Search for light bosons in the final state with muons and tau leptons with CMS Run II data

Dissertation

zur Erlangung des Doktorgrades

an der Fakultät für Mathematik,

Informatik und Naturwissenschaften

Fachbereich Physik

der Universität Hamburg

vorgelegt von

Sandra Consuegra Rodríguez

aus

Havanna, Kuba

Hamburg

2020

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

I hereby declare, on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids.

Hamburg, den 23. September 2020 Sandra Consuegra Rodríguez

Gutachter der Dissertation:	Dr. Alexei Raspereza Prof. Dr. Elisabetta Gallo
Zusammensetzung der Prüfungskommission:	Dr. Alexei Raspereza Prof. Dr. Elisabetta Gallo Prof. Dr. Robin Santra Prof. Dr. Gudrid Moortgat-Pick Prof. Dr. Johannes Haller
Vorsitzender der Prüfungskommission:	Prof. Dr. Robin Santra
Datum der Disputation:	29. Oktober 2020
Vorsitzender des Fach-Promotionsausschusses PHYSIK:	Prof. Dr. Günter H. W. Sigl
Leiter des Fachbereichs PHYSIK:	Prof. Dr. Wolfgang Hansen
Dekan der Fakultät MIN:	Prof. Dr. Heinrich Graener

ABSTRACT

A search for a pair of light bosons produced in decays of the 125 GeV Higgs boson, with one of the light states decaying into a pair of muons and the other into a pair of tau leptons, is presented. The search is based on a data sample corresponding to an integrated luminosity of 137.2 fb^{-1}, collected with the CMS detector at the CERN Large Hadron Collider in the years 2016, 2017, and 2018 at a center-of-mass energy of 13 TeV.
An extended Higgs sector is well motivated in a vast set of Beyond the Standard Model theories such as the two Higgs doublets plus one additional singlet (2HDM+S) and the Dark Photon Model. In the context of these models, the 125 GeV Higgs boson can decay into a pair of light bosons, which subsequently decay to pairs of Standard Model particles. Considering the enhanced coupling of the light bosons to leptons for some of the scenarios within these models, the final state considered in this work results of particular interest.
Masses of the light boson between 3.6 and 21 GeV are probed, which leads to an experimental signature in the detector with both the muon pair and visible decay products from the tau pair being highly collimated. The analysis benefits from the efficient identification and reconstruction of muons by the CMS detector. Using Multivariate Analysis Techniques, the information on several kinematic variables is exploited to enhance the sensitivity to the targeted topology. No significant excess of events is found above the Standard Model expectation. Therefore, model-independent upper limits at 95% confidence level on the 125 GeV Higgs boson production cross-section times the branching fraction into the studied final state are set. Model-specific upper bounds are obtained as constraints on the parameter space of the different benchmark scenarios within the 2HDM+S and the Dark Photon Model.

ZUSAMMENFASSUNG

In dieser Arbeit wird eine Suche für die Erzeugung eines Paares von neuen leichten Bosonen präsentiert, die aus dem Zerfall des 125 GeV HiggsBosons stammen und bei dem eines der neuen Teilchen in ein Paar von Muonen zerfällt und das andere in ein Paar von Tau-Leptonen. Die Suche basiert auf Daten die mit dem CMS Detektor in Proton-Proton Kollisionen am "Large Hadron Collider" am CERN in den Jahren 2016 bis 2018 aufgezeichnet wurden, bei einer Schwerpunktsenergie von 13 TeV. Die analysierte Datenmenge entspricht einer integrierten Luminosität von 137.2 fb^{-1}. Ein erweiterter Higgs Sektor wird von vielen Theorien jenseits des Standardmodells angenommen, wie zum Beispiel vom *"Two Higgs doublets plus one additional singlet (2HDM+S)"* Modell und dem *"Dark Photon"* Modell. In diesen Theorien kann das 125 GeV HiggsBoson in ein Paar von neuen leichten Bosonen zerfallen, die danach wiederum jeweils in Paare von Standardmodellteilchen zerfallen. Einigen Szenarios zufolge könnten die neuen Bosonen verstärkt an Leptonen koppeln und damit ist der hier untersuchte Zerfall in Leptonpaare von besonderem Interesse. Die Studie deckt einen Massenbereich von leichten Bosonen von 3.6 bis 21 GeV ab, was zu einer experimentellen Signatur im Detektor führt bei der sowohl das Muonpaar als auch die sichtbaren Zerfallsprodukte aus dem Taupaar stark kollimiert sind. Die Analyse profitiert von der effizienten Identifizierung und Rekonstruktion von Muonen mit dem CMS Detektor. Eine multivariate Analysetechnik wird auf ausgewählte kinematische Observablen angewandt um die Sensitivität auf ein mögliches Signal in der untersuchten Ereignistopologie zu erhöhen. Als Ergebnis wird kein signifikanter Überschuss von Ereignissen im Vergleich zu der Erwartung aus Standardmodellprozessen beobachtet. Stattdessen werden modellunabhängige obere Ausschlussgrenzen bei 95% Konfidenzniveau bestimmt für den Wirkungsquerschnitt für die Produktion des 125 GeV HiggsBosons multipliziert mit dem Verzweigungsverhältnis in den untersuchten Endzustand. Darüberhinaus werden modellspezifische obere Ausschlussgrenzen gesetzt im Parameterraum verschiedener Referenszenarien des *"2HDM+S"* Modells und des *"Dark Photon"* Modells.

CONTENTS

CHAPTER

1

INTRODUCTION

I am among those who think that science has great beauty

Marie Curie

The curiosity for unraveling the mysteries of nature surrounding us is an innate human condition. Since the move towards a rational understanding of nature began, philosophers struggled to find a first principle or element corresponding to the "ultimate underlying substance." The Greek philosopher Empedocles, nurtured by the knowledge gathered during the particularly prolific time in humankind in which he lived in, proposed what at that time was a revolutionary idea. He argued that all matter was composed of four elements: fire, air, water, and earth. He also stated that the ratio of these four elements would affect the properties of the matter. The suggestion that some substances that looked like pure materials could be made from a combination of different elements was an important development in scientific thinking. A few decades later, Democritus went a step further and stated that all matter is made of fundamental elements, which he called *atoms*, meaning indivisible.

From the very first ideas on the structure of matter, a long way has been transited. This path led us to an extraordinary development in the field of particle physics during the twentieth century. Fundamental questions found an answer or began to be answered. Are there fundamental, indivisible particles, and if so, what are they? How do they behave? How do they group to form the matter observed in nature? How do they interact with each other? Furthermore, the improvements in particle accelerators and detector technology allowed us to move from a picture with three fundamental particles (proton, neutron, and electron) to an extensive list of new particles, the so-called particle zoo. By the mid-1960s, the need for an explanation at a fundamental level that would provide simplicity and elegance to the crowded picture was clear. A theory known as the Standard Model of Particle Physics (SM) emerged.

The next several decades, with the discovery of the W [1, 2] and Z bosons [3, 4] in 1983, the top quark in 1995 [5], the tau neutrino in 2000 [6], among others, served to pave the way and

consolidate the Standard Model experimentally. Within this theory, three of the four known fundamental forces in the universe (electromagnetic, weak, and strong interactions), as well as the underlying structure behind the particle zoo, are described with remarkable accuracy. Despite the Standard Model success, many questions remain yet to be answered. Why a portion of the mass of the universe, bound up in the so-called Dark Matter, is not accounted for within the model? Why is the universe dominated by matter and not made of equals parts out of matter and antimatter? How can one add gravity into the picture? Moreover, the mass of the neutrinos, the hierarchy problem, and the strong CP-problem remain also to be explained.

To enable particle physicists to shed light on some of these questions, increasingly high energy regions began to be studied. The Large Hadron Collider (LHC) [7], the largest and up to now most powerful particle accelerator ever built, was constructed in a tunnel 100m underground between the Swiss and French borders. The first collisions were achieved in 2009, and already in 2012, the existence of the long thought Higgs boson was confirmed [8, 9]. Up to now, all the measurements of its properties have shown consistency with the SM within the uncertainties [10–24]. The Higgs discovery, along with the impressive amount of data collected at the LHC, have given physicists a sensitive tool to search for incompatibilities with the SM, which would result in new theoretical developments and could provide an answer to some of the questions stated above.

One of the simplest extensions of the Standard Model, the Two-Higgs-doublet Model (2HDM) [25], leads to a rich phenomenology featuring five physical states: two CP even neutral Higgs bosons h and H_0 (H_0 heavier than h by convention), one CP odd pseudoscalar A, and two charged Higgs bosons $H^{\pm}$. Exotic decays of the form $h \to AA$, $H_0 \to hh$, AA or $h \to ZA$, with subsequent decays of the daughter (pseudo)scalars to SM fermions or gauge bosons, are still possible in certain corners of parameter space but are now too constrained from existing data [26]. An additional one complex scalar singlet can be added to the 2HDM in the so-called 2HDM+S, which results in seven physical states: three scalars, two pseudoscalars, and two charged particles. The complex scalar singlet must only couple to the doublets in the potential. As a result, the mostly-singlet light pseudoscalar state has no direct Yukawa couplings, acquiring all of its couplings to SM fermions through the mixing between the singlet and the doublets. Under these two assumptions, exotic Higgs decays of the form h$\to aa \to X\overline{X}Y\overline{Y}$ are allowed, where a is one of the pseudoscalars, h is one of the scalars, compatible with the discovered Higgs boson, and X/Y are SM fermions or gauge bosons.

The Dark Photon Model constitutes another SM extension that includes an extended Higgs sector. In the context of the model, the four-fermion final state can be obtained through the exotic decay $h \to Z_D Z_D \to X\overline{X}Y\overline{Y}$, where Z_D is the dark photon candidate of the model.

The number of exotic Higgs decays that could be contained in the data collected at the LHC represents a considerable discovery potential. At one of the LHC collision points sits the Compact Muon Solenoid (CMS) [27] experiment. As the name "Compact Muon Solenoid" suggests, CMS was specifically designed to provide good muon detection and resolution. Therefore, it is especially suited to detect the decay of Higgs bosons with muons in the final state.

This doctoral dissertation focuses on the search for a pair of light bosons, produced in decays of the 125 GeV Higgs boson, in the final state with two muons and two tau leptons. The dataset used corresponds to an integrated luminosity of 137.2 fb^{-1}, collected with the CMS detector at the LHC during the Run II data-taking period, at a center-of-mass energy of 13

TeV. Masses of the light boson between 3.6 and 21 GeV are probed. Due to the large mass difference of the light bosons with respect to the 125 GeV Higgs boson, they are produced with a high boost and decay leaving a trace of collimated decay products. The experimental signature of the $\mu\mu\tau\tau$ final state is therefore characterized by the presence of a pair of muons and visible decay products from the tau lepton pair, produced with a high Lorentz boost. The final state profits from the characteristic resonance of the reconstructed dimuon mass spectrum in the $a_1(Z_D) \to \mu\mu$ decay and the dominant $\mathrm{a}_1 \to \tau\tau$ decay in many scenarios within the 2HDM+S.

This thesis is composed of nine chapters. Chapter 2 is dedicated to present a theoretical review of the Standard Model of particle physics. In Chapter 3, the 2HDM+S and the Dark Photon Model are introduced. The next chapter is focused on the description of the LHC machine and the CMS detector, used for the collection of the analyzed data. In Chapter 5, the generation of events, the simulation of the CMS detector, and the reconstruction of events are discussed. The global event description and the technical aspects of the reconstruction techniques to identify the physics objects of interest for this work are detailed. Chapter 6 is dedicated to introducing the statistical methods used for the analysis of the data. In Chapter 7, the main topic of this doctoral dissertation is presented, the search for a pair of light bosons in the final state with two muons and two tau leptons. Chapter 8 discusses the results of the analysis of the data and their interpretation in the context of the 2HDM+S and the Dark Photon Model. Chapter 9 is dedicated to present a summary of this work and a discussion on the prospects of this analysis after the second run of the LHC.

CHAPTER

2

THE STANDARD MODEL OF PARTICLE PHYSICS

Contents

The field of elementary particle physics blossomed in the last century. New particles were discovered one after another. At a first moment, the cosmic rays experiments were the only available source for high energy particles. But soon, the discovery of elementary particles in cosmic rays motivated the construction of high energy accelerators. Their intense and controlled beams of known energy allowed to reveal the quark substructure of matter. High precision measurements at LEP, SLC, Tevatron, LHC, and others have followed. The Standard Model of Particle Physics, describing the fundamental particles and (except for gravity) their interactions, is cemented in all these decades of prolific research. Through many experimental tests, the Standard Model has been established as a well-tested physical theory, encapsulating our best understanding of the link between the fundamental particles and three of the four fundamental forces. The comparative strengths of the force between two protons when just in contact, indicate the relative magnitudes of the four types of interaction:

strong	electromagnetic	weak	gravity
1	10^{-2}	10^{-7}	10^{-39}

with gravity being by far the weakest force and having no relevant effect in particle physics at accelerator energies. The Standard Model can be described as the union of Quantum Chromodynamics (QCD) and the Electroweak Theory. It is perturbative at sufficiently high energies and renormalizable. Therefore, it describes the electroweak and strong interaction at the quantum level. The construction of the model is cemented under principles of symmetry and the mathematical foundation of symmetry is given by the Group Theory. The gauge group of the model ($SU(3)_C \times SU(2)_L \times U(1)_Y$) can be divided into two sectors: $SU(2)_L \times U(1)_Y$ and $SU(3)_C$, associated to the electroweak and QCD theories, respectively. The next section is dedicated to present the algebraic foundation of the Standard Model, which motivates theoretically the particle content of the model presented in Subsec. 2.5.1. Starting from the definition of a group, the groups of particular significance in the formulation of the model are introduced.

2.1 Algebraic Foundations: The Standard Model Groups

A group G is formed by a set of elements $\{a, b, c, ...\}$, with a rule that combines any two elements of the group, so that the combined element is also an element of the group, satisfying the following conditions:
(i) The rule is associative: $a(bc) = (ab)c$.
(ii) G contains a unique identity element I such that, for every element a of G:

$$aI = Ia = a. \tag{2.1}$$

(iii) For every element a of G there is a unique inverse element a^{-1} such that:

$$aa^{-1} = a^{-1}a = I. \tag{2.2}$$

A group is said to be commutative or *Abelian* if: $ab = ba$ for all a, b; and is called a *Lie group* if its elements depend in a continuous and differentiable way on a set of real parameters

$\theta^a, a = 1, ..., N$ [28]. The Lie group theory is a powerful way of understanding and classifying symmetries. Subsecs. 2.1.1 - 2.1.4 introduce the Lie groups $U(1)$, $SU(2)$, and $SU(3)$, along with the $SL(2, C)$ group, which is related to the group of proper Lorentz transformations.

2.1.1 U(1)

An $n \times n$ matrix U is unitary if $UU^\dagger = U^\dagger U = I$ and the product of two unitary matrices is also unitary. Thus, $n \times n$ unitary matrices form a group under matrix multiplication, denoted by $U(n)$. Since:

$$\det(\boldsymbol{U}\boldsymbol{U}^\dagger) = \det \boldsymbol{U} \det \boldsymbol{U}^* = \det \boldsymbol{U} (\det(\boldsymbol{U})^*) = \det \boldsymbol{I} = 1, \tag{2.3}$$

$\det \boldsymbol{U}$ can be written as $\det \boldsymbol{U} = e^{in\alpha}$, where α is real. For the simple case in which n equals 1, the group $U(1)$ is obtained. $U(1)$ consists of all complex numbers with absolute value equal to 1 under the multiplication operation.

2.1.2 SU(2)

The Special Unitary group $SU(2)$ is the group of all 2×2 unitary matrices with determinant equal to 1. $SU(2)$ is a subgroup of $U(2)$, which is obtained for the case of $n = 2$ in $U(n)$, and it can be represented by $\boldsymbol{U} = exp(i\boldsymbol{H})$. $\boldsymbol{H}$ is a Hermitian matrix, taking the form:

$$H = \begin{pmatrix} \alpha^0 + \alpha^3 & \alpha^1 - i\alpha^2 \\ \alpha^1 + i\alpha^2 & \alpha^0 - \alpha^3 \end{pmatrix}. \tag{2.4}$$

The $\alpha^\mu (\mu = 0, 1, 2, 3)$ terms constitute four real parameters. Thus, $\boldsymbol{H}$ can be written as: $\boldsymbol{H} = \alpha^0 \boldsymbol{I} + \alpha^k \sigma^k$, with the index k running from 1 to 3 and the σ components being the Pauli spin matrices:

$$\sigma^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \sigma^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \qquad \sigma^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{2.5}$$

The unit matrix I commutes with all matrices. Hence, a general member of $U(2)$ in the fundamental representation can be formulated as:

$$\boldsymbol{U} = \exp i(\alpha^0 \boldsymbol{I} + \alpha^k \sigma^k) = \exp(i\alpha^0) \exp(i\alpha^k \sigma^k), \tag{2.6}$$

with the phase factor $exp(i\alpha^0)$ belonging to the group $U(1)$, so that the elements of $SU(2)$ take the form:

$$\boldsymbol{U_s} = \exp(i\alpha^k \sigma^k). \tag{2.7}$$

The three parameters α^k specify an element and the matrices σ^k are the corresponding generators of the group.

2.1.3 SL(2,C) and the proper Lorentz group

The Special Linear group $SL(2, C)$ is the group of all 2×2 matrices with complex elements and with determinant equal to 1. These matrices form a group under matrix multiplication.

A general Hermitian matrix is associated to each point $x = (x^0, \boldsymbol{x})$ in space-time:

$$\boldsymbol{X}(x) = \begin{pmatrix} x^0 + x^3 & x^1 - ix^2 \\ x^1 + ix^2 & x^0 - x^3 \end{pmatrix}. \tag{2.8}$$

The determinant of $\boldsymbol{X}$ takes the form: $\det \boldsymbol{X} = (x^0)^2 - x^k x^k$. Considering an element $\boldsymbol{M}$ of $SL(2, C)$, the matrix $\boldsymbol{X'}$ results from the operation:

$$\boldsymbol{M}^\dagger \boldsymbol{X'} \boldsymbol{M} = \boldsymbol{X} \quad \text{or} \quad \boldsymbol{X'} = (\boldsymbol{M^{-1}})^\dagger \boldsymbol{X} \boldsymbol{M^{-1}}. \tag{2.9}$$

$\boldsymbol{X'}$ is also Hermitian and, therefore, can be written as:

$$\boldsymbol{X'} = \begin{pmatrix} x'^0 + x'^3 & x'^1 - ix'^2 \\ x'^1 + ix'^2 & x'^0 - x'^3 \end{pmatrix}. \tag{2.10}$$

x'^μ and x^μ ($\mu = 0, 1, 2, 3$) are related by a real linear transformation. Additionally:

$$\det \boldsymbol{M}^\dagger \boldsymbol{X'} \boldsymbol{M} = \det \boldsymbol{M}^\dagger \det \boldsymbol{X'} \det \boldsymbol{M} = \det \boldsymbol{X'} = \det \boldsymbol{X}, \tag{2.11}$$

so that, $(x'^0)^2 - x'^k x'^k = (x^0)^2 - x^k x^k$. Thus, the matrix $\boldsymbol{M}$ corresponds to a Lorentz transformation matrix $\boldsymbol{L(M)}$. The matrices $\boldsymbol{L(M)}$ form a group that contains the identity transformation $\boldsymbol{L(I) = I}$ and, therefore, by continuity correspond to proper Lorentz transformations.
A general proper Lorentz transformation between two frames K and K' is defined by six parameters: three parameters for the velocity v of K' relative to K and three parameters for the orientation of K' relative to K. The condition $\det \boldsymbol{M} = 1$ reduces the eight real parameters of the general 2×2 complex matrix to six. Thus, a matrix $\boldsymbol{M}$ can be associated to every proper Lorentz transformation. The matrices $\boldsymbol{M}$ and $\boldsymbol{-M}$ give the same transformation. Hence, two elements of $SL(2, C)$ will be associated to each element of the proper Lorentz group.

2.1.4 SU(3)

The Special Unitary group $SU(3)$ is the group of all 3×3 unitary matrices with determinant equal to 1. Likewise the group $U(2)$ in Subsec. 2.1.2, $U(3)$ can be expressed as $\boldsymbol{U} = \exp(i\boldsymbol{H})$, with $\boldsymbol{H}$ in this case being a 3×3 Hermitian matrix. A general 3×3 matrix is specified by $3^2 = 9$ real parameters. The condition $\det \boldsymbol{U} = 1$, reduces this number of parameters by 1. In place of the Pauli matrices in Subsec. 2.1.2, eight traceless Hermitian matrices introduced by Gell-Mann are used:

$$\lambda_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \lambda_2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{2.12}$$

$$\lambda_4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \qquad \lambda_5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}, \qquad \lambda_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

$$\lambda_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \qquad \lambda_8 = (1/\sqrt{3}) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}.$$

Hence, $\boldsymbol{H}$ takes the form, $\boldsymbol{H} = \alpha_1\lambda_1 + \alpha_2\lambda_2 + ... + \alpha_8\lambda_8$

$$= \begin{pmatrix} \alpha_3 + \alpha_8/\sqrt{3} & \alpha_1 - i\alpha_2 & \alpha_4 - i\alpha_5 \\ \alpha_1 + i\alpha_2 & -\alpha_3 + \alpha_8/\sqrt{3} & \alpha_6 - i\alpha_7 \\ \alpha_1 + i\alpha_5 & \alpha_6 + i\alpha_7 & -2\alpha_8/\sqrt{3} \end{pmatrix}. \tag{2.13}$$

The matrices λ_a fulfill the commutation relations:

$$[\lambda_a, \lambda_b] = 2i \sum_{c=1}^{8} f_{abc}\lambda_c. \tag{2.14}$$

The structure constants f_{abc} are odd in the interchange of two indices, with non-vanishing components coming from the permutations of $f_{123} = 1$, $f_{147} = f_{246} = f_{257} = f_{345} = f_{516} = f_{637} = 1/2$, and $f_{458} = f_{678} = \sqrt{3}/2$. The structure constants also have the property:

$$\mathrm{Tr}(\lambda_a\lambda_b) = 2\delta_{ab}, \tag{2.15}$$

where δ_{ab} is the Kronecker δ.
Before introducing the Standard Model Lagrangian in Sec. 2.5, Sec. 2.2 and Sec. 2.3 are dedicated to present some topics of particular relevance in the formalism of the Standard Model: the Gauging of the Symmetry, Gauge Theories, and the Spontaneous Symmetry Breaking.

2.2 Gauging the Symmetry

Let's consider the Dirac Lagrangian:

$$\mathcal{L} = \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi, \tag{2.16}$$

which is invariant under the *global* $U(1)$ transformations of the form $\psi = e^{i\alpha}\psi$, introduced in Subsec. 2.1.1. A transformation is said to be global when it acts on the field in the exact same way at every point of spacetime. In this section it will be illustrated how a global symmetry can be made *local*, so that the factor α ($\alpha = \alpha(x^\mu)$) can depend on spacetime, and later on the Lagrangian will be forced to maintain its invariance, this time under a local $U(1)$ transformation. Making a global symmetry local is known as gauging the symmetry. As a result of applying the local $U(1)$ transformation to Eq. (2.16) the Lagrangian takes the form:

$$\mathcal{L} = \bar{\psi} e^{-\alpha(x)} (i\gamma^\mu \partial_\mu - m) e^{i\alpha(x)} \psi. \tag{2.17}$$

Extra terms from the differential operators acting on $\alpha(x)$ are obtained:

$$\begin{aligned} \bar{\psi} e^{-\alpha(x)} (i\gamma^\mu \partial_\mu - m) e^{i\alpha(x)} \psi &= \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi - \bar{\psi}\gamma^\mu \psi \partial_\mu \alpha(x) \\ &= \bar{\psi}(i\gamma^\mu \partial_\mu - m - \gamma^\mu \partial_\mu \alpha(x))\psi. \end{aligned} \tag{2.18}$$

For the Lagrangian invariance to be maintained under the applied local $U(1)$ transformation, the term $\bar{\psi}\gamma^\mu \partial_\mu \alpha(x)$ needs to be canceled. In order to cancel this term, an arbitrary field A_μ is defined, which transforms under the $U(1)$ transformation $e^{i\alpha(x)}$ according to:

$$A_\mu \to A_\mu - \frac{1}{q} \partial_\mu \alpha(x), \tag{2.19}$$

where q is a proportionality constant. The field A_μ is called *Gauge field*. Before introducing A_μ in Eq. (2.17), one first conveniently replaces the partial derivative ∂_μ with the *covariant derivative*:

$$D_\mu \equiv \partial_\mu + iqA_\mu. \tag{2.20}$$

Adding the field A_μ defined in Eq. (2.19) to Eq. (2.16) restores the $U(1)$ symmetry. The Lagrangian is now invariant under both the local and global transformations, with the same conserved $U(1)$ current $j^\mu = \bar{\psi} j^\mu \psi$. As a next step, the corresponding gauge-invariant kinetic term for an arbitrary field A_μ is added:

$$\mathcal{L}_{\text{Kin,A}} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu}, \tag{2.21}$$

where:

$$F^{\mu\nu} \equiv \frac{i}{q} [D^\mu, D^\nu] = \partial^\mu A^\nu - \partial^\nu A^\mu. \tag{2.22}$$

D^μ and D^ν are the covariant derivatives defined in Eq. (2.20) and q is the constant of proportionality already introduced in the transformation of A_μ in equation Eq. (2.19). With the addition of a source term J^μ for the field, the final Lagrangian takes the form:

$$\mathcal{L} = \bar{\psi}(i\gamma^\mu D_\mu - m)\psi - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} - J^\mu A_\mu. \tag{2.23}$$

Thus, in this section, starting from a Lagrangian for a spin-1/2 particle, invariant under global $U(1)$ transformations, the global $U(1)$ symmetry was promoted to a local symmetry, as treated in Ref. [29]. The gauge field A_μ came to fulfill the need of an additional term in the Lagrangian in order to get a consistent theory. Consequently, a kinetic term and a source term were also added. Starting with nothing but a non-interacting particle and requiring nothing but $U(1)$ symmetry at local level, one ends up with a field A_μ. Upon quantization, this field A_μ will be the photon. Therefore, the electromagnetism is said to be described by a $U(1)$ symmetry group, with the photon as a direct consequence of imposing $U(1)$. The type of theories in which one generates forces by specifying a Lie group, are called *Gauge Theories*, or *Yang-Mills Theories*.

2.3 Spontaneous Symmetry Breaking

The Lagrangian of a complex scalar field theory with a mass term and a quartic self-interaction can be expressed as:

$$\mathcal{L} = \partial_\mu \Phi^* \partial^\mu \Phi - V(\Phi), \tag{2.24}$$

where the potential $V(\Phi)$ is given by:

$$V(\Phi) = \mu^2 \Phi^* \Phi + \lambda |\Phi^* \Phi|^2. \tag{2.25}$$

The Lagrangian in Eq. (2.24) is invariant under the global $U(1)$ transformation [30]. Provided μ^2 is positive, the potential reaches its minimum value at $\Phi = 0$. Since the vacuum of any theory is at the minimum value of the potential, the vacuum of this theory is the state $\Phi = 0$. In terms of a Quantum Field Theory, where Φ is an operator, one would say that the operator Φ has zero *vacuum expectation value (VEV)*. If the potential in Eq. (2.25) is changed, by reverting the sign of μ^2, now -μ^2, it will no longer have a minimum at $\Phi = 0$, but at:

$$\Phi = \frac{v}{\sqrt{2}} = e^{i\theta} \sqrt{\frac{\mu^2}{2\lambda}}. \tag{2.26}$$

There exists now an infinite number of states, each with the same lowest energy, i.e., changing the potential results in a degenerate vacuum. θ can take any value from 0 to 2π, but for convenience $\theta = 0$ is chosen to be the vacuum, which results in $<\Phi> = v/\sqrt{2}$. The field Φ is said to have now a non-zero vacuum expectation value. In the language of Quantum Mechanics, *symmetry breaking* occurs when a field, or some components of a field, acquire a non-zero VEV. The symmetry breaking arises from the choice made for the value of the vacuum. It is said to be spontaneous because no external agent is responsible for it. One has just chosen one of many degenerate ground states to be the true vacuum. The theory would then have no obvious $U(1)$ symmetry. One can write the field in terms of fluctuations around the chosen vacuum. The spontaneous breaking of global symmetries always results in a massless boson, called *Goldstone Boson*.

2.4 The Higgs Mechanism

In the previous section, a global $U(1)$ symmetry was broken and the consequences examined. In this section, a similar procedure is followed, but this time for a local $U(1)$ symmetry, as treated in Ref. [29]. The Lagrangian for a complex scalar field with a gauged $U(1)$ field can be written as:

$$\mathcal{L} = -\frac{1}{2}[(\partial^\mu - iqA^\mu)\phi^\dagger][(\partial_\mu + iqA_\mu)\phi] - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} - V(\phi^\dagger, \phi). \tag{2.27}$$

The external source J^μ has been taken as 0. Assuming that the potential takes the form in Eq. (2.25), the vacuum has the $U(1)$ degeneracy at $|\phi| = \Phi$. Since $U(1)$ is now local, the factor $\alpha(x)$ of the arbitrary field A_μ from Eq. (2.19) is chosen so that both the vacuum and ϕ

are real. One can rewrite the field in terms of fluctuations around the chosen ground state:

$$\phi = \Phi + h, \tag{2.28}$$

where h is a real scalar field that represents the fluctuations around the vacuum. Now, introducing Eq. (2.28) in Eq. (2.27), the Lagrangian takes the form:

$$\mathcal{L} = -\frac{1}{2}\partial^\mu h \partial_\mu h - \frac{1}{2}4\lambda m^2 \Phi^2 h^2 - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} - \frac{1}{2}q^2\Phi^2 A^2 + \mathcal{L}_{\text{interactions}}. \tag{2.29}$$

Before the breaking of the local $U(1)$ symmetry, one had a complex scalar field ϕ and a massless vector field A^μ with two polarization states. Now, one has a single real scalar h and a field A^μ with masses equal to $\sqrt{4\lambda m^2\Phi^2}$ and $q\Phi$, respectively. The force-carrying particle A^μ has gained mass! Starting with a theory with no mass, and by means of spontaneous breaking of the symmetry, a mass has been introduced. This mechanism for introducing mass into a theory is called the *Higgs Mechanism*, built on the union of Gauge invariance and spontaneous symmetry breaking. The resulting field h is named the *Higgs Boson*. One can conclude that while the consequence of global symmetry breaking is a massless boson, as seen in Sec. 2.3, the effect of a local symmetry breaking is that the gauge field obtained as a result of the symmetry being local acquires mass.

2.5 Construction of the Standard Model Lagrangian

2.5.1 Particle content

In the Standard Model the fundamental particles are divided in two groups: fermions and bosons. The fermions constitute the building blocks of matter, while the bosons are the mediators of the interactions. The photons constitute the quanta of the electromagnetic interaction field between electrically charged fermions. The charged W^+ and W^- bosons and the neutral Z boson are the quanta of the weak interaction fields between fermions, while the quanta of the strong interaction field are the massless gluons. The fermions can be classified in two types: leptons and quarks. With spin 1/2, fermions are described by the Fermi-Dirac statistics. Bosons have an integer spin and obey the Bose-Einstein statistics.
The boson mediators are listed in Tab. 2.1 [31].

Table 2.1: The boson mediators

Interaction	Mediator	Spin/Parity
strong	gluon, G	1^-
electromagnetic	photon, γ	1^-
weak	$W^\pm$, Z^0	1^-, 1^+

Leptons interact exclusively through the electromagnetic and the weak interaction, while quarks interact through the electromagnetic, the weak, and the strong interaction. The leptons carry integer electric charge, with the electron and its antiparticle, the positron (e^+) being the only stable charged leptons. The muon, the τ lepton, and their antiparticles differ

from the electron and the positron only in their masses and finite lifetimes. The neutral leptons are called neutrinos, denoted by the generic symbol ν. Tab. 2.2 lists the symbol and the ratio of the electric charge Q to the elementary charge e of the electron, for each of the six quarks and six leptons, the area of influence of the fundamental interactions, and the lifetime for the leptons. The quarks carry a fractional charge. Their mass increases from left to right in Tab. 2.2, as is the case for the charged leptons. Quarks are grouped into two groups that differ only by one unit of electric charge. The quark type known as *flavor* is denoted by the symbols: u for "up", d for "down", s for "strange", c for "charm", b for "bottom", and t for "top". The algebra of the gauge group of the Standard Model was already introduced in

Table 2.2: The fundamental fermions

Particles	Generation/ Mass / Lifetime (s) for leptons			Q/|e|	Interaction (mediator)		
	I	II	III				
quarks	u $2.16^{+0.49}_{-0.26}$ (MeV)	c 1.27±0.02 (GeV)	t 172.76±0.30 (GeV)	$\frac{2}{3}$	strong (gluon)	el. mag. (photon)	weak ($W^{\pm}$, Z^0)
	d $4.67^{+0.48}_{-0.17}$ (MeV)	s 93^{+11}_{-5} (MeV)	b $4.18^{+0.03}_{-0.02}$ (GeV)	$-\frac{1}{3}$			
leptons	e 0.511 (MeV) stable	μ 105.66 (MeV) 2.20×10^{-6}	τ 1776.86 (MeV) 2.90×10^{-13}	-1			
	ν_e < 1.1 eV stable	ν_μ < 1.1 eV stable	ν_τ < 1.1 eV stable	0			

Sec. 2.1. The specific gauge bosons related to the generators of the gauge group of the model are:

$$\begin{array}{ccccc} SU(3)_C & \times & SU(2)_L & \times & U(1)_Y, \\ \downarrow & & \downarrow & & \downarrow \\ 8G^\alpha_\mu & & 3W^a_\mu & & B_\mu \\ \alpha = 1, ..., 8 & & a = 1, 2, 3 & & \end{array} \tag{2.30}$$

where the fields representing the spin-one gauge bosons and their transformation rules are denoted as follows:

$$\begin{array}{lll} G^\alpha_\mu & \text{transforms as} & (\mathbf{8}, \mathbf{1}, 0) \\ W^a_\mu & & (\mathbf{1}, \mathbf{3}, 0) \\ B_\mu & & (\mathbf{1}, \mathbf{1}, 0). \end{array} \tag{2.31}$$

The gluons are the eight spin-one particles, $G^{\alpha}_{\mu}(x)$, associated with the factor $SU(3)_C$. The subscript C denotes the *color*. The particles that transform with respect to the factor $G^{\alpha}_{\mu}(x)$ of the gauge group, and so which couple to the gluons, are said to be colored or to carry color. Any particle that couples to the gluons is said to be strongly interacting and the interaction is called *strong interaction*. Three spin-one particles, $W^{a}_{\mu}(x)$, are associated with the factor $SU(2)_L$. The subscript L is used to indicate that only the left-handed fermions carry this quantum number. An additional spin-one particle $B_{\mu}(x)$ is associated with the factor $U(1)_Y$. The subscript Y denotes the quantum number of the weak hypercharge and aims to differentiate this quantum number from the ordinary electric charge Q, defined as the sum of the weak hypercharge Y and the $SU(2)_L$ charge's T_3 component ($Q = T_3 + Y$). The electromagnetic group is not directly the $U(1)_Y$ component of the Standard Model gauge group. Thus, the electric charge Q is not one of the basic charges that particles carry under $SU(3)_C \times SU(2)_L \times U(1)_Y$. It is rather a derived quantity, arising after electroweak symmetry breaking, with the electromagnetic group written as $U_{em}(1)$.
The four spin-one bosons associated with the factor $SU(2)_L \times U(1)_Y$ are related to the physical bosons mediating the weak interactions $W^{\pm}$ and Z^0, and the photon [32]. Since the three lepton and quark families hold the same quantum numbers, respectively, and can only be distinguished through their masses, it is sufficient to consider only one family when discussing the gauge interaction. Tab. 2.3 summarizes the transformation behavior of the quark and lepton fields under the SM gauge groups for one generation [33].

Table 2.3: Transformation behavior under the SM gauge groups

Field	$SU(3)_C \times SU(2)_L \times U(1)_Y$
$Q_L = \begin{pmatrix} \mu_L \\ d_L \end{pmatrix}$	$(\mathbf{3}, \mathbf{2}, \frac{1}{3})$
u_R	$(\bar{\mathbf{3}}, \mathbf{1}, \frac{2}{3})$
d_R	$(\bar{\mathbf{3}}, \mathbf{1}, -\frac{4}{3})$
$L_L = \begin{pmatrix} e_L \\ \nu_{e_L} \end{pmatrix}$	$(\mathbf{1}, \mathbf{2}, -1)$
e_R	$(\bar{\mathbf{1}}, \mathbf{1}, 2)$

To construct the Standard Model Lagrangian, one first postulate the set of symmetries of the system. The most general renormalizable Lagrangian that fulfills these symmetries is constructed from the particle and field content introduced above. The SM Lagrangian density can then be written as a sum of Lagrangian densities [34]:

$$\mathcal{L} = \mathcal{L}_{\text{gauge}} + \mathcal{L}_{\text{fermion}} + \mathcal{L}_{\text{Higgs}} + \mathcal{L}_{\text{Yukawa}}. \tag{2.32}$$

The next subsections are dedicated to introducing each of the sectors that correspond to the four added Lagrangian densities.

2.5.2 Gauge sector

The Gauge sector of the SM Lagrangian ($\mathcal{L}_{\text{gauge}}$) describes the massless gauge bosons. The gauge boson dynamics is encoded in the Lagrangian and is expressed in terms of the field strength tensors as:

$$\mathcal{L}_{\text{gauge}} = -\frac{1}{4}G^a_{\mu\nu}G^{a\mu\nu} - \frac{1}{4}W^a_{\mu\nu}W^{a\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}, \tag{2.33}$$

where repeated indices are always taken as summed. The quadratic G term of the field strength tensor for $SU(3)_C$ takes the form:

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu + g_s f^{abc} G^b_\mu G^c_\nu, \tag{2.34}$$

where g_s is the strong interaction coupling strength, a, b, c run from 1 to 8, and f^{abc} are the structure constants of $SU(3)$, defined in Subsec. 2.1.4. W^a_μ are the $SU(2)_L$ gauge bosons already introduced in Subsec. 2.5.1 and $W^a_{\mu\nu}$ the corresponding field strength tensors, which take the form:

$$W^a_{\mu\nu} = \partial_\mu W^a_\nu - \partial_\nu W^a_\mu + g\epsilon^{abc} W^b_\mu W^c_\nu, \tag{2.35}$$

where g is the weak interaction coupling strength, a, b, c run from 1 to 3, and $f^{abc} = \epsilon^{abc}$ [35]. The field strength tensor corresponding to the $U(1)_Y$ interaction can be written as:

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu. \tag{2.36}$$

The hypercharge $U(1)_Y$ constitutes the underlying $U(1)$ symmetry of the model.

2.5.3 Fermion sector

The fermionic part of the Lagrangian ($\mathcal{L}_{\text{fermion}}$) describes the massless fermions and their interactions with the gauge bosons. It can be written as:

$$\mathcal{L}_{\text{fermion}} = \sum_{\text{quarks}} i\bar{q}\gamma^\mu D_\mu q + \sum_{\psi_L} i\bar{\psi_L}\gamma^\mu D_\mu \psi_L + \sum_{\psi_R} i\bar{\psi_R}\gamma^\mu D_\mu \psi_R. \tag{2.37}$$

The covariant derivative acts on ψ_R and ψ_L in the second and third term as:

$$D_\mu \psi_L = (\partial_\mu + igW_\mu + ig' Y_L B_\mu)\psi_L \qquad D_\mu \psi_R = (\partial_\mu + ig' Y_R B_\mu)\psi_R, \tag{2.38}$$

where g and g' are the $SU(2)_L$ and $U(1)_Y$ couplings, respectively. The $L(R)$ index refers to the left (right) chiral projections $\psi_{L(R)} = (1 \pm \gamma_5)\psi/2$. The left-handed quarks (d, u, s, c, b, t) and leptons $(e, \nu_e, \mu, \nu_\mu, \tau, \nu_\tau)$ are arranged in doublets, while the right-handed fermions are singlets [34, 36]. The following are the three generations of SU(2) doublet pairs of quarks and leptons:

$$\psi_L: \quad L^i = \begin{pmatrix}\nu_{eL}\\ e_L\end{pmatrix}, \begin{pmatrix}\nu_{\mu L}\\ \mu_L\end{pmatrix}, \begin{pmatrix}\nu_{\tau L}\\ \tau_L,\end{pmatrix} \qquad Q^i = \begin{pmatrix}u_L\\ d_L\end{pmatrix}, \begin{pmatrix}c_L\\ s_L\end{pmatrix}, \begin{pmatrix}t_L\\ b_L\end{pmatrix}, \tag{2.39}$$

where $i = 1, 2, 3$ correspond to the indexes of the generations [37], and the right-handed fermions can be indexed by the first-generation label:

$$\begin{aligned} \psi_R: \quad & e_R^i = \{e_R, \mu_R, \tau_R\}, \quad \nu_R^i = \{\nu_{eR}, \nu_{\mu R}, \nu_{\tau R}\}, \\ & u_R^i = \{u_R, c_R, t_R\}, \quad d_R^i = \{d_R, s_R, b_R\}. \end{aligned} \tag{2.40}$$

2.5.4 Higgs sector

The Higgs sector of the Lagrangian ($\mathcal{L}_{\text{Higgs}}$) is responsible for the *electroweak symmetry breaking (EWSB)*. Thus, it gives mass to the electroweak gauge bosons. The term $\mathcal{L}_{\text{Higgs}}$ of the SM Lagrangian can be expressed as:

$$\mathcal{L}_{\text{Higgs}} = (D^\mu \phi)^\dagger (D_\mu \phi) - V(\phi). \tag{2.41}$$

The first term of the Lagrangian contains the kinetic and gauge-interaction terms via the covariant derivative while the second term is a potential energy function of ϕ. The most general gauge-invariant potential involving ϕ can be written as:

$$V(\phi) = -\mu^2 \phi^\dagger \phi + \lambda (\phi^\dagger \phi)^2, \tag{2.42}$$

where the λ term describes quartic self-interactions among the scalar field. The Higgs field ϕ is a self-interacting $SU(2)_L$ complex doublet with weak hypercharge $Y = 1$, which takes the form:

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \sqrt{2}\phi^+ \\ \phi^0 + i a^0 \end{pmatrix}, \tag{2.43}$$

where ϕ^0 and a^0 are the CP-even and CP-odd neutral components, and ϕ^+ is the complex charged component of the Higgs doublet [38]. For negative values of the quadratic term μ^2, the neutral component of the scalar doublet acquires a non-zero vacuum expectation value:

$$<\phi> = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}. \tag{2.44}$$

The spontaneous breaking of the SM gauge symmetry $SU(3)_C \times SU(2)_L \times U(1)_Y$ into $SU(3)_C \times U(1)_{em}$ is induced by $\phi^0 = H + <\phi^0>$, with $<\phi^0> \equiv v$. The potential in Eq. (2.42) has a mexican hat shape, as depicted in Fig. 2.1. The axial symmetry is not violated at the top, but as the ball rolls down it will be spontaneously broken, with the rotational U(1) symmetry broken at any of the points located at the bottom of the potential.

The position of the minimum is determined with only one real parameter, the vacuum expectation value of the field ($v = (\sqrt{2} G_F)^{-\frac{1}{2}}$), which is fixed by the Fermi coupling (G_F) to approximately 246 GeV. The Higgs field couples to the W_μ and B_μ gauge fields associated with the $SU(2)_L \times U(1)_Y$ local symmetry through the electroweak covariant derivative:

$$D_\mu \phi = (\partial_\mu + i g T^i W_\mu^i + i \frac{1}{2} g' B_\mu) \phi, \tag{2.45}$$

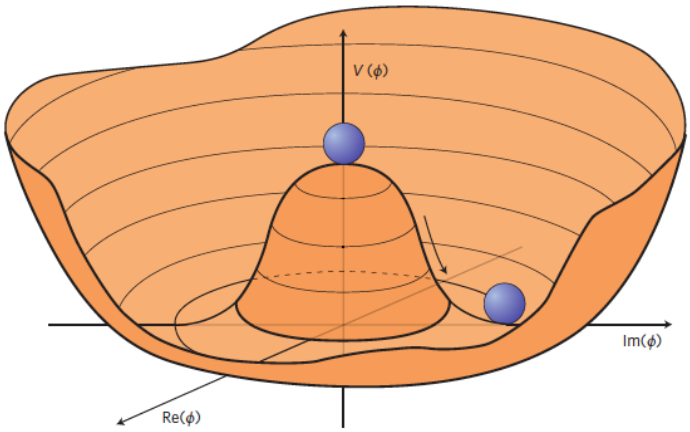


Figure 2.1: Illustration of the Higgs potential for $\mu^2 < 0$ [39].

where g and g' are the couplings already introduced in Eq. (2.38), and $T^i = \frac{\tau^i}{2}$ (τ^i are the three Pauli matrices). Working with the first term of the Lagrangian:

$$(D^\mu\phi)^\dagger(D_\mu\phi) = \frac{v^2}{8}\left[g^2\left((W_\mu^1)^2 + (W_\mu^2)^2\right) + (gW_\mu^3 - g'B_\mu)^2\right], \tag{2.46}$$

and defining the charged vector boson W_μ^- along with its complex conjugate as:

$$W_\mu^\pm \equiv \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2), \tag{2.47}$$

the g^2 term in Eq. (2.46) yields the W mass ($m_W = \frac{gv}{2}$). Thus, the combinations $W^1 \mp iW^2$ correspond to the charged W bosons. Additionally, the neutral gauge bosons Z and A are obtained:

$$Z_\mu = \frac{1}{\sqrt{g^2 + g'^2}}(gW_\mu^3 - g'B_\mu) \quad \rightarrow \quad m_Z = \frac{v}{2}\sqrt{g^2 + g'^2}, \tag{2.48}$$

$$A_\mu = \frac{1}{\sqrt{g^2 + g'^2}}(g'W_\mu^3 + gB_\mu) \rightarrow \quad m_A = 0.$$

Once the experimental values for the W and Z boson masses are determined, the electroweak mixing angle is obtained [40, 41]:

$$sin^2\theta_W = 1 - \frac{M_W^2}{M_Z^2} = 0.223. \tag{2.49}$$

From the four generators of the $SU(2)_L \times U(1)_Y$ SM gauge group, three are spontaneously broken and one remains unbroken, the one associated to the conserved $U(1)_{em}$ gauge symmetry. Thus, its corresponding gauge field, the photon, remains massless.

2.5.5 Yukawa sector

Fermions acquire mass through their Yukawa interaction with the single Higgs doublet ϕ, once EWSB occurs. The last missing piece of the SM Lagrangian to be introduced is the part that spans this interaction ($\mathcal{L}_{\text{Yukawa}}$), which takes the form:

$$\mathcal{L}_{\text{Yukawa}} = -\hat{h}_{d_{ij}} \bar{q}_{L_i} \phi \, d_{R_j} - \hat{h}_{u_{ij}} \bar{q}_{L_i} \tilde{\phi} \, u_{R_j} - \hat{h}_{l_{ij}} \bar{l}_{L_i} \phi \, e_{R_j} + h.c., \tag{2.50}$$

where $h.c.$ stands for hermitian conjugate of the previous terms, $\tilde{\phi} = i\sigma_2\phi^*$, while $q_L(l_L)$ and u_R, $d_R(e_R)$ are the quark (lepton) $SU(2)_L$ doublets and singlets, respectively. For each term a 3×3 matrix in family space parametrizes $\hat{h}_{X_{ij}}$ and after EWSB the Higgs-fermion interactions are diagonalized ($\hat{h}_{f_{ij}} \to h_{f_i}\delta_{ij}$). The masses of the fermions generated through the Yukawa interaction are proportional to the VEV of the Higgs field:

$$m_{f_i} = \frac{h_{f_i} v}{\sqrt{2}}, \tag{2.51}$$

with the index i referring to the three families in the up-quark, down-quark, and charged lepton sectors. The EWSB mechanism itself does not provide an insight on possible underlying reasons for the large variety of fermion masses. This, among other SM shortcomings, is discussed in Sec. 2.7.

Some observations can be made after working further with the first term of the Lagrangian in Eq. (2.41). All the Higgs couplings can be written in terms of the masses of the particles to which it couples. The dependence of the Higgs boson couplings on the mass of the fundamental particles makes this new type of interaction very weak for light particles like the electron but strong for heavy particles like the top quark. Being more precise, the SM Higgs couplings to fundamental fermions are linearly proportional to the fermion masses, while the couplings to bosons are proportional to the square of the boson masses. The following Lagrangian summarizes the SM Higgs boson couplings to gauge bosons and fermions, as well as the Higgs boson self-coupling:

$$\mathcal{L} = -g_{Hf\bar{f}} \bar{f} f H + \frac{g_{HHH}}{6} H^3 + \frac{g_{HHHH}}{24} H^4 + \delta_V V_\mu V^\mu (g_{HVV} H + \frac{g_{HHVV}}{2} H^2), \tag{2.52}$$

with

$$g_{Hf\bar{f}} = \frac{m_f}{v}, \;\; g_{HVV} = \frac{2m_V^2}{v}, \;\; g_{HHVV} = \frac{2m_V^2}{v^2}, \;\; g_{HHH} = \frac{3m_H^2}{v}, \;\; g_{HHHH} = \frac{3m_H^2}{v^2}, \tag{2.53}$$

where $V = W^\pm$ or Z, $\delta_W = 1$, and $\delta_Z = 1/2$. Thus, the dominant mechanisms for the SM Higgs boson production and decay involve the coupling of H to W, Z, and the third generation of quarks and leptons. The Higgs boson coupling to gluons is generated at leading order by a one-loop process in which H couples to a virtual $t\bar{t}$ pair, while the Higgs boson coupling to photons is generated via a one-loop graph, with a virtual W^+W^- pair having the dominant contribution and a smaller contribution from a virtual $t\bar{t}$ pair.

2.6 SM Higgs collider phenomenology

The collider phenomenology allows us to build a bridge between theory and experiment. On the one hand, one would like to study experimentally the consequences of a particular theoretical model. On the other hand, once the data is collected and processed, one would like to interpret the results and understand their implications.

2.6.1 Production mechanisms at hadron colliders

The main Higgs boson production mechanisms at hadron colliders are gluon fusion (ggF), vector boson fusion (VBF), associated production with a gauge boson (VH), and associated production with a pair of top quarks (ttH) or with a single top quark (tHq). The corresponding representative diagrams for these dominant Higgs boson production processes are depicted in Fig. 2.2.

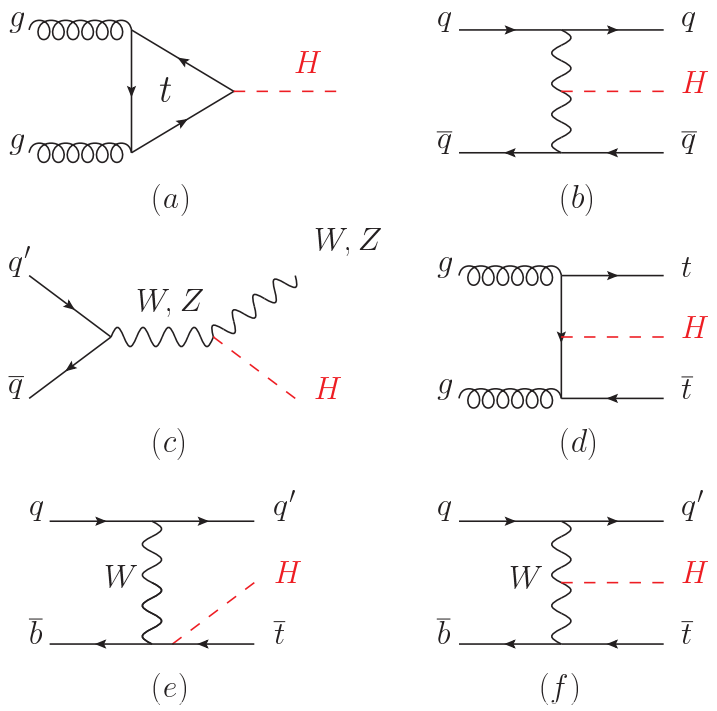


Figure 2.2: The main leading order Feynman diagrams contributing to the Higgs boson production: (a) gluon fusion, (b) Vector-boson fusion, (c) Higgs-strahlung (or associated production with a gauge boson), (d) associated production with a pair of top (or bottom) quarks, (e-f) production in association with a single top quark [38].

In parallel with the experimental effort, big progress has been achieved in the precision of the theoretical calculations for the Higgs boson production cross-sections. Higher order quantum corrections that result from the strong and electroweak interactions are now also considered. Tab. 2.4 summarizes the cross-sections for the production of the SM Higgs boson and the relative uncertainties as a function of the center-of-mass-energy for pp collisions. The quoted theoretical uncertainties largely emerge from unknown contributions from missing higher order corrections. The first listed center-of-mass-energy corresponds to the maximum center-of-mass-energy reached by the Tevatron accelerator, which operated until 2011 in the

premises of the Fermi National Accelerator Laboratory (FERMILAB). The additional values correspond to the centers-of-mass-energy reached or expected to be reached at different stages of operation of the LHC. More details on the LHC machine performance are given in Subsec. 4.1.2.

$\sqrt{s}$ (TeV)	Production cross-section (in pb) for m_H=125 GeV					
	ggF	VBF	WH	ZH	ttH	total
1.96	$0.95^{+17\%}_{-17\%}$	$0.065^{+8\%}_{-7\%}$	$0.13^{+8\%}_{-8\%}$	$0.079^{+8\%}_{-8\%}$	$0.004^{+10\%}_{-10\%}$	1.23
7	$16.9^{+4.4\%}_{-7.0\%}$	$1.24^{+2.1\%}_{-2.1\%}$	$0.58^{+2.2\%}_{-2.3\%}$	$0.34^{+3.1\%}_{-3.0\%}$	$0.09^{+5.6\%}_{-10.2\%}$	19.1
8	$21.4^{+4.4\%}_{-6.9\%}$	$1.60^{+2.3\%}_{-2.1\%}$	$0.70^{+2.1\%}_{-2.2\%}$	$0.42^{+3.4\%}_{-2.9\%}$	$0.13^{+5.9\%}_{-10.1\%}$	24.2
13	$48.6^{+4.6\%}_{-6.7\%}$	$3.78^{+2.2\%}_{-2.2\%}$	$1.37^{+2.6\%}_{-2.6\%}$	$0.88^{+4.1\%}_{-3.5\%}$	$0.50^{+6.8\%}_{-9.9\%}$	55.1
14	$54.7^{+4.6\%}_{-6.7\%}$	$4.28^{+2.2\%}_{-2.2\%}$	$1.51^{+1.9\%}_{-2.0\%}$	$0.99^{+4.1\%}_{-3.7\%}$	$0.60^{+6.9\%}_{-9.8\%}$	62.1

Table 2.4: The SM Higgs boson production cross-sections and relative uncertainties for m_H = 125 GeV as a function of the center-of-mass-energy, $\sqrt{s}$, for pp collisions ($p\bar{p}$ collisions at $\sqrt{s}$ = 1.96 TeV for the Tevatron). The upper (lower) percentage values of the relative uncertainties refer to the positive (negative) variation of the theoretical uncertainty [38].

The strength of the Higgs boson interaction with SM particles is dependent on their mass, as already discussed in Subsec. 2.5.4. Thus, the production of Higgs bosons at the LHC mainly involves heavy fermions and the massive vector bosons W and Z. These massive particles are produced via radiation processes from the incoming quarks or gluons in the colliding protons, as depicted in Fig. 2.2. The large abundance of gluons, together with the process being mediated by the exchange of a virtual heavy top quark, makes gluon-fusion the larger Higgs boson production mechanism at the LHC, producing Higgs bosons largely at rest and with no associated objects. In VBF, the second process in importance, the incoming quarks radiate a W or Z boson, which fuse to produce a Higgs boson. The scattered quarks are seen as two back-to-back hard jets in the forward and backward region of the detector. This characteristic fingerprint allows us to distinguish the VBF processes from the overwhelming QCD background and provides a clean environment for the determination of the Higgs boson couplings. The next most relevant Higgs boson production mechanisms are WH, ZH, and ttH associated production. WH and ZH provide a rather clean environment for the study of the Higgs boson decay into a pair of b-quarks, while ttH provides a direct probe of the top-Higgs Yukawa coupling, with the Higgs boson being radiated off top quarks.

2.6.2 Branching ratios and total width

Once produced, the Higgs boson decays promptly. The theoretical computation of all relevant Higgs boson decay widths and the total width is essential for a later interpretation of the experimental results. The decay branching fraction depends on the Higgs boson interaction strength with the particles it decays in, which in turn depends on the mass (Subsecs. 2.5.4 and 2.6.1). The Higgs boson discovery was accomplished essentially through production and decay channels related to the Higgs boson couplings to vector gauge bosons. Back in Summer 2012, the ATLAS and CMS experiments observed an excess of events near m_H = 125

GeV in the $H \to ZZ^* \to 4l$ and $H \to \gamma\gamma$, confirmed by the sensitive but low-resolution $H \to WW^* \to l\nu l\nu$ channel. About one year later, on October 8th of 2013, François Englert and Peter Higgs were awarded the Nobel Prize in physics: "for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider". The fermionic final states required a larger data sample to produce statistically significant measurements and had to wait until a partial Run II dataset was collected.
In the Born approximation, the partial width of the Higgs boson decay into fermion pairs can be written as:

$$\Gamma_{\text{Born}}(H \to f\bar{f}) = \frac{G_\mu N_c}{4\sqrt{2}\pi} M_H m_f^2 \beta_f^3, \tag{2.54}$$

where $\beta = (1 - 4m_f^2/M_H^2)^{1/2}$ is the velocity of the fermions in the final state and N_C is the color factor $N_C = 3(1)$ for quarks (leptons) [42]. The branching ratios into fermions can then be obtained as the ratio of the partial width to the total width $\Gamma(H \to f\bar{f})/\Gamma_{\text{tot}}$, where Γ_{tot} is given by a sum over all the partial decay widths of the Higgs boson. In the leptonic case, the predominant decay branching ratio is to $\tau^+\tau^-$ pairs, followed to a much lesser extent by the decay to $\mu^+\mu^-$ pairs. With a dataset corresponding to the 2016 data-taking period, the Higgs boson decay to fermions of the third generation (bottom quarks, tau leptons, and top quarks) was finally observed by the ATLAS and CMS experiments. Fig. 2.3 presents the SM Higgs branching ratios in the mass range 120 GeV $\leq M_H \leq$ 130 GeV, depicted as solid lines. The colored bands around the lines show the respective uncertainties. The fermionic decay $H \to b\bar{b}$ dominates over the entire mass range. Due to the values of the vector-boson masses, decays into two real massive vector bosons are not possible but decays into one real and one virtual boson can occur. Thus, the Higgs boson can decay into lower mass virtual W^* and Z^* bosons, which serve as mediators and decay promptly (the * is used to indicate that the particle is virtual). One of this kind of decays $H \to WW^*$, has the highest branching ratio after $H \to b\bar{b}$. The $H \to gg$, $H \to \tau^+\tau^-$, $H \to c\bar{c}$, and $H \to ZZ^*$ decays follow in that order. With much smaller rates, the $H \to \gamma\gamma$, $H \to \gamma Z$, and $H \to \mu^+\mu^-$ decays come behind. The decays into gluons, diphotons, and $Z\gamma$ are only possible via quantum-loop processes, involving heavy charged particles like W bosons, Z bosons, and top quarks. Hence, they provide indirect information on the Higgs boson couplings to WW, ZZ, and $t\bar{t}$. The total width of the SM Higgs boson is $\Gamma_H = 4.07 \times 10^{-3}$ GeV, with a relative uncertainty of $^{+4.0\%}_{+3.9\%}$ [38].

2.7 Standard Model shortcomings

The Standard Model is extremely well tested. It predicts the existence of particles that were subsequently found with precisely the foreseen properties such as the W and Z bosons, the gluon, two of the heavier quarks (the charm and the top quark), and the Higgs boson. The electroweak mixing angle presented in Eq. (2.45) has been found to maintain the same value for every electroweak process, as predicted in the SM. Several other precision measurements have provided stringent tests of the SM structure and an accurate determination of some of the 18 free parameters of the model. Despite the incredible success of the SM, the theory is unable to explain some phenomena and cannot even accommodate others. Thus, it

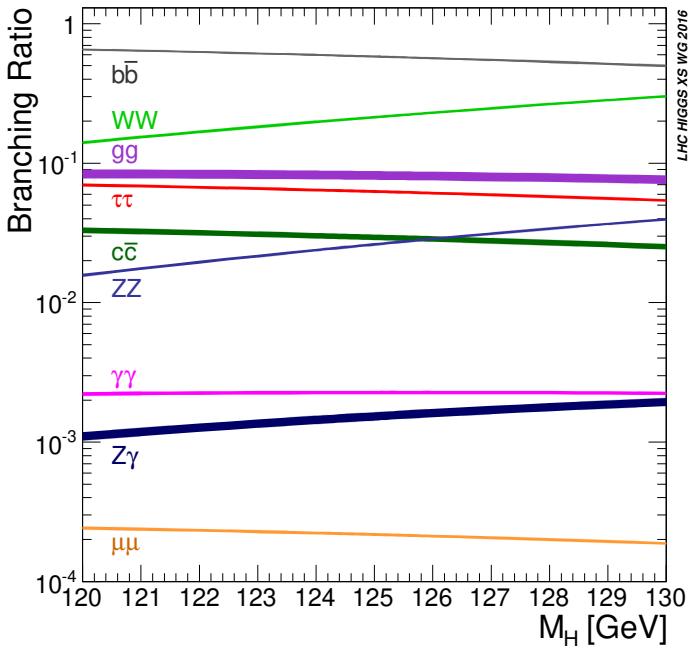


Figure 2.3: The branching ratios for the main decays of the SM Higgs boson. The bands represent the current knowledge of the theoretical uncertainties [38].

is strongly believed that the SM is an effective theory of a more fundamental one.
The SM cannot provide an explanation to the astronomical observations showing that the cosmological constant (the energy density of the vacuum) is smaller by many orders of magnitude than expected. This is the so-called cosmological constant problem. For a long time, the expansion of the universe was thought to be slowing down due to the mutual gravitational attraction of the matter in the universe. Nevertheless, it is known now that the expansion is accelerating, and the cause of the acceleration cannot be explained within the SM. Furthermore, the fields responsible for the extremely rapid expansion, which appears to have occurred in the first moments of the Big Bang, cannot be SM ones. This phenomenon of rapid expansion is called inflation. After the huge burst of energy in the Big Bang, matter and antimatter should have evolved into equal parts. Nevertheless, looking out at the stars, galaxies, gas clouds, clusters, superclusters, and the largest-scale structures of the universe, everything seems to be composed of matter and not antimatter. This matter-antimatter asymmetry cannot be explained within the SM. The ordinary matter described by the SM makes up less than five percent of the universe's energy density. About an additional quarter is made out of a new kind of unknown matter referred to as *dark matter* (DM), with the name coming from the fact that it doesn't interact with the electromagnetic force. Dark matter has not yet been observed experimentally, and no SM particle is found to meet the required properties to be a dark matter candidate. The remaining energy of the universe is accounted for by a hypothetical form of energy called dark energy, which seems to exert a negative and repulsive pressure. The existence of dark energy is thought to be the cause of the unexpected increase in the expansion rate of the universe.
Another puzzling question not answered by the SM refers to the so-called hierarchy problem: why the electroweak scale ($\approx 10^2$ GeV) and the Planck scale ($\approx 10^{19}$ GeV) differ in so many

orders of magnitude? The vast difference between both orders of magnitude is the hierarchy in the name hierarchy problem. It seems unnatural that, being the bare Higgs mass and the quantum corrections of the order of $\approx 10^{17}$ GeV, the cancellation of terms results in a renormalized Higgs mass reduced to its experimental value of 125 GeV.

In the SM, the neutrinos are very weakly interacting massless particles with no electric charge. However, experimental results show that neutrinos do have a mass, providing evidence that new physics Beyond the Standard Model (BSM) must exist. Although very low, the mass of the neutrinos gives rise to physical phenomena like neutrino oscillations. A neutrino of one type can change into one of a different type, which would not be possible if all three neutrinos have zero mass or the same mass. When a flavor state is produced by a weak interaction, e.g., a muon neutrino, the formed state is a mixture of states with different mass, which evolve at different rates. As a consequence of this, the muon neutrino can interact at a later time as a flavor state different from its original flavor. The term oscillation refers to this possibility of flavor change, namely that the neutrino is created in one flavor and can interact later on as another. The phenomenon of neutrino oscillation was observed by the Super-Kamiokande experiment in 1998 [43]. The observation allowed us to establish that neutrinos have non-zero and non-degenerate masses. Shortly after, the SNO experiment showed that electron neutrinos born in the core of the sun transition to a mixture of all three flavors, explaining the fewer-than-expected number of electron neutrinos detected on earth [44]. These observations were granted the 2015 Nobel Prize in Physics "for the discovery of neutrino oscillations, which shows that neutrinos have mass."

As in the case of the hierarchy problem mentioned above, another fine-tuning problem seems to appear within the SM. This time, the issue is related to CP transformations, which combine the charge conjugation C with the parity P. The CP violation is allowed in the SM weak and strong interactions. Nevertheless, it has only ever been observed experimentally in the weak interaction. Consequently, the physically observable angle $\bar{\theta}$, which parametrizes the CP violation in the strong interaction, has tight experimental limits set on $\bar{\theta} \ll 10^{-10}$ [38]. The unanswered question of why $\bar{\theta}$ is so small is known as the strong CP problem.

Another pending SM shortcoming, without a doubt one of the major ones, is the fact that the SM does not include gravity. In the context of some theories maintaining the particle-like interpretation of matter and interactions, gravity is mediated by the graviton, a massless spin-2 particle which constitutes the hypothetical quantum of gravity. The SM Lagrangian does not include a kinetic or an interaction term for the graviton and, therefore, does not describe quantum gravity.

An additional limitation of the SM comes from the fact that, while it describes three generations of fermions, it does not explain why more than one exists. The origin of the multiple generations of fermions, and the specific value of three, remains an open question. A final SM shortcoming of particular relevance is the high number of free parameters of the model, 18 numerical constants with unrelated and arbitrary values. The measured masses of quarks and leptons are within the 18 free parameters but, while their experimental values accommodate well into the model, they are not explained by the model itself.

In this section, an overview of the Standard Model limitations has been presented. The cosmological constant problem, the matter-antimatter asymmetry, the nature of dark matter and dark energy, the hierarchy problem, the neutrino oscillations, the strong CP problem, and the unrefined high number of free parameters of the SM Lagrangian, have been discussed. Some of the theoretical developments on BSM physics aiming to address and solve these SM

deficiencies are presented in Chapter 3.

2.8 The Higgs boson as a probe for new physics

The Higgs boson discovery opened a new path of exploration at the LHC. The study of its properties allows us to probe BSM models. Constraints for the new physics can be established through precision measurements of the SM Higgs boson properties, but also through direct searches for new phenomena. Within the BSM models, three categories can be identified according to the main assumption made: the existence of additional BSM Higgs bosons, the exotic production of the SM-like Higgs boson, and exotic decays of the SM-like Higgs boson. In the first category, direct constraints can be obtained by searching for the additional Higgs bosons. One can also set indirect constraints on BSM models that propose an extended Higgs sector through measurements of the SM Higgs boson properties. For instance, the BSM model might predict certain modifications in some of the SM parameters, e.g., the couplings of the SM Higgs boson to SM particles may be modified. If a precise measurement of these couplings is consistent with the SM expectation, indirect constraints on the BSM model can be derived. The second category groups models in which the BSM physics may alter the production rate of the SM-like Higgs boson at the LHC. For instance, the SM-like Higgs boson may be produced in association with undetectable particles, such as DM candidate particles. In this case, the DM particles would escape detection, and the signal left in the detector would consist of a SM-like Higgs boson and missing transverse energy. This specific kind of search is referred to as mono-Higgs. The third category refers to the set of BSM models in which new physics manifests itself by affecting the decays of the SM-like Higgs boson. Rare decays of the SM-like Higgs boson may be enhanced, e.g., decays to quarkonia. Furthermore, the SM-like Higgs boson may decay to BSM particles, which subsequently decay into SM particles. It might also be the case that the SM-like Higgs boson undergoes invisible decays to BSM particles that escape detection in the experimental apparatus, such as DM candidate particles.
The Higgs boson has definitely opened new possibilities for exploring the frontiers of the SM and its possible extensions at the LHC. The upcoming data-taking periods with increased luminosity and center-of-mass-energy (Subsec. 4.1.2), will allow continuing this exploration, searching for new physics with the Higgs boson.

CHAPTER

3

EXPLORING THE EXTENDED HIGGS SECTOR: 2HDM+S AND DARK PHOTON MODEL

Contents

3.1 Description of the models

The knowledge of the theoretical models giving rise to exotic Higgs decays allows the experimentalists to interpret in a model-dependent context the model-independent results obtained from the analysis of the data. Often, only a few parameters are enough to capture the model's relevant details, e.g. the non-SM four-body Higgs decays of the form $h \to \phi\phi \to (f\bar{f})(f'\bar{f}')$ (where ϕ is a singlet and f, f' are SM fermions) can be parametrized by $m_h = 125$ GeV, m_ϕ, $\mathcal{B}(h \to \phi\phi)$, and $\mathcal{B}(\phi \to ff')$. If the decays are displaced, or there are multistep cascades, more parameters can be added. Several of the so-called simplified models are able to encapsulate the main ingredients involved in more complicated BSM models in this way. The Higgs sector can be extended by adding a scalar to the SM, one or two fermions, or a vector.

The Dark Photon Model is a particular case of SM + Vector model. Another set of simplified models that extend the Higgs sector arise through the incorporation of a Higgs doublet, the so-called two-Higgs-doublet models, with the optional addition of a scalar. More complicated models, but with similar ingredients to the simplified models, are the Minimal Supersymmetric Standard Model (MSSM) [45], the Next to Minimal Supersymmetric Standard Model (NMSSM) [46], and the Little Higgs models [47]. There is also a rich phenomenology possible in Hidden Valley models [48].
The model-independent results obtained in the analysis presented in this thesis are interpreted within the 2HDM+S and the Dark Photon Model context. Therefore, the next two subsections are dedicated to describing the basic phenomenology of these models, in particular the one related to the presence of light bosons. Details on the characteristics of the $pp \to h \to XX \to l\bar{l}l'\bar{l}'$ process within these two theoretical frameworks are provided. Furthermore, the existing collider studies and dedicated searches exploring the $\mu\mu\tau\tau$ final state are presented.

3.1.1 Two Higgs Doublet models + Scalar

The most general 2HDM consists of two Higgs doublet fields and the corresponding Higgs potential can be written as:

$$V = m_1^2|H_1|^2 + m_2^2|H_2|^2 + \frac{\lambda_1}{2}|H_1|^2 + \frac{\lambda_2}{2}|H_2|^2 + \lambda_3|H_1|^2|H_2|^2 + \lambda_4|H_1^\dagger H_2|^2 + \tag{3.1}$$

$$\frac{\lambda_5}{2}\left((H_1H_2)^2 + \text{c.c.}\right) + m_{12}^2(H_1H_2 + \text{c.c.}) + \left(\lambda_6|H_1|^2(H_1H_2) + \text{c.c.}\right)$$

$$+ \left(\lambda_7|H_2|^2(H_1H_2) + \text{c.c.}\right).$$

The hypercharges of the Higgs fields are $-1/2$ for H_1 and $+1/2$ for H_2 [26]. The two scalar doublets expanded around their respective minima take the form:

$$H_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} v_1 + H_{1,R}^0 + iH_{1,I}^0 \\ H_{1,R}^- + iH_{1,I}^- \end{pmatrix} \qquad H_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} H_{2,R}^+ + iH_{2,I}^+ \\ v_2 + H_{2,R}^0 + iH_{2,I}^0 \end{pmatrix}. \tag{3.2}$$

The mass matrices for the charged scalar and pseudoscalar are diagonalized by a rotation angle β, with $\tan\beta = v_2/v_1$. The three real degrees of freedom left after EWSB yield one neutral pseudoscalar mass eigenstate (A), and two neutral scalar mass eigenstates (h, H^0):

$$A = H_{1,I}^0 \sin\beta - H_{2,I}^0 \cos\beta, \qquad \begin{pmatrix} h \\ H^0 \end{pmatrix} = \begin{pmatrix} -\sin\alpha & \cos\alpha \\ \cos\alpha & \sin\alpha \end{pmatrix}\begin{pmatrix} H_{1,R}^0 \\ H_{2,R}^0 \end{pmatrix}. \tag{3.3}$$

The scalar mass eigenstates are defined in terms of the real components of the doublets by the rotational angle α, with $-\pi/2 \leq \alpha \leq \pi/2$. The phenomenology of the 2HDMs is governed by the two parameters $\tan\beta$ and α, which determine the interactions of the various Higgs fields with the vector bosons and fermions. $\mathbb{Z}_2$ symmetries are imposed in order to avoid large

Flavor-Changing Neutral Currents (FCNCs), ensuring that fermions with the same quantum numbers couple to only one Higgs field. As a result of this requirement, four standard types of fermion couplings become available. The fermion Higgs couplings for each of the four types of 2HDMs are summarized in Tab. 3.1: Type I (all fermions couple to H_2), Type II (MSSM-like, d_R and e_R couple to H_1, u_R to H_2), Type III (lepton-specific, leptons and quarks couple to H_1 and H_2, respectively), and Type IV (flipped, with u_R, e_R coupling to H_2 and d_R to H_1).

Model	2HDM I	2HDM II	2HDM III	2HDM IV
u	H_2	H_2	H_2	H_2
d	H_2	H_1	H_2	H_1
e	H_2	H_1	H_1	H_2

Table 3.1: Fermion Higgs couplings for the 2HDMs Type I-IV [26].

From the two real mass eigenstates in Eq. (3.3), one is identified as the SM-like Higgs boson (h) and the other is taken to be heavy (H^0), which is easily achieved in the decoupling limit ($\alpha \to \pi/2 - \beta$). Tab. 3.2 lists all the couplings to fermions and gauge fields relative to the SM Higgs couplings of h, H_0, and the A mass eigenstates.

The complex scalar singlet added to the 2HDMs:

	Couplings	2HDM I	2HDM II	2HDM III	2HDM IV
h	g_{hVV}	$\sin(\beta-\alpha)$	$\sin(\beta-\alpha)$	$\sin(\beta-\alpha)$	$\sin(\beta-\alpha)$
	$g_{ht\bar{t}}$	$\cos\alpha/\sin\beta$	$\cos\alpha/\sin\beta$	$\cos\alpha/\sin\beta$	$\cos\alpha/\sin\beta$
	$g_{hb\bar{b}}$	$\cos\alpha/\sin\beta$	$-\sin\alpha/\cos\beta$	$\cos\alpha/\sin\beta$	$-\sin\alpha/\cos\beta$
	$g_{h\tau\bar{\tau}}$	$\cos\alpha/\sin\beta$	$-\sin\alpha/\cos\beta$	$-\sin\alpha/\cos\beta$	$\cos\alpha/\sin\beta$
H^0	g_{H^0VV}	$\cos(\beta-\alpha)$	$\cos(\beta-\alpha)$	$\cos(\beta-\alpha)$	$\cos(\beta-\alpha)$
	$g_{H^0t\bar{t}}$	$\sin\alpha/\sin\beta$	$\sin\alpha/\sin\beta$	$\sin\alpha/\sin\beta$	$\sin\alpha/\sin\beta$
	$g_{H^0b\bar{b}}$	$\sin\alpha/\sin\beta$	$\cos\alpha/\cos\beta$	$\sin\alpha/\sin\beta$	$\cos\alpha/\cos\beta$
	$g_{H^0\tau\bar{\tau}}$	$\sin\alpha/\sin\beta$	$\cos\alpha/\cos\beta$	$\cos\alpha/\cos\beta$	$\sin\alpha/\sin\beta$
A	g_{AVV}	0	0	0	0
	$g_{At\bar{t}}$	$\cot\beta$	$\cot\beta$	$\cot\beta$	$\cot\beta$
	$g_{Ab\bar{b}}$	$-\cot\beta$	$\tan\beta$	$-\cot\beta$	$\tan\beta$
	$g_{A\tau\bar{\tau}}$	$-\cot\beta$	$\tan\beta$	$\tan\beta$	$-\cot\beta$

Table 3.2: Couplings of the neutral scalar and pseudoscalar mass eigenstates, normalized to the SM Higgs boson couplings, for the 2HDMs Type I-IV with a $\mathbb{Z}_2$ symmetry [26].

$$S = \frac{1}{\sqrt{2}}(S_R + iS_I), \tag{3.4}$$

only couples to the two 2HDM Higgs fields ($H_{1,2}$) in the potential, with no direct Yukawa couplings. The resulting physical states acquire all of its couplings to SM fermions through its mixing with $H_{1,2}$. In order to preserve the SM-like nature of h, this mixing needs to be small. Assuming that the 2HDM is near or in the decoupling limit and that the mixing of the singlet to the doublets is small, exotic Higgs decays of the form:

$$h \to ss \to X\bar{X}Y\bar{Y}, \quad h \to aa \to X\bar{X}Y\bar{Y}, \quad h \to aZ \to X\bar{X}Y\bar{Y}, \tag{3.5}$$

are possible. The SM-like Higgs h couples to a singlet-like scalar s or pseudoscalar a, where $s(a)$ is a (pseudo)scalar mass eigenstate mainly composed of the real (imaginary) parts of the singlet field, and X, Y are SM fermions or gauge bosons. This setup, referred to as the 2HDM+S, opens the possibility for a large variety of searches with non-standard four-body final states.

For a given type of 2HDM, the exotic Higgs decay phenomenology is governed by the exotic branching ratios $\mathcal{B}(h \to aa)$ and $\mathcal{B}(h \to Za)$, as well as tanβ, which dictates the fermion couplings to a. Fig. 3.1 presents the branching ratios $\mathcal{B}(a \to X\bar{X})$, corresponding to the decay of the light pseudoscalar into a pair of SM particles for the Type III 2HDM+S, illustrating a rich decay phenomenology. For the Type I, the exotic decay branching ratios are independent

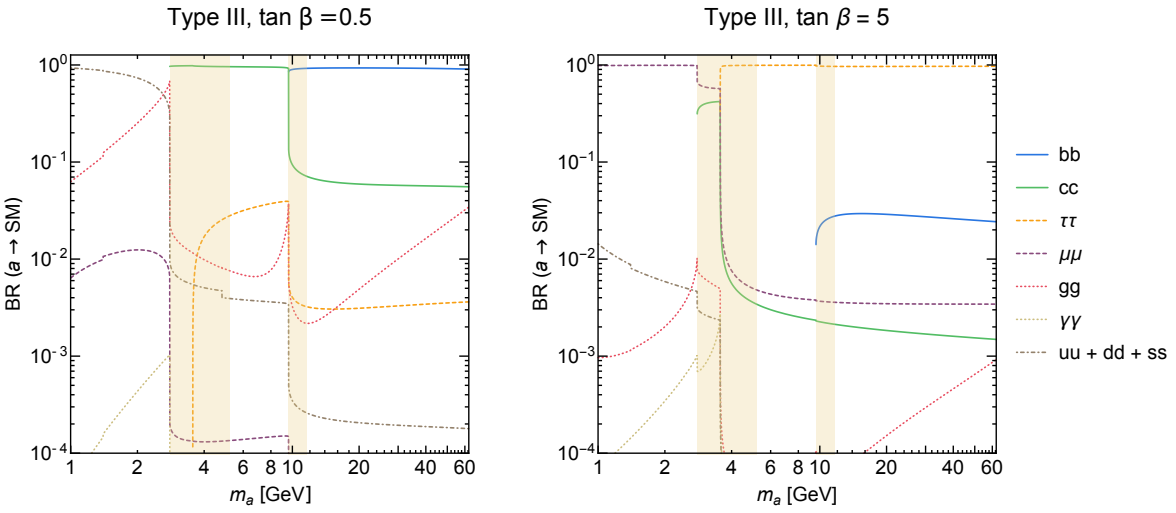


Figure 3.1: Branching ratios of a singlet-like pseudoscalar in the 2HDM+S for Type III Yukawa couplings. Decays to quarkonia probably invalidate theoretical calculations in the shaded regions [26].

of tanβ as a result of the fermions only coupling to H_2, while the couplings of the pseudoscalar to all fermions are proportional to the corresponding ones of the SM Higgs boson, with the same proportionality constant. Unlike the Type I models, the branching ratios depend on tanβ for Type II, III, and IV models. In the Type II models, the decays to down-type fermions are suppressed (enhanced) for tan$\beta < 1$ (tan$\beta > 1$). In the Type III models, the decays to quarks (leptons) are favored over the decays to leptons (quarks) for tan$\beta < 1$ (tan$\beta > 1$). Therefore, for tanβ higher than one, decays to $\tau^+\tau^-$ dominate over decays to $b\bar{b}$ above the $b\bar{b}$ threshold and $\mu^+\mu^-$ can dominate over decays to heavier, kinematically allowed quarks, as it can be seen in the right plot of Fig. 3.1. The $2b2\tau$ or $2c2\tau$ final states are specially sensitive to the Type IV models, for tan$\beta < 1$ since the branching ratio to up-type quarks and leptons

can be enhanced with respect to down-type quarks, resulting in similar branching ratios to $b\bar{b}$, $c\bar{c}$, and $\tau^+\tau^-$.

3.1.2 Dark Photon Model

Possible states almost decoupled from the SM particles constitute the so-called *hidden sector*. The leading interactions of the hidden sector with the SM sector may be through the hypercharge portal, via the kinetic mixing, or through the Higgs portal, via the Higgs mixing. The kinetic mixing is driven by the parameter ϵ and the Higgs mixing by the parameter κ. The introduction of a hidden sector and a very weak interaction with the SM fermions could explain some of the SM shortcomings discussed in Sec. 2.7, such as the antimatter excess in cosmic rays. The simplest hidden sector would be considering an additional $U(1)_{\mathrm{D}}$ gauge symmetry with a massive interaction carrier, a vector boson called dark photon and denoted as Z_{D}. The model is defined by the $U(1)_{\mathrm{D}}$ gauge sector and a SM singlet S with unit charge under $U(1)_{\mathrm{D}}$. The new physics contribution can be represented by the addition of one term to the Lagrangian in Eq. (2.32) [49], leading to:

$$\mathcal{L} = \mathcal{L}_{\mathrm{SM}} + \mathcal{L}_{\mathrm{D}}, \quad \text{with} \quad \mathcal{L}_{\mathrm{D}} = \mathcal{L}_{\mathrm{D}}^{\mathrm{KE}} + \mathcal{L}_{\mathrm{D}}^{H_{\mathrm{D}}}. \tag{3.6}$$

The only coupling to the SM of the new gauge sector with vector field X_μ is through kinetic mixing with the hypercharge gauge boson B_μ. The kinetic mixing coefficient ϵ determines the strength of the coupling of Z_{D} with the SM fermions. Two terms contribute to the kinetic energy term of $U(1)_{\mathrm{D}}$ so that $\mathcal{L}_{\mathrm{D}}^{\mathrm{KE}}$ takes the form:

$$\mathcal{L}_{\mathrm{D}}^{\mathrm{KE}} = -\frac{1}{4} X_{\mu\nu} X^{\mu\nu} + \frac{\epsilon}{2} X_{\mu\nu} B^{\mu\nu}, \tag{3.7}$$

where ϵ must be very small in order to keep consistency with the constraints from precision electroweak measurements. The dark Higgs sector Lagrangian ($\mathcal{L}_{\mathrm{D}}^{H_{\mathrm{D}}}$) takes the form:

$$\mathcal{L}_{\mathrm{D}}^{H_{\mathrm{D}}} = (D_\mu H_{\mathrm{D}})^\dagger (D^\mu H_{\mathrm{D}}) + \mu_{\mathrm{D}}^2 H_{\mathrm{D}}^\dagger H_{\mathrm{D}} - \lambda_{\mathrm{D}} (H_{\mathrm{D}}^\dagger H_{\mathrm{D}})^2 - \kappa H^\dagger H H_{\mathrm{D}}^\dagger H_{\mathrm{D}}. \tag{3.8}$$

The Higgs potential is minimized by the vacuum expectation values of the SM Higgs boson (H_0) and the dark Higgs (S):

$$H^0 = \frac{1}{\sqrt{2}}(h + v), \quad S = \frac{1}{\sqrt{2}}(s + w). \tag{3.9}$$

The non-zero vacuum expectation value acquired by S gives mass to Z_{D}. The connection between the dark and the SM sectors is given by the gauge kinetic mixing and the Higgs mixing, with the phenomenology of the model depending on the dominant mixing. The symmetry breaking of $U(1)_{\mathrm{D}}$ may lead to exotic Higgs decays, particularly if there is a mixing between the two Higgs sectors. The exotic Higgs boson decays can be through the Higgs portal with a Higgs-to-dark-Higgs mixing ($h \to Z_{\mathrm{D}} Z_{\mathrm{D}}$), as depicted in Fig. 3.2, or through the hypercharge portal Z-Z_{D} mass mixing ($h \to Z Z_{\mathrm{D}}$), which also drives the direct production of the dark photon in Drell-Yan events ($pp \to Z_{\mathrm{D}} \to l^+ l^-$).
The decay branching ratios of Z_{D} are ordered by the gauge couplings instead of the Yukawa couplings. Thus, decays to e^+e^- and $\mu^+\mu^-$ continue to be large above the τ threshold. If there are no hidden-sector states below the dark photon mass ($m_{Z_{\mathrm{D}}}$), which is a free parameter

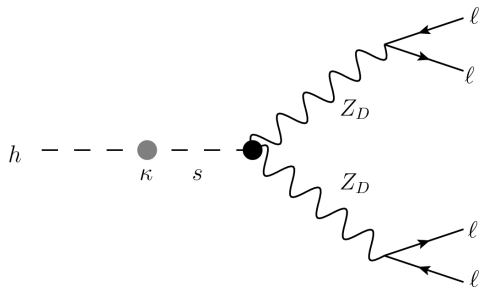


Figure 3.2: Exotic Higgs boson decay to four leptons induced by intermediate dark vector bosons via the Higgs portal, where s is a dark Higgs boson. The Z_{D} gauge bosons decay to SM particles through kinetic mixing with the hypercharge field or through mass mixing with the Z boson. The $HZ_{\mathrm{D}}Z_{\mathrm{D}}$ vertex factor is proportional to κ [50].

of the model, the dark photon will only decay to SM particles. The lowest Leading Order dark photon decay width to fermions can be written as:

$$\Gamma(Z_{\mathrm{D}} \to \bar{f}f) = \frac{N_c}{24\pi m_{Z_{\mathrm{D}}}} \sqrt{1 - \frac{4m_f^2}{m_{Z_{\mathrm{D}}}^2}} (m_{Z_{\mathrm{D}}}^2 (g_L^2 + g_R^2) - m_f^2(-6g_L g_R + g_L^2 + g_R^2)), \tag{3.10}$$

where $g_{L,R} = g_{Z_{\mathrm{D}} f_{L,R} \bar{f}_{L,R}}$, correspond to the $Z_{\mathrm{D}} f_{L,R} \bar{f}_{L,R}$ interaction. The $g_{L,R}$ terms are proportional to ϵ for $\epsilon \ll 1$. Eq. (3.10) constitutes a good approximation for dark photon masses above the $b\bar{b}$ threshold, but higher order QCD calculations and experimental information are needed in order to obtain consistent predictions across the entire mass range. From Eq. (3.10), the leptonic branching fraction:

$$\mathcal{B}(Z_{\mathrm{D}} \to ll) = \frac{\Gamma(Z_{\mathrm{D}} \to ll)}{\Gamma_{Z_{\mathrm{D}}}}, \tag{3.11}$$

can be obtained. Fig. 3.3 depicts the resulting branching ratios as a function of the dark photon mass. The branching ratios are roughly dependent on the square of the electric charge of the decay products and, therefore, leptons with charge -1 are very common.

3.2 General Motivation to Search for Exotic Higgs Decays

Exotic Higgs decays remain a well-motivated possibility, even after the discovery of a Higgs particle consistent with the SM expectations. At present, the search for exotic Higgs decays constitutes an important component of the LHC physics program. The upper limit at 95% CL on the branching fraction into exotic decay modes of the Higgs boson is currently set at 34%, from ATLAS and CMS Run I measurements of the Higgs boson production, decay rates, and constraints on its couplings [17]. In this joint analysis, a parameterization allowing contributions from BSM particles was made, resulting in the quoted upper limits. The dataset recorded by CMS during Run II might contain $\mathcal{O}(500,\,000)$ exotic Higgs decays, assuming a $\mathcal{B}(h \to \mathrm{BSM})$ of 10%. Thus, if an acceptable trigger efficiency for the final states of the exotic decays is achieved, dedicated searches offer a large discovery potential. The value of 10% is considered taking into account projections of the indirect measurement on Higgs coupling fits, which indicate that the reachable precision at the LHC on the $\mathcal{B}(h \to \mathrm{BSM})$

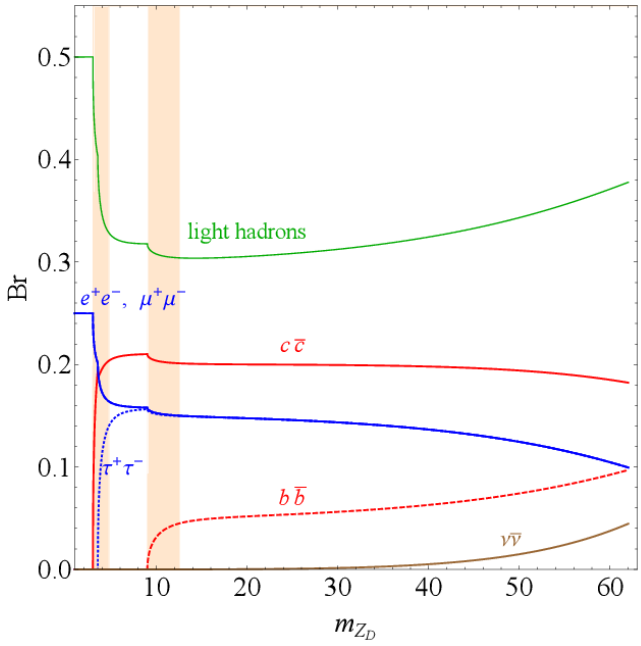


Figure 3.3: Branching ratios for dark photon decay at leading order and without QCD corrections from Ref. [26]. The label light hadrons refers to hadrons containing only up, down, and strange quarks. Decays of the dark photon to the dark sector are assumed to be kinematically forbidden. The simple theoretical calculations are probably invalidated in the shaded regions, due to missing threshold effects and QCD corrections, as well as the presence of hadronic resonances in these regions.

would be of $\mathcal{O}(5-10\%)$. Thus, branching fractions of $\mathcal{O}(10\%)$ into exotic decay modes are still allowed and will continue to be a reasonable target for the duration of the LHC physics program. Tab. 3.3 lists the number of exotic Higgs decays that could be contained in the datasets collected by CMS in Run II, Run III, and the High Luminosity LHC, according to the luminosity recorded or expected to be recorded during these data-taking periods. The numbers are reported separately for each Higgs production mechanism. The corresponding cross-section values for the centers of mass energy of 13 and 14 TeV are listed in Tab. 2.4.

Due to the tiny width of the SM Higgs boson (Subsec. 2.6.2), even extremely small couplings of the Higgs boson to new BSM particles can lead to potential signals that would be detectable at the LHC. The non-targeted analysis might be a possibility to constrain the large variety of exotic Higgs decay modes. However, typical LHC exotica searches apply cuts on the p_T threshold of the objects at analysis level, leaving decay modes with low p_T final states largely unconstrained. This is the case of the four-body exotic cascade decays, in which the characteristic p_T of the daughter particles in the dominant gluon fusion process is smaller than the usual analysis cuts applied. Therefore, *dedicated searches* are needed to discover or constrain the broad class of BSM theories. How sensitive to certain probed exotic decay a dedicated search is, heavily depends on the triggering strategy adopted.

Production	$\sqrt{s} = 13$ TeV		$\sqrt{s} = 14$ TeV	
	$N_{\text{ev}}^{10\%},\ 137.2\,fb^{-1}$	$N_{\text{ev}}^{10\%},\ 300\,fb^{-1}$	$N_{\text{ev}}^{10\%},\ 300\,fb^{-1}$	$N_{\text{ev}}^{10\%},\ 3000\,fb^{-1}$
ggF	666,792	1,458 $\times 10^3$	1,641$\times 10^3$	1,641$\times 10^4$
VBF	51,862	113,400	128,400	1,284$\times 10^3$
$hW^{\pm}$	18,796	41,100	45,300	453$\times 10^3$
$hW^{\pm}(l^{\pm}\nu)$	3,947	8,631	9,513	95,130
hZ	12,074	26,400	29,700	297$\times 10^3$
$hZ(l^{+}l^{-})$	809	1,769	1,990	19899
$t\bar{t}H$	6,860	15$\times 10^3$	18$\times 10^3$	297$\times 10^3$

Table 3.3: The number of exotic Higgs decays in LHC data, for Run II (137.2 fb^{-1}) at 13 TeV, Run III (300 fb^{-1}) at 13 TeV, Run III (300 fb^{-1}) at 14 TeV, and the High Luminosity LHC (3000 fb^{-1}) at 14 TeV, for the main production mechanisms: gluon-gluon fusion, vector-boson fusion, associated production ($hW^{\pm}$ and hZ), and associated production with a pair of top quarks, assuming the Standard Model production cross-section of a 125 GeV Higgs boson and a branching ratio $\mathcal{B}(h \to \text{BSM})$ of 10%.

3.3 Exotic Decay Modes of the 125 GeV Higgs Boson

Exotic decay modes of the SM Higgs boson can be obtained considering three main assumptions involving the observed Higgs at 125 GeV. First, that it decays to new particles beyond the SM; second, that it is responsible for the breaking of the electroweak symmetry and third, that the initial exotic decay is to two neutral BSM particles. Thus, the decay starts via the two-body process $h \to X_1 X_2$, where X_1 and X_2 can be identical BSM states. Many different exotic Higgs decay modes are possible depending on the properties of X_1 and X_2.

3.3.1 $h \to 2 \to 4$ decay topology

The cascade topology $h \to 2 \to 4$, depicted in Fig. 3.4, occurs in theories featuring additional singlet scalars, vector fields, in 2HDM+S and Little Higgs Models. The Higgs decays as $h \to aa'$, ss', V_1V_2, $aV_1 \to (xx)(yy)$, where a and a' (s and s', V_1 and V_2) can be either equal or different pseudoscalars (scalars, vectors). For this cascade topology, it is usually possible to reconstruct two resonances out of the (xx) (yy) systems, with $x, y =$ quarks, leptons, photons, or gluons for the case of scalar and pseudoscalars, and $x, y =$ quarks or leptons for vectors. The final state probed in the analysis presented in this thesis with two muons and two tau leptons corresponds to this topology and is further discussed in the following subsection.

3.3.2 $h \to aa(Z_{\mathbf{D}}Z_{\mathbf{D}}) \to \mu\mu\tau\tau$

The $h \to aa \to 2\mu 2\tau$ final state can arise in the set of 2HDM+S, as already discussed in Subsec. 3.1.1. Within the Type II 2HDM+S, a light a can correspond to the R-symmetry limit

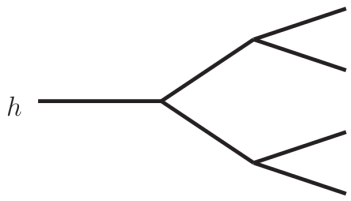


Figure 3.4: The exotic Higgs decay topology $h \to 2 \to 4$.

of the NMSSM, a kind of symmetry in which the generator has a non-trivial commutation with the fermionic generators. In the Type III 2HDMs, with or without the addition of an extra singlet field, the leptonic decays are enhanced for large values of $\tan\beta$ and will, therefore, dominate for new scalar or pseudoscalar states of nearly all masses. The main assumption besides the mass range of m_a is that the couplings of a are directly proportional to the lepton masses. Thus, the branching fractions to lepton pairs above the tau pair threshold have the following proportion: $\tau^+\tau^- : \mu^+\mu^- : e^+e^- \simeq m_\tau^2 : m_\mu^2 : m_e^2 \simeq 1 : 3.5 \times 10^{-3} : 8 \times 10^{-8}$. The dominant $2 \to 4$ fully leptonic branching fraction is to 4τ, with a roughly 1% relative branching ratio to $2\mu 2\tau$. The $h \to Z_\mathrm{D} Z_\mathrm{D} \to 2\mu 2\tau$ final state may arise in models as the Dark Photon Model. A $\mathcal{B}(h \to Z_\mathrm{D} Z_\mathrm{D}) \approx 10\%$ would also be possible in 2HDM+S models where one of the two Higgs doublets and the SM singlet is charged under $U(1)_\mathrm{D}$. In this subsection, the scenarios in which the Higgs decays into a pair of pseudoscalar bosons a or vector bosons Z_D, with an emphasis in the subsequent decay to a pair of muons and a pair of tau leptons, are considered.

Experimental signature

The dominant and sub-dominant branching fraction to leptons result in the 4τ and $2\mu 2\tau$ final states. With a mass of 1.777 GeV, the tau lepton is the only lepton heavy enough to decay into hadrons. The purely leptonic tau decays are to electrons and muons while the hadronic decays are typically to either one or three charged pions or kaons and up to two neutral pions (π_0) and one neutrino (ν_τ). This results in a large number of channels, in which the Higgs mass energy gets distributed between all the final-state particles, many of which are invisible neutrinos. Thus, even though each event is triply-resonant (H(125) and the two light bosons), the neutrinos from the tau decays complicate the reconstruction of the mass of at least one of the two a's or Z_D's and consequently of the 125 GeV Higgs boson. The lack of a resonance peak and the complicated triggering due to the low transverse momentum for electrons and muons constitute the major challenges of the 4τ final state. These difficulties unveil the major advantage of the $2\mu 2\tau$ channel with respect to 4τ. Despite the low branching fraction, the $2\mu 2\tau$ channel benefits from a clean final state with a narrow dimuon resonance, and it contains high p_T muons, which facilitates the triggering. In the so-called *boosted topologies* ($m_{a,\,Z_\mathrm{D}} \ll m_h$), the two tau leptons or muons from an individual a or Z_D decay can merge under standard isolation criteria. Therefore, special reconstruction techniques are needed to disentangle the merged objects. The difficulties associated with the τ reconstruction in the 4τ final state and the low rates of the $2\mu 2\tau$ final state made both final states hard to constrain until the reconstruction techniques were improved, the analysis strategies polished, and a significant amount of data was collected during the first two runs of the LHC. Considering both the experimental and theoretical side, viable search strategies

came about in the so-called *collider studies*. The next subsection is dedicated to present an overview of the collider studies that motivated the realization of dedicated searches for light bosons in the $2\mu 2\tau$ final state.

Collider Studies

The proposals for $h \to aa$ dedicated searches have exploited the $2a$ decay channels with one or more leptons in the final state. Ref. [51] studies the prospects of the $2\mu 2\tau$ channel, proposing a search strategy focused on the identification of the 2μ resonance and considering only the hadronic decays of the τ lepton. The two closeby hadronic taus are treated as a single jet with aligned missing transverse energy, composing a jet-like object characterized by a low track activity and a distinctive calorimeter pattern. The mass of the Higgs boson is approximately reconstructible but the Higgs resonance is not used for discrimination. The study estimated a 2σ sensitivity to $\mathcal{B}(h \to aa) < 10\%$ via ggF production, for a 125 GeV Higgs boson, $m_a = 7$ GeV, and 5 fb^{-1} of data taken at a center-of-mass-energy of 14 TeV. The D0 collaboration performed a search with the proposed strategy, providing the first limits from a dedicated search in the $2\mu 2\tau$ channel [52]. The most relevant details of this pioneering analysis are provided in the next subsection. Once the LHC started running and with the constraints from the analysis performed by the D0 collaboration already available, a preliminary collider study on the discovery potential at the LHC was done [26]. As an outcome of the study, it was suggested to supplement the D0 search by exploiting the final states with 3 and 4 leptons, which result in analysis channels with extra-low backgrounds. These final states were considered in the currently available CMS [53–56] and ATLAS [57,58] results for the $2\mu 2\tau$ channel.
The viability of probing $h \to Z_{\mathrm{D}} Z_{\mathrm{D}} \to 4l$ at Tevatron and the LHC was assessed in [49], prior to the Higgs boson discovery in 2012. An estimation for several benchmark scenarios of the LHC reach operating at 14 TeV was done. Among the scenarios considered, two scenarios labeled as "A" and "B", feature a Higgs mass $m_h = 120$ GeV and a dark photon mass $m_{Z_{\mathrm{D}}} =$ 5 (50) GeV, respectively. The study concluded that the prospects to observe this exotic decay were very good, even for small values of the branching ratio.

Experimental Searches and Limits

The first dedicated search for the Higgs boson production followed by the h $\to aa$ decay in the $2\mu 2\tau$ final state is based on a $p\bar{p}$ collision data sample corresponding to an integrated luminosity of 4.2 fb^{-1}, collected with the D0 detector at the Tevatron Collider during the Run II data-taking period at a center-of-mass energy of 1.96 TeV. The search is a bump hunt in the mass spectrum of the dimuon pair, over the range of the light boson between 3.6 and 19 GeV. The signal signature is either two pairs of collinear muons or one pair of collinear muons and large missing transverse momentum ($\not\!\!E_T$), accompanied by an additional muon or a loosely isolated electron from the $a \to \tau\tau$ decay. The majority of the events pass a dimuon trigger, with muon p_T thresholds of 4 and 6 GeV. Since the muon system of the D0 detector did not have sufficient granularity to reconstruct the two nearby muons reliably, the muon identification is relaxed for one of the muons of the $a \to \mu\mu$ candidate. Nevertheless, the inner track of this muon can still be reconstructed. For both muons, the track reconstructed from the hits in the muon system is required to match the track in the inner tracker. As suggested in [51], the $a \to \tau\tau$ leg is only loosely identified, through the requirement of large

missing transverse momentum near a low track multiplicity jet. Fig. 3.5 depicts the signal topology of the analysis.

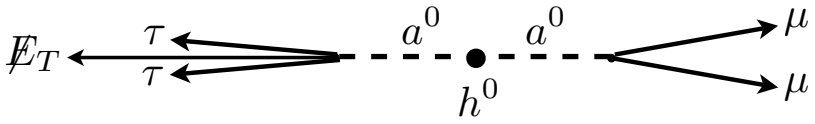


Figure 3.5: Illustration of the signal topology, with highly boosted nearly collinear muons and tau leptons. Most of the $\not\!\!E_T$ of the event is found in the direction of the tau leptons, reconstructed as one jet [51].

A few non-dedicated searches at the LHC with multilepton final states have some sensitivity to the $2\mu 2\tau$ final state. They were used to derive non-trivial constraints by reinterpreting and combining the individual results. These limits served as a starting point for the dedicated searches that came after, performed by the ATLAS and CMS collaborations with the dataset corresponding to the Run I data-taking period, in the mass ranges $5 < m_a < 62.5$ GeV [53] and $3.7 < m_a < 50$ GeV [57], respectively. More results have become available with a partial Run II dataset, corresponding to the 2016 data-taking period. CMS has explored the mass ranges $15 < m_a < 62.5$ GeV [54], $4 < m_a < 15$ GeV [55] and $3.6 < m_a < 21$ GeV [56] in three independent searches.
The analysis presented in this thesis constitutes the first search in the $2\mu 2\tau$ final state based on a data sample corresponding to an integrated luminosity of 137.2 fb^{-1}, collected with the CMS detector at the CERN Large Hadron Collider during the full Run II data-taking period at a center-of-mass energy of 13 TeV. The 2016 CMS analysis, probing masses of the light boson between 15 and 62.5 GeV, has been projected to integrated luminosities of up to 3000 fb^{-1}, expected at the High-Luminosity LHC [59]. The projection assumes that the Run II object reconstruction performance can be maintained with the upgraded CMS detector, operating under more demanding conditions due to the foreseen increase in luminosity. The effect of the difference in center-of-mass energy between the LHC and its high luminosity successor (13 vs 14 TeV) is neglected. The event yields are scaled to the integrated luminosity of 3000 fb^{-1} and two scenarios are considered in the treatment of the systematic uncertainties. In the so-called Run II systematic uncertainties scenario, the experimental and theoretical uncertainties keep the same values from the Run II analysis. In the YR18 systematics uncertainties scenario, an improvement by a factor of two of the theoretical uncertainties is considered and the experimental uncertainties are assumed to scale with the square root of the integrated luminosity until certain lower limit is reached. The uncertainties related to the limited size of the simulated samples are neglected. For both scenarios, the statistical uncertainty in the measurement is reduced by a factor $1/\sqrt{R_L}$, where R_L is the ratio of the projection of the integrated luminosity to the luminosity of the reference Run II analysis. The improvement in the sensitivity scales inverse-proportionally to the luminosity for low values of m_a and to the square root of the luminosity for high values of m_a. The difference in upper limits between the two scenarios considered for the systematic uncertainties reaches at most 5%, with the largest difference for high values of m_a.
The limits from CMS searches at 7 TeV [60], 8 TeV [61], and 13 TeV [62] in the $h \to aa \to 4\mu$ final state, which probed the mass range between 250 MeV and $2m_\tau$, are directly applicable to the $h \to Z_D Z_D \to 4\mu$ final state, since a significant difference in acceptance between h

$\rightarrow aa \rightarrow 4\mu$ and h $\rightarrow Z_{\mathrm{D}}Z_{\mathrm{D}} \rightarrow 4\mu$ is not expected within the probed mass range. For the mass range $4 < m_{Z_{\mathrm{D}}} < m_h/2$ GeV, CMS results are available in the four-lepton final state with $l = e, \mu$ ($4e$, 4μ, and $2e2\mu$ channels) [63]. The ATLAS collaboration also performed a search in the four-lepton final state at a center-of-mass-energy of 8 TeV in the mass range $15 < m_{Z_{\mathrm{D}}} < m_h/2$ GeV. This result was updated with a partial Run II dataset [58], building up from the experience of the 8 TeV analysis. Limits can also be derived from non-dedicated SM Higgs searches and ZZ cross-section measurements. The final states with τ leptons, $2\mu2\tau$ and 4τ, were not included in the ATLAS and CMS searches with four leptons, since the difficulties in the τ reconstruction together with the similar branching fraction of the leptonic final states for masses above 3.5 GeV (Subsec. 3.3.2, Subsec. 3.1.2), make the leptonic final states $4e$, 4μ, and $2e2\mu$ more suitable to provide stringent constraints for the Dark Photon Model.
The analysis presented in this thesis aims to provide a glance at the sensitivity of the $\mu\mu\tau\tau$ final state in the context of the Dark Photon Model and motivates the recent CMS dedicated search with Run-II dataset in the four-lepton final state with $l = e, \mu$ [63]. Furthermore, it allows a comparison of the results obtained with full Run II dataset and a multivariate analysis (MVA) approach, with the ATLAS results obtained with a partial Run II dataset and a cut-based approach, but more suitable four-lepton final states. Even with the 3000 fb^{-1} of data expected to be collected at the High-Luminosity LHC, the $\mu\mu\tau\tau$ final state will be statistically limited. Thus, there is still sufficient room left for further investigations in this channel and, in general, for exotic decays of the 125 Higgs boson into a pair of light pseudoscalars bosons a or vector bosons Z_{D}. A promising opportunity for this kind of search might come from future lepton colliders in which the main production mechanism of the SM-like Higgs boson (Zh) would result in a reduction of the so-called color backgrounds. Hence future lepton colliders such as the Circular Electron Positron Collider (CEPC), Future Circular Collider e^+e^- (FCC-ee), and the International Linear Collider (ILC) have the potential to become a powerful tool for the detection of exotic decays.

CHAPTER

4

THE CMS EXPERIMENT AT THE CERN LARGE HADRON COLLIDER

Contents

In this chapter, the LHC machine layout and performance are presented, followed by a brief introduction to the main detectors located in the LHC ring. A detailed description of the CMS detector used to collect the analyzed data is given. Key aspects of the data acquisition and offline computing systems are also discussed.

4.1 The LHC Machine

Colliders, machines where counter-circulating beams collide, present a big advantage over accelerators with beams directed to a stationary target, since for colliders the energy of the

collision is the sum of the energies of the two beams:

$$\sqrt{s} = E, \quad \text{where} \quad E = 2 \cdot E_{\text{beam}}. \tag{4.1}$$

The Large Hadron Collider, located underground in the border between France and Switzerland, is a two-ring-superconducting-hadron accelerator, at present the most powerful accelerator ever built [7]. It was installed in the already existing tunnel built at CERN for the operation of the Large Electron-Positron collider (LEP) machine [64], which was active from 1989 to 2000. The LHC is the last ring in a complex chain of particle accelerators, as shown in Fig. 4.1 (dark blue line). The purpose of the smaller machines is to consecutively increase the particles' energy up to the targeted final energy. At the same time, they also provide beams to smaller experiments, located outside of the LHC ring. A brief description of the main detectors currently in operation at the LHC and its physics goals is given in the next segment.

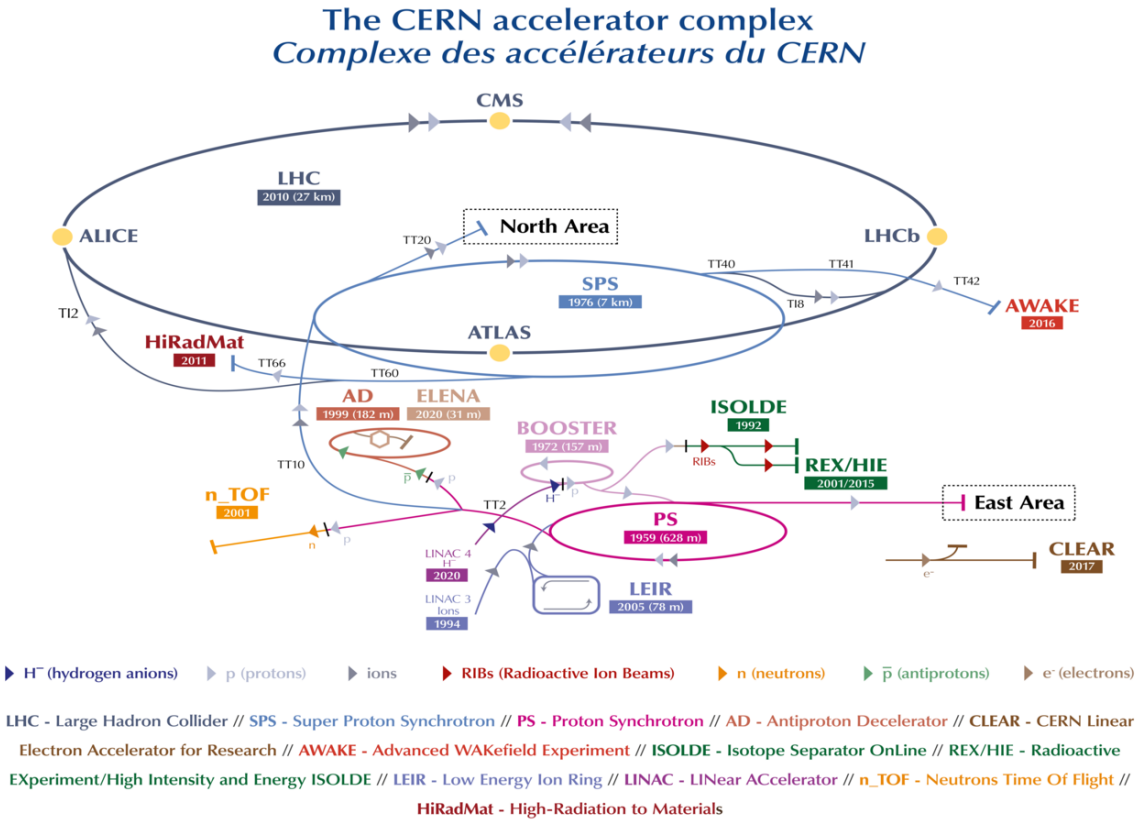


Figure 4.1: CERN accelerator complex [65].

4.1.1 Main experiments and their physics goals

Four main detectors have been constructed at the LHC: CMS, ATLAS (A Toroidal LHC Apparatus) [66], LHCb (LHC b-hadron experiment) [67], and ALICE (A Large Ion-Collider Experiment) [68]. These are complemented by three smaller experiments: LHCf (LHC forward experiment) [69], TOTEM (ToTal Elastic and diffractive cross-section Measurement) [70],

and MoEDAL (Monopole and Exotics Detector At the LHC) [71].
The two high luminosity general purpose particle detectors are CMS and ATLAS, both designed to reach a peak luminosity of L = 10^{34} $cm^{-2}\,s^{-1}$ for proton operation, already exceeded during the Run II data-taking period [72–74]. Their broad physics program ranges from studying the Standard Model to searching for extra dimensions and dark matter candidates. ALICE and LHCb, on the other side, have detectors specialized in the study of specific phenomena. LHCb is dedicated to precision measurements of CP violation and rare decays of b-hadrons. ALICE is a heavy-ion detector, aiming to study the physics of strongly interacting matter at extreme energy densities and temperature, where the quark-gluon plasma is formed. It is designed to reach a peak luminosity of L = 10^{27} $cm^{-2}\,s^{-1}$ for nominal lead-lead ion operation. The smaller LHCf and TOTEM experiments focus on forward physics. LHCf uses the particles thrown very forward by collisions in the LHC as a source for the simulation of cosmic rays in laboratory conditions. The collected data serve as input in the calibration of hadron interaction models used in the study of extremely high-energy cosmic-rays. LHCf has two detectors, along the LHC beamline, 140 meters either side of the ATLAS collision point. In a similar configuration, TOTEM experiment uses detectors positioned on either side of the CMS interaction point. TOTEM is designed to reach a peak luminosity of L = 2 $\cdot 10^{29}$ $cm^{-2}\,s^{-1}$ and is able to make precise measurements of protons emerging from collisions at small angles. It focuses on the study of elastic and diffractive scattering and measurements of the total pp cross-section.

4.1.2 Machine performance

The LHC accelerates two beams of particles which can be either protons or lead ions. It is designed to reach a center-of-mass energy of 14 TeV in pp collisions. This design energy is constrained by the size of the tunnel (26.7 km), the dipole magnetic field of the magnet, the cavities, and other essential elements of the machine. The number of events per second generated in the LHC collisions for certain physics process is given by:

$$N_{\text{events}} = L_M \cdot \sigma, \tag{4.2}$$

where σ is the cross-section for the process and L_M the machine luminosity. Since the machine luminosity only depends on the beam parameters, assuming a Gaussian beam distribution, it can be calculated as:

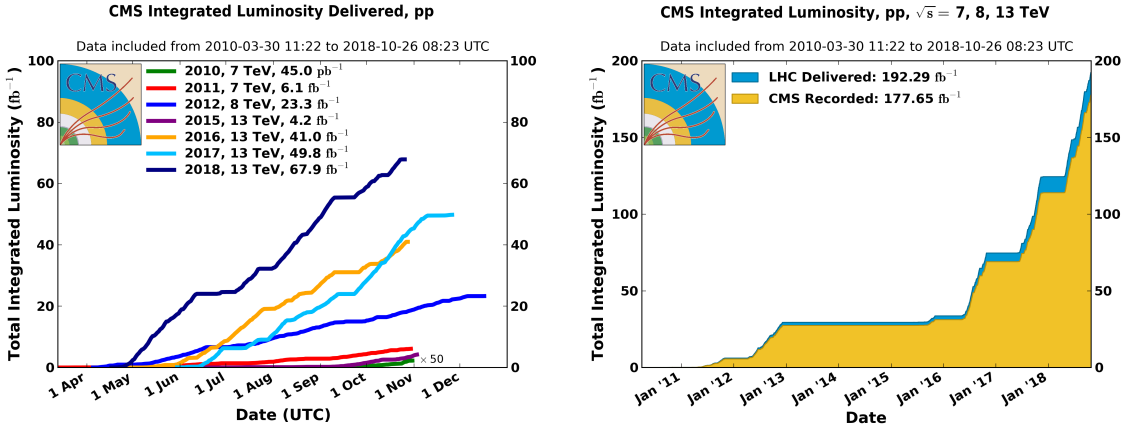
$$L_M = \frac{N_b^2 n_b f_{\text{rev}} \gamma_r}{4\pi\epsilon_n \beta^*} F, \tag{4.3}$$

where N_b is the number of particles per bunch, n_b the number of bunches per beam, f_{rev} the revolution frequency, γ_r the relativistic gamma factor, ϵ_n the normalized transverse beam emittance, β^* the beta function at the collision point, and F the geometric luminosity reduction factor due to the crossing angle at the interaction point (IP). The integral of the delivered luminosity over time, called *integrated luminosity*:

$$L = \int L_M dt, \tag{4.4}$$

constitutes a measurement of the collected data size and an important parameter to characterize the machine performance. Fig. 4.2a shows the integrated luminosity delivered to the

CMS experiment during Run I (period of time between 2010 and 2012) and Run II (period of time between 2015 and 2018). In ideal conditions, the amount of luminosity recorded should be the same as the amount delivered to the experiment by the LHC. Nevertheless, at certain moments during the machine operation, the detector may be unable to collect data. This can be caused by a busy data acquisition chain or the temporary unavailability of some detector subsystems. The recorded luminosity is shown in Fig. 4.2b (yellow), including only the luminosity that CMS was able to collect from the delivered luminosity (blue). If all subdetectors, triggers, and physics objects (electron, muon, photon, jet (collimated spray of hadrons), MET (missing transverse energy), etc) show a performance fulfilling certain quality requirements, the data is declared as good for physics analysis.



(a) (b)

Figure 4.2: a) Cumulative luminosity versus date delivered to CMS during stable beams for pp collisions at nominal center-of-mass energy. The best available offline calibrations for each year were used, b) Cumulative delivered and recorded luminosity versus time for the 2010-2012 and 2015-2018 data-taking periods (pp data only) [75–81].

Tabs. 4.1 and 4.2 summarize the luminosity information corresponding to pp runs taken during *stable beams* conditions in Run I and Run II. The term stable beams refers to LHC proton beams that are aligned, squeezed, focused, and finally directed to collide head-to-head. Special runs, in which requirements on the detector, trigger, and data acquisition are different with respect to standard data taking (e.g., low pileup runs), are not included. The term pileup refers to the additional collisions overlapping with the collisions of interest in the detector. The pp collisions data analyzed in this thesis corresponds to an integrated luminosity of 137.2 fb^{-1}, collected with the CMS detector during the Run II data-taking period, in the years 2016, 2017, and 2018, at a center-of-mass energy of 13 TeV.

Operation timeline

The LHC went online on 10 September 2008 [83], but due to a magnet quench incident 9 days later [84], the initial testings were delayed 14 months. On 20 November 2009 [85] a first circulating beam at 0.45 TeV was achieved. Three days later twin circulating beams

Run I			
	7 TeV		8 TeV
Year	2010	2011	2012
Total delivered luminosity (1/fb)	$44.96 \cdot 10^{-3}$	6.10	23.30
Recorded	$41.47 \cdot 10^{-3}$	5.55	21.79

Table 4.1: Integrated luminosity for Run I pp runs at 7 and 8 TeV during stable beams [81].

Run II (13 TeV)						
Year	2015	2016	2017	2018	2015-2018	2016-2018
Total delivered luminosity (1/fb)	4.21	40.99	49.79	67.86	162.85	158.64
Recorded and certified luminosity	2.26*	35.92	41.53	59.74	139.45	137.19

Table 4.2: Integrated luminosity for Run II pp runs at 13 TeV during stable beams. * 25 ns fills with magnet on only [82].

were established [86], and on 30 November 2009 [87] a new world record beam energy of 1.18 TeV was set. Before a short technical stop, which started in 16 December 2009 [88], the four major experiments managed to record over a million particle collisions. During this stop, the detector was prepared to achieve the higher energies needed for the beginning of the main research program. On 28 February 2010, beams circulated again, and on 19 March 2010 [89], 3.5 TeV proton beams were achieved. On 4 November 2010 [90], the first year of data taking at a center-of-mass energy of 7 TeV ended, followed by more data taking at 7 TeV in 2011. On 5 April 2012 [91] a new record collision energy of 8 TeV was established. The first long shutdown (LS1) planned for the end of 2012 was delayed to collect more data after the announcement of the discovery of the Higgs boson in July 2012 [92]. On 16 February 2013 [93] the machine entered the LS1, marking the end of Run I.
Run II started on 03 June 2015 [94]. The LHC was back online with a new record center-of-mass energy of 13 TeV. During an Extended-Year-End-Technical-Stop (EYETS) between 2016 and 2017, a new four-layer pixel detector was installed in CMS [95]. On 12 November 2018, pp collisions were stopped, followed by a lead-ion run, which finished on 03 December 2018 [96], marking the end of Run II. On 10 December 2018 [97], a last fixed target physics study with lead ions was stopped, signaling the start of the second long shutdown (LS2). CMS managed to record more than the planned 150 fb^{-1} during the Run II. A selection of dates from the above-described timeline is shown in Fig. 4.3.
Run III is (as of the time of writing this dissertation) scheduled to start in 2022 and finish between the end of 2024 and the beginning of 2025 [98, 99]. To prepare for this new period of data taking, the Phase-I upgrade of the detector is currently taking place. By the end of Run III, 300 fb^{-1} of data are expected to be collected.
A major upgrade of the LHC is scheduled for the third long shutdown (LS3), between the

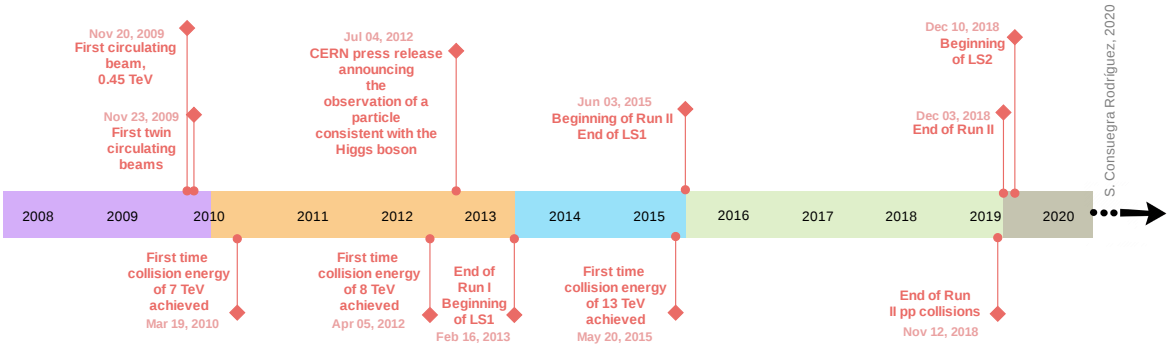


Figure 4.3: LHC operation timeline.

years 2025 and 2027. In preparation for the new project, known as the High-Luminosity LHC (HL-LHC) [100], the installation of some components have started during the LS2. The rest of the equipment and experimental components will be installed during the LS3. The Phase-II upgrade will allow the detector to face a higher instantaneous luminosity L = $5-7.5 \cdot 10^{27}$ cm^{-2} s^{-1}, which would bring an average number of pileup events per pp collision between 140 and 200. During the operation time of the HL-LHC, the 3000 fb^{-1} expected to be delivered will allow the study of extremely rare phenomena and improve the precision of already obtained measurements.

4.2 The CMS detector

CMS is a 21.6 meters long, 14.6 m in diameter, and 14 000 tonnes general-purpose detector, located about 100 meters underground in Point 5, CERN. It is designed to explore the physics of the Terascale, look for evidence of BSM physics, such as supersymmetry or extra dimensions, and study certain aspects of heavy-ion collisions. The overall layout of the detector, depicting its main subdetectors: the Silicon trackers, the Crystal Electromagnetic Calorimeter (ECAL) [101], the Hadron Calorimeter (HCAL) [102], and the Muon chambers [103], is shown in Fig. 4.4.
CMS uses a right-handed coordinate system, with the origin located at the nominal collision point. The z-axis points along the beam direction. In the x-y plane, the x-axis points to the center of the LHC ring and the y-axis to the experiment surface. The pseudorapidity angle (η), relative to the beam axis, is defined as:

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right), \tag{4.5}$$

where θ is the polar angle measured from the z-axis. In the x-y transverse plane, observables which are Lorentz invariant along the z-axis can be determined, such as the transverse energy (E_T), the transverse momentum (p_T), and the energy imbalance E_T^{miss}. Other important CMS related observables are the polar coordinates ϕ and r. The azimuthal angle ϕ is measured from the x-axis in the x-y plane, while the radial distance r corresponds to the radius in the same plane.

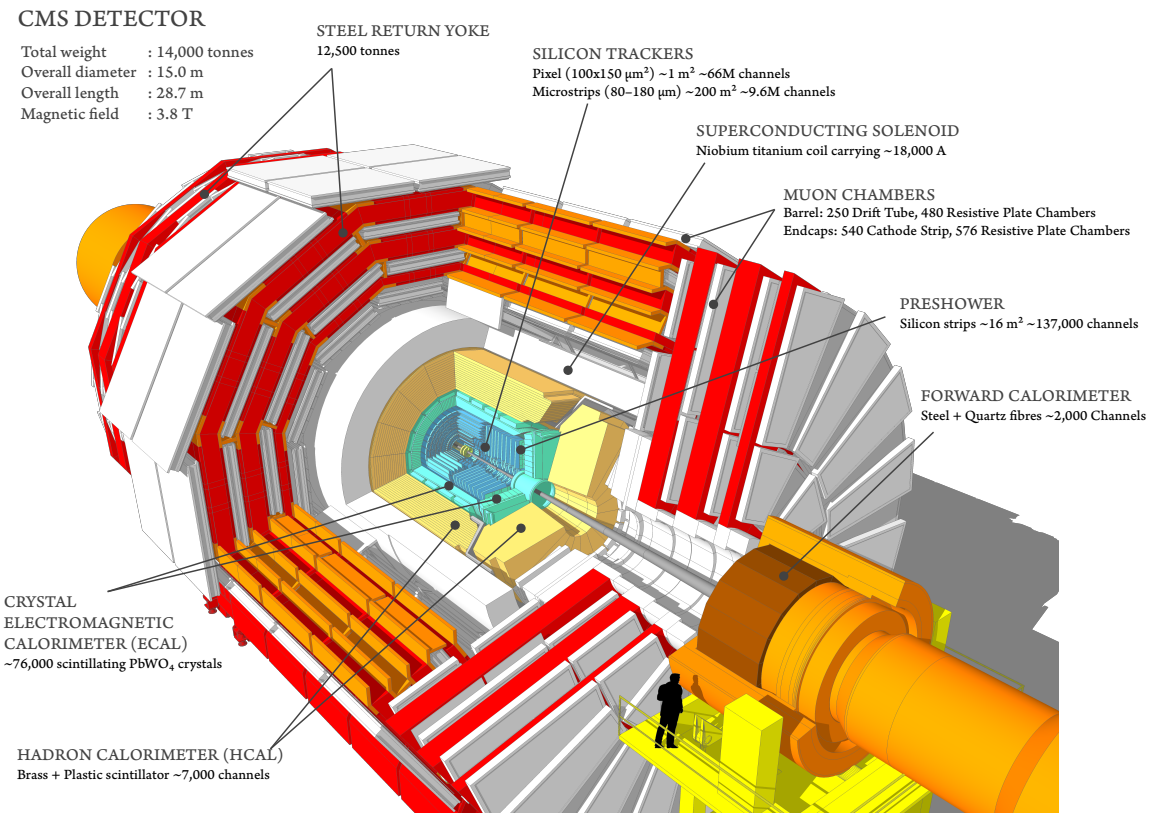


Figure 4.4: Sectional view of the CMS detector. [104].

4.2.1 Understanding CMS acronym

The description of the detector as compact in the acronym "CMS" comes from the fact that all CMS detector materials are quite concentrated, compared with the ATLAS detector, which is about only half the weight of CMS, but twice as long, with 1.5 times diameter. The central element of the acronym "muon" comes from the fact that CMS is specially designed to accurately detect muons, unique signatures of interesting physics. The final element of the acronym comes from the selection of the type of magnet, a solenoid, formed by a cylindrical coil of superconducting fibers. The choice of the magnet system is a fundamental difference between ATLAS and CMS. ATLAS has an air-cooled toroid system and a central solenoid [105], while CMS has a single superconducting solenoid [106]. This simplifies the reconstruction of tracks (charged particles) in CMS, since the particles only bend in the transverse plane. The solenoid is 13 m long, has an inner diameter of 6 m, and is designed to create an axial field of 4T. At the most outer part of the detector, a steel return yoke controls the field outside the solenoid and provides structural support to the detector.

4.2.2 Main features of CMS detector

The detector distinguishing features can be summarized as follows:

- Good muon identification and momentum resolution, good dimuon mass resolution, and unambiguous determination of the charge of muons with $p_T < 1$ TeV.

- Good charged-particle momentum resolution and reconstruction efficiency. Efficient triggering and offline identification of τ's and b-jets.

- Good electromagnetic energy resolution, good diphoton and dielectron mass resolution, wide geometric coverage, and efficient photon and lepton isolation at high luminosities.

- Good missing-transverse-energy and dijet-mass resolution.

Particle detection principles

The operation of a detector depends on how the particles to be detected interact with the detector's material [107]. Each subdetector of CMS is designed to stop, track, and measure a specific type of particle coming from the central collision. Tab. 4.3 shows the subdetectors in which each type of particle interacts and leave a signal. The combination of the information coming from all the subdetectors allows us to identify the particles and reconstruct their trajectory [108].

Particle	Tracker	ECAL	HCAL	Muon chambers
Muon	√			√
Electron	√	√		
Charged hadron	√	√*	√	
Neutral hadron			√	
Photon		√		

Table 4.3: Particles and signal left in each of the main subdetectors. [*] very weak signal.

Fig. 4.5 shows a transverse section of CMS with a pictorial representation of the information contained in Tab. 4.3, depicting the response of each of the main subdetectors to the particles passing through the different layers.

4.2.3 Tracking system

The inner tracking system of CMS, called the tracker, consists of two tracking devices: the inner tracker silicon pixel detector and the outer tracker silicon strip detector. The tracker, with a length of 5.8 m and a diameter of 2.5 m, is designed to provide a precise measurement of the charged particles' momentum and to reconstruct the event vertices. Primary vertices (points at which pp interactions occurred) and secondary vertices (common points of origin for a set of tracks produced in the decay of a particle within the detector) are successfully reconstructed [110]. Due to its high granularity, fast readout, radiation hardness, and low material budget (low-mass layers), the tracker can reconstruct the track trajectories reliably and attribute them to the correct bunch crossing in high pileup conditions, while deflecting them from their path as less as possible. Fig. 4.6 shows a section of the Phase-I CMS tracking system, depicting the pixel and the strip detectors.

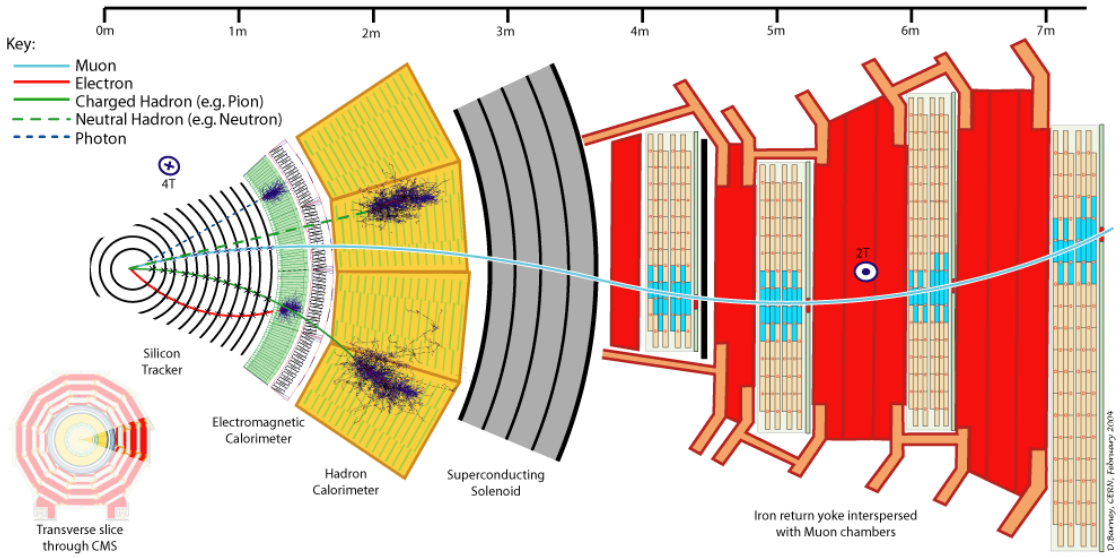


Figure 4.5: Transverse section of the CMS detector illustrating the interaction of the particles in each of the main subdetectors [109].

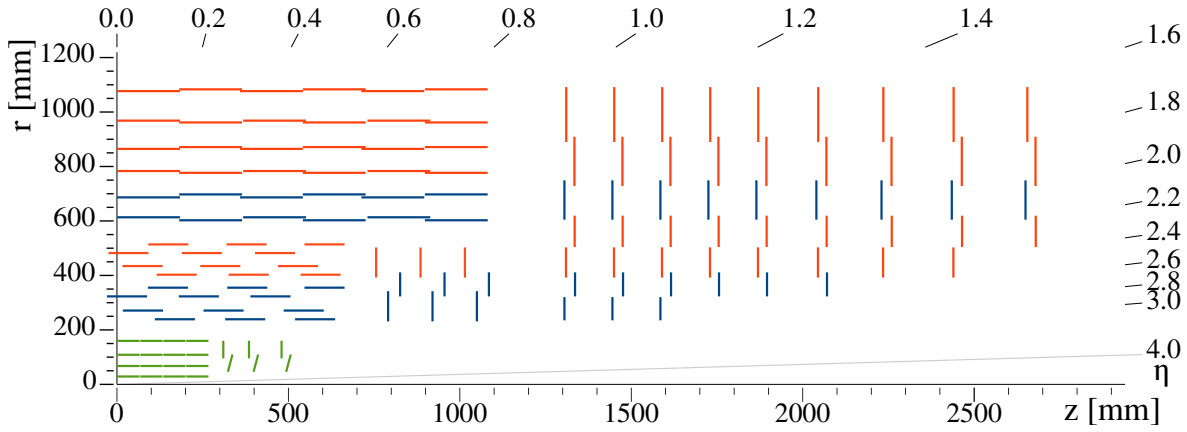


Figure 4.6: One quarter section of the Phase-I CMS tracking system in r-z view. The pixel detector is shown in green, single-sided strip modules in red and double-sided strip modules in blue [111].

Momentum measurement

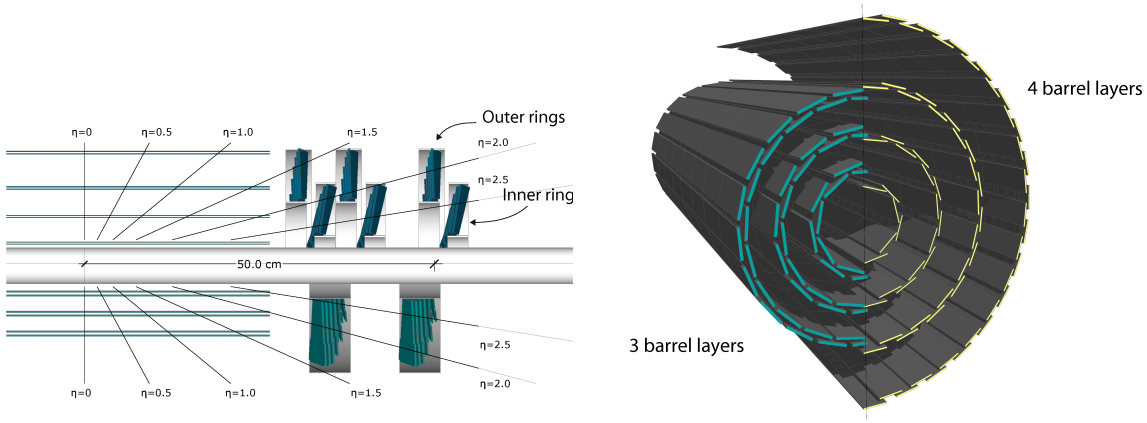
Charged particles passing through the tracker layers create hits (electrical signals registered in the detector modules), and their trajectory can be determined with high accuracy gathering the hits from all the tracker layers. Since the deflection of charged particles under the influence of a homogeneous magnetic field forms a circle of radius R, the transverse momentum of the charged particle (p_T) in GeV/c, can be calculated as:

$$p_T = 0.3\ (R \cdot B), \tag{4.6}$$

where R is the radius of curvature in meters and B the magnetic field strength in Tesla.

The tracker silicon pixel detector

The tracker silicon pixel is the first detector in proximity to the interaction point. The radiation tolerant sensors allow the detector to operate under high track rate conditions. With a pixel cell size of 100×150 μm^2, a similar track resolution in the $r - \Phi$ and z-direction is achieved. The pixel detector provides seed tracks, which serve as input for the outer track reconstruction and high-level triggering. The high impact parameter resolution achieved with the pixel detector is crucial for the reconstruction of secondary and primary vertices.
Figs. 4.7a and 4.7b show a comparison of the conceptual layout and transverse view of the original pixel detector, known as Phase-0 pixel detector (in operation until the end of the 2016 data-taking period), and the Phase-I pixel detector (in operation since 2017). The Phase-I pixel detector, installed during the EYETS between 2016 and 2017, is expected to be upgraded to the Phase-II pixel detector, to cope with the more challenging operation conditions foreseen at the HL-LHC [111, 112].



(a) (b)

Figure 4.7: a) Comparison of conceptual layout and hit coverage as a function of pseudorapidity for the Phase-0 (bottom) and Phase-I (top) pixel detector. b) Transverse view comparing the original three-layer geometry (blue) with the upgraded four-layer geometry (yellow) of the pixel detector [113].

Phase-0 pixel detector

The Phase-0 pixel detector was composed of three barrel layers (BPix) and two endcap disks (FPix). The 53-cm long BPix layers were located at 4.4, 7.3, and 10.2 cm, while the FPix disks extended from $\approx$ 6 to 15 cm in radius, at z = $\pm$34.5 and z = $\pm$46.5 cm on each side. The BPix had 48 million pixels in an area of 0.78 m^2, while the FPIX had 18 million pixels in an area of 0.28 m^2. This detector layout allowed us to count on three tracking points over almost the full η-range. The innermost layer of the Phase-0 pixel detector was designed to stay operational for a minimum of 2 years at LHC nominal luminosity design. Beyond this design peak luminosity, the high occupancy and trigger rate was expected to cause data loss in the readout chips (ROCs). Furthermore, to mitigate a lower tracking efficiency or higher fake rates at high pileup, an extra pixel layer was foreseen to be needed. Due to radiation damage, a degradation of performance (e.g., deterioration of the hit detection efficiency and resolution)

was expected. Moreover, a reduction of the material in the tracking volume with new light weight substitutes was known to be able to diminish the degradation in performance caused by the material. The Phase-0 upgrade addressed these key limitations. The characteristics of the resulting Phase-I pixel detector are described in the next segment.

Phase-I pixel detector

The Phase-I pixel detector is composed of 4 barrel layers (BPix) and three endcap disks (FPix). It has an additional layer and endcap compared to the Phase-0 pixel detector. Nevertheless, the mass of the detector is reduced with respect to its predecessor by using new ultra-lightweight support and cooling, as well as relocating further in z out of the tracker part of the passive material. The additional fourth layer is located at a radius of 16 cm, while the rest of the layers keep the same location from its predecessor. The fourth layer provides a safety margin in case the radiation damage of the first silicon strip layer of the Tracker Inner Barrel (TIB) becomes significant. The described layout allows us to count on four tracking points over the whole η-range. The track fake rate is reduced since now track seeds with four pixel hits (quadruplets) are available as input to the first tracking step of the track reconstruction, instead of the previous maximum of three. Thus, the installation of the Phase-I pixel detector allowed us to maintain the quality of the tracking by offsetting the effects of the radiation damage of the outer Tracker and led to an improvement in the tracking performance parameters.

The tracker silicon strip detector

The tracker silicon strip detector is the second detector in proximity to the interaction point after the tracker silicon pixel detector, located between 20 cm and 116 cm in the radial region of the tracker. It has 200 m^2 of active silicon area and a pseudorapidity coverage extending up to $|\eta| \approx 2.5$. It is composed of three subsystems: the Tracker Inner Barrel and Disks (TIB and TID), the Tracker Outer Barrel (TOB), and the Tracker EndCaps (TEC+ and TEC-), with the sign indicating the location along the z-axis.
The Tracker Inner Barrel and Disks are composed of four barrel layers and three disks at each end. Their 320 μm thick silicon micro-strip sensors are located parallel in the barrel and radial in the disk, providing a single point resolution between 23 μm and 35 μm in the r-ϕ direction. The Tracker Outer Barrel surrounds the TIB and TID, extending in the z-direction between $\pm$118 cm. The 500 μm thick micro-strip sensors provide a single point resolution between 35 μm and 53 μm in the r-ϕ direction. The Tracker Endcaps expand coverage beyond z=$\pm$118 cm. Each endcap is composed of nine disks, and each disk has seven rings of silicon microstrip detectors, with a width of 320 μm (inner four rings) or 500 μm (rings 5-7). A second micro-strip detector module, installed in the first two layers of TIB, TID, and TOB, as well as rings 1, 2, and 5 of the TECs, allows measuring the single point resolution in the z-direction, with 230 μm in the TIB and 530 μm in the TOB.
The strip tracker detector layout described in this segment ensures the detection of at least ≈ 9 hits for $|\eta| \approx 2.4$, estimated from a study of the number of hits in the strip tracker as a function of the pseudorapidity $|\eta|$.

4.2.4 Calorimeters

A calorimeter is composed of a block of matter in which particles get absorbed, and a fraction of the particle's energy is transformed into a measurable signal [114]. The term calorimeter comes from the fact that almost all the energy of the particles is converted to heat. They can be classified into two types: homogeneous and sampling calorimeters. In homogeneous calorimeters, the same medium acts as absorber and detector, while in sampling calorimeters a layer structure alternates passive absorber mediums with active mediums.

Electromagnetic calorimeter

The CMS electromagnetic calorimeter is a homogeneous calorimeter formed by 75 848 lead tungstate ($PbWO_4$) scintillating crystals [115], with a fast, high granularity, and radiation-resistant design. It is formed by a barrel part (EB), two endcaps (EE), and a Preshower detector (ES). The overall layout of the ECAL is shown in Fig. 4.8. The ECAL's main purpose is to provide a precise energy measurement, needed for many physics analyses [116]. In particular, a high resolution and efficient identification of photons was a driving criterion in its design and resulted crucial for the observation of the H$\rightarrow \gamma\gamma$ decay process. The electromagnetic energy resolution of the ECAL can be calculated as:

$$\frac{\sigma_E}{E} = \frac{a}{\sqrt{E}} \oplus \frac{b}{\sqrt{E}} \oplus c, \tag{4.7}$$

where a is a stochastic term, depending on event to event fluctuations, detector gain, etc. b is a noise term related to the electronic noise and pileup conditions, and c is a constant term resulting from a non-uniformity of the longitudinal light collection, energy leakage, and detector inter-calibration uncertainties.

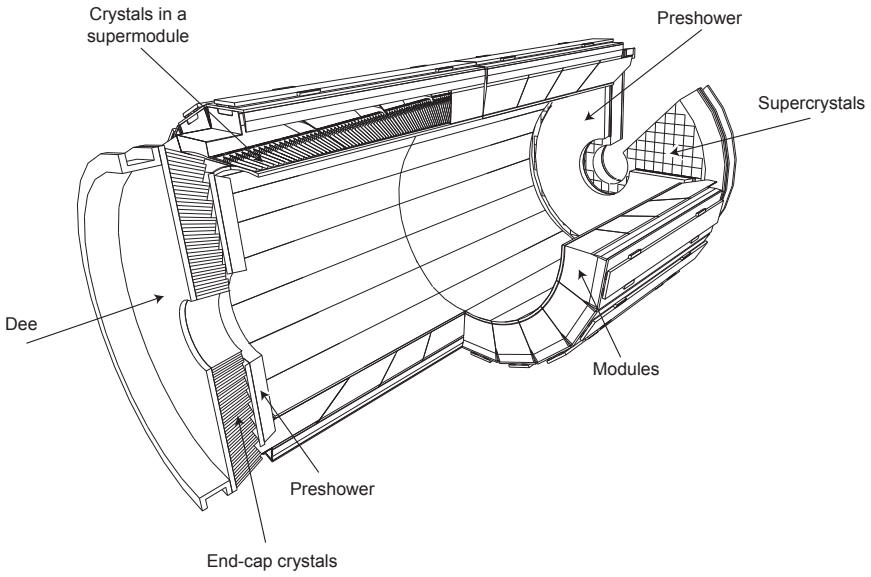


Figure 4.8: Layout of the CMS electromagnetic calorimeter, depicting the barrel, the two endcaps, and the preshower detectors [117].

Barrel and Endcaps

The ECAL barrels cover a region of pseudorapidity $|\eta| < 1.479$, with 36 identical supermodules. Each supermodule, formed by four modules, is equipped with crystals, avalanche photodiodes (APDs), and readout electronics. The modules are composed of submodules containing 400 or 500 crystals, according to their position in η. The scintillating light produced by the ionization radiation is converted in the APDs into an electrical current. In conventional photodiodes, the photons are converted to electron-hole pairs, and the later ones are simply collected. Nevertheless, in avalanches photodiodes as the ones installed in the ECAL barrels, an internal gain is incorporated by the use of higher electric fields that increase the number of charged carriers collected. The gain factor is very sensitive to the temperature and the applied voltage. Thus, the temperature of the ECAL has to be maintained constant with high precision to preserve the energy resolution. The cooling system of the ECAL is designed to extract the heat dissipated from the readout electronics and keep the temperature of the crystals and photodetectors, with a precision of ± 0.05 °C.
The ECAL endcaps cover the rapidity range $1.479 < |\eta| < 3$ and are divided into two Halves or *Dees*. Each endcap consists of 5×5 mechanical units of crystals, called supercrystals [118]. The scintillation light from the crystals is detected with vacuum phototriodes (VPTs), designed to operate with high reliability for at least ten years in the LHC environment. VPTs are single-stage photomultiplier tubes, consisting of an input window, a photocathode, focusing electrodes, an electron multiplier, and an anode. The photons excite the electrons, which are emitted to the vacuum. These so-called photoelectrons are focused by the electrodes into secondary electron emission surfaces, called dynodes. The secondary emission can be repeated several times in consecutive dynodes, to achieve a high gain. The secondary electrons coming from the last dynode are collected on the anode, and the signal is registered.

Preshower detector

The Preshower detector (ES) is a sampling calorimeter with two layers, lead absorbers, and silicon strip sensors, placed in front of the endcaps. The ES- and ES+ distinguish the two ends of the ES. The energy deposited by the electromagnetic showers in a lead layer is measured in the consecutive silicon strip sensor. Each sensor is divided into 32 strips, measures 63×63 mm^2, and is 300 μm thick. The main function of the Preshower detector is the identification of neutral pions within a fiducial region of $1.653 < |\eta| < 2.6$. An effective π_0 rejection is accomplished by measuring the transverse profile of electromagnetic showers after ≈ 3 radiation lengths (X_0), where a radiation length corresponds to the mean length in cm needed to reduce the energy of an electron by a factor $1/e$ [119]. The identification of electrons against minimum ionizing particles and the determination of the position of electrons and photons with high granularity are also relevant features of the Preshower detector. In the following subsection, the detector surrounding the ECAL, the Hadron calorimeter, is described.

Hadron calorimeter

The Hadron calorimeter consists of 4 subsystems: the HCAL Barrel (HB), the HCAL Endcap (HE), the HCAL Outer (HO), and the HCAL Forward (HF). Within the CMS radial configuration, it is located between the outer extent of the electromagnetic calorimeter (R =

1.77 m) and the inner extent of the magnet (R = 2.95 m). The information provided by the HCAL is especially relevant for the identification of hadron jets and particles whose signature in the detector is characterized by the presence of MET, as neutrinos or exotic particles. The four subsystems together provide a pseudorapidity coverage of $0 < |\eta| < 5$. A layout of the CMS detector, depicting the location of the HCAL four major sections, is shown in Fig. 4.9.

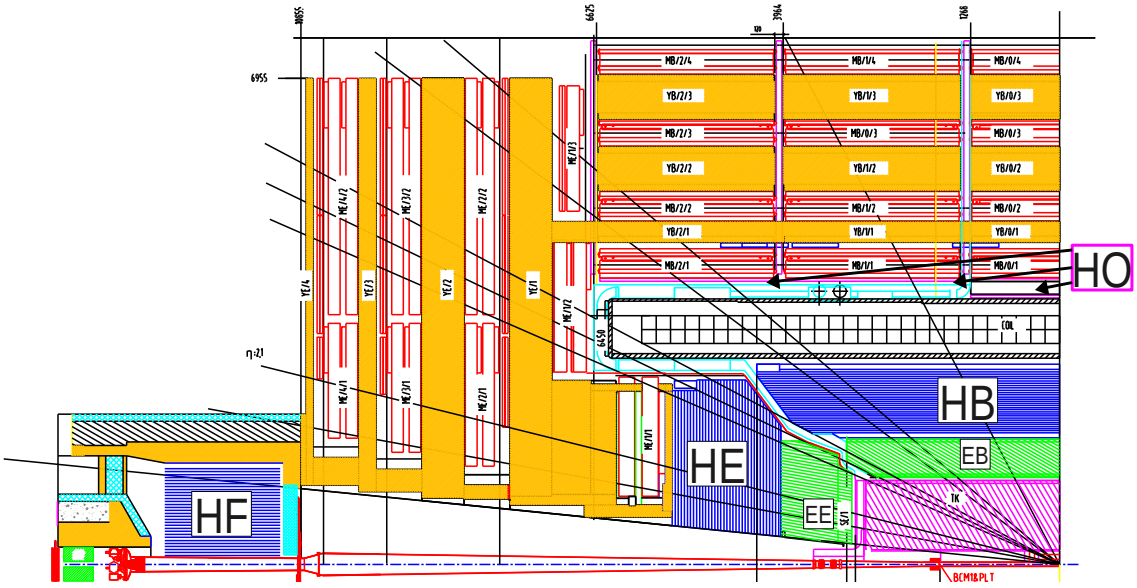


Figure 4.9: One quarter section of the CMS detector in r-z view, showing the location of the HB, HE, HO, and HF subsystems of the HCAL detector [120].

The HCAL Barrel is a sampling calorimeter, with a pseudorapidity coverage of $|\eta| < 1.3$, composed of 36 identical azimuthal wedges, distributed among two-half barrels sections (HB+ and HB-). Each wedge is divided into four azimuthal angular (Φ) sectors. The HCAL endcaps cover a pseudorapidity range of $1.3 < |\eta| < 3$. To be able to operate inserted into the ends of the magnet, the non-magnetic material brass was selected as absorber, while the active material is a plastic scintillator. The HCAL endcaps are able to contain the hadron showers within its pseudorapidity coverage, but for $|\eta| < 1.3$, the stopping power of EB and HB is not enough to contain them. Thus, to collect the energy of the showers that has not been deposited after the HB and the information on late starting showers, HO layers are located before each of the five 2.536 m wide rings of the iron yoke. Each HO layer is segmented in 12 Φ-sectors, with a sector having six slices in Φ. All *tiles* (the smaller scintillator units) of each Φ slice, group together to form a *try*.
The HF detector is a Cherenkov calorimeter formed by a cylindrical steel structure, segmented azimuthally into 20° Φ wedges. It is located 11.2 m from the interaction point. From the total 36 wedges, 18 are positioned at each side of the interaction point. The structure of the HF is constituted by grooved plates, 5 mm thick. The active medium of the detector is made of quartz fibers, which are inserted into the grooves. The HF is especially sensitive to the showers' electromagnetic component, with a signal generated by capturing a small fraction of the Cherenkov light emitted by the particles of the shower. While half of the fibers cover the full depth of the absorber, the other half is located after 22 cm of material. The electron and photon showers leave a significant fraction of their energy in the first 22 cm, while the

hadron shower energy deposits are nearly the same in both depth segments. The information provided by the two groups of fibers, which are independently read, allows us to distinguish the electron and photon showers from the hadron showers, due to the characteristic different energy deposit signature.
The ECAL and HCAL calorimeters were designed to operate over ten years up to the LHC machine luminosity design and to assimilate an integrated luminosity up to 500 fb^{-1}. After the Phase-I upgrade, the HCAL is prepared to operate up to twice the machine luminosity design. To meet the upcoming challenges in terms of longevity and performance, both calorimeters will face a Phase-II upgrade [115, 121], preparing the active material and the electronics for the operation conditions at the HL-LHC.

4.2.5 Muon system

The muon system of the CMS detector is designed to efficiently perform the tasks of muon identification, momentum measurement, and triggering, over the full kinematic range of the LHC. It is composed of three subsystems: drift tubes (DTs), cathode strip chambers (CSCs), and resistive plate chambers (RPCs), featuring three different types of gaseous particle detectors. The system has a cylindrical barrel section and two planar endcap regions. Fig. 4.10 shows an r-z view of the CMS detector, depicting the main muon subsystems. A high quality assurance of the muon momentum resolution is obtained through the cross-check of the measurement of the muon system with the one obtained in the inner Tracker. A distinctive characteristic of the DT and CSC subsystems is their ability to efficiently trigger on the p_T of the muons, in a process independent from the rest of the detector. The determination of the muon momentum resolution relies on a precise knowledge of the position with respect to each other of the components of the muon system and the inner Tracker. To accomplish this, an alignment system is in place.

Barrel region

The barrel region is characterized by a small neutron-induced background, a low muon rate, and a uniform 4T magnetic field. The DT chambers are suitable for this environment and, therefore, located in 4 stations in the barrel. In the first three stations, formed by 12 chambers, 8 chambers are dedicated to the measurement of the muon coordinate in the r-Φ plane, and 4 chambers provide the same measurement in the z-direction. In the last station, formed by 8 chambers, only the measurement in the r-Φ plane is performed. The number and orientation of the chambers in each station were chosen to efficiently link muon hits from different stations, to form the muon tracks, and to reject background hits.
Due to the uncertainty in the background rate and the measurement of the beam-crossing time at full luminosity, a third gaseous particle detector was installed. This detector, composed of resistive plate chambers, serve as dedicated trigger system. The double-gap RPC chambers operate in avalanche mode, with an efficient performance at high rates. Each of the first two stations and each of the two last stations of the barrel muon system contains one RPC layer, totaling 6 layers of RPCs. The RPCs have a fast response, which results in a good time resolution, though the position resolution is coarser than the one from DTs and CSCs [122, 123]. They support the task of constructing tracks from a group of hits, by helping to resolve pending ambiguities through the combination of their information with the one provided by the other subsystems.

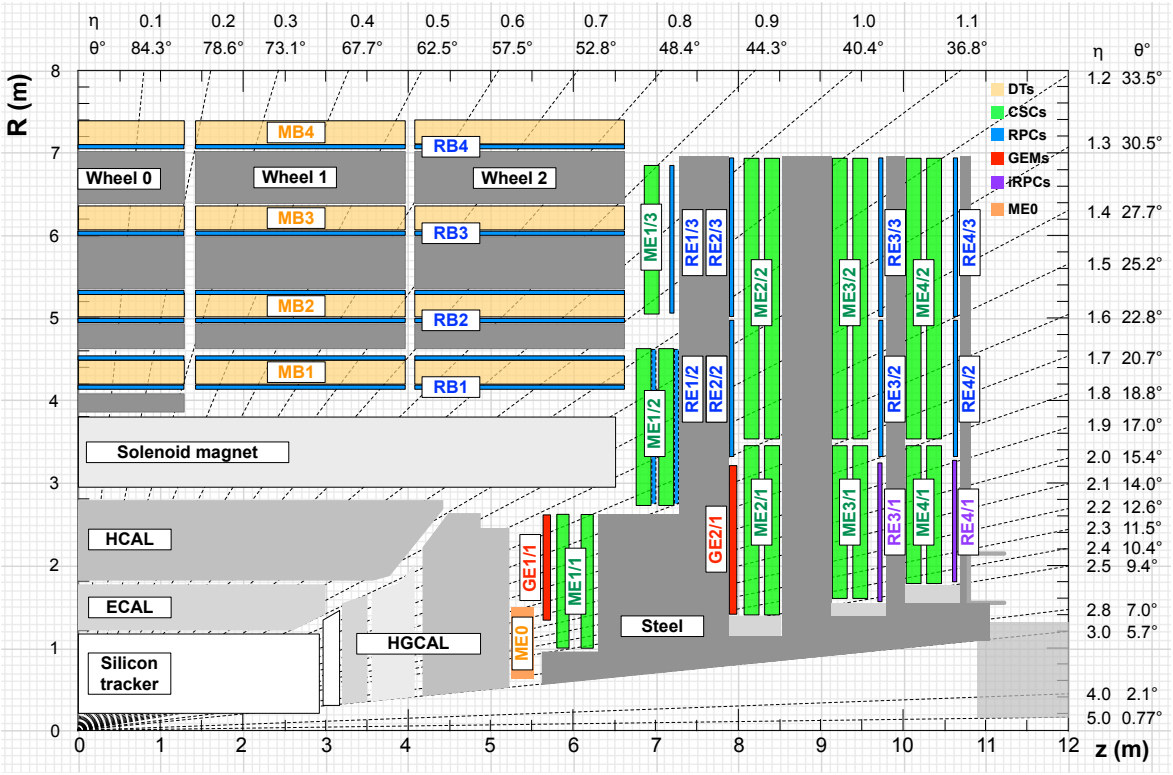


Figure 4.10: One-quarter section of the CMS detector in r-z view, featuring the four DT stations in the barrel (MB1–MB4, in orange), the four CSC stations in the endcap (ME1–ME4, in green), and the RPC stations (in blue). The Phase-II upgrades, some of them discussed below (RE3/1, RE4/1, GE1/1, GE2/1, ME0), are also included. MB = DT = Drift Tubes, ME = CSC = Cathode Strip Chambers, RB and RE = RPC = Resistive Plate Chambers, GE and ME0 = GEM = Gas Electron Multiplier, and iRPCs = improved RPC chambers) [124].

Endcaps region

A high muon and background rate and a non-uniform large magnetic field distinguish the endcaps region. The cathode strip chambers are suitable to operate under these conditions due to their fast response time, granularity, and radiation resistance. Four stations of CSCs, with chambers located perpendicular to the beamline, cover the pseudorapidity range $0.9 < |\eta| < 2.4$. A chamber is composed of 6 layers. The information provided by each layer allows the rejection of non-muon backgrounds and an efficient hit matching to those obtained in other muon stations and the inner Tracker. A plane of RPCs was installed in each of the first three stations of the endcaps, allowing the trigger to exploit the coincidence between the stations for background rejection. The location of the stations within the muon system allows us to have full pseudorapidity coverage for $|\eta| < 2.4$, or $10° < |\theta| < 170°$. The offline reconstruction efficiency of muons ranges from 95 to 99%, being lower only in the regions between the 2 DT wheels and between the DT and CSC.
The majority of the components of the four muon subsystems have been operating since 2008 when the LHC started running. Minor replacements occurred during the LS1 when the ME4/2 CSC stations, as well as the RE4/2 and RE4/3 RPC chambers, were substituted. In

2017, five testing pairs of chambers for a new type of detector in CMS, called gas electron multiplier (GEM), were installed in the forward region in order to assess its performance [125, 126]. The purpose of GEM is to improve the muon triggering in the forward region and the reconstruction in the pseudorapidity range $1.6 < |\eta| < 2.2$, in preparation for the operation under HL-LHC conditions. After the successful commissioning of the pilot GEM super-chambers, 72 super-chambers were added in 2019 and the installation of the very first GE11 station was finished in September of 2020. The remaining ones are expected to be installed before the end of the LS2. The Phase-II Upgrade of the muon system is step by step being accomplished.

4.2.6 Trigger and Data Acquisition

At LHC design luminosity, considering a beam crossing interval for pp collisions of 25 ns that corresponds to a crossing frequency of 40 MHz, an average of 20 collisions occur per bunch crossing. Since it is impossible to store for offline analysis the resulting amount of data in its totality, a decision whether to record an event or not has to be made online, event per event. In this way, only a small portion of the events are recorded and go into datasets to be used for physics analysis and calibration. The datasets to be saved on tape are determined according to the priorities of the CMS physics program. The needed drastic rate reduction of their size is performed in 2 steps, which together constitute the CMS trigger system [127]: the Level-1 (L1) trigger [128] and the High-Level Trigger (HLT) [129]. The L1 trigger is considered a hardware trigger since it is based on custom-designed electronics, and the HLT constitutes a software system, operating in a filter farm of commercial processor cores.

L1 trigger

The L1 trigger reduces the input rate of 40 MHz to an output rate of 100 kHz, with a fixed latency of 3.8 μs. In this time interval, the global decision whether to pass the event to the HLT or not is made, based on pieces of basic information coming from the muon chambers and the calorimeters, e.g., the presence of energy deposits compatible with physics objects known to be part of interesting final states. The high-resolution data of the events is kept during the short latency time in the front-end electronics so that if the event is accepted, all the event information can be quickly accessed by the HLT. Fig. 4.11 presents the architecture of the Level-1 trigger, composed of local, regional, and global components. At the bottom end of the hierarchy structure, the local component called Trigger Primitive Generators (TPG) is found, based on energy deposits in calorimeter trigger towers and track segments or hit patterns in muon chambers. The information from the TPGs is combined in the regional component, called regional trigger [130], so that trigger objects like electrons or muon candidates can be ranked and sorted. The candidates' rank is decided considering the energy, momentum, and quality. This last characteristic refers to the confidence in the parameters that come from the L1 measurements. The global calorimeters and global muon triggers decide, according to the information provided by the regional component, the objects to be passed to the global trigger [131], which constitutes the top step in the L1 trigger hierarchy. The global trigger makes the decision whether to reject the event at this point or pass it to the HLT. A relevant element for the decision is the readiness of the subdetectors and the Data Acquisition System (DAQ) [132], determined by the Trigger Control System (TCS) [133–135]. A positive decision,

known as Level-1 Accept (L1A), is sent to the subdetectors through the Timing, Trigger, and Control (TTC) system [136].

HLT trigger

The HLT trigger reduces the input rate of 100 kHz to around 1 kHz, performing more sophisticated calculations than the L1 trigger, similar to the ones done in offline analysis. The HLT event filter farm (EVF) is composed of filter-builder units [137]. In the builder units, fragments from different detectors are combined to form complete events. The assembled events enter the filter units for the unpacking of the raw data into detector-specific data structures, in order to reconstruct the events and apply trigger filters. Preliminary selections based on the information provided by the calorimeters and the muon detectors allow reducing the rate before feeding the events into the tracking reconstruction algorithms, which are quite CPU consuming. A specific sequence of steps with increasing complexity for the reconstruction and selection of physics objects is known as an HLT path. The HLT data processing is constructed around this concept. The accepted events are grouped into a set of non-exclusive streams, which are set taking into account the HLT decisions. The data events are first stored locally and shortly afterward transferred to the CMS-Tier 0 computing center [127], for offline processing and storage.

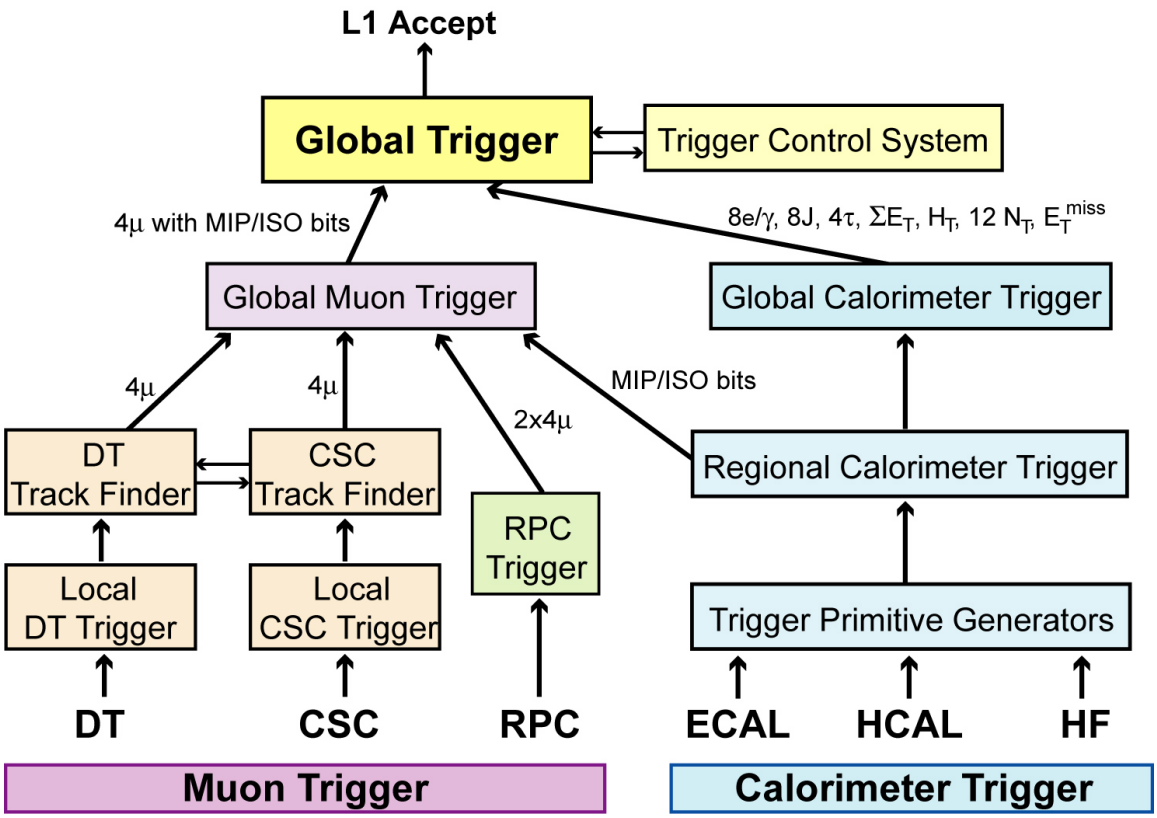


Figure 4.11: Architecture of the L1 trigger [27].

The excellent performance of the trigger system during the Run II data-taking period was guaranteed by the Phase-I upgrade of the trigger subsystems [138,139]. The main objective of

the upgrade was to assimilate the higher trigger rates while maintaining the trigger efficiency, such that the trigger thresholds would not need to be increased significantly in order to meet the 100 kHz limit output rate of the L1 trigger.

Data Acquisition

The main purpose of the CMS DAQ system is to transport the event data from the L1 trigger to the HLT, as depicted in Fig. 4.12, and to provide the computing power for the operation of the HLT. At CMS design luminosity, DAQ handles 100 kHz as input rate, i.e., the output rate of the L1 trigger, for a data flow of ≈ 100 GB/s. The DAQ system is deployed in up to 8 slices, which operate almost as autonomous systems. The full architecture of the CMS DAQ system is shown in Fig. 4.13. Once a synchronous L1 trigger arrives via the TTC, the data stored in the subdetector front-end systems (FES) is extracted from the FES by the Front-End Drivers (FEDs) and pushed into the DAQ system.

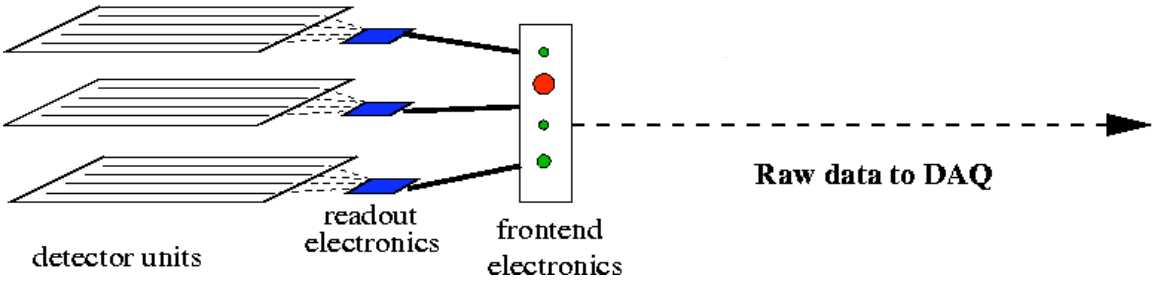


Figure 4.12: Path of the raw data on their way to entering the DAQ system [140].

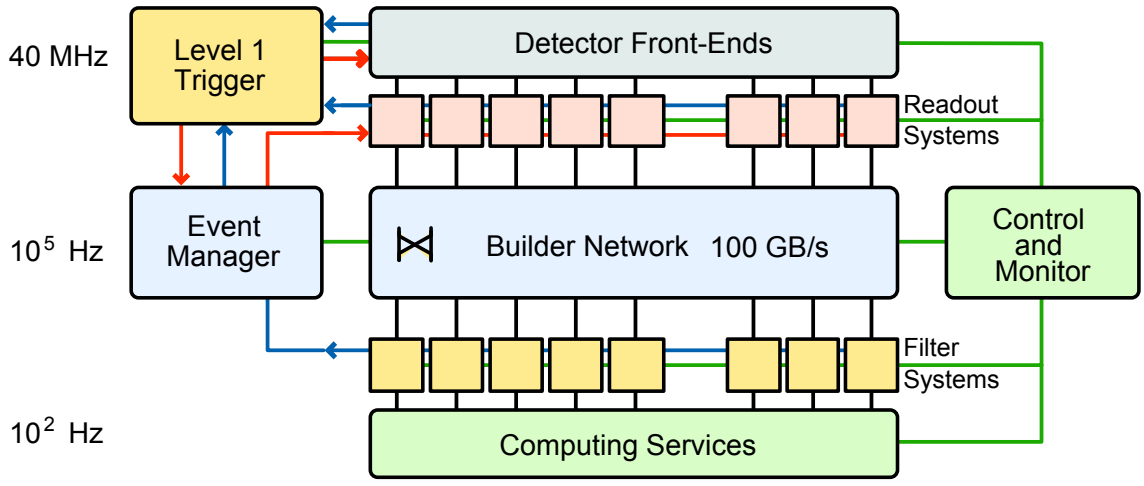


Figure 4.13: Architecture of the DAQ system [27].

The design of the FEDs depends on the subdetector they are associated with. However, the interface to the central DAQ is common. Each FED provides a signal indicating the status of the readout process, being the states: Ready, Warning, Busy, Out-Of-Sync, and Error.

The status information allows the DAQ shifter present at P5 to decide whether to stop a run or not if the acquisition and quality of the data is being compromised. The FEDs encapsulate the data from the FES in a common data structure and the data from each of them is read into the Front-end Read-out Links (FRLs). The information from two FEDs can be merged in the FRLs. At this point, DAQ has provided the HLT all the information from the event stored in the subdetectors, following its acceptance by the L1 trigger. After the assembling of the event fragments in the event builder, the transition to the Filter Units (FUs) in the Event Filter, and the selection of the portion of events for storage, the data is transferred from the local storage site at P5 to the mass storage at the Meyrin site.
The trigger and DAQ system will undergo a Phase-II upgrade [141, 142]. For the first time, tracking and high-granularity calorimeter information is expected to be included in the L1 trigger. The upgrade will allow the system to operate under the HL-LHC data-taking conditions, in which the event size will increase, with a foreseen L1 (HLT) output rate of 750 kHz (7.5 kHz) and a latency of 12.5 μs. The storage capacity at P5 would need to be increased consequently to be able to handle the increased amount of data.

4.2.7 Offline Computing

The main tasks of the CMS offline computing system are the storage, transfer, and manipulation of the recorded data [143,144]. The system provides access to conditions, calibrations, and supports the production of simulated events, using resources located around the world in collaborating institutes. The CMS hierarchical architecture composed by Tiered centers is shown in Fig. 4.14. For the elaboration of the diagram, the most updated information (as of the time of writing this dissertation) concerning site availability and status was used [145]. The Tier-0 center is located at CERN, a few Tier-1 centers are distributed at national computing facilities, and several Tier-2 centers are situated at research institutes.

CMS Data Hierarchy

The CMS data model is based on the concept of *event*. The recorded data from a single triggered bunch crossing along with the new data originated from it constitutes an event. The event serves as input to physics modules, which perform a specific selection, reconstruction, and analysis, while a so-called analyzer produces a piece of concise information from an event collection.
The following data formats are used for the analysis of the CMS data [146]:

- **RAW format:** ($\approx$ 1 MB/event) It contains the complete recorded information as it came out from the detector, including some trigger decision record and metadata.

- **RECO format:** ($\approx$ 3 MB/event) This format is obtained after performing the reconstruction, which constitutes the most CPU-intensive step of the data processing. It comprises low and high-level information. At the lowest level, it contains reconstructed hits, clusters, and segments. Based on them, it stores tracks and vertices. Finally, it comprises the high-level physics objects (jets, electrons, muons, etc) at the highest hierarchy level.

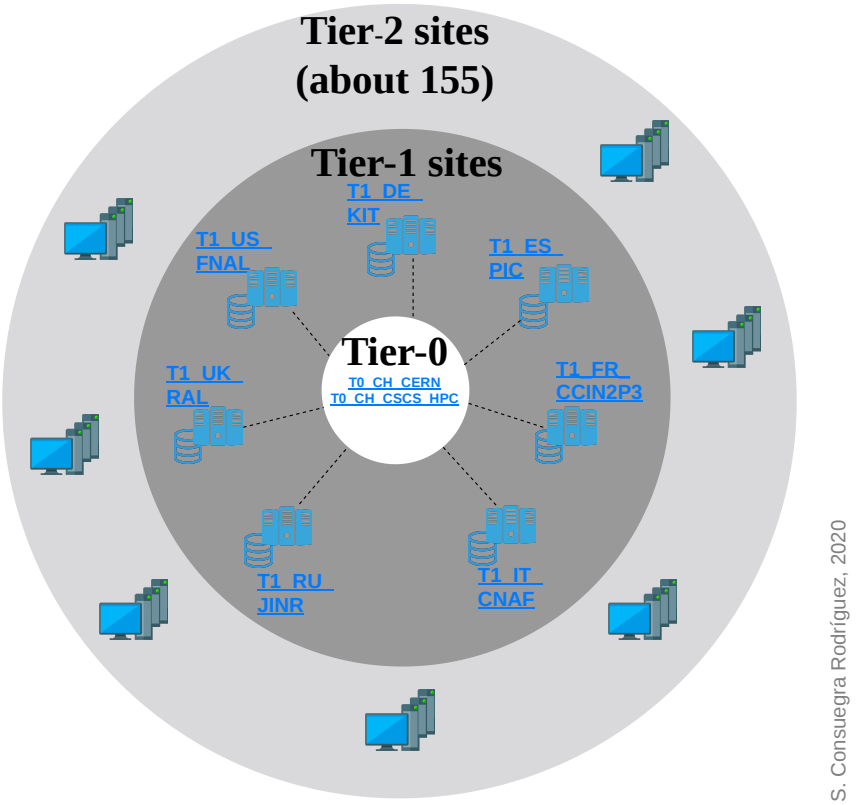


Figure 4.14: Overview of CMS computing model. The concentric circles represent the system of Tiers, with the Tier-0 center at CERN (in the internal circle), the several Tier-1 centers at regional computing centers (in the intermediate circle), and the many Tier-2 centers worldwide (in the external circle).

- **AOD format:** (400-500 kB/event) The RECO data is filtered to obtain the AOD format, which keeps the high-level physics objects and some additional information to perform kinematic refitting, containing only a portion of all the hits. During Run I, intermediate datasets denominated ntuples were produced by the physics analysis groups (PAGs) out of the AOD datasets. Additionally, individual groups made their custom ntuples out of these intermediate ntuples. With the increase of the data flow managed by the offline computing system in Run II, the production of the intermediate ntuples became unsustainable since they occupied a much-needed space and contained largely overlapping information among the different PAGs. Therefore, they were substituted with a new standard and condensed format called the MINIAOD format [147].

- **MINIAOD format:** ($\approx$ 100 kB/event) It was created to achieve a size of 10% of the Run I AOD format, while serving to about 80% of the CMS analysis. The MINIAOD contains high-level physics objects, comprising all the particle candidates, but only stores limited basic quantities (e.g. 4-vector, impact parameter, pdg id [38], and quality flags) with reduced numerical precision. The MINIAOD format also incorporates information on the simulated particles, trigger, and miscellaneous information (e.g., interaction vertices and E_T^{miss} filters).

- **NANOAOD format:** ($\approx$ 1 kB/event) A high-level of detail on the analyzed subset of collected events is often needed for calibration purposes, which implies the use of low-level detector information. On the other hand, high precision in low-level detector information is not required for searches and precision measurements. For this kind of analysis, the key is selecting a high number of events since they are often statistically limited. Thus, the flexibility of reducing the event size compared to the MINIAOD format depends on the type of analysis. The NANOAOD format was conceived to be used by at least 50% of the CMS physics analysis while reducing about 20 times the size compared to the MINIOAD format [148]. The format helps to deal more efficiently with the bigger dataset sizes that result from the consistent increase in luminosity during Run II and the foreseen luminosity increase in the upcoming periods of data taking. It contains only the top-level information usually employed in the final steps of the analysis, eliminating the track collection, considering thresholds for specific physics objects, and reducing the information stored on collections with many entries (e.g., jets). The NANOAOD format has already been used as the main data format for some Run II analysis, and its use is expected to expand further during the Run III.

The analysis presented in this thesis makes use of the MINIAOD format. A transition to the NANOAOD format was not contemplated due to the relevance of the track collection in the analysis.

CHAPTER

5

EVENT GENERATION, DETECTOR SIMULATION, AND RECONSTRUCTION

Contents

In this chapter, a description of the CMS simulation chain is presented, from the event generation to the simulation of the CMS detector and the readout electronics, up to the final simulated events. The CMS reconstruction techniques applied to both real data and simulated events are detailed, emphasizing in the techniques for the identification of the physics objects of interest for this work.

5.1 Event generation

The generation of events in CMS is handled in two steps. First, the pp collisions are simulated with a Monte Carlo event generator. Then, the passage through the different layers of the detector of the particles resulting from the step one is simulated in a second step referred to as detector simulation. Fig. 5.1 shows an overview of the complete CMS simulation chain, described hereafter.

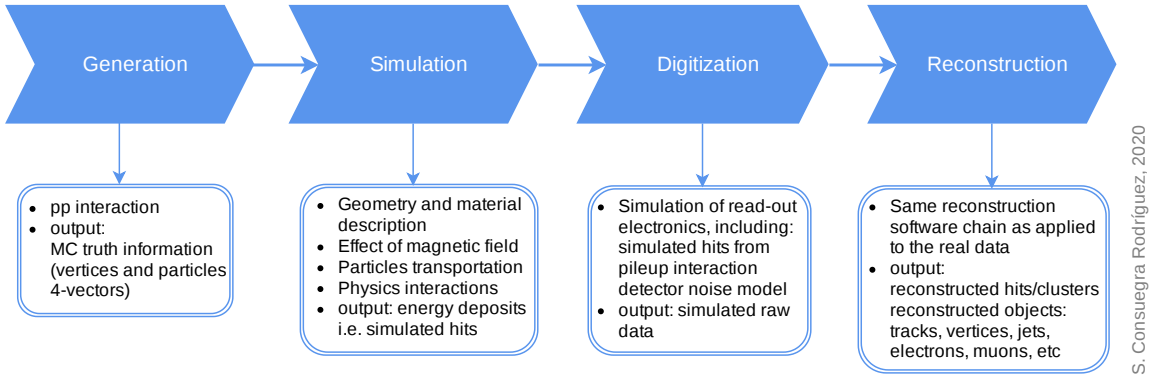


Figure 5.1: CMS simulation workflow.

5.1.1 Structure of the event

Over the past decades, the description with multi-purpose MC event generators of the final states obtained in high energy physics experiments has been improved [149]. The MC generators operate through many orders of magnitude of the momentum spectra, describing each of the phases of the process from the hard interaction to the final state, in which hundreds of particles are produced. Each of the phases of the process is associated with a step in the simulation chain of the MC event generator, as illustrated in Fig. 5.2.

Factorization of the cross-section

The treatment of the process of interest in the simulation is different according to the momentum transfer involved. The hardest parton-parton interaction is called primary hard process. If the momentum transfer is high, known as high scale, the partons inside one of the incoming hadrons interact with the ones from the other hadron and produce a small number of high energy partons, leptons, and gauge bosons. The cross-section for these processes with a large invariant momentum transfer can be written as:

$$\sigma_{h_1 h_2 \to X} = \sum_{a,b \in \{q,g\}} \int dx_a \int dx_b f_a^{h_1}(x_a, \mu_F^2) f_b^{h_2}(x_b, \mu_F^2) \int d\Phi_{ab\to X} \frac{d\hat{\sigma}_{ab}(\Phi_{ab\to X}, \mu_F^2)}{d\Phi_{ab\to X}}, \quad (5.1)$$

where X is the final state produced in the collision of the hadrons $h_{1,2}$, and $f_{a,b}^{h_{1,2}}(x_{a,b}, \mu_F^2)$ are the parton distribution functions in the collinear factorization.

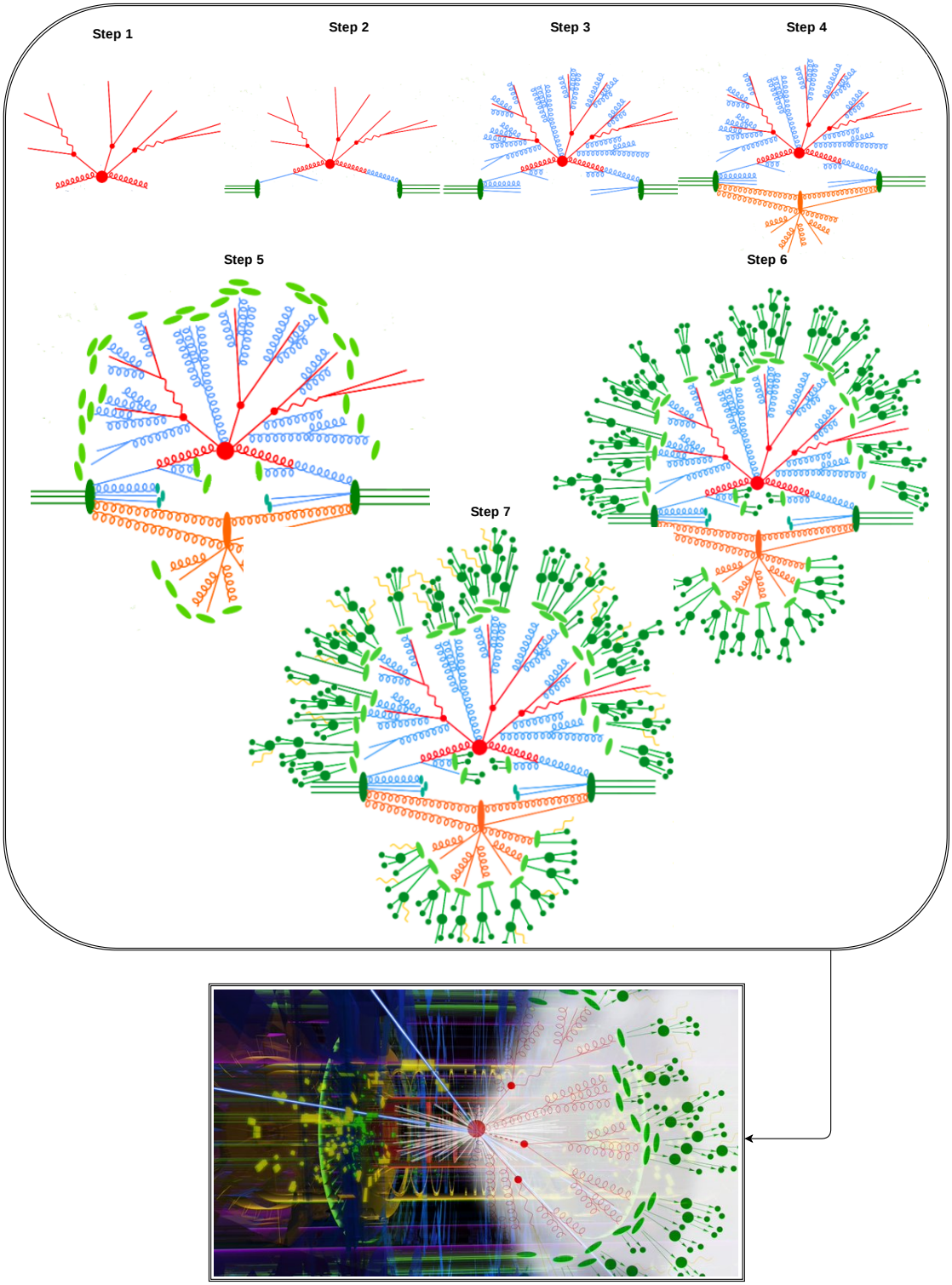


Figure 5.2: Overview of the event generation chain. Step 1: Hard scatter (red), matrix elements from first principles. Step 2: Incoming partons from the parton distribution functions (PDFs). Step 3: Radiative corrections, Initial State Radiation (ISR) and Final State Radiation (FSR) (blue). Step 4: Multi Parton Interaction (MPI) (orange). Step 5: Hadronization (light green). Step 6: Hadron decays from unstable resonances to final-state particles (dark green). Step 7: Photon radiation (occurs at any stage), QED corrections (yellow) [150, 151].

At leading order, the PDFs represent the probability of having a parton of flavor a, carrying a momentum fraction x in the parent hadrons $h_{1,2}$, at the factorization scale μ_F. The differential final-state phase-space element for the production of the final state X from the partonic initial state $d\Phi_{ab\to X}$ corresponds to the differential cross-section $d\sigma/d\Phi$. Fig. 5.3 shows the general picture of the primary hard process.

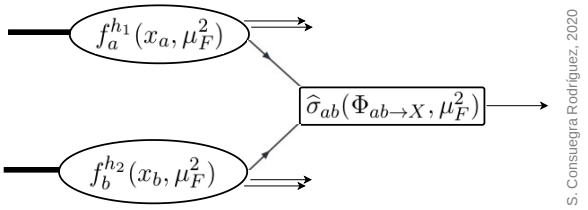


Figure 5.3: Schematic representation of a primary hard process.

Parton shower, evolution equation, and Sudakov factor

In the case of low momentum transfer, known as low scale (order of magnitude of 1 GeV) the partons from both incoming hadrons are confined, and the final-state hadrons are formed from the interaction of outgoing partons. An evolutionary process connects the low scale and the high scale. The many additional partons produced as a consequence of this scale evolution, in the form of the so-called parton showers, come from considering a probability for the addition of one or more partons to the final state during the evolution process in an interval of the evolution variable of the parton shower. The so-called parton shower algorithms are formulated as an evolution in a momentum-transfer-like variable and are simulated with a step-wise Markov chain, descending in momenta from a scale defined by the hard process. The evolution of the PDFs with changing factorization scale in collinear factorization takes the form:

$$\mu_F^2 \frac{df_a(x,\mu_F^2)}{d\mu_F^2} = \sum_{b\in\{q,g\}} \int_x^1 \frac{dz}{z} \frac{\alpha_s}{2\pi} \widehat{P}_{ba}(z) f_b(x/z, \mu_F^2), \tag{5.2}$$

where $\widehat{P}_{ba}(z)$ are the regularized Altarelli-Parisi splitting functions, characterizing the collinear splitting of parton b into parton a [152]. Eq. (5.3) constitutes the main equation to solve by the parton-shower MC generators:

$$\frac{d}{d\log t} \log \frac{f_a(x,t)}{\Delta_a(t_c,t)} = \sum_{b\in\{q,g\}} \int_x^{z_{\max}} \frac{dz}{z} \frac{\alpha_s}{2\pi} \widehat{P}_{ba}(z) \frac{f_b(x/z,t)}{f_a(x,t)}, \tag{5.3}$$

where $\Delta_a(t_c,t)$ is the Sudakov form factor (the probability for a parton not to undergo a branching process in an interval of time of the parton shower's evolution variable), calculated as:

$$\Delta_a(t,t') = \exp\{-\sum_{b\in\{q,g\}} \int_t^{t'} \frac{d\bar{t}}{\bar{t}} \int_{z_{\min}}^{z_{\max}} dz \frac{\alpha_s}{2\pi} \frac{1}{2} P_{ab}(z)\}. \tag{5.4}$$

The use of Sudakov factors for unresolved splittings and virtual corrections constitutes a common characteristic of the MC event generators. To fully reflect the complexity of the event

structure, the particles in the final state that do not come from the primary hard process are also considered and grouped into a component of the final state referred to as *underlying event*. An underlying event can be formed by:

- **Initial and Final State Radiation:** Gluon emission by the incoming partons before the hard interaction and gluon emission of the scattered partons.

- **Beam remnants (BBR):** Particles produced in the hadronization of the beam partonic constituents that did not engage in the hard interaction.

- **Multiple Parton Interactions:** The additional parton interactions to the main hard interaction. The two incoming hadrons are systems of strongly-interacting partons and therefore, it is likely that one or more pairs of partons interact with each other in the form of MPIs.

Hadronization and hadrons decay

Once an energy scale of the order of 1 GeV is reached, the perturbation evolution of the system can no longer be maintained, and a non-perturbative hadronization model is deployed to describe the confinement of the colored partons into hadrons. This process of obtaining hadrons out of quarks and gluons is called hadronization. The partons forming color-connected systems hadronize together, instead of each parton independently. Furthermore, the collective hadronization of color-connected systems is independent of how the system was produced in the first place. This means that once a model is tuned with data, it becomes predictive for new types of collisions and energy regimes. During the hadronization, unstable resonances are produced, unstable enough to decay inside the detector but stable enough to be detected before they decay. As a final step in the event simulation chain, the decay of these resonances into lighter hadrons is simulated.

Event generators

General-purpose event generators are used within CMS for massive event generation campaigns. They can be interfaced with Matrix Element (ME) generators, such that the output (parton-level events) can serve as input for the hadronization. Other more specific generators, such as the ME calculator MADGRAPH5_aMCATNLO, are also widely popular within the community. The event generators are incorporated in the software used for the analysis of the data (CMSSW) as external packages, with an interface provided by CMS-specific software. This allows to create the so-called configuration cards and tailor the generation task so that the obtained events correspond to the desired topology. A brief description of the most widely used MC event generators is given hereafter:

- **Pythia:** [153,154] This general-purpose event generator is quite popular within the LHC community, building up from the user experience at LEP, HERA, and the Tevatron. It has also been used in cosmic-ray and heavy-ion studies, comprising e^+e^-, pp, and $p\bar{p}$ collisions. All the physics aspects, hard and soft interactions, parton distributions, initial and final state parton showers, multiple parton interactions, fragmentation, and decay, are covered by PYTHIA in three main steps. The fist step, called Process Level,

uses a combination of matrix element expressions and phase space selection to choose the hard process. The second step, known as Parton Level, continues the evolution up to lower scales, including parton showering, MPIs, and beam remnants. The third step, called Hadron Level, handles the hadronization of the partons obtained as output from the previous step and simulates the decay of hadrons and leptons. PYTHIA is open to external input and, therefore, hard subprocesses can be added and the generation handled as an internal process. Other custom interfaces can also be interfaced with PYTHIA, e.g., FastJet for jet clustering.

- **Herwig++:** [155, 156] It stands for Hadron Emission Reactions With Interfering Gluons. HERWIG, its predecessor, developed during the era of LEP. Currently, HERWIG++ is a complete event generator that handles lepton-lepton, lepton-hadron, and hadron-hadron collisions. It comprises the automatic generation of hard processes for a comprehensive list of BSMs and the matching of the hard processes at NLO with the Positive Weight Hardest Emission Generator (POWHEG) method. The angular ordered parton showers, the cluster hadronization, and the modeling of the underlying events through hard and soft multiparton interactions, are also distinctive characteristics of HERWIG++. Well elaborated and advanced hadronic decay models, in particular for the case of bottom hadrons and τ leptons, are also integrated in HERWIG++.

The Shower MC (SMC) programs implement some approximate NLO corrections. Therefore, a possible overcounting needs to be considered when attempting to merge NLO calculations with parton shower simulations. MC@NLO and POWHEG constitute two different methods to overcome this overcounting problem:

- **MC@NLO:** [157] In this method the overcounting is avoided with the subtraction from the exact NLO cross-section of its approximation, as implemented in the SMC program to which the NLO computation is matched. Such an approximated cross-section is computed analytically and is SMC dependent. The events generated with MC@NLO can have negative weights in the cases in which the exact NLO cross-section minus the MC subtraction terms yield negative results. This does not imply a negative cross-section since physical distributions must turn out to be positive.

- **Powheg:** [158] The POWHEG Box implements the merging of NLO calculations and SMC programs following the so-called POWHEG method, which avoids negative weighted events. With this method, NLO calculations plus an initial state of parton shower are obtained and can then be fed into a SMC program for subsequent showering, without the problem of overcounting.

- **MadGraph5aMC@NLO:** [159, 160] It is a framework created from the merging of the MADGRAPH matrix element generator and the MC@NLO formalism. It comprises additional functionalities compared to the two predecessors, like the possibility to merge event samples with different light-parton multiplicities. Furthermore, it allows the computation of cross-sections at tree-level and next-to-leading order accuracy with MAD-

GRAPH, plus the parton shower with the MC@NLO formalism. The physical observables can be obtained with different perturbative accuracies and description of the final state, according to the following options:

- fLO: tree level + parton-level computation.
- fNLO: tree level and one-loop matrix elements + parton-level computation. For fLO and fNLO no parton shower is involved. The observables are reconstructed with the particles appearing as a result of the considered matrix elements.
- LO+PS/NLO+PS: matrix elements from fLO/fNLO computation, matched to parton showers.
- MLM-merged/FxFx-merged: a combination of LO+PS/NLO+PS samples, differing by the final-state multiplicity.

5.2 Detector simulation

The availability of a precise and realistic detector simulation contributes to the successful operation of a detector, more even so in the case of sophisticated detectors such as CMS. The detector simulation helps to test the design, supports the commissioning, and allows to assess the impact of planned upgrades. For the same collider, the simulation of pp collisions can be common for the different experiments since the physics processes are detector independent, but the detector simulation is custom for each experiment, adapting to its specific design characteristics. Among the experiment-specific characteristics are the geometrical structure of the detector, the detector response to the energy deposits (electric current and voltage signals), and the pileup. The MC simulation of the radiation transportation along the different detector layers starts from the beam pipe up to the end of the cavern. General codes grouped under the name of Monte Carlo radiation transportation codes are used. The simulation of the CMS detector is done with the GEometry ANd Tracking (GEANT4) package [161–163]. The output in the form of energy deposits provided by the transportation code is converted into electrical signals in a process known as digitization, followed by the calibration, in which these signals are translated into position, momentum, and energy measurements. Furthermore, a noise model for the detector and the pileup effect (studied through simulation of hits coming from pileup interactions) is taken into account. The resulting simulated events have the same format as the real collision events. Therefore, the same algorithms used in real data for the identification and reconstruction of particles and the computation of observables are used. A common final data format is thus obtained for real data and simulated events.

5.2.1 Main challenges

The balance between the precision of the simulation and the amount of computing resources used to obtain the results constitutes one of the challenges faced to provide the best detector simulation possible. Usually, more precise physics models would be slower and more resource consuming. The balance can be accomplished by simulating some processes in a condensed

form, such as the multiple scattering, incorporated as a net deflection resulting from all the small angle scatterings. A validation of the full simulation chain is performed by checking the agreement of physics observables between the MC prediction and the data measurement.

5.2.2 CMS event display

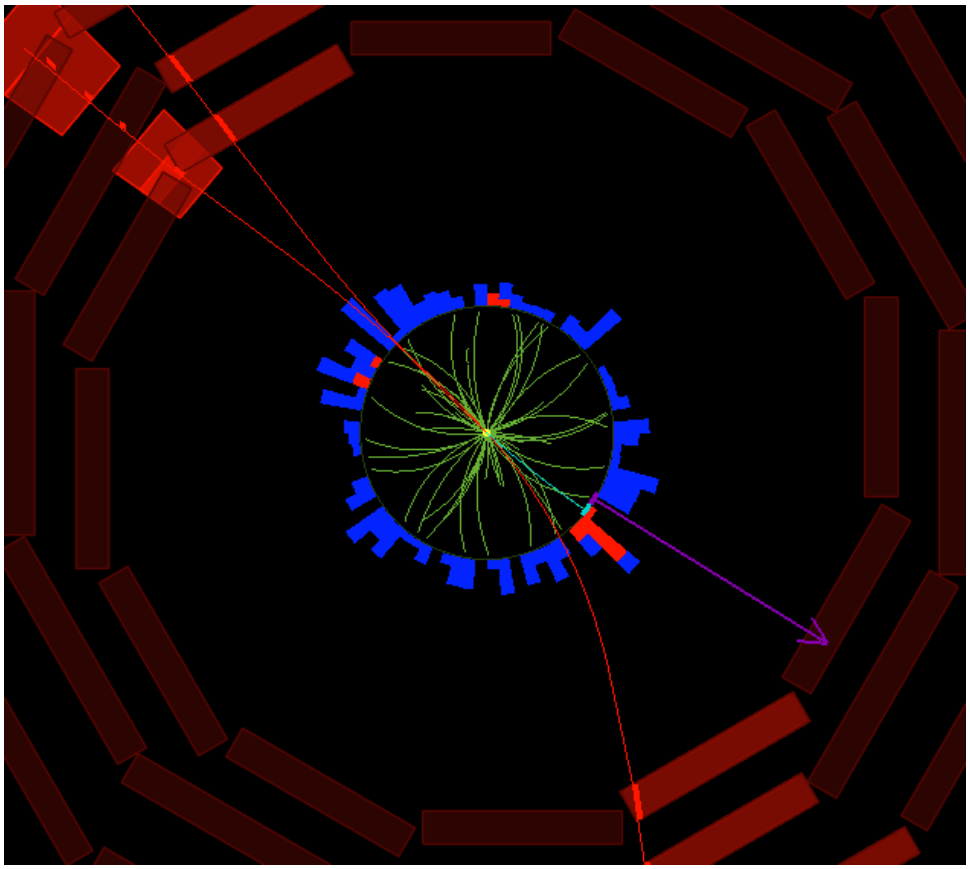
The simulated detector geometry together with the simulated or real data events can be inspected with a CMS event-display program [164]. The advanced geometry visualization package uses as input the MINIAOD format (Subsec. 4.2.7). Fig. 5.4a shows the event display of a $H \rightarrow a_1 a_1 (Z_D Z_D) \rightarrow \mu\mu\tau\tau$ simulated event, corresponding to the 2017 MC samples used for the analysis presented in this work, and Fig. 5.4b shows the energy deposit in the Φ-η plane of the four leptons in the final state.

5.3 Reconstruction of relevant physics objects

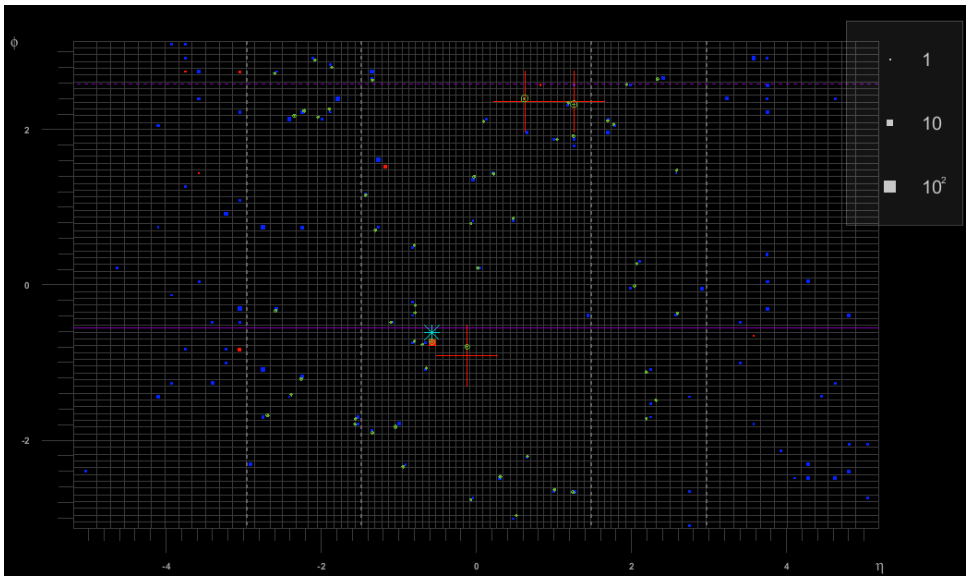
A comprehensive list of final-state particles can be reconstructed and identified with the CMS detector, due to the highly segmented subdetectors that provide enough separation between individual particles. A detailed description of the main physics objects of interest for the analysis presented in this work: primary vertices, muons, electrons, hadrons, jets, and τ leptons, is given below.

5.3.1 Primary vertex reconstruction

The location of an event and the associated uncertainty is measured in three consecutive steps: track selection, clustering, and a final fit for the extraction of the vertex parameters [110]. The reconstructed vertex with the largest total transverse momentum of physics objects is identified as the primary vertex. The physics objects refer in this case to the jets reconstructed from the tracks assigned to the vertex and the associated missing transverse momentum of these jets. The selected tracks used as input for the clustering most fulfill certain requirements, related to their prompt production in the primary interaction. The tracks are clustered on the basis of the value of their z-coordinate at the point of closest approach to the beam spot center. A compromise is made during the clustering process between the resolving power and a minimization of the incorrect splitting of true vertices.



(a)



(b)

Figure 5.4: a) CMS event display of a Higgs boson decaying to a pair of light bosons, which subsequently decay to a pair of muons and a pair of τ leptons. The signal signature is composed of a highly boosted pair of muons (upper left, red), acompanied by two nearly back to back decay products from the τ leptons, a third muon and an electron (light blue). The purple arrow denotes the MET of the event in the direction of the τ leptons decay. b) Energy deposit in the Φ-η plane.

To avoid the creation of fake vertices, each vertex is required to have at least two of their tracks incompatible with originating from other of the vertices. A weight between 0 and 1 is assigned to each track, corresponding to the likelihood for the track to belong to the vertex. An adaptive vertex fitter assigns the tracks to the vertex according to the compatibility characterized by the assigned weight. Finally, a fit is performed to determine the vertex parameters (e.g., x, y, and z coordinates).

5.3.2 Particle flow, link, and post-processing algorithms

The specific function of each of the CMS subdetectors in the reconstruction of physics objects makes possible to a large extent to reconstruct some of these objects using information from only one of the subdetectors [108], e.g., jets formed by hadrons and photons can be constructed exploiting calorimeter information without additional input from the tracker and the muon system. In addition, jets originated from hadronic τ decays or the hadronization of a b-quark mainly concern the tracking system since their identification is based on the properties of the charged tracks. The identification of isolated photons and electrons can be performed mainly with ECAL information, while the muon identification mostly concerns the muon system. A significant improvement on the event description comes from correlating all the basic elements obtained in the different subdetector layers and combining all the pieces of information to reconstruct the properties of the final state particles. This approach is implemented in the so-called *Particle Flow algorithm* (PF) [165]. The *link algorithm* connects the PF elements from the subdetectors that are assumed to belong to the same particle. The quality of a created *link* (geometrical connection between two PF elements) is quantified by the distance between the two elements. A representative example of the measurement of this distance would be the link between an inner track and a calorimeter cluster, where the distance is measured between the extrapolated track position and the cluster position in the η-Φ plane. The elements associated with a direct or indirect link are grouped into PF blocks. After processing the PF blocks and identifying all the particles, the global event description is obtained. The basic requirements for the identification of each of the main physics objects are the following:

- **charged hadrons:** Presence of a link between one track and one or more calorimeter clusters in the η-Φ plane. No associated signal in the muon system.
- **photons and neutral hadrons:** ECAL and HCAL clusters. No associated track link.
- **electrons:** A track and an ECAL cluster, momentum-to-energy ratio compatible with the unity, no associated HCAL cluster.
- **muons:** A track in the inner tracker connected to a track in the muon system.

The reconstruction and identification with the two algorithms described above allows an efficient combination of the information from all the subdetectors. Nevertheless, a remaining particle misreconstruction and misidentification rate is left, which is further reduced with the help of the *post-processing algorithm*. The algorithm investigates events with large artificial

missing transverse momentum (p_T^{miss}), caused by:

- cosmic-ray muons passing through the detector at the same time of an LHC beam crossing.

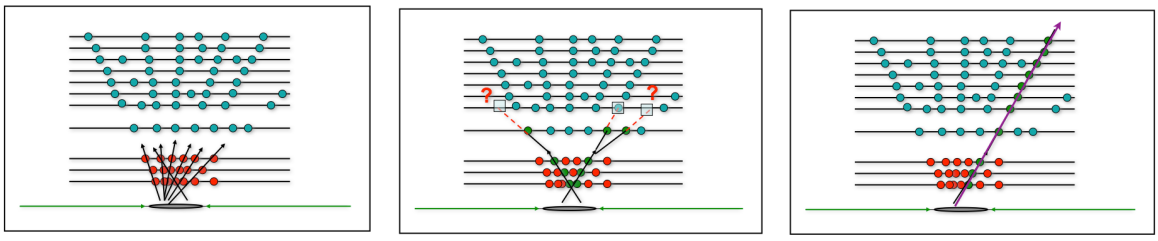
- severe misreconstruction of the muon momentum (incorrect determination of the momentum by the PF algorithm, caused among other reasons by an incorrect inner track association).

- particle misidentification (e.g., charged hadron with energy deposit in the muon system, misidentified as a muon. This results in the addition of a neutral hadron to the particle list, to account for the energy left by the charged hadron in the calorimeters. Unrecovered muon that fails the tight identification selection and overlaps with a neutral hadron. The overlap causes the neutral hadron to be lost from the particle list).

After the event post-processing, the artificial p_T^{miss} is largely reduced, without a relevant effect to the genuine high p_T^{miss}, which might be a sign of new physics phenomena.

Tracks

The collection of tracks is of central importance for the analysis presented in this work. Tracks are reconstructed in CMS in three stages, as shown in Fig. 5.5. A combinatorial track finder based on Kalman Filtering (KF) is used. First, an initial seed (a combination of 2-3 hits) is generated, containing only a few hits compatible with the path of a charged particle. Second, the trajectory is built by collecting the hits in all the detector layers compatible with this path. In the final step, a fit is performed to determine the origin, transverse momentum, and direction of the charged particles.



(a) Seeding (b) Building (c) Fitting

Figure 5.5: Illustration of the three steps of the CMS track reconstruction [166].

The reconstruction efficiency is calculated as the fraction of reconstructed tracks matched with a simulated track, measured in a sample of simulated events. At the same time, the misreconstruction rate is defined as the fraction of reconstructed tracks that can not be associated with a simulated track, both with respect to the total number of simulated tracks. The misreconstruction rate is kept low by applying tight track quality criteria at the expense of losing some reconstruction efficiency. Fig. 5.6 shows the reconstruction efficiency (left) and misreconstruction rate (right) measured in a sample of simulated $t\bar{t}$ events, with a mean number of 50 pileup (PU) interactions and the conditions of the CMS detector in June of 2018. The simulated tracks fulfill the following kinematic requirements, a $p_T > 0.9$ GeV and

$|\eta| < 2.5$. No significant difference is observed in the fake rate when considering a realistic detector with respect to the ideal detector, while a loss in tracking efficiency of about 5% is observed around 20 GeV.

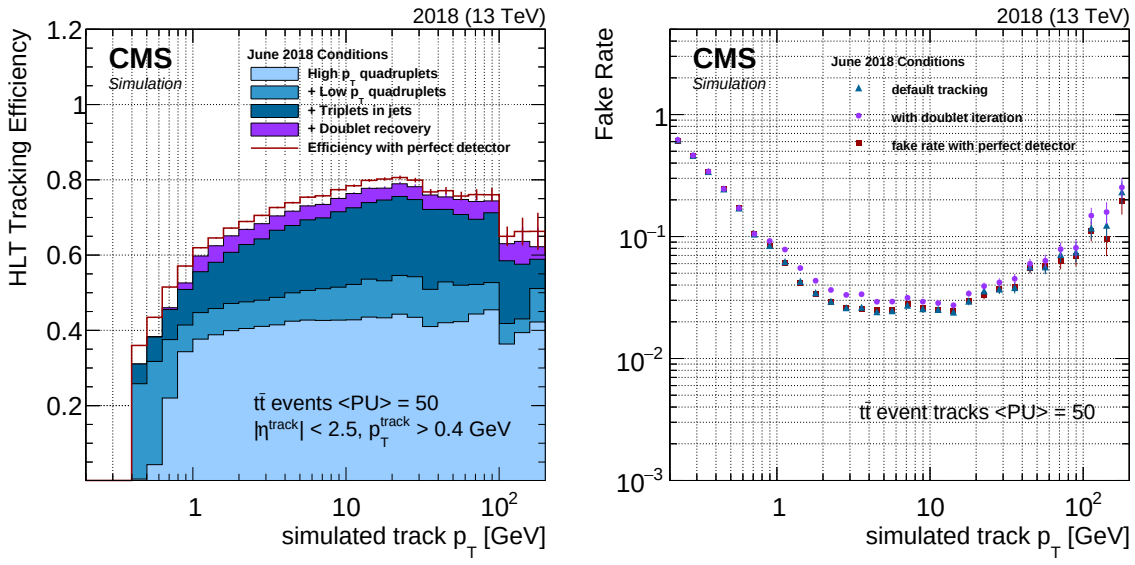


Figure 5.6: Tracking efficiency (left) and misreconstruction rate (right) as a function of the simulated track p_T. The contributions to the total efficiency resulting from different tracking iterations are shown in different colors. The performance that would be achieved with a perfect detector is shown as a red line (left) and squares (right) [167].

Fig. 5.7 shows the dependence of both quantities on the number of PU interactions. The fake rate increases with the additional interactions while the tracking efficiency shows no significant deterioration. To keep and increase the tracking efficiency while maintaining a similar fake rate constitutes a critical feature for a successful PF event reconstruction. This is accomplished with a combinatorial track finder applied in several iterations in a process known as iterative tracking. At each step of the iterative procedure, the hits connected to the tracks in the current iteration are removed before the next iteration, to reduce the random hit-to-seed association probability. Furthermore, the quality criteria are relaxed at each step, to increase the total tracking efficiency without degrading the purity. The seeds for the first three iterations are triplets of pixel hits, fulfilling additional criteria on the distance of closest approach to the beam axis. The forth and fifth iterations tailor the reconstruction of tracks with one or two missing hits in the pixel detector, while the sixth and seventh iterations are dedicated to the reconstruction of displaced tracks with no pixel hits. The eighth iteration aims to deal with tracks inside high p_T jets, where an attempt to disentangle the merged pixel hits is made, so that merged nearby tracks can be distinguished. Information from the muon system is added in the ninth and tenth iterations, to increase the muon-tracking efficiency. The iterations in which tracks are formed with a seed containing at least one pixel hit (1-7) allow to recover around half of the tracks with $p_T > 1$ GeV that are not reconstructed by the combinatorial track finder in a single step. A moderate improvement in the misreconstruction rate is also obtained. The iterative tracking is faster than a single step due to the smaller number of seeds at each stage. Thus, it became the default method in CMS.

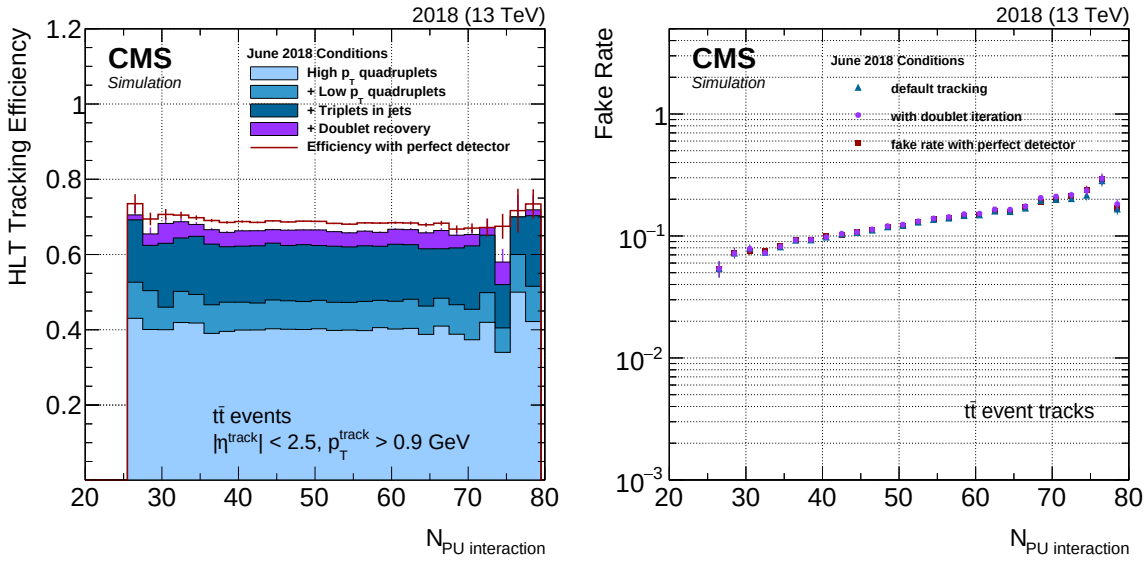


Figure 5.7: Tracking efficiency (left) and misreconstruction rate (right) as a function of the number of pileup interactions. The contributions to the total efficiency resulting from different tracking iterations are shown in different colors. The performance that would be achieved with a perfect detector is shown as a red line (left) and squares (right) [167].

5.3.3 Muons

Muons are identified with a high efficiency by the muon system and a precise measurement of their momentum involves the inner tracker. The following muon collections are relevant for the offline muon reconstruction:

- **standalone muon:** Muon-track resulting from the fit of the DT, CSC, and RPC hits found along the muon trajectory in the muon system. For the trajectory building of the track, fragments formed by seeds from DT and CSC detectors are used. Fig. 5.8 shows the standalone muon reconstruction efficiency for the 2016 and 2017 datasets as a function of the muon η, for muons with $p_T > 53$ GeV.

- **global muon:** Muon-track resulting from the matching of a standalone-muon track and an inner track. An updated global fit including the hits from both tracks is made, improving the momentum resolution of high p_T muons compared to the measurement obtained with a tracker-only fit. Fig. 5.9 shows the global muon reconstruction efficiency for the 2017 dataset as a function of the momentum, for muons with $p_T > 53$ GeV. A distinction is made between the muons produced in events without showering and with at least one shower.

- **tracker muon:** Muon-track resulting from the matching of an inner track and at least one muon segment. Their efficiency is higher than that of the global muons for muons with low momentum that do not satisfy the global muon requirement of having segments associated with more than one muon subsystem. If one tracker and global muon share the same inner track, they are merged into a unique candidate.

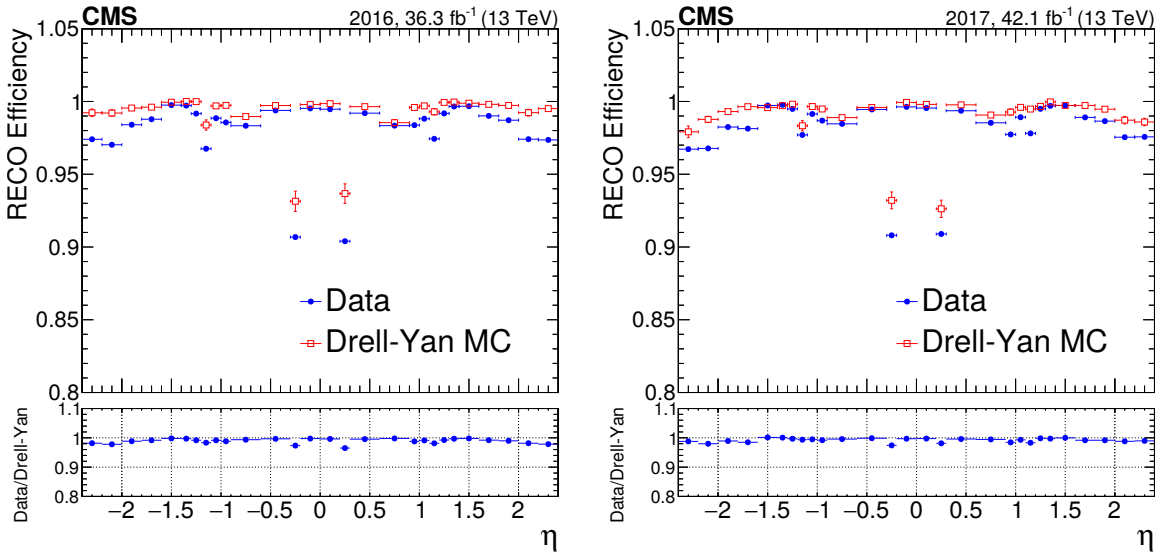


Figure 5.8: Standalone muon reconstruction efficiency as a function of the muon pseudorapidity for the 2016 (left) and 2017 (right) datasets. Data points are shown in blue, while the empty squares in red correspond to simulation [168].

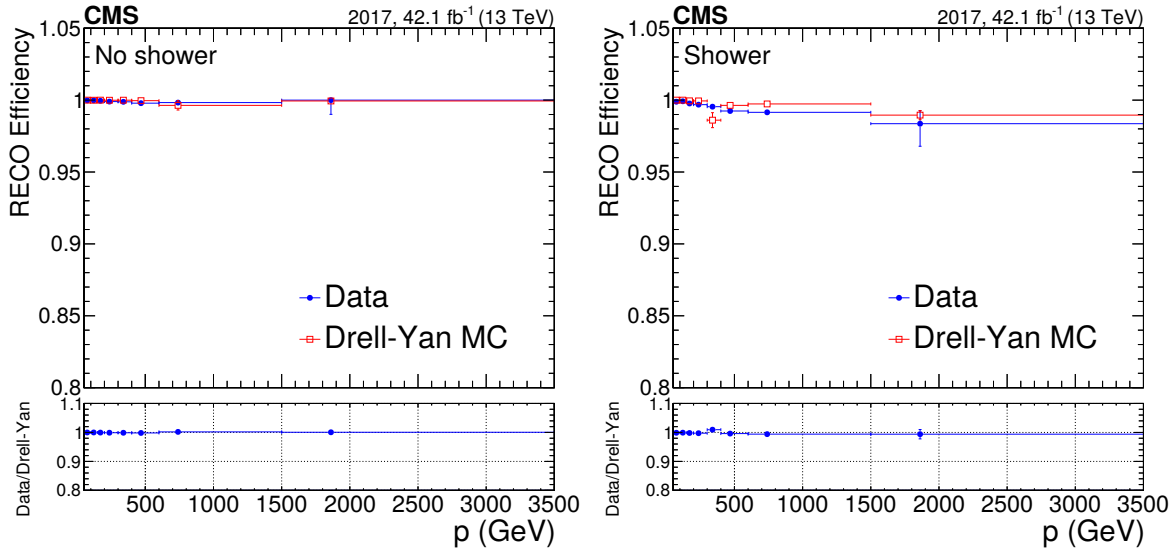


Figure 5.9: Global muon reconstruction efficiency as a function of the muon momentum for the 2017 dataset. The left plot corresponds to events without showering and the right plot to events with at least one shower. Data points are shown in blue, while the empty squares in red correspond to simulation [168].

The charged hadrons with energy deposits in the muon system due to hadron shower remnants can pass the muon identification criteria. This kind of misidentification, known as punch-through, is solved with the addition in the PF muon identification algorithm of information on the calorimeters' energy deposits. The procedure followed by the PF algorithm for the muon identification depends on the conditions of isolation of the muon. In the case of isolated global muons, the E_T of the energy deposits in the calorimeters and the p_T of the tracks within a

ΔR cone of 0.3 are required to be less than 10% of the p_T of the muon. These criteria allow a successful rejection of misidentified hadrons. A tight muon selection is applied to identify non-isolated global muons (e.g., muons inside jets). The accidental matching of tracker and standalone muons and the misidentified high-p_T hadrons due to the punch-through, are removed by requiring at least three matching track segments in the muon system. Muons may fail the tight-muon selection criteria if their inner track is poorly reconstructed (incorrect association of hits from nearby tracks) or because of a low-quality global fit. The muons lost due to a poor reconstruction can be recovered if they leave a significant amount of hits in the muon system and the track fit of the corresponding standalone muon is of high quality. The muons lost because of a low-quality global fit can be retrieved if a high-quality fit is obtained using at least 13 of the inner tracker hits. Finally, the PF elements that compose the identified muons are removed from the PF blocks to prevent their use in the reconstruction of other particles [123].

5.3.4 Electrons

The traditional electron seeding strategy, known as ECAL based electron seeding, exploits the information from the energetic ECAL clusters. Considering the electron bending in the magnetic field, a supercluster is formed gathering the ECAL clusters from a small window in η, but a wide window in Φ. The energy deposits from the electrons can largely overlap with the deposits from other particles and the backward propagation to the tracker is often compatible with hits that might not correspond to the electrons, causing a high misidentification rate. To reduce the misidentification, tight isolation requirements are applied to the ECAL based electron seeding, with the consequent loss of efficiency. To recover the lost electron tracking efficiency, a PF based reconstruction method with a tracker-based seeding is applied. The use of a Gaussian Sum Filter (GSF) allows us to deal with tracks experiencing large bremsstrahlung effects. All tracks with $p_T > 2$ GeV resulting from the iterative tracking are used as potential electron seeds. After the clustering in the calorimeter and the track-cluster matching, both collections are merged. The resulting collection serves as input to the full electron tracking. A sketch depicting the different elements considered for the reconstruction of the electrons is shown in Fig. 5.10.

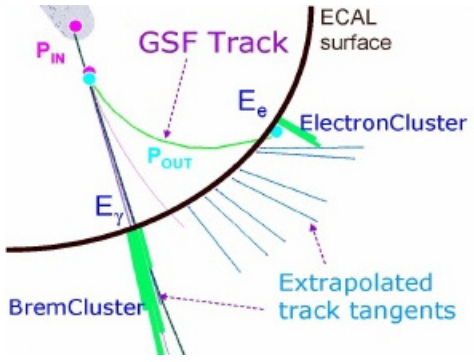


Figure 5.10: Sketch representing the components of CMS electron reconstruction [140].

Electrons often emit bremsstrahlung photons while interacting with the detector material. Therefore, the reconstruction of isolated photons and electrons is closely related. The electron candidate is seeded from a track electron if it matches an ECAL cluster not linked to three or

more additional tracks. Furthermore, the measured energy in the HCAL cells within a ΔR cone of 0.15 around the electron candidate must not exceed 10% of the supercluster energy. Additional criteria as the ratio between the HCAL and ECAL energies serve as input to a boosted decision tree (BDT), trained independently for isolated and non-isolated electrons. The training of the BDT is further differentiated for the ECAL barrel and endcaps. The electron reconstruction workflow is summarized in Fig. 5.11.

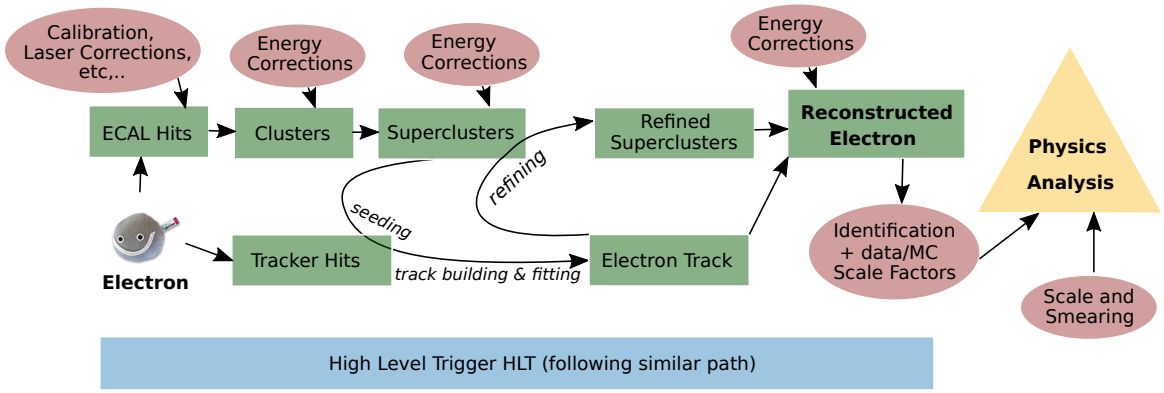


Figure 5.11: Simplified overview of CMS electron reconstruction workflow [169].

5.3.5 Hadrons

Hadrons constitute the remaining particles pending identification after the removal from the PF blocks of the already identified muons, electrons, and isolated photons. HCAL clusters with no associated tracks are identified as neutral hadrons (e.g., K_0^L or neutrons) and ECAL clusters not linked to any track as photons. After the allocation of these clusters, the remaining unassociated HCAL and ECAL clusters are linked to inner tracks to form single charged hadrons ($\pi^{\pm}$, $K^{\pm}$, and neutrons).
The clustering algorithm in the calorimeters contributes to the measurement of the energy and direction of the neutral hadrons. First, the cells with an energy standing out from their neighboring cells and with a value larger than a considered threshold are identified and cluster seeds formed. A topological cluster is formed by adding the cells with an energy that is twice the noise level and that share a corner in common with cells forming the seed cluster. Calorimeter clusters that do not match the extrapolated position of charged particle tracks constitute a signal of the presence of neutral particles. If the neutral particles overlap with charged particle clusters, the disentanglement can be difficult. In this case, the neutral particles can only be detected as calorimeter energy excesses. A good calibration maximizes the probability of a correct identification and energy determination of the neutral hadrons. Thus, if the response of the calorimeters to photons and hadrons, i.e., the calibration, is well known, the identification issues can be overcome. The calibration of electromagnetic and hadron clusters is described in the segment that follows.

Energy calibration

The calibrated energy for photons in the ECAL takes the form:

$$E^{\text{calib}} = \alpha(E^{\text{true}}, \eta^{\text{true}})E_{ECAL} + \beta(E^{\text{true}}, \eta^{\text{true}})[E_{PS1} + \gamma(E^{\text{true}}, \eta^{\text{true}})E_{PS2}], \tag{5.5}$$

where E_{PS1} and E_{PS2} are the energies measured in the two preshower layers, and α, β, and γ are calibration parameters that depend on the true photon energy and pseudorapidity. The calibration parameters are chosen so that the chi-square:

$$\chi^2 = \sum_{i=1}^{N_{\text{events}}} \frac{(E_i^{\text{calib}} - E_i^{\text{true}})^2}{\sigma_i^2}, \tag{5.6}$$

is minimized [108]. The σ in the denominator is the estimate of the energy measurement uncertainty. The response of the ECAL to hadrons is different than to photons. It depends on the fraction of the shower energy deposited by the hadron in the ECAL and the dependence is not linear with the energy. Therefore, an independent calibration specially tailored for hadrons is used, which takes the form:

$$E_{\text{calib}} = a + b(E)f(\eta)E_{ECAL} + c(E)g(\eta)E_{HCAL}, \tag{5.7}$$

where E and η are the true energy and pseudorapidity of the hadron. E_{ECAL} and E_{HCAL} constitute the fraction of the hadron energy measured in the ECAL (calibrated using Eq. (5.5)) and in the HCAL, respectively. The following chi-square is defined for a given value of a and for each energy bin:

$$\chi^2 = \sum_{i=1}^{N} \frac{(E_i^{\text{calib}} - E_i)^2}{\sigma_i^2}, \tag{5.8}$$

minimized with respect to b and c. E_i is the E^{true} energy defined above for the hadron i and σ_i is the corresponding energy resolution. Independent values of b and c are determined for the barrel and endcap regions. Separate coefficients are also calculated for hadrons that leave all their energy in the HCAL, with respect to those that leave some of it in the ECAL.

5.3.6 Jets

Jets are classified in CMS according to the type of input particle, the jet algorithm used for the clustering, and the jet size, as summarized in Fig. 5.12. Three different methods are available for the reconstruction of the jets: a calorimeter-based approach from which the calorimeter jets (CaloJets) are obtained, the Jet-Plus-Track approach, which improves the measurement of calorimeter jets with the addition of information from the associated tracks, and the Particle Flow approach, in which the particle flow candidates are used to build the jets [170]. Fig. 5.13a shows a sketch of a pp collision featuring the collision point, the hadronization of quarks and gluons, and the resulting collimated spray of particles clustered to form a jet.

Jets clustered from PF candidates can be classified according to the applied pileup reduction technique as Charge Hadron Subtraction (CHS) and Pileup per Particle Identification (PUPPI) [171] jets, both depicted in Fig. 5.13b. In the case of CHS jets, charged PF candi-

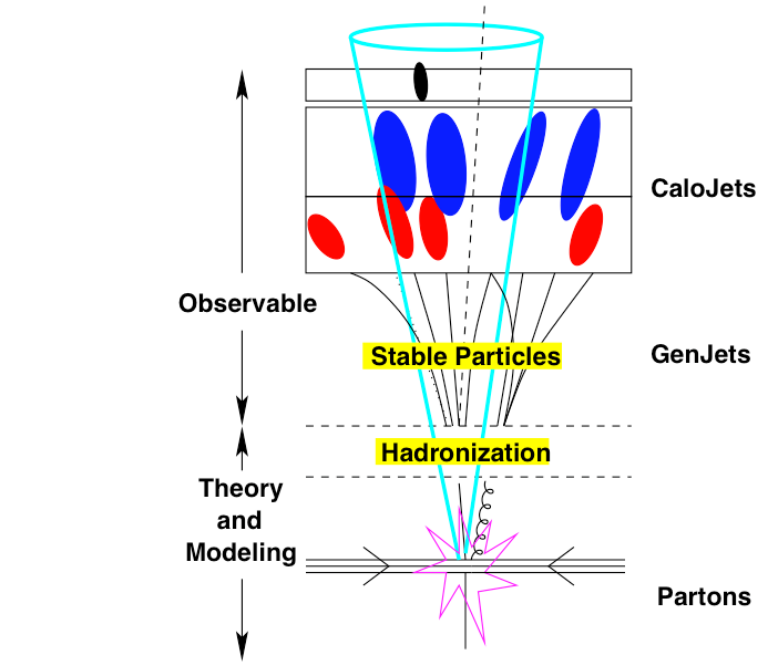
Figure 5.12: Overview of CMS jet classification criteria.

dates originating from pileup vertices are removed before proceeding with the clustering. For PUPPI jets, the PUPPI algorithm assigns a weight between 0 and 1 to each particle. Particles coming from pileup interactions will be assigned a low weight. The obtained weight is applied to each particle's four-momentum so that the ones with low weight or momentum get discarded. Finally, the pileup-corrected particles serve as input for the clustering, performed with algorithms such as the Cambridge Aachen (CA), k_T, and anti-k_T algorithms [172]. For the anti-k_T algorithm, the following parameters are defined:
d_{ij}: distance between entities i and j (particles, jets),

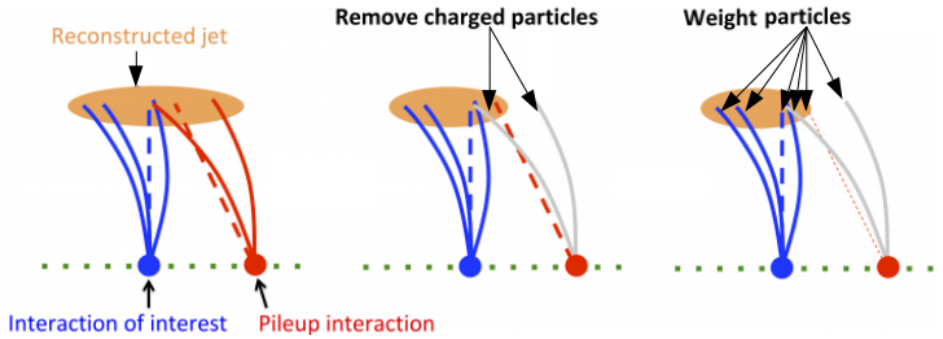
$$d_{ij} = \min(k_{ti}^{2p}, k_{tj}^{2p}) \frac{\Delta_{ij}^2}{R^2} \quad [\text{CA}\,(p=0),\ k_T\,(p=1),\ \text{anti}-k_T\,(p=-1)], \tag{5.9}$$

where:

$$\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2. \tag{5.10}$$



(a)



(b)

Figure 5.13: a) Sketch of a pp collision, depicting the generator level and calorimeter jets [173]. b) Sketch of an interesting interaction overlapping in the detector with a pileup interaction, along with a representation of the CHS and PUPPI pileup mitigation techniques [174].

d_{iB}: distance between entity i and the beam B,

$$d_{iB} = k_{ti}^{2p}. \tag{5.11}$$

The parameters k_{ti}, y_i, and ϕ_i are the transverse momentum, rapidity, and azimuthal angles of entity i. The first step of the algorithm is the determination of the smallest distance d_{ij} or d_{iB}. If it is a d_{ij}, the entities i and j are merged. If it is a d_{iB} the entity is called a jet and removed from the list of entities. The computation of the distance d_{ij} is done recursively until no entry is left in the list of entities. Once the clustering step is finished, the jet momentum is determined by summing up the momenta of all the particles within the jet. From simulation studies, this value is found to be within 5 and 20% of the true jet momentum, for the full p_T spectrum and detector pseudorapidity coverage.

Jet energy corrections

Jet energy corrections (JEC) are determined and applied to make the final measured response of the reconstructed jets match on average the response of the particle level jets [175]. To each component μ of the raw jet four momenta an overall correction is applied as a multiplicative factor:

$$p_{\mu}^{\text{cor}} = C \cdot p_{\mu}^{\text{raw}}, \tag{5.12}$$

where:

$$C = C_{\text{offset}}\left(p_T^{\text{raw}}\right) \cdot C_{\text{MC}}\left(p'_T, \eta\right) \cdot C_{\text{rel}}\left(\eta\right) \cdot C_{\text{abs}}\left(p''_T\right). \tag{5.13}$$

The various components of the overall correction factor C are applied in sequence and can be summarized as follows:

- C_{offset}: offset corrections, to subtract contribution from electronic noise and pileup, e.g., jet-area-based corrections to the four-momenta of CHS jets applied on an event-by-event basis.

- C_{MC}: MC calibration factor on the ratio of the reconstructed and generator level jet ($R = p_T^{\text{RECO}} / p_T^{\text{GEN}}$), known as jet response. This calibration is applied to match the reconstructed jet energy with the generator level jet energy. The average correction in each bin of p_T^{RECO} needed to restore the ratio to 1 ($C_{\text{MC}}\left(p_T^{\text{RECO}}\right) = 1/\langle \text{R} \rangle$) is called the *Jet Energy Scale* (JES). Issues as the fact that parts of jets can point towards regions beyond the detector acceptance, the existence of p_T thresholds, and the non-linear response of the CMS calorimeters, are accounted for by the JES.

- C_{rel} and C_{abs}: residual calibrations after the previous corrections.

  - C_{rel}: Variations of the response as a function of the pseudorapidity η.

  - C_{abs}: Absolute scale determined in the reference region ($|\eta| < 1.3$), which corresponds to the center of the detector, measured in Z+jet and γ+jet events.

In Eq. (5.13), p'_T and p''_T correspond to the jet transverse momentum after applying the offset and all the corrections, respectively. The *Jet Energy Resolution* (JER) is defined as the σ of a Gaussian fit to the distribution of R in the range $[m - 2\sigma,\, m + 2\sigma]$, where m and σ are the mean and width of the Gaussian.

b-jet identification

Once the jets are identified, different algorithms can be used for the identification or tagging of jets originating from the hadronization of b-quarks, exploiting the distinctive characteristics of b-jets. B-hadrons can travel millimeters away from the primary vertex before decaying, due to their relatively large lifetime, and the displaced tracks from their decay can form a secondary vertex. The magnitude of the displacement of the tracks is characterized by the

impact parameter (IP) value, defined as the distance between the primary vertex and the tracks at their point of closest approach to the primary vertex. Fig. 5.14 presents a sketch of the decay of a heavy flavor hadron, showing the main interaction vertex, the secondary vertex, and the displaced tracks. The impact parameter resolution d_0/σ, where d_0 represents the three spatial dimensions (3D) impact parameter measurement, has a large and positive tail for b-jets. A track is said to be produced upstream if it has a positive impact parameter. The impact parameter value can also be defined in the plane transverse to the beamline (2D) and in one dimension along the beamline, known as longitudinal impact parameter.

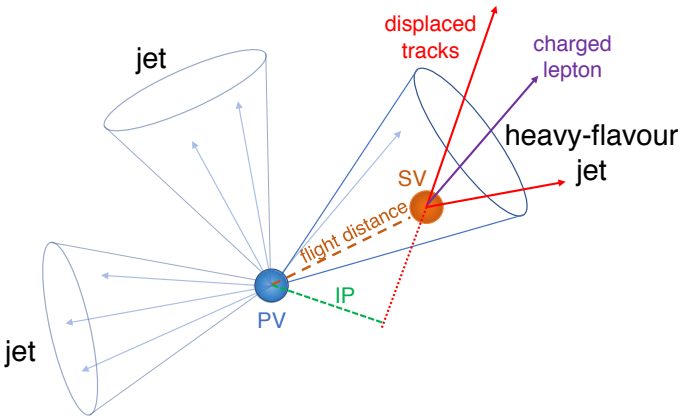
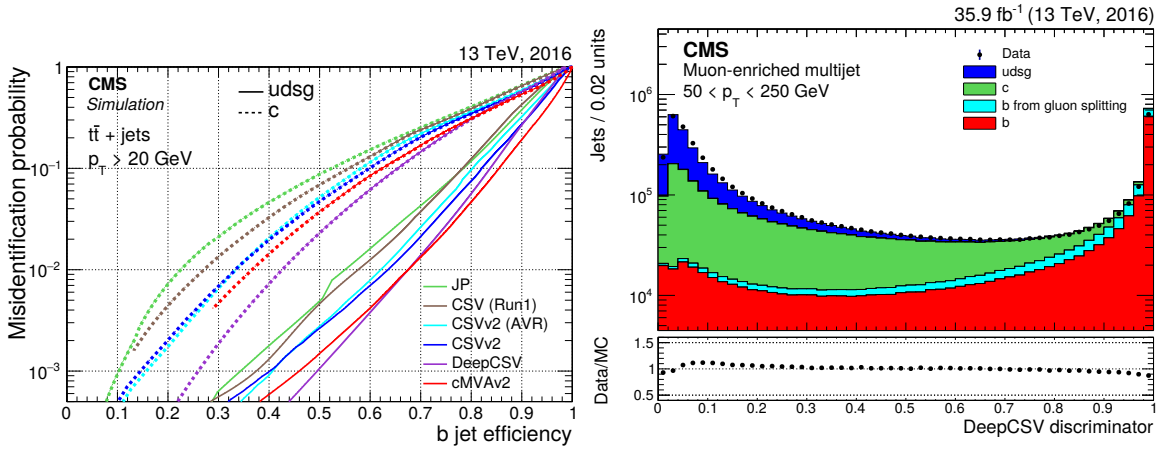


Figure 5.14: Illustration of a heavy flavor jet from the decay of a b or c hadron, depicting the main interaction vertex, the secondary vertex, and the displaced tracks [176].

The decay products from b-jets have a larger transverse momentum relative to the jet axis compared to decay products from light partons due to their relatively large mass and harder fragmentation. In addition to the properties of the reconstructed secondary vertex and the displaced tracks, the presence of soft leptons in the final state due to the sizable branching fraction for semileptonic decays constitutes another distinctive characteristic exploited by the b-jet identification techniques. In Run I, a Combined Secondary Vertex (CSV) algorithm based on the secondary vertex and track-based lifetime information was used for the identification of b-jets [177]. During Run II, the CSV algorithm was optimized, and in a new version referred to as CSVv2 [176], the discriminating variables serve as input to a neural network with one hidden layer. Another version of the CSV algorithm, operating with a very similar set of observables used in CSVv2, has become available. This new version is referred to as DeepCSV [178], and performs a deep neural network training with four hidden layers, having each layer a width of 100 nodes.
The different algorithms provide as output a value of the discriminant for each jet. A set of so-called working points are selected so that an average b-tagging efficiency for a given misidentification rate is obtained. The misidentification probability for c and light-flavor jets as a function of the b-jet identification efficiency, applied to jets in $t\bar{t}$ events, is shown in Fig. 5.15a. The different tagging algorithms used during Run I and Run II are included. The DeepCSV algorithm has the best performance among all the b-jet identification algorithms, except when discriminating against light-flavor jets for b-jet identification efficiencies above 70%, where the cMVAv2 tagger performs better. Fig. 5.15b shows the distribution of the DeepCSV discriminator in data compared to simulation for jets in a muon-enriched multijet

sample. The small discrepancies observed for low values of the discriminator are corrected by applying a data-to-simulation scale factor.



(a) (b)

Figure 5.15: a) Misidentification probability for c and light-flavor jets as a function of the b-jet identification efficiency applied to jets in $t\bar{t}$ events. b) DeepCSV discriminator distribution in data compared to simulation for jets in a muon-enriched multijet sample. The simulation is normalized to the number of entries in data [176].

5.3.7 Tau leptons

The tau leptons are the heaviest leptons, with a mass of 1.777 GeV/c^2 and a lifetime of 2.9 $\times 10^{-13}$ s ($c\tau = 87$ μm). They decay leptonically or hadronically, with a branching fraction of 35% and 65%, respectively. Due to their short life time, the decays occur inside the CMS beam pipe. The intermediate resonances and decay branching fractions of the tau lepton are summarized in Tab. 5.1. The decay products excluding the neutrinos are called visible decay products. The standard CMS electron or muon IDs are used for the identification of the leptonic decays, while for hadronic decays the Hadron Plus Strip Algorithm (HPS) is used. The dominant hadronic decays consist of one or three charged π mesons and up to two π_0 mesons in the final state. The π_0 mesons decay into pairs of photons that subsequently convert into e^+e^- pairs with a high probability, due to the multiple scattering and bremsstrahlung during the interaction with the detector material.
How often a true τ lepton is reconstructed at CMS, i.e., the τ identification efficiency, is calculated as follows:

$$\tau_{h\ \text{Efficiency}}^{\text{RECO.+ID.}} = \frac{\text{Denominator \& } p_T > 20\text{ GeV \& } |\eta| < 2.3 \text{ \& Isolation \& Decay Mode Finding}}{\text{Gen. visible } p_T > 20\text{ GeV \& Gen. visible } |\eta| < 2.3}. \tag{5.14}$$

The numerator corresponds to the reconstructed visible tau and the denominator to the generator level visible tau [180]. The term *Decay Mode Finding* refers to the identification of the tau decay mode. A matching of the generator level and the reconstructed taus is done. The reconstructed taus that matched some generator level tau are required to pass the identification requirements. Similarly, how often a jet fakes a τ lepton, i.e., the fake τ

Decay Mode	**Resonance**	$\mathcal{B}(\%)$
Leptonic decays		35.2
$\tau^- \to e^- \bar{\nu}_e \nu_\tau$		17.8
$\tau^- \to \mu^- \bar{\nu}_\mu \nu_\tau$		17.4
Hadronic decays		64.8
$\tau^- \to h^- \nu_\tau$		11.5
$\tau^- \to h^- \pi^0 \nu_\tau$	$\rho(770)$	25.9
$\tau^- \to h^- \pi^0 \pi^0 \nu_\tau$	$a_1(1260)$	9.5
$\tau^- \to h^- h^+ h^- \nu_\tau$	$a_1(1260)$	9.8
$\tau^- \to h^- h^+ h^- \pi^0 \nu_\tau$		4.8
Other		3.3

Table 5.1: Decays of the τ lepton, intermediate resonances, and branching fraction. For simplicity just the τ^- decays are shown. The values are also valid for the charge conjugate [179].

probability, is calculated as:

$$\tau_{h\ \text{Fake Prob.}}^{\text{RECO.+ID.}} = \frac{\text{Denominator \& } \tau_h\ p_T > 20\ \text{GeV \& } |\eta| < 2.3 \text{ \& Isolation}}{\text{Jet } p_T >\ 20\ \text{GeV \& Jet } |\eta| < 2.3}. \tag{5.15}$$

In this case, the denominator corresponds to the reconstructed visible jets. A matching of reconstructed jets and reconstructed taus is done. The reconstructed taus that matched some reconstructed jet are required to pass the identification requirements.

The HPS algorithm

The identification of hadronically decaying τ leptons (τ_h) is performed with the HPS algorithm in two steps [181]:

- **Reconstruction**: charged and neutral particles are combined to construct specific τ_h decays. The four-momentum of the τ_h candidates is computed.

- **Identification**: anti jet, electron, and muon discriminators are applied to separate true τ_h decays from jets, electrons, and muons faking hadronic decays of the tau lepton.

In the reconstruction step, particle flow jets clustered with the anti-k_T algorithm are used as seeds, and different τ_h decay channels are categorized through the reconstruction of their intermediate resonances. The so-called strips are formed by PF photons and electrons and allow to reconstruct the π^0 component of the τ_h via their decay products. The center of

the strips within the PF jets is set on the photon or electron (from photon conversions) with the highest energy. The HPS algorithm looks for additional photons or electrons in a window of $\Delta\eta \times \Delta\Phi$ around the strip center, adding the highest energetic electromagnetic particle found to the strip. For each iterative step, the strip four-momentum is recalculated, adding the contribution to the momentum of the integrated particle. The clustering process is repeated until no more particles can be associated to the strip within the defined $\Delta\eta \times \Delta\Phi$ window. Finally, the position of the strip, with a final size between 0.05 × 0.05 and 0.3 × 0.15, is recomputed. The p_T weighted average of all the PF e/γ constituents is summed up, so that:

$$\eta_{\text{strip}} = \frac{1}{p_T^{\text{strip}}} \sum p_T^{e/\gamma} \eta_{e/\gamma}, \qquad \Phi_{\text{strip}} = \frac{1}{p_T^{\text{strip}}} \sum p_T^{e/\gamma} \Phi_{e/\gamma}. \tag{5.16}$$

The clustering continues by forming new strips centered in the PF e and γ candidates with the highest p_T that are not yet associated with any strip. The strips with a p_T higher than 1 GeV/c and the charged hadrons are combined to reconstruct the individual τ_h decay modes [179]. The following decay topologies and the corresponding reconstructed decay modes are distinguished by the HPS algorithm:

- ***Single hadron:*** $h^-\nu_\tau$ and $h^-\pi^0\nu_\tau$ decays, low energy neutral pions are not reconstructed as strips.
- ***One hadron + one strip:*** $h^-\pi^0\nu_\tau$, presence of collimated photons from the π^0 decay.
- ***One hadron + two strips:*** $h^-\pi^0\nu_\tau$, presence of non-collimated photons from the π^0 decay.
- ***Three hadrons:*** $h^-h^+h^-$, the three charged hadrons are required to come from the same secondary vertex, created using a Kalman vertex fitter.

Fig. 5.16 shows a pictorial representation of the reconstruction process of the three main τ_h decay modes with the HPS algorithm.

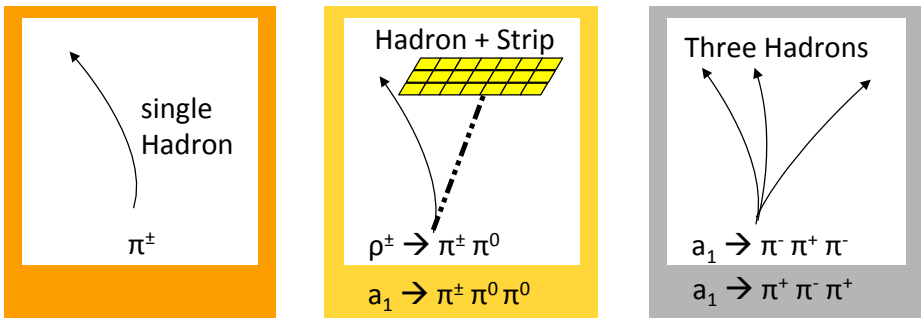


Figure 5.16: Three main decay modes of τ_h, reconstructed with the HPS algorithm.

The tau four-momentum is computed along with the isolation quantities of relevance for the identification step. For the isolation, each τ candidate is required to have no additional charged hadrons or photons within a ΔR cone of 0.5. The reconstructed mass of the visible hadronic component is required to be within a mass window around the masses of the inter-

mediate meson resonances listed in Tab. 5.1. The p_T thresholds of the particles considered in the isolation cone are adjusted to define different working points, with the loose working point corresponding to a probability for jets to be misidentified as τ_h of approximately 1%.

Discrimination against jets, electrons, and muons

Jets, electrons, and muons can pass the tau reconstruction algorithm. This fake probability is reduced by making tau leptons to meet specific requirements, using variable cuts and multivariate analysis discriminators, while maintaining a high identification efficiency.
Jets can contain exactly the same decay products as hadronically decaying taus, but they generally result in more constituents. Thus, the amount of jets faking taus can be reduced requiring additional isolation conditions. The isolation sum discriminant of τ_h candidates is defined as:

$$I_{\tau_h} = \sum p_T^{\text{charged}}(d_z \ < 0.2 \text{ cm}) + \max\left(0, \sum p_T^{\gamma} - \Delta\beta \sum p_T^{\text{charged}}(d_z \ > \ 0.2 \text{ cm})\right), \quad (5.17)$$

with the isolation cone centered on the direction of the τ_h candidate. $\sum p_T^{\text{charged}}$ and $\sum p_T^{\gamma}$ correspond to the sum of the scalar p_T of charged particles and photons respectively. The pileup contribution to the p_T sum of photons within the isolation cone is accounted for with the term $\Delta\beta \sum p_T^{\text{charged}}$. The scalar p_T of the charged hadrons located within a ΔR cone of 0.8, coming from a different vertex than the τ_h candidate vertex, is summed and weighted with the so-called $\Delta\beta$ factor. This factor accounts for the ratio between the charged hadrons and photons energy in inelastic collisions. The main working points of the isolation sum discriminant (loose, medium, and tight) are defined by requiring I_{τ_h} to be less than 2.5, 1.5, and 0.8 GeV, respectively.
The scalar p_T sum of the e and γ candidates located inside the strip, but outside the signal cone:

$$\sum p_T^{\text{strip, outer}} = \sum p_T^{e/\gamma}(\Delta R > R_{\text{sig}}), \quad (5.18)$$

is exploited to further reduce the jet $\rightarrow \tau_h$ misidentification probability. A 20% reduction is obtained by requiring $\sum p_T^{\text{strip,outer}}$ to be less than 10% of the $p_T^{\tau_h}$, while maintaining a similar efficiency. Apart from the isolation sum discriminants, MVA-based discriminants are also applied. The MVA classifier is based on BDTs trained to discriminate between τ_h decays and quark or gluon jets. Information such as the multiplicity of the photon and electron candidates, the isolation, and the τ lifetime serve as sensitive variables for the training of the BDT. The so-called perfect electrons (1 GSF track + ECAL deposit) might pass the tau-ID reconstruction algorithm and the isolation requirements mentioned above. Nevertheless, since they have a lower HCAL to ECAL ratio than charged hadrons, the electromagnetic energy fraction $E_{\text{ECAL}}/(E_{\text{ECAL}} + E_{\text{HCAL}})$ combined with other sensitive variables such as the occurrence of bremsstrahlung along the leading track, allow the separation of the electromagnetic showers from the hadronic showers, and the distinction of these electrons. Finally, only muons faking taus by passing the tau reconstruction algorithm are left, which are reduced by requiring no energy deposits in the muon system.

CHAPTER

6

STATISTICAL METHODS FOR DATA ANALYSIS IN HIGH ENERGY PHYSICS

Contents

This chapter is dedicated to present an overview of the basic set of statistical concepts needed for the analysis of the data and the interpretation of the results presented in this work. The chapter is mainly based on Refs. [182–189].

6.1 Introduction and fundamental concepts

The field of experimental high energy physics requires the analysis of large data samples. To successfully interpret the data, statistical methods are extensively applied at every step of the data analysis. First, the events (e.g., pp collisions events) are collected, and a set of characteristics as the particles' momentum, the number of muons, and jet energy are measured. The observed distribution of these characteristics can then be compared with the theoretical predictions. By comparing the observed and the theoretical distributions, the free parameters of the theory under scrutiny can be estimated. This comparison allows us to assess the level of agreement between the theory and the observed data. Nevertheless, a precise assessment of the agreement requires the determination of the uncertainty on the parameter estimates, quantified in terms of probability.

6.1.1 Bayes theorem and Total Law of probability

The theory of probability allows us to define the concept of probability in a rigorous mathematical manner. The definition of probability in terms of set theory was formulated in 1933 by the mathematician Kolmogorov. One first considers a set S called sample space, which consists of a group of elements, A is a subset of C, and the real number $P(A)$, called probability, is defined by three axioms. First, for every subset A in S, $P(A) \geq 0$. Second, for any two subsets A and B fulfilling the condition $A \cap B = \emptyset$ (called disjoint subsets), $P(A \cup B) = P(A) + P(B)$. The third axiom states that the probability assigned to the sample space S is $P(S) = 1$. The so-called *conditional probability* of A given B and of B given A, being A and B two subsets of the sample space S, is given by:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \tag{6.1a}$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)}. \tag{6.1b}$$

A and B are considered to be independent if:

$$P(A \cap B) = P(A)P(B). \tag{6.2}$$

If this condition is fulfilled, from the definition of conditional probability in Eqs. (6.1a) and (6.1b), follows that: $P(A|B) = P(A)$ and $P(B|A) = P(B)$. Since $A \cap B$ and $B \cap A$ are the same, by combining both equations, the so-called *Bayes' theorem* is obtained:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{6.3}$$

Thus, the Bayes' theorem relates the two conditional probabilities $P(A|B)$ and $P(B|A)$. If the sample space S is divided into disjoint subsets A_i and $P(A_i) \neq 0$ for all i, the *Law of total probability* can be defined as:

$$P(B) = \sum_i P(B|A_i)P(A_i), \tag{6.4}$$

where B is an arbitrary subset of S. From the combination of the Law of total probability and the Bayes' theorem, the conditional probability of A given B takes the form:

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}, \tag{6.5}$$

resulting especially convenient in cases where the probabilities for the subsets A_i are easy to calculate.

6.1.2 Interpretation of probability

The *relative frequency* and the *subjective or Bayesian probability* constitute two different interpretations of probability. They can be used, for instance, to assign a statistical error to a measurement or to quantify systematic uncertainties, respectively.
In the relative frequency interpretation, the elements of the sample space S are considered the possible outcomes of a measurement, i.e., subsets A, B, etc., correspond to the outcomes of a repeatable experiment. This interpretation allows us to experimentally test theories that assign to some process a certain probability value. The probability of the outcome to be A is calculated as the fraction of times that A occurs if the measurement were to be repeated an infinite number of times:

$$P(A) = \lim_{n\to\infty} \frac{\text{number of occurrences of outcome A in n measurements}}{\text{n}}. \tag{6.6}$$

Under this so-called frequentist approach of probability, each particle collision in a collider can be considered a repetition of an experiment. In the bayesian interpretation, the sample space corresponds to given hypotheses, statements that can be either true or false. The probability represents, in this case, a measure of the degree of belief:

$$P(A) = \text{degree of believe that hypothesis A is true.} \tag{6.7}$$

The subjective or Bayesian probability is, therefore, suitable to treat non-repeatable phenomena, and the value of probability assigned will reflect the state of knowledge of the system under study. Under the bayesian interpretation of probability, if a 95% probability for a particle to have a value of mass within a given mass interval is assigned, the 95% indicates the level of confidence that the mass value lies in the fixed interval. The concept of subjective probability constitutes the basis of *Bayesian statistics*. The subsets A and B in Eq. (6.3) can be interpreted as the hypothesis that a certain theory is true and that the outcome of an experiment would be a particular result, respectively. Under this interpretation, the Bayes' theorem acquires the following form:

$$P(\text{theory}|\text{data}) \propto P(\text{data}|\text{theory}) \cdot P(\text{theory}), \tag{6.8}$$

where $P(\text{theory})$ denotes the *prior probability* for the theory to be true, and $P(\text{data}|\text{theory})$ represents the probability of observing the data, under the assumption of the validity of the theory. After analyzing the results of the experiment, an updated probability $P(\text{theory}|\text{data})$ for the theory to be correct is calculated, known as *posterior probability*. The assignment of the prior probability does not concern Bayesian statistics, but once the value is assigned, the Bayesian statistics formulates how the value of the prior probability should evolve in the

light of the obtained experimental data.
After having introduced the concept of probability and the two main interpretations relevant for data analysis, the next subsection is dedicated to presenting the group of variables that describe the outcome of random processes and the functions that provide a probability prediction for this kind of variable.

6.1.3 Random variables and probability density functions

A variable which has a specific value for each element of the defined sample space S is called a *random variable*, and it can be discrete or continuous. Given the output of an experiment as a single continuous variable x and a sample space S composed by the possible values that x can assume, the probability to observe one of these values in an interval $[x, x + dx]$, called *probability density function (pdf)*:

$$\text{probability to observe x in an interval } [x, x + dx] = f(x)dx, \tag{6.9}$$

is defined. The probability density function is normalized so that the total probability of obtaining some outcome is one,

$$\int_S f(x)dx = 1. \tag{6.10}$$

For the case of a discrete outcome x_i of $\boldsymbol{x}$, with $i = 1, ..., n$; the probability to observe an specific x_i and the corresponding normalization condition are:

$$P(x_i) = p_i, \tag{6.11a}$$

$$\sum_i P(x_i) = 1. \tag{6.11b}$$

This set of n observations of x can be displayed in a histogram. The x axis of the histogram is divided into m subintervals or bins with a width Δx_i. The number of entries in a bin corresponds to the number of occurrences in the subinterval i, which is given in the y axis. The area under the histogram is obtained summing up the product of the number of entries per bin and the width of the bins:

$$\text{area} = \sum_{i=1}^{m} n_i \cdot \Delta x_i. \tag{6.12}$$

The histogram can then be normalized to the unity dividing each bin by the product of the total number of entries n_i in the bin and the bin width Δx_i.
Given the outcome of an experiment characterized by several values $(x_1, ..., x_n)$; a *joint probability density function* $f(x_1, ..., x_n)$ can be defined. The pdf of some (or one) of these components given the joint pdf is called *marginal probability density function*, and can be calculated as:

$$f_1(x_1) = \int ... \int f(x_1, ..., x_n)dx_2...dx_n. \tag{6.13}$$

Two of the x_1 and x_2 components are independent if the condition: $f(x_1,x_2)=f_1(x_1)f_2(x_2)$ is fulfilled. If the component x_2 is a constant, a conditional pdf for x_1 given x_2 can be defined

as:

$$g(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}. \tag{6.14}$$

The so-called *expectation value* of a variable constitutes another relevant parameter, defined for a random variable x with pdf $f(x)$ as:

$$E[x] = \int x f(x) dx, \quad E[x] = \mu. \tag{6.15}$$

The corresponding *variance* $V[x]$ of x takes the form:

$$V[x] = E[x^2] - \mu^2 = E[(x-\mu)^2], \quad V[x] = \sigma^2, \tag{6.16}$$

where the square root of $V[x]$ is the standard deviation of x. An estimate of the level of correlation between two random variables X and Y can be obtained through the covariance matrix:

$$\mathrm{cov}[x, y] = E[xy] - \mu_x \mu_y = E[(x-\mu_x)(y-\mu_y)], \tag{6.17}$$

which involves the expectation values of both variables and the expectation value of their product. Each dimensionless correlation coefficient can be computed as:

$$\rho_{xy} = \frac{\mathrm{cov}[x, y]}{\sigma_x \sigma_y}. \tag{6.18}$$

If x and y are independent, the expectation value of the product of the two random variables takes the form:

$$E[xy] = \int\int xy f(x, y) dx dy = \mu_x \mu_y. \tag{6.19}$$

Substituting Eq. (6.19) in Eq. (6.17) results in $\mathrm{cov}[x, y] = 0$. The variables x and y are then said to be uncorrelated. The estimation of the properties of a pdf, such as the mean and the variance, builds upon general concepts of parameter estimation, which are examined in the section that follows.

6.2 Parameter estimation

Two classes of statistical inference are relevant for the field of high energy physics: the estimation of parameters and their constraints and the testing of one or multiple hypotheses. The estimation of parameters from observed distributions called *fitting*, constitutes a fundamental task of the data analysis in high energy physics. It is used in every step of the measurement, from early stages as the track reconstruction (Subsec. 5.3.2) and detector energy calibrations (Subsec. 5.3.5), up to more advanced steps like the determination of relevant quantities for new particles, e.g., the mass, width, and signal strength. The two primary elements of parameter estimation are the estimation itself, i.e., the determination of the approximate true parameter values, and the estimation of the uncertainties on the estimated parameters. In the frequentist approach to parameter estimation, the two most commonly used methods are the *least squares* and the *maximum-likelihood*.
In this section, some fundamental concepts of parameter estimation are introduced. A subsection is dedicated to the description of the maximum-likelihood procedure, of particular

relevance for the analysis presented in this thesis, in which the signal is extracted through a binned maximum-likelihood fit applied to a BDT classification distribution.

6.2.1 Samples, estimators, and bias

The set of n independent observations of a random variable x, described by a pdf $f(x)$ is known as a *sample*. The n-dimensional vector $\boldsymbol{x} = (x_1, ..., x_n)$ where n is the sample size can be considered to be the output of a single random measurement, characterized by n quantities $x_1, ..., x_n$. Assuming that all x_i are independent and described by the same pdf $f(x)$, the joint pdf for the sample $f_{\text{sample}}(x_1, ..., x_n)$ becomes the product of n pdfs:

$$f_{\text{sample}}(x_1, ..., x_n) = f(x_1)f(x_2)...f(x_n). \tag{6.20}$$

A common situation is to have a set of n measurements of a random variable x, from which the properties of the corresponding pdf $f(x)$ need to be determined. The pdf $f(x, \theta)$ serves as hypothesis for the unknown $f(x)$, being $\theta = (\theta_1, ..., \theta_n)$ an n-dimensional vector of unknown parameters, which are typically constants that characterize the pdf shape. Thus, to obtain $f(x, \theta)$, the θ_i parameters need to be calculated. The value of the parameters θ_i can be obtained with a numerical function of the observed data x called an *estimator*. The estimator for the quantity θ is written as $\widehat{\boldsymbol{\theta}}$. The hat is used to differentiate the estimator from the true value θ, which would remain unknown, since $\widehat{\boldsymbol{\theta}}$ is, as the name indicates, just an estimate. The estimator constitutes itself a random variable. Thus, each time the experiment is repeated, a new set $x_1, ..., x_n$ of size n is obtained and the estimator $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ will have different values, distributed according to some pdf $g(\widehat{\boldsymbol{\theta}}, \theta)$, called *sampling distribution*. The expectation value of an estimator $\widehat{\boldsymbol{\theta}}$, which follows the sampling pdf distribution $g(\widehat{\boldsymbol{\theta}}, \theta)$, can be calculated as:

$$E[\widehat{\boldsymbol{\theta}}(\boldsymbol{x})] = \int \widehat{\boldsymbol{\theta}}_g(\widehat{\boldsymbol{\theta}}; \theta) d\widehat{\boldsymbol{\theta}} = \int ... \int \widehat{\boldsymbol{\theta}}(\boldsymbol{x}) f(x_1; \theta)...f(x_n; \theta) dx_1...dx_n, \tag{6.21}$$

where $f_{\text{sample}}(x_1, ..., x_n)$ was introduced as the joint pdf of the sample. The mean value of an estimator could be obtained after performing an infinite number of experiments, with a sample of size n for each experiment.

6.2.2 Properties of estimators

Two important indicators of the quality of an estimator are the *bias* and the *mean squared error*. The bias is obtained by subtracting the true parameter value θ from the expectation value of the estimator:

$$\boldsymbol{b} = E[\widehat{\boldsymbol{\theta}}] - \theta. \tag{6.22}$$

It constitutes an indicator of how close is the estimator from the true value. If b is zero for any sample size n, the parameter is said to be unbiased. The bias represents the systematic error on a measurement, while the variance represents the corresponding statistical error, indicating how much the estimator changes for different datasets. The sum of the variance and the bias squared is called *mean squared error* and takes the form:

$$\begin{aligned} \text{MSE} &= E[(\widehat{\boldsymbol{\theta}} - \theta)^2] = E[(\widehat{\boldsymbol{\theta}} - E[\widehat{\boldsymbol{\theta}}])^2] + (E[\widehat{\boldsymbol{\theta}} - \theta])^2, \\ &= V[\widehat{\boldsymbol{\theta}}] + \boldsymbol{b}^2. \end{aligned} \tag{6.23}$$

A good estimator would be one with small (or zero) bias, a small variance, and a small MSE. A small or zero bias is desired since the average of repeated measurements of a quantity should tend to the true value. Often, an estimator is said to be optimal if it has zero bias and a minimum variance, with certain trade-off between both characteristics. Fig. 6.1 shows the effects of bias (in red) and of a large variance (in green), with respect to the optimal estimator (in blue).

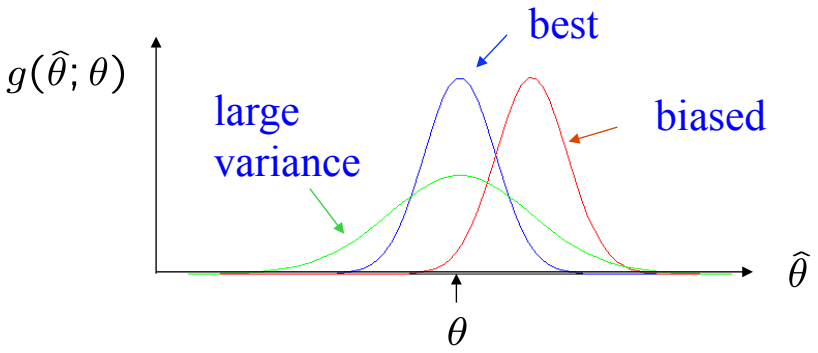


Figure 6.1: Distribution of a sampling pdf $g(\hat{\boldsymbol{\theta}},\theta)$ as a function of the estimator $\hat{\boldsymbol{\theta}}$. The estimator is: unbiased (in blue), biased (in red), and has a large variance (in green) [190].

The arithmetic mean of a sample of a random variable x with size n is called *sample mean* and constitutes an estimator of the expectation value μ of x:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{6.24}$$

For $n \to \infty$, $\overline{x}$ converges to μ. Therefore, $\overline{x}$ constitutes an unbiased estimator for the population mean μ. The *sample variance*, multiplied by a factor $\frac{1}{n-1}$:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n}{n-1}(\overline{x^2} - \overline{x}^2), \tag{6.25}$$

is an unbiased estimator for the population variance. Similarly, the quantity:

$$\widehat{V}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{n}{n-1}(\overline{xy} - \overline{x}\,\overline{y}), \tag{6.26}$$

constitutes an unbiased estimator for the covariance V_{xy} of two random variables x and y with unknown means.

6.2.3 The method of maximum-likelihood

The likelihood function

The maximum-likelihood is a method used to determine the parameter values for which a given observation would have the highest probability. Given the result of a set of measurements as a collection of numbers $\boldsymbol{x}$ and a joint pdf for the data $P(\boldsymbol{x}|\boldsymbol{\theta})$ that is a function of

a set of parameters $\boldsymbol{\theta}$, the so-called *likelihood function* is defined as:

$$\mathcal{L}(\boldsymbol{\theta}) = P(\boldsymbol{x}|\boldsymbol{\theta}). \tag{6.27}$$

Considering the n observations of $\boldsymbol{x}$ as independent and that $\boldsymbol{x}$ follows the distribution $f(\boldsymbol{x}, \boldsymbol{\theta})$, the joint pdf for the data sample can be calculated as the product of all the independent pdfs of x_i. Therefore, the likelihood function takes the form:

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) = f(x_1, ..., x_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}). \tag{6.28}$$

Maximum-likelihood estimators

The parameter values for which the likelihood reaches the maximum value are called *maximum-likelihood (ML) estimator(s)*. A low probability for the measurements is obtained if they deviate too much from the true values. If the pdf hypothesis and the parameter values are correct, a high probability for the data that was measured is expected. For differentiable likelihood functions, each independent estimator can be obtained by solving the equations:

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = 0, \quad i = 1, ..., m. \tag{6.29}$$

The estimation of parameters through a maximization of the likelihood function corresponds to the frequentist approach. Often, instead of maximizing the likelihood, the negative of the logarithm of the likelihood, known as *log-likelihood*, is minimized:

$$-\ln \mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(x_i; \boldsymbol{\theta}). \tag{6.30}$$

In the bayesian approach a *posterior probability density* for $\boldsymbol{\theta}$ given the data $\boldsymbol{x}$ is defined:

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int \mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')d\boldsymbol{\theta}'}. \tag{6.31}$$

The magnitude $\pi(\boldsymbol{\theta})$ denotes the prior probability density for $\boldsymbol{\theta}$ and constitutes a measure of the prior state of knowledge of $\boldsymbol{\theta}$, before updating this knowledge in light of the new data.

Maximum-likelihood fit with binned data

The computation of the log-likelihood function becomes difficult for large datasets, since the probability density function $f(x_i, \boldsymbol{\theta})$ must be calculated for each value of x_i. To overcome this difficulty the data can be binned. The expectation values $\boldsymbol{\nu} = (\nu_1, ..., \nu_N)$ of the numbers of entries per bin $\boldsymbol{n} = (n_1, ..., n_N)$ in N bins, can then be calculated as:

$$\nu_i(\boldsymbol{\theta}) = n_{\text{tot}} \int_{x_i^{\min}}^{x_i^{\max}} f(x, \boldsymbol{\theta})dx, \tag{6.32}$$

where $x_i^{\min}$ and $x_i^{\max}$ are the bin limits, and n_{tot} is the total number of entries. Considering the obtained histogram a single measurement of an N-dimensional random vector, the joint

pdf takes the form of a multinomial distribution:

$$f_{\text{joint}}(\boldsymbol{n};\boldsymbol{\nu}) = \frac{n_{\text{tot}}!}{n_1!...n_N!}(\frac{\nu_1}{n_{\text{tot}}})^{n_1}...(\frac{\nu_N}{n_{\text{tot}}})^{n_N}, \tag{6.33}$$

where each term ν_i/n_{tot} represents the probability for bin i. Assuming that the numbers of entries per bin are independent and follow the Poisson distribution, the data is instead described by the product of Poisson probabilities:

$$f_{\text{joint}}(\boldsymbol{n};\boldsymbol{\nu}) = \prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}. \tag{6.34}$$

Taking the logarithm and dropping additive terms that do not depend on the parameters, the log-likelihood function is obtained. A ML fit is performed to find the estimators $\widehat{\boldsymbol{\theta}}$. For a very small bin size, which corresponds to a large value of N, the likelihood function results the same as the unbinned case.

6.3 Hypothesis testing

The term hypothesis testing refers to the process used to decide on the acceptance or rejection of a hypothesis after the analysis of a set of measurements. Often, the process aims to determine if a dataset is consistent with the given hypothesis so that the hypothesis can be validated or disproved. One example would be to test the assumption that the observed data inspected in the search for indications of new physics corresponds only to Standard Model processes. In this section, the frequentist approach to hypothesis testing is presented, emphasizing the applications in the field of high energy physics. In Subsec. 6.3.2, the goodness-of-fit tests are introduced, a branch of hypothesis testing used to quantify how well a set of measurements agrees with a given hypothesis when there is no additional hypothesis. Subsecs. 6.3.3 and 6.3.4 are dedicated to discussing the testing of the background-only and signal hypotheses. Finally, the last subsection concerns the multivariate analysis methods used for event classification.

6.3.1 Frequentist statistical test

A test of a null hypothesis can be constructed considering both the null hypothesis H_0 and an alternative hypothesis H_1. Given a critical region W of the data space, the probability to observe the data in W, assuming the correctness of the null hypothesis can be calculated as:

$$P(x \in W|H_0) \leq \alpha, \tag{6.35}$$

where the region W is designed such that the *size* of the test α is small. If x is observed in the critical region, the null hypothesis is rejected. The critical region should be selected so that the probability to find x is low if H_0 is true, but at the same time high if H_1 is true.
Two kinds of errors can be made when deciding if the hypothesis H_0 is accepted or rejected. H_0 can be rejected when it is true, known as Type I error, with the probability for this to happen being the size of the test. Also, H_0 can be accepted when it is false, known as Type

II error, with an associated probability:

$$P(x \in S - W|H_1) = \beta. \tag{6.36}$$

Subtracting β from 1, the so-called *power* of the test with respect to the alternative H_1 is obtained:

$$\text{Power} = 1 - \beta. \tag{6.37}$$

The size and power of a test statistic are illustrated in Fig. 6.2.

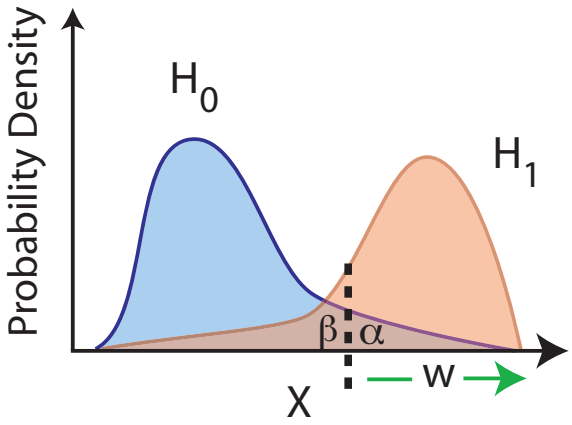


Figure 6.2: Sketch of the distribution of a test statistic under the null and alternative hypotheses. The size α and power β of the test are depicted [191].

Physics context of a statistical test

The frequentist statistical tests described above can be used in physics analysis in the event selection step to separate different types of particles (e.g., electron vs. muon) or to separate known events (e.g., $t\bar{t}$ vs. QCD multijet). For the separation of the different kinds of events, one assumes as null hypothesis H_0 that the event is a background event, and as alternative hypothesis H_1 that the event is a signal event. In the search for new physics, under the null or *background only* hypothesis all events correspond to Standard Model processes, while under the alternative or *signal-plus-background* hypothesis part of the events correspond to a type whose existence is not yet established. Thus, the background only hypothesis consists of the event type that one would like to reject, while the signal in the signal-plus-background hypothesis corresponds to the type one would like to retain after the selection. Typically, the available data sample contains a mix of the two kinds of events and, as result of the measurement, a collection of numbers $\boldsymbol{x}= (x_1, ..., x_n)$ per event is obtained:

x_1= number of muons

x_2= mean p_T of jets

x_3= missing energy...

where each event can be associated with a point in an $\boldsymbol{x}$ space. The n-dimensional joint pdf followed by $\boldsymbol{x}$ depends on the type of event produced. A so-called *decision boundary* whether to accept or reject the events that belong to certain event type can be drawn. One option would be to draw the decision boundary in the form of cuts $(x_1 < c_1, ..., x_n < c_n)$, as illustrated in Fig. 6.3a. Linear and non-linear decision boundaries can also be applied, as shown in Figs. 6.3b and 6.3c.

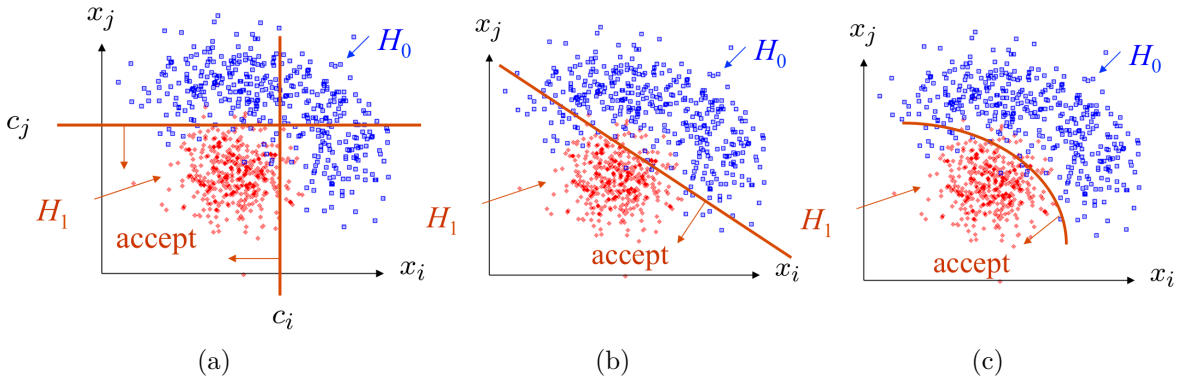


Figure 6.3: Visualization of decision boundaries. a) cut based decision boundary b) linear decision boundary c) non-linear decision boundary [190].

The optimal decision boundary is constructed through a statistical test in which the boundary of the critical region for the n-dimensional data space is defined as:

$$t(x_1, ..., x_n) = t_{\text{cut}}. \tag{6.38}$$

The decision boundary is translated into a single cut on t. Therefore, the n-dimensional problem has converted into a 1-d problem, with conditional pdfs $g(t|H_0)$ and $g(t|H_1)$. A representation of t_{cut}, $g(t|H_0)$, and $g(t|H_1)$ is shown in Fig. 6.4.

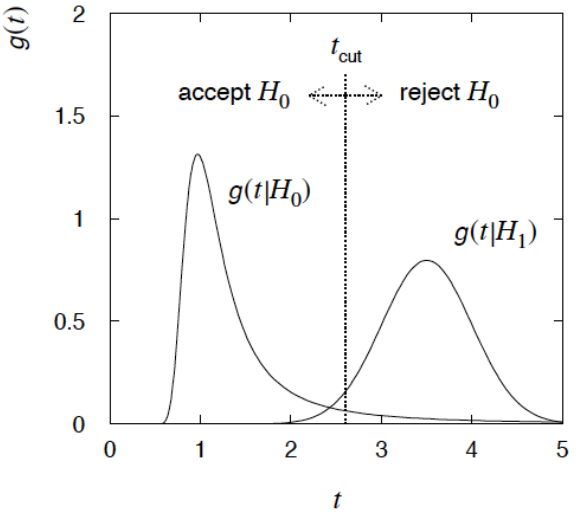


Figure 6.4: Illustration of a single cut decision boundary and the conditional pdfs $g(t|H_0)$ and $g(t|H_1)$ [190].

Neyman-Pearson lemma

The critical region for a test statistic can be selected so that the highest power at a size α is obtained. The Neyman–Pearson lemma, introduced by Jerzy Neyman and Egon Pearson in 1933, states that in order to get the highest power for a given size α in a test of H_0 (background) versus H_1 (signal), the critical region should fulfill the condition:

$$\frac{f(x|H_1)}{f(x|H_0)} > c, \tag{6.39}$$

inside the region, and have a smaller value than c outside. The constant c is chosen so that the test has a desired size.
The fraction in Eq. (6.39) is called the likelihood ratio for hypotheses H_0 and H_1. The corresponding optimal test statistic is:

$$t(x) = \frac{f(x|H_1)}{f(x|H_0)}. \tag{6.40}$$

The probability to reject H_0 if true, which constitutes the false discovery rate, can be calculated as:

$$\alpha = \int_W f(x|H_0)dx, \tag{6.41}$$

while the probability to accept H_0 if H_1 is true takes the form:

$$\beta = \int_{\overline{W}} f(x|H_1)dx. \tag{6.42}$$

Purity and misclassification rate

The purity of the event selection and the misclassification rate correspond to the probability for a signal event to be classified correctly as signal and the probability for the signal event to be wrongly classified as background, respectively. The background efficiency is the size of the test α in Eq. (6.41), while the signal efficiency is defined as follows:

$$\epsilon_s = \int_W f(x|H_1)dx = 1 - \beta. \tag{6.43}$$

Since one event can only be classified as signal or background, the sum of the signal purity and misclassification rate is one. Usually, the conditional pdfs $f(x|s)$ and $f(x|b)$ are not known and the likelihood ratio from Eq. (6.40) can not be evaluated. In this cases, simulated data is produced using Monte Carlo models for the signal and background processes, so that:

$$\text{generated} \quad \boldsymbol{x} \approx f(x|s) \quad \rightarrow \quad x_1, ..., x_n$$

$$\text{generated} \quad \boldsymbol{x} \approx f(x|b) \quad \rightarrow \quad x_1, ..., x_n$$

The production of MC events provides training data with events of known type, but comes with the inconvenience that the full simulation of the MC events at a large scale requires the availability of a considerable amount of computing resources.

6.3.2 Goodness-of-fit tests

The tests that estimate how well the hypothesis of a functional form describes certain data distribution are known as *goodness-of-fit tests* (GoF). The selected functional form constitutes the null hypothesis H_0 of the test statistic whose value is sensitive to the level of agreement between the observed measurements and the predictions of H_0. In a goodness-of-fit test, no alternative hypothesis is given. The probability of obtaining a value for the test statistic equal to or higher than the one obtained is known as *p-value* of the test and characterizes the level of agreement between the data and the null hypothesis.

6.3.3 Testing the background-only hypothesis: discovery

Poisson counting experiment

The goodness-of-fit tests introduced in the last subsection are used to assess whether an excess observed in the data with respect to the background expectation is enough to claim a discovery. Fig. 6.5 depicts the number of signal (n_s), background (n_b), and total events ($n = n_s + n_b$) of a Poisson counting experiment, which follow a Poisson distribution with mean values of ν_s, ν_b, and $\nu = \nu_s + \nu_b$, respectively.

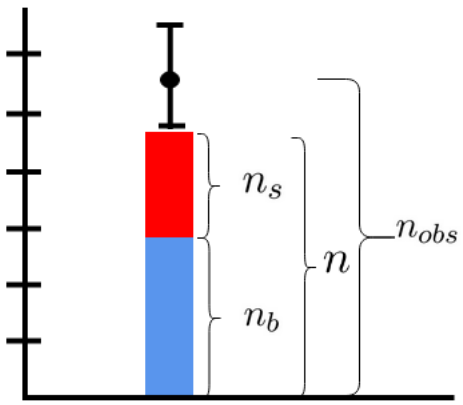


Figure 6.5: Sketch representing the number of signal, background, and observed events in a Poisson counting experiment.

Thus, the probability to observe n events can be calculated as:

$$f(n;\nu_s,\nu_b) = \frac{(\nu_s+\nu_b)^n}{n!}e^{-(\nu_s+\nu_b)}. \tag{6.44}$$

Given the number of observed events (n_{obs}) as the outcome of the experiment, the probability that a Poisson variable with mean ν_b will fluctuate so that the number of events would be equal or higher than n_{obs} under the assumption than $\nu_s = 0$, is:

$$P(n \geq n_{\text{obs}}) = \sum_{n=n_{\text{obs}}}^{\infty} f(n;\nu_s=0,\nu_b) = 1 - \sum_{n=0}^{n_{\text{obs}}-1} f(n;\nu_s=0,\nu_b) = 1 - \sum_{n=0}^{n_{\text{obs}}-1} \frac{\nu_b^n}{n!}e^{-\nu_b}. \tag{6.45}$$

Thus, the probability $P(n \geq n_{\text{obs}})$ constitutes the p-value for a given n and n_{obs} and quantifies how often, if the experiment is repeated many times, data as far away (or more) from the

null hypothesis as the observed data would be obtained, assuming the null hypothesis is true. Relatively small changes on the value of ν_b can result in an increase of even an order of magnitude of the p-value. Therefore, a precise estimation of the systematic uncertainty in the determination of the number of background events results of particular relevance for the assessment of the significance of an excess in the number of events observed in data. In the next subsection, the definition of *significance* and the procedure followed for its calculation are presented.

The significance of an observed signal

The determination of the significance of an observed signal constitutes one of the last steps in the statistical treatment of the data. The signal and background yields are written as $\mu \cdot s(\theta)$ and $b(\theta)$. The μ factor in the signal yield multiplies the expected SM cross-section such that:

$$\sigma = \mu \cdot \sigma_{SM}. \tag{6.46}$$

The parameters of the model that are not the parameters of interest (the expected signal and background yields) are called *nuisance parameters*. To each systematic uncertainty, a nuisance parameter is assigned, and the parameters of interest are functions of these nuisance parameters. The pdf $p_i(\widetilde{\theta}_i|\theta_i)$ denotes the probability to measure a value $\widetilde{\theta}_i$ of the ith nuisance parameter given its true value θ_i. The likelihood $\mathcal{L}$, given the data and the measurements of the nuisance parameters $\widetilde{\theta}$, takes the form:

$$\mathcal{L}(\text{data}|\mu \cdot s(\theta) + b(\theta)) = \mathcal{P}(\text{data}|\mu \cdot s(\theta) + b(\theta)) \cdot p(\widetilde{\theta}|\theta), \tag{6.47}$$

where $\mathcal{P}(\text{data}|\mu \cdot s(\theta) + b(\theta))$ is a product of probabilities over all bins for the case of a binned discriminant and over all the events for the unbinned case, and $p(\widetilde{\theta}|\theta)$ is the probability density function corresponding to all the measurements of the nuisance parameters. A test statistic is constructed to probe the hypothesis of the production of the new particle, with the result of the test comprising in a single number the information related to the observed data, expected signal, expected background, and the corresponding uncertainties. In order to infer the presence or absence of a signal in the data, the observed value of the test statistic is compared with the distribution of the test expected under the background-only and signal-plus-background hypotheses. The expected distributions for both hypotheses are obtained generating pseudo-datasets from the corresponding probability density functions. The value of the nuisance parameters used for the generation of the pseudo-datasets come from the maximization of the likelihood under the background-only and signal-plus-background hypotheses, respectively. The following test statistic is defined to quantify the statistical significance of an excess over the background-only expectation:

$$q_0 = -2\ln\frac{\mathcal{L}(\text{data}|b(\widehat{\theta}_0))}{\mathcal{L}(\text{data}|\widehat{\mu} \cdot s(\widehat{\theta}) + b(\widehat{\theta}))}, \quad \widehat{\mu} \geq 0, \tag{6.48}$$

where the values of the nuisance parameters in $b(\widehat{\theta}_0)$ come from the maximization of the likelihood in the numerator under the background-only hypothesis ($\mu = 0$), indicated with the subscript in $\widehat{\theta}_0$. Similarly, $\widehat{\theta}$ and $\widehat{\mu}$ are the values of the parameters θ and μ obtained from the maximization of the likelihood in the denominator under the signal-plus-background

hypothesis. For positive values of μ, which would denote a signal excess, the likelihood ratio q_0 is positive and if there is no excess ($\mu = 0$) the ratio is exactly equal to one. The significance of an observed excess is quantified with the p-value, which in the context of the test statistic q_0, denotes the probability to obtain a value of q_0 at least as large as the one observed in data q_0^{obs}, under the background only hypothesis:

$$p_0 = P(q_0 \geq q_0^{\text{obs}}|b). \tag{6.49}$$

The *significance* Z is defined as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the p-value p_0. Thus, the significance Z and the p-value are related through the one-sided tail integral of the Gaussian function:

$$p_0 = \int_Z^\infty \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)dx = 1 - \Phi(Z), \tag{6.50a}$$

$$\Phi(Z)^{-1} = (1 - p). \tag{6.50b}$$

A representation of the p-value and the statistical significance are depicted in Fig. 6.6. A value of the significance of 5 is called a 5 sigma effect and corresponds to a p-value of $2.9 \cdot 10^{-7}$. Fig. 6.7 shows the p-value as a function of the Z-value for a unit Gaussian probability density function. A discovery can be claimed if $Z > 5$.

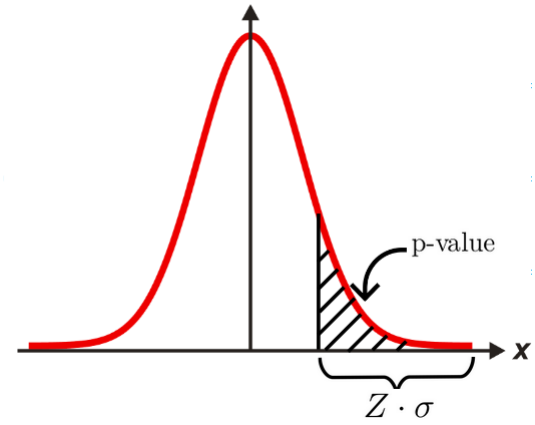


Figure 6.6: Illustration of the p-value and the statistical significance.

In the search for a signal peak with an unknown location in an invariant-mass spectrum, a local and global p-value can be defined. The local p-value represents the probability of finding a fluctuation in some particular location of the invariant-mass spectrum, while the global p-value is the probability of finding the fluctuation anywhere in the spectrum. For a mass spectrum spanned from m_1 to m_2 there is a probability boost to find a significant effect somewhere within the mass interval. This results in different local and global p-values and a boost factor $(m_2 - m_1)/W$, where W is the width of the signal. The effect is known as the *look-elsewhere-effect (LEE)* [192], and can be evaluated by assessing how the p-value distribution behaves as the signal mass hypothesis changes. The stronger the up and down variations of the p-value, the stronger is the Look-Elsewhere-Effect and, therefore, the corrections needed.

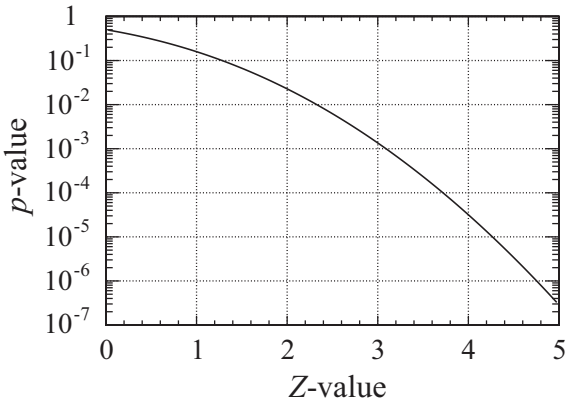


Figure 6.7: One-sided tail probability (p-value) vs. Z-value for a unit Gaussian probability density function [184].

6.3.4 Testing the signal hypothesis: setting limits

To quantify the absence of a signal a test statistic q_μ is defined:

$$q_\mu = -2\ln\frac{\mathcal{L}(\text{data}|\mu\cdot s(\widehat{\theta}_\mu)+b(\widehat{\theta}_\mu))}{\mathcal{L}(\text{data}|\widehat{\mu}\cdot s(\widehat{\theta})+b(\widehat{\theta}))}, \quad 0\leq\widehat{\mu}\leq\mu. \tag{6.51}$$

The subscript in μ indicates the dependence on the hypothesized signal rate μ, while the subscript in $\widehat{\theta}_\mu$ indicates that the maximization of the likelihood in the numerator is done under the hypothesis of a signal strength μ. The likelihood ratio q_μ is similar to Eq. (6.48), but uses the signal-plus-background instead of the background-only hypothesis in the numerator.

Exclusion limits

The modified frequentist construction CLs is used for the calculation of the exclusion limits in the case of the absence of a signal [193–195]. Two tail probabilities are defined:

$$CL_{s+b} = P(q_\mu \geq q_\mu^{\text{obs}}|\mu\cdot s+b), \tag{6.52a}$$

$$CL_b = P(q_\mu \geq q_\mu^{\text{obs}}|b), \tag{6.52b}$$

which represent the probability to obtain a value for the test statistic q_μ larger than the observed value q_μ^{obs}, for the signal-plus-background and background-only hypotheses, respectively. From the ratio of CL_{s+b} and CL_b, CL_s is obtained:

$$CL_s = \frac{CL_{s+b}}{CL_b}. \tag{6.53}$$

If $CL_s \leq \alpha$ for $\mu = 1$, the mass point is excluded at $1-\alpha$ confidence level. The procedure to quote the upper limit on μ at 95% confidence level is to adjust μ until $CL_s = 0.05$ is reached.

Upper limits results

The CLs procedure described in the previous subsection is usually carried out for each mass point within a scanned mass range, resulting in an upper limit μ_{up} per mass point. The distribution of μ_{up} assuming $\mu = 0$ is obtained for each mass point after generating an ensemble of pseudo-experiments, known as toy Monte Carlo. Each of the toys provides a value of μ_{up}. After generating certain amount of toys, a distribution of μ_{up} can be obtained. The median of μ_{up} as well as the 68% ($\pm 1\sigma$) and 95% ($\pm 2\sigma$) bands of these distributions can then be determined. Fig. 6.8 shows the median of μ_{up} and the green (yellow) bands that delimit the $\pm 1\sigma$ (2σ) regions for an individual mass point.

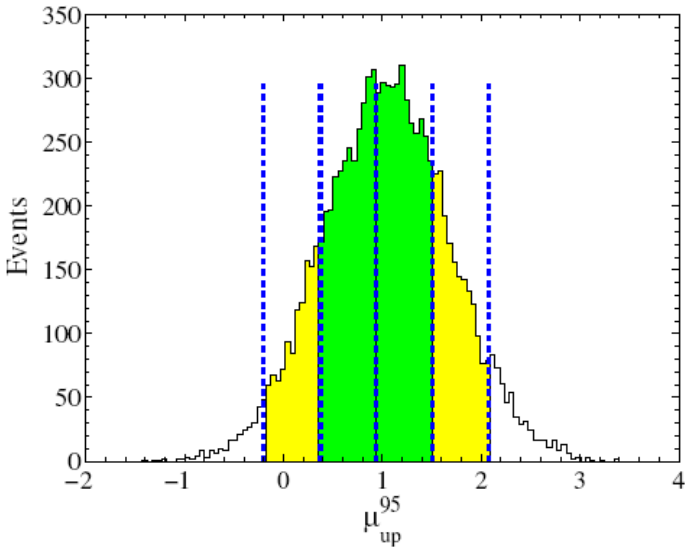


Figure 6.8: Median of μ_{up}, 68% and 95% CL bands, for mass point $m_H = 95$ GeV. The $\pm 1\sigma$ (2σ) regions are depicted in green (yellow) [190].

After performing the CLs procedure and determining the distribution of μ_{up} per mass point, an exclusion limit plot as the one presented in Fig. 6.9 from CMS Higgs discovery paper, back in 2012, can be obtained. The CLs values for the SM Higgs boson hypothesis as a function of the Higgs boson mass in the range from 110 to 145 GeV are presented. The dashed line corresponds to the median of μ_{up} under the background-only hypothesis. The corresponding 68% and 95% CL bands are also included.

6.4 Multivariate methods

The methods that examine multiple variables simultaneously are known as multivariate methods. The main goal of the analysis using multivariate techniques is to find an optimal decision boundary, which is obtained by minimizing a *loss function* in a process known as training of the classifier. The loss function evaluates the quality of the classification by returning large values when it deviates from the real event classification. The automated determination of a decision boundary according to a chosen algorithm is known as *machine learning*. The classifiers learn how the parameters of the decision boundaries should be chosen so that an optimal separation between signal and background events is obtained.

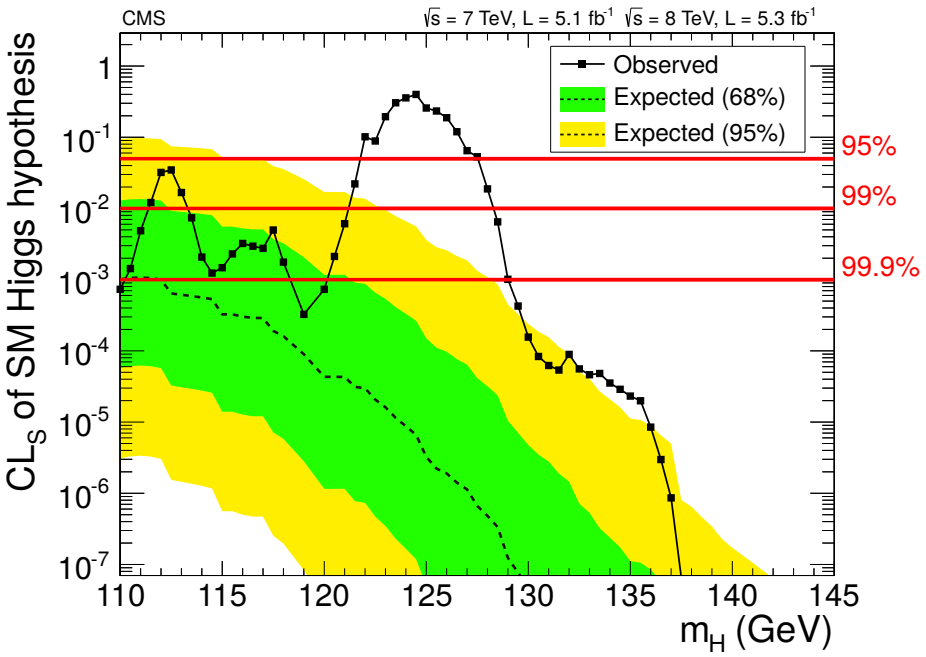


Figure 6.9: The CLs values for the SM Higgs boson hypothesis as a function of the Higgs boson mass in the mass range between 110 and 145 GeV. The expected upper limits are represented by their median (dashed line) and by the 68% (green) and 95% (yelow) CL bands [9].

In multivariate analysis, some general considerations need to be taken into account:

- The choice of the input variables.
- The functional form of the decision boundary or type of classifier (Fig. 6.3).
- The tradeoffs between:
  - sensitivity and complexity.
  - statistical and systematic uncertainty.

In the field of high energy physics, one is often searching for a small signal in a large dataset. Therefore, it is crucial to optimally exploit the available information on the data, for the reduction of the large background processes that contaminate the signal of interest. Multivariate classification methods based on machine learning techniques are extensively used for this purpose. Relevant examples of CMS analysis using machine learning techniques for object identification and event classification are the observation of the Higgs decaying into b-quarks [196], the associated Higgs production with top quarks [197], and the measurement of the Higgs boson decay to a pair of muons [198].

At present, the most popular machine learning algorithms used for physics analysis within the particle physics community involve the so-called supervised learning. For this kind of learning process, the type of events contained in the data samples needs to be known. Since nature does not provide a classifying event label for real data events, simulated data with known true class label is used for the training of the two possible event types (signal and background). Once the chosen algorithm has been trained to separate the signal from the background, the obtained decision boundary can be used to classify the real data events with an unknown event class. The boosted decision trees are one of the well-known learning

methods and constitute a fundamental component of the analysis presented in this thesis. Therefore, the next subsection is dedicated to introducing the learning method of decision trees.

6.4.1 Decision trees

A decision tree is a binary tree-structured classifier in which repeated cuts are made on a single variable at a time until a stop criterion is reached. The tree starts from a root node and goes down until a final node is reached. The final nodes are called *leaves*, and the set of nodes and splits leading to a leaf is called a *branch*. A diagram of a decision tree is shown in Fig. 6.10.

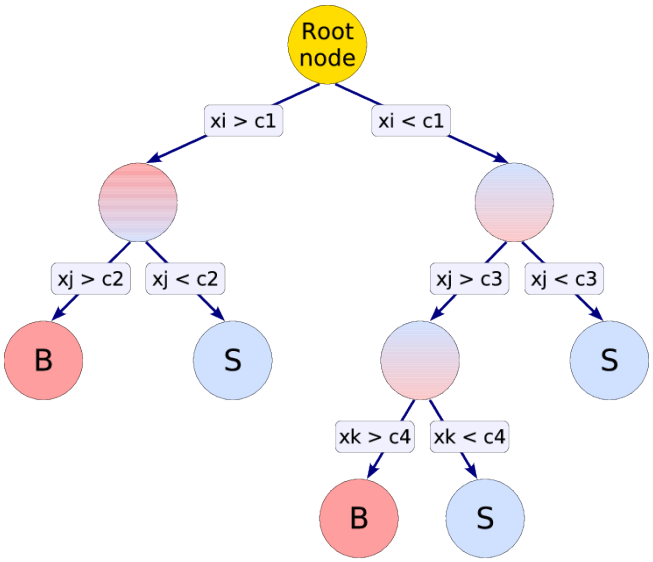


Figure 6.10: Schematic view of a decision tree [199].

The decision on the variable used at each split is based on the best-achieved separation between signal and background for the particular split. Thus, a same variable might provide the best separation in several nodes and be used multiple times, while other variables might end up not being used at all. The criteria to stop an iteration are based on signal purity and minimum number of events in a node. Once the iteration procedure stops, the resulting leaves are classified as signal or background according to a majority vote, e.g., the signal fraction greater than a specified threshold. Every point of the input variable space gets classified as signal or background. A so-called *decision tree classifier* is in place. Once the best variables and the corresponding cuts for each split are decided, i.e., the tree is trained, a set of events can be given as input to the decision tree for classification. Each event undergoes several yes or no decisions, starting at the root node until it ends up in certain leaf node, where is classified according to the class label of the leaf.

6.4.2 Separation measure and stability

The signal to background separation can be measured in terms of the *Gini index* or the *cross entropy*, defined as:

$$\text{Gini} = p\,(1-p), \tag{6.54a}$$

$$\text{Cross entropy} = -p\,\ln p - (1-p)\,\ln(1-p), \tag{6.54b}$$

where p denotes the purity. If the first training sample T consists of N events with:

- ($\boldsymbol{x_1}$,..., $\boldsymbol{x_N}$) event data vector.
- (y_1,..., y_N) true class labels, +1 for signal, −1 for background.
- (w_1,..., w_N) event weights.

and the ith event has a weight w_i, the purity at a given node can be calculated as:

$$p = \frac{\sum_{\text{signal}} w_i}{\sum_{\text{signal}} w_i + \sum_{\text{background}} w_i}. \tag{6.55}$$

The maximum values of the Gini and cross entropy indices are reached when the input sample is completely mixed and decrease monotonically for samples with a large signal or background composition. The separation indices before and after each split are compared to determine the variable with the highest separation for the splits. The misclassification error for a given value of purity p can be calculated as:

$$\text{misclassification error} = 1 - \max(p, 1-p). \tag{6.56}$$

Since the decision trees can be very sensitive to the statistical fluctuations in the training sample, methods such as boosting are used to overcome the stability problem, converting boosted decision trees in a robust and powerful classifier.

6.4.3 Boosting

The so-called boosting is a general method for enhancing the performance of a set of weak classifiers by combining them into a unique classifier. The obtained final classifier is more stable in time and has a smaller error than any of the individual ones. The combined decision trees form a so-called *forest*. After a first classifier is trained on a dataset and a set of misclassified events is obtained, further training iterations are performed. At each iteration, a new classifier is trained on a modified data sample, in which the misclassified events from the previous iteration are given a larger weight. An ensemble of training samples $(T_0, ..., T_M)$ corresponding to M iterations is then created, along with the classifiers $[b(\boldsymbol{x}, \boldsymbol{\alpha_0}), ..., b(\boldsymbol{x}, \boldsymbol{\alpha_M})]$ and the weights $(\alpha_0, ..., \alpha_M)$, calculated in the final average.

The final classifier is determined as a weighted sum of the so-called *base classifiers*:

$$y_{\text{Boost}}(\boldsymbol{x}; \alpha_0, ..., \alpha_M, \boldsymbol{\alpha_0}, ..., \boldsymbol{\alpha_M}) = \sum_{m=0}^{M} \alpha_m \cdot b(\boldsymbol{x}; \boldsymbol{\alpha_m}), \tag{6.57}$$

where the $\boldsymbol{\alpha_m}$ parameters are the parameters of the classifier m determined during the training and specify the various node splits for the decision tree m. The way in which the event weights are updated and the base classifiers are weighted in the final classifier varies according to the boosting algorithm used. In the *adaptative boost algorithm (AdaBoost)* the weights of previously misclassified events are multiplied by a common *boost weight* $\exp(\alpha_m)$, with the index m refering to the weights applied in the training steps before the training of the classifier m. The factor α_m:

$$\alpha_m = \ln(\frac{1 - \text{err}_{m-1}}{\text{err}_{m-1}}), \tag{6.58}$$

depends on the fraction of misclassified training events (err_{m-1}) of the previous classifier. All the event weights of the sample are then renormalized to keep constant the sum of weights:

$$\sum_{i=1}^{N} w_i = 1, \tag{6.59}$$

and the result of an individual classifier is defined as $b(\boldsymbol{x}) = +1$ or -1 for signal and background, respectively. The boosted event classification y_{Boost} is then calculated as:

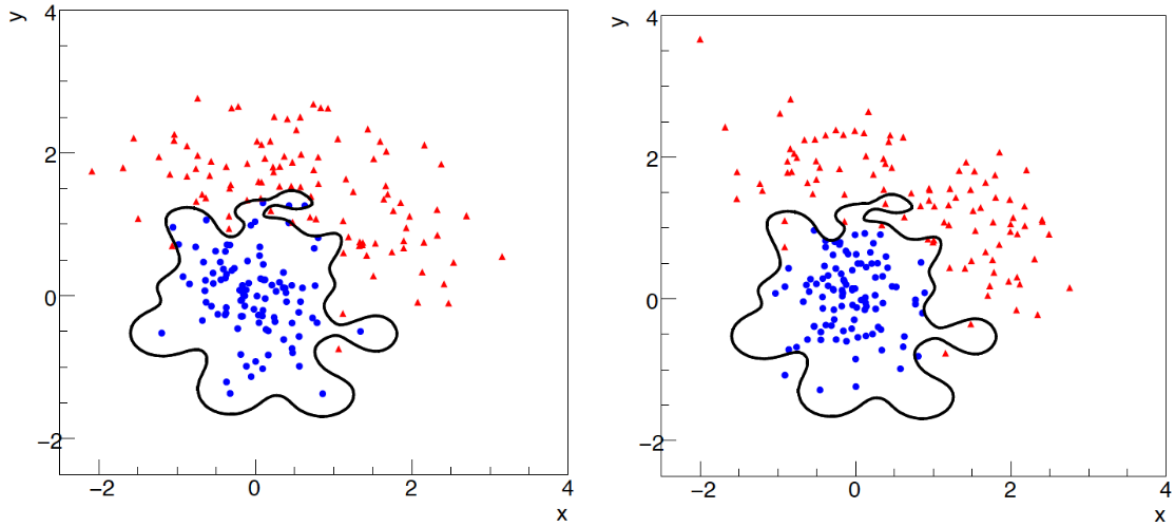
$$y_{\text{Boost}}(\boldsymbol{x}) = \frac{1}{M} \cdot \sum_{m}^{M} \alpha_m \cdot b_m(\boldsymbol{x}), \tag{6.60}$$

where the sum runs over all the individual M classifiers. A y_{Boost} value that tends to $+1(-1)$ indicates a signal(background)-like event.

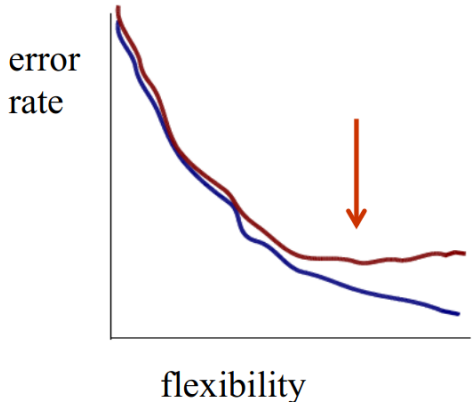
6.4.4 Overtraining

The decision boundary of a classifier becomes increasingly flexible when more parameters are included and might adapt too much to the training points. In this case, the decision boundary would show a deteriorated performance in an independent test data sample, as illustrated in Fig. 6.11a. This phenomenon is called *overtraining*. A set of solutions are available to solve it, such as removing insignificant nodes in the trees, known as tree pruning. The overtraining can be monitored by studying the dependence of the fraction of misclassified events with respect to the model flexibility, as shown in Fig. 6.11b. The optimal degree of model flexibility would be at the minimum error rate for the test sample. Usually, a physics analysis has more than one event category, and different optimal decision boundaries can be determined for each category. With enough training data, the classifier is able to learn the differences between categories. Nevertheless, individual classifiers are usually trained for each category. Even though the MVA training itself does not introduce a systematic uncertainty, the MVA output distribution can be influenced by the systematic uncertainties of the input variables. To estimate the influence of these systematic uncertainties, a new training for each possible variation is not needed, since one can estimate the change in the output distribution

for the already trained classifier by applying the classifier to test samples in which all possible systematic variations are considered.



(a)



(b)

Figure 6.11: a) Performance of an overtrained classifier in the training sample (left) and in an independent test sample (right). b) Overtraining monitoring: error rate as a function of the model flexibility for the training sample (blue) and the test sample (red). The red arrow indicates the optimum point of flexibility at minimum error rate for the test sample. The increase in error rate for the test sample indicates overtraining [190].

The statistical methods introduced in this chapter constitute the base of the tools used for the statistical analysis of the data in this work. The results of applying the techniques here discussed in the search for light bosons in the final state with two muons and two tau leptons are presented in Chapter 8.

CHAPTER

7

$H \to a_1 a_1 (Z_D Z_D) \to \mu\mu\tau\tau$ SEARCH WITH CMS RUN II DATA

Contents

The search for a pair of light bosons produced in decays of the 125 GeV Higgs boson in the final state with two muons and two tau leptons constitutes the core of this dissertation. In this chapter, the analysis steps undergone to successfully retrieve, reconstruct, and understand the data are described. The signal topology, the used datasets and simulated samples, and the corrections to simulation are examined. The MVA approach introduced to enhance the analysis sensitivity is discussed, addressing the use of the classification distribution of a BDT as the final discriminant of the analysis. The modeling of the signal and the background is examined in detail, with the corresponding treatment of the systematic uncertainties. Finally, the procedure followed for the statistical combination of the results from the three analyzed datasets is reviewed. The combined results of the search and their interpretation in the context of the 2HDM+S and the Dark Photon Model are presented in Chapter 8.

7.1 Signal Topology

The observed boson with a mass of 125 GeV may decay into a pair of the lightest CP-odd states a_1 of the 2HDM+S or into a pair of dark photons Z_D of the Dark Photon Model, as discussed in Chapter 3. One pair of pseudoscalar(vector) bosons $a_1(Z_D)$ (referred to as light bosons for simplicity in this chapter) subsequently decays into a pair of muons and the other into a pair of τ leptons. Due to the large mass difference between the light bosons and the 125 GeV Higgs boson (referred to as the Higgs for simplicity in this chapter), the first ones are expected to be highly boosted.

The muons of the muon pair can be easily identified with the CMS detector. The tau leptons of the tau pair are much more difficult to identify, with the complications in the reconstruction coming from their short lifetime, vast decay spectrum, and the presence of neutrinos in all their decay modes, as discussed in Subsec. 3.3.2. In this analysis, the 1-prong leptonic and hadronic decays of the tau lepton are considered, with the term prong referring to the number of final state charged particles. In the 1-prong hadronic mode, the tau lepton decays into one charged hadron and one or more neutral particles. The $\tau_{\text{1-prong}}\tau_{\text{1-prong}}$ decay is reconstructed in the analysis by selecting two opposite-charged tracks with no restriction on the neutral particles surrounding both tracks.

The final state of the analysis consists of two opposite-charged muons from the dimuon pair and two opposite-charged tracks from the decay of the pair of tau leptons, fulfilling certain kinematic requirements. A pictorial representation of the signal topology, depicting the highly collimated trk-trk and μ-μ systems, is shown in Fig. 7.1.

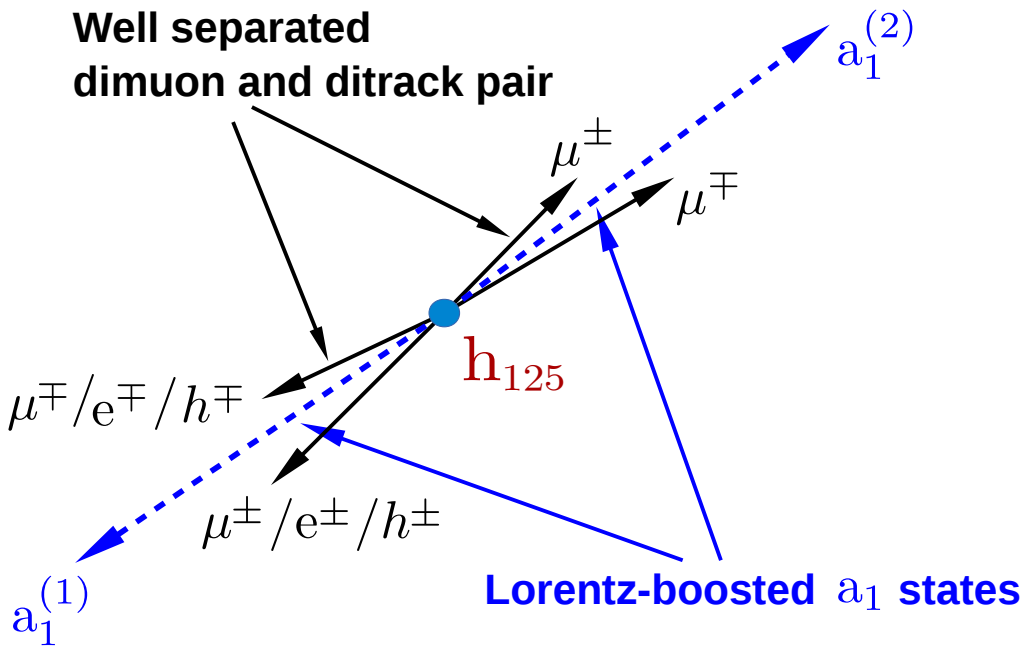


Figure 7.1: Illustration of the signal topology. The Higgs boson decays into two light bosons, with one of the light bosons decaying into a pair of muons and the other into a pair of tau leptons. The analyzed final state consists of two opposite-charged muons and two opposite-charged tracks.

The targeted Higgs boson production mode is the dominant gluon-gluon fusion process, in which the H(125) state is mainly produced with a relatively small transverse momentum, as discussed in Subsec. 2.6.1. Therefore, the light bosons are produced quite separated in the plane transverse to the beam, with the separation only reduced when the gluon is radiated off the top quark loop or any of the initial gluons. A conservative threshold was considered for the separation $\Delta R(\Delta\eta, \Delta\phi)$ between the light boson candidates, to take into account this last scenario. The topological characteristics illustrated in this section, together with the event kinematics features, are exploited in the event selection, described in Sec. 7.3.

7.2 Datasets and Simulated Samples

The analysis uses the LHC proton-proton collision datasets recorded with the CMS detector at $\sqrt{s} = 13$ TeV in the years 2016, 2017, and 2018, corresponding to an integrated luminosity of 35.9, 41.5, and 59.7 fb^{-1}, respectively. The total integrated luminosity is 137.2 fb^{-1}. The datasets include only the so-called *good runs*, runs certified as suitable for physics analysis where the LHC was providing stable beams and where the triggers, the CMS tracker, and the muon system performed well, according to certain established quality criteria. These so-called primary datasets are defined by a set of HLT paths and an event is only stored if it passes at least one of the paths. The conditions for an event to enter a primary dataset are

carefully defined, since the good performance and full exploitation of the physics potential of the experiment relies on the adequate selection of interesting events. Tab. 7.1 lists the primary datasets used in the analysis, the associated run ranges, and the corresponding integrated luminosities. The first element in the name of the listed datasets indicates the kind of triggers used to collect the events, in this case, Single Muon triggers.

Table 7.1: Overview of the datasets used in the analysis, the run ranges, and the corresponding integrated luminosities.

Dataset	Run range	Luminosity [fb^{-1}]
2016		
/SingleMuon/Run2016B-17Jul2018-ver2-v1/MINIAOD	272007-275376	5.788 /fb
/SingleMuon/Run2016B-17Jul2018-ver1-v1/MINIAOD	272007-275376	see above /fb
/SingleMuon/Run2016C-17Jul2018-v1/MINIAOD	275657-276283	2.573 /fb
/SingleMuon/Run2016D-17Jul2018-v1/MINIAOD	276315-276811	4.248 /fb
/SingleMuon/Run2016E-17Jul2018-v1/MINIAOD	276831-277420	4.009 /fb
/SingleMuon/Run2016F-17Jul2018-v1/MINIAOD	277772-278808	3.102 /fb
/SingleMuon/Run2016G-17Jul2018-v1/MINIAOD	278820-280385	7.540 /fb
/SingleMuon/Run2016H-17Jul2018-v1/MINIAOD	280919-284044	8.606 /fb
2017		
/SingleMuon/Run2017B-31Mar2018-v1/MINIAOD	297046-299329	4.792 /fb
/SingleMuon/Run2017C-31Mar2018-v1/MINIAOD	299368-302029	9.755 /fb
/SingleMuon/Run2017D-31Mar2018-v1/MINIAOD	302030-303434	4.319 /fb
/SingleMuon/Run2017E-31Mar2018-v1/MINIAOD	303824-304797	9.424 /fb
/SingleMuon/Run2017F-31Mar2018-v1/MINIAOD	305040-306462	13.50 /fb
2018		
/SingleMuon/Run2018A-17Sep2018-v2/MINIAOD	315252-316995	14.00 /fb
/SingleMuon/Run2018B-17Sep2018-v1/MINIAOD	317080-319310	7.10 /fb
/SingleMuon/Run2018C-17Sep2018-v1/MINIAOD	319337-320065	6.94 /fb
/SingleMuon/Run2018D-22Jan2019-v2/MINIAOD	320673-325175	31.93 /fb

For each year of data taking, separate signal samples with independent statistics were produced. The Higgs boson production is simulated at leading order with the MADGRAPH event generator, interfaced with PYTHIA8, followed by the decay into a pair of light bosons and the subsequent decay into a pair of muons and a pair of tau leptons. More precise spectrum calculations at Next-to next-to Leading Order (NNLO) with re-summation to next-to-next-to-Leading-Leading (NNLL) order are considered by reweighting the p_T distribution of the Higgs boson that emerges from the gluon-gluon fusion process. The factors for the re-weighting, called *k-factors*, are obtained with the program HqT [200, 201]. The weight

associated to each of the generated events is computed as:

$$\omega_{\mathrm{MC}} = \frac{\sigma \cdot \mathcal{L} \cdot \varepsilon_{\mathrm{MC}}}{N_{\mathrm{processed}}}, \tag{7.1}$$

where σ denotes the cross-section of the process considered, $\mathcal{L}$ the integrated luminosity, $\varepsilon_{\mathrm{MC}}$ the MC filter efficiency, and $N_{\mathrm{processed}}$ the number of processed MC events. The MC filter efficiency represents the fraction of generated events that pass a certain filter or filters (set of selection criteria) required at the generation level in order to enhance the acceptance to a probed signal. After the generation step, the events are passed through a full simulation of the CMS detector based on the GEANT4 package. The effects of the pileup are also included in this step. Upon completion of the simulation chain, the Monte Carlo samples are ready for the physics analysis.
The signal MC samples used in the analysis are summarized in Tab. 7.2, covering the mass range of the light boson between 3.6 and 21 GeV. The simulated datasets for the signal corresponding to the years 2016 and 2017 were created within CMS official production campaigns. Additional private signal MC samples for 2018 were also produced, following the same workflow and configuration of the centrally produced samples. The set of parameters of the Monte Carlo event generator for the modeling of the underlying-events, called tune, is changed between the years, as shown in the name of the MC samples. The tune CUETP8M1, which was created before data measured at $\sqrt{s} = 13$ TeV became available, is the one used for the production of the 2016 samples. For the 2017 and 2018 samples, the set of tunes CPX became available and the tune CP5 is used. The label CPX stands for CMS Pythia 8 and X is a progressive number related to the simulation of the multiparton interaction.
Tab. 7.3 summarizes the background MC samples used in the analysis. The QCD multijet background is simulated with PYTHIA8. For the production of the samples a filter that selects events containing at least one muon with transverse momentum higher than 5 GeV is added. The phase space was divided in bins of p_T, to improve the statistics of the generated samples. The top-pair and single top production are simulated using POWHEG interfaced to PYTHIA. The W/Z boson plus jets samples, with the subsequent leptonic decays of the W and Z bosons, and the Drell Yan samples, are produced using MADGRAPH interfaced to PYTHIA. The inclusive WW, WZ, and ZZ background processes are simulated with PYTHIA.

Table 7.2: Signal Monte Carlo samples used in the analysis.

Dataset
$gg \to H(125), H(125) \to a_1 a_1 (Z_D Z_D) \to 2\mu 2\tau$
2016 (centrally produced)
SUSYGluGluToHToAA_AToMuMu_AToTauTau_M-125_M-a_1_TuneCUETP8M1_MADGRAPH_PYTHIA8
2017 (centrally produced)
SUSYGluGluToHToAA_AToMuMu_AToTauTau_M-125_M-a_1_TuneCP5_13TeV_MADGRAPH_PYTHIA8
2018 (privately produced)
SUSYGluGluToHToAA_AToMuMu_AToTauTau_M-125_M-a_1_TuneCP5_13TeV_MADGRAPH_PYTHIA8

Table 7.3: Simulated datasets for the background processes. For each sample the corresponding cross-section is listed, times the filter efficiency in the case of the QCD multijet datasets. All samples have been created within the official production campaigns of 2016.

Description	$\sigma(\times\epsilon_{MC})$ at 13 TeV [pb]
Z + Jets, $m_{ll} < 50$ GeV	
1 GeV $< m_{ll} <$ 5 GeV, 70 $< p_T <$ 100 GeV	879 (LO)
1 GeV $< m_{ll} <$ 5 GeV, 100 $< p_T <$ 200 GeV	640.6 (LO)
1 GeV $< m_{ll} <$ 5 GeV, 200 $< p_T <$ 400 GeV	107 (LO)
1 GeV $< m_{ll} <$ 5 GeV, 400 $< p_T <$ 600 GeV	10.85 (LO)
1 GeV $< m_{ll} <$ 5 GeV, $p_T >$ 600 GeV	3.412 (LO)
5 GeV $< m_{ll} <$ 50 GeV, 70 $< p_T <$ 100 GeV	302.2 (LO)
5 GeV $< m_{ll} <$ 50 GeV, 100 $< p_T <$ 200 GeV	224.2 (LO)
5 GeV $< m_{ll} <$ 50 GeV, 200 $< p_T <$ 400 GeV	37.19 (LO)
5 GeV $< m_{ll} <$ 50 GeV, 400 $< p_T <$ 600 GeV	3.581 (LO)
5 GeV $< m_{ll} <$ 50 GeV, $p_T >$ 600 GeV	1.124 (LO)
$t\bar{t}$	831.76 (NNLO)
single top	
tW_antitop_5f_inclusiveDecays	35.85 (NLO)
tW_top_5f_inclusiveDecays	35.85 (NLO)
t_channel_top_4f_leptonDecays	136.02 (NLO)
t_channel_antitop_4f_leptonDecays	80.95 (NLO)
W + Jets	61526.7 (NNLO)
inclusive WW	118.7 (NNLO)
inclusive WZ	27.57 (NLO)
inclusive ZZ	12.14 (NLO)
QCD multijets (LO), $p_T^\mu > 5$ GeV	
15 $< p_T <$ 20 GeV	1273190000 × 0.003
20 $< p_T <$ 30 GeV	558528000 × 0.0053
30 $< p_T <$ 50 GeV	139803000 × 0.01182
50 $< p_T <$ 80 GeV	19222500 × 0.02276
80 $< p_T <$ 120 GeV	2758420 × 0.03844
120 $< p_T <$ 170 GeV	469797 × 0.05362
170 $< p_T <$ 300 GeV	117989 × 0.07335
300 $< p_T <$ 470 GeV	7820.25 × 0.10196
470 $< p_T <$ 600 GeV	645.528 × 0.12242
600 $< p_T <$ 800 GeV	187.109 × 0.13412
800 $< p_T <$ 1000 GeV	32.3486 × 0.14552
$p_T >$ 1000 GeV	10.4305 × 0.15544

7.3 Physics Objects and Event Selection

An event in this analysis is considered for further selection steps if it contains an isolated muon. The details of the offline event selection, the corrections to simulation, and the analysis event categorization are discussed in this section.

7.3.1 Trigger selection

The trigger system of CMS, as discussed in Subsec. 4.2.6, consists of two levels designed to select events with a potential physics interest. Tab. 7.4 lists the single muon triggers used in the analysis for each year of data taking.

Table 7.4: Triggers used in the analysis and the corresponding data-taking year.

Year	Trigger
2016	HLT_IsoTkMu24
2017	HLT_IsoMu27
2018	HLT_IsoMu24

The muons are identified by the HLT muon triggers using information from the muon system and the tracker subdetectors. The so-called L2 muons are reconstructed with information from the muon system only, while for the L3 muons, the information from both the tracker and muon subdetectors is exploited in a global fit of tracker and muon hits. The tracker hits for the L3 muon reconstruction come only from a portion of the tracker volume pointed by the L2 muons. For the tracker muons, the hit reconstruction is performed in the whole volume of the tracker, and then a match with DT and CSC segments is required, allowing the recovery from inefficiencies in the L2 muon reconstruction caused by the muon detector acceptance, among other reasons. A relative isolation requirement is applied for a further reduction of the rate after the muon identification. The energy deposits in the ECAL and HCAL of the tracks in a ΔR cone of 0.3 around the muons are summed and the relative isolation variable takes the form:

$$I_{\text{rel}} = \frac{1}{p_T^{\mu}} \left(\sum_i p_{T,\text{trk}}^i + \max(0, \sum_j E_{T,\text{ECAL}}^j + \sum_k E_{T,\text{HCAL}}^k - \pi(\Delta R)^2 \rho) \right). \tag{7.2}$$

A correction to the energy calorimeter deposits is introduced in order to reduce the dependence of this magnitude on the pileup conditions, using the average energy density ρ in the event [127]. The standard selection requires the relative isolation variable to be lower than 0.15. The logic of the triggers used in the analysis, taking as an example the IsoMu24 trigger, can then be summarized as follows: a single muon trigger is first seeded by a L1 trigger of $p_T > 16$ GeV, requirement of a L2 track of $p_T > 16$ GeV, requirement of a L3 track of $p_T > 16$ GeV, and isolation of the L3 track. A good trigger acceptance for the signal of the analysis is obtained despite having two muons produced with a small separation in (η, ϕ), because the online isolation sum does not include nearby muons. The additional muon within the isolation cone of the triggering muon can only be cleaned if it was indeed reconstructed as a muon,

with the identification efficiency varying for L3 and tracker muons. Therefore, the fact that IsoMu and IsoTkMu have different reconstructions, using L3 and tracker muons, respectively, explains the small difference observed in signal acceptance between both triggers.

7.3.2 Veto on b-tagged jets

The jets are reconstructed in the analysis from particle flow objects, using the anti-k_T clustering algorithm with a cone radius parameter of 0.4, and the jets originating from b-jets are identified with the DeepCSV algorithm (Subsec. 5.3.6). An early step in the event selection chain of the analysis consists of requiring the events that matched the trigger to have zero b-jets with p_T higher than 20 GeV. The vetoing of these events allows us to reject backgrounds with b-quark jets, like the $t\bar{t}$ background. The medium working point of the DeepCSV algorithm, which implies an efficiency for b-jets identification of about 70% and a misidentification rate for light-flavor jets of about 1%, is chosen. The next step in the analysis chain after the trigger selection and the veto on b-tagged jets is the identification of the pair of muons and the pair of tau leptons.

7.3.3 Muon identification and selection

The dimuon pair produced in the decay of one of the light bosons is formed by two opposite-charge muons identified by the CMS particle flow algorithm and reconstructed by the global reconstruction algorithm, as described in Subsec. 5.3.3. The muon with the highest p_T of the pair is regarded as leading muon and the other as sub-leading or trailing muon. Both muons must fulfill certain baseline identification selections and isolation cuts summarized in the so-called Muon ID, specifically, the medium working point (medium Muon ID). Furthermore, one of the muons of the pair is required to match the single muon trigger IsoTkMu24, IsoMu27, or IsoMu24 for the datasets of 2016, 2017, and 2018, respectively, as reported in Tab. 7.4. The trigger object must match the offline muon object within a ΔR cone of 0.5. Since the Medium ID was designed to be highly efficient for prompt muons, but also for muons from heavy quark decays, impact parameter cuts need to be applied when reconstructing the analysis signature, in order to reject the heavy quark decays. The impact parameter in the transverse (longitudinal) plane with respect to the primary vertex of the tracks corresponding to the two muons of the dimuon pair is required to be $|d_0| < 0.05$ cm ($|d_z| < 0.1$ cm). The following kinematic cuts are also imposed as a part of the selection:

- pseudorapidity of the leading and the sub-leading muons, $|\eta| < 2.4$.
- p_T of the muon matching the trigger higher than 25 GeV (2016 and 2018) or 28 GeV (2017) (the reconstructed muon must have a p_T higher than 1 GeV with respect to the corresponding online trigger threshold).
- p_T of the muon not matching the trigger higher than 3 GeV.

7.3.4 Track selection

The analysis uses a set of the reconstructed tracks known as *high purity* tracks, which pass strict selection criteria based on the χ^2/ndof of the track fit, the transverse and longitudinal impact parameters with respect to the primary vertex, the significance of the impact parameters ($d_0/\delta d_0$ and $d_z/\delta d_z$, where δd_0 and δd_z are the uncertainties on the impact parameters of the track fit), the number of tracker layers with a hit on the track, the number of tracker

"3D" layers with a hit on the track (either pixel layers or matched strip layers), the number of layers missing hits between the first and last hit on the track, and the $\delta p_T/p_T$ from the track fit. The tracks must fulfill the following kinematic requirements: $p_T > 1$ GeV, pseudorapidity $|\eta| < 2.4$, and impact parameter in the transverse (longitudinal) plane with respect to the primary vertex $|d_{xy}| < 1$ cm ($|d_z| < 1$ cm). The isolation criteria described in the next subsection start by counting the number of tracks around the muons of the dimuon pair and the tracks of the ditrack pair that fulfill the listed requirements. Due to the loose impact parameter cuts, the selected track collection is populated with tracks directly coming from the PV, but also with displaced tracks from decays of heavy flavor hadrons, which get reduced through a further set of topological cuts, discussed in the next subsection.

7.3.5 Topological selection

The topology of the probed signal is exploited in the selection of the light boson candidates, which are required to be separated by a ΔR higher than 1.5, according to the explanation given in Sec. 7.1.

Selection of the $a_1(Z_D) \to \mu\mu$ candidates. The $a_1(Z_D) \to \mu\mu$ candidates are formed by two opposite-sign muons, selected as described in Subsec. 7.3.3. The sum of the transverse momentum of the two muons:

$$p_T\text{Sum} = \sqrt{(p_{x_{\mu^+}} + p_{x_{\mu^-}})^2 + (p_{y_{\mu^+}} + p_{y_{\mu^-}})^2}, \tag{7.3}$$

is required to be higher than 45 GeV. The value was selected by optimizing the signal to background ratio in a sample of dimuon pairs. Moreover, considering that the total energy of each of the two light bosons in the reference frame of the Higgs boson is of roughly 62.5 GeV (half of the total invariant mass of the Higgs boson), the probed mass range of the light boson, and the component of the momentum carried away in the z-direction (p_z), it results natural to apply a cut around this value to the p_T distribution. Additional cuts on the impact parameter are imposed to the muons ($|d_{xy}| < 0.01$ cm and $|d_z| < 0.03$ cm). These cuts practically do not reject any signal and significantly reject background events with displaced tracks. If more than one $a_1(Z_D) \to \mu\mu$ candidate is found in an event, the pair with the highest p_TSum is selected.

Selection of the $a_1(Z_D) \to \tau\tau$ candidates. The $a_1(Z_D) \to \tau\tau$ candidates are formed by two opposite-sign high purity tracks, selected as described in Subsec. 7.3.4. The tracks identified as muons, electrons or hadrons, are required to fulfill a set of selection criteria, to be classified as *signal tracks*. The net charge of the ditrack pair must be zero ($q_{\text{trk}_1} + q_{\text{trk}_2} = 0$) and for both tracks the following kinematic requirements must be fulfilled: $p_T > 2.5$ GeV, $|\eta| < 2.4$, $|d_0| < 0.02$ cm, and $|d_z| < 0.04$ cm. The sum of the transverse momentum of the two tracks (defined as in Eq. (7.3), substituting μ^+ with $\texttt{trk}^+$ and μ^- with $\texttt{trk}^-$), is required to be higher than 10 GeV. A pair of tracks passing these signal track requirements form an $a_1(Z_D) \to \tau\tau$ candidate. If more than one $a_1(Z_D) \to \tau\tau$ candidate is found in an event, the candidate with the highest p_TSum is selected.

The mass of the dimuon system is required to be higher than the mass of the ditrack system. Since the $a_1(Z_D) \to \tau\tau$ decay is only kinematically allowed starting from masses of the light boson twice the mass of the τ lepton, a cut of 3.5 GeV is applied on the invariant mass of the $a_1(Z_D) \to \mu\mu$ candidate. The reconstruction of the two τ leptons is done under the

collinear approximation, considering the kinematic constraints on the mass of the τ lepton and the mass of the light boson ($m_{\tau_1,\tau_2} = m_{\mu_1,\mu_2}$). This approximation is based on two main assumptions. First, that the neutrinos from the τ decays are nearly collinear with the corresponding visible τ decay products:

$$\overrightarrow{p}_{\tau_1} = \alpha \cdot \overrightarrow{p}_{\tau_{1-vis}} \qquad \overrightarrow{p}_{\tau_2} = \beta \cdot \overrightarrow{p}_{\tau_{2-vis}}, \tag{7.4}$$

and second, that all the MET of the event comes from these neutrinos [202]. The assumptions lead to the equation:

$$\overrightarrow{\alpha}_{\beta} = \overrightarrow{1} + P^{-1} \cdot \overrightarrow{E}_T^{\text{miss}}, \tag{7.5}$$

where $\overrightarrow{\alpha}_{\beta}$ is the vector $(\alpha, \beta)^T$ and P the matrix with components $P_{00} = p^X_{\tau_{1-vis}}, P_{01} = p^X_{\tau_{2-vis}}, P_{10} = p^Y_{\tau_{2-vis}}$, and $P_{11} = p^Y_{\tau_{2-vis}}$ [203]. Under the collinear approximation, but considering only the absolute value of the MET and the p_T of the invisible τ decay products as equal, the equations to be solved take the form:

$$m^2_{\mu_1\mu_2} - 2m^2_\tau - 2\sqrt{m^2_\tau + \alpha^2 \overrightarrow{p}^2_{\tau_1^{vis}}}\sqrt{m^2_\tau + \beta^2 \overrightarrow{p}^2_{\tau_2^{vis}}} - \alpha\beta(\overrightarrow{p}_{\tau_1^{vis}} \cdot \overrightarrow{p}_{\tau_2^{vis}}) = 0, \tag{7.6}$$

$$(\alpha-1)^2 \overrightarrow{p}^2_{\tau_1^{vis}} + (\beta-1)^2 \overrightarrow{p}^2_{\tau_2^{vis}} + 2(\alpha-1)(\beta-1)(\overrightarrow{p}_{\tau_1^{vis}} \cdot \overrightarrow{p}_{\tau_2^{vis}}) - (\overrightarrow{E}_T^{\text{miss}})^2 = 0,$$

and are solved numerically. Once the contribution from the neutrinos is estimated, the invariant mass of the ditau system can be calculated as $M_{\tau_1\tau_2} = m_{\text{vis}}/\sqrt{x_1x_2}$, $x_{1,2} = p_{\text{vis}_{1,2}}/(p_{\text{vis}_{1,2}} + p_{\text{miss}_{1,2}})$, where m_{vis} is the invariant mass of the visible decay products and $x_{1,2}$ is the momentum fraction of the visible decay products out of the total momentum in the denominator.

The reconstructed visible mass of the two muons and the two tracks is required to be lower than 125 GeV, and a mass window around the mass of the dimuon-plus-ditau system of 75 GeV is applied. The value of 75 GeV was selected considering the resolution of the reconstructed Higgs mass, reached by means of the reconstruction algorithm for the ditau pair explained above.

The analysis of the dataset corresponding to one year will be referred hereafter with the nomenclature *YEAR analysis*, where $YEAR$ = 2016, 2017, or 2018, e.g., 2016 analysis. The dependence of the ΔR distributions on the mass of the light boson for the $a_1(Z_D) \to \mu\mu$ and $a_1(Z_D) \to \tau\tau$ candidates was studied at generator level on signal MC samples. Fig. 7.2 depicts the obtained results for one of the analysis categories, corresponding to the samples used for the 2018 analysis. The central value of the distribution is shifted towards larger ΔR values as the mass of the light boson increases, since the larger the mass of the light boson is, the less boosted the system becomes, which leads to less collimated pair of muons and therefore, larger ΔR cone. Only a loose requirement on the ΔR cone between the muons of the $a_1(Z_D) \to \mu\mu$ candidates of 1.5 is set, since as seen in Fig. 7.2, a unique optimized cut would not be possible for all the mass range of the light boson. The same applies to the tracks of the $a_1(Z_D) \to \tau\tau$ candidates.

The invariant mass distribution of the dimuon system peaks at the main value of the light boson, with the kinematic variables of the reconstructed muons being the main observables that affect the shape of the peak. One of these kinematic variables, the sum of the p_T of the two muons, is depicted in the left plot of Fig. 7.3. The same distribution for the ditrack

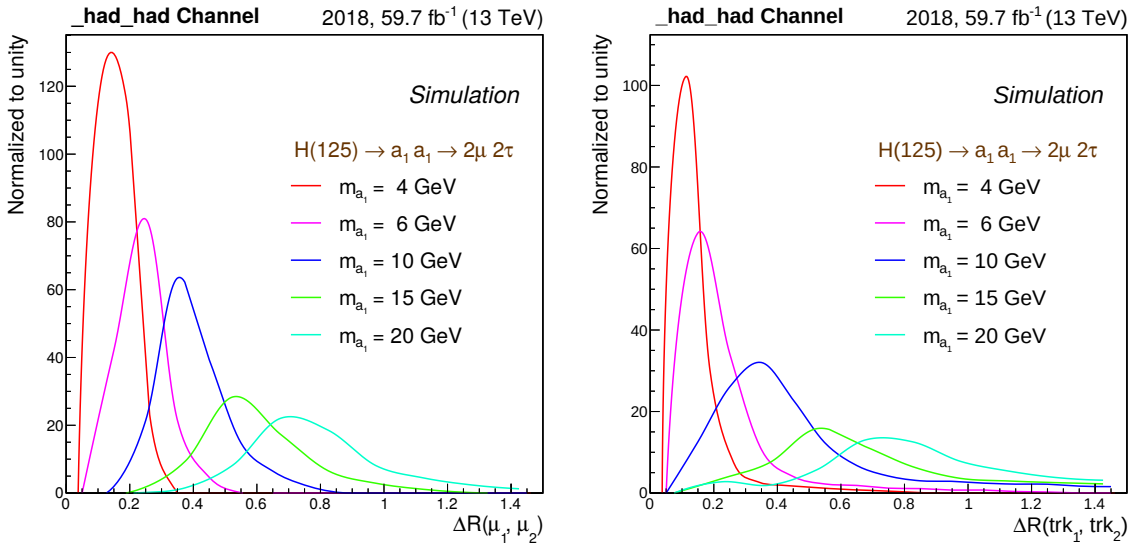


Figure 7.2: Distribution of ΔR for the $a_1(Z_D) \to \mu\mu$ (left) and $a_1(Z_D) \to \tau\tau$ (right) candidates. Different masses across the light boson spectrum are shown for the hadron-hadron category corresponding to the 2018 analysis.

system is illustrated in the right plot of Fig. 7.3. Both auxiliary distributions are obtained after the dimuon and ditrack selection without imposing any isolation requirement.

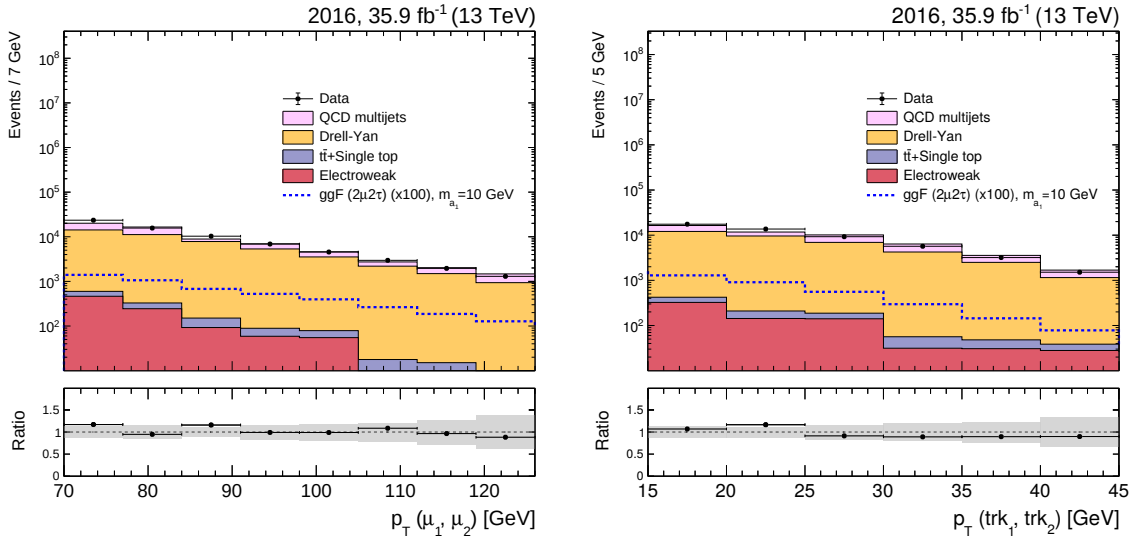


Figure 7.3: Distribution of p_T of the $a_1(Z_D) \to \mu\mu$ candidate (left) and the $a_1(Z_D) \to \tau\tau$ candidate (right), corresponding to the 2016 analysis. Data (dots) is compared with the expectation from simulation (histograms). The Electroweak label indicates the electroweak backgrounds W/Z boson plus jets, and diboson production. The blue-dashed histogram shows the signal distribution for the mass hypothesis $m_{a_1(Z_D)} = 10$ GeV. The lower panel shows the ratio between data and the background yield.

A good agreement between data and the predictions from simulation is observed. The Drell-Yan background is the dominant one, followed by the QCD multijet background, contributing roughly to 73 and 25% of the total background, respectively. The $t\bar{t}$ + single top and electroweak processes are estimated to represent around 1% each.

Isolation requirement. The final step of the selection is imposing an isolation requirement on the two tracks of the $a_1(Z_D) \to \tau\tau$ leg and the two muons of the $a_1(Z_D) \to \mu\mu$ leg. The number of tracks within a ΔR cone of 0.2 around the four objects (the two muons and the two tracks) with momentum $p_T > 1$ GeV, pseudorapidity $|\eta| < 2.4$, and impact parameters $|d_0| < 1$ cm and $|d_z| < 1$ cm, must be zero. Due to the expected boosting of the light bosons one of the objects might be within the isolation cone of its partner muon or track. Therefore, the tracks from the three additional objects are not included in the counting of the number of tracks within the ΔR cone of 0.2 of one of the objects, in a similar approach to the cleaning performed during the computation of the online isolation sum for the single muon triggers (Subsec. 7.3.1).
The region determined through all the selection criteria described above defines the so-called *signal region* of the analysis: a dimuon and a ditrack system, with both muons and both tracks isolated within a ΔR cone of 0.2. As one gradually approaches to the signal region of the analysis, the assessment of the background composition becomes more difficult due to a lack of statistics in the MC samples. Further phase spaces selected by relaxing the isolation requirement, so-called *control regions*, are used for the estimation of the background composition with the data themselves. The control regions used in the analysis are introduced in Sec. 7.7.

7.3.6 Event categorization

To increase the sensitivity of the analysis, the selected events are categorized according to the identification of the $a_1(Z_D) \to \tau\tau$ candidate tracks as lepton or hadron. This results in three categories: lepton-lepton, lepton-hadron, and hadron-hadron. The lepton-lepton category includes the cases in which both tracks are muons, both tracks are electrons, or one track is a muon and the other is an electron. The lepton-hadron category includes the cases in which one of the tracks is a muon or electron, and the other is a hadron, while the hadron-hadron category corresponds to the case in which both tracks are hadrons. A pictorial representation of all possible combinations and the three resulting categories is shown in Fig. 7.4.

7.4 Corrections to simulation

The accuracy of the description of the data with the MC simulation is affected by limitations in the physics event generation and the detector simulation, such as the limited computing capacity, limited accuracy of the MC generators in the theoretical calculations, and the complexity associated to the quite dynamic detector operation and beam conditions, which can not be fully reflected in the simulation. This section is dedicated to present the corrections to the MC simulation applied in the analysis to improve the description of the data. The corrections mainly concern the signal, since the background modeling is based on a data driven method, as further discussed in Sec. 7.7.

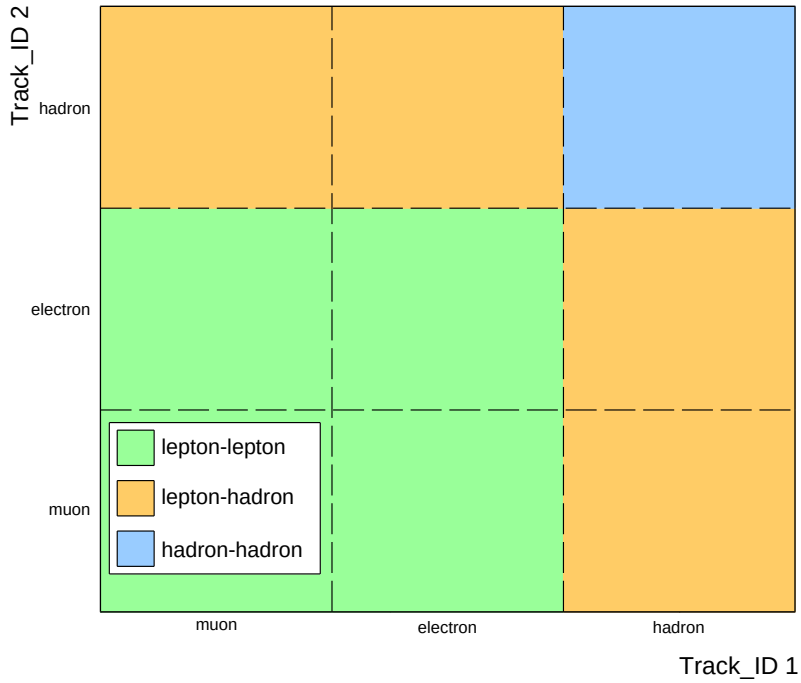


Figure 7.4: Definition of the analysis categories: lepton-lepton (green), lepton-hadron (orange), and hadron-hadron (blue).

7.4.1 Pileup reweighting

The MC samples are produced considering a distribution for the number of pileup interactions, which aims to emulate the conditions expected for each year of data taking. Nevertheless, the trigger and the applied offline event selection can bias this distribution. The number of pileup interactions is also sensitive to differences in the underlying events between data and simulation. Thus, a correction is applied to match the simulated distribution of reconstructed primary vertices with the corresponding distribution in data, given the recommended minimum bias cross-section of 69.2 mb. The so-called pileup reweighting is applied to each MC event in bins of pileup.

7.4.2 Muon ID and trigger efficiency

The efficiency of the muon identification is measured relative to the number of reconstructed muons [204]. The measurement is done in both data and simulation using the tag-and-probe method applied to a sample of $Z \to \mu\mu$ or $J/\Psi \to \mu\mu$ events [205, 206]. This method constitutes a generic tool for measuring any defined object efficiency, by exploiting di-object resonances like Z or J/Ψ. It uses an object that passes tight selection criteria prompting a minimal fake rate and an object passing loose selection criteria, known as tag and probe, respectively. The probes are paired with tags, to obtain an invariant mass consistent with the mass of the $Z \to \mu\mu$ or $J/\Psi \to \mu\mu$ resonance, and the number of signal pairings (N_{all}) is counted. A passing probe will satisfy the particular selection criteria from which one wants

to measure the efficiency. In the analysis, it would be a muon passing the medium ID. The number of tag + passing signal pairings (N_{pass}) constitutes the numerator of the ratio in the efficiency calculation. The efficiency of the probe selection criteria can then be calculated as the ratio of the number of probes that pass the selection criteria relative to the total number of signal pairings:

$$\varepsilon = \frac{N_{\text{pass}}}{N_{\text{all}}}. \tag{7.7}$$

A signal + background model is used to fit the two-line shapes (tag + passing probe) and (tag + failing probe) separately. The ratio of the signal yields is computed in bins of the kinematic variables p_T and $|\eta|$. An efficiency histogram is obtained for both data and simulation independently, and the simulation is then corrected by applying the weight:

$$w^{\mu,\text{Id}} = \frac{\varepsilon^{\mu,\text{Id}}_{\text{data}}}{\varepsilon^{\mu,\text{Id}}_{\text{MC}}} = \frac{\varepsilon^{\mu_1,\text{Id}}_{\text{data}} \cdot \varepsilon^{\mu_2,\text{Id}}_{\text{data}}}{\varepsilon^{\mu_1,\text{Id}}_{\text{MC}} \cdot \varepsilon^{\mu_2,\text{Id}}_{\text{MC}}}, \tag{7.8}$$

on an event-by-event basis. The weights $\varepsilon^{\mu,\text{Id}}_{\text{data}}$ and $\varepsilon^{\mu,\text{Id}}_{\text{MC}}$ are the products of the weights corresponding to the two muons of the $a_1(Z_D) \to \mu\mu$ candidate in data and MC, respectively. The trigger efficiencies of the single muon triggers used in the analysis are also measured in data and MC as a function of the muon p_T and $|\eta|$ with the tag-and-probe method applied to a sample of $Z \to \mu\mu$ events. The tag muons from the Z decays are required to match the single muon trigger, and the probes serve as a source of pure and unbiased muons to test the number of probes that pass the medium ID and the trigger requirement.

7.4.3 Track isolation and one-prong tau decay identification efficiency

The efficiency of the identification of the one-prong tau decay and the isolation of the tracks of the $a_1(Z_D) \to \tau\tau$ candidate is different for data and simulation. Therefore, a combined scale factor is calculated and applied to the simulation. The combined efficiency is measured using a $Z \to \tau_\mu \tau_{\text{1-prong}}$ sample selected with the HLT_IsoMu24 trigger. A muon that passes the medium Muon ID, with $p_T > 25$ GeV, $|\eta| < 2.4$, and a relative $\Delta\beta-$corrected particle-flow isolation $I_\mu < 0.15$, is required. Events are further selected if they have a track that fulfills the one-prong tau identification criteria of the main analysis (Subsec. 7.3.4). The muon and the track are required to be well separated ($\Delta\phi(\mu, \text{trk}) > 2.0$ rad), since the soft p_T spectrum of Drell-Yan events results in decay products from the two τ leptons having a large angular separation. Furthermore, the muon-track pair is required to have opposite charge and fulfill the same isolation requirement of the main analysis. If more than one track has zero tracks with $p_T > 1$ GeV, $|\eta| < 2.4$, $|d_{xy}| < 1$ cm, and $|d_z| < 1$ cm in the isolation cone of 0.2 around their momenta, the one that yields the highest separation in azimuthal angle is chosen. To clean the sample from $Z \to \mu\mu$ events, a muon veto requirement is imposed. Events containing an additional muon apart from the triggering muon are rejected. The top-pair, single-top, and diboson backgrounds are suppressed by requiring the event to have zero jets, while the W + jets background is suppressed by requiring the transverse mass of the muon and the missing transverse mass to satisfy the condition $m_{\text{T}} < 40$ GeV ($m^2_{\text{T}} = m^2 + p^2_x + p^2_y = E^2 - p^2_z$). The selected $Z \to \tau_\mu \tau_{\text{1-prong}}$ events are classified into four regions, which are defined in intervals of p_T of the track, as follows:

- $5 < p_T < 10$ GeV,

- $10 < p_T < 15$ GeV,
- $15 < p_T < 20$ GeV,
- $p_T > 20$ GeV.

A fit of the muon-track invariant mass ($m_{\mu,\mathrm{trk}}$) is done for each of the four track p_T ranges in order to extract the scale factor. Fig. 7.5 shows the distribution before the fit, known as prefit distribution, for the various track p_T ranges of the $Z \to \tau_\mu\tau_{\text{1-prong}}$ sample. The shapes of the $Z \to \tau_\mu\tau_{\text{1-prong}}$ signal, the top-pair, single-top, single-W, diboson, and $Z \to \mu\mu$ background templates are obtained from simulation while the QCD mutijet background is modeled in a data control region, reverting the opposite charge requirement of the muon and the track. An additional high transverse mass control region with $m_\mathrm{T} > 60$ GeV is defined for the determination of the single-W background normalization. The normalization of the QCD background is taken from the same-sign region, considering an extrapolation factor from the same-sign to the opposite-sign region, measured in an additional control region where the track is anti-isolated. The top-pair, single-top, and diboson backgrounds are normalized using the corresponding theoretical predictions for the cross-section. The normalization of the $Z \to \mu\mu$ background is allowed to float freely in the fit.
The following systematic uncertainties are considered in the fit. The uncertainty on the integrated luminosity which affects the yield of the processes estimated from the simulation is included. Additionally, a flat uncertainty for the identification, isolation, and trigger efficiency of 2% is assigned per muon. The uncertainties in the normalization of the backgrounds are 7% for the $\mathrm{t\bar{t}}$ background, 15% for the diboson background, 5 to 12% for the W+jets background, and 15% for the QCD multijet background. An uncertainty of 2% in the extrapolation factor from the $Z \to \mu\mu$ control region to the $Z \to \tau_\mu\tau_{\text{1-prong}}$ signal region for the yield of Drell-Yan events, is considered. The uncertainties in the muon and charged pion momentum scales, which correct the biases in the reconstructed p_T, are taken as 0.3% each. Due to the applied cut on m_T of 40 GeV, the estimation of the yield of simulated events in the $Z \to \tau_\mu\tau_{\text{1-prong}}$ sample is affected by the uncertainty in the reconstruction of the E_T^{miss}, which can be computed as [207]:

$$\overrightarrow{E}_T^{\mathrm{miss}} = -\sum_{\mathrm{j}\in\mathrm{jet}} \overrightarrow{p_T}^{\mathrm{j,\ JEC}} - \sum_{\mathrm{i}\in\mathrm{uncl}} \overrightarrow{p_T}^{\mathrm{i}}. \tag{7.9}$$

The first component represents the contribution from the jets in the event after applying the jet energy corrections (Subsec. 5.3.6) and the second component is associated to the unclustered energy, which comes from the particle flow candidates that are not clustered within jets. Therefore, the uncertainty on the MET is influenced by the uncertainties in the jet energy scale and the unclustered energy scale [170].
The signal and background shapes, the normalization, and the corresponding uncertainties are given as input to the fit together with the data. As a result of the fit, updated shapes and normalization are obtained for the signal and the background. The ratio of the postfit and prefit signal normalization:

$$SF = \frac{\text{Postfit signal norm}}{\text{Prefit signal norm}}, \tag{7.10}$$

constitutes the scale factor for the combined track isolation and one-prong tau decay identification efficiency, applied to the simulation. It is computed for each of the three years. The postfit distribution obtained for the 2018 analysis and the corresponding scale factor as a function of the p_T of the muon-track pair are depicted in Fig. 7.6 and Fig. 7.7, respectively. The scale factors for 2016 and 2017 are shown in Fig. A.1 in Appendix A. The dependence is fitted with a constant, and is applied in the main analysis to simulated events. The higher uncertainty on the value of the scale factor at low p_T is caused by the overwhelming background for low values of the p_T of the track, as observed in the upper left plot of Fig. 7.5 and Fig. 7.6. For higher p_T values, the signal to background ratio is much higher, which results in lower uncertainty values for the scale factor.
A *Fisher test* (F-test) is performed to check if the fit result can be significantly improved by fitting the scale factor dependence on the p_T with a more complex function, with more degrees of freedom. In the context of an F-test, the two models to be compared are said to be nested if the one with fewer degrees of freedom (the *restricted model*) is obtained from the other (the *full model*) by setting one or more parameters to zero [208]. In this case, the nested models, a constant and a linear function, are used to fit the dependence of the scale factor on the p_T and check if the additional degree of freedom of the linear function provides a significantly better fit than the chosen constant. The F-test for n data points takes the form:

$$F = \frac{\chi_1^2 - \chi_2^2}{\chi_2^2} \cdot (\frac{n - p_2}{p_2 - p_1}), \tag{7.11}$$

and follows the Fisher distribution with ($p_2 - p_1$, $n - p_2$) degrees of freedom, where p_1 (p_2) is the number of parameters for the restricted (full) model and $p_2 - p_1$ is the number of constraints on the parameters that reduce the full model to the restricted model. The null hypothesis of the test (model 2 does not provide a significantly better fit than model 1) is rejected if F is greater than the critical value. In the analysis, a linear function does not significantly improve the fit result with respect to a constant; thus, the null hypothesis is accepted.

7.4.4 Higgs p_T reweighting

The signal acceptance in the analysis is dependent on the kinematics of the Higgs boson, mainly on the p_T distribution. The estimate of the signal acceptance is improved after reweighting the p_T spectrum with H(125) p_T NNLO k-factors (Sec. 7.2), obtained with the program HqT. The k-factor corrections applied are shown in Fig. 7.8. Once the k-factors are considered, the corrected p_T distribution matches the higher order predictions for the H(125) p_T spectrum in the gluon-gluon fusion process.

7.4.5 b-tagging efficiency

The requirement on the events to have 0 b-tagged jets (Subsec. 7.3.2), makes necessary to apply a correction to simulation in order to account for differences in the b-tagging efficiency between data and MC. In general, the methods for the determination of this scale factor can be grouped into two general categories, according to the use of event reweighting or not [210]. One of the methods that apply event reweighting corrects the event yields in simulation by

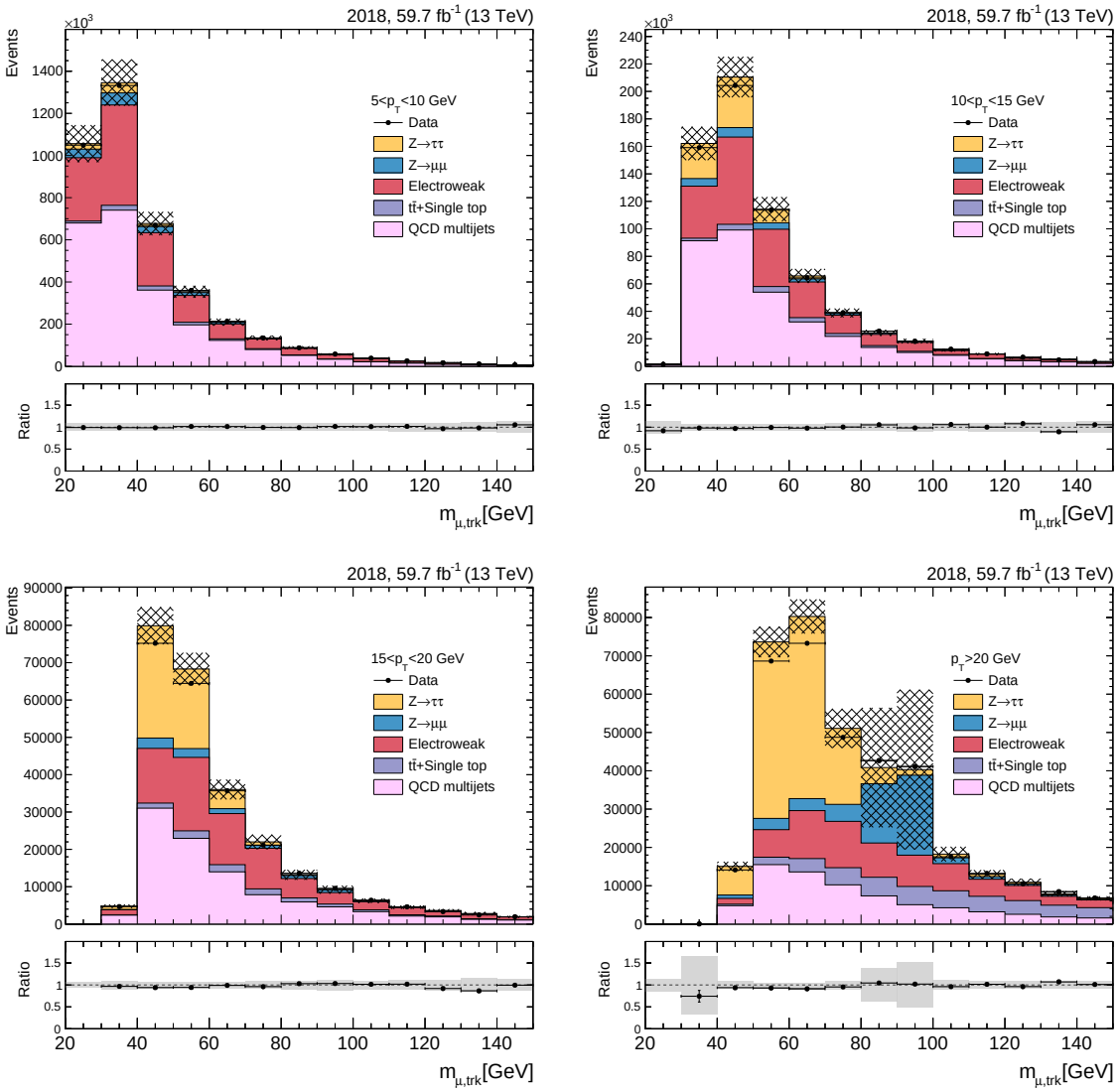


Figure 7.5: Prefit distributions of $m_{\mu,\text{trk}}$ in the selected sample of $Z \to \tau_\mu \tau_{\text{1-prong}}$ events, for the four different ranges of the track p_T: $5 < p_T < 10$ GeV (upper left), $10 < p_T < 15$ GeV (upper right), $15 < p_T < 20$ GeV (lower left), and $p_T > 20$ GeV (lower right), corresponding to the 2018 analysis. The lower panel shows the ratio between data and the signal+background yield.

just changing the weight of the selected MC events, which can be calculated as:

$$w = \frac{P(\text{DATA})}{P(\text{MC})}. \tag{7.12}$$

The terms in the numerator and denominator correspond to the probabilities of having a configuration of jets with i b-tagged jets and j not b-tagged jets, defined as:

$$P(\text{MC}) = \prod_{i=\text{tagged}} \varepsilon_i \prod_{\text{j= not tagged}} (1 - \varepsilon_j), \tag{7.13}$$

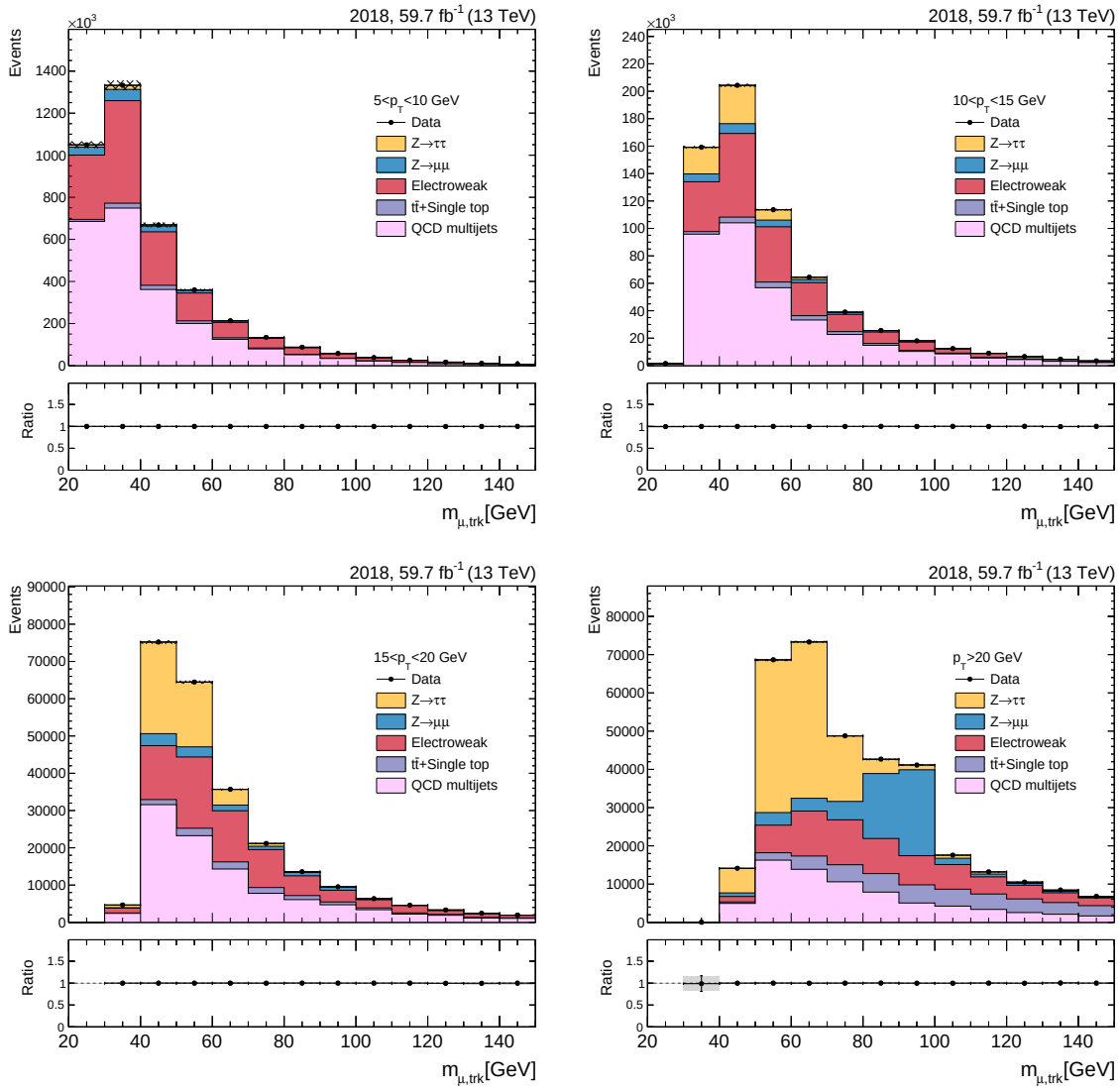


Figure 7.6: Postfit distributions of $m_{\mu,\text{trk}}$ in the selected sample of $Z \to \tau_\mu \tau_{\text{1-prong}}$ events for the four different ranges of the track p_T: $5 < p_T < 10$ GeV (upper left), $10 < p_T < 15$ GeV (upper right), $15 < p_T < 20$ GeV (lower left), and $p_T > 20$ GeV (lower right). The lower panel shows the ratio between data and the signal+background yield. The shown distributions correspond to the 2018 analysis.

$$P(\text{DATA}) = \prod_{i=\text{tagged}} \text{SF}_i\, \varepsilon_i \prod_{\text{j= not tagged}} (1 - \text{SF}_j\, \varepsilon_j),$$

where the scale factors SF_i take the form:

$$SF = \frac{\varepsilon_{\text{data}}(p_T, \eta)}{\varepsilon_{\text{MC}}(p_T, \eta)}, \tag{7.14}$$

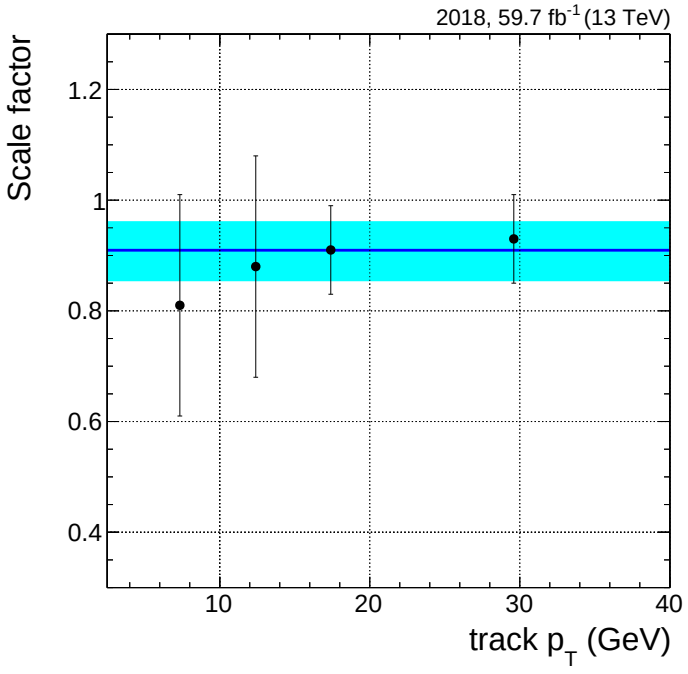


Figure 7.7: Scale factor for the combined track isolation and one-prong tau decay identification efficiency as a function of the track p_T, corresponding to the 2018 analysis. The symetric blue error band represents the 1-sigma error band around the scale factor value, obtained using a linear error propagation method [209]. The constant scale factor is applied in the main analysis to simulated events.

and the MC b-tagging efficiencies ε_i are functions of the jet flavor, the jet p_T, and jet η. The scale factors are available for the different taggers and official working points [176], but the MC b-tagging efficiencies need to be computed for each jet flavor in each of the specific MC samples used for the analysis, due to their dependence on the event kinematics. To overcome this difficulty, a different method was used in the analysis, which does not require the knowledge of the MC b-tagging efficiencies and, therefore, the events can be reweighted using the scale factors only. In this method, the scale factors are taken as pseudo-probabilities, applied only to b-tagged jets.

The same event with n b-tagged jets contributes to all the b-tag multiplicity bins for which the condition $m \leq n$ is fulfilled, where m corresponds to the number of b-tagged jets in a given bin. An event in which n jets are considered for b-tagging can contribute to events with 0 b-tagged jets (as the events in the analysis after the b-jet veto requirement), with the following event weight:

$$w(0|n) = \prod_{i=1}^{n} (1 - \mathrm{SF}_i). \tag{7.15}$$

Therefore, an event with one b-tagged jet contributes to events with 0 b-tagged jets with the event weight $w(0|1) = (1 - SF)$, while events with 0 b-tagged jets contribute to similar events with 0 b-tagged jets with an event weight $w(0|0) = 1$. The overall normalization of

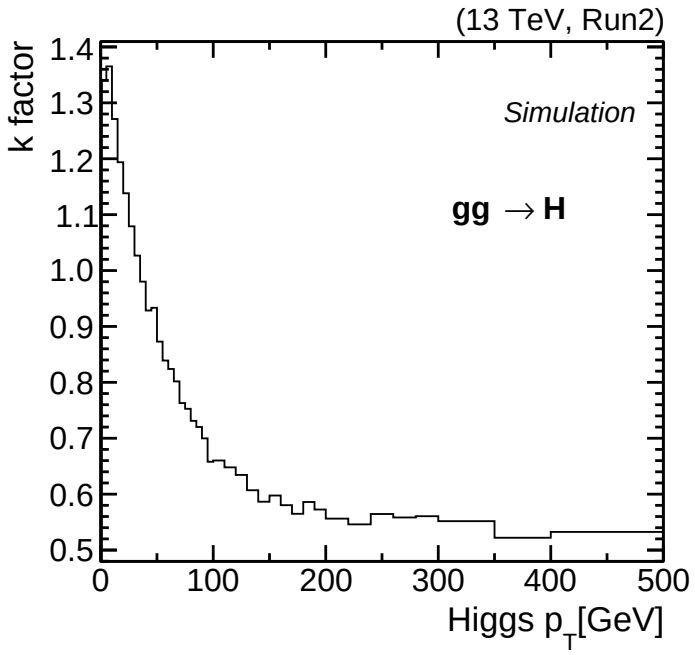


Figure 7.8: NNLO k-factor corrections applied to the H(125) p_T distribution obtained with PYTHIA8 at leading order for the gluon-gluon fusion process.

the event sample follows from:

$$\sum_{i=0}^{n} w(i|n) = 1, \tag{7.16}$$

and is kept unchanged, with only the event yields being changed in the different b-tag multiplicity bins. In practice, the scale factors SF, depend on the p_T and η of the jet. Therefore, the event weights need to be computed on an event-by-event basis, since two events with the same number of b-tagged jets can result in a different weight contribution to $w(0|n)$. The event weight in Eq. (7.15), including the contributions from all the different b-tagged jet multiplicities, is derived and applied on an event-by-event basis to the simulated events.

7.5 Final selected sample

The set of data events that enter the signal region, defined by the selection requirements outlined in Sec. 7.3, form the *final sample* of the analysis, which is used to extract the signal. The number of data events selected in the signal region is reported in Tab. 7.5. The signal acceptance and signal yields for a reference mass point of the light boson are listed in Tab. 7.6. The reported signal yields per category and per year are computed assuming the SM $H(125)$ production cross-section and considering a benchmark value for the branching fraction $\mathcal{B}(H(125) \to a_1 a_1(Z_D Z_D) \to 2\mu 2\tau)$ of 0.1%. Around 30% of the loss in efficiency comes from the detector acceptance and the requirement of having at least two well reconstructed muons in the event. An additional ∼35% efficiency loss comes from the trigger, which requires a high p_T muon and its respective isolation. The rest of the loss in efficiency is caused by the

offline selection. The higher p_T threshold (27 GeV) of the single muon trigger used in the 2017 analysis, with respect to the p_T threshold (24 GeV) of the triggers used for the 2016 and 2018 analysis, results in a reduction of the signal acceptance and, consequently, of the signal yield for the three categories in 2017, as it can be seen in Tab. 7.6. From the three triggers used, IsoTkMu24 has the highest acceptance for the probed signal.

Table 7.5: Number of data events after final selection per category and per year.

Sample	Category	Number of events		
		2016	2017	2018
Data	lepton-lepton	94	65	101
	lepton-hadron	4022	3057	4753
	hadron-hadron	44703	37809	57695

Table 7.6: The signal acceptance and expected signal yields after final selection per category and per year, for $m_{a_1(Z_D)} = 5$ GeV. The reported number of signal events assumes the SM H(125) production cross-section and a benchmark value of the branching fraction $\mathcal{B}(H(125) \to a_1a_1(Z_DZ_D) \to 2\mu2\tau) = 0.1\%$. The quoted uncertainties include only statistical errors.

Signal Acceptance (x100) $\mathcal{A}(pp \to H(125) + X, H(125) \to a_1a_1(Z_DZ_D) \to 2\mu2\tau)$, $gg \to$ H				
$m_{a_1(Z_D)}$ [GeV]	Category	2016	2017	2018
5	lepton-lepton	0.65 ± 0.01	0.54 ± 0.02	0.58 ± 0.01
	lepton-hadron	1.97 ± 0.02	1.47 ± 0.03	1.61 ± 0.02
	hadron-hadron	1.84 ± 0.02	1.31 ± 0.03	1.49 ± 0.02
Number of signal events $(pp \to H(125) + X, H(125) \to a_1a_1(Z_DZ_D) \to 2\mu2\tau)$, $gg \to$ H				
5	lepton-lepton	11.39 ± 0.24	10.91 ± 0.34	16.70 ± 0.30
	lepton-hadron	34.39 ± 0.42	29.56 ± 0.55	46.61 ± 0.50
	hadron-hadron	31.99 ± 0.41	26.47 ± 0.52	43.09 ± 0.48

7.6 Final discriminant: BDT output distribution

In previous analyses searching for light bosons in the $\mu\mu\tau\tau$ final state, the invariant mass distribution of the dimuon system, the ditau system, and the 4-body visible mass, have been used as final discriminant. In this analysis, a method based on a BDT is introduced for discriminating the signal from the background. The introduction of the BDT allows to better profit from the signal topology by adding information from several kinematic variables instead of using only one or two. The Toolkit for Multivariate Analysis (TMVA) was used to build the BDT classifier [199, 209]. The BDT output distribution, obtained for each of the three categories defined in Subsec. 7.3.6, and for each year, constitutes the final discriminant

of the analysis. One classifier is trained per mass hypothesis, per category, and per year, to maximize the discriminating power. The signal class for the training is generated with a multidimensional pdf, introduced in Sec. 7.8, while the background class is formed by the events selected in a control region defined in Sec. 7.7, which is orthogonal to the signal region. To avoid bias analyzing the data, a blinded region in the BDT spectra is defined, keeping the data that is most sensitive to a hypothetical signal blind. The non-blinded region, which corresponds to values of the BDT score smaller than 0.5, is used for the optimization steps of the analysis and for checking the agreement between the data and the background expectation before the "unblinding" procedure, which is performed once the techniques and procedures of the analysis are fully validated. The following subsections are dedicated to discuss the training and performance of the BDT classifier.

7.6.1 BDT input variables

The variables that serve as input to the BDT are: the mass of the dimuon system (m_{μ_1,μ_2}), the ΔR between the muons of the dimuon system ($\Delta R(\mu_1, \mu_2)$), the ΔR between the tracks of the ditrack system ($\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$), and the invariant mass of the four objects, adding the contribution from the MET ($m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$). The set of variables was selected taking into consideration the separation power and the low correlation among themselves.
Fig. 7.9 depicts the MVA input variable distributions for the signal and the background. The signal corresponds to a mass of the light boson of 5 GeV. A ranking for the four variables in terms of *separation power* (unspecific method) and *importance* (specific method) is reported in Tab. 7.7. The ranking corresponds to the BDT training performed for the 2016 analysis in the hadron-hadron category. In the unspecific method, the variables are organized according to their separation power, considering only one variable at a time. In contrast, in the specific method, the importance is calculated with all the variables available for the splitting. The separation power obtained through the preliminary ranking of the unspecific method is superseded by the algorithm-dependent ranking of the specific method. In the case of the specific method, the measure of the variable importance is computed, counting the number of times the variable is used to split decision tree nodes. A weight that considers the square of the separation gain achieved with the split, and the number of events in the node is applied. For a well-ranked variable, the separation between the signal and the background class is high. In contrast, the within-class dispersion, which describes the dispersion of events inside a class relative to the mean of the class, is minimal. The dimuon mass has the highest discriminating power in the analysis, being the first variable in the ranking for both the separation and the importance, mainly due to the low within-class dispersion given by the resonance structure. In contrast, the discriminating power of the mass of the four objects including the MET is limited by the poor MET resolution. This lack of resolution can be improved by using the MET computed after removing the pileup contamination with the Pileup Per Particle Identification method (Subsec. 5.3.6), instead of the corresponding particle flow quantity.

7.6.2 BDT configuration options

The tunning of the BDT configuration results in the following configuration options for the BDT classifier used in the analysis. The number of trees in the forest is set to 500 and the maximum allowed depth of one tree is set to 3. The adaptative boost algorithm

Table 7.7: Ranking of the MVA input variables in terms of separation power and importance.

Rank	Input variable	Separation	Importance
1	m_{μ_1,μ_2}	7.278e-01	3.735e-01
2	$\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$	3.091e-01	2.495e-01
3	$\Delta R(\mu_1, \mu_2)$	2.022e-01	2.274e-01
4	$m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$	4.947e-02	1.495e-01

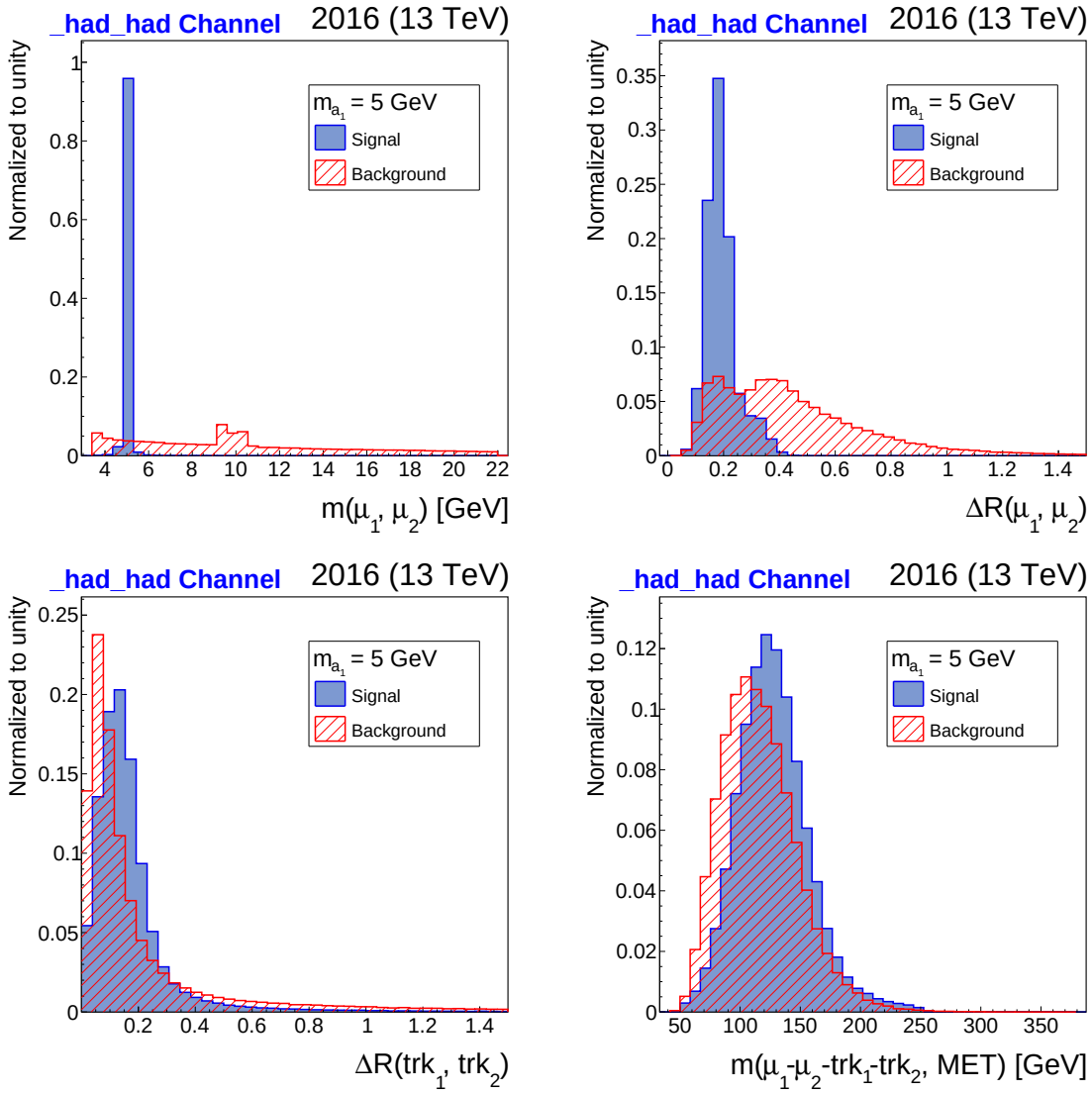


Figure 7.9: Signal (blue) and background (red) distribution of input variables in the training sample. The signal sample corresponds to $m_{a_1(Z_D)} = 5$ GeV, and the training of the classifier is done for the 2016 analysis in the hadron-hadron category.

(Subsec. 6.4.3) is selected for the boosting, with a learning rate of 0.5 and Gini index as separation criterion for the node splitting. The number of grid points used for finding the optimal cut in the node splitting is 20, while the minimum percentage of training events required in a leaf node is set to 5%. The signal samples generated for the training and testing have a size comparable to the number of observed events in the control region selected for the modeling of the background (Tab. 7.9). The events for the training and the testing are selected randomly from the source trees, and the training events are normalized to make the event weights of the signal and background classes equal to the number of events N_s and N_b. Thus, the overall renormalization of event-by-event weights in the training has an average weight of one per event, for both the signal and background.

7.6.3 Overtraining check

The overtraining phenomenon occurs when the classifier has too many degrees of freedom for the available number of training events. Due to the large number of nodes, boosted decision trees become often overtrained, with the overtraining degrading the performance by fitting the statistical fluctuations in the training sample. A check to detect the overtraining and measure its impact can be done by comparing the performance of the classifier in the training sample and in an additional test sample. If the classification performance is found to be deteriorated in the test sample with respect to the training sample, the BDT is said to be overtrained (Subsec. 6.4.4). Fig. 7.10 shows the training and testing samples, corresponding to the 2016 analysis in the hadron-hadron category, for both the signal ($m_{a_1(Z_D)} = 5$ GeV) and background classes superimposed.

7.6.4 Linear correlation coefficients

The input variables that show a high correlation constitute a source of information redundancy for the BDT. Linear correlation coefficients, with a range between $+1$ and -1, provide a first estimate of the shared information between pairs of variables, measuring the strength of the linear relation between them. Two variables are said to be linearly uncorrelated if the corresponding correlation coefficient is 0, positively correlated if it is higher than 0, and negatively correlated if it is lower than 0. The variables that serve as input to the BDT in the analysis are assumed to be uncorrelated for the construction of the signal pdf. The first check on this assumption comes from the inspection of the linear correlation coefficients between pairs of the four input variables. The linear correlation coefficients corresponding to the training of the classifier for the 2016, 2017, and 2018 analyses were computed with TMVA and found to be less than 15% for the three categories and all mass points, as shown in Fig. 7.11 for $m_{a_1(Z_D)} = 5$ GeV in the hadron-hadron category of the 2016 analysis. The lepton-hadron and hadron-hadron categories are shown in Fig. A.2 and Fig. A.3 in Appendix A. A further evaluation of the effect of this low correlations in the modeling of the signal is presented in Subsec. 7.8.2.

7.7 Background modeling

As discussed in Subsec. 7.3.5, the assessment of the background composition in the final selected sample becomes difficult due to limited statistics in the background MC samples.

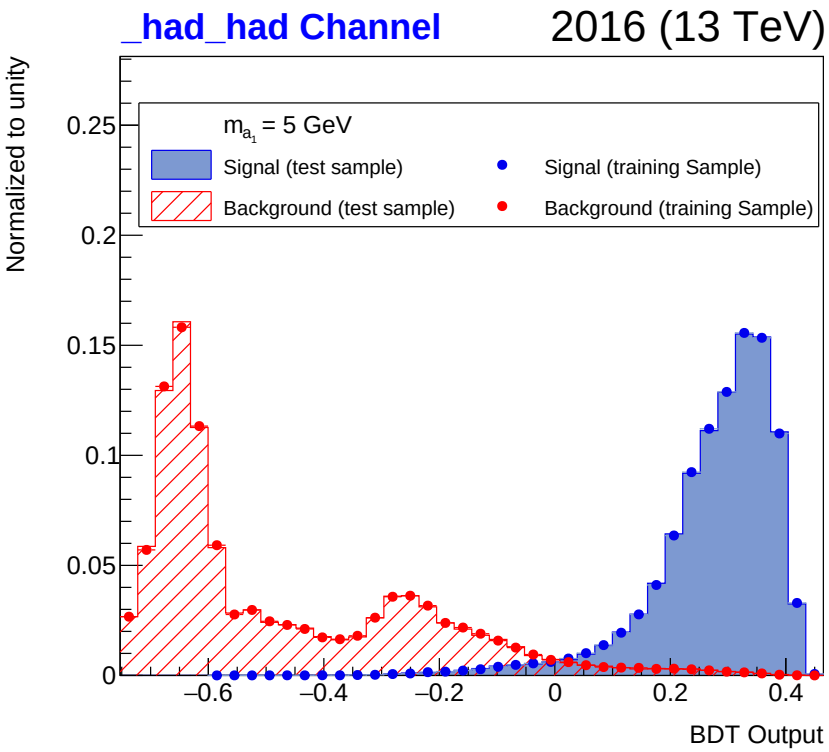


Figure 7.10: Signal and background distributions for the trained classifier. The test (histograms) and training (points with error bars) samples are superimposed for the signal and background classes, to check the overtraining of the classifier. The signal sample corresponds to $m_{a_1(Z_D)} = 5$ GeV and the training of the classifier is done for the 2016 analysis in the hadron-hadron category.

Hence, the background contribution in the signal region is estimated with a control region in data, in which the isolation requirement on the two muons and the two tracks is relaxed. The four objects are allowed to have any number of tracks within a ΔR cone of 0.2 and at least one of the muons or one of the tracks is required to be anti-isolated, to ensure that the control region, referred hereafter as NNNN, is orthogonal to the signal region. The expected signal yield assuming $\mathcal{B}(H(125) \to a_1 a_1(Z_D Z_D) \to 2\mu 2\tau) = 0.1\%$ is comparable with the statistical uncertainty on the number of selected data events. This control region consists of low mass DY, $t\bar{t}$, and QCD events. The QCD contribution includes the quarkonium states Ψ', $\Upsilon(1s)$, $\Upsilon(2s)$, and $\Upsilon(3s)$.

7.7.1 Validation regions

The control regions defined for the validation of the background model are known as *validation regions (VRs)*. The term distinguishes the set of control regions that are not used to determine the background model or constrain the background normalization, but to validate the model itself. For the validation of the background model of the analysis, three validation regions are defined. In two of these regions, the isolation requirement imposed on the dimuon and/or the ditrack pair is relaxed. The muons of the dimuon system and/or the tracks of the

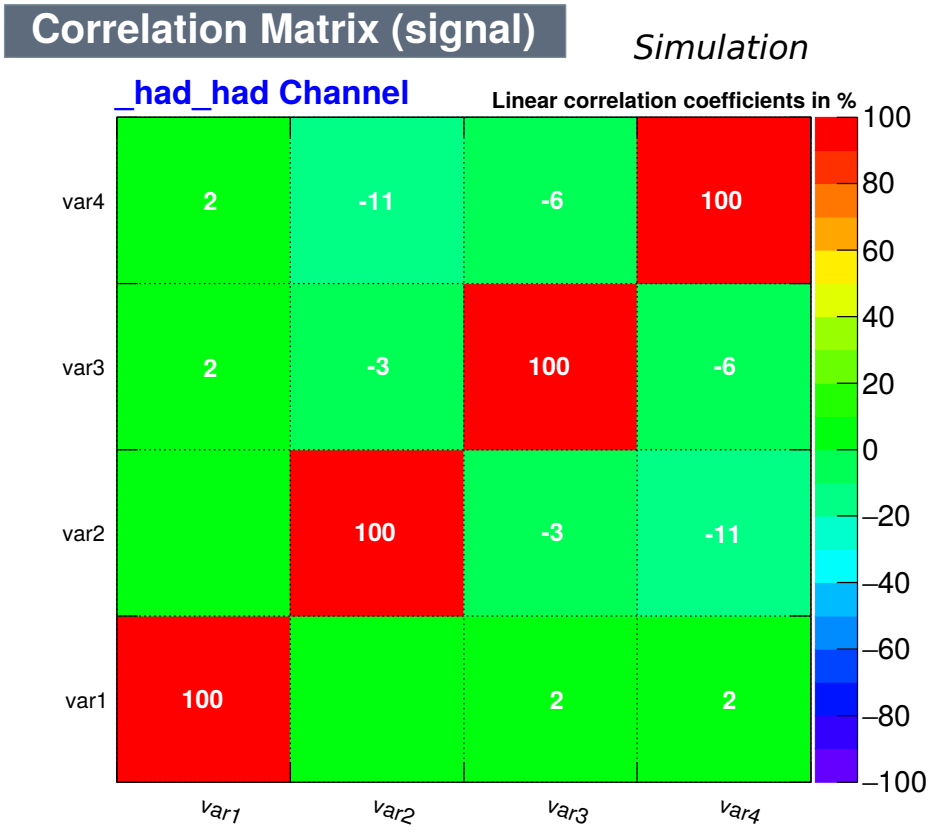


Figure 7.11: Matrix of linear correlation coefficients for the variables used in the BDT training: m_{μ_1,μ_2} (var1), $\Delta R(\mu_1, \mu_2)$ (var2), $\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$ (var3), and $m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$ (var4), for $m_{a_1(Z_D)} = 5$ GeV and the 2016 dataset in the hadron-hadron category.

ditrack system are allowed to have one or two *soft tracks* (tracks with $p_T > 1$ GeV, $d_{xy} > 1$ cm, and $d_z > 1$ cm) in a ΔR cone of 0.2. The two VRs are defined as follows:
Validation region Soft-Iso: The muons of the dimuon pair and the tracks of the ditrack pair are allowed to have one or two soft tracks within the ΔR cone of 0.2.
Validation region 00-Soft-Iso: The muons of the dimuon pair fulfill the isolation requirement of the signal region while the tracks of the ditrack pair are allowed to have one or two soft tracks within the ΔR cone of 0.2.
The third validation region (**Validation region Same-Sign**), contains events that pass the full signal selection of the main analysis, except that the opposite-sign requirement of the ditrack pair is inverted, i.e., the tracks of the ditrack system are required to be same sign. This validation region results in the highest difference in shape with respect to the CR NNNN from the three considered VRs. Therefore, it is used to estimate the shape uncertainty for the background in the signal region.
Tab. 7.8 summarizes the isolation requirement and the use given in the analysis to each of the three VRs, together with the CR NNNN. Tab. 7.9 reports the number of data events selected in each of the regions per category and per year. The signal contamination was

studied for the benchmark value of the branching fraction considered throughout the analysis ($\mathcal{B}(H(125) \to a_1a_1(Z_DZ_D) \to 2\mu2\tau) = 0.1\%$) and the value of the SM H(125) production cross-section corresponding to the ggF process. For the three VRs, the signal contamination is well below 1% for the hadron-hadron category and below 1% for the lepton-hadron category. In the lepton-lepton category, the signal to background ratio maintains also a low value. Hence, the signal contamination has a negligible effect on the final results.

Table 7.8: Control regions used to construct and validate the background model. The symbol N_{soft} denotes the number of soft tracks within a ΔR cone of 0.2 around the two muons and the two tracks momentum direction.

Control region	μ-μ	trk-trk	Purpose
NNNN	$N_{soft} > 0$	$N_{soft} > 0$	Bkgd model estimation
Soft-Iso	$N_{soft} = 1, 2$	$N_{soft} = 1, 2$	Bkgd model validation
00-Soft-Iso	$N_{soft} = 0$	$N_{soft} = 1, 2$	Bkgd model validation
Same-Sign	$N_{soft} = 0$	$N_{soft} = 0$	Extrapolation uncertainty

Table 7.9: The number of data events in the control regions used to construct and validate the background model.

Sample	Category	Observed events		
		2016	2017	2018
NNNN	lepton-lepton	301	167	293
	lepton-hadron	17407	13729	21202
	hadron-hadron	283693	250800	376039
Soft-Iso	lepton-lepton	29	7	26
	lepton-hadron	1467	1189	1842
	hadron-hadron	17481	15665	23357
00-Soft-Iso	lepton-lepton	23	5	18
	lepton-hadron	1135	888	1356
	hadron-hadron	14289	12437	18647
Same-Sign	lepton-lepton	43	29	44
	lepton-hadron	2466	1959	3024
	hadron-hadron	20724	16961	25703

7.7.2 Data-driven closure test

The validation of the background model is done through a data-driven *closure test*, in which the BDT output distribution for the background from CR NNNN is compared with the corresponding shapes in the VRs Soft-Iso and 00-Soft-Iso per category and per year, as one gradually approaches to the signal region. The high statistics of the CR NNNN allows to perform such a closure test, even though the VRs are not orthogonal with respect to NNNN, with the overlap representing less than 10% of the events of the CR NNNN, for each category, and for each year. An additional verification comes from the validation with the Same-Sign region, which is orthogonal to the CR NNNN, due to the same-sign requirement for the ditrack pair. Fig. 7.12 (Fig. 7.13) presents the comparison of the BDT output distribution between the CR NNNN and Soft-Iso (00-Soft-Iso) for the 2018 analysis. A good agreement is observed within the uncertainties. Similar distributions for the 2016 and 2017 analyses are shown in Figs. A.4 to A.7 in Appendix A.

7.8 Signal modeling

The modeling of the signal in the analysis is based on the construction of multidimensional pdfs out of the individual pdfs associated to each of the BDT input variables. For each of the available signal samples, a multidimensional probability density function is defined:

$$f_{\text{tot}} = f_{m_{\mu_1,\mu_2}} \cdot f_{\Delta R(\mu_1,\mu_2)} \cdot f_{\Delta R(\text{trk}_1,\text{trk}_2)} \cdot f_{m_{\mu_1,\mu_2,\text{trk}_1,\text{trk}_2,\text{MET}}}. \tag{7.17}$$

The construction of the multidimensional pdf as the product of the independent pdfs corresponding to each of the four BDT input variables is based on the assumptions of a low and non-significant correlation between the input variables. Checks on the validity of these assumptions are given in Subsec. 7.6.4 and Subsec. 7.8.2. The value of the parameters of the four multidimensional pdfs which are fixed, were determined by fitting the distribution of a reference mass point, leaving the parameters freely floating in this individual fit.
The dimuon mass distribution of the $H(125) \to a_1a_1(Z_DZ_D) \to 2\mu2\tau$ signal is parameterized with a Voigtian, a convolution of a Breit-Wigner [BW] and a Gaussian [G] [x= m_{μ_1,μ_2}], as shown in Eq. (7.18). The parameters of the pdf are the width, mean, and sigma of the Voigtian. The Breit-Wigner and the Gaussian have the same mean, while the width and the sigma of the Voigtian correspond to the width and the sigma of the Breit-Wigner and the Gaussian, respectively.

$$\text{pdf}(\bar{x}_V, \Gamma, \sigma_V) = \text{BW}(x; \bar{x}_V, \Gamma) * \text{G}(x; \bar{x}_V, \sigma_V). \tag{7.18}$$

The Gaussian modifies the Breit-Wigner-like signal by introducing the effect of the detector response. For all the variables parametrized with a convolution, the introduction of the Gaussian smearing accounts for electronic noise contributions and Gaussian components in the energy deposition. Therefore, the convolution expresses how the shape of the Breit-Wigner is modified by the Gaussian, resulting in an accurate function for the fit. The Breit-Wigner and the Gaussian take the form:

$$\text{BW}(x; \bar{x}_V, \Gamma) = \frac{1}{(x-\bar{x}_V)^2 + (\frac{\Gamma}{2})^2}, \;\; G(x; \bar{x}_V, \sigma_V) = \frac{1}{\sqrt{2\pi\sigma_V^2}} \times \exp[-(x-\bar{x}_V)^2/2\sigma_V^2]. \tag{7.19}$$

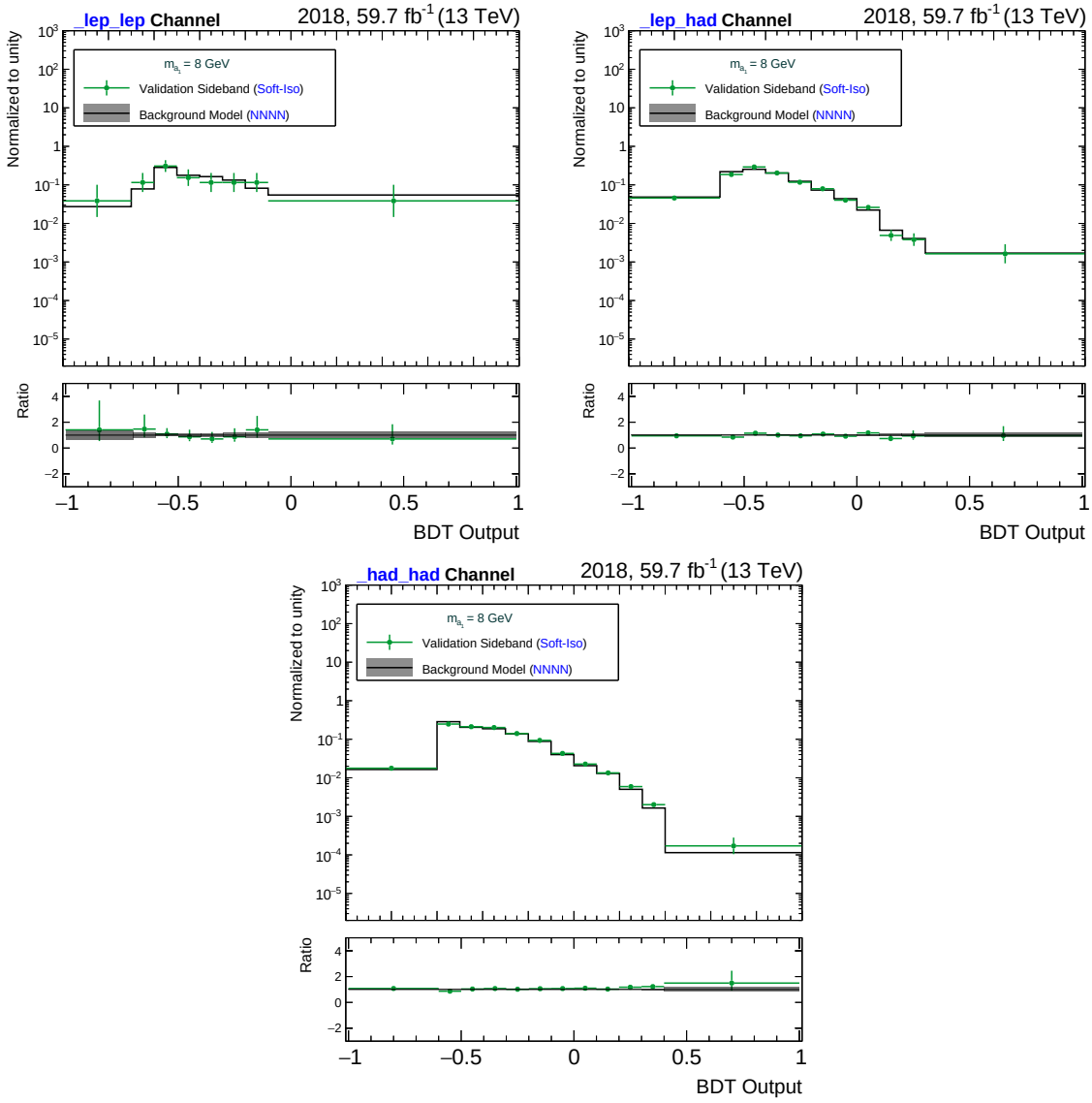


Figure 7.12: Data-driven closure test of background model. The BDT output distribution for events passing the signal selection in the validation region Soft-Iso is compared with the shape in CR NNNN, for $m_{a_1(Z_D)} = 8$ GeV and the 2018 dataset in the lepton-lepton (upper left), lepton-hadron (upper right), and hadron-hadron (lower) categories. The lower panel shows the ratio of the distribution observed in Soft-Iso to the distribution observed in NNNN for the corresponding category.

The fitted distribution of m_{μ_1,μ_2} for $m_{a_1(Z_D)} = 10$ GeV in the hadron-hadron category, corresponding to the 2018 analysis, is shown in the upper left plot of Fig. 7.14.
The $\Delta R(\mu_1, \mu_2)$ is modeled through a composite pdf, formed by the addition of two Gaussians [G_1 and G_2] [x= $\Delta R(\mu_1, \mu_2)$]:

$$\text{pdf}(\bar{x}_1, \sigma_1, \bar{x}_2, \sigma_2, \text{frac}_1) = G_1(x; \bar{x}_1, \sigma_1) + \text{frac}_1 \cdot G_2(x; \bar{x}_2, \sigma_2), \tag{7.20}$$

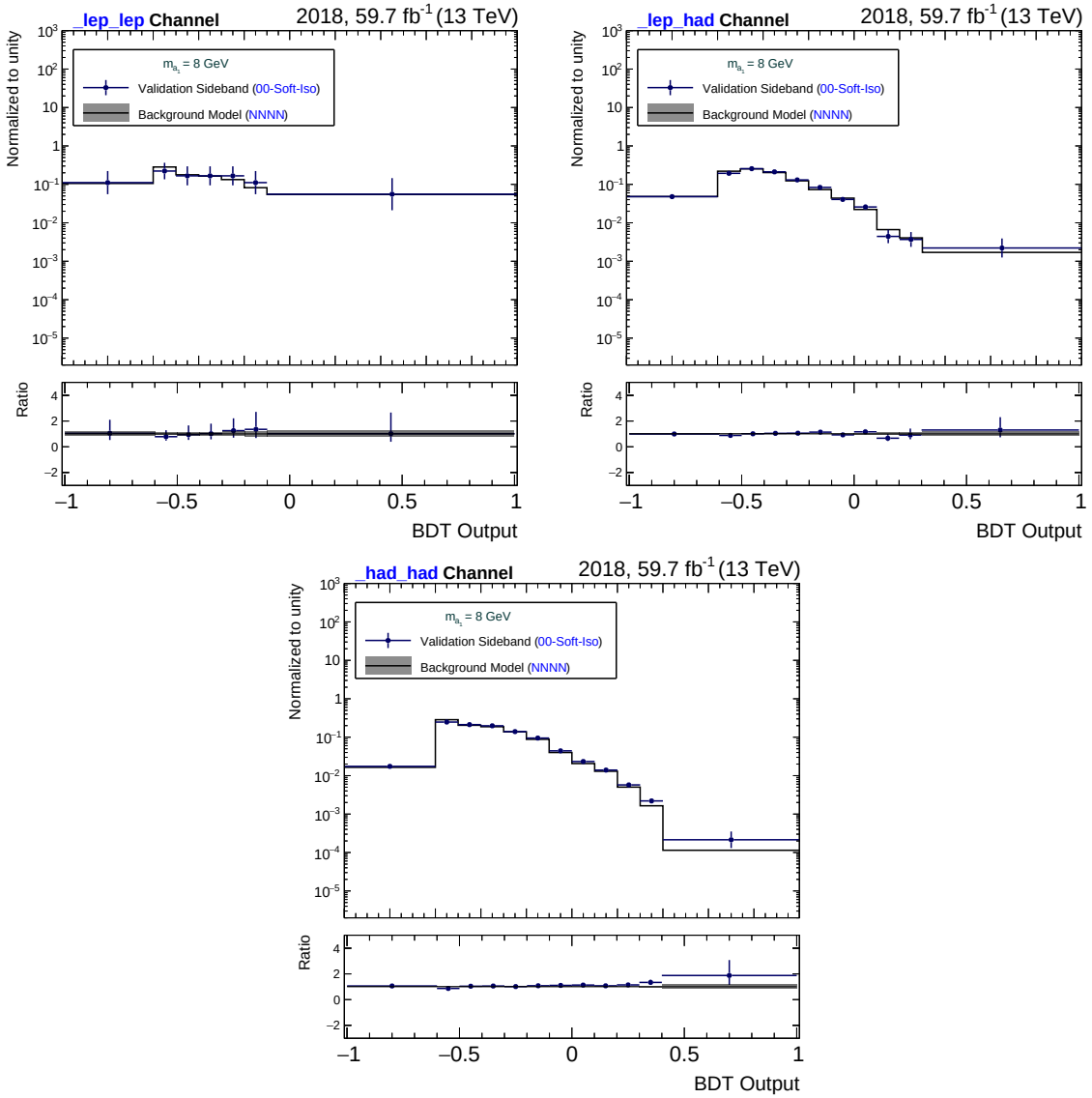


Figure 7.13: Data-driven closure test of background model. The BDT output distribution for events passing the signal selection in the validation region 00-Soft-Iso is compared with the shape in CR NNNN, for $m_{a_1(Z_D)} = 8$ GeV and the 2018 dataset in the lepton-lepton (upper left), lepton-hadron (upper right), and hadron-hadron (lower) categories. The lower panel shows the ratio of the distribution observed in 00-Soft-Iso to the distribution observed in NNNN for the corresponding category.

where the Gaussians take the form:

$$G_1(x; \bar{x}_1, \sigma_1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \times \exp[-(x - \bar{x}_1)^2 / 2\sigma_1^2], \tag{7.21}$$

$$G_2(x; \bar{x}_2, \sigma_2) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \times \exp[-(x - \bar{x}_2)^2 / 2\sigma_2^2].$$

The parameters of the composite pdf are the mean and sigma of the Gaussians and the fraction, which represents the proportion of the second Gaussian in the signal. The fitted distribution of $\Delta R(\mu_1, \mu_2)$ for the reference mass point and category is shown in the upper right plot of Fig. 7.14.
The $\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$ is also modeled with a composite pdf, a bit more elaborated with respect to the pdf for $\Delta R(\mu_1, \mu_2)$, since the same function needs to fit correctly the different proportion of fakes per category. The term fake is used to refer to the selection of a pair of tracks that do not come from the two tau leptons produced in the decay of one of the light bosons. The pair of tracks may be associated with some underlying event process, and they happen to be misidentified by the offline selection of the analysis as an $a_1(Z_D) \rightarrow \tau_{\text{1-prong}}\tau_{\text{1-prong}}$ candidate. As seen in the lower left plot of Fig. 7.14, the fakes are associated with a lower $\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$ since they tend to come from much higher boosted objects. The convolution of a Landau [L_1] and a Gaussian [G_3] is added to a Landau [L_2] [x=$\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$], resulting in the following composite pdf:

$$\begin{aligned}\mathrm{pdf}(\bar{x}_{L_1}, \sigma_{L_1}, \bar{x}_{G_3}, \sigma_{G_3}, \bar{x}_{L_2}, \sigma_{L_2}, \mathrm{frac}_2) =& L_1(x; \bar{x}_{L_1}, \sigma_{L_1}) * G_3(x; \bar{x}_{G_3}, \sigma_{G_3}) + \\ & \mathrm{frac}_2 \cdot L_2(x; \bar{x}_{L_2}, \sigma_{L_2}),\end{aligned} \tag{7.22}$$

where the Landau and the Gaussian functions have the following functional form:

$$L_1(x; \bar{x}_{L_1}, \sigma_{L_1}) = \frac{1}{\sigma_{L_1}}\phi(\lambda_1), \phi(\lambda_1) = \frac{1}{2\pi i}\int_{c-\infty}^{c+\infty} \exp\left[\lambda_1 s + s\log s\right]ds \quad \lambda_1 = \frac{x - \bar{x}_{L_1}}{\sigma_{L_1}}, \tag{7.23}$$

$$G_3(x; \bar{x}_{G_3}, \sigma_{G_3}) = \frac{1}{\sqrt{2\pi\sigma^2_{G_3}}} \times \exp[-(x - \bar{x}_{G_3})^2/2\sigma^2_{G_3}],$$

$$L_2(x; \bar{x}_{L_2}, \sigma_{L_2}) = \frac{1}{\sigma_{L_2}}\phi(\lambda_2), \quad \phi(\lambda_2) = \frac{1}{2\pi i}\int_{c-\infty}^{c+\infty} \exp\left[\lambda_1 s + s\log s\right]ds \quad \lambda_2 = \frac{x - \bar{x}_{L_2}}{\sigma_{L_2}}.$$

The parameters of the Landau function in the second term, which are found to be independent of the mass of the light boson and the category, are taken as constants. The fraction of the second Landau function is fixed within the mass range of the light boson per category, but is different for each of the three categories. This dependence of the fraction on the category is expected, since the proportion of fakes depends on the identification of the tracks as lepton or hadrons. The fraction of fakes reaches its minimum value ($\sim$0) in the lepton-lepton category and its maximum value in the hadron-hadron category ($\sim$0.25).
The $m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$ invariant mass is parametrized with the convolution of a Landau [L_3] and a Gaussian [G_4] [x= $m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$]:

$$\mathrm{pdf}(\bar{x}_{L_3}, \sigma_{L_3}, \bar{x}_{G_4}, \sigma_{G_4}) = L_3(x; \bar{x}_{L_3}, \sigma_{L_3}) * G_4(x; \bar{x}_{G_4}, \sigma_{G_4}), \tag{7.24}$$

where:

$$L_3(x; \bar{x}_{L_3}, \sigma_{L_3}) = \frac{1}{\sigma_{L_3}}\phi(\lambda_1), \ \phi(\lambda_1) = \frac{1}{2\pi i}\int_{c-\infty}^{c+\infty} \exp\left[\lambda_1 s + s\log s\right]ds \ \lambda_1 = \frac{x - \bar{x}_{L_3}}{\sigma_{L_3}}, \tag{7.25}$$

$$G_4(x; \bar{x}_{G_4}, \sigma_{G_4}) = \frac{1}{\sqrt{2\pi\sigma^2_{G_4}}} \times \exp[-(x - \bar{x}_{G_4})^2/2\sigma^2_{G_4}].$$

From the four parameters of the Landau and the Gaussian functions only the mean of the Gaussian is fixed, and the additional three parameters are found to be quite stable within one category for the entire mass range of the light boson. The lower right plot of Fig. 7.14 shows the representative fit result for $m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$.
Once the one-dimensional pdfs for each of the four input variables are constructed, the multidimensional pdf can be obtained for each mass point, for each category, and for each year. The division per category reduces the existing correlations between the input variables and improves the quality of the signal modeling, resulting in an increase on the classification performance. Each of the multidimensional pdfs consists of 19 parameters, 3 corresponding to m_{μ_1,μ_2}, 5 to $\Delta R(\mu_1, \mu_2)$, 7 to $\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$, and 4 to $m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$. The value of the 19 parameters are obtained by fitting the distribution of the variables in the signal simulated samples. The signal acceptance tends to drop for higher masses of the light boson, but the statistics is still enough to obtain good quality fits. The goodness of the fits are assessed through χ^2 tests, shown together with the fitted distributions. The value of the χ^2 square divided by the number of degrees of freedom ($ndof$), where ndof corresponds to the number of data points minus the number of parameters of the fit, is reported. Parametrized distributions of the additional two categories for $m_{a_1(Z_D)} = 10$ GeV, and for the three categories of a lower and a higher mass point are shown in Figs. A.8 to A.15 in Appendix A.

7.8.1 Signal interpolation method

In the analysis, the response model for the parameters of the signal corresponds to the simple case of linear interpolation, i.e., the 19 parameters of the multidimensional pdf show a linear dependence with respect to the mass of the light boson. The limited number of available Monte Carlo samples within the probed mass range of the light boson makes necessary to perform an interpolation procedure, in order to obtain the signal pdfs for intermediate-mass hypotheses. The parameters of the pdfs for the intermediate-mass points are obtained through a cubic spline [211] of each of the 19 fitted parameters. Four representative fitted parameters that serve as input fits to the signal spline procedure are depicted in Fig. 7.15, to illustrate the behavior of the parameters as a function of the mass of the light boson. Once the multidimensional pdfs for each category, for each year, and for each mass point are obtained, signal samples with a 0.2 GeV step can be generated. For the generation of the samples from the signal pdfs, a signal acceptance per category needs to be associated to each mass point. The information on the signal acceptance is available for the official mass points, while for the intermediate-mass hypotheses, a value is associated per category through a cubic spline interpolation.
The main advantage of the method comes from the possibility to generate training and testing samples for the signal region of much larger size compared to the ones existing in the nominal Monte Carlo samples, which do not have enough statistics to carry out an efficient training of the BDT. The signal interpolation method, together with an individualized training of the BDT for each mass point and each category, allows us to reach the maximum separation power.

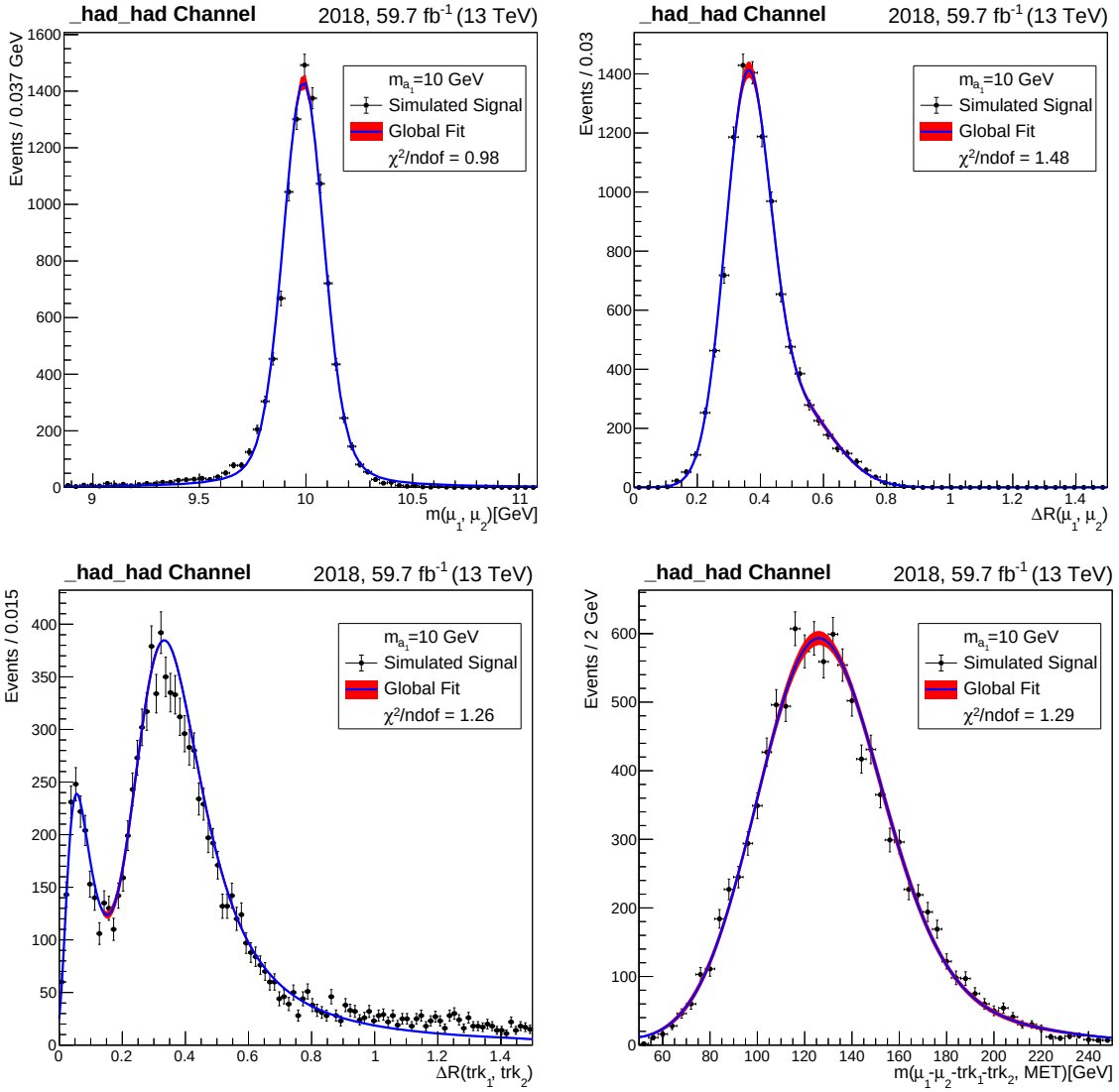


Figure 7.14: Parameterized distribution of the four BDT input variables: m_{μ_1,μ_2} (upper left), $\Delta R(\mu_1, \mu_2)$ (upper right), $\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$ (lower left), and $m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$ (lower right). The shown distributions correspond to $m_{a_1(Z_D)} = 10$ GeV in the hadron-hadron category for the 2018 analysis. The red shadow represents the fit error band within 2-sigma of confidence interval.

7.8.2 Validation of signal model

Two independent validations of the signal model presented in the previous section are carried out. The first validation assesses the quality of the interpolation by removing from the parametric fits of the fitted parameters the values of the 19 parameters corresponding to one mass point and predicting these values using the data from the rest of the mass points. The comparison of the predicted and observed values provides useful information on the robustness and stability of the proposed model for the construction of the multidimensional pdfs. The procedure is repeated for all mass points, removing the corresponding set of 19 param-

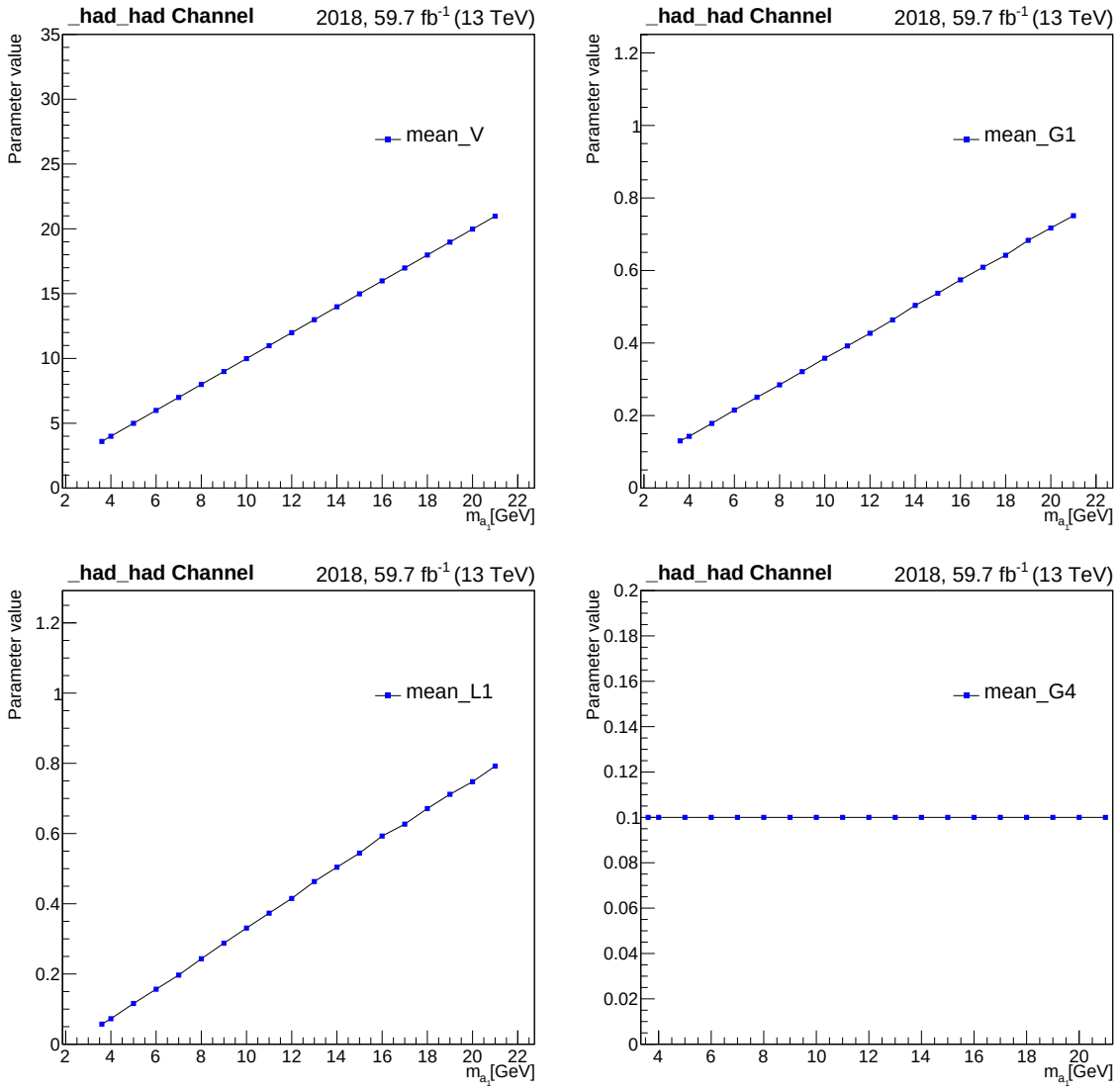


Figure 7.15: Fitted parameters for the signal spline procedure: mean of the Voigtian V from pdf$(\bar{x}_V, \Gamma, \sigma_V)$ (upper left), mean of the Gaussian G_1 from pdf$(\bar{x}_1, \sigma_1, \bar{x}_2, \sigma_2, \text{frac}_1)$ (upper right), mean of the Landau L_1 from pdf$(\bar{x}_{L_1}, \sigma_{L_1}, \bar{x}_{G_3}, \sigma_{G_3}, \bar{x}_{L_2}, \sigma_{L_2}, \text{frac}_2)$ (lower left), and mean of the Gaussian G_4 from pdf$(\bar{x}_{L_3}, \sigma_{L_3}, \bar{x}_{G_4}, \sigma_{G_4})$ (lower right), corresponding to the hadron-hadron category in the 2018 analysis.

eters one at a time and predicting the associated pdf. This first validation allows us to test the ability of the interpolation procedure to recreate the shape of the BDT input variables from the real sample.

Fig. 7.16 compares three of the four one-dimensional pdfs, constructed by fitting the distribution of the variables in the official signal sample with the corresponding pdfs obtained following the procedure for the intermediate-mass hypotheses described in the previous section. The actual and interpolated templates are found to be compatible within the MC statistical uncertainties, and the chi-square values do not deteriorate significantly with the

loss of information in the interpolated case. For the case of the fourth variable in the BDT ranking ($m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$) the interpolated and the real signal sample completely superimpose, leading to a same value for the chi-square test. This results from having a quite stable behavior of the parameters for the entire mass range of the light boson within one category for this variable. A first check on the assumption of a low correlation between m_{μ_1,μ_2}, $\Delta R(\mu_1,\mu_2)$, $\Delta R(\mathrm{trk}_1,\mathrm{trk}_2)$, and $m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$ was presented in Subsec. 7.6.4, in which the correlation coefficients between pairs of the variables were studied. Nevertheless, the presence of additional non-linear correlations is not discarded with this first check. If the correlations between input variables are ignored during the training step, the algorithm can suffer from a performance loss since the classifier's training might not be optimal. This can be fixed with pre-processing methods of decorrelation, applied to the input variables before the training. The decorrelation via the square-root of the covariance matrix or through a principal component analysis (a linear transformation that rotates the set of data points for a maximum visible variability) constitute examples of pre-processing methods used to reduce linear correlations. However, the correlations can not be completely removed by the decorrelation procedures, especially in the case of complex non-linear correlations, leaving space to some model inaccuracies. Even when the same BDT classifier is applied to a real sample (with the real correlation embedded) and a generated sample (produced assuming uncorrelated variables), a difference in the BDT response is observed. The effect of the embedded correlation results in the different BDT output distribution observed when compared to the one obtained in the case in which a multidimensional pdf, constructed as the product of individual one-dimensional pdfs, is considered. Therefore, if the effect of the correlations between the variables is not negligible and they are not considered in the construction of the multidimensional pdf, the model would not represent the real sample and would not produce a similar BDT output distribution.
Fig. 7.17 shows the comparison of the BDT output distributions obtained for the interpolated and the real signal sample and the results of the Kolmogorov-Smirnov (KS) test [212, 213], corresponding to $m_{a_1(Z_D)} = 15$ GeV in the 2018 analysis, for each of the three categories. The uncertainty on the parameters of the multidimensional pdf, obtained from the fitting procedure for the modeling of the signal, is propagated to the uncertainty on the bins of the BDT output distribution for the signal model. Each of the fitted parameters of the multidimensional pdf is shifted up and down by their corresponding fit uncertainty, and the net effect per bin of the up and down variation of all the parameters is assigned as uncertainty for the given bin. The KS test reports the maximum difference observed between the two distributions, and calculates a p-value from that and the size of the samples. From the results of the KS test it can be concluded that no significant discrepancies between the BDT responses are observed. Thus, the assumption of the negligible effect of the correlation between the four variables used to construct the multidimensional pdf is validated.

7.9 Binned shape analysis

The analysis presented in this work constitutes a *binned shape analysis*. The data and the expectations from the signal and background are provided as templates with the same binning to the statistical analysis, which is mathematically equivalent to performing N (number

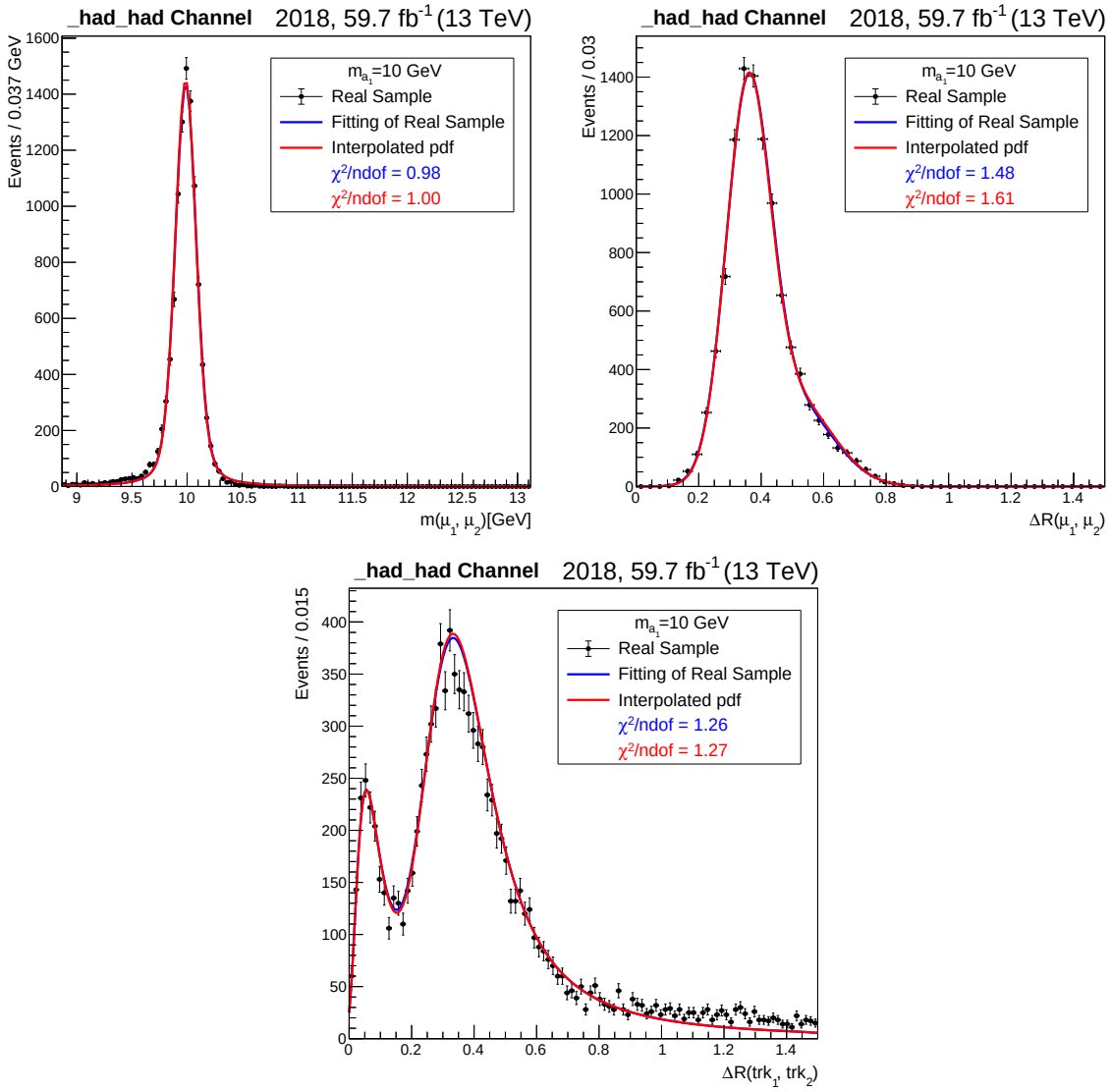


Figure 7.16: Parameterized distribution of three of the four BDT input variables for the interpolated and the real signal sample: m_{μ_1,μ_2} (upper left), $\Delta R(\mu_1, \mu_2)$ (upper right), and $\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$ (lower). The shown distributions correspond to $m_{a_1(Z_D)} = 10$ GeV in the hadron-hadron category for the 2018 analysis.

of bins) counting experiments (Subsec. 6.3.3), with one counting experiment per bin. The information on the shape allows to have a better discrimination power, compared to a counting experiment, since not only the information on the expected event yields for the signal and the background are exploited, but also the shape of the discriminating variable. For each of the three categories of the analysis, the BDT output distribution is provided for the observed and expected shapes to *Combine tool*, the software used to perform the statistical analysis [214]. The systematic uncertainties considered are discussed in detail in the next section. All the information on the observed events, expected yields, and systematic uncertainties is written into text files called *datacards*. The 9 datacards for a mass point of the light boson,

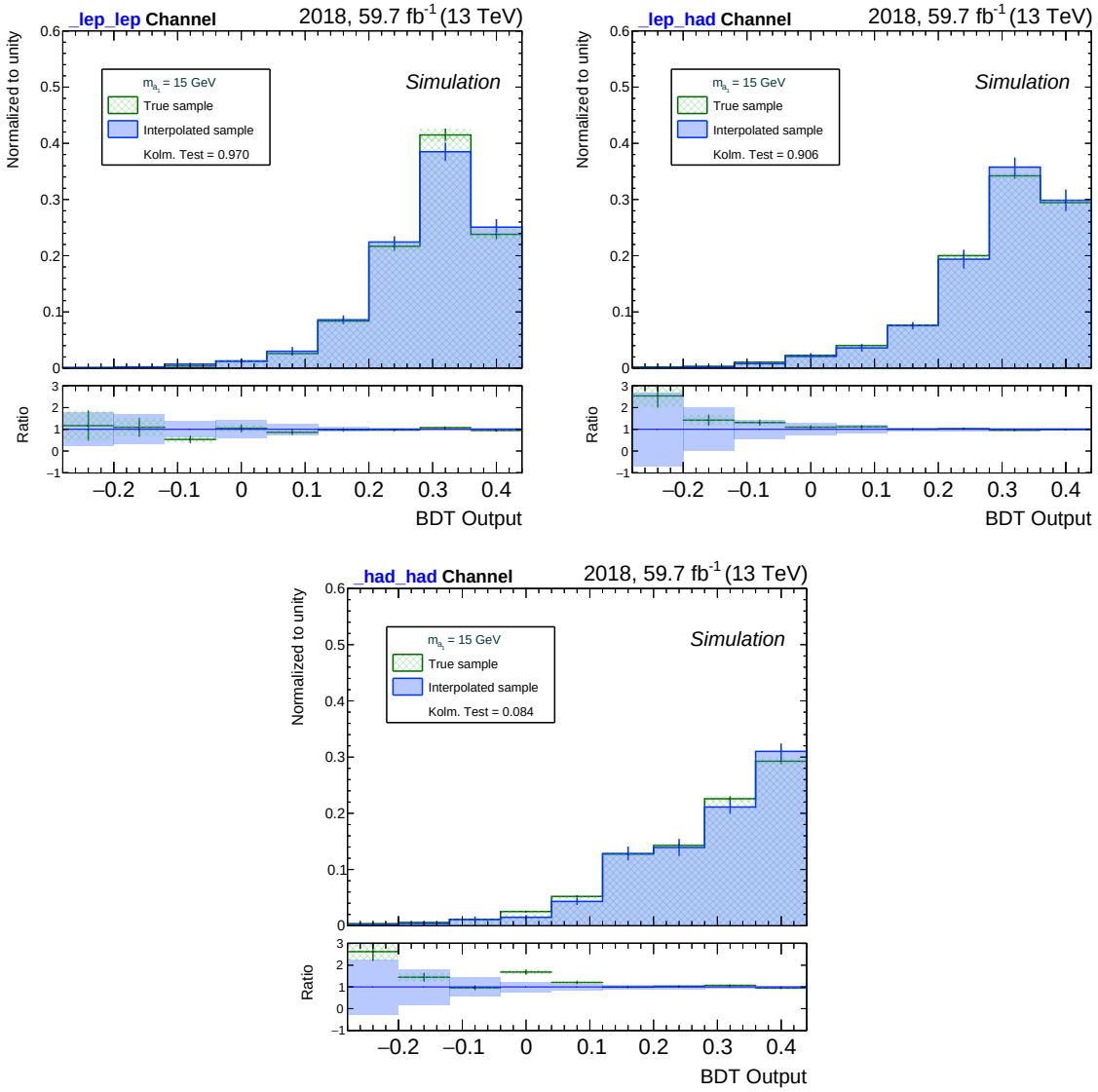


Figure 7.17: BDT output distribution of the interpolated and real signal sample, corresponding to $m_{a_1(Z_D)} = 15$ GeV, for the 2018 analysis in the lepton-lepton (upper left), lepton-hadron (upper right), and hadron-hadron (lower) categories.

which result from three categories per year of analyzed data, are combined into a single datacard. The information in the combined datacard is then converted into a likelihood function, constructed as the product over the bins. For each bin, the likelihood takes the form of Eq. (6.47), being a function of the signal strength modifier μ, defined in Subsec. 6.3.3. The construction of the likelihood constitutes the first step of the statistical analysis. The next step is to determine how different values of μ describe the observed data. The value that maximizes the likelihood, providing the best description of the data, corresponds to the $\hat{\mu}$ of the statistical test in Eq. (6.51). The CLs procedure described in Subsec. 6.3.4 is carried out, obtaining the distribution of upper limits for each mass of the light boson. The results of the CLs procedure performed in the analysis are presented in Subsec. 8.2.1.

7.10 Systematic uncertainties

The term systematic uncertainties comprises all the uncertainties that are not directly related to the statistics of the data, spanning measurement errors that are not caused by statistical fluctuations in real or simulated samples. Badly known detector acceptances, trigger efficiencies, detector resolutions, and background contributions constitute common sources of systematic uncertainties that arise when performing data analysis in the field of high energy physics. Other systematic errors come from incorrect detector calibrations, uncertainties in the simulation models, assumed theoretical models, and external input values such as cross-sections, branching fractions, and the luminosity of the experiment. From the technical machinery used to analyze the data, computational and systematic errors can arise. There can also be certain personal bias for the analysis to have a specific outcome. Furthermore, other unknown effects, which often result difficult to assess, might be present. By inspecting the general picture of the analysis, one can detect the systematic uncertainties coming from each step of the analysis and assess their impact in the final result. Through this check to the general analysis chain, one tries to make sure that every possible source of systematic uncertainty is being considered. In addition, the usual cross-checks of the data/MC agreement, performed from the initial steps of the analysis, allows to early detect inconsistencies and track their source.
Each of the systematic uncertainties in the analysis is modeled through a nuisance parameter θ (Subsec. 6.3.3) that affects the signal or the background, specifically the normalization and/or the shape of the observable distribution. The systematic uncertainties that affect the shape of the templates are shifted up and down by one standard deviation, and they are associated to a nuisance parameter with a unit Gaussian distribution [215]. Bin-wise statistical uncertainties are considered for every bin of every process. In order to prevent the large number of nuisance parameters resulting from creating all the up and down histograms with 1σ uncertainty, the Barlow-Beeston-lite technique is used [216]. The technique consists of assigning a single nuisance parameter with a Gaussian distribution to each bin instead of requiring a separate parameter per process and it is only applied when the number of events in the bin is above a certain threshold. If the number of events is below the threshold, a Poisson pdf is used to model the per-process uncertainties in that bin. The Barlow-Beeston-lite technique allows, when possible according to the number of events per bin, to minimize the number of parameters in the maximum-likelihood fit. Tab. 7.10 summarizes the systematic uncertainties related to the signal and the background considered in the analysis.

7.10.1 Uncertainties related to background

The estimation of the background in the analysis is not affected by imperfections in the simulation of the detector response or inaccuracies in the modeling of the reconstruction since it is based on a data-driven method. The uncertainty on the shape of the background, modeled as described in Sec. 7.7, is dominated by the bin-by-bin statistical uncertainty related to the size of the CR NNNN. An additional shape uncertainty derived in the CR Same-Sign is considered. One nuisance parameter is assigned per category and per year to account for each of these two sources of systematic uncertainty.

Table 7.10: Systematic uncertainties and their effect on the estimates of the background and signal in the analysis. The fourth column indicates if the source of systematic uncertainty is treated as correlated between the years in the fit.

Source	Value	Affected sample	Correlation	Type
Stat. unc. related to size of CR NNNN	-	bkg.	no	bin-by-bin
Extrapolation unc. in CR Same-Sign	-	bkg.	no	shape
Integrated luminosity	2016, 2017, 2018 2.5%, 2.3%, 2.3%	signal	no	norm.
Muon id. and trigger efficiency	2% per muon	signal	no	norm.
Track Iso.	6–9% per track	signal	no	norm.
MC stat. unc. propagated in parameters of the signal pdf (1 nuisance per parameter)	-	signal	no	shape
Theoretical uncertainties in the signal acceptance				
$\mu_{R,F}$ variations ($gg \to H(125)$)	0.8–2%	signal	yes	norm.
PDF ($gg \to H(125)$)	1–2%	signal	yes	norm.
Theoretical uncertainties in the signal cross-section				
$\mu_{R,F}$ variations ($gg \to H(125)$)	5–7%	signal	yes	norm.
PDF ($gg \to H(125)$)	3.1%	signal	yes	norm.

7.10.2 Uncertainties related to signal

The systematic uncertainties affecting the signal yield include the uncertainties related to the integrated luminosity and the trigger efficiency. For the case of the integrated luminosity, a value of 2.5% is assigned to the estimate for 2016 and 2018 ([75, 77]) and 2.3% [76] for 2017. The uncertainty in the muon identification and trigger efficiency (2% per muon) is estimated with the tag-and-probe technique applied to a sample of $Z \to \mu\mu$ events, as described in Subsec. 7.4.2. The 2% uncertainty per muon translates into a 4% systematic uncertainty in the signal acceptance, due to the presence of two muons in the final state. The track identification, isolation, and reconstruction uncertainty, which affects the shape of the signal estimate, amounts to 6 to 9% per track. This uncertainty is assessed through the study of the combined track isolation and one-prong tau decay identification efficiency, performed on a sample of Z bosons decaying into a pair of τ leptons (Subsec. 7.4.3). The muon and track momentum scale uncertainties are smaller than 0.3% and have a negligible effect in the analysis. The statistical uncertainty related to the limited size of the signal MC samples propagates into the parameterization of the signal. This uncertainty is accounted

for by 13 independent nuisance parameters, one for each of the non-fixed parameters used in the construction of the signal model pdf presented in Sec. 7.8.
The theoretical uncertainties affect the kinematic distributions of the Higgs boson, particularly its p_T spectrum, and therefore impact the signal acceptance of the analysis, as discussed in Subsec. 7.4.4. The theoretical uncertainty due to missing higher-order corrections to the gluon-gluon fusion process is estimated with the HQT program, varying the renormalization (μ_r) and factorization (μ_f) scales. The p_T-dependent k-factors are recomputed according to these variations and applied to the simulated signal samples. The impact on the signal acceptance of applying the recomputed k-factors, ranges between 1.2 and 1.5%, depending on the mass of the light boson. Furthermore, the imperfect knowledge of the proton composition has an impact on which parton is asigned a higher probability of initiating a high energy event. This is accounted for by the PDF uncertainties, evaluated with the HqT program. The NNPDF3.0 PDFs [217], used to compute the main k-factors, are varied within their uncertainties and produce a change in the signal acceptance of about 1%. The same exercise is performed with the CTEQ6L1 PDF set [218] and, in this case, the change in signal acceptance is of about 0.7%. Therefore, the variations on the signal acceptance caused by the PDF uncertainties are covered by the assigned uncertainty of 1%. In addition to the impact on the signal acceptance, the scale and pdf uncertainties also affect the cross-section for the various Higgs boson production mechanisms. The values of the theoretical uncertainty assigned to the cross-section due to variations of the normalization and factorization scales, as well as the PDFs, are reported in Tab. 7.10 for the targeted gluon-gluon fusion process. The impact of the nuisance parameters on the signal strength modifier is discussed in Subsec. 8.1.3.

CHAPTER

8

RESULTS

Contents

Throughout this thesis, a description of the theoretical motivation and the physics analysis chain needed to move on from collisions events to the physics result has been given. The data acquisition, calibration, reconstruction, and statistical tools needed to perform the analysis of the data, have been discussed in detail. This chapter is dedicated to present the results of the search for a pair of light bosons in the final state with two muons and two tau leptons, using a dataset collected with the CMS experiment during the Run II data-taking period, at a center-of-mass energy of 13 TeV. The chapter is organized as follows. Sec. 8.1 is dedicated to present the independent results of the 2016, 2017, and 2018 analyses, discussing the impact of the nuisance parameters on the signal strength modifier, the goodness-of-fit test to estimate the compatibility of the observed data with the null hypothesis, and the computation of the exclusion limits on the signal process. Sec. 8.2 is dedicated to present the Run II combination results. Finally, the interpretation of the results in the context of BSM scenarios in which the search in this final state is highly motivated, namely the 2HDM+S and the Dark Photon Model, is given in Sec. 8.3.

8.1 Analysis results with 2016, 2017, and 2018 data

8.1.1 Final discriminant

The existence of the signal is tested through a binned maximum-likelihood fit applied to the BDT classification distribution, with a likelihood function defined per category. The signal templates are derived from simulation, as explained in Sec. 7.8, and the background template is evaluated from data as described in Sec. 7.7. The background distribution is obtained after performing a fit to data under the background-only hypothesis. The systematic uncertainties that affect the yield are represented in the fit by nuisance parameters with a log-normal probability density function (pdf whose logarithm is normally distributed). The shape altering systematic uncertainties are incorporated via nuisance parameters with a Gaussian probability density function, and the bin-by-bin statistical uncertainties are assigned Gamma probability density functions. The fit is performed for each examined mass of the light boson, for each category, and for each year, with the normalization of the signal and the background kept freely floating. Fig. 8.1 shows the BDT output distribution for a mass of the light boson of 10 GeV, corresponding to the three categories of the 2018 analysis. The shown signal yields are computed assuming the SM $H(125)$ boson production cross-section and a branching fraction $\mathcal{B}(H(125) \to a_1 a_1(Z_D Z_D) \to 2\mu 2\tau)$ of 0.1%.

8.1.2 Goodness-of-fit test

The compatibility of the observed data with the background-only hypothesis is assessed through a goodness-of-fit test, based on the saturated model method [219]. Given the null hypothesis and the data, the method consists on constructing a likelihood ratio test with a so-called saturated model as an alternative hypothesis in the denominator, i.e., a model that fits the data exactly. The result of the test would, therefore, constitute a measure of how close is the null hypothesis to the data, described by the saturated model. The observed value of the goodness-of-fit test is compared with the distribution of the goodness-of-fit indicator obtained with an ensemble of Monte Carlo toy experiments (Subsec. 6.3.4), as depicted in Fig. 8.2. The observed data is found to be well described by the background-only hypothesis, with a probability of having in the ensemble of Monte Carlo toy experiments a value of the goodness-of-fit indicator higher than that observed in data of 9%, for a mass of the light boson of 7 GeV.

8.1.3 Impacts and pulls of nuisance parameters

A nuisance parameter is introduced for each systematic uncertainty in the likelihood function, as discussed in Sec. 7.10. The nuisance parameters act as additional constraint terms in the likelihood and model the a-priori knowledge of the parameters, estimated from auxiliary measurements. The constraint of a nuisance parameter is defined as the ratio of the postfit uncertainty to the initial prefit uncertainty. A ratio higher than one would indicate a conservative measurement, with a larger uncertainty than the initial estimation. If the ratio is much smaller than one, the nuisance parameter is said to be over-constrained. This usually happens when the initial variation of the nuisance parameter is large compared to the total statistical uncertainty, which results from the combination of the Poisson statistic fluctuation and the MC statistical uncertainty [220]. If the variation is strongly correlated with the signal shape,

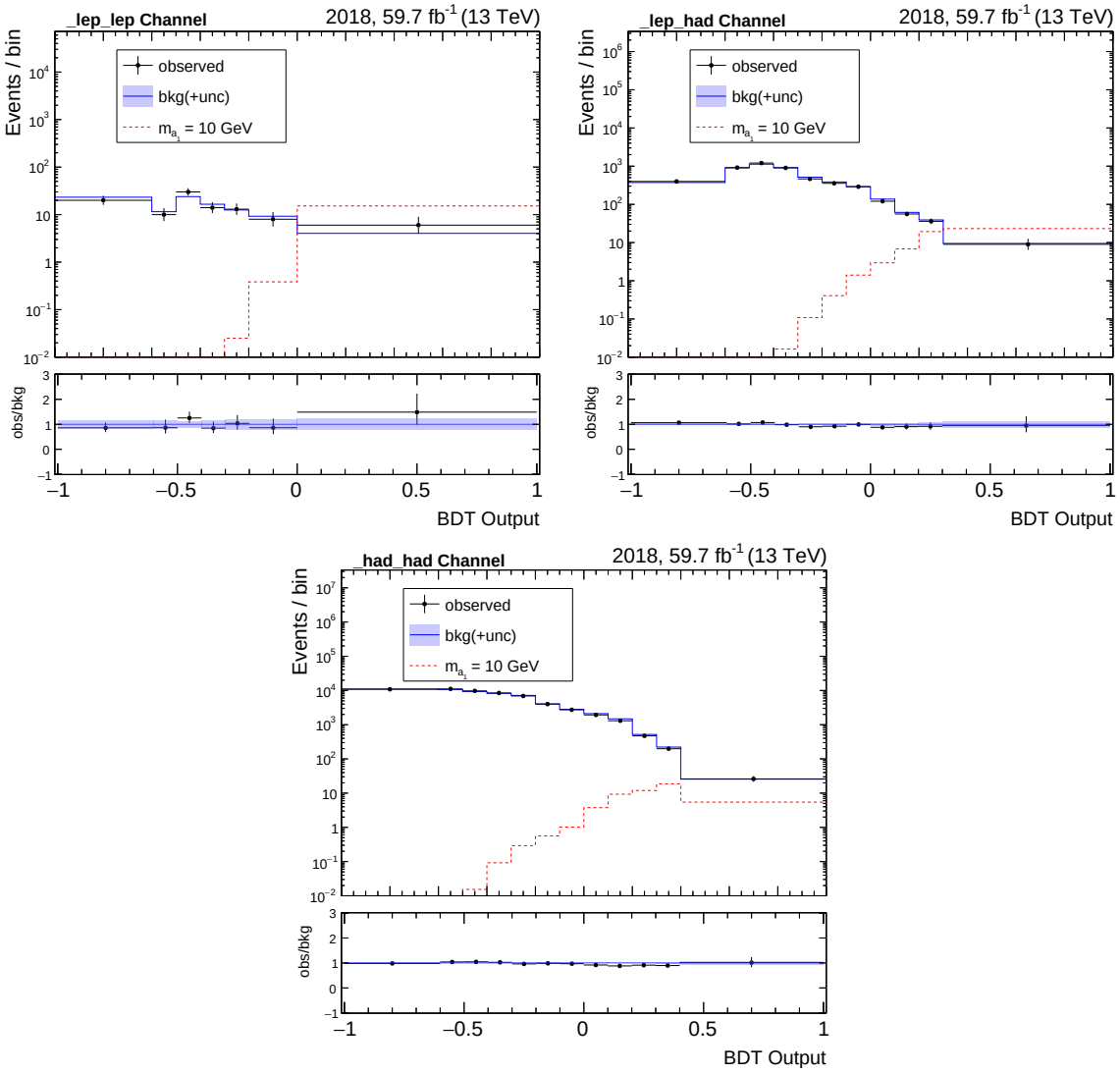


Figure 8.1: The BDT output distribution in the lepton-lepton (upper left), lepton-hadron (upper right), and hadron-hadron (lower) categories corresponding to the 2018 analysis. The shown background distribution is obtained after a fit to data under the background-only hypothesis. The signal expectation for a mass of the light boson of 10 GeV, represented by the dotted histogram, is normalized assuming a branching fraction $\mathcal{B}(H(125) \to a_1 a_1(Z_D Z_D) \to 2\mu 2\tau) = 0.1\%$. The production cross-section of the Higgs boson predicted in the SM is assumed.

the effect on the signal strength modifier will be large. The analysis of the postfit results allows to determine the sources of systematic uncertainty with the largest contribution to the uncertainty of the signal strength modifier. This is particularly relevant for physics analysis where the data statistics is not the main limiting factor, and an improvement of the leading sources of systematic uncertainties can lead to significantly better results. The *impact* of a nuisance parameter, which constitutes a measure of how the signal strength modifier varies

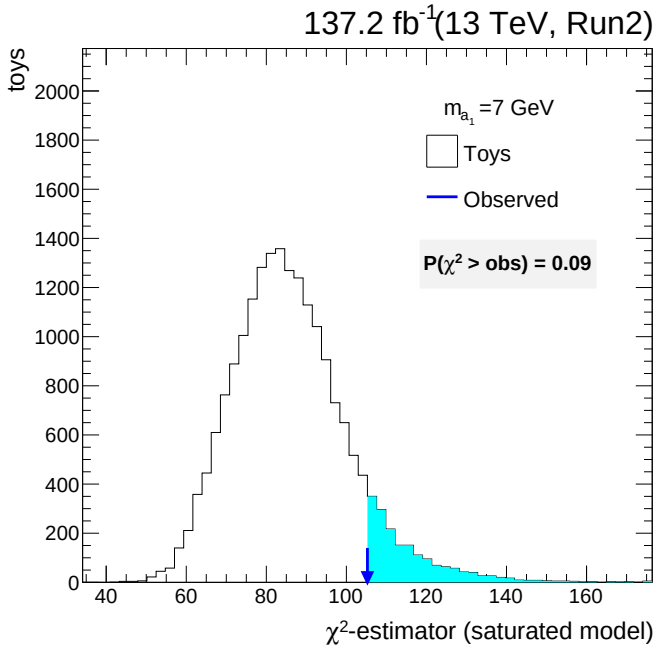


Figure 8.2: Results of goodness-of-fit test under the background-only hypothesis with the saturated model. The observed value of the goodness-of-fit, indicated by the blue arrow, is compared to the distribution of the goodness-of-fit indicator in an ensemble of Monte Carlo toy experiments, for a mass of the light boson of 7 GeV. The blue area represents the corresponding p-value.

as the nuisance parameters are changed one at a time, is defined as:

$$\text{impact}(\theta) = \Delta\hat{r}^{\pm} = \hat{\hat{\mu}}_{\theta_0 \pm \Delta_\theta} - \hat{\mu}, \tag{8.1}$$

where $\hat{\hat{\mu}}$ is the maximum likelihood estimator of μ when all the parameters are profiled except the one for which the impact is determined, and $\Delta\hat{r}$ is the shift induced as θ is considered to have a value $+1\sigma$ or -1σ from their real postfit value, with the rest of the nuisances kept profiled [221]. The direction of the $+1\sigma$ and -1σ impacts indicates whether the nuisance parameter is correlated or anti-correlated with the parameter of interest. The nuisance parameters with a low impact can be discarded or pruned to reduce the complexity of the fitting procedure and, therefore, the fitting time. The *pulls* constitute another relevant quantity to be monitored. For a given nuisance parameter θ, the pull takes the form:

$$\text{pull}(\theta) = \frac{\hat{\theta} - \theta_0}{\Delta\theta}, \tag{8.2}$$

where $\hat{\theta}$, θ_0, and $\Delta\theta$ are the postfit value, the prefit value, and the prefit uncertainty of the nuisance parameter, respectively [222]. The pull quantifies how far a parameter had to be pulled from its prefit value when finding the maximum likelihood estimator. Therefore, a pull average of zero with a standard deviation close to 1 would be the expected good behavior of a nuisance parameter, with further investigations needed if this is not the case.
Fig. 8.3 shows the result of the fit to data under the signal+background hypothesis, for a

mass of the light boson of 10 GeV, corresponding to the Run II combination, in terms of the fitted signal strength, postfit values, and uncertainties of the nuisance parameters and their impacts on the signal strength. The parameters with an error bar smaller than ±1 are the ones constrained by the fit. No nuisance parameter is pulled with respect to the prefit value more than 1.1σ of its prefit uncertainty. The largest effect on the uncertainty of the signal strength modifier is driven by the bin-by-bin statistical uncertainty in the less populated bins of the BDT output distribution.

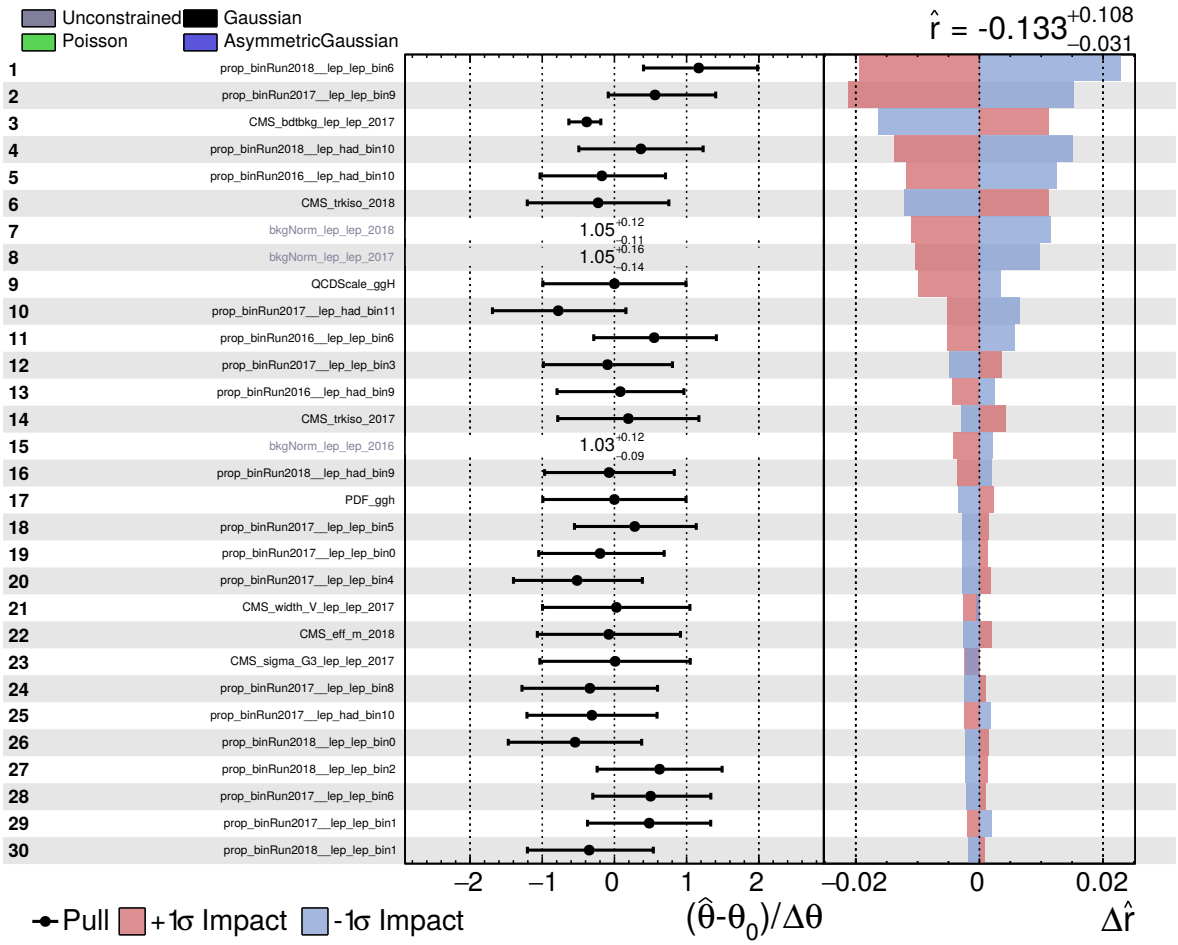


Figure 8.3: Pulls ($(\hat{\theta}-\theta_0)/\Delta\theta$) and impacts ($\Delta\hat{r}$) of the nuisance parameters with the major impact on the fitted value of the signal strength ($\hat{r} = \sigma \times \mathcal{B}/\sigma_{\mathrm{SM}}$), for a mass of the light boson of 10 GeV, corresponding to the Run II combination. The asymmetric error bars in the left panel show the ratio of the postfit to the prefit uncertainty.

8.1.4 Upper limits

Good agreement between the data and the background expectation is found in the BDT output distribution, with no observed evidence of a signal. Therefore, upper limits at 95% CL on $(\sigma_h/\sigma_{SM})\cdot\mathcal{B}(H(125)\to a_1a_1(Z_DZ_D)\to 2\mu 2\tau) = 2\cdot(\sigma_h/\sigma_{SM})\cdot\mathcal{B}(H(125)\to a_1a_1(Z_DZ_D))\cdot\mathcal{B}(a_1a_1(Z_DZ_D)\to\mu\mu)\cdot\mathcal{B}(a_1a_1(Z_DZ_D)\to\tau\tau)$ are set. The modified frequentist CL_s criterion (Subsec. 6.3.4), is used for the computation of the model-independent limits. The term

(σ_h/σ_{SM}) corresponds to the ratio of the Higgs boson cross-section for the gluon fusion production mode, divided by its SM prediction. Fig. 8.4 shows the limits obtained for the 2016, 2017, and 2018 analyses, for a mass of the light boson between 3.6 and 21 GeV.

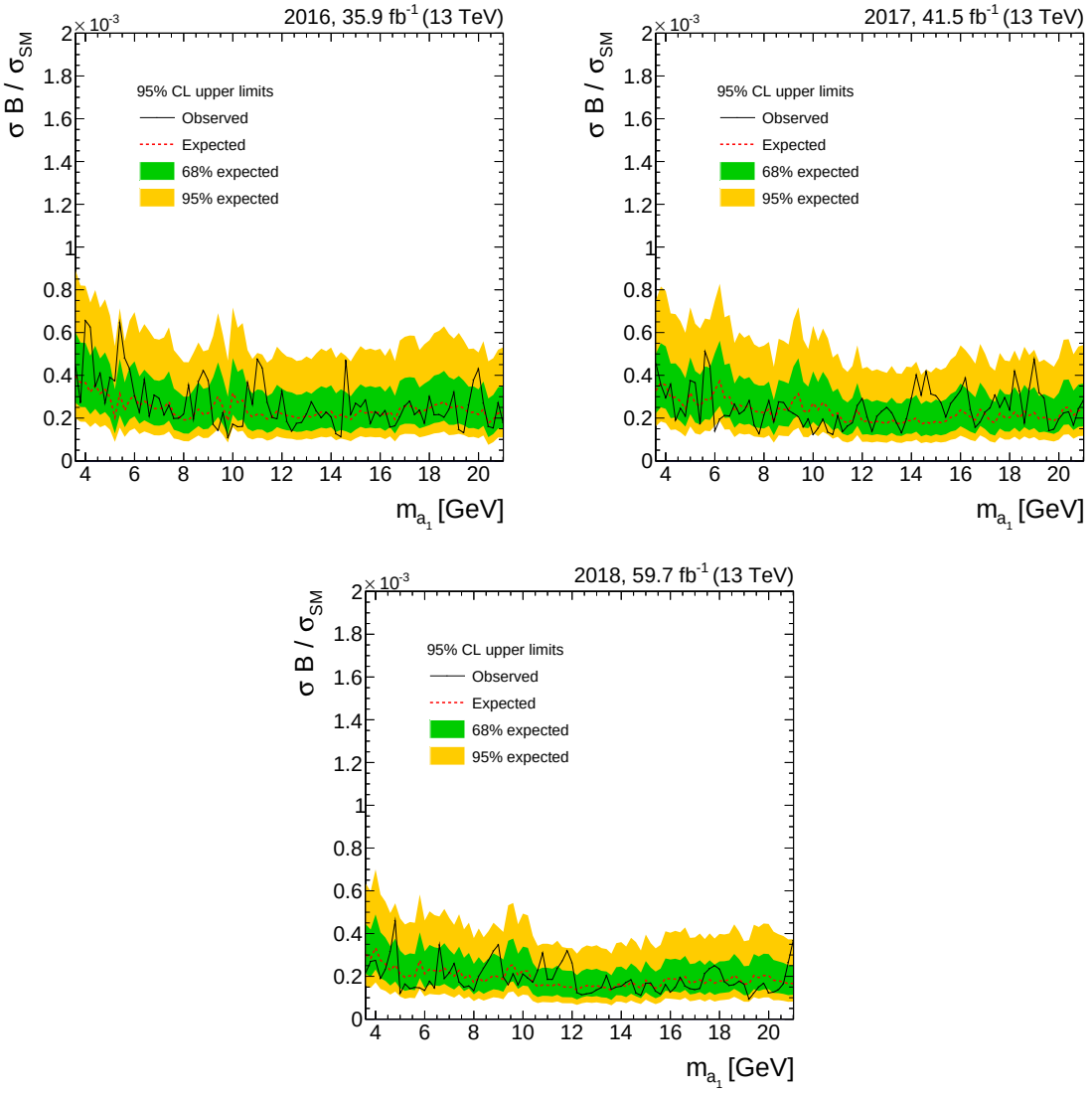


Figure 8.4: Upper limits at 95% confidence level on the signal cross-section times the branching ratio $\sigma(pp \to H(125) + X) \cdot \mathcal{B}(H(125) \to a_1 a_1(Z_D Z_D) \to \mu\mu\tau\tau)$ relative to the inclusive cross-section $\sigma(pp \to H(125)+X)_{\text{SM}}$ predicted in the SM, for the 2016 (upper left), 2017 (upper right), and 2018 (lower) analyses. The bands show the expected 68 and 95% probability intervals around the expected limit.

8.2 Run II combination results

A combination of the three individual analysis results is performed, considering all the statistical uncertainties, systematic uncertainties, and correlations between the years. The statistical procedure used to extract the combined results follows the guidelines of Ref. [[185]],

developed in the context of the Higgs boson search combination by the ATLAS and CMS collaborations back in 2012.

8.2.1 Upper limits

The combined results for the observed and expected limits on the signal cross-section times the branching ratio, relative to the total cross-section of the $H(125)$ boson production as predicted in the SM, are shown in Fig. 8.5. The obtained limits are also summarised in Tab. B.1 in Appendix B. The expected 95% CL limits range from 0.85×10^{-4} at $m_{a_1(Z_D)} = 14.8$ GeV to 1.93×10^{-4} at $m_{a_1(Z_D)} = 4$ GeV and the observed limits range from 0.57×10^{-4} at $m_{a_1(Z_D)} = 12.4$ GeV to 2.29×10^{-4} at $m_{a_1(Z_D)} = 8.8$ GeV. The observed limits are compatible with the expected limits within two standard deviations in the entire tested range of the mass of the light boson.

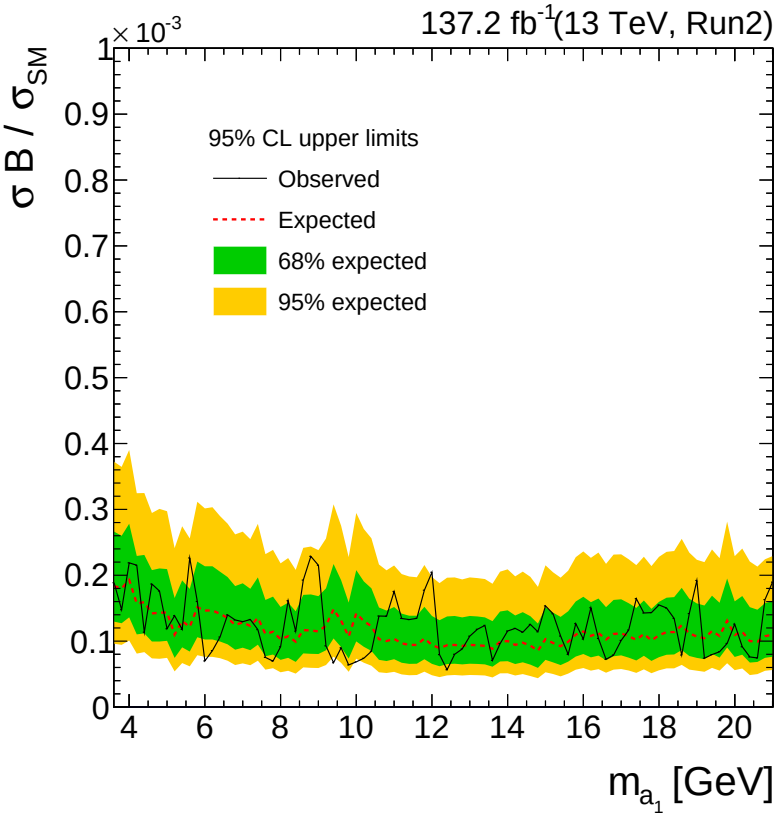


Figure 8.5: Upper limits at 95% confidence level on the signal cross-section times the branching ratio $\sigma(pp \to H(125) + X) \cdot \mathcal{B}(H(125) \to a_1 a_1(Z_D Z_D) \to \mu\mu\tau\tau)$ relative to the inclusive cross-section $\sigma(pp \to H(125) + X)_{\text{SM}}$ predicted in the SM, corresponding to the Run II combination of the results. The bands show the expected 68 and 95% probability intervals around the expected limit.

The most important contribution to the sensitivity of the combined results is provided by the 2018 analysis, owing to the larger data set collected during the 2018 data-taking period. The slight degradation of the sensitivity for low masses of the light boson is caused by a deteriorated signal to background separation of the BDT within the mass range. This is due

to the more difficult discrimination in the presence of the low mass background (Sec. 7.7). The two modulations of the expected limits observed around a mass of the light boson of 10 GeV are caused by the Υ resonances. The first resonance (Υ(1S)) is resolvable, while Υ(2S) and $\Upsilon(3S)$ are merged, resulting in the two observed modulations.
The analysis presented in this thesis provides the tightest constraints on exotic decays of the 125 GeV Higgs boson to a pair of light bosons in the final state with two muons and two τ leptons, for masses of the light boson between 3.6 and 21 GeV. Previous results probing the mass ranges of the light boson: $4 < m_{a_1(Z_D)} < 21$ GeV [55, 223] and $3.6 < m_{a_1(Z_D)} < 21$ GeV [56], are improved by a factor ranging from 4.3 to 9.2 and 1.6 to 4.2, respectively.

8.3 Interpretation of results for benchmark models

The model independent results presented in the previous section can be interpreted in the context of the benchmark models studied in this thesis, namely the 2HDM+S and the Dark Photon Model, as upper limits at 95% CL on $(\sigma_h/\sigma_{\text{SM}}) \cdot \mathcal{B}(h \to a_1 a_1)$ and $(\sigma_h/\sigma_{\text{SM}}) \cdot \mathcal{B}(h \to Z_D Z_D)$, respectively [224].

8.3.1 2HDM+S

The very rich phenomenology of the 2HDM+S, featuring four types of fermion couplings, was described in Subsec. 3.1.1. For a given type of 2HDM+S, assuming the decoupling limit, the phenomenology of the $h \to aa \to x\bar{x}y\bar{y}$ decays is determined by three independent parameters: $\mathcal{B}(h \to aa)$, tanβ, and m_a. Once the model type is specified, the branching fraction of the pseudoscalar to SM particles can be calculated for a given value of m_a and tanβ [42, 225]. The upper limits on the signal strength modifier obtained in a search for decays of the H(125) boson to a pair of light bosons ($h \to aa$), in which no evidence of signal is observed, can be translated into constraints on the free parameters of the model. The upper limits from the experimental search, corresponding to a discrete number of mass points, are first converted into a map $\mu_{\text{up}}(m_a)$. Then, given a specific type of 2HDM+S, a value of tanβ, and m_a, the branching fractions $\mathcal{B}(h \to x\bar{x})$ and $\mathcal{B}(h \to y\bar{y})$ are determined and the experimental limits transfered into upper limits on $\mathcal{B}(h \to aa)$, which take the form:

$$((\sigma_h/\sigma_{\text{SM}}) \cdot \mathcal{B}(h \to aa))_{\text{up}} = \frac{\mu_{\text{up}}(m_a)}{2 \cdot \mathcal{B}(a \to x\bar{x}; m_a, \tan\beta) \cdot \mathcal{B}(a \to y\bar{y}; m_a, \tan\beta)}. \tag{8.3}$$

This equation constitutes the basis of model-dependent interpretations in the context of the 2HDM+S. The procedure allows to map the phase space of tanβ and m_a, resulting in 95% CL upper limits on $\mathcal{B}(h \to aa)$ as a function of these two parameters. The branching fractions for the decay of the light boson into a pair of fermions in the denominator, which constitute theoretical predictions of the model, are provided by theorists as scans of the model parameters. These scans are made available in files containing the values of $a \to x\bar{x}(y\bar{y})$ and tanβ for certain number of mass points, which cover the regions of m_a: 1 GeV $< m_a <$ 70 GeV and tanβ: $0.5 <$ tan$\beta < 10$ [226]. In the analysis, upper limits at 95% CL on $(\sigma_h/\sigma_{\text{SM}}) \cdot \mathcal{B}(h \to a_1 a_1)$ are set for each type of 2HDM+S, as a function of tanβ and m_{a_1}, using Eq. (8.3). A cubic spline interpolation is applied between the mass points for which the upper limits were computed.
The upper limits obtained for selected values of tanβ (higher than one for Type II-III and

lower than one for Type IV) are shown in Fig. 8.6. For the Type I 2HDM+S, a value of tanβ is not specified since, in this case, the branching fraction of the decay of the light boson into a pair of fermions is independent of tanβ. In the case in which the observed limit is greater than one, the search has no sensitivity to the model type for the corresponding tanβ and m_{a_1} values.

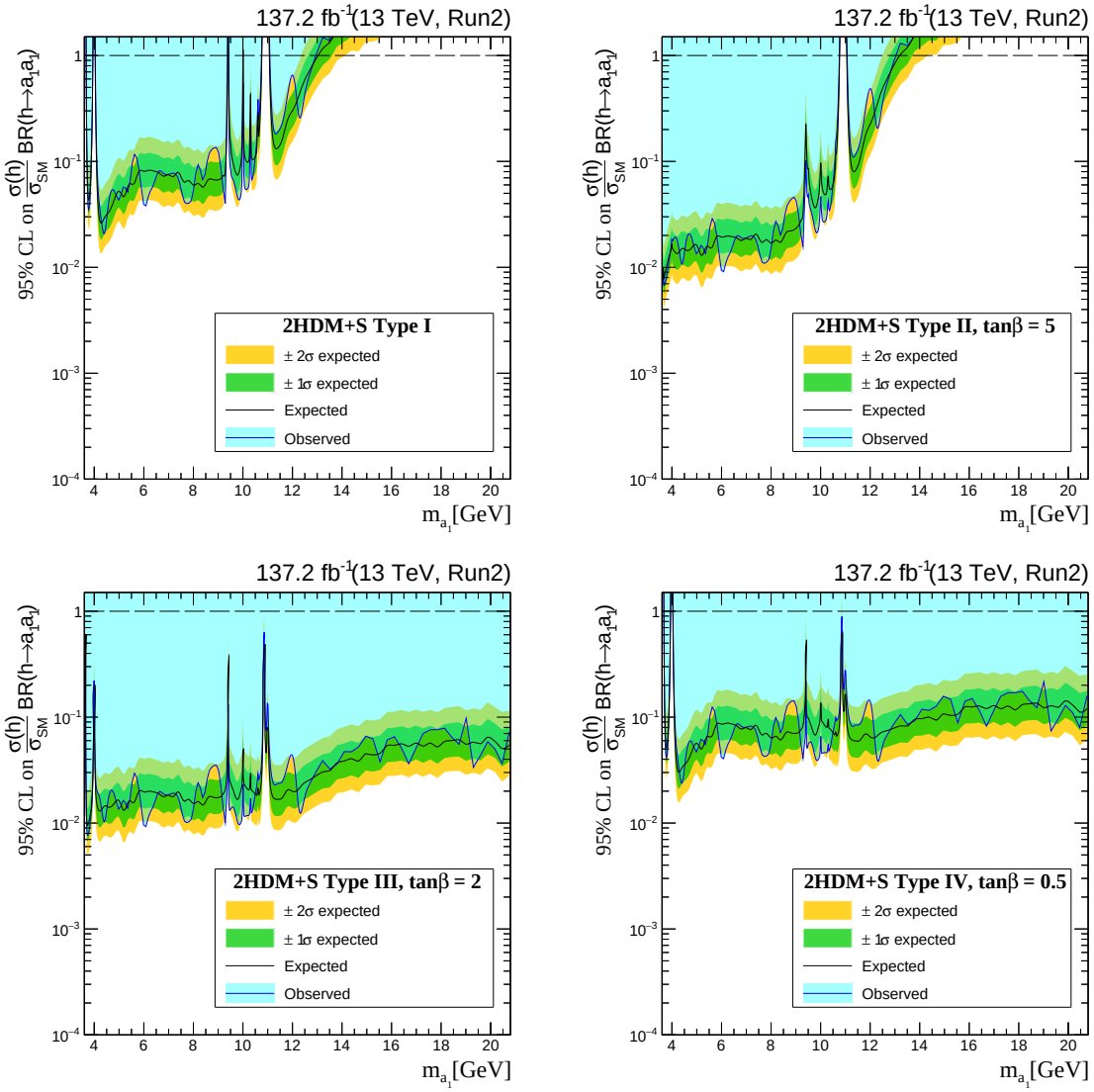


Figure 8.6: Upper limits on $(\sigma_h/\sigma_{\mathrm{SM}}) \cdot \mathcal{B}(h \to a_1 a_1)$ in the 2HDM+S Type I (upper left), Type II (upper right), Type III (lower left), and Type IV (lower right), for a given value of tanβ. The dotted line corresponds to $\mathcal{B}(h \to a_1 a_1) = 1$.

Upper limits for values of tanβ up to 10 are shown in Fig. 8.7 and Fig. 8.8. The peak-like shape of contours observed around 10 GeV is caused by the a_1-quarkonium mixing. The excluded region at 95% CL for the branching fraction of the decay of the 125 GeV Higgs boson into non-SM particles, placed at 34% from Run I combined ATLAS and CMS analysis [185], is reflected by the overlaid red contour line. The contour line runs over the set of values of

tanβ and m_{a_1}, which correspond to a branching fraction of the decay of the Higgs boson into a pair of light bosons equal to 34%, excluded at 95% CL. The additional blue contour line reflects the scenario in which the value of 34% is reduced to 5%. Such an improvement might come from fit results of parameterizations allowing contributions from BSM particles in loops and decays of the Higgs boson and further dedicated searches.
As it can be seen in Fig. 8.7b, the analysis has good sensitivity to the Type II 2HDM+S, for values of tan$\beta > 1$ and m_{a_1} below the b-quark pair threshold. The good sensitivity is explained by an enhanced decay to down-type fermions for the model type in the mentioned mass range. The analysis has exclusion power over the full probed mass range for the Type III 2HDM+S, especially for tan$\beta > 1$, as illustrated in Fig. 8.8a. In the Type III 2HDM+S, the decays to $\tau\tau$ dominate over decays to $b\bar{b}$ above the $b\bar{b}$ threshold, with the branching fraction to leptons increasing as a function of tanβ, which explains the stronger upper limits obtained as higher values of tanβ are probed. However, for the Type IV 2HDM+S, only the region corresponding to low values of tanβ can be effectively probed. The enhanced branching fraction to up-type quarks and leptons with respect to down-type quarks for tan$\beta < 1$ in the Type IV 2HDM+S, results in a similar branching ratio for $b\bar{b}$, $c\bar{c}$, and $\tau\tau$, as discussed in Subsec. 3.1.1. This makes the final states $2b2\tau$ and $2c2\tau$ more sensitive to the Type IV 2HDM+S and explains the low sensitivity observed in Fig. 8.8b for the $2\mu2\tau$ final state.
The analysis provides the tightest constraints within the probed mass range on exotic Higgs boson decays in scenarios with enhanced decays of the pseudoscalar boson to leptons. Within these scenarios, even if the excluded region at 95% CL for the branching fraction of the decay of the 125 GeV Higgs boson into non-SM particles is lowered to 5% from the current 34%, the results here presented will offer tighter constraints for values of tanβ above the corresponding contour line in Fig. 8.7b and Fig. 8.8a. This is particularly relevant considering that a branching fraction of $5-10$% into exotic decay modes will remain a reasonable target for the duration of the LHC physics program, as discussed in Sec. 3.2.

8.3.2 Dark Photon Model

The procedure described above is also followed to obtain model dependent results in the context of the Dark Photon Model. For sufficient low values of the kinetic mixing (ϵ) and the Higgs mixing (κ) parameters, the phenomenology of the $h \to Z_D Z_D \to x\bar{x}y\bar{y}$ decays is determined by the mass of the dark photon candidate (m_{Z_D}) (Subsec. 3.1.2). Once the discrete number of mass points from the experimental search is converted into the map $\mu_{\text{up}}(m_{Z_D})$, the equation for the transfer of experimental limits into upper limits in the context of the Dark Photon Model can be written as:

$$((\sigma_h/\sigma_{\text{SM}}) \cdot \mathcal{B}(h \to Z_D Z_D))_{\text{up}} = \frac{\mu_{\text{up}}(m_{Z_D})}{2 \cdot \mathcal{B}(Z_D \to x\bar{x}; m_{Z_D}) \cdot \mathcal{B}(Z_D \to y\bar{y}; m_{Z_D})}, \tag{8.4}$$

where the branching fractions for the decay of the dark photon into a pair of fermions in the denominator are provided by theorists as a function of m_{Z_D} [50]. In the analysis, upper limits at 95% CL on $(\sigma_h/\sigma_{\text{SM}}) \cdot \mathcal{B}(h \to Z_D Z_D)$ are set as a function of the mass of the dark photon candidate using Eq. (8.4).

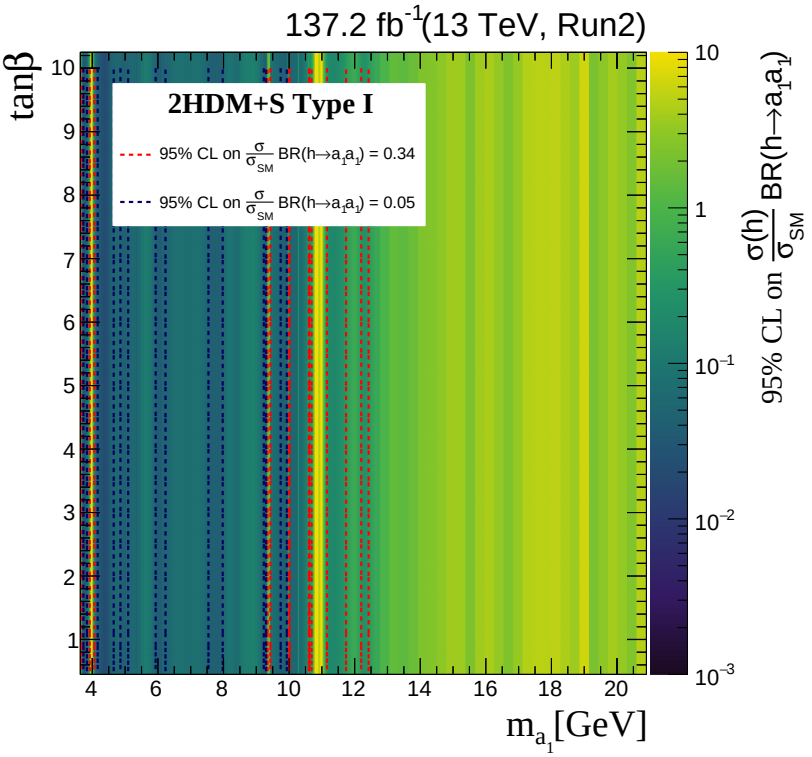


Fig. 8.7 a)

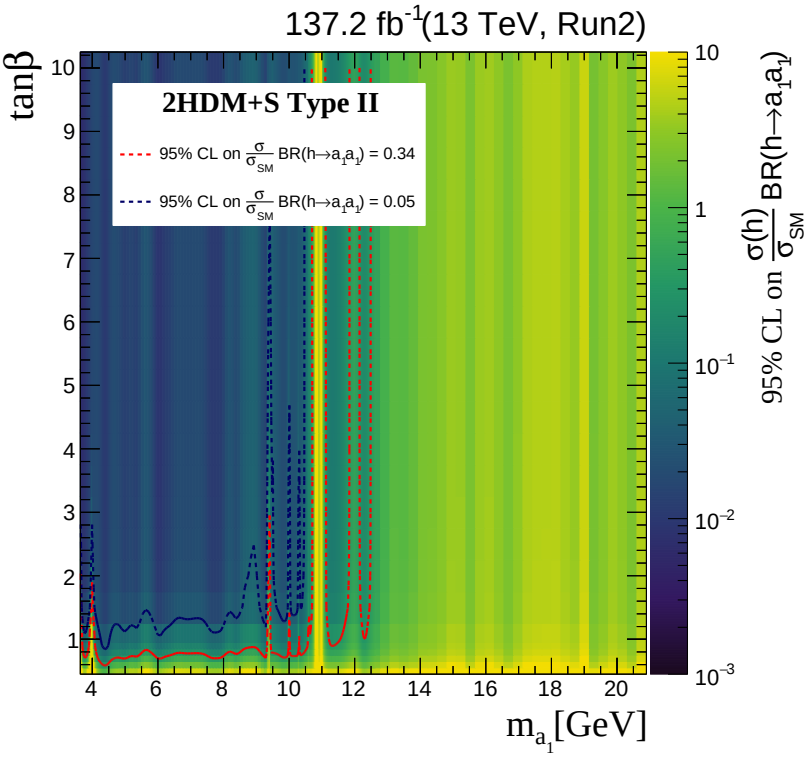


Fig. 8.7 b)

Figure 8.7: Upper limits on $(\sigma_h/\sigma_{\mathrm{SM}}) \cdot \mathcal{B}(h \to a_1 a_1)$ for the 2HDM+S Type I (a) and Type II (b) as a function of tanβ and m_{a_1}. Contour lines are shown for $\mathcal{B}(h \to a_1 a_1) = 0.05$ and 0.34.

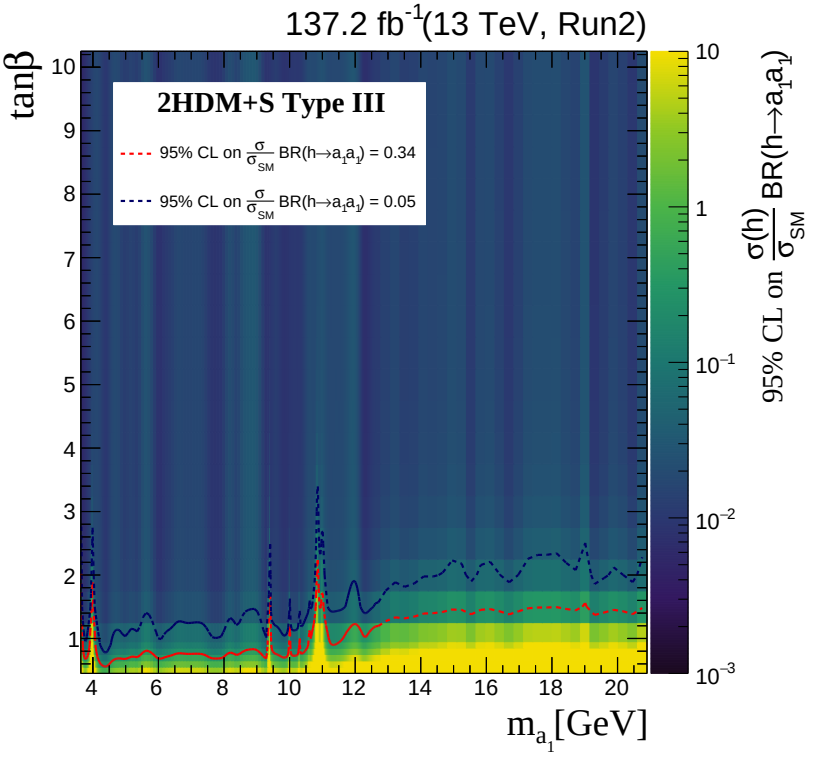


Fig. 8.8 a)

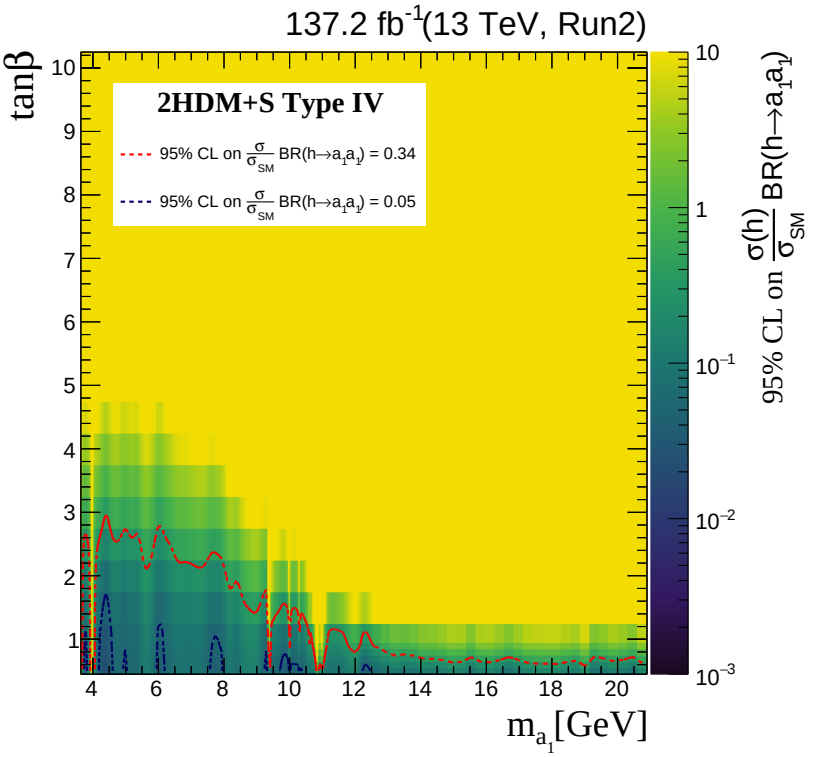


Fig. 8.8 b)

Figure 8.8: Upper limits on $(\sigma_h/\sigma_{\mathrm{SM}}) \cdot \mathcal{B}(h \to a_1 a_1)$ for the 2HDM+S Type III (a) and Type IV (b) as a function of tanβ and m_{a_1}. Contour lines are shown for $\mathcal{B}(h \to a_1 a_1) = 0.05$ and 0.34.

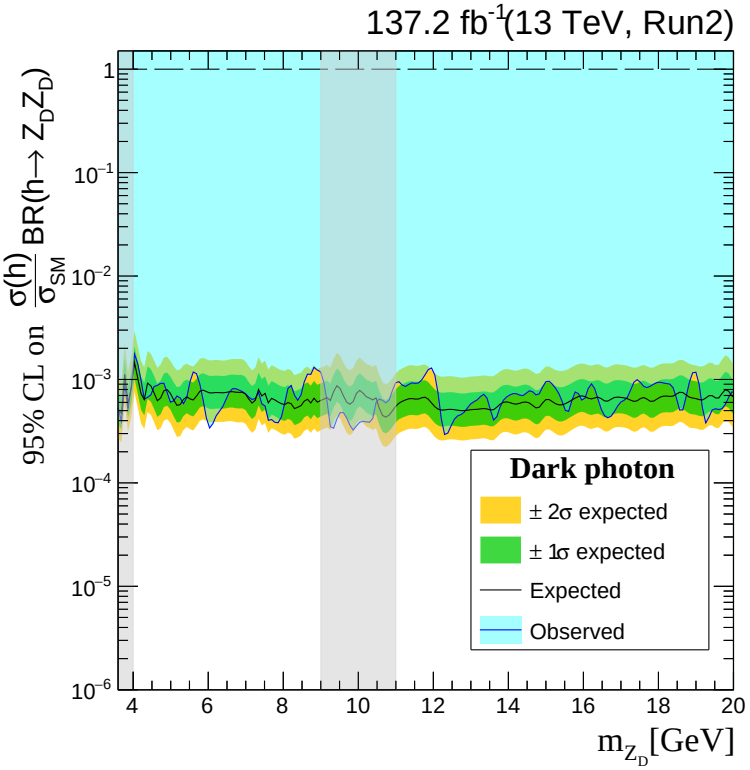


Figure 8.9: Upper limits on $(\sigma_h/\sigma_{\mathrm{SM}}) \cdot \mathcal{B}(h \to Z_D Z_D)$ for the Dark Photon Model as a function of the mass of the dark photon candidate. The shaded areas correspond to the quarkonia veto regions.

The ATLAS results with a partial Run II dataset, covering the $4l$ final state with $l = e, \mu$ [58], provide more stringent constraints in the context of the Dark Photon Model than the results presented in this thesis. The ATLAS analysis profits from the cleaner signature of electrons and muons in the detector, avoids the difficulties of the τ reconstruction, and the signal yield is maintained due to the similar branching fraction of the leptonic final states for masses of the dark photon candidate above 3.5 GeV (Subsec. 3.3.2). The recent CMS search with full Run II dataset in the $4l$ final state with $l = e, \mu$ [63], is therefore motivated by these conclusions and complements the previous CMS results for masses of the dark photon candidate below 3.5 GeV [62].

CHAPTER

9

SUMMARY AND CONCLUSIONS

The existence of exotic decays of the 125 GeV Higgs boson into a pair of light bosons is predicted by different theories beyond the Standard Model, such as the 2HDM+S and the Dark Photon Model. A search for such pair of light bosons, with one of the light bosons decaying into a pair muons and the other into a pair of τ leptons, is presented. Scenarios within the mentioned models in which the decays of the light boson to leptons are enhanced motivate the final state probed in this work.

The full Run II dataset collected with the CMS detector at the LHC during the years 2016, 2017, and 2018 at a center-of-mass energy of $\sqrt{s} = 13$ TeV is used. The data samples were accumulated with single muon triggers, corresponding to integrated luminosities of 35.9, 41.5, and 59.7 fb^{-1}, respectively, for a total of 137.2 fb^{-1}. The analysis profits from the good muon detection and momentum resolution of the CMS detector, whose design is tailored to search for signatures of new physics with muons in the final state. The final event selection is based on a pair of muons and a pair of tracks and each of the four objects is required to be isolated within a ΔR cone of 0.2. The SM background is estimated using a data-driven method in a control region orthogonal to the signal region, in which at least one of the two muons or one of the two tracks is anti-isolated within a ΔR cone of 0.2. The background model is validated in additional control regions, in which the isolation requirement on the four objects is relaxed, or the opposite sign requirement for the pair of tracks inverted.

A machine learning approach using boosted decision trees is applied to separate the signal from the background events. The output distribution of the BDT classifier constitutes the final discriminant of the analysis. To test for the existence of the signal, a binned maximum-likelihood fit is applied to the BDT classification distribution. The search covers the mass range of the light boson between 3.6 and 21 GeV. The lower end of the range is determined by the mass of the pair of τ leptons and the upper end by the transition to the non-boosted topology, in which different analysis techniques would need to be used. In the absence of a significant excess in data above the expected SM background, upper limits at 95% confidence level are set on the signal cross-section times the branching ratio

$\sigma(pp \to H(125) + X) \cdot \mathcal{B}(H(125) \to a_1 a_1(Z_D Z_D) \to \mu\mu\tau\tau)$ relative to the inclusive cross-section $\sigma(pp \to H(125) + X)_{\mathrm{SM}}$ predicted in the SM. Therefore, the main conclusion of this thesis is that no evidence for the production of the pair of light bosons is found. The observed limits range from 0.57×10^{-4} at $m_{a_1(Z_D)} = 12.4$ GeV to 2.29×10^{-4} at $m_{a_1(Z_D)} = 8.8$ GeV. The obtained model-independent exclusion limits are interpreted as model-specific upper limits on $(\sigma_h/\sigma_{\mathrm{SM}})\ \mathcal{B}(h \to a_1 a_1(Z_D Z_D))$ for different benchmark scenarios of the 2HDM+S and the Dark Photon Model. In the case of the 2HDM+S, the exclusion limits are derived for Type I to Type IV in the $[m_{a_1}, \tan\beta]$ plane and for the Dark Photon Model as a function of the mass of the dark photon candidate m_{Z_D}. The exclusion power of the analysis includes a wide range of the parameter phase space of the different types of 2HDM+S, especially in the scenarios with enhanced couplings to leptons. The most stringent limits are obtained for the Type III 2HDM+S at large values of $\tan\beta$.

The analysis constitutes the first search in the $\mu\mu\tau\tau$ final state with full Run II dataset at a center-of-mass energy of 13 TeV, complementing previous CMS searches at $\sqrt{s} = 7$, 8, and 13 TeV. It provides the tightest available constraints within the probed mass range on exotic decays of the 125 GeV Higgs boson for scenarios with an enhanced decay of the light boson to leptons.

The results of the analysis may be improved by adding the contributions from the VBF, VH, and ttH processes, and the search may be extended by considering masses of the Higgs boson above 125 GeV. Additional improvements might come with the application of machine learning techniques in several moments of the analysis workflow, especially in the identification of the ditau system. Possibilities are open in the event and particle identification, energy estimation, pileup suppression, physics performance of the reconstruction, analysis algorithms, execution time of some tasks as the event simulation, pattern recognition, and calibration as well as the management of the data in terms of data compression, placement, and access. In addition, the sensitivity reach of the analysis might be extended through optimizations of the techniques and the analysis workflow, as well as changes in the data-taking conditions such as the increased luminosity and center of mass-energy, expected for the upcoming periods of data taking at the LHC, namely the Run III and the High-Luminosity LHC. Further promising opportunities to search for exotic decays of the 125 Higgs boson into a pair of light bosons might come from future lepton colliders such as the Circular Electron Positron Collider, the Future Circular Collider, and the International Linear Collider.

To conclude, the analysis here presented and, in general, the exploration of an extended Higgs sector, offers exciting prospects for the future. Present and future colliders will allow us to perform more direct searches and precise measurements of the properties of the 125 Higgs boson. These new efforts will complement current results and pave the way to cement our knowledge of a possible extended Higgs sector.

APPENDIX

A

COMPLEMENTARY RESULTS WITH 2016, 2017, AND 2018 DATA

In this appendix, additional material corresponding to the $H \to a_1 a_1 (Z_D Z_D) \to \mu\mu\tau\tau$ search with CMS Run II data is provided, complementing the results presented in Chapters 7 and 8.

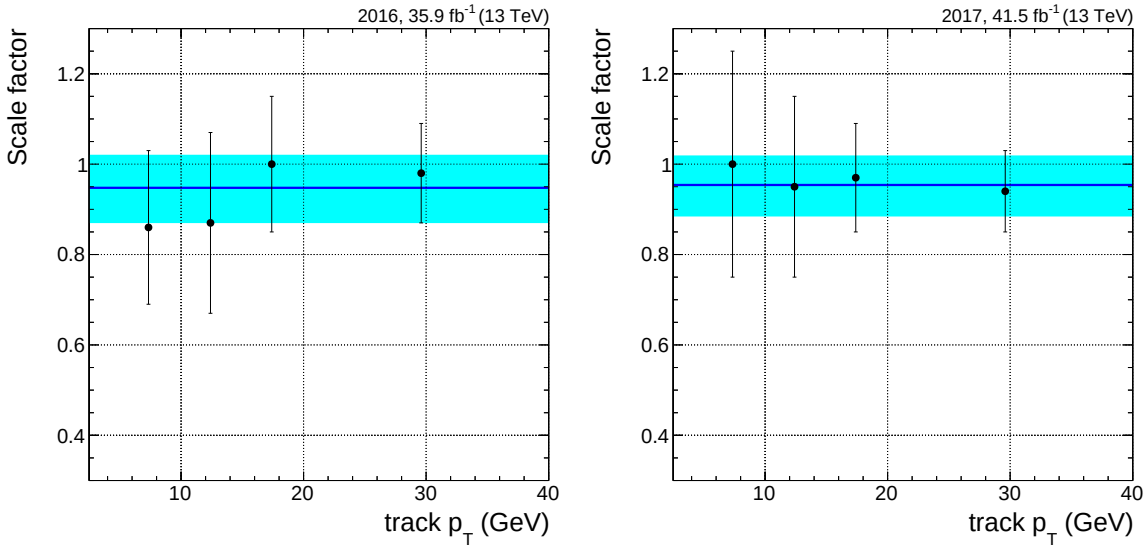


Figure A.1: Scale factor for the combined track isolation and one-prong tau decay identification efficiency as a function of the track p_T, corresponding to the 2016 (left) and 2017 (right) analysis. The constant scale factor is applied in the main analysis to simulated events.

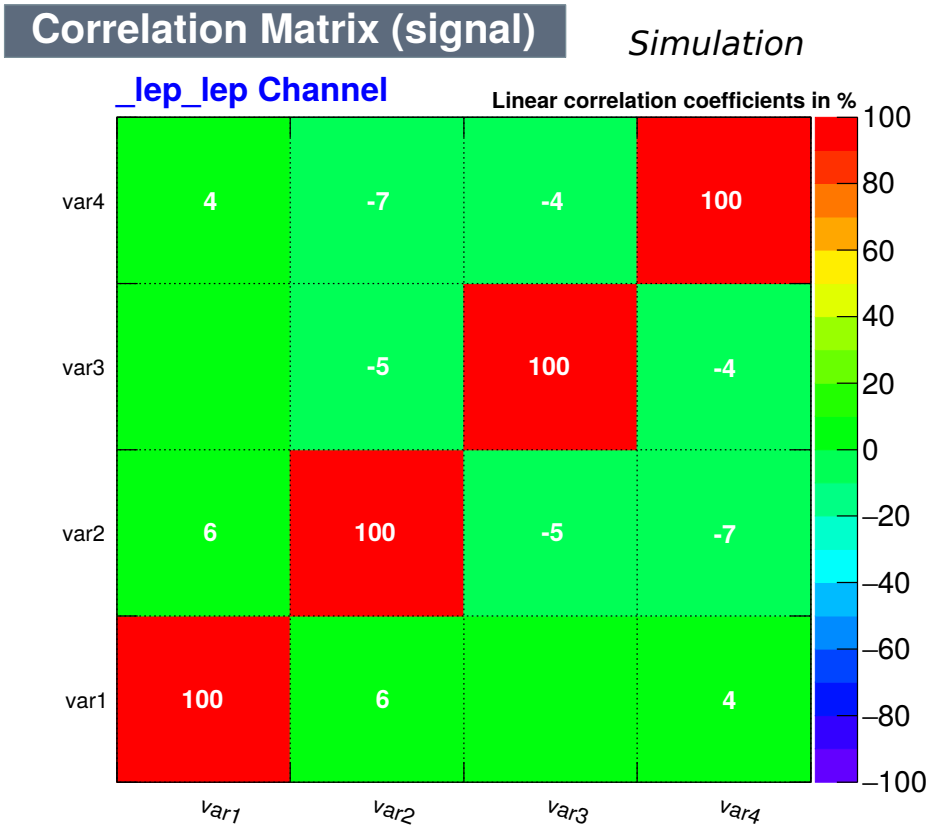


Figure A.2: Matrix of linear correlation coefficients for the variables used in the BDT training: m_{μ_1,μ_2} (var1), $\Delta R(\mu_1, \mu_2)$ (var2), $\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$ (var3), and $m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$ (var4), for $m_{a_1(Z_D)} = 5$ GeV and the 2016 dataset in the lepton-lepton category.

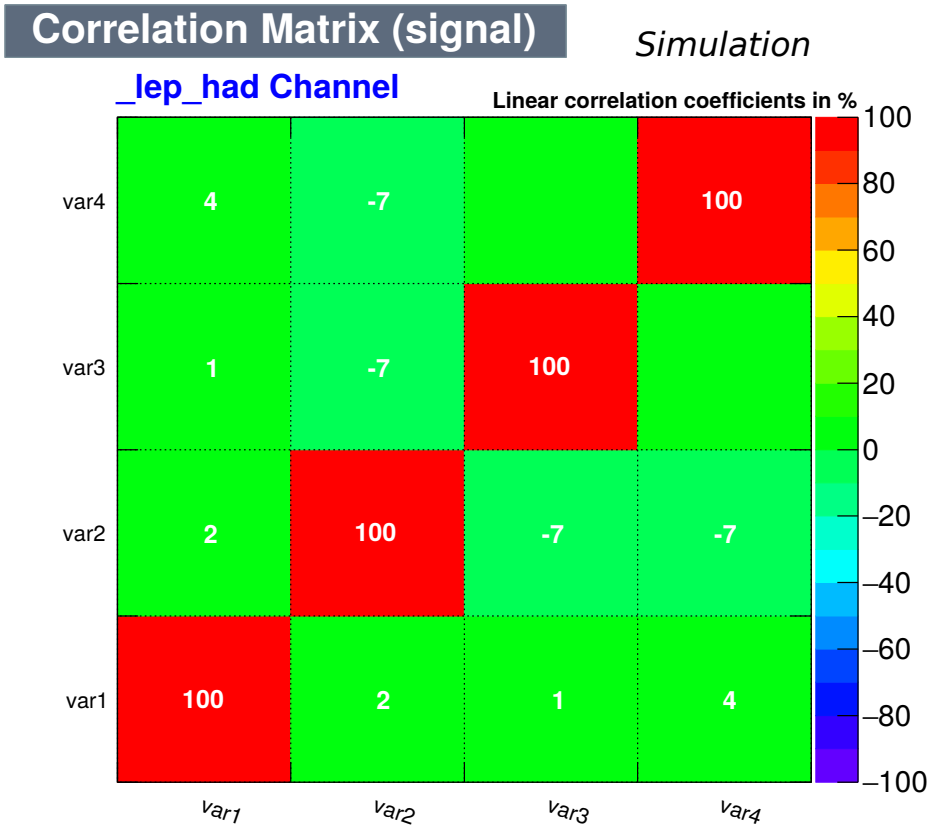


Figure A.3: Matrix of linear correlation coefficients for the variables used in the BDT training: m_{μ_1,μ_2} (var1), $\Delta R(\mu_1, \mu_2)$ (var2), $\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$ (var3), and $m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$ (var4), for $m_{a_1(Z_D)} = 5$ GeV and the 2016 dataset in the lepton-hadron category.

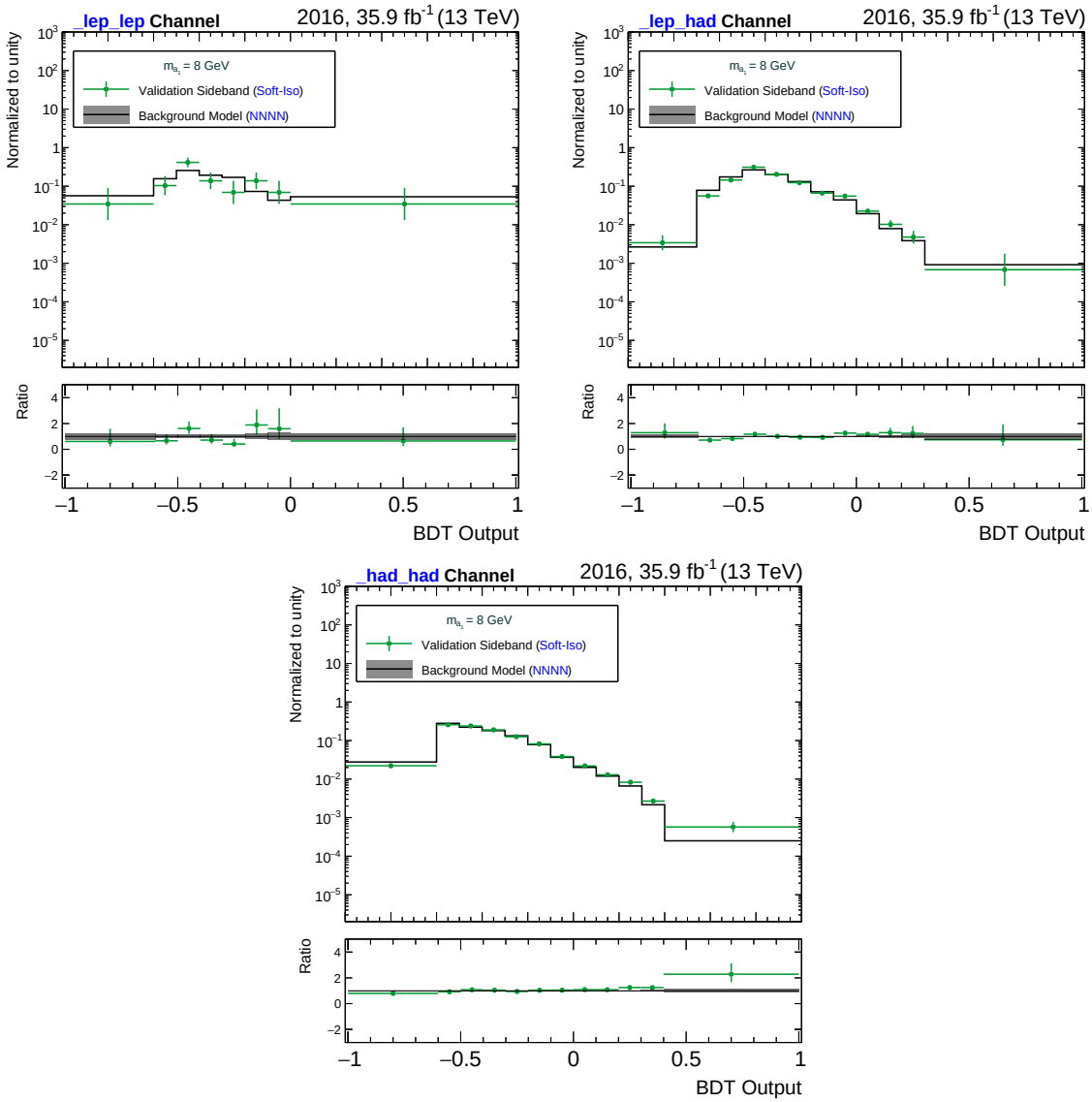


Figure A.4: Data-driven closure test of background model. The BDT output distribution for events passing the signal selection in the validation region Soft-Iso is compared with the shape in CR NNNN, for $m_{a_1(Z_D)} = 8$ GeV and the 2016 dataset in the lepton-lepton (upper left), lepton-hadron (upper right), and hadron-hadron (lower) categories. The lower panel shows the ratio of the distribution observed in Soft-Iso to the distribution observed in NNNN for the corresponding category.

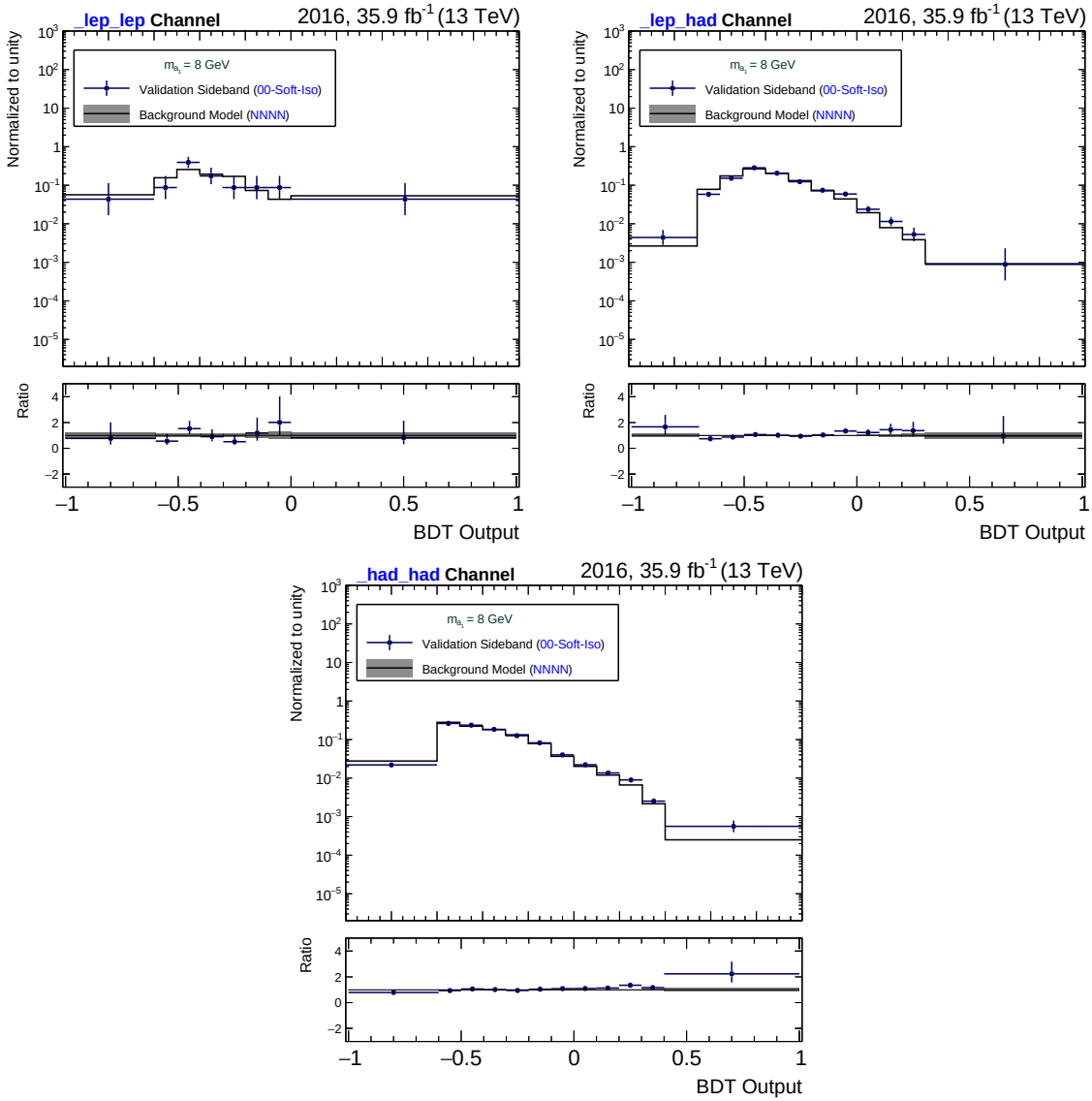


Figure A.5: Data-driven closure test of background model. The BDT output distribution for events passing the signal selection in the validation region 00-Soft-Iso is compared with the shape in CR NNNN, for $m_{a_1(Z_D)} = 8$ GeV and the 2016 dataset in the lepton-lepton (upper left), lepton-hadron (upper right), and hadron-hadron (lower) categories. The lower panel shows the ratio of the distribution observed in 00-Soft-Iso to the distribution observed in NNNN for the corresponding category.

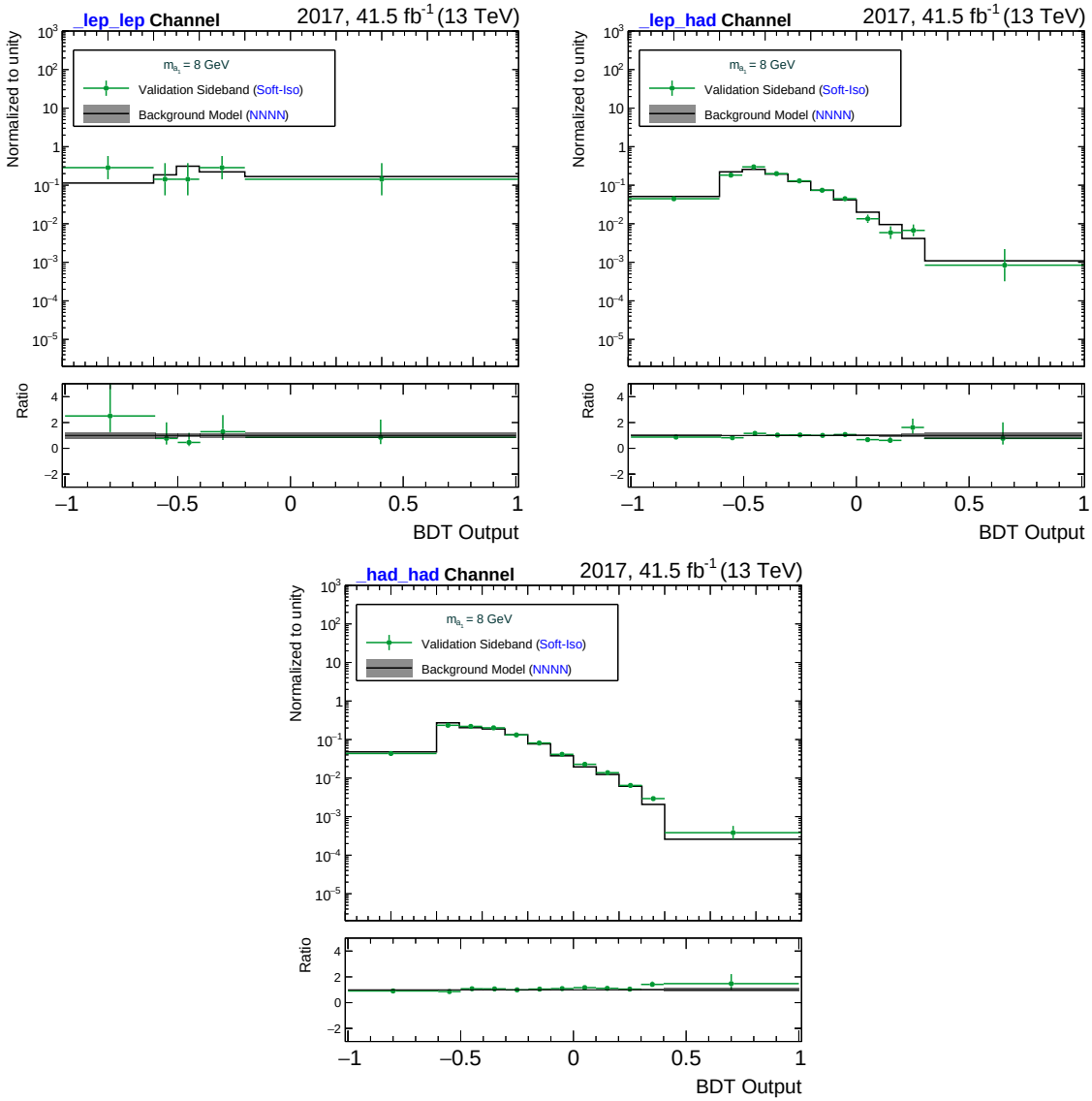


Figure A.6: Data-driven closure test of background model. The BDT output distribution for events passing the signal selection in the validation region Soft-Iso is compared with the shape in CR NNNN, for $m_{a_1(Z_D)} = 8$ GeV and the 2017 dataset in the lepton-lepton (upper left), lepton-hadron (upper right), and hadron-hadron (lower) categories. The lower panel shows the ratio of the distribution observed in Soft-Iso to the distribution observed in NNNN for the corresponding category.

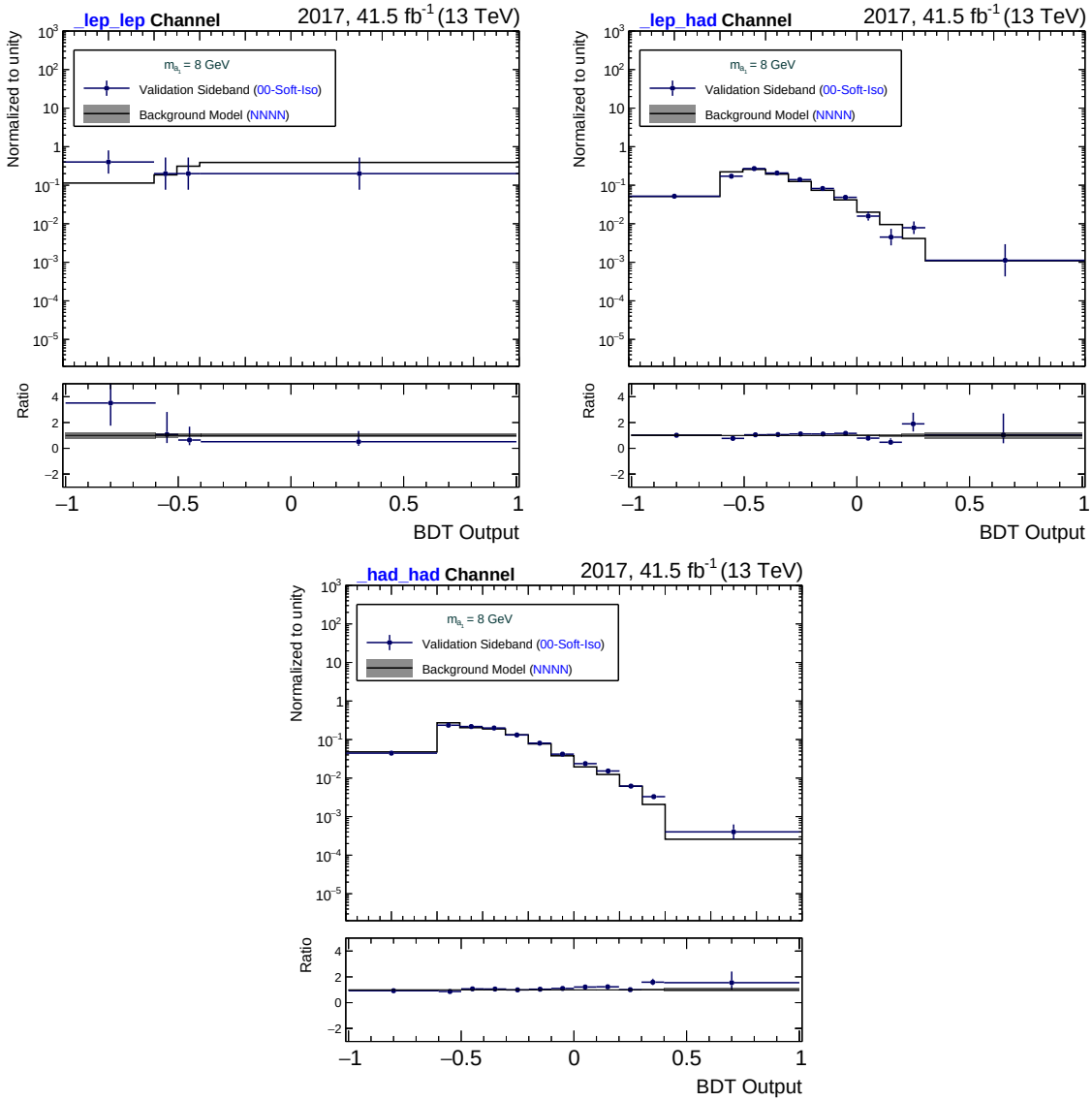


Figure A.7: Data-driven closure test of background model. The BDT output distribution for events passing the signal selection in the validation region 00-Soft-Iso is compared with the shape in CR NNNN, for $m_{a_1(Z_D)} = 8$ GeV and the 2017 dataset in the lepton-lepton (upper left), lepton-hadron (upper right), and hadron-hadron (lower) categories. The lower panel shows the ratio of the distribution observed in 00-Soft-Iso to the distribution observed in NNNN for the corresponding category.

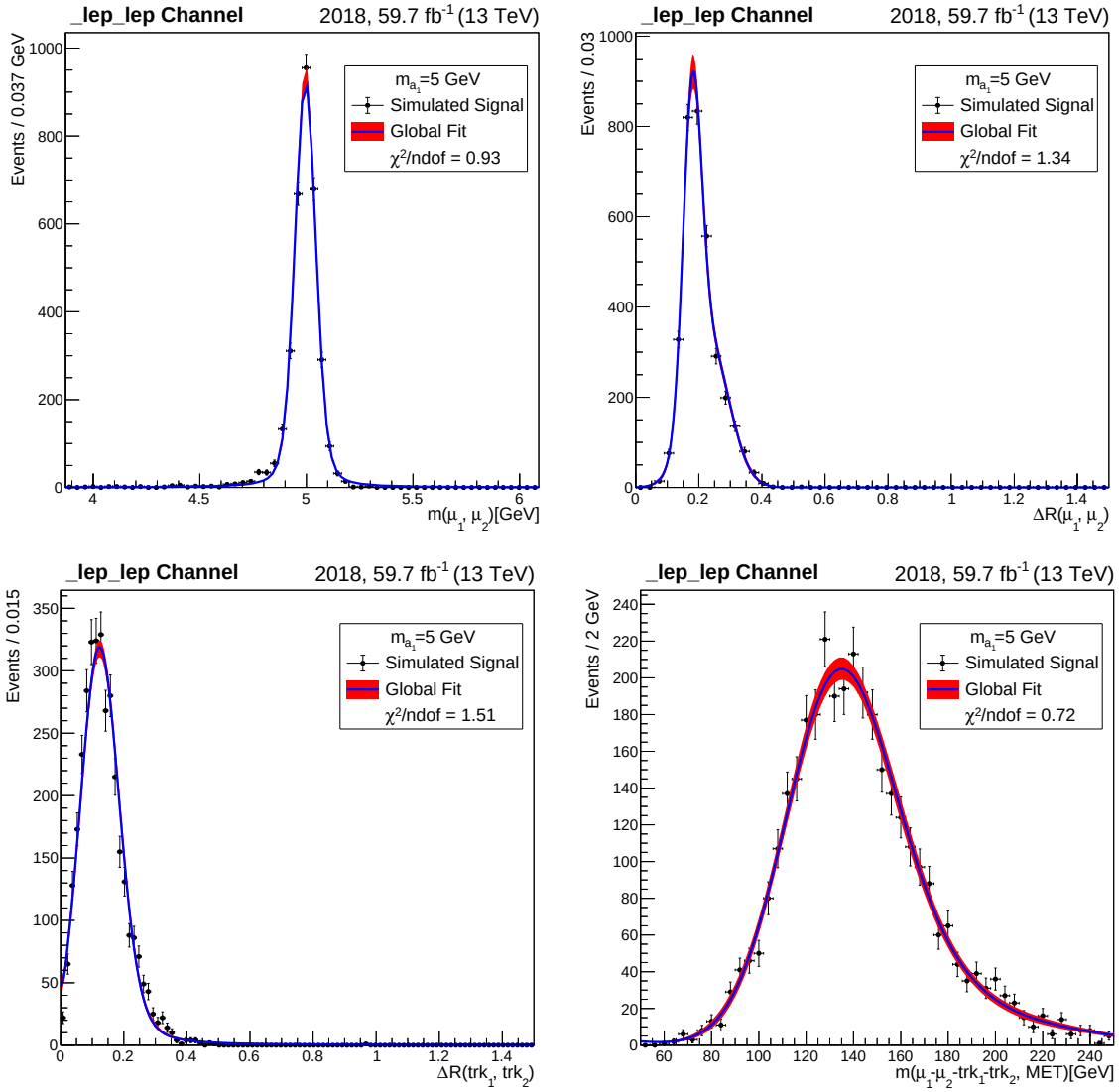


Figure A.8: Parameterized distribution of the four BDT input variables: m_{μ_1,μ_2} (upper left), $\Delta R(\mu_1, \mu_2)$ (upper right), $\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$ (lower left), and $m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$ (lower right). The shown distributions correspond to $m_{a_1(Z_D)} = 5$ GeV in the lepton-lepton category for the 2018 analysis. The red shadow represents the fit error band within 2-sigma of confidence interval.

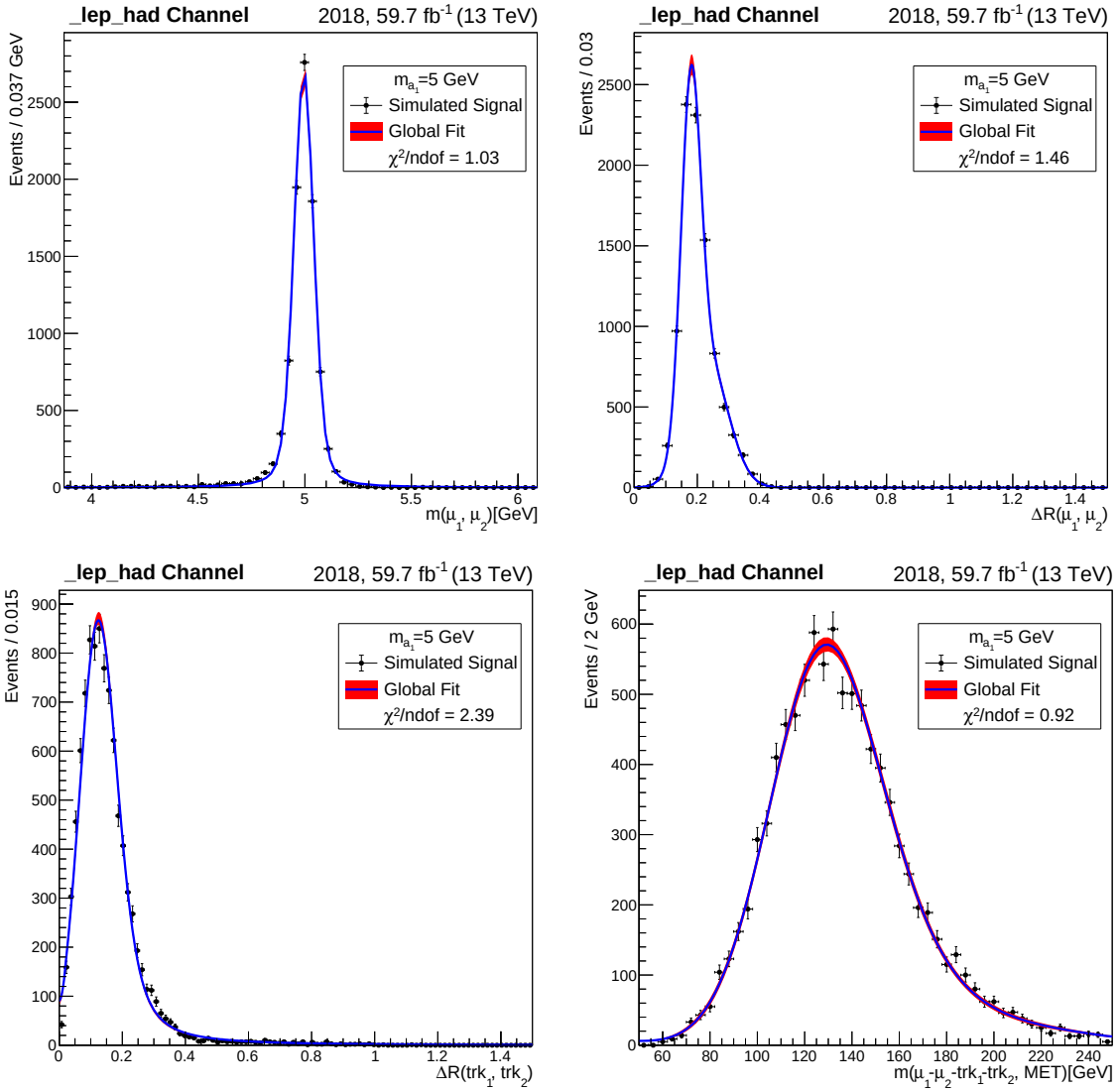


Figure A.9: Parameterized distribution of the four BDT input variables: m_{μ_1,μ_2} (upper left), $\Delta R(\mu_1, \mu_2)$ (upper right), $\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$ (lower left), and $m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$ (lower right). The shown distributions correspond to $m_{a_1(Z_D)} = 5$ GeV in the lepton-hadron category for the 2018 analysis. The red shadow represents the fit error band within 2-sigma of confidence interval.

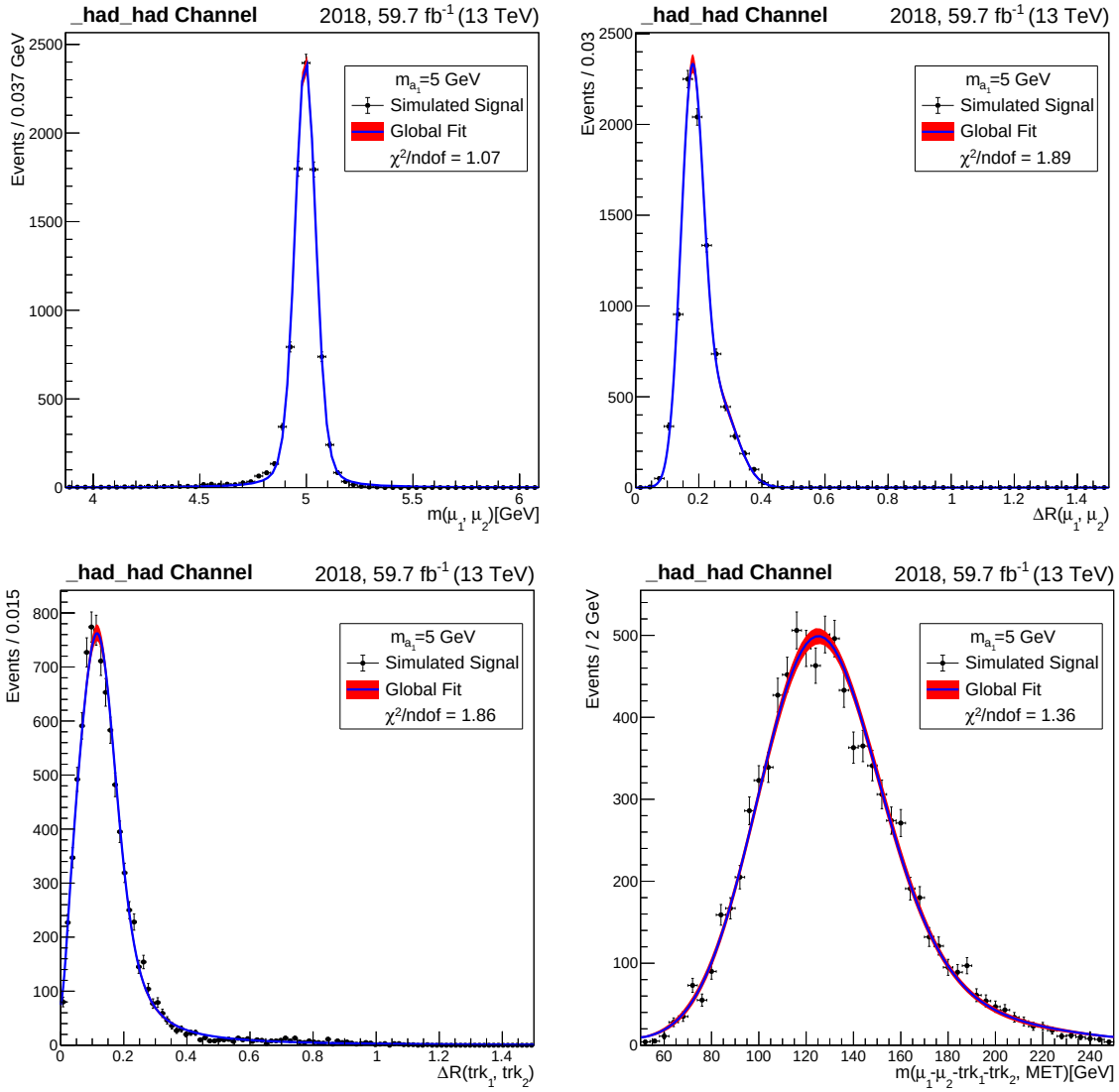


Figure A.10: Parameterized distribution of the four BDT input variables: m_{μ_1,μ_2} (upper left), $\Delta R(\mu_1, \mu_2)$ (upper right), $\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$ (lower left), and $m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$ (lower right). The shown distributions correspond to $m_{a_1(Z_D)} = 5$ GeV in the hadron-hadron category for the 2018 analysis. The red shadow represents the fit error band within 2-sigma of confidence interval.

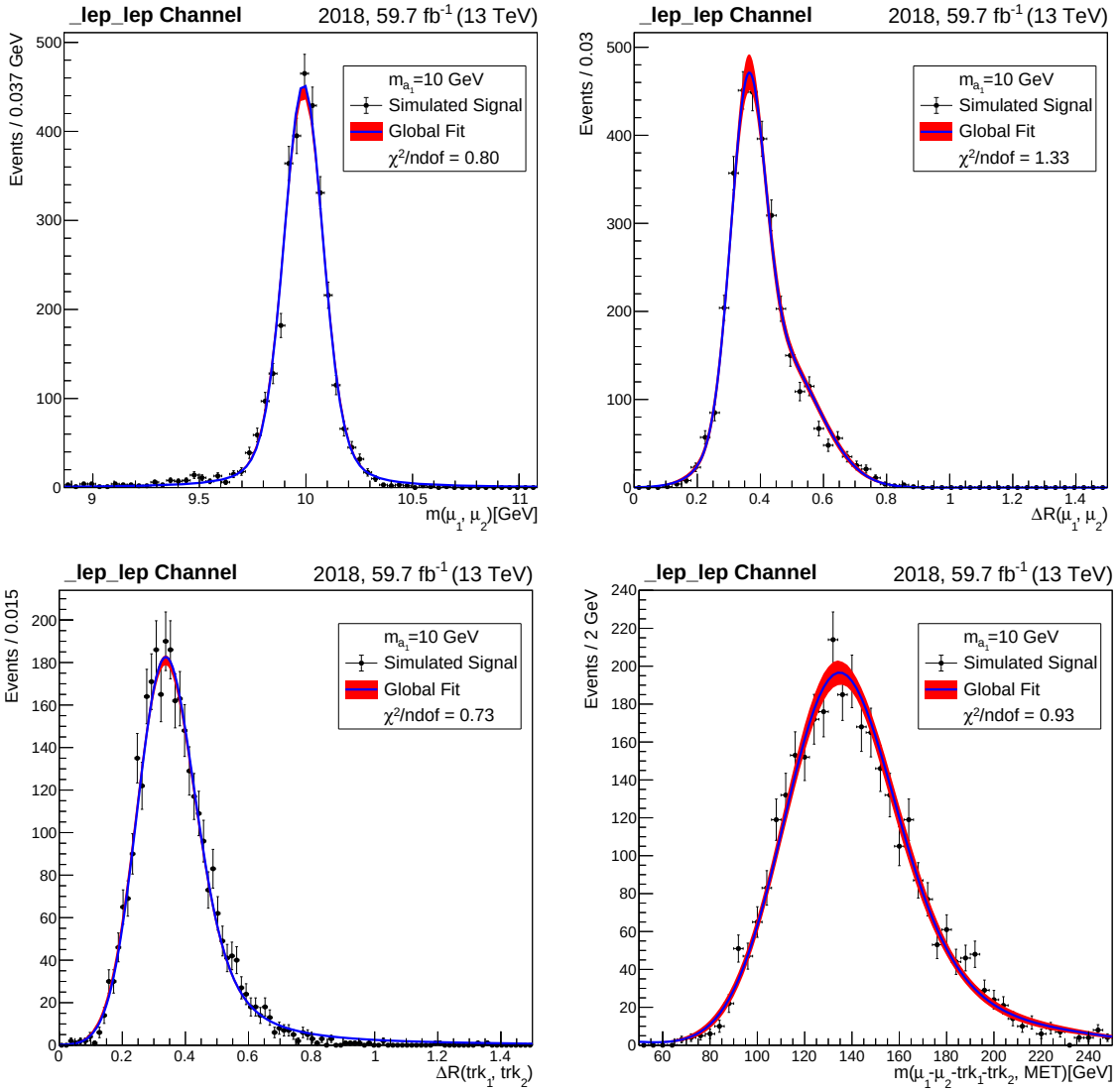


Figure A.11: Parameterized distribution of the four BDT input variables: m_{μ_1,μ_2} (upper left), $\Delta R(\mu_1, \mu_2)$ (upper right), $\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$ (lower left), and $m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$ (lower right). The shown distributions correspond to $m_{a_1(Z_D)} = 10$ GeV in the lepton-lepton category for the 2018 analysis. The red shadow represents the fit error band within 2-sigma of confidence interval.

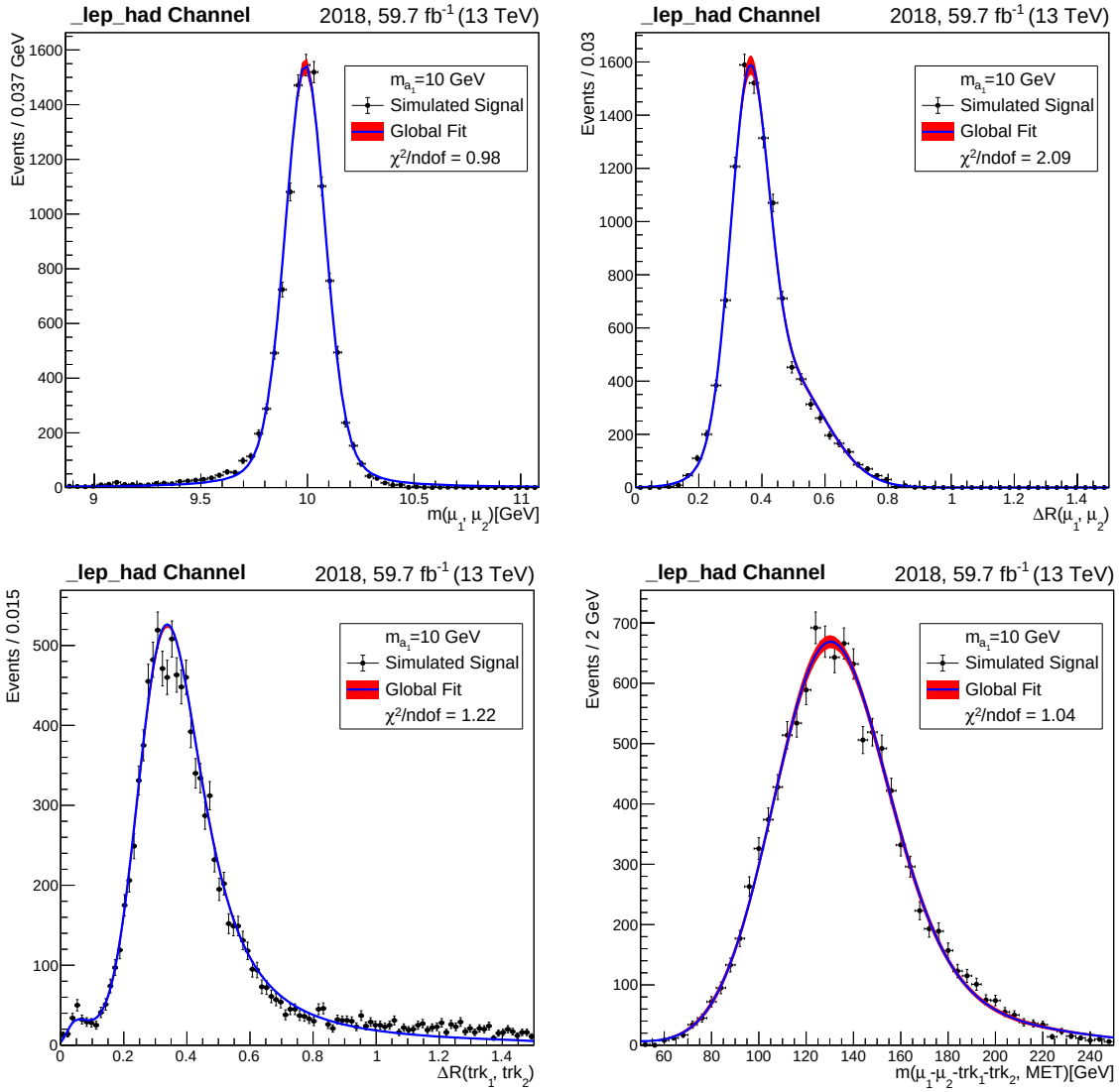


Figure A.12: Parameterized distribution of the four BDT input variables: m_{μ_1,μ_2} (upper left), $\Delta R(\mu_1, \mu_2)$ (upper right), $\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$ (lower left), and $m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$ (lower right). The shown distributions correspond to $m_{a_1(Z_D)} = 10$ GeV in the lepton-hadron category for the 2018 analysis. The red shadow represents the fit error band within 2-sigma of confidence interval.

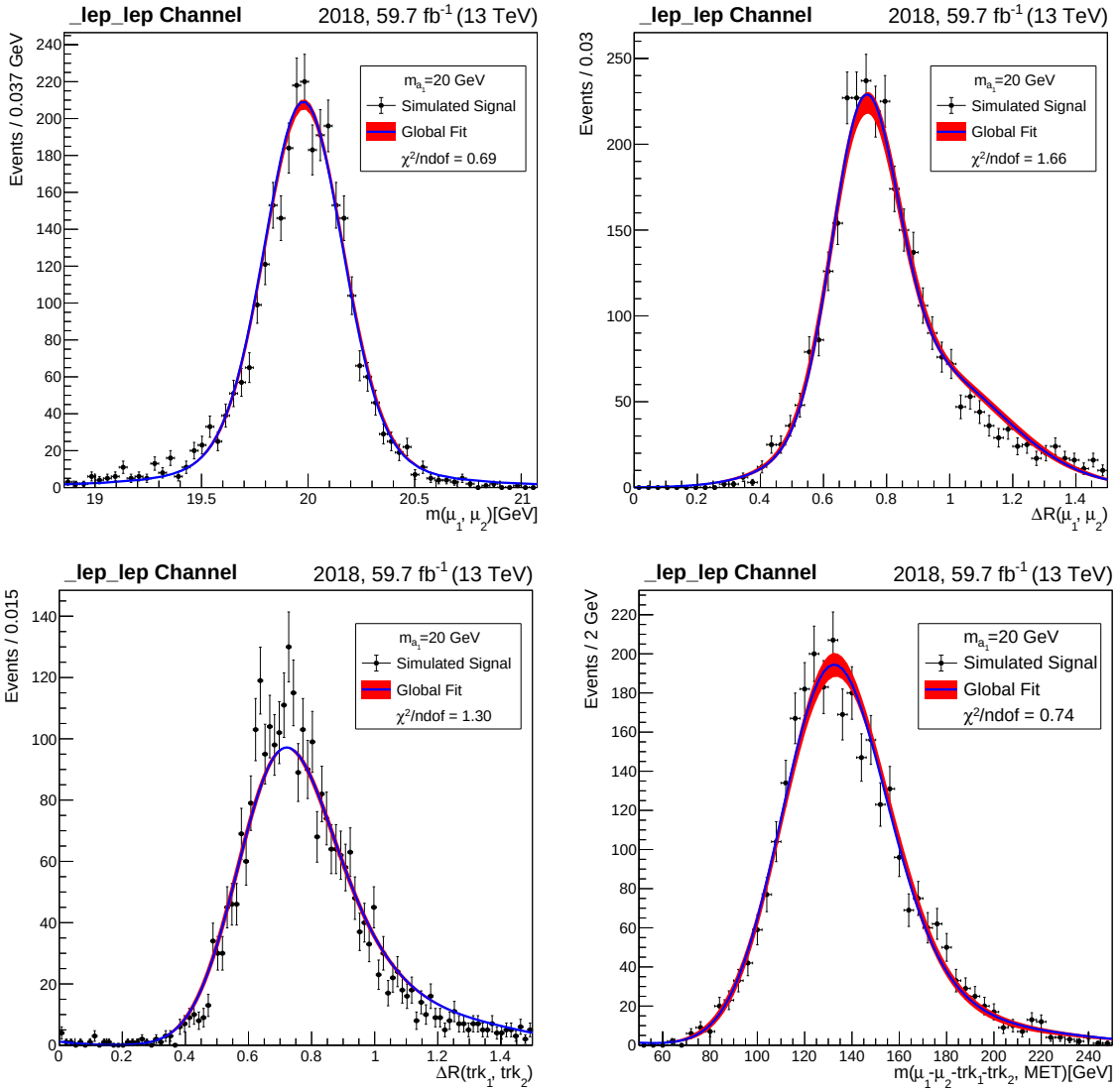


Figure A.13: Parameterized distribution of the four BDT input variables: m_{μ_1,μ_2} (upper left), $\Delta R(\mu_1, \mu_2)$ (upper right), $\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$ (lower left), and $m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$ (lower right). The shown distributions correspond to $m_{a_1(Z_D)} = 20$ GeV in the lepton-lepton category for the 2018 analysis. The red shadow represents the fit error band within 2-sigma of confidence interval.

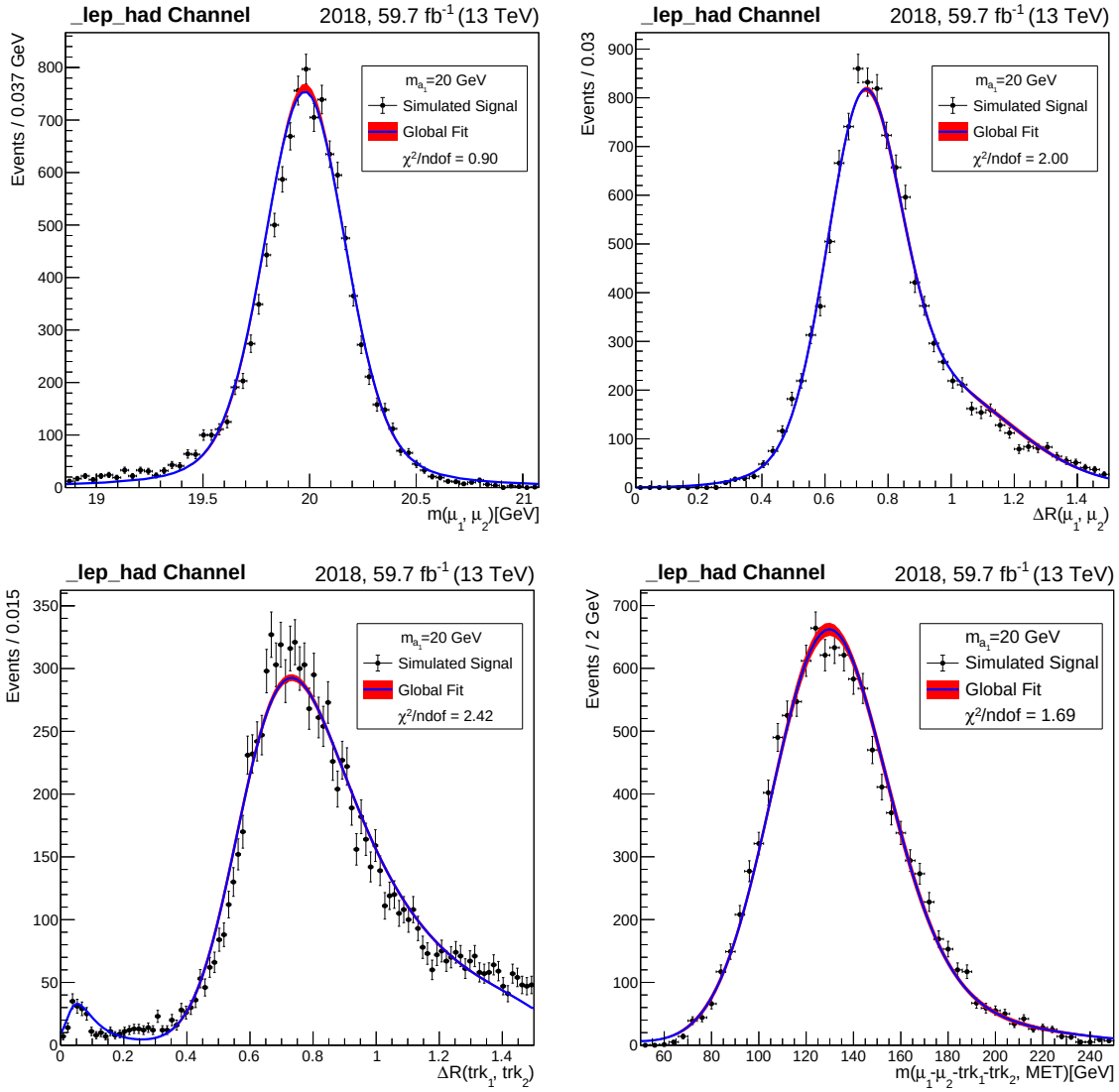


Figure A.14: Parameterized distribution of the four BDT input variables: m_{μ_1,μ_2} (upper left), $\Delta R(\mu_1, \mu_2)$ (upper right), $\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$ (lower left), and $m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$ (lower right). The shown distributions correspond to $m_{a_1(Z_D)} = 20$ GeV in the lepton-hadron category for the 2018 analysis. The red shadow represents the fit error band within 2-sigma of confidence interval.

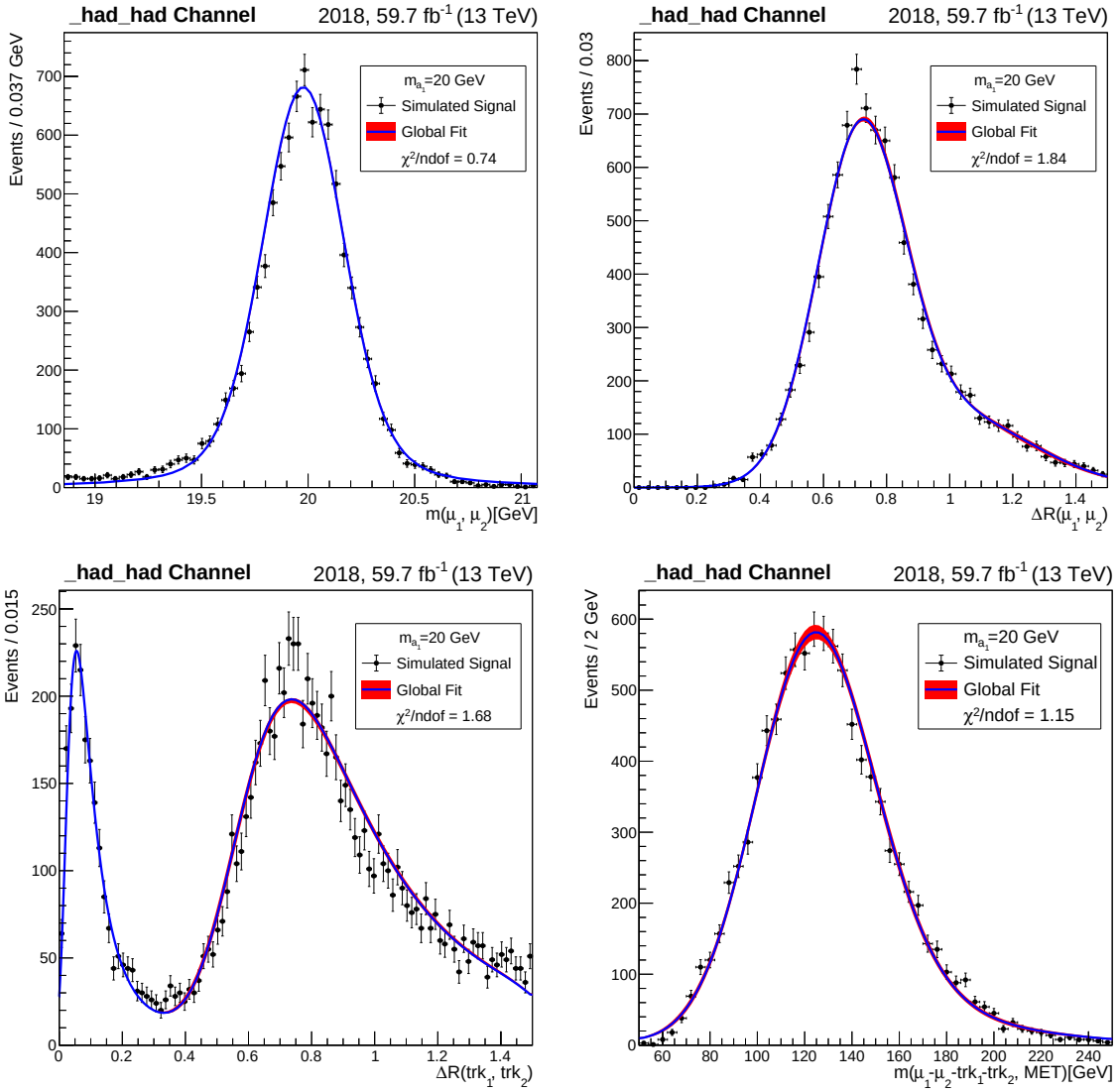


Figure A.15: Parameterized distribution of the four BDT input variables: m_{μ_1,μ_2} (upper left), $\Delta R(\mu_1, \mu_2)$ (upper right), $\Delta R(\mathrm{trk}_1, \mathrm{trk}_2)$ (lower left), and $m_{\mu_1,\mu_2,\mathrm{trk}_1,\mathrm{trk}_2,\mathrm{MET}}$ (lower right). The shown distributions correspond to $m_{a_1(Z_D)} = 20$ GeV in the hadron-hadron category for the 2018 analysis. The red shadow represents the fit error band within 2-sigma of confidence interval.

APPENDIX

B

EXCLUSION LIMITS

Table B.1: Expected and observed upper limits at 95% confidence level on the signal cross-section times the branching ratio $\sigma(pp \to H(125)+X)\cdot\mathcal{B}(H(125) \to a_1a_1(Z_DZ_D) \to \mu\mu\tau\tau)\times 10^{-4}$ relative to the inclusive cross-section $\sigma(pp \to H(125)+X)_{\mathrm{SM}}$ predicted in the SM, as a function of the mass of the light boson $m_{a_1(Z_D)}$.

Run II combination of results						
$m_{a_1(Z_D)}$ [GeV]	-2σ	-1σ	Median	+1σ	+2σ	Observed
3.60	0.96	1.30	1.84	2.66	3.72	1.89
3.80	0.94	1.27	1.80	2.60	3.65	1.48
4.00	1.01	1.36	1.93	2.78	3.90	2.19
4.20	0.81	1.10	1.57	2.29	3.25	2.15
4.40	0.83	1.12	1.59	2.31	3.25	1.14
4.60	0.74	0.99	1.42	2.08	2.95	1.87
4.80	0.73	0.99	1.43	2.10	3.01	1.76
5.00	0.75	1.00	1.44	2.09	2.97	1.19
5.20	0.54	0.75	1.10	1.65	2.41	1.39
5.40	0.67	0.91	1.30	1.92	2.74	1.18
5.60	0.62	0.84	1.22	1.79	2.56	2.26

Continued on next page

Table B.1 – *Continued from previous page*

$m_{a_1(Z_D)}$ [GeV]	-2σ	-1σ	Median	+1σ	+2σ	Observed
5.80	0.79	1.06	1.51	2.21	3.12	1.60
6.00	0.76	1.02	1.46	2.14	3.01	0.70
6.20	0.76	1.02	1.46	2.14	3.03	0.86
6.40	0.74	0.99	1.42	2.05	2.90	1.06
6.60	0.70	0.95	1.35	1.97	2.78	1.40
6.80	0.64	0.87	1.25	1.84	2.62	1.32
7.00	0.66	0.90	1.28	1.88	2.66	1.29
7.20	0.63	0.86	1.23	1.79	2.55	1.33
7.40	0.70	0.95	1.35	1.97	2.78	1.17
7.60	0.57	0.77	1.11	1.63	2.32	0.75
7.80	0.59	0.80	1.14	1.68	2.39	0.70
8.00	0.53	0.72	1.03	1.52	2.18	0.91
8.20	0.55	0.75	1.07	1.58	2.26	1.61
8.40	0.51	0.68	0.99	1.45	2.08	1.15
8.60	0.60	0.81	1.17	1.70	2.42	1.93
8.80	0.59	0.81	1.16	1.71	2.44	2.29
9.00	0.59	0.80	1.15	1.68	2.38	2.15
9.20	0.64	0.86	1.24	1.80	2.56	0.93
9.40	0.77	1.04	1.48	2.17	3.08	0.67
9.60	0.67	0.91	1.31	1.92	2.75	0.90
9.80	0.53	0.74	1.07	1.59	2.28	0.64
10.00	0.73	1.00	1.42	2.08	2.95	0.69
10.20	0.67	0.91	1.30	1.91	2.70	0.74
10.40	0.63	0.85	1.23	1.80	2.56	0.85
10.60	0.53	0.72	1.03	1.51	2.16	1.38
10.80	0.52	0.70	1.00	1.46	2.07	1.38
11.00	0.53	0.73	1.04	1.52	2.15	1.75
11.20	0.49	0.67	0.97	1.42	2.02	1.34

Continued on next page

Table B.1 – *Continued from previous page*

$m_{a_1(Z_D)}$ [GeV]	-2σ	-1σ	Median	+1σ	+2σ	Observed
11.40	0.48	0.65	0.94	1.39	1.98	1.33
11.60	0.49	0.66	0.94	1.39	1.97	1.35
11.80	0.54	0.73	1.04	1.52	2.16	1.77
12.00	0.48	0.65	0.94	1.38	1.97	2.04
12.20	0.45	0.62	0.89	1.32	1.88	0.80
12.40	0.48	0.65	0.94	1.37	1.96	0.57
12.60	0.49	0.66	0.94	1.38	1.97	0.80
12.80	0.48	0.65	0.93	1.36	1.94	0.88
13.00	0.49	0.66	0.94	1.38	1.97	1.08
13.20	0.48	0.65	0.94	1.37	1.96	1.18
13.40	0.48	0.65	0.93	1.36	1.94	1.24
13.60	0.45	0.61	0.88	1.30	1.87	0.71
13.80	0.51	0.69	0.99	1.45	2.05	0.96
14.00	0.51	0.70	1.00	1.47	2.09	1.15
14.20	0.48	0.65	0.94	1.39	1.98	1.19
14.40	0.50	0.68	0.98	1.44	2.06	1.13
14.60	0.47	0.64	0.93	1.38	1.98	1.25
14.80	0.44	0.59	0.85	1.27	1.82	1.14
15.00	0.53	0.72	1.03	1.51	2.15	1.54
15.20	0.50	0.68	0.97	1.43	2.05	1.42
15.40	0.47	0.64	0.92	1.37	1.96	1.08
15.60	0.51	0.69	0.99	1.45	2.07	0.80
15.80	0.57	0.77	1.10	1.60	2.27	1.27
16.00	0.59	0.80	1.15	1.67	2.36	1.03
16.20	0.55	0.75	1.08	1.59	2.26	1.51
16.40	0.57	0.78	1.12	1.65	2.34	1.04
16.60	0.52	0.71	1.03	1.52	2.17	0.72
16.80	0.57	0.77	1.11	1.63	2.31	0.79

Continued on next page

Table B.1 – *Continued from previous page*

$m_{a_1(Z_D)}$ [GeV]	-2σ	-1σ	Median	+1σ	+2σ	Observed
17.00	0.58	0.78	1.12	1.64	2.32	1.01
17.20	0.55	0.75	1.07	1.58	2.24	1.17
17.40	0.52	0.71	1.02	1.50	2.15	1.64
17.60	0.57	0.78	1.11	1.62	2.28	1.43
17.80	0.51	0.70	1.01	1.49	2.14	1.43
18.00	0.56	0.76	1.09	1.60	2.28	1.55
18.20	0.59	0.79	1.14	1.66	2.36	1.50
18.40	0.59	0.80	1.14	1.67	2.37	1.34
18.60	0.64	0.86	1.24	1.81	2.55	0.79
18.80	0.57	0.77	1.12	1.64	2.34	1.42
19.00	0.55	0.75	1.07	1.57	2.23	1.92
19.20	0.53	0.73	1.04	1.54	2.20	0.74
19.40	0.59	0.81	1.15	1.69	2.39	0.80
19.60	0.55	0.75	1.08	1.58	2.26	0.84
19.80	0.66	0.90	1.31	1.95	2.81	0.96
20.00	0.56	0.76	1.09	1.60	2.29	1.26
20.20	0.58	0.79	1.14	1.68	2.40	0.91
20.40	0.48	0.67	1.00	1.52	2.21	0.76
20.60	0.50	0.68	0.99	1.48	2.13	0.74
20.80	0.55	0.75	1.07	1.58	2.24	1.63
21.00	0.56	0.76	1.10	1.61	2.29	1.90

APPENDIX

C

MONTE CARLO MISALIGNMENT SCENARIOS FOR RUN II SIMULATION

Contents

C.1 Alignment of the CMS tracker

A good tracking performance results of particular relevance for the analysis presented in this thesis, which makes use of tracks and low-momentum muons. The measurement of the momentum of the particles depends on an accurate determination of the track curvature induced by the magnetic field, and the latter requires a carefully geometrically calibrated detector. The set of parameters that describe the geometrical properties of the tracker modules is known as *tracker geometry*. The accuracy in the knowledge of the geometrical position of the different tracker substructures upon installation does not reach the intrinsic resolution of the tracker sensors, of 10 to 30 μm [227]. This, together with the movements of the different substructures driven by the operation conditions during data taking, makes it necessary to correct the position, orientation, and curvature of the tracker sensors in a process known as

alignment of the CMS tracker. Once the mounting precision after assembly is reached, the alignment of the sensors is refined up to the order of their intrinsic hit resolution through a track-based alignment. The tracking performance relies on the precision of this alignment procedure, performed several times during data taking to detect and correct time-dependent effects. The limited knowledge of the position and orientation of the individual sensors due to misalignment effects is quantitatively assessed through the *alignment position errors* (APEs), which combined with the intrinsic hit resolution gives the total error of the hit position in the silicon modules. The APEs constitute a limiting factor for the performance and efficiency of track reconstruction, the track quality (χ^2), fake rate, momentum resolution, and vertexing resolution. The muon reconstruction is influenced by the tracker alignment since the tracks reconstructed in the tracker are extrapolated to the muon system for muon identification. The effect can be observed in the distribution of the reconstructed Z boson mass as a function of the azimuthal angle Φ of the positively and negatively charged muons, shown in Fig. C.1. The dependence of the reconstructed mass on the angular variables is reduced upon refinement of the alignment calibration. In addition, the alignment of the muon chambers is done relative to the tracker. Therefore, the alignment precision of the muon system is influenced by the alignment of the CMS tracker.

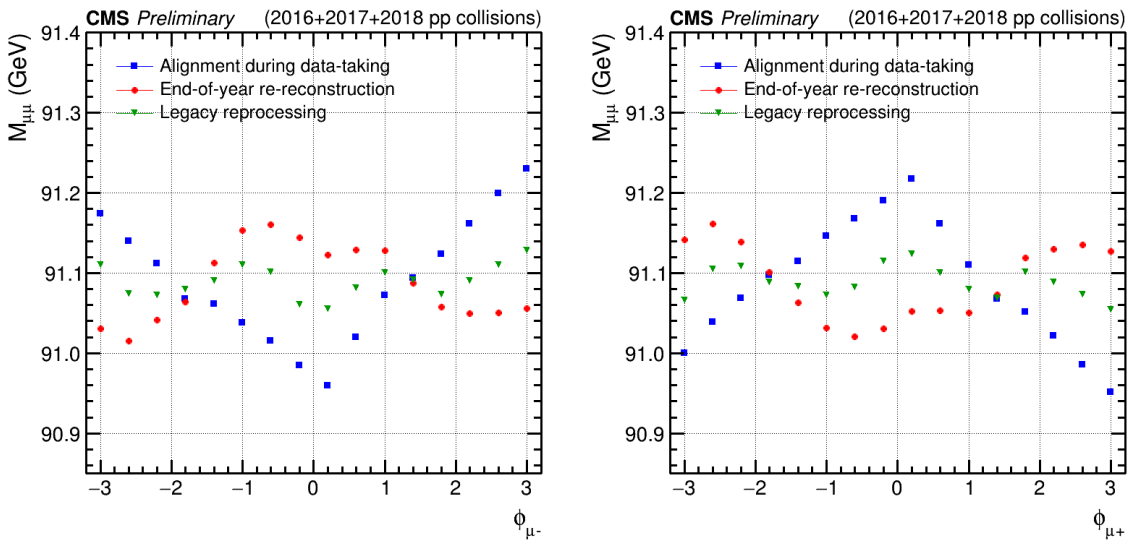


Figure C.1: Reconstructed $Z \to \mu\mu$ mass as a function of the azimuthal angle Φ of the negatively (left) and positively (right) charged muons. The blue, red, and green points show the alignment used during data taking, the end-of-year reconstruction, and the Run II legacy reprocessing, respectively. An improvement in the uniformity of the reconstructed $Z \to \mu\mu$ mass is observed in the Run II legacy reprocessing [228].

C.1.1 Track-based alignment

The track-based alignment of the CMS tracker constitutes a major challenge due to the enormous number of degrees of freedom involved. To every hit i registered in the detector modules a measured hit position (m_i) is assigned. A set of track parameters ($\mathbf{q}_j$) is associated

to the tracks formed from the combination of multiple hits. The fitted tracks depend on a set of so-called *alignment* or *global* parameters ($\mathbf{p}$) that define the space coordinates, orientation, and deformation of the different tracker components. Nine alignment parameters per sensor are needed to set the coordinates of the sensor center, the rotational angles, and the surface deformations. Considering the number of sensors of the tracker, the alignment parameters to be derived in order to fully align the detector may exceed 2×10^5, which requires a large number of tracks. The description of the trajectory of each of these tracks in the magnetic field makes use of $n_{\text{par}} = 5 + 2n_{\text{scat}}$ track parameters, where n_{scat} represents the number of scatterings undergone by the track in the detector material. This parametrization leads to a high number of parameters per track (e.g., $n_{\text{par}} > 50$ for cosmic ray tracks).
The track-based alignment method follows a least square approach, minimising the sum of squares of the normalised track-hit residuals from a set of tracks. A track hit-residual r_{ij} is obtained subtracting the projection of the track prediction f_{ij} from the measured hit position m_{ij}:

$$r_{ij}(\mathbf{p}, \mathbf{q}_j) = m_{ij} - f_{ij}(\mathbf{p}, \mathbf{q}_j), \tag{C.1}$$

and tends to become broader if the assumed geometry differs from the real one. The χ^2 function to be minimised takes the form:

$$\chi^2(\mathbf{p}, \mathbf{q}) = \sum_{j}^{\text{tracks}} \sum_{i}^{\text{measurements}} \left(\frac{m_{ij} - f_{ij}(\mathbf{p}, \mathbf{q}_j)}{\sigma_{ij}} \right)^2, \tag{C.2}$$

where σ_{ij} is the uncertainty of the measured hit position m_{ij} [229]. The hit position f_{ij} predicted by the track fit model depends on the assumed geometry ($\mathbf{p}$) and track parameters ($\mathbf{q}_j$). Fig. C.2 depicts the effect of assuming an incorrect position for a module on the reconstruction of one track, resulting in a displacement $\mathbf{r}$ between the measured hit and the fit of the real track. Since the alignment corrections are assumed to be small, the trajectory

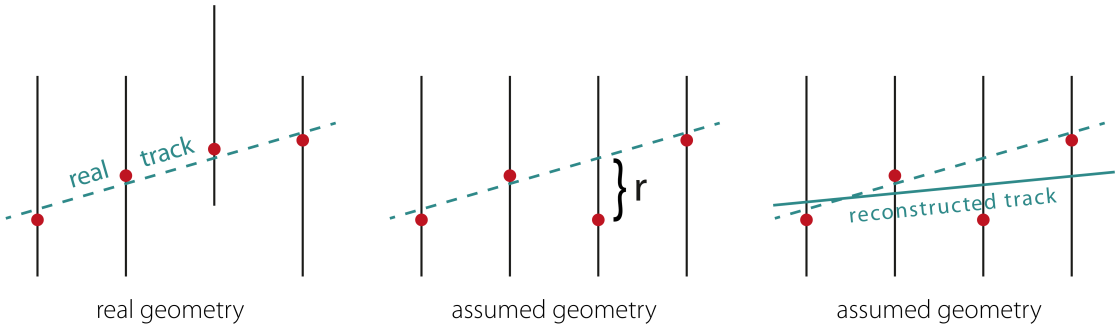


Figure C.2: Sketch representing the effect in the track reconstruction of an incorrect assumption on the position of one module. [230].

prediction can be linearised around an initial value that is usually available from surveys during assembly, design drawings, and previous alignment results. Given an initial geometry $\mathbf{p}_0$ and track parameters $\mathbf{q}_{0j}$, the χ^2 function can be expressed as:

$$\chi^2(\mathbf{p}, \mathbf{q}) \simeq \sum_{j}^{\text{tracks}} \sum_{i}^{\text{measurements}} \frac{1}{\sigma_{ij}^2} \left(m_{ij} - [f_{ij}(\mathbf{p}_0, \mathbf{q}_{0j}) + \frac{\partial f_{ij}}{\partial \mathbf{p}} \Delta\mathbf{p} + \frac{\partial f_{ij}}{\partial \mathbf{q}_j} \Delta\mathbf{q}_j] \right)^2. \tag{C.3}$$

The minimization leads to the linear system of equations $\mathbf{Ca} = \mathbf{b}$, with $\mathbf{a}^T = (\Delta\mathbf{p}, \Delta\mathbf{q})$, where $\Delta\mathbf{p}$ represents the corrections introduced to the initial geometry and $\Delta\mathbf{q}$ the corrections to the track parameters ($\Delta\mathbf{q}^T = (\Delta\mathbf{q}_1, ..., \Delta\mathbf{q}_n)$). The procedure is iterated several times for non-small alignment corrections. The system of equations is reduced to a smaller set of equations for the alignment parameters only and solved through a *global fit* [231], using the MILLE-PEDE II program [232]. After the alignment, the tracks are re-fitted with the new geometry (near to the real one) and the measurement of the track momentum is updated.
The alignment procedure can be carried out for different levels of complexity, making use of the hierarchical structure of the tracker. For instance, in the pixel barrel, individual modules are attached to ladders, which are further attached to half-shells. The half-shells form a half-barrel, and finally, two half-barrels form the entire barrel subdetector. All the individual modules can be aligned in a so-called module-level alignment or just the high-level structures, known as high-level alignment. In a high-level alignment, the high-level structures are treated as composite objects consisting of a set of individual modules and are aligned as a rigid body, i.e., the relative position of the individual modules within the composite objects is fixed and just the overall movements of the structures are considered [233]. The alignment at the highest hierarchy level is mainly performed in cases in which the number of tracks is not enough for a module-level alignment, the position uncertainty of the composite objects is much larger than the intrinsic resolution of the modules (e.g., after installation), or to monitor time-dependent distortions of the composite object that do not influence the relative position of individual modules.

C.1.2 Weak modes and bias

The described alignment algorithm aims to find the real detector geometry by minimizing the χ^2 of the track-hit residuals, but often the modules of the detector can be moved coherently ending with very different geometries and identical χ^2. The linear combinations of parameters that leave invariant the track-hit residuals and thus the χ^2 in Eq. (C.2) are known as *weak modes*. The cylindrical geometry of the CMS tracker results in the following set of weak modes:

- **radial**: misalignment in the Δr-direction as a function of r.
- **telescope**: misalignment in the Δz-direction as a function of r (offset of concentric rings in the longitudinal direction).
- **layer rotation**: misalignment in the $r\Delta\Phi$-direction as a function of r.
- **bowing**: misalignment in the Δr-direction as a function of z.
- **z-expansion (or contraction)**: misalignment in the Δz-direction as a function of z.
- **twist**: misalignment in the $r\Delta\Phi$-direction as a function of z (bias in the curvature of tracks).
- **elliptical**: misalignment in the Δr-direction as a function of Φ.

- **skew**: misalignment in the Δz-direction as a function of Φ.
- **sagitta**: misalignment in the $r\Delta\Phi$-direction as a function of Φ (off-centring of the barrel layers and endcap rings).

The characteristic module displacements Δr, Δz, and $r\Delta\Phi$ introduced by this set of global χ^2-invariant distortions are depicted in Fig. C.3 as a function of r, z, and Φ. Their control constitutes an important component in the strategy of the CMS alignment campaigns.
The movements in the directions of the weak modes are unconstrained. Tracks whose χ^2 is sensitive to them need to be included in the alignment procedure in order to reduce their effect. Therefore, a heterogeneous sample of tracks that cross the detector at different angles, cover their full active area, and relate the different detector components is selected to be used for the alignment procedure. The introduction of constraints in the χ^2 minimization helps to further reduce the effect of weak modes. Some examples are the constraints on the mass of resonances, the E/p ratio from calorimeter information, the use of hits that overlap in adjacent modules, and information on the mechanical structure of the detector.

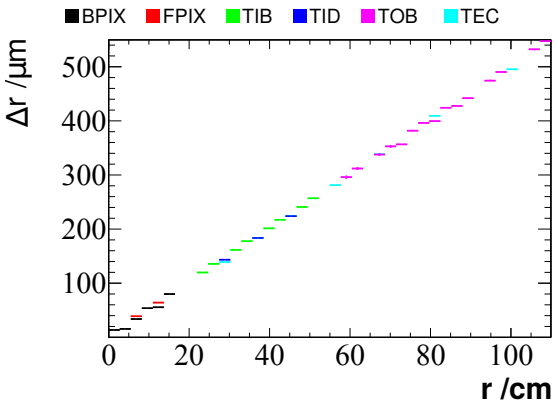


Fig. C.3 a)

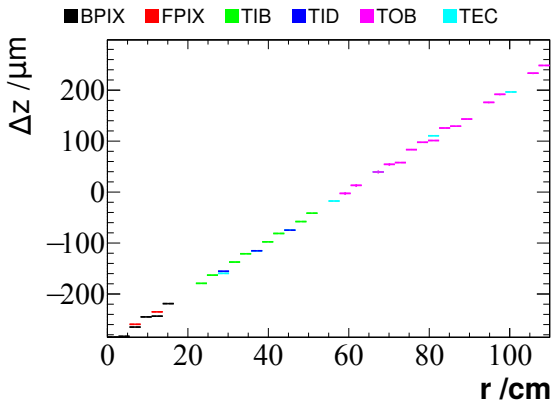


Fig. C.3 b)

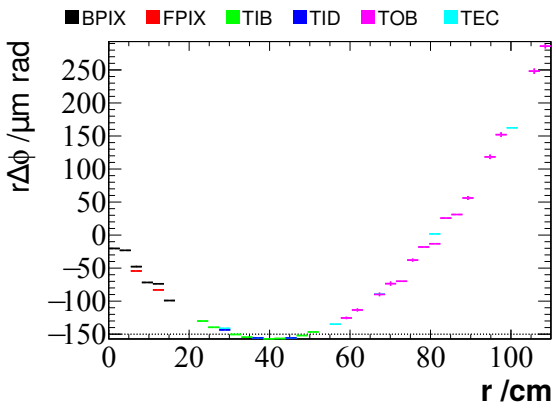


Fig. C.3 c)

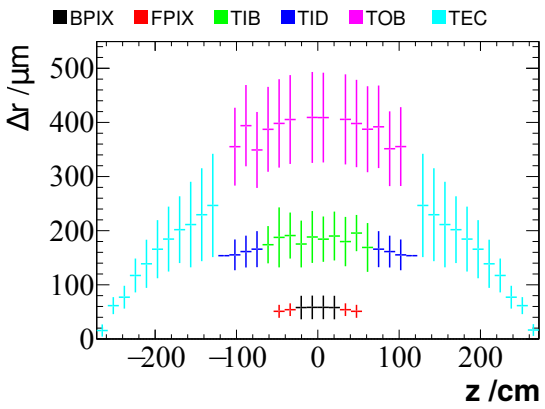


Fig. C.3 d)

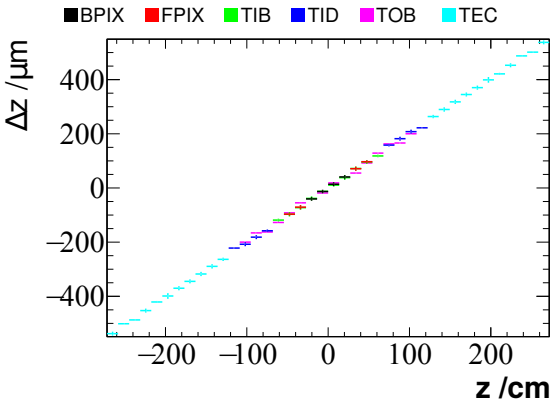


Fig. C.3 e)

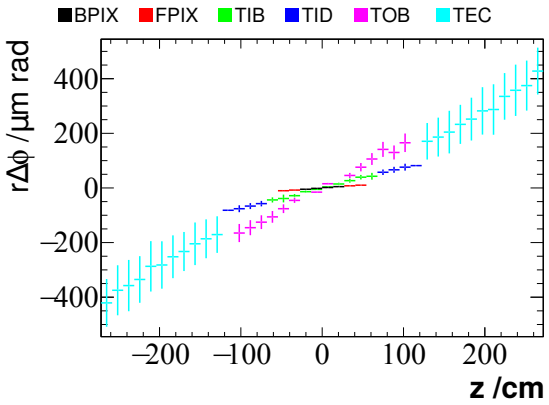


Fig. C.3 f)

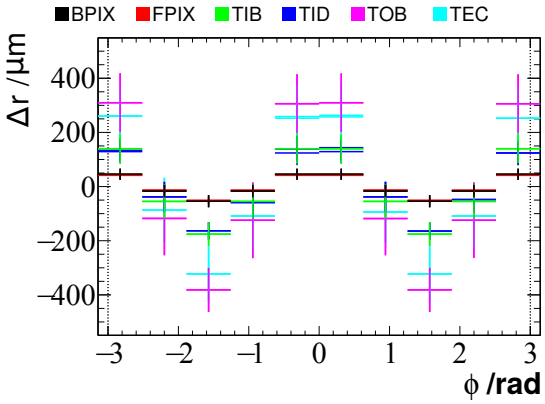


Fig. C.3 g)

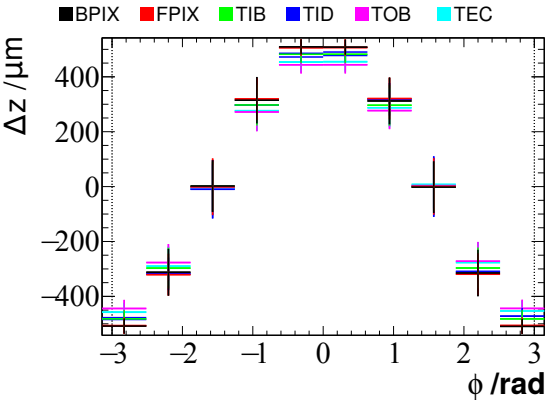


Fig. C.3 h)

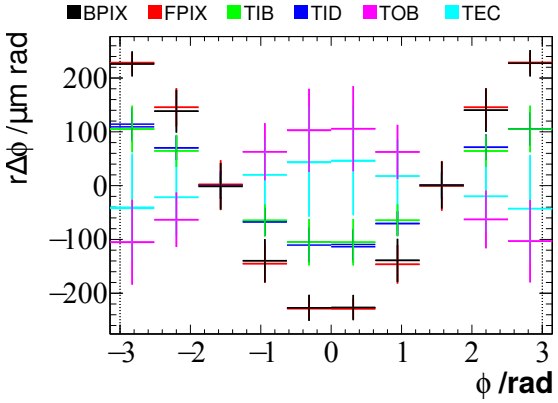


Fig. C.3 i)

Figure C.3: χ^2-invariant global distortions: a) radial, b) telescope, c) layer rotation, d) bowing, e) z-expansion, f) twist, g) elliptical, h) skew, and i) sagitta. The alignment procedure must be carefully tested against these systematic misalignments in order to insure one unique and stable solution.

In addition to weak modes, biases in the alignment calibration can arise from local reconstruction effects, the tracking, or an inaccurate knowledge of the magnetic field. An example of this kind of bias is the one caused by an incorrect determination of the Lorentz drift of charge carriers released by charged particles that pass through the silicon sensors. The pres-

ence of the magnetic field deflects the trajectory of the charges in the sensors by a Lorentz Angle θ_{LA} and, therefore, the collected charge is biased by a ΔX_{LA}:

$$\Delta X_{\mathrm{LA}} = w \cdot \tan\theta_{\mathrm{LA}}, \qquad \text{with} \qquad \theta_{\mathrm{LA}} = \mu_{\mathrm{H}} B, \tag{C.4}$$

where w is the sensor thickness and μ_{H} the charge mobility [234]. If the local reconstruction assumes an incorrect Lorentz angle, the hit positions determined from the collected charges in the affected sensors will be systematically biased. The alignment tries to compensate the ΔX shift by moving coherently the modules, which results in a biased geometry. This kind of effect does not constitute a weak mode since the biased position is really the one that minimizes the (biased) χ^2.

C.2 Tracker Alignment strategy for data and simulation in Run II

During data taking, CMS continuously monitors the high-level-structure movements of the pixel tracker with the prompt calibration loop (PCL) and automatically corrects the geometry if the alignment corrections exceed certain thresholds. A track-based alignment is periodically run offline to refine with a few manual updates the PCL alignment. The full statistics of the dataset collected during one year is exploited to provide a set of alignment conditions for the reprocessing of the data at the end of the year. The alignment calibrations are carried out individually for each interval of time during which they retain the same values, known as an interval of validity (IOV). To obtain the ultimate accuracy of the alignment calibration for the final or *ultra-legacy* (UL) reprocessing of the Run II data, the conditions derived at the end of each data-taking year, known as *end-of-year* (EOY) alignment, are used as starting geometry. CMS processes the simulated events through the same reconstruction chain that is used for data. The accuracy of the simulation in describing the data strongly relies on realistic reconstruction conditions. Thus, the full set of detector calibrations, including the tracker alignment conditions, needs to be derived also for the processing of the simulated events. Currently, one main difference between the data and MC conditions is that the latter does not include time dependence within one data-taking year. Therefore, the MC condition should reasonably reproduce the average performance observed in data during the year.

C.3 Ultra-Legacy MC Misalignment Scenarios for 2016, 2017, and 2018

In the context of this thesis, particular emphasis has been put on deriving realistic misalignment scenarios to be used in the ultra-legacy reconstruction of simulated events. The scenarios are derived separately for each data-taking year (2016, 2017, and 2018), and are tuned to emulate the effects of the residual misalignment left in data after the ultra-legacy alignment. The original contribution of this thesis to the derivation of the Run II MC misalignment scenarios is described hereafter.

C.3.1 Generation of the misaligned geometry

For the creation of the MC scenarios, a similar alignment strategy to the one used for the data alignment is followed. A Gaussian smearing consisting of random movements that follow a gaussian distribution with σ equal to the root mean square (RMS) of the distribution of track-hit residuals in each of the components corresponding to the EOY alignment in data is applied to the design tracker geometry. For the period of 2017, an additional systematic movement along the z-coordinate (± 30 μm) is applied to the high-level structures of the FPIX, to mimic an apparent systematic misalignment in the forward region of the tracker. The smeared geometry with the additional systematic movement emulates the corresponding EOY data alignment conditions and constitutes the starting geometry for the derivation of the 2017 ultra-legacy MC misalignment scenario. This scenario is used hereafter as a reference to illustrate the procedure also followed for the derivation of the 2016 and 2018 MC objects. Tab. C.1 reports the RMS values applied to each component of the tracker system for the 2017 scenario.

Substructure	RMS [μm] in local x-direction	RMS [μm] in local y-direction
BPIX	6.1	17.0
FPIX	5.3	2.7
TIB	13.7	13.7
TOB	30.9	30.9
TID	6.3	6.3
TEC	13.6	13.6

Table C.1: RMS values applied as Gaussian smearing to the design tracker geometry for the 2017 MC ultra-legacy campaign. The values are reported for each high-level structure and coordinate. The values of the y-coordinate that correspond to the substructures of the strip detector (not available from the 2017 EOY alignment) are taken as equal to those in the x-coordinate.

A module-level alignment is performed using MC samples of similar topology to those of the data alignment. For the MC samples with lower statistics with respect to the corresponding data samples, the contribution of each topology is tuned according to its real contribution in the data alignment by means of events weights, making the proportion of tracks of a given topology compatible with the UL data campaign.

C.3.2 Comparison with performance of the data alignment

The goodness of the MC conditions in describing the data is evaluated comparing distributions of quantities sensitive to the tracker alignment calibration. Three IOVs were chosen to be representatives of the conditions in the detector. The IOVs are associated with the runs *299370* (Data 18 July), *301417* (Data 18 August), and *304505* (Data 05 October), respectively. The set of validations performed are shown in Figs. C.4 to C.6.
The first validation in Fig. C.4 shows the distribution of the median of the unbiased track-hit residuals per module. The so-called local coordinates refer to the coordinates defined for

each module with the origin at the geometric center of the active area of the module. In the computation of the residuals, each track is refitted using the alignment constants under consideration without using the hit in question, resulting in unbiased track-hit residuals. The width of the distribution of the medians of residuals (DMR) constitutes a measure of the local precision of the alignment, with deviations from zero indicating possible biases. The width has an intrinsic component due to the limited number of tracks used for the alignment procedure. Therefore, the distributions can only be compared if they are produced with the same number of tracks, as is the case for both data and MC in the shown results.

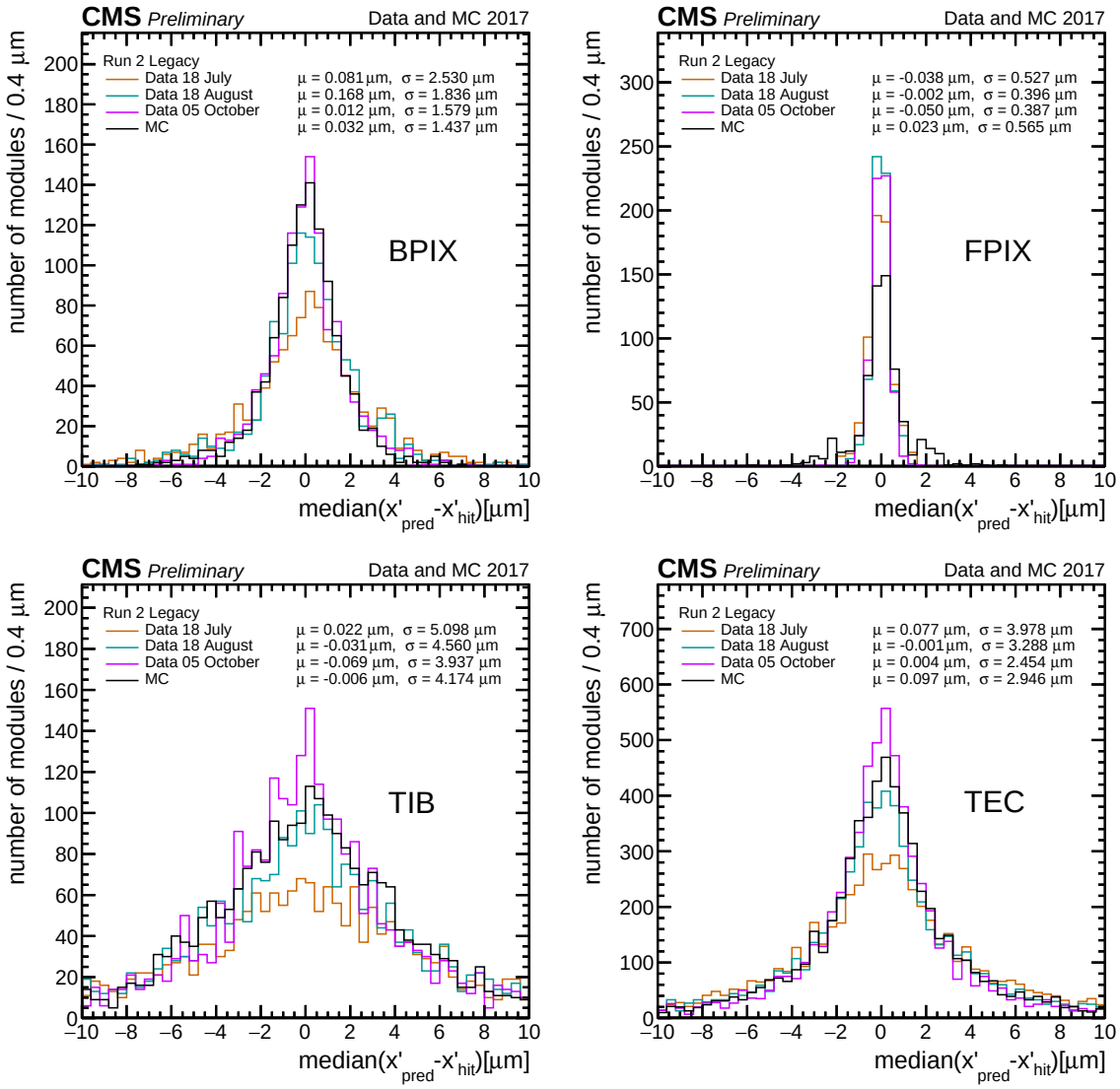


Figure C.4: The distribution of median residuals is plotted for the local-x coordinate in the barrel pixel (upper left), forward pixel (upper right), tracker inner barrel (lower left), and tracker endcaps (lower right). The orange, blue, and purple distributions show the performance of three IOVs corresponding to the 2017 ultra-legacy data alignment, derived using 3.8T cosmic ray and collision data from the 2017 proton-proton run. The black distribution shows the performance of the MC misalignment scenario. The quoted means μ and standard deviations σ correspond to the parameters of a Gaussian fit to the distributions [235].

The performance of the MC objects is quite balanced compared to the data IOVs; this tendency is observed in all the substructures. The larger width in the forward pixel is driven by the systematic misalignment of 30 μm applied in the beam direction to better describe the trends observed in the data alignment (right plot, Fig. C.5).
Another important test is the Primary Vertex Validation that studies the distance between the track and the vertex reconstructed without the track under scrutiny (unbiased track-vertex residual). This validation mainly evaluates the performance of the alignment in the pixel detector. The mean value of the unbiased track-vertex residuals in the transverse plane and the longitudinal direction extracted in bins of the azimuthal angle of the probe track are shown in Fig. C.5. Random misalignment of the modules may affect the resolution of the unbiased track-vertex residuals, increasing the width of the distribution without introducing a bias to their mean, while systematic movements of the modules can bias the distribution depending on the type and size of the misalignment and the topology of the tracks used for the alignment procedure. The MC object has comparable performance to the data IOVs.

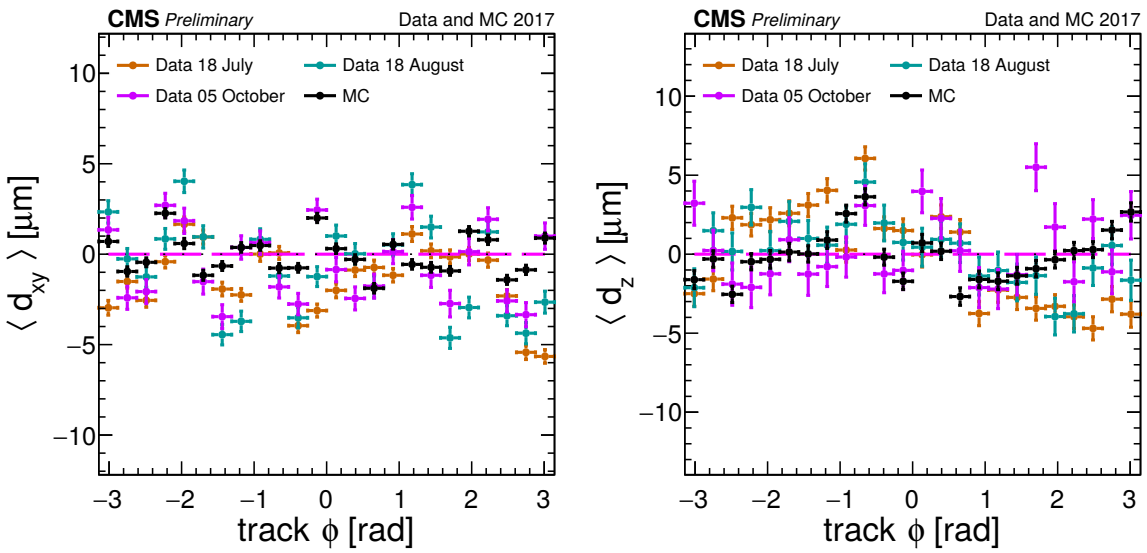


Figure C.5: The mean distance in the transverse plane d_{xy} (left) and longitudinal direction d_z (right) of the tracks at their point of closest approach to a refit unbiased primary vertex is studied in bins of the track azimuthal angle Φ. The orange, blue, and purple distributions show the performance of three IOVs corresponding to the 2017 ultra-legacy data alignment. The black distribution shows the performance of the MC misalignment scenario [235].

A further test known as track-split validation allows us to detect systematic misalignments in the form of the off-centring of the barrel layers and endcap rings. The cosmic track-splitting method uses cosmic tracks that pass through the pixel volume and splits them into two halves at their point of closest approach to the beamline. Both halves are treated as independent in the process of reconstruction. The validation checks for differences in the kinematic distributions of the top and bottom halves, as shown in Fig. C.6 for the impact parameter in the transverse plane.

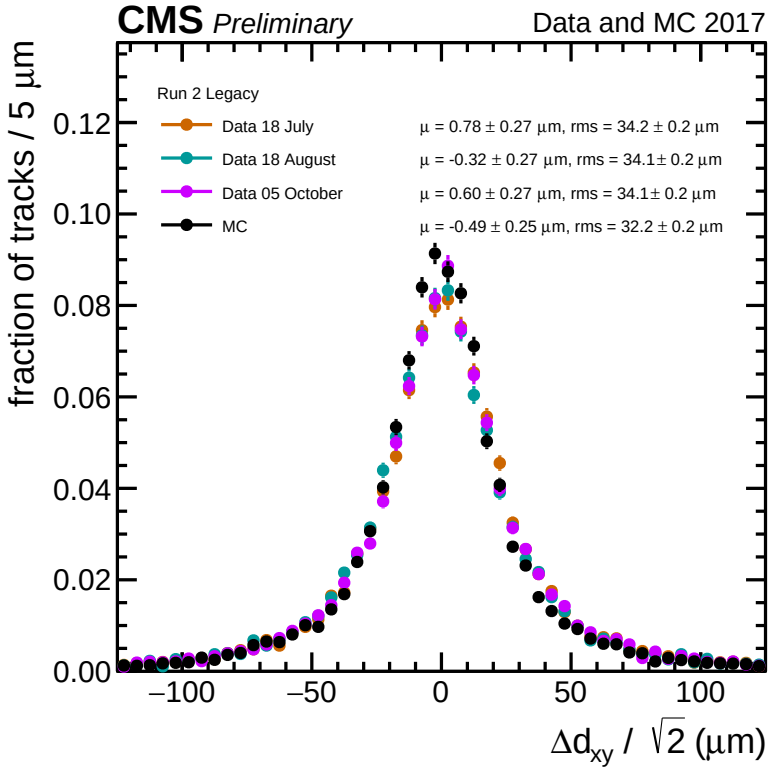


Figure C.6: Difference in impact parameter in the transverse plane d_{xy} between the top and bottom halves of cosmic tracks recorded with the CMS magnet at 3.8T. The orange, blue, and purple distributions show the performance of three IOVs corresponding to the 2017 ultra-legacy data alignment. The black distribution shows the performance of the MC misalignment scenario [235].

The set of validations performed allow us to conclude that the object derived for the 2017 MC misalignment scenario is consistent with the conditions observed in the real detector, therefore, it constitutes the final candidate for the tracker geometry used in the simulation of the CMS detector for the 2017-period ultra-legacy reprocessing.
In this appendix, the procedure used to obtain the misalignment scenarios for the Run II MC simulation has been described. The misalignment tools are available for the derivation of the MC objects in future data-taking periods and to study the impact of misalignment on physics measurements. The demanding operation conditions and the delivered luminosity expected at the HL-LHC will present new challenges to the alignment of the CMS tracker.

BIBLIOGRAPHY

[1] UA1 Collaboration, "Experimental Observation of Isolated Large Transverse Energy Electrons with Associated Missing Energy at $\sqrt{s} = 540$-GeV", *Phys. Lett. B* **122** (1983) 103–116, doi:10.1016/0370-2693(83)91177-2. 1

[2] UA2 Collaboration, "Observation of Single Isolated Electrons of High Transverse Momentum in Events with Missing Transverse Energy at the CERN anti-p p Collider", *Phys. Lett. B* **122** (1983) 476–485, doi:10.1016/0370-2693(83)91605-2. 1

[3] UA1 Collaboration, "Experimental Observation of Lepton Pairs of Invariant Mass Around 95-GeV/c^2 at the CERN SPS Collider", *Phys. Lett. B* **126** (1983) 398–410, doi:10.1016/0370-2693(83)90188-0. 1

[4] UA2 Collaboration, "Evidence for $Z^0 \to e^+e^-$ at the CERN $\bar{p}p$ Collider", *Phys. Lett. B* **129** (1983) 130–140, doi:10.1016/0370-2693(83)90744-X. 1

[5] C. Campagnari and M. Franklin, "The Discovery of the top quark", *Rev. Mod. Phys.* **69** (1997) 137–212, doi:10.1103/RevModPhys.69.137, arXiv:hep-ex/9608003. 1

[6] DONUT Collaboration, "Observation of tau neutrino interactions", *Phys. Lett. B* **504** (2001) 218–224, doi:10.1016/S0370-2693(01)00307-0, arXiv:hep-ex/0012035. 1

[7] L. Evans and P. Bryant, "LHC Machine", *JINST* **3** (2008) S08001, doi:10.1088/1748-0221/3/08/S08001. 2, 38

[8] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Phys. Lett. B* **716** (2012) 1–29, doi:10.1016/j.physletb.2012.08.020, arXiv:1207.7214. 2

[9] CMS Collaboration, "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC", *Phys. Lett. B* **716** (2012) 30–61, doi:10.1016/j.physletb.2012.08.021, arXiv:1207.7235. 2, 102

[10] CMS Collaboration, "Measurements of the Higgs boson width and anomalous HVV couplings from on-shell and off-shell production in the four-lepton final state", *Phys. Rev. D* **99** (2019), no. 11, 112003, doi:10.1103/PhysRevD.99.112003, arXiv:1901.00174. 2

[11] CMS Collaboration, "Measurement and interpretation of differential cross sections for Higgs boson production at $\sqrt{s} = 13$ TeV", *Phys. Lett. B* **792** (2019) 369–396, doi:10.1016/j.physletb.2019.03.059, arXiv:1812.06504. 2

[12] CMS Collaboration, "Combined measurements of Higgs boson couplings in proton–proton collisions at $\sqrt{s} = 13\,\text{TeV}$", *Eur. Phys. J. C* **79** (2019), no. 5, 421, doi:10.1140/epjc/s10052-019-6909-y, arXiv:1809.10733. 2

[13] CMS Collaboration, "Measurement of inclusive and differential Higgs boson production cross sections in the diphoton decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **01** (2019) 183, doi:10.1007/JHEP01(2019)183, arXiv:1807.03825. 2

[14] CMS Collaboration, "Measurements of properties of the Higgs boson decaying to a W boson pair in pp collisions at $\sqrt{s} = 13$ TeV", *Phys. Lett. B* **791** (2019) 96, doi:10.1016/j.physletb.2018.12.073, arXiv:1806.05246. 2

[15] CMS Collaboration, "Measurements of Higgs boson properties in the diphoton decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **11** (2018) 185, doi:10.1007/JHEP11(2018)185, arXiv:1804.02716. 2

[16] CMS Collaboration, "Measurements of properties of the Higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s} = 13$ TeV", *JHEP* **11** (2017) 047, doi:10.1007/JHEP11(2017)047, arXiv:1706.09936. 2

[17] ATLAS, CMS Collaboration, "Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV", *JHEP* **08** (2016) 045, doi:10.1007/JHEP08(2016)045, arXiv:1606.02266. 2, 30

[18] CMS Collaboration, "Measurement of the transverse momentum spectrum of the Higgs boson produced in pp collisions at $\sqrt{s} = 8$ TeV using $H \to WW$ decays", *JHEP* **03** (2017) 032, doi:10.1007/JHEP03(2017)032, arXiv:1606.01522. 2

[19] CMS Collaboration, "Measurement of differential and integrated fiducial cross sections for Higgs boson production in the four-lepton decay channel in pp collisions at $\sqrt{s} = 7$ and 8 TeV", *JHEP* **04** (2016) 005, doi:10.1007/JHEP04(2016)005, arXiv:1512.08377. 2

[20] CMS Collaboration, "Measurement of differential cross sections for Higgs boson production in the diphoton decay channel in pp collisions at $\sqrt{s} = 8$ TeV", *Eur.*

Phys. J. C **76** (2016), no. 1, 13, doi:10.1140/epjc/s10052-015-3853-3, arXiv:1508.07819. 2

[21] ATLAS, CMS Collaboration, "Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments", *Phys. Rev. Lett.* **114** (2015) 191803, doi:10.1103/PhysRevLett.114.191803, arXiv:1503.07589. 2

[22] CMS Collaboration, "Observation of the Diphoton Decay of the Higgs Boson and Measurement of Its Properties", *Eur. Phys. J. C* **74** (2014), no. 10, 3076, doi:10.1140/epjc/s10052-014-3076-z, arXiv:1407.0558. 2

[23] CMS Collaboration, "Measurement of the Properties of a Higgs Boson in the Four-Lepton Final State", *Phys. Rev. D* **89** (2014), no. 9, 092007, doi:10.1103/PhysRevD.89.092007, arXiv:1312.5353. 2

[24] CMS Collaboration, "Measurement of Higgs Boson Production and Properties in the WW Decay Channel with Leptonic Final States", *JHEP* **01** (2014) 096, doi:10.1007/JHEP01(2014)096, arXiv:1312.1129. 2

[25] G. Branco et al., "Theory and phenomenology of two-Higgs-doublet models", *Phys. Rept.* **516** (2012) 1–102, doi:10.1016/j.physrep.2012.02.002, arXiv:1106.0034. 2

[26] D. Curtin et al., "Exotic decays of the 125 GeV Higgs boson", *Phys. Rev. D* **90** (2014), no. 7, 075004, doi:10.1103/PhysRevD.90.075004, arXiv:1312.4992. 2, 26, 27, 28, 31, 34

[27] CMS Collaboration, "The CMS Experiment at the CERN LHC", *JINST* **3** (2008) S08004, doi:10.1088/1748-0221/3/08/S08004. 2, 54, 55

[28] W. Cottingham and D. Greenwood, "An introduction to the standard model of particle physics". Cambridge University Press, 4, 2007. 7

[29] M. B. Robinson, K. R. Bland, G. B. Cleaver, and J. R. Dittmann, "A Simple Introduction to Particle Physics. Part I - Foundations and the Standard Model", arXiv:0810.3328. 10, 11

[30] Stephen West, "The Standard Model". https://twiki.ph.rhul.ac.uk/twiki/pub/PP/Public/StephenWest/SM.pdf. Accessed: 2020-05-27. 11

[31] D. H. Perkins, "Introduction to high energy physics; 4th ed.". Cambridge Univ. Press, Cambridge, 2000. 12

[32] C. Burgess and G. Moore, "The standard model: A primer". Cambridge University Press, 12, 2006. 14

[33] M. M. Mühlleitner, "The Standard Model of Particle Physics". https://www.itp.kit.edu/~maggie/icise/standardmodel.pdf. Accessed: 2020-07-27. 14

[34] K. Nguyen, "The Higgs Mechanism". https://www.theorie.physik.uni-muenchen.de/lsfrey/teaching/archiv/sose_09/rng/higgs_mechanism.pdf. Accessed: 2020-06-10. 14, 15

[35] H. E. Logan, "TASI 2013 lectures on Higgs physics within and beyond the Standard Model", arXiv:1406.1786. 15

[36] G. Servant, "Introduction to HEP Theory". https://summerstudents.desy.de/sites2009/site_summerstudents/content/e69118/e246497/e246521/DESY_HEP_TH_Summer_students_02_08_2016.pdf. Accessed: 2020-06-10. 15

[37] A. Joseph, "Quantum Field Theory and the Standard Model". http://14.139.227.202/Faculty/anoshjoseph/courses/2020_even_qft2/lec47.pdf. Accessed: 2020-06-10. 16

[38] P. D. Group et al., "Review of Particle Physics", *Progress of Theoretical and Experimental Physics* **2020** (08, 2020) doi:10.1093/ptep/ptaa104. 083C01. 16, 19, 20, 21, 22, 23, 57

[39] J. Ellis, "Higgs Physics", in *2013 European School of High-Energy Physics*, pp. 117–168. 2015. arXiv:1312.5672. doi:10.5170/CERN-2015-004.117. 17

[40] K. Kumar, S. Mantry, W. Marciano, and P. Souder, "Low Energy Measurements of the Weak Mixing Angle", *Ann. Rev. Nucl. Part. Sci.* **63** (2013) 237–267, doi:10.1146/annurev-nucl-102212-170556, arXiv:1302.6263. 17

[41] M. Awramik, M. Czakon, and A. Freitas, "Electroweak two-loop corrections to the effective weak mixing angle", *JHEP* **11** (2006) 048, doi:10.1088/1126-6708/2006/11/048, arXiv:hep-ph/0608099. 17

[42] A. Djouadi, "The Anatomy of electro-weak symmetry breaking. I: The Higgs boson in the standard model", *Phys. Rept.* **457** (2008) 1–216, doi:10.1016/j.physrep.2007.10.004, arXiv:hep-ph/0503172. 21, 154

[43] Super-Kamiokande Collaboration, "Evidence for oscillation of atmospheric neutrinos", *Phys. Rev. Lett.* **81** (1998) 1562–1567, doi:10.1103/PhysRevLett.81.1562, arXiv:hep-ex/9807003. 23

[44] SNO Collaboration, "Direct evidence for neutrino flavor transformation from neutral current interactions in the Sudbury Neutrino Observatory", *Phys. Rev. Lett.* **89** (2002) 011301, doi:10.1103/PhysRevLett.89.011301, arXiv:nucl-ex/0204008. 23

[45] C. Csaki, "The Minimal supersymmetric standard model (MSSM)", *Mod. Phys. Lett. A* **11** (1996) 599, doi:10.1142/S021773239600062X, arXiv:hep-ph/9606414. 26

[46] U. Ellwanger, C. Hugonie, and A. M. Teixeira, "The Next-to-Minimal Supersymmetric Standard Model", *Phys. Rept.* **496** (2010) 1–77, doi:10.1016/j.physrep.2010.07.001, arXiv:0910.1785. 26

[47] M. Perelstein, "Little Higgs models and their phenomenology", *Prog. Part. Nucl. Phys.* **58** (2007) 247–291, doi:10.1016/j.ppnp.2006.04.001, arXiv:hep-ph/0512128. 26

[48] M. J. Strassler and K. M. Zurek, "Echoes of a hidden valley at hadron colliders", *Phys. Lett. B* **651** (2007) 374–379, doi:10.1016/j.physletb.2007.06.055, arXiv:hep-ph/0604261. 26

[49] S. Gopalakrishna, S. Jung, and J. D. Wells, "Higgs boson decays to four fermions through an abelian hidden sector", *Phys. Rev. D* **78** (2008) 055002, doi:10.1103/PhysRevD.78.055002, arXiv:0801.3456. 29, 34

[50] D. Curtin, R. Essig, S. Gori, and J. Shelton, "Illuminating Dark Photons with High-Energy Colliders", *JHEP* **02** (2015) 157, doi:10.1007/JHEP02(2015)157, arXiv:1412.0018. 30, 156

[51] M. Lisanti and J. G. Wacker, "Discovering the Higgs with Low Mass Muon Pairs", *Phys. Rev. D* **79** (2009) 115006, doi:10.1103/PhysRevD.79.115006, arXiv:0903.1377. 34, 35

[52] D0 Collaboration, "Search for NMSSM Higgs bosons in the h $\to$ aa $\to \mu\mu\mu\mu, \mu\mu\tau\tau$ channels using $p\bar{p}$ collisions at $\sqrt{s} = 1.96$-TeV", *Phys. Rev. Lett.* **103** (2009) 061801, doi:10.1103/PhysRevLett.103.061801, arXiv:0905.3381. 34

[53] CMS Collaboration, "Search for light bosons in decays of the 125 GeV Higgs boson in proton-proton collisions at $\sqrt{s} = 8$ TeV", *JHEP* **10** (2017) 076, doi:10.1007/JHEP10(2017)076, arXiv:1701.02032. 34, 35

[54] CMS Collaboration, "Search for an exotic decay of the Higgs boson to a pair of light pseudoscalars in the final state of two muons and two τ leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **11** (2018) 018, doi:10.1007/JHEP11(2018)018, arXiv:1805.04865. 34, 35

[55] CMS Collaboration, "Search for light pseudoscalar boson pairs produced from decays of the 125 GeV Higgs boson in final states with two muons and two nearby tracks in pp collisions at $\sqrt{s} = 13$ TeV", *Phys. Lett.* **B800** (2020) 135087, doi:10.1016/j.physletb.2019.135087, arXiv:1907.07235. 34, 35, 154

[56] CMS Collaboration, "Search for a light pseudoscalar Higgs boson in the boosted $\mu\mu\tau\tau$ final state in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **08** (2020) 139, doi:10.1007/JHEP08(2020)139, arXiv:2005.08694. 34, 35, 154

[57] ATLAS Collaboration, "Search for Higgs bosons decaying to *aa* in the $\mu\mu\tau\tau$ final state in *pp* collisions at $\sqrt{s} = 8$ TeV with the ATLAS experiment", *Phys. Rev. D* **92** (2015), no. 5, 052002, doi:10.1103/PhysRevD.92.052002, arXiv:1505.01609. 34, 35

[58] ATLAS Collaboration, "Search for Higgs boson decays to beyond-the-Standard-Model light bosons in four-lepton events with the ATLAS detector at $\sqrt{s} = 13$ TeV", *JHEP* **06** (2018) 166, doi:10.1007/JHEP06(2018)166, arXiv:1802.03388. 34, 36, 159

[59] CMS Collaboration, "Projection of searches for exotic Higgs boson decays to light pseudoscalars for the High-Luminosity LHC", 2019. 35

[60] CMS Collaboration, "Search for Light Resonances Decaying into Pairs of Muons as a Signal of New Physics", *JHEP* **07** (2011) 098, doi:10.1007/JHEP07(2011)098, arXiv:1106.2375. 35

[61] CMS Collaboration, "Search for a Non-Standard-Model Higgs Boson Decaying to a Pair of New Light Bosons in Four-Muon Final States", *Phys. Lett. B* **726** (2013) 564–586, doi:10.1016/j.physletb.2013.09.009, arXiv:1210.7619. 35

[62] CMS Collaboration, "A search for pair production of new light bosons decaying into muons in proton-proton collisions at 13 TeV", *Phys. Lett. B* **796** (2019) 131–154, doi:10.1016/j.physletb.2019.07.013, arXiv:1812.00380. 35, 159

[63] CMS Collaboration, "Search for a low-mass dilepton resonance in Higgs boson decays to four-lepton final states at $\sqrt{s} = 13$ TeV", (5, 2020). CMS-PAS-HIG-19-007. 36, 159

[64] C. Wyss, "LEP design report, v.3: LEP2". CERN, Geneva, 1996. Vol. 1-2 publ. in 1983-84. 38

[65] E. Mobs, "The CERN accelerator complex - 2019. Complexe des accélérateurs du CERN - 2019", (Jul, 2019). CERN-GRAPHICS-2019-002. 38

[66] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider", *JINST* **3** (2008) S08003, doi:10.1088/1748-0221/3/08/S08003. 38

[67] LHCb Collaboration, "The LHCb Detector at the LHC", *JINST* **3** (2008) S08005, doi:10.1088/1748-0221/3/08/S08005. 38

[68] ALICE Collaboration, "The ALICE experiment at the CERN LHC", *JINST* **3** (2008) S08002, doi:10.1088/1748-0221/3/08/S08002. 38

[69] LHCf Collaboration, "The LHCf detector at the CERN Large Hadron Collider", *JINST* **3** (2008) S08006, doi:10.1088/1748-0221/3/08/S08006. 38

[70] TOTEM Collaboration, "The TOTEM experiment at the CERN Large Hadron Collider", *JINST* **3** (2008) S08007, doi:10.1088/1748-0221/3/08/S08007. 38

[71] MoEDAL Collaboration, "Technical Design Report of the MoEDAL Experiment", Technical Report CERN-LHCC-2009-006. MoEDAL-TDR-001, CERN, Jun, 2009. 39

[72] "LHC performance reaches new highs". https://home.cern/news/news/accelerators/lhc-performance-reaches-new-highs. CERN news, 8 July 2016. 39

[73] "The LHC racks up records". https://home.cern/news/news/accelerators/lhc-racks-records. CERN news, 30 June 2017. 39

[74] "Record luminosity: well done LHC". `https://home.cern/news/news/accelerators/record-luminosity-well-done-lhc`. CERN news, 13 November 2017. 39

[75] CMS Collaboration, "CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV", 2019. CMS-PAS-LUM-18-002. 40, 145

[76] CMS Collaboration, "CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV", 2018. CMS-PAS-LUM-17-004. 40, 145

[77] CMS Collaboration, "CMS Luminosity Measurements for the 2016 Data Taking Period", 2017. CMS-PAS-LUM-17-001. 40, 145

[78] CMS Collaboration, "CMS luminosity measurement for the 2015 data-taking period", 2017. CMS-PAS-LUM-15-001. 40

[79] CMS Collaboration, "CMS Luminosity Based on Pixel Cluster Counting - Summer 2013 Update", 2013. CMS-PAS-LUM-13-001. 40

[80] CMS Collaboration, "CMS Luminosity Based on Pixel Cluster Counting - Summer 2012 Update", 2012. CMS-PAS-LUM-12-001. 40

[81] CMS Collaboration, "Integrated luminosity (Run I + Run II)". `https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults/`. Accessed: 2020-01-23. 40, 41

[82] CMS Collaboration, "Luminosity and uncertainty for Run II pp runs at $\sqrt{s} = 13$ TeV". `https://twiki.cern.ch/twiki/bin/view/CMS/TWikiLUM#SummaryTable`. Accessed: 2020-01-23. 41

[83] "First beam in the LHC - accelerating science". `https://home.cern/news/press-release/cern/first-beam-lhc-accelerating-science`. CERN press release, 10 September 2008. 40

[84] "Incident in LHC sector 3-4". `https://home.cern/news/press-release/cern/incident-lhc-sector-3-4`. CERN press release, 20 September 2008. 40

[85] "The LHC is back". `https://home.cern/news/press-release/cern/lhc-back`. CERN press release, 20 November 2009. 40

[86] "Two circulating beams bring first collisions in the LHC". `https://home.cern/news/press-release/cern/two-circulating-beams-bring-first-collisions-lhc`. CERN press release, 23 November 2009. 41

[87] "LHC sets new world record". `https://home.cern/news/press-release/cern/lhc-sets-new-world-record`. CERN press release, 30 November 2009. 41

[88] "LHC ends 2009 run on a high note". `https://home.cern/news/press-release/cern/lhc-ends-2009-run-high-note`. CERN press release, 18 December 2009. 41

[89] "LHC sets new record-accelerates beam to 3.5 TeV". https://home.cern/news/press-release/cern/lhc-sets-new-record-accelerates-beam-35-tev. CERN press release, 19 March 2010. 41

[90] "The LHC enters a new phase". https://home.cern/news/press-release/cern/lhc-enters-new-phase. CERN press release, 4 November 2010. 41

[91] "LHC physics data taking gets underway at new record collision energy of 8 TeV". https://home.cern/news/press-release/cern/lhc-physics-data-taking-gets-underway-new-record-collision-energy-8tev. CERN press release, 5 April 2012. 41

[92] "CERN experiments observe particle consistent with long-sought Higgs boson". https://home.cern/news/press-release/cern/cern-experiments-observe-particle-consistent-long-sought-higgs-boson. CERN press release, 4 July 2012. 41

[93] "The first LHC protons run ends with new milestone". https://home.cern/news/press-release/cern/first-lhc-protons-run-ends-new-milestone. CERN press release, 17 December 2012. 41

[94] "LHC experiments are back in business at a new record energy". https://home.cern/news/press-release/cern/lhc-experiments-are-back-business-new-record-energy. CERN press release, 3 June 2015. 41

[95] "CMS pixel tracker transplant: everything went well so far". https://home.cern/news/news/experiments/cms-pixel-tracker-transplant-everything-went-well-so-far. CERN news, 13 March 2017. 41

[96] "LHC Report: Another run is over and LS2 has just begun". https://home.cern/news/news/accelerators/lhc-report-another-run-over-and-ls2-has-just-begun. CERN news, 11 December 2018. 41

[97] "LHC prepares for new achievements". https://home.cern/news/press-release/accelerators/lhc-prepares-new-achievements. CERN press release, 3 December 2018. 41

[98] "A new schedule for the LHC and its successor". https://home.cern/news/news/accelerators/new-schedule-lhc-and-its-successor. CERN news, 13 December 2019. 41

[99] "LS2 Report: A new schedule". https://home.cern/news/news/accelerators/ls2-report-new-schedule. CERN news, 24 June 2020. 41

[100] A. Dainese et al., eds., "Report on the Physics at the HL-LHC,and Perspectives for the HE-LHC", volume 7/2019 of *CERN Yellow Reports: Monographs*. CERN, Geneva, Switzerland, 2019. 42

[101] CMS Collaboration, "The CMS electromagnetic calorimeter project: Technical Design Report", Technical Report CERN-LHCC-97-033, CERN, Geneva, 1997. 42

[102] CMS Collaboration, "The CMS hadron calorimeter project: Technical Design Report", Technical Report CERN-LHCC-97-031, CERN, Geneva, 1997. 42

[103] CMS Collaboration, "The CMS muon project: Technical Design Report", Technical Report CERN-LHCC-97-032, CERN, Geneva, 1997. 42

[104] CMS Collaboration, "Cutaway diagrams of CMS detector", (May, 2019). CMS-OUTREACH-2019-001. 43

[105] ATLAS Collaboration, "ATLAS magnet system: Technical Design Report, 1", Technical Report CERN-LHCC-97-018, CERN, Geneva, 1997. 43

[106] CMS Collaboration, G. Acquistapace et al., "The CMS magnet project: Technical Design Report". Number CERN-LHCC-97-0108 in Technical Design Report CMS. CERN, Geneva, 1997. 43

[107] G. Knoll, "Radiation Detection and Measurement (4th ed.)". John Wiley, Hoboken, NJ, 2010. 44

[108] CMS Collaboration, "Particle-flow reconstruction and global event description with the CMS detector", *JINST* **12** (2017), no. 10, P10003, doi:10.1088/1748-0221/12/10/P10003, arXiv:1706.04965. 44, 68, 75

[109] CMS Collaboration, "Interactive Slice of the CMS detector", (Aug, 2016). CMS-OUTREACH-2016-027. 45

[110] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", *JINST* **9** (2014), no. 10, P10009, doi:10.1088/1748-0221/9/10/P10009, arXiv:1405.6569. 44, 66

[111] CMS Collaboration, "The Phase-2 Upgrade of the CMS Tracker", Technical Report CERN-LHCC-2017-009. CMS-TDR-014, CERN, Geneva, Jun, 2017. 45, 46

[112] CMS Collaboration, "Technical proposal for the upgrade of the CMS detector through 2020", Technical Report CERN-LHCC-2011-006. LHCC-P-004, CERN, Jun, 2011. 46

[113] A. Dominguez et al., "CMS Technical Design Report for the Pixel Detector Upgrade", Technical Report CERN-LHCC-2012-016. CMS-TDR-11, CERN, Sep, 2012. 46

[114] R. Wigmans, "Calorimetry in High Energy Physics", *NATO Sci. Ser. B* **275** (1991) 325–379, doi:10.1007/978-1-4684-6006-3_6. 48

[115] CMS Collaboration, "The Phase-2 Upgrade of the CMS Barrel Calorimeters", Technical Report CERN-LHCC-2017-011. CMS-TDR-015, CERN, Geneva, Sep, 2017. 48, 51

[116] CMS Collaboration, "Time Reconstruction and Performance of the CMS Electromagnetic Calorimeter", *JINST* **5** (2010) T03011, doi:10.1088/1748-0221/5/03/T03011, arXiv:0911.4044. 48

[117] CMS Collaboration, "Performance and Operation of the CMS Electromagnetic Calorimeter", *JINST* **5** (2010) T03010, doi:10.1088/1748-0221/5/03/T03010, arXiv:0910.3423. 48

[118] K. W. Bell et al., "The development of vacuum phototriodes for the CMS electromagnetic calorimeter", *Nucl. Instrum. Meth. A* **469** (2001) 29–46, doi:10.1016/S0168-9002(01)00700-8. 49

[119] M. Gupta, "Calculation of radiation length in materials", Technical Report PH-EP-Tech-Note-2010-013, CERN, Geneva, Jul, 2010. 49

[120] J. Mans et al., "CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter", Technical Report CERN-LHCC-2012-015. CMS-TDR-10, CERN, Sep, 2012. 50

[121] CMS Collaboration, "The Phase-2 Upgrade of the CMS Endcap Calorimeter", Technical Report CERN-LHCC-2017-023. CMS-TDR-019, CERN, Geneva, Nov, 2017. 51

[122] CMS Collaboration, "Performance of CMS Muon Reconstruction in *pp* Collision Events at $\sqrt{s} = 7$ TeV", *JINST* **7** (2012) P10002, doi:10.1088/1748-0221/7/10/P10002, arXiv:1206.4071. 51

[123] CMS Collaboration, "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV", *JINST* **13** (2018), no. 06, P06015, doi:10.1088/1748-0221/13/06/P06015, arXiv:1804.04528. 51, 73

[124] CMS Collaboration, "The Phase-2 Upgrade of the CMS Muon Detectors", Technical Report CERN-LHCC-2017-012. CMS-TDR-016, CERN, Geneva, Sep, 2017. 52

[125] D. Contardo et al., "Technical Proposal for the Phase-II Upgrade of the CMS Detector", Technical Report CERN-LHCC-2015-010. LHCC-P-008. CMS-TDR-15-02, CERN, Geneva, Jun, 2015. 53

[126] A. Colaleo, A. Safonov, A. Sharma, and M. Tytgat, "CMS Technical Design Report for the Muon Endcap GEM Upgrade", Technical Report CERN-LHCC-2015-012. CMS-TDR-013, CERN, Jun, 2015. 53

[127] CMS Collaboration, "The CMS trigger system", *JINST* **12** (2017), no. 01, P01020, doi:10.1088/1748-0221/12/01/P01020, arXiv:1609.02366. 53, 54, 113

[128] CMS Collaboration, S. Dasu et al., "CMS TriDAS project: Technical Design Report, Volume 1: The Trigger Systems". Technical Design Report CMS. CERN, 2000. 53

[129] CMS Collaboration, S. Cittolin, A. Rácz, and P. Sphicas, "CMS The TriDAS Project: Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger. CMS trigger and data-acquisition project". Technical Design Report CMS. CERN, Geneva, 2002. 53

[130] P. R. Chumney et al., "Level-1 regional calorimeter trigger system for CMS", *eConf* **C0303241** (2003) THHT003, arXiv:hep-ex/0305047. 53

[131] C.-E. Wulz, "Concept of the First Level Global Trigger for the CMS Experiment at LHC", *Nucl. Instrum. Meth. A* **473** (2001) 231–242, doi:10.1016/S0168-9002(01)00809-9. 53

[132] G. Bauer et al., "Operational experience with the CMS Data Acquisition System", *Journal of Physics: Conference Series* **396** (dec, 2012) 012007, doi:10.1088/1742-6596/396/1/012007. 53

[133] M. Ashton et al., "Status report on the RD12 project: timing, trigger and control systems for LHC detectors", Technical Report CERN-LHCC-2000-002, CERN, Geneva, Jan, 2000. 53

[134] J. Varela, "CMS L1 Trigger Control System", Technical Report CMS-NOTE-2002-033, CERN, Geneva, Sep, 2002. 53

[135] B. Taylor, "Timing distribution at the LHC", in *8th Workshop on Electronics for LHC Experiments*, pp. 63–74. 9, 2002. 53

[136] RD12 Collaboration, "TTC distribution for LHC detectors", *IEEE Trans. Nucl. Sci.* **45** (1998) 821–828, doi:10.1109/23.682644. 54

[137] J. Andre et al., "Performance of the CMS Event Builder", *J. Phys. Conf. Ser.* **898** (2017), no. 3, 032020, doi:10.1088/1742-6596/898/3/032020. 54

[138] CMS Collaboration, "CMS Technical Design Report for the Level-1 Trigger Upgrade", Technical Report CERN-LHCC-2013-011. CMS-TDR-12, CERN, Jun, 2013. 54

[139] CMS Collaboration, "Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s} = 13$ TeV", doi:10.3204/PUBDB-2020-02629, arXiv:2006.10165. 54

[140] CMS Collaboration, "Software & Analysis in CMS". https://indico.cern.ch/event/92209/contributions/2114406/. Accessed: 2020-02-25. 55, 73

[141] CMS Collaboration, "The Phase-2 Upgrade of the CMS L1 Trigger Interim Technical Design Report", Technical Report CERN-LHCC-2017-013. CMS-TDR-017, CERN, Geneva, Sep, 2017. 56

[142] CMS Collaboration, "The Phase-2 Upgrade of the CMS DAQ Interim Technical Design Report", Technical Report CERN-LHCC-2017-014. CMS-TDR-018, CERN, Geneva, Sep, 2017. 56

[143] CMS Collaboration, G. Bayatyan et al., "CMS computing: Technical Design Report". Technical Design Report CMS. CERN, Geneva, 2005. 56

[144] C. Grandi et al., "CMS Computing Model: The "CMS Computing Model RTAG"", Technical Report CMS-NOTE-2004-031. CERN-LHCC-2004-035. LHCC-G-083, CERN, Geneva, Dec, 2004. 56

[145] CMS Collaboration, "CMS Site Status Readiness". https://twiki.cern.ch/twiki/bin/view/CMS/SiteSupportSiteStatusSiteReadiness. Accessed: 2020-02-09. 56

[146] CMS Collaboration, "Data Formats and Data Tiers". https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookDataFormats. Accessed: 2020-02-08. 56

[147] CMS Collaboration, "Mini-AOD: A New Analysis Data Format for CMS", *J. Phys. Conf. Ser.* **664** (2015), no. 7, 7, doi:10.1088/1742-6596/664/7/072052, arXiv:1702.04685. 57

[148] CMS Collaboration, "A further reduction in CMS event data for analysis: the NANOAOD format", *EPJ Web Conf.* **214** (2019) 06021, doi:10.1051/epjconf/201921406021. 58

[149] A. Buckley et al., "General-purpose event generators for LHC physics", *Phys. Rept.* **504** (2011) 145–233, doi:10.1016/j.physrep.2011.03.005, arXiv:1101.2599. 60

[150] T. Gleisberg et al., "Event generation with SHERPA 1.1", *JHEP* **02** (2009) 007, doi:10.1088/1126-6708/2009/02/007, arXiv:0811.4622. 61

[151] IPPP_Durham Collaboration, "Modelling the Invisible: computer simulation of a proton-proton collision at the LHC using Sherpa". https://twitter.com/IPPP_Durham/status/839850078496034816. Accessed: 2020-02-15. 61

[152] S. Höche, "Introduction to parton-shower event generators", in *Theoretical Advanced Study Institute in Elementary Particle Physics: Journeys Through the Precision Frontier: Amplitudes for Colliders*, pp. 235–295. 2015. arXiv:1411.4085. doi:10.1142/9789814678766_0005. 62

[153] T. Sjöstrand et al., "An introduction to PYTHIA 8.2", *Comput. Phys. Commun.* **191** (2015) 159–177, doi:10.1016/j.cpc.2015.01.024, arXiv:1410.3012. 63

[154] T. Sjöstrand, "The PYTHIA Event Generator: Past, Present and Future", *Comput. Phys. Commun.* **246** (2020) 106910, doi:10.1016/j.cpc.2019.106910, arXiv:1907.09874. 63

[155] M. Bahr et al., "Herwig++ Physics and Manual", *Eur. Phys. J. C* **58** (2008) 639–707, doi:10.1140/epjc/s10052-008-0798-9, arXiv:0803.0883. 64

[156] J. Bellm et al., "Herwig 7.0/Herwig++ 3.0 release note", *Eur. Phys. J. C* **76** (2016), no. 4, 196, doi:10.1140/epjc/s10052-016-4018-8, arXiv:1512.01178. 64

[157] S. Frixione et al., "The MCaNLO 4.0 Event Generator", arXiv:1010.0819. 64

[158] S. Frixione, P. Nason, and C. Oleari, "Matching NLO QCD computations with Parton Shower simulations: the POWHEG method", *JHEP* **11** (2007) 070, doi:10.1088/1126-6708/2007/11/070, arXiv:0709.2092. 64

[159] J. Alwall et al., "MadGraph 5 : Going Beyond", *JHEP* **06** (2011) 128, doi:10.1007/JHEP06(2011)128, arXiv:1106.0522. 64

[160] J. Alwall et al., "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations", *JHEP* **07** (2014) 079, doi:10.1007/JHEP07(2014)079, arXiv:1405.0301. 64

[161] J. Allison et al., "Recent developments in Geant4", *Nucl. Instrum. Meth. A* **835** (2016) 186–225, doi:10.1016/j.nima.2016.06.125. 65

[162] J. Allison et al., "Geant4 developments and applications", *IEEE Trans. Nucl. Sci.* **53** (2006) 270, doi:10.1109/TNS.2006.869826. 65

[163] GEANT4 Collaboration, "GEANT4: A Simulation toolkit", *Nucl. Instrum. Meth. A* **506** (2003) 250–303, doi:10.1016/S0168-9002(03)01368-8. 65

[164] L. Bauerdick et al., "Multiple-View, Multiple-Selection Visualization of Simulation Geometry in CMS", *J. Phys. Conf. Ser.* **396** (2012) 022052, doi:10.1088/1742-6596/396/2/022052. 66

[165] CMS Collaboration, F. Beaudette, "The CMS Particle Flow Algorithm", in *International Conference on Calorimetry for the High Energy Frontier*, pp. 295–304. 2013. arXiv:1401.8155. 68

[166] CMS Collaboration, "Tracking at High Level Trigger in CMS", *Nucl. Part. Phys. Proc.* **273-275** (2016) 2494–2496, doi:10.1016/j.nuclphysbps.2015.09.436. 69

[167] CMS Collaboration, "Tracking Performance in the CMS High Level Trigger - June 2018", (Jul, 2018). CMS-DP-2018-038. 70, 71

[168] CMS Collaboration, "Performance of the reconstruction and identification of high-momentum muons in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JINST* **15** (2020), no. 02, P02027, doi:10.1088/1748-0221/15/02/P02027, arXiv:1912.03516. 72

[169] CMS Collaboration, "CMS Electron and Photon Performance at 13 TeV", *J. Phys. Conf. Ser.* **1162** (2019), no. 1, 012008, doi:10.1088/1742-6596/1162/1/012008. 74

[170] CMS Collaboration, "Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector", *JINST* **14** (2019), no. 07, P07004, doi:10.1088/1748-0221/14/07/P07004, arXiv:1903.06078. 75, 121

[171] D. Bertolini, P. Harris, M. Low, and N. Tran, "Pileup Per Particle Identification", *JHEP* **10** (2014) 059, doi:10.1007/JHEP10(2014)059, arXiv:1407.6013. 75

[172] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k_t jet clustering algorithm", *JHEP* **04** (2008) 063, doi:10.1088/1126-6708/2008/04/063, arXiv:0802.1189. 76

[173] CMS Collaboration, "Sketch of a pp collision and resulting colimated spray of particles, a jet". https://twiki.cern.ch/twiki/pub/Sandbox/Lecture/Philipp_Schieferdeckers_Lecture.pdf. Accessed: 2020-02-27. 77

[174] CMS Collaboration, "Sketch of an interesting interaction overlapping in the detector with a pileup interaction". `https://cms.cern/news/how-cms-weeds-out-particles-pile`. Accessed: 2020-02-27. 77

[175] CMS Collaboration, "Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS", *JINST* **6** (2011) P11002, `doi:10.1088/1748-0221/6/11/P11002`, `arXiv:1107.4277`. 78

[176] CMS Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV", *JINST* **13** (2018), no. 05, P05011, `doi:10.1088/1748-0221/13/05/P05011`, `arXiv:1712.07158`. 79, 80, 125

[177] CMS Collaboration, "Identification of b-Quark Jets with the CMS Experiment", *JINST* **8** (2013) P04013, `doi:10.1088/1748-0221/8/04/P04013`, `arXiv:1211.4462`. 79

[178] D. Guest et al., "Jet Flavor Classification in High-Energy Physics with Deep Neural Networks", *Phys. Rev. D* **94** (2016), no. 11, 112002, `doi:10.1103/PhysRevD.94.112002`, `arXiv:1607.08633`. 79

[179] CMS Collaboration, "Performance of reconstruction and identification of τ leptons decaying to hadrons and ν_τ in pp collisions at $\sqrt{s} = 13$ TeV", *JINST* **13** (2018), no. 10, P10005, `doi:10.1088/1748-0221/13/10/P10005`, `arXiv:1809.02816`. 81, 82

[180] "LPC CMS DAS: Tau Leptons". `https://indico.cern.ch/event/662371/contributions/2704735/attachments/1514582/2496950/Tau_ID_tutorial_CMSDAS_2018.pdf`. Accessed: 2020-09-03. 80

[181] CMS Collaboration, "Performance of tau-lepton reconstruction and identification in CMS", *JINST* **7** (2012) P01001, `doi:10.1088/1748-0221/7/01/P01001`, `arXiv:1109.6034`. 81

[182] L. Lista, "Statistical Methods for Data Analysis in Particle Physics", volume 909. Springer, 2016. 85

[183] G. Cowan, "Statistical data analysis", *Oxford University Press, Oxford U.K.* (2002). 85

[184] O. Behnke, K. Kröninger, T. Schörner-Sadenius, and G. Schott, eds., "Data analysis in high energy physics: A practical guide to statistical methods". Wiley-VCH, Weinheim, Germany, 2013. 85, 100

[185] The ATLAS Collaboration, The CMS Collaboration, The LHC Higgs Combination Group Collaboration, "Procedure for the LHC Higgs boson search combination in Summer 2011", Aug, 2011. CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11. 85, 152, 155

[186] CMS Collaboration, "Combined results of searches for the standard model Higgs boson in *pp* collisions at $\sqrt{s} = 7$ TeV", *Phys. Lett. B* **710** (2012) 26–48, `doi:10.1016/j.physletb.2012.02.064`, `arXiv:1202.1488`. 85

[187] L. Lyons, H. B. Prosper, and A. De Roeck, eds., "Statistical issues for LHC physics. Proceedings, Workshop, PHYSTAT-LHC, Geneva, Switzerland, June 27-29, 2007", CERN Yellow Reports: Conference Proceedings. (3, 2008). doi:10.5170/CERN-2008-001. 85

[188] P. C. Bhat, "Multivariate Analysis Methods in Particle Physics", *Ann. Rev. Nucl. Part. Sci.* **61** (2011) 281–309, doi:10.1146/annurev.nucl.012809.104427. 85

[189] K. Albertsson et al., "Machine Learning in High Energy Physics Community White Paper", *J. Phys. Conf. Ser.* **1085** (2018), no. 2, 022008, doi:10.1088/1742-6596/1085/2/022008, arXiv:1807.02876. 85

[190] G. Cowan, "Lectures on Statistical Data Analysis". https://www.pp.rhul.ac.uk/~cowan/. Accessed: 2020-03-06. 91, 95, 101, 106

[191] P. K. Sinervo, "Signal significance in particle physics", in *Conference on Advanced Statistical Techniques in Particle Physics*, pp. 64–76. 6, 2002. arXiv:hep-ex/0208005. 94

[192] E. Gross and O. Vitells, "Trial factors for the look elsewhere effect in high energy physics", *Eur. Phys. J. C* **70** (2010) 525–530, doi:10.1140/epjc/s10052-010-1470-8, arXiv:1005.1891. 99

[193] A. L. Read, "Presentation of search results: The CL(s) technique", *J. Phys. G* **28** (2002) 2693–2704, doi:10.1088/0954-3899/28/10/313. 100

[194] T. Junk, "Confidence level computation for combining searches with small statistics", *Nucl. Instrum. Meth. A* **434** (1999) 435–443, doi:10.1016/S0168-9002(99)00498-2, arXiv:hep-ex/9902006. 100

[195] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics", *Eur. Phys. J. C* **71** (2011) 1554, doi:10.1140/epjc/s10052-011-1554-0, arXiv:1007.1727. [Erratum: Eur.Phys.J.C 73, 2501 (2013)]. 100

[196] CMS Collaboration, "Observation of Higgs boson decay to bottom quarks", *Phys. Rev. Lett.* **121** (2018), no. 12, 121801, doi:10.1103/PhysRevLett.121.121801, arXiv:1808.08242. 102

[197] CMS Collaboration, "Evidence for associated production of a Higgs boson with a top quark pair in final states with electrons, muons, and hadronically decaying τ leptons at $\sqrt{s} = 13$ TeV", *JHEP* **08** (2018) 066, doi:10.1007/JHEP08(2018)066, arXiv:1803.05485. 102

[198] CMS Collaboration, "Evidence for Higgs boson decay to a pair of muons", (9, 2020). arXiv:2009.04363. CMS-HIG-19-006, CERN-EP-2020-164. 102

[199] A. Hocker et al., "TMVA - Toolkit for Multivariate Data Analysis", arXiv:physics/0703039. 103, 127

[200] D. de Florian, G. Ferrera, M. Grazzini, and D. Tommasini, "Transverse-momentum resummation: Higgs boson production at the Tevatron and the LHC", *JHEP* **11** (2011) 064, doi:10.1007/JHEP11(2011)064, arXiv:1109.2109. 110

[201] G. Bozzi, S. Catani, D. de Florian, and M. Grazzini, "Transverse-momentum resummation and the spectrum of the Higgs boson at the LHC", *Nucl. Phys. B* **737** (2006) 73–120, doi:10.1016/j.nuclphysb.2005.12.022, arXiv:hep-ph/0508068. 110

[202] A. Elagin, P. Murat, A. Pranko, and A. Safonov, "A New Mass Reconstruction Technique for Resonances Decaying to di-tau", *Nucl. Instrum. Meth. A* **654** (2011) 481–489, doi:10.1016/j.nima.2011.07.009, arXiv:1012.4686. 116

[203] J. D'Hondt et al., "Fitting of event topologies with external kinematic constraints in CMS", (1, 2006). CERN-CMS-NOTE-2006-023. 116

[204] CMS Collaboration, "Muons in the CMS High Level Trigger System", *Nucl. Part. Phys. Proc.* **273-275** (2016) 2509–2511, doi:10.1016/j.nuclphysbps.2015.09.441. 119

[205] CMS Collaboration, "Tag and Probe method". https://twiki.cern.ch/twiki/bin/view/CMSPublic/TagAndProbe. Accessed: 2020-06-26. 119

[206] CMS Collaboration, "Generic tag and probe tool for measuring efficiency at cms with early data", (2009). CMS AN-2009/111. 119

[207] P. Das and K. Mazumdar, "Measurement of Missing Transverse Energy in CMS Experiment", *Springer Proc. Phys.* **203** (2018) 833–835, doi:10.1007/978-3-319-73171-1_202. 121

[208] "Model Comparisons and The F-test". http://people.reed.edu/~jones/Courses/P24.pdf. Accessed: 2020-06-26. 122

[209] I. Antcheva et al., "ROOT: A C++ framework for petabyte data storage, statistical analysis and visualization", *Comput. Phys. Commun.* **180** (2009) 2499–2512, doi:10.1016/j.cpc.2009.08.005, arXiv:1508.07749. 125, 127

[210] CMS Collaboration, "Methods to apply b-tagging efficiency scale factors". https://twiki.cern.ch/twiki/bin/view/CMS/BTagSFMethods#1c_Event_reweighting_using_scale. Accessed: 2020-07-08. 122

[211] B. Garrett and G. H. L., "Smooth Surface Interpolation", *Journal of Mathematics and Physics* **39** (1960), no. 1-4, 258–268, doi:10.1002/sapm1960391258, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/sapm1960391258. 138

[212] A. Kolmogorov, "Foundations of the Theory of Probability". Chelsea Publishing, New York, 1956. 141

[213] F. J. M. Jr., "The Kolmogorov-Smirnov Test for Goodness of Fit", *Journal of the American Statistical Association* **46** (1951), no. 253, 68–78, doi:10.1080/01621459.1951.10500769. 141

[214] ATLAS, CMS, LHC Higgs Combination Group Collaboration, "Procedure for the LHC Higgs boson search combination in summer 2011", (8, 2011). ATL-PHYS-PUB-2011-011, CMS-NOTE-2011-005. 142

[215] J. Conway, "Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra", in *PHYSTAT 2011*, pp. 115–120. 2011. arXiv:1103.0354. doi:10.5170/CERN-2011-006.115. 144

[216] R. J. Barlow and C. Beeston, "Fitting using finite Monte Carlo samples", *Comput. Phys. Commun.* **77** (1993) 219–228, doi:10.1016/0010-4655(93)90005-W. 144

[217] NNPDF Collaboration, "Parton distributions for the LHC Run II", *JHEP* **04** (2015) 040, doi:10.1007/JHEP04(2015)040, arXiv:1410.8849. 146

[218] J. Pumplin et al., "New generation of parton distributions with uncertainties from global QCD analysis", *JHEP* **07** (2002) 012, doi:10.1088/1126-6708/2002/07/012, arXiv:hep-ph/0201195. 146

[219] R. D. Cousins, "Generalization of Chisquare Goodness-of-Fit Test for Binned Data Using Saturated Models, with Application to Histograms". http://www.physics.ucla.edu/~cousins/stats/cousins_saturated.pdf. Accessed: 2020-07-08. 148

[220] L.-G. Xia, "Study of constraint and impact of a nuisance parameter in maximum likelihood method", *J. Phys. G* **46** (2019) 085004, doi:10.1088/1361-6471/ab02c0, arXiv:1805.03961. 148

[221] E. Gross, "Practical Statistics for High Energy Physics", *CERN Yellow Rep. School Proc.* **3** (2018) 199–221, doi:10.23730/CYRSP-2018-003.199. 150

[222] "Combine tool: Advanced Use Cases". https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/part3/nonstandard/. Accessed: 2020-07-08. 150

[223] D. Perez Adan, "Search for Light Bosons in Exotic Decays of the 125 GeV Higgs Boson". Dissertation, Hamburg University, Hamburg, 2020. doi:10.3204/PUBDB-2020-02143. 154

[224] "Interpretation of h->aa results in the context of the 2HDM+S". https://twiki.cern.ch/twiki/bin/viewauth/CMS/HaaInterpretations. Accessed: 2020-07-10. 154

[225] U. Haisch, J. F. Kamenik, A. Malinauskas, and M. Spira, "Collider constraints on light pseudoscalars", *JHEP* **03** (2018) 178, doi:10.1007/JHEP03(2018)178, arXiv:1802.02156. 154

[226] "Higgs Exotic Decay". https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWGExoticDecay. Accessed: 2020-07-10. 154

[227] CMS Collaboration, "Alignment of the CMS Silicon Tracker during Commissioning with Cosmic Rays", *JINST* **5** (2010) T03009, doi:10.1088/1748-0221/5/03/T03009, arXiv:0910.2505. 183

[228] CMS Collaboration, "CMS Tracker Performance results for full Run 2 Legacy reprocessing", Feb, 2020. CMS-DP-2020-012. 184

[229] CMS Collaboration, "Alignment of the CMS tracker with LHC and cosmic ray data", *JINST* **9** (2014) P06009, doi:10.1088/1748-0221/9/06/P06009, arXiv:1403.2286. 185

[230] "Tracker Alignment". https://etpwww.etp.kit.edu/~mschrode/research_tkal.html. Accessed: 2020-08-28. 185

[231] V. Blobel and C. Kleinwort, "A New method for the high precision alignment of track detectors", in *Conference on Advanced Statistical Techniques in Particle Physics*. 6, 2002. arXiv:hep-ex/0208021. 186

[232] V. Blobel, "Software alignment for tracking detectors", *Nucl. Instrum. Meth. A* **566** (2006) 5–13, doi:10.1016/j.nima.2006.05.157. 186

[233] T. Lampén, "Detector Alignment Studies for the CMS Experiment". Dissertation, University of Helsinki, Helsinki, 2007. 186

[234] A. Bonato, "Weak modes in alignment". https://indico.cern.ch/event/137973/contributions/1362329/attachments/115109/163450/WS2011_AB_WeakModes_v2.pdf. Accessed: 2020-08-28. 189

[235] CMS Collaboration, "Additional Run 2 CMS Tracker Alignment Performance Results", Jul, 2020. CMS-DP-2020-038. 191, 192, 193

ACKNOWLEDGMENTS

After the completion of this work it is time to pause for a bit, and before heading for new challenges, have a look behind and thank all the people who helped to make it possible.
The first steps of my education shaped the paths I took later on. From these times, I am especially thankful to my teachers Anabel for her life lessons and Francisco Gutiérrez Pérez (Pancho) for sharing his contagious passion for Physics with his students. The science fairs and experiments carried out during the physics lessons under his attentive eye were a major motivation for me to study Physics. I would also like to thank the professors at Higher Institute of Technologies and Applied Sciences in Havana, who inspired me with their personal example as scientists and educators, and from which I learned the "tú sí puedes" spirit. They created an optimal environment to gather knowledge and valuable skills, that I treasure with great affection. In particular, I would like to thank Prof. Lic. Katia D'Alessandro Rodríguez, Prof. Dr. César E. García Trápaga for the great lessons on Statistics, and Prof. Dr. Fernando Guzmán Martínez for his supervision during my undergraduate studies, and for being always in constant communication during these years.
I am grateful to PD Dr. Hannes Jung, Dr. Daniela Domínguez Damiani, and Dr. Armando Bermúdez Martínez who guided me through my first steps at DESY as a summer student in 2016. As a recent physics graduate at the time, the opportunity of joining the QCD DESY-CMS group and involving in the day to day scientific work left a lasting impression on me and motivated me to pursue doctoral studies in the field of Particle Physics. I am especially grateful to Hannes for his guidance and words of encouragement during these first steps of my scientific career. Later on, during my PhD time, he was kind enough to always find the time for a chat and has been someone I could rely on for precious advice on future career steps.
I would like to thank my PhD supervisor Dr. Alexei Raspereza for the opportunity to join the Higgs DESY-CMS group under his supervision. Back in the summer of 2016, when we first met, I was impressed by the knowledge and enthusiasm for research he transmitted during the lectures on Electroweak and Higgs Physics he taught as part of the student program. During my PhD time, he provided precious guidance on the critical moments while allowing

me to gain confidence working independently. It has been truly my honor to work under his supervision. On equal foot goes my gratitude to my co-supervisor Prof. Dr. Elisabetta Gallo for her valuable feedback and suggestions. CMS DESY group has greatly benefited from her knowledge, experience, kind attitude, and wise leadership.
I am grateful to the jury members who joined my supervisors in the examination committee, Prof. Dr. Robin Santra, Prof. Dr. Gudrid Moortgat-Pick, and Prof. Dr. Johannes Haller, for the time dedicated to read the thesis and evaluate the work.
I would also like to express gratitude to my colleagues in the Higgs group, Dr. Teresa Lenz, Dr. Mareike Meyer, Dr. Yiwen Wen, Dr. Merijn van de Klundert, Dr. Rainer Mankel, Andrea Cardini, Oleg Filatov, and Maryam Bayat Makou for all the help with technicalities, feedback, and fruitful discussions during our work meetings. I benefited a lot from the group's gathered knowledge and for that, I am really grateful. I am thankful to my office mates Henriette Petersen and Rafael Eduardo Sosa Ricardo for the nice time shared at the office while working on our PhD projects. I am also grateful to Dr. Gregor Mittag who helped me to get familiar with the alignment tasks at the beginning of my PhD. From the tracker alignment group, I would also like to thank Dr. Patrick Connor, Dr. Adinda de Wit, and Dr. Valeria Botta for the fruitful collaboration. I am especially thankful to Valeria for finding the time to read the appendix of my thesis related to alignment and providing very helpful comments on short notice, and to PD Dr. Olaf Behnke for the valuable help with the german version of the abstract. I would also like to thank DESY CMS secretaries for the kind and very efficient support on administrative matters. A big thanks goes also to the members of the DESY International Office, from which I received as a summer student a warm welcome and facilitated a lot the first days at DESY as a PhD student. A nice complement to my workday was the german lessons offered by the PIER Helmholtz Graduate School. I am grateful to the german teachers Gisa Guenther and Alexander Hoeppner for their friendliness and for convincing me that "Das Leben ist nicht zu kurz, um Deutsch zu lernen". I also thank DESY colleagues who joined the lessons and helped to create a relaxed and friendly learning atmosphere.
The biggest thanks of all go to my parents Orgilia and Sergio and my sister Dania for having encouraged me and accompanied me in every project that I have undertaken in my life. From my grandparents Orgilia and Julián, as well as Tita and Raúl, I cherish their life lessons on sacrifice and the importance of studying hard for a better future. From my grandfather Sergio I treasure his "always do your best" advice. I also thank my aunts Gladyta and Freida and my cousins Javier and Ariel for the encouragement from the distance and for celebrating with me every accomplishment. My gratitude also goes to my mother-in-law Leticia for supporting me during my studies and for also giving me occasionally funny hard times. I thank my lifetime friends. Time has shown our bond is stronger than distance and I am always looking forward to meeting again, remember the "old times", and happily talk about our hopes for the future. I thank the little Europa Kolleg community in Hamburg for being a valuable support network and making our stay away from home much more enjoyable. I also especially thank my sister Dania for being an inspiration to follow this path. Finally, I thank my life partner Danyer for the unique connection and mutual understanding we have built along this long route, for his contribution to this work, and for helping me to bring out the best of me, sometimes knowing me better than I do myself. Please, know all that I am humbly grateful!