Vector Boson Fusion trigger and study of events featuring di-tau pairs and b jets in the CMS experiment

Chiara Amendola

université PARIS-SACLAY

NNT : 2019SACLX103

Thèse de doctorat

Vector Boson Fusion trigger and search for Higgs boson pair production at the LHC in the bb$\tau\tau$ channel with the CMS detector

Déclenchement pour la production par fusion de bosons vecteurs et recherche de production de paires de bosons de Higgs se désintégrant en bb$\tau\tau$ dans CMS auprès du LHC

Thèse de doctorat de l'Université Paris-Saclay
préparée à École polytechnique

École doctorale n°576 Particules, Hadrons, Énergie, Noyau, Instrumentation, Imagerie, Cosmos et Simulation (PHENIICS)

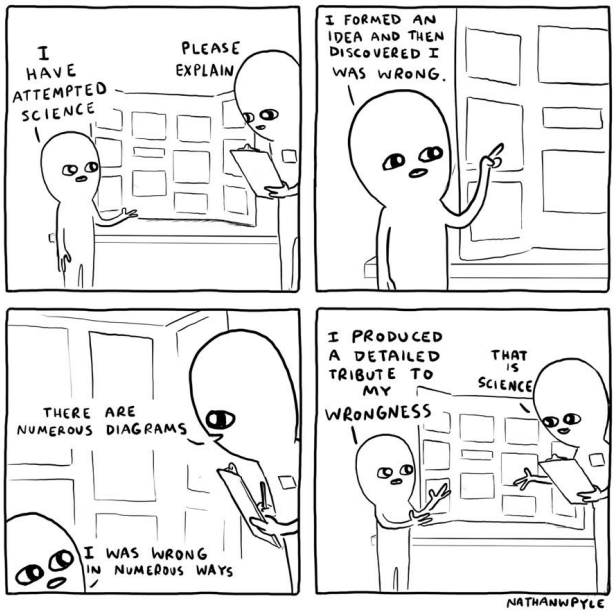
Spécialité de doctorat : Physique des particules

Thèse présentée et soutenue à Palaiseau, le 28 Novembre 2019, par

Chiara Amendola

Composition du Jury :

Pascal Paganini	LLR, Palaiseau	Président du Jury
Isabelle Wingerter-Seez	LAPP, Annecy	Rapportrice
Marco Delmastro	LAPP, Annecy	Rapporteur
Chiara Rovelli	INFN, Rome	Examinatrice
Somnath Choudhury	IISc, Bangalore	Examinateur
Alex Tapper	Imperial College London	Examinateur
Florian Beaudette	LLR, Palaiseau	Directeur de thèse


I HAVE ATTEMPTED SCIENCE
PLEASE EXPLAIN
I FORMED AN IDEA AND THEN DISCOVERED I WAS WRONG.
THERE ARE NUMEROUS DIAGRAMS
I WAS WRONG IN NUMEROUS WAYS
I PRODUCED A DETAILED TRIBUTE TO MY WRONGNESS
THAT IS SCIENCE
NATHANWPYLE

Abstract

This thesis describes a search for events with a pair of Higgs bosons (HH) in proton-proton collisions at 13 TeV, provided by the Large Hadron Collider, with the CMS (Compact Muon Solenoid) experiment at CERN (Geneva).

The study of the Higgs boson pair production allows its trilinear self-coupling (λ_{HHH}) to be measured; moreover, the HH production through Vector Boson Fusion (VBF) gives access to the measurement of the coupling between two Higgs bosons and two vector bosons ($\lambda_{2\mathrm{V}}$). The values of these parameters are particularly sensitive to the existence of physics beyond the Standard Model: even small variations from the values of the couplings predicted by the theory can lead to a large modification of the cross section.

The HH production at the LHC is a very rare process. The production through the main mechanism, by gluon fusion, has a cross section of about 30 fb, followed by the VBF process, which is about 20 times less likely. Therefore, the optimisation of the signal selection is essential. Hence, the first part of the thesis work was devoted to the study of algorithms for the Level-1 (L1) trigger system of CMS and a dedicated algorithm for the VBF process was optimised, targeting possible improvements for the search for HH$\rightarrow$bb$\tau\tau$ events. This is the first VBF algorithm for the L1 trigger system and it was included for the data-taking starting as of summer 2017. The events thus selected are available for the ongoing Higgs boson searches, including the search described in this thesis.

The rest of the thesis work was dedicated to the analysis of events HH$\rightarrow$bb$\tau\tau$ with the data collected in 2017, starting from a comprehensive study of the data-over-simulation agreement and the development of specific corrections for tau leptons. In addition to the inclusive study of the HH$\rightarrow$bb$\tau\tau$ events, specific event categories for the VBF production were included. The study presented in this thesis is the first dedicated measurement for this production mechanism in the context of the HH$\rightarrow$bb$\tau\tau$ analyses: it lead to the measurement of $\lambda_{2\mathrm{V}}$, constrained by the observed data between -0.8 and 2.8 times the theoretical prediction.

Résumé

Cette thèse présente une recherche d'événements avec paires de bosons de Higgs (HH) dans les collisions proton-proton à 13 TeV, fournies par le Large Hadron Collider (LHC), au sein de l'expérience CMS (*Compact Muon Solenoid*) du CERN (Genève). L'étude de la production de paires de bosons de Higgs permet la mesure de la constante d'auto-couplage trilinéaire (λ_{HHH}). En plus, la production HH par fusion de bosons vecteurs (*Vector Boson Fusion* ou *VBF*) donne accès à la mesure de la constante de couplage entre deux bosons de Higgs et deux bosons vecteurs ($\lambda_{2\mathrm{V}}$). La valeur de ces deux paramètres est particulièrement sensible à l'existence de physique au-delà du Modèle Standard : même des faibles variations par rapport aux valeurs des couplages prévus par la théorie peuvent induire un changement important de la section efficace. Cette thèse cible le cas où un des bosons de Higgs se désintègre dans deux leptons tau et l'autre dans deux jets de particules engendrés par des quarks de type beau : cet état final permet de conjuguer une statistique importante garantie par la désintégration en quarks et la pureté de la désintégration en leptons tau. Le lot de données analysé correspond à la prise de données de 2017, qui correspond à environ 45 fb^{-1}, soit une fraction des donnés du Run 2 du LHC qui s'est déroulé de 2016 à 2018.

Cependant, la production de HH au LHC est un processus très rare. La production par le mécanisme principal, de fusion de gluons (gg→HH), a une section efficace d'environ 30 fb, suivie par le processus de VBF HH, lequel est environ 20 fois moins probable. Ainsi, optimiser l'efficacité de la sélection du signal est essentiel. Par conséquent, la première partie du travail de thèse a été dédiée à l'étude d'algorithmes pour le premier niveau du système de déclenchement de CMS (*Level-1* ou *L1 trigger*), en implémentant des sélections similaires à celles qui sont appliquées au niveau de l'analyse finale, et un algorithme dédié au processus VBF a été mis au point en ciblant des possibles améliorations pour la recherche d'événements HH $\rightarrow$ bb$\tau\tau$. Sa topologie, caractérisée par la présence de deux jets reconstruits dans des régions opposées du détecteur, est un levier puissant pour le distinguer des autres processus.

L'algorithme de sélection d'évènements de VBF au L1 a été optimisée utilisant des données prises en 2016, en faisant des projections réalistes pour les conditions de prise de données du 2017. Il sélectionne les événements en exploitant les propriétés cinématiques des jets, indépendamment des caractéristiques de la désintégration du boson de Higgs. D'ailleurs, utilisé en complément des algorithmes classiques, ciblant les produits de désintégration du boson de Higgs, l'algorithme VBF permet d'étendre la couverture de l'espace des phases de manière significative. Il s'agit du premier algorithme VBF pour le système de déclenchement : grâce à ses bonnes performances, il a été inclus dans l'ensemble des sélections pour la prise de données à partir de l'été 2017. Les données ainsi sélectionnées sont accessibles pour les recherches du boson de Higgs en cours et, en particulier, celle qui est présentée dans cette thèse. En effet, la sélection d'événements HH $\rightarrow$ bb$\tau\tau$ en bénéficie largement : il est mesuré que 17% en plus d'événements de

signal sont sélectionnés grâce à l'algorithme VBF.

La suite du travail de thèse a été consacrée à l'analyse d'événements HH $\rightarrow$ bb$\tau\tau$ avec les données collectées en 2017. L'extraction du signal requiert une bonne connaissance des bruits de fond, dont les principaux sont la production de paires de quarks top ($t\bar{t}$) qui est irréductible dans la mesure où il produit une paire de quark b et une paire de leptons taus ; le bruit de fond d'événements multi-jet génériques, où les jets peuvent être identifiés à tort comme leptons taus ; et le bruit de fond de Z $\rightarrow \tau\tau$ où des jets additionnels peuvent être identifiés comme des jets de b. Ce dernier bruit de fond a été étudié en détail pour en améliorer la modélisation. Les variables discriminants entre le signal de gg$\rightarrow$HH et le bruit de $t\bar{t}$ sont exploitées par une méthode multivariée ; son résultat est utilisé pour vérifier la présence de signal. Aucun excès significatif n'est observé ; des limites supérieures à 95% de niveau de confiance sont déterminées pour plusieurs valeurs du couplage λ_{HHH}. Dans le cas de la valeur de λ_{HHH} prédite par le Modèle Standard, la limite supérieure sur la section efficace correspond à environ 20 fois la prédiction théorique ; les valeurs de k_λ, soit le rapport entre le couplage λ_{HHH} et sa valeur dans le Modèle Standard, sont restreintes par les données observées entre -9.1 et 15.4.

En plus de l'étude inclusive des événements de type HH $\rightarrow$ bb$\tau\tau$, des catégories d'événements dédiées à la production par VBF ont été introduites, conçues à partir de la topologie typique du processus et de de l'espace des phases couvert par l'algorithme VBF du système de déclenchement ; des techniques de apprentissage automatique sont utilisées pour mieux rejeter le signal gg$\rightarrow$HH en faveur de celui de VBF HH. L'introduction de telles catégories vise la première mesure dédiée à ce mécanisme de production dans le cadre des analyses de HH $\rightarrow$ bb$\tau\tau$. Des limites supérieures à 95% de niveau de confiance sont déterminées en fonction du couplage $\lambda_{2\mathrm{V}}$: sa valeur est restreinte entre -0.8 et 2.8 fois la prédiction théorique.

Acknowledgements

I have been lucky enough to be able to count on the support of many people during my PhD; as this journey comes to an end, I would like to thank those who helped to set this milestone.

First of all, I would like to thank the president of the jury Pascal Paganini, the referees Isabelle Wingerter-Seez and Marco Del Mastro, and the examiners Chiara Rovelli, Alex Tapper and Somnath Choudhury for their thorough review of this manuscript; the extensive feedback and the curiosity showed towards my work were highly appreciated and motivating.

I am deeply grateful to my supervisor Florian for his dedicated and patient guidance through these (inevitably) intense years. The work presented in this manuscript reflects his long-sighted vision; I owe it to his unfailing support, to the frequent discussions and also to a remarkable efficiency in anticipating troubles.

Even though I was hardly present during the writing process, the kindness and solidarity of colleagues and friends in Laboratoire Leprince-Ringuet did not go unnoticed. Among the others, I would like to thank the administration and the IT teams for their crucial work behind the scenes of a thesis. To all the post-docs and PhD students, I wish the best of luck for the continuation of their careers and for their upcoming theses; in particular, I would like to thank Cristina, Artur and Marina, with whom I shared the most, for their support and for their friendship.

I would like to thank Yves, Roberto and the rest of the CMS team in LLR; I was given many occasions for growth, often at the edge of my comfort zone but always backed up. Thanks to Jean-Baptiste for his precious advice in preparation for my thesis defence.

I was brought into the L1 trigger world by Alex's enthusiasm; I am very thankful for the trust he always showed towards me. On the side of the PhD, I had the pleasure of teaching in his TREX dream team. With him, I also thank my fellow TREX *moniteurs* for the nice time that we spent together.

None of the work on trigger algorithms would have been possible without Olivier's expertise, passion and great patience. Working together was always pleasant and extremely instructive.

I cannot thank Luca enough for his help through all the phases of my PhD. I have found in him a generous senior PhD student first, then a dedicated physicist, and a good friend all along.

I would like to thank the friends from the CMS team in Milano-Bicocca involved in the $HH \to bb\tau\tau$ analysis. Thanks to Giacomo for his support and precious advice. Thanks to Pietro for his insights and pragmatism that so often got me out of dead ends; and for bringing good mood (and candies!) to the lab. Finally, I could not have found a better

match than Francesco as a work partner: despite all the struggle, we always made it through with a laugh.

I am also thankful for the enormous support that I found outside the lab. I am very grateful for the bonds that lasted in spite of the distance and I feel lucky for the friends that I met during these years.

My deepest gratitude goes to Jean-Baptiste, who has been sharing his life with me and gave me strength in the difficult moments. His thoughtfulness was determinant for this achievement and for my happiness.

Infine, ringrazio la mia famiglia, lontana ma mai distante. Questo traguardo lo devo a voi: il vostro sostegno e il vostro orgoglio sono stati la motivazione più forte.

Contents

Introduction

After the discovery of the Higgs boson in 2012 by the CMS and ATLAS collaborations, the Large Hadron Collider (LHC) physics program entered a new phase of searches: on one hand, it called for precision measurements and stress-test of the consistency of the Standard Model of particle physics; on the other hand, it stimulated searches for New Physics.

The Higgs boson, indeed, represents the last tile in the experimental exploration of the Standard Model: through the Brout-Englert-Higgs mechanism, mediated by the Higgs boson, the electroweak symmetry is spontaneously broken, providing the bosons and fermions of the theory with mass; with its discovery, the Standard Model is confirmed to be extremely predictive and self-consistent.

The success of the description of the fundamental interactions given by the Standard Model is accompanied by open questions on our understanding of the Nature: however consistent and precise, it does not incorporate known phenomena such as the abundance of dark matter in the Universe and the large asymmetry between matter and antimatter; it describes only three out of the four known fundamental interactions, as its formalism is not compatible with the general relativity; as for the consistency of the Standard Model itself, the accidental cancellation of divergencies allowing the Higgs boson itself to have a mass at the reach of the energy scale that we can probe appears unnatural. These and other theoretical problems lead to the belief that the Standard Model is part of more general fundamental theories valid at higher scales, which motivates the searches for physics beyond the Standard Model.

The Higgs boson pair production (HH) searches, such as the study described in this thesis, provide a unique independent test of the electroweak symmetry breaking mechanism: the HH production cross section directly depends on the Higgs trilinear self-coupling, whose value is currently indirectly determined by its relation to the mass of the Higgs boson and the vacuum expectation value of the Higgs field. Moreover, the HH production through Vector Boson Fusion (VBF) gives access to the measurement of the coupling between two Higgs bosons and two vector bosons.

At the LHC, the HH production is extremely rare. The production through the main mechanism, the gluon fusion, has a cross section of about 30 fb, followed by the VBF process, which is about 20 times less likely. A precise measurement of the Higgs couplings cannot be accomplished in the near future; however, effects of physics beyond the Standard Model can arise through a modification of the values of the couplings from the Standard Model prediction. The HH searches are particularly sensitive to these effects: even small variations from the values of the couplings predicted by the theory can lead to a large modification of the HH production cross section; in some scenarios, its enhancement is large enough that stringent constraints can be set on anomalous values of the couplings.

This thesis is focused on the scenario where one of the Higgs bosons decays in two tau leptons and the other in two jets of particles generated by b quarks: the study of this final state benefits from the high statistics of the decay in quarks and from the purity of the decay in tau leptons. Both the search for gluon fusion and for VBF production searches are performed.

Given the rarity of the process, the optimisation of the signal selection is essential. A consistent part of the thesis work was devoted to specific strategies for the VBF HH event selection. In the first place, the capability of the CMS Level-1 (L1) trigger system to handle complex correlations between physics objects is exploited to implement online selections similar to those applied at the final analysis stage, targeting the VBF topology. In the second place, an offline selection that discriminates at best the VBF HH signal from the background processes is defined and incorporated in the consolidated inclusive analysis. Finally, a DNN-based technique is implemented to disambiguate the VBF HH signal against the gluon fusion HH production.

This thesis is structured as follow. The theoretical context of the HH production within the Standard Model formulation, the phenomenology of the HH processes and the experimental status are presented in Ch. 1. In Ch. 2, the LHC operations are detailed and a brief description of the CMS detector is given. The design and optimization of the first dedicated L1 VBF trigger selection is detailed in Ch. 3, along with its performance. The complete event selection for the $HH \to bb\tau\tau$ search and the definition of the signal regions are detailed in Ch. 4, and the background and signal modelling are described in Ch. 5. Finally, the results are shown in Ch. 6.

Chapter 1

Theoretical context of the double Higgs boson production

The Standard Model (SM) of particle physics [1] is the theoretical framework that describes the fundamental interactions between elementary particles, formulated as a renormalizable and Lorentz-invariant quantum field theory. It offers a consistent representation of the fundamental interactions, verified experimentally to a high level of precision. Its formulation is briefly described in Sec. 1.1. The phenomenology of the double Higgs boson production is introduced in Sec. 1.2, in the context of searches for SM physics and beyond. Finally, a summary of the experimental status of the double Higgs searches is given in Sec. 1.3.

1.1 The Standard Model

1.1.1 Fields and particle content

Three of the four known forces are incorporated in the SM: the strong interaction, represented as a gauge theory based on a $SU(3)_C$ symmetry; the electroweak interaction, represented as the unification of the weak and electromagnetic force as a $SU(2)_L \times U(1)_Y$ symmetry, where the electromagnetic interaction appears as a residual $U(1)_{em}$ group from the spontaneous symmetry breaking, as clarified in Sec. 1.1.2; the gravitational interaction cannot be accounted for in the SM formulation (besides, it does not play a significant role compared to the others at the sub atomic scale). Together, the strong and electroweak interactions produce a gauge symmetry of the form $SU(3)_C \times SU(2)_L \times U(1)_Y$ [2, 3].

The elementary particles are the ones that, to our knowledge, do not have internal structure; they are classified as fermions and bosons. The former have spin $s = 1/2$ and they are basic constituents of the matter; in turn, they are classified as quarks and leptons. The latter have $s = 1$ and mediate the interactions between fermions.

An overview of the particle content of the SM is given in Fig. 1.1; in the following, the SM formulation of the interactions between particles is summarised.

Fermions

Twelve fermions are experimentally observed and included in the SM. Each of them has an "antiparticle", e.g. a particle with identical mass and opposite quantum numbers; the existence of antiparticles arose from Dirac's equation, describing an electron moving at relativistic speed in a form consistent both with quantum mechanics and special

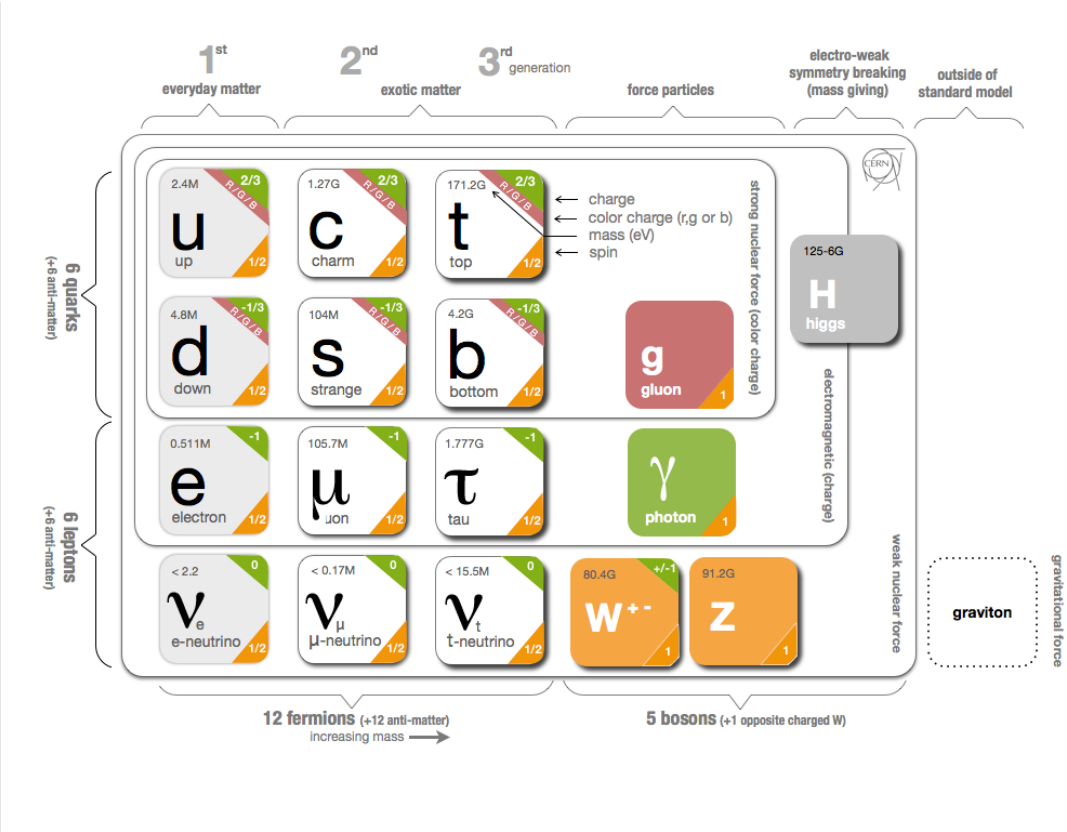


Figure 1.1 – Diagram representing the particle content of the Standard Model [4]. The graviton, represented outside of the Standard Model in this figure, is not observed. The mass hierarchy of the neutrinos is unknown.

relativity: in addition to the solution corresponding to the electron, a solution interpreted as that of an identical positively charged particle exists [5]. Omitting their antiparticles, the known electrically charged leptons are the electron e, the muon μ and the tau lepton τ, with charge $Q = -1$; each is paired to a neutrino (ν_e, ν_μ, ν_τ), electrically neutral, to form a "generation". Their masses span a wide range: the electron has mass $m_e = 511\,\text{MeV}$ and it is the lightest of the charged leptons; the muon and the tau lepton have masses about 200 and 3500 times larger; neutrinos are massless in the classical SM formulation. However, the observation of neutrino oscillations implies that also neutrinos are massive [6] ($m_\nu < 2\,\text{eV}$ [7]); to account for it, a modification is introduced in the SM formalism through the Pontecorvo-Maki-Nakagawa-Sakata (PMNS) matrix [8].

Each quark carries a quantum number called "flavor", subjected to the electroweak interactions; like leptons, the quark flavors are paired in three generations: up (u) and down (d), charm (c) and strange (s), top (t) and beauty (b). Within each pair, the first quark has $Q = 2/3$ and the second $Q = -1/3$. In addition to the other quantum numbers, quarks carry a "color", denoted as $q_i = 1, 2, 3$, on which the strong interaction operates by the rules described by the Quantum Chromodynamics (QCD) theory. Quarks are only experimentally observed as color-neutral bonds, called "hadrons"; this QCD property is known as "confinement". Hadrons have quantum numbers determined by their "valence" quarks: the ones made of a quark-antiquark valence pair are named "mesons" and those made of three valence quarks are called "baryons"; the strong force within hadrons generates a "sea" of virtual quarks and gluons.

The quarks of the first generation are the lightest, with masses of a few MeV. Therefore,

they cannot spontaneously decay via electroweak interaction and their bound states constitute the ordinary matter: protons and neutrons are uud and udd bound states. Like charged leptons, quarks also have masses that differ by several orders of magnitude: the heaviest quark, the top, has a $m_{\mathrm{t}} = 172.9\,\mathrm{GeV}$ mass; thus, it has extremely short lifetime and, before an hadron can be formed, it decays through weak interaction into a W boson and a lighter quark, predominantly b ($|V_{tb}| = 1.010 \pm 0.025$) [7]. Although they are confined in hadrons, quarks are "asymptotically free" particles [9] in high energy collisions, i.e. the strong coupling becomes weaker when the momentum transfer is large; this feature allows the fundamental interactions between them to be investigated in High Energy Physics with proton colliders such as the LHC.

Bosons

The gauge sector of the SM contains gluons, which mediate the strong interaction and correspond to the generators of the $SU(3)_C$ group, and the γ, Z and $\mathrm{W}^{\pm}$ particles, which are the gauge bosons of the $SU(2)_L \times U(1)_Y$ electroweak group. $SU(N)$ groups with dimension N require $(N^2 - 1)$ generators; therefore, $SU(3)_C$ has eight generators corresponding to eight gluons.

Gluons carry color quantum numbers for red, green, blue and antiquarks carry anticolors; therefore, they interact with colored particles as quarks and other gluons. They can have eight different colors and anticolors, they are massless and electrically neutral.

The two $\mathrm{W}^{\pm}$ bosons have identical mass $m_{\mathrm{W}} = 80.3\,\mathrm{GeV}$ [7] and opposite electrical charge $Q = \pm 1$; the Z boson has $m_{\mathrm{Z}} = 91.2\,\mathrm{GeV}$ [7] and is neutral; finally, the photon is massless and has no electric charge.

As will be detailed in Sec. 1.1.2, the $SU(2)_L \times U(1)_Y$ symmetry must be broken to allow the Z and $\mathrm{W}^{\pm}$ bosons to have mass; the Brout-Englert-Higgs mechanism represents a symmetry breaking process that provides the SM with a renormalizable [10] and gauge-invariant Lagrangian that accounts for the mass of the gauge bosons. The existence of a Higgs boson with mass of approximately 125 GeV, mediator of the Higgs scalar field, was discovered by the CMS and ATLAS collaborations in 2012 [11, 12].

Strong interaction

The QCD is a gauge theory that describes the strong interactions, based on a non-abelian Lie group $SU(3)_C$; the "C" subscript denotes the color symmetry. Its Lagrangian density is obtained by applying gauge conditions on the Dirac field, so that it becomes invariant under color transformations.

The Dirac Lagrangian describes free massive particles with $s = 1/2$ such as quarks and electrons, and it has the form

$$\mathcal{L}_f = \overline{\psi}(x)(i\partial\!\!\!/ - m)\psi(x) \tag{1.1}$$

where the notation $\partial\!\!\!/ = \gamma_\mu \delta^\mu$ is used, γ are the Dirac matrices and ψ is the fermionic field. A "global" SU(3) transformation

$$\psi'(x) = U\psi(x), \tag{1.2}$$

i.e. a transformation through a unitary 3×3 matrix U that operates on the initial field in the same way over all the spacetime coordinates, leaves $\mathcal{L}_f$ unchanged. Such transformation can be written as

$$U = e^{ig_s \theta^a \frac{\lambda^a}{2}} \tag{1.3}$$

where g_s is a constant and $\lambda^a/2$ denote the eight Gell-Mann matrices that generate $SU(3)$ rotations.

The global invariance can be promoted to "local", i.e. dependent on the spacetime coordinate x. In quantum field theories, the expedient to achieve local invariance is the introduction of "gauge" fields corresponding to vector bosons; thus, the original free theory turns into a theory that involves interactions between the fermions.

In this case, eight gauge potentials A_a^μ that transform under spacetime-dependent rotations $U(x)$ like

$$A^\mu \to U(A^\mu + \frac{i}{g_s}\partial^\mu)U^\dagger, \tag{1.4}$$

where $A^\mu = A_a^\mu \lambda^a/2$, must be introduced in $\mathcal{L}_f$; to do so, the derivative ∂^μ needs to be replaced with one defined as

$$D^\mu = \partial^\mu + ig_s A_a^\mu \frac{\lambda^a}{2}, \tag{1.5}$$

which transforms covariantly under the gauge transformation, i.e.

$$D^\mu \psi(x) \to D'^\mu \psi'(x) = U(x) D^\mu \psi(x) i. \tag{1.6}$$

Thus, the quark Lagrangian reads

$$\mathcal{L}_q = \overline{\psi}(i\not{D} - m)\psi \tag{1.7}$$

The propagation of the gluon field requires a kinematic term to complete the QCD Lagrangian. Such kinematic term, which also needs to be gauge-invariant, is built by introducing the field strength tensor $F^{\mu\nu}$, or Yang-Mills tensor, as

$$\mathcal{L}_g = -\frac{1}{4} F_a^{\mu\nu} F^a_{\mu\nu} \tag{1.8}$$

with

$$F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu - g[A^\mu, A^\nu]. \tag{1.9}$$

Finally, the complete QCD Lagrangian is obtained from the sum of Eq. 1.7 and Eq. 1.8:

$$\mathcal{L}_{QCD} = \overline{\psi}(i\not{D} - m)\psi - \frac{1}{4} F_a^{\mu\nu} F^a_{\mu\nu}. \tag{1.10}$$

The quark Lagrangian incorporates a term

$$\mathcal{L}_{int}^{QCD} = -g_s \overline{\psi}_j A_a^\mu \frac{\lambda^a_{ij}}{2} \gamma_\mu \psi \tag{1.11}$$

corresponding to the interaction between quarks and gluons through a coupling constant g_s; the interaction with a gluon changes the color of a quark from i to j.

However, the QCD does not only accounts for the interactions between quarks through the mediation of gluons. Indeed, the $\mathcal{L}_g$ term is not linear in terms of the gluon potential, as it contains a three-gluons and a four-gluons term. Thus, gluons also interact with themselves.

Electroweak interaction

The electroweak group represents the unified description of the electromagnetic and weak interactions.

The electromagnetic interaction is described by a gauge theory called Quantum Electrodynamics (QED), whose Lagrangian can be written as

$$\mathcal{L}_{QED} = \overline{\psi}(i\not{D} - m)\psi - \frac{1}{4}F_a^{\mu\nu}F^a_{\mu\nu} \tag{1.12}$$

following the same procedure described for the QCD: starting from the free massive fermion Lagrangian in Eq. 1.1, which represents the electron field, a local invariance under $U(1)_{em}$ is achieved by introducing a gauge potential $A^\mu(x)$ corresponding to the photon; to do so, the derivative is replaced by a covariant derivative, which is

$$D^\mu = \partial - ieA^\mu; \tag{1.13}$$

finally, a kinematic term, also invariant under local $U(1)$ transformations and defined as

$$F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu, \tag{1.14}$$

is introduced to account for the propagation of the photon. As a result, an interaction term

$$\mathcal{L}_{int}^{QED} = e\overline{\psi}A^\mu\gamma_\mu\psi \tag{1.15}$$

arises from the electron Lagrangian, representing the mediation through the photon.

The description of weak interactions requires a more complex structure to account for experimental facts. For instance, it should describe the behaviour of particles of several fermionic flavors, appearing in doublets. Also, it should account for the fact that the $W^\pm$ bosons only interact with left-handed particles or right-handed antiparticles; the property of being left- or right-handed is said "chirality". The unification of weak and electromagnetic interactions under a $SU(2)_L \times U(1)_Y$ group gives a satisfactory representation of these features; however, a gauge symmetry is not suited to motivate the existence of massive mediators such as the $W^\pm$ and Z boson; it is accounted for in the SM through a symmetry breaking mechanism, which will be discussed in the next section.

The weak isospin group $SU(2)_L$ is non-abelian group. The "L" subscript in $SU(2)_L$ denotes that it describes left-handed doublets, while right-handed particles are singlets. For instance, the electron-neutrino pair is

$$\psi_L(x) = \begin{pmatrix}\nu_e \\ e\end{pmatrix}_L; \qquad \psi_R(x) = \nu_{eR}; \qquad \psi'_R(x) = e_R; \tag{1.16}$$

the other fermion doublets are expressed analogously.

The three generators of the transformation are $T_i = \sigma_i/2$, where σ_i denotes the Pauli's matrices.

The hypercharge $U(1)_Y$ group is abelian and it has one generator $Y/2$. Its relation with the $U(1)_{em}$ group is expressed by

$$Q = T_3 + \frac{Y}{2}, \tag{1.17}$$

where Q is the electric charge, T_3 is the weak isospin and Y is the weak hypercharge.

As done earlier with the QCD and QED, the Lagrangian is built by imposing gauge conditions. Using the chirality projectors $P_L = (1-\gamma^5)/2$ and $P_R = (1+\gamma^5)/2$ where $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$, the free Lagrangian of each doublet can be decomposed as

$$\begin{aligned}\mathcal{L}_f = &\overline{\psi_L}(i\not{\partial})\psi_L + \overline{\psi_R}(i\not{\partial})\psi_R + \overline{\psi'_R}(i\not{\partial})\psi'_R + \\ &- m(\overline{\psi_R}\psi_L + \overline{\psi_L}\psi_R) - m'(\overline{\psi'_R}\psi_L + \overline{\psi_L}\psi'_R).\end{aligned} \tag{1.18}$$

Under this form, it is clear that the Lagrangian is invariant under global $SU(2)$ transformations only if the masses m and m' are set to zero: the $SU(2)$ transformation only acts on the doublet fields ψ_L; hence, the mass terms, combining left- and right-handed fields, would spoil the symmetry. Therefore, all fermions must be considered massless at this stage. The $SU(2)_L \times U(1)_Y$ symmetry is promoted from global to local by introducing the covariant derivative

$$D^\mu = \partial^\mu - ig\vec{T}\vec{W}^\mu - ig\frac{Y}{2}B^\mu, \tag{1.19}$$

where $W_i^\mu (i = 1, 2, 3)$ are the fields of the three gauge bosons corresponding to the generators of the $SU(2)_L$ group and B^μ corresponds to $U(1)_Y$; the former are invariant under $SU(2) \times U(1)$ transformations, whereas the latter transforms covariantly. Applied to left- and right-handed fields, the covariant derivative acts like

$$\begin{aligned} D^\mu\psi_L &= (\partial^\mu - ig\frac{\vec{\sigma}}{2}\vec{W}^\mu + ig\frac{1}{2}B^\mu)\psi_L; \\ D^\mu\psi_R = (\partial^\mu + ig'B^\mu)\psi_R; &\quad D^\mu\psi'_R = (\partial^\mu + ig'B^\mu)\psi'_R. \end{aligned} \tag{1.20}$$

Thus, the Lagrangian acquires an interaction term in the form

$$\mathcal{L}_{int}^{EW} = -g\overline{\psi_L}\gamma_\mu\frac{\sigma^i}{2}\psi_L W_i^\mu + ig'\overline{\psi_L}\gamma_\mu\frac{Y}{2}\psi_L B^\mu + g'\overline{\psi_R}\gamma_\mu\frac{Y}{2}\psi_R B^\mu + g'\overline{\psi'_R}\gamma_\mu\frac{Y}{2}\psi'_R B^\mu. \tag{1.21}$$

The four gauge bosons cannot be directly identified as the $W^\pm$, Z and γ bosons of the electroweak interactions, although they are connected. By comparing the currents associated to the first term of Eq. 1.21 to the ones needed to describe the weak interaction, it can be seen that the charged $W^\pm$ bosons are linear combinations of the W_1 and W_2 gauge bosons, defined as

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2); \tag{1.22}$$

analogously, the corresponding Pauli's matrices can be expressed as

$$\sigma^\pm = \frac{1}{\sqrt{2}}(\sigma^1 \pm i\sigma^2). \tag{1.23}$$

The physical Z_μ and A_μ fields, corresponding to the Z boson and the photon, are obtained by applying a rotation by the weak mixing angle θ_w to the W_μ^3 and B_μ fields as

$$\begin{pmatrix} A_\mu \\ Z_\mu \end{pmatrix} = \begin{pmatrix} \cos\theta_w & \sin\theta_w \\ -\sin\theta_w & \cos\theta_w \end{pmatrix} \begin{pmatrix} B_\mu \\ W_\mu \end{pmatrix}. \tag{1.24}$$

Using the relations Eq. 1.22 and Eq. 1.24, three components can be identified in the interaction term of the electroweak Lagrangian (Eq. 1.21):

$$\mathcal{L}_{int}^{EW} = \mathcal{L}_{CC} + \mathcal{L}_{NC}^Z + \mathcal{L}_{NC}^\gamma, \tag{1.25}$$

where "CC" and "NC" stand for "charged current", associated to interactions that modify the electric charge of the particles such as those mediated by $W^\pm$, and "neutral current", with no exchange of charge; individually, they read

$$\begin{aligned} \mathcal{L}_{CC} &= \frac{g}{2\sqrt{2}}\left(W_\mu^+\overline{\psi_L}\gamma^\mu\sigma^+\psi_L + W_\mu^-\overline{\psi_L}\gamma^\mu\sigma^-\psi_L\right) \\ \mathcal{L}_{NC}^Z &= \overline{\psi_L}\gamma^\mu Z_\mu\left(g\frac{\sigma_3}{2}\cos\theta_w - g'\frac{Y}{2}\sin\theta_w\right)\psi_L \\ \mathcal{L}_{NC}^\gamma &= \overline{\psi_L}\gamma^\mu A_\mu\left(g\frac{\sigma_3}{2}\sin\theta_w + g'\frac{Y}{2}\cos\theta_w\right)\psi_L. \end{aligned} \tag{1.26}$$

Finally, $\mathcal{L}_{NC}^{\gamma}$ can be identified, through the relation Eq. 1.17, to the interaction term of the QED Lagrangian (Eq. 1.15), giving

$$e = g' \sin\theta_w = g \cos\theta_w; \tag{1.27}$$

since $\sin^2\theta_w + \cos^2\theta_w = 1$, one also gets

$$e = \frac{g'g}{\sqrt{g'^2 + g^2}}, \quad \text{i.e.} \quad \cos\theta_w = \frac{g}{\sqrt{g'^2 + g^2}} \quad \text{and} \quad \sin\theta_w = \frac{g'}{\sqrt{g'^2 + g^2}}. \tag{1.28}$$

As for the kinematic term, the strength fields are defined as

$$\begin{aligned} B^{\mu\nu} &= \partial^\mu B^\nu - \partial^\nu B^\mu \\ W_i^{\mu\nu} &= \partial^\mu W_i^\nu - \partial^\nu W_i^\mu + g\epsilon^{ijk} W_j^\mu W_k^\nu \end{aligned} \tag{1.29}$$

where ϵ^{ijk} is the Levi-Civita tensor. They give rise, thus, to

$$\mathcal{L}_{kin}^{EW} = -\frac{1}{4} W_i^{\mu\nu} W_{\mu\nu}^i - \frac{1}{4} B^{\mu\nu} B_{\mu\nu}, \tag{1.30}$$

which contains the trilinear (ZW^+W^-, γW^+W^-) and quadrilinear (ZZW^+W^-, $\gamma\gamma W^+W^-$, γZW^+W^-, $W^+W^-W^+W^-$) self-interaction terms.

Thus, a gauge theory that describes the weak interactions and incorporates the QED is achieved; it encompasses the four physical bosons needed to mediate the interactions among fermions, as well as describing the interactions among the bosons themselves. However, none of these particles are given mass, which is a prediction incompatible with the experimental evidence.

1.1.2 Electroweak symmetry breaking

The massless gauge bosons predicted at this stage have two degrees of freedom associated to their two transverse polarizations; a third degree of freedom, corresponding to the mass, must be introduced by breaking the symmetry without spoiling the gauge invariance. The mass terms of the electroweak Lagrangian arise from a so-called "spontaneous symmetry breaking": it occurs when the potential of a system is invariant under a symmetry transformation, while its ground state is not.

In the SM, the spontaneous symmetry breaking is achieved through the Brout-Englert-Higgs (BEH) mechanism [13, 14, 15]. By the Goldstone theorem [16], for every generator of a continuous global symmetry that is spontaneously broken, there is a scalar field term representing a massless spin-0 boson, or a "Goldstone boson". When local gauge conditions are applied to such system, massive bosons are formed by the combination of the gauge and the Goldstone bosons.

Goldstone Lagrangian

The Lagrangian of a complex scalar field $\phi(x) = (\phi_1(x) + i\phi_2(x))/\sqrt{2}$ can be written as

$$\mathcal{L} = \partial_\mu \phi^\dagger \partial^\mu \phi - V(\phi), \tag{1.31}$$

with a scalar potential chosen in a form that is appropriate to generate the spontaneous symmetry breaking, given by

$$V(\phi) = -\mu^2 \phi^\dagger \phi + \lambda(\phi^\dagger \phi)^2, \tag{1.32}$$

where μ is real or purely imaginary. It can be easily verified that the Lagrangian thus defined is invariant under global phase transformations of $U(1)$.

To guarantee the stability of the theory, the potential needs to have a bound from below. In the free case, with $\lambda = 0$, this condition is verified only if $\mu^2 < 0$; otherwise, the condition of stability is $\lambda > 0$, allowing for both the $\mu^2 > 0$ and the $\mu^2 < 0$ scenarios.

In the case of $\mu^2 < 0$, the potential is a concave function of ϕ_1 and ϕ_2, with minimum in $V(\phi_0) = 0$ given by the trivial solution $\phi_0 = 0$. Such minimum is unique and symmetrical under phase transformations; therefore, the system is said to have an "exact symmetry".

The case with $\mu^2 > 0$ is more interesting: the potential still has an extreme in $V(\phi_0 = 0) = 0$, but it corresponds to an unstable local maximum in this scenario; the minimum, instead, is given by the configurations with

$$|\phi_0| = \sqrt{\frac{\mu^2}{2\lambda}} = \frac{v}{\sqrt{2}}. \tag{1.33}$$

Because the Lagrangian is invariant under U(1) phase-transformations, the potential has a shape commonly referred to as a "Mexican hat": it has an infinite number of degenerate states of minimum potential

$$V(\phi_0) = -\frac{\lambda}{4}v^4 \tag{1.34}$$

corresponding to the solutions $\phi_0 = (v/\sqrt{2})e^{i\theta}$ and lying on a circle centered on the origin. Therefore, the system is spontaneously broken: if one choses an arbitrary minimum as ground state, the application of a phase transformation ends up to another point of the circle. The shape of the potential is illustrated in Fig. 1.2.

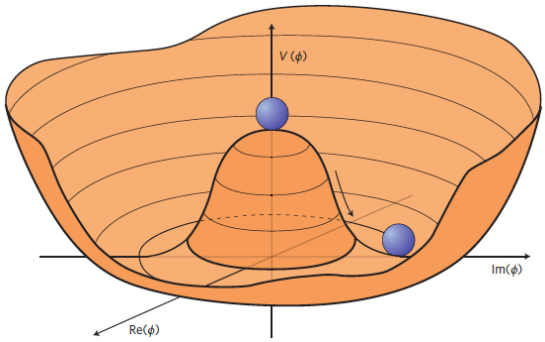


Figure 1.2 – Illustration of the Higgs potential (Eq. 1.32) in the case $\mu^2 > 0$, i.e. with minimum at $|\phi_0| = v/\sqrt{2}$. Choosing any of the points at the minimum of the potential, $V(\phi_0) = -\lambda v^4/4$, the U(1) symmetry is broken [17].

By introducing small excitations to ϕ around the ground state, or the "vacuum expectation value" (VEV), the spectrum of the particles can be determined. Without loss of generality, one can choose the ground state

$$\phi_{1,\text{VEV}} = \frac{v}{\sqrt{2}}, \phi_{2,\text{VEV}} = 0 \tag{1.35}$$

and describe the perturbations around it using the convenient parametrization

$$\phi(x) = \frac{v}{\sqrt{2}} + \frac{\phi_1(x) + i\phi_2(x)}{\sqrt{2}}, \tag{1.36}$$

where $\phi_1(x)$ and $\phi_2(x)$ represent small excitations. By expanding the Eq. 1.36 and inserting it in the original Lagrangian Eq. 1.31, it can be seen that ϕ_1 describes a particle with mass $2\lambda v^2$; the field ϕ_2, instead, is massless and represents an excitation around the flat direction of the potential, i.e. along the curvature of the minimum.

Higgs Lagrangian

In the electroweak case, a field that leads to a spontaneous symmetry breaking while preserving the gauge symmetry of the QED, following the scheme

$$SU(2)_L \times U(1)_Y \to U(1)_{em}, \tag{1.37}$$

should be introduced. To have bosons that are sensitive to the gauge group, a good choice is a weak hypercharge doublet of complex scalar fields

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}. \tag{1.38}$$

with hypercharge $Y_\phi = 1$. The Lagrangian Eq. 1.31 becomes, imposing the gauge conditions allowing for the local symmetry to become global,

$$\mathcal{L}_{BEH} = (D_\mu \phi^\dagger)(D^\mu \phi) - V(\phi) \tag{1.39}$$

with

$$\begin{aligned} V(\phi) &= -\mu^2 \phi^\dagger \phi + \lambda(\phi^\dagger \phi)^2 \quad (\mu^2 > 0, \lambda > 0) \\ D_\mu &= \partial_\mu + ig\frac{\sigma_i}{2} W_\mu^i + ig'\frac{Y}{2} B_\mu. \end{aligned} \tag{1.40}$$

The W_μ and B_μ are the gauge fields corresponding to the generators of the electroweak group and g and g' are the associated coupling constants.

As in the U(1) case, the potential of this system has an unstable local maximum in $\phi = 0$ if $\mu^2 > 0$; its minimum is not symmetric under arbitrary $SU(2) \times U(1)$ transformation and corresponds to a non-zero VEV. Thus, the symmetry is spontaneously broken. It can be seen that, for instance, the choice

$$\phi^+_{\text{VEV}} = 0, \phi^0_{\text{VEV}} = \frac{v}{\sqrt{2}}, \tag{1.41}$$

with notation analogous to that used in the $U(1)$ case, breaks both the $SU(2)_L$ and $U(1)_Y$ symmetries, while preserving the $U(1)_{em}$ symmetry; thus the Eq. 1.37 scheme is achieved. Since three of the four generators are broken spontaneously, by the Goldstone theorem three massless bosons should appear.

The excitations around the ground state can be parametrized as

$$\phi(x) = \exp\left(\frac{i\sigma_i}{2}\theta^i(x)\right)\frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v + H(x) \end{pmatrix} \tag{1.42}$$

where $\theta^i(x)$ and $H(x)$ are four real fields representing small perturbations. By expanding the Eq. 1.42 and injecting it in the Lagrangian of Eq. 1.39, one can see that the $\theta^i(x)$ play the role of the fields of the three massless Goldstone bosons. However, given that the Lagrangian is invariant under local $SU(2)_L$ transformations, one can rotate the system by

$$U(\theta) = \exp\left(\frac{-i\sigma_i}{2}\theta^i(x)\right) \tag{1.43}$$

so that the transformed ϕ' reads

$$\phi \to \phi'(x) = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v + H(x) \end{pmatrix} \tag{1.44}$$

and any unphysical $\theta^i(x)$ dependence is canceled.

Thus, the only field left is $H(x)$ and it correspond to a massive boson, i.e. the Higgs boson (H) that finally closes the SM. Indeed, using the ϕ' parametrization in Eq. 1.42, as well as the physical $W^{\pm}_{\mu}$, Z_{μ} fields introduced in Eq. 1.22 and Eq. 1.24, one gets

$$\mathcal{L}_{BEH} = \frac{1}{2}\partial_{\mu}H\partial^{\mu}H + \frac{g^2v^2}{4}W^{+}_{\mu}W^{\mu-} + \frac{1}{2}\frac{(g^2+g'^2)}{4}v^2Z_{\mu}Z^{\mu} + \frac{1}{2}2\mu^2H^2 + \tag{1.45a}$$

$$+ \frac{g^2}{2}vHW^{+}_{\mu}W^{-\mu} + \frac{g'^2}{2}vHZ_{\mu}Z^{\mu} + \tag{1.45b}$$

$$+ \frac{g^2}{4}H^2W^{+}_{\mu}W^{-\mu} + \frac{g'^2}{4}H^2Z_{\mu}Z^{\mu} + \tag{1.45c}$$

$$+ \frac{\mu^2}{v}H^3 + \frac{\mu^2}{4v^2}H^4 \tag{1.45d}$$

The Eq. 1.45a terms show that the degrees of freedom associated to the $\theta^i(x)$ fields turned into longitudinal degrees of freedom, which allow the $W^{\pm}$ and Z bosons to acquire mass terms

$$m_{\mathrm{Z}} = \frac{\sqrt{g^2+g'^2}}{2}v \quad \text{and} \quad m_{\mathrm{W}} = \frac{gv}{2} = m_{\mathrm{Z}}\cos\theta_w; \tag{1.46}$$

a Higgs boson arises with mass

$$m_{\mathrm{H}} = \sqrt{2}\mu; \tag{1.47}$$

finally, the absence of terms with A_{μ} reflects the fact that the $U(1)_{em}$ is not broken, and the photon remains massless. The interactions among bosons also arise in Eq. 1.45: the trilinear interactions HZZ and $\mathrm{HW^+W^-}$ (Eq. 1.45b); the quadrilinear interactions HHZZ and $\mathrm{HHW^+W^-}$ (Eq. 1.45c); and the trilinear and quadrilinear self-interactions HHH and HHHH (Eq. 1.45d).

The terms belonging to the potential of the BEH Lagrangian can be expressed more conveniently as

$$\begin{aligned} V(\phi) &= \frac{1}{2}(2\mu)^2H^2 + \frac{\mu^2}{v}H^3 + \frac{\mu^2}{4v^2}H^4 = \\ &= \frac{1}{2}m^2_{\mathrm{H}}H^2 + \lambda_{\mathrm{HHH}}vH^3 + \frac{\lambda_{\mathrm{HHHH}}}{4}H^4 \end{aligned} \tag{1.48}$$

where the self-couplings of the Higgs boson are defined as

$$\lambda_{\mathrm{HHH}} = \lambda_{\mathrm{HHHH}} = \frac{m^2_{\mathrm{H}}}{v^2}. \tag{1.49}$$

The Higgs boson self-coupling, thus, is related to the mass of the Higgs boson and it shapes the Higgs potential. As detailed in the following sections, the studies presented in this thesis are fully relevant for testing the electroweak symmetry breaking through the measurement of the trilinear self-coupling involved in the double Higgs production.

As for the interaction with the gauge bosons, the trilinear and quadrilinear couplings can be written under the forms

$$\lambda_{\mathrm{V}} = \frac{2}{v}m^2_{\mathrm{V}} \quad \text{and} \quad \lambda_{2\mathrm{V}} = \frac{1}{v^2}m^2_{\mathrm{V}} \tag{1.50}$$

where "V" stands for "vector" and indicates Z or W, because they correspond to vector fields; in the following, they will be often globally referred to as "vector bosons".

Finally, the mass of the fermions, null until now, is also generated by the Higgs field. Using the notation of Eq. 1.16 accounting for the first lepton generation, their interaction is described by the Yukawa Lagrangian

$$\mathcal{L}_{\text{Yukawa}} = -y_e \left(\overline{\psi}_R \phi^\dagger \psi_L + \overline{\psi}_L \phi \psi_R \right), \tag{1.51}$$

where y_e is the coupling constant between an electron and a Higgs boson. By injecting the scalar doublet with parametrization Eq. 1.44, one has

$$\mathcal{L}_{\text{Yukawa}} = -y_e \frac{v+H}{\sqrt{2}} \left(\overline{e}_R e_L + \overline{e}_L e_R \right); \tag{1.52}$$

re-grouping the electron fields as $\overline{e} = (\overline{e}_R, \overline{e}_L)$, $e = (e_R, e_L)$, the Lagrangian can be written more conveniently as

$$\mathcal{L}_{\text{Yukawa}} = -y_e \frac{v}{\sqrt{2}} \overline{e}e - y_e \frac{H}{\sqrt{2}} \overline{e}e, \tag{1.53}$$

where the mass of the electron $m_e = y_e v/\sqrt{2}$ arises, as well as an interaction term between the electron and the Higgs boson. The same mechanism affects all fermions.

In conclusion, the BEH mechanism allows for both bosons and fermions to acquire masses through the introduction of the Higgs boson scalar field. All the masses are related to a parameter v and to the gauge couplings; the interaction of the Higgs boson with fermions and gauge bosons is also related to the gauge couplings and the masses: as for fermions, the interaction is proportional to their mass, while the interaction with gauge bosons depends quadratically on their mass. As a consequence, the Higgs boson decays preferentially in the heaviest kinematically accessible particles.

1.1.3 Phenomenology and experimental status of the Higgs sector

The BEH mechanism has only two free parameters, both determined experimentally: the VEV and the mass of the Higgs boson.

The numerical value of v is extracted from the charged current interaction in the muon decay $\mu \to e\overline{\nu}_e\nu_\mu$: the transferred momentum is much smaller than m_{W}^2, so that the interaction through the exchange of a W boson can be well approximated through a local four-fermion interaction; thus, the Fermi's constant $G_{\text{F}} = 1.1663788(7) \times 10^{-5}\,\text{GeV}^{-2}$ [18] is determined with high precision and one can identify

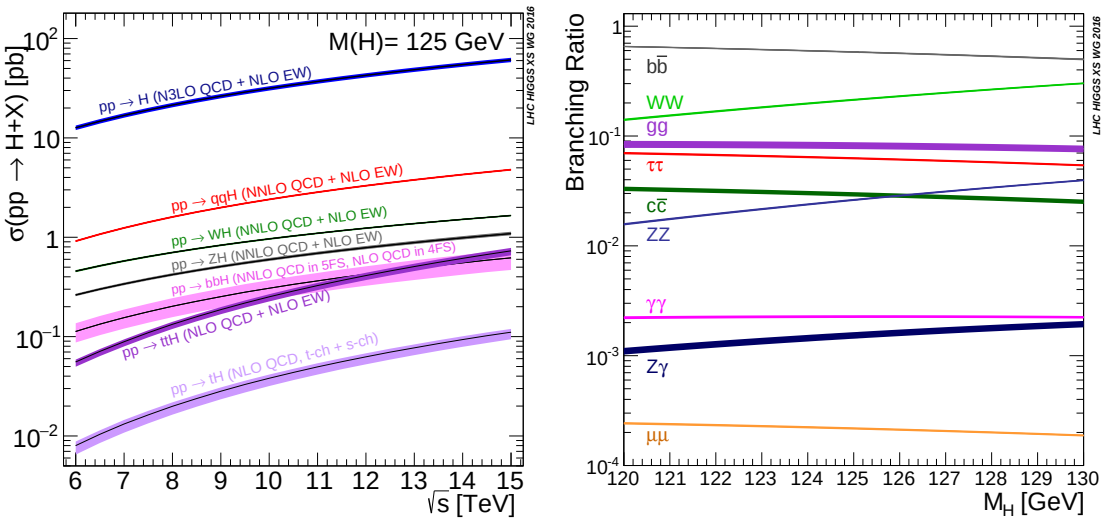
$$\frac{G_{\text{F}}}{\sqrt{2}} = \frac{g^2}{8m_{\text{W}}^2} = \frac{1}{2v^2} \tag{1.54}$$

which gives

$$v = \frac{1}{(\sqrt{2}G_{\text{F}})^{1/2}} = 246\,\text{GeV}. \tag{1.55}$$

As for the mass of the Higgs boson, it was determined by the CMS and ATLAS collaborations through the data analysis of proton-proton collisions with energy in the center of mass $\sqrt{s} = 8\,\text{TeV}$ delivered by the LHC (see Sec. 2.1.4), during the phase of the LHC physic program known as Run 1. The Higgs phenomenology is clarified in the following.

The interaction of the Higgs boson with other particles, as mentioned in Sec. 1.1.2, is predicted in the SM as a function of the values of the free parameters of the theory. Several Higgs boson production modes can occur at the LHC, as shown in Fig. 1.3a for $m_{\text{H}} = 125\,\text{GeV}$, which is approximately the mass of the Higgs boson eventually measured



(a) (b)

Figure 1.3 – On the left, SM Higgs boson ($m_{\mathrm{H}} = 125\,\mathrm{GeV}$) production cross sections in proton-proton interactions as a function of the centre-of-mass-energies; on the right, branching fraction of a Higgs boson of approximately $m_{\mathrm{H}} = 125\,\mathrm{GeV}$ as a function of its mass [19]. The "NLO" acronym for "next-to-leading-order", and similarly "NNLO" and "N3LO", indicates the order of expansion used for the cross section computation in perturbation theory.

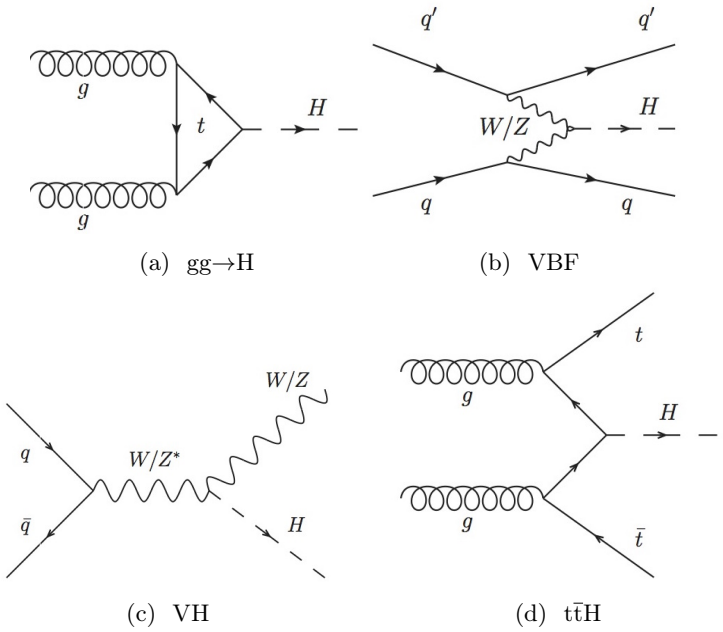
by the two collaborations. The dominant mechanism is the gluon fusion (gg→H), where the Higgs boson is produced via loops of heavy quarks, with leading contribution from the t quark; its cross section is about 21 pb at $\sqrt{s} = 8\,\mathrm{TeV}$ and 49 pb at $\sqrt{s} = 13\,\mathrm{TeV}$. The second-largest contribution, with a cross section about 10 times smaller, is the vector boson fusion (VBF), where the Higgs boson is produced in association with two quarks with large invariant mass. Follows the associate production of Higgs boson with a vector boson (VH). Finally, the Higgs boson can be produced in association with a pair of t quarks ($\mathrm{t\bar{t}H}$) or with a single t quark (tH). The Feynman diagrams representing the main production modes, at leading order in perturbative expansion, are shown in Fig. 1.4.

The branching fractions (or branching ratios, BR) of the Higgs boson decay are shown in Fig. 1.3b as a function of the Higgs boson mass and summarised in Tab. 1.1 for a Higgs boson of $m_{\mathrm{H}} = 125\,\mathrm{GeV}$. The decay to pairs of b quarks has the largest BR, corresponding to about 58% for $m_{\mathrm{H}} = 125\,\mathrm{GeV}$; experimentally, thanks to the excellent invariant mass resolution of the final state objects, the most convenient decay modes are given by the H→ $\gamma\gamma$ decay and the H→ZZ*→ $\ell^+\ell^-\ell^+\ell^-$ ($\ell = e, \mu$), in spite of their very small cross section (about 0.2 and 0.01%). These are, indeed, the two final states with the highest sensitivity in the combination of searches that lead to the observation of a scalar boson compatible to the Higgs boson prediction, announced in July 2012 [11, 12]. As for the production mode, the searches were performed inclusively.

A precise measurement of the mass of the Higgs boson, commonly quoted in the Higgs physics searches, is given by the combination of the CMS and ATLAS H→ $\gamma\gamma$ and H→ZZ*→ $\ell^+\ell^-\ell^+\ell^-$ searches performed in Run 1 [20], i.e.

$$m_{\mathrm{H}} = 125.09 \pm 0.21\ (stat.) \pm 0.11\ (syst.)\,\mathrm{GeV}; \tag{1.56}$$

the most precise measurement up to date, obtained by combining the CMS results in the H→ $\gamma\gamma$ and H→ZZ*→ $\ell^+\ell^-\ell^+\ell^-$ channels in Run 1 and 2016 Run 2 data, gives $m_{\mathrm{H}} = 125.35 \pm 0.12\ (stat.) \pm 0.9\ (syst.)\,\mathrm{GeV}$ [21].



(a) gg→H (b) VBF

(c) VH (d) tt̄H

Figure 1.4 – Main Higgs boson production modes in proton-proton collisions; their cross sections at the LHC are summarised in Fig. 1.3.

Table 1.1 – Branching fraction of the main Higgs boson decay modes in the SM scenario with $m_{\mathrm{H}} = 125.09\,\mathrm{GeV}$. The theoretical uncertainties take into account for missing higher-order corrections to the partial widths, the uncertainties on the mass of the quarks and on the value of α_s [19].

Decay mode	Branching ratio [%]
H→bb	$58.09^{+0.72}_{-0.73}$
H→WW*	21.52 ± 0.33
H→gg	8.18 ± 0.42
H→ $\tau\tau$	6.27 ± 0.10
H→cc	$2.88^{+0.16}_{-0.06}$
H→ZZ*	2.641 ± 0.040
H→ $\gamma\gamma$	0.2270 ± 0.0047
H→ Zγ	0.1541 ± 0.0090
H→ $\mu\mu$	$0.02171^{+0.00036}_{-0.00037}$

The mass measurement is compatible across many different searches; however, the confidence in the fact the observed particle is actually the Higgs boson predicted by the BEH mechanism comes from additional measurements. Firstly, the interaction of the Higgs boson with fermions and vector bosons must be, respectively, linearly and quadratically proportional to their masses. As shown in Fig. 1.5, the measured couplings are indeed observed to have the predicted dependency on the mass of the particles over a wide range. Secondly, it was shown that the observed Higgs boson is a spin-0 and CP-even particle, as predicted in the SM [22, 23].

The observation of the Higgs boson opened a phase of exploration of the electroweak symmetry breaking sector focused on testing its consistency and on precision measurements of the characteristics of the Higgs boson [26]. In this context, the measurement of the Higgs boson self-coupling is one of the goals of the LHC physics program.

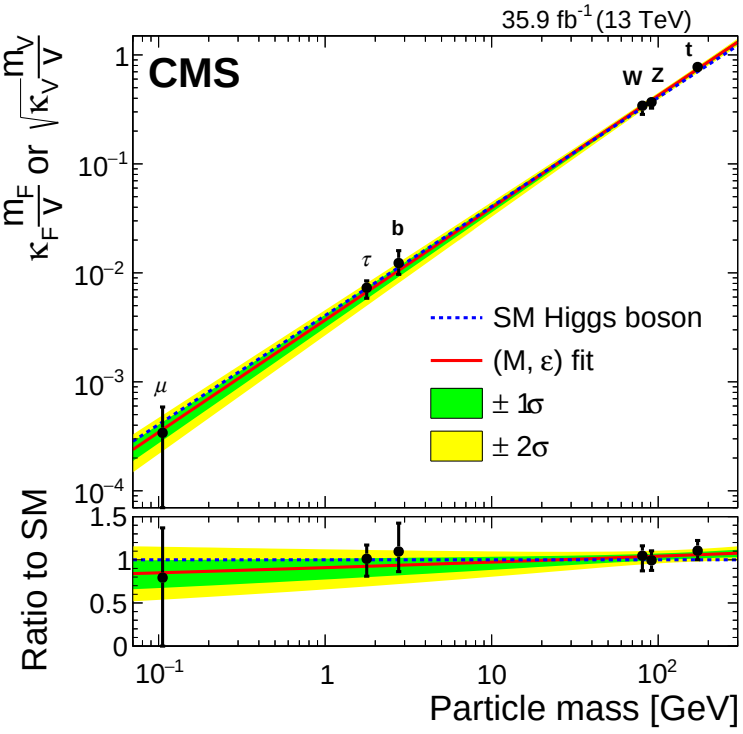


Figure 1.5 – Normalized Higgs boson coupling constants as a function of the boson or fermion masses [24]. The coupling to muons is measured through a H→ $\mu\mu$ search; however, the H→ $\mu\mu$ decay was not observed yet and the upper limit on the cross section times the branching fraction of this process is set to about 2.9 times the SM prediction [25].

The Higgs boson self-coupling, indeed, is responsible for the mass of the Higgs boson itself. By the Eq. 1.49, the value of the self-coupling is now indirectly determined to be $\lambda \sim 0.13$ from the values of v and of m_{H}; however, the direct measurement of the coupling is a unique test of the consistency of the theory. The λ_{HHH} coupling can be probed in events where two Higgs bosons are produced; the cross section of the double Higgs production amounts to about 40 fb in proton-proton collisions with center-of-mass energy of $\sqrt{s} = 13$ TeV, as those occurring the LHC in nominal conditions. The study of these events is the topic of this thesis. Processes where three Higgs bosons are produced are much rarer ($\sigma = \mathcal{O}(0.1)$ fb at 14 TeV [27]); therefore, the measurement of the quadrilinear self-interaction λ_{HHHH} is currently out of reach.

1.2 The double Higgs boson production

The production of pairs of Higgs bosons (HH) is a rare process predicted by the SM with a cross section of about 31 fb at $\sqrt{s} = 13$ TeV for the dominant production mode.

The trilinear self-coupling is not the only interaction that contributes to the HH production: as detailed in Sec. 1.2.1, specific production modes allow the Yukawa coupling y_{t} with a top quark to be probed, as well as the trilinear and quadrilinear couplings with vector bosons λ_{V} and $\lambda_{2\mathrm{V}}$.

Precise measurements of the value of the couplings are far from being performed directly; the trilinear self-coupling is expected to be determined with a precision of about 50% by the end of the LHC operations [28], including the High Luminosity phase. However, effects from physics Beyond the Standard Model (BSM) can be probed with the current LHC configuration (see Sec. 2.1): even small deviations of the values of the couplings

from the SM prediction can lead to a large modification of the HH production cross section and of the kinematics of specific production mechanisms. An overview of these effects is given in Sec. 1.2.2.

1.2.1 Overview of the HH production modes

The main production modes of Higgs boson pairs in proton-proton collisions are described in the following. They are similar to the single-H production modes for topology and cross section hierarchy, although each HH production mode is about $O(1000)$ times rarer of its single-H counterpart. Their cross section is represented in Fig. 1.6 as a function of the center-of-mass energy $\sqrt{s}$; the search performed in this thesis uses data collected during the LHC Run 2 phase, carried out with proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$.

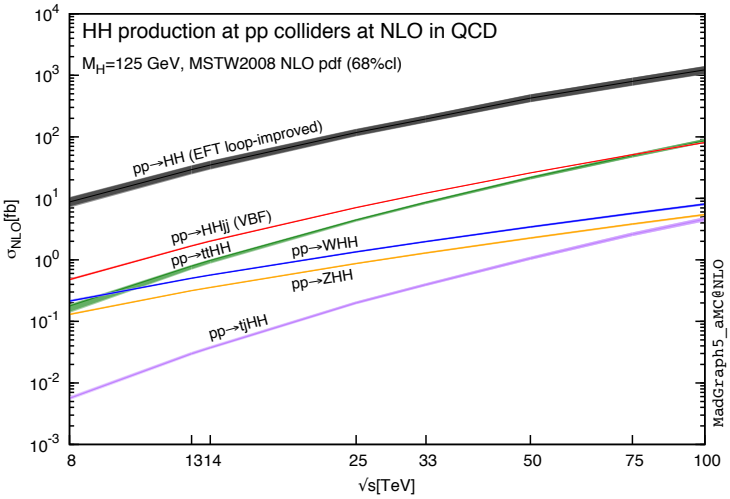


Figure 1.6 – Total cross sections at NLO precision for the six largest HH production channels at proton-proton colliders. The thickness of the lines corresponds to the scale and PDF uncertainties added linearly [29].

Gluon fusion Higgs boson pairs are dominantly produced in the heavy quarks loop-induced gluon fusion (gg→HH) mechanism; the top quark contribution to the production is the largest, followed by the b quark loop, whose contribution amounts to about 1%. The two diagrams represented in Fig. 1.7, denoted in the following as "box" and "triangle" diagrams, participate to the double Higgs production with destructive interference. While the triangle contribution depends linearly on the value of the trilinear self-coupling and of the y_t Yukawa coupling, the box contribution depends quadratically on y_t.

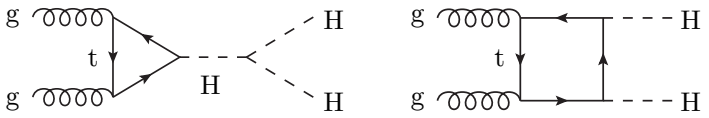


Figure 1.7 – Diagrams contributing to the Higgs pair production via gluon fusion.

Vector Boson Fusion The VBF process ($qq' \to qq'$HH) involves the production of a single Higgs boson splitting into a Higgs boson pair, and that of two Higgs bosons

radiating off virtual W or Z bosons; the most distinctive VBF feature is the presence of a pair of quarks, produced with large angular separation and large invariant mass. The main diagrams are represented in Fig. 1.8: while the first diagram involves the trilinear H self-interaction, the Higgs pair production occurs through the trilinear and quadrilinear interaction with vector bosons in the other processes.

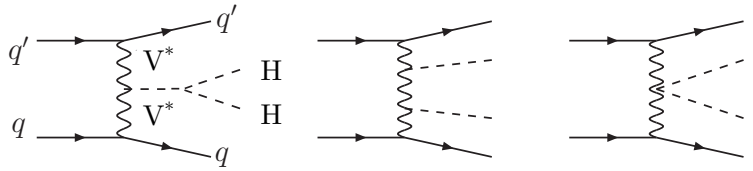


Figure 1.8 – Diagrams contributing to the Higgs pair production via VBF.

Top quark pair associated production The diagrams corresponding to the associated production of Higgs boson pairs with top quark pairs ($qq'/gg \rightarrow t\bar{t}HH$) are represented in Fig. 1.9: two Higgs bosons are either produced in a $t\bar{t}H$ process from the Higgs boson self-coupling, or are radiated from the top quarks; thus, the y_t and λ_{HHH} couplings are involved. The total cross section is smaller than that of the VBF process in the range of center-of-mass energy already explored and up to $\sqrt{s} \sim 75\,\text{TeV}$.

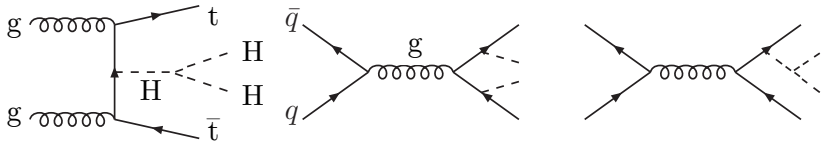


Figure 1.9 – Diagrams contributing to the Higgs pair production in association with a $t\bar{t}$ pair.

Vector boson associated production The associated production of Higgs pairs with a W or Z boson ($qq' \rightarrow VHH$), or "double Higgstrahlung", involves the couplings λ_V and λ_{2V} in addition to the λ_{HHH} coupling, as represented in Fig. 1.10. Its cross section is significantly smaller than that of the VBF mechanism, involving the same interactions.

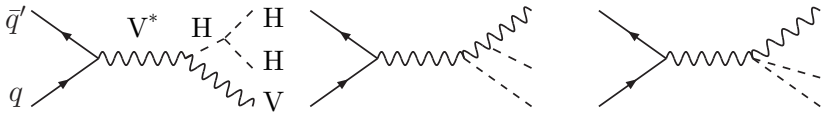


Figure 1.10 – Diagrams contributing to the Higgs pair production in association with a vector boson.

Single top quark associated production The associated production with a single top ($qq' \rightarrow tqHH$) occurs either through t- or s-channel, represented respectively in the top and bottom row of in Fig. 1.11. Among the HH production modes, this is the only one that is sensitive to the Higgs boson self-coupling, to the coupling with vector bosons and to the interaction with the top quark; however, due to its extremely small cross

section (about $3 \cdot 10^{-2}$ fb at $\sqrt{s} \sim 13$ TeV), it currently cannot be investigated at the LHC.

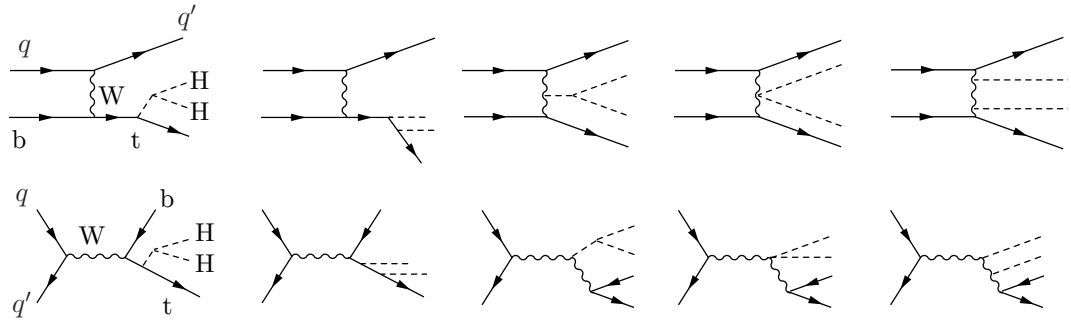


Figure 1.11 – Diagrams contributing to the Higgs pair production in association with a top quark.

Due to the rarity of the HH production processes, the past HH searches are focused on the dominant production mechanism. In this thesis, in addition to the gluon fusion search, a strategy for VBF-specific studies is outlined: in spite of its small cross section, its peculiar signature can be exploited to define regions with large signal-over-backgroud ratio.

Within each production mode, the final cross section is the result of the interference of several diagrams. For instance, the differential cross section corresponding to the triangle and box diagram participating to the gluon fusion mechanism, as well as that of their interference, is represented in Fig. 1.12 as a function of the Higgs pair invariant mass: due to the large destructive interference, the global cross section is reduced by about 50% compared to the box-only contribution, which is the largest.

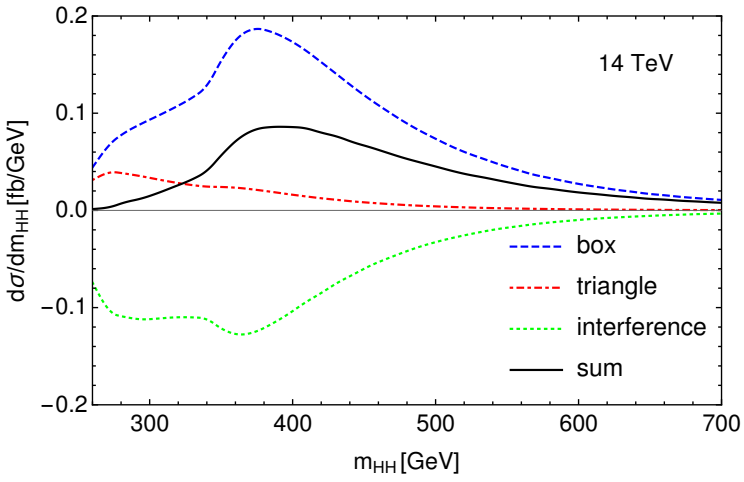


Figure 1.12 – Differential cross section corresponding to each of the gluon fusion production diagrams, represented in Fig. 1.7, and their interference, as a function of the invariant mass of the Higgs pair [30].

Similarly, the differential cross section of the VBF production results from large cancellations due to the interference among the single diagrams. In Fig. 1.13, it is represented as a function of the Higgs pair invariant mass.

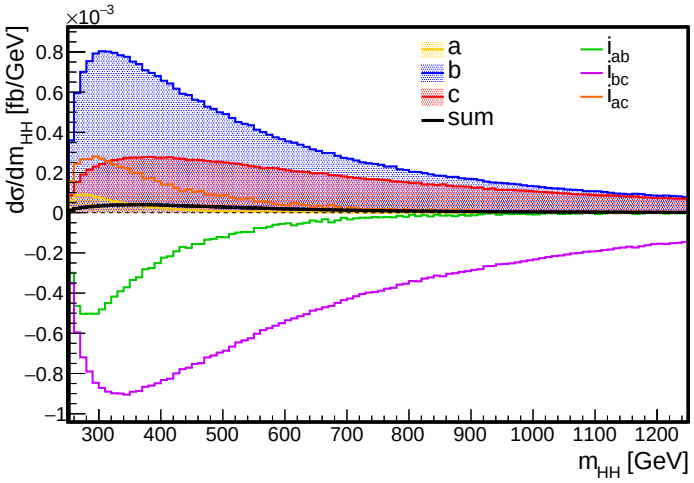


Figure 1.13 – Differential cross section corresponding to each of the VBF production diagrams, represented in Fig. 1.8, and their interference, as a function of the invariant mass of the Higgs pair.

1.2.2 BSM production

The elaboration of BSM theories is motivated by the awareness on the limitations of the SM in explaining some features specific to the electroweak sector itself and in the description of the Nature in a wider context.

In the context of the present SM formulation, although the BEH mechanism provides a satisfactory model to solve the lack of massive particles due to the electroweak symmetry, many questions stay unanswered.

Indeed, a number of experimental evidences are not accounted for by the SM. For instance, there is no explanation for the large dominance of matter over antimatter in the Universe [31]; there is also no indication for the nature of the dark matter [32]. Also, the current SM formalism only accounts for three of the fundamental interactions: the gravity is not included.

As for the electroweak sector, in the first place, the fermions follow a scheme replicated over three generations that essentially differ from each other because for their masses, i.e. because of the strength of their interaction with the Higgs boson; this degeneracy is an experimental evidence rather than consequence of the theory, although the BEH mechanism copes extremely well with it. Secondly, the mass of the Higgs boson is not protected by a fundamental symmetry; the value that we observe is the result of large divergences canceled out by regularisation mechanisms that call for extreme fine-tuning [33] if the SM is valid up to the Plank scale; out of these large cancellations that depend on the cutoff of the theory, it seems unnatural that the resulting m_{H}, which is a phenomenological parameter, conveniently happens to be in a range that can be probed at the LHC.

In practice, none of these observations invalidate the current SM as structurally complete. However, they can point to the existence of more fundamental theories at higher scales; ideally, these theories should be more general and incorporate the existing mechanisms.

Two approaches can be followed to measure the possible deviations from the SM predictions: studies can be performed either in the context of a specified model or using a model-independent effective field theory.

The former case usually involves a new particle X that decays in two Higgs bosons, so that the signature is that of a resonance with mass $m_X > 2m_H$: the resonant production is predicted by many extensions of the SM such as the Singlet model [34, 35, 36], the Two-Higgs Doublet Model (2HDM) [37], the Minimal Supersymmetric Standard Model (MSSM) [38, 39], and the Warped Extra Dimensions (WED) [40, 41] model. For example, the WED models, inspired by the string theory, are based on the hypothesis that a finite extra spatial dimension exists; for instance, while the SM belongs to the four-dimensional space, gravity propagates also in higher dimensions in this scenario, so that its interaction in the four-dimensional space appears weaker than that of the other fundamental forces. Additional particles belonging to a higher dimensional space and with sizeable branching fractions to the HH final state are predicted by WED models, such as the spin-0 radion [42] and the spin-2 first Kaluza-Klein (KK) excitation of the graviton [43]. Although the other models describe different phenomenologies, they all predict the existence of a CP-even scalar spin-0 or spin-2 particles, with an intrinsic width that is negligible with respect to the detector resolution and with similar signatures as those for the graviton and radion; however, the mass hypotheses span over a wide range: from 250 to 350 GeV for 2HDM and MSSM, from 250 GeV to 1 TeV for the Singlet Model, and from 250 GeV to 3 TeV for WED. Therefore, they can all be explored at once with similar strategies, but a large phase space needs to be covered. Compared to the non-resonant searches, similar analysis techniques are implemented in the searches for new particles decaying into a HH pair, with the additional handle that the invariant mass of the two Higgs boson candidates is a natural final observable.

The latter case is that of the non-resonant production and it is the choice made in this thesis. In general, an imbalance of the relative yield of the diagrams participating to a given production mode can result in a large modification of the cross section: even small deviations of the value of each coupling from the SM prediction can lead to effects large enough to be probed within the LHC program. From Fig. 1.7, one can also infer that a possible imbalance modifies the the differential cross section distribution that results from the sum of the contributions of different processes.

In the simple case of the gluon fusion production, represented in Fig. 1.7, the ratio of the HH production cross section over its SM prediction can be parametrised at leading order from the square of the amplitude of the box and triangle diagrams as

$$\frac{\sigma_{HH}}{\sigma_{HH}^{SM}} = 0.28\ k_\lambda^2 k_t^2 - 1.37\ k_\lambda k_t^3 + 2.09\ k_t^3 \tag{1.57}$$

where the deviations from the SM prediction of the λ_{HHH} and y_t are denoted respectively as $k_\lambda = \lambda_{HHH}/\lambda_{HHH}^{SM}$ and $k_t = y_t/y_t^{SM}$, so that their value is equal to 1 in the SM; the coefficients of the Eq. 1.57 are computed in [44], in the context of the effective field theory parametrization described later in this section.

The effect of a deviation from the SM prediction of the value of λ_{HHH} is represented for all the HH production mechanisms in Fig. 1.14. The cross section of all processes is sensitive to the value of the trilinear self-coupling, each with a different dependency. The curve of the gluon fusion, as by the Eq. 1.57, has a minimum in $k_\lambda/k_t = 2.45$, corresponding to the maximum destructive interference between the box and triangle diagram; by moving only by 1 from the SM prediction along the x axis in the negative direction, the production cross section is enhanced by about a factor 20.

As for the VBF production, as mentioned earlier, it also involves the trilinear and quadrilinear couplings with the vector bosons, denoted in the following as $c_V = \lambda_V/\lambda_V^{SM}$ and $c_{2V} = \lambda_{2V}/\lambda_{2V}^{SM}$; the value of c_V and c_{2V} is assumed to be stable beyond leading order

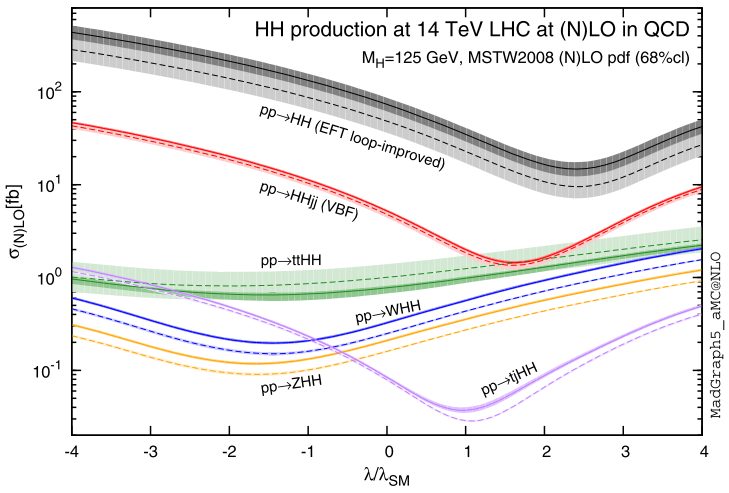


Figure 1.14 – Total cross sections at NLO precision for the six largest HH production mechanisms at proton-proton colliders as a function fo the variation from the λ_{HHH} prediction. The thickness of the lines corresponds to the scale and PDF uncertainties added linearly [29].

precision. The cross section of the HH production through VBF can be expressed (see Sec. 5.3) under the form

$$\begin{aligned}\sigma(c_{\mathrm{V}}, c_{2\mathrm{V}}, k_\lambda) \sim 0.9\ c_{\mathrm{V}}^2 k_\lambda^2 + 31.4\ c_{\mathrm{V}}^4 + 16.5\ c_{2\mathrm{V}}^2 + \\ - 8.6\ c_{\mathrm{V}}^3 k_\lambda + 5.5\ c_{\mathrm{V}} c_{2\mathrm{V}} k_\lambda - 44.0\ c_{\mathrm{V}}^2 c_{2\mathrm{V}}.\end{aligned} \tag{1.58}$$

As shown in Fig. 1.15, the sensitivity on the quartic coupling $\mathrm{c}_{2\mathrm{V}}$ is much larger than that on modifications of the trilinear-self coupling: variations of $c_{2\mathrm{V}}$ itself of about 0.5 correspond to an enhancement of the cross section by about a factor 40. Moreover, the measurement of $\mathrm{c}_{2\mathrm{V}}$ can provide information on the nature of the electroweak symmetry breaking dynamics: specific BSM models where the Higgs is a composite boson lead to $\mathrm{c}_{2\mathrm{V}} \neq \mathrm{c}_{\mathrm{V}}^2$ scenarios [45].

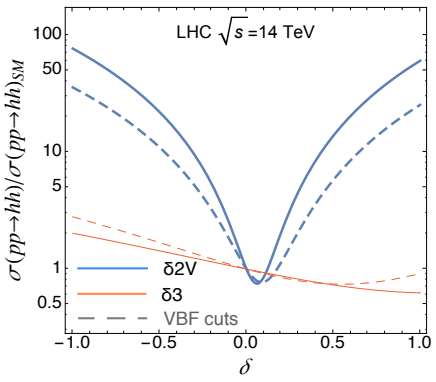


Figure 1.15 – VBF HH production cross section, in units of the SM value, as a function of $\delta_{c_{2\mathrm{V}}} = c_{2\mathrm{V}} - 1$ after cuts reproducing the realistic acceptance of LHC detectors (solid) and after specific VBF analysis selections (dashed) [45].

Towards Effective Field Theories

Since no new particles beyond the SM have been observed so far by the LHC experiments, one can make the hypothesis that new particles exists but they are significantly heavier than the kinematically accessible mass region; in this case, an effective field theory (EFT)

approach can be adopted: New Physics valid at a new scale $\Lambda \gg v$ can be described by complementing the Lagrangian of the SM with additional fields that act only at short distances and at large energy scales, whose substructure within the current reach of the LHC can be ignored. In practice, to describe physics at scales below Λ, the fields of a renormalizable extended theory, interacting only at high energy can be integrated out; they give rise to residual higher dimensions non-renormalizable operators built from those fields.

For the Lagrangian to be renormalizable, all the terms must be operators of dimension 4 or less; thus, in the EFT context, operators with with dimension $d > 4$ are included; they must have a coefficients proportional to Λ^{4-d}, i.e. they are highly suppressed at the current energy scale.

Thus, the effective Lagrangan is written under the form

$$\mathcal{L}_{eff} = \mathcal{L}_{\mathrm{SM}} + \mathcal{L}_{D=5} + \mathcal{L}_{D=6} + ... \tag{1.59}$$

where each part consists of operators invariant under $SU(3)_C \times SU(2)_L \times U(1)_Y$ local transformations of dimension D; the SM term only contains operators up to $D = 4$. The $\mathcal{L}_{D=5}$ corresponds to operators that account for the mass of neutrinos and it violates lepton number conservation; since it has no impact on the Higgs phenomenology, it is removed from the study. The dominant term, thus, is the one with $D = 6$ operators. While in the VBF case the EFT operator is only responsible for the modification of the SM couplings, additional couplings arise in the case of the gluon fusion.

Gluon fusion EFT couplings

The gluon fusion production BSM production can be described by five parameters controlling the Higgs boson interactions. These parameters are represented in Fig. 1.16: the y_{t} and λ_{HHH} are the regular SM couplings; additional parameters representing contact interactions, i.e. interactions with a substructure that is unknown, are the c_2 effective coupling between a Higgs boson pair and a heavy quarks pair, the c_{2g} effective coupling between a Higgs pair and a pair of gluons, and the c_g effective coupling between one Higgs boson and two gluons. Because the triangle diagram is dominated by the t quark (99%) and the Yukawa coupling y_{b} with a b quark already has strong constraints in EFT [46], anomalous values of the y_{b} coupling are omitted.

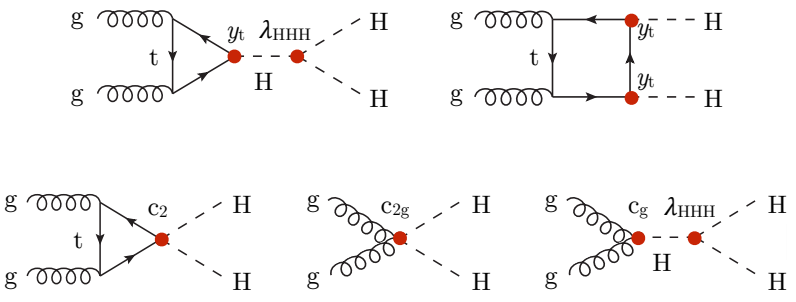


Figure 1.16 – Effective field theory couplings contribution to the HH production through gluon fusion.

From the squared amplitudes of the diagrams in Fig. 1.16, the ratio of the gluon fusion

HH cross section over its SM prediction can be parametrised as [44]

$$\begin{aligned} R_{\mathrm{HH}} = \frac{\sigma_{\mathrm{HH}}}{\sigma_{\mathrm{HH}}^{\mathrm{SM}}} = {} & A_1 k_{\mathrm{t}}^4 + A_2 c_2^2 + (A_3 k_{\mathrm{t}}^2 + A_4 c_g^2) k_\lambda + A_5 c_{2g}^2 + \\ & + (A_6 c_2 + A_7 k_\lambda k_{\mathrm{t}}) k_{\mathrm{t}}^2 + (A_8 k_{\mathrm{t}} k_\lambda + A_9 c_g k_\lambda) c_2 + \\ & + (A_{10} c_2 c_{2g} + (A_{11} c_g k_\lambda + A_{12} c_{2g}) k_{\mathrm{t}}^2 + \\ & + (A_{13} k_\lambda c_g + A_{14} c_{2g}) k_{\mathrm{t}} k_\lambda + A_{15} c_g c_{2g} k_\lambda, \end{aligned} \tag{1.60}$$

which is equivalent to the Eq. 1.57 in the scenario where $c_g = c_2 = c_{2g} = 0$. The coefficients A_i are determined from a simultaneous fit to the cross section obtained from MadGraph5_aMC@NLO [47] simulations at leading order (LO) precision. As mentioned earlier, the value of the couplings does not only affect the total cross section of the gluon fusion HH process, but also the kinematics of the HH pair, i.e. it modifies the differential cross section as a function of the relevant observables related to the event topology.

For easier modelling of these effects in physics analyses, a finite set of benchmarks, each representative of a cluster of several combinations of couplings producing similar kinematic features, can be identified; their use in the search performed in this thesis is discussed in Sec. 5.1. The couplings corresponding to the benchmarks identified in [44] are summarised in Tab. 1.2; as shown in Fig. 1.17, the distribution of the invariant mass of the HH system in signal samples within the same cluster is reasonably close, while it is dramatically different in separate clusters.

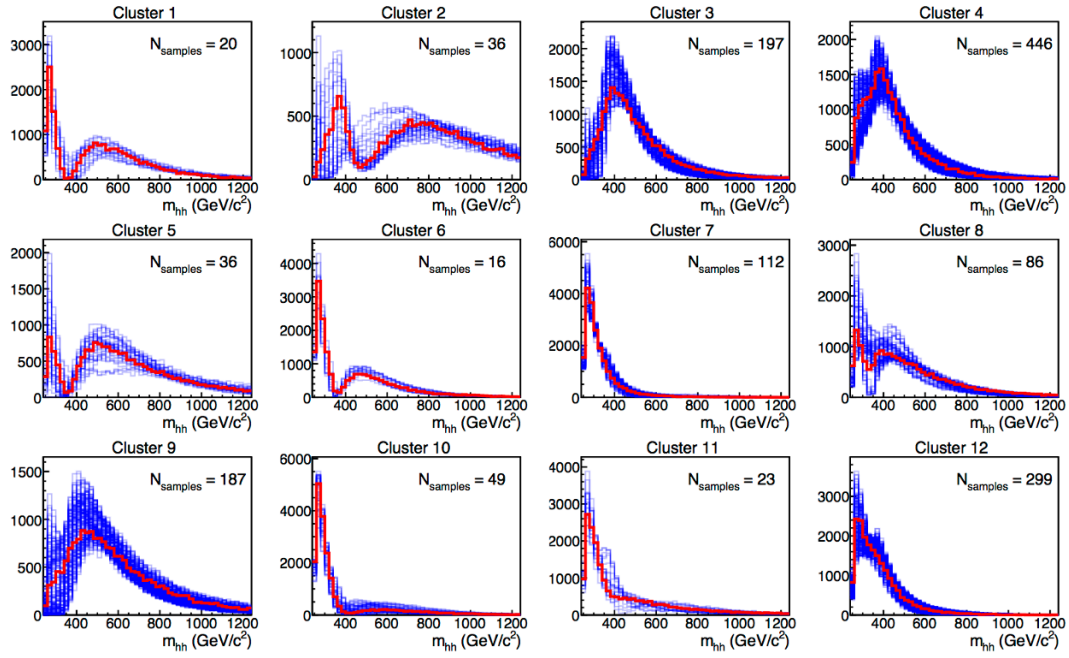


Figure 1.17 – Invariant mass of the Higgs boson pair for each of the identified clusters; the red curves correspond to the benchmark describing at best the kinematic of the signals in the cluster, corresponding to the couplings listed in Tab. 1.2; the blue curves represent the other samples in the cluster. The SM simulation is represented in Cluster 3 [44].

1.3 Double Higgs searches at the LHC

A rich set of HH final states is accessible at the LHC; their branching ratio is summarised in Fig. 1.18. As the HH production is a very rare process, the signal event collection

Table 1.2 – Values of the EFT couplings corresponding to the benchmarks identified in clusters of processes with similar kinematic features [44].

Cluster	k_λ	k_t	c_2	c_g	c_{2g}
1	7.5	1.0	-1.0	0.0	0.0
2	1.0	1.0	0.5	-0.8	0.6
3	1.0	1.0	-1.5	0.0	-0.8
4	-3.5	1.5	-3.0	0.0	0.0
5	1.0	1.0	0.0	0.8	-1.0
7	5.0	1.0	0.0	0.2	-0.2
8	15.0	1.0	0.0	-1.0	1.0
9	1.0	1.0	1.0	-0.6	0.6
10	10.0	1.5	-1.0	0.0	0.0
11	2.4	1.0	0.0	1.0	-1.0
12	15.0	1.0	1.0	0.0	0.0
SM	1.0	1.0	0.0	0.0	0.0

needs to be preserved by targeting final states with a sizeable branching ratio, such as those where at least one of the Higgs bosons decays in a pair of b quarks. The search in each final state presents different experimental challenges. Some examples are given in the following.

HH→bbbb is the final state with largest branching fraction (BR = 33.7%). However, the contamination from processes faking the signal signature is extremely high; in particular, it suffers from the contamination of generic processes with large hadronic activity (multijet QCD).

HH→bb$\tau\tau$ can profit from a sizeable branching fraction (BR = 7.3%) and from the high purity guaranteed by the $\tau\tau$ pair. Therefore, it is a good compromise, although it suffers from large contamination from multijet QCD events and from events where a top-antitop quark pair ($\mathrm{t\bar{t}}$), each having a large probability of decaying in a W and a quark b, is produced.

HH→bbVV has a sizeable branching fraction; CMS searches have been performed in the VV(ZZ or WZ)→ $2\ell 2\nu$ case, which globally has BR = 2.7%. This final state suffer from a large $\mathrm{t\bar{t}}$ background.

HH→bb$\gamma\gamma$ is a very clean final state, thanks to the excellent resolution on photons achieved by the CMS experiment (see Ch. 2). Its branching ratio, though, is very small (BR = 0.3%).

The choice of the channels that are explored in current searches, thus, is the result of the trade-off between the purity that can be achieved by the event selection and the branching ratio of the HH decay for the given channel.

The final state investigated in this thesis is the one where the two Higgs bosons decay in a pair of b quarks and a pair of tau leptons. The strategies implemented for the analysis of $41.6\,\mathrm{fb}^{-1}$ of data collected in 2017 with the CMS detector are presented in Ch. 4.

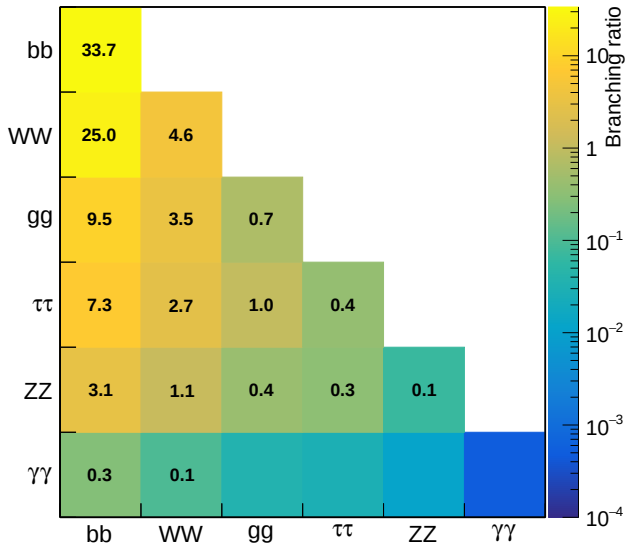


Figure 1.18 – Branching fraction of the HH final states; the value of the branching ratio is omitted for the rarest final states. The SM Higgs branching fractions are used, summarised in Tab. 1.1.

1.3.1 Summary of the past HH searches

Several HH searches have been performed by both the CMS and ATLAS collaborations at the LHC. The full set of proton-proton collisions at $\sqrt{s} = 8\,\text{TeV}$ (Run 1) has already been exploited in many HH final states. Run 2 data corresponding to collisions at $\sqrt{s} = 13\,\text{TeV}$ are still being analysed; most of the existing searches cover the 2016 data-taking only. The statistical method used for interpretation of the results is described in Sec. 6.3; in the following, the results are quoted in the form of a signal strength with respect to the prediction $\sigma_{\text{SM}}(\text{gg} \to \text{HH}) = 33.5\,\text{fb}$.

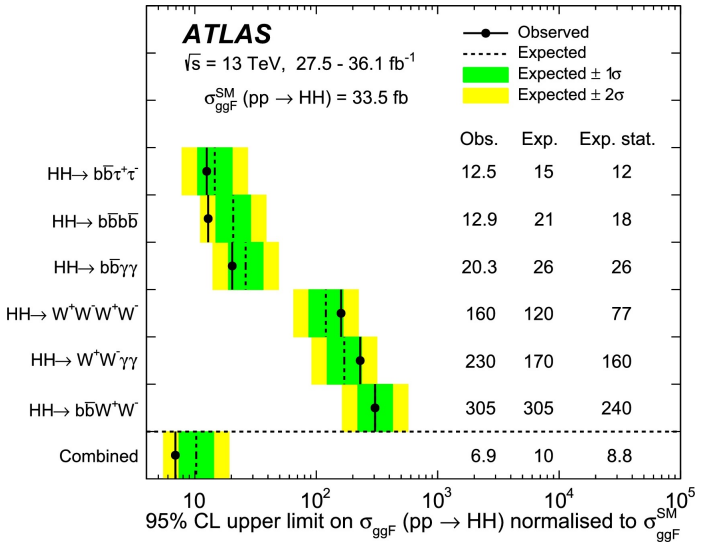


Figure 1.19 – Observed and expected upper limit at 95% CL on $\sigma(\text{gg} \to \text{HH}) \cdot \text{BR}$ obtained in each ATLAS analysis performed with 2016 data and with their combination [48]. The $\pm 1\sigma$ and $\pm 2\sigma$ deviations from the expected value are represented by green and yellow bands.

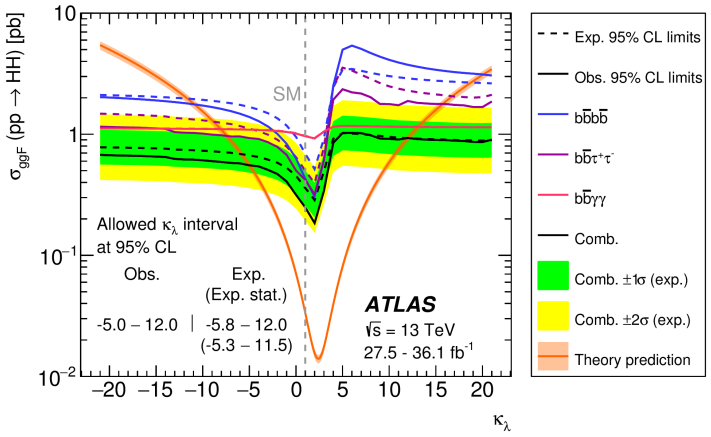


Figure 1.20 – Observed and expected upper limit at 95% CL on $\sigma(\mathrm{gg} \to \mathrm{HH})$ as a function of k_λ obtained in each ATLAS analysis performed with 2016 data and with their combination [48]. The $\pm 1\sigma$ and $\pm 2\sigma$ deviations from the expected value are shown only for the combined expected limit; they are represented by green and yellow bands.

The ATLAS HH results from the analysis of 2016 data (36.1 fb^{-1} at $\sqrt{s} = 13$ TeV) are summarised and combined in [48]: the bbbb, bb$\tau\tau$, bbWW, WWWW, bb$\gamma\gamma$ and WW$\gamma\gamma$ channels are explored. The results are summarised in Fig. 1.19: the most sensitive result is that obtained in the bb$\tau\tau$ final state, giving an observed (expected) upper limit at 95% CL on the $\sigma(\mathrm{gg} \to \mathrm{HH}) \cdot \mathrm{BR}$ cross section about 12.5 (15) larger than the SM prediction; ordered by sensitivity on the SM prediction, the bbbb search and the bb$\gamma\gamma$ follow with similar results, while the other analyses are less competitive. The ATLAS analyses combined set an upper limit of 6.9(10) times the SM prediction. As for the value of k_λ, it is constrained between -5 and 12 by observed data and between -5.8 and 12 through the expected upper limit; the limit scan over k_λ is represented in Fig. 1.20. A dedicated VBF search, including constraints on c_{2V}, is performed in the bbbb final state [49] using about 126 fb^{-1} of data collected in Run 2 operations ($\sqrt{s} = 13$ TeV); the observed (expected) upper limits exclude $-1 < c_{2V} < 2.8$.

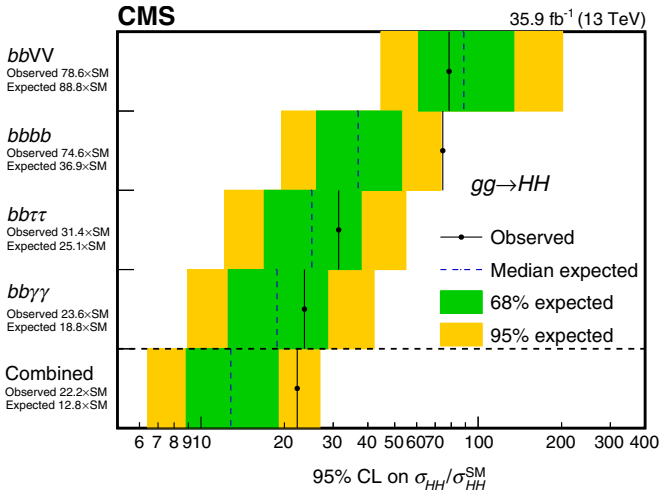


Figure 1.21 – Observed and expected upper limit at 95% CL on $\sigma(\mathrm{gg} \to \mathrm{HH}) \cdot \mathrm{BR}$ obtained in each CMS analysis performed with 2016 data and with their combination [50]. The $\pm 1\sigma$ and $\pm 2\sigma$ deviations from the expected value are represented by green and yellow bands.

The CMS combination of the HH searches with 2016 data is documented in [50]. In

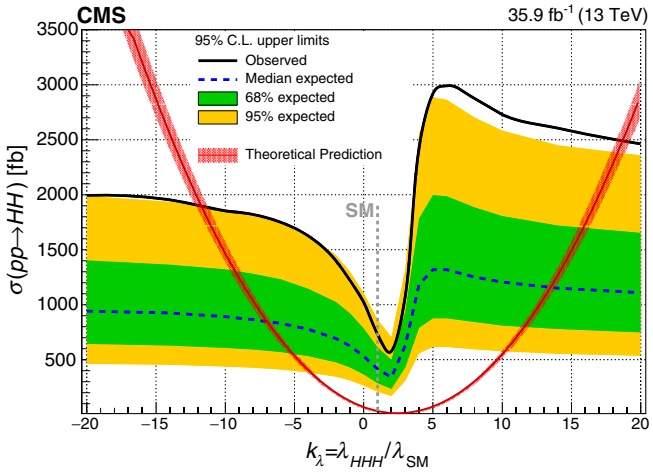


Figure 1.22 – Observed and expected upper limit at 95% CL on $\sigma(gg \to HH)$ as a function of k_λ obtained with the combination of the CMS analyses performed with 2016 data [50]. The $\pm 1\sigma$ and $\pm 2\sigma$ deviations from the expected value are represented by green and yellow bands.

Fig. 1.21, the results for each analysis in the SM scenario are summarised. The hierarchy of the sensitivity of the analyses is different from that of the ATLAS searches: the most sensitive search is the one in the $bb\gamma\gamma$ final state, followed by $bb\tau\tau$ and bbbb. The observed (expected) combined upper limit is about 22(13) times larger than the SM prediction; the observed (expected) constraints on the trilinear self-coupling are $-11.8(-7.1) < k_\lambda < 18.8(13.6)$, as shown in Fig. 1.22.

Each search has a different coverage of the phase-space, so that a different sensitivity can be achieved by each analysis across the BSM scenarios investigated. The complementarity of the searches in the various final states is illustrated in Fig. 1.23; the upper limits are shown for each of the benchmarks in Tab. 1.2. For instance, one can compare the benchmark no. 2 and benchmark no. 7 with the help of Fig. 1.17: the former has a much harder m_{HH} spectrum and the $bb\tau\tau$ search performs better in this scenario; in the latter case, the $bb\gamma\gamma$ analysis is significantly more sensitive.

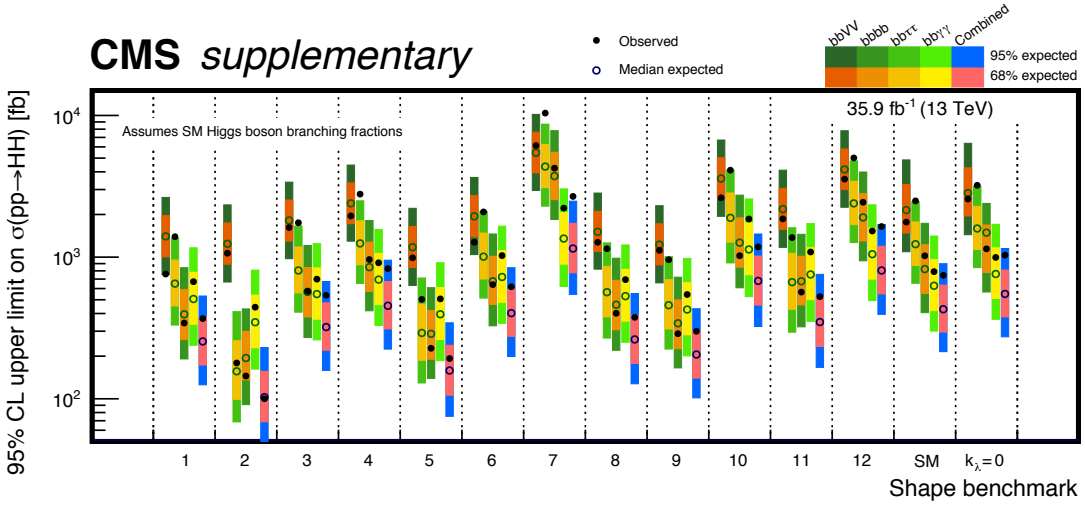


Figure 1.23 – Observed and expected upper limit at 95% CL on $\sigma(gg \to HH) \cdot BR$ obtained in each CMS analysis performed with 2016 data and with their combination [50], shown separately for each of the benchmarks listed in Tab. 1.2.

Searches for resonant HH production were also performed in CMS in the spin-0 and

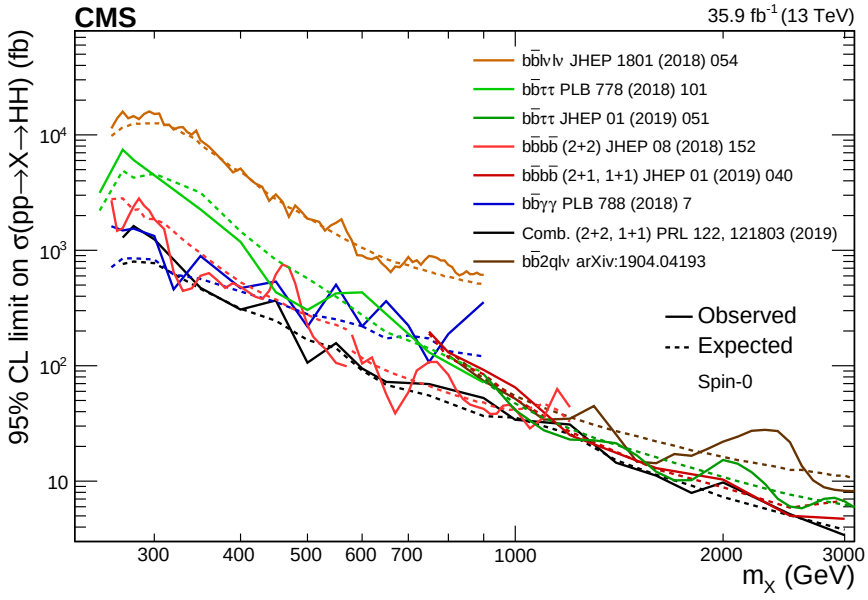


Figure 1.24 – Observed and expected upper limit at 95% CL on the cross section $\sigma(\mathrm{pp} \to \mathrm{X} \to \mathrm{HH})$ of a spin-0 particle X decaying in a Higgs boson pair, obtained in each CMS analysis performed with 2016 data and with their combination [51].

spin-2 scenarios. The limits obtained in all the final states are summarised in Fig. 1.24 as a function of the mass of a new CP-even particle X of spin-0 (radion). The comparison among the analyses reflects the sensitivity change in different regions: the curves cross each other several times over the range; the bb$\gamma\gamma$, for instance, has a good performance at low m_{X}, consistently with that observed in Fig. 1.23.

Chapter 2

The Large Hadron Collider and the CMS detector

The Large Hadron Collider (LHC) is the most powerful particle accelerator ever built, operated by CERN (*Conseil Européen pour la Recherche Nucléaire*, European Organization for Nuclear Research) and representing the cutting edge technologies for the High Energy Physics. The CERN accelerators complex, of which the LHC is the latest and largest unit, is installed at the border between Swiss and France, in the vicinity of Geneva.

The main physics goals of the LHC are to guarantee interactions with a center-of-mass energy large enough to produce new heavy particles and to deliver a number of collisions that allows very rare processes to be observed. These requirements drive the machine design and the beam structure.

The LHC collides the particle beams in four interaction points, where four detectors with various purposes are installed. Among them, the Compact Muon Solenoid (CMS) is the detector used to collect the data analysed in the context of this thesis.

In this chapter, the experimental apparatus is described. The LHC operations are discussed in Sec. 2.1. The CMS detector and its structure are described in Sec. 2.2. The algorithms for the detection of particles and the reconstruction of their experimental signature are described in Sec. 2.3. Finally, the CMS trigger system and the relative experimental challenges are described in Sec. 2.4.

2.1 The Large Hadron Collider

The LHC [52] was primarily designed for proton-proton (pp) collisions and it was conceived to investigate the nature of the spontaneous symmetry breaking by pursuing the search for the Higgs boson and to scan the full accessible space for new phenomena; however, a physics program of heavy ions collisions is also carried out, focusing on the study of the collective behaviour of quarks and gluons in the form of plasma. The LHC is installed in an underground tunnel of circumference 26.7 km situated at a depth between about 45 and 170 m; the tunnel was originally built for the Large Electron-Positron Collider (LEP), which was in service until 2000. The LHC physics operations started in 2009 and will last until 2023 (possibly 2024) in the current configuration; then, the machine will undergo a major upgrade towards the High Luminosity LHC phase (HL-LHC). The LHC timeline is shown in Fig. 2.1 and will be detailed in the following.

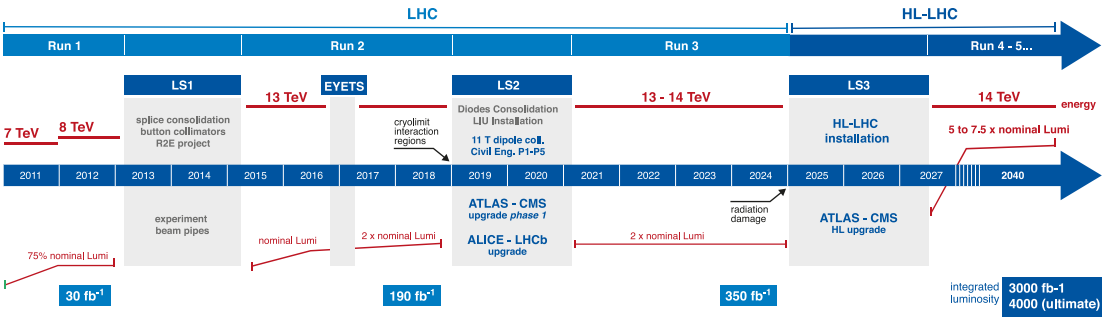


Figure 2.1 – The LHC and the HL-LHC baseline plan for the next decade and beyond. The Large Hadron Collider (LHC) operation is a sequence of interleaved operating runs of 3-4 years each and long shutdowns (LS) of 2 years. "EYETS" indicates an extended year-end technical stop. After LS3 (2025-2027), the machine will be in the high-luminosity LHC (HL-LHC) configuration [53].

2.1.1 Design

Before accessing the LHC, the proton beams are accelerated in a complex system of pre-existing machines; at each step of the acceleration chain, the beams are boosted to higher energy exploiting the previous machine as injector. The CERN accelerator complex is represented in Fig. 2.2.

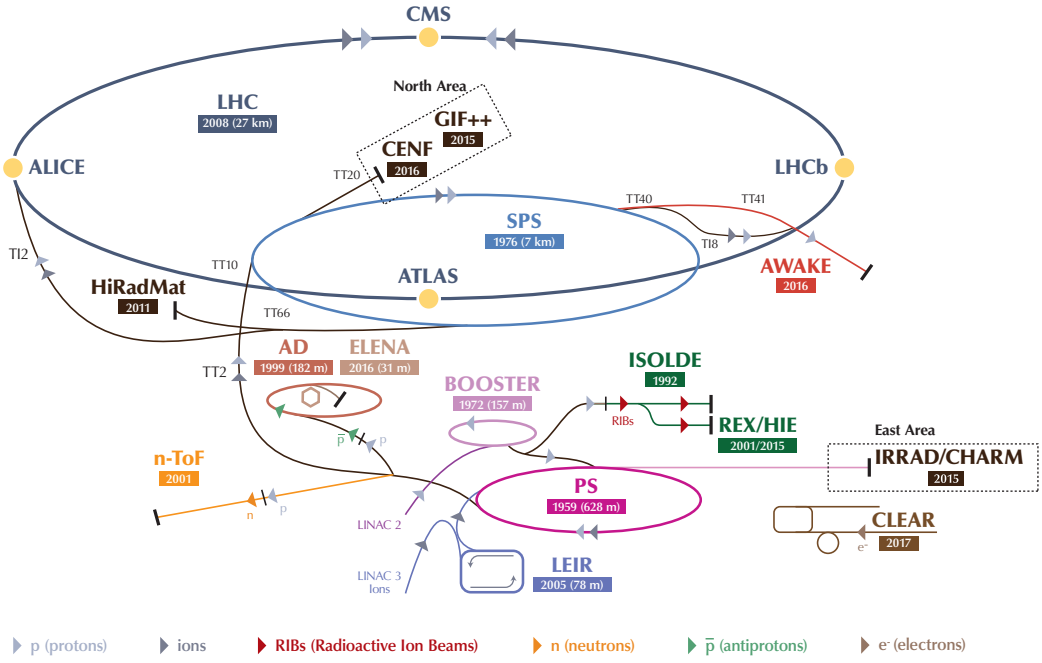


Figure 2.2 – Scheme of the CERN accelerator complex [54]

At the beginning of the chain, electrons are torn off the atoms of a hydrogen gas using a strong magnetic field, so that only protons are left. The first acceleration stage is performed at the Linear Accelerator (LINAC 2), which brings the proton beam to an energy of about 50 MeV. The first circular collider is the Proton Synchrotron Booster (PSB), which accelerates the beam up to 1.4 GeV and passes it to the Proton Synchrotron (PS) for three to four cycles that bring the beam to about 25 GeV. The last pre-acceleration step before the LHC takes place in the Super Proton Synchrotron (SPS), where the beam is accelerated to 450 GeV in about twelve cycles.

Finally, the beam is split in two adjacent parallel beamlines within the LHC tunnel,

so that they are kept travelling in opposite directions before colliding in the designated interaction points once they have reached a target energy; although the design center-of-mass energy for proton-proton collisions is $\sqrt{s} = 14\,\text{TeV}$, the target energy of the LHC acceleration stage changed over the research phases, as detailed in Sec. 2.1.4: 3.5 TeV until 2011, 4 TeV in 2012, 6.5 TeV from 2015 on.

The LHC ring is made of eight 2.45 km long arcs, where bending dipoles are placed, and and eight 545 m long straight sections. Because the tunnel is inherited from LEP, the LHC design faces some space limitations: in the arcs, for instance, the tunnel has small internal diameter of 3.7 m. Therefore, instead of implementing separate rings of magnets for each of the beams, the two counter-rotating beamlines need to be hosted within a single cryogenic and mechanic assembly; the antiparallel magnetic fields are generated by two independent sets of coils, although they are are placed close enough to each other that they are coupled both magnetically and mechanically.

The LHC electromagnets are made of copper-clad niobium-titanium superconductor, amounting globally to about 470 tonnes. The beams are kept on their circular path through the magnetic field of 1232 dipoles operating at 8.33 T, while 392 quadrupole magnets keep the particles focused in narrow beams. Dedicated quadrupoles squeeze the beams in proximity of the interaction points and magnets of higher multipole orders correct smaller imperfections in the magnetic field. The operating temperature of the magnets is 1.9 K (-271.25°C); approximately 96 tonnes of superfluid helium-4 are needed to keep them cool.

2.1.2 Parameters

The nominal beam parameters and their meaning are summarised in Tab. 2.1.

Within the beam, protons are packed in bunches distanced in time by Δt and distributed in a structure prepared along the injection chain. Larger accelerators can accommodate more bunches; therefore, at each stage of the chain, they are accumulated in longer trains as closely packed as possible. However, to move out of the injector and get to the following accelerator, the bunches need to be "kicked" by dedicated magnets, which have rise and fall times larger than the Δt separation, leading to complicated bunch schemes. A "fill" is complete when the LHC cannot accommodate any more bunches.

For a given process, the total number N_{evt} of instances where that process is reproduced ("events") is proportional to its cross section σ as

$$N_{evt} = \sigma \int L \, \mathrm{d}t \tag{2.1}$$

where the instantaneous luminosity L is a measure of the number of collisions occurring per second and the integral goes over the time of activity of the experiment. The luminosity is a key parameter of a machine such as the LHC: if large, it allows processes with low probability to be produced, so that physics searches are possible. However, handling a large instantaneous luminosity is experimentally challenging for the data-taking, as discussed in Sec. 2.4.

Under the assumption that the two beams are identical, that they have round transverse section and that they are highly collimated, L can be written as

$$L = \frac{N_p^2 n_b f_{LHC} \gamma_r}{4\pi \epsilon_n \beta^*} F, \tag{2.2}$$

where

$$F = \left[1 + \left(\frac{\theta_c \sigma_z}{2\sigma^*}\right)^2\right]^{-1/2} \tag{2.3}$$

is a geometric factor accounting for the luminosity reduction due to the crossing angle at the point of interaction between the beams.

The duration of a fill can widely change from a few minutes to about twelve hours; however, in nominal conditions, it is limited by the luminosity lifetime. Indeed, the instantaneous luminosity decreases along the fill mainly due to beam losses from collisions. In 2017 and 2018, levelling strategies were applied to maintain the initial luminosity by tuning dynamically β^* or the factor F.

The average number of simultaneous interactions per bunch crossing, i.e. the pileup, is given by

$$< \text{PU} >= \frac{L \cdot \sigma_{pp}^{inel}}{n_b \cdot f_{\text{LHC}}}, \tag{2.4}$$

where σ_{pp}^{inel} is the inelastic pp cross section; at $\sqrt{s} = 13\,\text{TeV}$, it amounts to about 69 mb [55]. A large n_b is needed to keep the pileup under control; in this sense, the LHC machine design is excellent: the nominal number of bunches is as high as 2808. Inversely, the higher the instantaneous luminosity, the larger the pileup. The nominal LHC instantaneous luminosity at the beginning of the fill is $L = 10^{34}\,\text{cm}^{-2}\,\text{s}^{-1}$; given the design number of bunches, the average pileup in nominal conditions is about 22. However, the nominal values were exceeded during ordinary 2017 operations, as summarised in Sec. 2.1.4.

Table 2.1 – Nominal LHC parameters in proton-proton collisions.

Parameter	Meaning	Nominal value
$\sqrt{s}$	Center-of-mass energy	14 TeV
Δt	Bunch separation	25 ns
n_b	Number of bunches	2808
N_p	Number of protons per bunches	$1.15 \cdot 10^{11}$
f_{LHC}	Revolution frequency	11245 Hz
σ^*	Transverse bunch r.m.s. at the interaction point	16.7 µm
σ_{xy}	Longitudinal bunch r.m.s.	7.55 cm
β^*	Beta function at the interaction point	0.55 m
θ_c	Crossing angle at the interaction point	285 μrad
ϵ_n	Transverse emittance	3.75 µm

2.1.3 Experiments

The beam crossing at the LHC occurs in multiple regions, so that data can be collected simultaneously in several points. The four main experiments, with detectors installed at the collision points, are indicated in Fig. 2.2.

"A Toroidal LHC ApparatuS" (ATLAS) [56] and the "Compact Muon Solenoid" (CMS) [57] are two general-purpose experiments with similar physics programs, covering both proton-proton and heavy ions collisions. Their detectors were optimally designed for the Higgs boson search. Because their searches are focused on rare processes, a large instantaneous luminosity is required at their collision points.

"A Large Ion Collider Experiment" (ALICE) [58] is a low luminosity experiment, collecting heavy ion collisions data to investigate the strong interactions sector of the SM and the quark-gluon plasma physics.

"LHC beauty" (LHCb) [59] is also a low luminosity experiment and it is almost exclusively devoted to the heavy flavour quarks physics, with the primary goal of searching for evidence of new physics in CP violation and rare decays of hadrons of b and c quarks.

2.1.4 Operations

Two eras of LHC physics operations are already concluded: the Run 1 and Run 2 phases, lasted respectively from 2009 to 2013 and from 2015 to 2018. The center-of-mass energy (see Fig. 2.1), as well as the instantaneous luminosity, has been increased over the years towards the nominal performance level. The evolution of the instantaneous and integrated luminosity recorded by CMS is represented in Fig. 2.3; the average pileup distribution per year is represented in Fig. 2.4.

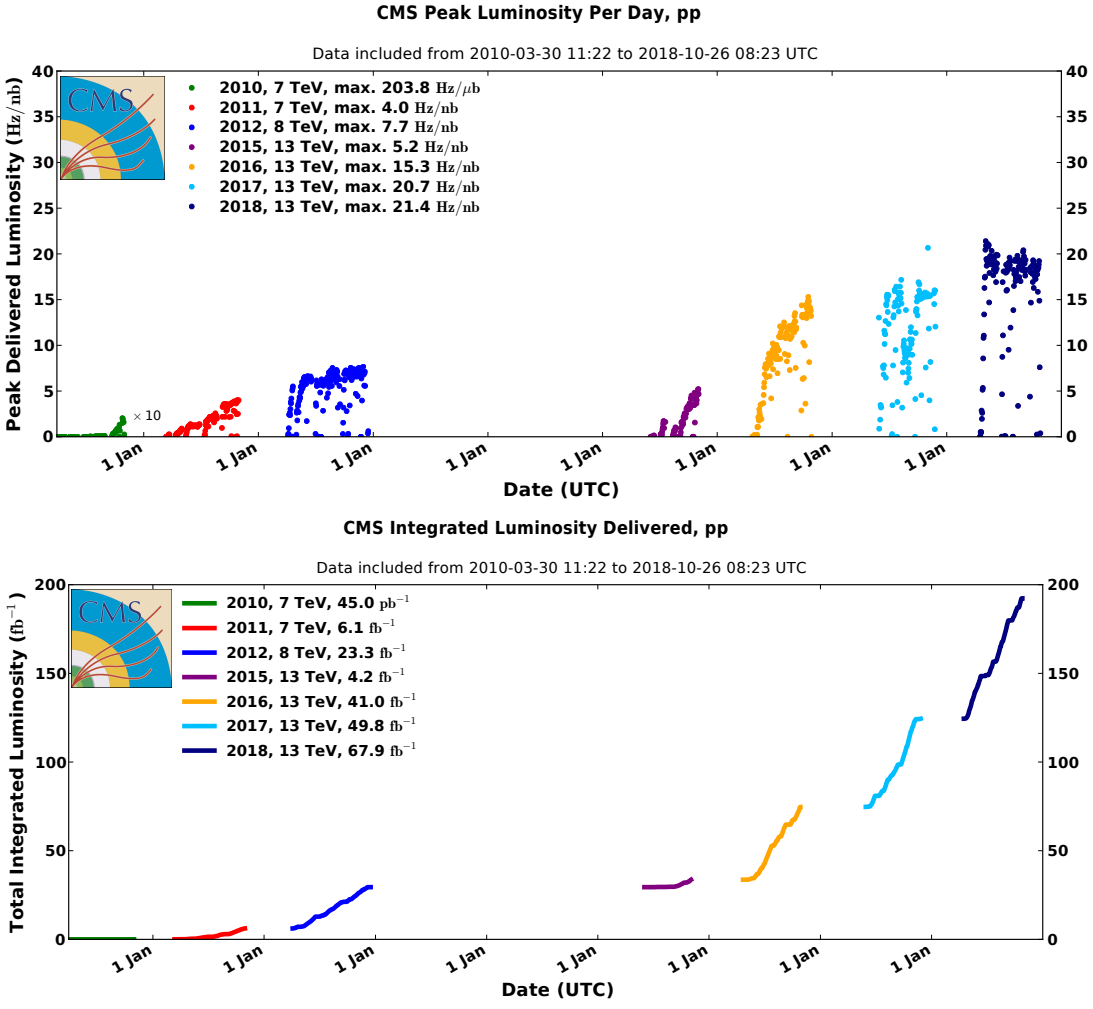


Figure 2.3 – Peak instantaneous luminosity (top) and total integrated luminosity (bottom) versus day delivered to CMS during stable beams and for pp collisions in Run 1 and Run 2, using the best offline calibrations for each year [60].

At the beginning of the Run 1, the proton beams used to be accelerated to 3.5 TeV each to achieve $\sqrt{s} = 7\,\text{TeV}$ collisions; before the end of 2011, about $6\,\text{fb}^{-1}$ of collisions were delivered to the high luminosity experiments. In 2012, the collisions took place at

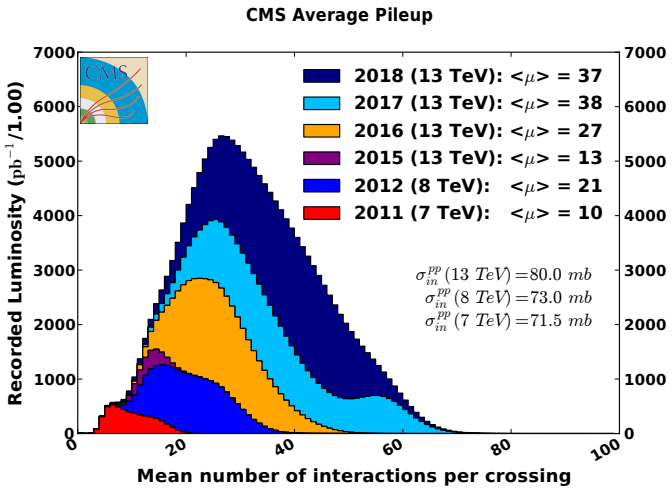


Figure 2.4 – Average pileup profile, represented in stacked histograms split by year of data-taking. The average pileup is denoted as $< \mu >$; it is computed using conventional values of σ_{pp}^{inel}, indicated on the plot [60].

$\sqrt{s} = 8\,\text{TeV}$ and a larger set of about $23\,\text{fb}^{-1}$ was delivered. These operations lead to the observation of the Higgs bosons, announced in July 2012.

The Run 1 was followed by the first Long Shutdown (LS1), which lasted about 2 years. During this time, LHC consolidation work was performed so that the beams energy could be raised to 6.5 TeV. It was also an occasion for the experiments to make the necessary detector upgrades to cope with the harsher data-taking conditions expected for the Run 2 LHC operations; for instance, part of the CMS trigger electronics were replaced.

All along the Run 2, started in 2015 and concluded in 2018, the proton-proton collisions took place at $\sqrt{s} = 13\,\text{TeV}$. The 2015 operations correspond to a phase of commissioning of the new energy configuration and the instantaneous luminosity was reduced compared to the typical values reached at the end of Run 1 ($L \sim 8 \cdot 10^{33}\,\text{cm}^{-2}\,\text{s}^{-1}$). Starting from 2016, the instantaneous luminosity was brought beyond the original LHC design value, reaching up to $L = 2.14 \cdot 10^{34}\,\text{cm}^{-2}\,\text{s}^{-1}$ in 2018. In 2016, 2017 and 2018, data sets of about 41, 50 and $68\,\text{fb}^{-1}$ were delivered by the LHC.

The LHC is currently in a phase of maintenance and upgrade (LS2) in preparation for the Run 3, when the collisions may take place at the nominal center-of-mass energy $\sqrt{s} = 14\,\text{TeV}$; by the end of 2023 or 2024, the integrated luminosity delivered by the LHC should amount to $300\,\text{fb}^{-1}$.

The Run 3 will conclude the LHC Phase 1; the potential for new discovery without a significant luminosity increase would become almost negligible. Therefore, the LHC and the accelerator complex will undergo a profound upgrade towards the High Luminosity LHC (HL-LHC) during the LS3, entering the Phase 2. The goal of the upgraded machine is to achieve a peak luminosity of $L = 5 \cdot 10^{34}\,\text{cm}^{-2}\,\text{s}^{-1}$, so that about $250\,\text{fb}^{-1}$ per year can be delivered. After about a decade of operations, the total integrated luminosity should exceed $3000\,\text{fb}^{-1}$.

2.2 The Compact Muon Solenoid experiment

The CMS experiment [57] uses a multi-purpose detector, originally designed aiming at the best resolution for the reconstruction of the Higgs boson decay products. Like most

of the detectors built to detect the final states out of particle collisions, it is made of several subdetectors placed concentrically around the interaction point, in a cylindrical configuration, and covering most of the solid angle around it. Each of the subdetectors is dedicated to the detection of a different kind of particle or to the measurement of an observable.

As its name suggests, the CMS detector is quite compact, given its size and the amount of material that it is made of: being 15 m tall and 21 m long, it is smaller than the ATLAS detector [56] by about a factor two; however, it is almost twice as heavy, with a weight of about 14000 tonnes.

2.2.1 Coordinate system

The CMS experiment uses a right-handed coordinate system, centered in the nominal interaction point. The x axis points to the geometrical center of the LHC; the y axis points upwards perpendicularly to the LHC plane, which is tilted of about 1.41% with respect to the horizontal; finally, the z axis points in the anticlockwise beam direction. The longitudinal and transverse coordinates are respectively the ones along the z axis and on the xy plane.

Polar coordinates are also used, given the cylindrical shape of the detector. The azimuthal angle ϕ is defined in the xy plane as the angle formed with respect to the x axis; the radial coordinate in this plane is denoted with r. The polar angle θ is defined in the rz plane as the angle formed with z.

The conventional CMS Cartesian and polar coordinates are illustrated in Fig. 2.5.

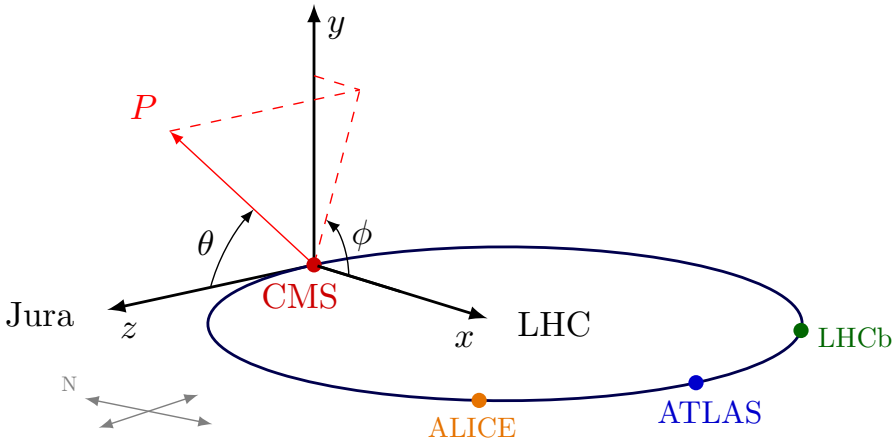


Figure 2.5 – Illustration of the CMS coordinate system [61].

In hadron colliders, the proton-proton interaction takes place at the level of their fundamental constituents; the fraction of momentum carried along z by each quark or gluon is unknown and so is the longitudinal boost of the rest frame of the collision, which does not usually match the detector rest frame. Therefore, observables that are not distorted by the center-of-mass boost are preferred.

Rather than θ, the polar angle coordinate is usually expressed as a pseudorapidity, the approximation of the rapidity for ultra-relativistic particles:

$$\eta = -\ln \tan \frac{\theta}{2}; \tag{2.5}$$

it varies from 0 for $\theta = \pi/2$ and infinity for $\theta = 0$, as represented in Fig. 2.6. The regions of the detector with large η are often referred in the following as the "forward"

direction. It can be seen that differences of η are invariant under Lorentz boosts along the z axis. Thus, the spatial separation between two particles is expressed as $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$.

The momentum of a particle, as well as its energy, is also often given in a Lorentz boost-invariant form, i.e. as a $p_{\mathrm{T}} = \sqrt{p_x^2 + p_y^2}$ component in the transverse plane.

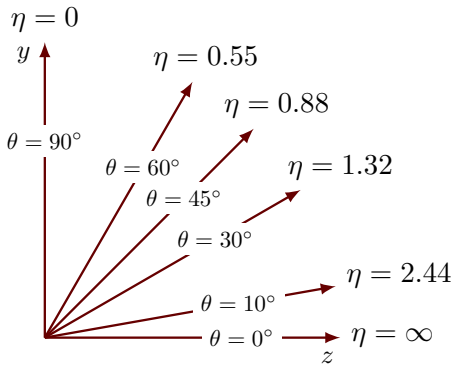


Figure 2.6 – Illustration of the relation between the pseudorapidity η and the polar angle θ [61].

2.2.2 Detector structure

The structure of the CMS detector is shown in a schematic view in Fig. 2.7. It has the shape of a cylinder with axis along the beamline; a variety of subdetectors, nested around the interaction point, are installed to probe various properties of different kinds of particles. The central part of the CMS apparatus is called "barrel", while the extremities covering the forward regions are called "endcaps".

Ordered by vicinity to the interaction point, the subdetectors composing the CMS can be grouped as follows.

The tracking system is where charged-particle trajectories, said "tracks", and their origin, or "vertices", are reconstructed by connecting the so-called "hits", i.e. the signals in the tracker in layers.

The electromagnetic calorimeter (ECAL) is the volume where electromagnetic showers, initiated by electrons and photons, occur; therefore, electrons and photons are detected as "clusters" of energy in neighbouring ECAL cells.

The hadron calorimeter (HCAL) is the volume where most of the hadron showers are exhausted, after depositing on average one third of their energy in ECAL.

The muon chambers are additional tracking layers placed in the most external region of the detector, as muons have little interaction with the rest of the detector material.

The central feature of the CMS design is a large solenoid magnet, about 13 m long and 3 m wide, surrounding the volume of the tracker and the calorimeters. Its role is to provide an intense magnetic field so that the electric charge of charged particles can be inferred, as well as their transverse momentum, from the bending of their trajectories.

A description of the CMS detector components is given in the following.

The solenoid magnet

The niobium-titanium superconducting solenoid magnet [63], operating at a 4.5 K temperature, provides an axial and uniform magnetic field of 3.8 T within its volume; the

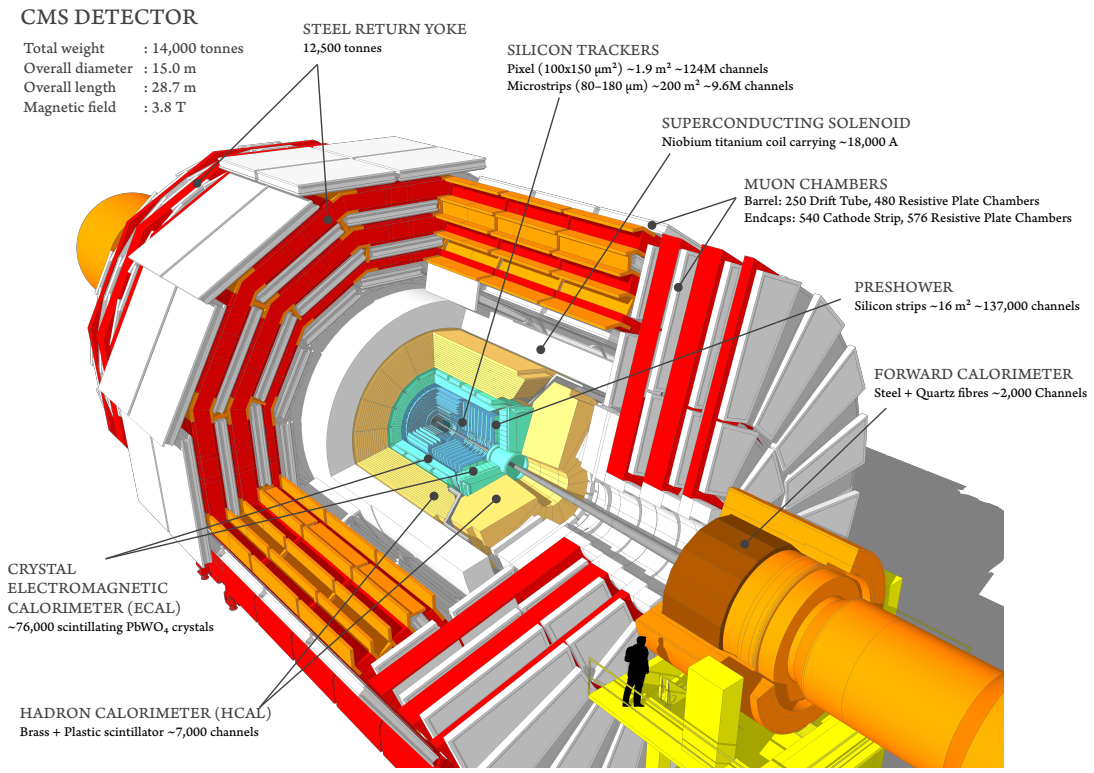


Figure 2.7 – Schematic view of the CMS detector [62].

muon system surrounds the magnet and is embedded in a steel return yoke, which confines the high magnetic field to the volume of the detector. Thus, the muon system is immersed in a magnetic field of about 2 T intensity. In addition to returning the magnetic flux of the solenoid, the steel plates play the role of absorber between the layers of the muon chambers.

The whole CMS detector design is driven by the solenoid constraints. By placing the magnet outside the volume of the tracker and calorimeters, the amount of material placed in front of the calorimeters is minimised and the link between tracks and calorimeter clusters is facilitated; the density of the calorimeters material must then be very high, to allow for the electromagnetic and hadronic showers to be exhausted within the volume bounded by the solenoid.

The track bending guaranteed by the intense magnetic field is powerful enough to provide a strong separation between neutral and charged particles: for instance, a charged particle with $p_{\mathrm{T}} = 20\,\mathrm{GeV}$ is deviated in the transverse plane by about 5 cm at the ECAL surface, which is a distance sufficiently large to resolve its energy deposit from that of a photon emitted in the same direction.

The inner tracking system

The inner tracking detectors [64, 65] are placed directly around the interaction point; they are confined within an outer radius of about 1.2 m and a length of 5.6 m.

Their design is motivated by several challenging requirements. The detectors need to be finely segmented to provide precise spatial measurement and a high vertex resolution. The "primary vertex" corresponding to the hard interaction should be well discriminated from additional interactions in the event; as shown in Fig. 2.4, the number of simultaneous interactions went up to 40 during Run 1 and exceeded 60 during Run 2. Additionally,

it should be possible to identify "secondary vertices" corresponding to the decays of heavy particles, such as τ leptons and hadrons made of b and c quarks; such secondary vertices are displaced by up to a few mm with respect to the interaction point. In a busy environment such as the region immediately close to the interaction point, the detector needs to be safe against the extremely high level of radiation. Finally, the substructure of the tracking system is optimized to minimize the amount of material, so that the performance of the energy measurement in the calorimeters is preserved. To cope with these experimental necessities, silicon detectors with diverse granularity based on their position are deployed.

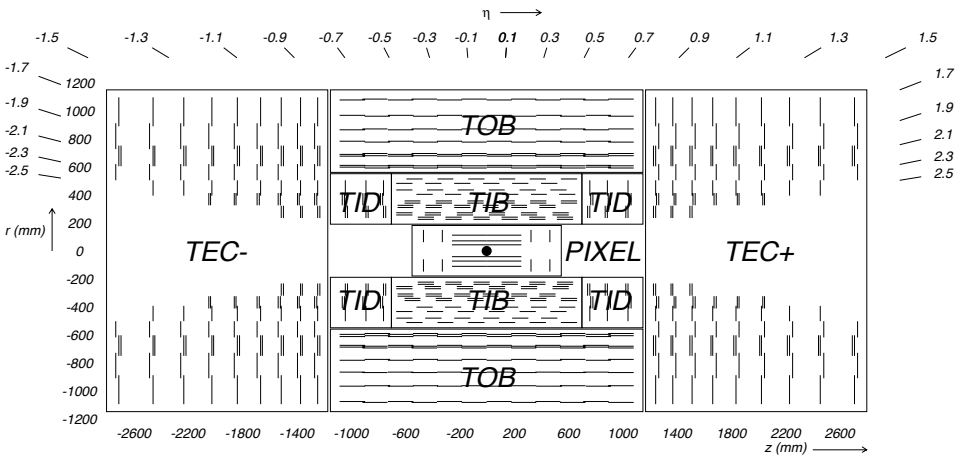


Figure 2.8 – Schematic longitudinal view of the CMS tracker in the rz plane. The original pixel detector is represented; however, the global layout of the tracker did not change after the pixel detector upgrade installed in 2017. Each line-element represents a detector module. Closely spaced double line-elements indicate back-to-back silicon strip modules, in which one module is rotated through a 'stereo' angle, as to permit reconstruction of the hit positions in 3D. The strip tracker detector consists of the tracker inner barrel (TIB) and the tracker inner disks (TID); and by the outermost tracker outer barrel (TOB) and the tracker endcaps (TEC) [66].

The longitudinal view of the inner tracking system is shown in Fig. 2.8. It consists of two main detectors: the inner pixel detector and the silicon strip detector.

The pixel detector covers a pseudorapidity up to $|\eta| = 2.5$ in the innermost region of the tracker (in the barrel, $29\,\text{mm} < r < 10\,\text{cm}$), where the flux of particles produced from the collisions is larger. To cope with the increase of luminosity foreseen for the 2017 and 2018 LHC operations, a pixel detector upgrade was installed in March 2017, profiting of the extended year-end technical stop that followed the 2016 data-taking. The layout of the original and upgraded pixel detectors is compared in Fig. 2.9. The current pixel detector consists of pixel cells of size $100 \times 150\,\mu\text{m}^2$ installed over four layers in the barrel and three disks in the endcap, providing a vertex spacial resolution in the range of 15-20 μm.

An intermediate region ($20 < r < 55\,\text{cm}$ in the barrel) is occupied by microstrip silicon detectors, typically large $10\,\text{cm} \times 80\,\mu\text{m}$. Finally, larger silicon strip detectors with typical size of $25\,\text{cm} \times 180\,\mu\text{m}$ are installed in the most external region ($55 < r < 120\,\text{cm}$). The resolution on the single point ranges form 20 to 500 μm in the radial direction and from 200 to 500 μm in the longitudinal direction.

Within a given layer, each module is shifted slightly in r or z with respect to its neighbouring modules; the overlap thus obtained allows the holes in the acceptance to be minimised.

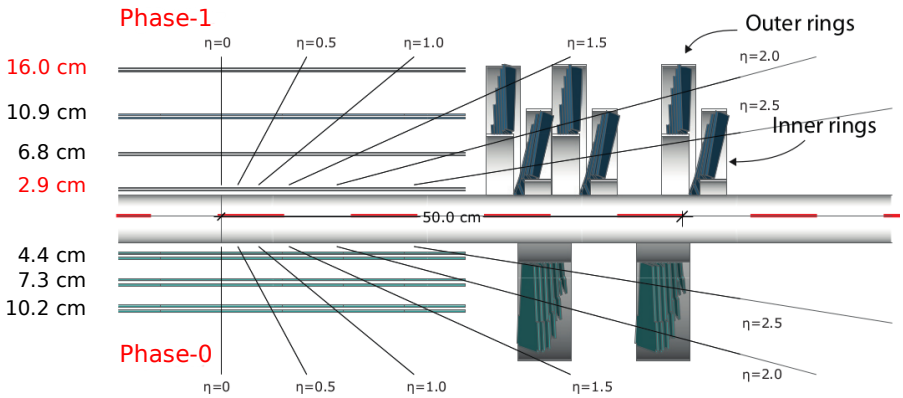


Figure 2.9 – Comparison of the layouts of the current upgraded pixel detector (top) and the original pixel detector (bottom).

A detailed view of the barrel tracking system in the xy plane is obtained through a hadrography technique, consisting in using reconstructed nuclear interactions to precisely map the positions of inactive elements surrounding the proton-proton collision point. The beam pipe position, the pixel detector and the first layer of the tracker inner barrel are visible in Fig. 2.10.

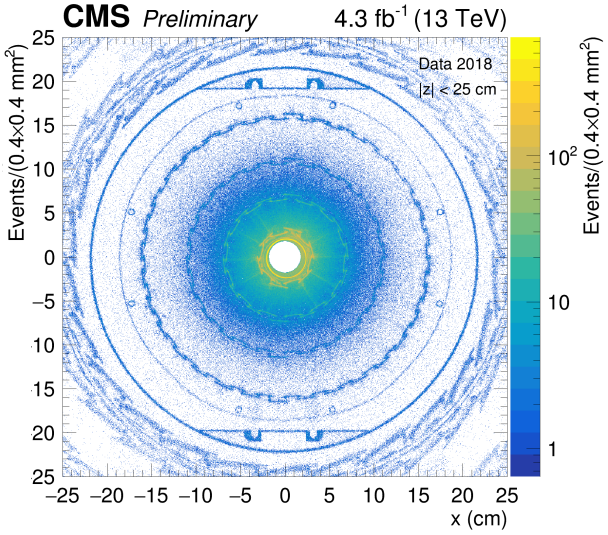


Figure 2.10 – Hadrography of the tracking system in the xy plane in the barrel region ($|z| < 25\,\text{cm}$). The density of nuclear interaction vertices is indicated by the color scale. The signatures of the beam pipe, the 4 layers of the barrel pixel detector with its support, and the first layer of the tracker inner barrel (TIB) detector can be observed above the background of misreconstructed nuclear interactions [67].

As shown in Fig. 2.11, the tracking material and the relative services, such as cables, support, and cooling system, represent a substantial amount of material in front of the calorimeters, up to 1.6 radiation lengths. The pixel material was significantly reduced by about 40% in the endcaps and by 10% in the barrel with the 2017 upgrade; thus, the impact parameter (IP) is better determined: the technical design studies show an expected IP resolution improved by up to a factor 1.5 in the longitudinal direction [68]. The pixel detector upgrade lead, for instance, to an increase of about 10% of the b tagging efficiency [69].

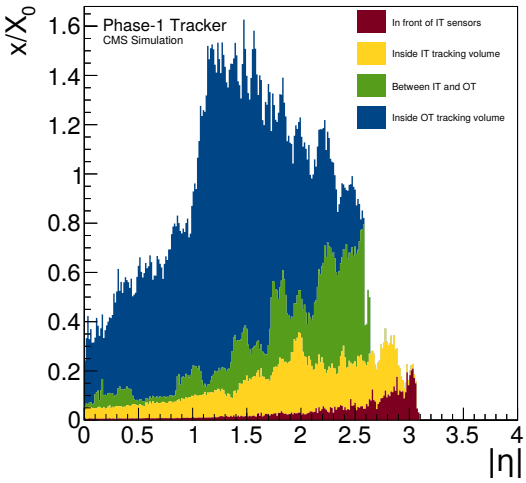


Figure 2.11 – Material budget simulation in as a function of η of the CMS tracker in units of radiation lengths X_0 [70].

The electromagnetic calorimeter (ECAL)

The ECAL [71] is a hermetic calorimeter dedicated to the measurement of the energy of electrons and photons; it consists of an array of lead tungstate ($PbWO_4$) scintillating crystals. The crystals are homogeneous: the medium is at once the absorber and the active material. It is designed so that the incident electron or photon initiates an electromagnetic shower and deposits most of its energy within the ECAL volume itself; the energy measurement is based on the light produced by the particles of the electromagnetic shower.

The longitudinal extension of the shower depends on the radiation length X_0 of the medium, i.e. the distance that an electron needs to cross before its energy is reduced by $1/e$ of its initial value; about 90% of its transversal dispersion lies within the Molière radius

$$R_M = \frac{X_0 \cdot 21.1[\,\mathrm{MeV}]}{E_C[\,\mathrm{MeV}]}\,, \tag{2.6}$$

where E_C is the critical energy for which the average energy loss by ionization equals the average energy loss by radiation.

Given its high density ($8.28\,\mathrm{g\,cm^{-3}}$), small radiation length (8.9 mm) and small Molière radius (22 mm), the lead tungstate allows a very compact calorimeter with high granularity to be deployed; moreover, it produces fast signals: in an average size ECAL crystal, 80% of the light is emitted within 25 ns, which is the time spacing between bunch crossings. However, the $PbWO_4$ has relatively low light yield (EY= 30 γ/ MeV); therefore, the crystal readout needs internal amplification.

The ECAL layout is shown in Fig. 2.12. The barrel covers the region $|\eta| < 1.479$ and is instrumented with about 61200 trapezoidal crystals. Each crystal covers a $22 \times 22\,\mathrm{mm^2}$ surface, equivalent to 0.0174×0.0174 in the $\eta - \phi$ plane, which matches the $PbWO_4$ Molière radius. The two endcap disks have acceptance $1.479 < |\eta| < 3.0$ and each consists of 7324 crystals of $28.6 \times 28.6\,\mathrm{mm^2}$ surface and 22 cm length. The average crystal length in the barrel and in the endcaps correspond to about 25.8 and 24.7 radiation lengths, sufficient to contain more than 98% of the shower produced by electrons and photons of energy up to 1 TeV. Such fine transverse granularity allows hadron and photon energy deposits as close as 5 cm from each other to be fully resolved.

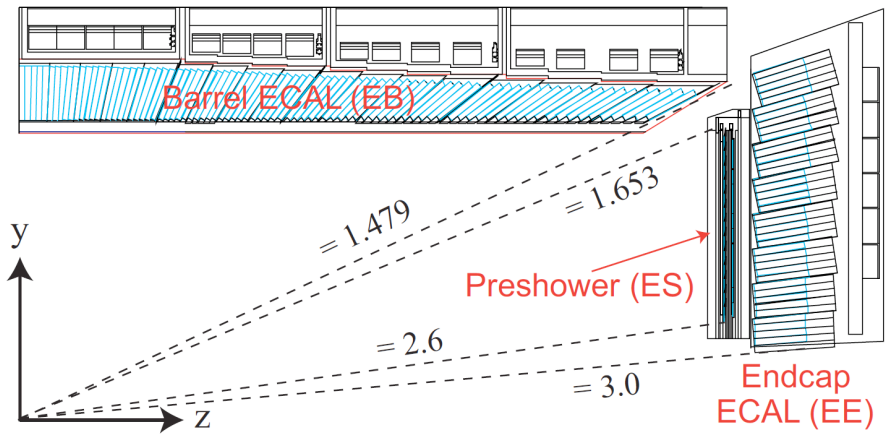


Figure 2.12 – Schematic longitudinal layout of a quadrant of the ECAL detector.

All crystals are oriented with axis tilted up to 3° compared to the position of the nominal interaction point to avoid acceptance gaps between the crystals. Silicon avalanche photodiodes (APD), designed be resistant to high radiation levels and to the intense magnetic field, collect and amplify the crystal scintillation light in the barrel, while vacuum phototriodes (VPT) are used in the endcaps. As detailed in Sec. 3.3 in the context of the 2017 data-taking conditions, each crystal is regularly monitored along the collision runs for transparency loss due to radiation damage; thus, adequate corrections are applied to compensate for the change in the crystal response.

ECAL performance measurements were performed using $\sqrt{s} = 7\,\text{TeV}$ proton-proton collisions, as documented in [72]. The measured energy resolution of electrons with transverse energy $E_\text{T} \sim 45\,\text{GeV}$ from Z boson decays is better than 2% in the central region with $|\eta| < 0.8$, and it ranges between 2% and 5% elsewhere; as for the photons, their resolution at $E_\text{T} \sim 60\,\text{GeV}$ ranges from 1.1% to 2.6% across the barrel and from 2.2% to 5% in the endcaps.

In front of each endcap, a much finer-graned preshower detector is installed to improve the discrimination of single photons from π^0 decays; it consists of a $1X_0$ and a $2X_0$ thick lead plates alternate with two layers of silicon detectors. However, parasitic signals originating from the large quantity of neutral pions produced by hadron interactions within the tracker material significantly affect the preshower identification and separation capabilities; therefore, its response is only marginally exploited in the reconstruction phase.

The hadron calorimeter (HCAL)

The HCAL [73] measures the energy deposited by hadrons; they typically lose about 30% of their energy in ECAL. As mentioned in Sec. 1.1, quarks cannot exist as free particles as a consequence of the QCD confinement; therefore, quarks produced out of the LHC collisions immediately hadronize, i.e. they fragment and produce hadrons. Similar hadronic showers are initiated by gluons. As a result, narrow jets made mostly of hadrons and photons are produced in the same direction as the quark or gluon that initiated the shower.

The global layout of the HCAL is shown in Fig. 2.13. The ensemble of the HCAL subdetectors has a large pseudorapidity coverage, up to $|\eta| = 5.2$. While the ECAL detector's crystals play at once the role of absorber and active scintillating material, the HCAL subdetectors use patterns of heavy absorbers and scintillator layers.

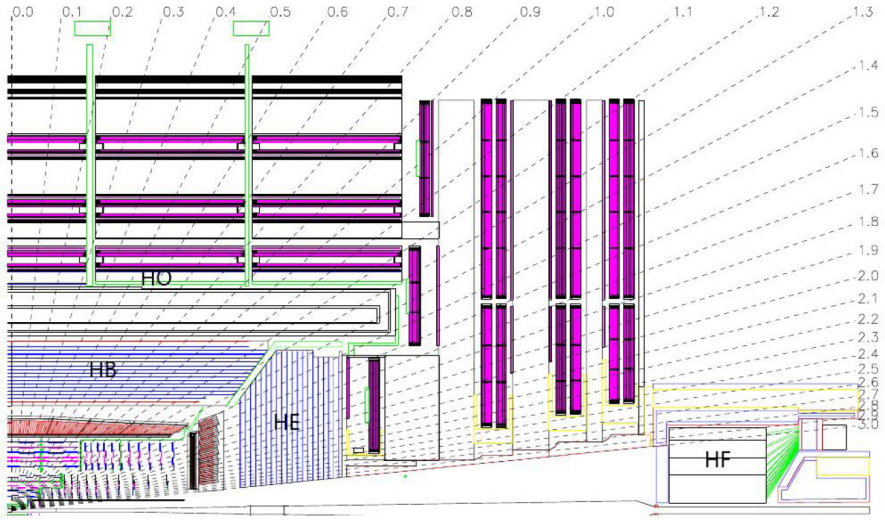


Figure 2.13 – Schematic longitudinal layout of a quadrant of the HCAL detector.

The Barrel Hadronic Calorimeter (HB), covering up to $|\eta| < 1.4$, consists of alternate layers of brass absorbers and plastic scintillator tiles. The relative size of the alternate materials is optimised to maximize the hadron interaction length λ_i within the volume constrained by the solenoid. Wavelength shifter fibers are embedded in the tiles and transmit the collected light to hybrid photodiodes. The Endcap Hadronic Calorimeter (HE) has similar design and covers the endcaps in the $1.3 < |\eta| < 3.0$ region. Globally, HB and HE are between $7\lambda_i$ and $10\lambda_i$ thick.

To improve the longitudinal confinement of the hadronic showers, a Outer Hadronic Calorimeter (HO) is placed outside the solenoid volume, covering the $|\eta| < 1.4$ region; it consists of scintillating material rings embedded in the yoke structure, read out by silicon photomultipliers (SiPM).

Finally, the $3 < |\eta| < 5.2$ region is covered by the Forward Hadronic Calorimeter (HF) calorimeter, deploying steel absorbers and quartz fibers, more resistant to the intense radiation acting on the forward detectors; the Cherenkov light produced by the quartz medium is collected by photomultiplier tubes (PMT).

From beam test analyses, the measured resolution of the ECAL + HCAL barrel calorimeters as a function of the energy E of the incident particles (pions) is

$$\frac{\Delta E}{E} = \frac{84.7\%}{\sqrt{E}} \oplus 7.4\% \tag{2.7}$$

in the energy range between 2 and 350 GeV [74]. The first term accounts for stochastic effects such as statistical fluctuations in the shower development; the second constant term is due to detector effects independent from the energy, such as imperfect calorimeter calibration. The modest energy resolution degrades the calorimeter-based reconstruction of jets and hadronic tau leptons; therefore, the reconstruction of such objects relies on the Particle Flow algorithm described in Sec. 2.3, which optimally exploits the whole detector to achieve improved energy and angular resolution.

The muon chambers

Muons produced in LHC hard scattering interactions, interesting for physics searches, have energy of a few to hundreds of GeV; in this range, their energy loss in the passage through the calorimeters and the detector inactive material is minimal, while all the

other charged particles are absorbed. Therefore, the muon detector system [75], whose role is the reconstruction of the muons momentum and electric charge, is the outermost CMS subdetector, placed outside the solenoid volume. As sketched in Fig. 2.14, three different detector technologies, all based on gas ionization, are deployed: Drift Tubes (DT) chambers, Cathode Strip Chambers (CSC) and Resistive Plate Chambers (RPC). Globally, their coverage extends up to $|\eta| = 2.4$. Performance studies performed with $\sqrt{s} = 7\,\text{TeV}$ proton-proton collisions can be found in [76].

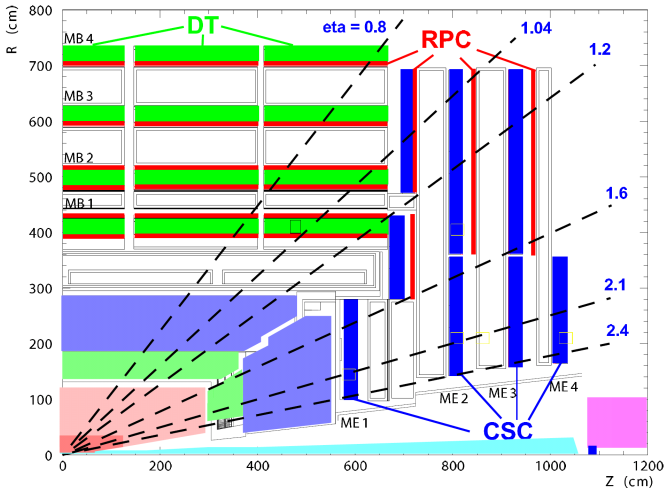


Figure 2.14 – Schematic longitudinal layout of a quardrant of the muon detector system.

The DT detectors surround the barrel and cover the central region with $|\eta| < 1.2$, where the muon occupancy and neutron background are low, and the magnetic field is uniform and as weak as 0.4 T. Each chamber is made of twelve levels of drift tubes filled with a mix of argon (85%) and CO_2 (15%). Their spatial resolution ranges form about 80 to 120 µm in the (r, ϕ) plane and from 200 to 300 µm in the z direction.

The CSC detectors can tolerate the strong and non-uniform magnetic field of the endcap regions and have very fast response, which is fundamental in regions with large muon rate. They consist of high granularity multi-wire chambers filled with a mix of argon (45%), CO_2 (50%) and CF_4 (10%) and provide a space resolution ranging from 30 µm for the finer chambers, placed at large $|\eta|$, to 150 µm for the wider chambers.

Finally, RPC detectors, partially overlapped to the other muon chambers in pseudorapidity, have excellent time resolution (smaller than 3 ns) and complement the measurement of the correct beam crossing time. Each level consists of two plates electrically charged and interlaid with a mixture of gas (95.2% $C_2H_2F_4$, 45% i-C_4H_10 and 0.3% SF_6), operated in avalanche mode.

2.3 Physics objects identification and reconstruction

A summary of the experimental signature of particles in the CMS subdetectors is given in Fig. 2.15.

Due to the high density of the detector materials, muons lose some energy in the innermost part of the detector. However, the fraction of energy lost is small; for instance, muons with 170 GeV momentum lose on average about 15% of their energy in the ECAL [77]. Being charged particles, muons are detected both in the inner tracker system and in the dedicated muon detectors. Electrons and photons are absorbed within the ECAL volume, where their energy is measured; while photons are neutral, electrons

are also detected in the inner tracker. Hadrons, after crossing ECAL with small energy loss, deposit most of their energy in HCAL; the trajectory of the charged components is detected in the inner tracker. Finally, neutrinos have negligible interaction with the detector material and escape undetected; however, the presence of neutrinos appears as an energy imbalance in the transverse plane, as clarified later.

A significantly improved reconstruction and identification of the physics objects produced in the event is achieved through the "particle flow" (PF) approach [78], which consists in reconstructing the stable particles by combining the information of all subdetectors. The nature of the particle (charged or neutral hadron; photon; electron; muon) is deduced and a combination of the different subdetector measurements is carried out to infer its momentum and direction. Thus, the resulting list of particles can be used to build higher level objects such as jets and tau leptons, to compute the missing transverse momentum, or to quantify the isolation of an energetic particle.

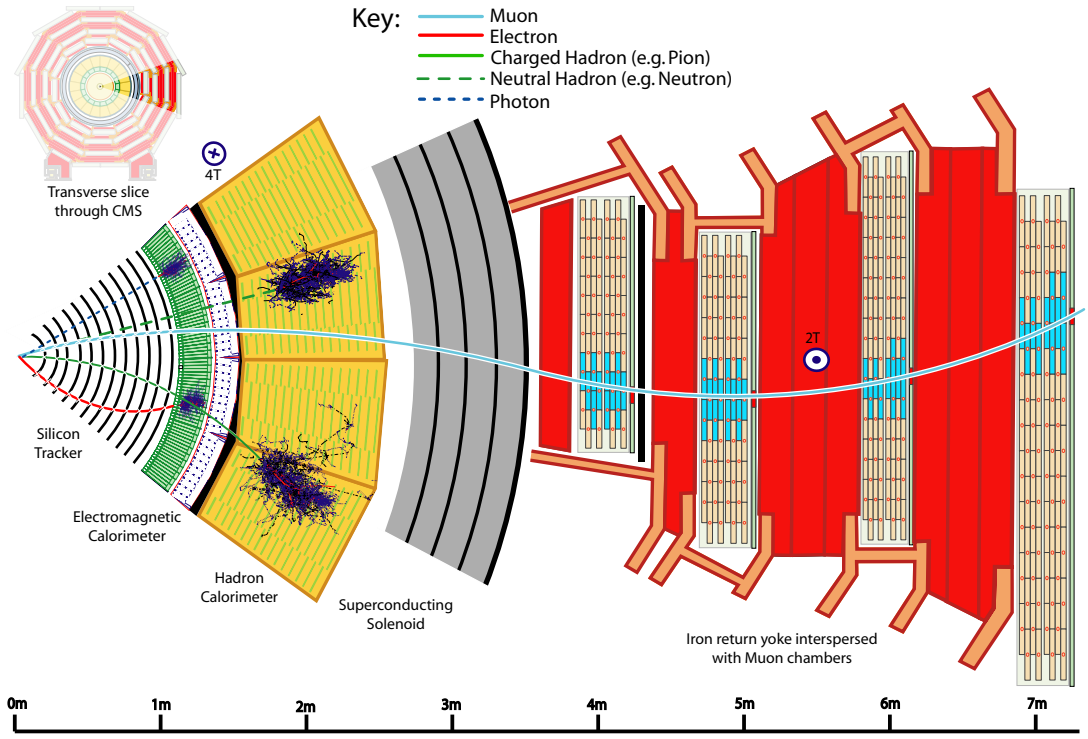


Figure 2.15 – Slice of CMS in the transverse view and experimental signature of particles in the subdetectors.

Given the intense 3.8 T magnetic field hosting the large tracker and the high resolution electromagnetic calorimeter, the CMS detector is ideally suited for a PF reconstruction. A simplified description is given in the following; an exhaustive documentation can be found in [78].

2.3.1 Particle flow basic elements

At a hadron collider, the production of hadrons is ultra-dominant; within jets, the fractions of energy globally carried by each particle type is about 65% for charged hadrons, 25% for photons (mostly π^0) and 10% for neutral hadrons. The transverse momentum of the charged hadrons rarely exceeds 100 GeV; at this energy, the relative resolution of the transverse momentum is better than 2% in the tracker, while the calorimeter relative

resolution on hadrons is about 10%. Therefore, the momentum measurement of charged hadrons relies primarily on the tracker.

The energy of photons is measured in the ECAL with high precision, while the one of neutral hadrons is measured in HCAL.

Tracking

An iterative approach is adopted to achieve an efficient tracking with low fake rate. In a first instance, only the reconstruction of isolated tracks compatible with the beam position, with clear associations of hits, is attempted. The fitted track is required to pass quality criteria. If selected, the hits of the corresponding tracks are masked and a second iteration of track reconstruction is launched. At each iteration, the constraints are relaxed and/or non-trivial pixel hits associations are attempted to form the track seeds. At the end of each iteration, the track properties should match specific quality criteria. The iterative approach allows the combinatorics when associating the hits to be limited; thus, smaller CPU time is needed.

The track reconstruction goes over about twelve iteration steps, some of them targeting the core of the jets and some optimized for the muon reconstruction.

Clustering

The PF algorithm requires that the energy deposits of the particles are properly associated to tracks; the matching efficiency depends on the ability to reconstruct the shower into a cluster. The PF should efficiently reconstruct particles within jets, which implies that the clustering algorithm needs to be able to disentangle overlapping showers. This goal is achieved through an iterative clustering technique with a built-in lateral shower profile, whose role is sharing the energy of the cells belonging to nearby clusters. In practice, less than five iterations are needed for the algorithm to converge. The same algorithm, with different parameters, is used in the ECAL and in the HCAL.

2.3.2 Muon reconstruction

The regular muon reconstruction used in CMS is not PF-specific as they have a very clean signature in the dedicated muon trackers, which provide high efficiency over the full detector acceptance; moreover, the CMS design is optimised so that all the other particles (except neutrinos) are absorbed in the calorimeters; therefore, the muon purity in the outermost detector is very high.

Three reconstruction algorithm are implemented to exploit the subdetectors information.

Standalone muon is reconstructed by fitting only hits in the muon detectors.

Tracker muon is reconstructed in the inner tracker; the extrapolated track must be compatible with at least one track segment reconstructed in the muon detectors.

Global muon if the parameters of the track reconstructed in the inner tracker and in the muon chambers, propagated to a common surface, are compatible, they are merged into a global muon candidate.

About 99% of the muons produced within the geometrical acceptance of the muon system are reconstructed either as a global muon or a tracker muon or as both, in which case they are merged in a single candidate.

Low p_{T} (below 10 GeV) muons often fail the global-muon reconstruction requirements, due to larger multiple scattering in the yoke material; they are often reconstructed only as tracker muons. Charged hadrons can also be reconstructed as muons if the hadron shower exceeds the HCAL volume; thus, they contaminate the low p_{T} tracker muons collection. Besides, the default identification criteria turns out to be tighter than necessary at high p_{T}. Therefore, different criteria are applied to the muon tracks in the PF algorithm in order to achieve a better balance between identification efficiency and purity.

2.3.3 Particle flow particle reconstruction

The association of the particle signatures in different subdetectors is essentially geometric: a track is associated to a cluster if its extension is within the boundaries of the calorimeter cluster. Similarly, two clusters overlapping in the (η, ϕ) plane are associated. Thus, the PF algorithm does not need to tackle the entire detector at once: it acts on a closed set of objects connected together. Once all the ingredients are prepared, the flow of the PF reconstruction can be simplified as follows.

The muons are selected first: their track and possible associated clusters are put aside for the rest of the algorithm.

Then, electrons are reconstructed. Due to the material in the tracker and the intense magnetic field, electrons can irradiate photons so that several ECAL clusters are produced. These clusters, spread in ϕ, and are grouped together into a super-cluster, which is used to seed a specific track reconstruction algorithm based on a Gaussian Sum Filter [79, 80]. Thus, the track is correctly reconstructed in spite of the curvature changes due to photon emission. A second tracker-based approach seeds this tracking algorithm [80] and increases its efficiency. The electron candidate is finally required to pass a quality cut. If it is selected, its constituting elements (the track, the clusters) are put aside for the subsequent treatment. An additional dedicated step targeting the isolated high p_{T} photons is described in [78] and it is not applied at this stage.

The remaining tracks are natural charged hadrons candidates. The compatibility between the track momentum and the calorimeter energy, taking into account the track reconstruction uncertainty and the estimated cluster energy resolution, is evaluated. Mismatches are typically due to an excess of a calorimeter energy measurement due to an overlapping neutral particle. In this case, a photon candidate carrying the difference of energy and, if needed, an additional neutral hadron candidate are created. The momentum of the charged hadrons is a combination of the tracker and the calorimeter measurements, dominated by the former.

Finally, the remaining isolated clusters are turned into photon candidates in the ECAL and into neutral hadron candidates in the HCAL.

2.3.4 Higher level objects reconstruction

The collection of particles reconstructed by the PF algorithm is fed to the higher level object reconstruction algorithms, as if they were generated simulated particles. Since the inputs are already calibrated, both the resolution and the response are improved with respect to a solely calorimeter-based approach.

Jets

Jets are reconstructed by clustering the PF candidates through the "anti-k_{T}" algorithm [81]. It is a recursive algorithm that clusters PF candidates pairwise to build

a larger object. At each iteration, starting from the PF particles, the two candidates closest to each other by a certain metric are paired and recombined in a composite candidate; a jet is build up from the recombination until no pair has distance smaller than a parameter R. The metric used by the anti-k_T algorithm clusters the high p_T pairs first and disfavours clusters of low p_T candidates, so that the jet is built in a conical shape around a hard core and the soft radiation at the borders is suppressed.

The jet resolution is significantly improved by the PF approach, as shown in Fig. 2.16. For instance, for 30 GeV jets in the barrel, the energy resolution is improved by a factor two and the response is 90%, against the 50% obtained from the calorimeter measurements alone. The jet direction resolution is also improved by more than a factor two thanks to the track measurements [78].

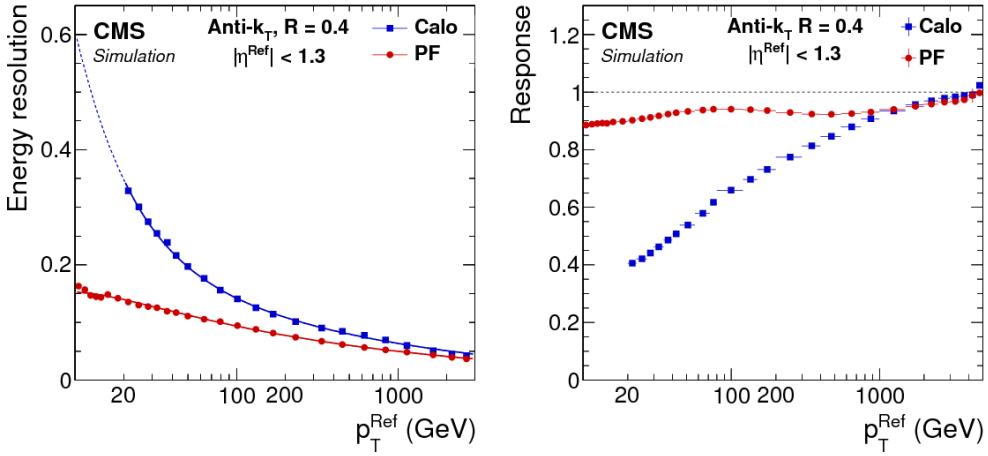


Figure 2.16 – Jet p_T relative resolution (left) and response (right) in the barrel obtained through a PF-based approach and from measurements solely calorimeter-based [78]. The jet response is defined as the ratio of arithmetic means of matched reconstructed and particle-level jets transverse momenta in bins of particle-level p_T.

Tau leptons

The tau lepton is an unstable particle and it decays immediately after its production. Its decay produces either a lighter lepton and two neutrinos or a few hadrons and one neutrino. While the leptonic tau decay is reconstructed as a regular electron or muon, the hadronic tau lepton decay gives rise to a composite object, similar to a jet, reconstructed from its PF constituents. The tau lepton decay branching fractions are listed in Tab. 2.2. In the following, the neutrino is omitted from the hadronic tau decay notation.

For the hadronic tau lepton (τ_h) reconstruction, PF jets reconstructed as described in Sec. 2.3.4 are used as input to the hadrons-plus-strips (HPS) algorithm [82]. The PF jet must have $p_T > 14$ GeV and $|\eta| < 2.5$; its constituent particles must match either of the $h^\pm$, $h^\pm\pi^0$, $h^\pm h^\mp h^\pm$ decay modes. Photons from the π^0 decay are likely to convert before reaching the ECAL, due to the significant amount of tracker material; the conversion probability is larger than 50% for $|\eta| > 1.5$. Through an iterative procedure, photons are dynamically clustered into ECAL "strips" of $\Delta\phi$ and $\Delta\eta$ size that depends on the E_T of the cluster itself.

Missing transverse energy

As mentioned in Sec. 2.2.1, the momentum of the constituents of the protons involved in the fundamental interaction carry an unknown fraction of the proton momentum;

Table 2.2 – Tau lepton branching fractions [7].

Decay mode	BR [%]
$e\nu_e\nu_\tau$	17.8
$\mu\nu_\mu\nu_\tau$	17.4
$h^\pm\nu_\tau$	11.5
$h^\pm\pi^0\nu_\tau$	26.0
$h^\pm\pi^0\pi^0\nu_\tau$	10.8
$h^\pm h^\mp h^\pm\nu_\tau$	9.8
$h^\pm h^\mp h^\pm\pi^0\nu_\tau$	4.8
other hadronic modes	1.8
all hadronic decays	64.8

however, their momentum is negligible in the transverse plane. Therefore, by the conservation of the momentum, particles escaping the detection, such as neutrinos, are observed as an energy imbalance in the transverse plane. The missing transverse momentum $p_{\mathrm{T}}^{\mathrm{miss}}$ is computed from PF particles as

$$\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}} = -\sum_{i=1}^{N_{part}} \vec{p}_{\mathrm{T}}^{\,i}. \tag{2.8}$$

Its resolution is better by about 40% compared to the computation only with based on calorimeter measurements, as shown in Fig. 2.17.

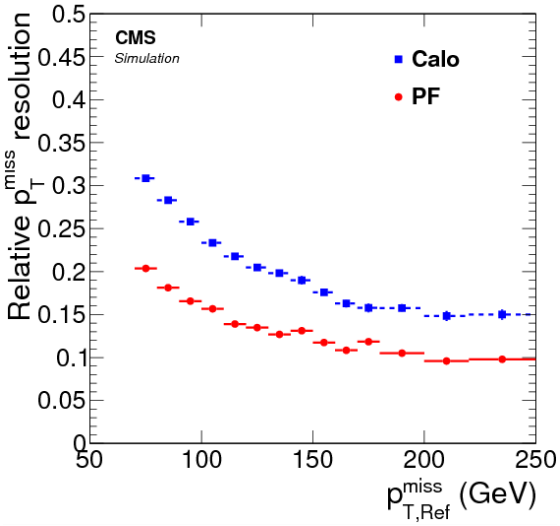


Figure 2.17 – Relative $p_{\mathrm{T}}^{\mathrm{miss}}$ resolution obtained through a PF-based approach and from measurements solely calorimeter-based [78].

2.4 Trigger system

For a high-luminosity experiment such as CMS, a trigger system is essential to probe the most interesting processes. Indeed, the number of events collected per unit of time for any process is given by its cross section σ and the instantaneous luminosity achieved by the collider (see Sec. 2.1.2). Not surprisingly, many of the processes of interest for the

LHC physics program have small cross section, as shown in Fig. 2.18. The cross section of the dominant Higgs boson production mechanism, for instance, is about 50 pb, which is 9 orders of magnitude smaller than the total proton-proton interaction cross section $\sigma_{pp} = 69\,\text{mb}$ [55].

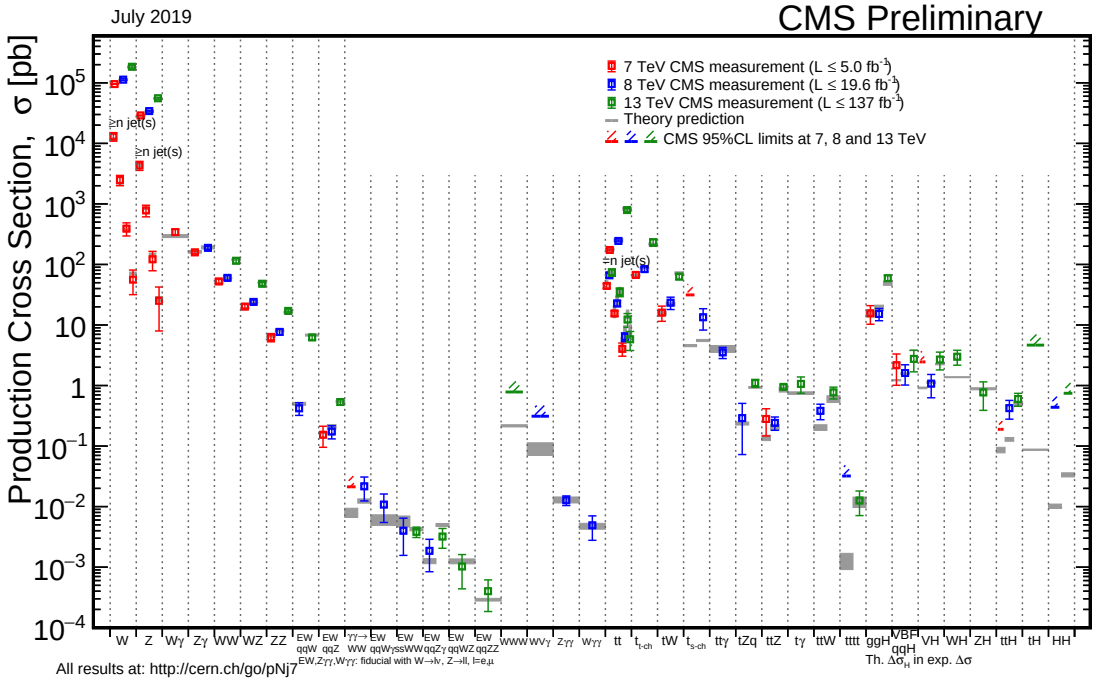


Figure 2.18 – Summary of the cross section measurements of Standard Model processes [83].

On one hand, a high luminosity is needed in order to enhance the number of signal events; in nominal conditions, the LHC provides a $L = 10^{34}\,\text{cm}^{-2}\,\text{s}^{-1}$ instantaneous luminosity. On the other hand, the resulting rate of bunch crossing, of about 40 MHz, is far beyond the sustainable 1 kHz event output rate and the $\sim 3\,\text{Gb/s}$ throughput rate. Reducing the event rate while retaining those that are interesting for physics searches is the main goal of the trigger system, described in the following.

The CMS experiment implements a trigger system [84] operating on two consecutive levels, making increasingly sophisticated decisions in an extremely short time. Globally, it achieves a 10^5 event rate reduction.

The first, fast operations of skimming are performed by the Level-1 (L1) trigger. It operates only at the hardware level and uses the information of the calorimeters and the muon chambers to make a decision in about 3.8 μs, providing input events for the next trigger level with a event rate of 100 kHz.

The High Level Trigger (HLT) exploits the full detector information and makes a software-based decision in about 200 ms. The selected events, reduced to a 1 kHz rate, are then available for the storage and the full reconstruction.

2.4.1 The Level-1 trigger

The L1 trigger is the first system in the data acquisition chain; therefore, it needs to be as flexible and scalable as possible to accommodate widely changing data-taking conditions and to reliably collect data for physics searches in the harshest scenarios.

As mentioned in Sec. 2.1.4, starting from 2015 the LHC has been delivering collisions with an instantaneous luminosity well beyond its design specifications. A high luminosity

poses several challenges to the data-taking: it implies a large event rate and a large pileup (see Sec. 2.1.2).

The pileup interactions give rise mainly to soft collisions, uninteresting to the physics searches. They however result in additional energy in the calorimeters with respect to the energy deposited by the hard interaction products; therefore, the object reconstruction at trigger level largely depends on pileup.

However, as the luminosity increases, the boundaries for the output rate fo the L1 trigger (100 kHz) remain the same; to fit in this budget, a harder event selection needs to be applied.

Therefore, the original L1 trigger system, designed to face the nominal LHC operations, could not ensure an adequate physics acceptance in the extreme conditions of the ordinary 2017-2018 collisions runs, with foreseen luminosity as large as $L = 2.2{\cdot}10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ and pileup up to 60 simultaneous collisions. A major upgrade [85, 86] of the L1 trigger was installed and commissioned between 2015 and 2016; the electronic boards were replaced and advanced mezzanine cards (AMC) mounting powerful field-programmable gate arrays (FPGA) were introduced.

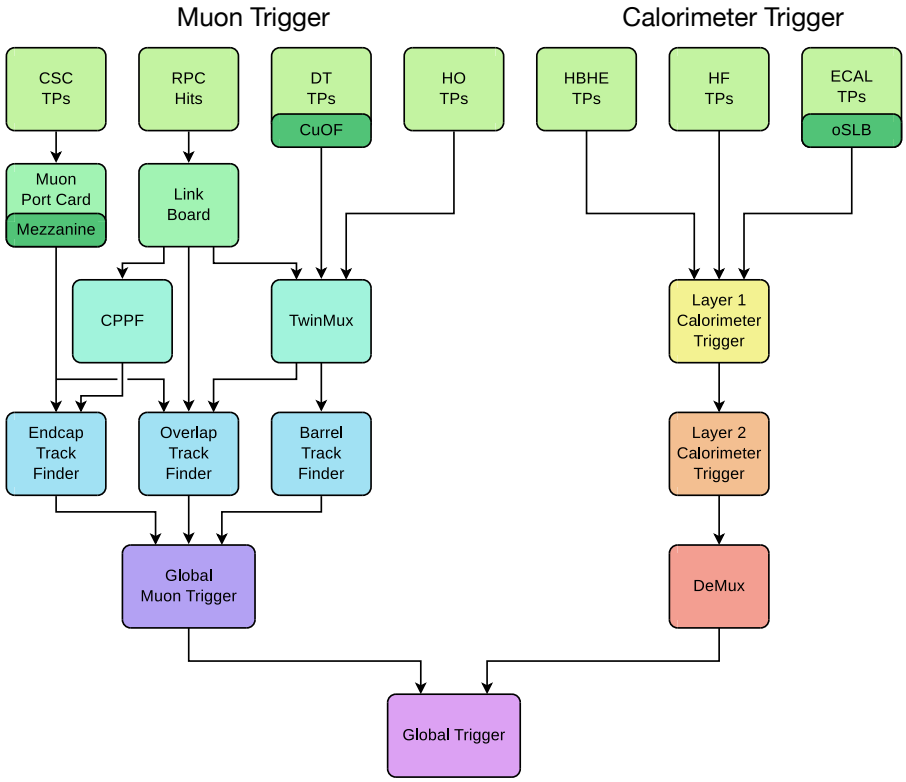


Figure 2.19 – Layout of the L1 trigger system [86].

A flowchart of the upgraded L1 trigger logic is shown in Fig. 2.19. To operate a decision, the L1 trigger collects the information from the calorimeters and from the muon detectors separately. The calorimeter information is read in units called "trigger towers" (TT); the transverse energy measured in the calorimeters is transmitted to the L1 Calorimeter Trigger in the form of "trigger primitives". The information is transmitted through the Time Multiplexed Trigger (TMT) architecture, represented schematically in Fig. 2.20, which enables the whole calorimeter data to be processed at once by a single trigger processor. Compared to the Run 1 L1 trigger, processing regions of 4×4 TT in parallel,

the TMT architecture allows the energy to be sampled with full TT granularity, so that the resolution is significantly improved.

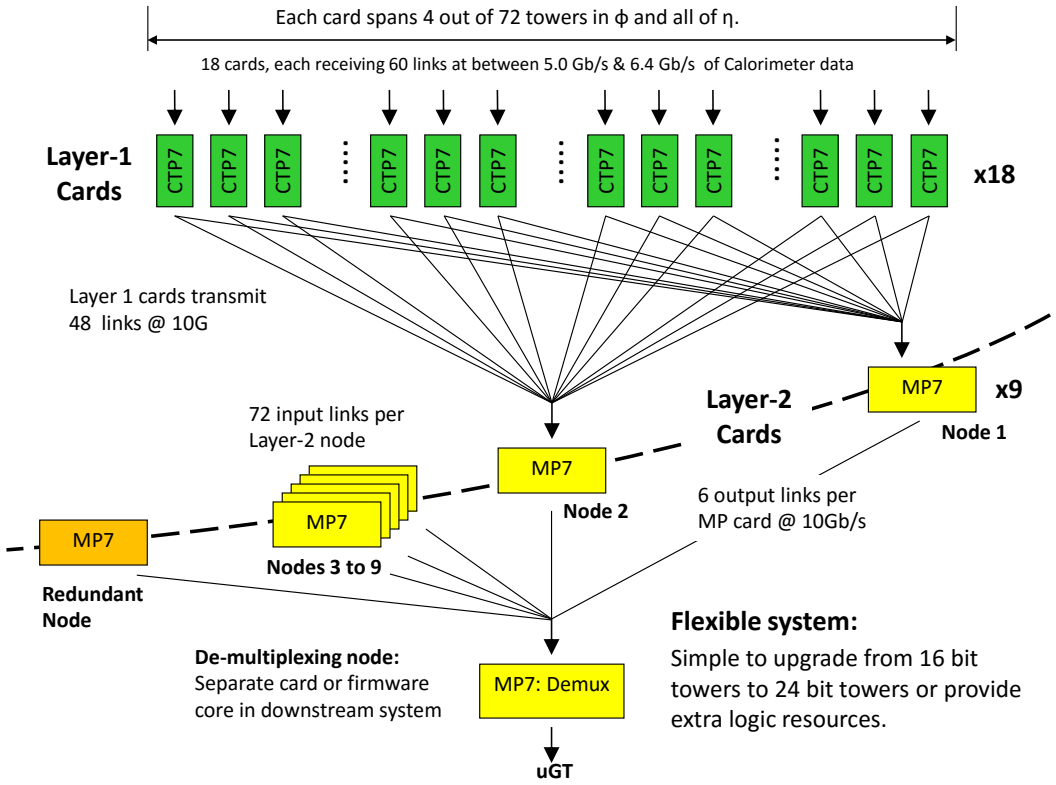


Figure 2.20 – Layout of the Time Multiplexed Architecture deployed in the upgraded L1 Calorimeter Trigger. The inputs from the calorimeter subdetectors are pre-processed in the Layer-1 and distributed to the processing nodes in the Layer-2. The output is collected by a demultiplexer node and transmitted to the μGT. "MP" stands for "Master Processor"; "CTP" stands for "Calorimeter Trigger Processor" [85].

The trigger primitives are then combined to reconstruct calorimeter trigger objects: electrons, photons, hadronic tau leptons, jets and energy sums.

As the L1 trigger does not use the inner tracker information, electron and photons are reconstructed together as L1 e/γ objects. The shape of the corresponding clusters in ECAL and HCAL is exploited to discriminate them against jet-like clusters. The L1 e/γ algorithm implements an isolation computation based on Look-Up-Tables (LUT) as a function of the position and of the pileup, which results in a reconstruction efficiency larger than 90% over a wide pileup range.

The L1 hadronic tau candidates are reconstructed through the dynamic clustering of basic L1 e/γ objects, using both the ECAL and HCAL energy information; neighbour clusters are merged into a single L1 hadronic tau candidate to take into account for multi-particle decays. The key feature of the L1 τ_h trigger is the isolation: it allows the background to be kept under control and the purity to be increased. The isolation thresholds are η, E_T and pileup dependent, so that the rate is reduced in challenging data-taking conditions and the p_T thresholds are kept suitable for Higgs physics. In spite of the high pileup, the thresholds were maintained in the 32-38 GeV range throughout the 2017 and 2018 data taking.

The L1 jet candidates are reconstructed through a sliding-window algorithm as 9×9 TT clusters centered on local maxima of deposited E_T; a pileup subtraction using the energy

in the surrounding region is implemented. The L1 jets are also used to compute energy sums, providing a fair estimate of the missing transverse momentum.

As for the muons, the upgraded trigger implements a coordinate-oriented architecture: the Barrel/Overlap/Endcap Muon Track Finders (BMTF, OMTF, EMTF) assign η, ϕ, p_{T} and quality criteria to each candidate, reconstructed from the combined information of the three muon detectors; the muon candidates are transmitted to the Global Muon Trigger (μGMT), which ranks them by p_{T} and quality and removes the duplicates. In Fig. 2.21, the di-muon invariant mass computed with L1 objects is shown with (orange curve) and without (blue curve) track extrapolation to the vertex, implemented since 2017; the orange curve shows a clear improvement in the di-muon resolution, which peaks towards the low mass resonances.

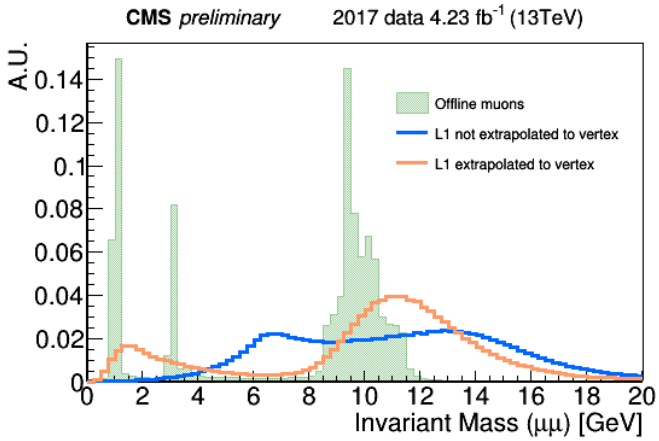


Figure 2.21 – The di-muon offline spectrum, peaking at the mass of the Φ and ϒ resonances, is compared to that obtained from L1 muons with and without L1 track extrapolation to the vertex. The resonances are well visible in the orange curve [87].

Finally, the event is kept or discarded based on the decision made by the Global Trigger (μGT), which takes in input the information processed by the L1 Calorimeter Trigger and the L1 Muon Trigger. The L1 selection algorithms are commonly called "seeds".

The L1 trigger selection can thus rely on a versatile architecture, allowing complex correlations among trigger objects to be computed and sophisticated dynamic clustering algorithms to be implemented. Moreover, the L1 trigger potential can be exploited to design specific analysis-targeted algorithms: the first dedicated L1 VBF trigger, extensively described in Ch. 3, makes profit of these capabilities. The L1 seeds design and monitoring requires access to the L1 trigger information for data and simulations; the emulated L1 trigger response in simulated events and the production of L1 trigger objects are performed through a bit-by-bit emulation of every L1 trigger subsystem by the CMS L1 emulation software, described in [88].

2.4.2 The High Level Trigger

The HLT online object reconstruction is largely based on the algorithms used offline and shares a large fraction of the code and of the framework with them. The computing intensive steps, such as the tracking, are run with simplified configurations to speed them up; for example, in the case of the tracking, the algorithms are simplified by requiring less hits. The PF approach is largely used for the HLT candidates reconstruction. The complex operations performed at the HLT are possible thanks to its implementation in

a single processor farm ("Event Filter Farm") with more than 13000 CPUs (up to 30000 in 2018), which can handle at once the full L1 trigger information.

The HLT candidates are generally seeded by L1 objects and reconstructed around them. The set of steps and requirements performed by a HLT reconstruction algorithm is commonly referred to as "HLT path"; within a path, the binary decisions on the object or event quality are called "filters". The ordering of the filters in a path is tuned so that the most discriminant filters are applied first, so that the computation of higher level observables used to apply the subsequent filters is less expensive.

The full HLT selection uses more than a hundred paths, split between physics paths and monitoring paths; the latter are selected with a rate reduced by a "prescale" factor, so that they collect only enough event statistics for efficiency measurements and background studies. The HLT paths are run one after the other or in parallel when possible.

The HLT paths exploited for the data analysis presented in this thesis are described in Sec. 4.2, together with their performance.

Chapter 3

The Vector Boson Fusion Trigger

As the HH production cross section at the LHC is very small compared to the background processes, one of the major experimental challenges in this search is maximising the event collection efficiency. As mentioned in Sec. 1.2.1, specific Vector Boson Fusion studies in the HH $\rightarrow$ bb$\tau\tau$ search can provide information on λ_{HHH} and additional couplings, even if the cross section of the HH production mode through Vector Boson Fusion is about 20 times smaller than that of the production through gluon fusion. Furthermore, the most sensitive HH $\rightarrow$ bb$\tau\tau$ analysis channel, where the two tau leptons decay hadronically, used to rely only on di-τ_{h} triggers until 2016. The corresponding analysis selection, mainly driven by the trigger requirements, is critical for tau leptons and it is even harder on VBF signal events than on those where the Higgs boson pair is produced through gluon fusion: in the former case, the taus have softer p_{T} spectra, populating a region that is inaccessible for the di-τ_{h} triggers and, thus, for the analysis. After trigger and object selection, the VBF signal event yield is about 45 times smaller than that of the gluon fusion signal. The kinematics of the signals and the event selection are detailed in Ch. 4.

The VBF peculiar topology is a useful handle to identify signal events and a strategy can be defined exploiting its signature to extend the acceptance of the signal, starting from the earliest stages of the event selection. The activities regarding the L1 VBF trigger algorithm were a substantial part of my PhD work, covering the phases of the design of the algorithm, of its optimisation towards the 2017 data taking and of its maintenance online; I had the opportunity to show this study in a poster at the 2017 European Physics Society (EPS2017) conference [89]. To make available the events collected with the L1 VBF trigger in the HH $\rightarrow$ bb$\tau\tau$ analysis, I computed the HLT VBF H $\rightarrow \tau\tau$ trigger efficiencies that are necessary to derive correction scale factors to be applied to simulated events. This effort finally allowed a VBF event category in the HH $\rightarrow$ bb$\tau\tau$ analysis to be populated by events that could not have been selected through the existing trigger strategies; although the VBF seed potential was eventually not fully exploited, the full selection chain based on it realistically brings about 17% of additional events.

The benefit of a VBF trigger strategy goes well beyond its use in the HH $\rightarrow$ bb$\tau\tau$ analysis: for example, the VBF event category is one of the most sensitives for the H $\rightarrow \tau\tau$ search [90]. For this reason, the acceptance gain on the VBF H $\rightarrow \tau\tau$ signal was one of the main figures of merit in the design of the L1 VBF algorithm. In this chapter, the steps that lead to the design and the commissioning of the L1 VBF algorithm will be described, as well as the subsequent measurement of the HLT VBF H $\rightarrow \tau\tau$ trigger performance.

3.1 The L1 VBF trigger

3.1.1 Trigger design

The studies for the design of the VBF trigger were performed in 2016, targeting a proposal to be included in the 2017 L1 trigger selection. As mentioned in Ch. 2, the 2017 LHC operations were expected to be particularly challenging and the L1 trigger strategies were re-arranged in 2016 to get ready for harsh data taking conditions: the typical instantaneous luminosity foreseen for 2017 was $2 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, exceeding the nominal value of the LHC design and giving a number of simultaneous interactions per bunch crossing (pileup) as high as 60. The major L1 system upgrade was installed and commissioned in this context, in 2015 and 2016. The luminosity peak was eventually $2.07 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ in 2017 [91] and $2.14 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ in 2018 [92].

The optimisation of a L1 algorithm must reflect technical and physics needs. The role of the trigger system is to collect, out of extremely busy collision runs, those events that are potentially interesting for physics analysis, recording as many of them as affordable by the data acquisition, storage and Tier0 throughput rate while using the least possible resources.

The full set of L1 trigger selections, called "L1 menu", is configured to perform optimally in different luminosity and pileup conditions, and the thresholds are constantly adjusted to have a total rate smaller than 100 kHz. Different luminosity settings are available, with identical copies of each L1 seed differing only in their thresholds, to allow the L1 menu to be adapted to the LHC evolving running conditions along the collision fill. For an algorithm to be included in the L1 menu, it is of utmost importance that it provides a high rate reduction, rejecting the background that arises mainly from generic QCD processes, so that it can fit in the total available L1 rate bandwidth.

The rate produced by a L1 seed can be reduced by tightening the selection criteria on the objects involved in the algorithm. Typical criteria are the minimum transverse energy of the objects or the accepted η range, but also sophisticated reconstruction algorithms using dynamic clustering and isolation working points targeting different kinematic regimes are implemented in L1 seeds.

The selection criteria shall also achieve a high efficiency on the signal, as passing the L1 selection is a necessary condition for the events to be kept and move along the rest of the acquisition system chain, and to provide a high-purity event collection, as the time and storage resources are limited and are not to be wasted on events that will not make it to the final stages of the physics analyses. The main challenge in the design of a L1 algorithm is to tune the selection criteria to keep a manageable rate while preserving the efficiency on the signal: the tighter the selection, the smaller the acceptance.

In the case of the VBF Higgs boson production, the most discriminating characteristic of the signal signature consists of the jets produced in the process. A Higgs boson can be produced from the fusion of two vector bosons (Z or W) radiated by the incoming quarks; despite the emission of the vector bosons, the incoming quarks usually only lose a very small amount of their longitudinal energies and the energetic outgoing quarks hadronize producing jets ($p_{\mathrm{T}} \sim m_{\mathrm{H}}/2$) in the forward regions of the detector. As a result, the invariant mass of the two jets thus produced is large, as well as their angular separation. These two variables are very discriminating, as can be observed in Fig. 3.2, where the VBF $\mathrm{H} \to \tau\tau$ signal is compared to $\mathrm{H} \to \tau\tau$ events produced through gluon fusion and to the $\mathrm{Z} \to \tau\tau$ process, which is the dominant background in the $\mathrm{H} \to \tau\tau$ analysis. A L1 trigger aiming at increasing the VBF event collection efficiency will logically make use

of one of these variables; although both options are valid, only the choice of designing a VBF seed with a jet-jet invariant mass was explored.

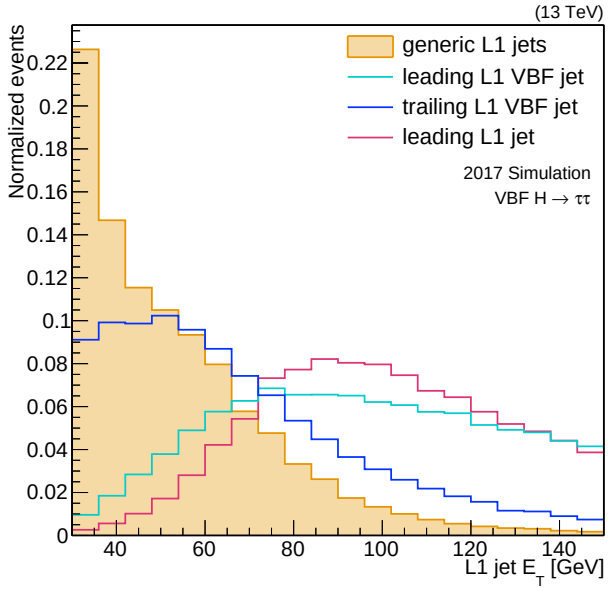
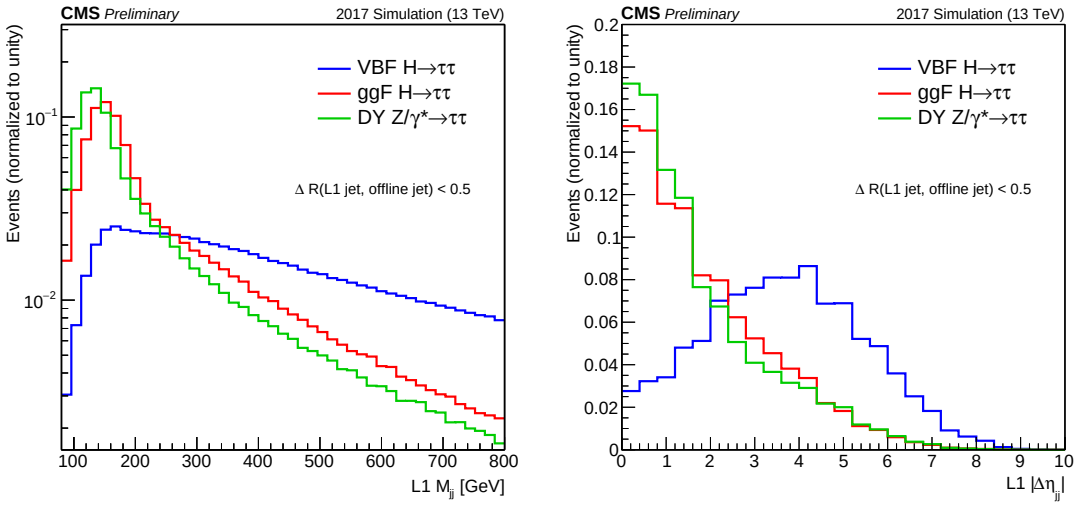
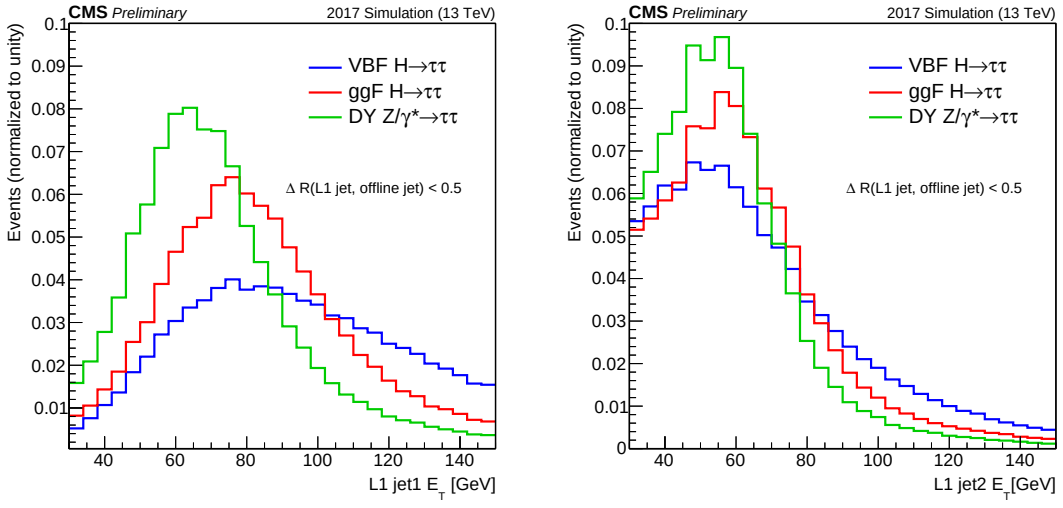


Figure 3.1 – Transverse energy distributions of L1 jets in VBF H $\rightarrow \tau\tau$ simulated events, produced with the MadGraph5_aMC@NLO generator at LO precision [47]. The yellow curve represents all the L1 jets in the event; the leading and trailing L1 VBF jets (blue and cyan curves) are L1 jets that are geometrically matched with $\Delta R(\text{L1jet}, \text{VBF quark}) < 0.3$ to the simulated VBF quarks at the generator level; the red curve represents the E_{T} of the leading L1 jet in the event.



(a) (b)

Figure 3.2 – Invariant mass (a) and angular separation (b) of the online $E_{\text{T}} > 20\,\text{GeV}$ jets geometrically matched to the highest invariant mass pair of offline jets, in simulated events. The distribution for the VBF H$\rightarrow \tau\tau$ signal is presented together with those of the Drell-Yan $Z/\gamma^* \rightarrow \tau\tau$ and gluon fusion Higgs processes. The simulated events are produced with the MadGraph5_aMC@NLO generator at LO precision [47]. The jets produced in the VBF process typically have large η separation and larger invariant mass [93].



(a) (b)

Figure 3.3 – Transverse energy of the online jet geometrically matched to the leading (a) and subleading jet (b) in the highest invariant mass offline pair, in simulated events. The distribution for the VBF H $\rightarrow \tau\tau$ signal is presented together with those of the Drell-Yan $Z/\gamma^* \rightarrow \tau\tau$ and gluon fusion Higgs processes. The simulated events are produced with the MadGraph5_aMC@NLO generator at LO precision [47]. The jets produced in the VBF process typically have larger p_{T} [93].

However, the complexity of the operations that can be performed by the L1 algorithm is also limited by computing power constraints. The L1 trigger is a hardware-based system with a fixed latency of 3.8 μs: the decisions performed by the system must be fast to fit in this time. As a matter of fact, the data transfer from the detector to the L1 system and vice versa is not negligible, and only $\sim 1\,\mu$s is actually available for the algorithms. Thanks to the recent architecture upgrade (see Ch. 2), the CMS trigger system outstanding capabilities allow complex variables among L1 objects, such as m_{jj} [94], to be computed. This is the main feature exploited by the L1 VBF seed. In the highly relativistic regime, the invariant mass of each pair of jets with transverse energies E_{T1} and E_{T2}, pseudorapidities η_1 and η_2 and azimuthal angles ϕ_1 and ϕ_2 is can be approximated to

$$m_{\mathrm{jj}}^2 = 2 \cdot E_{\mathrm{T1}} E_{\mathrm{T2}} [\cosh(\eta_1 - \eta_2) - \cos(\phi_1 - \phi_2)]. \tag{3.1}$$

Under this form, m_{jj}^2 can easily be computed in the FPGAs as the trigonometry can be implemented in Look Up Tables (LUT). The FPGAs are not suited for arithmetic computations that go beyond additions and multiplications; more complex operations can be implemented at the expense of a large resource consumption. In the present case, it is sufficient to store the cosh and cos response in LUTs. Besides, there is no need to extract the square root of m_{jj}^2. The validity of the m_{jj} computation through LUTs is proven indirectly later, in Fig. 3.9.

In Fig. 3.1, the E_{T} distributions of the L1 jets (or L1 taus) in VBF H $\rightarrow \tau\tau$ simulated events are shown: the yellow curve represents all the L1 jets in the event, that have mostly low E_{T}; the blue and cyan curves correspond to L1 jets geometrically matched with $\Delta R(\mathrm{L1jet}, \mathrm{VBF\ quark}) < 0.3$ to the leading and trailing VBF quarks produced in the simulation at the generator level, i.e. before undergoing the detector response simulation step; the red curve represents the E_{T} of the leading L1 jet in the event.

The VBF jets typically have moderate E_{T}: although their E_{T} is generally larger than that

of additional jets in other processes (Fig. 3.3), they are not the most energetic objects in VBF events (Fig. 3.1). At the LHC, the number of events containing moderate E_{T} L1 jets is extremely high, hence keeping the jet E_{T} thresholds as low as the typical VBF jet transverse energy leads to an unreasonable trigger rate, even with a requirement on the jet-jet invariant mass or the angular separation. Therefore, an additional handle is needed. For instance, a tight threshold can be applied to the leading jet E_{T}: a threshold around 90 GeV can serve the purpose of discriminating against the main background, as shown in Fig. 3.3a, and at the same time it can help to reduce the rate.

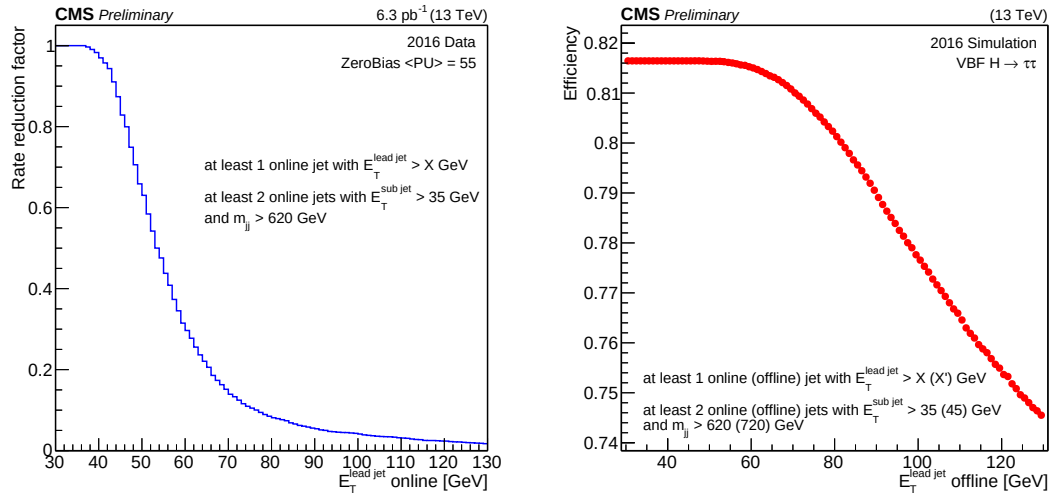
On the other hand, a hard cut on the E_{T} of the leading jet in the VBF pair would not produce the anticipated and desired result. Indeed, the object identification is ambiguous at L1: as they are all similar purely calorimetric objects, all L1 taus are also included in the L1 jets collection (while not all the L1 jets are included in the L1 taus collection). Therefore, large E_{T} L1 jets can be produced by tau leptons. In this sense, a truly general VBF algorithm cannot be achieved at L1: as a matter of fact, a fraction of the events selected by this jet-based algorithm will actually feature tau leptons producing L1 jets. A requirement of large L1 m_{jj} is enough to mitigate this effect, as there is no reason for a pair made by a tau and a jet to have large invariant mass. Besides, as long as the target is to collect VBF $\mathrm{H} \to \tau\tau$ events, the accidental presence of tau leptons can impact the performance only positively. An additional algorithm implementing a L1 objects disambiguation is described in Sec. 3.4.

Given the presence of tau leptons in the L1 jets collection, it is clear that the jet pair E_{T} and m_{jj} requirements need to be disentangled: a tight jet E_{T} requirement would drive the selection of the L1 jets towards those that are actually produced by taus, missing the VBF pair with moderate E_{T}; the high m_{jj} threshold, then, would likely cause the event to be discarded, as the selected pair actually contains a tau lepton, and the selection efficiency would be spoiled.

In conclusion, an event should fulfil two independent requirements in order to be selected: there must be at least two jets with moderate transverse energy $E_{\mathrm{T}} > Y$ and large invariant mass $m_{\mathrm{jj}} > Z$, so that the acceptance on VBF events is preserved; in addition, in order to ensure a large rate reduction, there must be at least one jet with large transverse energy $E_{\mathrm{T}} > X$ ($X > Y$). The large E_{T} L1 jet does not necessarily belong to the highest m_{jj} pair: it can also happen to be produced by a tau lepton. However, the algorithm is based on the sole jet topology and no additional requirements are made on the objects coming from the Higgs boson decay, even though tau leptons reconstructed as L1 jets can contribute to the trigger efficiency.

As illustrated in Fig. 3.4, where the rate and the efficiency of the VBF trigger are shown as a function of the leading jet E_{T} threshold, a compromise is needed: the highest the cut is placed, the highest the rate reduction; that said, the chosen working point should preserve a high efficiency at the same time. Nonetheless, the rate drops faster than the efficiency: for example, a 90 GeV threshold provides about 95% rate reduction, while only 5% efficiency is lost. In reality, the optimisation is multi-dimensional and several combinations of thresholds are tested. Two figures of merit are used: the event rate and the acceptance gain on the VBF $\mathrm{H} \to \tau\tau$ signal.

The optimisation of the L1 algorithms is performed using unbiased data, available in the so-called "ZeroBias" data sets. The ZeroBias events are collected using a trigger that selects valid bunch crossings using the coincidence of signals in the two beam position monitors installed along the beamline at the opposite sides of the CMS, i.e. using the timing information rather than the event content.



(a) (b)

Figure 3.4 – Rate reduction factor (a) for the L1 VBF seed as a function of the X threshold with fixed $Y = 35\,\text{GeV}$ and $Z = 620\,\text{GeV}$, computed using 2016 data from a ZeroBias high pileup data set and normalised to the rate obtained with threshold $X = 30\,\text{GeV}$. Its efficiency, computed on a 2016 VBF $H \to \tau\tau$ simulation, is shown in (b) as a function of the offline $p_T > X + 10\,\text{GeV}$ threshold with fixed offline thresholds $Y = 45\,\text{GeV}$ and $Z = 720\,\text{GeV}$ [95].

The actual event rate of the VBF trigger is computed with $6.3\,\text{pb}^{-1}$ of ZeroBias high-pileup data events collected in 2016, commonly used in the trigger community for the commissioning of the trigger menu for the 2017 run. It is reminded that the average pileup per bunch crossing is related to the luminosity L and the number n_b of colliding bunches in the CMS interaction point as

$$< \text{PU} >= \frac{L \cdot \sigma_{pp}^{inel}}{n_b \cdot f_{\text{LHC}}} \tag{3.2}$$

where σ_{pp}^{inel} is the inelastic pp cross section and $f_{\text{LHC}} = 11245.6\,\text{Hz}$ is the bunch revolution rate at the LHC. The measured inelastic pp cross section at 13 TeV is 69 mb [55]; however, the value commonly used in the rate estimations is 80 mb, for convention shared with the ATLAS experiment and for consistency with a previous estimation from simulation tools.

The run configuration of the data under study has $n_b = 99$ and initial luminosity $L = 7.6 \times 10^{32}\,\text{cm}^{-2}\,\text{s}^{-1}$, leading to $< \text{PU} >= 55$. As the luminosity decreases along the collision run, the $< \text{PU} >$ is also reduced, hence only the events collected at the beginning of the considered data taking, before the $< \text{PU} >$ gets to the 98% of its initial value, are used for the rate measurement.

Using L1 objects to emulate the trigger algorithm response, the VBF seed total rate per bunch is

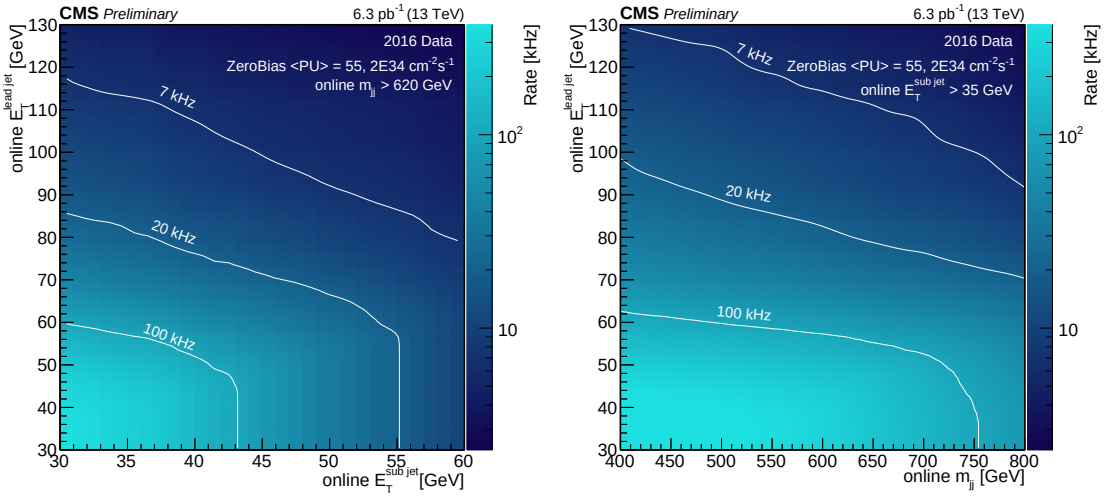
$$rate = x_{\text{VBF}} \cdot f_{\text{LHC}} \tag{3.3}$$

where $x_{\text{VBF}} = x_{\text{VBF}}(X, Y, Z)$ is the fraction of events firing the VBF trigger. The result is extrapolated to the 2017 data taking conditions by scaling

$$rate_{2017} = rate \cdot n_{b,2017}$$

where $n_{b,2017}$ is the number of bunches needed to have $< \mathrm{PU} >= 55$ with instantaneous luminosity $L_{2017} = 2 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, which is the typical value that was foreseen in 2016 for 2017 LHC operations. From the Eq. 3.2, $n_{b,2017} = 2592$.

The resulting extrapolated rate is shown in Fig. 3.5 as a function of the X-Y and X-Z thresholds. The study of the total rate provides a good understanding of the rate reduction handles but, in practice, it is firstly the pure rate that matters for a L1 seed to be added in the menu. The pure rate is the additional rate with respect to the other existing algorithms included in the full L1 selection; its measurement of the pure rate is not trivial: it requires the knowledge of the selection algorithm of all the seeds included in the L1 menu, as well as the most updated tuning proposed for the optimisation and maintenance of each set of seeds. For this reason, its computation is a task performed within the CMS trigger studies group and its measurement served as a feedback to the thresholds proposed throughout the optimisation procedure. For example, the choice $X = 110\,\mathrm{GeV}$, $Y = 35\,\mathrm{GeV}$ and $Z = 620\,\mathrm{GeV}$ was, at the time of this study, giving a 7.7 kHz total rate and 2.7 kHz pure rate.



(a) (b)

Figure 3.5 – Rate for the L1 VBF seed, selecting events where at least one L1 jet with $p_{\mathrm{T}} > X$ is found, as well as a pair of L1 jets with $p_{\mathrm{T}} > Y$ and $m_{\mathrm{jj}} > Z$. The rate is shown as a function of the X and Y thresholds with fixed $Z = 620\,\mathrm{GeV}$ (a) and as a function of the X and Z thresholds with fixed $Y = 35\,\mathrm{GeV}$ (b), computed using a 2016 ZeroBias high-pileup data set and extrapolating to the 2017 data taking conditions [95].

3.1.2 Evaluation of expected performance

While the usual trigger strategies exploit the Higgs boson decay mode, in the case of the VBF seed that I designed the selection targets the production mode, although also tau leptons incidentally contribute to the efficiency. The VBF trigger is meant to be complementary to the classic triggers, allowing the phase-space to be extended and the Higgs boson signal event yield to be increased. The $\mathrm{H} \to \tau\tau$ analysis, for the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel, relies only on the dedicated hadronic tau trigger, selecting a pair of isolated hadronic taus with $|\eta| < 2.1$. The p_{T} threshold of the taus selected by the L1 double-τ_{h} trigger is tuned to maintain its total event rate below 14 kHz and it was set to 32 GeV for the 2017 data taking conditions. In order to evaluate the acceptance gain on the signal with respect to the existing $\mathrm{H} \to \tau\tau$ analysis, suitable offline selections are defined for the events triggered by the VBF seed (with thresholds $X = 110\,\mathrm{GeV}$, $Y = 35\,\mathrm{GeV}$

and $Z = 620\,\text{GeV}$) and for those triggered by the double-τ_h trigger, choosing thresholds as low as possible, while still on the efficiency plateau, as clarified in Sec. 3.2. The taus reconstructed offline are required to pass identification criteria, as well as jets, and to be well separated from electrons and muons (cf. Sec. 4.3). Since the tau leptons are also reconstructed as jets, a jet-tau disambiguation has to be applied and the jets reconstructed offline fulfilling $\Delta R(jet, \tau_\text{h}) < 0.3$ with the leading and trailing selected tau leptons are excluded from the invariant mass computation. In the following, the selected offline jets are reconstructed with $|\eta| < 5$.

For the expected performance estimation, the goal is to reproduce realistic analysis-like scenarios, starting from the VBF category of the $\text{H} \to \tau\tau$ analysis [96], and compare the offline event yield of the sole double-τ_h trigger case with that of the joint use of the VBF and double-τ_h trigger. On one hand, the VBF offline selection has a high m_jj threshold ($m_\text{jj} > 700\,\text{GeV}$), but the lower m_jj is covered by the double-τ_h selection ($m_\text{jj} > 400\,\text{GeV}$). On the other hand, as the VBF trigger does not impose selection criteria on the Higgs decay products, it allows the p_T offline threshold on the tau leptons ($p_\text{T}^{sub\,\tau} > 20\,\text{GeV}$) to be lower than that of the double-τ_h selection ($p_\text{T}^{sub\,\tau} > 35\,\text{GeV}$), as represented in Fig. 3.6. The gain is evaluated from the number of events passing the trigger selections

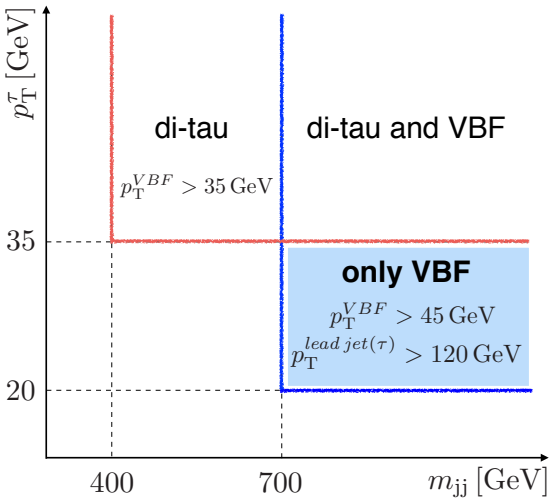


Figure 3.6 – Definition of the categories for the computation of the event yield gain from the use of the VBF algorithm to complement the double-τ_h trigger. The minimum transverse momentum of the VBF jet candidates is p_T^{VBF}, while $p_\text{T}^{lead\,jet(\tau)}$ is referred to the leading jet or τ in the event. The light blue region is a new portion of phase-space that cannot be covered by a selection relying on the sole double-τ_h trigger.

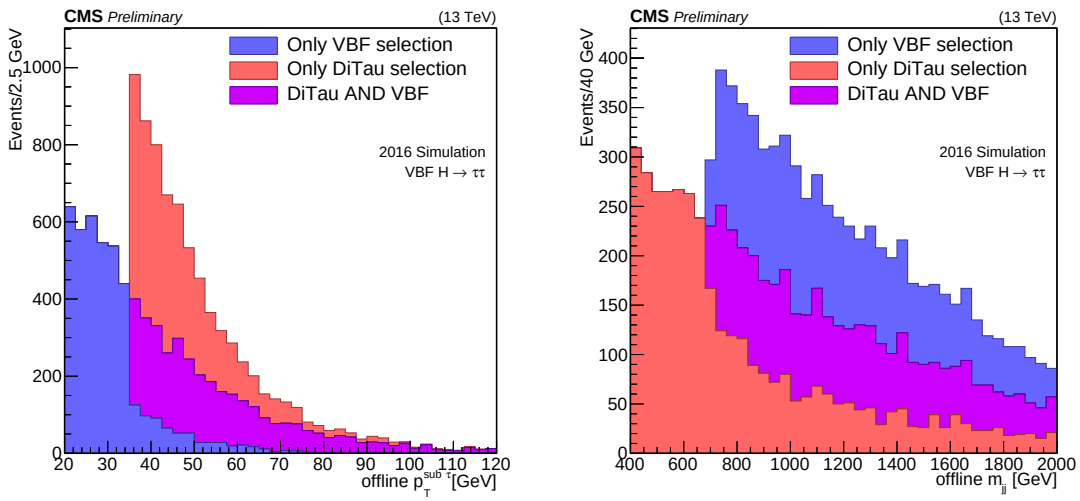
and the corresponding offline selections: the events in each of the categories "Only VBF selection" and "Only di-tau selection" are events that fired only one of the triggers and pass only the offline selection corresponding to that trigger; the events in the category "di-tau and VBF" fire both triggers and pass both offline selections. The distribution of the VBF $\text{H} \to \tau\tau$ signal in the event categories is shown in Fig. 3.7 both as a function of m_jj and of $p_\text{T}^{sub\,\tau}$. The red and magenta histograms correspond to regions that are already covered by the di-τ_h trigger. One can clearly notice the structures created by the VBF seed thresholds with the corresponding acceptance increase, appearing in blue in the plot.

The event yield gain obtained by using both triggers rather than the sole di-τ_h seed is

$$\frac{N_\text{Only VBF}}{N_\text{di-tau}} = 58\%$$

coming from the coverage of new portions of the phase-space by the "Only VBF selection" category. The low $p_{\mathrm{T}}^{sub\,\tau}$ region, indeed, is out of reach for the double-τ_{h} trigger, while the VBF trigger does not have an explicit tau selection. The performance estimated with different sets of thresholds are compared in Tab. 3.1. It is clear that the tighter is the VBF online selection, the smaller is the additional phase-space covered. Such trade-off is illustrated in Fig. 3.8 and detailed in Sec. 3.2, where the rate reduction on the vertical axis and acceptance gain on the horizontal axis, estimated using 2017 high-pileup data events and a 2017 VBFH $\rightarrow \tau\tau$ simulation, are represented in as a function of the highest jet E_{T} indicated as colour scale markers.

With the same strategy, the event yield gain is also computed on the HH $\rightarrow$ bb$\tau\tau$ signal. The only modification to the event selection, coherently with the criteria in the HH $\rightarrow$ bb$\tau\tau$ analysis, is the requirement of two offline jets with $p_{\mathrm{T}} > 20\,\mathrm{GeV}$, $|\eta| < 2.4$ and with minimal distance $\Delta R(jet, \tau_{\mathrm{h}}) > 0.5$ from the selected taus. Although the HH $\rightarrow$ bb$\tau\tau$ analysis is based on the number of jets passing the b tag selection (see Sec. 4.4.4), this estimation is inclusive and at least one b tagged jet is required. The selection of the b jet candidates is applied both to the events triggered by the VBF seed and for those triggered by the di-τ_{h} seed. The p_{T} spectrum of the tau leptons in the HH $\rightarrow$ bb$\tau\tau$ is softer than that of the tau leptons in H $\rightarrow \tau\tau$. Therefore, in this case there is even more to gain in reducing the tau p_{T} thresholds: the additional event yield is $N_{\mathrm{Only\,VBF}}/N_{\mathrm{di\text{-}tau}} = 69\%$.



(a) (b)

Figure 3.7 – Additional events from the L1 VBF seed (with at least one jet with $E_{\mathrm{T}} > 110\,\mathrm{GeV}$ and at least two jets with $E_{\mathrm{T}} > 35\,\mathrm{GeV}$ and $m_{\mathrm{jj}} > 620\,\mathrm{GeV}$) with respect to the double-τ_{h} seed (at least 2 isolated taus with $E_{\mathrm{T}} > 32\,\mathrm{GeV}$ and $|\eta| < 2.1$) as a function of the p_{T} of the offline subleading tau (a) and of the invariant mass of the offline jets with highest m_{jj} (b). The events in each of the categories "Only VBF selection" and "Only di-tau selection" are events that fired only one of the triggers and pass only the offline selection corresponding to that trigger. The events in the category "di-tau AND VBF" fire both triggers and pass both offline selections [95].

3.2 L1 VBF trigger online performance

In the following, the total rate is expressed per bunch, rather than being scaled to the actual number of bunches through the Eq. 3.3, as this was the standard in the trigger community for easier comparison between different collision runs. In order to achieve a

Table 3.1 – Total and pure rate for some L1 VBF seeds, computed using 2016 data from a ZeroBias high pileup data set and extrapolated to the 2017 data taking conditions, and the corresponding acceptance gain on a VBFH $\rightarrow \tau\tau$ signal simulation using the VBF trigger and the double-τ_h trigger rather than the double-τ_h trigger only.

Thresholds			Rate		
Leading jet E_T [GeV]	Jet pair E_T [GeV]	Highest m_{jj} [GeV]	Total rate [kHz]	Pure rate [kHz]	Acceptance gain [%]
110	35	620	7.7	2.7	58
110	40	620	6.6	2.2	55
115	35	620	6.0	1.9	53
115	40	620	5.4	1.5	50

high luminosity, the LHC can operate with different "bunch schemes", i.e. with different configurations of number of proton bunches, of spacing between consecutive bunches and of number of bunches packed together in a "train". The nominal bunch scheme chosen for the Run 2 operations, with 2544 bunches per colliding beam in the CMS interaction point, packed in trains of 48 (scheme "48b"), could not be maintained throughout the whole the data-taking period due to technical limitations. After the technical stop between the 2016 and 2017 operations, "electron cloud" effects were observed [97]: when some residual gas is trapped in the LHC beam pipes, the electrons generated by its ionisation get accelerated in the vicinity of the beam, positively charged, and hit the walls of the pipes with enough energy to produce secondary electrons and initiate an avalanche; the passage of several proton bunches leads to the formation of electron clouds that are dense enough to degrade the vacuum and make the beam unstable, reducing its durability. This effect is one of the main challenges for the beam operations and many actions are programmed during the technical stops to prevent the formation of such clouds; the most effective is the "scrubbing" technique, which consists in injecting high-intensity beams to release as much as possible of the gas trapped in the walls of the pipe. In spite of these measures, the beam quality was degraded in 2017. Starting from September 2017, the LHC moved to a special configuration with up to 1909 higher density bunches, packed in 8 filled batches followed by 4 empty slots (scheme "8b4e"), to mitigate the electron clouds formation; for constant instantaneous luminosity, a smaller number of proton bunches compared to nominal conditions implies a larger pileup (cf. Eq. 3.2). During this period, the instantaneous luminosity rarely exceeded $1.55 \cdot 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, while the nominal conditions were restored in 2018.

The L1 VBF trigger was included in the CMS online selection in summer 2017 with thresholds $X = 115\,\mathrm{GeV}$, $Y = 40\,\mathrm{GeV}$ and $Z = 620\,\mathrm{GeV}$, together with other luminosity settings with larger X and Y for harsh data-taking conditions. The event rate computed at μGT through the m_{jj} approximation from Eq. 3.1 is in excellent agreement with the full computation, as shown in Fig. 3.9. Nonetheless, it turned out to be much higher than expected from the 2016 extrapolations in slightly different conditions. For instance, in collision runs with higher pileup ($< \mathrm{PU} >= 60$) and same instantaneous luminosity $L = 2 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, the chosen VBF seed gives up to 10.7 kHz of total rate, while it was estimated to give 5.4 kHz using 2016 data with $< \mathrm{PU} >= 55$ (cf. Tab. 3.1). The origin of this discrepancy and the solutions are discussed in Sec. 3.3.

In Fig. 3.8, the trade-off between the achieved rate reduction and the expected signal event yield gain is represented as a function of the threshold X on the leading L1 jet E_T, while the values of the other thresholds are fixed to $Y = 40\,\mathrm{GeV}$ and $Z = 620\,\mathrm{GeV}$. Using this configuration, the VBF trigger total rate per bunch from the VBF trigger

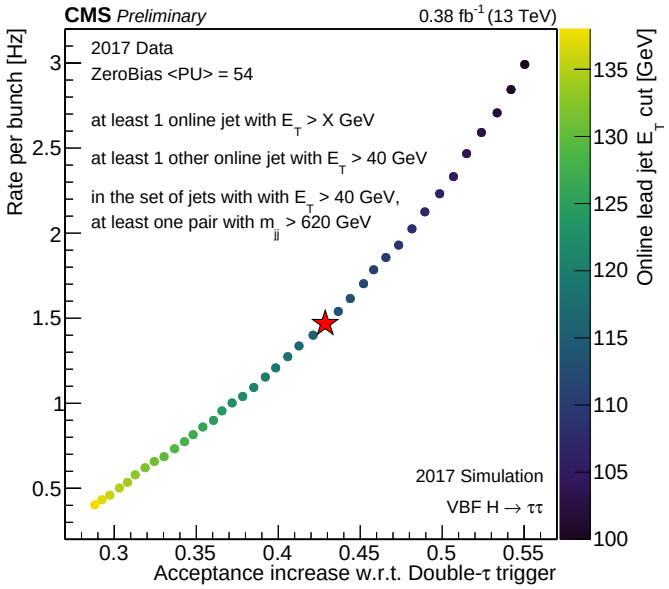


Figure 3.8 – The L1 VBF algorithm is defined with $Y = 40$ GeV and $Z = 620$ GeV, while X (the threshold on the leading L1 jet E_{T}) is varied. Using this configuration, the total rate of the VBF trigger is computed in a ZeroBias 2017 sample with $<$ PU $>= 54$, as function of the X cut. In the plot, the x-axis values represent the net increase of VBFH $\rightarrow \tau\tau$ signal events, computed as the number of events passing only the VBF trigger and the corresponding offline selection over the number of events passing the double-τ_{h} trigger. The red star indicates the VBF trigger configuration that was used in the data taking throughout 2017: $X = 115\,\mathrm{GeV}$, $Y = 40\,\mathrm{GeV}$, $Z = 620\,\mathrm{GeV}$. In this run configuration, with $n_b = 1909$, 1.5 Hz per bunch crossing corresponds to 2.9 kHz [93].

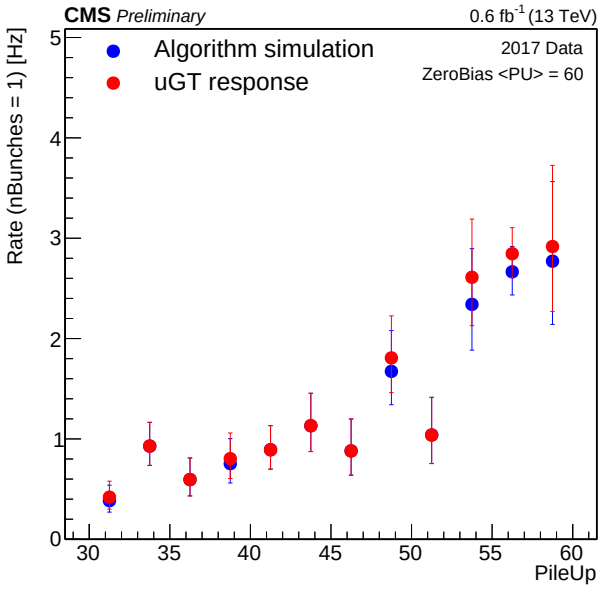


Figure 3.9 – Total rate per bunch of the VBF trigger with at least one jet with $E_{\mathrm{T}} > 115\,\mathrm{GeV}$ and at least two jets with $E_{\mathrm{T}} > 40\,\mathrm{GeV}$ and $m_{\mathrm{jj}} > 620\,\mathrm{GeV}$ computed in a 2017 ZeroBias high pileup sample ($<$ PU $>= 60$, $n_b = 1866$) as a function of the pileup. The rate is estimated using L1 objects to emulate the trigger algorithm and compared to that obtained using the trigger decision.

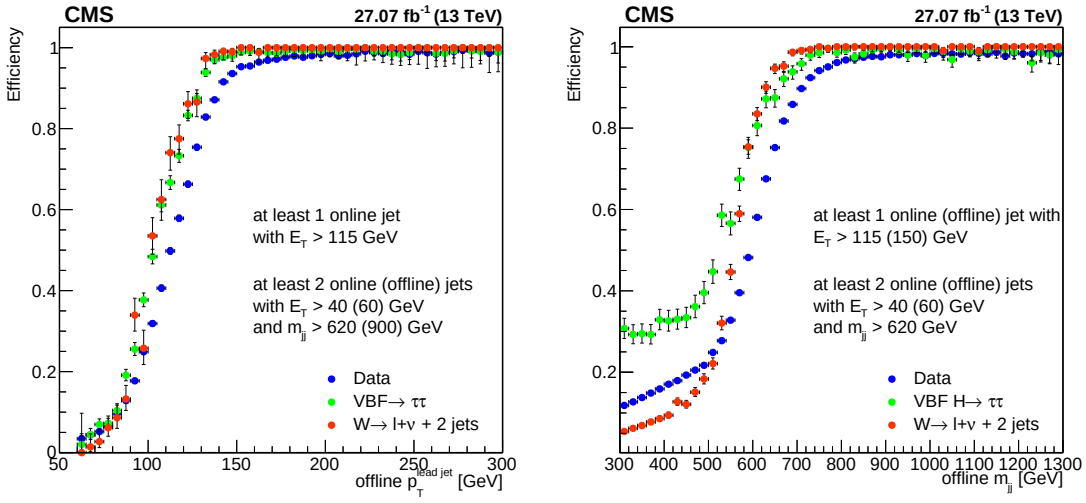
is computed in a ZeroBias 2017 sample with $<$ PU $>= 54$, while the net increase of VBFH $\rightarrow \tau\tau$ signal events, computed as $N_{\text{Only VBF}}/N_{\text{di-tau}}$, is estimated in a 2017 VBF H $\rightarrow \tau\tau$ simulation. As in the studies performed on 2016 simulation, $N_{\text{Only VBF}}$ is the number of events passing only the VBF trigger and the corresponding offline selection, while $N_{\text{di-tau}}$ is the number of events passing the double-τ_{h} trigger with $p_{\text{T}} > 32\,\text{GeV}$ cut on each L1 tau, in addition to an isolation requirement, and the corresponding offline selection. The offline event selection for the VBF (double-τ_{h}) trigger requires at least one jet with $p_{\text{T}} > X + 10(35)\,\text{GeV}$, at least one other jet with $p_{\text{T}} > 45(35)\,\text{GeV}$ and, in the set of jets with $p_{\text{T}} > 45(35)\,\text{GeV}$, at least one pair with $m_{\text{jj}} > 700(400)\,\text{GeV}$. At least two well-identified offline hadronic taus with $p_{\text{T}} > 20(35)\,\text{GeV}$ are required as well. By raising the leading jet E_{T} threshold, a better rate reduction is achieved, but the acceptance gain coming from the use of the VBF trigger is reduced. The VBF trigger configuration that was used in the data-taking throughout 2017 is highlighted: it provides 1.5 Hz total rate per bunch, corresponding to 2.9 kHz in this run configuration (1909 colliding bunches per beam) and 3.8 kHz in nominal 2017 conditions, allowing 44% additional VBFH $\rightarrow \tau\tau$ events to be collected.

The performance of a trigger is measured using the offline quantities corresponding to those exploited from the online algorithms as reliable discriminators for the event selection, since they are computed with higher precision through the whole reconstruction chain. The fraction of events passing both the offline and online requirements over those passing the offline requirement, computed as a function of the offline quantities, is a good estimator of the selection efficiency of a given trigger. The shape of such efficiency, in the simple 1D case, is that of a step function (at L1, the event is selected if passing a fixed threshold) convolved with a Gaussian: the sharpness of the efficiency rise depends on the online vs. offline resolution of the relevant quantities; if the online reconstruction were identical to the offline one, the efficiency would be a step function. The shape obtained through this measurement is commonly called a "turn-on" efficiency curve.

The efficiency of the VBF trigger used throughout the 2017 data taking is shown in Fig. 3.10 as a function of the leading jet p_{T} and the highest jet-jet invariant mass. It is computed as the fraction of offline events passing the VBF L1 trigger selection in 2017 data (27.07 fb, corresponding to the period covered by the VBF trigger implementation) in a sample of unbiased events triggered by a single muon selection, as well as in 2017 VBF H $\rightarrow \tau\tau$ and W$\rightarrow \ell\nu$ + 2 jets simulations. In order to have a good description of the 1D efficiencies as a function of each variable, tight offline thresholds are set alternately on the two other significant variables: the offline selection for the computation of the efficiency as a function of the leading jet offline p_{T} requires at least two jets with $p_{\text{T}} > 60\,\text{GeV}$ and with $m_{\text{jj}} > 900\,\text{GeV}$; in the computation of the efficiency as a function of the largest invariant mass, the jets in the highest m_{jj} pair must have $p_{\text{T}} > 60\,\text{GeV}$ and at least one jet with $p_{\text{T}} > 115\,\text{GeV}$ is required. All the selected offline jets must fulfil tight identification criteria and be well discriminated against e/μ leptons and taus ($\Delta R(\text{jet}, \tau_{\text{h}}) > 0.5$); besides, no online nor offline leptons are required. The efficiencies for the simulated events are weighted using the profile of the number of vertices in data and simulation, so that the profile of PU, estimated as the number of reconstructed vertices, matches between data and simulation.

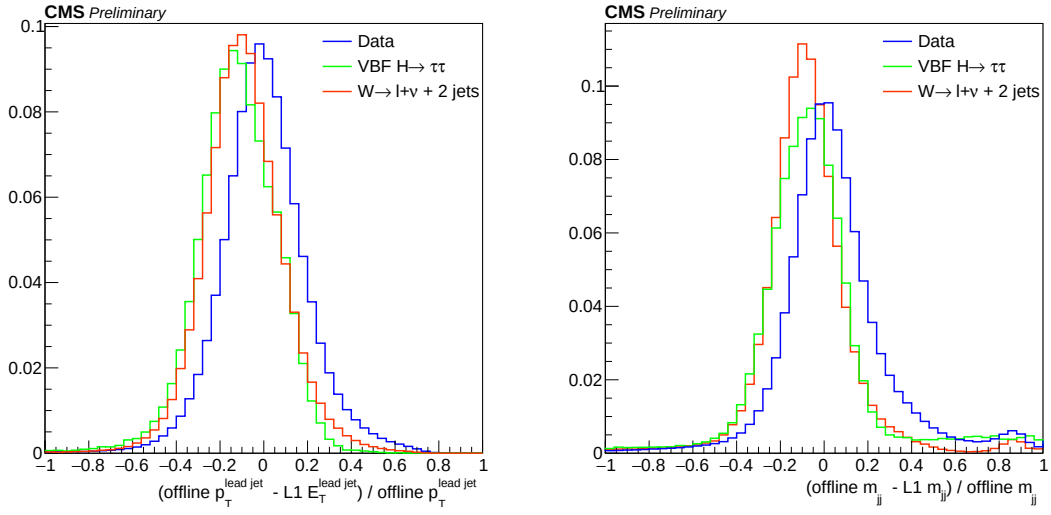
A flat 100% efficiency is achieved at the plateau and the turn-on as a function of the leading jet p_{T} (Fig. 3.10a) is remarkably steep. However, two anomalies emerge from the measurement.

Firstly, the efficiency on the simulation samples appears shifted with respect to that on data. This inconsistency is understood and it is due to a different L1 jets energy scale,



(a) (b)

Figure 3.10 – The efficiency of the VBF L1 trigger, defined with $X = 115$ GeV, $Y = 40$ GeV and $Z = 620$ GeV, is computed for offline events as a function of the offline leading jet p_{T} (a) and as a function of the offline m_{jj} of the highest invariant mass jet pair (b). The efficiency is computed as the fraction of offline events passing the VBF L1 trigger selection in early 2017 data, in $W \to l\nu$ + 2 jets simulated events and in VBF H$\to \tau\tau$ simulated events.



(a) (b)

Figure 3.11 – The p_{T} resolution of the leading jet (a) and of the m_{jj} resolution (b), using L1 jets geometrically matched to those selected offline.

leading to a different resolution in data and simulations: it is not a flaw of the trigger algorithm, but rather the effect of early calibrations used for the simulation samples production. The p_{T} resolution of the leading jet and the m_{jj} resolution, using L1 jets geometrically matched to those selected offline, are shown in Fig. 3.11. A $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ is required for the L1 jets, while the offline jets must have $p_{\mathrm{T}} > 60\,\mathrm{GeV}$. The shift between each of the simulations p_{T} resolution curves and that of the data (Fig. 3.11a), computed as the difference of the mean values obtained through Gaussian fits, is -0.11 for VBF H $\to \tau\tau$ and -0.10 for W$\to \ell\nu$ + 2 jets. A satisfactory agreement is achieved by applying a relative correction, equivalent to the shift measured from the resolution

curves, to the L1 jets E_{T} in the computation of the efficiencies on the simulated events, as shown in Fig. 3.12.

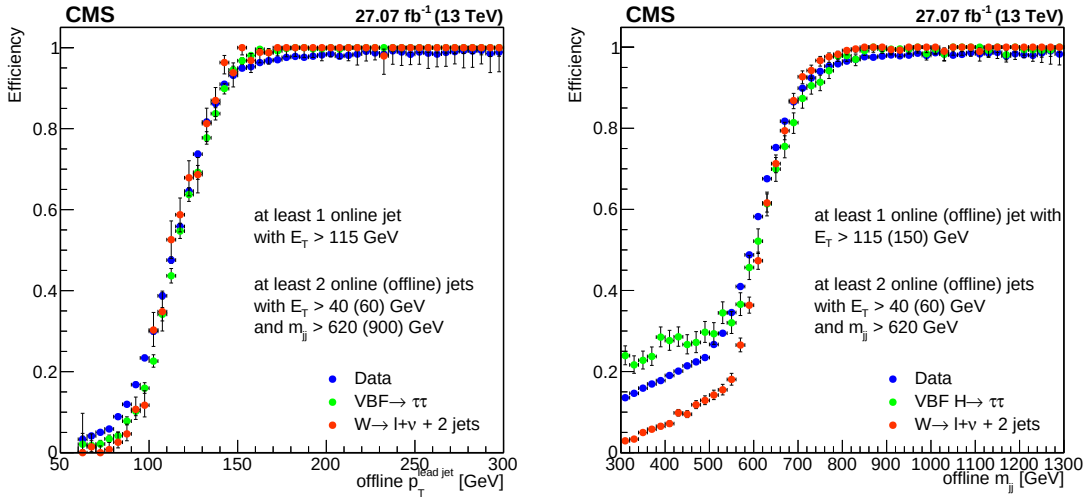


Figure 3.12 – An energy scale correction, computed from Fig. 3.11, is applied in the computation of the efficiency on the VBF H $\rightarrow \tau\tau$ and W$\rightarrow \ell\nu$ + 2 jets simulation samples already shown in Fig. 3.10.

Secondly, the $m_{\mathrm{jj}} < 500\,\mathrm{GeV}$ region (Fig. 3.10b) presents an efficiency plateau for the three curves. Given the handling of physics object in the L1 trigger system, this is an expected feature: the L1 jets can also be produced from tau leptons or electrons that contribute to the efficiency of the VBF trigger, while the more sophisticated offline reconstruction, reveals a lower invariant mass for the best VBF jet pair candidate, selected with tight jet identification requirements, in events that passed the high m_{jj} threshold required online. The nature of the considered processes justifies different trends: the two final state taus in the VBF H $\rightarrow \tau\tau$ simulation inevitably contaminate the L1 jets, causing an offset in the efficiency that is not negligible; the W$\rightarrow \ell\nu$ + 2 jets simulation has a cleaner L1 jets collection, and it has a smaller efficiency in the low m_{jj} region; finally, the data composition gives an efficiency closer to that of the W$\rightarrow \ell\nu$ + 2 jets simulation, with some additional contribution from Z$\rightarrow \ell\ell$ events. In any case, given the thresholds applied in the L1 VBF seed and offline, in the phase space used in the analysis, at large m_{jj} values, there is a good agreement between simulation and data.

In conclusion, the VBF seed is proven to have excellent performance. The m_{jj} computation at the firmware level is confirmed to agree with the simulation of the algorithm response and the narrow resolution between online and offline jets shapes the efficiency in sharp turn-ons.

3.3 Treatment of the problematic Trigger Tower 28

As shown in Fig. 3.13, representing the event rate of the VBF seed with $X = 115\,\mathrm{GeV}$, $Y = 40\,\mathrm{GeV}$ and $Z = 620\,\mathrm{GeV}$ computed by the online monitoring tool of CMS in a 2017 run, the VBF seed has a large, non-linear pileup dependence, and its rate explodes beyond $< \mathrm{PU} >= 55$. Similar effects, indeed, were observed in the majority of the seeds in the L1 menu selecting L1 jets, especially in the large-η region, and, in general, making use of calorimetric L1 objects. On one hand, it is not surprising that the forward regions of the detector, where the low-p_{T} pileup jets are abundant, are sensitive to the pileup increase. On the other hand, a few circumstances concurred to aggravate this problem.

The effort of several teams was invested in understanding the causes, summarised in the following, of the unexpected L1 trigger selection response.

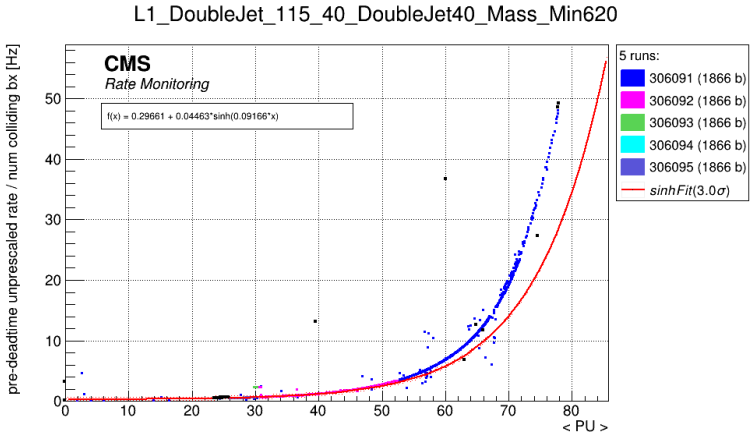


Figure 3.13 – Total rate per bunch of the VBF seed with $X = 115\,\text{GeV}$, $Y = 40\,\text{GeV}$ and $Z = 620\,\text{GeV}$, as a function of the pileup as displayed in the online rate monitoring tool during a 2017 collision run with $< \text{PU} >= 78$ and $L = 2 \cdot 10^{34}\,\text{cm}^{-2}\,\text{s}^{-1}$.

The constant radiation acting on the ECAL crystals, produced by the ordinary LHC operations, has the expected effect of reducing their transparency, affecting their response to the deposited energy. The changes in transparency are measured all along the data taking through a dedicated laser monitoring system, injecting light with wave length close to the peak of the scintillation light spectrum of the lead tungstate ($\lambda = 447$ nm) in every crystal making cycles of about 40 minutes; the corrections for the response variation are available to be applied online within 48 hours. The history of relative response to laser light over the Run 1 and Run 2 is shown in Fig. 3.14. A partial recovery is observed in absence or radiation; however, the damage keeps degenerating over time. Although the extent of the transparency loss during 2017 was not exceptional, it was large enough to start being critical. The crystals in the forward regions of the detectors are the most exposed to the radiation and the change in their performance is more acute: the laser response in the highest η bin in Fig. 3.14 decreased of about factor 3 in 2017. Since the degradation of this area evolves faster, the laser corrections are 10 to 20 times larger at large η than in the rest of the detector. The noise produced by the ADCs that read the response of the crystals is unrelated to the transparency loss, but it is also amplified by the laser corrections. Therefore, as the corrections become larger, the noise contribution becomes more and more important at large η.

The signals in the ECAL and in the HCAL are sampled every 25 ns, which is also the separation between two colliding bunches at the LHC. This mechanism can lead to an overlap in the readout of several pulses and to a significant contribution to the event rate from a neighbour bunch crossing. To mitigate this effect, called "out-of-time" pileup (OOTPU), a weighting strategy is implemented in the ECAL sampling: two sets of weights, derived separately for endcaps and barrel from data collected in beam tests, are applied to shape the reconstructed pulse; the first weight, negative, subtracts the average early OOTPU contributions to the pulse [99]. These weights, however, are no longer optimal: since they were computed, the typical pulse shape has changed due to the radiation damage of the crystals; therefore, their application fails subtracting the OOTPU contribution and leads to an amplitude bias especially pronounced at large $|\eta|$ and increasing over time [100]. The impact of the amplitude bias is also larger at the beginning of the trains, within the first about ten bunches: since there are no bunches immediately before, the compensation computed in stable beam test conditions is off.

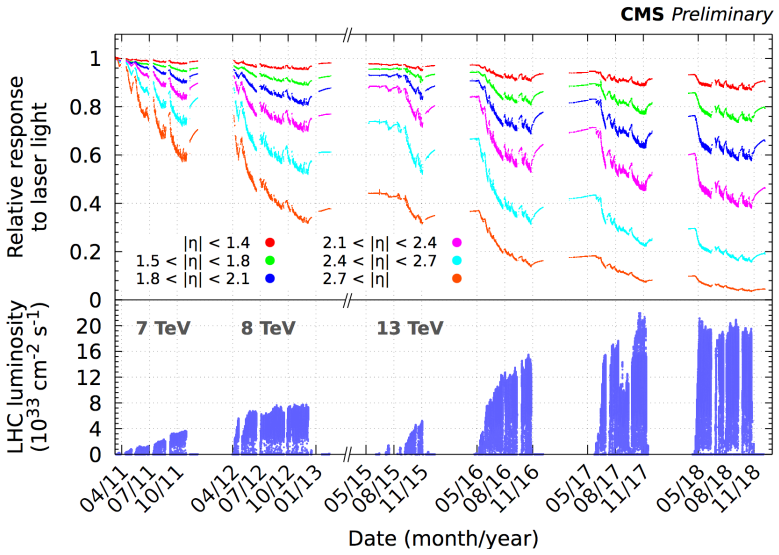


Figure 3.14 – Relative response to laser light (440 nm in 2011 and 447 nm from 2012 onwards) injected in the ECAL crystals, measured by the ECAL laser monitoring system, averaged over all crystals in bins of η during Run 1 and Run 2. The response change observed in the ECAL channels is up to 96% in the region closest to the beam pipe. The recovery of the crystal response during the periods without collisions is visible. These measurements, performed every 40 minutes, are used to correct the physics data. The bottom pad shows the instantaneous LHC luminosity delivered during this time period [98].

In typical LHC bunch schemes, structured in tens of consecutive bunches, this effect is averaged out. In spite of the online corrections, the amplitude bias became more important with the 8b4e bunch scheme, introduced in 2017 to face technical problems of the LHC (cf. Sec. 3.2) and structured in shorter and more frequent trains.

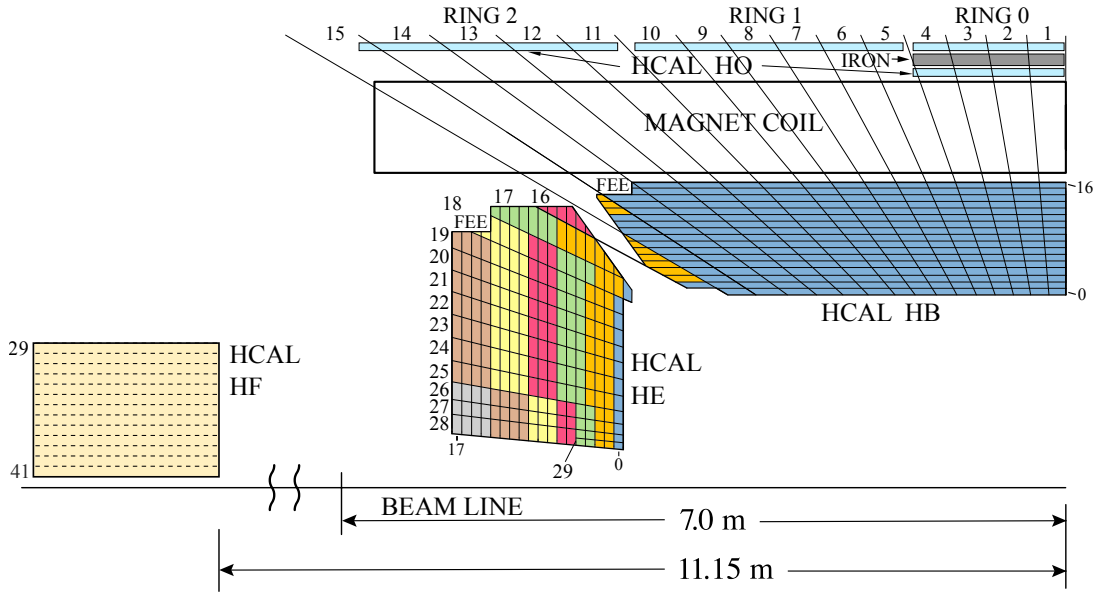


Figure 3.15 – Detail of the HCAL subdetectors in the r-z plane, displaying the trigger tower segmentation. The shape of TT28 is irregular, while the logic TT29 physically covers two different towers in HE and HF.

A section of the HCAL detector, divided in trigger towers (TT), is shown in Fig. 3.15. The TT28 and TT29, covering the regions in the range $\eta \sim 2.5 - 3.0$, have a different

structure than the rest of the trigger design, which cannot be fully properly taken into account in the firmware and enhances the trigger event rate. In particular, the TT28 is about 2.5 times bigger than TT27 and TT26, and 4 times bigger than the majority of the other trigger towers, while the region that is logically covered by TT29 is actually displaced over two different subdetectors.

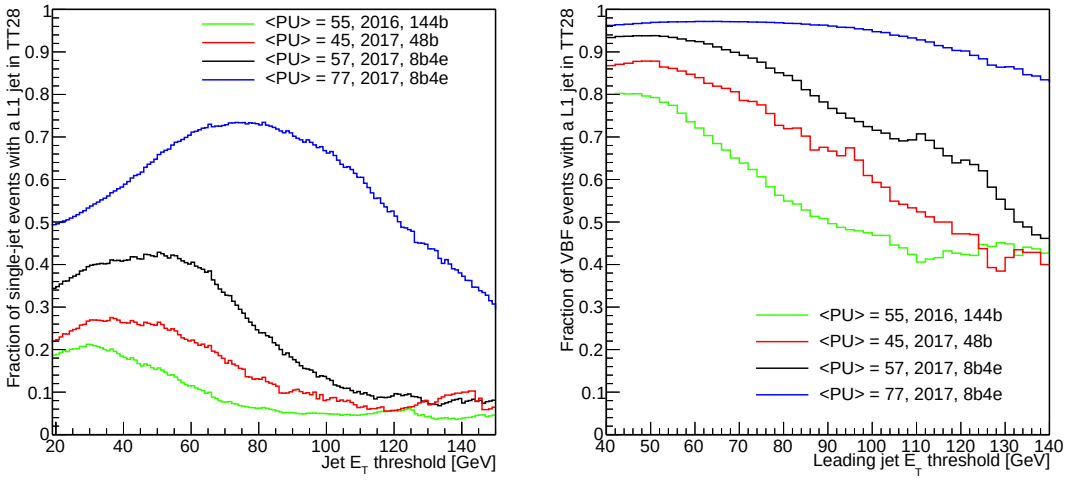
These trigger towers, already larger by design, are those that suffer the most from the effects described, being placed at large η. As the VBF process is characterised by jets in the forward parts of the detectors, the rate of the VBF trigger is inevitably related to the event rate in these trigger towers. The TT28 was found to be the most sensitive to the out-of-time pileup increase.

The impact of the TT28 in the event rate is illustrated qualitatively in Fig. 3.16, where ZeroBias data sets corresponding to different collision runs are used: a 2016 run with bunch scheme "144b" (144 bunches packed in each train) and $<$ PU $>= 55$; three 2017 runs with "48b" and $<$ PU $>= 45$, "8b4e" and $<$ PU $>= 57$, and "8b4e" and $<$ PU $>= 77$. In Fig. 3.16a the fraction events with the leading jet seed in TT28 over all the events firing a single jet selection is shown as a function of the leading jet E_{T} threshold. In general, the fraction of jets detected in TT28 is sizeable at low E_{T}, reaching a maximum in the moderate E_{T} region; as the threshold is further raised, it becomes less and less likely that the leading jet is produced at large η. The green curve, corresponding to 2016 data, shows the smallest contribution in TT28, with a fraction that is never higher than 20% and decreases to a few percents with tight E_{T} threshold. This fraction is significantly higher in 2017. The red curve corresponds to data taken in conditions similar to those of the green curve: the impact of the laser corrections updates from 2016 to 2017 is probably the dominant effect leading to the higher contribution of jets in TT28. The black curve corresponds to data with higher pileup, collected after the introduction of the "8b4e" scheme. Together with the laser corrections, these changes contribute to make the fraction of jets in TT28 higher. The blue curve represents a collision run with extremely high pileup, which significantly contributes to the impact of the TT28 in the event rate. In the moderate E_{T} region, from 40 to 80 GeV, a dramatic evolution is observed: the fraction of events triggered by jets in TT28 changes substantially in different conditions and its maximum is also shifted towards tighter E_{T} thresholds. In Fig. 3.16b, the fraction of events where one of the three jets in the VBF algorithm is in TT28 over the events that fired the VBF seed with $Y = 40\,\mathrm{GeV}$ and $Z = 620\,\mathrm{GeV}$, computed using the same data sets, is shown as a function of the threshold on the leading jet E_{T}. In 2017 conditions, almost 100% of the events that are fired by the VBF seed have at least one jet in TT28 in the low p_{T} region, and this fraction decreases slowly with tighter E_{T} thresholds.

The pileup became progressively higher along the year of data taking. In 2016 and partially 2017 data, the instantaneous luminosity, thus the pileup, decreases along each collision run. Starting from 2017, a levelling strategy, consisting in progressively changing the collision angle to increase the effective section of collision and recover some luminosity, was put in place in the LHC: during most of the late 2017 operations, the pileup was thus smoothly maintained at the same level for several hours.

As a result of the evolving beam conditions and due to the features of the TT28, the VBF seed became unsustainable for the 2017 operations and some specific actions were required to keep it online.

As mentioned earlier, most of the L1 seeds using calorimetric objects were affected to some extent. To mitigate the noise at large η, the trigger primitive thresholds were raised in ECAL from 1 GeV to 3, 5 and 6 GeV in TT26, TT27 and TT28 during the 2018 data taking. Thorough studies on the optimization of the pileup mitigation weights for Run



(a) (b)

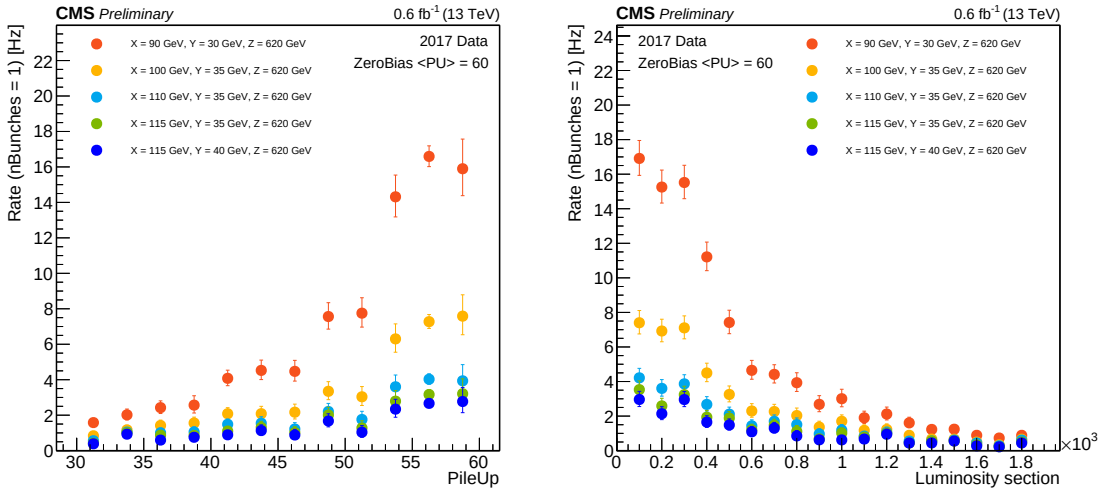
Figure 3.16 – Fraction of events with the leading jet in TT28 as a function of the jet E_{T} threshold, using ZeroBias samples from 2016 and 2017 with different pileup and bunch scheme configurations (left) and fraction of events fired by the VBF trigger where at least one of the seed jets is in TT28, as a function of the E_{T} threshold on the leading jet (right).

3 were performed [100].

As for specific measures for the VBF trigger, the most immediate solution to reduce the rate is to further raise the thresholds. As the m_{jj} threshold is already very tight, only higher p_{T} thresholds were tested, and it is observed that a tighter selection on the leading jet E_{T} is more effective than a tighter selection on the highest invariant mass pair p_{T}. The total rate per bunch for a few options is shown in Fig. 3.17 as a function of the pileup and as a function of time (expressed in "luminosity sections", each corresponding to about 30 s), in a high-pileup run without luminosity levelling. The rate is much higher at the beginning of the collision run, when the pileup is still at its nominal value, and it decreases with time. Although tighter thresholds do mitigate the pileup dependence, this strategy obviously has consequences on the acceptance on the VBFH $\rightarrow \tau\tau$ signal, which is not desirable.

Another solution would be to apply pre-scales to the VBF seed, which would drastically reduce the physics potential of the L1 VBF selection; therefore, it is not acceptable.

Another choice is to discard the L1 jets falling in the TT28 region, i.e. with $2.7 < |\eta| < 3$, from the VBF algorithm selection. This is beneficial to keep the rate of the L1 VBF trigger under control: indeed, a large fraction of events triggered by the VBF seed contain at least a jet detected in TT28 (Fig. 3.16b). The rate and the efficiency of the regular VBF seed compared to those of the modified VBF seed are shown in Fig. 3.19. While a good rate reduction is achieved, the full rejection of the TT28 jets also leads to an efficiency loss, as can be observed in Fig. 3.19b. As expected, a progressive improvement is observed for the efficiency as a function of the p_{T} of the VBF seed modified to reject all the jets in TT28: as already argued with the interpretation of Fig. 3.16, as the threshold is raised the impact of jets in the TT28 decreases and, as a consequence, the efficiency is slowly recovered at high p_{T}. The opposite tendency is observed for the efficiency as a function of m_{jj}, shown in Fig. 3.20a. Indeed, the higher the invariant mass of the jet-jet system, the higher the probability that one of the jets of the selected pair is produced with large η, hence possibly in TT28. This is illustrated in Fig. 3.20b, where the $|\eta|$



(a) (b)

Figure 3.17 – Total rate per bunch of some VBF seeds computed in a 2017 ZeroBias high pileup sample ($<$ PU $>= 60$, $n_b = 1866$) as a function of the pileup (left) and as a function of the luminosity section (right). Each luminosity section lasts about 30 s, hence the considered time range is about 15 h.

distribution of the subleading jet in the highest m_{jj} pair reconstructed offline is shown for different bins of m_{jj}: when the invariant mass is larger, the distribution is shifted towards higher $|\eta|$, and the fraction of selected jets in TT28 increases. Overall, the TT28 rejection leads to a loss of about 20%. Rather than a full rejection of the TT28 jets, a tighter E_{T} threshold can be set for jets in this region only. It was observed that the rejection of the jets in TT28 that have $E_{\mathrm{T}} < 60\,\mathrm{GeV}$, which is the most populated $|\eta| - E_{\mathrm{T}}$ region (Fig. 3.18), leads to an efficiency loss that is only minor with respect to the original L1 VBF algorithm, while providing a rate reduction very similar to that obtained with a full TT28 rejection, as shown in Fig. 3.19b and Fig. 3.20a. Therefore, this treatment of the TT28 is preferred.

In Fig. 3.21, the reduction achieved with respect to a baseline VBF seed with $X = 90$ GeV, $Y = 30$ GeV and $Z = 620$ is represented as a function of the pileup for the two different strategies: the application of tighter thresholds and the rejection of the jets in TT28. In both cases, as the pileup gets larger, the rate dependency becomes flatter (cf. Fig. 3.17a and Fig. 3.16a), resulting in a higher rate reduction. A similar rate reduction can be obtained by raising the $X(Y)$ threshold by 10(5) GeV or by rejecting the jets with $E_{\mathrm{T}} < 60\,\mathrm{GeV}$ in TT28: when $<$ PU $>\sim 35$, the rate can be reduced to about the 60% of the baseline value, and almost to its 40% as the pileup approaches 60. In conclusion, a combination of tighter thresholds and the application of the special TT28 treatment is a good compromise to achieve an acceptable rate, while giving up a reasonably small acceptance coverage.

In spite of the advantages of the improved algorithm, its implementation is computationally expensive. The original version of the algorithm requires, at μGT level, the computation of the invariant mass among a single L1 jet collection, namely that of L1 jets with $E_{\mathrm{T}} > Y$. The proposed TT28 treatment, instead, requires several L1 jets collections to be defined for the algorithm to be readable in the firmware logic: the L1 jets with $E_{\mathrm{T}} > 60\,\mathrm{GeV}$; the L1 jets with $E_{\mathrm{T}} > Y$ and $|\eta| > 3.0$; and the L1 jets with $E_{\mathrm{T}} > Y$ and $|\eta| < 2.7$. As a result, six instances of invariant mass computations must be performed. An example of the firmware code is shown in List. 3.1. However, together with

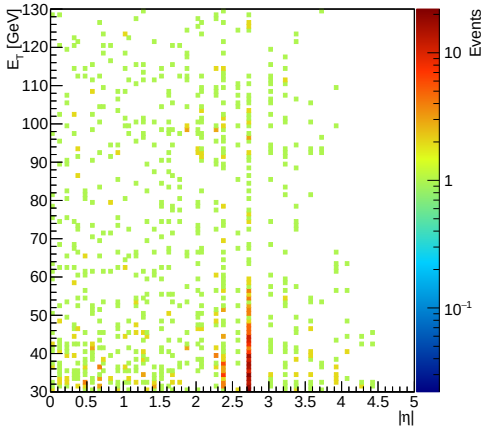
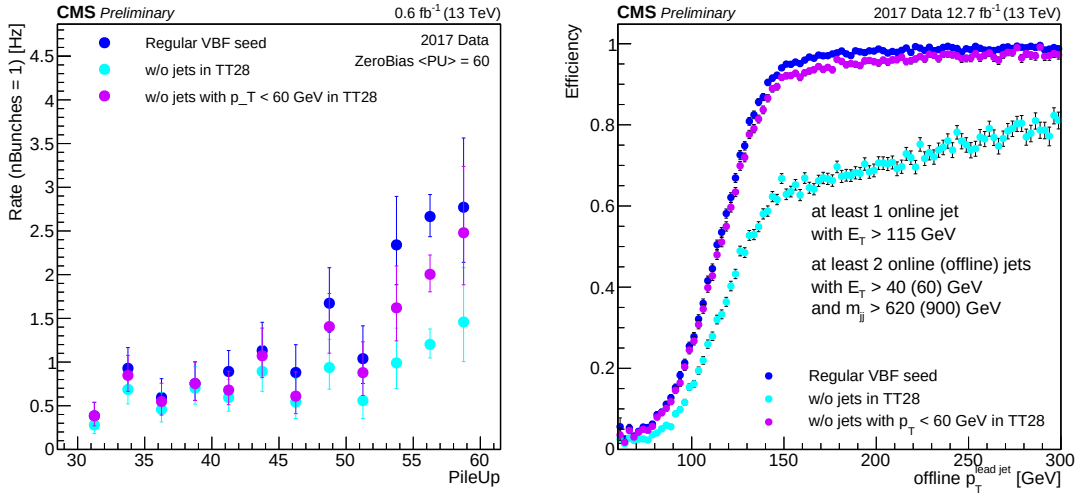


Figure 3.18 – p_T-$|\eta|$ distribution of the jets in the highest invariant mass pair of the VBF seed in 2016 data from a ZeroBias high pileup data set.



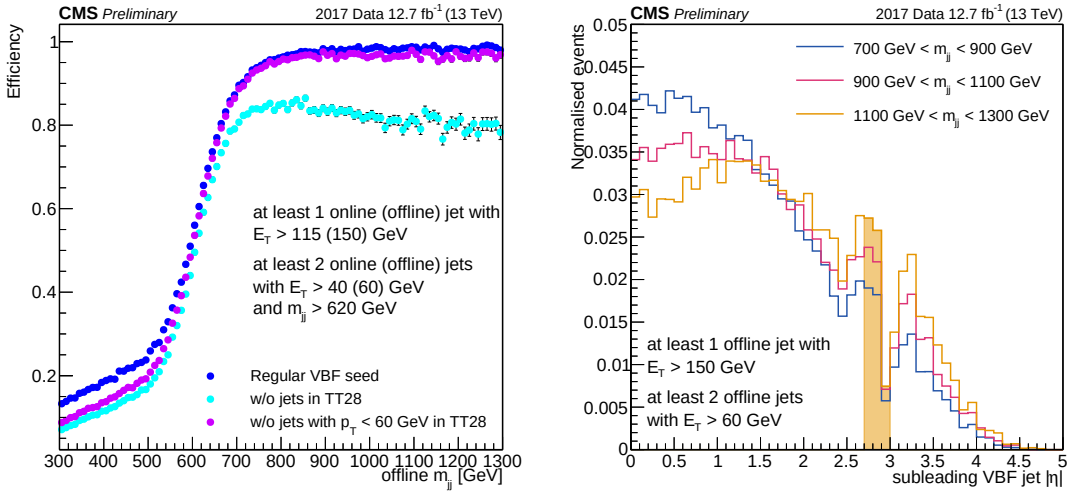
(a) (b)

Figure 3.19 – Total rate per bunch of the VBF seed with $X = 115\,\text{GeV}$, $Y = 40\,\text{GeV}$ and $Z = 620\,\text{GeV}$, computed in a 2017 ZeroBias high pileup sample ($< \text{PU} >= 60$, $n_b = 1866$) as a function of the pileup, without any TT28 treatment, with the exclusion of the entire TT28 region and with the exclusion of the jets with $E_T < 60\,\text{GeV}$ in the TT28 area (left). The corresponding efficiencies as a function of the leading jet p_T, computed in 2017 data triggered by a single muon selection, are also shown (right).

a set of standard VBF algorithms, a set of two VBF seeds with special TT28 treatment was included in the 2018 trigger selection. Their total rate computed in a 2017 ZeroBias data set with $< PU >= 60$ and extrapolated to $L = 2 \times 10^{34}\,\text{cm}^{-2}\,\text{s}^{-1}$ using $n_b = 2600$ bunches (cf. Eq. 3.2) is shown in Tab. 3.2, compared to that of the corresponding seeds without TT28 treatment.

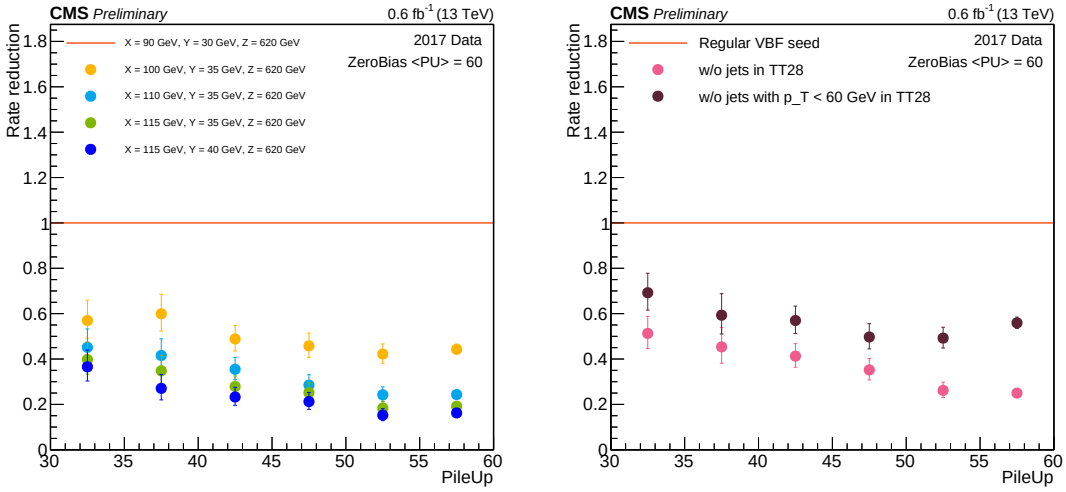
3.4 The VBF+τ_h L1 trigger

The VBF trigger strategy sets a successful example of how the CMS trigger system can be exploited to extend the acceptance on specific topologies and make analysis-targeted selections, which is especially useful for analyses that have offline thresholds



(a) (b)

Figure 3.20 – Efficiency of the regular VBF trigger, of the VBF trigger with the exclusion of the entire TT28 and of that with the exclusion of the jets with $E_{\mathrm{T}} < 60\,\mathrm{GeV}$ in TT28 as a function of the highest jet-jet invariant mass, computed in 2017 data triggered by a single muon selection (right). The $|\eta|$ distribution, normalised to unity, of the subleading jet in the highest m_{jj} pair is shown in bins of m_{jj}. The region covered by the TT28 is highlighted in yellow (left).



(a) (b)

Figure 3.21 – Rate reduction of some VBF seeds with tighter thresholds with respect to the baseline $X = 90$ GeV, $Y = 30$ GeV and $Z = 620$ (left) and rate reduction achieved by applying the TT28 special treatment (right), as a function of the pileup. The rate is computed in a 2017 ZeroBias high pileup sample ($PU =< 60 >$), as a function of the pileup.

only limited by the trigger requirements. Starting from the original VBF seed, a cross trigger (combining the information from different L1 object collections) even more specific to the VBF H $\rightarrow \tau\tau$ signal was designed for the 2018 data taking. The VBF+τ_{h} seed requires a pair of jets with moderate transverse energy $E_{\mathrm{T}} > Y$ and large invariant mass $m_{\mathrm{jj}} > Z$; in addition to targeting the VBF production, a L1 isolated tau with moderate $E_{\mathrm{T}} > X$ is required. It is yet another compromise between acceptance gain and rate reduction: the requirement of an isolated tau helps reducing the event rate, so that lower thresholds can be set on the VBF jets side, while tighter thresholds are set on the

Listing 3.1 – VBF algorithm in the firmware syntax, with and without the special TT28 treatment. The TT28 treatment makes the VBF algorithm significantly more complex.

```
% Original VBF algorithm
comb{JET115,JET40} AND mass{JET40,JET40}[MASS_MIN_620]
% Exclusion of jets in TT28 with ET < 60 GeV
(JET115 AND mass_inv{JET40_out,JET40_out}[MASS_MIN_620])
OR
(JET115 AND mass_inv{JET40_in,JET40_in]}[MASS_MIN_620])
OR
(comb{JET115,JET60} AND mass_inv{JET60,JET60}[MASS_MIN_620])
OR
(comb{JET115,JET40_out} AND mass_inv{JET60,JET40_out}[MASS_MIN_620])
OR
(comb{JET115,JET40_in} AND mass_inv{JET60,JET40_in}[MASS_MIN_620])
OR
(JET115 AND mass_inv{JET40_out,JET40_in}[MASS_MIN_620])
```

Table 3.2 – Total rate for some L1 VBF seeds implemented for the 2018 data taking, computed using 2017 ZeroBias high-pileup data ($<$ PU $>= 60$) and extrapolated to $L = 2 \cdot 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ using $n_b = 2600$.

Thresholds				
Leading jet E_T [GeV]	Jet pair E_T [GeV]	Highest m_{jj} [GeV]	TT28 rejection	L1 total rate [kHz]
115	40	620	no	10.7
120	45	620	no	7.8
115	40	620	yes	7.3
120	45	620	yes	5.7

$H \to \tau\tau$ side.

Although it is possible at L1 to make a decision based on quantities computed across different L1 object collections, the aforementioned ambiguity of physics objects entering more than one collection needs to be handled in the case of the VBF+τ_h trigger: all the genuine taus also enter the L1 jet collection, while a fraction of jets generated from quarks and gluons enter the L1 tau. If the chances that a moderate E_T tau can be part of the highest invariant mass jet-jet pair are small, the risk that a jet ends up both in the L1 jets and in the L1 taus collection is sizeable. In this case, the L1 jet could meet the jet requirements and its duplicate in the L1 taus collection could meet the tau requirements, thus events with only two jets and no taus, which are the vast majority of the events at the LHC, could fire the VBF+τ_h trigger. To avoid this effect, a minimal distance $\Delta R = 0.2$ between the L1 jets and tau firing the seed is required. As the L1 jet-tau overlap is removed, a lower m_{jj} threshold can be afforded. The total rate of the VBF+τ_h is shown as a function of the pileup in Fig. 3.22 for a few choices of the p_T thresholds X and Y. The seed was included in the 2018 data taking with thresholds $X = 35\,\mathrm{GeV}$, $Y = 45\,\mathrm{GeV}$ and $Z = 450\,\mathrm{GeV}$.

Like the inclusive VBF seed, the VBF+τ_h has a strong pileup dependence and it can benefit from the TT28 treatment. Indeed, as shown in Fig. 3.23, a $\sim 50\%$ rate reduction is achieved both by raising the tau and jets p_T thresholds by 5 and 10 GeV and by removing the jets with $p_T < 60\,\mathrm{GeV}$ in the TT28 region. Due to the complexity of the resulting algorithm (cf. List. 3.1), the VBF+τ_h seed with the TT28 exclusion was not implemented; however, such seeds are envisioned for the upcoming Run 3.

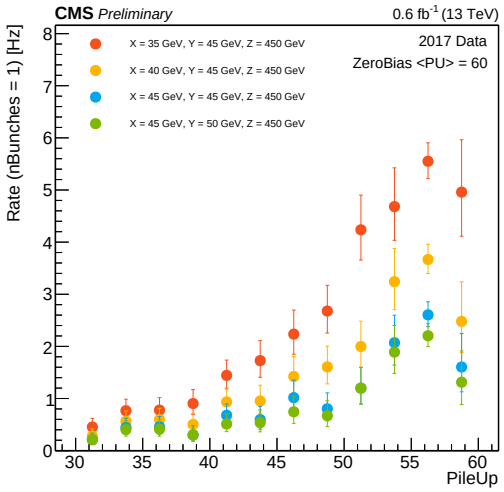


Figure 3.22 – Total rate per bunch of some VBF+τ_h seeds computed in a 2017 ZeroBias high pileup sample ($< PU >= 60$) as a function of the pileup.

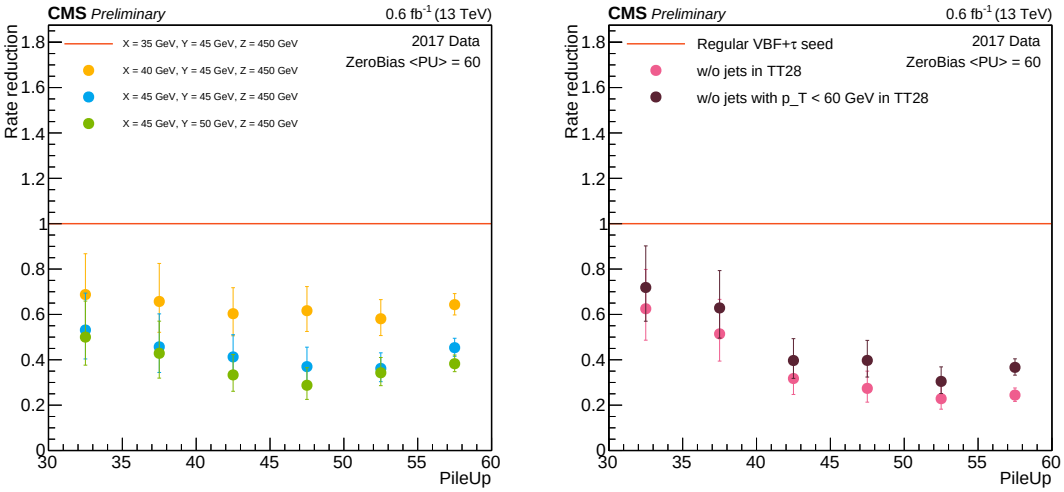


Figure 3.23 – Rate reduction of some VBF+τ_h seeds with tighter thresholds with respect to the baseline $X = 35$ GeV, $Y = 45$ GeV and $Z = 450$, as a function of the pileup (left) and rate reduction achieved by applying the TT28 special treatment (right). The rate is computed in a 2017 ZeroBias high pileup sample ($< PU >= 60$), as a function of the pileup.

3.5 The HLT VBF paths

Thanks to its general nature, the L1 VBF algorithm has many physics cases in the Higgs searches. At this moment, two dedicated sets of HLT selection criteria, or HLT "paths", are available.

The L1 VBF seed, targeting only the production mode of the Higgs boson, is a helpful handle in the search for invisible decays. The Run 2 invisible Higgs analysis relies on a trigger based on the missing transverse energy, which drives the analysis-level offline p_T^{miss} threshold to about 200 GeV. A HLT path was designed on top of the L1 VBF seed with the addition of a $E_T^{miss} > 110$ GeV requirement, and it was implemented from 2017, allowing a new region of the phase-space with lower missing transverse momentum to be probed.

The L1 VBF seed was primarly optimised to increase the H $\to \tau\tau$ acceptance. A HLT VBF H $\to \tau\tau$ path was put online as of August 2017, covering about 27 fb of the 2017 data taking and the whole 2018 data set. Although I did not participate to the its design, I contributed to the measurement of the HLT VBF H $\to \tau\tau$ efficiency.

3.5.1 The HLT VBF H $\to \tau\tau$ trigger

The selection of a VBF event at HLT is implemented as follows. In order to be selected, an event must fire at least one of the L1 VBF seeds copies available for the different luminosity configurations. On top of the L1 requirements, a typical HLT di-τ_h requirement [84] is applied: each L1 tau or central L1 jet candidate is used to build a level-2 τ-jet with a $\Delta R = 0.2$ cone using the calorimeter information; the level-2.5 step requires the L2 tau candidates to be isolated, using matching tracks reconstructed from the pixel detector; the last step, called level-3, uses the particle flow algorithm [101] to reconstruct a tau using the information from the different subdetectors. At each step, at least a pair of tau leptons needs to survive the selection. Since there is no explicit L1 selection applied to taus, the HLT tau p_T threshold can be as low as it is allowed by the reconstruction algorithms, namely 20 GeV.

Finally, the HLT particle flow jets in the highest invariant mass jet-jet pair must have transverse momenta larger than 115 GeV and 40 GeV; moreover, they should not be in overlap with any HLT hadronic tau lepton of $p_T > 20$ GeV. A simplified flowchart of the jet selection implemented in the HLT VBF H $\to \tau\tau$ trigger is represented in Fig. 3.24. A difference needs to be stressed between the L1 VBF trigger algorithm and its implementation in the HLT path: the L1 seed is potentially fired by events with 3-jet topology, as the required leading jet does not need to be one of those participating to the highest invariant mass; in the flow of the HLT path, instead, the high p_T requirement and that of large jet-jet invariant mass are not correctly implemented as independent selections: the leading jet within the highest invariant mass pair needs to pass a tight p_T threshold. This mismatch with the L1 VBF trigger requirements can be considered as a flaw and has a major effect on the additional event yield brought by the VBF trigger, as discussed in Sec. 3.6.

For a trigger path to be available for use in the analyses, the L1+HLT selection efficiency needs to be computed to provide adequate correction scale factors for the simulated events. The efficiency measurement performed for the 2017 conditions is described in the following. Additional information on the measurement of the tau legs efficiency can be found in [102]. The jet legs efficiency, for continuity with the studies on the L1 VBF trigger, was computed within this thesis work. In both cases, all the 2017 data collected since the implementation of the VBF H $\to \tau\tau$ trigger and a 2017 VBF H $\to \tau\tau$ simulation are used for the efficiency computation.

L1+HLT efficiency on tau legs

The efficiency of the di-τ_h legs can be measured through a "tag-and-probe" technique, using events triggered by a single muon selection. To do so, a HLT cross trigger requiring a muon (tag) and a hadronic tau lepton (probe), where the latter is selected with an algorithm similar to that used for the τ_h legs of the VBF H $\to \tau\tau$ trigger, is available and can be manipulated to extract the efficiency for the single tau selection. The efficiency on each τ_h-leg of the di-τ_h trigger is considered independent.

In practice, the $\mu\tau_h$ monitoring trigger implements global algorithm for the tau reconstruction, while the VBF H $\to \tau\tau$ trigger tau reconstruction algorithm is regional, i.e. it only selects taus reconstructed around the L1 candidate. In data, this technical problem

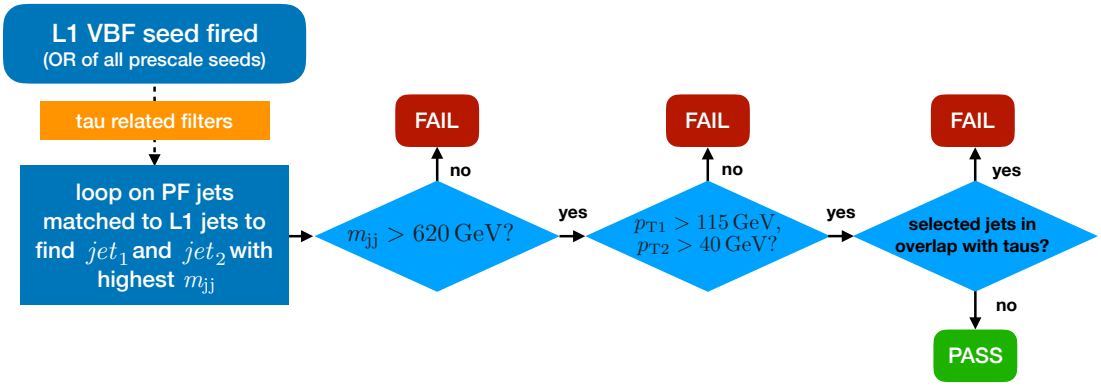


Figure 3.24 – Simplified flowchart of the VBF H $\to \tau\tau$ path jets selection. A tight p_{T} requirement is applied to the leading jet in the highest invariant mass pair, and no 3-jet topology is considered.

was overcome by re-running part of the reconstruction chain, applying a different HLT skim with a regional $\mu\tau_{\mathrm{h}}$ seed.

This operation was, unfortunately, not feasible for the Monte Carlo sample, as the intermediate steps of the simulation production chain were not available. Therefore, while the efficiency on data can be measured through with tag-and-probe technique, an alternative strategy is applied for the signal simulation efficiency measurement.

In simulated signal events, the HLT VBF H $\to \tau\tau$ path is used to compute a 2D di-τ_{h} efficiency as a function of the p_{T} of the two legs, excluding from the selection the last HLT filter with the final requirements for the particle flow jet legs; assuming that the efficiency is only a function of the kinematics, the 2D efficiency can be interpreted as a bin-by-bin product of the 1D efficiencies of each tau leg. In order to map the result from the 2D histogram to a 1D efficiency, a fit with parameters $\epsilon^{\mathrm{1D}}_{1,\ldots,N} = \epsilon^{\mathrm{1D}}_{1,\ldots,N}(p^{\tau}_{\mathrm{T}})$, corresponding to the 1D efficiency in each bin as a function of the tau p_{T}, is performed minimising the quantity

$$\chi^2 = \sum_{i,j<i}^{bins} \frac{(\epsilon^{\mathrm{1D}}_i \cdot \epsilon^{\mathrm{1D}}_j - \epsilon^{\mathrm{2D}}_{ij})^2}{\sigma^2_{ij}}$$

where $\epsilon^{\mathrm{2D}}_{ij}$ and σ_{ij} are the content of the (i,j)-th bin of the 2D efficiency and its statistical uncertainty.

As two different strategies are used for the measurement of the efficiency on data and on the signal simulation, the possible bias induced by this procedure needs to be taken into account. The outcomes of the two measurement methods are compared by applying both on an outdated signal sample configured with the regional $\mu\tau_{\mathrm{h}}$ monitoring trigger, reweighed to the most recent pileup conditions. A bin-by-bin correction is derived from the ratio of these efficiencies.

To summarise, a tag-and-probe technique, exploiting a monitoring $\mu\tau_{\mathrm{h}}$ trigger, is used to measure the HLT VBF H $\to \tau\tau$ trigger efficiency on each τ_{h} leg; in simulated events, as the monitoring trigger $\mu\tau_{\mathrm{h}}$ is not available, a 2D di-τ_{h} efficiency is computed using directly the HLT di-τ_{h} requirements of the VBF H $\to \tau\tau$ path, and a 1D efficiency is extrapolated from it; the difference in the performance of the two efficiency measurements is taken into account for the scale factor computation. The final selection efficiencies on each tau leg of the HLT VBF H $\to \tau\tau$ path are shown in Fig. 3.25 for data and simulated events. Their ratio is the correction scale factor to be applied to each tau in simulated events. The efficiency on data is significantly lower than that on simulated samples over

most of the considered p_{T} range: in particular, the scale factor is as low as 0.7 in for taus of $p_{\mathrm{T}} \sim 30\,\mathrm{GeV}$. However, this trend is compatible with that observed for the measured scale factors of other di-τ_{h} triggers.

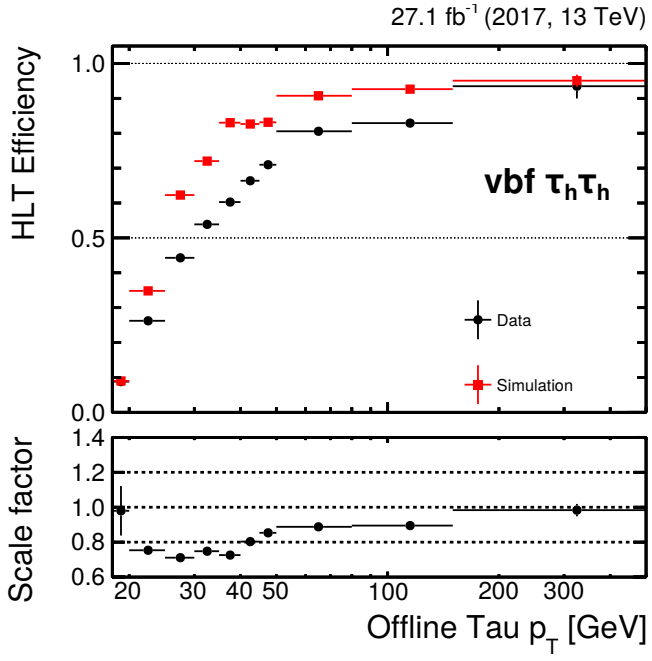


Figure 3.25 – Single tau leg selection efficiency of the HLT VBF H $\to \tau\tau$ trigger as a function of the p_{T} of the tau reconstructed offline. The correction scale factor for simulated events is shown in the bottom pad. The vertical error bars are to small to be appreciated [102].

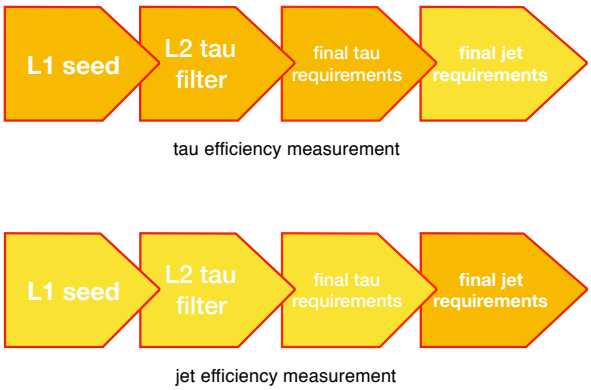


Figure 3.26 – Tau leg and jet leg efficiency measurement strategy: the tau selection efficiency is computed using only the HLT filters corresponding to the HLT H $\to \tau\tau$ requirements, while the jet selection efficiency is computed using only the final jet filters.

L1+HLT efficiency on jet legs

As for the jet legs efficiency, it is computed both for data and signal simulation in events triggered by the regular di-τ_{h} triggers and it is defined as the efficiency of passing the final jet requirements of the HLT VBF H $\to \tau\tau$ trigger, while no requirement is made on the filters corresponding to the di-τ_{h} selection within the HLT VBF H $\to \tau\tau$ trigger: assuming that the efficiency of the tau selection is uncorrelated from that of the jets

selection, they can be factorised and computed separately, as sketched in Fig. 3.26. The offline event selection is similar to that used in the 2016-2017 H $\rightarrow \tau\tau$ analysis [96]. The pair with the most isolated hadronic taus within the $|\eta| < 2.1$ region is selected; in order to have a close to 100% di-τ_{h} trigger selection efficiency, a $p_{\mathrm{T}} > 40\,\mathrm{GeV}$ threshold is applied to both offline taus, each of them being required to be geometrically matched to HLT taus; very loose identification criteria are required, to preserve the statistics, while the regular discriminants against e/μ leptons are applied.

The jet pair selected as VBF pair candidate is chosen as the one with the highest invariant mass, coherently with the VBF trigger requirements, even though the two highest p_{T} jets are preferred in the existing H $\rightarrow \tau\tau$ analyses. The latter strategy is not optimal to exploit the VBF trigger acceptance and the trigger design should drive the choice of the jet pair selection criteria in the future analyses. Besides, the VBF signature is characterised by small hadronic activity between the tagging jets, and the two highest p_{T} jets are often those with the highest invariant mass. Events with additional leptons are rejected, as well as those where one of the VBF jets candidates is reconstructed with $p_{\mathrm{T}} < 50\,\mathrm{GeV}$ in the noisy area of the detector with $2.65 < |\eta| < 3.14$.

The efficiency is a function of three variables: the p_{T} of each of the jets in the highest invariant mass pair, and their m_{jj}. The following strategies are considered:

a. if they are uncorrelated, the 3D efficiency can be factorised and a 1D×1D×1D efficiency would provide a good description of the 3D efficiency while keeping a fine binning;

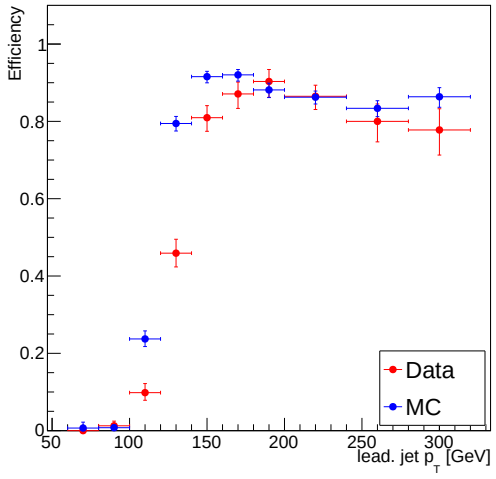
b. the 3D efficiency has the advantage of capturing the correlations, but at the cost of using a coarse binning to minimise the statistical fluctuations, which results in a MC correction that is averaged over wide regions of the phase-space;

c. a 1D×2D factorisation can be suitable as a good compromise between statistics and physics description.

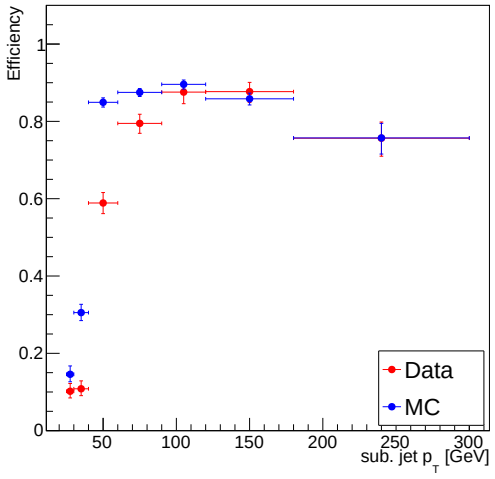
The 1D efficiencies are shown in Fig. 3.27. Due to the design of the HLT path, the variables that are relevant for the efficiency measurement are the p_{T} of the two jets in the highest invariant mass pair and their invariant mass. For each 1D efficiency, a tight offline threshold is applied to the other two relevant variables to exclude their turn-on regions and factor out the potential inefficiency in the other dimensions: the leading and subleading p_{T} must be larger than 140 GeV and 60 GeV, and the invariant mass should be greater than 800 GeV. All the 1D efficiencies present a sharp turn-on but, in spite of the tight selections, their plateau never reaches the value of 1. Moreover, the efficiency on the two jets selection tends to decrease for high p_{T}. This observation is only qualitative: the 1D efficiencies under study are strongly correlated among them and a selection inefficiency smaller than 4% can translate in an observed 20% inefficiency in the 1D projection. A few hypotheses can be made to understand this behaviour:

- as a consequence of the suboptimal jet matching performed in the HLT path, when the L1 VBF trigger is fired by a 3-jet topology, i.e. it has a high p_{T} jet distinct from those in the largest m_{jj} pair, the HLT selection is not at the efficiency plateau since there is no explicit threshold on the third jet;
- the efficiency of the jets selection could have some sizeable correlation with that of the tau selection.

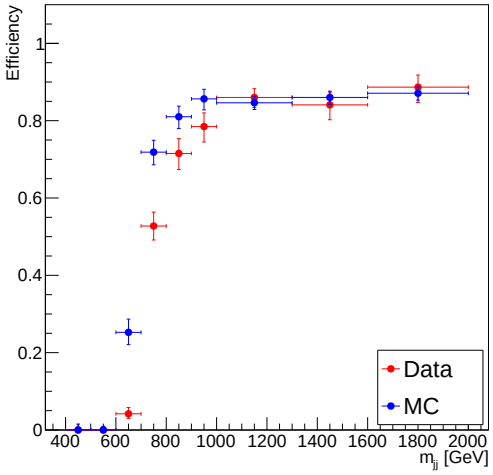
The cause of the efficiency drop was not identified; however, a flat plateau is observed in 2018 HLT VBF H $\rightarrow \tau\tau$ efficiency measurements.



(a)



(b)

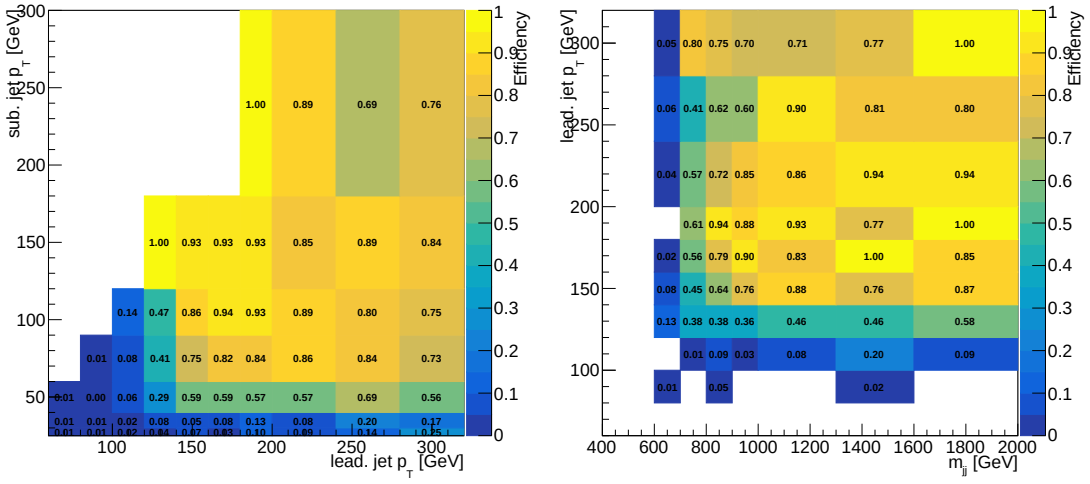


(c)

Figure 3.27 – 1D efficiencies as a function of the leading jet p_{T} using events with subleading jet $p_{\mathrm{T}} > 60\,\mathrm{GeV}$ and $m_{\mathrm{jj}} > 800\,\mathrm{GeV}$ (a), of the subleading jet p_{T} using events with leading jet $p_{\mathrm{T}} > 140\,\mathrm{GeV}$ and $m_{\mathrm{jj}} > 800\,\mathrm{GeV}$ (b), and of highest invariant mass pair m_{jj} using events with leading jet $p_{\mathrm{T}} > 140\,\mathrm{GeV}$ and subleading jet $p_{\mathrm{T}} > 60\,\mathrm{GeV}$ (c). The efficiency is computed on events firing the classic di-τ_{h} triggers in 2017 data and in a 2017 VBF $\mathrm{H} \to \tau\tau$ simulation.

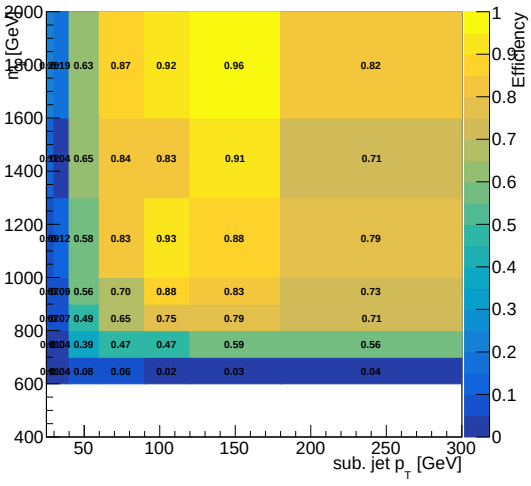
The most simple study of the validity of the approach **a.** is to carry out the comparison of the 1D×1D to the 2D efficiencies, which are shown for data and VBF $\mathrm{H} \to \tau\tau$ signal simulation in Fig. 3.28 and Fig. 3.29, where one can see that for all the combinations, the product of the efficiencies is overall underestimated with respect to those computed in 2D. It is however not sufficient to conclude on the validity of the factorised approach. Indeed, if the 1D×1D over 2D ratio is similar for data and Monte Carlo, this effect can be cancelled, and the scale factors could still be acceptable corrections for the Monte Carlo.

To test the incompatibility of the two approaches, for each combination an estimator is



(a)

(b)



(c)

Figure 3.28 – 2D efficiencies as a function of the leading and subleading jet p_{T} (a), of the highest invariant mass pair m_{jj} and the leading jet p_{T} (b), and of the subleading p_{T} and the highest invariant mass pair m_{jj} (c) computed on events firing the classic di-τ_{h} triggers in 2017 data.

built computing, for each of the bins with efficiency at plateau,

$$\mathrm{P} = \frac{\epsilon_{1\mathrm{D}x} \cdot \epsilon_{1\mathrm{D}y} - \epsilon_{2\mathrm{D}}}{\sqrt{\sigma^2_{1\mathrm{D}x\times1\mathrm{D}y} + \sigma^2_{2\mathrm{D}}}} \tag{3.4}$$

where $\sigma_{2\mathrm{D}}$ is the uncertainty on the 2D efficiency and

$$\sigma_{1\mathrm{D}x\times1\mathrm{D}y} = \sqrt{\sigma^2_{1\mathrm{D}x} + \sigma^2_{1\mathrm{D}y}}$$

is that of the 1D×1D efficiency. The distributions, represented in Fig. 3.30, show that the 1D×1D approximation is closer to the corresponding 2D efficiency for data than for Monte Carlo. This suggests that the strategy **a.** is to be discarded: the estimator P (Eq. 3.4) would need to be computed for every possible simulated process to ensure the approximation works, which makes this solution unpractical. The P distributions of 1D×1D×1D vs. the 3D efficiency, shown in Fig. 3.31, are made by extending the Eq. 3.4 and the comparison confirms that this approximation is not suitable.

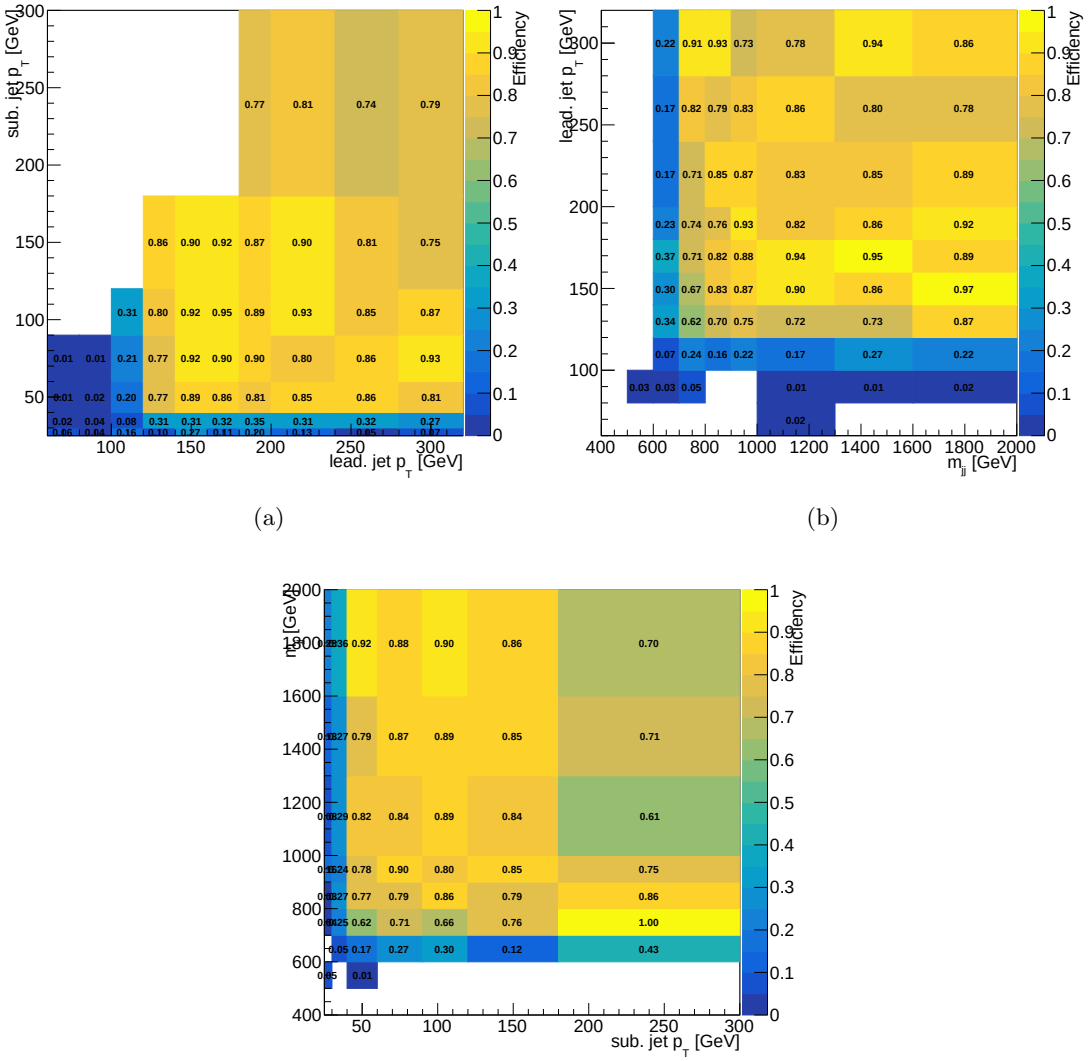


Figure 3.29 – 2D efficiencies as a function of the leading and subleading jet p_{T} (a), of the highest invariant mass pair m_{jj} and the leading jet p_{T} (b), and of the subleading p_{T} and the highest invariant mass pair m_{jj} (c) computed on events firing the classic di-τ_{h} triggers in a 2017 VBF H $\rightarrow \tau\tau$ simulation.

Nonetheless, it is possible that some 2D efficiency captures the correlation and can be used for a 1D×2D approximation of the 3D efficiency. The estimator P distributions in Fig. 3.32 show that none of the permutations gives a satisfactory agreement for data and Monte Carlo; hence, the approach **c.** is also ruled out.

Therefore, a 3D parametrisation of the efficiency is necessary. The slices on 2D planes of the 3D efficiency, with a tight offline threshold on the variable on the third dimension not represented, are shown in Fig. 3.33. The systematic uncertainty is estimated by sliding by ±5 GeV the bins edges on the p_{T} axes and by ±50 GeV those on the m_{jj} axis, alternately. The relative difference with respect to the nominal binning ranges from 1 to 4%; thus, a conservative 4% systematic uncertainty is associated to all scale factors.

In Fig. 3.34, the data over signal simulation efficiency scale factors are shown. Overall, the efficiency on the simulated VBF H $\rightarrow \tau\tau$ signal is higher than that in the data. In

particular, comparing to Fig. 3.33, the efficiency rise as a function of the p_T of the VBF jet candidates appears sharper for data. Moving towards the plateau, the efficiencies are similar and their ratio approaches the value of 1. These are the final corrections accepted by the H $\rightarrow \tau\tau$ working group in view of the usage of the VBF H $\rightarrow \tau\tau$ trigger in the analysis.

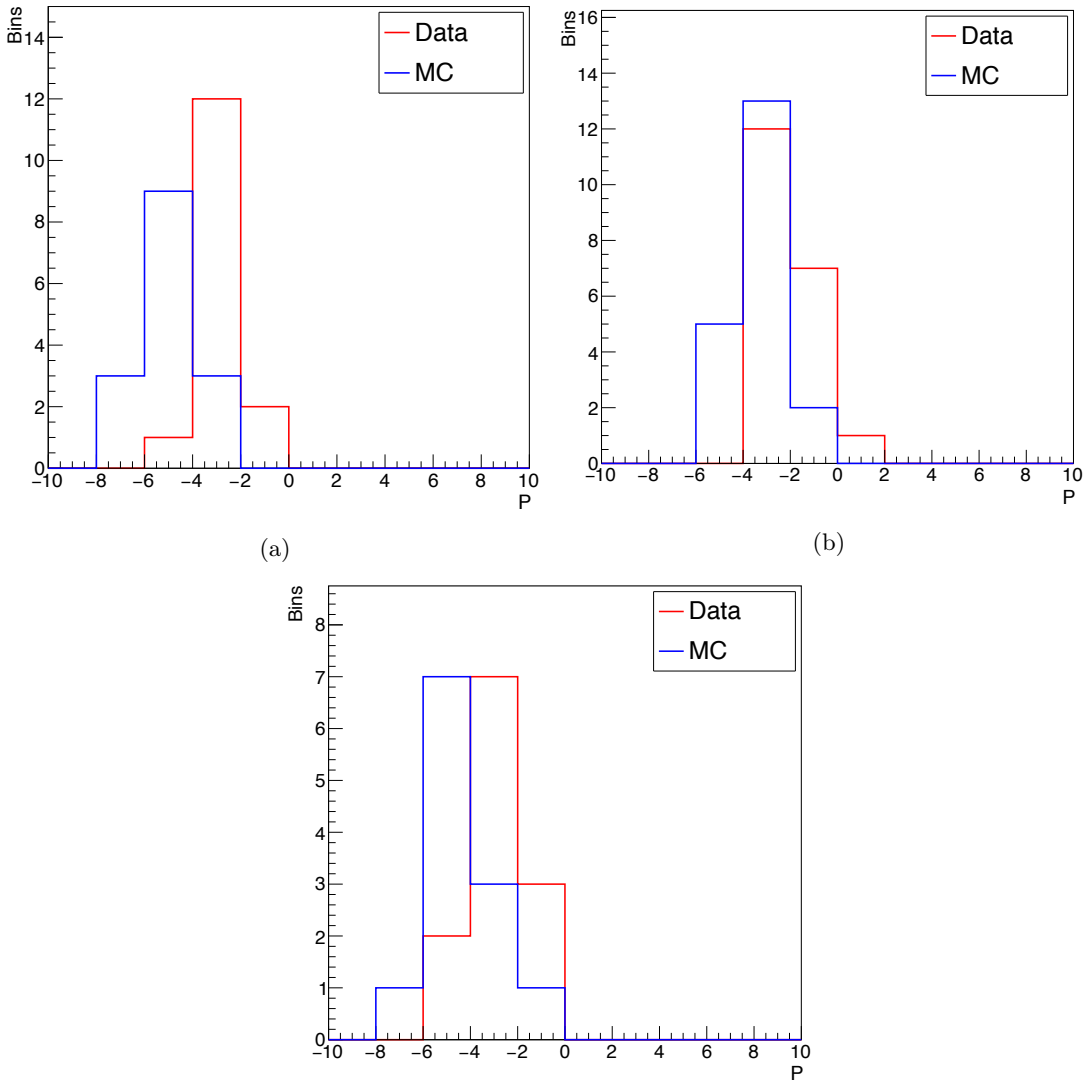


Figure 3.30 – Distribution of Eq. 3.4 for 1Dx1D efficiencies vs. 2D efficiencies as a function of the leading and subleading jet p_T (a), of the highest invariant mass pair m_{jj} and the leading jet p_T (b), and of the subleading p_T and the highest invariant mass pair m_{jj} (c) computed on events firing the classic di-τ_h triggers in 2017 data and in a 2017 VBF H $\rightarrow \tau\tau$ simulation.

3.6 Evaluation of the performance in the 2017 H $\rightarrow \tau\tau$ analysis

The preliminary estimation of the VBF H $\rightarrow \tau\tau$ signal event yield brought by the use of the L1 VBF seed to complement the classic di-τ_h trigger, illustrated in Sec. 3.1.2, gives 50% additional events for the loosest VBF trigger selection that was online throughout 2017 (cf. Tab. 3.1). The measurement underestimates the impact of the HLT resolution and of the HLT path design described in Sec. 3.5.1, which has a two-fold effect. Firstly,

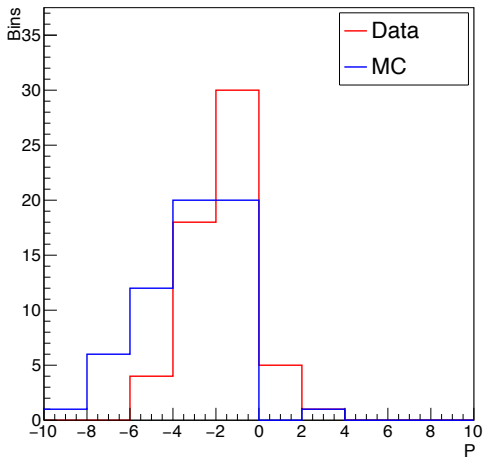


Figure 3.31 – Distribution of Eq. 3.4 for 1Dx1Dx1D efficiencies vs. 3D efficiency as a function of the leading and subleading jet p_{T} and the highest invariant mass pair m_{jj}, computed on events firing the classic di-τ_{h} triggers in 2017 data and in a 2017 VBF H → ττ simulation.

the HLT VBFH → ττ jet selection efficiency is never 100%. Secondly, the VBF L1 jet legs matching performed in the HLT path is suboptimal: the leading jet in the highest m_{jj} pair is required to have large p_{T}, while typically moderate E_{T} jets (see, for instance, Fig. 3.1) are produced by the VBF process. While the inefficiency of the jet selection at HLT has a minor effect, the jet selection scheme itself has a major impact on the final acceptance. As a consequence of the tight p_{T} threshold set on one of the VBF jet candidates at HLT level, a tighter threshold must be applied in the analysis offline selection. Consequently, the additional phase space brought by the L1 VBF trigger selection is dramatically reduced due to the unfortunate jet selection scheme applied at HLT.

The event yield gain is computed with a 2017 VBF H → ττ signal simulation. Rather than offline selections tailored to those that are used online by the L1 VBF trigger (Sec. 3.1.2), the full H → ττ analysis selection is used, including the HLT di-τ_{h} and VBF H → ττ requirements recommended for the 2017 data analyses. The coverage of the L1 VBF seed as estimated in Sec. 3.1.2 is compared to the actual coverage of the HLT VBFH → ττ trigger in Fig. 3.35 in a view complementary to that shown in Fig. 3.6.

The H → ττ analysis event selection is similar to the baseline selection used in the HH → bbττ analysis for the H → ττ system, detailed in Sec. 4.3. The taus reconstructed offline are required to pass tight identification criteria and to be well discriminated from electrons and muons; the selected ττ pair is the one with the most isolated hadronic tau leptons in the event, required to have opposite charge. The jets constituting the pair giving the highest invariant mass are chosen as VBF jet candidates; they are also required to pass tight isolation criteria, to be separated from the selected hadronic tau leptons by $\Delta R(jet, \ell) > 0.5$, and not to be reconstructed with $p_{\mathrm{T}} < 50\,\mathrm{GeV}$ and $2.65 < |\eta| < 3.14$, as recommended for 2017 analyses (see Sec. 4.3). All events where at least one electron or a muon is reconstructed with loose identification criteria are rejected.

The events in the category "Only di-tau" must fire at least one of the di-tau HLT paths recommended for the 2017 analysis. The selected taus are required to be geometrically matched to HLT taus and to have $p_{\mathrm{T}} > 40\,\mathrm{GeV}$, while the VBF jet candidates must have $p_{\mathrm{T}} > 35\,\mathrm{GeV}$ and $m_{\mathrm{jj}} > 400\,\mathrm{GeV}$. The "Only VBF" category contains events that fired the HLT VBF H → ττ path, where the selected hadronic tau leptons and the jets are geometrically matched to the corresponding HLT objects. Coherently with the HLT

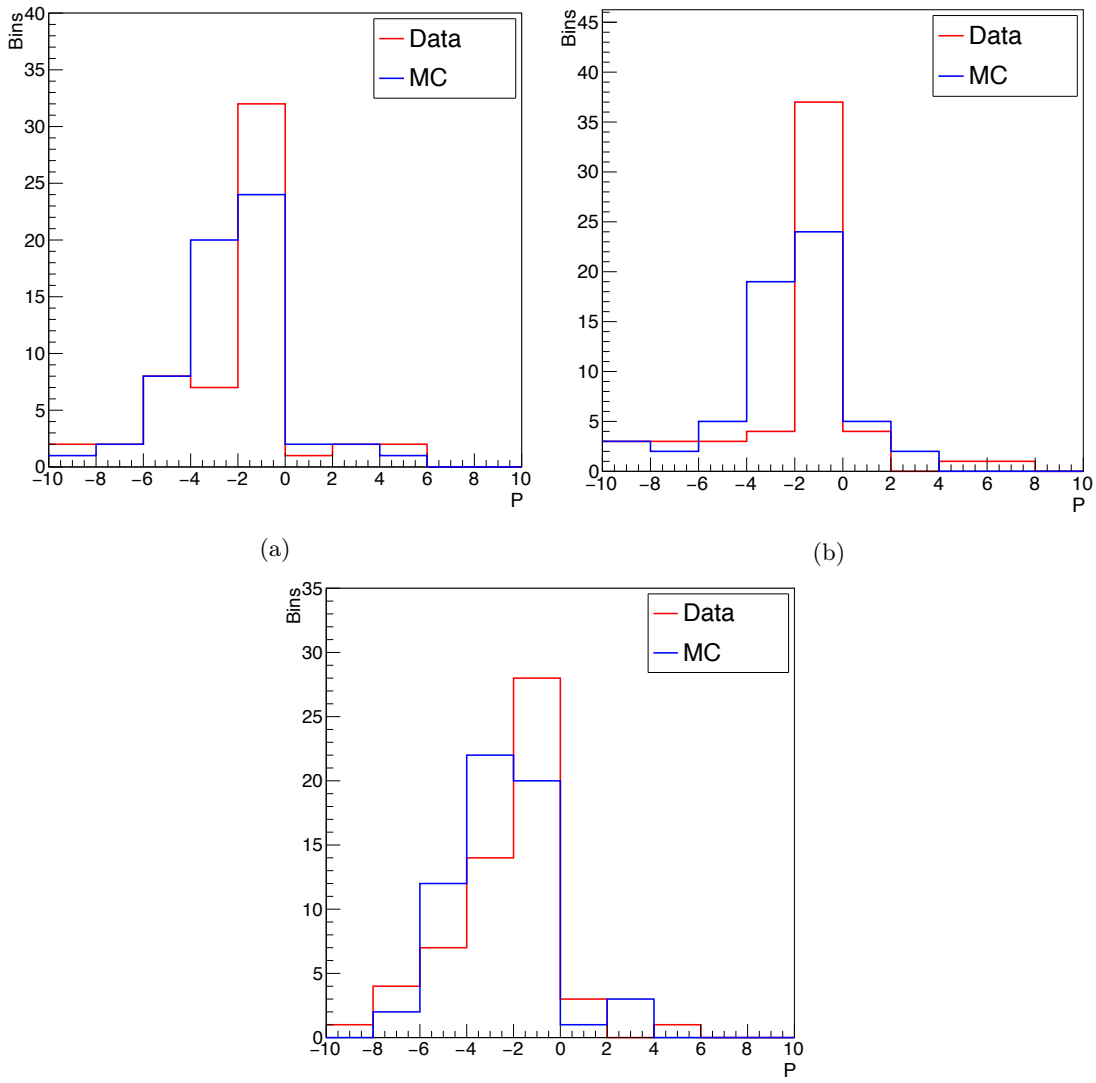
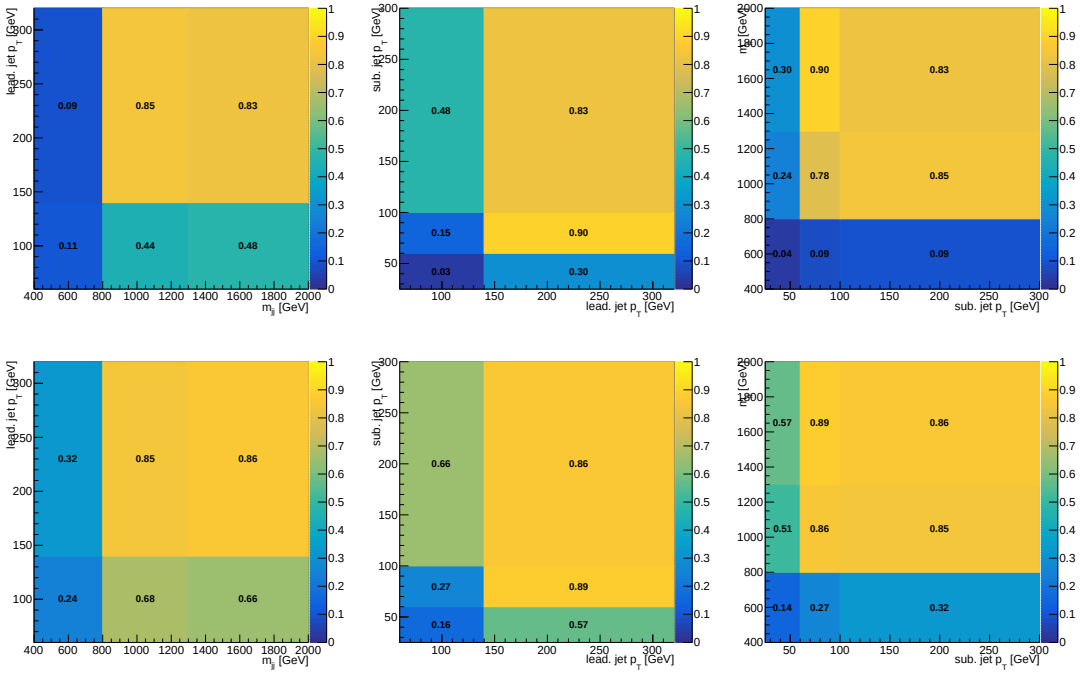


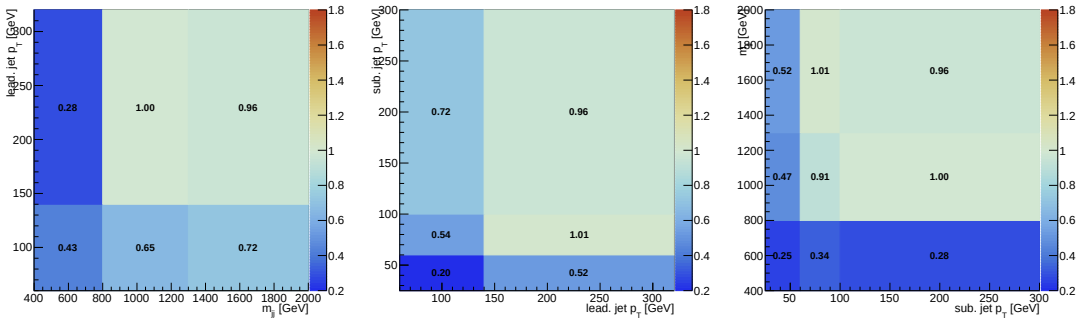
Figure 3.32 – Distribution of Eq. 3.4 for 1Dx2D efficiencies vs. 3D efficiencies, where the 2D efficiency considered are that as a function of the leading and subleading jet p_{T} (a), of the highest invariant mass pair m_{jj} and the leading jet p_{T} (b), and of the subleading p_{T} and the highest invariant mass pair m_{jj} (c), computed on events firing the classic di-τ_{h} triggers in 2017 data and in a 2017 VBF H $\rightarrow \tau\tau$ simulation.

selection and its efficiency computation, the VBF jet candidates must have respectively $p_{\mathrm{T}} > 140$ and 60 GeV and invariant mass $m_{\mathrm{jj}} > 600$ GeV; the selected hadronic tau lepton p_{T} threshold is as low as 25 GeV, allowing the "Only VBF" category to cover a region inaccessible to the "Only di-tau" events. These two categories are exclusive, while "di-tau and VBF" is filled with events that pass both the di-tau and the VBFH $\rightarrow \tau\tau$ triggers and the corresponding offline selections. Correction scale factors corresponding to the di-τ_{h} trigger efficiency are applied to each tau in the events of the "Only di-tau" and "di-tau and VBF" categories, while the scale factors discussed in Sec. 3.5.1 are applied to taus and jets selected in the events of the "Only VBF" category. The distribution of the events in the categories thus defined is shown in Fig. 3.36, as a function of m_{jj} and of $p_{\mathrm{T}}^{sub\,\tau}$. These plots can be compared to those in Fig. 3.7: the VBF H $\rightarrow \tau\tau$ trigger additional events, represented in blue histograms, come from the low tau p_{T} region. This additional region, smaller than what could be obtained by fully exploiting the potential



(a) (b) (c)

Figure 3.33 – Slices on 2D planes of the 3D efficiency, represented as a function of the leading and subleading jet p_{T} (a), of the highest invariant mass pair m_{jj} and the leading jet p_{T} (b), and of the subleading p_{T} and the highest invariant mass pair m_{jj} (c), computed on events firing the classic di-τ_{h} triggers in 2017 data (top) and in a 2017 VBF H → ττ simulation (bottom).



(a) (b) (c)

Figure 3.34 – Slices on 2D planes of the scale factor between the 3D efficiencies on data and VBF H → ττ Monte Carlo, represented as a function of the leading and subleading jet p_{T} (a), of the highest invariant mass pair m_{jj} and the leading jet p_{T} (b), and of the subleading p_{T} and the highest invariant mass pair m_{jj} (c), computed on events firing the classic di-τ_{h} triggers.

of the L1 VBF seed, corresponds to a $N_{\mathrm{Only\,VBF}}/N_{\text{di-tau}} = 14\%$ event yield gain.

The additional event yield is also computed using a 2017 VBF HH → bb$\tau\tau$ simulation. On top of the selections already mentioned, two offline jets with $p_{\mathrm{T}} > 20\,\mathrm{GeV}$, $|\eta| < 2.4$ and with minimal distance $\Delta R(jet, \tau_{\mathrm{h}}) > 0.5$ from the selected taus are required in all categories; at least one of the selected jets must pass a b tag requirement. The measured event yield gain is $N_{\mathrm{Only\,VBF}}/N_{\text{di-tau}} = 17\%$: for a signal as rare as the VBF HH production, the collection of this sizeable fraction of additional events is a major achievement.

In conclusion, the suboptimal jet legs matching between the L1 VBF seed and the HLT

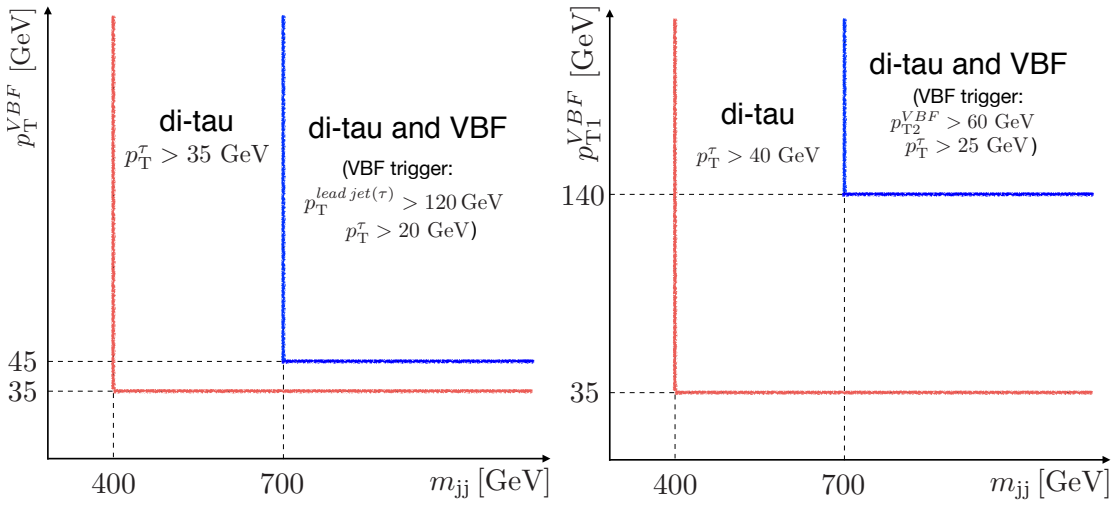


Figure 3.35 – Definition of the categories, in a view complementary to that shown in Fig. 3.6, for the computation of the event yield gain from the use of the L1 VBF trigger (left) or the HLT VBF H $\to \tau\tau$ trigger (right) to complement the double-τ_h trigger. The minimum transverse momentum of the VBF jet candidates is p_T^{VBF}, while $p_T^{lead\ jet(\tau)}$ is referred to the leading jet or τ in the event.

VBF H $\to \tau\tau$ path induces a large reduction of signal acceptance throughout the selection chain. In Tab. 3.3, the expected event yield gain computed in Sec. 3.1.2 is compared to the realistic estimate computed in this section. However, the L1 VBF seed, described extensively in this chapter, brings a sizeable event yield gain both for the VBF H$\to \tau\tau$ and the VBF HH$\to$bb$\tau\tau$ signal.

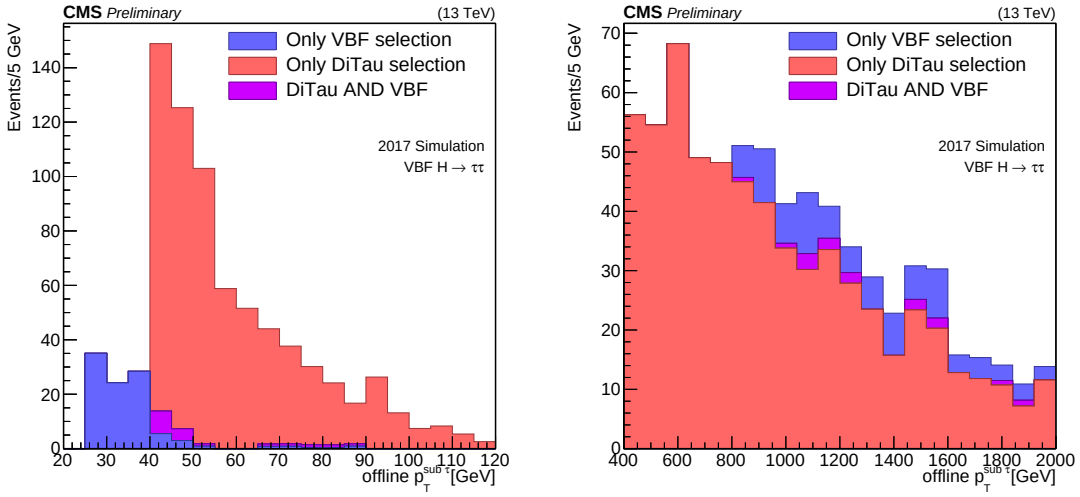
Table 3.3 – Event yield gain of VBF H$\to \tau\tau$ and VBF HH$\to$bb$\tau\tau$ signals. The L1 estimate is recomputed using the strategy described in Sec. 3.1.2, with L1 VBF seed thresholds used throughout the 2017 data taking. The L1+HLT estimate is computed consistently in Sec. 3.6.

Signal	L1 estimate	L1+HLT estimate
VBF H$\to \tau\tau$	43%	14%
VBF HH$\to$bb$\tau\tau$	51%	17%

3.7 Conclusion and perspectives

The CMS L1 trigger system upgrade made possible the design of sophisticated correlation algorithms. Its flexibility can be used to make trigger algorithms that are tailored to the physics needs, as done for the first L1 VBF trigger algorithm, described in this chapter. In its original version, the L1 VBF trigger only targets the production mode of the Higgs boson, with the two-fold advantage of covering a new portion of the phase-space by expanding overall the acceptance on the signal, and of being general enough to benefit searches for different Higgs boson decay channels. Thanks to further developments made at HLT, it is available already from 2017 for the use in analyses featuring Higgs bosons decaying in two hadronic tau leptons, and those of the Higgs boson invisible decay. Within this thesis work, the L1+HLT VBF trigger strategy has great relevance: the ultimate target of increasing the VBF HH $\to$ bb$\tau\tau$ event collection was successfully met with an estimated event yield gain of 17%.

Moreover, an additional set of L1 VBF seeds was provided to guarantee its online main-



(a) (b)

Figure 3.36 – Additional events from the HLT VBF H $\rightarrow \tau\tau$ path with respect to the OR of the double-τ_h paths as a function of the p_T of the offline subleading tau (a) and of the invariant mass of the offline jets with highest m_{jj} (b). The events in each of the categories "Only VBF selection" and "Only di-tau selection" are events that fired only one of the triggers and pass only the offline selection corresponding to that trigger. The events in the category "di-tau AND VBF" fire both triggers and pass both offline selections.

tenance in harsh data taking conditions and it is available for 2018 physics analysis. A L1 VBF+τ_h algorithm, even more analysis specific, was implemented for testing in 2018: the required hadronic tau lepton allows to bring down the thresholds on the p_T and m_{jj} of the L1 jets, an important limitation in the original L1 VBF seed.

The L1 trigger menu follows the physics program the collaboration and evolves with it. As the community moves on to the next phases of data taking, such as Run 3 and Phase 2, the trigger strategy is re-designed to face the new challenges and adjust to the physics priorities. The design of dedicated triggers will become increasingly important for statistics limited analyses: the double Higgs search, for example, is an excellent case. The experience acquired during Run 2 is a strong base to be exploited in the next phases.

Chapter 4

The HH$\to$ b$\bar{\text{b}}\tau^+\tau^-$ event selection

In this chapter, the analysis strategies aimed at the search for pairs of Higgs bosons -one decaying in two of tau leptons, the other in two jets originating from b quarks- are presented. Other searches for the production of Higgs boson pairs in the bb$\tau\tau$ final state were already performed with the Run 1 data by both the ATLAS [103] and CMS [104] collaborations. These searches used LHC data collected at $\sqrt{s} = 8\,\text{TeV}$, by the CMS collaboration using $2.7\,\text{fb}^{-1}$ of data collected during 2015 at $\sqrt{s} = 13\,\text{TeV}$ [105, 106] and by the CMS collaboration using $35.9\,\text{fb}^{-1}$ of data collected during 2016 at $\sqrt{s} = 13\,\text{TeV}$ [107]. The latter result has been included in the CMS HH combination [108] released in 2018. In this chapter and in the following, I will sometime refer to the H $\to \tau\tau$ analysis [96], which is a flagship for the tau identification; therefore, it is a useful comparison.

In this thesis, the 2017 data set is analyzed; it corresponds to $41.6\,\text{fb}^{-1}$ of proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$. Building on the 2016 HH $\to$ bb$\tau\tau$ search published in [107], I have worked to introduce extensions and improvements to boost the analysis sensitivity and broaden the physics interpretation.

The analysis flow essentially follows the strategy used in the previous search: the tau lepton candidates are identified, consistently with the trigger requirements; the selected events need to have two b jet candidates and they are classified based on the number of jets passing specific b tag selections; finally, a selection based on the mass of the reconstructed Higgs boson candidates is applied to further reject the background processes.

The physics searches are often the result of a team work. I have been the main analyzer of the group, leading the effort of performing the HH $\to$ bb$\tau\tau$ analysis with 2017 data; I have been developing and maintaining the existing analysis software, running the analysis of the data and their statistical interpretation. My personal contribution is summarised in the following.

The trigger strategy is improved by complementing single-lepton triggers with cross triggers, i.e. triggers combining multiple objects, and by including the VBF trigger described in Ch. 3. The mixture of various kind of triggers requires an attentive handling of the corresponding efficiencies and of the offline object selection. The trigger requirements are described in Sec. 4.2.

The H$\to \tau\tau$ candidates selection is also bound to the trigger requirements, as detailed in Sec. 4.3. Due to the observation of a large data-over-prediction disagreement affecting events with pairs of hadronic tau leptons, I carried out an extensive investigation on several elements of the object identification, of the background processes modelling and

of the analysis flow. In this context, I have computed a dedicated correction for the tau lepton identification efficiency for the HH → bb$\tau\tau$ analysis phase space. These results have stimulated the improvement of the tau efficiency measurement and correction within the CMS experiment. This work is summarised in Appendix A. As mentioned in Sec. 1.3, most of the existing HH searches are optimized for the gluon fusion HH signal only; this is also the case of the 2016 HH → bb$\tau\tau$ search. I have introduced a new dedicated category to study the VBF production, giving access to additional interactions of the Higgs boson, as described Sec. 1.2.1. In this chapter, the definition of specific strategies for the VBF HH signal extraction are described. The selection of VBF jet candidates was optimized and, inevitably, the b jet candidates selection was tuned for a better jet assignment.

A key aspect in the enhancement of the sensitivity compared to the the previously published results is the introduction of an improved multivariate discriminant for the $t\bar{t}$ rejection, described in Sec. 4.6. Its training was not performed within this thesis work and it is documented in [109]. I have tested it and implemented it in the current analysis.

4.1 The HH → bb$\tau\tau$ signal

As discussed in Sec. 1.3, the signal consists of two Higgs bosons produced either through gluon fusion or vector boson fusion. The case where one Higgs boson decays in a pair of b quarks and one in a pair of tau leptons is a good compromise between efficiency and purity. Indeed, it combines a sizeable branching ratio, of BR(H → $\tau\tau$) = 7.3%, with the high $\tau\tau$ pair selection purity; it is the HH final state studied in this search. The tau lepton being an unstable particle, the $\tau\tau$ pair itself can be reconstructed in several final states. The tau lepton mainly decays hadronically, with an overall branching ratio of 64.79%; the leptonic decays $\tau \to e + \nu_e + \nu_\tau$ and $\tau \to \mu + \nu_\mu + \nu_\tau$, denoted in the following as τ_e and τ_μ, have similar branching ratio of 17.82% and 17.39% [7]. These values are used to compute the branching ratio of the $\tau\tau$ final states, shown in Fig. 4.1. In this search, only the three dominant final states, with at least one hadronic tau lepton, are considered for the signal extraction; altogether, they cover the 87.6% of the cases. Events with a $\mu\mu$ pair in the final state are only used as define sideband regions for the Drell-Yan background modelling (Sec. 5.4).

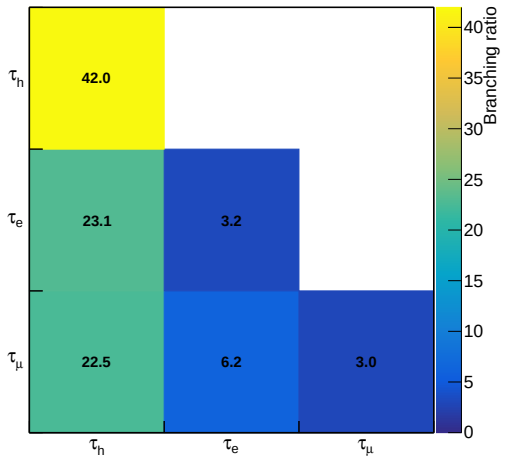


Figure 4.1 – Branching ratio, expressed in %, of the $\tau\tau$ final states as combinations of τ_h, τ_e, τ_μ.

The non-resonant HH production, introduced in Sec. 1.2, is explored in this search: its cross section through gluon fusion and through VBF amounts to about 31 fb and 1.7 fb, respectively. These processes are predicted in the Standard Model, but a variety of anomalous couplings scenario brought by BSM effects are also explored.

The gluon fusion signal simulations correspond to different values of the coefficients of the effective Lagrangian model introduced in Ch. 1: y_t, λ_{HHH}, c_2, c_{2g} and c_g. Six sets of anomalous couplings, among those identified in the benchmark model, are considered in this analysis. As for the VBF production, simulations with different combinations of the λ_{HHH}, c_V and c_{2V} couplings are used for the signal extraction. The modelling of these signal processes is discussed in Sec. 5.1 and Sec. 5.3.

Several processes have final state similar to the signal under study and the extraction of the signal from data requires a meticulous study of their features. The background processes can be separated into two categories: the "irreducible" backgrounds have the same final state of the signal, while for the "reducible" backgrounds the final state similarity results from the misidentification of an object. In general, processes with leptons, that can be wrongly attributed to a Higgs boson candidate, and featuring jets that can either pass the b jet or hadronic tau lepton selections, are backgrounds for the considered signal.

The signal selection described in this chapter is defined with the aim of rejecting a large fraction of the background contamination; at the same time, as the signal is rarer than the background processes by several orders of magnitude, the selection is the result of a compromise to keep an acceptable coverage of the signal region. Tight object identification and isolation criteria are a useful handle to reject reducible backgrounds. On the other hand, the discrimination from irreducible backgrounds can only rely on the kinematic features of the processes; a multivariate analysis (MVA) technique exploits these properties. Although the background processes and modelling will be detailed in Ch. 5, their general description is presented below.

The major irreducible background, with an inclusive cross section of 832.71 pb, is the $t\bar{t}$ production: each of the top quarks has a large probability ($|V_{tb}| = 1.010 \pm 0.025$ [7]) of decaying in a b quark and a W boson, which in turn can decay in a lepton and a neutrino, with $BR(\tau, \nu_\tau) = BR(\mu, \nu_\mu) = BR(e, \nu_e) = 10.8$. Its contamination is particularly important in the $\tau_\mu\tau_h$ and $\tau_e\tau_h$ final states, as it contributes predominantly both through the decay W→ $\ell + \nu_\ell$ ($\ell = e, \mu$) and through the mediation of a tau lepton in W→ $\tau + \nu_\tau$ that decays in a muon or electron.

Generic QCD multijet events constitute a large reducible background for the $\tau_h\tau_h$ channel when two jets pass the τ_h identification selection, described in Sec. 4.3.3; its contribution is small in the $\tau_\mu\tau_h$ and $\tau_e\tau_h$ channels, as it is less likely for muon or electrons within jets to pass the corresponding object selection (Sec. 4.3.2 and 4.3.1).

Events with pairs of leptons produced through Drell-Yan processes $Z/\gamma^* \to \ell\ell$ ($\sigma = 6225.42$ pb) in association to jets are a large source of background; a small fraction of these events, where the pair of leptons is produced in association with two b jets, constitute an irreducible background.

Finally, W+jets events, with one genuine lepton from the W decay and one from the misidentification of a jet, have a minor contribution, further suppressed by the b tag requirements; like the $t\bar{t}$ and Drell-Yan events, they can enter the $\tau_\mu\tau_h$ and $\tau_e\tau_h$ selections also through the mediation of a tau lepton; hence, their contribution is larger in those channels. Other minor backgrounds, denoted as "Others" in the plots shown in the following, consist of single Higgs boson production through Standard Model processes,

single top production and events where two or three vector bosons are produced.

4.2 Trigger requirements

The trigger selection for the HH $\rightarrow$ bb$\tau\tau$ events targets the H $\rightarrow \tau\tau$ decay and no selection related to the H_{bb} system is required at this stage. To maximise the efficiency, it is convenient to use several different HLT paths. For a given final state resulting from the $\tau\tau$ decay, it is sufficient that one of the relevant HLT paths selects the event; technically, the trigger selection it is implemented with a logic OR. Scale factors (SF) that account for the different trigger selection efficiency on data and on simulated events are applied consistently.

The choice of the trigger paths aims at the largest possible signal acceptance coverage. Indeed, the selection used at HLT drives the offline thresholds: as clarified later in the description of the H $\rightarrow \tau\tau$ pair selection, the objects corresponding to those required by the trigger selection must pass the same p_T threshold used online with an additional margin that is chosen based on the HLT vs. offline resolution (2 GeV for muons, 3 GeV for electrons, 5 GeV for hadronic taus).

Three types of HLT sequences targeting the final states are used: the double-τ_h triggers, i.e. those where the presence of two hadronic tau leptons is required; the cross triggers, i.e. those where only one hadronic tau is required, accompanied by a muon or an electron; and the single-lepton triggers, i.e. those where only a muon or an electron is required.

The τ_h-legs in the double-τ_h paths are reconstructed using the classic HLT sequence described in Sec. 3.5.1: the L1 tau trigger output rate is too high to run directly the sophisticated and resource-expensive L3 step of the HLT reconstruction; therefore, the sequence needs to go through the L2 and L2.5 intermediate steps and to be regionally centred around a L1 tau candidate. In the case of the cross triggers, based on L1 seeds that also require the presence of a muon or electron and that have HLT selections for these objects, the tau reconstruction sequence can move directly to the L3 step; also, their HLT tau reconstruction uses the full detector acceptance ("global" reconstruction).

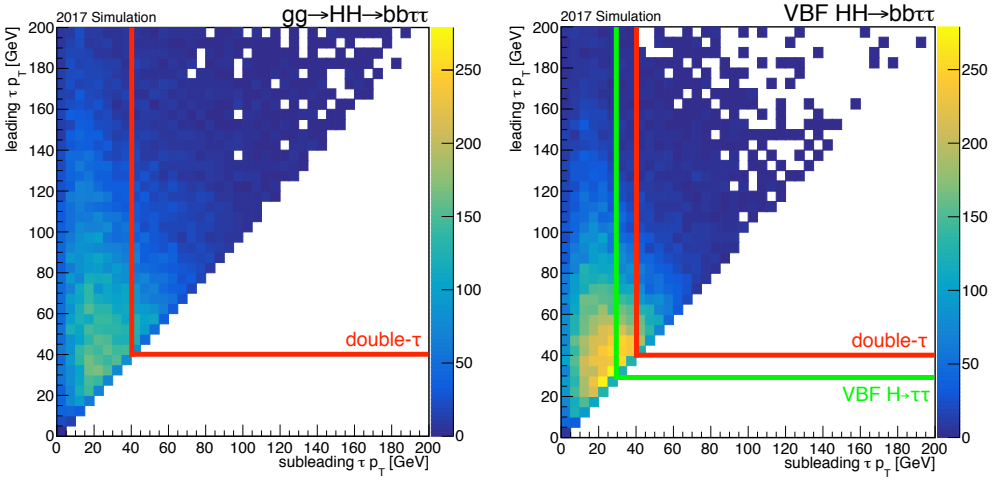
4.2.1 Triggers for the $\tau_h\tau_h$ final state

As anticipated in Ch. 3, the $\tau_h\tau_h$ channel selection mainly relies on double-τ_h triggers. The trigger requirements of the paths used in the 2017 H $\rightarrow \tau\tau$ search and in this analysis are summarised in Tab. 4.1. Hadronic tau leptons are only selected in $|\eta| < 2.1$ and need to fulfil different isolation and identification working points. Among the three paths, the lowest available p_T threshold is 35 GeV. Hence, the events in the $\tau_h\tau_h$ channel are required to have at least two tau candidates, geometrically matched with $\Delta R < 0.5$ to taus reconstructed online, with $p_T > 40$ GeV and $|\eta| < 2.1$.

In Fig. 4.2, the distributions of gg$\rightarrow$HH $\rightarrow$ bb$\tau\tau$ and VBF HH $\rightarrow$ bb$\tau\tau$ simulated events in the $\tau_h\tau_h$ final state are shown as a function of the p_T of the leading and subleading signal taus, i.e. those coming from the Higgs boson decay, as computed in the Monte Carlo generation. The p_T threshold used for hadronic tau leptons reconstructed offline is highlighted in red: the fraction of HH $\rightarrow$ bb$\tau\tau$ events produced through gluon fusion where both tau leptons have $p_T > 40$ GeV and $|\eta| < 2.1$ at generated level is 25%, whereas in the case of VBF production the tau leptons tend to have lower p_T and the fraction of events passing the selection is only 14%. In order to increase the acceptance on the VBF signal, the dedicated VBF H $\rightarrow \tau\tau$ trigger largely described in this thesis is used, when available, in addition to the double-τ_h paths: it selects events where hadronic

Table 4.1 – Trigger selections used in the $\tau_h\tau_h$ channel and corresponding integrated luminosity. The VBF H → $\tau_h\tau_h$ is only used to collect events in the corresponding VBF category.

double-τ_h triggers		
Kinematic selection	τ_h isolation and ID	Int. lumi [fb^{-1}]
Two τ_h, $p_T > 35$ GeV, $\|\eta\| < 2.1$	Tight, Tight	41.6
Two τ_h, $p_T > 40$ GeV, $\|\eta\| < 2.1$	Medium, Tight	41.6
Two τ_h, $p_T > 40$ GeV, $\|\eta\| < 2.1$	Tight, none	41.6
VBF H → $\tau_h\tau_h$ trigger		
Kinematic selection	τ_h isolation and ID	Int. lumi [fb^{-1}]
Two τ_h, $p_T > 20$ GeV, $\|\eta\| < 2.1$, two jets, $p_{T1} > 115$ GeV and $p_{T1} > 40$ GeV, $m_{jj} > 620$ GeV	Loose, none	27.1



(a) (b)

Figure 4.2 – Distribution of events from a gg→ HH → bb$\tau\tau$ (left) and a VBF HH → bb$\tau\tau$ signal sample as a function of the p_T of the leading and subleading hadronic tau originated by a Higgs boson, retrieved from the generator event information. The offline thresholds allowed by the use of the double-τ_h triggers are indicated in red, while those corresponding to the VBF H → $\tau\tau$ trigger are indicated in green. The number of events, expressed on the z-axis, is given in arbitrary units.

tau leptons are reconstructed online with $p_T > 20$ GeV, allowing the acceptance on the signal to be extended. The offline threshold of $p_T > 25$ GeV, driven by the VBF H → $\tau\tau$ trigger selection, is indicated in green in Fig. 4.2b. The VBF jet candidates topology cannot be appreciated in this view; however, the estimation of the event yield gain using full analysis-like selections described in Sec. 3.6 show that a non negligible gain comes from this trigger strategy.

The VBF category design and optimisation for the analysis will be described later in this chapter. However, some considerations can be made about the corresponding trigger selection: a relevant matter is to establish whether the categorisation should depend on the trigger or not. The HLT VBF H → $\tau\tau$ path is only available for the H → $\tau_h\tau_h$ final state. A natural choice would be to design a VBF category only based on offline quantities, using events collected with the logic OR of the double-τ_h and VBF H → $\tau\tau$ triggers in the $\tau_h\tau_h$ channel and the regular triggers in the $\tau_e\tau_h$ and $\tau_\mu\tau_h$ channels. This

choice has the advantage of being simple and consistent over the three $\tau\tau$ considered final states; however, anomalous structures would arise in the tau p_{T} and m_{jj} distributions, similar to those shown in Fig. 3.36. Moreover, the HLT VBF H $\to \tau\tau$ path was only implemented starting from August 2017: the composition of the VBF category population, thus, would change over time. This adds up to the practical difficulty of the trigger scale factors implementation: it is not trivial to capture the different kinematics and the correlation between different objects in the computation of the efficiencies of the logic OR of double-τ_{h} and VBF H $\to \tau\tau$ triggers. Therefore, an approach based on the trigger selection is chosen. In the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ final state, two VBF categories are considered: a loose VBF category, designed with loose VBF jet candidates offline selection and populated by events firing the double-τ_{h} path; and a tight VBF category, populated by events that only passed the VBF H $\to \tau\tau$ trigger, which drives the offline VBF jet candidates selection. In the former case, the double-τ_{h} trigger selection efficiency scale factors are applied; in the latter, the corrections presented in Sec. 3.5.1 are used.

4.2.2 Triggers for the semi-leptonic final states

Cross $\ell\tau_{\mathrm{h}}$ triggers are used in the corresponding semi-leptonic channels together with single-lepton triggers. The HLT paths used for the $\tau_{\mathrm{e}}\tau_{\mathrm{h}}$ and $\tau_{\mu}\tau_{\mathrm{h}}$ event selection are listed in Tab. 4.2 and Tab. 4.3.

Table 4.2 – Trigger selections used in the $\mu\tau_{\mathrm{h}}$ channel and corresponding integrated luminosity.

single-μ triggers		
Kinematic selection		Int. lumi [fb^{-1}]
One isolated μ, $p_{\mathrm{T}} > 24\,\mathrm{GeV}$		3.6
One isolated μ, $p_{\mathrm{T}} > 27\,\mathrm{GeV}$		41.6
cross $\mu\tau_{\mathrm{h}}$ trigger		
Kinematic selection	τ_{h} isolation	Int. lumi [fb^{-1}]
One τ_{h}, $p_{\mathrm{T}} > 27\,\mathrm{GeV}$, $\lvert\eta\rvert < 2.1$, one isolated μ, $p_{\mathrm{T}} > 20\,\mathrm{GeV}$, $\lvert\eta\rvert < 2.1$	Loose	41.6

Table 4.3 – Trigger selections used in the $\mathrm{e}\tau_{\mathrm{h}}$ channel and corresponding integrated luminosity.

single-e triggers		
Kinematic selection		Int. lumi [fb^{-1}]
One isolated e, $p_{\mathrm{T}} > 32\,\mathrm{GeV}$, $\lvert\eta\rvert < 2.1$		41.6
One isolated e, $p_{\mathrm{T}} > 35\,\mathrm{GeV}$, $\lvert\eta\rvert < 2.1$		41.6
cross $\mathrm{e}\tau_{\mathrm{h}}$ trigger		
Kinematic selection	τ_{h} isolation	Int. lumi [fb^{-1}]
One τ_{h}, $p_{\mathrm{T}} > 30\,\mathrm{GeV}$, $\lvert\eta\rvert < 2.1$, one isolated e, $p_{\mathrm{T}} > 24\,\mathrm{GeV}$, $\lvert\eta\rvert < 2.1$	Loose	41.6

The single-muon triggers, used in the $\tau_{\mu}\tau_{\mathrm{h}}$, select events with a L1 muon candidate and reconstruct a HLT candidate first using only the muon system information, then propagating the trajectories inwards to the tracker subdetectors. The muon isolation is then evaluated considering the additional tracks and the calorimeter energy deposits in a $\Delta R = 0.3$ cone around the HLT muon candidate. The loosest single-muon trigger used

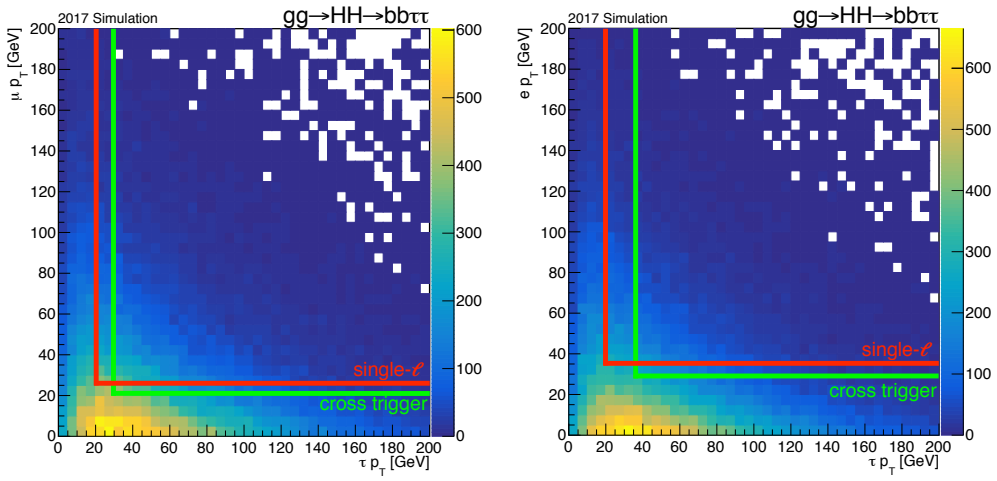


Figure 4.3 – Distribution of events from a gg→ HH → bb$\tau\tau$ signal sample as a function of the p_{T} of the muon (left) or electron (right) and the hadronic tau originated by an Higgs boson, retrieved from the generator event information. The offline thresholds allowed by the use of the single-lepton triggers are indicated in red, while those corresponding to the cross triggers are indicated in green. The number of events, expressed on the z-axis, is given in arbitrary units.

in this analysis, with $p_{\mathrm{T}} > 24\,\mathrm{GeV}$, was not enabled during the full 2017 data taking; it is accounted for by normalising the simulated events in the p_{T} region only accessible by this path to the integrated luminosity that it covered.

In the single-electron triggers, the electron candidate is reconstructed as energy deposits in ECAL grouped in clusters around the L1 electron and matched with a track; an isolation criterion is applied using the particle flow clusters in ECAL, in HCAL and in the tracker around the electron candidate. The loosest available p_{T} threshold is 35 GeV. The corresponding path was only enabled for about half of the data taking; however, the presence of a similar monitoring trigger allowed the path to be emulated in data with minor manipulations.

The use of cross triggers in addition to the single-lepton paths, with combined requirements on the hadronic tau candidate and on the lepton, allows larger coverage to be achieved. On one hand, the single-lepton triggers do not set an explicit threshold on the tau candidate; thus, the tau offline selection can be as loose as the offline reconstruction algorithms permit. On the other hand, by giving up some of the low tau p_{T} region already covered by the events firing the single-lepton paths, the lepton p_{T} threshold is lowered and a new region is accessed. The offline thresholds corresponding to single-lepton triggers are marked in red in Fig. 4.3, where the distribution of the events in a simulated gg→ HH → bb$\tau\tau$ is represented as a function of the p_{T} of the muon or electron and of the hadronic tau at generator level. It can also be observed that, due to the presence of two neutrinos in the decay cascade, the p_{T} spectra of the electron and muon are particularly soft. The thresholds corresponding to the cross triggers are indicated in green on the same figures. In the $\tau_\mu\tau_{\mathrm{h}}$ channel, the 30% of the signal events can be collected by using only the single-muon trigger and the 26% by using the cross $\mu\tau_{\mathrm{h}}$ trigger only, while the region covered by the OR of these paths gives a 34% coverage. Similarly, the single-electron trigger allows 23% of the $\tau_{\mathrm{e}}\tau_{\mathrm{h}}$ to be collected and the cross eτ_{h} trigger alone covers the 21% of the signal region; in this case, their combined use gives a 28% coverage.

4.2.3 Trigger efficiency

In the semi-leptonic channels, the trigger efficiency scale factors are combined to take into account the use of single-lepton and cross triggers. Assuming that the efficiency on each of the two cross trigger legs is uncorrelated, the efficiency of the OR of the single-lepton and cross trigger paths can be factorised and easily computed from the efficiencies on each object as

$$\epsilon = \epsilon_L(1 - \epsilon_\tau) + \epsilon_l\epsilon_\tau \tag{4.1}$$

where ϵ_L is the single-lepton trigger efficiency; ϵ_l is the cross trigger efficiency for the electron or muon leg; and ϵ_τ is the cross trigger efficiency for the τ_h leg. The non-relevant efficiencies are cancelled when the event fired only the single-lepton or only the cross trigger. The trigger scale factor is, thus, computed event by event as

$$\mathrm{SF} = \frac{\epsilon_{data}}{\epsilon_\mathrm{MC}}$$

where ϵ_{data} and ϵ_MC are computed using the Eq. 4.1 in data and simulated events.

The trigger selection efficiency for the ℓ-legs in the single-lepton triggers and in the cross triggers were computed within the $\mathrm{H} \to \tau\tau$ analysis. The μ-leg trigger efficiency is computed in $\mathrm{Z} \to \mu\mu$ data and in a simulated sample using a tag-and-probe technique, where a muon selected by a tight muon trigger requirement represents the tag and another muon, the probe, selected using the $\mathrm{Z} \to \mu\mu$ kinematic properties, is used to measure the efficiency of firing the logic OR of the HLT single muon paths under study. The efficiencies are given as a function of the p_T and η of the muon. A similar strategy is used in $\mathrm{Z} \to \mathrm{ee}$ events for the measurement of the efficiency of the single-e trigger. The trigger efficiency on leptons is shown in Fig. 4.4. In general, the efficiency on simulated events is larger than that on data: a large difference is observed, for instance, in Fig. 4.4a, where the highest efficiency among the considered η bins (red) is about 0.95 for Monte Carlo and 0.80 for the corresponding data curve. In all the measurements, the efficiency plateau appears lower at larger η. For example, in Fig. 4.4a, the efficiency on simulated events is 0.95 for muons in $|\eta| < 0.9$ and 0.78 for muons with $|\eta| > 2.1$. The discrepancies between data and simulations are due to the imperfect simulation modelling.

The trigger selection efficiency for each τ_h-leg in the double-τ_h and cross $\ell\tau_\mathrm{h}$ triggers is provided by the CMS tau trigger team. The measurements are preformed selecting $\mathrm{Z} \to \tau_\mu\tau_\mathrm{h}$ events in data, using events collected with a single muon trigger requirement, and in a $\mathrm{Z} \to \tau\tau$ simulated sample.

Three monitoring $\mu\tau_\mathrm{h}$ HLT paths are available for the double-τ_h efficiency measurement: unlike the $\mu\tau_\mathrm{h}$ cross triggers used in the analysis, they implement the full tau reconstruction sequence used by the double-τ_h triggers, with the only difference of being global and not regional; this difference was considered negligible in the context of this efficiency computation. Each of the monitoring $\mu\tau_\mathrm{h}$ paths implements a τ_h-leg selection that corresponds to those of the HLT paths listed in Tab. 4.1. The efficiency of each τ_h-leg corresponding to the logic OR of the double-τ_h paths is computed using a tag-and-probe approach; it is defined as the fraction of events, among those that pass the single muon trigger and the corresponding muon selection (tag), that fire any of the three monitoring HLT paths and the corresponding tau offline selection (probe). A similar strategy is used for the τ_h-leg efficiency computation for the cross triggers used in the analysis: in this case, the regular $\mu\tau_\mathrm{h}$ path is used for the probe selection.

This measurement is meant to give scale factors that are accurate for taus with different features; therefore, the efficiency is computed as a function of the offline tau p_T of the for each of the MVA tau identification working points and for each of the three hadronic

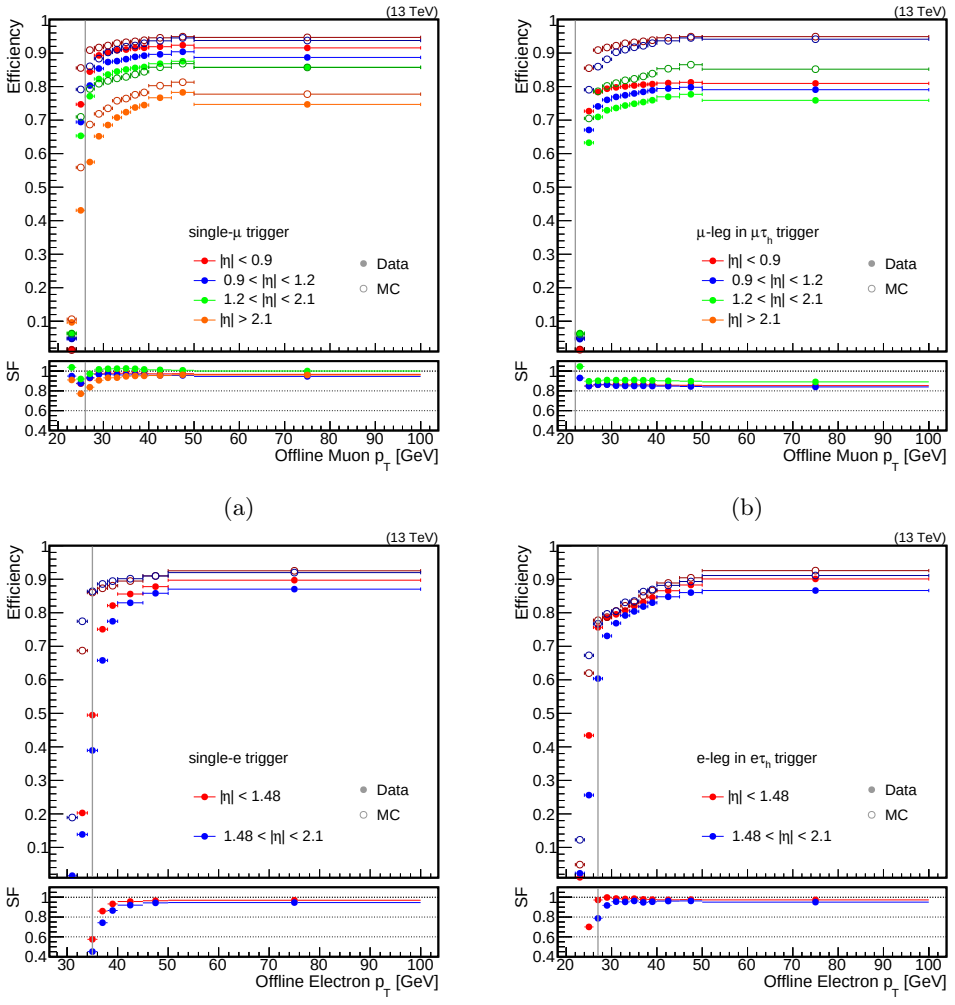


Figure 4.4 – Trigger efficiencies on ℓ-legs of the single-ℓ triggers and of cross triggers: in the top row, muon efficiencies as a function of the offline muon p_{T} in the single-μ trigger (a) and in the cross $\mu\tau_{\mathrm{h}}$ trigger (b); in the bottom row, electron efficiencies as a function of the offline electron p_{T} in the single-e trigger (c) and in the cross eτ_{h} trigger (d). The measurements were performed within the H $\to \tau\tau$ analysis. The grey lines represent the thresholds set in this analysis for the object selection. The errors along the y axis are too small to be appreciated.

tau decay modes identified with the current tau offline reconstruction algorithm. The efficiencies corresponding to the tau selection used in this search, requiring the medium tau identification working point, is shown for all the τ_{h}-leg types and for each hadronic tau decay mode in Fig. 4.5, Fig. 4.6 and Fig. 4.7. In the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel, the scale factor is simply obtained as the ratio between the efficiency on data and on simulated events. Moreover, η and ϕ-dependent corrections are considered in the final scale factors computation.

The impact of the decay mode is not negligible: for instance, the selection efficiency on tau leptons reconstructed as $\mathrm{h}^{\pm}\mathrm{h}^{\mp}\mathrm{h}^{\pm}$ turns on slowly compared to the other decay modes, but its plateau for simulated events is higher. A comparison of the data over simulation efficiency scale factors for each decay mode is shown in Fig. 4.8 as a function of the tau p_{T} for τ_{h}-legs in each trigger typology. The double-τ_{h} legs show a pronounced decay mode dependency: the low p_{T} region, below 40 GeV, is not interesting for the analysis as it is not covered by the event selection; over 40 GeV, the $\mathrm{h}^{\pm}\mathrm{h}^{\mp}\mathrm{h}^{\pm}$ decay mode scale factor is rather flat over the considered range, while those corresponding to

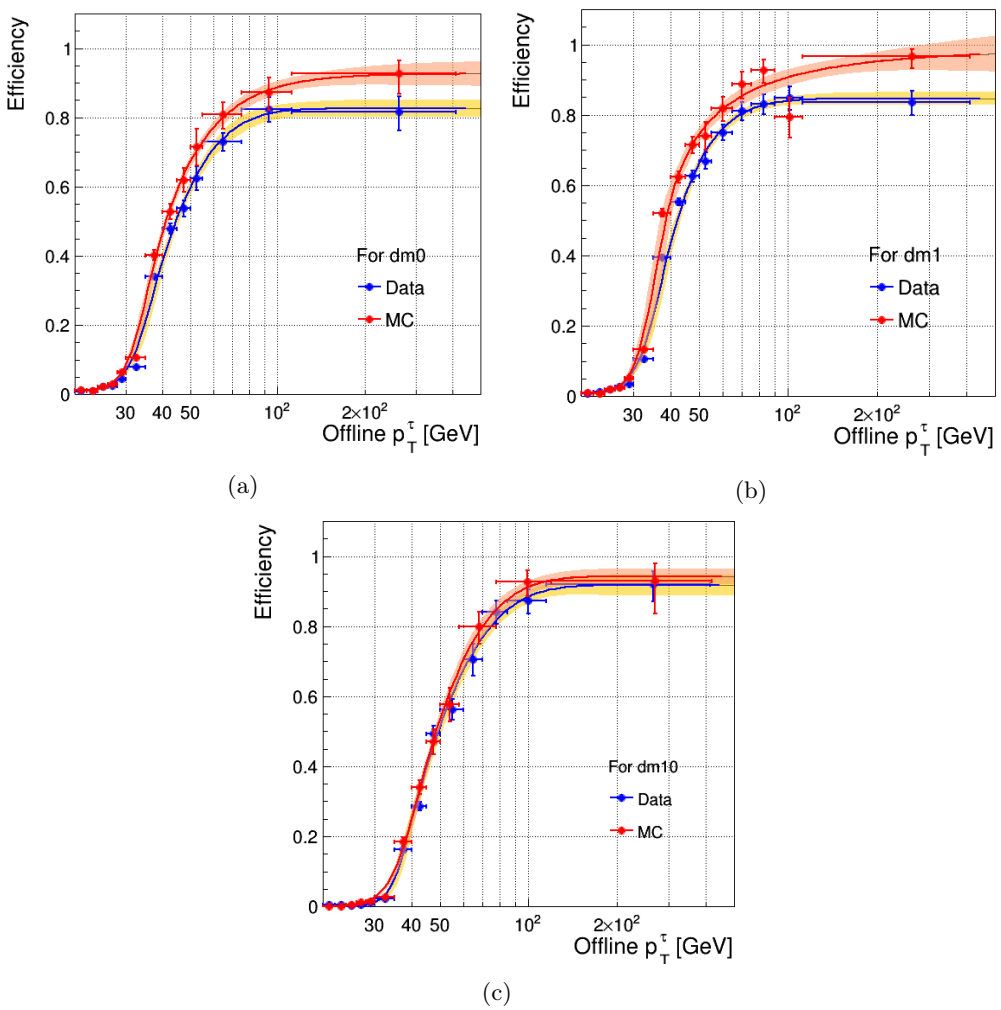


Figure 4.5 – Trigger selection efficiency curves, corresponding to taus reconstructed offline with medium identification working point, on data and on simulated Z$\rightarrow \tau_\mu\tau_h$ events for each τ_h-leg in the logic OR of the three double-τ_h trigger paths used in the $\tau_h\tau_h$ final state, as a function of the offline tau p_T. The efficiencies are computed separately for each hadronic tau decay mode: $h^\pm$ decay mode (a); $h^\pm\pi^0$ decay mode (b); and $h^\pm h^\mp h^\pm$ decay mode (c).

the $h^\pm$ and $h^\pm\pi^0$ decay modes range between 0.8 and 0.95. For the τ_h-legs in the cross triggers, instead, a good agreement among the decay modes is observed in the p_T region between 30 and 70 GeV, while the $h^\pm h^\mp h^\pm$ decay has higher scale factors than the other decay modes at higher p_T. The decay mode dependency on the tau selection efficiency was found to be relevant also for the tau isolation and identification in the $\tau_h\tau_h$ channel, as discussed in Appendix A.

An efficiency loss in the data collection due to a so-called "L1 trigger prefiring" was observed in 2017. The ECAL pulse reconstruction timing had a gradual shift starting from 2016 until 2017; as a consequence, L1 trigger primitives are misassinged by one bunch crossing with a rate that exceeds 20% for e/γ triggers in the late 2017 data taking. The L1 trigger system does not allow firing in consecutive bunch crossings; thus, if the objects firing a L1 trigger selection are misassinged by one bunch crossing too early, the bunch crossing that actually contains that object is missed and, most likely, the event gets discarded by the HLT. This effect is not described in the simulation and it does not appear in offline vs. online turn-on efficiencies.

The impact of the prefiring issue was found to be larger for objects in the forward regions

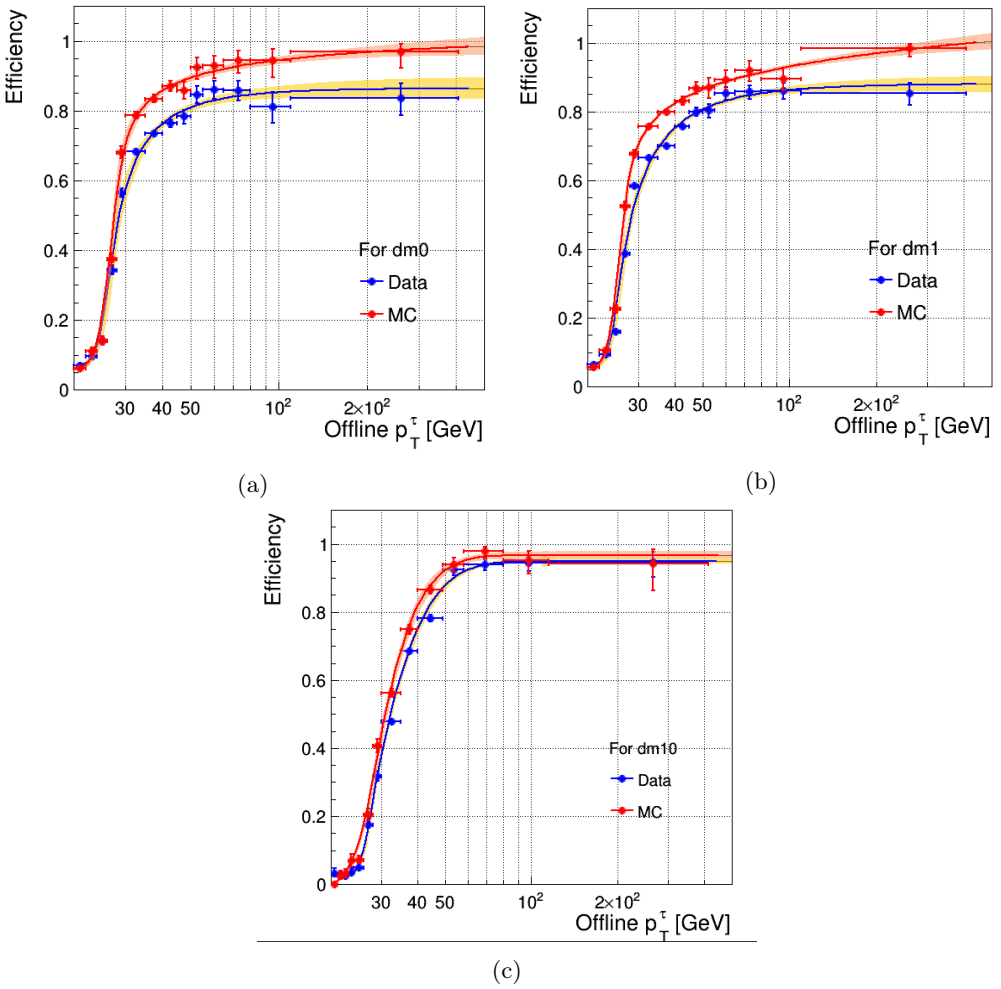


Figure 4.6 – Trigger selection efficiency curves, corresponding to taus reconstructed offline with medium identification working point, on data and on simulated Z→ $\tau_\mu\tau_h$ events for the τ_h-leg in the $\mu\tau_h$ trigger path, as a function of the offline tau p_T. The efficiencies are computed separately for each hadronic tau decay mode: $h^\pm$ decay mode (a); $h^\pm\pi^0$ decay mode (b); and $h^\pm h^\mp h^\pm$ decay mode (c).

of the detector. In this analysis, all the lepton and b jet candidates lie within the region with $|\eta| < 2.4$; therefore, the efficiency loss is considered small. An inefficiency can still affect the VBF jet candidates, which are reconstructed up to $|\eta| = 5$; however, as it can be seen in Fig. 4.21d, a large data-over-prediction disagreement going in the opposite direction is observed. In conclusion, no correction is implemented in this analysis to compensate for the trigger prefiring inefficiency.

4.3 H → ττ pair selection and categorization

The first step in the event selection is the choice H → ττ pair. First, a baseline selection is applied to select e, μ and τ_h candidates. At this stage, only minimal p_T and η thresholds are applied to ensure an efficient object reconstruction. Isolation and identification criteria are also applied. The τ_h, μ and e preselection requirements are detailed in Sec. 4.3.1, Sec. 4.3.2 and Sec. 4.3.3.

The ττ combinations are then examined and the best pair is chosen. The H → ττ pair selection is bound to the trigger selection, not only because the classification of the

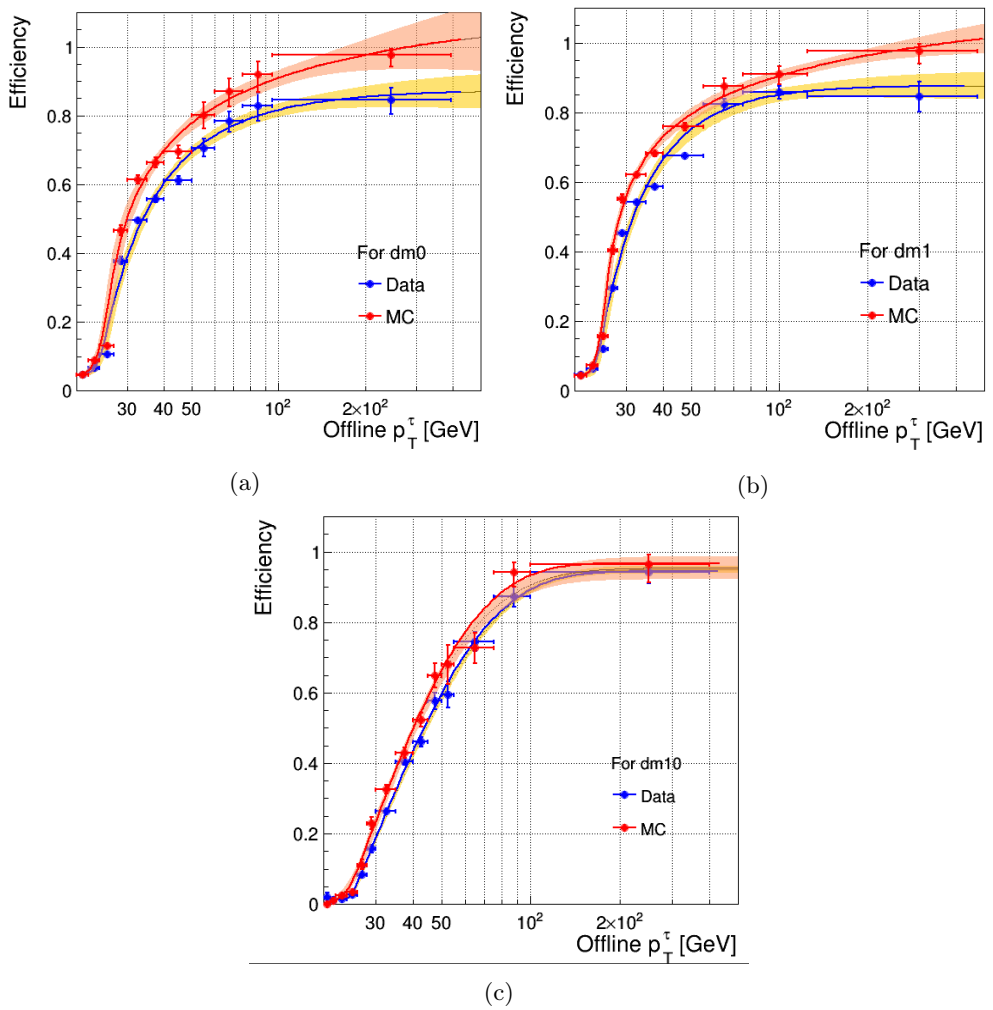


Figure 4.7 – Trigger selection efficiency curves, corresponding to taus reconstructed offline with medium identification working point, on data and on simulated Z→ $\tau_\mu\tau_h$ events for the τ_h-leg in the eτ_h trigger path, as a function of the offline tau p_T. The efficiencies are computed separately for each hadronic tau decay mode: $h^\pm$ decay mode (a); $h^\pm\pi^0$ decay mode (b); and $h^\pm h^\mp h^\pm$ decay mode (c).

events in the $\tau_e\tau_h$, $\tau_\mu\tau_h$ and $\tau_h\tau_h$ channels requires the corresponding triggers to be fired, but also because the trigger selection drives the offline selection of the $\tau\tau$ candidates: event by event, the offline lepton candidates need to be geometrically matched to a HLT trigger object of the same type and to pass a selection that is considered compatible with the trigger paths that are fired. The assessment of the best $\tau\tau$ pair and its subsequent selection are described in Sec. 4.3.5.

4.3.1 Electron preselection

Electrons are reconstructed through the standard CMS algorithm described in Sec. 2.3.3, combining the information from ECAL and from the tracker. A MVA approach is used to identify genuine prompt electrons: the main contributions to the background come from jets and from electrons originating from the photon conversion in e^+e^- pairs. The discriminator is based on a Boosted Decision Tree (BDT) that combines purely calorimetric variables, that are sensible to the shape of the shower and the amount of energy deposited in ECAL and HCAL, as electrons tend to generate narrow showers mostly contained in ECAL; observables that combine the information from the tracker and the

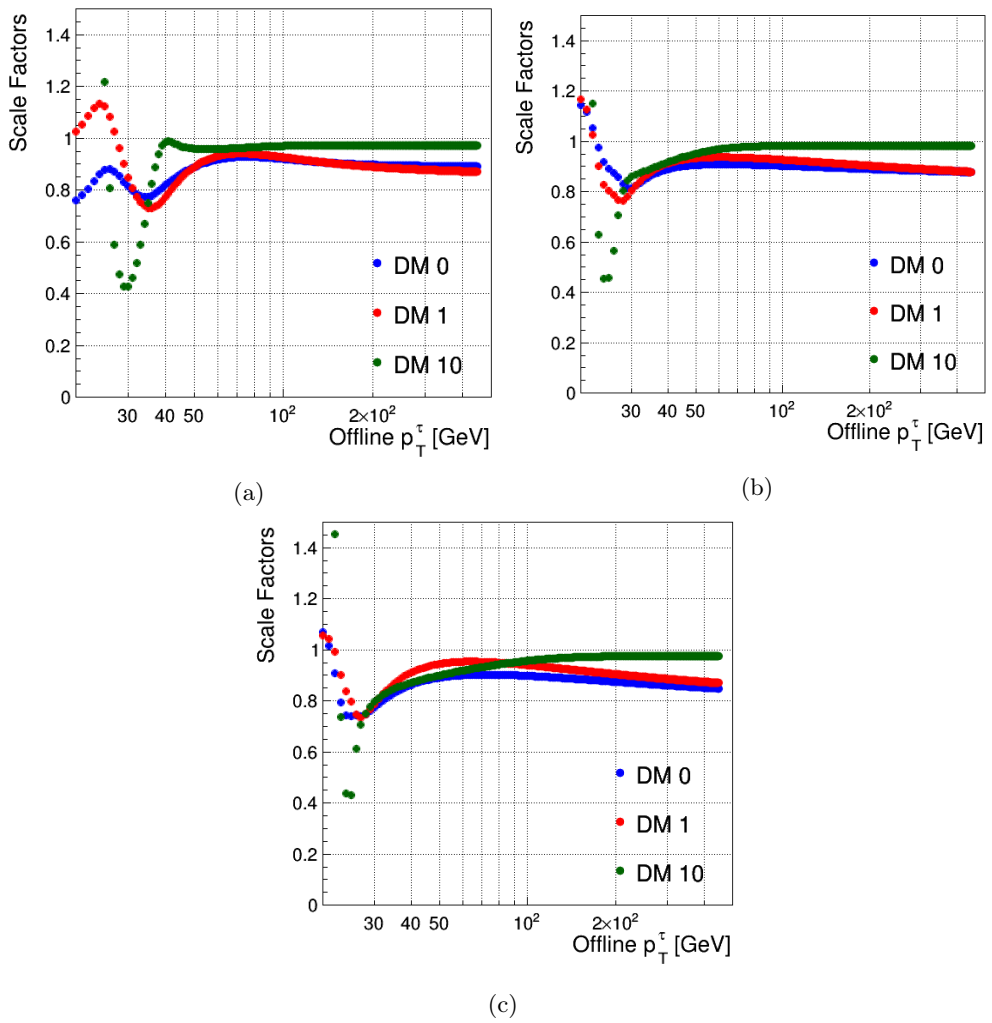


Figure 4.8 – Trigger scale factors for each τ_h-leg in the logic OR of the three double-τ_h trigger paths used in the $\tau_h\tau_h$ final state (a); for the τ_h-leg of the $\mu\tau_h$ trigger path (b); and for the τ_h-leg of the eτ_h trigger path (c). They are represented separately for each hadronic tau decay mode and as a function of the offline tau p_T: "DM0" denotes the $h^\pm$ decay; "DM1" denotes the $h^\pm\pi^0$ decay mode; and "DM10" denotes the $h^\pm h^\mp h^\pm$ decay. These scale factors correspond to the medium working point of the offline tau identification, used for the tau selection in this analysis.

calorimeters, as the geometrical and momentum matching between the candidate's reconstructed trajectory and the associated calorimeter clusters; and purely track-based observables as, for instance, the fraction of energy lost through Bremsstrahlung.

In addition to the use of the identification discriminator, isolation requirements are applied. A significant fraction of background to isolated signal electrons is composed by jets, either because they contain charged particles misidentified as electrons or because of genuine electrons within the jet resulting from semileptonic decays of b or c quarks. Requiring electron candidates to be isolated from nearby activity reduces significantly this source of background. The isolation is computed from the scalar sum of the transverse momenta of particles inside a cone of size $\Delta R = 0.3$ around the electron candidate, relative to its transverse momentum, as

$$I^{\ell}_{rel} = \left(\sum p_{\mathrm{T}}^{ch} + \max\left[0, \sum p_{\mathrm{T}}^{n} + \sum p_{\mathrm{T}}^{\gamma} - p_{\mathrm{T}}^{PU}\right]\right) \Big/ p_{\mathrm{T}}^{\ell} \,, \tag{4.2}$$

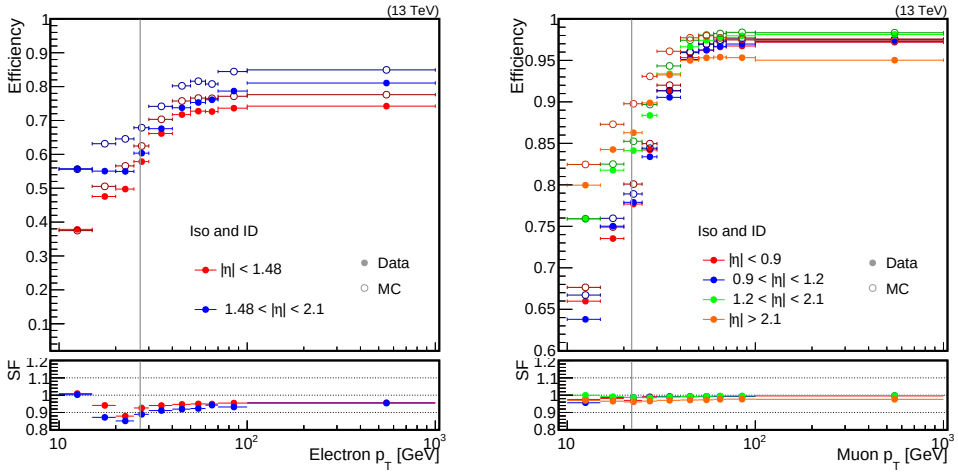
where $\ell = e, \mu$. Because the charged particles, accounted for in the first term of the denominator of Eq. 4.2, are required to originate from the primary vertex, they have

negligible dependence on the pileup. The second term, instead, corresponding to the contribution of photons and neutral particles, that cannot be associated to the primary vertex, is corrected by subtracting the estimated p_{T} associated to pileup. The pileup contribution is computed in Z→ $e^{+}e^{-}$ events assuming that $p_{\mathrm{T}}^{PU} = \rho A^{eff}$, where ρ is the energy density in the isolation cone and A^{eff} is the effective area of the cone for each component of the isolation (photons and neutral particles); the effective area is the area of the cone scaled by the ratio of the slopes for ρ and for each isolation component as a function of the number of reconstructed vertices.

A $I^{e}_{rel} < 0.10$ threshold is required for electrons to be selected in $\tau_{\mathrm{e}}\tau_{\mathrm{h}}$ pairs. The MVA identification discriminant is tuned for isolated electrons with $p_{\mathrm{T}} > 10\,\mathrm{GeV}$ in different eta regions. The *Tight* working point, identified by the EGamma Physics Object Group (POG) and corresponding to a 80% efficiency, is used in this analysis. The efficiency of the electron selection is shown, as a function of the electron p_{T} and in bins of η, in Fig. 4.9a. Overall, a higher efficiency is observed in the endcaps; as the selection efficiency on electrons in data appears always lower than for electrons reconstructed in simulated events by about 5%, the scale factors shown in the bottom plot of Fig. 4.9a are applied to each simulated electron to account for the different performance.

Finally, the electron candidate is required to be compatible with the primary vertex. The distance between the electron track and the primary vertex must fulfil $d_{xy} < 0.045\,\mathrm{cm}$ in the transverse plane and $d_{z} < 0.2\,\mathrm{cm}$ in the longitudinal plane.

The transverse momentum and pseudorapidity distributions of electrons selected in $\tau_{\mathrm{e}}\tau_{\mathrm{h}}$ events are shown in Fig. 4.10. The very first part of the distribution, with $p_{\mathrm{T}} < 35\,\mathrm{GeV}$, corresponds to events that can only fire the cross $e\tau_{\mathrm{h}}$ trigger; for $p_{\mathrm{T}} > 35\,\mathrm{GeV}$, the single-$e$ trigger is also enabled and the number of selected events increases. An excellent data-over-prediction agreement, within the level of the 10%, is achieved in all the regions.

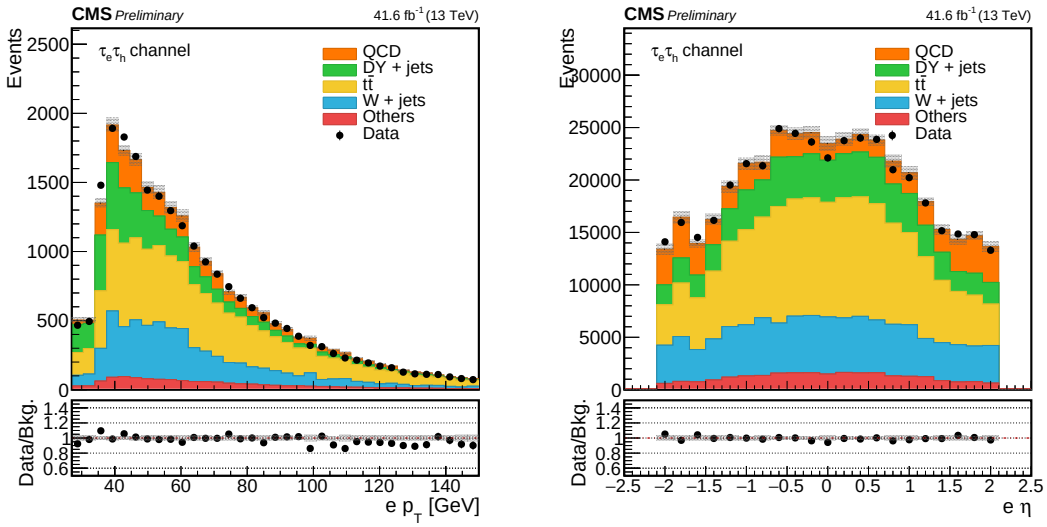


(a) (b)

Figure 4.9 – Efficiency of the combined selection of isolation and identification discriminant on electrons (a) and muons (b). The measurements were performed within the H → ττ analysis. The grey lines represent the thresholds set in this analysis for the object selection. The errors along the y axis are too small to be appreciated.

4.3.2 Muon preselection

Muons are reconstructed, as described in Sec. 2.3.2, by the tracker muon algorithm or the global muon reconstruction algorithm, or they are merged in a single candidate



(a) (b)

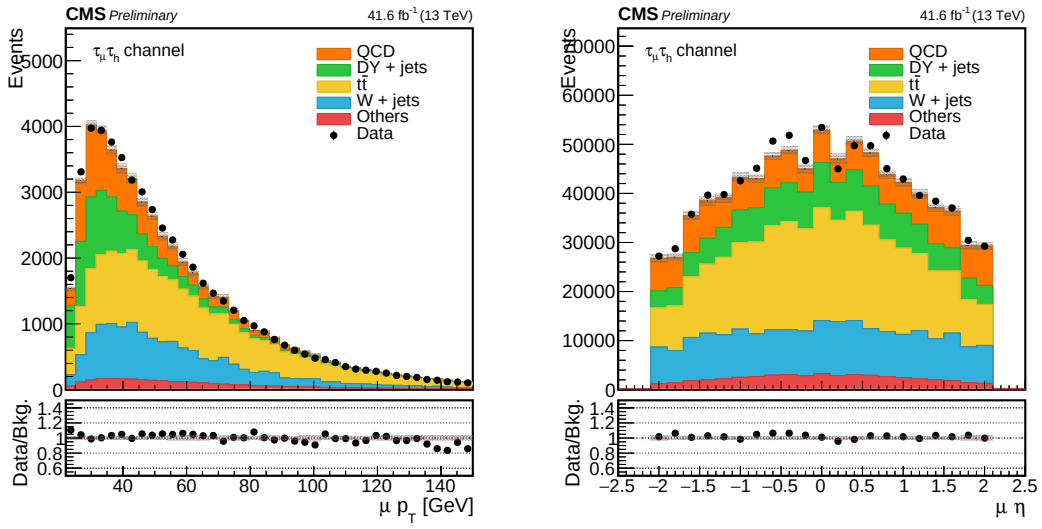
Figure 4.10 – Transverse momentum (left) and pseudorapidity (right) distributions of electrons selected in the $\tau_e\tau_h$ final state. The error bars along y for the data distribution and for the data-over-prediction ratio are too small to be appreciated.

if reconstructed with both algorithms. The latter condition is required by the *Tight* identification discriminant used to select signal muons in this search. The identification discriminant, aimed to suppress misidentification of charged hadrons escaping the calorimeters, entails additional criteria as the number of hits in the inner tracker used to reconstruct the considered muon track and the quality of the global muon fit. Like the selected electrons, the muon candidate must also be compatible with the primary vertex: its track is required have distance $d_{xy} < 0.045\,\text{cm}$ in the transverse plane and $d_z < 0.2\,\text{cm}$ in the longitudinal plane from the primary vertex. Similarly to electrons, in order to suppress the background contribution from weak decays within jets, the muons are required to be isolated. The surrounding activity is quantified through the computation of I^{μ}_{rel}, using the definition Eq. 4.2 with a cone of size $\Delta R < 0.4$ around the direction of the candidate. In this case, the pileup contribution to the neutral particles is estimated as the sum of charged hadron deposits originating from pileup interactions; this quantity is then scaled it by the estimate of the ratio of neutral particle to charged hadron production and subtracted from the neutral hadron and photon sums. The *Tight* isolation working point $I^{\mu}_{rel} < 0.15$, optimised targeting a 98% efficiency, is used to select signal muons. The efficiency of the combined efficiency of the muon isolation and identification is shown in Fig. 4.9b as a function of the muon p_{T} and in bins of η. A high efficiency, larger than 95% and similar for data and simulation in all the η regions, is achieved at plateau. The scale factors shown in the bottom plot of Fig. 4.9b, very close to 1 in all the considered range, are applied to each selected muon.

The resulting data-over-prediction agreement can be appreciated in Fig. 4.11. The distributions of muons selected in $\tau_\mu\tau_h$ events as a function of the transverse momentum and the pseudorapidity show an excellent agreement. Similarly to the electron p_{T} distribution, the increase of selected events for $p_{\text{T}} > 27\,\text{GeV}$ corresponds to the activation of the single-μ trigger.

4.3.3 Hadronic tau lepton preselection

The tau lepton decays into hadrons and a neutrino are reconstructed through the HPS identification algorithm described in Sec. 2.3.4. Only tau leptons reconstructed in the



(a) (b)

Figure 4.11 – Transverse momentum (left) and pseudorapidity (right) distributions of muons selected in the $\tau_\mu\tau_h$ final state. The error bars along y for the data distribution and for the data-over-prediction ratio are too small to be appreciated.

$h^\pm$, $h^\pm\pi^0$ and $h^\pm h^\mp h^\pm$ decay modes are selected.

The largest source of background contamination to the hadronic tau reconstruction originates from quark and gluon jets. The isolation of the candidates is one of the main handles to reject these backgrounds. It is computed from the sum of the transverse momenta of charged particles not identified as constituent of the hadronic tau decay. Within a cone of size $\Delta R = 0.5$ around the reconstructed tau lepton direction, the candidate isolation is

$$I_\tau = \sum_{d_Z<0.2\,\mathrm{cm}} p_\mathrm{T}^{ch} + \max\left(0, \sum p_\mathrm{T}^{\gamma} - \Delta\beta \sum_{d_Z>0.2\,\mathrm{cm}} p_\mathrm{T}^{ch}\right). \tag{4.3}$$

The smallest the value of I_τ, the more the candidate is considered isolated. The first term of Eq. 4.3 represents the isolation from charged particles; its computation is restricted to tracks originated within distance $d_Z < 0.2\,\mathrm{cm}$ along the longitudinal direction from the vertex of the hadronic tau production. Thus the pileup contribution to the isolation is suppressed. The second term represents the contribution from photons. In this case, the effect of pileup is computed as follows. The sum of the momenta sum of charged particles originating from pileup interactions is multiplied by the $\Delta\beta = 0.2$ factor representing the neutral to charged hadron production ratio in inelastic proton-proton collisions. The quantity thus obtained is then subtracted from the sum of the transverse momenta of the photons within the isolation cone.

In this analysis, rather than requiring the isolation to be lower than a fixed threshold, a tau identification discriminant determined through a MVA-based approach is used [110]. A BDT is trained using, in addition to the isolation defined in Eq. 4.3, the reconstructed τ_h decay mode, the main kinematic variables, and variables sensitive to the tau lepton lifetime. Several MVA discriminator working points, corresponding to different identification efficiency, are identified. Each working point is defined by thresholds on the BDT discriminant adjusted as a function of the hadronic tau lepton p_T, so that the resulting efficiency is uniform.

The *Medium* working point is chosen to define the signal region. However, events with

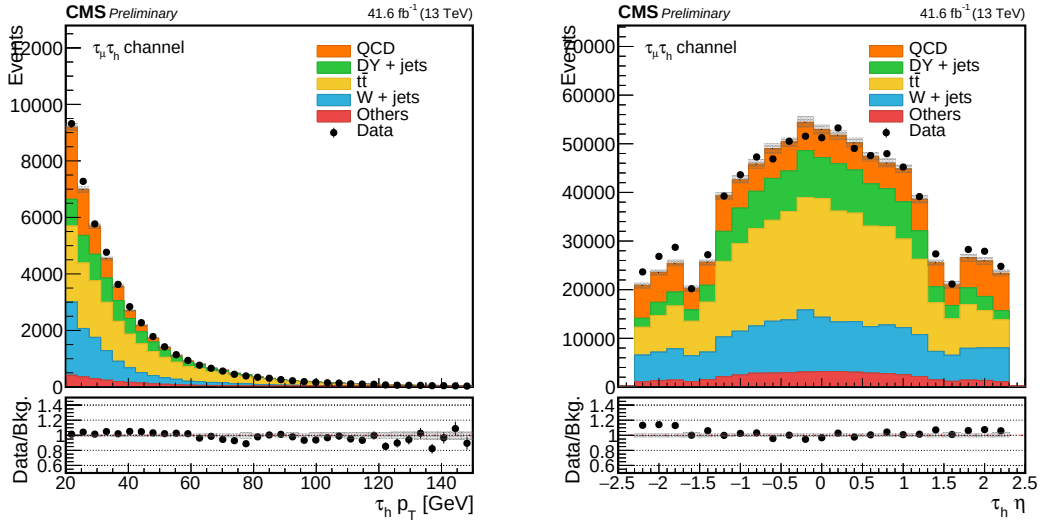
hadronic tau leptons that pass the *VLoose* identification working point and not the *Medium* are used in the analysis to populate the sidebands for the QCD background estimation.

Additional discriminators are applied to separate hadronic tau leptons from electrons and muons. The probability that an electron or a muon is misidentified as a charged product of the hadronic tau lepton decay is sizeable; furthermore, electrons can emit photons when crossing the detector material and, thus, fake the π^0 particles that can occur in the $h^{\pm}\pi^0$ decay mode. The probability of misidentification of a muon or electron into an hadronic tau lepton ($e/\mu \to \tau_h$), or "fake rate", is commonly computed through a tag-and-probe technique. For instance, the $e \to \tau_h$ fake rate is measured in $Z \to ee$ events with a well identified electron (tag) an electron reconstructed as an hadronic tau (probe). An MVA-based discriminator is used to reduce the $e \to \tau_h$ misidentification probability: the discriminator uses a BDT that takes as input variables the number of photons associated to the candidate together with the fraction of energy carried by them, the distance of the photons from the leading track of the candidate, and variables that are sensitive to the fraction of energy deposited in ECAL and HCAL. In this analysis, two anti-electron discriminant working point are used. In the $\tau_\mu\tau_h$ and $\tau_h\tau_h$ channels, the hadronic tau lepton candidate is required to pass the *Very loose* working point of the anti-electron discriminant, that reduces the probability of $e \to \tau_h$ misidentification to $8 - 5\%$; in the $\tau_e\tau_h$ channel, in order to reduce the $Z \to ee$ background, the τ_h candidate must pass the anti-ele *Tight* working point, giving a fake rate of about 0.2%. The anti-muon discriminant is based on the presence of signals in the muon system in the vicinity of the hadronic tau lepton direction. Two working points are provided: the *Loose* working point gives a fake rate at the per mille level, while the use of the *Tight* working point further reduces it of one order of magnitude. Similarly to the anti-electron discriminant, the anti-muon is used in this analysis with both working points: hadronic tau leptons in the $\tau_e\tau_h$ and $\tau_h\tau_h$ channels should pass the *Loose* discriminant; in the $\tau_\mu\tau_h$ channel, the *Tight* working point is applied.

The output of the discriminating algorithms, as well as the evaluation of their performances, are provided by the CMS Tau POG. Correction scale factors, accounting for the different performance on data and simulation, are computed as well and are applied in this analysis. The origin of each of the selected hadronic tau candidates in simulated events is assessed as the type of the closest generator level particle within a $\Delta R < 0.2$ cone: if the reconstructed hadronic tau lepton is associated to a generated tau decaying hadronically, it is considered genuine; if it is matched to a prompt electron or an electron originated from the leptonic tau decay, it is considered as a $e \to \tau_h$ misidentification; similarly, if it is matched to a muon, prompt or originated from a tau lepton, it is ascribed to $\mu \to \tau_h$ misidentification; in all other cases, it is considered as a jet-to-tau misidentification. Event-by-event, a scale factor is applied based on the outcome of the matching for each tau candidate: the scale factor corresponding to the tau identification discriminant is applied to all genuine hadronic tau leptons, while those for the anti-e and anti-mu discriminants are applied to hadronic tau leptons faked by electrons and muons.

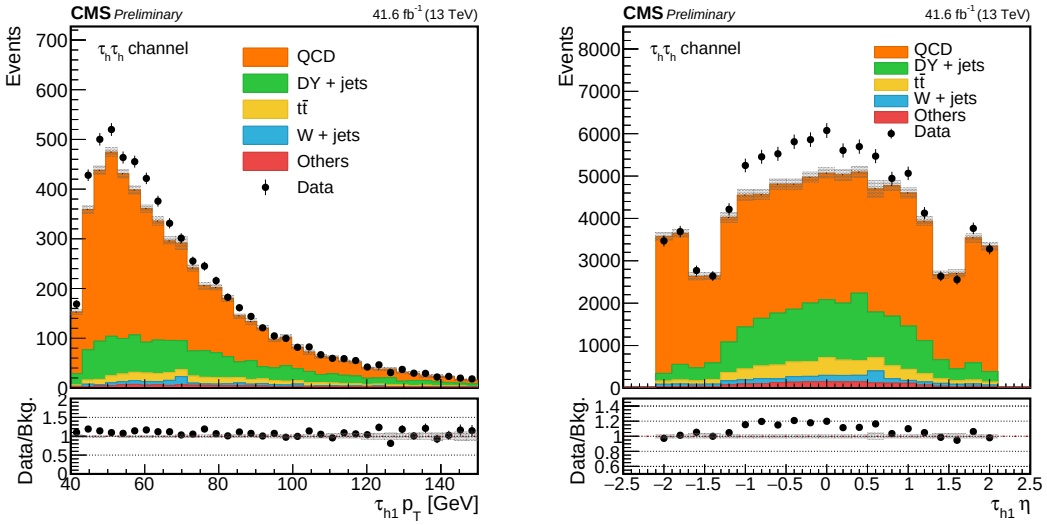
All the identification scale factors for hadronic tau leptons discriminators are inclusive with respect to the reconstructed tau decay mode, although several studies have shown a non-negligible decay mode dependence. A 0.89 scale factor is applied each genuine hadronic tau lepton in simulated $\tau_e\tau_h$ and $\tau_\mu\tau_h$ events, following the recommendation given by the Tau POG. The resulting distributions show a satisfactory data-over-prediction agreement, as shown in Fig. 4.12 for hadronic tau leptons in the $\tau_\mu\tau_h$ channel. However, these corrections turned out to be not suited for the $\tau_h\tau_h$ channel, leading to a large disagreement observed in several event distributions and most accentuated in

regions populated by genuine hadronic tau leptons. The poor background modelling can be appreciated in Fig. 4.13, where the p_T and η distributions of the leading hadronic tau lepton in $\tau_h\tau_h$ events are represented. Therefore, an alternate set of scale factors, also accounting for the tau identification efficiency, was computed within this analysis and is applied in the $\tau_h\tau_h$ channel. The resulting data and background distributions, with satisfactory agreement, are shown in Fig. 4.14 for the leading τ_h in $\tau_h\tau_h$ events. Together with an extensive analysis of the possible origins of the observed disagreement, the necessity of specific corrections is discussed in Appendix A.



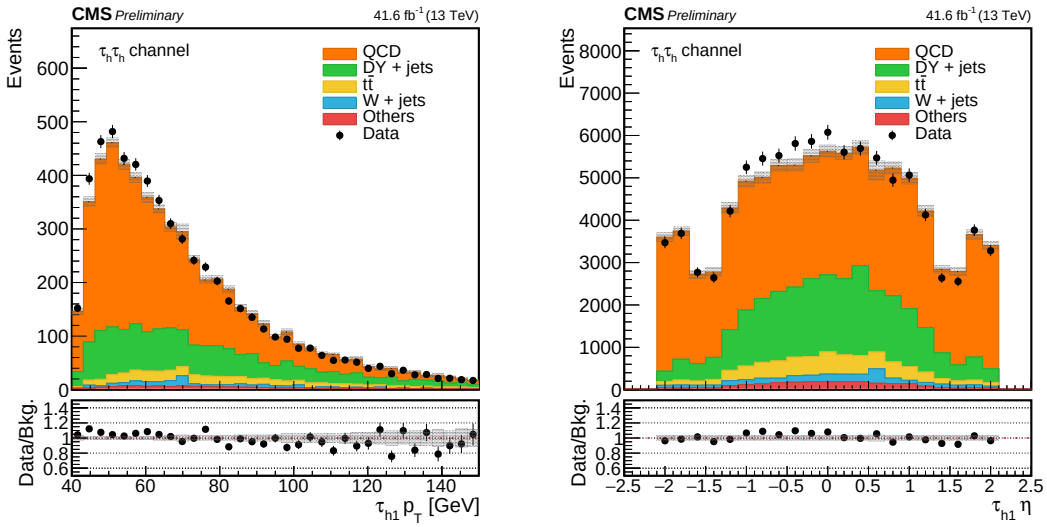
(a) (b)

Figure 4.12 – Transverse momentum (left) and pseudorapidity (right) distributions of hadronic tau leptons selected in the $\tau_\mu\tau_h$ final state. The error bars along y for the data distribution and for the data-over-prediction ratio are too small to be appreciated.



(a) (b)

Figure 4.13 – Transverse momentum (left) and pseudorapidity (right) distributions of hadronic tau leptons selected in the $\tau_h\tau_h$ final state. All the recommended scale factors are applied event-by-event; to recover the disagreement, specific corrections are derived (see Appendix A).



(a) (b)

Figure 4.14 – Transverse momentum (left) and pseudorapidity (right) distributions of hadronic tau leptons selected in the $\tau_h\tau_h$ final state. Instead of the recommended tau identification scale factor, a correction computed within this analysis is applied event-by-event for each tau lepton (see Appendix A).

4.3.4 Missing transverse momentum

In events featuring tau leptons, missing transverse momentum arises in all the decay modes due to the production of neutrinos, which escape without interacting with the detector material. As detailed in Sec. 2.3.4, the reconstruction of the missing transverse momentum is performed using the vectorial sum of the transverse momenta of the objects reconstructed with the particle flow algorithm.

No requirement is set on the $p_{\mathrm{T}}^{\mathrm{miss}}$ at the object selection level; instead, the missing transverse momentum is one of the variables used by the MVA technique for the discrimination against irreducible backgrounds and by the algorithm for the $\mathrm{H}_{\tau\tau}$ candidate reconstruction. However, some filters are applied to reject events where a large $p_{\mathrm{T}}^{\mathrm{miss}}$, unphysical or uninteresting, is reconstructed. The corresponding algorithms use the timing, pulse shape and topology of the signals from the subdetectors to identify, for example, events where $p_{\mathrm{T}}^{\mathrm{miss}}$ arises from the reconstruction of particles that are produced in the interaction of protons from low density tails of the beam (or "beam halo"); events with artificial $p_{\mathrm{T}}^{\mathrm{miss}}$ due to dead cells in ECAL; events with anomalous signals produced by the HCAL read-out; and events with low quality muons that participate to the $p_{\mathrm{T}}^{\mathrm{miss}}$ computation as charged hadron candidates or as muons with inconsistent p_{T}. These effects are not modelled by the simulations; hence, the application of the filters guarantees a good agreement between the $p_{\mathrm{T}}^{\mathrm{miss}}$ distribution in data and background events.

Additional corrections turned out to be necessary to mitigate a large disagreement observed by several analyses in the tail of the $p_{\mathrm{T}}^{\mathrm{miss}}$ distributions with 2017 data. As discussed in Sec. 3.3, the data collected in 2017 are affected by an interplay between the ageing of ECAL crystals at large η, the increase of pileup due to the ordinary LHC operations, and the ineffective out-of-time-pileup mitigation due to the unforeseen LHC bunch scheme choice. These effects result in large noise in the forward regions of the detector, leading to an artificial imbalance of energy in the transverse plane in data and, subsequently, to a large disagreement at large $p_{\mathrm{T}}^{\mathrm{miss}}$ with respect to the background distribution. Consistently with the identified causes of this effect, the observed disagree-

ment evolves with time, becoming larger at the end of the data taking. Following the recommendations from the Jet-MET POG, a corrected version of the $p_{\mathrm{T}}^{\mathrm{miss}}$ was used in the analysis, where jets and unclustered particle flow candidates with $p_{\mathrm{T}} < 50\,\mathrm{GeV}$ in the region with $2.65 < |\eta| < 3.14$ are excluded from the computation.

The missing transverse momentum distribution in $\tau_\mu\tau_\mathrm{h}$ events is shown in Fig. 4.15. A good agreement is achieved over all the range, including the tails. A similar level of agreement is obtained in the other channels.

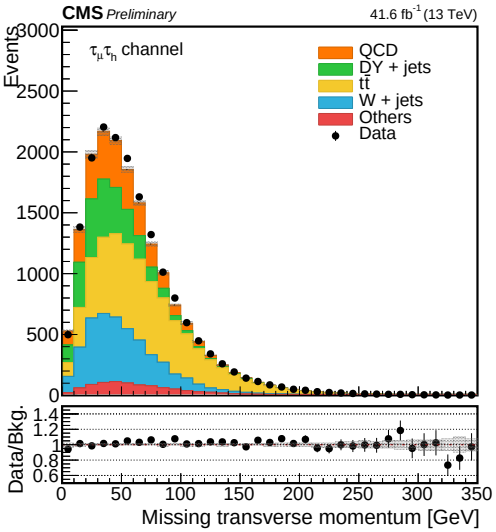


Figure 4.15 – Missing transverse momentum in $\tau_\mu\tau_\mathrm{h}$ events. The error bars along y for the data distribution and for the data-over-prediction ratio are too small to be appreciated.

4.3.5 Assessment of the H → ττ pair

To classify the event in one of the considered H → ττ final states, the possible pairs of the selected τ_h, μ and e are built and compared, and the three considered final states go through orthogonal selections.

The ττ pair type is assessed using offline information only. Selected events are required to have at least one tau lepton candidate that decayed hadronically. The type of the other leg is determined according to the object purity: if a muon is found, the event is classified as $\tau_\mu\tau_\mathrm{h}$; otherwise, if an electron is found, it is classified as $\tau_\mathrm{e}\tau_\mathrm{h}$; finally, if a second τ_h is found and there is not an electron nor a muon, it is classified as a $\tau_\mathrm{h}\tau_\mathrm{h}$. Although the identification requirements were listed along with the object selection description for consistency, it should be mentioned that, as the events with non-isolated hadronic taus are needed to populate sidebands for the QCD estimation, the hadronic tau identification discriminator is not applied prior to the assessment of the H → ττ pair. As a consequence, several hadronic tau candidates are available at this stage, which motivates the following ordering strategy.

Once the pair type is assessed, all the possible pairs of the same type in the event are examined. The μ or e is placed as first leg; if the event is classified as $\tau_\mathrm{h}\tau_\mathrm{h}$, both legs permutations are considered. Pairs are at first sorted according to the isolation of their first leg. If two pairs are built using the same object as first leg, the pair with the most isolated second leg is preferred; if there is still ambiguity, priority is given to the pair with the highest second leg p_{T}. The first ττ pair in the collection thus sorted is the selected H → ττ pair.

A subsequent check of the compatibility between the selected final state and the trigger path, as detailed in Tab. 4.1, Tab. 4.2 and Tab. 4.3, is performed, requiring a geometrical matching between the online and offline objects. Thus, the correct treatment of the trigger efficiencies is ensured. Each reconstructed offline lepton is required to pass a p_{T} threshold depending on the HLT trigger path fired by the event:

$$p_{\mathrm{T}}^{offline} \geq p_{\mathrm{T}}^{HLT} \; + \; offset$$

where $p_{\mathrm{T}}^{offline}$ is the transverse momentum of the offline selected object, p_{T}^{HLT} is the p_{T} threshold applied at trigger level and *offset* is different for each object: 2 GeV for muons, 3 GeV for electrons and 5 GeV for tau leptons.

For instance, the single-e trigger selects electrons with $p_{\mathrm{T}} > 32\,\mathrm{GeV}$ or $p_{\mathrm{T}} > 35\,\mathrm{GeV}$ and $|\eta| < 2.1$, with no requirements for the hadronic tau candidate. The corresponding offline selection requires an electron with $p_{\mathrm{T}} > 35\,\mathrm{GeV}$ and $|\eta| < 2.1$; there is no check on wether the path with lowest p_{T} threshold was fired or not: the efficiency scale factors already account for the logic OR of the paths of the same type. As no tau is required at trigger level, there is no trigger requirement driving the corresponding selection and the loosest thresholds allowed by the hadronic tau lepton identification algorithm are used: it should have $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ and $|\eta| < 2.3$. The cross $e\tau_{\mathrm{h}}$ trigger selects electrons with $p_{\mathrm{T}} > 24\,\mathrm{GeV}$ reconstructed within the region with $|\eta| < 2.1$. As for the τ_{h} leg of the cross trigger, it must have $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $|\eta| < 2.1$. These requirements drive the selection for both objects: the electron reconstructed offline must have $p_{\mathrm{T}} > 27\,\mathrm{GeV}$ and $|\eta| < 2.1$, while the hadronic tau lepton must have $p_{\mathrm{T}} > 35\,\mathrm{GeV}$ and $|\eta| < 2.1$. If none of the combination of triggers and their corresponding offline selection is fulfilled, the event is rejected.

The object selections for each final state follow the same strategy and they are summarised in Tab. 4.4, Tab. 4.5 and Tab. 4.6. Additional selections, common to all the final states, are applied for consistency among the final states and to further increase the purity of the selection. The candidates forming the $\tau\tau$ pair must to be separated by $\Delta R > 0.1$; thus, it is ensured that none of the selected particles is reconstructed from identical particle flow candidates. Moreover, to reject the QCD contribution, the candidates forming $\tau\tau$ pairs are required to have opposite charge; the events failing this selection are exploited in the QCD background estimation.

Finally, a veto is applied to events where an additional electron or muon is found besides those selected in the $\tau\tau$ pair. This selection helps reducing the Drell-Yan contribution; moreover, it ensures that the three $\tau\tau$ categories are mutually exclusive. The object selection for veto leptons is looser than the one used for the $\tau\tau$ candidates: the event is rejected if an additional electron with $p_{\mathrm{T}} > 10\,\mathrm{GeV}$ and $|\eta| < 2.5$, passing the *Loose* MVA identification working point and the *Loose* relative isolation working point ($I^{e}_{rel} < 0.3$) is found, or if there is an additional muon with $p_{\mathrm{T}} > 10\,\mathrm{GeV}$ and $|\eta| < 2.4$, passing the *Loose* identification working point and the *Loose* relative isolation working point ($I^{\mu}_{rel} < 0.3$).

4.4 H→bb categorization

The H→bb event categories definition, described in Sec. 4.4.4, follows the strategy previously defined for the HH → b$\bar{\text{b}}\tau\tau$ analysis on 2016 data [107]. The VBF categories, instead, are entirely designed within this thesis work; their optimisation is detailed in Sec. 4.7.

Table 4.4 – Offline selection for the $\tau_e\tau_h$ final state.

$\tau_e\tau_h$ channel		
	HLT paths	Selection
e	all	$I^e_{rel} < 0.10$, *Tight* MVA ID $d_{xy} < 0.045\,\text{cm}$, $d_z < 0.2\,\text{cm}$
	single-e	$p_T > 35\,\text{GeV}$, $\lvert\eta\rvert$<2.1
	cross $e\tau_h$	$p_T > 27\,\text{GeV}$, $\lvert\eta\rvert$<2.1
τ_h	all	Decay: $h^\pm$, $h^\pm\pi^0$, $h^\pm h^\mp h^\pm$ $d_{xy} < 0.045\,\text{cm}$, $d_z < 0.2\,\text{cm}$ *Medium* MVA ID *Loose* anti-μ and *Tight* anti-e
	single-e	$p_T > 20\,\text{GeV}$, $\lvert\eta\rvert$<2.3
	cross $e\tau_h$	$p_T > 35\,\text{GeV}$, $\lvert\eta\rvert$<2.1
Pair		$\Delta R(e, \tau_h) > 0.1$ Opposite charge

Table 4.5 – Offline selection for the $\tau_\mu\tau_h$ final state.

$\tau_\mu\tau_h$ channel		
	HLT paths	Selection
μ	all	$I^\mu_{rel} < 0.15$, *Tight* ID $d_{xy} < 0.045\,\text{cm}$, $d_z < 0.2\,\text{cm}$
	single-μ	$p_T > 26\,\text{GeV}$, $\lvert\eta\rvert$<2.1
	cross $\mu\tau_h$	$p_T > 22\,\text{GeV}$, $\lvert\eta\rvert$<2.1
τ_h	all	Decay: $h^\pm$, $h^\pm\pi^0$, $h^\pm h^\mp h^\pm$ $d_{xy} < 0.045\,\text{cm}$, $d_z < 0.2\,\text{cm}$ *Medium* MVA ID *Tight* anti-μ and *Very loose* anti-e
	single-μ	$p_T > 20\,\text{GeV}$, $\lvert\eta\rvert$<2.3
	cross $\mu\tau_h$	$p_T > 32\,\text{GeV}$, $\lvert\eta\rvert$<2.1
Pair		$\Delta R(\mu, \tau_h) > 0.1$ Opposite charge

The final categorisation of the selected events is based on the number and the nature of the reconstructed jets: two b jets result from a Higgs boson decay, while two forward jets are the signature of a VBF production. The jet assignment flow was conceived as to allow a reasonably simple matching between the categories of the 2016 published analysis and the categories of the 2017 (and 2018) analysis. Although the expected kinematic features are different enough, the probability of jet mistagging is sizeable. The chosen jet sorting procedure, described in Sec. 4.4.3, is as conservative as possible towards the choice of the eligible events for the inclusive categories, while selecting efficiently the VBF events.

Table 4.6 – Offline selection for the $\tau_h\tau_h$ final state. The VBF H→ $\tau_h\tau_h$ trigger entails additional selections on the VBF jet pair; only events eligible for the VBF signal category use this HLT path.

$\tau_h\tau_h$ channel		
	HLT paths	Selection
Both τ_h	all	Decay: $h^\pm$, $h^\pm\pi^0$, $h^\pm h^\mp h^\pm$ $d_{xy} < 0.045\,\text{cm}$, $d_z < 0.2\,\text{cm}$ *Medium* MVA ID *Loose* anti-μ and *Very loose* anti-e
	di-τ_h	$p_T > 40\,\text{GeV}$, $\lvert\eta\rvert<2.1$
	VBF H→ $\tau_h\tau_h$	$p_T > 25\,\text{GeV}$, $\lvert\eta\rvert<2.1$
Pair		$\Delta R(\tau_h, \tau_h) > 0.1$ Opposite charge

4.4.1 Selection of the jets

Jets are reconstructed using the "anti-k_t" algorithm, described in Sec. 2.3.4. Regular jets, or "AK4 jets", are reconstructed with a distance parameter $R = 0.4$. The signal topology such as in specific BSM scenarios with large m_{HH} [44] features highly Lorentz-boosted b jets; in these cases, the hadronization cones are likely to be partially in overlap and cannot be resolved as separate AK4 jets. Therefore, large-area jets or "AK8 jets", reconstructed with $R = 0.8$, are also used in this analysis to recover the boosted topologies. Unless explicitly stated otherwise, the mention of "jets" in the following always denotes AK4 jets.

Particle flow-based identification criteria are applied to reject poorly reconstructed jets as well as jets induced by calorimeter noise. The *Tight* working point of the identification, required for the jet selection, corresponds to a set of thresholds on quantities related to the type of particle flow constituents of the jets, such as the number of candidates of each type clustered into a jet and the fraction of jet energy that is carried by them.

The jet identification efficiency is estimated in data with events where at least a pair of jets well separated and with large invariant mass are found. The measurement is performed through a tag-and-probe technique: one of the two leading jets from the dijet selection (tag) is required to fulfil the *Tight* working point criteria; the efficiency is defined as the number of events where the opposite jet (probe) also passes the *Tight* identification over the number of selected events. The measured efficiency is larger than 99.5% in all the pseudorapidity regions.

In addition to the jet identification requirement, a pileup jet discriminator is used in specific regions. The so-called "pileup jets" are jets not originating from the primary vertex, produced with high p_T or reconstructed as a result of the overlap of several low p_T jets. They can be identified using variables related to the jet-shape and tracking observables: they tend to be broader than jets originating from quarks and gluons in the hard interaction, and the majority of the tracks of their constituents are not associated with the primary vertex. Such variables are given as input for the BDT technique used to build the pileup jet discriminant.

The noise contribution at large η discussed in Sec. 4.3.4 leads to an increase of the multiplicity of low p_T jets in data. Similarly to the p_T^{miss} distribution, jet-related quantities

as the jet multiplicity present a disagreement between the data and background distributions; within the H $\to \tau\tau$ analysis, it was observed that the data-over-prediction ratio gets larger as the number of jets grows, reaching a 90%-95% disagreement for $N_{jets} = 10$. The exclusion of jets reconstructed in the the noisy regions from the $p_{\mathrm{T}}^{\mathrm{miss}}$ computation motivates a similar cleaning of the jet collection. The strategy used in the H $\to \tau\tau$ analysis consists in rejecting jets with $p_{\mathrm{T}} < 50\,\mathrm{GeV}$ reconstructed with $2.65 < |\eta| < 3.14$. Although these jets are not entirely produced by pileup interactions, the same features

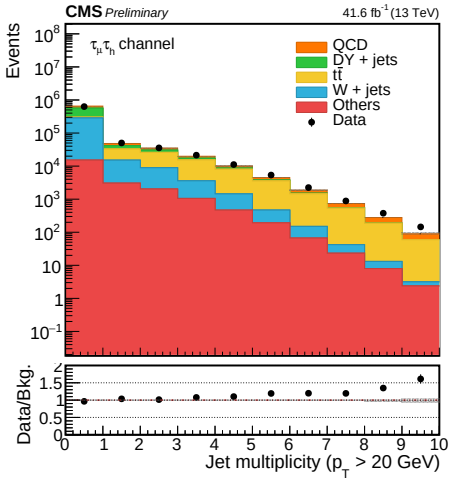


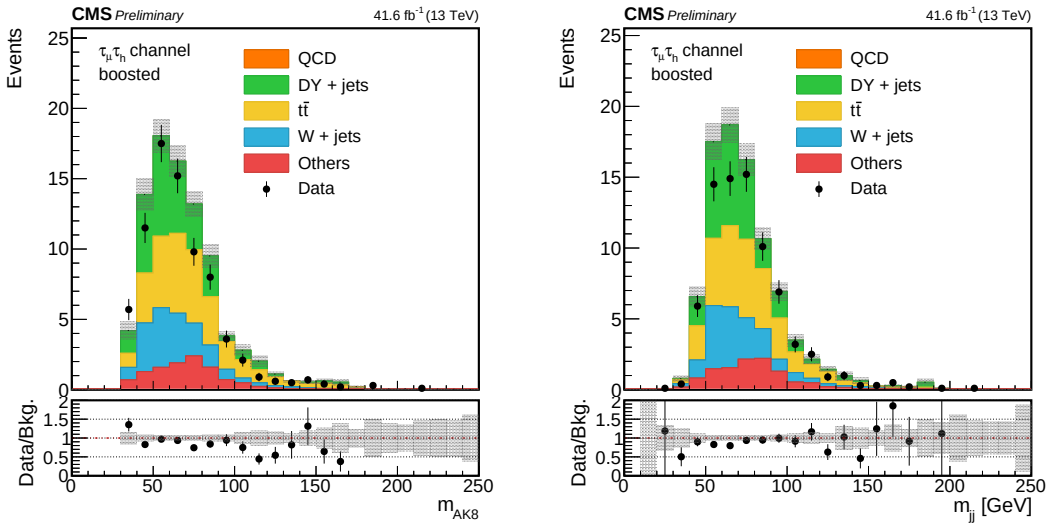
Figure 4.16 – Multiplicity of jets with $p_{\mathrm{T}} > 20\,\mathrm{GeV}$. Jets reconstructed with $2.65 < |\eta| < 3.14$ and not passing the *Loose* working point of the pileup identification are rejected. The error bars along y for the data distribution and for the data-over-prediction ratio are too small to be appreciated.

can be exploited to distinguish them from jets originating from the primary interactions. Thus, the application of the pileup jet discriminant was found to be effective in recovering the agreement between the data and background distributions of jet-related quantities. The distribution of the number of jets, excluding only those with $p_{\mathrm{T}} < 50\,\mathrm{GeV}$ that do not pass the *Loose* pileup identification working point and that are reconstructed with $2.65 < |\eta| < 3.14$, is shown in Fig. 4.16. The performance of such jet cleaning is similar to that of the strategy used in H $\to \tau\tau$. Besides, the use of the pileup discriminant gives a looser selection than a $p_{\mathrm{T}} < 50\,\mathrm{GeV}$ threshold and allows a larger acceptance on VBF jet candidates to be preserved. Therefore, this strategy is preferred and it is implemented for the jet selection. The data-over-prediction ratio for $N_{\mathrm{jets}} = 10$ is reduced to 1.5.

The "soft drop declustering" grooming algorithm is used to interpret the substructure of AK8 jets and identify the hadronization products of the two b quarks, while mitigating the contribution from initial state radiation and pileup. First, the AK8 constituents are reclustered: rather than being ordered by $1/k_t$, they are combined in pairs sorted by increasing ΔR separation. The large-area jet thus obtained is then broken in two subjets, i.e. it is reverted to the last step of the clustering procedure; the softer subjets are recursively dropped while declustering, until a pair of hard subjets is found. The invariant mass of the AK8 system is efficiently computed though the grooming algorithm; its distributions, showing a good data-over-prediction agreement, is shown in Fig. 4.17 and it is compared to the invariant mass of the regular AK4 jets in the same events.

4.4.2 b jets selection

The most distinctive features of the hadronization of b quarks, exploited to build b tagging variables, are related to the lifetime of B hadrons, typically of about 1.5 ps,



(a) (b)

Figure 4.17 – Event distribution in the $\tau_\mu\tau_h$ channel as a function of the mass m_{AK8} of AK8 jets computed through the "soft drop declustering" algorithm (left) and of the invariant mass AK4 jets in the same events (right).

present in jets originating by b quarks; such lifetime is considerably longer than the one of hadrons produced by c quarks or light quarks, which is below 1 ps. Depending on its momentum, the B hadron can travel from the primary vertex for a few millimetres up to about 1 cm before decaying; therefore, the b jets typically contain displaced tracks, consistent with a secondary vertex. Moreover, an electron or a muon is produced within the 20% of b jets, whereas the fraction c jets containing leptons is 10%: even less leptons are produced within jets coming from light jets.

In this analysis, a b tagger based on a deep neural network (DeepCSV) [111] is used to select the b jet candidates. The DeepCSV output is interpreted as a probability for a given jet to belong to one of the flavour categories, related to the number of b and c hadrons within the jet. The probability P(b) that at least one B hadron is produced within the jet and the probability P(bb) that exactly two B hadrons are produced are summed together to define a single DeepCSV discriminator. As shown in Fig. 4.18, a good separation is achieved between jets originated from b quarks, whose distribution is peaked towards large values of the discriminator, and other flavours. A jet is considered b-tagged if its DeepCSV score is larger than a threshold, chosen based on the discriminating power against jets originated by other quark flavours. Only jets with $p_T > 20$ GeV and $|\eta| < 2.4$ are selected as b jet candidates. The pseudorapidity range is restricted to the central region because, for the b tagging discriminant to be built, jets must be reconstructed predominantly within the acceptance of the tracker; however, the decay products of the Higgs boson are typically located in the central part of the detector.

As detailed below, the b tag efficiency on simulated events is measured using reconstructed jets that are matched to generated quarks produced in the hard interaction; therefore, pileup jets do not enter the computation. To avoid biases from pileup jets in the efficiency estimate, the b jet candidates are required to pass the *Loose* pileup identification working point.

The distribution of the p_T and of the DeepCSV score of the selected jet with highest DeepCSV score is shown in Fig. 4.19. An excellent data-over-prediction agreement is observed in the p_T distribution. The DeepCSV discriminant efficiently separates the

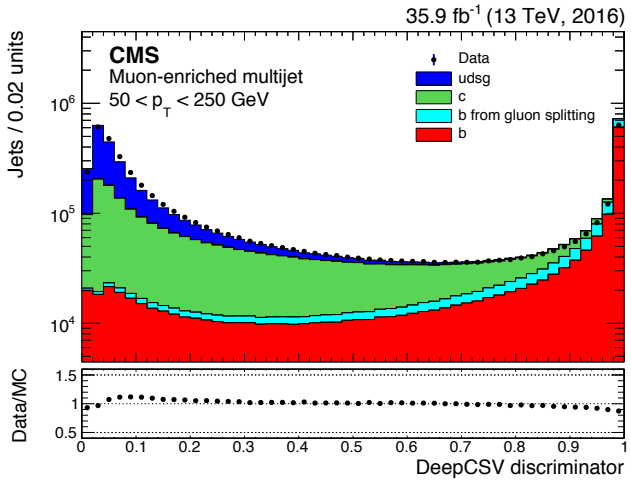
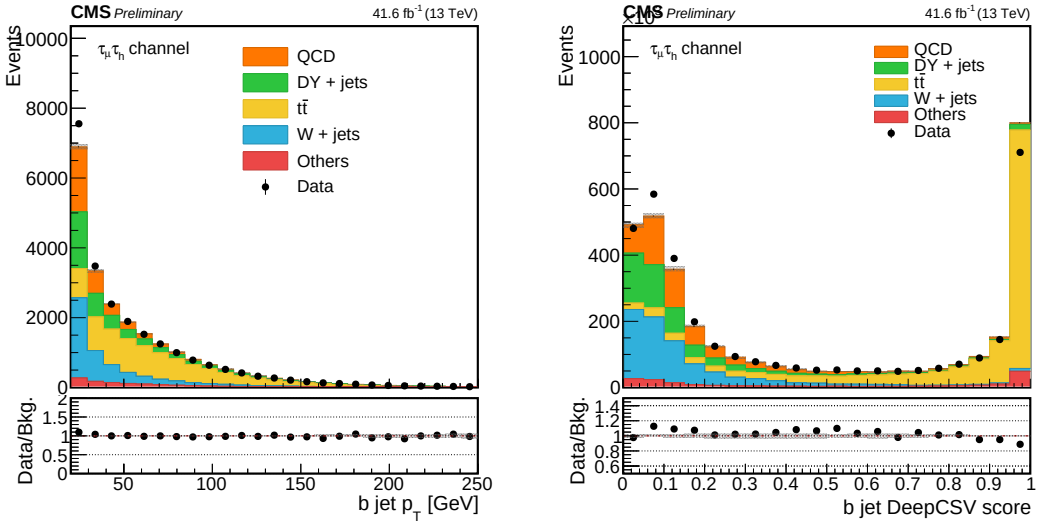


Figure 4.18 – Distribution of the DeepCSV P(b)+P(bb) discriminator value for jets of different flavours in muon enriched QCD events [111].

$t\bar{t}$ process, featuring jets originating from b quarks, from other backgrounds; in this analysis, the *Medium* (DeepCSV score > 0.15) and *Loose* (DeepCSV score > 0.8) working points are used. However, a disagreement of about 15% is observed for very low and very high values of the discriminant: the b jet candidates in simulated events tend to be over ranked. To correct this trend, a weight ω is applied event-by-event in simulated



(a) (b)

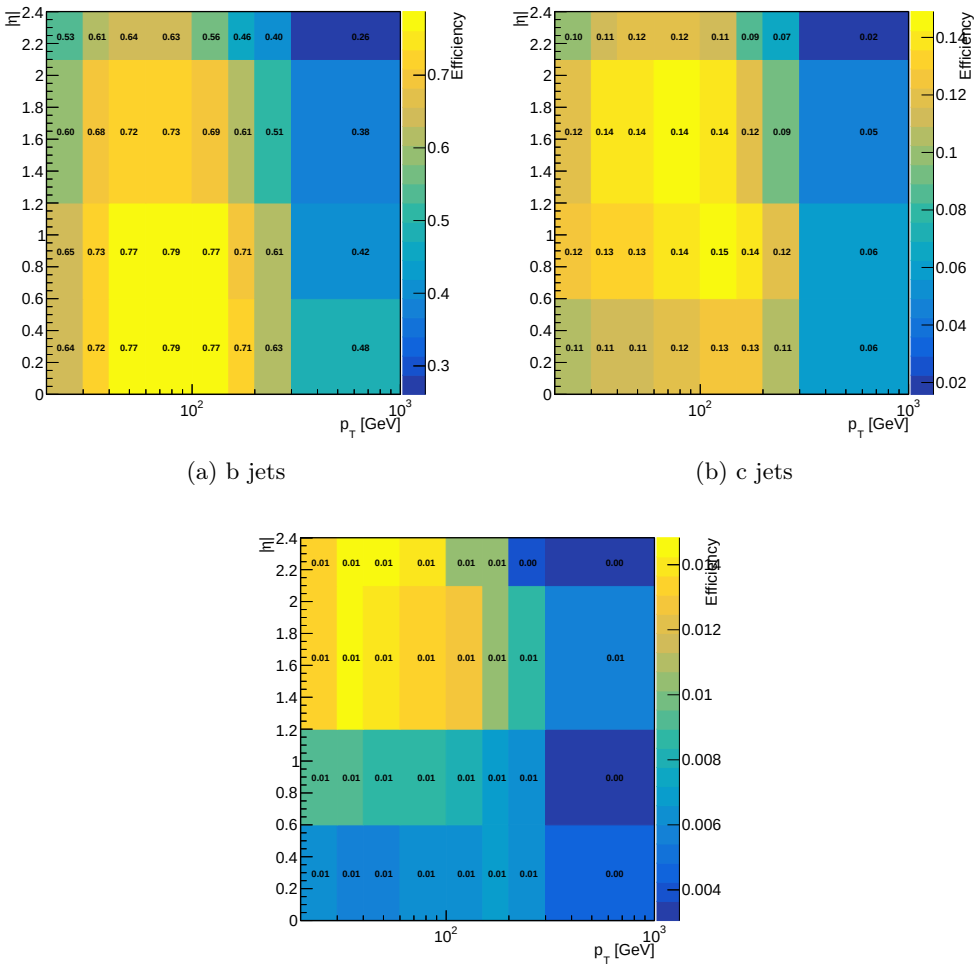
Figure 4.19 – Transverse momentum (left) and b tag score (right) of the highest DeepCSV score b jet candidate in the $\tau_\mu\tau_h$ channel. The *Medium* and *Loose* b tag working points, used in this analysis, correspond respectively to DeepCSV score > 0.15 and > 0.8. The error bars along y for the data distribution and for the data-over-prediction ratio are too small to be appreciated.

samples; it accounts for the b tagging performance and the mistagging probability, i.e. the probability that a jet originated by a quark of different flavour is tagged as a b jet, in data and simulations.

Scale factors are provided by the b tag and vertexing POG; they are computed using different event topologies, chosen according to the their flavour composition. Generic QCD events are used to compute the light jet mistagging probability; muon-enriched QCD events and $t\bar{t}$ events are used to compute the b and c tagging efficiency. In simulated

events, the b and c tagging efficiency is computed as the number of jets tagged with a given DeepCSV working point over the number of jets matched to generated b or c quarks. In data, the denominator of the efficiency has tight selections to identify regions very pure in flavour composition. The mistagging probability computation is performed with a similar strategy.

The flavour tagging efficiencies strongly depend on the kinematics of the considered processes; hence, an additional measurement of the efficiency ϵ as a function of p_{T} and η is performed within this analysis in $\mathrm{t\bar{t}}$ simulated events that pass the $\mathrm{H} \to \tau\tau$ selection described in Sec. 4.3. The resulting efficiencies are shown in Fig. 4.20.



(a) b jets

(b) c jets

(c) light jets

Figure 4.20 – Efficiency of the DeepCSV *Medium* working point selection on b jets, c jets and light jets in simulated $\mathrm{t\bar{t}}$ events, as a function of the jet $|\eta|$ and p_{T}.

The probability of a given configuration of jets, "tagged" and "not-tagged" by a flavour tagger, in an event is thus defined as

$$P(MC) = \prod_{i \in \text{tagged}} \epsilon_i \prod_{j \in \text{not-tagged}} (1 - \epsilon_j) \tag{4.4}$$

$$P(Data) = \prod_{i \in \text{tagged}} \mathrm{SF}_i \epsilon_i \prod_{j \in \text{not-tagged}} (1 - \mathrm{SF}_j \epsilon_j) \tag{4.5}$$

where SF is the p_{T} dependent scale factor provided centrally and ϵ is the efficiency on

simulated $t\bar{t}$ events. Finally, for each simulated event, the event weight is computed as

$$\omega = \frac{P(Data)}{P(MC)}. \tag{4.6}$$

The resulting correction, applied to events selected through a b tag requirement, is global: it takes into account all the jets in the event and their flavour, rather than only the b-tagged jets.

4.4.3 b jets and VBF jets assignment

To reconstruct two b jets corresponding to the decay of a boson and simultaneously search for VBF jet candidates, a jet arbitration procedure has to be set up.

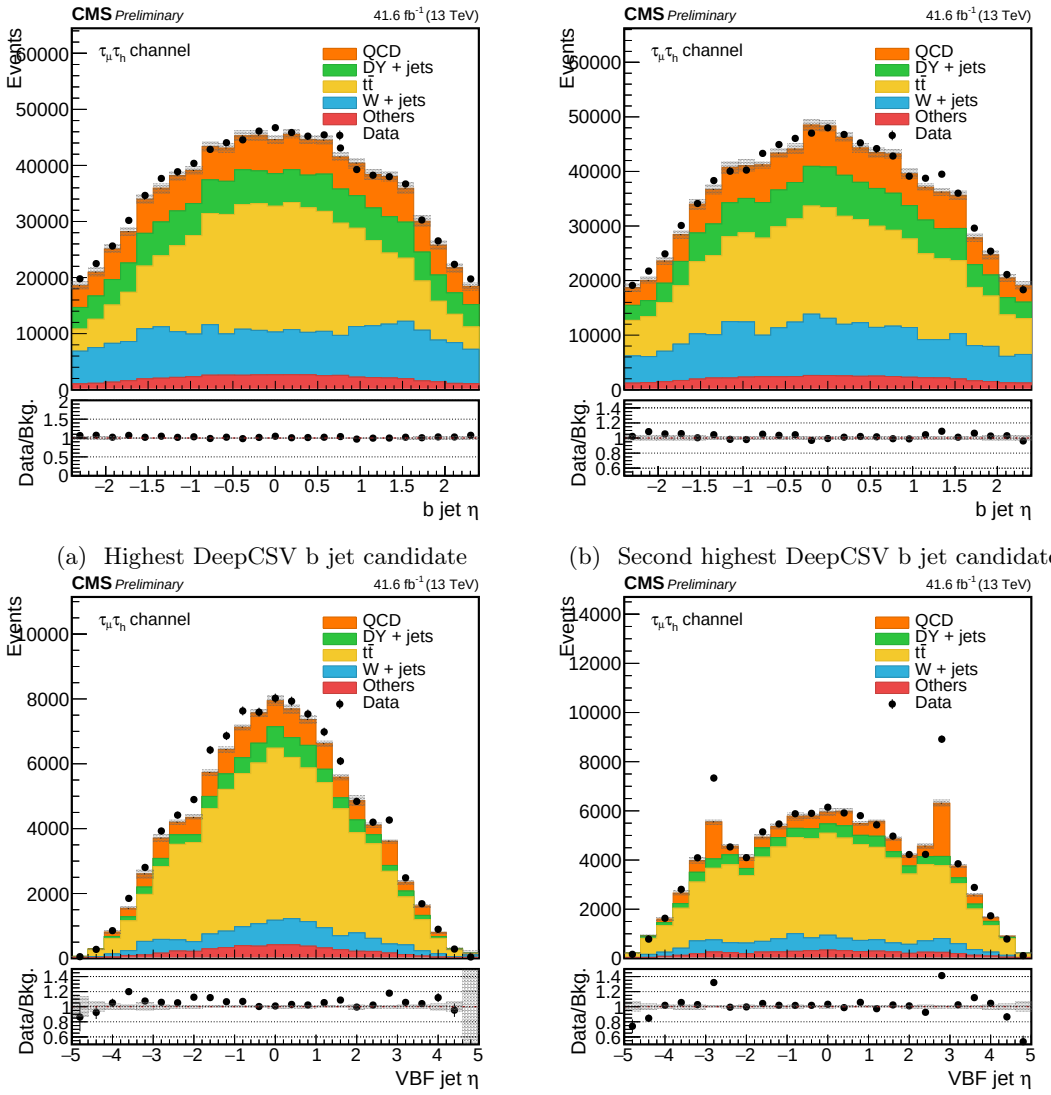
Jets are ordered according to their DeepCSV discriminator output and the one with the highest score is chosen as the first b jet candidate. The second ordered jet is selected as the other b jet candidate if it passes the *Medium* DeepCSV working point or if there are not additional jets in the event. Otherwise, the VBF jets pair is assessed first.

Additional jets with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and *Tight* jet identification are selected as VBF jet candidates. Since the VBF jets are typically produced in the forward regions of the detector, no η restriction is required; however, jets reconstructed in the noisy region with $2.65 < |\eta| < 3.14$ are required to pass the *Loose* pileup discriminant, as mentioned in Sec. 4.4.1. The selected VBF jets candidates should be well separated ($\Delta R > 0.5$) from the leptons of the $\mathrm{H} \to \tau\tau$ pair. If more than two additional jets fulfilling this requirement are found in the event, the pair of jets with the highest highest jet-jet invariant mass is selected.

If the second b jet is not assessed at this stage, it is now selected as the next jet by DeepCSV score among those that are not yet assigned; there are some chances, indeed, that the second jet by DeepCSV score is selected as one of the VBF jet candidates. Lastly, if after the VBF jet selection no jet is left fulfilling the b jet selection criteria, the VBF pair is discarded and the original sorting by DeepCSV score is restored.

This assignment procedure adopted has a strong dependence on the tag score of the second b jet candidate and does not seem very natural. As a matter of fact, it was conceived with the aim of being consistent with respect to the previous $\mathrm{HH} \to \mathrm{bb}\tau\tau$ analysis, where the two highest b tag score jets are selected as b jet candidates. Indeed, the H→bb final categories definition, described in Sec. 4.4.4, depends on the b tag score of the b jet candidates and it should be preserved at best in spite of the introduction of the VBF category. In practice, the current jets assignment is performed considering that if a b jet candidate does not have a b tag score large enough to enter the final inclusive categories, it might as well be considered as a VBF jet candidate; at the same time, the b jet selection in the existing inclusive categories is preserved. The jet sorting described allows about the 10% more of VBF jets to be correctly assigned with respect to a procedure consisting in selecting the VBF jets after the two b jets.

In Fig. 4.21, the pseudorapidity of the jets thus selected in $\tau_\mu\tau_\mathrm{h}$ events is shown. As expected, the b jet candidates are mostly reconstructed in the central region of the detectors. Between the VBF jet candidates, the one with highest p_{T} is the most central. Both VBF jet candidates show a residual disagreement in the noisy area of the detector, not fully cleaned by the rejection of pileup jets in that region; the η distribution of the second VBF jet candidate is the most affected. Besides, the m_{jj} and $|\Delta\eta_{\mathrm{jj}}|$ distribution of the VBF jet candidates, shown in Fig. 4.22, present a satisfactory agreement.



(a) Highest DeepCSV b jet candidate

(b) Second highest DeepCSV b jet candidate

(c) Leading VBF jet candidate

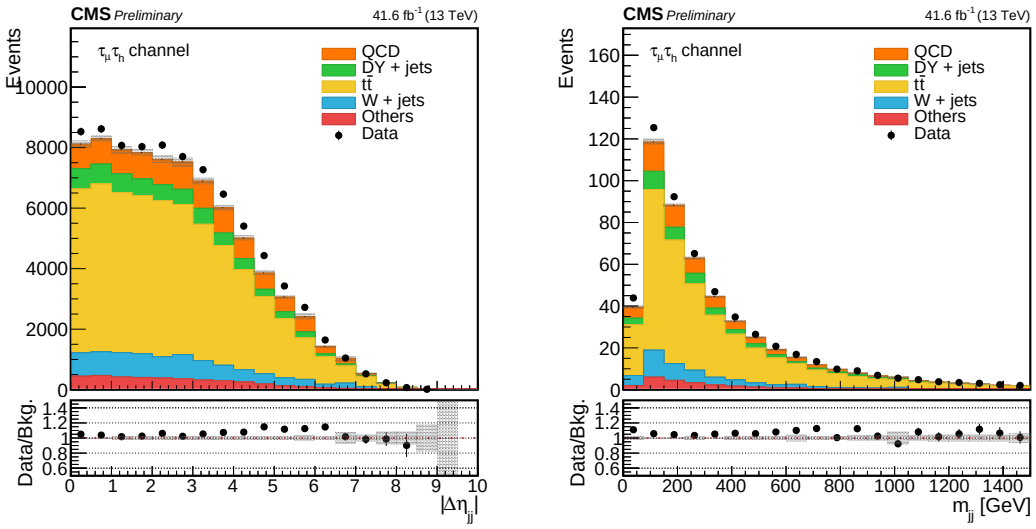
(d) Trailing VBF jet candidate

Figure 4.21 – Pseudorapidity distribution of the two b jet candidates (top) and of the two VBF jet candidates (bottom) in $\tau_\mu\tau_\mathrm{h}$ events. The error bars along y for the data distribution and for the data-over-prediction ratio are too small to be appreciated.

4.4.4 H→ bb categories

For events to be selected, at least two jets that are compatible with the b jet candidate selection presented in Sec. 4.4.2 prior to the b tag requirements, i.e. jets with $p_T > 20\,\mathrm{GeV}$ and $|\eta| < 2.4$ passing the *Tight* identification and the *Loose* pileup discriminant, must be found. The b jets candidates are also required to have a $\Delta R > 0.5$ separation with the leptons selected in the H $\to \tau\tau$ pair candidate. The first b jet candidate is the one with highest DeepCSV score; the selection the second b jet candidate is bound to that of the VBF jets, as detailed in Sec. 4.4.3. In the following, this stage of the selection will be often referred to as "baseline".

Three different regimes are possible for the reconstruction of jet pairs: if their separation ΔR is larger than 0.8, they are reconstructed as separate AK4 jets; in the intermediate regime, when $0.4 < \Delta R < 0.8$, the jets are partially in overlap and they are reconstructed both as separated AK4 jets and as one merged AK8 jet; finally, highly boosted jets with separation smaller than 0.4, they can only be reconstructed as AK8 jets. The scenario



(a) (b)

Figure 4.22 – Invariant mass (left) and angular separation (right) distributions of the two selected VBF jet candidates in $\tau_\mu\tau_\mathrm{h}$ events. The error bars along y for the data distribution and for the data-over-prediction ratio are too small to be appreciated.

with $\Delta R < 0.4$ is found to be not relevant for most of the signals under study. Therefore, events are only classified into "resolved" or "boosted" categories, corresponding to the $\Delta R > 0.8$ and $0.4 < \Delta R < 0.8$ topologies and that use AK4 jets and AK8 jets with substructure requirements.

The inclusive categories described in the following are orthogonal to the VBF categories: if a pair of b jet candidates and a pair of VBF jet candidates that fulfil the requirements described in Sec. 4.7 are found, priority is given to the categorisation as VBF event.

Events in the boosted category should have a AK8 jet with mass $m_{\mathrm{AK8}} > 30\,\mathrm{GeV}$ and they should be geometrically matched to the two previously selected AK4 b jet candidates with $\Delta R < 0.4$. If these criteria are not fulfilled, the event is assigned to the resolved categories.

Finally, b tag requirements are applied in each category, so that backgrounds with jets originating from light quarks are rejected. As summarised in Tab. 4.7, the final categories are defined based on the DeepCSV score of the selected b jets: the high-purity resolved category ("resolved 2b0j") contains event where both b jet candidates pass the DeepCSV *Medium* working point; a second resolved category with higher statistics ("resolved 1b1j") contains events where only one of the b jet candidates passes the DeepCSV *Medium* working point; lastly, the two b jet candidates selected in the boosted category should have DeepCSV score larger than the *Loose* working point, to preserve larger statistics.

4.5 HH signal region

Once the tau lepton pair and the b jet candidates are identified, they can be exploited to reconstruct observables related to the Higgs boson candidates $\mathrm{H}_{\tau\tau}$ and H_{bb} from which they are originating; their invariant mass, indeed, can be used to define a tight signal region and further reject the background contributions.

The invariant mass of the $\tau\tau$ pair is reconstructed using the SVfit algorithm [112], based on a likelihood function which quantifies the level of compatibility between a Higgs mass

Table 4.7 – Offline jets selection in the inclusive categories. The H→bb classification is performed after the H → ττ selections summarised in Sec. 4.3. To be selected in the inclusive categories, the event should not pass the VBF category requirements summarised in Tab. 4.7.

ττ pair type	H→bb categories		
	2 b jet candidates with $p_T > 20\,GeV$ and $\|\eta\| < 2.4$ and passing *Tight* ID, *Loose* pileup ID		
	Resolved 2b0j	**Resolved 1b1j**	**Boosted**
all	both b jet candidates have *Medium* DeepCSV tag	only one b jet candidate has *Medium* DeepCSV tag	AK8 jet with $m_{AK8} > 30\,GeV$ and $p_T > 170\,GeV$, matched to 2 b jet candidates with 2 *Loose* DeepCSV tags

hypothesis and the measured momenta of the visible tau lepton decay products plus the missing transverse momentum reconstructed in the event. The resolution on the invariant mass of the ττ pair thus computed ($m_{\tau\tau}^{SVfit}$) is improved compared with the visible mass ($m_{\tau\tau}^{vis}$), e.g. the invariant mass computed using only the visible particles. The corresponding distributions in $\tau_h\tau_h$ simulated events are compared in Fig. 4.23; the SVfit algorithm gives a resolution improved by about 35%.

The mass of the H_{bb} candidate is computed as the invariant mass of the two selected b jet candidates. Its distribution and the one of $m_{\tau\tau}^{SVfit}$ is shown for the different channels and different categories in Fig. 4.25 and Fig. 4.24.

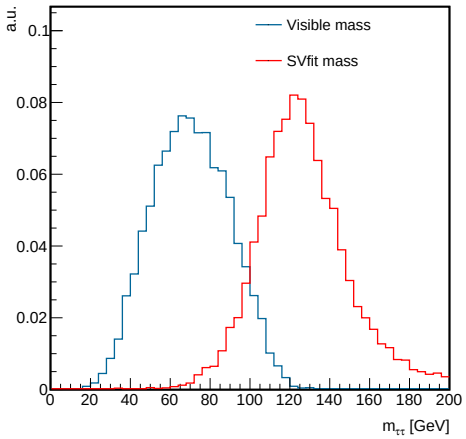
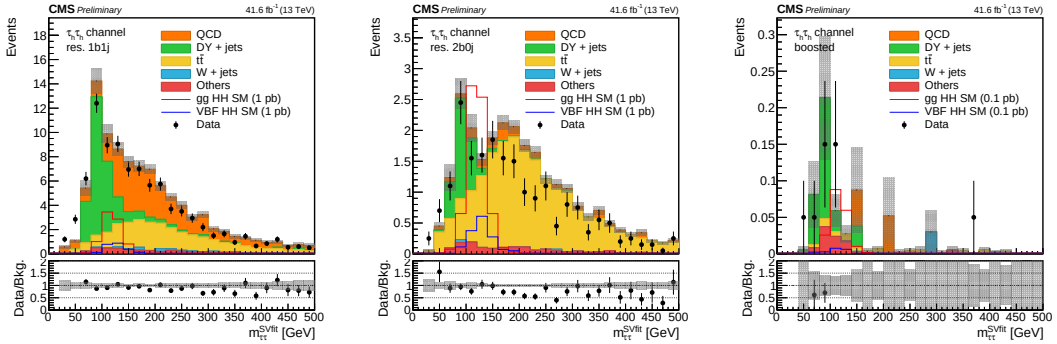


Figure 4.23 – Mass of the $H_{\tau\tau}$ candidate in $\tau_e\tau_h$ and $\tau_\mu\tau_h$ events in a simulated gg→HH sample, computed as the invariant mass of the reconstructed hadronic tau leptons and through the SVfit algorithm.

Within the resolved categories, a signal region is designed through an elliptical cut on the $m_{\tau\tau}^{SVfit}$ vs. m_{bb} plane. The ellipse is defined by

$$\frac{\left(m_{\tau\tau}^{SVfit} - 116\,GeV\right)^2}{\left(35\,GeV\right)^2} + \frac{\left(m_{bb} - 111\,GeV\right)^2}{\left(45\,GeV\right)^2} < 1 \tag{4.7}$$

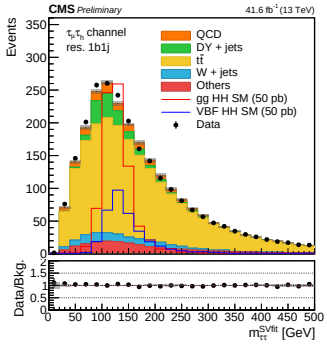
where the values of 35 and 45 GeV are the measured resolution on the invariant mass for the ττ and bb objects; the ellipse is centred on the position of the expected reconstructed



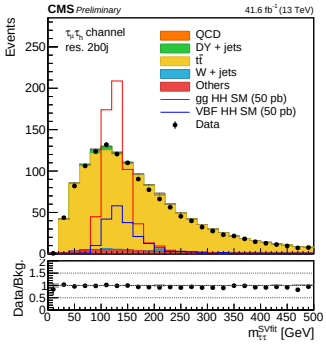
(a) $\tau_h\tau_h$ res. 1b1j

(b) $\tau_h\tau_h$ res. 2b0j

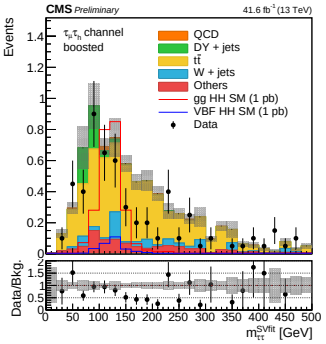
(c) $\tau_h\tau_h$ boosted



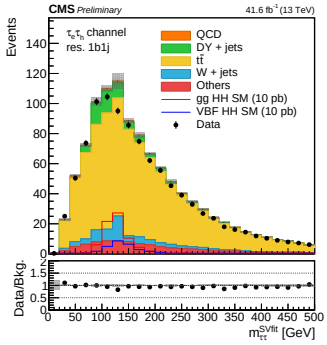
(d) $\tau_\mu\tau_h$ res. 1b1j



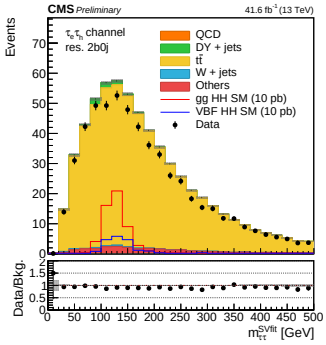
(e) $\tau_\mu\tau_h$ res. 2b0j



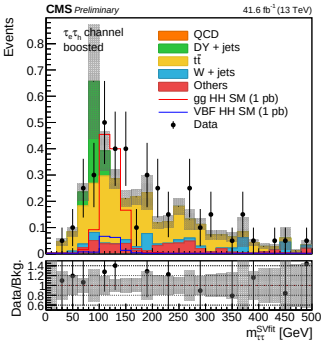
(f) $\tau_\mu\tau_h$ boosted



(g) $\tau_e\tau_h$ res. 1b1j



(h) $\tau_e\tau_h$ res. 2b0j



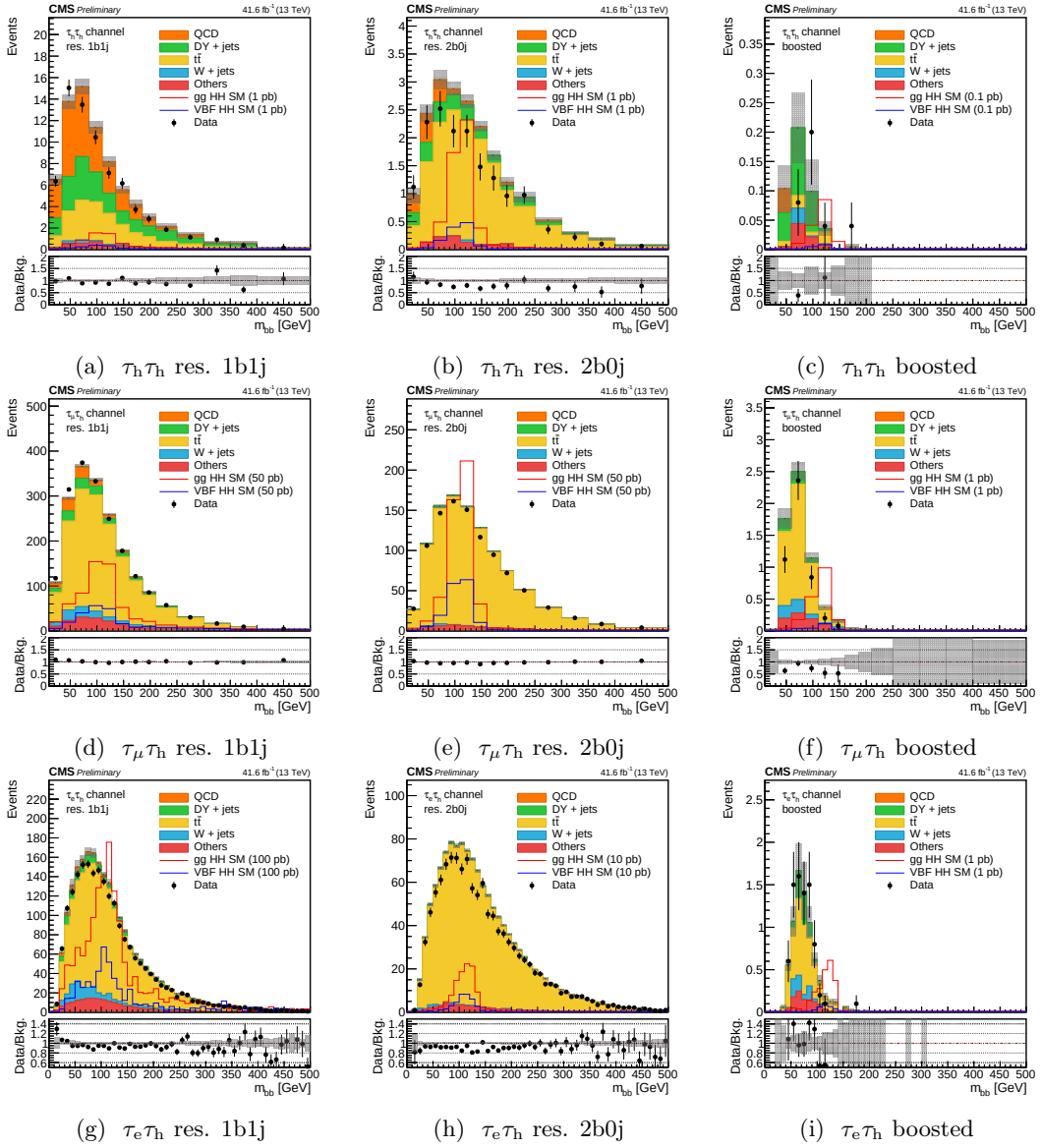
(i) $\tau_e\tau_h$ boosted

Figure 4.24 – Data and background event distribution as a function of the mass of the $H_{\tau\tau}$ candidate, computed through the SV fit tool, in the three channels and in the tree $H \to bb$ categories. No mass cut is applied. The SM gluon fusion and VBF signals are represented; their normalisation is chosen in each plot to facilitate their visualisation; the relative gluon fusion-over-VBF normalisation is not preserved.

125 GeV Higgs boson peak in the $m_{\tau\tau}$ and m_{bb} distributions, i.e. 116 and 111 GeV. In the $\tau_\mu\tau_h$ channel, for instance, this selection allows to reject about 87% of the $t\bar{t}$ background in the resolved 2b0j category, while the gluon fusion SM signal is only suppressed by about the 30%. The 2D $m_{\tau\tau}^{\text{SVfit}}$ vs. m_{bb} signal and background event distributions are shown in Fig. 4.26 for the different channels in the resolved 2b0j category; the ellipse is indicated in red.

Because of the different kinematics, the mass signal region in the boosted categories is defined by a square selection

$$\begin{aligned} 80\,\text{GeV} < m_{\tau\tau}^{\text{SVfit}} < 160\,\text{GeV} \\ 90\,\text{GeV} < m_{\text{AK8}} < 160\,\text{GeV}. \end{aligned} \tag{4.8}$$



(a) $\tau_h\tau_h$ res. 1b1j (b) $\tau_h\tau_h$ res. 2b0j (c) $\tau_h\tau_h$ boosted

(d) $\tau_\mu\tau_h$ res. 1b1j (e) $\tau_\mu\tau_h$ res. 2b0j (f) $\tau_\mu\tau_h$ boosted

(g) $\tau_e\tau_h$ res. 1b1j (h) $\tau_e\tau_h$ res. 2b0j (i) $\tau_e\tau_h$ boosted

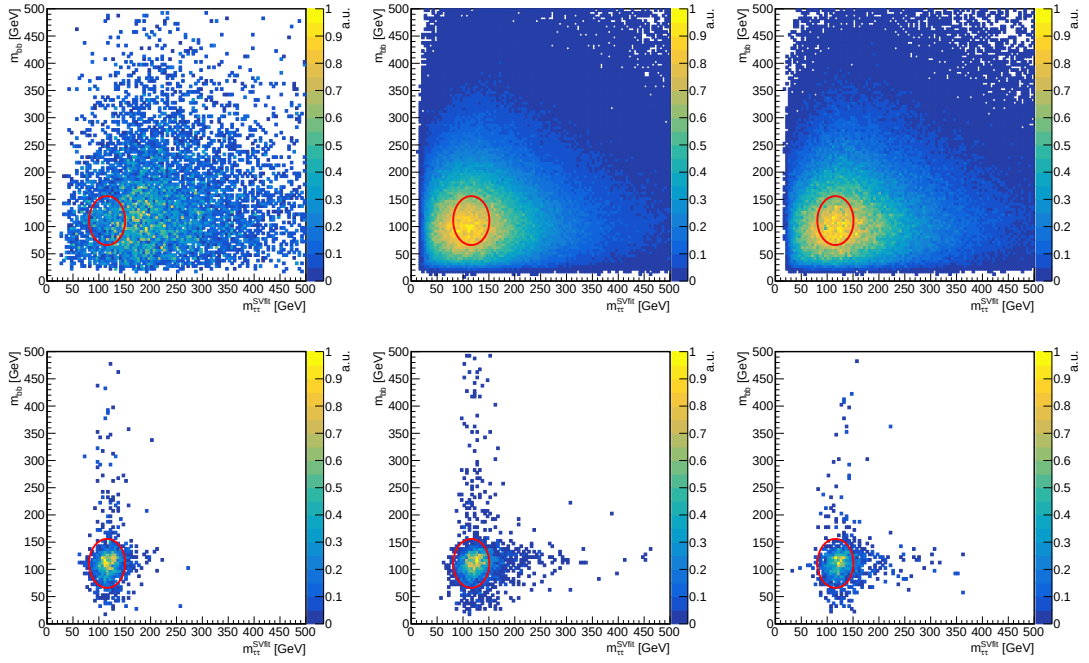
Figure 4.25 – Data and background event distribution as a function of the mass of the H_{bb} candidate, in the three channels and in the tree H → bb categories. No mass cut is applied. The SM gluon fusion and VBF signals are represented; their normalisation is chosen in each plot to facilitate their visualisation; the relative gluon fusion-over-VBF normalisation is not preserved.

4.6 Multivariate method for the tt̄ background rejection

After the b jets categorisation and the HH signal region selection, the background contributions are significantly reduced. However, a dedicated multivariate algorithm based on a BDT approach is designed to efficiently reject the residual contamination from the irreducible tt̄ background in all the channels by fully exploiting its kinematic features.

The multivariate classifier is developed using the TMVA toolkit [113], fully integrated in the ROOT analysis framework [114]. As it was found to give the best discrimination performance and to be the most robust against overtraining, a gradient boost algorithm was chosen after testing various algorithms available in TMVA.

The design of the MVA strategy is documented in [109]. It is optimized for the three $\tau\tau$ channels using the analysis strategies and the 2016 data sets used in [107], showing



(a) $\tau_h\tau_h$ (b) $\tau_\mu\tau_h$ (c) $\tau_e\tau_h$

Figure 4.26 – Event distribution of $t\bar{t}$ background (top) of gluon fusion signal (bottom) events, as a function of the $H_{\tau\tau}$ mass computed though the SVfit tool and of the invariant mass of the selected b jet candidates in the resolved 2b0j category. The red ellipse represents the mass cut applied to define the final signal region.

an improvement in sensitivity of about 30%. A similar multivariate technique was also implemented in the 2016 $HH \to bb\tau\tau$, although only available in the $\tau_e\tau_h$ and $\tau_\mu\tau_h$ final states. Instead of implementing a selection based on the BDT discriminant, as it was done in the 2016 $HH \to bb\tau\tau$ search, its output is meant to be the variable for the final signal extraction. Within the same study, several MVA discriminants are also optimised for resonant HH production; their performance is not detailed in the following, but they will be implemented in the future $HH \to bb\tau\tau$ analyses.

4.6.1 Choice of the input variables

The BDT algorithm is initially trained using an extensive set of over 100 potentially discriminant variables. Only a few of them are described in the following.

The visible mass of the $\tau\tau$ pair, its visible mass plus the missing transverse energy and its mass reconstructed through the SVfit algorithm are used as input variable; all the variables that contain information about the mass of the $H_{\tau\tau}$ candidate are also computed for each version of the mass estimate.

The mass of the HH system is reconstructed both using the KinFit algorithm and as "reduced mass" m_X. The KinFit algorithm [115] computes the mass m_{HH}^{KinFit} through a fit that takes as input the four-momenta of the selected $\tau\tau$ and b jet candidates and the missing momentum. The reduced mass is defined by

$$m_X = m_{\tau\tau bb} - (m_{bb} - m_H) - (m_{\tau\tau} - m_H), \tag{4.9}$$

where $m_H = 125\,\text{GeV}$ is the mass of the Higgs boson, while the other masses are computed from the selected objects in the hypothesis that they are originating from the Higgs bosons decay.

In Fig. 4.29 a schematic view of typical simulated SM HH and $t\bar{t}$ events, reconstructed as HH $\rightarrow$ bb$\tau\tau$, is shown. The two Higgs bosons are usually produced back-to-back; hence, the bb and the $\tau\tau$ systems are produced in opposite hemispheres of the detector. The Higgs decay product pairs also have small separation, for example, in (z, y) plane. In t$\bar{\text{t}}$ events, instead, two top quarks are produced back-to-back, and each of them decays in a b quark and a τ in association with neutrinos; the reconstructed bb and $\tau\tau$ systems, therefore, are not boosted and can have large separation in the (z, y) plane. Moreover, in signal events, the missing momentum vector $\vec{p}_{\text{T}}^{miss}$ typically has the same direction as the reconstructed $\text{H}_{\tau\tau}$ candidate; in background events, since each top quark produces a bτ pair candidate, the transverse momentum is randomly distributed. Therefore, signal events have a small angular separation in the transverse plane between the $\vec{p}_{\text{T}}^{miss}$ and the lepton momentum vector $\vec{p}^{\ell}$. The transverse mass, defined as

$$m_{\text{T}}(\ell, p_{\text{T}}^{\text{miss}}) = \sqrt{2p_{\text{T}}^{\text{miss}} p_{\text{T}}^{\ell}(1 - \cos\Delta\Phi)} \tag{4.10}$$

is sensitive to this difference in the event topology: signal events usually have small values of m_{T}, while $t\bar{t}$ events usually have a transverse mass close to the mass of the W boson; this consideration holds for both the $\tau\tau$ legs and for the reconstructed $\text{H}_{\tau\tau}$ candidate. Additional variables exploiting these kinematic properties are the total transverse mass

$$m_{\text{T}}^{\text{TOT}} = \sqrt{m_{\text{T}}(\ell, p_{\text{T}}^{\text{miss}})^2 + m_{\text{T}}((\tau_{\text{h}}, p_{\text{T}}^{\text{miss}})^2 + m_{\text{T}}(\ell, \tau_{\text{h}})^2} \tag{4.11}$$

and the "stransverse mass" [116]. The latter is a particularly powerful handle for the HH vs. t$\bar{\text{t}}$ discrimination; it was, indeed, used for the final signal extraction in the 2016 HH $\rightarrow$ bb$\tau\tau$ analysis. It exploits the topology of processes like the t$\bar{\text{t}}$, where a pair of identical mass parents produce visible products (the b jets and the leptons) and invisible products (the neutrinos coming from the W or lepton decay). The objects involved are denoted as follow: $\vec{b}$, $\vec{b'}$ and m_b, $m_{b'}$ indicate the vector momenta of the two b jet candidates; the quantities corresponding to the leptons and neutrinos are globally denoted as $\vec{c}$,$\vec{c'}$ and m_c, $m_{c'}$. The m_{T2} variable is thus constructed as

$$m_{\text{T2}}(m_B, m_{B'}, \vec{\Sigma}_T, m_c, m_{c'}) = \min_{\vec{c}_{\text{T}} + \vec{c'}_{\text{T}} = \vec{\Sigma}_T} [\max(m_{\text{T}}, m_{\text{T'}})], \tag{4.12}$$

where the kinematic constraint is represented by the minimization over $\vec{\Sigma}_{\text{T}}$, i.e. the sum of the measured lepton momenta and the missing transverse momenta. The minimisation needed for the m_{T2} computation is performed using the method provided in [117].

Two topological discriminant variables [119] related to the momentum of the objects are defined as

$$p_\zeta = (\vec{p}_{\text{T}}(\ell) + \vec{p}_{\text{T}}(\tau_{\text{h}}) + \vec{p}_{\text{T}}^{miss}) \cdot \hat{\zeta} \qquad \text{and} \qquad p_\zeta^{vis} = (\vec{p}_{\text{T}}(\ell) + \vec{p}_{\text{T}}(\tau_{\text{h}})) \cdot \hat{\zeta} \tag{4.13}$$

where on the axis $\hat{\zeta}$ is the direction of the bisector of the $\vec{p}_{\text{T}}$ of the two reconstructed tau leptons. The distance between the neutrino produced by the tau lepton decay and the visible tau lepton products is typically small; therefore, the missing momentum vector $\vec{p}_{\text{T}}^{miss}$ points in the direction of the reconstructed tau leptons.

Finally, angular variables are computed among all the reconstructed objects, also in different reference frames.

The final set of input variables for the BDT is identified by taking into account the similarities between all the potential discriminating observables and their capability to bring informations as different as possible. The input variables selection is done, as detailed in [109], by ranking them through a statistical method known as "Jensen Shannon divergence" (JSD), based on the "Kullback-Leibler divergence" [120].

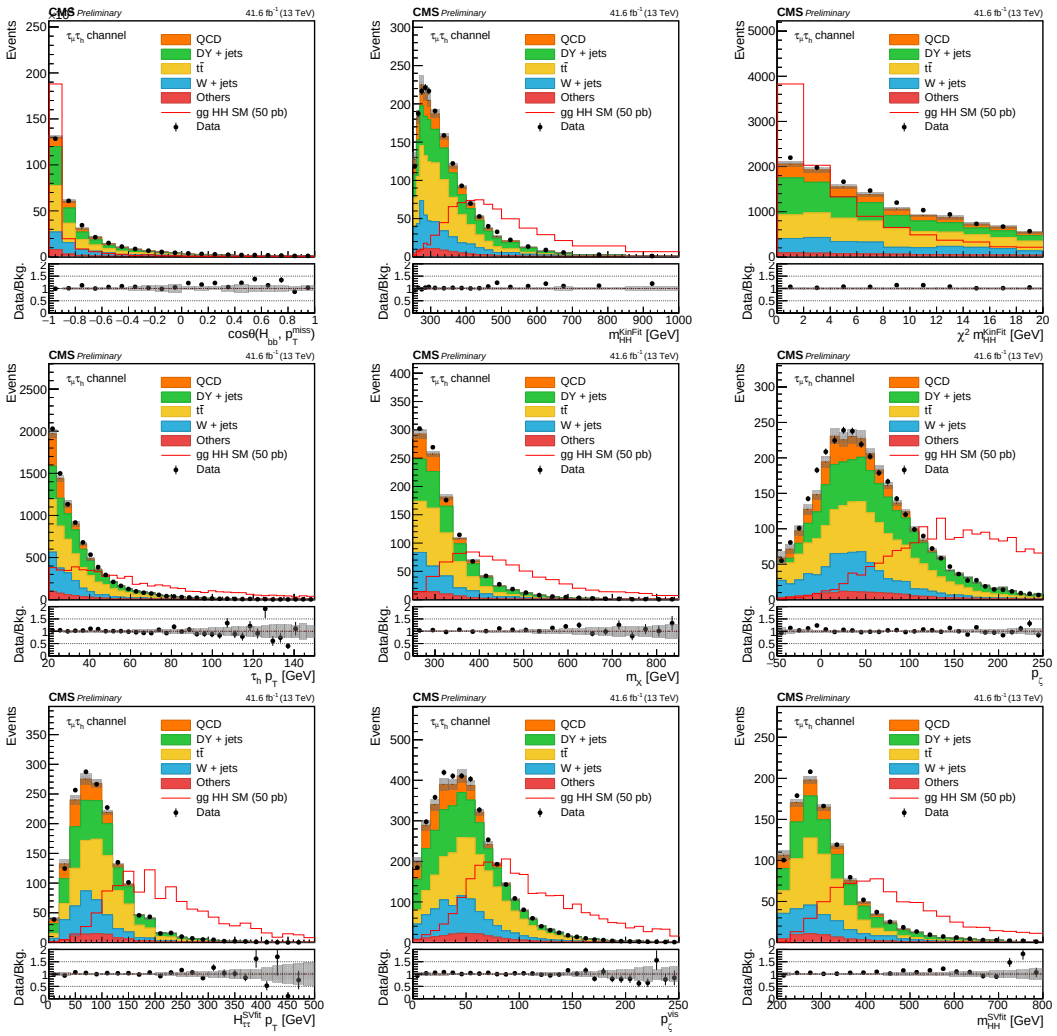


Figure 4.27 – Distribution of events passing the baseline selection and the elliptic mass cut in the $\tau_\mu\tau_\mathrm{h}$ channel, as a function of the input variables of the BDT training. Their description can be found in Tab. 4.8. The gluon fusion signal is represented; the corresponding $\sigma \cdot \mathrm{BR}$ is normalized to 50 pb to facilitate its visualisation.

The twenty variables thus selected are listed in Tab. 4.8, ordered by importance as defined in Sec. 4.6.3. Their distribution show a satisfactory data-over-prediction agreement, as shown in Fig. 4.27 and Fig. 4.28, which is essential for the consistency of the BDT response. To preserve suitable statistics, the BDT is trained using events from the three final states simultaneously, merging gluon fusion signal samples corresponding to different k_λ hypotheses. However, this choice can impact negatively the discrimination power within each channel and in different regimes: as argued in [44], the production involving anomalous couplings changes drastically the kinematics of the process. Therefore, k_λ and the $\tau\tau$ pair type are introduced as input variables; this expedient is called "parametrised learning".

4.6.2 Training

The BDT configuration is tuned through a "grid search": a set of values is identified for the most relevant hyperparameters of the algorithm (e.g., the number of trees or the maximum tree depth); the performance of the algorithm is evaluated trying all the possible combinations of the hyperparameter values to be tested. To minimise the risk

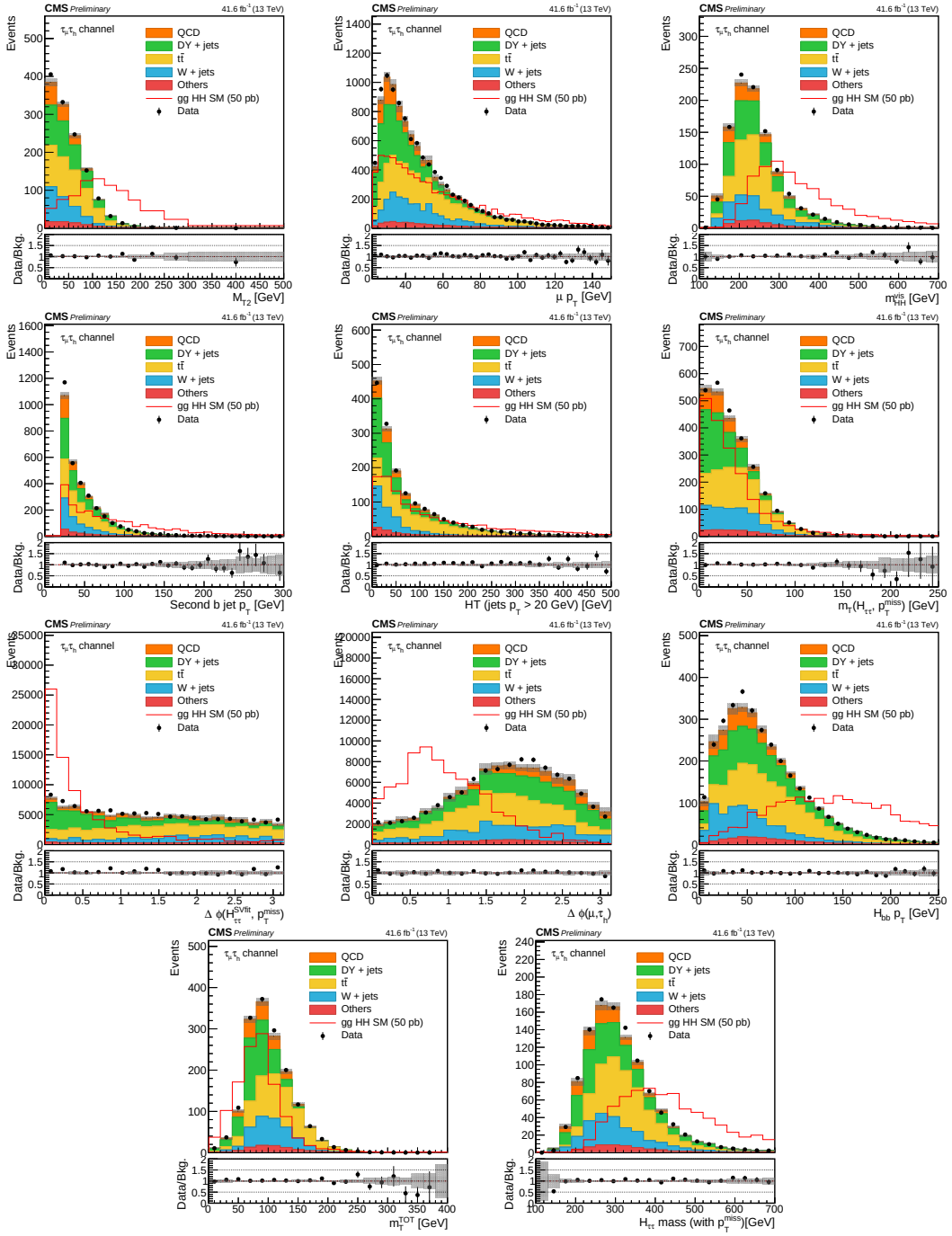
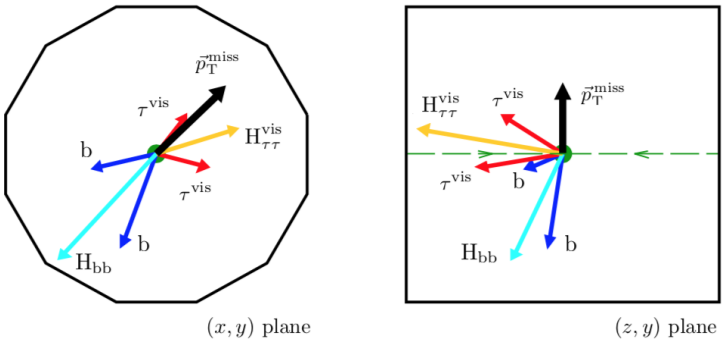


Figure 4.28 – Distribution of events passing the baseline selection and the elliptic mass cut in the $\tau_\mu\tau_\mathrm{h}$ channel, as a function of the input variables of the BDT training. Their description can be found in Tab. 4.8. The gluon fusion signal is represented; the corresponding $\sigma \cdot \mathrm{BR}$ is normalized to 50 pb to facilitate its visualisation.

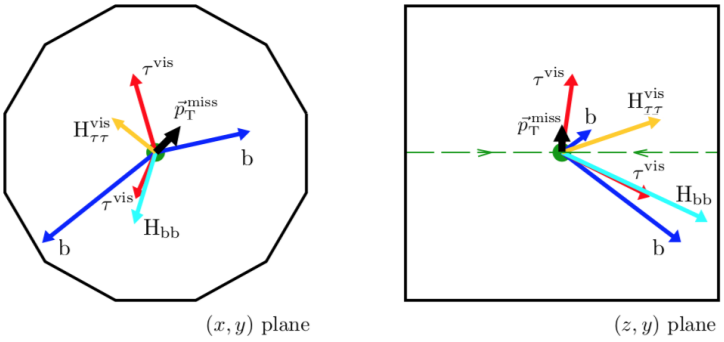
of overtraining, occurring when a machine learning algorithm such as the BDT learns from the statistical fluctuations of the training data set, the data set is split in two independent subsets. Within the primary subset, a cross validation method is applied and a search over the grid of the hyperparameters is performed; the configuration that provides the largest area under the ROC curve, which is taken as a measurement of the performance, is chosen. The corresponding hyperparameters are listed in Tab. 4.9.

The equivalent procedure is replicated on the secondary data set; as the optimization



(x, y) plane (z, y) plane

(a) HH event



(x, y) plane (z, y) plane

(b) $t\bar{t}$ event

Figure 4.29 – Graphical representation of two simulated events. In the upper row, a HH event is shown, and in the lower row a $t\bar{t}$ event is shown. Each event is represented in the transverse plane (left) and in the (z, η) plane (right) with respect to the beam line. Blue arrows denote b jets, red arrows τ leptons visible decay products and black arrow missing transverse momentum. Cyan and orange arrows denote the Higgs bosons reconstructed from the bb and $\tau\tau$ systems, respectively. Their lengths are proportional to the magnitude of the spatial momentum of the corresponding object [118].

points to the same configuration chosen using the primary data set, the procedure is validated.

4.6.3 Performance

The importance of each variable is determined as the number of times that the BDT algorithm uses it for the splitting of a binary tree; each splitting occurrence is weighted by the square of the gain achieved by that separation and the number of events in the node. The position of each variable by importance is shown in Tab. 4.8. The *score* resulting by the evaluation of the BDT algorithm on the features of the events that it evaluates is given between -1 and 1; the background events should have scores distributed towards the value of -1, and the distribution of signal events should peak at 1. The event distribution in the baseline selection and after the application of the elliptic mass cut are shown in Fig. 4.30 as a function of a the BDT output corresponding to a few κ_λ values and in the three final states. An excellent agreement is achieved over all the range in $\tau_e\tau_h$ and $\tau_\mu\tau_h$. The $\tau_h\tau_h$ channel suffers from a suboptimal modelling in the regions populated by genuine tau leptons (see Appendix A); however, the data-over-prediction agreement is satisfactory overall and very good for high BDT score, i.e. in the most sensitive bins.

Table 4.8 – Lists of chosen input variables used by the BDT for the $t\bar{t}$ background rejection, ordered by their importance determined in the training phase.

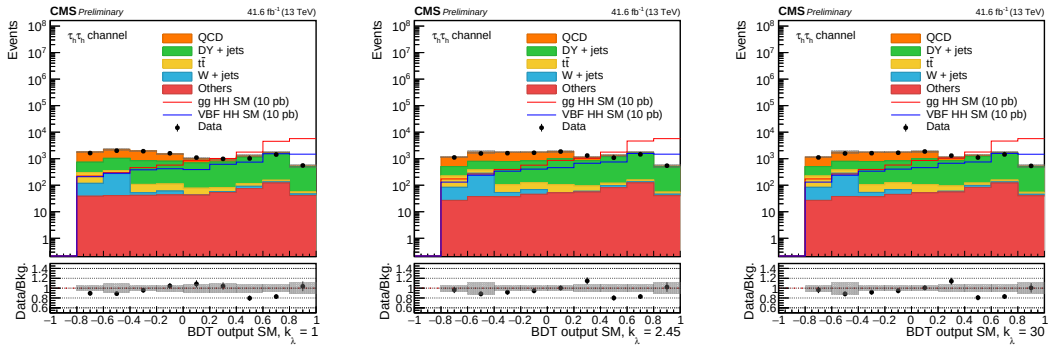
Variable	Description
p_ζ	Projection of the transverse momenta of the $\tau\tau$ candidates and the missing momentum along the direction of the $\tau\tau$ pair, defined by Eq. 4.13
$m_{\mathrm{T}}^{\mathrm{TOT}}$	Total transverse mass, defined by Eq. 4.11
$\chi^2(m_{\mathrm{HH}}^{\mathrm{KinFit}})$	χ^2 associated to the HH mass computation through KinFit
$m_{\mathrm{HH}}^{\mathrm{KinFit}}$	Mass of the HH system computed with the KinFit algorithm [115]
$p_{\mathrm{T}}(\tau_{\mathrm{h}})$	Transverse momentum of the (second) selected τ_{h} leg
$m(\mathrm{HH}^{p_{\mathrm{T}}^{\mathrm{miss}}})$	Invariant mass of the HH system, using the visible objects and the $p_{\mathrm{T}}^{\mathrm{miss}}$ for the reconstruction of the $\mathrm{H}_{\tau\tau}$ candidate
κ_λ	Trilinear coupling expressed as $\lambda_{\mathrm{HHH}}/\lambda_{\mathrm{HHH}}^{\mathrm{SM}}$; for parametrized learning
p_ζ^{vis}	Projection of the transverse momenta of the $\tau\tau$ candidates and the missing momentum along the direction of the $\tau\tau$ pair, defined by Eq. 4.13
$p_{\mathrm{T}}(\mathrm{H_{bb}})$	Transverse momentum of the $\mathrm{H_{bb}}$ candidate
$p_{\mathrm{T}}(\mathrm{H}_{\tau\tau}^{\mathrm{SVfit}})$	Transverse momentum of the $\mathrm{H}_{\tau\tau}$ candidate computed through SVfit
$m_{\mathrm{T}}(\mathrm{H}_{\tau\tau}, p_{\mathrm{T}}^{\mathrm{miss}})$	Transverse mass of the $\mathrm{H}_{\tau\tau}$ system
$p_{\mathrm{T}}(\ell)$	Transverse momentum of the first leg of the $\tau\tau$ pair
$\Delta\phi(H_{\tau\tau}^{SVfit}, p_{\mathrm{T}}^{\mathrm{miss}})$	Angular separation in the transverse plane between the $\mathrm{H}_{\tau\tau}$ system reconstructed through SVfit and the missing transverse momentum
$m(\mathrm{HH}^{vis})$	Invariant mass of the HH system, using the visible objects for the reconstruction of the $\mathrm{H}_{\tau\tau}$ candidate
$m(\mathrm{HH}^{\mathrm{SVfit}})$	Invariant mass of the HH system, using the SVfit algorithm for the reconstruction of the $\mathrm{H}_{\tau\tau}$ candidate
m_{X}	Reduced mass of the HH system, defined by the Eq. 4.9
final state	$\tau\tau$ pair type; for parametrized learning
$\cos\theta(\mathrm{H_{bb}}, p_{\mathrm{T}}^{\mathrm{miss}})$	Cosinus of the angle between the $\mathrm{H_{bb}}$ candidate and the $p_{\mathrm{T}}^{\mathrm{miss}}$ in the direction of flight of $\mathrm{H_{bb}}$ in its rest frame
p_{T}(b jet 2)	Transverse momentum of the second selected b jet candidate
$\Delta\phi(\ell, \tau_{\mathrm{h}})$	Angular separation in the transverse plane between the $\tau\tau$ candidates
m_{T2}	Stransverse mass [116], defined by Eq. 4.12
H_{T}	Sum of the transverse momenta of all the additional jets with $p_{\mathrm{T}} > 20\,\mathrm{GeV}$

4.7 VBF categories

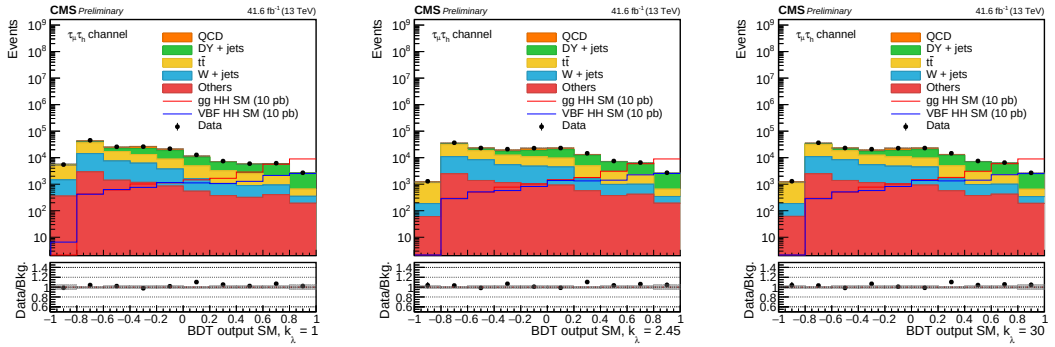
The VBF HH → $b\bar{b}\tau\tau$ signal extraction has several experimental challenges. The most obvious is given by the rarity of the VBF HH production: its cross section is 1.6 fb, i.e. about 20 times smaller than the gluon fusion and, for instance, six orders of magnitude

Table 4.9 – Selected hyperparameters values. Their meaning is clarified in [109].

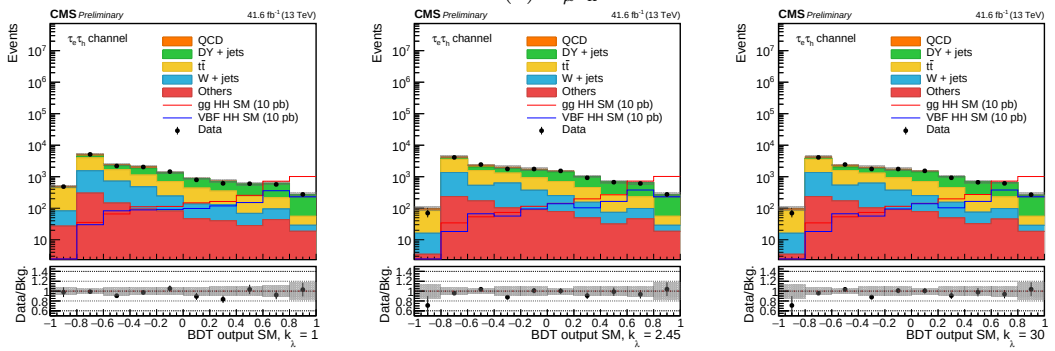
Name	Value
Number of trees	700
Maximum tree depth	3
Minimum node size	0.03
Number of cuts	500
Shrinkage	0.05
Bagged sample fraction	0.5



(a) $\tau_h\tau_h$



(b) $\tau_\mu\tau_h$



(c) $\tau_e\tau_h$

Figure 4.30 – Distributions of events passing the baseline selection for each channel and the elliptic mass cut, as a function of the BDT score corresponding to $k_\lambda = 1$ (SM), 2.45 (maximum interference) and 30.

smaller than the $t\bar{t}$ background. Moreover, as shown in Sec. 4.2, the leptons originating from the decay of a Higgs boson produced through the VBF HH mechanism have softer p_T spectra than the ones resulting from the gluon fusion HH production; therefore, the

$H \to \tau\tau$ selection described in Sec. 4.3.5 suppresses the VBF HH signal more than the one of gluon fusion. Gluon fusion $HH \to bb\tau\tau$ events accompanied by jets, which can mimic the VBF topology, are expected to give a large contamination in the VBF signal region. On one hand, the largest possible acceptance on the VBF $HH \to bb\tau\tau$ signal is needed; on the other hand, even tight selections bring in the VBF region a large gluon fusion signal, as it has a larger cross section and it is selected more efficiently.

The VBF category strategy can be outlined in three points. Firstly, the VBF topology must be efficiently exploited at every step of the event selection. The design of a L1 VBF trigger and, subsequently, of the HLT VBF $H \to \tau\tau$ path goes in this direction; the Ch. 3 is dedicated to this topic. Secondly, a VBF HH event selection consistent with the signal kinematics is designed. The most discriminating variables are the invariant mass m_{jj} of the VBF jets and their angular separation $|\Delta\eta_{jj}|$; hence, they are important handles in the VBF event selection. The corresponding thresholds are tuned targeting a good compromise between a large VBF HH signal acceptance and an efficient background rejection. The VBF category selection is only optimised using HH production SM signals.

Finally, in view of specific VBF HH studies, a VBF vs. gluon fusion signal disambiguation can be attempted, for example exploiting a machine learning technique. Indeed, the contribution form the gluon fusion HH production is irreducible, having the same final state of the signal under consideration. A preliminary DNN-based strategy is presented, along with the results, in Sec. 6.6.2.

4.7.1 VBF event selection

As argued in Sec. 4.2, a VBF category containing events firing the VBF $H \to \tau\tau$ path needs to be exclusive. The VBF $H \to \tau\tau$ path covers the region with very high m_{jj}, with tight p_T thresholds on the VBF jet candidates. Therefore, it corresponds to a region that is very pure from the background contamination. However, the statistics is very limited; moreover, it was only enabled for about 27 over the $42\,fb^{-1}$ of collisions recorded by CMS during 2017. Since the kinematic distributions of the events collected by the VBF $H \to \tau\tau$ path and the di-τ_h path are different, making separate categories is a natural choice; in addition, given their different kinematic features, the computation of the trigger efficiency of their logic OR of is not trivial. Hence, a tight VBF category in the $\tau_h\tau_h$ channel is populated only by events firing the dedicated trigger. A looser category, relying on the regular triggers, is designed for the $\tau_h\tau_h$, $\tau_\mu\tau_h$ and $\tau_e\tau_h$ channel.

To pass any of the VBF selections, an event should have at least two b jet candidates, not necessarily b-tagged; by construction, given the assignment described in Sec. 4.4.3, it is the case anytime a pair of VBF jet candidates is found. In order to preserve a large statistics and for consistency with the jet sorting procedure, it is sufficient that one of the b jet candidates passes the *Medium* DeepCSV score.

In Fig. 4.31, the distributions of generated VBF quarks in a signal simulation sample is shown as a function of their separation $|\Delta\eta_{jj}|$ and their invariant mass m_{jj}. While the thresholds of the tight VBF category in the $\tau_h\tau_h$ channel are driven by the selection applied at trigger level, those of the loose VBF category can be optimised using the offline information only; a simple squared selection in the m_{jj} vs. $|\Delta\eta_{jj}|$ plane is applied to events that have two VBF jet candidates.

In Fig. 4.32, the event yield normalised to the integrated luminosity recorded in 2017, i.e. $41.6\,fb^{-1}$, is represented for the VBF HH signal, the gluon fusion HH signal and the sum of all the backgrounds in the $\tau_h\tau_h$ channel and in each of the jet categories. The angular separation threshold is fixed to $|\Delta\eta_{jj}| > 3$, while the m_{jj} threshold varies along

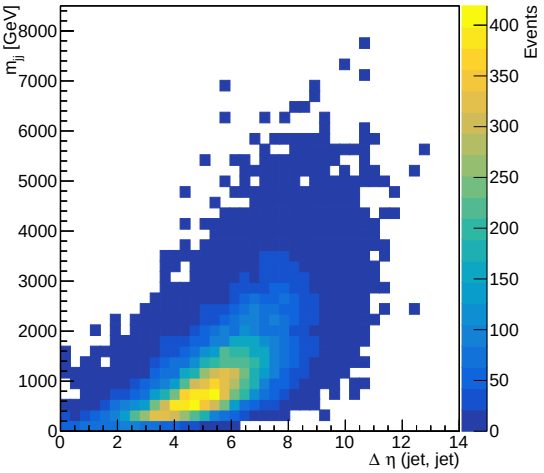


Figure 4.31 – Event distribution as a function of the invariant mass and the angular separation between the two generated VBF quarks in a SM VBF signal simulation.

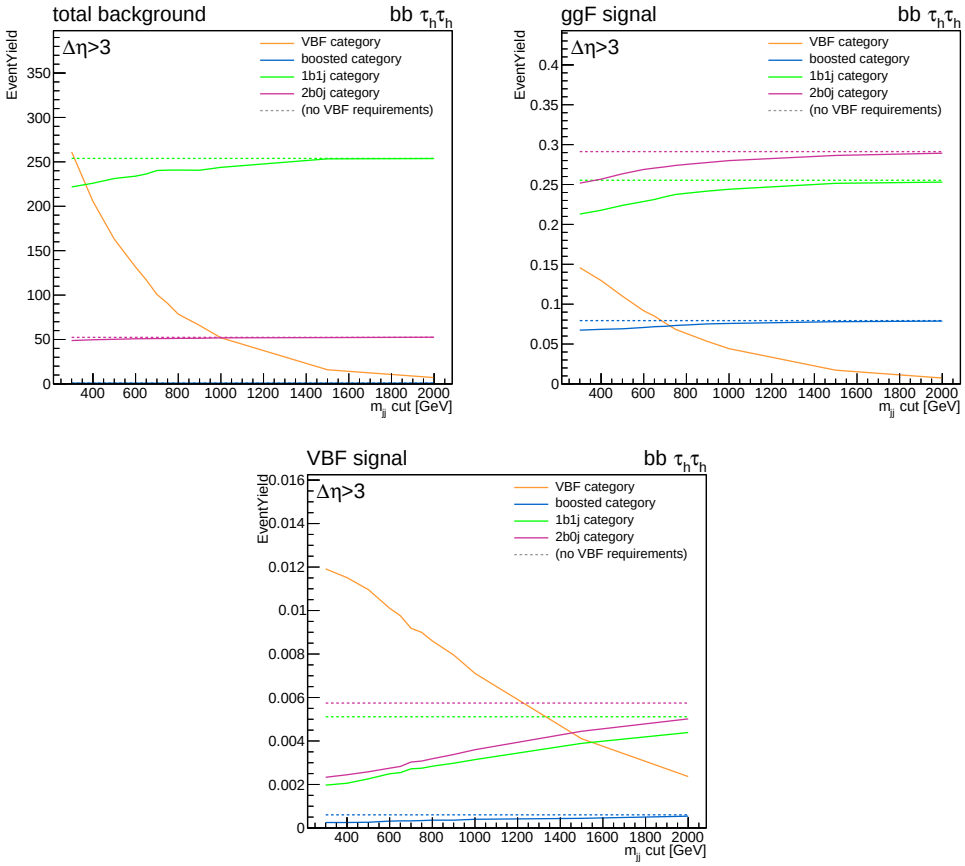


Figure 4.32 – Number of background, gluon fusion signal and VBF signal $\tau_h\tau_h$ events in each category, as a function of the m_{jj} threshold of the VBF category selection. The dashed lines correspond to the number of events in the inclusive categories if no VBF category was implemented. Only events firing the di-τ_h triggers are considered. The event yield is normalised to an integrated luminosity of 41.6 fb^{-1}.

the x axis. The solid orange line represents the number of events in the VBF category thus defined. The solid violet, green and blue lines correspond to the event yield in the

resolved 2b0j, resolved 1b1j and boosted categories after the application of the mass cut introduced in Sec. 4.5. Since the VBF selection has priority on the inclusive categories, their event yield depends on the invariant mass of the additional jets: the higher the m_{jj} cut, the smaller the event yield in the VBF category and, thus, the bigger the event yield in the inclusive ones. The dashed lines represent the event yield in the b jet categories described in Sec. 4.4.4 in the scenario of the selection used in the 2016 analysis, i.e. when no VBF category is implemented. The dashed lines are constant, as their definition does not change as a function of the m_{jj} cut. By construction, the solid lines tend to the dashed ones for tight m_{jj} thresholds, i.e. when progressively reducing the phase-space of the VBF category.

A different composition of the events populating the four categories can be observed from the plots in Fig. 4.32. The VBF HH signal events are predominantly collected by the VBF category; most of them would be collected in the inclusive categories if no VBF category was implemented. A small fraction of gluon fusion events enters the VBF selection; these events mainly share the features of those collected in the resolved inclusive categories. Although the gluon fusion contamination rapidly decreases as a function of the m_{jj} threshold, it is very large compared to the event yield of the VBF HH signal. Finally, the background contribution is large for loose m_{jj} threshold and decreases in the high invariant mass region.

In Fig. 4.33, the significance of the VBF and of the gluon fusion signal over the sum of the backgrounds, computed as $S/\sqrt{S+B}$, is shown as a function of the thresholds placed on the m_{jj} and the $|\Delta\eta_{jj}|$ of the VBF candidates. As expected, the significance of the gluon fusion signal is higher with looser selections; as for the VBF signal, instead, the tighter the thresholds the higher the sensitivity.

It is legitimate to ask whether the implementation of a VBF category in addition to the consolidated inclusive analysis categories improves the global sensitivity on HH signals or not; in other words, one should question where should the m_{jj} and $|\Delta\eta_{jj}|$ edges of the VBF selection should be placed for signal events to be worth being collected in the VBF categories rather than in the inclusive ones.

A few pieces of information come from preliminary studies carried out with 2016 data sets, summarised in the following. It was observed that the gluon fusion contamination in the VBF category is so large that optimizing aiming at a global improvement of the sensitivity on the HH signal is basically equivalent to optimizing targeting an improvement on the gluon fusion signal only. This observation drives the VBF category selection towards tight m_{jj} cuts: only for m_{jj} thresholds larger than 1 TeV the combination of VBF and inclusive categories starts to be competitive over the use of the inclusive categories only. However, the acceptance on the VBF signal itself is penalised by such a tight selection; moreover, the integrated luminosity in 2017 only is not large enough to have a significant number of events passing that hypothetical selection.

Finally, it should be noted that the gluon fusion signal simulations provided centrally in CMS account only for the processes where no additional jets are produced in the hard interaction; hence, the additional jets that allow gluon fusion signal events to pass the VBF selection are mainly produced by pileup and parton shower effects. Therefore, it is more consistent to target a large efficiency on the VBF HH selection and reject at best the gluon fusion signal contamination.

To preserve large statistics in the VBF categories, the selection chosen for the loose VBF categories is $m_{jj} > 500\,\text{GeV}$ and $|\Delta\eta_{jj}| > 3$. The final VBF categories selections are summarised in Tab. 4.10.

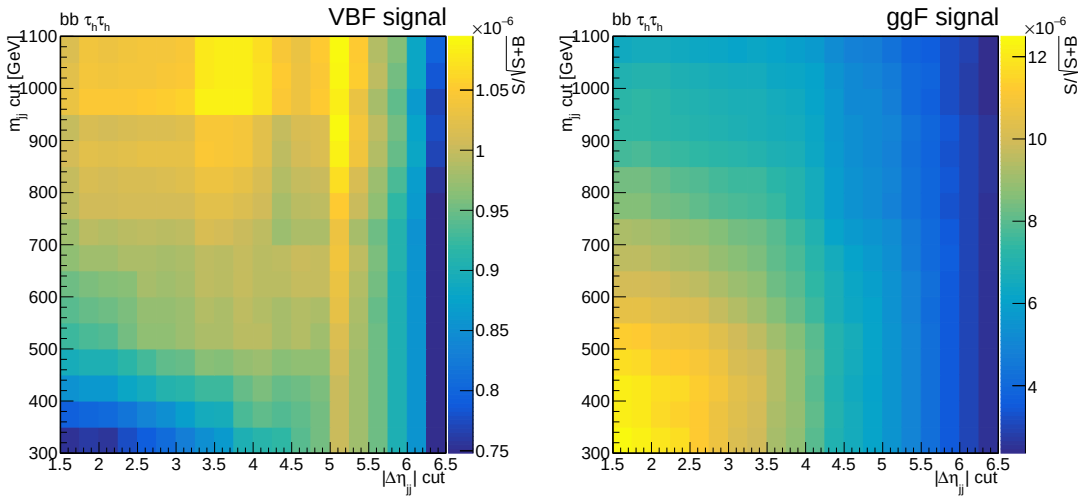


Figure 4.33 – Significance $S/\sqrt{S+B}$ of the VBF signal (left) and the gluon fusion signal (right) computed with events in the $\tau_h\tau_h$ channel passing the baseline selection and with at least one b-tagged jet, shown as a function of the m_{jj} and the $|\Delta\eta_{jj}|$ of the VBF jet candidates. Only events firing the di-τ_h triggers are considered.

Table 4.10 – Offline jets selection in the VBF categories. The tight VBF category only contains events that fired the VBF H $\to \tau\tau$ trigger.

$\tau\tau$ pair type	VBF	
all	2 b jet candidates with $p_T > 20\,\text{GeV}$ and $\lvert\eta\rvert < 2.4$ and passing *Tight* ID, *Loose* pileup ID at least one b jet candidate has *Medium* DeepCSV tag at least two additional jets with $p_T > 20\,\text{GeV}$ and *Tight* ID	
$\tau_e\tau_h$	**Loose VBF** $m_{jj} > 500\,\text{GeV}$ and $\lvert\Delta\eta_{jj}\rvert > 3$	
$\tau_\mu\tau_h$	**Loose VBF** $m_{jj} > 500\,\text{GeV}$ and $\lvert\Delta\eta_{jj}\rvert > 3$	
$\tau_h\tau_h$	**Loose VBF** $m_{jj} > 500\,\text{GeV}$ and $\lvert\Delta\eta_{jj}\rvert > 3$, fail Tight VBF selection	**Tight VBF** jets $p_T > 140\,\text{GeV}$ and $60\,\text{GeV}$, $m_{jj} > 800\,\text{GeV}$, $\lvert\Delta\eta_{jj}\rvert > 3$

Chapter 5

Background and signal modelling

The exploration of various BSM scenarios requires the modelling of the signal for several different sets of couplings. Since only a limited set of simulated samples can be produced, weighting techniques are implemented both for the gluon fusion and the VBF signals, in order to model additional BSM scenario starting from a small set of fully simulated BSM signals. The corresponding modelling strategies are detailed in Sec. 5.1 and Sec. 5.3. The modelling of BSM gluon fusion signals is performed through the consolidated method used in the previous HH $\rightarrow$ bb$\tau\tau$ analysis [107] and it is shared by all the CMS HH searches; the VBF modelling strategy, emerged from discussions within the HH group and with colleagues from the ATLAS experiment, is tested and implemented for the first time within this search.

A brief description of the backgrounds is given in Sec. 4.1; a list of the processes and their cross section is given in Tab. 5.1. In this chapter, the modelling techniques of the major backgrounds are described. Most of the processes are entirely modelled from simulated samples, making use of the generators listed in Tab. 5.1; the exceptions are the QCD multijet background estimation (Sec. 5.3), which is data-driven, and the Drell-Yan processes modelling (Sec. 5.4), which is tuned using weights computed in control regions. The performance of the QCD estimation was already validated in the previously published results; the Drell-Yan modelling, whose strategy is detailed in [121], is improved by applying p_{T}-dependent corrections.

5.1 Gluon fusion HH signal modelling

The weighting technique for the gluon fusion modelling allows various scenarios to be explored in terms of the effective Lagrangian parametrization described in [44], where the Higgs pair production is regulated by the five couplings y_{t}, λ_{HHH}, c_2, c_{2g} and c_g; the variations from the SM values of the standard model are expressed as $k_\lambda = \lambda_{\mathrm{HHH}}/\lambda_{\mathrm{HHH}}^{\mathrm{SM}}$ and $k_{\mathrm{t}} = y_{\mathrm{t}}/y_{\mathrm{t}}^{\mathrm{SM}}$.

The two Higgs bosons are produced back-to-back in the reference frame of the center of mass; before any hadronization effect, they have identical transverse momenta and opposite azimuthal angle. At this level, the kinematics of the event is totally determined by two parameters: the invariant mass of the HH system and the angle $\cos\theta^*$ between one Higgs boson and the beam axis [44]. These variables are exploited by the weighting procedure.

For each of the identified BSM benchmarks, signal samples are produced centrally by CMS at leading order (LO) precision with MadGraph5_aMC@NLO [47]; the seven

Table 5.1 – Background processes and corresponding cross sections; details on the generators can be found in [47, 122].

Process	Modelling	Cross section [pb]
W$\to \ell\nu_\ell$ + jets	MadGraph5_aMC@NLO, NLO precision	6.15×10^4
Z/$\gamma^* \to \ell\ell$ + jets	MadGraph5_aMC@NLO, LO precision	6225.42
Electroweak	MadGraph5_aMC@NLO, LO precision	
W^+ + jj		29.69
W^- + jj		20.25
Z + jj		3.98
$t\bar{t}$	powheg, NLO precision	832.71
QCD multijet	Data-driven (see Sec. 5.3)	
Single t($\bar{t}$)	powheg, NLO precision	
W channel		35.9(35.9)
t-channel		80.95(136.02)
WW	MadGraph5_aMC@NLO, LO precision	
$\to 2\ell 2\nu$		12.18
$\to 2\ell 2q$		50.0
$\to 4q$		51.7
ZZ	MadGraph5_aMC@NLO, LO precision	
$\to 2\ell 2\nu$		0.564
$\to 2\ell q\nu$		3.22
$\to 4\ell$		1.21
$\to 4q$		7.06
ZW	MadGraph5_aMC@NLO, LO precision	
$\to \ell\ell\ell\nu$		0.564
$\to \nu\nu\ell\nu$		3.22
$\to qq\ell\nu$		1.21
$\to \ell\ell qq$		7.06
WWW	MadGraph5_aMC@NLO, LO precision	0.21
ZZZ	MadGraph5_aMC@NLO, LO precision	0.014
WWZ	MadGraph5_aMC@NLO, LO precision	0.16
WZZ	MadGraph5_aMC@NLO, LO precision	0.056
ZH	powheg, NLO precision	0.884

samples used in this search are listed in Tab. 5.2. It should be mentioned that the signal description is not ideal: the ratio between the production cross section computed at LO over that computed at next-to-leading-order (NLO) of the perturbative expansion varies by about 35% over the range $-1 < k_\lambda < 5$ [123]; signal samples at NLO precision will be produced for future analyses.

The events of all the gluon fusion HH signal samples are combined to build a 2D distribution as a function of m_{HH} and $|\cos\theta^*|$, computed using simulated Higgs boson properties after the hard scatter and before hadronization effects; an identical histogram is filled using the SM signal sample only. The content of a bin j in the two bidimensional distributions, normalised to unity, is denoted as f^j_{comb} and f^j_{SM}.

Table 5.2 – Combinations of the couplings available in gg→HH → bb$\tau\tau$ simulation data sets. All samples are generated centrally by CMS with MadGraph5_aMC@NLO at LO precision and each contains about 300M events. The identification of the benchmarks is discussed in [44].

Benchmark	k_λ	k_{t}	c_2	c_g	c_{2g}
2	1.0	1.0	0.5	-0.8	0.6
3	1.0	1.0	-1.5	0.0	-0.8
4	-3.5	1.5	-3.0	0.0	0.0
7	5.0	1.0	0.0	0.2	-0.2
9	1.0	1.0	1.0	-0.6	0.6
12	15.0	1.0	1.0	0.0	0.0
SM	1.0	1.0	0.0	0.0	0.0

The ratio of the total HH cross section over the SM prediction, whose parametrization can be found in Eq. 1.60, can be expressed as a function of the j bin number as

$$\begin{aligned} R^j_{\mathrm{HH}} = \frac{\sigma^j_{\mathrm{HH}}}{\sigma^{j,\mathrm{SM}}_{\mathrm{HH}}} = {} & A^j_1 k^4_{\mathrm{t}} + A^j_2 c^2_2 + (A^j_3 k^2_{\mathrm{t}} + A^j_4 c^2_g) k_\lambda + A^j_5 c^2_{2g} + \\ & + (A^j_6 c_2 + A^j_7 k_\lambda k_{\mathrm{t}}) k^2_{\mathrm{t}} + (A^j_8 k_{\mathrm{t}} k_\lambda + A^j_9 c_g k_\lambda) c_2 + \\ & + (A^j_{10} c_2 c_{2g} + (A^j_{11} c_g k_\lambda + A^j_{12} c_{2g}) k^2_{\mathrm{t}} + \\ & + (A^j_{13} k_\lambda c_g + A^j_{14} c_{2g}) k_{\mathrm{t}} k_\lambda + A^j_{15} c_g c_{2g} k_\lambda. \end{aligned} \tag{5.1}$$

The ratio R^j_{HH} is computed using simulated events with different sets of couplings; thus, the A^j_i coefficients are extracted from its interpolation as a function of the couplings.

Finally, event-by-event weights are computed as

$$\omega = \frac{\Omega}{\sum_n \Omega} \tag{5.2}$$

where

$$\Omega(k_\lambda, k_{\mathrm{t}}, c_2, c_g, c_{2g}; j) = \frac{f^j_{\mathrm{SM}}}{f^j_{comb}} \cdot \frac{R^j_{\mathrm{HH}}(k_\lambda, k_{\mathrm{t}}, c_2, c_g, c_{2g})}{R_{\mathrm{HH}}(k_\lambda, k_{\mathrm{t}}, c_2, c_g, c_{2g})} \tag{5.3}$$

and the sum goes over the number of simulated events; thus, only the differential event distribution is modified and not the global normalization.

5.2 VBF HH signal modelling

The HH production cross section via VBF can be written as the square of the amplitude of the LO diagrams represented in Fig. 5.1, i.e.

$$\begin{aligned}\sigma(c_{\mathrm{V}}, c_{2\mathrm{V}}, k_\lambda) \sim |\mathrm{A}c_{\mathrm{V}}k_\lambda + \mathrm{B}c_{\mathrm{V}}^2 + \mathrm{C}c_{2\mathrm{V}}|^2 = \\ = \mathrm{a}c_{\mathrm{V}}^2k_\lambda^2 + \mathrm{b}c_{\mathrm{V}}^4 + \mathrm{c}c_{2\mathrm{V}}^2 + \mathrm{i}_{\mathrm{ab}}c_{\mathrm{V}}^3k_\lambda + \mathrm{i}_{\mathrm{ac}}c_{\mathrm{V}}c_{2\mathrm{V}}k_\lambda + \mathrm{i}_{\mathrm{bc}}c_{\mathrm{V}}^2c_{2\mathrm{V}}\end{aligned} \tag{5.4}$$

where $\mathrm{a} = |\mathrm{A}|^2$, $\mathrm{b} = |\mathrm{B}|^2$, $\mathrm{c} = |\mathrm{C}|^2$ and i_{ij} are the interference terms. Therefore, the cross section, as well as any differential distribution $\mathrm{d}\sigma/\mathrm{d}x$, depends on six components. Thus, the VBF signal can be modelled through the sum of six components

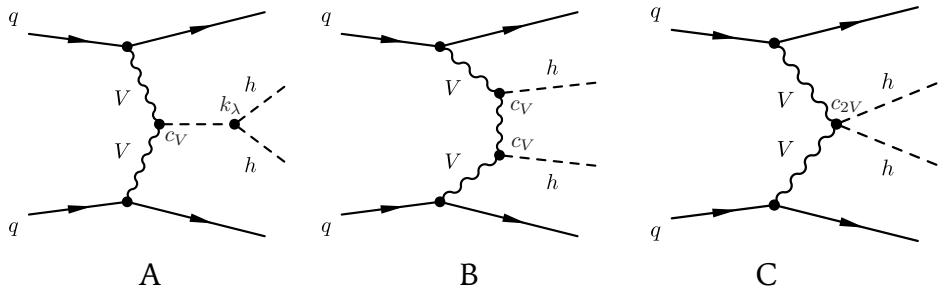


Figure 5.1 – Leading order diagrams participating to the HH production via VBF.

$\mathbf{V} = \{\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{i}_{\mathrm{ab}}, \mathrm{i}_{\mathrm{ac}}, \mathrm{i}_{\mathrm{bc}}\}$, each scaled by a function of c_{V}, $c_{2\mathrm{V}}$ and k_λ; denoting as $\mathbf{K} = \{c_{\mathrm{V}}^2k_\lambda^2, c_{2\mathrm{V}}^4, c_{2\mathrm{V}}^2, c_{\mathrm{V}}^3k_\lambda, c_{\mathrm{V}}c_{2\mathrm{V}}k_\lambda, c_{\mathrm{V}}^2c_{2\mathrm{V}}\}$ the vector of the functions of the couplings, the Eq. 5.4 can be expressed as

$$\sigma = \mathbf{K}^{\mathrm{T}}\mathbf{V}. \tag{5.5}$$

However, the generator used does not allow for generating the individual $\mathbf{V}_i$ components. Instead, they can be determined by solving a system of equations using six samples corresponding to different combinations of $(c_{\mathrm{V}}, c_{2\mathrm{V}}, k_\lambda)$. Denoting these samples as $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5, \sigma_6\}$, where $\sigma_i = \sigma(c_{\mathrm{V},i}, c_{2\mathrm{V},i}, k_{\lambda,i})$, they can be represented as

$$\boldsymbol{\sigma} = \mathbf{M}\ \mathbf{V} \tag{5.6}$$

where $\mathbf{M}$ is the 6×6 coefficients matrix; its solution is

$$\mathbf{V} = \mathbf{M}^{-1}\boldsymbol{\sigma}. \tag{5.7}$$

Thus, the cross section σ_{target} of a given $(c_{\mathrm{V}}, c_{2\mathrm{V}}, k_\lambda)$ combination can be computed as

$$\sigma_{target} = [\mathbf{K}^{\mathrm{T}}\mathbf{M}^{-1}]\boldsymbol{\sigma}. \tag{5.8}$$

The Eq. 5.8 can be equally applied to build the differential distribution as a function of a given observable and for a given $(c_{\mathrm{V}}, c_{2\mathrm{V}}, k_\lambda)$ combination; in that case, the unknowns $\mathbf{V}(x)$ are a function of the observable x and

$$h(x)_{target} = [\mathbf{K}^{\mathrm{T}}\mathbf{M}^{-1}]\mathbf{h}(x), \tag{5.9}$$

where $\mathbf{h}(x)$ contains the differential distributions. Thus, the shape of a signal can be easily obtained by manipulating a few input histograms, rather than going through an event-by-event reweighting; to do so, only six fully simulated combinations of $(c_{\mathrm{V}}, c_{2\mathrm{V}}, k_\lambda)$ are needed.

Five of such signal samples are provided centrally by CMS; a sixth signal sample was generated privately using MadGraph5_aMC@NLO. The combinations of couplings available in fully simulated events and the corresponding cross sections are listed in Tab. 5.3. In Tab. 5.4, the elements of **V** resulting from Eq. 5.7, using the cross section of those samples, are listed.

Table 5.3 – Combinations of the (c_V, c_{2V}, k_λ) couplings available in VBF HH $\to$ bb$\tau\tau$ simulation data sets. The data sets corresponding to the first five combinations are provided centrally by CMS; the last one was produced privately. All samples are generated with MadGraph5_aMC@NLO at LO precision and each contains 300M events. Except for the SM scenario, where the theoretical prediction is used, the cross sections are computed with MadGraph5_aMC@NLO.

c_V	c_{2V}	k_λ	σ [fb]
1	1	1	1.726 [124]
1	1	0	3.9
1	1	2	1.2
1	2	1	12.7
1.5	1	1	57.9
1	0	2	17.8

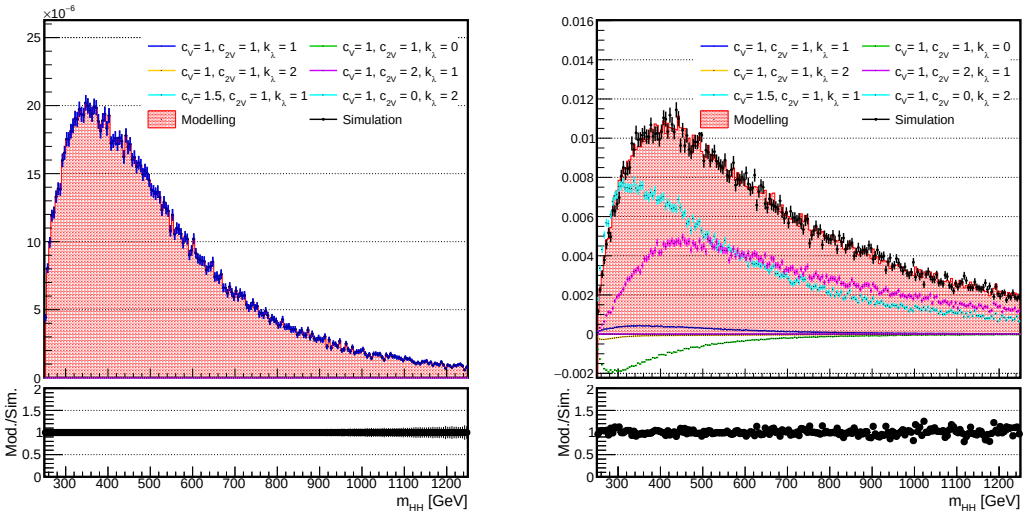
Table 5.4 – Solution of the Eq. 5.6, computed with the cross sections of the signals listed in Tab. 5.3.

Coefficient	Value [fb]
$a = \lvert A\rvert^2$	0.9
$b = \lvert B\rvert^2$	31.4
$c = \lvert C\rvert^2$	16.5
$i_{ab} = \lvert A \cdot B + B \cdot A\rvert$	-8.6
$i_{ac} = \lvert A \cdot C + C \cdot A\rvert$	5.5
$i_{bc} = \lvert B \cdot C + C \cdot B\rvert$	-44.0

Two target distributions as a function of the mass of the generated HH pair and the corresponding six terms at the second member of the Eq. 5.9 are shown in Fig. 5.2. The target distribution, represented by the red histogram, is compared to the black line, which corresponds to the event distribution obtained through a full simulation. In the left plot, the target distribution corresponds to one of the (c_V, c_{2V}, k_λ) combinations used as an input, showing identical distributions as expected from the closure test. In the right plot, a new (c_V, c_{2V}, k_λ) is targeted; the excellent agreement validates the procedure.

5.3 Multijet background

Generic multijet QCD events can enter the event selection if one or two jets are misidentified as hadronic tau leptons. As a prohibitively large number of simulated events would be needed for a solid QCD background estimation, its contribution is instead fully estimated though a data-driven strategy. It consists in building an event categorisation that allows estimating the number of events in the signal region A from the number of events in a "sideband" B, i.e. an independent region, through an extrapolation factor N_C/N_D; this factor is computed from a C and D region with orthogonal definition. This strategy is commonly called "ABCD method".



(a) Target: $c_V = 1$, $c_{2V} = 1$, $k_\lambda = 1$ (b) Target: $c_V = 3$, $c_{2V} = 2$, $k_\lambda = 4$

Figure 5.2 – Event distribution as a function of the mass of the generated HH pair; the "Modelling" curve is obtained from the sum of distributions of simulated events. The corresponding input distributions are shown, scaled by the coefficients of the Eq. 5.9. The black line represent the distribution of the target signal obtained through the full simulation.

The definition of the ABCD regions used in this analysis for the QCD estimation is sketched in Fig. 5.3. The signal region A is the one defined in Sec. 4.3: a pair of leptons with opposite charge, one of them being an hadronic tau lepton, are selected; the hadronic tau lepton needs to pass the *Medium* identification working point discriminant. The selection of the events in the B region is made following the same analysis flow, but it is made orthogonal to the signal region by requiring the tau leptons from the selected $\tau\tau$ pair to have same charge; thus, the jet contamination in the B region is enhanced.

The C and D regions made orthogonal to A and B by inverting the tau identification requirement on the τ_h-leg in the $\tau_e\tau_h$ and $\tau_\mu\tau_h$ channel, or the second hadronic tau lepton (ordered by isolation) in the $\tau_h\tau_h$ channel: it should pass the *Loose* working point, but not the *Medium*.

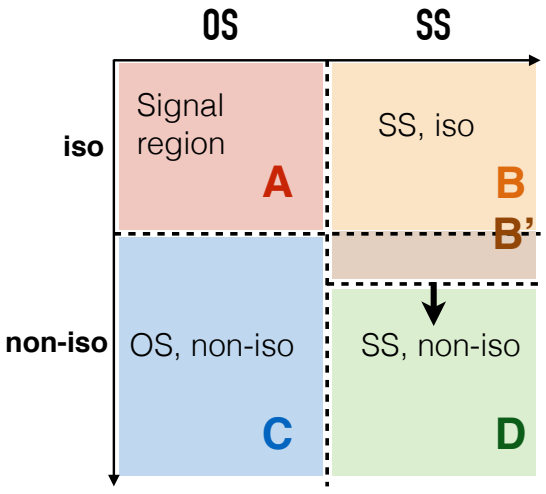


Figure 5.3 – Sketch of the ABCD regions definition for the QCD background estimation.

In each of the B, C and D regions, the number of QCD events is estimated as

$$\mathrm{N}_i = \mathrm{N}_i^{data} - \mathrm{N}_i^{\mathrm{MC}}, \tag{5.10}$$

where N_{MC} is the number of background events modelled with simulations; the background contamination other than QCD, however, is negligible outside the signal region (see Fig. A.8).

Thus, the QCD event yield is extracted as

$$\mathrm{N}_{\mathrm{A}} = \mathrm{N}_{\mathrm{B}} \times \frac{\mathrm{N}_{\mathrm{C}}}{\mathrm{N}_{\mathrm{D}}}. \tag{5.11}$$

The QCD differential distributions, instead, are directly taken from the region with a pair of same-sign, well identified hadronic tau leptons; to enhance the statistical precision and, thus, smooth the templates, a region B' with a relaxed tau identification selection on the τ_{h}-leg that defines the ABCD separation is used for the shape estimation, rather than the previously introduced B region. Signal leakage in the control regions is treated as a source systematic uncertainty.

The final QCD distributions as a function of each variable and computed for each selection are derived from the corresponding data distributions in the B' region, after the subtraction of the residual contamination from other backgrounds; it is then normalised to the result of the Eq. 5.10. For a validation of the ABCD method, see the investigations in Appendix A.

5.4 Drell-Yan background

The contribution of $\mathrm{Z}/\gamma^* \to \ell\ell$ ($\ell = e, \mu, \tau$) background, including the processes that involve jets produced in the hard interaction, is estimated using simulated events.

Three kind of data sets are produced by CMS to model this background: simulated events produced at LO precision with the MadGraph5_aMC@NLO generator; simulated events generated with MadGraph5_aMC@NLO at NLO precision with FxFx merging [125]; and real data with embedded simulated tau leptons. The number of events in each data set is shown in Tab. 5.5.

Table 5.5 – Drell Yan modelling data sets available for 2017 analyses and corresponding number of events.

Data set	Generator	Number of events
$\mathrm{Z}/\gamma^* \to \ell\ell$	MadGraph5_aMC@NLO, LO precision	
inclusive, up to 4 jets		9.38×10^7
+1 jet		7.60×10^7
+2 jets		9.09×10^6
+3 jets		1.15×10^6
+2 b jets		5.11×10^6
$\mathrm{Z}/\gamma^* \to \ell\ell$	MadGraph5_aMC@NLO, NLO precision	
inclusive, up to 2 jets		2.09×10^8
+1 jet		6.06×10^7
+2 jet		2.58×10^8
$\mathrm{Z}/\gamma^* \to \tau\tau$	Data with embedded τ leptons	1.73×10^7

The generation at LO precision allows large samples to be produced, where Drell-Yan processes involving up to four jets produced at matrix element are simulated; the inclusive Drell-Yan samples are complemented by exclusive data sets with 1, 2, 3 and 4 jets produced at matrix element, or with 2 jets originating from b quarks. Thus, a large statistics is guaranteed in the signal region. However, the LO modelling of the jets emission in different flavors is imperfect. A better modelling is achieved with the use of NLO simulated events; the bottleneck, though, is that they are only generated with up to two additional jets. Moreover, there is not an exclusive NLO data set with 2 b jets; given that the fraction of those events is at the permille level, the size of the NLO data set is not sufficient to cover the signal region.

In the H $\to \tau\tau$ analysis [96], the Z/$\gamma^* \to \tau\tau$ contribution is estimated using embedded samples. They consist of selected data in a $Z \to \mu\mu$ enriched region, where the energy deposits and charged tracks associated with the muons are replaced with those from simulated tau leptons. Thus, the underlying physics and the additional jets production are fully estimated from data, providing a very precise modelling. However, since these samples are based on data it is not possible to enhance the number of events needed to model the Drell Yan background in the most sensitive regions of this search. Therefore, this is not a viable solution.

Practically, the only suitable data sets are the ones generated with LO precision. The Drell-Yan modelling is improved through a data-driven technique [121], using events selected Z/$\gamma^* \to \mu\mu$ events to correct the LO simulation for the jet emission.

The Z$\to \mu\mu$ event selection flow is the same as the one used to identify the $\tau\tau$ final states. The events are selected through the single-μ trigger; the most isolated muon is required to have $p_\mathrm{T} > 23\,\mathrm{GeV}$ and $|\eta| < 2.1$, while the second one should have $p_T > 10$ GeV and $|\eta| < 2.4$; both should pass the *Tight* muon identification and isolation criteria and have tracks compatible with the primary vertex. The two selected muons should be separated by $\Delta R > 0.1$ and have opposite charge. Finally, events where additional leptons are found are rejected. All the trigger and identification scale factors are implemented.

A pair of b jet candidates, selected as detailed in Sec. 4.4.4, is also required. The correction should be effective in the signal region of the analysis; therefore, the selected jet pair and muon pair are required to satisfy an elliptical mass cut

$$\frac{(m_{\mu\mu} - 116\,\mathrm{GeV})^2}{(35 + 5\,\mathrm{GeV})^2} + \frac{(m_{\mathrm{bb}} - 111\,\mathrm{GeV})^2}{(45 + 5\,\mathrm{GeV})^2} < 1, \tag{5.12}$$

relaxed by 5 GeV compared to the mass cut defining the signal region (Eq. 4.7).

The $\mathrm{t\bar{t}}$ background is further suppressed by applying a $p_\mathrm{T}^\mathrm{miss} < 45\,\mathrm{GeV}$ selection. The small residual QCD contribution extrapolated using the strategy described in Sec. 5.3; in this case, the AB and CD regions are delimited by the isolation of the least isolated muon.

The selected data and simulated events are split in three categories based on the number of b jet candidates passing the *Medium* DeepCSV working point: the "2b0j" category contains events where both jets are b tagged; the "1b1j" is populated by events with only one b tagged jet; all the other events populate the "0b2j" category. The simulated Drell-Yan events are further split in bins of generator level quantities: the p_T of the Z boson and the number of b quarks (0, 1 or at least 2) produced in the hard interaction.

A correction is assigned to each Drell-Yan bin thus defined; such scale factor is defined through a simultaneous fit to the correct the normalization in the three event categories; the QCD contribution, built from the sidebands of the ABCD method, is varied at the

same time. In Fig. 5.4, the $Z \to \mu\mu$ events distribution as a function of the mass of the di-muon system, i.e. the mass of the reconstructed Z candidate, is shown in the three event categories after the corrections summarised in table Tab. 5.6. An excellent data-over-prediction agreement is achieved.

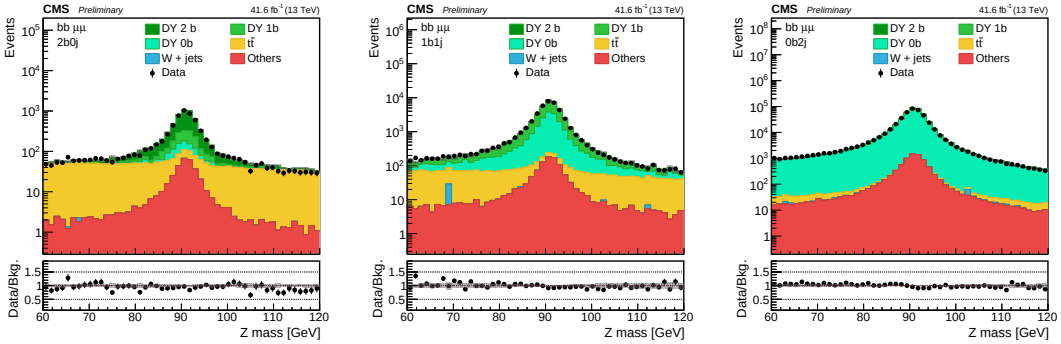


Figure 5.4 – Invariant mass of the two muon candidates in each of the categories used for the determination of the Drell-Yan scale factors: events with 2 b-tagged jets; events with only 1 b-tagged jets; events without b-tagged jets. The selection applied is the one used for the computation of the normalization corrections; the cut on the di-muon mass is removed. The QCD contribution is negligible and is not represented.

Table 5.6 – Scale factors assigned to Drell-Yan simulated events based on observables extracted at the hard interaction level of the simulation: the p_{T} of the Z boson and the flavor of the additional quarks [121].

	0 b quarks	1 b quark	at least 2 b quarks
$0 < p_{\mathrm{T,Z}} < 10\,\mathrm{GeV}$	1.03	1.25	0.51
$10 < p_{\mathrm{T,Z}} < 30\,\mathrm{GeV}$	1.24	1.25	0.83
$30 < p_{\mathrm{T,Z}} < 50\,\mathrm{GeV}$	1.19	1.25	0.92
$50 < p_{\mathrm{T,Z}} < 100\,\mathrm{GeV}$	1.14	1.28	0.94
$100 < p_{\mathrm{T,Z}} < 200\,\mathrm{GeV}$	1.02	1.45	0.85
$p_{\mathrm{T,Z}} > 200\,\mathrm{GeV}$	0.8	1.69	0.89

Chapter 6

Results

The statistical methods for the analysis of the 2017 data are described in this chapter. While in the previous HH $\to$ bb$\tau\tau$ search the stransverse mass m_{T2} (cf. Eq. 4.12) was used as a discriminating observable for the signal extraction, the differential distribution of the BDT discriminant described in Sec. 4.6 is used in the following.

Compared to the previous HH $\to$ bb$\tau\tau$ search performed with the 2016 data set, an additional interpretation is given for the VBF HH signal extraction, exploring various BSM c_{2V} coupling scenarios; together with the trigger and analysis strategies dedicated to the VBF HH signal selection and described in the previous chapters, these results are part of my main area of contribution. Moreover, to improve the disambiguation between the gluon fusion and the VBF HH signals, I have optimized a dedicated multivariate method, described in Sec. 6.6.2.

Those presented in this chapter represent the first CMS HH $\to$ bb$\tau\tau$ set of results with 2017 data. In Sec. 6.7, perspectives on the HH search are given for the coming phases of the LHC program.

6.1 Data set analysed

The results presented in this chapter exploit the data collected with the CMS detector in 2017; the integrated luminosity corresponding to good quality data, recorded in good conditions in terms of beam and of sub detectors integrity, amounts to $41.6\,\mathrm{fb}^{-1}$. The simulated events are weighted so that the event distribution of the generated number of vertices in Monte Carlo data sets matches the 2017 data pileup profile, shown in Fig. 6.1.

6.2 Signal extraction and categories

The signal extraction strategy is the result of choices of compromise and of overall consistency. Inclusive categories, already defined in the previous analyses, are optimised for the search of HH $\to$ bb$\tau\tau$ events and exploit the signal selection efficiency of different b tag requirements. In addition, VBF categories, designed within this thesis work, are optimised targeting a large acceptance on the VBF signal.

Ideally, limits on the SM HH production cross sections can be set by using the gluon fusion and the VBF signal together; moreover, the signal modelling strategies described in Sec. 5.1 and Sec. 5.3 allow any differential event distribution to be reproduced in various k_λ scenarios. The biggest limitation in such strategy is the VBF vs. gluon fusion

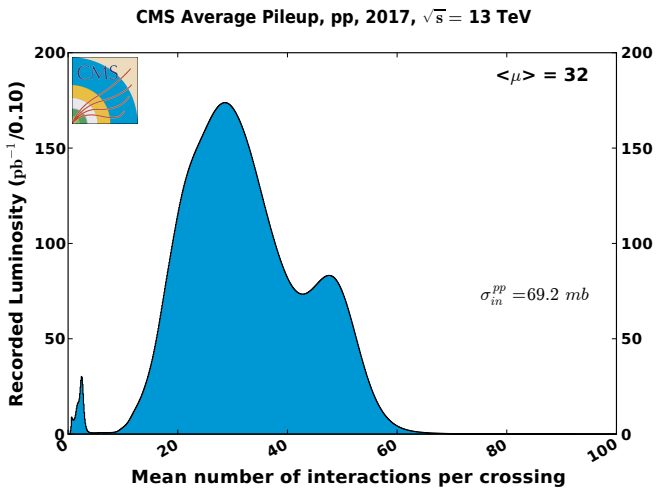


Figure 6.1 – Mean number of interactions per bunch crossing for the 2017 pp collision runs at $\sqrt{s} = 13\,\text{TeV}$ [60]

.

disambiguation. Indeed, as discussed in Sec. 4.7, the VBF selection is also efficient on the gluon fusion signal, given that all the analysis selections target HH $\rightarrow$ bb$\tau\tau$ processes.

Therefore, additional selections are needed to define the VBF categories so that a better separation is achieved between gluon fusion and VBF signals. Within this scope, a DNN-based discriminant was optimised. However, as much as this strategy has a good potential to achieve the desired disambiguation, the event statistics is still a limitation for the implementation of tighter VBF selections such as a DNN-discriminant based cut.

In these conditions, the results presented in this chapter are mostly previsional and set the strategies to be used with increased statistics, in view of the full Run 2 analysis and of the Run 3. As k_λ variations do not bring an enhancement of the VBF cross section large enough to be appreciated, the gluon fusion signal only is used for the signal extraction as a function of k_λ. However, the VBF results are interpreted as a function of the c_{2V} coupling, specific of this production mechanism; as detailed in [45], a large sensitivity to little deviations of c_{2V} from the standard model value is expected. The potential of a DNN-based selection is also shown in this context.

6.3 Statistical interpretation

The statistical methods used for the interpretation of the results commonly used in High Energy Physics consists in quantifying the incompatibility between the observed data (n) and the signal (s) hypothesis or between data and background (b) only hypothesis, expressed in terms of "confidence level" (CL). In this analysis, a modified frequentist approach, often referred to as CL_s, is used to set an upper limit on the presence of the HH signal. To do so, a null hypothesis (usually denoted with H_0) is defined, describing a model with signal plus background, and it is tested against the alternative hypothesis H_1, describing background-only processes. The strategy implemented is the one used for the 2011 combination of the CMS and ATLAS results in the context of the Higgs boson observation, documented in [126, 127]. Following the convention used in SM-like Higgs boson searches, the signal an has arbitrary normalization ($\sigma \cdot \text{BR} = 1\,\text{pb}$) and the results are expressed as a limit on a signal strength modifier $\mu = (\sigma \cdot \text{BR})_{obs}/(\sigma \cdot \text{BR})_{\text{SM}}$.

Both the background and signal event yields are affected by various sources of uncer-

tainties, modelled in the statistical method by introducing nuisance parameters globally denoted as θ; they have a probability density function $\rho(\theta|\tilde{\theta})$ associated to some estimate of the nominal nominal value $\tilde{\theta}$ and other parameters regulating its shape. By the Bayes' theorem, the probability density function can be re-interpreted as a posterior arising from auxiliary measurements of $\tilde{\theta}$, i.e. as

$$\rho(\theta|\tilde{\theta}) \sim p(\theta|\tilde{\theta}) \cdot \pi_\theta(\theta) \tag{6.1}$$

where $p(\theta|\tilde{\theta})$ is the probability density function for the auxiliary measurement of $\tilde{\theta}$ and it can be shown that the prior $\pi_\theta(\theta)$ in the chosen probability density functions used in the following.

6.3.1 Observed limit

The frequentist method consists in defining a "test statistic" q_μ that allows a discrimination of the signal-like event from those background-like. In this search, binned distributions are used for the signal extraction; therefore, the signal and background are vectors of components s_i and b_i corresponding to the content of each bin i. The likelihood function of the hypothesis that the the data observed are compatible with the signal plus background is the product of the Poisson probabilities to observe n_i events in the bins i

$$\mathcal{L}(n|\mu s + b) = \prod_i \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-\mu s_i - b_i}. \tag{6.2}$$

In its classic formulation, the frequentist method does not include systematic uncertainties. However, they are introduced in the following by requiring that the signal and background are functions of the nuisances θ; the likelihood is scaled by the nuisance probability density function as

$$\mathcal{L}(n|\mu, \theta) = \mathcal{L}(n|\mu s(\theta) + b(\theta)) \cdot p(\theta, \tilde{\theta}). \tag{6.3}$$

By the Neyman-Pearson lemma, the most discriminant observable is the ratio of the likelihoods corresponding to the signal plus background and to the background only hypotheses; in case of small signal strength, it is more appropriate to build the test statistic from the profile likelihood function

$$\lambda(\mu) = \frac{\mathcal{L}(n|\mu, \hat{\theta}_\mu)}{\mathcal{L}(n|\hat{\mu}, \hat{\theta})} \tag{6.4}$$

where $\hat{\theta}_\mu$ is the so-called "conditional maximum-likelihood estimator", i.e. the value of θ that maximizes the likelihood for the specified μ; the denominator represents the maximised likelihood. The ratio $\lambda(\mu)$ can, thus, assume values $0 \leq \lambda \leq 1$, where $\lambda(\mu) = 1$ implies a good compatibility between the observed data and the signal plus background hypothesis with strength μ. Commonly, the test statistic uses its logarithm and is defined as

$$q_\mu = -2 \ln \lambda(\mu) \quad \text{with} \quad \mu_0 < \hat{\mu} < \mu, \tag{6.5}$$

where higher values of q_μ correspond to increasing incompatibility with the signal plus background hypothesis; under this form, it can be shown that q_μ follows a χ^2 distribution, so that it is easily readable in terms of probability. The higher constraint on the value of $\hat{\mu}$ is a protection against the effect of downward background statistical fluctuations, so that they are not interpreted as evidence against the hypothesis of a signal with small strength; such protection is one of the main features of the modified frequentist method.

As for the lower constraint, the signal rate is usually assumed to be positive and $\mu_0 = 0$ is chosen.

The probability density functions $f(q_\mu|\mu, \hat{\theta}_\mu^{obs})$ and $f(q_\mu|0, \hat{\theta}_0^{obs})$ are obtained by generating pseudo-data, i.e. simulations that follow the same Poisson probability distribution, in the signal plus background and background-only ($\mu = 0$) hypotheses; the nuisance parameters are fixed to $\hat{\theta}_\mu^{obs}$ and $\hat{\theta}_0^{obs}$ by fitting the observed data. Thus, the probabilities that the observed value q_μ^n is compatible with the signal plus background or with the background-only hypotheses are

$$\begin{aligned} \mathrm{CL}_{s+b} = P(q_\mu \geq q_\mu^n|\mu s + b) = \int_{q_\mu^n}^{\infty} f(q_\mu|\mu, \hat{\theta}_\mu^{obs})\,\mathrm{d}q_\mu, \\ \mathrm{CL}_b = P(q_\mu \geq q_\mu^n|b) = \int_{q_0^n}^{\infty} f(q_0|0, \hat{\theta}_0^{obs})\,\mathrm{d}q_\mu. \end{aligned} \tag{6.6}$$

Their ratio

$$\mathrm{CL}_s = \frac{\mathrm{CL}_{s+b}}{\mathrm{CL}_b} \tag{6.7}$$

is the final figure of merit of the hypothesis test: if, for $\mu = 1$, $\mathrm{CL}_s < \alpha$, the presence of a signal is excluded with a $(1-\alpha)\mathrm{CL}_s$ confidence level; by convention, the 95% confidence level upper limit on μ ($\mu^{95\%\mathrm{CL}}$) is quoted, obtained by raising μ until the $\mathrm{CL}_s = 0.05$ value is reached.

6.3.2 Expected limit

The median of the expected upper limit and the $\pm 1\sigma$ and $\pm 2\sigma$ bands are determined through the generation of a large set of pseudo-data reflecting the background-only hypothesis and compute CL_s and $\mu^{95\%\mathrm{CL}}$ for each sample as if they were data. Hence, a distribution of cumulative probability is built to define the expected limits: the value of $\mu^{95\%\mathrm{CL}}$ for which the cumulative probability distribution crosses the quantile 0.50 is the median of the expected value; the edges of the $\pm 1\sigma$ band are determined from the values corresponding to the crossing of the quantiles 0.16 and 0.84; finally, the values corresponding to the crossings at 0.025 and 0.975 define the $\pm 2\sigma$ band.

6.3.3 Systematic and statistical uncertainties

All sources of uncertainties are considered either fully correlated or fully independent. Uncertainties that are partially correlated need to be broken down in components so that they can be factorised in a clean way; if this disentanglement is not possible, they are treated as fully correlated or fully independent, depending on which is the most conservative assumption.

The systematic uncertainties are propagated to the limit computation after generating the θ nuisances following a $\rho(\theta|\tilde{\theta})$ probability density function centered on the nominal value $\tilde{\theta}$. The log-normal probability density function

$$\rho(\theta|\tilde{\theta}) = \frac{1}{\sqrt{2\pi}\ln(\kappa)} \exp\left(-\frac{(\ln(\theta/\tilde{\theta})^2}{2(\ln\kappa)^2}\right)\frac{1}{\theta} \tag{6.8}$$

is commonly used for this application, as is more appropriate than a Gaussian for large uncertainties and for positively defined observables such as the cross section of a process or the integrated luminosity; however, for small uncertainties, the Gaussian with a relative uncertainties ϵ and the log-normal distribution with $\kappa = 1 + \epsilon$ are asymptotically identical.

To account for the statistical uncertainties associated to the estimation $\mathrm{N}_{bkg} = \alpha\mathrm{N}$ of the number of events in the signal region N_{bkg} from the number N of simulated events or data events from sidebands, gamma distributions are included in the test statistic. The uncertainty on the predicted event rate N_{bkg} is described by

$$\rho(\mathrm{N}_{bkg}) = \frac{1}{\alpha}\frac{(\mathrm{N}_{bkg}/\alpha)^{\mathrm{N}}}{\mathrm{N}!}e^{-\mathrm{N}_{bkg}/\alpha}\,, \tag{6.9}$$

with mean value $\alpha(\mathrm{N}+1)$ and dispersion $\alpha\sqrt{\mathrm{N}^2+1}$.

6.4 Systematics uncertainties

Several sources of systematic uncertainties are considered in the analysis. The systematic uncertainties affecting only the yield of a given process (either signal or background) are detailed in Section 6.4.1; the ones affecting the shape of the final discriminating variable are detailed in Section 6.4.2.

6.4.1 Normalisation uncertainties

- An uncertainty of 2.3% on the event yield is associated to the measurement of the integrated luminosity corresponding to the analysed data [128]. This value is obtained from dedicated Van-der-Meer scans and stability of detector response during the data taking. Since the normalization of the background depend on the measured integrated luminosity, the corresponding uncertainty is considered fully correlated among all the final states and all the processes, except for the ones estimated from data-driven methods such as the QCD and the Drell-Yan background.

- The uncertainties on electron and muon trigger identification and isolation affect separately the $\tau_\mathrm{e}\tau_\mathrm{h}$ and $\tau_\mu\tau_\mathrm{h}$ channels, while the uncertainties corresponding to the hadronic tau lepton selection are fully correlated between the three final states. The normalization uncertainties amount to 2% for muons, 3% for electrons and 7% for hadronic tau leptons; in the $\tau_\mathrm{h}\tau_\mathrm{h}$ final state, where alternative identification scale factors are applied (see Appendix A), a statistical uncertainty of 10% is applied to hadronic tau leptons.

- The hadronic tau lepton energy scale is defined as the average of the ratio of the reconstructed and simulated energies; it is measured in $\mathrm{Z}\rightarrow\tau\tau\rightarrow\tau_h\nu_\tau\mu\nu_\mu\nu_\tau$ events by fitting the hadronic tau reconstructed mass and the invariant mass of the $\mu\tau_\mathrm{h}$ system [82]. For each reconstructed hadronic tau lepton, up and down variations of its energy, ranging from -0.2 to 0.7% based on the decay mode [129], are computed and propagated to the observables making use of the hadronic tau lepton reconstruction. As the final selection depends on the hadronic tau lepton p_T and the mass of the $\mathrm{H}_{\tau\tau}$ candidate, the tau energy scale affects the signal and background event yield in the signal regions. The impact of this uncertainty on the final selection, fully correlated among the final states, amounts to up to 3% depending on the process.

- Similarly to the tau energy scale, uncertainties on the jet energy scale are taken into account by measuring the changes in the acceptance for the different simulated processes when the two selected jets momenta are varied in the range defined by their uncertainty. Changes in the acceptance arise from both the threshold on the jet momenta and the selection on the m_{bb} invariant mass. The impact of

this uncertainty amounts to about 3% for the signal and 4% for the background processes in the inclusive categories; in the loose VBF categories, it goes up to 7%, while in the VBF tight category it can be as large as 20%.

- The extent of the uncertainties on b-tagging efficiencies as function of jet p_T and η is evaluated together with the data-over-prediction scale factors and ranges from 2% for most of the backgrounds to 6% for the processes with genuine b jets in the final states.
- The cross section of the HH pair production via gluon fusion has uncertainties arising from the the scale variations, whose impact amount to $+0.66/-2.8\%$, and other theoretical uncertainties such as PDFs and α_S, giving an additional $\pm 3.3\%$ [130].
- As for the VBF HH cross section, the uncertainty due to the scale variations amounts to +0.03/-0.04%, while the PDF+α_S uncertainty is $\pm 2.1\%$ [124].
- Other theoretical uncertainties affecting the signal event yield come from the imperfect knowledge on the 125 GeV Higgs boson branching ratios in a bb and a $\tau\tau$ pair, of $+3.21/-3.27\%$ and $+5.71/-5.67\%$ respectively [131]. These uncertainties are only taken into account when quoting the limit on the HH production compared to the SM prediction.
- For the signal samples, as well as for $t\bar{t}$, W+jets, single top, single Higgs, and di-boson backgrounds, uncertainties due to the imperfect knowledge of the process normalizations and simulation are considered. For the signal case, they yield an effect of about 5% due to renormalization and factorization scales variation, and about 3% due to the uncertainties on parton distribution functions and α_S.
- The Drell-Yan + jets background contribution is estimated from Monte Carlo samples and tuned using data in $\mu\mu$ sidebands (5.4). The uncertainties on the SFs are obtained through the comparison with another set of SFs computed with a relaxed mass selection. The SFs are computed in several bins of Z p_{T}; the uncertainties corresponding to the average p_{T} of each of the Z+0b, Z+1b, Z+2b distributions, of 3%, 6% and 20%, are applied separately to those processes.
- The QCD background is extrapolated from sidebands as detailed in Sec. 5.3. Its estimation is affected by several sources of systematic uncertainty: the number of events estimated in the control region, that is subject to Poissonian fluctuation and modelled through the Eq. 6.9; additional uncertainty sources from the number of events measured in each of the ABCD regions are taken into account in the simultaneous binned maximum likelihood fit, performed for the computation of the final limit on HH production.

6.4.2 Shape uncertainties

- The gluon fusion HH signal extraction uses a BDT discriminant for each k_λ point of the scan. The shape uncertainty on the BDT response is estimated by building an envelope of the BDT score distributions: alternative "up" and "down" histograms are filled bin-by-bin using the maximum of the minimum among all the BDT distributions in the corresponding bin.
- Uncertainties on the measurement of the energy of jets and hadronic tau leptons do not only affect the signal acceptance, but also the differential distribution of the BDT score used for the signal extraction. A BDT discriminant can be re-evaluated using, instead of the nominal observables, those shifted by the possible variations

of the energy scales, so that the uncertainty is propagated to the BDT score shape. For the VBF HH signal extraction, one BDT discriminant only ($k_\lambda = 1$) is used; in this case, the alternative up and down shapes are introduced in the final fit. These uncertainties are considered fully correlated with the corresponding normalisation uncertainties. As for the gluon fusion HH signal extraction, this shape uncertainty is considered negligible compared to the one described in the previous item.

- The uncertainty affecting the kinematic distributions in the $\mathrm{t\bar{t}}$ background simulation is taken into account by building alternative shapes of the BDT discriminant, obtained by varying the p_T distribution of the top quark and antiquark according to the differential p_T measurements described in [132].

6.5 Gluon fusion HH production

The gluon fusion signals are modelled, as detailed in Sec. 5.1, through the parametrization of the effective Lagrangian couplings introduced in [44]. In the following, the $c_2 = c_{2g} = c_g = 0$ scenario is considered; this is the case where only the triangle diagram, whose cross section is governed by the $k_\lambda k_\mathrm{t}$ product, and the box diagram, regulated by the value of k_t^2, participate to the HH production though gluon fusion. The HH production cross section is proportional to the square of the corresponding matrix elements, which can be written as

$$|Bk_\mathrm{t}^2 + Tk_\lambda k_\mathrm{t}|^2 = |B|^2 k_\mathrm{t}^4 + |T|^2 k_\lambda^2 k_\mathrm{t}^2 + 2BTk_\lambda k_\mathrm{t}^3 \tag{6.10}$$

where B and T are the coefficients of the box and triangle diagram. Thus, the differential cross section with respect to a generic observable x can be expressed under the form

$$\frac{\mathrm{d}\sigma}{\mathrm{d}x} \sim \frac{\mathrm{d}|T(x)|^2}{\mathrm{d}x}\frac{k_\lambda}{k_\mathrm{t}}^2 + 2\frac{\mathrm{d}B(x)T(x)}{\mathrm{d}x}\frac{k_\lambda}{k_\mathrm{t}} + \frac{\mathrm{d}|B(x)|^2}{\mathrm{d}x} \tag{6.11}$$

so that the event distribution is a function of the ratio k_λ/k_t.

The discriminant variable used for the gluon fusion signal extraction is the score of the BDT discriminant implemented in the MVA technique for the $\mathrm{t\bar{t}}$ background rejection. The value of k_λ is one of the inputs of the training; therefore, the signal extraction can profit from a specific BDT discriminant for each point in the k_λ scan.

The search is performed in all the final event categories: the resolved 1b1j, resolved 2b0j and boosted, orthogonal to the VBF categories and optimised for the sensitivity on inclusive HH $\to \mathrm{bb}\tau\tau$ signals; and the VBF categories, as the purity of these regions is beneficial also to the gluon fusion event selection. It should be reminded, however, that the gluon fusion signals currently employed in CMS HH searches do not integrate the production Higgs boson pairs in association to jets; therefore, the jets in the gluon fusion signal that allow it to enter the VBF selection are produced by parton shower processes.

6.5.1 Results

The upper limit on the cross section of the HH production via gluon fusion times the HH $\to \mathrm{bb}\tau\tau$ branching fraction, normalised to $(\sigma \cdot \mathrm{BR}) = 1\,\mathrm{pb}$ as mentioned in Sec. 6.3.1, is set as a function of the k_λ/k_t ratio. The result is shown in Fig. 6.2. The green and yellow bands represent the $\pm1\sigma$ and $\pm2\sigma$ deviations from the expected value; the observed limit lies within the $\pm1\sigma$ band over all the considered range. The red parabolas represent the theoretical predictions for $k_\mathrm{t} = 1$ and $k_\mathrm{t} = 2$.

The shape assumed by the limit reflects experimental and theoretical elements. The point $k_\lambda/k_t = 2.46$ corresponds to the maximum disruptive interference between the box and triangle diagram; therefore, the sensitivity changes dramatically around that value. An edge with lower sensitivity is observed on the right; indeed, the mass spectra of gluon fusion HH signals with positive k_λ/k_t up to about 10 are softer, so that the experimental acceptance on them is smaller. For $|k_\lambda/k_t| > 10$, the limits tend asymptotically to similar values.

By comparing the observed upper limits to the theoretical curve with $k_t = 1$, the value of k_λ is constrained at 95% CL in the range between -9.1 and 15.4; the expected limits give $-10.1 < k_\lambda < 17.2$. In the $k_t = 2$ scenario, the observed (expected) constraints are $-0.5(-0.5) < k_\lambda < 6.2(6.7)$.

The value of $k_\lambda/k_t = 1$ corresponds to the SM prediction. The observed 95% CL upper limit is 46.6 fb, i.e. about 20 times the HH production cross section and HH $\to$ bb$\tau\tau$ branching ratio predicted by the SM. The $\tau_h\tau_h$ channel is the most sensitive: the observed limit in this channel alone is equivalent to about $32 \times (\sigma \cdot \mathrm{BR})_{\mathrm{SM}}$. As for the categories, the most sensitive is the one where two b tagged jets are selected. The VBF tight category is the least sensitive; however, it should be noted that it exists only for the $\tau_h\tau_h$ channel. The observed and expected limits on the SM HH production cross section are listed per channel and per category in Tab. 6.1 and Tab. 6.2.

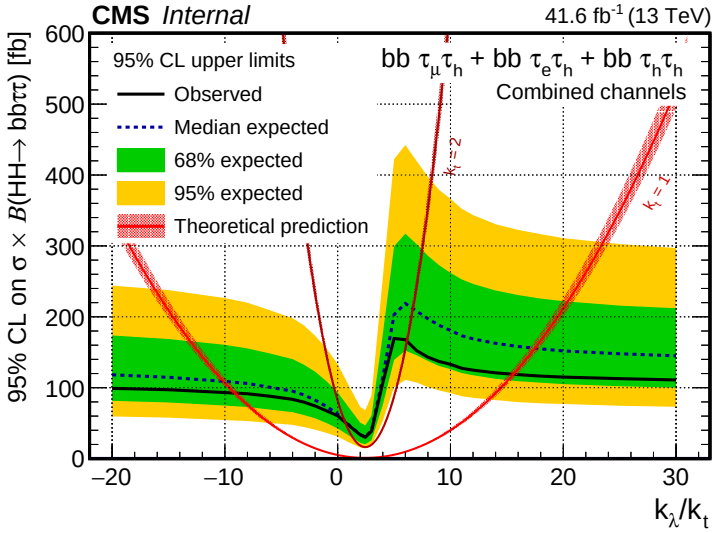


Figure 6.2 – 95% CL upper limits for on $\sigma(gg \to HH) \cdot BR(HH \to bb\tau\tau)$ as a function of the ratio k_λ/k_t. The signal (σ·BR) is normalised to 1 fb. The $\pm1\sigma$ and $\pm2\sigma$ deviations from the expected value are represented by green and yellow bands; the red curves represent the theoretical predictions for $k_t = 1$ and $k_t = 2$.

6.5.2 Comparison with earlier LHC results

In the following, the most recent results produced by ATLAS and CMS in non-resonant gluon fusion HH production searches are summarised.

- The most recent Run 2 result from CMS was obtained from the analysis of about 35.9 fb^{-1} of data collected in 2016, documented in [107]. It was the first Run 2 paper on this analysis. In this search, the signal is extracted by fitting the m_{T2} distribution (see Eq. 4.12). The observed (expected) 95% upper limit set on the SM signal is 75.4(61.0) fb, which corresponds to about 33(27) times the

Table 6.1 – 95% CL upper limits for $k_\lambda/k_t = 1$ on $\sigma(gg \to HH) \cdot BR(HH \to bb\tau\tau)$ by $\tau\tau$ channel; in the third column, the limit is expressed in terms of the SM prediction as a signal strength μ.

Channel	Obs. (exp.) upper limit on $\sigma \cdot BR$ [fb]	Obs. (exp.) $\mu = (\sigma \cdot BR)/(\sigma \cdot BR)_{SM}$
$\tau_h\tau_h$	72.1 (59.1)	31.5 (25.8)
$\tau_\mu\tau_h$	85.3 (114.7)	37.3 (50.1)
$\tau_e\tau_h$	124.7 (146.5)	54.5 (64.1)
Combined	46.6 (47.4)	20.4 (20.7)

SM prediction using the most updated theoretical gg→ HH cross section quoted in this thesis [130]. For searches limited by the statistical uncertainties, such as HH → bb$\tau\tau$, the sensitivity to a given signal is expected to improve with the integrated luminosity $\mathcal{L}$ following a $1/\sqrt{\mathcal{L}}$ trend: under the same data quality conditions and analysis strategies used in the 2016 analysis, a realistic estimate of the result expected for 41.6 fb^{-1}, i.e. the integrated luminosity of the data analysed in this thesis, is 31(25) times the SM cross section. Therefore, the improvement brought by the changes in the analysis strategy amounts to about 30%; the largest impact is expected to come from the change of signal extraction variable. As for the scan as a function of k_λ, the observed (expected) constraints set by the 2016 search are $-18(-14) < k_\lambda < 26(22)$.

- A HH → bb$\tau\tau$ non-resonant search, documented in [133], was also performed by the ATLAS collaboration on 2016 data (35.9 fb^{-1}). Similarly to the analysis strategy described in this thesis, the discriminant observable used for the signal extraction is the score of a BDT; in this case, the BDT is trained against all the major backgrounds ($t\bar{t}$, Drell-Yan, QCD). Thus, a stringent limit is set on the SM model HH production cross section: the observed (expected) limit is 30.9 fb (36.0 fb), i.e. about 14(16) times the state-of-the-art theoretical SM cross section prediction [130].

- The combination of the CMS HH searches with 2016 data is documented in [108]; it should be noted that the observed constraints to k_λ obtained in this thesis ($-9.1 < k_\lambda < 15.4$) are more stringent than those obtained through the combination of the 2016 analyses ($-11.8 < k_\lambda < 18.18$).

Table 6.2 – 95% CL upper limits for $k_\lambda/k_t = 1$ on $\sigma(gg \to HH) \cdot BR(HH \to bb\tau\tau)$ by category; in the third column, the limit is expressed in terms of the SM prediction as a signal strength μ. It should be noted that the VBF tight category is only implemented in the $\tau_h\tau_h$ channel and for a fraction of data corresponding to about 27 fb^{-1}.

Channel	Obs. (exp.) upper limit on $\sigma \cdot BR$ [fb]	Obs. (exp.) $\mu = (\sigma \cdot BR)/(\sigma \cdot BR)_{SM}$
res. 2b0j	54.8 (64.9)	24.0 (28.4)
boosted	110.1 (105.5)	48.2 (46.2)
VBF loose	264.2 (205.1)	115.5(89.7)
res. 1b1j	291.3(211.9)	127.4(92.7)
VBF tight ($\tau_h\tau_h$)	2505.9 (2476.6)	1096.4 (1083.5)
Combined	46.6 (47.4)	20.4 (20.7)

6.6 VBF HH production

The VBF search is also performed in all the final event categories. The gluon fusion contribution, in this case, is considered as a background. Although the multivariate BDT-based technique against the $t\bar{t}$ contamination is trained with a gluon fusion sample, it also provides a good discrimination between $t\bar{t}$ and the VBF HH signal. However, the VBF modelling of BSM scenario is based on the sum of histograms with potentially large cancellations; therefore, a shape analysis is delicate in low statistics region. Therefore, a counting experiment is performed, equivalent to the shape analysis described in Sec. 6.3 performed in one bin only: the discriminant is exploited for background rejection by applying a BDT *score* > 0 selection.

6.6.1 Preliminary results

In Tab. 6.3 and Tab. 6.4, the upper limits in the SM scenario are shown separately by channel and by category. The resolved 1b1j category, due to the large background contamination, is the least sensitive; its sensitivity is similar to the one of the VBF tight category, which only exists for the $\tau_h\tau_h$ channel.

Table 6.3 – 95% CL upper limits for $k_\lambda/k_t = 1$ on $\sigma(\text{VBF HH})\cdot\text{BR}(\text{HH} \to \text{bb}\tau\tau)$ by $\tau\tau$ channel; in the third column, the limit is expressed in terms of the SM prediction as a signal strength μ.

Channel	Obs. (exp.) upper limit on $\sigma \cdot \text{BR}$ [fb]	Obs. (exp.) $\mu = (\sigma \cdot \text{BR})/(\sigma \cdot \text{BR})_{\text{SM}}$
$\tau_h\tau_h$	216 (153)	1717 (1217)
$\tau_\mu\tau_h$	209 (267)	1660 (2116)
$\tau_e\tau_h$	309 (357)	2452 (2836)
Combined	139 (107)	1122 (848)

Table 6.4 – 95% CL upper limits for $c_V = 1$, $c_{2V} = 1$ and $k_\lambda = 1$ on $\sigma(\text{VBF HH}) \cdot \text{BR}(\text{HH} \to \text{bb}\tau\tau)$ by category; in the third column, the limit is expressed in terms of the SM prediction as a signal strength μ. It should be noted that the VBF tight category is only implemented in the $\tau_h\tau_h$ channel and for a fraction of data corresponding to about 27 fb^{-1}.

Channel	Obs. (exp.) upper limit on $\sigma \cdot \text{BR}$[fb]	Obs. (exp.) $\mu = (\sigma \cdot \text{BR})/(\sigma \cdot \text{BR})_{\text{SM}}$
VBF loose	151 (128)	1218 (1037)
res. 2b0j	301 (441)	2426 (3557)
boosted	595 (682)	4792 (5492)
res. 1b1j	1569 (1484)	12640 (11961)
VBF tight ($\tau_h\tau_h$)	1445 (1449)	11644 (11961)
Combined	139 (107)	1122 (848)

The VBF signal modelling described in Sec. 5.3 allows various k_λ, c_V and c_{2V} scenarios to be explored. As the greatest sensitivity is expected on the c_{2V} coupling, the upper limits are computed as a function of this parameter. In Fig. 6.3, the upper limit on the HH production via VBF is shown. As expected, a striking sensitivity is observed to small variations of the c_{2V}: in the SM scenario, the observed 95% CL upper limit is 139 fb, i.e. 1122 times the state-of-the-art SM prediction [124]; by moving only by 0.5 along the x-axis in the positive direction, for instance, the upper limit is reduced by about a factor

5. The red line represents the theoretical prediction for $k_\lambda = 1$, $c_V = 1$, from the Eq. 5.4 with the solution shown in Tab. 5.4; thus, the observed (expected) constraint to c_{2V} are set to $-0.9(-0.8) < c_{2V} < 2.9(2.8)$.

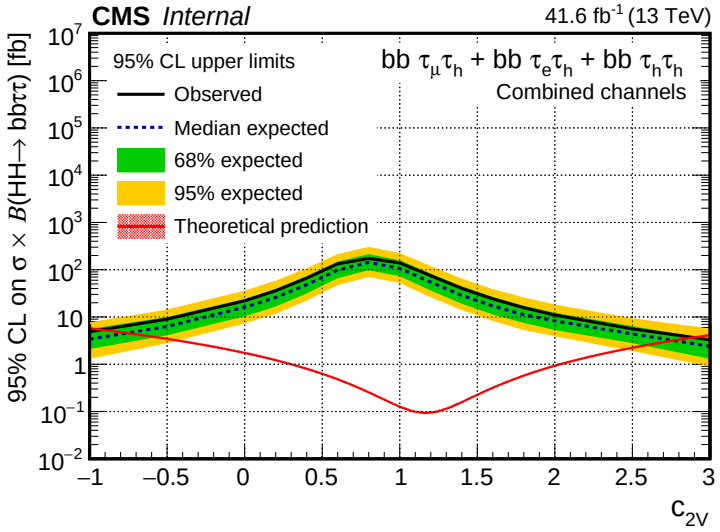


Figure 6.3 – 95% CL upper limits for $k_\lambda = 1$, $c_V = 1$ on $\sigma(\text{VBF HH}) \cdot \text{BR}(\text{HH} \to \text{bb}\tau\tau)$ as a function of the coupling c_{2V}. The signal (σ·BR) is normalised to 1 fb. The signal (σ·BR) is normalised to 1 fb. The $\pm 1\sigma$ and $\pm 2\sigma$ deviations from the expected value are represented by green and yellow bands; the red curves represent the theoretical predictions for $k_\lambda = 1$ and $c_V = 1$.

This result can be further improved by implementing a dedicated method for the VBF vs. gluon fusion discrimination; in the next section, a preliminary DNN-based discriminant is presented.

6.6.2 VBF vs. gluon fusion discriminant

The discrimination of two HH production signals with different production mechanism is particularly challenging; rather than implementing further kinematic selections, it is convenient to implement a machine learning (ML) technique. In general, ML algorithms build a prediction, expressed in a single discriminant, by learning the features of the samples contained in large data sets and by capturing their correlations. This application is a case of "supervised learning": fed with a VBF HH and a gluon fusion signal data set, the algorithm can compare the input features (the event observables) to the desired output (whether the event belongs to the VBF HH data set or not). Although a parametrization of the ML discriminant as a function of c_{2V} could be relevant, it is not included in the training presented in the following.

A review of the machine learning techniques is out of the scope of this study; however, a general description of the BDT and DNN algorithms is given in [134, 135].

Data sets

A private production of signal samples was necessary for the training. The VBF HH and gluon fusion samples produced centrally within the CMS collaboration contain 300k events. At the end of the H $\to \tau\tau$ selection chain, the VBF HH event statistics is reduced to $\mathcal{O}(3000) - \mathcal{O}(30000)$ events for each training; to improve the ML performance, I produced an additional identical data set of 3M events. Each training, depending on the corresponding selection, can thus rely on more than 10k VBF HH events. Although the

gluon fusion signal profits from a higher $H \to \tau\tau$ selection efficiency, a larger simulation sample was also produced for this process. However, as mentioned in Sec. 4.7, the pre-existing gluon fusion sample only contains the simulation of processes without additional jets. Therefore, rather than reproducing the same sample with increased statistics, a data set of 600k gluon fusion events where the Higgs boson pair is produced in association to one or two jets was generated; thus, the sensitivity of the ML trainings to jets coming from the hard interactions is improved.

Both the signal data sets were produced following the CMS guidelines for Monte Carlo generation, using MadGraph5_aMC@NLO; the quark hadronization, underlying event and pileup effects are modelled through pythia 8. While the VBF data set is perfectly consistent with the one centrally provided, gluon fusion data sets with additional jets were never produced centrally by the CMS Monte Carlo generators experts; due to the lack of guidelines, the computation of the relative cross section of events with one and with two jets is not trivial. Therefore, this sample is only used for the training of the ML algorithms.

Choice of the algorithm

The performance of the DNN and BDT approaches are compared in the $\tau_h\tau_h$ channel. The training is performed using a very loose selection, without any trigger requirement. Indeed, even though the kinematics of the HH decay present some differences for VBF and gluon fusion events, the main target is to train the ML algorithms to discriminate the jet topology; although the events that do not fire the trigger do not pass the analysis event selection, they are useful for the learning of the jets features. The presence of two hadronic taus with $p_T > 20\,\text{GeV}$, passing the *Loose* identification working point, is required; events with additional leptons are discarded. At least two b jet candidates must be found, without requirements on their b tag score. A loose threshold $m_{jj} > 300\,\text{GeV}$ on the jet-jet invariant mass of the VBF jet candidates is applied, so that the phase-space probed by the algorithm is similar to that of the analysis. About 37k events from each signal sample are used.

At this stage, the aim is only to verify whether a DNN approach is more performant than a BDT or not; therefore, a large number of observables is given as an input to both algorithms. The variables used for the trainings are those relative to the kinematics of the hadronic tau candidates, of the b jet candidates, of each of the Higgs bosons reconstructed from them, of the HH system and, if found, of the VBF jet candidates; since the pseudorapidity distribution of the second VBF jet shows an important data-over-prediction disagreement in the noisy region at large η (see Fig. 4.21d), it is not given as an input to the ML algorithms. The p_T, m_{jj} and $|\Delta\eta_{jj}|$ of the first two additional jets ordered by p_T, rather than by m_{jj} like the VBF selected jets, are also considered.

In addition to having large m_{jj} and large $|\Delta\eta_{jj}|$, as the VBF jets are typically produced in opposite regions of the detector, a useful discriminating variable is the the product of the pseudorapidity of the two jets; this quantity, typically negative for VBF jets, is computed for all the mentioned jet pairs (VBF jets candidates, b jet candidates, first two additional jets by p_T) and included in the training.

A distinguishing feature of the VBF process is that there is no color exchange between the quarks involved in the interaction, so that the hadronic activity is suppressed central region between the two VBF jets except for the Higgs decay products. The Zeppenfeld [136] centrality is a useful handle to exploit this information. It is defined as

$$z = \frac{\eta - 1/2(\eta_1^{\text{VBF}} + \eta_2^{\text{VBF}})}{|\eta_1^{\text{VBF}} - \eta_2^{\text{VBF}}|} \tag{6.12}$$

where η is the pseudorapidity of a given object and η_1^{VBF} is η_2^{VBF} are those of the two VBF jet candidates; in practice, it is a measurement of the pseudorapidity of an object with respect to the bisector of the two VBF jets along η. While the Higgs decay products in VBF events are expected to have small ζ in absolute value, additional jets with respect to the selected VBF jets are expected to be produced in the forward region. These features are not reflected in gluon fusion events *a priori*, although the centrality distribution of the HH system and its products can be biased by the selection of the VBF jets candidates as those with highest invariant mass. The centrality of the Higgs candidates, of the HH system, of all the decay product candidates and of the first additional jet by p_{T} besides the VBF jet pairs are given as input to the ML algorithms. A variable similar to the Zeppenfeld centrality, defined as [137]

$$\zeta_{\text{HH}} = \min[\Delta\eta_-, \Delta\eta_+] \tag{6.13}$$

where

$$\begin{aligned} \Delta\eta_- &= \min[\eta(\text{H}_{\text{bb}}), \eta(\text{H}_{\tau\tau})] - \min[\eta(jet_1^{\text{VBF}}], \eta(jet_2^{\text{VBF}})] \\ \Delta\eta_+ &= \max[\eta(jet_1^{\text{VBF}}), \eta(jet_2^{\text{VBF}})] - \max[\eta(\text{H}_{\text{bb}}), \eta(\text{H}_{\tau\tau})] \end{aligned} \tag{6.14}$$

is also used; its value is large when the VBF jets have large η separation and the Higgs candidates are produced in the region between them.

Finally, the kinematics of the VBF process lead the HH pair to be boosted; in addition to the p_{T} of the HH system, a p_{T} balance observable is used in the training, defined as [137]

$$A_{\text{HH}} = \frac{|\vec{p_{\text{T}}}(\text{H}_{\text{bb}}) + \vec{p_{\text{T}}}(\text{H}_{\tau\tau})|}{p_{\text{T}}(\text{H}_{\text{bb}}) + p_{\text{T}}(\text{H}_{\tau\tau})}. \tag{6.15}$$

All the variables used for the training are listed in Tab. 6.5.

Table 6.5 – Input variables of the ML algorithms, grouped by physics objects.

Object	Observable
τ_1, τ_2 candidates	p_{T}, η, isolation, z (Eq. 6.12)
First and second b jet candidate	p_{T}, η, z (Eq. 6.12), $\Delta\eta_{\text{bb}}$, $\eta_1 \cdot \eta_2$
First and second VBF jet candidate	p_{T}, η (first VBF jet only) , z (Eq. 6.12), $\Delta\eta_{\text{jj}}$, $\eta_1 \cdot \eta_2$, m_{jj}
First and second additional jet by p_{T}	p_{T}, η, z (Eq. 6.12), $\Delta\eta_{\text{jj}}$, $\eta_1 \cdot \eta_2$, m_{jj}
First jet by p_{T} additional w.r.t. VBF jet candidates	p_{T}, η, z (Eq. 6.12)
$\text{H}_{\tau\tau}$ candidate	mass, p_{T}, z (Eq. 6.12)
H_{bb} candidate	mass, p_{T}, z (Eq. 6.12)
HH system	mass, p_{T}, ζ_{HH} (Eq. 6.13), A_{HH} (Eq. 6.15)

The BDT and DNN training were performed using the XGBoost [138] and Keras [139] python libraries respectively. In both cases, area under the ROC curve is used as estimator of the performances; to improve it, the parameters of the algorithms were tuned by making small variations around values commonly used for this kind of applications. The chosen configurations are summarised in Tab. 6.6; the meaning of the parameters is detailed in [138, 139] The ROC curves thus obtained are compared in Fig. 6.4: for all the possible thresholds applied to the discriminant, the fraction of correctly assigned and rejected background events are represented on the x and y axis. The DNN training

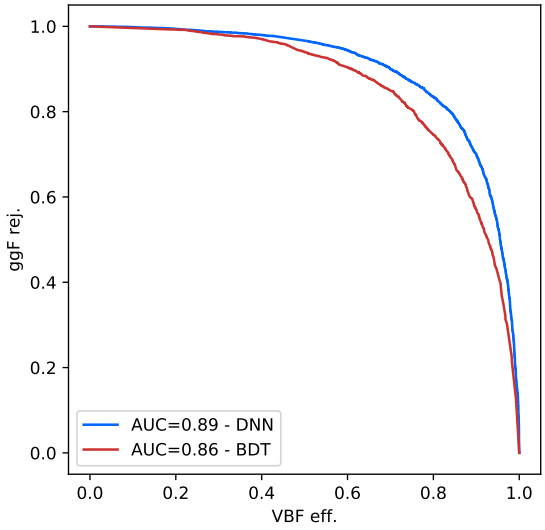


Figure 6.4 – ROC curves representing the discrimination performance of a DNN and a BDT discriminant in the VBF loose $\tau_h\tau_h$ category.

Table 6.6 – Parameters of the BDT (left) and DNN trainings. Their meaning is clarified in [138, 139].

(a) BDT training

Number of trees	100
Maximum tree depth	3
Minimum children weight	1
Learning rate	0.1
Gamma	0

(b) DNN training

Number of layers	4
Nodes per layer	100,100, 10,1
Activation functions	relu, relu, relu, sigmoid
Learning rate	0.1
Dropout rate	0.2

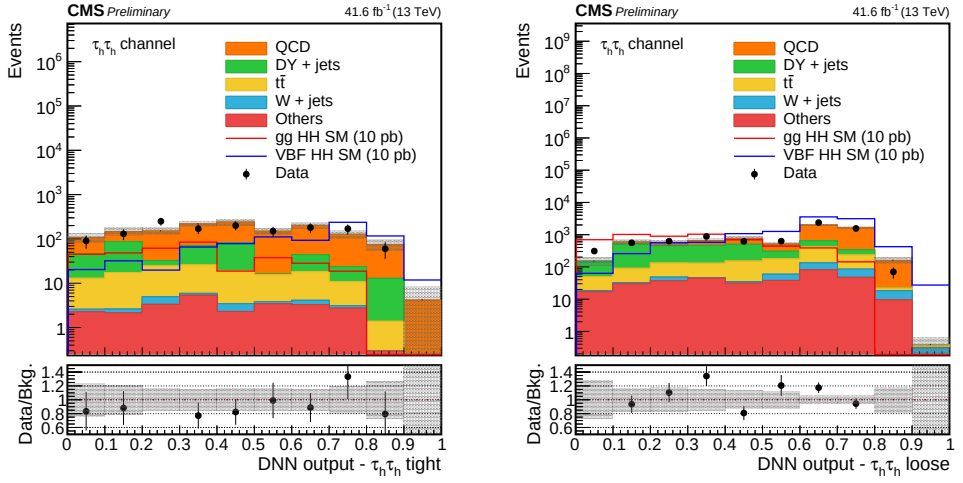
shows a better performance, with an area under the ROC curve of 0.89; the ROC curve of the BDT, instead, spans an area of 0.86.

VBF categories discriminators

Given the better performance of the DNN approach, this method is chosen to build the final discriminators. All the variables listed in Tab. 6.5 are used to train the algorithm. Four different trainings are performed, one for each of the VBF categories: the VBF loose category in the $\tau_\mu\tau_h$ channel, the VBF loose category in the $\tau_e\tau_h$ channel, the VBF loose category in the $\tau_h\tau_h$ channel, and the VBF tight category in the $\tau_h\tau_h$ channel. The distributions of the DNN outputs are shown in Fig. 6.5; a good data-over-prediction agreement, as well as an efficient VBF vs. gluon fusion separation, is achieved. The ROC curves of each DNN discriminant, with efficiencies computed within the corresponding VBF categories, are shown in Fig. 6.6. For instance, in the VBF loose category of the $\tau_h\tau_h$ channel, selecting events with DNN score larger than 0.4 allows the 69% of the gluon fusion events to be rejected, while preserving the 87% of VBF signal events. Loose working points of the DNN-discriminant are identified, so that a reasonable statistics is preserved in the final signal regions and a similar signal efficiencies is achieved across different VBF categories. In Tab. 6.7, the corresponding VBF selection and gluon fusion rejection efficiencies are shown.

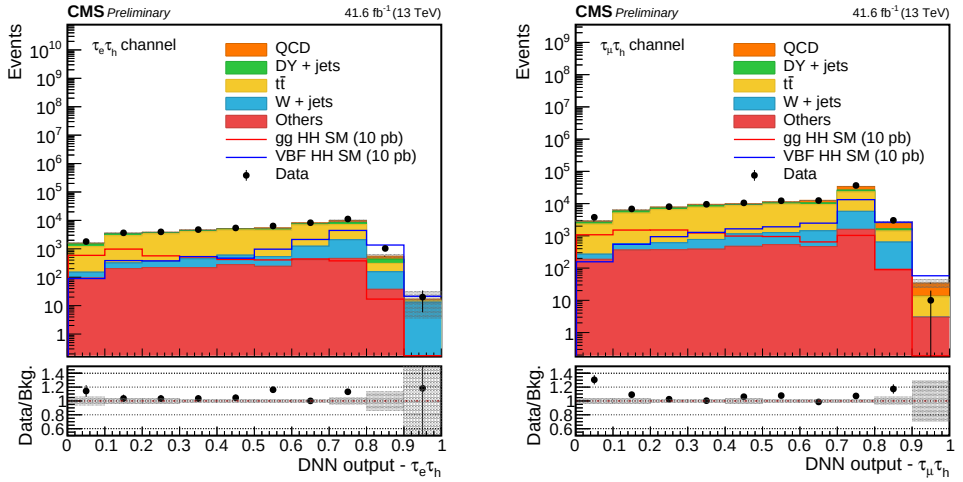
Table 6.7 – Efficiency of the DNN output selection for the working points chosen for each category.

Category	DNN output threshold	Selection efficiency on VBF events	Rejection of gluon fusion events
VBF tight $\tau_h\tau_h$	0.3	91%	45%
VBF loose $\tau_h\tau_h$	0.4	87%	69%
VBF loose $\tau_e\tau_h$	0.4	87%	71%
VBF loose $\tau_\mu\tau_h$	0.4	88%	59%



(a) VBF tight $\tau_h\tau_h$

(b) VBF loose $\tau_h\tau_h$



(c) VBF loose $\tau_e\tau_h$

(d) VBF loose $\tau_\mu\tau_h$

Figure 6.5 – Distributions of events passing the baseline selection and the m_{jj} and $|\Delta\eta_{jj}|$ requirements of each VBF category, as a function of the output of the dedicated DNN training.

6.6.3 Results

Purer VBF categories are defined by implementing the corresponding DNN-based selection, listed in Tab. 6.7, while the existing inclusive categories are unchanged. The resulting 95% upper limits on the VBF production cross section are shown in Fig. 6.7 as a function of the c_{2V} coupling. Although the DNN discriminants are only optimised for the SM scenario, without any c_{2V} parametrization, an improvement of about 10% with respect to the results shown in Fig. 6.3 is observed over all the considered range. Thus,

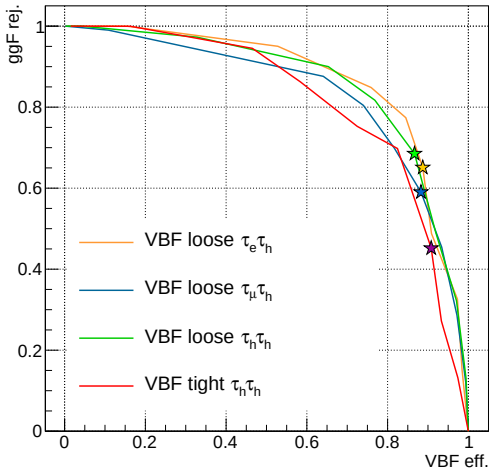


Figure 6.6 – ROC curves representing the discrimination performance of the DNN discriminants in the corresponding VBF categories. The star markers indicate the working points listed in Tab. 6.7.

the combined observed (expected) limit for the SM production cross section is reduced to 62.0 (93.8) fb, i.e. 492 (744) times the theoretical prediction. In Tab. 6.8 and Tab. 6.9, the results corresponding to $c_{2V} = 1$ are shown separately by channel and by category. As for the c_{2V} limit, it is constrained to the range $-0.8 < c_{2V} < 2.8$ both by observed and expected data.

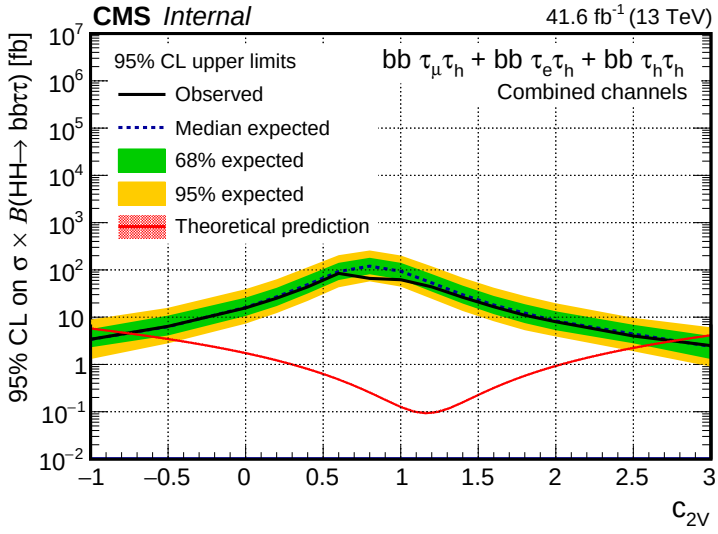


Figure 6.7 – 95% CL upper limits for $k_\lambda = 1$, $c_V = 1$ on $\sigma(\text{VBF HH}) \cdot \text{BR}(\text{HH} \to \text{bb}\tau\tau)$ as a function of the coupling c_{2V}; in the VBF categories, a DNN based selection is implemented for the VBF vs. gluon fusion disambiguation. The $\pm1\sigma$ and $\pm2\sigma$ deviations from the expected value are represented by green and yellow bands; the red curve represents the theoretical predictions for $k_\lambda = 1$ and $c_V = 1$.

6.6.4 Comparison with earlier results

The only earlier VBF HH specific search was performed by the ATLAS collaboration in the HH→bbbb channel and is documented in [49]; the full Run 2 statistics, correspond-

Table 6.8 – 95% CL upper limits for $k_\lambda/k_t = 1$ on $\sigma(\text{VBF HH}) \cdot \text{BR}(\text{HH} \to \text{bb}\tau\tau)$ by $\tau\tau$ channel; in the VBF categories, a DNN based selection is implemented for the VBF vs. gluon fusion disambiguation. In the third column, the limit is expressed in terms of the SM prediction as a signal strength μ.

Channel	Obs. (exp.) upper limit on $\sigma \cdot \text{BR}$ [fb]	Obs. (exp.) $\mu = (\sigma \cdot \text{BR})/(\sigma \cdot \text{BR})_{\text{SM}}$
$\tau_h\tau_h$	112 (137)	885 (1092)
$\tau_\mu\tau_h$	107 (231)	851 (1836)
$\tau_e\tau_h$	321 (286)	2550 (2271)
Combined	62 (93)	492 (744)

Table 6.9 – 95% CL upper limits for $c_V = 1$, $c_{2V} = 1$ and $k_\lambda = 1$ on $\sigma(\text{VBF HH}) \cdot \text{BR}(\text{HH} \to \text{bb}\tau\tau)$ by category; in the VBF categories, a DNN based selection is implemented for the VBF vs. gluon fusion disambiguation. In the third column, the limit is expressed in terms of the SM prediction as a signal strength μ. The resolved 1b1j, resolved 2b0j and boosted categories are quoted for completeness, although their definition is unchanged with respect to the results presented in Tab. 6.4. It should be noted that the VBF tight category is only implemented in the $\tau_h\tau_h$ channel and for a fraction of data corresponding to about $27\,\text{fb}^{-1}$.

Channel	Obs. (exp.) upper limit on $\sigma \cdot \text{BR}$ [fb]	Obs. (exp.) $\mu = (\sigma \cdot \text{BR})/(\sigma \cdot \text{BR})_{\text{SM}}$
VBF loose	59 (109)	465 (868)
res. 2b0j	301 (441)	2426 (3557)
boosted	595 (682)	4792 (5492)
res. 1b1j	1569 (1484)	12449 (11781)
VBF tight ($\tau_h\tau_h$)	1129 (1469)	8961 (11657)
Combined	62 (93)	492(744)

ing to an integrated luminosity of $126\,\text{fb}^{-1}$, is exploited in this search. The observed (expected) 95% upper limit on the SM VBF HH production cross section is set to 1600 (1000) fb, i.e. about 941 (588) times the theoretical prediction.

6.7 Conclusion and perspectives

Two sets of results were obtained in the context of the non-resonant production: upper limits on the gluon fusion HH production cross section were computed, complemented by a limit scan as a function of the k_λ coupling; with a similar strategy, a specific VBF HH search was performed, using a dedicated DNN-approach and exploring $c_{2V} \neq 1$ scenarios. Given the small cross section and small acceptance, the VBF HH search is more ambitious; yet, the result is encouraging and competitive.

The search was performed using proton-proton collisions collected during 2017, corresponding to an integrated luminosity of $41.6\,\text{fb}^{-1}$. In table Tab. 6.10, the expected SM results quoted in Tab. 6.1 and Tab. 6.8, combined over channels and categories and quoted as signal strengths, are shown in the two cases and compared to their extrapolation in higher statistics scenarios. It should be noted that the extrapolation is only based on the integrated luminosity increase, without attempt to estimate the additional sensitivity brought by improved strategies or purer selections, which will be possible with higher statistics; therefore, it is a pessimistic estimate.

Table 6.10 – Upper limit at 95% CL on the signal strength, as quoted in Tab. 6.1 and Tab. 6.8, compared to the extrapolation corresponding to the full Run 2 integrated luminosity ($137.2\,\text{fb}^{-1}$), the integrated luminosity expected by the end of the Run 3 ($300\,\text{fb}^{-1}$) and that expected by the end of the Phase 2 of the LHC ($3000\,\text{fb}^{-1}$).

Data set	$\mathcal{L}$ [fb^{-1}]	$\mu = (\sigma_{\text{gg}\to\text{HH}}\cdot\text{BR})/(\sigma\cdot\text{BR})_{\text{SM}}$	$\mu = (\sigma_{\text{VBF HH}}\cdot\text{BR})/(\sigma\cdot\text{BR})_{\text{SM}}$
2017	41.6	20.7	744
Run 2	137.2	11.4	410
Run 3	300	7.7	277
Phase 2	3000	2.4	88

Exhaustive projection studies, taking into account the upcoming detector upgrades [140] and state-of-the-art machine learning analysis techniques, were carried out in ATLAS [141] and CMS [142] to estimate the HH searches potential in the Phase 2 physics program.

The reconstruction algorithms for HL-LHC conditions are not fully developed yet; therefore, the projection of the analyses performance rely on parametric object resolutions, efficiencies and misidentification rates. The CMS projections use a parametric simulation performed through the DELPHES [143] software. Similarly, in ATLAS projections, the parametric description of the objects is achieved by smearing the generated particles according to the expected response of the upgraded ATLAS detector.

The projected sensitivities of the gluon fusion HH CMS existing searches, of an additional $\text{HH} \to \text{bbZZ} \to 4\ell$ search and of their combination are summarised in Fig. 6.8. An integrated luminosity of $\mathcal{L} = 3000\,\text{fb}^{-1}$ and an average pileup PU = 200 are assumed. The projected upper limit at 95% with $k_\lambda = 1$ of the combination of the searches under

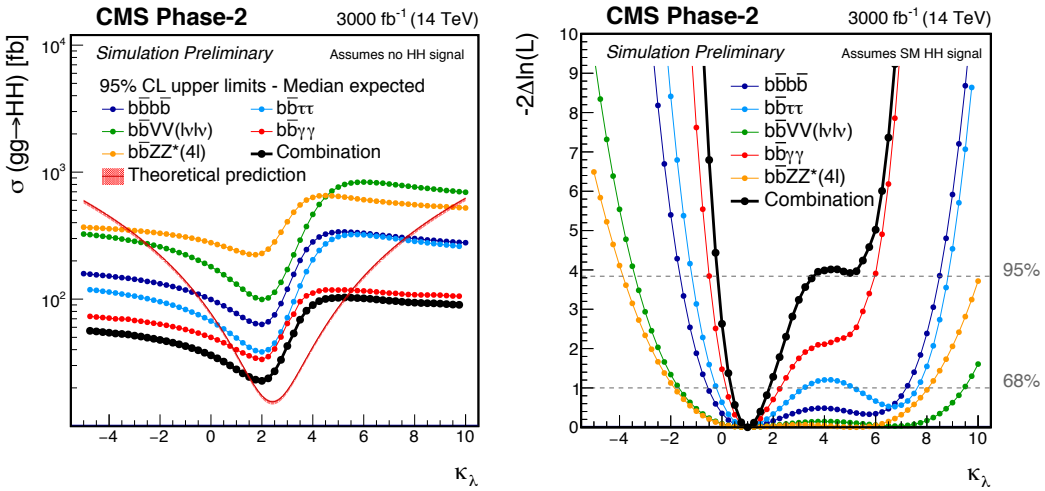


Figure 6.8 – Projected upper limit at 95% CL on $\sigma(\text{gg} \to \text{HH})$ (left) and minimum negative-log likelihood (right) with $\mathcal{L} = 3000\,\text{fb}^{-1}$ as a function of k_λ, obtained through CMS HH searches and their combination [142].

exam corresponds to $0.77 \times \sigma_{\text{SM}}$. More interestingly, by the projected results, k_λ will be constrained between 0.35 and 1.9 at 68% CL and between -0.18 and 3.6 at 95% CL: even before being able to observe the HH production, the of the $k_\lambda = 0$ scenario could be excluded, implying that the Higgs boson has a self-coupling constant and that the HH production can be observed; moreover, the second minimum of the likelihood will be excluded.

Combined ATLAS and CMS projections are also available in [28]: as represented in Fig. 6.9, the k_λ value will be constrained between about -0.5 and 0.5 at at 95% CL.

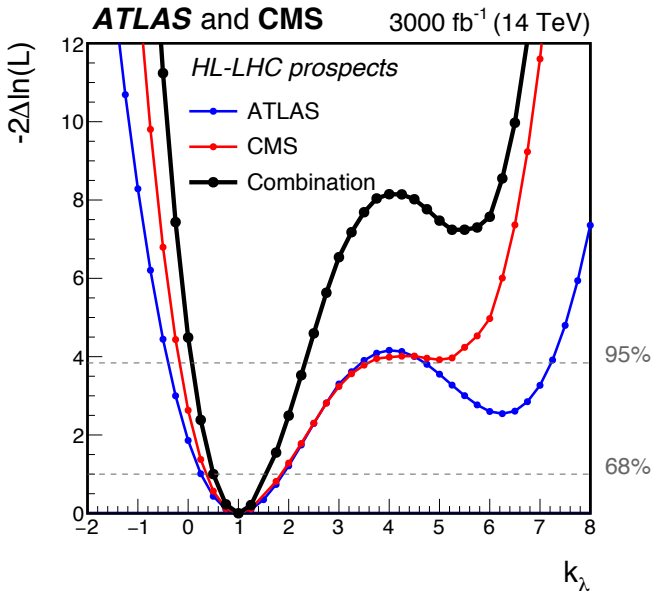


Figure 6.9 – Projected minimum negative-log likelihood, as a function of k_λ, of the combination of ATLAS HH searches and the combination of CMS HH searches; the black line represent their combined result. An integrated luminosity $\mathcal{L} = 3000\,\text{fb}^{-1}$ and a center-of-mass energy $\sqrt{s} = 14\,\text{TeV}$ are assumed [28].

It is worth observing that also these projection can be considered conservative. Previous projections of the HH $\to$ bb$\tau\tau$ search sensitivity, extrapolated from about $2.7\,\text{fb}^{-1}$ of early Run 2 collision runs for the 2016 European Committee for Future Accelerators (ECFA) workshop, are shown in Fig. 6.10. The most realistic scenario in this projection, taking into account for the estimated improvements in the detector performance and in the theoretical description of the physics processes, is represented with the red curve; the result obtained in the present search, of about $20 \times \sigma_{\text{SM}}$, was expected using an integrated luminosity larger by about a factor three.

In conclusion, the HH production observation is definitely within reach of the HL-LHC program.

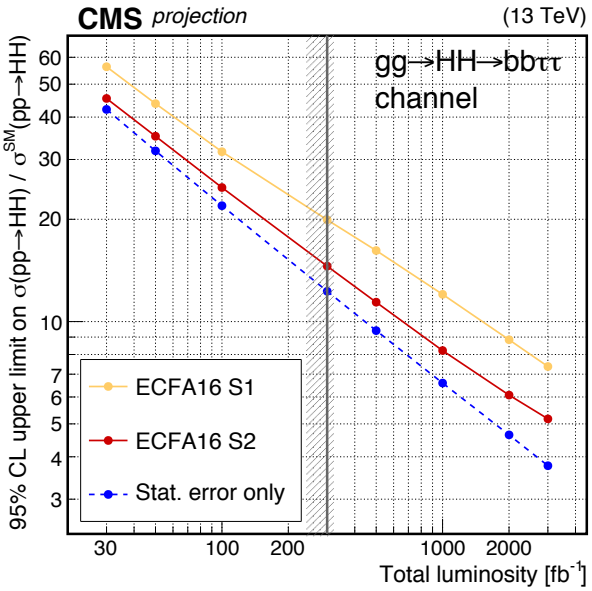


Figure 6.10 – Projected signal strength at 95% CL on the gg → HH production with early 2015 data. The grey line indicates the end of the LHC Phase 1; beyond it, the detector performance estimation are less reliable. The S1 curve is obtained under the assumption that all the systematic uncertainties constant as a function of the integrated luminosity and that the detector performances are unchanged; for the S2 curve, improvements are assumed in the detector performance and in the theoretical description of the physics processes; the dashed blue line is obtained under the assumption of systematic errors only [144].

Conclusion

The work presented in this thesis is focused on the search for production of two Higgs bosons at the LHC, either via gluon fusion or Vector Boson Fusion, in the decay channel where two b jets and two tau leptons are produced; about $41\,\mathrm{fb}^{-1}$ of data collected with the CMS detector in proton-proton collisions with center-of-mass energy $\sqrt{s} = 13\,\mathrm{TeV}$ are analyzed.

The analysis of the 2017 data set has proven to be particularly challenging and it required the whole analysis flow used in the previously published HH $\rightarrow$ bb$\tau\tau$ search, performed with 2016 data, to be re-examined and validated; however, the encountered experimental challenges lead to a stronger confidence in the analysis strategies.

Many improvements are introduced in the current search, from the object selection to the interpretation of the results. In particular, a large effort in the exploration of the VBF HH production mode is described in this manuscript; although its cross section is extremely small, a promising strategy is outlined for future searches.

In this context, the VBF topology is exploited starting from the L1 trigger operations by the first dedicated L1 VBF trigger algorithm, designed within this thesis work and included in the L1 trigger menu starting from 2017; although it is optimized for H $\rightarrow \tau\tau$ searches, the general nature of its selection make it useful also for searches relying on triggers targeting the production mode, such as H$\rightarrow$inv; dedicated HLT paths are built on top of it. The rewarding outcome of this work is the chance of implementing in the HH $\rightarrow$ bb$\tau\tau$ analysis a VBF category entirely populated by events collected by the L1 VBF trigger.

A VBF selection is implemented offline so that the inclusive HH $\rightarrow$ bb$\tau\tau$ analysis is complemented by additional VBF categories, independent from the L1 VBF trigger. In addition to the kinematic selection, a preliminary multivariate discriminant for a better disambiguation between the VBF and the gluon fusion signals is presented.

Thus, in addition to the gluon fusion HH search, allowing the k_λ value to be constrained, a VBF search is performed and limits are set on the $c_{2\mathrm{V}}$ coupling. The upper limit at 95% on the gluon fusion HH production is set to about 20 times the Standard Model prediction and the k_λ value is constrained at 95% CL by observed data to be in the range from -9.1 to 15.4. The observed (expected) limit for the VBF HH production cross section corresponds to about 500 (750) times the theoretical prediction; the $c_{2\mathrm{V}}$ coupling is constrained between -0.8 and 2.8; this is the first CMS dedicated VBF HH search.

Appendix A

Investigation on data-over-prediction disagreement in the $\tau_h\tau_h$ channel

The work of optimisation of the analysis strategies for the 2017 data was significantly slowed down due to the observation of a large data-over-prediction disagreement. It consists in a lack of background compared to the selected $\tau_h\tau_h$ data events, especially pronounced in regions populated by genuine hadronic tau leptons. A few months were devoted to the understanding of the origin of this problem; the corresponding studies were presented to the hadronic tau identification and trigger experts in several occasions. Eventually, a set of alternative tau identification scale factors was computed within this analysis to recover the agreement. In the following, a comprehensive summary of the possible causes of the disagreement and of the corresponding studies is presented; the computation of the alternative scale factor is described, along with its performance.

A.1 Description of the problem

As shown, for instance, in Sec. 4.3.1 and Sec. 4.3.2, the implementation of all the corrections that account for the different trigger, isolation and identification efficiencies in data and in simulations result in a satisfactory agreement over all the phase-space both in the $\tau_e\tau_h$ and $\tau_\mu\tau_h$ channels. Instead, when all the recommended corrections are applied in the $\tau_h\tau_h$ channel, a disagreement at the level of the 20% arises (see Fig. 4.13 and Fig. 4.13).

A.1.1 Reminders of the event selection and the simulation corrections

Some elements of the information detailed in Sec. 4.2 and Sec. 4.3, regarding the hadronic tau selection in each channel, are summarised in the following.

In the $\tau_e\tau_h$ channel and $\tau_\mu\tau_h$ channel, events are collected using single-ℓ and cross $\ell\tau_h$ triggers ($\ell = e, \mu$), while the $\tau_h\tau_h$ events are collected using di-τ_h triggers. The HLT sequence for the reconstruction of the τ_h-legs of the di-τ_h triggers goes over three levels (L2, L2.5 and L3) with increasingly tight requirements; this algorithm is regional, i.e. it reconstructs the HLT τ_h candidate in the direction of the L1 τ_h candidate. The τ_h-leg in the cross triggers have similar HLT sequence, although they miss the L2 and L2.5 filters and their reconstruction is global.

The offline thresholds are driven by the trigger requirements: each object should be produced in a region of the phase-space efficiently covered by the trigger. The selection is the result of the logic OR, within the same $\tau\tau$ final state, of the available HLT paths and the corresponding offline selections.

The efficiency of the τ_h-leg selection in cross $\ell\tau_h$ triggers is measured through a tag-and-probe method in $Z\to\tau\tau$ events. The computation of the efficiency of the tau τ_h-legs is performed with a similar strategy, exploiting a monitoring $\mu\tau_h$ trigger with a hadronic tau HLT sequence that is identical to the one used for the τ_h legs of the di-τ_h trigger; the trigger scale factor in $\tau_h\tau_h$ events is given by the product of the tau trigger scale factors corresponding to each leg, as they are considered independent.

Event-by-event scale factors are applied to simulated events to take into account the differences of the trigger efficiency compared to data. For what concerns the τ_h-legs, the scale factors are applied based on the transverse momentum and decay mode of each of the selected τ_h legs; the differences among the three decay modes can be appreciated in Fig. 4.8.

At the analysis level, the hadronic tau leptons are required to pass a MVA-based identification; additionally, electrons and muons misidentified as hadronic tau leptons are rejected through dedicated discriminators. Tau identification and anti-ℓ discrimination scale factors are proposed by the Tau POG; they should be used in simulated events based on the geometric match between the reconstructed hadronic tau leptons and the generated objects: for each hadronic tau lepton in the selected $\tau\tau$ pair correctly matched, a scale factor $SF_{tauID} = 0.89$ should be applied. This scale factor does not depend on the reconstructed hadronic decay mode, nor on the p_T of the hadronic tau. It is applied indifferently on hadronic tau leptons in the $\tau_h\tau_h$, $\tau_e\tau_h$ and $\tau_\mu\tau_h$ channels.

A.1.2 Comparison with the $H\to\tau\tau$ analysis

The event selection in the $H\to\tau\tau$ analysis of the 2017 data [96] has only very negligible differences with respect to the $\tau\tau$ pair selection described in Sec. 4.3. However, different strategies are implemented for the background modelling.

In the $HH\to bb\tau\tau$ search, a data-driven ABCD method is used for the QCD computation (see Sec. 5.3). In the $H\to\tau\tau$ analysis, all the backgrounds with jets misidentified as taus are estimated at once through a data-driven method through a "fake factor" strategy. This background is denoted as "jet$\to\tau_h$" in Fig. A.1 and it incorporates the QCD background, the W+jets background and, partially, the $t\bar{t}$ background. As they are evaluated from data and not from simulation, they are not subjected to the application of tau trigger nor tau identification scale factors.

As for the Drell-Yan background, in this analysis it is modelled using LO simulations with up to 4 jets produced in the hard interaction; a data set with simulated events where two b quarks are produced within the DY process are also included. The $H\to\tau\tau$ analysis, instead, uses embedded Drell-Yan samples. These samples are produced from recorded $Z\to\mu\mu$ events and undergo a complex treatment where the real muons are replaced with fully simulated hadronic tau leptons. One important advantage of this method is that the additional jets, the recoil, the pileup and, in general, the underlying physics, are by construction extremely well modelled since taken from data.

The tau leptons in embedded data sets are just like those in simulated events. However, the underlying physics in data can have an effect on the isolation of the simulated tau lepton. Therefore, independent tau trigger and tau identification scale factors are com-

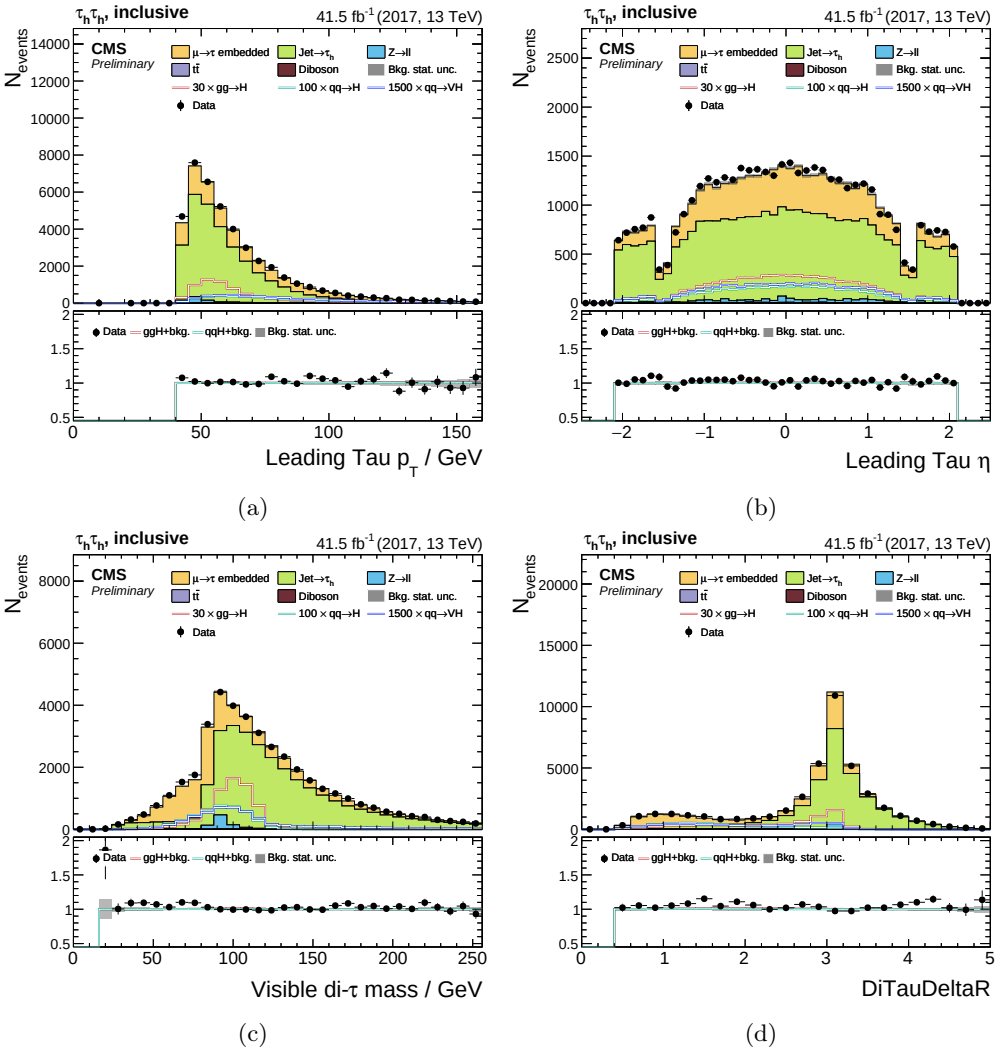


Figure A.1 – Data and background event distributions in the $\tau_h\tau_h$ channel with the $H \to \tau\tau$ 2017 analysis strategies and selection [145]: backgrounds with jets misidentified as hadronic tau leptons are modelled through the fake factor method; the Drell-Yan background is estimated from embedded data sets. Events are required to pass the selection described in Sec. 4.3. The event distribution is shown as a function of the main kinematic observables that exhibit a large disagreement in the $HH \to bb\tau\tau$ analysis. All the correction scale factors (trigger and identification) recommended by the Tau POG are applied.

puted for the embedded samples. These scale factors are, in average, closer to 1 than the ones applied for fully simulated events.

As the "jet$\to \tau_h$" background and the Drell-Yan background are largely dominant, the $H \to \tau\tau$ analysis is not subjected to the regular tau trigger and tau identification scale factors used in this search.

In conclusion, a good agreement is observed in Fig. A.1. However, the embedded Drell-Yan data sets are not suited for this analysis: they do not guarantee adequate statistics for the modelling of the irreducible DY+2 b jets background, which is essential in the most sensitive regions of the $HH \to bb\tau\tau$ search.

A.1.3 Data vs. simulation with recommended scale factors

Preliminary considerations can be drawn from $\tau_h\tau_h$ event distributions as a function of some of the main kinematic variables, as the transverse momenta the pseudorapidity of the two selected hadronic tau leptons, as well as their ΔR separation and their invariant mass $m_{\tau\tau}^{vis}$. They are shown in Fig. A.2, Fig. A.3, Fig. A.4 and Fig. A.5 using different selections: events passing the $\tau_h\tau_h$ selection described in Sec. 4.3; events passing the $\tau_h\tau_h$ selection, where at least two b jet candidates are found; $\tau_h\tau_h$ events passing the resolved 1b1j requirements; $\tau_h\tau_h$ events passing the resolved 2b0j requirements.

Firstly, the extent of the disagreement observed with the $H \to \tau\tau$ selection (Fig. A.2) and the $H \to \tau\tau$ + 2 jets selection (Fig. A.3) is comparable. Undesired effects due to the jet selection, therefore, are excluded.

Secondly, the largest difference between the background and data distribution corresponds to events with low hadronic tau p_T. A bad modelling of the QCD background can have this effect; also, the QCD background contamination is larger in the $\tau_h\tau_h$ channel than in the semileptonic final states, which would explain why the data-over-prediction agreement is degraded in the $\tau_h\tau_h$ channel only. In the most inclusive selections, the distribution of the leading hadronic tau as a function of η seems to point to an inefficiency on the background in the barrel (e.g., Fig. A.2b); however, this trend is less evident for the second hadronic (Fig. A.2d) tau and is mitigated by additional b tag requirements (e.g. Fig. A.4b).

Finally, a large disagreement is observed in regions mostly contaminated by Drell-Yan background, as observed in the $\Delta R(\tau_h, \tau_h)$ and the $m_{\tau\tau}^{vis}$ distributions (e.g., Fig. A.2f and Fig. A.2e); in those regions, with $\Delta R(\tau_h, \tau_h) < 2$ and $50\,\text{GeV} < m_{\tau\tau}^{vis} < 120\,\text{GeV}$, the disagreement is mitigated by applying tight b tag criteria, i.e. by suppressing the Drell-Yan contribution (Fig. A.5d and Fig. A.5c); however, although the agreement is globally satisfactory in Fig. A.5, a closer inspection in the Drell-Yan dominated regions shows a systematic underestimation of the event yield. On one hand, this behaviour can attributed to a Drell-Yan background mismodelling. On the other hand, since the Drell-Yan background is composed mainly of events with genuine hadronic tau leptons, a mismodelling of the Drell-Yan background arising only in the $\tau_h\tau_h$ channel can point to a mismatch of the tau identification efficiency between data and simulation. The corresponding scale factor of 0.89, indeed, is only applied in events with genuine hadronic taus; therefore, most of the Drell-Yan events are scaled by $\text{SF}_{\text{tauID}} \times \text{SF}_{\text{tauID}} = 0.89^2 = 0.79$. A tau identification, however, is also performed at trigger level; the corresponding requirements are listed in Tab. 4.1. Therefore, an inefficiency affecting well identified hadronic tau leptons can also arise from the trigger efficiency measurement. Finally, an unexpected interplay between the tau trigger and tau identification efficiency measurements, like the double counting of an inefficiency, can cause this behaviour. It should be remarked that the effect of the combined use of the tau trigger and identification scale factor is quite large: in Fig. 4.8a, it can be seen, for instance, that for hadronic tau leptons decaying in $h^\pm$ or $h^\pm\pi^0$, which are the majority, and with with $p_T \sim 50\,\text{GeV}$, the tau trigger scale factor is ~ 0.90; therefore, if the are also genuine hadronic tau leptons, the final scale factor is $0.89^2 \cdot 0.90^2 = 0.64$.

Changes in data-taking conditions

As largely discussed, the 2017 data were collected using several different beam configurations, to which the data acquisition chain coped with differing performance; moreover, the degradation of the ECAL crystals placed in the forward regions of the detectors became relevant and it increasingly affected the data quality. However, no significant

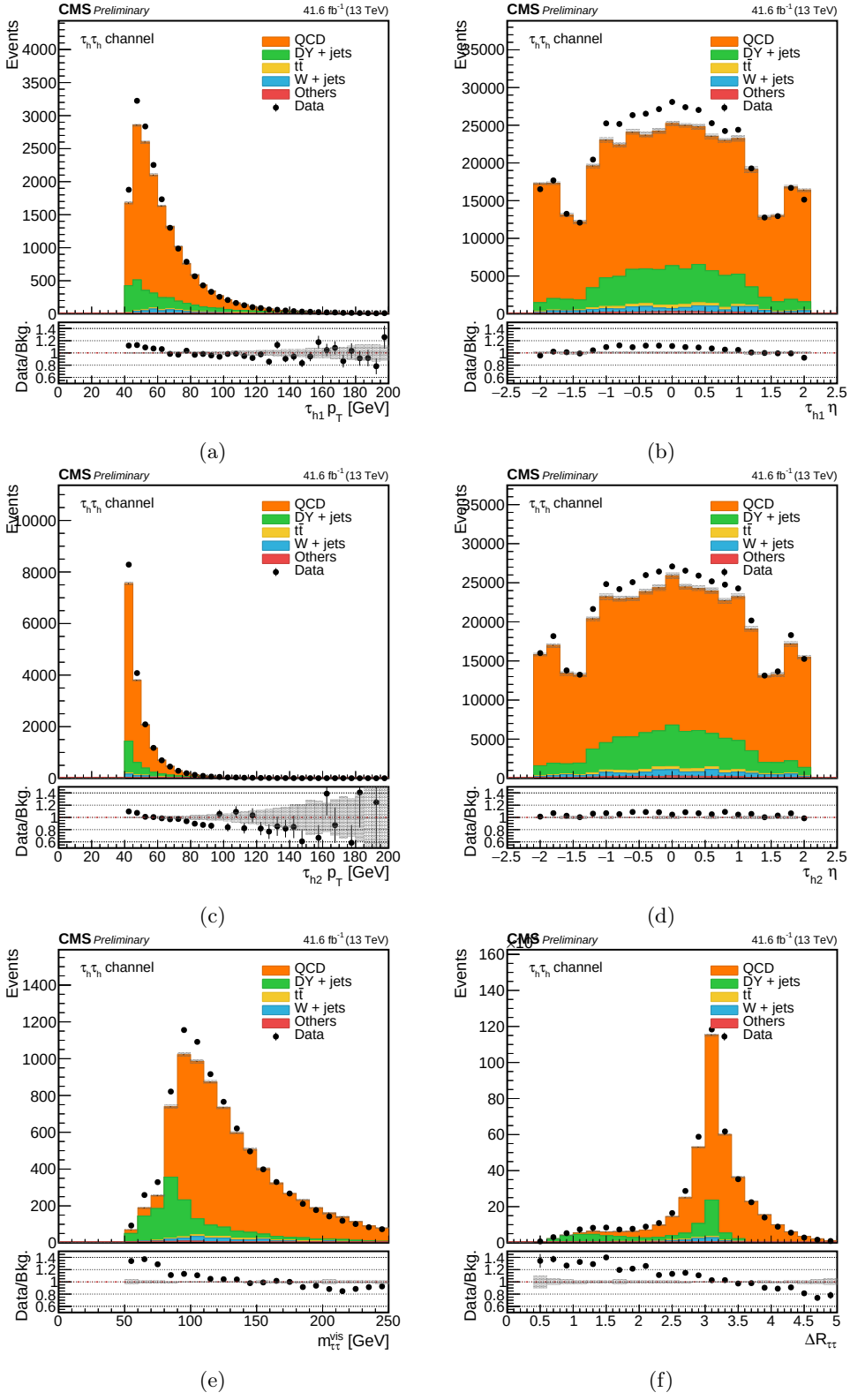


Figure A.2 – Distribution of $\tau_h\tau_h$ data and background events, selected as described in Sec. 4.3, as a function of the main kinematic variables. All the correction scale factors (trigger and identification) recommended by the Tau POG are applied.

dependence on time was observed; therefore, the disagreement is not to be ascribed to changes in the data-taking conditions, although it is not excluded that they minorly impact the background modelling.

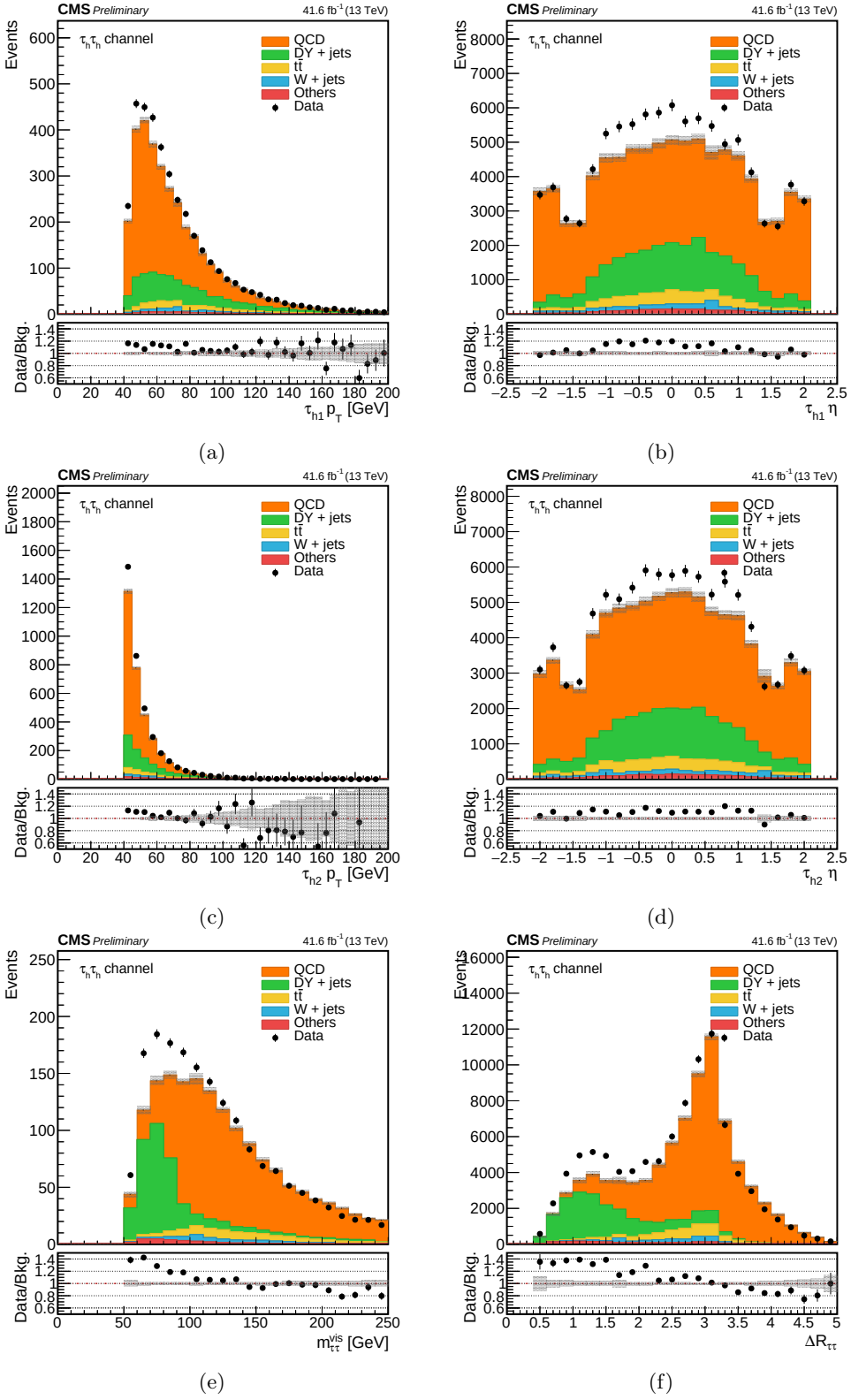


Figure A.3 – Distribution of $\tau_h\tau_h$ data and background events passing the baseline selection of this search, i.e. where two b jet candidates are found. All the correction scale factors (trigger and identification) recommended by the Tau POG are applied.

Drell-Yan modelling

The description of the modelling and reweighting technique applied to the Drell-Yan data

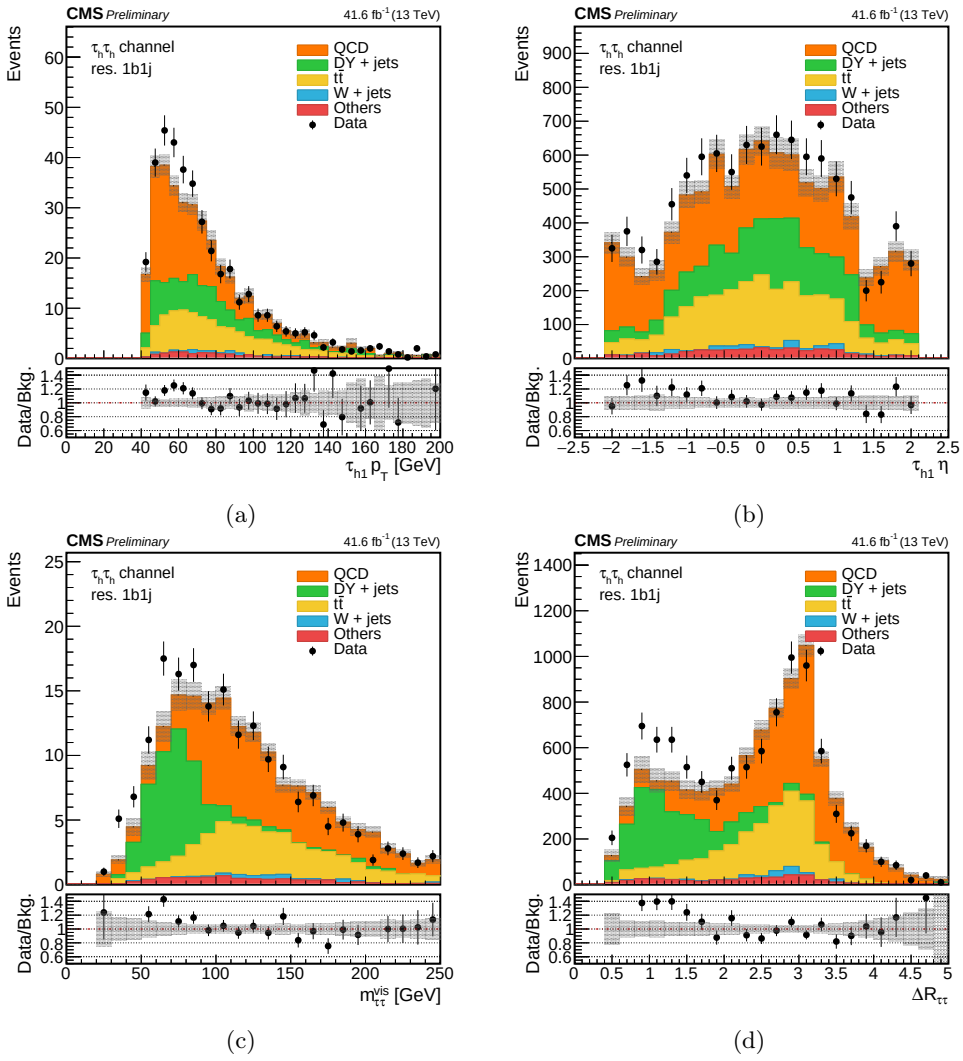


Figure A.4 – Distribution of $\tau_h\tau_h$ data and background events passing the res. 1b1j requirements. All the correction scale factors (trigger and identification) recommended by the Tau POG are applied.

sets used in this search can be found in Sec. 5.4. In 2017, the Drell-Yan inclusive cross section measurement was improved, giving a result about 7% larger than the one used in the previous analyses. The Drell-Yan background events are correctly normalised to the most up to date recommendation. However, the joint use of several Drell-Yan simulation samples requires a relative normalisation of the different processes to be applied; a wrong estimation of the exclusive cross sections could lead to an imbalance and spoil the data-over-prediction agreement in some regions.

On one hand, it was shown already in Sec. 5.4 that a very good agreement is achieved in $\mu\mu$ events; the event distribution as a function of the transverse momentum of the reconstructed Z boson is shown in Fig. A.6a, in the selection used for the derivation of the Drell-Yan weights. Also, the Drell-Yan modelling is proven to have good performance in the $\tau_e\tau_h$ and $\tau_\mu\tau_h$ channels. On the other hand, Drell-Yan events selected in $\tau_h\tau_h$ belong to a different region of the phase-space, as τ_h-legs have harder p_T spectra. In Fig. A.6b, the distribution of the p_T of the di-muon system, e.g. the p_T of the Z candidate, is shown for $\mu\mu$ events where a $p_T > 60\,\text{GeV}$ selection is applied. The data-over-prediction agreement is excellent in all the considered range.

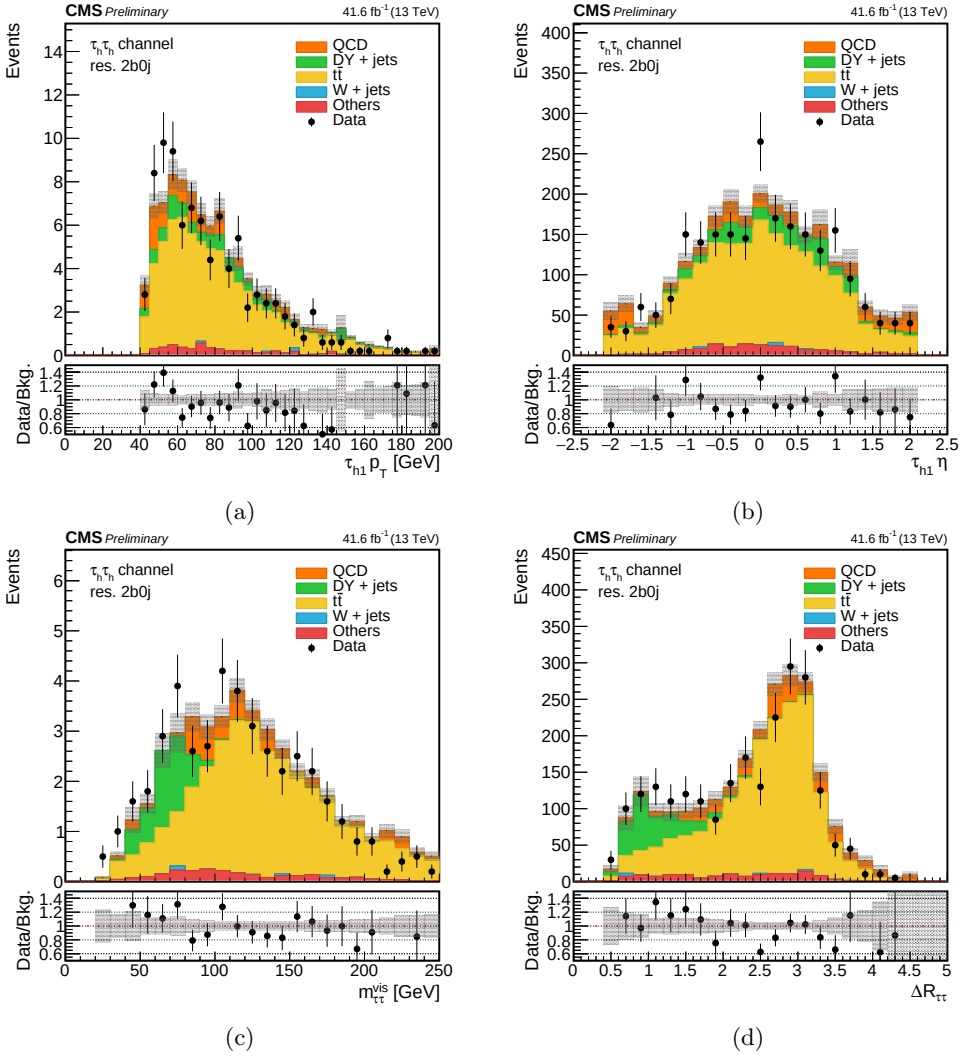
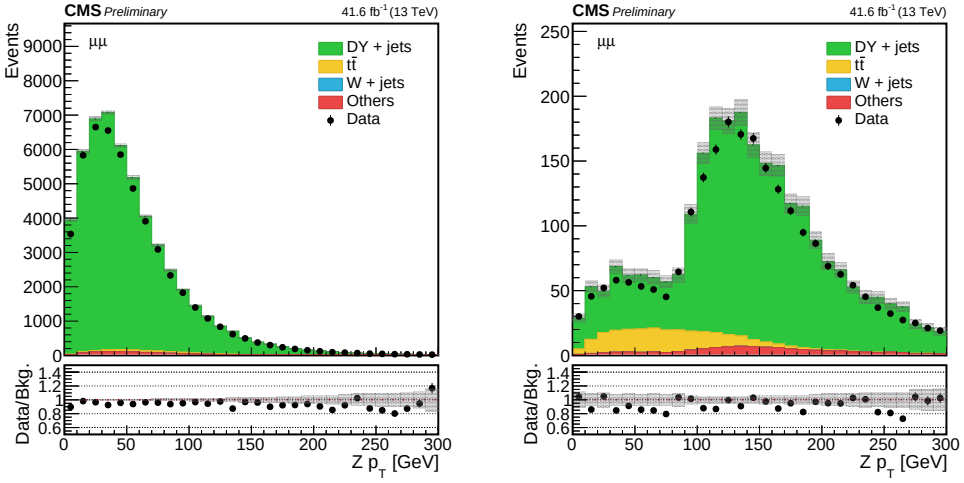


Figure A.5 – Distribution of $\tau_h\tau_h$ data and background events passing the res. 2b0j requirements. All the correction scale factors (trigger and identification) recommended by the Tau POG are applied.

In figure Fig. A.7, some event distributions in the $\tau_h\tau_h$ channel are replicated using an inclusive NLO Drell-Yan simulation; no recover of the data-over-prediction agreement is observed. This check provides two useful pieces of information: in the first place, it is unlikely that a bad modelling affects two independent data sets, produced with different Monte Carlo generators, as the NLO and LO Drell-Yan simulations; in the second place, the comparison between the inclusive NLO sample and the joint use of several LO samples shows that the overall normalization is consistent.

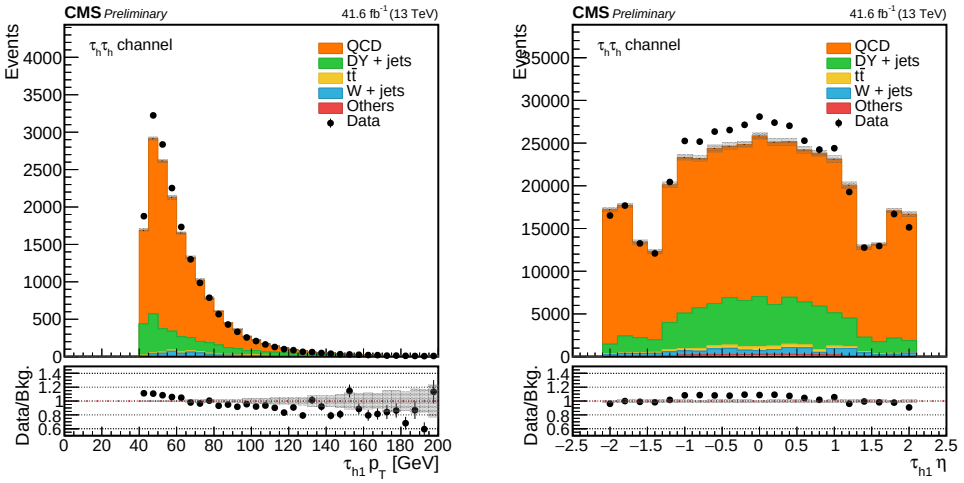
Multijet background estimation

The multijet QCD data-driven estimation is described in Sec. 5.3. It consists of a ABCD method, where the distribution of the QCD background in the signal region A is estimated from a jet-enriched region B; the A and B regions are orthogonal: the $\tau\tau$ candidates should have opposite-sign charge in the signal region and same-sign charge in the B region. To compute the QCD event yield, an extrapolation factor is computed in C and D regions, where the τ_h-leg identification requirement is inverted. In practice, it exploits the fact that the ratio of opposite-sign and same-sign events is expected to be the same whether the jet$\rightarrow\tau_h$ candidates pass tight tau identification or not. In the



(a) (b)

Figure A.6 – Distribution of $\mu\mu$ data and background events as a function of the transverse momentum of the Z candidate, selected as described in Sec. 5.4 (left) and with the additional $p_T > 60\,\text{GeV}$ requirement on both muons (right). The Drell-Yan weights used in the analysis are implemented.

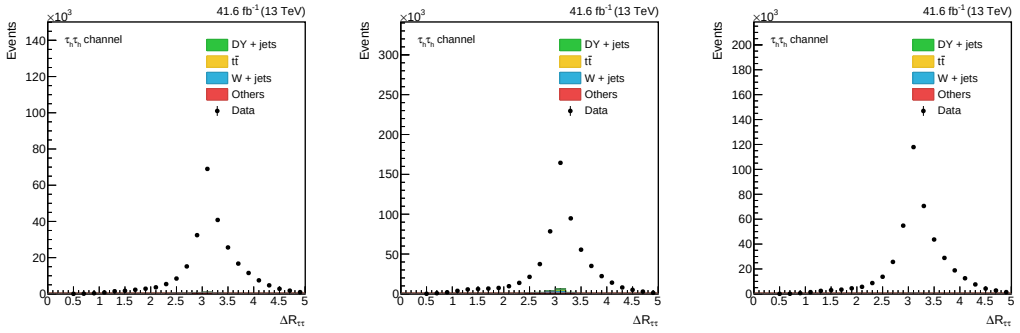


(a) (b)

Figure A.7 – Distribution of $\tau_h\tau_h$ data and background events, selected as described in Sec. 4.3, as a function of the transverse momentum and the pseudorapidity of the first hadronic tau lepton. All the correction scale factors (trigger and identification) recommended by the Tau POG are applied. Instead of the usual Drell-Yan simulations generated with LO precision, Drell-Yan NLO simulations are used.

control B, C and D regions, the number N of QCD events is estimated by subtracting the Monte Carlo events to the data distribution; the QCD event yield is computed as $N_B \times N_C/N_D$. Various checks were performed to ensure that there is not a large QCD mismodeling.

The data and simulation distributions in the B, C and D regions are compared in Fig. A.8 as a function of $\Delta R(\tau_h, \tau_h)$. These are the regions from which the QCD contribution in A is extrapolated: the large gap between data and simulated events corresponds to QCD contamination. As expected, the contribution of simulated background events is minimal: some Drell-Yan and W+jets contamination can barely be observed in the C region (Fig. A.8b).



(a) B - *Medium* ID, SS (b) C - not *Medium* ID, OS (c) D - not *Medium* ID, SS

Figure A.8 – Event distribution in the B, C and D sidebands of the ABCD method for the QCD estimation described in Sec. 5.3, as a function of the ΔR separation of the two hadronic tau leptons selected as described in Sec. 4.3. The contribution of backgrounds estimated from simulations is negligible, as expected.

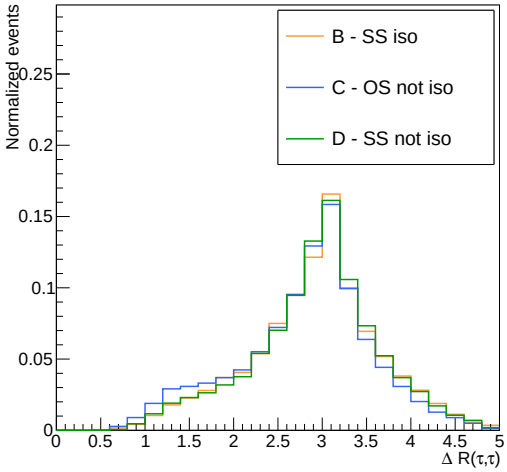


Figure A.9 – Normalised $h_{data} - h_{MC}$ distributions as a function of $\Delta R(\tau_h, \tau_h)$ in the B, C and D regions of the ABCD method (Sec. 5.3). The events are required to pass the baseline selection, i.e. to pass the $\tau_h\tau_h$ selection and to have two b jet candidates.

In Fig. A.9, the differential distributions of $N_i = N_i^{data} - N_i^{MC}$ as a function of $\Delta R(\tau_h, \tau_h)$ in B, C and D are shown for $\tau_h\tau_h$ events where two b jet candidates are found, i.e. for events passing the baseline selection used in this search. The shape of the distributions is reasonably similar; however, the ABCD definition used in this search could be suboptimal in capturing the effect of gluon splitting, occurring in the low $\Delta R(\tau_h, \tau_h)$ region: rather than estimating the QCD background from the B region, with same-sign candidates, this effect could be better modelled by using the C region, with opposite-sign not isolated candidates. Event distributions obtained through such alternative ABCD method design are represented in Fig. A.10: the shape of the QCD distribution is estimated from C; the QCD event yield is computed as $N_C \times N_B/N_D$. Although the data-over-prediction ratio becomes flatter as a function of angular variables, as the pseudorapidity and of the $\Delta R(\tau_h, \tau_h)$, the trend of the disagreement becomes more acute as a function of the transverse momentum of the hadronic tau candidate and of the invariant mass of the $\tau\tau$ pair candidates. Aside from a better QCD shape in some event distributions, the overall data-over-prediction agreement does not improve. Therefore, although the ABCD method is perfectible, a strategy change does not seem necessary.

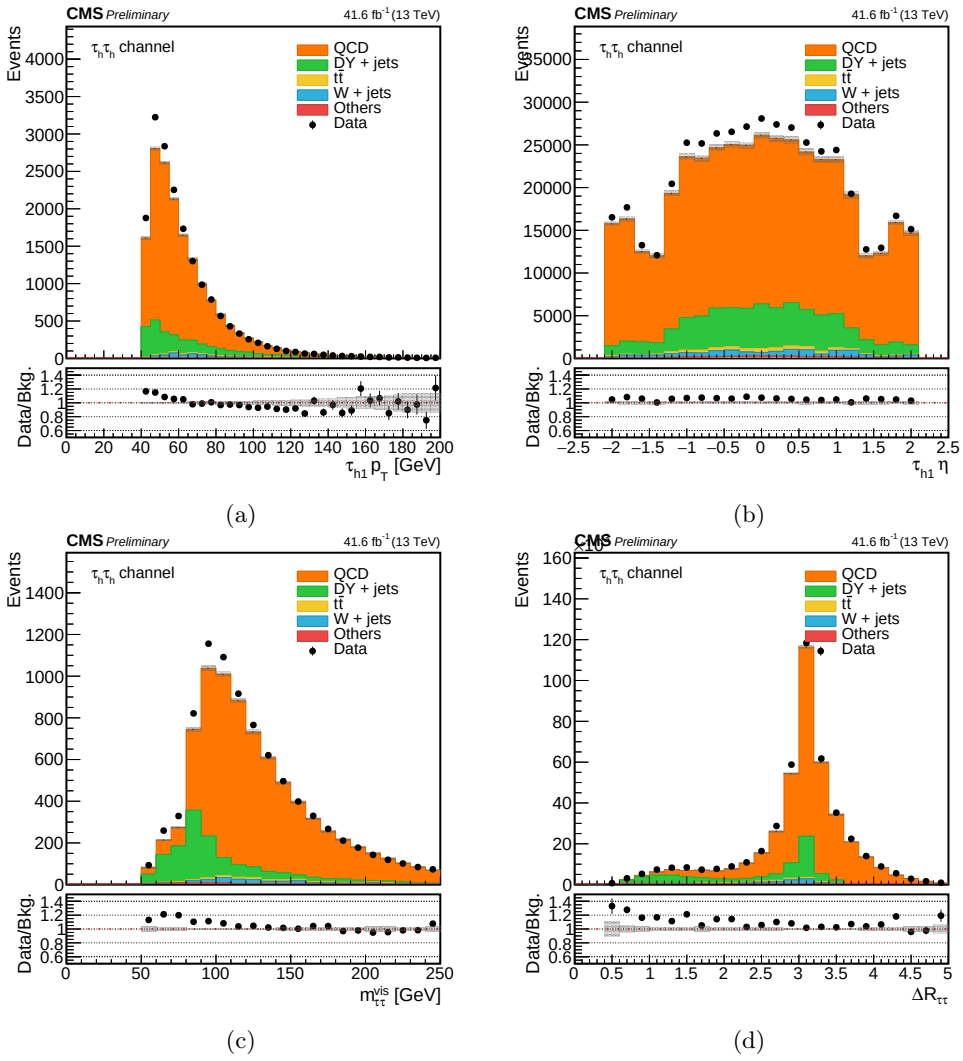


Figure A.10 – Distribution of $\tau_h\tau_h$ events passing the selection described in Sec. 4.3. The extrapolation of the QCD from ABCD regions is rearranged compared to the method described in Sec. 5.3: the shape of the QCD distribution is estimated from C; the QCD event yield is computed as $N_C \times N_B/N_D$. All the correction scale factors (trigger and identification) recommended by the Tau POG are applied.

A more conclusive test on the validity of the ABCD method comes from the comparison with the $H \to \tau\tau$ analysis.

Synchronization with the $H \to \tau\tau$ analysis

A synchronization with the $H \to \tau\tau$ analysis framework helped revealing a disagreement in $H \to \tau\tau$ control plots of the same extent of the one observed in the $HH \to bb\tau\tau$ analysis.

Some $H \to \tau\tau$ control plots, where the $HH \to bb\tau\tau$ analysis strategy is reproduced, are shown in Fig. A.11: instead of the fake method, the QCD estimation used in this analysis (and in past $H \to \tau\tau$ searches) was implemented; also, instead of the embedded data set, the LO samples used in this search were used for the Drell-Yan background estimation. A lack of data is observed, also in this case, in the region mostly populated by Drell-Yan events, i.e. with genuine hadronic tau leptons.

The versatility of the $H \to \tau\tau$ framework allows to compare singularly the effects of the analysis strategy changes. In Fig. A.12, the event distributions are reproduced using the

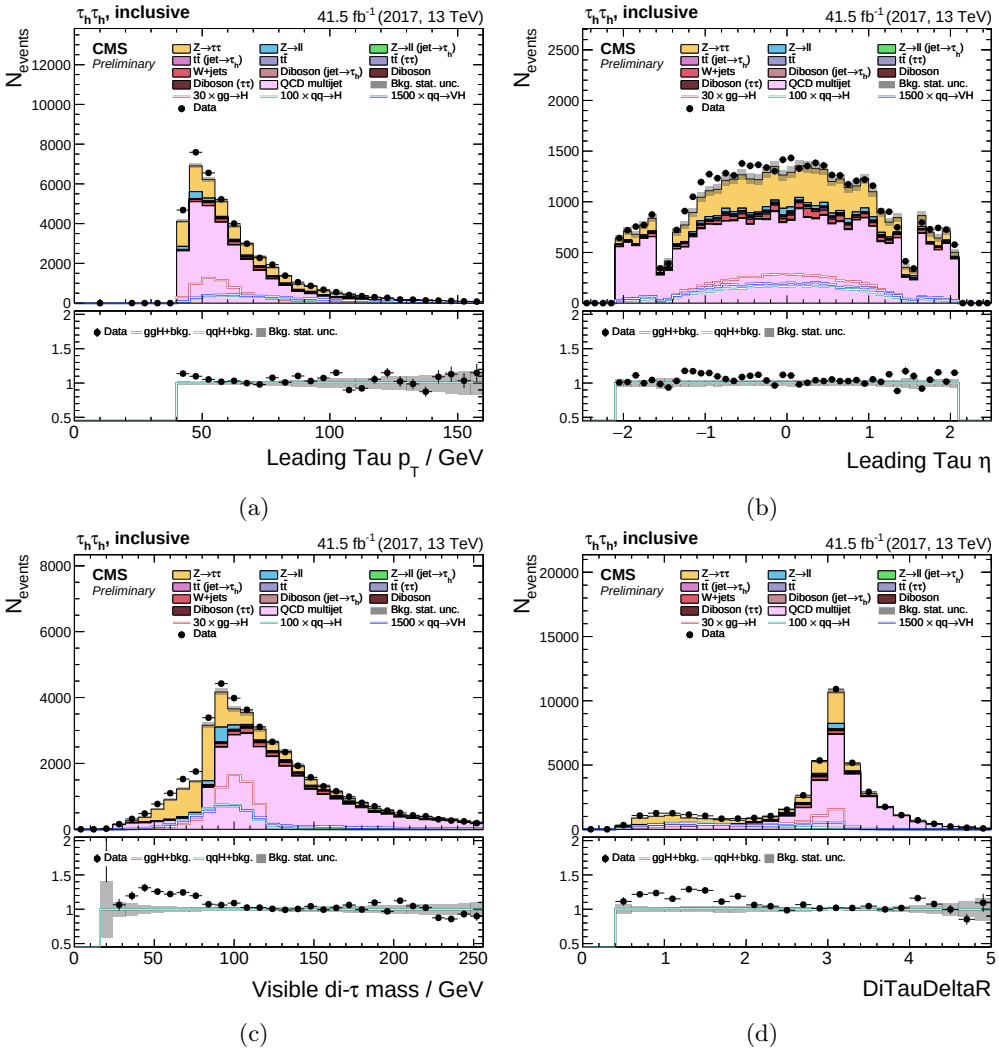


Figure A.11 – Control plots of the H $\to \tau\tau$ 2017 analysis [145]. Instead of the analysis strategies used in [96], the HH $\to$ bb$\tau\tau$ strategies are implemented: the ABCD method is used to estimate the QCD background contribution; the Drell-Yan background is estimated from LO simulations. Events are required to pass the selection described in Sec. 4.3. All the correction scale factors (trigger and identification) recommended by the Tau POG are applied.

fake factor method for the estimation of backgrounds with fake jets; LO simulations are implemented for the Drell-Yan background estimation. No improvement is observed with the use of the fake factors method compared to the use of the ABCD method; therefore, it seems unlikely that the origin of the disagreement is a bad multijet background modelling.

For completeness, the other possible combination, consisting in using the ABCD method for the QCD estimation and the embedded samples to model the the Drell-Yan background, is represented in Fig. A.13. The use of embedded Drell-Yan samples, together with their tau trigger and tau identification scale factors, makes the difference in recovering the data-over-prediction agreement.

Tau trigger and tau identification scale factors

The tau trigger efficiency is measured exploiting the τ_h-leg in monitoring cross $\mu\tau_h$ triggers. It is assumed that the efficiencies on the two τ_h legs of the di-τ_h trigger are independent; thus, the scale factor to be applied to the event is the product of the scale

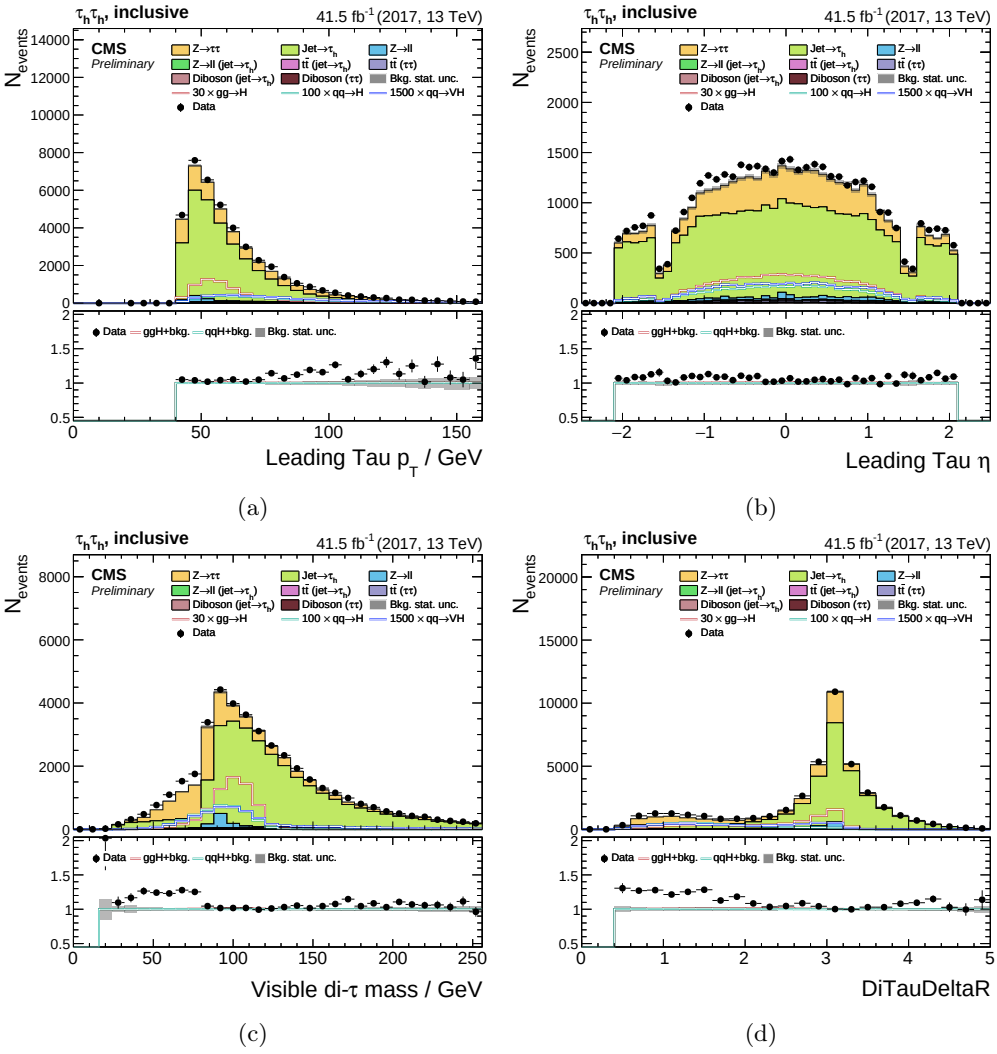


Figure A.12 – Control plots of the H $\to \tau\tau$ 2017 analysis [145]. The jet$\to \tau_h$ background contribution is modelled through the fake factor method described in [96]; the Drell-Yan background is estimated from LO simulations. Events are required to pass the selection described in Sec. 4.3. All the correction scale factors (trigger and identification) recommended by the Tau POG are applied.

factors corresponding to each τ_h-leg. This assumption is generally safe, as long as the inefficiencies are local; it is not to be excluded, though, that a global inefficiency is counted twice, once for each leg. The same holds for the tau identification efficiency.

Moreover, as mentioned earlier, both the tau trigger and tau identification select well identified hadronic tau leptons, like the majority of those reconstructed in the Drell-Yan background, and they are not fully independent: an object passing the HLT hadronic tau lepton requirements, including identification and isolation, is more likely than other objects to also pass the offline tau identification. At the offline stage of the analysis, it is not trivial to disentangle these effects.

However, as a strong decay mode dependency affects the tau trigger efficiency, the decay mode should also be relevant for the tau identification. In Fig. A.16, the number of data and background events is represented in bins of $\tau\tau$ decay mode combinations. An improved data-over-prediction is observed anytime one of the hadronic tau leptons decays in $h^{\pm}h^{\mp}h^{\pm}$. In Fig. A.14, the event distribution of events where the second τ_h leg decays in $h^{\pm}h^{\mp}h^{\pm}$ are shown as a function of the main kinematic variables; for these events, the

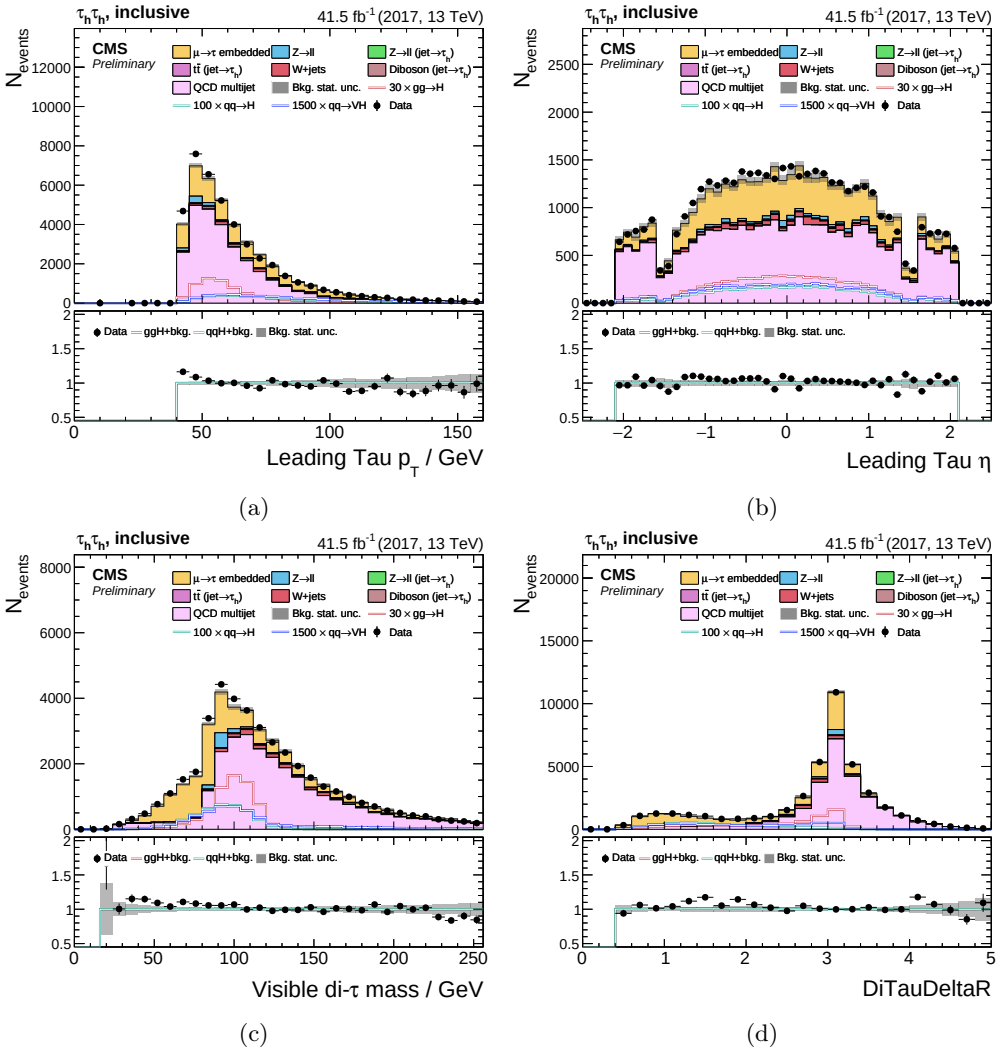


Figure A.13 – Control plots of the H $\to \tau\tau$ 2017 analysis [145]. The QCD background contribution is modelled through the ABCD method used in this search; the Drell-Yan background is estimated from embedded data sets. Events are required to pass the selection described in Sec. 4.3. All the correction scale factors (trigger and identification) recommended by the Tau POG are applied.

agreement is much improved. From this observation, decay mode dependent scale factors used instead of the tau identification scale factors are computed within this analysis. A measurement of tau identification scale factors decay mode dependent was also provided by the Tau POG at the time of the writing of this manuscript; therefore, they could not be implemented in this search.

A.2 Alternative tau identification scale factors

As no satisfactory SF were centrally provided, I tackled the problem fixing simultaneously two issues:

- a direct determination of the efficiency in the $\tau_h\tau_h$ channel, i.e. in events triggered in the same way as in the analysis;
- a DM-dependent measurement.

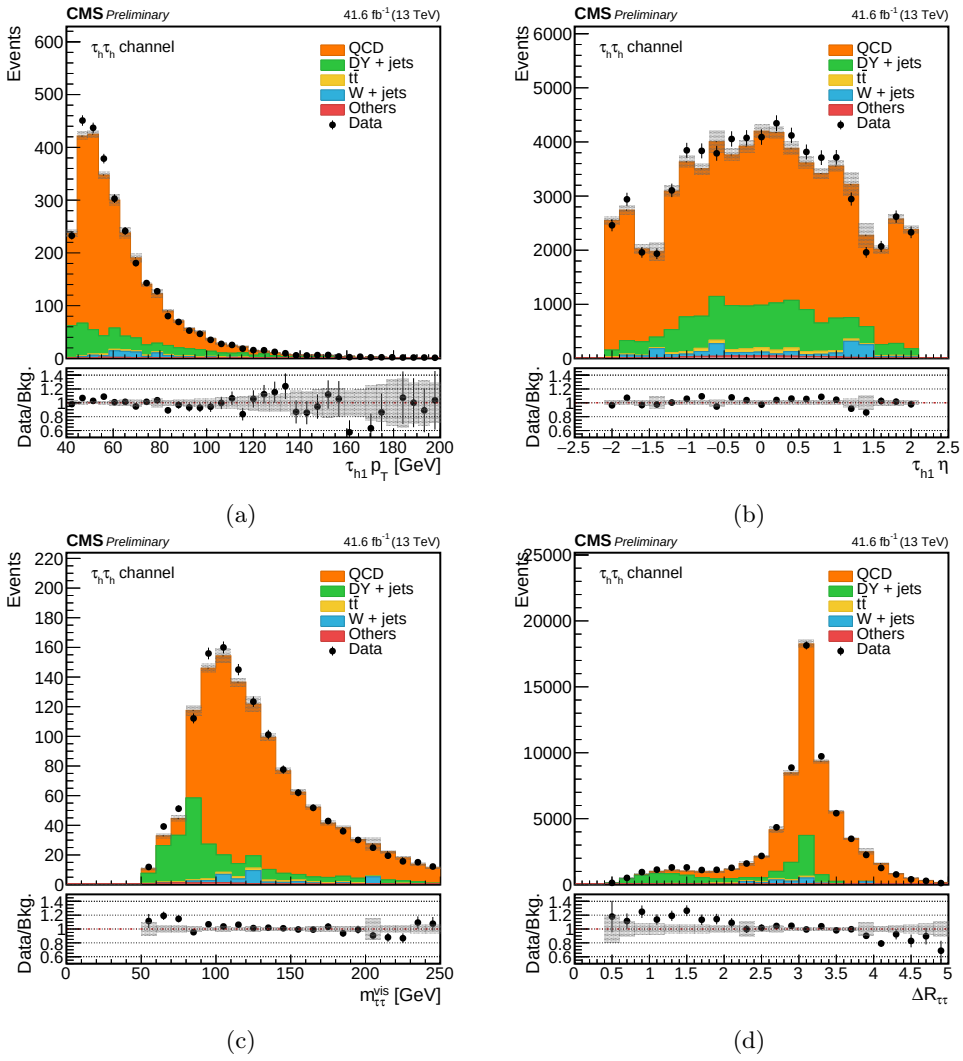


Figure A.14 – Distribution of $\tau_h\tau_h$ data and background events, selected as described in Sec. 4.3, as a function of the main kinematic variables. Only events where the second hadronic tau lepton decays in $h^{\pm}h^{\mp}h^{\pm}$ are selected. All the correction scale factors (trigger and identification) recommended by the Tau POG are applied.

Alternative scale factors (SF_{DM}) are computed in events passing the regular $\tau_h\tau_h$ selection and with $\Delta R(\tau_h, \tau_h) < 2$. Thus, a region relatively pure region relatively pure in genuine τ_h is defined: as can be seen looking at the "inclusive" bin of Fig. A.15, the fraction of Drell-Yann events thus selected amounts to about 63%; a small residual QCD background contamination can also be seen.

The recommended tau identification scale factor is not applied; all the other corrections, including the tau trigger scale factors, are implemented as recommended. A correction SF_{DM} is extrapolated within each of the bins represented in Fig. A.15 where the tau-legs have same decay mode, i.e. in fully independent decay mode combinations. The simulated events are split in events with one, two and zero genuine hadronic tau leptons; global event yield of simulated events in these category should match the number of data events, through the variation of the multiplicative factors SF^2_{DM}, for events with two real tau leptons, and SF_{DM} for events with one real tau lepton. As the QCD estimation is affected by the changes in the event yield of the simulated backgrounds in the sidebands, the measurement is performed through a simultaneous fit in the four ABCD regions.

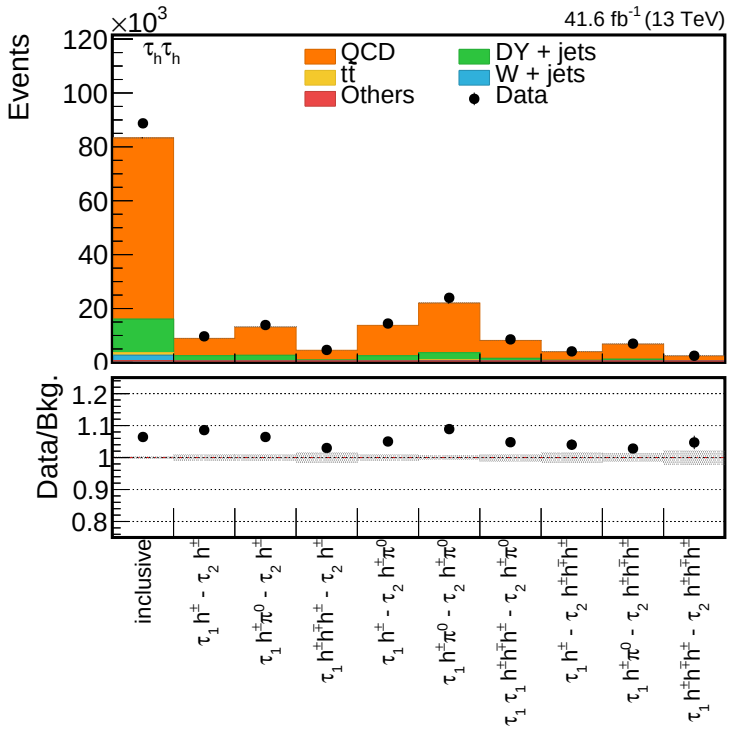


Figure A.15 – Number of data and background events passing the $\tau_h\tau_h$ selection described in Sec. 4.3 as a function of the hadronic decay mode of the two selected tau leptons. All the correction scale factors (trigger and identification) recommended by the Tau POG are applied.

The values of the corrections thus obtained are listed in Tab. A.1. The corresponding uncertainties are computed by repeating the scale factor computation with a tighter selection. Two of the corrections are larger than 1, which is rather unusual. However, they are relative to the event yield obtained after the application of other corrections; therefore, they compensate various effects.

A.2.1 Final kinematic distributions

In Fig. A.17, the event distribution of $\tau_h\tau_h$ events in bins of decay mode pairs are shown after the application of the decay mode dependent scale factors to each real hadronic tau in all the simulated events. A satisfactory data-over-prediction agreement is recovered over all the bins; a good closure is observed in the bins that did not enter the scale factor computation, i.e. those where the two hadronic taus have different decay mode. The final kinematic distributions are shown in Fig. A.18, Fig. A.19, Fig. A.20 and Fig. A.21: the data-over-prediction ratio is rather flat in all the regions that were exhibiting a disagreement, see for example $m_{\tau\tau}^{vis}$ and $\Delta R(\tau_h, \tau_h)$ distributions. On the other hand, a $t\bar{t}$ excess is observed, especially in the resolved 2b0j category; however, the $t\bar{t}$ global normalization is also exceeding in the other channels; the same excess is observed also in the previous analysis, documented in [107].

A.3 Conclusion

Although the investigation on the data-over-prediction disagreement is performed only through the implementation of offline techniques, its outcome points to possible inefficiency not correctly accounted by the method used for the tau trigger and tau identification computation; to recover a good background modelling, alternative corrections with

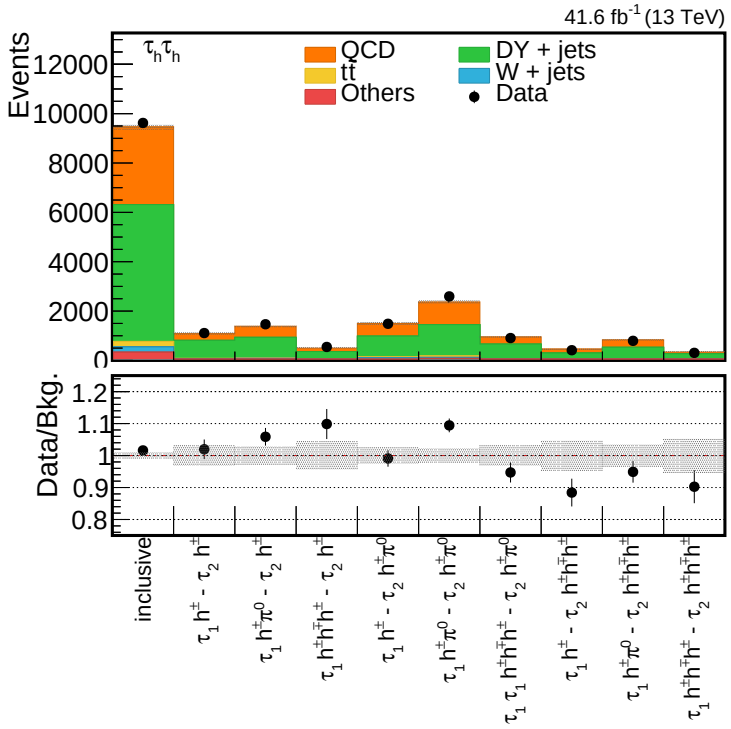


Figure A.16 – Number of data and background events in a Drell-Yan enriched region defined for the computation of corrections alternative to the tau identification $\mathrm{SF}_{\mathrm{tauID}} = 0.89$ proposed by the Tau POG. Events passing the $\tau_h\tau_h$ criteria described in Sec. 4.3 and where the selected tau leptons have separation $\Delta R(\tau_h, \tau_h)$ are selected; they are represented as a function of the hadronic decay mode of the two selected tau leptons. No tau identification scale factors are applied in this plot; all the other recommended corrections are implemented.

Table A.1 – Decay mode dependent corrections $\mathrm{SF}_{\mathrm{DM}}$ alternative to the recommended tau identification scale factor $\mathrm{SF}_{\mathrm{tauID}} = 0.89$, computed in a Drell-Yan enriched region within this analysis selection.

τ_h decay mode	$\mathrm{SF}_{\mathrm{DM}}$
$h^{\pm}$	1.02 ± 0.04
$h^{\pm}\pi^0$	1.09 ± 0.03
$h^{\pm}h^{\mp}h^{\pm}$	$0.93^{+0.05}_{-0.06}$

satisfactory performance are computed within the $HH \to bb\tau\tau$ analysis. Most likely, these corrections cover a few different effects together: some inefficiency coming from the tau trigger or tau identification efficiency computation, or both; and the decay mode dependency not captured by the recommended tau identification scale factor. Indeed, measurements decay mode dependent of the tau identification efficiency, recently delivered by the Tau POG [146], give scale factors smaller than the corrections computed within this analysis; their use cannot fully recover the gap between data and simulation.

The corrections computed as described in this section are applied to $\tau_h\tau_h$ events, whereas a good agreement is achieved using the tau corrections for the τ_h-legs in $\tau_e\tau_h$ and $\tau_\mu\tau_h$ events.

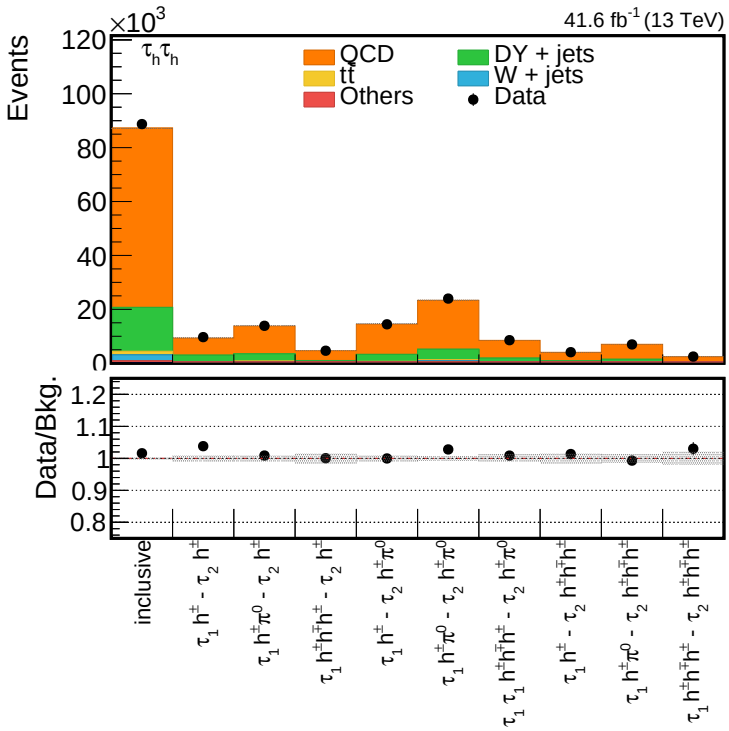


Figure A.17 – Number of data and background events passing the $\tau_h\tau_h$ selection described in Sec. 4.3 as a function of the hadronic decay mode of the two selected tau leptons. The corrections computed within this analysis and listed in Tab. A.1, alternative to the tau identification scale factor, are applied; besides the Tau POG tau identification scale factor, all the other recommended corrections are implemented.

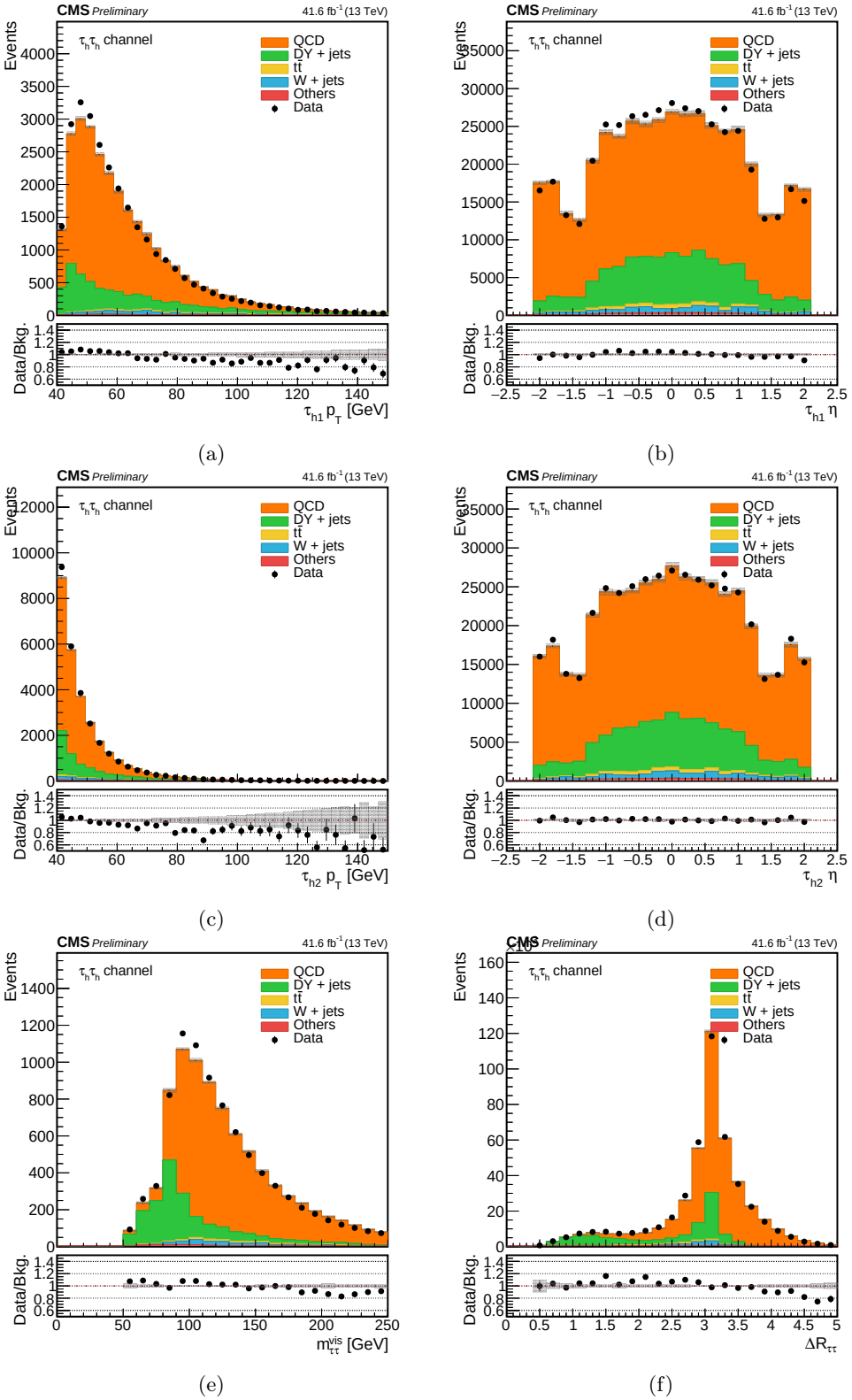


Figure A.18 – Distribution of $\tau_h\tau_h$ data and background events, selected as described in Sec. 4.3, as a function of the main kinematic variables. The corrections computed within this analysis and listed in Tab. A.1, alternative to the tau identification scale factor, are applied; besides the Tau POG tau identification scale factor, all the other recommended corrections are implemented.

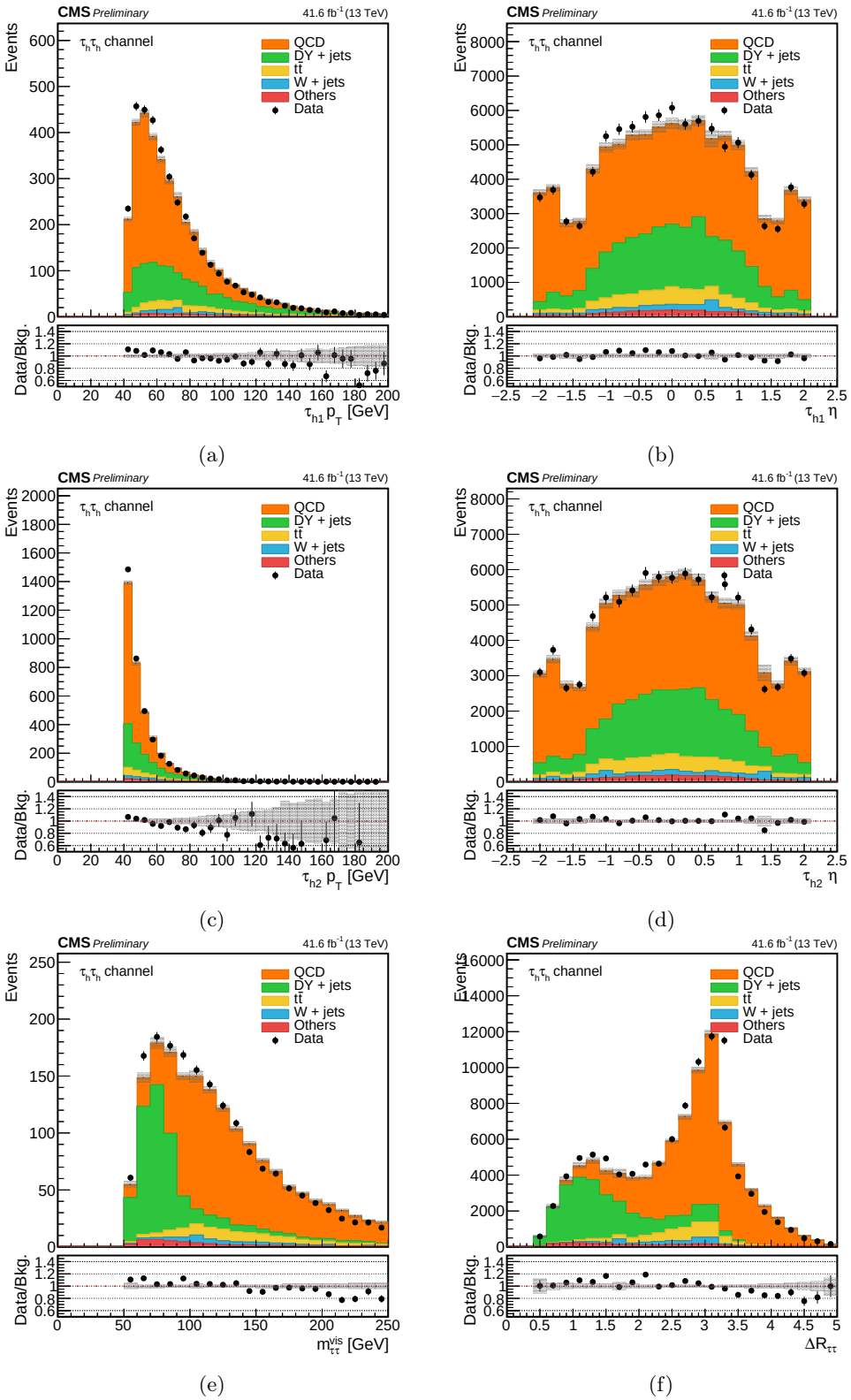


Figure A.19 – Distribution of $\tau_h\tau_h$ data and background events passing the baseline selection of this search, i.e. where two b jet candidates are found.The corrections computed within this analysis and listed in Tab. A.1, alternative to the tau identification scale factor, are applied; besides the Tau POG tau identification scale factor, all the other recommended corrections are implemented.

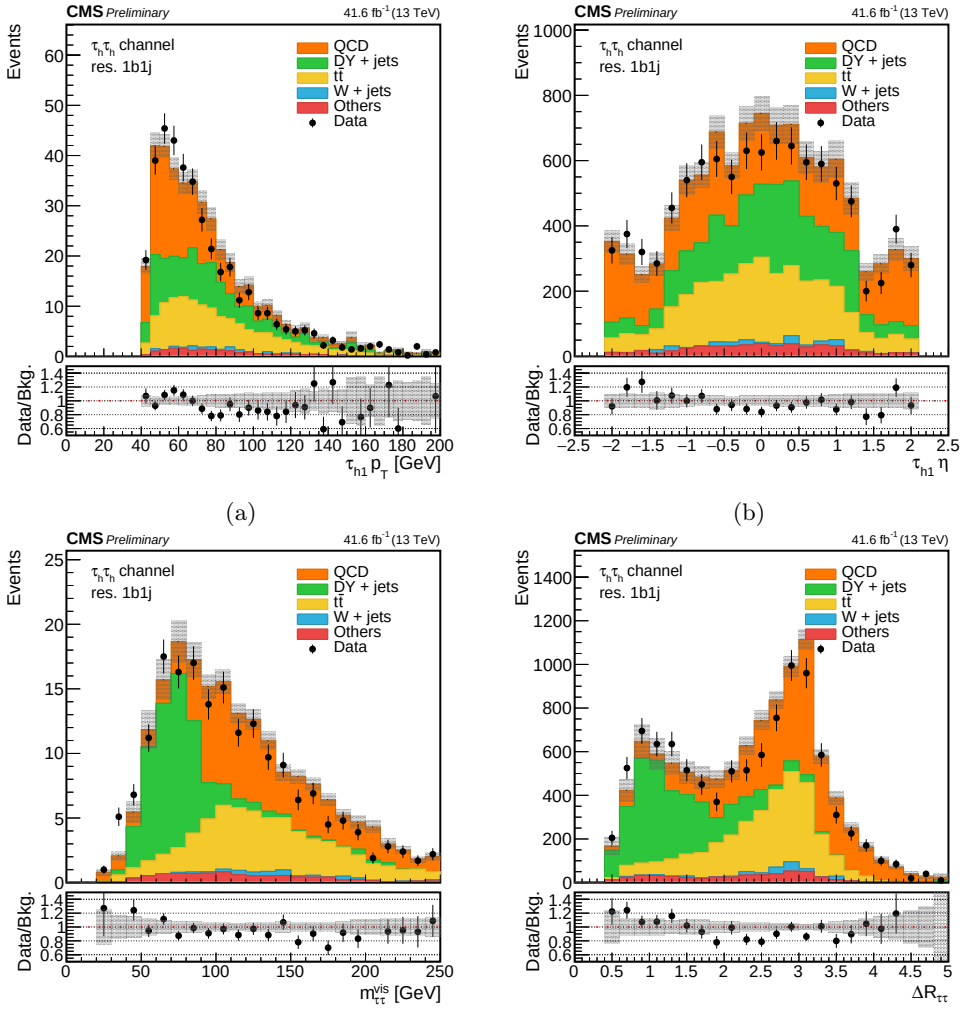


Figure A.20 – Distribution of $\tau_h\tau_h$ data and background events passing the resolved 1b1j requirements. The corrections computed within this analysis and listed in Tab. A.1, alternative to the tau identification scale factor, are applied; besides the Tau POG tau identification scale factor, all the other recommended corrections are implemented.

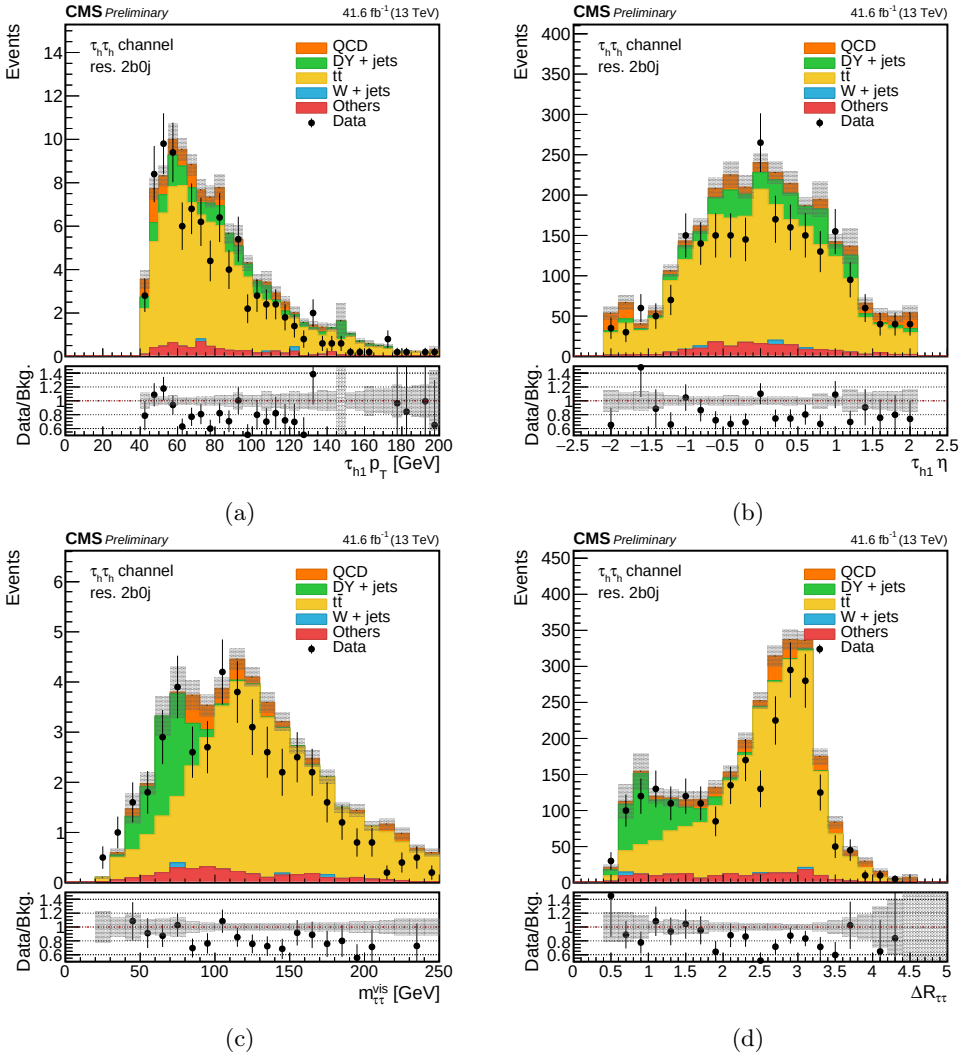


Figure A.21 – Distribution of $\tau_h\tau_h$ data and background events passing the res. 2b0j requirements. The corrections computed within this analysis and listed in Tab. A.1, alternative to the tau identification scale factor, are applied; besides the Tau POG tau identification scale factor, all the other recommended corrections are implemented.

Bibliography

[1] M. J. Herrero, "The Standard Model" arXiv:hep-ph/9812242v1.

[2] H. Georgi, "Lie algebras in particle physics: from isospin to unified theories". Advanced Book Program. Benjamin/Cummings Pub. Co., Advanced Book Program, 1982.

[3] M. D. Schwartz, "Quantum Field Theory and the Standard Model". Cambridge University Press, 2014.

[4] D. Galbraith, "Standard Model of the Standard Model". http://davidgalbraith.org/portfolio/ux-standard-model-of-the-standard-model/. Last visited on April 29, 2020.

[5] P. A. M. Dirac, "Nobel Lecture (1933): Theory of Electrons and Positrons", *Nobel Media* **AB** (2019).

[6] T. Kajita, "Nobel Lecture (2015): Discovery of atmospheric neutrino oscillations", *Rev. Mod. Phys.* **88** (2016), no. 3, 030501. doi:10.1103/RevModPhys.88.030501.

[7] Particle Data Group, M. Tanabashi et al., "Review of Particle Physics", *Phys. Rev. D* **98** (Aug, 2018) 030001. doi:10.1103/PhysRevD.98.030001.

[8] Z. Maki, M. Nakagawa, and S. Sakata, "Remarks on the Unified Model of Elementary Particles", *Progress of Theoretical Physics* **28** (11, 1962) 870–880. doi:10.1143/PTP.28.870.

[9] D. J. Gross and F. Wilczek, "Ultraviolet Behavior of Non-Abelian Gauge Theories", *Phys. Rev. Lett.* **30** (Jun, 1973) 1343. doi:10.1103/PhysRevLett.30.1343.

[10] G. 't Hooft, "Renormalizable Lagrangians for Massive Yang-Mills Fields", *Nucl. Phys.* **B35** (1971) 167–188. doi:10.1016/0550-3213(71)90139-8.

[11] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", *Physics Letters B* **716** (2012), no. 1, 30 – 61. doi:10.1016/j.physletb.2012.08.021.

[12] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Physics Letters B* **716** (2012), no. 1, 1 – 29. doi:10.1016/j.physletb.2012.08.020.

[13] P. W. Higgs, "Broken symmetries and the masses of gauge bosons", *Phys. Rev. Lett.* **13** (Oct, 1964) 508–509. doi:10.1103/PhysRevLett.13.508.

[14] P. Higgs, "Broken symmetries, massless particles and gauge fields", *Physics Letters* **12** (1964), no. 2, 132 – 133. doi:https://doi.org/10.1016/0031-9163(64)91136-9.

[15] F. Englert and R. Brout, "Broken Symmetry and the Mass of Gauge Vector Mesons", *Phys. Rev. Lett.* **13** (Aug, 1964) 321–323. doi:10.1103/PhysRevLett.13.321.

[16] J. Goldstone, "Field theories with "Superconductor" solutions", *Il Nuovo Cimento* **19** (Jan, 1961) 154–164. doi:10.1007/BF02812722.

[17] J. Ellis, "Higgs Physics" KCL-PH-TH-2013-49. doi:10.5170/CERN-2015-004.117.

[18] MuLan Collaboration, "Measurement of the Positive Muon Lifetime and Determination of the Fermi Constant to Part-per-Million Precision", *Phys. Rev. Lett.* **106** (Jan, 2011) 041803. doi:10.1103/PhysRevLett.106.041803.

[19] LHC Higgs Cross Section Working Group, D. de Florian, C. Grojean, F. Maltoni et al., "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector" FERMILAB-FN-1025-T. doi:10.23731/CYRM-2017-002.

[20] ATLAS and CMS Collaborations, "Combined Measurement of the Higgs boson mass in pp collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS experiments", *Phys. Rev. Lett.* **114** (May, 2015) 191803. doi:10.1103/PhysRevLett.114.191803.

[21] CMS Collaboration, "A measurement of the Higgs boson mass in the diphoton decay channel" CMS-PAS-HIG-19-004, Geneva, 2019.

[22] ATLAS Collaboration, "Evidence for the spin-0 nature of the Higgs boson using ATLAS data", *Physics Letters B* **726** (2013), no. 1, 120 – 144. doi:https://doi.org/10.1016/j.physletb.2013.08.026.

[23] CMS Collaboration, "Study of the Mass and Spin-Parity of the Higgs Boson Candidate via its decays to Z boson pairs", *Physical Review Letters* **110** (Feb, 2013). doi:10.1103/physrevlett.110.081803.

[24] CMS Collaboration, "Combined measurements of Higgs boson couplings in proton-proton collisions at $\sqrt{s} = 13$ TeV", *Eur. Phys. J. C* **79** (Sep, 2018) 421, arXiv:1809.10733. doi:10.1140/epjc/s10052-019-6909-y.

[25] CMS Collaboration, "Search for the Higgs boson decaying to two muons in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$", *Physical Review Letters* **122** (jan, 2019). doi:10.1103/physrevlett.122.021801.

[26] S. Dawson, C. Englert, and T. Plehn, "Higgs Physics: It ain't over till it's over", *Physics Reports* **816** (Jul, 2019) 1–85. doi:10.1016/j.physrep.2019.05.001.

[27] F. Maltoni, E. Vryonidou, and M. Zaro, "Top-quark mass effects in double and triple Higgs production in gluon-gluon fusion at NLO", *Journal of High Energy Physics* (2014), no. 11, 79. doi:10.1007/JHEP11(2014)079.

[28] Physics of the HL-LHC Working Group Collaboration, "Higgs Physics at the HL-LHC and HE-LHC" arXiv:1902.00134, Geneva, 2018.

[29] R. Frederix, S. Frixione, V. Hirschi et al., "Higgs pair production at the LHC with NLO and parton-shower effects", *Physics Letters B* **732** (May, 2014) 142–149. doi:10.1016/j.physletb.2014.03.026.

[30] B. D. Micco, M. Gouzevitch, J. Mazzitelli et al., "Higgs boson pair production at colliders: status and perspectives" arXiv:1910.00012, FERMILAB-CONF-19-468-E-T.

[31] A. D. Sakharov, "Violation of CP Invariance, C asymmetry, and baryon asymmetry of the universe", *Pisma Zh. Eksp. Teor. Fiz.* **5** (1967) 32–35. doi:10.1070/PU1991v034n05ABEH002497.

[32] G. Bertone, D. Hooper, and J. Silk, "Particle dark matter: evidence, candidates and constraints", *Physics Reports* **405** (2005), no. 5, 279 – 390. doi:https://doi.org/10.1016/j.physrep.2004.08.031.

[33] A. R. Vieira, B. Hiller, M. C. Nemes et al., "Naturalness and theoretical constraints on the Higgs boson mass", *International Journal of Theoretical Physics* **52** (Jun, 2013) 3494–3503. doi:10.1007/s10773-013-1652-x.

[34] T. Binoth and J. J. van der Bij, "Influence of strongly coupled, hidden scalars on Higgs signals", *Z. Phys.* **C75** (1997) 17–25, arXiv:hep-ph/9608245. doi:10.1007/s002880050442.

[35] R. M. Schabinger and J. D. Wells, "A Minimal spontaneously broken hidden sector and its impact on Higgs boson physics at the large hadron collider", *Phys. Rev.* **D72** (2005) 093007, arXiv:hep-ph/0509209. doi:10.1103/PhysRevD.72.093007.

[36] B. Patt and F. Wilczek, "Higgs-field portal into hidden sectors" arXiv:hep-ph/0605188.

[37] G. C. Branco, P. M. Ferreira, L. Lavoura et al., "Theory and phenomenology of two-Higgs-doublet models", *Phys. Rept.* **516** (2012) 1–102, arXiv:1106.0034. doi:10.1016/j.physrep.2012.02.002.

[38] P. Fayet, "Supergauge Invariant Extension of the Higgs Mechanism and a Model for the electron and Its Neutrino", *Nucl. Phys.* **B90** (1975) 104–124. doi:10.1016/0550-3213(75)90636-7.

[39] P. Fayet, "Spontaneously Broken Supersymmetric Theories of Weak, Electromagnetic and Strong Interactions", *Phys. Lett.* **69B** (1977) 489. doi:10.1016/0370-2693(77)90852-8.

[40] K. Agashe, H. Davoudiasl, G. Perez et al., "Warped Gravitons at the LHC and Beyond", *Phys. Rev.* **D76** (2007) 036006, arXiv:hep-ph/0701186. doi:10.1103/PhysRevD.76.036006.

[41] A. L. Fitzpatrick, J. Kaplan, L. Randall et al., "Searching for the Kaluza-Klein Graviton in Bulk RS Models", *JHEP* **09** (2007) 013, arXiv:hep-ph/0701150. doi:10.1088/1126-6708/2007/09/013.

[42] C. Csáki, M. L. Graesser, and G. D. Kribs, "Radion dynamics and electroweak physics", *Phys. Rev. D* **63** (Feb, 2001) 065002. doi:10.1103/PhysRevD.63.065002.

[43] H. Davoudiasl, J. L. Hewett, and T. G. Rizzo, "Phenomenology of the Randall-Sundrum Gauge Hierarchy Model", *Phys. Rev. Lett.* **84** (Mar, 2000) 2080–2083. doi:10.1103/PhysRevLett.84.2080.

[44] A. Carvalho and M. Dall'Osso and T. Dorigo and F. Goertz and C. A. Gottardo and M. Tosi, "Higgs pair production: choosing benchmarks with cluster analysis", *Journal of High Energy Physics* **2016** (2016), no. 4,. doi:10.1007/jhep04(2016)126.

[45] F. Bishara, R. Contino, and J. Rojo, "Higgs pair production in vector-boson fusion at the LHC and beyond", *Eur. Phys. J.* **C77** (2017), no. 7, 481, arXiv:1611.03860. doi:10.1140/epjc/s10052-017-5037-9.

[46] CMS Collaboration, "Combined Higgs boson production and decay measurements with up to 137 fb^{-1} of proton-proton collision data at $\sqrt{s} = 13\,\text{TeV}$" CMS-PAS-HIG-19-005, Geneva, 2020.

[47] J. Alwall, R. Frederix, S. Frixione et al., "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations", *Journal of High Energy Physics* **2014** (Jul, 2014). doi:10.1007/jhep07(2014)079.

[48] ATLAS Collaboration, "Combination of searches for Higgs boson pairs in pp collisions at $\sqrt{s}$ =13 TeV with the ATLAS detector" CERN-EP-2019-099, 2019.

[49] ATLAS Collaboration, "Search for the $HH \to b\bar{b}b\bar{b}$ process via vector boson fusion production using proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector" ATLAS-CONF-2019-030, Geneva, Jul, 2019.

[50] CMS Collaboration, "Combination of searches for Higgs boson pair production in proton-proton collisions at $\sqrt{s} = 13$ TeV", *Phys. Rev. Lett.* **122** (Nov, 2018) 121803. 18 p, arXiv:1811.09689. doi:10.1103/PhysRevLett.122.121803.

[51] CMS Collaboration, "Higgs PAG Summary Plots". https://twiki.cern.ch/twiki/bin/view/CMSPublic/SummaryResultsHIG. Last visited on April 29, 2020.

[52] L. Evans and P. Bryant, "LHC Machine", *Journal of Instrumentation* **3** (aug, 2008) S08001–S08001. doi:10.1088/1748-0221/3/08/s08001.

[53] HiLumi Collaboration, "The HL-LHC project". https://hilumilhc.web.cern.ch/content/hl-lhc-project. Last visited on April 29, 2020.

[54] E. Mobs, "The CERN accelerator complex - August 2018" OPEN-PHO-ACCEL-2018-005, Aug, 2018.

[55] CMS Collaboration, "Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 13\,\text{TeV}$", *JHEP* **07** (2018) 161, arXiv:1802.02613. doi:10.1007/JHEP07(2018)161.

[56] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider", *JINST* **3** (2008) S08003. 437 p. doi:10.1088/1748-0221/3/08/S08003.

[57] CMS Collaboration, "The CMS experiment at the CERN LHC. The Compact Muon Solenoid experiment", *JINST* **3** (2008) S08004. 361 p. doi:10.1088/1748-0221/3/08/S08004.

[58] ALICE Collaboration, "The ALICE experiment at the CERN LHC. A Large Ion Collider Experiment", *JINST* **3** (2008) S08002. 259 p. doi:10.1088/1748-0221/3/08/S08002.

[59] LHCb Collaboration, "The LHCb Detector at the LHC", *JINST* **3** (2008), no. LHCb-DP-2008-001. CERN-LHCb-DP-2008-001, S08005. doi:10.1088/1748-0221/3/08/S08005.

[60] CMS Collaboration, "Public luminosity results". https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults. Last visited on April 29, 2020.

[61] I. Neutelings, "CMS Wiki Pages, How to draw diagrams in LaTeX with TikZ". https://wiki.physik.uzh.ch/cms/latex:tikz. Last visited on April 29, 2020.

[62] CMS Collaboration, "Cutaway diagrams of CMS detector" CMS-OUTREACH-2019-001.

[63] CMS Collaboration, Technical Report CMS-TDR-1 "The CMS magnet project", CERN-LHCC-97-010, Geneva, 1997.

[64] CMS Collaboration, V. Karimaki et al., technical report "The CMS tracker system project", CERN-LHCC-98-006, Geneva, 1997.

[65] CMS Collaboration, technical report "The CMS tracker: addendum to the Technical Design Report", CERN-LHCC-2000-016, Geneva, 2000.

[66] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", *JINST* **9** (May, 2014) P10009. 80 p, CMS-TRK-11-001. doi:10.1088/1748-0221/9/10/P10009.

[67] CMS Collaboration, "Precision measurement of the structure of the CMS inner tracking system using nuclear interactions with data collected in 2018" CMS-DP-2019-001, Feb, 2019.

[68] CMS Collaboration, Technical Report CMS-TDR-11 "CMS Technical Design Report for the Pixel Detector Upgrade", CERN-LHCC-2012-016, sep, 2012.

[69] CMS Collaboration, "Performance of b tagging algorithms in proton-proton collisions at 13 TeV with Phase 1 CMS detector" CMS-DP-2018-033, jun, 2018.

[70] CMS Collaboration, Technical Report CMS-TDR-014 "The Phase-2 Upgrade of the CMS Tracker", CERN-LHCC-2017-009, CERN, Geneva, Jun, 2017.

[71] CMS Collaboration, Technical Report CMS-TDR-4 "The CMS electromagnetic calorimeter project", CERN-LHCC-97-033, Geneva, 1997.

[72] CMS Collaboration, "Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at $\sqrt{s} = 7$ TeV", *Journal of Instrumentation* **8** (Sep, 2013) P09009–P09009. doi:10.1088/1748-0221/8/09/p09009.

[73] CMS Collaboration, Technical Report CMS-TDR-2 "The CMS hadron calorimeter project", CERN-LHCC-97-031, Geneva, 1997.

[74] CMS Collaboration, "The CMS barrel calorimeter response to particle beams from 2 to 350 GeV/c", *The European Physical Journal C* **60** (Apr, 2009) 359–373. doi:10.1140/epjc/s10052-009-0959-5.

[75] CMS Collaboration, Technical Report CMS-TDR-3 "The CMS muon project", CERN-LHCC-97-032, Geneva, 1997.

[76] CMS Collaboration, "The performance of the CMS muon detector in proton-proton collisions at $\sqrt{s} = 7$ TeV at the LHC", *Journal of Instrumentation* **8** (Nov, 2013) P11002. doi:10.1088/1748-0221/8/11/p11002.

[77] CMS Collaboration, "Measurement of the Muon Stopping Power in Lead Tungstate. ", *JINST* **5** (Dec, 2009). doi:10.1088/1748-0221/5/03/P03007.

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

[79] W. Adam, R. Frűhwirth, A. Strandlie et al., "Reconstruction of Electrons with the Gaussian-Sum Filter in the CMS Tracker at the LHC" CMS-NOTE-2005-001, 2005.

[80] CMS Collaboration, "Performance of electron reconstruction and selection with the CMS detector at $\sqrt{s} = 8$ TeV", *J. Instrum.* **10** (Feb, 2015) P06005. 63 p, arXiv:1502.02701.

[81] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-ktjet clustering algorithm", *Journal of High Energy Physics* **2008** (Apr, 2008) 63. doi:10.1088/1126-6708/2008/04/063.

[82] CMS Collaboration, "Performance of reconstruction and identification of τ leptons decaying to hadrons and ν_τ in pp collisions at $\sqrt{s} = 13$ TeV", *JINST* **13** (2018), no. 10, P10005, arXiv:1809.02816. doi:10.1088/1748-0221/13/10/P10005.

[83] CMS Collaboration, "Summaries of CMS cross section measurements". https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsCombined. Last visited on April 29, 2020.

[84] CMS Collaboration, "The CMS trigger system", *JINST* **12** (2017), no. 01, P01020, arXiv:1609.02366. doi:10.1088/1748-0221/12/01/P01020.

[85] A. Zabi, F. Beaudette, L. Cadamuro et al., "The CMS Level-1 Calorimeter Trigger for the LHC Run II", *Journal of Instrumentation* **12** (Jan, 2017) C01065–C01065. doi:10.1088/1748-0221/12/01/c01065.

[86] CMS Collaboration, Technical Report CMS-TDR-12 "CMS Technical Design Report for the Level-1 Trigger Upgrade", CERN-LHCC-2013-011, Jun, 2013.

[87] CMS Collaboration, "Dimuon Level-1 invariant mass in 2017 data" CMS-DP-2018-002, Feb, 2018.

[88] CMS Collaboration, "The CMS L1 Trigger emulation software", *Journal of Physics: Conference Series* **219** (apr, 2010) 032009. doi:10.1088/1742-6596/219/3/032009.

[89] C. Amendola, CMS Collaboration, "The CMS Level-1 tau lepton and Vector Boson Fusion triggers for the LHC Run II", in *Proceedings, 2017 European Physical Society Conference on High Energy Physics (EPS-HEP 2017): Venice, Italy, July 5-12, 2017*, volume EPS-HEP2017, p. 773. 2017. doi:10.22323/1.314.0773,

[90] CMS Collaboration, "Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector", *Phys. Lett.* **B779** (2018) 283–316, arXiv:1708.00373. doi:10.1016/j.physletb.2018.02.004.

[91] CMS Collaboration, "Peak day-by-day instantaneous luminosity, 2017". https://cmslumi.web.cern.ch/cmslumi/publicplots/peak_lumi_per_day_pp_2017NormtagLumi.png. Last visited on April 29, 2020.

[92] CMS Collaboration, "Peak day-by-day instantaneous luminosity, 2018". https://cmslumi.web.cern.ch/cmslumi/publicplots/peak_lumi_per_day_pp_2018NormtagLumi.png. Last visited on April 29, 2020.

[93] CMS Collaboration, "Selecting VBF Higgs in the Level-1 Trigger at CMS" CMS-DP-2018-005, Feb, 2018.

[94] C.-E. Wulz, G. Aradi, B. Arnold et al., CMS Collaboration, "Data analysis at the CMS level-1 trigger: migrating complex selection algorithms from offline analysis and high-level trigger to the trigger electronics", in *Proceedings, 2017 European Physical Society Conference on High Energy Physics (EPS-HEP 2017): Venice, Italy, July 5-12, 2017*, volume EPS-HEP2017, p. 807. 2017. doi:10.22323/1.314.0807,

[95] CMS Collaboration, "Level 1 Tau trigger performance in 2016 data and VBF seeds at Level 1 trigger" CMS-DP-2017-022, Jul, 2017.

[96] CMS Collaboration, "Measurement of Higgs boson production and decay to the $\tau\tau$ final state" CMS-PAS-HIG-18-032, 2019.

[97] G. Iadarola, G. Rumolo, P. Dijkstal et al., "Analysis of the beam induced heat loads on the LHC arc beam screens during Run 2" CERN-ACC-NOTE-2017-0066, Dec, 2017.

[98] CMS Collaboration, "CMS ECAL Response to Laser Light" CMS-DP-2019-005, Mar, 2019.

[99] R. Brunelière and A. Zabi, "Reconstruction of the signal amplitude of the CMS electromagnetic calorimeter" CMS-NOTE-2006-037, Geneva, Feb, 2006.

[100] CMS Collaboration, "CMS ECAL trigger plots" CMS-DP-2019-031, Sep, 2019.

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

[102] Burkart, Maximilian, MS, Karlsruher Institut of Technologie (KIT). Master's thesis defended on 29th of May 2019. EKP-2019-00022.

[103] ATLAS Collaboration, "Searches for Higgs boson pair production in the $hh \to bb\tau\tau$, $\gamma\gamma WW^*$, $\gamma\gamma bb$, $bbbb$ channels with the ATLAS detector", *Phys. Rev.* **D92** (2015) 092004, arXiv:1509.04670. doi:10.1103/PhysRevD.92.092004.

[104] CMS Collaboration, "Model independent search for Higgs boson pair production in the bb$\tau\tau$ final state" CMS-PAS-HIG-15-013, 2016.

[105] CMS Collaboration, "Search for non-resonant Higgs boson pair production in the b$\bar{\text{b}}\tau^+\tau^-$ final state" CMS-PAS-HIG-16-012, Geneva, 2016.

[106] CMS Collaboration, "Search for resonant Higgs boson pair production in the b$\bar{\text{b}}\tau^+\tau^-$ final state" CMS-PAS-HIG-16-013, Geneva, 2016.

[107] CMS Collaboration, "Search for Higgs boson pair production in events with two bottom quarks and two tau leptons in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$", *Phys. Lett.* **B778** (2018) 101–127, arXiv:1707.02909. doi:10.1016/j.physletb.2018.01.001.

[108] CMS Collaboration, "Combination of searches for Higgs boson pair production in proton-proton collisions at $\sqrt{s} = 13$ TeV" CMS-PAS-HIG-17-030, Geneva, 2018.

[109] A. Giraldi, "Optimisation of a multivariate analysis technique for the $\mathbf{t\bar{t}}$ background rejection in the search for Higgs boson pair production in $\mathbf{b\bar{b}}\tau^+\tau^-$ decay channel with the CMS experiment at the LHC", Università di Pisa. Master's thesis defended on 10th of April 2018. CERN-THESIS-2018-070.

[110] CMS Collaboration, "Tau Identification Performance in 2017 Data at $\sqrt{s} = 13$ TeV" CMS-DP-2018-026, Jun, 2018.

[111] A. Sirunyan, A. Tumasyan, et al., "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV", *Journal of Instrumentation* **13** (May, 2018) P05011–P05011. doi:10.1088/1748-0221/13/05/P05011.

[112] L. Bianchini, J. Conway, E. K. Friis et al., "Reconstruction of the Higgs mass in H to tautau Events by Dynamical Likelihood techniques", *Journal of Physics: Conference Series* **513** (2014), no. 2, 022035.

[113] A. Hocker et al., "TMVA — Toolkit for Multivariate Data Analysis", *PoS* **ACAT** (2007) 040, arXiv:physics/0703039.

[114] R. Brun and F. Rademakers, "ROOT: An object oriented data analysis framework", *Nucl. Instrum. Meth.* **A389** (1997) 81–86. doi:10.1016/S0168-9002(97)00048-X.

[115] CMS Collaboration, "Searches for a heavy scalar boson H decaying to a pair of 125 GeV Higgs bosons hh or for a heavy pseudoscalar boson A decaying to Zh, in the final states with h$\to \tau\tau$", *Physics Letters B* **755** (2016) 217 – 244. doi:https://doi.org/10.1016/j.physletb.2016.01.056.

[116] A. J. Barr, M. J. Dolan, C. Englert et al., "Di-Higgs final states augMT2ed - Selecting *hh* events at the high luminosity LHC", *Physics Letters B* **728** (Jan, 2014) 308–313. doi:10.1016/j.physletb.2013.12.011.

[117] C. G. Lester and B. Nachman, "*Bisection-based asymmetric M_{T2} computation: a higher precision calculator than existing symmetric methods*", *JHEP* **03** (2015) 100, arXiv:1411.4312. doi:10.1007/JHEP03(2015)100.

[118] L. Cadamuro, "Search for Higgs boson pair production in the $b\bar{b}\tau^+\tau^-$ decay channel with the CMS detector at the LHC", Université Paris-Saclay. PhD thesis defended on 5th October 2017. CERN-THESIS-2017-231.

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

[120] S. Kullback and R. A. Leibler, "On Information and Sufficiency", *Ann. Math. Statist.* **22** (1951), no. 1, 79–86. doi:10.1214/aoms/1177729694.

[121] R. Bhattacharya, "Drell-Yan estimation for the HH $\to$ bb$\tau\tau$ analysis (slides)". https://indico.cern.ch/event/818643/contributions/3419662/attachments/1838946/3014569/Rajarshi_22_04_2019.pdf. Last visited on April 29, 2020, CMS restricted.

[122] C. Oleari, "The POWHEG BOX", *Nuclear Physics B - Proceedings Supplements* **205-206** (Aug, 2010) 36–41. doi:10.1016/j.nuclphysbps.2010.08.016.

[123] G. Heinrich, S. P. Jones, M. Kerner et al., "Probing the trilinear Higgs boson coupling in di-Higgs production at NLO QCD including parton shower effects",

Journal of High Energy Physics **2019** (2019), no. 6, 66. doi:10.1007/JHEP06(2019)066.

[124] LHC Higgs Cross Section HH Sub-group, "Latest HH cross section recommendations (TWiki)". https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWGHH. Last visited on April 29, 2020.

[125] R. Frederix and S. Frixione, "Merging meets matching in MC@NLO", *Journal of High Energy Physics* **2012** (Dec, 2012). doi:10.1007/jhep12(2012)061.

[126] G. Cowan, K. Cranmer, E. Gross et al., "Asymptotic formulae for likelihood-based tests of new physics", *The European Physical Journal C* **71** (Feb, 2011) 1554. doi:10.1140/epjc/s10052-011-1554-0.

[127] ATLAS Collaboration, CMS Collaboration, LHC Higgs Combination Group, "Procedure for the LHC Higgs boson search combination in Summer 2011" CMS-NOTE-2011-005, Geneva, Aug, 2011.

[128] CMS Collaboration, "CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV" CMS-PAS-LUM-17-004, Geneva, 2018.

[129] CMS Collaboration, "Tau identification recommendations at $\sqrt{s} = 13$ TeV". https://twiki.cern.ch/twiki/bin/view/CMS/TauIDRecommendation13TeV. Last visited on April 29, 2020.

[130] L.-B. Chen, H. T. Li, H.-S. Shao et al., "Higgs boson pair production via gluon fusion at N^{3}LO in QCD" arXiv:1909.06808.

[131] LHC Higgs Cross Section Working Group, J. R. Andersen et al., "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties" arXiv:1307.1347. doi:10.5170/CERN-2013-004.

[132] CMS Collaboration, "Measurement of differential cross sections for top quark pair production using the lepton + jets final state in proton-proton collisions at 13 TeV", *Phys. Rev. D* **95** (May, 2017) 092001. doi:10.1103/PhysRevD.95.092001.

[133] ATLAS Collaboration, "Search for Resonant and Nonresonant Higgs Boson Pair Production in the $b\bar{b}\tau^+\tau^-$ Decay Channel in pp Collisions at $\sqrt{s} = 13$ TeV with the ATLAS Detector", *Phys. Rev. Lett.* **121** (Nov, 2018) 191801. doi:10.1103/PhysRevLett.121.191801.

[134] L. Breiman, J. Friedman, R. Olshen et al., "Classification and Regression Trees". Chapman & Hall, New York, 1984.

[135] Y. LeCun, B. Boser, J. S. Denker et al., "Backpropagation Applied to Handwritten Zip Code Recognition", *Neural Comput.* **1** (Dec, 1989) 541–551. doi:10.1162/neco.1989.1.4.541.

[136] D. Rainwater, R. Szalapski, and D. Zeppenfeld, "Probing color-singlet exchange in $Z + 2$-jet events at the CERN LHC", *Phys. Rev. D* **54** (Dec, 1996) 6680–6689. doi:10.1103/PhysRevD.54.6680.

[137] ATLAS, "Search for anomalous electroweak production of WW/WZ in association with a high-mass dijet system in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector", *Physical Review D* **95** (Feb, 2017) 032001. doi:10.1103/PhysRevD.95.032001.

[138] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System" arXiv:1603.02754.

[139] "Keras documentation". https://keras.io/. Last visited on April 29, 2020.

[140] CMS Collaboration, Technical Report CMS-TDR-15-02 "Technical Proposal for the Phase-II Upgrade of the CMS Detector", CERN-LHCC-2015-010, Geneva, Jun, 2015.

[141] ATLAS Collaboration, "Measurement prospects of the pair production and self-coupling of the Higgs boson with the ATLAS experiment at the HL-LHC" ATL-PHYS-PUB-2018-053, Geneva, Dec, 2018.

[142] CMS Collaboration, "Prospects for HH measurements at the HL-LHC" CMS-PAS-FTR-18-019, Geneva, 2018.

[143] The DELPHES 3 Collaboration, "DELPHES 3: a modular framework for fast simulation of a generic collider experiment", *Journal of High Energy Physics* **2014** (Feb, 2014) 57. doi:10.1007/JHEP02(2014)057.

[144] CMS Collaboration, "Projected performance of Higgs analyses at the HL-LHC for ECFA 2016" CMS-PAS-FTR-16-002, Geneva, 2017.

[145] A. Gottmann, "H $\to \tau\tau$ control plots". Private communications, 2019.

[146] CMS Collaboration, "Github TauID scale factors repository". https://github.com/cms-tau-pog/TauIDSFs. Last visited on April 29, 2020.

université PARIS-SACLAY | ÉCOLE DOCTORALE PHENIICS

Titre : Déclenchement pour la production par fusion de bosons vecteurs et recherche de production de paires de bosons de Higgs se désintégrant en bb$\tau\tau$ dans CMS auprès du LHC

Mots clés : Boson de Higgs, déclenchement , Vector Boson Fusion, leptons tau

Résumé :
Cette thèse présente une recherche d'événements avec paires de bosons de Higgs (HH) en collisions proton-proton à 13 TeV, fournies par le Large Hadron Collider, au sein de l'expérience CMS (Compact Muon Solenoid) du CERN (Genève). L'étude de la production de paires de bosons de Higgs permet la mesure de la constante d'auto-couplage trilinéaire (λ_{HHH}) ; en plus, la production HH par fusion de bosons vecteurs (*Vector Boson Fusion* ou *VBF*) donne accès à la mesure de la constante de couplage entre deux bosons de Higgs et deux bosons vecteurs ($\lambda_{2\mathrm{V}}$). La valeur de ces deux paramètres est particulièrement sensible à l'existence de physique au-delà du Modèle Standard : même des faibles variations par rapport aux valeurs des couplages prévus par la théorie peuvent induire un changement important de la section efficace. Cependant, la production de HH au LHC est un processus très rare. La production par le mécanisme principal, de fusion de gluon, a une section efficace d'environ 30 fb, suivie par le processus de VBF, lequel est environ 20 fois moins probable. Ainsi, optimiser l'efficacité de la sélection du signal est essentiel. Par conséquent, la première partie du travail de thèse a été dédiée à l'étude d'algorithmes pour le premier niveau du système de déclenchement de niveau 1 (*Level-1* ou *L1 trigger*) de CMS et un algorithme dédié au processus VBF a été mis au point en ciblant des possibles améliorations pour la recherche d'événements HH→bb$\tau\tau$. Il s'agit du premier algorithme VBF pour le système de déclenchement et il a été inclus dans la prise de données à partir de l'été 2017. Les données ainsi sélectionnées sont accessibles pour les recherches du boson de Higgs en cours et celle présentée dans cette thèse. La suite du travail de thèse a été consacrée à l'analyse d'événements HH→bb$\tau\tau$ avec les données collectées en 2017, en commençant par une étude approfondie de l'accord data-simulation et le développement de corrections spécifiques aux leptons taus. En plus de l'étude inclusive des événements de type HH→bb$\tau\tau$, des catégorie d'événements dédiées à la production par VBF ont été introduites et l'algorithme VBF est exploité. Il s'agit de la première mesure dédiée à ce mécanisme de production dans le cadre des analyses de HH→bb$\tau\tau$: la valeur de $\lambda_{2\mathrm{V}}$ est ainsi restreinte par les données observées entre -0.8 et 2.8 fois la prédiction théorique.

Title: Vector Boson Fusion trigger and search for Higgs boson pair production at the LHC in the bb$\tau\tau$ channel with the CMS detector

Keywords: Higgs boson, trigger, Vector Boson Fusion, tau leptons

Abstract:
This thesis describes a search for events with a pair of Higgs bosons (HH) in proton-proton collisions at 13 TeV, provided by the Large Hadron Collider, with the CMS (Compact Muon Solenoid) experiment at CERN (Geneva). The study of the Higgs boson pair production allows its trilinear self-coupling (λ_{HHH}) to be measured; moreover, the HH production through Vector Boson Fusion (VBF) gives access to the measurement of the coupling between two Higgs bosons and two vector bosons ($\lambda_{2\mathrm{V}}$). The values of these parameters are particularly sensitive to the existence of physics beyond the Standard Model: even small variations from the values of the couplings predicted by the theory can lead to a large modification of the cross section. The HH production at the LHC is a very rare process. The production through the main mechanism, by gluon fusion, has a cross section of about 30 fb, followed by the VBF process, which is about 20 times less likely. Therefore, the optimisation of the signal selection is essential. Hence, the first part of the thesis work was devoted to the study of algorithms for the Level-1 (L1) trigger system of CMS and a dedicated algorithm for the VBF process was optimised, targeting possible improvements for the search for HH→bb$\tau\tau$ events. This is the first VBF algorithm for the L1 trigger system and it was included for the data-taking starting as of summer 2017. The events thus selected are available for the ongoing Higgs boson searches, including the search described in this thesis. The rest of the thesis work was dedicated to the analysis of events HH→bb$\tau\tau$ with the data collected in 2017, starting from a comprehensive study of the data-over-simulation agreement and the development of specific corrections for tau leptons. In addition to the inclusive study of the HH→bb$\tau\tau$ events, specific event categories for the VBF production were included. The study presented in this thesis is the first dedicated measurement for this production mechanism in the context of the HH→bb$\tau\tau$ analyses: it lead to the measurement of $\lambda_{2\mathrm{V}}$, constrained by the observed data between -0.8 and 2.8 times the theoretical prediction.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France