Università degli Studi di Milano - Bicocca

DIPARTIMENTO DI FISICA "GIUSEPPE OCCHIALINI"

Scuola di Dottorato in Fisica e Astronomia
Curriculum in Fisica Subnucleare e Tecnologie Fisiche
Ciclo XXXI

Search for the $HH \to b\bar{b}\tau^+\tau^-$ decay with the CMS experiment

Relatore:
Dott.ssa Sandra Malvezzi

Coordinatore:
Dott.ssa Marta Calvi

Candidato:
Francesco Brivio
Matricola 726663

Anno Accademico 2017-2018

Abstract

This thesis describes the search for Higgs boson pairs in the final state composed of two b quarks and two τ leptons. The structure of this dissertation closely follows the workflow of the analysis and the strategies adopted to identify and reconstruct the $b\bar{b}\tau^+\tau^-$ signal candidates. Both the resonant and the non-resonant double Higgs production mechanisms are explored with the statistics collected by the CMS experiment during the 2016 and 2017 data taking periods at a center of mass energy of $\sqrt{s} = 13\ TeV$.

After the discovery of the Higgs boson by the ATLAS and CMS collaborations in 2012, the collective efforts of the high energy physics community have been focused on a precise characterization of this particle. In this context, HH searches play a fundamental role as they represent the favourite channel to measure the Higgs boson trilinear self coupling (λ_{HHH}). Only three parameters, the Higgs boson mass (m_H), the vacuum expectation value and the Higgs trilinear coupling (λ_{HHH}), shape the Higgs field potential in the Standard Model and the last one is the only remaining that has not been directly measured experimentally. Its determination is a crucial point for a proper understanding of the spontaneous electroweak symmetry breaking, which is at the base of the mechanism that gives masses to bosons and fermions. At the same time, any deviation from the theoretical predictions of the Standard Model would lead to sizeable changes in both the kinematics and in the production rate of HH events, thus making double Higgs searches extremely sensitive to New Physics effects.

The $b\bar{b}\tau^+\tau^-$ final state represents one of the most interesting channels to explore double Higgs processes, because of the high branching ratio and the relatively small background contamination. At the same time, however, this final state poses some non trivial experimental challenges such as the reconstruction of the τ lepton decay that involves the presence of undetectable neutrinos, and the discrimination of signal events from background contributions. These challenges prompted the development of specific algorithms and techniques to identify and reconstruct the signal candidates and to maximize the analysis sensitivity.

No excess of events is observed in the analysis of 2016 data, corresponding to an integrated luminosity of 35.9 fb^{-1}, and the results are found to be consistent with the Standard Model background predictions [1]. Exclusion upper limits at 95% Confidence Level are thus set on the product $\sigma_{HH} \times \mathcal{B}(HH \to bb\tau\tau)$.

In the resonant search case, the limits vary from 500 to 5 pb, depending on the mass of the signal resonance hypothesized, while in the non-resonant search, the observed and expected exclusion limits are set to about 30 and 25 times the theoretical Standard Model prediction, respectively. A further interpretation of the non resonant results is given in the context of effective field theories (EFT), in order to explore models that predict the modification

of the Higgs couplings values.

As the double Higgs production rate is very small at the LHC, current HH analyses are mainly limited by the available statistics and are expected to become more and more sensitive with the increase of the integrated luminosity collected. After the success of the 2016 results and in order to fully exploit the statistical power offered by the data, the $bb\tau\tau$ analysis strategy is now set on the combination of the full $2016-2018$ statistics, that amounts to about 160 fb^{-1}.

I was involved in this search since the beginning of Run II and I actively participated in the changes and developments put in place during the 2016 data analysis. This gave me the opportunity to understand and learn the most critical aspects on which to focus our efforts in the future. Hence, each Section of this thesis is complemented with the changes that I introduced in the analysis workflow in 2017 and with ideas on how to further improve the performance and sensitivity on the path to the study of the full Run II statistics and, in a wider perspective, of the High Luminosity phase of the LHC.

Sommario

Questa tesi descrive la ricerca di coppie di bosoni di Higgs nel canale di decadimento in cui uno dei due bosoni decade in una coppia quark-antiquark b e l'altro in una coppia leptone-antileptone tau. L'organizzazione della tesi segue la stessa struttura dell'analisi e delle strategie adottate per l'identificazione e la ricostruzione di eventi di segnale $b\bar{b}\tau^+\tau^-$. Sia l'ipotesi di produzione risonante che quella non risonante di una coppia HH sono considerate e analizzate utilizzando i dati raccolti dall'esperimento CMS nel 2016 e nel 2017, a un'energia nel centro di massa di $\sqrt{s} = 13\ TeV$.

Dopo la scoperta del bosone di Higgs da parte delle collaborazioni ATLAS e CMS nel 2012, lo sforzo collettivo della comunità scientifica di fisica delle alte energie si è concentrato sulla caratterizzazione di questa nuova particella. In tale contesto, le ricerche relative alla produzione di coppie di bosoni di Higgs giocano un ruolo fondamentale, dal momento che permettono di misurare direttamente l'accoppiamento triplo dell'Higgs (λ_{HHH}). Solamente tre parametri, la massa del bosone di Higgs (m_H), il valore di aspettazione del vuoto a l'accoppiamento triplo dell'Higgs, sono necessari per descrivere il potenziale del campo di Higgs all'interno del modello Standard. Di questi tre, λ_{HHH} è l'unico a non essere stato ancora misurato sperimentalmente. La precisa determinazione del suo valore permetterebbe una migliore comprensione della rottura spontanea della simmetria elettrodebole, che sta alla base dell'origine della massa di bosoni e fermioni. Inoltre, qualsiasi deviazione dalla predizione teorica del Modello Standard implicherebbe notevoli cambiamenti nella cinematica e nel *rate* di produzione di eventi con coppie bosoni di Higgs, rendendo le analisi HH estremamente sensibili a effetti di Nuova Fisica.

Il canale di decadimento $b\bar{b}\tau^+\tau^-$ rappresenta uno dei più interessanti stati finali nell'esplorazione di eventi con due bosoni di Higgs, poichè caratterizzato da un *branching fraction* piuttosto elevato e relativamente poco affetto dalla contaminazione dei fondi. D'altro canto, questo canale presenta al contempo alcune problematiche di non semplice soluzione, come ad esempio la ricostruzione del decadimento dei leptoni τ che coinvolge l'emissione di neutrini non rilevabili sperimentalmente a CMS, e la discriminazione degli eventi di segnale dai processi di fondo. Tali difficoltà hanno reso necessario lo sviluppo di specifiche tecniche per l'identificazione e ricostruzione dei candidati di segnale, e per l'ottimizzazione della sensitività dell'analisi stessa.

Nessun eccesso di eventi è osservato nell'analisi di 35.9 fb^{-1} raccolti nel 2016 e i risultati sono consistenti con l'ipotesi di solo fondo predetta dal Modello Standard [1]. Limiti superiori di esclusione al 95% di livello di confidenza sono dunque calcolati relativamente alla quantità $\sigma_{HH} \times \mathcal{B}(HH \to bb\tau\tau)$. Nel caso dell'analisi risonante, tali limiti variano da 500 a 5 pb in funzione dell'ipotesi di massa della risonanza testata, mentre nel caso non risonante i limiti attesi ed osservati corrispondono a 25 e 30 volte, rispettivamente, la

predizione del Modello Standard. Un'ulteriore interpretazione dei limiti non risonanti è data nel contesto di teorie effettive (EFT), al fine di esplorare modelli in cui gli accoppiamenti del bosone di Higgs sono differenti da quelli previsti dal Modello Standard.

Data la ridotta sezione d'urto di produzione di coppie di bosoni di Higgs, le analisi HH sono attualmente limitate dalla statistica disponibile. Per questo motivo, dopo il promettente risultato ottenuto dallo studio dei dati 2016, l'attenzione delle analisi HH a CMS è rivolta all'inclusione nelle ricerche della totalità dei dati raccolti durante il Run II e corrispondenti a circa 160 fb^{-1}.

L'aver lavorato a questa analisi sin dal suo inizio mi ha permesso di partecipare attivamente allo sviluppo di nuove strategie e di capire quali sono gli aspetti più critici dell'analisi stessa sui quali è necessario concentrare gli sforzi per massimizzare la sensitività. In ogni Sezione di questa tesi quindi, affiancherò alla descrizione delle tecniche utilizzate nel 2016, i cambiamenti e i miglioramenti apportati durante il mio studio dei dati 2017 in vista dei risultati futuri, prima per il Run II e, in una prospettiva più ampia, per la fase ad alta luminosità di LHC.

Contents

Chapter 1

Double Higgs Production

The Standard Model of particle physics (SM) is a renormalizable quantum field theory of fundamental interactions that describes phenomena at the subnuclear scale. Over the last decades it has been tested at many collider experiments and the model has proven to be an accurate description of particle physics, up to the TeV scale, by offering an unified vision of the strong, weak, and electromagnetic forces. After the discoveries of the W and Z bosons (1983), and of the top quark (1995), the observation of the Higgs boson by the ATLAS and CMS experiments at the LHC [2, 3] is one of the most recent results corroborating the predictions of the SM.

Despite the excellent agreement with experimental results, some observations, ranging from subnuclear to astrophysical scale, are in contrast with the theoretical predictions and suggest the presence of physics beyond the SM (BSM). The exploration of the scalar sector, started with the discovery of the Higgs boson, provides a new exciting way to study possible hints of new physics at the TeV scale and beyond.

Being intimately related to the nature of the scalar sector, the study of double Higgs production and the measurement of its cross section offer a unique opportunity to explore the structure of the Higgs field potential through the determination of the Higgs boson self interaction and to explore the electroweak symmetry breaking mechanism.

In order to properly understand the mechanisms related to double Higgs production and their study, in Sections 1.1 and 1.2 I will introduce the principles on which the Standard Model is based and the spontaneous symmetry breaking concept that is the foundation of the Higgs boson theory.

1.1 The Standard Model of particle physics

The Standard Model [4–6] is a renormalizable quantum field theory based on the local gauge invariance of its Lagrangian under the gauge group $SU(3)_C \times SU(2)_L \times SU(1)_Y$ that explains strong, weak and electromagnetic interactions. Invariance under $SU(3)_C$ results in the presence of gluons (g), mediators of the strong force and described by quantum chromodynamics (QCD); the weak and electromagnetic forces are explained by the $SU(2)_L \times SU(1)_Y$ symmetry and they are mediated by the spin-1 $W^{\pm}$ and Z bosons, and by the spin-0 photon (γ). Matter is described in the SM by spin-$\frac{1}{2}$ fermion fields and experimental observations show the existence of twelve of these fields: six "quarks" and six "leptons". To each fermion corresponds an antiparticle with identical properties but opposite quantum numbers.

Quarks posses a "colour" charge, are subject to all three forces and they are divided in three families: the first is composed of up (u) and down (d) quarks with a mass of few MeV, while charm (c) and strange (s) quarks compose the second family, with masses of about 1.28 GeV and 95 MeV respectively. Finally the third family is made of the top (t) and bottom (b) quarks with masses of 173 and 4.2 GeV, respectively. u, c and t quarks are collectively known as the "up" family and have a positive electric charge of $+\frac{2}{3}$, while the remaining three compose the "down" family with a negative electric charge of $-\frac{1}{3}$. Due to QCD confinement properties, quarks can not appear as free states, but are always bounded in quark-antiquark pairs ("mesons") or triples ("barions") that are collectively denoted "hadrons". The process that creates an hadron from a single quarks takes the name "hadronization" and it happens on timescales of the order of 10^{-24} s [7].

Leptons, that have no colour charge and are subject only to the electromagnetic and weak forces, are also divided in three families, each one containing a charged and a neutral particle. The charged lepton of the first family is the electron (e), which is a stable particle with a mass of 511 keV, while the electonic neutrino (ν_e) represents its neutral counterpart. The second leptonic family is composed by the muon (μ) and the muonic neutrino (ν_μ). The muon has a mass of 105.7 MeV and a lifetime of 2.2 μs, long enough for the particle to cross all the subdetectors of the LHC experiments, and can thus be considered stable. From the experimental point of view, the tau (τ) lepton and the tauonic neutrino (ν_τ) compose the third leptonic family. The tau lepton has a mass of 1.8 GeV and a life time too short (2.9×10^{-13} s) to be detected directly by any of the LHC experiments: only its decay products can be reconstructed. Since neutrinos have no "colour" or electric charge, they can only interact with matter through the weak force, making their detection at collider experiments almost impossible. Their only tangible experimental signature is the imbalance of the total transverse momentum vector sum, often referred to as missing transverse energy, or "MET". The

observation of neutrino flavour oscillations prove that their mass in not zero, nonetheless no experiment has been able to directly measure it yet.

QCD is based on the local gauge invariance under the $SU(3)_C$ group, where the Lagrangian density of a massless spin-$\frac{1}{2}$ fermion is:

$$\mathcal{L} = \bar{\psi}(x)\left(i\gamma^\mu\partial_\mu\right)\psi(x) \tag{1.1}$$

where ψ is the fermion field and γ^μ are the Dirac matrices. The product $\gamma^\mu\partial_\mu$ is also written as $\not{\partial}$. The fermion fields transform under the $SU(3)_C$ group as:

$$\psi(x) \rightarrow e^{ig\frac{\lambda^a}{2}\theta_a(x)}\psi(x) \tag{1.2}$$

where $\frac{\lambda^a}{2}$ are the Gell-Mann matrices that generate the group. Nonetheless, the derivatives $\partial_\mu\psi(x)$ do not transform in the same way and need to be re-defined as *covariant derivatives* in order to maintain the Lagrangian density invariance under the transformation in Equation 1.2:

$$D_\mu = \partial_\mu - igA^a_\mu(x)\frac{\lambda^a}{2} \tag{1.3}$$

where the gauge vector fields $A^a_\mu(x)$ correspond to the eight gluons that are the mediators of the strong force. The introduction of the vector fields ensures the invariance, under the local gauge transformation, of the Lagrangian that can be completed with a kinetic term for the gluon fields. The complete QCD Lagrangian density thus becomes:

$$\mathcal{L}_QCD = \bar{\psi}\left(i\gamma^\mu\partial_\mu\right)\psi - g\bar{\psi}\gamma^\mu\frac{\lambda_a}{2}\psi A^a_\mu - \frac{1}{4}F^{\mu\nu}_a F^a_{\mu\nu} \tag{1.4}$$

with a summation over all quark fields. The first term represents the free-field quark propagation, while the second originates from the introduction of the covariant derivatives and describes the interactions of quarks with gluons. The strength of this interaction is denoted by g in Equation 1.4, but is commonly referred to as the strong coupling constant $\alpha_s = g^2/4\pi$. The third term is instead the kinetic term of the vector field. The introduction of gauge bosons (gluons) and the description of their interaction with the fermion fields (quarks) is a direct consequence of requiring the invariance of the Lagrangian under a local gauge transformation. Finally, any explicit mass term for the gauge bosons in the form $A^a_\mu A^\mu_a$ would break the Lagrangian invariance and this poses a question that will be addressed in Section 1.2.

The same local invariance mechanism under the $SU(2)_L \times U(1)_Y$ group is used to explain the electroweak interaction. The $SU(2)_L$ gauge group is associated to the weak isospin quantum number (I_3) and results in the

presence of 3 gauge fields W_μ^i, while the $U(1)_Y$ group is linked to the weak hypercharge Y and by imposing the gauge invariance a single field arises, denoted as B_μ. The Gell-Mann-Nishijima formula exhibits the relationship between these quantum numbers and the electric charge:

$$Q = I_3 + \frac{Y}{2} \tag{1.5}$$

Experimental results show that parity is violated by weak interaction, thus fermionic fields are expressed with right or left chirality. Fermion fields can be therefore represented as left chirality doublets and right chirality singlets:

$$\Psi_L = \frac{1-\gamma^5}{2}\begin{pmatrix}\psi\\ \psi'\end{pmatrix} = \begin{pmatrix}\psi_L\\ \psi'_L\end{pmatrix} \psi_R = \frac{1+\gamma^5}{2}\psi\psi'_R = \frac{1+\gamma^5}{2}\psi' \tag{1.6}$$

where $\frac{1-\gamma^5}{2}$ and $\frac{1+\gamma^5}{2}$ are respectively the left and right projection operators. The fields ψ and ψ' represent either the neutrino and lepton or the up- and down-type quarks.

With this notation, the Lagrangian density can be written as

$$\mathcal{L} = i\bar{\Psi}_L \not{D} \Psi_L + i\bar{\psi}_R \not{D} \psi_R + i\bar{\psi}'_R \not{D} \psi'_R \tag{1.7}$$

where the covariant derivative is

$$D_\mu = \partial_\mu - igW_\mu^i T_i - ig\frac{Y}{2}B^\mu \tag{1.8}$$

Upon substitution of the derivative in Equation 1.7, a charged current interaction appears that couples the ψ_L and ψ'_L fields and is mediated by the $W^\pm$ bosons defined as

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2) \tag{1.9}$$

Since both W_μ^3 and B_μ couple to neutral fields and neither of them can therefore be interpreted as the photon field, a linear superposition can be used to express them as combination of the physical Z_μ (the neutral Z boson field) and the A_μ (the photon field):

$$B_\mu = A_\mu cos\theta_W - Z_\mu sin\theta_W \qquad W_\mu^3 = A_\mu sin\theta_W + Z_\mu cos\theta_W \tag{1.10}$$

The final electroweak Lagrangian density can be expressed in a compact form

as

$$\mathcal{L} = i\bar{\Psi}_L \not{D} \Psi_L + i\bar{\psi}_R \not{D} \psi_R + i\bar{\psi}'_R \not{D} \psi'_R - \frac{1}{4} B^{\mu\nu} B_{\mu\nu} - \frac{1}{4} W_i^{\mu\nu} W^i_{\mu\nu} \qquad (1.11)$$

and it contains the free fermion Dirac Lagrangian as well as charged and neutral currents. As it was the case with the strong interaction Lagrangian in Equation 1.4, any explicit mass term of the gauge fields would break the invariance, while mass terms for the fermions are also not allowed due to the left and right chiralities of the fields that would generate a mass term $m\psi\bar{\psi} = m(\bar{\psi}_R\psi_L + \bar{\psi}_L\psi_R)$ that would incorrectly mix singlets and doublets.

An overview of the SM particles and their quantum numbers is given in Table 1.1. Fermions and bosons are shown in their $SU(2)_L$ representation with the corresponding spin, hypercharge Y and the electromagnetic charge Q.

	field			spin	Y	Q
Leptons	$\begin{pmatrix} e \\ \nu_e \end{pmatrix}_L$	$\begin{pmatrix} \mu \\ \nu_\mu \end{pmatrix}_L$	$\begin{pmatrix} \tau \\ \nu_\tau \end{pmatrix}_L$	1/2	1/3	$\begin{pmatrix} -1 \\ 0 \end{pmatrix}$
	e_R	μ_R	τ_R	1/2	4/3	1
Quarks	$\begin{pmatrix} u \\ d \end{pmatrix}_L$	$\begin{pmatrix} c \\ s \end{pmatrix}_L$	$\begin{pmatrix} t \\ b \end{pmatrix}_L$	1/2	1/3	$\begin{pmatrix} -2/3 \\ 1/3 \end{pmatrix}$
	u_R	c_R	t_R	1/2	4/3	2/3
	d_R	s_R	b_R	1/2	-2/3	-1/3
Gauge bosons		W^i_μ		1	0	0, ±1
		B_μ		1	0	0
		G^a_μ		1	0	0

Table 1.1: Summary of the Standard Model particles and their quantum numbers.

1.2 The Brout-Englert-Higgs mechanism

Since any explicit mass term in the Standard Model Lagrangian described so far would violate the gauge invariance, the theory also bonds all fermions and gauge bosons to have zero mass, in clear contrast with the experimental observation of massive weak bosons and fermions.

The solution, proposed in 1964 independently by physicists Englert and Brout [8], and Higgs [9], is known as the Brout-Englert-Higgs (BEH) mechanism and it is based on the concept of spontaneous symmetry breaking.

If a system, whose Lagrangian $\mathcal{L}$ possesses a particular symmetry, is considered, two situations can occur when considering a particular energy level: either the energy level is non-degenerate and gives rise to a unique eigenstate, or it is degenerate and the corresponding eigenstates are not invariant but transform linearly amongst themselves under the symmetry transformations of $\mathcal{L}$. In field theory, the state of the lowest energy is the vacuum, when one of the degenerate states is arbitrarily selected as the ground states, then it will no longer share the symmetries of $\mathcal{L}$: the asymmetry thus obtained is not due to adding a non-invariant asymmetric term to $\mathcal{L}$, but to the arbitrary choice of one out of the continuum of possible ground states.

In the BEH mechanism, the simplest way to break the symmetry is to introduce a complex scalar doublet, invariant under translation and Lorentz transformations

$$\Phi = \begin{pmatrix} \phi_a \\ \phi_b \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} \tag{1.12}$$

where ϕ_i represent four scalar fields that contribute to the field Lagrangian

$$\mathcal{L}_{BEH} = (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(\Phi^\dagger \Phi) \tag{1.13}$$

and the potential $V(\Phi^\dagger \Phi)$ is defined as:

$$V(\Phi^\dagger \Phi) = -\mu^2 \Phi^\dagger \Phi + \lambda (\Phi^\dagger \Phi)^2 \qquad \text{with } \lambda > 0 \tag{1.14}$$

If $\mu^2 < 0$ the potential assumes the shape shown in Figure 1.1 and the field acquires a non-null vacuum expectation value (VEV)

$$v = \sqrt{\frac{\mu^2}{\lambda}} \tag{1.15}$$

Since any choice of a specific ground state is related to the others by a global phase transformation, to break the symmetry a particular value Φ_0 of the field can be chosen without loosing generality. It can be shown that for a specific choice, out of the four scalar fields only one massive physical field remains, the Higgs field, whose quanta correspond to a new physical massive particle, the Higgs boson. The three remaining mass-less degrees of freedom are Goldstone bosons and can be considered as the longitudinal polarizations of the gauge bosons Z and W, which, in turn, acquire mass.

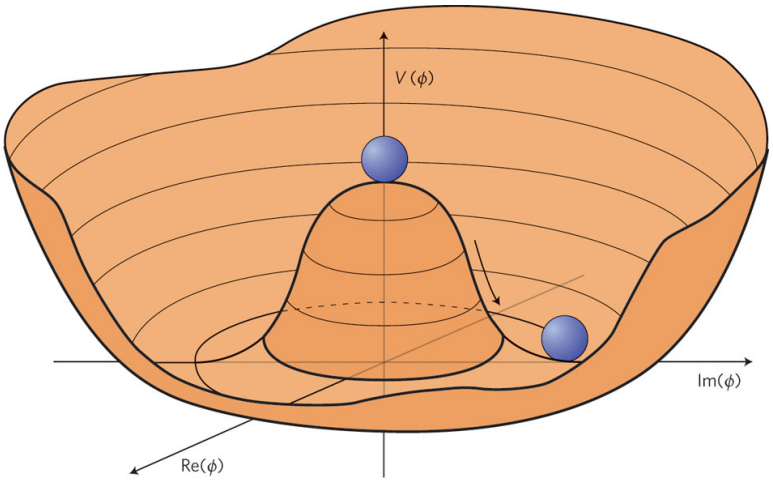


Figure 1.1: Schematic representation of the Higgs field potential in case of $\mu^2 < 0$. Despite being symmetric with respect to zero, a non-zero value has to be assumed in order to minimize $V(\Phi)$ [10].

The resulting Lagrangian for the BEH mechanism is

$$\begin{aligned}\mathcal{L}_{BEH} = &\frac{1}{2}\partial_\mu H \partial^\mu H - \frac{1}{2}\left(2\lambda v^2\right) H^2 \\ &+ \left[\left(\frac{gv}{2}\right)^2 W_\mu^+ W^{\mu-} + \frac{1}{2}\frac{(g^2+g'^2)^2 v^2}{4} Z_\mu Z^\mu\right]\left(1+\frac{H}{v}\right)^2 \\ &+ \lambda v H^3 + \frac{\lambda}{4}H^4 - \frac{\lambda}{4}v^4\end{aligned} \tag{1.16}$$

Where g and g' are the coupling constants and are usually expressed in terms of the Weinberg mixing angle as $\sin(\theta_W) = g'/\sqrt{g^2+g'^2}$ and $\cos(\theta_W) = g/\sqrt{g^2+g'^2}$. The first line in Equation 1.16 describes the evolution of the Higgs field and the associated boson that has a mass $m_H^2 = 2\lambda v^2 = 2\mu^2$, while the second line represents not only the mass terms of the weak interaction bosons

$$\begin{aligned} m_W^2 &= \frac{g^2 v^2}{4} \\ m_Z^2 &= \frac{(g^2+g'^2)v^2}{4} = \frac{m_W^2}{cos^2(\theta_W)} \end{aligned} \tag{1.17}$$

but also the interaction of the weak bosons with one or two Higgs bosons: HWW, HZZ, $HHWW$ and $HHZZ$. Finally the third line in equation 1.16 predicts the cubic and quartic self-interactions of the Higgs boson. The BEH

potential can thus be written as

$$V(H) = \frac{1}{2}m_H^2 H^2 + \lambda v H^3 + \frac{1}{4}\lambda H^4 - \frac{\lambda}{4}v^4 \tag{1.18}$$

At this point only two free parameters, directly related to the scalar potential, are present in the BEH mechanism: the VEV v and the mass of the Higgs boson m_H. Given the fact that the masses of the Z and W bosons are experimentally well know and the VEV can be extracted from the measurement of the Fermi constant G_F

$$\frac{G_F}{\sqrt{2}} = \left(\frac{g}{2\sqrt{2}}\right)^2 \frac{1}{m_W^2} \quad \Rightarrow \quad v = \sqrt{\frac{1}{\sqrt{2}G_F}} \approx 246\ GeV \tag{1.19}$$

the Higgs boson self coupling, responsible for the mass of the boson itself, is the only missing piece to fully understand the scalar sector of the Standard Model.

In the fermion sector, masses are generated by the interaction with the Higgs field through a Yukawa interaction that couples the left and right chiral fields. The Yukawa Lagrangian is Lorentz and gauge invariant and can therefore be included in the Standard Model Lagrangian; after the electroweak symmetry breaking, the Yukawa term can be written as

$$\mathcal{L} = -\sum_f m_f \left(\psi_L \psi_R + \psi_R \psi_L\right)\left(1 + \frac{H}{v}\right) \tag{1.20}$$

where the sum runs on both up- and down-type fermions, and the masses (m_f) and couplings to the Higgs boson (y_f) are connected by the relation

$$m_f = y_f \frac{v}{\sqrt{2}} \tag{1.21}$$

The strenghts of the interactions, directly proportional to the mass of the fermions themselves, are free parameters of the theory, which however does not explain neither their origin nor the hierarchy of the fermion families.

In conclusion, upon spontaneously breaking the electroweak symmetry, the scalar field generates Goldstone bosons that are absorbed as degrees of freedom by the vector bosons fields, which in turn become massive. Moreover, the Higgs field couples the right and left chiral components of the fermion fields in a Yukawa interaction that introduces the fermion masses without breaking the gauge invariance.

1.3 Double Higgs production

A very powerful way to investigate the Higgs sector is the study of the double Higgs production, predicted both by the Standard Model and by many BSM scenarios; it allows for the determination of the Higgs boson self-interaction, and provides a fertile ground to search for hints of BSM physics. Despite being an extremely rare process, recent results on searches for double Higgs production (detailed in Section 1.4), show a promising view for future studies.

It has been known since a long time that the trilinear Higgs self coupling (λ_{HHH}) can be extracted from the measurement of Higgs boson pair production cross section.

However, in a proton-proton collider, there are five main mechanisms through which an HH pair can be produced: gluon fusion, vector boson fusion, vector or top quark pair associated production and single top quark associated production.. Beside the Higgs self coupling, each of these processes involves different interactions of Higgs boson with other particles: their effect, thus, must be properly taken into account in order to obtain a valid measure of the λ_{HHH} parameter. Table 1.2 summarizes the cross sections of these processes at a center of mass energy of 13 TeV, as it was in LHC during the data taking period $2015-2018$, while Figure 1.2 shows a graphical comparison of the cross sections as function of the center of mass energy.

Production mode	$\sigma[fb]$ for $\sqrt{s} = 13\ TeV$
Gluon fusion	$33.49^{+4.3\%}_{-6.0\%}$(scale)$\pm$2.1%(PDF)$\pm$2.3%(α_s)$\pm$5.0%(top)
VBF	$1.62^{+2.3\%}_{-2.7\%}$ (scale) $\pm$2.3% (PDF + α_s)
$t\bar{t}$HH	$0.772^{+1.7\%}_{-4.5\%}$ (scale) $\pm$3.2% (PDF + α_s)
W^+HH	$0.329^{+0.32\%}_{-0.41\%}$ (scale) $\pm$2.2% (PDF + α_s)
W^-HH	$0.173^{+1.2\%}_{-1.3\%}$ (scale) $\pm$2.8% (PDF + α_s)
ZHH	$0.362^{+3.4\%}_{-2.6\%}$ (scale) $\pm$1.9% (PDF + α_s)
tjHH	$0.0281^{+5.2\%}_{-3.2\%}$ (scale) $\pm$4.5% (PDF + α_s)

Table 1.2: Cross section for different HH production mechanisms assuming a Higgs boson mass of 125.09 GeV and a $\sqrt{s} = 13\ TeV$ [11].

Given that HH processes are in general very rare at the LHC, only the first two production channels, that have the highest cross section values, are currently investigated:

- The **Gluon fusion production ($gg \to HH$)** involves the production of Higgs pairs either through the trilinear self coupling, or through the radiation of two Higgs bosons from a heavy quark loop (Figure 1.3).

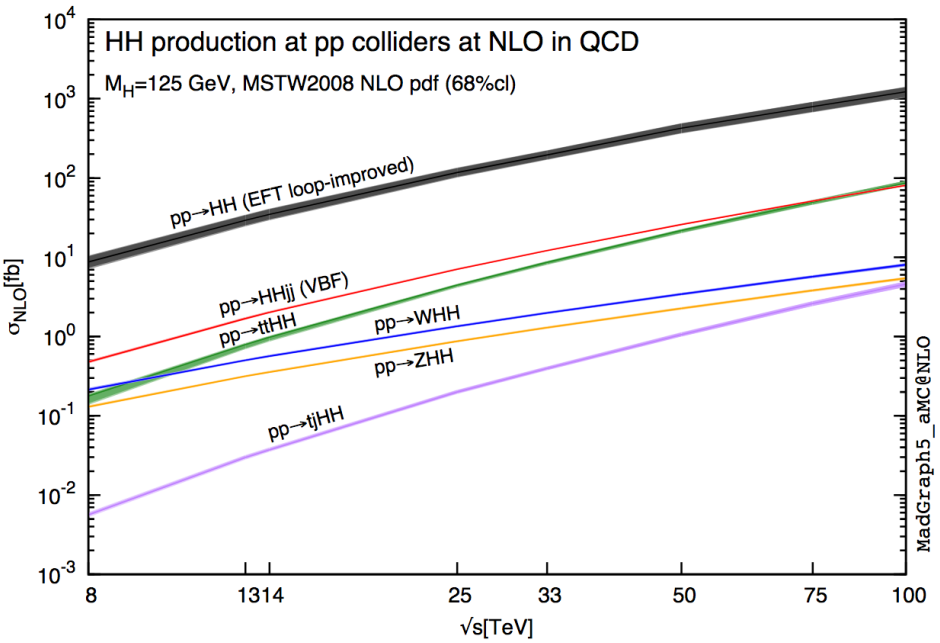


Figure 1.2: Total cross section for HH production in proton-proton collisions for the production modes described in Section 1.3. The cross sections are computed at the NLO accuracy and the bands shown the linear combination of the theoretical errors on the scale and PDF uncertainties. The plot is taken from [12].

The production cross section consequently depends on λ_{HHH} and y_t, *i.e.* the Higgs coupling to the top quark: since contribution from the b quark are less than 1% and can be temporarily neglected. The two diagrams contributing to the gluon fusion production have amplitudes of nearly the same magnitude, which interfere destructively and result into the small cross section reported in Table 1.2.

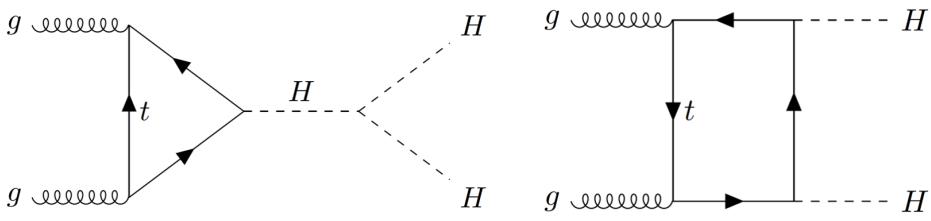


Figure 1.3: Feynman diagrams of the production of double Higgs pairs for the gluon fusion mechanism.

- The **Vector boson fusion (VBF) production** ($qq' \to jjHH$) depends on the trilinear Higgs self-coupling and on the interaction of vector boson pair with either a single Higgs boson or a HH pair (Figure 1.4). The production cross section is about 20 times smaller than the gluon fusion one, but the presence of two jets in the final state provides a very peculiar signature that can be exploited to discriminate signal from background events.

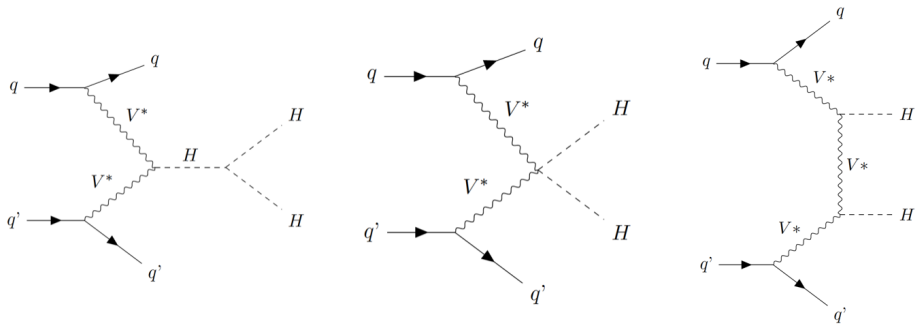


Figure 1.4: Feynman diagrams of the production of double Higgs pairs for the vector boson fusion mechanism.

1.3.1 HH Beyond the Standard Model

Both theoretical considerations and experimental results indicate that the Standard Model is not a complete theory. For example it does not provide an explanation for the large matter/anti-matter asymmetry in the universe, nor it does include possible candidates for dark matter, and it does not even provide a description of gravity. In addition, the SM does not predict the existence of exactly three fermion families as we observe in nature and does not provide an explanation of the couplings values that span many orders of magnitude.

In this context it is conceivable to assume that the Standard Model is just a manifestation of a more comprehensive theory able to explain all these phenomena and at the same time preserve the undeniable success of the SM in describing the phenomenology observed at collider experiment so far.

Being strictly linked to the SM scalar sector, double Higgs production studies offer a probe to test different BSM scenarios and discriminate between possible alternatives. Moreover, exactly because of the destructive interference of the two gluon fusion Feynman diagrams that results in a small production cross section (as discussed in Section 1.3), the HH channels become extremely sensitive to any physics beyond the Standard Model (BSM).

Variations of the Higgs couplings or the presence of heavy resonances might alter the destructive interference and enhance the double Higgs production rate, thus revealing the presence of new phenomena, outside the description of the SM.

If new resonances exist at the TeV scale, they can be produced at the LHC and directly observed by experiments such as ATLAS and CMS; on the other hand, even if BSM effects might be related to energies much higher than those reachable at LHC, their presence can be inferred from non-resonant enhancement of the cross sections, due to anomalous Higgs coupling values or new particles entering the quantum loops. These possible different scenarios are described in the next paragraphs.

Resonant BSM HH production

Higgs boson pairs can be produced, in the context of BSM physics, from a new resonance X with mass $m_X > 2m_H$ that couples significantly to the Higgs boson. This generates, in the invariant mass distribution, a peak at m_X that represent a signature common to this kind of processes and allows for model-independent searches that can subsequently be interpreted in more specific BSM models.

Resonant HH signatures are most commonly known to appear in models predicting either an extended scalar sector with respect to the SM, or the presence of warped extra dimensions that might alter the relations between the Higgs boson and the matter fields. Some representative examples are here described in order to show how they can be simultaneously probed in HH studies even though they move from different theoretical assumptions.

The Higgs Singlet Model [13] represents the simplest extension of the scalar sector as it contemplates the existence of an Higgs singlet in addition to the Standard Model Higgs doublet. After the electroweak symmetry breaking, the Higgs Single Model implies the presence of two physical fields that correspond to an heavy and a light scalar boson, commonly denoted as H and h, respectively. The lightest of the two is interpreted as the SM Higgs boson. Both the Higgs trilinear self-coupling (hhh) and a new interaction, Hhh, are predicted in the model, where, especially for the latter, the branching fraction is sizeable for resonances of mass up to the TeV scale, as shown in Figure 1.5.

A slightly more complex approach to extend the scalar sector is represented by the class of Two-Higgs-doublet models [14] (2HDM) that postulate the existence of an additional Higgs doublet. In their most general form, 2HDMs have a very complex phenomenological structure; the focus will be here restricted only to the minimal supersymmetric extension of the Standard Model (MSSM) which represents a specific case of 2HDM. In 2HDMs,

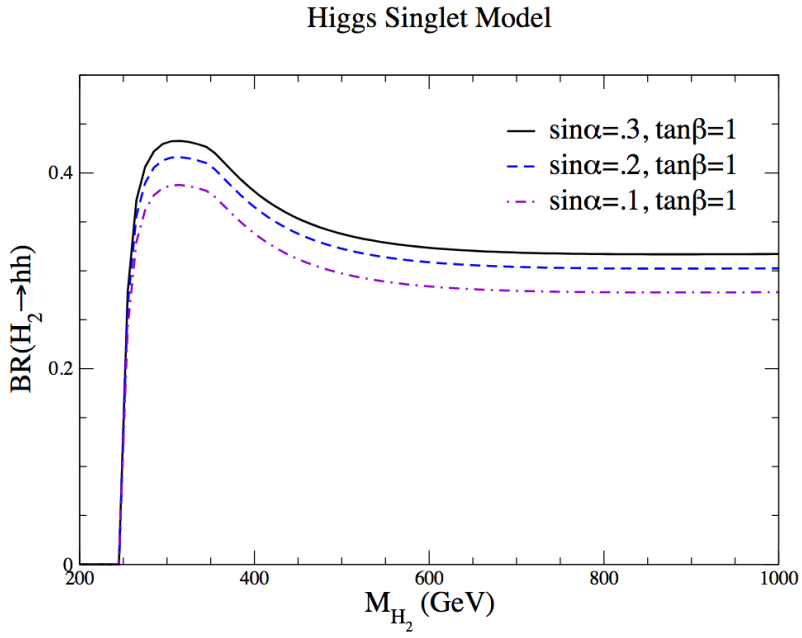


Figure 1.5: Leading order branching ratio of $H \to hh$ in the Higgs Singlet Model for representative values of the model parameters [11]. $tan\beta$ is the ratio of the vacuum expectation values of the two fields, while α is the mixing angle of the two fields.

five physical fields arise from the presence of a second doublet: two neutral scalars h and H, one neutral pseudoscalar A and two charged scalars H^+ and H^-. As it was the case in the Higgs Singlet Model, the lightest scalar h is usually assumed to be the boson observed at the LHC in 2012. At tree level, this model can be completely described by two parameters: the mass of the pseudoscalar (m_A) and the ratio of the VEV of the two fields (tanβ). Deviations of the Higgs boson couplings are induced by the non triviality of the MSSM scalar sector, and are investigated in several contexts other than double Higgs searches (Figure 1.6).

Finally, since long time the idea of a space-time with more than three spatial dimensions has been proposed in order to unify gravity and quantum mechanics. The most interesting case, for the analysis presented in this thesis, is the model proposed by Randall and Sundrum [16] that contemplates extra dimensions compactified between two points of space ("warped extra dimensions"). The relevant consequence for HH searches is the presence of new particles of spin 2 ("graviton", G) and of spin 0 ("radion", R) that can decay in two Higgs bosons. The graviton is the mediator of the gravitational force and its branching faction to an HH pair can be as large as 10% and remain constant as function of m_G, while the radion stabilizes the size of the

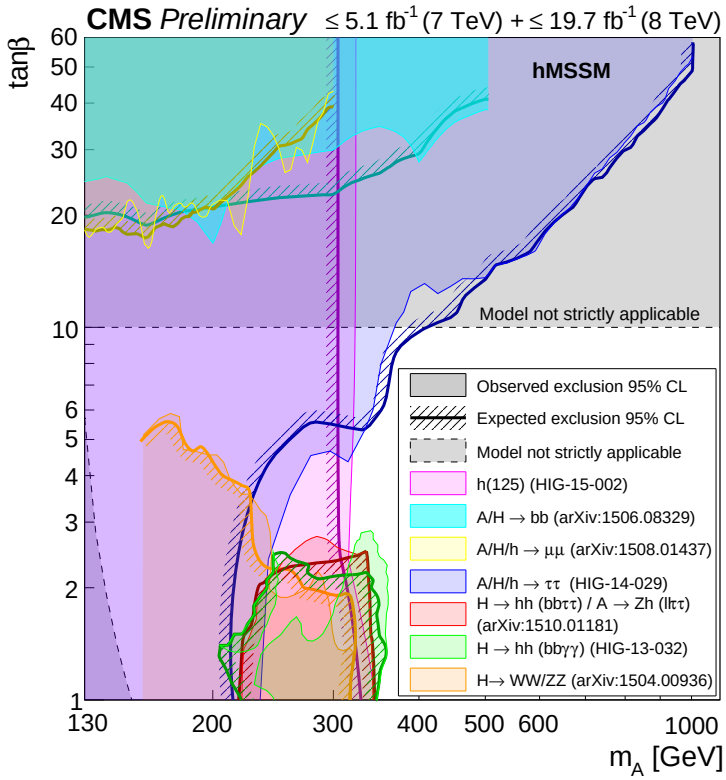


Figure 1.6: Summary of the $(m_A, tan\beta)$ regions excluded at 95% confidence level from Run I analyses performed by the CMS experiment [15].

extra dimensions with a $\mathcal{B} \approx 25\%$ and very little dependence on the model parameters.

Non-resonant BSM HH production

Non-resonant double Higgs production studies represent also an interesting field to study BSM physics, whose effects can be observed as contributions to the quantum loops that concur to HH production. In this case, the experimental signature is not as clear as it is in the resonant case, where a peak in the invariant mass spectrum m_{HH} is expected, but it occurs via enhancement of the production cross section and large modifications to the events kinematics.

In the SM, the value of λ_{HHH} is completely determined by m_H and the value of VEV, but many BSM models predict modifications to this parameter, which are often quantified with $k_\lambda = \lambda_{HHH}/\lambda^{SM}_{HHH}$. As an example, the variation of the HH production cross section for different mechanisms is

reported in Figure 1.7 as function of k_λ.

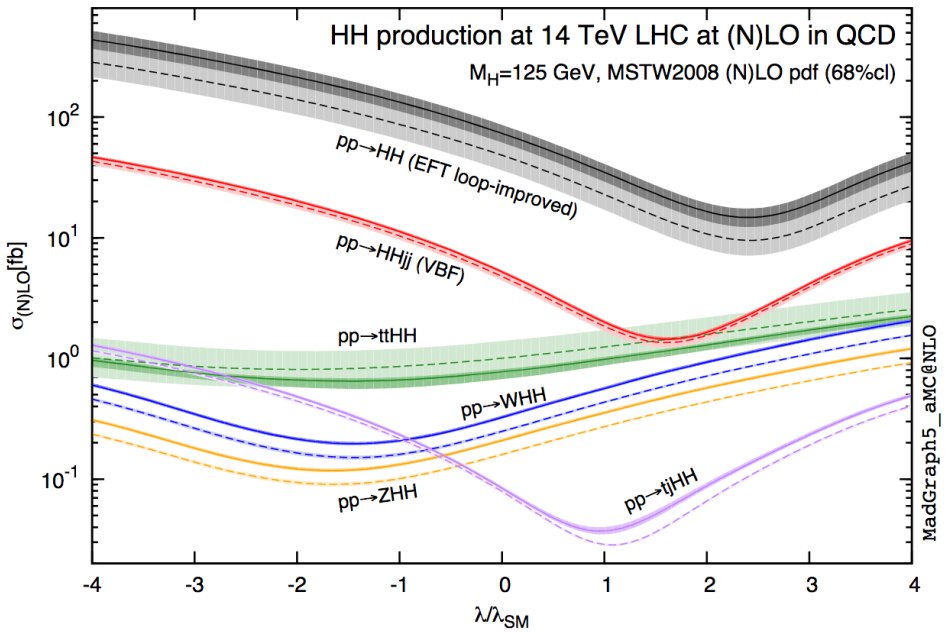


Figure 1.7: HH production cross section as a function of the coupling modifier k_λ for several production mechanisms. The dashed and solid lines denote respectively the LO and NLO predictions and the bands indicate the PDF and scale uncertainties added linearly [12].

The Higgs self coupling is not the only parameter to be modified by BSM effects, for example the Higgs coupling to the top quark can be parametrized with $k_t = y_t/y_t^{SM}$, where y_t represents the Yukawa coupling. In order to appropriately account for the dependence of the production cross section on all the possible Higgs interactions, a generalization of this approach is provided by the effective field theory (EFT) [17]. BSM effects can be approximated using higher order operators that are added to the Lagrangian: if no CP violation is assumed, the only dimension-5 operator can be neglected for HH purposes,, and the Lagrangian can be rewritten as

$$\mathcal{L} = \mathcal{L}_{SM} + \sum_i \frac{c_i}{\Lambda^2}\mathcal{O}_i^6 + \dots \tag{1.22}$$

where c_i are the Wilson coefficients and Λ is the EFT scale.

In the context of HH searches [18], the EFT Lagrangian can be expressed in terms of effective Higgs boson couplings, with the aforementioned k_λ and k_t parameters joined by three new BSM contact interactions vertices: $ttHH$

parametrized with c_2, $ggHH$ parametrized with c_{2g} and ggH parametrized with c_g. The Feynman diagrams involved in the gluon fusion process thus become five and are illustrated in Figure 1.8.

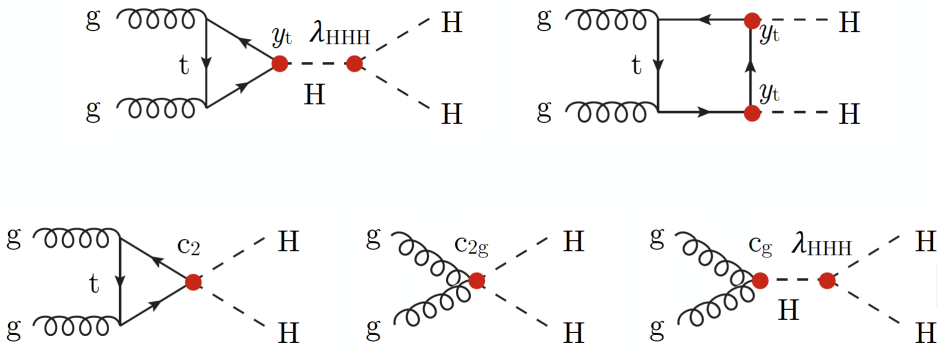


Figure 1.8: Leading order Feynman diagrams contributing to the EFT gluon fusion process $gg \rightarrow HH$. The red points highlight the BSM couplings.

As mentioned at the beginning of this Section, anomalies in the Higgs couplings modify profoundly both the cross section and the kinematic properties of HH events. Since exploring all the possible combinations of the five couplings is not feasible in terms of computing time, the optimal solution is to define some "shape benchmarks" with different combinations of the five EFT parameters that lead to distributions representative of large portions of the five-dimensional parameters space. Twelve benchmarks are identified [19] and their m_{HH} distributions are shown in Figure 1.9, while the respective couplings values are reported in Table 1.3.

1.4 Experimental status on HH searches

The cross section for HH production is very small because of the destructive interference that originates from the two concurrent Feynman diagrams contributing to the gluon fusion mechanism and the number of expected HH events produced at the LHC is low. Table 1.4 reports an approximate number of expected double Higgs events produced at the LHC for different values of integrated luminosity collected. On top of these numbers one must take into account the detector acceptance, the decay branching fractions, and the trigger and reconstruction efficiencies that all together result in a small number of detectable signal events and complicate the experimental searches for double Higgs production.

The phenomenology offered by double Higgs decays is very rich and can be explored in different final states. The choice of the decay channel, used

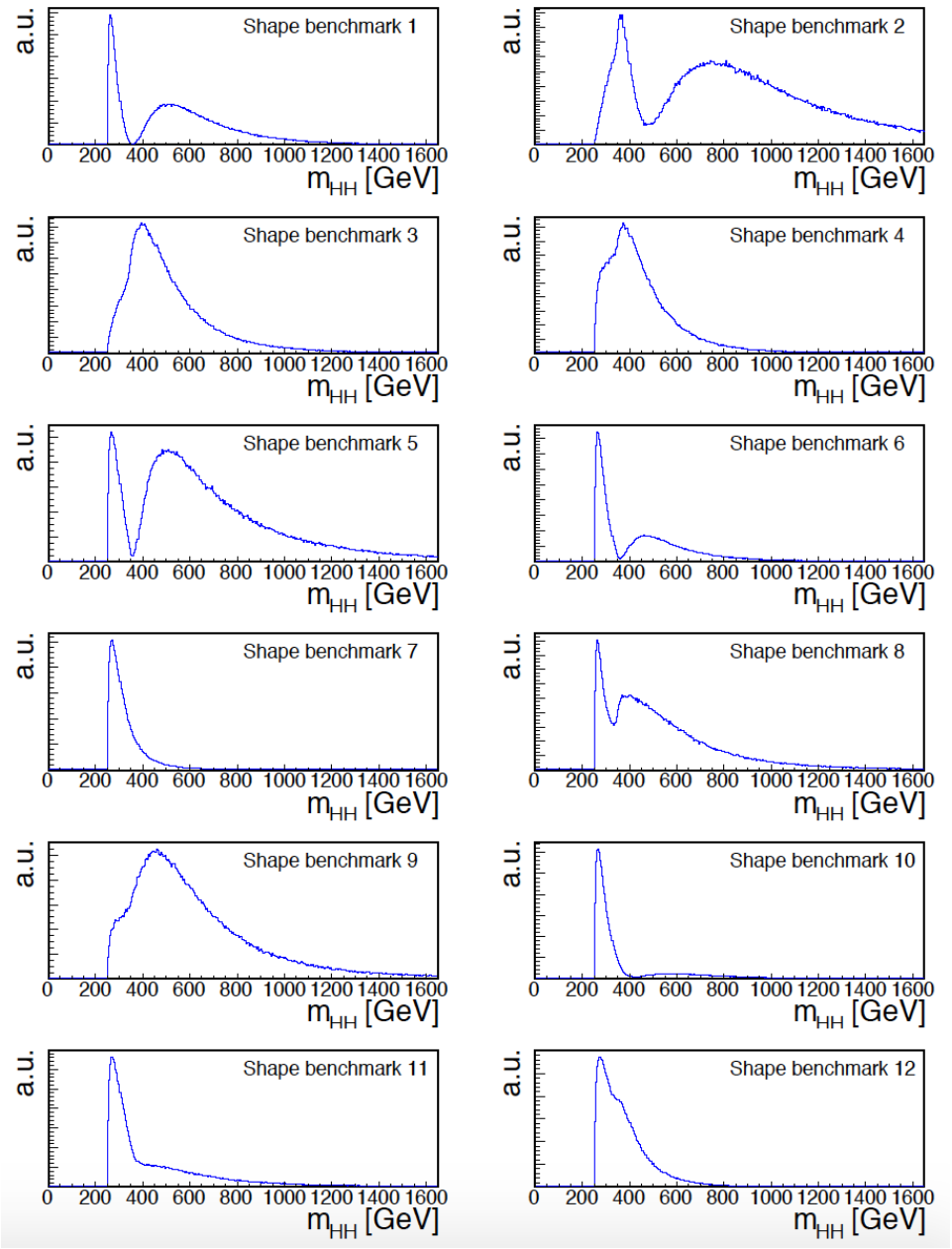


Figure 1.9: Double Higgs invariant mass distributions for the 12 shape benchmarks identified in [19].

to detect and reconstruct HH signal events, is crucial and always involves a trade-off between the branching fraction and the background contamination, which in case of double Higgs events is very large. The decay branching fractions for HH final states are shown in Figure 1.10.

Benchmark	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
6	2.4	1.0	0.0	0.2	−0.2
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

Table 1.3: Values of the effective coupling that define the 12 shape benchmarks. The last line reports the Standard Model values for comparison.

N. of HH events produced			
End of	Int. Lumi.$[fb^{-1}]$	Gluon fusion	VBF
Run I	30	300	15
Run II	150	5000	250
Run III	300	10000	500
HL-LHC	3000	100000	5000

Table 1.4: Expected number of Higgs boson pairs produced via the gluon fusion and the VBF production mechanisms at the end of different LHC "milestones". For the Run-I case, when the center of mass energy was $\sqrt{s} = 8\ TeV$, the production cross sections used are $\sigma_{GF} = 10.1\ fb$ and $\sigma_{VBF} = 0.46\ fb$.

At the moment, the LHC searches are carried out in events where at least one of the two Higgs bosons decays to either a $b\bar{b}$ or W^+W^- pair, the two channels with the highest branching fractions.

The four final states which provide the highest sensitivity in the searches and the largest coverage of possible HH topologies are:

- $b\bar{b}b\bar{b}$ has the largest branching fraction of all channels, but its sensitivity is spoiled by the large background contamination, especially at low masses
- $b\bar{b}VV$ $(V = W^{\pm}, Z)$ profits from the second highest branching ratio,

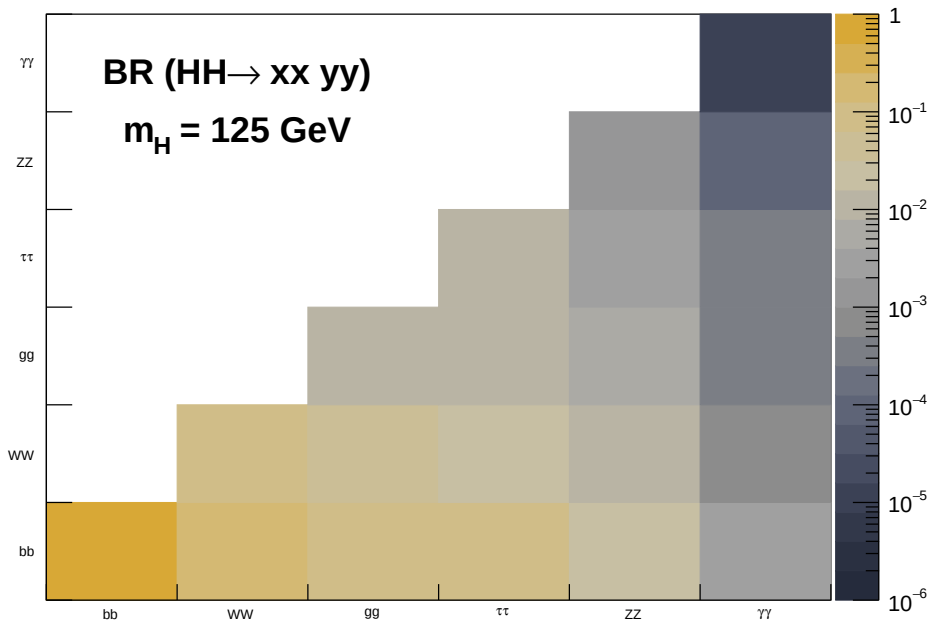


Figure 1.10: Branching factions for the decay of Higgs pairs: the plot assumes the Standard Model single Higgs values and $m_H = 125\ GeV$.

but it is affected by the irreducible $t\bar{t}$ background, especially for HH events decaying into $b\bar{b}W^+W^-$

- $b\bar{b}\tau^+\tau^-$ is the optimal trade-off between the branching fraction ($\sim 7.3\%$) and the background contamination. Several final states for the $\tau\tau$ pair are considered, but the unavoidable presence of neutrinos prevent the full reconstruction of the events.

- $b\bar{b}\gamma\gamma$ profits from the clean signature of the two photons, whose invariant mass is a powerful tool to discriminate the signal events against irreducible backgrounds. Nonetheless the small branching fraction is the primary limiting factor in the analysis sensitivity.

Searches for double Higgs production have been conducted at the LHC with data collected both at $\sqrt{s} = 8\ TeV$ and at $\sqrt{s} = 13\ TeV$.

During Run I the ATLAS collaboration investigated and combined 4 different channels $bbbb$, $bb\tau\tau$, $bb\gamma\gamma$ and $WW\gamma\gamma$ [20]. The observed and expected upper exclusion limits for the non-resonant production correspond to 70 and 48 times the Standard Model prediction, respectively. The CMS collaboration instead focused on the $bbbb$, $bb\tau\tau$ and $bb\gamma\gamma$ final states [21] with an observed limit on the non-resonant production set to 43 times the SM prediction, for an expected upper limit of 47. A comparison of both ATLAS and

CMS results on the search for resonant double Higgs production is presented in Figure 1.11 for the Spin-0 hypothesis.

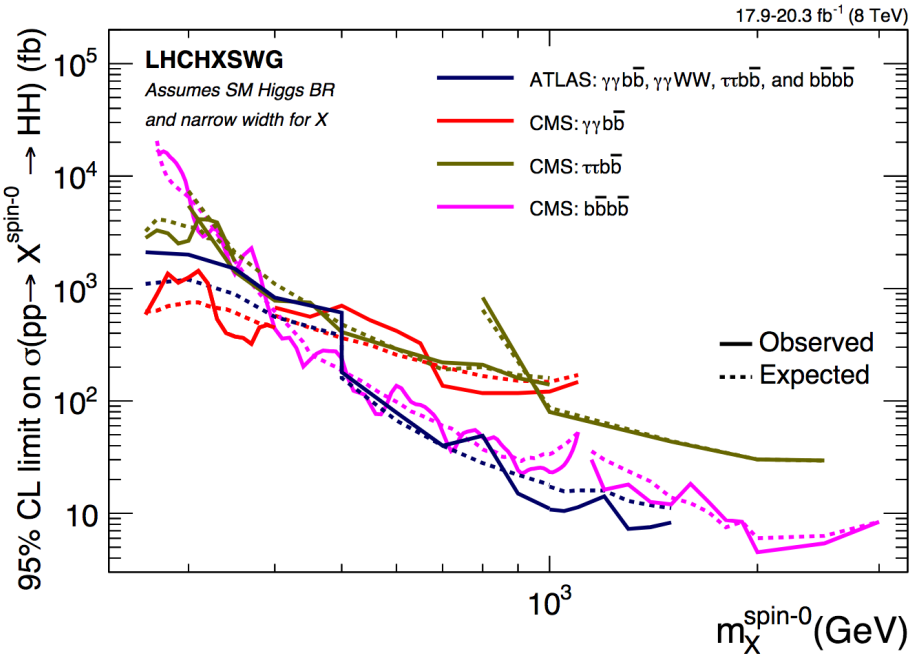


Figure 1.11: Comparison of the observed and expected 95% confidence level upper limits on $\sigma(pp \to X^{spin-0}) \times \mathcal{B}(X^{spin-0} \to hh)$ from Run I [11].

Using data collected during 2016 both the ATLAS and CMS collaborations extended their research program including other final states, such as $bbWW$ and $bbZZ$, and by exploring different topologies, such as the boosted production of Higgs bosons. In Chapter 6 a detailed description of the CMS combination using the 2016 datasets is given.

Chapter 2

Experimental Setup and Physics Object Reconstruction

The European Organization for Nuclear Research (CERN) is an international laboratory for fundamental physics. Founded in 1954, the CERN laboratory sits astride the Franco-Swiss border near Geneva and has nowadays 22 member states. More than 10000 people, including physicists, engineers and technicians, from 76 different countries, work and cooperate in synergy to explore the forefront of particle physics research at the edge of the most advanced technological development.

The CERN laboratories host the Large Hadron Collider (LHC), the largest and most powerful particle accelerator ever built. The LHC collides bunches of protons in four interactions points: in one of these points the Compact Muon Solenoid (CMS) experiment is located. More than 3500 scientists from 47 different countries are involved in the CMS experiment.

In this Chapter I will briefly describe the LHC structure and operation (Section 2.1) with particular focus on the design and concept of the Compact Muon Solenoid (CMS) experiment (Section 2.2). The $HH \to bb\tau\tau$ search described in this thesis is conducted on data collected by CMS and it involves many different physics objects, ranging from leptons to jets, whose identification and reconstruction is detailed in Section 2.3.

2.1 The Large Hadron Collider

Located in an underground 26.7 km long circular tunnel which formerly housed the CERN Large Electron Positron (LEP) collider, the Large Hadron

Collider is designed to collide protons at a center of mass energy of $\sqrt{s} = 14\ TeV$ with an instantaneous luminosity of $\mathcal{L} \simeq 10^{34} cm^{-2} s^{-1}$, as well as lead ions at a center of mass energy of $\sqrt{s} = 2.76\ TeV$ per nucleon and $\mathcal{L} \simeq 10^{27} cm^{-2} s^{-1}$ [22] [23].

In the LHC, 1232 superconductive Niobium-Titanium coils generate an $8.3\ T$ magnetic field that keeps two counter-rotating particle beams in orbit. Each dipole measures 15 meters and weights around 35 tons, while an $11\ kA$ currents runs through them in order to generate the magnetic field.

The stability of the beams is obtained thanks to 392 quadrupoles magnets that focus the particles in a narrow beam. Special quadrupoles, installed on both sides of the collision points, are used to focus the beams in order to maximize the proton density at the moment of the collision. The LHC magnets are cooled to a $1.9\ K$ temperature by a superfluid Helium-4 cryogenic system.

Particles are accelerated by Radio Frequency (RF) cavities, located in a dedicated cavern (IP4), that operate at a frequency of 400 MHz. The RF cavities are also responsible for shaping the beams into proton bunches and for the distribution of the clock to all LHC experiments.

Hydrogen atoms are stripped from electrons and accelerated to an energy of 50 MeV in the Linear Accelerator (LINAC2), the first stage of the CERN accelerator complex. Protons are then fed into the Proton Synchrotron Booster (PBS) and subsequently into the Proton Synchrotron (PS), which accelerate the particles to 1.4 GeV and 25 GeV respectively. The last acceleration step before the LHC, is the Super Proton Synchrotron (SPS) where protons reach an energy of 450 GeV. Finally, beams are injected in the two beam pipes of the LHC and accelerated in opposite directions to the target energy of 6.5 TeV. A schematic representation of the CERN injection and acceleration chain is reported in Figure2.1.

Four main experiments are installed around the interaction points where the LHC collides the proton bunches. The "Compact Muon Solenoid" (CMS) and "A Toroidal LHC ApparatuS" (ATLAS) experiments are two multi-purpose detectors, located in Points 5 and 1, where the highest instantaneous luminosity of collisions is reached. The "LHC beauty" (LHCb) experiment, a forward spectrometer built to study the B hadrons decay, is located at Point 8, while "A Large Ion Collider Experiment" (ALICE) surrounds the interaction point 2 and is dedicated to the study of heavy ion collisions and of the quark-gluon plasma.

A fundamental parameter for the LHC accelerator is the integrated luminosity, that represents a measure of the total amount of collisions produced. The luminosity is the coefficient of proportionality between the number N

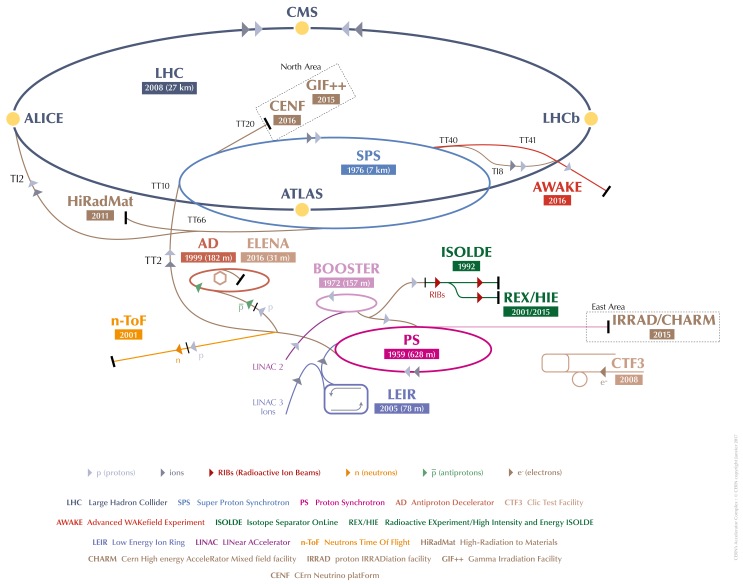


Figure 2.1: Schematic representation of the CERN accelerator facilities, protons are injected in LHC with a 450 GeV energy after a multilevel acceleration chain. In the picture all the boost stages are visible as well as the four main LHC experiments: CMS, ATLAS, ALICE and LHCb. [24]

of events produced for a specific process and its cross section σ:

$$N = L \times \sigma \tag{2.1}$$

At the LHC, the luminosity L is calculated integrating over time the instantaneous luminosity $\mathcal{L}$ of the collisions: $L = \int \mathcal{L}$. Diffferent parameters of the beam contribute to the definition of the instantaneous luminosity [25]:

$$\mathcal{L} = \frac{N_b^2 n_b f_{rev} \gamma_r}{4\pi \epsilon_n \beta^*} F \tag{2.2}$$

Table 2.1 describes the LHC parameters and their nominal value. The factor F represents the geometric reduction of the instantaneous luminosity due to the beams crossing angle θ_c and the transverse and longitudinal sizes, σ_{xy} and σ_z, of the Beam Spot (BS), the 3-dimensional region of space that envelopes

the collision of the proton bunches.

$$F = \left(1 + \frac{\theta_c \sigma_z}{2\sigma_{xy}}\right)^{-\frac{1}{2}} \tag{2.3}$$

$\sqrt{s}$	center of mass energy	14 TeV
Δt_b	bunch spacing	25 ns
N_b	particles per bunch	1.15×10^{10}
n_b	bunches per beam	2808
f_{rev}	revolution frequency	11.2 kHz
ϵ_n	transverse beam emittance	3.75 μm
β^*	beta function	0.55 m
θ_c	crossing angle at i.p.	285 μrad
σ_{xy}	BS transverse size	16.7 μm
σ_z	BS longitudinal size	7.55 cm

Table 2.1: Nominal LHC parameters during proton-proton collisions.

The first proton beams circulated in the LHC on September 10^{th} 2008, while the fist high energy collisions took place on March 30^{th} 2010. The data collected during the Run-I period, from 2010 to 2012, amount to about 6 fb^{-1} at $\sqrt{s} = 7\ TeV$ and 23 fb^{-1} at $\sqrt{s} = 8\ TeV$. After a two years long shutdown, the LHC physics program resumed in 2015 at an higher energy of $\sqrt{s} = 13\ TeV$ and marked the beginning of the so-called Run-II data taking period (2015 – 2018). A summary of the LHC performance in terms of integrated luminosity delivered to the experiments can be seen in Figure 2.2.

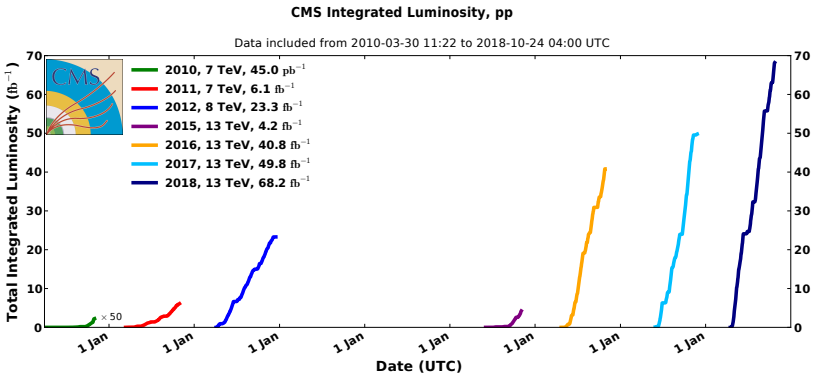


Figure 2.2: Total integrated luminosity of the LHC as function of the year, as measured by the CMS experiment [26].

In 2016 the integrated luminosity recorded by the CMS experiment amounts to a dataset of 35.9 fb^{-1} validated for the use in physics analyses and can be regrouped in data taking eras labeled from A to H. Details of the eras from B to H can be seen in Table 2.2, while era A is excluded from the Table as it was devoted to the commissioning of the machine, thus not suitable for physics analyses.

2016 data taking			
Era	Time	LHC fills	L delivered $[fb^{-1}]$
B	28 Apr-21 Jun	4879-5030	6.1
C	24 June-4Jul	5038-5071	3.2
D	4 Jul-15 Jul	5702-5095	4.6
E	15 Jul-25 Jul	5096-5117	4.6
F	29 Jul-14 Aug	5134-5198	3.4
G	14 Aug-16 Sep	5199-5303	8.5
H	16 Sep-28 Oct	5304-5471	10.0

Table 2.2: Summary of the 2016 data taking periods. For each period the time, LHC fill ranges and the integrated luminosity delivered to CMS are reported.

In 2017 the integrated luminosity recorded by the CMS experiment amounts to a dataset of 41.6 fb^{-1} validated for the use in physics analyses and can be regrouped in data taking eras labeled from A to H. As in 2016, data collected during the period A were devoted to the commissioning of the LHC machine, thus not useful for physics analyses. Collisions during era G happened at a center of mass energy of $\sqrt{s} = 5\ TeV$, while during era H at a center of mass energy of 13 TeV, but in low pile-up conditions: these periods are thus excluded from the physics analyses. Details about eras B to F can be found in Table 2.3.

2017 data taking			
Era	Time	LHC fills	L delivered $[fb^{-1}]$
B	16 Jun-18 Jul	5839-5960	4.8
C	18 Jul-30 Aug	5962-6147	9.7
D	30 Aug-13 Sep	6147-6193	4.3
E	24 Sep-11 Oct	6239-6291	9.3
F	13 Oct-10 Nov	6297-6371	13.5

Table 2.3: Summary of the 2017 data taking periods. For each period the time, LHC fill ranges and the integrated luminosity delivered to CMS are reported.

On 24^{th} of October 2018, the proton-proton data taking period of Run II officially ended after collecting over 160 fb^{-1} of data. The LHC operations will then be paused for two years, in order to replace the accelerator injectors in view of the high luminosity phase and to upgrade some detectors of the four main experiments. Run-III is planned to take place between 2021 and 2023 at a center of mass energy of $\sqrt{s} = 14\ TeV$ with an instantaneous luminosity twice as higher as the design value. At the end of Run-III the integrated luminosity delivered to the experiments should reach 300 fb^{-1} and mark the end of the so called LHC Phase I.

A 30 months stop will then take place in order to prepare the LHC to the High Luminosity Phase with the installation of new superconducting quadrupole magnets at the ATLAS and CMS interaction points. Together with the quadrupoles, newly installed compact superconducting cavities ("crab cavities") will be able to enhance the factor F in Eq.2.3 and increase the instantaneous luminosity by a factor five with respect to the original design value. The high luminosity LHC (HL-LHC) is expected to deliver about 3000 fb^{-1} of data.

Since double Higgs searches are mainly limited by statistical factors, the large amount of data collected during the HL-LHC phase will represent a unique opportunity to explore these rare processes.

A schematic representation of the past and future schedule of the LHC and HL-LHC can be seen in Figure 2.3.

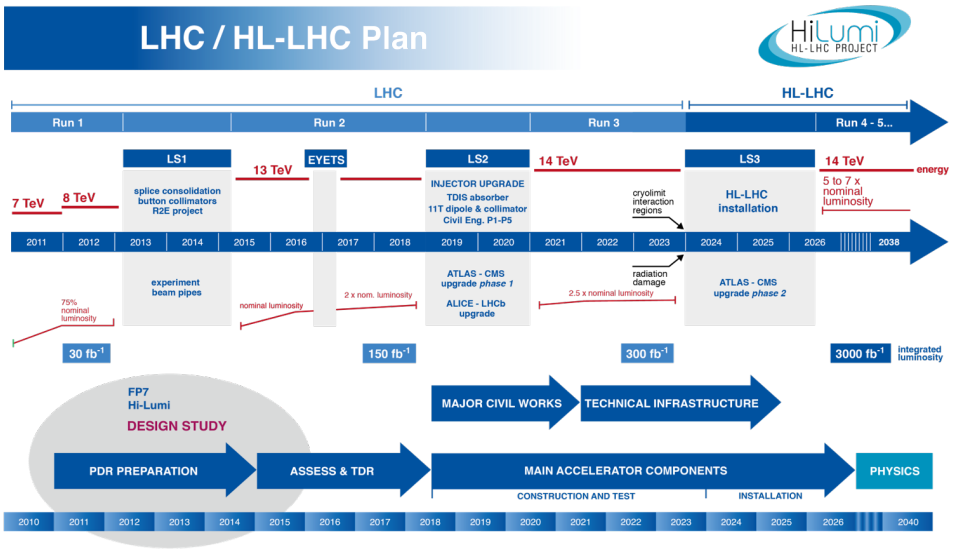


Figure 2.3: Schematic of the past and future LHC schedule [27].

2.2 The CMS Experiment

The Compact Muon Solenoid (CMS) experiment [28] [29] is located at LHC interaction point 5 and is designed as a multi-purpose detector to explore the physics at the TeV scale.

The detector is composed of a barrel section enclosed by two endcaps with a diameter of 15 m, a length of 21.5 m, and an overall weight of about 12500 t. The main feature of CMS is a large superconducting magnet, operated at a temperature of 4.5 K, which produces a 3.8 T magnetic field along the beam axis direction. The CMS solenoid surrounds the tracker system and the calorimeters, while the muon systems are placed outside between the iron return yokes of the magnetic field which, in this region, reaches a value of about 2 T. Such high values of magnetic field are used to bend the trajectories of the charged particles produced in the proton-proton collisions, and thus measure their momentum. Since tracking and muon systems are placed in two different regions of the experiment, where the magnetic field has an opposite direction, muon tracks have a double and opposite curvature that is a characteristic feature of the CMS detector.

The coordinate system chosen in CMS is defined with the center in the interaction point, the z axis along the beam direction, the x axis directed towards the center of the LHC ring and the y axis pointing upwards, orthogonally to the z and y axes. Given the cylindrical structure of the detector, a different set of coordinates is most commonly used. The azimuthal angle ϕ is defined in the (x, y), or "transverse", plane as the angle formed with respect to the positive x axis, and the radial coordinate r represents the distance from the beam axis. The polar angle θ with respect to the z axis is usually expressed as pseudorapidity $\eta = ln(\ tan\ \theta/2)$, since the difference between the pseudorapidity of two particles ($\Delta\eta$) is invariant under Lorentz boosts along the beam axis. The spatial separation of two particles in the tridimensional space is defined in terms of their angular distance $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$, while the transverse momentum p_T of a particle is the projection of the momentum onto the transverse plane.

Like many other collider experiments, the CMS experiment has a cylindrical design that encloses hermetically the interaction point and it is composed of several layers of subsystems built to identify and measure the particles produced in proton-proton collisions.

2.2.1 Inner tracking system

The CMS inner tracking system [30] is the closest subdetector to the interaction point and it is based on silicon sensors sensitive to charged particles. Being located inside the 3.8 T magnetic field, the information on the position of charged particles, also referred to as "hits", is combined in order to

reconstruct the particle trajectory, momentum and charge. Moreover, the information provided by the tracking system allows to determine the position of the hard scatter interaction point ("primary vertex") and the reconstruction of additional "secondary vertexes" originating from the decay of long lived particles such as B hadrons and tau leptons.

Given the presence of highly energetic particle beams, the tracking system must survive in an hard radiation environment where the high particle flux contributes to the detector occupancy that decreases with the distance from the interaction point as r^{-2}. In order to minimize the radiation damages to the silicon sensors, to maintain an optimal performance and to dissipate the heath produced by the electronics, the tracker system is operated at a temperature of around $-20\ ^{\circ}C$ thanks to a CO_2 cooling system.

Due to different particle flux in the tracking volume, two different systems are deployed in CMS, as illustrated in Figure 2.4 : the silicon Pixels detector in the region where $r < 16\ cm$ and the silicon Strip detector in the radial region between 20 and 120 cm.

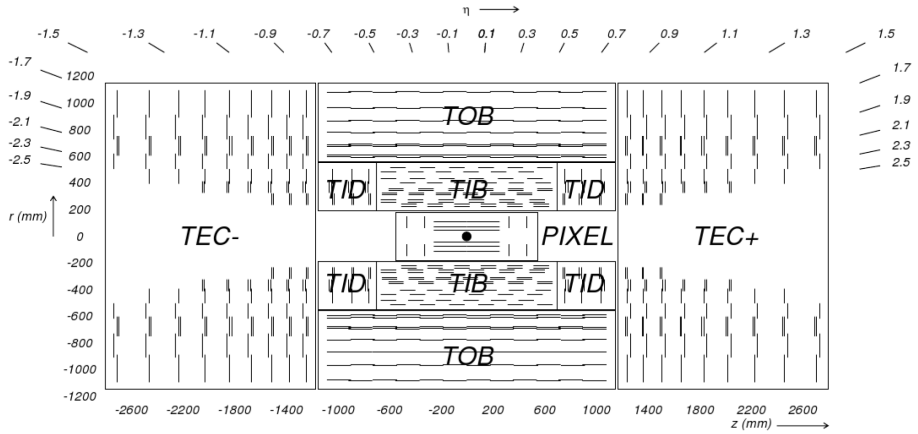


Figure 2.4: Cross section of the original design of the CMS tracking system, showing the nomenclature used to identify different sections. Each line represents a detector module. The strip tracker is composed of the tracker inner barrel (TIB), the inner disks (TID), the outer barrel (TOB) and the endcaps (TEC) [30].

After almost ten years of operations, during the End of Year Technical Stop (EYETS) $2016 - 2017$, the Pixel detector has undergone a thorough upgrade [31] from the so called Phase-0 to the Phase-I, designed to operate until the beginning of the high luminosity phase of LHC in 2026.

In the original design, the Pixel detector was composed of three layers of silicon sensors in the barrel region and two endcap disks on each side of the interaction point. Instead, the Phase-I layout of the detector comprehends

four barrel layers with the innermost at a radial distance of $r = 3\ cm$ from the interaction point and a further disk on each endcap side, providing an additional "hit" in both regions (Figure 2.5).

Despite being the smallest subsystem ($\sim 40 \times 100\ cm$), the Pixel detector has the largest number of modules: it is composed of about 125 million silicon cells (compared to the 66 millions of Phase-0) measuring $100 \times 150\ \mu m^2$ for a total active area of $\sim 2\ m^2$.

The barrel layers are located at a radial distance of 3, 6.8, 10.9 and 16 cm from the beam line and measure 54.9 cm each, while the three endcap disks have a radial coverage from 4.5 to 16.1 cm and are located at 29.1, 39.6 and 51.6 cm from the interaction point. The spatial resolution of the pixels is 10 μm in the (r, ϕ) plane and 20 μm along the longitudinal coordinate, in a pseudorapidity region of $|\eta| < 2.5$.

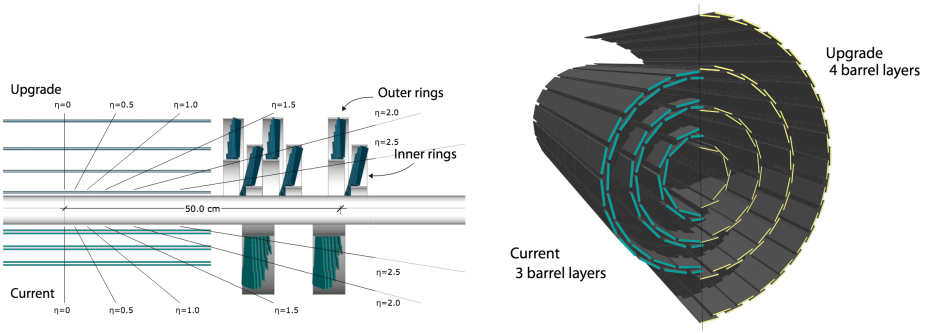


Figure 2.5: Comparison of the Phase-0 and Phase-1 of the Pixel detector. On the left longitudinal view of the upgraded layout (top) with respecto to the original design (bottom). On the right comparison of the pixel barrel layers [31].

In the Strip detector two different sensor geometries are used: at intermediate distance from the interaction point ($20 < r < 50\ cm$) silicon micro-strips with a cell size of 10 cm $\times$ 80 μm are used, while for radial distances greater than 50 cm the reduced flux of particles allows for the deployment of larger strip cells of size 25 cm $\times$ 180 μm. The entire Strip tracker system is composed of about 9.6 million silicon strips with an overall active are of $\sim 200\ m^2$.

The Strip tracker is organized in four sectors (see Figure 2.4) and, as the Pixel detector, covers a pseudorapidity region between -2.5 and $+2.5$. Four layers, that extend up to $|z| < 65\ cm$, shape the Tracker Inner Barrel (TIB), composed of modules parallel to the beam line with a strip pitch that varies from 80 to 120 μm. The Tracker Outer Barrel (TOB) is instead composed of six layers, made up of modules with the same orientation as those in TIB, but they cover the region up to $|z| < 110\ cm$. The Tracker Inned Disk (TID)

and Tracker EndCap (TEC) are arranged in ring modules centered on the beam line: the TEC comprises nine disks that extend in the region from $124 < |z| < 282\ cm$, while the TID is composed of three smaller rings that fill the gap between the TIB and the TEC in the region $75 < |z| < 110\ cm$.

The two innermost layers of TIB and TOB, the two innermost rings of TEC and TID and the fifth ring of the TEC are composed of double sided (stereo) modules with a tilt angle of 100 μrad that provide a measurement in both $r - \phi$ and $r - z$ planes. The resolution on the single "hit" ranges from 20 to 50 μm in the radial direction and from 200 to 500 μm in the longitudinal one, depending on the value of r.

The main drawback of the CMS silicon inner tracker system is the large amount of material due to detector modules, support structures, cooling plants, cables and electronic devices that particles have to cross before reaching other subdetectors. The upgrade of the Pixel detector to Phase-I played a major role in the reduction of the material budget of the tracker system, as illustrated in Figure 2.6, especially thanks to the reduction of the diameter of the cooling tubes and the deployment of longer connecting cables that allowed to relocate part of the passive material outside of the tracker acceptance.

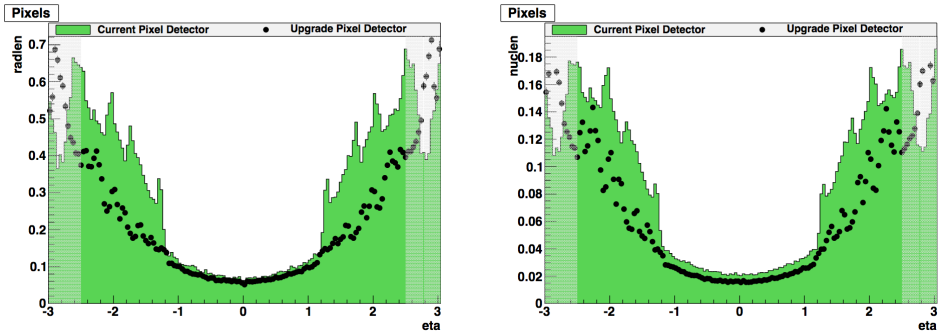


Figure 2.6: The amount of material in the Pixel detector is shown in units of radiation length (left), and in units of nuclear interaction length (right) as a function of η. The plots show the Phase-0 situation (green histogram), compared to the Phase-I upgrade (black points). The shaded regions at high values of η are outside the tracker acceptance [31].

2.2.2 Electromagnet calorimeter

The CMS electromagnetic calorimeter (ECAL) [32] is designed to measure the energy of electrons and photons produced in the proton-proton collisions. The ECAL is an hermetic and highly granular detector able to provide an excellent energy resolution thanks to the lead tungstate crystals ($PbWO_4$) that convert the incident electrons and photons into an electromagnetic shower which is in turn converted into scintillation light by the crystals themselves. The choice of such design was originally driven by the search for an Higgs boson decaying in two photons ($H \rightarrow \gamma\gamma$), where a peak in the di-photon invariant mass has to be distinguished from a continuous background.

Lead tungstate crystals are chosen as scintillating material due to its high density ($8.28 g/cm^3$), short radiation lenght ($X_0 = 0.89\ cm$) and small Molière radius ($r = 2.2\ cm$). Moreover, $PbWO_4$ is radiation hard and has very fast light emission properties, around 25 ns, allowing for a fast response, adequate to the bunch spacing in the high luminosity collisions of the LHC. The only drawback of $PbWO_4$ is the poor light yield, about 30 photons per MeV, which requires the usage of photomultipliers with high internal amplification that must be operated in a strong magnetic field. In the barrel region light is read out by silicon avalanche photodiodes (APDs), while vacuum phototriodes (VPTs) have been installed in the endcaps. Since both the light yield of the crystals and the response of the photodetectors strongly depends on temperature changes, ECAL deploys a water cooling system that guarantees a long term temperature stability.

61200 lead tungstate crystals compose the barrel region of ECAL (EB), each with a transverse section of $22 \times 22\ mm^2$ which corresponds to a granularity of $\Delta\eta \times \Delta\phi = 0.0175 \times 0.0175$, and a length of 23 cm which corresponds to a total radiation length of 25.8 X_0. This region has a coverage in pseudorapidity up to $|\eta| = 1.479$ and an inner radius of 1.29 m. Crystals are organized in 36 supermodules, each containing up to 1700 scintillating crystals and covering an angle of 20° in ϕ.

Each of the two ECAL endcaps (EE) is composed of 7324 crystals with a slightly larger section with respect to the barrel modules, $28.6 \times 28.6\ mm^2$, and a length of 22 cm, that corresponds to a total radiation length of 24.7 X_0. The EE disks are located at 3.14 m from the interaction point and cover the pseudorapidity range of $1.479 < |\eta| < 3.0$.

In order to avoid the leak of particles through the dead regions of the crystals, these are mounted, both in the barrel and in the endcaps, in a geometry which is off-pointing with respect to the mean position of the primary interaction vertex, with a 3° tilt in both ϕ and η. The layout of the crystals in the ECAL is illustrated in Figure 2.7.

A sampling preshower (ES) made of two layers of lead radiator and two silicon strip detectors is installed in front of the two endcaps to cover the

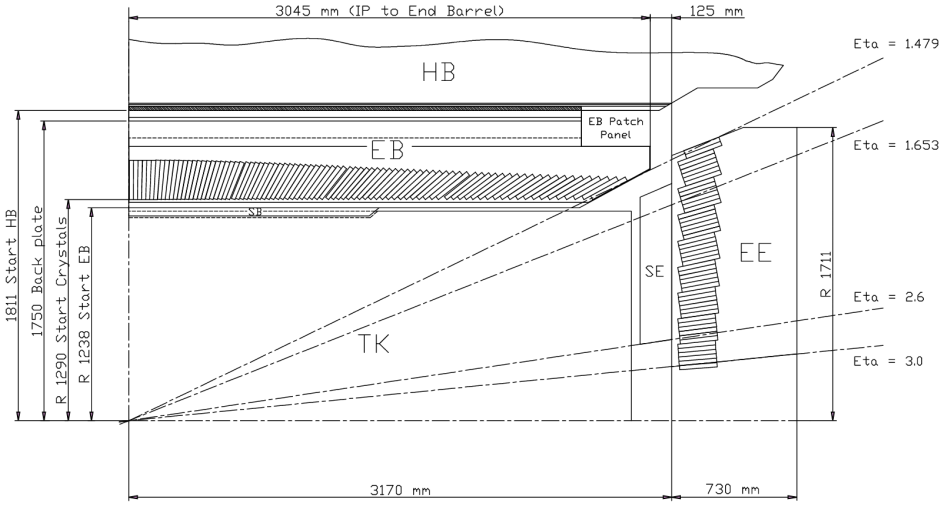


Figure 2.7: Longitudinal section of one quarter of the CMS electromagnetic calorimeter [32].

pseudorapidity region $1.6 < |\eta| < 2.6$ and is used to enhance the discrimination of single photons from $\pi_0 \to \gamma\gamma$ decays. In the passive lead absorbers the electromagnetic shower is initiated and subsequently sampled by the silicon strips in order to measure the deposited energy and the shower transverse shape.

Information collected by the ECAL detector is complementary to the data coming from the tracking system as it can cope with neutral particles, such as π_0, and its resolution increases with the particle energy itself. Indeed, the resolution of a generic calorimeter can be parametrized as:

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{S}{\sqrt{E}}\right)^2 + \left(\frac{N}{E}\right)^2 + c^2 \tag{2.4}$$

The first term (S), in the right member of Equation 2.4, represents the stochastic term depending on the number of scintillation photons emitted, which in turn depends on the incident particle energy E. N accounts for the noise in the detector and does not depend on the energy, while c is related to detector inhomogeneities and results in an error that is a constant fraction of the energy E.

The large dose of radiation coming from the beams affects the natural transparency of the crystals, thus spoiling the performance of the ECAL detector. This effect is mitigated by the natural recovery of transparency at the operating temperature of ECAL during the interval of time that is

necessary to the LHC to refill the machine with proton beams. The long term effect of radiation on the crystals is also monitored and corrected through the injection of a laser light in each crystal in order to derive time-dependent correction factors.

2.2.3 Hadronic calorimeter

The CMS hadronic calorimeter (HCAL) [33] is a sampling calorimeter, located mostly inside the magnetic coil, designed to absorb hadrons and measure their energy. With respect to what happens in ECAL, nuclear and hadronic interactions are more difficult to measure from hadron showers as they give rise to non-Gaussian tails in the energy resolution due to the production of undetectable particles such as neutrinos.

Despite this limitation, the HCAL is fundamental to measure the jets and the imbalance in the transverse momentum sum (MET) of the event. This is achieved maximizing both the material inside the magnetic coil, in terms of interaction lengths, and the geometrical coverage of the detector.

As well as the tracker and the ECAL systems, also the HCAL detector is divided in barrel (HB) and endcap (HE) sections, covering the regions $|\eta| < 1.3$ and $1.3 < |\eta| 3.0$, respectively. The absorber layer of the calorimeter is made of brass that has a short interaction length, is easy to manipulate and is a non-ferromagnetic material, while the active layer of the detector is made of plastic scintillators tiles, 3 mm thick. Wavelenght shifting fibers, embedded in the tiles, are used to collect the scintillation light that is subsequently read by hybrid photodiodes (HPDs). Modules are organized in cells, or "towers", with a spatial coverage $\Delta\eta \times \Delta\phi$ of about 0.087×0.087 in HB and 0.17×0.17 in HE.

The limited space allocated to HCAL, between ECAL and the solenoid, prevents the full containment of the hadronic showers, thus, to ensure the highest possible hermeticity of the system, the detector is complemented by two more systems. The outer hadron calorimeter (HO), lining the outer vacuum tank of the magnet coil, is composed by thicker scintillators (10 mm) and its main purpose is to sample the energy from hadron showers leaking through the rear of the calorimeters and thus to increase the effective thickness of HCAL. In the forward region, the energy measurement is improved by the forward hadronic calorimeter (HF), located at 11.2 m from the interaction point and covering the pseudorapidity region $3 < |\eta| < 5.2$. Given the high radiation level in the forward region, HF is designed with steel absorbers and quartz fibers where Cherenkov light is produced and collected by photomultiplier tubes.

An overall view of the HCAL detector is schematized in Figure 2.8.

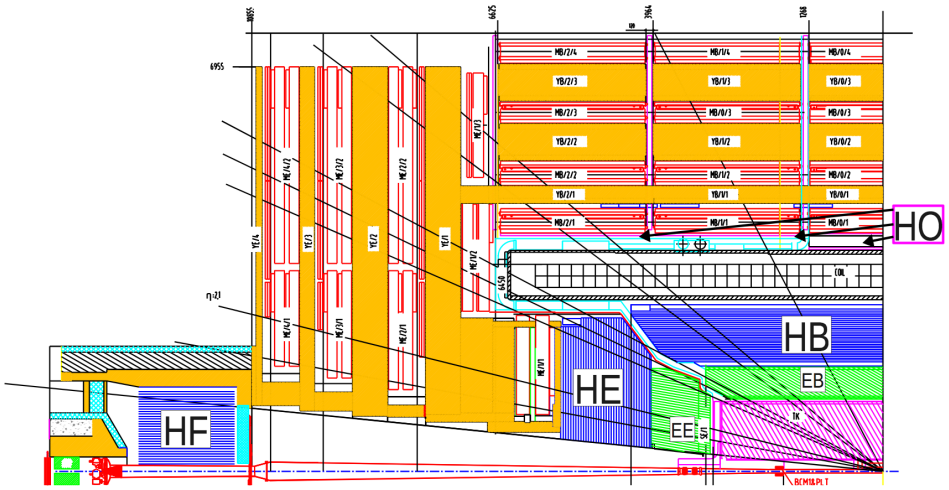


Figure 2.8: Transverse view of the CMS experiment showing the location of the HB, HE, HF and HO calorimeters [34].

2.2.4 Muon detectors

At the LHC, muons produced during collisions have minimal energy loss rates and a lifetime long enough to cross all the subdetectors and escape CMS itself. In order to identify them and measure their momentum, three different detectors [35] are employed in CMS: drift tubes (DT) in the barrel region, cathode strip chambers (CSC) in the endcap, and resistive plate chambers (RPC) in both regions (Figure 2.9). In the muon chambers, muon tracks are bent by the return magnetic field conveyed by the iron yokes; this information is combined with the inner tracker "hits" in order to measure the muon momentum.

In the barrel region ($|\eta| < 1.2$) where the neutron-induced background is small, the muon rate is relatively low and the magnetic field is uniform, CMS deploys 250 DTs uniformly spread across five barrel "wheels". Each wheel contains four concentric DT stations, composed by 60 (70 for the outermost station) DT chambers. In every chamber, twelve planes of drift tubes are organized in three super layers (SL) made of four planes each with parallel wires: two SLs measure the coordinate in the (r, ϕ) plane, the third measures the track coordinate along the z direction.

The basic element of the drift tubes is a rectangular cell of section $4.0 \times 1.3\ cm^2$ containing an anode wire and filled with a Ar/CO_2 gas mixture. Cathodes are placed on the sides of the cell, while electrodes on the top and bottom ensure a constant field and a uniform drift velocity of about 55 $\mu m/s$. A muon traversing the cell ionizes the gas and creates free electrons that drift toward the anode wires: the position and angle of the muons

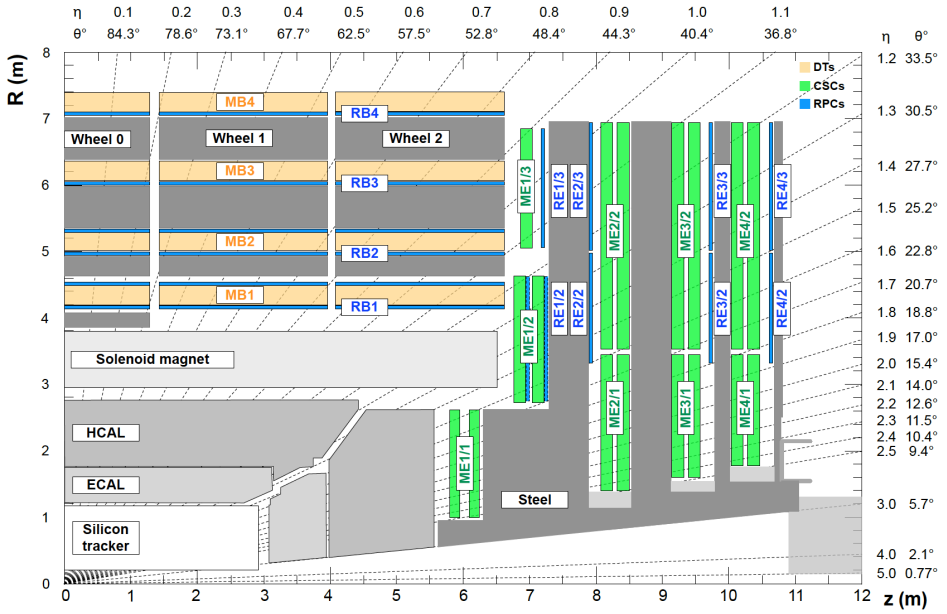


Figure 2.9: Transverse view of the CMS experiment highlighting the muon detectors. DTs, CSCs and RPCs are shown in yello, green and blue, respectively [36].

are reconstructed from the drift time information. Each DT cell has a resolution of about 200 μm, for a total resolution of chamber of about $80-120$ μm.

Cathode strip chambers are used to reconstruct muons in the endcap regions of CMS ($0.9 < |\eta| < 2.4$), where radiation levels are high and the magnetic field is non-uniform. Four CSC stations are installed in each endcap, perpendicular to the beam line and interspersed between the magnetic field return plates. CSCs are designed in a trapezoidal shape and composed of six layers of anode wires interposed between seven cathode strips plates running radially from the beamline and disposed in perpendicular direction with respect to the wires. CSC chambers contain a gas mixture of $Ar/CO_2/CF_4$ which is ionized when crossed by muons, the signal induced on the wires and the strips is collected and analyzed in order to provide a position measurements in the (r, ϕ) plane and along the z direction. The closely spaced wires make CSCs a fast detector, able to identify the bunch crossing and to achieve a spatial resolution of $40 - 150$ μm.

A complementary muon system, composed of resistive plate chambers, is implemented in both the barrel and endcap regions for a coverage in pseudorapidity up to $|\eta| < 1.6$. RPCs are highly segmented double gap chambers made of two resistive Bakelite layers separated by a $2mm$ volume filled with a

$C_2H_2F_4 - C_4H_10/SF_6$ gas mixture. They are operated in avalanche mode so that, when crossed by a muon, the high electric field in the gas volume generates an avalanche which is read out by strips located on the surface of the gap. Although the coarse spatial resolution, between 0.8 and 1.2 cm, RPCs have a sharp p_T threshold and excellent timing properties, with a resolution of the order of the nanosecond, that allow for an optimal determination of the proton-proton bunch crossings.

2.2.5 The CMS Trigger system

At the nominal LHC luminosity of $10^{34} cm^{-2} s^{-1}$ and with collisions happening every 25 ns, the information generated by the CMS detector is about 70 terabytes per second: the order of magnitude of this figure largely exceeds any achievable capability of processing and storing the data. However, each bunch crossing results in multiple soft proton-proton interaction that generate particle with low p_T and that are of no interest for the CMS physics program. This effect is called "pileup" and its evolution during the first 10 years of LHC operations can be seen in Figure 2.10.

The task of the CMS trigger system is to identify and select only the interesting collision events, thus reducing the data acquisition rate to a sustainable level. The trigger represents the interface between the "online" data taking and the "offline" data analyses, as it must satisfy the technical constraints of the former without spoiling the efficiency of the latter.

In CMS a two-tiered triggering system [37,38] is deployed and it is based on the Level-1 Trigger (L1) and the High Level Trigger (HLT). Custom hardware compose the Level-1 Trigger and processes information from calorimeters and muon systems with a reduced granularity. The decision on whether to accept or reject an event is made in 3.8 μs and the data taking rate is reduced to about 100 kHz. At the HLT level, a second selection is applied, based on the result of complex algorithms that have access to the full detector information and run on 22000 CPU cores in order to produce a decision in an average time of about 220 μs and reduce the trigger rate below 1 kHz. Both trigger systems use configurable algorithms, "seeds", to identify and reconstruct physics objects on which the accept/reject decision is made: each seed is assigned a "prescale factor" f that further reduces the trigger rate of $1/f$ by retaining only one accepted event every f occurrences. Events that are accepted by the trigger systems include data used in physics analyses, detector calibration, alignment and monitoring, and differ by the amount of detector information stored.

To deal with the time constraints coming from the 25 ns bunch crossing, object reconstruction in the L1 trigger is performed separately using only inputs from the calorimeters and the muon systems. Arrays of 5 crystals are grouped in ECAL to match the HCAL granularity and compose the so called

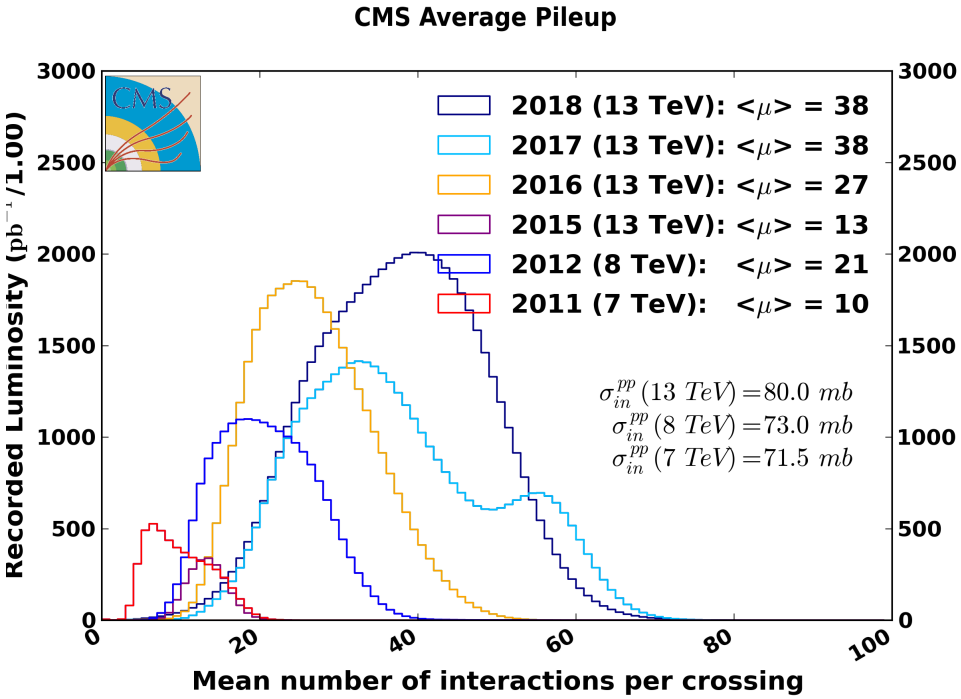


Figure 2.10: Distributions of the average number of interactions per bunch crossing (in-time pileup) for proton-proton collisions in 2011 (red), 2012 (blue), 2015 (purple), 2016 (orange), 2017 (light blue), and 2018 (navy blue). The overall mean values and the minimum bias cross sections are also shown [26].

trigger towers (TT), which represent the calorimeter trigger readout units. From these calorimetric regions, candidates for the main physics objects are built: jets, electrons, photons and hadronically decaying taus (τ_h); since no information from the tracker systems is available at this stage, electrons and photons have almost the same experimental signature and are reconstructed as e/γ objects. Information from the muon detectors is at first processed separately for each subdetector and then merged in order to reconstruct muon tracks. Finally, a Global Trigger combines information from muon and calorimetric triggers to take a decision on whether to accept or reject the event accordingly to the reconstructed objects and their properties.

Opposed to the hardware-based L1, the High Level Trigger is a software system that reduces the output rate down to 100 Hz. The basic idea of the HLT is an online reconstruction and selection that is a slightly simplified version of the offline reconstruction algorithms. This process is usually applied only locally around the objects already build by the L1 trigger, thus reducing the time needed to read and process the raw information from the detectors.

The processing time is optimized by applying selections on the most signal over background discriminating variables as soon as possible, and priority is given to the least time consuming algorithms. Jets are formed by clustering together candidates with the $anti - k_t$ algorithm, while the presence of displaced or secondary vertexes is used to asses if the jets are compatible with the hadronization and decay of a b-quark. Muons are initially built using information from the CSC and DT systems and only later are matched to tracks reconstructed in the inner silicon detector, while the trigger rate is reduced using isolation criteria based on the number of objects around the muon candidate. As for muons, also electrons and photons are reconstructed combining the ECAL information with the tracks built in the Strips and Pixel detectors, and isolation constraints are applied in order to reduce the pileup contribution and to lower the trigger rate. Finally, hadronically decaying taus are reconstructed starting from 3 charged candidates clustered inside a jet and combining them with electron/γ information. The reconstruction of HLT τ_h objects is very similar to the offline analysis, but due to timing constraints it does not carry information about the tau decay, so that the overall HLT efficiency is very high, but also the background contamination is almost one order of magnitude larger with respect to offline.

2.3 Physics object reconstruction in CMS

As described in Section 2.2, in CMS the tracker is immersed in a magnetic field that bends the charged particles trajectories and allows the electric charges and momenta of charged particles to be measured. Electrons and photons are absorbed in an electromagnetic calorimeter, while charged and neutral hadrons that may initiate a hadronic shower in the ECAL as well are subsequently fully absorbed in the hadron calorimeter: the recorded clusters are used to estimate their energy and direction. While neutrinos escape the experiment undetected, muons produce "hits" both in the inner tracker systems and in the muon detectors. A simplified view of all these processes and their experimental signature is displayed in Figure 2.11.

The information from the individual subdetectors can be combined in order to improve the reconstruction and identification of each final-state particle, and to form a complete and unambiguous description of the event. The algorithm that performs such reconstruction is called Particle Flow [39]. The PF approach was developed and used for the first time by the ALEPH experiment at LEP [40] and is now driving the design of detectors for possible future colliders. As coarse-grained detectors may cause the signals from different particles to merge thus reducing the particle identification and reconstruction capabilities, a key ingredient of the Particle Flow is the fine spatial granularity of the detector subsystems, as in the case of the CMS

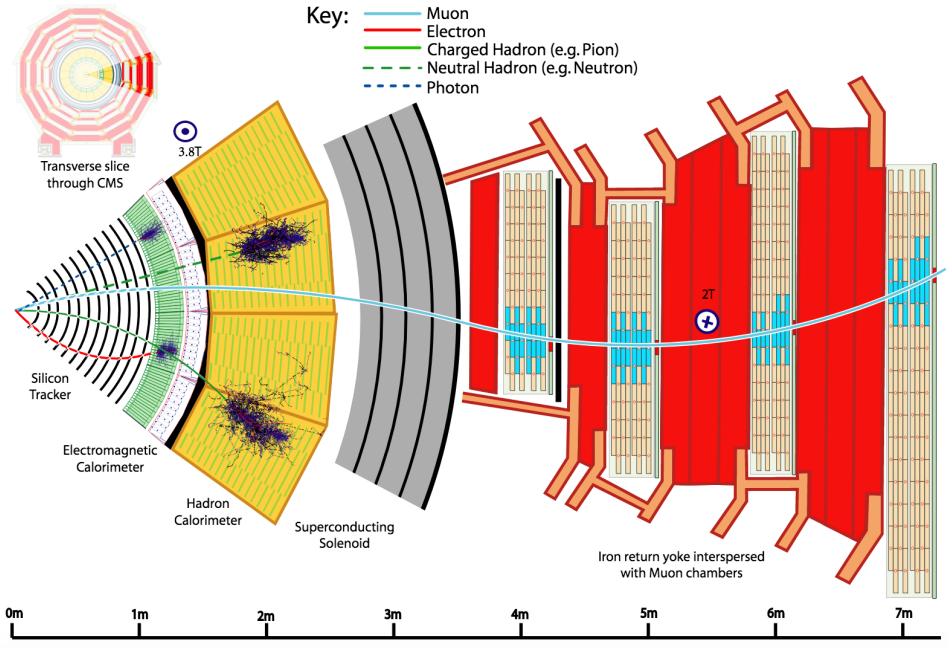


Figure 2.11: A sketch of the specific particle interactions in a transverse slice of the CMS detector, from the beam interaction region to the muon detectors. The muon and the charged pion are positively charged, while the electron is negatively charged [39].

experiment.

The reconstruction of a particle first proceeds with a *link algorithm* that connects the PF elements from different subdetectors. This algorithm can test any pair of elements, but, in order to prevent the computing time from growing quadratically with the number of particles, the pairs of elements considered by the link procedure are restricted to the nearest neighbors in the (η, ϕ) plane. Next, a distance between two linked objects is defined to quantify the quality of the link. Using the links, the algorithm produces *PF blocks* of elements associated by direct or indirect links through common elements. Thanks to the high granularity of the CMS subdetectors, most of the *PF blocks* contain few elements originating from one particle: thus, the algorithm is not affected by the particle multiplicity and the computing time needed only increases linearly with multiplicity. In each *block*, first the muon candidates are identified and removed from the block itself; secondly, electron identification is performed together with photon reconstruction, in order to collect the energy of all bremsstrahlung processes. Photons and electron PF candidates are also removed from the *PF block*. Before the last step, tracks with a p_T uncertainty in excess of the calorimetric energy resolution expected for charged hadrons are masked. The remaining elements in

the *block* are then subject to a cross-identification between charged hadrons, neutral hadrons and photons, arising from parton fragmentation, hadronization, and decays in jet.

The Particle Flow candidates are categorized as muons, electrons, photons and neutral and charged hadrons: their detailed reconstructions is described in the next Sections of this Chapter. These candidates are further on used to reconstruct higher-level analysis objects such as jets and hadronically decaying taus.

The holistic approach of the Particle Flow algorithm also gives a quick way to cross-calibrate the various subdetectors, to validate their measurements, and to identify and mask problematic modules.

2.3.1 Electrons

The reconstruction of electrons is complicated by the difficulty of modeling their interaction with the inner tracker material before they reach ECAL. Electron candidates are built starting from clusters of energy deposits in ECAL and "hits" in the tracker systems: algorithms must take into account both the non-Gaussian energy loss and the bremsstrahlung photon energy deposits that can be located outside the calorimeter.

The standard CMS electron reconstruction algorithm [41] regroups PF ECAL clusters in "superclusters" and gathers together the energy deposits associated to photons: the aggregation process depends on the cluster transverse energy E_T and exploits the $\eta - \phi$ correlations of the candidates. Electron tracks are refitted using a Gaussian sum filter (GSF) method that approximates the energy loss probability with a sum of Gaussia distributions. Complementary algorithms are use to seed the GSF tracking: the PF ECAL supercluster position and the silicon system tracks. Finally GSF tracks and PF superclusters are associated into an electron candidate and used to estimate its charge and momentum.

In order to distinguish electrons coming from the hard scatter process from those originating in "soft collisions", a multivariate approach (MVA) is used, based on a Boosted Decision Tree (BDT) and developed within the TMVA-Toolkit framework [42]. The algorithm combines observables sensitive to the amount of bremsstrahlung along the electron trajectory, the geometrical and momentum matching between the electron trajectory and associated clusters, shower-shape observables, and electron conversion variables. Furthermore, signal electrons are usually required to be isolated by applying selections on the relative isolation of the Particle Flow candidate,

defined as:

$$\mathrm{I}^{\ell}_{rel} = \left(\sum p_T^{\mathrm{charged}} + \max \left[0, \sum p_T^{\mathrm{neutral-had}} + \sum p_T^{\gamma} - \frac{1}{2} \sum p_T^{\mathrm{PU}} \right] \right) / p_T^{\ell}, \tag{2.5}$$

where $\sum p_T^{\mathrm{charged}}$, $\sum p_T^{\mathrm{neutral-had}}$, and $\sum p_T^{\gamma}$ are the scalar sums of the transverse momenta of charged hadrons originating from the primary vertex, neutral hadrons and photons, respectively; the $\sum p_T^{\mathrm{PU}}$ is the sum of transverse momenta of charged hadrons not originating from the primary vertex.

2.3.2 Muons

Muons are the only detectable particles that cross the full CMS detector and leave a clean signature in the muon systems. As a consequence, their reconstruction is done using a Kalman filter method that accounts for the energy loss in the other subdetectors materials. Three different alghoritms are deployed in CMS [43]

- **Standalone Muons** are reconstructed using the information of the muon systems only. "Hits" in the DTs, CSCs and RPCs are combined and fitted in a common muon track.

- **Tracker Muons** reconstruction starts from the inner silicon detector tracks and is extrapolated to the muon systems, requiring the presence of at least one muon segment at a compatible position.

- **Global Muons** rely on both muon and tracking systems information. Inner tracks and standalone muon tracks are propagated to a common surface and the two collections of "hits" are fitted together to form a global muon track.

Thanks to the high efficiency of both the tracker and the muon systems, about 99% of muons produced in proton-proton collisions are reconstructed either as Tracker muons or as Global muons, while Standalone muons have worse momentum resolution and higher admixture of cosmic-ray muons. Candidates found both by the Global Muon and the Tracker Muon approaches, and that share the same track, are merged into a single candidate. Depending on the muon transverse momentum, charge and momentum itself are assessed using either the inner tracking system (for soft muons with $p_T < 200\ GeV$) or the muon chambers (for $p_T > 200\ GeV$).

As for electrons, also muons are often required to be isolated in order to be distinguished from background contributions, the relative isolation of muons is defined in equation 2.5.

2.3.3 Taus

Tau leptons have a mean lifetime of about 2.9×10^{-13} s, thus they decay away from the primary interaction point creating a secondary vertex. The decay process involves the emission of a tauonic neutrino (ν_τ) and a W boson, which in turn creates a lepton-neutrino or a quark-antiquark pair (Figure 2.12).

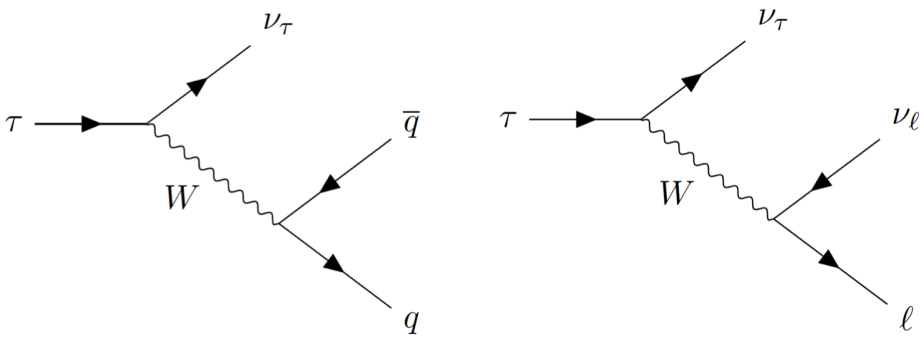


Figure 2.12: On the left the Feynman diagram of the hadronic decay of the tau, on the right the diagram for the leptonic decay.

Leptonic decays to a muon or an electron are reconstructed from the respective object algorithms detailed in Sections 2.3.1 and 2.3.2 and they always involve large values of missing transverse momentum coming from the presence of two neutrinos. On the other side, hadronic decays of the taus (τ_h) produce small and collimated hadron jets. The decay can occur through the intermediate resonances $\rho(770)$ or $a_1(1260)$ which result in different multiplicities of charged and neutral hadrons (Table 2.4). The charged hadrons produced in the decay are usually referred to as *prongs*.

The reconstruction of hadronic tau decays is performed by the Hadrons Plus Strips (HPS) algorithm [44–46] that determines the tau decay mode, identifies the PF candidates associated to the process and regroups them to estimate the τ_h kinematic properties. The HPS process begins with the analysis of the constituents of all PF jet candidates in order to verify their compatibility with a τ_h object. The contribution of $\pi^0 \to \gamma\gamma$ appears either directly as photons or as PF electrons inside the jet due to the high $\gamma \to e^+e^-$ conversion probability; photon and electron candidates are thus clustered into "strips". The strip position is recomputed as a p_T-weighted average as more candidates from a clustering region around the strip are added to the strip itself. When no more candidates are found within the clustering region the strip association is complete. During Run-II, in order to optimize the energy collection and the background rejection, a dynamic

Tau Decay Mode	Meson Resonance	$\mathcal{B}r(\%)$
$\tau^{\pm} \to e^{\pm}\nu_e\nu_\tau$		17.8
$\tau^{\pm} \to \mu^{\pm}\nu_\mu\nu_\tau$		17.4
$\tau^{\pm} \to h^{\pm}\nu_\tau$		11.5
$\tau^{\pm} \to h^{\pm}\pi^0\nu_\tau$	$\rho(770)$	25.9
$\tau^{\pm} \to h^{\pm}\pi^0\pi^0\nu_\tau$	$a_1(1260)$	9.5
$\tau^{\pm} \to h^{\pm}h^{\mp}h^{\pm}\nu_\tau$	$a_1(1260)$	9.8
$\tau^{\pm} \to h^{\pm}h^{\mp}h^{\pm}\pi^0\nu_\tau$		4.8
Other modes with hadrons		3.3
All modes containing hadrons		64.8

Table 2.4: Decay modes and relative branching fractions of a τ lepton. The symbol h generically refers to a charged pion or kaon.

strip reconstruction was deployed to define the clustering window in the $\Delta\eta \times \Delta\phi$ plane as function of the stip p_T.

The presence of extra particles within the jet, not compatible with the reconstructed decay mode of the τ, is used as criterion to discriminate hadronic τ decays from quark and gluon jets. Charged hadrons ($h^{\pm}$) and strips are combined to reconstruct one of the following decay modes:

- $h^{\pm}$
- $h^{\pm} + \mathrm{n}\pi^0$
- $h^{\pm}h^{\mp}h^{\pm}$

The decay $h^{\pm}h^{\mp}h^{\pm}\pi^0$ is not considered due to a too small branching fraction and a too large contamination from quark and gluon jets. The $h^{\pm} + \mathrm{n}\pi^0$ comprehends $h^{\pm}$ in association with both one or two π^0, since these channels are often analyzed together.

Different quality criteria are applied to all the valid decay mode hypotheses to test the compatibility with a real tau: in case of multi-prong decay, for example, all tracks must originate in the same vertex, the combined invariant mass must be compatible to that of the meson resonance and the total electric charge has to be ± 1.

2.3.4 Jets

As already mentioned in Section 1.1, quarks and gluon do not exist as free states, but they undergo an hadronization process that give life to short lived hadrons, which in turn decay in jets of lighter particles. The momentum measurement and reconstruction of jets thus, must take into account all hadronization products.

In CMS, jets are reconstructed [47, 48] using the $anti-k_T$ algorithm [49] that iteratively combines the PF candidates close to each other according to a metric defined to produce jets of an approximate conic shape around the hardest particles in the event. The size of the jet is determined by the parameter R at which the algorithm is operated: in CMS, a common choice of $R = 0.8$ is used for the boosted topologies, *i.e.* when two objects have a high momentum and their separation in $\eta - \phi$ is small, while for non-boosted jets the most commonly used value is $R = 0.4$. One of the major strengths of the $anti - k_T$ algorithm is being resilient against infrared and collinear effects, which means that its performance is not affected by soft radiation or collinear parton splitting. A set of corrections is applied in order to calibrate the detector response to the jets, especially about the jet energy scale that has to take into account pileup contributions, non-linearities of the calorimetric detectors and the residual differences between data and the simulated events.

Jets originating from b quarks are identified using either the Combined Secondary Vertex (CSV) [50] algorithm (for data recorder before 2017), or the Deep Combined Secondary Vertex (DeepCSV) [51] algorithms. Both methods exploit the long lifetime of hadrons containing b quarks that usually decay far from the primary interaction vertex. Information from the secondary vertex is combined with track-based variables into a single discriminant using a multivariate technique. In CMS, the CSV and DeepCSV algorithms are defined only for jets with $|\eta| < 2.4$ since this is the acceptance region where tracking information is available [52]. A jet is qualified as b-tagged if the value of the discriminant is larger that a fixed threshold ("working point") that determines the efficiency of correctly identified b jets and the misidentification probability for gluon and light flavour jets.

2.3.5 Missing Transverse Energy

Since neutrinos in the final state can not be detected, their presence must be inferred from the imbalance of the total transverse momentum vector sum. The negative projection of this vector onto the transverse plane is commonly referred to as missing transverse momentum (p_T^{miss}) or missing transverse energy (MET). In the context of the $HH \to \bar{b}b\tau^+\tau^-$ analysis, the decay of tau leptons always involves at least one neutrino, two if the tau decays leptonically, and additional contribution to the MET may come from the presence of additional neutrinos in the b-jets.

In CMS, two different approaches are used to reconstruct the missing transverse momentum [53]. In the first case (Particle Flow MET), the p_T^{miss} is reconstructed by PF algorithm as the negative vectorial sum of the transverse momenta of all PF candidates in the event. Due to hardware and software limitations such as tracking inefficiencies, energy thresholds in the

calorimeters and non-linearities in the detector responses, a correction is applied to the p_T^{miss} by propagating the jet energy corrections to the MET, in order to take into account the initial jet p_T and its corrected value:

$$\vec{p}_T^{\,miss,corr} = \vec{p}_T^{\,miss} - \sum_{jets} (\vec{p}_T^{\,corr} - \vec{p}_T) \tag{2.6}$$

The second possibility to estimate the MET is know as MVA-MET and it is based on a multivariate technique that separates the PF candidates associated to the signal process from the rest of the candidates in the event to improve the p_T^{miss} resolution.

In the $HH \to \bar{b}b\tau^+\tau^-$ analysis, the Particle Flow approach is employed.

Chapter 3

The $HH \to b\bar{b}\tau^+\tau^-$ Analysis Strategy

As discussed in Chapter 1, double Higgs processes represent a fertile ground to investigate the scalar sector of the Standard Model and search for hints of BSM effects. In order to fully exploit the potential offered by the data collected by the CMS experiment, the strategy of the HH analyses is to combine the full $2016 - 2018$ statistics, that amounts to about 160 fb^{-1}, and publish the results in a so-called "legacy paper".

The $b\bar{b}\tau^+\tau^-$ final state represents one of the most interesting channels to explore double Higgs processes, given the high branching ratio and the relatively small background contamination. On the other hand, this decay mode poses some quite difficult experimental challenges, especially the need to reconstruct several types of objects, ranging from leptons to jets originating from b quarks. In order to select signal-like events, special techniques must be put in place to reconstruct the $H \to \tau^+\tau^-$ and $H \to b\bar{b}$ candidates and exploit their kinematical properties to reject the background contributions.

While the analysis of the 2016 dataset represents a fundamental step to prove the manageability of data collected at the previously unexplored 13 TeV energy regime, given the small cross sections of the investigated HH processes, the most significant results are expected with the increase of statistics. As a matter of fact, double Higgs searches are amongst the most anticipated physics results of the High Luminosity phase of the LHC and are also one of the physics cases driving the design and development of future collider machines [54].

In this thesis I will describe the results obtained analyzing data collected in 2016 at a center of mass energy of $\sqrt{s} = 13\ TeV$ and published on March 2018 in *Physics Letters B* [1], to which I contributed to one of the most crucial aspects, the background estimation. I will complement each Section with the improvements put in place during the study of the 2017 dataset and with ideas on how to further enhance the analysis sensitivity in view of the ultimate exploitation of the full statistical power of the Run II data. I am

leading the 2017 data analysis and I am the contact person for the related documentation.

The structure of this Chapter follows closely the analysis flow itself. It begins with the trigger requirements (Section 3.1), used to store events for the subsequent analysis and continues with the preselection of the objects in the final state (Section 3.2). Starting from these objects, candidates for the Higgs boson decays into τ or b pairs are identified and their properties and topologies are exploited to categorize events and improve the analysis sensitivity (Sections 3.3.1 and 3.3.2). Finally, dedicated techniques are used to enhance the separation of signal events from background contributions 3.3.3.

3.1 Trigger requirements

Events are recorded and stored using a set of HLT triggers, or "paths", that require the presence of specific objects in the final state: for this analysis the trigger paths used are tuned to look for the decay products of the tau leptons.

As discussed in Section 2.3.3, τ leptons have a short mean life and decay producing either a lighter lepton or a quark pair, in association to neutrinos. The study of events containing $H \to \tau^+\tau^-$ decays, thus require the reconstruction of different possible final states. For sake of simplicity, and where no ambiguity appears, in the following the lepton or quark charge will be omitted, the leptonic decay of the tau will be denoted as τ_ℓ (with $\ell = e$ or μ) and the hadronic decay will be denoted with τ_h. The decay of a $\tau\tau$ pair can happen in six different channels, reported in Table 3.1 togheter with the relative branching fractions.

Decay Mode	$\mathcal{B}r(\%)$
$\tau_h\tau_h$	42.0%
$\tau_e\tau_h$	23.1%
$\tau_\mu\tau_h$	22.5%
$\tau_\mu\tau_e$	6.2%
$\tau_e\tau_e$	3.2%
$\tau_\mu\tau_\mu$	3.0%

Table 3.1: Decay modes and relative branching fraction of a $\tau\tau$ pair.

The $HH \to bb\tau\tau$ search is performed exclusively in three final states: $\tau_\mu\tau_h$, $\tau_e\tau_h$ and $\tau_h\tau_h$, which in total cover about 88% of the decays. Fully leptonic channels are neglected in this search due to their smaller branching

fractions and the overwhelming contamination of background events coming from the Drell-Yan processes $Z \to \mu\mu/ee$.

In the $\tau_\mu\tau_h$ and $\tau_e\tau_h$ channels in 2016, events with a single lepton, either a muon or an electron, were selected, while in the $\tau_h\tau_h$ case the presence of two hadronically decaying taus was requested in the event. Table 3.2 documents the names of the paths used and the data taking periods during which they were deployed.

2016 analysis		
Channel	HLT path name	Runs
$\tau_\mu\tau_h$	`HLT_IsoMu22_eta2p1_v*`	all runs
	`HLT_IsoTkMu22_eta2p1_v*`	all runs
	`HLT_IsoMu22_v*`	all runs
	`HLT_IsoTkMu22_v*`	all runs
$\tau_e\tau_h$	`HLT_Ele25_eta2p1_WPTight_Gsf_v*`	all runs
$\tau_h\tau_h$	`HLT_DoubleMediumIsoPFTau35_Trk1_eta2p1_Reg_v*`	from run B to G
	`HLT_DoubleMediumCombinedIsoPFTau35_Trk1_eta2p1_Reg_v*`	run H

Table 3.2: Trigger paths used in the $\tau_\mu\tau_h$, $\tau_e\tau_h$, and $\tau_h\tau_h$ channels in 2016. In cases where multiple paths covering the same runs are listed, the logical OR of these paths is used.

In the $\tau_\mu\tau_h$ channel, the logical OR of two paths with different isolation definition is used: at the HLT level, the reconstructions of muons starts from L1 trigger μ candidates and computes the isolation of the lepton either from ECAL and HCAL detectors information, or from the tracks, built in the tracker system, around the μ candidate.

In the $\tau_e\tau_h$ final state, the electron required by the HLT path is reconstructed with a similar approach to the offline strategy and its isolation is computed from the scalar sum of the energy clusters and tracks in a cone of size $\Delta R < 0.3$ around the candidate.

Finally, in the $\tau_h\tau_h$ channel, two τ_h objects are required at trigger level. The candidates are built from charged hadrons and π^0 candidates in an approach similar, but simplified due to timing constraints, to the offline HPS algorithm, detailed in Section 2.3.3. The isolation of the tau lepton, computed from tracks within a cone of radius $R = 0.4$, is the only parameter distinguishing the two paths listed in Table 3.2: during run G the background rejection was enhanced by loosening the requested quality of the tracks and tightening the selection criteria on the scalar p_T sum of HLT neutral candidates in the isolation cone.

The same triggering philosophy was adopted in 2017 regarding the single object triggers (muons and electrons), but with an increase in the p_T thresholds of the objects in order to reduce the trigger rate and cope with the higher instantaneous luminosity of the collisions. The loss of acceptance due to the tighter selections was mitigated by the introduction of the so-called "cross-lepton" triggers, which save events only if both an isolated lepton,

electron or muon, and one τ_h object are found. Table 3.3 documents the names of the paths used and the data taking periods during which they were deployed.

2017 analysis		
Channel	HLT path name	Runs
$\tau_\mu\tau_h$	HLT_IsoMu24_v*	all runs
	HLT_IsoMu27_v*	all runs
	HLT_IsoMu20_eta2p1_LooseChargedIsoPFTau27_eta2p1_CrossL1_v*	all runs
$\tau_e\tau_h$	HLT_Ele32_WPTight_Gsf_v*	all runs
	HLT_Ele35_WPTight_Gsf_v*	all runs
	HLT_Ele24_eta2p1_WPTight_Gsf_LooseChargedIsoPFTau30_eta2p1_CrossL1_v*	all runs
$\tau_h\tau_h$	HLT_DoubleTightChargedIsoPFTau35_Trk1_TightID_eta2p1_Reg_v*	all runs
	HLT_DoubleMediumChargedIsoPFTau40_Trk1_TightID_eta2p1_Reg_v*	all runs
	HLT_DoubleTightChargedIsoPFTau40_Trk1_eta2p1_Reg_v*	all runs

Table 3.3: Trigger paths used in the $\tau_\mu\tau_h$, $\tau_e\tau_h$, and $\tau_h\tau_h$ channels in 2017. In cases where multiple paths covering the same runs are listed, the logical OR of these paths is used.

The same trigger selections used to store data are also applied to MC simulated events. To account for systematic differences in the data and MC efficiencies, "scale factors" (SFs) computed with a tag and probe technique, using $Z \to \mu\mu/ee/\tau\tau$ events, are applied in MC events to the selected electron or muon candidate for the $\tau_\mu\tau_h$ and $\tau_e\tau_h$ final states, and to both the selected τ_h candidates for the $\tau_h\tau_h$ channel.

The tag and probe technique uses tight trigger, reconstruction and identification selections to identify one "tag" lepton, while it exploits the kinematics of the $Z \to \ell\ell$ decay to identify the other "probe" lepton, without directly applying trigger criteria on it. The probe leptons are thus unbiased with respect to the trigger requirements and their fraction satisfying the trigger requirements can be used to compute the trigger efficiency itself.

For the muon triggers, the SF are computed as a function of the p_T and η of the reconstructed lepton. In 2016, two separate sets of scale factors were derived for the data taking eras B to F and G-H due to the different performance of the strip tracker detector. When applying the SF to the simulated events, an average was used, based on the relative integrated luminosities collected, *i.e.* $\mathrm{SF} = f_{\mathrm{BF}}\mathrm{SF}_{\mathrm{BF}} + f_{\mathrm{GH}}\mathrm{SF}_{\mathrm{GH}}$, where $f_{\mathrm{BF}} = 0.55$ and $f_{\mathrm{GH}} = 0.45$.

The scale factors for the single electron trigger are derived as a function of the electron transverse momentum, separately for the barrel and endcap regions.

In the $\tau_h\tau_h$ channel, the trigger efficiencies are measured using $Z \to \tau\tau \to \tau_\mu\tau_h$ events, as function of the τ_h candidate transverse momentum, as well as the decay mode of the tau lepton. Due to the changes occurred in the τ_h isolation to cope with the higher instantaneous luminosity, data efficiencies

are measured separately for the B to G and H data taking periods. The SFs are thus combined according to integrated luminosity collected with a formula equivalent to the one used for the muon SFs, using $f_{\mathrm{BG}} = 0.76$ and $f_{\mathrm{H}} = 0.24$.

The trigger scale factors for muons, electrons and τ_h candidates for 2016 data are shown in Figure 3.1.

As mentioned previously in this Section, in 2017, cross-lepton triggers were introduced for the $\tau_\mu\tau_h$ and $\tau_e\tau_h$ final states in order to increase the signal acceptance. The trigger scale factors must take into account the efficiency of the logical OR between single- and cross-lepton triggers: assuming the efficiencies of the two legs to be independent, the efficiency of the logic OR can be factorized and easily computed from the single objects efficiencies. The resulting event by event SF formula is:

$$SF = \frac{Eff_{DATA}}{Eff_{MC}} \tag{3.1}$$

where the efficiency for both data and MC simulation is defined as

$$Eff = \varepsilon_L(1 - \varepsilon_\tau) + \varepsilon_l\varepsilon_\tau \tag{3.2}$$

and
ε_L = single lepton trigger efficiency
ε_l = cross lepton trigger efficiency for the τ_μ or τ_e leg
ε_τ = cross lepton trigger efficiency for the τ_h leg

The trigger scale factors for 2017 data are shown in Figure 3.2.

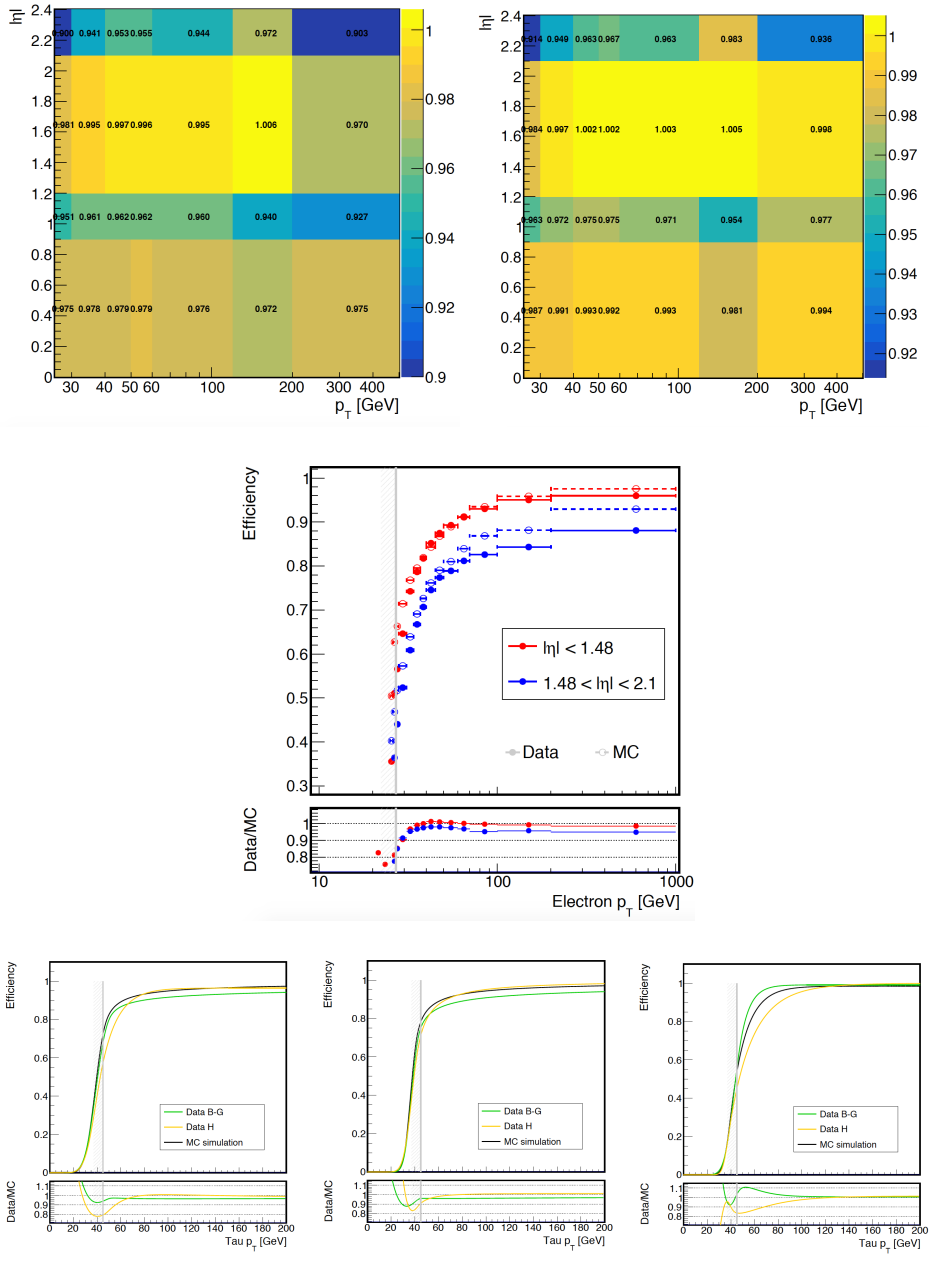


Figure 3.1: SFs for the single muon trigger (top row), for single electron trigger (central row), and for the tau triggers (bottom row) measured with 2016 data. Muon scale factors are computed for different data taking periods as function of the muon p_T and η: periods B to F (left) and G,H (right). Electron SFs are computed separately for the barrel and endcap regions as functions of the electron p_T. Tau SFs are computed for different data taking periods as function of the p_T as well as the decay mode of the τ lepton: "1-prong" (left), "1-prong+π^0" (center), and "3-prongs" (right). For electrons and taus the SFs are represented in the plots as the ratio between the data and the MC simulation.

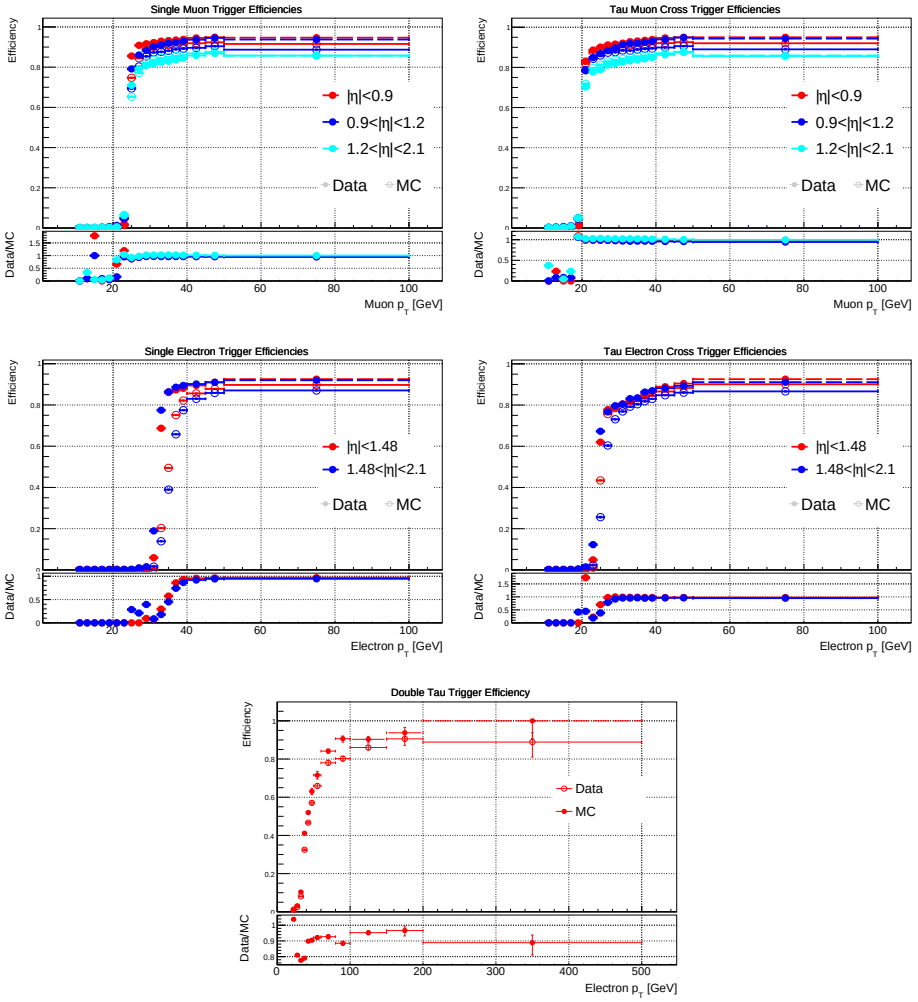


Figure 3.2: SFs for the muon triggers (top row), for electron triggers (central row), and for the tau triggers (bottom row) measured with 2017 data. For the muon and electron cases both the single lepton (left) and cross tau-lepton (right) triggers are shown. The tau trigger efficiency instead is displayed for the logical OR of the three double tau trigger paths used. The SFs are represented in the plots as the ratio between the data and the MC simulation.

3.2 Objects selections

Quality criteria are applied to the reconstructed muons, electrons, τ_h objects, jets and missing transverse momentum, in order to optimize the selection of real $HH \to bb\tau\tau$ events. This Section describes the specific choices of the final state object as well as the correction applied to the Monte Carlo simulations to eliminate any possible discrepancy.

3.2.1 Electrons

Electron reconstruction is performed using information from the ECAL and tracker detectors, as described in Section 2.3.1, while their identification is based on a BDT classifier trained for electrons with $p_T > 10\ GeV$ in three different regions: two in the barrel and one in the endcap. The relative isolation of the electron candidates is described by the Equation 2.5 as the sum of the transverse momenta of PF candidates reconstructed within a distance $\Delta R < 0.3$ from the electron, normalized to its transverse momentum.

In this search, in order to reduce the hadron jet background contamination, electron candidates must satisfy the "tight" working point of the BDT, that corresponds to a signal efficiency of about 80%, and have $\mathcal{I}^e_{rel} < 0.1$. In 2016, selection cuts on the transverse momentum and the eta of the electron were fixed at $p_T > 27\ GeV$ and $|\eta| < 2.1$, while in 2017 these values depend on the trigger path "fired", as described in Section 3.1. Finally, the electron associated track must have a distance from the primary vertex of $\Delta_{xy} < 0.045\ cm$ in the transverse plane and $\Delta_z < 0.2\ cm$ in the longitudinal direction.

A correction factor is applied to the MC simulation to take into account differences with respect to data in the isolation and identification efficiencies of electrons. These factors are derived from $Z \to e^+e^-$ events selected with a tag and probe technique. Figure 3.3 illustrates the agreement between data and MC simulation in the $\tau_e\tau_h$ channel after the application of the correction factors.

3.2.2 Muons

As described in Section 2.3.2, muons can be reconstructed as Standalone Muons, Tracker Muons or Global Muons. In order to suppress the erroneous identification of hadrons escaping HCAL and cosmic rays, the identification of muons exploits several different quality requirements such as the minimal number of hits in the muon, strips and pixel detectors and selections on the χ^2 of the associated track fit.

The quality of the reconstructed muon track and the number of hits are

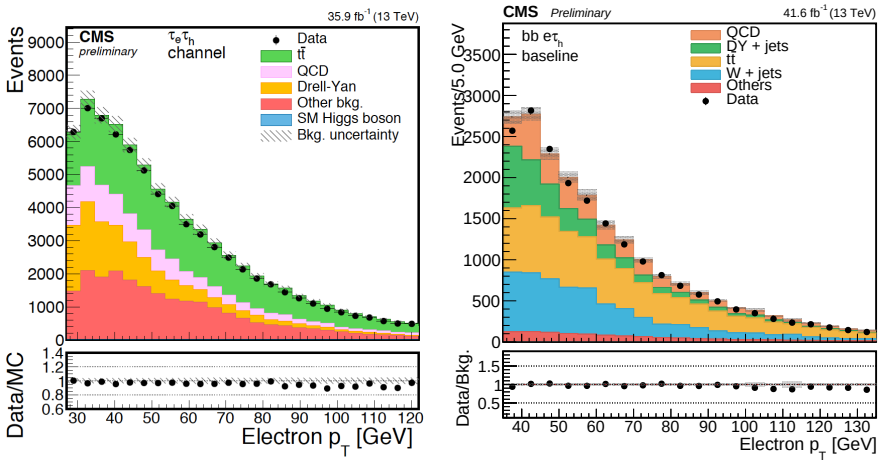


Figure 3.3: Transverse momentum distributions of the electron in the $\tau_e\tau_h$ channel. On the left events selected in 2016, on the right in 2017.

used to define different identification working points. In this analysis, the signal muon candidates are required to pass the "tight" identification criteria, while "veto muons" are defined using the "loose" WP. The efficiencies of these working points are about 96% and 99%, respectively.

As for electrons, in 2016, fixed selection cuts on the transverse momentum and the eta of muon candidates were set at $p_T > 23\ GeV$ and $|\eta| < 2.1$, while in 2017 these selections depend on the on the trigger path "fired" (Section 3.1). Also the quality selections applied to the associated muon tracks are similar to those used for electrons: $\Delta_{xy} < 0.045\ cm$ and $\Delta_z < 0.2\ cm$, while the relative isolation requirement is $\mathcal{I}^{\mu}_{rel} < 0.15$.

Differences in isolation and identification efficiencies between data and MC simulation are corrected by applying scale factors derived from $Z \rightarrow \mu^+\mu^-$ events selected with a tag and probe technique.

In 2016, due to an inefficiency in the strip tracker, the correction factors were split in two different sets, one for the periods B to F, and one for periods G and H. In the analysis the different corrections are combined in an average weighted on the relative luminosity of the two datasets, reported in Table 2.2. In addition, due to a change in the Particle Flow reconstruction, some events contained extra non-physical muons, that are removed from the analysis using the recipe provided by the CMS Muon Pysics Object Group (Muon POG).

In 2017, neither the tracking, nor the PF issues were observed, and the correction factors are derived for the full dataset. Figure 3.4 shows the data and MC efficiencies as function of the transverse momentum of the muon, divided in four different pseudorapidity region: the correction factors are

obtained from the ratio of 2017 data over MC efficiencies.

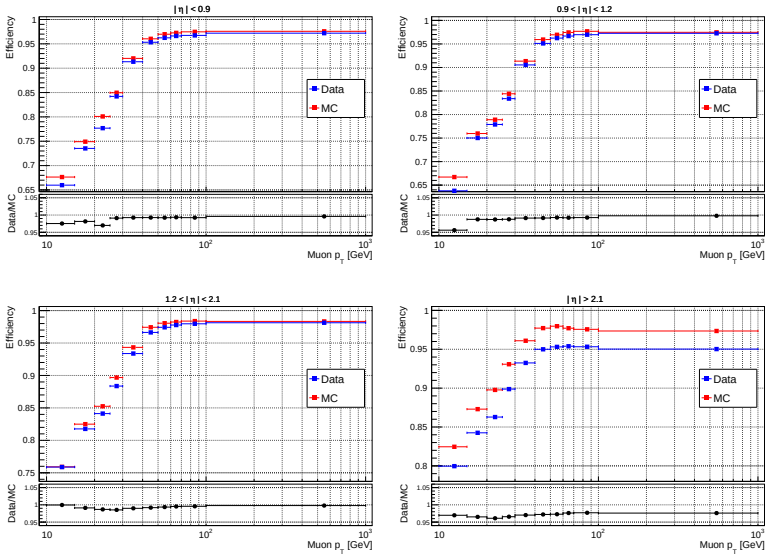


Figure 3.4: Data (blue) and MC simulation (red) efficiencies for the combination of identification and isolation criteria applied to muon candidates (top panel) and their ratio (bottom panel) used to correct the MC simulation in 2017.

The agreement between the simulation and the observed data, after the application of the correction factors, is shown in Figure 3.5 for the transverse momentum distribution of the muon.

3.2.3 Hadronic Taus

As detailed in Section 2.3.3, the decays of tau leptons into hadrons and a neutrino are reconstructed through the HPS algorithm that combines information from the PF jet constituents to identify real τ_h objects and discard fake candidates originating from a quark or gluon decay.

One of the main handles to reject quark and gluon jets is the selection criteria applied to the isolation of the τ candidate. These criteria are based on the PF candidates reconstructed inside a cone around the τ_h object itself.

In this analysis, the isolation of τ_h candidates is determined by an MVA-based approach using a Boosted Decision Trees (BDT) algorithm. The most discriminating variables used as inputs to the BDT descriminator are [45]:

- The scalar p_T sums of the charged- and neutral-particles within an isolation cone $\Delta R < 0.5$
- The τ_h decay mode

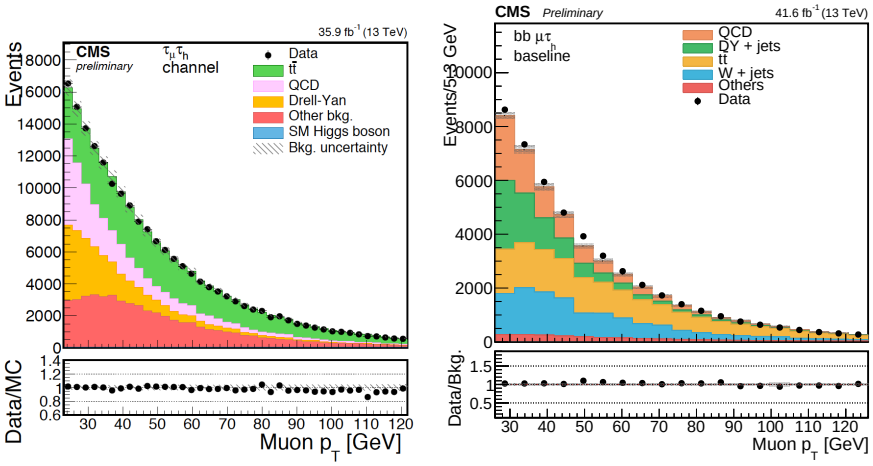


Figure 3.5: Transverse momentum distributions of the muon in the $\tau_\mu\tau_h$ channel. On the left events selected in 2016, on the right in 2017.

- The transverse impact parameter of the leading track of the τ_h and its significance
- The distance between the tau production and decay vertices and its significance

Other definitions of the τ isolation can be adopted, for example requiring that the transverse momentum sum of all the particles in the isolation cone is smaller than a fixed threshold, but the MVA approach has proven to be the most powerful in discriminating real τ_h objects from background contributions, and has thus been adopted in the $bb\tau\tau$ analysis.

Different working points are defined for the MVA discriminator which, in CMS, is maintained by the Tau Physics Object Group (Tau POG). The "medium" working point is chosen in this analysis as it represents the best compromise between background rejection and signal efficiency: for a genuine τ_h the efficiency is $\sim 60\%$ almost flat as function of the candidate p_T, while the misidentification probability ranges from 2 to 0.1% depending on the tau transverse momentum [46].

The performance of the isolation discriminator as a function of p_T for the different working points was measured by the Tau POG and is reported in Figure 3.6.

Additional discriminators are exploited to suppress the contamination coming from muons and electrons. Electrons are rejected using a BDT that is based on the fraction of energy deposited in the ECAL and HCAL subdetectors, the multiplicity, topology, and energy of the photons inside the τ_h signal cone, and the curvature of the tracks associated with the reconstructed

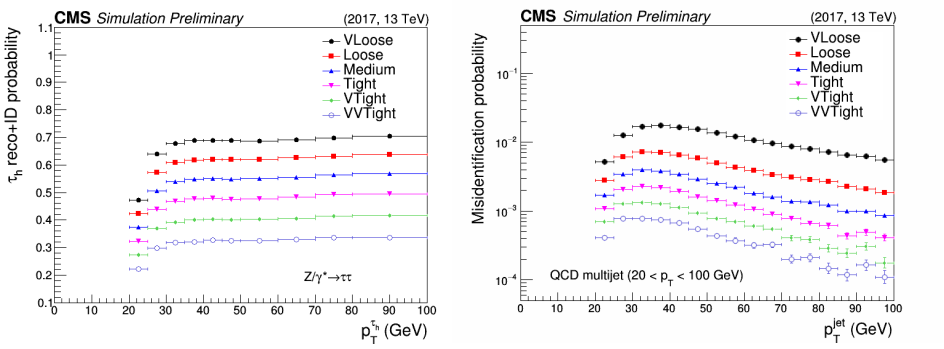


Figure 3.6: Expected τ reconstruction and identification efficiency (left) and $jet \to \tau$ misidentification probability (right) for the MVA based tau isolation discriminator in 2017 [55].

candidate. The anti-muon discriminator instead rejects τ_h candidates in case tracks in the muon systems are found aligned to the tau candidate direction. For the anti-electron discriminator, the very loose ("VLoose") working point is used in the $\tau_\mu\tau_h$ and $\tau_h\tau_h$ final states, while the "tight" WP is deployed in the $\tau_e\tau_h$ channel. Vice versa, for the anti-muon discriminator, the loose WP is used in the $\tau_e\tau_h$ and $\tau_h\tau_h$ final states, while the tight WP is applied in the $\tau_\mu\tau_h$ channel.

A detailed list of the applied selections for each final state is reported in Section 3.3.1, while the transverse momentum distributions of τ_h candidates are displayed in Figure 3.7 for 2016 and 2017 data.

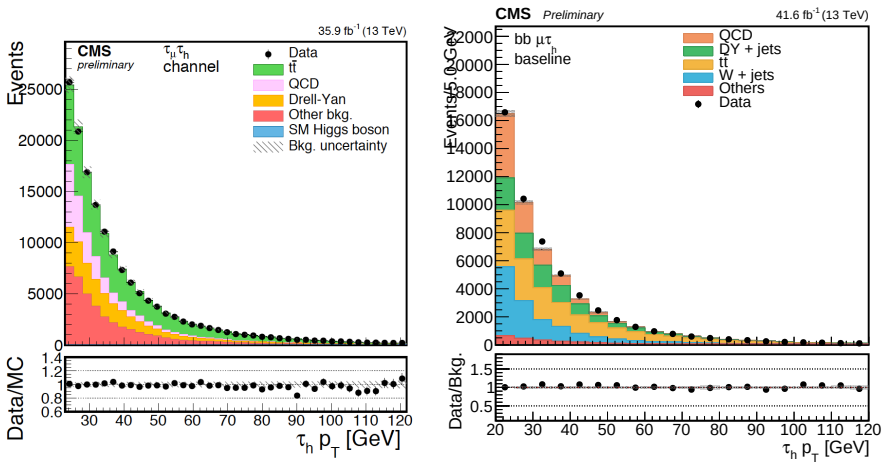


Figure 3.7: Transverse momentum distributions of the τ_h candidate in the $\tau_\mu\tau_h$ channel. On the left, events selected in 2016, on the right, in 2017.

The efficiencies for reconstruction, identification and isolation have been measured on $Z \to \tau\tau \to \tau_\mu\tau_h$ events using a tag and probe technique. In 2016 the efficiencies of data and MC simulation were found to be compatible within a 5% uncertainty, thus no further correction is needed. In 2017 instead, a constant SF of 0.95, with no dependence on p_T, η or ϕ of the τ_h candidate, has been computed in a $t\bar{t}$-enriched region. This control region has been obtained using events selected in the $\tau_\mu\tau_h$ channel and requiring both jets to pass the medium working point of the b tag discriminator described in Section 3.2.6. The invariant mass requirement defined in 3.3.3 are inverted in order to compute the SF in a phase space free from signal events.

3.2.4 MET

In $bb\tau\tau$ events, MET originates mainly from the neutrinos emitted during the τ leptons decays, while a small fraction also comes from the decays of B hadrons produced in the hadronization process of the two b quarks. The imbalance of the transverse momentum sum due to neutrinos from the B hadrons is distributed over a multitude of final products, thus reducing the dependence on the original b quarks momentum. On the other hand, the momentum of neutrinos from τ decay is directly related to the tau momentum itself and even increases with the mass of the resonances for the resonant production mechanism.

In order to ensure a good quality of the reconstructed MET, the CMS JetMET Physics Object Group (JetMET POG) provides filters, to be applied on an event by event basis, that check the quality of the reconstructed primary vertex, the effect of high energy halo muons in the calorimeters and the possible anomalous responses observed in the HCAL and ECAL detectors.

Despite the magnitude and direction of the MET are not explicitly used to select $bb\tau\tau$ events, they are combined with other relevant objects in the analysis in order to discriminate the signal from the dominant $t\bar{t}$ background contribution (Section 3.3).

As described in Section 2.3.5, in the $bb\tau\tau$ analysis, the missing transverse momentum is reconstructed using the Particle Flow algorithm, and its magnitude can be observed in Figure 3.8 for the 2016 dataset.

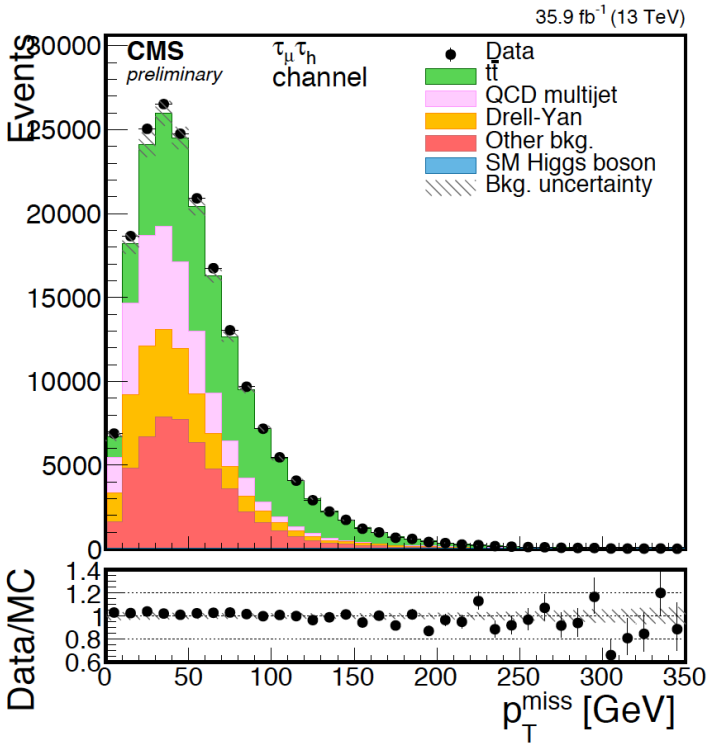


Figure 3.8: Distribution of the MET magnitude in the $\tau_\mu\tau_h$ final state for the 2016 dataset.

In 2017 data, some CMS physics analyses reported an issue with the modeling of the missing transverse momentum, especially pronounced in the last three data taking eras, 2017D, E and F, as shown in Figure 3.9). The origin of the disagreement between data and MC simulation was tracked to unexpected responses of the ECAL detector: the loss of transparency in the endcap region ($|\eta| > 2.5$) of the electromagnetic calorimeter resulted in the amplification of noise and large correction factors that subsequently generated an excess of photon candidates in this region.

Thus, we investigated the possible effect of this issue on events selected in the $bb\tau\tau$ analysis and found no significant excess of events. The agreement between data and MC simulation is good both in the magnitude and in the direction of the missing transverse momentum distributions, as shown in Figure 3.10 for the $\tau_\mu\tau_h$ final state.

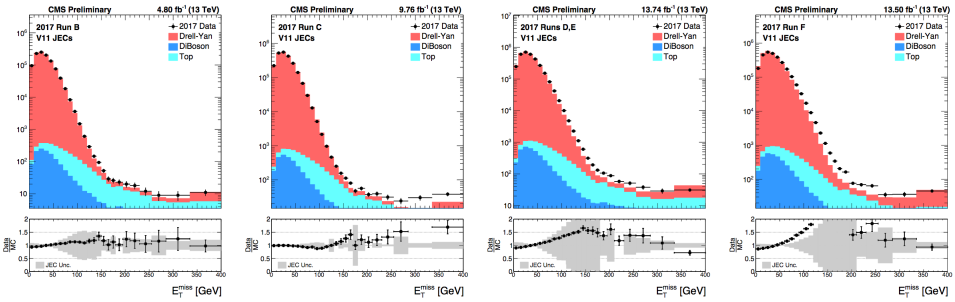


Figure 3.9: Data and MC simulation comparison of the MET magnitude divided in data taking eras, as reported by the JetMET Physics Object Group. The grey band in the bottom plots shows the jet energy correction (JEC) uncertainties obtained by varying the distributions and taking the difference of the resulting ratio plots.

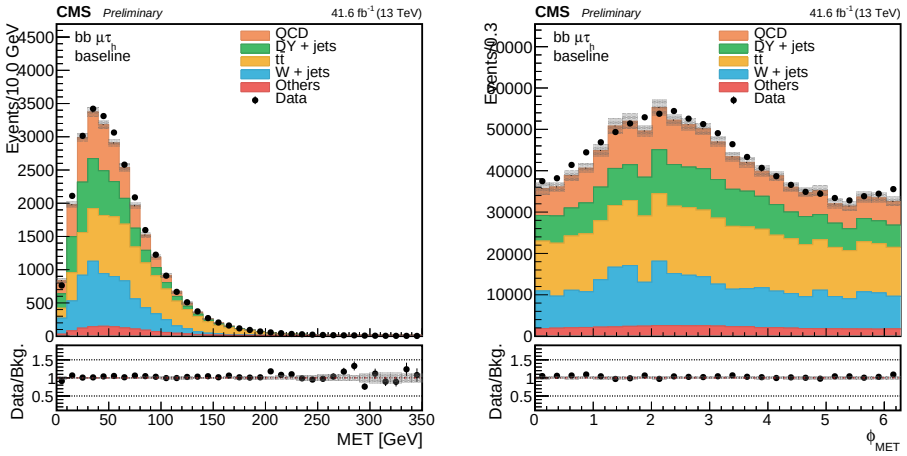


Figure 3.10: Distribution of the MET magnitude (left) and direction (right) in the $\tau_\mu\tau_h$ final state for the 2017 dataset.

3.2.5 Jets

Jets selected in the $HH \to bb\tau\tau$ analysis are recontructed through the $anti-k_T$ algorithm with a distance parameter $R = 0.4$ ($AK4$ jets) or $R = 0.8$ ($AK8$ jets), as described in Section 2.3.4.

In 2016, $AK4$ jets were required to have a transverse momentum larger that 20 GeV and a pseudorapidity $|\eta| < 2.4$, dictated by the fact that the b-tagging criteria can be applied only in the region where the silicon tracking systems information is available. In 2017, beside the jets originating from the hadronization of b quarks, also the VBF jets are taken into account, thus, different selection criteria are needed to select these objects. Jets originating from vector boson fusion processes are usually quite energetic and well separated on the azimutal plane (high $\Delta\eta_{jj}$), hence the p_T threshold can be tightened to 30 GeV and the cover range of pseudorapidity extended to $|\eta| < 5$. More details on the VBF jets selection and identification can be found in Section 3.3.2.

In order to avoid ambiguities in the event, jets reconstructed within a distance $\Delta R < 0.5$ from the two selected tau leptons are discarded in all three channels.

Selected jets are required to pass the loose (in 2016) or tight (in 2017) working point of the Particle Flow jet identification criterion: the change in the working point follows the recommendation of the JetMET physics object group as the tight jet identification efficiency has been found to be $> 99\%$. The jet identification criterion is built on jet related observables such as the fraction of charged and neutral hadron clusters, the charged hadron multiplicity and the energy deposited in the ECAL detector, and it is used to suppress jets poorly reconstructed or affected by noise in the detector.

The agreement between data and MC simulation of the jet position and transverse momentum can be seen in Figure 3.11, for 2016, and in Figure 3.12, for 2017.

The $AK8$ jets are used to achieve a better reconstruction of the events where the $H \to b\bar{b}$ decay has a high Lorentz boost and the two b quarks are produced close to each other, generating overlapping jets. In order to improve the identification efficiency of the two b quarks inside the $AK8$ jets and to reduce the contamination from initial state radiation and pileup effects, a jet substructure technique, denominated Soft Drop algorithm [56], is used: the technique iteratively decomposes the jet in sub-jets and removes the soft and wide-angled radiation.

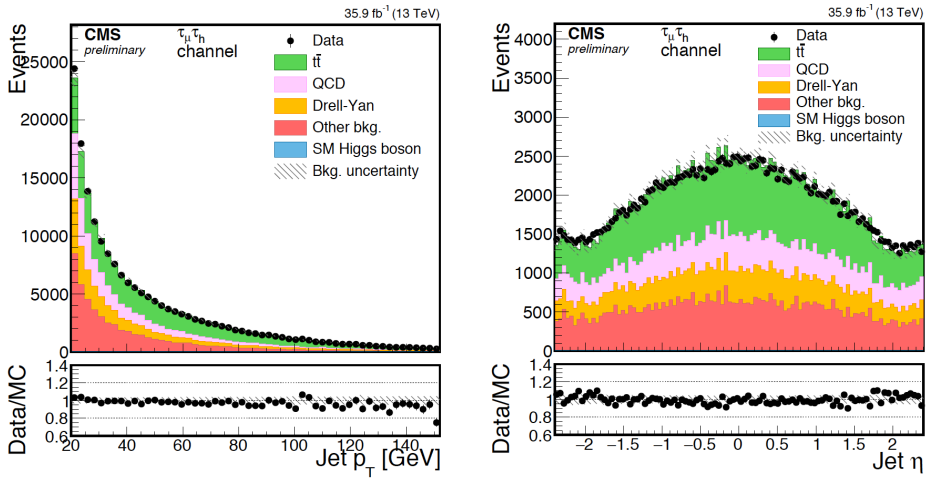


Figure 3.11: Jet p_T (left) and η (right) distributions for events selected in the $\tau_\mu\tau_h$ final state in 2016.

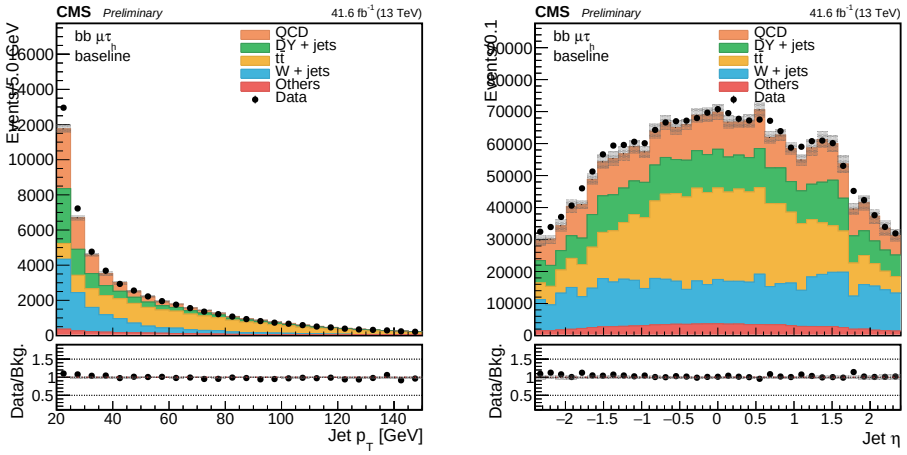


Figure 3.12: Jet p_T (left) and η (right) distributions for events selected in the $\tau_\mu\tau_h$ final state in 2017.

3.2.6 b-tagging

As described in Section 2.3.4, jets originating from b quarks are reconstructed using the CSV, for 2016 data, or the DeepCSV, for 2017 data, algorithms.

In 2016 the $bb\tau\tau$ analysis used two different working points (WP): the "loose WP", corresponding to a b jet identification efficiency of $\sim 80\%$ and misidentification rate of 10%, and the "medium WP" with an identification efficiency of about 65% for a background misidentification rate of 1%. The former WP corresponds to a selection $CSV > 0.5426$, while the latter to $CSV > 0.8484$.

To account for discrepancies in the b tagging performance, Scale Factors (SF) are defined as the ratio of the efficiencies observed in data and MC simulated events. The b Tag and Vertexing Physics Object Group (b-POG) provides to the whole CMS experiment the SFs as function of the jet transverse momentum and pseudorapidity.

MC tagging efficiencies are computed combining all the three final states in the $t\bar{t}$ sample that composes the largest background contribution. Measured efficiencies for the medium tagging WP, for the 2016 analysis, are shown in Figure 3.13.

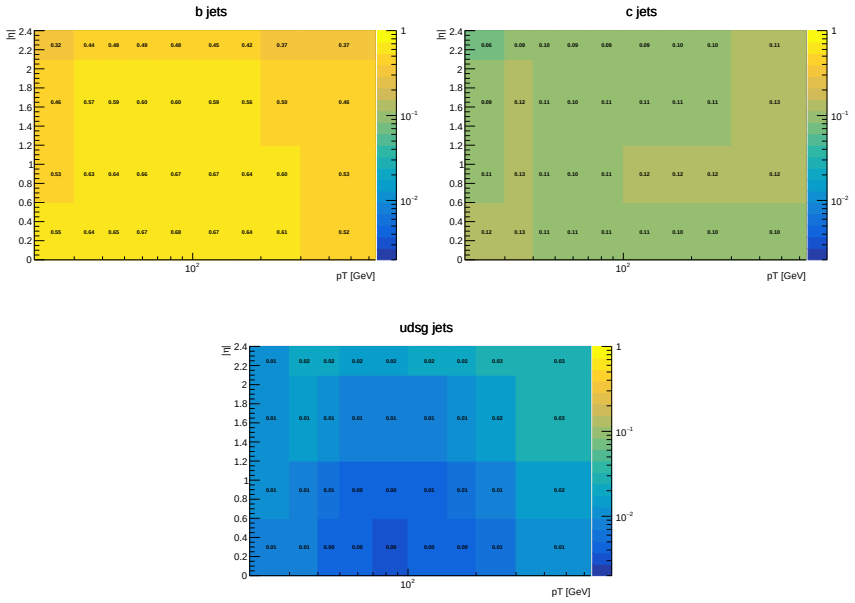


Figure 3.13: MC b tagging efficiencies derived for the CSV medium tagging working poin in 2016.

In 2017, the DeepCSV algorithm, based on a Neural Network architecture, defines five jet flavour categories. Two of these categories, P(b) and P(bb), are summed together to define a single discriminator (shown in Fig-

ure 3.14) used to identify b-jets in physics analyses. The first categoy defines the jets that contain exactly one b hadron, while the second tags the jets containing at least two b hadrons.

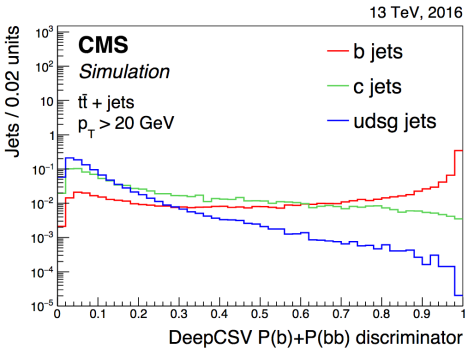


Figure 3.14: Distribution of the DeepCSV P(b)+P(bb) discriminator value for jets of different flavours in $t\bar{t}$ events. The distributions are normalized to unit area [51].

In the $bb\tau\tau$ analysis, the loose and the medium WPs, corresponding to a selection $DeepCSV > 0.1522$ and $DeepCSV > 0.4941$, respectively, are used. The measured efficiencies of the medium tagging WP for the 2017 analysis are shown in Figure 3.15.

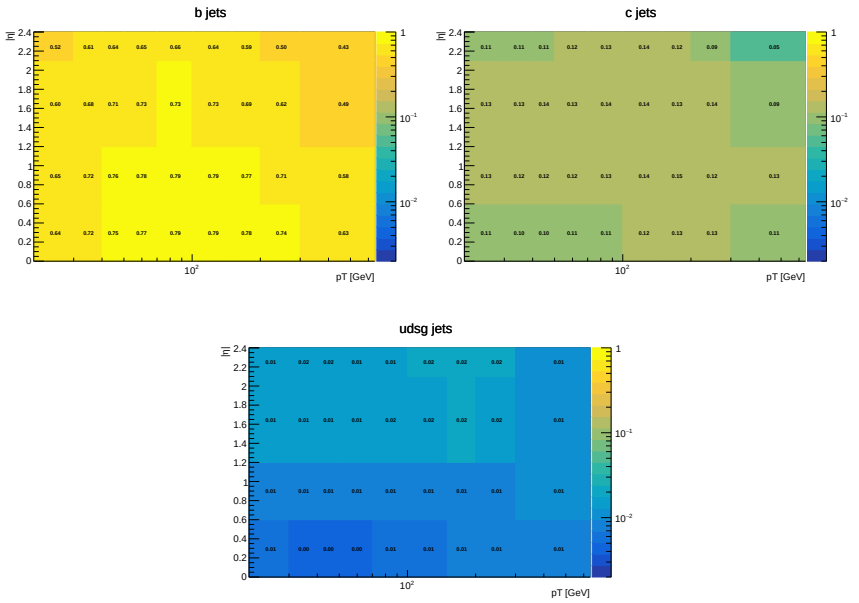


Figure 3.15: MC b tagging efficiencies derived for the DeepCSV medium tagging WP in 2017.

3.3 The "HH tag"

Once the final state objects are selected, it is necessary to determine whether suitable $H \to \tau^+\tau^-$ and $H \to b\bar{b}$ candidates are present in each event.

Sections 3.3.1 and 3.3.2 describe the efforts to reconstruct and identify the Higgs boson candidates using dedicated selections and event categorization. The events are classified in three $\tau\tau$ final states and in three (four for 2017) categories of b jets quality and topology.

Finally, in Section 3.3.3, the different methods used to reduce the background contribution and improve the analysis sensitivity are detailed. The kinematic properties of the HH signal, such as the invariant mass of the jet and lepton pairs, are exploited to further reject events not compatible with the hypothesis of a Higgs boson pair decay and multivariate techniques are used to suppress the residual backgrounds events.

3.3.1 $H \to \tau^+\tau^-$ candidates

In this Section, the selection and identification of the visible decay products of one 125 GeV Higgs boson decaying into a τ pair are described.

The $\tau\tau$ decay mode is assessed using offline information only. Selected signal events are required to have at least one τ candidate that decays hadronically and has been reconstructed by the HPS algorithm. For these events, a first loop is performed over the offline objects looking for muon and electron candidates passing the baseline selection criteria. An event is classified as $\tau_\mu\tau_h$ if a muon is found, otherwise it is classified as $\tau_e\tau_h$ if an electron is found, or as $\tau_h\tau_h$ if a second hadronic τ is present. In the semi leptonic final states, the particles, or "legs", inside each pair are ordered by assigning to the leptonic leg (μ or e) the first position. On the contrary, in the $\tau_h\tau_h$ channel all permutations are built and compared as described in the next paragraph. After the pair type has been assessed, all the pairs of the same type are sorted according to the algorithm described in the next paragraph.

All the possible pair candidates are at first sorted according to the isolation of their first leg. If the two first legs have the same isolation, the highest first leg p_T is used to order the pair. If also the transverse momentum is the same (*i.e.*, the pairs share the same first leg) the pair with the most isolated second leg is preferred, and, if the ambiguity persists, priority is given to the pair with the highest second leg p_T. This strategy has been chosen because it maximizes the purity of the event and removes any possible event overlap between the three different final states.

The two reconstructed tau leptons are required to be separated by a distance $\Delta R > 0.5$ in order to reduce cases where the same PF candidate is associated to different objects, and to have opposite charge: the charge as-

signment is very precise, especially for electrons and muons, and it represents an efficient method to enhance the signal selection efficiency.

A subsequent check that the event is firing the trigger path associated to the selected final state and that the selected offline leptons are geometrically matched to the candidates built at HLT level is performed. The correspondence is ensured by requiring that the offline and the HLT objects are within a distance $\Delta R < 0.5$.

Finally, events containing additional isolated electrons or muons, besides the two leptons used to build the $\tau\tau$ pair, are discarded. As no additional leptons are expected in signal events, this proves to be a highly efficient background rejection requirement.

2016 **analysis**

A detailed summary of all the selections applied to identify the $H \to \tau\tau$ candidates, for the three final states considered in the 2016 analysis, is given in the next paragraphs.

$\tau_\mu\tau_h$ **channel**

Events in the $\tau_\mu\tau_h$ channel are selected by requiring:

- A muon of $p_T^\mu > 23\ GeV$ and $|\eta_\mu| < 2.1$ passing tight Particle Flow muon identification criteria plus the relative isolation requirement $I_{rel}^\mu < 0.15$. The reconstructed muon production vertex must be close to the main primary vertex within a distance $\Delta_{xy} < 0.045\ mm$ and $\Delta_z < 0.2\ mm$.

- A hadronic tau of $p_T > 20\ GeV$ and $|\eta_{\tau_h}| < 2.3$ and passing the anti-electron and anti-muon discriminators. The isolation requirement on the hadronic τ is the medium working point of the MVA isolation algorithm. The same requirements on the reconstructed vertex as in the case of the muon are applied.

- Muon and hadronic tau are required to have opposite electric charge.

- In case multiple combinations of muon and τ_h exist in the event, the pair with the most isolated muon is preferred. If the muon isolation is the same, the pair with the highest p_T muon is chosen. If both previous requirements are not enough to select one pair, the pair with the most isolated τ_h is preferred.

- The event is required to pass any of the $\tau_\mu\tau_h$ triggers and the offline leptons to match the candidates reconstructed at HLT.

- The event is required to satisfy an additional lepton veto. Events are rejected if at least one additional lepton is present and passes the following selections:

  1. A muon with $p_T > 10\ GeV$, $|\eta| < 2.4$, $I^{\mu}_{rel} < 0.3$, passing the same vertex requirement as the selected muon candidate and the loose PF identification.
  2. An electron with $p_T > 10\ GeV$, $|\eta| < 2.5$, $I^{e}_{rel} < 0.3$, passing the loose MVA identification, whose reconstructed vertex is close to the main primary vertex within $\Delta_{xy} < 0.045\ mm$ and $\Delta_z < 0.2\ mm$.

$\tau_e\tau_h$ channel
Events in the $\tau_e\tau_h$ channel are selected by requiring:

- An electron of $p^{\mu}_T > 27\ GeV$ and $|\eta_e| < 2.1$ passing the tight MVA identification criteria (80% efficiency WP) and the relative isolation requirement $I^{e}_{rel} < 0.1$. The reconstructed electron production vertex must be close to the main primary vertex within a distance $\Delta_{xy} < 0.045\ mm$ and $\Delta_z < 0.2\ mm$.

- A hadronic tau of $p_T > 20\ GeV$ and $|\eta_{\tau_h}| < 2.3$ and passing the anti-electron and anti-muon discriminators. The isolation requirement on the hadronic τ is the medium working point of the MVA isolation algorithm. The same requirements on the reconstructed vertex as in the case of the electron are applied.

- Electron and hadronic tau are required to have opposite electric charge.

- In case multiple combinations of electron plus τ_h exist in the event, the pair with the most isolated electron is preferred. If the electron isolation is the same, the pair with the highest p_T electron is chosen. If both previous requirements are not enough to select one pair, the pair with the most isolated τ_h is preferred.

- The event is required to pass any of the $\tau_e\tau_h$ triggers and the offline leptons to match the candidates reconstructed at HLT.

- The event is required to satisfy an additional lepton veto. The same selections on additional leptons are applied as in the $\tau_\mu\tau_h$ channel.

$\tau_h\tau_h$ channel
Events in the $\tau_h\tau_h$ channel are selected by requiring:

- Two hadronic τ_h candidates with $p_T > 45\ GeV$ and $|\eta_{\tau_h}| < 2.1$, and passing the anti-electron and anti-muon discriminators. The isolation

requirement on the hadronic τ is the medium working point of the MVA isolation algorithm. The usual vertex requirements $\Delta_{xy} < 0.045\ mm$ and $\Delta_z < 0.2\ mm$ are applied.

- The two hadronic τ_h candidates are required to have opposite electric charge.

- The two hadronic τ_h candidates are ordered by decreasing p_T inside the pair (i.e. $p_T\tau_1 > p_T\tau_2$). In case multiple pairs satisfy the previous requirements, the pair with the most isolated τ_1 is preferred. In case the two isolation values are equal, the pair with the highest $p_T\tau_1$ is chosen. If also this requirement does not allow to select a pair, the one with the most isolated τ_2 is chosen.

- The event is required to pass any of the $\tau_h\tau_h$ triggers and the offline leptons to match the candidates reconstructed at HLT.

- The event is required to satisfy a third lepton veto. The same selections on additional leptons as in the $\tau_\mu\tau_h$ and $\tau_e\tau_h$ channels are applied.

2017 **analysis**

The main difference in the selection of the $H \to \tau\tau$ candidate, with respect to the 2016 data analysis, is the use of cross-lepton triggers that allow for a much lower threshold on the transverse momentum of the τ leptons, hence the request of a fixed cut on the lepton p_T is dropped from the $\tau\tau$ pair selection. Instead, each reconstructed offline lepton is required to pass a p_T threshold that depends on the HLT trigger path fired by the event:

$$p_T^{offline} \geq p_T^{HLT} \ + \ threshold \tag{3.3}$$

where $p_T^{offline}$ is the transverse momentum of the offline selected lepton, p_T^{HLT} is the p_T threshold applied at trigger level and $threshold$ is a fixed number depending on the lepton type: 2 GeV for muons, 3 GeV for electrons and 5 GeV for taus. The thresholds are chosen to be conservative with respect to the turn-on curves of the triggers used in the 2017 analysis and listed in Table 3.3.

A detailed summary of all the selection applied to identify the $H \to \tau\tau$ candidates, for the three final states considered in the 2017 analysis, is given in the next paragraphs.

$\tau_\mu\tau_h$ **channel**

Events in the $\tau_\mu\tau_h$ channel are selected by requiring:

- A muon of $|\eta_\mu| < 2.1$ passing tight Particle Flow muon identification criteria and the relative isolation requirement $I^{\mu}_{rel} < 0.15$. The reconstructed muon production vertex must be close to the main primary vertex within a distance $\Delta_{xy} < 0.045\ mm$ and $\Delta_z < 0.2\ mm$.
- A hadronic tau of $|\eta_{\tau_h}| < 2.3$ (2.1 for events firing only the cross-lepton trigger), passing the anti-electron and anti-muon discriminators. The isolation requirement on the hadronic τ is the medium working point of the MVA isolation algorithm. The same requirements on the reconstructed vertex as in the case of the muon are applied.
- Muon and hadronic tau are required to have opposite electric charge.
- In case multiple combinations of muon plus τ_h exist in the event, the pair with the most isolated muon is preferred. If the muon isolation is the same, the pair with the highest p_T muon is chosen. If both previous requirements are not enough to select one pair, the pair with the most isolated τ_h is preferred.
- The event is required to pass any of the $\tau_\mu\tau_h$ triggers and the offline leptons to match the candidates reconstructed at HLT.
- The event is required to satisfy an additional lepton veto. Events are rejected if at least one additional lepton is present and passes the following selections:
  1. A muon with $p_T > 10\ GeV$, $|\eta| < 2.4$, $I^{\mu}_{rel} < 0.3$, passing the same vertex requirement as the selected muon candidate and the loose PF identification.
  2. An electron with $p_T > 10\ GeV$, $|\eta| < 2.5$, $I^{e}_{rel} < 0.3$, passing the medium MVA identification (90% efficiency WP), whose reconstructed vertex is close to the main primary vertex within $\Delta_{xy} < 0.045\ mm$ and $\Delta_z < 0.2\ mm$.

$\tau_e\tau_h$ channel
Events in the $\tau_e\tau_h$ channel are selected by requiring:

- An electron of $|\eta_e| < 2.1$ passing the tight MVA identification criteria (80% efficiency WP) and the relative isolation requirement $I^{e}_{rel} < 0.1$. The reconstructed electron production vertex must be close to the main primary vertex within a distance $\Delta_{xy} < 0.045\ mm$ and $\Delta_z < 0.2\ mm$.
- A hadronic tau of $|\eta_{\tau_h}| < 2.3$ (2.1 for events firing only the cross-lepton trigger), passing the anti-electron and anti-muon discriminators. The isolation requirement on the hadronic τ is the medium working point of the MVA isolation algorithm. The same requirements on the reconstructed vertex as in the case of the electron are applied.

- Electron and hadronic tau are required to have opposite electric charge.
- In case multiple combinations of electron plus τ_h exist in an event, the pair with the most isolated electron is preferred. If the electron isolation is the same, the pair with the highest p_T electron is chosen. If both previous requirements are not enough to select one pair, the pair with the most isolated τ_h is preferred.
- The event is required to pass any of the $\tau_e\tau_h$ triggers and the offline leptons to match the candidates reconstructed at HLT.
- The event is required to satisfy an additional lepton veto. The same selections on additional leptons are applied as in the $\tau_\mu\tau_h$ channel.

$\tau_h\tau_h$ channel
Events in the $\tau_h\tau_h$ channel are selected by requiring:

- Two hadronic τ with $|\eta_{\tau_h}| < 2.1$ and passing the anti-electron and anti-muon discriminators. The isolation requirement on the hadronic τ is the medium working point of the MVA isolation algorithm. The usual vertex requirements $\Delta_{xy} < 0.045\ mm$ and $\Delta_z < 0.2\ mm$ are applied.
- The two hadronic τ are required to have opposite electric charge.
- The two hadronic τ are ordered by decreasing p_T inside the pair (i.e. $p_T\tau_1 > p_T\tau_2$). In case multiple pairs satisfy the previous requirements, the pair with the most isolated τ_1 is preferred. In case the two isolation values are equal, the pair with the highest $p_T\tau_1$ is chosen. If also this requirement does not allow to select a pair, the one with the most isolated τ_2 is chosen.
- The event is required to pass any of the $\tau_h\tau_h$ triggers and the offline leptons to match the candidates reconstructed at HLT.
- The event is required to satisfy a third lepton veto. The same selections on additional leptons as in the $\tau_\mu\tau_h$ and $\tau_e\tau_h$ channels are applied.

3.3.2 $H \to b\bar{b}$ candidates

In this Section, the selection and identification of one 125 GeV Higgs boson decaying into a $b\bar{b}$ quark pair is described.

The two b quarks are experimentally observed as hadron jets and their reconstruction must take into account the contamination coming from jets that originate from gluons or light flavour quarks. Additionally, in the 2017 data analysis, two VBF-jet candidates are selected.

For events to be selected in the $bb\tau\tau$ analysis, at least two jets with $p_T > 20\ GeV$ and $|\eta| < 2.4$ must be present. Moreover, the distance between each jet and both selected τ candidates must be $\Delta R > 0.5$.

In order to maximize the analysis sensitivity, events are categorized in separate topologies depending on the spatial overlap of the selected jets and on the number of jets that are identified as coming from b quarks accordingly to the discriminant described in Section 3.2.6.

Due to the fact that in 2017 also the double Higgs VBF production mechanism is considered in the analysis, different strategies to select the jets and categorize the events were adopted in 2016 and 2017.

2016 strategy

In order to increase the probability that the two jets selected are actually the two b-jets originating from the $H \to bb$ decay, and to enhance the signal over background ratio, those with the largest output of the b tagging discriminant are chosen. This selection criterion has proven to be the most effective in selecting $H \to bb$ events with respect to other possible solutions such as the two jets with the highest transverse momentum, or the two with combined invariant mass closest to 125 GeV.

Depending on the event topology, events are classified into resolved or boosted categories in order to improve the analysis sensitivity over the entire mass range studied for resonant production. The separation of b quarks, produced in the Higgs boson decay, depends on the Lorentz boost of the Higgs itself ($\gamma = E/m_H$) as $\Delta R(bb) \simeq 2/\gamma$. In CMS, three different regimes can be observed experimentally:

- $\Delta R(b,b) > 0.8$: for low values of γ, jets are reconstructed as separated objects with the AK4 algorithm (resolved jets)
- $0.4 < \Delta R(b,b) < 0.8$: at intermediate values of γ, the two jets can be reconstructed both as separated objects and as a single "fat jet"
- $\Delta R(b,b) < 0.4$: at high values of γ, the separation of jets is small and the AK4 algorithm is unable to distinguish them. Jets are thus merged and reconstructed only as "fat jets"

For resonances that are explored in this search with masses up to 900 GeV, the highly boosted regime ($\Delta R(b,b) < 0.4$) is never reached, events are thus categorized into a "resolved" and a "boosted" category. Events containing an AK8 jet with

- $m_{AK8} > 30$ GeV
- $P_T > 170$ GeV
- a structure with two subjets matched to two independently reconstructed AK4 jets
- both subjets passing the "loose" working point of the b-tagging discriminator

fall in the boosted category, otherwise the events are assigned to the resolved category. The fraction of events classified as boosted is reported in Figure 3.16 as function of the resonance mass and for the 12 EFT benchmarks defined in Section 1.3.1.

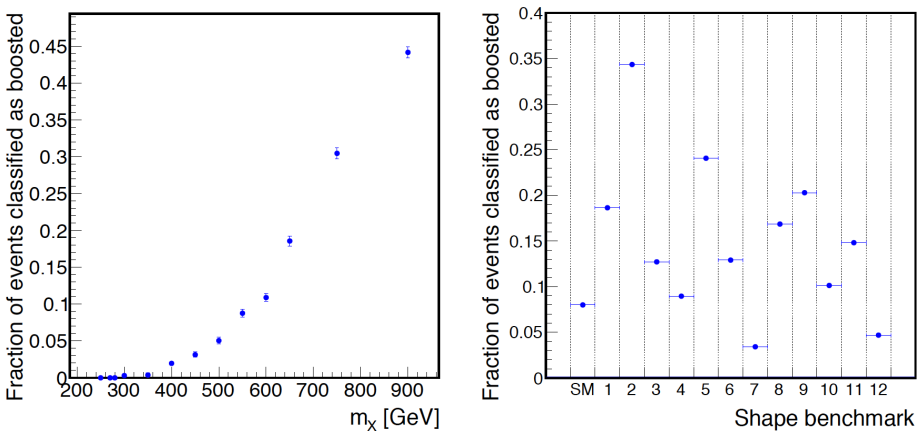


Figure 3.16: Fraction of events classified as boosted as a function of the resonance mass hypothesis for resonant HH production (left) and for the 12 EFT shape benchmarks for non-resonant production (right) .

The benefit of using AK8 jets in boosted topologies mainly originates from the clear separation of signal events from $t\bar{t}$ contributions obtained in the reconstruction of the invariant mass of the two jets, as illustrated in Figure 3.17.

Events classified as "resolved" are further split in two categories, depending on the number of b-tagged jets:

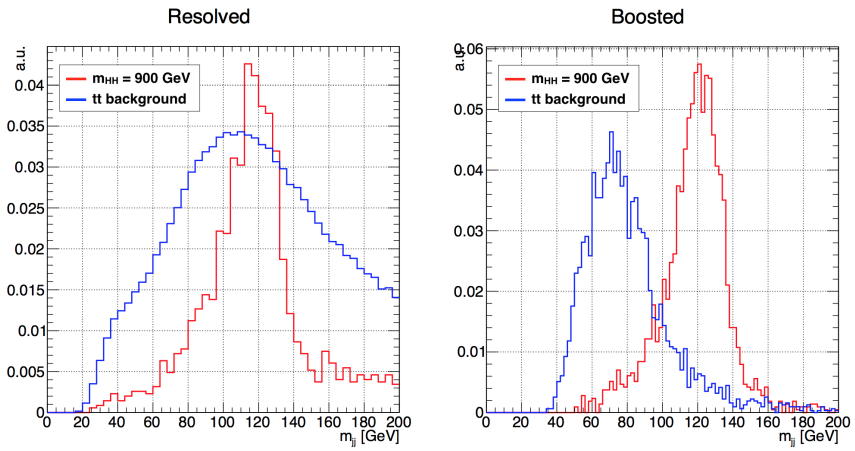


Figure 3.17: Distributions of the invariant mass of the bb pair for the resolved (left) and boosted (right) categories, assuming a signal with $m_{HH} = 900\ GeV$.

- **resolved 2jet–1tag (1b1j)** Events in this category are such that only the leading but not the subleading jet passes the "medium" WP (CSV > 0.8484) for all the final states
- **resolved 2jet–2tag (2b0j)** Events in this category are such that both the leading and subleading jets pass the "medium" WP for all the final states. This is the most signal-sensitive category.

2017 strategy

With respect to the 2016 data analysis, the inclusion of the VBF topology (with two additional jets inthe final state) in the 2017 the search, implies a series of problems in the identification and assignment of the b jets coming from the Higgs boson decay and the two VBF jets originating from the hard interaction.

As shown in Figure 3.18, the second jet by DeepCSV score often does not fulfill the minimal b-tag requirement (the "medium" working point described in Section 3.2.6). As a consequence, the identification of the two jets with the highest discriminant score as the b jets originating from the Higgs boson, is not always the optimal choice and, in order to minimize the jet mistagging probability, a new strategy is needed.

Given the distributions shown in Figure 3.18, we developed in 2017 a new way to properly identify and assign the jets:

1. Jets are ordered by their DeepCSV discriminator score and the one

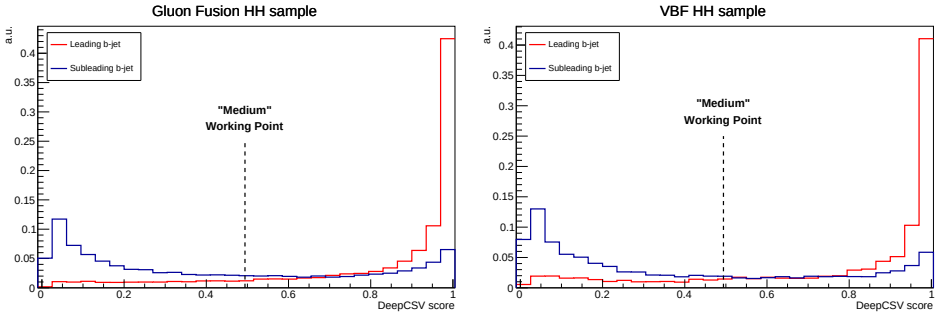


Figure 3.18: Distribution of the DeepCSV score for the first two jets ordered by DeepCSV score for a gluon fusion (left) and a VBF (right) sample.

with the highest value is chosen as the leading b-jet coming from the Higgs decay

2. The second ordered jet is selected as the subleading b-jet only if it passes the "medium" working point of the discriminator (DeepCSV $>$ 0.4941)

3. Amongst all the remaining jets with $p_T > 30\ GeV$, all possible pairs are built and the pair with the highest invariant mass (m_{jj}) is chosen as the VBF-jets candidate pair

4. If the subleading b-jet was not selected in step 2 it is now assigned:

   - If the second jet by DeepCSV score was already selected as one of the two VBF-jets, the subleading b-jet is the next jet by DeepCSV score, excluding the VBF-jets already assigned
   - If, after the assignment of the two VBF candidates, there are no more jets left to assign the subleading b-jet, the VBF pair is discarded and the jet with the second highest DeepCSV score is selected as subleading b-jet

The effect of this procedure can be appreciated in the $m_{jj}, \Delta\eta_{jj}$ distributions illustrated in Figure 3.19

Events containing a VBF jets pair and passing the selection criteria that define the "VBF tag", described in Section 3.3.3, compose the VBF category, while those failing one of the requirements are classified in the same categories that were used in the 2016 analysis: one boosted and two resolved, $1b1j$ and $2b0j$.

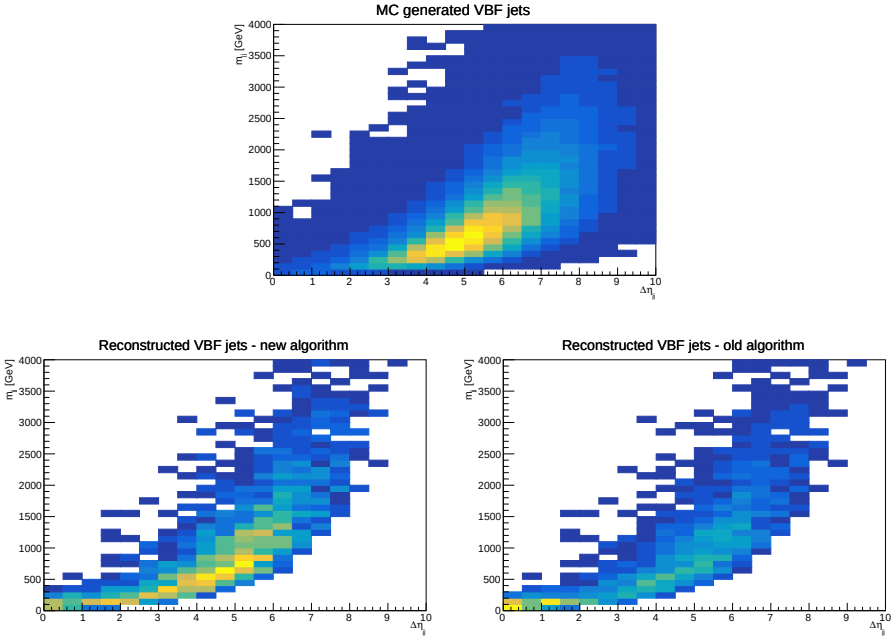


Figure 3.19: Distributions of $m_{jj}, \Delta\eta_{jj}$ for a Standard Model VBF HH signal sample. The top plot shows the distribution for Monte Carlo generated VBF jets, while the bottom left plot shows the distribution of reconstructed VBF jets identified with the algorithm described in this Section. The bottom right plot shows the distribution of reconstructed VBF jets identified as the pair with the highest invariant mass m_{jj} after assigning the b jets with the DeepCSV discriminator.

3.3.3 HH candidates

The aim of the previous steps of the analysis was to determine whether in each event suitable $H \to bb$ and $H \to \tau\tau$ candidates were present. Nevertheless, the events selected after those steps are still expected to be background dominated. The goal of this Section is to exploit the kinematics of the $HH \to bb\tau\tau$ decay to reduce the background contributions and maximize the signal purity in the event categories defined in the previous Sections, thus improving the analysis sensitivity.

Three different techniques are here described: the first two were developed in the 2016 data analysis, while the last was deployed in 2017 in order to select events where the Higgs pairs are generated through the VBF mechanism.

HH invariant mass

Since the bb and $\tau\tau$ pairs originate from Higgs bosons, in signal events the invariant mass distributions are expected to peak around 125 GeV. Thus, by applying a selection on this variable the background contamination is largely suppressed in favor of events compatible with a $HH \to bb\tau\tau$ decay.

The invariant mass of the $\tau\tau$ pair is reconstructed using the SVfit algorithm [57], a dynamical likelihood technique that quantifies the level of compatibility between a Higgs mass hypothesis and the measured momenta of the visible τ decay products plus the missing transverse energy reconstructed in the event. The kinematic properties of a $\tau \to \tau_h \nu_\tau$ decay are described by six parameters, while an additional parameter is needed to describe a $\tau \to \tau_\ell \nu_\tau \nu_\ell$ process due to the presence of a second neutrino. Only four of these variables can be measured experimentally, thus the observables do not provide sufficient information to solve for the tau pair mass $m_{\tau\tau}$ analytically. As a result, there are two or three unconstrained parameters in the decay of a $\tau\tau$ pair in the fully hadronic or semi-leptonic channels, respectively.

The SVfit algorithm computes a conditional probability $P(y|\mathbf{q}, \mathbf{x})$ using the measured lepton momenta $\mathbf{x}$ and a τ kinematic decay model that includes the measured missing transverse momentum. y is the parameter of interest and can be assumed to be either the invariant mass $m_{\tau\tau}$ or the complete four-momentum of the $\tau\tau$ pair.

The resolution on $m_{\tau\tau}$ is improved by about 30% with respect to the visible invariant mass, as illustrated in Figure 3.20, and thus allowing for a better signal to background discrimination. The agreement between data and MC simulation for the three channels in the $2b0j$ category, which is the most sensitive one, is shown in Figure 3.21 for 2016 and 2017.

The invariant mass of the Higgs decaying in a bb pair is computed from

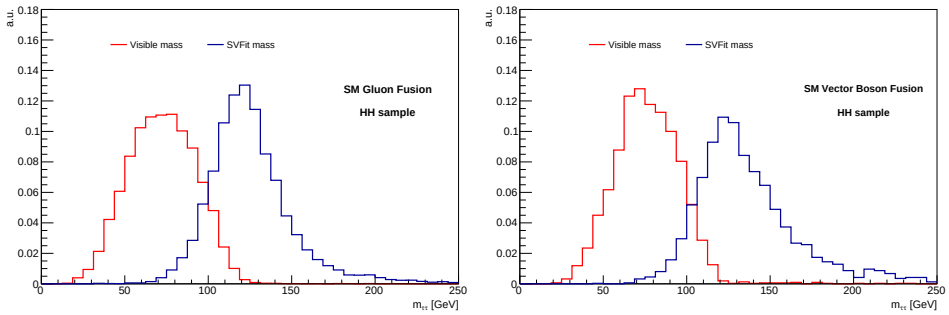


Figure 3.20: Distributions of the invariant mass of the $\tau\tau$ pair reconstructed from the visible decay products (red) and with the SVfit algorithm. The Plot on the left shows the distributions for a SM gluon fusion HH sample, the plot on the right instead shows the distributions for a SM vector boson fusion HH sample.

the invariant mass of the reconstructed AK4 jets in the resolved category, and from the mass of the AK8 jet in the boosted category. The agreement between the observed data and the MC simulation is illustrated in Figure 3.22 for the three final states studies in the $bb\tau\tau$ analysis in the $2b0j$ category.

To define the signal region, a selection on the invariant mass of the two Higgs boson candidates is applied simultaneously.

In the resolved category, the invariant mass criterion corresponds to an ellipse in the $(m_{\tau\tau}, m_{bb})$ plane defined as:

$$\frac{(\mathrm{m}_{\tau\tau} - 116\ GeV)^2}{(35\ GeV)^2} + \frac{(\mathrm{m}_{bb} - 111\ GeV)^2}{(45\ GeV)^2} < 1 \tag{3.4}$$

where 116 and 111 GeV are the expected peak positions of the reconstructed invariant masses of the two Higgs candidates, while 35 and 45 GeV have been chosen accordingly to the resolution of the distributions. These selections are optimized to give a signal efficiency around $75-80\%$, depending on the final state and on the category, and a rejection of the $t\bar{t}$ background around 85%. The $m_{\tau\tau}, m_{bb}$ distributions in the $2b0j$ category are shown, together with the elliptical selected region, in Figure 3.23 for three samples: two different double Higgs signal samples, gluon fusion and VBF, and the $t\bar{t}$ background.

Due to different event kinematics, in the boosted category a different invariant mass selection criterion is defined:

$$\begin{aligned} 80\ GeV < m_{\tau\tau} < 160\ GeV \\ 90\ GeV < m_{\mathrm{AK8}} < 160\ GeV \end{aligned} \tag{3.5}$$

The signal efficiency for the selections in the boosted category is about 85% for a background rejection of 80%. The $m_{\tau\tau}, m_{bb}$ distributions for the boosted category are shown in Figure 3.24

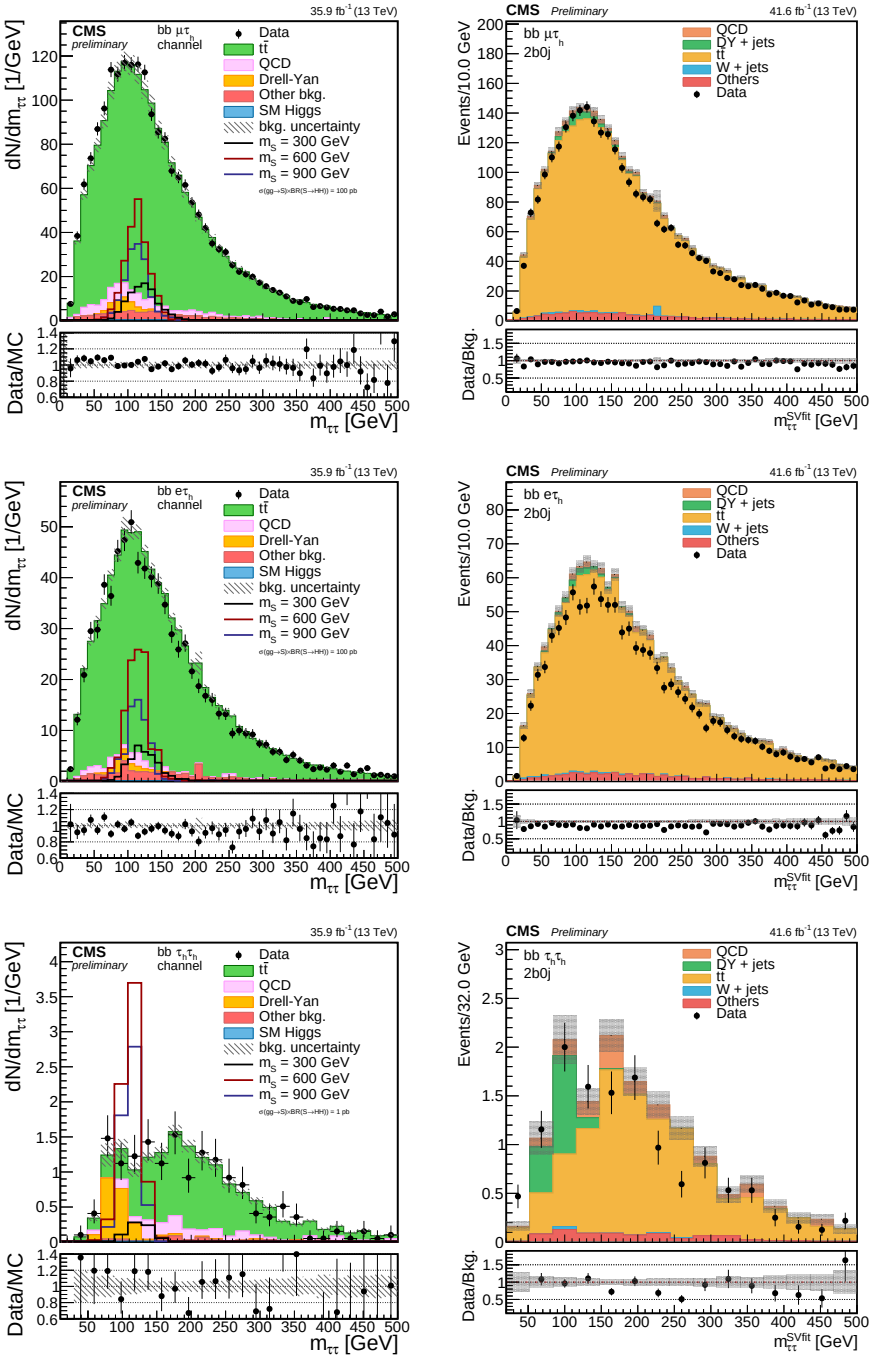


Figure 3.21: Distributions of $m_{\tau\tau}$ computed with the SVfit algorithm in the $2b0j$ category. The left column displays the 2016 data, while the right column represents the 2017 data. The three different channels investigated are shown: $\tau_\mu\tau_h$ in the top row, $\tau_e\tau_h$ in the middle row and $\tau_h\tau_h$ in the bottom row.

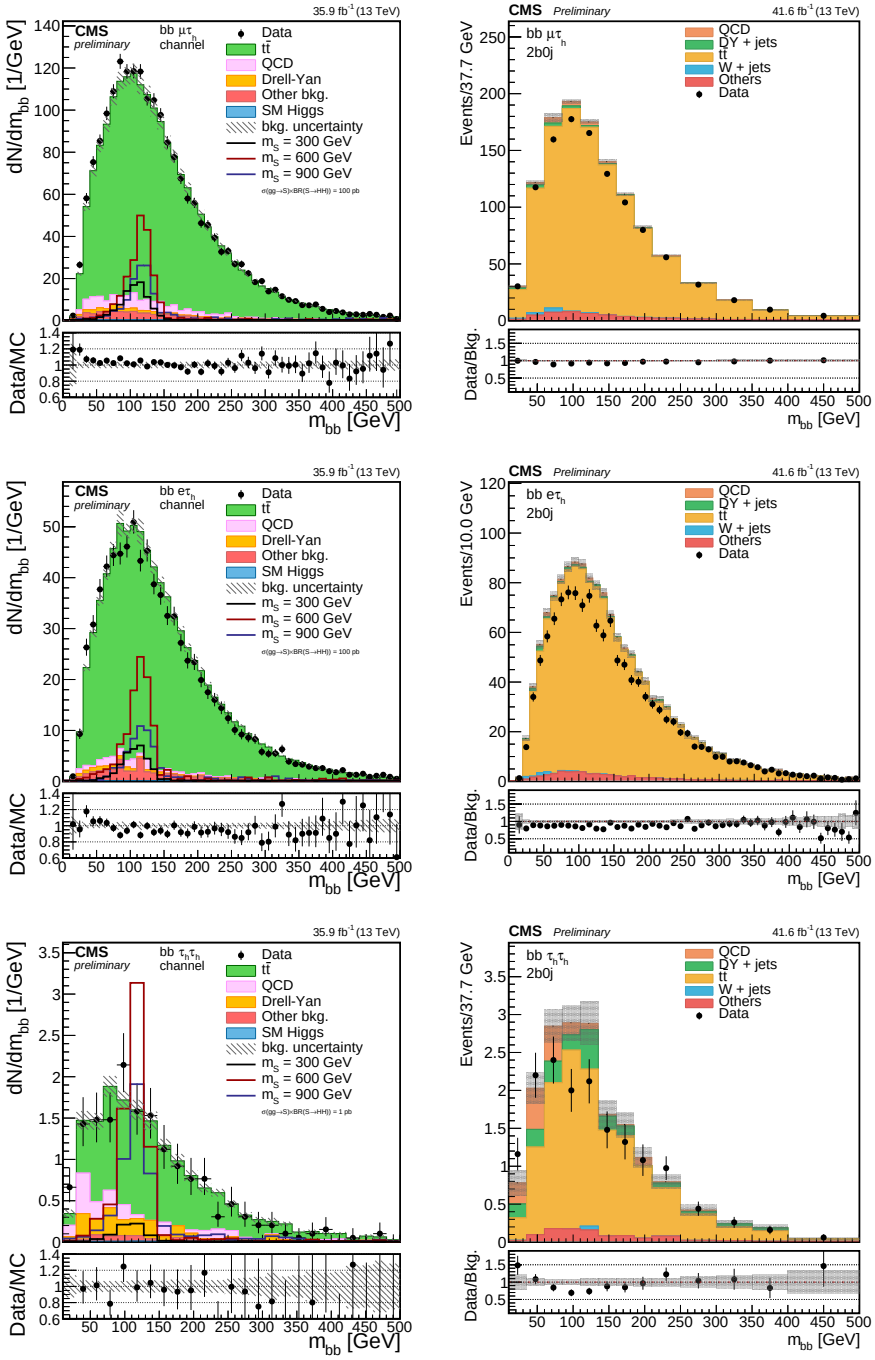


Figure 3.22: Distributions of m_{bb} in the $2b0j$ category. The left column displays the 2016 data, while the right column represents the 2017 data. The three different channels investigated are shown: $\tau_\mu\tau_h$ in the top row, $\tau_e\tau_h$ in the middle row and $\tau_h\tau_h$ in the bottom row. A slight disagreement between data and MC simultation can be observed in the right column distributions and it is related to the SF computed for the tau identification efficiency: the official recommendation from the Tau POG has not been prodived yet for 2017 data.

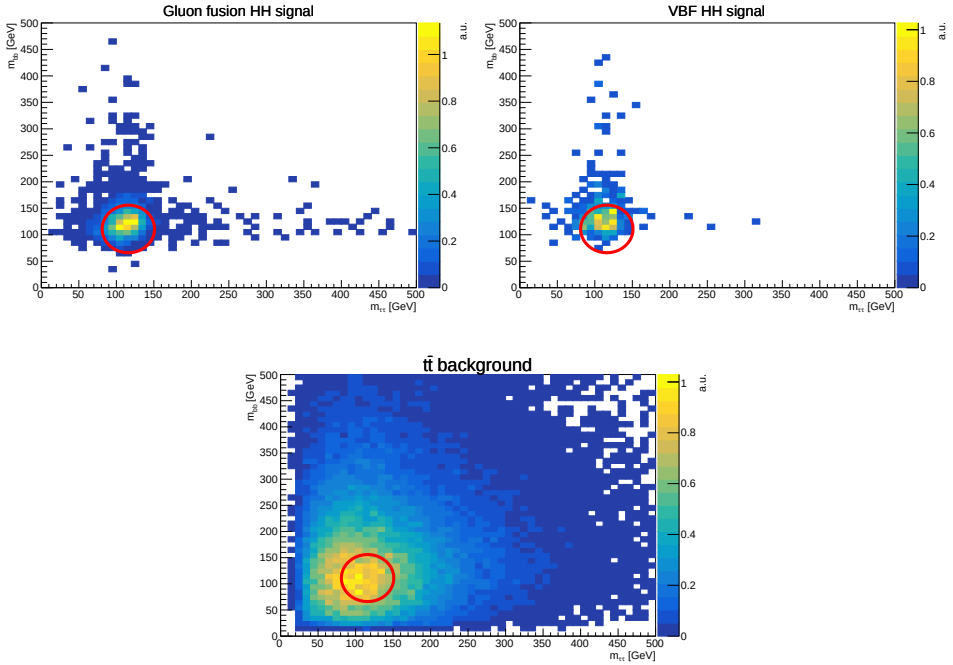


Figure 3.23: 2D distributions of $m_{\tau\tau}, m_{bb}$ for the SM gluon fusion HH signal (top left), for the SM vector boson fusion HH signal (top right), and for the $t\bar{t}$ background (bottom). The distributions are show for the $2b0j$ category only and for the combination of all three $\tau\tau$ final states. The red ellipse superimposed to the plots represents the selected area of events selected for the $HH \to bb\tau\tau$ analysis, as described in Equation 3.4.

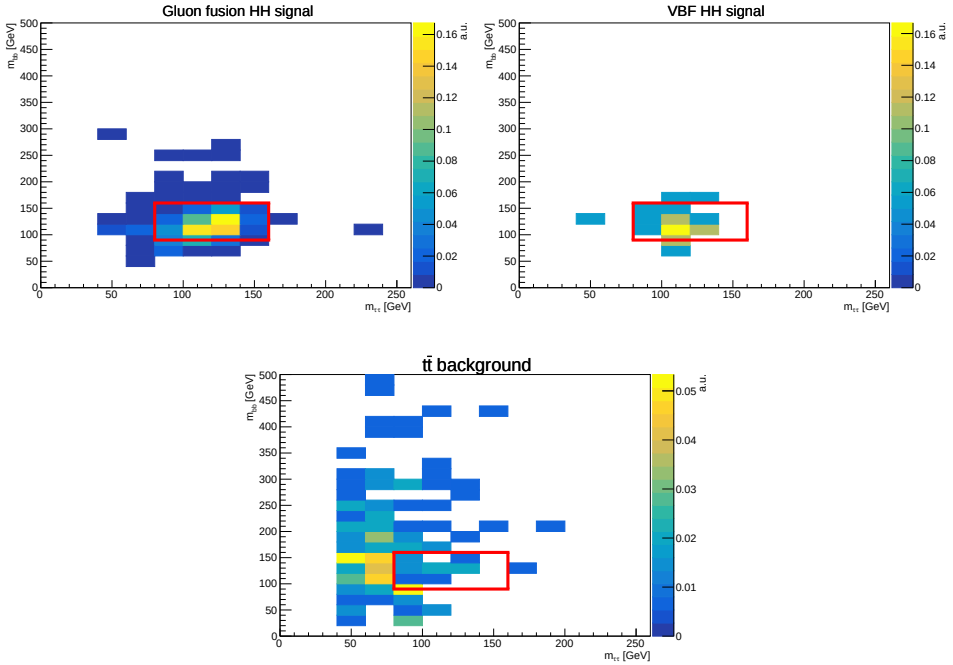


Figure 3.24: 2D distributions of $m_{\tau\tau}, m_{bb}$ for the SM gluon fusion HH signal (top left), for the SM vector boson fusion HH signal (top right), and for the $t\bar{t}$ background (bottom). The distributions are show for the *boosted* category only and for the combination of all three $\tau\tau$ final states. The red square superimposed to the plots represents the selected area of events selected for the $HH \to bb\tau\tau$ analysis, as described in Equation 3.5.

BDT for $t\bar{t}$ rejection

The contribution of different background processes in the $bb\tau\tau$ analysis has a strong dependence on the channel and category considered. Especially in the semileptonic final states, $\tau_\mu\tau_h$ and $\tau_e\tau_h$, the dominant contribution originates from $t\bar{t}$ decays due to the direct production of a prompt muon or electron. This consideration urged the design of a dedicated technique that I personally developed to suppress the $t\bar{t}$ contribution and enhance the analysis sensitivity.

The selected method is a multivariate analysis in the form of a boosted decision tree (BDT) that aims at exploiting the information from different kinematic variables by combining them and evaluating their correlations.

The most important characteristics of the BDT must be a large background rejection efficiency with a simultaneous high signal efficiency for different processes, and little correlation with the final variables used to evaluate the presence of signal events and set the exclusion limits,

In the fully hadronic channel ($\tau_h\tau_h$) the BDT discriminant is not applied due to the limited statistics available after the selections and the sizeable contributions from the QCD and Drell-Yan backgrounds, described in Chapter 4.

The Boosted Decision Tree technique combines the information of all input variables in a single value, or score (s_{BDT}), that classifies background events with low values and signal event with an high score. The TMVA toolkit [42] is used to generate a multitude (a "forest") of binary decision trees that apply a series of selections on the input variables that better separate signal and background events. The optimal variables and the corresponding threshold are selected using the Gini index $G = p(1-p)$ where p indicates the fraction of signal events correctly classified ("purity"). The number of consecutive selections applied in each tree is set to 3 in this analysis. Even if the single binary trees do not have a very large discriminating power, if a large set of trees (500 in this analysis) is created by making them aware of the events erroneously classified in the previous iteration, very good level of discrimination can be reached. This is achieved by assigning the erroneously identified events a larger weight and estimating their rate through the minimization of a loss function.

Training events are selected applying all the identification selections described in Section 3.2 as well as the invariant mass cuts, but with no requirement on the b tag discriminators for the jets, in order to maintain enough statistics. To minimize statistical fluctuations, that can interfere with the "learning" process of the BDT, events from the $\tau_\mu\tau_h$ and $\tau_e\tau_h$ channels can be combined since the distributions of both final states (shown in Figures 3.29 and 3.30) have been observed to be the same.

Finally, to further increase the number of events and to ensure a good coverage of all possible signal processes, the signal samples are divided in two

categories and two separate BDT trainings are performed for the resonant search. Resonances with $m_X < 350\ GeV$ compose the Low Mass (LM) topology, while those with masses $350 < m_X < 900\ GeV$ are part of the High Mass (HM) one. The mass separation value of the two regimes is chosen in correspondence of the mass of a top quark pair as it represents the value that guarantees the most similarity of the BDT input variables for the different signal hypotheses inside the two regimes.

All the signal samples that are input to the BDT are normalized according to the Radion cross section and the branching fraction of its decay to a HH pair. $\sigma^{Rad} \times \mathcal{B}(Rad \to HH)$ decreases with the increase of the resonance mass, thus a larger importance is assigned to the mass values close to the m_{HH} kinematic threshold. Other normalization methods have been tested and found to bring an insufficient gain at high masses for a too large loss in performance at low masses.

The separation of the BDT in two regimes represents a compromise between the complexity and the performance of the search. Even though the ideal situation would be to have a specific training for each signal, in case of the $bb\tau\tau$ analysis, where many possible signals are tested, this would add a sizeable amount of complexity to the analysis itself and would suffer from the limited statistics available in the single samples. Furthermore, the variety of inputs used in the trainings allows to have two BDT discriminants that are reasonably efficient across all the mass range studied without being hyper-optimized for a single m_X value but sub-optimal for the others.

When optimizing a BDT, one of the most concerning issues is the so-called "overtraining", that is the individuation of statistical fluctuation as discriminating features typical either of the signal or of the background events. In order to minimize this effect one has to rely on as much statistics as possible and a fine tuning of the BDT parameters is needed. To check for the presence of overtraining in the developed BDT, the input samples are split in a "training" and "testing" subsamples: the former actually serves the purpose of training the BDT, while the latter is used to compare the BDT output distribution. In addition, different algorithms provided by the TMVA package have tested and the gradient boost algorithm proved to be the most robust against overtraining. The comparison of the training and testing outputs is reported in Figure 3.25, where a good agreement is observed.

The Low Mass training is also found to be very effective in the non resonant search. Thus, a single LM BDT is trained and the applied to the resonant and non-resonant cases with a a different selection on the BDT output, as described later in this Section.

When optimizing a BDT, one of the most important parameters is the choice of the input variables to the training. The choice is restricted to those observables that provide the best separation between HH and $t\bar{t}$ events according to the kinematic differences of the two processes.

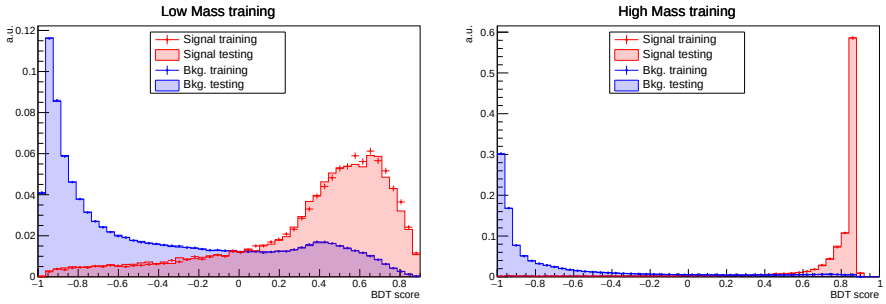


Figure 3.25: Output distributions for the Low Mass (left) and High Mass (right) BDT. No overtraining is observed as the training and testing samples are compared and found to be in good agreement.

The transverse momentum distributions are potential good candidates, but they are largely correlated with the final discriminant observables $MT2$ and m_{HH}^{KinFit}, have a strong dependence on the signal hypothesis and are affected by higher order effects that are not always properly modeled in the MC simulation, thus are not selected for the BDT training. On the other hand, the topologies of HH and $t\bar{t}$ events present multiple differences, as illustrated in Figure 3.26. In signal events the Higgs bosons tend to be produced back-to-back and their decay products, either the bb or the $\tau\tau$ pair, tend to be emitted, in the transverse plane, in opposite hemispheres of the detector. Furthermore, the missing transverse energy is expected to originate mainly from the neutrinos involved in the τ decays, thus its separation from the $H \to \tau\tau$ candidate, and it's decay products, is usually small and lays in the same detector hemisphere. Conversely, in $t\bar{t}$ events, the decay of top quarks produces τb pairs that are randomly distributed in the transverse plane, as is the MET that originates from the presence of $W \to \ell\nu_\ell$ decays. Another variable strictly related to these observation is the "transverse mass", m_T, of the lepton candidates that is defined as:

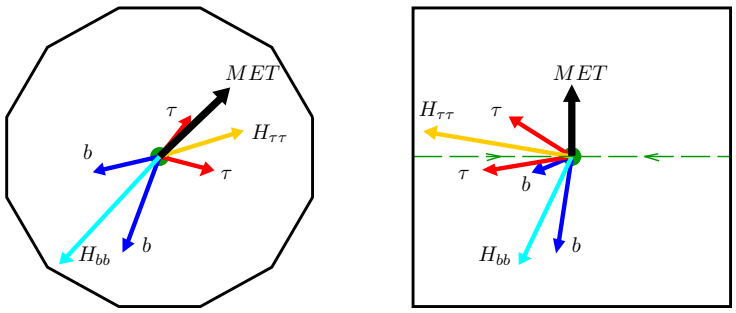
$$m_T(\ell) = \sqrt{2p_T^{miss}p_T^{\ell}\,(1-\Delta\varphi)} \tag{3.6}$$

where $\Delta\varphi$ describes the angular separation in the transverse plane between the lepton and the missing transverse momentum.

Two final sets, with eight variables each, have been chosen as input to the Low Mass and High Mass trainings and are reported in Table 3.4.

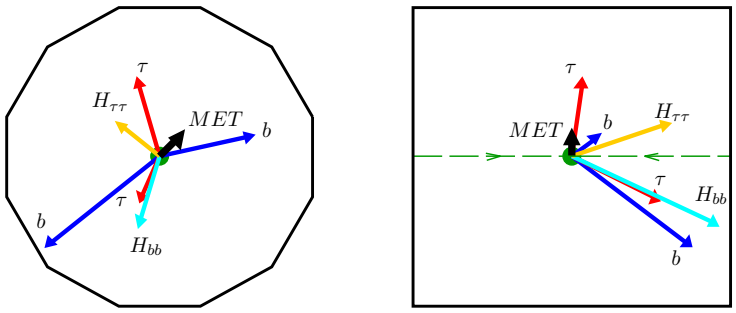
Different sets of variables have been studied and recursively pruned in order to maintain only the subset that leads to the best BDT performance, as illustrated in Figure 3.27.

The linear correlations between the input variables is shown in Figure 3.28, while the agreement between the observed data and the MC simu-



(x, y) plane (z, y) plane

(a) HH event



(x, y) plane (z, y) plane

(b) $t\bar{t}$ event

Figure 3.26: A schematic representation of a SM HH (a) and of a $t\bar{t}$ (b) event.

lation is shown in Figure 3.29 and Figure 3.30 for the $\tau_\mu\tau_h$ and $\tau_e\tau_h$, respectively.

The BDT output distributions for the $\tau_\mu\tau_h$ and $\tau_e\tau_h$ channels are shown in Figure 3.31 for both the LM and HM BDT trainings.

The performances of the two BDT trainings are plotted in Figure 3.32 as function of the signal efficiency and expected $t\bar{t}$ background rejection (ROC curves). ROC curves are reported, for resonant signals with masses between 250 GeV and 900 GeV, as function of the the selection applied to the BDT output for both the LM and HM trainings. The performances of the BDT trainings are also compared to the method adopted in the Run I $bb\tau\tau$ analysis to reject the $t\bar{t}$ background, that is to require the transverse mass (m_T) of the event to be smaller than a certain threshold, set to $m_T < 30$ GeV. It can be observed that the BDT behavior reflects exactly the division point ($m_X = 350$ GeV) chosen for the two training regimes as the LM ROCs present, for low mass samples, an higher efficiency than the m_T cut, while

NR and LM inputs	HM inputs
$\Delta\varphi(H_{bb}, H_{\tau\tau})$	$\Delta\varphi(H_{bb}, H_{\tau\tau})$
$\Delta\varphi(H_{\tau\tau}, p_T^{miss})$	$\Delta\varphi(H_{\tau\tau}, p_T^{miss})$
$\Delta\varphi(H_{bb}, p_T^{miss})$	$\Delta\varphi(H_{bb}, p_T^{miss})$
$\Delta\varphi(\ell, p_T^{miss})$	$\Delta\varphi(\ell, p_T^{miss})$
$m_T(\ell)$	$m_T(\ell)$
$m_T(\tau_h)$	$m_T(\tau_h)$
$\Delta R(b,b) \cdot p_T(H_{bb})$	$\Delta R(b,b)$
$\Delta R(\ell,\tau_h) \cdot p_T(H_{\tau\tau})$	$\Delta R(\ell,\tau_h)$

Table 3.4: Input variables of the BDT discriminant. Left column: inputs for the non-resonant (NR) and low-mass (LM, $m_X \leq 350\ GeV$) training. Right column: inputs for the high-mass (HM, $m_X > 350\ GeV$) training. $H_{\tau\tau}$ and H_{bb} denote the Higgs boson candidates reconstructed with the SVFit algorithm in the first case, and as invariant mass of the two selected jets in the second case. ℓ represents either the electron or the muon, while τ_h is the hadronically decaying tau.

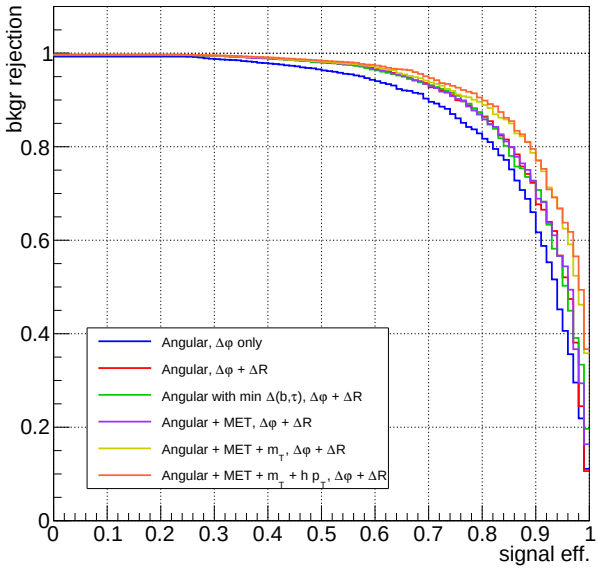


Figure 3.27: Signal efficiency versus background rejection for different sets of variables used as input to the BDT.

they are sub-optimal for high mass signals. On the other hand, the HM ROCs behave excellently for high mass resonances and show a less efficient

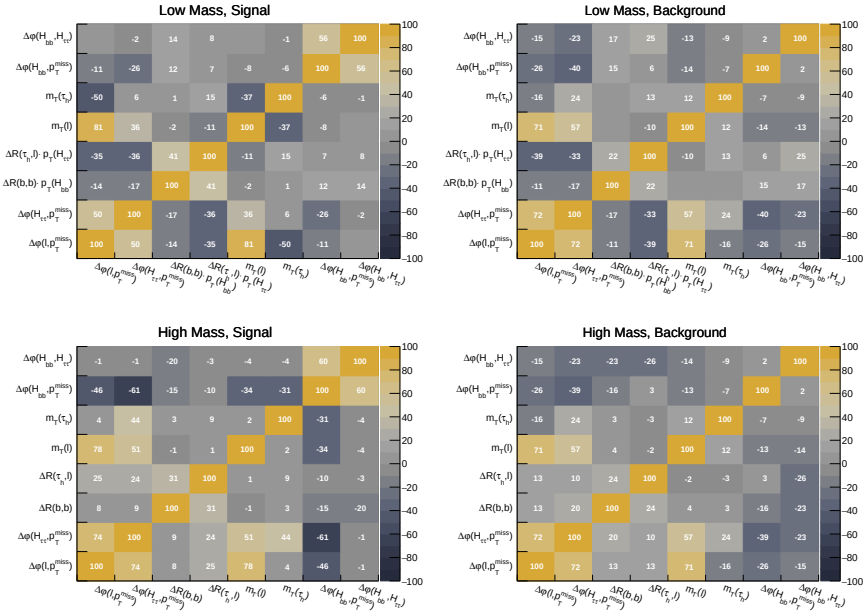


Figure 3.28: Linear correlation coefficients of the BDT input variables for the low mass training (upper row) and high mass training (lower row) for signal sample (left) and background sample (right).

performance for smaller values of m_X. Figure 3.32 also reports the behavior of the LM BDT training in presence of non-resonant samples with different values of the Higgs self coupling parameter modifier k_λ. The performance is found to be compatible with the cut on the transverse mass for all the possible signals investigated in the analysis.

Using events selected in $\tau_\mu\tau_h$ channel from the $2b0j$ category as benchmark, different selections on the BDT output have been tested and compared in the LM and HM training in order to select the working point that lead to the highest analysis sensitivity. The gain achieved with the usage of the BDT discriminant is estimated from the exclusion upper limits on $\sigma \times \mathcal{B}$ for the $HH \to bb\tau\tau$ process: a lower limits implies that a larger region of the parameter space is excluded, thus meaning an increase of the analysis sensitivity. The related plots are illustrated in Figure 3.33, where it is clear that the optimal working point for resonant signals, both in the Low and High Mass regimes, are obtained selecting events with $s_{BDT} > 0.477$ (LM) or $s_{BDT} > 0.0188$ (HM), both WPs corresponding to about 90% $t\bar{t}$ rejection and a signal efficiency that ranges between 65 and 95% depending on the resonance mass. In presence of non-resonant signals instead, the maximum sensitivity is reached by applying on the BDT score a selection ($s_{BDT} > 0.0764$) on the LM training, that corresponds to a 70% $t\bar{t}$ rejection for a signal efficiency of about 80%,

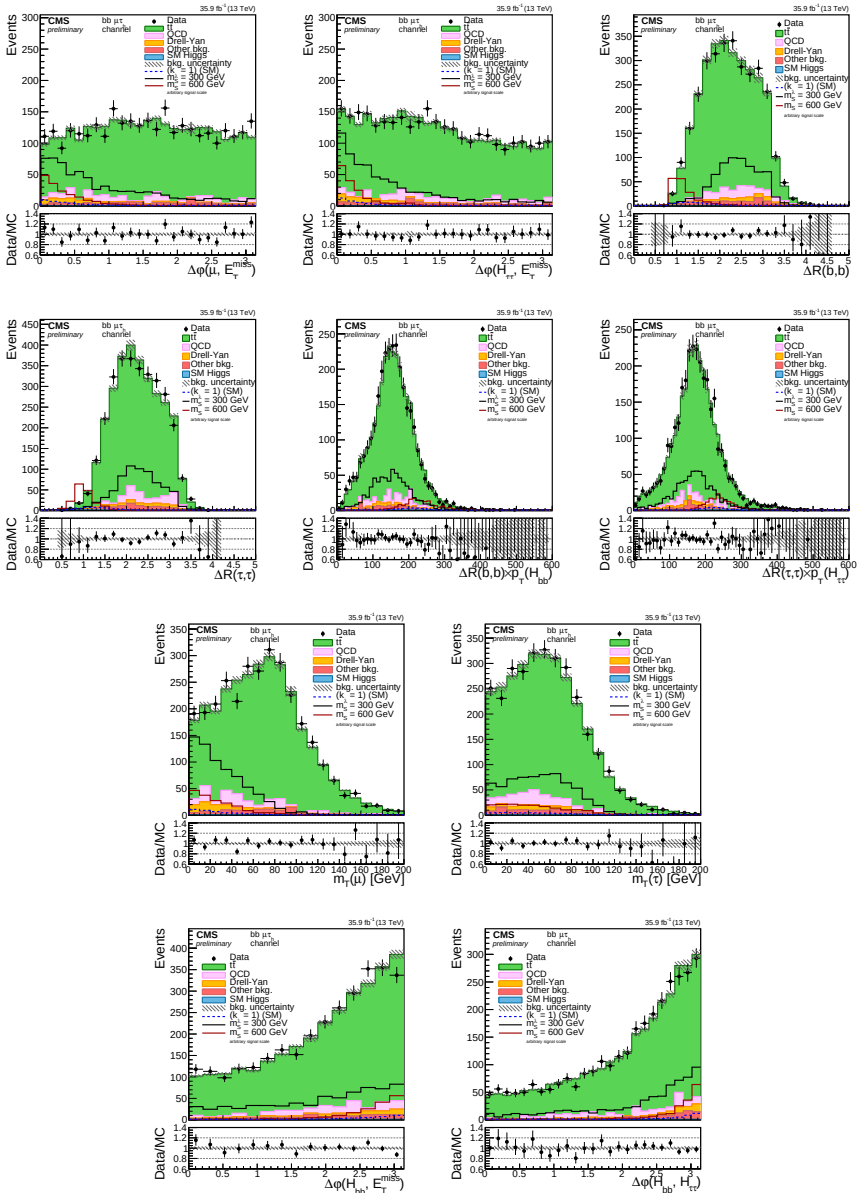


Figure 3.29: Distributions of BDT input variables in the $\tau_\mu\tau_h$ channel after the $\tau\tau$ and bb candidates selections after the invariant mass requirements.

The working points thus selected are compared for the resonant and non-resonant samples in Figure 3.34. In the resonant search (left plot in Figure 3.34), a clear transition between the LM and HM trainings is observed around 350 GeV, we therefore decided to apply the selection $s_{BDT} > 0.477$ for signals with $m_X \leq 350\ GeV$ and $s_{BDT} > 0.0188$ for signals with $m_X >$

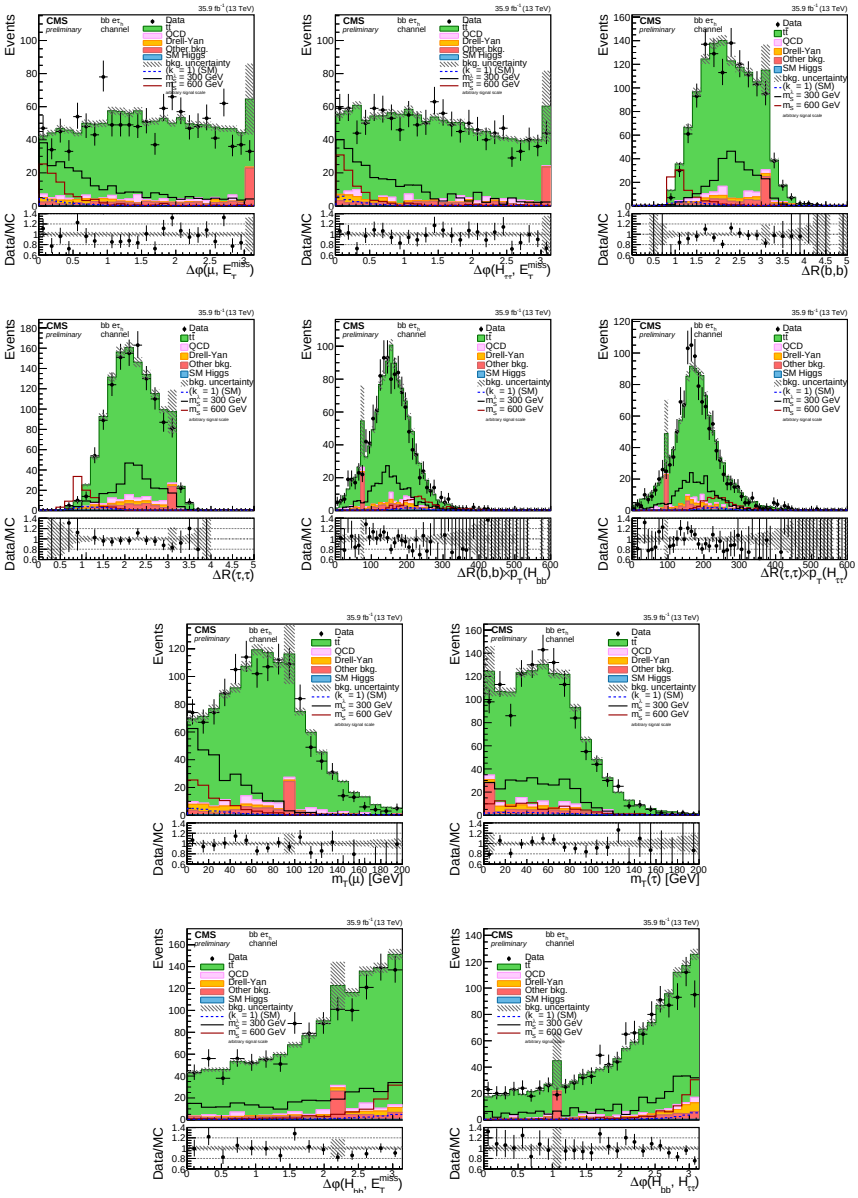


Figure 3.30: Distributions of BDT input variables in the $\tau_e\tau_h$ channel after the $\tau\tau$ and bb candidates selections after the invariant mass requirements.

350 GeV in order to obtain a high efficiency over the whole m_X spectrum. For the non-resonant case, the selection $s_{BDT} > 0.0764$ ensures a better performance with respect to what was achieved in previous $bb\tau\tau$ searches.

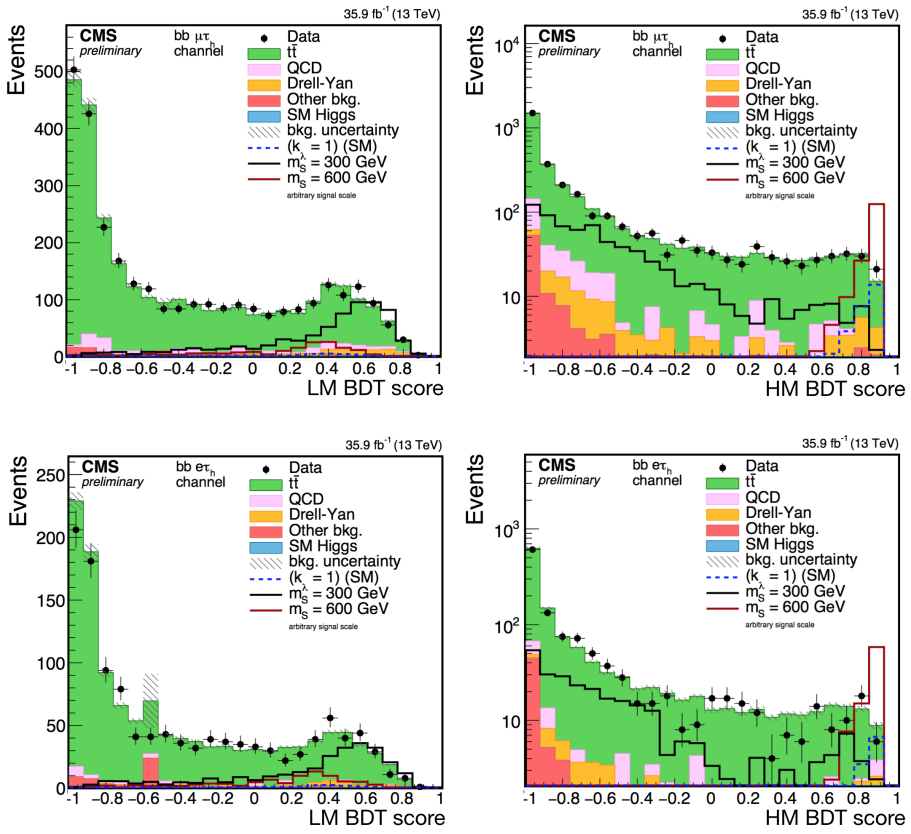


Figure 3.31: BDT output distribution for the $\tau_\mu\tau_h$ final state (upper row) and the $\tau_e\tau_h$ final state (bottom row). The LM BDT output is shown on the left, while the HM one is shown on the right.

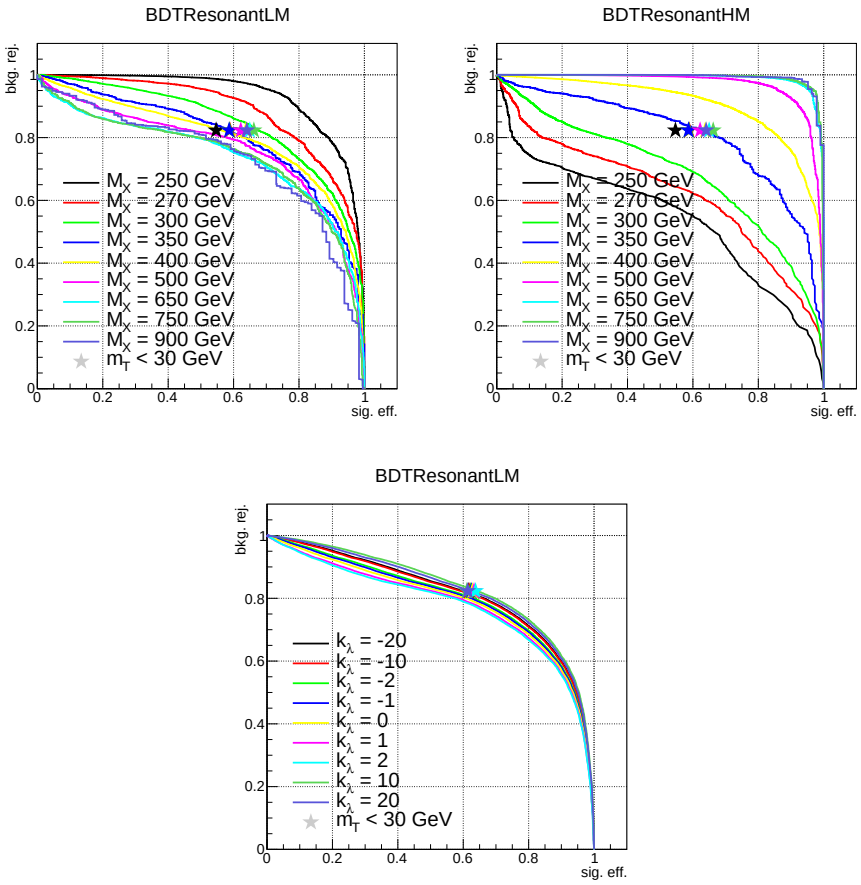


Figure 3.32: $t\bar{t}$ rejection as a function of signal efficiency for resonant signals (LM and HM trainings, top row) and for non-resonant signals (LM training, bottom row). The star marker denotes the performance of a selection $m_T < 30\ GeV$.

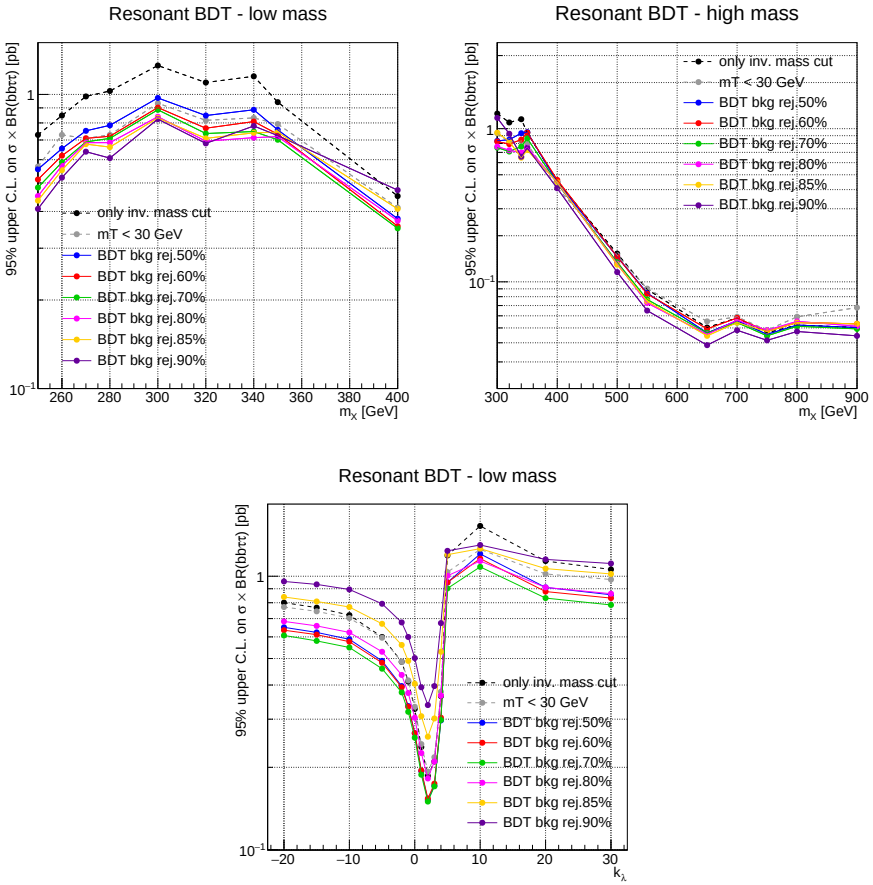


Figure 3.33: Comparison of the expected sensitivity in the $\tau_\mu\tau_h$ channel $2b0j$ category for differents working points of the BDT. In the top row, the LM and HM BDTs applied to resonant samples is shown, while in the bottom row the LM BDT is applied to the non-resonant signals.

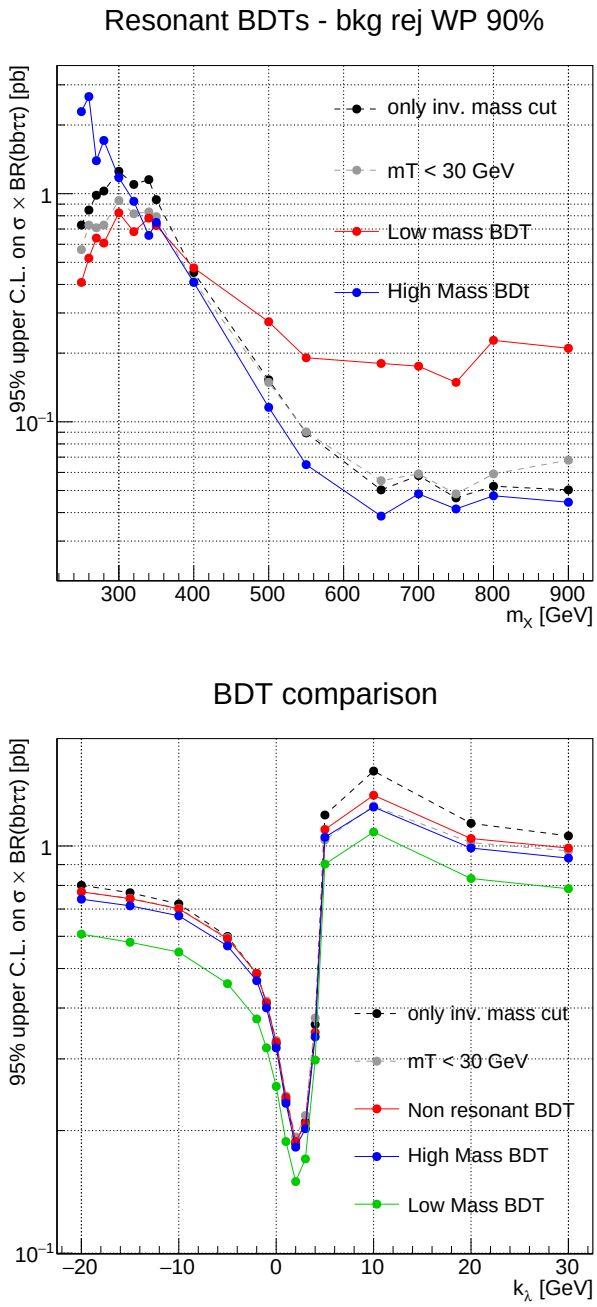


Figure 3.34: Comparison of the expected sensitivity in the $\tau_\mu\tau_h$ channel $2b0j$ category for differents working points of the BDT. In the top row, the LM and HM BDTs applied to resonant samples is shown, while in the bottom row the LM BDT is applied to the non-resonant signals.

VBF tagging

In order to properly study the double Higgs Vector Boson Fusion production mechanism, the definition of specific selection is of primary importance in order to be able to identify the VBF events. The task is not easy and the performance is mainly limited by three factors:

1. The final state of VBF events contains two more objects, that originate from the signal event, with respect to gluon fusion process. This increases the probability of misidentification or mis-assignment of all the objects in the event.

2. Jets originating from a VBF process have particular kinematic properties that, due to experimental limitations, may reduce the signal acceptance.

3. As reported in Section 1.3 of Chapter 1, the cross section of HH VBF production is very small ($\sigma_{HH}^{VBF} \simeq 1.62\ fb$), even when compared to the gluon fusion cross section: $\sigma_{HH}^{VBF}/\sigma_{HH}^{GF} \sim 1/20$).

In order to identify and properly assign the jets reconstructed in the event to those actually originating in the hard scatter interactions (b and VBF jets) a new algorithm has been developed and is descibed in Section 3.3.2. This procedure increase the probability to correctly identify the four jets by $\sim 10\%$.

In VBF events, two quarks, one for each of the colliding protons, undergo an hard scattering interaction that generates a Higgs pair through the fusion of two vector bosons and shifts the initial quarks between the final state objects. As already discussed in Sections 1.1 and 2.3.4, due to the QCD confinement properties, quarks can not exist as free states, but they hadronize giving life to short lived particles, which in turn decay in "showers" of lighter leptons or hadrons. Despite the complexity of the hadronization process, most of the times the final state jets maintain the same kinematic properties of of their parent quarks.

Figure 3.35 illustrates the main kinematic properties of the VBF jets generated in a double Higgs VBF MC signal sample.

In order to properly reconstruct the jets, High Energy Physiscs experiments apply some minimal selections on the objects, either due to detectors limitations, such as the pseudorapidity coverage, or because quality cuts are needed to guarantee an high efficiency of reconstruction and identification. In CMS the main selections are:

$$\begin{aligned} Jet \quad & p_T > 20\ GeV \\ Jet \quad & |\eta| < 5 \end{aligned} \tag{3.7}$$

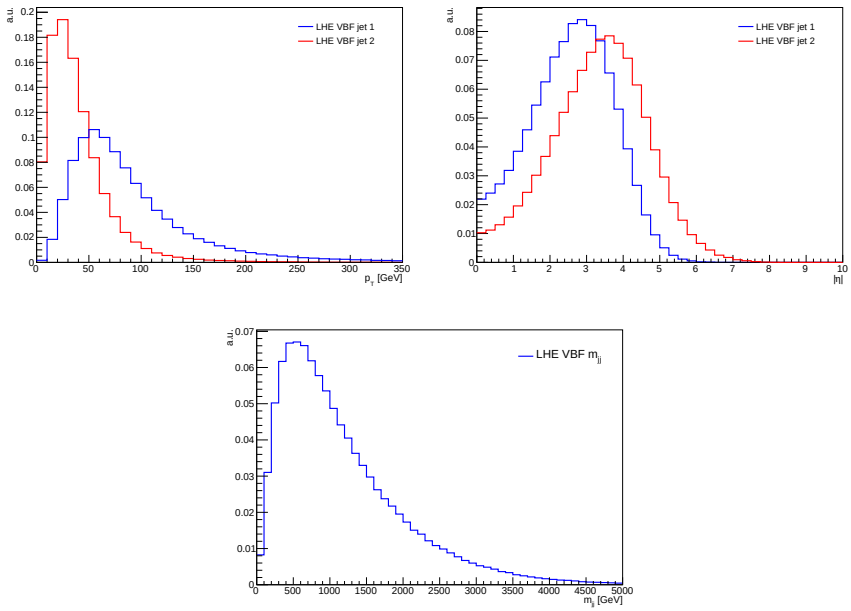


Figure 3.35: Main kinematic properties of the VBF jets in a MC simulated HH sample. The top row illustrates the transverse momentum (left) and the pseudorapiduty (right) of the two jets, while the bottom row displays their invariant mass.

which, given the distributions observed in Figure 3.35, already imply an unavoidable reduction of signal acceptance of about 30%. Despite this experimental drawback, VBF jets maintain a set of specific characteristics that may be helpful in the discrimination against other processes. For example, as seen from the long tail in the invariant mass distribution in Figure 3.35, VBF jets are usually very energetic, with almost half of the events having an invariant mass m_{jj} larger than 1 TeV.

The spatial distributions of the two VBF jets are shown in Figure 3.36. Having origin in a scattering process that involves the two colliding protons, the VBF jets are usually characterized by a large spatial separation along the beam axis direction: more than 70% of the events contain a jet pair with $|\Delta\eta_{jj}| > 5$, which is more or less equivalent to the spatial coverage of the tracking system in the CMS detector. However, this doesn't mean that the jets are necessarily produced both "forward", *i.e.* one in each of the two endcap regions of the detector. As a matter of fact, in around 45% of the events that have $|\Delta\eta_{jj}| > 5$, at least one of the two jets is produced in the barrel ($|\eta_j| < 2.5$). Moreover, some additional information can be inferred from the pseudorapidity positions. As displayed in the right plot of Figure 3.36, the product of the pseudorapidity values has a peak around $\eta_{j_1} \cdot \eta_{j_2} \simeq -6$ and a large left tail: around 95% of the events lay one the

negative semi-axis. This means that, despite one of the two jet has an high probability of being produced in the "central" region, the two jets almost certainly lay on two different sides of the interaction point (that in CMS is taken as the origin of the coordinate system).

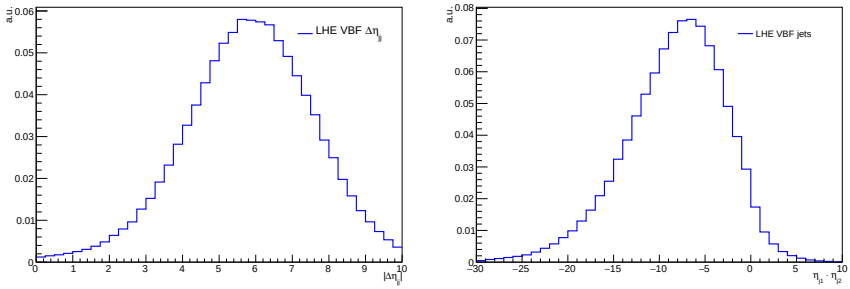


Figure 3.36: Absolute difference in pseudorapidity for the two VBF jets (left) and product of the two pseudorapidity values (right).

Based on this observations, a two step procedure is proposed to identify the VBF events: a selection on the most discriminating variables and the development of a BDT discriminator. The former is used to reject the background contributions, while the latter is exploited to distinguish VBF from gluon fusion events.

As already discussed, the most characteristic features of the VBF jets are their invariant mass and the relative pseudorapidity distance, Figure 3.37 illustrates the m_{jj},$\Delta\eta_{jj}$ distributions for the SM VBF HH signal and for some of the main backgrounds, including the gluon fusion HH sample, that in the case of VBF searches represents a background itself.

From the plots in Figure 3.37 it is evident that a simple rectangular selection should be able to suppress most of the background contributions without reducing too much the signal acceptance. Table 3.5 reports, as an example, the percentage of events for each process, that fall in the "VBF region".

As already mentioned in Section 1.3, in order to obtain a valid estimate of the Higgs self interaction λ_{HHH} in HH analyses, it is necessary to disentangle its effect from the other Higgs boson couplings that enter double Higgs diagrams. In the context of gluon fusion searches this is accomplished, in the SM case, by setting exclusion limits as a function of the ratio between λ_{HHH} and the Higgs-top quark coupling y_t, and in BSM searches by exploring different EFT benchmarks, as detailed in Section 5.4. Thus, when considering the VBF production mode, which is already complicated by the effects of Higgs-Vector boson couplings, it is of the uttermost importance to minimize the contamination of gluon fusion events in the "VBF region".

Experimentally, the double Higgs gluon fusion production cross section

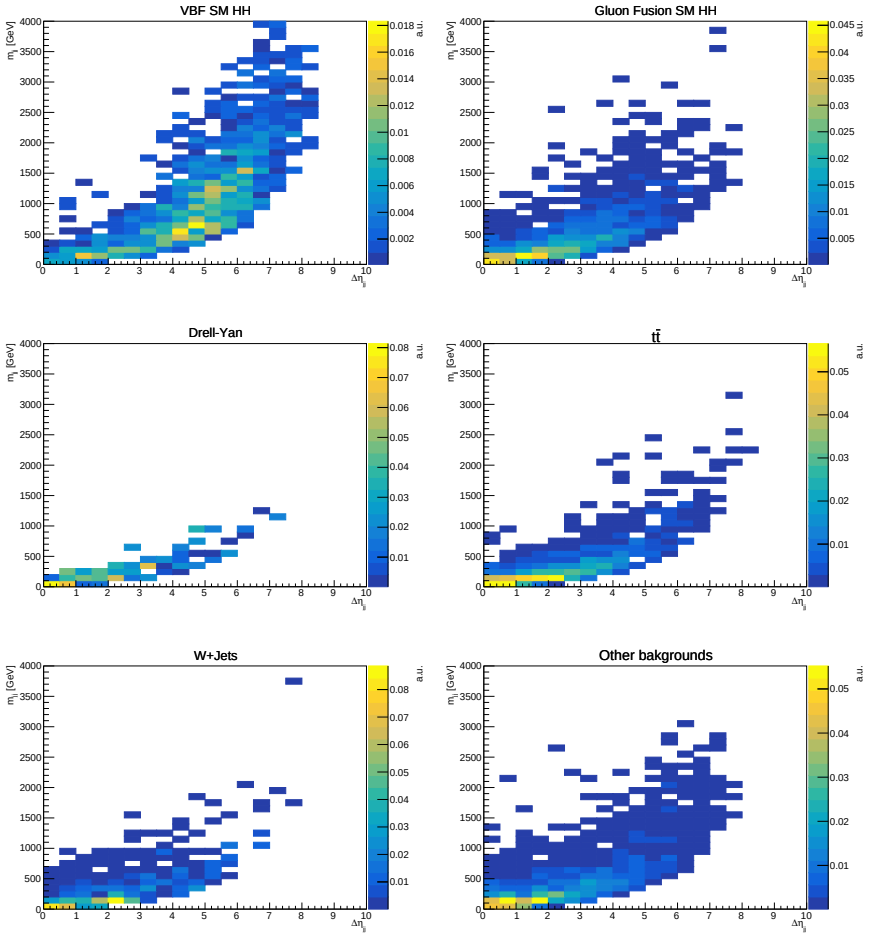


Figure 3.37: The first row shows the m_{jj},$\Delta\eta_{jj}$ distributions for HH signals in the VBF (left) and gluon fusion (right) production modes. The central and bottom rows report the contributions from DY, $t\bar{t}$, $W + jets$ and the sum of the remaining backgrounds.

is about twenty times larger than the vector boson fusion one and, moreover, the two processes lead to the same final state objects. This means that, even after the selections applied to identify the $H \to bb$ and $H \to \tau\tau$ candidates, and after the "VBF region" requirements on the invariant mass and the pseudorapidity separation of the jets, a large contamination of gluon fusion events still spoils the VBF measurement.

A multivariate approach is adopted with the development of a BDT capable of discriminating gluon fusion and VBF events. Boosted decision trees techniques have been already widely discussed in this Section and in Section 5.1.3 of Chapter 5, thus here I will expose only the main features studied to characterize this BDT.

% of events in "VBF region" ($m_{jj} > X\ GeV$ and $\Delta\eta_{jj} > Y$)			
Sample	$X = 200, Y = 2$	$X = 300, Y = 3$	$X = 500, Y = 4$
VBF HH	88.5%	81.7%	68.5%
Gluon Fusion HH	54.3%	35.7%	19.0%
DY	47.5%	33.6%	16.6%
$t\bar{t}$	46.9%	30.4%	15.4%
W+jets	39.8%	27.4%	13.3%
Oth. Bkgs.	47.7%	32.1%	17.8%

Table 3.5: Percentage of events surviving the "VBF region" selections.

Most of the inputs variables selected as input to the VBF BDT, and reported in Table 3.6 together with the ranking assigned during the training, are related to the spatial separation of different combination of objects. As a matter of fact, the presence of two additional jets in the final state affects the event kinematic and the relative positions of the reconstructed Higgs candidates, of the τ leptons and of the b jets. Moreover, to further exploit the VBF jets properties discussed earlier in this Section, a new variable, denoted "boson centrality" (ζ_V), is added to the BDT inputs and defined as:

$$\zeta_V = \min[\Delta\eta_-,\ \Delta\eta_+] \tag{3.8}$$

where

$$\begin{aligned} \Delta\eta_- &= \min[\eta_{H_{\tau\tau}},\ \eta_{H_{bb}}] - \min[\eta_{VBFjet_1},\ \eta_{VBFjet_2}] \\ \Delta\eta_+ &= \max[\eta_{VBFjet_1},\ \eta_{VBFjet_2}] - \max[\eta_{H_{\tau\tau}},\ \eta_{H_{bb}}] \end{aligned} \tag{3.9}$$

The variable ζ_V is a topological variable that tends towards large positive values when the VBF jets have a large separation in η and both Higgs boson candidates are in the pseudorapidity gap between the VBF jets.

The BDT training is performed using SM VBF events as signal, and SM gluon fusion events as background, selected in all three $\tau\tau$ final states and without any specific requirement on the b tagging of the jets. A loose "VBF region" selection is applied to the events before the training by requiring $\Delta\eta_{jj} > 1$ and $m_{jj} > 600\ GeV$.

The BDT training parameters are reported in Table 3.6, while the correlation matrices and the overtrain check plot are illustrated in Figure 3.38.

A more detailed study of the VBF selections and BDT performances is reported in Section 5.5 of Chapter 5.

Ranking	VBF BDT inputs
1	$p_T(VBFjet_1)$
2	$p_T(H_{bb})$
3	$\eta_{bjet_1} \cdot \eta_{bjet_2}$
4	$\Delta R(\tau_1, \tau_2)$
5	$\Delta R(H_{bb}, H_{\tau\tau})$
6	$\zeta_V(H_{bb}, H_{\tau\tau})$
7	$\eta_{VBFjet_1} \cdot \eta_{VBFjet_2}$
8	m_{jj}

Training Parameter	Value
NTrees	300
MaxDepth	2
MinNodeSize	0.15
nCuts	500
Shrinkage	0.04
BaggedSampleFraction	0.5

Table 3.6: Input variables to the VBF BDT discriminant (left). $H_{\tau\tau}$ and H_{bb} denote the Higgs boson candidates reconstructed with the SVFit algorithm in the first case, and as invariant mass of the two selected jets in the second case. Selected values of the parameter for the VBF BDT training (right).

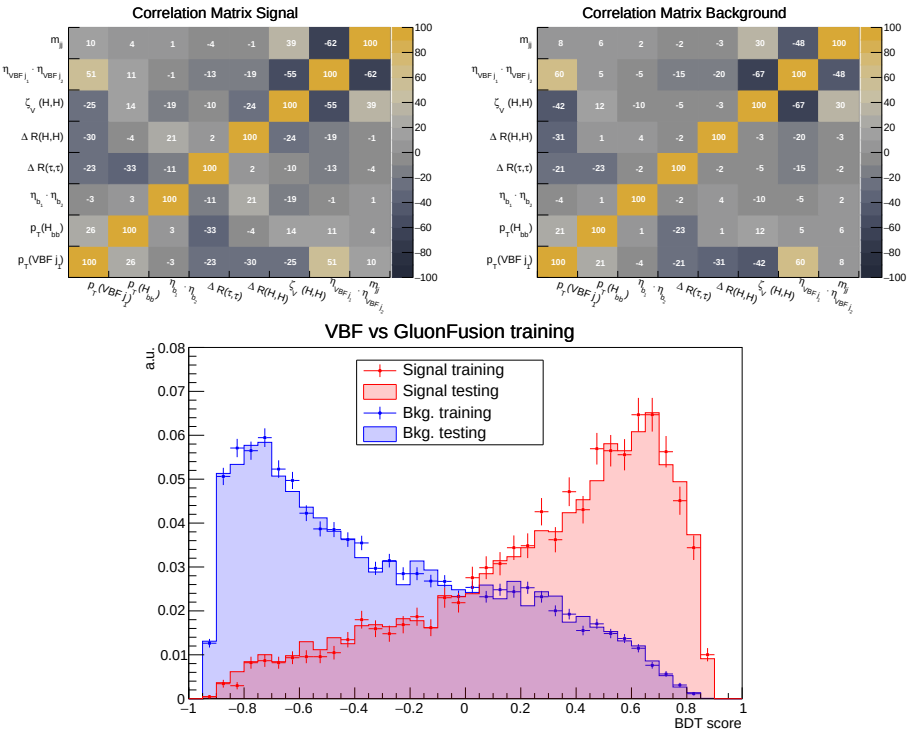


Figure 3.38: The top row shows the correlation matrices of the input variable for signal (left) and background (right). The plot in the bottom row is the overtrain check: the performances of the BDT on the training and testing samples is in good agreement, thus no overtrain is observed.

Chapter 4

Monte Carlo simulation

Since multiple sources of background affect the $bb\tau\tau$ analysis and different signal topologies (resonant and non-resonant) are explored in the search, an accurate modeling of all processes involved is crucial in order to optimize the analysis strategy and techniques, and to obtain a valid comparison between the observed data and the theoretical predictions.

The backgrounds can be classified in two main categories, either as "irreducible" or as "reducible" contributions. The former are composed of processes that lead to the exact same final state as in the $HH \to bb\tau\tau$ decay, that is the object of the search described in this thesis. The two most important contributions in this category originate from the $t\bar{t} \to bbWW \to bb\ell\nu_\ell\tau\nu_\tau$ decay and from the Drell-Yan production of a tau pair in association to a b quark pair. On the other hand, reducible backgrounds arise from the misidentification of objects due to experimental detector effects, the most striking case being the erroneous identification of gluon or light quark initiated jets with wither a τ_h candidate or as a b jet. The perfect example of reducible background is the QCD multi-jet contributions, especially relevant in the $\tau_h\tau_h$ channel.

In order to handle these two categories of background sources, different strategies are exploited. The reducible contributions are suppressed through the application of tight quality selection requirements that aim at guaranteeing an high background rejection efficiency, but, as a consequence, often result in a loss in signal acceptance: the optimal working point is the balanced trade off between the two effects. Irreducible background sources can instead only be tackled by exploiting the kinematic differences with respect to HH signal events. These strategies put in place in order to reject background events are thoroughly described in Chapter 3: Section 3.2 and Section 3.3 for the reducible and irreducible contributions, respectively.

Monte Carlo samples used to simulate events are produced starting from the hard scatter interaction simulated with the MadGraph5_aMC@NLO [58]

or POWHEG 2.0 [59] generators and with the NNPDF3.0 [60] parton distribution function set. The hadronization and fragmentation effects, and the pileup conditions are simulated with PYTHIA 8.212 [61], while the simulation of the CMS detector response is based on GEANT4 [62].

In the $bb\tau\tau$ analysis, the estimation of background processes is performed mainly through Monte Carlo simulation, while some known flaws of the hard scatter and detector response modeling are corrected using data-driven techniques, as described in this Chapter.

4.1 HH signal

Both resonant and non-resonant double Higgs production mechanisms are modeled using a Monte Carlo simulation based on the MADGRAPH5_aMC@NLO generator at leading order precision.

For the resonant case, samples are generated under the assumption of a narrow width resonance, *i.e.* negligible when compared to the experimental detector resolution. The production of a resonance decaying into a HH pair is simulated for both the spin-0 and the spin-2 hypotheses, in a mass range that varies from 250 to 900 GeV.

In the context of non-resonant searches, in addition to the SM signal, the BSM models predict a wide variety of processes that have very different kinematic properties. The effective field theory allows to parametrize the Lagrangian according to five Higgs boson couplings, λ_{HHH}, y_t, c_2, c_g and c_{2g}, as already discussed in Section 1.3.1. The generation of samples for every combination of a five-dimensional hyperspace is clearly not feasible, thus an event reweighting approach is adopted to model the specific combinations of BSM couplings studied in this search. Out of the 12 EFT benchmarks described in Section 1.3.1, only six shapes are used as their combined statistics and kimeatics represent a sufficient input to the reweighting process.

To properly model the signal process for a particular set of EFT couplings, each event is reweighted according to the kinematic properties at matrix element level. At leading order, HH production is a $2 \to 2$ scattering process where the Higgs bosons are produced back-to-back in the azimuthal plane with the same transverse momentum. Since isotropy is assumed in the azimuthal direction, the process is fully described by two parameters represented by the invariant mass of the HH pair, m_{HH}, and the polar angle between one Higgs boson and the beam axis, $cos\theta^*$.

The bi-dimensional distribution $f(m_{HH}, cos\theta^*)$ for the combination of all the generated samples is shown in Figure 4.1 and the same distribution is produced for or every coupling combination of the effective Lagrangian parametrization that is explored in this search and denoted $f'(m_{HH}, cos\theta^*)$. For each event, the weight is computed from $\frac{f'(m_{HH}, |cos\theta^*|)}{f(m_{HH}, cos\theta^*)}$ as function of

m_{HH} and $cos\theta^*$. In order to ensure the the application of these weight modify only the shape, but not the yield of the distributions, a normalization corresponding to the sum over all the events of f'/f is applied to each event as well. An example of the resulting distribution is shown in Figure 4.1, where the reweighted sample for the SM hypothesis couplings is compared to the generated SM sample: the distributions are found to be in agreement inside the errors.

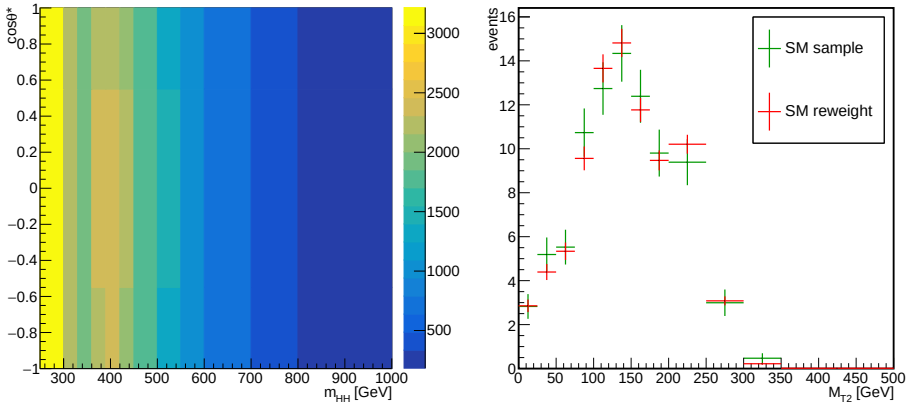


Figure 4.1: On the left, distribution of the simulated events for the combination of all the shape benchmark samples. On the right, comparison of the m_{HH} distributions obtained for a MC generated sample (green) and for the sample reweighted under the hypothesis of SM Higgs couplings (red).

4.2 QCD multi-jet background

One of the main difficulties in the $bb\tau\tau$ analysis is the correct identification of hadronically decaying taus. Thus, together with the $t\bar{t}$ processes described in Section 4.4, the multi-jet QCD events represent the main background source, especially in the fully hadronic $\tau_h\tau_h$ final state.

Most of the CMS analyses, involving $\tau\tau$ pairs in the final state, evaluate this contribution either with techniques related to the estimation of the $jet \rightarrow \tau_h$ rate in data sidebands (*e.g.* in [63]), or evaluating the yield and shape of multi-jet distributions in jet enriched regions in data, as in the case of the $HH \rightarrow bb\tau\tau$ search here described.

The use of MC simulation to evaluate the QCD contribution is discarded in favor of the data-driven method due to two main factors: firstly, the probability for a quark or a gluon jet to be identified as a τ_h object is very low

(between 10^{-2} and 10^{-3}) and it has to be combined with the equally poor probability to have in the event two additional jets that pass the medium working point of the b tagging discriminator. To cope with these rates, the QCD sample generated with a MC simulation would require a too large statistics to ensure a sufficient presence of events in the phase space considered in the analysis. The second reason why the data-driven method is preferred, is due to the fact that the misidentification rate for τ_h objects is mainly lead by detector effects that are very complex to simulate properly and can change over time in account of many external factors impossible to predict in advance.

In the analysis described in this thesis, the so-called ABCD method is adopted in order to model and estimate the QCD multi-jet background from jet enriched regions in data. The phase space of the events is divided in four regions, whose schematic representation can be seen in Figure 4.2:

- **Region A** Represents the signal region as defined in Section 3.3.1 and contains a pair of opposite sign electric charge (OS) tau leptons (either $\tau_{e/\mu}\tau_h$ or $\tau_h\tau_h$) and where all τ_h objects pass the medium working point of the tau isolation discriminant.

- **Region B** Represents the region where the multi-jet background is actually estimated and then extrapolated to the signal region. It is defined with the same isolation selections, but the pair charge requirement is inverted (same sign or SS).

- **Region C** It is composed by events with an opposite sign tau pair where τ_h objects pass the very loose working point of the tau isolation discriminant, but are required to fail the medium WP that defines the signal region. In the $\tau_\mu\tau_h$ and $\tau_e\tau_h$ channels this tau isolation selection is applied to the only τ_h candidate present in the event, while in the $\tau_h\tau_h$ final state it is applied only to the lowest p_T τ_h candidate selected. Together with region D, region C is used to evaluate the extrapolation factor from the SS to the OS region.

- **Region D** It is the region most different from the signal phase space as it has the same tau isolation criteria of region C, but it also requires that the leptons in the tau pair have the same electric charge.

In order to properly estimate the multi-jet yield, firstly the contributions coming from the other backgrounds estimated with MC simultaion, are subtracted from the data yield in the B,C and D regions. The QCD background yield in the signal region A is then estimated from region B through the ($k^{OS/SS}$) extrapolation factor:

$$N_A = N_B \times k^{OS/SS} = N_B \times \frac{N_C}{N_D} \tag{4.1}$$

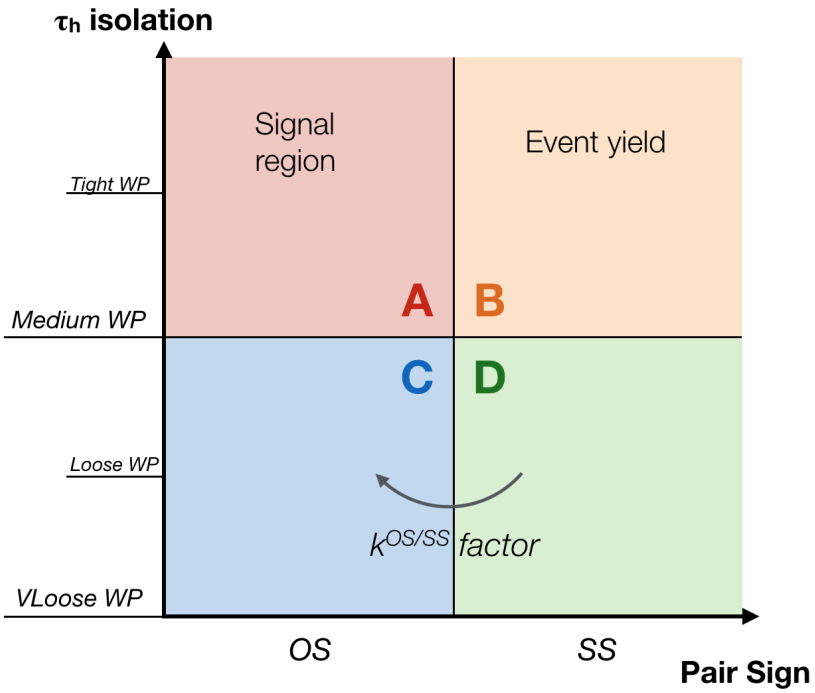


Figure 4.2: Schematic representation of the four regions used to estimate the QCD multi-jet background as described in Section 4.2.

The differential distribution (or "shape") of the QCD background is evaluated by subtracting from the data all the remaining MC simulation contributions in each bin of the distributions. As the statistics in the B region is typically limited, the shape of the multi-jet background is derived from a new region, denoted B', that is an extension of the previously defined B region. B' is obtained relaxing the τ_h isolation criterion to accept candidates passing the very loose working point of the discriminator. An example of the observed data and expected MC distributions in the B and B' regions for the three final states is shown in Figure 4.3 for the $2b0j$ category.

In the final maximum likelihood procedure described in Section 5.2, the multi-jet estimation is expressed as parametric function of the observed data and residual background contribution and is simultaneously fitted in the B,C, and D regions in order to fully correlate the background subtraction and to take into account the constrains that come from the signal region.

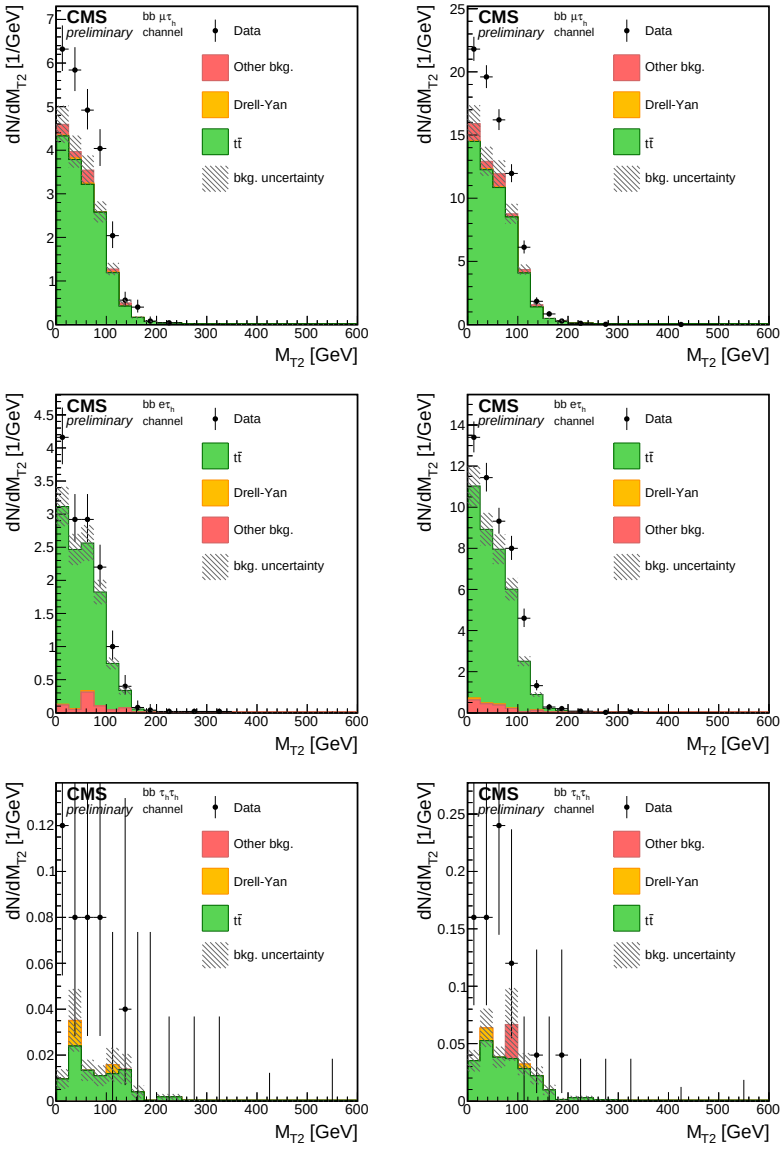


Figure 4.3: Distributions of the $MT2$ variable (defined in Section 5.1.2) for the $\tau_\mu\tau_h$ (top row), $\tau_e\tau_h$ (central) and $\tau_h\tau_h$ (bottom row) channels. The distributions show events in the $2b0j$ category for the B region (SS-isolated on the left) and for the B' region (SS-relaxed isolation on the right).

4.3 Drell-Yan $Z/\gamma^* \to \tau\tau$ background

As already mentioned in the introduction to this Chapter, Drell-Yan decays in association with the production of two jets represent one of the main

background contributions for the $bb\tau\tau$ analysis. The MC simulation used for the DY sample is based on the MadGraph5_aMC@NLO generator at Leading Order precision and the total theoretical cross section is computed at NNLO precision as $\sigma(Z/\gamma^* \to \ell\ell) = 5765\ pb$. Given the narrow phase space studied in the analysis, the DY statistics is increased by combining the inclusive sample with complementary ones, where the emission of 1, 2, 3 or 4 additional jets, or the emission of 1 or 2 b jets, is required. Since the full MC simulation process is quite consuming, in terms of computing time and power, in all the events of these samples the invariant mass $m_{\ell\ell}$ is forced to be larger than 50 GeV without loosing any information from low mass events that would be in any case be excluded from the analysis by the selections used to define the signal region.

If, on one hand, the differential distributions of the DY events show a good agreement with the observed data, as shown in Figure 4.4, the modelization of the yield, especially when the production is in association with multiple jets, is known to be imperfect and thus require a correction, that in the case of the $bb\tau\tau$ analysis is computed from data in a control region.

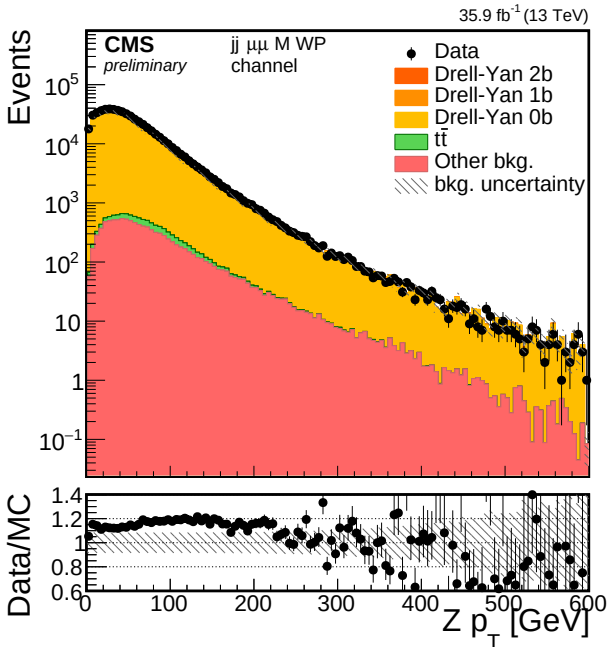


Figure 4.4: Differential distribution of the Z boson p_T for $Z \to \mu\mu$ events in association to two "light" jets that do not pass the b tagging medium WP. The data/MC ratio is flat as function of the transverse momentum and only the normalization factor presents a disagreement.

$Z \to \mu\mu$ events are selected using the same $\tau_\mu\tau_h$ triggers described in Section 3.1 and requiring the presence of two muons with $p_T > 23\ GeV$

and $|\eta| < 2.1$ for the leading one and $p_T > 10\ GeV$ and $|\eta| < 2.4$ for the trailing one, both isolated and passing the tight working point of the muon identification discriminator. In addition, the invariant mass of the muons must be in the window $60 - 120\ GeV$ and the event must contain two jets with $p_T > 20\ GeV$ and $|\eta| < 2.4$. The invariant mass requirement on the jet pair is $80 < m_{jj} < 160\ GeV$, in order to select a phase space as similar as possible to the signal region.

Both data and MC simulated events thus selected are split in three regions according to the number of jets that pass the medium working point of the b tag discriminant.

- **Z + light jets** This region is obtained by inverting the b tagging request on the two jets selected, so that the main contribution of events arises from the erroneous identification of light jets as b jets. Event if the contamination from other backgrounds is negligible a further selection is applied by requiring that the missing transverse momentum of the event is smaller than 45 GeV in order to fully reject $t\bar{t}$ contributions.

- **Z + one b jet** This region is optimized in order to enhance the contribution coming from $Z+1\ b\ jet$ processes, thus one of the selected jets is required to pass the medium b tagging WP, while the second is required to fail it.

- **Z + two b jets** This represent the most important source of background as its final state is composed of exactly the same particles as in $HH \to bb\tau\tau$ events, with the only exception that the tau and b pairs do not originate from Higgs bosons. Both selected jets are required to pass the medium b tagging working point.

Scale factors for each of these regions is estimated through a likelihood fit on the $m_{\mu\mu}$ distribution in the range $60 < m_{\mu\mu} < 120\ GeV$ and they are used to correct the Drell-Yan background yield. Thus, three templates obtained from the three control regions and one additional for the complex residual background contributions are simultaneously fitted and their normalization is allowed to float around the initial value estimated from the simulation. The result of the likellihood fit is composed by three scale factors that are reported in Table 4.1, while the distributions of $m_{\mu\mu}$ for the three control regions, before and after the application of the scale factors, is shown in Figure 4.5

Given the large cross section of $Z+$ light jets, quite large contamination of this process is expected in the other regions that thus can not be considered independent. To account for this effect, the fit introduces correlations between the correction factors and their errors and covariance matrix are handles as systematic uncertainties and included in the final limit setting procedure (detailed in Section 5.2) as nuisance parameters of the model.

Process	Scale Factor
$Z+$ light jets	1.1412 ± 0.0017
$Z+$ one b jet	1.187 ± 0.015
$Z+$ two b jets	1.170 ± 0.029

Table 4.1: Scale factors for the three Drell-Yan components.

4.3.1 2017 LO to NLO reweighting

The procedure followed in the 2016 analysis provides a good agreement between data and Monte Carlo for some basic kinematic variables, however, some more complex observables, such as the invariant mass of the four bodies, may still suffer from the known imperfect modelization provided by the LO Drell-Yan samples.

The analysis of the differential cross section measurement of a Z boson production in association with jets [64] shows that the use of the NLO sample leads a much better modeling of data. Nonetheless, this sample has a maximum of 2 jets at generator level, while, in the $bb\tau\tau$ final state, a major contribution from DY $+ \ > 2$ jets is expected, especially in the $2b0j$ category which represents the most sensitive region of the analysis. In addition to this, the limited DY NLO statistcs renders the sample almost useless for this analysis purpose.

In order to overcome this issue, a reweighting procedure to match the LO sample to the generator level quantities of the NLO simulation, has been implemented after the publication of the 2016 analysis, and is part of the improvements that are been developed in view of the Run II $bb\tau\tau$ legacy paper.

Two sets of scale factors are determined.
The first is used to match the fractions of events with a particular combination of light- and b-jets at generator level from the LO to the NLO Drell-Yan Monte Carlo sample. Since the NLO sample at generator level involves the emission of at most 2 jets, in order to compute these scale factors, six concurring processes can be identified:

$$\begin{aligned} &Z \to \mu\mu \ + \ n_{light} + \ n_b \\ &\quad \text{with } 0 \leq n_{light}, n_b \leq 2 \text{ and } n_{light} + n_b \leq 2 \end{aligned} \tag{4.2}$$

where n_{light} and n_b represent the number of light and b jets emitted at generator level. The $p_T(Z)$ distributions of these processes defined in Equation 4.2 are reported in Figure 4.6 simply to show the difference between the LO and NLO simulation.

The second set of corrections is determined as the ratio between the LO

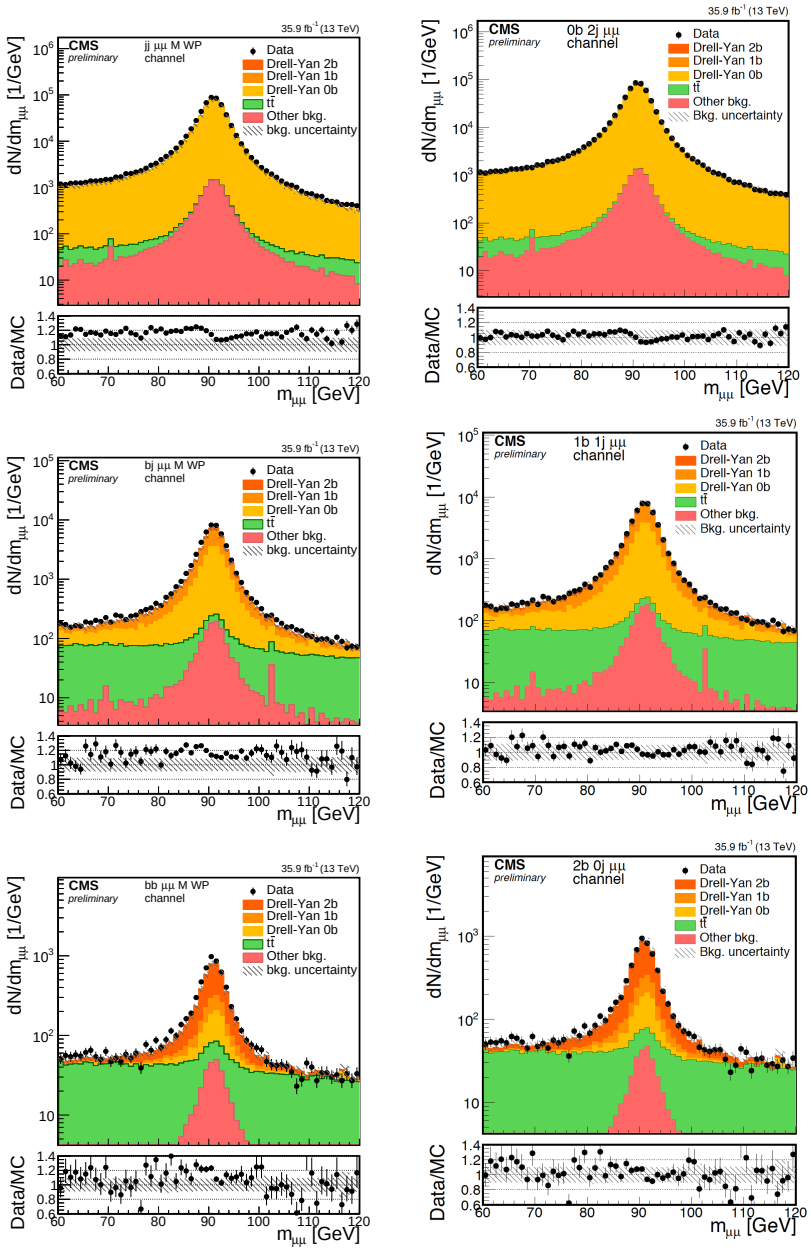


Figure 4.5: Distributions of the $m_{\mu\mu}$ variable before (left column) and after (right column) the application of the scale factors. The top row shows the $Z+$ light jets category, the central one shows the $Z+$ one b jet and the last row the $Z+$ two b jets.

and NLO Z boson transverse momentum distributions of all the possible contributions to the DY + jets background, which of course are not limited at the emission of just two jets (as it was the case with the NLO sample).

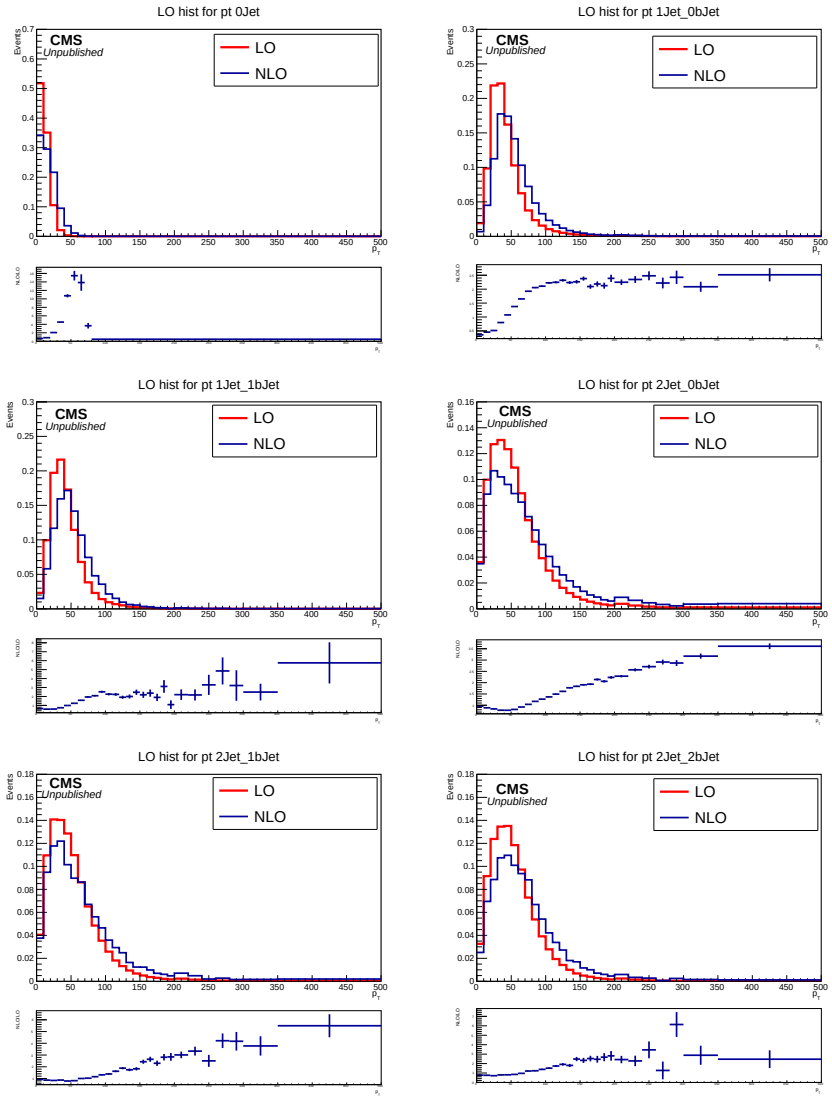


Figure 4.6: Comparison of the Z boson p_T distributions at generator level between the LO and NLO Drell-Yan Monte Carlo samples.

Six event categories are defined:

- Z + two b jets
- Z + two b jets + any number of light jets
- Z + one b jet + $\leq$ one light jet
- Z + one b jet + $>$ one light jet
- Z + $\leq$ two light jets

- $Z + >$ two light jets

The $p_T(Z)$ distributions of these categories are used as input to a simultaneous likelihood fit where the normalization of the processes is left floating.

The data control regions used for the SF extraction are the same used also in 2016, but with an additional requiremens on $p_T^{miss} < 45\ GeV$ to further reduce the QCD and $t\bar{t}$ contributions. Furthermore, events are required to pass an invariant mass requirement that is similar to the selection defined in Equation 3.4 in Section 3.3.3, but with relaxed conditions:

$$\frac{(m_{\mu\mu} - 116\ GeV)^2}{(35 + 5\ GeV)^2} + \frac{(m_{bb} - 111\ GeV)^2}{(45 + 5\ GeV)^2} < 1 \tag{4.3}$$

Both sets of scale factors are reported in Table 4.2.

Drell-Yan LO to NLO reweight				
Gen process	Event fraction SF	-	DY contribution	Z_{p_T} SF
0 light + 0 b jets	1.36	-	$Z+ \leq$ two light jets	1.1465 ± 0.002
1 light + 0 b jets	1.50	-	$Z+ >$ two light jets	0.01 ± 0.0002
0 light + 1 b jets	2.02	-	$Z+$ one b jet + $\leq$ one light jet	1.577 ± 0.066
2 light + 0 b jets	0.7	-	$Z+$ one b jet + $>$ one light jet	0.01 ± 0.01
1 light + 1 b jets	0.86	-	$Z+$ two b jets	1.903 ± 0.568
0 light + 2 b jets	0.59	-	$Z+$ two b jets + any num of light jets	0.189 ± 0.519

Table 4.2: Scale factors for the Drell-Yan LO to NLO reweight. The first two columns report the SFs relative to the the fraction of events with a particular combination of number of light jets and number of b jets at generator level, while the tlast two columns report the SFs based on the transverse momentum of the Z boson.

Even though this procedure was developed using 2016 samples and its deployment on 2017 DY samples may be sub-optimal, the improvement in the comparison between observed data and MC simulation is clearly visible in Figure 4.7.

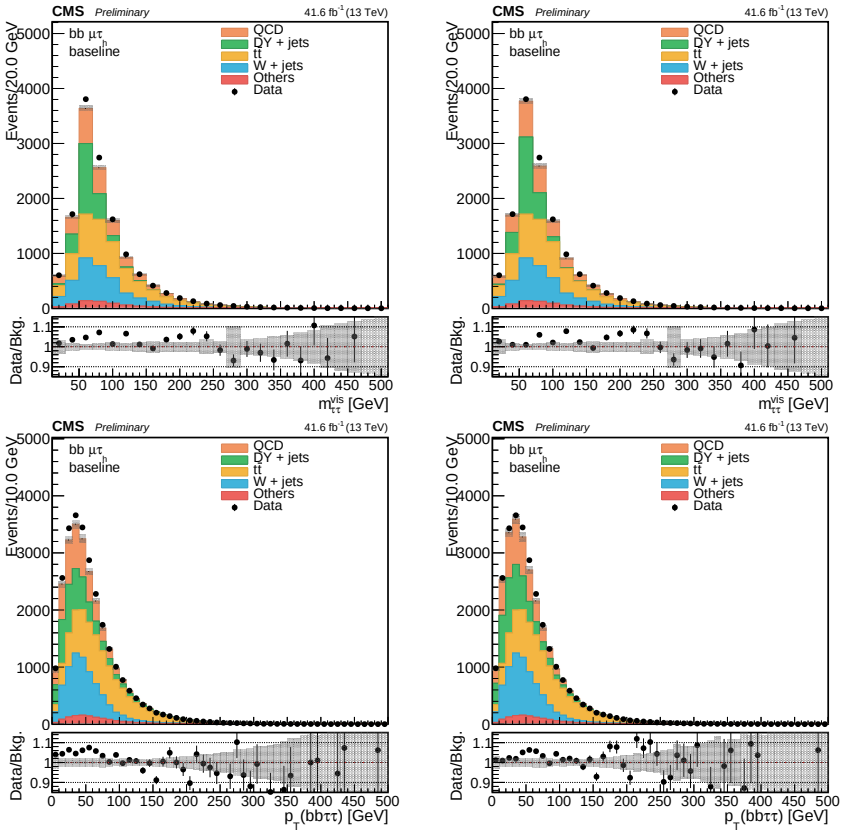


Figure 4.7: Distributions in the $\tau_\mu\tau_h$ final state of the visible mass of the tau pair ($m_{\tau\tau}^{vis}$, top row) and of the transverse momentum of the HH system reconstructed as sum of the visible decay products ($p_T(bb\tau\tau)$, bottom row). No selection is applied on the b tagging score of the jets. The left column represents the distributions before applying the scale factors, the right one the distributions after the SF application.

4.4 $t\bar{t}$ background

The principal contribution of background, especially in the semileptonic channels, originates from $t\bar{t}$ events, that in the $bb\tau\tau$ analysis are modeled through a Monte Carlo simulation with the POWHEG 2.0 generator at NLO precision. Events are simulated assuming a top quark mass of 172.5 GeV and normalized to the inclusive coss section at NNLO precision which corresponds to 831.8 pb.

In 2016, in order to obtain a satisfying coverage of the phase space analyzed, the inclusive $t\bar{t}$ sample was supplemented with two additional samples generated for the fully leptonic and semileptonic decays of the top quark pair. During the MC production campaign related to 2017 instead, the CMS Collaboration decided upon generating three different samples that cover the total phase space (fully hadronic, fully leptonic and semileptonic) and that are thus used in the $bb\tau\tau$ 2017 analysis.

Measurements of the $t\bar{t}$ differential production cross section in CMS [65] show that this process is well described by the MC simulation and the excellent agreement with the observed data is shown in Section 3.3.3 for some kinematic variables of interest. In particular, the plots in Figures 3.29 and 3.30, for the $\tau_\mu\tau_h$ and $\tau_e\tau_h$ channels respectively, show the distributions in the $2b0j$ category, which becomes dominated by $t\bar{t}$ events once the request of two b tagged jets is applied.

Residual differences are nonetheless present, especially in the transverse momentum distributions, and, given the importance of the $t\bar{t}$ background in the $bb\tau\tau$ analysis, must be taken into account. An event reweight technique is used to derive a systematic uncertainty, that is based on the generated top quark p_T following the recommendations of the Top Physics Analysis Group (Top PAG) and is described in Section 5.3.

4.5 Other backgrounds

Other background contributions show a very limited presence in the phase space considered in the $bb\tau\tau$ analysis and their contribution and modeling, both in shape and event yield, are assessed relying solely on Monte Carlo simulation.

The next Paragraphs describe the contributions originating from W bosons production in association with jets, single top quark production, pair production of vector bosons, electroweak production of a vector boson in association with jets and Standard Model single Higgs boson production.

W+jets
In this analysis the contribution of $W + jets$ background is highly sup-

pressed by the requirement applied to b tagging discriminator value of the jets The production of $W \to \ell\nu_\ell$ (with $\ell = e, \mu, \tau$) is simulated with the MadGraph5_aMC@NLO generator at Leading Order precision and normalized to the theoretical NNLO cross section $\sigma(W \to \ell\nu_\ell) = 6.15 \times 10^4$ *pb*.

Single top
The contribution of single top quark production in association with a W boson is extremely small and is simulated with the Powheg 2.0 generator at NLO precision, normalized to the NNLO theoretical cross section $\sigma(tW) = 71.7$ *pb*.

VV
The background contributions arising from vector boson pairs, included in this search, are ZZ, ZW and WW and are generated using both MadGraph5_aMC@NLO and Powheg 2.0. The ZZ process is split in four different samples according to the final decay state simulated ($\ell\ell\ell\ell$, $\ell\ell\nu_\ell\nu_\ell$, $\ell\ell qq$ and $qqqq$) and normalized to the NNLO inclusive production cross section $\sigma(ZZ) = 16.5$ *pb*. Samples for the WW process are generated for the $\ell\nu_\ell\ell\nu_\ell$, $\ell\nu_\ell qq$ and $qqqq$ channels, and are normalized to $\sigma(WW) = 118.7$ *pb*, while those for ZW, normalized to $\sigma(ZW) = 45$ *pb*, comprehend the $\ell\ell\ell\nu_\ell$, $\nu_\ell\nu_\ell\ell\nu_\ell$, $qq\ell\nu_\ell$ and $\ell\ell qq$ final states.

Electroweak V+jets
Electroweak production of W^+, W^- or Z in association with two jets is simulated with the MadGraph5_aMC@NLO generator and each process is normalized to the LO cross section obtained from the MC generator: $\sigma^{EWK}(W^+) = 25.69$ *pb*, $\sigma^{EWK}(W^-) = 20.25$ *pb* and $\sigma^{EWK}(Z) = 3.987$ *pb*.

Single Higgs
The single Higgs production cross section is small compared to the other backgrounds. However the associated production of single Higgs with a vector boson (VH) or with a pair of top quarks (ttH) have similar final states with the $HH \to bb\tau\tau$ signal and therefore are considered as backgrounds in this search. The MC simulation of these processes is realized through the Powheg 2.0 generator assuming a $m_H = 125$ *GeV* Higgs boson in the final states $ZH \to \ell\ell bb/qqbb$ and $Z \to any\ H \to \tau\tau$. The samples are normalized to the inclusive cross section computed at NNLO precision of the QCD corrections and at the NLO precision of electroweak corrections, that amounts to $\sigma(ZH) = 0.884$ *pb*.

In 2017, the inclusion of the vector boson fusion HH production mechanism in the search prompted us to evaluate the presence of contributions coming from other single Higgs processes. The gluon and vector boson fusion samples, with the Higgs boson decaying to a $\tau\tau$ pair, were added to

the list of possible backgrounds. The MC simulation is performed through the POWHEG 2.0 generator and the samples are normalized to the single Higgs cross section scaled by the branching fraction of the $\tau\tau$ final state: $\sigma^{VBF}(H) \times \mathcal{B}(H \to \tau\tau) = 0.24\ pb$ and $\sigma^{gg}(H) \times \mathcal{B}(H \to \tau\tau) = 1.35\ pb$.

4.6 Pileup treatment

In order to obtain a valid comparison between observed data and Monte Carlo simulation, a correct treatment of the pileup must be taken into account. The interaction point environment at the LHC, especially in experiments like ATLAS and CMS where particle beams collide "head-on", is not easily reproducible in MC and the simulated distribution of the number of real proton-proton interactions in a bunch crossing is slightly different than the observed data, as shown in Figure 4.8 for the 2016 analysis. The MC events therefore must be reweighted in a manner such that the distributions of the additional energy and tracks from the extra pileup interactions are adjusted to be the same as in data. This is accomplished by giving each MC event a weight that corresponds to the probability that the number of interactions in a given MC event occurs in the data sample.

In 2017, the CMS Monte Carlo production campaign suffered from a problem related to the multi-threaded infrastructure of the CMS software (CMSSW) that affected some of the samples produced. As a consequence, the pileup distributions of the affected samples erroneously display sharp peaks randomly distributed over the spectrum. In order to properly account for these features in the MC samples, dedicated weights were computed for each MC sample used in the 2017 $bb\tau\tau$ analysis and applied to the simulated events accordingly.

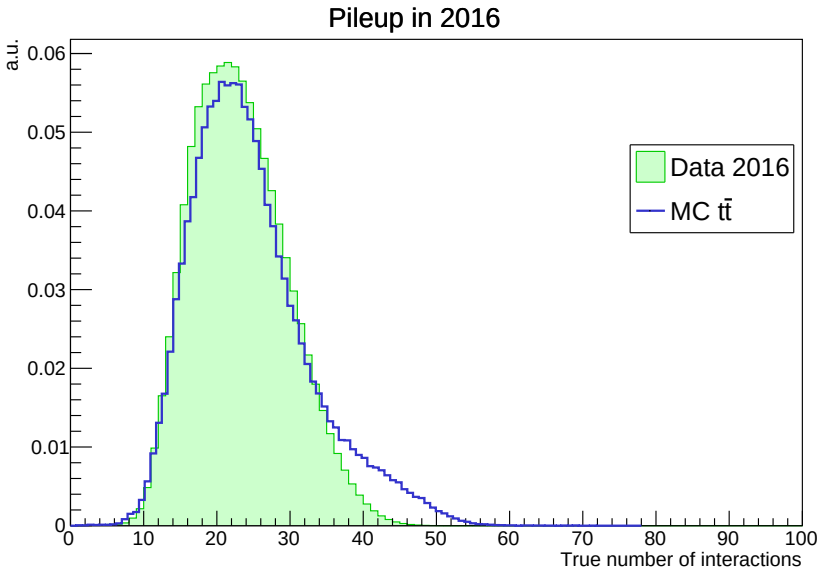


Figure 4.8: Distributions of the number of proton-proton interactions as measured in 2016 data (green shadowed histogram) and as generated by the Monte Carlo simulation for a $t\bar{t}$ sample (blue solid line). The data distribution is obtained assuming a minimum bias cross section of 69.2 mb and it is derived for $35.9 fb^1$ which correspond to the full integrated luminosity collected in 2016. The ratio of the distributions is used to define the weight to be applied to the MC simulation.

Chapter 5

Results of $b\bar{b}\tau^+\tau^-$ Searches

The aim of the $bb\tau\tau$ search is to explore both the resonant and the non-resonant double Higgs production mechanisms in order to explore the electroweak sector of the Standard Model and probe the existence of New Physiscs predicted by BSM models, as described in Chapter 1.

The exploration of HH production with CMS data requires the definition of variables with high discriminating power between signal and background events (Section5 5.1). A proper statistical model and a good knowledge of the systematic uncertainties affecting the measurements (Sections 5.2 and 5.3, respectively) are also needed to evaluate the presence of a signal, or, in case of its absence, to set exclusion limits based on the observed data and the MC simulations.

In Section 5.4 I will present the final results obtained from the analysis of the data collected in 2016 and published in [1]. Finally, in Section 5.5 I will give an overview of the improvements put in place for the 2017 analysis and I will evaluate the enhancement in sensitivity. The future CMS combined (2016+2017+2018 data) legacy paper will benefit from the studies performed on 2017 data and illustrated in this thesis.

5.1 Discriminating variables

The definition of sensitive observables to explore double Higgs production is a crucial step to maximize the sensitivity of the analysis.

In 2016, different variables were used to search for the presence of signal events in the resonant and non-resonant searches and they are detailed in Sections 5.1.1 and 5.1.2, respectively. The distributions of these variables for the different channels and categories of the analysis are shown in Section 5.4.

In 2017 instead, as described in Section 5.1.3, a similar approach was adopted in both searches.

5.1.1 2016 resonant search

The expected signal signature of two Higgs bosons originating from a high mass resonance, a sharp peak over a continuous background, drives the choice of the discriminating variable adopted in the resonant search, that is the invariant mass m_X of the resonance itself.

The resolution of m_X is spoiled by the presence of neutrinos in the final state, thus requiring the usage of a kinematic fit to correct for the missing energy information and obtain a valid m_{HH} estimate. The use of a kinematic fit was already adopted in the same search with Run I data [66].

The constraints imposed in the kinematic fit follow from the hypothesis of two 125 GeV bosons decaying in bb and $\tau\tau$ pairs:

$$m(\tau_1, \tau_2) = m(b_1, b_2) = m_H = 125\ GeV \tag{5.1}$$

For the reconstructed b jets, it is assumed that the measurement of the directions $\eta_{b_{1,2}}$ and $\phi_{b_{1,2}}$ is very accurate compared to the b jet energy. As any measurement of the jet momentum applies also for the jet energy, the ratio $\vec{\beta} = \vec{p}/E$ does not change in first approximation and the same holds for $\gamma = 1/\sqrt{1-\beta^2}$. The energy of one b jet can directly be calculated from the other's using the invariant mass constraint:

$$\begin{aligned} m_h^2 &= p_{b_1}^2 + p_{b_2}^{2,new} + 2p_{b_1}p_{b_2}^{new} \\ &= m_{b_1}^2 + E_{b_2}^{2,new}\gamma_{b_2}^{-2} + 2E_{b_1}E_{b_2}^{new}k \end{aligned} \tag{5.2}$$

where $k = 1 - \vec{\beta_{b_1}}\vec{\beta_{b_2}}$ is assumed constant and can therefore be calculated from the pre-fit event kinematics. Equation 5.2 can thus be solved as:

$$E_{b_2}^{new} = E_{b_1}k\gamma_{b_2}^2\left(-1 + \sqrt{1 + \frac{m_h^2 - m_{b_1}^2}{(E_{b_1}k\gamma_{b_2})^2}}\right) \tag{5.3}$$

Since the tau leptons originate from a heavy object, compared to their own mass ($m_H/m_\tau \simeq 70$), they are highly boosted and the collinear approximation holds: the reconstructed direction of the visible decay products of the tau leptons is assumed to point into the direction of the original τ leptons. Again, only the energy of one tau is a free fit parameter, as the other can be constrained analogously to Equation 5.3.

Furthermore, it is also assumed that the reconstruction of the η and ϕ directions of the two b jets and of the reconstructed tau decay products are accurately determined with uncertainties negligible when compared to those arising from their energy measurement.

These assumptions reduce the number of parameters needed to describe the $bb\tau\tau$ system to two, which are the energies of the first b jet (E_{b_1}) and of the first tau lepton (E_{τ_1}). By varying these parameters, a χ^2 minimization is performed and the best estimate is used to reconstruct the mass of the heavy resonance.
For the two measured b jets, the χ^2 terms can be written as

$$\chi^2_{b_1,b_2} = \left(\frac{E^{fit}_{b_1,b_2} - E^{meas}_{b_1,b_2}}{\sigma_{b_1,b_2}} \right)^2 \tag{5.4}$$

where $E^{fit}_{b_1,b-2}$ and $E^{meas}_{b_1,b_2}$ are the fitted and reconstructed b jet energies, respectively, while σ_{b_1,b_2} represent the energy resolution.

For the χ^2 term coming from from the $\tau\tau$ pair, the presence of neutrinos prevents any accurate measurement of the original tau energies, but contributes to the missing transverse energy reconstructed in the event. The MET can therefore be exploited to constrain the *tau* lepton energies by comparing the expected transverse momentum of the resonance

$$\vec{p}^{\,fit}_{T,X} = \vec{p}^{\,fit}_{T,H_1} + \vec{p}^{\,fit}_{T,H_2} = \vec{p}^{\,fit}_{T,b_1} + \vec{p}^{\,fit}_{T,b_2} + \vec{p}^{\,fit}_{T,\tau_1} + \vec{p}^{\,fit}_{T,\tau_2} \tag{5.5}$$

with the transverse momentum measured experimentally

$$\vec{p}^{\,meas}_{T,X} = \vec{p}^{\,meas}_{T,H_1} + \vec{p}^{\,meas}_{T,H_2} = \vec{p}^{\,meas}_{T,b_1} + \vec{p}^{\,meas}_{T,b_2} + \vec{p}^{\,meas}_{T,\tau_1} + \vec{p}^{\,meas}_{T,\tau_2} + \vec{MET} \tag{5.6}$$

which by definition corresponds to the reconstructed transversal recoil of the resonance

$$\vec{p}^{\,meas}_{T,recoil} = -\vec{p}^{\,meas}_{T,X} \tag{5.7}$$

Any nonzero residual recoil vector can thus be written as

$$\vec{p}^{\,res}_{T,recoil} = \vec{p}^{\,fit}_{T,X} - \vec{p}^{\,meas}_{T,X} = \vec{p}^{\,fit}_{T,X} + \vec{p}^{\,meas}_{T,recoil} \tag{5.8}$$

and contributes to the χ^2 as

$$\chi^2_{recoil} = \left(\vec{p}^{\,res}_{T,recoil}\right)^T \cdot V_{recoil} \cdot \vec{p}^{\,res}_{T,recoil} \tag{5.9}$$

where V_{recoil} is the covariance matrix of the reconstructed recoil vector.

The complete χ^2 function finally reads

$$\chi^2 = \chi^2_{b_1} + \chi^2_{b_2} + \chi^2_{recoil} \tag{5.10}$$

After minimization of this function by varying E_{b_1} and E_{τ_1}, their best estimate is used to compute the final value of m_X.

The usage of the kinematic fit allows for a very accurate reconstruction of the heavy resonance mass on an event-by-event basis with a distribution of the reconstructed m_{HH}^{KinFit} centered around the the mass of the resonance. The invariant mass reconstructed with the kinematic fit is compared to the invariant mass obtained considering only the visible decay products of the τ leptons and the b jets, and the distributions for possible resonances of different masses are shown in Figure 5.1.

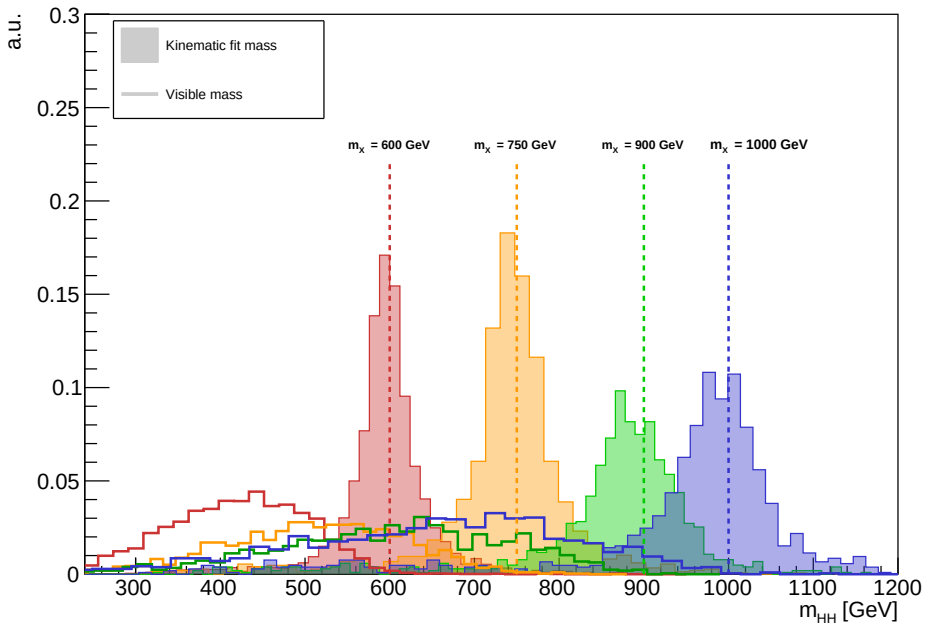


Figure 5.1: Comparison of m_{HH} obtained with the kinematic fit (shadowed histograms) and as the visible $bb\tau\tau$ invariant mass (solid lines). The correct value of the resonance mass is reported in the plot for the four samples displayed. The events belong to the $2b0j$ category for the $\tau_\mu\tau_h$, $\tau_e\tau_h$ and $\tau_h\tau_h$ channels, after the mass selections described in Section 3.3.3 are applied.

5.1.2 2016 non-resonant search

In the non-resonant production mechanism, double Higgs events do not have a signature as clear as the resonance peak described in Section 5.1.1. A different variable, denoted as "stransverse mass", or $MT2$, is thus used to disentangle the signal and background contributions.

Originally proposed for supersymmetry searches [67, 68] and later ad-

justed for $HH \to bb\tau\tau$ analyses [69], $MT2$ is developed to exploit the kinematic information of events containing two equal mass particles undergoing a two-body decay into a visible and invisible particle. The stransverse mass is a generalization of the transverse mass m_T and is defined as the largest mass of the parent particle that is compatible with the kinematic constraints of the event.

In the context of the $HH \to bb\tau\tau$ analysis, the $t\bar{t}$ process is one of the main background and it represents the perfect example to showcase the power of the $MT2$ variable. A top quark is interpreted as the parent particle, while its decay products, a b quark and a W boson, are considered as the two daughter particles.

In the following description $\vec{b}$, $\vec{b'}$, m_b and $m_{b'}$ denote the momenta of the two selected b jets and their masses, respectively. The remaining particles produced in the top quark decay, *i.e.* the measured leptons and the neutrinos, are globally denoted as $\vec{c}$ and $\vec{c'}$, while their masses, due to the undetectable neutrinos, are set, as suggested in [69], to $m_c = m^{vis}(\tau_1)$ and $m_{c'} = m^{vis}(\tau_2)$, where m^{vis} denotes the invariant mass of the measured leptons or τ_h. The components $\vec{c}$ and $\vec{c'}$ are defined as the sum of the decay products of the W boson and contain both visible and invisible components so that they form the transverse momentum constraint

$$\vec{c}_T + \vec{c'}_T = \vec{p}_T^{\Sigma} = \vec{p}_T^{vis}(\tau_1) + \vec{p}_T^{vis}(\tau_2) + M\vec{E}T. \tag{5.11}$$

For each particle the "transverse energy" e is defined as

$$e = \sqrt{m^2 + p_T^2} \tag{5.12}$$

while the transverse mass is:

$$m_T\left(\vec{b}_T, \vec{c}_T, m_b, m_c\right) = \sqrt{m_b^2 + m_c^2 + 2\left(e_b e_c - \vec{b}_T \cdot \vec{c}_T\right)} \tag{5.13}$$

Starting from Equation 5.13, the stransverse mass can be defined as:

$$MT2\left(m_b, m_{b'}, \vec{b}_T, \vec{b'}_T, \vec{p}_T^{\Sigma}, m_c, m_{c'}\right) = \min_{\vec{c}_T + \vec{c'}_T = \vec{p}_T^{\Sigma}} \left\{\max\left(m_T, m'_T\right)\right\} \tag{5.14}$$

where m_T is the transverse mass constructed from m_b, m_c, $\vec{b}_T$ and $\vec{c}_T$, while m'_T is the transverse mass constructed from $m_{b'}$, $m_{c'}$, $\vec{b'}_T$ and $\vec{c'}_T$. In the $bb\tau\tau$ analysis, the minimization in Equation 5.14 is over the measured momenta of the tau leptons and the missing transverse momentum previously defined $\vec{p}_T^{\Sigma} = \vec{c}_T + \vec{c'}_T$.

Equation 5.14 is computed with the bisection method provided in [70], which yields a fast and stable minimization procedure, and can achieve machine numerical precision.

The $MT2$ variable has, by construction, a large discrimination power of HH events against the irreducible $t\bar{t} \to bbWW \to bb\tau\nu_\tau\tau\nu_\tau$ background which is bounded above at the top quark mass, while it has no such limitation for the HH signal where the τ and b jet do not originate from the same parent particle. Any presence of signal events would thus appear as an enhancement of the event yield in the tail of the $MT2$ distribution.

Finally, it should be noted that the distributions of the stransverse mass reported in Section 5.4 show a small contribution of $t\bar{t}$ events also at high values of $MT2$ due to detector resolution effects and other decay modes of the $t\bar{t}$ system that result in an extension of the tail of the $MT2$ $t\bar{t}$ background distribution.

5.1.3 2017 search

Given the excellent BDT performance in rejecting the $t\bar{t}$ background during the analysis of 2016 data, the method was improved and extended to include also the fully hadronic channel $\tau_h\tau_h$. In this multivariate approach, the information from kinematic variables is fully exploited in order to produce a single score to be used as final discriminating variable.

Three separate trainings are performed for the resonant samples in the low mass (LM) region (from 250 GeV to 320 GeV), in the medium mass (MM) region (from 320 GeV to 450 GeV), and in the high mass (HM) region, which covers the mass range up to 900 GeV. These regions are identified studying the compatibility among the probability density functions (PDFs) of the variables at different mass points. An additional training is performed for non-resonant signals (NR). To further improve the available statistics, in each mass region the BDT is trained using events from both spin hypotheses and combining the three final states. A parametrized learning approach is chosen to add information (*e.g.* the mass of the resonance and the $\tau\tau$ final state) to keep track of the origin of the events.

For each one of the four mass regions training, the set of input variables must be carefully selected taking into account both the kinematic differences between signal and background events, and considering the necessity to provide to the BDT as much information as possible. A statistical method based on "mutual information" (MI) [71], is used to determine, starting from an extensive set of variables, the optimal ones to be used as input to the BDT. The MI is a measure of the information that two variables share and it can quantify how much information can be obtained about one variable, through the other one. The MI is used to sort the variables in descending order for their ability to distinguish signal from background, either individually or in pair with other selected variables. Out of a broader set of possible inputs, which has been recursively pruned of the least discriminating and most correlated variables, the twenty highest performing ones are chosen for each training and are reported in Table 5.1.

Most of the variables reported in Table 5.1 have already been described in this thesis, the remaining ones are listed here:

- Starting from m_T, defined in Equation 3.6, a generalized variable can be defined as $m_T^{total} = \sqrt{m_T(\tau_1, MET)^2 + m_T(\tau_2, MET)^2 + m_T(\tau_1, \tau_2)^2}$
- ϕ_1 and ϕ_2 are the angles defined between the decay plane of the first jet or the first lepton, respectively, and a plane defined by the vector of H_{bb} or $H_{\tau\tau}$ in the four final objects rest frame and the positive direction of the collision axis.
- $\phi(H_{bb}, H_{\tau\tau})$ is the angle between the decay planes of the four final state objects expressed in their rest frame.

Input variables			
LM $[250-320]$ GeV	MM $[320-450]$ GeV	HM $[450-900]$ GeV	NR
$m_T(\tau_1, MET)$	$\lvert\Delta\phi(H_{\tau\tau}^{SVfit}, MET)\rvert$	$\Delta R(\tau_1, \tau_2)$	$\cos\theta(H_{bb}, MET)$
m_{HH}^{KinFit}	m_{HH}^{KinFit}	$\boldsymbol{\chi^2(m_{HH}^{KinFit})}$	m_{HH}^{KinFit}
$\cos\theta(H_{\tau\tau}^{SVfit}, MET)$	$\boldsymbol{\chi^2(m_{HH}^{KinFit})}$	$P_T(H_{\tau\tau}^{SVfit})$	$\boldsymbol{\chi^2(m_{HH}^{KinFit})}$
$\phi_2(H_{bb}, H_{\tau\tau}^{SVfit})$	$\lvert\Delta\phi(H_{bb}, MET)\rvert$	H_T	$P_T(\tau_2)$
$m(X^{MET})$	$m_T(\tau_1, MET)$	$\mathbf{m(m_{HH}^{KinFit})}$	$m_X^{reduced}$
$\boldsymbol{\chi^2(m_{HH}^{KinFit})}$	$\Delta R(\tau_1, \tau_2)$	p_ζ^{vis}	p_ζ
$\mathbf{m(X^{vis})}$	$\lvert\Delta\phi(b_1, b_2)\rvert$	$\Delta R(H_{bb}, H_{\tau\tau}^{MET})$	$P_T(H_{\tau\tau}^{SVfit})$
$\Delta R(H_{bb}, MET)$	$\lvert\Delta\eta(\tau_2, MET)\rvert$	$m_X^{reduced}$	p_ζ^{vis}
$MT2$	$\lvert\Delta\eta(b_1, b_2)\rvert$	$m(X^{SVfit})$	$m(X^{SVfit})$
$m_T(H_{\tau\tau}^{SVfit}, MET)$	$\lvert\Delta\eta(H_{bb}, MET)\rvert$	p_ζ	$MT2$
$\Delta\phi(\tau_1, \tau_2)$	$\phi(H_{bb}, H_{\tau\tau}^{SVfit})$	$P_T(\tau_2)$	$P_T(\tau_1)$
p_ζ^{vis}	$m_X^{reduced}$	$\mathbf{m(X^{vis})}$	$\mathbf{m(X^{vis})}$
$\Delta\phi(\tau_1, MET)$	$\cos\theta(H_{\tau\tau}^{SVfit}, MET)$	$\lvert\Delta\phi(H_{\tau\tau}^{SVfit}, MET)\rvert$	$P_T(b_2)$
$P_T(H_{bb})$	$\phi_1(H_{bb}, H_{\tau\tau}^{SVfit})$	$P_T(\tau_1)$	H_T
$\Delta R(\tau_1, \tau_2) \cdot p_T(H_{\tau\tau}^{SVfit})$	$m_T(H_{\tau\tau}^{SVfit}, MET)$	$m(X^{MET})$	$m_T(H_{\tau\tau}^{MET})$
$\Delta\phi(b_1, b_2)$	p_ζ	$m_T(H_{\tau\tau}^{MET}, MET)$	$\Delta\phi(H_{\tau\tau}^{SVfit}, MET)$
$P_T(H_{\tau\tau}^{MET})$	m_T^{total}	$P_t(H_{bb})$	$\Delta\phi(\tau_1, \tau_2)$
m_{top_1}	$P_T(H_{\tau\tau}^{vis})$	$MT2$	$P_T(H_{bb})$
$P_T(MET)$	$\cos\theta(b_1, H_{bb})$	$m(H_{\tau\tau}^{SVfit})$	m_T^{TOT}
$\Delta R(\tau_1, \tau_2) \cdot p_T(H_{\tau\tau}^{MET})$	$\mathbf{m(X^{vis})}$	$\Delta\phi(H_{bb}, MET)$	$m(X^{MET})$

Table 5.1: Input variables for the four different BDT trainings, reported in order of discriminating power between signal and background events. Variables highlighted in bold are common among the three resonant and the non-resonant training.

- $\theta_1(\tau_2, H_{\tau\tau})$ and $\theta_2(b_2, H_{bb})$ are the angles between the second lepton or jet and the direction of flight of $H_{\tau\tau}$ or H_{bb} in their rest frame.

- p_ζ and p_ζ^{vis} take into account the relative direction of flight of the two τ leptons and the MET, and are defined as:

$$p_\zeta = (\vec{p}_T(\tau_1)+\vec{p}_T(\tau_2)+\vec{p}_T^{miss})\cdot\hat{\zeta} \qquad p_\zeta^{vis} = (\vec{p}_T(\tau_1)+\vec{p}_T(\tau_2))\cdot\hat{\zeta} \quad (5.15)$$

  where $\hat{\zeta}$ is a unit vector oriented as the bisector of the $\vec{p}_T$ vectors of the two leptons.

The agreement between observed data and MC simulation for some of the BDT input variables is shown in Figure 5.2 for events selected in the $\tau_\mu\tau_h$ final state in the $1b1j$ category, *i.e.* only the leading b jet is required to pass the medium working point of the b-tagging discriminant. A more extensive set of BDT input variables is reported in Appendix A.

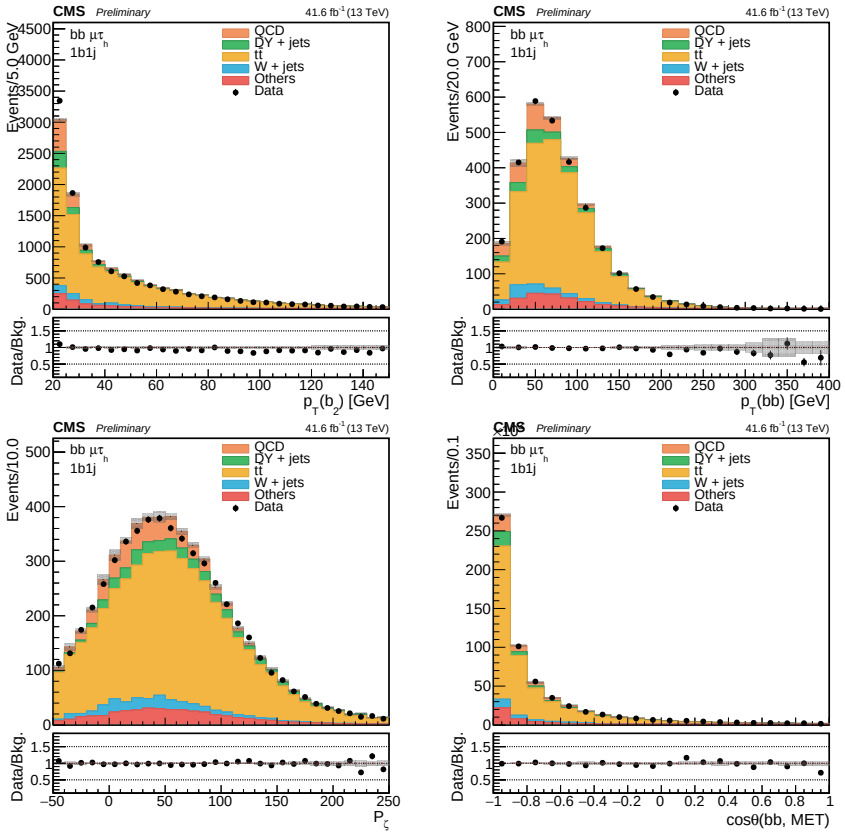


Figure 5.2: Distributions of some representative BDT input variables for events selected in the $\tau_\mu\tau_h$ channel in the $1b1j$ category. A good level of agreement between data and MC can be observed.

As previously described, in order to increase the number of events, signal samples with different mass and spin hypothesis are joined together in the

tree mass regions used to train the resonant BDTs. This merging, however, can in principle affect the performance of the BDTs themselves by deteriorating the discrimination power of the selected input variables. A parametrized learning is thus introduced by including in the trainings additional knowledge to keep track of the origin of the events. For each event, supplementary information is provided in the form of the mass and spin of the resonance, and the reconstructed final state of the $\tau\tau$ pair. The same considerations hold for the non-resonant BDT, where the $\tau\tau$ pair final state and the generated value of h_λ are provided to the training. The additional inputs are added to each list of selected variables described in Table 5.1, so that, even mantaining a single training, the discrimination of signals, whose kinematic properties vary depending on those parameters, is enhanced. Technically, the parameters variables are treated by the BDT as all the other inputs, but a proper reweighting of the events is needed not to assign too much importance to those signal hypotheses whose samples have a larger statistics.

As already done in the 2016 data analysis, the GradBoost algorithm is chosen to perform the trainings since it represents the most robust alternative against overtraining issues. Within the TMVA package, the BDT classifiers can be customized with several configuration options that can assume a wide variety of values and can affect the performance of the BDT classifiers themselves. In the parameters hyperspace more than 900 different points have been identified and tested in order to reject all those that lead to an overtrained BDT. This pruning process is performed through the comparison of the training and testing output distributions via the χ^2 test. Between the 100 point surviving this skimming procedure, the final set of parameters is chosen by evaluating the area under the ROC curve (AUC) since a larger value of AUC implies a better performance of the BTD. The BDT training parameters selected are reported in Table 5.2: $NTrees$ is the total number of trees that constitute the BDT "forest", $MaxDept = 3$ indicates the number of consecutive selections applied in each tree, while $MinNodeSize$ is the minimum percentage of training events required in a leaf node and it is fixed to 3% in this case. The number of grid points used to define the optimal cut in a variable range is denoted $nCuts$, while the $Shrinkage$ parameters defines the learning rate for the GradBoost algorithm. Finally, the "bagging" fraction ($BaggedSampleFraction$) defines the randomly chosen subset of events on which each tree is trained.

Some example distributions for the BDT output scores are illustrated in Figure 5.3. There is a general good agreement between observed data and MC simulation and, especially in the last bins close to 1, the $t\bar{t}$ background is suppressed. In addition, QCD multi-jet events are pushed to the lowest values of the BDT score, while $DY + jets$ contributions become dominant in the most sensitive bins.

The BDT output provided by the TMVA package is usually limited between -1 and 1, however, the conditions imposed by the classifier for events

Parameter	Value
NTrees	700
MaxDepth	3
MinNodeSize	0.03
nCuts	500
Shrinkage	0.05
BaggedSampleFraction	0.5

Table 5.2: Selected values of the parameter for the BDT training.

to be considered as signal or background can alter the shape of the BDT score so that the minimal or maximal values might be shifted inwards, as can be seen in Figure 5.3. In order to compare the distributions from different classifiers and with different *learning parameter* values (*i.e.* mass, spin, k_λ), the BDT *score* with values between min and max can be easily renormalized in the range from $min' = 0$ to $max' = 1$ with the formula:

$$score' = \frac{max' - min'}{max - min} \cdot (score - min) + min' \tag{5.16}$$

As this represents a simple mono-dimensional translation, for sake of simplicity, the plots in the following will be shown before such transformation.

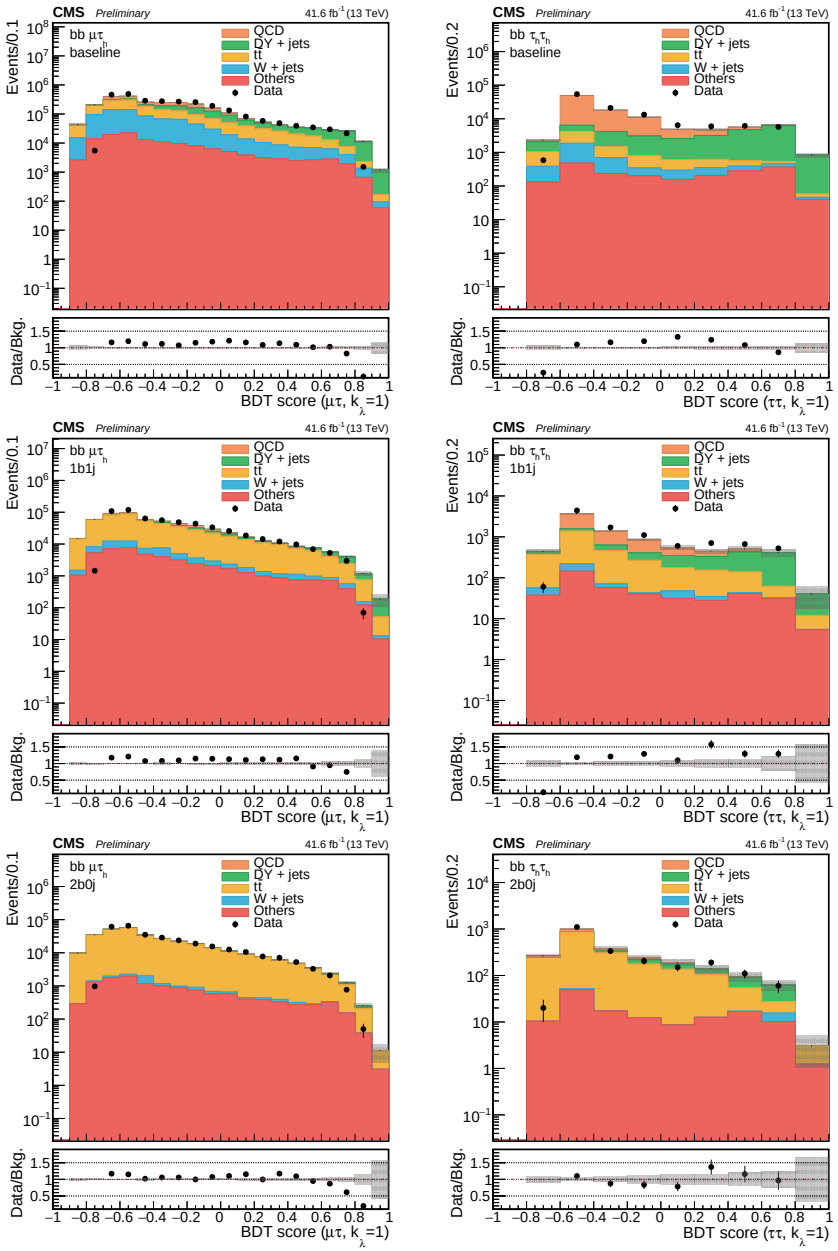


Figure 5.3: Distributions of some representative BDT output scores for the non-resonant training where the learning parameter k_λ is set to 1 (SM case). Events are selected in the $\tau_\mu\tau_h$ (left column) and $\tau_h\tau_h$ (right column) final states. Plots in the top row are obtained without any selection on the b tag score of the jets; in the central row only the leading jet is required to pass the b tag medium WP ($1b1j$ category), and in the bottom row both jets are required to be b-tagged ($2b0j$ category).

5.2 Statistical treatment

In order to asses the presence or absence of signal events in the selected final distributions, a statistical procedure is needed. In the context of Higgs searches, the ATLAS and CMS collaborations use a modified frequentist approach referred to as CL_s and defined in [72].

A binned maximum likelihood fit is performed on the discriminating variables described in Section 5.1. The expected event yield of the signal and the total background are denoted as s and b, respectively, and, in order to perform a model-independent search, the signal normalization is arbitrarily fixed to $\sigma \times \mathcal{B} = 1\ pb$ and scaled by a signal strength modifier μ. Given that the variables used are binned, s and b are vectors containing the yield expectations in each bin of the distributions.

The systematic uncertainties, described in Section 5.3, are included in this model as nuisance parameters θ_i, collectively denoted as θ, that affect the expected event yield for both signal and background processes which can thus be written as $s(\theta)$ and $b(\theta)$.

The likelihood function can be written as

$$\mathcal{L}(n, \tilde{\theta}|\mu, \theta) = P(n|\mu s + b) \cdot p(\tilde{\theta}|\theta) \tag{5.17}$$

where P denotes the probability density function of the observation of n events in a particular bin given by the sum of signal μs and background b expected events. For binned distributions, P is the product of the Poisson distributions for every bin:

$$P(n|\mu s + b) = \prod_j \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \tag{5.18}$$

The second term on the right side of Equation 5.17 represents the knowledge about the values of the nuisance parameters: each term reflects the probability for the true value to be equal to θ_i, given the best estimate $\tilde{\theta}_i$ obtained from auxiliary measurements on control region events or directly from the MC simulation. Since all systematic uncertainties are assumed to be uncorrelated, the combined term is the product of the single uncertainties $p(\tilde{\theta}|\theta) = \prod_i p_i(\tilde{\theta}_i|\theta_i)$.

The functional form of $p_i(\tilde{\theta}_i|\theta_i)$ depends on the type of uncertainty described. Uncertainties which arise from independent measurements, such as

luminosity or trigger efficiencies, are modeled with log-normal functions:

$$\rho(\theta) = \frac{1}{\sqrt{2\pi}ln(k)} exp\left(-\frac{(ln(\theta/\tilde{\theta}))^2}{2(lnk)^2}\right)\frac{1}{\theta} \tag{5.19}$$

where k is the parameter that defines the width of the log-normal distribution and thus represents the interval of possible variations of the observable. Systematic uncertainties that are of statistical origin instead, like the number of events observed in a control region, are represented by a Gamma distribution

$$\rho(n) = \frac{1}{\alpha}\frac{(n/\alpha)^N}{N!} exp(-n/\alpha) \tag{5.20}$$

where N is the number of events observed in the control region and α is the extrapolation factor by which the expected event yield in the signal region, n, is determined: $n = N \cdot \alpha$. The extrapolation factor is a nuisance parameter itself and is accounted for as an additional log-normal term.

Uncertainties on the template shapes are taken into account during the fit procedure, using the *Vertical Template Morphing* technique: for each quantity that affects the shape, multiple instances of the templates are produced from the simulated events by varying that quantity by $\pm 1\sigma$ and bin-by-bin interpolation is performed between them. A nuisance that represents the variation of such quantities from the nominal value, is added to the likelihood model.

In order to quantify whether the observed data supports the presence or absence of signal events, two hypotheses are tested for the signal plus background or background only cases, $H_{\mu s+b}$ and H_b respectively. To set an exclusion limit on the presence of a signal, one has to find the value of μ that allows to reject the $H_{\mu s+b}$ in favor of H_b.

The test statistic chosen to set the exclusion limit is the likelihood ratio:

$$q_\mu = -2ln\frac{\mathcal{L}(data|\mu, \hat{\theta}_\mu)}{\mathcal{L}(data|\hat{\mu}, \hat{\theta})} \qquad \text{with } 0 \leq \hat{\mu} \leq \mu \tag{5.21}$$

where $\hat{\theta}_\mu$ refers to the conditional maximum likelihood estimators of θ given the signal strength modifier μ, while "data" refers to the set of event yields n_i observed in all bins of the observed variables. The pair of parameter estimators $\hat{\mu}$ and $\hat{\theta}$ correspond to the global maximum of the likelihood defined in Equation 5.17. The lower constraint on $\hat{\mu}$ is dictated by physics (positive signal rate only), while the upper constraint is imposed by hand in order to guarantee a one-sided confidence interval. This definition of the test statistic implies that larger values of q_μ represent an increasing

incompatibility between the data and the hypothsized value of μ.

Given a signal strength modifier μ, the observed value q_μ^{obs} is obtained through the test statistic using the observed data n. In order to quantify the degree of compatibility of the observed data to the signal plus background or background only hypotheses, the probabilities for q_μ to be equal or larger than q_μ^{obs} are computed:

$$\begin{aligned} CL_{s+b}(\mu) &= P(q_\mu \geq q_\mu^{obs} | H_{\mu s+b}) \\ CL_b(\mu) &= P(q_\mu \geq q_\mu^{obs} | H_b) \end{aligned} \tag{5.22}$$

Finally, CL_s is computed, for the given value of μ under test, by the ratio of these probabilities:

$$CL_s(\mu) = \frac{CL_{s+b}(\mu)}{CL_b(\mu)} \tag{5.23}$$

A signal of strength μ is said to be excluded at a Confidence Level $1-\alpha$ if $CL_s(\mu) \leq \alpha$. In the $bb\tau\tau$ search we adopted $\alpha = 0.05$ and varied the parameter μ until the condition $CL_s(\mu) \leq \alpha$ is met so that the exclusion limits are always quoted with a 95% Confidence Level. The value of μ thus obtained is converted into a limit on $\sigma_{HH} \times \mathcal{B}(HH \rightarrow bb\tau\tau)$ by simply rescaling the signal normalization initially fixed.

5.3 Systematic uncertainties

Residual differences between data and MC simulation, due to uncertainties on theoretical predictions, on unforeseen detector responses as well as on statistical uncertainties affecting the data-driven methods, result in an imperfect knowledge of the modeling of signal and background processes. In order to properly account for all these effects they are included in the final fit as systematic uncertainties, described in the likelihood model by nuisance parameters, as described in Section 5.2.

Section 5.3.1 describes the so-called "normalization uncertainties" that only the yield of a given process, either signal or background, while in Section 5.3.2 the systematics affecting the differential distributions of the final discriminating variables ("shape uncertainties") are listed.

A summary of the systematic uncertainties considered together with their value and the processes affects is reported in Table 5.3.

5.3.1 Normalization Uncertainties

Luminosity
An uncertainty is applied to all MC simulated processes that are normalized by the total integrated luminosity collected. Since the normalization value is the same this uncertainty is considered fully correlated across all sample, channels and categories. The normalizations of the multi-jet and Drell-Yan backgrounds are obtained directly from data and thus are not subject to the luminosity uncertainty. The value is obtained through special Van Der Meer scans performed during the data taking period and it is measured to be 2.6% in 2016 [73] and 2.3% in 2017 [74].

Trigger, isolation and identification efficiencies
The trigger, identification and isolation efficiences of electron, muon and τ_h candidates are measured from $Z \to ee/\mu\mu/\tau_\mu\tau_h$ events as described in Section 3.2. The uncertainties related to these measurements are considered uncorrelated across channels since they are specific for each final state and values or 3%, 2% and 3% are measured for electrons, muons and τ_h, respectively.

Tau energy scale
An uncertainty coming from the τ energy scale knowledge is applied to all τ_h candidates. Different values are observed depending on the decay mode of the candidate and vary between 0.2 and 2.3%. A conservative approach is adopted by assuming a single value for the energy scale, varying its uncertainty by 3% and evaluating the changes in acceptance after the invariant mass selections: the overall impact on the analysis is between 3 and 10% depending on the process and channel considered. A shape uncertainty in the $MT2$ and m_{HH}^{KinFit} distributions is also defined and fully correlated to the normalization uncertainty.

Jet energy scale
As for the tau energy scale, also jet energy scale uncertainties are taken into account by measuring the changes in acceptance that occur when the selected jet energies are shifted inside the boundaries defined by the uncertainty and by estimating the possible induced changes in the process normalization. In the $bb\tau\tau$ analysis an inclusive uncertainty is applied that represents the combination in quadrature of 27 different sources that affect the jet energy scale. The effect of the single sources has been anyway studied in preparation for the combination of CMS HH searches, as detailed in Chapter 6.

b tagging scale factors
Uncertainties from the b tagging efficiencies as function of jet transverse

momentum and pseudorapidity are estimated by propagating the uncertainty on the MC-to-data scale factors and range from 2 to 6% for samples with true b jets in the final state.

Cross section
Uncertainties due to the imperfect knowledge of the normalization and simulation are considered for the MC simulated processes. The backgrounds affected are $t\bar{t}$, $W + jets$, single top, single Higgs and di-boson, with values ranging between 1 and 6%.

Data-driven techniques
The QCD background, estimated from data in a relaxed control region, is affected by statistical fluctuations of the number of events observed in the same-sign sidebands. The uncertainty is modeled with a Gamma function that depends on both the number of events observed and the $k^{OS/SS}$ extrapolation factor described in Section 4.2 and ranges from 5 to 30% depending on the channel and the category.

The uncertainties related to the three correction factors derived in control regions with 0, 1 or 2 b-tagged jets in order to correct the Drell-Yan $Z/\gamma^* \to \tau\tau$ contribution are propagated to the signal regions taking into account the covariance matrix that describes their correlations.

5.3.2 Shape Uncertainties

Tau and jet energy scales
Shape uncertainties are defined fully correlated to the normalization ones for the jet and τ_h candidates energy scales. Alternative shapes for the simulated processes are computed by varying the scale of the selected objects and considering the effect of the distributions of the final discriminating variables $MT2$ and m_{HH}^{KinFit}. Uncertainties on the energy scales of other objects are very small given the distributions binning chosen and their total impact is thus non taken into account.

Top quark p_T reweighting
To account for the residual differences in the transverse momentum distribution an event reweight technique is used to derive a systematic uncertainty. In $t\bar{t}$ events, the weight is computed as $w = \sqrt{SF(p_T^1) \cdot SF(p_T^2)}$ where $SF(p_T) = e^{a+bp_T}$, while the parameters a and b are provided from the Top PAG and correspond to 0.0615 and 0.0005, respectively. The nominal distribution shape is obtained when no reweighting is performed, while the alternative shape includes the weights and affects mostly the high mass tails of the distributions.

Low statistic bins
For any background process, if the ratio between the bin uncertainty and the bin content itself is larger than 10%, an additional shape uncertainty is added to the model by allowing to the bin to independently fluctuate around the observed value. Two shapes are created by shifting the bin content up and down accordingly to the bin uncertainty value. These uncertainties are denoted "bin-by-bin" (bbb) uncertainties and mainly affect the multi-jet process because of the statistical fluctuations observed in the data sidebands used to estimate the QCD background.

Uncertainty	Value	Background
Normalization		
Luminosity	2.5%	all except DY and QCD
Lept. trig., ID and isolation	$2-6\%$	all except QCD
τ energy scale	$3-10\%$	all except QCD
Jet energy scale	$2-4\%$	all except QCD
b tagging	$2-6\%$	all except QCD
Cross section	$1-6\%$	all except DY and QCD
DY scale factors	$0.1-2.5\%$	DY
Multi-jet	$5-30\%$	QCD
Shape		
τ and jet energy scale	-	all except QCD
Top p_T reweighting	-	$t\bar{t}$
Bin-by-bin	-	all

Table 5.3: Systematic uncertainties affecting the normalization or the shape of differential distributions. For each uncertainty the corresponding value and the processes to which it is applied is listed. DY and QCD represent the Drell-Yan $Z/\gamma^* \to \ell\ell$ and the mult-jets backgrounds, respectively.

5.4 2016 analysis results

In this Section, the results of the $bb\tau\tau$ analysis, obtained with data collected in 2016 at an energy of $\sqrt{s} = 13\ TeV$ and corresponding to and integrated luminosity of 35.9 fb^{-1}, are reported.

The non-resonant double Higgs production search is conducted both in the context of the Standard Model and in an effective Lagrangian framework where it is characterized by anomalous Higgs boson couplings. The production of two Higgs bosons through the decay of an heavy resonance instead is explored in the mass range between 250 and 900 GeV under the hypotheses of either a spin-0 or a spin-2 particle.

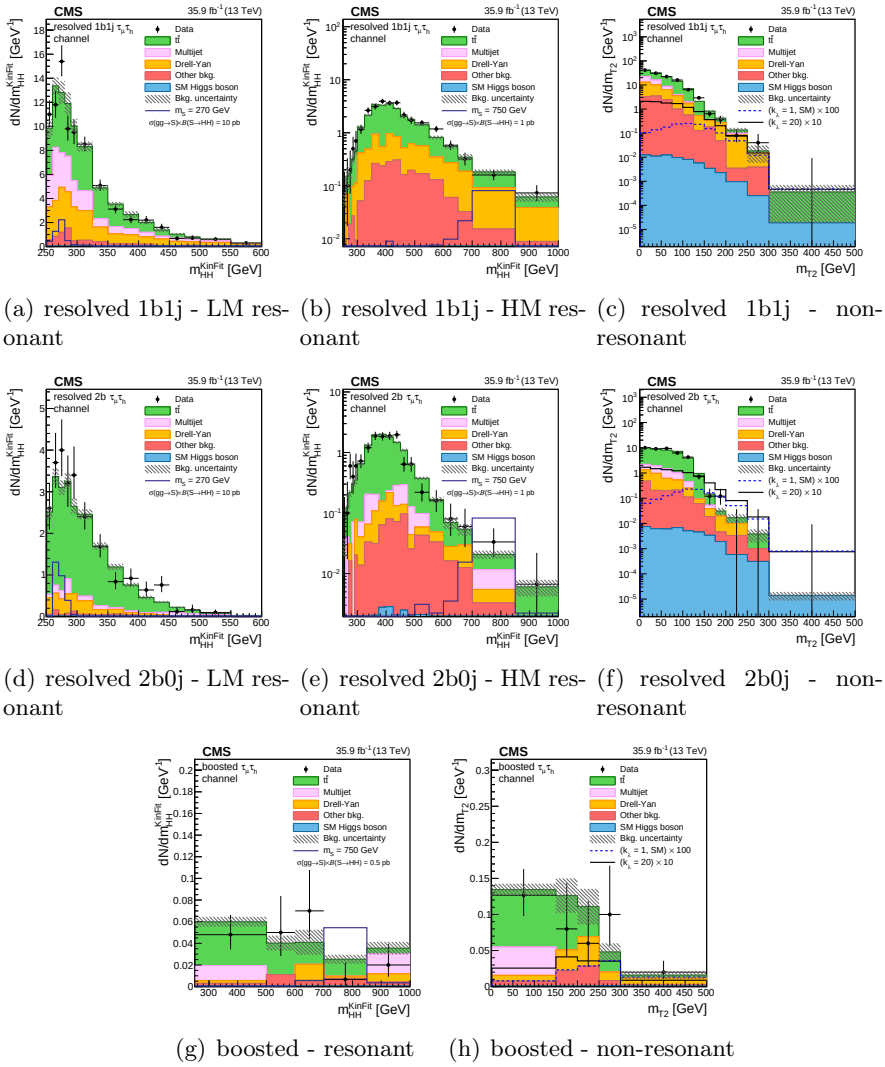
In both cases, the exclusion limits are derived starting from general assumptions on the signal kinematics so that a subsequent reinterpretation of the limits is possible in different specific BSM models.

5.4.1 Event yields and final distributions

Both resonant and non-resonant searches are performed in three final states ($\tau_\mu\tau_h$, $\tau_e\tau_h$ and $\tau_h\tau_h$) and three categories based on the b jet topologies. In case of the semileptonic channels two differents BDT, Low and High Mass, are applied to the signal hypotheses $m_X \leq 350\ GeV$ and $m_X > 350\ GeV$, respectively.

The number of expected and observed events in each category is summarized in Tables 5.4 and 5.6 for the $\tau_\mu\tau_h$ channel for the resonant and non-resonant searches, respectively. Tables 5.5 and 5.7 reports the numbers for the $\tau_e\tau_h$ channel, while for the $\tau_h\tau_h$ final state, since no BDT is used and thus the definition of the signal region is the same for both searches, the event yields are detailed in Table 5.8.

The distributions of $MT2$ and m_{HH}^{KinFit}, used to compute the event yields and later on to set the exclusion limits, are reported in Figures 5.4, 5.5 and 5.6 for the three channels.

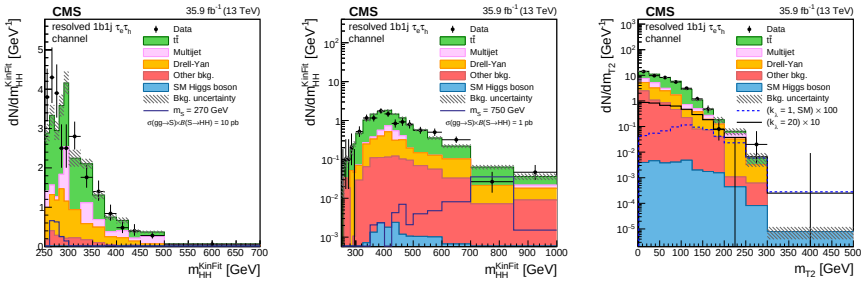


(a) resolved 1b1j - LM resonant (b) resolved 1b1j - HM resonant (c) resolved 1b1j - non-resonant

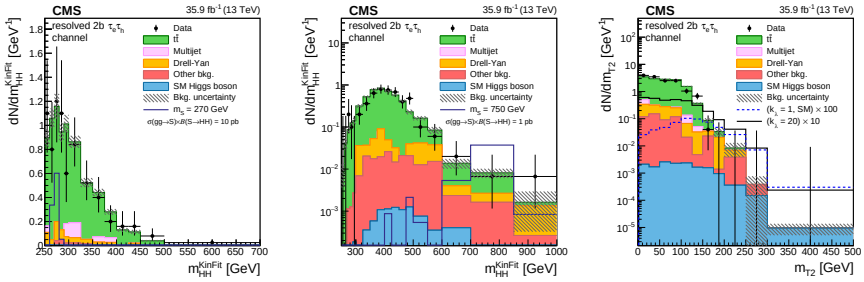
(d) resolved 2b0j - LM resonant (e) resolved 2b0j - HM resonant (f) resolved 2b0j - non-resonant

(g) boosted - resonant (h) boosted - non-resonant

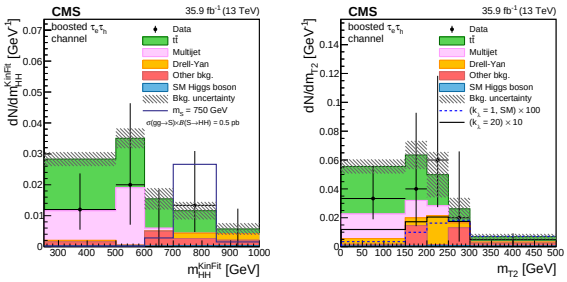
Figure 5.4: Distributions of the events observed in the signal regions of the $\tau_\mu\tau_h$ final state. The first, second, and third row show the $1b1j$, $2b0j$ and boosted regions respectively. Figures (a),(b),(d),(e),(g) show the distribution of the m_{HH}^{KinFit} variable and Figures (c),(f),(h) show the distribution of the $MT2$ variable. Points with error bars represent the observed data, while shaded histograms represent the backgrounds; finally, solid lines represent the expected signal yields and are not stacked to the background histograms. The dashed areas correspond to the systematic uncertainty band of the background estimates.



(a) resolved 1b1j - LM resonant (b) resolved 1b1j - HM resonant (c) resolved 1b1j - non-resonant

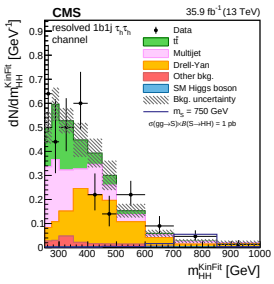


(d) resolved 2b0j - LM resonant (e) resolved 2b0j - HM resonant (f) resolved 2b0j - non-resonant

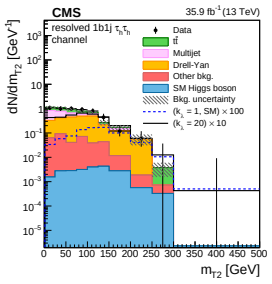


(g) boosted - resonant (h) boosted - non-resonant

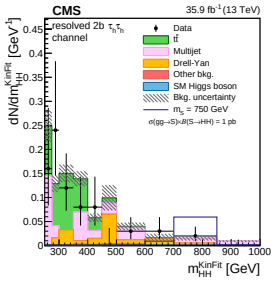
Figure 5.5: Distributions of the events observed in the signal regions of the $\tau_e\tau_h$ final state. The first, second, and third row show the $1b1j$, $2b0j$ and boosted regions respectively. Figures (a),(b),(d),(e),(g) show the distribution of the m_{HH}^{KinFit} variable and Figures (c),(f),(h) show the distribution of the $MT2$ variable. Points with error bars represent the observed data, while shaded histograms represent the backgrounds; finally, solid lines represent the expected signal yields and are not stacked to the background histograms. The dashed areas correspond to the systematic uncertainty band of the background estimates.



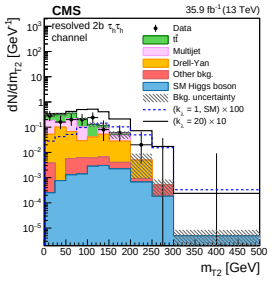
(a) resolved 1b1j - resonant



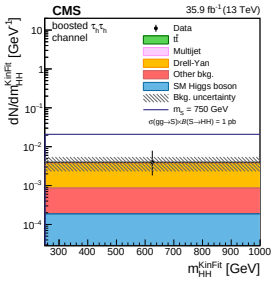
(b) resolved 1b1j - non-resonant



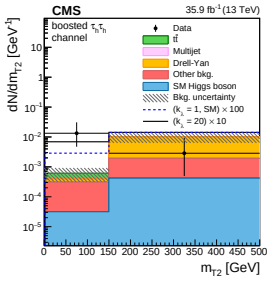
(c) resolved 2b0j - resonant



(d) resolved 2b0j - non-resonant



(e) boosted - resonant



(f) boosted - non-resonant

Figure 5.6: Distributions of the events observed in the signal regions of the $\tau_h\tau_h$ final state. The first, second, and third row show the $1b1j$, $2b0j$ and boosted regions respectively. Figures (a),(b),(d),(e) show the distribution of the m_{HH}^{KinFit} variable and Figures (c),(f) show the distribution of the $MT2$ variable. Points with error bars represent the observed data, while shaded histograms represent the backgrounds; finally, solid lines represent the expected signal yields and are not stacked to the background histograms. The dashed areas correspond to the systematic uncertainty band of the background estimates.

$\tau_\mu\tau_h$ final state - Resonant search					
Process	res. 1b1j		res. 2b0j		boosted
	LM	HM	LM	HM	
$t\bar{t}$	523.1 ± 19.2	507.4 ± 15.5	263.5 ± 11.2	267.1 ± 8.4	18.2 ± 1.0
QCD	266.2 ± 29.2	-	24.5 ± 2.7	19.0 ± 3.7	6.3 ± 1.6
Z+jets	373.8 ± 15.9	160.0 ± 6.8	40.8 ± 1.7	16.5 ± 1.0	3.7 ± 0.1
W+jets	45.0 ± 2.1	14.1 ± 1.3	1.5 ± 0.07	2.8 ± 0.1	0.76 ± 0.04
single top	38.3 ± 3.2	36.8 ± 2.5	7.6 ± 0.7	10.8 ± 0.7	2.3 ± 0.2
di-boson	7.5 ± 0.5	7.5 ± 0.6	1.5 ± 0.1	1.4 ± 0.1	0.75 ± 0.05
EWK W/Z	4.6 ± 0.2	5.1 ± 0.3	0.77 ± 0.04	0.85 ± 0.05	0.15 ± 0.01
SM Higgs	0.72 ± 0.04	0.97 ± 0.06	0.46 ± 0.02	0.68 ± 0.04	0.14 ± 0.01
Tot. exp. bkg.	1259 ± 39	732 ± 17	340 ± 12	319 ± 9	32.2 ± 1.9
Expected signal for $\sigma(gg \to X) \times \mathcal{B}(X \to HH \to bb\tau\tau) = 1$ pb					
$m_X = 300\ GeV$	59.6	11.5	47.3	10.2	0.6
$m_X = 600\ GeV$	38.6	264.2	36.5	237.1	55.4
$m_X = 900\ GeV$	23.0	176.3	12.2	127.9	419.6
Observed data	1252	782	363	318	28

Table 5.4: Observed and expected event yields in different signal regions of the resonant search for the $\tau_\mu\tau_h$ final state. Quoted uncertainties represent the combination of statistical and systematic uncertainties.

$\tau_e\tau_h$ final state - Resonant search					
Process	res. 1b1j		res. 2b0j		boosted
	LM	HM	LM	HM	
$t\bar{t}$	187.5 ± 6.8	227.4 ± 7.3	95.2 ± 4.0	118.7 ± 4.0	8.1 ± 0.4
QCD	62.7 ± 6.9	16.8 ± 3.3	6.8 ± 2.1	-	7.34 ± 2.2
Z+jets	106.7 ± 5.0	59.6 ± 2.2	8.2 ± 0.7	8.3 ± 0.4	0.69 ± 0.03
W+jets	10.4 ± 0.9	10.3 ± 1.1	0.029 ± 0.001	0.099 ± 0.004	0.45 ± 0.02
single top	14.6 ± 1.2	15.9 ± 1.2	2.2 ± 0.2	4.2 ± 0.4	0.68 ± 0.05
di-boson	3.7 ± 0.2	3.9 ± 0.4	0.56 ± 0.06	0.61 ± 0.06	0.27 ± 0.02
EWK W/Z	1.2 ± 0.1	0.63 ± 0.02	0.093 ± 0.004	0.43 ± 0.01	0.14 ± 0.01
SM Higgs	0.26 ± 0.01	0.48 ± 0.03	0.14 ± 0.01	0.29 ± 0.02	0.10 ± 0.01
Tot. exp. bkg.	387 ± 11	335 ± 9	113 ± 5	133 ± 4	17.7 ± 2.2
Expected signal for $\sigma(gg \to X) \times \mathcal{B}(X \to HH \to bb\tau\tau) = 1$ pb					
$m_X = 300\ GeV$	21.2	6.8	16.2	5.1	0.1
$m_X = 600\ GeV$	15.5	127.5	16.1	118.5	28.0
$m_X = 900\ GeV$	10.6	100.3	5.1	57.3	213.8
Observed data	388	316	114	123	7

Table 5.5: Observed and expected event yields in different signal regions of the resonant search for the $\tau_e\tau_h$ final state. Quoted uncertainties represent the combination of statistical and systematic uncertainties.

$\tau_\mu\tau_h$ final state - Non-resonant search			
Process	res. 1b1j	res. 2b0j	boosted
$t\bar{t}$	1617.6 ± 38.7	802.2 ± 22.4	20.0 ± 0.9
QCD	443.9 ± 38.2	80.9 ± 7.0	5.6 ± 1.9
Z+jets	629.6 ± 22.3	64.8 ± 2.9	7.1 ± 0.3
W+jets	124.7 ± 6.7	4.9 ± 0.2	0.95 ± 0.04
single top	121.9 ± 7.8	22.0 ± 1.5	2.5 ± 0.2
di-boson	18.3 ± 1.2	2.9 ± 0.3	0.89 ± 0.06
EWK W/Z	9.4 ± 0.5	1.2 ± 0.1	0.15 ± 0.01
SM Higgs	1.7 ± 0.1	1.1 ± 0.1	0.18 ± 0.01
Tot. exp. bkg.	2967 ± 60	980 ± 24	38 ± 2
Expected signal			
$k_\lambda = 1$ (SM)	0.38	0.33	0.08
$k_\lambda = 20$	25.75	20.88	1.12
Observed data	3020	996	35

Table 5.6: Observed and expected event yields in different signal regions of the non-resonant search for the $\tau_\mu\tau_h$ final state. Quoted uncertainties represent the combination of statistical and systematic uncertainties.

$\tau_e\tau_h$ final state - Non-resonant search			
Process	res. 1b1j	res. 2b0j	boosted
$t\bar{t}$	631.8 ± 16.3	311.1 ± 9.3	8.9 ± 0.4
QCD	135.9 ± 11.7	6.7 ± 2.1	6.5 ± 2.1
Z+jets	213.3 ± 7.0	20.2 ± 0.8	2.2 ± 0.1
W+jets	70.2 ± 3.2	0.42 ± 0.02	0.47 ± 0.02
single top	48.9 ± 3.2	10.5 ± 0.8	0.82 ± 0.05
di-boson	7.9 ± 0.5	1.1 ± 0.1	0.42 ± 0.03
EWK W/Z	3.3 ± 0.1	0.91 ± 0.03	0.33 ± 0.02
SM Higgs	0.69 ± 0.04	0.41 ± 0.03	0.12 ± 0.01
Tot. exp. bkg.	1112 ± 22	351 ± 10	19.7 ± 2.1
Expected signal			
$k_\lambda = 1$ (SM)	0.16	0.14	0.04
$k_\lambda = 20$	10.28	8.26	0.55
Observed data	1057	355	11

Table 5.7: Observed and expected event yields in different signal regions of the non-resonant search for the $\tau_e\tau_h$ final state. Quoted uncertainties represent the combination of statistical and systematic uncertainties.

$\tau_h\tau_h$ final state - Resonant and non-resonant searches			
Process	res. 1b1j	res. 2b0j	boosted
$t\bar{t}$	33.6 ± 1.5	16.5 ± 1.1	0.068 ± 0.004
QCD	40.6 ± 7.9	14.5 ± 2.8	0.012 ± 0.012
Z+jets	48.7 ± 6.2	9.1 ± 1.0	2.2 ± 0.1
W+jets	1.11 ± 0.06	-	0.031 ± 0.002
single top	4.2 ± 0.3	0.026 ± 0.002	-
di-boson	2.3 ± 0.4	0.57 ± 0.08	0.33 ± 0.03
EWK W/Z	0.78 ± 0.04	-	0.15 ± 0.01
SM Higgs	0.63 ± 0.08	0.38 ± 0.05	0.14 ± 0.01
Tot. exp. bkg.	132 ± 10	41 ± 3	2.9 ± 0.1
Expected signal for $\sigma(gg \to X) \times \mathcal{B}(X \to HH \to bb\tau\tau) = 1$ pb			
$m_X = 300\ GeV$	20.48	15.03	0.08
$m_X = 600\ GeV$	185.27	165.44	40.51
$m_X = 900\ GeV$	126.17	105.13	379.10
$k_\lambda = 1$ (SM)	0.24	0.21	0.05
$k_\lambda = 20$	9.20	7.88	0.60
Observed data	140	33	3

Table 5.8: Observed and expected event yields in different signal regions of the $\tau_h\tau_h$ final state. Quoted uncertainties represent the combination of statistical and systematic uncertainties.

5.4.2 Exclusion limits

No evidence for the presence of signal events is found in the channels and categories considered, neither in the kinematic fit (m_{HH}^{KinFit}) nor in the $MT2$ case, thus, the distributions are used to set upper exclusion limits at 95% confidence level on $\sigma(gg \to HH) \times \mathcal{B}(HH \to bb\tau\tau)$.

Resonant production

In the resonant search the exclusion limits are set on the production cross section of the resonance times the branching fraction of the decay of the resonance itself into two Higgs boson ($\sigma(gg \to X) \times \mathcal{B}(X \to HH)$), as a function of the mass m_X.

The limits, for the combination of all channels and categories considered, are shown in Figure 5.7 under the radion (spin-0) and graviton (spin-2) hypotheses. The values of $\sigma \times \mathcal{B}$ excluded vary from 500 to 5 pb depending on the resonance mass.

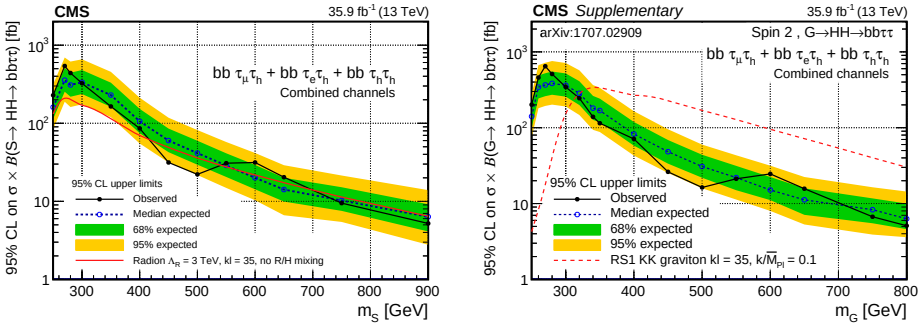


Figure 5.7: 95% CL upper limits on $\sigma(gg \to X) \times \mathcal{B}(X \to HH)$ for a spin-0 (left) and spin-2 (right) resonance. The green and yellow bands indicate the regions containing 68 and 95% of the distribution of limits expected under the background-only hypothesis. The red curves indicate the theoretical prediction for the production of a radion (spin-0) or of a graviton (spin-2) decaying to a HH pair [1] [75].

Figure 5.8 reports the separate contribution of each channel and category in the case of a spin-0 category. Thanks to the higher signal purity, the $\tau_h\tau_h$ final state has the best sensitivity for $m_X > 300\ GeV$, while the semileptonic channels, having lower p_T thresholds and thus larger acceptance, are more sensitive for lower values of m_X. As expected, the resolved $2b0j$ is the dominant category for masses up to $\sim 700\ GeV$ when the fraction of events containing boosted b jets becomes dominant and the boosted category achieves best results.

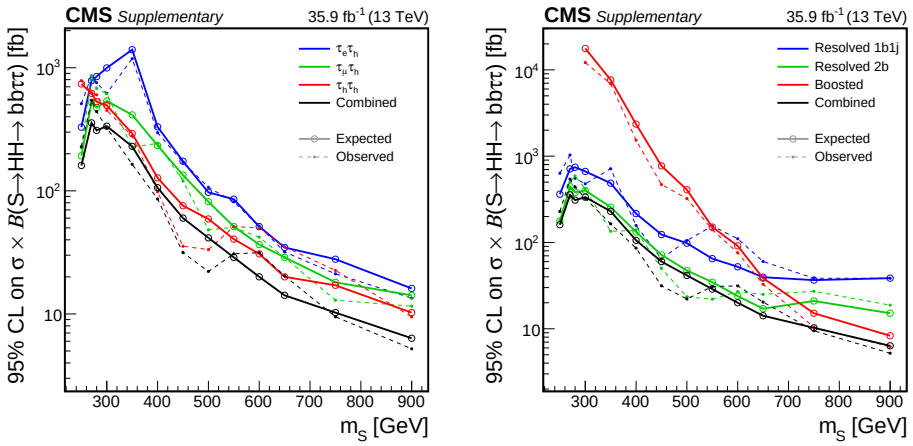


Figure 5.8: Comparison of the 95% CL upper limits separately for the three final states (left) and for the three categories (right) [75].

These results are reinterpreted in the context of a specific supersymmetric model, the so-called $hMSSM$. As described in Chapter 1, in models with two Higgs doublets, three physical neutral Higgs bosons are predicted: h, assumed to be the 125 GeV scalar boson observed at the LHC; H, a heavy CP-even scalar assumed to be the heavy resonance object of this search; A, a heavy CP-odd scalar. The observed results are tested against the model predictions and a portion of the parameter space, corresponding to 230 $GeV < m_A < 360$ GeV and $\tan\beta \leq 2.5$ is excluded at 95% CL, as shown in Figure 5.9.

Non-Resonant production

The non-resonant HH production can be parametrized, in an EFT context, with the five Higgs boson couplings described in Chapter 1. Two sets of results are derived: in the first case assuming as function of the ratio k_λ/k_t with $c_2 = c_g = c_{2g} = 0$, in the second one an exclusion limit is set for each of the 12 EFT benchmarks.

Figure 5.10 displays the 95% CL exclusion limit on $\sigma(gg \rightarrow HH) \times \mathcal{B}(HH \rightarrow bb\tau\tau)$ as function of the ratio k_λ/k_t. The observed constraints on k_λ, assuming all the other couplings to be $k_t = 1$ and $c_2 = c_g = c_{2g} = 0$, are $-18 < k_\lambda < 26$, for the expected ones being $-14 < k_\lambda < 22$. The observed exclusion limit in the Standard Model case ($k_\lambda = 1$) is $\sigma^{SM}_{HH} \times \mathcal{B}(HH \rightarrow bb\tau\tau) \leq 75.4\ fb$, while the expected limit is $\leq 61\ fb$. These values correspond to about 30 and 25 times the SM prediction, respectively.

Assuming $c_2 = c_g = c_{2g} = 0$, double Higgs production is completely determined by the interference between the "triangle" and the "box" diagrams

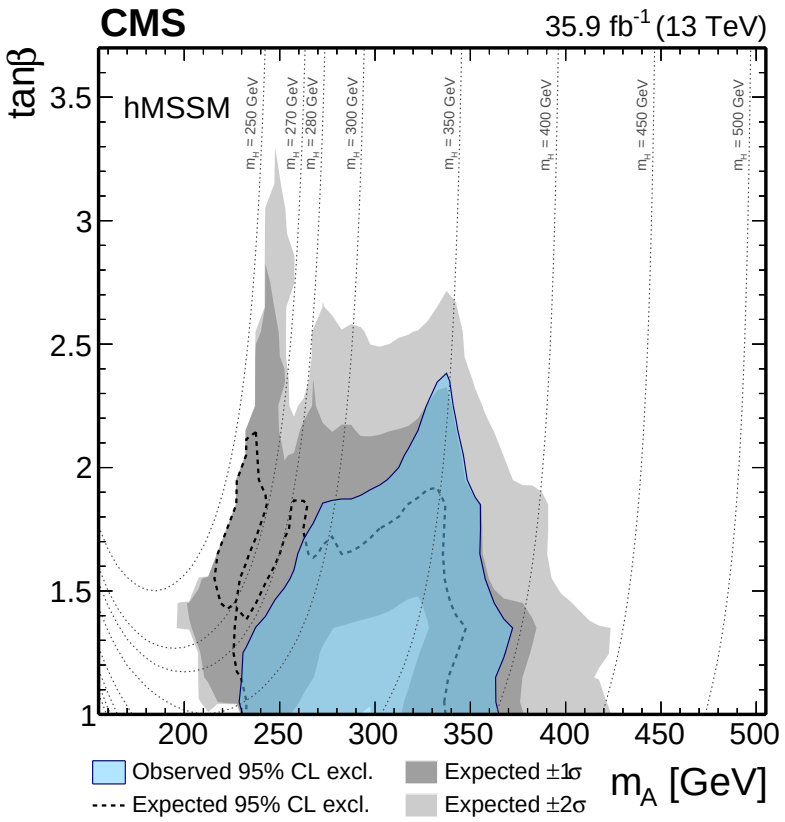


Figure 5.9: Interpretation of the exclusion limit in the context of the hMSSM model. The blue transparent curve denotes the region of the $tan\beta$ and m_A parameters space excluded by the observation, while the dashed line and the grey bands denote the expected exclusion and its associated 68 and 95% exclusion intervals. The dotted lines indicate trajectories in the plane corresponding to equal values of the mass of the CP-even heavier scalar of the model m_H [1].

described in Section 1.3. This gives rise to the peculiar shape obtained for the limit as different values of k_λ/k_t are investigated. At the edges of the distribution in Figure 5.10, where $|k_\lambda/k_t| \gg 1$, the "triangle" diagram becomes dominant and the limits asymptotically tend to the same vale. On the opposite, towards the value of maximum interference ($k_\lambda/k_t = 2.46$), even small modification of the parameters are responsible for profound changes in the event kinematic, giving rise to the sharp feature in the limit plot.

The contributions to the exclusion limit of the single final states and categories are shown in Figure 5.11. As it was the case in the resonant production, also in the non-resonant analysis the $\tau_h\tau_h$ channel and the resolved $2b0j$ category are the most sensitive.

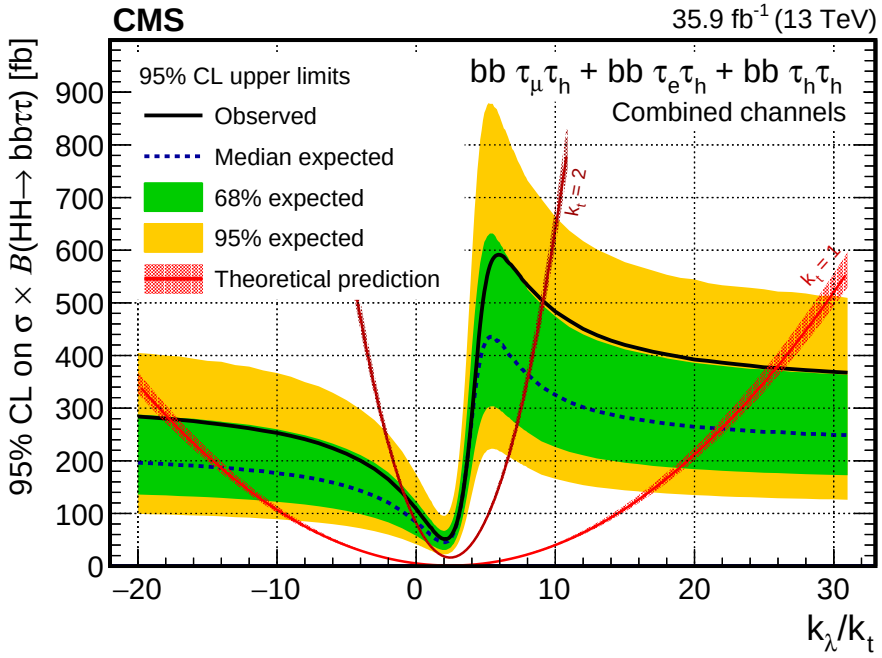


Figure 5.10: 95% CL confidence level upper limits on $\sigma(gg \rightarrow HH) \times \mathcal{B}(HH \rightarrow bb\tau\tau)$ for the combination of the three final states and three event categories. The two red lines denote the theoretical cross section times the branching fraction for a value of k_t of 1 (SM prediction) and 2 [1].

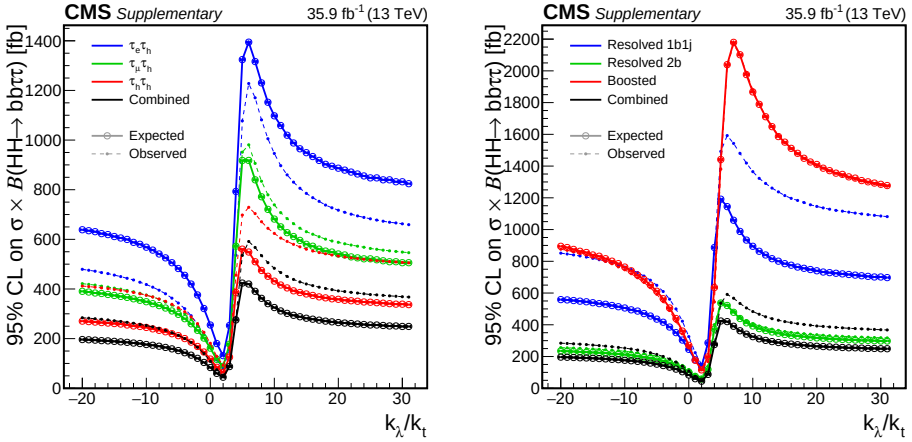


Figure 5.11: Comparison of the 95% CL expected upper limits as a function of the ratio k_λ/k_t , separately shown for the three final states (left) and the three event categories (right) [75].

The results can be also plotted in the bi-dimensional phase space of the two parameters as simultaneous exclusion of k_λ e k_t values, as shown in Figure 5.12.

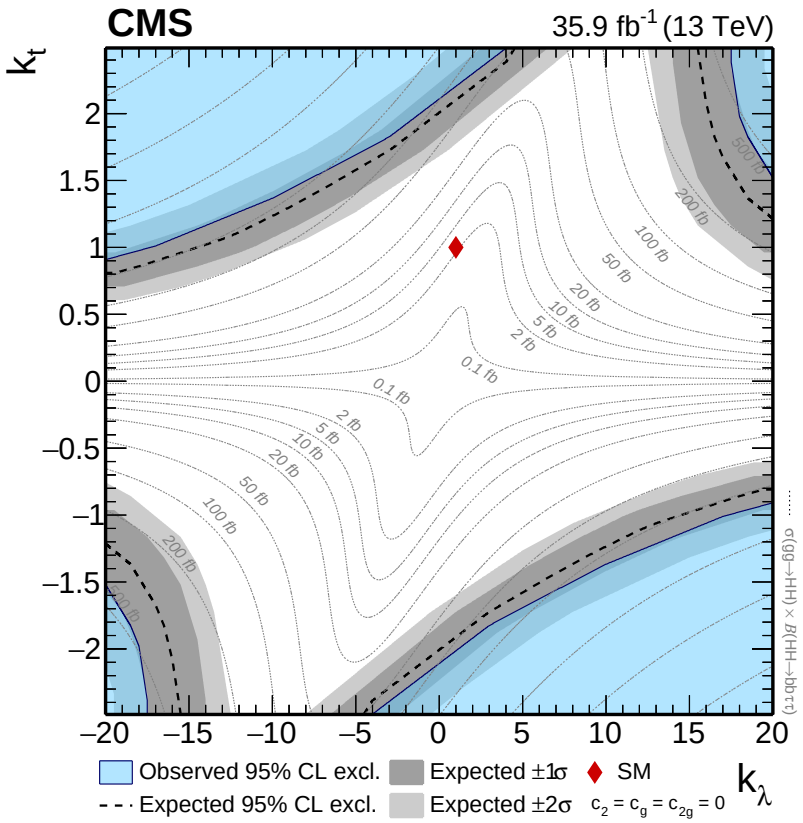


Figure 5.12: Exclusion limits in the k_λ and k_t plane. The blue region denotes the parameters excluded at 95% CL with the observed data, while the dashed black line and the grey regions denote the expected exclusions and the 1 and 2σ bands. The dotted lines indicate trajectories in the plane with equal values of cross section times branching fraction that are displayed in the associated labels. The SM couplings, corresponding to $k_\lambda = k_t = 1$, are indicated by the red diamond-shaped marker [1].

Finally, Figure 5.13 reports the exclusion limits set on $\sigma \times \mathcal{B}$ for the twelve EFT benchmarks described in Section 1.3.1. The different values for the limits, obtained in each scenario, are a direct consequence of the variety of kinematic properties that is generated by different assumptions of the Higgs boson couplings.

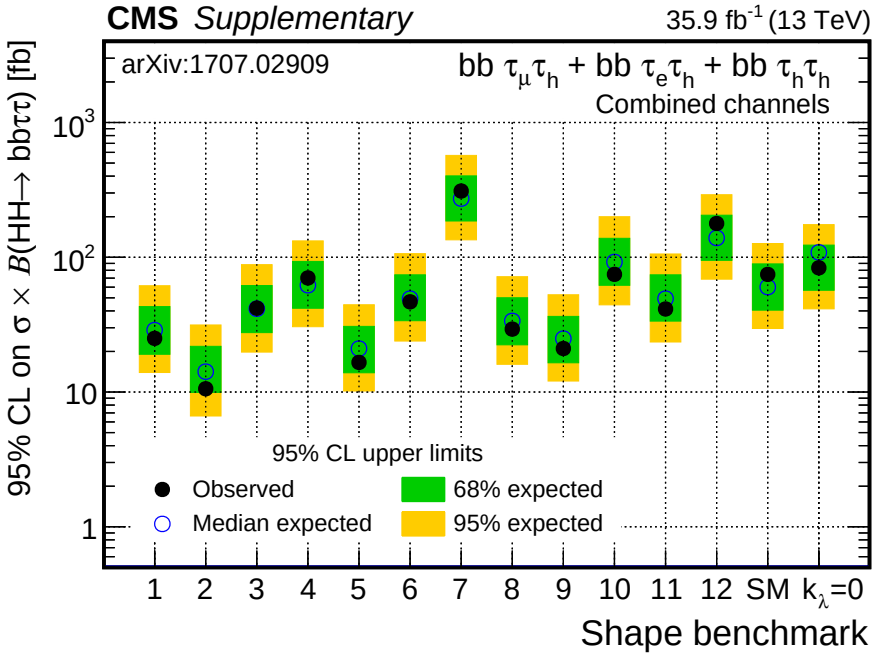


Figure 5.13: 95% CL upper limits on $\sigma(gg \to HH) \times \mathcal{B}(HH \to bb\tau\tau)$ for the twelve EFT shape benchmarks, the SM and the $k_\lambda = 0$ signals [75].

5.5 2017 analysis

The final analysis strategy for the legacy paper, including all the Run II data, is not yet completely defined. Nevertheless, the study of 2017 data allows to design and develop new techniques and algorithms to improve the search.

In order to asses the presence of signal events or, in case of their absence, set the 95% confidence level exclusion limits, the CL_s method, described in Section 5.2, is commonly used by CMS physics analysis. However, its full computation is quite computing expensive, especially in cases like the $HH \rightarrow bb\tau\tau$ analysis where many different signal hypotheses are tested against the observed data. In addition, a complete knowledge of all the uncertainties and their effect on the processes is necessary to obtain proper exclusion limits. Despite some of the systematics for the 2017 analysis are already known (*e.g.* the uncertainty on the integrated luminosity collected), many more are still to be properly evaluated: in particular regarding the energy scales of tau leptons and jets, which are known from the 2016 analysis to be the among the most influential.

Thus, in order to give an estimate of the improvements coming from the new tools described in this thesis, I will adopt as figure of merit for the sensitivity the estimator Z_A described in Section 5.5.1. Section 5.5.2 describes the performances of the new BDT discriminant developed for the case of gluon fusion searches and described in Section 5.1.3 of this Chapter, while Section 5.5.3 contains some considerations on the VBF selections performances. Finally, in Section 5.6 I report some suggestion on how to further optimize these new techiniques and the analysis strategy.

5.5.1 Sensitivity estimators

In particle physics, the quantity $s/\sqrt{b}$ has been used to measure the expected discovery significance. For a process with Poisson distributed events, the likelihood function can be written

$$\mathcal{L}(s) = \frac{(s+b)^n}{n!} e^{-(s+b)} \tag{5.24}$$

Using the test statistic q_0 defined in Equation 5.21 and the Wilks' theorm [76], the significance can be approximated in this case as:

$$Z = \sqrt{q_0} = \sqrt{2\left(n \ln \frac{n}{b} + b - n\right)} \tag{5.25}$$

Using the Asimov dataset, *i.e.* substituting the number of observed data with the expected $s + b$, the significance becomes:

$$Z_A = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right) - s\right)} \tag{5.26}$$

which, when expanding the logarithms in s/b, returns the aforementioned formula:

$$Z_A = \frac{s}{\sqrt{b}}(1 + \mathcal{O}(s/b)) \tag{5.27}$$

As shown in Figure 5.14, the formula $s/\sqrt{b}$ is usually a good approximation of the sensitivity obtained with the full statistical CL_s method described in Section 5.2 of this Chapter. Even if the results thus obtained can't be used to set proper exclusion limits at 95% of Confidence Level, being much less computing intensive, the $s/\sqrt{b}$ approximation represents an optimal approach to design algorithms and optimize the selections to improve the sensitivity of the analysis.

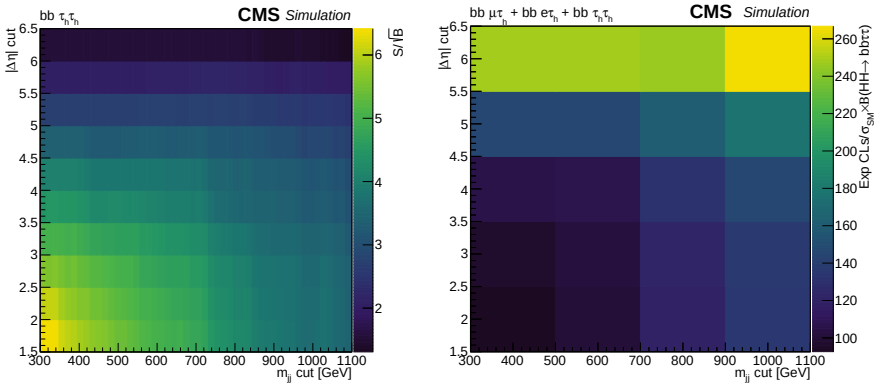


Figure 5.14: Comparison of the sensitivity to a 2016 gluon fusion HH signal sample as function of the selection applied to m_{jj} and $\Delta\eta_{jj}$. On the left the sensitivity is obtained through the approximation $s/\sqrt{b}$, on the right the 95% CL upper limits on $\sigma \times \mathcal{B}(HH \to bb\tau\tau)$. The estimate given by the method is the same as higher sensitivity is indicated by high values of $s/\sqrt{b}$ (left plot) and lower values of CL_s (right plot). In this specific case, given the fact that a gluon fusion sample is considered as signal, the two additional jets, to which the selections (m_{jj} and $\Delta\eta_{jj}$) are applied, originate mainly from the parton shower process or from pileup contributions and are thus focused at low values. When the selections are tightened, more and more events are rejected and the sensitivity drops.

As demonstrated in [77] the $s/\sqrt{b}$ approximation holds when $s \ll b$. Moreover, if the expected number of background events, b, is not known one must treat it as a nuisance parameter in the likelihood function. Since b could be adjusted to accommodate any observed number of events, it would be impossible to reject the background-only hypothesis unless some additional information is introduced that constrains b. Usually this is done by means of a control measurement by counting the number of events in a data sample where signal events are believed to be absent ("control region"), and where the mean number of events can be related to the number of background events in the signal region.

This turns out to be exactly the case of the $HH \to bb\tau\tau$ analysis, where the QCD multi-jet background is estimated from control regions (Section 4.2) and where in the final signal regions of some categories, such as the $2b0j$ and the *boosted* ones, the expected backgrounds yield is also very small.

Following [77], the signal significance can be modified to to account both the statistical and systematic error on b (σ_b):

$$Z_A = \left[2\left((s+b)\ln\left[\frac{(s+b)(b+\sigma_b^2)}{b^2+(s+b)\sigma_b^2}\right] - \frac{b^2}{\sigma_b^2}\ln\left[1+\frac{\sigma_b^2 s}{b(b+\sigma_b^2)}\right]\right)\right]^{\frac{1}{2}} \quad (5.28)$$

Figure 5.15 shows that both methods agree with the Monte Carlo values for sufficiently large values of b, but the formula defined in Equation 5.28 is clearly in far better agreement for low b.

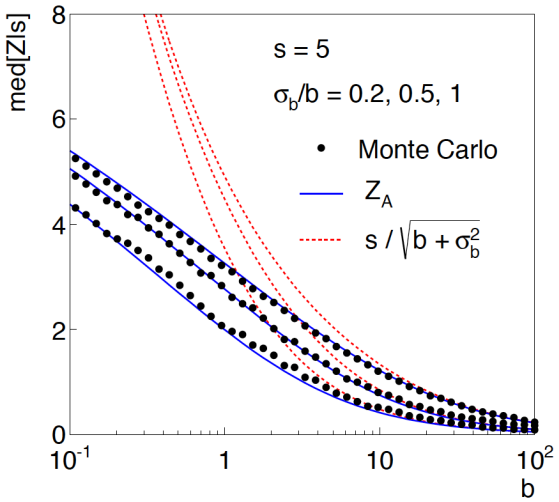


Figure 5.15: Comparison, as function of b, of the median discovery significance for $s = 5$ obtained both with the $s/\sqrt{b}$ approximation and with Z_A. Different values of σ_b/b are shown: the upper set of curves (points) corresponds to the smaller σ_b/b [77].

In Sections 5.5.2 and 5.5.3 thus, the expected significance Z_A defined in Equation 5.28 will be used to evaluate the performances of the BDT developed in 2017 for the signal extraction (Section 5.1.3) and of the VBF selections described in Section 3.3.3.

5.5.2 Performanced of the gluon fusion BDT

The BDT discriminant described in Section 5.1.3 has been developed to maximize the discrimination of $bb\tau\tau$ events originating from the decay of a HH pair from those initiated by other background processes. A parametrized learning is introduced in the training depending on the characteristic properties of the signals investigated: the mass m_X for the resonant case, and the Higgs boson self-coupling strength modifier k_λ for the non-resonant one.

As an example, Figure 5.16 shows how this approach proves to be an ideal choice when investigating different signal hypothesis. The BDT output score distributions are displayed for two out of the three mass regimes investigated: Low Mass regime with learning parameter $m_X = 280\ GeV$ on the left and High Mass regime with learning parameter $m_X = 650\ GeV$ on the right. In both plots the distributions for three signals with different resonance masses (280, 400 and 750 GeV), are shown superimposed to the backgrounds stack. It is evident how the introduction of parametrized learning helps the BDT in discriminating the signal events that correctly match the input parameter by assigning them a score closer to +1.

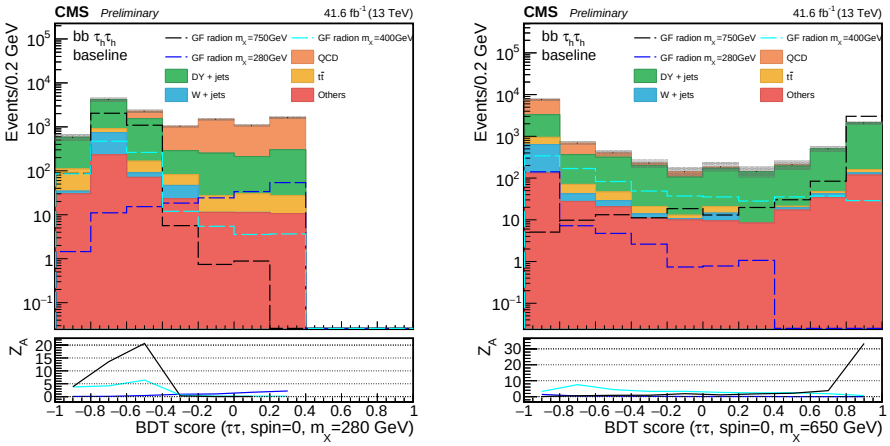


Figure 5.16: Comparison of the BDT score distributions for three resonant signals (radion 280, 400 and 750 GeV). The left plot shows the BDT output from a training with learning parameter $m_X = 280\ GeV$, the right one shows the BDT score obtained with learning parameter $m_X = 650\ GeV$.

Usually two different approaches are possible in order to exploit the dis-

crimination power offered by the BDT: either selecting only the high tail of the output distribution, thus removing the background dominated low tail ("cut-based approach"), or feeding the full output distribution to the likelihood fit described in Section 5.2. The former case is usually employed when the final discriminating distributions are dominated by a specific background that can be removed with a selection on the BDT score, this was the case of the BDT developed for $t\bar{t}$ rejection in the 2016 analysis. In the latter case instead, the full BDT output distributions is fed to the likelihood model and used as discriminating variable: the main advantage of this approach lays the fact that the signal events that have a low BDT score aren't exluded from the analysis, thus increasing the signal acceptance. This second approach is the one adopted for the development of the 2017 BDT.

In order to qualitatively evaluate its effectiveness in terms of sensitivity, the estimator Z_A can be computed in each bin of the distribution and compared to the same value computed for the variables used in the 2016 analysis. As an example, this comparison is illustrated in Figure 5.17 for the non-resonant case. Two signal hypotheses are displayed: the SM gluon fusion (black line), with parameters $k_\lambda = 1, k_t = 1, c_2 = c_g = c_{2g} = 0$, and the EFT beanchmark 7 (blue line), with parameters $k_\lambda = 5, k_t = 1, c_2 = 0, c_g = 0.2, c_{2g} = -0.2$. The distributions show events selected in the $\tau_h\tau_h$ channel in two regions, one where no selection on the jets b-tagger is applied, and the other where both jets are required to pass the medium b-tag working point. It can be noticed that, as expected, in both the BDT and the $MT2$ distributions the sensitivity increases when moving to the resolved $2b0j$ category. Furthermore, the Z_A values computed from the BDT score distributions are higher than those obtained from the $MT2$ distribution, which is indicative of an increase in the sensitivity to $HH \to bb\tau\tau$ events.

5.5.3 Performanced of VBF selections

As described in Section 3.3.3, when defining the "VBF region", the easiest way to get rid of background contributions without spoiling the signal acceptance is to apply a selection on the invariant mass and on the spatial separation of the two jets identified as VBF jets. Figure 5.18 illustrates the sensitivity, evaluated with Z_A, as a function of the cuts applied on m_{jj} and $\Delta\eta_{jj}$. The plot is obtained using a SM double Higgs VBF sample as signal and the sum of all other the processes as background. As expected from the VBF event kinematics, the sensitivity increases when the selections are tightened until a certain value is reached (around $\Delta\eta_{jj} \sim 6$ or $m_{jj} \sim 1300\ GeV$), after this threshold the statistics becomes too small also for the signal and the sensitivity drops.

Once most of the backgrounds are rejected, a BDT is applied to discriminate $bb\tau\tau$ events that originate in vector boson fusion processes from those

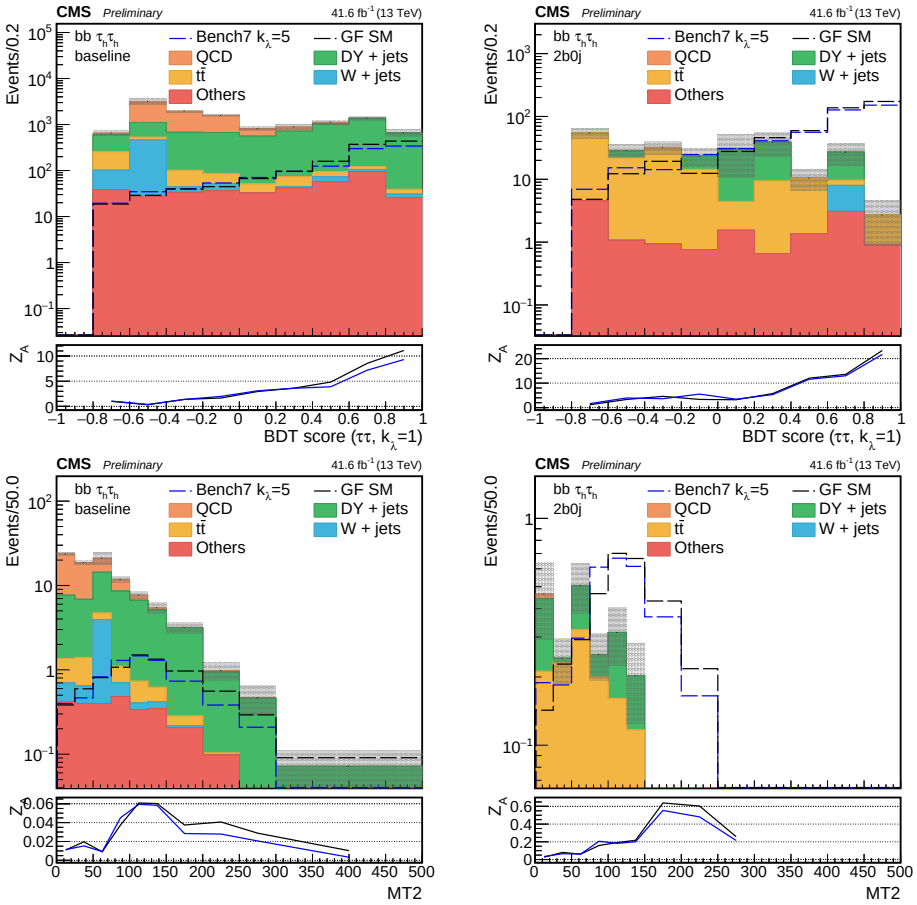


Figure 5.17: Comparison of the Z_A estimator values computed in each bin of the distributions. The upper plots report the BDT output obtained for a training with learning parameter $k_\lambda = 1$ (SM case), the lower plots show the $MT2$ variable which was used to set the exclusion limits in the 2016 analysis. Left distributions contain events selected without any requirement on the b-tagging of the jets, while right distributions show the resolved $2b0j$ category.

initiated by gluon fusion. The distribution of the BDT score is shown in Figure 5.19 for events selected in the $\tau_h\tau_h$ channel. A good agreement, although not yet optimal, is observed between the data and the MC simulation.

To evaluate its performance, the BDT score distributions for the backgrounds and for some signals are reported in the top plot of Figure 5.20 for events in the $\tau_h\tau_h$ channels selected after the invariant mass cuts on $m_{\tau\tau}^{SVfit}$ and m_{bb}. As an example, three different signals are considered, one resonant and two non resonant. The former is the VBF production of a spin-0 radion ($m_X = 280\ GeV$) decaying in a HH pair, while the latter are two

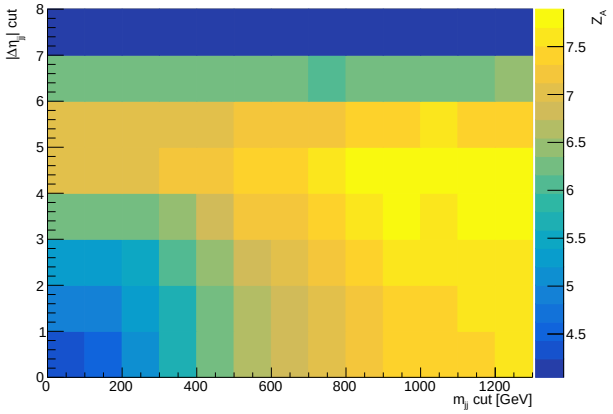


Figure 5.18: Distribution of Z_A as function of the cuts applied to m_{jj} and $\Delta\eta_{jj}$ and used to define a "VBF region".

non-resonant VBF samples: one is produce accordingly to the predictions of the Standard model, the other represents a BSM model where the self coupling of the Higgs boson (k_λ) is set to zero.

The same signals, with the addition of the heavy VBF radion ($m_X = 750\ GeV$) are used in the bottom plot of Figure 5.20, where the sensitivity, estimated with Z_A, is illustrated as function of the cut on the BDT score. While for the non-resonant samples the sensitivity increases as expected for tighter selections, for the resonant signals the sensitivity decreases with tighter selections: this reflects the fact that the BDT was trained using non-resonant samples only and its application to radion and graviton searches may not be the the optimal choice.

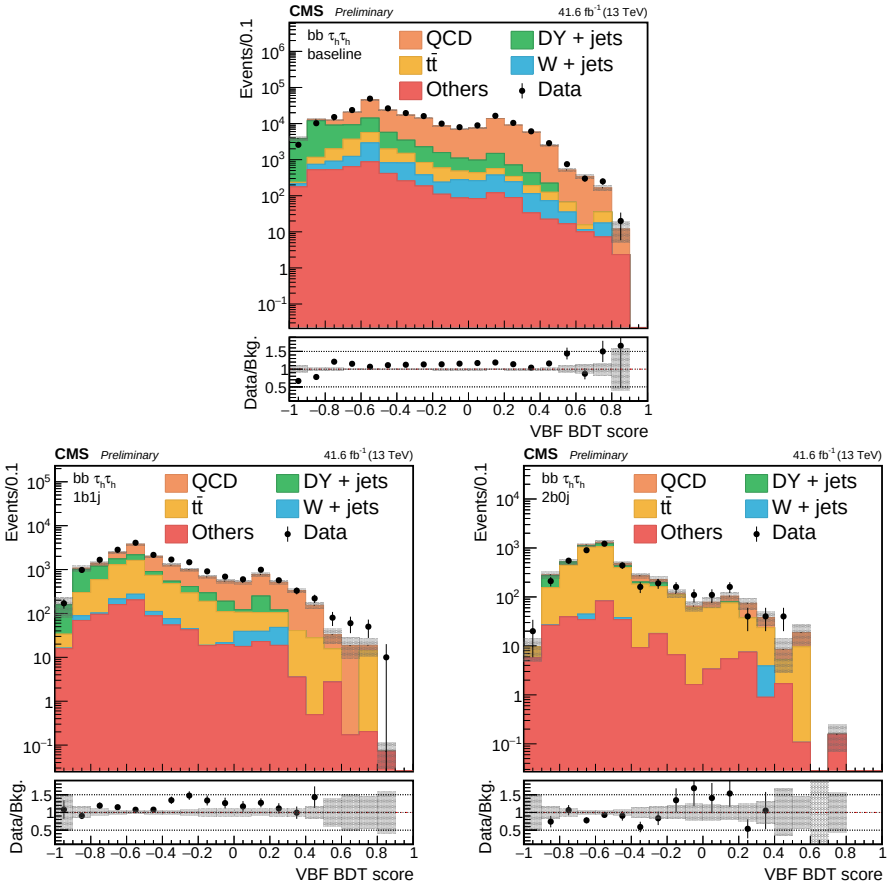


Figure 5.19: Distributions of the VBF BDT output scores for events in the $\tau_h\tau_h$ channel. In the top row no selection on the b-tag of the jets is applied, on the bottom row in the left plot only one jet is required to pass the medium b tag working point, while in the right one both jets must be b tagged.

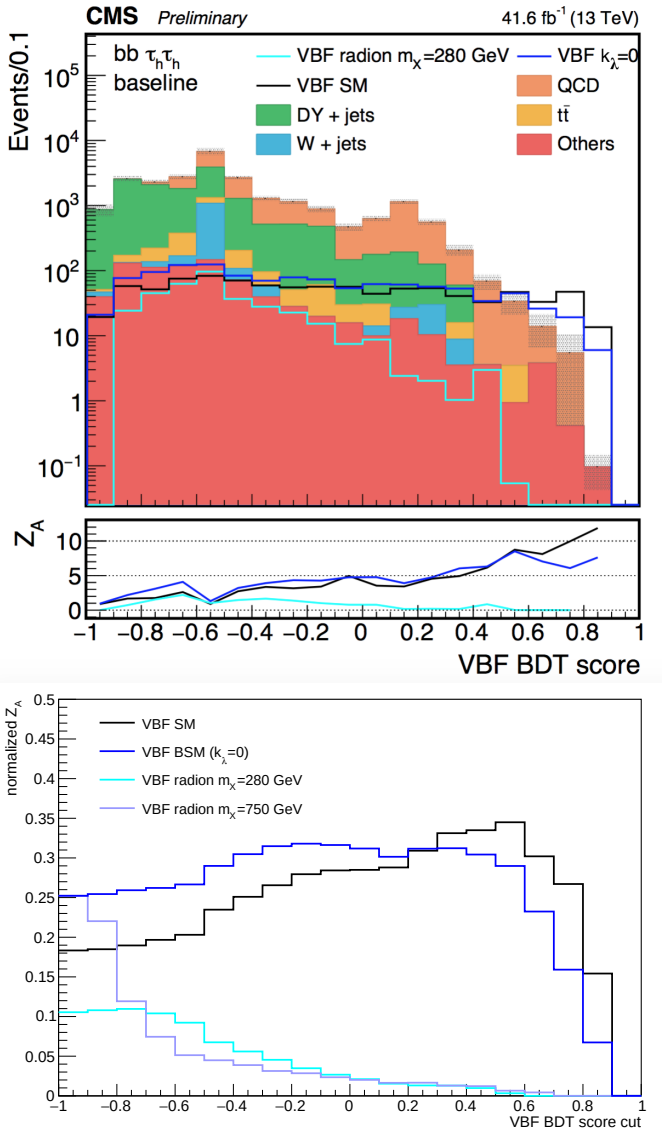


Figure 5.20: On the top, comparison of the VBF BDT output scores for background and signal events selected after the invariant mass cuts on $m_{\tau\tau}^{SVfit}$ and m_{bb} in the $\tau_h\tau_h$ channel with no additional requirement on the b tagging of the jets. On the bottom, Z_A as function of the selection applied to the BDT score: as expected, for tighter selection the sensitivity in non-resonant searches increases. The decrease observed for resonant signals is due to the fact that no VBF resonant samples where included in the training of the BDT.

5.6 Final remarks and future prospects on the $HH \to bb\tau\tau$ analysis

Having worked on the $HH \to bb\tau\tau$ channel since 2016 I was lucky enough to have the opportunity to participate in the analysis design and development from the initial approach to the first $\sqrt{s} = 13\ TeV$ data, until the end of the LHC proton-proton data taking period of Run II, officially ended on the 24^{th} October 2018.

This also gave me the chance of understanding what the most critical aspects of this analysis are and what main issues could represent an obstacle in its future development. So far, the algorithms and techniques put in place for the $bb\tau\tau$ search have led to excellent results when analyzing 2016 data; the improvements introduced in 2017, that I described in this thesis, represent their natural continuation on the path to the analysis of the full Run II statistics and, in a wider perspective, to the High Luminosity phase of the LHC (HL-LHC).

Here I wish to briefly summarize some specific aspects that I think should be the focus of the efforts in the years to come in order to fully exploit the opportunities offered by an interesting channel such as the $bb\tau\tau$ final state.

Following the structure of this thesis, which in turn mirrors the structure of the analysis itself, I first comment on the trigger selection and object reconstruction. Unfortunately, both aspects are strictly driven by the limitations and needs of the experimental apparatus. Despite the upgrade foreseen for the CMS detector, as we enter the environment shaped by the HL-LHC collisions, a mean pileup of about 200 proton-proton interactions every bunch crossing will force the trigger threshold to be raised, with a subsequent loss of acceptance. Thus, the techniques and algorithms implemented for the reconstruction and identification of physics objects will play a major role in maintaining an high selection efficiency.

The SVfit algorithm, applied to reconstruct the $H \to \tau\tau$ decays, already provides very good performances and will profit from the improvement in the reconstruction of objects like hadronically decaying taus and the missing transverse momentum of the event. For the former, new discriminators, based on Deep Neural Network infrastructures, are being studied to reduce the contamination originating from gluon and quark jets misidentified as τ lepton and to include different tau decay modes other than those already considered and detailed in Section 3.1. The reconstruction of MET instead will benefit from the upgrade of the HCAL calorimeter (HGCAL) that will provide a finer granularity and a better control of the observable related to the jets and to the multiplicity of objects in the event.

Regarding the selection of $H \to bb$ candidates, the possible improvements are mainly related to the correct identification of b jets originating from

an Higgs boson. Beside a better understanding of the object multiplicity and variables related the object energy, the improvements will come from the development of new algorithms, either for the b tagging, such as the DeepFlavour tagger that is being commissioned with Run II data, or through dedicated MVA techniques designed to identify $H \to bb$ candidates.

In searches like $HH \to bb\tau\tau$, that are characterized by a small expected number of signal events, the choice of the "working points" is always a compromise between the performance of the discriminators and the statistics available. Thus, the increase of data collected in Run II, then in Run III and finally in HL-LHC, will allow CMS to tighten the requirements on the discriminators to select only the purest events (the $2b0j$ category in this specific case) and obtain a better object identification maintaining a good statistical power.

As anticipated at the beginning of this Section, the techniques developed in 2017, and described in this thesis, represent a further step in the optimization of the final analysis. The introduction of a BDT discriminator to be used as input to the likelihood fit has proven to be very effective when tested on 2016 data, with a gain in sensitivity up to 40% in some channels and categories. The training of the BDT has been performed on 2016 samples but, due to unavoidable changes in the detector and reconstruction performances, the input variable distributions and the agreement between the observed data and the MC simulation may change during the years, thus the BDT training needs to be upgraded to include the full 2016+2017+2018 statistics.

Moreover, the study of the vector boson fusion (VBF) production mode is a relatively new field in double Higgs searches; its impact on the overall $bb\tau\tau$ analysis will be better understood and evaluated only when the full Run II statistics will be analyzed.

CMS prospects of the $bb\tau\tau$ analysis in the HL-LHC phase (at 3000 fb^{-1} integrated luminosity), have been recently studied in the context of the European Strategy for Particle Physics (EuSPP) and reported in the document CMS-FTR-18-019 [78] and in the Yellow Report CERN-LPCC-2018-04 [79].

Results are projected under the assumptions that events are collected with triggers and selections similar to those used in Run II collisions, but with a trigger efficiency of 100% for the reconstructed objects. These assumptions appear reasonable considering the improved capabilities of the upgraded CMS detector, the usage of track information in the L1 trigger, and the possibility to develop more sophisticated kinematic triggers to specifically target the $HH \to bb\tau\tau$ signal.

No specific fitting technique is applied to reconstruct the $H \to \tau\tau$ candidates which are simply defined as the sum of the four momenta of the visible decay products of the tau leptons and the missing transverse energy. Events are selected for this study only if they contain at least two b-tagged jets, *i.e.*

mimicking the $2b0j$ category as defined in Section 3.3.2.

A neural network discriminant (DNN) is developed to separate the signal contribution from the background processes and its output is used to determine the expected discovery significance and cross section upper limit at 95% confidence level. An upper limit on the HH cross section times the branching fraction of 1.4 times the SM prediction is obtained, corresponding to a significance of 1.4σ.

The $bb\tau\tau$ final state is combined with other HH decay channels (*i.e.* $bbbb$, $bbWW$, $bb\gamma\gamma$, $bbZZ$) to estimate the overall sensitivity. The exclusion limits are summarized in Table 5.9: the combined 95% CL upper limit on the SM HH cross section amounts to 0.77 times the SM prediction, with a corresponding significance of the signal of 2.6σ. This result can be directly compared with the exclusion limit of 12.8 times the SM prediction, obtained from the combination of all the CMS HH analyses performed with 2016 data and discussed in Section 6.3.

Channel	Significance		95% CL limit on $\sigma_{HH}/\sigma_{HH}^{SM}$	
	Syst. + Stat.	Stat. only	Syst. + Stat.	Stat. only
$bbbb$	0.95	1.2	2.1	1.6
$bb\tau\tau$	1.4	1.6	1.4	1.3
$bbWW(\ell\nu\ell\nu)$	0.56	0.59	3.5	3.3
$bb\gamma\gamma$	1.8	1.8	1.1	1.1
$bbZZ(\ell\ell\ell\ell)$	0.37	0.37	6.6	6.5
Combination	2.6	2.8	0.77	0.71

Table 5.9: Upper limit at the 95% confidence level, significance, projected measurement at 68% confidence level of the Higgs boson self coupling λ_{HHH} for the five channels studied and their combination. Systematic and statistical uncertainties are considered. [78]

Prospects for the measurement of the λ_{HHH} coupling are also studied and, under the assumption that no HH signal exists, 95% CL upper limits on the SM HH production cross section are derived as function of $k_\lambda = \lambda_{HHH}/\lambda_{HHH}^{SM}$, where λ_{HHH}^{SM} denotes the SM prediction. The results are illustrated in Figure 5.21 and can be compared to the same exclusion limit obtained for the combination of all the CMS HH analyses performed with 2016 data and reported in Figure 6.3.

The High-Luminosity LHC will thus provide a unique opportunity to study HH production as predicted in the SM and identify possible deviations induced by BSM physics in the signal cross section and kinematic properties.

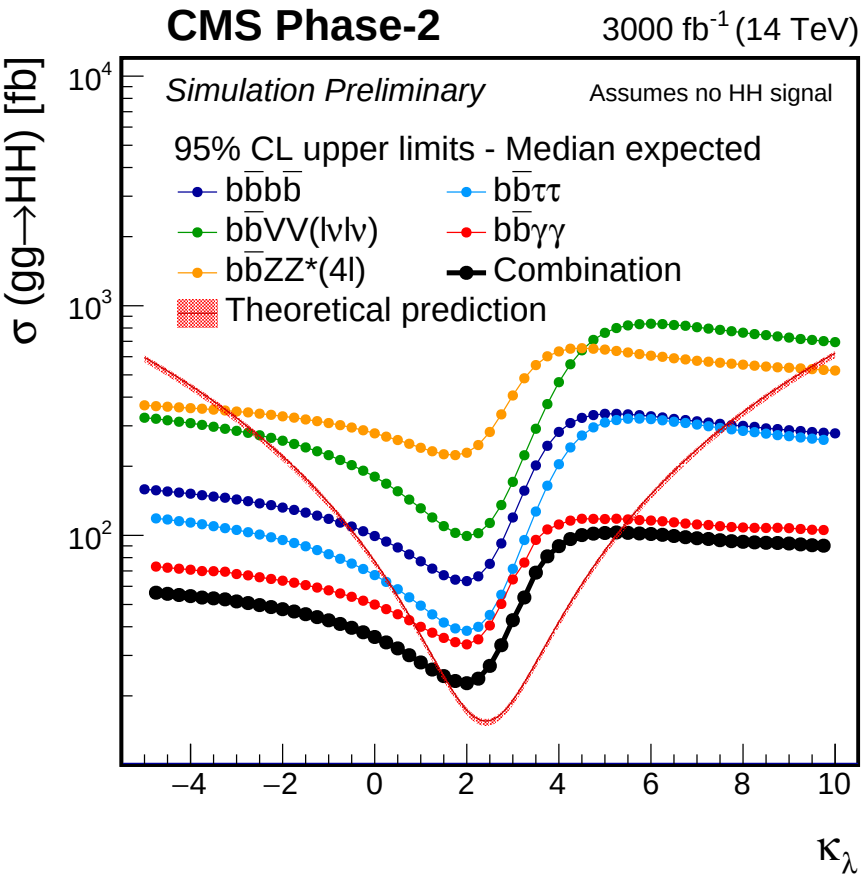


Figure 5.21: Upper limit at the 95% CL on the HH production cross section as a function of $k_\lambda = \lambda_{HHH}/\lambda^{SM}_{HHH}$ for the five decays channels investigated and their combination. The red band indicates the theoretical production cross section. [78]

Chapter 6

Combination of 2016 HH Analyses

6.1 Introduction

Higgs boson pairs decay in several channels, each with specific topological and kinematic features, that result in different sensitivities to different regions of the anomalous coupling parameter space and of the resonant invariant mass spectrum. A fundamental characteristic of the decay channels, is the complementarity offered in terms of sensitivity, as displayed in Figure 6.1 for the non-resonant case. The exploration and combination of several channels is therefore necessary in order to probe, in the most efficient and effective way, BSM physics in the context of HH processes.

This Chapter describes the combination of double Higgs searches performed by the CMS collaboration using an integrated luminosity of 35.9 fb^{-1}, collected for each final state in 2016 [81]. Four different decay channels are considered, where one Higgs boson decays to a bb pair, and the other decays to $\gamma\gamma$ [82], $\tau\tau$ [1], bb [83–86] or VV [87], where V stands for a vector boson decaying leptonically. I worked in the context of this combination as the contact person of the $bb\tau\tau$ final state.

6.2 Analyses description

In Sections 6.2.1, 6.2.2 and 6.2.3, the $bb\gamma\gamma$, $bbbb$ and $bbVV$ analyses are briefly described, while the $bb\tau\tau$ search, which is the object of this thesis, has been already described in detail in the previous Chapters.

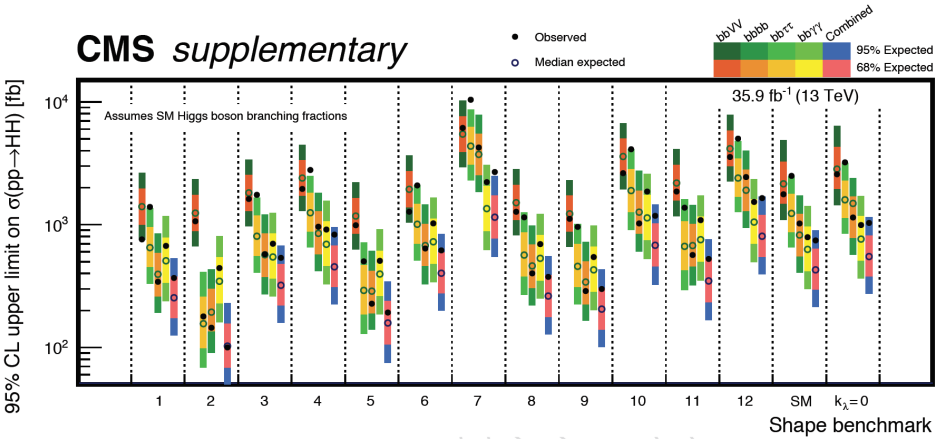


Figure 6.1: The 95% CL exclusion limits on non-resonant Higgs boson pair production cross sections for different EFT benchmark topologies (bins 1 to 12). The last two bins show the 95% CL exclusion limits for the SM and for the $k_\lambda = 0$ scenarios. Limits are shown for each of the four final states separately and for the combination [80].

6.2.1 $b\bar{b}\gamma\gamma$

The $bb\gamma\gamma$ analysis is currently the most sensitive to double Higgs production in CMS. Despite the low branching fraction ($\sim$ 0.26%), the small background contamination and the excellent di-photon mass resolution of the CMS detector allow for very clean event signatures. In order to maximize the sensitivity, events are split in categories according to two variables. The first is the "reduced mass", defined as $\tilde{M}_X = M_{jj\gamma\gamma} - M_{jj} - M_{\gamma\gamma} + 250\ GeV$, and the second is a BDT discriminator, built from the b-tagging probabilities of the jets, the angles in the HH systems, and the transverse momentum of the two Higgs boson candidates. The analysis relies on a $2D$ fit to the $H \to bb$ and $H \to \gamma\gamma$ invariant mass distributions to calculate the exclusion limits. The main background contribution comes from the continuum $N\gamma + jets$ and is estimated from mass sidebands, while the modelization of the other backgrounds, mainly coming from single Higgs events, relies on Monte Carlo simulations. Given the small event yield, the analysis is dominated by statistical uncertainty.

6.2.2 $b\bar{b}b\bar{b}$

Among all the possible HH final states, the one with four b quarks has the highest branching fraction, about 33%. Two different analyses are developed for the non-resonant and resonant cases. The former is performed targeting

four b-jets in the final state, and the sensitivity is enhanced by the use of a BDT technique that exploits jet-related and Higgs kinematic variables. Further sensitivity is gained adding the final state where one or two Higgs bosons have a boosted topology. For the resonant case, the analysis is further optimized to target different resonance masses: below 700 GeV four b-tagged jets are required in the final state, while for resonances with $m_X > 1200\ GeV$, events are selected only if they contain two "fat jets", where the products of the Higgs boson decays are actually merged in a single large jet. Finally, in the intermediate region ($700 < m_X < 1200\ GeV$), the search sensitivity is improved by considering both the final states with four b-jets and the case with one "fat jet" and two b-jets. The dominant background in the *bbbb* searches comes from QCD multijet production that is estimated using sideband regions and an hemisphere mixing technique.

6.2.3 $b\bar{b}VV$

The $bbVV$ analysis include the $bbWW \to bb\ell\nu\ell\nu$ and $bbZZ \to \ell\ell\nu\nu$ final states, for a total branching fraction of $\sim 2.7\%$. Two main backgrounds affect this analysis: $t\bar{t}$ events, estimated from MC simulation, and Drell-Yan processes in association with jets, estimated from data. In order to reduce the background contamination in the signal regions, a parametrized Deep Neural Network (DNN) approach is exploited: in the resonant case the DNN is parametrized accordingly to the resonance mass, while k_λ and k_t are used as parameters in the non-resonant search. Signal extraction relies on the shape of the neural network output binned in three different regions of the m_{jj} spectrum, while the main source of systematic uncertainty arises from the b-tagging efficiency and the electron identification.

6.2.4 Analyses cross-checks: the $b\bar{b}\tau^+\tau^-$ case

When combining various analyses in a single result it is important to take into account all the possible correlations. In this Section, as an example, only the $bb\tau\tau$ case is detailed.

Special care must be given to the estimate of the phase space of the different searches involved in the combination, due to the fact that any overlapping event represents a potential double count contribution. By construction the $bb\tau\tau$ and $bbVV$ channels are mutually exclusive, since the former applies a "third lepton veto" in its selections, rejecting every event where a second electron or muon is found, while the latter explicitly requires in the final state the presence of two leptons (e or μ) coming from the decay of the Z or W bosons.
Even if the fraction of events firing the trigger requirements for the *bbbb* and $bb\tau\tau$ analyses is significant, less than 1% of $bb\tau\tau$ selected events have four

b tagged jets and, at the same time, less than 1% have hadronic taus that could be misidentified as b jets. The few $bb\tau\tau$ events still passing the $bbbb$ analysis selections are anyway kinematically very different with respect to pure $bbbb$ events, and are therefore pushed to low values of the BDT score where their impact on the final result is very limited, if not null at all.

The treatment of systematic uncertainties and its inclusion in the final fit is better detailed in Section 6.3, while here only the special case of the $bb\tau\tau$ jet energy scale (JES) systematics is described.

Jet energy corrections are a common feature in CMS double Higgs searches, since all of them require the presence of at least two jets in the final state. The systematic uncertainty related to these corrections can be split in 28 different sources that affect the JES estimation in both the rate and the shape of the final observables considered. In the $bb\tau\tau$ result published in [1], the jet energy scale systematics were applied inclusively: the cumulative effect of all the JES sources both on yield and shape was applied to the different signal and background processes. In order to properly correlate the JES uncertainties across the different channels, the effect of the 28 sources has been evaluated independently. The dominant effect comes from the yield variation, but it remains well contained to less than 3 and 4% for the signal and $t\bar{t}$ background process, respectively. The shape effects from the individual sources have been verified to be negligible, as shown in Figure 6.2 for a signal sample and for the $t\bar{t}$ background.

It was thus decided to introduce in the final fit 28 nuisances affecting the normalization of the processes and one uncorrelated shape uncertainty that covers the cumulative shape effect of the JES.

6.3 Statistical combination and results

For both the resonant and non-resonant searches, likelihood fits are performed using as parameter of interest the signal strength modifier μ, defined as the ratio between the observed and expected signal rates, and estimated with its corresponding confidence interval via the profile likelihood ratio test statistic. The expected value for the signal strength is assumed to be the Standard Model gluon fusion double Higgs cross-section, which corresponds to 33.49 fb. For all measurements, the Higgs boson mass is fixed at $m_H = 125\ GeV$ and its branching fractions are assumed to be equal to the Standard Model predictions. Systematic uncertainties and their correlations are modeled in the test statistic by introducing nuisance parameters described by likelihood functions that express the various experimental and theoretical uncertainties.

A proper handling of the systematic uncertainty sources and their correlations is particularly important when combining the various analyses entering

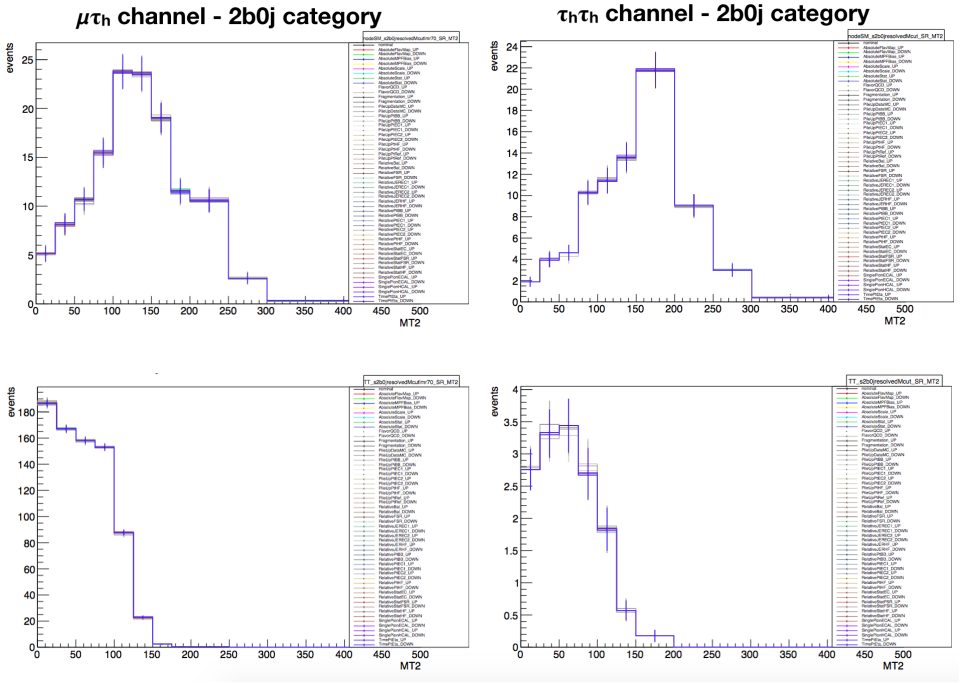


Figure 6.2: The $bb\tau\tau$ alternative shapes of the $MT2$ variable in the $2b0j$ category. The $\mu\tau_h$ and $\tau_h\tau_h$ decay channels are shown on the left and on the right, respectively, for the Standard Model HH signal (top) and the $t\bar{t}$ background (bottom).

the double Higgs combination. Some of the systematics are assumed to be fully correlated across the different channels, for example the uncertainties on the integrated luminosity and on the total cross sections of the common background processes, or the theoretical uncertainties affecting α_s, the PDFs and the finite top mass effects in next-to-next-to-leading order calculations. Some uncertainties, related to the reconstruction and identification efficiencies, and to the energy scale corrections, are assumed to be fully correlated across the channels that use the same objects. Others are related specifically to a single final state, such as the uncertainty on the same sign to opposite sign candidate ratio used in the $bb\tau\tau$ analysis, or the photon identification, selection and resolution uncertainties, relevant only for the $bb\gamma\gamma$. Uncertainty sources related to the b tagging process are considered the same across all analyses except for the non-resonant $bbbb$ search that makes use of a different b tagging algorithm.

In total, more than 450 individual nuisance parameters are identified and included in the final fit. The main sources of systematic uncertainties come from the $bbbb$ and $bb\tau\tau$ analyses, that have similar sensitivity. In particular, the normalization fluctuations in the most sensitive bins of the $bbbb$ BDT

and the τ energy scale effects in the $bb\tau\tau$ analysis are the systematics with the largest impact.

Once all the correlations across the different channels are included, the observed and expected exclusion limits at 95% confidence level on the non-resonant double Higgs production signal strength are measured to be 22.2 and 12.8 times the Standard Model predictions, respectively. A scan is performed for different values of the k_λ parameter, that impacts not only the HH production cross section but also the kinematic properties of HH events. When keeping all the other parameters fixed to their SM value, the k_λ parameter is observed to be constrained in the interval $11.8 < k_\lambda < 18.8$ at 95% CL, for an expected value of $7.1 < k_\lambda < 13.6$. Figure 6.3 illustrates the exclusion limits in the SM case for the individual channels and their combination, and the scan as function of k_λ.

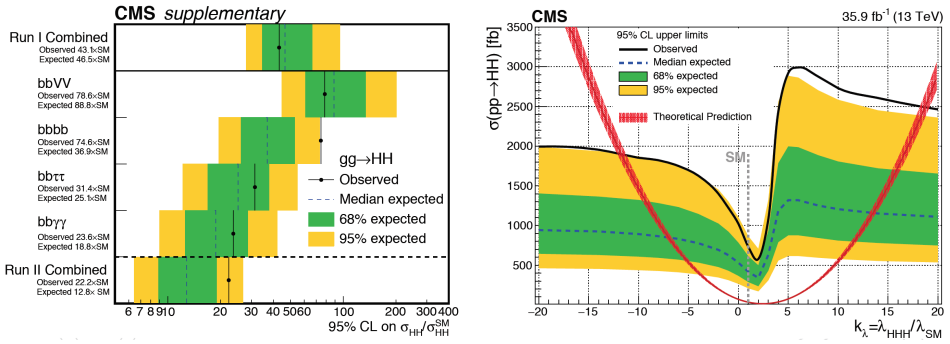


Figure 6.3: On the left, 95% CL exclusion limits on non-resonant Higgs boson pair production for the individual channels and their combination. On the right, the exclusion limit as value of k_λ. The green and the yellow bands indicate the regions containing 68 and 95%, respectively, of the expected limits distribution [81].

The resonant search is performed for either a spin-0 or a spin-2 resonance and no significant excess of events is found. The range of masses investigated in the different searches spans from 250 to 3000 GeV. The results are combined and displayed in Figure 6.4 for both the spin hypothesis.

In light of these results, it is clear that the combination of different HH analyses is a fundamental tool in order to explore the scalar sector of the Standard Model by fully exploiting the Run II collected statistics, and even more so in view of the High Luminosity phase of the Large Hadron Collider.

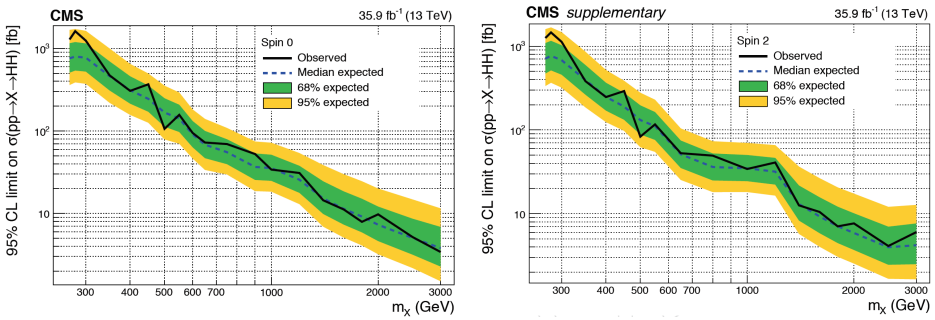


Figure 6.4: Expected (dashed) and observed (solid line) 95% CL exclusion limits on the production of a narrow, spin-0 (left) of spin-2 (right) resonance decaying into a pair of Higgs bosons. The green and the yellow bands indicate the regions containing 68 and 95%, respectively, of the expected limits distribution [81] [80].

Appendices

Appendix A

2017 search: BDT input variables

This Appendix reports the agreement between observed data and MC simulation for a more extensive set of the input variables of the 2017 analysis BDT. Plots are shown for events selected in the $\tau_\mu\tau_h$ final state in the $1b1j$ category. A general good level of agreement between data and MC can be observed.

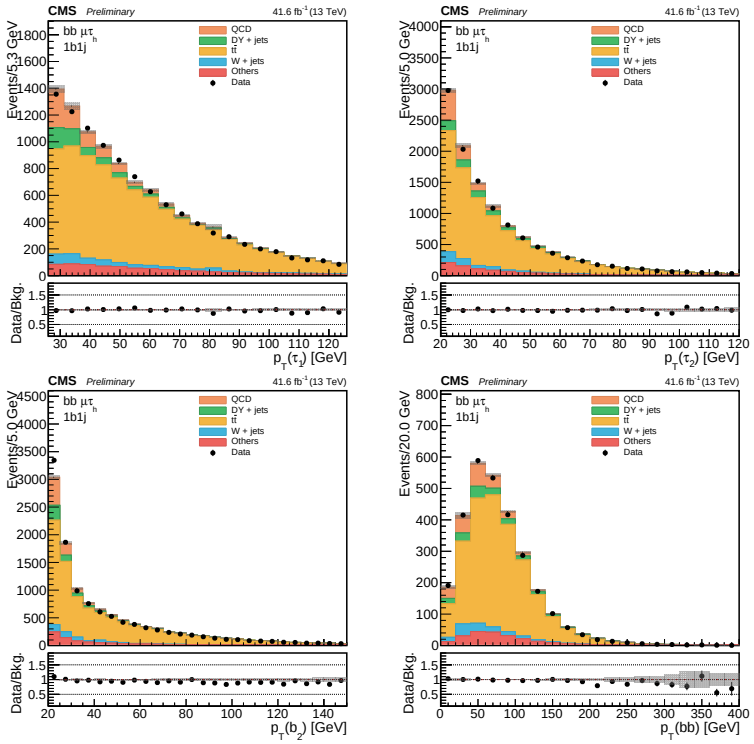


Figure A.1: Distributions of some representative BDT input variables for events selected in the $\tau_\mu\tau_h$ channel in the $1b1j$ category.

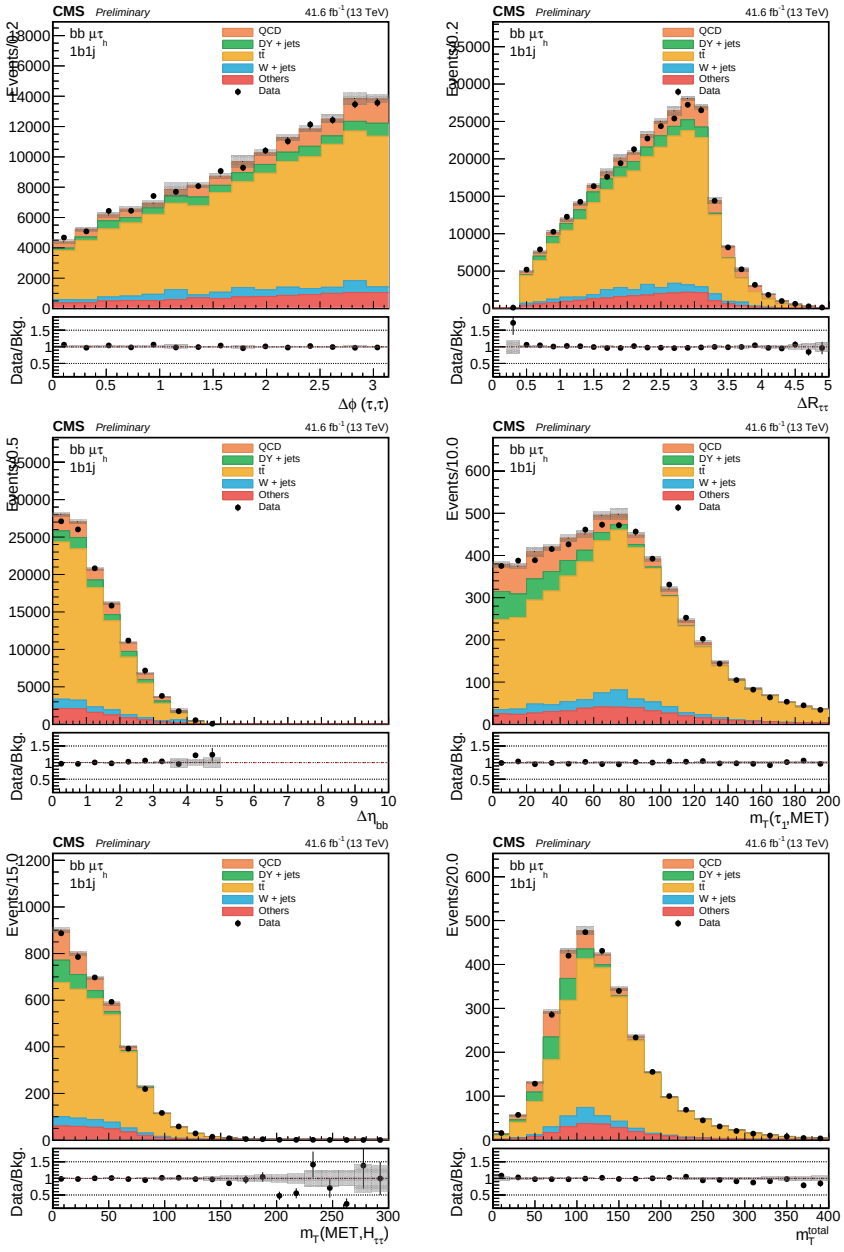


Figure A.2: Distributions of some representative BDT input variables for events selected in the $\tau_\mu\tau_h$ channel in the $1b1j$ category. Here m_T is computed as $m_T(x,y) = \sqrt{2 \cdot p_T(x) \cdot p_T(y) \cdot (1 - cos\theta(x,y))}$, while $m_T^{total} = \sqrt{m_T^2(\tau_1, MET) + m_T^2(\tau_2, MET) + m_T^2(\tau_1, \tau_2)}$.

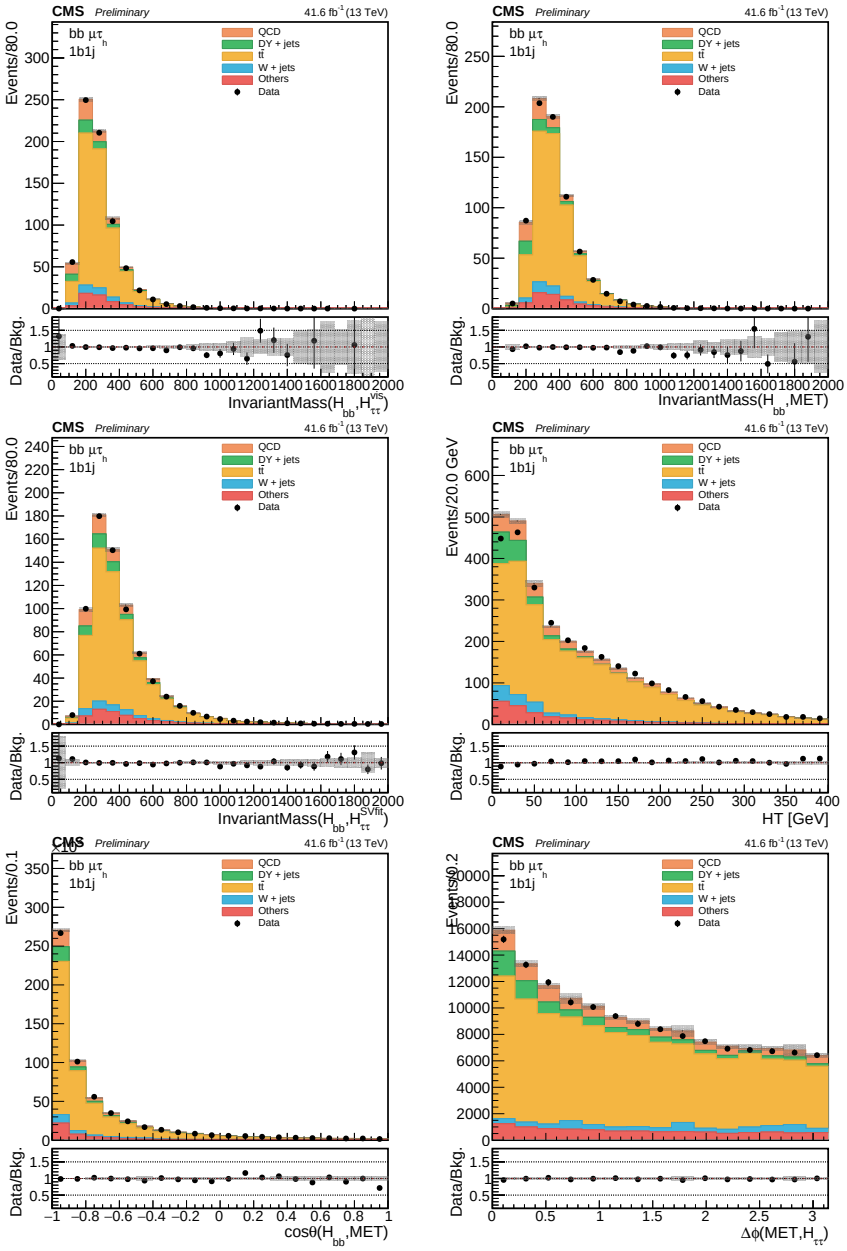


Figure A.3: Distributions of some representative BDT input variables for events selected in the $\tau_\mu\tau_h$ channel in the $1b1j$ category. Here HT represents the scalar sum of the transverse momentum of all jets with $p_T > 20$ GeV, excluding the two selected b-jets.

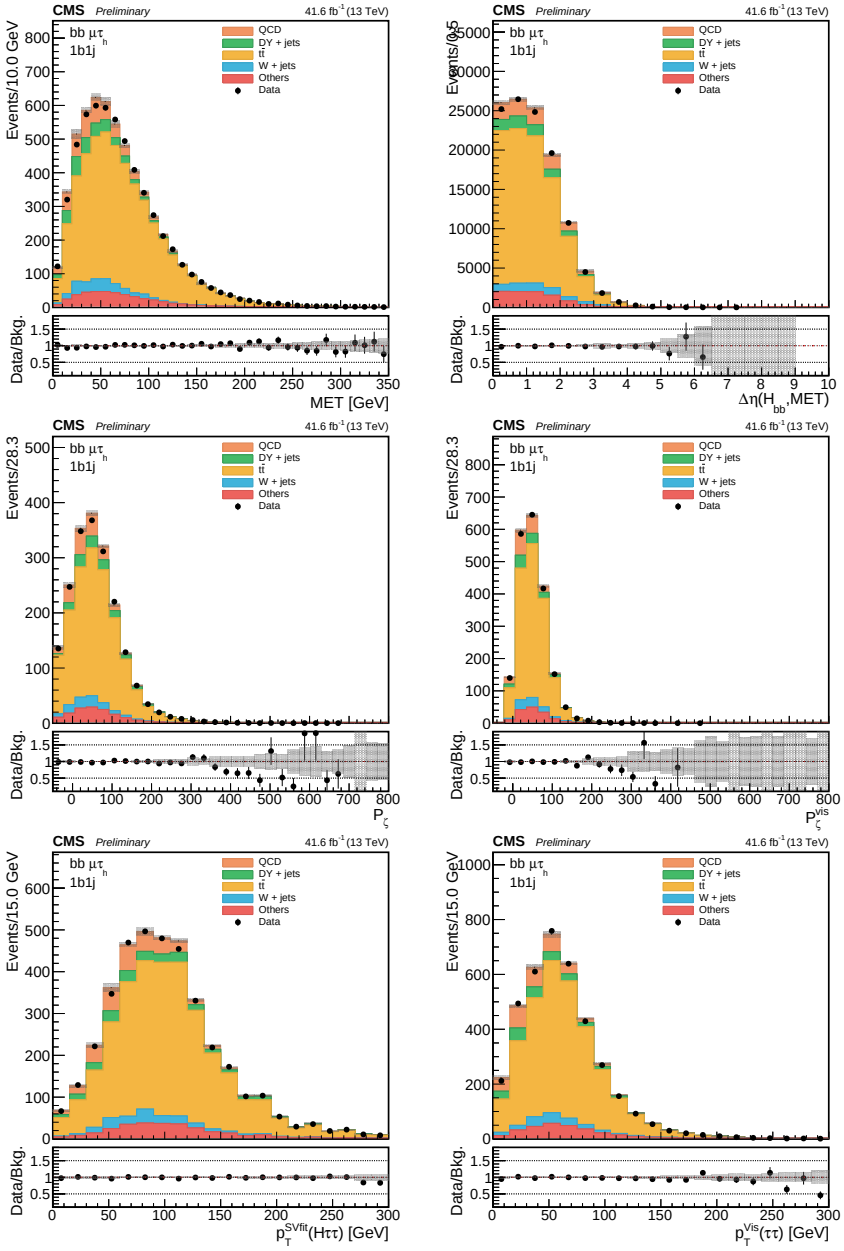


Figure A.4: Distributions of some representative BDT input variables for events selected in the $\tau_\mu\tau_h$ channel in the $1b1j$ category. Here p_ζ and p_ζ^{vis} are defined as $p_\zeta = (\vec{p}_T(\tau_1)+\vec{p}_T(\tau_2)+MET)\cdot\hat{\zeta}$ and $p_\zeta^{vis} = (\vec{p}_T(\tau_1)+\vec{p}_T(\tau_2))\cdot\hat{\zeta}$, where $\hat{\zeta}$ is a unit vector in the direction of the bisector of the $\vec{p}_T$ vectors of the two tau leptons.

Bibliography

[1] Albert M Sirunyan et al. Search for Higgs boson pair production in events with two bottom quarks and two tau leptons in proton-proton collisions at $\sqrt{s}$ =13TeV. *Phys. Lett.*, B778:101–127, 2018.

[2] Georges Aad et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett.*, B716:1–29, 2012.

[3] Serguei Chatrchyan et al. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett.*, B716:30–61, 2012.

[4] S. L. Glashow. Partial Symmetries of Weak Interactions. *Nucl. Phys.*, 22:579–588, 1961.

[5] Steven Weinberg. A Model of Leptons. *Phys. Rev. Lett.*, 19:1264–1266, 1967.

[6] Abdus Salam. Weak and Electromagnetic Interactions. *Conf. Proc.*, C680519:367–377, 1968.

[7] Werner Bernreuther. Top quark physics at the LHC. *J. Phys.*, G35:083001, 2008.

[8] F. Englert and R. Brout. Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.*, 13:321–323, Aug 1964.

[9] Peter W. Higgs. Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.*, 13:508–509, Oct 1964.

[10] John Ellis. Higgs Physics. In *Proceedings, 2013 European School of High-Energy Physics (ESHEP 2013): Paradfurdo, Hungary, June 5-18, 2013*, pages 117–168, 2015.

[11] D. de Florian et al. Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector. 2016.

[12] R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, P. Torrielli, E. Vryonidou, and M. Zaro. Higgs pair production at the LHC with NLO and parton-shower effects. *Phys. Lett.*, B732:142–149, 2014.

[13] T. Binoth and J. J. van der Bij. Influence of strongly coupled, hidden scalars on Higgs signals. *Z. Phys.*, C75:17–25, 1997.

[14] G. C. Branco, P. M. Ferreira, L. Lavoura, M. N. Rebelo, Marc Sher, and Joao P. Silva. Theory and phenomenology of two-Higgs-doublet models. *Phys. Rept.*, 516:1–102, 2012.

[15] Summary results of high mass BSM Higgs searches using CMS run-I data. Technical Report CMS-PAS-HIG-16-007, CERN, Geneva, 2016.

[16] Lisa Randall and Raman Sundrum. A Large mass hierarchy from a small extra dimension. *Phys. Rev. Lett.*, 83:3370–3373, 1999.

[17] Florian Goertz, Andreas Papaefstathiou, Li Lin Yang, and Jose Zurita. Higgs boson pair production in the D=6 extension of the SM. *JHEP*, 04:167, 2015.

[18] Alexandra Carvalho, Martino Dall'Osso, Pablo De Castro Manzano, Tommaso Dorigo, Florian Goertz, Maxime Gouzevich, and Mia Tosi. Analytical parametrization and shape classification of anomalous HH production in the EFT approach. 2016.

[19] Alexandra Carvalho, Martino Dall'Osso, Tommaso Dorigo, Florian Goertz, Carlo A. Gottardo, and Mia Tosi. Higgs Pair Production: Choosing Benchmarks With Cluster Analysis. *JHEP*, 04:126, 2016.

[20] Georges Aad et al. 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.*, D92:092004, 2015.

[21] Albert M Sirunyan et al. Search for Higgs boson pair production in the $bb\tau\tau$ final state in proton-proton collisions at $\sqrt{(s)} = 8$ TeV. *Phys. Rev.*, D96(7):072004, 2017.

[22] Yves Baconnier, Giorgio Brianti, P Lebrun, A G Mathewson, R Perin, and Yves Baconnier. *LHC: the Large Hadron Collider accelerator project.* CERN, Geneva, 1993.

[23] Thomas Sven Pettersson and P Lefevre. The Large Hadron Collider: conceptual design. Technical Report CERN-AC-95-05-LHC, Oct 1995.

[24] Esma Mobs. The CERN accelerator complex. Complexe des accelerateurs du CERN. Jul 2016. General Photo.

[25] Lyndon Evans and Philip Bryant. Lhc machine. *Journal of Instrumentation*, 3(08):S08001, 2008.

[26] Cms luminosity - public results
. https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[27] Hl-lhc project. https://project-hl-lhc-industry.web.cern.ch.

[28] G. L. Bayatian et al. CMS Physics. 2006.

[29] G. L. Bayatian et al. CMS technical design report, volume II: Physics performance. *J. Phys.*, G34(6):995–1579, 2007.

[30] V Karimaki, M Mannelli, P Siegrist, H Breuker, A Caner, R Castaldi, K Freudenreich, G Hall, R Horisberger, M Huhtinen, and A Cattai. *The CMS tracker system project: Technical Design Report.* Technical Design Report CMS. CERN, Geneva, 1997.

[31] A at al. Dominguez. CMS Technical Design Report for the Pixel Detector Upgrade. Technical Report CERN-LHCC-2012-016. CMS-TDR-11, Sep 2012.

[32] *The CMS electromagnetic calorimeter project: Technical Design Report.* Technical Design Report CMS. CERN, Geneva, 1997.

[33] *The CMS hadron calorimeter project: Technical Design Report.* Technical Design Report CMS. CERN, Geneva, 1997.

[34] J Mans, J Anderson, B Dahmes, P de Barbaro, J Freeman, T Grassi, E Hazen, J Mans, R Ruchti, I Schimdt, T Shaw, C Tully, J Whitmore, and T Yetkin. CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter. Technical Report CERN-LHCC-2012-015. CMS-TDR-10, Sep 2012.

[35] *The CMS muon project: Technical Design Report.* Technical Design Report CMS. CERN, Geneva, 1997.

[36] The CMS collaboration. The performance of the cms muon detector in proton-proton collisions at âĹŽ s = 7 tev at the lhc. *Journal of Instrumentation*, 8(11):P11002, 2013.

[37] S. Dasu et al. CMS. The TriDAS project. Technical design report, vol. 1: The trigger systems. 2000.

[38] P. Sphicas. CMS: The TriDAS project. Technical design report, Vol. 2: Data acquisition and high-level trigger. 2002.

[39] A. M. Sirunyan et al. Particle-flow reconstruction and global event description with the CMS detector. *JINST*, 12(10):P10003, 2017.

[40] D. Buskulic et al. Performance of the ALEPH detector at LEP. *Nucl. Instrum. Meth.*, A360:481–506, 1995.

[41] Vardan Khachatryan et al. Performance of Electron Reconstruction and Selection with the CMS Detector in Proton-Proton Collisions at $\sqrt{s} =$ 8 TeV. *JINST*, 10(06):P06005, 2015.

[42] Andreas Hocker et al. TMVA — Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007.

[43] Serguei Chatrchyan et al. Performance of CMS muon reconstruction in *pp* collision events at $\sqrt{s} = 7$ TeV. *JINST*, 7:P10002, 2012.

[44] CMS Collaboration. Performance of ÏĎ-lepton reconstruction and identification in cms. *Journal of Instrumentation*, 7(01):P01001, 2012.

[45] Vardan Khachatryan et al. Reconstruction and identification of ÏĎ lepton decays to hadrons and Î¡ at CMS. *JINST*, 11(01):P01019, 2016.

[46] Performance of reconstruction and identification of tau leptons in their decays to hadrons and tau neutrino in LHC Run-2. Technical Report CMS-PAS-TAU-16-002, CERN, Geneva, 2016.

[47] Vardan Khachatryan et al. Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV. *JINST*, 12(02):P02014, 2017.

[48] Jet algorithms performance in 13 TeV data. Technical Report CMS-PAS-JME-16-003, CERN, Geneva, 2017.

[49] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti- k t jet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063, 2008.

[50] Serguei Chatrchyan et al. Identification of b-quark jets with the CMS experiment. *JINST*, 8:P04013, 2013.

[51] A.M. Sirunyan and A. Tumasyan et al. Identification of heavy-flavour jets with the cms detector in pp collisions at 13 tev. *Journal of Instrumentation*, 13(05):P05011, 2018.

[52] Identification of b quark jets at the CMS Experiment in the LHC Run 2. Technical Report CMS-PAS-BTV-15-001, CERN, Geneva, 2016.

[53] Performance of missing energy reconstruction in 13 TeV pp collision data using the CMS detector. Technical Report CMS-PAS-JME-16-004, CERN, Geneva, 2016.

[54] CMS Collaboration. TECHNICAL PROPOSAL FOR A MIP TIMING DETECTOR IN THE CMS EXPERIMENT PHASE 2 UPGRADE. Technical Report CERN-LHCC-2017-027. LHCC-P-009, CERN, Geneva, Dec 2017.

[55] Tau Identification Performance in 2017 Data at $\sqrt{s} = 13$ TeV. Jun 2018.

[56] Andrew J. Larkoski, Simone Marzani, Gregory Soyez, and Jesse Thaler. Soft Drop. *JHEP*, 05:146, 2014.

[57] Lorenzo Bianchini, John Conway, Evan Klose Friis, and Christian Veelken. Reconstruction of the higgs mass in h âĘŠ ÏĎÏĎ events by dynamical likelihood techniques. *Journal of Physics: Conference Series*, 513(2):022035, 2014.

[58] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 07:079, 2014.

[59] Emanuele Re. Single-top Wt-channel production matched with parton showers using the POWHEG method. *Eur. Phys. J.*, C71:1547, 2011.

[60] Richard D. Ball et al. Parton distributions for the LHC Run II. *JHEP*, 04:040, 2015.

[61] TorbjÃűrn SjÃűstrand, Stefan Ask, Jesper R. Christiansen, Richard Corke, Nishita Desai, Philip Ilten, Stephen Mrenna, Stefan Prestel, Christine O. Rasmussen, and Peter Z. Skands. An Introduction to PYTHIA 8.2. *Comput. Phys. Commun.*, 191:159–177, 2015.

[62] S. Agostinelli et al. GEANT4: A Simulation toolkit. *Nucl. Instrum. Meth.*, A506:250–303, 2003.

[63] Albert M Sirunyan et al. Search for additional neutral MSSM Higgs bosons in the $\tau\tau$ final state in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 09:007, 2018.

[64] Albert M Sirunyan et al. Measurement of differential cross sections for Z boson production in association with jets in proton-proton collisions at $\sqrt{s} = 13$ TeV. 2018.

[65] Measurement of the inclusive and differential tt production cross sections in lepton + jets final states at 13 TeV. Technical Report CMS-PAS-TOP-15-005, CERN, Geneva, 2015.

[66] Vardan Khachatryan et al. 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 $h \to \tau\tau$. *Phys. Lett.*, B755:217–244, 2016.

[67] C. G. Lester and D. J. Summers. Measuring masses of semiinvisibly decaying particles pair produced at hadron colliders. *Phys. Lett.*, B463:99–103, 1999.

[68] Alan Barr, Christopher Lester, and P. Stephens. m(T2): The Truth behind the glamour. *J. Phys.*, G29:2343–2363, 2003.

[69] Alan J. Barr, Matthew J. Dolan, Christoph Englert, and Michael Spannowsky. Di-Higgs final states augMT2ed – selecting hh events at the high luminosity LHC. *Phys. Lett.*, B728:308–313, 2014.

[70] Chrisopher G. Lester and Benjamin Nachman. Bisection-based asymmetric M_{T2} computation: a higher precision calculator than existing symmetric methods. *JHEP*, 03:100, 2015.

[71] Ralf Steuer, Carsten O. Daub, Joachim Selbig, and Jürgen Kurths. *Measuring Distances Between Variables by Mutual Information.* In *Innovations in Classification, Data Science, and Information Systems*, pages 81–90, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[72] Procedure for the LHC Higgs boson search combination in Summer 2011. Technical Report CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11, CERN, Geneva, Aug 2011.

[73] CMS Luminosity Measurements for the 2016 Data Taking Period. Technical Report CMS-PAS-LUM-17-001, CERN, Geneva, 2017.

[74] CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} =$ 13 TeV. Technical Report CMS-PAS-LUM-17-004, CERN, Geneva, 2018.

[75] Cms hig-17-002 supplementary material. http://cms-results.web.cern.ch/cms-results/public-results/publications/HIG-17-002/index.html.

[76] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9(1):60–62, 03 1938.

[77] Glen Cowan. Discovery sensitivity for a counting experiment with background uncertainty, May 2012. http://www.pp.rhul.ac.uk/ cowan/stat/medsig/medsigNote.pdf.

[78] Prospects for HH measurements at the HL-LHC. Technical Report CMS-PAS-FTR-18-019, CERN, Geneva, 2018.

[79] Maria Cepeda, Stefania Gori, Philip James Ilten, Marumi Kado, and Francesco Riva. Higgs Physics at the HL-LHC and HE-LHC. Technical Report CERN-LPCC-2018-04, CERN, Geneva, 2018.

[80] Cms hig-17-030 supplementary material. http://cms-results.web.cern.ch/cms-results/public-results/publications/HIG-17-030/index.html.

[81] Combination of searches for Higgs boson pair production in proton-proton collisions at $\sqrt{s} = 13$ TeV. Technical Report CMS-PAS-HIG-17-030, CERN, Geneva, 2018.

[82] Albert M Sirunyan et al. Search for Higgs boson pair production in the $\gamma\gamma\mathrm{b\bar{b}}$ final state in pp collisions at $\sqrt{s} = 13$ TeV. 2018.

[83] Albert M. Sirunyan et al. Search for resonant pair production of Higgs bosons decaying to bottom quark-antiquark pairs in proton-proton collisions at 13 TeV. *JHEP*, 08:152, 2018.

[84] A. M. Sirunyan et al. Search for a massive resonance decaying to a pair of Higgs bosons in the four b quark final state in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Phys. Lett.*, B781:244–269, 2018.

[85] Search for resonant and non-resonant production of Higgs boson pairs in the four b quark final state using boosted jets in proton-proton collisions at sqrt(s)=13 TeV. Technical Report CMS-PAS-B2G-17-019, CERN, Geneva, 2018.

[86] Search for Non-Resonant Higgs Pair-Production in the $\mathrm{b\bar{b}b\bar{b}}$ Final State with the CMS detector. Technical Report CMS-PAS-HIG-17-017, CERN, Geneva, 2018.

[87] Albert M Sirunyan et al. Search for resonant and nonresonant Higgs boson pair production in the $\mathrm{b\bar{b}}\ell\nu\ell\nu$ final state in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 01:054, 2018.