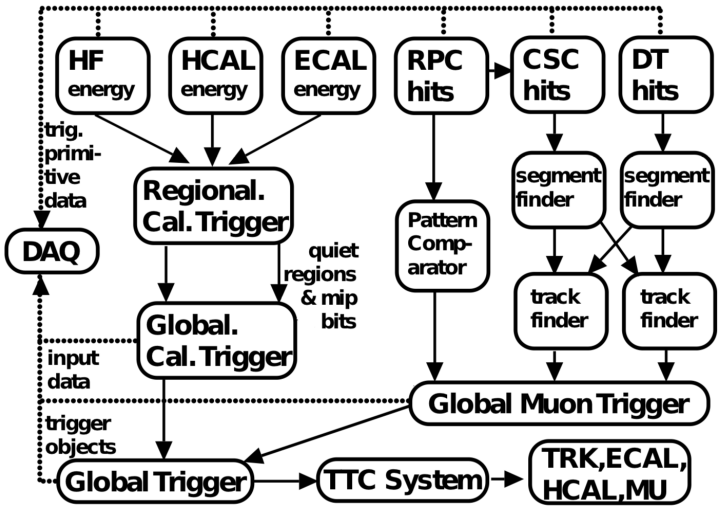

Figure 2.6: Overview of the CMS L1 trigger system. Data from the HF, HCAL, and ECAL are processed first regionally and then globally. Energy deposits from RPC, CSC, and DT are processed either via a pattern comparator or via a system of segment- and track-finders and sent onwards to a global muon trigger. The information from the global calorimeter and muon triggers are combined in a global trigger, which makes the final trigger decision. This decision is sent to the tracker (TRK), ECAL, HCAL or muon systems (MU) via the trigger, timing and control (TTC) system. The data acquisition system (DAQ) reads data from various subsystems for offline storage.

and transmit to the global muon trigger (GMT) up to 4 (CSCTF, DTTF) or 8 (RPC) candidates every bunch crossing. Each candidate is assigned a p_T and quality code as well as a (η, ϕ) position in the muon system. The GMT then merges muon candidates found by more than one system to eliminate a single candidate passing multiple-muon triggers (with several options on how to select p_T between the candidates). The GMT also performs a further quality assignment so that, at the final trigger stage, candidates can be discarded if their quality is low and they are reconstructed only by one muon track finder.

The GT is the final step of the CMS L1 trigger system and implements a menu of triggers, a set of selection requirements applied to the final list of objects (i.e., electrons/photons, muons, jets, or τ leptons), required by the algorithms of the HLT algorithms to meet the physics data-taking objectives. This menu includes trigger criteria ranging from simple single-object selections with E_{T} above a preset threshold to selections requiring coincidences of several objects with topological conditions among them.

The Level-1 trigger system mainly consists of customised hardware, such as application specific integrated circuits (ASICs), in order to ensure a fast processing of the data. Nevertheless, also programmable hardware, like field programmable gate arrays (FPGAs), is used. Until the response of the L1 trigger is returned, the entire information of the events is stored in pipelined memory given by the buffers of the frontend electronics. The time period from sending the data to the Level-1 trigger system until the response is received adds up to about 4 μs, where about 1 μs is reserved for the decision making in the L1 trigger system. Selected events are released for further processing, while the rejected events are dropped. At this stage, the event rate is reduced to less than 100 kHz. The data passed on by the front-end electronics is further merged, before it is transferred to

the CMS computing installations on the surface. There, an event-builder network collects the data of each event and distributes them to various processing units.

HLT

The second stage of data reduction is the HLT software running on each of these processing units. It is implemented in software running on a farm of commercial computers which includes about 16 000 CPU cores, and reduces the L1 output rate to the sustainable level for storage and physics analysis of about 1 kHz. The HLT software consists of a streamlined version of the offline reconstruction algorithms; it exploits the same sophisticated software used for offline reconstruction and analysis, optimised in order to comply with the strict time requirements of the online selection. The software follows a strategy of rejecting events as soon as possible. This is achieved by sequentially reconstructing analysis objects in different paths that together form a modular structure. At different stages of this reconstruction procedure, events are checked for selection criteria. Collision events passing this selection process are transferred to the CERN Tier-0 computing facility for further processing and storage.

2.2.8 Computational Infrastructure

The LHC experiments make use of a Tier-organized distributed system for the storage and analysis of recorded and simulated data, called the Worldwide LHC Computing Grid (WLCG) [50]. The WLCG consists of over 170 centres distributed across 41 countries. The Tier-0 is located at the main CERN site at Meyrin in Switzerland. The data recorded by the experiments are directly transferred to Tier-0 enabling a fast transfer of data through the enormous storage resources provided by the Tier-0. The main processing of the data is carried out in the distributed Tier-1 centres. These centres are spread out all over the world and connected via high-speed networks. The Tier-1 centres additionally provide a backup for the data stored at the Tier-0 centres. The over 160 Tier-2 centres provide a platform for data analysis performed by scientists all over the world.

Chapter 3

Analysis objects

In high energy physics, when particles collide together it is necessary to reconstruct the image of the collision for real data as well as for simulated data. In the second case, simulations involve all the physics phenomena that take place, from the proton-proton collision to the interaction of the produced particles with the materials. The Monte Carlo (MC) simulation mechanism is presented in Section 3.1. Final states of hadronic topologies, like the one studied in this work, typically involve jets, showers of hadrons produced by the strong interaction. In Section 3.2, a description of jets is provided. The identification of jets originating from bottom quarks is referred to as b-tagging and presented in Section 3.3. Jets can be also boosted when the p_T of the particle originating the jet is quite large. This implies that particles are clustered inside a cone with a proper radius parameter according to a specific algorithm which is presented in Section 3.4. Finally, it is not easy to identify jets and recognizing which particle originates them. Multivariate analysis techniques have been employed to perform this job, as well as to discriminate signal events from background ones. In Section 3.5, some of the most important multivariate analysis techniques which will be adopted in the following chapter are reported. Finally, the standard HEP tagger used by CMS is described in Section 3.6.

3.1 Monte Carlo Simulation

The structure of a proton-proton collision at the LHC needs to be reproduced by the MC event generators using the existing knowledge of SM and guesses on BSM. The understanding of the final state particles in proton-proton collisions is a very challenging problem. The simulation of a proton-proton collision by MC event generators consist of the following steps:

1. Hard process. It is defined by the collision of two beam constituents at a high momentum scale and consists of the most energetic final states. It is denoted as the central red blob in Fig. 3.1. This process involves large invariant momentum transfers and it is the first step of any simulation through MC event generators. The implementation is not straightforward since it involves non-perturbative calculations. According to the asymptotic freedom of QCD, hadrons interact weakly at high energies corresponding to a smaller coupling constant, α_S, so that the constituents of the hadron can be regarded as free particles. Whereas, at low energies the interaction becomes stronger as the α_S becomes larger and partons confine into hadrons. The high-energetic interactions, also called short-distance interactions, can be calculated perturbatively while in case of low-energy, long-distance, inter-

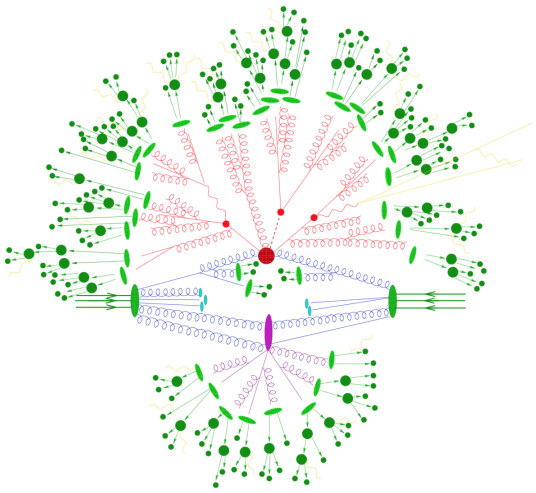

Figure 3.1: Representation of a pp collision at LHC.

actions this is not possible due to the large value of α_S. Therefore the so-called factorisation theorem brings a solution to this problem by resolving the short distance parton cross section from the long distance interactions. Accounting for the factorisation theorem for partons a and b, from hadrons 1 and 2, scattering to c and d partons, the following equation can be written

$$d\sigma^{h_1 h_2 \to cd} = \int_0^1 \mathrm{d}x_1 \int_0^1 \mathrm{d}x_2 \sum_{a,b} f_{a/h_1}(x_1, \mu_F^2) f_{b/h_2}(x_2, \mu_F^2) \mathrm{d}\hat{\sigma}^{ab \to cd}(\mu_R^2, \mu_F^2) \quad (3.1)$$

where $f_{a/h_i}(xi, \mu_F^2)$ is the parton distribution function (PDF) which gives the probability of finding a parton of flavour a with momentum fraction x_i of the hadron h_i at the energy scale μ_F. The parameter μ_F is the factorisation scale, which characterises the hard scattering and can be thought as the scale that separates the long- and short-distance interactions, while μ_R is the renormalisation scale, which is a scale used to fix the divergences of loop diagrams. The PDFs can not be obtained via perturbative QCD calculations, so they are computed by fitting the data from several experiments and many different processes. This is possible due to the fact that the PDFs are process-independent meaning that they are universal. They can be measured in one process and can be applied to other processes. Their evolution to any scale can be calculated by DGLAP evolution functions once they are measured in one scale. The hard interaction differential cross section for a and b scattering to c and d is denoted by $\mathrm{d}\hat{\sigma}^{ab \to cd}(\mu_R^2, \mu_F^2)$. This term contains only hard emissions above the factorisation scale μ_F and can be calculated by perturbative QCD.

2. Parton shower. The simulation of the proton-proton collision is followed by the parton shower. The partons carrying a colour charge can emit gluons (QCD radiation) and can also interact with each other emitting further gluons. This process is called parton shower, denoted by the red spiral tree structure surrounding the hard interaction. It evolves until the partons lose energy due to gluon emission and they go into the hadronisation phase.

3. Hadronization. In the process of partons losing their energy by QCD radiation, at

some energy level the interaction among the coloured partons become stronger, i.e. α_S becomes large, and they are bounded into colourless hadrons. This transition is called hadronization. Hadrons are the first experimental observables of the event generation in an event. The hadronization process roughly happens at an energy of 1 GeV where this energy depends on the hadronization model. The most common hadronization models are the cluster model and the Lund string model. The transition of partons to hadrons are denoted as the light green blobs in Fig. 3.1.

4. Hadron decays. Most of the hadrons produced in the previous step are unstable and they subsequently decay, until a set of particles is obtained that can be considered stable on time scales relevant to the given measurement. These stable hadrons are the final observables detected. Therefore the decay modelling has an important impact on the final state yields and spectra. The hadron decays are shown as the dark green blobs in Fig. 3.1.

5. Secondary interactions. Up to this point, the interactions of the partons that are not coming from the hard collisions have not been considered. At first approximation, it can be assumed that these partons do not interact and just fly away undisturbed. But actually this is not the case and partons not coming from the hard collision can also interact with each other. These interactions are called multiparton interactions. In a proton-proton collision the primary spectator partons (beam remnants) can split or emit gluons and hadronize. In addition, the initial- and final-state gluon radiations not connected to the hard collisions and the multiparton interactions are called underlying event. The illustration of a secondary interaction is denoted as a purple blob in Fig. 3.1.

3.2 Jets

Jets are collimated showers of particles originated from quarks and gluons produced in collision events, due to the special properties of the strong interaction. In order to deduce the properties of the initially produced strongly interacting particle, all of its secondary particles are combined. However, in most cases the grouping of all reconstructed particles is ambiguous. For this reason, clusters of particles are formed based on special rules given by jet algorithms. The reconstruction of the original particle is complicated by the presence of additional particles in the event stemming from other sources. Especially in case of hadron colliders, where a very large fraction of interactions are based on QCD processes, a huge multiplicity of additional hadrons can be produced by the underlying event or additional proton-proton collisions. Pile-up effects due to multiple collisions in the same bunch crossing can be partly mitigated by identifying reconstructed particles stemming from additional proton-proton collisions and removing them from the set of reconstructed particles.

3.2.1 Jet reconstruction

The properties of quarks and gluons produced in a collision event are deduced by analysing the collection of particles resulting from the hadronization process. Dedicated algorithms provide a set of rules for collecting these particles and merging them into a single object. Jet algorithms can be applied on a variety of different input objects: partonic calculations, the output of parton-shower simulations, measured quantities like calorimeter deposits, or reconstructed particles. In this analysis, the particles obtained by the CMS particle-flow (PF) event reconstruction [51] serve as input to the jet algorithms applied. This set of

input objects is cleaned from pile-up particles using a dedicated procedure. There is a huge variety of different jet algorithms. However, in most cases, there is no single optimal way for clustering particles into jets and the choice of a jet algorithm is ambiguous. Still, an important property determining the quality of a jet algorithm is the infrared and collinear (IRC) safety. Jet algorithms are considered infrared or collinear safe if the radiation of a soft particle or a collinear splitting of partons does not change the outcome. Jet algorithms can be grouped into two major classes: cone algorithms and sequential recombination algorithms.

- Cone algorithms feature a top-down approach relying on the approximation that QCD branching and hadronization leave the energy-flow unchanged. Generally, the procedure is based on clustering all particles in a cone of a given size. However, most cone algorithms suffer from IRC unsafety, which is one of the reasons, why they are not considered by CMS.

- Sequential recombination algorithms feature a bottom-up approach relying on the iterative combination of the closest particles based on a specific distance measure. An advantage of these algorithms is their clustering sequence, which in some cases resembles QCD branching. This property is especially important for the analysis of the substructure of jets. Another important characteristic of these algorithms is their IRC safety.

Three sequential recombination algorithms represent the jet algorithms most commonly used at the LHC:

- the k_T algorithm [52, 53],
- the Cambridge/Aachen algorithm [54, 55],
- the anti-k_T algorithm [56].

The three algorithms follow the same procedure and only differ in the definition of the distance measure between two particles. The single steps of the algorithms are the following:

1. Compute the particle-particle distances d_{ij} for a pair of input particles i and j with transverse momenta $p_{T,i}$ and $p_{T,j}$

$$d_{ij} = \min(p_{T,i}^{2p}, p_{T,j}^{2p}) \frac{\Delta R_{ij}^2}{R^2} \tag{3.2}$$

   and the particle-beam distances for all input particles i with respect to the beam B

$$d_{iB} = p_{T,i}^{2p} \tag{3.3}$$

   The value of ΔR_{ij} is the distance in the η-ϕ plane, defined as:

$$\Delta R_{ij}^2 = (\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2 \tag{3.4}$$

   The cone-size parameter R defines at which angular distance particles are still combined or declared as final jets. Accordingly, it can be interpreted as the radius of the jet in the η-ϕ plane. The different algorithms are defined by the choice of the parameter p:

   - $p = 1$: k_T algorithm
   - $p = 0$: Cambridge/Aachen algorithm

- $p = -1$: anti-k_T algorithm

2. Compute the minimum among all particle-particle and particle-beam distances.

3. If the minimum is given by a particle-particle distance, the particle i and j are combined into a single object by adding their four-vector momenta. Else if the minimum is given by a particle-beam distance, particle i is declared a jet and removed from the set of particles. In both cases, the algorithm continues with step 1.

4. If this step is reached, no particles remain in the set of particles and all final jets are found. Accordingly, the clustering process is stopped.

In the Cambridge/Aachen algorithm, which uses $p = 0$, the particle-particle distance reduces to a term with the only angular distance, and the particle-beam distance reduces to $d_{iB} = 1$. Hence, the clustering is fully independent of the momenta of the particles and only relies on their angular distances. This results in a clustering sequence resembling the QCD branching at different angular scales. Due to this property, the Cambridge/Aachen algorithm is well suited for the investigation of jet substructure. The jets resulting from the Cambridge/Aachen algorithm feature non-circular geometries.
The clustering procedure of the k_T algorithm relies on the transverse momenta of the particles in addition to their angular distances. Due to its distance measure, it favours combinations that involve soft particles. The clustering sequence obtained by the k_T algorithm resembles the QCD branching at different energy scales. For this reason, the k_T algorithm is also suited for the investigation of jet substructure. As for the Cambridge/Aachen algorithm, the geometries of jets clustered with the k_T algorithm are typically non-circular.
The anti-k_T algorithm also depends on the transverse momenta of the particles in addition to their angular distances. Nevertheless, the clustering behaviour of the anti-k_T algorithm is the opposite of the k_T algorithm as it favours the combination of hard particles. The clustering sequence does not resemble QCD branching in any way. For this reason, the anti-k_T algorithm is not suited for the investigation of jet substructure. In contrast to the other two jet algorithms, the jets resulting from the anti-k_T algorithm feature circular geometries, which is a reason why this algorithm is in some cases preferred over the other ones.
The three algorithms are all implemented in the FASTJET package [57] which uses reconstructed PF candidates as input for the clustering of the jets. In this analysis jets are clustered using the anti-k_T algorithm and distinguished in two categories according to the cone-size of the algorithm: AK8 when $R = 0.8$, AK4 for $R = 0.4$.

3.3 b-tagging

Jets originating from b quarks can be distinguished by the presence of b-hadrons, i.e. hadrons containing a valence b quark. These hadrons show several characteristic features. b-hadrons are the heaviest hadrons with a rest mass of more than 5 GeV. They decay via the weak interaction into hadrons containing c-quarks, a decay that is suppressed by a small CKM matrix element, which results in a long lifetime of $\tau = 0.5$ mm/c. Their decay products usually include multiple charged leptons and in about 20% of the decays an electron or muon. Furthermore, in the fragmentation of b quarks most of the energy is passed on to the b-hadron, so that they carry a large fraction of the total jet momentum. They can thus have a large Lorentz boost and increased lifetimes of several mm/c. The displaced decay of the hadron gives rise to a secondary vertex (SV) that has a significant distance from the primary vertex (PV). Tracks originating from

charged particles produced in this decay tend to have large impact parameters, which is the distance between the primary vertex and the tracks at their points of closest approach. A sketch of a b-hadron decay is shown in Fig. 3.2. Jets arising from the hadronization

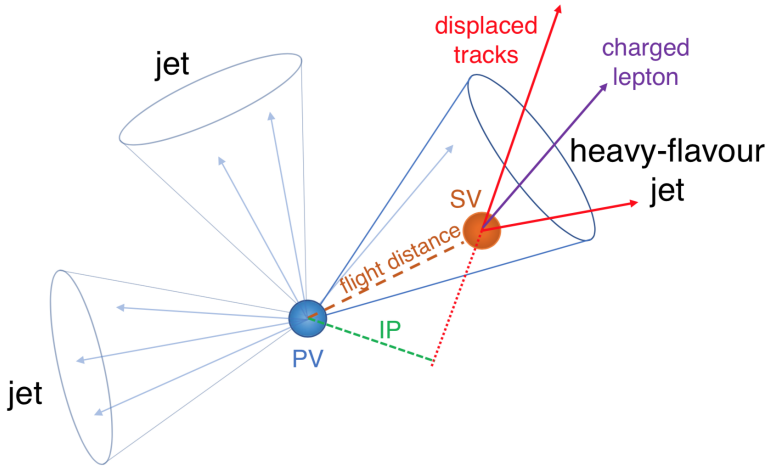

Figure 3.2: Diagram showing the common principle of identification of jets initiated by b-hadron decays.

of bottom quarks (b-jets) characterise many physics processes, such as the decay of top quarks, the Higgs boson, and several particles predicted by supersymmetric models. For this reason, the ability to identify b-jets accurately is crucial in reducing the dominant background to these channels, from processes involving jets from gluons (g) and light quarks (u, d, s), and from c-quark fragmentation. The CMS detector is well suited for the task of b-jet identification (b-jet tagging), thanks to its precise charged-particle tracking and robust lepton identification systems.

3.3.1 b-tagging algorithms

A variety of reconstructed objects (such as tracks, vertices and identified leptons) can be used to build observables that discriminate between b- and light-parton jets [58]. Algorithms for heavy-flavour jet identification use variables connected to the properties of heavy-flavour hadrons present in jets resulting from the radiation and hadronization of b or c quarks. For instance, the lifetime of hadrons containing b quarks is of the order of 1.5 ps, while the lifetime of c hadrons is 1 ps or less. This leads to typical displacements of a few mm to one cm for b hadrons, depending on their momentum, thus giving rise to displaced tracks from which a SV may be reconstructed. The displacement of tracks with respect to the primary vertex is characterised by their impact parameter, which is defined as follows. The vector pointing from the primary vertex to the point of closest approach is referred to as the impact parameter vector. The impact parameter value can be defined in three spatial dimensions (3D) or in the plane transverse to the beam line (2D). The longitudinal impact parameter is defined in one dimension, along the beam line. The impact parameter is defined to be positive or negative, with a positive sign indicating that the track is produced "upstream". This means that the angle between the impact parameter vector and the jet axis is smaller than $\pi/2$, where the jet axis is defined by the primary vertex and the direction of the jet momentum. In addition, b and c quarks have a larger mass and harder fragmentation compared to the light quarks and massless gluons.

As a result, the decay products of the heavy-flavour hadron have, on average, a larger p_T relative to the jet axis than the other jet constituents. In approximately 20% (10%) of the cases, a muon or electron is present in the decay chain of a heavy b (c) hadron. Hence, apart from the properties of the reconstructed secondary vertex or displaced tracks, the presence of charged leptons is also exploited for heavy-flavour jet identification techniques and for measuring their performance in data. A single observable is used by several simple and robust algorithms, while others combine a few of these objects, in order to achieve a higher discrimination power. In either cases, these CMS algorithms provide a single discriminator value for each jet. The minimum thresholds on these discriminator values define loose ("L"), medium ("M") and tight ("T") operating points with a misidentification probability for light-parton jets of close to 10%, 1% and 0.1%, respectively, at an average jet p_T of about 80 GeV.

3.3.2 The CSV algorithm

The presence of a secondary vertex, and the kinematic variables associated with it, can be used to discriminate between b and non-b jets. The most significant variables used for the discrimination are the flight distance and direction, using the vector between primary and secondary vertices. The remaining variables are related to some properties of the system of associated secondary tracks, such as the multiplicity, the mass, or the energy. In order to enhance the "b-purity", secondary vertex candidates should meet the following requirements:

- secondary vertices must share less than 65% of their associated tracks with the primary vertex and the significance of the radial distance between the two vertices has to be $> 3\sigma$, with σ being the uncertainty on the distance;
- secondary vertex candidates with a radial distance of more than 2.5 cm with respect to the primary vertex, with masses compatible with the mass of K^0, or exceeding 6.5 GeV are rejected, thus reducing the contamination by vertices corresponding to both interactions of particles with the detector material, and decays of long-lived mesons;
- the flight direction of each candidate has to lie within a cone of $\Delta R = 0.5$ around the jet direction.

The Combined Secondary Vertex (CSV) algorithm [59] involves the use of secondary vertices, together with track-based lifetime information. By using these additional variables, the CSV algorithm provides discrimination also in cases when no secondary vertices are found, increasing the maximum efficiency with respect to the so-called "Simple Secondary Vertex" algorithms - these using only the flight distance as discriminating variable. In many cases, tracks with an impact parameter significance SIP - that is the ratio of the IP to its estimated uncertainty - that is > 2 can be combined into a "pseudo-vertex", allowing the computation of a subset of secondary-vertex-based quantities even without an actual vertex fit. Finally, when even this is not possible, a "no vertex" category reverts to track-based variables and the discrimination is conducted in a way similar to that of the track-based algorithms. The CSV algorithm uses a set of variables with high discriminating power and low correlations, such as: the vertex category (which can be real, pseudo, or no vertex), the flight distance significance in the transverse plane, the vertex mass (i.e. the invariant mass of the particles associated to the vertex), the number of tracks at the vertex, the pseudorapidities of the tracks at the vertex with respect to the jet axis, the mass of the secondary vertex with the smallest uncertainty on its flight distance, the number of tracks from SV, the ratio of the energy carried by tracks at the

vertex with respect to all tracks in the jet, the 3D IP significances of the first four tracks. Then, two likelihood ratios are built from these variables and used to discriminate between b- and c-jets, and between b- and light-parton jets. Finally, they are combined with prior weights of 0.25 and 0.75, respectively. The CSV algorithm has evolved into the CSVv2 (CSV version 2) algorithm [58] in Run 2. Just like the CSV, the CSVv2 is based on secondary vertex and track-based lifetime information. The training is performed on inclusive multijet events in three independent vertex categories:

- RecoVertex: The jet contains one or more secondary vertices.
- PseudoVertex: No secondary vertex is found in the jet but a set of at least two tracks with a 2D impact parameter significance above two and a combined invariant mass at least 50 MeV away from the K_0^S mass are found. Since there is no real secondary vertex reconstruction, no fit is performed, resulting in a reduced number of variables.
- NoVertex: Containing jets not assigned to one of the previous two categories. Only the information of the selected tracks is used.

The following discriminating variables are combined in the algorithm:

- The "SV 2D flight distance significance", defined as the 2D flight distance significance of the secondary vertex with the smallest uncertainty on its flight distance for jets in the RecoVertex category.
- The "number of SV", defined as the number of secondary vertices for jets in the RecoVertex category.
- The "track η_{rel}", defined as the pseudorapidity of the track relative to the jet axis for the track with the highest 2D impact parameter significance for jets in the RecoVertex and PseudoVertex categories.
- The "correctedSVmass", defined as the corrected mass of the secondary vertex with the smallest uncertainty on its flight distance for jets in the RecoVertex category or the invariant mass obtained from the total summed four-momentum vector of the selected tracks for jets in the PseudoVertex category.
- The "number of tracks from SV", defined as the number of tracks associated with the secondary vertex for jets in the RecoVertex category or the number of selected tracks for jets in the PseudoVertex category.
- The "SV energy ratio", defined as the energy of the secondary vertex with the smallest uncertainty on its flight distance divided by the energy of the total summed four-momentum vector of the selected tracks.
- The "ΔR(SV, jet)", defined as the ΔR between the flight direction of the secondary vertex with the smallest uncertainty on its flight distance and the jet axis for jets in the RecoVertex category, or the ΔR between the total summed four-momentum vector of the selected tracks for jets in the PseudoVertex category.
- The "3D IP significance of the first four tracks", defined as the signed 3D impact parameter significances of the four tracks with the highest 2D impact parameter significance.
- The "track $p_{T,\text{rel}}$", defined as the track p_T relative to the jet axis, i.e. the track momentum perpendicular to the jet axis, for the track with the highest 2D impact parameter significance.

- The "ΔR(track, jet)", defined as the ΔR between the track and the jet axis for the track with the highest 2D impact parameter significance.
- The "track $p_{T,\text{rel}}$ ratio", defined as the track p_T relative to the jet axis divided by the magnitude of the track momentum vector for the track with the highest 2D impact parameter significance.
- The "track distance", defined as the distance between the track and the jet axis at their point of closest approach for the track with the highest 2D impact parameter significance.
- The track decay length, defined as the distance between the primary vertex and the track at the point of closest approach between the track and the jet axis for the track with the highest 2D impact parameter significance.
- The "summed tracks E_T ratio", defined as the transverse energy of the total summed four-momentum vector of the selected tracks divided by the transverse energy of the jet.
- The "ΔR(summed tracks, jet)", defined as the ΔR between the total summed four-momentum vector of the tracks and the jet axis.
- The "first track 2D IP significance above c threshold", defined as the 2D impact parameter significance of the first track that raises the combined invariant mass of the tracks above 1.5 GeV. This track is obtained by summing the four-momenta of the tracks adding one track at the time. Every time a track is added, the total four-momentum vector is computed.
- The number of selected tracks.
- The jet p_T and η.

The discriminating variables in each vertex category are combined into a neural network, specifically a feed-forward multilayer perceptron with one hidden layer. The number of nodes in the hidden layer is different for the three different vertex categories and is set to twice the number of input variables. The discriminator values of the three vertex categories are combined with a likelihood ratio taking into account the fraction of jets of each flavour expected in $\mathrm{t\bar{t}}$ events. The fraction of jets of each flavour is obtained as a function of the jet p_T and $|\eta|$, using 19 exclusive bins in total. Two dedicated trainings are performed, one with c jets, and one with light-flavour jets as background. The final discriminator value is a linear combination of the output of these two trainings with relative weights of 1 and 3 for the output of the network trained against c and light-flavour jets, respectively. The value of these relative weights is inspired by $\mathrm{t\bar{t}}$ events where one of the two W bosons decays into quarks and the other into leptons, and provides the best performance for a wide variety of physics topologies compared to alternative relative weights. The more refined CSVv2 algorithm provides a better efficiency at the same working points.

3.4 Boosted jets

Massive particles with large transverse momenta decaying into strongly interacting particles form boosted objects called boosted jets. When decaying, these particles pass their momentum to the decay products, which form collimated showers of hadrons. This type of topologies is mostly beyond being resolvable with standard jet reconstruction algorithms.

Yet, such configurations bear the advantage that all decay products are locally accumulated instead of being spread out in all directions. Specialised clustering and substructure algorithms do not only allow the analysis of boosted objects, but also make use of the collimated decay products in the reconstruction of the massive particles. The application of these dedicated algorithms results in large reconstruction efficiencies for massive particles with large transverse momenta. In most cases, the reconstruction efficiencies achieved exceed the ones reached in the reconstruction of fully resolved events. The main reason is that resolved events involve combinatorial permutations, which are reduced in the boosted-object reconstruction. In the resolved reconstruction, the ambiguous assignment of jets to the decay products of massive particles leads to a huge source of incorrect identification possibilities.

3.4.1 Boosted jets clustering

The clustering of boosted jets aims at merging all decay products of boosted massive particles into a single object. Boosted jets can be clustered with the aforementioned Cambridge/Aachen algorithm, which clusters objects solely based on their angular distance. The corresponding clustering sequence resembles the sequential ordering of the parton splitting process, which is a crucial feature for obtaining meaningful results by the declustering algorithms. A similar behaviour is provided by the k_T algorithm but not by the anti-k_T algorithm. The Cambridge/Aachen algorithm has some advantages in fat-jet clustering and substructure investigation compared to the k_T algorithm, like the fact that the fat-jet mass is less prone to soft radiation.
An important parameter in the clustering of boosted jets is the distance parameter R. This parameter needs to be chosen large enough to cluster all decay products of a boosted massive particle into a single jet. The distances between the decay products depend on the type of the decay, the mass, and the transverse momentum of the massive particle. A simple example is the two-body decay of a Higgs boson into two bottom quarks. In this case, the angular distance between the two bottom quarks is approximately given by

$$\Delta R_{bb} \simeq \frac{1}{\sqrt{z(1-z)}} \frac{m_H}{p_T} \quad \text{with}\, p_T \gg m_H. \tag{3.5}$$

In this equation z and $1-z$ are the momentum fractions of the two bottom quarks. For fixed z, ΔR_{bb} scales as the inverse of p_T. This characteristic can be observed in a plot which shows the angular distance between the bottom quarks from Higgs boson decay in simulated $\mathrm{t\bar{t}H}$ events as a function of the Higgs boson transverse momentum.

3.4.2 Substructure Algorithms

The study of the internal structure of hadronic jets has become in recent years a very active area of research in particle physics. Jet substructure techniques are increasingly used in experimental analyses by the LHC collaborations, both in the context of searching for new physics and for SM measurements. On the theory side, the quest for a deeper understanding of jet substructure algorithms has contributed to a renewed interest in all-order calculations in QCD [60].
In addition to particles stemming from the hard interaction of the pp collision, particles originating from various other sources, like pile-up, the underlying event, and initial-state radiation, can be found in the final state. Even though a major part of this contamination is removed by the selection and cleaning steps, boosted jets remain prone to these effects because of their large cone size. Impurities clustered into boosted jets hide the distinctive features of massive-particle decays, as the distributions of reconstructed observables

are washed out. In order to obtain more information about the process underlying the particles clustered into the fat jet, substructure algorithms are applied. These algorithms exploit the fact that when a boosted massive object decays into partons, all the partons typically carry a sizeable fraction of the initial jet transverse momentum, resulting in multiple hard cores in the jet. Conversely, quark and gluon jets are dominated by the radiation of soft gluons, and are therefore mainly single-core jets. "Prong finders" look for multiple hard cores in a jet, hence reducing the contamination from standard QCD jets. This is often used to characterise the boosted jets in terms of their "pronginess", i.e. to their expected number of hard cores: QCD jets would be 1-prong objects, W/Z/H jets would be two-pronged, boosted top jets would be three-pronged, an elusive new resonance with a boosted decay into two Higgs bosons, both decaying to a $b\bar{b}$ pair would be a 4-prong object. In this way, the contamination is removed from the jet and the substructure of the fat jet can be extracted.

Mass-Drop Declustering

Mass-drop declustering [61] algorithm was originally proposed as a tagger to isolate boosted Higgs bosons decaying to $b\bar{b}$ pairs from the QCD background. It decreases the invariant mass of the two individual subjets with respect to the mother jet, when splitting the decay products of a massive particle with the Cambridge/Aachen algorithm. The first step in each iteration is splitting the mother jet j into two daughter subjets j_1 and j_2 by undoing the last step of the clustering history. The two subjets are labelled according to their invariant mass, where the more massive subjet is denoted by j_1 and the remaining one by j_2. The second step of each iteration is to check if the mass-drop criterion,

$$m_{j_1} < \mu m_j, \tag{3.6}$$

is fulfilled. The parameter μ represents the mass-drop threshold as a fraction of the invariant mass of the mother jet j. Its value is chosen based on the mass and the decay type of the massive particle, for which the reconstruction is optimized. If the equation is not fulfilled, subjet j_2 is considered soft radiation, not originating from the massive-particle decay, and is discarded. In this case, subjet j_1 is declared the mother jet j for the next iteration and the declustering is continued. One can also set the splitting of the two jets to be sufficiently symmetric,

$$\min\left(p_{T,j_1}^2, p_{T,j_2}^2\right) \Delta R_{j_1,j_2} > y_{\text{cut}}\ m_j^2, \tag{3.7}$$

where y_{cut} acts as the symmetry cut parameter. Based on the application of the algorithm, the declustering is continued or stopped if the criteria are fulfilled. The two conditions imposed by the mass-drop tagger exploit the fundamental properties for tagging two-pronged boosted objects: the symmetry cut requires that one indeed finds two hard prongs and the mass-drop condition imposes that one goes from a massive boson jet to two jets originated from massless QCD partons. Although it was originally introduced as a tagger, the mass-drop tagger also iteratively removes soft radiation at the outskirts of the jet, hence reducing the pileup/UE contamination.

Soft-Drop Declustering

Soft-drop declustering [62] algorithm is a jet substructure technique which aims at recursively removing wide-angle soft radiation from a jet. In the first step of each iteration, the mother jet j is split into two daughter subjets j_1 and j_2 based on the last step of the clustering history. The two subjets are labelled according to their transverse momentum,

with the harder one denoted as j_1 and the softer as j_2. The second step of each iteration is to check if the "soft-drop condition",

$$\frac{p_{T,j_2}}{p_{T,j_1} + p_{T,j_2}} > z_{cut}\Big(\frac{\Delta R_{j_1,j_2}}{R}\Big)^{\beta}, \tag{3.8}$$

is fulfilled. The parameter R represents the cone size used for the clustering of the fat jet. This equation basically depends on two parameters, z_{cut} and β. z_{cut} is the soft-drop threshold, which determines the transverse momentum of particles to be removed and plays an equivalent role as the mass-drop threshold μ in mass-drop declustering. The β exponent determines the influence of the angular distance of the subjets. For $\beta \to \infty$, the last term of Eq. 3.8 becomes zero, as $\Delta R_{j_1,j_2} < R$, and the algorithm returns the ungroomed jet. The case $\beta = 0$ results in a behaviour equivalent to mass-drop declustering. For $\beta > 0$, wide-angle soft radiation is removed, while keeping some of the soft-collinear radiation controlled by the β parameter. This configuration is called "grooming mode", meaning that it changes the constituents of a jet without affecting the overall jet production cross section. It is infrared and collinear safe even for jets with only one constituent. For $\beta < 0$, two separated hard subjets are required to satisfy the soft-drop condition. This configuration is therefore called the "tagger mode" since it vetoes jets that do not have two well-separated hard prongs. In this mode, soft-drop declustering can remove both soft and collinear radiation. As for the mass-drop declustering, the softer of the two subjets is discarded if the soft-drop condition is not fulfilled. In this case, the subjet j_1 is declared the mother jet j for the next iteration and the declustering is continued. If Eq. 3.8 is fulfilled, on the other hand, depending on the application, the soft-drop declustering is stopped or continued with both subjets.

3.4.3 Jet Grooming

Jet grooming represents a further way of cleaning contamination from boosted jets and uncovering the underlying substructure. The algorithms in this category of substructure algorithms rely on reclustering the constituents of the fat jet with a different clustering configuration and applying additional criteria. Unlike the declustering algorithms, which are adapted to the hypothesis of a massive-particle decay, the jet grooming algorithms are completely independent of information on the massive particle. In the following, three different algorithms from this category of substructure algorithms are described.

Filtering and Trimming

Filtering [61] and trimming [63] are grooming techniques, which aim at resolving the fat jet at a finer angular scale. Both algorithms start by reclustering the constituents of the fat jet with a sequential recombination algorithm and a small cone-size parameter. A typical choice of the cone-size parameter used for filtering and trimming is $R = 0.3$. The reclustering of the fat-jet constituents results in a number of subjets determined by its substructure. While filtering retains only the N subjets with the largest transverse momentum for further analysis, trimming discards all subjets below a chosen transverse momentum threshold. In this way, the filtering and trimming methods remove soft radiation in form of subjets with small transverse momenta. The degree of grooming is steered by the grooming parameters, the subjet multiplicity N for filtering and the transverse momentum threshold for trimming.

Pruning

Pruning [64, 65] is a technique designed for removing soft and wide-angle radiation. Just like filtering and trimming, pruning is based on the reclustering of the fat-jet constituents. Yet, unlike these algorithms, pruning does not necessarily aim at finding subjets. Instead of discarding soft subjets, pruning removes contamination by vetoing soft and large-angle recombinations during reclustering. The requirements for vetoing recombination of two constituents j_1 and j_2 with $p_{T,j_1} > p_{T,j_2}$ to a resulting jet j are

$$\frac{p_{T,j_2}}{p_{T,j}} < z_{cut} \quad \text{and} \quad \Delta R_{j_1,j_2} > D_{cut}. \tag{3.9}$$

The pruning method is steered by two parameters. The parameter z_{cut} represents a lower threshold for the transverse momentum of the softer constituent with respect to the combined jet. Hence, it determines how soft the constituents may be in order to be recombined. The parameter D_{cut} determines the minimum angular distance for recombination to be pruned. If both requirements are fulfilled, the constituents are not combined and the softer one is discarded. In all other cases, the two constituents are merged. If the pruning is performed with the Cambridge/Aachen algorithm, a typical choice for the transverse-momentum threshold is $z_{cut} = 0.1$. The application of the k_T-jet-clustering algorithm requires slightly larger values, e.g. $z_{cut} = 0.15$, to achieve similar performance. This fact can be explained by the transverse momentum ordering of the recombinations in the k_T-clustering process. Concerning the parameter determining the minimum angular distance for pruning D_{cut}, too small values should be avoided as this would cause the pruning away of particles stemming from the original massive particle. Removing such particles would result in a degradation of the scale of the reconstructed particles observables, as fractions of the energy of the initially produced particle would be dismissed. Pruning with too large values of D_{cut}, on the other hand, would not take full advantage of the procedure, as particles from other sources would not be efficiently removed. A typical choice is $D_{cut} = 0.5$.

3.4.4 N-Subjettiness

N-subjettiness is a jet-shape procedure that aims to discriminate jets according to the number N of subjets they are made of. It is an inclusive jet-shape variable investigating the energy-flow properties of fat jets. Unlike the substructure algorithms previously presented, N-subjettiness is only based on the constituents of the fat jet and does not necessarily depend on any clustering algorithm. N-subjettiness can be interpreted as a form of counting the number of hard subjets inside the fat jet by calculating the deviation of the energy flow from N subjet axes. It is calculated as the sum of the minimum angular distances of all $N_{\text{particles}}$ to the N subjet axes weighted by their transverse momentum,

$$\begin{aligned} \tau_N = \frac{1}{d_0} \sum_{i}^{N_{\text{particles}}} p_{T,i} \min\{\Delta R_{1,i}, \Delta R_{2,i}, ..., \Delta R_{N,i}\} \\ d_0 = \sum_i p_{T,i} R \end{aligned} \tag{3.10}$$

where R represents the cone size used for fat-jet clustering. Eq. 3.10 is linear in the particles transverse momenta, which causes the results to be infrared and collinear safe. In cases with $\tau_N \simeq 0$, all of the fat-jet constituents are aligned with the N subjet axes. Hence, the fat jet features N or fewer hard subjets. The other extreme, $\tau_N \gg 0$, implies that there is a large fraction of the constituents which lie away from the N subjet axes. Accordingly,

the fat jet features at least $N+1$ hard subjets. Due to varying degrees of contamination, the absolute value of N-subjettiness is biased for each fat jet individually. For this reason, the ratio of successive values of N-subjettiness τ_N/τ_{N-1} is better suited for discriminating between different hard subjet multiplicities. The ratio τ_2/τ_1, for example, is a well-performing variable for the identification of two-prong decays, as they appear in hadronic W boson and Higgs boson decays. The fraction τ_3/τ_2, on the other hand, is well suited to identify three-prong decays. Examples are hadronic top quarks decays. A main issue when calculating N-subjettiness values is finding the directions of the N subjets axes. An optimal approach would be the minimisation of τ_N over all possible subjet directions. In this case, the values of N-subjettiness would be strictly decreasing with increasing N. However, this approach is computationally intensive. A more practical way of finding the directions of the N subjet axes is reclustering the fat-jet constituents with the k_T-algorithm. For this approach, the clustering is stopped as soon as exactly N subjets are clustered. Furthermore, energy-correlation functions (ECFs) achieve essentially the same objective than N-subjettiness without requiring the selection of N reference axes. Compared to N-subjettiness, ECFs have the advantage of not requiring a potentially delicate choice of reference axes [66, 67].

3.5 Multivariate analysis

The search for $t\bar{t}H$ production is very challenging. An overwhelming number of background events featuring a signature very similar to the one of $t\bar{t}H$ production complicates the extraction of signal events. In order to isolate these events, a machine-learning approach, which exploits different kinematic properties in form of different variables and the correlations between them, has to be applied. Different machine-learning techniques have been tested in this work: MultiLayerPerceptron (MLP), Boosted Decision Trees (BDT), Fisher discriminant. These MultiVariate Analysis (MVA) methods combine the signal and background separation abilities of a set of variables into a single observable. The construction of these observables is based on supervised learning, which aims at an optimal separation of signal from background. The supervised learning approach makes use of datasets with well-known properties, in order to train the MVA method. Those methods are implemented in the TMVA ROOT package [68] and will be used in the next chapter for event classification.

3.5.1 Fisher discriminant

An event selection based on Fisher discriminants is performed in a transformed variable space with zero linear correlations, by distinguishing the mean values of the signal and background distributions. The linear discriminant analysis determines an axis in the (correlated) hyperspace of the input variables such that, when projecting the output classes (signal and background) upon this axis, they are pushed as far as possible away from each other, while events of a same class are confined in a close vicinity. The linearity property of this classifier is reflected in the metric with which "far apart" and "close vicinity" are determined: the covariance matrix of the discriminating variable space. The classification of the events in signal and background classes relies on the following characteristics: the overall sample means $\bar{x}_k$ for each input variable $k = 1, ..., n_{\text{var}}$, the class-specific sample means $\bar{x}_{S(B),k}$, and total covariance matrix C of the sample. The covariance matrix can be decomposed into the sum of a within- (W) and a between-class matrix (B). The within-class matrix describes the dispersion of events relative to the

means of their own class and it is given by

$$W_{k\ell} = \sum_{U=S,B} \langle x_{U,k} - \bar{x}_{U,k} \rangle \langle x_{U,\ell} - \bar{x}_{U,\ell} \rangle = C_{S,k\ell} + C_{B,k\ell}, \tag{3.11}$$

where C_S is the covariance matrix of the signal sample and C_B is the covariance matrix of the background sample. The between-class matrix describes the dispersion of events relative to the overall sample means and it is given by

$$B_{k\ell} = \frac{1}{2} \sum_{U=S,B} (\bar{x}_{U,k} - \bar{x}_k)(\bar{x}_{U,\ell} - \bar{x}_\ell), \tag{3.12}$$

where $\bar{x}_{S,k}$ is the average of variable x_k for the signal sample and $\bar{x}_{B,k}$ is the average of variable x_k for the background sample and $\bar{x}_k$ denotes the average for the entire sample. The Fisher coefficients, F_k, are then given by

$$F_k = \frac{\sqrt{N_S N_B}}{N_S + N_B} \sum_{\ell=1}^{n_{\text{var}}} W_{k\ell}^{-1} (\bar{x}_{S,\ell} - \bar{x}_{B,\ell}), \tag{3.13}$$

where N_S is the number of signal events and N_B is the number of background events in the training sample. The Fisher discriminant $y_{\text{Fi}}(i)$ for event i is given by

$$y_{\text{Fi}}(i) = F_0 + \sum_{k=1}^{n_{\text{var}}} F_k x_k(i). \tag{3.14}$$

The offset F_0 centers the sample mean $\bar{y}_{\text{Fi}}$ of all $N_S + N_B$ events at zero.

3.5.2 Neural networks

An Artificial Neural Network (ANN) is most generally speaking any simulated collection of interconnected neurons, with each neuron producing a certain response at a given set of input signals. By applying an external signal to some (input) neurons the network is put into a defined state that can be measured from the response of one or several (output) neurons. One can therefore view the neural network as a mapping from a space of input variables x_1, ..., $x_{n_{\text{var}}}$ onto a one-dimensional (e.g. in case of a signal-versus-background discrimination problem) or multi-dimensional space of output variables y_1, ..., $y_{m_{\text{var}}}$. The mapping is nonlinear if at least one neuron has a nonlinear response to its input.

Multilayer Perceptron

While in principle a neural network with n neurons can have n^2 directional connections, the complexity can be reduced by organising the neurons in layers and only allowing direct connections from a given layer to the following layer. This kind of neural network is termed multi-layer perceptron. The first layer of a multilayer perceptron is the input layer, the last one is the output layer, and all the others are hidden layers. For a classification problem with n_{var} input variables the input layer consists of n_{var} neurons that hold the input values, x_1, ..., $x_{n_{\text{var}}}$, and one neuron in the output layer that holds the output variable, the neural net estimator y_{ANN}.
In ANN it is common to refer to the neuron response function ρ which maps the neuron input i_1, ..., i_n onto the neuron output. Often it can be separated into a $R^n \to R$ synapse function κ, and a $R \to R$ neuron activation function α, so that ρ is the composition of the α and κ functions, $\rho = \alpha \circ \kappa$. The functions κ can be in different forms: sum,

sum of squares, sum of absolutes; the function α can be: linear, sigmoid, tanh or radial. When building a network two rules should be kept in mind. The first is the theorem by Weierstrass, which if applied to neural nets, ascertains that for a multilayer perceptron a single hidden layer is sufficient to approximate a given continuous correlation function to any precision, provided that a sufficiently large number of neurons is used in the hidden layer. If the available computing power and the size of the training data sample suffice, one can increase the number of neurons in the hidden layer until the optimal performance is reached. The same performance can likely be achieved with a network of more than one hidden layer and a potentially much smaller total number of hidden neurons. This would lead to a shorter training time and a more robust network. The most common algorithm for adjusting the weights that optimise the classification performance of a neural network is the so-called back propagation. It belongs to the family of supervised learning methods, where the desired output for every input event is known. Back propagation is used by all neural networks in TMVA. The output of a network is given by

$$y_{ANN} = \sum_{j=1}^{n_h} y_j^{(2)} w_{j1}^{(2)} = \sum_{j=1}^{n_h} \tanh(\sum_{i=1}^{n_{\text{var}}} x_i w_{ij}^{(1)}) \cdot w_{j1}^{(2)}, \tag{3.15}$$

where n_{var} and n_h are the numbers of neurons in the input layer and in the hidden layer, respectively, $w_{ij}^{(1)}$ is the weight between input-layer neuron i and hidden-layer neuron j, and $w_{j1}^{(}2)$ is the weight between the hidden-layer neuron j and the output neuron.
During the learning process, the network is supplied with N training events $\mathbf{x}_a = (x_1, ..., x_{n_{\text{var}}})_a$, $a = 1, ..., N$. For each training event a the neural network output $y_{ANN,a}$ is computed and compared to the desired output $\hat{y}_a$ ranges from 0 to 1 (in classification 1 for signal events and 0 for background events). An error function E, measuring the agreement of the network response with the desired one, is defined by

$$E(\mathbf{x}_1, ..., \mathbf{x}_N|\mathbf{w}) = \sum_{a=1}^{N} E_a(\mathbf{x}_a|\mathbf{w}) = \sum_{a=1}^{N} \frac{1}{2}(y_{ANN,a} - \hat{y}_a)^2 \tag{3.16}$$

where $\mathbf{w}$ denotes the ensemble of adjustable weights in the network. The set of weights that minimises the error function can be found using the method of gradient descent, provided that the neuron response function is differentiable with respect to the input weights. Starting from a random set of weights $\mathbf{w}^{(\rho)}$ the weights are updated by moving a small distance in $\mathbf{w}$-space into the direction $-\nabla_{\mathbf{w}}\mathbf{E}$ where E decreases most rapidly

$$\mathbf{w}^{(\rho+1)} = \mathbf{w}^{(\rho)} - \eta \nabla_{\mathbf{w}} E \tag{3.17}$$

where the positive number η is the learning rate. The weights connected with the output layer are updated by

$$\Delta w_{j1}^{(2)} = -\eta \sum_{a=1}^{N} \frac{\partial E_a}{\partial w_{j1}^{(2)}} = -\eta \sum_{a=1}^{N} (y_{ANN,a} - \hat{y}_a) y_{j,a}^{(2)}, \tag{3.18}$$

and the weights connected with the hidden layers are updated by

$$\Delta w_{ij}^{(1)} = -\eta \sum_{a=1}^{N} \frac{\partial E_a}{\partial w_{ij}^{(1)}} = -\eta \sum_{a=1}^{N} (y_{ANN,a} - \hat{y}_a) y_{j,a}^{(2)} (1 - y_{j,a}^{(2)}) w_{j1}^{(2)} x_{i,a}, \tag{3.19}$$

where it has been used $\tanh' x = \tanh x(1 - \tanh x)$. This method of training the network is denoted bulk learning, since the sum of errors of all training events is used to update

the weights. An alternative choice is the so-called online learning, where the update of the weights occurs at each event. The weight updates are obtained from the previous equations by removing the event summations. Online learning is the learning method implemented in TMVA and used in this work.

3.5.3 Boosted Decision Trees

A decision tree is a binary tree-structured classifier with repeated binary yes/no decisions performed on one single variable at a time until a stop criterion is fulfilled. The phase space is split this way into many regions that are eventually classified as signal or background, depending on the majority of training events that end up in the final leaf node. The training of a decision tree is the process that defines the splitting criteria for each node. The training starts with the root node, where an initial splitting criterion for the full training sample is determined. The split results in two subsets of training events that each goes through the same algorithm of determining the next splitting iteration. This procedure is repeated until the whole tree is built. At each node, the split is determined by finding the variable and corresponding cut value that provides the best separation between signal and background. A variety of separation criteria can be defined to assess the performance of a variable and a specific cut requirement:

- Gini Index, defined by $p \cdot (1 - p)$;
- Cross entropy, defined by $-p \cdot \ln(p) - (1 - p) \cdot \ln(1 - p)$;
- Misclassification error, defined by $1 - \max(p, 1 - p)$;
- Statistical significance, defined by $S/\sqrt{S + B}$;

Decision trees can be boosted. The boosting of a decision tree extends this concept from one tree to several trees which form a forest. The trees are derived from the same training ensemble by reweighing events and are finally combined into a single classifier which is given by the average of the individual decision trees. Boosting stabilises the response of the decision trees with respect to fluctuations in the training sample and is able to considerably enhance the performance with respect to a single tree.

Boosting

Boosting is a way of enhancing the classification performance of typically weak MVA methods by sequentially applying a MVA algorithm to reweighed versions of the training data and then taking a weighted majority vote of the sequence of MVA algorithms thus produced. It has been introduced to classification techniques in the early '90s and in many cases this simple strategy results in dramatic performance increases. These so-called 'weak classifiers' are small trees, limited in growth to a typical tree depth of as small as two, depending on the how much interaction there is between the different input variables. By limiting the tree depth during the tree building process (training), the tendency of overtraining for simple decision trees, which are typically grown to a large depth and then pruned, is almost completely eliminated.

Adaptive Boosting

Adaptive Boosting (AdaBoost) is the most popular boosting algorithm. In a classification problem, events that were misclassified during the training of a decision tree are given a higher event weight in the training of the following tree. Starting with the original event weights when training the first decision tree, the subsequent tree is trained using a

modified event sample where the weights of previously misclassified events are multiplied by a common boost weight α. The boost weight is derived from the misclassification rate, err, of the previous tree,

$$\alpha = \frac{1 - err}{err}. \tag{3.20}$$

The weights of the entire event sample are then renormalised such that the sum of weights remains constant. We define the result of an individual classifier as $h(\mathbf{x})$, with ($\mathbf{x}$ being the array of input variables) encoded for signal and background as $h(\mathbf{x}) = +1$ and -1, respectively. The boosted event classification $y_{\text{Boost}}(\mathbf{x})$ is then given by

$$y_{\text{Boost}}(\mathbf{x}) = \frac{1}{N_{\text{collection}}} \cdot \sum_{i}^{N_{\text{Collection}}} \ln(\alpha_i) \cdot h_i(\mathbf{x}) \tag{3.21}$$

where the sum is over all $N_{\text{collection}}$ classifiers in the collection. Small (large) values for $y_{\text{Boost}(\mathbf{x})}$ indicate background-like (signal-like) event. Equation 3.21 represents the standard boosting algorithm. The performance is often further enhanced by forcing a "slow learning" and allowing a larger number of boost steps. The learning rate of the AdaBoost algorithm is controlled by a parameter β giving as an exponent to the boost weight $\alpha \to \alpha^{\beta}$, which can be modified using the configuration option string of the MVA method to be boosted.

Gradient Boosting

The boosting procedure adjusts the parameters such that the deviation between the model response $F(\mathbf{x})$ and the true value y obtained from the training sample is minimised. The deviation is measured by the so-called loss-function $L(F, y)$. It can be shown that the loss function fully determines the boosting procedure. The AdaBoost method is based on exponential loss, $L(F, y) = e^{-F(\mathbf{x})y}$, which leads to the well-known reweighing algorithm previously described. Exponential loss has the shortcoming that it lacks robustness in presence of outliers or mislabelled data points. The performance of AdaBoost, therefore, is expected to degrade in noisy settings. The GradientBoost algorithm attempts to cure this weakness by allowing for other, potentially more robust, loss functions without giving up on the good out-of-the-box performance of AdaBoost. The current TMVA implementation of GradientBoost uses the binomial log-likelihood loss:

$$L(F, y) = \ln(1 + e^{-2F(\mathbf{x})y}) \tag{3.22}$$

for classification. As the boosting algorithm corresponding to this loss function cannot be obtained straightforwardly, one has to resort to a steepest-descent approach to do the minimisation. This is done by calculating the current gradient of the loss function and then growing a regression tree whose leaf values are adjusted to match the mean value of the gradient in each region defined by the tree structure. Iterating this procedure yields the desired set of decision trees which minimises the loss function. GradientBoost can be adapted to any loss function as long as the calculation of the gradient is feasible. Just like AdaBoost, GradientBoost works best on weak classifiers, meaning small individual decision trees with a depth of often just 2 to 4. Given such small trees, they are much less prone to overtraining compared to simple decision trees. Its robustness can be enhanced by reducing the learning rate of the algorithm through the "Shrinkage" parameter. In certain settings, GradientBoost may also benefit from the introduction of a bagging-like resampling procedure using random subsamples of the training events for growing the trees. This is called stochastic gradient boosting and can be enabled by selecting the UseBaggedGrad option. The sample fraction used in each iteration can be controlled

through the "BaggingSampleFraction" parameter, where typically the best results are obtained for values between 0.5 and 0.8.

3.5.4 k-Nearest Neighbour (k-NN)

The k-Nearest Neighbour method compares an observed event from the test sample to reference events from a training dataset. An event is classified by a plurality vote of its neighbours, with the event being assigned to the class most common among its k nearest neighbours. Unlike other MVA methods, which use a fixed-sized multidimensional volume surrounding the test event, the k-NN algorithm is intrinsically adaptive and it defines a volume for the metric used looking at the adjacent events. The k-NN classifier has the best performance when the boundary that separates signal and background events has irregular features that cannot be easily approximated by parametric learning methods.

The k-NN algorithm searches for k events that are closest to the test event. Closeness is thereby measured using a metric function. The simplest metric choice is the Euclidean distance

$$R = \Big(\sum_{i=1}^{n_{\text{var}}} |x_i - y_i|^2 \Big)^{\frac{1}{2}} \tag{3.23}$$

where n_{var} is the number of input variables used for the classification, x_i are coordinates of an event from a training sample and y_i are variables of an observed test event. The k events with the smallest values of R are the k-nearest neighbours. The value of k determines the size of the neighbourhood for which a probability density function is evaluated. Large values of k do not capture the local behaviour of the probability density function. On the other hand, small values of k cause statistical fluctuations in the probability density estimate. A case study with real data suggests that values of k between 10 and 100 are appropriate and result in similar classification performance when the training sample contains hundreds of thousands of events (and n_{var} is of the order of a few variables). The classification algorithm finds k-nearest training events which can be of two types:

$$k = k_S + k_B \tag{3.24}$$

where k_S represents the number of the signal events in the training sample and k_B is the number of the background events in the training sample. The relative probability that the test event is of signal type is given by

$$P_S = \frac{k_S}{k_S + k_B} = \frac{k_S}{k} \tag{3.25}$$

The choice of the metric governs the performance of the nearest neighbour algorithm. When input variables have different units a variable that has a wider distribution contributes with a greater weight to the Euclidean metric. This feature is compensated by rescaling the variables using a scaling fraction which applies a factor to variable i determined by the width w_i of the x_i distribution for the combined sample of signal and background events. The input variables are then rescaled by $1/w_i$, leading to the rescaled metric

$$R = \Big(\sum_{i=1}^{d} \frac{1}{w_i^2} |x_i - y_i|^2 \Big)^{\frac{1}{2}} \tag{3.26}$$

The output of the k-NN algorithm can be interpreted as a probability that an event is of signal type, if the sum of event weights of signal and background in the training sample are equal. This can be enforced via trimming. If set training events of the overabundant type are randomly removed until parity is achieved.

Like all TMVA classifiers, the k-NN estimate suffers from statistical fluctuations in the training data. The typically high variance of the k-NN response is mitigated by adding a weight function that depends smoothly on the distance from a test event. The current k-NN implementation uses a polynomial kernel

$$W(x) = \begin{cases} (1 - |x|^3)^3 & \text{if |x| < 1,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.27}$$

If R_k is the distance between the test event and the k-th neighbour, the events are weighted according to the formulas:

$$W_S = \sum_i^{k_S} W\Big(\frac{R_i}{R_k}\Big), \quad W_B = \sum_i^{k_B} W\Big(\frac{R_i}{R_k}\Big) \tag{3.28}$$

where k_S is the number of the signal events and k_B is the number of the background events in the neighbourhood. Finally, the weighted signal probability for the test event is then given by

$$P_S = \frac{W_S}{W_S + W_B}. \tag{3.29}$$

3.6 HEP top tagger

The HEP top tagger was first designed to reconstruct mildly boosted top quarks in a busy event environment [60], i.e. for the reconstruction of top quarks in the process $t\bar{t}H$ with semi-leptonic top quark decays and $H \to b\bar{b}$ [69]. The hadronically decaying top quark was expected to be boosted in the p_T range around 250-500 GeV. This first incarnation of the tagger was augmented by cuts on observables that were manifestly Lorentz-invariant, and thus boosting between reference frames were no longer necessary. It proceeds as follows (for the details of analytic calculations see Appendix A of [60]):

1. one first defines very fat jets setting $R = 1.5$ with the Cambridge/Aachen algorithm,

2. for a given boosted jet, one recursively undoes the last step of the clustering, i.e. decluster the jet j into subjets j_1 and j_2 with the convention $m_{j_1} > m_{j_2}$, until we observe a mass-drop $m_{j_1} < 0.8\, m_j$. The declustering procedure with j_1 keeps going on until the mass-drop condition is met.

3. For subjets which have passed the mass-drop condition and which satisfy $m_j > 30$ GeV, one further decomposes the subjet recursively into smaller subjets.

4. The next step is to apply a filter similarly to what is done by the mass-drop tagger. One considers all pairs of hard subjets, defining a filtering radius $R_{\text{filt}} = \min(0.3, \Delta R_{ij})$. We then add a third hard subjet - considering again all possible combinations - and apply the filter on the three hard subjets keeping (at most) the 5 hardest pieces and use that to compute the jet mass. Amongst all possible triplets of the original hard subjets, we keep the combination for which the jet mass − calculated after filtering − gives the mass closest to the top mass and is within a mass window around the true top mass, e.g. in the range 150 − 200 GeV.

5. Out of the 5 filtered pieces, one extracts a subset of 3 pieces, j_1, j_2, j_3, ordered in p_T and accept it as a top quark candidate if the masses satisfy at least one of the following 3 criteria:

$$0.2 < \arctan\Big(\frac{m_{13}}{m_{12}}\Big) < 1.3 \qquad \text{and} \qquad R_{\min} < \frac{m_{23}}{m_{123}} < R_{\max} \tag{3.30}$$

$$R^2_{\min}\left(1+\frac{m^2_{13}}{m^2_{123}}\right) < 1-\frac{m^2_{23}}{m^2_{123}} < R_{\max^2}(1+\frac{m^2_{13}}{m^2_{123}}) \quad \text{and} \quad \frac{m_{23}}{m_{123}} > 0.35 \quad (3.31)$$

$$R^2_{\min}\left(1+\frac{m^2_{12}}{m^2_{123}}\right) < 1-\frac{m^2_{23}}{m^2_{123}} < R_{\max^2}(1+\frac{m^2_{12}}{m^2_{123}}) \quad \text{and} \quad \frac{m_{23}}{m_{123}} > 0.35 \quad (3.32)$$

6. the combined p_T of the 3 subjets constructed in the previous step is imposed to be at least 200 GeV.

Physically, the first three steps above try to decompose a massive object into its hard partons, in a spirit similar to what the mass-drop condition used in the mass-drop tagger does. The filtering step also plays the same role of further cleaning the contamination from the underlying event as in the mass-drop tagger (see Section 3.4.2). Finally, the set of constraints in Equations 3.30, 3.31, 3.32 is meant as a cut on the 3-subjets, mimicking a 3-parton system, to match the kinematics of a top quark decay and further suppress the QCD background. The whole procedure can be visualised as shown in Fig 3.6.

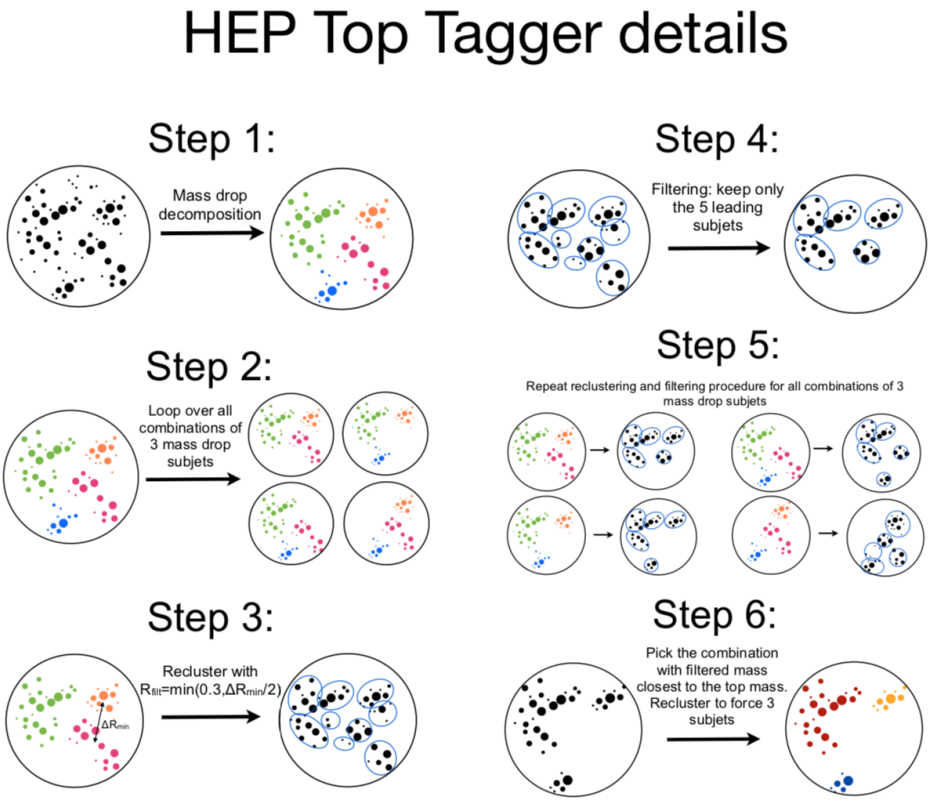

Figure 3.3: Visualisation of the HEP top tagger algorithm.

Version 2 of the HEPTopTagger [70] brings several improvements by using an extended set of variables and cuts. We just list those modifications without entering into the details.

First, it introduces a variable radius by repeatedly reducing the jet radius, starting from $R = 1.5$, until we see a drop in the reconstructed top quark mass. This is meant to reduce possible combinatorial effects where the softest of the W decays is mistaken with a hardish QCD subjet in the fat top candidate jet. Then, the tagger includes additional shape variables:

- N-subjettiness values computed both on the plain, ungroomed, jet and on the filtered jet;
- Q-jet information: the reconstructed top quark mass obtained from 100 Q-jet [71, 72] histories based on the Cambridge/Aachen algorithm, as well as the fraction of positive top tags one would obtain with version 1 of the HEPTopTagger.

In the end, the tagger uses a BDT multivariate analysis based on the series of kinematic variables - subjet transverse momenta and masses - the optimal jet radius, and the shape values.

Chapter 4

Data Analysis

This chapter describes the characterisation of the $t\bar{t}H$ production in the multijet final states. In Section 4.1, we present the samples used for the analysis and their compositions. A preselection of the events based on jet triggers can be found in Section 4.2. Some preliminary categories based on resolved (AK4) and boosted (AK8) jet multiplicities are defined in Section 4.3 and it is observed which are the most promising according to the signal-over background ratio. Then the analysis is subdivided into two parts, the resolved topology, in which all particles originate resolved jets (Section 4.4), and the boosted one (Section 4.5), where at least one jet is clustered as boosted. The two analyses are studied separately, and proper signal categories are defined. Different MVA techniques are adopted, in order both to enhance signal contribution from the very large background and to classify boosted jets according to which particle they are originated.

4.1 Event samples

The analysis has been conducted using several MC samples, simulating both signal and background processes. These simulations are normalised to the full Run2 dataset, corresponding to a centre-of-mass energy of $\sqrt{s} = 13\,\text{TeV}$ and an integrated luminosity $L = 140\,\text{fb}^{-1}$ (referring to the 2016+2017+2018 dataset). Signal events correspond to the $t\bar{t}H$ production, while $t\bar{t}$ and QCD are treated as backgrounds. There are two $t\bar{t}H$ samples, one where the Higgs boson decays into a b quark pair, $t\bar{t}H_{H\to b\bar{b}}$, and another sample accounting for all the other decay possibilities, $t\bar{t}H_{H\to W^+W^-,\tau^+\tau^-,ZZ,\gamma\gamma,...}$. In both cases, $t\bar{t}H$ events have been simulated using MADGRAPH + aMC@NLO [28], while $t\bar{t}$ background events have been simulated using POWHEG [73, 74]. QCD multijet production has been simulated with MADGRAPH. QCD simulated events are divided into slices of H_T, which stands for the scalar sum of parton transverse momenta. In all cases, the parton shower simulation is performed using PYTHIA [75] with the CUETP8M2 tuning [76, 77], developed by CMS with an updated strong coupling α_S for initial-state radiation, to better model the jet multiplicity spectrum. Table 4.1 summarizes all the MC samples used in the analysis and the corresponding cross sections of the involved processes.

4.1.1 Samples weighting

The events of all the plots in this work are weighed considering the reference integrated luminosity $L_{ref} = 140\,\text{fb}^{-1}$ and the physical cross section σ_{th} of the processes. Considering the total number of MC generated events, N_{MC}, and the physical cross section σ_{th},

Sample	σ_{th}**(pb)**	**Events**
$t\bar{t}H_{H\to b\bar{b}}$	0.2934	9794226
$t\bar{t}H_{H\to W^+W^-,\tau^+\tau^-,ZZ,\gamma\gamma,...}$	0.2151	3860872
$t\bar{t}$	832	76079906
QCD multijet (300<H_T<500 GeV)	3.477 $\times 10^5$	54537903
QCD multijet (500<H_T<700 GeV)	3.21 $\times 10^4$	62271343
QCD multijet (700<H_T<1000 GeV)	6.831 $\times 10^3$	45412780
QCD multijet (1000<H_T<1500 GeV)	1.207 $\times 10^3$	15127293
QCD multijet (1500<H_T<2000 GeV)	119.9	11826702
QCD multijet (H_T>2000 GeV)	25.24	6039005

Table 4.1: MC samples used in the analysis: $t\bar{t}H$, $t\bar{t}$ and QCD multijet events. $t\bar{t}H$ simulated samples are divided into two samples according to the $H \to b\bar{b}$ decay mode or all the other decay modes.

both listed in Table. 4.1, for each sample the integrated luminosity L_{MC} is defined by

$$N_{MC} = \sigma_{th} \cdot L_{MC}. \tag{4.1}$$

The number of generated events expected after a certain selection and corresponding to L_{ref} is obtained from the generic number of events N passing the selection by the following relation

$$N_{ref} = N \cdot \frac{L_{ref}}{L_{MC}} = N \cdot \frac{L_{ref} \cdot \sigma_{th}}{N_{MC}}. \tag{4.2}$$

4.1.2 Samples composition

Lepton composition

The $t\bar{t}H$ sample offers different topologies to be studied according to the number of prompt leptons: all-hadronic ($0\,\ell$), single lepton ($1\,\ell$) and dilepton ($\geq 2\,\ell$). A general distinction between events is accomplished using the number of identified charged leptons. The inclusive $t\bar{t}H$ sample, that is the sum of $t\bar{t}H_{H\to b\bar{b}}$ and $t\bar{t}H_{H\,Not\,b\bar{b}}$ samples, contains 70325 weighted events. For the three topologies we expect the following number of events:

$$t\bar{t}H(0\ell) = 47103 \qquad t\bar{t}H(1\ell) = 20213 \qquad t\bar{t}H(2\ell) = 3009 \tag{4.3}$$

If we divide for the total number of events, we get the following fractions $\mathcal{F}$:

$$\mathcal{F}(0\ell) = 67\% \qquad \mathcal{F}(1\ell) = 28.7\% \qquad \mathcal{F}(2\ell) = 4.3\% \tag{4.4}$$

A small fraction of events is characterised by the presence of three identified leptons. In Fig. 4.1 events are divided according to the number of leptons and the individual contribution of the two signal samples can be appreciated.

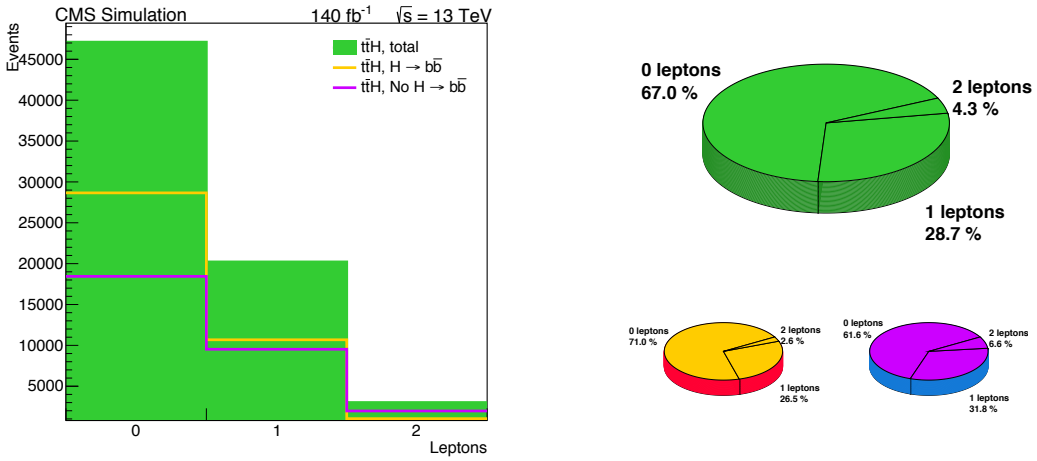

Figure 4.1: Lepton composition for the $t\bar{t}H$ sample shown as events (left) and percentages (right), for the $t\bar{t}H_{H\to b\bar{b}}$ (orange), $t\bar{t}H_{H\,Not\,b\bar{b}}$ (violet) and $t\bar{t}H$ (green) samples.

AK8 jets

The so-called AK8 jets are jets reconstructed with a large R parameter (see the previous chapter). The AK8 jet multiplicity is shown in Fig. 4.2 (left), while the AK8 b-jets multiplicity is shown in Fig. 4.2 (right) for the different simulated samples.

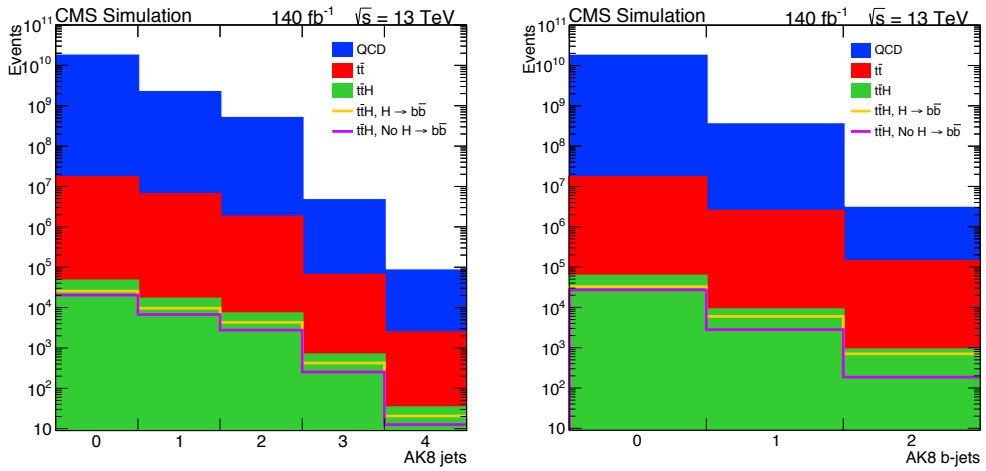

Figure 4.2: Number of AK8 jets (left) and AK8 b-jets (right) for the different simulated samples.

AK4 Jets

The so-called AK4 jets are jets reconstructed with a small R parameter (see the previous chapter). Here, the events are then divided according to the number of the AK4 jets. The AK4 jet multiplicity is shown in Fig. 4.3 (left), while the AK4 b-jets multiplicity is shown in Fig. 4.3 (right) for the different simulated samples.

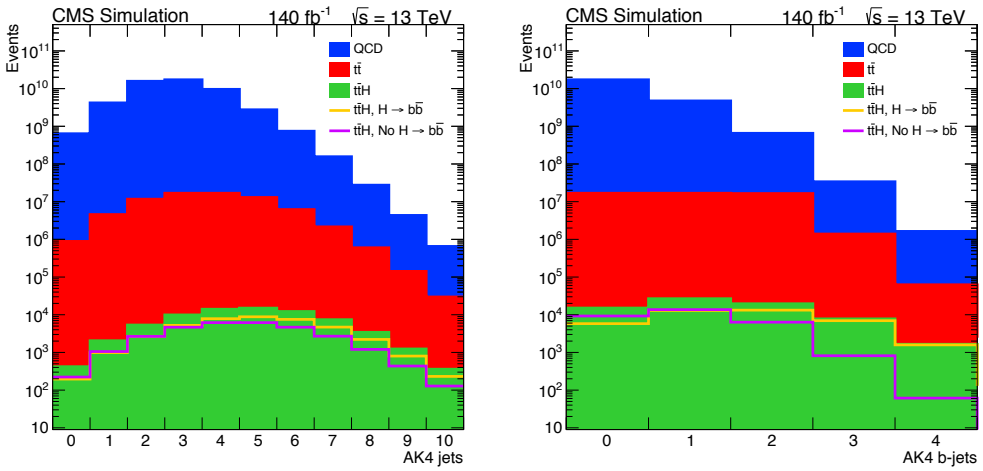

Figure 4.3: Number of AK4 jets (left) and AK4 b-jets (right) for the different simulated samples.

4.2 Preselection: lepton veto and multijet topologies

From now on, the analysis will focus on the all-hadronic topologies, namely the topologies without any leptons. The lepton veto (0ℓ) is so applied until the end of the work. Considering the two kinds of jets, we define two all-hadronic topologies: a "resolved topology", with no AK8 jets but only several AK4 jets, and the "boosted topology", with at least one AK8 jet and additional AK4 jets. A schematic picture of the resolved topology can be seen in Fig. 4.4 (left), while in Fig 4.4 (right) it is represented an example of the boosted topology. These two topologies are so different that they deserve to be studied separately, after a common introduction.

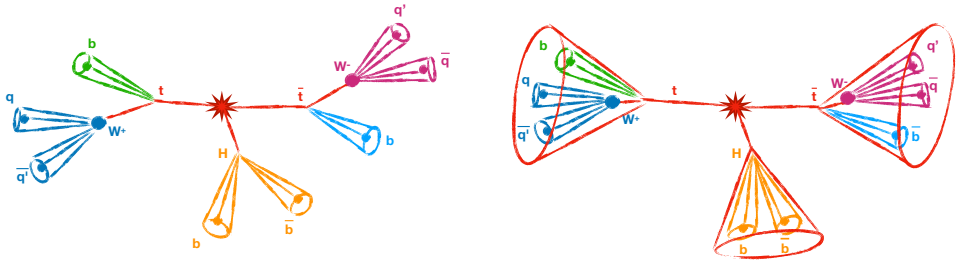

Figure 4.4: Schematic picture of the resolved (left) and boosted (right) tt̄H multijets topologies.

4.2.1 Multijet triggers

A combination of three different trigger has been employed in order to efficiently select signal events for the resolved analysis. Events are required to pass the OR of the triggers:

- `HLT_PFHT450_SixJet40_BTagCSV_p056_v`, requiring the presence of six leading PF jets with $p_T > 40$ GeV and one b-jet (with CSV discriminator > 0.56). It also requires $H_T > 450$ GeV in the event;

- `HLT_PFHT400_SixJet30_DoubleBTagCSV_p056_v`, requiring the presence of six leading PF jets with $p_T > 30$ GeV and two b-jets (with CSV discriminator > 0.56). It also requires $H_T > 400$ GeV in the event;

- `HLT_QuadJet45_TripleBTagCSV_p087_v`, requiring the presence of four jets with $p_T > 45$ GeV and three b-jets (with CSV discriminator > 0.87). It also requires $H_T > 450$ GeV in the event.

The first two triggers have been developed by CMS to efficiently select signal events, both requiring at least six jets with $|\eta| < 2.6$, and subsequent complementary requirements. The first trigger has more strict kinematic requirements in jet p_T and H_T than the second one, but it requires at least one of the jets to be b-tagged while the second one requires at least two of the jets to be b-tagged. The b-jets are tagged online by the triggers at an efficiency of $\approx$ 70–80%, with a misidentification rate of $\approx$ 6% for light-flavour quark and gluon jets [79]. The efficiency in data and simulation for the either of the first two triggers is measured in bins of the number of b-jets, the p_T of the jet with the sixth-highest p_T, and the H_T in control samples collected using single-muon triggers. A bin-by-bin scale factor has been applied to simulated events to correct for any remaining differences. The overall trigger efficiency for signal events that pass the offline event selection is 99%. In our analysis, a third additional trigger is employed. To ensure that the trigger selection is close to full efficiency relative to the offline selection, thereby reducing the uncertainty in any efficiency differences between data and simulation, the offline analysis should selects simulated events that contain at least six jets, at least 2 b-tagged jets, $|\eta| < 2.4$ and requiring $H_T > 500$ GeV.

4.2.2 AK8 jet trigger

For all the boosted jet topologies, the trigger used is:

- `HLT_AK8PFHT700_TrimR0p1PT0p03Mass50_v`, requiring the presence of one AK8 PF jet with $H_T > 700$ GeV and invariant (trimmed) mass > 50 GeV.

The AK8 jet trigger efficiency is shown in the left plot of Fig. 4.5 as a function of S_T, which stands for the sum of AK4 jets H_T and AK8 jets H_T. The trigger efficiency has been computed as the ratio of events passing both the aforementioned trigger and a reference trigger, and events passing only the reference trigger, which has less strict kinematic requirements. As a reference trigger, `HLT_AK8PFJet260_v` trigger is used, requiring the presence of one AK8 PF jet with $p_T > 260$ GeV. It can be noticed a monotonic behaviour of the trigger efficiency as a function of the S_T variable, increasing with increasing values of S_T. If one requires $S_T > 900$ GeV, the trigger efficiencies for the three samples are all $> 80\%$. The S_T distribution for the three different samples is shown in the right plot of Fig. 4.5.

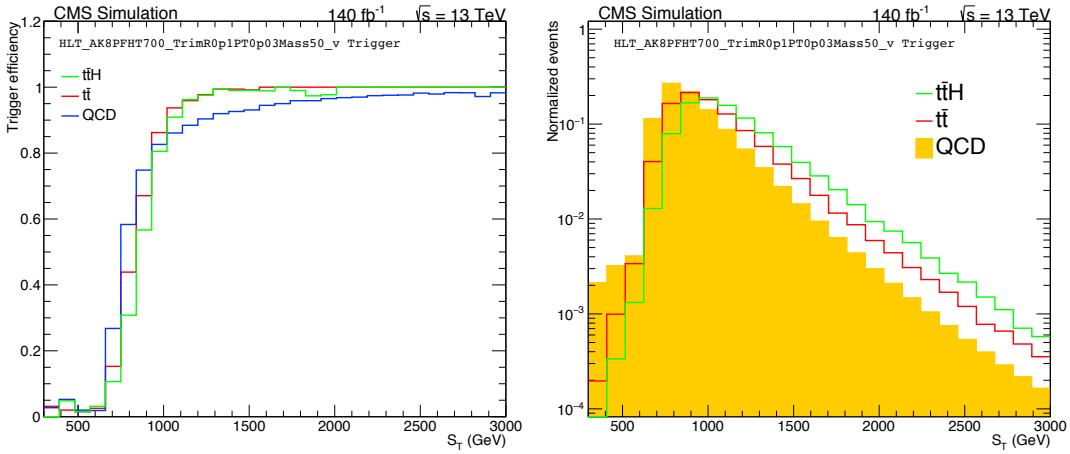

Figure 4.5: Trigger efficiency as a function of AK8 jet S_T (left) and AK8 jet S_T distributions (right) for the different samples.

4.2.3 The effect of parton p_T on the jet topology

At the recent 13 TeV run of LHC, one encounters jets with large p_T. The decay products can be so collimated that the standard reconstruction technique of AK4 jets is not effective because, along with these, AK8 jets come into play in the final states. The number of AK4 jets as a function of the number of AK8 jets is shown in Fig. 4.6, using events from the $\mathrm{t\bar{t}H}$ sample. In the left plot, there is no request on the p_T of the particle originating the jets. In the right plot, the minimum p_T between t, $\bar{\mathrm{t}}$ and H particles is computed and required to be > 300 GeV. The change in the 2D distributions clearly indicates that top quarks or Higgs bosons with large p_T tend to produce jets collimated by Lorentz boost which are clustered as AK8 jets, increasing their multiplicity.

As discussed in Eq. 3.5, for a fixed value of z, the angular distance between the bottom quarks from the Higgs boson decay, $\Delta R_{b\bar{b}}$, is predicted to have a $1/p_T$ behaviour. This feature can be observed in Fig. 4.7, which shows $\Delta R_{b\bar{b}}$ as a function of the Higgs boson p_T in simulated $\mathrm{t\bar{t}H_{H\to b\bar{b}}}$ events.

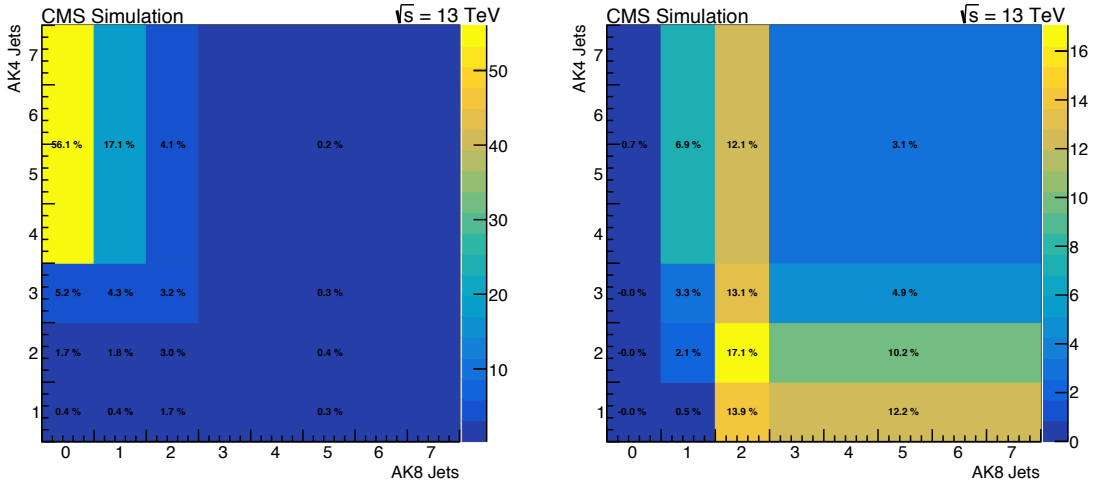

Figure 4.6: AK4-AK8 jets without (left) and with (right) p_T >300 GeV request. Lepton veto imposed.

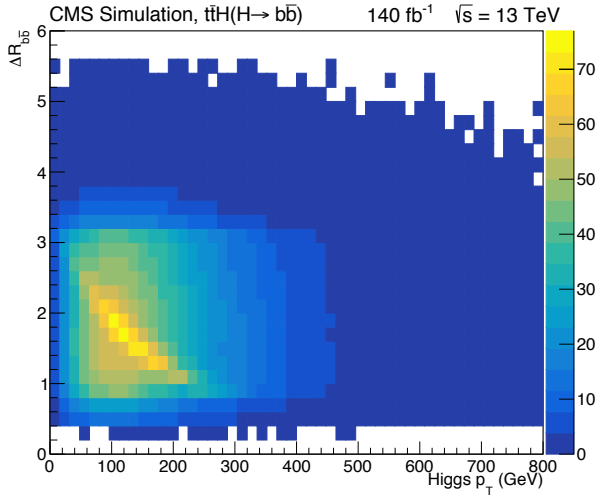

Figure 4.7: Distance between the two bottom quarks coming from the Higgs boson decay as a function of the Higgs boson p_T.

4.3 Jet multiplicities

In the all-hadronic regime, the number of jets in the final states varies according to the jet topology, resolved or boosted. For this reason, it is important to try to ideally estimate the number of events which we expect for different jet multiplicities.

4.3.1 Expected composition of jet multiplicity

If the p_T of the particles originating the jets is relatively small, no boosted jets are present and all jets are resolved, thus 8 resolved jets, four of which b-tagged, are expected (see Fig. 4.8). If a jet coming from the top quark decay is boosted, then we expect in principle

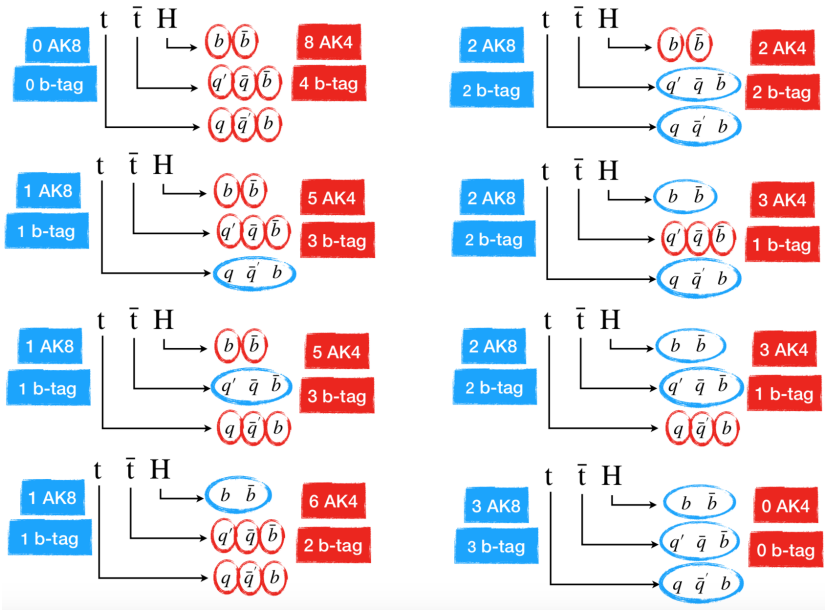

Figure 4.8: Jet categories, with the assumption that jets from H → bb̄ or W → qq̄′ originate resolved (red) or boosted (blue) jets.

5 resolved jets, 3 of which b-tagged; if a jet from the Higgs boson decay is boosted, 6 resolved jets, 2 of which b-tagged, are expected. The analysis considers also the possibility that more than one jet is boosted: if the jets coming from the two top quarks are boosted, 2 resolved jets also b-tagged are expected; if a jet from the Higgs boson and a jet from a top quark are boosted, 3 resolved jets with one b-tagged are expected. If the particle p_T is very large, all jets are boosted, and no resolved jet is present in the final state. All of the boosted jets, independently if they come from a top quark or a Higgs boson, should be b-tagged (the Higgs boson twice b-tagged).

4.3.2 Jet multiplicity for the simulated samples

Figure 4.9 displays how the distributions of the AK8 and AK4 jet multiplicities, respectively, change with the trigger request in addition to the lepton veto with respect to the previous AK8 and AK4 jet multiplicity distributions of Figs. 4.2 and 4.3. The QCD multijet background is suppressed especially for the topology with 0 AK8 jets by about 2 orders of magnitude, and significantly also for topologies with 1 or 2 AK8 jets. For

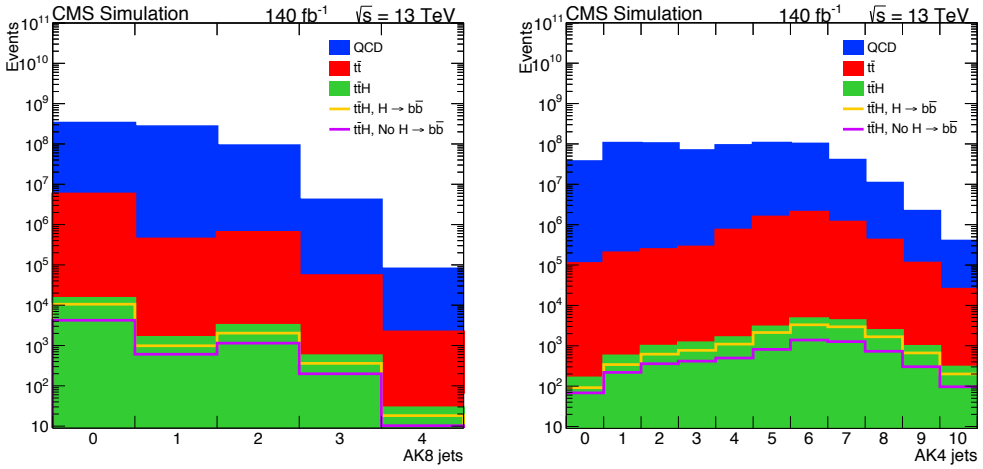

Figure 4.9: AK8 (left) and AK4 (right) jet multiplicities, for events passing both the boosted trigger request and the lepton veto.

different numbers of AK8 jets, the multiplicity of AK4 jets varies. In Fig. 4.10 AK4 jet multiplicity distributions are shown for 0, 1, 2 or ≥ 3 AK8 jets with events taken from the signal samples.

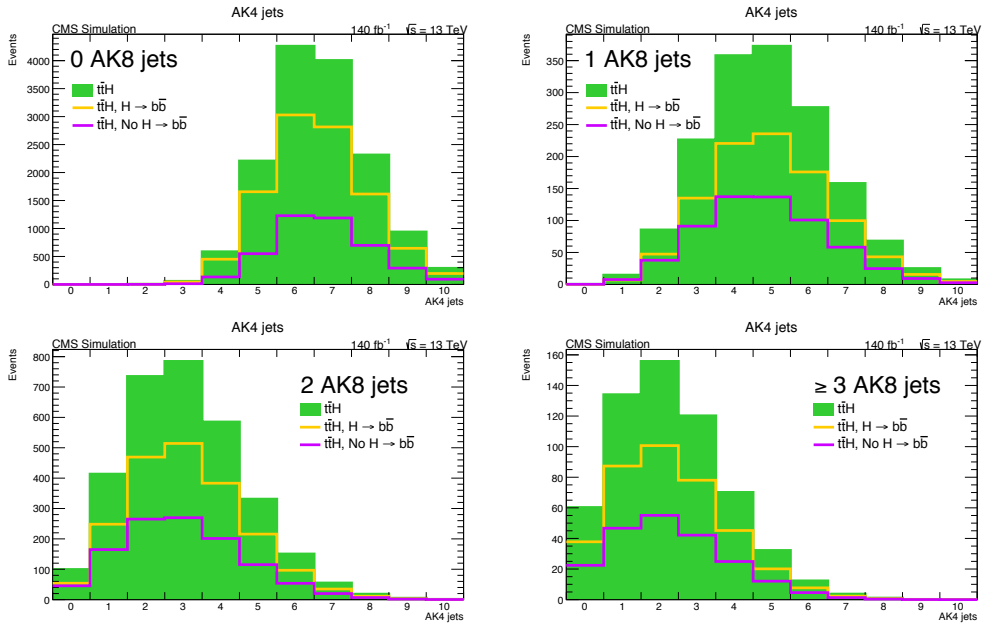

Figure 4.10: AK4 jets distributions for different AK8 jet multiplicity requests.

The topology with 0 AK8 jets is the most populated in comparison to the others and has been carefully studied by the CMS collaboration. In this region, low jet multiplicities are suppressed from the multijet trigger combination, as can be seen. Increasing the number of AK8 jets from 1 to ≥ 3, one can see that the AK4 jets distributions tend to have

fewer and fewer jets, as expected. Moreover, as can be inferred from the distributions, the region with ≥ 3 jets is the least populated, as it is rare that all three jets can be simultaneously boosted and therefore originate from a particle with a remarkably high p_T, given the high mass of the particle. This suggests from now on to merge the regions with 2 and ≥ 3 AK8 jets into a single ≥ 2 AK8 jet topology.

4.3.3 Categories based on jet multiplicity

We categorise events according to the number of jets, studying how many AK4 jets are expected for the 0, 1 and ≥ 2 AK8 jet topologies. The all-boosted topology with 3 AK8 jets is included in the ≥ 2 AK8 jet topology. Events with no identified leptons based on jet multiplicities can be chosen to have:

- 0 AK8 jets, with ≤ 3, $4/5$, ≥ 6 AK4 jets;
- 1 AK8 jet, with ≤ 2, $3/4$, ≥ 5 AK4 jets;
- ≥ 2 AK8 jets, with $0/1$, $2/3$, ≥ 4 AK4 jets.

Simulated signal and background yields for various combinations of AK8 and AK4 jets are displayed in Fig. 4.11, along with the signal-over-background ratio in Fig. 4.11 and both reported in Table 4.2.

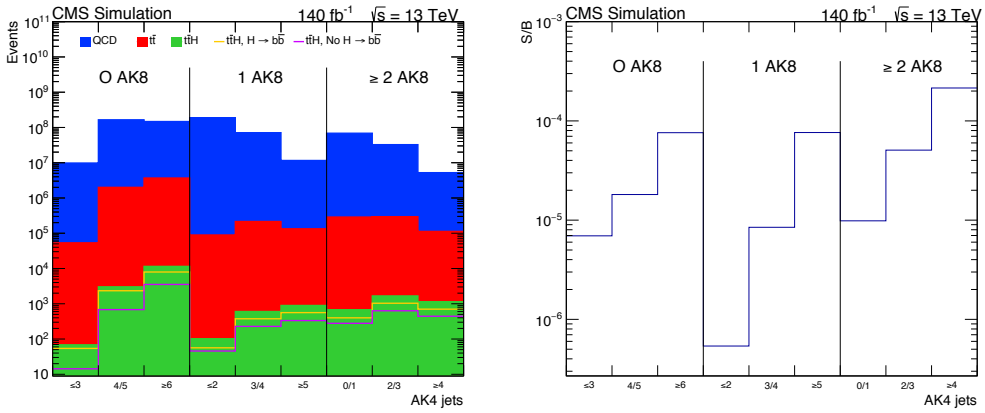

Figure 4.11: Simulated signal and background yields (left) and signal-over-background ratio (right) for various combinations of AK8 and AK4 jets.

For the topology with 0 AK8 jets, signal events are expected to populate mainly the category with ≥ 6 AK4 jets. In fact, if the Higgs boson does not decay into a $b\bar{b}$ pair, 6 AK4 jets will be produced, while if the H boson decays into a $b\bar{b}$ pair, 8 AK4 jets will be produced. Regardless of the Higgs boson decay mode, at least 6 jets are expected and required in the resolved topology.
For the topology with 1 AK8 jet, signal events are expected to populate mainly the categories with ≥ 5 AK4 jets or with $3/4$ AK4 jets. In fact, if the Higgs boson decays into a $b\bar{b}$ pair, 5 or 6 AK4 jets will be produced (depending on which particle forms the AK8 jet); if the Higgs boson does not decay into a $b\bar{b}$ pair, $3/4$ AK4 jets will be produced.
For the topology with ≥ 2 AK8 jets, signal events can feed into the category with $0/1$ AK4 jets if the two top quarks are boosted and the Higgs boson does not decay into a $b\bar{b}$ pair, or if we have 3 AK8 jets. The majority of the events are expected to populate the

Expected yields, $0\,\ell$, 140 fb^{-1}, $\sqrt{s} = 13$ GeV					
AK8 jets	AK4 jets	$t\bar{t}H$ sample	$t\bar{t}$ sample	QCD sample	S/B
0	≤ 3	69	53507	9807391	1/143959
	4/5	3024	2004059	164933568	1/55203
	≥ 6	11494	3684231	147397056	1/13144
1	≤ 2	102	89321	189395440	1/1857035
	3/4	604	215491	71099840	1/118082
	≥ 5	896	133862	11593893	1/13096
≥ 2	≤ 1	679	288104	68698736	1/101625
	2/3	1668	295098	32570502	1/19700
	≥ 4	1147	112899	5229359	1/4658

Table 4.2: Simulated signal and background yields for various combinations of AK8 and AK4 jets with the request of zero leptons. The signal-over-background ratio is also computed.

category with 2/3 AK4 jets if the Higgs boson decays into a $b\bar{b}$ pair.

4.3.4 Categories based on jet multiplicity including b-tag

The following step is taken by defining some categories considering also b-tagging. A boosted jet is considered b-tagged when it has the output discriminator of the b-tagging algorithm (CSVv2) greater than a threshold. CSVv2 discriminators for the $t\bar{t}H$, and $t\bar{t}$ and QCD samples are shown respectively in the left and right panel of Fig. 4.12. The b-jets composition is enhanced requiring a high CSVv2 score value, especially in the $t\bar{t}$+QCD case where for low value the c and udsg jets compositions are predominant. In this analysis a value of CSVv2 threshold of 0.8484 is used, corresponding to the medium working point of CSVv2 (CSVv2 M), having an efficiency of b-identification of about 63%, and of c and udsg quarks, respectively, of about 12% and 0.9% [78].

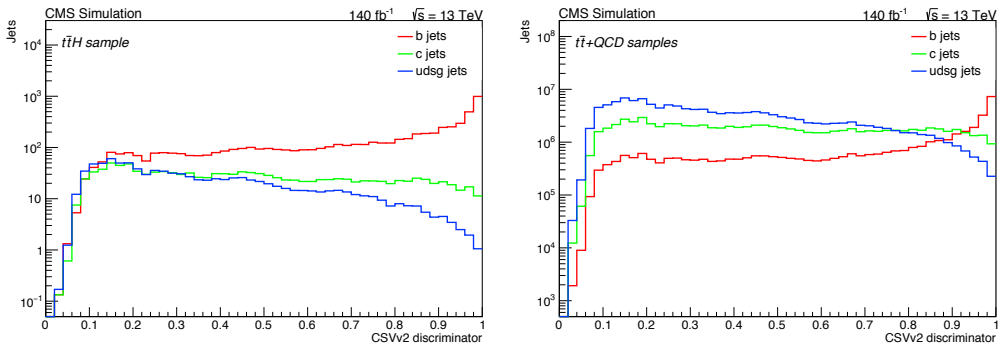

Figure 4.12: $t\bar{t}H$ jets composition (left) and $t\bar{t}$+QCD jets composition (right) as a function of the CSVv2 discriminator.

Events without any identified lepton based on jet multiplicities also including b-tag are chosen to have:

- 0 AK8 jets, with ≤ 3, 4/5, ≥ 6 AK4 jets;

- 1 AK8 jet, 0 AK8 b-jet, with ≤ 2, 3/4, ≥ 5 AK4 jets;
- 1 AK8 jet, 1 AK8 b-jet, with ≤ 2, 3/4, ≥ 5 AK4 jets;
- ≥ 2 AK8 jets, 0 AK8 b-jet, with 0/1, 2/3, ≥ 4 AK4 jets.
- ≥ 2 AK8 jets, 1 AK8 b-jet, with 0/1, 2/3, ≥ 4 AK4 jets.
- ≥ 2 AK8 jets, ≤ 2 AK8 b-jet, with 0/1, 2/3, ≥ 4 AK4 jets.

Expected signal and background yields for various combinations of AK8 and AK4 jets are shown in the upper panel of Fig. 4.13 with their signal-over-background ratios in the lower panel of Fig. 4.13 and both reported in Table 4.2.
The topology with 0 AK8 jets has already been discussed in the previous section, here reported for completeness. At least 6 AK4 jets will be required in the following resolved analysis.
The topology with 1 AK8 jet is most populated by signal events in the categories with ≥ 5 AK4 jets, both with the further request of 0 and 1 AK8 b-jet. The b-tagging request on the AK8 jet improves the signal-over-background ratio if compared to the corresponding category in which the AK8 jet is not b-tagged, mainly suppressing the QCD background contribution.
The topology with 2 AK8 jets is most populated by signal events in the categories with 2/3 AK4 jets, for all b-tagging requests on AK8 jet multiplicity. The b-tagging requests improve the signal-over-background ratio as the number of AK8 b-jets increases, helping to suppress the background.

Expected yields, $0\,\ell$, 140 fb^{-1}, $\sqrt{s} = 13$ GeV						
AK8 jets	AK8 b-jets	AK4 jets	$t\bar{t}H$ sample	$t\bar{t}$ sample	QCD sample	S/B
0	0	≤ 3	69	53507	9807391	1/143959
		4/5	3024	2004059	164933568	1/55203
		≥ 6	11494	3684231	147397056	1/13144
1	0	≤ 2	63	59354	159848144	1/2531801
		3/4	390	147771	66657804	1/171186
		≥ 5	600	96748	10557010	1/17757
	1	≤ 2	39	30044	15657642	1/403354
		3/4	213	67432	6850022	1/32415
		≥ 5	295	37222	1053600	1/3694
≥ 2	0	≤ 1	255	134707	58028956	1/227964
		2/3	704	153531	26866512	1/38403
		≥ 4	555	65588	4357993	1/7977
	1	≤ 1	312	123398	11457253	1/37064
		2/3	759	118913	5059581	1/6823
		≥ 4	487	41710	835437	1/1802
	≥ 2	≤ 1	105	28939	641676	1/6418
		2/3	201	21684	279069	1/1494
		≥ 4	105	5731	44770	1/483

Table 4.3: Expected signal and background yields for various combinations of AK8 jets, AK8 b-jets and AK4 jets with the request of zero leptons. The signal-over-background ratio is also computed.

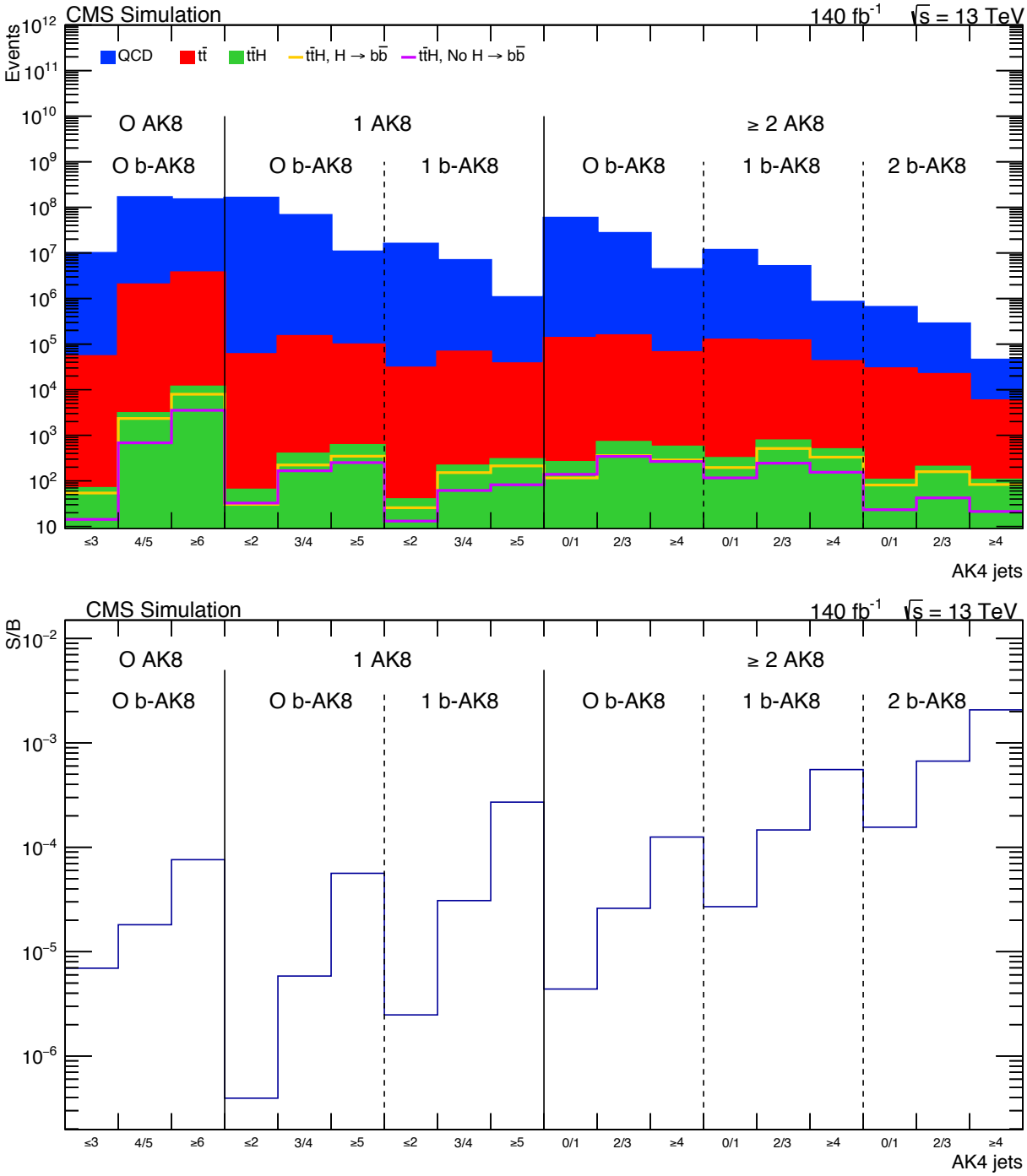

Figure 4.13: Yields (upper panel) and signal-over-background ratio (lower panel) for the defined categories.

4.4 Resolved analysis

The resolved analysis focuses on the presence of resolved jets only, corresponding to the previously mentioned category of 0 AK8 jets. In this case, it is expected a number of resolved jets greater or equal to six, at least two of which are b-tagged. Some preliminary cuts are performed and two MVAs have been employed, the first one discriminating the $t\bar{t}H$ signal events from the very large QCD multijet background, the second one discriminating signal events from the $t\bar{t}$ background. Events are then subdivided into categories according to the resolved jets and b-jet multiplicities.

4.4.1 Event preselection

In order to maximise the signal efficiency, a preselection common to all samples is needed. The preselection of events has to be coherent with the requirements of the multijet triggers and the kinematic properties of the final topology considered. The choice of preselection is a delicate balance between maximising the trigger efficiency and minimising the signal loss. We require the presence of at least 6 resolved jets, of which at least 2 are b-tagged. The resolved jet H_T is required to be ≥ 500 GeV and jet $|\eta|$ is required to be ≤ 2.4. The combination of the three triggers discussed in section 4.2.1 is employed. A schematic view of the baseline cuts is provided in Fig 4.14.

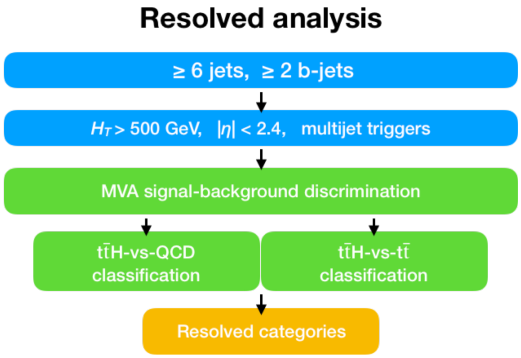

Figure 4.14: Resolved analysis strategy.

4.4.2 MVA signal-background discrimination

Input variables for the $t\bar{t}H$-vs-QCD classification

The following variables have been used as discriminating ones in the $t\bar{t}H$-vs-QCD MVA:

- number of resolved jets (see Fig. 4.15-left);
- H_T of resolved jets (see Fig. 4.15-right);

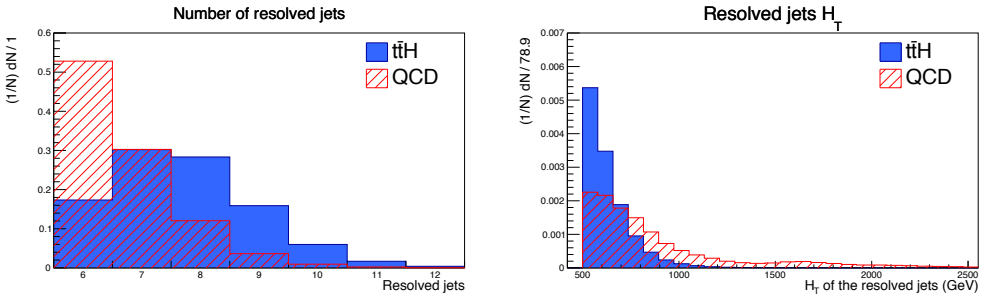

Figure 4.15: Number of resolved jets (left) and H_T of resolved jets (right).

- leading jet p_T (see Fig. 4.16-left);
- second jet p_T (see Fig. 4.16-right);

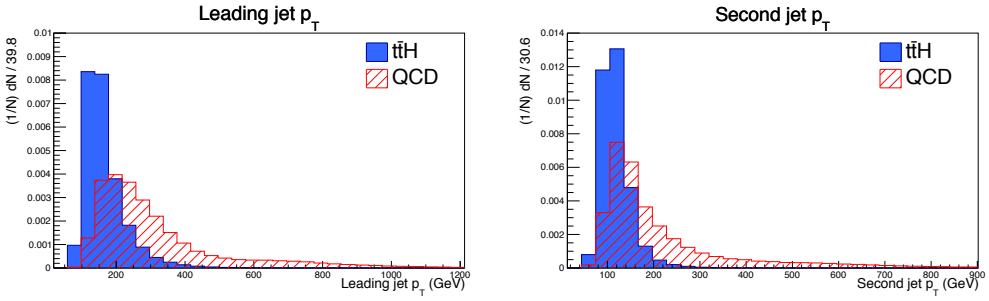

Figure 4.16: p_T of the leading resolved jet (left) and p_T of the second resolved jet (right).

- Minimum ΔR for the $b\bar{b}$ pairs (see Fig. 4.17-left);
- Mass of the $b\bar{b}$ pair with the minimum ΔR (see Fig. 4.17-right);

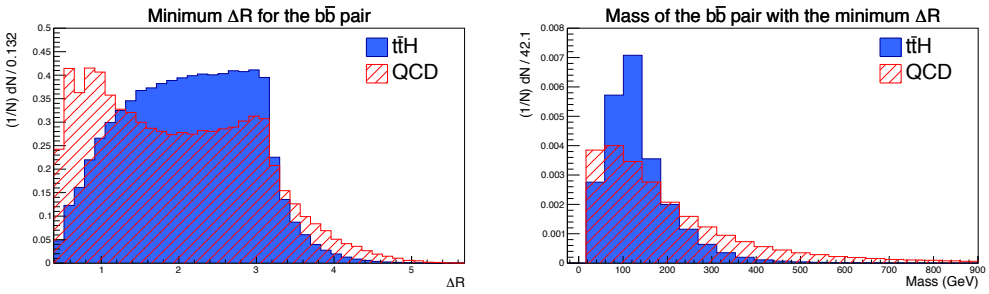

Figure 4.17: Minimum ΔR of all the b-tagged jet pairs (left) and mass of the b-tagged jet pairs which have the minimum ΔR (right).

- Minimum ΔR for the dijet pairs (see Fig. 4.18-left);
- Mass of the dijet system for minimum ΔR (see Fig. 4.18-right);

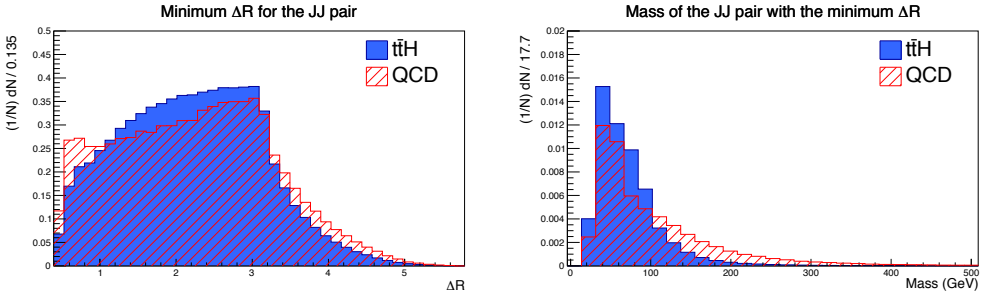

Figure 4.18: Minimum ΔR of all the resolved jets pairs (left) and mass of the resolved jets pairs which have the minimum ΔR (right).

- centrality, defined as $\sum_i p_T^i / E_{TOT}$ (see Fig. 4.19);

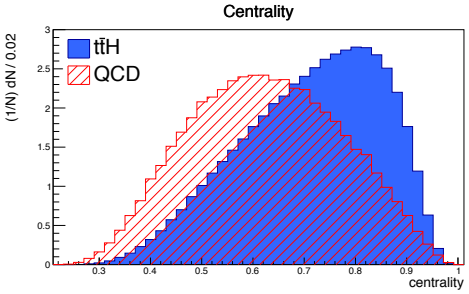

Figure 4.19: Centrality.

- sphericity, defined as $S = \frac{3}{2}(\lambda_2 + \lambda_3)$ where $\lambda_1 \geq \lambda_2 \geq \lambda_3$ are the three eigenvalues of the sphericity tensor defined as $S^{\alpha\,\beta} = \frac{\sum_i p_i^\alpha p_i^\beta}{\sum_i |p_i|^2}$. $S \simeq 1$ is expected for events where the jets are produced isotropically in space (see Fig. 4.20-left);
- aplanarity, defined as $A = \frac{3}{2}\lambda_3$. It essentially measures the transverse momentum component out of the event plane: a planar event has $A = 0$ and an isotropic one $A \simeq \frac{1}{2}$ (see Fig. 4.20-right).

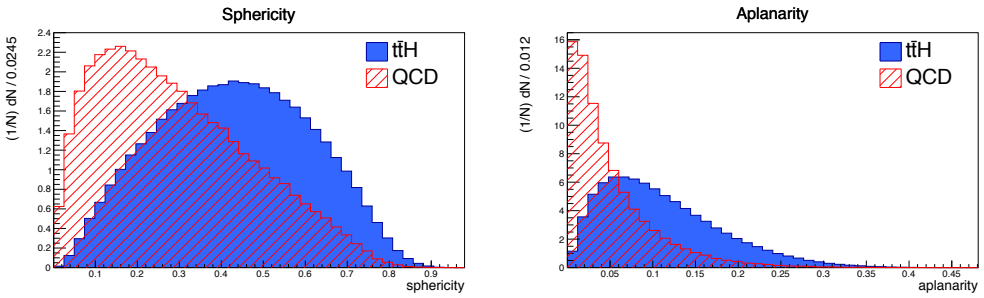

Figure 4.20: Sphericity (left), aplanarity (right).

Input variables for the $t\bar{t}H$-vs-$t\bar{t}$ classification

The following variables have been used as discriminating ones in the $t\bar{t}H$ vs $t\bar{t}$ MVA:

- number of resolved jets (see Fig. 4.21-left);
- H_T of resolved jets (see Fig. 4.21-right);

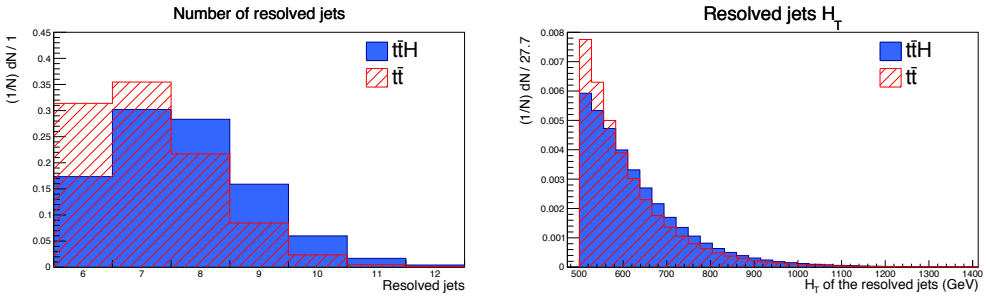

Figure 4.21: Number of resolved jets (left) and H_T of resolved jets (right).

- Minimum ΔR for the $b\bar{b}$ pairs (see Fig. 4.22);

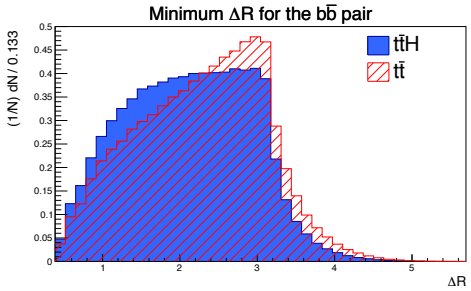

Figure 4.22: Minimum ΔR of the b-tagged jet pairs.

- cos θ_L^*, where θ_L^* is the angle for the leading jet and the z-axis in the centre-of-mass framework of the multijet system (see Fig. 4.23-left);
- cos θ_S^*, where θ_S^* is the angle for the second jet and the z-axis in the centre-of-mass framework of the multijet system (see Fig 4.23-right);

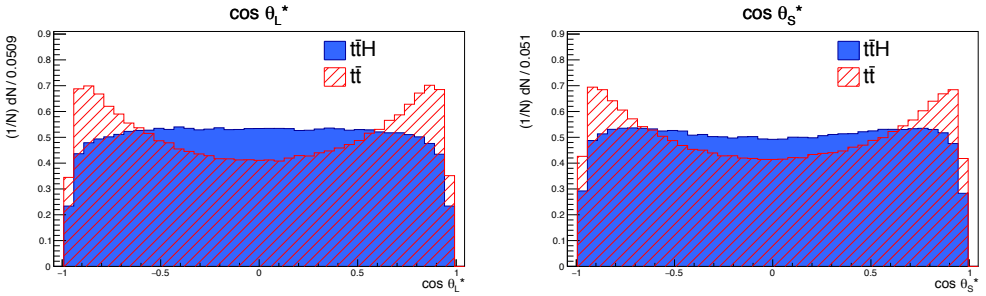

Figure 4.23: cos θ_L^* (left) and cos θ_S^* (right) between resolved jets.

- centrality (see Fig. 4.24)

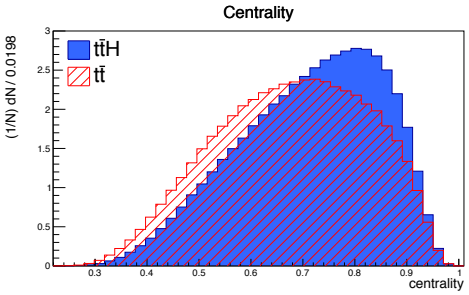

Figure 4.24: Centrality.

- sphericity (see Fig. 4.25-left);
- aplanarity (see Fig. 4.25-right).

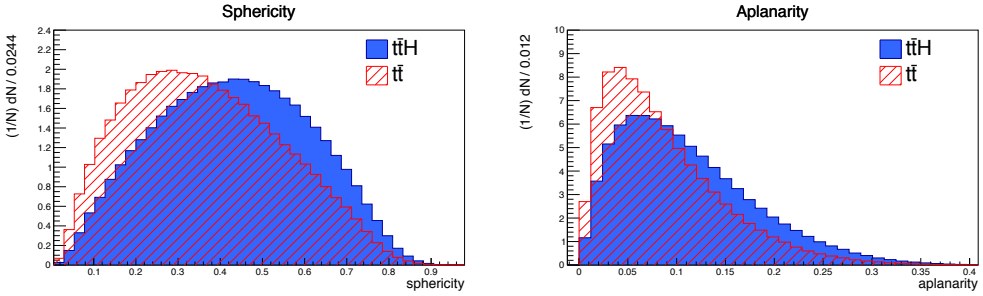

Figure 4.25: Sphericity (left), aplanarity (right).

MVA methods

Several MVA algorithms have been adopted in our analysis for the $t\bar{t}H$-vs-QCD classification:

- Fisher method, also in the boosted version (see section 3.5.1);
- ANN (see section 3.5.2);
- BDT, with Adaptive or Gradient boosting (see section 3.5.3);
- KNN (see section 3.5.4).

For each MVA method, some configuration parameters need to be set, apart from the Fisher method which is quite simple and without configuration parameters. The parameters values are reported for ANN, BDT and KNN in Table 4.4, 4.5 and 4.6, respectively.

ANN method	
Number of training cycles	500
Number of hidden layers	5
Neuron activation function type	Sigmoid
Random seed for initial synapse weights	Yes
Training method	Back-Propagation (BP)
Learning rate	0.02
Decay rate	0.01
Regolator to avoid over-training	Yes

Table 4.4: Configuration parameters for the ANN adopted in the resolved analysis.

BDT method		
Number of trees in the forest	1000	1000
Max depth of the decision tree allowed	3	3
Shrinkage (learning rate GradBoost)	-	0.3
AdaBoostBeta (learning rate AdaBoost)	0.5	-
Separation criterion for node splitting	Gini index	Gini index
Negative event weigths	Ignored	Ignored
Fraction of events used in each iteration	0.6	0.6

Table 4.5: Configuration parameters for the BDT adopted in the resolved analysis.

KNN method	
Number of k-nearest neighbours	25
Kernel function	Gaussian
Fraction of events used to compute variable width	0.8
Use weight to count kNN events	Yes

Table 4.6: Configuration parameters for the KNN adopted in the resolved analysis.

The choice of the best classifier has been made considering the "receiver operating characteristic", also called ROC curve, which shows the background rejection efficiency as a function of signal efficiency for the different MVA methods employed. The ROC curve for the $t\bar{t}$H-vs-QCD classification is shown in Fig. 4.26. The boosting of the Fisher discriminator clearly improves the standard Fisher performances. The KNN seems not to work very well for the classification. The ANN shows a good signal efficiency and background rejection and could be improved by better setting parameters. However, one should use some values of the parameters which do not favour one method with respect to another, and the choice adopted is to set the configuration values similar to the standard TMVA ones. The BDTs have the best background rejection versus signal efficiency power, the choice between the two is not straightforward since they quite overlap and for different ranges their relative performance changes. It seems that the GradBoostBDT performs better, especially in the regions of high- and medium-signal efficiency, as can be seen in the bottom panel of the figure, and for this reason it is chosen as the MVA method. Such behaviour is also reproduced in the $t\bar{t}$H-vs-$t\bar{t}$ discrimination, accordingly GradBoostBDT has been chosen for both the $t\bar{t}$H-vs-$t\bar{t}$ classification and the $t\bar{t}$H-vs-QCD classification.

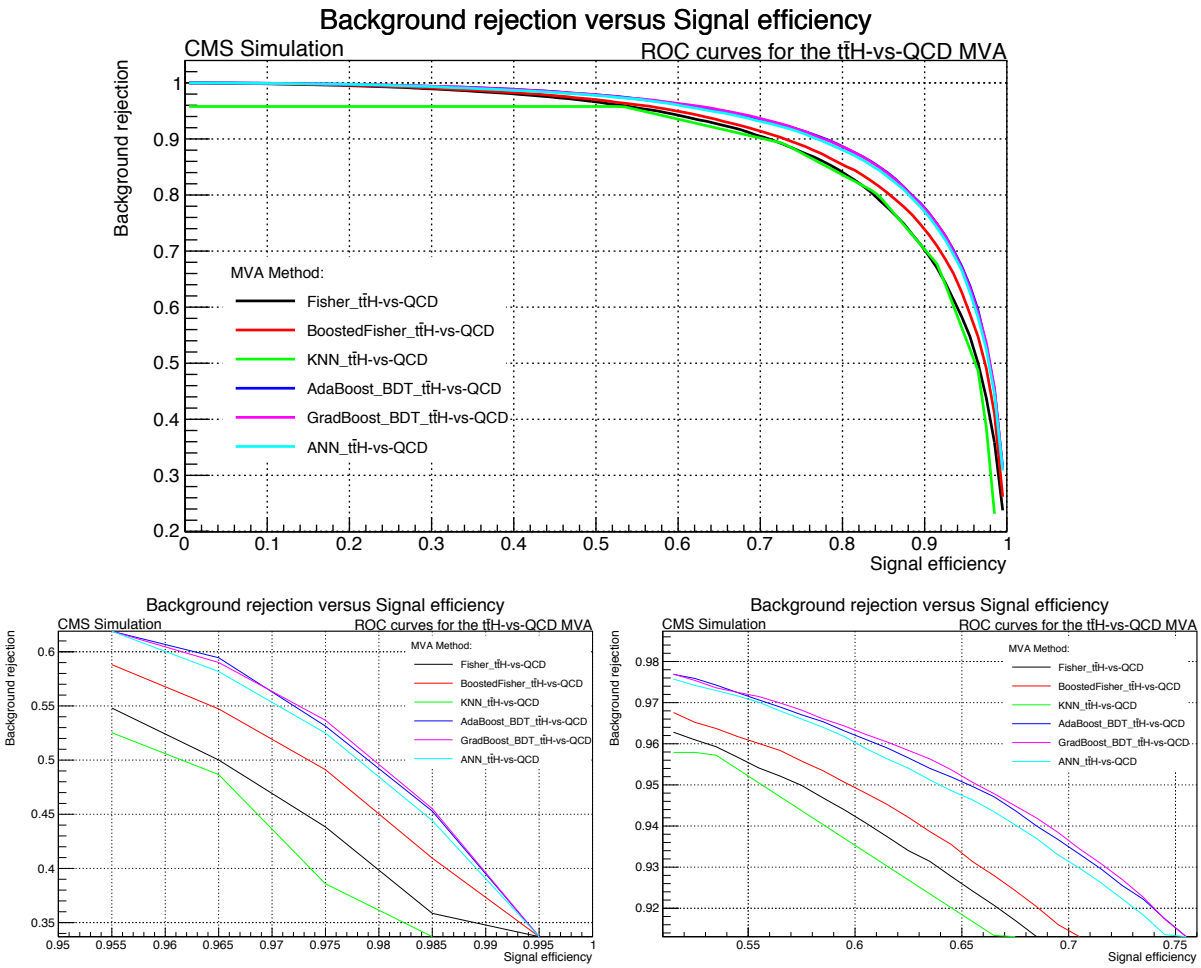

Figure 4.26: ROC curves for the $t\bar{t}$H-vs-QCD classification (upper panel), and zoom in high- and medium-signal efficiency intervals (lower panel).

Outputs scores

The input events for the two MVA analyses are split into one training sample and one test sample. This guarantees a statistically independent evaluation of the MVA algorithms based on the test sample. The number of events used in the two samples is the same. The $t\bar{t}$H-vs-QCD and $t\bar{t}$H-vs-$t\bar{t}$ output score distributions for the resolved analysis are shown in Figs. 4.27 and 4.28, respectively, with the training and test samples superimposed. The BDT output score ranges between -1 and 1, with 1 corresponding to a signal-like region and -1 corresponding to a background-like region. Each of these two output scores allows us to perform a single cut which essentially synthesises many cuts performed on the input variables with potentially complex relationships and leads to an improvement of the signal-over-background.
A very good separation can be appreciated for the $t\bar{t}$H-vs-QCD classification and a quite good separation can be appreciated for the $t\bar{t}$H-vs-$t\bar{t}$ classification.

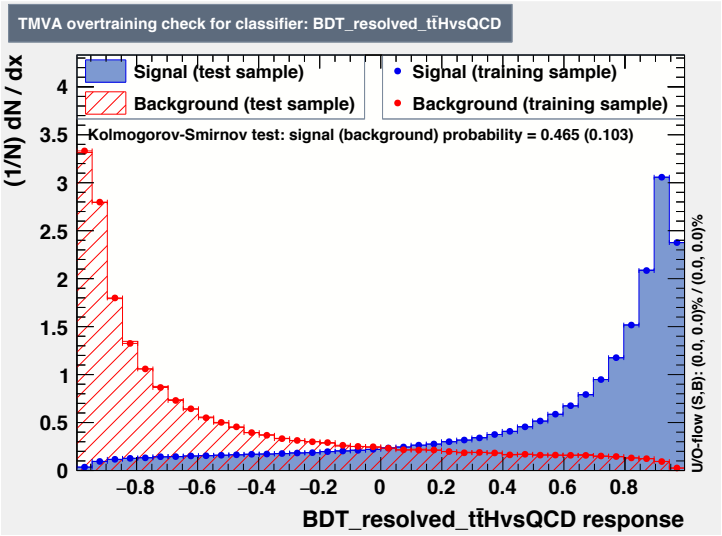

Figure 4.27: BDT output score for the $t\bar{t}$H-vs-QCD classification.

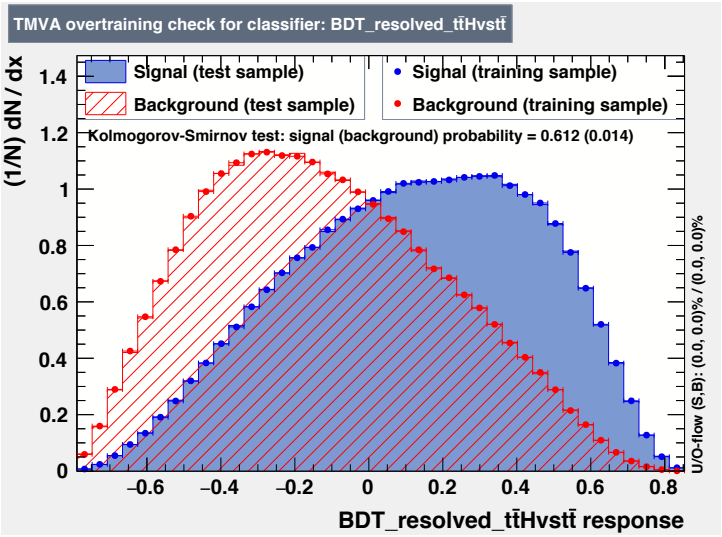

Figure 4.28: BDT output score for the $t\bar{t}$H-vs-QCD classification.

The $t\bar{t}$H and the total background ($t\bar{t}$+QCD) distributions are displayed in Fig. 4.29 as a function of the $t\bar{t}$H-vs-QCD output score. The optimal value for discrimination is chosen considering the $S/\sqrt{B}$ figure of merit (Fig. 4.29) and it is 0.8. The $t\bar{t}$H and the total background ($t\bar{t}$+QCD) distributions are displayed in Fig. 4.30 as a function of the $t\bar{t}$H-vs-$t\bar{t}$ output score. The optimal value for discrimination is chosen considering the $S/\sqrt{B}$ figure of merit (Fig. 4.30) and it is -0.1.

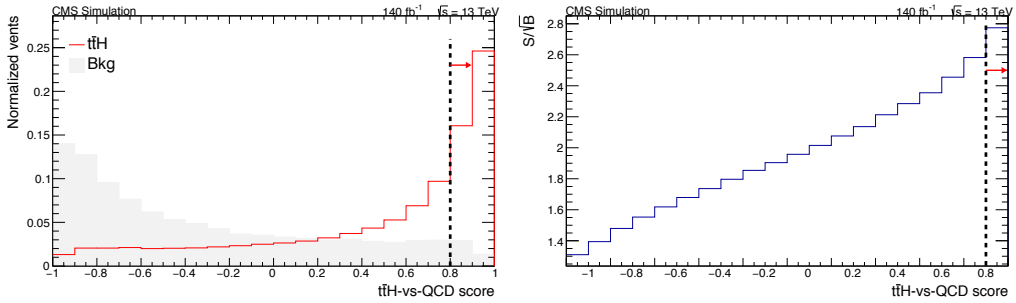

Figure 4.29: $t\bar{t}H$ and $t\bar{t}$+QCD background (left) and $S/\sqrt{B}$ (right) as a function of $t\bar{t}H$-vs-QCD output score.

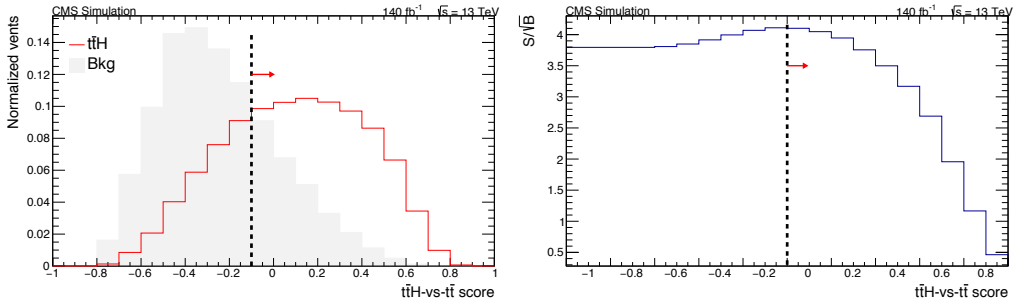

Figure 4.30: $t\bar{t}H$ and $t\bar{t}$+QCD background (left) and $S/\sqrt{B}$ (right) as a function of $t\bar{t}H$-vs-$t\bar{t}$ output score.

4.4.3 Signal categories

Signal categories are defined with further requirements on the number of resolved jets and b-jets in addition to the previously discussed selections for events, which are chosen requiring the BDT($t\bar{t}H$-vs-QCD) score to be larger than 0.8 and the BDT($t\bar{t}H$-vs-$t\bar{t}$) score to be larger than -0.1. Events are divided into 6 orthogonal categories differing for jet and b-jet multiplicities, as reported in Table 4.7.

Resolved analysis categories		
Category	Resolved jets	Resolved b-jets
1	7	3
2	7	≥ 4
3	8	3
4	8	≥ 4
5	≥ 9	3
6	≥ 9	≥ 4

Table 4.7: Resolved categories.

The expected yields and signal-over-background ratios in these categories are displayed in Fig. 4.31 and reported in Table 4.8.

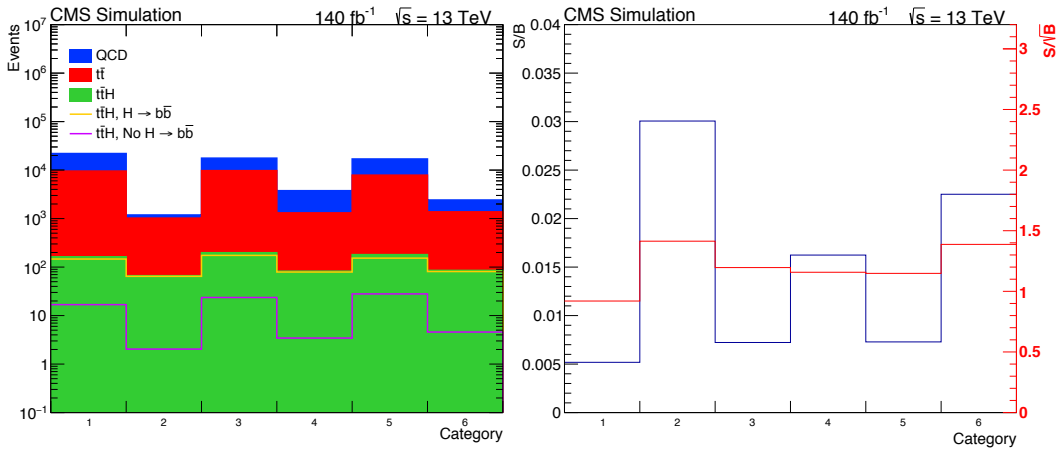

Figure 4.31: Resolved categories events yields (left), S/B and S/$\sqrt{\text{B}}$ (right).

Resolved categories				
Expected yields $L_{ref} = 140\,\text{fb}^{-1}, \sqrt{s} = 13\,\text{TeV}$				
Category	S	B	S/B	S/$\sqrt{\text{B}}$
1	163	31537	1/193	0.92
2	67	2212	1/33	1.42
3	198	27395	1/138	1.20
4	83	5079	1/61	1.16
5	181	24838	1/137	1.15
6	86	3798	1/44	1.40
Total	777	94860	1/122	2.52

Table 4.8: Signal and background expected yields, signal-over-background ratio and signal significance for the defined resolved categories.

The distributions of leading jet p_T for the defined categories are shown in Fig. 4.32 while the distributions of leading jet η for the defined categories are shown in Fig. 4.33. The distributions of leading jet p_T are peaked around a value which correctly matches the typical values of p_T of resolved jets and they show a similar shape for all categories. It can be noticed that the QCD yields have relevant fluctuations in the high-value tails because of the reduced size of the simulated sample. The jet η distributions are uniform. We also show the two BDT distributions in Figs. 4.34 and 4.35. For $\text{t}\bar{\text{t}}$H-vs-QCD discrimination, the signal is enhanced in the region close to 1 of the output score, and the requirement of being greater than 0.8 increases the signal S/B. For $\text{t}\bar{\text{t}}$H-vs-$\text{t}\bar{\text{t}}$ discrimination, the $\text{t}\bar{\text{t}}$ background is suppressed with the requirement for the output score of being greater than -0.1, with a small loss on signal events.

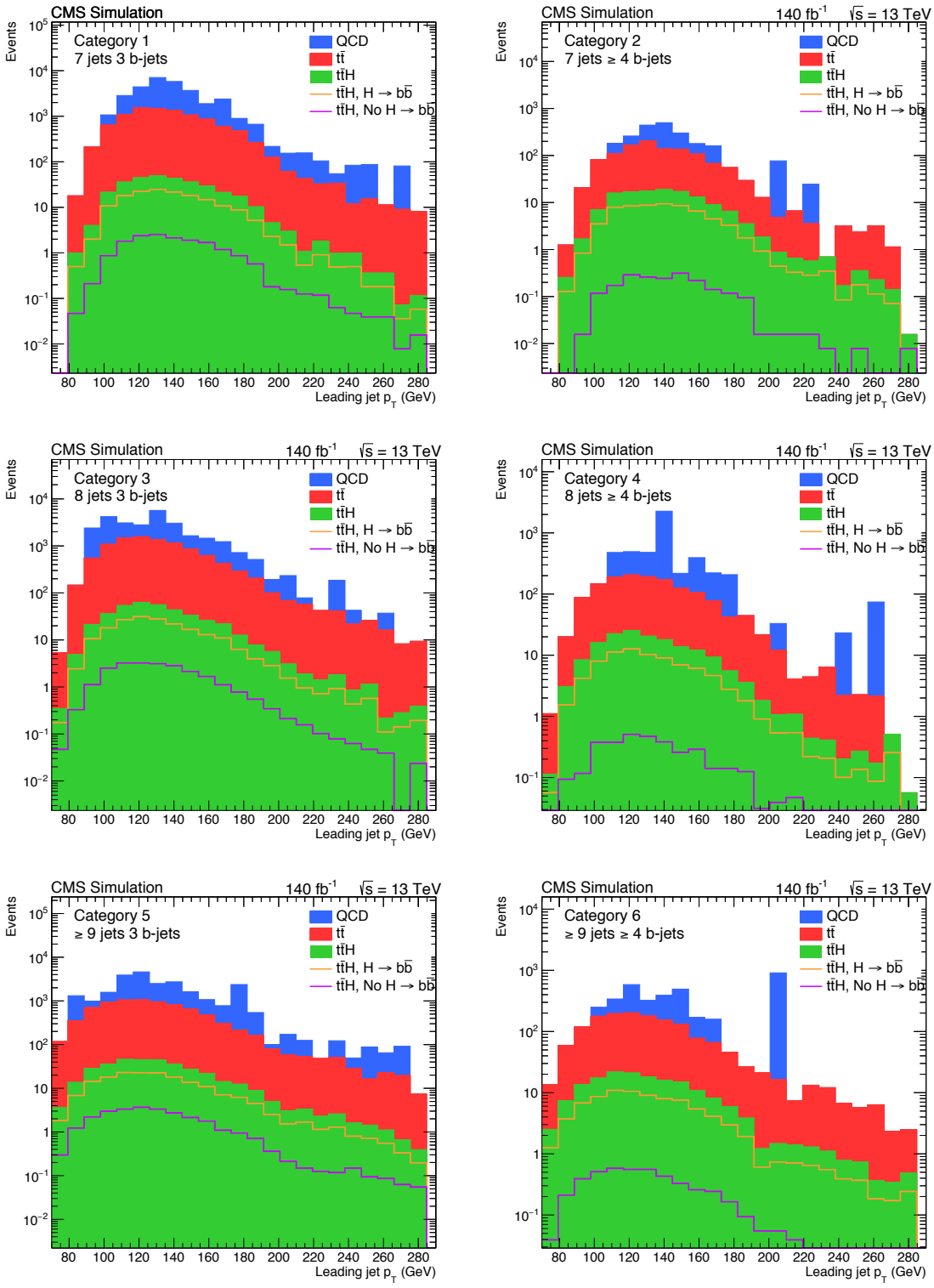

Figure 4.32: Leading jet p_T distributions for the resolved categories.

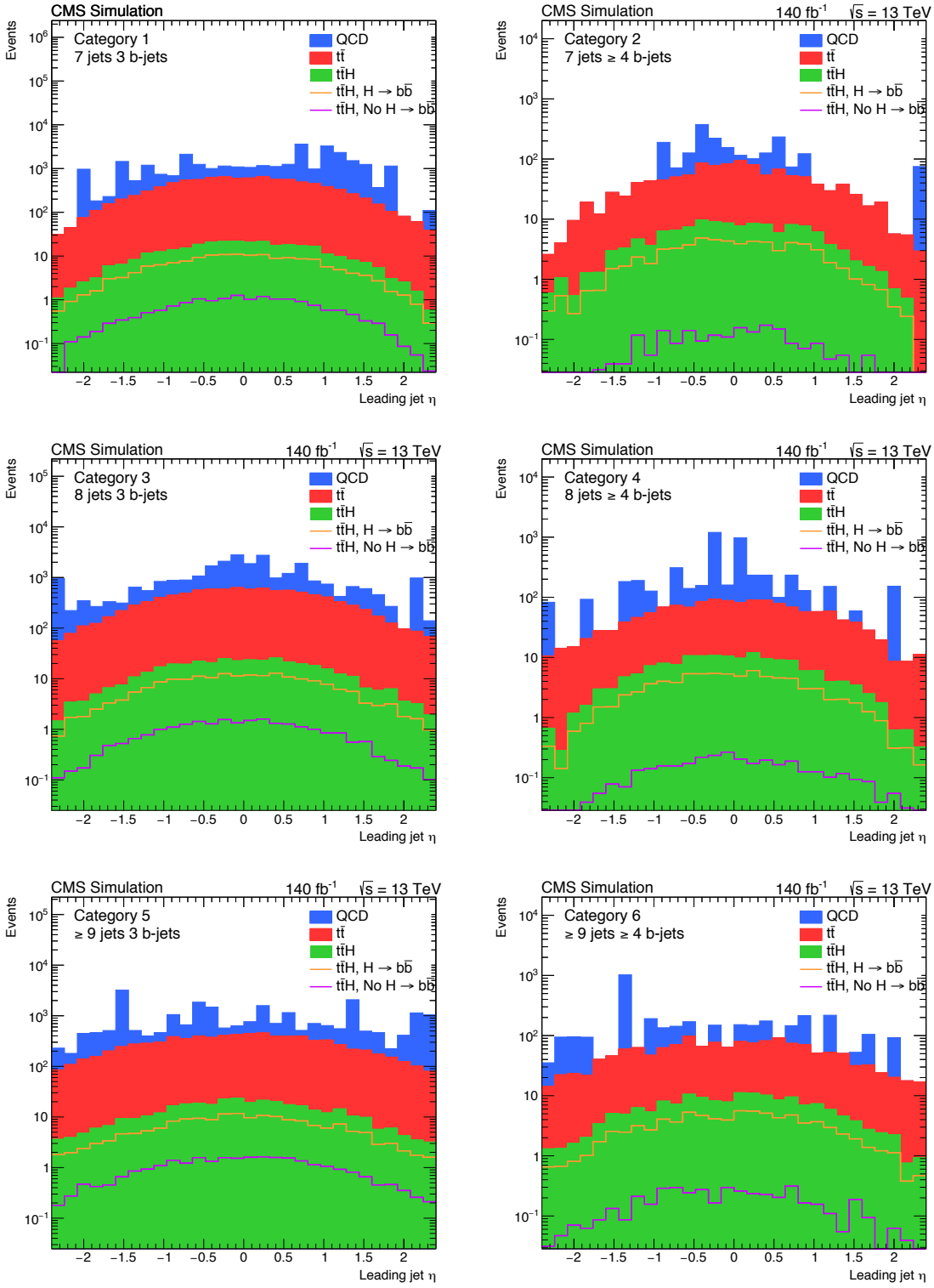

Figure 4.33: Leading jet η distributions for the resolved categories.

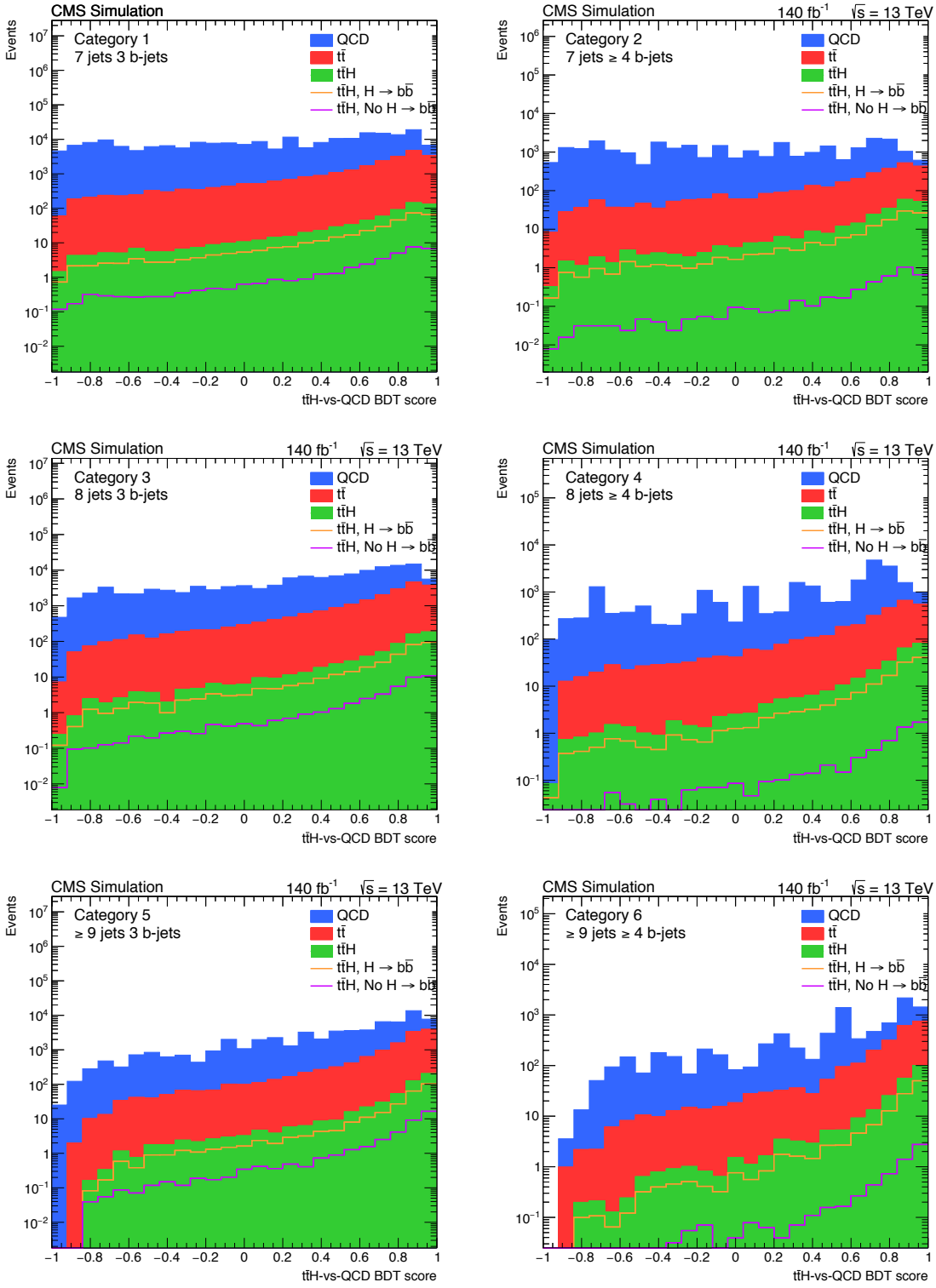

Figure 4.34: tt̄H-vs-QCD BDT output for the resolved categories.

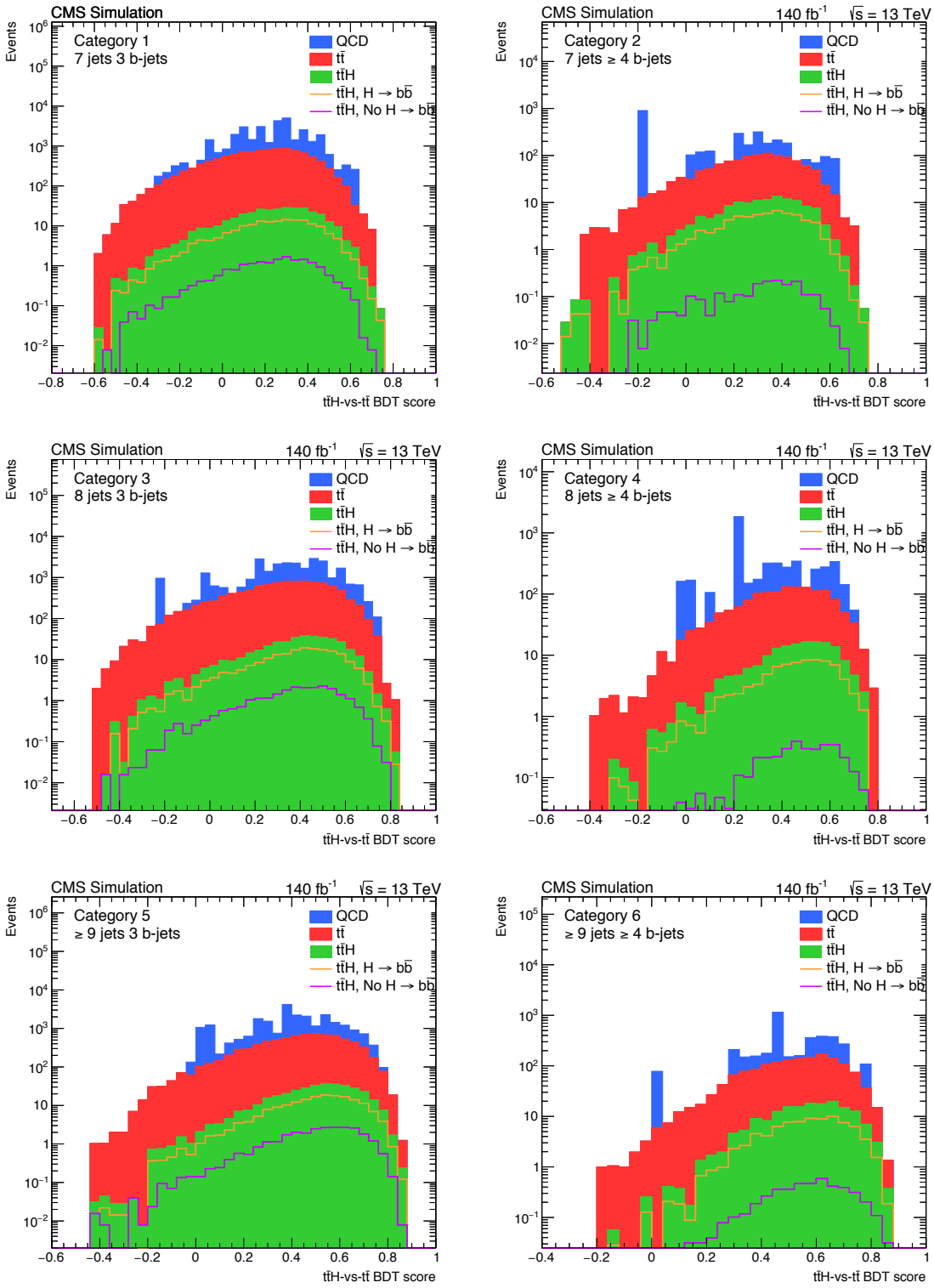

Figure 4.35: tt̄H-vs-tt̄ BDT output for the resolved categories.

4.5 Boosted analysis

The boosted analysis focuses on the presence of boosted jets along with resolved jets. In this regime, fewer events are expected. After a preselection of the events based on p_T, $|\eta|$, trigger, three MVA have been adopted which take subjet variables as input variables and are used with the aim of tagging the boosted jets, i.e. identify from which particle the jet has been originated. The three different MVAs, labelled H-vs-QCD, T-vs-QCD, H-vs-T, are discussed in the following. Signal categories based on boosted and resolved jets multiplicities are defined.

4.5.1 Event preselection

For the boosted analysis, the presence of at least 1 boosted jet is required. The trigger discussed in section 4.2.2 is also employed, requiring the AK8 PF jet $H_T > 700$ GeV with an invariant mass > 30 GeV. In order to have an overall good trigger efficiency a threshold on the S_T, which stands for the sum AK8 H_T plus the AK4 H_T, is set to ≥ 900 GeV. Furthermore, we require $p_T > 300$ GeV and $|\eta| < 2.4$ for the leading jet, jet soft-drop mass m_{SD} at least of 70 GeV. A schematic view of the analysis strategy is in Fig 4.36.

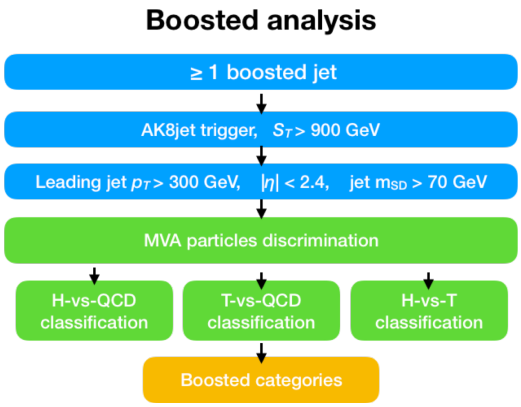

Figure 4.36: Boosted analysis strategy.

4.5.2 Higgs boson or top quark taggers

The idea is to define an algorithm which identifies the AK8 jet as associated to a Higgs boson or to a top quark, again using MVA analysis techniques. Using the three different samples, we will build three classification scores:

- H-vs-QCD score,
- T-vs-QCD score,
- H-vs-T score.

The three MVAs use jet events taken from the three samples, the $t\bar{t}H$ sample is used to define the H-tagger, the $t\bar{t}$ sample is used to define the T-tagger and the QCD sample is used as a multijet background whose jets are generic and neither H- nor T-tagged. The matching of a jet to its corresponding particle, the Higgs boson or a top quark, is performed at the generator level, requiring the particle to be closer to the jet axis within $\Delta R < 0.3$. Subjet variables are then used as input for the three MVAs and are reported in the following, for the three different classifications.

Input variables for the H-vs-QCD classification

The following subjet variables have been used for the H-vs-QCD classification:

- jet τ_1, τ_2, τ_3 (see Fig. 4.37):

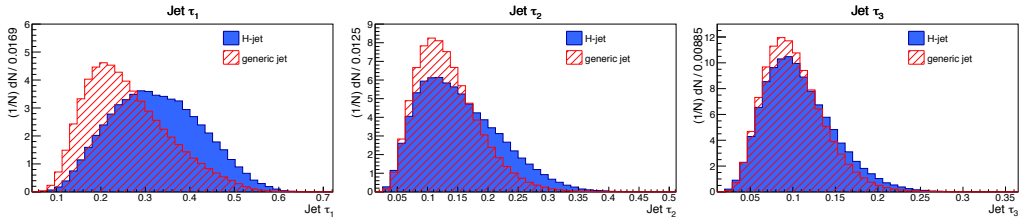

Figure 4.37: Jet τ_1 (left), τ_2 (center), τ_3 (right), for jets originated from the Higgs boson or generic jets.

- b-tagging score of the jet and of the the leading and second subjets (see Fig. 4.38):

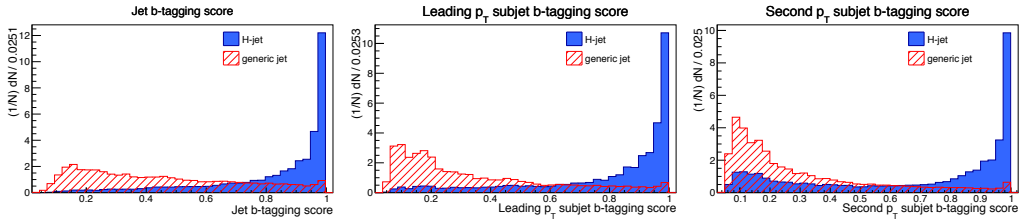

Figure 4.38: b-tagging score of the jet (left), of the leading subjet (center), of the second subjet (right), for jets originated from the Higgs boson or generic jets.

- mass of the leading and second subjects (see Fig. 4.39):

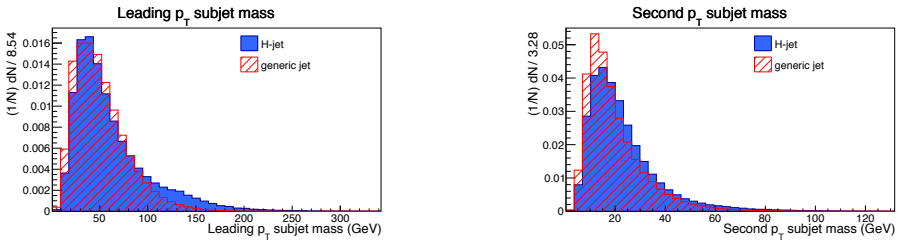

Figure 4.39: Mass of the leading (left) and second (right) subjects, for jets originated from the Higgs boson or generic jets.

Input variables for the T-vs-QCD classification

The following subjet variables have been used for the T-vs-QCD classification:

- jet τ_1, τ_2, τ_3 (see Fig. 4.40):

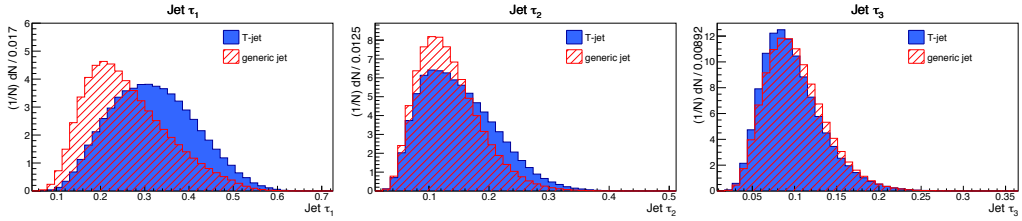

Figure 4.40: Jet τ_1 (left), τ_2 (center), τ_3 (right), for jets originated from the top quark or generic jets.

- b-tagging score of the jet and of the the leading and second subjets (see Fig. 4.41):

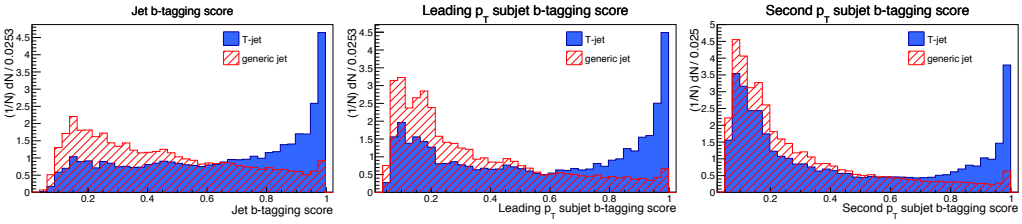

Figure 4.41: b-tagging score of the jet (left), of the leading subjet (center), of the second subjet (right), for jets originated from the top quark or generic jets.

- mass of the leading and second subjects (see Fig. 4.42):

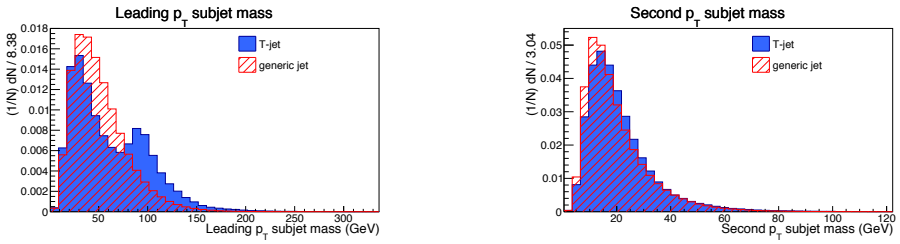

Figure 4.42: Mass of the leading (left) and second (right) subjects, for jets originated from the top quark or generic jets.

Input variables for the H-vs-T classification

The following subjet variables have been used for the H-vs-T classification:

- jet τ_1, τ_2, τ_3 (see Fig. 4.43):

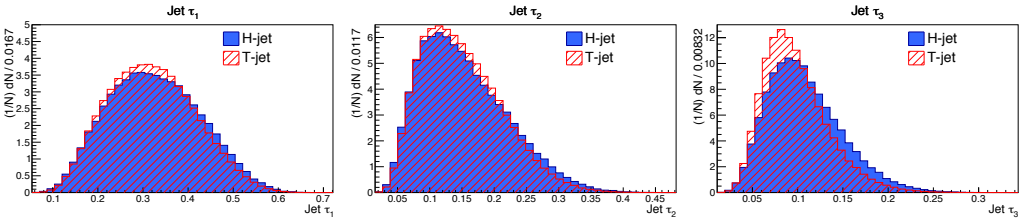

Figure 4.43: Jet τ_1 (left), τ_2 (center), τ_3 (right), for jets originated from the Higgs boson or from the top quark.

- b-tagging score of the jet and of the the leading and second subjets (see Fig. 4.44):

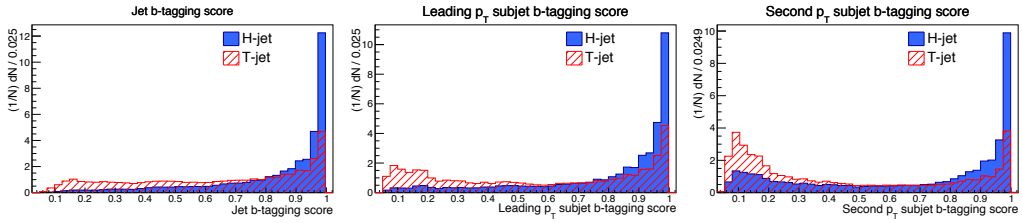

Figure 4.44: b-tagging score of the jet (left), of the leading subjet (center), of the second subjet (right), for jets originated from the Higgs boson or from the top quark.

- mass of the leading and second subjets (see Fig. 4.45):

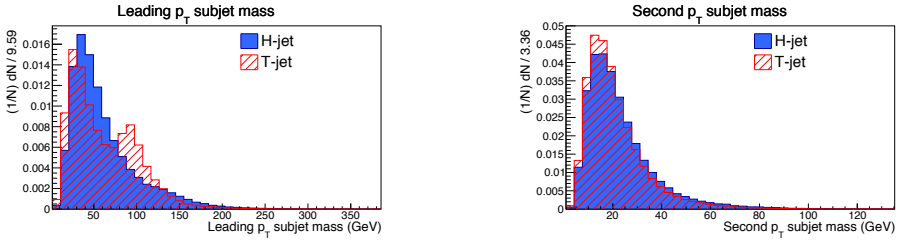

Figure 4.45: Mass of the leading (left) and second (right) subjets, for jets originated from the Higgs boson or from the top quark.

MVA methods

Different MVA methods are employed for the H-vs-QCD, T-vs-QCD and H-vs-T classifications: Fisher (also in the boosted version), ANN, BDT (in the AdaBoost or GradBoost versions) and KNN. These methods and their configuration parameters reflect those previously mentioned in the resolved analysis. A great effort has been made for the H-vs-T classification, which is the most difficult one since subjet variables for $t\bar{t}H$ and $t\bar{t}$ events are more similar than QCD ones, and the b-tagging is expected both for a Higgs boson and for a top quark. The ROC curves for the H-vs-T classification are shown in Fig. 4.46. Once again the BDT with gradient boosting has been chosen as a classifier for the three multivariate analyses. Events are equally divided in trees for training and testing, with 1000 trees for the H-vs-QCD and T-vs-QCD classifications and 3000 trees for the H-vs-T classification. For the H-vs-T classification, the shrinkage parameter is reduced from 0.3 to 0.1 since a low value of it can significantly improve the accuracy of the prediction in difficult settings, even though it demands more trees to be grown. The larger boosting for the H-vs-T permits to increase the statistical separation of the classifier and thus the performance using the same separation criterion for node splitting. Negative event weights entering in the BDTs are again cut off.

Output scores

The output scores of the three BDT discriminators are shown in Figs. 4.47-4.49. Good discrimination power can be appreciated for the H-vs-QCD analysis and quite good for the T-vs-QCD and H-vs-T analyses.

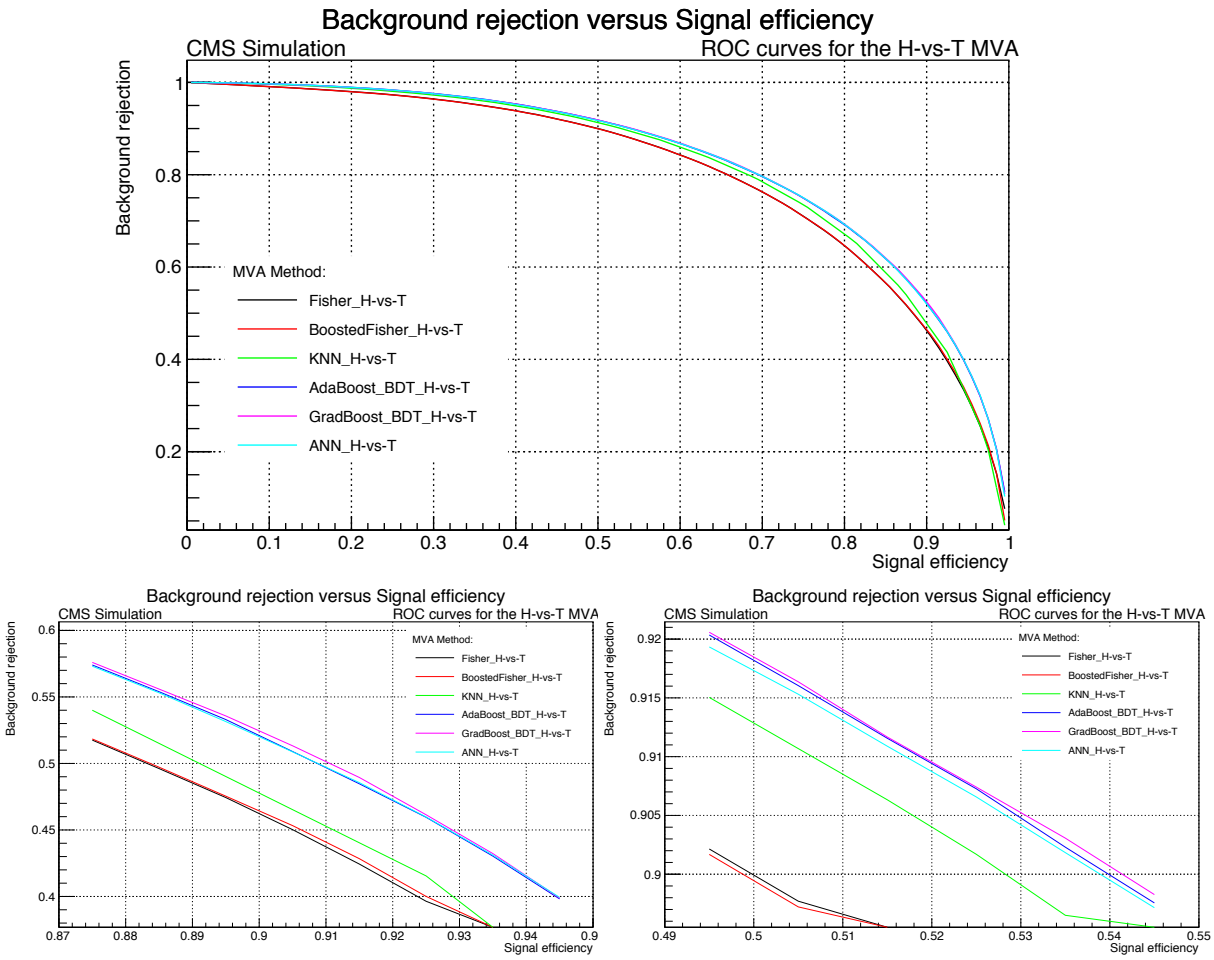

Figure 4.46: ROC curves for the H-vs-T classification (upper panel), and zoom in high- and medium-signal efficiency intervals (lower panel).

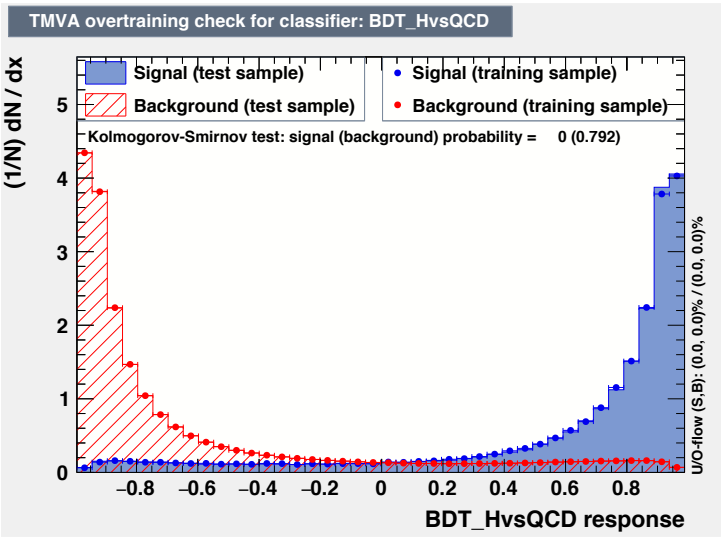

Figure 4.47: BDT output score for the H-vs-QCD classification.

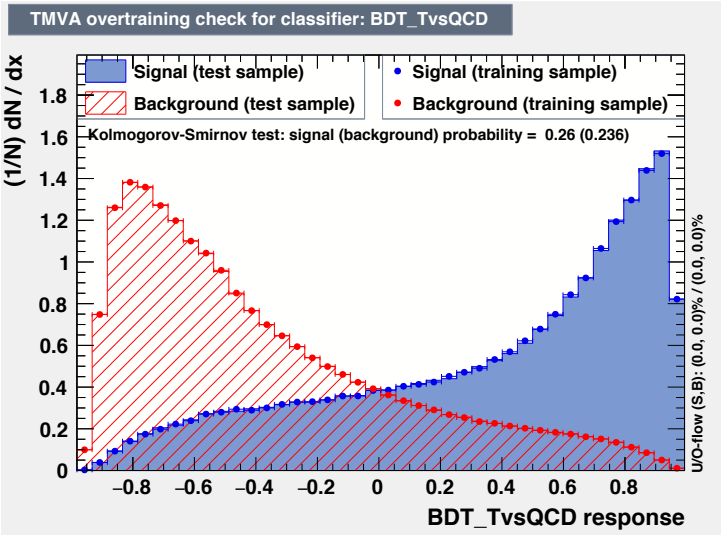

Figure 4.48: BDT output score for the T-vs-QCD classification.

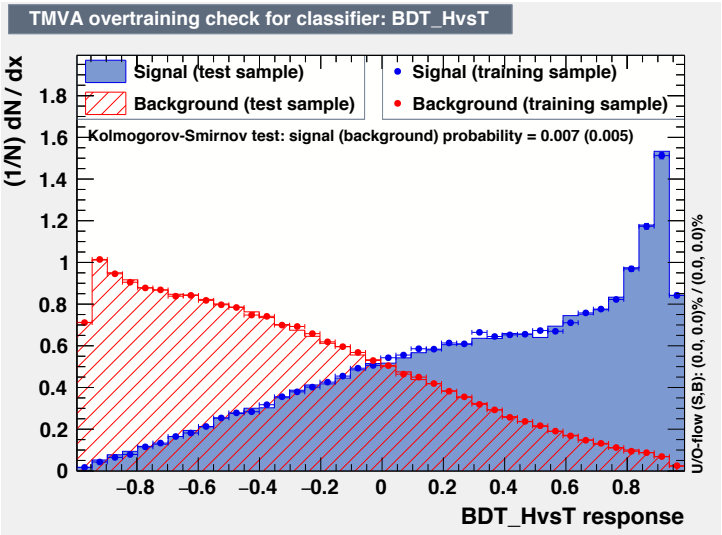

Figure 4.49: BDT output score for the H-vs-T classification.

4.5.3 H- and T-tagging

The identification of a jet as coming from the Higgs boson, from the quark top, or neither one nor the other, is performed through the output scores of the three MVAs, H-vs-QCD, T-vs-QCD, H-vs-T. The $t\bar{t}H$ signal and $t\bar{t}$+QCD background distributions are shown in Figs. 4.50-4.52 as a function of the three BDT output scores. Optimal cuts on these BDT output scores are chosen considering the $S/\sqrt{B}$ figure of merit with values that maximise it. The $S/\sqrt{B}$ figure of merit distributions are also shown in the right panels of Figs. 4.50-4.52 as a function of the three BDT output scores. The optimal values result to be:

- BDT(H-vs-QCD) > 0.8

- BDT(T-vs-QCD) > 0.6
- BDT(H-vs-T) > 0.5

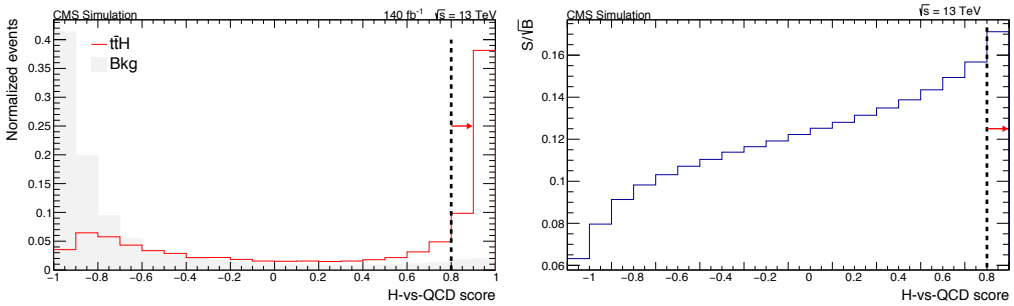

Figure 4.50: (Left) BDT(H-vs-QCD) score for the background ($t\bar{t}$ + QCD) and $t\bar{t}H$ signal samples. (Right) $S/\sqrt{B}$ as a function of the BDT(H-vs-QCD) score.

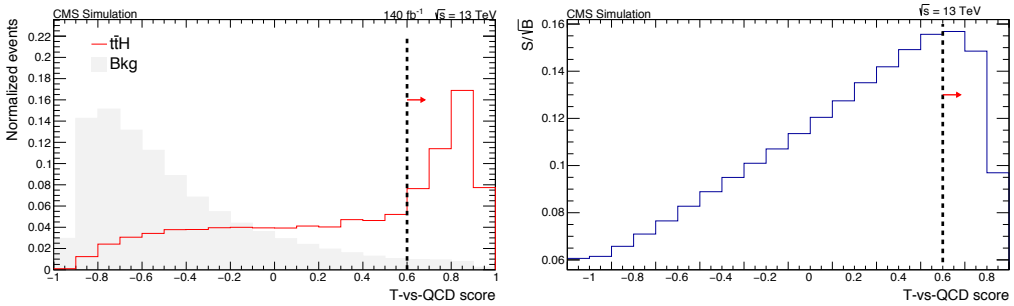

Figure 4.51: (Left) BDT(T-vs-QCD) score for the background ($t\bar{t}$ + QCD) and $t\bar{t}H$ signal sample. (Right) $S/\sqrt{B}$ as a function of the BDT(T-vs-QCD) score.

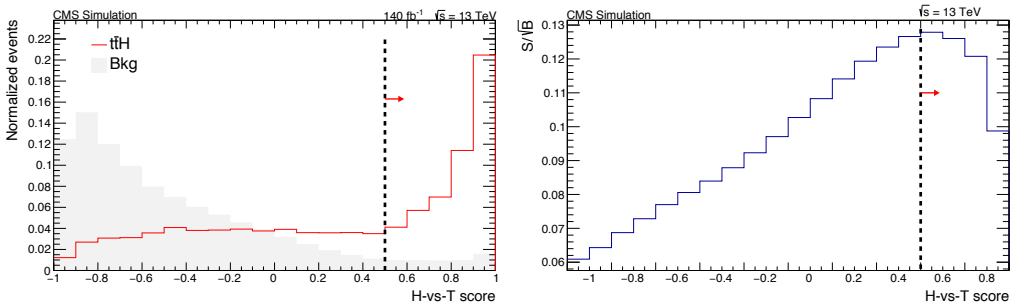

Figure 4.52: (Left) BDT(H-vs-T) score for the background ($t\bar{t}$ + QCD) and $t\bar{t}H$ signal sample. (Right) $S/\sqrt{B}$ as a function of the BDT(H-vs-T) score.

The three BDTs can be used to define two different boosted jet topologies, associated to a Higgs boson or a top quark. In particular, they can be defined as following:

Higgs boson identification (H-tagging)

- jet $p_T > 300$ GeV and jet $m_{SD} > 70$ GeV;

- jet with the highest BDT(H-vs-QCD)+BDT(H-vs-T);
- BDT(H-vs-QCD) > 0.8, BDT(H-vs-T) > 0.5.

Top quark identification (T-tagging)

- jet $p_T > 300$ GeV and jet $m_{SD} > 70$ GeV;
- jet with the highest BDT(T-vs-QCD);
- BDT(T-vs-QCD) > 0.6, BDT(H-vs-T) < 0.5.

The the soft-drop mass distributions for jets associated to the Higgs boson or to the top quark are shown Fig. 4.53, with events taken from the $t\bar{t}H_{H\to b\bar{b}}$ sample. The matching is performed at the generator level, requiring the jet selected by the tagger to be closer to the corresponding particle within $\Delta R < 0.3$. Concerning the Higgs boson tagging, a narrow peak is present in the region of the Higgs boson mass, but part of the events are instead associated to the top quark. This is not surprising since $t\bar{t}H$ consists for the majority of the events of top quarks jets, and it is not so easy to discriminate jets associated to the Higgs boson from jets associated to the top quark (e.g. they are both b-tagged jets, but differing in the number of b-subjets). The QCD multijet production has values more concentrated in the low region of the soft-drop mass. Concerning the top tagging, it can be noticed that the majority of the events are correctly matched, suppressing events that are not associated to the top quark particles at the generator level. Also, a secondary peak in the distribution for jets matched to the top quarks can be noticed around 90 GeV, accounting for the prompt W coming from the decays of top quark. The total number of events for the defined above regimes of the H- and T-tagging is shown in Fig. 4.54. The overall H-tagging efficiency is $\simeq 43\%$, whereas T-tagging efficiency is $\simeq 72\%$.

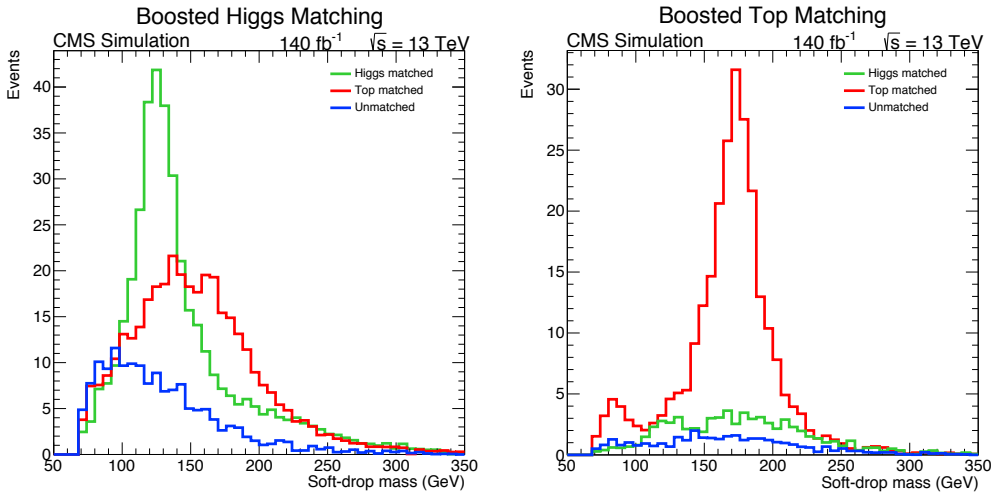

Figure 4.53: Soft-drop mass for H-tagged (left) or T-tagged (right) jets in $t\bar{t}H_{H\to b\bar{b}}$ events. Lines of different colours correspond to jets matched or not at generator level to the Higgs boson or to the top quark.

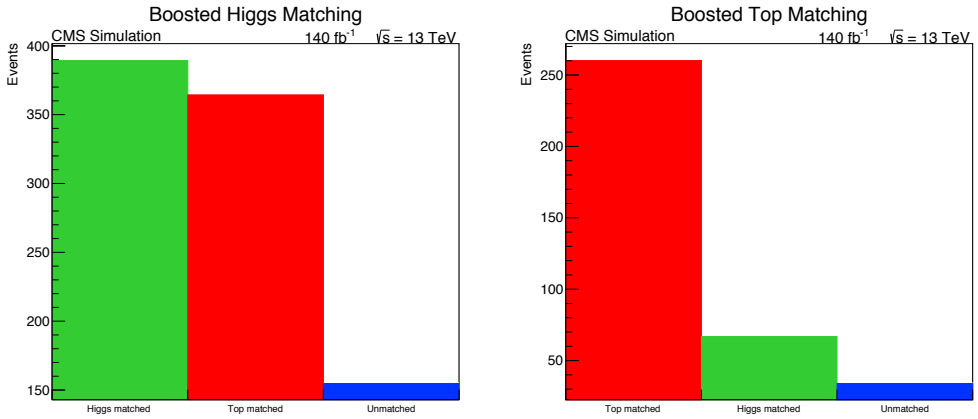

Figure 4.54: H-tagged (left) or T-tagged (right) jets, matched (or not) at generator level to the Higgs boson or to the top quark, in $t\bar{t}H_{H\to b\bar{b}}$ events.

4.5.4 Categories based on H- and T-tagging

The particle taggers can be used to devise new signal categories for the boosted topologies. Six orthogonal categories are defined according to the presence of at least one boosted jet, identified as H- or T-tagged, with the tagging of boosted jets performed according to the previously discussed criteria. Boosted jet multiplicity is split in exactly 1 jet or ≥ 2 jets. It is considered the possibility of having 1 AK8 H-tagged jet (cat. 1), or 1 AK8 T-tagged jet (cat. 2), or having 2 or more AK8 jets, 1 H-tagged and another one T-tagged (cat. 3), at least 1 H-tagged jet (cat. 4), 2 T-tagged jets (cat. 5) or at least 1 T-tagged jet (cat. 6). It is very difficult to have three AK8 jets, and all of them H- or T- tagged, thus it is neglected here the category in which H- tagging is applied once and T-tagging is applied twice. Note that the category with 0 AK8 is neglected, since H- and T- taggers, which are constructed from jet sub-jettiness, can not obviously be constructed for this topology. These categories are indexed and summarised in Table 4.9.

Boosted analysis categories			
Category	Boosted jets	H-tag	T-tag
1	1	$\surd$	$\times$
2	1	$\times$	$\surd$
3	≥ 2	$\surd$	$\surd$
4	≥ 2	$\surd$	$\times$
5	≥ 2	$\times$	$\surd\surd$
6	≥ 2	$\times$	$\surd$

Table 4.9: Preliminary definition of categories, basing on the tagged particles inside boosted jets.

For each category, AK4 jets and b-jets multiplicities are reported in Fig. 4.55, with events taken from $t\bar{t}H_{H\to b\bar{b}}$ sample. The first two categories have higher multiplicities since only 1 AK8 is required, while the other categories show a lower multiplicity of AK4 jets since $\geq$ 2 AK8 jets are required. Given the multiplicity of AK8 jets, the H- or T-tagging selections further modify the resolved multiplicity distributions. These distributions should be considered to define more efficient categories, that will be referred to as signal categories.

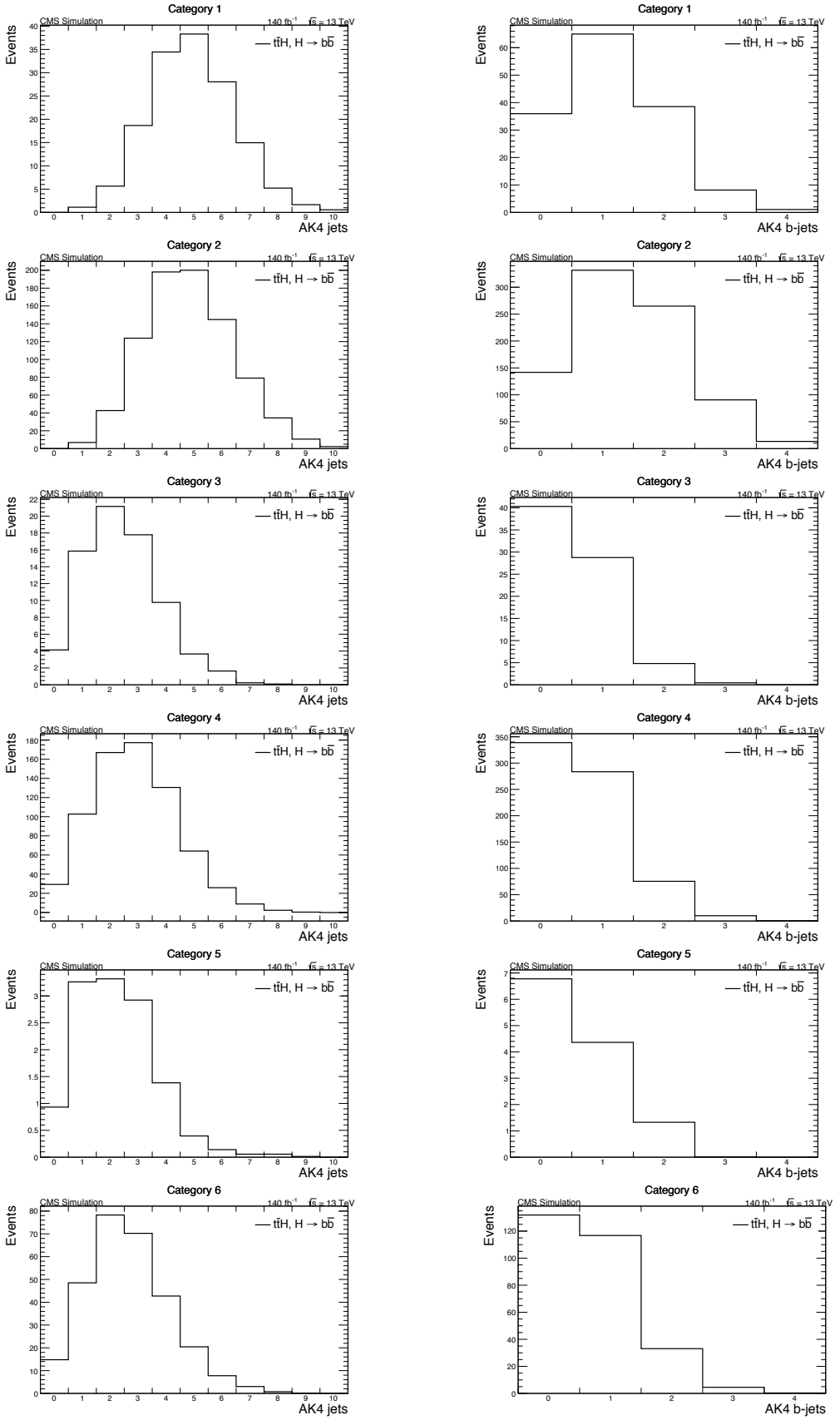

Figure 4.55: Resolved jets (left) and b-jets (right) compositions for the boosted categories, in $t\bar{t}H_{H\to b\bar{b}}$ events.

4.5.5 Signal categories

Signal categories are defined with further requirements on the number of resolved jets and b-jets in addition to the previously discussed selections of Table 4.9. All signal categories require the presence of at least 1 boosted jet that has also been identified, either as coming from a Higgs boson or as coming from a top quark.
More precisely, if a boosted jet is H-tagged, 6 resolved jets are expected in the final states, 2 of which should be b-tagged. The first category is then defined, requiring ≥ 5 resolved jets and ≥ 2 resolved b-jets. The selection on the number of resolved jets also corresponds to the most populated bin, as can be seen for the first category in Fig. 4.55. Instead, if a boosted jet is T-tagged, 5 resolved jets are expected in the final states, 3 of which should be b-tagged, assuming that the Higgs boson decays into a $b\bar{b}$ pair. The second category is then defined, requiring a lower resolved jet multiplicity, corresponding to ≥ 4 jets and ≥ 2 b-jets.
If 2 or more jets are boosted, other categories are considered. If we have a boosted Higgs boson and a boosted top quark, 3 resolved jets are expected, 1 b-tagged. The third category is defined requiring ≥ 2 resolved jets, with ≥ 1 resolved b-jet. The fourth category is related to the third one, requiring the same jet and b-jet multiplicity but is less strict on the tagger, with events required to have only 1 H-tagged jet. If we have 2 boosted jets coming from the top quark, 2 resolved jets, b-tagged, are expected, in the case that the Higgs boson decays into a $b\bar{b}$ pair. The fifth category is defined requiring ≥ 1 resolved jet, with ≥ 1 resolved b-jet. The sixth category is related to the fifth one, with a less strict request on the T-tagger, only 1 T-tagged jet, and ≥ 2 resolved jets, at least 1 b-tagged. The requirements of the signal categories are reported in Table 4.10.

Boosted analysis categories					
Category	Boosted jets	H-tag	T-tag	Resolved jets	Resolved b-jets
1	1	$\surd$	$\times$	≥ 5	≥ 2
2	1	$\times$	$\surd$	≥ 4	≥ 2
3	≥ 2	$\surd$	$\surd$	≥ 2	≥ 1
4	≥ 2	$\surd$	$\times$	≥ 2	≥ 1
5	≥ 2	$\times$	$\surd\surd$	≥ 1	≥ 1
6	≥ 2	$\times$	$\surd$	≥ 2	≥ 1

Table 4.10: Boosted categories.

The expected yields for the six categories are displayed in Fig. 4.56 along with the S/B and $S/\sqrt{B}$. In each category, the different samples contribution are shown superimposed (not stacked), with the y-axis providing the actual category yield. Their values are reported in Table 4.11. It can be noticed that the first two categories have similar S/B. The third category, benefiting from a double (H- and T-) tagging, has the largest S/B compared to all categories and also a good signal significance. The fourth category is the category with the largest number of events, but actually the S/B is the lowest. Not requiring T-tag worsens the S/B but improves $S/\sqrt{B}$ because of the large number of signal events. Double T-tagging for the fifth category results in a very good S/B. The sixth category requires only only T-tag but more resolved jets than the fifth category: this worsens the S/B but improves $S/\sqrt{B}$ because of the large number of signal events.

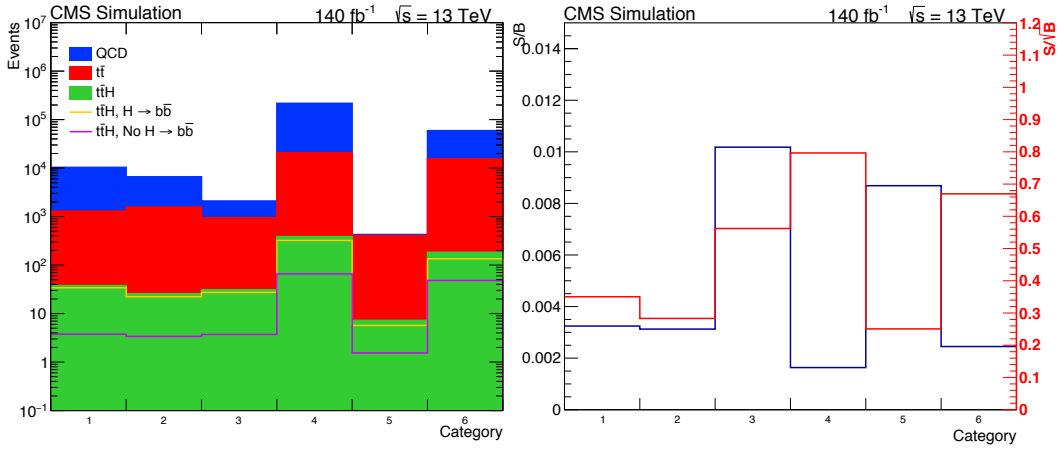

Figure 4.56: Boosted categories events yields (left), S/B and S/$\sqrt{\text{B}}$ (right).

Boosted categories				
Expected yields $L_{ref} = 140\,\text{fb}^{-1}, \sqrt{s} = 13\,\text{TeV}$				
Category	S	B	S/B	S/$\sqrt{\text{B}}$
1	38	11663	1/307	0.35
2	26	8203	1/316	0.29
3	31	3049	1/98	0.56
4	388	237440	1/612	0.80
5	7.2	832.7	1/115	0.25
6	183	74693	1/408	0.67
Total	673	335882	1/499	1.16

Table 4.11: Signal and background expected yields, signal-over-background ratio and signal significance for the defined boosted categories.

Figs. 4.57 and 4.58 show the resolved jet and b-jet multiplicities, respectively, that have been accepted with the definition of the boosted categories, for all the samples. More stringent requirements than those that have been made on the jet and b-jet multiplicity do not lead to an improvement in the sensitivity of the analysis, since a higher multiplicity of jets and b-jets is permitted by the initial and final state radiation, as much as for signal than for background processes. Fig. 4.59 shows jet p_T for the jet that has been H- or T-tagged, depending on the specific category, through the tagging strategy defined in this analysis.

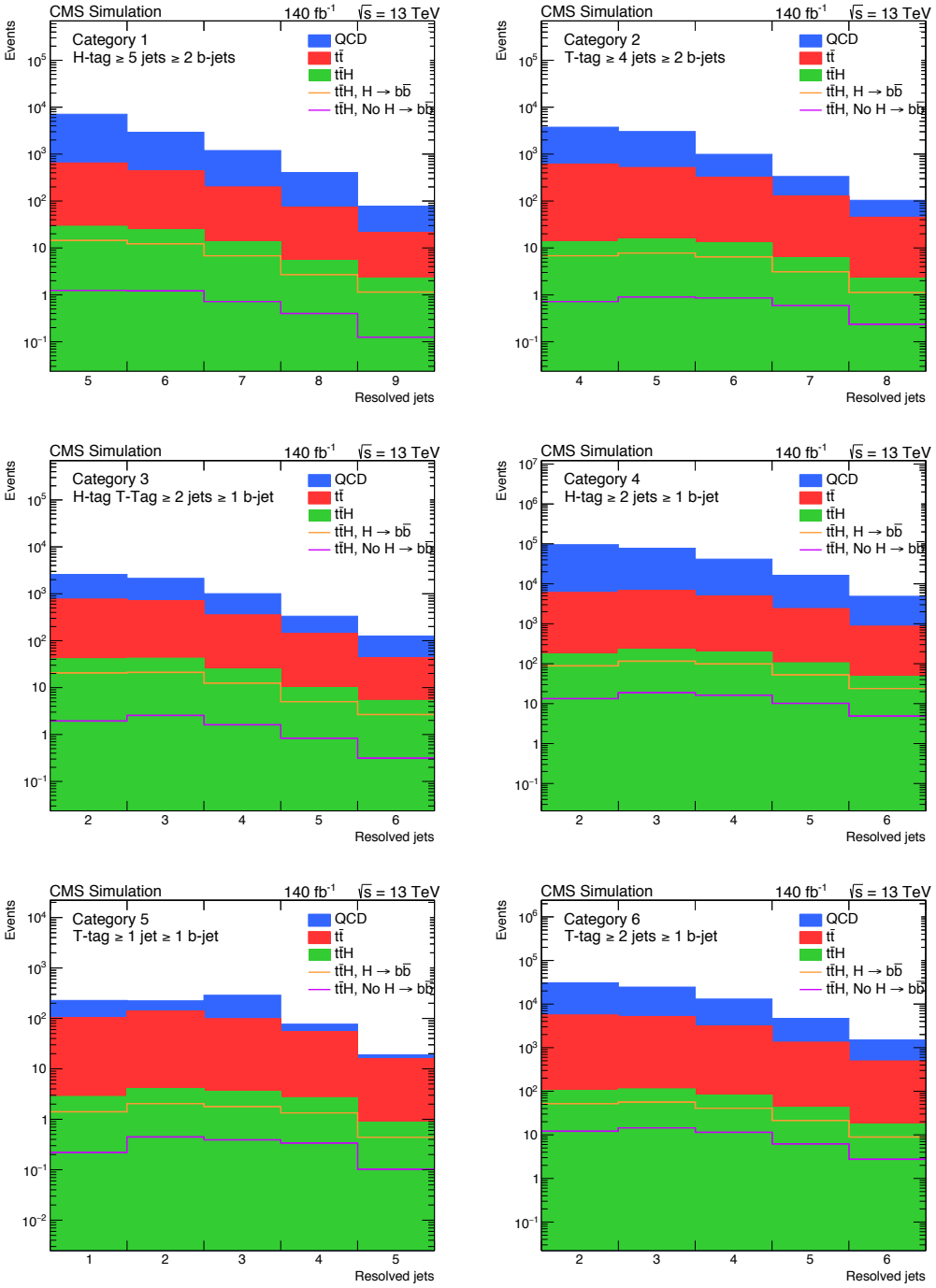

Figure 4.57: Number of resolved jets for the defined boosted categories.

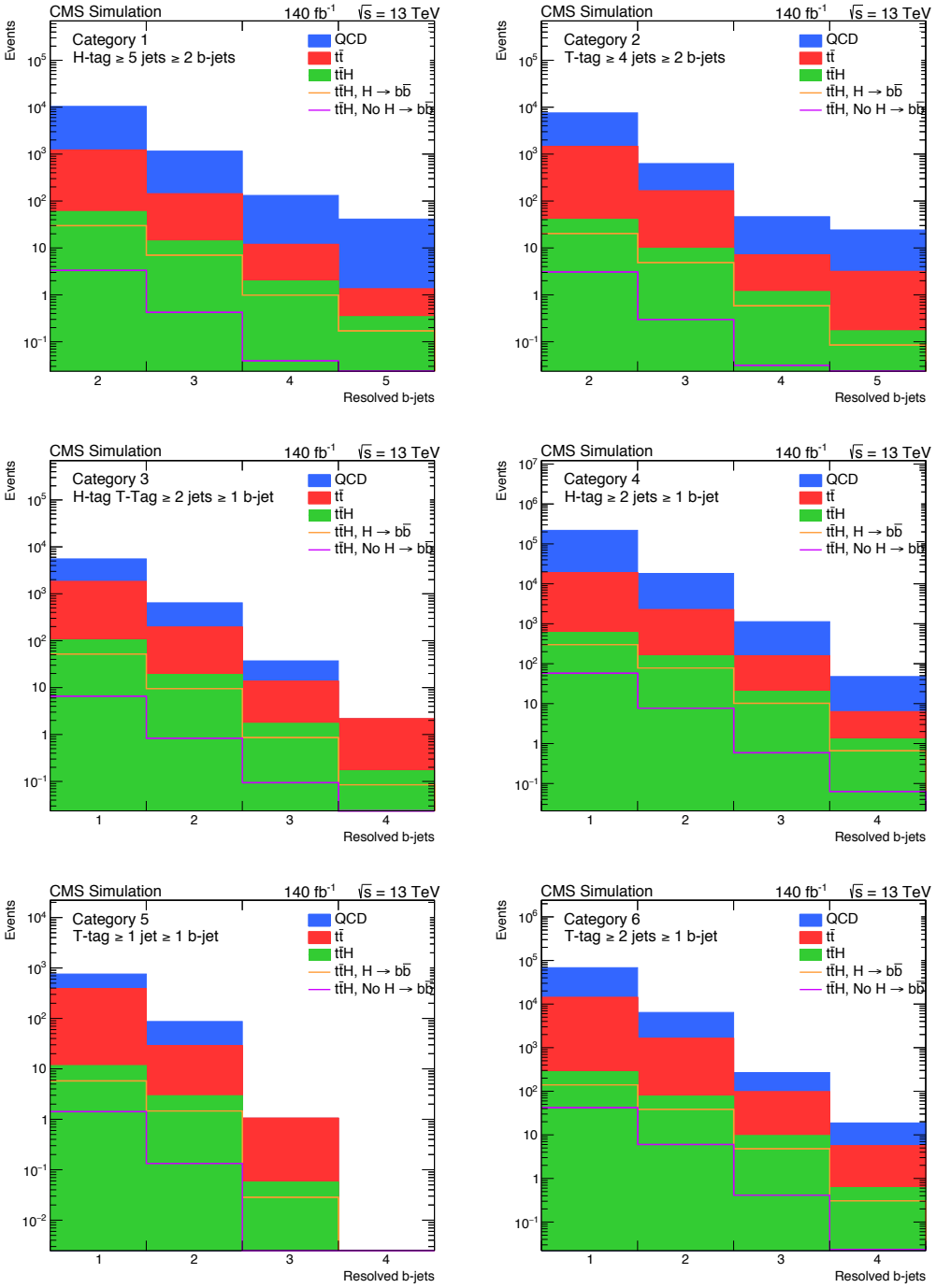

Figure 4.58: Number of resolved b-jets for the defined boosted categories.

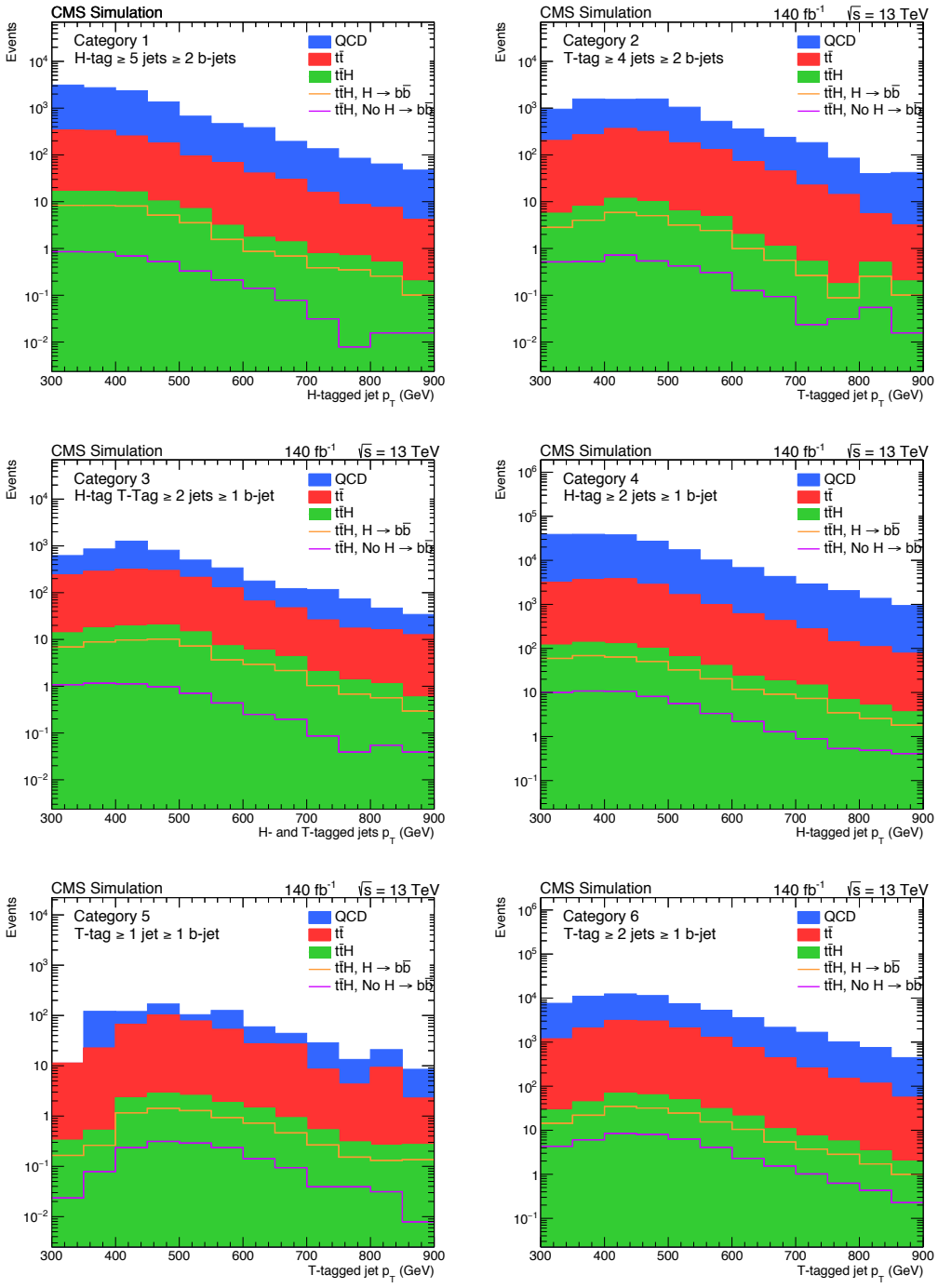

Figure 4.59: p_T of the H- or T- tagged jet for the defined boosted categories.

Chapter 5

Statistical treatment of the expected signal

The expected signal for the $t\bar{t}H$ production in the all-hadronic decay channel is overwhelmed by the QCD multijet and $t\bar{t}$ backgrounds, as shown in the previous chapters. The procedure adopted for the event selection strongly improved the signal-over-background ratio, and increased the signal significance as well. In addition to this, it is convenient to express the expected signal yields in terms of a relevant parameter, the so-called signal strength μ. It is defined as the ratio of the measured $t\bar{t}H$ production cross section $\sigma_{t\bar{t}H}$ to the SM prediction $\sigma_{t\bar{t}H^{SM}}$,

$$\mu = \sigma_{t\bar{t}H}/\sigma_{t\bar{t}H^{SM}}. \tag{5.1}$$

The parameter μ modifies the strength of the signal process, with $\mu = 0$ corresponding to the background-only hypothesis and $\mu = 1$ being the nominal signal hypothesis. In Section 5.1 of this chapter, the statistical procedure for setting upper limits on μ is presented. Expected upper limits on μ are computed under the m_H =125 GeV assumption and reported at 95% CL in Section 5.2, separately both for the resolved and boosted analyses, and for the combination of the two analyses.

5.1 Statistical formalism

The first step is to introduce a binned likelihood function $L(\mu, \boldsymbol{\theta})$ given by the product of Poisson probabilities for all N bins [80],

$$L(\mu, \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-(\mu s_i + b_i)} \prod_{j=1}^{M} \frac{u_j^{m_j}}{m_j!} e^{-u_j} \tag{5.2}$$

with s_i and b_i representing the number of expected signal and background events in each bin and n_i the observed number of the events whose expected value is $E[n_i] = \mu s_i + b_i$. More precisely, $s_i = s_{\text{tot}} \int_{\text{bin } i} f_s(x, \boldsymbol{\theta}_s)$ and $b_i = b_{\text{tot}} \int_{\text{bin } i} f_b(x, \boldsymbol{\theta}_b)$, where s_{tot} and b_{tot} are the total mean numbers of signal and background events, $f_s(x, \boldsymbol{\theta}_s)$, $f_b(x, \boldsymbol{\theta}_b)$ are the probability density functions (pdfs) of the variable x for signal and background events, and $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_b$ represent nuisance parameters that might affect the shapes of pdfs.
In addition to the N values, one often makes further subsidiary measurements that help constrain the set of nuisance parameters $\boldsymbol{\theta}$, selecting some kinematic variables. This then gives a set of values m_j for the number of entries in each of the M bins of these variables whose expectation value is $E[m_j] = u_j(\boldsymbol{\theta})$.

In this analysis, there are also several categories that need to be combined. In this case, a likelihood function $L_k(\mu, \boldsymbol{\theta}_k)$ is defined for each category k, with μ representing the signal strength and $\boldsymbol{\theta}_k$ representing the set of nuisance parameters for the k-th category. The signal strength is assumed to be the same for all categories but in general the set of nuisance parameters can vary between categories. Assuming the categories to be statistically independent, the combined likelihood function is given by the product over all of them,

$$L(\mu, \boldsymbol{\theta}) = \prod_k L_k(\mu, \boldsymbol{\theta}_k). \tag{5.3}$$

To test a hypothesised value of μ, the profile likelihood ratio,

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}, \tag{5.4}$$

is considered, where $\hat{\hat{\boldsymbol{\theta}}}$ in the numerator denotes the value of $\boldsymbol{\theta}$ that maximises L for the specified μ, i.e., it is the conditional maximum-likelihood (ML) estimator of $\boldsymbol{\theta}$ (and thus is a function of μ). The denominator is the maximised likelihood function, i.e., $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$ are their ML estimators. The profile likelihood ratio $\lambda(\mu)$ assumes values between 0 and 1 (at $\mu = \hat{\mu}$), with λ close to 1 implying a good agreement between data and the hypothesised value of μ. The presence of the nuisance parameters broadens the profile likelihood as a function of μ relative to what one would have if their values were fixed. This reflects the loss of information about μ due to the systematic uncertainties. In our analyses, the contribution of the signal process to the mean number of events is assumed to be non-negative. However, it is convenient to define an effective estimator $\hat{\mu}$ as the value of μ that maximises the likelihood, allowing for $\hat{\mu} < 0$, but providing that the Poisson mean values, $\mu s_i + b_i$, remain nonnegative. This will allow us to model $\hat{\mu}$ as a Gaussian-distributed variable, and in this way we can determine the distributions of the test statistics that we consider.
For the purpose of establishing an upper limit on the strength parameter μ, we consider the test statistic $\tilde{q}_\mu$, defined as

$$\tilde{q}_\mu = \begin{cases} -2 \ln \tilde{\lambda}(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$

where $\tilde{\lambda}(\mu)$ is the profile likelihood ratio as defined in Eq. 5.4. The reason for setting $\tilde{q}_\mu = 0$ for $\hat{\mu} > \mu$ is that when setting an upper limit, one would not regard data with $\hat{\mu} > \mu$ as representing less compatibility with μ, and therefore this is not taken as part of the rejection region of the test. From the definition of the test statistic one sees that higher values of $\tilde{q}_\mu$ represent greater incompatibility between the data and the hypothesised value of μ. In our case, $\mu > 0$, so considering $\tilde{\lambda}(\mu)$, we have

$$\tilde{q}_\mu = \begin{cases} -2 \ln \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0, \hat{\hat{\boldsymbol{\theta}}}(0))} & \hat{\mu} < 0 \\ -2 \ln \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})} & 0 \leq \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu. \end{cases}$$

Assuming the Wald approximation [81], we find that

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

where $\hat{\mu}$ follows a Gaussian distribution centred about μ' with a standard deviation σ. The pdf $f(\tilde{q}_\mu|\mu')$ is found to be

$$f(\tilde{q}_\mu|\mu') = \Phi\Big(\frac{\mu'-\mu}{\sigma}\Big)\delta(\tilde{q}_\mu) + \begin{cases} \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{q}_\mu}}\exp\Big[-\frac{1}{2}\Big(\sqrt{\tilde{q}_\mu}-\frac{\mu-\mu'}{\sigma}\Big)^2\Big] & 0<\tilde{q}_\mu\leq\mu^2/\sigma^2 \\ \frac{1}{\sqrt{2\pi}(2\mu/\sigma)}\exp\Big[-\frac{1}{2}\frac{(\tilde{q}_\mu-(\mu^2-2\mu\mu')/\sigma^2)^2}{(2\mu/\sigma)^2}\Big] & \tilde{q}_\mu>\mu^2/\sigma^2 \end{cases} \tag{5.5}$$

where Φ is the cumulative distribution function of the standard normal distribution. The cumulative distribution function corresponding to the pdf $f(\tilde{q}_\mu|\mu')$ is given by

$$F(\tilde{q}_\mu|\mu') = \begin{cases} \Phi(\sqrt{\tilde{q}_\mu}-\frac{\mu-\mu'}{\sigma}) & 0<\tilde{q}_\mu\leq\mu^2/\sigma^2 \\ \Phi(\frac{\tilde{q}_\mu-(\mu^2-2\mu\mu')/\sigma^2}{2\mu/\sigma}) & \tilde{q}_\mu>\mu^2/\sigma^2 \end{cases} \tag{5.6}$$

In the special case $\mu=\mu'$, we have

$$F(\tilde{q}_\mu|\mu) = \begin{cases} \Phi(\sqrt{\tilde{q}_\mu}) & 0<\tilde{q}_\mu\leq\mu^2/\sigma^2 \\ \Phi\Big(\frac{\tilde{q}_\mu+\mu^2/\sigma^2}{2\mu/\sigma}\Big) & \tilde{q}_\mu>\mu^2/\sigma^2 \end{cases} \tag{5.7}$$

The p-value of the hypothesised μ is given by the formula

$$p_\mu = 1 - F(\tilde{q}_\mu|\mu) \tag{5.8}$$

with a corresponding significance

$$Z_\mu = \begin{cases} \sqrt{\tilde{q}_\mu} & 0<\tilde{q}_\mu\leq\mu^2/\sigma^2 \\ \frac{\tilde{q}_\mu+\mu^2/\sigma^2}{2\mu/\sigma} & \tilde{q}_\mu>\mu^2/\sigma^2 \end{cases} . \tag{5.9}$$

If the p-value is found below a specified threshold α (often taken as $\alpha=0.05$), then the value of μ is said to be excluded at a confidence level (CL) of $1-\alpha$. The observed upper limit on μ is the smallest μ such that $p_\mu\leq\alpha$. Therefore, the observed upper limit on μ at CL $1-\alpha$ is found by setting $p_\mu=\alpha$ and solving Eq. 5.8 for μ, that is

$$\mu_{up} = \hat{\mu} + \sigma\Phi^{-1}(1-\alpha). \tag{5.10}$$

If $\alpha=0.05$, then $\Phi^{-1}(1-\alpha)=1.64$. Moreover, σ depends in general on the hypothesised value of μ. Upper limits closer to the hypothesised value of μ correspond to stricter constraints on the hypothesised value of μ and thus higher significance.

It is customary CMS policy not to look at data candidates from the signal selection unless specific parts of the analysis have been approved. For this reason, in the following we will refer to candidates obtained as a sum of simulated signal and backgrounds (the so-called Asimov dataset), and we will compute the expected (more precisely, median) upper limits on μ. In this case, one has to consider a specific Asimov dataset, where data are set to their expectation values and are the same that would be estimated from the MC model using a very large data sample. Using the statistic $\tilde{q}_\mu$ one finds the same expression for the upper limit at a confidence level of $1-\alpha$, namely Eq. 5.10. Therefore, the median upper limit assuming a strength parameter μ' is found simply by substituting this for $\hat{\mu}$, and the $\pm N\sigma$ error bands are found similarly by substituting the corresponding values of $\mu'\pm N\sigma$. The median upper limit is given by

$$\text{median } \mu_{up} = \mu' + \sigma\Phi^{-1}(1-\alpha), \tag{5.11}$$

and the $\pm N\sigma$ error band is given by

$$\text{band } N\sigma = \mu' + \sigma(\Phi^{-1}(1-\alpha)\pm N). \tag{5.12}$$

The standard deviation σ of $\hat{\mu}$ can be obtained from the Asimov value of the test statistic $\tilde{q}_\mu$ recovering the known properties of $-\lambda_A(\mu)$, such that

$$-2\ln\lambda_A(\mu) \approx \frac{(\mu-\mu')^2}{\sigma^2} = \Lambda, \tag{5.13}$$

where Λ is the non-centrality parameter. For the special case $\mu' = \mu$ one has $\Lambda = 0$ and $-2\ln\lambda_A(\mu)$ approaches a χ^2 distribution for one degree of freedom, a result shown by Wilks [82]. Equivalently, one can use Eq. 5.13 to obtain an estimate of the variance σ^2 which characterises the distribution of $\hat{\mu}$, namely,

$$\sigma_A^2 = \frac{(\mu-\mu')^2}{\tilde{q}_{\mu,A}}, \tag{5.14}$$

where $\tilde{q}_{\mu,A} = -2\ln\lambda_A(\mu)$. For the important case where one wants to find the median exclusion significance for the hypothesis μ assuming that there is no signal, then one has $\mu' = 0$ and therefore

$$\sigma_A^2 = \frac{\mu^2}{\tilde{q}_{\mu,A}}. \tag{5.15}$$

5.2 Expected upper limits on the signal strength

All of the formulas presented in Section 5.1 are implemented inside a CMS tool named "Combine". In our case, we use it to perform two kinds of fit:

- one considering only the yields for each category (counting experiment), corresponding to assuming $N = 1$ and no subsidiary measurements in Eq. 5.2;
- one which considers also the distribution of a given variable (shape analysis).

The expected upper limits on the signal strength μ_{up} are computed setting $\alpha = 0.05$ in Eq. 5.11, corresponding to upper limits at 95% CL. The smaller this value, the more sensitive is the experiment to the signal hypothesis.

5.2.1 Upper limits from the counting experiment

The results of the expected upper limits from the counting fit, obtained from the yields of each category, are reported in Table 5.1 for the resolved categories and in Table 5.2 for the boosted categories. The median expected value for the upper limit on the signal strength combining all the defined categories is reported in Table 5.3 and results to be $\mu_{up} = 0.61$.

Resolved categories			
Expected upper limits (counting fit)			
Category	Median μ_{up}	$\pm1\sigma$	$\pm2\sigma$
1	2.14	[1.54, 2.97]	[1.15, 3.95]
2	1.41	[1.01, 1.96]	[0.76, 2.62]
3	1.65	[1.19, 2.29]	[0.89, 3.04]
4	1.71	[1.23, 2.39]	[0.92, 3.18]
5	1.71	[1.23, 2.39]	[0.93, 3.17]
6	1.43	[1.03, 1.99]	[0.77, 2.67]
Combined	0.66	[0.48, 0.92]	[0.36, 1.22]

Table 5.1: Expected upper limits on the signal strength μ_{up} computed from the yields in each resolved category.

Boosted categories			
Expected upper limits (counting fit)			
Category	Median μ_{up}	$\pm1\sigma$	$\pm2\sigma$
1	5.63	[4.05, 7.84]	[3.03, 10.41]
2	6.97	[5.01, 9.72]	[3.76, 12.95]
3	3.52	[2.54, 4.93]	[1.90, 6.57]
4	2.46	[1.78, 3.42]	[1.34, 4.55]
5	8.00	[5.72, 11.22]	[4.31, 15.06]
6	2.94	[2.11, 4.07]	[1.58, 5.42]
Combined	1.46	[1.05, 2.03]	[0.79, 2.69]

Table 5.2: Expected upper limits on the signal strength μ_{up} computed from the yields in each boosted category.

Expected upper limits (counting fit)			
Category	Median μ_{up}	$\pm1\sigma$	$\pm2\sigma$
Resolved categories	0.66	[0.48, 0.92]	[0.36, 1.22]
Boosted categories	1.46	[1.05, 2.03]	[0.79, 2.69]
All categories combined	0.61	[0.44, 0.84]	[0.33, 1.12]

Table 5.3: Expected upper limit on the signal strength μ_{up} computed combining categories.

5.2.2 Upper limits from the shape analysis

In the case of the shape analysis fit, one chooses a variable whose distribution is used for the computation of the upper limits, in addition to the numerical values of the yields (which are also present, being equal to the normalisation of the input histograms). The variable used for the resolved analysis is the mass of the jet pair which has the minimum separation ΔR, while the variable used for the boosted analysis is the soft-drop mass of the H- or T-tagged jets. For the boosted categories, we have one tagged jet (H or T) apart from the third category, where two jets are tagged (one H and one T), and the fifth category, with two T-tagged jets. In these two cases, we choose the soft-drop mass of only one tagged jet, that is the H-tagged one for the third category and the T-tagged jet with the highest p_T for the fifth category. The distributions used for the computation of the expected upper limit on the signal strength with the shape analysis are presented in Fig 5.1 for the resolved categories and in Fig 5.2 for the boosted categories. The results of the upper limits are reported in Table 5.4 for the resolved categories and in Table 5.5 for the boosted categories. The median expected value for the upper limit on the signal strength using all the defined categories at 95% CL is reported in Table 5.6 and results to be $\mu_{up} = 0.49$.

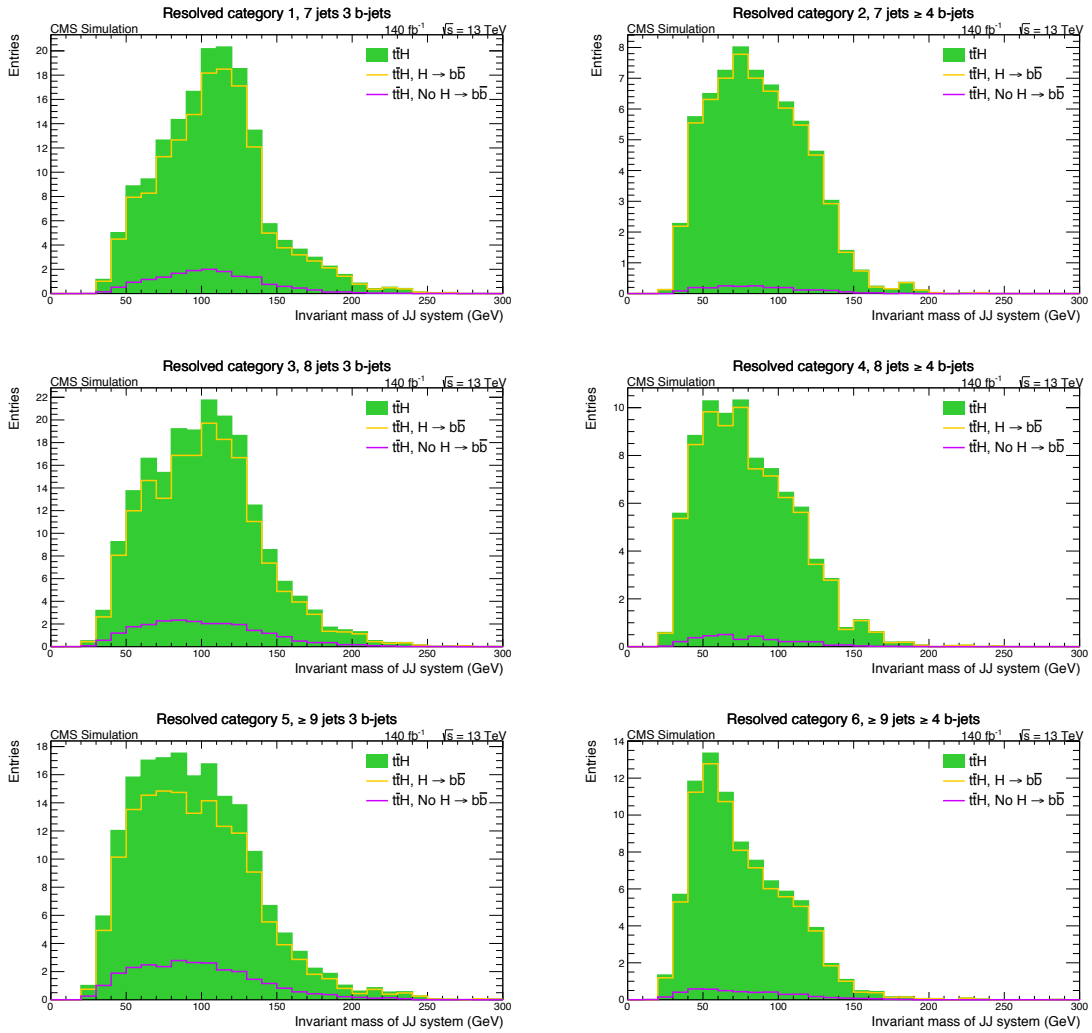

Figure 5.1: Invariant dijet mass distributions in $t\bar{t}H$ events for the defined resolved categories.

Resolved categories			
Expected upper limits (shape fit)			
Category	Median μ_{up}	$\pm1\sigma$	$\pm2\sigma$
1	1.70	[1.21, 2.39]	[0.89, 3.20]
2	1.20	[0.84, 1.70]	[0.62, 2.29]
3	1.42	[1.01, 1.99]	[0.75, 2.65]
4	1.37	[0.97, 1.93]	[0.72, 2.58]
5	1.52	[1.09, 2.11]	[0.82, 2.81]
6	1.11	[0.79, 1.56]	[0.59, 2.09]
Combined	0.53	[0.37, 0.74]	[0.27, 1.00]

Table 5.4: Expected upper limits on the signal strength μ_{up} computed from the shape of the invariant mass of the closest jet pair in each resolved category.

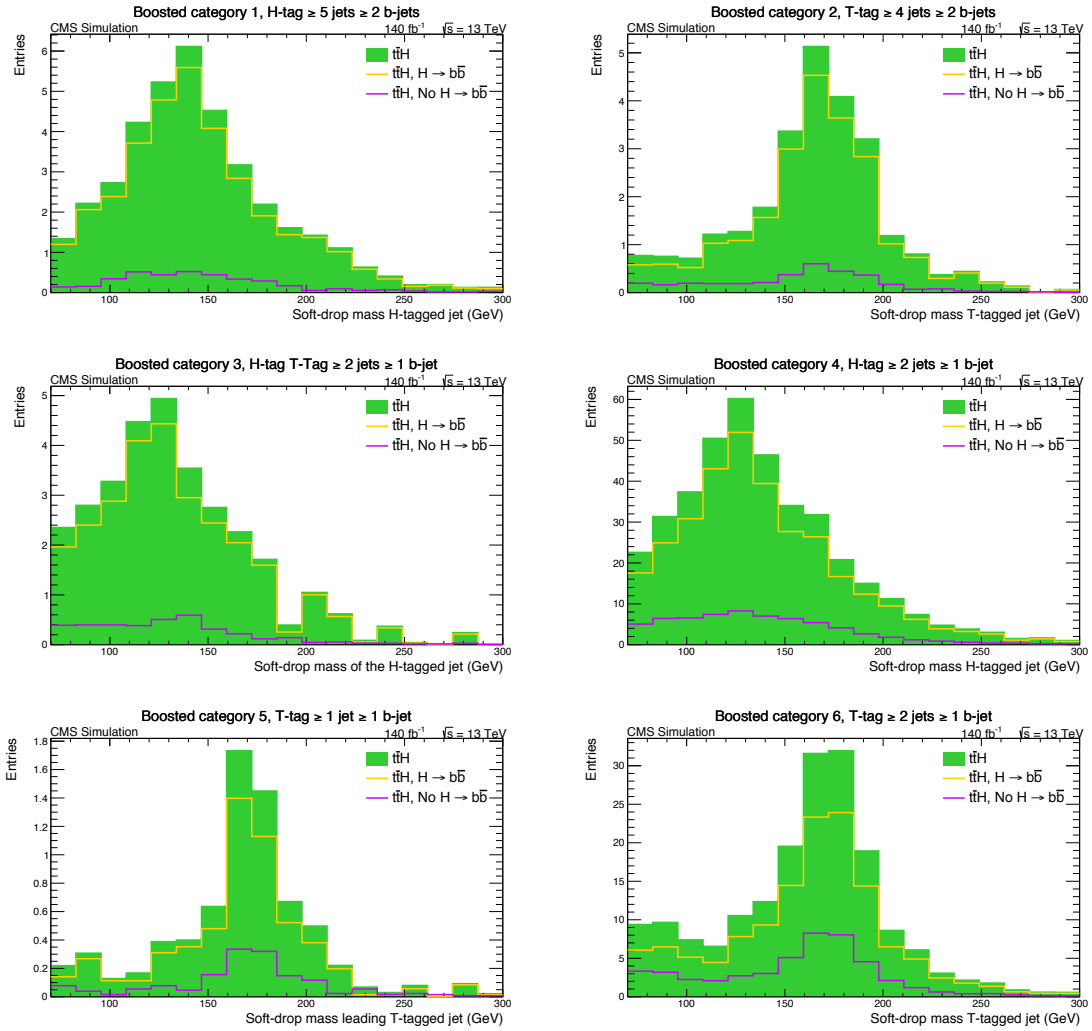

Figure 5.2: Soft-drop mass distributions in $t\bar{t}H$ events for the boosted categories.

Boosted categories			
Expected upper limits (shape fit)			
Category	Median μ_{up}	$\pm1\sigma$	$\pm2\sigma$
1	5.14	[3.70, 7.17]	[2.77, 9.58]
2	6.47	[4.65, 8.97]	[3.49, 12.02]
3	3.14	[2.25, 4.38]	[1.69, 5.87]
4	2.32	[1.67, 3.24]	[1.26, 4.30]
5	6.34	[4.53, 9.00]	[3.37, 12.17]
6	2.73	[1.97, 3.79]	[1.47, 5.05]
Combined	1.39	[1.01, 1.94]	[0.76, 2.58]

Table 5.5: Expected upper limits on the signal strength μ_{up} computed from the shape of the soft-drop mass in each boosted category.

Expected upper limits (shape fit)			
Category	Median μ_{up}	$\pm1\sigma$	$\pm2\sigma$
Resolved categories	0.53	[0.37, 0.74]	[0.27, 1.00]
Boosted categories	1.39	[1.01, 1.94]	[0.76, 2.58]
All categories combined	0.49	[0.35, 0.69]	[0.25, 0.93]

Table 5.6: Expected upper limit on the signal strength μ_{up} computed combining categories.

5.2.3 Expected significance

With the Combine tool, we can compute the a-priori expected significance, which does not depend on the observed data, and so is a good metric to optimize an analysis when still blinded, like in our case. The $S/\sqrt{B}$ expected signal significance is widely used in particle physics data analyses, but it is an approximation strictly valid only in cases where $S \ll B$. The expected significance is computed releasing this approximation, with the general formula of Eq. 5.9. The values of expected significance coming from the counting fit are reported in Table 5.7 for both the resolved and boosted categories. The combination of all categories yields a signal significance of 3.236.

Expected significance (counting fit)					
Resolved categories			Boosted categories		
Category	$S/\sqrt{B}$	Z_μ	Category	$S/\sqrt{B}$	Z_μ
1	0.920	0.919	1	0.350	0.350
2	1.414	1.407	2	0.283	0.283
3	1.196	1.195	3	0.562	0.561
4	1.157	1.154	4	0.797	0.796
5	1.148	1.147	5	0.250	0.250
6	1.387	1.382	6	0.666	0.679
Combined	2.522	2.968	Total	1.161	1.348
All categories combined: $S/\sqrt{B} = 2.209$, $Z_\mu = 3.236$					

Table 5.7: Expected significance from the resolved (left) and boosted (right) categories.

5.2.4 Upper limits including systematic uncertainties

Expected values of the upper limits of the signal strength must take into account the systematic uncertainties of the involved processes. Systematic uncertainties are handled by introducing nuisance parameters θ on observables with a pdf $\rho(\theta)$ associated with the best estimate of the nuisance $\tilde{\theta}$ and some other parameter characterising the overall shape of the pdf, and in particular its width. The simplest choice of the pdf for systematic uncertainties is the Gaussian distribution

$$\rho(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(\theta - \tilde{\theta})^2}{2\sigma^2}\Big). \tag{5.16}$$

However, the Gaussian distribution reveals to be not suitable for positively defined observables, like cross sections, cut efficiencies, integrated luminosity, etc. A valid alternative is represented by the so-called "log-normal" distribution (sometimes referred to as lnN)

$$\rho(\theta) = \frac{1}{\sqrt{2\pi}\ln(\kappa)} \exp\Big(- \frac{(\ln(\theta/\tilde{\theta}))^2)}{2(\ln\kappa)^2}\Big)\frac{1}{\theta}. \tag{5.17}$$

The width of the log-normal pdf is characterised by the parameter κ. For example, if $\kappa = 1.10$, the observable can be larger or smaller by a factor 1.10 (both deviations having a chance of 16%). For small uncertainties, the Gaussian distribution with a relative uncertainty ϵ and the log-normal with $\kappa = 1 + \epsilon$ are asymptotically identical, however the log-normal pdf is a more appropriate choice for large uncertainties. The log-normal distribution has a longer tail compared to the Gaussian and goes to zero at $\theta = 0$ [36]. It is the log-normal pdf that was chosen for all uncertainties that are deemed to be correlated between ATLAS and CMS when the Higgs boson was searched at LHC.
Another possibility is represented by the "log-uniform" distribution (sometimes referred to as lnU). This distribution is useful when one wants to set a large a-priori uncertainty on a given background and then rely on the correlation between channels to constrain it. A value of $1 + \epsilon$ for the systematic uncertainty implies that the yield x of a certain background is allowed to float freely between $x(1 + \epsilon)$ and $x/(1 + \epsilon)$ and, in particular, if ϵ is small, then this is approximately $(x - \Delta x, x + \Delta x)$ with $\epsilon = \Delta x/x$.
Lacking a complete determination of the systematic uncertainties, here we use some preliminary values. The systematic uncertainty on the amount of $t\bar{t}$ background is set to be 1.10 log-normal, while the systematic uncertainty on the amount of QCD background is set to be 2.00 log-uniform. The systematic uncertainty for the expected $t\bar{t}H$ yield is set to be 1.10 log-normal, however the yields are so small with the respect to QCD and $t\bar{t}$ yields, that it does not produce any variation on the values of the upper limits. Expected upper limits of the signal strength are computed again using the shape fit which has revealed to be the best choice, with the aforementioned systematic uncertainties included. The results of the upper limits are reported in Table 5.8 for the resolved categories and in Table 5.9 for the boosted categories. The median expected upper limit on the signal strength using all the categories at 95% CL is reported in Table 5.10 and amounts to $\mu_{up} = 0.93$. The expected significance is computed again using the shape fit and including the systematic uncertainties, and reported in Table 5.11 for both the resolved and boosted categories. The combination of all the categories, including systematic uncertainties, yields a signal significance of 3.217. Both for the upper limit and the significance, the inclusion of systematic uncertainties has worsened the values, as expected.

Resolved categories			
Expected upper limits (shape fit) systematic uncertainties included			
Category	Median μ_{up}	$\pm1\sigma$	$\pm2\sigma$
1	6.13	[4.20, 8.79]	[2.97, 11.86]
2	2.66	[1.85, 3.77]	[1.33, 5.03]
3	6.11	[4.23, 8.67]	[3.03, 11.57]
4	2.98	[2.12, 4.16]	[1.56, 5.52]
5	6.88	[5.03, 9.37]	[3.81, 12.21]
6	2.63	[1.90, 3.62]	[1.42, 4.76]
Combined	0.97	[0.66, 1.42]	[0.47, 1.97]

Table 5.8: Expected upper limits on the signal strength μ_{up} including systematic uncertainties in each resolved category.

Boosted categories			
Expected upper limits (shape fit) systematic uncertainties included			
Category	Median μ_{up}	$\pm1\sigma$	$\pm2\sigma$
1	12.06	[8.70, 16.82]	[6.55, 22.33]
2	17.19	[12.44, 23.76]	[9.40, 31.45]
3	6.38	[4.57, 8.94]	[3.41, 12.00]
4	8.59	[6.22, 11.85]	[4.70, 15.70]
5	9.59	[6.85, 13.61]	[5.10, 18.35]
6	14.44	[10.54, 19.79]	[7.95, 25.91]
Combined	3.45	[2.45, 4.87]	[1.83, 6.67]

Table 5.9: Expected upper limits on the signal strength μ_{up} including systematic uncertainties in each boosted category.

Expected upper limits (shape fit) systematic uncertainties included			
Category	Median μ_{up}	$\pm1\sigma$	$\pm2\sigma$
Resolved categories	0.97	[0.66, 1.42]	[0.47, 1.97]
Boosted categories	3.45	[2.45, 4.87]	[1.83, 6.67]
All categories combined	0.93	[0.63, 1.34]	[0.45, 1.86]

Table 5.10: Expected upper limit on the signal strength μ_{up} computed combining categories, including systematic uncertainties.

Expected significance (shape fit) systematic uncertainties included			
Resolved categories		Boosted categories	
Category	Z_μ	Category	Z_μ
1	1.0681	1	0.162
2	1.814	2	0.113
3	1.020	3	0.314
4	1.207	4	0.224
5	0.271	5	0.215
6	1.053	6	0.131
Combined	3.164	Total	0.584
All categories combined: $Z_\mu = 3.217$			

Table 5.11: Expected significance from the resolved (left) and boosted (right) categories, including systematic uncertainties.

Chapter 6

Conclusions

The Higgs boson production mode in association with top quarks provides access to a direct measurement of the top quark Yukawa coupling to the Higgs boson, which is a strong test of the SM consistency. In this work, the SM $t\bar{t}H$ production has been investigated using simulated samples both for signal and for background events. Events have been scaled to the integrated luminosity of the full Run2 dataset, corresponding to 140 fb^{-1} with a center-of-mass energy of 13 TeV. CMS official measurements for the $t\bar{t}H$ production at 140 fb^{-1} have not yet been released and this is a preliminary study of the all-hadronic final state for $t\bar{t}H$ that would be useful when the analysis of the Run2 full dataset will be finalised and expected yields for the events will be compared with the observed ones.

The $t\bar{t}H$ channel has been studied in the all-hadronic topology, characterised by the presence of jets in the final states. A lepton veto ensures that leptonic final states in $t\bar{t}H$ production are not considered, as these are covered by separate searches at CMS. Jets can be resolved or boosted, depending on the particle p_T which originates them. Some preliminary categories based on jet multiplicity have been defined and the most promising have been evaluated. Then the analysis has been subdivided in two parts, one based on the resolved topology, in which all particles originate resolved jets, and one based on the boosted topology, where at least one jet is clustered as boosted. For each of the two topologies, preliminary cuts related to the kinematic properties of the events and to the trigger efficiencies have been made.

In the case of the resolved topology, two multivariate analyses have been performed, in order to discriminate as efficiently as possible the signal from the background. The first multivariate analysis considers the $t\bar{t}H$ signal with respect to the multijet QCD production, the second one the signal with respect to the $t\bar{t}$ background. Orthogonal categories have been defined with respectively 7, 8 and $\geq$9 jets, each of these split into 3 or $\geq$4 b-jets categories. For each category, the expected upper limit on the signal strength μ has been reported, together with the combined value for all the resolved categories. These upper limits show a behaviour qualitatively similar to those of the CMS published analysis [41], with the ≥ 4 b-jets categories yielding tighter constraints on the signal strength. The CMS published analysis uses the same multiplicity of jets in the definition of the aforementioned categories; however, the numerical values reported here cannot be directly compared as the events are scaled to a different integrated luminosity, a complete treatment of the systematic uncertainties is missing and different techniques to discriminate the signal from the background are employed.

In the case of the boosted topology, we have tried to identify the boosted jet (coming from the Higgs boson or from the top quark) using multivariate algorithms that use prop-

erties of subjets that make up the jet itself as input variables for the training. This is a new approach which differs from the jet-based use of the quark-gluon likelihood discriminator or the matrix element method employed in the past analyses, and here has proved to be useful for tagging the particle which originates the jets. Subsequent categories according to the jets multiplicity have been defined and for each of those, the expected upper limits on μ have been calculated, together with the combined value for all the boosted categories.

The results of the two analyses are reported in Fig. 6.1, including preliminary values for the associated systematic uncertainties.

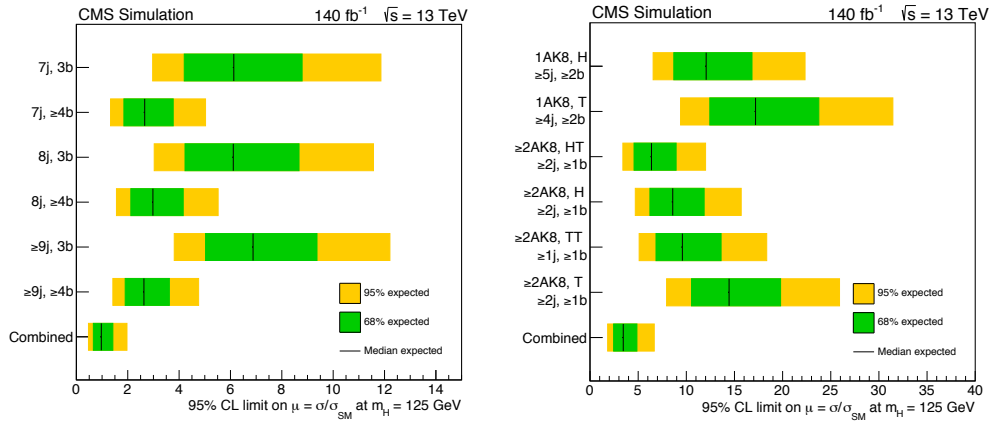

Figure 6.1: 95% CL expected upper limits on μ. The median expected limits are displayed together with their 68% and 95% CL intervals for the resolved (left) and boosted (right) analyses. For each of the two analyses the combined upper limit is also shown. A preliminary assumption of the systematic uncertainties is used.

The resolved categories have lower expected upper limits, and this is due to the fact that the signal contribution with respect to the background is higher, resulting as the most promising. The overall result indicates a median expected upper limit on the signal strength of 0.93 at 95% CL, which is lower than that obtained from the resolved analysis alone, a clear evidence that considering also boosted topologies together with resolved ones improve the efficiency of the analysis in fixing the upper limits. Also, the expected significance improves, resulting to be 3.2 when all the categories are combined. Clearly, the gain of including the boosted categories is still minimal, but hopefully the performance of the H- and T-taggers could be improved by further studies, along with the background rejection.
The future of this analysis appears very promising, and when the candidates of the full Run2 dataset will be analysed, the expected limits and signal significance can be compared with the observed ones. Future datasets will lead to more stringent limits and a higher significance.

List of Figures

List of Tables

Bibliography

[1] I. J. R. Aitchison, A. J. G. Hey, "Gauge Theories in Particle Physics: A Practical Introduction, Volume 1: From Relativistic Quantum Mechanics To QED", 2013 (4th ed), cds.cern.ch:1507184.

[2] I. J. R. Aitchison, A. J. G. Hey, "Gauge Theories in Particle Physics: A Practical Introduction, Volume 2: Non-Abelian Gauge Theories: QCD and The Electroweak Theory", 2013 (4th ed), cds.cern.ch:1507184.

[3] F. Halzen, A. D. Martin, "Quarks and leptons: An introductory course in modern particle physics", 1984, cds.cern.ch:111880.

[4] C. Quigg, "Gauge Theories of the Strong, Weak, and Electromagnetic Interactions", 2013 (2nd ed).

[5] M. Tanabashi et al., "Particle Data Group", Phys. Rev. D 98, 2019, pdg.lbl.gov.

[6] D. J. Griffiths, "Introduction to elementary particles", 2008 (2n ed), cds.cern.ch:111880.

[7] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", Phys.Lett. B 716 (2012) 1-29, doi:10.1016/j.physletb.2012.08.020, arXiv:1207.7214.

[8] 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, doi:10.1016/j.physletb.2012.08.021, arXiv:1207.7235.

[9] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons", Phys. Rev. Lett. 13 (1964), doi: 10.1103/PhysRevLett.13.321, PhysRevLett.13.321.

[10] P. W. Higgs, "Broken Symmetries and the Masses of Gauge Bosons", Phys. Rev. Lett. 13 (1964), doi: 10.1103/ PhysRevLett.13.508, arXiv:PhysRevLett.13.508.

[11] CDF Collaboration, "Observation of Top Quark Production in $p\bar{p}$ Collisions", Phys. Rev. Lett. 74 (1995) 2626, doi:10.1103/PhysRevLett.74.2626, arXiv:hep-ex/9503002.

[12] D0 Collaboration, "Observation of the Top Quark", Phys. Rev. Lett. 74 (1995) 2632, doi:10.1103/PhysRevLett.74.2632, arXiv:hep-ex/9503003.

[13] The four LEP collaborations, ALEPH, DELPHI, L3 and OPAL, "Search for the Standard Model Higgs Boson at LEP", Phys. Rev. Lett. B565 61, (2003), doi: 10.1016/S0370-2693(03)00614-2, arXiv:hep-ex/0306033.

[14] CDF and D0 Collaborations, "Higgs Boson Studies at the Tevatron", Phys. Rev. Lett. D88, 052014 (2013), doi: 10.1103/PhysRevD.88.052014, arXiv:1303.6346.

[15] ATLAS Collaboration, "Measurement of Higgs boson production in the diphoton decay channel in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector", Phys. Rev. D 90, 112015, (2014), doi: 10.1103/PhysRevD.90.112015, PhysRevD.90.112015.

[16] ATLAS Collaboration, "Measurements of Higgs boson properties in the diphoton decay channel with 36.1 fb^{-1} pp collision data at the center-of-mass energy of 13 TeV with the ATLAS detector", 2017 European Physical Society Conference on High Energy Physics, Venice, Italy, 05 - 12 Jul 2017, cds.cern.ch/record/2273852.

[17] CMS Collaboration, "Observation of the diphoton decay of the Higgs boson and measurement of its properties", Eur. Phys. J. C 74 (2014) 3076, doi: 10.1140/epjc/s10052-014-3076-z, arXiv:1407.0558.

[18] CMS Collaboration, "Measurements of properties of the Higgs boson in the diphoton decay channel with the full 2016 data set", CMS-PAS-HIG-16-040, cds.cern.ch/record/2264515.

[19] ATLAS Collaboration, "Measurements of Higgs boson production and couplings in the four-lepton channel in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector", Phys. Rev. D 91, 012006 (2015), PhysRevD.91.012006.

[20] CMS Collaboration, "Observation of Z decays to four leptons with the CMS detector at the LHC", JHEP 12 (2012) 034, doi:10.1007/JHEP12(2012)034, arXiv:1210.3844.

[21] ATLAS Collaboration, "Measurement of the Higgs boson coupling properties in the $H \to ZZ^* \to 4\ell$ at $\sqrt{s} = 13$ TeV with the ATLAS detector", ATLAS-CONF-2017-043, cds.cern.ch/record/2273849.

[22] 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, doi:10.1007/JHEP11(2017)047, arXiv:1706.09936.

[23] CMS Collaboration, "Measurement of the top quark Yukawa coupling from $\text{t}\bar{\text{t}}$ kinematic distributions in the lepton+jets final state in proton-proton collisions at $\sqrt{s} =$ 13 TeV", Submitted to Phys. Rev. D, arXiv:1907.01590

[24] F. Bezrukov, M. Shaposhnikov, "Why should we care about the top quark Yukawa coupling?", J. Exp. Theor. Phys. 120 (2015) 3, 335-343, arXiv:1411.1923.

[25] CMS Collaboration, "Measurement of the cross section for top quark pair production in association with a W or Z boson in proton-proton collisions at $\sqrt{s} = 13$ TeV", JHEP 08 (2018) 011, doi:10.1007/JHEP08(2018)011, arXiv:1711.02547.

[26] B. Grzadkowski, M. Iskrzynski, M. Misiak, and J. Rosiek, "Dimension-six terms in the Standard Model Lagrangian", JHEP 10 (2010) 085, doi:10.1007/JHEP10(2010)085, arXiv:1008.4884.

[27] F. Maltoni, E. Vryonidoua and C. Zhangb, "Higgs production in association with a top-antitop pair in the Standard Model Effective Field Theory at NLO in QCD", CP3-16-39, MCnet-16-29, arXiv:1607.05330.

[28] J. Alwall et al., "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations", JHEP 07 (2014) 079, doi:10.1007/JHEP07(2014)079, arXiv:1405.0301.

[29] CMS Collaboration, "Search for the associated production of the Higgs boson with a top-quark pair", JHEP 09 (2014) 087, doi:10.1007/JHEP09(2014)087, arXiv:1408.1682.

[30] 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, doi:10.1140/epjc/s10052-015-3454-1, arXiv:1502.02485.

[31] CMS Collaboration, "Search for Higgs boson production in association with top quarks in multilepton final states at $\sqrt{s}$ = 13 TeV", 2017, CMS-PAS-HIG-17-004, cds.cern.ch:2256103.

[32] 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, doi:10.1007/JHEP11(2017)047, arXiv:1706.09936.

[33] CMS Collaboration, "Search for the associated production of a Higgs boson with a top quark pair in final states with a τ lepton at $\sqrt{s}$ =13 TeV", 2017, CMS-PAS-HIG-17-003, cds.cern.ch:2257067.

[34] CMS Collaboration, "Observation of $t\bar{t}H$ production", Phys. Rev. Lett. 120, (2018), doi: 10.1103/PhysRevLett.120.23180, arXiv:1804.02610.

[35] ATLAS Collaboration, "New ATLAS result establishes production of Higgs boson in association with top quarks", Phys. Lett. B 784 (2018) 173, doi: 10.1016/j.physletb.2018.07.035, arXiv:1806.00425.

[36] The ATLAS Collaboration, The CMS Collaboration, The LHC Higgs Combination Group, "Procedure for the LHC Higgs boson search combination in Summer 2011", Technical Report CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11, 2011. cds:1379837.

[37] A. Djouadi, "The Anatomy of Electro-Weak symmetry Breaking. I: The Higgs boson in the Standard Model", Phys. Rept. 457 (2008) 1, doi:10.1016/j.physrep.2007.10.004, arXiv:hep-ph/0503172

[38] ATLAS Collaboration, "Search for the Standard Model Higgs boson decaying into $b\bar{b}$ produced in association with top quarks decaying hadronically in pp collisions at $\sqrt{s}$ = 8 TeV with the ATLAS detector", Phys. Rev. D 97, 072016 (2018), doi:10.1103/PhysRevD.97.072016, arXiv:1712.08895.

[39] 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, arXiv:1502.02485

[40] ATLAS Collaboration, "Search for the Standard Model Higgs boson decaying into $b\bar{b}$ produced in association with top quarks decaying hadronically in pp collisions at $\sqrt{s}$=8 TeV with the ATLAS detector", JHEP 05 (2016) 160, doi:10.1007/JHEP05(2016)160, arXiv:1604.03812

[41] CMS Collaboration, "Search for $t\bar{t}H$ production in the all-jet final state in proton-proton collisions at $\sqrt{s} = 13$ TeV", JHEP 06 (2018) 101, 10.1007/JHEP06(2018)101, arXiv:1803.06986.

[42] CMS Collaboration, "The CMS experiment at the CERN LHC", JINST 3 (2008) S08004, doi:10.1088/1748-0221/3/08/S08004, 2008 JINST 3 S08004.

[43] CMS Collaboration, "The CMS tracker system project: Technical Design Report", CERN-LHCC-98-006, CMS-TDR-5, cds.cern.ch:368412.

[44] CMS Collaboration, "The CMS electromagnetic calorimeter project: Technical Design Report", CERN-LHCC-97-033,CMS-TDR-4, cds.cern.ch:349375.

[45] CMS Collaboration, "The CMS hadron calorimeter project: Technical Design Report", CERN-LHCC-97-031, CMS-TDR-2, cds.cern.ch:357153.

[46] CMS Collaboration, "The CMS magnet project: Technical Design Report", CERN-LHCC-97-010, CMS-TDR-1, cds.cern.ch:331056.

[47] CMS Collaboration, "The CMS muon project: Technical Design Report", CERN-LHCC-97-032, CMS-TDR-3, cds.cern.ch:343814.

[48] CMS Collaboration, "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV", CERN-EP-2018-058, doi: 10.1088/1748-0221/13/06/P06015, arXiv:1804.04528.

[49] CMS Collaboration, "The CMS trigger system", JINST 12 (2017) P01020, doi:10.1088/1748-0221/12/01/P01020, arXiv:1609.02366.

[50] K. Bos, N. Brook, D. Duellmann et al., "LHC computing Grid: Technical Design Report", CERN-LHCC-2005-024, LCG-TDR-001, cds.cern.ch:840543.

[51] CMS Collaboration, "Particle-flow reconstruction and global event description with the CMS detector", JINST 12 (2017) P10003, doi:10.1088/1748-0221/12/10/P10003, arXiv:1706.04965.

[52] S. Catani, Y. L. Dokshitzer, M. H. Seymour, B. R. Webber, "Longitudinally-invariant k_T-clustering algorithms for hadron-hadron collisions", Nucl. Phys. B406 (1993) 187–224, doi:10.1016/0550-3213(93)90166-M, cds.cern.ch:246812.

[53] S. D. Ellis and D. E. Soper, "Successive Combination Jet Algorithm For Hadron Collisions", Phys. Rev. D48 (1993) 3160–3166, doi:10.1103/PhysRevD.48.3160, arXiv:9305266.

[54] Yu. L. Dokshitzer, G.D. Leder, S. Moretti, B.R. Webber, "Better jet clustering algorithms", JHEP 9708 (1997) 001, doi:10.1088/1126-6708/1997/08/001, arXiv:9707323.

[55] M. Wobisch and T. Wengler, "Hadronization Corrections to Jet Cross Sections in Deep-Inelastic Scattering", PITHA 99/16, arXiv:9907280.

[56] M. Cacciari, G. P. Salam, G. Soyez, "The anti-k_T jet clustering algorithm", JHEP 04 (2008) 063, doi:10.1088/1126-6708/2008/04/063, arXiv:0802.1189.

[57] M. Cacciari, G. P. Salam, G. Soyez, "FastJet user manual", Eur. Phys. J. C 72 (2012) 1896, doi:10.1140/epjc/s10052-012-1896-2, arXiv:1111.6097.

[58] CMS Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV", JINST 13 (2018) P05011, doi:10.1088/1748-0221/13/05/P05011, arXiv:1712.07158.

[59] CMS Collaboration, "Identification of b-quark jets with the CMS experiment", JINST 8 (2013) P04013, doi:10.1088/1748-0221/8/04/P04013, ttps://arxiv.org/abs/1211.4462arXiv:1211.4462.

[60] S. Marzani, G. Soyez, M. Spannowsky, "Looking inside jets: an introduction to jet substructure and boosted-object phenomenology", Lecture Notes in Physics, volume 958 (2019), doi:10.1007/978-3-030-15709-8, arXiv:1901.10342.

[61] J. M. Butterworth, A. R. Davison, "Jet substructure as a new Higgs search channel at the LHC", Phys.Rev.Lett.100:242001,2008, doi:10.1103/PhysRevLett.100.242001, arXiv:0802.2470.

[62] A. J. Larkoski, S. Marzani, G. Soyez, J. Thaler, "Soft Drop", JHEP 1405 (2014) 146, doi:10.1007/JHEP05(2014)146, arXiv:1402.2657.

[63] D. Krohn, J. Thaler, L.-T. Wang, "Jet Trimming", JHEP 1002:084,2010, doi:10.1007/JHEP02(2010)084, arXiv:0912.1342.

[64] S. D. Ellis, C. K. Vermilion, J. R. Walsh, "Techniques for improved heavy particle searches with jet substructure", Phys. Rev. D80 (2009) 051501, doi:10.1103/PhysRevD.80.051501, arXiv:0903.5081

[65] S. D. Ellis, C. K. Vermilion, J. R. Walsh, "Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches", Phys. Rev. D81 (2010) 094023, doi:10.1103/PhysRevD.81.094023, arXiv:0912.0033

[66] A. J. Larkoski, G. P. Salam and J. Thaler, "Energy Correlation Functions for Jet Substructure", JHEP 1306 (2013) 108, doi:10.1007/JHEP06(2013)108, arXiv:1305.0007.

[67] I. Moult, L. Necib and J. Thaler, "New Angles on Energy Correlation Functions", JHEP 12 (2016) 153, doi:10.1007/JHEP12(2016)153, arXiv:1609.07483.

[68] A. Hocker, P. Speckmayer, J. Stelzer, "TMVA - Toolkit for Multivariate Data Analysis with ROOT TMVA", CERN-OPEN-2007-007, cds.cern.ch:1019880.

[69] T. Plehn, G. P. Salam, M. Spannowsky, "Fat Jets for a Light Higgs", Phys. Rev. Lett. 104 (2010) 111801, doi:10.1103/PhysRevLett.104.111801, arXiv:0910.5472.

[70] G. Kasieczka, T. Plehn, T. Schell, T. Strebler, G. P. Salam, "Resonance Searches with an Updated Top Tagger", JHEP 06 (2015) 203, doi:10.1007/JHEP06(2015)203, arXiv:1503.05921.

[71] S. D. Ellis, A. Hornig, T. S. Roy, D. Krohn and M. D. Schwartz, Qjets: "A Non-Deterministic Approach to Tree-Based Jet Substructure", Phys. Rev. Lett. 108 (2012) 182003, doi:10.1103/PhysRevLett.108.182003, arXiv:1201.1914.

[72] S. D. Ellis, A. Hornig, D. Krohn, T. S. Roy, On Statistical Aspects of Qjets, JHEP 01 (2015) 022, doi:10.1007/JHEP01(2015)022, arXiv:1409.6785.

[73] S. Frixione, P. Nason, C. Oleari, "Matching NLO QCD computations with parton shower simulations: the POWHEG method", JHEP 11 (2007) 070, doi:10.1088/1126-6708/2007/11/070, arXiv:0709.2092.

[74] S. Alioli et al., "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX", JHEP 06 (2010) 043, doi:10.1007/JHEP06(2010)043, arXiv:1002.2581.

[75] T. Sjöstrand et al., "An introduction to PYTHIA 8.2", Comput. Phys. Commun. 191 (2015) 159, doi:10.1016/j.cpc.2015.01.024, arXiv:1410.3012.

[76] CMS collaboration, "Event generator tunes obtained from underlying event and multiparton scattering measurements", Eur. Phys. J. C (2016) 76:155, doi: 10.1140/epjc/s10052-016-3988-x, arXiv:1512.00815.

[77] P. Skands, S. Carrazza, J. Rojo, "Tuning PYTHIA 8.1: the Monash 2013 tune", Eur. Phys. J. C 74 (2014) 3024, doi:10.1140/epjc/s10052-014-3024-y, arXiv:1404.5630.

[78] CMS collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV", CERN-EP-2017-326, doi: 10.1088/1748-0221/13/05/P05011, arXiv:1712.07158.

[79] CMS collaboration, "Search for $t\bar{t}H$ production in the all-jet final state in proton-proton collisions at $\sqrt{s} = 13$ TeV", JHEP06(2018)101, arXiv:1803.06986.

[80] G. Cowan, K. Cranmer, E. Gross, O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics", Eur. Phys. J. C (2011) C71:1554, doi:10.1140/epjc/s10052-011-1554-0, arXiv:1007.1727.

[81] A. Wald, "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large", Transactions of the American Mathematical Society, Vol. 54, No. 3, 1943, https://doi.org/10.1090/S0002-9947-1943-0012401-3, Trans. Amer.Math. Soc. 943-054-03.

[82] S.S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses", Ann. Math. Statist. 9 (1938) 60-2, doi:10.1214/aoms/1177732360, inspirehep:1247197.