Search with the CMS experiment for a heavy scalar boson decaying into a pair of Standard-Model-like Higgs bosons in the final states $b\bar{b}\tau^+\tau^-$ with one $\tau$ decaying hadronically and the other leptonically

Tesi di Dottorato di Ricerca in Fisica Sperimentale
In partial fulfillment of the requirements for the Ph.D. Thesis in Experimental Physics

Candidato: Konstantin Androsov
Supervisor: Dott.sa Maria Agnese Ciocci

Ciclo di dottorato XXVIII
# Contents

<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>5</td>
</tr>
<tr>
<td>Introduction</td>
<td>7</td>
</tr>
<tr>
<td>1 Theoretical basis and motivations</td>
<td>9</td>
</tr>
<tr>
<td>1.1 The Higgs mechanism</td>
<td>9</td>
</tr>
<tr>
<td>1.1.1 Introduction to the Higgs mechanism</td>
<td>9</td>
</tr>
<tr>
<td>1.1.2 The Higgs mechanism in the SM</td>
<td>10</td>
</tr>
<tr>
<td>1.2 The discovery of the Higgs boson and beyond</td>
<td>12</td>
</tr>
<tr>
<td>1.3 Supersymmetry and the two Higgs doublet models</td>
<td>13</td>
</tr>
<tr>
<td>1.3.1 Supersymmetry</td>
<td>14</td>
</tr>
<tr>
<td>1.3.2 The minimal supersymmetric standard model</td>
<td>15</td>
</tr>
<tr>
<td>1.3.3 The MSSM Higgs sector and the two Higgs doublet models</td>
<td>16</td>
</tr>
<tr>
<td>1.4 Heavy Higgs production at the LHC</td>
<td>18</td>
</tr>
<tr>
<td>1.5 Motivation of the search of $H \rightarrow hh \rightarrow bb\tau\tau$ at CMS</td>
<td>20</td>
</tr>
<tr>
<td>2 The experimental framework</td>
<td>25</td>
</tr>
<tr>
<td>2.1 The Large Hadron Collider</td>
<td>25</td>
</tr>
<tr>
<td>2.2 The Compact Muon Solenoid detector</td>
<td>28</td>
</tr>
<tr>
<td>2.2.1 The inner tracking system</td>
<td>30</td>
</tr>
<tr>
<td>2.2.2 Calorimeter system</td>
<td>32</td>
</tr>
<tr>
<td>2.2.3 Superconducting magnet</td>
<td>36</td>
</tr>
<tr>
<td>2.2.4 The muon system</td>
<td>37</td>
</tr>
<tr>
<td>2.2.5 Luminosity measurement</td>
<td>39</td>
</tr>
<tr>
<td>2.2.6 Triggers and data processing</td>
<td>40</td>
</tr>
<tr>
<td>3 Event reconstruction and physical objects</td>
<td>43</td>
</tr>
<tr>
<td>3.1 High level reconstruction objects</td>
<td>43</td>
</tr>
<tr>
<td>3.1.1 Track reconstruction</td>
<td>43</td>
</tr>
<tr>
<td>3.1.2 Primary vertex</td>
<td>44</td>
</tr>
<tr>
<td>3.1.3 Particle-flow</td>
<td>46</td>
</tr>
<tr>
<td>3.2 Physical objects</td>
<td>49</td>
</tr>
<tr>
<td>3.2.1 Electrons</td>
<td>49</td>
</tr>
</tbody>
</table>
Abstract

This thesis reports a search for a heavy scalar boson $H$ predicted by the Two-Higgs-Doublet Models (2HDM) and, in particular, by the Minimal Supersymmetric Extension of the Standard Model (MSSM). We consider a decay mode where the heavy Higgs decays into a pair of lighter “Standard-Model-like” Higgs bosons $h$ with mass $125$ GeV. The final state studied consists of a pair of $\tau$ leptons and a pair of $b$ jets from either of the $h$. One of the $\tau$ leptons decays into an electron or muon, while the second $\tau$ lepton decays hadronically. The search is done using proton–proton collision data collected by the Compact Muon Solenoid experiment in 2012, corresponding to an integrated luminosity of $19.7 \text{ fb}^{-1}$. No evidence of the heavy boson $H$ was found, and the observed results are compatible with the predictions of the Standard Model. The upper limits on the $H$ production cross section times the branching fraction of $hh \to bb\tau\tau$ in the mass range $260 \leq m_H \leq 350 \text{ GeV}$ and the exclusion limits for the parameter regions of the 2HDM Type 2 and of the MSSM phase space have been set.
Introduction

The discovery of the Higgs boson at the Large Hadron Collider (LHC) in 2012 by the CMS and ATLAS experiments [1][2] was an important confirmation of the Standard Model (SM) [3]. Experimental results show that the SM provides precise predictions in the low energy scale [4].

However, the SM cannot be a full description of Nature for a series of reasons. First of all, it does not contain a description of the gravitational interactions that is compatible with the general theory of relativity [5] and, therefore, will not be valid at the Plank energy scale and higher. Secondly, cosmological measurements [6][7] are in good agreement with a cold dark matter ($\Lambda$-CDM) model [8] predictions, which indicates the presence in the universe of a matter that does not emit electromagnetic radiation. In the SM, there are no potential candidate particles that could be constituents of dark matter. Thirdly, the SM predicts that neutrinos are massless particles, therefore it cannot incorporate the neutrino oscillation effect [9], whose existence is experimentally confirmed [10][11]. Fourthly, the level of charge−parity symmetry (CP-symmetry) violation predicted by the SM is too small to explain the baryon asymmetry in the observed universe [12]. And, lastly, quantum corrections to the Higgs mass, within the SM, contain quadratic divergencies at a high energy scale, that can vanish only if the bare Higgs mass is fine-tuned, so that contributions from Yukawa and gauge coupling constants cancel each other up to the Plank scale [13][14].

This opens up opportunities to develop various Beyond Standard Model (BSM) theories to solve at least some of these problems, and to search for experimental signatures of new physics. Supersymmetry (SUSY) is such model, which predicts an unification of the fundamental interactions at the Plank energy scale and may explain baryon asymmetry and the large gauge hierarchy [4][15]. It introduces an additional symmetry for fermion–boson transformations and provides mechanisms to break this symmetry.

One particular case of SUSY is the Minimal Supersymmetric Standard Model (MSSM), which extends the SM with the minimal amount of super-partner particles. To achieve the unification of the fundamental interactions within the MSSM, it is necessary that the masses of the super-partners lie below the scale of a few TeV. On the other hand, MSSM could be considered as a particular case of the Two-Higgs-Double Models (2HDM) [16] that generalizes the properties of theories, which effective low energy model predicts existence of two Higgs doublets. The Higgs sector of a 2HDM consists of 5 physical particles: two CP-even neutral Higgses ($h$, $H$), one CP-odd neutral Higgs ($A$), and two charged Higgses ($H^+$, $H^-$).
A sizable part of the parameter space of the MSSM and 2HDM, which has not been fully explored by previous experimental searches, becomes available with the data delivered by Run 1 of the LHC. Using these data, a direct search for the signature of the second Higgs doublet can cover additional part of the parameter space, not accessible by other complementary BSM searches. Depending on the mass range, the $H \to hh$ channel can be one of the most sensitive direct searches for an additional Higgs doublet.

The work described in this thesis is the search for a heavy Higgs $H$ decaying into two SM-like Higgses $h$ in the mass range $260 \leq m_H \leq 350$ GeV. The final state of one $h$ is a pair $b$ jets, $h \to b\bar{b}$\footnote{Throughout this thesis, $h \to bb$ denotes $h \to b\bar{b}$.} and that of the other is a $\tau$ pair, $h \to \tau_\tau\tau_\tau$\footnote{Throughout this thesis, $h \to \tau_\tau\tau_\tau$ denotes $h \to \tau^+\tau^-\tau^+\tau^-$.} where one $\tau$ decays hadronically, $\tau_\tau$\footnote{By $\tau_h$ we denote all possible hadronic decays of $\tau^-$ or their charge conjugates.}, and the other decays leptonically, $\tau_\tau$\footnote{By $\tau_l$ we denote the leptonic $\tau$ decays: $\tau^- \to l^- \bar{\nu}_l \nu_\tau$, where $l$ is an electron or a muon, or the charge conjugate state.}. The data used for this search were collected at the Compact Muon Solenoid (CMS) experiment\footnote{The proceedings of the poster session will be published online by the Proceedings of Science.} in 2012. The results of this search provided model-independent limits for a heavy scalar $H$ as well as model dependent interpretations for the MSSM and 2HDM Type 2 phase spaces.

In the first chapter, the theoretical basis and motivations for a $H \to hh \to bb\tau\tau$ search are presented. The second chapter describes the principal components and features of the CMS detector that are relevant to this analysis. In the third chapter, reconstruction algorithms of the main objects of analysis are introduced. The analysis strategy for event selection and signal signature extraction is explained in the fourth chapter. The fifth chapter contain the description of the techniques used to model the signal and background contributions. In the sixth chapter, algorithms to extract the exclusion limits are defined and the final results are presented. At the end, the conclusions from the work are presented and outlooks for possible future research are discussed.

The results of the analysis described in this thesis are meant to be published in the journal, *Physics Letters B*, and are currently available on arXiv\footnote{The proceedings of the conference will be published by World Scientific Publ. Co, Singapore.}.

I presented this analysis, on behalf on the CMS collaboration, as part of the CMS results for the analyses with $h \to \tau\tau$ at the 17th Lomonosov Conference on Elementary Particle Physics, 20–26 Aug 2015, Moscow State University, Moscow. The proceedings of the conference will be published by World Scientific Publ. Co, Singapore.

I am also a co-author of the poster dedicated to the analyses of $H \to hh \to bb\tau\tau$ and $A \to Zh \to ll\tau\tau$ presented at the XXVII International Symposium on Lepton Photon Interactions at High Energies, 17–22 Aug 2015, Ljubljana, Slovenia. The proceedings of the poster session will be published online by the Proceedings of Science.
Chapter 1

Theoretical basis and motivations

1.1 The Higgs mechanism

The mechanism that allows the gauge bosons to obtain their masses by means of the spontaneous symmetry breaking for the non-Abelian theories was introduced by P. W. Higgs [19,20], and by F. B. Englert and R. Brout [21], and is now an essential ingredient of the Standard Model (SM). In this section we will summarize the key points of this mechanism, which are indispensable for understanding the beyond standard model searches in the Higgs sector discussed in the following sections.

The development of the SM was a collaborative effort of many theoreticians and experimentalists over several decades. There are many publications that summarize this development and describe the SM as a whole, including the Higgs mechanism. The summary presented in this section is essentially inspired by the book An Introduction to quantum field theory by Peskin and Schroeder [22], and by the article The anatomy of electro–weak symmetry breaking. I: The Higgs boson in the Standard Model by Djouadi [23].

1.1.1 Introduction to the Higgs mechanism

To understand the foundation of the Higgs mechanism, we will first consider the general case.

Let the Lagrangian of some model be invariant under the actions of a local gauge symmetry group $G$ and let this Lagrangian depend on several real scalar fields $\phi_i$ for which the group $G$ is represented by the transformation

$$\phi_i \rightarrow (1 - \alpha^a T^a)_{ij} \phi_j,$$

(1.1)

where $iT^a$ are generators of $G$ and $T^a$ are some real antisymmetric matrices.
In this case, the covariant derivative of $\phi_i$ is

$$D_\mu \phi_i = (\partial_\mu + g A_\mu^a T^a) \phi_i.$$  \hfill (1.2)

Therefore, the kinetic energy associated with the scalar fields is

$$E_{\text{kin}} = \frac{1}{2} (D_\mu \phi_i)^2 = \frac{1}{2} (\partial_\mu \phi_i)^2 + g A_\mu^a (\partial_\mu \phi_i T^a) + \frac{1}{2} g^2 A_\mu^a A_\nu^b (T^a \phi)_i (T^b \phi)_i.$$  \hfill (1.3)

Now, if the fields acquire non-zero vacuum expectation values, the last term in Equation (1.3) will be responsible for gauge bosons’ becoming massive when the Lagrangian is expanded around the vacuum expectation:

$$\Delta \mathcal{L}_m = \frac{1}{2} m^2_{ab} A_\mu^a A_\nu^b,$$  \hfill (1.4)

where the mass matrix $m^2_{ab} = g^2 (T^a \langle \phi \rangle)_i (T^b \langle \phi \rangle)_i$.

The diagonal elements of $m^2_{ab}$ correspond to the squared masses of the bosons:

$$m^2_{aa} = g^2 (T^a \langle \phi \rangle)^2 \geq 0$$  \hfill (1.5)

It follows from Equation (1.5) that if the vacuum state is invariant under some generator $i T^a$, i.e. $T^a \langle \phi \rangle = 0$, the corresponding gauge boson will remain massless, while the others will acquire masses. These results are valid for a theory with any number of scalar fields, independently of their exact form.

### 1.1.2 The Higgs mechanism in the SM

In the Standard Model, the introduction of one scalar $SU(2)$ doublet results in the $SU(2) \times U(1)_Y$ symmetry breaking to $U(1)_{EM}$. The vacuum expectation of a scalar $SU(2)$ doublet with a hypercharge 1 is invariant under the action of one of the combinations of the $SU(2) \times U(1)_Y$ generators. If we choose the basis where the vacuum expectation of the scalar field is

$$\langle \phi \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix},$$  \hfill (1.6)

after following steps similar to eqs. (1.2) to (1.4) for the $SU(2) \times U(1)_Y$ case, the part of the Lagrangian that is related to the boson masses will be

$$\Delta \mathcal{L}_m = \frac{v^2}{8} \left( g^2 (A_\mu^1)^2 + g^2 (A_\mu^2)^2 + (-g A_\mu^3 + g' B_\mu)^2 \right),$$  \hfill (1.7)

where $A_\mu^a$ are the three $SU(2)$ gauge bosons with a coupling $g$, and $B_\mu$ is the $U(1)$ gauge boson with a coupling $g'$.
1.1. The Higgs mechanism

By rotating the basis, the four physical vector bosons can be introduced. Three of them became massive, while one remains massless:

\[
W_\mu^{\pm} = \frac{1}{\sqrt{2}} \left( A_\mu^1 \mp i A_\mu^2 \right) \quad m_W = \frac{v}{2} g \\
Z_\mu^0 = \frac{1}{\sqrt{g^2 + g'^2}} \left( g A_\mu^3 - g' B_\mu \right) \quad m_Z = \frac{v}{2} \sqrt{g^2 + g'^2} \\
A_\mu = \frac{1}{\sqrt{g^2 + g'^2}} \left( g' A_\mu^3 + g B_\mu \right) \quad m_A = 0
\]

The Higgs mechanism also provides masses to the fermions by means of the Yukawa couplings, without explicitly breaking the \( SU(3) \times SU(2) \times U(1) \) symmetry of the SM.

The Lagrangian part related to the fermion masses, expended around the \( \phi \) vacuum expectation, has the form:

\[
\Delta \mathcal{L}_f = - \sum_f m_f \bar{f}_L f_R + \text{h.c.,} \quad (1.9)
\]

where \( m_f = \frac{1}{\sqrt{2}} \lambda_f v \).

To describe the properties of the quantum of the scalar field, which is called the Higgs boson, the scalar field \( \phi \) in the unitary gauge may be parametrized by

\[
\phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}, \quad (1.10)
\]

where \( h(x) \) is a real scalar field with \( \langle h(x) \rangle = 0 \), so that the condition Equation (1.6) is satisfied.

The \( \phi \) potential that leads to a renormalizable Lagrangian, which is compatible with Equation (1.6), is

\[
V(\phi) = \mu^2 \phi^\dagger \phi + \lambda (\phi^\dagger \phi)^2. \quad (1.11)
\]

It has a minimum at \( v = \left( -\frac{\mu^2}{\lambda} \right)^{1/2} \).

The Lagrangian part related to this potential energy is:

\[
\Delta \mathcal{L}_V = -\lambda v^2 h^2 - \lambda v h^3 - \frac{\lambda}{4} h^4, \quad (1.12)
\]

from which the mass of the Higgs boson is obtained as \( m_h = \sqrt{2\lambda v} \).
Chapter 1. Theoretical basis and motivations

By expanding \( \phi(x) \) in Equation (1.10) around the vacuum state, in addition to the boson mass terms Equation (1.7) and the fermion mass terms Equation (1.9), we obtain the Higgs boson couplings to the other bosons and fermions:

\[
\Delta \mathcal{L}_h = \left( m_W^2 W^\mu W^-\mu + \frac{1}{2} m_Z^2 Z^\mu Z^-\mu \right) \frac{h}{v} \left( 2 + \frac{h}{v} \right) - \sum_f m_f \bar{f} f \frac{h}{v}.
\] (1.13)

1.2 The discovery of the Higgs boson and beyond

The discovery of the Higgs boson was announced in 2012 [1,2], and as of today, the combined measurement performed by the CMS and ATLAS experiments in proton–proton collisions at 7 and 8 TeV centre of mass energy shows the Higgs boson mass to be 125.09 ± 0.21 (stat.) ±0.11 (syst.) GeV [24].

The Higgs boson couplings to the other bosons and fermions, measured in \( H \to \gamma\gamma, H \to ZZ, H \to WW, H \to bb \) and \( H \to \tau\tau \) decay channels, are consistent within ±1σ confidence interval with the SM predictions, as shown in Figure 1.1. The measurement of the Higgs charge–parity (CP) properties, its spin, and the upper limit on the Higgs decay width, are also consistent with the predictions of the SM [25–29].

One way to further verify the consistency the predictions of the SM in the Higgs sector is to proceed with precision measurements of the SM parameters, including the Higgs mass and the couplings to the fermions and other bosons, and to directly measure the Higgs self-coupling (\( \lambda_{hhh} = 3m_h^2/v^2 \), from Equation (1.12)) by studying double-Higgs production modes. However, the data collected so far at LHC by ATLAS and CMS (approximately 25fb\(^{-1}\) in total per experiment) are insensitive to the self-coupling in the SM [31,32], because of the expected small signal rates [33,34] and large backgrounds.

On the other hand, the available measurements do not exclude the possibility that the discovered Higgs boson belongs to the Higgs sector predicted by models beyond the SM (BSM). As discussed in the next sections, the Higgs sector of some of these models, requires two Higgs doublets [16,35], and one of the neutral heavy Higgs of these models can decay into two lighter “SM-like” Higgs, with a signal rate, depending on the model parameters, significantly larger than that of the SM double Higgs production. Furthermore, for these models, the invariant mass of the two SM-like Higgses is another handle to discriminate the signal of decaying Higgs. The search for a resonant pair of “SM-Higgs-like” Higgses is a probe for New Physics Beyond the Standard Model, since the current achieved statistics at LHC can be sufficient to cover some part of the parameter spaces of these BSM models, which was not yet excluded from the other measurements [36,37].
1.3 Supersymmetry and the two Higgs doublet models

As was shown in Section 1.1.1, the Higgs mechanism may be realized with an arbitrary number of scalar fields. The Standard Model realizes the simplest case by introducing only one scalar field. The main motivations for thinking that in Nature there could be more than one scalar field comes from a theory called supersymmetry (SUSY). On one hand, SUSY solves most of the issues that the Standard Model leaves open, as discussed in the Introduction. On the other hand, SUSY is more appealing than the SM from the point of view of mathematical consistency, because, as was proved by Haag, Sohnius and Lopuszanski, the supersymmetry algebra is the only graded Lie algebra of symmetries of the $S$-matrix consistent with relativistic quantum field theory \cite{38}.

While up to now there has been no experimental evidence for SUSY, there is still a significant part of the SUSY parameter space that is not yet excluded by the experimental measurements. Moreover, one of the simplest SUSY realizations, the minimal supersymmetric standard model (MSSM), requires that the masses of the

Figure 1.1: Negative log-likelihood contours of the relative to the SM predictions Higgs couplings to fermions ($k^f_F$) versus the relative to the SM predictions Higgs couplings to bosons ($k^f_V$) for the combined ATLAS and CMS measurements for $H \rightarrow \gamma\gamma$, $H \rightarrow ZZ$, $H \rightarrow WW$, $H \rightarrow bb$ and $H \rightarrow \tau\tau$ decay channels, without any assumptions on the signs of $k^f_F$ and $k^f_V$. The two quadrants with negative $k^f_V$ are symmetric with respect to the point $(0,0)$. \cite{30}.
new particles, the “superpartners” of the SM particles, should not be greater than a few GeV, to avoid a fine tuning of the model.

In this section there will be described the few key properties of the Higgs sector of the MSSM that motivated this thesis’s analysis of the experimental data. A detailed description of SUSY can be found in *Supersymmetry and supergravity* by Wells and Bagger [39] or in *A supersymmetry primer* by Martin [40].

### 1.3.1 Supersymmetry

Supersymmetry states that there is a symmetry that transforms a half-spin particle to a particle with an integer spin, and vice versa. Irreducible representations of the supersymmetry algebra are called supermultiplets. Each supermultiplet consist of a half-spin particle and its integer-spin superpartner.

Supermultiplets may be divided into two types: chiral, a combination of a complex scalar field \( \phi_i \) with a left-handed Weyl fermion \( \psi_i \); and gauge, a combination of a spin-1 gauge boson and its spin-1/2 superpartner. Fermions may only have scalar superpartners, because their left-handed and right-handed parts transform differently under the gauge group, and therefore they are represented only within the chiral supermultiplets.

Invariance under the supersymmetry transformations puts many restrictions on the possible form of a renormalizable Lagrangian. As a consequence, the part of a renormalizable SUSY Lagrangian that corresponds to the interactions of chiral supermultiplets should have the following general form [40]:

\[
\Delta \mathcal{L}_{\text{int}} = -\frac{1}{2} \left( W_{ij} \psi_i \psi_j + W_{ij}^* \psi_i^\dagger \psi_j^\dagger \right) - W_i^* W_i, \tag{1.14}
\]

where \( W \) is the superpotential of the scalar fields \( \phi_i \):

\[
W^i = \frac{\delta W}{\delta \phi_i}, \quad W^{ij} = \frac{\delta^2 W}{\delta \phi_i \delta \phi_j}. \tag{1.15}
\]

The form of the superpotential for a SUSY theory without gauge singlet chiral supermultiplets is

\[
W = \frac{1}{2} M^{ij} \phi_i \phi_j + \frac{1}{6} y^{ijk} \phi_i \phi_j \phi_k, \tag{1.16}
\]

where \( M^{ij} \) is the symmetric mass matrix for the fermion fields and \( y^{ijk} \) is the symmetric tensor of Yukawa couplings of a scalar \( \phi_k \) and two fermions \( \phi_i, \phi_j \).

Therefore, if the supersymmetry is an exact symmetry, the masses of the particles and their superpartners must be the same. Since no experimental evidence for such superpartners for the observed SM particles has been found, the SUSY symmetry must be broken.
1.3.2 The minimal supersymmetric standard model

There is no commonly accepted dynamical way to break SUSY, so symmetry-breaking terms should be placed manually in the Lagrangian to construct an effective low-energy SUSY model. The minimal supersymmetric standard model (MSSM) is one such model, which extends the SM by introducing the minimal number of new particles needed to be compatible with supersymmetry. The assumptions that define MSSM may be found, for example, in [35,40].

As in the SM, the MSSM gauge group is $SU(3)_C \times SU(2)_L \times U(1)_Y$ and only three generations of quarks and fermions are present, as shown in Table 1.1. The supersymmetry is broken by explicitly adding mass terms and trilinear couplings for scalar particles in such a way that quadratic divergencies do not appear, without any need for fine tuning, which is called soft-SUSY-breaking. The lepton and baryon numbers are conserved by imposing the conservation of a discrete symmetry called R-parity:

$$R_p = (-1)^{2s+3B+L},$$

where $s$, $L$ and $B$ are, respectively, the spin, the leptonic, and the baryon number. For the standard model particles, $R_p = +1$, while for their superpartners, $R_p = -1$.

<table>
<thead>
<tr>
<th>Superfield</th>
<th>SM particles</th>
<th>Superpartners</th>
<th>Quantum numbers $SU(3)_C \times SU(2)_L \times U(1)_Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tilde{G}_a$</td>
<td>$G^a$</td>
<td>gluinos, $G_a$</td>
<td>8 \ 1 \ 0</td>
</tr>
<tr>
<td>$\tilde{W}_a$</td>
<td>$W^a$</td>
<td>winos, $\tilde{W}_a$</td>
<td>1 \ 3 \ 0</td>
</tr>
<tr>
<td>$\tilde{B}$</td>
<td>$B^a$</td>
<td>bino, $\tilde{B}$</td>
<td>1 \ 1 \ 0</td>
</tr>
<tr>
<td>$\tilde{Q}$</td>
<td>$(u_L,d_L)$</td>
<td>$(\tilde{u}_L,d_L)$</td>
<td>3 \ 2 \ 1/3</td>
</tr>
<tr>
<td>$\tilde{U}^c$</td>
<td>$\tilde{u}_R$</td>
<td>$\tilde{u}^*_R$</td>
<td>3 \ 1 \ -4/3</td>
</tr>
<tr>
<td>$\tilde{D}^c$</td>
<td>$\tilde{d}_R$</td>
<td>$\tilde{d}^*_R$</td>
<td>3 \ 1 \ 2/3</td>
</tr>
<tr>
<td>$\tilde{L}$</td>
<td>$(\nu_L,e_L)$</td>
<td>$(\tilde{\nu}_L,\tilde{e}_L)$</td>
<td>1 \ 2 \ -1</td>
</tr>
<tr>
<td>$\tilde{E}^c$</td>
<td>$\tilde{e}_R$</td>
<td>$\tilde{e}^*_R$</td>
<td>1 \ 1 \ 2</td>
</tr>
</tbody>
</table>

Table 1.1: The list of the MSSM superfields composed by the SM particles and their superpartners, and their quantum numbers [35]. The MSSM Higgs sector is composed of the two chiral superfields.

However, without additional assumptions, this leaves 105 new unknown parameters in the MSSM that are introduced by soft-SUSY-breaking. Therefore, to increase the experimental predictability of the model, some additional assumptions are usually made.

To describe the phenomenological properties, only a constrained part of the parameters phase space is considered. This allows defining the so-called benchmark
Chapter 1. Theoretical basis and motivations

1.3.3 The MSSM Higgs sector and the two Higgs doublet models

The SUSY superpotential (Equation 1.16) depends only on the superfields, but not on their conjugates. The form of the SUSY Lagrangian (Equation 1.14) implies that in SUSY, one scalar field cannot provide masses for both $1/2$ and $-1/2$ isospin fermions without explicitly breaking the SUSY symmetries. For that reason, in the MSSM, two chiral superfields are introduced, in such a way that their scalar components $\phi_1$ and $\phi_2$ make the other particles massive.

However, MSSM is only one of the many phenomenologically interesting theories and at this stage more general reasoning may be performed to cover all hypothetical theories, whose effective low-energy model contains two scalar doublets that provide masses to the other particles by means of the Higgs mechanism.

An ensemble of such effective models is called the two Higgs doublet models (2HDM).

If no additional assumptions are made, the scalar 2HDM potential depends on 14 parameters and is quite complicated from a phenomenological point of view. However, to construct models that could be tested in experiments that could be carried out in the near future, several reasonable simplifications might be applied [16]:

- CP-symmetry is not spontaneously broken and is conserved in the Higgs sector.
- Tree level flavour-changing neutral currents (FCNC) are absent. Introduced symmetries, which ensure the absence of FCNC, will also eliminate quadratic terms (see Equation (1.20)), where one of the doubles appears an odd number of times.

Both 2HDM scalar fields are $SU(2)$ doublets. If we assume, without loss of generality, that both doublets have hypercharge $+1$, the vacuum expectation of each doublet can be written in a way similar to the standard model case (Equation 1.6):

$$\langle \phi_a \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v_a \end{pmatrix},$$

(1.18)

where $a = 1, 2$.

Therefore, each doublet can be written as a combination of four scalar fields:

$$\phi_a = \frac{1}{\sqrt{2}} \begin{pmatrix} \varphi_a^+ \sqrt{2} \\ v_a + \rho_a + i\eta_a \end{pmatrix}.$$ 

(1.19)
Following a procedure similar to the one used for the SM (eqs. (1.7) and (1.8)), three out of the eight fields will be fixed to give mass to $W^\pm$ and $Z$ bosons. The other 5 form the 2HDM physical Higgs sector: two neutral scalars (the heavy Higgs $H$ and the light Higgs $h$), one neutral pseudoscalar ($A$), and two charged scalars ($H^\pm$).

The 2HDM scalar potential, assuming the simplifications mentioned above, is

\[
V = \sum_{a,b=1}^{2} (-1)^{a+b} m_{ab}^2 \phi_a^\dagger \phi_b + \frac{\lambda_a}{2} (\phi_a^\dagger \phi_a)^2 \\
+ \lambda_3 \phi_1^\dagger \phi_1 \phi_2^\dagger \phi_2 + \lambda_4 \phi_1^\dagger \phi_2 \phi_1^\dagger \phi_1 + \frac{\lambda_5}{2} \left[ (\phi_1^\dagger \phi_2)^2 + h.c. \right].
\]  

(1.20)

After expanding $\phi_a$ (Equation (1.19)) into the 2HDM potential, the squared mass matrices for $\varphi_a^+, \rho_a$ and $\eta_a$ fields might be obtained. The mixing angles that diagonalize these matrices are the two most important 2HDM parameters: the angle $\alpha$ that diagonalizes the neutral scalars matrix and the angle $\beta$ ($\tan \beta \equiv \frac{v_2}{v_1}$) that diagonalizes the pseudoscalars and the charged scalars matrices.

By rotating the basis, we can express the neutral physical fields as functions of $\rho_a$ and $\eta_a$ using the definitions of $\alpha$ and $\beta$:

\[
h = \rho_1 \sin \alpha - \rho_2 \cos \alpha, \quad H = -\rho_1 \cos \alpha - \rho_2 \sin \alpha, \quad A = \eta_1 \sin \beta - \eta_2 \cos \beta.
\]  

(1.21)

The 2HDM models with neutral flavor conservation should satisfy the conditions of the Paschos–Glashow–Weinberg theorem, which states that a necessary and sufficient condition for the absence of FCNC at the tree level is that all the fermions of a given charge and helicity transform according to the same irreducible representation of $SU(2)$ and correspond to the same eigenvalue of $T_3$, and that a basis exists in which they receive their contributions in the mass matrix from a single source [16].

This condition is satisfied by introducing a continuous or discrete symmetry. As a result, the 2HDM models without FCNC at the tree level can be divided into four types, depending on how the quarks and leptons are coupled to the Higgs field:\n
- **Type 1**: all quarks and leptons couple to $\phi_2$;

- **Type 2**: up-type quarks ($Q_{EM} = 2/3$) couple to $\phi_2$, down-type quarks ($Q_{EM} = -1/3$) and leptons couple to $\phi_1$;

- **Type 3**: all quarks couple to $\phi_2$, leptons couple to $\phi_1$;

- **Type 4**: up-type quarks and leptons couple to $\phi_2$, down-type quarks couple to $\phi_1$.

\footnote{The choice between $\phi_1$ and $\phi_2$ in the following definitions is conventional.}
Chapter 1. Theoretical basis and motivations

The 2HDM Type 2

In this thesis, we will focus our attention on the 2HDM Type 2\(^2\) because its parameter space also covers the Higgs sector of the MSSM. However, the supersymmetry puts additional constraints on the MSSM Higgs sector.

The mixing angle \(\alpha\) and the masses of the charged and neutral scalars \(m_{H^\pm}\) and \(m_{h,H}\), which are free parameters in the general 2HDM Type 2 phase space, can be expressed as functions of \(m_A, m_W, m_Z\) and \(\beta\) in the MSSM [35]:

\[
\alpha = \frac{1}{2} \arctan \left( \tan 2\beta \frac{m_A^2 + m_Z^2}{m_A^2 - m_Z^2} \right),
\]

\[
m_{h,H}^2 = \frac{m_A^2 + m_Z^2}{2} \left( 1 \mp \sqrt{1 - \left( \frac{2m_A m_Z}{m_A^2 + m_Z^2} \right)^2 \cos 2\beta} \right),
\]

\[
m_{H^\pm}^2 = m_A^2 + m_W^2.
\]

For this reason, the present thesis will provide an interpretation of the results for two benchmark scenarios: one for the general 2HDM Type 2 and one for the MSSM.

1.4 Heavy Higgs production at the LHC

The parameter spaces of the 2HDM and MSSM models are large and it is possible to interpret the Higgs boson observed at LHC (Section 1.2) within these models. To achieve this, we suppose that the discovered LHC Higgs is the light 2HDM \(h\) and its mass is \(\approx 125\) GeV. In the MSSM case, to be independent of the possible unaccounted high order corrections in the theoretical calculations, the part of the MSSM parameter space considered corresponds to a mass of the light Higgs \(122 \leq m_h \leq 128\) GeV [37].

In addition it is also required that the \(h\) coupling should be compatible with the LHC measurements and, therefore, with the couplings of the SM Higgs (Equation (1.13)).

In the 2HDM, the \(h\) Yukawa couplings to the SM fermions are defined by

\[
- \sum_{f=u,d,l} \xi_f^h m_f \tilde{f} \frac{h}{v},
\]

where the \(\xi_f^h\) are constants that define how different the 2HDM \(h\) couplings are from the SM Higgs couplings. If \(h\) is exactly an SM-like Higgs, these constants are equal to one.

\(^2\)In the following text, if not explicitly specified, the 2HDM Type 2 model is always assumed when speaking about the 2HDM.
1.4. Heavy Higgs production at the LHC

\( \xi_h^l \) can be expressed in terms of \( \alpha \) and \( \beta \):

\[
\xi_h^l = \xi_h^d = -\frac{\sin \alpha}{\cos \beta}, \quad \xi_h^u = \frac{\cos \alpha}{\sin \beta}.
\]  

(1.24)

The region of the 2HDM parameter space where the \( h \) couplings are similar to the SM Higgs couplings lies near the alignment limit, where \( \cos(\beta - \alpha) = 0 \) and \( h \) becomes the SM Higgs.

In the low \( \tan \beta \) region of the MSSM (2HDM) parameter space, the expected production rates of the \( A \) and \( H \) Higgses through gluon fusion (\( gg \rightarrow A \) and \( gg \rightarrow H \)) become large enough that the data collected in the 2012 LHC could be sensitive to their presence. In this thesis we will focus on \( gg \rightarrow H \) production, while different scenarios of the CP-odd Higgs boson \( A \) production were considered in the other searches [18,41–43].

The branching fractions of the different decay modes of the heavy CP-even Higgs \( H \) in the 2HDM parameter space strongly depend on its mass \( m_H \) and \( \tan \beta \). Near to the alignment limit, for low \( \tan \beta \) and \( m_H \) between 2\( m_h \) and \( 2m_{\text{top}} \), where on-shell \( H \rightarrow t\bar{t} \) decay is kinematically forbidden, the \( H \rightarrow hh \) decay mode is dominant [36], as shown in Figure 1.2. Similarly, Figure 1.3 shows that the \( H \rightarrow hh \) mode is one of the dominant decay modes for the MSSM parameter space [37].

The relatively high values of the predicted \( gg \rightarrow H \) cross-section and of the \( H \rightarrow hh \) branching fraction in the part of the MSSM and 2HDM parameter spaces with low \( \tan \beta \) makes this production channel very appealing. For this reason, the \( gg \rightarrow H \rightarrow hh \) process will be investigated in this thesis.

Another important motivation is that it is essential to quantify the sensitivity of the current datasets and to develop new tools and strategies (trigger and analysis techniques) for future measurements of the SM Higgs self-coupling for the phase of high luminosity of HL-LHC [44], foreseen after 2024.

For the work presented in this thesis, to interpret the analysis results, two benchmark scenarios were chosen:

- The MSSM low \( \tan \beta \) scenario with 122 \( \leq m_h \leq 128 \) GeV and 1 \( \leq \tan \beta \leq 4 \); the chosen region of the phase space is parametrized by \( \tan \beta \) and \( m_A \).

- The 2HDM Type 2 scenario considered near to the alignment limit with \( m_h = 125 \) GeV, \( m_H = m_A = m_{H^\pm} \) and \( m_{12} = m_A^2 \tan \beta/(1 + \tan^2 \beta) \); for the given \( m_H \) value, the phase space is parametrized by \( \tan \beta \) and \( \cos(\beta - \alpha) \).
Figure 1.2: Cross section times branching ratio of the heavy Higgs $H$ for different decay modes for $\sqrt{s} = 8$ TeV p–p collisions as a function of $m_H$ for two 2HDM Type 2 scenarios. 

In the first scenario (top) $\tan\beta = 1$ and $\cos(\beta - \alpha) = -0.11$, in the second scenario (bottom) $\tan\beta = 10$ and $\cos(\beta - \alpha) = -0.02$. $\lambda_5,6,7 = 0$ and $m_A = m_H$ for both scenarios. [36]

1.5 Motivation of the search of $H \to hh \to bb\tau\tau$ at CMS

The main motivations for a search of $gg \to H \to hh$ were described in the previous section. Now, assuming that the light Higgs $h$ is an SM-like Higgs in the chosen region of the parameter space of 2HDM and MSSM. It then has branching fractions that are very similar to the SM Higgs. Several decay modes of the $h$ pair produced in the process $gg \to H \to hh$ can be investigated using the data collected by the CMS detector at the LHC.

According to the predictions of the SM, the $h \to b\bar{b}$ decay mode has the highest branching fraction, $57.5 \pm 1.9\%$ [45], but, on the other hand, the $b$ jet reconstruction and identification has a substantial background from the other quark and gluon...
jets from the QCD processes. The other decay mode, \( h \rightarrow \gamma\gamma \), which was one of the original Higgs discovery channels, has a very good experimental signature, but a low branching fraction, \( 0.228 \pm 0.011 \) \cite{45}. The \( H \rightarrow hh \rightarrow b\bar{b}b\bar{b} \) and \( H \rightarrow hh \rightarrow \gamma\gamma b\bar{b} \) final states have already been studied in other CMS analyses \cite{46,47}.

In this thesis we study \( H \rightarrow hh \rightarrow b\bar{b}\tau^+\tau^- \) \cite{3} final state (see Figure 1.4), requiring that one of \( \tau \) (\( \tau^- \)) decays to a lighter lepton (\( e \) or \( \mu \)) and neutrinos, while the other \( \tau \) (\( \tau^+ \)) decays hadronically. The \( h \rightarrow \tau\tau \) branching fraction of \( 6.30 \pm 0.36\% \) \cite{45} is higher than that of \( h \rightarrow \gamma\gamma \). Furthermore, CMS shows good performance in electron, muon and \( \tau \) reconstruction and identification, as will be described in Chapter 3. Therefore, the results of the analysis of the \( H \rightarrow hh \rightarrow b\bar{b}\tau\tau \) final state will be complementary to the results of the other two analyses mentioned.

\footnote{For simplicity, from now on the charge of the leptons and the particle–antiparticle nature of quarks will be omitted, if the absence of their indication will not cause ambiguities.}
Figure 1.4: Lowest order Feynman diagram for the $gg \to H \to hh \to bb\tau\tau$ process at LHC.

above, with a comparable sensitivity.

The main backgrounds for the $H \to hh \to bb\tau\tau_h$ at the LHC are the SM processes. The cross-sections for several important SM processes according to the CMS measurements and their theoretical predictions are shown in Figure 1.5.

Figure 1.5: Production cross sections for different SM processes measured by the CMS at $\sqrt{s} = 7$ and 8 TeV p–p collisions in comparison with the theoretical predictions [48].

At LHC, the high number ($\sim 20$) of simultaneous proton–proton collisions per one proton–proton bunch crossing is the source of a considerable amount of background quark and gluon jets that could be misidentified as an electron, or muon, or $\tau_h$ or a $b$ jet, which is referred to as the pileup background. This background contributes cumulatively with the backgrounds listed below, by enhancing the
misidentification rates of all four final state objects, especially of the two b jets, in the $H \rightarrow hh \rightarrow bb\tau\tau$ signature.

Considering the $H \rightarrow hh \rightarrow bb\tau\tau$ signature (two b jets, a $\tau_l$ and a $\tau_h$) and the cross-sections of the SM processes, the main background processes expected for this analysis are:

- **$Z/\gamma^* \rightarrow l^+l^- + \text{jets}$, $l = e, \mu, \tau$**: The Z boson has a mass of 91.1876 ± 0.0021 GeV [4], which is close to the mass of $h$, therefore, due to the finite resolution of the detector, the signal from $Z/\gamma^* \rightarrow \tau\tau$ will be very similar to the $h \rightarrow \tau\tau$. $Z/\gamma^* \rightarrow ee$ and $Z/\gamma^* \rightarrow \mu\mu$ may contribute when the selected signal electron (muon) comes from Z decay, while the signal $\tau_h$ is a misidentified jet or the other electron (muon) from Z decay. Furthermore, two quark or gluon jets originating from the same vertex or from the pileup can be misidentified as two b jets from the $h \rightarrow bb$ signal.

- **$W \rightarrow l\nu_l + \text{jets}$, $l = e, \mu, \tau$**: The signal $\tau_l$ or $\tau_h$ can come from W decay, while 3 other signal objects are misidentified quark or gluon jets originating from the same vertex or from the pileup. Due to the high W+jets production cross-section, such a “combinatoric” contribution will be considerable.

- **QCD multijets**: the cross-section of the QCD processes is very big, therefore, even if the probability that all four signal objects will be faked by the QCD jets is low, the total QCD yield will be sizable.

- **$t\bar{t} + \text{jets}$**: $t\bar{t} \rightarrow b\bar{b}W^+W^-$ has two genuine b jets and when two W decay into $\tau_l$ and $\tau_h$, the $t\bar{t}$ contribution is irreducible. The hadronic W decay or jet from the same vertex or from the pileup can also be misidentified as a $\tau_h$ signal.

- **Di-boson + jets**: WW + jets, WZ + jets and ZZ + jets, where one boson decays into leptons and the other into quarks, or when both bosons decay into leptons. In all cases, in combination with the jets from the same vertex or pileup, the signature of di-boson production can be similar to the double-Higgs signature.

- **Single top, $tW$-channel**: there are two W produced in the $tW \rightarrow bWW$ process. This background is similar to that of the di-boson processes. In addition, in this case, a genuine b jet, coming from the top decay, can fake one of the signal b jets.

In the next chapters, after the description of the CMS detector and the algorithms for the reconstruction and identification of the physical objects, we will introduce the techniques used to reduce the presence of these backgrounds and to extract a possible signal contribution from the data. The reduction of the physical background and its modelling are important challenges for this search.
Chapter 2

The experimental framework

The experimental framework used for the search conducted in this work consists of the Compact Muon Solenoid (CMS) detector [17] and the Large Hadron Collider (LHC) [49]. This chapter contains a brief description of the principal features of the LHC and CMS, focusing on the elements relevant for the analysis.

2.1 The Large Hadron Collider

The LHC is a two-ring hadron beam accelerator and collider operated by the European Organization for Nuclear Research (CERN). It was built in the tunnel constructed for the LEP machine [50], which was dismantled in 2000. The tunnel lies 45–170 m below ground level and is 26.7 km long with eight straight sections and eight arcs. The LHC is designed to accelerate proton beams with centre of mass energy $\sqrt{s}$, up to 14 TeV, as well as lead ion ($^{208}\text{Pb}^{82+}$) beams with beam kinetic energy up to 2.76 TeV/nucleon, which corresponds to $\sqrt{s} = 1.15$ PeV for ion–ion collisions. This thesis uses data from proton–proton collisions.

The LHC is the main part of the CERN accelerator complex shown in Figure 2.1. Protons are collected by ionization of the Hydrogen gas source and accelerated in steps by a chain of accelerators. The first step is the Linear Accelerator (LINAC) 2, which is followed by the Proton Synchrotron Booster (PSB), the Proton Synchrotron (PS) and, at the end, the Super Proton Synchrotron (SPS), where the protons are accelerated up to the energy of 450 GeV and grouped into beams. Each beam is split into 2808 bunches with a nominal number of protons per bunch $1.15 \times 10^{11}$. The initially accelerated proton beams are then injected into the two LHC rings with opposite directions of rotation, due to the opposite magnetic fields in the two rings. The further beam acceleration, up to the collision energy, is done by a radio frequency (RF) cavity system, which at the same time adjusts the shape of the beams. The beams are guided within the LHC rings by superconducting niobium-titanium (Nb/Ti) magnets cooled to a temperature below 2 K and operating with a magnetic field above 8 Tesla.
The resulting centre of mass energy of the collisions delivered by the LHC was 7 TeV in 2011, 8 TeV in 2012, and 13 TeV in 2015.

There are four detectors at the beams-crossing points at the LHC. Two of them, CMS [17] and ATLAS [51], are general purpose detectors that were designed to study the nature of electroweak symmetry breaking and possible signatures of new physics up to the TeV scale. The LHCb [52] detector is dedicated to study rare decays of B hadrons and to perform precision measurements of CP violation. The ALICE [53] detector is designed for heavy-ion collisions to study the physics of strongly interacting matter at high energy density.

![Diagram of CERN accelerator complex](image)

Figure 2.1: A schematic representation of the CERN accelerator complex [54].

One of the key parameters that characterizes the collider’s performance is the machine instantaneous luminosity ($\mathcal{L}$), since the production rate for events of a given physical process is proportional to it, in particular at LHC:

$$\nu_{\text{event}}^{\text{LHC}} = \sigma_{\text{process}}^{\text{LHC}} \mathcal{L},$$

(2.1)

where $\nu_{\text{event}}^{\text{LHC}}$ and $\sigma_{\text{process}}^{\text{LHC}}$ are, respectively, the rate and the production cross section of a process at the LHC.

$\mathcal{L}$ is a function of the beam’s parameters. For beams composed of Gaussian distributed bunches with equal number of protons per bunch, it can be written as

$$\mathcal{L} = f_{\text{rev}} n_b N_p^2 \frac{\gamma F}{4\pi \varepsilon_{\text{norm}} \beta^*},$$

(2.2)

where $f_{\text{rev}}$ is the beam revolution frequency, $n_b$ the number of bunches per beam, $N_p$ the number of protons per bunch, $\varepsilon_{\text{norm}}$ the normalized transverse beam emittance, $\beta^*$ the beta function at the collision point, $\gamma$ the relativistic factor, and $F$ the geometric reduction factor due to the bunch crossing angle.
2.1. The Large Hadron Collider

The analysis described in these thesis is based on the LHC Run 1 proton–proton collision data collected by the CMS in 2012. In 2012, LHC delivered proton collisions at $\sqrt{s} = 8$ TeV with 50 ns bunch spacing, which corresponds 20 MHz collision rate. During this period, the maximal instantaneous luminosity registered at the CMS collision point was $7.67 \times 10^{33}$ cm$^{-2}$s$^{-1}$, as shown in Figure 2.2a, while the LHC design peak luminosity is $10^{34}$ cm$^{-2}$s$^{-1}$. The total integrated luminosity delivered by LHC in 2012 was 23.30 fb$^{-1}$, out of which 21.79 fb$^{-1}$ was recorded by the CMS, as in Figure 2.2b.

![CMS Peak Luminosity Per Day, pp, 2012, $\sqrt{s} = 8$ TeV](image1)

![CMS Integrated Luminosity, pp, 2012, $\sqrt{s} = 8$ TeV](image2)

Figure 2.2: LHC performance at $\sqrt{s} = 8$ TeV in 2012 in terms of (a) instantaneous luminosity delivered and (b) integrated over the year 2012. The blue curve corresponds to the delivered integrated luminosity while the yellow curve corresponds to the data recorded under stable beam conditions.

For the data recorded by CMS in 2012, the number of proton–proton interactions per bunch crossing is reported in Figure 2.3. Its average value is 21. The probability that more than one proton–proton interaction could produce an interesting process is usually considered to be negligible. Therefore, at the analysis level, only one most energetic proton–proton collision per event is selected and is referred to
as the primary hard interaction, while the other collisions in the event are called pileup interactions.

Figure 2.3: The mean number of inelastic proton–proton interactions per bunch crossing recorded by CMS in 2012 \[55\].

2.2 The Compact Muon Solenoid detector

The CMS detector is a multipurpose detector aiming to fully reconstruct proton–proton collision events. The main physical motivations for its design and construction were to investigate the consistency of the Standard Model at the TeV scale, in particular to search for a Higgs boson, and to search for possible beyond Standard Model signatures.

The high pileup operation conditions, as shown in Figure 2.3, implies an excessive amount of particles emerging from the interaction region. To maintain good physical performance in such conditions and reduce pileup effects, the CMS was designed so that all main sub-systems provide large geometric coverage and are composed of high granularity detectors. This allows obtaining good momentum resolution for all charged particles and good muon identification, in particular. An accurate electromagnetic energy resolution provides an efficient photon and lepton isolation, while a fine segmented hadronic calorimeter ensures good missing transverse energy and di-jet mass resolution. On the other hand, all detectors and front-end electronics, especially those that are close to the interaction region, are designed to be radiation-hard, so as to minimize the performance decrease over time under the large particle flux.

The layout of the CMS detector is shown in Figure 2.4. As with most high energy physics detectors, it has a cylindrical form with a central barrel enclosed by endcaps on each side. It is 28.7 m long and 15 m in diameter with a total weight of \(1.4 \times 10^7\) kg. The main distinguishing features of the CMS design are a silicon-based inner tracker, an electromagnetic calorimeter made of scintillating crystals, and a 3.8 T superconducting solenoid. An overview of these features,
2.2. The Compact Muon Solenoid detector

... together with the other important CMS sub-systems, is presented in the following subsections.

Figure 2.4: A cutaway view of the CMS detector [56].

Coordinate system and useful quantities

The CMS uses two coordinate systems. The origin of both systems is centred at the nominal collision point of the beams, which is located an the centre of the detector. The first coordinate system is Cartesian, with the left-handed orientation: the $x$-axis points radially towards the centre of the LHC, the $y$-axis points vertically upward, and the $z$-axis points along the beam direction.

The second coordinate system is $(r, \phi, \eta)$, which is more suitable for the cylindrical shape of the detector. It can be expressed through the $xyz$ coordinate system described above:

$$r = \sqrt{x^2 + y^2}, \quad \phi = \tan^{-1} \frac{y}{x},$$

$$\theta = \tan^{-1} \frac{r}{z}, \quad \eta = -\ln \tan \frac{\theta}{2},$$

(2.3)

where $\theta$ is the polar angle and $\eta$ is the pseudo-rapidity.

The pseudo-rapidity is the relativistic limit of the rapidity of a particle, $y_z \equiv \tanh^{-1}(p_z/E)$, which depends on the particle’s energy $E$ and longitudinal momentum $p_z$. The difference between the rapidities of two particles, $\Delta y_z$, is invariant under a Lorentz boost along the $z$-axis. The same is true for the difference between their pseudo-rapidities, $\Delta \eta$, in the relativistic limit, which is a good approximation...
to the experimental high energy physics case.

Since $\eta$ does not depend on either $E$ or $p_z$, but only on $\theta$, the use of the $\eta$ coordinate is more advantageous than using $z$ or $\theta$. The pseudo-rapidity is used to identify different detector regions according to their position with respect to the beamline and to the nominal beam collision point.

Since both $\Delta \phi$ and $\Delta \eta$ are invariant under a Lorentz boost along the $z$-axis, it is useful to introduce a separation in the $\eta \phi$ space, $\Delta R = \sqrt{\Delta \eta^2 + \Delta \phi^2}$. $\Delta R$ is used to define a $z$-boost invariant cone around a single particle or detector position, in order to study nearby detector activity.

A luminous region near the coordinates’ origin, where two proton beams intersect each other, is referred to as the beam spot. A characteristic width of this region is $16 \mu m$ in the $x$ direction and $100 \mu m$ in the $y$ direction, while its length in the $z$ direction is approximately $15 cm$. The beam spot is the origin for all proton–proton interaction products and its location plays an important role during the event reconstruction, as described in Chapter 3.

### 2.2.1 The inner tracking system

The inner tracking system is the part of the CMS detector closest to the beam spot. It is designed to provide precise measurements of charged particle trajectories and momenta, as well as the positions of the primary and secondary vertices, which is crucial for the subsequent physical object reconstruction.

At the design operation conditions, the average number of particles produced by proton–proton interactions in one bunch crossing is approximately 1000. To achieve good precision in such big flux environment, the tracking system is composed of high granularity silicon detectors. A strong, fairly homogeneous, magnetic field of 3.8 T, produced by the superconducting solenoid, described in Section 2.2.3 allows measuring the transverse momentum of charged particles by bending their trajectories. The chosen tracker design allows the detection, with a good efficiency and accuracy, of the trajectories of all charged particles with $1 < p_T < 100 \text{ GeV}$ originating from the beam spot region with $|\eta| < 2.5$.

The layout of the inner tracking system is shown in Figure 2.5. The innermost part of the tracker is composed of silicon pixel detectors, while the outer parts use silicon strip detectors. Each of these parts is divided into barrel and endcap regions.

The pixel tracker includes three cylindrical barrel layers and two endcap disks on each side [58]. The barrel layers are 53 cm long with average radii of 4.4, 7.3 and 10.2 cm. The endcap disks are located at $|z|$ equal 34.5 and 46.5 cm extending in radius from 6 to 15 cm.

Both barrel and endcap use 285 $\mu$ m thick silicon modules with pixel size $100 \times 150 \mu m^2$. A charged particle that traverses a module produces a signal on several of the nearby pixels, which are combined into a hit. The fine pixel granularity
allows obtaining a three-dimensional measurement of the hit position, which is a key component for precise vertex reconstruction. By calculating the weighted average through the charge collected within a hit and by taking into account corrections due to the Lorentz drift of the collected electrons, the resulting hit position resolution is approximately 10 µm in the transverse direction and 20–40 µm in the longitudinal direction, depending on η.

The pixel modules are arranged into half-ladders in the barrel part and into blades in the endcap. The total number of modules in a pixel barrel and endcap is 768 and 672, respectively, which represents an overall area of approximately 1 m² with about $6.6 \times 10^7$ readout channels.

The strip tracker is divided into four parts: Tracker Inner Barrel (TIB), Tracker Outer Barrel (TOB), Tracker Inner Disks (TID) and Tracker End Cap (TEC).

The inner parts of the strip tracker, TIB and TID, are located in the radial region 20–55 cm with $|z| < 124$ cm and are composed of 4 layers and $3 \times 2$ disks, respectively. The TIB and TID use 320 µm thick micro-strip silicon sensors with pitch size varying between layers from 80 to 120 µm for TIB and from 100 to 141 µm for TID. The resulting TIB spatial resolution is 23–35 µm in the transverse direction.

The outer strip tracker comprises 6 TOB layers, which occupy the region $55 < r < 116$ cm with $|z| < 110$ cm, and $9 \times 2$ TEC disks, which occupy the region $124 < |z| < 282$ cm. In the TOB and 3 outer TEC disks, 500 µm thick strip silicon sensors are used to improve the signal-to-noise ratio, while 4 inner TEC disks using 320 µm sensors, the same as the TIB and TID. The pitch size varies from 122 to 183 µm for the TOB and from 97 to 184 µm for the TEC. The resulting TOB spatial resolution is 35–53 µm in the transverse direction.
Chapter 2. The experimental framework

In the first two layers of TIB and TOB, and in the transition region between the different parts of the tracker, as shown in Figure 2.5, back-to-back modules with 0.1 rad ‘stereo’ rotation angle are used. Such a configuration allows simultaneously measuring the transverse and longitudinal hit positions. The additional longitudinal resolution for the ‘stereo’ modules is 230 \( \mu \text{m} \) for the TIB and 530 \( \mu \text{m} \) for the TOB.

The inner tracker covers the region \(|\eta| < 2.5\), providing up to 10 high resolution measurements for charged particles with \(|\eta| < 2.4\).

While the tracker provides key information for the physical object reconstruction, it also adds a significant amount of material budget. The total material budget added by the tracker is shown in Figure 2.6 and, depending on \( \eta \), it is equivalent to \( \approx 0.4 - 2 \) radiation length or \( \approx 0.14 - 0.56 \) nuclear interaction length.

2.2.2 Calorimeter system

The inner tracker is followed by a calorimeter system. The CMS uses four calorimeters to efficiently measure the energies of different particles over a large \( \eta \) range.

The electromagnetic calorimeter (ECAL) \[59\] is designed to measure the energy of photons and electrons. The main benchmark channel considered during the design was \( H \rightarrow \gamma \gamma \), which was considered as one of the ‘golden’ channels for the Higgs boson discovery. Therefore, in order to obtain a good resolution for the invariant mass of a gamma–gamma object, both the energy and angular resolutions of the calorimeter are important. A high angular resolution is also important to separate two closely located photons to suppress \( \pi^0 \rightarrow \gamma \gamma \) background.

The hadron calorimeter (HCAL) \[60\] is designed to measure the energies of particle jets that are mostly composed of hadrons. Precise measurements of the jet energies

Figure 2.6: Material budget of the tracker traversed by a particle produced at the nominal interaction point, as function of \( \eta \), in radiation length, \( X_0 \), units (a) and in nuclear interaction length, \( \lambda_I \) units (b). [57]
2.2. The Compact Muon Solenoid detector

and directions is a key component for determining the direction of the missing energy flow produced by neutrinos. HCAL provides independent measurements for hadrons in the event, improving the identification of muons, electrons and photons by the ECAL and the muon system.

Two forward calorimeters, the centauro and strange object research calorimeter (CASTOR) and the zero degree calorimeter (ZDC), have been designed to complement the CMS measurements in the very forward region [17]. CASTOR covers the baryon-free mid-rapidity region, providing important information for heavy ion collisions. In proton–proton collisions, CASTOR is used in the diffractive physics measurements. The ZDC is located between the two LHC beam pipes and complements the CASTOR measurements for heavy ion and diffractive proton–proton physics studies.

Electromagnetic calorimeter and preshower

The ECAL is a hermetic homogeneous calorimeter composed of lead tungstate (PbWO₄) crystals with an array of photodetectors placed after a preshower detector. It is separated into barrel (EB) and endcap (EE) sections, as in Figure 2.7.

Taking into account the requirements of high granularity and compact size, the use of lead tungstate as the material for the scintillating crystals is a logical choice, because of its short radiation length ($X_0 = 0.98\,\text{cm}$) and small Molière radius (2.2 cm). The other two important advantages of lead tungstate are its radiation hardness and its fast attenuation of scintillation light, which is about 80% for the LHC design bunch spacing interval of 25 ns.

However, it has a relatively low light output, of 4.5 photoelectrons per MeV at 18°C, with a strong temperature dependence of $-2.1\%/\degree\text{C}$. Therefore, to ensure good energy resolution, sensitive photodetectors and thorough temperature control are required. For that reason, silicon avalanche photodetectors (APD) and vacuum phototriodes (VPT) are used in the barrel and endcap sections, respectively. On
the other hand, a water flow based cooling system delivers a stable operating temperature of $18 \pm 0.05^\circ\text{C}$ for both the crystals and the APD, whose yield is also temperature dependent.

The EB covers the region $0 < |\eta| < 1.479$ and has an inner radius of 1.29 m. It is composed of $6.12 \times 10^4$ crystals, which are 23 cm long ($25.8\ X_0$) and $22 \times 22\ \text{mm}^2$ at the front face cross section.

The preshower detector is placed in front of the EE, allowing it to identify $\pi^0$ and improving the electron identification against minimum ionizing particles. The preshower is a sampling calorimeter with a layer of lead radiators followed by silicon strip sensors. It covers the region $1.653 < |\eta| < 2.6$ with a radiation length of about 3 $X_0$.

The EE covers the region $1.479 < |\eta| < 3.0$ and are located at $|z| = 3.154\ \text{m}$, however, due to the increase of pileup and radiation dose, fine precision measurements are limited to $|\eta| < 2.5$. It is composed of $7.324 \times 10^3$ crystals, which are 22 cm long ($24.7\ X_0$) and $28.62 \times 28.62\ \text{mm}^2$ at the front face cross section. The crystals are grouped in $5 \times 5$ mechanical units, called supercrystals.

The combined ECAL and preshower energy resolution $\sigma$, in the energy range $25 < E < 500\ \text{GeV}$, can be parametrized as a function of a stochastic term $a$, noise $\sigma_n$ and a constant term $c$:

$$
\left( \frac{\sigma_E}{E} \right)^2 = \left( \frac{\sigma_n}{E} \right)^2 + \left( \frac{a}{\sqrt{E}} \right)^2 + c^2.
$$

The noise term is composed of noise from the preamplifier, digitization, and pileup noises. The basic contributions to the stochastic term are the photo-statistics, fluctuation in the lateral containment, and fluctuations in the energy deposited by the preshower. The constant term is present due to calibration errors, geometrical effects, non-uniformity in the light collection, and energy leakage from the back of the crystal. Estimates of the contributions of these terms to the energy resolution based on the beam test results and simulations are shown in Table 2.1.

<table>
<thead>
<tr>
<th>Contributing term</th>
<th>Barrel ($\eta = 0$)</th>
<th>Endcap ($\eta = 2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise term $\sigma_n$, MeV</td>
<td>210</td>
<td>915</td>
</tr>
<tr>
<td>Stochastic term $a$, % $\cdot \sqrt{\text{GeV}}$</td>
<td>2.7</td>
<td>5.7</td>
</tr>
<tr>
<td>Constant term $c$, %</td>
<td>0.55</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Table 2.1: Total contributions from terms of different origins to the ECAL energy resolution at the design LHC luminosity for barrel and endcap ($5 \times 5$ crystal array) [59].

The hadron calorimeter

The HCAL is a sampling calorimeter situated after the ECAL, whose aim is to maximize the amount of energy deposited by the strongly interacting particles before they reach the magnet coil. It is divided into four sub-detectors: the
2.2. The Compact Muon Solenoid detector

Hadron Barrel (HB), the Hadron Endcap (HE), the Hadron Outer (HO), and the Hadron Forward (HF), as shown in Figure 2.8.

![Figure 2.8: Longitudinal view of one quadrant of the CMS showing the location of the HCAL sub-detectors [17].](image)

The HCAL detectors have a layered structure, with the absorbing layers alternating with the active material layers. The HB and HE absorber material is C26000/cartrige brass, a metal alloy made of copper (70%) and zinc (30%). Brass was chosen because of its relatively low nuclear interaction length of 16.42 cm and because it is not ferromagnetic. Plastic scintillator tiles are used as the HB and HE active material. The signal from the scintillators is extracted by wavelength shifting fibres (WLS), which deliver light to the optical decoder units, where it is read out by hybrid photodiodes.

The HB covers the region $|\eta| < 1.3$. In addition to the 14 layers of 50.5–56.5 mm thick brass absorbers, two steel plates in the front and in the back of the HB are used to ensure structural strength. The absorbing layers are interspaced by 17 layers of 3.7–9 mm scintillators. The HB is divided into segments, each with an $\eta\phi$-size of 0.087 × 0.087.

However, the HB do not provide sufficient hadron shower containment. The HO, which covers the same $\eta$ region as the HB, is used to catch the tails of the hadron showers. The HO is placed outside the solenoid coil, using it as an additional absorber with an interaction length of about $1.4\lambda_I/\sin\theta$. The HO allows increasing the minimal interaction length of the HCAL to $11.8\lambda_I$.

The HE covers the region $1.3 < |\eta| < 3$. The composition of its layers is similar to that of the HB, providing a total interaction length of $\approx 10\lambda_I$. The granularity of the HE in $\eta\phi$ units varies from 0.087 × 0.087 to 0.17 × 0.17, depending on $\eta$.

The forward part, HF, is located at $|z| = 11.2$ m and extends the HCAL coverage up to $|\eta| = 5$. It is designed to sustain very high particle fluxes, which in the region $4.5 < |\eta| < 5$ will result in a radiation dose that approaches 100 Mrad/year for the design LHC operating luminosity. The HF is composed of a 1.65 m long cylindrical steel absorber ($\approx 10\lambda_I$) and radiation hard quartz fibers made of fused-silica core and polymer hard-cladding. In the transverse direction, the HF forms
a ring around the beam line with an inner radius of 12.5 cm and an outer radius of 130 cm. The fibers run in the longitudinal direction and are organized to form towers with $0.175 \times 0.175 \eta \phi$ granularity.

Because the HF provides good precision for energy measurements in the forward region, it has been used for the CMS luminosity measurements, as described in Section 2.2.5.

The energy resolution of the central and endcap HCAL parts can be fit with a parametrization similar to the one used for ECAL and shown in Equation (2.4). In the case of HCAL, the contribution from the noise $\sigma_n$ is negligible. In the energy range 2–300 GeV, based on test beam results, the stochastic and constant terms are equal $(84.7 \pm 1.6)\% \cdot \sqrt{\text{GeV}}$ and $(7.4 \pm 0.8)\%$, respectively [61].

**Forward detectors**

The CASTOR is a Cerenkov-based sampling calorimeter, which is designed to provide a fast response and to measure shower developments in an environment exposed to a high level of radiation. It is installed at $|z| = 14.38$ m and covers the region $5.2 < |\eta| < 6.6$.

CASTOR is divided into electromagnetic and hadronic sections. Each section is composed of tungsten plate absorbers and fused silica quartz as an active material. The total radiation length of the electromagnetic section is $20.1 X_0$ and the total interaction length of the hadronic section is $10 \lambda_I$. Test beam results show good constant term resolution of $\lesssim 1\%$ for the electromagnetic part, with the stochastic term $\approx 50\% \cdot \sqrt{\text{GeV}}$ [17].

The ZDC is a quartz-tungsten sampling calorimeter with a design similar to that of the CASTOR. It is placed at $|z| \approx 140$ m, covering the region $|\eta| \geq 8.3$. As with the CASTOR, the ZDC is divided into an electromagnetic ($19 X_0$) and a hadronic part, with total interaction length $\approx 7.5 \lambda_I$. Test beam results measured the stochastic and constant terms of the electromagnetic (hadronic) energy resolution to be $70\% \cdot \sqrt{\text{GeV}}$ ($138\% \cdot \sqrt{\text{GeV}}$) and $8\%$ ($13\%$), respectively [17].

**2.2.3 Superconducting magnet**

Bending the trajectory of charged particles in a magnetic field allows measuring their transverse momentum, which is especially important for muons. The bending force and, therefore, the precision of the measurements of the momentum, is proportional to the magnetic field strength. One of the most distinguishing features of the CMS detector is its superconducting magnet, which provides a strong field of $3.8 \text{ T}$ [62].

The magnet is composed of a superconducting solenoid surrounded by an iron yoke. A cryogenic system maintains the magnet at an operating temperature of $\approx 4.5 \text{ K}$ using flows of liquid helium.
The solenoid consists of a winding with 4 layers made of the conductor NbTi. Its length, 12.5 m, provides a homogeneous field over the entire central region of the CMS, as shown in Figure 2.9. In addition, the cold bore diameter of 6.3 m permits placing the inner tracker system, the ECAL and most of the HCAL, inside the magnet, in order to not disturb the energy measurements.

![Figure 2.9: The predicted absolute value of the magnetic field (left) and field lines (right) in a longitudinal section][63]

To minimize the radial extent, the magnet has a thin coil design with 0.312 m radial thickness, so that $\Delta R/R \sim 0.1$. As a result, during the operation of the magnet, the coil has a very high energy density per unit mass (11.6 kJ/kg). To maintain such an energy density without damaging the coil, and to ensure a stable magnet performance, the cryogenic system should fulfill a series of strict requirements, including a maximum temperature margin of 1.8 K with respect to the operating temperature.

The magnetic field loop is closed by a very massive iron yoke of $10^7$ kg, which requires a well designed support structure. It is composed of 5 barrel wheels and 6 endcap disks. Each barrel wheel and endcap disk also hosts muon chambers, as described in Section 2.2.4. The return field is $\approx 2$ T, pointing in the direction opposite to the direction of the field inside the coil. This is used to improve the momentum resolution of the muons, whose trajectories are bent in opposite directions in the tracking region and in the region of the muon system.

### 2.2.4 The muon system

Muons in the energy range from 1 GeV to 1 TeV have only small radiative losses or ionization energy losses. That allows them to pass through the entire detector volume, while most of the other interaction products are stopped by the ECAL and HCAL. For this reason, the muon system is the outermost part of the CMS detector.
Because the contribution from the background is significantly reduced by the calorimeters, highly efficient standalone muon identification by the muon system is possible, which is essential for the triggering purposes, as discussed in Section 2.2.6. On the other hand, the muon system and the inner tracker are the only sources of the muon momentum measurements. Therefore, to achieve good physical performance, the muon system is designed to deliver a fine transverse momentum resolution.

The muon system is divided into barrel and endcap sections, as shown in Figure 2.10.

The barrel section covers the region $|\eta| < 1.2$. In the barrel, standard drift tube (DT) chambers are used, because of the low muon and neutron-induced background rates and the fairly uniform magnetic field, which is mostly contained in the iron yokes, in this region. The DT chambers are organized into 4 stations. The first three stations are hosted in between the magnet flux return plates, with each station composed of 8 chambers. Four of the chambers provide measurements in the $r\phi$-plane, while the other four measure a coordinate in the $rz$-plane. The fourth station is placed outside the yoke. It provides only $r\phi$-coordinate measurements and is composed of 8 chambers, which are separated as great an extent as possible, so as to achieve the best angular resolution.

The two endcap sections allow identifying muons in the region $0.9 < |\eta| < 2.4$. In the endcap, where the muon rates are higher than in the barrel, cathode strip chambers (CSC) are used to ensure good performance in the presence of a high background level and the condition of a nonuniform magnetic field. The advantages of the CSC are its fine segmentation, fast response time, and radiation hardness. Each endcap section consists of four CSC stations interlayered between...
the magnet flux return plates. In each CSC station, 6 layers of approximately perpendicular anode wires and cathode strips provide efficient muon tracking and robust pattern recognition for background rejection. Both the cathodes and anodes are readout delivering $r\phi$ and $\eta$ measurements, respectively. The anodes also provide good temporal resolution, which is used to identify the beam-crossing time of a muon.

In addition to the DC and CSC, resistive plate chambers (RPC) are used in the region $|\eta| < 1.6$. The RPC are gaseous parallel-plate detectors with a high intrinsic temporal resolution of $\approx 2$ ns made of double-gap modules operating in avalanche mode. Six RPC stations in the barrel region and three RPC stations in the endcap region provide additional $p_T$ trigger capabilities, due to their fast RPC response time, and provide extra information for tracking ambiguity resolution.

The spatial and temporal resolutions of the muon system were measured using $\sqrt{s} = 7$ TeV p–p collisions data collected by the CMS in 2010 [65]. The average DT chamber resolution is $77 - 123 \mu m$ in the $r\phi$ direction and $133 - 393 \mu m$ in the $rz$ direction. The spatial resolutions of the CSC and RPC chambers vary from 58 to 136 $\mu m$ and from 8.1 to 13.2 mm, respectively, depending on the chamber type.

The overall RMS of the DT temporal resolution measured for a high $p_T$ muons is $\lesssim 2.6$ ns. The signals from the CSC were sampled at 50 ns intervals, allowing obtaining an offline single hit resolution of 5 ns. The overall RPC temporal resolution, which also includes the additional time uncertainties coming from the electronics and cable lengths, is better than 3 ns.

### 2.2.5 Luminosity measurement

The luminosity measurements provide an overall normalization for the yields of all physical processes. The precision of the luminosity measurements is essential to reduce the systematic uncertainties in the expected number of events at the analysis level.

In the CMS, the luminosity is measured on a bunch-by-bunch basis, which allows real-time monitoring of the performance of the LHC. For the online luminosity measurements, the CMS uses the HF detector, the forward part of the HCAL (Section 2.2.2). In addition to online methods, an offline algorithm based on pixel cluster counting is available for more precise luminosity estimates. [66, 67]

To be able to collect luminosity bunch-by-bunch, the HF is equipped with an independent high rate data acquisition system. CMS used two online HF-based algorithms. Both algorithms allow obtaining the average luminosity with a statistical accuracy of 1% within a time interval $\lesssim 1$ s. The first method uses the ‘zero counting’ technique, estimating the mean number of interactions per bunch crossing from the average fraction of empty HF towers. The second method is based on the linear relation between the mean luminosity and the average transverse...
energy. For the collected data, these algorithms provide results that are within 5% of each other.

The accuracy of HF-based methods is limited by calibration drift over a long period of time and the non-linearity of the pileup dependence. The offline algorithm used for precision luminosity measurements is based on the measurements provided by the pixel tracker silicon detectors (Section 2.2.1). Since the average number of pixel clusters per event, $\langle n \rangle$, is proportional to the number of $p$–$p$ collisions in the 2012 LHC run, the instantaneous luminosity can be estimated by

$$ L = \frac{f_{\text{rev}} \langle n \rangle}{\sigma_{\text{vis}}}, $$

(2.5)

where $f_{\text{rev}}$ is the beam revolution frequency and $\sigma_{\text{vis}}$ is the visible cross section, which is calibrated using the Van der Meer scan technique [67].

The total integrated luminosity, calculated by the offline method, recorded by the CMS in 2012, is $21.79 \, \text{fb}^{-1}$, as shown in Figure 2.2b. The overall uncertainty in the integrated luminosity is estimated to be 2.5% (syst.) $\pm$ 0.5% (stat.) [67].

### 2.2.6 Triggers and data processing

During Run 1, the LHC delivered proton beams with 50 ns bunch spacing, which corresponds to a bunch crossing frequency of 20 MHz. Technical limitations do not allow storing the information for all the bunch crossing events, imposing a reduction in the stored event rate. Since the majority of collisions are not of interest, the reduction is done by selecting events that have a higher probability of containing $p$–$p$ interactions that resulted in processes of physical interest. This led to the implementation of a trigger system, which makes a decision to accept or to reject an event based on the detector response.

The trigger decision should be taken in a very limited amount of time, to ensure a quasi continuous data taking over long time intervals. This imposes serious restrictions on the complexity of the trigger algorithms. On the other hand, the trigger algorithms should be very selective, in order to reduce the data flow from $2 \times 10^7$ to $\mathcal{O}(10^3)$ events per second, while keeping a good efficiency in selecting physically interesting events.

To achieve an optimal performance, the CMS uses two level of triggers: Level-1 (L1) triggers [68] and High-Level triggers (HLT) [69].

The L1 trigger is implemented in custom electronics and allows reducing the event rate from 20 MHz to $\lesssim 100$ kHz. The design latency for the L1 trigger decision, which is $3.2 \, \mu\text{m}$, makes it impossible to use the entire detector output at the Level-1. At this stage, only information from the calorimeters and the muon system is available.

The architecture of the L1 trigger, as shown in Figure 2.11, is composed of local, regional and global components. During the local step, trigger primitives for
each sub-detector are generated. The calorimeter trigger primitive combines the energies deposited in the ECAL crystals or HCAL towers to obtain the trigger tower $E_T$. The muon trigger primitives are composed of track segments, for the CSC and φ-projection in the DT, or hit patterns, for the RPC and $\eta$-projection in the DT. The local information is combined by Regional Triggers, where the calorimeter electron/photon candidates are defined and combined DT/CSC tracks are assigned physical parameters. In the next step, the Global Calorimeter Trigger (GCT) and Global Muon Trigger (GMT) determine the L1 trigger objects based on the combined sub-detector information. The GCT trigger objects include jets and $\tau$-jets, the total and missing transverse energies, and the isolated and non-isolated $e/\gamma$ candidates.

![Figure 2.11: Architecture of the Level-1 Trigger at the CMS](image)

Then, at the last step, the L1 Global Trigger (GT) decides to accept or to reject an event based on the GCT and GMT trigger objects. The decision to accept is taken if the event satisfies all the criteria of at least one of the GT algorithms. The GT may execute in parallel up to 128 algorithms, including both the basic algorithms, which consist of some simple $p_T$ or $E_T$ thresholds to a single object, and the complex algorithms based on topological selections. The result of each algorithm is represented by one bit, which indicates whether an event passed the algorithm’s requirements or not. If L1 decides to accept, the entire detector information is read out and passed to the HLT.

The HLT is implemented in software\(^1\) and is executed in the Event Filter Farm (EVF), composed of common PCs. The HLT event selection algorithms are similar to those of the offline reconstruction, with some simplifications made so as to reduce the execution time. During 2012, the average HLT output rate was 300–350 Hz of events reconstructed by the prompt HLT algorithms. To extend the physics program, the output was enhanced by an extra 300–350 Hz rate of events, which did not undergo a complete prompt reconstruction immediately, but were reconstructed only after the end of the run\(^2\).

---

\(^1\)The CMS software (CMSSW) used for the HLT and offline reconstruction is open-source and its source code is available online\(^3\).

\(^2\)During 2012, the average HLT output rate was 300–350 Hz of events reconstructed by the prompt HLT algorithms. To extend the physics program, the output was enhanced by an extra 300–350 Hz rate of events, which did not undergo a complete prompt reconstruction immediately, but were reconstructed only after the end of the run.

\(^3\)The CMS software (CMSSW) used for the HLT and offline reconstruction is open-source and its source code is available online.
The HLT architecture is designed to reduce the average CPU time needed to process an event, by rejecting, in the earliest possible stages, events that certainly will not pass the acceptance criteria. For that reason, the sequence of HLT algorithms is run in order of increasing complexity, starting from the least CPU consuming steps.

The HLT execution sequence is organized in HLT paths, which combine physical objects reconstruction and selection steps that are run in a predefined order. An HLT path corresponds to a set of requirements that an event has to fulfill at Level 1 and at HLT. The CMS trigger system implements a great number of HLT trigger paths, as described in [69].

Since information from all sub-detectors is available at the HTL level, by combining all the measurements of the various subsystems it is possible to efficiently record, at the same time, events characterized by different signatures. Indeed, the data collected by the CMS can be used to study the properties of the Higgs boson, the top quark, weak bosons, as well as the decays of $b$ and $c$ hadrons, tau leptons, and more generally SM processes, or to search for evidence of physics beyond the SM. The trigger path used in this analysis requires a central isolated electron or muon candidate and a hadronic $\tau$ candidate, and will be briefly described in Section 4.2.

To optimize the execution time, the HLT does not execute all the paths for every event. Which HLT paths should be executed is determined by the L1 results. One HLT path may be seeded by several L1 algorithms. The HLT execution starts with a local calorimeter and muon system reconstruction and selection, using L1 objects as seeds. This allows of reducing the event rate before the tracking and full reconstructions steps, which are CPU-consuming.

If an event is accepted by the HLT, it is transferred for permanent storage to the CMS Tier-0 computing centre. While the size of one event is about 0.5–1 megabyte, the total size of the CMS data, including measurements and simulations, is $\mathcal{O}(10)$ petabytes. Such a huge amount of information is managed by the Worldwide LHC Computing Grid (WLCG), a global computing infrastructure that provides the computing resources to store, distribute, and analyze the data generated by the LHC [73,74].
Chapter 3

Event reconstruction and physical objects

In this chapter we describe how the information from the different sub-detectors is used to reconstruct the outgoing particles from the p–p interactions. In the first step, the collected raw information is used to reconstruct the high level objects, including tracks, vertices, and stable particles. Then the high level objects are used in the reconstruction of the physical objects, including the object of interest for the final state investigated in this thesis: electrons, muons, jets, b-jets, missing transverse energy for neutrinos, and taus.

3.1 High level reconstruction objects

3.1.1 Track reconstruction

Track reconstruction is a sophisticated algorithm to estimate the trajectory and the momentum of the charged particles passing through the inner tracking system (Section 2.2.1).

Inside the tracker region, the magnetic field induced by the superconducting magnet (Section 2.2.3) is sufficiently uniform and the amount of matter introduced by the tracker is small enough to allow fitting a charged particle trajectory by an helix. In the CMS a track trajectory is parametrized by five parameters: $d_0$, $z_0$, $\phi$, cot $\theta$, and the transverse momentum $p_T$ of the track. $p_T$ is defined at the impact point $(x_0, y_0, z_0)$, the point of the track trajectory closest to the nominal beam axis ($z = 0$). $\theta$ and $\phi$ are, respectively, the polar and the azimuthal angles of the track momentum vector at the impact point. $d_0$ is the impact parameter of the track in the transverse plane and is defined through the coordinate of the impact point, $d_0 = -y_0 \cos \phi + x_0 \sin \phi$.

To estimate these parameters, the spatial positions of the charge deposits (hits) produced by a charged particle when it passes through the layers of silicon de-
tectors are used. Due to the high particle multiplicity inside the tracker, the track reconstruction is a computationally challenging procedure. To reduce the combinatorial complexity, an iterative tracking process is used. This allows excluding from further iterations those hits that were assigned to tracks in previous iterations.

Six iteration steps are defined. The first three iterations (0–2) are designed to reconstruct the prompt tracks: the tracks that originate near the beam spot. The last three iterations (3–5) reconstruct the remaining tracks, those with origins that lie outside the beam spot [57].

To reconstruct the tracks within each iteration step, the combinatorial track finder algorithm (CTF) [75] is used with configuration parameters that depend on the iteration step. The CFT is composed of four logical steps [57]:

- **Seed generation**, which provides initial track candidates based on two or three hits in the pixel tracker.

- **Trajectory building** by extrapolating the seeds to all tracker layers and updating the track parameter estimates by taking into account new matched hits using a Kalman filter [76].

- **Final track fitting and smoothing** using a Kalman filter and a Runge–Kutta propagator [77] to obtain the optimal track parameters by taking into account material effects and accommodating possible inhomogeneities in the magnetic field.

- **Track selection** by applying quality requirements to reject fake tracks, which are not associated with a charged particle.

The physical performance of the track reconstruction is defined by its efficiency, fake rate, and resolution of the track parameters. Simulations for $t\bar{t}$ events in a pileup environment show that for the high-purity quality tracks with $0.7 < p_T < 100\text{ GeV}$, the efficiency is above 85%, while the fake rate is below 10%, as shown in Figures 3.1a and 3.1b. For the tracks with $p_T > 2\text{ GeV}$, the resolution of the transverse ($d_0$) and longitudinal ($z_0$) impact parameters is better than 100 and 300 $\mu$m, respectively, as shown in Figures 3.1c and 3.1d.

Most of the inefficiency in the general tracks is due to the tracks produced by pions and kaons, which undergo elastic and inelastic nuclear reactions, and electrons, which lose energy through bremsstrahlung. On the other hand, the tracking efficiency for the isolated muons is much higher and the efficiency measured for muons originating from $Z \rightarrow \mu\mu$ events is above 99%. [57]

### 3.1.2 Primary vertex

A precise reconstruction of the positions of all p–p interactions in the event (vertices) allows significantly improving the efficiency of the separation between the
primary hard interaction products and the objects originated in the pileup collisions. On the other hand, the accurate location of the primary vertex, in combination with other information, e.g. the secondary vertex position, plays an important role in the determination of the properties of the physical objects.

During LHC Run 1 in 2012, the average number of p–p interactions per event was 21, as shown in Figure 2.3. The position of each interaction vertex in the event is reconstructed in three steps [57,78]. Firstly, tracks are preselected to be consistent with the primary interaction region. Then, using the deterministic annealing algorithm [79], the selected tracks are split into several groups based on a possible common vertex of origin. In the last step, the position of each vertex is obtained by fitting the associated tracks with an adaptive vertex fitter [80].

Figure 3.1: Track reconstruction efficiency (a), fake rate (b) and resolutions of transverse (c) and longitudinal (d) impact parameters as a function of $p_T$ for charged particles from simulated $t\overline{t}$ events. The number of pileup interactions is modelled by a Poisson distribution with mean value of 8. The efficiency and fake rate plots are produced for tracks with $|\eta| < 2.5$ that pass the high-purity quality requirements. Resolution distributions are with pileup and solid (open) symbols correspond to the half-width of the 68% (90%) central confidence intervals. [57]
Chapter 3. Event reconstruction and physical objects

The number of degrees of freedom of the vertex ($n_{dof}$) indicates the success of the final vertex fit. It is defined by

$$n_{dof} = -3 + 2 \sum_{\text{tracks}} w_i,$$  \hspace{1cm} (3.1)

where $w_i \in [0; 1]$ is a weight associated to each track, representing the likelihood that the track actually belongs to the vertex.

At the analysis level, the primary hard interaction vertex is selected as the vertex with the highest sum of $p_T^2$ of the associated tracks that satisfies the following quality requirements: the absolute value of its longitudinal position ($z$) is less than 24 cm, its transverse distance from the $z$-axis, $d = \sqrt{x^2 + y^2}$, is less than 2 cm, and $n_{dof} > 4$.

For the vertices with at least 5 associated tracks, the reconstruction efficiency is measured to be close to 100%, while the transverse and longitudinal position resolutions are less than 110 and 150 $\mu$m, respectively [57].

3.1.3 Particle-flow

The particle-flow (PF) algorithm [81] [82] reconstructs all stable particles in the event by combining information from all sub-detectors. The particles reconstructed by the particle-flow algorithm (PF particles) are used to reconstruct the jets and the hadronic $\tau$ decays, as well as to calculate the isolation of the physical objects (see Sections 3.2.1, 3.2.2 and 3.2.4 for the definition of ‘isolation’). The PF electrons and PF muons are some of the main ingredients for the analysis-level reconstruction of the electrons and muons.

Since one particle usually produces signals in more than one sub-detector, the core of the PF is a link algorithm and the subsequent particle reconstruction and identification is based on the linked blocks. The PF link algorithm combines the charged particle tracks reconstructed by the inner tracker, the calorimeter clusters, and the muon tracks. To obtain an optimal combined precision, the standalone reconstruction of each linked element should deliver a good accuracy.

The iterative track reconstruction in the inner tracker is covered in Section 3.1.1. The standalone muon reconstruction in the muon system starts by building track segments in the DT and CSC using pattern recognition and linear fit techniques [83]. Next, the track segments and RPC clusters are combined and refitted by a Kalman filter forming standalone muon tracks [17].

Calorimeters with fine granularity and an advanced calorimeter clustering algorithm are essential for the PF reconstruction. The information provided by the calorimeters allows measuring the energy and direction of the photons and stable neutral hadrons, identifying the electrons and accompanying bremsstrahlung photons, and improving the energy resolution of the charged particles that have low quality or high $p_T$ tracks.
3.1. High level reconstruction objects

The clustering algorithm is run on each part of the PS, ECAL and HCAL, separately, except the HF. As a consequence, each HF cell is considered as a separate cluster. In the first step of the algorithm, the cluster seeds are identified as those cells with a local maximum of deposited energy. Then ‘topological’ clusters are defined by recursively adding to the seeds those adjacent cells that have energy deposits above a certain threshold. For both ECAL and HCAL, the threshold value was set as two standard deviations above the electronics noise level, which corresponds to 80–300 MeV and 800 MeV, respectively. After collecting all adjacent cells, one topological cluster may contain more than one seed. Therefore, in the last step of the algorithm, the topological clusters are split by sharing the energy deposit of each cell between the PF clusters, which amount is equal to the number of seeds. The fine granularity of the calorimeters granularity is the key component that allows efficiently splitting the topological clusters into PF clusters.

**PF link algorithm**

Charged particle tracks, calorimeter clusters, and muon tracks are linked together into PF blocks. Each block links a group of elements from two sub-detectors.

Each charged particle track is extrapolated to the PS, ECAL and HCAL and linked with all calorimeter clusters that contain an extrapolated track position within their boundaries. To incorporate reconstruction uncertainties, gaps between calorimeter cells and cracks between calorimeter modules, the cluster envelope of a track can be enlarged by up to the size of one cell. Additional links between charged particle tracks and calorimeter clusters are established to take into account bremsstrahlung by electrons. In this case, tangents to the tracks are extrapolated to the ECAL from the intersection points between the track and each of the tracker layers. If the extrapolated tangent position is within the boundaries of a cluster, as defined above, then the cluster is linked to the track as a potential bremsstrahlung photon.

A link between the ECAL and HCAL clusters is created if the position of the ECAL cluster, which have higher granularity, is within the HCAL cluster envelope. A similar logic is followed to link the PS and ECAL clusters.

A link between inner tracker tracks and muon system tracks is referred to as a global muon. It is established if the $\chi^2$ of the global fit of the two tracks is less than $\chi^2_{\text{max}}$. If more than one combination of tracks satisfy this criterion, the combination with the smallest $\chi^2$ is chosen.

**PF particle reconstruction and identification**

At the final step of the particle-flow algorithm, the PF particles are reconstructed and identified starting from the blocks of linked PF elements. PF particles are created in several iterations. Blocks assigned to a PF particle are excluded in the subsequent iterations.
In the first iteration, the PF muons are created from the global muons, requiring that the combined global muon momentum is within three sigma of the momentum estimated using only the inner tracker.

In the second iteration, the PF electrons are reconstructed and identified. The algorithm starts with a pre-identification step in which electron energy losses by bremsstrahlung are considered. It continues by refitting the pre-identified electron trajectories to a Gaussian-Sum Filter (GSF) \cite{84} and final selection criteria that improves the electron identification. The reconstruction and identification of electrons is discussed in more detail in the following Section 3.2.1.

In the following steps, the calibrated energy from the calorimeter clusters is matched with the charged track energies in the block. All possible combinatoric combinations within the blocks are analysed (e.g. one cluster that is associated with several tracks or one track associated to several clusters), considering possible physical cases that may correspond to the observed block composition. At this stage, tighter quality criteria are applied to the remaining tracks, which allows significantly reducing fake rate, while keeping high the efficiency of the hadronic jets reconstruction.

If for a given block, the combined calorimeter cluster energy is more than three standard deviations less than the sum of the momenta of the associated tracks, then by exploiting redundancies from the sub-detector measurements, a recurrent search with relaxed matching criteria may give rise to additional PF muons without increasing the fake rate. Each track remaining in the block gives rise to a PF charged hadron with 4-momentum determined from the track, assuming the charged pion mass hypothesis. In case the calorimetric energy is compatible, within the uncertainties, with the track’s momentum, the PF charged hadron energy is refitted using both the calorimeter and tracker measurements, which allows improving the energy resolution for high-$p_T$ and large-$\eta$ particles.

In the inverse case, when the combined calorimeter cluster energy is significantly larger than the sum of the momenta of the associated tracks, in addition to the PF charged hadrons, first a PF photon is created with an energy that correspond to the total energy excess in the ECAL and then if the remaining energy excess is larger than the calorimeter energy resolution, a PF neutral hadron is also created. The priority in reconstruction of PF photons over PF neutral hadrons is justified by the fact that in jets, the fraction of the energy carried by photons is significantly larger than the fraction of the energy carried by neutral hadrons.

All the remaining ECAL and HCAL clusters produce PF photons and PF neutral hadrons, respectively.
3.2 Physical objects

3.2.1 Electrons

Reconstruction

A distinguishing feature of the electron reconstruction is the significant amount of energy that an electron can emit via bremsstrahlung before reaching the ECAL. Depending on the amount of the tracker material traversed, an electron radiates, on average, from 33% of its energy in the central barrel region ($|\eta| \approx 0$) up to 86% in the $|\eta| \approx 1.4$ region. To achieve an optimal performance that takes into account this effect, the electron reconstruction and identification are based on a combination of standalone and PF algorithms \[85\].

In the standalone approach, two ECAL clustering techniques are used: a hybrid algorithm and a multi-$5 \times 5$ algorithm. The first is used in the ECAL barrel while the second one is used in the ECAL endcaps. Both are applied, taking into account the fact that radiated photons are mostly spread in the $\phi$-direction, while their spread in the $\eta$-direction is very small.

**Hybrid algorithm.** In this case, clusters are formed starting from crystal seeds by adding arrays of $5 \times 1 \eta \phi$ crystals in both $\phi$-directions. The energy threshold below which crystals are not clustered is 100 MeV. Then the reconstructed clusters belonging to the same region are grouped into a Super Cluster (SC). The idea of the hybrid algorithm is illustrated in Figure 3.2.

![Figure 3.2: Illustration of the hybrid clustering algorithm used in the ECAL barrel for the electron reconstruction \[86\].](image)

**Multi-$5 \times 5$ algorithm.** This algorithm is used for the ECAL endcaps, where crystals are not arranged in the $\eta \phi$-directions. In this case, clusters are formed starting from seeds by adding to them $5 \times 5$ neighboring crystal arrays with possible overlaps and, if the total energy of the nearby clusters are above a threshold, the clusters are grouped into a SC.
The energy-weighted average position of the supercluster is computed, which is the equivalent of the impact point in the ECAL of a non-radiating electron with energy equal to the supercluster energy.

The standard track reconstruction procedure (Section 3.1.1) is not optimal for electrons. Due to the high radiative losses, a fit of the electron trajectory with a helix using a Kalman filter will result in hit losses or in poor estimation of the track parameters. To avoid such inefficiencies, in the standalone approach, a dedicated electron track reconstruction algorithm is used. It starts from the extrapolation of the electron trajectory from the SC position to the interaction vertex. Each SC defines a $\phi z$-window, which will be searched for compatible track seeds in the pixel tracker. The compatible track seeds are propagated in the tracker using a Kalman filter with energy losses modelled by a Bethe-Heitler function. To optimize the estimation of the electron track parameters, final tracks are refitted by a Gaussian-Sum Filter (GSF) \[84\], with the energy losses in each layer modelled as a combination of Gaussian functions, building tracks up to ECAL $\text{ECAL}^1$. Then the GSF tracks are associated with ECAL clusters, requiring a $\Delta \eta$ and $\Delta \phi$ matching.

The PF approach (Section 3.1.3) is based on the standard particle-flow calorimeter clustering and allows sharing a single crystal energy among two clusters. In this approach, tracks reconstructed by the standard charged-track reconstruction algorithm (Section 3.1.1) are then refitted with the GSF. The identification as a PF electron of the PF block of ECAL clusters and GSF tracks is based on multivariate (MVA) technique \[87\] that combines information of track and cluster observables into a single variable. MVA identification uses track momentum, Kalman filter fit quality ($\chi^2_{KF}$), GSF fit quality ($\chi^2_{GSF}$), and the ECAL cluster’s shape and pattern.

The reconstruction efficiency of the combined standalone and PF approaches is $>90\%$ for electrons from $Z/\gamma^*$ data samples and simulated events, with the transverse energy of the SC ($E_{TSC}^\text{SC} > 10\text{ GeV}$) decay \[85\] shown in Figure 3.3.

Identification

The main sources of misidentification of the prompt electrons are jets and electrons coming from photon conversion or semileptonic b-quark and c-quark decays. In order to separate good electrons from jets and to increase the identification sensitivity, the identification of the electron is based on a multivariate Boosted Decision Three (BDT) discriminator \[87\].

The BDT discriminator \[85\] uses a set of the most discriminating variables which exploits the track-cluster matching, the associated SC substructure and its shape, the kinematic observables, and the fraction of energy loss through bremsstrahlung ($f(brem)$). The discriminator was trained in two bins of $p_T$ and three bins of $\eta$, for genuine electrons from Z decays into $e^+e^-$, using data and simulated events.

\[1\]The fraction of the energy lost through bremsstrahlung ($f(brem)$) is estimated using the momentum at the point of closest approach to the beam spot ($p_{in}$) and the momentum extrapolated to the ECAL surface ($p_{out}$) and is defined as $f_{brem} = (p_{in} - p_{out})/p_{in}$. 
3.2. Physical objects

Figure 3.3: Electron reconstruction efficiency measured in di-electron events in data and DY simulations, as a function of the electron $E_{T}^{SC}$ for $|\eta| < 0.8$ (a) and $1.57 < |\eta| < 2$ (b) \[85\]. The bottom panels show the corresponding data-to-simulation scale factors.

and for misidentified electrons reconstructed in W+jets events in the data.

The BDT output is a real number in the interval $[-1; 1]$ defined in such a way that the genuine electrons have a tendency to obtain higher output values, while the non-genuine electrons obtain lower output values with a bigger probability. The comparison of the BDT performance for electrons from $Z$ decays and for misidentified electrons is shown in Figure 3.4. Two working points (WP) that have been used in this thesis for the BDT-based electron identification are summarized in Table 3.1. Each WP defines the acceptance threshold for the BDT output as a function of the $p_T$ and $\eta$ of the electron.

An alternative to the BDT-based electron identification, the cut-based electron identification, is used at the HLT level. It has lower efficiency, but provides more robust results and can also be used in the veto definitions.

| Working point | Electron $p_T$ | $|\eta| \leq 0.8$ | $0.8 < |\eta| \leq 1.479$ | $|\eta| > 1.479$ |
|---------------|---------------|-----------------|-----------------|-----------------|
| Loose ID      | $p_T \leq 20$ GeV | 0.925           | 0.915           | 0.965           |
|               | $p_T > 20$ GeV  | 0.905           | 0.955           | 0.975           |
| Tight ID      | $p_T > 20$ GeV  | 0.925           | 0.975           | 0.985           |

Table 3.1: Definition of loose and tight working points for the BDT electron identification as a function of electron $p_T$ and $\eta$. To pass the WP identification criterion, BDT output should be greater than the threshold value.

Electron isolation

In this thesis, we are interested in electrons from $\tau$ decays: these electrons tend to have a small activity of particles around the lepton direction in the tracker and in the calorimeter. In order to reduce significantly the contamination from electrons.
originating from b- or c-quark decays within jets or decays in flight, the selected electrons are required to be isolated.

The electron isolation ($I_{PF}$) is calculated considering all PF particles, excluding the electron candidate, within a cone of size $\Delta R = 0.4$ around the direction of the lepton, excluding an innermost region, the veto cone, where tracks coming from photon conversions may spoil the isolation.

The PF isolation ($I_{PF}$) is a function of the sums of the $p_T$ of the charged PF candidates ($\sum p_T^{charged}$), PF neutral hadrons ($\sum p_T^{neutral-had}$), PF photons ($\sum p_T^\gamma$) and pileup corrections ($p_T^{PU}$) \cite{85}:

$$I_{PF} = \sum p_T^{charged} + \max [0, \sum p_T^{neutral-had} + \sum p_T^\gamma - p_T^{PU}] .$$

The veto cones $\Delta R > 0.01$ are applied to the photons and the charged PF particles in the EB region. For the charged PF particles in the EE region, the veto cone $\Delta R > 0.015$ is applied.

The isolation value is very sensitive to the pileup effects. Therefore, all charged PF particles are required to originate from the primary hard interaction vertex, requiring $\Delta z < 2$ mm. The contribution from the neutral PF particles is corrected to take into account the presence of pileup by using the FastJet technique \cite{88}.

In Figure 3.5, it is shown that the PF isolation relative to the electron $p_T$ ($I_{PF}/p_T$) has a good discriminating power against misidentified electrons.

To further improve the identification performance, the electrons that have a high probability of originating from photon conversions are rejected. This is achieved by requiring that the electron trajectory should not have any missing hits in the inner tracker layers and by applying topological constraints on $e^+e^-$ pairs in the event, by fitting their tracks to a common vertex and considering the quality of...
3.2. Physical objects

Figure 3.5: PF electron isolation relative to the electron’s $p_T$ with pileup corrections applied, for electrons from $Z \rightarrow ee$ data and simulated events, and for misidentified electrons reconstructed in $W+\text{jets}$ events in data, in the barrel (a) and endcap (b) [85].

the fit ($\chi^2$). The secondary electrons are also rejected by applying cuts on the $d_0$ and $z_0$ of the electron track.

3.2.2 Muons

Muons have a mean lifetime of $\approx 2.2 \, \mu s$ and, in most cases, only a small amount of energy loss from the interaction with the tracker material, over a wide range of $p_T$. This allows the muons to pass through all the detector creating tracks in both the inner tracker and the muon system, and depositing a minimal amount of energy in the calorimeters.

Two offline muon reconstruction algorithms are available in the CMS [89], the Global Muon and the Tracker Muon:

- **Global Muon**: the reconstruction starts from the standalone muon tracks (Section 3.1.3) and matches them with the inner tracks. The final Global Muon tracks are refitted with a Kalman filter.

- **Tracker Muon**: the reconstruction starts from the inner tracks, considering all tracks with $p_T > 0.5 \, \text{GeV}$ and $|\vec{p}| > 2.5 \, \text{GeV}$, and extrapolates them to the muon system. If an extrapolated track position lies within the uncertainties of at least one muon segment, the corresponding track is considered as a Tracker Muon track.

The Global Muon reconstruction provides better momentum resolution for the tracks with $p_T > 200 \, \text{GeV}$, while the Tracker Muon reconstruction is more efficient for the low momentum tracks with $p_T \lesssim 5 \, \text{GeV}$. 
Chapter 3. Event reconstruction and physical objects

In this thesis, two muon identification techniques are used: the PF muon selection and the tight muon selection. Each selection combines both reconstruction algorithms, providing different combinations of muon identification efficiency and misidentification probability.

- **PF muon** The selection described in Section 3.1.3. It combines the Global Muon and Tracker Muon reconstructions, maintaining a very high reconstruction and identification efficiency \( \varepsilon_{\text{rec+id}} \) for the prompt muons. \( \varepsilon_{\text{rec+id}} \) was measured by the tag-and-probe technique in \( Z \rightarrow \mu\mu \) events: it is above 97%, as shown in Figures 3.6a and 3.6b for muons with \( 20 < p_T < 100 \) GeV. Furthermore, the PF identification has a small probability of misidentifying a charged hadron as a muon, and it is highly efficient in identifying muons within jets. The probability of misidentifying a pion or a kaon with \( |\vec{p}| > 15 \) GeV as a muon is below 0.3% and 0.4%, respectively [89].

- **Tight muon** This selection applies stringent requirements for the tracks and matching qualities, which allows rejecting a significant part of the secondary muons that originated in decays within the detector. It is achieved by requiring that the muon be reconstructed as a Global Muon and as a Tracker Muon with more than one matched muon station and more than 10 hits in the inner tracker, including at least one pixel hit; a Global Muon fit, with at least one muon chamber included, should have \( \chi^2/\text{ndof} < 10 \); the transverse impact parameter of the muon track with respect to the primary vertex should be less than 2 mm [89]. \( \varepsilon_{\text{rec+id}} \) of the tight muon selection is above 93%, as shown in Figures 3.6c and 3.6d. The probability of misidentifying a pion or a kaon with \( |\vec{p}| > 15 \) GeV as a muon is below 0.1% and 0.3%, respectively [89].

### Muon isolation

In this case, as for electrons from tau decays, we are interested in isolated muons.

Requiring that the prompt muons should be isolated allows rejecting muons within jets and improving the purity of the muon selection. The muon isolation is defined using the same PF technique as for the electrons (Equation (3.2)). In this case, \( I_{\text{PF}} \) is calculated, excluding the candidate muon, within a cone of \( \Delta R = 0.4 \) around the direction of the lepton, excluding an innermost region, muon’s veto cone. For the charged PF candidates, the veto cone is \( \Delta R > 0.0001 \), while for the photons, neutral hadrons, and pileup contributions, the veto cone is \( \Delta R > 0.01 \). In addition to the requirements for electrons, for muon isolation, the photon and neutral hadron candidates are required to have a transverse energy \( E_T > 0.5 \) GeV.

The isolation efficiency was measured by the tag-and-probe technique in \( Z \rightarrow \mu\mu \) events. For the muons with \( 20 < p_T < 50 \) GeV and \( I_{\text{PF}}^{\text{rel}} < 0.1 \), the efficiency is above 92%, while the background efficiency measured in the QCD-enhanced dataset is below 2.5%, as shown in Figure 3.7.
3.2. Physical objects

![Graphs showing muon reconstruction and identification efficiency](image)

Figure 3.6: Muon reconstruction and identification efficiency ($\varepsilon_{\text{rec+id}}$) measured by the tag-and-probe method in $J/\psi \rightarrow \mu\mu$ events for $p_T < 20\,\text{GeV}$ and in $Z \rightarrow \mu\mu$ events for $p_T > 20\,\text{GeV}$. The plots show the efficiency as a function of $p_T$ in the barrel and endcap regions for the PF muon selection (a,b) and the tight muon selection (c,d). [89]

3.2.3 Jets

QCD processes constitute the dominant part of the underlying processes in the p–p collisions at the LHC. Quarks and gluons have a very short lifetime and almost immediately undergo fragmentation and hadronisation, becoming a part of a meson or baryon, with the top quark, which has a lifetime long enough to decay into $Wb$, being the only exception. In the highly energetic QCD processes, due to the Lorentz boost, most of the final state products are concentrated in a narrow $\eta\phi$-cone and are reconstructed as jets.

The jet reconstruction in CMS starts from the PF particles (Section 3.1.3), which are clustered using the anti-$k_t$ algorithm [90] to form jets. The anti-$k_t$ algorithm belongs to the class of sequential recombination algorithms with a continuous
parametrization, which generalize the $k_t$ [91] and Cambridge/Aachen [92,93] jet finding algorithms.

The anti-$k_t$ algorithm provides a jet reconstruction that is robust against infrared and collinear soft radiation, which is an important phenomenological property, and, at the same time, the reconstructed jets have cone shapes without any irregularities. Due to the large number of particles in the event ($N$), jet reconstruction can be very CPU consuming, therefore, another advantage of the anti-$k_t$ algorithm is that CMS uses the FastJet package [88], which implements the anti-$k_t$ algorithm with an algorithmic complexity of $O(N \ln N)$ [94], which significantly reduces the time requirements for the jet reconstruction.

The anti-$k_t$ algorithm starts with a list of the PF particles. In each iteration, it recombines the current set of particles and jet candidates, both of which are referred to as entries. The distances between pairs of entries ($d_{ij}$) and the parameter used to represent the distance between an entry and the beam line ($d_{iB}$) are updated in each step and are defined as [90]

\[
\begin{align*}
    d_{ij} &= \min \left( p_{T_i}^{-2}, p_{T_j}^{-2} \right) \frac{\Delta R_{ij}^2}{R^2}, \\
    d_{iB} &= p_{T_i}^{-2},
\end{align*}
\]

where $p_{T_i}$ is the transverse momentum of entry $i$, $\Delta R_{ij}$ is the distance in the $\eta \phi$-plane between entries $i$ and $j$, and $R$ is a radius parameter that defines balance between the power of geometrical and kinematical variables and is fixed in the CMS to 0.5.
If $d_{iB}$ is smaller than $\min_j d_{ij}$, entry $i$ is promoted to a reconstructed jet and excluded from subsequent iterations. Otherwise, entries $i$ and $j' = \arg \min_j d_{ij}$ are merged into a new entry $i'$ and the algorithm proceeds to the next iteration. The algorithm stops when all candidate jets are promoted to be a reconstructed jet or when $\min(d_{ij}, d_{iB})$ passes below the threshold $d_{\text{cut}}$.

To distinguish the jets that come from the pileup interactions, a multivariate approach, the BDT, has been used [95]. Using the precise tracking information, and the knowledge of the primary vertex, the variables that define the proximity of the charged PF particles in a jet to the primary vertex are used as BDT discriminating variables. Other discriminating variables associated to the diffuseness of the jet are used in the discriminator, since jets coming from pileup are generally wider. The most discriminating shape variable is a $p_T$-weighted average $\eta\phi$ distance of PF candidates within the jet with respect to the jet direction:

$$\langle \Delta R^2 \rangle = \frac{\sum_i \Delta R_{ij}^2 p_T^2}{\sum_i p_T^2}.$$  \hspace{1cm} (3.4)

The BDT was trained on Z+jets simulated events.

In this thesis, jets are required to pass the loose working point of the discriminator. At the loose WP, the pileup jet identification for the central jets with $30 < p_T < 50$ GeV has a signal efficiency of $\approx 99\%$ with a background rejection of 90\%–95\%, while in the endcap and forward regions it has a signal efficiency of $\approx 80\%$ with a background rejection is around 60\% [95].

**Jet energy scale**

Due to the non-linearities in the detector response, imperfect detector modelling, noise, and pileup effects, the raw jet four-momentum obtained by the anti-$k_t$ algorithm ($p_{\mu}^{\text{raw}}$) does not represent an optimal estimate of the true jet four-momentum ($p_{\mu}^{\text{true}}$). To take these effects into account and improve, on average, the jet momentum estimation, a multiplicative correction factor $C$ (the jet energy scale factor) is applied to the $p_{\mu}^{\text{raw}}$ [96]:

$$p_{\mu}^{\text{cor}} = C(p_{\mu}^{\text{raw}}, \eta) \cdot p_{\mu}^{\text{raw}}.$$  \hspace{1cm} (3.5)

$C$ is composed of four sequential correction factors: the offset correction ($C_{\text{offset}}$) allows to remove the energy excesses due to pileup and noise; it is followed by the Monte Carlo correction ($C_{\text{MC}}$) that aims to remove the non-linearity in $p_T$ and non-uniformity in $\eta$ of the detector response, using MC truth; after that, the relative ($C_{\text{rel}}$) and the absolute ($C_{\text{abs}}$) residual energy scale corrections are applied to fix small disagreements between the data and the simulations.

The correction factors were estimated using calibration data [96]. For the PF jets with $p_T > 20$ GeV, the combined correction is below 20\%. Its dependency on $\eta$ is
Chapter 3. Event reconstruction and physical objects

Figure 3.8: The combined jet energy correction factor $C$ as a function of jet $\eta$ for jets with $p_T = 50$ GeV (a) and $p_T = 200$ GeV (b). The correction factor is shown for three different jet reconstruction algorithms: PF jets, JPT jets and CALO jets. Only PF jets are used in the analysis presented in this thesis. For PF jets, the total uncertainty varies between 3% and 5%, depending on $p_T$. [96]

**b jet tagging**

The $b$ quark has a relatively big mass of $(4.66 \pm 0.03)$ GeV [4] and has a high probability of hadronising into long-lived hadrons, i.e. $b$ hadrons. Therefore, the information about the secondary vertices, kinematic properties and composition of the jets can be exploited in $b$ jet identification ($b$ jet tagging) against jets originating from light quarks or gluons.

To reconstruct the secondary vertices (SV), the same algorithm used for the primary vertex (PV) reconstruction (Section 3.1.2) is applied to the tracks within the given jet, requiring the $\Delta R$ between the track direction and the jet axis to be less than 0.3 [97].

After the SV candidates have been reconstructed, additional requirements are applied in order to improve the discriminating power of the $b$-tagging: the fraction of the shared tracks between the primary and secondary vertices should be less than 65%, the significance of the radial distance between the PV and SV should be greater than $3\sigma$, and the SV should be inside the cone $\Delta R < 0.5$ around the jet axis; lastly, to reduce the contamination due to interactions with the detector and decays of heavy long-lived mesons, secondary vertices with a radial distance $> 2.5$ cm with masses $> 6.5$ GeV or compatible with the $K^0$ mass are rejected [97].

The Combined Secondary Vertex (CSV) algorithm uses a combination of secondary vertex information and tracker-based variables in order to separate $b$ jets...
3.2. Physical objects

from other jets \[97\]. Use of the tracker-based variables allows restoring the efficiency of the b tagging, when the SV cannot be properly reconstructed.

The CSV uses a set of more than 9 variables with high discriminating power, including information about the quality of the SV reconstruction, the vertex mass, the significance of the flight distance, the significances of the impact parameters of the tracks relative to the primary vertex ($S_{IP} \equiv \frac{d_{xy}}{\sigma d_{xy}}$), etc. These variables are incorporated into two likelihood ratios that are used to discriminate b jets from c jets ($Q_c$) and b jets from light-parton jets ($Q_{light}$). The $Q_c$ and $Q_{light}$ are then combined into a single CSV discriminator with the prior weights of 0.25 and 0.75, respectively.

The distribution of the CSV discriminator for different types of jets and its performance against c jet misidentification obtained in the simulations are shown in Figure 3.9. Three WPs were defined, based on the expected misidentification rate, as summarized in Table 3.2 allowing flexible b jet selection at the analysis level.

![Figure 3.9: Distribution of the CSV discriminator for different jet flavors (a), and the c jet misidentification probability as a function of the b jet efficiency obtained from simulations for different b tagging algorithms (b) \[97\]. The analysis presented in this thesis uses the CSV algorithm.](image)

<table>
<thead>
<tr>
<th>Working point</th>
<th>CSV threshold</th>
<th>Expected misidentification probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loose</td>
<td>0.244</td>
<td>0.1</td>
</tr>
<tr>
<td>Medium</td>
<td>0.679</td>
<td>0.01</td>
</tr>
<tr>
<td>Tight</td>
<td>0.898</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Table 3.2: Definition of b tag working points with the corresponding CSV value thresholds and the expected misidentification probabilities. To pass the WP identification criterion, a jet should have a CSV tag value greater than the threshold value.

The b tagging efficiency measured in multijet (QCD) and $t\bar{t}$ events and the b tagging misidentification probability measured using the ‘negative tag rate’ technique allowed establishing scale factors to correct the simulation predictions \[97\]. The Data/MC scale factors for b tag efficiency and misidentification probability for the medium CSV working point as a function of the b jet $p_T$ are shown in Figure 3.10.
Based on these scale factors, at the analysis level, the b tag efficiency and misidentification probability were independently scaled within the uncertainties to take into account the corresponding systematics.

\[ SF_b(p_T) = \frac{\alpha + \beta p_T}{1 + \gamma p_T} \]

A polynomial fit has been used for the b jet misidentification probability scale factor. [97]

3.2.4 Taus

\( \tau \) is the heaviest SM lepton, with a mass of \((1.77686 \pm 0.00012) \text{ GeV}\) and a relatively short mean lifetime of \((2.903 \pm 0.005) \times 10^{-13} \text{ s}\) [4], which means that it decays before reaching the detector volume. The main \( \tau \) decay modes are listed in Table 3.3. In \( \approx 35.2\% \) of the cases, when a \( \tau \) decays into an electron or a muon and neutrinos (these are referred to as leptonic decays) it is difficult to separate the \( \tau \) signature from the other electron and muon production processes, therefore the \( \tau \) is reconstructed as an electron Section 3.2.1 or as a muon Section 3.2.2, respectively. In the other \( \approx 64.8\% \) of the cases, when the \( \tau \) decay contains hadrons, the \( \tau \) is reconstructed as a hadronic \( \tau \) jet (\( \tau_h \)) using the Hadron Plus Strips algorithm [98,99].

The Hadron Plus Strips algorithm

The \( \tau \) reconstruction starts from PF jets (Section 3.2.3) with \( p_T > 14 \text{ GeV} \) and \( |\eta| < 2.5 \) as seeds. To properly reconstruct \( \pi^0 \) candidates within a \( \tau \) jet, \( \gamma \to ee \) conversion of the photons originated in \( \pi^0 \to \gamma\gamma \) decays should be taken into
3.2. Physical objects

<table>
<thead>
<tr>
<th>Decay mode</th>
<th>Branching fraction, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau^{-} \rightarrow \mu^- \bar{\nu}<em>\mu \nu</em>\tau$</td>
<td>$17.41 \pm 0.04$</td>
</tr>
<tr>
<td>$\tau^{-} \rightarrow e^- \bar{\nu}<em>e \nu</em>\tau$</td>
<td>$17.83 \pm 0.04$</td>
</tr>
<tr>
<td>$\tau^{-} \rightarrow l^- \bar{\nu}<em>l \nu</em>\tau$</td>
<td>$35.24 \pm 0.08$</td>
</tr>
<tr>
<td>$\tau^{-} \rightarrow h^- \bar{\nu}_\tau$</td>
<td>$11.53 \pm 0.06$</td>
</tr>
<tr>
<td>$\tau^{-} \rightarrow h^- \pi^0 \nu_\tau$</td>
<td>$25.95 \pm 0.09$</td>
</tr>
<tr>
<td>$\tau^{-} \rightarrow h^- \pi^0 \pi^0 \nu_\tau$</td>
<td>$9.53 \pm 0.11$</td>
</tr>
<tr>
<td>$\tau^{-} \rightarrow h^- h^- h^+ \nu_\tau$</td>
<td>$9.80 \pm 0.06$</td>
</tr>
<tr>
<td>Other modes with hadrons</td>
<td>$\approx 7.95$</td>
</tr>
<tr>
<td>All hadronic modes</td>
<td>$\approx 64.76$</td>
</tr>
</tbody>
</table>

Table 3.3: Branching fractions of different $\tau$ decay modes. $\tau^+$ modes are charge conjugates of the listed $\tau^-$ modes; $h^\pm$ stands for $\pi^\pm$ or $K^\pm$; $l^-$ stands for $e^-$ or $\mu^-$. \[4\]

account. For that reason, the PF electrons and PF photons within the jet are clustered into strips. These strips are created by combining the electrons and photons in a $\eta\phi$-window of size $0.05 \times 0.20$, starting from the most energetic candidates as the seeds. All reconstructed strips with $\sum p_T > 2.5$ GeV are considered as $\pi^0$ candidates.

After clustering, the strips are combined with the charged PF particles within the jet, requiring the $p_T$ of the selected particles to be greater than 0.5 GeV. Contributions from the pileup are reduced by considering only particles that are compatible with the estimated $\tau_h$ production vertex, i.e. when the longitudinal ($|\Delta z|$) and transverse ($|d_{xy}|$) distances between the particle track and the position of the production vertex are less than 0.4 and 0.03 cm, respectively.

To choose between possible combinations of strips and charged particles, multiple $\tau_h$ hypotheses are considered. Each hypothesis includes one or three charged particles and 0–2 strips, and corresponds to the expected signature from one of the $\tau_h$ decay modes listed in Table 3.3: $h^\pm$, $h^\pm \pi^0$, $h^\pm \pi^0 \pi^0$ or $h^\pm h^\mp h^\mp$.

The compatibility between the hypothesis and the expected signature is tested, requiring that the charge of the $\tau_h$ should be $\pm 1$, that all particles and strips are within the signal cone ($\Delta R_{\text{sig}}$), and that the mass of the $\tau_h$ candidate is inside the expected mass window. The signal cone is defined as \[99\]

$$\Delta R_{\text{sig}} = \min \left[ 0.1, \max \left( 0.05, \frac{3.0 \text{ GeV}}{p_T} \right) \right].$$  \tag{3.6}$$

The definition of the mass window depends on the signature with which the compatibility is to be tested. For the $h^\pm$ signature test, no mass window is used; for the $h^\pm h^\pm h^\mp$ signature test, the reconstructed mass ($m_{\tau_h}$) is required to satisfy $0.8 < m_{\tau_h} < 1.5$ GeV; for the $h^\pm \pi^0$ and $h^\pm \pi^0 \pi^0$ signature tests, the mass window is defined by the following formula:

$$m_{\text{down}} < m_{\tau_h} < \min \left[ m_{\text{up}}^\text{low}, \max \left( m_{\text{up}}^\text{high}, m_{\text{up}}^\text{low} \sqrt{\frac{p_T}{100 \text{ GeV}}} \right) \right],$$  \tag{3.7}$$
where $m_{\text{down}}$, $m_{\text{up}}^\text{low}$ and $m_{\text{up}}^\text{high}$ are, respectively, equal to 0.3, 1.3 and 4.2 GeV for $h^+\pi^0$ and to 0.4, 1.2 and 4.0 GeV for $h^+\pi^0\pi^0$.

To ensure that one seeding jet does not give rise to multiple $\tau_h$ candidates, in the case that several hypothesis pass the compatibility requirements, the hypothesis with the highest $p_T$ is chosen as the reconstructed $\tau_h$ object.

$\tau$ and $\tau'$ decay products are usually isolated from other p–p interaction products. This property can be used to separate $\tau$ candidates from the jets originating from quarks and gluons, which, in general, are not isolated.

The $\tau_h$ isolation ($I_\tau$) is calculated using charged particles and photons within the cone $\Delta R < 0.5$ around the $\tau$ direction that satisfy $p_T > 0.5$ GeV and were not used to form the $\tau_h$ candidate. The pileup effects are reduced by requiring charged the particles’ tracks to be compatible with the $\tau_h$ production vertex ($|\Delta z| < 0.2$ cm and $|d_{xy}| < 0.03$ cm) and applying statistical corrections for contributions from the pileup photons ($\Delta \beta$). Moreover, to take into account the particles with associated tracks not well reconstructed, track quality requirements for the charged particles within the isolation cone are relaxed, requiring only 3 hits, instead of 8, in the tracker.

The $\tau_h$ isolation is defined as:

$$I_\tau = \sum_{\text{charged}} p_T + \max \left( 0, \sum_{\gamma} p_T - \Delta \beta \right).$$

(3.8)

$\Delta \beta$ is defined as a scaled (0.4576) sum of the charged particles’ $p_T$ within the cone $\Delta R < 0.8$ around the $\tau_h$ direction that satisfy $\Delta z > 0.2$ cm with respect to the $\tau_h$ vertex:

$$\Delta \beta = 0.4576 \sum_{\text{charged}} p_T.$$  

(3.9)

Three working points for $\tau_h$ isolation have been defined, requiring $I_\tau$ to be less than 2.0 GeV (loose), 1.0 GeV (medium) and 0.8 GeV (tight). The identification efficiency measured in $Z \to \tau \tau$ data and simulations is above 40% for all $\tau_h$ candidates with $|\eta| < 2.4$ passing the medium WP selection, as shown in Figure 3.11.

At the same time, the probability to misidentify a quark or a gluon jet as a $\tau_h$, measured in W+jets and QCD events, is below 2% and 0.6%, respectively (Figure 3.12).

To further improve the $\tau_h$ identification, two additional discriminators against electrons and muons were introduced. They help to reduce the probability of misidentifying isolated electrons and muons from W and Z decays as $h^\pm$ or $h^\pm\pi^0$ decay modes.

To identify $\tau_h$ against electrons, observables that allow separating hadronic and electromagnetic showers, including quantities that characterize calorimeter energy patterns and the matching between tracks and calorimeter clusters, are combined in the BDT-based discriminator. The BDT was trained using different W, Z,
3.2. Physical objects

Figure 3.11: \( \tau_h \) identification efficiency measured in \( Z \rightarrow \tau \mu \tau_h \) events as a function of (a) \( p_T \) and (b) \( \eta \) for different isolation working points \[99\].

\( t\bar{t} \) and SM H samples that include a \( \tau \) or electron in the final state. Different working points were defined to provide flexibility between identification efficiency and misidentification probability for the \( \tau_h \) selection.

To discriminate \( \tau_h \) against muons, two working points, loose and tight, with a cut-based selection were defined \[99\]. To pass the loose WP selection, there is required no more than one muon station that matches the direction of the \( \tau_h \) candidate within the cone \( \Delta R < 0.3 \) and with total deposited energy in the calorimeters equal to at least 20% of the momentum of the leading track in the \( \tau_h \). In the tight WP selection, in addition to the loose WP constraints, it is required that there be no hits in the detectors of the two outer muon stations within the cone \( \Delta R < 0.3 \) around the direction of the \( \tau_h \) candidate.

The measurements of the performance of the \( \tau_h \) identification and reconstruction in \( Z \rightarrow \tau \tau \) and W+jets \[98,99\] showed that the overall systematic uncertainty in the simulated \( \tau_h \) energy scale is 3%. Based on these measurements, in order to improve, on average, the agreement between the data and simulations, three correction factors were introduced. A factor of 0.88 was introduced to correct the reconstruction efficiency of the \( h^\pm \) decay mode predicted by the simulations. Another factor of 1.012 was introduced to correct simulated the \( \tau_h \) energies for the \( h^\pm \pi^0, h^\pm \pi^0 \pi^0 \) and \( h^\pm h^\pm h^\mp \) decay modes. The last factor \( \alpha_{h^\pm}^{mis-id} \) is needed to correct the misidentification rate for jets as \( \tau_h \) in the W+jets simulations.
Figure 3.12: Probability of misidentifying quark and gluon jets as a \(\tau_h\) as a function of \(p_T\) measured in (a) W+jets and (b) QCD multijets data and simulations for different isolation working points [99].

The \(w_{\text{jet-}\tau}^{\text{mis-id}}\) is fitted by a polynomial function of \(p_T\) of the \(\tau_h\) candidate:

\[
\begin{align*}
  w_{\text{jet-}\tau}^{\text{mis-id}} &= 1.15743 - 7.36136 \times 10^{-3} \cdot p_{T\text{eff}} \\
                             &\quad + 4.3699 \times 10^{-5} \cdot p_{T\text{eff}}^2 - 1.188 \times 10^{-7} \cdot p_{T\text{eff}}^3,
\end{align*}
\]

where \(p_{T\text{eff}} = \min(200, p_T[\text{GeV}])\). The simulated events are weighted by this factor.

### 3.2.5 Missing transverse energy

Because neutrinos do not undergo either electromagnetic or strong interactions, the rate of their interactions with matter is extremely small and, therefore, they cannot be directly detected by the CMS. However, the presence of neutrinos can be inferred from an imbalance in the sum of the transverse momenta of all the visible products in the event. The transverse momentum of the colliding proton pair is negligibly small, and therefore, from the conservation of 4-momentum it follows that the sum of the transverse momenta of all the p–p interaction products should be 0.

The missing transverse energy \(\vec{E}_T\) is defined as

\[
\vec{E}_T = - \sum_{\text{visible}} \vec{p}_T,
\]

where the sum is over all observed particles in the event.
The missing energy plays an important role in the analysis of $H \rightarrow hh \rightarrow bb\tau\tau$, due to the presence of at least three neutrinos from $\tau$ decay and eventually other neutrinos in the b jets. Its use in the analysis will be described in the next chapter.

Since $\vec{E}_T$ uses the transverse momenta of all the particles in the event, the accuracy of $\vec{E}_T$ is very sensitive to the calibration of the sub-detectors, the reconstruction efficiency, and the pileup effects. The high sensitivity to the overall detector performance may result in anomalously large $\vec{E}_T$ values in some events. Special noise-rejection techniques that monitor the ECAL, HCAL, and tracker measurements, are used to reject the events with a fake $\vec{E}_T$.

In order to achieve an optimal resolution, this analysis uses an MVA-based technique, the MVA PF $\vec{E}_T$ algorithm, which was developed to combine 5 possible methods to estimate $\vec{E}_T$ [101]. The MVA PF $\vec{E}_T$ algorithm consider the hypothesis that the hard scattering process in the event contains a boson $V_{\text{hard}}$ (e.g., a Z or an SM Higgs) that decays into a pair of leptons and that the dependence of the main uncertainties in the $\vec{E}_T$ measurements on the other hard scattering products, excluding $V_{\text{hard}}$, can be neglected. This hypothesis is reasonable for the $H \rightarrow hh \rightarrow bb\tau\tau$ process, because most of the invisible momentum in the final state is due to $h \rightarrow \tau\tau$.

Under this hypothesis, Equation (3.11) can be written into the following form:

$$\vec{q}_T + q_T + \vec{u}_T = 0,$$

(3.12)

where $\vec{q}_T = \sum p_T$ over all visible $V_{\text{hard}}$ decay products and $\vec{u}_T$ is the boson recoil. To avoid a dependency on the global coordinate system, it is convenient to define the recoil vector in the $\vec{q}_T$-based coordinates as parallel ($u_\parallel$) and perpendicular ($u_\perp$) to the $\vec{q}_T$ direction.

The 5 initial estimates of the missing transverse energy ($\vec{E}_T^i$) are based on different selections of the PF particles (Section 3.1.3) in the event, which are defined by varying the requirements on the vector of origin for the charged PF particles, and on the jet MVA pileup identification (Section 3.2.3) for the neutral PF particles. These selections were chosen in such a way that from linear combinations of $\vec{E}_T^i$, one could obtain 5 uncorrelated quantities that describe the charged PF particles from (not from) the hard scattering process, the neutral PF particles in jet passing (failing) the MVA pileup identification, and the unclustered neutral PF particles [101].

The MVA PF $\vec{E}_T$ algorithm uses a BDT regression to compute $\vec{E}_T^{\text{MVA}}$. For each $\vec{E}_T^i$, $\vec{u}_T$ is calculated and its magnitude and azimuthal angle are used as the BDT inputs. The other BDT inputs are $\sum_j |p_T^j|$ over all PF particles included in $\vec{E}_T^i$, the momenta of the two jets with the highest $p_T$, and the number of primary vertices in the event [101]. The output of the BDT is the corrected recoil ($\vec{u}_T^{\text{MVA}}$), which is added to $\vec{q}_T$ in order to obtain $-\vec{E}_T^{\text{MVA}}$.

Data and simulated samples of the $Z \rightarrow \mu\mu$ process were chosen for the BDT training, which is an optimal choice because of the absence of neutrinos in the
Chapter 3. Event reconstruction and physical objects

hard scattering process. The obtained resolutions for $u_\parallel$ and $u_\perp$ for $q_T < 200$ GeV are smaller than 20 GeV and 15 GeV, respectively, as shown in Figure 3.13.

Figure 3.13: Resolution of the (a) parallel and (b) perpendicular recoils ($\vec{u}$) as a function of the visible transverse $Z$ momentum ($q_T$) for the MVA PF $\vec{E}_T$ in $Z \rightarrow \mu\mu$, $Z \rightarrow ee$ and direct-photon events. The upper frame shows the resolution in data, while the lower frame shows the ratio between the data and simulations and corresponding systematic uncertainties. [101]

To reduce, on average, the discrepancies between the simulated recoil and the recoil calculated from the data, additional corrections were established. The corrections are parametrized as a function of the generated $V_{\text{hard}}$ boson $p_T$ and affect only simulations that contain generated $h$, $Z$ or $W$. The parameterization was obtained from the discrepancies in the BDT responses between the $Z \rightarrow \mu\mu$ data and the simulations. For each generated boson $p_T$ range, the $\vec{u}_{\text{MVA}}$ distribution from the simulations was smeared to match the distribution from the data. The smearing function is defined as $\sqrt{w_{\text{data}}^2 - w_{\text{MC}}^2}$, where $w_{\text{data}}$ and $w_{\text{MC}}$ are the widths of the $\vec{u}_{\text{MVA}}$ distribution in the data and the simulations, respectively.
Chapter 4

$H \to hh$ candidate selection and reconstruction

Starting from the physical objects described in the previous chapter, this chapter describes the selection and the reconstruction steps of the heavy Higgs $H$ candidate decay mode $H \to h[bb][\tau_l\tau_h]$, where $l$ indicates an electron or a muon. The decay modes of $h$, $\tau_e\tau_h$ and $\tau_\mu\tau_h$ covers approximately 45.6% of the possible di-$\tau$ final states. The advantage of studying these two final states together is that most of the techniques used in the analysis are the same for both leptons.

The goal of the applied selection and reconstruction techniques is to select the signal in the most efficient way by reducing the presence of the backgrounds that share a signature similar to the signal. The strategy followed makes use of the datasets employed by CMS for the searches for the SM and of MSSM Higgs decaying into a pair of taus. An accurate modelling of the data sample selected in these datasets is done: the signal is modelled using simulation while the modelling of the physical background uses simulated samples are data-driven techniques. The result of the selection and reconstruction steps is a set of distributions of reconstructed $H$ candidate masses for the signal, the modelled backgrounds, and the data, that are used to fit the model to the data, as described in the next chapter.

4.1 Data sample

The total number of proton–proton collisions recorded by the CMS detector in 2012 during stable beam periods corresponds to an integrated luminosity of $21.79 \pm 1$ fb at centre of mass energy of $\sqrt{s} = 8$ TeV. The collected data are separated into datasets according to a set of satisfied trigger requests designed to efficiently record, at the same time, interesting processes characterized by different

---

1Since we do not select events based on the $b$ quark hadronization mode for $h \to bb$, in the further text the ‘$\tau_e\tau_h$ channel’ will mean the $bb\tau_e\tau_h$ final state, and by the ‘$\tau_\mu\tau_h$ channel’ we will mean the $bb\tau_\mu\tau_h$ final state. Also, in some places in the text, an alternative notation is used, where these two channels are, respectively, denoted by $e\tau_h$ and $\mu\tau_h$. 

67
signatures. Furthermore, since over time the trigger configurations were slightly changed, each dataset is separated into run periods.

This analysis uses the TauPlusX dataset. It collects events with a $\tau_h$ plus an additional physical object (namely, electron + $\tau_h$, muon + $\tau_h$) and events with muon + large $E_T$. The run ranges and their integrated luminosities are summarized in Table 4.1. For this analysis, events were selected from the TauPlusX dataset by requiring a muon or an electron plus a $\tau_h$ physical object, as described in Section 4.2. In addition, quality requirements on the full functionality conditions for each CMS sub-detector and accelerator have been in order to accept the runs used in this analysis. The resulting integrated luminosity of the analysed data sample is $\approx 19.7 \text{ fb}^{-1}$.

<table>
<thead>
<tr>
<th>Dataset name</th>
<th>Run range</th>
<th>Luminosity, fb$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>/TauPlusX/Run2012A-22Jan2013-v1/AOD</td>
<td>190456 – 193621</td>
<td>0.887</td>
</tr>
</tbody>
</table>

Table 4.1: TauPlusX datasets that were used in this thesis with ranges of covered LHC runs and corresponding integral luminosity, computed after applying quality requirements to the events in the dataset.

4.2 Trigger requirements

Trigger selections were applied for both the data and the simulations, requiring events to pass the L1 and the HLT paths listed in Table 4.2. These paths have the following requirements at L1 and at HLT (Section 2.2.6), depending on the period in which the data was taken.

- **L1**: a single $e/\gamma$ candidate with $E_T > 18 – 22$ GeV and $|\eta| < 2.17 – 3$ with or without the isolation requirements, depending on the trigger path; or a single muon candidate with $p_T > 14 – 16$ GeV and $|\eta| < 2.1$.

- **HLT**: the candidate seeded by an L1 electron (or muon) should be loosely identified and isolated, and have a $p_T > 20 – 22$ GeV ($p_T > 17 – 18$ GeV). In addition, there is required the presence of a loosely isolated PF $\tau_h$ candidate (Section 3.2.4) with $p_T > 20$ GeV.

Over the entire period of collecting data, the output of the HLT trigger paths used were not pre-scaled, collecting the full events statistic and, therefore, the corresponding data sample normalization corrections must not be applied.

The same trigger paths were used for the SM $H \rightarrow \tau_\ell\tau_h$ channel.

---

2 The list of runs that were certified to pass the quality requirements is specified in a json file (Cert_190456-208686_8TeV_22Jan2013ReReco_Collisions12_JSON.txt).
4.2. Trigger requirements

<table>
<thead>
<tr>
<th>Channel</th>
<th>HLT path</th>
<th>L1 seeds</th>
<th>Luminosity, fb⁻¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_e \tau_h$</td>
<td>HLT_Ele20_CaloIdVT_CaloIsoRhoT_TrkIdT_TrkIsoT_LooseIsoPFTau20</td>
<td>L1_SingleIsoEG18er, L1_SingleEG20</td>
<td>0.883</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>18.900</td>
</tr>
<tr>
<td>$\tau_\mu \tau_h$</td>
<td>HLT_IsoMu18_eta2p1_LooseIsoPFTau20</td>
<td>L1_SingleMu14er, L1_SingleMu16er</td>
<td>0.889</td>
</tr>
<tr>
<td></td>
<td>HLT_IsoMu17_eta2p1_LooseIsoPFTau20</td>
<td>L1_SingleMu14er, L1_SingleMu16er</td>
<td>18.900</td>
</tr>
</tbody>
</table>

Table 4.2: HLT trigger paths and corresponding L1 seeds used for $\tau_e \tau_h$ and $\tau_\mu \tau_h$ channels with the effective integrated luminosity.

The trigger efficiencies were measured using the tag-and-probe technique [104] in the $Z \to ll$ data and simulated events. This is the same technique used in the CMS SM Higgs search for the decay mode mentioned above. Assuming that the final trigger selection efficiency can be estimated as the product of the efficiencies of selecting each trigger component (trigger leg), the probability of passing the trigger selection for electrons, muons and hadronic taus were estimated separately. For this purpose, different $Z \to ll$ samples of data and simulated events were used: $Z \to ee$ to estimate the trigger efficiency for electrons, $Z \to \mu\mu$ for muons, and $Z \to \tau_\mu \tau_h$ for hadronic taus. In each case, after a pre-selection, which ensures a purity of the events, one lepton is chosen as a ‘tag’ and should pass the corresponding single electron (muon) trigger requirements, while the other lepton is used as a ‘probe’ leg for the triggers reported in Table 4.2.

The efficiencies for the probe at passing the trigger condition of the electron, muon and $\tau_h$ legs of the $\tau_e \tau_h$ and $\tau_\mu \tau_h$ triggers were measured as functions of $p_T$ in bins of $\eta$, separately for the $Z \to ll$ data and Monte Carlo simulated events. This function is called the ‘turn-on curve’, because it has a sharp rise and then a flat plateau after a certain $p_T$. Turn-on curves, in each $\eta$ region, have been fitted using the crystal-ball function [100] for both data ($T^\text{data}_\eta(p_T)$) and simulations ($T^\text{MC}_\eta(p_T)$). An example in the $\eta$ region $0 < \eta < 0.8$ of the turn-on curve fits of both channels, $\tau_e \tau_h$ and $\tau_\mu \tau_h$ is shown in Figure 4.1.

To reproduce the experimental trigger performance, in the $H \to hh \to b\bar{b}\tau_\mu \tau_h$ analysis, each simulated event (see Section 4.3 for the MC samples used) is weighted by the measured ratio of turn-on curves for each trigger leg $i$:

$$w_i^{\text{trigger}} = \frac{T^\text{data}_\eta(p_T)}{T^\text{MC}_\eta(p_T)}.$$  (4.1)
Figure 4.1: Example of fitted, using crystal ball function, distributions of trigger efficiencies for objects in the $\eta$ region $0 < \eta < 0.8$, as a function of the object’s $p_T$ for the (a) electron and (b) $\tau_h$ HLT legs in the $\tau_e\tau_h$ channel and for the (c) muon and (d) $\tau_h$ HLT legs in the $\tau_\mu\tau_h$ channel. The fit is performed for data (red) and MC simulations (blue).

### 4.3 Sample composition modelling

After the selections mentioned above, the data sample contains a large variety of processes including processes with signatures similar to that of the $H \to hh \to bb\tau_l\tau_h$ process, referred to as the signal process, as was discussed in Section 1.5.

To model the composition of the data, both data-driven and simulation-based techniques were used.

The simulation-based techniques allow a modelling of some particular processes that entirely reflects the currently achieved level of precision and of details in the description of these processes and of the detector responses. This gives the opportunity to perform detailed studies of the process signatures using simulated events, in order to separate the signal from the other processes or to apply a specific selection to reduce the contribution of a background process. On the other hand, the simulation-based techniques are significantly affected by the systematic uncertainties due to an imprecise modelling of the detector response, by the limited accuracy in the pileup and QCD generator-level description, by uncertainties in the luminosity measurements, and by uncertainties in the theoretical cross-sections.
4.3 Sample composition modelling

For some processes \((Z \rightarrow \tau\tau + \text{jets}, \ W \rightarrow ll + \text{jets}, \ l = e, \mu, \tau\ \text{and QCD multijets events})\), where the uncertainties due to the modelling are big, and/or necessary statistics are prohibitive, data driven techniques were used. In the data driven methods, the contribution from a particular process is estimated in a sideband region of the event parameter space where the presence of this process is expected to be dominant, while the presence of the signal is expected to be negligible. After that, the obtained contribution is extrapolated from the sideband region to the region used by the final model fit for the signal extraction, the signal region. To avoid big systematic uncertainties due to the extrapolation, the parameter space of the sideband region should be as close as possible to that of the signal region.

Therefore, data driven methods were applied after the initial pre-selection, also referred to as the baseline selection, and will be described in Section 4.4.3, while in the next subsection the modelling of contributions using simulated samples will be described.

4.3.1 Monte Carlo based modelling

Monte Carlo (MC) generators were used to model different processes, as summarized in Table 4.3.

<table>
<thead>
<tr>
<th>Process</th>
<th>Generator</th>
<th>Cross-section, pb</th>
</tr>
</thead>
<tbody>
<tr>
<td>(gg \rightarrow H(\rightarrow h[bb]h[\tau^+\tau^-]))</td>
<td>PYTHIA</td>
<td>-</td>
</tr>
<tr>
<td>(Z/\gamma^* (\rightarrow l^+l^-) + \text{jets})</td>
<td>MADGRAPH</td>
<td>3504</td>
</tr>
<tr>
<td>(W(\rightarrow ll) + \text{jets})</td>
<td>MADGRAPH</td>
<td>36257</td>
</tr>
<tr>
<td>(tt + \text{jets})</td>
<td>MADGRAPH</td>
<td>249.5</td>
</tr>
<tr>
<td>Single top (tW-channel)</td>
<td>POWHEG</td>
<td>22.2</td>
</tr>
<tr>
<td>(W(\rightarrow l\nu) + Z(\rightarrow l^+l^-) + \text{jets})</td>
<td>MADGRAPH</td>
<td>5.824</td>
</tr>
<tr>
<td>(Z(\rightarrow l^+l^-) + Z(\rightarrow q\bar{q}) + \text{jets})</td>
<td>MADGRAPH</td>
<td>2.502</td>
</tr>
<tr>
<td>(W(\rightarrow q\bar{q}) + Z(\rightarrow l^+l^-) + \text{jets})</td>
<td>MADGRAPH</td>
<td>2.207</td>
</tr>
<tr>
<td>(W(\rightarrow l\nu) + Z(\rightarrow l^+l^-) + \text{jets})</td>
<td>MADGRAPH</td>
<td>1.058</td>
</tr>
<tr>
<td>(Z(\rightarrow l^+l^-) + Z(\rightarrow \nu\bar{\nu}) + \text{jets})</td>
<td>MADGRAPH</td>
<td>0.716</td>
</tr>
<tr>
<td>(Z(\rightarrow l^+l^-) + Z(\rightarrow l^+l^-) + \text{jets})</td>
<td>MADGRAPH</td>
<td>0.181</td>
</tr>
</tbody>
</table>

Table 4.3: List of the processes simulated using Monte Carlo based techniques, specifying the MC generator used and the corresponding production cross-sections for 8 TeV p–p collisions. \(l\) stands for \(e, \mu\) or \(\tau\).

The \(H \rightarrow hh \rightarrow b\bar{b}\tau\tau\) process was modelled with the PYTHIA 6.4.26 \[105\] in the range of \(260 \leq m_H \leq 350\) GeV with a step of 10 GeV in \(m_H\). \(Z/\gamma^*(\rightarrow l^+l^-) + \text{jets}, \ W(\rightarrow l\nu) + \text{jets}, \ tt + \text{jets}\) and di-boson + jets processes were modelled with the MADGRAPH 5.1 \[106\] matrix element generator. The contribution from the single top quark processes (tW channels) were modelled with POWHEG 1.0 \[107\].
For all simulated samples, PYTHIA with the $Z2^*$ tune [111] has been used to model parton showering and fragmentation. The $\tau$ decays were simulated with TAUOLA [112], which was interfaced to all generators. Interactions of the long-lived event products with the detector were simulated with GEANT4 [113], using a highly detailed description of the CMS detector. After that, the simulated detector responses were digitized and the standard reconstruction procedure, starting from the trigger algorithms, was applied.

The process $Z(\rightarrow ll)$ where $l$ is $e$ or $\mu$ or $\tau$ uses the same value of the cross section calculated to the next-to-next-leading order (NNLO) [114]. For $Z \rightarrow ee$ and $Z \rightarrow \mu\mu$ the normalization uses the cross section value. To model the $Z \rightarrow \tau\tau + \text{jets}$ process, which is one of the main backgrounds, there was used a mixed modelling based on simulation and on data-driven techniques, as discussed in Section 4.4.3.

W+jets events are normalized using a data-driven technique (see Section 4.4.3), while the MC simulations were used to model the shapes of the distributions of the physical quantities of the events.

Normalization of the $t\bar{t}$ yield is based on the combined ATLAS and CMS cross section measurements [115].

The cross section values calculated to the next-to-leading order (NLO) were used for di-boson [116,117] and single-top [118] processes.

According to the prescriptions of PDF4LHC [119], the parton distribution functions (PDFs) CT10 [120] or CTEQ6L1 [121] for the proton are used, depending on the generator in question, together with MSTW2008 [122].

**Modelling of pileup in MC samples**

To model the pileup contribution, to each generated MC event a random number of minimum bias events, modelled by soft QCD processes in inelastic p–p collisions, were added. In order to better match the distribution of the average number of pileup interactions in the data (Figure 2.3), the simulated number of underlying pileup events follows a Poisson-like ‘S10′ distribution [123] defined by CMS. To account for the remaining differences between the data and the MC, a pileup-related weight ($w^{PU}$) is applied to each MC event. $w^{PU}$ is a function of the number of the ‘on-time’ pileup interactions ($n_{MC}^{PU}$) in the event, i.e. the pileup interactions that come from the same bunch-crossing as the hard scattering interaction. This weight is defined as:

$$w^{PU}(n_{MC}^{PU}) = H(60 - n_{MC}^{PU}) \cdot \frac{\nu_{data}(n_{MC}^{PU})}{\nu_{MC}(n_{MC}^{PU})},$$

(4.2)

where $H(n)$ is the Heaviside step function, $\nu_{data}(n)$ and $\nu_{MC}(n)$ are, respectively, the average frequencies of having $n$ pileup interactions in the data and in the MC events. The resulting dependency of $w^{PU}$ on $n_{MC}^{PU}$ is shown in Figure 4.2.
4.4 \( H \) candidate reconstruction

Overview

In order to identify events that contain potential \( H \rightarrow hh \rightarrow b\tau l\tau h \) candidates, a special selection procedure is defined.

At the beginning of the selection, each event should pass the trigger requirements (Section 4.2) and have a primary hard interaction vertex that passes the selection defined in Section 3.1.2.

The first step of the procedure, also referred to as the baseline selection, is the selection of events with a \( h \rightarrow \tau l \tau h \) candidate reconstructed, requiring a well identified \( l\tau h \)-pair, passing vetoes that constrain the event parameter space to be close to the signal one. The baseline selection is the same as used for the SM and MSSM \( H \rightarrow \tau\tau \) searches \cite{103,124}. Applying the baseline selection only on \( h \rightarrow \tau l \tau h \) candidates allows validating the physical object reconstruction (Section 3.2) and the corrections applied to the MC simulations by reproducing the physical objects (kinematic distributions and properties) and the selected set of events when using the same data sample.

In the second step, it is required that the event have, in addition, at least two well identified jets, not necessarily identified as b jets. Then, these two jets are used to form a \( h \rightarrow bb \) candidate. At this stage, to improve the sensitivity, the selected events are separated into several orthogonal categories. After that, the physical backgrounds that require the use of a data-driven modelling are modelled in each category.
In the final step, additional kinematic selections are applied to further constrain the parameter space towards similarity with the signal and the invariant mass of the $H \to hh$ candidate is computed.

### 4.4.1 $h \to \tau_l \tau_h$ selection and reconstruction

To select an $h \to \tau_l \tau_h$ candidate, the $l\tau_h$-pair is selected by following the physical objects reconstruction and identification procedures and recommendations (Section 3.2).

The kinematic constraints on the $p_T$ threshold and on the $\eta$ region applied to electron, muon and $\tau_h$ candidates that will be used for the signal selection were chosen in order to ensure good reconstruction performance and stable HLT trigger operation, away from the boundary conditions. The direction of each selected electron, muon or $\tau_h$ candidate should match the direction of the corresponding HLT trigger object within $|\Delta R| < 0.5$. To pass the signal selection or any of the veto selections, the electrons, muons and hadronic taus should be compatible with the primary hard interaction vertex, requiring $|\Delta z| < 0.2$ and $|d_{xy}| < 0.045$.

#### Signal selection

In the $\tau_e \tau_h$ channel, the electron ($\tau_e$) should satisfy $p_T > 24$ GeV and $|\eta| < 2.1$, be isolated with $I_{PF}^{rel} < 0.1$, pass the tight WP of the BDT-based identification (Table 3.1), and pass the requirements of the discriminator against electrons from photon conversions (Section 3.2.1). The $\tau_h$ should satisfy $p_T > 20$ GeV and $|\eta| < 2.3$, be isolated with $I_\tau < 1.5$ GeV, pass the medium WP of the discriminator against electrons, and the loose WP of the discriminator against muons (Section 3.2.4).

In the $\tau_\mu \tau_h$ channel, the muon ($\tau_\mu$) should satisfy $p_T > 20$ GeV and $|\eta| < 2.1$, pass the tight muon identification, and be isolated with $I_{PF}^{rel} < 0.1$. The $\tau_h$ should satisfy $p_T > 20$ GeV and $|\eta| < 2.3$, be isolated with $I_\tau < 1.5$ GeV, pass the loose WP of the discriminator against electrons, and the tight WP of the discriminator against muons (Section 3.2.4).

One important variable is the transverse mass variable ($M_T$), which is defined by

$$M_T = \sqrt{(p_T + E_T^\perp)^2 - |p_T + E_T^\perp|^2},$$  \hspace{1cm}(4.3)$$

where $p_T$ is the transverse momentum of the $\tau_l$ electron or muon and $E_T^\perp$ is the missing transverse energy computed using the MVA PF $E_T$ algorithm (Section 3.2.5).

$M_T$ has a good discriminating power against $W+$jet and $t\bar{t}$ backgrounds that have higher average $M_T$ values, with respect to the expected $M_T$ values in the $H \to hh \to bb\tau\tau$ signal, as shown in Figure 4.3. For that reason, in both the $\tau_e \tau_h$ and $\tau_\mu \tau_h$ channel selections, it is required that $M_T < 30$.

The two selected signal objects ($e\tau_h$ or $\mu\tau_h$) should have opposite charges and their directions should be separated by $\Delta R > 0.5$ to avoid possible misidentifications.
between the signal objects. If in some event more than one $e\tau_h$ ($\mu\tau_h$) pair passes all previous selection steps, only that pair with the highest scalar sum of $p_T$ is chosen to form the $h \rightarrow \tau\tau_h$ candidate ($h_{\tau\tau_h}$).

**Vetoes**

To reduce the presence of $Z \rightarrow ee$ background in the $e\tau_h$ channel and $Z \rightarrow \mu\mu$ background in the $\mu\tau_h$ channel, a veto is applied, excluding any event where a pair of electrons (muons) with opposite charges have their directions separated by $\Delta R > 0.15$ and pass the selection described below. In the $e\tau_h$ channel, the selection to veto the $Z \rightarrow ee$ requires that an electron has $p_T > 15$ GeV and $|\eta| < 2.5$, passes the cut-based electron identification (Section 3.2.1), and is isolated with $\Gamma_{PF}^{el} < 0.3$. In the $\mu\tau_h$ channel, the selection for the $Z \rightarrow \mu\mu$ veto requires that a muon has $p_T > 15$ GeV and $|\eta| < 2.4$, is identified as a Global Muon, Tracker Muon and PF Muon, and it is isolated with $\Gamma_{PF}^{el} < 0.3$.

To exclude possible contributions from the background processes that have more than one prompt electron or muon in the final state, vetoes are applied for events in both the $e\tau_h$ and $\mu\tau_h$ channels, requiring the absence of any additional electron or muon that passes the selection described below. The selection for the background electron veto requires that an electron satisfy $p_T > 10$ GeV and $|\eta| < 2.5$, passes the loose WP of the BDT-based identification, passes the requirements of the discriminator against photon conversion, and is isolated with $\Gamma_{PF}^{el} < 0.3$. The
selection for the background muon veto requires that a muon satisfy $p_T > 10$ GeV and $|\eta| < 2.4$, passes the tight muon identification WP, and is isolated with $I_{PF}^{\text{rel}} < 0.3$.

$h_{\tau\tau\tau}$ mass reconstruction

One of the discriminating properties of the $h \rightarrow \tau\tau\tau$ signal is the invariant mass of the $h_{\tau\tau\tau}$ candidate ($m_{\tau\tau}$). However, due to the presence of three neutrinos from $\tau$ decays in the final state, the invariant mass of the visible $\tau\tau\tau$ pair ($m_{\text{vis}}$) has a resolution which does not allow well separating the $h \rightarrow \tau\tau\tau$ signal from the background, especially from $Z \rightarrow \tau\tau\tau$, since $m_Z \approx 91.2$ GeV is close to the $h$ mass. To improve the $m_{\tau\tau}$ resolution, the kinematic parameters from the visible $\tau$ candidate decay products are combined with the $E_T$ by using the SVfit algorithm, which is based on the dynamic likelihood method \[125\].

SVfit provides estimates of the invariant mass of $h_{\tau\tau\tau}$ ($m_{\text{sv}}^{\tau\tau}$) on an event-by-event basis. In addition to the observed momentum, the $\tau$ decay is parametrized in the detector coordinate frame by the fraction of the total $\tau$ energy carried by the visible decay products ($X$) and the angle between the original $\tau$ direction and the visible $\tau$-candidate direction. To parametrize $\tau_l$ decay, in addition to the $\tau_h$ parameterization, the invariant mass of the system of two neutrinos ($m_{\nu\nu}$) is used. $m_{\tau\tau}^{\text{sv}}$ is estimated by maximizing the probability of having measured $E_T$, given the observed $\tau_l\tau_h$ momenta ($\vec{y} = (p_{\text{vis}}^{\tau_l}, p_{\text{vis}}^{\tau_h})$) and marginalizing the unknown decay parameters ($\vec{a} = (X_1, \phi_1, m_{\nu\nu}, X_2, \phi_2)$) using the following equation \[125\]:

$$
m_{\tau\tau}^{\text{sv}} = \arg \max_{m_{\tau\tau}} P \left( m_{\tau\tau} | \vec{E}_T, \vec{y} \right)
= \arg \max_{m_{\tau\tau}} \int \delta \left( m_{\tau\tau} - m_{\tau\tau}(\vec{y}, \vec{a}) \right) p \left( \vec{E}_T | \vec{y}, \vec{a} \right) d\vec{a}.
$$

The probability $P \left( m_{\tau\tau} | \vec{E}_T, \vec{y} \right)$ is estimated in $\delta m_{\tau\tau}$ steps in the range $5 \leq m_{\tau\tau} \leq 2000$ GeV, using Markov Chain Monte Carlo (MCMC) to perform numerical integration. The advantage of using MCMC instead of VEGAS-based integration, as in \[125\], is that the MCMC-based method allows reconstructing the full 4-momentum of the $h$-candidate. The probability $p \left( \vec{E}_T | \vec{y}, \vec{a} \right)$ is estimated using the matrix elements for the leptonic and hadronic $\tau$ decays, combined with a likelihood for the $E_T$ calculated assuming a Gaussian resolution of the measurements and using the $E_T$ covariance matrix.

The effect of the SVfit algorithm on the $h \rightarrow \tau\tau$ mass resolution for the $H \rightarrow hh \rightarrow bb\tau\tau$ and $Z \rightarrow \tau\tau$ events is shown in Figure 4.4.
4.4. H candidate reconstruction

Figure 4.4: Distributions renormalized to the unit area of the invariant mass of the $h \rightarrow \tau_l \tau_h$ candidate for the $\tau_e \tau_h$ and $\tau_\mu \tau_h$ channels calculated as a sum of 4-momenta of visible $\tau$ decay products (a, c) and using the SVfit algorithm (b, d) after the full $h \rightarrow \tau_l \tau_h$ candidate selection is applied. After applying the SVfit algorithm, the separation between the $Z \rightarrow \tau\tau$ process (blue) and the signal $H \rightarrow hh \rightarrow bb\tau\tau$ with $m_H = 300$ GeV (red) is significantly improved.

4.4.2 $h \rightarrow bb$ and b jet based categorization

After the $h \rightarrow \tau_l \tau_h$ candidate has been selected, pre-selection criteria to select the jets in the event are determined, in order to choose jet candidates to form the $h \rightarrow bb$ candidate.

The jet pre-selection requires that jets pass the loose working points of the PF and pileup identifications (Section 3.2.3). In addition, to ensure a good performance of the CSV algorithm used for b tagging (Section 3.2.3), the jet must satisfy $p_T > 20$ GeV and $|\eta| < 2.4$. To avoid possible misidentifications between the signal objects, the direction of each selected jet should be separated by $\Delta R > 0.5$ from the directions of the $\tau_l$ and $\tau_h$ that form $h_{\tau_l\tau_h}$ candidate.
For an event to be accepted, it is required that at least two jets pass the pre-selection requirements.

By applying the b tag criteria to the jets, one may control the expected misidentification probability (Table 3.2). For that reason, the b tag based criterion is used to separate the selected events into three orthogonal categories, so that each category has a different background–signal composition, which allows improving the sensitivity of the analysis in the signal extraction. The medium CSV b tag WP (CSVM), which have an expected misidentification probability rate $\sim 1\%$, is applied to each pre-selected jet with the following split into the categories:

- $2\text{jets-0tag}$: selected events with zero jets that pass CSVM criterion;
- $2\text{jets-1tag}$: selected events with only one jet that pass CSVM criterion;
- $2\text{jets-2tag}$: selected events with at least two jets that pass CSVM criterion;
- $2\text{jets-inclusive}$: no b tag requirements on the selected events, so it combines all events from $2\text{jets-0tag}$, $2\text{jets-1tag}$ and $2\text{jets-2tag}$ categories.
- inclusive: only $h \rightarrow \tau_\ell \tau_h$ selection is applied without any requirement on jets.
- $2\text{jets-1tag-loose}$ and $2\text{jets-2tag-loose}$: two supplementary categories that are used in data-driven background modelling; they are defined in the same way as $2\text{jets-0tag}$ and $2\text{jets-1tag}$, but using the loose CSV b tag WP (CSVL) instead of CSVM.

The first three categories are the three orthogonal categories.

Different methods to select the b jet candidates to form the $h \rightarrow bb$ candidate ($h_{bb}$) were studied, the using MC truth in the signal sample $H \rightarrow hh \rightarrow bb\tau_\ell \tau_h$ with $m_H = 300\text{ GeV}$. In the first two methods, the two b jets with the highest CSV values (the CSV-based method) or the highest $p_T$ values (the $p_T$-based method) in the event are chosen to form the $h \rightarrow bb$ candidate ($h_{bb}$). In the third, the $\chi^2$-based method, the b jet pair that results in the smallest $\chi^2$ value of a kinematic fit applied to reconstruct the $m_H$ mass (Section 4.4.4) is selected. The performance of these three methods was studied using the fraction of the signal events that have two pre-selected jets that match two MC truth b jets from the $h \rightarrow bb$ signal within the cone $\Delta R < 0.3$. The fraction of such events out of the total number of events in the $2\text{jets-inclusive}$ category is $\approx 69\%$ for both $\tau_\ell \tau_h$ channels.

The performance of each selection method was quantified by considering the ‘purity’ of the selected b jet pairs, i.e. the ratio between the number of events where both selected b jets match MC truth b jets to the total number of events that have two pre-selected jets that match two the MC truth b jets. The results of the performance studies are reported in Table 4.4. The $2\text{jets-2tag}$ event category shows the best $h_{bb}$ candidate selection performance, but it has a low events statistic, due to the strict requirements on CSV discriminator. On the other hand, the
2jets-0tag event category has a higher events statistic, but a significantly lower sensitivity to the $H \rightarrow hh \rightarrow bb\tau\tau_h$ signal.

The CSV-based method, which shows better overall performance, was chosen to select the $h_{bb}$ candidates. To distinguish the b jets to form an $h_{bb}$ candidate, the jet with the highest CSV value in the event is referred to as the leading b jet, while the jet with the second highest CSV value is referred to as the sub-leading b jet.

<table>
<thead>
<tr>
<th>Channel</th>
<th>Event category</th>
<th>Selection ‘purity’, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CSV-based</td>
</tr>
<tr>
<td>$\tau_e\tau_h$</td>
<td>2jets-inclusive</td>
<td>84.1</td>
</tr>
<tr>
<td></td>
<td>2jets-0tag</td>
<td>72.2</td>
</tr>
<tr>
<td></td>
<td>2jets-1tag</td>
<td>75.9</td>
</tr>
<tr>
<td></td>
<td>2jets-2tag</td>
<td>96.8</td>
</tr>
<tr>
<td>$\tau_\mu\tau_h$</td>
<td>2jets-inclusive</td>
<td>86.5</td>
</tr>
<tr>
<td></td>
<td>2jets-0tag</td>
<td>70.0</td>
</tr>
<tr>
<td></td>
<td>2jets-1tag</td>
<td>81.2</td>
</tr>
<tr>
<td></td>
<td>2jets-2tag</td>
<td>96.9</td>
</tr>
</tbody>
</table>

Table 4.4: Performance of the b jet pair selection methods in the different event categories for $\tau_e\tau_h$ and $\tau_\mu\tau_h$ channels. The selection methods and the selection ‘purity’ are defined in the text.

### 4.4.3 Data driven background modelling

The data driven modelling is used to estimate $Z(\rightarrow \tau\tau_h) +$ jets, $W(\rightarrow l\nu_l) +$ jets ($l = e, \mu, \tau$) and QCD multijets events. The next sections describe the techniques used.

$Z \rightarrow \tau\tau +$ jets

An ‘embedded’ sample is used to model the contribution.

The embedding technique was developed for SM and MSSM $H \rightarrow \tau\tau$ searches [103, 124]. It combines data driven and simulation based approaches, providing good pileup background modelling in the events together with well-modelled kinematics of $Z \rightarrow \tau\tau$ decays. In the ‘embedded’ samples, events from $Z \rightarrow \mu\mu$ data are used, exploiting the good performance of CMS in muon reconstruction and identification (Section 3.2.2). In each event, two reconstructed muons from Z decay are replaced by two MC generated $\tau$ with the same momentum at the particle-flow level of the reconstruction (Section 3.1.3). After that, the full event reconstruction sequence is repeated.

An additional statistical weight ($w^{emb}$) is applied to each embedded event to take into account a possible bias introduced due to $Z \rightarrow \mu\mu$ selection in data. $w^{emb}$ is defined using $Z \rightarrow \mu\mu$ MC simulations as the ratio of events that pass the applied
Chapter 4. $H \to hh$ candidate selection and reconstruction

$Z \to \mu\mu$ selection to the total number of events in the sample and depends on $p_T$ and $\eta$ of each muon.

The embedded $Z \to \tau\tau$ event selection should be compensated for a contamination from the $t\bar{t}$ events in the original data selection, in particular in the $2jets-2tag$ category, where two well identified b jets are present, as shown in Figure 4.3. For that reason, the same embedding technique was applied to the MC-modelled $t\bar{t}$ events, requiring the same selection that has been used to select $Z \to \mu\mu$ data. Then, in each event category, the average contribution from the $t\bar{t}$-embedded events is subtracted from the average contribution of the $Z \to \tau\tau$ embedded events to model the shapes of the distributions of the event parameters in $Z \to \tau\tau$ events, or simply ‘$Z \to \tau\tau$ shape’.

To normalize the contribution from the $Z \to \tau\tau$ events, the yield from MC-modelled $Z \to \tau\tau$ events in the inclusive category is normalized to the NNLO cross-section (Table 4.3) and multiplied by the efficiency to pass the category selection for the events in the embedded sample ($\varepsilon_{\text{cat}}(Z \to \tau\tau)$):

$$\varepsilon_{\text{cat}}(Z \to \tau\tau) = \frac{N_{\text{emb}}^{\text{cat}}(Z \to \tau\tau)}{N_{\text{inc}}^{\text{cat}}(Z \to \tau\tau)}, \quad (4.5)$$

where $\text{cat}$ is the category where the $Z \to \tau\tau$ contribution should be estimated, and $N_{\text{emb}}^{\text{cat}}(Z \to \tau\tau)$, respectively, $N_{\text{inc}}^{\text{cat}}(Z \to \tau\tau)$, are the yields of the $Z \to \tau\tau$ embedded events that pass the $\text{cat}$, respectively, inclusive category selections.

**W+jets**

The W+jets yield is estimated using high-$M_T$ data regions, were W+jets is the dominating process, while the expected contribution from the signal is negligible (Figure 4.3). Due to the low events statistic and marginal increase of the presence of $t\bar{t}$, the W+jets estimation in $2jets-2tag$ category is slightly different than in the other categories.

For all other event categories, the high-$M_T$ sideband event region is defined by applying the same category selection, except the cut on $M_T$, which is replaced by $M_T > 70\text{GeV}$. The contribution from all other backgrounds (except QCD) in the sideband region is subtracted using the MC modelled predictions. The extrapolation factor to scale the yield obtained in the high-$M_T$ region to the yield in the signal region ($k_{W}^{\text{cat}}$) is estimated using simulated MC W+jets events:

$$k_{W}^{\text{cat}} = \frac{N_{\text{MC}}^{\text{cat}}(W + \text{jets})}{N_{\text{high-}M_T-\text{cat}}^{\text{MC}}(W + \text{jets})}, \quad (4.6)$$

where $\text{cat}$ is the category where W+jets contribution should be estimated, and $N_{\text{cat}}^{\text{MC}}(W + \text{jets})$, respectively, $N_{\text{high-}M_T-\text{cat}}^{\text{MC}}(W + \text{jets})$, are the yields of the W+jets MC events that pass the $\text{cat}$ category selection, respectively, the $\text{cat}$ high-$M_T$ sideband region selection.
4.4. $H$ candidate reconstruction

In $2\text{jets-2}\text{tag}$, the high-$M_T$ sideband region is defined by $60 < M_T < 120\text{GeV}$. To decrease the statistical uncertainties, $k_{W+\text{jets-2}\text{tag-loose}}$ has been used to extrapolate from the sideband to the signal region.

The ‘$W$+jets shape’ is modelled using the MC simulated $W$+jet events that pass the category selection, except for $2\text{jets-1}\text{tag}$ and $2\text{jets-2}\text{tag}$, where the shape is modelled using, respectively, $2\text{jets-1}\text{tag-loose}$ and $2\text{jets-2}\text{tag-loose}$ category selections, in order to avoid irregularities in the distribution shape due to statistical fluctuations.

Figure 4.5: Contamination of $Z \rightarrow \tau \tau$ embedded sample (blue) by the contribution from the $t\bar{t}$ process (red) as a function of $m_{\tau\tau}$ in the $2\text{jets-1}\text{tag}$ (a) and $2\text{jets-2}\text{tag}$ (b) event categories of the $\tau_e\tau_h$ channel and in the $2\text{jets-1}\text{tag}$ (c) and $2\text{jets-2}\text{tag}$ (d) event categories of the $\tau_\mu\tau_h$ channel.
The set of QCD multi-jet processes has a high production cross-section in p–p collisions, but their contribution is significantly reduced by requiring the presence of well identified and isolated electron or muon in the event. This discriminating property of the isolation criterion is used to define the sideband regions for the QCD modelling.

It is also assumed that the properties of the parameter space of the QCD events are almost insensitive to the sign compatibility requirement for the $\tau_l\tau_h$ candidate pair in the $h \rightarrow \tau\tau$ selection, since the correlation between the parameters of the jets that are misidentified as the signal electron (muon) and $\tau_h$ candidates should be minimal. Therefore, it is assumed that only the amount of the QCD contribution may change between the selection made by requiring that the $\tau_l$ and $\tau_h$ have opposite signs (OS) and the selection made by requiring that the $\tau_l$ and $\tau_h$ have the same sign (SS).

The QCD OS/SS normalization factor ($f_{OS/SS}$) is defined as the ratio between QCD contributions in the OS and SS selections. The value of $f_{OS/SS}$ was measured for the SM and MSSM $H \rightarrow \tau\tau$ searches \cite{103,124} by inverting the isolation requirements on the $\tau_l$ and was found to be compatible with $1.06 \pm 0.05$ in all event categories used in the analyses. We use the same $f_{OS/SS}$ value, relaying on these measurements, because the parameter space of the $H \rightarrow hh \rightarrow bb\tau_l\tau_h$ analysis is close to those where these measurements were performed and because the events statistic in the final selection of this analysis is rather low for estimating $f_{OS/SS}$ with good accuracy.

To estimate the QCD contribution in all event categories, except $2jets-2tag$, SS sideband regions were defined, using the same event category selection, but requiring that the $\tau_l$ and $\tau_h$ candidates have the same sign. The QCD yield in the SS regions ($N_{cat}(QCD)$) is estimated by subtracting the contributions from all the other backgrounds using the data driven technique for W+jet (similar to the one described in the previous section, but using SS requirement) and MC modelling for the remaining backgrounds. The distribution of $m_{\tau\tau}$ for the data in the SS sideband region of the $2jets$-inclusive category is shown in Figure 4.6 with the contributions from all modelled backgrounds superimposed (except QCD). The final estimation of the QCD yield in the signal region ($N_{cat}(QCD)$) is obtained by applying the extrapolation factor: $N_{cat}(QCD) = N_{cat}^{SS}(QCD) \cdot f_{OS/SS}$.

Due to the low events statistic in $2jets-2tag$, in order to avoid big statistical uncertainties, the QCD yield in the SS sideband of the $2jets$-inclusive category is extrapolated using the anti-isolated SS sideband regions (SS-$\overline{iso}$) defined below, where the events statistic is higher, to estimate the category selection efficiency. For a given event category, the SS-$\overline{iso}$ sideband region is defined using the selection as for the SS sideband region, replacing the isolation requirements for the $\tau_l$ by $0.2 < I_{PF}^\tau < 0.5$. 

QCD
4.4. H candidate reconstruction

The last step of the event reconstruction is the reconstruction of the $H \rightarrow hh$ candidate from the $h \rightarrow \tau_\tau h$ ($h_{\tau\tau}$) and $h \rightarrow bb$ ($h_{bb}$) candidates.

Before reconstructing the $H$ mass, the events that have signatures that are significantly different from the hypothesis $m_h = 125\text{ GeV}$ are rejected by applying requirements to the invariant masses of the $h_{\tau\tau}$ and $h_{bb}$ candidates. To fix the cut values, the distributions for of the invariant mass of the $h_{\tau\tau}$ candidate reconstructed by the SVfit algorithm ($m_{\tau\tau}^{sv}$) and the invariant mass of the $h_{bb}$ candidate...

Figure 4.6: Background composition modelled in the SS sideband region of the of the 2jets-inclusive category for $\tau_\tau h$ (a) and $\tau_\mu h$ (b) channels as a function of $m_{\tau\tau}^{sv}$. All unaccounted contributions from data in the region is considered to originate from the QCD multijet processes. The ‘Electroweak’ in the legends (red) includes W+jets, diboson and single top processes; the $Z \rightarrow ll$ (blue) includes $Z \rightarrow ee$ and $Z \rightarrow \mu\mu$ processes.

Therefore, the QCD yield in the 2jets-2tag signal region is estimated by

$$N_{\text{2jets-2tag}}(QCD) = N_{\text{2jets-inclusive}}^{SS}(QCD) \cdot \frac{N_{\text{2jets-2tag}}^{SS-iso}(QCD)}{N_{\text{2jets-inclusive}}^{SS-iso}(QCD)} \cdot f_{OS/SS}. \ (4.7)$$

The modelling of the ‘QCD shape’ depends on the category. For the inclusive and 2jets-inclusive categories, the ‘QCD shape’ is modelled using the same technique that is used to estimate the QCD yield in these categories. In the other categories, to reduce the statistical uncertainties, the SS-iso sideband regions have been used for the shape estimations without removing the contributions from the other backgrounds. Moreover, since in 2jets-1tag and 2jets-2tag the events statistic is very low, the ‘QCD shape’ is estimated using the 2jets-1tag-loose, respectively, 2jets-1tag-loose event categories.

4.4.4 $H \rightarrow hh$

The last step of the event reconstruction is the reconstruction of the $H \rightarrow hh$ candidate from the $h \rightarrow \tau_\tau h$ ($h_{\tau\tau}$) and $h \rightarrow bb$ ($h_{bb}$) candidates.

Before reconstructing the $H$ mass, the events that have signatures that are significantly different from the hypothesis $m_h = 125\text{ GeV}$ are rejected by applying requirements to the invariant masses of the $h_{\tau\tau}$ and $h_{bb}$ candidates. To fix the cut values, the distributions for of the invariant mass of the $h_{\tau\tau}$ candidate reconstructed by the SVfit algorithm ($m_{\tau\tau}^{sv}$) and the invariant mass of the $h_{bb}$ candidate...
Chapter 4. $H \rightarrow hh$ candidate selection and reconstruction

![Figure 4.7](image)

Figure 4.7: Distributions of the $h$ candidates masses in modelled background and data events in the 2jets-2tag category: $m_{\tau\tau}$ (a) and $m_{\tau\tau}$ (b) for the $\tau\tau\tau\tau$ channel, $m_{\tau\tau}$ (c) and $m_{\tau\tau}$ (d) for the $\tau\tau\tau\tau$ channel. Contribution from the modelled signal distribution is superimposed on the backgrounds and multiplied by 30, assuming an MSSM $H$ with $m_H = 300$ GeV and $\tan\beta = 2$. The ‘Electroweak’ in the legends (red) includes $W+$jets, di-boson and single top processes; the $Z \rightarrow ll$ (blue) includes $Z \rightarrow ee$ and $Z \rightarrow \mu\mu$ processes.

($m_{\tau\tau}$) for the modelled signal and the background events are considered. Here, $m_{\tau\tau}$ is defined as $\|p_{b1} + p_{b2}\|$, where $p_{b1}$ and $p_{b2}$ are the 4-momenta of the leading, respectively, the sub-leading, $b$ jet candidates. An example of the distributions of $m_{\tau\tau}$ and $m_{\tau\tau}$ in the most sensitive 2jets-2tag category is shown in Figure 4.7.

The chosen requirements for the $h$ candidates’ masses are:

$$90 < m_{\tau\tau} < 150 \text{ GeV},$$
$$70 < m_{\tau\tau} < 150 \text{ GeV},$$

(4.8)

Two approaches are considered to reconstruct the mass of the $H$ candidate ($m_H$) from different estimates of the 4-momentum of the $H$ ($p_H, m_H = \|p_H\|$).
In the first approach, $p_H$ is estimated as the sum of the 4-momenta of all the reconstructed signal objects ($p_{vis}$) plus the $E_T$ 4-momentum ($p_{mis} = (E_{vis,x}, E_{vis,y}, 0, 0)$):

$$p_H^{bb\tau\tau} = p_{vis} + p_{mis}.$$  

(4.9)

Furthermore, $p_{vis}$ is defined as

$$p_{vis} = p_{\tau 1} + p_{\tau 2} + p_{b 1} + p_{b 2},$$

(4.10)

where $p_{\tau 1}$ and $p_{\tau 2}$ are the 4-momenta of the $\tau_l$ and $\tau_h$ candidates, and $p_{b 1}$ and $p_{b 2}$ are the 4-momenta of the two b jet candidates.

This approach yields a robust estimator of the signal, the $H$ candidate mass ($m_{H}^{bb\tau\tau}$), however it has relatively low resolution, mostly due to the MET, as is shown in Figures 4.8 and 4.9 (left plots), where $m_{H}^{bb\tau\tau}$ is reported for $bb\tau_l\tau_h$ and $bb\tau_{l\mu}\tau_h$.

In the second approach, the $m_{H}$ resolution is improved by applying a kinematic fit to the reconstructed signal objects and $E_T$, constraining the masses of the $h_{\tau\tau}$ and $h_{bb}$ candidates to be 125 GeV, exploiting the fact that the expected natural decay width of $h$ is on the order of a few MeV, so that the differences between the measured $h$ candidates’ masses and 125 GeV are assumed to be entirely due to the detector resolution effects. Furthermore, two approximations are applied to simplify the fit procedure. First, since the measurements of the $\eta$ and $\phi$ directions of the b jets are usually more precise than the measured b jet energy, to fit the b jet pair 4-momentum to the $m_h = 125$ GeV constraint, only the energies of the b jets are scaled, keeping fixed the $\eta$ and $\phi$ directions. Second, because the taus in $h \rightarrow \tau\tau$ decays have, on average, an high $p_T$, the collinear approximation between the visible and invisible $\tau$ decay products is applied for both the $\tau_l$ and the $\tau_h$ candidates.

Under such assumptions and constraints, in the kinematic parameter space that defines the $m_{H}$ momentum, only two independent parameters are left free to float in the fit. The two independent free parameters chosen are the energy of the $\tau_l$ candidate ($E_{\tau 1}$) and the energy of the leading b jet ($E_{b 1}$).

The $h_{bb}$ candidate is fit independently, with the quality of the fit defined by the $\chi^2$ of the fit:

$$\chi^2_{bb} = \sum_{i=1}^{2} \frac{E_{bi}^{fit} - E_{bi}^{meas}}{\sigma_{bi}},$$

(4.11)

where, for each b jet candidate, $\sigma_{bi}$ is the measured energy resolution, $E_{bi}^{fit}$ and $E_{bi}^{meas}$ are respectively the fitted and the measured energies of the b jet.

The $h_{\tau\tau}$ candidate is fit together with the recoil ($\vec{p}_{recoil}$), defined as the momentum in the transverse plane that is opposite to the $H$ candidate momentum, so that the measured recoil is defined as

$$\vec{p}_{recoil} = (-p_{vis,x}, -p_{vis,y}).$$

(4.12)
Chapter 4. $H \rightarrow hh$ candidate selection and reconstruction

The quality of the $h_{\tau\tau}$ fit is quantified by $\chi^2_{\text{recoil}}$, using the covariance matrix of the recoil measurement ($\text{COV}_{\text{recoil}}$), which combines the covariance matrices for $E_T$ and the b jets. $\chi^2_{\text{recoil}}$ is given by

$$\chi^2_{\text{recoil}} = (\vec{p}^\text{fit}_{\text{recoil}} - \vec{p}^\text{meas}_{\text{recoil}}) \cdot \text{COV}_{\text{recoil}}^{-1} (\vec{p}^\text{fit}_{\text{recoil}} - \vec{p}^\text{meas}_{\text{recoil}}),$$

(4.13)

where $\vec{p}^\text{fit}_{\text{recoil}}$ is the recoil value fit result.

The overall fit quality is characterized by the sum of $\chi^2_{bb}$ and $\chi^2_{\text{recoil}}$:

$$\chi^2_{\text{kinfit}} \equiv \chi^2_{bb} + \chi^2_{\text{recoil}}.$$  

(4.14)

The kinematic fit procedure consists of an iterative minimization of the $\chi^2_{\text{kinfit}}$ using Newton’s method and is implemented in the HHKinFit package [126].

In some cases, when the momenta combination of the signal objects is significantly incompatible with the $m_h = 125$ GeV hypothesis, the kinematic fit may not converge. However, for the $H \rightarrow hh \rightarrow bbl\ell_h$ signal events, such cases are very rare, since after applying the requirements on the masses of $h_{\tau\tau}$ and $h_{bb}$ (Equation (4.8)), the fraction of signal events where the kinematic fit does not converge is below 1%. Therefore, to ensure that the kinematic parameter space of the selected events are close to that for the signal, the events for which the kinematic fit does not converge are rejected for both the data and the simulations.

The mass of the $H$ candidate estimated by the kinematic fit ($m_H^{\text{kinfit}}$) is defined from the fitted signal 4-momenta ($\vec{p}_{\text{obj}}^\text{fit}$) that minimizes $\chi^2_{\text{kinfit}}$:

$$m_H^{\text{kinfit}} = \left\| \vec{p}_{\tau 1}^\text{fit} + \vec{p}_{\tau 2}^\text{fit} + \vec{p}_{b 1}^\text{fit} + \vec{p}_{b 2}^\text{fit} \right\|.$$ 

(4.15)

The $m_H^{\text{kinfit}}$ has a better resolution than the $m_H^{bb\tau\tau}$, as shown in Figures 4.8 and 4.9, and, therefore has a better discriminating power between signal and background. For that reason, the distribution of $m_H^{\text{kinfit}}$ in the data and in the modelled composition of the data sample have been chosen to be used in the final model fit procedure, described in the next chapter, for the signal extraction.
Figure 4.8: Distributions of $m_{H^+\tau\tau}$ (a, c and e) and $m_{H^\text{kinfit}}$ (b, d and f) for modelled contributions and data events in the 2jets-0tag, 2jets-1tag and 2jets-2tag categories for the $\tau\tau\tau\tau$ channel. Contribution from the modelled signal distribution is superimposed on the backgrounds. The signal yield is multiplied by 10, assuming an MSSM $H$ with $m_H = 300$ GeV and $\tan\beta = 2$. The ‘Electroweak’ in the legends (red) includes W+jets, di-boson and single top processes; the $Z \to ll$ (blue) includes $Z \to ee$ and $Z \to \mu\mu$ processes.
Figure 4.9: Distributions of $m_H^{\ell\ell}$ (a, c and e) and $m_H^{\text{kinfit}}$ (b, d and f) for modelled contributions and data events in the 2jets-0tag, 2jets-1tag and 2jets-2tag categories for the $\tau\tau\tau\tau$ channel. Contribution from the modelled signal distribution is superimposed on the backgrounds. The signal yield is multiplied by 10, assuming an MSSM $H$ with $m_H = 300\,\text{GeV}$ and $\tan\beta = 2$. The ‘Electroweak’ in the legends (red) includes W+jets, di-boson and single top processes; the $Z \rightarrow ll$ (blue) includes $Z \rightarrow ee$ and $Z \rightarrow \mu\mu$ processes.
Chapter 5

Model fit procedure

In this step of the analysis, the modelled contributions from all backgrounds and possible contribution from the signal are fitted to the data using the statistical method adopted by the CMS. To provide reliable fit results, all systematic and statistical uncertainties in the modelling should be taken into account. The applied fit procedure will result in the exclusion limits with the chosen level of confidence for the $gg \rightarrow H \rightarrow hh \rightarrow bb\tau\tau$ signal production cross-section times branching ratio as a function of $m_H$, which will also be interpreted in the parameter space of the MSSM and 2HDM models.

5.1 Systematic uncertainties

The systematic uncertainties in the modelling originate from many different sources. We divide them into two groups: normalization and shape uncertainties. Uncertainties that with a good approximation influence only the total yield of the modelled contribution are referred to as normalization uncertainties. If an uncertainty also influences the shape of the parameter distribution chosen for the fit ($m_H^{\text{kinfit}}$), it is called a shape uncertainty.

5.1.1 Normalization uncertainties

Luminosity and cross-section

To set a limit on the signal production cross-section, the modelled $H \rightarrow hh \rightarrow bb\tau\tau$ event yield should be rescaled by the luminosity, according to Equation (2.1). Therefore, the signal normalization uncertainty equals the uncertainty in the luminosity measurements (Section 2.2.5), which is 2.6%.

The simulated processes that are normalized using their cross-sections, are also affected by uncertainties in the value of their cross-section arising from measurements or theoretical predictions. The combined contribution from the luminosity
Chapter 5. Model fit procedure

and cross-section uncertainties to the normalization uncertainty in the event yields is 3% for event yields from \( Z \rightarrow ll \) (\( l = e, \mu, \tau \)) plus jets, 10% for \( t \bar{t} \) plus jets and 15% for di-boson plus jets and single top.

Isolation, identification and trigger efficiencies

The identification, isolation and trigger efficiencies of the signal electrons, muons and \( \tau_h \) were measured for the data and simulations, as discussed in Sections 3.2.1, 3.2.2, 3.2.4 and 4.2.

According to these measurements, some statistical weights were applied for each simulated event to take into account the discrepancies between the simulations and the data in the identification (\( w_{id} \)), isolation (\( w_{iso} \)) and trigger (\( w_{trig} \)) performances, where \( l = e, \mu, \tau_h \). The corresponding uncertainties in the modelled identification (\( u_{id} \)), isolation (\( u_{iso} \)) and trigger (\( u_{trig} \)) efficiencies are assumed to be independent and are joined in the combined identification-isolation-trigger uncertainty (\( u_{id-isol-trig} \)) as

\[
u_{id-isol-trig}^l = \sqrt{(u_{id}^l)^2 + (u_{iso}^l)^2 + (u_{id-isol-trig}^l)^2}, \tag{5.1}
\]

where \( l \) is \( \tau_e \) or \( \tau_{\mu} \) or \( \tau_h \).

The \( u_{id-isol-trig} \) for \( \tau_e \) and \( \tau_{\mu} \) is 2% and 1%, respectively, and it affects all the processes for which an MC-based simulation has been used, namely \( H \rightarrow hh \rightarrow bb\tau\tau \), \( Z \rightarrow ll \) (\( l = e, \mu, \tau \)), \( t \bar{t} \), di-boson and single top.

The measured uncertainty of the \( \tau_h \) identification and isolation efficiencies of 6% is combined with the uncertainty in the \( \tau_h \) trigger efficiency of 3%, giving a \( u_{\tau_h}^{id-isol-trig} \approx 8\% \) (Equation (5.1)). \( u_{\tau_h}^{id-isol-trig} \) is applied to those MC samples where \( \tau_h \) are present with a high probability: \( H \rightarrow hh \rightarrow bb\tau\tau \), \( Z \rightarrow \tau\tau \), \( t \bar{t} \), di-boson and single top.

b jet tagging

The b jet identification efficiency and misidentification probability were measured for data and simulations, as discussed in Section 3.2.3.

According to these measurements, to take into account the discrepancies between the simulations and the data, the number of jets that pass the CSVM working point in the simulated events was re-evaluated. In each simulated event, jets that do not originally pass the CSVM criterion are ‘promoted’ to pass the CSVM with the probability \( P_{\text{promote}}(\text{jet}) \), while jets that originally passed the CSVM criterion are ‘demoted’ from CSVM selection with probability \( P_{\text{demote}}(\text{jet}) \).

\(^1\)From this section, the processes \( Z \rightarrow ll \) plus jets, \( t \bar{t} \) plus jets and di-boson plus jets are, for simplicity, denoted by \( Z \rightarrow ll \), \( t \bar{t} \) and di-boson, respectively.
5.1. Systematic uncertainties

These probabilities are

\[ P_{\text{promote}}(\text{jet}) = (k_{\text{btag}} - 1) \cdot H(k_{\text{btag}} - 1) \frac{\varepsilon^{\text{MC}}_{\text{btag}} - 1}{k_{\text{btag}}}, \]
\[ P_{\text{demote}}(\text{jet}) = (1 - k_{\text{btag}}) \cdot H(1 - k_{\text{btag}}), \]

(5.2)

where \( H(n) \) is the Heaviside step function, \( \varepsilon^{\text{MC}}_{\text{btag}} \) is the b jet identification efficiency measured in the simulations, and \( k_{\text{btag}} \) is the scale factor that is estimated to correct for disagreements in the b tagging performance between the data and the simulations as a function of the jet’s \( p_T \) and \( \eta \).

To estimate how the b tagging uncertainties affecting the yields of the analysis, the b jet identification efficiency and misidentification probability were independently modified by \( \pm 1\sigma \), according to the measurements of their uncertainties, which are discussed in Section 3.2.3. Then, the whole event selection of the analysis was repeated four additional times, using the modified values of \( P_{\text{promote}}(\text{jet}) \) and \( P_{\text{demote}}(\text{jet}) \). The obtained systematic uncertainty in the normalization is 0–72% due to the uncertainties in the simulations of the b jet identification efficiency and in the range 0–5% due to the uncertainties in the simulations of the b jet misidentification probability, depending on the process sample and on the event category.

These uncertainties mostly affect the MC generated samples. However, due to the contamination of the W plus jet sideband region by \( t\bar{t} \) events, especially in the 2jets-2tag event category, the modelled W plus jet yield is also affected by b jet identification and misidentification uncertainties.

The \( E_T \) energy scale

The performance of the PF MVA \( E_T \) reconstruction in \( Z \rightarrow \mu\mu \) events was measured and recoil corrections were applied to improve the agreement between the data and the simulations, as described in Section 3.2.5. The \( E_T \) uncertainty affects all MC modelled processes, which are \( H \rightarrow hh \rightarrow bb\tau\tau, Z \rightarrow ll \ (l = e, \mu), t\bar{t}, \) di-boson and single top.

The changes in the final yields of these processes are estimated by varying the \( h_{\tau\tau} \) candidate recoil within its uncertainties and repeating the analysis steps that are influenced by this variation. The effect of the \( E_T \) energy scale uncertainty on the normalization is within the range 0–7%, depending on the process and on the event category.

\(^2\)If not specified, by ‘event category’ we indicate one of the three orthogonal categories defined in Section 4.4.2: 2jets-0tag, 2jets-1tag, 2jets-2tag, depending on the number of jets tagged as b jets.
\( \tau_h \) misidentification

The uncertainties in the \( \tau_h \) misidentification probabilities are taken from the measurements that were discussed in Section 3.2.4. The uncertainty in the probability of misidentifying an electron or a muon as a \( \tau_h \) are respectively 20% and 30%. The probability of misidentifying a quark or gluon jet as a \( \tau_h \) has an uncertainty of 20%. For electron and muon misidentification, this uncertainty is 20% and 30%, respectively.

The \( \tau_h \) misidentification uncertainty mainly affects the \( Z \to ee, Z \to \mu\mu \) and part of the \( Z \to \tau\tau \) processes, where the selected \( \tau_h \) is a misidentified electron, muon or jet. The information about how the fake \( \tau_h \) was misidentified is accessible through the MC truth. Based on this information, the selected \( Z \to ll \) events are split into the three groups: ZTT, ZL and ZJ. The ZL includes the \( Z \to ee \) and \( Z \to \mu\mu \) events where \( \tau_h \) is faked by an electron or muon. The ZJ includes the \( Z \to ll \) \( (l = e, \mu, \tau) \) events where the \( \tau_h \) is faked by a quark or a gluon jet. The ZTT includes the remaining \( Z \to \tau\tau \) events, which were not included in the ZJ. This allows a separate estimate of the systematics for each group, depending on the type of misidentification.

The resulting ZL and ZJ normalization uncertainties are in the range 20–90%, depending on the channel \((\tau_e \tau_h, \tau_\mu \tau_h)\) and on the event category.

Normalization extrapolation for data driven modelling

For the processes modelled using a data driven technique, the statistical errors in the extrapolation factor combined with the statistical error in the yield of the sideband region is considered a source of the systematic uncertainty for normalization Section 4.4.3.

The resulting systematic uncertainty of the \( Z \to \tau\tau \) normalization from the inclusive category (see Section 4.4.2) is in the range 5–6%, depending on the event category \( (2jets-0tag, 2jets-1tag, 2jets-2tag) \). Another contribution to the systematic uncertainty of the \( Z \to \tau\tau \) normalization arises from \( t\bar{t} \) contamination in the embedded sample, resulting in an additional systematic uncertainty in the range 0–32%, depending on the channel \((\tau_e \tau_h, \tau_\mu \tau_h)\) and on the event category.

The statistical error in the extrapolation uncertainty of the W+jets normalization is 10% in the \( 2jets-0tag \) category but increases significantly, to 100%, in the \( 2jets-2tag \) category, where the W+jets yield is low. The additional uncertainty in the \( 2jets-1tag \) category is due to the presence of \( t\bar{t} \) in the sideband region, where the W+jets are estimated. The corresponding normalization uncertainty of 25% is measured by extrapolating the uncertainty of 10% in the \( t\bar{t} \) normalization.

The extrapolation factor uncertainty of the QCD normalization is composed of the statistical error in the sideband region and the uncertainty of 10% in the \( f_{OS/SS} \) extrapolation factor (Section 4.4.3), resulting in the overall systematic uncertainty in the range 20–100%, depending on the channel and on the event category.
5.1. Systematic uncertainties

5.1.2 Shape uncertainties

The energy scales of the $\tau_h$ and the jets have a direct influence on the $m_H$ reconstruction (Section 4.4.4), and are considered to be a source of systematic uncertainties, which affecting the shape of the $m_H$, as well as the normalization. Therefore, the corresponding uncertainties are applied as shape uncertainties, which are defined on a bin-by-bin basis. The uncertainties in the simulations of the energies of the $\tau_h$ and jets, 3% and 3-5%, respectively, were determined in the measurements discussed in Sections 3.2.3 and 3.2.4.

Their impact on the final $m_H^{\text{kinfit}}$ distribution in each event category is estimated by independently varying the energy scales of $\tau_h$ and the jets within their uncertainties and repeating the full analysis selection and reconstruction procedure.

The jet energy scale is considered for all the MC modelled processes, while the $\tau_h$ energy scale has a considerable effect on the $H \rightarrow hh \rightarrow bbt\tau$ and $Z \rightarrow \tau\tau$ processes.

An example of how the $m_H^{\text{kinfit}}$ distribution changes, depending on varying the energy scales of $\tau_h$ and the jet within their uncertainties, for the $\tau_\mu\tau_h$ channel in the 2jets-2btag category, is shown in Figure 5.1.

![Figure 5.1](image_url)

Figure 5.1: An example of the impact of the uncertainties in the energy scales of $\tau_h$ and jets on the distribution of $m_H^{\text{kinfit}}$ for the $\tau_\mu\tau_h$ channel in the 2jets-2btag category. (a) $m_H^{\text{kinfit}}$ distributions for the $H \rightarrow hh \rightarrow bbt\tau$ process with $\tau$ energy scaled by $\pm1\sigma$. (b) $m_H^{\text{kinfit}}$ distributions for the $t\bar{t}$ process with jet energy scaled by $\pm1\sigma$.

5.1.3 Summary

A summary of all the sources of the systematic uncertainties and their values considered in this analysis is reported for the $\tau_e\tau_h$ and $\tau_\mu\tau_h$ channels in Tables 5.1 and 5.2.
### Table 5.1: Summary of the systematic uncertainties used in the $\tau_e\tau_h$ channel.

<table>
<thead>
<tr>
<th>Source</th>
<th>Uncertainty value [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H</td>
</tr>
<tr>
<td>$\tau_e\tau_h$ $2jets-0tag$, $2jets-1tag$, $2jets-2tag$</td>
<td></td>
</tr>
<tr>
<td>Luminosity and cross-section</td>
<td>2.6</td>
</tr>
<tr>
<td>$\tau_h$ energy scale</td>
<td>shape</td>
</tr>
<tr>
<td>Jet energy scale</td>
<td>shape</td>
</tr>
<tr>
<td>Electron trigger, identification and isolation efficiencies</td>
<td>2</td>
</tr>
<tr>
<td>$\tau_h$ trigger, identification and isolation efficiencies</td>
<td>8</td>
</tr>
<tr>
<td>$E_T$ energy scale</td>
<td>1</td>
</tr>
<tr>
<td>b jet identification efficiency</td>
<td>7</td>
</tr>
<tr>
<td>Lepton misidentified as $\tau_h$ probability</td>
<td>-</td>
</tr>
<tr>
<td>Jet misidentified as $\tau_h$ probability</td>
<td>-</td>
</tr>
<tr>
<td>Extrapolation to the signal region</td>
<td>-</td>
</tr>
<tr>
<td>$E_T$ energy scale</td>
<td>1</td>
</tr>
<tr>
<td>b jet identification efficiency</td>
<td>-</td>
</tr>
<tr>
<td>b jet misidentified probability</td>
<td>-</td>
</tr>
<tr>
<td>Lepton misidentified as $\tau_h$ probability</td>
<td>-</td>
</tr>
<tr>
<td>Jet misidentified as $\tau_h$ probability</td>
<td>-</td>
</tr>
<tr>
<td>$tt$ contamination</td>
<td>-</td>
</tr>
<tr>
<td>Extrapolation to the signal region</td>
<td>-</td>
</tr>
<tr>
<td>$E_T$ energy scale</td>
<td>1</td>
</tr>
<tr>
<td>b jet identification efficiency</td>
<td>4</td>
</tr>
<tr>
<td>b jet misidentified probability</td>
<td>-</td>
</tr>
<tr>
<td>Lepton misidentified as $\tau_h$ probability</td>
<td>-</td>
</tr>
<tr>
<td>Jet misidentified as $\tau_h$ probability</td>
<td>-</td>
</tr>
<tr>
<td>$tt$ contamination</td>
<td>-</td>
</tr>
<tr>
<td>Extrapolation to the signal region</td>
<td>-</td>
</tr>
</tbody>
</table>

Process labels: H stands for $H \rightarrow hh \rightarrow b\bar{b}\tau\tau$ signal; $Z \rightarrow ll$ events are split into three groups, ZTT, ZL and ZJ, according to the definitions provided in the text; VV+t uncertainties are applied to the total event yield of di-boson and single top processes; W stands for W+jets.
### Table 5.2: Summary of the systematic uncertainties used in the $\tau_\mu \tau_h$ channel. Process labels: H stands for $H \rightarrow hh \rightarrow bb\tau\tau$ signal; Z → ll events are split into three groups, ZTT, ZL and ZJ, according to the definitions provided in the text; VV+t uncertainties are applied to the total event yield of di-boson and single top processes; W stands for W+jets.

<table>
<thead>
<tr>
<th>Uncertainty source</th>
<th>Uncertainty value, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H</td>
</tr>
<tr>
<td><strong>$\tau_\mu \tau_h$, 2jets-0tag, 2jets-1tag, 2jets-2tag</strong></td>
<td></td>
</tr>
<tr>
<td>Luminosity and cross-section</td>
<td>2.6</td>
</tr>
<tr>
<td>$\tau_h$ energy scale</td>
<td>shape</td>
</tr>
<tr>
<td>Jet energy scale</td>
<td>shape</td>
</tr>
<tr>
<td>Muon trigger, identification and isolation efficiencies</td>
<td>1</td>
</tr>
<tr>
<td>$\tau_h$ trigger, identification and isolation efficiencies</td>
<td>8</td>
</tr>
<tr>
<td><strong>$\tau_\mu \tau_h$, 2jets-1tag</strong></td>
<td></td>
</tr>
<tr>
<td>$E_T$ energy scale</td>
<td>1</td>
</tr>
<tr>
<td>b jet identification efficiency</td>
<td>8</td>
</tr>
<tr>
<td>b jet misidentified probability</td>
<td>-</td>
</tr>
<tr>
<td>Lepton misidentified as $\tau_h$ probability</td>
<td>-</td>
</tr>
<tr>
<td>Jet misidentified as $\tau_h$ probability</td>
<td>-</td>
</tr>
<tr>
<td>Extrapolation to the signal region</td>
<td>-</td>
</tr>
<tr>
<td><strong>$\tau_\mu \tau_h$, 2jets-2tag</strong></td>
<td></td>
</tr>
<tr>
<td>$E_T$ energy scale</td>
<td>1</td>
</tr>
<tr>
<td>b jet identification efficiency</td>
<td>1</td>
</tr>
<tr>
<td>b jet misidentified probability</td>
<td>-</td>
</tr>
<tr>
<td>Lepton misidentified as $\tau_h$ probability</td>
<td>-</td>
</tr>
<tr>
<td>Jet misidentified as $\tau_h$ probability</td>
<td>-</td>
</tr>
<tr>
<td>$ll$ contamination</td>
<td>-</td>
</tr>
<tr>
<td>Extrapolation to the signal region</td>
<td>-</td>
</tr>
<tr>
<td><strong>$\tau_\mu \tau_h$, 2jets-0tag, 2jets-1tag, 2jets-2tag</strong></td>
<td></td>
</tr>
<tr>
<td>$E_T$ energy scale</td>
<td>1</td>
</tr>
<tr>
<td>b jet identification efficiency</td>
<td>5</td>
</tr>
<tr>
<td>b jet misidentified probability</td>
<td>-</td>
</tr>
<tr>
<td>Lepton misidentified as $\tau_h$ probability</td>
<td>-</td>
</tr>
<tr>
<td>Jet misidentified as $\tau_h$ probability</td>
<td>-</td>
</tr>
<tr>
<td>$ll$ contamination</td>
<td>-</td>
</tr>
<tr>
<td>Extrapolation to the signal region</td>
<td>-</td>
</tr>
</tbody>
</table>
5.2 Signal extraction

The signal is extracted using the CL$_s$-based method developed by the LHC Higgs Combination Group [127].

The procedure considers the distributions of the chosen variable ($m_{H}^{\text{kin}}fit$), the reference distributions, for the measured data and its modelled composition in all the event categories. The number of events predicted by the modelling in the bin $j$ of the reference distribution of the event category $i$ ($n_{ij}$) is defined by

$$n_{ij}(\mu, \theta_i) = \mu \cdot s_{ij}(\theta_i) + b_{ij}(\theta_i),$$  \hspace{1cm} (5.3)

where $s_{ij}(\theta_i)$ and $b_{ij}(\theta_i)$ are the modelled signal and background yields in the given bin, $\mu$ is the signal strength, and $\theta_i$ represents the values of all nuisance parameters considered in the event category $i$.

The signal strength parameter $\mu$ is introduced to represent different signal cross-section hypotheses, by varying the signal yield, with $\mu = 0$ corresponding to the background-only hypothesis, when the signal contribution is not present.

The probability of observing $n_{ij}^{\text{obs}}$ data events in bin $j$ of event category $i$, given $n_{ij} > 0$, has the Poisson distribution:

$$P(n_{ij}^{\text{obs}} | n_{ij}) = \text{Poiss}(n_{ij}^{\text{obs}}, n_{ij}) = \frac{\exp \left( -n_{ij} + n_{ij}^{\text{obs}} \ln n_{ij} \right)}{n_{ij}^{\text{obs}} !}. \hspace{1cm} (5.4)$$

$n_{ij}$ depends on the nuisance parameters that represent sources of the systematic and of the statistical uncertainties. The summary of the nuisance parameters considered in each event category, together with their best estimates obtained using the auxiliary measurements ($\theta_i^{\text{obs}}$), were presented in Tables 5.1 and 5.2.

Our beliefs concerning the real values of the nuisance parameters, based on their best estimates are quantified using probability density functions (pdfs) $p_k(\theta_k | \theta_k^{\text{obs}})$, which represents a posterior probability that the real value of the nuisance parameter $k$ equals $\theta_k$, given its best estimate $\theta_k^{\text{obs}}$.

For the normalization uncertainties, which correspond to multiplicative factors for the signal or background yields, the $\theta_k$ posterior probabilities are represented by log-normal pdfs.

From the histograms of the shape uncertainties (Section 5.1.2), for each nuisance parameter $\theta_k$ we may obtain three yield estimates: $n_{ij}(\theta_k^{\text{obs}} - \sigma_k, ...)$, $n_{ij}(\theta_k^{\text{obs}}, ...)$ and $n_{ij}(\theta_k^{\text{obs}} + \sigma_k, ...)$. Using these three estimates, each $n_{ij}(\theta_k, ...)$ is smoothly interpolated in the region $\theta_k^{\text{obs}} \pm \sigma_k$ and the new uniformly distributed nuisance parameter $\theta'_k$ is introduced to parameterize this yield variation within the $\theta_k^{\text{obs}}$ uncertainty.
5.2. Signal extraction

By using the flat prior, the posterior probability \( p_k(\theta_k | \theta_k^{\text{obs}}) \) is transformed to the prior probability \( p_k(\theta_k^{\text{obs}} | \theta_k) \), allowing the use of the frequentist approach in the subsequent steps.

The probability models from all bins and event categories are combined in the likelihood function \( \mathcal{L} \):

\[
\mathcal{L}(\text{data} | \mu, \theta) = \prod_{ij} P(n_{ij}^{\text{obs}} | n_{ij}) p(\theta | \theta), \tag{5.5}
\]

where ‘data’ may refer to either the observed data or generated ‘pseoudo-data’, as discussed below, and, respectively, \( \theta \) is either \( \theta^{\text{obs}} \) or \( \theta^{\text{pseudo-obs}} \).

To compare the compatibility of the data with different signal + background hypotheses against the background-only hypothesis, the test statistics \( \tilde{q}_\mu \) are defined using the profile likelihood ratio, defined as:

\[
\tilde{q}_\mu = -2 \ln \frac{\mathcal{L}(\text{data} | \mu, \hat{\theta}_\mu)}{\mathcal{L}(\text{data} | \bar{\mu}, \hat{\theta})}. \tag{5.6}
\]

\( \hat{\mu}, \hat{\theta} \) and \( \hat{\theta}_\mu \) are defined as

\[
(\hat{\mu}, \hat{\theta}) = \arg \max_{(\mu, \theta)} \mathcal{L}(\text{data} | \mu, \theta),
\]

\[
\hat{\theta}_\mu = \arg \max_{\theta} \mathcal{L}(\text{data} | \mu, \theta), \tag{5.7}
\]

where the maximizations are subject to two constraints: the physical constraint \( \hat{\mu} \geq 0 \), and the constraint that \( \hat{\mu} \leq \mu \). The second constraint allows setting a one-sided limit on \( \mu \).

The \( \tilde{q}_\mu \) pdfs \( p(\tilde{q}_\mu | \mu, \theta^{\text{obs}}) \) can be estimated by generating ‘pseoudo-data’ with toy MC experiments. However, this is a very CPU consuming task, therefore for this analysis we have used an asymptotic approximation to evaluate \( p(\tilde{q}_\mu | \mu, \theta^{\text{obs}}) \), according to the formulas defined in \[128\]. In this approximation, to evaluate \( \tilde{q}_\mu \), only a particular dataset (the Asimov dataset, defined in the previous reference), where the observations are equal to the predictions and the nuisance parameters are equal to their nominal values, is used. This approach does not require the use of toy MC experiments, which are so expensive in terms of CPU time. On the other hand, in the studies performed for the MS and MSSM \( H \to \tau\tau \) searches \[103, 124\], it was demonstrated that the results obtained using the asymptotic profile likelihood ratio are very close to the results obtained by using the profile likelihood ratio without the asymptotic approximation, even for a moderately limited events statistic.

From the \( \tilde{q}_\mu \) pdfs, the probability of observing \( \tilde{q}_\mu \geq \tilde{q}_\mu^{\text{obs}} \) for a given value of \( \mu \) is

\[
p_\mu \equiv P(\tilde{q}_\mu \geq \tilde{q}_\mu^{\text{obs}} | \mu) = \int_{\tilde{q}_\mu^{\text{obs}}}^{\infty} p(\tilde{q}_\mu | \mu, \hat{\theta}^{\text{obs}}) d\tilde{q}_\mu. \tag{5.8}
\]
The $p_\mu$ with $\mu > 0$ indicates the significance of the measurement under the hypothesis of a signal with strength $\mu$ ($CL_{s+b}(\mu)$), while $p_\mu$ with $\mu = 0$ indicates the significance of the measurement under the background-only hypothesis ($CL_b$).

In order to avoid rejecting a signal with strength $\mu$ when both $CL_{s+b}(\mu)$ and $CL_b$ are similarly small, the $CL_s(\mu)$ value is defined as

$$CL_s(\mu) = \frac{CL_{s+b}(\mu)}{CL_b}.$$  \hspace{1cm} (5.9)

The value $1 - CL_s$ is used in the CMS to cite the confidence level (CL), which is a conservative approach, since the values of $CL_s$ are usually bigger than the actual significance. The 95% CL value, i.e. $CL_s = 0.05$, was chosen as a default for reporting the CMS results.

To obtain the observed upper exclusion limits for the signal strength (signal cross-section times branching fraction) $\geq \mu_{up}$, the value of $\mu_{up}$ is adjusted until $CL_s(\mu_{up}) = 0.05$ is reached.

The observed exclusion limits are compared with the expected exclusion limits, and are supposed to be within the $\pm 1\sigma$ or $\pm 2\sigma$ error bands of the expected limits (corresponding respectively to $\approx 68\%$ and to $\approx 95\%$ confidence intervals) to check the agreement between the observations and the model predictions. In the case of the asymptotic approximation approach used here, the $N\sigma$ error bands $\mu_{up+N}^{exp}$ and the median $\mu_{up}^{exp} = \mu_{up+0}^{exp}$ of the expected limits are obtained using the Asimov dataset and are given by \[127,128\]

$$\mu_{up+N}^{exp} = \sigma_A \cdot \left[ \Phi^{-1}(1 - \alpha\Phi(N)) + N \right], \hspace{1cm} (5.10)$$

where $\alpha = 0.05$, $\sigma_A^2 = \mu^2/q_{\mu,A}$ is evaluated using the Asimov dataset and $\Phi(\cdot)$ is the cumulative standard Gaussian distribution.

As the result of the fit, the yields of all the modelled processes are adjusted according to the differences between $\theta^{obs}$ and $\hat{\theta}$. The parameter distributions obtained after this adjustment are called the ‘post-fit’ distributions, and represent the final model predictions for the observed data composition.
5.3 Results

The simultaneous maximum likelihood fit is carried out for all channels and categories. The purpose of the fit is to select the values of the nuisance parameters which maximize the likelihood of the signal plus background fit. This results in small changes in the shape and yield of the contributing terms. In order to visualize the agreement between the results of the fit and the data, we show, in following figures, the plots of several quantities that have a direct impact on the \( H \) candidate reconstruction and the \( m_H^{\text{kin,fit}} \) distribution. These plots are shown for each event category of the \( \tau_e \tau_h \) and \( \tau_\mu \tau_h \) channels:

- Figure 5.2: \( p_T \) of the selected electron (muon) candidate;
- Figure 5.3: \( p_T \) of the selected \( \tau_h \) candidate;
- Figure 5.4: \( p_T \) of the leading b jet candidate;
- Figure 5.5: \( p_T \) of the sub-leading b jet candidate;
- Figure 5.6: the MVA PF \( \mathcal{E}_T \);
- Figure 5.7: the invariant \( H \) candidate mass reconstructed using the kinematic fit, \( m_H^{\text{kin,fit}} \).

All post-fit control plots shows a reasonable agreement between the observed data and the model.
Chapter 5. Model fit procedure

Figure 5.2: The post-fit distributions of the selected electron (muon) candidate $p_T$ in the 2jets-0tag, 2jets-1tag and 2jets-2tag event categories of the $\tau_\ell\tau_h$ (left) and $\mu\tau_h$ (right) channels. Contribution from the modelled signal distribution is superimposed on the backgrounds. The signal yield is multiplied by 10, assuming an MSSM $H$ with $m_H = 300$ GeV and $\tan \beta = 2$. The ‘Electroweak’ in the legends (red) includes W+jets, di-boson and single top processes; the $Z \rightarrow ll$ (blue) includes $Z \rightarrow ee$ and $Z \rightarrow \mu\mu$ processes. Expected background contributions are shown for the values of nuisance parameters (systematic uncertainties) obtained after fitting the signal plus background hypothesis to the data.
5.3. Results

Figure 5.3: The post-fit distributions of the selected $\tau_h$ candidate $p_T$ in the $2\text{jets-0tag}$, $2\text{jets-1tag}$ and $2\text{jets-2tag}$ event categories of the $e_\tau^+2\tau_h$ (left) and $\mu_\tau^+2\tau_h$ (right) channels. Contribution from the modelled signal distribution is superimposed on the backgrounds. The signal yield is multiplied by 10, assuming an MSSM $H$ with $m_H = 300\text{GeV}$ and $\tan\beta = 2$. The ‘Electroweak’ in the legends (red) includes $W+\text{jets}$, di-boson and single top processes; the $Z \rightarrow ll$ (blue) includes $Z \rightarrow ee$ and $Z \rightarrow \mu\mu$ processes. Expected background contributions are shown for the values of nuisance parameters (systematic uncertainties) obtained after fitting the signal plus background hypothesis to the data.
Figure 5.4: The post-fit distributions of the leading b jet candidate $p_T$ in the $2\text{jets-0tag}$, $2\text{jets-1tag}$ and $2\text{jets-2tag}$ event categories of the $\tau\tau_h$ (left) and $\tau\mu\tau_h$ (right) channels. Contribution from the modelled signal distribution is superimposed on the backgrounds. The signal yield is multiplied by 10, assuming an MSSM $H$ with $m_H = 300$ GeV and $\tan \beta = 2$. The ‘Electroweak’ in the legends (red) includes $W+jets$, di-boson and single top processes; the $Z \to ll$ (blue) includes $Z \to ee$ and $Z \to \mu\mu$ processes. Expected background contributions are shown for the values of nuisance parameters (systematic uncertainties) obtained after fitting the signal plus background hypothesis to the data.
Figure 5.5: The post-fit distributions of the sub-leading b jet candidate $p_T$ in the 2jets-0tag, 2jets-1tag and 2jets-2tag event categories of the $\tau_\tau$, $t\tau_b$ (left) and $t\bar{t}\tau_b$ (right) channels. Contribution from the modelled signal distribution is superimposed on the backgrounds. The signal yield is multiplied by 10, assuming an MSSM $H$ with $m_H = 300$ GeV and $\tan \beta = 2$. The ‘Electroweak’ in the legends (red) includes $W+jets$, di-boson and single top processes; the $Z \rightarrow ee$ (blue) includes $Z \rightarrow \mu\mu$ processes. Expected background contributions are shown for the values of nuisance parameters (systematic uncertainties) obtained after fitting the signal plus background hypothesis to the data.
Figure 5.6: The post-fit distributions of the MVA PF $E_T$ in the $2\text{jets-0tag}$, $2\text{jets-1tag}$ and $2\text{jets-2tag}$ event categories of the $\tau_\tau T_h$ (left) and $\tau_\mu T_h$ (right) channels. Contribution from the modelled signal distribution is superimposed on the backgrounds. The signal yield is multiplied by 10, assuming an MSSM $H$ with $m_H = 300$ GeV and $\tan\beta = 2$. The ‘Electroweak’ in the legends (red) includes W+jets, di-boson and single top processes; the $Z \rightarrow ll$ (blue) includes $Z \rightarrow ee$ and $Z \rightarrow \mu\mu$ processes. Expected background contributions are shown for the values of nuisance parameters (systematic uncertainties) obtained after fitting the signal plus background hypothesis to the data.
Figure 5.7: The post-fit distributions of the reconstructed invariant mass of the H candidate ($m_{H}^{\text{kinfit}}$) in the $2\text{jets-0tag}$, $2\text{jets-1tag}$ and $2\text{jets-2tag}$ event categories of the $\tau_{e}\tau_{h}$ (left) and $\tau_{\mu}\tau_{h}$ (right) channels. Contribution from the modelled signal distribution is superimposed on the backgrounds. The signal yield is multiplied by 10, assuming an MSSM $H$ with $m_{H} = 300$ GeV and $\tan \beta = 2$. The ‘Electroweak’ in the legends (red) includes $W+$jets, di-boson and single top processes; the $Z \rightarrow ll$ (blue) includes $Z \rightarrow ee$ and $Z \rightarrow \mu\mu$ processes. Expected background contributions are shown for the values of nuisance parameters (systematic uncertainties) obtained after fitting the signal plus background hypothesis to the data.
The model independent 95% CL exclusion limits for the $gg \to H$ production cross-section of the heavy neutral scalar $H$ ($\sigma_{ggH}$) times the branching fraction $H \to hh \to bb\tau\tau$ ($\text{BR}_{bb\tau\tau}$) are calculated for all the MC signal samples, which were generated for $m_H$ values between 260 and 350 GeV with a step of 10 GeV. These limits have been linearly interpolated to cover the full range $260 \leq m_H \leq 350$ GeV. The observed $\sigma_{ggH} \cdot \text{BR}_{bb\tau\tau}$ exclusion limits are $\approx 0.4 - 0.8$ pb for the $\tau_\ell \tau_h$ channel and $\approx 0.2 - 0.7$ pb for the $\tau_\mu \tau_h$ channel, depending on $m_H$, as shown in Figure 5.8. For both channels, the observed limits lie within the $\pm 2\sigma$ confidence interval around the expected limits.

Figure 5.8: The observed and expected model independent 95% CL exclusion limits for the $gg \to H \to hh \to bb\tau\tau$ production cross-section times branching fraction of the heavy neutral scalar $H$ as a function of $m_H$ in $260 \leq m_H \leq 350$ GeV region obtained in the $\tau_\ell \tau_h$ (a) and $\tau_\mu \tau_h$ (b) channels.
5.3. Results

The $\tau\tau$ and $\tau\mu\tau\tau$ channel results were combined with the results of the $H \rightarrow hh \rightarrow b\bar{b}\tau\tau$ CMS analysis [18], where both $\tau$ decay hadronically (the $\tau\tau\tau\tau$ channel), using the same signal extraction technique described in Section 5.2. The observed $\sigma_{ggH} \cdot BR_{b\bar{b}\tau\tau}$ exclusion limits obtained from the combination of these three channels are $\approx 0.13 - 0.3$ pb, depending on $m_H$, as shown in Figure 5.9. The observed combined limits lie within the $\pm 2\sigma$ confidence interval around the expected limits.

Figure 5.9: The observed and expected model-independent 95% CL exclusion limits for the $gg \rightarrow H \rightarrow hh \rightarrow b\bar{b}\tau\tau$ production cross-section times branching fraction of the heavy neutral scalar $H$ as a function of $m_H$ in $260 \leq m_H \leq 350$ GeV region obtained by combining $H \rightarrow hh \rightarrow b\bar{b}\tau\tau$ CMS analyses results, considering $b\bar{b}\tau\tau\tau\tau$, $b\bar{b}\tau\tau\mu\tau\tau$ and $b\bar{b}\tau\tau\tau\tau$ final states. The analysis of the $b\bar{b}\tau\tau\tau\tau$ final state is described in [18].

The combined $\sigma_{ggH} \cdot BR_{b\bar{b}\tau\tau}$ exclusion limits were interpreted in the parameter space of the MSSM low tan $\beta$ benchmark scenario and the 2HDM Type 2 benchmark scenario with $m_H = m_A = m_{H^\pm}$ (Section 1.4). For both interpretations, the correspondence between the parameter space values and the value of $\sigma_{ggH} \cdot BR_{b\bar{b}\tau\tau}$ expected from the theory is made according to the recommendations of the LHC Higgs Cross Section Working Group [129,130]. The obtained exclusion limits in the $(m_A \times \tan \beta)$ MSSM parameter space are shown in Figure 5.11. An example of the obtained exclusion limits in the $(\cos(\beta - \alpha) \times \tan \beta)$ 2HDM parameter space for $m_H = 300$ GeV is shown in Figure 5.11.
Figure 5.10: The observed and expected 95% CL exclusion limits for the $m_A \times \tan \beta$ parameters region of the MSSM low $\tan \beta$ benchmark scenario (Section 1.3.3) phase space based on the combination of the $H \to hh \to b\bar{b}\tau\tau$ CMS analyses, considering $b\bar{b}\tau_e\tau_h$, $b\bar{b}\tau_\mu\tau_h$ and $b\bar{b}\tau_h\tau_h$ final states.

Figure 5.11: The observed and expected 95% CL exclusion limits for the $\cos(\beta-\alpha) \times \tan \beta$ parameters region of the 2HDM Type 2 benchmark scenario (Section 1.3.3) phase space with $m_H = m_A = m_{H^\pm} = 300$ GeV based on the combination of the $H \to hh \to b\bar{b}\tau\tau$ CMS analyses, considering $b\bar{b}\tau_e\tau_h$, $b\bar{b}\tau_\mu\tau_h$ and $b\bar{b}\tau_h\tau_h$ final states.
Conclusions and outlooks

My thesis consisted in the search for a heavy scalar neutral Higgs $H$ in the decay mode $H \rightarrow hh \rightarrow bb\tau\tau$ in the final states $bb\tau_+\tau_-$ and $bb\tau_-\tau_+$. The data analysed for this work were collected by the CMS in 2012 in pp collisions, at a centre of mass energy of 8 TeV, and correspond to an integrated luminosity of 19.7 fb$^{-1}$.

In this work, model-independent 95% CL exclusion limits on the $gg \to H$ production cross section times the branching fraction $\text{BR}(H \to hh \to bb\tau\tau)$ of a scalar neutral boson $H$ as a function of its mass in the range $260 \leq m_H \leq 350$ GeV was set for the $bb\tau_+\tau_-$ and $bb\tau_-\tau_+$ channels. These results were combined with the $bb\tau_+\tau_-$ channel, obtaining 95% CL exclusion limits in the range $\approx 0.13 - 0.3$ pb, depending on $m_H$.

The obtained model-independent exclusion limits are complementary to the results of other double-Higgs searches performed by the CMS collaboration, using different final states [46,47,131], as shown in Figure 5.12. To perform the comparison, the SM BR($h \to bb$) $\approx 0.575$ and BR($h \to \tau\tau$) $\approx 0.063$ are used, to obtain the exclusion limits for $\sigma(gg \to H) \cdot \text{BR}(H \to hh)$. The corresponding 95% CL exclusion limits are in the range $\approx 1.7 - 4.1$ pb, depending on $m_H$.

From Figure 5.12, the sensitivity of the $bb\tau\tau$ is slightly worse than the $bb\gamma\gamma$, but better than the $bbbb$ channel, in the range of mass investigated in this thesis.

The $H \rightarrow hh$ searches, similar to that of the CMS, were performed by the ATLAS collaboration, considering the $hh \rightarrow bb\tau\tau$, $\gamma\gamma WW^*$, $\gamma\gamma bb$ and $bbbb$ channels [132]. The range of mass explored by ATLAS to extract the limits for the $bb\tau\tau$ channel is $260 \leq m_H \leq 1000$ GeV. To compare the results of ATLAS and CMS for this channel, the common range of $m_H$ for the two experiments has to be used. In this range, the resulting 95% CL exclusion limits set by ATLAS are $\approx 1.7 - 4.2$ pb, which are comparable to the results obtained in this thesis.

The model dependent interpretations of the exclusion limits based on the combined results for two benchmark scenarios, MSSM low $\tan\beta$ and the 2HDM Type 2, were set, excluding small regions of the MSSM $m_A \times \tan\beta$ and the 2HDM $\cos(\beta - \alpha) \times \tan\beta$ parameter space.
Figure 5.12: The observed and expected 95% CL exclusion limits for the heavy neutral scalar $H$ production times $H \rightarrow hh$ branching fraction set by CMS searches, considering $\gamma\gamma(bb, bbb\gamma, \gamma+\text{leptons and } b\tau\tau$ final states, and assuming the SM Higgs branching fractions for the $h$ decays.

For the current Run 2 of the LHC with $\sqrt{s} = 13\text{ TeV}$, the searches for the double-Higgs production should be able to cover a significant part of the MSSM and 2HDM parameter space. On the other hand, for the High Luminosity LHC (HL-LHC, after 2024), where the instantaneous luminosity is expected to increase by a factor of 10, one of the most important goals is the precision measurement of the SM Higgs self-coupling to probe for new physics.

The tools and techniques developed for the work presented in this thesis and that will be improved for Run 2, will be an important legacy for the search of the SM double-Higgs production. Therefore, the pursuit of the analysis of this thesis with the new Run 2 and future HL-LHC data seems to be very appealing.
Bibliography


[27] ATLAS Collaboration, “Measurement of the Higgs boson mass from the $H \to \gamma\gamma$ and $H \to ZZ^\ast \to 4\ell$ channels with the ATLAS detector using 25 fb$^{-1}$ of pp collision data”, *Phys. Rev.* **D90** (2014), no. 5, 052004, \texttt{doi:10.1103/PhysRevD.90.052004}, \texttt{arXiv:1406.3827}.


[93] M. Wobisch and T. Wengler, “Hadronization corrections to jet cross-sections in deep inelastic scattering”, in Monte Carlo generators for...


[132] ATLAS Collaboration, “Searches for Higgs boson pair production in the $hh \to b\bar{b}\tau\tau, \gamma\gamma WW*, \gamma\gamma bb, b\bar{b}bb$ channels with the ATLAS detector”, [arXiv:1509.04670], CERN-PH-EP-2015-225, 2015.