UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DE PHYSIQUE ET CHIMIE PHYSIQUE

Institut Pluridisciplinaire Hubert Curien

THÈSE

présentée par :

Xavier COUBEZ

soutenue le : 15 septembre 2017

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline/ Spécialité : Physique des particules

Recherche du boson de Higgs standard produit en association avec une paire de quarks top dans le canal multi-leptons dans l'expérience CMS

THÈSE dirigée par :

M. Daniel Bloch	Institut Pluridisciplinaire Hubert Curien
Mme Anne-Catherine Le Bihan (co-encadrante)	Institut Pluridisciplinaire Hubert Curien

RAPPORTEURS :

Mme Lucia Di Ciaccio	Laboratoire d'Annecy le Vieux de Physique des Particules
M. Laurent Vacavant	Centre de Physique des Particules de Marseille

AUTRES MEMBRES DU JURY :

M. Pascal Paganini	Laboratoire Leprince-Ringuet
Mme Isabelle Ripp-Baudot	Institut Pluridisciplinaire Hubert Curien

Université de Strasbourg

Thèse

pour l'obtention du diplôme de

Docteur de l'Université de Strasbourg

spécialité : physique des particules

présentée par

Xavier Coubez

Recherche du boson de Higgs standard produit en association avec une paire de quarks top dans le canal multi-leptons dans l'expérience CMS

présentée le 15 septembre 2017 devant la commission d'examen composée de

M. Daniel Bloch,	directeur de thèse
Mme Anne-Catherine Le Bihan,	co-encadrante
M. Laurent Vacavant,	rapporteur
Mme Lucia Di Ciaccio,	rapporteur
M. Pascal Paganini,	examinateur
Mme Isabelle Ripp-Baudot,	examinatrice

«Le savant n'étudie pas la nature parce que cela est utile; il l'étudie parce qu'il y prend plaisir et il y prend plaisir parce qu'elle est belle. Si la nature n'était pas belle, elle ne vaudrait pas la peine d'être connue, la vie ne vaudrait pas la peine d'être vécue. Je ne parle pas ici, bien entendu, de cette beauté qui frappe les sens, de la beauté des qualités et des apparences; non que j'en fasse fi, loin de là, mais elle n'a rien à faire avec la science; je veux parler de cette beauté plus intime qui vient de l'ordre harmonieux des parties, et qu'une intelligence pure peut saisir.»

Henri Poincaré, Science et méthode (1908)

« The most incomprehensible thing about the Universe is that it is comprehensible. »

Albert Einstein

Acknowledgements

Contents

Introduction

«It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way...»

Charles Dickens, *A Tale of Two Cities*

For centuries, the nature of the Universe and of its building blocks was a matter of philosophical debates. However in the last century, the birth of cosmology and particle physics led to a new understanding of the world we live in. Cosmology extended our understanding of the Universe based on general relativity while particle physics produced in the sixties a model aiming at a description of fundamental particles and their interactions, based on the quantum field theory. The Standard Model was then tested with a remarquable accuracy at accelerator facilities. All the particles predicted were found but one was still eluding the physicists in 2009 at the start of the Large Hadron Collider, the biggest accelerator operating today. In only three years, the Higgs boson was discovered and the announcement in July 2012 of its observation is the last step toward the completness of the model.

Since this discovery, particle physics is facing an unprecedented situation. The model which rose during decades is now complete. However, theoretical considerations as well as experimental observations indicate that the Standard Model is only an effective theory from which a more general theory could be constructed. Testing the validity of this theory is one of the goal of the LHC but it is possible that no new physics will be accessible at colliders, the scale of new physics being at energies beyond our reach.

Before moving to the search for new physics, it is therefore important to get a better understanding of the physics of the Higgs boson. During the second period (Run 2) of data taking which is ongoing since 2015 and will last until 2018, one of the key studies allowing to test the validity of the Standard Model in the Higgs sector is the measurement of the coupling of the Higgs boson to the most massive particle, the top quark. Because of the role of the Higgs boson in the generation of the mass, this coupling is expected to be important. The small cross-section of the associated production of the Higgs boson with a pair of top quarks made it challenging to study this process during Run 1. The analysis benefits from the rise in collision energy of the LHC between Run 1 (7 - 8 TeV) and Run 2 (13 TeV) which leads to an increase by a factor four of the production cross section, allowing to probe with increasing accuracy the top - Higgs coupling.

This document presents the search for the associated production of the Higgs boson with a top-antitop quark pair with the data taken in 2016 in the multilepton channel and the first evidence for such a production.

The theoretical context will first be introduced by reviewing the main constituents of the Standard Model. Physics in the top and Higgs sectors will be motivated and described. After a brief review of the shortcomings of the Standard Model, searches for new physics in the top and Higgs sectors will be mentioned.

The experimental apparatus will be described as well as the improvements introduced to cope with increasingly complex conditions of data taking. A short description of the accelerator and its operation will be followed by the description of the Compact Muon Solenoid (CMS) detector.

The reconstruction of particles and their identification inside the CMS experiment will then be presented. Because of its architecture, CMS is ideally suited to deploy an identification based on Particle-Flow (PF) which combines information from all the sub-detectors to achieve an optimal reconstruction of individual particles.

The focus will then move to the reconstruction and identification of jets produced in the hadronization of bottom quarks. The b quarks are of particular interest in many physics analyses at the LHC and play an important role in top quark physics. The identification of jets originating from b quarks has benefited in the last years from algorithms with increasing complexity and from detector upgrades which will both be described. The use of such algorithms at trigger level makes it possible to filter events during the data taking, keeping only events relevant to the analyses among the millions of events produced every second. Their use at trigger level and the constraints related to their deployment will be discussed.

Finally, the study of the top-Higgs coupling in CMS will be presented. After going through the various channels in which this coupling is being studied, the search for the Higgs boson production in association with top quarks in the multilepton final states will be described. The study, targeting the decay of the Higgs to WW* or ZZ* and the leptonic decay of either one or the two top quarks, was performed with the 35.9 fb^{-1} of data taken in 2016 and led to the first evidence of the coupling of the Higgs boson to the top quark.

Chapter 1

The Standard Model

Contents

This chapter intends to introduce the Standard Model, the theory describing the behavior and interactions between elementary particles. Section 1.1 will describe the structure of the Standard Model. Section 1.2 will briefly present the limits of the model and introduce the various ways this model might be extended, focusing on some possible tests of new physics in the top quark and Higgs sectors.

1.1 The Standard Model

In the Standard Model the fundamental entities are quantum fields. An excitation of the field corresponds to the associated observable elementary particle which can either be a component of matter or a mediator of the interaction. Two kinds of particles can be distinguished, the fermions, particles of spin 1/2, satisfying the Pauli exclusion principle and the bosons, particles of spin 0 or 1 which satisfy Bose-Einstein statistics.

1.1.1 Fields and symmetries

The Standard Model is a quantum field theory and relies on two key elements: fields and symmetries.

Fields are associated with particles and the evolution of such particles is described by the lagrangian. For a classical particle, the equation of motion can be derived from the lagrangian which is function of the generalized coordinates. In quantum field theory, a particle is described by a field associated to the probability to find this particle at a particular point of space-time. The evolution of the field is given by integration on space-time of the lagrangian density (action):

$$\mathcal{S} \equiv \int_V \mathcal{L}(\phi, \partial_\mu \phi) d^4x \tag{1.1}$$

The lagrangian density is composed of a kinetic part T and a potential part V:

$$\mathcal{L}(\phi, \partial_\mu \phi) = T(\phi, \partial_\mu \phi) - V(\phi, \partial_\mu \phi) \tag{1.2}$$

For particles, the potential part is associated to the mass and to the interactions.

The Noether theorem involves a correspondance between any symmetry leaving the lagrangian invariant and a conserved quantity. Invariance of the lagrangian by translation is associated with momentum conservation, invariance of the lagrangian by time evolution is associated with energy conservation. But beyond space time symmetry, a set of internal

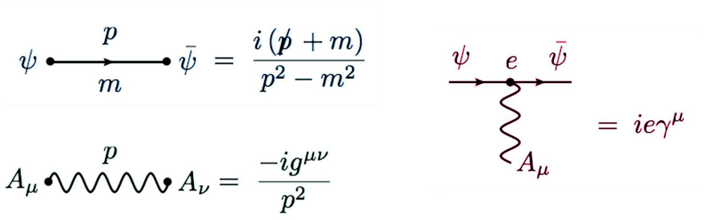


Figure 1.1: Feynman diagrams associated with the QED lagrangian.

symmetries is needed.

1.1.2 QED and QCD

The Standard Model is a gauge theory based on the symmetry $SU(3)_C \times SU(2)_L \times U(1)_Y$ describing strong and electroweak interactions.

Quantum electrodynamics (QED)

The first key component to have been fully described in the context of quantum field theory is the electromagnetic interaction. The theory called quantum electrodynamics is based on a Lagrangian which can be decomposed in three parts as follows:

$$\begin{aligned} \mathcal{L}_{QED} &= \mathcal{L}_{Dirac} + \mathcal{L}_{Maxwell} + \mathcal{L}_{int} && (1.3) \\ &= \bar{\psi}(i\not{\partial} - m)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} - e\bar{\psi}\gamma^\mu\psi A_\mu && (1.4) \\ &= \bar{\psi}(i\not{D} - m)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} && (1.5) \end{aligned}$$

where A_μ is the electromagnetic vector potential, $F_{\mu\nu} = \delta_\mu A_\nu - \delta_\nu A_\mu$ is the field tensor strength, $\not{D}_\mu = \delta_\mu + ieA_{mu}$ is the covariant derivative introduced to preserve the local gauge invariance. The electromagnetic coupling constant is related to the electric charge as $\alpha = e^2/4\pi$. The U(1) group of QED is abelian as the photon carries no charge.

From this lagrangian, we can derive the possible Feynman rules (Figure 1.1). Because of the structure of the lagrangian, electrons and photons can propagate, and an electron can either radiate or absorb a photon. According to the Noether theorem, the gauge symmetry of the action implies that the electric charge is conserved.

Quantum chromodynamics (QCD), the theory of the strong interaction

Quantum chromodynamics relies on the quarks introduced to explain the pattern of the meson and baryon states and the conservation of the new quantum number, colour, which was needed to satisfy the Fermi-Dirac statistics. Mesons and baryons are described as colour-singlet combinations.

The strong interaction holding together quarks within hadrons is based on the $SU(3)$ group. The non-abelian nature of the group implies that the gluons are carrying colours and anti-colours and interact with each other.

Quarks can be combined in two ways into colour singlets of the SU(3) group. Mesons are composed of an even number of quarks and antiquarks of opposite colour. Baryons are composed of three quarks of different colours.

The QCD lagrangian can be written as:

$$\mathcal{L}_{\text{QCD}} = \sum \bar{\psi}(i\not{D} - m_f\delta_{ij})\psi - \frac{1}{4}F^a_{\mu\nu}F^{\mu\nu}_a, \tag{1.6}$$

where the covariant derivative is now:

$$D^\mu_{ij} = \partial^\mu\delta_{ij} + ig_s t^a_{ij}A^\mu_a \tag{1.7}$$

and the field strength is:

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g_s f_{abc}A^b_\mu A^c_\nu. \tag{1.8}$$

The strength of the strong interaction is given by the parameter g_s. Its presence in the field strength describes the self-interaction between gluons (Figure 1.2). Similarly to QED, the strong coupling constant is the only free parameter of QCD and is defined as $\alpha_s = g_s^2/4\pi$.

Generators of SU(3) are colour matrices corresponding to gluons. From the N^2 complex values of the matrix (with N=3), to which correspond $2N^2$ real values, the requirement of unitarity and the unitarity of the determinant leads to N^2-1 hence 8 generators, the 8 gluons.

The colour matrices t^a can be expressed as a function of the Gell-Mann matrices λ^a by $t^a = \frac{1}{2}\lambda^a$. The matrices do not commute and the associated group is thus non-abelian.

Strong interaction exhibits two specific features (Figure 1.3). At low energy (large distance), the strong coupling constant increases, which involves the confinement of quarks and gluons within hadrons and the evolution of their density function (PDF) with respect to the available energy. At high energy (short distance), the value of the strong coupling constant decreases, leading to a regime called asymptotic freedom in which quarks close to each other are free. In this regime, perturbative QCD can be applied to provide precise predictions.

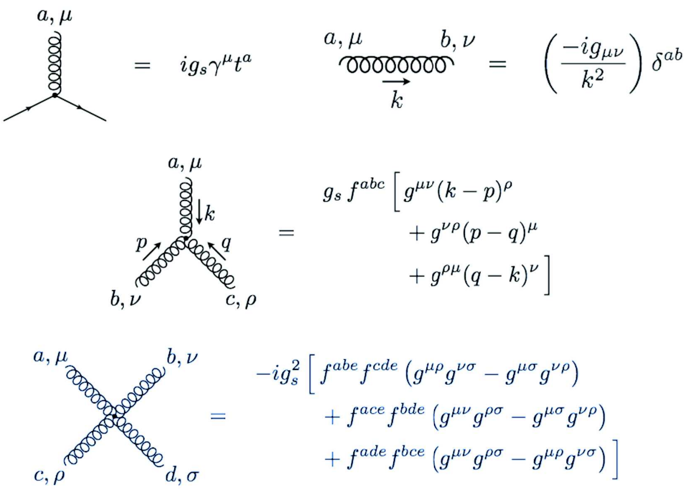

a, μ
$= \quad i g_s \gamma^\mu t^a$
a, μ
b, ν
k
$= \quad \left(\frac{-i g_{\mu\nu}}{k^2}\right) \delta^{ab}$
a, μ
k
p
q
b, ν
c, ρ
$= \quad g_s f^{abc} \Big[g^{\mu\nu}(k-p)^\rho + g^{\nu\rho}(p-q)^\mu + g^{\rho\mu}(q-k)^\nu \Big]$
a, μ
b, ν
c, ρ
d, σ
$= \quad -i g_s^2 \Big[f^{abe} f^{cde} \left(g^{\mu\rho} g^{\nu\sigma} - g^{\mu\sigma} g^{\nu\rho}\right) + f^{ace} f^{bde} \left(g^{\mu\nu} g^{\rho\sigma} - g^{\mu\sigma} g^{\nu\rho}\right) + f^{ade} f^{bce} \left(g^{\mu\nu} g^{\rho\sigma} - g^{\mu\rho} g^{\nu\sigma}\right) \Big]$

Figure 1.2: Feynman diagrams associated with the QCD lagrangian.

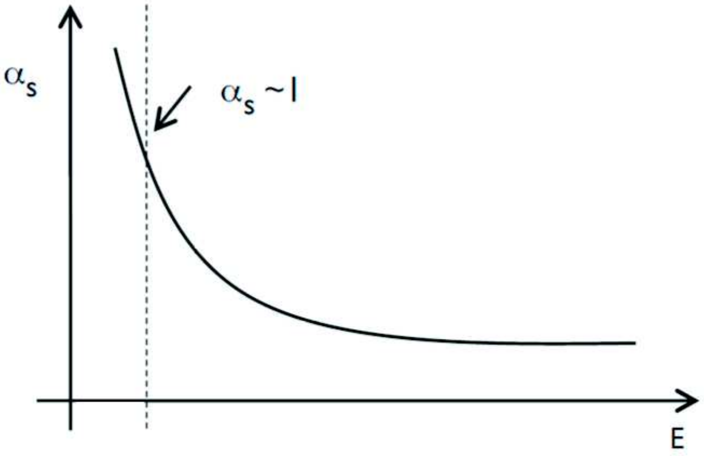


Figure 1.3: Evolution of the strong coupling constant.

1.1.3 Electroweak interaction

The weak interaction is the interaction describing the β decay, first introduced as a contact interaction in 1933 by Fermi. It becomes a short range interaction through the exchange of a vector boson.

The weak interaction is a SU(2) model giving rise to three gauge bosons W^+, W^- and Z. The gauge bosons are mediating two kinds of interactions. The charged current contains only left-handed fermionic fields and is non diagonal in the quark flavour space. The neutral current contains both left and right handed fermionic fields and is diagonal in the quark flavour space.

The gauge invariance implies that both the gauge bosons and the fermions are massless.

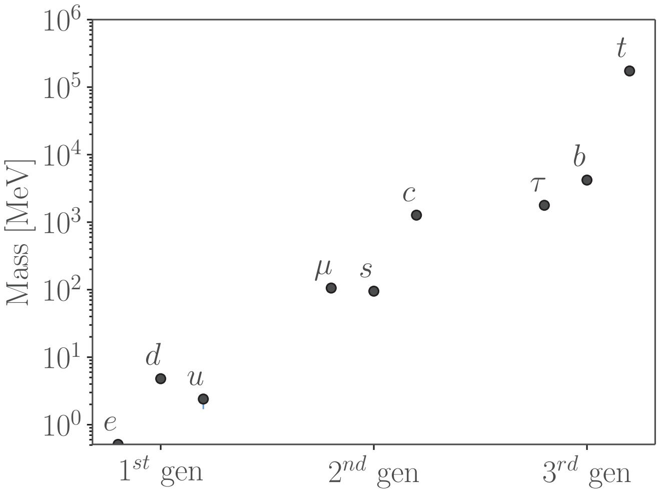


Figure 1.4: Fermion mass spectrum.

However experimental observations show that the weak interactions are short-ranged, making it necessary for the bosons to have a mass. Moreover, W and Z bosons were then observed and their masses measured.

The fermion mass spectrum of the three families is shown in Figure 1.4 while Table 1.1 summarises charges associated to the interactions of the particle generations.

1st gen.	2nd gen.	3rd gen.	Electric charge	Weak charge	Number of colour charges
u	s	t	2/3	+1/2	3
d	c	b	-1/3	-1/2	3
$\nu_{e,L}$	$\nu_{\mu,L}$	$\nu_{\tau,L}$	0	-1/2	-
e	μ	τ	-1	1/2	-

Table 1.1: Elementary particle generation and associated charges.

Electroweak unification

The electroweak unification is the first success in merging two of the fundamental interactions by providing a unified description of the electromagnetism and of the weak interaction. The unification of charged current, neutral current and electromagnetic interactions is done starting from four massless bosons W^1, W^2, W^0 and B^0 from which emerge the physical bosons W^+, W^-, Z^0 and γ through the mixing of the neutral bosons.

$$\begin{pmatrix} Z^0 \\ \gamma \end{pmatrix} = \begin{pmatrix} \cos\theta_W & -\sin\theta_W \\ \sin\theta_W & \cos\theta_W \end{pmatrix} \begin{pmatrix} W^0 \\ B^0 \end{pmatrix} \tag{1.9}$$

The weak mixing angle is also called Weinberg angle θ_W. The electromagnetic and weak couplings are related as $e = g\sin\theta_W$.

The remaining puzzle of the electroweak model is the mass of the W and Z bosons which acquire their masses through the Brout-Englert-Higgs mechanism.

Spontaneous symmetry breaking and Higgs mechanism

In order to give a mass to the W and Z bosons while keeping a massless photon, a symmetry breaking must be introduced. The mechanism proposed by Brout, Englert and Higgs [3, 4] allows to do so by introducing a doublet of complex scalar fields of the form:

$$\phi = \begin{pmatrix} \phi_+ \\ \phi_0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}. \tag{1.10}$$

By choosing a non-zero value for the Higgs field expectation value in the vacuum, the fields and interactions remain symmetric under SU(2) × U(1) while the vacuum is not. The degrees of freedom can be used to generate the masses of the W and Z bosons. The residual degree of freedom corresponds to a new spin 0 particle, the Higgs boson. The vacuum expectation value of the Higgs field is related to the W boson mass and weak coupling as: $v = 2\ M_W\ /\ \mathrm{g} \simeq 246$ GeV.

Yukawa couplings to fermions

While the masses of the W and Z vector bosons arise directly from their interaction with the Higgs boson, the masses of the fermions originate in a Yukawa coupling with the Higgs boson. Introduced by Yukawa to describe the nuclear force mediated by pions, the Yukawa coupling describes the interaction between a scalar field and a Dirac field. In the case of the Higgs boson, its coupling to quarks and leptons involves a mass term in the lagrangian of the form:

$$\mathcal{L}_{Yukawa}(\Psi_f) = -\frac{m_f}{v}\bar{\Psi}_f\Psi_f. \tag{1.11}$$

The couplings of the Higgs boson have been measured [9] and found in agreement in the limit of the achieved precision with the expected quadratic evolution with respect to the mass of the gauge bosons and linear with respect to the fermion masses (Figure 1.5).

1.1.4 Internal coherence and tests of the Standard Model

Since the discovery of the Higgs boson, the 19 free parameters of the Standard Model (three coupling constants, m_H, m_f (9), m_Z or m_W, and the four parameters of the CKM matrix) have been measured, allowing for global tests of the coherence of the Standard Model (Figure 1.6). All the observables are compatible with their prediction within 2.5σ.

Moreover, the overall production cross sections measured in CMS at the LHC for the processes predicted by the Standard Model are in agreement with the predictions

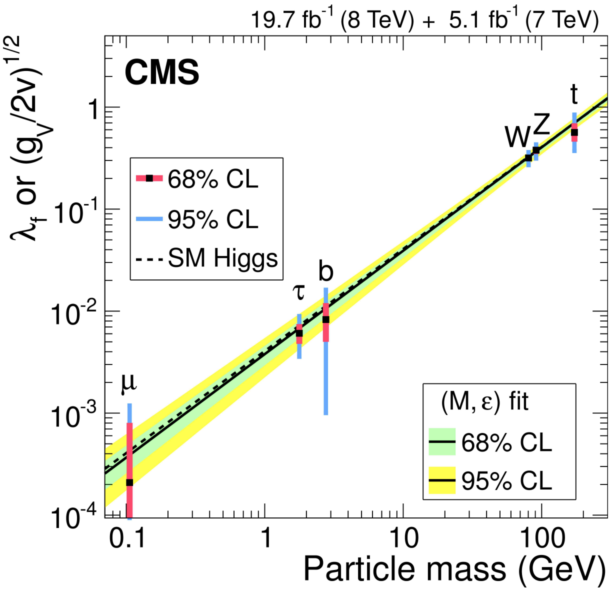


Figure 1.5: CMS measurements of the couplings of the Higgs boson to fermions and bosons as a function of the particle mass ($\lambda_f = \frac{m_f}{v}$, $\sqrt{\frac{g_v}{2v}} = \frac{m_{W/Z}}{v}$) [9].

(Figure 1.7).

1.1.5 Top quark physics

The existence of the top quark was necessary to make the electroweak theory consistent after the discovery of the tau lepton in 1975 and of the b quark in 1977. Discovered at the Tevatron in 1995, the top quark is the heaviest particle in the Standard Model with a mass of 173,34 ± 0,76 GeV [1]. Because of its short lifetime, the top quark is the only quark which doesn't undergo hadronization before decay, allowing for the study of a free quark. Moreover, the V$_{tb}$ element of the matrix element being close to one [1], the top quark decays quasi exclusively to a b quark and a W boson. Finally, its high mass makes it as a window to electroweak symmetry breaking, and maybe to new physics, with a Yukawa coupling to the Higgs boson which is close to one.

Production

At hadron colliders (such as the Tevatron, colliding protons and antiprotons and the Large Hadron Collider colliding protons and protons), the production of top quarks is possible either in pair via the strong interaction ($\sigma_{t\bar{t}}$ = 832 pb at 13 TeV) or singly via the electroweak interaction ($\sigma_{singletop}$ = 217 pb at 13 TeV). At LHC, two main production

[1] $\frac{BR(t \to Wb)}{BR(t \to Wq)} \approx 0.9982$ [31]

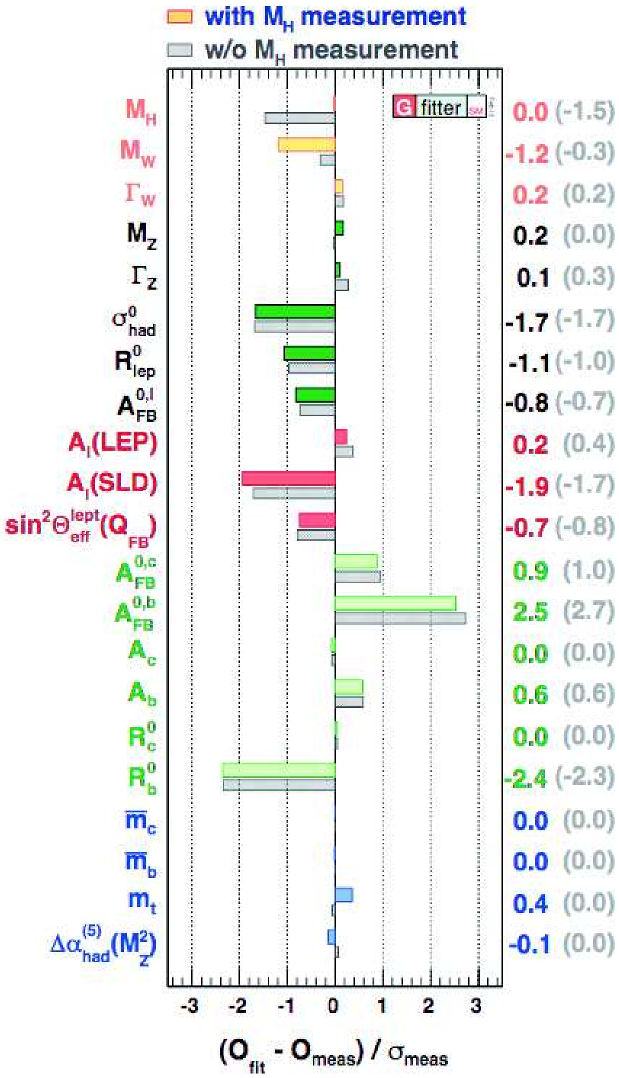


Figure 1.6: Global fit of the Standard Model observables after the discovery of the Higgs boson [2].

processes are involved in the top pair production which is the main contribution to top production. The top pairs can be produced by gluon fusion and by quark anti-quark annihilation. At 13 TeV, the top pair production is dominated by gluon fusion (80%). In 2016, the top pair production rate is about 10 Hz, allowing to make precise measurements in the top quark sector and to perform the study of rare decays as well as differential cross sections.

Decay

When considering the $t\bar{t}$ final state, three channels can be defined based on the decay products of the W bosons. Each leptonic decay (e, μ, τ) contributes to the total W decay width as about 10%, while hadronic decays of the W are shared between $u\bar{d}$ and $c\bar{s}$ quark

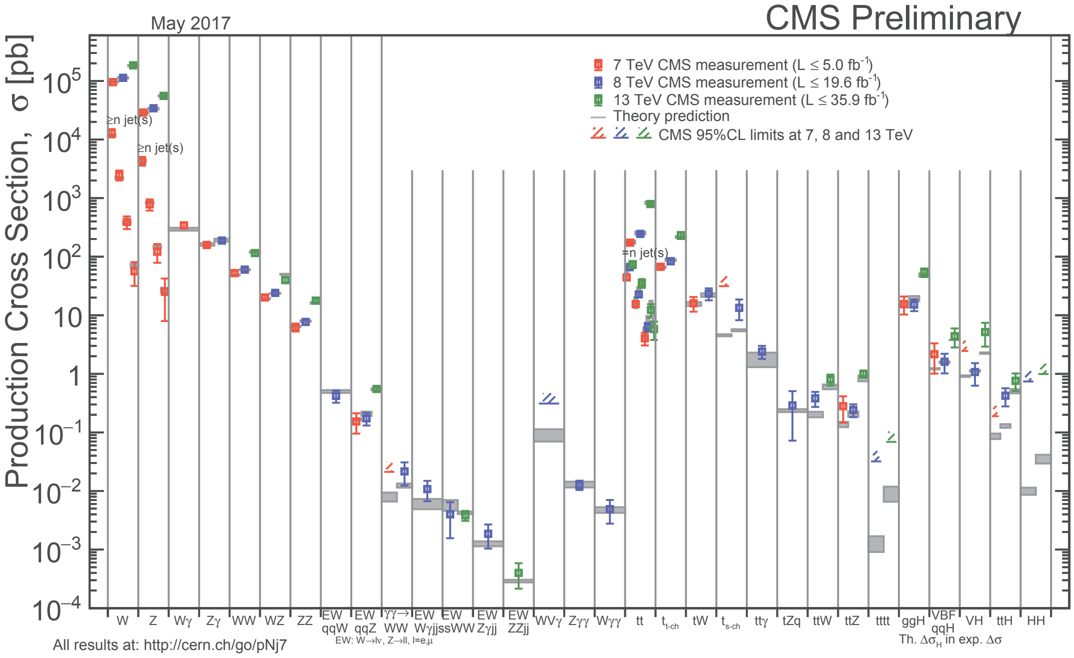


Figure 1.7: Production cross sections of Standard Model processes measured in CMS and compared with theory.

pairs. In the dileptonic channel, both W decay leptonically, leading to a final state with two leptons (e, μ, τ), two b jets, and some large missing transverse energy due to the presence of two neutrinos. While being a clean signature, this decay channel suffers from a relatively low branching ratio, leading to a contribution to the overall decays of the $t\bar{t}$ pair of the order of 9%, dropping to 4% when considering only the electrons and muons. In the $t\bar{t}$ semileptonic channel, only one of the two Ws decays leptonically, leading to a final state with one lepton and missing transverse energy from the neutrino, two jets from the hadronic decay of the second W and two b jets. Each leptonic decay (e, μ, τ) contributes to the overall $t\bar{t}$ yields as 15%. In the hadronic channel, both Ws decay hadronically leading to a final state with four light (u,d,s,c) jets and two b jets. While being the main decay, it features a challenging signature at hadron colliders where a huge hadronic activity is expected in the events.

Mass and production cross sections

Among the many top quark properties which can be studied at the LHC, three are of particular interest, either for the techniques which are used to measure them or for the consequences they have on the analysis which will be described in the next sections.

Mass

The measurement of the top quark mass has been performed in many ways, from the

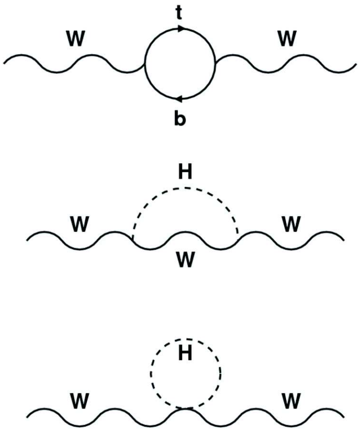


Figure 1.8: Loop processes leading to constraints on the Higgs mass.

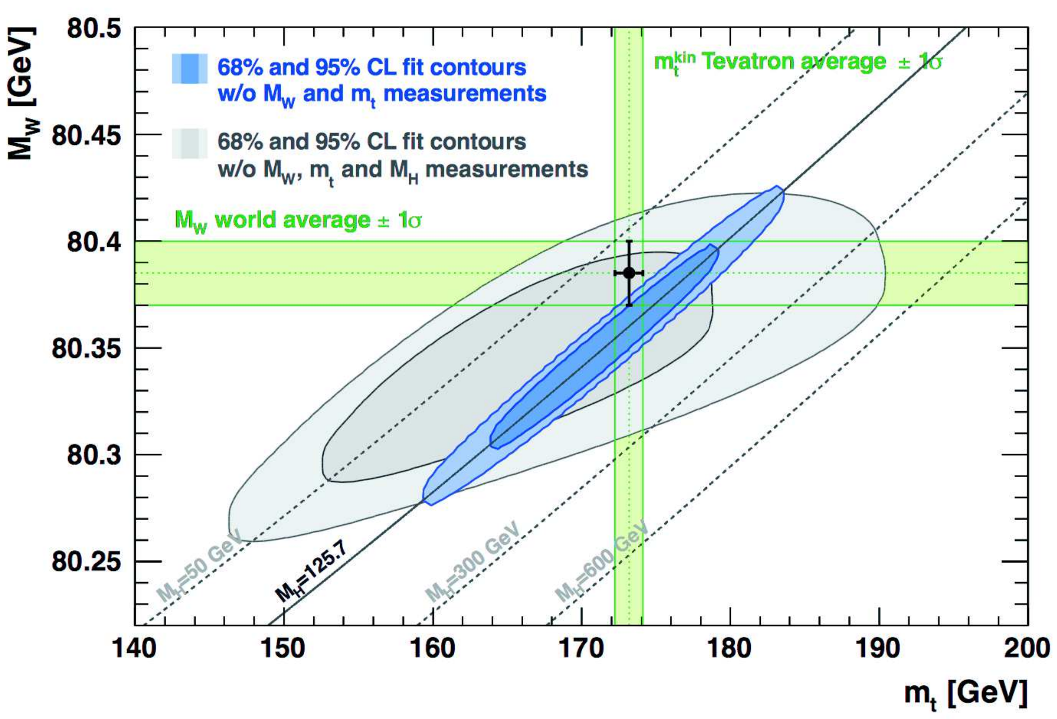


Figure 1.9: Interplay between the top, W and Higgs boson.

standard measure of the invariant mass of three jets in lepton + jets events, to the use of more sophisticated methods such as the Matrix Element Method.

One of the interest of measuring the top mass (together with the W mass) was the possibility to infer the mass of the Higgs boson from electroweak measurements (Figure 1.6 and Figure 1.9), the W propagator being sensitive to corrections related to the top and Higgs loops (Figure 1.8).

Now that the Higgs boson has been discovered, the top mass together with the Higgs

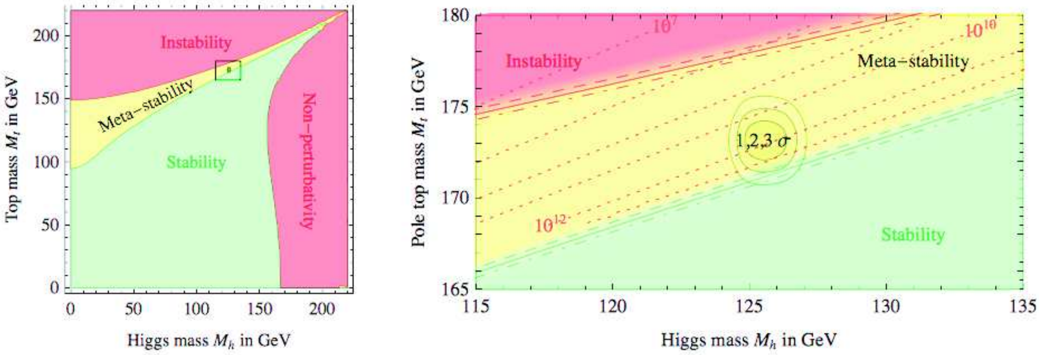


Figure 1.10: Regions of absolute stability, meta-stability and instability of the SM vacuum [21].

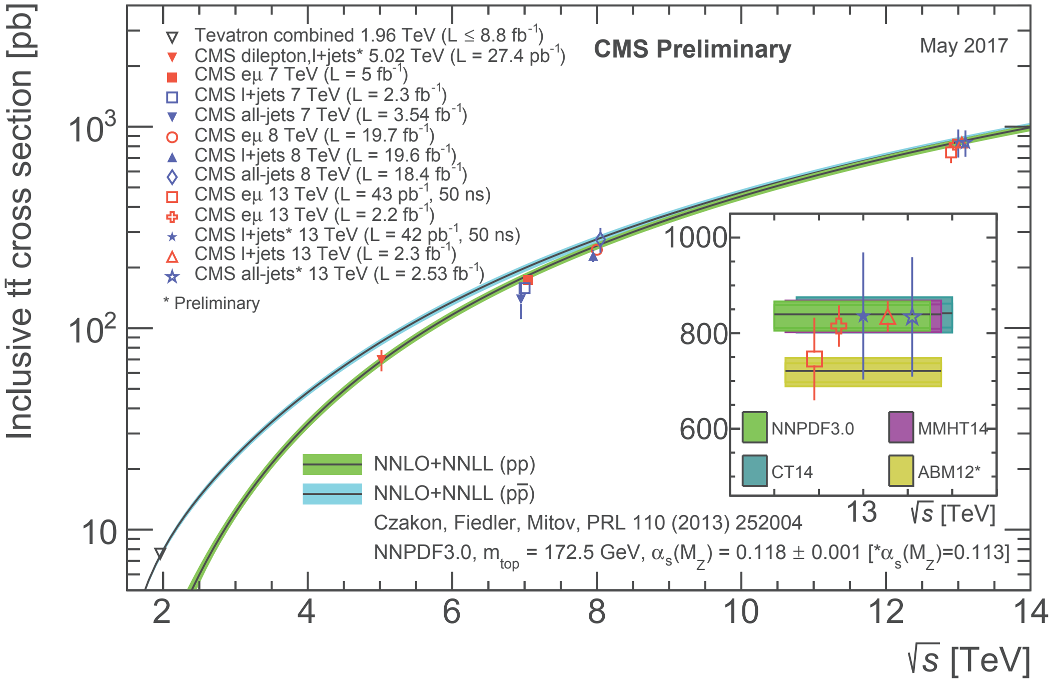


Figure 1.11: Measured t$\bar{t}$ cross-sections at different collision energies.

mass allow to evaluate the stability of the vacuum in our universe (Figure 1.10). Based on the latest measurements of the top and Higgs masses, the SM vacuum is meta-stable, at about 2-3 σ of the stability frontier.

Production cross section

The production cross sections of the top-quark pair (Figure 1.11) and single-top processes have been measured at several energies in $p\bar{p}$ and pp collisions, showing a good agreement with the prediction from the Standard Model.

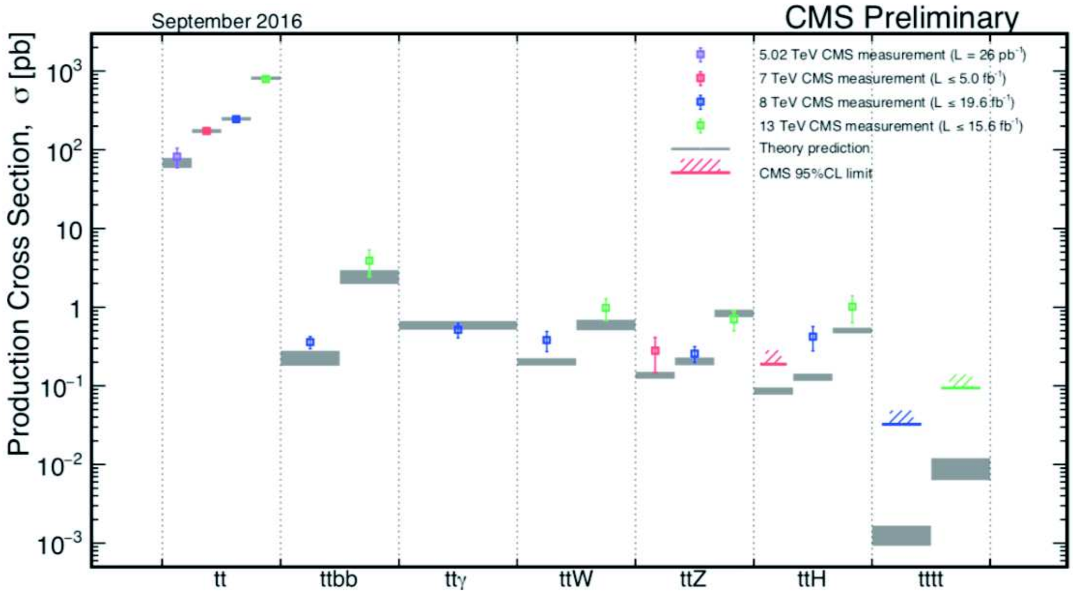


Figure 1.12: Production cross section of the different $\mathrm{t\bar{t}X}$ processes.

Other production cross sections which are relevant are the associated production with vector bosons or additionnal quarks (light or b). The theoretical value of such processes is now computed at higher orders in perturbation (NLO and NNLO). The precise measurements of these associated productions is important for a comparison with increasingly precise predictions (Figure 1.12).

The associated productions are of particular importance both to confirm the coupling of the top quark to bosons and to constrain the cross section of these processes which have a large contributions as background to the Higgs physics, mainly in the $\mathrm{t\bar{t}H}$ searches. From a theoretical point of view, the precision on the production cross sections of processes such as $t\bar{t}b\bar{b}$ is one of the main systematics when searching for $\mathrm{t\bar{t}H, H \rightarrow b\bar{b}}$. In the same way, the theoretical limits on the knowledge of the production cross section of both $\mathrm{t\bar{t}Z}$ and $\mathrm{t\bar{t}W}$ is among the main theoretical systematics in the search for $\mathrm{t\bar{t}H, H \rightarrow WW^*, ZZ^*}, \tau\tau$.

$\mathrm{t\bar{t}H}$ **production**

While more details will be provided about the motivation to study the associated production of the Higgs boson with a pair of top quarks, an overview of the cross sections and associated uncertainties [10] is provided in Table 1.2. The main backgrounds in the study of the multileptonic channel of $\mathrm{t\bar{t}H}$ are the $\mathrm{t\bar{t}}$, $\mathrm{t\bar{t}W}$ and $\mathrm{t\bar{t}Z}$ processes.

The scales uncertainty is related to the value chosen for the renormalization and factorization scales. The PDF uncertainty comes from the limited knowledge of the parton distribution functions derived from global fits to data from deep-inelastic scattering, Drell-Yan and multijet data. The α_s uncertainty is related to the choice of the value of the

Process	cross section [fb]	Scale[%]	PDF [%]	α_s [%]	Best theory prediction
$t\bar{t}H$	507.1	+5.8 -9.2	±2.0	±3.0	NLO QCD+EW
$t\bar{t}Z$	839.3	+9.6 -11.3	±2.8	±2.8	NLO QCD+EW
$t\bar{t}W^+$	397.6	+12.7 -11.4	±2.0	±2.6	NLO QCD+EW
$t\bar{t}W^-$	203.2	+13.3 -11.7	±2.1	±2.9	NLO QCD+EW
$t\bar{t}$	$831.76\ 10^3$	+2.4 -3.5	± 4.2		NNLO+NNLL

Table 1.2: Production cross sections of the $t\bar{t}$ processes at 13 TeV.

strong coupling constant in the calculation of the cross section. The interplay between PDF and α_s explains the common uncertainty provided in the case of the prediction of the $t\bar{t}$ cross section.

1.1.6 Higgs physics

Postulated to solve the issue of the mass of the vector bosons from the electroweak interaction, it took nearly fifty years before the predicted Higgs boson was discovered at the LHC [7, 8]. In the last years preceding its discovery, experimental constraints were set on the mass range at which it could be found. The experiments at LEP2 were able to set a lower limit on the Higgs boson mass through Higgsstrahlung ($e^+e^- \to Z^* \to ZH$) up to a mass of about 114.5 GeV [5]. Searches at the Tevatron further constrained the mass of the Higgs boson between 114 and 185 GeV [6].

Production

Four main production mechanisms are available for the Higgs boson at the LHC. The leading process is the production through gluon-gluon fusion, followed by vector-boson fusion, Higgsstrahlung and associated production with heavy quarks (Figure 1.13).

The production cross section through gluon-gluon fusion amounts to 50 pb a 13 TeV, while the associated production with heavy quark, despite increasing by a factor four between 8 and 13 TeV, is equal to 0.5 pb (Figure 1.14).

Decay modes

Once the mass of the Higgs boson is known, the partial decay widths can be fully predicted from the mass of the decay products. The coupling of the Higgs to massive gauge bosons scales with the square of their mass, the coupling to fermions scales with their masses,

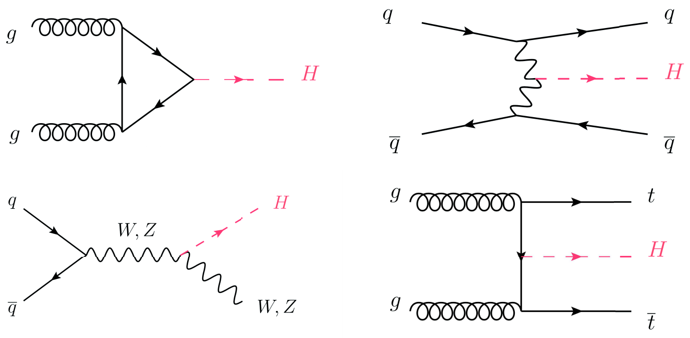


Figure 1.13: Higgs production modes: gluon-gluon fusion (top left), vector-boson fusion (top right), Higgsstrahlung (bottom left) and associated production with heavy quarks (bottom right).

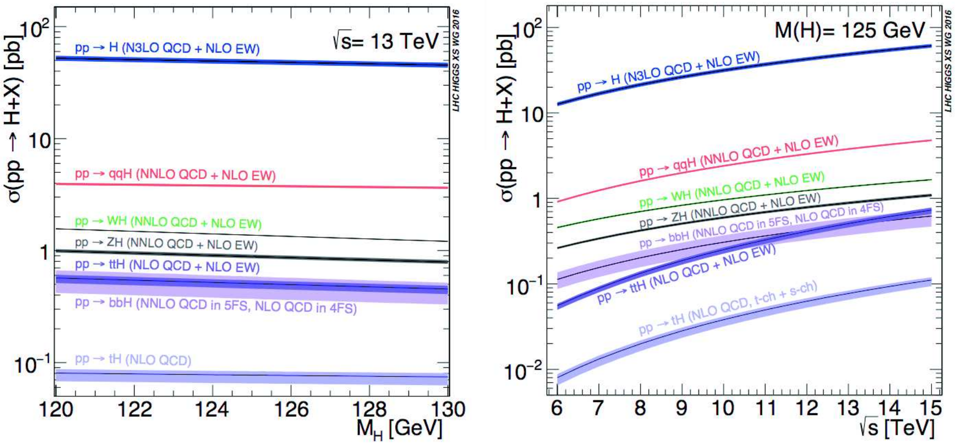


Figure 1.14: The SM Higgs boson production cross sections at $\sqrt{s} = 13$ TeV as a function of the Higgs boson mass (left) and as a function of the LHC centre-of-mass energy (right).

leading to a privileged coupling of the Higgs boson to the most massive particles.

The mass of the Higgs boson having been found to be 125.26 ± 0.21 GeV [58], its main observable decay is into b$\bar{\text{b}}$ (57 %), followed by the decay to WW*, $\tau\tau$, ZZ* and $\gamma\gamma$.

The mass of the Higgs boson being smaller than the mass of the pair of W or Z bosons, the decay goes to WW* (22%) or ZZ*(3%), meaning that one of the bosons will be offshell (Figure 1.15).

The Standard Model doesn't allow couplings of the Higgs boson to massless particles and the decay of the Higgs boson to $\gamma\gamma, Z\gamma, gg$ can only happen through heavy particle loops. Because these decay modes imply higher orders in the coupling constants, the branching ratio is expected to be smaller. The privileged coupling of the Higgs boson to heavy particles (especially top quark) keeps the branching ratio of the Higgs boson to $\gamma\gamma$ at the level of 2‰.

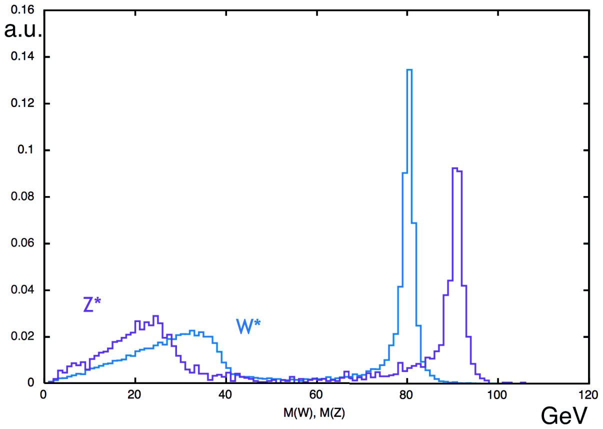


Figure 1.15: Mass spectrum of the WW* and ZZ* boson pairs in Higgs decay.

Analysis of the Higgs decays

In order to place the studies of the associated production of the Higgs boson with heavy quarks in the broader context of the Higgs physics, a brief review of the results obtained by CMS in 2012 and updated with the data taken at 13 TeV seems useful.

H $\rightarrow \gamma\gamma$

The $H \rightarrow \gamma\gamma$ analysis relies on the fit of a narrow peak in the invariant mass distribution of the diphoton system. A score is attributed to the photon identification based on a multivariate analysis using variables such as the shower shape, the isolation, the median energy density and the photon kinematics. The score is then used together with the diphoton mass resolution and kinematic variables related to the diphoton pair as input to a second multivariate analysis aiming at discriminating between signal and background. The main backgrounds come from $\gamma\gamma$, $\gamma + jet$ and $jet + jet$ production (Figure 1.16). The measured signal strength (ratio of the observed over expected contribution from the Higgs boson in the Standard Model) with the full 2016 statistics is $\mu = 1.16^{+0.15}_{-0.14}$ [57].

H $\rightarrow$ ZZ*

The $H \rightarrow ZZ^* \rightarrow 4l$ is the *golden channel* for the study of Higgs properties. The analysis relies on the reconstruction of the two Z bosons and the splitting in lepton categories ($4e$, 4μ, $2e2\mu$) with low background. In 2016, the analysis was performed with 36 fb^{-1} of data collected at 13 TeV (Figure 1.17). The total and differential cross sections are in agreement with the Standard Model expectations with a signal strength of $\mu = 1.05^{+0.19}_{-0.17}$. The analysis led in 2016 to the best measurement of the Higgs mass with a value of $m_H = 125.26 \pm 0.20(stat.) \pm 0.08(syst.)$ GeV [16]. Further analyses are performed to complete the decay channels of the Higgs to ZZ* such as $H \rightarrow ZZ^* \rightarrow 2l2\nu$, $H \rightarrow ZZ^* \rightarrow 2l2q$, ...

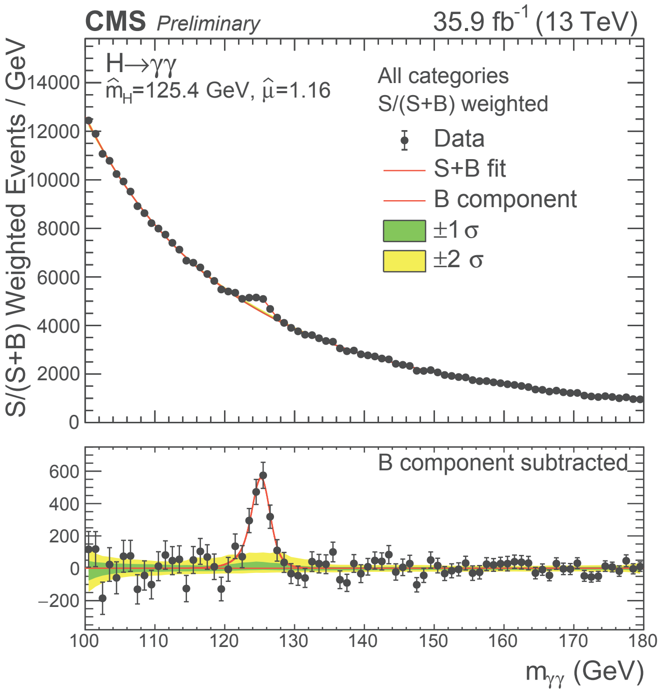


Figure 1.16: Distribution of the diphoton invariant mass for all categories in 2016 [57].

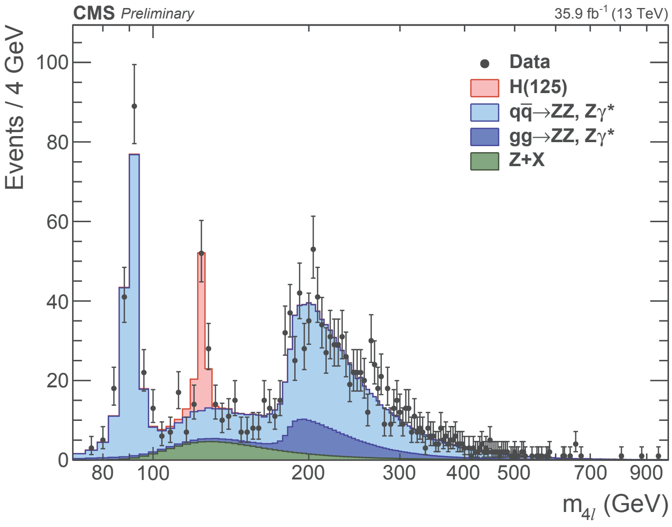


Figure 1.17: Distribution of the four lepton invariant mass in the H $\rightarrow$ ZZ* $\rightarrow$ 4 ℓ channel [16].

H → WW∗

The $H \rightarrow WW^*$ decay has a large branching ratio but suffers from a low mass resolution due to the presence of two neutrinos. The $2\ell2\nu$ final state has a relatively high purity and moderate backgrounds especially in the eμ channel. In the analysis, based on the first 2.3 fb^{-1} collected at 13 TeV in 2015 and 12.9 fb^{-1} collected in 2016 [18], the signal strength is found to be 1.05 ± 0.26.

$\mathbf{H} \rightarrow \tau^+\tau^-$

The search for the decay of the Higgs boson in a pair of τ lepton relies on the reconstruction of the di-tau mass spectrum in categories based on the different final states (Figure 1.18). The best fit of the signal cross section times branching ratio in 2016 is $1.06^{+0.25}_{-0.24}$ times the standard model expectation and leads to the first observation of the Higgs boson decaying to a pair of τ leptons [19].

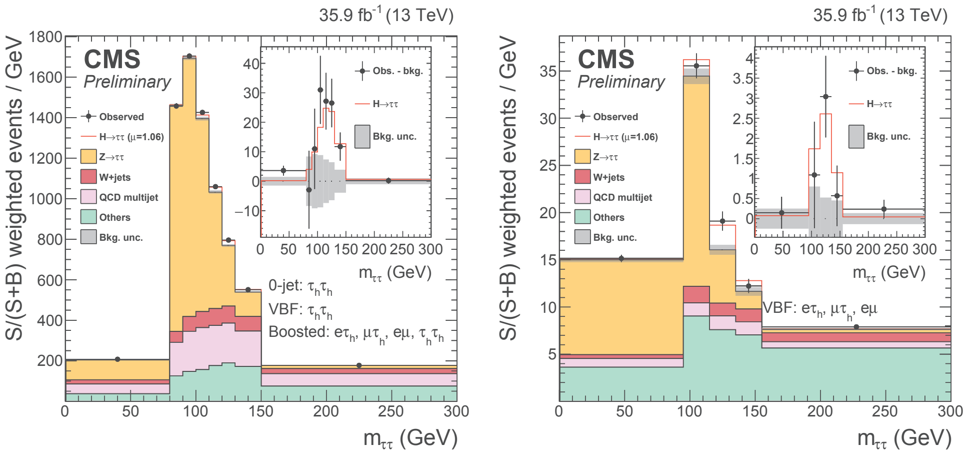


Figure 1.18: Distribution of the di-tau mass spectrum in 2016 [19] in fully hadronic modes and boosted topologies (left), and eμ and semileptonic modes (right).

$\mathbf{H} \rightarrow \mathrm{b\bar{b}}$

The last decay channel which can be considered with a large branching ratio is the decay of the Higgs to a $\mathrm{b\bar{b}}$ quark pair. The analysis relies on the performance of the identification of a jet produced by hadronization of a b quark. During Run 1, the analysis strategy was to study the associated production of the Higgs boson with a W or Z boson decaying leptonically, leading to a signal strength of 1.0 ± 0.5. The latest results [20] are using the production by fusion of vector bosons. This production mode has as a higher cross section but the final state which contains four jets, among which two are coming from b quarks suffers from the large QCD background. The signal extraction is performed through a fit of the di-jet invariant mass spectrum. The measured signal strength when combined with 8 TeV data is $\mu = -1.3^{+1.2}_{-1.1}$.

Properties

Following its discovery, studies have been performed to check the compatibility of the observed Higgs boson with the properties predicted by the Standard Model and to improve these measurements.

Mass of the Higgs boson

The mass of the Higgs boson was until its discovery the last missing parameter of the Standard Model. The mass is determined by the combination of the measurements in analyses allowing a full reconstruction of the decay products, namely $H \to \gamma\gamma$ and $H \to ZZ^{*}$. While the first measurements already had a resolution of the order of 1 to 2% per channel, the combination of results at 7 and 8 TeV and the latest result from CMS in the ZZ* channel brings the precision on the Higgs mass at the level of 0.2%.

Spin parity

The spin parity of the Higgs boson can be studied in various channels, testing different hypotheses depending on the kinematic of the final states. Thus, considering the decay to $\gamma\gamma$ allows to test 0^{+} against 2^{+}, based on the angle between the two photons. The decay to WW* allows to test 0^{+} against both 0^{-} and 2^{+}. The richest channel is the decay to ZZ* into four leptons which allows, by studying the kinematic of the decay angles of leptons in both Z bosons to test 0^{+} against 0^{-}, $1^{\pm}$ and $2^{\pm}$. The pseudoscalar, spin-1 and spin-2 hypotheses are excluded at a confidence level above 99.9%, 99.9% and 95% respectively, in agreement with a standard Higgs boson of spin 0^{+}.

Width of the Higgs boson

In the same way that the measurement of the total decay width of the Z boson allows to probe the existence of a fourth light neutrino (and thus a fourth family of fermions), the width of the Higgs boson is of particular interest in order to study the coupling of the Higgs boson to all the massive particles. It allows to probe the existence of invisible particles.

While too narrow (4.2 MeV) to be measured at the LHC, indirect methods developed in the last few years allowed to constrain the Higgs width by using its decay into $\gamma\gamma$, WW* and ZZ*. Based on the on-shell and off-shell relative productions of the Higgs boson, a constraint of $\Gamma_H < 22$ MeV at 95% confidence level was set by CMS, corresponding to 5.4 times the expectation of the Standard Model [11].

Couplings of the Higgs boson

Based on the various analyses, targeting dedicated production or decay modes of the Higgs boson, combinations are performed to check the overall compatibility with the Standard Model. Most of these combinations are done in the narrow width approximation, fixing the expected width to the nominal value predicted by the Standard Model. The couplings are found to be compatible with the prediction of the Standard Model. While most of the couplings are measured through the decay of the Higgs boson in pairs of fermions or bosons, the large mass of the top quark forbids the decay of the on-shell

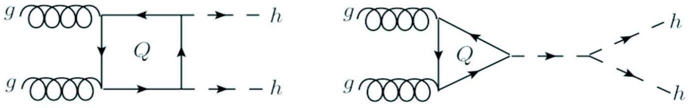


Figure 1.19: Double Higgs production through gluon fusion mediated by loops of heavy quarks splitting of the Higgs to two Higgs bosons.

Higgs boson into a top-antitop pair. The decay of the Higgs boson into a pair of photons provides a probe of the top-Higgs coupling via the heavy quark loop dominated by top quarks.

Self coupling of the Higgs boson

A more difficult study related to the Higgs boson will be the measurement of its self coupling. The trilinear coupling of the Higgs might be studied through the production of one off-shell Higgs boson splitting into two real Higgs bosons but this very rare process will suffer from a background coming from the associated production of two Higgs bosons, making this study particularly challenging at the LHC (Figure 1.19). The trilinear coupling might be studied at HL-LHC while the quartic coupling could be studied at a future collider such as the ILC.

1.2 Beyond the Standard Model

Despite being incredibly succesful, the Standard Model of particle physics shows some limits, both from a theoretical and an experimental point of view.

1.2.1 Weaknesses of the Standard Model

The reason for the structure $SU(3)_c \times SU(2)_L \times U(1)_Y$ of the Standard Model remains mysterious. In the same way, the masses are free parameters of the model and the large difference between the masses of light and heavy quarks or leptons is not explained. The reason for the arbitrary number of families and the flavour structure is unknown. The neutrino masses, as infered from the observed neutrino oscillations, are not included in the Standard Model.

For a scalar field such as the Higgs field, no symmetry can remove the mass term. As such, the mass of the Higgs boson should diverge with the loop corrections if no mechanism is introduced to cancel them.

The asymmetry between matter and anti-matter cannot be explained in the context of the Standard Model. Despite the CP violation described in the electroweak sector of the Standard Model, the measured violation is too weak to explain such an asymmetry in the universe.

The Standard Model describes only three of the four fundamental interactions as it is not able to account for gravity.

The electroweak scale is of the order of 100 GeV while the gravitational scale is sixteen orders of magnitude larger.

Cosmological observations indicate the presence of Dark Matter which is not predicted by the Standard Model and is four times more abundant in the Universe than the matter accounted for in the Standard Model. Moreover, Dark Energy is expected to be the main contribution to the energy of the universe.

Various theories were proposed in order to solve one or several of those shortcomings. Extending the SM lagrangian can be done in several ways. By adding a new particle with a weak coupling to SM particles, by adding a new force in the lagrangian or by adding a new symmetry.

One of the appealing extension of the Standard Model is the Supersymmetry. The number of particles is doubled by adding a symmetry between fermions and bosons, each fermion (boson) from the Standard Model being associated to a supersymmetric boson (fermion). This extension helps stabilizing the mass of the Higgs boson if the symmetry between particles and their superpartners is weakly broken, and provides a candidate for Dark Matter if the lightest neutral supersymmetric particle is stable. Since none of the masses of the new particles is predicted in the theory, searches for Supersymmetry rely on a simplified mass spectrum of the supersymmetric particles, probing the mass spectrum for one heavy supersymmetric particle decaying into the lightest ones predicted by the model.

1.2.2 New physics in the top sector

In the top sector, new physics can be studied via direct searches or indirect searches.

Direct searches target production or decay modes which are not predicted by the Standard Model. An example is looking at the $\mathrm{t\bar{t}}$ mass spectrum, allowing to probe the existence of high mass resonances such as $Z' \rightarrow \mathrm{t\bar{t}}$ (Figure 1.20) or $W' \rightarrow \mathrm{t\bar{b}}$.

Indirect searches use the precise measurements of the top quark properties to set limits on the phase space in which new physics could arise. The absence of deviation of the $\mathrm{t\bar{t}}$ production cross section from that expected in the Standard Model allows to set limits on

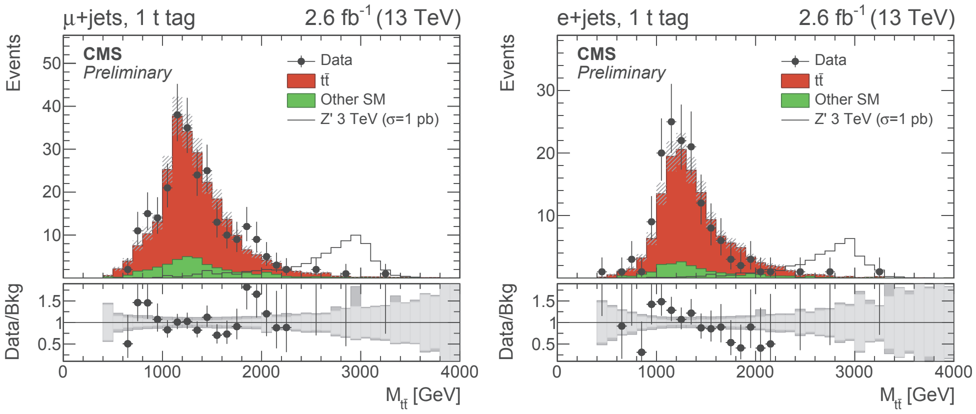


Figure 1.20: Distribution of the invariant mass of the reconstructed $t\bar{t}$ pair for data and expected background for events passing the signal selection in the muon (left) and electron (right) channel, in the one top tagged category [22].

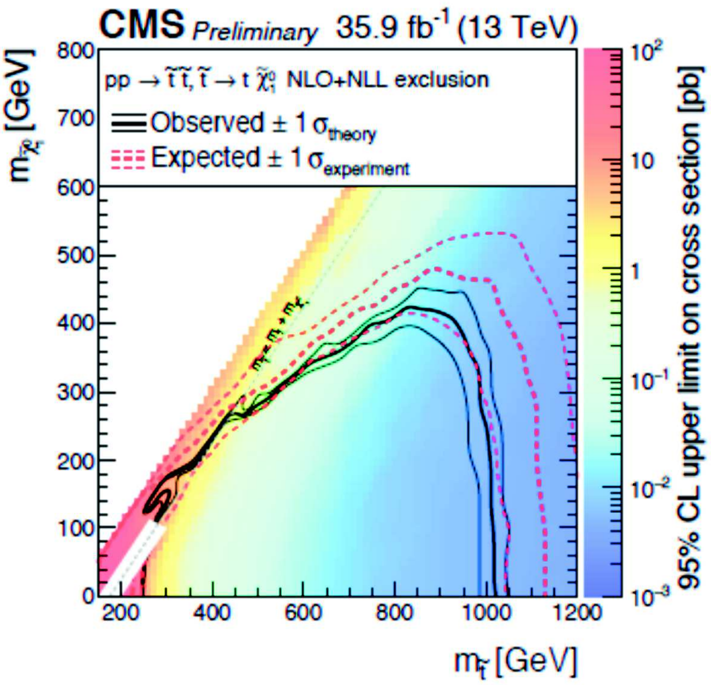


Figure 1.21: Exclusion limit obtained in the neutralino mass / stop mass plane from direct stop searches [23].

the cross section of stop pair production. In the same way, the existence of a new particle with mass lower than the top and compatible with the decay $t \to Xb$ would reduce the branching ratio of the known decays of a $t\bar{t}$ pair [22].

The white area of Figure 1.21 corresponds to a mass of the stop close to the mass of the neutralino plus the mass of the top, which would lead to a signature similar to a $t\bar{t}$ pair production. Limit on the cross section of this part of the phase space can therefore come from reinterpretation of the measurement of the $t\bar{t}$ production cross section.

1.2.3 New physics in the Higgs sector

Among the avenues proposed to solve some of the problems raised by the Standard Model, some are directly related to the Higgs sector and the hierarchy problem. A possibility in order to explain the difference between the electroweak and the Planck scales is to have a composite Higgs. In this case, the Higgs boson can be a bound state of two fermions which would lead to properties different than the ones predicted by the Standard Model. Another possibility is the existence of extra space dimensions. The last possibility is a new (broken) symmetry between fermions and bosons which would lead to an extended Higgs sector. In the context of the minimal extension of the Standard Model to a Supersymmetric Standard Model, four additional physical Higgs states would exist.

Many searches are ongoing for Higgs bosons beyond the Standard Model. Such searches can be performed looking for a resonance in the diboson spectra at higher masses. In the Neutral MSSM Higgs model, the coupling to down-type fermions is increased, leading to higher branching ratio to τ leptons and b quarks. Searches are as well ongoing for a charged Higgs, with masses both higher and lower than the top mass. For masses of the charged Higgs below the top mass, the main production mode would be through top decay, while for masses above the top mass, the main production mode would be through gluon fusion with a possible decay into top. In the context of a Type I Two Higgs Doublet Model, the charged Higgs can be lighter than the top quark. In this case, the signal would overlap with the $\mathrm{t\bar{t}H}$ production, making them accessible by looking for a deviation of the $\mathrm{t\bar{t}H}$ cross section. Finally, doubly charged Higgs are predicted by extending the SM with a scalar triplet which could explain the neutrino masses. Doubly Charged Higgs searches are looking for excess in events with same signe lepton pairs. Behind the possibility of a new particle, searches are done in order to explore the possibility of non SM decays of the Higgs boson. Among the many final states and exotic particles which could emerge from non SM decays of the Higgs boson, one of the study which is ongoing at the LHC is the search for lepton flavour violation of the Higgs decays, with the Higgs boson decaying into $\mathrm{e}\tau$ or $\mathrm{e}\mu$ final states.

Despite a large zoology of searches for new physics in the Higgs sector, no hint has been found for such new physics, neither in looking for new particles, nor in non Standard Model decays of the Higgs boson.

Chapter 2

Experimental apparatus: Accelerator, detector and operation

Contents

After a first period of operation from 2010 to 2012 that led to the discovery of the Higgs boson by the CMS and ATLAS experiments, the Large Hadron Collider (LHC), the world's largest particle accelerator, stopped for two years of upgrade. The LHC restarted operation at higher collision energy (13 TeV) in 2015. In the first section, we will shortly describe the accelerator and its operation. In the second section, we consider the design of the Compact Muon Solenoid, focusing on the different sub-detectors, their performances and the upgrades they have undergone to sustain the new data taking conditions starting in 2015.

2.1 The Large Hadron Collider

Built in the former LEP tunnel, the Large Hadron Collider is a proton-proton collider of 27 km of circumference designed to produce collisions at an energy of 14 TeV in the centre-of-mass. It reached 7 TeV in 2010-2011, 8 TeV in 2012 and is running at 13 TeV since 2015. The LHC also allows collisions between heavy ions and between protons and heavy ions. The maximum energy the LHC can reach has been constrained by the strength of the magnets used at the time it was designed and by the size of the LEP tunnel.

2.1.1 The acceleration complex

The LHC is the last step of an acceleration complex beginning with a simple hydrogen bottle and composed of linear (LINAC) and circular (PS, SPS) accelerators (Figure 2.1). The LINAC accelerates protons up to 50 MeV. The PS inherits from 1.5 GeV protons accelerated further in the PS Booster and raises the energy to 25 GeV. The SPS then raises the energy up to 450 GeV before injecting protons in the LHC where they are accelerated up to the design energy using radiofrequency cavities.

2.1.2 Energy and luminosity

The performance goals of accelerators as the LHC are based on a physics program at the highest possible energy in order both to understand physics through high precision measurements and possibly to discover physics beyond the Standard Model. The collision energy had to allow the discovery of the long awaited Higgs boson, whatever the value of its mass between 100 and 1000 GeV. Based on this purpose, the LHC has been designed to provide beams of protons colliding at a centre-of-mass energy of 14 TeV and an instantaneous luminosity of $10^{34} cm^{-2} s^{-1}$ (excluding the possible use of anti-proton

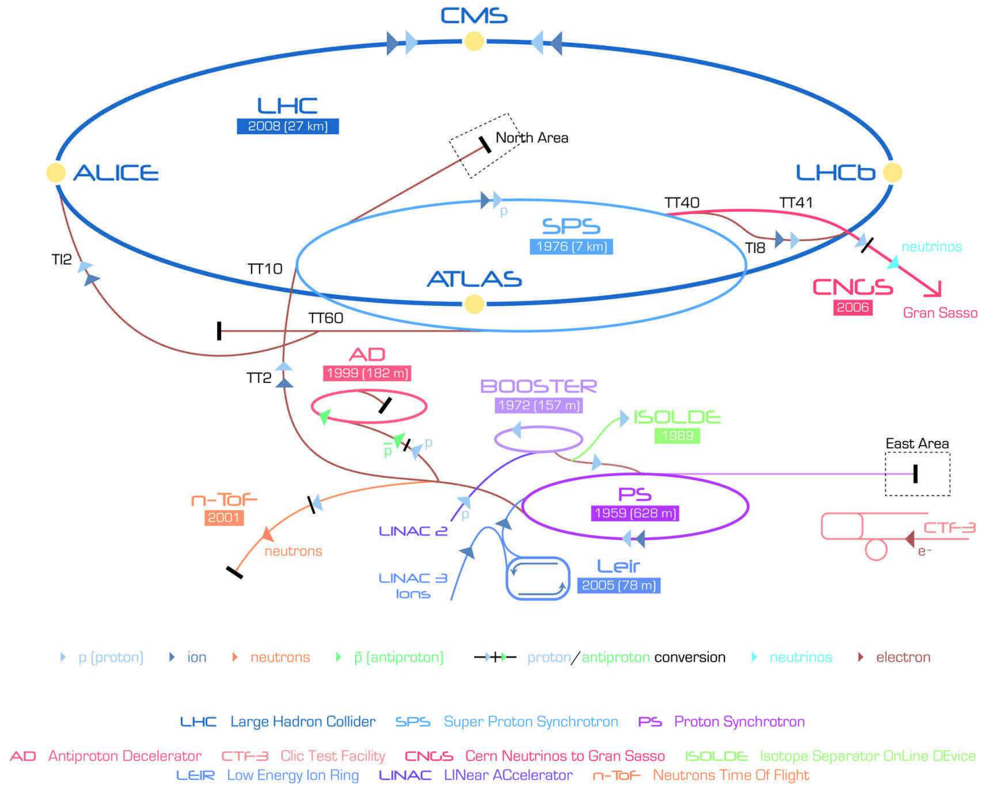


Figure 2.1: The LHC acceleration complex.

beams). Reaching this energy allows to probe many signals whose thresholds are in the TeV region. Reaching this luminosity allows for enough statistics to be collected in order to study rare processes, such as for instance Higgs boson production during the first years of data taking and associated production of top and Higgs or production of two Higgs during the following years.

The number of events produced at the LHC every second is given by:

$$N_{events} = L \times \sigma_{event} \qquad (2.1)$$

where σ_{event} is the cross section of the considered process and L is the instantaneous luminosity.

Processes	$\sigma\cdot$ BR [pb]	Rate at 1.0 10^{34}	Rate at 1.5 10^{34}
$W \to \ell\nu$	20 400 ($\pm$ 5 %)	204 Hz	306 Hz
$W^+ \to \ell^+\nu$	11 700 ($\pm$ 5 %)	117 Hz	175 Hz
$W^- \to \ell^-\nu$	8 640 ($\pm$ 5 %)	86 Hz	129 Hz
$Z \to \ell^+\ell^-$	1 930 ($\pm$ 5 %)	19 Hz	30 Hz
$t\bar{t}$	830 ($\pm$ 6 %)	8 Hz	14 Hz
H	50 ($\pm$ 5%)	0.5 Hz	0.75 Hz

Table 2.1: Process cross sections and rates at 13 TeV.

Table 2.1 shows the number of events produced every second for two instantaneous luminosities reached during the data taking in 2016. A W boson is produced and decays to a lepton between 200 and 300 times per second while one Higgs boson is produced every 2 seconds.

The instantaneous luminosity depends on the beam parameters in the following way:

$$L = \frac{n_1 \cdot n_2 \cdot k_b}{4 \cdot e_n \cdot \beta^*} F \tag{2.2}$$

where N is the number of protons per beam, k_b the number of bunches, e_n the emittance, β^* the amplitude function and F the frequency.

During the Run 1, the LHC produced collisions at 7 and 8 TeV with a luminosity up to 7.7 $10^{33} cm^2 s^{-1}$. The successful operation during the three years of data taking led to an integrated luminosity of nearly 30 fb^{-1}.

The goal for Run 2 was twofold, to go both to a higher energy and a higher luminosity, the two challenges relying on different machine requirements.

Going to higher energy necessitates a dedicated training of the magnets, but will allow for an increase of the production cross sections of interesting processes (Figure 2.2).

Going to higher instantaneous luminosity implies a linear increase of the number of simultaneous pp collisions (pileup) happening during a single bunch crossing, but increases the production rate of events useful for physics. To go to higher luminosities, several strategies can be considered based on the different parameters of Equation (2.2):

- increase the total number of bunches (k_b) by reducing the spacing between bunches (from 50 ns at Run 1 to 25 ns at Run 2);
- increase the number of particles per bunch (N);
- decrease the focal length at the impact parameter from 60 cm to 40 cm (β^*);
- reduce the transverse size of the bunches.

To better emphasize the new challenges related to Run 2, Table 2.2 summarizes the parameters associated to each period of data taking.

The number of collisions per bunch crossing (pile-up) is a key feature at the LHC. The high pile-up leads to many tracks and clusters in the detector which makes the object reconstruction more difficult (Figure 2.3).

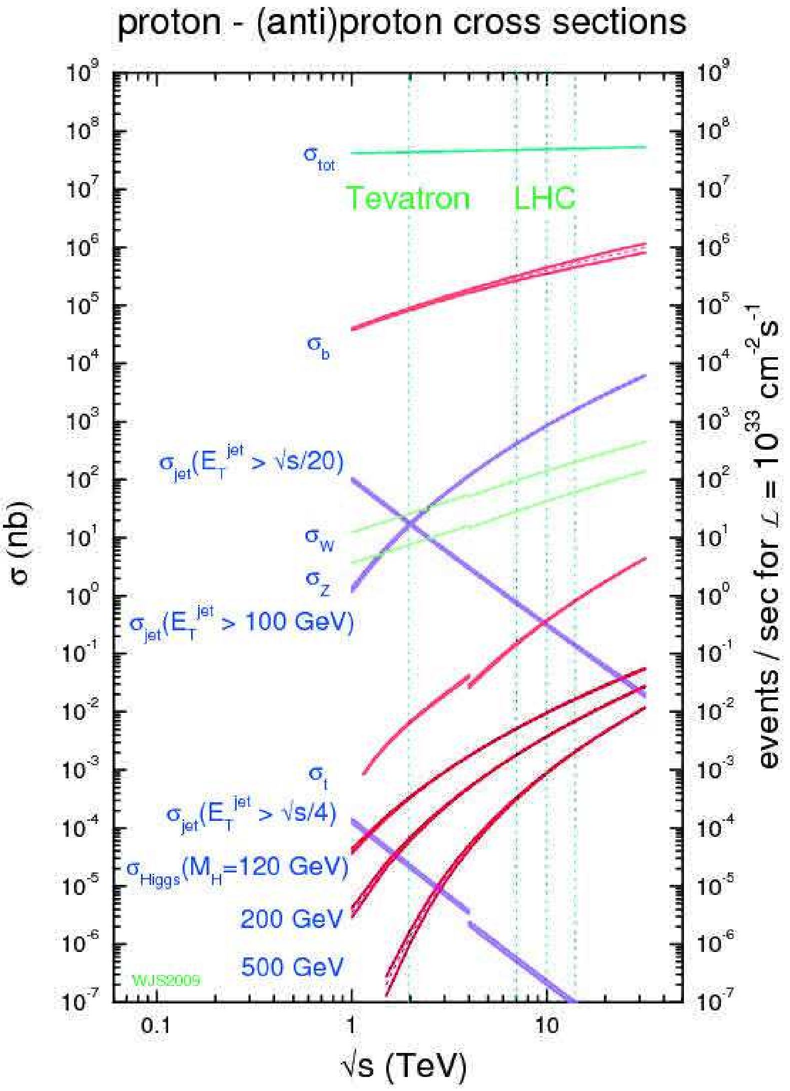


Figure 2.2: Cross sections of different physics processes vs centre-of-mass energy.

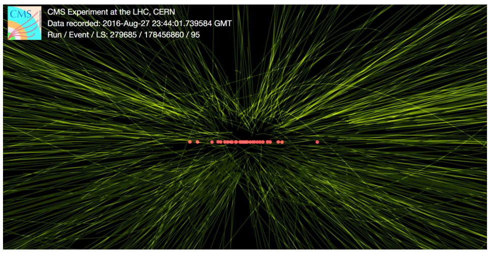

CMS
CMS Experiment at the LHC, CERN
Data recorded: 2016-Aug-27 23:44:01.739584 GMT
Run / Event / LS: 279685 / 178456860 / 95

Figure 2.3: Event display at PU=30.

Since the pile-up condition during the data taking can not be predicted for the whole year before the production of the simulation, the pile-up is measured from data (Figure 2.4) and the simulation is reweighted based on the observed difference between the

Parameter	Run I (2012)	Run II (2016/2017)	Design value
Beam energy [TeV]	4.0	6.5	7.0
Average bunch intensity (10^{11} p)	1.5	1.1	1.15
Number of bunches	1380	2244	2808
β^* [m]	0.6	0.4	0.55
Bunch spacing [ns]	50	25	25
Emittance at start of collision [μm]	2.4	3.4	3.75
Peak luminosity [$cm^{-2}s^{-1}$]	$7.7\ 10^{33}$	$1.58\ 10^{34}$	$1\ 10^{34}$
Number of collisions per bunch crossing	≈ 27	≈ 40	20

Table 2.2: Beam parameters during Run I and Run II.

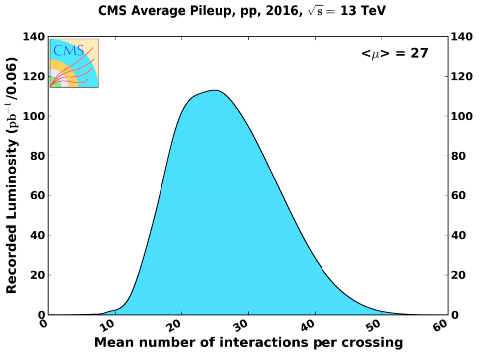


Figure 2.4: Mean number of interactions per crossing.

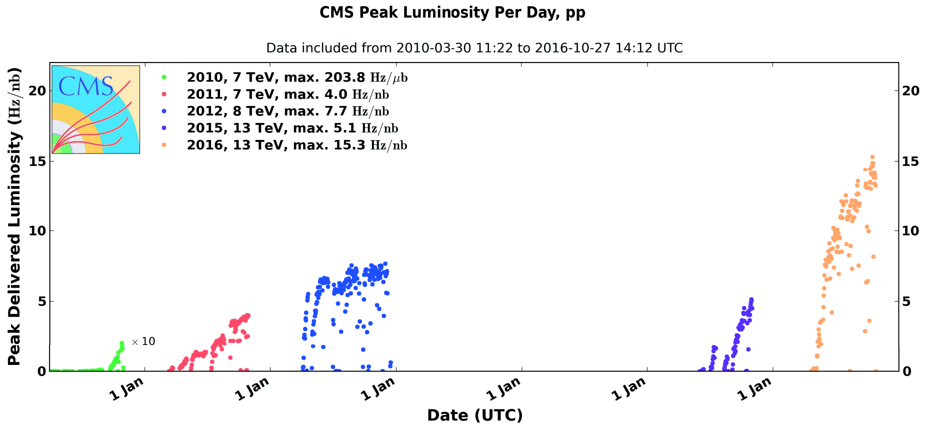
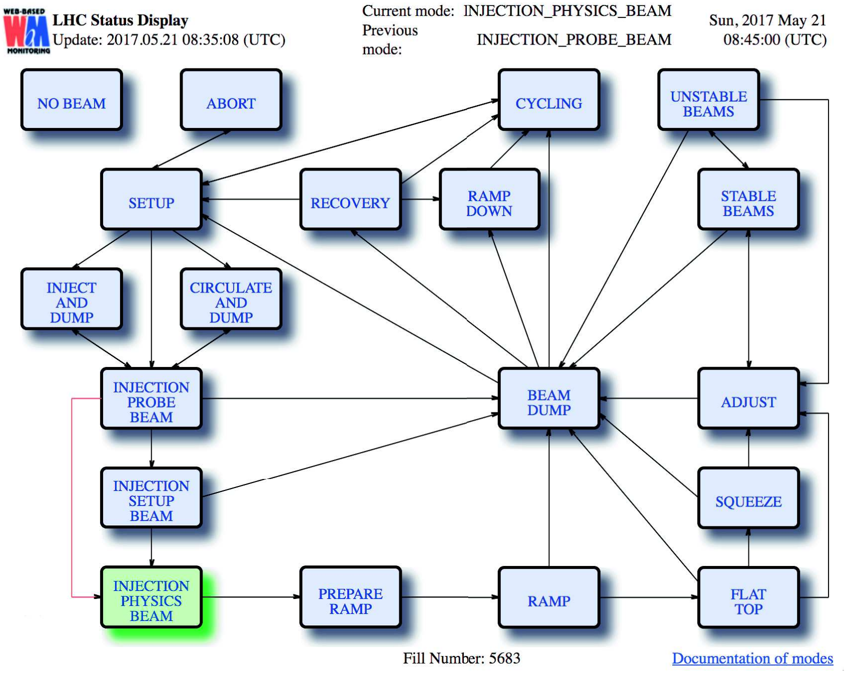


Figure 2.5: Instantaneous luminosity along years.

distributions in data and simulation.

Thanks to the modifications applied during LS1 (dipole training through training quenches), the new LHC conditions were reached, the energy during Run 2 rose to 13 TeV in the centre-of-mass and the nominal luminosity expected at the LHC (10^{34}) was reached and surpassed in 2016, leading to challenging data taking conditions for the detectors (Figure 2.5).



LHC Status Display
Update: 2017.05.21 08:35:08 (UTC)

Current mode: INJECTION_PHYSICS_BEAM
Previous mode: INJECTION_PROBE_BEAM

Sun, 2017 May 21
08:45:00 (UTC)

Fill Number: 5683

Figure 2.6: LHC running operation.

In 2017 the luminosity rose to $1.58 \cdot 10^{34}$ in the first weeks of data taking and is expected to reach twice the nominal luminosity, $2 \cdot 10^{34}$ $\text{cm}^2.\text{s}^{-1}$ by the end of the year. All the developments made for the data taking in CMS are driven by this unexpected challenge.

2.1.3 Running the LHC

From a detector point of view, the LHC operation (Figure 2.6) can be split in four main operation modes.

When no beam is circulating, the detector can perform tests of the sub-detectors or take cosmic data which will be used for alignement, calibration, ...

In injection mode, the LHC is filled with a probe beam and then according to a given bunch spacing scheme with 50 ns or 25 ns between bunches.

After injection, the accelerator prepares the ramp up, which rises the energy of the protons to 6.5 TeV (from 2015). When this energy is reached, the accelerator is in flat top and the detector moves to collision mode.

The beam is then squeezed and ajusted, leading to stable beams at nominal energy which will be used for collisions.

At any time, the beam can be dumped to ensure the integrity of the accelerator and

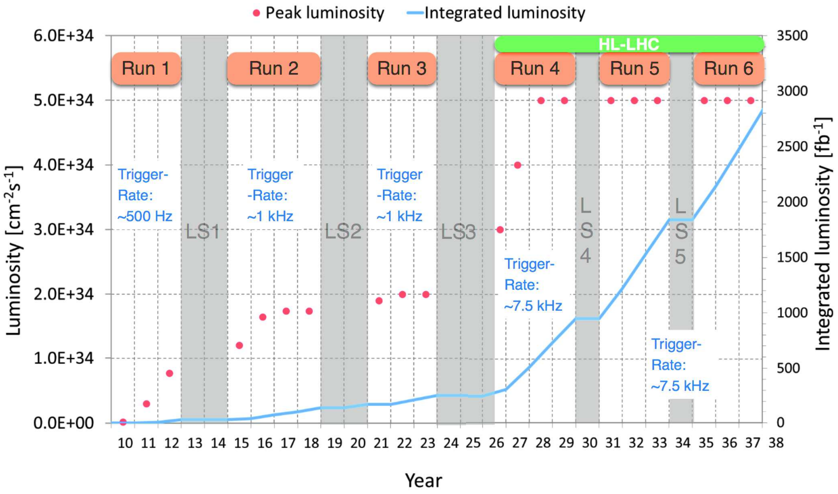


Figure 2.7: Timeline of the LHC upgrade schedule.

the detectors. Many reasons can lead to a beam dump, ranging from beam unstability to sub-detectors (such as the Roman Pots) moving too close to the beam axis.

2.1.4 The (successful) beginning of a much longer adventure

Despite the tremendous success of the physics program at LHC, the ten first years from 2008 to 2018 are just the beginning of a much longer adventure.

The Run 1 lasted from 2010 to 2012 and led to the discovery of the Higgs boson. The Run 2 began in 2015 and allowed us to probe the structure of the electroweak sector while putting stronger limits on new physics. The Run 3 will begin in 2021, with a similar instantaneous luminosity than at the end of the Run 2 and should lead to twice the integrated luminosity recorded during the Run 2.

In 2026, the second era of exploitation of the LHC, the High Luminosity LHC will begin at 14 TeV, targeting an integrated luminosity ten times larger and a data taking rate increased by a factor 7.5 (Figure 2.7).

2.2 The CMS detector

CMS (Compact Muon Solenoid) is one of the two multi-purpose experiments operating at the LHC (Figure 2.8). The "S" of CMS emphasizes the choice of using a solenoid producing an intense magnetic field (3.8 T) in order to have a good resolution in momentum. It is the largest superconducting coil in the world (6 m in diameter and 12.5 m long). The term Compact is related to the relatively small size of the detector and its high density, necessary to confine and measure the energy of particles produced at high energy. The size is constrained by the internal radius of the solenoid in which several sub-detectors are placed. Muon sub-detectors are placed in the return yoke of the magnet outside of the solenoid to measure the momentum of muons escaping the detector.

2.2.1 Physics motivation and detector requirements

Detector performance requirements derive from the sensitivity necessary to detect dedicated physics signals at the LHC.

A good charged particle momentum resolution and reconstruction is, for instance, necessary for an efficient tagging of τ leptons and b quark jets, in order to be able to study the decay of B_s to J/Ψ, the decay of the MSSM heavy neutral Higgs H(A) to a pair of τ leptons and the study of the coupling of the Higgs to the top quark through $t\bar{t}H, H \to b\bar{b}$.

A good electromagnetic energy resolution is needed to reconstruct the decays $H \to \gamma\gamma$ and $H \to ZZ^* \to 4$ electrons, the two main channels leading to the discovery of the Higgs boson. The energy resolution allows for a precise measurement of the Higgs mass. Moreover, a good electromagnetic resolution at high energy is necessary for the study of the potential decay $Z' \to e^+e^-$.

A good E_T^{miss} (see Section 3.7) resolution is necessary for new physic searches, as supersymmetric models provide dark matter candidates with a stable weakly interacting neutral particle giving high E_T^{miss}. A good dijet invariant mass resolution is necessary to explore new resonances at high mass (from extra-dimension theories for instance).

Finally, a good muon identification and momentum resolution is necessary for a high efficiency of reconstruction of processes such as $H \to ZZ^* \to 4$ muons, $H \to WW^* \to 2$ muons and high mass resonances in the dimuon spectrum such as $Z' \to \mu\mu$.

The constraints on the energy, momentum and spatial resolutions of the sub detectors were thus imposed by the expected signatures of the Higgs boson decay products and expected new physics channels. To study the decay of the Higgs boson into two photons, it is necessary to have a constant term on the energy resolution in the electromagnetic calorimeter of less than 0.5%. Less good resolution would lead to larger $\gamma\gamma$ peak making it more difficult to observe the Higgs boson in the diphoton mass spectrum. To study the decay of the Higgs boson into four leptons, we need an excellent track resolution. A 90% efficiency on each lepton would lead to an acceptance for the Higgs boson of 66%, which is far too low. It is therefore necessary to be able to reconstruct individual particles with reconstruction efficiency close to 100%. All these performances have been achieved and allowed to carry out the physics program of CMS.

The detector requirements can therefore be summarized as follows:

- Good charged particle momentum resolution and reconstruction efficiency,
- Good electromagnetic energy resolution,
- Good missing transverse energy (E_T^{miss}) and dijet mass resolution, hence a large coverage of the hadron calorimeter,
- Good muon identification and momentum resolution.

2.2.2 Coordinate system

The coordinate system is defined from the center of the detector. The y-axis points to the top of the detector, the x-axis points to the center of the LHC. The z axis is along the beam axis and oriented towards the Jura mountains. The polar angle Θ is defined relative to the z axis, the azimuthal ϕ angle being in the xy transverse plane. We define the pseudorapidity in order to have a polar angle quantity whose difference is a relativistic invariant: $\eta = -\ln(\tan\frac{\theta}{2})$. The angular distance between two objects is thus defined as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$.

2.2.3 Magnet

The main specificity of the CMS detector is the choice made to use a 3.8 T superconducting solenoid of 12.000 tons, aiming at bending the muons in a magnetic field such that the momentum resolution is better than 10% for p_T less than 1 TeV.

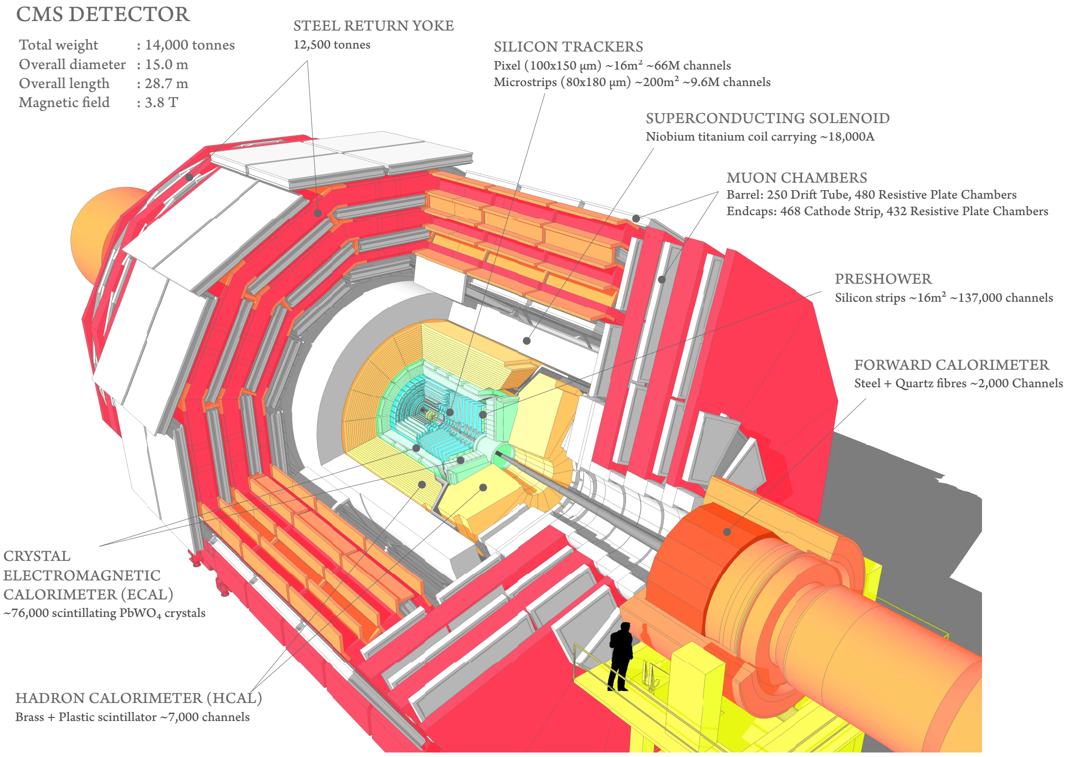

CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~16m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO₄ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Figure 2.8: Overview of the CMS detector.

2.2.4 Tracker

Closest to the beam axis, the tracker makes it possible to determine the trajectory of the charged particles, curved in the magnetic field. It consists of several silicon layers. The first layers are made of pixels in order to determine the position of the interaction point, seed the tracking and provide a good handle on the determination of possible secondary vertices. The following layers, composed of silicon strips, makes it possible to measure a track composed of up to 13 points before reaching the calorimeter (Figure 2.9). A good resolution in position will allow an accurate determination of the transverse momentum of charged particles.

Before being replaced during the winter shutdown at the end of 2016, the three layers of pixel detector contained 65 millions pixels. Each pixel had a depth of 250 μm and a size of 100 $\times$ 150 μm^2. The technology is based on semiconductors which will create a signal through the creation and displacement of an electron-hole pair in the silicon. The following strip layers provide a two-dimension position thanks to the stereo angle of 100 mrad between the strips. The width of the strip ranges from 80 to 141 μm in order to limit the rate and provide a good position resolution. Their length is of the order of 10 cm. The CMS tracker is the largest active surface in silicon ever built with an area of nearly 200 m^2, shared between the 66 millions of pixels and the 9.6 millions of strips.

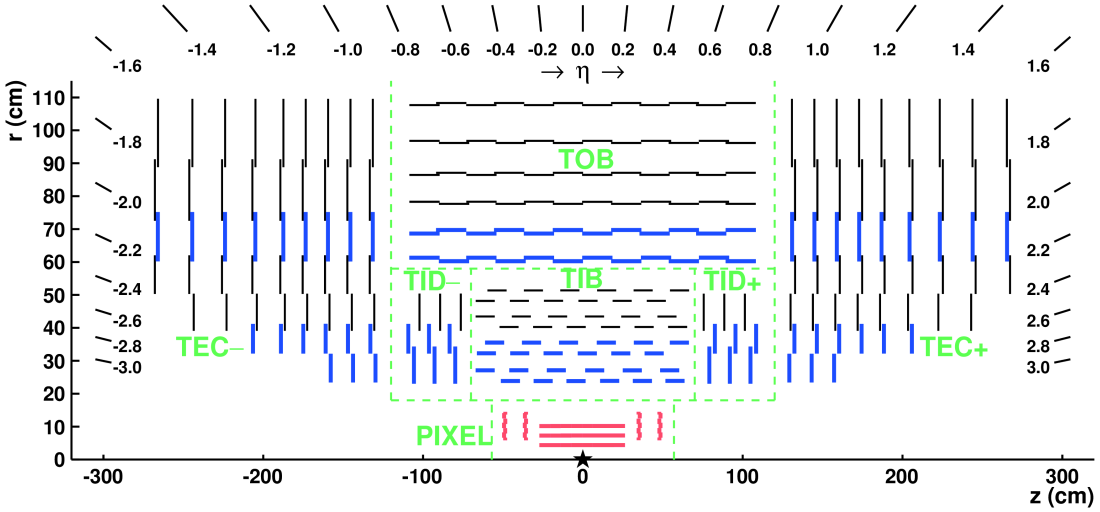


Figure 2.9: CMS Tracker (2016).

Being the detector closest to the beam pipe, it has to cope with a very high particle flux. The combination of two (doublets) or three (triplets) hits in the layers are the initial blocks in the reconstruction of tracks. The ten layers of silicon strip detectors are then used to reconstruct the full track. The low level of noise of the pixels and their low occupancy makes it possible to run the tracking despite the very high particle flux.

Any tracker technology corresponds to a particular compromise between read-out speed, spatial resolution, material budget, power dissipation and radiation hardness. The aim is to have the best spatial resolution and the lowest material budget, in order to reduce secondary interactions of the particles with the layers.The choice done by CMS was to use silicon pixels and silicon strips allowing for a spatial resolution below 10 μm thanks to an analog reading of the pixels, a temporal resolution below 20 ns for a material budget not exceeding 1 % X_0 per point (Figure 2.10 and Figure 2.11).

The performance of the tracker relies on the quality of the alignement. In order to determine the relative and absolute position of the pixels and strips, global and local fits are performed using cosmic-ray and tracks from the proton-proton collisions.

2.2.5 Electromagnetic and hadronic calorimeters

The purpose of the calorimeters is to measure particle energy. Their segmentation is chosen to minimize the stacking of two particles in the same cell, which would have the effect of deforming the signal.

Electromagnetic calorimeter

The electromagnetic calorimeter (Figure 2.12) consists of 76.200 very dense scintillator

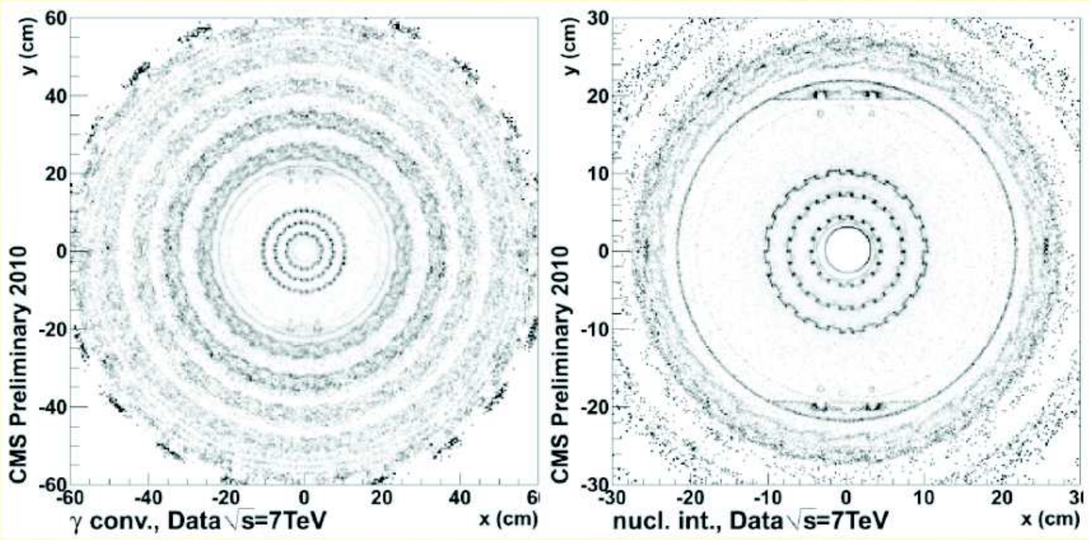


Figure 2.10: Material budget visualisation through photon conversions (left) and nuclear interactions (right).

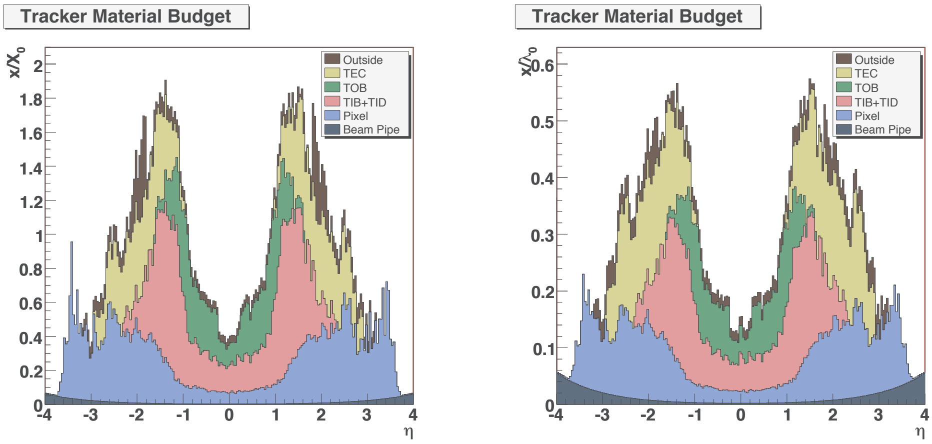


Figure 2.11: Material budget profile of the tracker: fraction of radiation length (left) and nuclear interaction length (right) as a function of pseudo-rapidity.

crystals of lead tungstate ($PbWO_4$) which measure the energy of electrons and photons that will create electromagnetic showers, due to the successive interactions of photons which produce electron-positron pairs and of electrons and positrons which radiate photons up to total absorption. The radiation length in such a dense material is 0.89 cm leading to a good energy measurement for crystals of few cm thickness. The characteristic dimensions of the crystals used were chosen in order to contain the entire shower. Their size (up to 26 X_0 depth, 1 Molière radius [1]) makes it possible to recover most of

[1] The Molière radius is the radius of a material containging 90% of the electromagnetic shower transverse deposition. For lead tungstate, the Molière radius is 2.2 cm leading to a good shower position

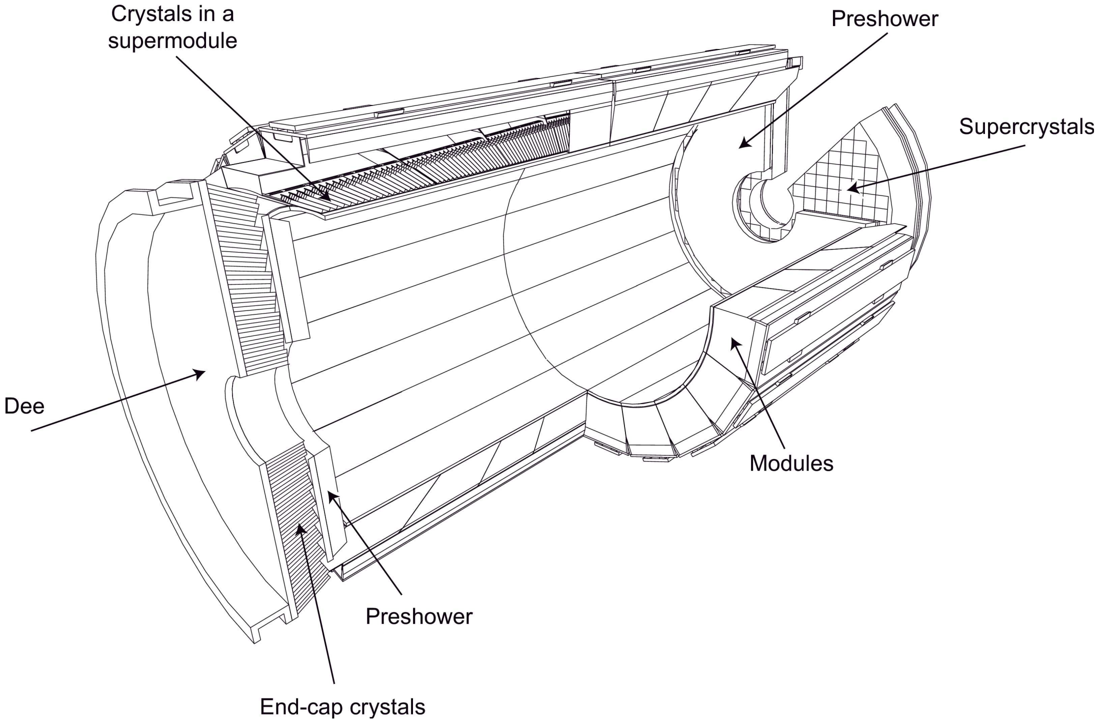


Figure 2.12: Electromagnetic calorimeter design.

the energy and thus reach an excellent resolution in energy while limiting the fluctuations outside the active zone. The cristals are quasi-projective, pointing towards the interaction point.

One of the main caracteristics of the electromagnetic calorimeter is the energy resolution which will, together with the angular resolution, drive the sensitivity to resonances decaying to either two photons or two electrons. The energy resolution can be parametrized as:

$$\left(\frac{\sigma}{E}\right) = \left(\frac{S}{\sqrt{E}}\right) \oplus \left(\frac{N}{E}\right) \oplus C \tag{2.3}$$

where $S = 0.027$ GeV$^{1/2}$ is the stochastic term, $N = 0.12$ GeV the noise term and $C = 0.005$ the constant term in the barrel. The parameters were measured in data during Run 1 and were found compatible with the ones measured during the beam tests performed before the beginning of the data taking.

One of the drawback of the chosen technology is the variation of the crystal transparency which requires a frequent monitoring with a laser system.

The performances of the electromagnetic calorimeter rely on the ability to properly calibrate the detector and to monitor its response during the data taking. The calibration

resolution.

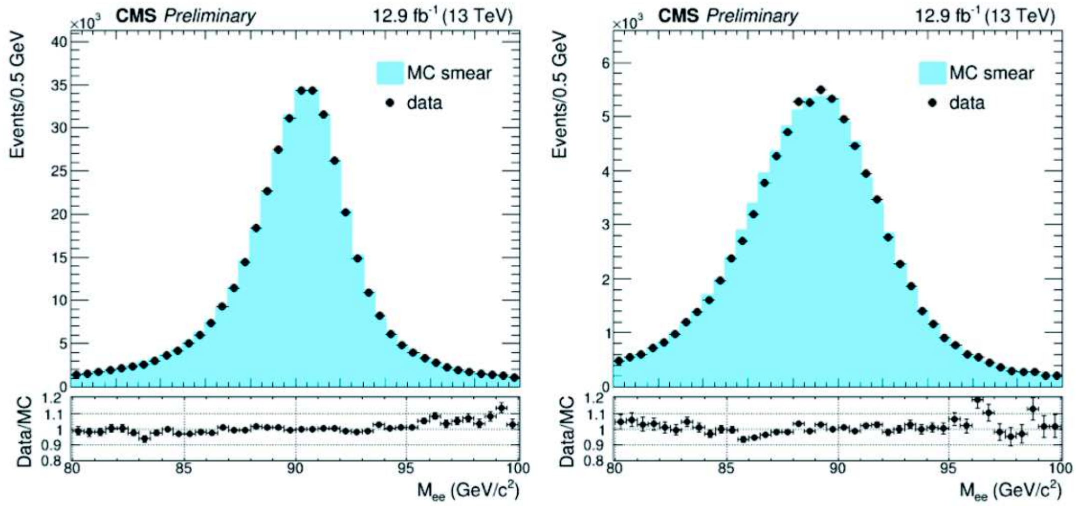


Figure 2.13: Invariant mass of Z → ee decay candidates in barrel-barrel (left) and endcap-endcap (right).

is performed using the expected ϕ-symmetry of the detector in bins of η, with photons from $\pi^0 \to \gamma\gamma$ decays and electrons from W and Z decays (Figure 2.13).

Hadron calorimeter

The hadron calorimeter measures the hadron and jet energy. Its size, limited by the space inside the solenoid, led to the choice of a heterogeneous calorimeter composed of dense layers of copper leading to the interaction of particles with matter and layers allowing the measurement of energy. This configuration does not allow an energy resolution as good as the one of the electromagnetic calorimeter, the total energy being derived from the part lost in the sensitive part of the detector.

The hadron calorimeter is divided in three main parts, the barrel (HB) covering pseudo-rapidities up to 1.4, the endcap (HE) from 1.4 to 3 and the forward (HF) covering pseudo-rapidities from 3 to 5. The barrel part, being placed between the electromagnetic calorimeter and the solenoid, is only one meter thick, corresponding to 5.8 interaction lengths. A second calorimeter H0 is placed outside of the solenoid and allows to measure the remaining energy up to 11 interaction lengths.

The energy resolution is parametrized as:

$$\left(\frac{\sigma}{E}\right) = \left(\frac{S}{\sqrt{E}}\right) \oplus C \tag{2.4}$$

with S = 0.085 GeV$^{1/2}$ and C = 0.074 in the barrel.

The performances of the hadronic calorimeter are related to the calibration of the

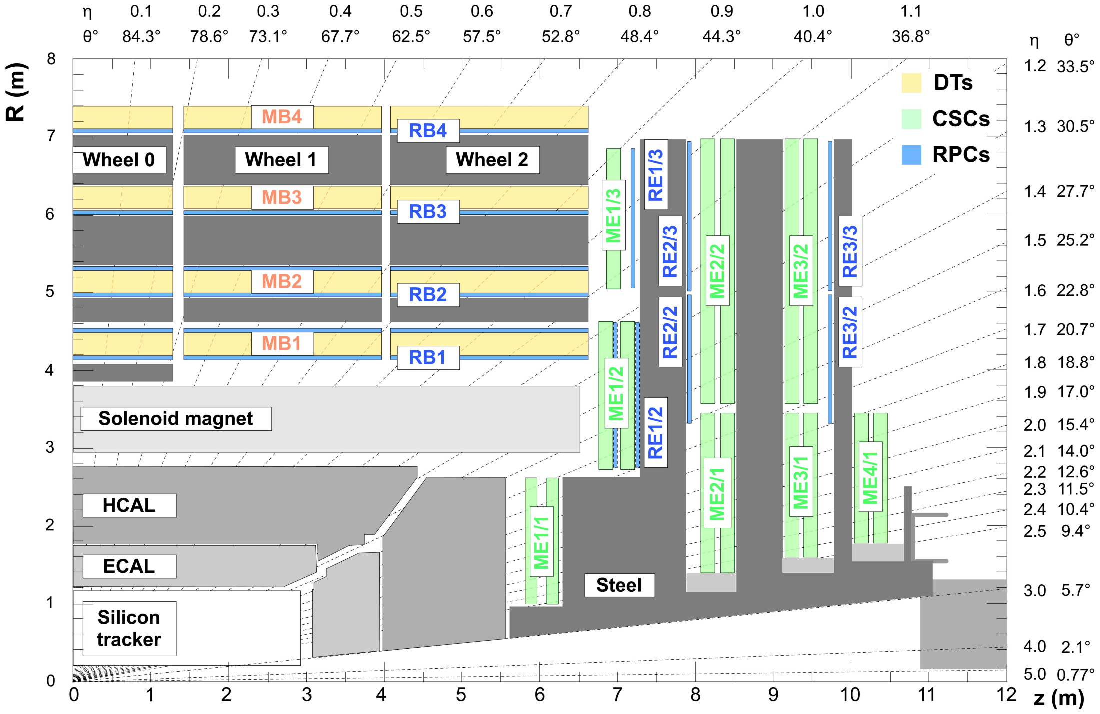


Figure 2.14: CMS Muon system.

energy response of HB, HE, HF and H0. The calibration of HB and HE is done using collision data with isolated charged hadrons. The momentum of the isolated charged hadron is measured in the tracker while its energy is measured in HCAL from all cells in a cone around the impact point of the track. The calibration is then performed in bins of η ring and depth segments.The calibration of HF is done using the decay of a Z boson to an electron-positron pair and the calibration of H0 is done using muons. The calibration of the HF response is done selecting one electron detected in ECAL allowing for a precise measurement of its energy and a second electron candidate detected in HF. By reconstructing the invariant mass of the di-electron system and comparing the position of the peak in data and simulation, the energy scale can be derived with a precision of the order of 3%.

2.2.6 Muon system

The muon system is installed outside the magnet. By combining this information with that from the tracker it is possible to obtain an excellent resolution on the measurement of the momentum of these particles.

Three technologies are used, based on gaseous detectors, covering different parts of the cylinder (Figure 2.14).

In the barrel, drift tubes are used in an arrangement of 5 wheels within $0 < |\eta| < 1.2$. The position in r-ϕ and r-z of the muon is measured from the time needed for the signal to travel in the tube. In the endcaps, cathode strip chambers (CSC) are used to reconstruct muons within $1{,}2 < |\eta| < 2{,}4$ in a higher radiation environnement. An array of positive wires perpendicular to negative strips placed in a gas chamber allows to determine the position in r-ϕ of the incoming muons. Because the CSCs provide a fast response, they are used in the trigger process to register events containing a muon candidate. In both the barrel and the endcaps, resistive plate chambers (RPC) are used in addition to the two previous technologies. Composed of parallel plates of opposite charges separated by a gas volume, they register the passing of a muon through the ionization of the gas which will create a delayed signal in external strips, allowing for a position measurement in r-ϕ with a good spatial resolution. They also provide a fast response with a time resolution of the order of the nanosecond which can be used for trigger purposes.

The performance of the muon detector depends on the quality of the alignement of the muon sub-detectors. In 2016, the non-zero Alignement Position Errors is used [2], leading to an improvement of the performances during the first steps of the data-taking.

[2] The non-zero Alignement Position Error accounts for the residual uncertainties on the position of the muon chambers and allows to increase the efficiency of the track reconstruction when the alignement is applied.

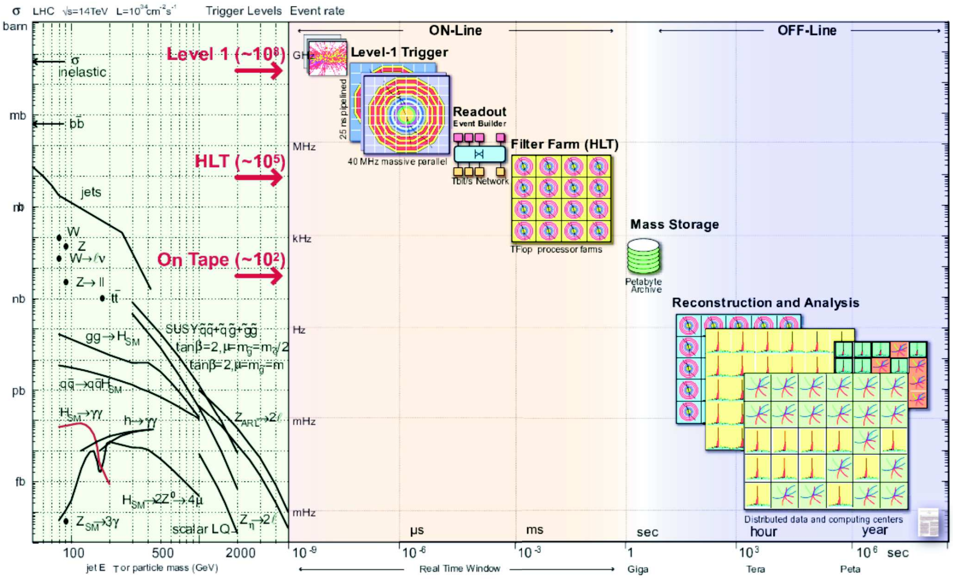


Figure 2.15: From online to offline.

2.3 CMS trigger system

In order to select events useful for physics analyses among the millions produced every second, a two level trigger has been developed in CMS. The Level-1 allows a rate reduction from 40 MHz (rate at which the bunches are crossing) to 100 kHz. The High Level Trigger allows a further reduction down to about 1 kHz, the maximal rate at which the data can be processed and stored (Figure 2.15).

2.3.1 Level-1 Trigger

The Level-1 (L1) trigger is based on electronic cards developed for CMS in order to perform dedicated calculations. At this level, a decision has to be taken for every bunch crossing in order to select any potentially interesting event. The time between two bunch crossings being too short to allow for a decision to be taken, the data can be stored during 3.2 μs, the time for the decision to keep or reject the event to be taken. The need for the whole L1 chain to operate at a rate of 40 MHz forbids the use of iterative algorithms. For the same reason, complex calculations can not be done and the use of Look-Up Tables (LUT) stored in memory allows the results to be associated to a specific set of inputs. At this step, only the calorimeters and muon detectors can be used, running the tracking taking too long to stay in the time constraint. The main objects created are electron/photon candidates which share the same signature in the electromagnetic calorimeter, tau leptons that can be reconstructed through the study of patterns in the

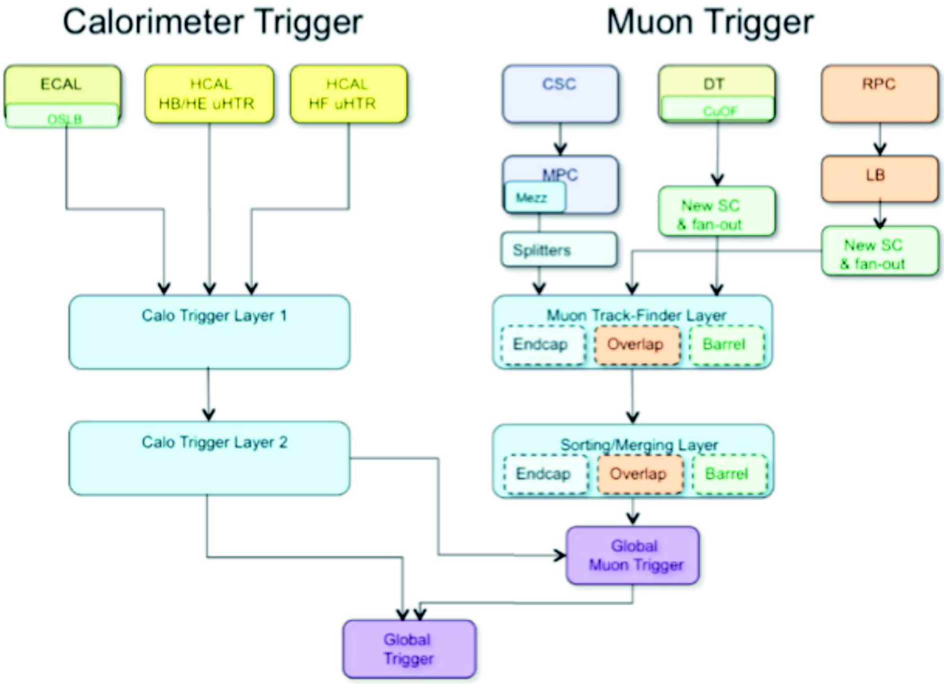


Figure 2.16: L1 architecture during Run 2.

calorimeters, jets that are reconstructed from energy deposits in both the electromagnetic and hadronic calorimeters, and muon candidates that are observed based on the response of the DT and CSC subdetectors (Figure 2.16). A decision is taken after reconstruction of the event based on a set of conditions called the L1 menu. Few hundreds conditions such as the presence of two muons with a momentum above few tens of GeV or the presence of a high jet multiplicity can lead the event to be accepted, the constraint on the number of events being accepted coming from the ability for the next step to process all these events.

2.3.2 High Level Trigger

The High Level Trigger is using the seeds provided by the L1 and runs a more precise reconstruction of the event. The input rate of the HLT (output rate of the L1) is constrained by the average time needed to read and process the events by the processor farm. Because the HLT input rate (of the order of 100 kHz) is much lower than the L1 input rate, the timescale accessible to the High Level Trigger allows it to use all the informations provided by the CMS detector to take the decision to keep the event (Figure 2.17). Objects of increasing complexity are sequentially reconstructed and filtered. Energy deposits are now matched to tracks to sort the candidates between charged and neutral. The information from the tracker being accessible, new objects can be reconstructed such

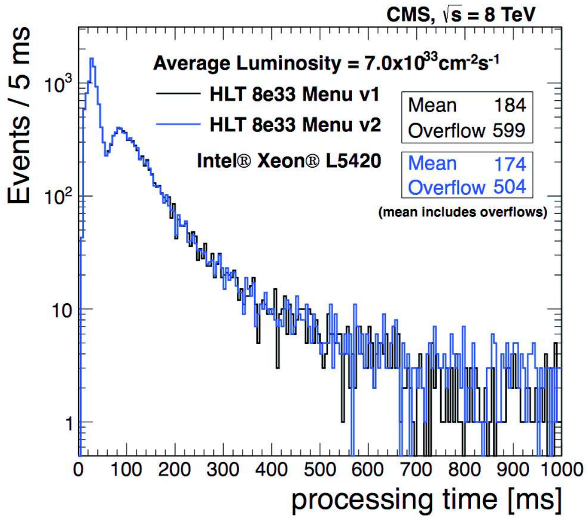


Figure 2.17: Processing time of the HLT paths during Run 1.

as jets originating from b quarks which have a displaced vertex. A new set of conditions will lead to the final decision to store the event, based on the HLT menu defined to share between analyses the final rate of 1 kHz which can be recorded.

2.4 Upgrades and challenges

In order to maintain a rich physics program in everchanging data taking conditions, upgrades of the detector are regularly done. The two main upgrades during this thesis work were done during the Long Shutdown 1 (LS1) between 2013 and 2015 and the End of the Year Extended Technical Stop which lasted from November 2016 to May 2017.

2.4.1 Upgrades during LS1

After the first period of data taking which led to the Higgs boson discovery in 2012, the LHC and CMS had undergone a two years technical stop.

An important part of the work has been to repair sub-detectors to make them ready for three more years of data-taking. The tracker, ECAL, DT, RPC had undergone such a revision.

The L1 trigger architecture was modified to allow for some more complex calculation to be made already at electronic level, such as invariant mass calculation which might begin to be used in 2017.

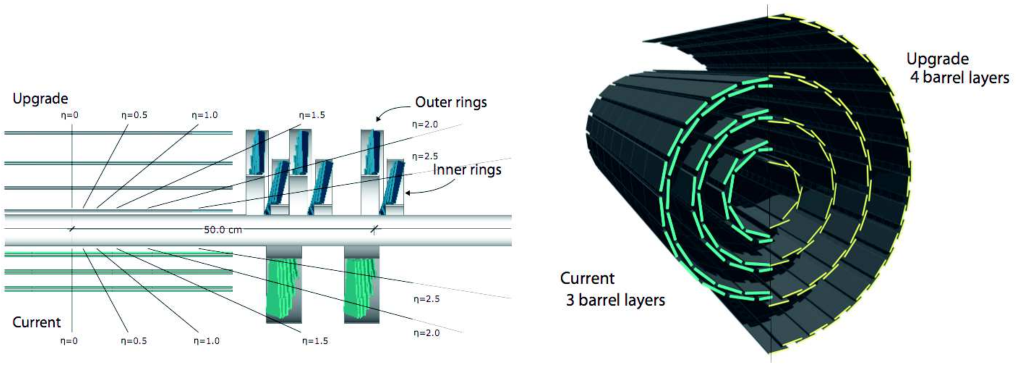


Figure 2.18: Left: Conceptual layout comparing the different layers and disks in the current (2016) and upgraded (2017) pixel detectors. Right: Transverse-oblique view comparing the pixel barrel layers for the two detectors.

2.4.2 Upgrades during EYETS

Between 2016 and 2017, the annual Winter stop of the LHC was extended to allow for some upgrades of the detector. The main change was at the heart of the CMS detector with the three pixel layer tracker being replaced by a four pixel layer tracker (Figure 2.18). The first layer moved closer to the beam with a distance reduced from 4.4 cm to 3.0 cm.

The main goal of this upgrade is to mitigate:

- data loss at high occupancy,
- lower tracking efficiency or higher fake rates at high pileup,
- degradation of performances due to radiation damage,
- degradation of performances due to material budget.

While the hit efficiency was decreasing from 99% at 4×10^{33} cm^{-2}.s^{-1} to 94% at 1.5×10^{34} cm^{-2}.s^{-1} in the first layer, the hit efficiency is found to be above 99% at the highest instantaneous luminosity recorded in 2016, of the order of 1.5×10^{34} cm^{-2} $\times$ s^{-1}.

Passing from three to four layers could have increased the material budget in the pixel detector, leading to a loss of performances through Bremsstrahlung, conversions and nuclear interactions. The architecture of the upgraded detector has been thought to relocate the passive material out of the tracking volume, leading to a lower material budget, both per layer and overall (Figure 2.19).

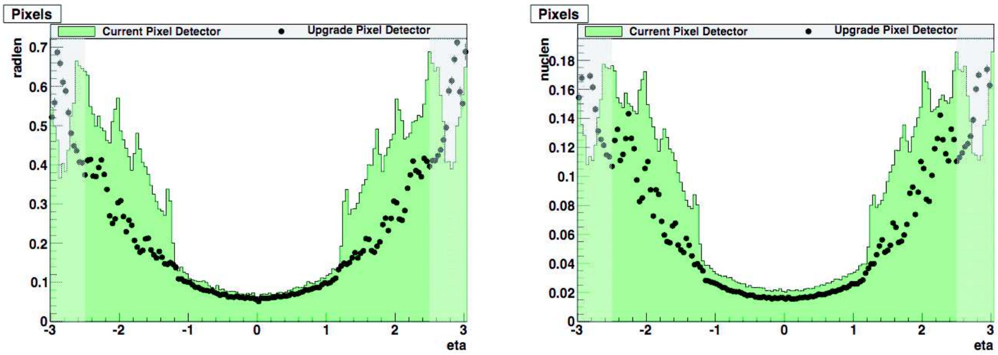


Figure 2.19: The amount of material in the pixel detector shown in units of radiation length (left), and in units of nuclear interaction length (right) as a function of η; this is given for the current pixel detector (green histogram), and the Phase 1 upgrade detector (black points). The shaded region at high $|\eta|$ is outside the region for track reconstruction.

Chapter 3

Object reconstruction and identification

Contents

Events are reconstructed using the Particle Flow (PF) algorithm which aims for an optimal combination of the information coming from all the subdetectors [28]. The goal of this algorithm is to identify all the stable particles (electrons, muons, photons, charged and neutral hadrons) and to recluster hadrons into jets.

Taking benefit of the experience of many detectors running at previous colliders, the CMS detector has been designed as a cylindrical detector placed in a strong magnetic field and composed of successive layers around the beam pipe, beginning with the tracker, followed by the electromagnetic and hadronic calorimeters and the muon system. Most of the objects reconstructed inside such a detector can be matched to a particular subdetector. The jets are reconstructed based on the energy they deposit in the calorimeters. The photons and electrons are identified by the shape of the energy deposit in the electromagnetic calorimeter. The tagging of τ lepton and b quark jets is achieved based on the tracker informations. Finally, the muons can be reconstructed with the standalone information from the muon detector (Figure 3.1).

However, the combination of information coming from all the subdetectors allows for a better reconstruction of single particles and leads to better performances in the determination of the energy and momentum of a particle, the tracker giving a better resolution on the momentum for charged particles at low energy while the calorimeters will provide a better energy measurement for particles at higher energy. Such a combination of subdetector informations allows as well for a cross-calibration of subdetectors and a validation of their measurements.

3.1 Particle Flow

First used in the ALEPH experiment at LEP [29], the Particle Flow algorithm is based on few key elements allowing a combination of subdetector informations. While perfectly suited to be used in the clean environnement of an e^+e^- collider, its adaptation to the hadronic environment of a proton-proton collider leads to new challenges.

The key components needed to deploy a Particle Flow algorithm can be listed as follows. A large magnetic field will separate the neutral and charged contributions in the calorimeter. A fine-grained tracker will allow the momentum measurement of the charged component (70%) of the jets. A highly-segmented electromagnetic calorimeter will allow to match the energy deposit to the track of charged particles. A hermetic hadronic calorimeter will allow to complete the measurement of the energy of neutral hadrons. A good muon tracking system will provide an unambiguous muon identification.

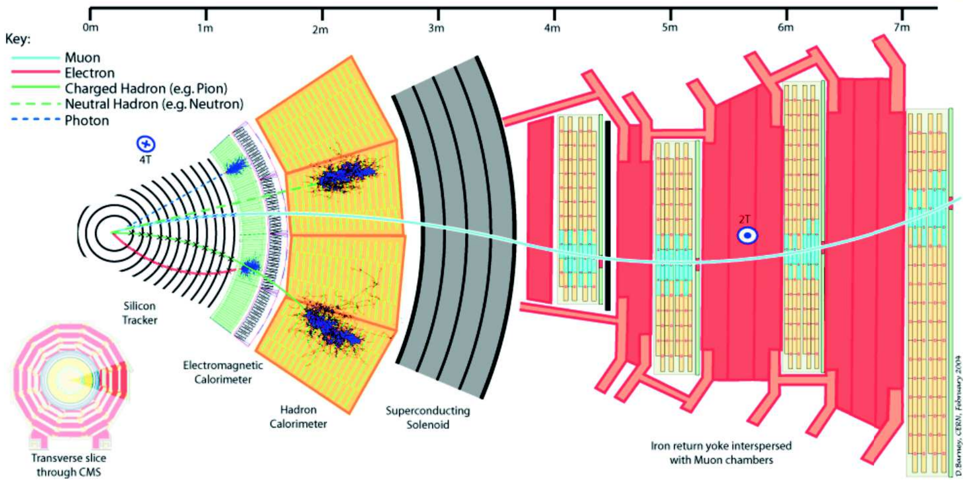

0m
1m
2m
3m
4m
5m
6m
7m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
4T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
Transverse slice through CMS
D.Barney, CERN, February 2004

Figure 3.1: Particle interactions in a transverse slice of the CMS detector.

3.2 Building blocks: tracks and clusters

The two main building blocks of the PF algorithm, which will be linked together and allow for the reconstuction of individual particles, are tracks and clusters.

Tracks are reconstructed based on the three layers of silicon pixels and the ten layers of silicon strips. A combinatorial track finder based on the requirement of a seed of two hits in consecutives pixel layers, followed by the requirement of a reconstructed track of at least eight hits with p_T above 0.9 GeV originating from few mm around the beam axis was first proposed. While this algorithm allowed to keep a low misreconstruction rate, it led to a loss of efficiency, most of the tracks failing to pass the requirement on the number of hits. Another strategy was therefore adopted, based on successive track reconstruction steps, allowing to recover most of the efficiency while keeping a low misreconstruction rate. The steps target specific sets of tracks (high p_T, displaced, low p_T, muon tracks) and the quality criteria can be loosened after the removal of the hits corresponding to reconstructed tracks in order to keep the combinatorial and the fake rate under control. The three first steps target high quality tracks: prompt high p_T tracks, from b hadron decay and prompt at low p_T respectively. Once reconstructed, the quality criterion wich required three pixel hits can be loosened to two for the next steps, allowing to recover high p_T tracks with one missing hit in the pixel detector, or displaced tracks. The following steps are targeting very displaced tracks, and at last muon tracks by using the muon information in the seeding step. This procedure was adapted to the new pixel detector, triplets in the seeding step replaced by quadruplets. A recovery procedure for missing hits in the pixel detector has been developed to cope with some inefficiencies in the pixel layers.

Although for electrons, the main identification comes from the ECAL cluster information and is used to infer the expected position of the corresponding hits in the tracker, this approach leads to inefficiencies when the energy of the electron gathered in a supercluster [1] doesn't span over a large enough range. In this case, the expected position of the associated hits might be misreconstructed, leading to a mismatch and the wrong identification of the electron as a photon. Because of this, another complementary tracking is done for electron, seeding the matching between cluster and track from the tracker side. To do so, all the tracks from the iterative tracking with p_T above 2 GeV are used as potential seeds for electrons.

Muons can be reconstructed from the muon subdetectors alone (DT, CSC, RPC), leading to a collection of muon candidates called *standalone muons*. These muons are propagated to the tracker and added to the collection of *global muons* if a matching can be done with an inner track. Finally, inner tracks with p_T above 5 GeV are extrapolated to the muon system and produce a *tracker muon* candidate if they are matched to at least one muon segment, muon segments being short tracks made of DT and CSC hits. The combination of these muon collections guarantees a reconstruction efficiency above 99% while keeping the misreconstruction rate under control. The main contribution to the muon misidentification comes from hadron shower remnants escaping the hadronic calorimeter and reaching the muon system (*punch-through*).

Calorimeter clusters are seeded by a local maximum of energy deposit, extended to neighbouring cells with a larger extension in ϕ to recover for bremsstrahlung.

After reconstruction of tracks and clusters, these elements are grouped together by a link algorithm based on their relative geometrical compatibility. Tracks are propagated to ECAL clusters and matched if they satisfy criteria based on their distance in the η/ϕ plane. ECAL and HCAL clusters are matched if they share the same η/ϕ position and the closest elements from each other are chosen to be linked in case they are shared with several others.

Muons are reconstructed first and elements associated to the muons candidates are removed. Electrons and photons are then identified, leading to the removal of the corresponding tracks and electromagnetic deposits in ECAL. Charged and neutral hadrons are finally identified before a last post-processing of the event. This post-processing looks for objects contributing to more than half of the missing transverse momentum and re-run the identification when such an object is found. When it is the case, the object found is often misidentified (hadron *punch-through* identified as muon and leading to a large miss-

[1] superclusters are clusters merged together based on a cluster used as a seed and the addition of nearby cluster in a narrow η but wider ϕ window.

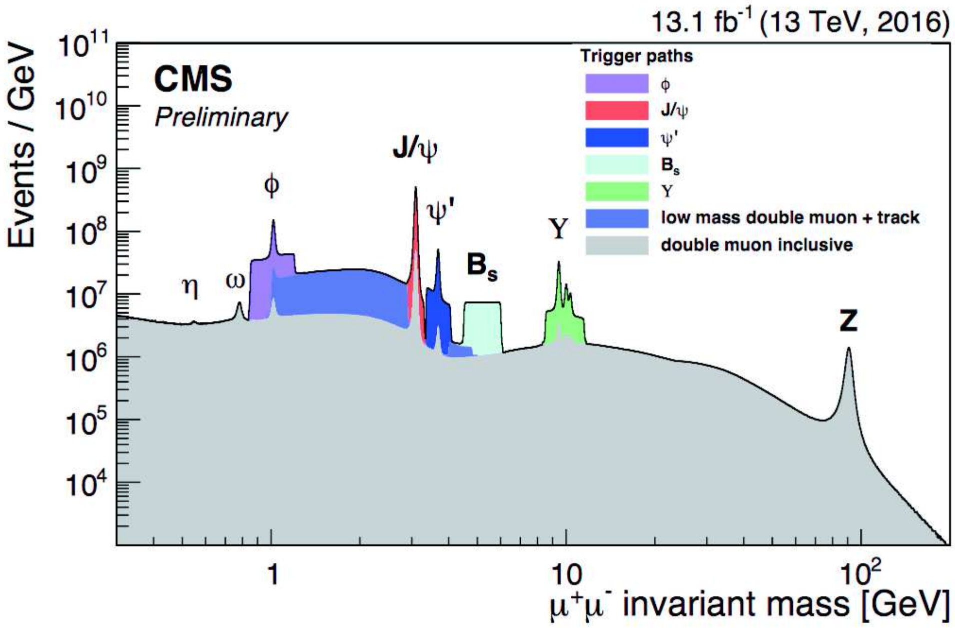


Figure 3.2: Dimuon invariant mass spectrum.

ing transverse momentum, the neutral component of the jet being lost when identified as a muon).

3.3 Muons

A relevant illustration of the range over which muons are reconstructed is the dimuon invariant mass (Figure 3.2).

The efficiency of identifying a muon is studied using the Tag-and-Probe method at the Z mass (Figure 3.3). Tag-And-Probe uses dilepton events, requiring one to pass the tight identification criterion (Tag), the other lepton (Probe) being the one from which the efficiency is derived. The tight muon has to be a PF muon fulfilling selection cuts requiring at least one pixel hit, five hits in the tracker and at least one hit in a muon chamber included in the global muon track fit. Moreover, the normalised global muon track chi-square is required to be below 10 and transverse impact parameter of the track in the tracker to be lower than 2 mm with respect tos the primary vertex.

3.4 Photons and electrons

The main information allowing for the reconstruction of electron and photon is the similar pattern they produce in the electromagnetic calorimeter. The distinction between electron

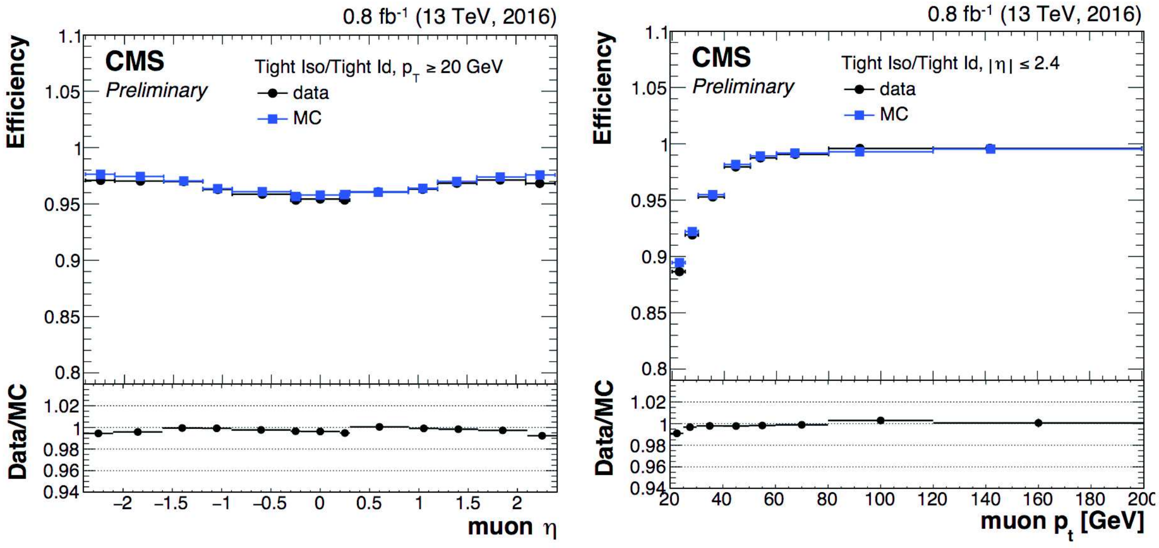


Figure 3.3: Muon efficiency with respect to η (left) and p_T (right) with early 2016 data.

and photon is done by the matching between a track and the energy cluster for electrons. The matching is done in two ways, a first one from the calorimeter to the tracker, the second one from the tracker to the calorimeter.

Electrons

Figure 3.4 shows the agreement between data and simulation for the p_T, η, ϕ and MVA distributions related to electrons [30]. The MVA distribution corresponds to the value associated to the electron based on a multivariate technique, aiming at discriminating between genuine electrons and fake electrons. Among the main variables used as input to the MVA are the shower shape and the isolation.

The efficiency is studied by Tag-And-Probe at the Z peak. The efficiency based on a multi-variate discriminant shows better results than the one based on a serie of cuts aiming at identifying electron while rejecting fakes (Figure 3.5).

Photons

Figure 3.6 shows the agreement between data and simulation for the p_T, η, and MVA distributions related to photons with the first tens of fb^{-1} from 2016.

The efficiency is again better when using MVA techniques to identify photons (Figure 3.7).

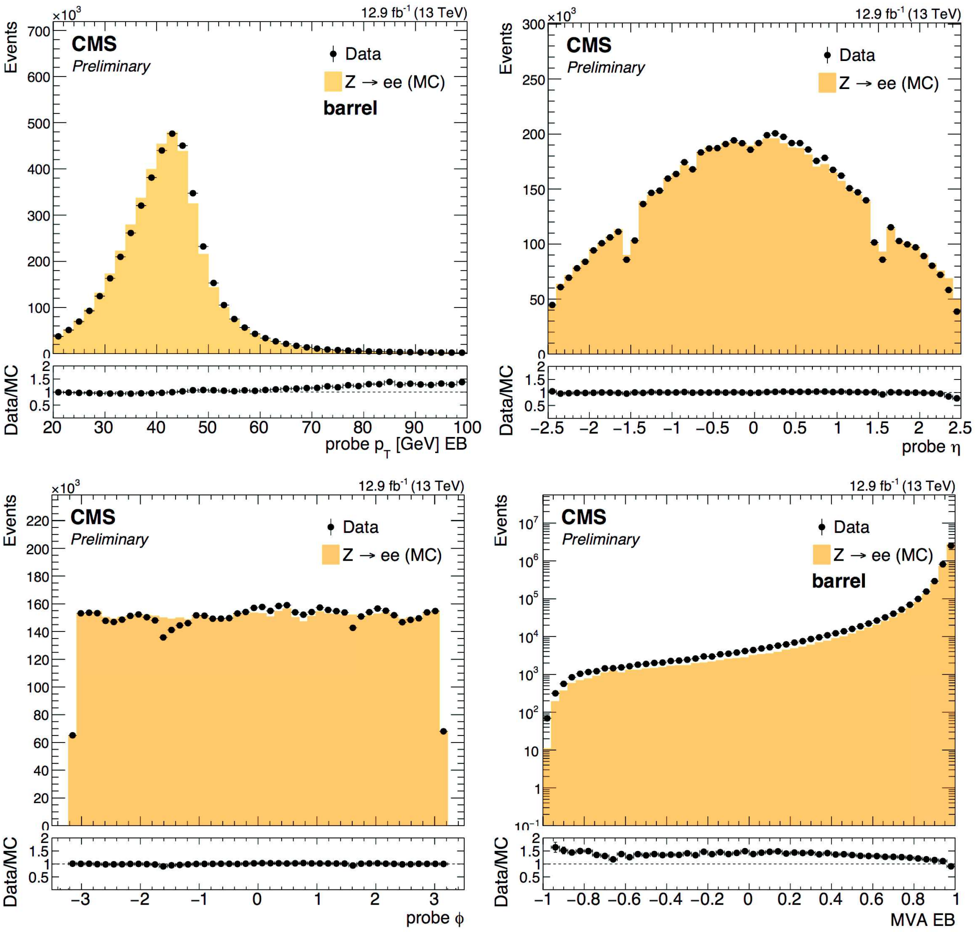


Figure 3.4: Electron p_T (top left), η (top right), ϕ (bottom left) and MVA (bottom right).

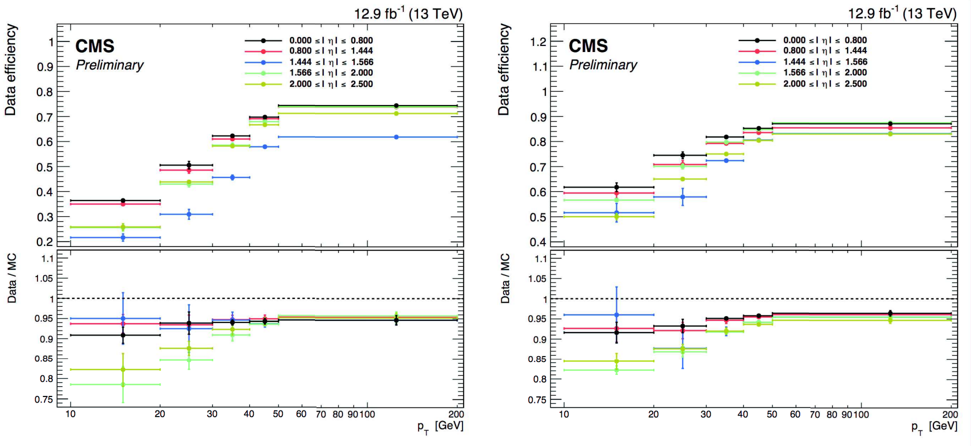


Figure 3.5: Electron efficiency for cut-based (left) and MVA-based (right) identification.

3.5 Taus

Tau leptons can decay either leptonically to an electron or muon and two neutrinos or hadronically, leading to final states with few hadrons and a neutrino (Table 3.1). The

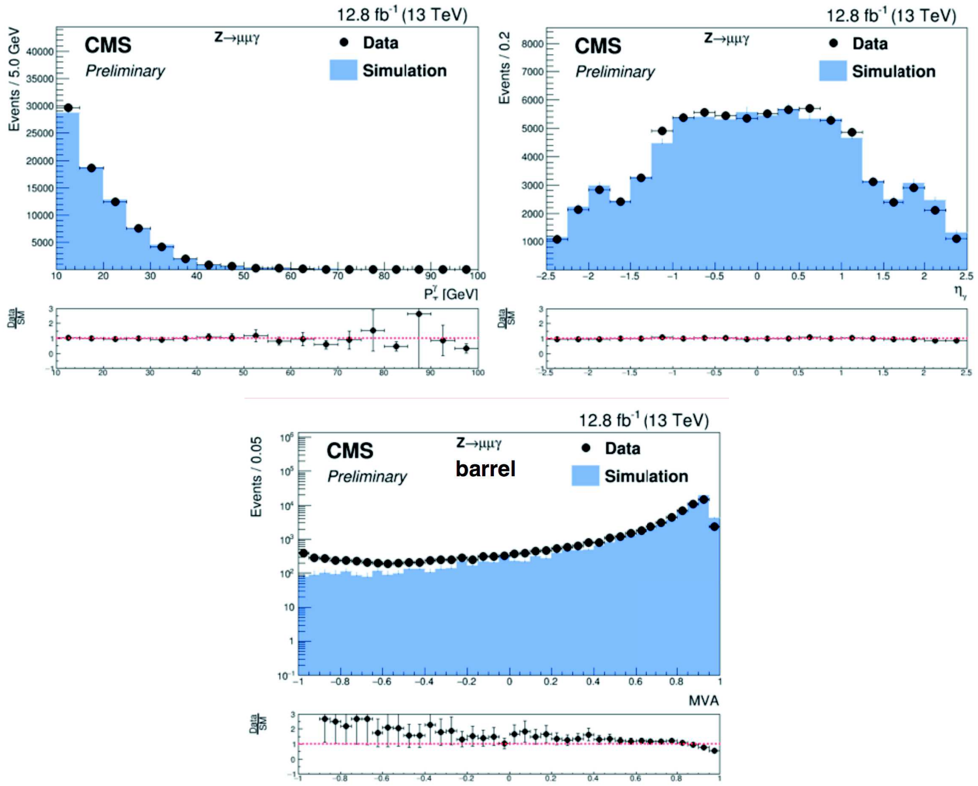


Figure 3.6: Photon p_T (top left), η (top right) and MVA (bottom).

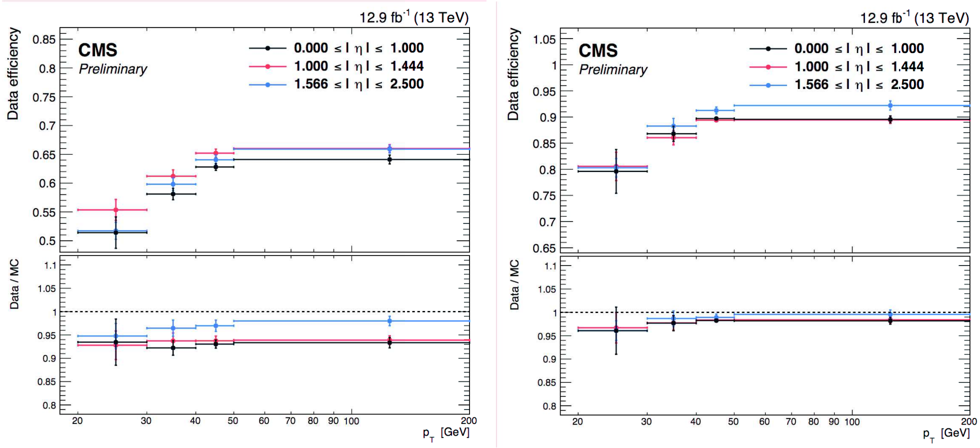


Figure 3.7: Photon efficiency of tight working cut-based (left) and MVA (right) identification.

hadrons are composed of a combination of charged and neutral mesons (π and K). Several observables can be used to identify jets produced by tau decays, such as the multiplicity of tracks, the region over which the energy is deposited and the isolation of decay products.

Taus decaying hadronically are reconstructed by the hadron-plus-strips algorithm [32, 33] which targets a reconstruction of the decay mode (1 or 3 prongs and π^0's). Charged hadrons are reconstructed using the PF algorithm while π^0 are reconstructed as strips in the ECAL. The geometry of the strip chosen for the reconstruction of the π^0's is related to the photon conversion in the tracker material which will lead to a larger extension in ϕ due to the electron track bending in the magnetic field. The mass reconstructed from the strip is required to be consistent with the π^0 mass. The hadronic decay mode is then reconstructed based on the number of PF candidates for the charged hadrons (either 1 or 3) and the number of strips. Taus are required to have p_T above 20 GeV and $|\eta| < 2.3$. While the hadron-plus-strip algorithm provided good results during Run 1, it has been updated during Run 2 [34]. The strip reconstruction algorithm has been modified to account for the electromagnetic leakage of the τ_h decay. Two multivariate analyses were introduced to suppress the misidentification of jets and electrons as τ_h.

Decay mode	Meson resonance	B [%]
$\tau^- \to e^-\bar{\nu}_e\nu_\tau$		17.8
$\tau^- \to \mu^-\bar{\nu}_\mu\nu_\tau$		17.4
$\tau^- \to h^-\nu_\tau$		11.5
$\tau^- \to h^-\pi^0\nu_\tau$	$\rho(770)$	26.0
$\tau^- \to h^-\pi^0\pi^0\nu_\tau$	a1(1260)	10.8
$\tau^- \to h^-h^+h^-\nu_\tau$	a1(1260)	9.8
$\tau^- \to h^-h^+h^-\pi^0\nu_\tau$		4.8
Other modes with hadrons		1.8
All modes containing hadrons		64.8

Table 3.1: Branching fraction of the main τ decay modes [31].

The performance of the tau identification was studied with the first data from 2016 looking at the decay $Z \to \tau_\mu\tau_h$ [34]. The visible mass is in this case the invariant mass between the muon and the τ_h candidates.

3.6 Jets

Jets are built by clustering candidates in a cone in order to reconstruct the kinematics of the parton from which they originated. Two kinds of clustering can be performed depending on the information used. *Calo jets* are built from energy deposits in the calorimeter while *PF jets* are built using the PF candidates. Because using PF candidates lead to a better efficiency, *Calo jets* are only used in the online reconstruction in order to reduce the computing time by providing a fast, raw information about the energy and direction of the jet. The $anti - k_T$ [36] algorithm provides a infrared safe and collinear safe way of clustering PF candidates.

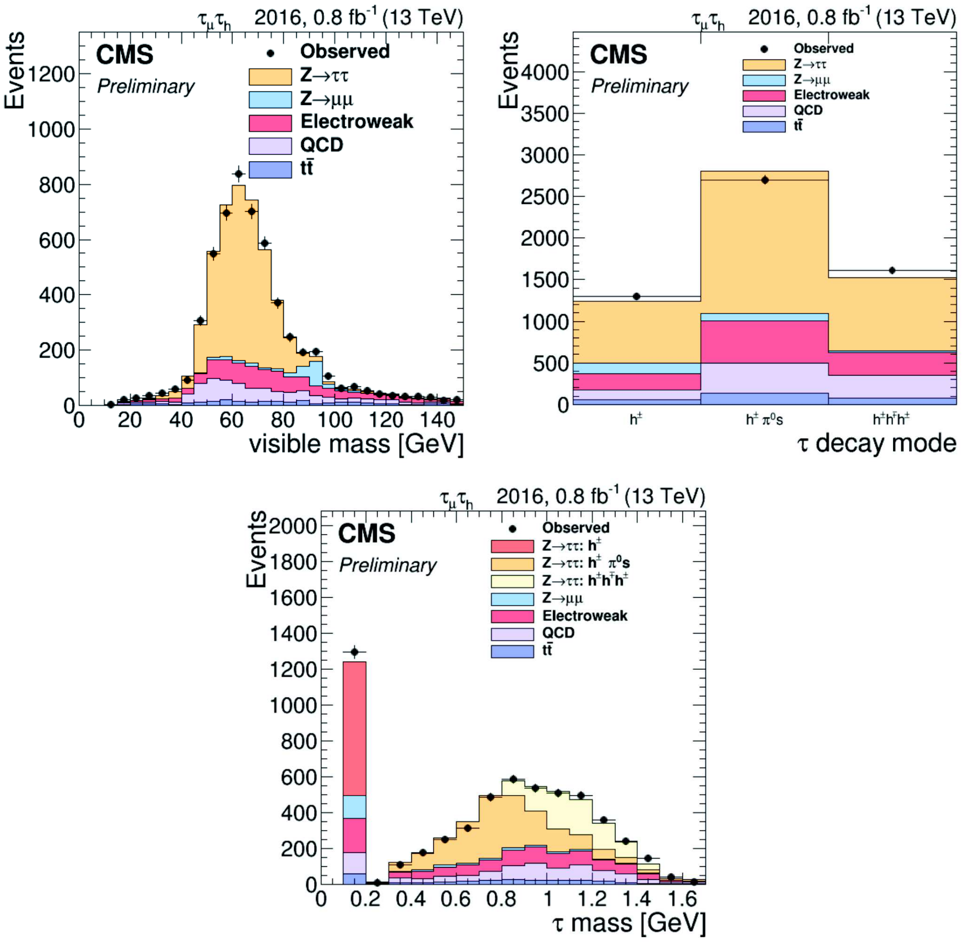


Figure 3.8: Tau pair visible mass, tau decay modes and mass.

The Charged Hadron Subtraction (CHS) limits the contribution from charged particles coming from pileup by removing from the clustering procedure charged hadrons which are not associated to the reconstructed primary vertex.

The goal of jet identification is to discriminate between physical jets and instrumental noise in the calorimeters. The efficiency to reconstruct a PF jet in a p_T range from 30 to 2.5 TeV is above 98% (Figure 3.9).

Standard jets are reconstructed with a ΔR = 0.4 anti-kt parameter (AK4 jets) which was shown sufficient to recover most of the jet constituents while decreasing the contribution from pile-up jets with respect to the ΔR = 0.5 parameter which was used during Run 1.

Jet reconstruction and energy measurement suffer from the increase in pileup leading to an increasing number of objects and an offset of energy. The jet energy is therefore corrected via a hybrid jet area method using the effective area of the jet and the average energy density in the event in order to calculate the value of the offset to be subtracted

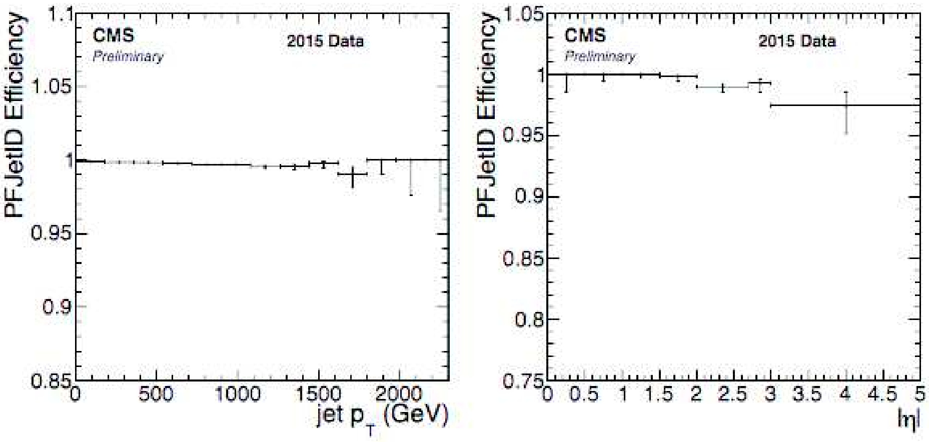


Figure 3.9: Tight PF jet identification efficiency as a function of p_T for central jets ($|\eta|$<0.5) (left) and as a function of $|\eta|$ for 30 < p_T < 100 GeV with the first data at 13 TeV [38].

from the energy of the jet. A second set of correction is applied to the jet based on the expected difference between the generated and reconstructed jet. A last correction is applied based on the difference between data and simulation.

The jet energy resolution can be extracted from the p_T asymmetry in dijet events and the p_T balancing in Z+jets events. The resolution is below 10% for jets with p_T above 100 GeV and better than 5% for jets with p_T above 1 TeV.

Fat jets, i.e. jets originating from the hadronization of quarks coming from boosted objects, are reconstructed in a cone of ΔR = 0.8 (AK8 jets), allowing to recover in one jet all the decay products of the two initial quarks.

The future of jets might be in the use of sophisticated machine learning techniques. While most of the current algorithms going in this direction are considering the jet as a picture, trying to identify it through geometrical considerations, recent results using natural language processing and considering jets as words have led to increased performances [39].

3.7 Missing transverse energy and corrections

While most of the objects can be reconstructed based on their interaction with one or several subdetectors, weakly interacting particles such as the neutrinos or particles predicted by new physics models will escape the detector without being detected. They will however contribute to an imbalance in the transverse plane. The missing transverse energy is thus defined as the negative vector sum of the transverse momenta of the PF candidates and

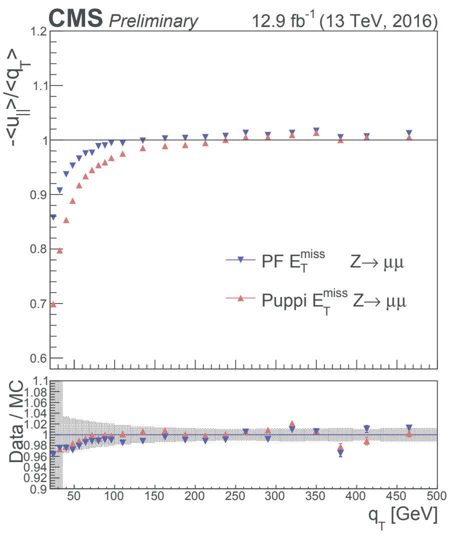


Figure 3.10: Response in Z → $\mu^+\mu^-$ events for the PF E_T^{miss} and Puppi E_T^{miss}.

will account for the presence of such particles.

The standard PF E_T^{miss} is computed using the PF candidates while the pileup per particle identification (PUPPI) [37] computes for each particle a weight based on the local shape information. This weight based on the difference in shapes between the collinear structure of QCD and the diffuse radiation from pileup can be interpreted as pileup likeliness and is used to rescale the 4-momenta of the particles.

The E_T^{miss} response can be measured from a data sample containing Z boson decaying into lepton whose p_T can be measured precisely. The transverse momentum of the Z boson $\vec{q}_T$ can be compared to the transverse momentum $\vec{u}_T$ computed from all the other particles in the event projected on the $\vec{q}_T$ axis (Figure 3.10).

The resolution of the response can be studied with respect to the pileup approximated by the number of vertices (Figure 3.11). As expected, the resolution of the E_T^{miss} decreases with an increasing pileup. Mitigation strategies have been studied in the reconstruction of the E_T^{miss} in order to reduce the contribution of the pileup, through better corrections to the energy and momentum of the different objects.

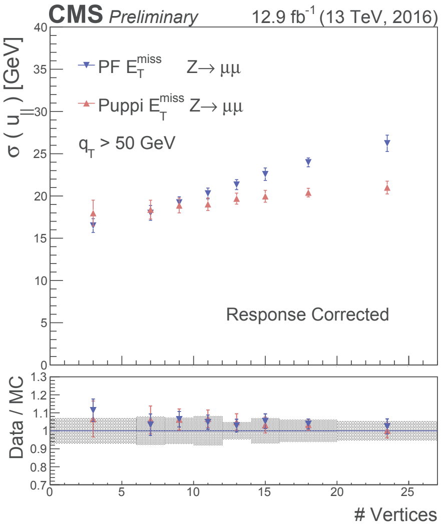


Figure 3.11: Resolution of the response in Z → $\mu^+\mu^-$ events with respect to the number of vertices for PF E_T^{miss} and Puppi E_T^{miss}.

Chapter 4

b-tagging

Contents

In the previous chapter, we saw that particle identification in CMS relies on complex combination of the information coming from all subdetectors. The last object which can be identified inside the CMS detector and which is the topic of this chapter is jet coming from the hadronization of b quarks. Such a b jet is a perfect example of the increasing complexity of the identification and of the interplay between detector upgrades and software (r)evolutions. From a simple variable used for the identification of b jets (namely the impact parameter of the highest p_T track), the b-tagging algorithms moved to the combination of several variables using multivariate techniques such as neural network and deep neural network in the last few months. After introducing the motivation behind the development of the taggers (Section 4.1), we will quickly go through the tracking and vertexing (Section 4.2) before introducing different types of taggers (b-tagger Section 4.3, c-tagger Section 4.4, double b-tagger Section 4.5). The b-tagging efficiency measured in data is compared to the simulation in Section 4.6. Finally, we will focus on the triggers needed for b-tagging performance measurements and on the use of b-tagging at trigger level (Section 4.7).

4.1 Motivation

Discovered at Fermilab in 1977 through the observation of the Upsilon resonance [40], the b quark has been ever since a very useful tool to probe physics at higher energy. In the hadronic environnement of the LHC, the ability to identify jets originating from the hadronization of b quarks is of particular importance to study the hadronic decay of Z and Higgs bosons and in the processes involving a top quark. Such identification of b quark jets is called b-tagging and relies on the properties of the b hadron produced.

The main physical properties of b hadrons are:

- a long lifetime, $\tau = 1.5$ ps and thus a large decay length $\gamma\beta c\tau = 1.8$ mm for a b hadron of 20 GeV,
- a large mass $\simeq 5$ GeV,
- a rather large multiplicity of charged particles from b hadron decay;
- a high probability of leptonic decay ($BR(b \to \mu X) \approx 20\%$).

When translated in terms of observables (Figure 4.1), these properties lead to two main goals in order to properly identify jets originating from b quarks:

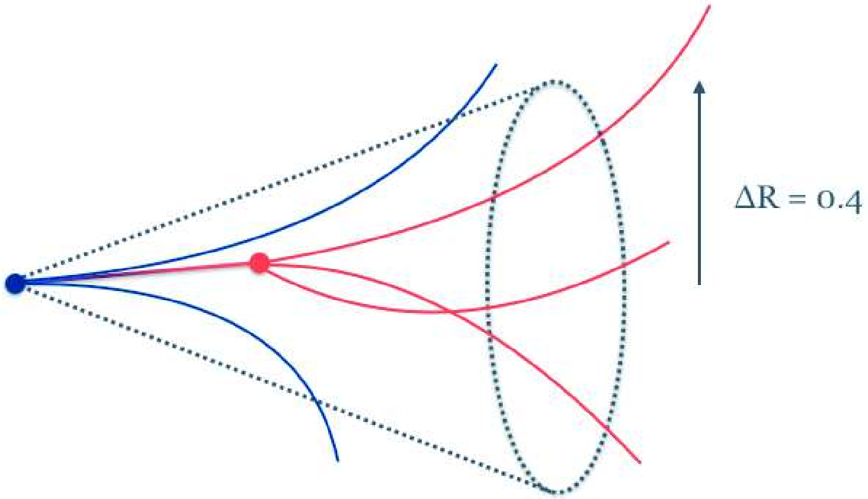


Figure 4.1: Topology of a b-hadron decay.

- good tracking to associate the correct tracks to each vertex and measure their properties,
- good vertexing to measure the position of the first and secondary vertices.

4.2 Tracking and Vertexing

The b jet identification relies on tracking and vertexing. Before moving to the use of the variables associated with tracks and vertex, a short description of the tracking and vertexing is necessary. In the 3.8 T magnetic field of CMS, the curvature of the charged particles allows for a precise measurement of their transverse momentum. Association of tracks will then allow to reconstruct the vertices from which they originated.

Tracking

The tracking has been described in the previous chapter. Some steps of the iterative producedure are specifically targeting the displaced decay of b or c hadrons. During Run 1, the tracks used were expected to have three hits in the pixel layers, but some inefficiency in the first part of the data taking in 2016 led to a relaxation of the constraint to two hits. While the inefficiency in the tracker was solved, the constraint was kept relaxed.

Track selection

In order to select tracks likely to originate from heavy hadron decay, selections are applied to the tracks before considering them in the tagging step. Those global selections have to be valid over a wide range of jet p_T (given that the track density is larger at large p_T) and $|\eta|$ (given that the number of available layers of the silicon tracker is reduced at high pseudorapidity). A first quality cut is applied requiring at least one hit in the pixel layers. Further cuts are applied on the transverse momentum which has to be above 1 GeV and the normalized χ^2 of the trajectory fit which has to be lower than 5. Furthermore, selection cuts are applied on the distance between the track and the primary vertex, between the track and between the jet axis and on the track decay length (distance between the primary vertex and the point at which the track is closest from its jet axis). All these requirements are optimized to keep a good track selection efficiency while reducing the contribution from fake tracks and from pileup.

Some studies were performed in order to improve the track selection by using multivariate techniques. However, defining a figure of merit is not trivial. Track selection follows a set of goals leading to the cuts aforementioned. The two mains goals are to reject tracks from pileup, nuclear interactions, conversions, kaons K^0_S and Λ decays (denoted V^0) and to protect the downstream algorithms against mismeasured tracks and variation in the data taking conditions. Attempt to implement a track selection through the use of multivariate technique did not lead to a clear improvement in the b-tagging efficiency.

Vertexing

The primary vertex reconstruction aims at measuring the position and uncertainties on the position of the vertex. Since the position of the primary vertex is determined quite accurately in the x-y plane, the goal of the reconstruction is to achieve a good z resolution. In order to determine the position of the primary vertex, tracks are sorted with respect to the accuracy of their impact parameter (IP). The impact parameter is the distance from the primary vertex to the track at its point of closest approach. The IP can be computed either in the transverse xy plane or in space. It can be signed according to its angle relative to the jet direction (positive if less than 90°, negative otherwise). After removing tracks failing selection cuts such as the number of hits, a low p_T or a large impact parameter, tracks are sorted based on the z value error. The z value (z coordinate of the trajectory at point closest to the beam axis) of the first track is considered as the seed and each following track is either associated with a vertex or leads to a new seed if its distance to the other seeds is large. If the new track is associated to a vertex, the position of the vertex is recomputed. After all the tracks have been considered, all the vertices are refitted in three dimensions and the procedure is repeated until a stable configuration is found.

The secondary vertex reconstruction during Run 1 was based on an adaptative vertex

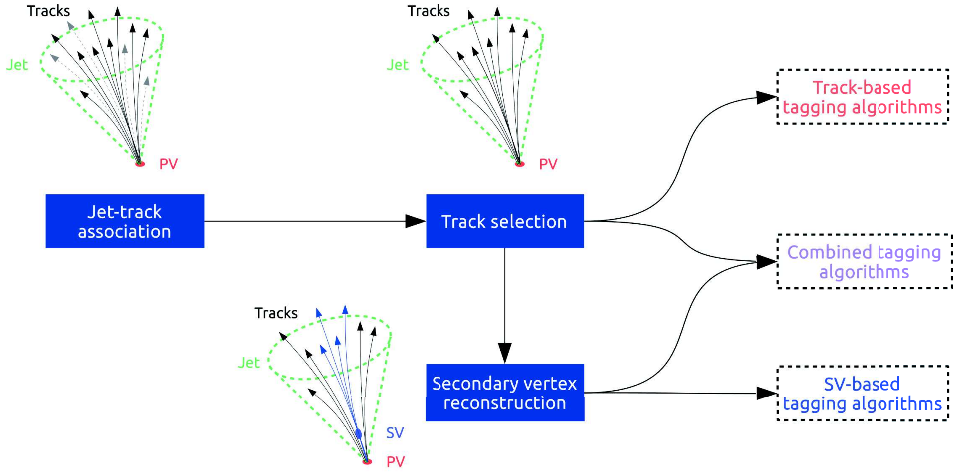

Tracks
Jet
PV
Tracks
Jet
PV
Jet-track association
Track selection
Track-based tagging algorithms
Combined tagging algorithms
Tracks
Jet
SV
PV
Secondary vertex reconstruction
SV-based tagging algorithms

Figure 4.2: b-tagging strategy.

reconstruction (AVR) using a jet of tracks [43]. Selections were applied to the set of reconstructed vertices to remove vertices likely to originate from pileup or from long-lived particles such as K^0.

During Run 2, the default algorithm is the inclusive vertex finder (IVF) which is making use of all tracks in the event, thus being independent from the jet direction and size. From a seed constructed using a track with IP above 50 μm and an IP significance (IP divided by its uncertainty) above 2, a clustering is done by adding further tracks to the secondary vertex. The clustered tracks are then fitted with the adaptative vertex fitter. Reconstructed vertices are filtered based on their distances and flight distance (distance to the primary vertex). The use of IVF algorithm increases the secondary vertex reconstruction efficiency by about 10% with respect to AVR, while increasing the probability to find a secondary vertex for light jet by about 8%. If the improvement provided by this algorithm is non trivial in the case of b vs light jets, the lack of prior jet-track association allows a better determination of the secondary vertices in the case of double-b jets emerging from the decay of a boosted resonance decaying into two b quark jets close to each other [45].

4.3 Taggers

Many algorithms have been developed allowing to identify jets originating from b quarks. All rely on vertices, tracks or a combination of these informations (Figure 4.2).

The Combined Secondary Vertex (CSV) [46] has been the main tagger used for the last years. It combines track information, such as impact parameter, and vertex information, such as the mass of the secondary vertex (invariant mass of particles attached to the secondary vertex), in a neural network. The output of the neural network is a discriminant ranging from 0 for jets likely to come from light jets to 1 for jets likely to have been induced by the hadronization of b quarks. Three vertex categories are defined depending on the presence of a secondary vertex and the set of variables used as input to the tagger is adapted accordingly. The first vertex category corresponds to the optimal configuration for b-tagging. At least one reconstructed secondary vertex is found, allowing to use informations from both tracks and secondary vertex. In the case where several secondary vertices are found, the vertex mass and flight distance significance are taken from the vertex with the lower uncertainty on the flight distance. The second category corresponds to cases where a pseudo-vertex is found, based on at least two tracks with a signed IP significance above 2. Since no vertex fit is performed, the secondary vertex position is not defined and the downstream tagging is using a reduced set of variables. The third vertex category corresponds to cases where no secondary vertex of pseudo-vertex is found but tracks are associated to the jets. In this case, the set of variables used in the tagging is further reduced. In case no tracks are associated to the jet, the CSV discriminant is given a negative value.

The following variables are used as input to the tagger:

- jet p_T , η,
- number of tracks,
- track related variables (significance of the impact parameter, transverse momentum of the track relative to the jet axis, ...),
- vertex related variables (invariant mass of the charged particles, number of tracks, ...),
- number of secondary vertices.

The full list of variables used in each category can be found in Annex.

Figure 4.3 shows the distribution of some input variables of the CSV tagger. The first plot shows the 3D IP significance of the tracks on a $\mathrm{t\bar{t}}$ selection requiring an isolated electron and an isolated muon. The contribution of b jets is therefore important. The second plot shows the secondary vertex mass for jets of different flavours from a multijet sample. Such a sample is dominated by light jets with a lower secondary mass than the one of c or b jets. The third plot shows the distribution of the SV 3D flight distance significance in a muon enriched sample obtained by requesting the presence of a muon

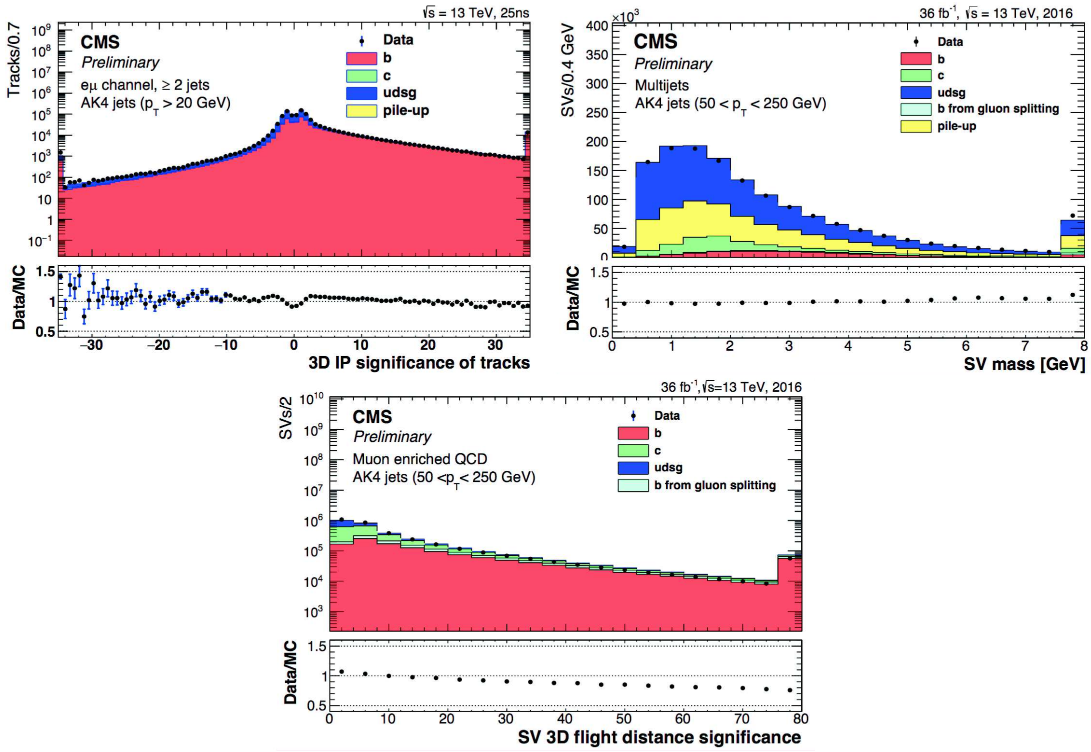


Figure 4.3: Input variables of the CSV tagger.

in a jet. Figure 4.4 shows the distribution of the discriminant in two samples. The first sample is an inclusive multijet sample and a contribution from pileup jet can be seen at low value of the discriminant. The second sample is a muon enriched sample. Because of the high probability for the decay of a b hadron to contain a muon, the sample is enriched in b jets and no pileup jets are present.

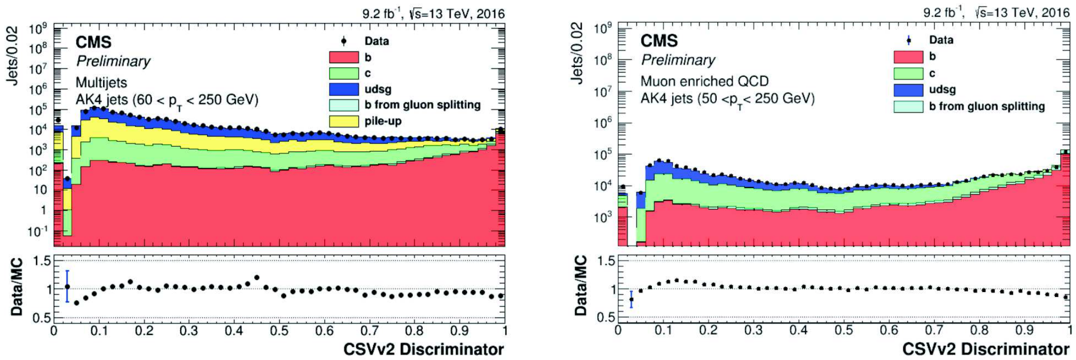


Figure 4.4: CSV distribution in multijet and muon enriched samples.

The figure of merit of a tagger is a ROC (Receiver Operating Characteristic) curve (Figure 4.5) illustrating the performance of the tagger with respect to the chosen value of the threshold applied to the discriminant. For the standard tagger (CSV), three working points, Loose, Medium and Tight are chosen corresponding to 10, 1 and 0.1% mistag rate

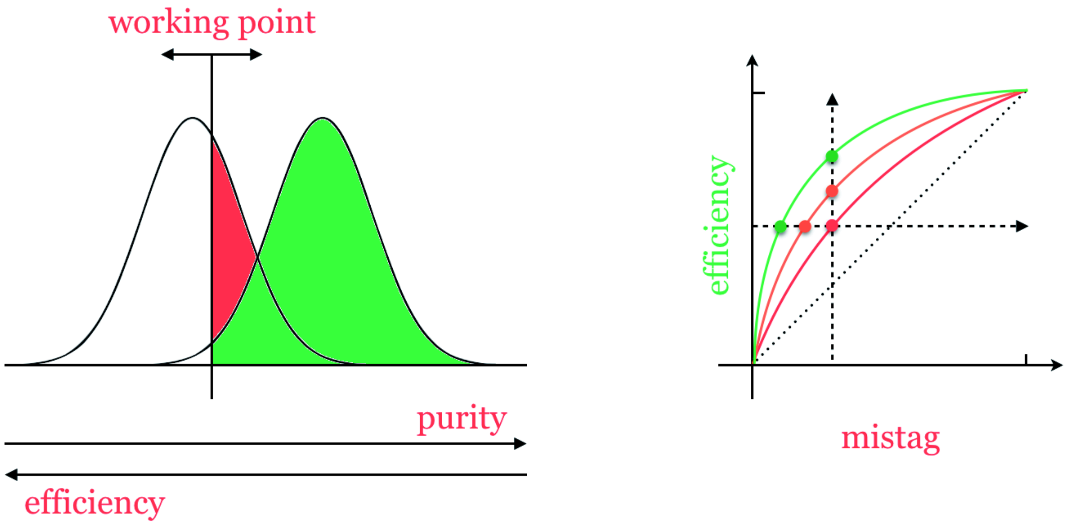


Figure 4.5: ROC curves and performances of a binary classifier.

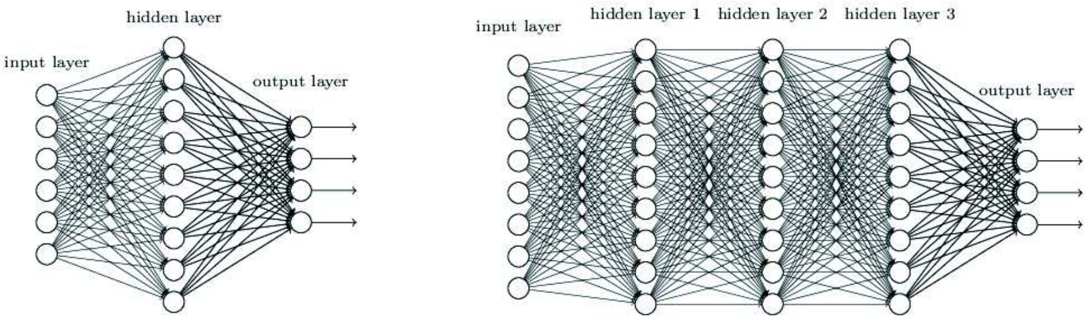


Figure 4.6: Comparison between a neural network such as CSV and a deep neural network such as DeepCSV.

respectively, the mistag rate being the probability to identify jets originating from light quarks (u, d, s) and gluons as b jets.

Last year, a new tagger was introduced, leading to better performances in the identification of jets originating from b quark hadronization. This tagger called Deep CSV is making use of the same set of variables than the standard CSV tagger (extending only the number of tracks used) but is based on a deep neural network of four hidden layers of 100 nodes each. Deep Neural Networks are based on the same idea than shallow neural networks such as the one used for CSV but allow to catch more correlations between the variables by using several hidden layers (Figure 4.6).

The input value of a node is the sum of the weighted inputs. The weights are derived during the training, fixed in the network and associated to the edges [1]. The output value of a node is produced from the input value trough an activation function. The activation

[1] edges are the link between two nodes

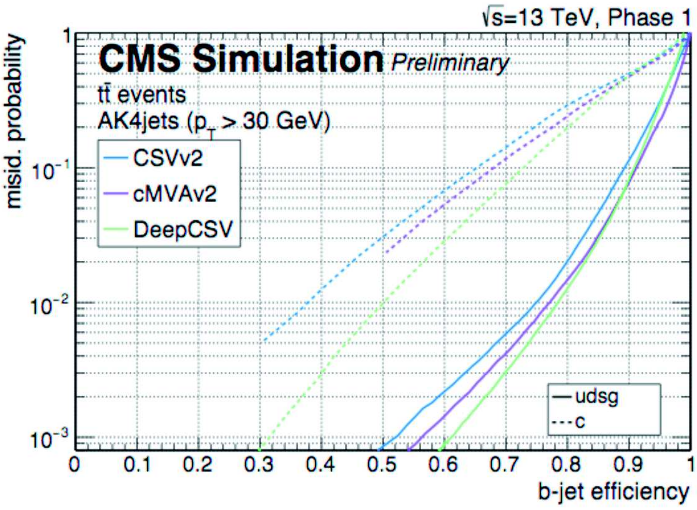


Figure 4.7: Performances of the CSV, cMVA and Deep CSV taggers.

function can be identity in which case the input and output are equivalent but is usually a rectified linear unit (x = max(0, x)) or an hyperbolic tangent. In a Deep Neural Network, the first layers will catch raw features whereas later layers will be specialized in the recognition of higher level features leading to a better discrimination at the output level. The training had to be adapted to this new algorithm and used around 50 millions of jets [52].

Another tagger called cMVA (combined MVA) was developed making use of more variables (adding soft lepton information) in order to help identifying jets originating from b hadron decay. The use of more variables led to an improvement in the tagging efficiency but shows lower performances than the DeepCSV tagger.

Using the same variables, Deep CSV shows a relative improvement in b-tagging efficiency of the order of 20% with respect to CSV at the tight working point for simulated $t\bar{t}$ events (Figure 4.7). An extension of Deep CSV using the additional variables used in cMVA is in progress.

4.4 Interlude: the challenge of c-tagging

Thanks to the good tracking and vertexing of CMS, another challenging identification can be performed: c-tagging. Hadron decays from c quarks share to a lesser extent some of the properties of hadron decays from b quarks. Their identification is therefore more complicated, all the input variable distributions used in the tagging ranging between the ones from light and b jets. A dedicated c-tagger was developed and used in 2016 [48].

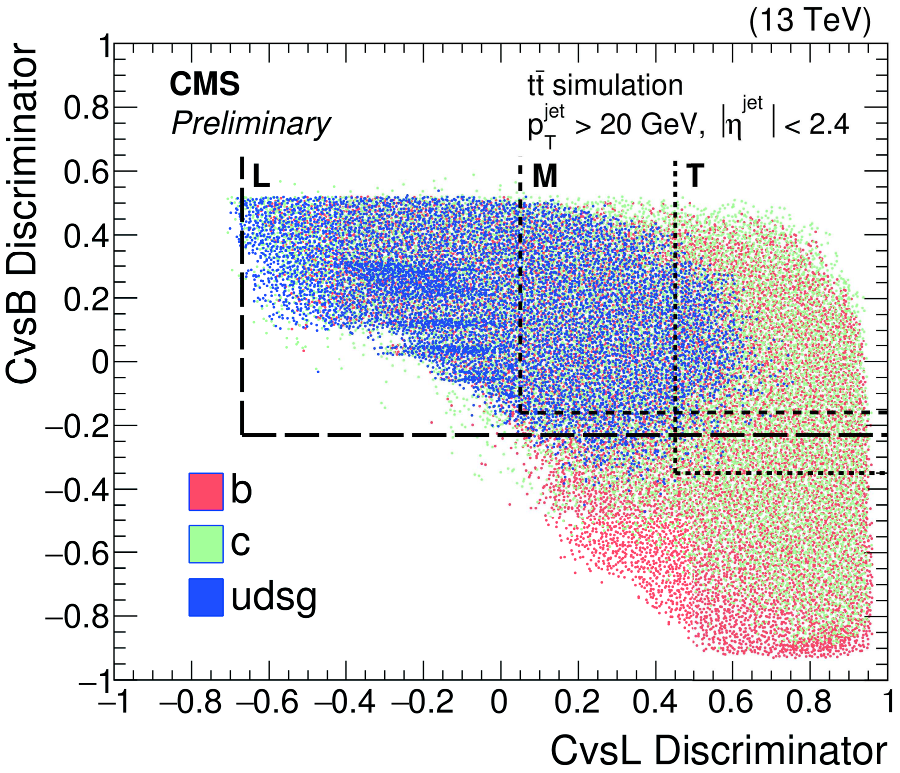


Figure 4.8: Two dimensional overlay of the BDT discriminator for b(red), c(green) and light (blue) jets for the tagging of c jets.

Figure 4.8 shows the difficulty to discriminate between light, c, and b jets. A classifier is trained against light jets (CvsL), a second one against b jets (CvsB) and the final c jet identification is based on a two-dimensional cut in the plane covered by the two classifiers.

However, a new integration of a c-tagger based on a Deep Neural Network seems to lead to better performances than the dedicated tagger and might be used as the default c-tagger in the future.

4.5 Double b-tagger and boosted environnement

The last tagging which was developed related to identification of hadron coming from b-quarks was the double-b tagger [50]. This tagger is targeting boosted decay of resonances decaying into two b jets such as H → $b\bar{b}$, allowing to probe processes such as HH and VH (V = W or Z) production, $t\bar{t}$H and boosted single objects. For particles produced with a momentum higher than their mass, the decay products are collimated and are merged into a fat jet (typically AK8). During Run 1, such kind of boosted topology was making use of subjet tagging (jets reclustered from fat jets) and fat jet tagging. During Run 2, a specific tagger was developed to reconstruct 2 b jets within a single fat jet (Figure 4.9).

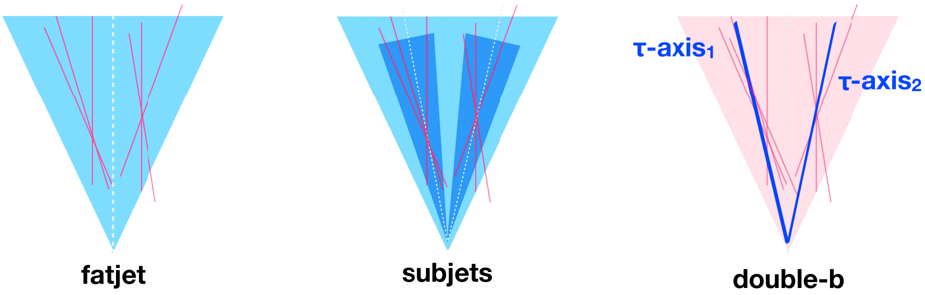


Figure 4.9: Sketch of fat jet, subjet and double-b tagger.

Based on variables such as the ones used in CSV but adapted to the τ-axes [2] of the subjets and on some specific variables such as the smallest distance between the secondary vertices, the double-b tagger shows better performances in the case of H $\rightarrow b\bar{b}$ than the previously used subjet and fat jet CSV taggers (Figure 4.10). The process considered for the mistaging efficiency here is the gluon splitting into a $b\bar{b}$ which will be an important background of the search for boosted H $\rightarrow b\bar{b}$ because of the similar final state with two collimated b jets.

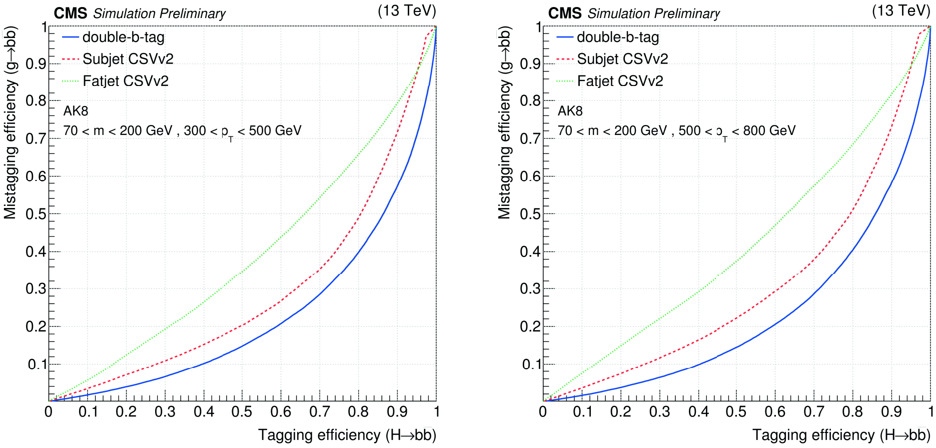


Figure 4.10: Double b-tagger efficiency compared to subjet and fat jet taggers in simulated $H \rightarrow b\bar{b}$ events with p_T(Higgs) > 300 GeV.

The efficiency measurement in data is based on the study of events containing an AK8 jet and two muons, one associated to each of the subjet. The selected events are expected to come mainly from quark production from gluon splitting which, in the absence of events from the processes under study such as H → b$\bar{\text{b}}$, provide a good sample to study the performances of the tagger.

[2] τ-axes are related to the standard subjets reconstruction derived from the computation of τ_N (parametrized k-mean algorithm targeting the optimal clustering in subjets) [49]

4.6 Scale factors

Differences in performances of the taggers between collected data and simulation may arise from changes in the data taking conditions or from the simulation itself. In order to correct for these differences in the analyses, different sets of data and different selections are used to acertain the level of discrepancy between data and simulation. The comparison allows to produce corrections to be applied to the simulation. Scale factors are defined as the ratio of the tagging efficiency in data to the tagging efficiency in simulation. The scale factors are produced in bins of p_T, η and depending on the jet flavour. While in the simulation, the jet flavour is derived from the matching to the generated b or c hadron, the measurement of the efficiency in data requires to select a sample of jets enriched in a given flavour.

Misidentification scale factors

The scale factors for light jets are derived from inclusive multijet events using a negative tag method. This method relies on positive and negative taggers using the same set of input variables as the standard tagger, but using tracks with positive and negative impact parameters and secondary vertices with positive and negative decay lengths respectively. The distributions for positive and negative taggers are expected to be the same for light jets, modulo some corrections due to secondary interactions, V0 decays and fake tracks which are infered from the simulation.

Scale factors for c jets

Enriched c jet samples are using the associated production of W boson with c quark and the semileptonic $t\bar{t}$ decay with one of the W decaying to $c\bar{s}$.

For the associated production of W and jet, with $W \to \ell\nu$ decay ($\ell =$ e or μ), events are selected requesting one isolated lepton satisfying some identification criteria. The presence of a neutrino in the W decay is used to reject the multijet background through a selection cut on the lepton transverse momentum and on the missing transverse energy. Finally, the leading jet is requested to contain a non isolated soft muon (from charm hadron decay). The signal purity is above 60% in both the electron and muon channels after the selection cuts. The c quark efficiency is defined as the ratio of the number of tagged c jets over the total number of c jets in the sample. The scale factor is derived from the efficiencies measured in both data and simulation.

The semileptonic $t\bar{t}$ selection takes profit of the presence of a c quark in half of the W hadronic decays. Events are selected requiring an isolated muon and four jets within the tracker acceptance. The leptonic decay of the W is used to reduce the multijet background by requiring the transverse mass of the isolated muon and the missing transverse energy to be above 50 GeV. Jets are assigned to the two top quarks by combination of two jets compatible with the W mass and three jets compatible with the top quark mass. The two b jet candidates are then required to pass the tight and loose working point of the CSV tagger respectively. The two remaining jets are used to determine the efficiency of the tagging algorithm.

Scale factors for b jets

The scale factors for b jets are derived selecting events enriched in b jets. Such events can be selected by requiring a muon inside a jet (from b hadron semileptonic decay) in multijet events, or by a $t\bar{t}$ event (in the dilepton final state) requiring two leptons of opposite charge and of different flavour in order to reduce the contamination from Drell-Yan events.

In the case of muon enriched samples, three methods are used to derive scale factors for b-jets: PtRel, LifeTime and System-8. This sample can only be used to derive scale factors of taggers which are not using lepton information. Moreover, this sample composed of jets containing a muon is slightly biased, the multiplicity of tracks in this b-jets being lower than in an inclusive b-jets collection and the muon track being better reconstructed than another track in a standard b-jet.

The PtRel method relies on the distribution of the transverse momentum (p_T^{rel}) of the muon relative to the jet axis. Because of the large mass of the heavy hadrons, this variable is expected to have a larger value for b hadrons. The b-tagging efficiency is derived from the ratio of the number of b jets tagged over the total number of b jets, estimated from the fit to the p_T^{rel} distributions.

The LifeTime method is based on the same strategy, namely deriving the efficiency from the ratio of b-tagged jets over the total number of b-jets, estimated here from the fit of a tagger (Jet Probability, which relies on the IP significance of all selected tracks in the jet with positive IP values, those tracks with negative IP values being used to build a probability distribution).

The System-8 method relies on the usage of weakly correlated b-taggers and sub-sets of the muon enriched sample. The b-tagging efficiency can be derived by numerically solving a set of equations relating tagging efficiencies for b and non b-jets content of two sub-sets of the muon enriched samples.

These three methods are combined, taking into account the statistical and systematic correlations.

The achieved relative precision on the b jet scale factor is 1 to 1.5% for jets with a p_T

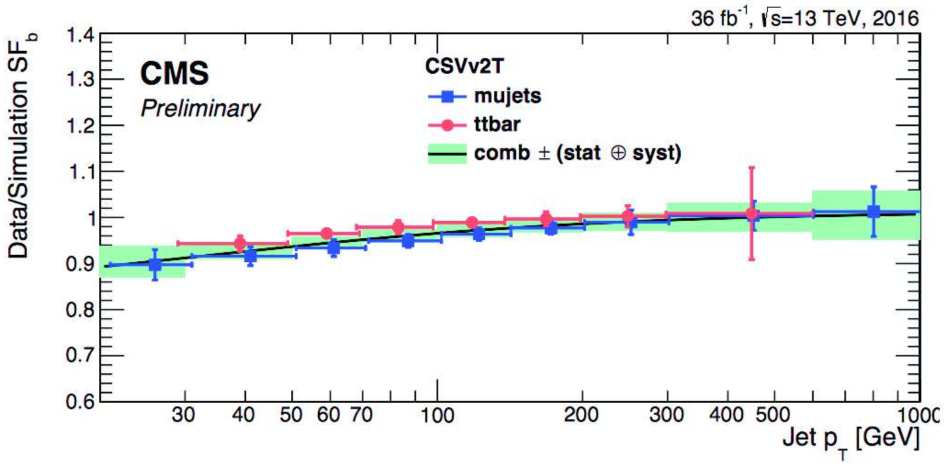


Figure 4.11: Inclusive b-jet scale factors in 2016 [47].

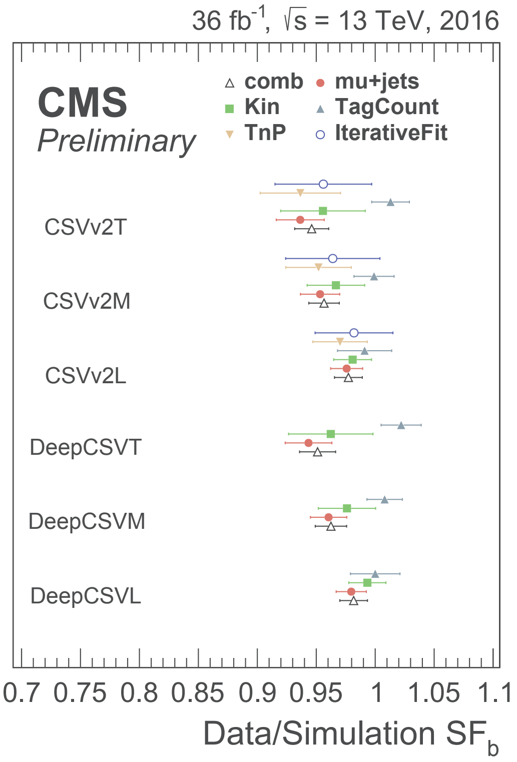


Figure 4.12: Scale Factors for ICHEP 2016. The Kin and TnP correspond to scale factors derived from $t\bar{t}$ in the dilepton final state while TagCount and IterativeFit correspond to scale factors derived from $t\bar{t}$ in the semileptonic final state.

between 70 and 100 GeV and rises to about 4 to 9% at the highest considered jet p_T.

The comparison of the b jet scale factors for the tight working point of the CSV tagger in 2016 (Figure 4.11) shows that the data/simulation ratio ranges from 0.9 for lower p_T values of the jets to 1 at higher values. The scale factor from $t\bar{t}$ is combined to the one from muon enriched.

4.7 B-tagging at trigger level

In this section, we will go through the interplay between b-tagging and trigger. The first part will be dedicated to triggers used to collect data needed for the commissionning of offline taggers and for the derivation of scale factors. The second part will be dedicated to the online b-tagging, used for analyses such as $t\bar{t}$ decaying fully hadronically or for Higgs decaying into $b\bar{b}$ for which the identification of b jets is needed to collect the event among the overwhelming hadronic background.

4.7.1 Control paths

In order to be able to commission the various input variables of the taggers and to derive scale factors for correcting the data / simulation differences, dedicated sets of events have to be taken. This section will go through the various algorithms used to take such data.

The first set of paths is composed of inclusive AK4 jet paths targeting the collection of events containing light jets (from u, d, s quarks and gluons) and pileup jets. This set of paths is used for the commissionning of the taggers and the derivation of mistag rate scale factors. The lowest p_T threshold is seeded by ZeroBias (random trigger based on beam bunch crossing time) while all the following paths are seeded by L1 jets.

The second set of paths is targeting the collection of events enriched in b jets in order to commission the input variables of the b-tagger and to derive scale factors for the tagging. It uses the frequent (around 20% when including the cascade to c) decay of b-hadrons to muon in the final state, leading to a presence of a jet containing a muon in the event. At L1, this set of paths requires a loose muon ($p_T > 3$ GeV) to be inside an AK4 jet. At HLT, the threshold is raised to 5 GeV for the muon and a second jet is required in the event in order to reduce the rate. All these paths are prescaled meaning that only a fraction of the events satisfying the trigger conditions will be kept. The prescales are derived during the data taking based on a target rate allowing to get enough events to provide scale factors with a low statistical uncertainty. Because these paths are prescaled, the coverage of the p_T range is done by several triggers with different thresholds. In 2016, a new path was added to offer a better coverage of the p_T range between 200 and 300 GeV (Figure 4.13).

The following distributions (Figure 4.14) show the value of the CSV and Deep CSV discriminants in samples collected with these triggers.

This dataset provides one of the most precise derivation of b-tagging scale factors

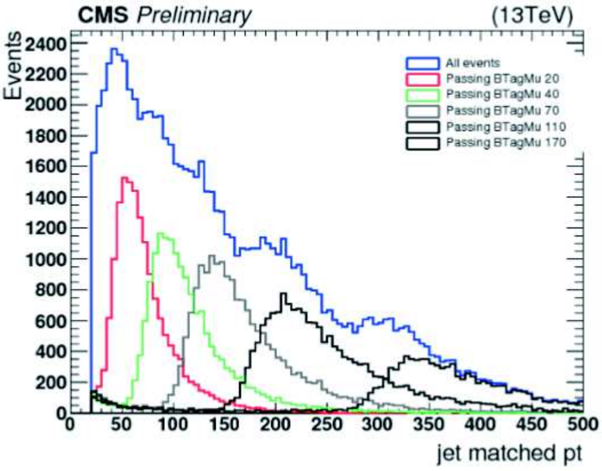


Figure 4.13: p_T distribution of jets collected with the BTagMu paths set.

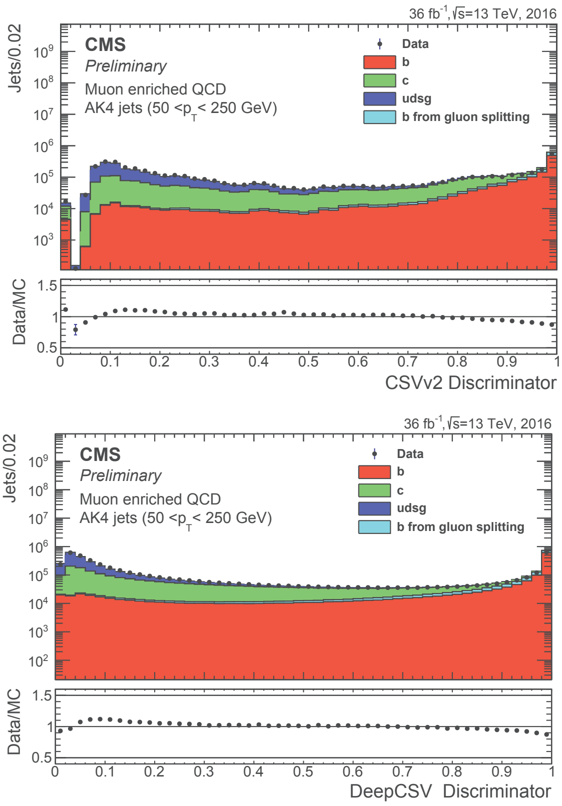


Figure 4.14: CSV and Deep CSV distribution of events collected requiring a muon in a jet at trigger level.

(Figure 4.12).

The third set of paths is composed of AK8 jets containing a muon. A new path was added in 2016, targeting boosted events. It requires a muon inside an AK8 jet of p_T above 300 GeV at HLT and is seeded at L1 by a high threshold single jet ($p_T > 200$ GeV). This trigger allowed to collect data for the commissionning (Figure 4.15) and scale

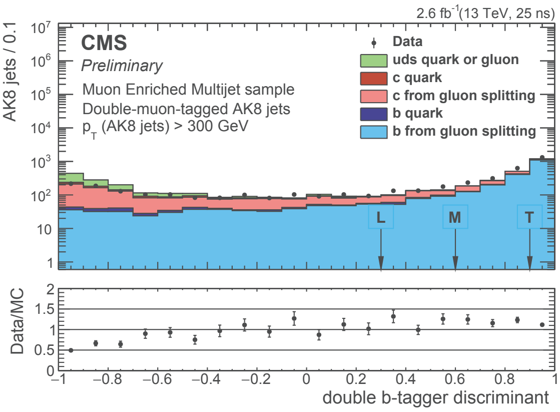


Figure 4.15: Distribution of the double-b tagger in data collected by BtagMu triggers.

factor derivation (Figure 4.16) of the double-b tagger, later used in the inclusive analysis of H → $b\bar{b}$ [42].

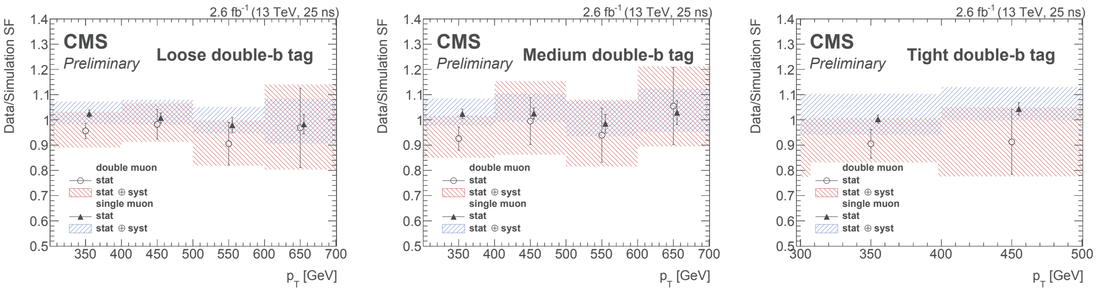


Figure 4.16: Scale factors derived from BTagMu (AK4/AK8) triggers

In 2017, one new path was added to lower the threshold on the jet containing the muon, requiring two AK8 jets . Work is ongoing to integrate a last path which will help collecting data for the boosted double-muon tagged jets, which is one of the two categories used in the derivation of the double b-tagger scale factors.

4.7.2 b-tagging at HLT

The b-tagging at HLT relies on two sequences [44] (Figure 4.17). The first one is based on the local reconstruction of calorimetric jets after a fast determination of the primary vertex. The second one is based on the reconstruction of Particle Flow jets, giving a more precise estimation of the energy of the jet but requiring more computing time to be performed.

The calorimetric sequence begins with the fast reconstruction of the primary vertex using the FastPV algorithm. While the position in the transverse plane is constrained according to the beam spot, its position along the beam line has to be determined. Clusters

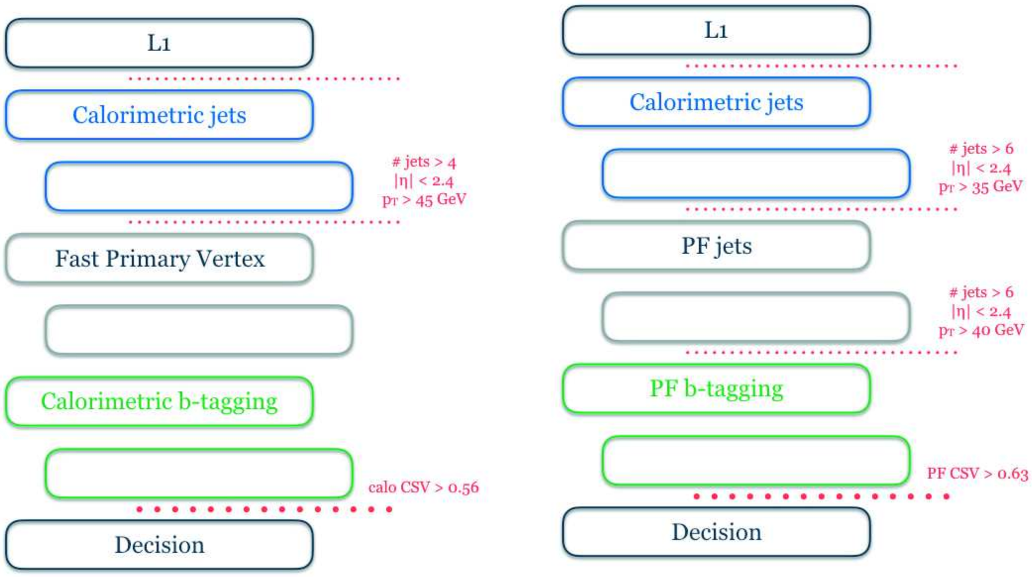


Figure 4.17: Structure of paths using the calorimetric (left) and PF (right) b-tagging sequence at trigger level.

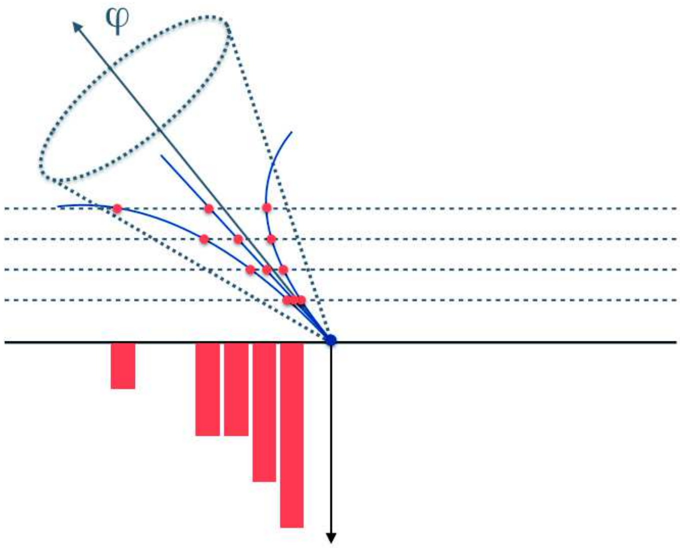


Figure 4.18: Fast PV strategy.

of pixels compatible with the direction in ϕ of the jets are selected. A set of requirements is applied to the cluster shape and direction in order to reject pileup contributions. The clusters are then weighted based on their probability to be associated to the jet. The weight is derived at each of the three consecutive steps, leading to an increased resolution while keeping the timing under control. After projection on the z-axis, the primary vertex appears as a peak in the cluster distribution (Figure 4.18).

The Fast Primary Vertex algorithm allows for a resolution along z of the order of 2.5 mm. From this first constraint, tracks originating from an area of 1.5 cm around the primary vertex and compatible with one of the eight leading jets are reconstructed using only the pixel information. The pixel tracks are used to put a new constraint on the primary vertex position, leading to the determination of the position of the pixel primary vertex.

Based on this primary vertex position, a regional reconstruction of the full tracks is done close to the axis of the jet. An iterative tracking is performed, close to the one used offline but using different steps and seedings in order to be more resilient against changes in the data taking conditions. The first step is seeded by pixel tracks, the next two steps are using triplets and pairs respectively and a reduced requirement on the p_T of the tracks in order to recover tracks with low p_T and one missing hit in the pixel detector. From the produced set of tracks, the position of the primary vertex is determined for the last time and the IVF algorithm [45] is used to find the position of the secondary vertex. Tracks and vertex are then provided to the CSV algorithm and a selection cut on the value of the discriminant is applied.

The PF sequence is used for trigger paths whose rate is low enough to allow for a longer computing time at HLT. While used standalone in some rare cases, the path has usually a more elaborate structure: beginning with calo jet reconstruction and removing events if the p_T and $|\eta|$ don't satisfy selection cuts, following with fast calo b-tagging, removing events if the CSV value doesn't satisfy the new selection cut, and following with PF jet reconstruction and PF b-tagging. Each step reduces the number of events through new selection cuts based on refined estimation of the variables associated to the jets (Figure 4.19).

With respect to offline b-tagging where three values of the discriminant cut can be used, online b-tagging allows for more flexibility in the choice of the selection cut on the discriminant. The goal when using b-tagging online is to have more handle on the rate by adding a requirement on the content of the event. Whereas offline b-tagging is used to improve the separation between signal and background, online b-tagging is a trade-off between rate and efficiency/purity (Figure 4.20). A better understanding of this trade-off can come from studying two exemples of the use of online b-tagging.

In the case of the fully hadronic decay of $t\bar{t}$, the expected event content is six jets among which two are b jets. Since there is no lepton in the final state, b-tagging needs to be applied online to select the interesting events in the hadronic environment. Triggers are therefore designed by requiring a sum of the transverse momentum of all jets above a threshold X, jet transverse momentum above a threshold Y and the value

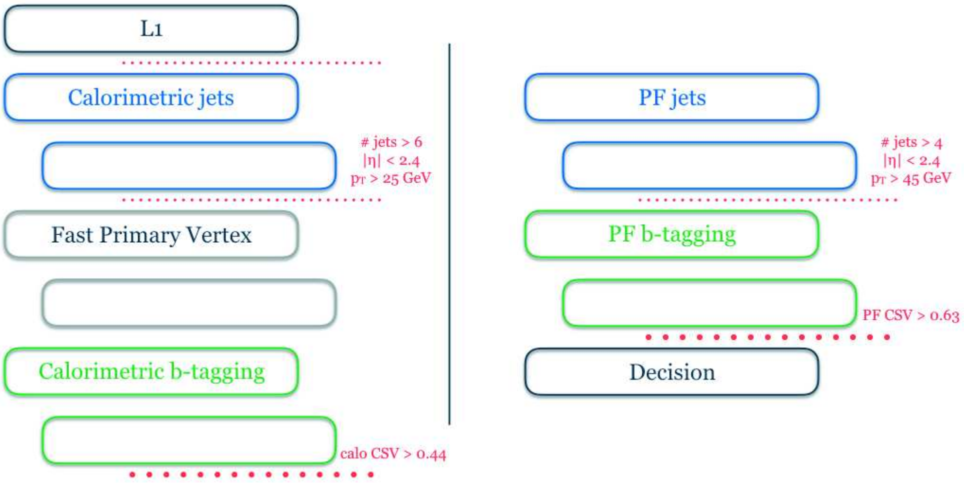


Figure 4.19: Structure of paths using both the calorimetric and PF b-tagging sequences at trigger level.

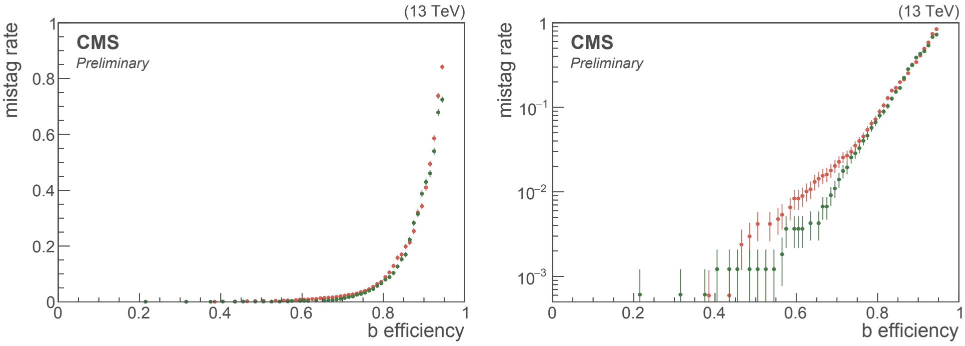


Figure 4.20: Performances (mistag rate vs. efficiency) of the calo (red) and PF (green) sequences with PU in linear (left) and logarithmic (right) scales.

of the b-tag discriminant above a threshold Z. The corresponding paths are labelled HLT_PFHTX_SixJetY_(Double)BTagCSV_pZ, Double denoting the fact that the condition on the b-tag discriminant has to be fulfilled by two jets. The goal of a trigger study is to find the X, Y, Z combination leading to the better efficiency / purity at a given rate. The rate is constrained by the total bandwidth at the output of the HLT farm / input of the Tier0. The efficiency is the fraction of signal events that would fire the trigger, the purity is the part of signal events among the events firing the trigger. In 2017, two triggers will be used: HLT_PFHT380_SixJet32_DoubleBTagCSV_p075 and HLT_PFHT430_SixJet40_BTagCSV_p080. Adding a requirement on the value of the CSV discriminant of the second jet allows for a reduction of the p_T threshold of all the jets at constant rate.

In the case of the fully hadronic decay of $\mathrm{t\bar{t}H}, \mathrm{H} \rightarrow \mathrm{b\bar{b}}$, the same optimization has to be done. From the six jets of the fully hadronic decay of $t\bar{t}$, we move to eight jets among

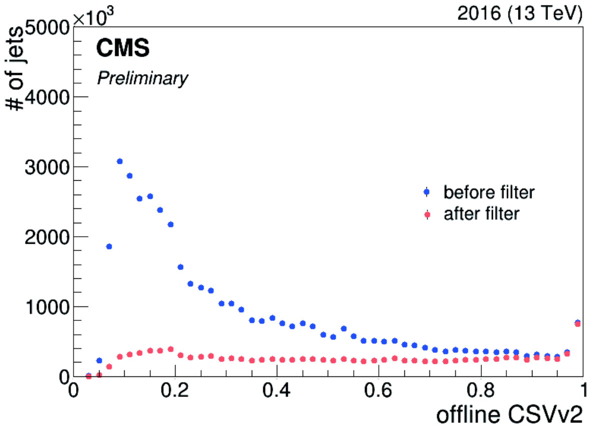


Figure 4.21: Offline CSV of jets before and after the selection cut on online CSV for the calo sequence.

which we have 4 b jets. Two options are then possible, based on the use of online b-tagging. Tagging three among the four expected b jets reduces further the rate. Increasing the total number of jets would reduce the rate as well but makes the trigger specific to this analysis. The rate reduction provided by the use of three b-tagged jets allows to reduce the minimum number of requested jets and creates a trigger of four jets among which three are b-tagged, which could be used for HH → 4b and other processes with four b jets in the final state.

One way of considering the performances of the online b-tagger is to determine the fraction of jets of a given offline CSV fulfilling the online cut on the discriminant. For this, we can look at the value of the discriminant before and after the cut is applied, using data collected with a high multiplicity jet requirement without any condition on the value of the discriminant. From data collected requiring the sum of the transverse momenta of jets to be above 800 GeV (condition fulfilled by high multiplicity jets events), we can study the impact of a trigger requiring the presence of six jets of p_T above 30 GeV and requiring one of the jet to have a value of the discriminant above a certain threshold.

Figure 4.21 shows the distribution of the offline CSV discriminant before (blue) and after (red) the cut applied at trigger level. A large reduction of the fraction of jets with low offline CSV can be observed. Figure 4.22 shows the ratio of the two curves and confirms a reduction of the rate at low offline CSV value while keeping an efficiency above 80% for jets passing the tight offline working point. Finally, Figure 4.23 shows that the efficiency is stable at 100% for high offline CSV values.

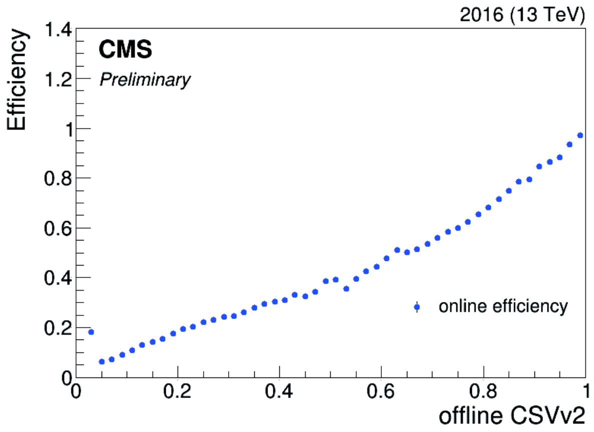


Figure 4.22: Fraction of offline jet passing the online cut for calo sequence with respect to the value of their CSV discriminant.

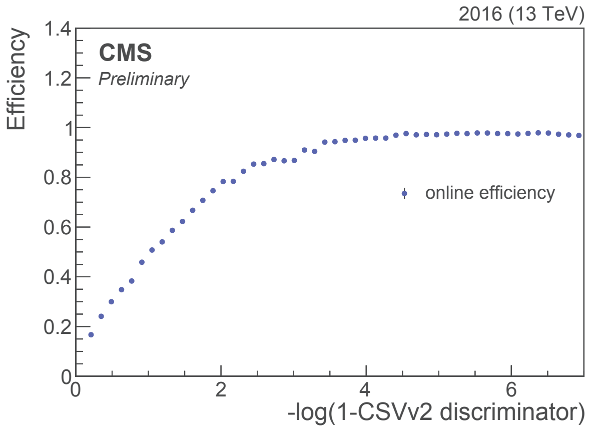


Figure 4.23: Fraction of offline jet passing the online cut for calo sequence with respect to the logarithm of the value of their CSV discriminant.

4.7.3 Impact of the new pixel detector at trigger level

The new pixel detector consisting of four layers of pixels has led to a new tracking strategy, impacting the b-tagging both offline and at trigger level.

The first impact the new pixel detector deployment had is the modification of the procedure for the track seeding. Seeds were previously reconstructed from the three layers by using a doublet propagated to the third layer. The increase in the number of pixels would lead to an important increase in the timing if the same procedure was to be used in 2017. The tracking therefore moved to the use of cellular automaton (CA, Figure 4.24), a technique which allows for important parallelization of the algorithm and which lead to a reduced computing time.

One can tune the CA algorithm in order to get the same efficiency than the one expected when deriving the quadruplets by propagation of the triplets in the previous

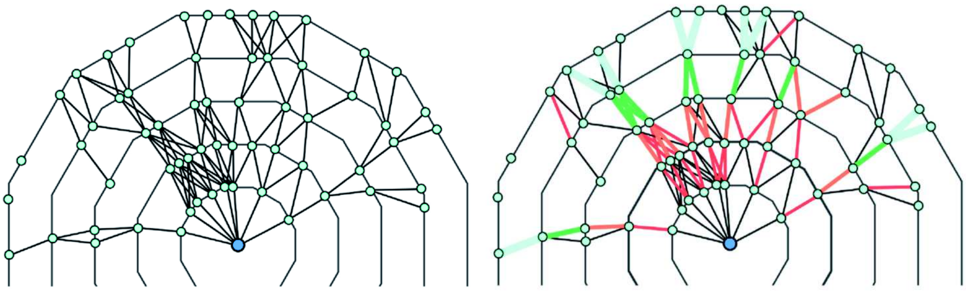


Figure 4.24: Cellular automaton propagation from layer to layer.

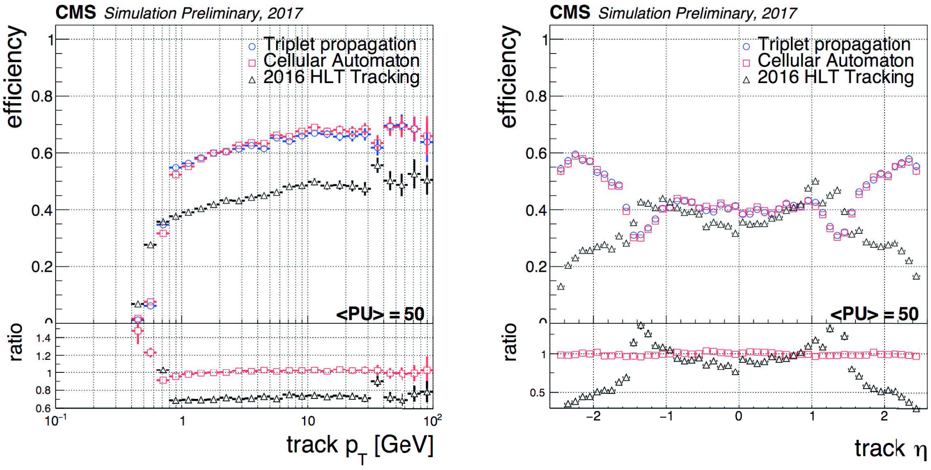


Figure 4.25: Efficiency of various tracking scenarii vs. p_T (left) and η (right).

tracking. In this case, the HLT pixel tracking efficiency (Figure 4.25) is better for CA while the fake rate (Figure 4.26) is reduced.

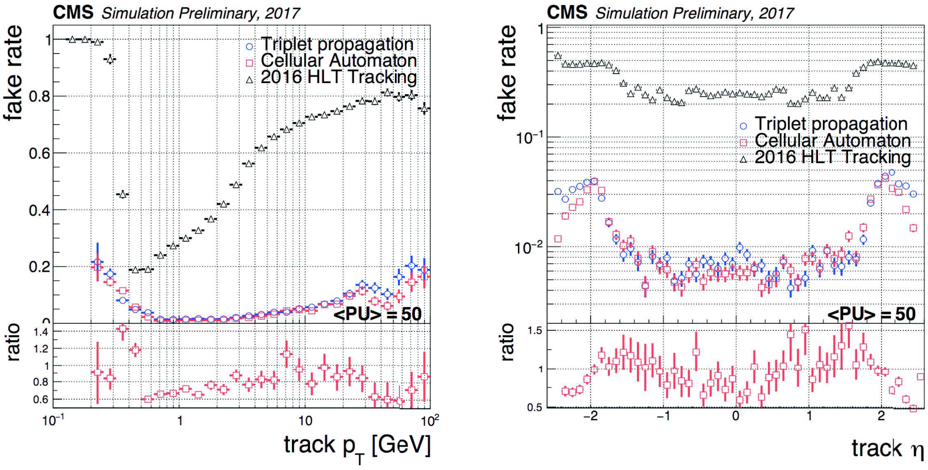


Figure 4.26: Fake rate of various tracking scenarii vs p_T (left) and η (right).

Moving to b-tagging, the use of the fourth layer of pixels leads to an increased resolution of the fast Primary Vertex (Figure 4.27), while keeping the timing within constraints.

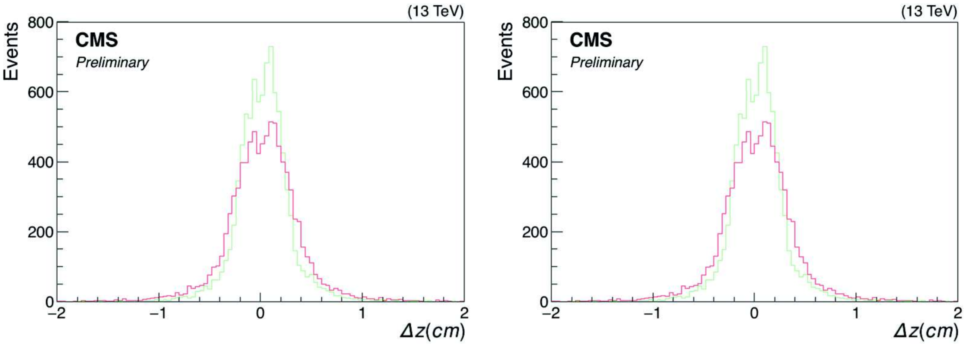


Figure 4.27: Resolution of the Fast PV in 2016 (red) and 2017 (green) without pileup (left) and with pileup 25 (right) in simulated TTbar events.

The b-tagging sequence benefits from the better spatial resolution of the primary vertex and the better tracking. When working with simulation, we can access the jet flavor and study the performances in terms of efficiency with respect to mistag rate, the goal of any development in the tagger being to maximize the discrimination between light and b jets (Figure 4.28).

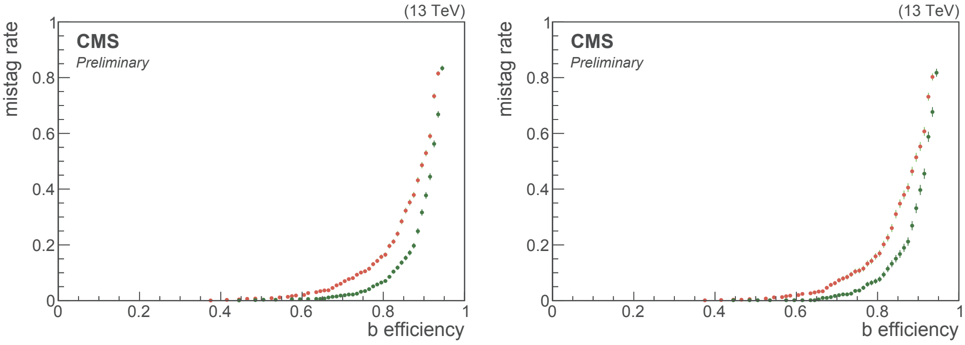


Figure 4.28: Performances (mistag rate vs. efficiency) of the CSV tagger in 2016 (red) and 2017 (green) against gluon (left) and light jet (right) for the calo sequence with pileup 25 in simulated TTbar events.

The optimal scenario of the whole pixel detector being fully working and aligned after deployement would have led to nominal performances for the b-tagging at HLT. However, few failure scenarii (incomplete alignement procedure, random or structured loss of parts of the tracker) had to be studied in order to acertain the impact they would have on the performances. The main problems related to tracking are the loss of pixels (randomly distributed or in part of the detector due to common mode failure) and problems in the alignement of the detector.

4.7.4 Towards a new tagger at HLT

The new Deep CSV tagger shows better performances than the standard CSV tagger. Being able to use it at trigger level would allow to reduce the rate further for the same efficiency. Some modifications were made to be able to run it in the two sequences (calo and PF).

In order to commission the use of this tagger in 2017, a copy of paths making use of the standard calo and PF sequences based on CSV is in progress. The sequences will then be replaced with a DeepCSV version. At first, this duplicated path will be deployed in the shadow of the one used for the data taking for the beginning of the year. After few weeks of commissionning, analyses will be able to move to this new tagger in order to improve the purity of their data.

4.8 Conclusion

After a preliminary work during LS1 on the study of paths aiming at collecting data for the commissionning of the taggers and the derivation of the scale factors, I was given the opportunity to work as convener of the group in charge of the b-tagging at trigger level. The main task was to follow the evolution of the performances of the online b-tagging during the data taking, working on mitigation procedures for failure scenarii and following the data taking of events used for the derivation of the scale factors of all the taggers.

I added a new algorithm to the set of pre-existing ones in order to provide a better coverage of the p_T spectrum in the muon enriched sample. Based on the need of analyses with boosted objects decaying to two b jets in the final state, I introduced a new trigger algorithm. Based on the presence of a muon in a fat jet, the algorithm is dedicated to the commissionning of the double-b tagger and the derivation of the scale factors. I followed the integration of a second algorithm based on the same needs and aiming at reducing the threshold on the p_T of the jet containing a muon.

In prevision of the change of pixel detector, I worked on and supervised the adaptation of the sequences to profit from the new geometry. I followed the offline development, making sure that the online algorithms stay as close as possible to the latest evolution of the offline taggers, despite the specific restrictions existing at trigger level. I followed the reintegration of the algorithms dedicated to the collection of events for the commissionning in the L1 and HLT menu at the beginning of 2017, supervising the rate and timing studies needed for their validation. I finally tested the performances of the online tagger with the first data of 2017, making sure that the online b-tagging was performing according to the expectations with the new detector and using the recently deployed cellular automaton.

I finally initiated the integration of the new DeepCSV tagger at HLT, following the first steps of its testing at HLT.

Thanks to the succesfull installation of the new pixel detector during the 2016-2017 winter shutdown, a new area now opens for b-tagging in CMS. The more precise determination of the secondary vertex and the better tracking shows already increased performances which will help in challenging analyses such as $\mathrm{H} \to b\bar{b}$ and $\mathrm{t\bar{t}H}, \mathrm{H} \to \mathrm{b\bar{b}}$.

At trigger level, the online b-tagging performed well during the beginning of the Run 2. The online b-tagging sequences have been updated to make use of the new pixel detector and provide to analyses a better handle on the selection of events. A new tagger will be deployed soon, relying on an improved algorithm.

Chapter 5

Search for $\mathrm{t\bar{t}H}$ production at $\sqrt{s} = 13$ TeV

Contents

5.1 Introduction

After the discovery of the Higgs boson by the ATLAS and CMS collaborations [53, 54], a new era opened to study its properties.

Among the main production modes of the Higgs boson (Figure 5.1), the associated production of a Higgs boson with one top quark (tHq) or a pair of top quarks ($t\bar{t}H$) is the only direct probe of the coupling between Higgs and top. While the $t\bar{t}H$ process allows to measure the coupling, tHq allows to access the sign of the coupling (Figure 5.2). An indirect probe is provided by the quark loop in the gluon fusion production mode of the Higgs boson and the loop in the Higgs decay to a pair of photons. While the fit of the couplings in the κ framework shows a positive coupling of the Higgs to fermions, this fit relies on hypotheses and the direct measurement of the coupling is awaited in order to confirm this result.

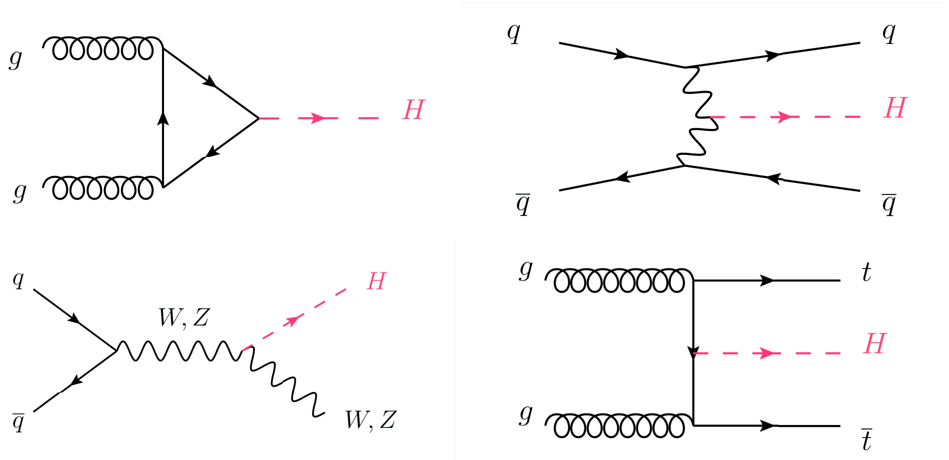


Figure 5.1: Higgs main production modes, tHq not included.

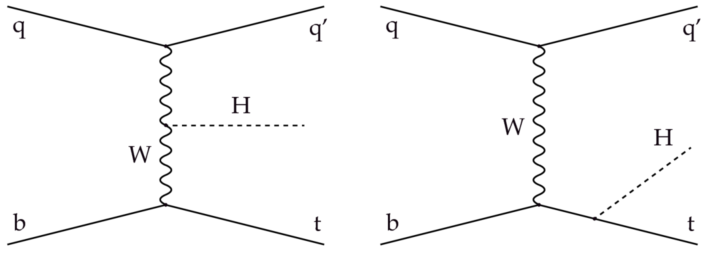


Figure 5.2: Feynman diagrams of the tHq processes.

One of the challenges of studying the top-Higgs coupling through t$\bar{\text{t}}$H production is the production cross section (0.50 pb at 13 TeV) which is two orders of magnitude smaller than the inclusive production cross section of the Higgs boson (50 pb at 13 TeV, dominated by gluon fusion).

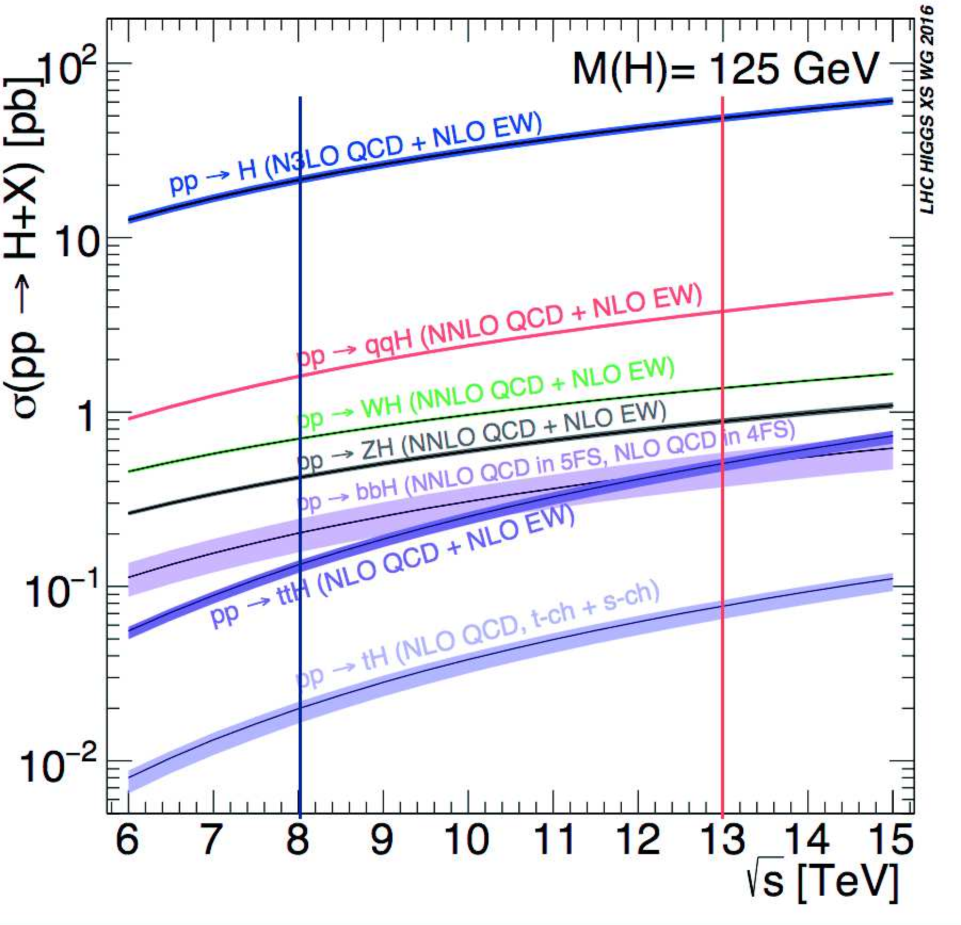


Figure 5.3: Evolution of the production cross sections of the main Higgs boson production modes as function of the centre-of-mass energy.

From Run 1 (8 TeV) to Run 2 (13 TeV), the production cross section of tt̄H increases by a factor four (Figure 5.3), while the production cross section of the main backgrounds (tt̄+X) increases by a factor of roughly three, leading to a better expected sensitivity.

During Run 1, the study of the tt̄H production led to results compatible with the Standard Model [55]. However, a small tension was observed in the decay of the Higgs boson to WW (Figure 5.4) with an observed excess of about 2σ with respect to the expected value from the Standard Model. This excess was driven by the study of Higgs boson decaying to multilepton, which motivated to pursue with higher statistics this channel at 13 TeV. The updated analysis at 13 TeV offers thus an interesting opportunity for this thesis.

Several strategies are targeting the study of the tt̄H production through the different decay modes of the Higgs boson, using the hadronic, semi-leptonic or fully leptonic final state of the tt̄ pair (Figure 5.5). While the fully hadronic decays of the $t\bar{t}$ pair has a large branching ratio, it represents in the tt̄H analysis a challenging signature because of the

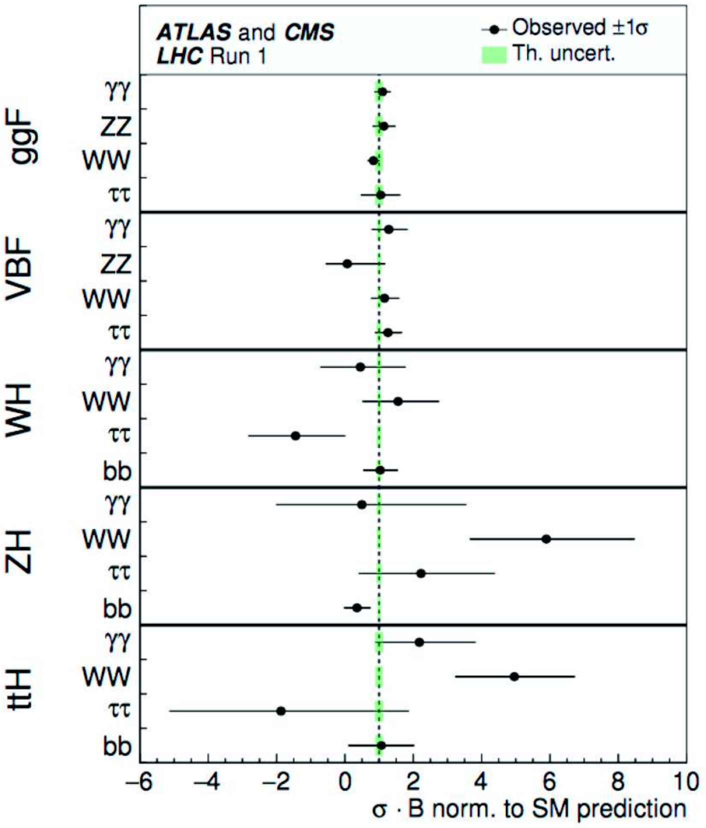


Figure 5.4: Results from Run 1 on Higgs production and decay.

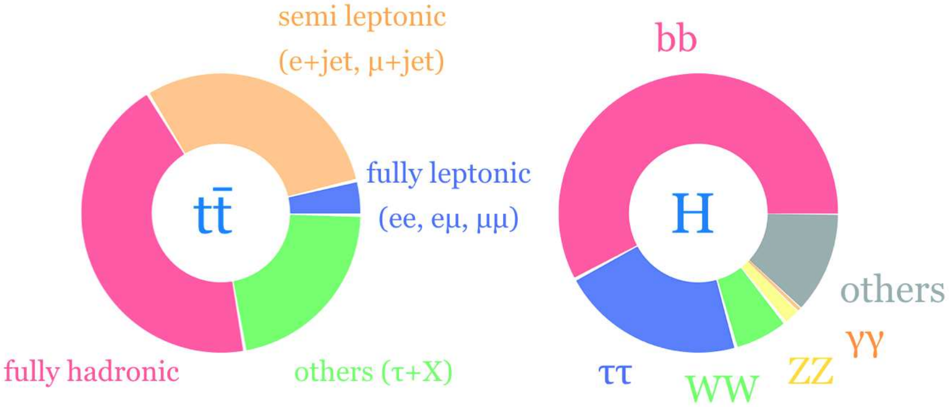


Figure 5.5: Branching ratios of t$\bar{\text{t}}$ and H.

large QCD background. The semi-leptonic final state of the $t\bar{t}$ pair has a large branching ratio and allows for an easier reconstruction of the pair. The fully leptonic decay of the t$\bar{\text{t}}$ pair has a small branching ratio (4%) but a clean signature. The semi-leptonic and leptonic decays provide an easy trigger through the presence of a hight p_T isolated lepton in the final state.

The decays of the Higgs boson to two photons or to two Z bosons decaying leptonically provide a very clean signature and allow for an unambiguous distinction between the $\mathrm{t\bar{t}}$ pair and the Higgs part of the event. However, these decay channels suffer from a low branching ratio. They were studied in 2016 data, the $\mathrm{t\bar{t}H}$ production mode being tagged in the context of the general Higgs to $\gamma\gamma$ and ZZ analyses.

In the analysis of the decay of the Higgs boson into two photons [57], two categories are defined in order to probe the top-Higgs coupling. The leptonic $\mathrm{t\bar{t}H}$ category relies on the presence of two jets among which one is passing the medium working point of the CSV tagger and the presence of at least one lepton in the event. The hadronic $\mathrm{t\bar{t}H}$ category relies on the presence of at least three jets among which one is passing the medium working point of the CSV tagger. The diphoton mass spectrum is fitted in these categories (Figure 5.6).

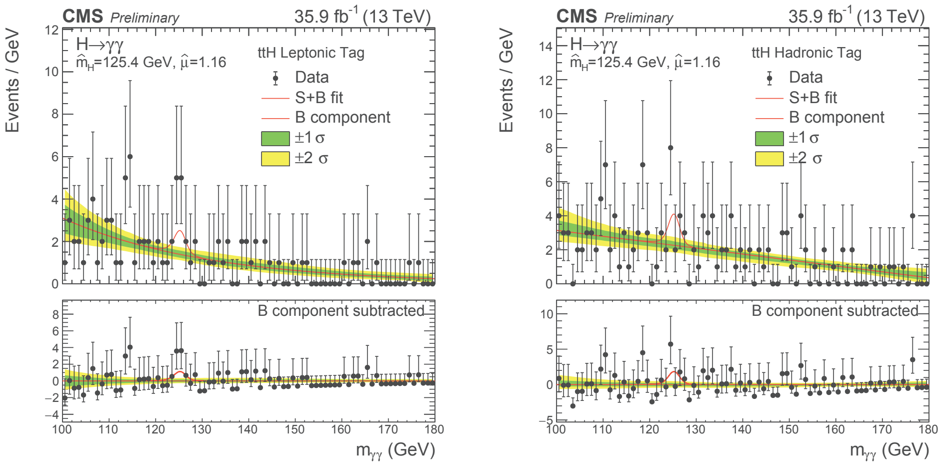


Figure 5.6: Fit of $\mathrm{t\bar{t}H}$ tagged categories: leptonic (left) and hadronic (right).

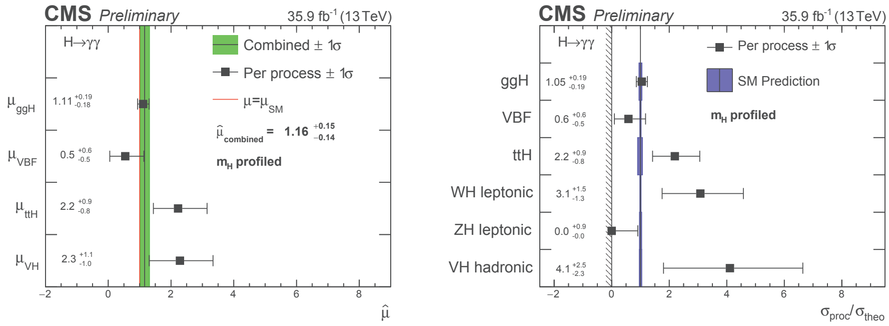


Figure 5.7: Signal strength modifiers of the various production modes of the Higgs boson with $H \to \gamma\gamma$ (left) and cross section ratios (right) at Run 2.

The best fit value is found to be $\mu_{\mathrm{t\bar{t}H}} = 2.2^{+0.9}_{-0.8}$ [1] when the fit is done in the two $\mathrm{t\bar{t}H}$ categories only (Figure 5.7).

[1] $\mu_{\mathrm{t\bar{t}H}}$ is the signal strength defined as the ratio between the observed rate and the one predicted under the SM only hypothesis.

The decay of the Higgs boson to WW/ZZ/$\tau\tau$ provides a clean signature and has a low reducible background. The next part of this chapter will be devoted to this decay mode.

The decay of the Higgs boson to a pair of b quarks has the highest branching ratio but suffers from the difficult modeling of the main backgrounds and from the combinatorics related to the high number of jets and b jets. The strategy adopted by the analysis [59] relies on a classification according to the number of jets and b jets. After the event selection, a discriminant based on a BDT aims to distinguish between $t\bar{t}$H and $t\bar{t}$ while the use of the Matrix Element Method (detailed in Section 5.9.2) in two bins of the BDT (low and high score) allows a better discrimination between $t\bar{t}$H and $t\bar{t}$+bb in the high score region (Figure 5.8).

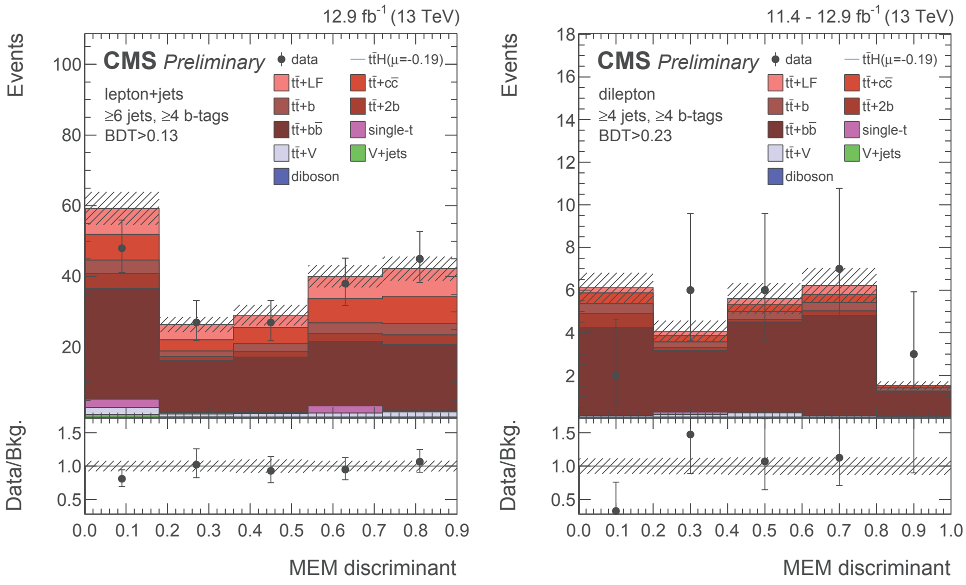


Figure 5.8: Post fit distribution of the MEM discriminant in the most sensitive categories in the lepton + jet (left) and dilepton (right) categories for the $t\bar{t}$H, H → $b\bar{b}$ search.

The best fit value is found to be $\mu_{t\bar{t}H}$ = -0.19$^{+0.45}_{-0.44}$ (stat)$^{+0.66}_{-0.68}$ (syst) (Figure 5.9).

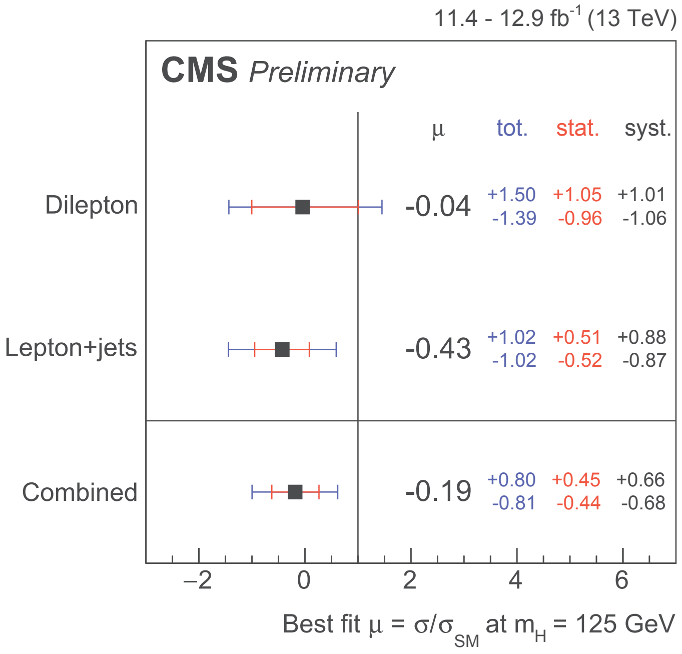


Figure 5.9: Signal strength modifier in the two categories considered in the $\mathrm{t\bar{t}H, H \to b\bar{b}}$ analysis based on the final state of the $\mathrm{t\bar{t}}$ decay.

5.2 Analysis strategy of $\mathrm{t\bar{t}H}$ to multilepton

The analysis strategy of the $\mathrm{t\bar{t}H}$ to multilepton search, based on the 35.9 fb^{-1} of data collected in 2016 [56], is to use both semi-leptonic and fully leptonic decays of the $\mathrm{t\bar{t}}$ system. Events are then categorized into two, three and four leptons, requiring the presence of additionnal b jets. The discrimination between the signal and the reducible and irreducible backgrounds is maximized through the use of multivariate techniques. This analysis aims to probe the decays of the Higgs boson to WW*, ZZ* or $\tau\tau$ (Figure 5.10).

After selection, the main remaining irreducible backgrounds are $\mathrm{t\bar{t}Z}$ and $\mathrm{t\bar{t}W}$ (denoted $\mathrm{t\bar{t}V}$). The main reducible background comes from $\mathrm{t\bar{t}}$ + jets leading to a signal similar to $\mathrm{t\bar{t}H}$ when one of the jet is wrongly identified as a lepton. Multivariate techniques are trained to separate $\mathrm{t\bar{t}H}$ and $\mathrm{t\bar{t}}$ on one side and $\mathrm{t\bar{t}H}$ and $\mathrm{t\bar{t}V}$ on the other. The $\mathrm{t\bar{t}H}$ signal yield is extracted from a fit to the 2D plane covered by both discriminants.

5.3 Monte Carlo samples

Different Monte Carlo simulations are used in the analysis. Some are applied in the training of the multivariate techniques used to discriminate between signal $\mathrm{t\bar{t}H}$ and back-

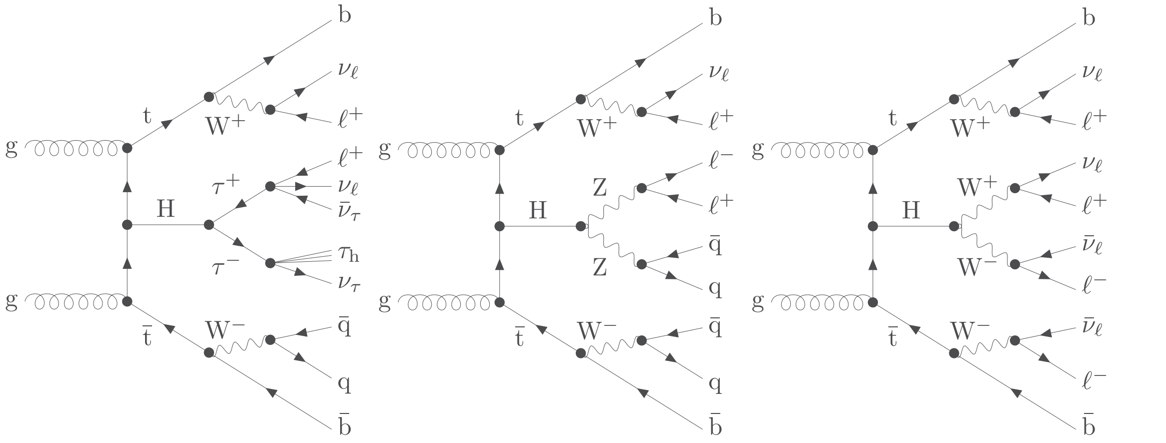


Figure 5.10: Examples of leading Feynman diagrams for tt̄H production in the multilepton final state.

grounds tt̄Z, tt̄W, tt̄, the others are used for the control and signal regions.

The tt̄H Monte Carlo signal relies on Powheg and Pythia8. The main irreducible backgrounds are generated using Amcatnlo and Pythia8. Table 5.1 shows the different generators involved for the signal, the background estimated from simulation and the background used for the estimation from data.

5.4 Event online selection

Events are recorded using single lepton, dilepton and trilepton triggers. While most of the events are passing the dilepton triggers, adding a single lepton trigger allows to recover events for which the second lepton p_T is below the threshold of the second leg of the dilepton triggers. Adding the trilepton trigger allows to reduce the threshold on the p_T of the leading lepton.

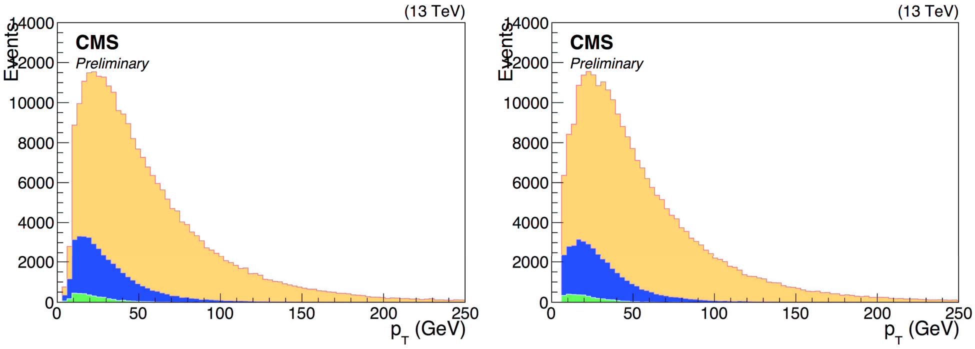


Figure 5.11: p_T of the leading (yellow), subleading (blue) and third (green) muons (left) and electron (right) produced in tt̄H process

In the same-sign dilepton category based on the presence of 2 muons, triggers are com-

Sample	Generators	cross section [pb]
t$\bar{\text{t}}$H (without $b\bar{b}$)	powheg + pythia8	0.215
t$\bar{\text{t}}$W $\rightarrow \ell\nu$	amcatnloFXFX + madspin + pythia8	0.2043
t$\bar{\text{t}}$Z $\rightarrow \ell\ell\nu\nu$	amcatnlo + pythia8	0.2529
$W\gamma \rightarrow \ell\nu\gamma$	amcatnloFXFX + pythia8	585.8
$Z\gamma \rightarrow \ell\ell\gamma$	amcatnloFXFX + pythia8	131.3
$t\gamma jets$	amcatnlo + madspin + pythia8	2.967
t$\bar{\text{t}}\gamma jets$	amcatnloFXFX + madspin + pythia8	3.697
Rares $W^+W^+jetjet$	madgraph + pythia8	0.03711
Rares $WW \rightarrow \ell\ell\nu\nu$	pythia8	0.1729
Rares WWW	amcatnlo + pythia8	0.2086
Rares WWZ	amcatnlo + pythia8	0.1651
Rares WZZ	amcatnlo + pythia8	0.05565
Rares ZZZ	amcatnlo + pythia8	0.01398
Rares tZq	amcatnlo + pythia8	0.0758
Rares $tttt$	amcatnlo-pythia8	0.009103
t$\bar{\text{t}}$Jets $\rightarrow \ell$	madgraphMLM + pythia8	182.18
t$\bar{\text{t}}$Jets $\rightarrow \ell\ell$	madgraphMLM + pythia8	87.3
Single top W	powheg + pythia8	35.6
Single top t-channel	powhegV2 + madspin + pythia8	136.02
Single top s-channel	amcatnlo + pythia8	80.95
$DYJets \rightarrow \ell\ell$ (M-10to50)	madgraphMLM + pythia8	18610
$DYJets \rightarrow \ell\ell$ M-50	madgraphMLM + pythia8	6025.2
$WJets \rightarrow \ell\nu$	amcatnloFXFX + pythia8	61526.7
$WZ \rightarrow \ell\ell\ell\nu$	powheg + pythia8	4.42965
$WW \rightarrow \ell\ell\nu\nu$	powheg	10.481
$ZZ \rightarrow 4\ell$	powheg + pythia8	1.256

Table 5.1: List of signal and background samples and generators used in the analysis.

bined requiring isolated muon with transverse momentum above 24 GeV or two isolated muons with transverse momentum above 17 and 8 GeV for the the leading and subleading lepton respectively.

In the same-sign dilepton category based on the presence of 2 electrons, triggers are combined requiring electron restricted to $|\eta| < 2.1$ with a transverse momentum above 27 GeV or two electrons with transverse momentum above 23 and 12 GeV for the leading and subleading ones respectively.

The rate of events containing one genuine lepton is mostly related to the production of W+jets whose production rate was given in Table 2.1. At 1.5 10^{34} cm^2.s^{-1}, the HLT rate is close to 100 Hz, to be compared with the total rate of 1 kHz which will be stored. The threshold on the electron is higher than for muon, the probability to catch a jet faking an electron being higher than the probability to catch a fake muon.

In the same-sign dilepton category based on the presence of one electron and a muon, the previously quoted single lepton triggers are used in combination with a low p_T lepton. For high p_T muon and low p_T electron, the muon p_T has to be above 23 GeV while the electron p_T has to be above 8 GeV. The thresholds are inverted in the case of a high p_T electron and a low p_T muon.

In the three lepton category, all these triggers are used in combination with trilepton triggers. In the case of a pure electron trigger, the three first electron p_T must be above 16, 12 and 8 GeV respectively while in the case of pure muon trigger, the three first muon p_T are required to be above 12, 10 and 5 GeV. The two last triggers require the presence of two muons and one electron with p_T above 9 GeV or two electron with a p_T above 12 GeV and one muon with a p_T above 8 GeV.

5.5 Event reconstruction and object identification

Events are reconstructed by using the particle-flow (PF) algorithm which offers an optimal combination of the information coming from all the sub-detectors, as described in Chapter 3.

Muons are reconstructed through a global fit based on information coming from the tracker and the muon spectrometer. Muons have to be in the acceptance of the muon system $|\eta| < 2.4$ and have a $p_T > 5$ GeV.

Electrons are reconstructed combining information from the tracker and the electromagnetic calorimeter. Electrons have to be in the tracking acceptance ($|\eta| < 2.5$) and have a $p_T > 7$ GeV. Electrons are discarded if they are within a ΔR distance lower than 0.4 from an already reconstructed muon candidate.

An additionnal step was added in order to increase the discrimination between *signal leptons* originating from W, Z and τ decays and *background leptons* coming from b-hadron decays or misidentification of jets. Three selection criteria are used for the electron and muon candidates: Loose, Fakeable and Tight. The selection corresponding to each criterion can be found in Annex.

Tau leptons decaying hadronically are reconstructed by the hadron-plus-strips algorithm. They are required to have $p_T > 20$ GeV and $|\eta| < 2.3$. In the latest version of the analysis, events with hadronic taus are vetoed to ensure orthogonality with the dedicated $\mathrm{t\bar{t}H}, \mathrm{H} \to \tau\tau$ analysis.

Jets are reconstructed by a clustering of the PF candidates in a cone of $\Delta\mathrm{R} = 0.4$. The overlap removal requires every jet to be separated from any lepton by a distance ΔR > 0.4.

The jets are considered as b jet candidates according to the value of the associated CSV discriminant. Two working points are used in this analysis, the loose one corresponds to an efficiency of around 85% and a mistag rate of around 10% for u, d, s, g jets while the medium working point has an efficiency of approximately 70% and a mistag rate of around 1.5%.

The missing transverse energy is computed as the negative vector sum of the transverse momenta of the PF candidates. An additionnal variable ($E_T^{miss}LD$) less sensitive to pileup is computed, using a linear combination of the missing transverse energy and the missing transverse momentum (H_T^{miss}) based on the vector sum of the transverse momenta of jets and leptons. This variable does not take into account energy which is not associated to any particle candidates (Equation (5.1)).

$$E_T^{miss}LD = 0.6E_T^{miss} + 0.4H_T^{miss} \tag{5.1}$$

The working point was tuned for an optimal separation of $\mathrm{t\bar{t}H}$ signal and Z + jets background events and is $E_T^{miss}LD > 30$ GeV.

5.6 Event selection and categorization

In order to reject events which don't correspond to the considered t$\bar{\text{t}}$H final state, the following selections are applied. All events are required to contain two leptons passing the tight selection. A further condition is applied to remove background from Z decays by vetoing events containing same flavour opposite charge leptons with an invariant mass in a 10 GeV window around the Z mass. Requiring $E_T^{miss}LD$ to be above 30 GeV reduces further the Z decay contribution. At least two hadronic jets with $p_T > 25$ GeV and $|\eta| < 2.4$ are required. Because of the presence of the t$\bar{\text{t}}$ pair, two jets are required to pass the loose working point of the CSV discriminant or one jet is required to pass the medium working point of the CSV discriminant. Several event categories are defined (Figure 5.13).

Same sign dilepton selection (2lss)

The same sign dilepton channel is based on the leptonic decay of two W bosons (one from the top, one from the Higgs) while the other top quark and W boson decay into jets. This leads to a final state with two same sign leptons, six jets among which two b-jets and some missing transverse energy produced by the presence of two neutrinos.

For those events with two tight leptons, the p_T of the leading and subleading leptons are required to be greater than 25 and 15 GeV respectively.

At least four jets with transverse momentum greater than 25 GeV and $|\eta|$ below 2.4 are required, allowing up to two missing jets in the event.

Three lepton selection (3l)

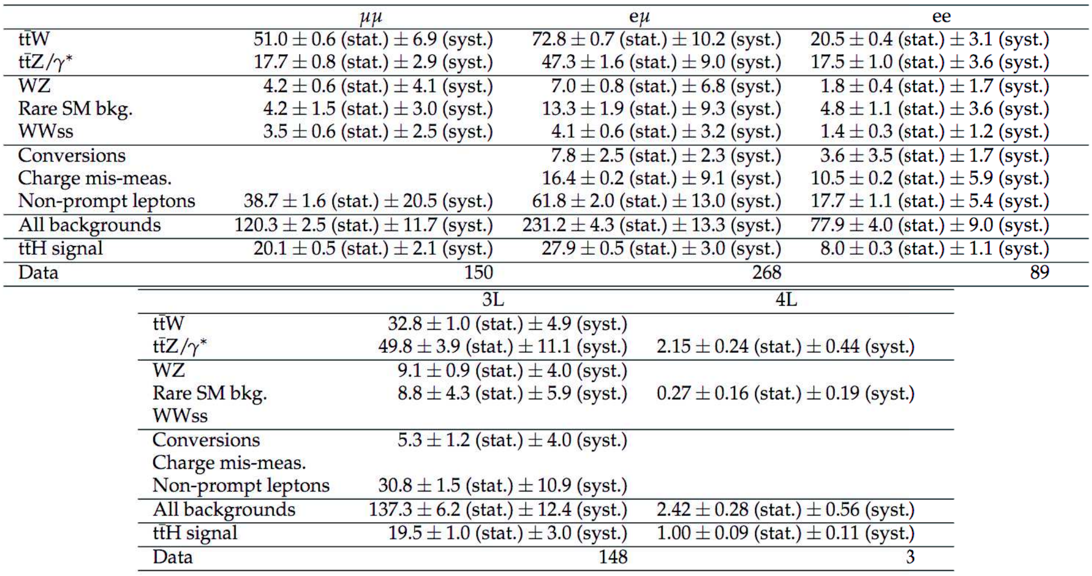
The three lepton category contains events from various processes: the fully leptonic final state of the Higgs to WW* pair with a semi-leptonic final state from the t$\bar{\text{t}}$ pair, or the semileptonic final state of the WW* pair from Higgs with a fully leptonic final state from the t$\bar{\text{t}}$ pair, or finally the leptonic decay of the off-shell Z from the Higgs boson together with a semileptonic final state from the t$\bar{\text{t}}$ pair.

The p_T of the three leptons are required to be greater 15 GeV, the leading one being greater than 25 GeV.

The sum of the lepton charges has to be +1 or -1 as expected for the signal.

Four lepton selection

In 2016, the statistics allowed for a four lepton category to be defined, based on the same requirements than the three lepton selection, with a minimum p_T of 10 GeV for the fourth lepton. Furthermore, the sum of charges is required to be 0 as expected for the



	$\mu\mu$	eμ	ee
t$\bar{t}$W	51.0 ± 0.6 (stat.) ± 6.9 (syst.)	72.8 ± 0.7 (stat.) ± 10.2 (syst.)	20.5 ± 0.4 (stat.) ± 3.1 (syst.)
t$\bar{t}$Z/γ^*	17.7 ± 0.8 (stat.) ± 2.9 (syst.)	47.3 ± 1.6 (stat.) ± 9.0 (syst.)	17.5 ± 1.0 (stat.) ± 3.6 (syst.)
WZ	4.2 ± 0.6 (stat.) ± 4.1 (syst.)	7.0 ± 0.8 (stat.) ± 6.8 (syst.)	1.8 ± 0.4 (stat.) ± 1.7 (syst.)
Rare SM bkg.	4.2 ± 1.5 (stat.) ± 3.0 (syst.)	13.3 ± 1.9 (stat.) ± 9.3 (syst.)	4.8 ± 1.1 (stat.) ± 3.6 (syst.)
WWss	3.5 ± 0.6 (stat.) ± 2.5 (syst.)	4.1 ± 0.6 (stat.) ± 3.2 (syst.)	1.4 ± 0.3 (stat.) ± 1.2 (syst.)
Conversions		7.8 ± 2.5 (stat.) ± 2.3 (syst.)	3.6 ± 3.5 (stat.) ± 1.7 (syst.)
Charge mis-meas.		16.4 ± 0.2 (stat.) ± 9.1 (syst.)	10.5 ± 0.2 (stat.) ± 5.9 (syst.)
Non-prompt leptons	38.7 ± 1.6 (stat.) ± 20.5 (syst.)	61.8 ± 2.0 (stat.) ± 13.0 (syst.)	17.7 ± 1.1 (stat.) ± 5.4 (syst.)
All backgrounds	120.3 ± 2.5 (stat.) ± 11.7 (syst.)	231.2 ± 4.3 (stat.) ± 13.3 (syst.)	77.9 ± 4.0 (stat.) ± 9.0 (syst.)
t$\bar{t}$H signal	20.1 ± 0.5 (stat.) ± 2.1 (syst.)	27.9 ± 0.5 (stat.) ± 3.0 (syst.)	8.0 ± 0.3 (stat.) ± 1.1 (syst.)
Data	150	268	89

	3L	4L
t$\bar{t}$W	32.8 ± 1.0 (stat.) ± 4.9 (syst.)	
t$\bar{t}$Z/γ^*	49.8 ± 3.9 (stat.) ± 11.1 (syst.)	2.15 ± 0.24 (stat.) ± 0.44 (syst.)
WZ	9.1 ± 0.9 (stat.) ± 4.0 (syst.)	
Rare SM bkg.	8.8 ± 4.3 (stat.) ± 5.9 (syst.)	0.27 ± 0.16 (stat.) ± 0.19 (syst.)
WWss		
Conversions	5.3 ± 1.2 (stat.) ± 4.0 (syst.)	
Charge mis-meas.		
Non-prompt leptons	30.8 ± 1.5 (stat.) ± 10.9 (syst.)	
All backgrounds	137.3 ± 6.2 (stat.) ± 12.4 (syst.)	2.42 ± 0.28 (stat.) ± 0.56 (syst.)
t$\bar{t}$H signal	19.5 ± 1.0 (stat.) ± 3.0 (syst.)	1.00 ± 0.09 (stat.) ± 0.11 (syst.)
Data	148	3

Figure 5.12: Yields for expected signal and background processes, and observed yields in data, for the 2LSS (top), 3L and 4l (bottom) channels. The predictions for the non-prompt lepton and charge mis-measurement contributions are extracted from data. Yields are shown after a fit to data, with all processes constrained to the SM expectation..

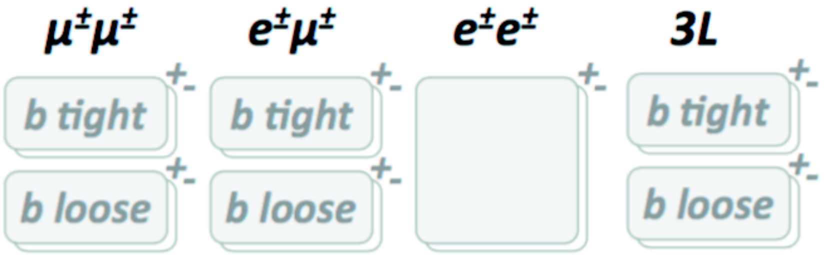


Figure 5.13: Sketch of the event categories. Events are splitted based on the multiplicity of b jets between *b tight* (at least 2 medium b jets) and *b loose* (at least 1 medium or 2 loose b jets). Events are further splitted based on the sign of the sum of charges.

signal.

Further categorization

The events are further split in categories according to the number of medium b-jets and to the sum of lepton charges. (Figure 5.13).

The signal for two and three leptons could be extracted from a fit to the distribution of the number of events shown in Figure 5.14 but the important contribution from fakes and irreducible backgrounds calls for a more sophisticated signal extraction.

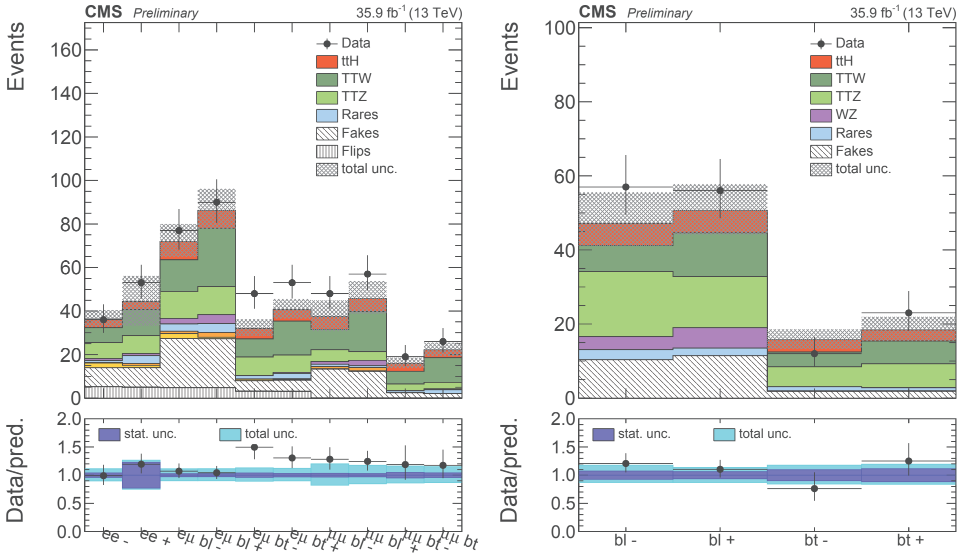


Figure 5.14: Event content in the 2lss (left) and 3l (right) categories.

5.7 Background predictions

Three kinds of background contributions are evaluated. The $t\bar{t}$V irreducible background is estimated from simulation, while both the WZ irreducible background and the reducible backgrounds are estimated from data.

5.7.1 Irreducible backgrounds

Beyond $t\bar{t}$V which is the main contribution to the irreducible background, additional irreducible backgrounds come from diboson production where jet radiation in the final state can lead to a signature similar to the one produced by the signal. The main process contributing to the signal is the WZ production. While $t\bar{t}$V contribution is estimated from MC, the WZ contribution is fitted from data. The WZ contribution is estimated in a WZ control region based on the three lepton selection but adding a b-jet veto (i.e. removing events for which one jet passes the loose working point of the CSV tagger) and reverting the Z veto condition.

The extraction of the WZ yield in the control region is performed via a one dimensional negative log likelihood fit of the shape of transverse mass of the W using the lepton not associated to the Z boson (Figure 5.15). The shape and normalization of the residual backgrounds are fixed to the expectations from simulations. The measurement yields a WZ scale factor of 0.96 ± 0.06.

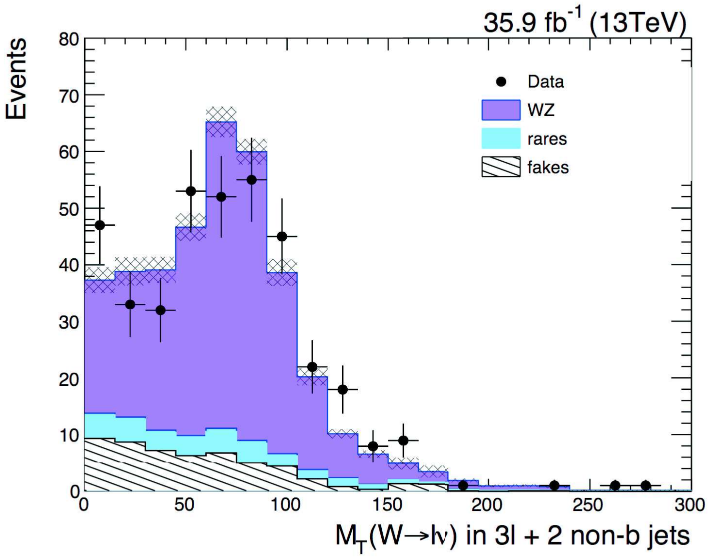


Figure 5.15: Distribution of the transverse mass of the W boson used for the fit.

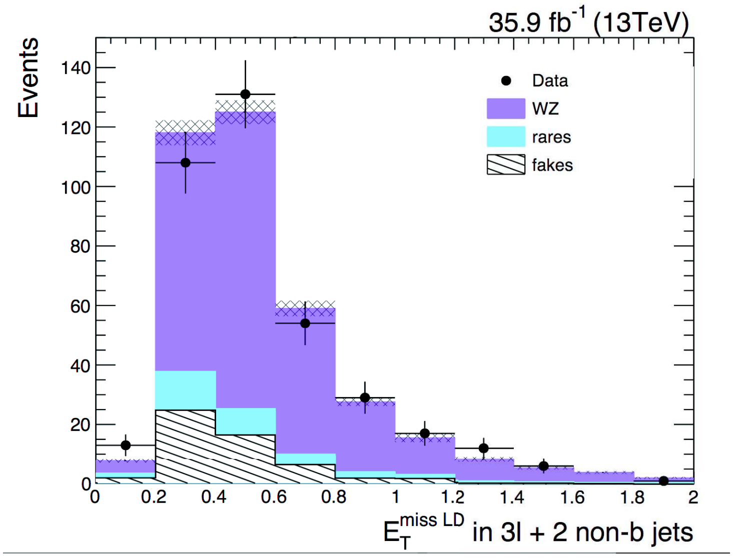


Figure 5.16: Distribution of the linear discriminant of the missing transverse energy in the WZ control region.

The post fit distribution of the variable related to leptons and MET (Figure 5.16) shows a good agreement between data and simulation.

The WZ scale factor can then be propagated to the signal region. The ratio of the WZ event yields between signal region and control region is measured in the simulation. Applying the scale factor to the signal region requires to remove the b-jet veto and the main systematic coming from the application of the scale factor is coming from the b-tagging. Theoretical uncertainties arise from the modelling of the heavy flavour content of the jets in diboson plus multijet events.

The ZZ contribution is estimated from simulation and its contribution to the signal region is much lower.

5.7.2 Reducible backgrounds

Two of the main contributions to reducible backgrounds are the charge misreconstruction of electrons, leading to events with opposite sign leptons to enter in the same sign dilepton category, and the non prompt leptons where either a lepton is produced inside a jet or a hadronic jet is misidentified as a lepton.

Electron charge misassignment

The cross section of processes such as t$\bar{\text{t}}$ and DY+jets (pair of oppositely charged leptons from the decay of a virtual photon or Z boson produced in a quark-antiquark annihilation), being orders of magnitude above the one from t$\bar{\text{t}}$H, they can have a large contribution to the signal region through charge misassignment of one of the leptons. It is thus important to be able to calculate the probability of such a misassignment in order to determine the expected event yield in the signal region coming from this opposite-sign lepton pairs. The charge misassignment is caused by electron bremsstrahlung followed by photon conversion in the tracker material, leading to a wrong reconstruction of the primary electron trajectory.

This probability is derived by a data-driven method looking at same-sign electrons with invariant mass close to the Z mass, expected to come from opposite-sign electron pairs. The charge misassignment probability is calculated from the ratio of same-sign and opposite-sign events in bins of lepton p_T and $|\eta|$. The event yield by bin is estimated by a fit of the invariant mass shape.

The probability of charge misassignment is then applied to events from a control region based on the same selection than the same-sign dilepton but requiring the charge of the two leptons to be opposite. Events will reintegrate the signal region with a weight based on the p_T and $|\eta|$ of the two electrons in the ee channel and of the electron in the eμ channel. While for muons, the probability of such charge misassignment is found to be negligible, the probability of charge misassignment for electron ranges from 0.05% to 0.4% depending on the p_T and $|\eta|$ of the considered electron.

Fake lepton background

The main source of background comes from events containing non-prompt leptons, i.e. originating from semileptonic b-hadron decay or from jets misidentified as leptons. Despite a tight cut on the lepton discriminant aiming at a high purity, the contribution of such events remains important. The determination of the probability for a non-prompt lepton to pass the tight working point of the lepton identification is determined in a control region enriched in multijet events, aiming at rejecting any source of genuine prompt leptons, and applied to weighted events in an application region.

The probability is derived from a control region containing one loose lepton and a hadronic jet separated by $\Delta R > 0.7$. The trigger used is relying on the presence of a non-isolated lepton and a jet. The fake rate will be derived from the probability for the lepton passing the fakeable working point to pass the tight selection. The fake rate derived in this way is based on a lepton passing the trigger requirement while in the signal region, leptons can have failed the trigger. The online identification criteria of the lepton have therefore to be looser than the offline ones in order to avoid any bias. While not being a problem for muons as long as no isolation is applied on the online muon, it can lead to a bias in the measurement for electrons for which the MVA-based identification allows for looser cuts on some parameters while providing a better discrimination between *signal electrons* and *background electrons*. For this reason, a set of cuts is applied to electron to emulate the online selection, leading to a loss of around 3% of the events in both the dielectron and electron-muon final states.

In order to select only fake leptons, the challenge consists in removing events with prompt lepton originating from W and Z production as well as from leptonic decays from $\mathrm{t\bar{t}}$. Z events can be removed by asking for exactly one loose lepton while the W events removal can be achieved through constraints on the E_T^{miss}. The standard variable to be used in order to reject the W contribution is the transverse mass of the W reconstructed from the lepton and E_T^{miss}. However, this variable is using the p_T of the lepton and can thus bias the estimation of the fake rate. By defining a new transverse mass replacing the lepton p_T by a fix value of 35 GeV, the correlation with the lepton p_T, is reduced. Once defining a discriminating variable helping to separate QCD events from W+jet, the contribution from genuine prompt leptons to the control region can be estimated by a fit of the discriminating variable distribution. Once the prompt contribution is removed, the fake rate is estimated using the ratio of the number of leptons passing the fakeable selection and failing the tight selection over all the leptons passing the fakeable selection, in bins of p_T and η.

In the same way as for the probability of charge misassignment, the fake rate is applied to events whose selection is the same as in the signal region, but requiring at least one of the lepton to fail the tight lepton requirement.

As a summary, fake leptons estimation is derived from multijet events, removing all possible sources of prompt leptons, measuring the probability for a fakeable lepton to

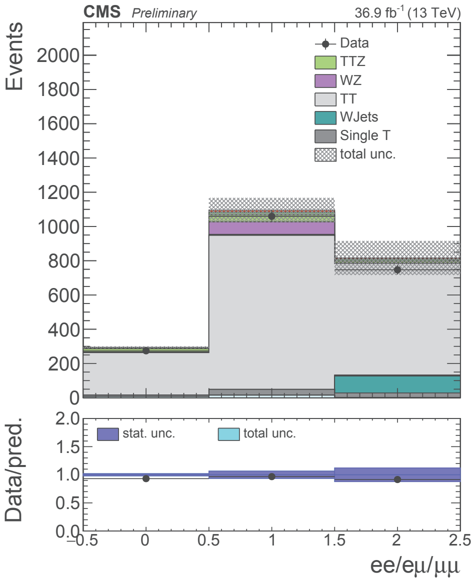


Figure 5.17: Number of events in the 2lss categories (one tight + one fakeable leptons) control region.

pass the tight identification, and applying this probability as a weight in events from an application region corresponding to the selection of the signal region

5.8 Control regions

Beyond the control regions used to derive the charge misassignment of leptons and the fake lepton rate, several control regions have been defined to check our understanding of the lepton identification, the jet multiplicity, and the contribution of the major backgrounds.

Lepton MVA

A first control region targeting the study of the lepton MVA is defined, based on the two leptons selection and requesting one of the two leptons to fail the tight lepton request but pass the fakeable lepton request. This selection is dominated by $\mathrm{t\bar{t}}$ events (Figure 5.17).

Jet multiplicity

A second control region targeting the study of jets is defined, based on the two leptons selection requiring exactly three jets in the final state (Figure 5.18).

WZ → 3l

A third control region is defined in order to validate the three leptons signal region. By inverting the Z veto and requiring that no selected jet satisfies the medium working

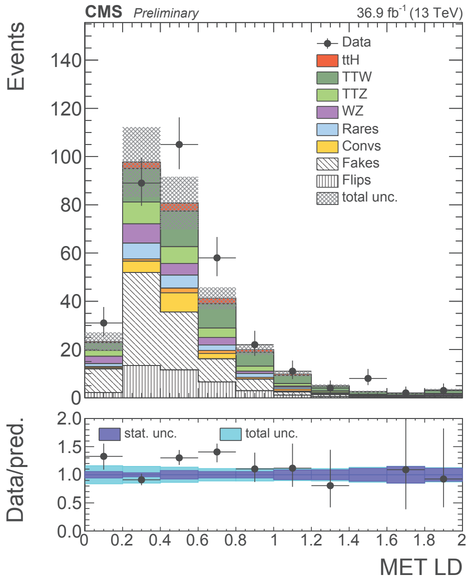


Figure 5.18: Distribution of the missing transverse energy linear discriminant in the 2lss control region with 3 jets.

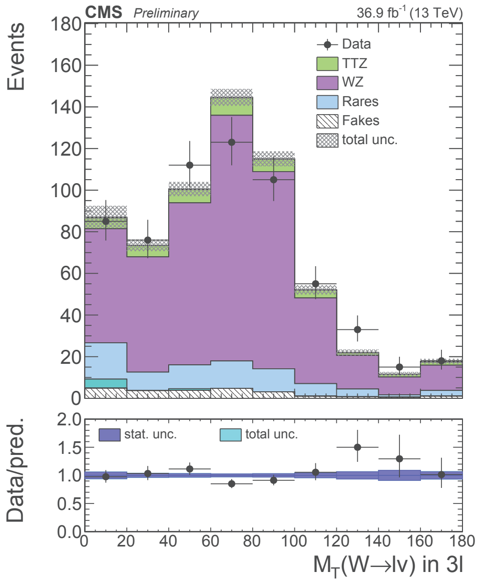


Figure 5.19: Distribution of the W transverse mass in the WZ control region.

point of the CSV b-tagging discriminator, this region is enriched in WZ to 3l events. This region differs from the one used in the derivation of WZ cross-section by the fact that the requirement on the lepton identification criteria is loosened (Figure 5.19).

All distributions show a good agreement between data and simulation.

t$\bar{\text{t}}$Z → **3l**

A fourth control region is defined, requiring two leptons of same flavour opposite charge to fall in a 10 GeV range from the Z mass. By requiring at least two loose and one medium b-tagged jets, the region is enriched in $\mathrm{t\bar{t}Z}$ events. Requiring at least four selected jets increases the purity of the region in $\mathrm{t\bar{t}Z}$ events (Figure 5.20).

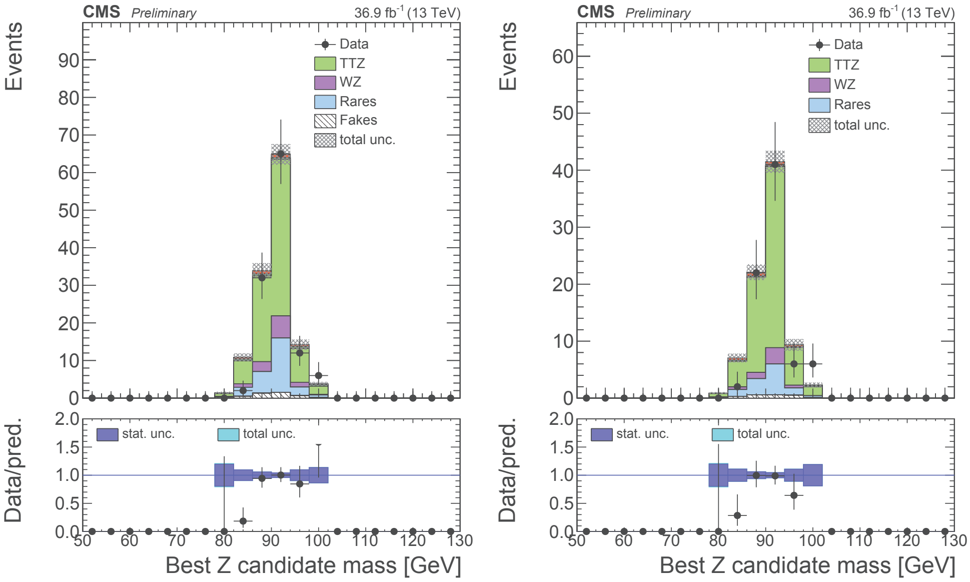


Figure 5.20: Distribution of the best Z candidate invariant mass in the $\mathrm{t\bar{t}Z} \to 3\ell$ control region for >=3 jets (left) and 4 jets (right).

Both the objects and the signal regions having been thoroughly validated in the control regions, the next step aims at the optimization of the discrimination between signal and background.

5.9 Discrimination between signal and background

After the event selection, all remaining events are used to evaluate the signal contribution. The two main backgrounds originate from $\mathrm{t\bar{t}}$ through fakes and from $\mathrm{t\bar{t}V}$ (V=W,Z). Since the signal region remains dominated by the background, a simple fit in the various categories wouldn't yield the best sensivity. Further discrimination between signal and background helps evaluating the signal contribution. To do so, multivariate techniques are used aiming to discriminate between $\mathrm{t\bar{t}H}$ and $\mathrm{t\bar{t}}$, and between $\mathrm{t\bar{t}H}$ and $\mathrm{t\bar{t}V}$.

5.9.1 Discrimination by Boosted Decision Trees

Standard discrimination can be achieved by using a Boosted Decision Tree. Such a multivariate technique relies on the training of a tree based on input variables from labelled events ($\mathrm{t\bar{t}H}$, $\mathrm{t\bar{t}V}$, $\mathrm{t\bar{t}}$). By using different Monte Carlo events for the training and for the application, one guarantees that a possible overtraining (BDT catching unphysical features related to the sample used) does not lead to a bias and would just lower the discriminating power of the method.

The interest of using BDT and such multivariate techniques based on the training of a discriminator is the possibility to use any variable which might help in the discrimination. These variables can be related to the isolation of the lepton, the value of the b-tagging discriminant or the event kinematics. The training can be done on MC at any order in perturbation, provided that the simulated samples exist at this order. However, limits come from the number of events needed for the training, which becomes limited when tight selections are applied, leading to too few simulated events in the signal region. Moreover, a good modeling of the input variables is needed for the discrimination to be performed optimally.

The basic idea behind the use of Decision Trees is that most events do not have clear characteristics of either signal or background. Based on this assumption, the optimal use of events does not rely on a simple cut-based selection in which events are thrown away when found unlikely to be signal.

Creating a tree consists in sorting all events with respect to each input variable, finding the splitting value providing the best separation between signal and background, selecting the variable which maximizes the discrimination and splits the population of events according to the optimal value, producing two branches at which the process is repeated. A criterion can then be defined to stop the process, such as the ratio of signal over background. The tree is then frozen after training and the score of an event passing through the tree can either be the purity ranging from 0 to 1, or can be a binary answer leading to classification in one or the other category.

The boosting part of the algorithm consists in producing several trees in order to get some improvement for outliers. The first method developed consisted in training a first classifier T1, training a second one T2 containing half of events misclassified by T1, training a last one on events for which T1 and T2 disagree. The output classifier was based on the majority of votes in T1, T2, T3. Today, more sophisticated methods are available in order to boost decision trees.

More details will be given concerning the input variables used to achieve discrimination in the various categories and some of the physics motivation for doing so in section 5.9.5. Figure 5.21 illustrates the performances of various MVA based classifiers (columns in the figure) for three different input data configurations (rows in the figure). The classification of blue (signal) and red (background) dots is performed based on a clustering algorithm (column 2), a linear discriminant (column 3) such as the one used for the $E_T^{miss} LD$ introduced previously, based on a simple Decision tree (colum 4), based on a combination

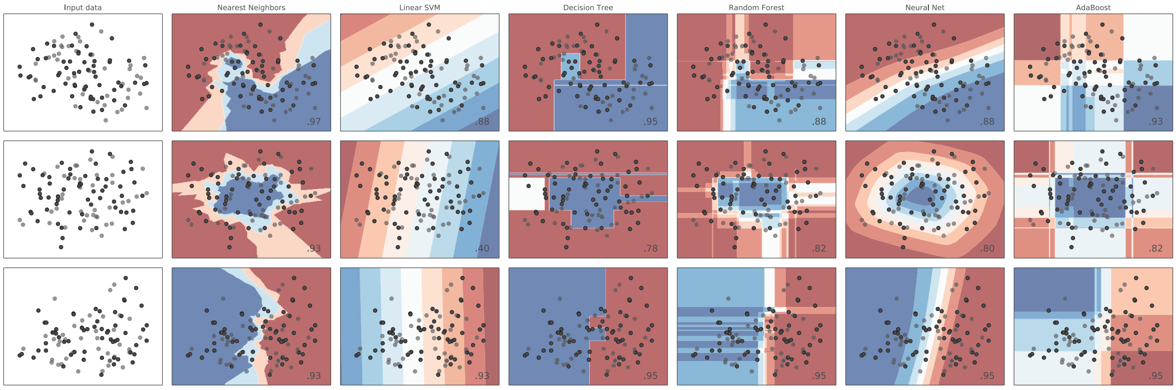


Figure 5.21: MVAs (horizontal axis), and different input data configurations (vertical axis).

of decision trees (column 5) and on a BDT (last column). While the performances of the algorithms might change depending on the tuning of their parameters, this figure shows the way the classification is performed in the plane.

5.9.2 Discrimination through Matrix Element Method

Further discrimination can be achieved by using the Matrix Element Method. While BDTs are based on the training of an algorithm to recognize patterns related to some hypotheses, the Matrix Element Method relies on discrimination based on Leading Order Feynman diagrams of the considered theoretical processes (here t$\bar{\text{t}}$H, t$\bar{\text{t}}$W, t$\bar{\text{t}}$Z, and t$\bar{\text{t}}$).

The main advantages of this method are the absence of need for training, which makes the performances of the method independent from the available statistics in simulation (which will become a serious limitation when moving to possible use of Deep Learning for discrimination between signal and background) and the good discrimination it provides against irreducible backgrounds. The use of this method is however limited because of the computing time it requires and because the matrix element is computed at LO only, higher orders entering in the discrimination in an effective way.

A weight is computed for each of the considered hypotheses:

$$w_{i,\alpha}(\Phi') = \frac{1}{\sigma_\alpha}\int d\Phi_\alpha \cdot \delta^4\Big(p_1^\mu + p_2^\mu - \sum_{k\geq 2} p_k^\mu\Big) \cdot \frac{f(x_1,\mu_F)f(x_2,\mu_F)}{x_1 x_2 s} \cdot \Big|\mathcal{M}_\alpha(p_k^\mu)\Big|^2 \cdot W(\Phi'|\Phi_\alpha)$$

The weight can be decomposed in three parts based on the different inputs they represent in the MEM (Figure 5.22).

First are the variables associated to the incoming particles: $f(x,\mu_F)$ is the parton

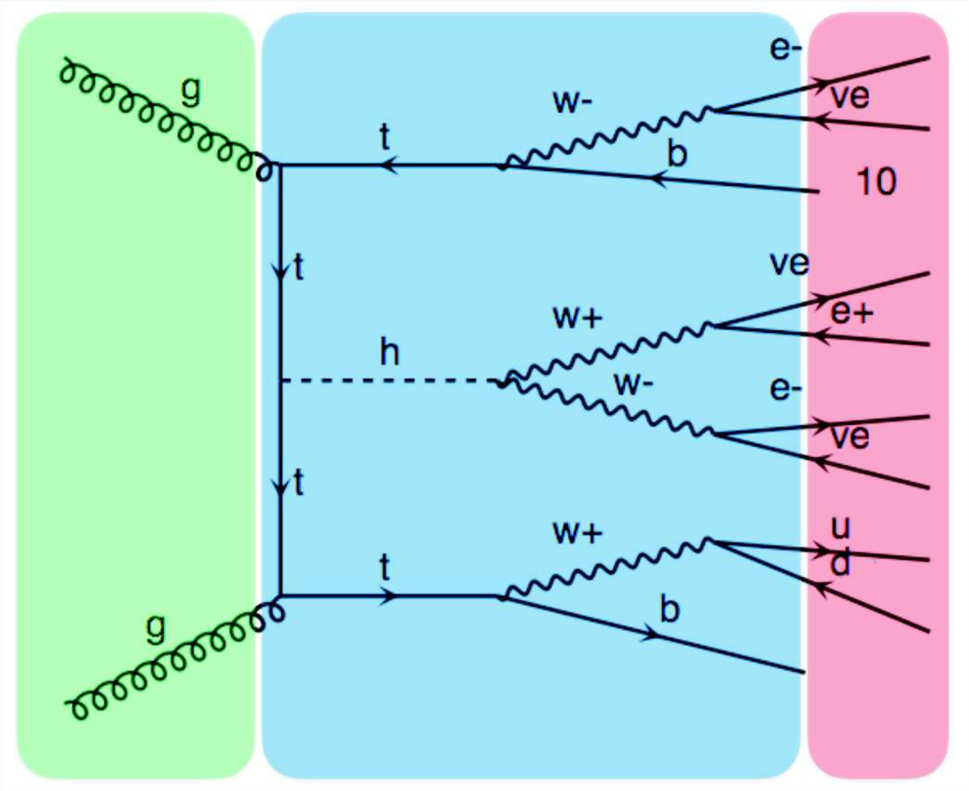


Figure 5.22: Sketch of the Matrix Element Method.

density function of the proton, x_1, x_2 being the fraction of proton energy carried by the partons.

Then comes the part related to the considered process (t$\bar{\text{t}}$H, t$\bar{\text{t}}$W, t$\bar{\text{t}}$Z ou t$\bar{\text{t}}$): σ_α is the cross section of the process α, $\left|\mathcal{M}_\alpha(p_k^\mu)\right|^2$ is the squared matrix element of the process. Φ' is the 4-momenta of the reconstructed particles in the event, $d\Phi_\alpha$ are the process-dependent integration variables, corresponding to the 4-momenta of all the particles at the vertex in the hypothesis α. The Dirac function δ represents the momentum conservation between incoming and final state particles.

Finally, W stands for the transfer functions needed to pass from the energy of ME particles at the vertex to the energy of the reconstructed leptons, jets and b jets.

Before looking at the performances of the MEM, a preliminary study tried to determine the quality of the input provided. The leptons are associated to the ME leptons. For quarks, two options are available. Either the two reconstructed jets with the highest CSV values are assigned to the two b quarks from top decays, or reconstructed jets are assigned to b quarks if they pass the loose working point of the tagger. In the first case, the matching between b quarks and b jets is unique. In the second case, the quark to jet association can run over several combinations.

Leptons and b jets as input to the MEM

One of the challenges of using the MEM in the case of the t$\bar{\text{t}}$H or t$\bar{\text{t}}$V processes is to provide as input all the final state objects (leptons and jets) to achieve an optimal result. For the MEM to perform optimally, all the final state objects should be provided as input. Focusing on the decay of a Higgs to two Ws which is expected to be the main contribution to the signal region, two leptons, six jets among which 2 b jets have to be provided in the same-sign dilepton category and three leptons, four jets among which 2 b jets have to be provided in the three lepton category.

In the three lepton category, the third lepton and second b-jet provided to the MEM are more likely to come from non-prompt leptons (3%) and light jets (21%) respectively than the first two leptons and the highest CSV jet.

While the tight cut on the lepton MVA guarantees a high purity for the third lepton, it appears that in nearly 30% of the cases, the second b jet is originating from non b quark. This is partly related to the selection which allows events with only one jet passing the medium working point of the b-tagger without any requirement on the value of the discriminant of the other jet. Thus it appears interesting to allow the MEM to loop over various matchings between reconstructed jets and generated quarks instead of forcing the two jets with the highest CSV value to be matched with the two b quarks.

In the case where one or two jets are missing from the reconstructed final state, the MEM will adjust the quadri-vector of *ex-nihilo* jets. The constraint on the mass of the W will lead to an adjustement of the quadri-vector of these new jets.

Standalone performances

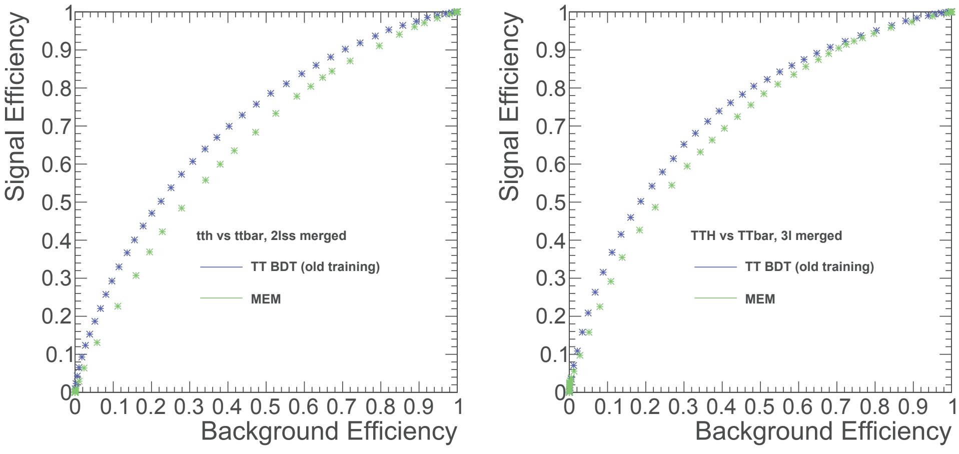


Figure 5.23: Comparison of BDT and MEM performances in the 2l (left) and 3l (right) categories

Figure 5.23 shows the performance of the BDT and MEM in discriminating between t$\bar{\text{t}}$H and t$\bar{\text{t}}$. While some discrimination is observed in both the two lepton and three lepton categories, a better discrimination is achieved by the MEM in the three lepton case because of the easier reconstruction of the event thanks to the third lepton.

5.9.3 MEM and event reconstruction

The Matrix Element Method is able to provide weights associated to each hypothesis which, when combined into a likelihood, leads to a better discrimination between signal and backgrounds. However, based on the quadrimomenta provided for each final state object, the MEM can provide the most likely event reconstruction associated to each hypothesis, thus allowing for the reconstruction of higher level objects such as the t$\bar{\text{t}}$ pair and the Higgs boson, whose kinematics can help discriminating further between signal and background. By maximizing the integrand instead of integrating, we can reconstruct the most probable kinematic configuration.

Focusing on the reconstruction of the t$\bar{\text{t}}$H hypothesis in the three lepton category where the MEM is expected to perform the best, it is possible to study the compatibility of the reconstructed top and Higgs with the generated ones (Figure 5.24, Figure 5.25).

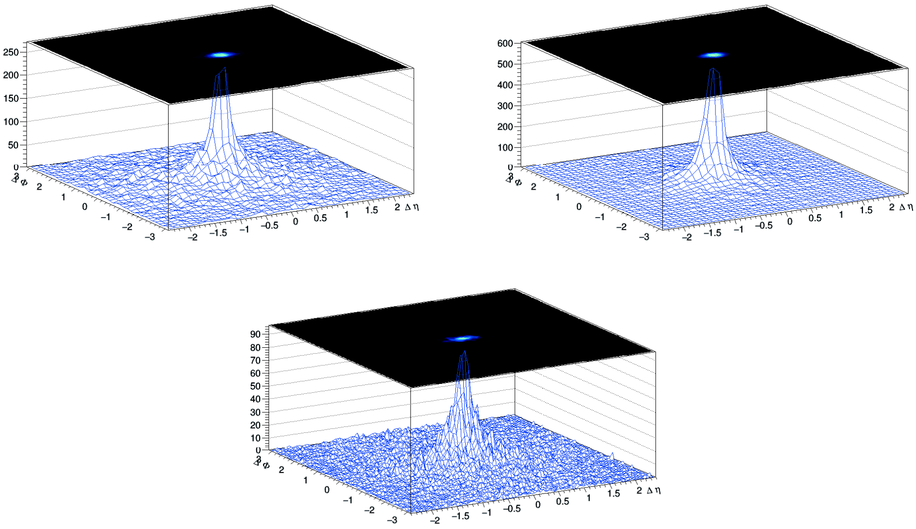


Figure 5.24: $\Delta\eta$ vs $\Delta\phi$ between the generated and MEM reconstructed hadronic top (left), Higgs(center), leptonic top (right).

5.9.4 BDT MEM hybridization

While both BDT and MEM yield some discrimination, a combination of these two strategies can allow a better discriminant, the performances of both techniques being optimal

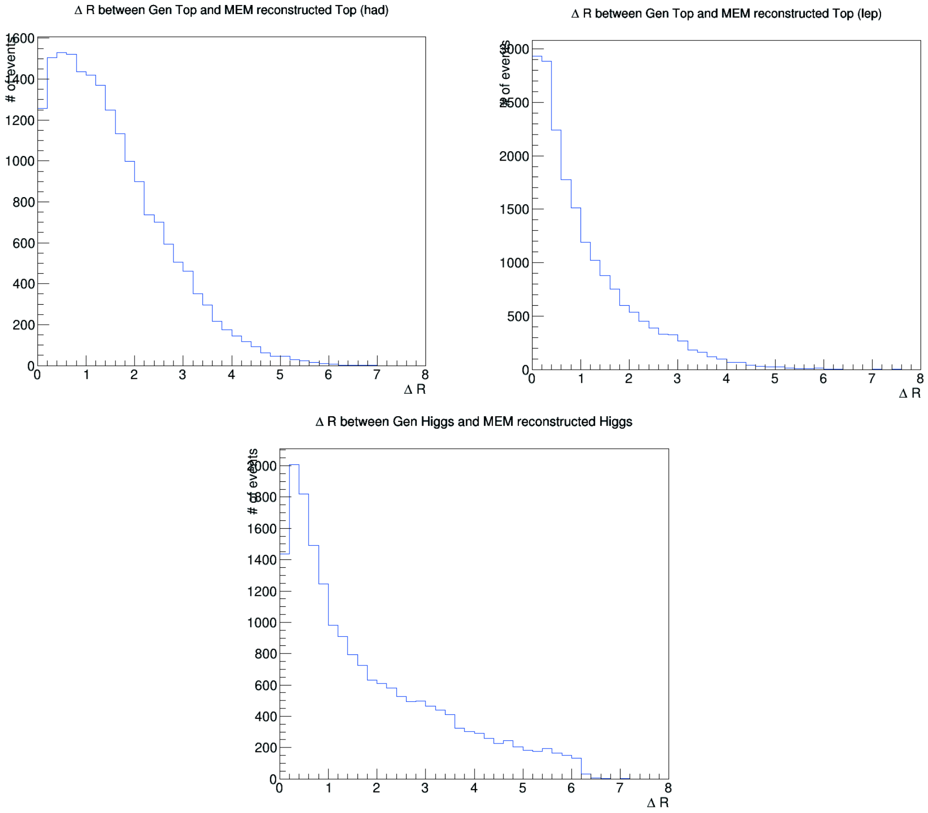


Figure 5.25: ΔR between the generated and MEM reconstructed hadronic top (left), Higgs(center), leptonic top (right).

in different parts of the phase space. Various hybridizations were already tried in the $\text{t}\bar{\text{t}}\text{H}, \text{H} \rightarrow \text{b}\bar{\text{b}}$ analysis. BDT and MEM can be used standalone depending on their respective performances, MEM can be used as an input to the BDT, or categorization based on the BDT output can be used to define a region in which the proper reconstruction of the event will help the MEM discriminating between signal and background. In this analysis, after comparing the performances of the two methods in each category, we decided to introduce variables produced by the MEM into the training of the BDT in order to improve the discrimination.

The first iteration of the analysis making use of the MEM was presented at ICHEP 2016. The MEM weights were then directly used as input to the BDT and led to an improvement of the order of 10% in the discrimination against $\text{t}\bar{\text{t}}\text{V}$ in the three lepton category. This category is the easiest one for the MEM, three leptons providing a good constraint on the system, the two b-tagged jets being associated to the top and the remaining two jets being associated to the W decay from the top or the Higgs.

In the latest iteration of the analysis, following the studies of various combination of MEM outputs as input to the BDT, MEM weights were combined in a likelihood

(Equation (5.2)):

$$\mathcal{L}_{t\bar{t}Hvst\bar{t}V} = -log\left(\frac{\sigma_{t\bar{t}V}w_{t\bar{t}V}}{\sigma_{t\bar{t}H}w_{t\bar{t}H} + \sigma_{t\bar{t}V}w_{t\bar{t}V}}\right) \tag{5.2}$$

This change of variable led to a further improvement of the order of a few percent. While improvements were shown in both two and three leptons (Figure 5.26 and Figure 5.27) categories, the MEM was only used as input in the three leptons category due to the computation time needed to run over the large number of events in the two lepton category. The performances are shown for events with same flavor opposite sign (SFOS) leptons for which the t$\bar{t}$Z hypothesis is relevant and events without same flavor opposite sign (noSFOS) leptons.

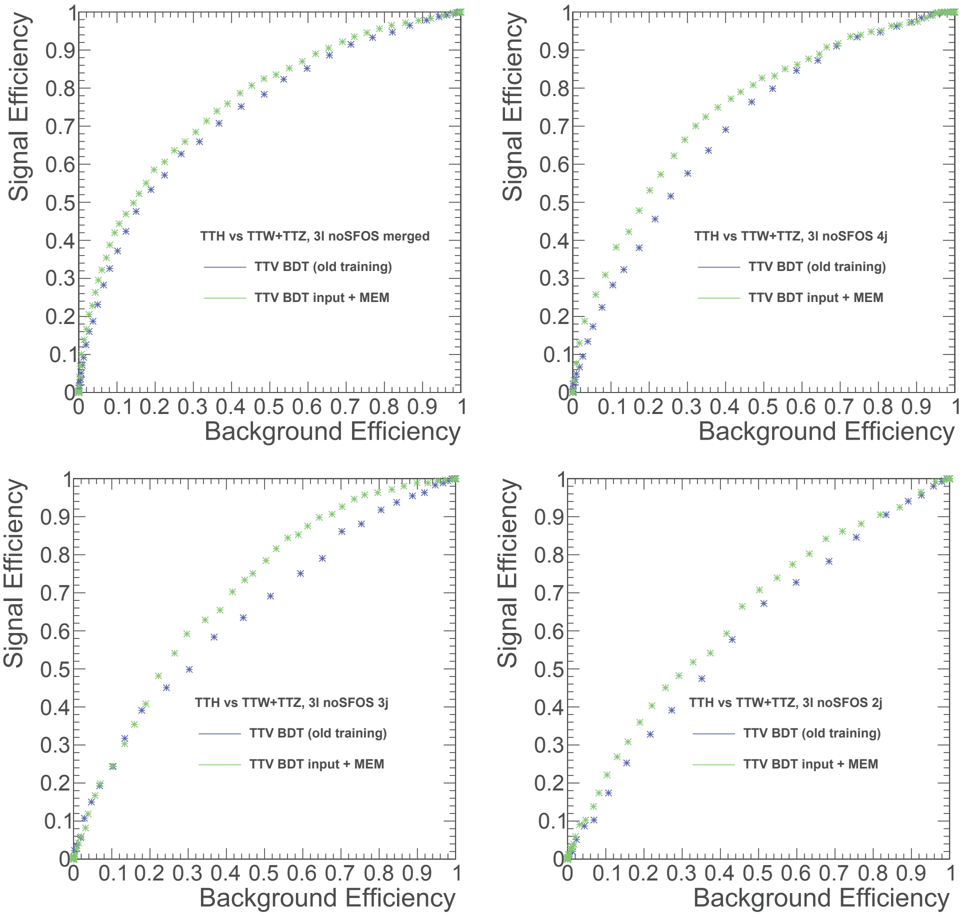


Figure 5.26: Comparison of t$\bar{t}$V BDT and new BDT including MEM in the 3l signal region, without SFOS lepton pair, for (a) all events (b) 0 missing jets, (c) 1 missing jets, (d) 2 missing jets.

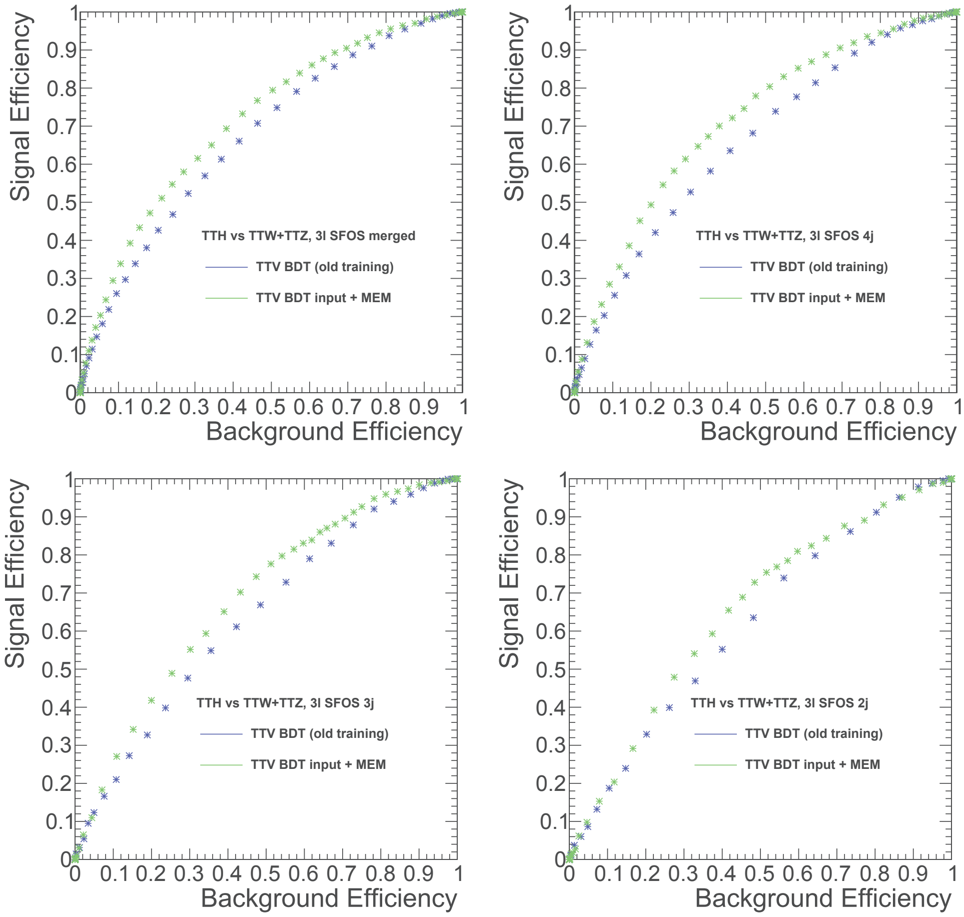


Figure 5.27: Comparison of t$\bar{\text{t}}$V BDT and new BDT including MEM in the 3l signal region, with a SFOS lepton pair, for (a) all events (b) 0 missing jets, (c) 1 missing jets, (d) 2 missing jets.

5.9.5 Discrimination between t$\bar{\text{t}}$, t$\bar{\text{t}}$V vs t$\bar{\text{t}}$H

While the t$\bar{\text{t}}$V contribution to the signal region is irreducible, the t$\bar{\text{t}}$ contribution to the signal region is mostly related to fake leptons. The discrimination between signal and both backgrounds is achieved by using a BDT trained separately in the two lepton and three lepton categories. Beyond the standard kinematic variables, two variables were added recently to the analysis and introduced in the BDT, the hadronic top tagger and the Hj tagger.

The hadronic top reconstruction aims at providing a proper reconstruction of the t$\bar{\text{t}}$H system through the constraint on the hadronic top in order to later help in the discrimination against backgrounds which don't contain hadronic top. In the same sign dilepton category, one lepton is expected to come from the decay of the Higgs, while the second one comes from the leptonic decay of the top, leading to a final state with one leptonic and one hadronic top.

The Hj tagger aims at identifying the hadronic part of the decay of H → WW* → $\ell\nu_\ell jj$. Based on a BDT using jet variables (minimum and maximum ΔR between the jet and one of the lepton, jet p_T, jet CSV and jet quark-gluon discriminator), the tagger allows for a better discrimination between t$\bar{\text{t}}$H and t$\bar{\text{t}}$W.

All variables used in the two lepton and three lepton categories to discriminate the signal against both t$\bar{\text{t}}$ and t$\bar{\text{t}}$V are listed in Table 5.2 and Table 5.3. The variables associated with the MEM (likelihood) were added to the BDT against t$\bar{\text{t}}$V in the three lepton category and have shown improvements to the BDT discrimination in the three lepton category against t$\bar{\text{t}}$ and in the two lepton category against t$\bar{\text{t}}$V.

	2 leptons
Against $t\bar{t}$	maximum absolute pseudorapidity of the two leptons
	multiplicity of hadronic jets
	minimum distance between the leading lepton and the closest jet
	minimum distance between the trailing lepton and the closest jet
	transverse mass of the leading lepton and missing energy
	hadronic top reconstruction
Against $t\bar{t}V$	maximum absolute pseudorapidity of the two leading leptons
	multiplicity of hadronic jets
	minimum distance between the leading lepton and the closest jet
	minimum distance between the trailing lepton and the closest jet
	transverse mass of the leading lepton and missing energy

Table 5.2: Variables used in the two BDTs for the discrimination between signal and background in the two lepton categories.

The 2D plane covered by the 2 BDT against t$\bar{\text{t}}$ and t$\bar{\text{t}}$V is then binned based on the likelihood ratio between signal and background. From a fine splitting of the plane, larger regions are defined for bins presenting a similar ratio of signal to background. The bins are then ordered by increasing signal / background ratio.

5.10 Systematics

The main theoretical uncertainties arise from the uncertainties due to unknown higher orders of the inclusive production cross section of t$\bar{\text{t}}$H, t$\bar{\text{t}}$Z, t$\bar{\text{t}}$W, which are of the order of 10% or more (Table 1.2). These uncertainties are propagated to the final normalization of the event yields. The cross section is fixed to the theoretical predictions.

Another uncertainty comes from the limited knowledge of the parton distribution functions in the proton. This uncertainty is evaluated reweighting the event using the

	3 leptons
Against $t\bar{t}$	maximum absolute pseudorapidity of the two leading leptons multiplicity of hadronic jets minimum distance between the leading lepton and the closest jet minimum distance between the trailing lepton and the closest jet transverse mass of the leading lepton and missing energy leading lepton transverse momentum trailing lepton transverse momentum Hj tagger score after hadronic top jet triplet removal
Against $t\bar{t}V$	maximum absolute pseudorapidity of the two leading leptons multiplicity of hadronic jets minimum distance between the leading lepton and the closest jet minimum distance between the trailing lepton and the closest jet transverse mass of the leading lepton and missing energy (W) leading lepton transverse momentum trailing lepton transverse momentum Matrix Element Method related variables

Table 5.3: Variables used in the two BDTs for the discrimination between signal and background in the three leptons categories.

error sets associated to the PDF.

Finally, the last source of theoretical uncertainty is related to the normalisation and factorisation scales. The impact on the analysis is acertained by varying both scales by a factor two upward and downward. The variation leads to a variation in amplitude of 2 to 4 % in the shape of the classifier.

The first source of experimental uncertainty comes from the integrated luminosity. This uncertainty is of the order of 2%.

Another source of uncertainty is the scale factors derived for the trigger efficiency and lepton selections. For the trigger efficiency, the scale factors are derived using an orthogonal dataset and are found to be less than 3%. The lepton selection efficiency is derived by a Tag-and-Probe method and is of the order of 4%.

The uncertainty associated to the fake yield in the signal region is of the order of 20 to 40%, partially due to the statistical uncertainty in the region used to derive the fake rate, and to the systematic uncertainty from the method.

The last important uncertainty is related to the b-tagging scale factor (explained in chapter 4), derived for each jet with respect to its p_T, η and flavour. The event weight is calculated from the combination of jet weights and the variation of the uncertainty on the scale factor is propagated to the event weight.

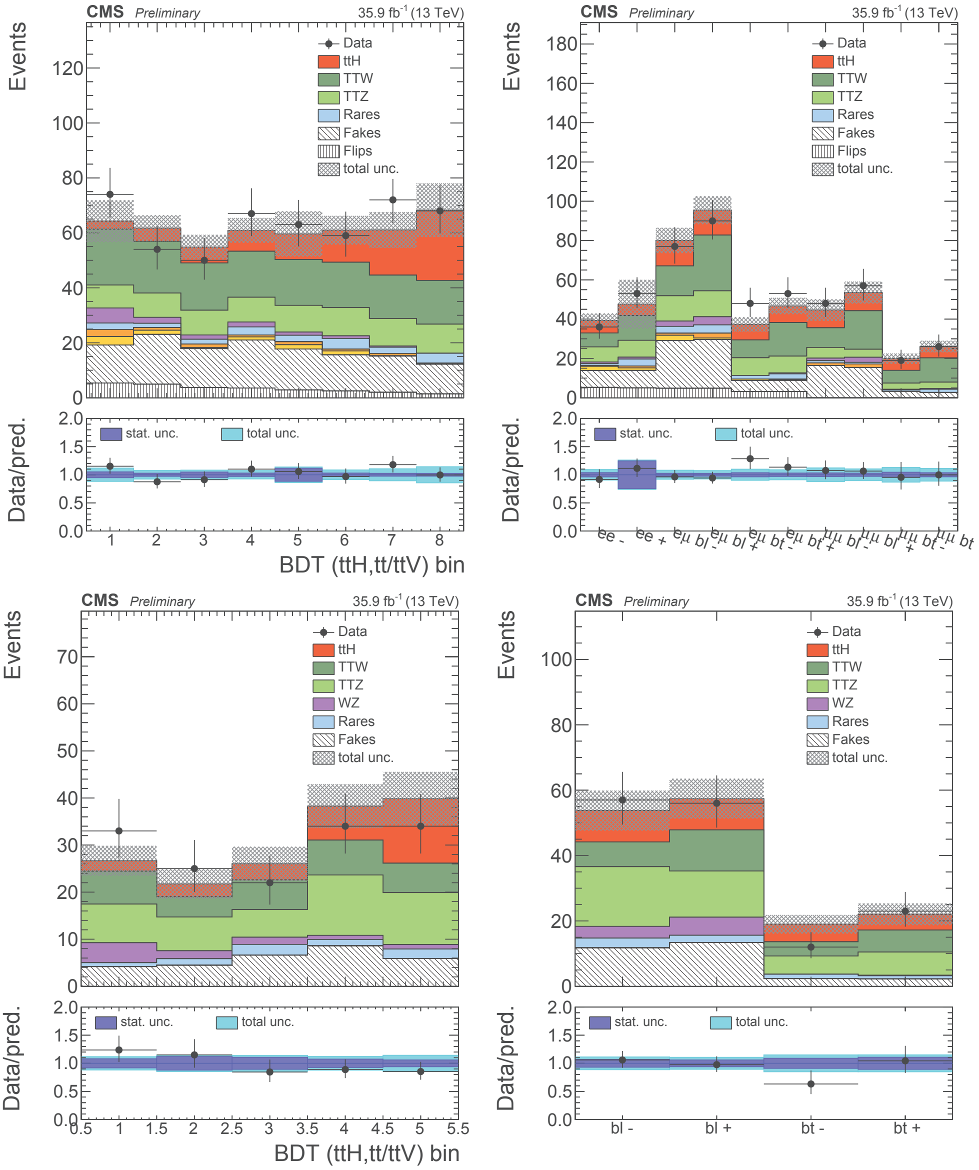


Figure 5.28: Post-fit distributions of discriminating variables and category population for 2lss (top row) and 3l (bottom row) For the categories, refer to Figure 5.13.

The final contributions of statistical, theoretical and experimental uncertainties to the total uncertainty on μ are 0.29, 0.24 and 0.32 respectively.

5.11 Results

The results are summarized in terms of an upper limit on the signal strength modifier (Table 5.4) and of a best fit on the signal strength (Table 5.5). The observed best fit signal strength is $1.5^{+0.5}_{-0.5}$ with an observed significance of 3.3σ (against 2.4 expected). The observed 95% CL exclusion limit on μ is 2.5 in the context of the background-only

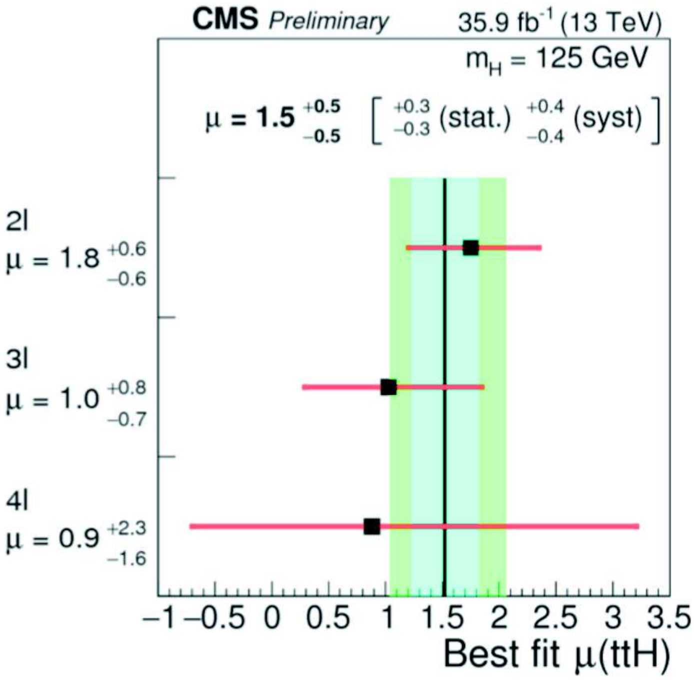


Figure 5.29: Best fit of the signal strength modifier.

hypothesis.

Category	Observed limit	Expected limit $\pm 1\sigma$
same-sign di-lepton	2.8	0.86 (−0.25) (+0.39)
three leptons	2.7	1.34 (−0.41) (+0.64)
four leptons	6.1	4.70 (−1.66) (+2.96)
combined	2.5	0.76 (−0.23) (+0.34)

Table 5.4: Asymptotic 95% CL upper limits on μ under the background-only hypothesis.

Category	Observed μ fit $\pm 1\sigma$	Expected μ fit $\pm 1\sigma$
same-sign di-lepton	1.78 (−0.54) (+0.60)	1.00 (−0.47) (+0.51)
three leptons	1.16 (−0.76) (+0.84)	1.00 (−0.67) (+0.76)
four leptons	1.05 (−1.58) (+2.35)	1.00 (−1.56) (+2.29)
combined	1.56 (−0.48) (+0.54)	1.00 (−0.42) (+0.46)

Table 5.5: Best fit of the signal strength parameter.

5.12 Conclusion

After a preliminary study of the WZ production cross section with the early data from 2015 [60], I began to work on the $t\bar{t}H$ analysis. A first implementation of the object and of the various signal regions was done for Moriond 2016, and I derived the scale factor of the WZ in the three lepton signal region [61]. From there, I moved to the test of the Matrix Element Method as a way to improve the sensitivity of the analysis. The first results showed an initial improvement of the discrimination in the three lepton category against $t\bar{t}V$ and led to the integration of the MEM for ICHEP 2016 [62]. Afterwards, I worked on getting a deeper understanding of the MEM and on providing the optimal input objects. The next step focused on the best output of the MEM to be provided to the BDT. The improvement shown when moving from MEM weights to a likelihood led to the use of this new variable for Moriond 2017 [56]. Despite the improvements shown in other categories, the integration was not possible because of a tight schedule. However, a broader use can be made of the MEM in the future.

With the data taken in 2016, the $t\bar{t}H$ analysis was able to provide the first evidence for $t\bar{t}H$ production. To go further, a look at the sharing of the uncertainties is useful (Table 5.6).

Category	Expected uncertainty on μ
Statistical sources	$(-0.26)\,(+0.27)$
Theoretical sources	$(-0.21)\,(+0.24)$
Experimental sources	$(-0.25)\,(+0.28)$
Total	$(-0.42)\,(+0.46)$

Table 5.6: Split of expected uncertainty in statistical, theoretical and experimental contributions.

Uncertainties are evenly distributed between statistical, theoretical and experimental.

The data samples will increase in 2017 and 2018 during which the expected integrated luminosity could reach 100 fb^{-1}. From a theoretical point of view, the main uncertainty comes from the production cross section contribution from higher orders. From an experimental point of view, the leading uncertainties are related to lepton identification and b-tagging. Both rely on dedicated triggers to perform efficiency and scale factor derivations and the limited number of events of the collected samples might become a limiting factor in the near future.

New ideas could be studied in order to increase the sensitivity of the analysis. Deep learning could provide a better discrimination between $t\bar{t}H$, $t\bar{t}Z$, $t\bar{t}W$ and $t\bar{t}$. In the current form, the analysis does not separate well $t\bar{t}V$ from $t\bar{t}$, focusing on the discrimination between $t\bar{t}H$ signal and backgrounds. However, a better separation between the four components would allow a simultaneous measurement of all the $t\bar{t}$ + boson couplings, including $t\bar{t}H$. Moving to boosted topology would increase the accessible phase space. However, by defining a boosted regime in the analysis, part of the statistics is moving from the resolved unboosted to the boosted category and the gain in sensitivity is not so trivial. The analysis could as well help probing light extended Higgs sector with a charged Higgs mass between the top and standard Higgs masses leading to a signature close from the one of $t\bar{t}H$.

Conclusion

After the two-year technical stop following the announcement of the Higgs boson discovery by the ATLAS and CMS experiments, a new data taking period has begun in 2015.

The new data taking conditions at higher energy and higher luminosity are challenging. The trigger was succesfully adapted to these new conditions both through the development of new algorithms and by using the possibilities offered by the upgrade of the detector. In 2017, the tracking of CMS is benefiting from the installation of a new pixel detector which should soon reach its nominal performances. It will then allow for a better reconstruction of objects such as τ leptons and b quarks which rely heavily on tracking and vertexing. At the same time, new machine learning techniques are used offline and being commissionned online to improve the reconstruction and identification of complex objects through an optimal combination of information provided by all the subdetectors.

One of the physics goal of the Run 2 is to study the properties of the Higgs boson and their compatibility with the predictions of the Standard Model. In a few months, the precision on the couplings of the new boson to bosons and leptons have reached a new level of precision, confirming its compatibility with the standard Higgs boson. While some tensions were observed during the Run 1 between observation in data and the theoretical prediction in the associated production of the Higgs boson with a top quark pair, the data accumulated in 2016 allowed for the first evidence of such a process with a result compatible with the prediction. This analysis is of particular interest, allowing to probe directly the coupling of the Higgs boson to the top quark.

In 2017 and 2018, the integrated luminosity could rise up to 100 fb^{-1}, allowing for more and more precise tests in the electroweak sector. At the same time, the generalization of increasingly complex machine learning techniques and the use of matrix element method seem to offer a way to increase the sensitivity of the analyses.

Bibliography

[1] The ATLAS, CDF, CMS, D0 Collaborations, *First combination of Tevatron and LHC measurements of the top-quark mass.* [March 2014, arXiv:1403.4427]

[2] Max Baak, Roman Kogler, *The global electroweak Standard Model fit after the Higgs discovery.* [19 June 2013, arXiv:1306.0571]

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

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

[5] ALEPH and DELPHI and L3 and OPAL Collaborations and LEP Working Group for Higgs boson searches, *Search for the standard model Higgs boson at LEP* . [Phys.Lett. B565 (2003) 61-75, doi:10.1016/S0370-2693(03)00614-2]

[6] CDF and D0 Collaborations, *Combination of Tevatron searches for the standard model Higgs boson in the W+W- decay mode.* [Phys.Rev.Lett. 104 (2010) 061802, doi:10.1103/PhysRevLett.104.061802]

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

[8] 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), doi:10.1016/j.physletb.2012.08.020, arXiv:1207.7214]

[9] CMS Collaboration, *Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV.* [December 2014, CMS-PAS-HIG-14-009, cds:1979247]

[10] LHC Higgs Cross Section Working Group, *Handbook of LHC Higgs cross sections: 4. Deciphering the nature of the Higgs sector.* [May 2017, arXiv:1403.4427]

[11] CMS Collaboration, *Constraints on the Higgs boson width from off-shell production and decay to Z-boson pairs.* [May 2014, CMS Phys. Lett. B736 (2014) 64, arXiv:1405.3455]

[12] CMS Collaboration, *Evidence for a new state decaying into two photons in the search for the standard model Higgs boson in pp collisions.* [July 2012, CMS-PAS-HIG-12-015, cds:1460419]

[13] CMS Collaboration, *Measurement of differential fiducial cross sections for Higgs boson production in the diphoton decay channel in pp collisions at $\sqrt{s}$=13* TeV. [July 2012, CMS-PAS-HIG-17-015, cds:2257530]

[14] CMS Collaboration, *Properties of the Higgs-like boson in the decay H to ZZ to 4l in pp collisions at* $\sqrt{s}$*=13* TeV. [March 2013, CMS-PAS-HIG-13-002, cds:1523767]

[15] CMS Collaboration, *Measurements of properties of the Higgs boson decaying into four leptons in pp collisions at* $\sqrt{s}$*=13* TeV. [July 2012, CMS-PAS-HIG-14-041, cds:2256357]

[16] CMS Collaboration, *Measurements of properties of the Higgs boson decaying into the four-lepton final state in pp collisions at* $\sqrt{s}$*=13* TeV. [June 2017, CMS-PAS-HIG-16-041, cds:2272260]

[17] CMS Collaboration, *Evidence for a new state decaying into two photons in the search for the standard model Higgs boson in pp collisions.* [July 2012, CMS-PAS-HIG-12-015, cds:1460419]

[18] CMS Collaboration, *Higgs to WW measurements with 15.2 fb*$^{-1}$ *of 13 TeV proton-proton collisions.* [July 2017, CMS-PAS-HIG-16-021, cds:2273908]

[19] CMS Collaboration, *Observation of the SM scalar boson decaying to a pair of* τ *leptons with the CMS experiment at the LHC.* [April 2017, CMS-PAS-HIG-16-043, cds:2264522]

[20] CMS Collaboration, *Search for the standard model Higgs boson produced through vector boson fusion and decaying to bb with proton-proton collisions at* $\sqrt{s}$*=13* TeV. [June 2016, CMS-PAS-HIG-16-003, cds:2160154]

[21] Degrassi et al., *Higgs mass and vacuum stability in the Standard Model at NNLO.* [May 2012, arXiv:1205.6497]

[22] CMS Collaboration, *Search for* $\mathrm{t\bar{t}}$ *resonances in boosted semileptonic finale states in pp collisions at* $\sqrt{s}$*=13* TeV. [March 2016, CMS-PAS-B2G-15-002, cds:2138345]

[23] CMS Collaboration, *Search for supersymmetry using hadronic top quark tagging in 13 TeV pp collisions.* [May 2017, CMS-PAS-SUS-16-050, cds:2262651]

[24] Daniele S. M. Alves et al., *Charged Higgs Signals in* $\mathrm{t\bar{t}H}$ *Searches.* [March 2017, arXiv:1703.06834]

[25] CMS Collaboration, *Technical proposal for the upgrade of the CMS detector through 2020.* [cds:1355706]

[26] CMS Collaboration, *CMS technical design report for the pixel detector upgrade.* [cds:1481838]

[27] CMS Collaboration, *Technical Proposal for the Phase-II upgrade of the CMS Detector.* [cds:2020886]

[28] The CMS Collaboration, *Particle-flow reconstruction and global event description with the CMS detector.* [June 2017, arXiv:1706.04965]

[29] The Aleph Collaboration, *Performance of the ALEPH detector at LEP.* [June 1995, cds:272484]

[30] The CMS Collaboration, *Electron and photon performance in CMS with first 12.9/fb of 2016 data.* [Jul 2016, CMS-DP-2016-049, cds:2203016]

[31] Particle Data Group, *Review of Particle Physics.* [Chin.Phys.C 38 (2014), doi:10.1088/1674-1137/38/9/090001]

[32] The CMS Collaboration, *Performance of tau-lepton reconstruction and identification in CMS.* [JINST 7 (2012) P01001, doi:10.1088/1748-0221/7/01/P01001, arXiv:1109.6034]

[33] The CMS Collaboration, *Reconstruction and identification of τ lepton decays to hadrons and ν_τ at CMS.* [JINST 11 (2016), no. 01, P01019, doi:10.1088/1748-0221/11/01/P01019, arXiv:1510.07488]

[34] CMS Collaboration, *Performance of reconstruction and identification of τ leptons in their decays to hadrons and $\nu\tau$ in LHC Run-2.* [June 2016, CMS-PAS-TAU-16-002, cds:2196972]

[35] CMS Collaboration, *Performance of tau identification with 2016 data at $\sqrt{s}$=13* TeV. [March 2017, CMS-DP-17-006, cds:2255737]

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

[37] D. Bertolini, P. Harris, M. Low, and N. Tran, *Pileup Per Particle Identification.* [JHEP 10 (2014) 059, doi:10.1007/JHEP10(2014)059, arXiv:1407.6013]

[38] CMS Collaboration, *Jet algorithms performance in 13 TeV data.* [March 2017, CMS-PAS-JME-16-003, cds:2256875]

[39] G.Louppe, K. Cho, C. Becot, K. Cranmer, *QCD-Aware Recursive Neural Networks for Jet Physics.* [February 2017, arXiv:1702.00748]

[40] L. Lederman, *Discovery of the Bottom Quark.* [Fermilab News Release]

[41] C. Patrigani et al., *Particle Data Geoup.* [Chin.Phys.C, 40, 1000001 (2016)]

[42] CMS Collaboration, *Inclusive search for the standard model Higgs boson produced in pp collisions at $\sqrt{s}$=13* TeV *using $H \to b\bar{b}$ decays.*. [May 2017, CMS-PAS-HIG-17-010, cds:2266164]

[43] R. Fruhwirth, W. Waltenberger and P. Vanlaer, *Adaptive vertex fitting.* [J.Phys. G34 (2007) N343]

[44] S. Donato, *Search for the Standard Model Higgs boson decaying to b quarks with the CMS experiment.*. [Ph.D, INFN Pisa, 2016]

[45] V. Khachatryan et al., *Measurement of $B\bar{B}$ Angular Correlations based on secondary vertex reconstruction at $\sqrt{s}$=7* TeV.. [February 2011, arXiv:1102.3194]

[46] The CMS collaboration, *Identification of b-quark jets with the CMS experiment.*. [May 2013, 2013 JINST 8 P04013, CMS-PAS-BTV-16-001, arXiv:1211.4462]

[47] The CMS collaboration, *Performance of heavy flavour identification algorithms in proton-proton collisions at 13 TeV at the CMS experiment.*. [May 2017, CMS-DP-2017/013]

[48] CMS Collaboration, *Identification of c-quark jets at the CMS experiment.*. [August 2016, CMS-PAS-BTV-16-001, cds:2205149]

[49] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness.* [JHEP 03 (2011) 015, doi:10.1007/JHEP03(2011)015, arXiv:1011.2268]

[50] CMS Collaboration, *Identification of double-b quark jets in boosted event topologies.*. [July 2016, CMS-PAS-BTV-15-002, cds:2195743]

[51] CMS Collaboration, *CMS Phase 1 heavy flavour identification performance and developments.* [CMS-DP-2017/013, cds:2263802]

[52] CMS Collaboration, *Heavy flavor identification at CMS with deep neural networks.* [CMS-DP-2017/005, cds:2255736]

[53] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC.* [Phys. Lett. B716 (2012) 1-29, doi:10.1016/j.physletb.2012.08.020, arXiv:1207.7214]

[54] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC.* [Phys. Lett. B716 (2012) 30-61, doi:10.1016/j.physletb.2012.08.021, arXiv:1207.7235]

[55] ATLAS and CMS Collaborations, *Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at* $\sqrt{s}$*=7* TeV *and 87* TeV. [J. High Energ. Phys. (2016), doi:10.1007/JHEP08(2016)045, arXiv:1606.02266]

[56] CMS Collaboration, *Search for Higgs boson production in association with top quarks in multilepton final states at* $\sqrt{s}$*=13* TeV. [July 2016, CMS-PAS-CMS-17-004, cds:2256103]

[57] CMS Collaboration, *Measurements of properties of the Higgs boson in the diphoton decay channel with the full 2016 dataset.* [May 2017, CMS-PAS-HIG-16-040, cds:2264515]

[58] CMS Collaboration, *Measurements of properties of the Higgs boson decaying into four leptons in pp collisions at* $\sqrt{s}$*=13* TeV. [May 2017, CMS-PAS-HIG-16-041, cds:2256357]

[59] CMS Collaboration, *Search for ttH production in the H bb decay channel with 2016 pp collision data at* $\sqrt{s}$*=13* TeV. [May 2017, CMS-PAS-HIG-16-038, cds:2231510]

[60] CMS Collaboration, *Measurement of the WZ production cross section in pp collisions at* $\sqrt{s}$*=13* TeV. [December 2015, CMS-PAS-SMP-15-006, cds:2114821]

[61] CMS Collaboration, *Search for ttH production in multilepton final states at* $\sqrt{s}$*=13* TeV. [March 2016, CMS-PAS-HIG-15-008, cds:2141078]

[62] CMS Collaboration, *Search for associated production of Higgs bosons and top quarks in multilepton final states at* $\sqrt{s}$*=13* TeV. [August 2016, CMS-PAS-HIG-16-022, cds:2205282]

Chapter A

CSV variables

Variable	RecoVertex	PseudoVertex	NoVertex
jet p_T	X	X	X
jet η	X	X	X
number of tracks	X	X	X
trackSip3dSig	X	X	X
trackSip2dSigAboveCharm	X	X	X
trackPtRel	X	X	
trackEtaRel	X	X	
trackDeltaR	X	X	
trackPtRatio	X	X	X
trackJetDist	X	X	X
trackDecayLenVal	X	X	X
trackSumJetEtRatio	X	X	X
trackSumJetDeltaR	X	X	X
vertexMass	X	X	
vertexNTracks	X	X	
vertexEnergyRatio	X	X	
vertexJetDeltaR	X	X	
flightDistance2dSig	X		
JetNSecondaryVertices	X		

Table A.1: Variables used in the CSV + IVF algorithm.

Chapter B

Lepton identification in $\mathrm{t\bar{t}H}$ multilepton analysis

Cut	**Loose**	**Fakeable Object**	**Tight**
$\|\eta\| < 2.4$	✓	✓	✓
p_T	> 5	> 15	> 15
$\|d_{xy}\| < 0.05$ (cm)	✓	✓	✓
$\|d_z\| < 0.1$ (cm)	✓	✓	✓
$\mathrm{SIP}_{3D} < 8$	✓	✓	✓
miniIso < 0.4	✓	✓	✓
is Loose Muon	✓	✓	✓
jet CSV	–	< 0.8484	< 0.8484
is Medium Muon	–	–	✓
tight-charge	–	–	✓
lepMVA > 0.90	–	–	✓

Table B.1: Requirements on each of the three muon selections.

Cut	**Loose**	**Fakeable Object**	**Tight**
$\|\eta\| < 2.5$	✓	✓	✓
p_T	> 7	> 15	> 15 2lss(3l)
$\|d_{xy}\| < 0.05$ (cm)	✓	✓	✓
$\|d_z\| < 0.1$ (cm)	✓	✓	✓
$\text{SIP}_{3D} < 8$	✓	✓	✓
miniIso < 0.4	✓	✓	✓
MVA ID $> (0.0, 0.0, 0.7)$	✓	✓	✓
$\sigma_{i\eta i\eta} < (0.011, 0.011, 0.030)$	–	✓	✓
H/E $< (0.10, 0.10, 0.07)$	–	✓	✓
$\Delta\eta_{\text{in}} < (0.01, 0.01, 0.008)$	–	✓	✓
$\Delta\phi_{\text{in}} < (0.04, 0.04, 0.07)$	–	✓	✓
$-0.05 < 1/E - 1/p < (0.010, 0.010, 0.005)$	–	✓	✓
$\text{p}_T Ratio$	–	> 0.5† / –	–
jet CSV	–	< 0.3† / < 0.8484	< 0.8484
tight-charge	–	–	✓
conversion rejection	–	–	✓
Number of missing hits	< 2	$== 0$	$== 0$
lepMVA > 0.90	–	–	✓

Table B.2: Requirements on each of the three electron selections.

Xavier Coubez

Recherche de la production $t\bar{t}H$ dans l'expérience CMS auprès du LHC

Résumé (étendu)

Pendant des siècles, la nature de l'Univers et de ses composants élémentaires a été l'objet de débats philosophiques. Lors du siècle dernier, la naissance de la cosmologie et de la physique des particules a conduit à une nouvelle compréhension du monde que nous habitons. La cosmologie a étendu notre compréhension de l'Univers basée sur la relativité générale alors que la physique des particules proposa dans les années soixante un modèle pour décrire les particules élémentaires et leurs interactions à l'aide de la théorie quantique des champs. Ce modèle, le Modèle Standard fut ensuite testé avec une précision remarquable auprès d'accélérateurs de particules. Toutes les particules prédites furent observées mais la dernière, le boson de Higgs, dut attendre l'avènement du Large Hadron Collider (LHC), le plus grand accélérateur de particules en fonctionnement aujourd'hui. En seulement trois ans d'opération, le boson de Higgs fut découvert et l'annonce en 2012 de son observation vint compléter le modèle.

Depuis cette découverte, la physique des particules se trouve dans une situation inédite. Le modèle qui s'affirma pendant des décennies est maintenant complet. Cependant, des considérations théoriques ainsi que des observations expérimentales indiquent que le Modèle Standard n'est qu'une théorie effective à partir de laquelle une théorie plus générale devrait pouvoir être construite. Tester la validité du modèle est un des buts du LHC mais il est possible qu'aucune nouvelle physique ne soit accessible aux accélérateurs, l'échelle de la nouvelle physique étant à une énergie au-delà de nos possibilités techniques.

Tout en poursuivant la recherche d'une incertaine nouvelle théorie, il est important d'obtenir une meilleure compréhension de la physique du boson de Higgs. Pendant la deuxième période (Run 2) de prise de données qui se déroule depuis 2015 et qui durera jusqu'à la fin de l'année 2018, une des études permettant de tester la validité du Modèle Standard dans le secteur du Higgs est la mesure du couplage du boson de Higgs à la particule la plus massive, le quark top. En raison du rôle joué par le boson de Higgs dans la génération de la masse, le couplage attendu est important. La faible section efficace de production du boson de Higgs avec une paire de quarks top rendait cette étude difficile avec les données prises lors du Run 1, pour une énergie de collision proton-proton de 7-8 TeV. L'analyse bénéficie au Run 2 de l'augmentation en énergie de collision à 13 TeV qui conduit à une augmentation de la section efficace de production, permettant pour la première fois de sonder directement le couplage top Higgs.

Ce document présente une partie du travail effectué dans le cadre de ma thèse, travail portant sur l'étiquetage des jets issus de quarks b dès le déclenchement, ainsi que l'analyse de la production associée d'un boson de Higgs avec une paire de quarks top dans le canal multilepton.

Le premier chapitre introduit le contexte théorique dans lequel s'inscrit le travail effectué. Le Modèle Standard est une théorie quantique des champs, reposant sur deux éléments : les champs et les symétries. Les champs sont associés aux particules et l'évolution des particules est décrite par un lagrangien. Le Modèle Standard peut être décomposé en deux parties, l'une décrivant l'interaction forte, l'autre l'interaction électrofaible. Le premier élément à avoir été décrit dans le cadre de la théorie quantique des champs est l'interaction électromagnétique. Le lagrangien de l'électrodynamique quantique permet la dérivation des diagrammes de Feynman possibles, rendant compte comment les électrons et les photons

peuvent se propager et comment un électron peut rayonner ou absorber un photon. La chromodynamique quantique étend le formalisme précédent à l'interaction forte, basée sur l'échange de gluons colorés et sans masse et sur l'existence de quarks permettant d'expliquer la structure des états des différents mésons et baryons. Enfin, l'interaction faible décrivant la désintégration β, est décrite comme une interaction à courte portée due à l'échange de bosons massifs. L'unification de l'interaction électromagnétique et de l'interaction faible au sein de l'interaction électrofaible constitua un premier succès vers une unification des interactions fondamentales. Afin de donner une masse aux bosons W et Z tout en conservant un photon de masse nulle, un processus de brisure de symétrie était nécessaire. Le mécanisme de Brout, Englert et Higgs permet de briser la symétrie en introduisant un champ scalaire complexe conduisant à l'apparition d'une nouvelle particule, le boson de Higgs.

Découvert en 1995, le quark top est la particule la plus massive du Modèle Standard. Son existence était rendue nécessaire par la découverte du quark b en 1977 afin de conserver une théorie électrofaible cohérente. Produit essentiellement en paire top-antitop, le quark top se désintègre principalement en un quark b et un boson W. Sa masse a été mesurée avec précision et sa section efficace de production déterminée à différentes énergies. Du fait de sa masse élevée, le couplage de cette particule au boson de Higgs est important.

Postulé pour résoudre le problème de la masse des bosons vecteurs de l'interaction électrofaible, le boson de Higgs a été découvert au LHC en 2012. Des contraintes indirectes sur sa masse étaient disponibles suite à sa recherche au LEP et au Tevatron. Depuis sa découverte, l'étude de sa production et de sa désintégration est l'objet d'un intérêt particulier. Quatre modes de production sont possibles : la fusion de gluons, la fusion de boson vecteurs, le Higgsstrahlung (radiation d'un boson de Higgs par un boson W ou Z), et la production associée avec des quarks lourds. Du fait de sa masse de l'ordre de 125 GeV, le boson de Higgs se désintègre essentiellement en paire de quark-antiquark b (57%). Viennent ensuite des modes de désintégration en dibosons, avec un des bosons hors couche de masse, et la désintégration en paire de lepton τ. Le Modèle Standard ne permet pas la désintégration du boson de Higgs en particules sans masse et la désintégration du boson de Higgs en paire de photons n'est possible qu'à travers une boucle de particule lourde.

Le second chapitre introduit le contexte expérimental. La physique des particules repose sur deux éléments principaux. Un accélérateur de particules, ici le Large Hadron Collider, permettant d'obtenir des collisions à hautes énergies et ainsi de produire durant un court instant des particules instables. Un détecteur, tel le Compact Muon Solenoid, permet d'observer ces collisions et d'en déduire les particules produites en reconstruisant les produits de leur désintégration.

Construit dans l'ancien tunnel du LEP, le LHC est un collisionneur proton-proton de 27 km de circonférence, construit pour produire des collisions à une énergie de 14 TeV dans le centre de masse. Après une période de collisions à 7 TeV en 2010-2011, l'énergie dans le centre de masse a atteint 8 TeV en 2012 puis 13 TeV depuis 2015.

Les performances attendues du LHC sont liées au programme de physique couvrant mesures de précision ainsi que possible découverte de nouvelles particules. Afin de permettre cet ambitieux programme, deux paramètres sont d'une importance fondamentale : l'énergie dans le centre de masse des collisions produites et la luminosité. Atteindre une énergie élevée permet de sonder de nombreux signaux dont le seuil en masse est dans la région du TeV. Atteindre une haute luminosité permet d'accumuler suffisamment d'événements pour étudier des processus rares tels que la production du boson de Higgs durant les premières années de prise de données, puis la production associée d'un boson de Higgs avec une paire de quark top pendant les années qui suivent.

Le détecteur CMS (Compact Muon Solenoid) est l'une des deux expériences généralistes opérant au LHC. Sa conception repose sur le choix d'un solénoïde produisant un champ magnétique intense afin d'obtenir une mesure de l'impulsion des particules chargées avec une bonne résolution. Le terme Compact est lié à la taille relativement faible du détecteur et à sa haute densité, nécessaire pour confiner et mesurer l'énergie des particules produites à haute énergie. La taille du détecteur est contrainte par le rayon interne du solénoïde au sein duquel doivent être placés une partie des sous-détecteurs. Le détecteur de muon est placé dans le retour de champ du solénoïde afin de mesurer l'impulsion des muons s'échappant du détecteur.

Les caractéristiques techniques et les performances attendues du détecteur dérivent du programme de physique et des signaux recherchés au LHC. Afin de mener à bien des analyses couvrant un large spectre d'états finaux, le détecteur doit répondre aux contraintes suivantes : permettre une reconstruction efficace et une mesure précise de l'impulsion des particules chargées, offrir une bonne résolution en énergie et en position des dépôts électromagnétiques, une bonne résolution sur l'énergie transverse manquante, une bonne identification des muons avec une bonne résolution en impulsion.

Le détecteur CMS se compose d'un trajectographe au plus proche du faisceau, permettant la mesure de la trajectoire des particules chargées. Il consiste en couches de silicium situées autour de l'axe du faisceau, les premières composées de pixels, les suivantes de pistes. Viennent ensuite les calorimètres. L'objectif des calorimètres est de mesurer l'énergie des particules incidentes. Leur segmentation est choisie pour minimiser l'empilement de deux particules dans la même cellule, ce qui déformerait le signal. Le calorimètre électromagnétique permet de mesurer l'énergie des électrons et des photons qui vont y créer des gerbes électromagnétiques. Le calorimètre hadronique permet la mesure de l'énergie des hadrons et des jets. Enfin, les chambres à muons permettent de mesurer l'impulsion des muons s'échappant du détecteur.

Le troisième chapitre présente rapidement la reconstruction des différents objets au sein du détecteur. Les événements sont reconstruits à l'aide d'un algorithme appelé de "flux de particules" qui combine les informations provenant des différents sous-détecteurs pour identifier des particules stables (électrons, muons, photons, hadrons) et regrouper les hadrons à l'intérieur de jets. Les jets pourraient être reconstruits à l'aide du dépôt en énergie laissé dans les calorimètres, les photons et les électrons identifiés à l'aide de leur dépôt dans le calorimètre électromagnétique, les τ et b identifiés à l'aide des informations du trajectographe et les muons à l'aide des chambres à muons. Cependant, une combinaison des informations provenant des différents sous-détecteurs permet une meilleure reconstruction des particules et conduit à une meilleure détermination de leur énergie et impulsion.

Au LHC, 40 millions de croisements de faisceaux ont lieu chaque seconde, chaque croisement donnant lieu à quelques dizaines de collisions. De ces 40 millions d'événements produits, seuls quelques centaines pourront être sauvegardés. Afin de ne conserver que les événements intéressants pour la physique, une stratégie en deux étapes a été mise en place au sein de l'expérience CMS. L'électronique de déclenchement (L1) va effectuer à l'aide de cartes électroniques une reconstruction grossière des objets et prendre une première décision qui va réduire le nombre d'événements de 40 millions à environ 100 000. Les événements ainsi sélectionnés vont ensuite être envoyés à une ferme d'ordinateurs qui va effectuer des reconstructions plus fines, proches de celles effectuées hors-ligne. Ce second niveau de déclenchement, appelé HLT (High Level Trigger) va réduire le nombre d'événements de 100 000 à quelques centaines qui pourront être sauvegardés. Au niveau L1, il n'est pas possible d'utiliser les informations provenant du trajectographe et seules les informations calorimétriques (dépôts en énergie) et provenant des chambres à muons peuvent servir. Au HLT, les informations provenant de l'ensemble du détecteur sont utilisables et il devient alors possible d'effectuer des reconstructions permettant l'identification des jets issus de quark b. Le b-tagging est un processus d'étiquetage qui utilise les propriétés des jets issus de quark b afin de les distinguer des jets issus de quarks plus légers. Certains événements intéressants attendus au LHC comporteront un nombre important de jets issus de quark b dans un environnement hadronique riche en jets légers. La sélection de tels événements nécessitera une identification des jets de quark b dès le système de déclenchement. C'est le cas par exemple de la production ZH dans laquelle le boson Z se désintègre en deux neutrinos non détectés et le boson de Higgs se désintègre en une paire de quark-antiquark b ou de la production associée de quark top et d'un boson de Higgs se désintégrant en une paire de quark b.

Le travail présenté dans ce document a tout d'abord porté sur l'étude de voies de déclenchement permettant de créer des collections d'événements enrichis en jets issus de quarks b. En utilisant la désintégration fréquente (environ 20%) de quark b en lepton μ dans l'état final, il est possible de créer un échantillon enrichi en requérant la présence de deux jets, l'un contenant un muon. Cette collection d'événements sera ensuite utilisée pour étudier hors ligne l'efficacité d'étiquetage des jets de quark b. Suite au travail effectué et fort d'une première expérience dans l'étiquetage des jets de quark b et le système de déclenchement, la responsabilité de la coordination du groupe chargé de l'étiquetage des jets issus de quarks b au niveau du déclenchement m'a été confiée. Dans ce cadre, j'ai travaillé à la mise en place de la stratégie pour 2016 et 2017, ainsi qu'au suivi des performances des algorithmes déployés pour la prise de données. Le suivi des performances nécessite de porter une attention particulière à l'état des sous-détecteurs utilisés pour effectuer l'étiquetage des jets issus de quarks b afin de pouvoir mofidier les algorithmes en réaction à une perte d'efficacité. La mise en place d'un nouveau sous-détecteur à pixels au coeur de CMS durant l'arrêt technique qui a eu lieu entre 2016 et 2017 a également nécessité l'adaptation des algorithmes afin que les performances d'étiquetage profitent de cette mise à jour. Enfin, la mise en ligne d'un nouvel algorithme fondé sur l'utilisation d'un réseau de neurones profond devrait permettre une amélioration supplémentaire des performances en ligne.

L'analyse principale développée dans ce document porte sur l'étude du couplage du boson de Higgs au quark top par l'observation de la production associée d'une paire de quarks top-antitop avec le boson de Higgs (processus ttH).

La stratégie de l'analyse est d'étudier un état final composé de deux, trois ou quatre leptons (électron ou muon). Le choix de cet état final permet de réduire les bruits de fonds et de sélectionner aisément les événements. Une première partie du travail a porté sur la détermination de la contribution du bruit de fond WZ. Dans la catégorie à trois leptons, le bruit de fond WZ est important et une mesure à partir des données en permet une meilleure estimation. Une mise en place de la sélection des trois leptons ainsi que d'une sélection orthogonale permettant d'enrichir les données en événements WZ a permis d'estimer cette contribution WZ.

La méthode des éléments de matrice (MEM) a été implémentée au laboratoire. Elle permet de mettre à profit la connaissance théorique des processus attendus en basant la discrimination sur les diagrammes de Feynman à l'ordre dominant des processus théoriques considérés (signal ttH et bruits de fond tt ou ttZ/ttW, notés ttV). La mise en place de l'analyse effectuée précédemment a permis les premières études de performances de cette méthode qui a été utilisée dans la dernière itération de l'analyse, présentée à la conférence ICHEP 2016. Par la suite, mon travail a porté sur l'amélioration de la discrimination en

combinant les sorties possibles de la méthode des éléments de matrice avec les variables déjà utilisées dans l'analyse standard pour distinguer le signal des bruits de fond. Des études plus détaillées ont également porté sur la reconstruction complète des événements et l'utilisation de variables cinématiques associées à des particules telles que le quark top et le boson de Higgs qui ne sont détectées qu'à travers leurs produits de désintégration. Ces différentes études ont montré une amélioration de la sensibilité par l'introduction de la MEM pour la discrimination du signal ttH et des bruits de fonds tt et ttV dans la catégorie trois leptons.

Dans le cas à deux leptons, une amélioration de la discrimination entre ttH et ttV a également été obtenue. Dans la dernière itération de l'analyse, la MEM qui nécessite un temps de calcul important n'a été utilisée que dans le cadre de la discrimination entre ttH et ttV dans le cas trois leptons mais son intégration pourrait avoir lieu dans les autres catégories dans lesquelles son efficacité a été démontrée. L'analyse présentée a fourni la première évidence expérimentale du couplage du boson de Higgs au quark top, présentée aux Rencontres de Moriond en Mars 2017.

Les conditions actuelles de prises de données à plus haute énergie et luminosité sont difficiles et ambitieuses. Le système de déclenchement a été adapté avec succès à travers le développement de nouveaux algorithmes et en utilisant au mieux les possibilités offertes par les améliorations du détecteur. En 2017, la trajectographie de CMS bénéficie de l'installation d'un nouveau détecteur à pixels qui devrait atteindre bientôt ses performances nominales. Il permettra alors une meilleure reconstruction des leptons τ et des jets issus de quarks b dont l'identification se fonde sur la trajectographie et la reconstruction de vertex. En parallèle, de nouvelles techniques d'apprentissage automatique telles que les réseaux de neurones profonds sont utilisées hors ligne et testées en ligne pour améliorer la reconstruction et l'identification d'objets complexes à travers une combinaison optimale des informations fournies par l'ensemble des sous-détecteurs.

Un des objectifs des analyses de physique lors du Run 2 est d'étudier les propriétés du boson de Higgs et leur compatibilité avec les prédictions du Modèle Standard. En quelques mois, la précision sur la mesure des couplages du nouveau boson a atteint un nouveau niveau de précision, confirmant sa compatibilité avec le boson de Higgs. Alors que certaines tensions avaient été observées lors du Run 1 entre les données et les prédictions théoriques dans la production associée d'un boson de Higgs avec une paire de quark top, les données prises en 2016 ont permis la première mise en évidence expérimentale d'un tel processus avec une mesure compatible avec la prédiction du Modèle Standard. Cette analyse est d'un intérêt particulier, permettant de sonder directement la couplage du boson de Higgs au quark top.

En incluant les prises de données de 2017 et 2018, la luminosité intégrée pourrait atteindre 100 fb^{-1}, permettant de plus en plus de tests de précision du secteur électrofaible. Dans la même période, la généralisation de techniques d'apprentissage automatique de complexité croissante et l'utilisation de la méthode des éléments de matrice semblent offrir une voie permettant d'augmenter la sensibilité des analyses.

Mot-clés : physique des particules, système de déclenchement, identification de jets issus de quark b, boson de Higgs, couplage top-Higgs