# Search for Higgs boson production in association with b-quarks in final states with leptons with machine learning techniques at CMS

## Dissertation

zur Erlangung des Doktorgrades an der Fakultät für

Mathematik, Informatik und Naturwissenschaften

Fachbereich Physik

der Universität Hamburg

vorgelegt von

**Maryam Bayat Makou**

aus

Teheran, Iran

Hamburg

2024

# Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich die Arbeit eigenständig verfasst habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen und die eingereichte schriftliche Fassung entspricht der auf dem elektronischen Speichermedium.

Hamburg, den May 13, 2024.

Maryam Bayat Makou

MARYAM BAYATMAKOU

| | |
|---|---|
| Gutachter der Dissertation: | Prof. Dr. Elisabetta Gallo |
| | Prof. Dr. Peer Stelldinger |
| Zusammensetzung der Prüfungskommission: | Prof. Dr. Sven-Olaf Moch |
| | Prof. Dr. Elisabetta Gallo |
| | Prof. Dr. Peer Stelldinger |
| | Dr. Rainer Mankel |
| | Prof. Dr. Christian Schwanenberger |
| Vorsitzender der Prüfungskommission: | Prof. Dr. Sven-Olaf Moch |
| Datum der Disputation: | 26. April 2024 |
| Vorsitzender des Fach-Promotionsausschusses Physik: | Prof. Dr. Markus Drescher |
| Leiter des Fachbereichs Physik: | Prof. Dr. Wolfgang J. Parak |
| Dekan der Fakultät MIN: | Prof. Dr.-Ing. Norbert Ritter |

# Abstract

One of the primary objectives of the CMS experiment is to measure the mechanisms of the Higgs boson production precisely and to determine its coupling structure. The Yukawa interaction is the mechanism via which the Higgs boson couples to fermions in the Standard Model. Due to the small cross-section and the very large background processes, the Yukawa coupling to b-quarks ($y_b$) has only been measured in the decay process and not in the production mechanism. The present work is devoted to analysing the Higgs boson production in association with b-quarks in final states with leptons. The search for Higgs boson decays to tau leptons and W bosons was performed using data collected by the CMS experiment in proton-proton collisions at a center-of-mass energy of 13 TeV during LHC Run 2. The final states covered in this search are those in which the Higgs boson is produced via the bbH channel and decays into two tau leptons, which subsequently decay either fully hadronically ($\tau_h\tau_h$), semi-leptonically ($\tau_e\tau_h$, $\tau_\mu\tau_h$), or fully leptonically ($\tau_e\tau_\mu$). The latter channel is also studied from the decay of Higgs boson into two W bosons. The fully leptonic channel is the primary focus of this work, and as such, it will be described extensively. Other channels will be briefly presented to discuss the combination of the results. This analysis employs machine learning techniques to improve the sensitivity to the signal process, which brings almost factor 2 improvement with respect to the traditional cut-based approach.

Final constraints are derived on the Higgs boson production cross section as an upper limit on the signal strength of the process. The observed (expected) upper limits on the signal strength modifier obtained for the fully leptonic channel for the Run 2 data is 18.7 (19.1) times the Standard Model prediction at 95% confidence level. For the combination of all the channels, fully hadronically ($\tau_h\tau_h$), semi-leptonically ($\tau_e\tau_h$, $\tau_\mu\tau_h$), fully leptonically ($\tau_e\tau_\mu$), the upper limit at 95% confidence level is observed (expected) to be 3.7 (6.1) times the Standard Model prediction. Furthermore, constraints on the Higgs Yukawa coupling to the b-quark are determined in the kappa model interpretation, and the results are consistent with the Standard Model.

# Zusammenfassung

Eines der Hauptziele des CMS-Experiments ist es, die Mechanismen der Higgs-Boson-Produktion präzise zu messen und seine Kopplungsstruktur zu bestimmen. Die Yukawa-Wechselwirkung ist der Mechanismus, durch den das Higgs-Boson im Standardmodell an Fermionen koppelt. Aufgrund des kleinen Wirkungsquerschnitts und der sehr großen Untergrundprozesse wurde die Yukawa-Kopplung an b-Quarks ($y_b$) nur im Zerfallsprozess und nicht im Produktionsmechanismus gemessen. Die vorliegende Arbeit widmet sich der Analyse der Higgs-Boson-Produktion in Verbindung mit b-Quarks in Endzuständen mit Leptonen. Die Suche nach Zerfällen des Higgs-Bosons in Tau-Leptonen und W-Bosonen wurde mit Daten durchgeführt, die vom CMS-Experiment in Proton-Proton-Kollisionen bei einer Schwerpunktsenergie von 13 TeV während des LHC-Laufs 2 gesammelt wurden. Die in dieser Suche abgedeckten Endzustände sind solche, in denen das Higgs-Boson über den bbH-Kanal produziert wird und in zwei Tau-Leptonen zerfällt, die anschließend entweder vollständig hadronisch ($\tau_h\tau_h$), semi-leptonisch ($\tau_e\tau_h$, $\tau_\mu\tau_h$) oder vollständig leptonisch ($\tau_e\tau_\mu$) zerfallen. Der letztere Kanal wird auch vom Zerfall des Higgs-Bosons in zwei W-Bosonen untersucht. Der vollständig leptonische Kanal ist der Hauptfokus dieser Arbeit und wird daher ausführlich beschrieben. Andere Kanäle werden kurz vorgestellt, um die Kombination der Ergebnisse zu diskutieren. Diese Analyse verwendet Techniken des maschinellen Lernens, um die Empfindlichkeit für den Signalprozess zu verbessern, was eine fast doppelte Verbesserung im Vergleich zum traditionellen Schnittbasierten Ansatz bringt.

Abschließend werden Einschränkungen für den Higgs-Boson-Produktionsquerschnitt als obere Grenze für die Signalstärke des Prozesses abgeleitet. Die beobachteten (erwarteten) oberen Grenzen für den Modifikator der Signalstärke für den vollständig leptonischen Kanal für die Run2-Daten betragen das 18,7 (19,1)-fache der Vorhersage des Standardmodells mit einer Vertrauenswahrscheinlichkeit von 95%. Für die Kombination aller Kanäle, vollständig hadronisch ($\tau_h\tau_h$), semi-leptonisch ($\tau_e\tau_h$, $\tau_\mu\tau_h$), vollständig leptonisch ($\tau_e\tau_\mu$), wird die obere Grenze mit einer Vertrauenswahrscheinlichkeit von 95% beobachtet (erwartet), das 3,7 (6,1)-fache der Vorhersage des Standardmodells zu sein. Darüber hinaus werden Einschränkungen für die Higgs-Yukawa-Kopplung an das b-Quark im Kappa-Modell-Interpretation bestimmt, und die Ergebnisse sind konsistent mit dem Standardmodell.

# Contents

CHAPTER 1

# Introduction

Humans have long been fascinated by the universe and its origin, sparking a quest for understanding that has shifted from myths to scientific research. Philosophers from Ancient Greece, such as Democritus, postulated the existence of atoms as the fundamental units of matter. Since this first step towards the foundation of modern science, the experimental efforts resumed only after the scientific revolution.

In the early 19th century, John Dalton's atomic theory became well-established. According to Dalton, all the elements are composed of indivisible, tiny particles known as atoms, each with unique properties. Later, the discovery of the electron by J.J. Thomson, the introduction of Rutherford's model of the nucleus, and the identification of the neutron by James Chadwick modified our core understanding of the atomic structure.

Furthermore, discoveries such as the photoelectric effect and black body radiation paved the way for modern physics and the formulation of the quantum theory, to which special relativity was later added. The quantum theory reveals the complex nature of particles and forces, which led to today's understanding of particle physics based on the Standard Model framework. The Standard Model (SM) of particle physics was developed in the $20^{th}$ century after decades of experimental work in high-energy physics [1]. This phenomenological model captures our entire understanding of the fundamental non-gravitational interactions that control our universe and the building blocks of matter. The SM describes three of the four fundamental forces of our universe: the electromagnetic, strong, and weak forces. Also, the SM integrates the electroweak and the strong forces into a single theoretical framework described in a quantum gauge field theory. Even though the Standard Model (SM) is, for many aspects, a successful theory, there are still unanswered theoretical problems and evidence from experiments that suggest the Standard Model is incomplete. The latter can not yet provide an explanation to account for the neutrino oscillation discovery or the cosmic observation of non-baryonic cold dark matter. A thorough treatment of Beyond Standard Model (BSM) theories is beyond the scope of this work; for further details, see the external sources present in the literature

such as [2].

Many precise experiments have been performed using particle collider technology, reaching the TeV scale and verifying the validity of the SM in this regime. The existence of the W and Z bosons, gluons, top quarks, and charm quarks, as well as the anomalous gyromagnetic moment of the electron with precisions better than one part in a billion, are only a few of the findings that, up to this point, have been determined to be consistent with the predictions of the Standard Model of particle physics. Furthermore, the discovery of a scalar particle compatible with the Higgs boson of the Standard Model [3,4], with a mass of 125 GeV, represents one of the most outstanding achievements of the SM model and the experimental efforts.

The existence of a new elementary particle was postulated in 1964 by Robert Brout, Francois Englert and Peter Higgs (BEH) [5, 6] to explain the origin of the masses of force-carrying particles via the so-called Electroweak Symmetry Breaking (EWSB). In the 1960s, it was still uncertain whether a self-consistent theory that contained massive force carriers could be built, and this query was also mentioned by both nuclear physicists and in the scope of condensed matter physics. According to Peter Higgs, Brout and Englert (BEH), such a notation could eventually be realized if one could construct a non-zero value for the "Higgs" field and introduce an interaction between the force carriers and the Higgs field.

The interactions between particles and a Higgs field were to play a central role in developing the electroweak part of the Standard Model. This was especially crucial to generate masses for the W and Z bosons, which were needed to be consistent with experimental observations, while photons and gluons remained massless. Interestingly, interactions with the Higgs field also gave rise to a viable theoretical mechanism for the fermion mass generation: the stronger the interaction (or coupling) between a fermion and the Higgs field, the larger the particle resulting mass. The interaction is referred to as a "Yukawa" interaction in the Standard Model.

Measuring the Yukawa coupling structure is crucial to comprehend the coupling structure of the Higgs Boson to fermions. It can be measured experimentally from either Higgs Boson production or decay modes. All of the main Higgs Boson production modes have been observed so far, with the exception of the production mode of interest in this work, i.e. the b-associated production (bbH). Due to its limited cross-section and the overwhelmingly large background, the Yukawa coupling to b-quarks ($y_b$) has only been measured in the decay process. Nevertheless, it is yet to be measured in the production mechanism, presented for the first time in the present thesis.

The largest circular proton-proton collider ever built, the Large Hadron Collider (LHC), is the experimental apparatus used to conduct searches for the Higgs production modes. The high centre-of-mass energy of 13 TeV was attained in 2016 for the first time to also search for Higgs production modes and precisely measure the Higgs boson production to clarify its coupling structure. One of the two multipurpose detectors placed at the LHC, the Compact Muon Solenoid (CMS), has been engineered to precisely identify

the decay products of a particle similar to a Higgs boson. The detector achieves excellent vertexing, tracking, and calorimeter resolution, which is crucial for reconstructing the final state particles of interest.

This work presents the measurement of the b-associated Higgs production (bbH) using data collected by the CMS experiment during Run 2 data taking (2016-2018) at the center of mass $\sqrt{s} = 13$ TeV, corresponding to an integrated luminosity of 138 fb$^{-1}$. The study covers events where the Higgs boson is produced through the bbH channel and further decays into two tau leptons, subsequently fully leptonically ($\tau_e\tau_\mu$), semi leptonically ($\tau_e\tau_h$ , $\tau_\mu\tau_h$), and fully hadronically ($\tau_h\tau_h$). Machine learning techniques are deployed to distinguish the signal from other processes with a similar signature, namely background events. The present work will discuss the first results of this search. The primary focus of the present work is the analysis of the fully leptonic final state which is produced from the Higgs decay to tau leptons or W bosons.The other channels will be briefly covered to show how they are included in the result combination.

The format of this work is as follows: A brief overview of the Standard Model is given in chapter two, with particular attention to the Higgs boson and the properties of production of b-associated Higgs bosons. The CMS experiment and the CERN Large Hadron Collider (LHC) are introduced in the third chapter. The CMS experiment and its primary physics object reconstruction are covered in the fourth chapter. The multivariate techniques used for the analysis are covered in Chapter Five, along with strategies for handling imbalanced data sets, explainable AI techniques like SHAP, and boosted decision trees. The analysis approach for the study in the fully leptonic channel is described in Chapter Six, together with background modelling, multivariate analysis results, the modelling of the background processes and uncertainties relevant to the analysis. The combined analysis results from all the channels will be covered in Chapter Seven, and the summary and conclusion will be covered in the last chapter.

CHAPTER 2

# Theoretical framework

Elementary particle physics tackles the question, "What is the structure of matter?" at a fundamental level. The standard model of particle physics is the theory explaining the building blocks of matter and three of the universe's four fundamental forces. Despite many successful predictions by the standard model, which are verified by numerous experiments [1], the standard model still faces unanswered questions, such as the nature of dark matter, and a quantized theory of gravity. This chapter presents a concise phenomenological overview of the standard model of particle physics, with the main focus on the Higgs boson and its properties.

## 2.1 The Standard Model

The Standard Model (SM) of particle physics [7] was primarily formulated during the 1970s and early 1980s. The SM summarizes our current understanding of all fundamental particles and non-gravitational interaction in a comprehensive way. Mathematically, the Standard Model is described by the Quantum Field Theory (QFT) [8]. At its core lies the concept of quantization applied to both the constituents of matter and the particles responsible for mediating forces, all of which are observable physical entities. Within this framework, every elementary particle is viewed as an excitation of the corresponding quantum fields. Three of the four fundamental forces of our universe are described by the SM, namely the electromagnetic, weak and strong forces. In contrast, gravity is not included, owing to the difficulties of formulating a renormalizable quantization of the gravitational force.

In the SM framework, particles are divided into three generations of fermions, further classified into leptons and quarks. The latter categorization of fermions corresponds to the twelve fundamental spin half particles listed in Figure 2.1. The lepton families of fermions ($e$, $\mu$ and $\tau$) are negatively charged, and each has a corresponding neutral particle, the

neutrinos ($\nu_e$, $\nu_\mu$ and $\nu_\tau$). The quark generations are each composed of an up- type quark (u, c, t) and a down-type quark (d, s, b), with the charge of the up-type quark being $+2/3$ $e$ and for the down-type quark with $-1/3$ $e$ charge, where $|e|$ denotes the fundamental absolute value of the electron charge. Additionally, quarks also carry a colour charge, which is conventionally labelled as red, green and blue. In the particle content of the SM, the mass of the fermions increases with a higher generation index, resulting in the third generation containing the heaviest family of particles. The heaviest quark observed is the top quark with a mass of 172.52 GeV [9], whereas the heaviest lepton is the $\tau$ lepton with a mass of 1.776 GeV [9]. Within the SM framework, the neutrinos are usually massless[1], which is in contrast with the observation of neutrino oscillation.

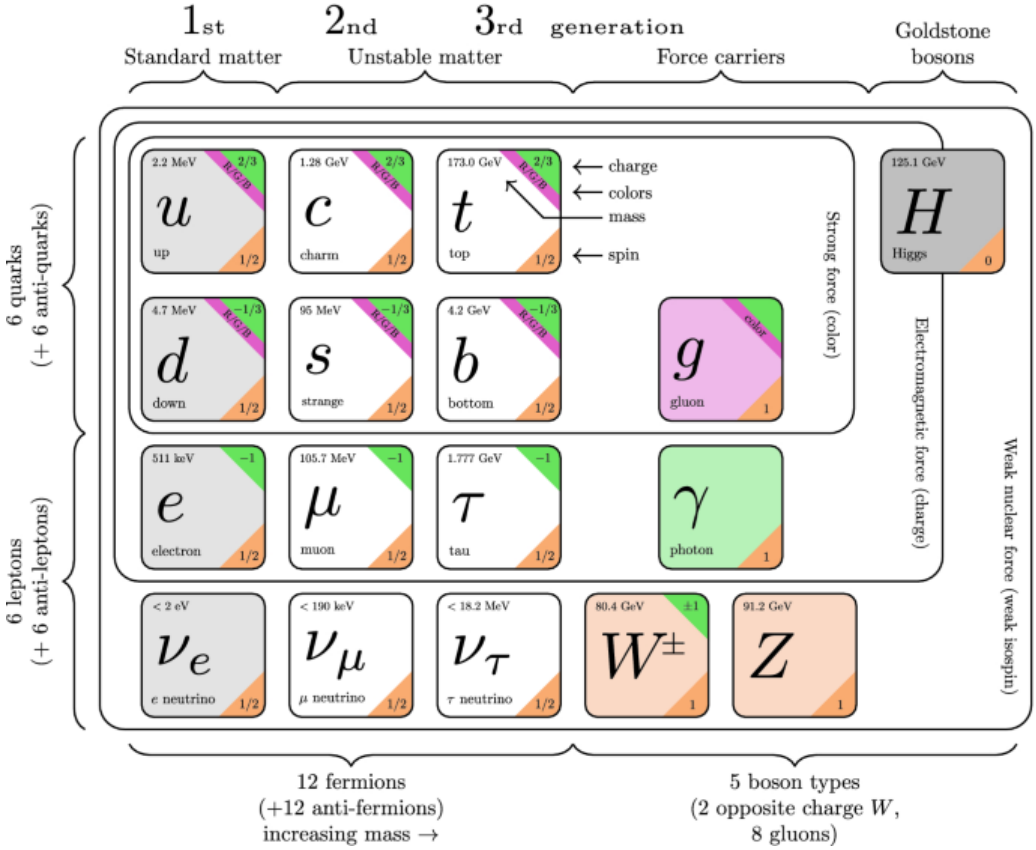


Figure 2.1: The three generations of particles & force carriers, including their mass charge and spin, are shown. [10]

For each of the twelve fermions, a corresponding antiparticle state exists with an opposite charge and identical mass. The existence of an anti-particle is a direct consequence

[1]since no right-handed neutrino term is observed, which is needed to generate the Dirac mass term for neutrinos.

of both positive and negative energy states, representing an equally viable solution to the Dirac equation, which describes the dynamics of the fermions.

The particle scope of the SM shown in figure 2.1 also contains the gauge bosons, with an integer spin of 1, which are the mediators of the interactions between matter fields. The only scalar (spin-0) particle in the SM is the Higgs boson, which is responsible for generating particle mass. The spin-one gauge bosons, namely the photon $\gamma$, the $W^{\pm}$ and the $Z$ gauge bosons and the eight gluons ($g$) mediate respectively, the electromagnetic, the weak and strong force. All fermions are influenced by the weak force as summarised in Figure 2.1. The electrically charged particles take part in the electromagnetic interaction (QED), and since only quarks are colour-charged, they feel a strong force (QCD). The following section discusses the interaction of particles and their formulation, which are the core of the SM.

### 2.1.1 Symmetries and interactions

Despite the large number of degrees of freedom, i.e. 28 bosonic and 90 fermionic ones [11], the Standard Model (SM) stands out as a remarkably elegant Quantum Field Theory (QFT) model. The foundation of the SM lies in Lagrangian formalism, and the concept of symmetries fundamentally drives all of its interactions. The consequences of the invariance of a system under a continuous symmetry were first formulated by Emmy Noether in her groundbreaking theorem [12]. In terms of the Lagrangian formalism, every conserved physical quantity, such as energy and momentum of a particle, is associated with the invariance of the equations of motions under a continuous symmetry, respectively, time and spatial translations.

The Standard Model also features additional symmetries other than space-time transformations, as it is a local gauge theory, which requires the particles' dynamics to remain unchanged under local gauge transformations. The invariance under the gauge symmetries of the theory and the subsequent quantisation of the model gives rise to the presence of gauge bosons, thus establishing a profound link between each force mediator and the corresponding gauge symmetry group.

### 2.1.2 Electromagnetic interaction

The Standard Model describes the electromagnetic force using Quantum Electrodynamics (QED), a relativistic quantum gauge theory. Mathematically, QED follows the symmetry group $U(1)$, which is commutative, and is hence defined as an Abelian gauge theory. The Lagrangian density of a free fermion can be written in the following way:

$$\mathscr{L}_f = i\bar{\psi}(x)\gamma^{\mu}\partial_{\mu}\psi(x) - m\bar{\psi}(x)\psi(x), \tag{2.1}$$

where $m$ is the mass of the fermion, $\psi(x)$ is a spinor of a fermionic field and $\bar{\psi}(x)$ is its adjoint spinor and $\gamma^{\mu}$ are the Dirac matrices. We require the Lagrangian density of

QED to be invariant under the local unitary Lie group, i.e. $U(1)$ transformation, which means it does not change under the gauge transformations of the field:

$$\psi(x) \rightarrow \exp^{iq\theta(x)}\,\psi(x), \tag{2.2}$$

with $q$ being the elementary electromagnetic charge and $\theta(x)$ the phase of the local transformation. Applying the 2.2 in the Lagrangian equation from 2.1 gives the following:

$$\mathscr{L} \rightarrow \mathscr{L}_f - q\bar{\psi}(x)\gamma^\mu\partial_\mu\theta(x)\psi(x), \tag{2.3}$$

which indicates that the Equation in 2.1 is not invariant under the $U(1)$ transformation, therefore extra terms namely the gauge field $A_\mu(x)$ and the covariant derivative $D_\mu$ are introduced:

$$A_\mu(x) \rightarrow A_\mu(x) + \partial_\mu\theta(x), \tag{2.4}$$

$$D_\mu = \partial\mu - iqA_\mu. \tag{2.5}$$

Substituting the above gauge field and the covariant derivative in the Lagrangian and adding the interaction of the fermion field with the scalar field, the Lagrangian of the field $A$ can be cast in the following form:

$$\mathscr{L}_A = -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} + m_A A^\nu(x)A_v(x) + q\bar{\psi}(x)\gamma^\mu\psi(x)A_\mu(x), \tag{2.6}$$

where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the field tensor and the term $m_A A^\nu(x)A_v(x)$ refers to the mass. The Lagrangian 2.6 is invariant under the local $U(1)$ transformation only in the case of the massless quantum field, $m_A = 0$, which is experimentally confirmed in the case of the photon. Therefore, the QED Lagrangian can be noted as follows:

$$\mathscr{L}_{QED} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi - ej^\mu A_\mu, \tag{2.7}$$

where $j^\mu = \bar{\psi}\gamma^\mu\psi$, the electromagnetic current density arises from Noether's theorem.

The electromagnetic interaction has a finite-range fine structure constant, namely $\alpha_{em}$, resulting from the neutral massless field. The $\alpha_{em} = \frac{e^2}{4\pi}$ determines the strength of the electromagnetic interaction. Since QED is a renormalizable theory, $\alpha_{em}$ varies with the energy of the interaction, making it a running coupling.

### 2.1.3 Strong interaction

The theory explaining the strong interaction is Quantum Chromodynamics (QCD), a gauge field theory that obeys the symmetry group $SU(3)$. QCD describes the strong interactions between particles carrying a colour charge, namely gluons and quarks. Similarly to QED, one can write the Lagrangian density for QCD as follows:

$$\mathscr{L}_{QCD} = \mathscr{L}_f + g_s\bar{\psi}\gamma^\mu\lambda_a\psi G^a_\mu - \frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a. \tag{2.8}$$

Where the eight component vectors $\lambda_a$ are the so-called Gell-Mann matrices and are derived from the $SU(3)$ generators $T^a$, with the following represntation below:

$$T_a = \frac{\lambda_a}{2}. \tag{2.9}$$

Unlike QED, the spinor $\psi$ is a three-component vector $\psi = (\psi_b, \psi_r, \psi g)$, where each of the components correspond to the Dirac spinor of the colours blue, red and green. In the QCD Lagrangian equation 2.8, $G^a_\mu$ represents the eight gauge or gluon fields with the following conserved colour current $J^a_\mu = g_s\bar{\psi}\gamma_\mu\lambda^a\psi$.

Similarly to the QED mediators, i.e. photons, the force carriers in QCD, i.e. gluons, are also massless. Nonetheless, since the QCD is a non-Abelian gauge theory, contrary to the QED, an extra term appears in the field strength tensor as follows:

$$G^{\mu\nu}_a = \partial^\mu G^\nu_a - \partial^\nu G^\mu_a - g_s f^{bc}_a G^\mu_b G^\nu_c. \tag{2.10}$$

This additional term comes from the non-commutativity of the gluon fields and represents the self-interactions (encoded in the $f^{abc}$, which are the structure constants of the $SU(3)$) of the gluon fields. Similar to QED electric charge, the quantity $g_s$ ( or $\alpha_s = \frac{g_s^2}{4\pi}$) represents the coupling constant in QCD. The latter is a function of the momentum transfer $Q$ as follows:

$$\alpha_s(Q^2) = \frac{12\pi}{(33 - 2N_f)\ln(\frac{Q^2}{\Lambda^2_{QCD}})}, \tag{2.11}$$

where $N_f$ is the number of active quark flavours and $\Lambda_{QCD} \sim 200MeV$ is an experimentally defined parameter, which describes the scale of the non-perturbative QCD regime. The QCD coupling constant $\alpha_s$ (2.11) decreases with $Q^2$, consequently for the very large $Q^2$ quarks are considered to be free; this phenomenon is understood as asymptotic freedom. At low $Q^2$ and distances of the order of a few fm, the inter-quark coupling increases; this makes the quarks confined into colour-neutral states, known as hadrons, a phenomenon called confinement.

### 2.1.4 Weak interaction & the electroweak theory

The third fundamental interaction discussed in this section is the weak interaction, historically starting from the Fermi model [13]. The latter was built to explain the $\beta$ decay of neutrons and muons at low energies. The massive gauge bosons $W^\pm$ and the neutral $Z$ boson mediate [14–16] the weak interaction and are responsible for the charged and neutral currents, respectively. The $W^\pm$ bosons couple fermions together, differing by one

unit of electric charge. Experimentally it was observed that this fermionic current is the only interaction in the SM in which parity is not conserved. Wu [17] observed for the first time parity violation in the experiment on the $\beta$ decay of cobalt. Parity violation is incorporated into the so-called vector-axial ($V - A$) form of the charged weak current as follows:

$$J^{(CC)\mu} = \bar{\psi}\gamma^{\mu}\frac{1}{2}(1-\gamma^{5})\psi. \tag{2.12}$$

Where the $J^{(CC)\mu}$ is the weak charged current. The above equation 2.12 indicates that the weak interaction only couples to the left-handed fermions (or right-handed anti-fermions). Another marker of the weak interaction is that it does not couple to the mass eigenstates of the quarks but rather to a linear combination of the quarks, i.e. the mixed weak states [14, 18, 19]. The link between the mixed weak states and the mass eigenstates of the three generations of quarks is delivered by the Cabibbo-Kobayashi-Maskawa (CKM) matrix as follows:

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{\mathrm{ud}} & V_{\mathrm{us}} & V_{\mathrm{ub}} \\ V_{\mathrm{cd}} & V_{\mathrm{cs}} & V_{\mathrm{cb}} \\ V_{\mathrm{td}} & V_{\mathrm{ts}} & V_{\mathrm{tb}} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}.$$

Where the $V_{ij}$ is proportional to the coupling of quark $i$ to quark $j$ and the $d', s', b'$ are the quark's weak eigenstates, whereas $d, s, b$ are the mass eigenstates. The elements of the CKM matrix above are derived experimentally [9], and the SM delivers no predictions. The existence of two charged heavy mediators $W^{\pm}$ is the consequence of the observation of the short range of the weak force, which corresponds to a heavy mediator and the presence of $\beta^{+}$ decays in addition to $\beta^{-}$ interactions, which suggests the presence of at least two mediators $W^{\pm}$.

In the theory of weak interactions, the neutral current mediator is the $Z$ boson, which is not involved in any flavour-changing currents:

$$J^{(NC)\mu} = \bar{\psi}\gamma^{\mu}\frac{1}{2}(g_{V}^{f} - g_{A}^{f}\gamma^{5})\psi, \tag{2.13}$$

where $g_{V}^{f}$ and $g_{A}^{f}$ are the vector and axial fermionic couplings. The Glashow-Weinberg-Salam (GWS) [14–16] electroweak unification theory requires a neutral current, as it will be discussed in the next section.

### 2.1.5 Electroweak unification

The unification of electromagnetic and weak interaction was formulated in the 1960s by Glashow, Weinberg and Salam (GWS) [15, 16, 20] . One of the main consequences of the

GWS model is the prediction of the weak neutral current equation 2.13 mediated by $Z$ boson. The unified electroweak theory is gauge invariant under the $SU(2)_L \times U(1)_Y$ gauge group structure. The conservation of the weak isospin $I$ is a consequence of the invariance under $SU(2)_L$, while the conservation of the weak hypercharge $Y$ finds its origins in the $U(1)_Y$ symmetry.

The essence of the EW unification theory is the connection between the fermionic electromagnetic charge $Q_f$, the third component of isospin $I_3$ and the weak hypercharge as follows:

$$Q_f = I_3 + Y/2. \tag{2.14}$$

The Lagrangian density of the EW theory under the $SU(2)_L \times U(1)_Y$ local gauge invariance can be written as follows:

$$\begin{aligned} \mathscr{L} =& \bar{\psi}_L(x)\gamma^\mu(i\partial_\mu + \frac{g_w}{2}\tau W_\mu + \frac{g'_w}{2}YB_\mu)\psi_L(x) + \\ &+ \bar{\psi}_R(x)\gamma^\mu(i\partial_\mu + \frac{g'_w}{2}YB_\mu)\psi_R(x) - \\ &- \frac{1}{4}W^i_{\mu\nu}W^{\mu\nu}_i - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}. \end{aligned} \tag{2.15}$$

Here $\tau$ represents the Pauli matrices [21], the $2 \times 2$ generators of the $SU(2)$ symmetry. The local gauge invariance of $SU(2)_L \times U(1)_Y$ is satisfied by introducing the $W_\mu$ and $B_\mu$ gauge fields, which correspond to respectively three non-abelian weak isospin fields that couple only to left-handed fermions with the coupling strength $g_w$ and the Abelian gauge field ($B_\mu$) that couples to both right- and left-handed fermions with the $g'$ coupling.

The physical states of the charged weak bosons is defined as the superposition of the $W^1_\mu$ and $W^2_\mu$ as follows:

$$W^\pm_\mu = \frac{1}{\sqrt{2}}(W^1_\mu \mp iW^2_\mu), \tag{2.16}$$

and the neutral current boson fields are derived as linear combinations of the $W^3_\mu$ and $B_\mu$:

$$\begin{pmatrix} \gamma \\ Z^0 \end{pmatrix} = \begin{pmatrix} \cos\theta_\mathrm{W} & \sin\theta_\mathrm{W} \\ -\sin\theta_\mathrm{W} & \cos\theta_\mathrm{W} \end{pmatrix} \begin{pmatrix} B_\mu \\ W^3_\mu \end{pmatrix},$$

where $\theta_W$ is the weak mixing angle. The Lagrangian of the Electroweak (EW) is invariant under its symmetry group only if all EW gauge bosons are massless, which is in contrast to the experimentally observed massive $W^\pm$ (80.379 GeV [9]) and $Z$ (91.188 GeV [9]) bosons. The masses of the bosons mentioned earlier are introduced to the SM Lagrangian via spontaneous symmetry breaking, as discussed in the following.

### 2.1.6 Spontaneous symmetry breaking & the Higgs mechanism

As discussed above, the SM electroweak interactions are invariant under the $SU(2)_L \times U(1)_Y$ group symmetries; however, gauge bosons are required to be massless under the symmetry mentioned above. The mechanism that provides a general framework to generate the masses of the $W$ and $Z$ bosons is the Electroweak Symmetry Breaking (EWSB) [22]. The EWSB mechanism suggests the presence of a self-interacting complex scalar field in the electroweak sector. This field's CP-even neutral component gains a vacuum expectation value (VEV)$\sim$246 GeV, which defines the scale at which the electroweak symmetry breaks. Consequently, this process generates three massless Goldstone bosons, which are absorbed to provide mass to the W and Z gauge bosons. The remaining part of this complex scalar doublet transforms into the Higgs boson, thus far the only known scalar particle. Furthermore, the masses of all fermions originate from the EWSB mechanism, as the Higgs doublet couples to fermions via Yukawa interactions. So far, the Yukawa interactions with the heaviest fermions, such as top quark, have been experimentally verified.

The rest of this section is dedicated to the formulation of the spontaneous symmetry-breaking, also known as the Brout-Englert-Higgs mechanism (BEH), which was proposed independently by Higgs [6], Brout and Englert [5] in 1964. The Higgs mechanism introduces two complex scalar fields, which provide the necessary four degrees of freedom as follows:

$$\Phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}. \tag{2.17}$$

Since the Higgs mechanism generates the masses of the electroweak gauge bosons, $\phi^0$ must be neutral, and $\phi^+$ must be charged in a way that $(\phi^+)^* = \phi^-$ give the longitudinal degrees of freedom of the $W^+$ and $W^-$. In Equation 2.17, the lower and upper components of the doublet differ by one unit of charge. The corresponding Lagrangian density of the isospin doublet above takes the following form:

$$\mathscr{L}_{Higgs} = (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(\Phi), \tag{2.18}$$

where $D_\mu = \partial_\mu - i g_w \frac{\tau}{2} W_\mu - i g'_w \frac{Y}{2} B_\mu$ is the covariant derivative which couples the scalar doublet to the SM lagrangian and $V(\Phi)$ is the associated Higgs potential as follows:

$$V(\Phi) = \mu^2 \Phi^\dagger \Phi + \lambda (\Phi^\dagger \Phi)^2, \tag{2.19}$$

where $\mu^2$ and $\lambda$ are two new free parameters of the theory, the above form of the $V(\Phi)$ not only maintains the renormalizability of the SM but also is gauge invariant under $SU(2)_L \times U(1)_Y$ group symmetries. In the case of $\mu^2 > 0$, the potential maintains the symmetries of the Lagrangian with the minimum occurring at $\Phi^\dagger = \Phi = 0$. Hence, the masses for the $Z$ and $W$ bosons are not generated, and the theory is equivalent to QED

with massless photon and charged scalar field $\phi$ and mass of $\mu$. However, for $\mu^2 < 0$, the potential $V$ has an infinite set of degenerate minima with the so-called "Mexican Hat" shape illustrated in Figure 2.2 that satisfies the following condition:

$$\Phi^\dagger\Phi = \frac{1}{2}(\phi_1^2 + \phi_2^2 + \phi_3^2 + \phi_4^2) = \frac{v^2}{2} = -\frac{\mu^2}{2\lambda}, \tag{2.20}$$

where $v$ is the vacuum expectation value (VEV) of $\Phi$ and $v = \sqrt{-\mu^2/\lambda} \sim 246$ GeV is the VEV of the Higgs field.

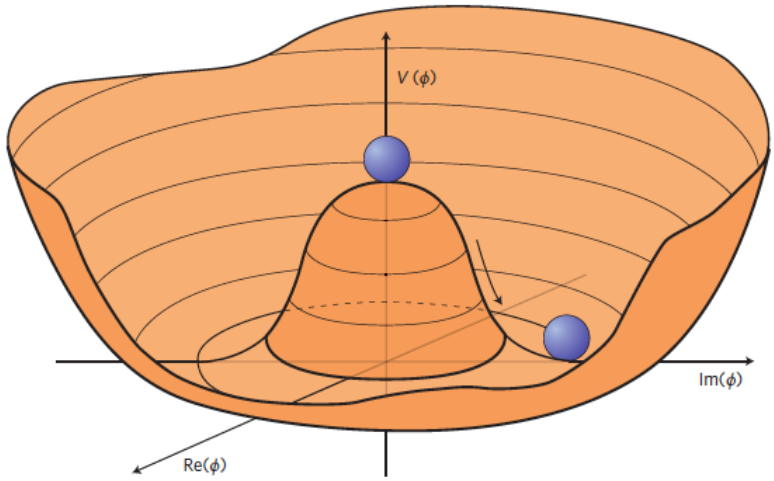


Figure 2.2: Figure shows the Higgs potential shape for $\mu^2 < 0$ with the minimum at $v = \sqrt{-\mu^2/\lambda}$. [23]

Choosing one minimum lying around the circle of the radius of VEV breaks the symmetry spontaneously. The minimum is selected as follows:

$$\langle 0|\Phi|0\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\v\end{pmatrix}. \tag{2.21}$$

By perturbing the chosen ground state 2.21 as follows:

$$\Phi = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\v+h\end{pmatrix}, \tag{2.22}$$

$h = h(x)$ represents the excitation of the Higgs field or, in other words, a massive Higgs boson and, substituting it in the Lagrangian density, one can infer the mass of the Higgs as $M_H = \sqrt{-2\mu^2}$, which is a free parameter of the SM. Additionally, the masses of the massive gauge bosons can be derived from the Higgs coupling to weak vector boson terms in the Lagrangian density as follows:

$$m_W = \frac{1}{2}|g|v, \quad m_{Z^0} = \frac{1}{2}v\sqrt{g^2 + g'^2} = \frac{m_W}{\cos\theta_W}. \tag{2.23}$$

The electroweak symmetry-breaking mechanism only accounts for the generation of the vector boson masses. To obtain the mass of fermions, one has to introduce by hand in the Lagrangian a new term describing the interaction of the fermions with the Higgs field. This process is known as the Yukawa interaction, and it is a renormalizable interaction between the scalar Higgs field and the fermionic fields that can be written as follows:

$$\mathscr{L}_{Yukawa} = -g_Y \bar{\psi}\Phi\psi, \tag{2.24}$$

where $g_Y$ is the Yukawa coupling constant. The mass term of the fermions can be derived after the spontaneous symmetry breaking:

$$m_f = \frac{1}{\sqrt{2}} g_Y^f v \tag{2.25}$$

where $g_Y^f$ is the coupling constant directly proportional to the fermion mass.

The Higgs boson discovered almost ten years ago at the LHC, is the only fundamental scalar (zero-spin) state in the Standard Model (SM). So far, the precise measurements of the Higgs boson properties, such as its mass, spin and decay width, have been shown to align with the SM's theoretical expectations. In the upcoming section, a comprehensive discussion of the production and decay processes associated with the Standard Model Higgs boson will be discussed.

### 2.1.7 The SM Lagrangian

The SM Lagrangian can be written as follows by combining all the above-discussed components:

$$\begin{aligned}
\mathscr{L} =& \mathscr{L}_{gauge} + \mathscr{L}_{EW} + \mathscr{L}_{QCD} + \mathscr{L}_H + \mathscr{L}_{Yukawa} = \\
& -\frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a - \frac{1}{4}W^{i,\mu\nu}W^i_{\mu\nu} - \frac{1}{4}B^{\mu\nu}B_{\mu\nu} \\
& + \sum_{\psi\in q,l} \bar{\psi}_L i D^L_\mu \gamma^\mu \psi_L + \sum_{\psi\in u,d,e} \bar{\psi}_R i D^R_\mu \gamma^\mu \psi_R \\
& + (D^\mu\Phi)^\dagger D_\mu\Phi - \mu^2\Phi\Phi^\dagger + \frac{1}{4}\lambda(\Phi\Phi^\dagger)^2 \\
& [-y_u \bar{u}_R \Phi^\dagger q_L + y_l \bar{l}_L \Phi^\dagger l_R + h.c.],
\end{aligned} \tag{2.26}$$

where $D_\mu = \partial_\mu + igW^i_\mu\frac{\sigma^i}{2} + ig'B_\mu Y + ig_s G^a_\mu\frac{t^a}{2}$ is the covariant derivative and satisfies the $SU(3)_C \times SU(2)_L \times U(1)$ symmetry.

## 2.2 Higgs production modes & decays

After decades of effort, the Higgs boson was discovered in July 2012 at the Large Hadron Collider (LHC) by the ATLAS and CMS experiments [3, 4]. The Feynman diagrams corresponding to the main production mechanisms for the Higgs boson at the LHC, are illustrated in Figure 2.3. Ranked by their cross sections, the main production mechanism for the SM Higgs boson are listed below:

1. Gluon-Gluon Fusion (ggH): The primary production mode of the SM Higgs boson at the LHC is the gluon-gluon fusion, denoted as ggH (Figure 2.3 a). This process happens through intermediary quark loops since there are no direct couplings between SM Higgs boson and gluons. Given that the fermionic Yukawa couplings are proportional to the quark masses, the dominant contribution to the loop is attributed to the top quarks. As shown in Figure 2.5 on the left, the cross-section for ggH production exceeds other production mechanisms by an order of magnitude.

2. Vector Boson Fusion (VBF): The second largest cross-section for Higgs boson production at the LHC is the vector boson fusion process 2.3 b). Each initial quark emits a W or Z boson in this production mode, which fuses to produce the Higgs boson. This process is characterized experimentally by the existence of two forward or backward light-flavour jets, offering a means to reduce background noise and improve experimental sensitivity effectively.

3. Higgsstrahlung (VH): The Higgs production in association with a vector boson, shown in Figure 2.3 c, has the third largest cross-section at the LHC. In this procedure, a virtual W or Z boson decays to produce the Higgs. The Higgs recoils against leptons with high momentum or jets from the decay of the vector boson.

4. Higgs production in association with top quarks ($\bar{t}tH$): This production mechanism involves the Higgs boson associated production with a $t\bar{t}$ pair (Figure 2.3 d), in which the Higgs boson production cross section is small, roughly equal to 0.5 pb. Although the $t\bar{t}H$ production has a lower cross-section production than the other production modes listed, it can offer a unique way to test the direct Yukawa coupling between the Higgs boson and the top quark, the heaviest particle.

5. Higgs production in association with bottom quarks (bbH): This production mechanism is similar to the ttH one, but the main distinction is that the Higgs is produced with a bottom anti-bottom pair rather than a top anti-top pair. This production mechanism will be presented for the first time in the current work and has never been studied for the SM Higgs boson.

The Higgs production mechanisms discussed above and their cross-sections for a proton-proton collision are plotted versus center of mass energy $\sqrt{s}$ in Figure 2.4. The

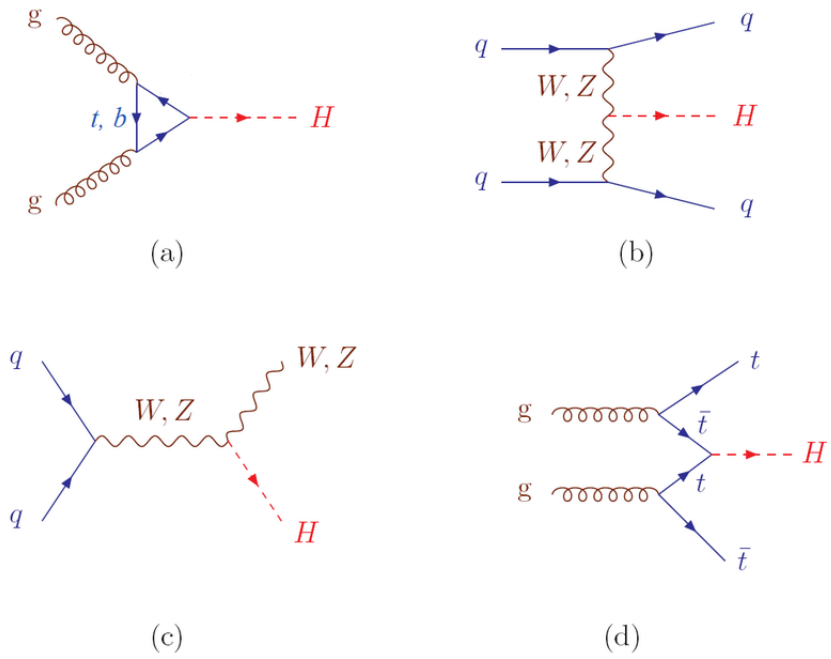


Figure 2.3: Feynman diagram of the different Higgs production: gluon-gluon fusion (top left, a), vector boson fusion (top right, b), Higgs-strahlung (bottom-left, c) and top associated production (bottom right, d) [24].

ggH production labelled $pp \to H$ is one of the most frequent hard scattering processes generated at the LHC, which is a direct consequence of gluons dominating the proton parton distributions function (PDF) [25] at a low momentum fraction and energy scale of interest, which is around 100 GeV. At the LHC, the VBF generation process with the second greatest cross-section is $pp \to qqH$ (Fig. 2.4). The W- and Z-strahlung production modes, denoted as $pp \to WH$ and $pp \to ZH$ in Fig. 2.4, are additional processes where the Higgs boson is produced through coupling with vector bosons. Competitive cross-sections are seen in the heavy-quark related productions with top and bottom quarks; however, only the top-associated production has been observed [26].

The SM Higgs boson is an unstable and massive particle that currently can be detected at the LHC through its decay products. Within the Standard Model, a range of potential decay channels for the Higgs boson exists with their respective probabilities, quantified as branching ratios, illustrated in Figure 2.5 on the right.

The dominant decay mode for the Higgs boson is the decay into a $\bar{b}b$ pair. The observation of this hadronic final state is challenging due to the large QCD multi-jet background. In spite of this, the $H \to \bar{b}b$ decay was discovered in association with vector boson-associated Higgs production mode by both ATLAS [28] and CMS [29].

The second largest branching ratio among the Higgs boson decay modes is the decay

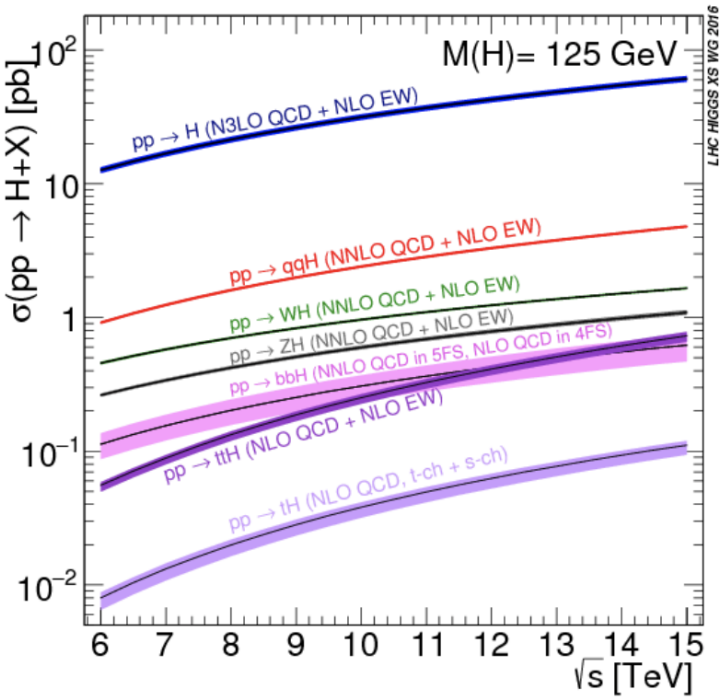


Figure 2.4: The SM Higgs boson production cross section versus the center of mass energy $\sqrt{s}$ for proton proton collisions at the LHC, the theoretical uncertainties are showed as bands [27].

into a W boson pair. This decay mode allows a very competitive measurement to detect directly the Higgs boson at the LHC, and it represents one of the channels of interest for the present work.

The second largest branching ratio among the fermionic decays is $H \to \tau^+\tau^-$, which happens approximately 6.3 % of the time. This channel is especially attractive due to its high branching ratio and lower background contributions compared to $H \to b\bar{b}$ and the potential of probing the Higgs coupling to fermions. The $H \to \tau^+\tau^-$ channel is also used to analyse the present work.

Among the Higgs decay channels, $H \to ZZ$ is known as one of the cleanest processes at the LHC. Despite the small branching ratio, the low background associated with the four charged lepton final states leads to a signal-to-background ratio significantly greater than one.

Since the SM Higgs boson couples only to massive particles, it decays into a pair of photons, or gluons can only occur via intermediate boson or quark loop, respectively. Although the gluon decay mode remains inaccessible at the LHC due to the large background, the $H \to \gamma\gamma$ decay mode, despite its relatively low branching ratio, stands out as one of the most sensitive due to its relatively low background and high resolution of

the $m_{\gamma\gamma}$ distribution. This is due to the excellent electromagnetic energy and momentum resolution provided by the ATLAS and CMS detectors.

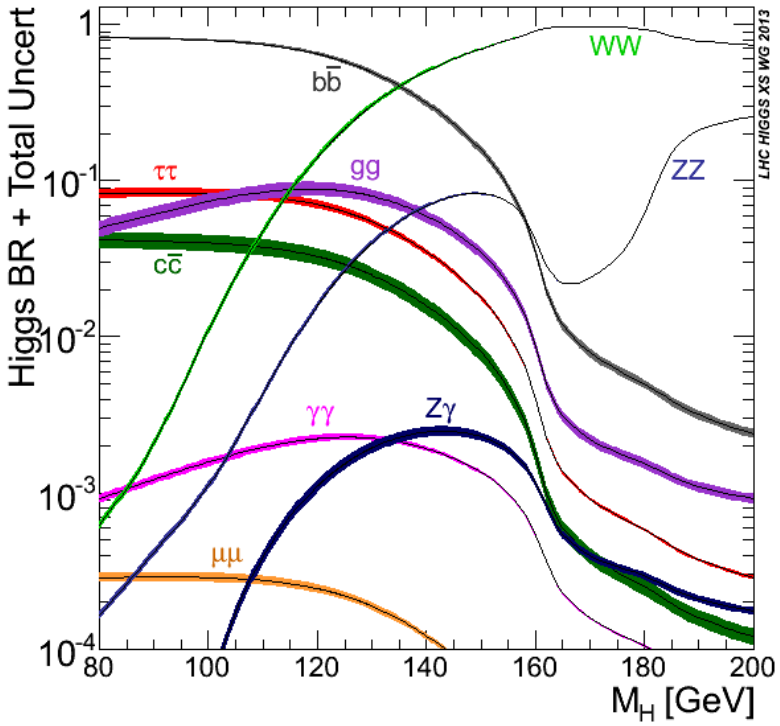


Figure 2.5: The main branching ratios of the SM Higgs boson. [27]

## 2.3 Status of 10 years experimental Higgs boson

Nearly 12 years since the discovery of the Higgs boson, a significant understanding of its properties has been achieved, thanks to data from producing 30 times more Higgs bosons over the past decade.

The general-purpose detectors at CERN (ATLAS [30] & CMS [31]) have detected the Higgs boson in diverse fermionic and bosonic decay channels. The characteristics of the Higgs boson, such as its spin-parity quantum numbers, mass and production cross-sections across different modes, were determined. This section will discuss the most recent results regarding the Higgs boson characteristics. These results are based on data from proton-proton collisions (measured by the CMS collaboration [32]) at a centre-of-mass energy of 13 TeV corresponding to the Run 2 data. The Higgs boson mass was measured to be $125.25 \pm 0.17$ GeV [9] using $H \to \gamma\gamma$ and $H \to ZZ \to 4l$ decay channels combining the results from Run 1 and Run 2. The natural width of the Higgs boson was determined

to be $\Gamma_H = 3.2^{+2.4}_{-1.7}$ [33] using the ratio of off-shell and on-shell Higgs boson production cross sections, which are distinguished by the emission of a virtual and a real Higgs boson, respectively. Additionally, with the Run 2 data set, CMS has observed the Higgs boson decaying into $\tau$ leptons with a statistical significance of 5.9 sigma($\sigma$), into a bottom quark pair with (5.6 $\sigma$), and in the $ttH$ production mode (5.2 $\sigma$). The Higgs boson decay channel into muons was also measured to be (3 $\sigma$).

In CMS experiments, the signal strength parameter, commonly referred to as $\mu$, is crucial for assessing how well the observed signal yields align with the predictions of the Standard Model (SM). This parameter $\mu$ represents the ratio of the observed signal to the expected SM prediction.

The interpretation of $\mu$ changes based on the analysis being conducted. Specifically, for each production channel, denoted as $\mu_i$, and each decay mode, indicated by $\mu_f$, in transitions from an initial state $i$ to a final state $f$ through a process $i \to H \to f$, we define $\mu_i$ as the ratio of the observed production cross-section $\sigma_i$ to the production cross-section predicted by the Standard Model $(\sigma_i)_{SM}$. Similarly, $\mu^f$ is defined as the ratio of the observed branching fraction $B^f$ to the branching fraction predicted by the Standard Model $(B^f)_{SM}$. A $\mu$ equal to 1 indicates a perfect agreement with the Standard Model predictions.

At the time of the Higgs discovery, the first test of compatibility was performed by fitting the data from all production and decay modes with a common signal strength $\mu$, which was found to be $0.87 \pm 0.23$. The latter equals $\mu = 1.002 \pm 0.057$ for the combination of the Run 2 data, and it aligns with the SM. By introducing different $\mu_i$ and $\mu^f$, corresponding to the signal strength parameter for different production modes and decay modes, one can summarise the other measurements, as illustrated in Figure 2.6.

The interaction between the Higgs boson and the coupling modifier can be analyzed using the $\kappa$ framework, as detailed in [34]. The $\kappa$ parameter scales the Higgs boson's interaction with specific particles, influencing its production by affecting its cross-section and decay by affecting the decay width. A $\kappa$ value equal to one signifies the predictions of the Standard Model (SM). Observed research on the Higgs boson's couplings, as parameterized by $\kappa$ with both fermions and gauge bosons, demonstrates a significant consistency with the SM predictions across various mass ranges, as illustrated in Figure 2.7. Due to the mass-dependent nature of the Higgs boson's coupling, studying its interaction with first and second-generation fermions presents a significant experimental challenge. Since these fermions' masses are substantially lower, by at least an order of magnitude, compared to those of the third-generation fermions.

Nevertheless, observing and measuring these processes, whether in production or decay, remains vital to confirm further that the observed particle with a mass of approximately 125 GeV is indeed the SM Higgs boson. Recent evidence regarding the Higgs boson's decay into $\mu^+\mu^-$ [35] and the search for $VH \to cc$ decays [36] represent promising steps in this direction.

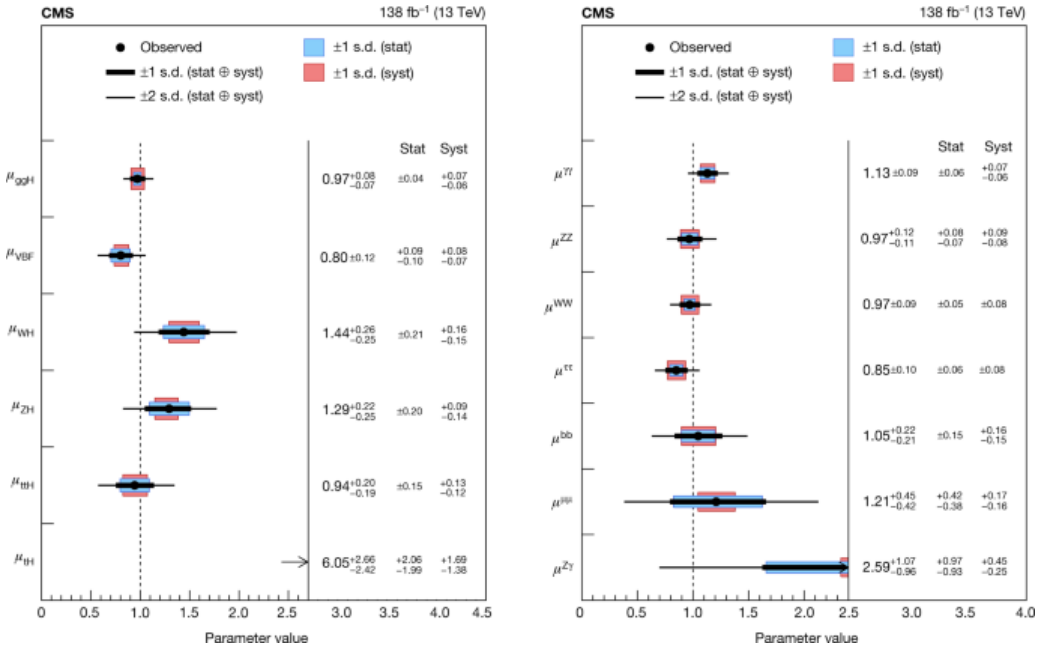


Figure 2.6: The figures show the signal strength and agreement with the SM for different production modes $\mu_i$ on the left and decay channels $\mu^f$ on the right. The thick and thin black lines represent the 1 and 2 standard deviations, and the blue and red bands, respectively, indicate the statistical and systematics components of the 1 standard deviations [33].

## 2.4 Higgs production in association with bottom quarks

So far, almost all of the Higgs boson production modes shown in Figure 2.5 have been observed at the LHC. This thesis presents the first search for one of the rare production mechanisms of the SM Higgs boson, namely the b-associated production of the Higgs (bbH). The bbH production with a predicted cross-section of 0.48 pb [37] features a comparable rate to the $ttH$ production mode (0.51 pb), yet, with more background contribution compared to $ttH$. Hence, this makes studying bbH more challenging due to the sizeable irreducible background. Feynman diagrams for the bbH productions are illustrated in Figure 2.8.

The experimental study of bbH production is of particular interest due to the possibility of performing a direct measurement of the Higgs coupling to bottom quarks (bottom quark Yukawa coupling $y_b$). A direct sensitivity to $y_b$ can be obtained by studying the $H \to b\bar{b}$ decay or the associated production of the Higgs boson with the bottom quark pair. The $H \to b\bar{b}$ has already been observed [39], and no LHC measurements have been dedicated to the bbH production so far [2]. Consequently, a search for bbH can be used

[2]Investigations into beyond the Standard Model (BSM) Higgs production in final states with numerous

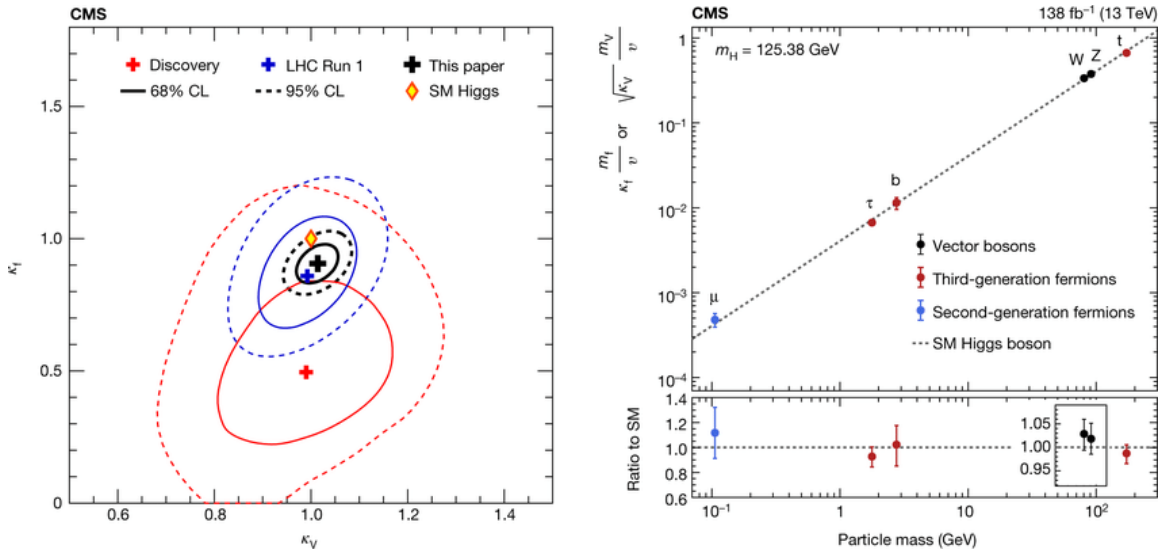


Figure 2.7: (left) Figure presents the constraints on the coupling modifiers of the Higgs boson to fermions ($\kappa_f$) and heavy gauge bosons ($\kappa_V$) across different data sets: the initial discovery set (depicted in red), the dataset from LHC Run 1 (illustrated in blue), and the data from Run 2 (represented in black). The point where $\kappa_V = \kappa_f = 1$, aligning with the Standard Model (SM) predictions, is denoted by a diamond symbol. (Right) This figure illustrates the calculated modification factors for the couplings of the Higgs boson to both fermions and heavy gauge bosons, shown as a function of their masses. Here, $\nu$ symbolizes the vacuum expectation value of the BEH field. For the heavy gauge bosons, the square root of the coupling modifier is displayed to maintain the linear relationship with the mass of the gauge boson, which is consistent with SM predictions. [33]

in addition to $H \to b\bar{b}$ results to enhance bottom Yukawa coupling property results and complete the investigation of SM Higgs boson production modes.

Besides the experimental motivation, the bbH process is also exciting from the theoretical point of view and in the context of higher-order QCD corrections.

Figure 2.8 displays the primary Feynman diagrams representing the production of the Higgs boson in association with b-quarks. The diagrams on the left side represent the production of a Higgs boson and a $b\bar{b}$ pair through the interaction of gluons and b-quarks ($y_b$) via Yukawa coupling.

The diagram located in the middle can be identified as gluon fusion (or VBF), with the presence of an extra gluon splitting into a pair of $b\bar{b}$ particles. The figure depicts

b quarks have taken place within the framework of theories that propose an extended scalar sector, including the two-Higgs-doublet model (2HDM) and the minimal supersymmetric Standard Model (MSSM). However, the coupling of the Higgs boson to bottom quarks is enhanced in both the MSSM and Type-II and Flipped 2HDM models compared to the SM search, performed for the first time in the present work.

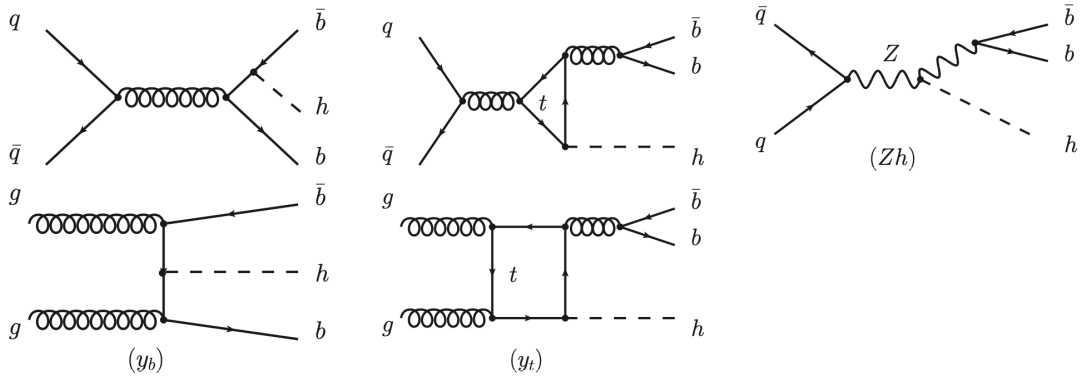


Figure 2.8: Feynman diagram of the bbH production modes, grouped by their amplitude, with the direct coupling of the Higgs to bottom quarks on the left, Higgs coupling to top loop in the middle and the amplitude corresponding to the Higgs coupling to $Z$ on the right [38].

the production of the Higgs particle through a quark loop, with a significant contribution from the top quark. The cross-section is primarily influenced by the Higgs Yukawa interaction to the top quark ($y_t$), resulting in a dominant contribution. The cross-section for this process is obtained by taking a portion of the NLO gluon fusion cross-section, with calculations conducted in the four-flavour scheme [40]. This requires the conversion of the gluon into a $b\bar{b}$ pair. This process destructively interferes with the diagram on the left side of Figure 2.8.

The diagram on the right side shows the Higgs-strahlung process, where the production of b-quarks occurs through the decay of a Z boson on-shell. This diagram exclusively includes electroweak terms in leading order. The interference observed in the diagram on the left is minimal compared to the cross-section of the individual processes. In this analysis, the Higgs-strahlung mechanism is considered a background in the search for bbH production.

The notation used to differentiate the various contributions to the production of the Higgs boson in association with b-quarks is as follows: The term "bbH" is used to refer to the production of the bottom-associated Higgs, excluding only the Higgs-startling process. "bbH($y_b^2$)" refers to the production mechanisms where the Higgs is produced through the Higgs Yukawa coupling to bottom quarks. "bbH($y_t^2$)" indicates the Higgs production through the top quark loop. "bbH($y_t y_b$)" represents the interference term between the top-mediated production and the b-quark mediated one.

The prediction for the bbH production cross-section in the Standard Model (SM) is divided into three parts, each corresponding to specific cross-sections [40], which are summarized in Table 2.4.

| term | cross-section (pb) |
|---|---|
| $y_t^2$ (NLO reweighted to N3LO) | 1.040 |
| $y_b^2$ (NLO) | 0.482 |
| $y_t y_b$ | - 0.033 |
| total | 1.489 |

In this analysis, the SM Higgs associated with b-quarks has been studied, where the Higgs boson further decays to $\tau\tau$ and $WW$.

CHAPTER 3

# The CMS experiment at the LHC

Particle accelerators are the key experimental tools designed to accelerate and collide high-energy particles in a controlled environment. These collisions provide valuable insights into the fundamental particles and the forces governing them. This chapter describes the Large Hadron Collider (LHC) [41], the most advanced and largest particle accelerator built to date, followed by a concise discussion of one of the general-purpose experiments at the LHC, the Compact Muon Solenoid (CMS) [31].

## 3.1 The Large Hadron Collider

The Large Hadron Collider (LHC), situated at CERN near Geneva, Switzerland, is the largest and very powerful particle accelerator built with the aim to study the fundamental structure of matter. LHC is also, at present, the largest accelerator, with a circumference of approximately 27 kilometres, and it is situated underground. Todays LHC tunnel, which lies 50 to 175 meters underground at the France-Switzerland border, was once used for the Large Electron-Positron (LEP) [42] collider.

The Large Hadron Collider (LHC) facilitates collisions between protons, protons and heavy ions (p-Pb), and heavy ions themselves (Pb-Pb, Xe-Xe). To optimize the rate of interactions per unit of time, particles within each beam are grouped into bunches.

The LHC accelerates particles through two separate beam pipes, maintaining a vacuum pressure between $10^{-10}$ and $10^{-11}$ mbar. To bend the particle beams, LHC employs 1232 dipole magnets, each 15 meters in length. Moreover, 392 quadrupole magnets, ranging from 5 to 7 meters in length, are used for beam focusing. Additional sextupole, octuple, and decapole magnets correct minor magnetic field imperfections. These superconducting magnets are made from niobium-titanium coated in copper and kept at a temperature 1.9K using superfluid helium-4 and a sophisticated vacuum system for insulation.

Prior to entering the LHC, protons are step-by-step energized through multiple stages

by other accelerators. This process boosts the protons to the desired energy level for experiments at the LHC, as shown in Figure 3.1.

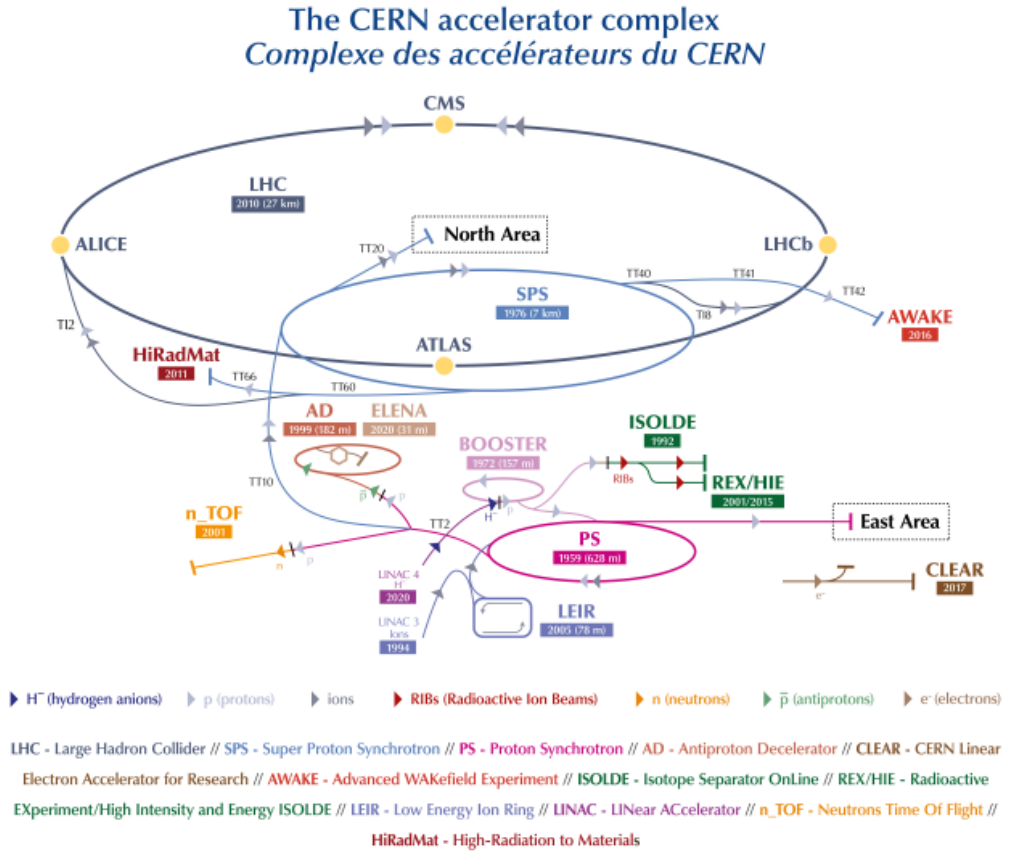


Figure 3.1: Schematic of the CERN accelerator complex including different experiments [43].

The process of acceleration at the LHC starts with acquiring the protons. The procedure starts from the generation of hydrogen ions $H^-$ at an energy of 160 MeV using the Linac 4 linear accelerator, which began operations in 2020, replacing the formerly operated Linac 2. Later, these ions enter the Proton Synchrotron Booster (PSB), where electrons are expelled from the atom, leaving only a nucleus with a single proton. Now, at 2 GeV, these protons proceed to the Proton Synchrotron (PS), which their energy is boosted up to 25 GeV. Following the PS, the protons enter the Super Proton Synchrotron (SPS), where their energy is further risen to 450 GeV before being injected into the ring of the LHC. In this ring, they accumulate and are accelerated to achieve the desired center-of-mass energy.

During the first data-taking period since the start of the operations, LHC Run1, the LHC reached proton-proton collision at a centre-of-mass energy of $\sqrt{s} = 7$ TeV in 2010 and 2011, later by an increase to $\sqrt{s} = 8$ TeV in 2012. In the subsequent Run 2 period from 2016 to 2018, the LHC operated at a centre-of-mass energy of $\sqrt{s} = 13$ TeV. The ongoing Run 3 period involves the LHC working at a centre-of-mass energy of $\sqrt{s} = 13.6$ TeV. Two Long Shutdowns (LS) took place between the Runs mentioned earlier to upgrade the LHC and its detectors to enable their operation at the higher centre-of-mass energy and collision rate. The collision rate is an essential parameter for the LHC and is referred to as instantaneous luminosity, defined as follows:

$$\mathscr{L} = \frac{N_b^2 n_b f_{rev} \gamma_r}{4\pi \epsilon_n \beta *} F. \tag{3.1}$$

Where $N_b$ denotes the number of particles in each bunch, $n_b$ represents the number of bunches in each beam, $f_{rev}$ is the beam's revolution frequency, $\gamma_r$ stands for the beam's relativistic gamma factor, $\epsilon_n$ is the normalized transverse beam emittance, and $\beta*$ refers to the beta function at the collision point. Moreover, $F$ signifies the geometric reduction factor, defined as follows:

$$F = (1 + (\frac{\theta_c \sigma_z}{1\sigma *})^2)^{-1/2}. \tag{3.2}$$

$\theta_c$ represents the full crossing angle at the interaction point, while $\sigma_z$ denotes the root mean square (RMS) value of the bunch length. Additionally, $\sigma*$ indicates the RMS value of the transverse beam size at the interaction point. Given the instantaneous luminosity, it's possible to calculate the total event rate $N$ for a given physical process using the equation below:

$$\dot{N} = \mathscr{L} \cdot \sigma, \tag{3.3}$$

where $\sigma$ defines the cross-section of the physical process of interest.The LHC was initially designed to reach the instantaneous luminosity of $\mathscr{L} = 10^{34} cm^{-2} s^{-1}$, which was reached in 2017 and twice the value mentioned above was reached during the Run 2 period. The total amount of collision data delivered by the collider is quantified by integrating the instantaneous luminosity over a specified time period as follows:

$$L = \int \mathscr{L} dt. \tag{3.4}$$

At the Large Hadron Collider (LHC), nine experiments employ detectors to examine the diverse range of particles generated by collisions within the accelerator. Global collaborations of scientists from many institutes conduct these experiments. Every experiment is unique and differentiated by its detector. Some of the experiments at the LHC are seen in Figure 3.2; a brief description of the nine experiments is provided below:

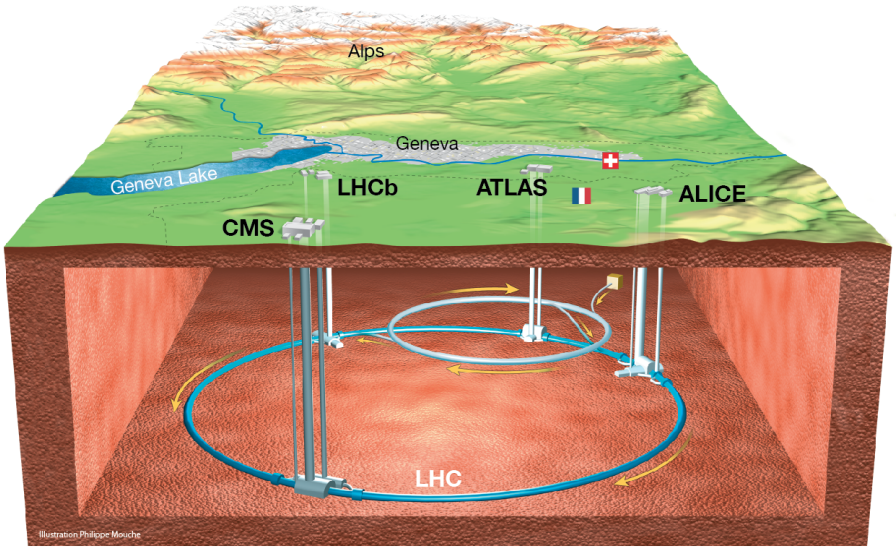


Figure 3.2: general view of the five experiments at the interaction points and the LHC tunnel [44].

1. CMS ("Compact Muon Solenoid") [31] and ATLAS ("A Toroidal LHC ApparatuS") [30] are positioned at Point 5 (P5) and Point 1 (P1) respectively, serving as two general-purpose detectors. These detectors are engineered to investigate a wide range of phenomena, spanning from understanding the Higgs mechanism to probing for new physics, such as candidates for dark matter and extra dimensions.

2. LHCb ("Large Hadron Collider beauty") [45], situated at Point 8 (P8), specializes in the study of heavy flavor physics. Its scope contains studies ranging from CP violation to the exploration of exotic hadron spectroscopy.

3. ALICE ("A Large Ion Collider Experiment") [46], located at Point 2 (P2), is purposefully designed to investigate the heavy-ion collisions. Its primary objective is to provide valuable insights into the properties of the quark-gluon plasma.

4. TOTEM ("The Total, elastic and diffractive cross-section measurement") [47] and LHCf ("Large Hadron Collider forward") [48] constitute the two smallest experiments at the LHC, with a specific focus on "forward particles" – protons or heavy ions – that slide past each other rather than colliding head-on at the beam intersection points. LHCf comprises two detectors positioned along the LHC beamline, 140 meters apart from the ATLAS collision point, while TOTEM utilizes detectors placed on either side of the CMS interaction point.

5. Three additional experiments at CERN include SND@LHC ("Scattering and Neutrino detector at the LHC") [49], FASER ("Forward Search Experiment") [50], and MoEDAL-MAPP ("Monopole and Exotics detector at the LHC") [51]. MoEDAL-MAPP focuses on the search for a hypothesized particle called the magnetic monopole, employing detectors positioned near LHCb. The two newest experiments, FASER and SND@LHC, are situated close to the ATLAS collision point, aiming to explore neutrinos and search for new light particles.

The analysis in this work uses data collected by the CMS detector during the Run 2 period (2016-2018), amounting to an integrated luminosity of 137.62 $fb^{-1}$. Figure 3.3 provides a comparison of the luminosity delivered by the CMS experiment in both Run 1 and Run 2. Subsequent sections of this chapter focus exclusively on the CMS experiment, as it is the primary experiment of interest for this analysis.

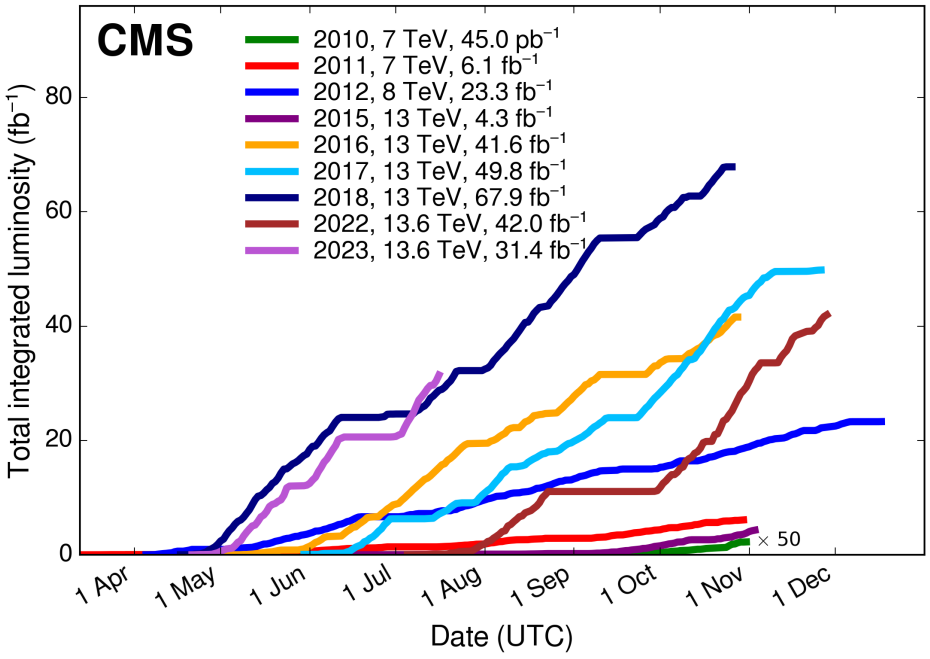


Figure 3.3: Delivered Luminosity for proton-proton collision at CMS experiment versus time, for Run 1 and Run 2 [52].

The continuous growth of instantaneous luminosity, as shown in Figure 3.3 from Run 1 till the current Run 3, is due to the increase in the number of proton bunches, which leads to a higher frequency of inelastic pp collisions. Subsequently, this leads to the generation

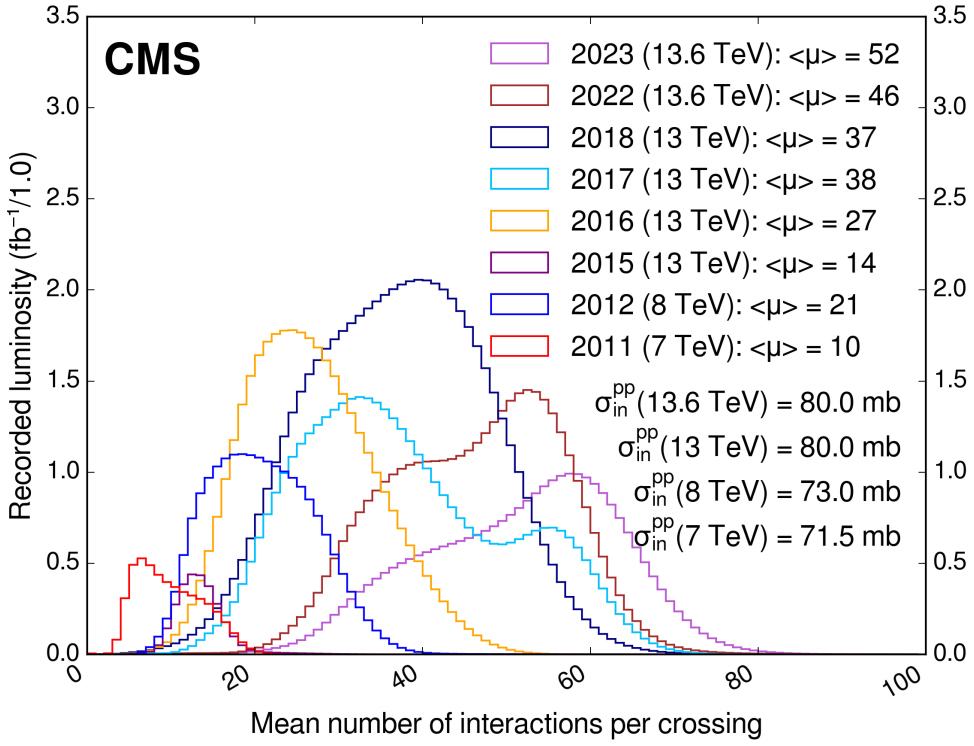


Figure 3.4: Figure shows the mean number of pileup events (number of interactions per bunch crossing) for Run 1 (2011-2015), Run 2 (2016-2018) and the ongoing Run 3 (2022-2023). Additionally, the number of pp inelastic cross-sections are shown for the different centres of mass energies corresponding to each Run [52].

of the so-called Pileup (PU) events, which arise when the detectors record signals from not just the primary collision of interest but also from other secondary collisions occurring simultaneously. The mean number of PU interactions is denoted as $< \mu >$, and the comparison of the mean number of PU events from Run 1, Run 2 and Run 3 are given in Figure 3.4.

Generally, two types of pile-up interactions can be identified: in-time (IT) and out-of-time (OOT) pileups. The in-time pileup happens simultaneously with the hard scattering vertex within the same bunch crossing. The out-of-time pileup originates from soft particles generated in previous collisions, reaching the detector during subsequent collisions. The high pileup environment affects the accurate identification of the genuine hard scattering vertex and the reconstruction of kinematic quantities, such as missing transverse energy and jets. Luckily, methods have been developed at CMS to mitigate the effect of pileup.

## 3.2 The CMS experiment

The Compact Muon Solenoid detector [31] is one of the two general-purpose apparatuses at the LHC developed to cover a broad range of research in particle physics at high energies, such as tests of the SM at the TeV scale, study of the Higgs boson and searches for the physics beyond the SM. The CMS detector is located around 100 meters underground at the Point 5 collision area of the LHC near Cessy, France. The CMS detector weighs around 14000 t, has a length of 21.6 m, and has a diameter of 15 m. However, given these dimensions, it is named a "compact" detector due to the wide range of other fundamental components fitted in its magnet. Figure 3.5 illustrates the CMS detector, including its subdetectors.

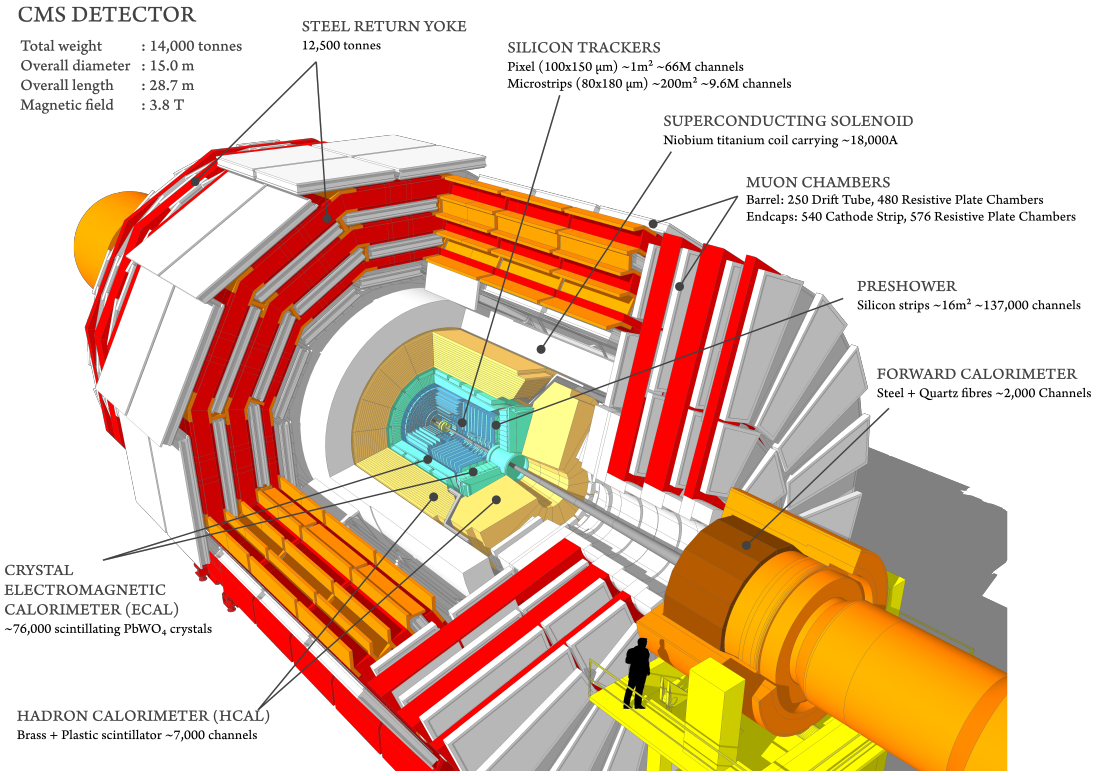


Figure 3.5: Illustration of the CMS detector [53].

The CMS detector includes several subsystems, each developed for a different purpose, and it will be explained in detail in the following sections. The CMS detector was assembled with a layered structure resembling an onion shape as shown in Figure 3.5, which facilitates distinguishing between different particle signatures. Each subdetector within this structure is designed to target a specific particle. The innermost subdetector is the silicon tracker, followed by a pre-shower detector. The latter is positioned before the brass-scintillator electromagnetic calorimeter (ECAL) and the crystal-scintillator hadronic

calorimeter (HCAL). The tracker, the preshower detector, and the calorimeters are positioned within the solenoid magnet. Gas-ionizing muon chambers are placed between iron plates within a steel return yoke to measure muon energy. The overall design of the CMS detector is rather complex, driven by the suitability for a diverse range of measurements. The high number of particles recorded per pp bunch crossing — a bunch crossing occurs every 25 ns — also contributes to this complexity. As such, achieving the necessary resolution should depend critically on the granularity of the detector subsystems. The overall detector requirements for CMS to achieve precise physical measurements across the different parameters can be summarised as follows:

- Excellent muon identification and momentum resolution up to 1 TeV with a precise charge assignment.
- Reconstruction of electrons and photons within a large geometrical acceptance, guaranteeing adequate (di)electron/photon energy resolution, which is mainly reached by the electromagnetic calorimeter's design.
- The tracking system's design ensures an excellent resolution of the charged-track momentum and reconstruction efficiency. This is important, for the jet and tau lepton reconstruction.
- Hermeticity to contain all particles resulting from collisions, acquired through the appropriate design of the hadron calorimeter. This directly affects the accuracy of the missing transverse energy reconstruction (MET).

Figure 3.5 illustrates the cylindrical symmetry of the CMS detector around the beam axis. Consequently, to represent physical quantities invariant under Lorentz boosts along the beam axis, it is advantageous to use the polar coordinate system alongside Cartesian coordinates. The nominal interaction point, positioned at the detector's center, serves as the origin of the CMS frame of reference. This frame establishes a right-handed coordinate system with the x, y, and z axes directed toward the center of the LHC ring, vertically, and along the beam axis, respectively.

Conventionally, the coordinate system is denoted as (r,$\eta$,$\phi$), where $r = \sqrt{x^2 + y^2}$ represents the radial distance from the z-axis (the beam axis). The azimuthal angle, $\phi$, is measured in the x-y plane relative to the x-axis, while $\eta$ signifies pseudorapidity. Pseudorapidity ($\eta$) is determined from the polar angle ($\theta$) as follows:

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right). \tag{3.5}$$

In the ultra-relativistic limit, the pseudorapidity aligns with rapidity, denoted as $y$ and defined by the particle's energy $E$ and its momentum projection $p_z$ along the beam

axis:

$$y = \frac{1}{2} \ln \frac{E + p_z}{E - p_z}. \tag{3.6}$$

A valuable property of rapidity $y$ is the Lorentz invariance under boosts for pseudorapidity differences in the ultra-relativistic limit. Hence, $\eta$ is a helpful variable for describing particle distances in proton-proton collisions. In this coordinate system, the angular separation between two objects $i$ and $j$ can be computed as follows:

$$\Delta R(i,j) = \sqrt{(\Delta\eta(i,j))^2 + (\Delta\phi(i,j))^2}. \tag{3.7}$$

Another crucial variable within particle physics is the transverse momentum, denoted as $p_T$, of a detected particle. It is defined as the $x-$ and $y$-components of the particle's total momentum:

$$p_T = \sqrt{p_x^2 + p_y^2}. \tag{3.8}$$

To distinguish hard collisions between partons from the underlying event, characterized by a considerably lower transverse momentum transfer, one can rely on the particle's transverse momentum. This quantity remains invariant under Lorentz boosts in the beam direction.

### 3.2.1 Tracking system

The innermost component of the CMS detector is referred to as the tracking system. Due to the large number of particles that will emerge per pp bunch crossing, the tracking system must possess specific attributes such as high granularity, excellent timing resolution and radiation hardness. These requirements are the driving force behind opting for silicon detectors for the whole tracking system.

The inner silicon tracker of CMS provides precise measurements of the trajectory of charged particles, known as tracks. The tracks are reconstructed by combining the electrical signals (hits) produced by the charged particle as it passes through the different layers of the tracker made of silicon modules. Moreover, bending the particle trajectory caused by a magnetic field allows for measuring the charged particle's transverse momentum. The electric charge of the particle can also be determined from the sign of the bent track. Accurate track reconstruction is also essential for determining the position of primary and secondary vertices. The tracking detector is furthermore helpful in measuring particle lifetimes and determining track impact parameters for intermediate states that cannot be linked to a primary vertex.

The active area of the CMS inner silicon tracker covers a total surface area of 200 $m^2$ and extends up to $|\eta| < 2.5$ in the pseudorapidity range. The CMS inner tracker comprises two subdetectors: a high-granularity pixel detector closest to the interaction point and silicon strip detectors positioned at a radial distance between 0.2 and 1.2 m from the interaction point (Figure 3.6).

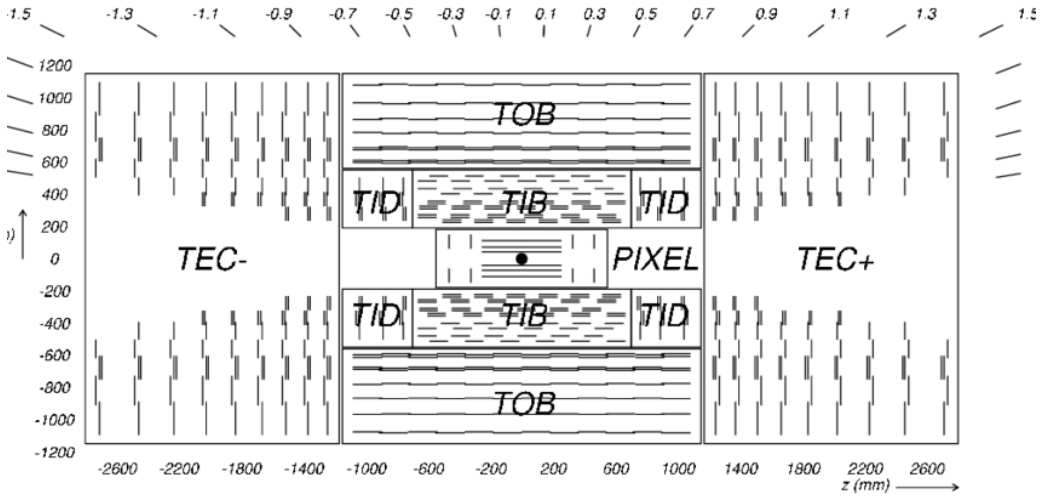


Figure 3.6: View of a transverse cross-section of the inner tracker system [31].

The inner tracking system's high-occupancy environment primarily causes the detector's radiation ageing. This can lead to an increase in leakage currents, subsequent loss of tracking data, and up to 10% loss of tracking efficiency. In order to address these issues, the CMS pixel, known as the CMS Phase-I Pixel Upgrade, was upgraded during the technical shutdown in 2016/2017. This upgrade aimed to maintain excellent tracking performance under conditions of increased peak luminosity, specifically with an average of 50 primary vertices. The Phase I pixel detector consists of the Barrel Pixel (BPIX) and the Forward Pixel (FPIX) detectors. The BPIX has four coaxial barrel layers at radial distances of 2.9, 6.8, 10.9, and 16.0 cm. The FPIX has three end cap disks per side with blades positioned at 29.1, 39.6, and 51.6 cm from the nominal interaction point (IP). Figure 3.7 compares the so-called Phase-0 and the upgrade during Phase-I.

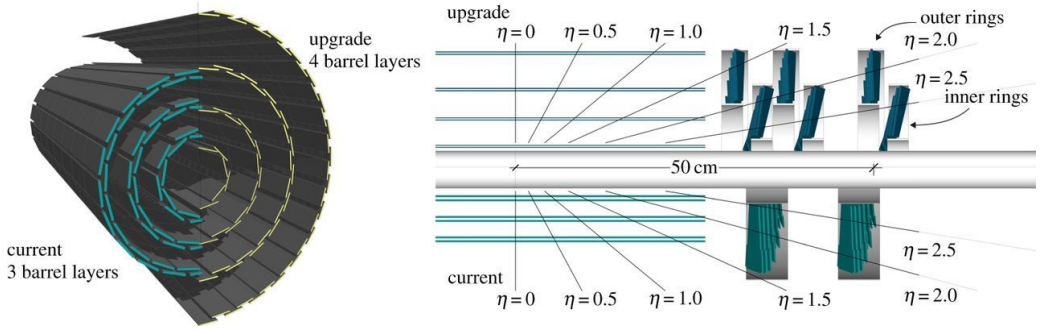


Figure 3.7: The updated pixel detector schematic. Phase-I design is compared to Phase-0 in the areas labelled "upgrade", which are the yellow area on the left and the top area on the right of the illustration [54].

The Phase I pixel detector has nearly 124 million silicon sensors, each with a pitch of 100 $\mu$m $\times$ 150 $\mu$m. By doubling the number of pixel channels compared to the previous design, it is now possible to detect four individual hits per trajectory within the coverage range of $|\eta| < 2.5$.

The silicon strip tracker is positioned at a radial distance from 20 to 116 cm outside the pixel detector. It is comprised of 15148 silicon modules, containing approximately 9.3 million strips. Like the pixel detector, the strip subsystem also has a barrel-endcap partition. The barrel region is divided into the Tracker Inner Barrel (TIB) and the Tracker Outer Barrel (TOB). The end-cap system consists of the Tracker Inner Disks (TID) and the Tracker EndCaps (TEC).

Inside the tracker barrel are four cylindrical layers of the TIB subdetector, extending up to $\pm$65 cm along the z-axis, and three TID disks covering up to a radius of 55 cm. The first two TIB layers have double-sided modules arranged back-to-back, allowing for hit measurement in cylindrical coordinates with a spatial resolution of 23-34 $\mu m$ in the $r - \phi$ plane and 530 $\mu m$ in the z-direction. The TIB/TID system is surrounded by six barrel layers of the TOB, which extend to a radius of 116 cm and up to $\pm$118 cm in the direction of the proton beam. Due to lower fluence expectations, TOB sensors are almost a factor two wider than the TIB. The Tracker Endcaps (TEC) subdetector, consisting of nine different sizes of disks, provides hit positional information in the $r - \phi$ plane within the $124 \leq |z| \leq 282$ cm range. Its resolution falls between 18 to 47 $\mu m$, ensuring coverage in this region.

The tracker system's module alignment is another crucial factor to take into account. The real position, orientation, and curvature of the modules are slightly different from the detector design. Numerous factors contribute to these shifts, ranging from straightforward misalignment during module installation and construction to the ageing effects of the detector. Reconstructing a track involves linking each hit in the tracker layers to a particle trajectory.

Figure 3.8 illustrates a comparison between two scenarios: the realistic scenario is shown on the right side of the figure, featuring distortions in the position, orientation, and curvature of the modules. On the left side of the figure, the ideal scenario showcases layers positioned according to the detector design, with tracks reconstructed by fitting a trajectory through the various hits.

A systematic misalignment might result in bias in the track reconstruction and hence in the physical measurements, whereas a random misalignment would lead to an overall decrease in the track reconstruction efficiency. CMS uses tracks gathered from pp-collisions and cosmic rays to accomplish a track-based alignment. The following $\chi^2$ is minimised in order to carry out the alignment process:

$$\chi^2(p,q) = \sum_{j}^{tracks} \sum_{i}^{hits} (\frac{m_{ij} - f_{ij}(p,q)}{\sigma_{ij}})^2, \tag{3.9}$$

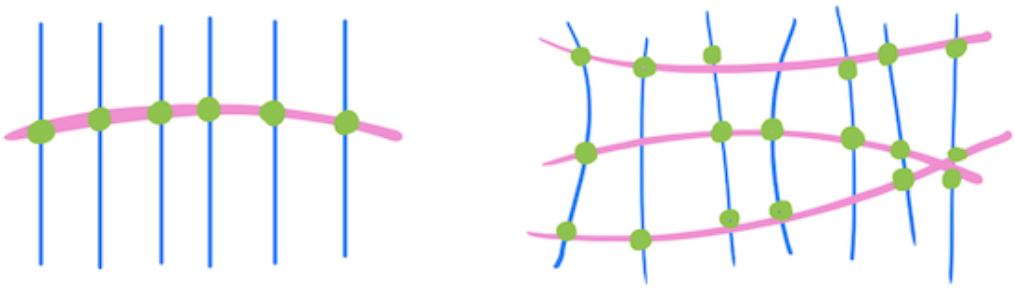

Figure 3.8: The silicon modules are represented transversally by the blue straight lines; the pink curving line represents tracks; and track hits are represented by the green circles. Idealised representation of the silicon modules on the left. Realistic case on the right, in which the modules are twisted or out of alignment.

where $m$ marks the measurements and $f$ predictions, $\sigma$ denotes the measurement uncertainties, and $p$ and $q$ stand for the alignment and track parameters, The algorithms Millepede-II [55] and HipPY [56] are used to operate the minimization. This makes it possible to obtain a hit measurement precision of the design one.

### 3.2.2 The calorimeter system

Calorimeters measure the total energy of charged and neutral particles by fully absorbing their electromagnetic or hadronic showers. When particles enter the calorimeter, they engage with an "active material" and release energy in the form of a particle shower. This shower comprises secondary particles produced when a highly energetic primary particle interacts within the dense layers of the calorimeter. The CMS detector has two categories of calorimeters: electromagnetic (ECAL) and hadronic calorimeters (HCAL), which are employed to reach an optimal measurement of the electrons, photons and hadrons within the coverage range of $|\eta| < 2.5$.

Electromagnetic calorimeters are positioned before hadron calorimeters due to the longer characteristic length of hadronic showers and the presence of an electromagnetic component. The development of ECAL focused on achieving high photon and electron energy and angular resolutions, with a primary emphasis on the $H \to \gamma\gamma$ channel. On the other hand, HCAL was explicitly designed to identify and quantify the energies of highly interacting particles and particle jets composed mainly of hadrons. Furthermore, precise energy measurement is necessary to calculate the missing energy from neutrinos.

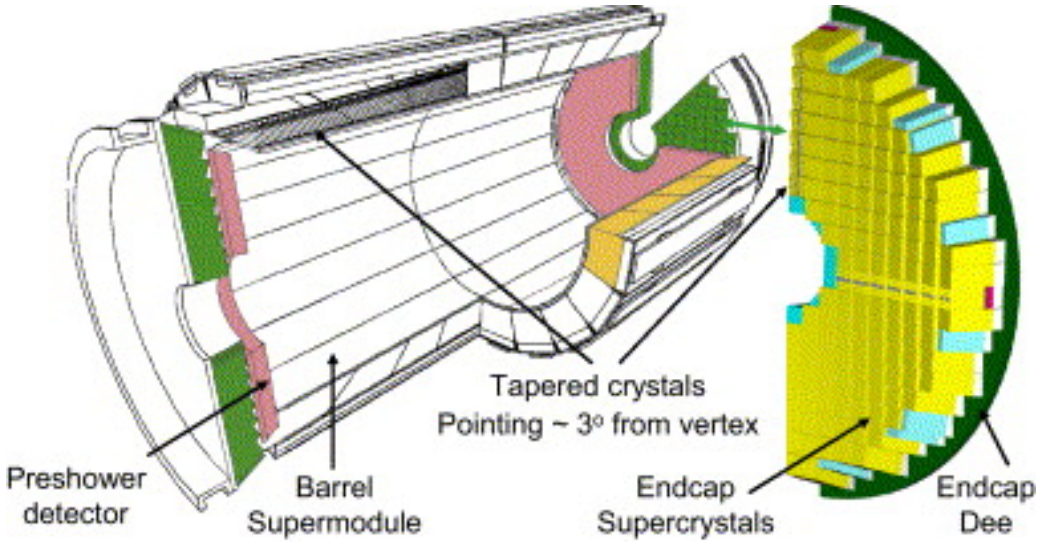


Figure 3.9: Schematic of CMS electromagnetic calorimeter [57].

### Electromagnetic calorimeter

CMS is equipped with a hermetic and homogeneous electromagnetic calorimeter, which covers the pseudorapidity range of $|\eta| < 3$, shown in Figure 3.9. The design must meet specific requirements due to general LHC conditions, such as fast scintillation time, fine granularity, and radiation resistance—choosing $PbWO_4$ crystals as the main calorimeter material fulfils all these criteria. The scintillation time is similar to the time between two consecutive bunch crossings ($\sim 25$ ns). Furthermore, the small Moliere radius [1] (2.2 cm) and short radiation length $X_0$ [2] (0.89 cm), allow for constructing a compact calorimeter with fine granularity.

The ECAL (EB) barrel section has 61,200 crystals with a volume of 8.14 $m^3$ (weighing 67.4 tons). These crystals cover the range $|\eta| < 1.479$ in pseudorapidity and are positioned at a radius of 1.29 m from the collision point. The cross-section of the crystals in the $\eta\phi$ plane is approximately $0.0174 \times 0.0174$. The length of the crystals is 25.8 $X_0$ (230 mm), allowing for over 98% energy containment for electrons and photons with energies up to 1 TeV.

The endcap section of the ECAL (EE) consists of two parts (referred to as dees) on both sides, each containing 3,662 crystals. These crystals are arranged in a rectangular $xy$ grid with off-pointing angles ranging from 2 to 8 degrees. The front and rear cross-sections

[1] The Molière radius is a defined measurement referring to the radius of a cylinder that encompasses approximately 90% of the total energy deposition of a shower. This radius plays a crucial role in determining the typical transverse size of the shower.

[2] The depth of electromagnetic showers is quantified using the $X_0$. This represents the average distance an electron travels in the detecting material before its total energy decreases by $1/e$.

of the crystals are $28.62 \times 28.62$ mm$^2$ and $30 \times 30$ mm$^2$ respectively, with a length of 24.7 $X_0$ (220 mm). The total volume of EE crystals is 2.90 m$^3$ with a weight of 24 tonnes.

A pre-shower detector is installed before each endcap disk to distinguish between $\pi^0$ decays and prompt photons. This detector comprises two layers, each consisting of a lead radiator followed by a silicon strip sensor plane. The dimensions of the two radiators are approximately one and two radiation lengths, respectively.

The silicon sensors have a pitch of 1.9 mm, with each sensor plane divided into 32 strips, each measuring $61 \times 61$ mm$^2$. Avalanche photodiodes (vacuum phototriodes) are employed in the EB (EE) regions to capture the scintillation light, chosen for their high radiation tolerance, rapid response time, and ability to function in a 4T magnetic field. A dedicated water cooling system maintains the operating temperature of the ECAL subsystems at 18$^\circ$ to ensure consistency in the counts of scintillated photons.

#### Hadron calorimeter

The hadron calorimeter is positioned around the electromagnetic calorimeter (ECAL) to enhance the energy deposition of strongly interacting particles before they reach the magnet coil. Comprising four sub-detectors, the HCAL is a sampling calorimeter that consists of alternating layers of absorbers and scintillators. Figure 3.10 depicts the longitudinal layout of the HCAL. Notably, the Hadron Barrel (HB) encompasses the central pseudo-rapidity region up to $\eta = 1.3$, and the Hadron Outer (HO) covers the same range as HB. The Hadron Endcap (HE) extends from $1.3 < |\eta| < 3.0$ on both sides, while the Hadron Forward (HF) provides coverage up to $|\eta| = 5.2$ in the forward region.

The HB and HE calorimeters are located inside the solenoid magnet, hence require the use of non-ferromagnetic materials. Brass has been selected as the absorbing material, while plastic scintillating tiles are the active material. The HO is located outside the magnet and completes the structure of the calorimeter in the barrel, capturing the tails of the hadron shower. The HF is positioned outside the muon chambers and covers the high-pseudo rapidity forward region. The absorbing plates consist of steel to withstand high particle fluxes, while quartz-based fibers collect Cherenkov radiation emitted by showers in the absorbers.

The purpose of HCAL is not only to measure the energy of hadronic jets accurately but also to determine the missing transverse momentum for each event and reduce errors in identifying muons. When all data from the detector is combined, the resolution for jet energy is 15-20% at 30 GeV, 10% at 100 GeV, and 5% at 1 TeV [59].

### 3.2.3 Solenoid magnet

A crucial part of the CMS detector is the superconducting solenoid magnet. A strong magnetic field must be generated in order to bend the trajectory of charged particles and enable the measurement of their momenta and charge. The magnet is the largest of its

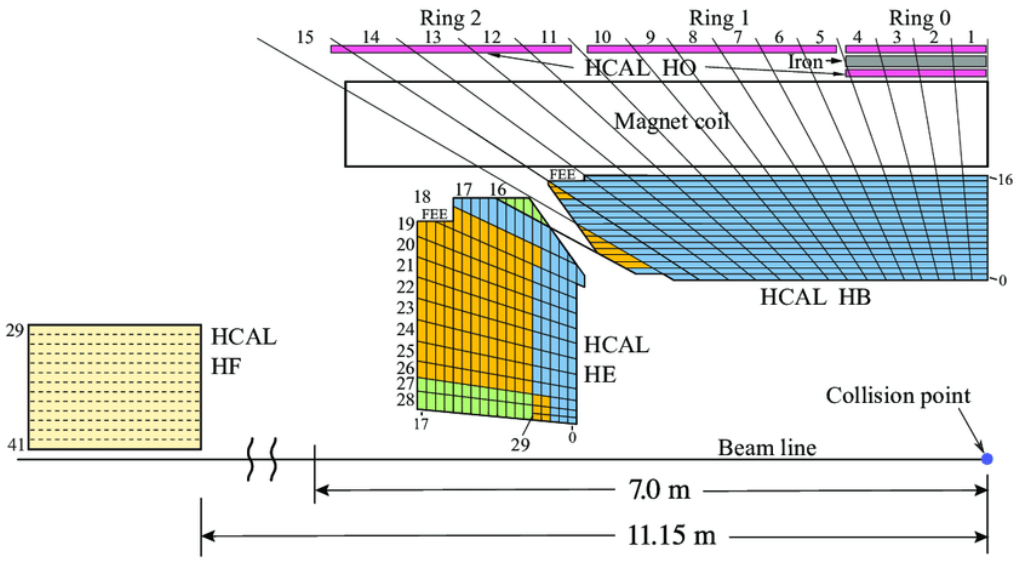


Figure 3.10: Figure shows a longitudinal perspective of a single quadrant of the CMS HCAL in $r - \phi$ plane. The HB, HE, HO, and HF elements of the HCAL detector are shown. [58]

sort ever built, falling within the range of 2.9 m $< r <$ 3.8 m, $|\eta| < 1.5$. All additional detectors, apart from the muon chambers, the Hadron Forward and the Hadron 0uter calorimeter, are kept inside the solenoid coil. As a result, tracking and energy measurements can be carried out without the particles interacting with the magnet material's bulk.

By passing a current of about 18 kA through the aluminium-coated NbTi superconducting wires that comprise the solenoid winding, the magnet creates a magnetic field of 3.8 T at the detector's centre. This consequently necessitates using a liquid helium cooling device to keep the solenoid at 4 K. A 21.6-meter-long and 14-meter-diameter steel yoke closes off the magnetic field lines. Over 10,000 tons of weight make up the yoke. With respect to the magnetic field in the centre region, the return yoke's magnetic field is directed in the opposite direction and has an intensity of 1.8 T. A more accurate reconstruction of the muon track is made possible by the return yoke's layers being spread out between the muon chambers.

### 3.2.4 Muon system

The outermost component of the CMS detector includes muon chambers that employ gaseous detection principles to detect muons. Within these chambers, charged particles ionize gas atoms within the detector volume. An external electric field directs the resulting ions and free electrons towards the cathode and anode, generating a signal. Combining

these signals from the muon chambers with data from the tracker makes the reconstruction of muon momenta feasible. Figure 3.11 shows the illustration of the CMS muon system's sub-detectors.

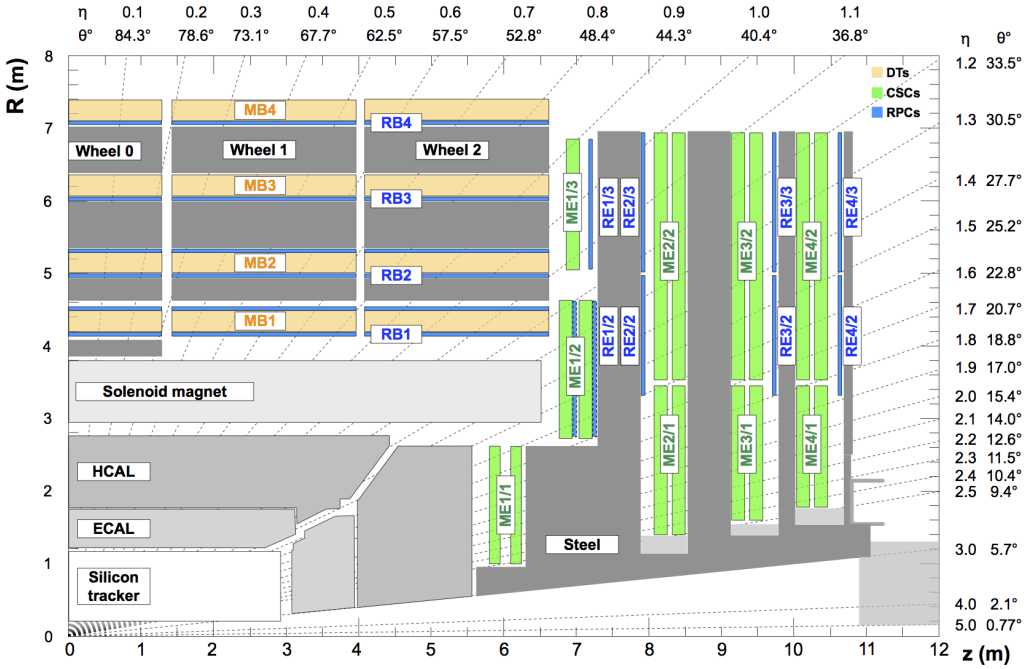


Figure 3.11: A description of the CMS detector's cross-section in the r-z plane showcases the muon system's subdetectors. The drift tube stations, identified as MB and highlighted in yellow; the cathode strip chambers denoted as ME and presented in green and the resistive plate chambers, labelled RB and RE, shown in blue; are visible within this illustration. [60]

The CMS muon system was initially designed to accurately identify and measure muons across a wide range of experimentally determined kinematics. In the barrel region, where particle rates are low, drift tube (DT) chambers are utilized. These chambers are strategically positioned between the layers of the flux return plates to cover the pseudo-rapidity range $|\eta| < 1.2$, with a total of 4 stations in this region.

The initial three stations consist of eight chambers divided into two groups of four chambers each. The first group of four chambers measures the coordinate in the $r\phi$ plane, while the second group measures the z coordinate.

Cathode strip chambers (CSC) are utilized in the endcap region to withstand a high particle rate. These CSCs cover the pseudorapidity range $0.9 < |\eta| < 2.4$ and offer high radiation resistance and fast response. Each endcap contains four CSC stations positioned between the flux return plates. The cathode strips are arranged radially within each chamber to measure the coordinate $r\phi$. In contrast, the anode wires are arranged

perpendicularly to measure both the $\eta$ coordinate and the beam crossing time of the muon.

Both the drift tubes (DT) and the cathode strip chambers (CSC) can function independently for event triggering, as they have a relatively good transverse momentum ($p_T$) resolution and are effective at rejecting background signals. However, to enhance the triggering capabilities, resistive plate chambers (RPC) are incorporated in both the barrel and endcap regions due to their ability to improve the resolution of beam crossing time. These double-gap RPC chambers, which operate in avalanche mode, cover the pseudorapidity range $|\eta| < 1.6$. The barrel region consists of 6 RPC layers, with two layers positioned in the first two stations and one layer placed in each of the last two stations. An additional RPC plate is added to each of the first three stations in the endcap region. The RPC chambers offer enhanced time resolution and rapid response with a lower coordinate resolution.

During the phase-2 upgrade of the CMS detector, the muon system is going through several changes to utilize the operation at higher luminosity [61]. The DT chambers and CSC electronics will be adjusted to maintain consistent latency and readout rates within the trigger system. Enhancements include two new segments featuring upgraded RPCs in the forward region, broadening the RPC pseudorapidity coverage from 1.9 to 2.4. Furthermore, installing three detector sets using Gas Electron Multiplier (GEM) technology aims to extend the pseudorapidity range for offline muon reconstruction up to $|\eta| < 2.8$ while improving trigger capabilities.

### 3.2.5 Trigger system

The Large Hadron Collider (LHC) produces proton-proton collisions at a rate of 40 MHz, posing challenges for real-time data analysis in the CMS subdetectors. Moreover, the high granularity of the CMS detector generates extensive amounts of data and overwhelming storage capacities. A trigger system is implemented to manage this, reducing the data size significantly. The trigger system consists of two primary stages. In the initial stage, which includes the Level-1 (L1) trigger and Data Acquisition (DAQ) system, the collision rate is reduced from 40 MHz to 100 kHz. It uses basic information from the calorimeter and muon systems to identify high-level physics objects, e.g. jets, and decide which collisions should undergo further processing. All computations happen in hardware to handle large data volumes at high speeds. The second stage, the High-Level Trigger (HLT), further reduces the collision rate from 100 to 1 kHz. Unlike the L1 trigger, this stage uses software computations on a sequence of processors to reconstruct physics objects from the selected collisions. Reconstruction paths are defined to identify specific collision types, gradually building complex objects from raw detector-level data.

Events approved by the trigger system are forwarded to a "storage manager" process for subsequent data transfer to the CMS Tier-0 computing centre located at CERN. The CMS computing system functions through three tiers:


Figure 3.12: Diagram above shows the different types of Data formats utilized by CMS and their estimated sizes for each occurrence. The formats follow a consistent sequence, where, for instance, NanoAOD derives from MiniAOD.

- Tier-0 (CERN): Handles online data transfer, initial raw data processing, and sends data to Tier-1 centres.
- Tier-1 (national computing facilities): Provides secondary data storage, transfers data to Tier-2 centres, conducts additional data processing, and supports data analysis.
- Tier-2 (research institutes): Manages local data storage, aids final-stage data analysis, and produces simulated data for Tier-1 centres.

### 3.2.6 Data processing

The events that pass the triggers are sent to the Worldwide LHC Computing Grid (WLGC), which employs standardized storage and CPU-access frameworks. The grid system allows the CMS-specific software to remotely access and analyze data efficiently. The grid is also dispersed across various computing centres worldwide and is structured into four tiers. The process starts at Tier-0 at CERN, where the raw data undergoes reconstruction to form initial objects like electrons, pertinent for subsequent analysis known as RECO data.

Following this, the Tier-1 and subsequent tiers take charge. Tier-1 centres generate the Analysis Object Data (AOD) from Tier-0 data. This AOD format contains high-level physics entities and a summary of the low-level information, which is sufficient for most CMS analyses. AOD employs ROOT file format to organize its content, utilizing the CMS Event Data Model's structure. Accessing this data necessitates the CMS software framework CMSSW [62].

After the AOD stages, two further steps are done in order to refine the data; these are the so-called Mini-AOD [63] and Nano-AOD. The Mini-AOD data sets serve as input to generate the Nano-AOD. Figure 3.12 summarizes the chain of the data formats at CMS, with Nano-AOD being the most portable and refined format, averaging around 1 kB per event. However, for the Analysis of the present work, the Mini-AOD data sets are used.

### 3.2.7 Event Simulation at CMS

A critical aspect of experimental particle physics involves the production of realistic event simulations for diverse physical processes. The event simulation is important for several reasons, such as the calibration of the detector, refinement of the trigger selection, and optimization of the physics analysis. Furthermore, simulation allows the study of possible future experiments. Simulation of the events for the proton-proton collisions is performed with the Monte Carlo event generator.
To simulate the event, not only the hard scattering process, which involves the interaction between two partons at high transverse momentum but also the underlying events should be taken into consideration. The underlying event usually refers to the interaction between other partons at lower transverse momentum and the emission of additional particles by the hard scattering partons, occurring either in the initial or final state. The visualization of a simulated event is shown in Figure 3.13, which contains the hard scattering process and the underlying event simulation.

To reflect the full complexity of the events produced at the proton-proton collision, the Monte Carlo event generators should simulate both hard scattering and underlying events. Starting with the factorization theorem [64], which enables the calculation of the cross-section for hadronic processes by incorporating the parton distribution functions (PDFs) [25] of the two incoming partons. According to the factorization theorem, any hadronic cross section can be expressed as a convolution of the PDFs of the two incoming partons, evaluated at the factorization scale. This convolution is then multiplied by a parton-parton cross-section, calculable in perturbation theory as an expansion in the strong coupling constant $\alpha_s$. Typically, the factorization scale is set to the scale of the hard scattering, coinciding with the renormalization scale where the running constant $\alpha_s$ is evaluated.

Due to the complexity of the calculations, which largely disregard analytical solutions, numerical integration methods are commonly used to obtain results. Therefore, the Monte Carlo methods are used for the event simulation, especially since, in quantum mechanics, only the probability for a specific final state can be computed. Various programs dedicated to event generation are available. Normally the simulated events are generated using the MADGRAPH5_aMC@NLO [65], POWHEG [66] and PYTHIA [67] generators.

The MADGRAPH5_aMC@NLO can calculate the hard scattering process at either Leading Order (LO) or Next to Leading Order (NLO), allowing for the inclusion of up to four additional real partons in the matrix element calculation. However, when utilized for the more computationally demanding NLO computation, the number of additional partons is restricted to two. A $p_T$ cut is essential for the simulated partons in the final state to mitigate potential divergences from soft gluon radiation.

On the other hand, POWHEG specializes in computing $2 \rightarrow 2$ processes at NLO, allowing for the inclusion of a maximum of one additional parton in the matrix element calculation. It is optimized for simulations involving heavy quarks. Like MAD-

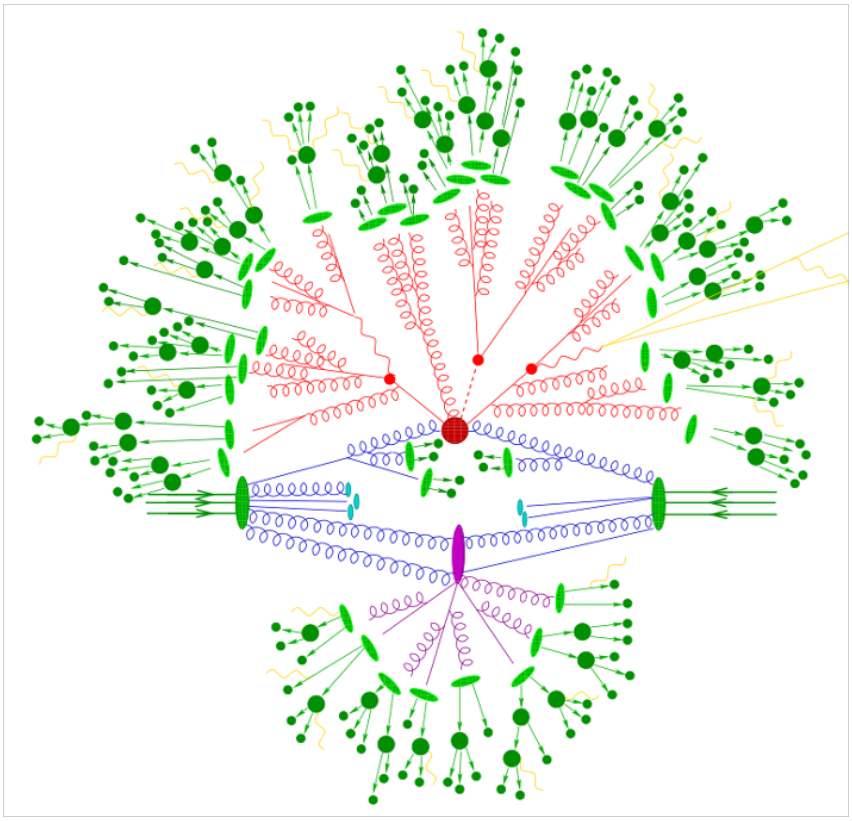

Figure 3.13: Illustration of a simulated event produced from a proton-proton collision. The red blob in the middle represents the hard scattering process, and the underlying events are shown in purple. Transitions from partons to hadrons are denoted by light green blobs, dark green blobs signify hadron decays, and soft photon radiation is indicated by yellow lines. [64]

GRAPH5_aMC@NLO, POWHEG does not simulate the underlying event.

Furthermore, PYTHIA is designed to calculate tree-level $2 \to 2$ matrix elements, with higher-order corrections approximated using the parton shower algorithm. While not commonly used as a hard scattering event generator, Pythia simulates parton showers and hadronization.

The hard scattering and Parton shower steps are linked, necessitating proper interfacing through dedicated jet matching and merging procedures to prevent double counting or unexplored regions of the phase space. It is crucial to note that the jet matching and merging processes are implemented differently for LO and NLO matrix element calculations as discussed further in [64].

In the context of these interconnected processes, the potential for double-counting arises when combining partons produced at the matrix element level with those generated

due to the parton shower. To address this challenge, Matrix Element + Parton Shower (ME + PS) matching and merging techniques have been produced [68]. This matching procedure involves the separate generation of Parton-level events for each jet multiplicity, followed by showers. Subsequently, the showered partons are clustered into jets, and each jet is matched to its corresponding particle-level parton. If all showered partons find a match with the ME partons, the events are retained rather than discarded.

Furthermore, it is essential to consider pileup interactions to secure a realistic simulation of the pp collision.

Following the realisation of event simulation, the GEANT4 [69] is employed to simulate the response of the CMS detector. The latter simulates the propagation of particles through the detector, accounting for their interaction with the materials. Using the detector response obtained from this simulation, signals from various particles are simulated, and the identical reconstruction algorithm employed in real data is applied. This ensures a consistent and meaningful comparison between the data and the simulation.

CHAPTER 4

# Event reconstruction at CMS

This chapter provides an overview of the reconstruction of high-level physics objects, such as muons, electrons and jets, using the standard CMS algorithm, namely the Particle Flow [70] algorithm. In this physics analysis, the final state particles are electrons, muons, taus, and b-tagged jets, which will be described in detail, following a brief discussion of the reconstruction techniques employed for tracking and vertexing.

## 4.1 Particle Flow algorithm

The ALEPH collaboration at LEP initially designed the Particle-Flow (PF) algorithm [70], while the CMS collaboration later adopted it to reconstruct the final state particles with high precision. Particle-flow event reconstruction aims to identify and reconstruct all stable particles within an event, i.e. electrons, muons, photons, charged and neutral hadrons, by integrating information from all CMS sub-detectors. This global approach aims for an optimal determination of their direction, energy, and particle type. Figure 4.1 illustrates the typical signature of different particles in the detector. Muons are reconstructed by connecting a track in the silicon tracker with hits in the muon chambers. In contrast, electrons and charged hadrons are identified based on a track matched to an energy deposit in the ECAL and HCAL, respectively. For photons and neutral hadrons, the identification involves searching for clusters in the ECAL and HCAL, respectively, which are not associated with any track. Combining the measurements from different sub-detector components by the PF algorithm ensures a higher resolution compared to employing only a single component of the detector.

The features of the CMS detector ensuring outstanding PF algorithm performance are the high granularity in the electromagnetic calorimeter, the hermicity of the hadron calorimeter, and the large magnetic field. This algorithm relies on precise track reconstruction, an effective clustering technique to distinguish overlapping showers, and a reliable

linking process to merge information from various sub-detectors measuring a single particle energy deposit. The PF algorithm is simplified for the online reconstruction. It is employed at the High-Level Trigger (HLT) level to maintain consistency between offline and online object reconstructions, minimizing trigger inefficiencies. The Particle-Flow event reconstruction unfolds in three steps: firstly, fundamental elements such as tracks and calorimetric clusters are reconstructed; secondly, these fundamental elements are connected into blocks potentially originating from the same particle; and finally, particles are reconstructed and identified based on these blocks.

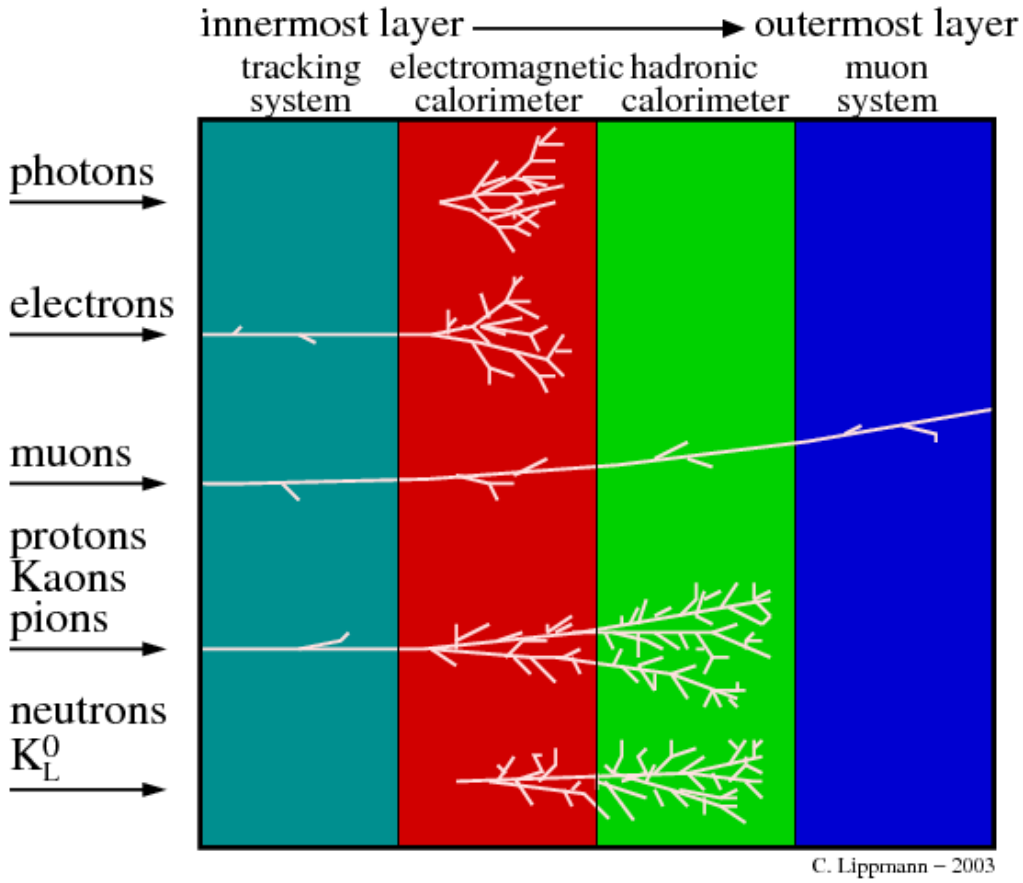


Figure 4.1: Figure shows the signature of each particle in the different components of the detector [71].

The elementary components in the Particle-Flow event reconstruction contain tracks and calorimetric clusters reconstructed using advanced algorithms to ensure high efficiency and purity. The following section will briefly discuss the reconstruction of tracks and calorimeter clusters.

### 4.1.1 Track reconstruction

In the initial stage of track reconstruction, known as local reconstruction, only digitized signals in the readout channels of the Inner Tracker above fixed thresholds are transmitted, a mechanism known as zero-suppression. The hits are then combined together to form clusters, and their barycenter is computed to determine the cluster position and their corresponding uncertainty, using a local coordinate system $(u, \nu)$ defined relative to the plane of each sensor.

The hits obtained from the local reconstruction are then fitted to reconstruct the trajectories or *tracks* needed to estimate the momentum and position of the associated charged particles. This process involves translating between the local coordinate system of the hits and the global track coordinate system, accounting for discrepancies between the assumed and the actual location of each detector partition, as well as the potential surface deformation of the individual modules revealed during the alignment process [72]. Furthermore, the uncertainty in the position of the detector element contributes to the intrinsic uncertainty in the local hit position.

To reconstruct the path of the charged particles, CMS employs the Combinatorial Track Finder (CTF) [73], an adaptation of the combinatorial Kalman filter that extends its capabilities to combine pattern recognition and track fitting in a single framework. The CTF iteratively produces reconstructed tracks through multiple steps, a process known as iterative tracking. Initially, the algorithm focuses on locating easily identifiable tracks (e.g., tracks with higher transverse momentum produced near the interaction region), gradually refining the search to encompass more complex tracks (e.g., low-$p_T$ or significantly displaced tracks). The main iterative tasks carried out by the tracking algorithm can be summarised in four following steps:

1. **Seed generation** delivers the initial track candidates from a minimal number of hits (typically a doublet or triplet), establishing the starting point for estimating the track parameters and their associated uncertainties.

2. **Track finding** employs a Kalman filter to extend the seed trajectories along the anticipated path of a charged particle, aiming to identify additional hits that can be assigned to the track candidate as well as their spatial correlations.

3. **Track-fitting and refinement** is performed using the clustered hits using a Kalman-filter technique to provide the most accurate estimate of trajectory parameters for each track.

4. **Track quality selection** involves setting quality flags and rejecting tracks not meeting specified criteria, such as a $\chi^2$ goodness-of-fit test, the compatibility of the observed hit position with the reconstructed track trajectory and the hit multiplicity associated with the track.

### 4.1.2 Calorimeter clusters

The goal of clustering algorithms in the calorimeters is to identify and measure the energy of neutral hadrons and photons, distinguishing them from the energy deposits of charged hadrons. Additionally, these algorithms aim to identify and reconstruct electrons and their associated bremsstrahlung photons. The calorimetric energy measurement is linked to the momentum measurement of the associated tracks to enhance the overall energy resolution. In PF reconstruction, a dedicated clustering algorithm is employed separately in the ECAL(HCAL) barrel and endcaps, respectively, and in the two pre-shower layers. In the initial step, cluster seeds are formed as cells with the maximal energy deposit, i.e. 8 (4) neighbouring cells in the case of the ECAL (HCAL) detector, respectively. Subsequently, topological clusters are formed starting from these seeds by adding cells with at least one common corner and energy above a given threshold. Finally, individual clusters are identified within each topological cluster using a specific algorithm [74]. An accurate calibration of the calorimeter's response to hadrons and photons is employed to minimize false reconstruction in energy excesses and effectively identify neutral particles.

### 4.1.3 Muon track reconstruction

The standard method for reconstructing muon tracks consists of two stages: first, fitting tracks independently for muon hits within the inner tracker to produce tracker tracks, and second, fitting hits in the muon chamber to generate standalone muon tracks [75]. Two different algorithms, each with distinct approaches, are used to identify and reconstruct Global and Tracker muons.

Global muons are reconstructed via an "outside-in" approach. This involves matching a standalone muon track with a tracker track. Before this assignment, both tracks were extended to a shared surface via a Kalman-filter technique [76].

Tracker muon tracks are constructed with a complementary "inside-out" strategy. These tracks are formed with less strict criteria, starting from the tracker track candidates. All candidate tracks with transverse momentum $p_T > 0.5$ GeV and total momentum $p > 2.5$ GeV are propagated to the muon system considering the tracking material's multiple Coulomb scattering, magnetic field intensity and the expected energy losses. The extrapolated track must accurately coincide with a muon segment in either the CSC or DT system of the muon spectrometer. The inner tracker and muon system exhibit outstanding performance, allowing for the reconstruction of roughly 99% of the detected muons within the defined tracker volume as either tracker or global muons.

### 4.1.4 Particle Flow link algorithm

The Particle Flow Link Algorithm aims to connect elements from local reconstruction in various subdetectors to generate Particle Flow (PF) blocks. The latter is crucial for

achieving a global event description. Normally, a given particle is associated with multiple particle-flow elements, which are connected through a link algorithm, representing the detector's global reconstruction of a single particle. The link algorithm begins by selecting a pair of elements within the event and determining the distance between them. Subsequently, the linked algorithm assesses the quality of these paired elements and organizes them into blocks. This linking process is carried out iteratively for each pair of elements in the event, with measures in place to avoid double counting. Due to the fine granularity of the CMS detectors, the formed blocks typically contain only a small number of elements. These blocks then serve as input for the particle reconstruction and identification algorithm.

### 4.1.5 Particle reconstruction and identification

The essential factor of the particle-flow algorithm involves the reconstruction and identification of a set of particles from each block of elements. This process produces a list of reconstructed particles that are required and used for physics analysis. The blocks produced by the link algorithm are used for particle reconstruction and identification. First, each global muon with momentum consistent with the tracker-only measurement is identified as a particle-flow muon, and the corresponding track is excluded from the block. Subsequently, electrons are identified, considering their short tracks and energy loss through bremsstrahlung in the tracker layers. Tracks are traced to the ECAL and matched with the related clusters in the calorimeter. When an electron is identified, the associated track and the corresponding ECAL clusters are withdrawn from further processing.

Furthermore, the remaining tracks in the block are linked to a particle-flow charged hadron, with momentum and energy derived from the tracker and the pion mass hypothesis taken into account. To find the neutral particles, such as photons and neutral hadrons, the momentum of the tracks is compared to the corrected energy noticed on the calorimeter. Suppose the energies of the nearest ECAL and HCAL clusters related to the tracks exceed expectations based on the total momentum of associated charged particles. In that case, these are considered photons or neutral hadrons. Furthermore, the remaining ECAL and HCAL clusters, which are not connected to any track, are also considered photons and neutral hadrons.

## 4.2 Physics objects reconstruction

### 4.2.1 Primary Vertex

Out of the set of collisions occurring simultaneously at the LHC during proton-proton data-taking, only a few results in hard scattering processes referred to as pile-up inter-

action vertices. The positions of the primary products of the interaction, referred to as primary vertices (PV), are reconstructed with high precision by looking into the high-energy (high $p_T$) products from the pile-up collisions. Additionally, precise identification of final state particles requires excellent primary and secondary vertex position measurements.

The PV reconstruction involves a two-step process after track reconstruction using the CTF algorithm. Initially, tracks associated with a single primary vertex candidate are grouped using the Deterministic Annealing (DA) algorithm [77]. This algorithm assigns each track to a primary vertex candidate based on the track's transverse position with respect to the nominal interaction point. Subsequently, it probabilistically determines whether each primary vertex candidate should be subdivided into two vertices. The existence of outlier tracks, often linked to poorer fit quality, can generate false primary vertices. To counter this effect, they are proactively given less weight when calculating the overall count of primary vertices.

In the second step, the selected grouped vertex candidates, with a minimum of at least two associated tracks, undergo fitting using the Adaptive Vertex Filter [78]. During this stage, both the vertex position and covariance matrix are computed. The adaptive vertex reconstruction assigns a specific probability and a corresponding weight $w_i$ to measure the extent of compatibility between the track and the fitted vertex.

The vertices must satisfy specific criteria: their z position should be located within 24 cm of the beam spot, their radial position within 2 cm from the interaction point, and the vertex fit must have a minimum of four degrees of freedom. The primary vertex of the interaction is determined as the vertex with the highest sum of $p_T^2$ from its associated tracks. Other reconstructed vertices can be designated secondary vertices or interpreted as contributions from pileup events.

### 4.2.2 Muons

The CMS detector was designed to ensure a highly precise measurement of muon momentum and charge for a large range of kinematic parameters. Due to their long lifetime and minimal energy deposits in the electromagnetic calorimeter, muons can travel the entire CMS apparatus, reaching the muon system positioned beyond the superconducting solenoid. The muons displayed in Figure 4.1 are identified using tracks within the silicon tracker system and registering hits within the muon chambers.

#### Muon identification

After the muon track reconstruction, potential muon candidates are identified using the PF algorithm, which integrates muon track information and calorimeter energy deposits. This reconstruction effectively minimizes most instances of hadronic punch-through, where

highly energetic charged hadrons, such as pions, generate electromagnetic showers that penetrate the muon system.

Muon identification criteria are applied within the muon reconstruction sequence using variables such as track hit count, track fit quality, and track impact parameters. The compatibility of muon segments between tracker tracks and standalone muon tracks, quantified as a probability between 0 and 1, is also employed. Additional criteria for muon identification relate to global variables associated with the muon track and other particles in the reconstructed event. One of these factors is the relative isolation, which measures the number of particles reconstructed in the vicinity of the muon. It is needed to identify muons stemming from the in-flight decay of baryons generated inside a jet. The definition of muon relative isolation is [79]:

$$I^{\mu}_{rel} = \frac{\sum E_T(charged) + max(\sum E_T(neutral) - 0.5 \sum E_T(charged, PU), 0)}{p^{\mu}_T}, \qquad (4.1)$$

where $E_T$ is the transverse energy of the particles contained in a $\Delta R$ cone with radius 0.4 around the muon direction of flight, and $p^{\mu}_T$ is the muon transverse momentum. The types of particles are, respectively, the charged hadrons (charged), neutral hadrons and photons (neutral), and charged hadrons from pile-up interactions (charged, PU).

Charged particles moving through the tracker system have the potential to emit photons via bremsstrahlung radiation. Therefore, the second term is needed to adjust the energy of the neutral particles in the vicinity of the muon by subtracting the energy of charged particles stemming from the PU vertices. This further improves the accuracy of the muon isolation for prompt muons. For a correct estimation, the adjusted energy from neutral particles must be positive-definite. In this analysis, the selected muons must pass the $I_{rel} < 0.15$ relative isolation requirement.

Establishing muon identification working points (WP) allows refining the trade-off between selection efficiency and the purity of reconstructed muons, tailored to the specific requirements of individual physics analyses. Reconstructed muon candidates are subject to additional identification (ID) criteria. The CMS Collaboration employs multiple identification methods, each with its own efficiency and misidentification rates, listed below:

- *Loose muon ID* aims to identify prompt muons from the primary vertex and muons from light and heavy flavour decays while maintaining a low misidentification rate of charged hadrons as muons. It encompasses muons selected by the PF algorithm and identified as either Tracker or Global muons.

- *Medium muon ID*, optimized for prompt muons and those stemming from heavy-flavour decays, builds upon loose muons, further requiring a tracker track with hits from over 80% of the pierced inner tracker layers.

- *Tight muon ID* suppresses the number of muons produced from in-flight decays and hadronic punch-through. The reconstructed PF muon must be classified as a Global muon with a normalised track-fit $\chi^2 < 10$. Additionally, it must have at least one muon-chamber hit included in the global muon track fit, muon segments in at least two muon stations, track hits in at least five layers of the tracker (including at least one pixel hit). The track transverse $d_{xy}$ (longitudinal $d_z$) impact parameter with respect to the primary vertex must be less than 0.2 (0.5) cm.

### 4.2.3 Electrons

Electron reconstruction in CMS relies on linking their associated tracks to the energy clusters in the ECAL. Electrons and photons deposit most of their energy in the ECAL, while hadrons mainly do so in the HCAL. In addition, electrons register hits in the tracker layers as illustrated in Figure 4.1. Electron reconstruction is intrinsically linked to that of photons and vice versa due to the bremsstrahlung radiation, which causes electrons to lose energy as they traverse the tracking system. The accurate determination of the electron incident energy implies collecting the energy radiated by the emitted photons. The electron mostly propagates along the $\phi$ axis due to the bending of its path in the presence of a magnetic field. Highly energetic photons are affected by pair production, leading to two electromagnetic showers with a similar spread in the $\phi$ direction that may initiate before crossing the ECAL system. A clustering approach computes the initial energy of the primary electron or photon, in which groups of crystals with energy levels above a specific threshold are identified.

Thresholds are set through calibration based on the subdetectors noise[1]. The cluster with the highest energy in a certain $\eta \times \phi$ region is used as the seed. Its neighbouring clusters are combined to create a super-cluster (SC), which is reconstructed with two algorithms: the *mustache algorithm* using ECAL and preshower data to categorise clusters dispersed around the seed, and the *refined algorithm* merging the ECAL clusters with the tracks reconstructed in the inner tracker.

The compatibility between these tracks and the supercluster (SC) is assessed using a Gaussian Sum Filter (GSF) algorithm [81]. GSF-tracks are combined with the associated supercluster to determine the electron properties using a linking algorithm. The latter employs both cut- and multivariate-based techniques and is optimised for detecting low energy electrons [80]. Subsequently, the PF algorithm builds electron and photon candidates based on whether the linked GSF-track has a hit in the innermost pixel layer.

Common electron identification criteria, including cut-based and Boosted Decision Tree (BDT) techniques, are used by many physics analyses at CMS. Several features of the reconstructed electron candidates are used by the BDT-based algorithm. The

[1]The typical values for Electron Barrel (Electron Endcap) are approximately 80 MeV (approximately 300 MeV) [80]

parameters for the ECAL shower are similar in the identification processes for photons and electrons, for further details see [80]. One key variable is the relative electron isolation defined as:

$$I^e_{rel} = \frac{\sum E_T(charged) + max(\sum E_T(neutral) - \rho \times A_{eff,0})}{p^e_T}, \tag{4.2}$$

where $p^e_T$ is the transverse momentum of the electron, and $E_T$ is the transverse energy of the charged hadrons or neutral hadrons and photons. Additionally, $\rho$ is the average transverse density in the event, and $A_{eff}$ is the isolation region. The $\rho \times A_{eff}$ correction factor is adjusted to consider the dependence of the transverse energy on $\eta$ from pile-up events. The sum index iterates over the particles inside a cone with $\Delta R = 0.3$ around the flight path of the electron.

### 4.2.4 Jets

The production of a jet, or collimated shower of particles, results from the hadronization of quarks and gluons in proton-proton collisions. A jet's distinctive shape is a tight cone parallel to its mother parton's initial flight path; Figure 4.2 illustrates a simplified jet sketch. The process of jet reconstruction is carried out by a jet reconstruction algorithm that groups the spray of particles surrounding the mother parton candidate to assign a jet observable to it.

The main characteristics of jet clustering techniques are infrared and collinear safety (IRC), which must be fulfilled concurrently to prevent divergences in perturbative QCD calculations of the jet shape. To ensure collinear safety, the jet observable cannot be changed by splitting a single particle over a group of particles moving simultaneously, each carrying a portion of the original momentum. According to infrared safety regulations, the latter must likewise be maintained even in the case of soft particles. Additionally, the $p_T$ spectrum's measured offset before the pile-up correction scales with the radius parameter $R$ used for the clustering.

The reconstruction algorithm for jets used in the CMS experiment is the so-called anti-$K_T$ algorithm [82], which satisfies the IRC safety and is iterative. The two parameters used by the anti-$k_T$ algorithm to cluster jets are $d_{ij}$ defined as the distance between particle $i$ and particle $j$, and $d_{iB}$, which is the distance between the proton beam axis and the particle $i$. These two parameters are defined as follows:

$$d_{ij} = \min\left(k^{2p}_{T,i}, k^{2p}_{T,j}\right)\frac{\Delta^2_{ij}}{R^2}, \qquad \text{where} \qquad \Delta^2_{ij} = (y_i - y_j)^2 + (\phi_i - \phi_j)^2, \tag{4.3}$$

$$d_{iB} = k^{2p}_{T,i}, \tag{4.4}$$

In this context, the cluster variable $K_{T,i}$ denotes the transverse momentum of the particle $i$, while $\phi_i$ and $y_i$ denote the azimuthal angle and the rapidity of the particle $i$,

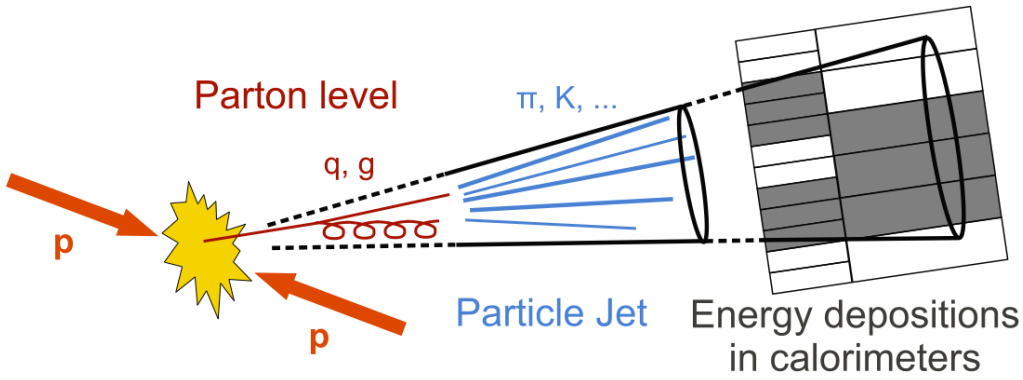


Figure 4.2: Figure illustrates a simplified sketch of a jet produced in pp collision.

which can be classified as either a pseudo-jet or a particle. The parameter $R$ represents the radius parameter, whereas the parameter $p$ is specified as $-1$ in the anti-$k_T$ algorithm [82]. The parameter $p$, in general, represents the different algorithms.

Initially, the algorithm calculates all possible values of $d_{ij}$ and $d_{iB}$, selecting the smallest one. If $d_{iB}$ is the smallest value, the object $i$ is identified as a jet and removed from the object list for clustering. Contrarily, if $d_{iB}$ is not the smallest value, objects $i$ and $j$ are combined into a single cluster. This process is repeated iteratively until no particles remain to be identified as jet constituents. Figure 4.3 displays the cone-like shape of jets clustered by the anti– $k_T$ algorithm.

Due to the non-uniform and non-linear response of the calorimeter, the reconstructed jet energy does not match the original parton energy. Hence, corrections are applied at the reconstruction level for data and the Monte Carlo simulation. The two types of corrections applied are the so-called Jet Energy Scale corrections (JES) and the Jet Energy Resolution corrections (JER), which will be discussed in detail in the following sections.

### Pileup mitigation

Several techniques are developed at CMS to mitigate the effect of PU events. One of these methods is the so-called charged-hadron subtraction (CHS) [70], which mitigates the impact of PU on jet reconstruction. The CHS algorithm uses the tracking information to distinguish particles stemming from the PU after particle flow reconstruction and before any jet clustering. Within this procedure, the charged particle candidates related to a PU vertex are excluded. In the case of charged particles not affiliated with any PU vertex, along with all neutral particles, they are retained.

On the other hand, the other pileup mitigation algorithm, the PUPPI [83], aims to address the impact of PU on observables of clustered hadrons, such as jets, missing transverse momentum ($p_T^{miss}$), and lepton isolation. The PUPPI algorithm is applied at

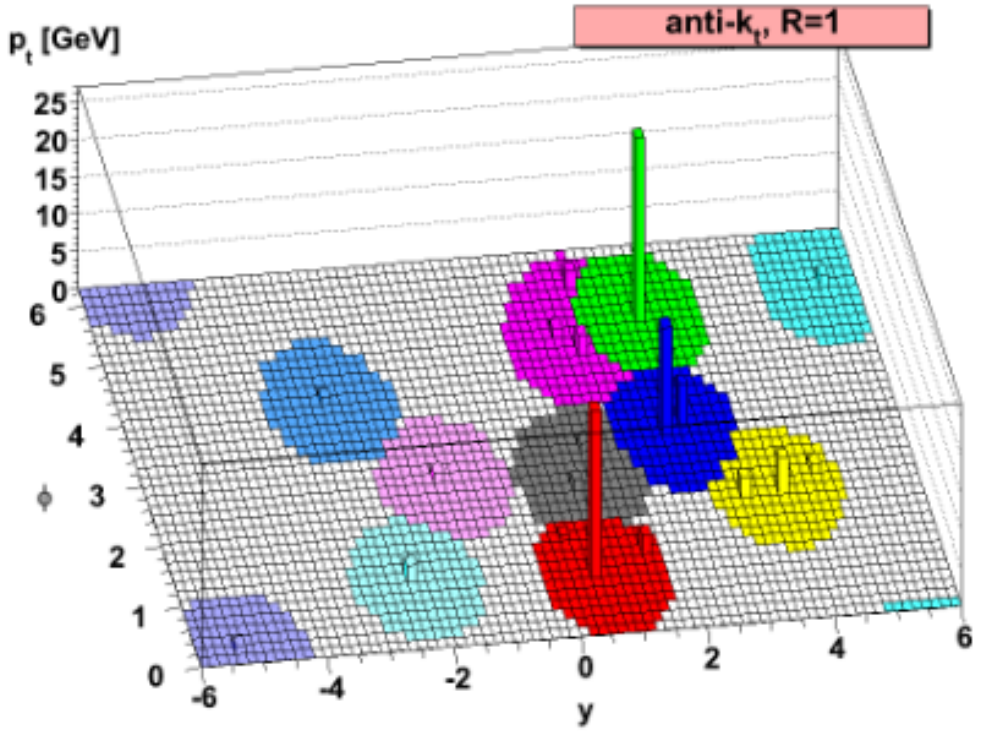


Figure 4.3: Illustration of ant-$k_T$ jets [82]

the particle level before any clustering technique is applied. The basic idea behind the PUPPI algorithm is to assign a weight in the range of 0 to 1 to each particle based on the information of the neighbouring particles. A value of 1 is assigned to particles considered to originate from the primary vertex. The assigned per particle weight by PUPPI is employed to rescale the particle four-momentum to correct the PU effect at the particle level hence diminishing the contribution of PU of the relevant observabels. The sketch 4.4 illustrates both CHS and PUPPI algorithms.

### Jet Energy Scale

It is necessary to match the measured energy of each reconstructed jet to the true energy of the corresponding parton at the generator level. To this end, Jet Energy Scale (JES) corrections are computed as a multiplicative factor on the four-momentum of the observed jet and applied both in data and simulation. The ratio of the measured reconstructed jet $p_T$ to the generated jet $p_T^{ptcl}$ at the particle level is defined as the observed jet transverse momentum response. The CMS JetMet group [85] provides centrally this set of correction factors and their associated uncertainty. The jets are calibrated using a factorized approach, and the individual corrections are applied one after the other in the sequence

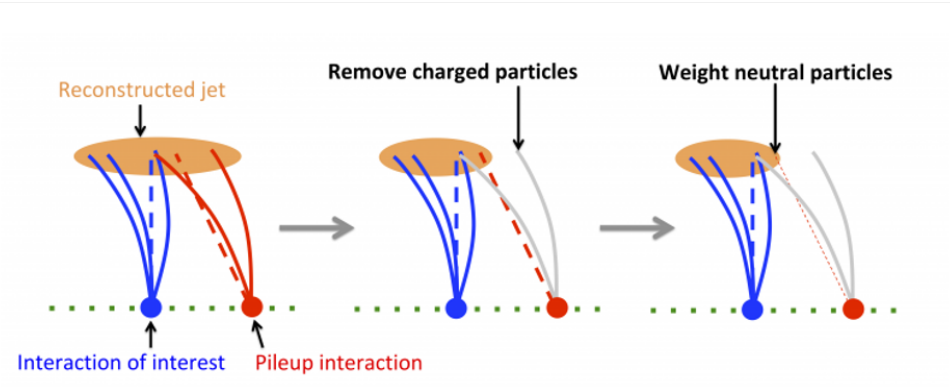


Figure 4.4: Sketch of the Pile Up mitigation techniques (CHS & PUPPI) used at CMS [84].

shown in Figure 4.5.

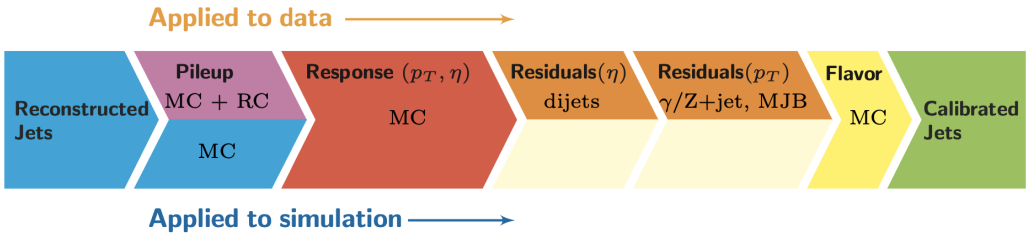


Figure 4.5: Overview of the stepwise corrections applied for the jet energy calibration in data and simulation [85]

The first step consists of applying the *pileup offset* corrections, which take into consideration the additional energy of the jet that is measured as a result of in-time (IT) and out-of-time (OOT) pileup (discussed in section 3.1). They are applied to data and MC samples obtained from the simulation of QCD multi-jet events processed with or without pile-up injection. Secondly, the *simulated response* corrections, parameterized as a function of these jet kinematic variables, compensate for the non-uniform angular response in the tracker transition zone and the non-linear response in $p_T$ from the calorimeters.

The next step is to apply the *residual data to Monte Carlo* corrections to enhance the data to simulation agreement as a function of the jet kinematics. The absolute scale of the recorded jet $p_T$ response is adjusted in the data using multi-jet and Drell-Yan plus jets events, which benefit from the high precision measurement of the Z-boson and the photon energy in the ECAL. In di-jet events, the residual $\eta$ correction is obtained by exploiting the $p_T$ imbalance present in the di-jet system. Finally, *flavour* corrections are applied due to the different $p_T$ responses for light and heavy flavour jets. They are retrieved from data containing Z+b-quark events for jets initiated by b-quarks.

### Jet Energy Resolution

After adjusting the jet energy scale, it is also necessary to apply Jet Energy Resolution (JER) corrections. A certain dose of smearing in the simulated jet transverse momentum is introduced to align it with the actual detector resolution. Differently from the scale corrections, which lead to an overall shift in the jet $p_T$ spectrum, those on the jet resolution impact exclusively the width on the residual jet $p_T$ response $\Delta$, defined as the ratio:

$$\Delta = \frac{p_T - p_T^{\text{ptcl}}}{p_T}. \tag{4.5}$$

To establish JER corrections, a technique similar to the aforementioned one adopted to assess the residual corrections in Drell-Yan+jets events is employed. These corrections are determined in simulation by parameterizing the width of the residual jet $p_T$ response $\Delta$ with a double-sided Crystal-Ball function. The latter accounts for the distinct behaviour of the Gaussian bulk and the non-Gaussian tail of the distribution, which highlights potential detector resolution effects, such as non-uniformities in the response in the tracker transition region. This is derived for various pileup scenarios as a function of the jet transverse momentum. To achieve this, a series of data-driven methods have been devised to extract the data-to-simulation scale factors, which are, in turn, used to smear the simulated jet resolution.

### B-tagged jets

The b jets are produced in the final state by the incident bottom-quark fragmentation. They are present in the signature of a wide range of physical processes, including top-quark and the Higgs boson decays. Accurately identifying b jets is essential to minimize the background from processes involving jets from gluons (g), light-flavour quarks (u,d,s), and from the fragmentation of c-quarks. A dedicated tagger algorithm [86] must be used to identify the b-jets, which relies on the characteristics of the b hadrons, including jet composition (the decay products are more energetic than those of light-quark hadrons), kinematic properties (long lifetime and relatively large mass $\sim 5$ GeV), and the presence of secondary vertices of the interaction. Figure 4.6 illustrates diagrammatically the topology of the b-jet production and decay.

One of the defining features of b-jets is the presence of a secondary vertex (SV). Therefore, it is crucial to reconstruct accurately SVs to differentiate b-jets from other jets. The algorithm employed for reconstructing the secondary vertex (SV) is the Inclusive Vertex Finder (IVF). This algorithm can identify at least one vertex within a jet without any jet-related information. It first conducts a pre-clustering of the tracks in the event by using displaced tracks as starting points. Simultaneously, it clusters the remaining tracks based on their pairwise separation. The IVF algorithm takes all reconstructed tracks from the event as its input. The secondary vertex candidates must meet certain requirements

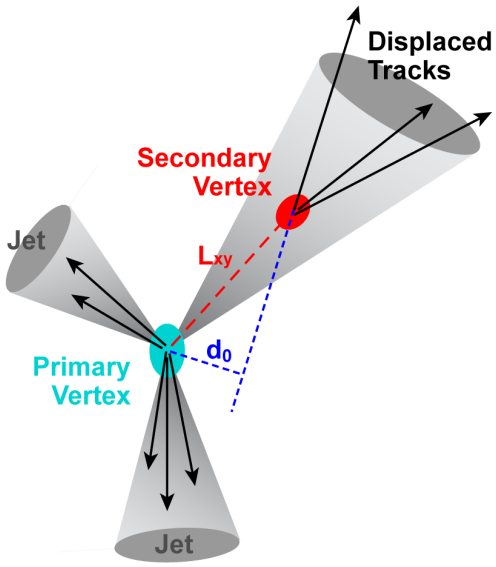


Figure 4.6: Simplified view of a heavy flavour jet. The main characteristics of these jets are the large impact parameter ($d_0$) and the secondary vertex ($L_{xy}$).

to be classified as a reconstructed vertex. The transverse distance between the primary and secondary vertex must be between 0.1 to 25 mm. The invariant mass of any charged particle whose associated track is compatible with the secondary vertex assignment must not exceed 6.5 GeV. In addition, secondary vertices linked to the decay of a $K_S^0$ meson are excluded. The Inclusive Vertex Finder is the algorithm conventionally employed by all b-taggers implemented for physics analyses at CMS.

During LHC Run 1 and 2, the CMS Collaboration developed several b-tagging algorithms. High-level variables were initially employed as inputs for b-jet recognition using multivariate techniques. These features are incorporated into more sophisticated discriminators using state-of-the-art machine learning algorithms. For every jet, each approach yields a single discriminator value, for which higher values denote a higher probability of a b-initiated jet. In comparison, lower values are associated with light-flavoured jets.

The two quantities that characterise the performance of various b-tagging algorithms are the *misidentification rate*, which takes into account the possibility of a light-flavoured jet being mistakenly identified as a b-jet, and the *b-tagging efficiency*, the probability of a real b-jet being identified correctly by the b-tagger. The b-tagging and vertexing group (BTV) at CMS [87] defines three operational working points (WP) to describe various b-tagger accuracy levels for jets with $p_T > 20$ GeV and $|\eta| < 2.5$.

The latter are the *tight, medium* and *loose* working points, corresponding to a non-b-jet misidentification probability of 0.1, 1 and 10%, respectively. The algorithm used to select b-tagged jets for the present work is the so-called DeepJet [88] algorithm, which

| Working point | 2018 | 2017 | 2016 preVFP | 2016 postVFP | Misidentification Rate% |
|---|---|---|---|---|---|
| loose | 0.0490 | 0.0532 | 0.0508 | 0.0480 | 10% |
| medium | 0.2783 | 0.3040 | 0.2598 | 0.2489 | 1% |
| tight | 0.7100 | 0.7476 | 0.6502 | 0.6377 | 0.1% |

Table 4.1: Definition of the working points of the DeepJet b-tagger as a function of the b-jet misidentification rate used for each Run 2 data-taking period.

uses deep neural networks for efficient jet flavour identification. The working points and the corresponding misidentification rates of the DeepJet algorithm are summarized in table 4.1. The working point adopted for this analysis is the medium working point due to the optimal balance between signal retention and fake b-jet rejection.

### 4.2.5 Tau leptons

Tau leptons belong to the third generation of lepton families and possess a mass of $m_\tau = 1776.86 \pm 0.12$ MeV [9] and a lifetime of $2.9 \times 10^{-13}$ s. Unlike other leptons, tau leptons have many potential decay modes. The primary decay modes, their branching ratios, and the intermediate meson resonances dominating certain hadronic channels are shown in Table 4.2.

| **Decay channel** | **Main resonance** | **BR (%)** |
|---|---|---|
| **Leptonic decays** | | 35.2 |
| $\tau^- \to \nu_\tau e^- \bar{\nu}_e$ | | 17.8 |
| $\tau^- \to \nu_\tau \mu^- \bar{\nu}_\mu$ | | 17.4 |
| **Hadronic decays** | | 64.8 |
| $\tau^- \to \nu_\tau \pi^-$ | $\pi(140)$ | 10.8 |
| $\tau^- \to \nu_\tau \pi^- \pi^0$ | $\rho(770)$ | 25.9 |
| $\tau^- \to \nu_\tau \pi^- \pi^0 \pi^0$ | $a_1(1260)$ | 9.3 |
| $\tau^- \to \nu_\tau \pi^- \pi^0 \pi^0 \pi^0$ | | 1.0 |
| $\tau^- \to \nu_\tau \pi^- \pi^- \pi^+$ | $a_1(1260)$ | 9.3 |
| $\tau^- \to \nu_\tau \pi^- \pi^- \pi^+ \pi^0$ | | 4.6 |
| $\tau^- \to other$ | | 3.9 |

Table 4.2: Main tau lepton decay channels and their branching ratios

The leptonic decay of taus, which happens in about 35% of the cases, is one of the tau decay channels studied for the present work. Since the lifetime of the tau lepton is relatively short, electrons and muons stemming from $\tau$ decays are hard to distinguish from those emitted promptly from the hard scattering vertex. The electron and muons

originating from tau leptons are identified employing the techniques explained in section 4.2.3 and 4.2.2, respectively.

To reconstruct the hadronic decay of taus ($\tau_h$), the so-called hadron-plus-strip (HPS) [89] algorithm is employed, which is seeded with anti-$k_T$ jets. This algorithm reconstructs the candidates using the number of charged hadrons and ECAL crystal strips with energy deposits in various decay modes. The charged decay products of tau leptons are commonly referred to as *prongs*, and the decay channels are divided based on their multiplicity. Leptonic decay includes only one prong, while hadronic ones are categorised based on the specific mesonic resonance involved. This results in the following labels: $\tau_h$ with $h \in \{\pi, \rho, a_1^{1Pr}, a_1^{3Pr}\}$. The names $a_1^{1Pr}$ and $a_1^{3Pr}$ are both assigned to decays that involve the $a_1$ meson as a resonance. These labels respectively represent decays with one ($\tau^- \to \nu_\tau \pi^- \pi^0 \pi^0$) or three ($\tau^- \to \nu_\tau \pi^- \pi^- \pi^+$) prongs.

The following decay modes are taken into account while reconstructing tau leptons with the HPS algorithm:

- *DM0*: one-prong + 0 $\pi^0$
- *DM1*: one-prong + 1 $\pi^0$
- *DM10*: three-prong + 0 $\pi^0$
- *DM11*: three-prong + 1 $\pi^0$

Depending on the number of charged and neutral hadrons, the decay mode coding used by the HPS algorithms is determined by the formula $DM = 5 \times (N_{charged} - 1) + N_{neutral}$. The following criteria must be met by the $\tau_h$ candidates, i.e. a $p_T > 30$ GeV threshold, an $\eta < 2.3$ acceptance window, unit elementary charge, and a longitudinal impact parameter $d_z < 0.2$.

The HPS method can mistakenly identify particular physics objects, such as jets, muons, and electrons, as tau leptons that are decaying hadronically. An algorithm based on neural networks, known as the Deep Tau [90], minimizes the tau-lepton misidentification rate. The multiclass convolutional neural network (NN) that serves as the foundation for the Deep Tau algorithm receives low- and high-level features from the hadronic tau candidates. The tracks and energy deposits of their decay products are classified as low-level features, whereas their transverse momenta and angular variables $\eta$ and $\phi$ are considered high-level features. The neural network gives four output scores for each class, i.e. genuine taus, jets, muons, and electrons.

### 4.2.6 Missing transverse energy

The CMS detector can identify most particles generated in proton-proton collisions. Electrically neutral and weakly interacting particles like neutrinos constitute the only exception. These particles do not leave a direct trace in the detector because of their extremely

low interaction cross-section with the detector material. Therefore, it is necessary to deduce the existence of these particles from the imbalance in total momentum, also known as the missing transverse energy (MET). The latter is defined as:

$$\vec{p}_T^{miss} = -\sum_i^{reco} \vec{p}_{T,i}, \tag{4.6}$$

where $\vec{p}_{T,i}$ represents the reconstructed transverse momentum of each particle, and the sum now spans all of the reconstructed final state particles. Detecting particles that leave the detector without interaction, such as neutrinos or neutral, weakly interacting particles predicted by BSM theories —collectively called invisible particles — combines the effects of detector inefficiencies with the MET itself. There is always at least one neutrino among the decay products of tau leptons that decay via a weakly charged current. Thus, an accurate reconstruction of the MET is necessary for the processes examined within this thesis work. The missing transverse energy $E_T^{miss}$ is defined according to two distinct criteria:

- Particle Flow MET (PF-MET) [91], which is related to equation 4.6, where the total transverse momentum is computed from all PF candidates present in the event.
- PileUP per Particle Identification MET (PUPPI-MET) [92] infers the MET component of the hard scattering process. To this end, the four-momenta of every particle $\vec{p}_{T,i}$ in the event are scaled by a weight $w_i$. This is assigned a value ranging from 0 for particles originating from pile-up vertices and 1 for the decay products of the hard scattering vertex. In this case, the MET is calculated as $\vec{p}_T^{miss} = -\sum_i^{reco} \vec{w}_i p_{T,i}$.

The weights $w_i$ are calculated using a shape $\alpha$ parameter for every particle to discriminate particles originating from the either primary or the additional pile-up vertices. Its functional form can change [92] depending on which subdetectors are used to reconstruct the particle, on the particle transverse momentum, as well as the type of neighbouring particles.

The $\alpha$ distribution is subsequently derived using the charged particles allocated to pile-up vertices on an event-by-event basis. Since the tracker data is not accessible in the very forward region, its distribution is obtained by correcting the root-mean-squared ($RMS_{PU}$) and median ($\bar{\alpha}_{PU}$) of the corresponding one in the barrel region using transfer factors obtained from simulation. The value $\alpha_i$ relative to the $i^{th}$ particle is compared to the $\alpha$ distribution for PU events to determine the probability it stems from pile-up interaction vertices. To this end, the signed chi-squared is computed as:

$$\chi_i^2 = \frac{(\alpha_i - \bar{\alpha}_{PU}|(\alpha_i - \bar{\alpha}_{PU})|)}{(RMS_{PU})^2} \tag{4.7}$$

Subsequently, the cumulative distribution of the $\chi^2$ is used to convert the $\chi^2$ term to a weight as follows:

$$w_i = F_{\chi^2,ndf=1}(\chi_i^2), \tag{4.8}$$

where $F_{\chi^2,ndf=1}$ is the cumulative distribution for a $\chi^2$ with one degree of freedom. These weights are used directly in CMS for neutral particles and assigned a value of either 0 or 1 for charged particles, depending on whether they are allocated to the primary or the additional pile-up vertices [93].

CHAPTER 5

# Machine Learning techniques

The subject of Artificial Intelligence (AI) has expanded at an outstanding pace over the past few years, with numerous applications in data analysis, pattern recognition, and decision-making. In high-energy physics, machine learning tools have been devised for event reconstruction, feature extraction, and anomaly detection, further enhancing the discovery potential of subtle signals in a complex multi-particle environment. This chapter will cover the machine learning (ML) techniques relevant to and used in the current work, assuming that the reader is already familiar with the basic concepts of ML algorithms. A thorough explanation of ML fundamentals from external sources is available in the literature, such as [94].

## 5.1 Classification in Machine Learning

The history of machine learning applications in high-energy physics dates back to the first advanced physics analyses in the early 1990s and 2000s. An exponential growth followed this period in the techniques used in particle reconstruction and event identification in 2010 [95]. Nowadays, many physics analyses employ machine learning algorithms in regression and classification problems to extend the sensitivity reach and improve the discovery potential. Both classification and regression problems fall under the category of supervised learning. In case the target variable is continuous, the prediction task is a regression, while when discrete, it corresponds to a classification problem. The current study employs a classification task to improve the analysis sensitivity.

Machine learning involves a variety of classification tasks, most of which may be subdivided into four groups: binary, multi-class, multi-label, and imbalanced classification. The objective of a *binary classification* problem is to divide the input data into two mutually exclusive categories. Depending on the type of task being addressed, the training data in such a scenario is labelled in a binary format, such as true or false, positive or

negative, 0 or 1, etc. On the other hand, *multi-class* classification aims to predict which class a given input example belongs to by using at least two mutually different class labels. In the case of *multi-label* classification, the goal is to predict zero or more classes for every input feature. In this instance, the input feature may have multiple labels, such that mutual exclusion is absent. It can be applied to various fields, such as tagging in natural language processing, where a text may contain several themes. In the case of the *imbalanced classification*, there may be significantly more instances of one class in the training data than the others due to the unequal distribution of examples within each class. The present study deals also with an unbalanced dataset, i.e. there are significantly fewer events in the class of interest "signal" than in other classes, namely "background".

### 5.1.1 Imbalanced classification methods

Imbalanced classifications present a challenge for predictive modelling, as most classification machine learning algorithms were developed assuming an equal distribution of examples across classes. Consequently, these models exhibit poor predictive accuracy, particularly for the minority class. This is problematic because the minority class is typically of greater importance, making the problem more vulnerable to misclassification errors in the minority class compared to the majority class. Multiple strategies can be employed to tackle the problem of imbalanced data sets. The most commonly used methods employ alternative customised algorithms, techniques to change the input data or both approaches.

Data-level techniques focus on transforming the training set to ensure its compatibility with a traditional learning algorithm. Regarding balancing distributions, there are two distinct approaches: oversampling, which involves generating new items for minority groups, and undersampling, which consists of removing samples from majority groups. Traditional methods employ a random selection process to choose target samples for preprocessing. However, this frequently eliminates crucial samples or includes unimportant new data points.

One of the several techniques available to perform this approach at the data level is the so-called Synthetic Minority Over-sampling Technique (SMOTE) [96]. To provide a brief overview of how SMOTE operates, the first step involves randomly selecting a minority class instance, denoted as 'a', and identifying its k nearest neighbours from the minority class. A synthetic instance is generated by randomly selecting one of the k nearest neighbours, denoted as b, and linking it to the original instance, denoted as a, to construct a line segment in the feature space. The synthetic instances are created by combining two selected input features, a and b. The SMOTE oversampling and undersampling techniques were applied to the input data of the current project and evaluated on the Multi classifier developed using the Deep Neural Network. Figure 5.1 illustrates the application of SMOTE to a subset of data for this analysis.

This method led to an enhancement in the Neural Network's output and to the res-

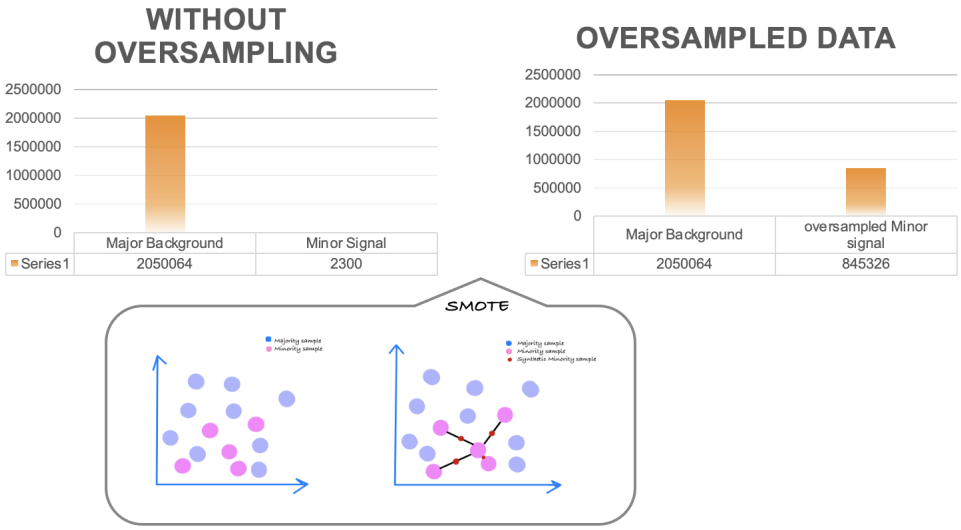


Figure 5.1: The bar chart on the left shows the number of the input data for the major background class and the minor signal class before oversampling, and the chart on the right shows the oversampled data after applying the SMOTE oversampling method.

olution of the bias towards the majority class, as discussed in the Appendix. However, ultimately it was not used due to the requirement in the HEP domain to have a physically accurate description of the input data distribution. Even a minor alteration in the distribution or shape of the input data, also referred to as *sculpting* could potentially lead to artifacts and to an incorrect physical interpretation of the results [97]. To address the issue of imbalanced data sets, existing machine learning algorithms might be adapted at the algorithmic level.

This thesis initiated the classification process by employing neural networks to categorise signals from various background classes. Subsequently, the Boosted decision tree approach (XGBoost) [98] replaced the Neural Network (NN) due to its superior performance in terms of speed and robustness. By adjusting the weights allocated to each class and achieving a balanced weight distribution, the problem of an imbalanced data set is solved, and the desired outcome was successfully achieved using the XGBoost algorithm. The comparison of the output of NN and XGBoost applied on the same data is discussed in Appendix B. The following section briefly explains why XGBoost outperformed the NN with minimal optimisation required. Moreover, a hierarchical classification approach was developed to address the issue of imbalanced data sets and to enhance the outcome. This method will be briefly discussed in the remaining part of this chapter.

### 5.1.2 Choice of Machine Learning algorithm

One crucial aspect of addressing a data science problem is the selection of an appropriate algorithm, which normally falls into one of two categories: using deep neural networks or boosted decision trees. In the case of tabular datasets, tree-based models such as XGBoost are known to perform better than deep learning models, a through discussion can be found in external sources, the brief description below is taken from [99, 100].

One of the primary factors contributing to this is the inductive bias of decision trees, which allows them to learn non-linear patterns in tabular data effectively. Applying a *smoothing* technique to the target function in the training set leads to a notable reduction in accuracy for tree-based models. This suggests that tabular data frequently include non-smooth patterns that neural networks find challenging to explain. Moreover, tabular datasets may include uninformative features, and eliminating up to half of these features does not substantially affect the accuracy of gradient-boosting trees (GBT). Multi-layer perceptron (MLP) designs have less resilience to uninformative data, and the difference in performance between MLPs and other models increases with the inclusion of uninformative features. Currently, a lot of research is being conducted [100] to evaluate various MLP models with boosted decision tree models. The benchmark study conducted in the article [99] analyses the performance of boosted trees against different algorithms on tabular datasets of varying sizes. Figure 5.2 shows one of the key results discussed in the article.

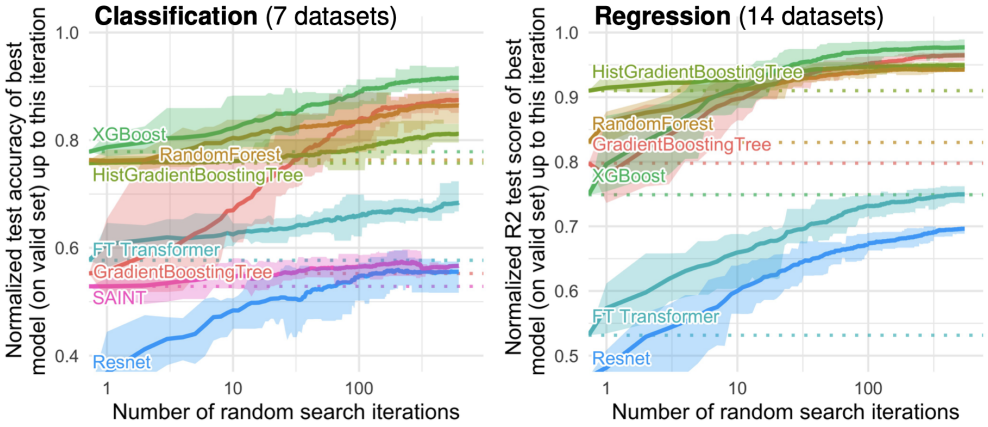


Figure 5.2: Performance comparison of various ML methods on a dataset of moderate size, containing both numerical and categorical features, for both classification and regression tasks [99].

### 5.1.3 Flat versus hierarchical classification

Flat classification refers to the conventional binary or multi-class classification techniques. Nevertheless, numerous classification tasks in the real world involve hierarchical structures, which imply a direct relationship between the classes. This is particularly evident in facial recognition algorithms, where hierarchy can be employed. Consider a given collection of biometric photographs. At a fundamental level, these images may be divided into the background and the foreground, representing the head. Additionally, the head can be further distinguished by features such as hair and eyebrows. While our brain can easily express such complexities, Machine Learning algorithms often overlook these relationships. Currently, the hierarchical technique is receiving more attention and numerous ongoing studies in machine learning, such as the article [101], are being conducted.

Classes are arranged in a class hierarchy in hierarchical classification tasks, usually as a tree or a directed acyclic graph (DAG).

Various methods exist to establish the hierarchical classification, including the local classifier approach. This approach may be further categorised into the local classifier per node or per level. The local classifier per node strategy involves the training of a binary classifier for each individual node inside the class hierarchy. This methodology trains a multi-class classifier or binary classifier such as the One-Against-One scheme Single Vector Machine (SVM) for each parent node in the class hierarchy to differentiate between its child nodes. The illustration of this approach is shown in the Figure 5.3 on the left site.

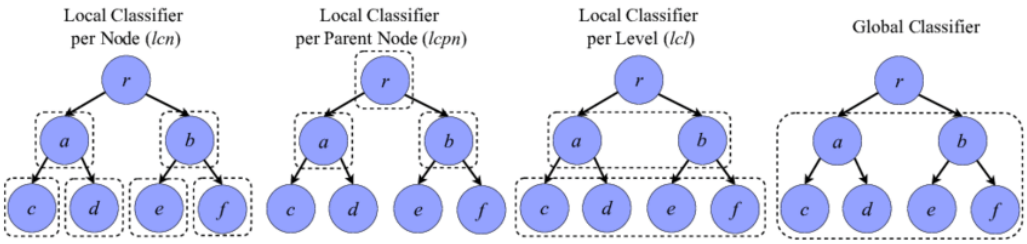


Figure 5.3: The diagram illustrates many hierarchical classifier approaches, including the local classifier per node, local classifier per parent node, local classifier per level, and the global classifier. The dashed boxes indicate the number of classifiers required for each strategy. The figure is sourced from the publication referenced as [102].

The local classifier per level strategy involves the training of a separate multiclass classifier for each level inside the class hierarchy. Furthermore, it is the preferred methodology for analysing the data in the current thesis.

The hierarchical model for this work is organised in the following manner. The initial model is a binary classifier that distinguishes signal events from background events. It is important to note that signal events represent the minority class, while background events represent the majority. The next level employs a multi-classifier to differentiate subclasses

within the signal and background classes. Additionally, the output from the training of the initial model is utilised as additional variables for the multi-classifier. Different subsets of the training are used, and at both levels, the class weights are adjusted to balance to address the problem of imbalance. The frameworks as mentioned earlier are created and used in conjunction with the XGBoost algorithm. The outcome will be discussed in detail in the analysis section.

## 5.2 Shapley Additive Explanations

There is a common practice of referring to the outcome of machine learning models as a "black box". In fact, it can prove challenging to explain the output of complex machine learning models like deep neural networks and to interpret how such a result was achieved. The field of Explainable AI (xAI) aims to clarify and comprehend the output of ML models by humans through explanation and interpretation. Over the last few years, experts in this subject have developed several approaches, including the SHapley Additive exPlanations (SHAP) [103] method. The latter uses a game-theoretic approach to quantify the individual contribution of each player to the final result. In machine learning, every feature is assigned a significance value that indicates its impact on the model output. SHAP values provide insights into the impact of each feature on the final prediction, the relative importance of each feature compared to others, and the extent to which the model depends on the interaction between features.

The SHAP technique decomposes the model prediction into parts that can be assigned to specific variables. The primary concept is to evaluate the significance of a variable by examining its impact on the model response when it is added to a subset. This approach relies on a value function considering a particular set of variable indices. This function represents the model output when specific variables in the subset are given. The term can be defined as the expected value for a conditional distribution or as the model prediction after eliminating values of variables outside the subset as follows:

$$f_s = E_{conditional\ distribution}, \tag{5.1}$$

where $s$ is a subset of variable indexes, and $E$ is the expected value typically used for a tabular data set. Moreover, the objective is to break down the prediction differences into parts assigned to individual variables. Reviewing variable importance involves analysing the impact of adding a variable to the subset of the value function equation 5.1. The contribution of a variable is represented as a weighted average across all potential subsets, as indicated in the equation below:

$$\phi_i(s) = \frac{1}{p}\sum_{\sigma}[f(S \cup (i)) - f(S \cup (\sigma_1, \sigma_2, ...\sigma_{i_1}))]. \tag{5.2}$$

The equation 5.2 calculates the contribution $\phi_s(s)$ of a variable $i$ to the value of function $f(s)$, where $\sigma$ represent a set of all orderings of $p$ variables. SHAP values are computed by averaging the contributions of variables over all possible orderings. SHAP values allocate differences in predictions across variables, and their calculation requires looking at all possible orderings. The addition of SHAP values to the model prediction enables the forecast to be broken down into elements that can be assigned to different variables. It is essential to comprehend the significance of each variable to fully understand this feature contribution to the model output.

CHAPTER 6

# Analysis of bbH production in final states with leptons

Following the discovery of a boson with a mass close to 125 GeV compatible with the Higgs boson in the Standard Model (SM) [5,6], significant knowledge has been gained about this particle through the experimental and theoretical efforts. The agreement of its properties with those of the Higgs boson in the Standard Model was verified upon discovering this particle. The existence of interactions between the Higgs boson and other bosons and third-generation fermions has been confirmed through the detection of Higgs boson decays into $\gamma\gamma$, ZZ, WW, tt, and bb, as well as the observation of Higgs boson production in association with a top-antitop quark pair (ttH) [104].

So far, most of the Higgs boson production modes with large cross sections, such as gluon-gluon fusion, vector boson fusion, Higgsstrahlung and ttH, have been experimentally studied except for the production of Higgs boson in association with a pair of b-quarks (bbH). The methods employed in this analysis to search for bbH production will be discussed in this chapter.

## 6.1 Analysis strategy

The present analysis uses the full Run 2 data set corresponding to an integrated luminosity of 138 $\text{fb}^{-1}$ collected by the CMS experiment at the center of mass energy 13 TeV.

This search studies the Higgs boson's production associated with b-jets and its further decay in final states involving leptons. Leptons are chosen because they allow for an effective rejection of background events while having a relatively moderate branching fraction. The final states studied for this analysis are the Higgs boson decay to $\tau\tau$ and $WW$. For the $H \to \tau\tau$ pair, the $e\mu$, $\mu\tau_h$, $e\tau_h$ and $\tau_h\tau_h$ final states are studied, while for the $H \to W^+W^-$ case only the $e\mu$ channel has been investigated. Given that the analysis

primarily focuses on the bbH production mode rather than the decay process, to increase the sensitivity to this production mode, both contributions to the $e\mu$ final state arising from the $H \to \tau\tau$ and $H \to W^+W^-$ decay chains are taken into account. These two decay chains yield a comparable event topology. The main focus of this chapter is the analysis performed in the fully leptonic channel $e\mu$.

The first step of the analysis includes identifying and characterizing the final state physics objects within the experimental data. This includes particles relevant to the present study, such as tau leptons and b-tagged jets. Events are filtered based on specific criteria, requiring the presence of tau lepton pairs and at least one b-tagged jet. This selection helps to isolate events that are relevant to the analysis. The next step of the analysis is the modelling of the signal and background processes. The signal of interest related to b-quark associated production of the Higgs is modelled using simulated samples. Background processes are modelled either with simulated samples or by using dedicated data-driven techniques. A Boosted Decision Tree (BDT) algorithm is employed to improve signal-to-background discrimination. This machine learning technique enhances the analysis sensitivity to the signal process of interest. The next step involves categorising the events based on their BDT scores., which involves separating events into distinct signal and background categories.The final step involves extracting the signal from the simultaneous fit to the BDT score distributions across all event categories. The primary objectives of the analysis are to derive stringent constraints on the inclusive cross-section of b-quark associated production and on the Higgs Yukawa couplings to third-generation quarks through a 2D likelihood scan of $(\kappa_t, \kappa_b)$. This chapter provides a thorough discussion of the analysis strategy followed by the analysis of the results.

## 6.2 Signal and background processes

### 6.2.1 Signal definition

Within the SM, two primary production modes contribute to the bbH production as exemplified in Figure 6.1, displaying a Feynman diagram of either case. The dominant contribution to the b-associated Higgs boson production comes from the gluon-gluon fusion (ggF), where the two b-quarks in the final state are produced by means of gluon splitting (Figure 6.1 left) [105, 106]. The second most relevant production mechanism occurs via b-quark fusion, as shown in Figure 6.1 right. This second production mode provides access to the b-quark Yukawa coupling.

The analysis of bbH production poses experimental challenges for two primary reasons. Firstly, while the bbH production mode has a predicted cross-section of 0.48 (+0.097,-0.1157) pb [40], which is comparable to the ttH production rate of 0.51 pb, it is more challenging to suppress the accompanying irreducible background processes with respect to the case of ttH production. Furthermore, the bbH production mode interferes with

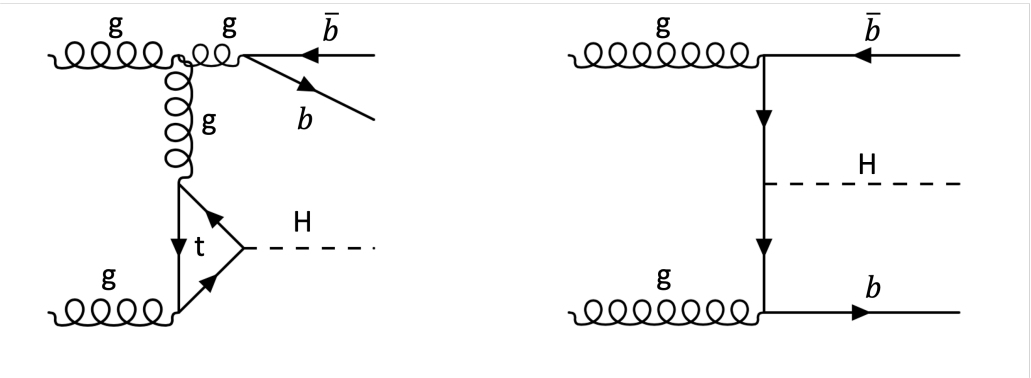


Figure 6.1: Feynman diagrams of the two main production modes of Higgs boson in association with a pair of b-quarks, (left) gluon-gluon fusion (ggF) with two additional b-quarks in the final state, (right) b-quark fusion.

other Higgs production processes, explicitly exhibiting negative interference with the ggF process. This diminishes the overall signal yield and restricts the ability to constrain the Higgs Yukawa coupling to bottom quarks directly.

In this analysis, various notations will be employed to distinguish between the bbH production mode and the production mode in which the Higgs is generated through the Higgs Yukawa coupling to the bottom quark. The latter, displayed on the right side of the figure 6.1, is denoted as bbH($y_b^2$) and corresponds to a cross-section of 0.482 pb at next-to-leading order (NLO) [40]. The production mechanism of the Higgs boson through its coupling to the top quark is known as bbH($y_t^2$) (Figure 6.1 left), with a cross-section of 1.040 pb (NLO reweighted to N3LO) [40]. The negative interference term is also called bbH($y_b y_t$), with a cross-section of -0.033 pb. Therefore, the overall cross-section of bbH is equal to 1.489 pb [40]. The section 2.4 provides a comprehensive analysis of the diagrams illustrating the bbH production.

Two different methods, the so-called 4-flavour scheme (4FS) and 5-flavour scheme (5FS), can be used to estimate the overall cross-section for the bbH production mode. Due to the considerable mass of the bottom quark compared to the QCD scale,i.e. $m_b \gg \Lambda_{QCD}$, the production of bottom quarks can be treated as a perturbative process. In the four-flavour scheme (4FS), where b-quarks are not considered as partons in the proton, gluons participate mainly in the initial state; however, when the mass of the Higgs boson denoted as $m_H$ is much larger than the mass of the bottom quark $m_b$ ($m_H \gg 4m_b$), the bottom quark distribution function should be considered, which corresponds to the calculations within the five flavour scheme (5FS). The two schemes are complementary and have strengths in describing distinct aspects of b-quark processes. However, knowing the limitations of each scheme, it is possible to combine both calculations using a weighting procedure [107], or other approaches such as [108] or decide on one of the schemes. The

Yukawa coupling to b-quarks necessitates using the 4FS approach, as a massless b-quark implies no Yukawa coupling to bottom quarks (case of 5FS); hence the cross-section for the bbH samples in this thesis is calculated within the 4FS scheme [40]; signal samples and their cross-sections' times branching fraction are summarized in table 6.1. Furthermore, to cover the b-quarks from the Parton Distribution Function (PDF), the Standard Model (SM) Higgs to $\tau$ sample, calculated with the 5FS, is used.

Table 6.1: Signal samples and their respective cross-section times branching fraction

| Sample name | $\sigma \cdot B$ (pb) |
|---|---|
| bbHToTauTau M-125 4FS yt2 TuneCP5-13TeV-amcatnlo-pythia8 | 1.040 · 0.06208 |
| bbHToTauTau M-125 4FS yb2 TuneCP5-13TeV-amcatnlo-pythia8 | 0.4822 · 0.06208 |
| GluGluHToTauTau M125 TuneCP5 13TeV-amcatnlo-pythia8 | 48.52 · 0.06208 |
| bbHToTauTau M-125 4FS ybyt TuneCP5-13TeV-amcatnlo-pythia8 | -0.033 · 0.06208 |
| bbHWWTo2L2Nu M-125 4FS yt2 TuneCP5 13TeV-amcatnlo-pythia8 | 1.040 · 0.0231 |
| bbHWWTo2L2Nu M-125 4FS yb2 TuneCP5 13TeV-amcatnlo-pythia8 | 0.4822 · 0.0231 |
| GluGluHWWTo2L2Nu M125 TuneCP5 13TeV-amcatnlo-pythia8 | 48.52 · 0.0231 |
| bbHWWTo2L2Nu M-125 4FS ybyt TuneCP5 13TeV-amcatnlo-pythia8 | -0.033 · 0.0231 |

### 6.2.2 Background processes

The main irreducible SM background processes contributing to the analysis of bbH are discussed below.

#### top anti-top production

The top anti-top production ($t\bar{t}$) contributes to many analyses at the LHC as a primary background process due to its large cross-section. This process can occur through gluon-gluon fusion and $q\bar{q}$ annihilation, and at the LHC, gluon fusion is the leading mechanism for this process. The $t\bar{t}$ process is the main background process in all the channels; however, it is more prominent in the $e\mu$ channel. A top quark often decays to a W-boson that can further decay leptonically and to a b-quark. An example Feynman diagram of the $t\bar{t}$ process is shown in Figure 6.2.

#### Single top production

Single-top quark electroweak production is possible in three channels: t-channel, s-channel and tW. The respective Feynman diagrams are shown in Figure 6.3. This process has a relatively low cross-section; however, it mimics the $H \rightarrow W^+W^-$ kinematics and is mainly relevant in the $e\mu$ channel.

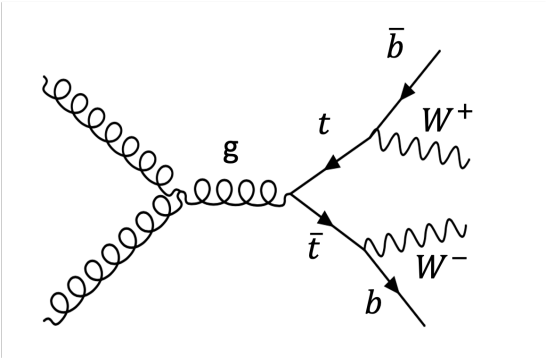


Figure 6.2: Example of one of the Feynman diagrams of top anti-top production mechanism

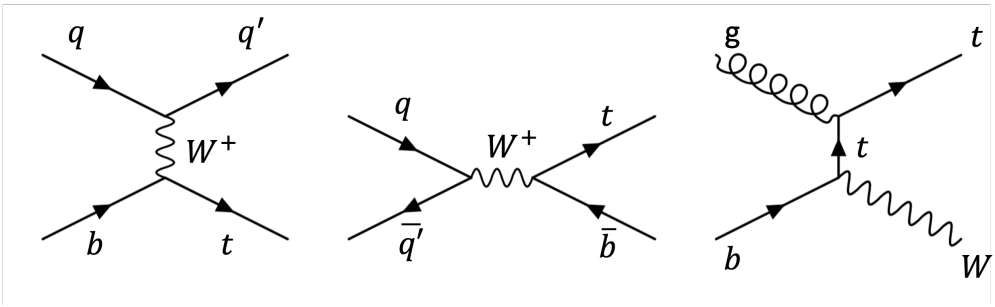


Figure 6.3: Feynman diagrams of the single top production process for the three different channels respectively from left to right: t-channel, s-channel and tW.

### V+jets

The V+jets process can produce a similar final state to the signal in all the channels of this analysis. The V boson decays as $Z \to ee, \mu\mu, \tau\tau$ or $W \to e\nu, \mu\nu, \tau\nu$. The background resulting from the V+jets is irreducible and is associated with a large production cross-section. However, since the jets in the final state of the V+jets process can be categorized based on their flavour, light flavour jets in the background can be substantially reduced using the b-tagging algorithm. Therefore, to reduce the Z+jets background, at least a b-tagged jet is required in the present analysis. The Z+jets process is denoted as Drell-Yan+Jets in the present work. In the case of the W+jets background process setting a cut on the transverse mass can reduce the background.

### Diboson production

The diboson production processes (WW, WZ, ZZ) are considered in this analysis. The diboson process can mimic the same final state as the bbH signal when a Z-boson decays (semi)-leptonically, including one or more neutrinos.

**QCD multi-jet production**

The QCD multi-jet process contributes to all the channels of this analysis due to the highly hadronic environment of pp collision events at the LHC, as b-quark pairs can be emitted in the QCD events.

**other backgrounds**

The other two background processes considered in the $e\mu$ channel are the associated production of the W and Z bosons with top-quark pairs ($t\bar{t}V$) and the associated production of the W boson and photon (W$\gamma$), in which the W boson subsequently decays into leptons.

Furthermore, in this search, other Higgs processes, such as the ttH production, are considered as background.

## 6.3 Data and simulated samples

The data used for this search corresponds to the pp collisions recorded by the CMS experiment during Run 2, i.e. from 2016 to 2018, at the centre of mass energy $\sqrt{s} = 13$ TeV. The collected data has a combined integrated luminosity of 137.6 fb$^{-1}$. The total integrated luminosity corresponding to each year corresponds to 36.3 fb$^{-1}$ for 2016[1], 41.5 fb$^{-1}$ for 2017 and 59.7 fb$^{-1}$ for 2018.

For the signal samples used in this work, which are shown in table 6.1, the simulation was done at next-to-leading (NLO) order with the MadGraph5-aMC@NLO 2.6.1 generator [65]. As explained in section 6.2.1, there are three different samples corresponding to the Yukawa coupling ($y_b^2, y_t^2, y_b y_t$) of the bbH production process.

For the events coming from the b-quark fusion samples and the interference term, i.e. samples containing $y_b^2$ and $y_b y_t$ amplitudes in table 6.1, events are selected with a least one generator level bottom quark jets. The events with zero-generated jets are considered to be background processes. From the gluon-gluon fusion samples, i.e. $y_t^2$ coupling, events are selected either with two generator level b jets or, in the case of one generator level b jet, with at least two generator level b hadrons matched to it with $\Delta R < 1.5$. The events with at least one generator level b quark outgoing from the hard scattering vertex contribute to the overall bbH sample ($y_t^2$). They are stitched to the samples of gluon fusion with two generator-level b-jets. For the signal samples, the parton showers are simulated with Pythia 8.240 [67].

The primary background process for all the channels is $t\bar{t}$ production, and it is simulated with the Powheg 2.0 generator [66] at NLO. The second most important back-

[1]The 2016 data is divided into two eras of 2016preVFP and 2016 postVFP due to the changes in detector conditions. The low signal-to-noise ratio and fewer hits on tracks in 2016preVFP were linked to saturation effects occurring in the preamplifier of the APV25 readout chip. Hence, this effect was mitigated by changing the feedback preamplifier bias voltage (VFP) in 2016postVFP.

ground among all the channels, the V+jets, is produced with the MadGraph5-aMC@NLO 2.6.5 generator. The single top process, more relevant for the $e\mu$ channel, is generated with the Powheg 2.0 at NLO. For the diboson process relying on the specific combination of vector boson studied, the production has been performed either with MadGraph5-aMC@NLO 2.6.5 at NLO or with the Powheg generator. The associated production of vector bosons and top-quark pairs ($t\bar{t}V$) and the associated production of the W boson and photon ($W\gamma$) are simulated with MadGraph5-aMC@NLO at LO level. However, to match the NLO(LO) matrix element calculations with the parton shower model, the MLM [109] prescription is used. For all the simulated samples, the NNPDF 3.1 [110] set of Parton Distribution Functions (PDFs) is used. The parton shower and hadronization processes are performed with Pythia 8.240 with the CP5 [111] tune. The simulation of the CMS detector response is achieved with GEANT4 [69] for all the processes. Furthermore, the simulation of additional pp interactions per bunch crossing, i.e. pileup, is performed with Pythia and corrected to match the pileup measured in recorded data for each year. An overview of the background samples and their respective cross-section is given in table 6.2. Additionally, the Higgs boson samples considered as background in this search are shown in table 6.3.

## 6.4 Trigger and event selection

A two-stage online trigger is used to select the events at CMS, as previously discussed in chapter 3.2.5. The Level 1 (L1) trigger is a hardware-based trigger which reduces the event rate from 40 MHz to 100 kHz and conducts primary selection and counting of the physics objects. In the second stage, the High-Level Trigger (HLT) performs a more accurate reconstruction of the events that passed the L1 trigger and were acquired by the DAQ system. The so-called trigger path can identify the selection list performed at the HLT level. The $e\mu$ channel events are selected online from a combination of $e + \mu$ cross trigger, which has asymmetric thresholds. The cross triggers in this channel have a $p_T$ threshold of 8 GeV and 23 GeV for muons or 23 GeV and 12 GeV for electrons. The trigger path for the $e\mu$ channel and all other three channels are shown in table 6.4.

### 6.4.1 Event selection

The signal event definition consists of a series of selections applied to data and simulation. The online event selection in the $e\mu$ channel requires the presence of the electron-muon pair, while the presence of two hadronically decaying tau leptons is needed in the $\tau_h\tau_h$ channel. Exactly one muon (one electron) must be present per event in the $\mu\tau_h$, $e\tau_h$ channels respectively. The selected hadronically or leptonically tau leptons should match the objects used for the online selections in all the channels. This criterion is fulfilled by requiring small angular separation with respect to the online objects and that the

Table 6.2: Simulated background samples and their respective production cross sections

| **Sample Name** | **Cross Section (pb)** |
|---|---|
| TTTo2L2Nu | 88.29 |
| TTToHadronic | 377.96 |
| TTToSemiLeptonic | 365.34 |
| ST s-channel antitop leptonDecays | 3.97 |
| ST s-channel top leptonDecays | 6.35 |
| ST t-channel antitop 4f InclusiveDecays | 80.95 |
| ST t-channel top 4f InclusiveDecays | 136.02 |
| ST tW antitop 5f inclusiveDecays | 35.85 |
| ST tW top 5f inclusiveDecays | 35.85 |
| WWToLNuQQ | 45.99 |
| WWTo2L2Nu | 11.08 |
| WWTo4Q | 47.73 |
| WZTo3LNu | 5.052 |
| WZTo2L2Q | 6.331 |
| WZTo1L3Nu | 3.3 |
| WZTo1L1Nu2Q | 11.66 |
| ZZTo4L | 1.369 |
| ZZTo2L2Q | 3.688 |
| ZZTo2Q2Nu | 4.561 |
| WGToLNuG | 464.4 |
| ttZJets | 0.252 |
| ttWJets | 0.204 |
| DYJetsToLL M-50 | 6077.22 |
| DYJetsToLL 0J | 5125 |
| DYJetsToLL 1J | 951.4 |
| DYJetsToLL 2J | 358.6 |
| WJetsToLNu | 61526.7 |
| WJetsToLNu 0J | 53300 |
| WJetsToLNu 1J | 8949 |
| WJetsToLNu 2J | 3335 |

selected leptonic or hadronic candidates have a transverse momentum higher than the corresponding trigger threshold of 1 GeV for the leptonic decay of taus and of 5 GeV for the hadronic decay of taus.

Among all the channels ($e\mu$, $\tau_h\tau_h$, $\mu\tau_h$, $e\tau_h$), common standard selections are applied. In all the channels, the presence of two leptons or two hadronic tau candidates with the

| **Sample name** | **Process Name** | $\sigma \times B$ **(pb)** |
|---|---|---|
| VBFHToTauTau | qqH, H $\to \tau\tau$ | 3.771 $\times$ 0.06208 |
| WminusHToTauTau | $W^-H, H \to \tau\tau$ | 0.831 $\times$ 0.06208 |
| WplusHToTauTau | $W^+H, H \to \tau\tau$ | 0.527 $\times$ 0.06208 |
| ZHToTauTau | $ZH, H \to \tau\tau$ | 0.877 $\times$ 0.06208 |
| ttHToTauTau | $t\bar{t}H, H \to \tau\tau$ | 0.503 $\times$ 0.06208 |
| VBFHToWWTo2L2Nu | qqH, H $\to WW \to 2\ell 2\nu$ | 3.771 $\times$ 0.0231 |
| HWminusJ HToWW | $W^-H, H \to WW$ | 0.831 $\times$ 0.2152 |
| HWplusJ HToWW | $W^+H, H \to WW$ | 0.527 $\times$ 0.2152 |
| HZJ HToWW | $ZH, H \to WW$ | 0.877 $\times$ 0.2152 |
| ttHJetToNonbb* | $t\bar{t}H \to WW$ | 0.212 |

Table 6.3: Background samples for the Higgs production processes generated with POWHEG with the Tune CP5 and with the parton shower and hadronization simulated with PYTHIA and their corresponding production cross sections times branching ratio. The ttH* sample is simulated with MADGRAPH5-aMC@NLO and MADSPIN with the FXFX MERGING SCHEME. This sample is only considered in the $e\mu$ channel considering the decay of Higgs to W bosons and a smaller fraction of Z bosons. The decay of Higgs to $\tau\tau$ and $bb$ are excluded from this sample.

opposite electric charge is required. Furthermore, the lepton or $\tau_h$ candidate pair should be separated by an angle $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$ greater than 0.3 in the $e\mu$ channel and greater than 0.5 for the other three channels. Additionally, one or two b-tagged jets are required. The following sections will discuss the specific event selection for the $e\mu$ channel since it is the main focus of this thesis.

### Event selection in the $e\mu$ channel

In the $e\mu$ channel, events are targeted by a trigger containing electron and muon pairs with opposite signs (OS) of electric charge. In the case that several pairs exist, the pair with the most isolated leptons are selected based on the following criteria: the pair with the most isolated muon is preferred; however, if the isolation of the muon is the same in both pairs, the pair with the most isolated electron is preferred. If the isolation of the electron is the same in both pairs, the pair with the highest muon $p_T$ is selected. In case the $p_T$ of the muon is the same in both pairs, the pair with the highest electron $p_T$ is chosen.

In this channel, both electrons and muons are required to have $p_T > 15$ GeV and $|\eta| < 2.4$ and impact parameters $d_z < 0.2$ cm and $d_{xy} < 0.045$ cm. The electron and muon should be separated by $\Delta R > 0.3$.

Muons are further required to pass the isolation criteria of $I_\mu < 0.2\ p_T^\mu$ and pass the medium muon ID points. The isolation criteria for electrons is required to be $I_e < 0.15\ p_T^e$

| channel | HLT paths |
|---|---|
| $e\mu$ | HLT Mu23 TrkIsoVVL Ele12 CaloIdL TrackIdL IsoVL DZ<br>HLT Mu8 TrkIsoVVL Ele23 CaloIdL TrackIdL IsoVL DZ |
| $\tau_h\tau_h$ | HLT DoubleMediumChargedIsoPFTau40 Trk1 TightID eta2p1 Reg v<br>HLT DoubleTightChargedIsoPFTau40 Trk1 eta2p1 Reg v<br>HLT DoubleTightChargedIsoPFTau35 Trk1 TightID eta2p1 Reg v<br>HLT DoubleMediumChargedIsoPFTauHPS35 Trk1 eta2p1 Reg v |
| $\mu\tau_h$ | HLT IsoMu24<br>HLT IsoMu27 |
| $e\tau_h$ | HLT Ele32 WPTight Gsf<br>HLT Ele35 WPTight Gsf |

Table 6.4: The HLT trigger path used in the $e\mu$, $\tau_h\tau_h$, $\mu\tau_h$ and $e\tau_h$ for 2018.

and electrons have to pass the 90% efficiency working point of the electron ID MVA with the exclusion of isolation variables in the MVA ID training. Furthermore, the electron tracks are required to have less than two missing hits in the vertex detector and not to be identified as the product of photon conversion. In addition, either the electron or muon must match the high $p_T$ leg of the HLT trigger path with $p_T > 24$ GeV.

To veto the selection of a second muons or electrons, the event is required not to contain other muons with $p_T^\mu > 10$ GeV and $|\eta| < 2.4$ passing the muon medium ID working points. Additionally, the additional muons are vetoed if they have impact parameters of $d_z < 0.2$ cm and $d_{xy} < 0.045$ cm with the requirement for isolation variable of $I_\mu < 0.3 p_T^\mu$.

The second electrons are vetoed for events not containing electrons with $p_T^e > 10$ GeV and $|\eta| < 2.5$, which passes the 90% efficiency working point of the electron ID MVA with the exclusion of isolation variables in the MVA ID training. The vetoed electrons have impact parameters of $d_z < 0.2$ cm and $d_{xy} < 0.045$ cm and isolation requirement of $I_e < 0.3 p_T^e$ which passes the photon conversion veto and have less than two missing hits in the vertex detector.

### Control distributions

The comparison of the data and simulated control distributions in the $e\mu$ channel after applying the inclusive selection, i.e. with no b-tagging selection applied, are shown in Figure 6.6 and 6.7 for the 2018 data-taking year. Overall, a good agreement is observed among all data taking years. The analogous control plots for the 2016 and 2017 data-taking years are shown in Appendix A.

### Reconstruction of high-level variables

The data and simulation control distributions shown in Figures 6.6 and 6.7 illustrate some of the most important high-level variables that are used for the analysis of this work. This section gives a short description of the reconstruction of variables such as invariant mass transverse momentum, etc.

### Invariant mass

Given the presence of the neutrinos in the $H \to \tau^+\tau^-$ final state, reconstruction of the full four-momentum of the $\tau\tau$ system is not feasible. This is because neutrinos do not interact with the detector material and hence leave no sign in the detector. To account for this effect, the sum of missing energy in the transverse plane, the so-called $E_T^{miss}$ can be measured, and to reconstruct the invariant mass of the $\tau\tau$ system, one can consider only the visible decay products of the $\tau$ leptons and compute their invariant mass. The latter is the visible mass ($m_{vis}$). The distribution of the visible mass is shown for the 2018 data in Figure 6.6 and for 2017 and 2016 in Figure 6.4.

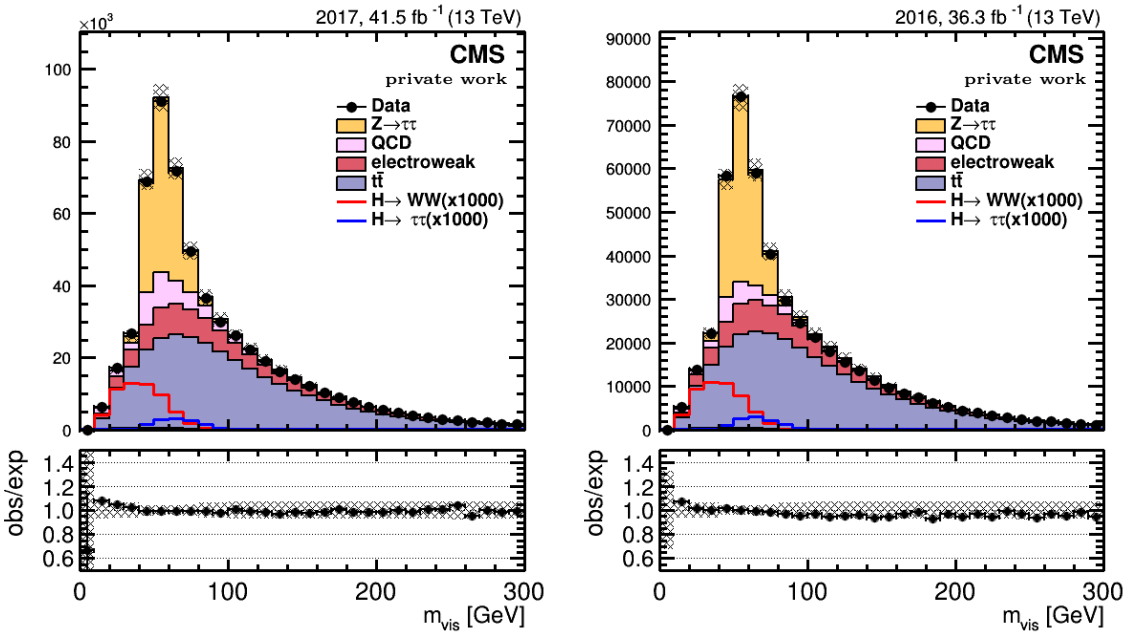


Figure 6.4: Data and Monte Carlo distributions in $e\mu$ channel for $m_{vis}$ variable for 2017 (left) and 2016 (right) data taking years.

A more precise method to asses the invariant mass of the di-$\tau$ system is performed by the so-called SVFit algorithm [112]. The latter is a dedicated algorithm based on a maximum likelihood approach to derive the di-$\tau$ invariant mass on an event-by-event basis. A detailed explanation of this approach can be found in [112].

### Variable $D_\zeta$

Another essential variable used in the present work to analyse $e\mu$ is the $D_\zeta$ [113] variable. The term $D_\zeta$ is defined as the sum of the projection of the missing transverse momentum $p_T^{miss}$ along the bisector of the leptons or $\tau_h$ candidates ($\hat{\zeta}$) and the momentum of the dilepton (or lepton and $\tau_h$ candidate) system ($\vec{p}_T^{tot}$) as follows:

$$D_\zeta = p_\zeta^{miss} - 0.85 p_\zeta^{vis}, \tag{6.1}$$

with

$$p_\zeta^{miss} = \vec{p}_T^{miss} \cdot \hat{\zeta}; \quad p_\zeta^{vis} = \vec{p}_T^{tot} \cdot \hat{\zeta}. \tag{6.2}$$

In the equation 6.1, the value of 0.85 was fine-tuned to enhance the discrimination ability between resonant $\tau\tau$ decays, namely those originating from Higgs or Z bosons and processes with more evenly distributed neutrino emissions, such as $t\bar{t}$ or Higgs decays to W bosons [113].

This variable efficiently distinguishes Higgs decays to $\tau$ leptons from events in which the neutrinos are emitted with less collimation to the leptons, such as $t\bar{t}$ or Higgs decays to W bosons. As a result, it is particularly effective for the separation of signal $H \to \tau^+\tau^-$ form $H \to W^+W^-$ in the $e\mu$ channel. The data and MC distribution for the $D_\zeta$ variable is shown in Figure 6.7 for 2018 and in Figure 6.5 for 2017 (left) and 2016 (right).

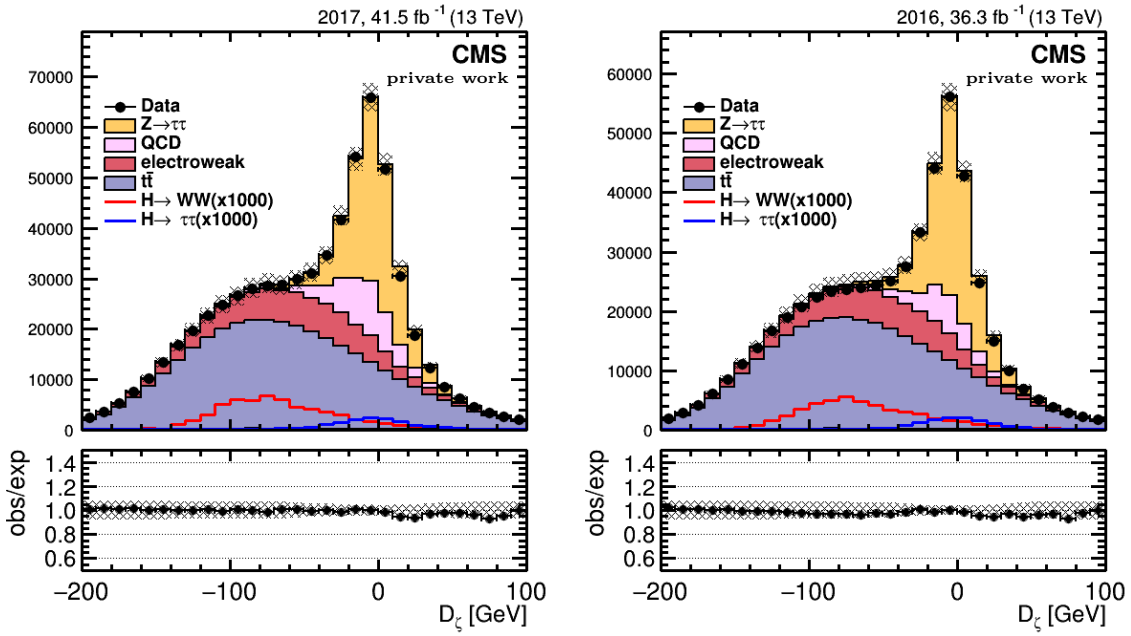


Figure 6.5: Data and Monte Carlo control distributions in $e\mu$ channel for the $D_\zeta$ variable for 2017 (left) and 2016 (right) data taking years.

**Transverse mass**

The transverse mass $m_T$ is defined either for a pair of leptons (in the case of $e\mu$ channel) or for a lepton and the missing transverse momentum (in semi-leptonic channels) is as follows:

$$m_T(\ell_1, \ell_2) = \sqrt{2p_T^1 p_T^2(1 - \cos\Delta\phi(\vec{\ell_1}, \vec{\ell_2}))}, \tag{6.3}$$

with $\vec{\ell}$ being a vector in the transverse plane. The transverse mass $m_T$ is calculated in semileptonic channels using the charged lepton and the missing transverse momentum, aiming to mitigate the contribution from electroweak background processes. However, in the $e\mu$ channel, it serves a different purpose by determining the total transverse mass for the electron, muon, and missing transverse momentum system defined as follows:

$$m_T^{tot} = \sqrt{m_T(\mu, e)^2 + m_T(e, p_T^{miss})^2 + m_T(\mu, p_T^{miss})^2}, \tag{6.4}$$

which serves as an estimator for the Higgs boson transverse mass in both $H \to \tau^+\tau^-$ and $H \to W^+W^-$ decays. The data-MC distribution for the $m_T^{tot}$ is shown in Figure 6.6 for the 2018 data-taking period.

## 6.5 Background estimation in the $e\mu$ channel

The dominant backgrounds in the $e\mu$ channel and their corresponding Feynman diagrams are discussed in section 6.2.2. The $e\mu$ channel considers several background processes: Drell-Yan jets (Z+jets), $t\bar{t}$, W+jets, QCD, Diboson, single-top and electroweak W and Z productions. About 85-90% of the background events in this channel contain genuine prompt leptons, which are estimated from Monte Carlo (MC) simulation. For the processes $t\bar{t}$, W/Z+jets, single top, and diboson, in which one or both leptons are misidentified as jets, the contribution from these backgrounds are estimated from simulations and corrected for the difference in misidentification rate between data and MC. The only data-driven background process in this channel is the QCD multijet background. Table 6.5 summarizes background events in the $e\mu$ channel and their estimation method, which will be discussed in the following section.

### 6.5.1 QCD background estimation

A data-driven *ABCD* [114] method is employed to estimate the QCD background contribution to the $e\mu$ channel to prevent inaccuracies arising from potential incorrect model assumptions in simulation. In this data-driven method, the background process yield (QCD multi-jet production) in the signal region is estimated by extrapolating the contribution of the same process from adjacent control regions obtained by inverting one or

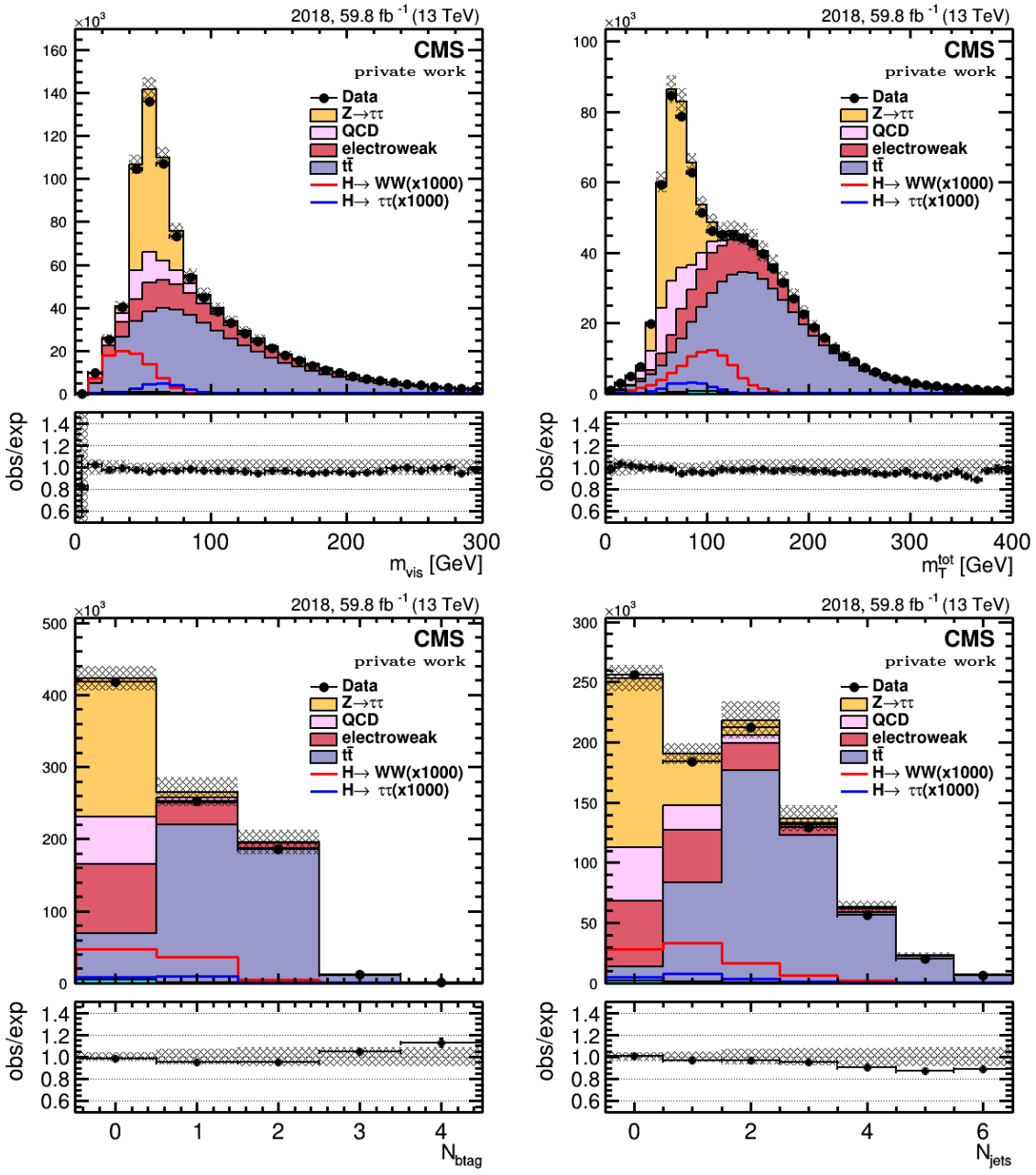


Figure 6.6: Data and Monte Carlo distributions in the $e\mu$ channel in the control region for the 2018 dataset. From top left to bottom right: visible mass of the electron-muon system, total transverse mass, b-tagged jet multiplicity, jet multiplicity.

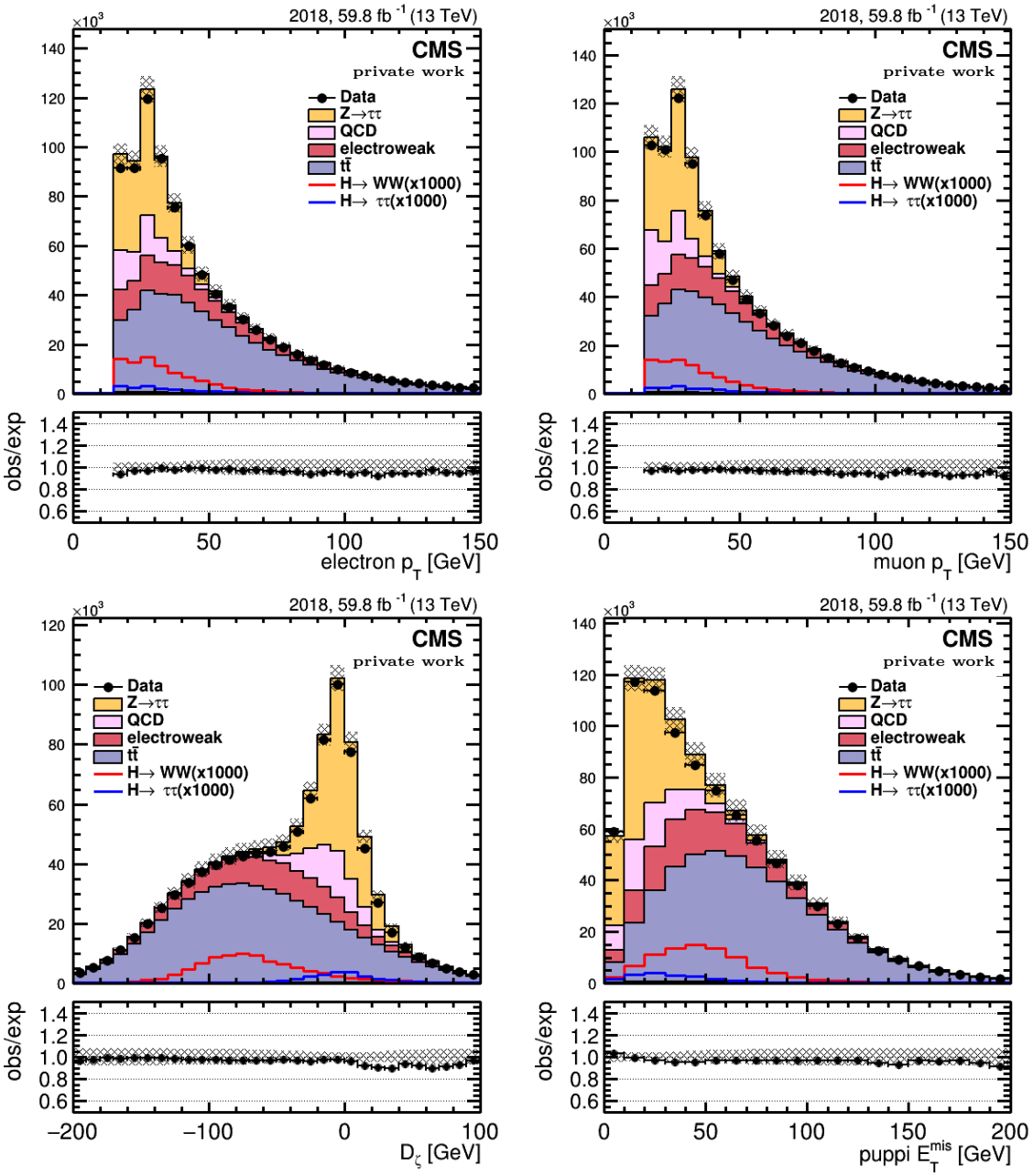


Figure 6.7: Data and Monte Carlo control distributions in the $e\mu$ channel for 2018 data set. From top left to bottom right: $p_T$ of electron , $p_T$ of muon, $D_\zeta$, PUPPI $E_T^{miss}$ (discussed in 4.2.6).

| Background | Estimation method |
|---|---|
| DY+jets, $t\bar{t}$ , diboson. single-top (Genuine prompt leptons) | simulation (MC) |
| QCD multi jets | data-driven |
| non-QCD with non-prompt leptons (W+jets, $t\bar{t}$, single-top) | simulation (MC) corrected with jet $\rightarrow e/\mu$ fake rate scale factors |

Table 6.5: Background processes in the $e\mu$ channel and their respective estimation method

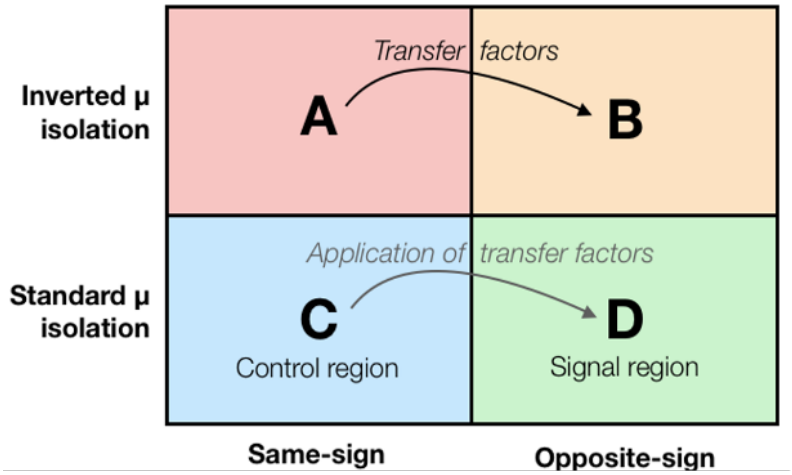


Figure 6.8: Diagram of the control and signal regions definition in the ABCD method used for QCD background estimation.

more selection cuts (lepton isolation or same-sign charge requirement). The leading assumption is that the shape of the QCD distribution of any observable in the signal region is the same as the equivalent one in the control region. However, the overall normalisation can vary from the signal to the control region. A dedicated QCD extrapolation factor must be introduced to account for the difference in the respective yields. Figure 6.8 illustrates diagrammatically the simplest version of the ABCD method to estimate the QCD background. The primary background sources are QCD events with jets misreconstructed as leptons and are estimated from data in the $e\mu$ channel using events with same-sign lepton charge to enrich the control region in QCD events. The transfer factors OS/SS are the ratio of the yields in the opposite- to the same-sign regions, respectively, and are computed by requiring the presence of an anti-isolated muon. The latter is a muon candidate that fails the standard isolation requirement but must satisfy $0.2 < I_{rel} < 0.5$. The transfer factors can be used in the application region with the standard muon isolation to extrapolate the QCD contribution in the signal region as follows:

$$N(\mathrm{QCD_{OS}}) = \mathrm{N(QCD_{SS})_{data}}\frac{\mathrm{OS}}{\mathrm{SS}}. \tag{6.5}$$

This method implies simply scaling the QCD distribution using a constant extrapolation factor from the same-sign region to an opposite-sign region. However, this approach was found to be suboptimal [114], and a more refined extrapolation factor is needed to address the main limitation, i.e. the assumption that the shape of the QCD distribution of interest does not vary between the signal and control regions. To this end, the extended extrapolation factor $\epsilon_{QCD}$, is introduced. The QCD distribution is obtained in the same-sign control region from data after subtracting all the predicted non-QCD contributions in bins of the relevant observables. The modified version of Equation 6.5 then becomes:

$$N(\mathrm{QCD_{OS}}) = \mathrm{N(QCD_{SS})_{data}} \cdot \epsilon_{\mathrm{QCD}} - \mathrm{N(nonQCD_{SS})_{MC}} \cdot \epsilon_{\mathrm{QCD}}. \tag{6.6}$$

This extended factor is computed as a function of the muon and electron transverse momenta ($p_T$), the number of jets ($N_{jets}$), and the angular separation ($\Delta R$) between the electron and muon. The extended approach defines three additional regions (R1, R2,R3) other than the aforementioned same- and opposite-sign regions, based on the relative isolation of the muon and that of the electron, as summarized in table 6.6.

| Region | Isolation $\mu$ | Isolation $e$ |
|---|---|---|
| Opposite sign $e\mu$ (OS) | $I_{\mathrm{rel}}(\mu) < 0.2$ | $I_{\mathrm{rel}}(e) < 0.15$ |
| Same-sign $e\mu$ (SS) | $I_{\mathrm{rel}}(\mu) < 0.2$ | $I_{\mathrm{rel}}(e) < 0.15$ |
| R1 | $I_{\mathrm{rel}}(\mu) \in [0.2, 0.5)$ | $I_{\mathrm{rel}}(e) < 0.15$ |
| R2 | $I_{\mathrm{rel}}(\mu) \in [0.2, 0.5)$ | $I_{\mathrm{rel}}(e) \in [0.15, 0.5)$ |
| R3 | $I_{\mathrm{rel}}(\mu) < 0.2$ | $I_{\mathrm{rel}}(e) \in [0.15, 0.5)$ |

Table 6.6: Isolation criteria for different regions for QCD estimation in $e\mu$ channel.

As the transfer factors depend on the number of jets, they are calculated as a function of the electron-muon separation in bins of the jet multiplicity in the first region (R1).Three different jet multiplicity bins correspond to the 0-, 1 and $\geq 2$ jet categories, respectively. The $\Delta R$ dependence of the OS/SS scale factors is parametrized with a quadratic polynomial function for each jet category, as shown in Figure 6.9.

Furthermore, two-dimensional weights are derived as a function of the $p_T$ of the electron and the muon to account for the dependence of the QCD scale factors on the lepton kinematics. These weights shown in Figure 6.10 are obtained by applying the scale factors from the previous step to the same sign (SS) events with an anti-isolated muon.

The distribution of $(p_T(\mu), p_T(e))$ bins is compared to the expected QCD distribution, obtained from opposite-sign events in the data and adjusted by removing non-QCD contributions from the simulation. The resulting correction factor denoted as $C[p_T(\mu), p_T(e)]$ reflects the ratio between the expected and observed distributions.

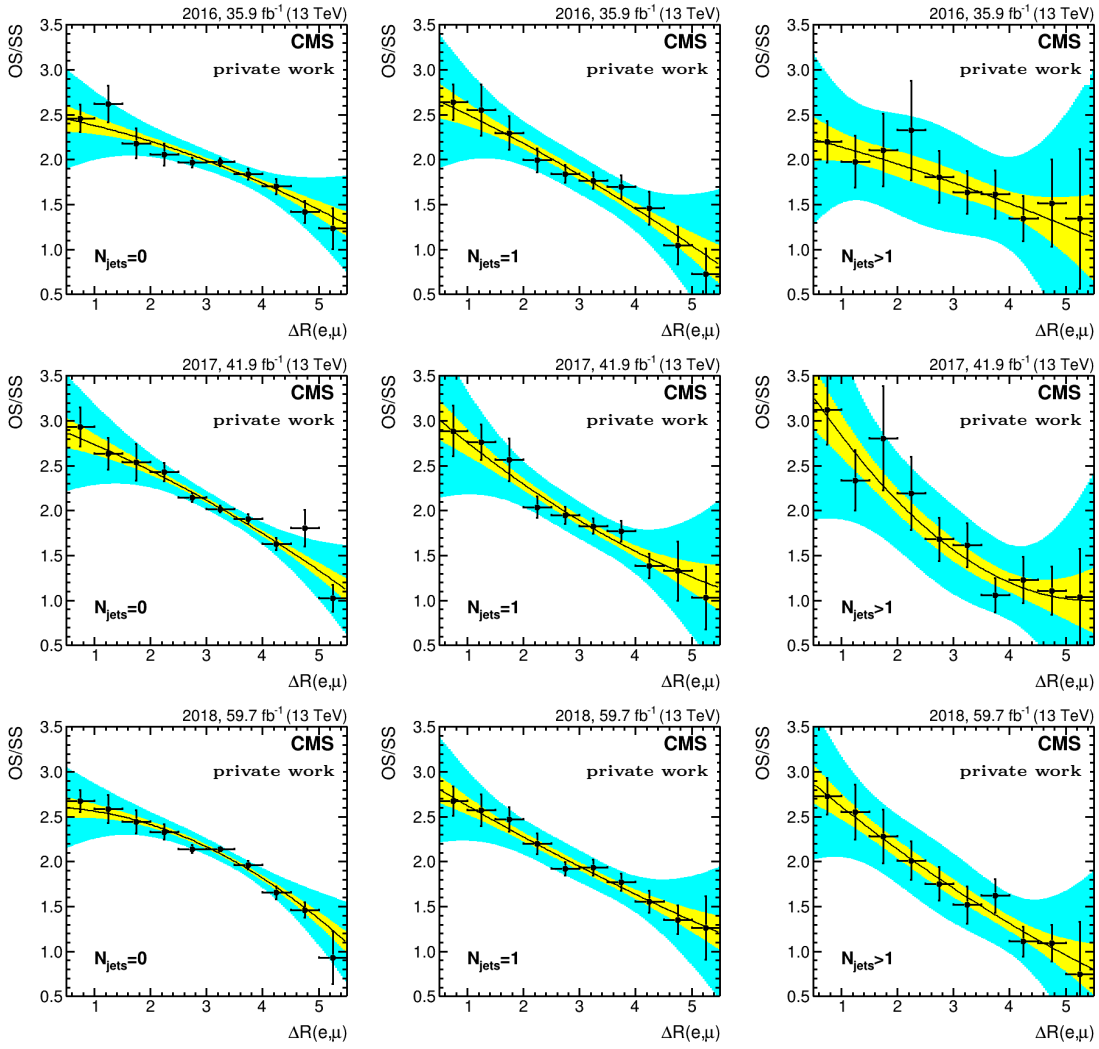


Figure 6.9: Dependencies of the OS/SS scale factors derived in an anti-isolated muon region versus $\Delta R$ between electron and muon are shown for the 0, 1 and $\geq$ 2-jet category for each data-taking year: 2016 (first row), 2017 (second row) and 2018 (last row). The central fit is shown in the black line, and the yellow and green areas correspond to the 68% and 95% error bands, respectively.

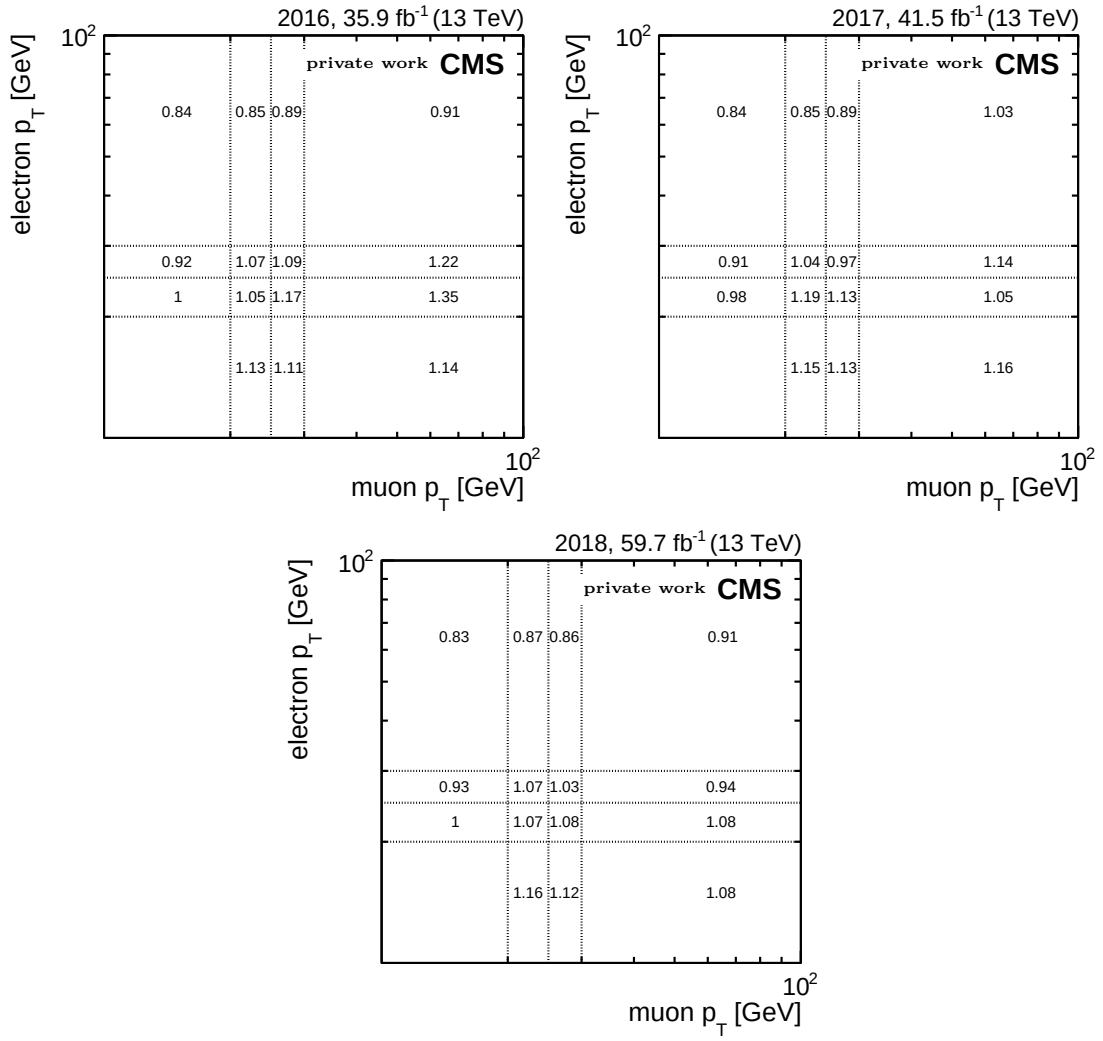


Figure 6.10: Corrections to the QCD OS/SS scale factors were computed in an anti-isolated muon region as a function of the electron and muon $p_T$, using data collected in 2016 (top left), 2017 (top right), and 2018 (bottom).

Moreover, to address potential inaccuracies in the scale factors' dependency on anti-muon isolation, the R2 and R3 regions are introduced. Corrections are implemented from regions with loosely isolated muons, specifically R1 and R2, and extended to those with isolated muons (R3, OS, and SS). These corrections are established using two regions with less stringent electron isolation, R2 and R3. Within each of these regions, QCD extrapolation factors are computed in bins of $(p_T(\mu), p_T(e))$, following the outlined procedure. Figure 6.11 illustrates the derived weights. Subsequently, the correction factor for the extrapolation between isolation regions denoted as $C_{iso}[p_T(\mu), p_T(e)]$, is determined by calculating the ratio of the QCD extrapolation factor in region R3 to that in region R2.

The product of the aforementioned three weights is applied as a correction factor to the QCD extrapolation factor to estimate the QCD multi-jet background process in the signal region. The final extrapolation factor is the following:

$$\epsilon_{QCD} = \epsilon_{QCD}[N_{jets}, \Delta R(e,\mu)] \cdot C[p_T(\mu), p_T(e)] \cdot C_{iso}[p_T(\mu), p_T(e)]. \tag{6.7}$$

**Closure tests**

Closure tests were conducted after determining the OS/SS extrapolation factors by comparing the distributions of the most representative variables in the OS region with the extrapolated distributions from the SS sideband region. The comparison is done in the OS/SS determination region. The results are provided for several variables in Figures 6.12 for the inclusive $e\mu$ event selection, i.e. without requiring b-tagged jets.

To ensure consistency between the distributions extrapolated from the same-sign (SS) region and those in the opposite-sign (OS) region, an additional inclusive scale factor ($r_{b-tag}$) is necessary for events including one or more b-tagged jets. This can be inspected in Figure 6.13, which displays the distribution of the total transverse mass of electrons, muons, and missing transverse momentum ($m_T^{tot}$) in both opposite-sign (OS) and same-sign (SS) samples for events that have at least one jet classified as a b-jet.

The $r_{b-tag}$ factors for each year are initially estimated by fitting the ratio of the total transverse mass $m_T^{tot}$ distribution obtained in the OS region using a constant to the one extrapolated from the SS sideband.

Furthermore, to address a potential bias in $r_{b-tag}$ arising from the requirement of a non-isolated muon in the OS/SS determination region, comparisons are made for values of $r_{b-tag}$ in two control regions:

1. Isolated muon ($Iso_\mu < 0.2$) and non-isolated electron ($0.15 < Iso_e < 0.5$). In this region, $r_{b-tag}$ is systematically lower, e.g., 2016 ($0.72/0.76 = 0.95$).

2. Non-isolated muon ($0.2 < Iso_\mu < 0.5$) and non-isolated electron ($0.15 < Iso_e < 0.5$).

The final $r_{b-tag}$ estimates are obtained by inserting these correction factors with systematic uncertainties treated as correlated between datasets, shown in the table below:

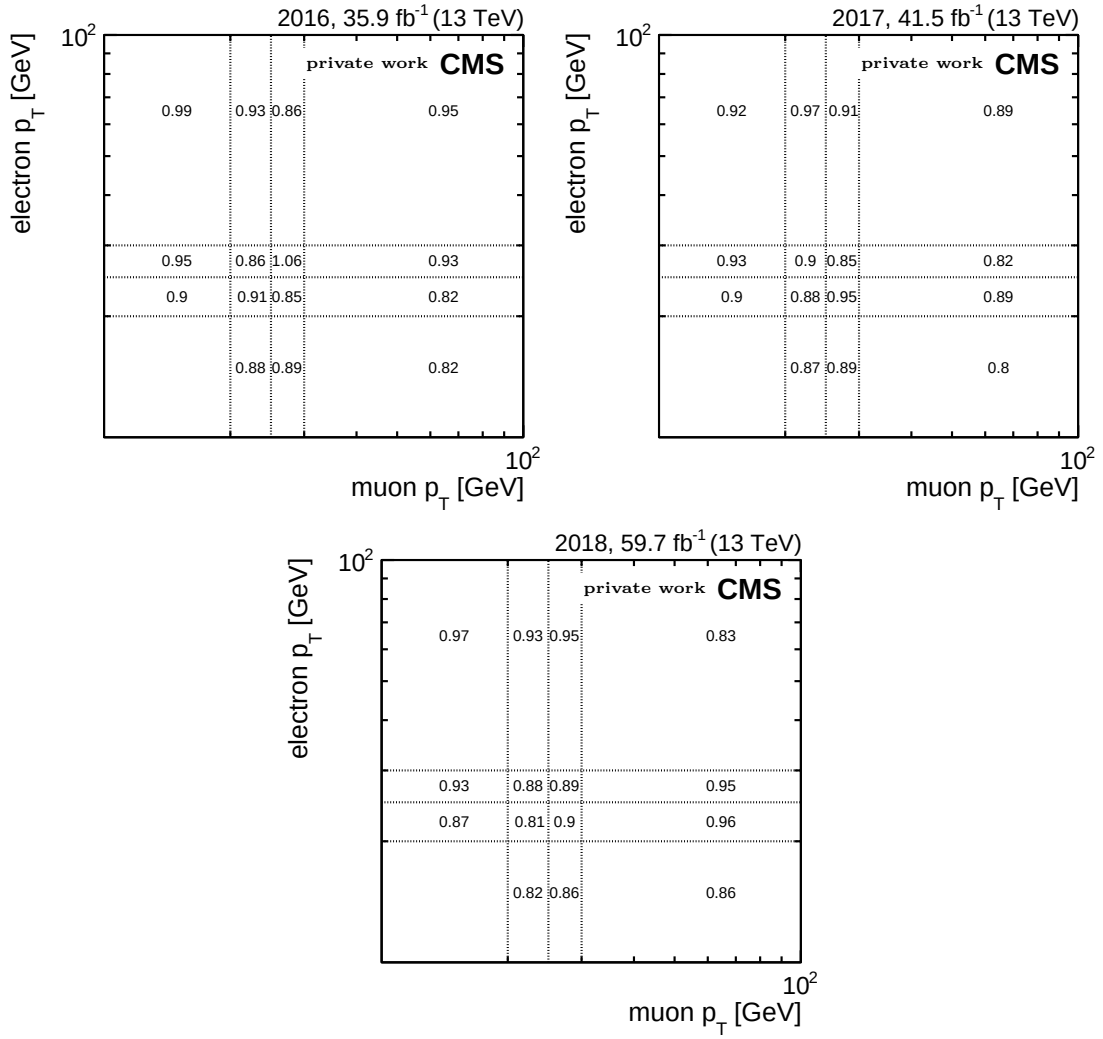


Figure 6.11: Corrections to the QCD OS/SS scale factors address the inaccuracies introduced by anti-isolating muon to measure the OS/SS scale factors. This is done using data collected in 2016 (top left), 2017 (top right), and 2018(bottom)

.

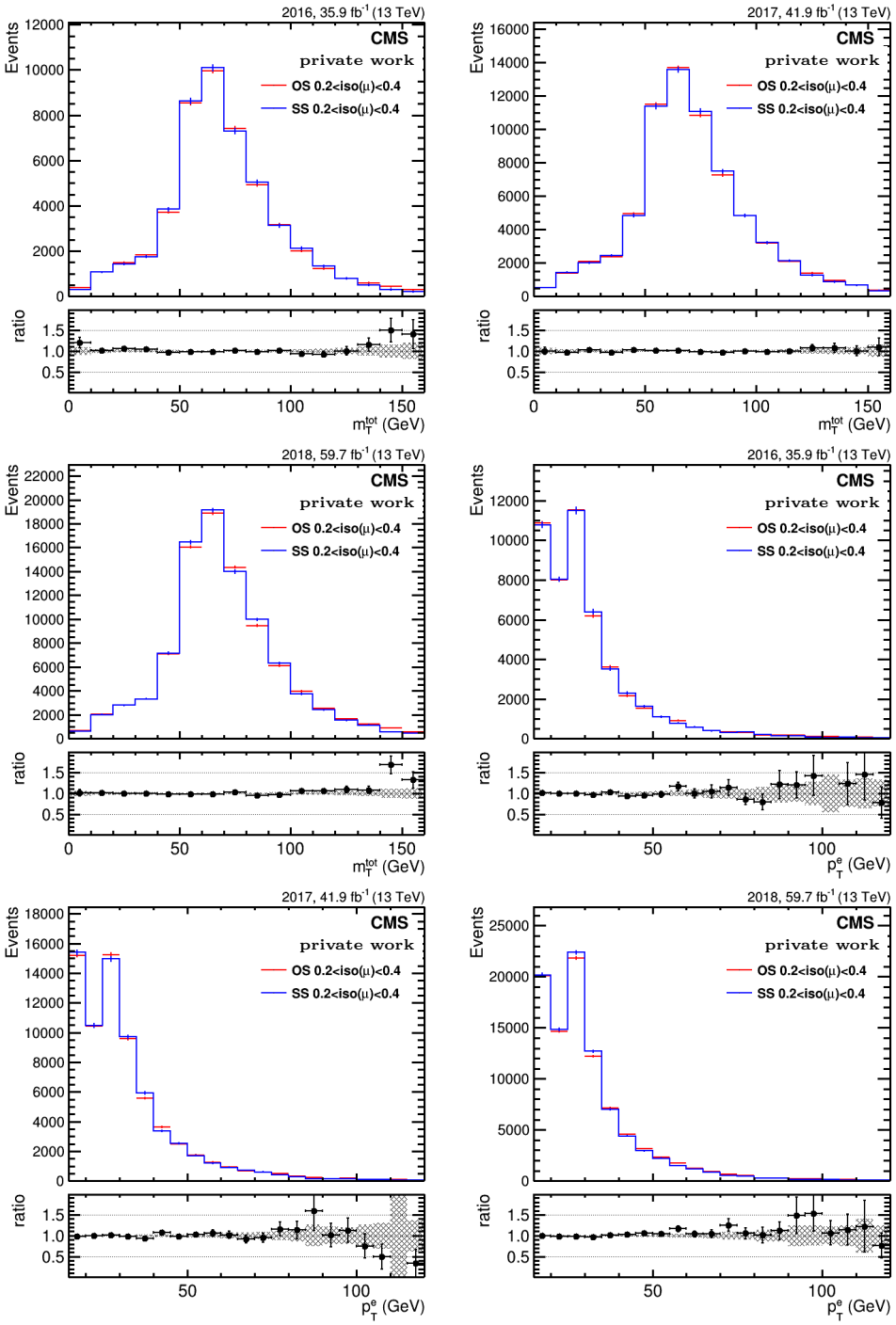


Figure 6.12: Comparison of the distributions for Opposite Sign electron and muon (OS) to the distributions derived from the Same Sign (SS) region. The distribution of $m_T^{tot}$ and $p_T^e$ are shown as an example for the three different data taking years. Generally, a good closure is observed for the inclusive selection in the $e\mu$ channel.

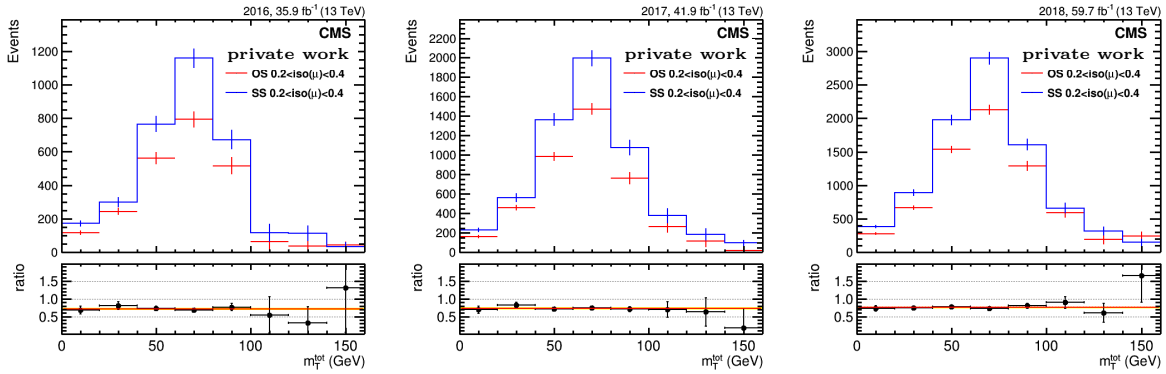


Figure 6.13: Figures show the $m_T^{tot}$ distribution comparing a sample of opposite-sign (OS) electron-muon pairs with the distribution extrapolated from the same-sign (SS) sideband region. These distributions are acquired within the OS/SS determination region, specifically for events featuring one or more b-tagged jets. The lower panels display the outcomes of fitting the OS/SS ratio distributions with a constant.

| Dataset | $r_{b-tag} \pm$ (stat.) $\pm$ (sys.) |
|---|---|
| 2016 | $0.67 \pm 0.04 \pm 0.05$ |
| 2017 | $0.69 \pm 0.04 \pm 0.05$ |
| 2018 | $0.71 \pm 0.03 \pm 0.05$ |

Table 6.7: Correction factor $r_{b-tag}$ for each data-taking year.

The scale factors above allow predicting the QCD multijet background in the b-tag categories of the $e\mu$ channel. To verify the need for an additional extrapolation factor of 0.7 in the b-tag category, the primary mechanisms behind OS and SS electron-muon pairs in QCD multijet events should be considered. In the no-bag category, electron-muon pairs mainly come from heavy flavour di-jets ($pp \to b\bar{b}/c\bar{c}$). Opposite-sign lepton pairs arise when leptonic decay of $B(D)$-mesons occure in b(c)-jets, and the leptonic decay of $\bar{B}(\bar{D})$-mesons decay occure in $\bar{b}(\bar{c})$-jets. Same-sign lepton pairs result from the decay of $B(\bar{B})$-mesons in one b-jet and $\bar{B}(B)$-mesons into $\bar{D}(D)$-mesons in the other b-jet. Processes with heavy flavour quarkonium decay yield OS and SS lepton pairs equally. Studies show that when vetoing additional b-jets, OS lepton-pair events are about twice as frequent as SS lepton-pair events: $(OS/SS)_{no-btag} \approx 2$.

In the events where at least one identified b-jet accompanies an electron-muon pair, the significant contribution comes from triple-jet production ($pp \to b\bar{b}g$). Opposite- and same-sign electron-muon pairs are equally produced in this case, leading to an average OS/SS value close to unity: $(OS/SS)_{btag} \approx 2 \times 0.7 = 1.4$.

After applying the $r_{b-tag}$ scale factors, a closure test is conducted. The closure test

compares the distributions of selected key variables in the OS region with those from the SS region. This comparison is performed for events with at least one b-tagged jet, as shown in Figure 6.14.

### 6.5.2 Non-QCD background with jets misidentified as prompt leptons

Non-QCD background with non-prompt leptons, where one or two leptons are mimicked by hadronic jets, constitutes approximately 4-7% of the total number of events selected in the final sample in the electron-muon ($e\mu$) channel. This background primarily comes from $t\bar{t}$, single-top, W+jets production, and Z+jets and di-boson processes with smaller contributions. It is estimated using Monte Carlo (MC) simulation containing scale factors to adjust for differences in misidentification rates observed in data versus simulated events.

Misidentification rates are separately measured for light-flavour and heavy-flavour jets. The light-flavour and gluon jets mistakenly selected as prompt electron or muon samples of Z+1jet events with the Z boson decaying into a muon pair are used. The jet recoiling against the Z boson is considered a probe. It is employed to assess the probabilities of misidentification for jets as electrons (jet $\rightarrow$ e) and muons (jet $\rightarrow \mu$) in both data and MC. No jet PF ID is applied to the probed jets to ensure an unbiased measurement, allowing for the genuine evaluation of (jet $\rightarrow$ e) and (jet $\rightarrow \mu$) fake rates. These rates are calculated as the ratio of the number of passing jets to the sum of the number of passing and failing jets as follows:

$$F(jet \rightarrow l) = \frac{N_p(jet)}{N_p(jet) + N_f(jet)}, \tag{6.8}$$

where passing jets ($N_p$) are those matching reconstructed leptons within specified criteria of the nominal identification and isolation within a cone of $\Delta R < 0.3$. In contrast, failing jets ($N_f$) do not match the reconstructed leptons. The number of passing (failing) probes is determined by subtracting contributions of genuine prompt leptons and heavy flavour jets observed in data from the number estimated with MC simulation. The purity of the sample of light-flavoured jets in the numerator (denominator) of the ratio was found to lie between $60 - 85\%$(93-97%), with the lowest purity observed for harder jet $p_T$ spectra. The misidentification probabilities and the corresponding data-simulation scale factors are measured for the Run 2 eras as a function of jet $p_T$ and fitted with a linear function, shown in Figure 6.15.

Scale factors are applied as jet $p_T$-dependent weights to simulated event samples, where the reconstructed lepton matches the generator jet from light-flavour quarks or gluons. Uncertainties in the fit function, dominated by statistical uncertainties related to data and simulation sizes, are considered uncertainties in the udsg-jet misidentification probabilities, de-correlated between data-taking periods.

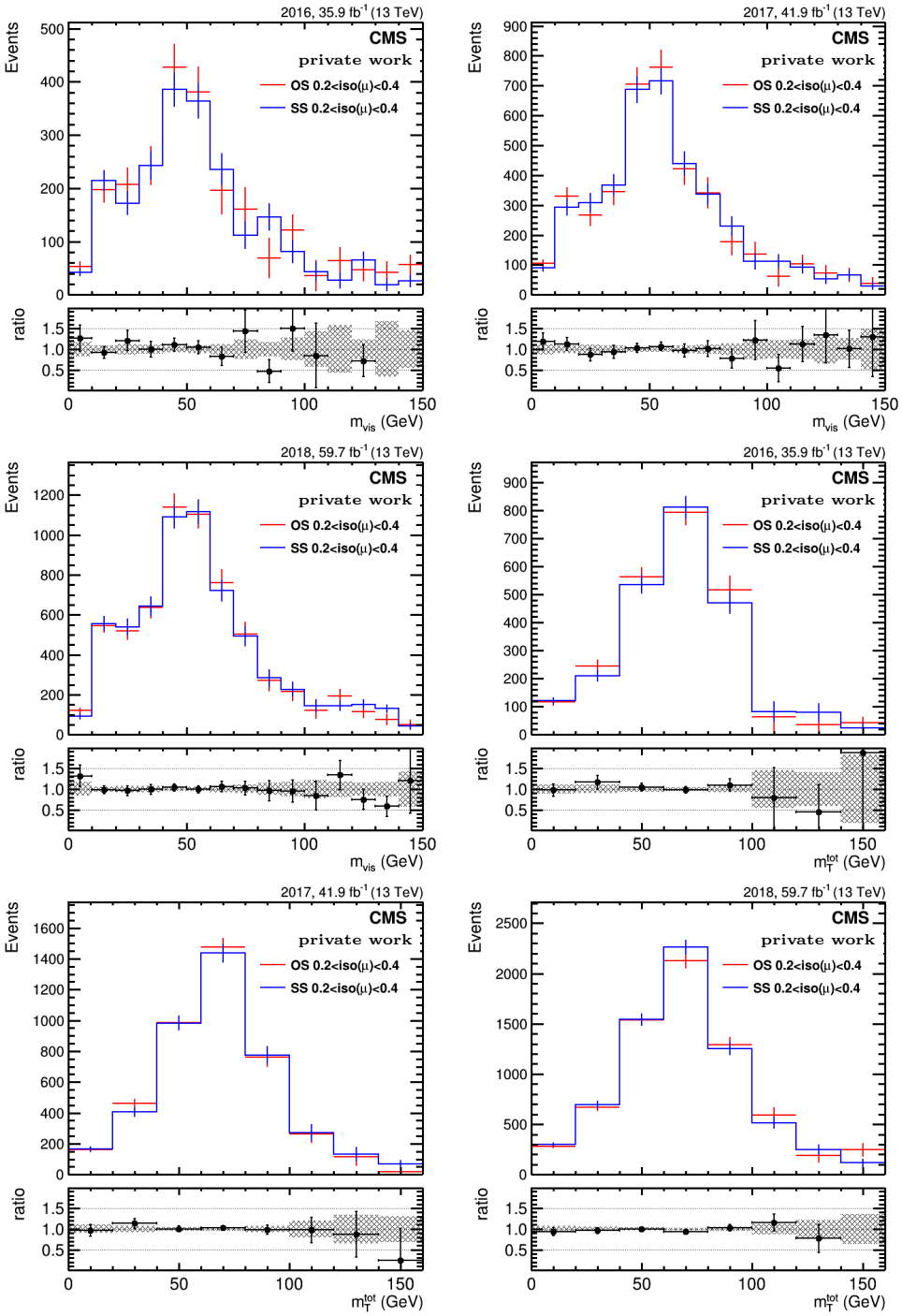


Figure 6.14: Shown are the distributions within the sample of opposite-sign (OS) electron-muon pairs involving comparing them with distributions extrapolated from the same-sign (SS) sideband region. These distributions are acquired within the OS/SS determination region for events featuring one or more b-tagged jets.

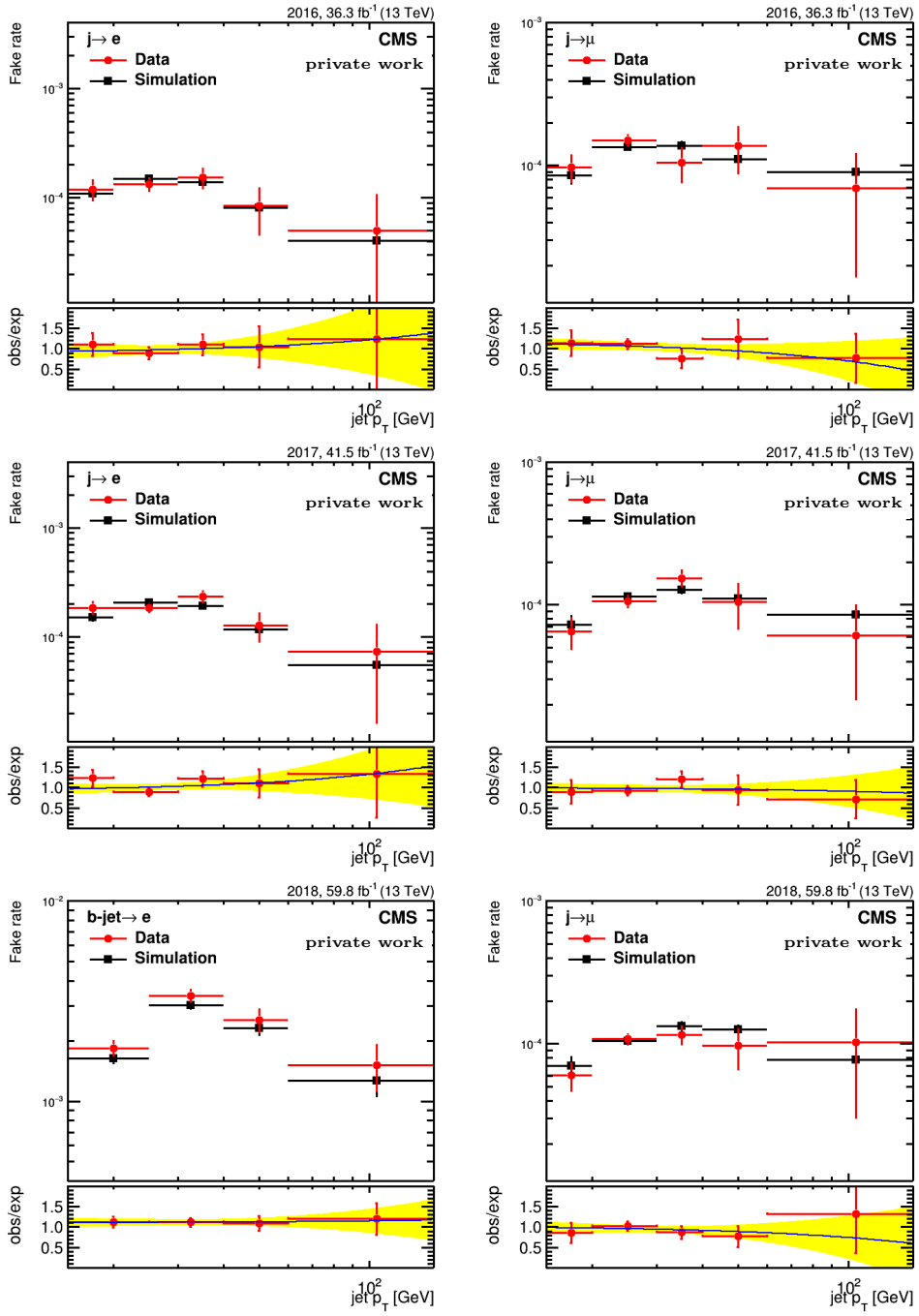


Figure 6.15: Misidentification probabilities for light-flavour or gluon jets to be misidentified as prompt electrons (left plots) or prompt muons (right plots) are presented. The upper plots correspond to the 2016 data-taking period, the middle plots to the 2017 data-taking period, and the lower plots to the 2018 data-taking period.These probabilities are measured in the sample of $Z \to \mu\mu + 1$ jet events. In the lower panels of the plots, data-simulation scale factors for the misidentification probabilities of light-flavour or gluon jets as electrons (udsg $\to$ jet $\to$ e) and muons (udsg $\to$ jet $\to \mu$) are displayed, along with a fitted linear function. The yellow shaded area indicates the 68% uncertainty band in the fitted function.

Similar measurements are conducted for b-jets, using leptonic $t\bar{t}$ decays as the sample. The purity in the sample of probed b-jets ranges between 75-90% (90-95%), with lower purity at higher jet $p_T$ values. The dependencies of the scale factors on jet $p_T$ are fitted with a linear function and are shown in Figure 6.16. The uncertainty of the fitted function is considered a systematic uncertainty. Furthermore, the precision of this evaluation is affected by statistical uncertainties, which are connected to the sample size of the data and MC. Therefore, the uncertainties in the b-jet $\rightarrow$ e and b-jet $\rightarrow \mu$ misidentification probabilities are de-correlated between data-taking periods. The scale factors and related uncertainties are then applied to simulated events involving reconstructed prompt electrons (muons) matching generator jets from the fragmentation of bottom and charm quarks, with the uncertainty for c-jets doubled in size due to the absence of a dedicated measurement.

### 6.5.3 Background from muons misidentified as electron

Backgrounds in which muons are misidentified as electrons arise predominantly from Z bosons with subsequent $Z \rightarrow \mu^+\mu^-$ decay, in which one of the muons mimics a prompt electron. The scale factors for the probability of muon-to-electron misidentification ($\mu \rightarrow e$) in data and simulation are determined using events where electrons fake muons. This is conducted on a specific event sample with an invariant mass of the electron-muon pair compatible with the nominal mass of the Z boson, $80 < m_{e\mu} < 100$ GeV. Furthermore, additional jets are excluded to enhance the purity of Drell-Yan+jets events involving $Z \rightarrow \mu^+\mu^-$ decay. The scale factor for the muon-to-electron misidentification is calculated as:

$$f(\mu \rightarrow e) = \frac{N_{Data}(e\mu) - N_{MC}(e_{gen}, jet_{gen})}{N_{MC}(\mu \rightarrow e)}, \tag{6.9}$$

where $N_{Data}(e\mu)$ is the observed $e\mu$ events in data, $N_{MC}(e_{gen}, jet_{gen})$ is the subtracted contribution from $e\mu$ events where the reconstructed prompt electron matches a genuine prompt electron or a genuine hadronic jet, estimated from Monte Carlo (MC), and $N_{MC}(\mu \rightarrow e)$ the number of $e\mu$ events in the DY+jets MC sample where a prompt muon mimics a prompt electron.

The results in Table 6.8 involve applying these scale factors as weights to simulated events where a prompt-isolated muon is mistakenly identified as a prompt-isolated electron. The scale factors are determined independently for the 2016, 2017, and 2018 datasets, with precision ranging from 14% (2018 dataset) to 19% (2016 dataset). The statistical dominance of the overall uncertainty in the measurements ensures decorrelated uncertainties between data-taking periods. Simulation studies indicate a negligible impact of electron-to-muon misidentification on backgrounds in the $e\mu$ channel.

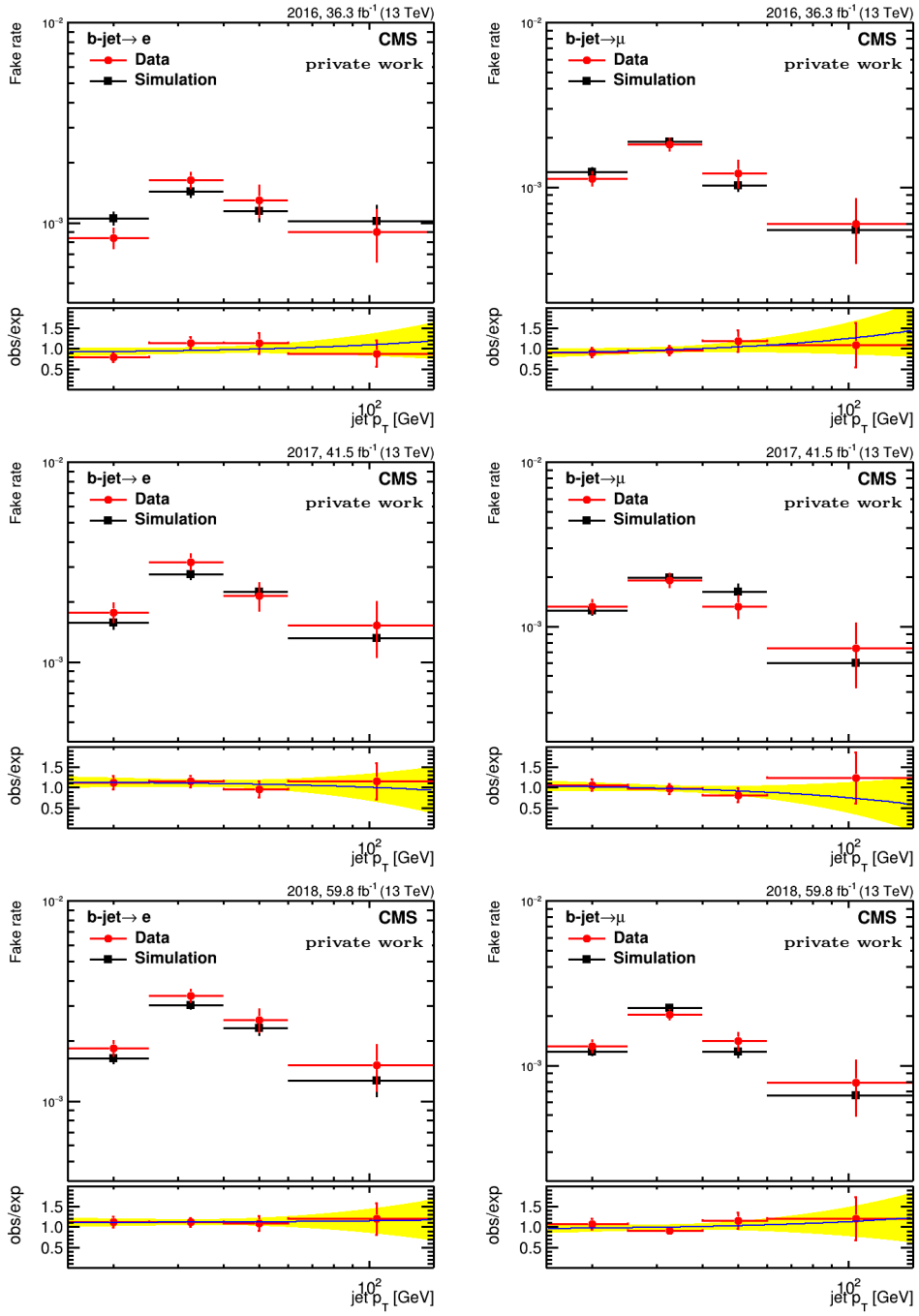


Figure 6.16: The misidentification probabilities for b-jets to be misidentified as electrons (left plots) or muons (right plots) are illustrated. The upper plots represent the 2016 data-taking period, the middle plots correspond to the 2017 data-taking period, and the lower plots depict the 2018 data-taking period. These probabilities are determined in the sample of $t\bar{t}$ + 2 jets events with electrons and muons in the final state. In the lower panels of the plots, data-simulation scale factors for the misidentification probabilities of b-jets as electrons (bjet → e) and muons (bjet → m) are depicted, along with a fitted linear function. The yellow shaded area denotes the 68% uncertainty band in the fitted function.

| Dataset | $\mu \to e$ misidentification probability |
|---|---|
| 2016 | $1.30 \pm 0.25$ |
| 2017 | $1.58 \pm 0.23$ |
| 2018 | $1.54 \pm 0.22$ |

Table 6.8: Scale factor measured for each data taking year for $\mu \to e$ misidentification probability.

## 6.6 Event classification

Classification tasks are introduced for all the channels of this analysis to improve the signal-to-background ratio. The chosen algorithm to perform the classification is the Gradient Boosted Decision Tree implemented with the dedicated libraries, namely XGBoost [98] and LightGBM [115] depending on the channel[2]. A separate model was trained and evaluated for each channel and data-taking year. This section will discuss the event classification specifically for the $e\mu$ channel.

The classification has been performed to distinguish signal classes, respectively $b\bar{b}H \to W^+W^-$, $b\bar{b}H \to \tau^+\tau^-$, from the dominant background processes combined in one class ($t\bar{t}$, single top production) as well as DrelL-Yan+jets (DY+jets) for the $e\mu$ channel. There are two signal categories in this channel, but the sensitivity is dominated by the $b\bar{b}H \to W^+W^-$ channel. The main challenge is to suppress the overwhelming $t\bar{t}$ background. To this end, boosted decision trees are employed to increase the separation power between this background and the signal events.

The Monte Carlo samples are generated separately for each data-taking period; hence, different training was carried out for 2016 pre- and post-VFP, 2017 and 2018 samples, respectively. The simulated samples are used for both the training and the evaluation of the models, but due to their limited statistics, a *two-fold training* has been used. The latter is commonly adopted to prevent overfitting and infer physically accurate results from the training step. In this approach, the training of the model is performed on a subset of data, while the evaluation is done using a disjoint subset not used during the training. The data partitioning is performed based on the event number such that the events assigned an even index are used to check the predictions of a model trained exclusively on the events with an odd index and vice versa. Additionally, the subset of the data used for the model training is further divided into training, testing and validation subsets. Exactly 30% of the input data set is used as the test set, while 20% is used as a validation set. Having different subsets of the input data is essential to evaluate the model correctly and to avoid potential overfitting.

Furthermore, a tuning of the predefined settings of the model established before the

[2]The algorithm used in the $e\mu$ channel is XGBoost and for $\tau_h\tau_h$, $\mu\tau$ and $e\tau$ LightGBM is used.

training, or *hyperparameters*, is performed to guide the optimization via the so-called Grid search within sklearn library [116]. The details of the model hyperparameter are summarized in table 6.9, and their exact definition can be found in the XGBoost official documentation [98].

Table 6.9: XGBoost hyperparameters after the fine-tuning of the model

| **Parameter** | **Value** |
|---|---|
| learning_rate | 0.1 |
| n_estimators | 150 |
| objective | multi:softprob |
| max_depth | 3 |
| min_child_weight | 2 |
| max_delta_step | 6 |
| subsample | 0.8 |

A set of balanced weights is used so that the signal and background classes contribute equally to the training. This approach was selected after testing several combinations of class weights and was shown to deliver the best result. The training performance is assessed by means of the confusion matrix and ROC curve, which is shown for the 2018 data in Figure 6.17.

The input feature used to train the model is preprocessed with the standardization technique, which involves adjusting the input data for each feature, ensuring that the mean of observed values is zero and that the corresponding standard deviation is set to unity. The standardization of the input data before training is less important for XGBoost compared to Neural Network models owing to the inherently different model architecture. This standardization process before NN training helps maintain consistency in the neural network's learning process across various input features. It reduces the likelihood of misinterpreting statistical fluctuations as meaningful discriminating features. Despite the fact that models like XGBoost are less sensitive to feature scaling than NNs, standardizing the input before the training can still provide several benefits, such as improved convergence speed, interpretability and regularization of the model.

The kinematic information of the events is used as input features for the training task, which is slightly different among different channels and includes high and low-level features. The list of the input features used for the training in the $e\mu$ channel is given in the table 6.10.

The selection of the input variables for the training is based on the discrimination power of the variable between the signal and background process and also the quality of the variable modelling. Only well-modelled variables have been chosen as input to the BDT, which implies a good level of data-to-simulation agreement in the shape distributions and a quality criterion based on the goodness-of-fit (GoF) test. The input variables with the

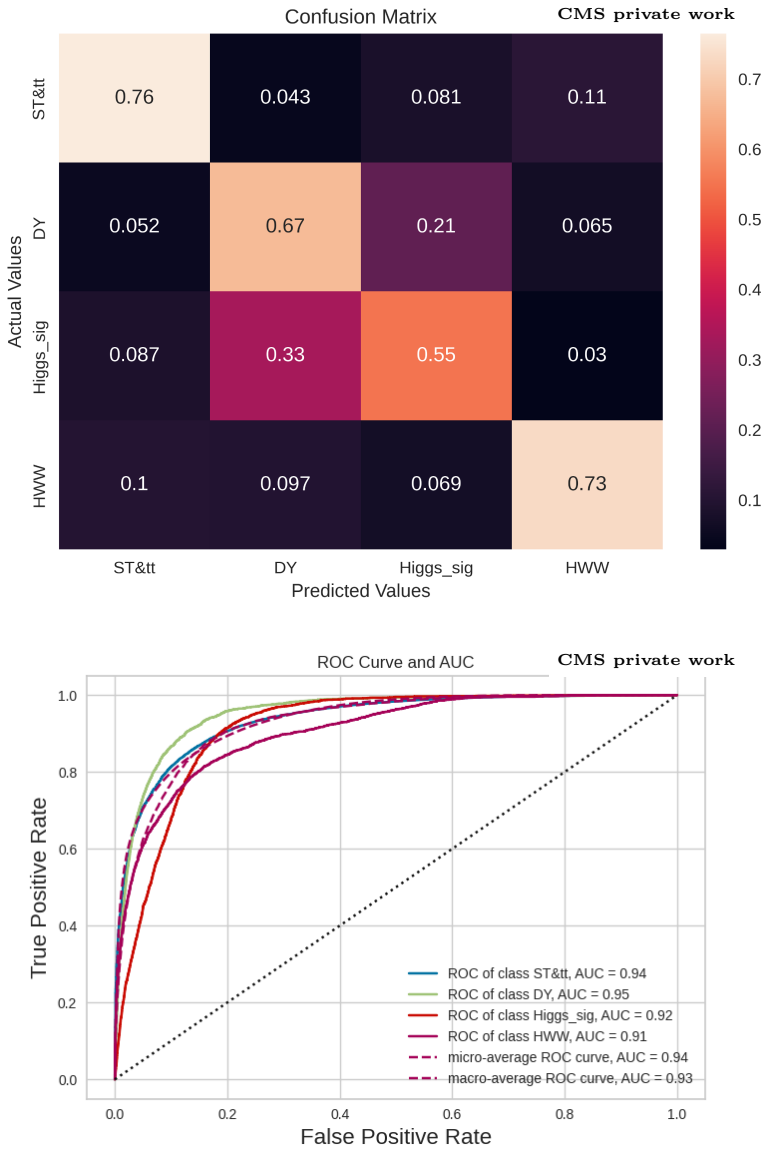


Figure 6.17: The confusion matrix (top) for the training of the 2018 data set. The multi-classification involves four classes: $t\bar{t}$ and single top (combined in one class) labelled as ST & TT, DY+jets class labelled as DY, $b\bar{b}H \to \tau^+\tau^-$ class labelled as Higgs signal, and $b\bar{b}H \to W^+W^-$ labelled as HWW (additional signal class). The matrix shows the training outcome by displaying the actual value versus the predicted class; large coefficients on the diagonal axis point to an optimal separation power. The ROC curve (bottom) and the corresponding value of the area under the curve (AUC) are shown for the aforementioned classes. The largest AUC value indicates the maximal separation power.

| Input Variables used in the $e\mu$ channel |
|---|
| Visible di-$\tau$ mass |
| $D_\zeta$ |
| Total transverse mass |
| Di-$\tau$ $p_T$ |
| Electron $p_T$ |
| Muon $p_T$ |
| Number of b-jets |
| $p_T$ of leading b-jet |
| Number of jets |
| $p_T$ of leading jet |
| $p_T$ of trailing jet |
| Di-jet $\Delta\eta$ |

Table 6.10: Input features used for the classification task in the $e\mu$ channel

GoF test lower than 0.05 are excluded since, in this case, the agreement between the data and simulation is considered inadequate. Furthermore, different set of variables was tested for the BDT training. The SHAP method introduced in section 5.2 was used to monitor the importance of those features and their effect on the classification task. The most effective set of input features for the training was chosen, and it is listed in table 6.10. The ranking of the input features after the training calculated with the SHAP method is shown in Figure 6.18.

The common selection explained in section 6.4.1 is applied before the training of BDT. Additionally, since a slight disagreement is observed in the distribution of the $D_\zeta$ variable for $D_\zeta > 20$ GeV, this region which was not well modelled is excluded from the analysis of all data sets. Events are selected if they contain 1 or 2 b-tagged jets $1 \leq N_{btag} \leq 2$. Since $t\bar{t}$ background events dominate the final selected samples, additional selections are $10 < m_{vis} < 100$ GeV and $m_T^{tot} < 200$ GeV are applied to suppress this background further. The distributions of various BDT input variables after requiring the $1 \leq N_{btag} \leq 2$ selection is shown in Figure 6.19 and 6.20 for 2018 data. The additional figures for 2016 and 2017 can be found in the Appendix.

The trained models are then saved in a JSON file format to produce predictions on an unseen dataset. The probability score and the predicted class should be stored for all the events and the systematics tree in each sample. Therefore, the final format is a root file which contains the probability scores for each event that can be used to produce the result based on a likelihood fit.

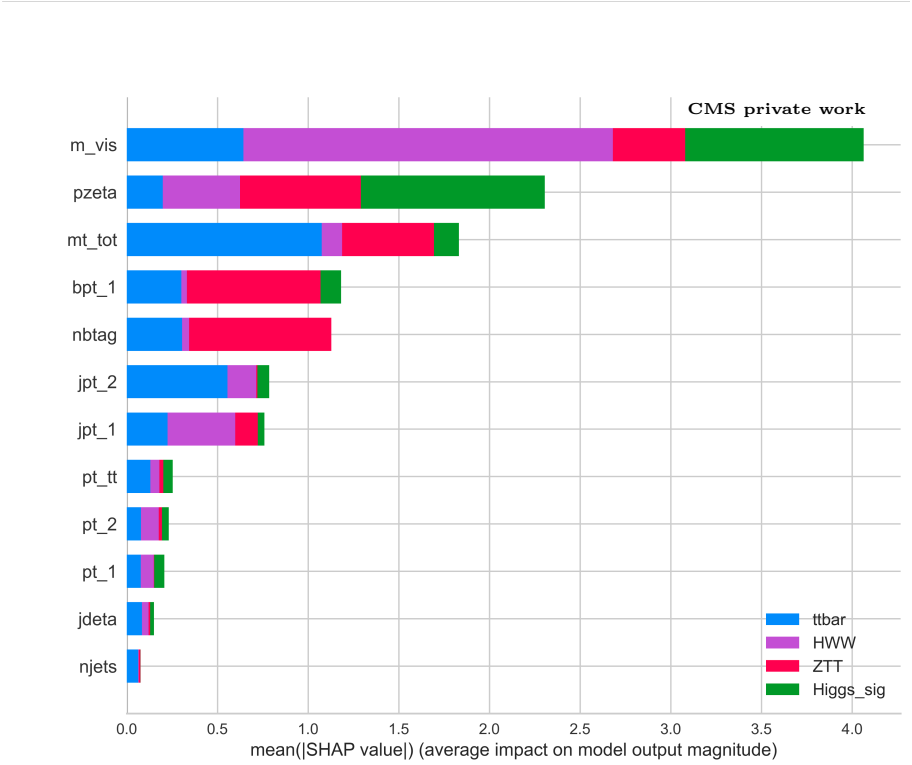


Figure 6.18: Input features ranked in decreasing order of their impact on the model output magnitude using the corresponding SHAP value, with the separation power of each class versus the rest highlighted with a different colour for each feature.

### 6.6.1 Further optimization of the BDT

To further optimize the BDT, training based on the hierarchical classification approach has been designed in addition to the classification result presented in the previous section.

This approach was chosen to improve the signal-to-background ratio by suppressing further the $t\bar{t}$ background and to study in detail the effect of separation of signal samples corresponding to amplitudes of $y_b$ and $y_t$ on the result. The hierarchy definition between the classes was tested for several combinations from three- to two-level hierarchy. The most optimal result is achieved by defining the two-level hierarchy. The first level of the hierarchy with a dedicated model is defined as a binary classifier to separate all the signal events in the $e\mu$ channel ($b\bar{b}H \rightarrow \tau\tau$ and $b\bar{b}H \rightarrow W^+W^-$ combination) from all the background events ($t\bar{t}$, single top and DY+jets). The output of the model from this level is then used as an additional input feature for the training of the model in the second level of the hierarchy. The latter then categorizes all the subclasses from one another, separating the signal events corresponding to the bottom Yukawa coupling ($y_b$) and top

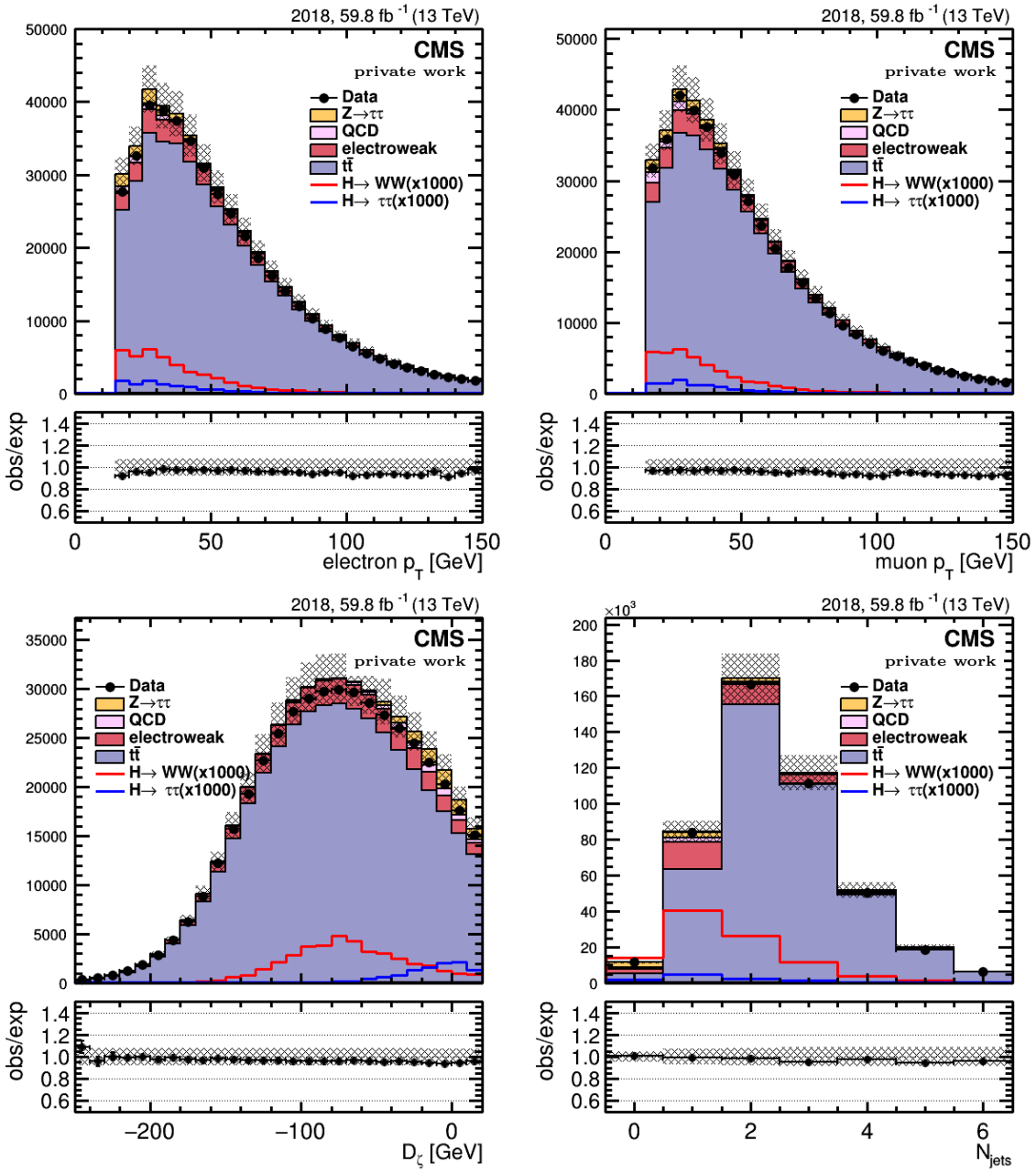


Figure 6.19: Data and Monte Carlo distributions in the $e\mu$ channel for 2018 data set after requirring ($1 \leq N_{btag} \leq 2$). Upper-left: $p_T$ of electron , Upper-right: $p_T$ of muon, Lower-left: $D_\zeta$ , Lower-right: number of jet multiplicity.

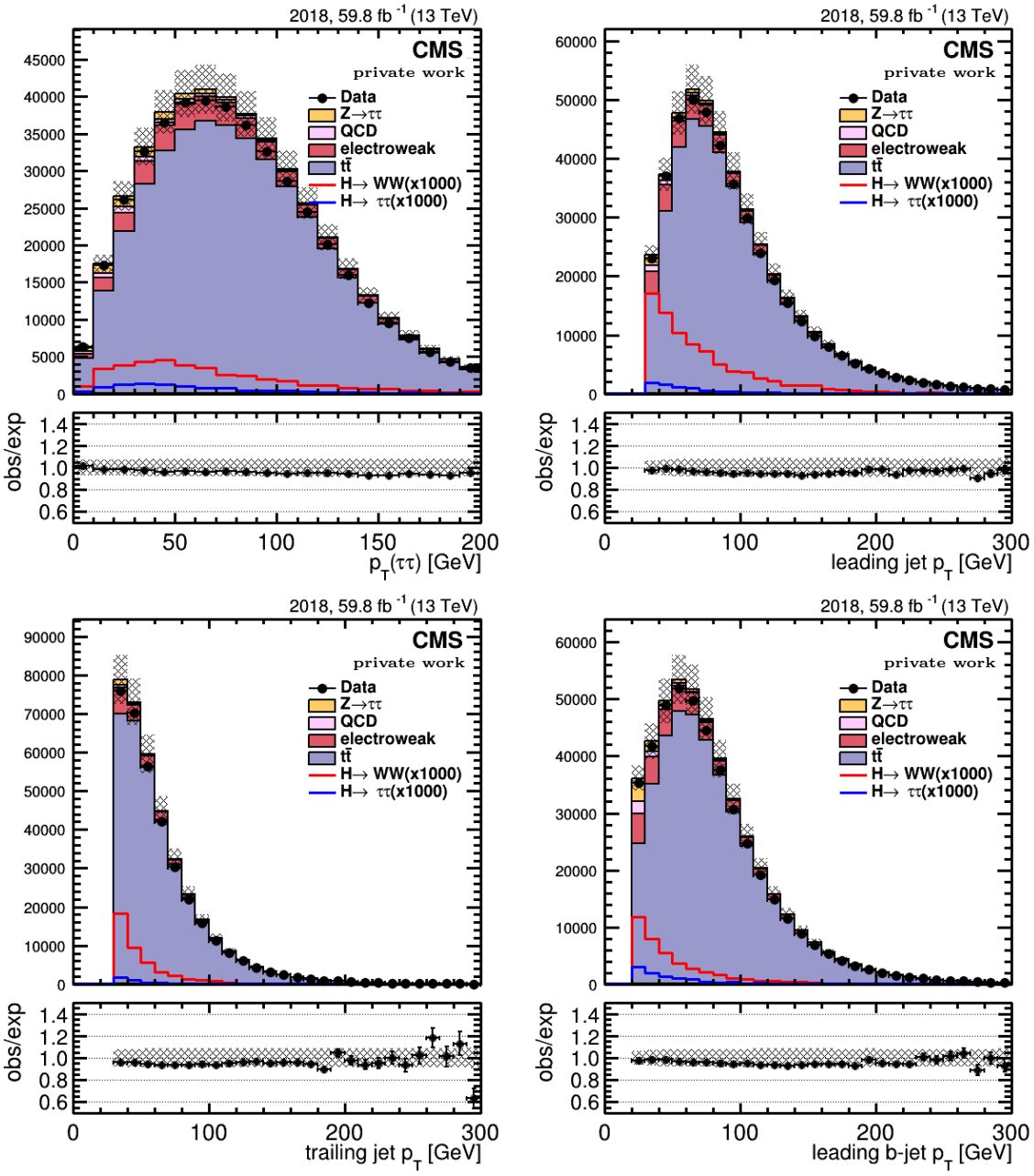


Figure 6.20: Data and Monte Carlo distributions in the $e\mu$ channel for 2018 data set after requiring ($1 \leq N_{btag} \leq 2$). Upper-left: $p_T$ of Higgs candidate, Upper-right: leading jet $p_T$, Lower-left: trailing jet $p_T$, Lower-right: leading b-tagged jet $p_T$ distribution of pseudorapidity between two leading jets.

Yukawa coupling ($y_t$), respectively. The hierarchy structure illustrated in Figure 6.21 uses the same techniques and input features explained in the previous section.

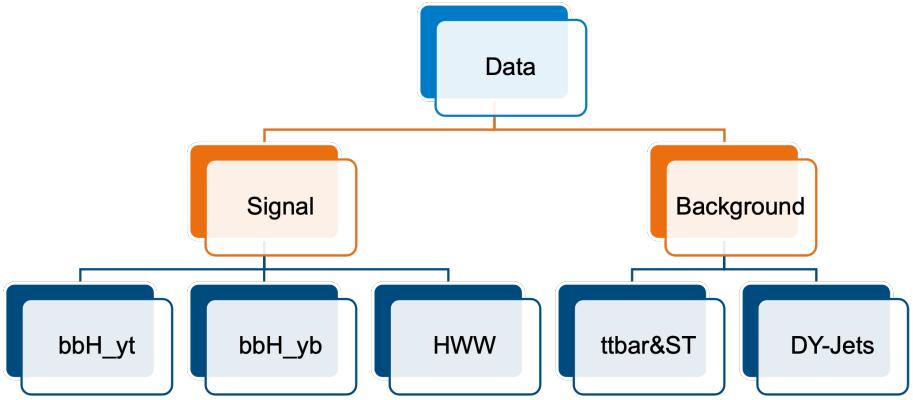


Figure 6.21: Figure showing the hierarchy levels: The first level has a dedicated model to separate signal from background events, and the second level uses a dedicated model to separate all the subclasses from each other

The results of the training for both levels are shown in Figure 6.22. The overall result compared to the flat classification approach improves the separation power between the $t\bar{t}\&ST$ class and the $b\bar{b}H \rightarrow W^+W^-$ class particularly. The best result is also achieved by combining the $t\bar{t}$ and single top into one class. Also, the $b\bar{b}Hy_b$ and $b\bar{b}Hy_t$ classes are combined in this confusion matrix to draw a better comparison with the flat classification case. Further discussion of the hierarchy approach and the corresponding results obtained are discussed in Appendix B, since the result obtained in combination with other channels and presented in the present work are derived with the initial flat training results.

## 6.7 Event categorization

The trained BDT models produce a set of output scores for each event in which the number of scores corresponds to the output classes. Hence, in the case of the $e\mu$ channel, four output scores are produced. Furthermore, an event is assigned to a class with the highest BDT score. To enable a probabilistic interpretation of the results, the output scores are

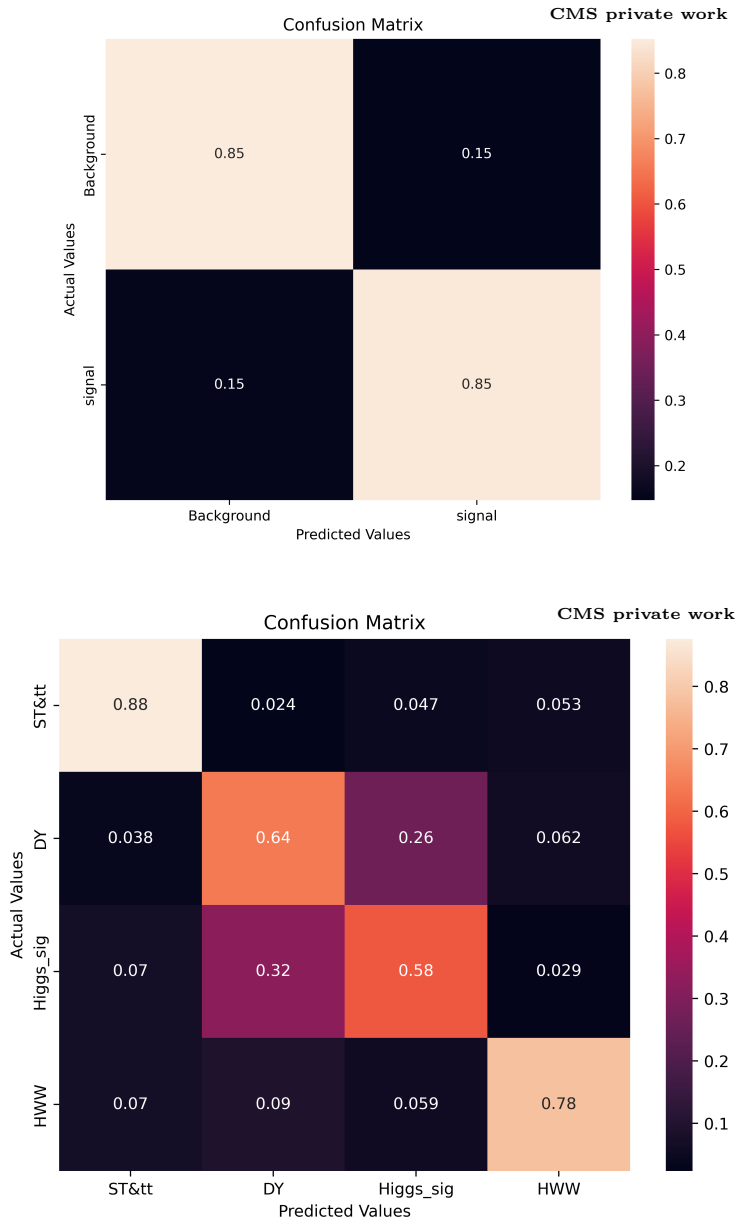


Figure 6.22: Confusion matrix for the two hierarchy levels are shown. Left confusion matrix of the first level separating all background from signal events. The right confusion matrix shows the results of the second level. for a better comparison with the flat case, the two subclasses single-top and $t\bar{t}$ production and $b\bar{b}H_{yb}$ and $b\bar{b}H_{yt}$ are combined.

normalized such that their sum equals one. Each output score represents the probability of the event arising from a specific physical process connected to the corresponding class. Subsequently, events are ordered based on the highest output score. The extraction of the cross-section limits on the bbH process involves a combined fit across all BDT distributions, which will be discussed in the next chapter. This fitting process uses BDT score distributions for both signal and background categories. The distribution of the BDT scores for signal and background classes in the $e\mu$ channel is shown in Figure 6.23. The binning of these distributions is optimized to ensure the maximal sensitivity in terms of the expected upper limit on the signal strength and in such a way that the statistical uncertainties in each bin of the combined backgrounds are lower than 20%.

## 6.8 Systematics uncertainties

The uncertainty model encompasses both theoretical and experimental uncertainties, as well as the statistical precision associated with simulated events. Some uncertainties are common among various channels, while others are channel-specific. Theoretical uncertainties include parameters affecting the production cross-section, such as the strong coupling constant, higher-order QCD and EW corrections, and variations in the renormalization and factorization scales. The scale variation uncertainties impact the overall process normalization and event topology, affecting the shape of the kinematic distributions. Scale variations also affect parton shower simulations independently. The normalization of distributions used for statistical inference is affected by uncertainties in background processes, trigger efficiencies, luminosity, and reconstruction and identification efficiencies for leptons, taus and jets. These rate uncertainties are parameterised using a LogNormal (lnN) distribution, and a detailed description is given in the following section.

### 6.8.1 Normalization uncertainties

This section outlines the sources of systematic uncertainty having an impact on the normalization of both signal and background processes. These uncertainties are generally correlated across different channels and data-taking periods. The main exception is the luminosity uncertainty, which contains both correlated and uncorrelated components across different data-taking periods; see for further details [117–119]. The uncertainties associated with the electron and muon identification efficiency are estimated to be 2%, and they are found to be correlated among the $e\tau_h$,$\mu\tau_h$, and $e\mu$ channels across different data-taking periods. Other sources of systematic uncertainty arise from the cross-section calculation of the main background processes, including $t\bar{t}$ (6%), di-boson (5%), single-top (5%), W boson (4%), and Z boson production (2%). The theoretical uncertainties related to the cross sections of the Higgs boson production mechanisms, referred to as *QCD scale uncertainties*, arise from the missing higher-order QCD corrections in fixed-order calculations.

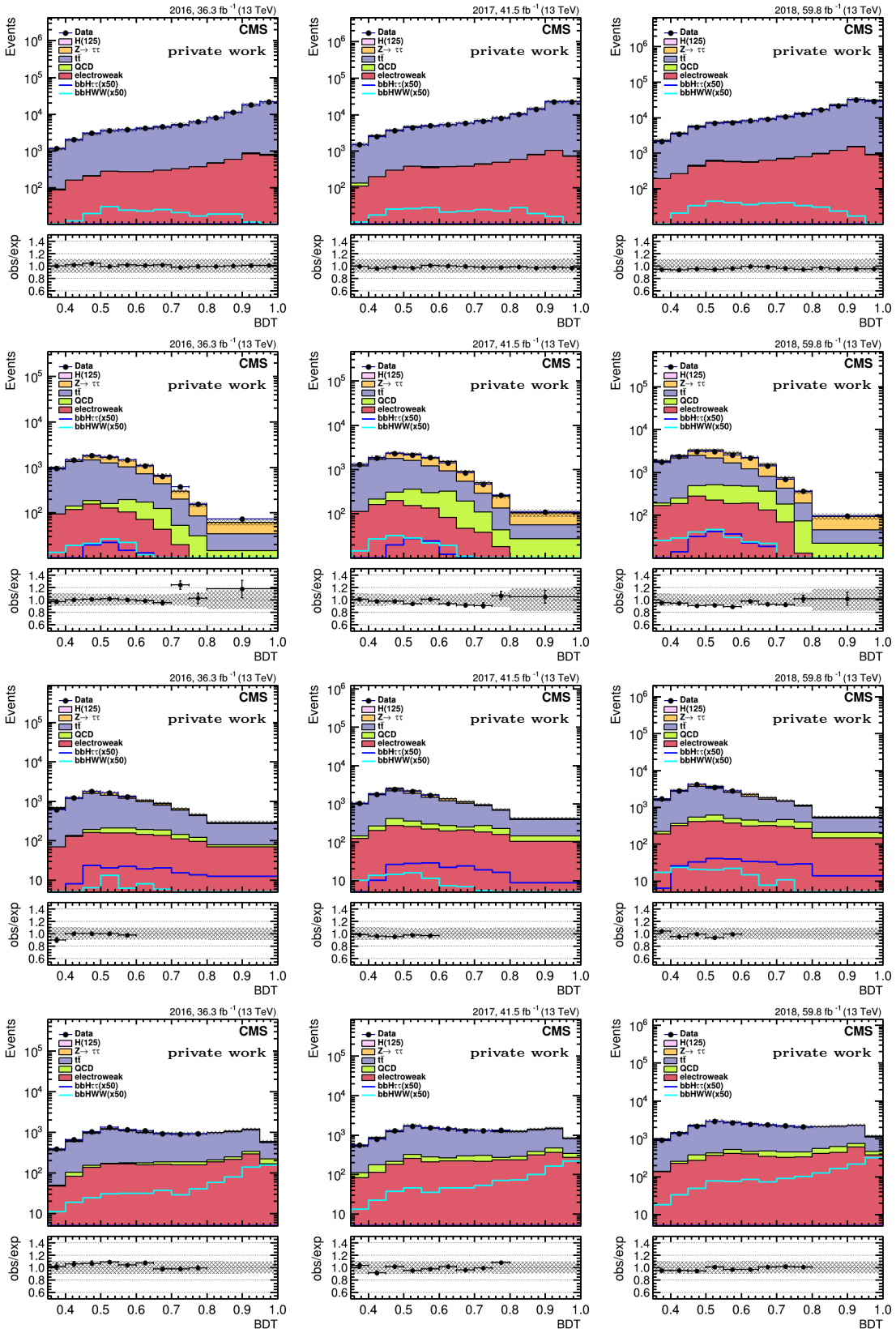


Figure 6.23: BDT score distributions, first row: $t\bar{t}$&*singletop* class, second row: DY+jets class, third row: $b\bar{b}H \to \tau^+\tau^-$, fourth row: $b\bar{b}H \to W^+W^-$. The signal distributions are shown super imposed and multiplied by 50 for better visibility.

They lead to a significant variation of the normalization of the following processes: bbH (20–24%), ZH (0.9%), WH (0.8%), ttH (8%), VBF (0.5%), and ggH (4.5–7%). Additionally, a 40% uncertainty is assigned specifically to the ggH production accompanied by b-jets. The Parton Distribution Function (PDF) and the strong coupling constant ($\alpha_s$) related uncertainties have a smaller impact on the rate, namely for ZH (1.3%), WH (1.8%), ttH (3.6%), VBF (2.1%), and ggH (3.2%). The total theoretical uncertainty in the branching ratio of $b\bar{b}H \to W^+W^-$ is estimated to be 1.5%, while for the branching ratio of $b\bar{b}H \to \tau\tau$ it is found to be 2.1%.

### 6.8.2 Shape uncertainties

This section provides a comprehensive overview of uncertainties that have common treatment across all the channels, impacting both the shape and normalization of simulated signal and background processes.

- **Tau ID Efficiency** [120] The uncertainties on the efficiency for hadronic taus are applied as a function of $p_T$ and tau decay mode. These uncertainties are correlated among $e\tau$, $\mu\tau$, and $\tau_h\tau_h$ channels while being uncorrelated across different data-taking periods. They lead to variations of 2–3% in the BDT distributions of the simulated signal and background samples.

- **Lepton to Tau Fake Rate** [120] uncertainties are defined, as a function of the lepton pseudorapity $\eta$. Applied to both semi-leptonic and fully hadronic channels. These uncertainties are uncorrelated across different years and lead to variations of up to 2% in the BDT distributions of simulated samples with misidentified leptons as $\tau_h$.

- **Tau Energy Scale** uncertainties are based on the decay mode of hadronic taus. Uncorrelated among different years, albeit correlated between $e\tau$, $\mu\tau$, and $\tau_h\tau_h$ channels, they result in variations up to 3% in the BDT distributions.

- **Electron Energy Scale** [121] are evaluated by varying the electron energy scale uncertainties. They are correlated across different years and between $e\tau_h$ and $e\mu$ channels. They induce variations to the BDT distributions of simulated signal and background samples ranging from 0.2% to 0.5%.

- **Jet Energy Scale** uncertainties are split into groups followed the recommendation [122]. Correlated uncertainties across data-taking periods include a set of corrections to the jet energy depending on the flavour.Uncorrelated uncertainties across years but correlated across channels alter the shape of the BDT distributions of the simulated samples.

- **Jet Energy Resolution** [122] uncertainties are uncorrelated across the years and correlated across channels. They lead to variations in the BDT distributions of the simulated samples ranging from 1% to 4%.

- **B-Tagging** uncertainties are presented as variations to the shape of b-tag scale factor and are applied to b-jets and light jets. These include a set of corrections to the jet energy scale depending on the jet flavour, extended to the case of c-jets. Some uncertainties are correlated between eras, while others are decorrelated. For each of the JES corrections, a respective up and down template corresponding to one-sigma variation is applied. These variations cause changes to the BDT distribution in simulated samples within the range of 0.5%–3%.

- **Unclustered Missing Transverse Energy Uncertainty** arising from the variation of $E_T^{miss}$ with unclustered energy uncertainty [122]. Uncorrelated across years, this uncertainty is correlated between all the channels.

- **Top $p_T$ Reweighting** uncertainty related to top $p_T$ dependent corrections is derived in [123] and applied to the simulated $t\bar{t}$ samples. This uncertainty is estimated with a conservative approach and leads to variations up to 7% in the BDT distributions of the simulated $t\bar{t}$ events.

- **Drell-Yan Di-lepton Mass, $p_T$ Reweighting** are applied to simulated Drell-Yan samples, uncorrelated between one and two b-tagged jets.

- **Prefiring** uncertainties related to the effect of erroneous bunch crossing pre-firing of the triggers are applied to simulated samples of the 2016 and 2017 data-taking periods.

### 6.8.3 Channel-specific uncertainties

Several uncertainties are implemented specifically for the $e\mu$ channel. These comprise the uncertainty related to the efficiency of the electron-plus-muon cross trigger. This trigger uncertainty is characterized by a normalization uncertainty of 1.5%, which remains uncorrelated across different eras. The uncertainties addressing the scale factors for the fake rate of udsg-jets mimicking electrons are applied based on the jet transverse momentum ($p_T$) in simulated events. Similarly, the uncertainty in the scale factor for the misidentification rate of b/c-jets mimicking electrons is applied in bins of the jet transverse momentum in simulated events involving heavy-flavour jets. For the fake rate of udsg-jets faking muons, an uncertainty is applied as a function of jet $p_T$ in simulated events. Likewise, an uncertainty related to the scale factor for the fake rate of heavy flavour jets mimicking muons is applied, which varies with the jet transverse momentum in simulated events. Uncertainties in the extrapolation of the opposite-sign to same-sign

(OS/SS) transfer factors used to estimate the QCD multi-jet background are characterized by shape uncertainties, containing both statistical and non-closure uncertainties as discussed thoroughly in subsection 6.5.1.

CHAPTER 7

# Results

In this chapter, the sensitivity of the analysis to the b-associated production of Higgs boson in final states with leptons is assessed. An overview of the methods employed for the statistical inference and the treatment of the systematic uncertainties is presented. Stringent constraints are derived on the Higgs boson production cross-section as an upper limit on the signal strength of the process with respect to the SM prediction. Due to the limited expected sensitivity, an upper bound is obtained for the production cross-section of the bbH process for all the channels considered. The final fit result combines the semi-, fully-leptonic and fully hadronic channels using the full CMS dataset collected during LHC Run 2. Furthermore, an interpretation of the results in the parameter space of the Yukawa coupling modifiers to the bottom and top quark is performed within the kappa framework [124].

## 7.1 Statistical Analysis

The main objective of this search is to constrain the inclusive cross-section of b-associated production of the Higgs boson by performing a simultaneous likelihood fit to the BDT score distributions. Additional constraints on the Higgs Yukawa couplings to the third-generation quarks are obtained by means of a two-dimensional likelihood scan in the parameter space of the corresponding coupling modifiers. The statistical analysis is performed within the Combine-Tool [125] software, which allows a statistical interpretation of the results and a detailed treatment of the systematic uncertainties.

In general, the statistical analysis in the Higgs sector consists of several steps [126]. The ultimate goal is to isolate the Higgs production mechanism in the final state of interest from the irreducible background processes. A widely used method for the analytical approach is the so-called confidence level for the signal (CLs), which evaluates the compatibility of observed data with the theoretical predictions. The statistical analysis

involves fitting theoretical models to experimental data. The first step is performed on a pseudo-data, also known as the Asimov dataset, where the distribution of the expected background processes is generated, which allows the comparison to the experimental data. The signal hypothesis to test is whether the observed data is compatible with a genuine signal from the Higgs boson or whether it could be attributed to background processes (background-only hypothesis). The main parameter in the signal extraction fit is $\mu$, representing the strength of the signal with relation the SM prediction. Upper limits on the production of Higgs bosons are achieved by setting the compatibility of the observed data with a background-only hypothesis, where the signal strength $\mu$ is set to zero. The CLs method enables this comparison, providing a confidence level for the presence of a signal by using the following ratio:

$$CL_s(\mu) = \frac{p_{s+b}}{1 - p_b}, \tag{7.1}$$

which can be adapted to reject a signal hypothesis with a signal given strength $\mu$ associated with a $1 - CL_s(\mu)$ confidence level. The $p_{s+b}$ and $p_b$ given in the equation 7.1 are the respective $p$-values for the two hypothesis signal plus background and background only; detailed definitions are given in [126]. This analytical approach adopts the negative log-likelihood fit to enable the modelling of uncertainties and potential statistical fluctuations in the measurement and is described in the following section.

### 7.1.1 Likelihood fit

The likelihood function quantifies the degree of agreement between an experimental observation and the theoretical model of reference. The latter is a mathematical function that depends on several parameters, namely the observable physical quantities $\vec{x}$ and the vector of nuisance parameters $\theta = (\theta_1, ..., \theta_m)$. The nuisance parameters are included to account for the systematic errors, such as the theoretical uncertainties that impact the main measurement. Therefore, the overall likelihood function for N distinct measurement of the main variables $\vec{x}$ under observation can be derived as:

$$L(\vec{x}, \theta) = \prod_{i=1}^{N} f(x_1^i, ..., x_n^i; \theta_i, ..., \theta_m) \tag{7.2}$$

For Poissonian counting experiments, as in the present case, the expected total event yield ($\nu$) can be expressed as a function of the signal strength ($\mu$), the signal ($s$), and background ($b$) yields as follows:

$$\nu = \mu \cdot s(\theta) + b(\theta), \tag{7.3}$$

where both signal and background yields are affected by specific sources of systematic uncertainty $\theta$.

### 7.1.2 Maximum likelihood fit

In the present work, the maximum likelihood fit has been performed by minimizing the negative logarithmic likelihood (NLL). The mathematical formulation of the negative likelihood is the following:

$$NLL = -log(L), \tag{7.4}$$

where $L$ is the maximum likelihood function defined as:

$$L(data, Asimov|\mu, \theta) = \prod_i P(n_i|\mu \cdot s_i(\theta) + b_i(\theta) \prod_j p(\tilde{\theta_j}|\theta_j)). \tag{7.5}$$

In this analysis, the maximum likelihood function is computed based on the boosted decision tree (BDT) score distributions for each category $i$, as described in the previous chapter. The likelihood function, denoted as $L$ in equation 7.5 is expressed as a product over all bins in the BDT score distributions across all the channels and categories. It includes two key components, namely a Poisson distribution ($P(n_i|\mu \cdot s_i(\theta) + b_i)$ representing the probability of observing a certain event yield in each bin (i-th bin $n_i$ events), assuming the expected contributions from the signal ($s_i$), i.e. the bbH process, and the background sources ($b_i$). An additional term related to the nuisance parameters ($\theta$) parameterizes the effects of the systematic uncertainties from the experimental setup and the theoretical model.

This likelihood function is used in a fit, where all the BDT categories are simultaneously considered. Initially, the fit is performed on a background-only Asimov dataset, i.e. a pseudo-dataset where the sum of all predicted background processes determines the content of each bin. The fit aims to evaluate the compatibility of the observed data with this background-only hypothesis ($\mu = 0$) by including the parameter signal strength $\mu$.

### 7.1.3 Treatment of nuisance parameters

Systematic uncertainties, discussed in detail in the previous chapters, can be generally classified into two categories: those that exclusively impact the event yield, known as normalization uncertainties, and those that also influence the shape of the predicted distribution. Concerning the rate-altering systematic uncertainties, the log-normal distribution is used for the parameterization:

$$p(\tilde{\theta}|\theta) = \frac{1}{\sqrt{2\pi} ln k} \cdot \frac{1}{\tilde{\theta}} \cdot exp(-\frac{(ln(\tilde{\theta}/\theta_m))^2}{2ln^2 k}). \tag{7.6}$$

The above equation represents a random variable whose logarithm follows a normal distribution, characterized by a mean $\tilde{\theta} = ln(\theta_m)$ and a standard deviation $\tilde{\sigma} = ln(k)$. The preference for the log-normal over the Gaussian distribution originates because it ensures a positive-definite normalization for the modelled nuisance by construction. This

is particularly relevant when the nuisance is applied as a multiplicative factor to the overall signal strength, hence it cannot acquire negative values, as they are considered to be unphysical.

The second type of uncertainties are referred to as shape-altering systematic uncertainties, which have a significant impact on a subset of the shape parameters in the signal fit. A parametric approach is implemented to accommodate the systematic shift in the affected shape parameters $\alpha$:

$$\alpha = \alpha_{ctrl} + \sum_k \frac{(\alpha_{up}^{\theta_k} - \alpha_{down}^{\theta_k})}{2n} \cdot \theta_k, \tag{7.7}$$

where the index $k$ runs over all shape-altering systematic uncertainties. The symbol $n$ stands for the number of systematic variations in standard deviation units, and the parameter $\alpha_{ctrl}$ denotes the fitted value of the shape parameter for the central template fit. The $\alpha_{up\{down\}}$ parameters are extracted in the systematic up- (down-) variations of the signal template associated with the nuisance parameter $\theta_k$. To incorporate the systematic shape variations, a value of two standard deviations from the central signal template is chosen.

### 7.1.4 Test statistics

One of the most reliable methods for hypothesis testing of the value of $\mu$ is using the profile likelihood ratio, a statistical estimator used to quantify the likelihood of different parameters in a statistical model:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})}, \tag{7.8}$$

where $\hat{\hat{\theta}}$ is the conditional maximum-likelihood of $\theta$, i.e. the value of $\theta$ that maximizes the likelihood $L$ for a specific $\mu$. The $\hat{\mu}$ and $\hat{\theta}$ are the unconditional ML estimators which do not impose any constraints on the corresponding parameters. To set an upper limit on the signal strength parameter $\mu$, the test statistics definition used in this work is:

$$q_\mu = -2ln\lambda(\mu) = -2ln\frac{L(data|\mu, \hat{\theta_\mu})}{L(data|\mu = 0, \hat{\theta_0})}, \tag{7.9}$$

where higher values of $q_\mu$ indicate a more significant discrepancy between the data and the hypothesized value of $\mu$. In the present work, the upper limits are defined by the CLs method [127] described in the first section. The latter are derived for both the observed and expected signal strength whilst requiring the 95% confidence level (C.L.) criterion, corresponding to a value of $CL_s(\mu) \leq 0.05$.

## 7.2 Results in the $e\mu$ channel

The upper limits on the inclusive signal process cross-section are derived by fitting all event categories simultaneously. The event categories in the $e\mu$ channel correspond to the $t\bar{t}$ and single top production, the Drell-Yan-plus-jets and two dedicated signal categories, namely $b\bar{b}H \to \tau^+\tau^-$ and $b\bar{b}H \to W^+W^-$, respectively.

The BDT score distributions are fitted for both the signal and background categories. The data and MC distributions for the BDT categories in the $e\mu$ channel before the final likelihood fit (prefit) in the background and the signal categories are shown in Figure 7.1. The data-to-simulation agreement observed for all categories is optimal within the statistical uncertainties. In general, the need for a prefit plot is motivated by an observed low goodness of fit (GoF) value to identify potential discrepancies. As this is not the case for the current analysis, this set of plots are shown for completeness.

Upper limits are derived on the bbH process signal strength in which the Higgs boson is produced either by coupling to the bottom quarks directly bbH ($y_b$), or via the top-quark mediated loop ($y_t$), by taking into account the interference term bbH ($y_b y_t$). The 95% C.L. expected upper limits are derived using a signal-plus-background Asimov dataset where the bbH signal corresponding to a signal strength of $\mu = 1$ is injected. The expected median limit from the SM prediction and the uncertainty interval of one- (68%) and two-standard deviations (95% C.L.) are computed. The observed upper limit on the signal strength are derived from the actual data collected by the CMS detector, corresponding to the total luminosity of 138 fb$^{-1}$ for the LHC Run 2 data taking.

The observed (expected) bounds for the $e\mu$ channel are shown in table 7.1 for each data-taking year and for the full Run 2 dataset. The observed upper limits for the full Run 2 dataset on the signal strength of the bbH process in the $e\mu$ channel is equal to 18.7, which is compatible, yet slightly lower than the expected value of 19.1. The equivalent set of the plots after the final fit (postfit) for the background and signal categories in the $e\mu$ channel are shown in Figure 7.2.

The goal of having prefit and postfit set of plots can be summarized as follows: Prefit plots are produced to ensure the model accurately describes the data before unblinding. Since the fitting procedure adjusts nuisance parameters related to systematics, it is crucial to fine-tune the trade-off between the accuracy of the data description and the introduction of additional penalty terms in the likelihood fit. If the goodness-of-fit (GoF) shows a high p-value (p$>$ 10%) and postfit plots indicate no systematic effects, the statistical inference is considered to be robust, as in the case of the present analysis. A set of exemplary plots of the goodness-of-fit test are listed in the Appendix A.

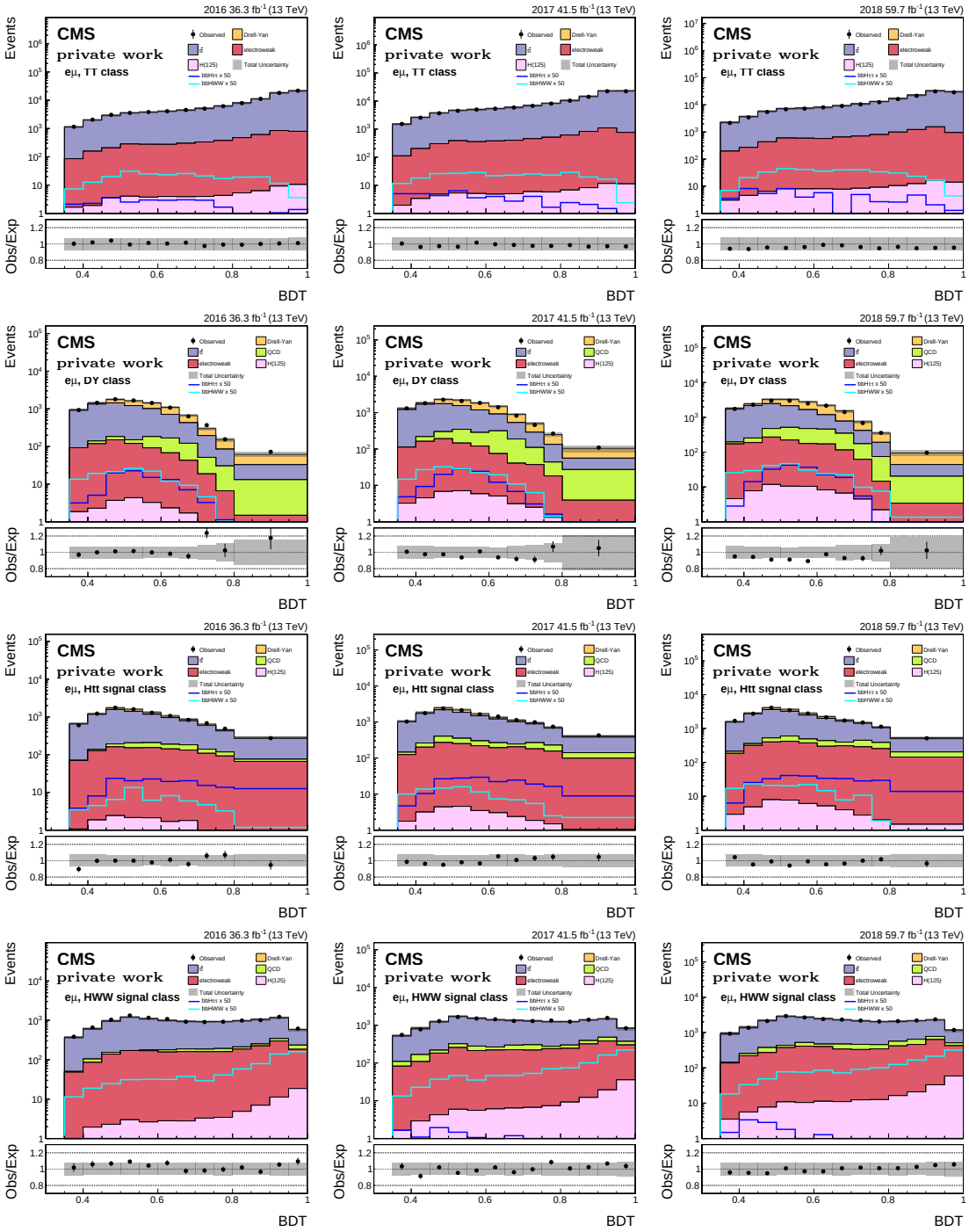


Figure 7.1: Prefit plots for all the categories in $e\mu$ channel for 2016, 2017 and 2018 years separataly. The first row shows the $t\bar{t}$ and single top class, the second one the DY+jet background, while the third the $bbH \to \tau\tau$ signal and the fourth the $bbH \to WW$.

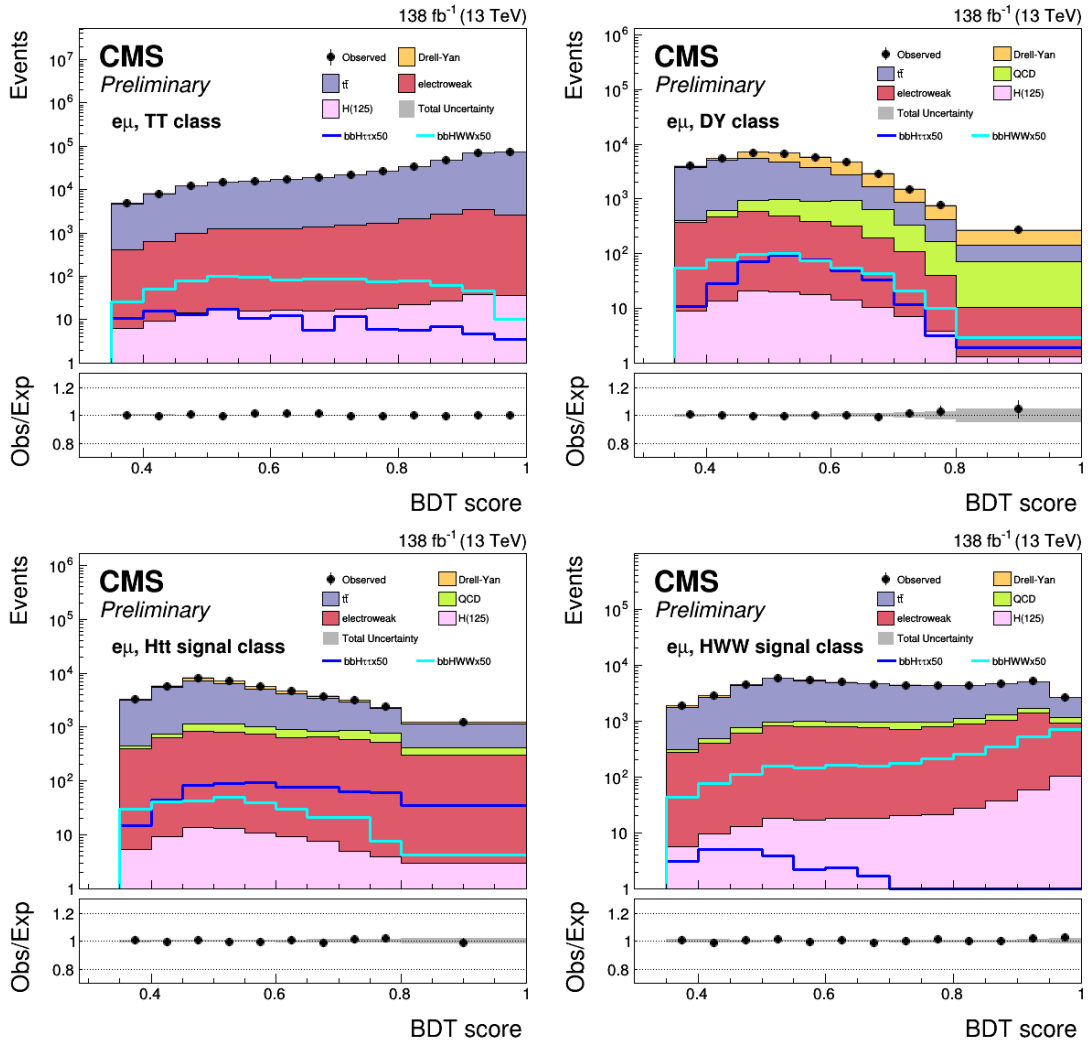


Figure 7.2: Postfit plots for all the categories in $e\mu$ channel for Run 2 data. The first row shows the $t\bar{t}$ and single top class (Top left) and the DY+jet background (top right), while the second row shows the $bbH \rightarrow \tau\tau$ signal (bottom left) and the fourth the $bbH \rightarrow WW$ signal (bottom right).

| $e\mu$ **channel** | | | | | | |
|---|---|---|---|---|---|---|
| | $-2\sigma$ | $-1\sigma$ | Median | $+1\sigma$ | $+2\sigma$ | **Observed** |
| 2016 | 13.7 | 19.5 | 30.5 | 50.6 | 84.9 | 43.1 |
| 2017 | 13.8 | 19.5 | 30.5 | 50.4 | 83.6 | 31.7 |
| 2018 | 13 | 18.5 | 28.8 | 47.9 | 79.4 | 23.9 |
| **Run 2** | 8.6 | 12.3 | 19.1 | 31.8 | 52.7 | 18.7 |

Table 7.1: Expected and observed upper limits derived on bbH production in $e\mu$ channel.

### 7.2.1 Impacts of nuisance parameter in the $e\mu$ channel

The impact plots are generated for a signal-plus-background Asimov dataset where the bbH signal corresponding to a signal strength of $\mu = 1$ is injected. The aim is to study the contributions from the largest nuisance parameter to the overall expected signal strength for the $e\mu$ channel. The impact and pull distributions are shown in Figure 7.3, after combining all the data-taking periods in LHC Run 2. The leading uncertainty for this channel is the Monte-Carlo (MC) bin-by-bin statistical uncertainty affecting the signal template of the fourth category, which corresponds to the $H \to W^+W^-$ signal category in $e\mu$, pointing to the fact that this measurement is statistically limited. The second largest uncertainties are, respectively, the theory uncertainties related to the QCD scale of bbH and the experimental reconstruction efficiency of b-tagged jets.

## 7.3 Combination of the results for all the channels

The upper limits at 95 % confidence level are computed on the signal strength for the bbH process for each channel and data-taking year separately and further combined. Table 7.2 shows the result of the bounds calculated on an Asimov dataset labelled as median expected and for the LHC Run 2 data set labelled as observed. The observed (expected) 95 % upper limits for the combination of all the channels for Run 2 data is 3.7 (6.1).

The calculated upper limits on the signal strength account for the diagrams in which the Higgs boson is generated through Yukawa interaction with the top ($y_t$) or bottom quarks ($y_b$), further including the interference term ($y_b\ y_t$). The resulting limits are shown in Figure 7.4 for all the channels and their combination. The theory predictions are also displayed with a red line set at 1, corresponding to an estimated value of cross-section 1.489 pb. The observed limits align with the expected values within the 68% confidence level. It is interesting to note that by combining the leptonic and fully-hadronic channels, the overall sensitivity is significantly improved with respect to the single most-sensitive channel, i.e. the $\tau_h\tau_h$ one. For this fully-hadronic channel, the associated observed (expected) upper limit on the bbH signal strength is found to be 8.5 (7.8) pb. The observed

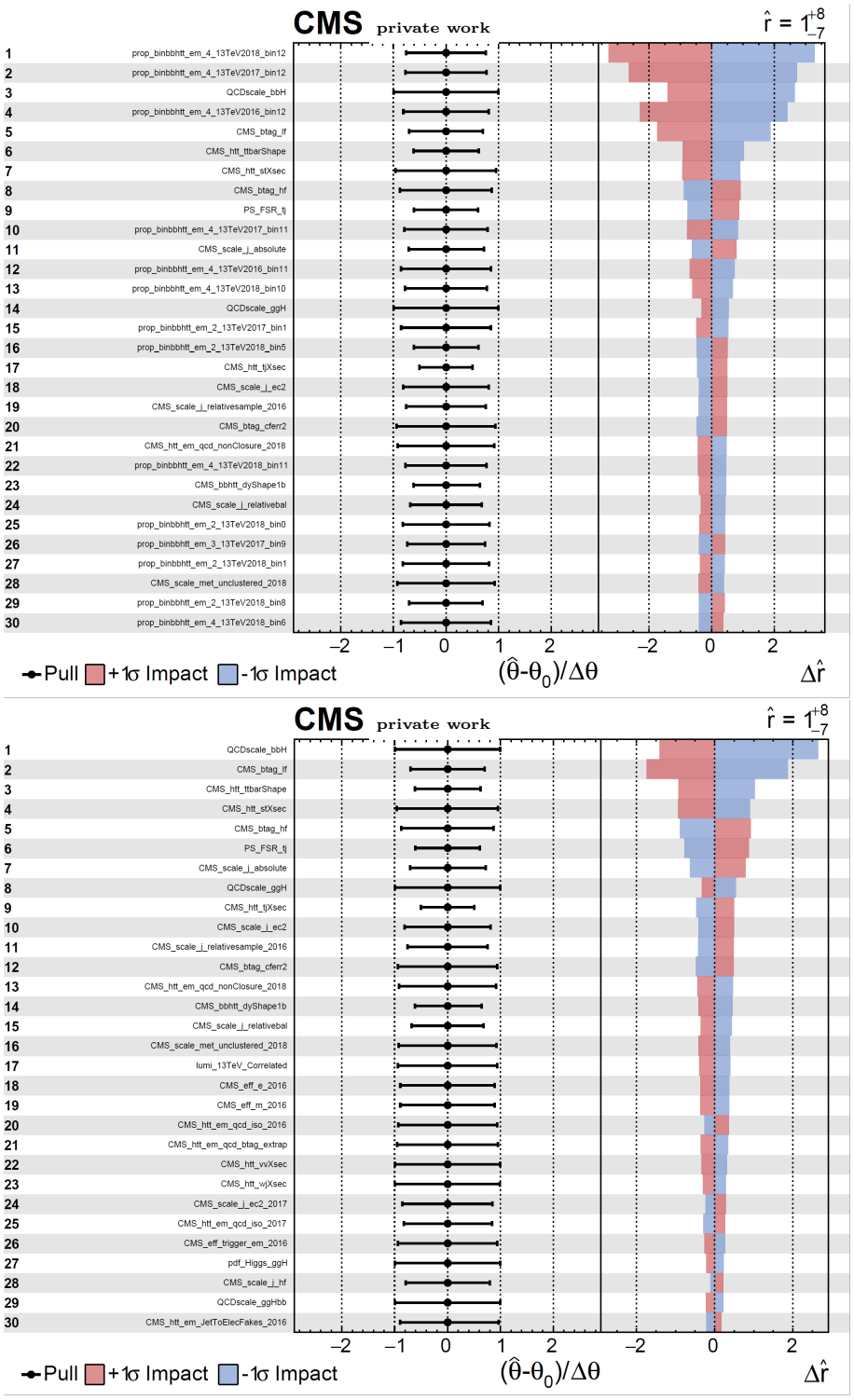


Figure 7.3: List of the uncertainties ranked in decreasing order by their impact on the overall signal strength uncertainty in the $e\mu$ channel for the full Run 2 dataset in the case of all the systematic uncertainties (top) and excluding the bin-by-bin statistical uncertainties (bottom).

| Channel | $-2\sigma$ | $-1\sigma$ | Median Expected | $+1\sigma$ | $+2\sigma$ | Observed |
|---|---|---|---|---|---|---|
| $\tau_h\tau_h$ | 3.7 | 5.6 | 8.5 | 14.6 | 25.5 | 7.8 |
| $\mu\tau_h$ | 5.1 | 7.3 | 11.6 | 19.8 | 34.4 | 8.6 |
| $e\tau_h$ | 8.0 | 11.5 | 18.0 | 30.8 | 53.2 | 10.1 |
| $e\mu$ | 8.6 | 12.3 | 19.1 | 31.8 | 52.7 | 18.7 |
| 2016 | 4.5 | 6.4 | 10.0 | 17.0 | 29.2 | 9.8 |
| 2017 | 4.9 | 7.1 | 11.3 | 19.5 | 34.2 | 8.5 |
| 2018 | 4.2 | 6.1 | 9.6 | 16.4 | 28.2 | 4.8 |
| Run 2 | 2.7 | 3.9 | 6.1 | 10.4 | 18.0 | 3.7 |

Table 7.2: Upper limits derived on the signal strength modifier for all the channels and their combination. Both expected and observed limits are shown. The first four rows (labelled by the name of each specific channel) correspond to the combined data taking years for each channel separately. The limits for the combination of all channels together are shown for each data-taking year denoted by 2016, 2017 and 2018. Finally, the expected limits for the combination of Run 2 are shown in the last row. The last column corresponds to the observed limits predicated on data for each channel separately, for a combination of channels per year and Run 2 data combined.

cross-section upper limit obtained from the combination is further reduced to 3.7 pb, while the expected value yields a value of 6.1 pb, which is a result beyond initial sensitivity estimations [106]. Since this measurement is statistically limited, this upper bound could be further improved by complementing the analysis using LHC Run 3 data. Assuming the initial projections of luminosity delivered by LHC during Run 3 (300 $\text{fb}^{-1}$), combining the Run 2 and Run 3 datasets could yield an overall improvement of approximately a factor $\sqrt{2} \approx 1.4$ in the overall statistical precision.

The likelihood scan of the signal strength, shown in Figure 7.5, reveals a negative best-fit signal strength. This arises from an under-fluctuation of the observed data relative to the expected signal, indicating a reduction in the observed Higgs signal compared to the Standard Model prediction. Despite this negative best-fit value, the observed results remain consistent with the $\mu = 1$ hypothesis at a $2\sigma$ (two standard deviations) level. This indicates that, within the statistical and systematic uncertainties considered, the data is compatible with the expected Higgs signal strength, and any deviation is within the expected range of fluctuations.

The impact plots, which show the contribution from the main systematics uncertainties for combining all the channels studied in the search, are shown after unblinding the data in Figure 7.6. The full list of observed impacts is shown in A. The major systematic uncertainty in this analysis arises from the QCD scale uncertainty on the gluon splitting process, followed by the statistical uncertainties, mainly the bin-by-bin statistical uncer-

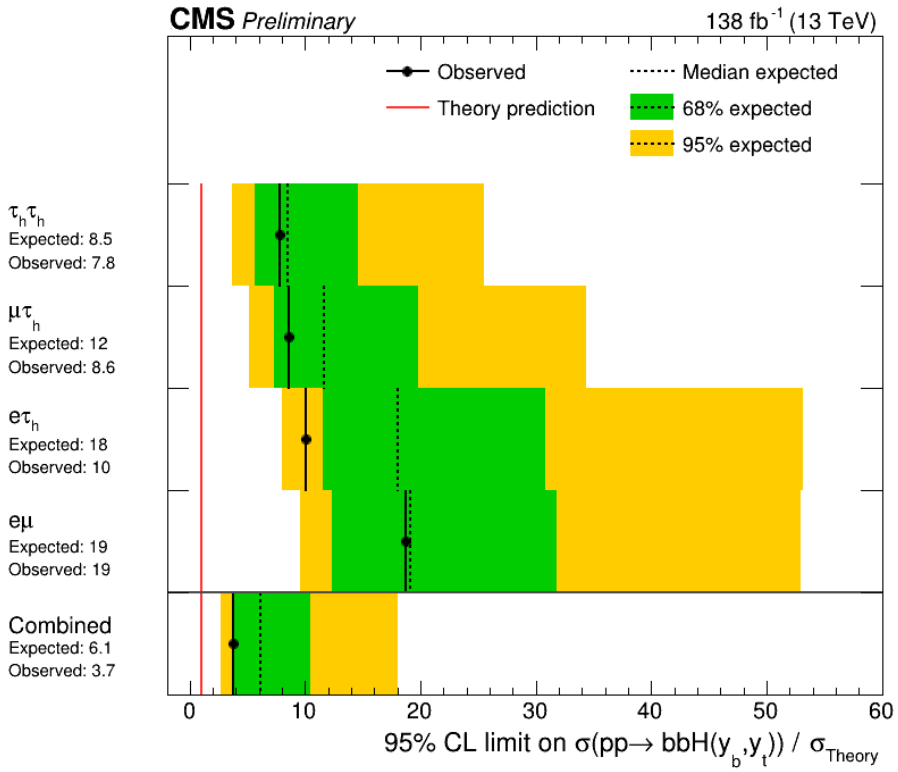


Figure 7.4: Upper limits at 95% C.L. on the signal strength of the bbH process. Both expected (dashed lines with respective $1\sigma$ and $2\sigma$ uncertainty bands in yellow and green) and observed limits (black marker) are shown [128].

tainties affecting the signal template in all the channels. In general, the ranking of the systematic uncertainties by their impact on the sensitivity agrees with the expected one from the signal-plus-background Asimov dataset shown in Figure 7.3.

### 7.3.1 Two-dimensional likelihood scan

As mentioned previously, the analysis of the bbH production mode is also motivated from the point of view of constraining the coupling of Higgs boson to bottom quarks. To this end, the so-called $\kappa$ framework [124] is used to characterize the Higgs coupling properties. Within this framework, the Higgs couplings are scaled by the corresponding $\kappa_i$ parameters. These parameters are defined as the ratios of the couplings of the Higgs bosons to particles $i$ to their respective Standard Model values. Within the $\kappa$ framework, a single narrow resonance is assumed, which allows using the zero-width approximation [124]

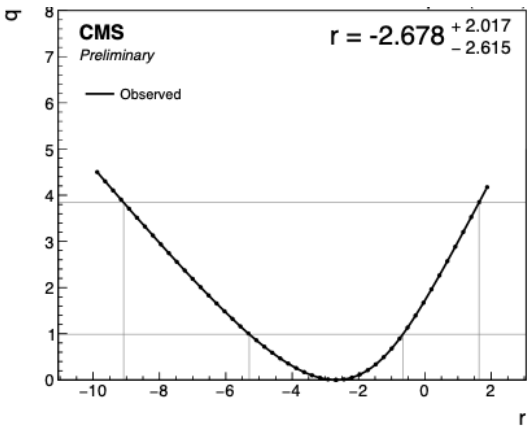


Figure 7.5: Best-fit value for the likelihood scan of the signal strength.

to decompose the cross-section as follows:

$$(\sigma \cdot BR)(i \to H \to f) = \frac{\sigma_i \cdot \Gamma_f}{\Gamma_H},, \tag{7.10}$$

where $\sigma_i$ is the production cross-section through the initial state $i$, $\Gamma_f$ is the partial decay width into the final state $f$, and $\Gamma_H$ is the total width of the Higgs boson. To introduce the $\kappa$ parameters, each term of the equation 7.10 is expressed as their Standard Model expectation and multiplied by the square of a coupling strength modifier for the corresponding process at leading order:

$$(\sigma \cdot BR)(i \to H \to f) = \frac{\sigma_i^{SM}\kappa_i^2 \cdot \Gamma_f^{SM}\kappa_f^2}{\Gamma_H^{SM}\kappa_H^2} \to \mu_i^f \equiv \frac{\sigma \cdot BR}{\sigma_{SM} \cdot BR_{SM}} = \frac{\kappa_i^2 \cdot \kappa_f^2}{\kappa_H^2}. \tag{7.11}$$

When the parameter $\kappa_i$ equals one, there is a perfect alignment with the SM prediction. In equation 7.11, the term $\mu_i^f$ is defined as the rate relative to the SM expectation, and $\kappa_H$ is the term that accommodates the SM Higgs width to take into account the changes of the SM Higgs coupling strengths ($\kappa_i$).

To extrapolate the limits on the coupling structure of the Higgs boson production in association with b-quarks coupling scaling parameters, $\kappa_t$ and $\kappa_b$ are introduced to perform the likelihood ratio scan. The bbH production mode induced by the top Yukawa loop (bbH($y_t^2$)) is represented by $\kappa_t^2$, the direct coupling of the Higgs to bottom quarks (bbH($y_b^2$)) is scaled by $\kappa_b^2$ and the interference term by $\kappa_b\kappa_t$. Additionally, to constrain the $\kappa_t$ parameter, this current analysis is incorporated with the results of the inclusive Higgs production cross-section measurement in final states with tau leptons [129]. The latter combines three analyses targeting the dominant Higgs production modes, i.e., the gluon fusion, the vector boson fusion, and the vector-boson-associated production. Since all of these analyses employ a veto on the presence of b-tagged jets in the event, they

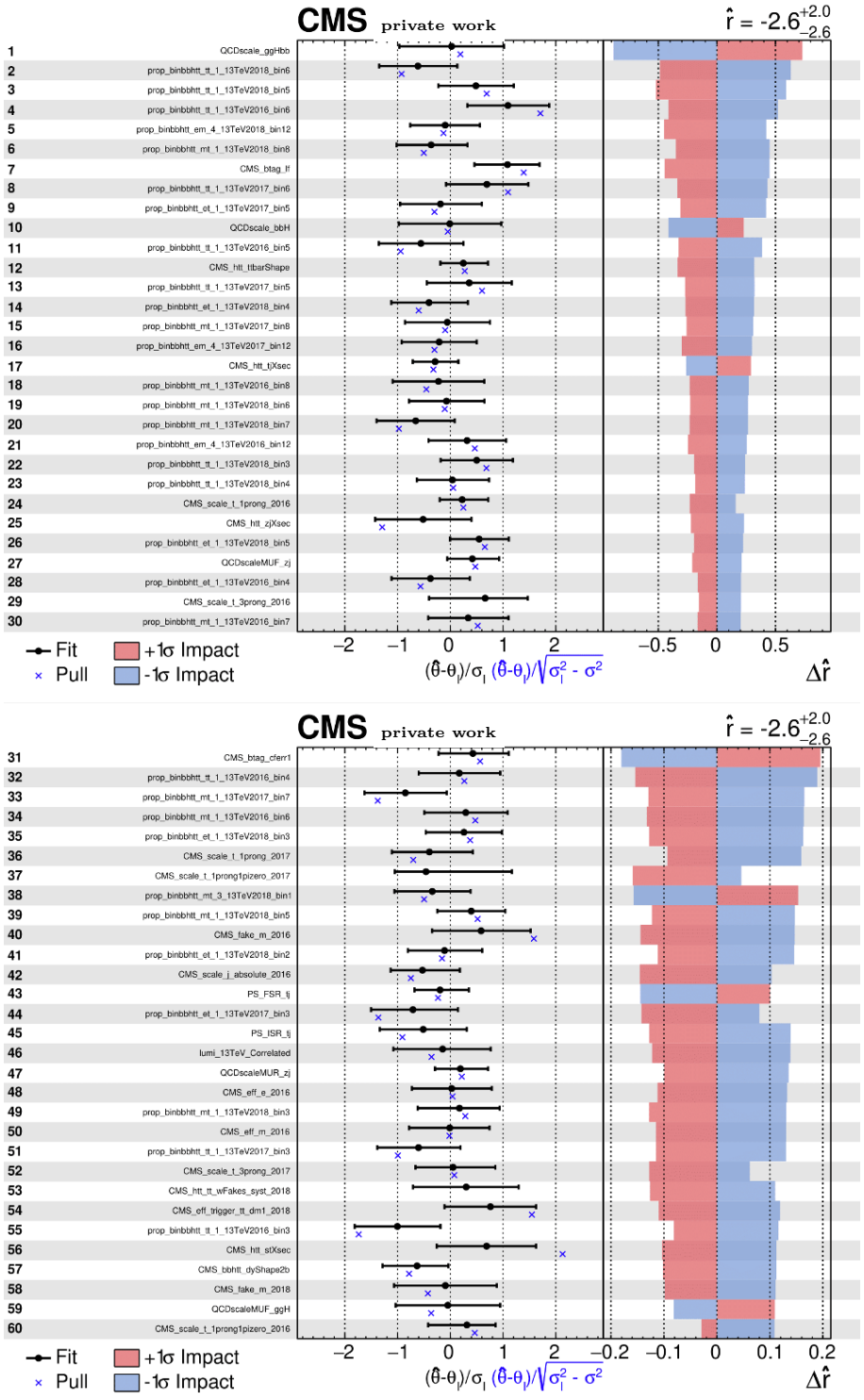


Figure 7.6: The list of uncertainties with the largest impacts after combining all the semi-(fully-) leptonic and fully-hadronic channels for the whole Run 2 dataset.

have an orthogonal selection with respect to the bbH analysis, hence the measurements of all production modes can be combined.
The two-dimensional $(\kappa_t, \kappa_b)$ scan is shown in Figure 7.7. The results from the Higgs production cross-section measurement in final states with tau leptons [129] are shown in red and the combination of this analysis with the previous result for $\kappa_t$ is in green. The green diamond denote the SM expectation corresponding to $(\kappa_t, \kappa_b) = (1.0, 1.0)$. The observed limits are shown with the blue line. The best-fit point is shown with a blue cross and it corresponds to $(\kappa_t, \kappa_b) = (-0.73, 1.58)$. The overall result on constraining $\kappa_b$ reveals a significant improvement, which can be evinced from the more stringent constraints, highlighted by the 68% CL contour. The derived limits are compatible with the SM expectation at 95% CL. The fit has been performed by treating the $\kappa_\tau$ parameter as a free parameter, i.e. it is freely-floating. In contrast, the other coupling parameters are set to their SM value (unity). Therefore, this statistical treatment allows us to derive indirect constraints to the Yukawa coupling modifier of the third-generation fermions.

In conclusion, the first analysis of the bbH production mode in the final states with leptons performed at LHC was presented with the CMS Run 2 data. So far, no dedicated analysis has been performed for the bbH at the ATLAS experiment. The sensitivity of the analysis allows stringent limits on the bbH inclusive cross section and competitive constraints on the Yukawa coupling structure of both the top and bottom quark.

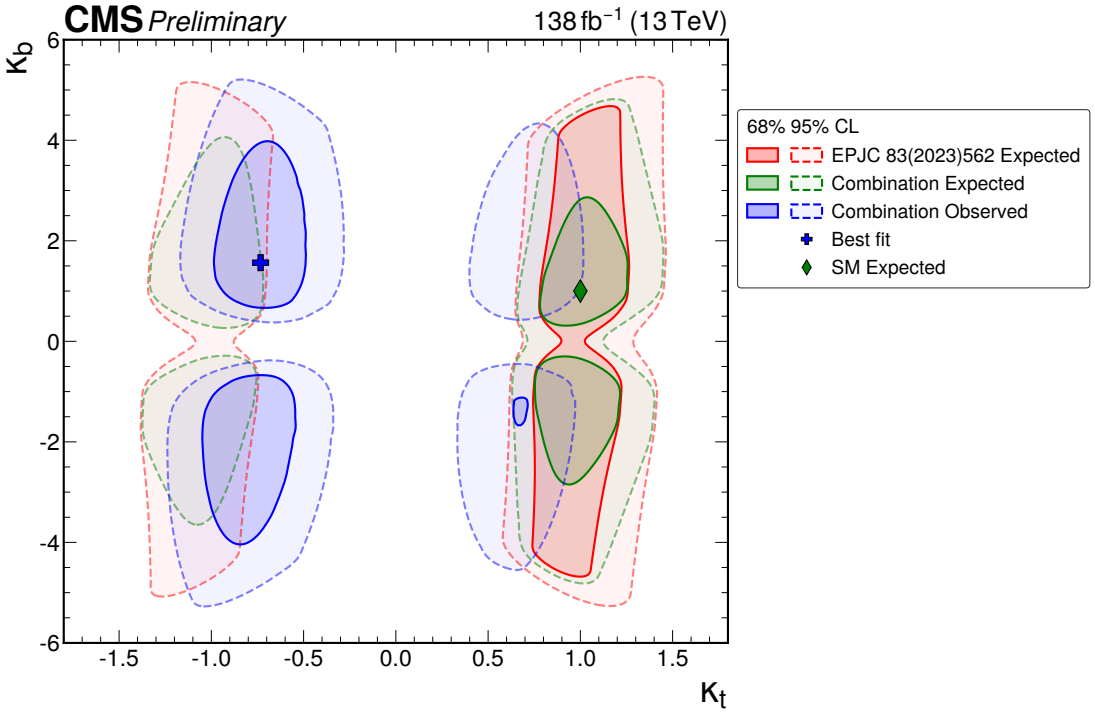


Figure 7.7: Two-dimensional scan of $(\kappa_b, \kappa_t)$ for the inclusive Higgs measurement in the tau decay channel for the dominant Higgs production modes [129](labelled in the legend as EPJC) in red and its combination with the present bbH analysis in green. The observed limits using the full Run 2 dataset are shown with blue lines. The dashed (continuous) lines indicate respectively the 95(68)% confidence areas [128].

CHAPTER 8

# Summary & Conclusion

The discovery of the Higgs boson in July 2012 by ATLAS [30] and CMS [31] collaborations at the Large Hadron Collider opened a new era of precision measurements in particle physics. Since then, many of the properties of this scalar boson have been studied, and experimental and theoretical efforts are still being conducted to refine the measurements of the most rare Higgs boson production and decay modes. Accurate measurement of all the properties of the Higgs boson is crucial to confirm the Standard Model predictions, as it could otherwise provide indirect evidence for Beyond SM physics, as in the case of models with extended Higgs sectors, such as the Minimally Supersymmetric Standard Model, as well as dark matter Higgs portal models.

This study is the first dedicated Standard Model (SM) search for the b-associated Higgs production (bbH) mode, which is yet to be observed experimentally. The discovery of the bbH process is the last missing piece of the puzzle of the leading Higgs production modes at the LHC. One of the main motivations behind the search for the different production modes of the Higgs boson is to clarify its coupling structure. The coupling of the Higgs boson to fermions is introduced via the Yukawa interactions, and with the bbH process, one can measure the Higgs coupling to b-quarks ($y_b$) directly. The latter was measured only in the decay process, namely the Higgs decay into $b\bar{b}$. Therefore, this search can complement and extend the sensitivity reach to the bottom Yukawa coupling. However, the bbH search is very challenging from the experimental perspective due to the large irreducible amount of background from other Higgs production modes such as the gluon fusion (ggF). In addition, from the theoretical point of view, this analysis was thought to have too little sensitivity at the LHC to be able to access the bottom Yukawa ($y_b$) coupling directly, as described thoroughly in [106]. However, in the present analysis, multivariate methods were employed and proved to enhance the signal purity and suppress the irreducible background.

This search was conducted with the Run 2 data collected by the CMS experiment at the centre of mass energy of $\sqrt{s} = 13$ TeV, corresponding to an integrated luminosity of

138 fb$^{-1}$. The bbH search was performed in the Higgs boson decay channel to a pair of $\tau$ leptons. Initially, the Higgs decay to tau leptons was studied in four dedicated final states: fully-leptonic ($\tau_e\tau_\mu$), semi-leptonic ($\tau_e\tau_h$, $\tau_\mu\tau_h$), and fully-hadronic ($\tau_h\tau_h$). The decay of Higgs to the pair of tau leptons has a branching ratio of only 6%. However, owing to its relatively clean signature, it represents one of the most sensitive fermionic channels at the LHC. The tau leptons decay mostly hadronically and only 6% of the times fully leptonically, i.e. via the $e\mu$ channel. Therefore, to enhance the sensitivity in the fully leptonic channel ($e\mu$), the Higgs decay channel to a pair of W bosons was included, which dominates the main sensitivity to the $e\mu$ final state. The Higgs decay to W bosons has a larger branching ratio of 21% with respect to the other signal process. Since the main focus of the search is the bbH production mode rather than the decay mode, the combination of both decay modes in the ($e\mu$) channel significantly enhanced this analysis sensitivity, making it comparable to the analogous bbH analysis in the $e\tau_h$ channel. In addition, from the experimental perspective, the $H \to \tau^+\tau^-$ and $H \to W^+W^-$ channels feature a similar event topology which makes the inclusive analysis of both processes viable.

The main focus of this thesis is the analysis of the $e\mu$ channel, which was later combined with other channels to derive the final results. This channel has been thoroughly studied, from the sample production and generation to the event selection, data-driven methods for the background modelling, and statistical inference. A core part of this thesis work has been implementing a dedicated framework to enhance the signal-to-background ratio using Machine Learning models, followed by studying the channel-specific optimization techniques to improve the classification results.

The final results of this search are derived from the Higgs boson production cross-section in the form of an upper limit on the bbH signal strength process. The upper limits of 3.7 (6.1) are observed at 95 % confidence level for the combination of all the channels. Furthermore, this search places constraints on the Higgs Yukawa coupling to b-quarks within the kappa-model interpretation, resulting in a best-fit value for the couplings of $(\kappa_t, \kappa_b) = (-0.73, 1.58)$, which is compatible with the predictions of the Standard Model at 95 % confidence level. This analysis is undergoing the final revision process before its publication in Physics Letters B. The results obtained for the bbH production search are in part due to the precise data analysis techniques employed and optimal data quality. The latter necessitates sophisticated detector hardware and a precise set of calibration and alignment constants for all subsystems during the whole data-taking period. The maintenance and calibration of the detector require a significant amount of work from all the members of the CMS collaboration. For this reason, I made a leading contribution to monitor and enhance the alignment of the CMS silicon tracker modules during this thesis work. Part of these efforts was directed at the Run 3 data-taking calibration.

The search for the b-associated production of Higgs boson did not yield evidence at the LHC; however, the result goes beyond the initial sensitivity estimations. The main limiting factors are two-fold. Firstly, due to the large theory uncertainties affecting the signal process normalisation, namely the gluon splitting QCD scale uncertainties.

This implies that better results can be achieved with further refinement of the theoretical uncertainties by performing the calculation in the next-to-next-to-leading order. Secondly, the analysis sensitivity is also affected by the statistical uncertainty of the simulated events, which can be reduced in the future. Therefore, observing the bbH process at the LHC might be possible by combining the Run 2 and Run 3 datasets and concurrently improving analytical methods and by re-adapting this search during the high-lumi LHC era, which is planned to start in 2029. Additionally, a combination of this analysis with other final non-leptonic states can be performed to enhance further the sensitivity reach to the b-associated Higgs production mode.

Besides these additions, a refinement of the analysis techniques, such as a more precise background estimation, using, for instance, data-driven techniques for Drell-Yan jets background estimation with the embedding technique [130]), could potentially improve the results. Using a more sophisticated algorithm to enrich the signal-to-background ratio, such as the hierarchical training introduced in this thesis, in combination with more statistics, can also prove advantageous. Finally, a further refinement of the final state particle reconstruction with improved b-tagging techniques and lepton identification efficiency could also improve this analysis.

In conclusion, the first dedicated search for the b-associated production of Higgs boson in final states with leptons was carried out using the data from CMS Run 2 at the Large Hadron Collider. With the current precision of this analysis, stringent upper limits on bbH inclusive cross-section were imposed, followed by establishing competitive constraints on the Yukawa couplings of the top and bottom quarks. This result contributes to the current knowledge regarding the b-associated production of Higgs boson and sets the stage for ongoing efforts and further improvement. With further optimization in the experimental methods and data analysis techniques, a potential discovery of the b-associated production of the Higgs boson could be at hand well during the High-Lumi LHC phase.

APPENDIX A

# Additional Figures

## A.1 Control distributions

Additional Data-MC figures for the $e\mu$ channel are shown in this appendix for the 2016 and 2017 data-taking periods. The control plots with inclusive (i.e. no b-tag cut applied) selection are given in A.1, A.2, A.3 and A.4.

Further control plots after applying the selection $1 \leq N_{btag} \leq 3$ are shown in A.5, A.6, A.7 and A.8.

## A.2 Impacts and Goodness of Fit test

The full list of observed impacts for the unblinded signal strength is shown in Figures A.11 to A.20. These impacts follow the impacts shown in chapter 7.3.1.

The GoF tests are derived individually for each channel and are shown for the $e\mu$ in Figure A.9. Furthermore, the combined GoF test for all the channels in each era and the GoF test for the combined Run 2 data are shown in A.10.

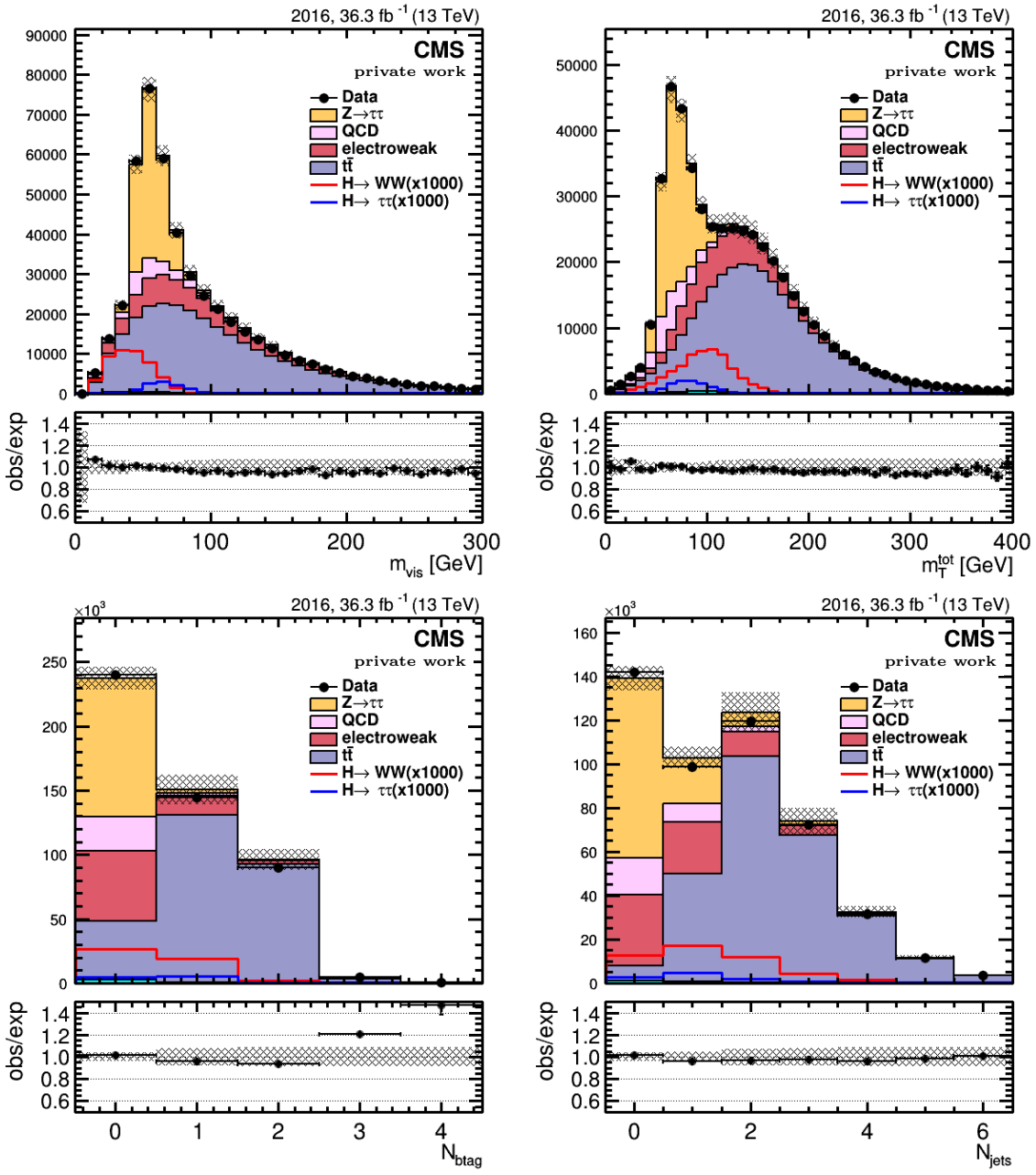


Figure A.1: Data and Monte Carlo distributions in $e\mu$ channel for 2016 data set. Upper-left: visible mass of the electron-muon, Upper-right: total transverse mass, Lower-left: b-tagged jet multiplicity, Lower-right: jet multiplicity.

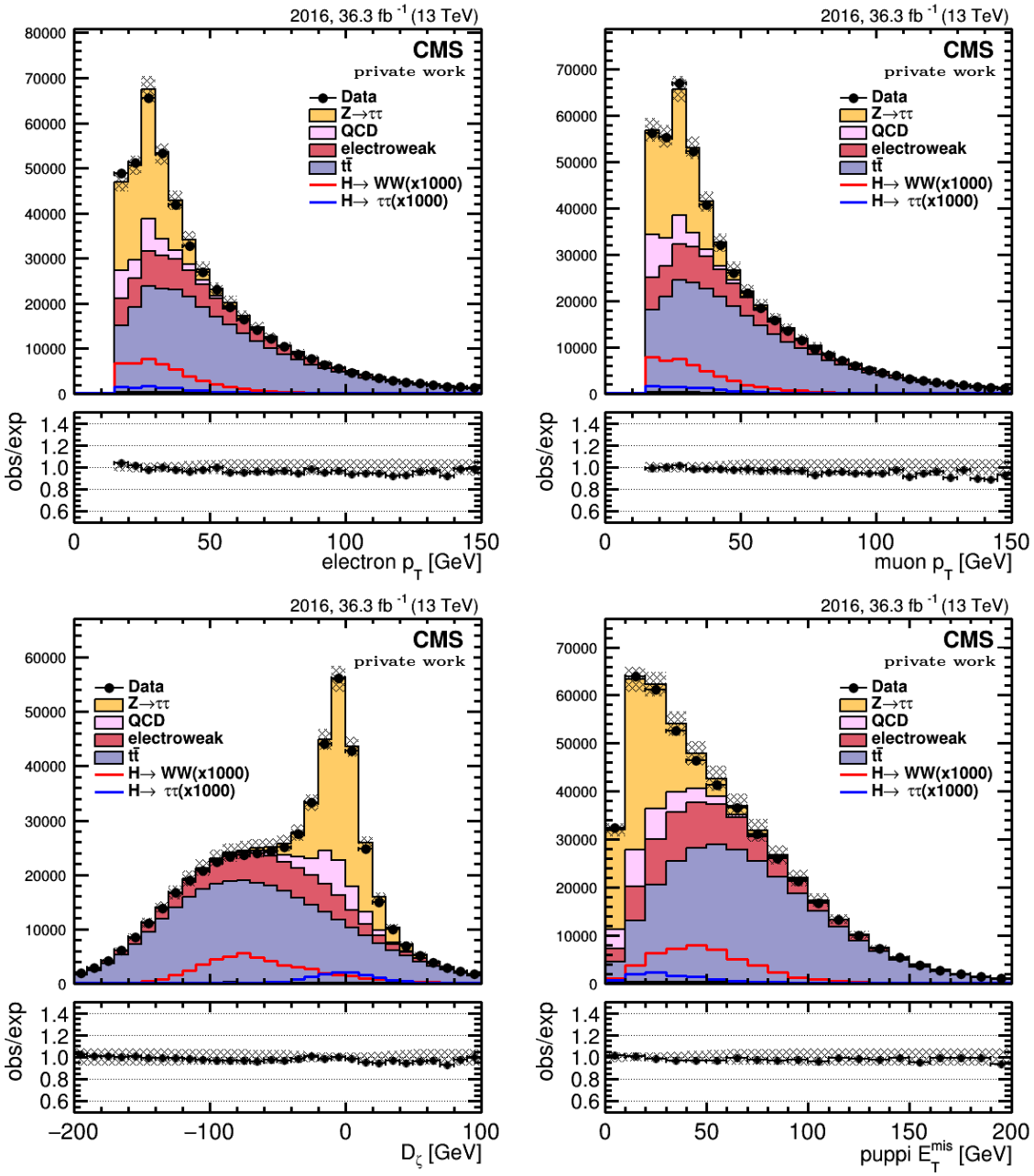


Figure A.2: Data and Monte Carlo distributions in $e\mu$ channel for 2016 data set. Upper-left: $p_T$ of electron , Upper-right: $p_T$ of muon, Lower-left: $D_\zeta$ , Lower-right: puppi $E_T^{miss}$.

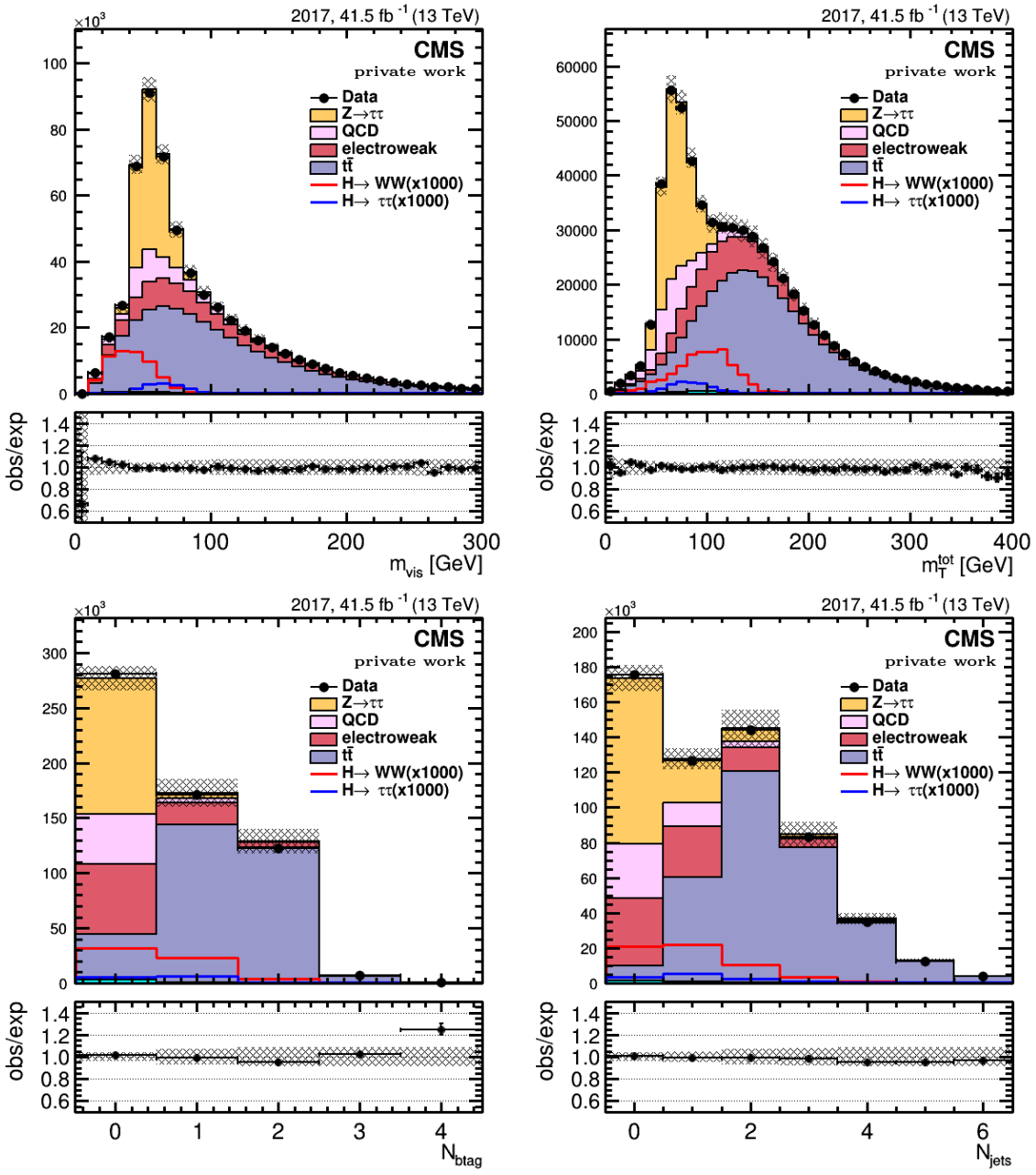


Figure A.3: Data and Monte Carlo distributions in $e\mu$ channel for 2017 data set. Upper-left: visible mass of the electron-muon, Upper-right: total transverse mass, Lower-left: b-tagged jet multiplicity, Lower-right: jet multiplicity.

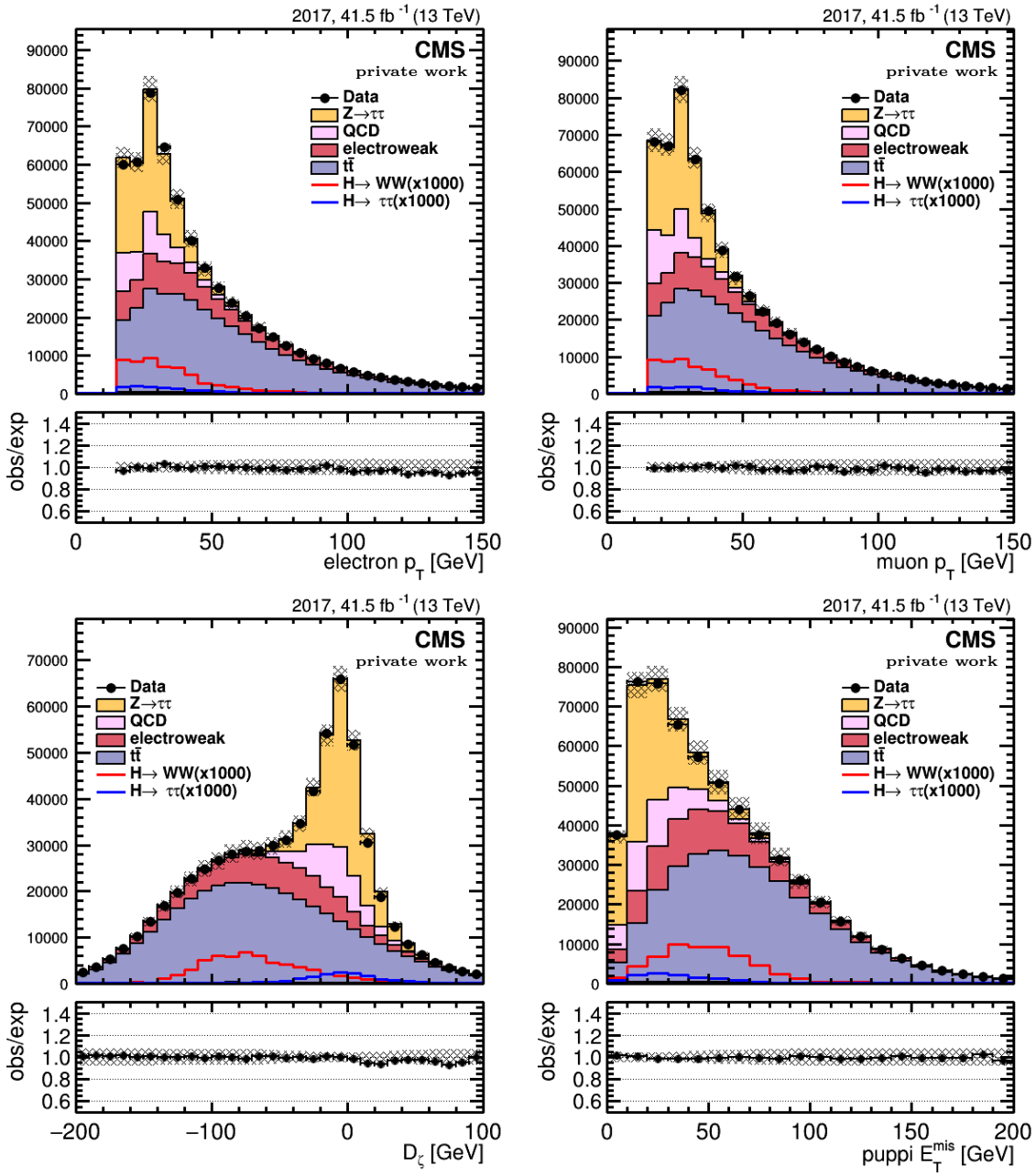


Figure A.4: Data and Monte Carlo distributions in $e\mu$ channel for 2017 data set. Upper-left: $p_T$ of electron , Upper-right: $p_T$ of muon, Lower-left: $D_\zeta$ , Lower-right: puppi $E_T^{miss}$.

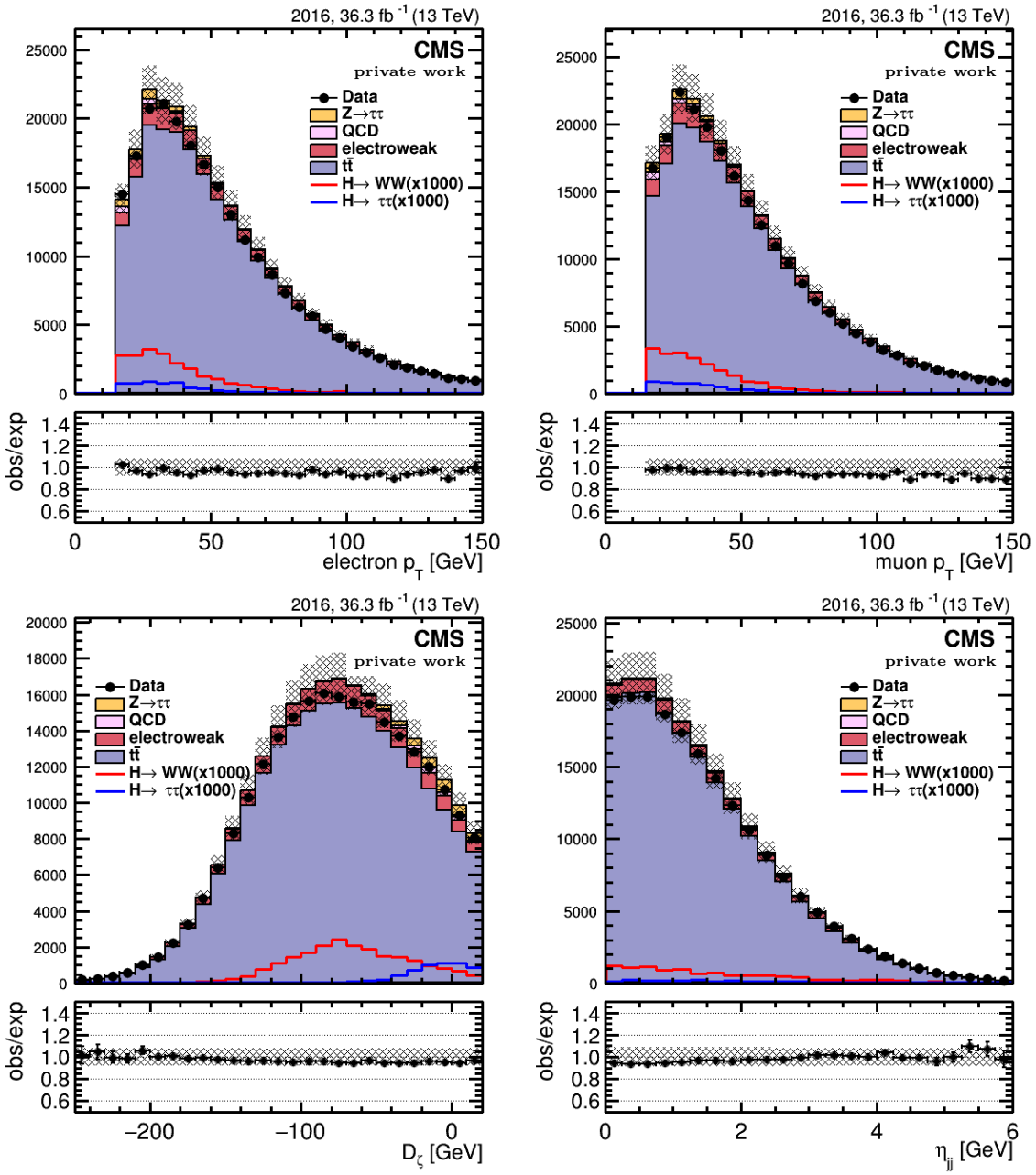


Figure A.5: Data and Monte Carlo distributions in the $e\mu$ channel for 2016 data set after requiring ($1 \leq N_{btag} \leq 2$). Upper-left: $p_T$ of electron , Upper-right: $p_T$ of muon, Lower-left: $D_\zeta$ , Lower-right: distribution of pseudorapidity between two leading jets.

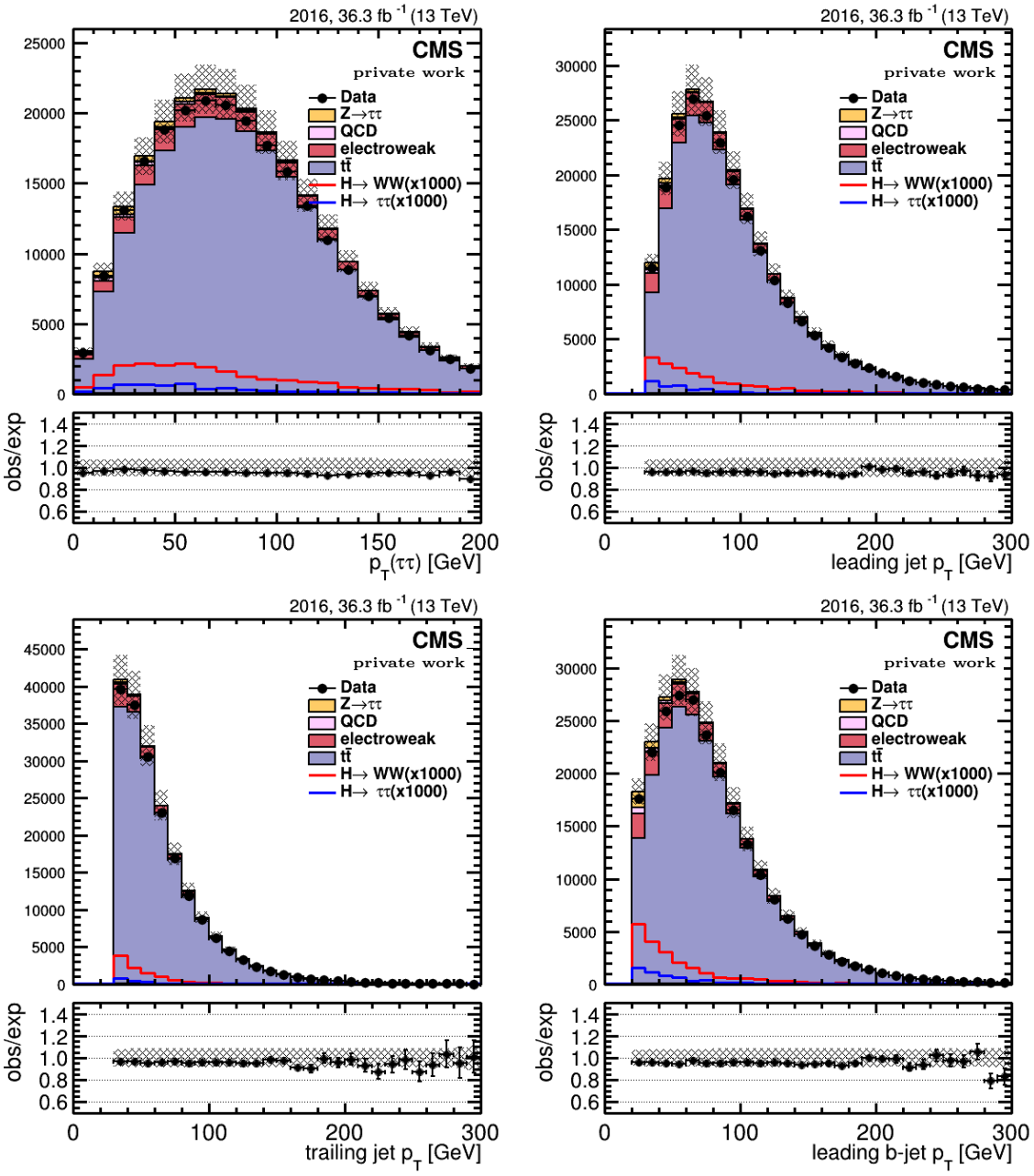


Figure A.6: Data and Monte Carlo distributions in the $e\mu$ channel for 2017 data set after requiring ($1 \leq N_{btag} \leq 2$). Upper-left: $p_T$ of Higgs candidate, Upper-right: leading jet $p_T$, Lower-left: trailing jet $p_T$, Lower-right: leading b-tagged jet $p_T$ distribution of pseudorapidity between two leading jets.

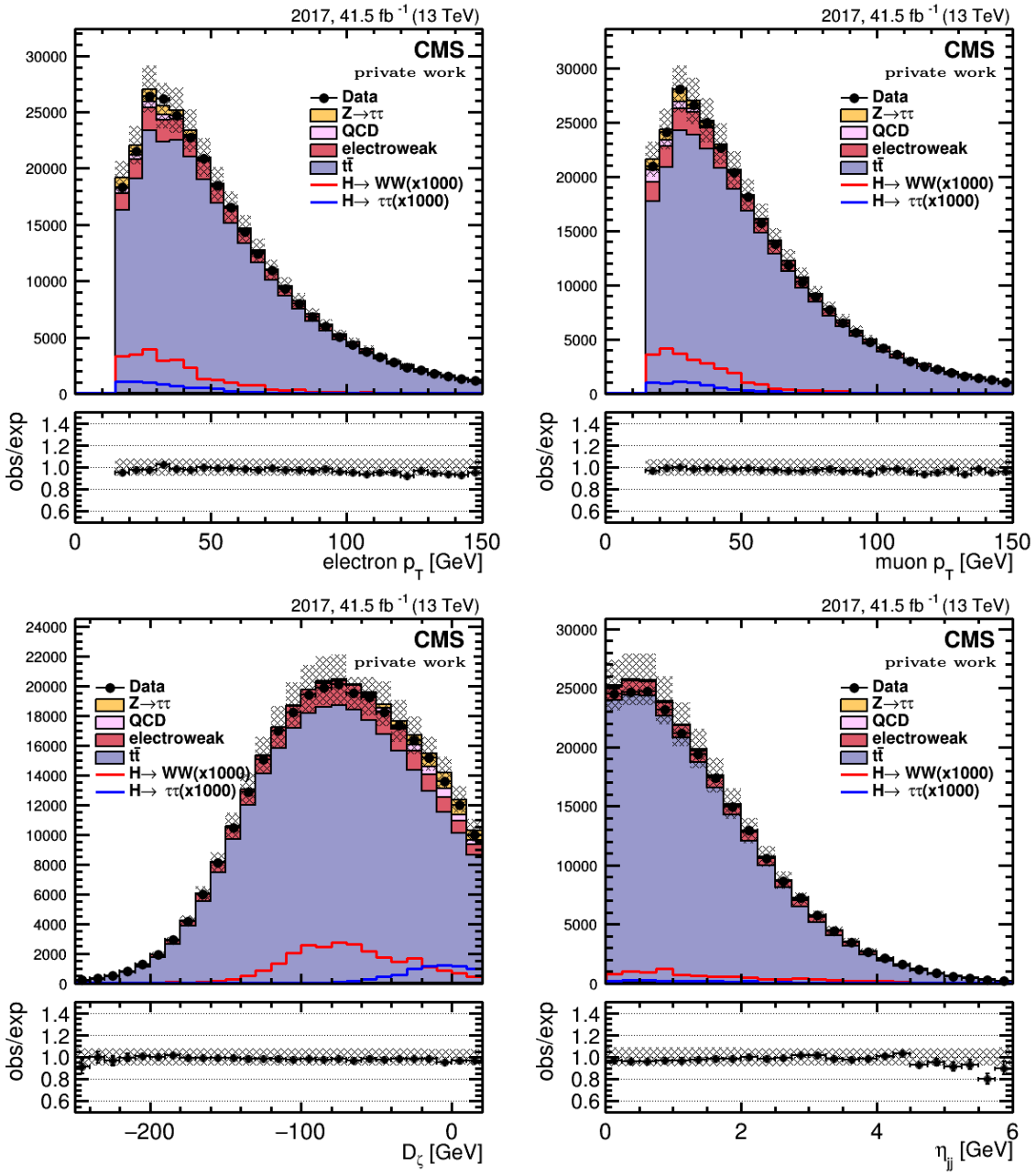


Figure A.7: Data and Monte Carlo distributions in the $e\mu$ channel for 2017 data set after applying inclusive selections ($1 \leq N_{btag} \leq 2$). Upper-left: $p_T$ of electron, Upper-right: $p_T$ of muon, Lower-left: $D_\zeta$, Lower-right: distribution of pseudorapidity between two leading jet.

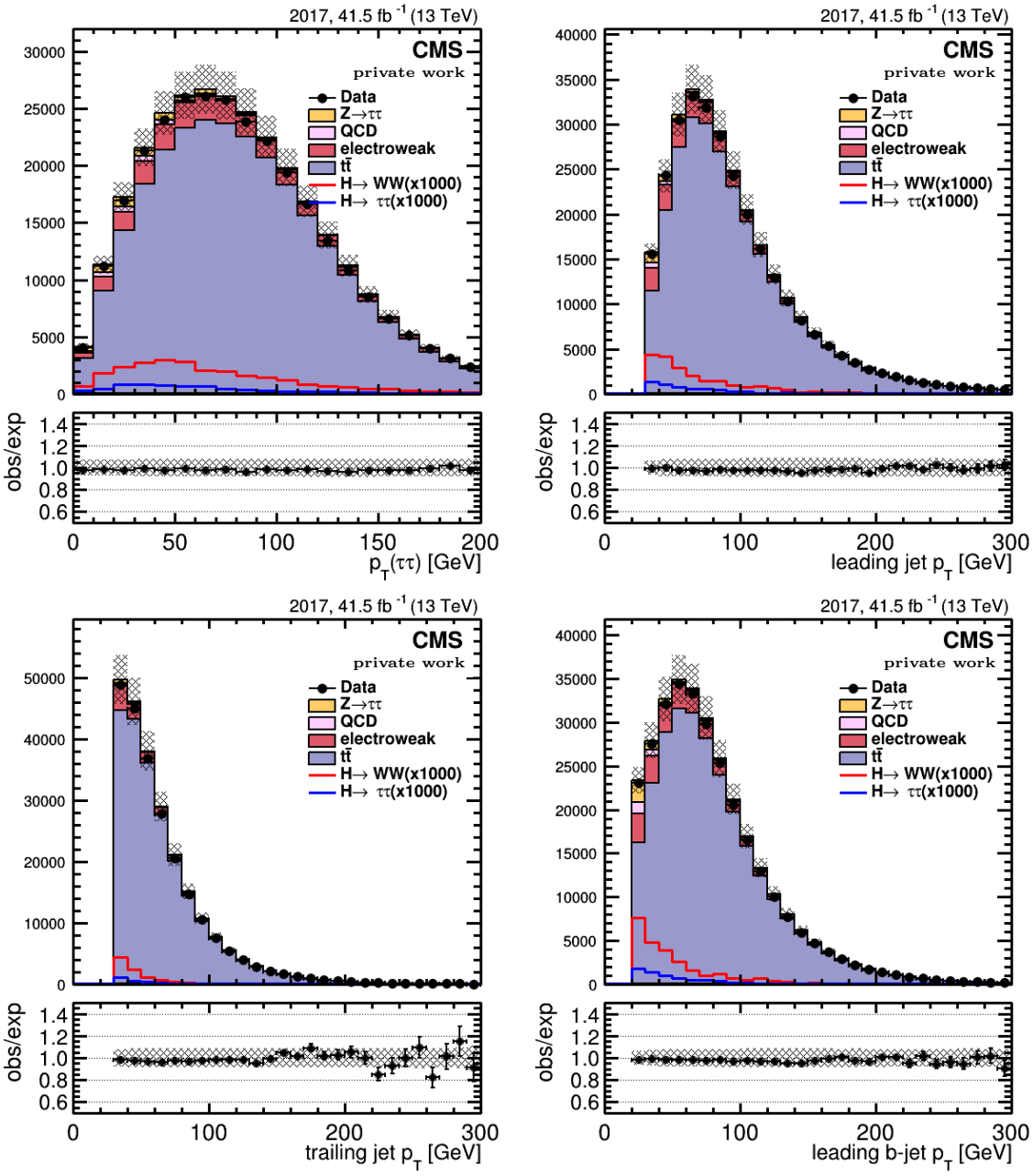


Figure A.8: Data and Monte Carlo distributions in the $e\mu$ channel for 2017 data set after applying inclusive selections ($1 \leq N_{btag} \leq 2$). Upper-left: $p_T$ of Higgs candidate, Upper-right: leading jet $p_T$, Lower-left: trailing jet $p_T$, Lower-right: leading b-tagged jet $p_T$ distribution of pseudorapidity between two leading jets.

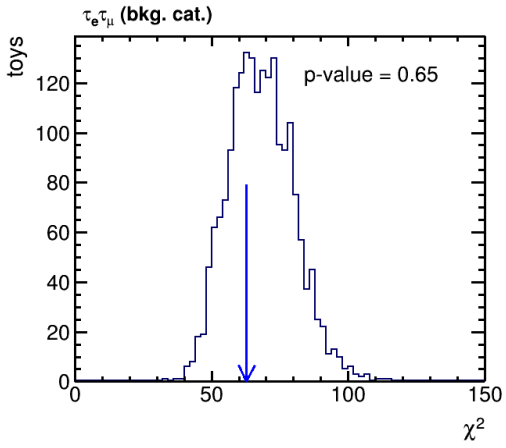


Figure A.9: Reuslts of GoF test for the $e\mu$ channel.

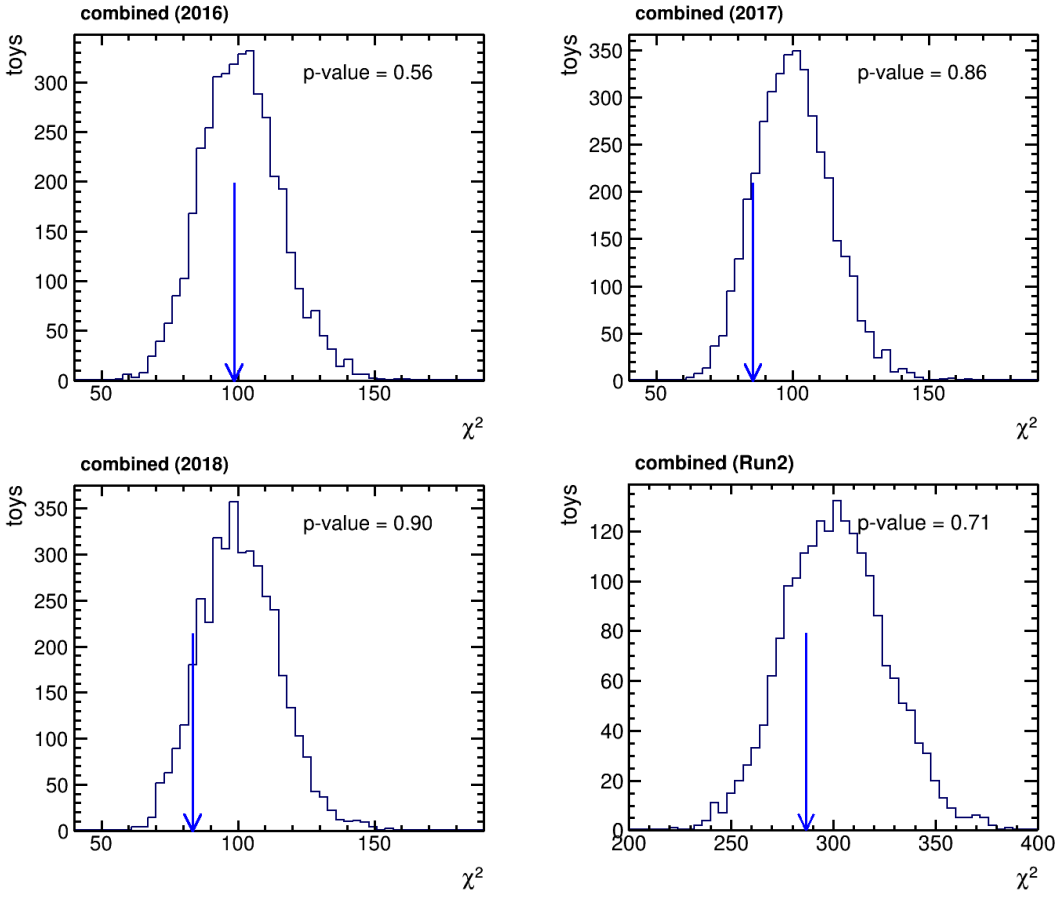


Figure A.10: Reults of GoF test for the 2016 (upper-left), 2017 (upper-right), 2018 (lower-left) and for the Run 2 combination (lower-right).

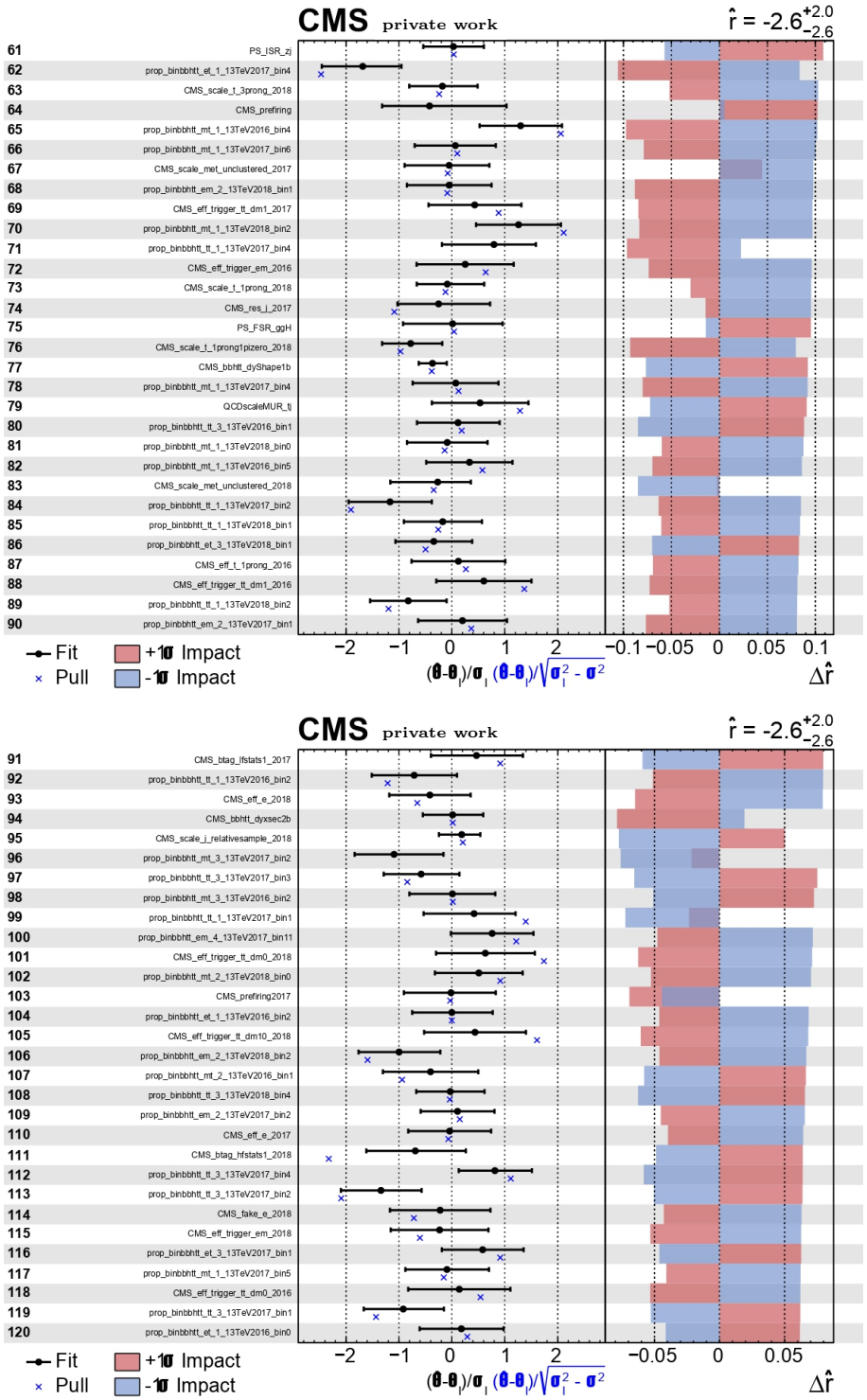


Figure A.11: The list of uncertainties with the largest impacts after combining all the semi- (fully-) leptonic and fully-hadronic channels for the whole Run 2 dataset.

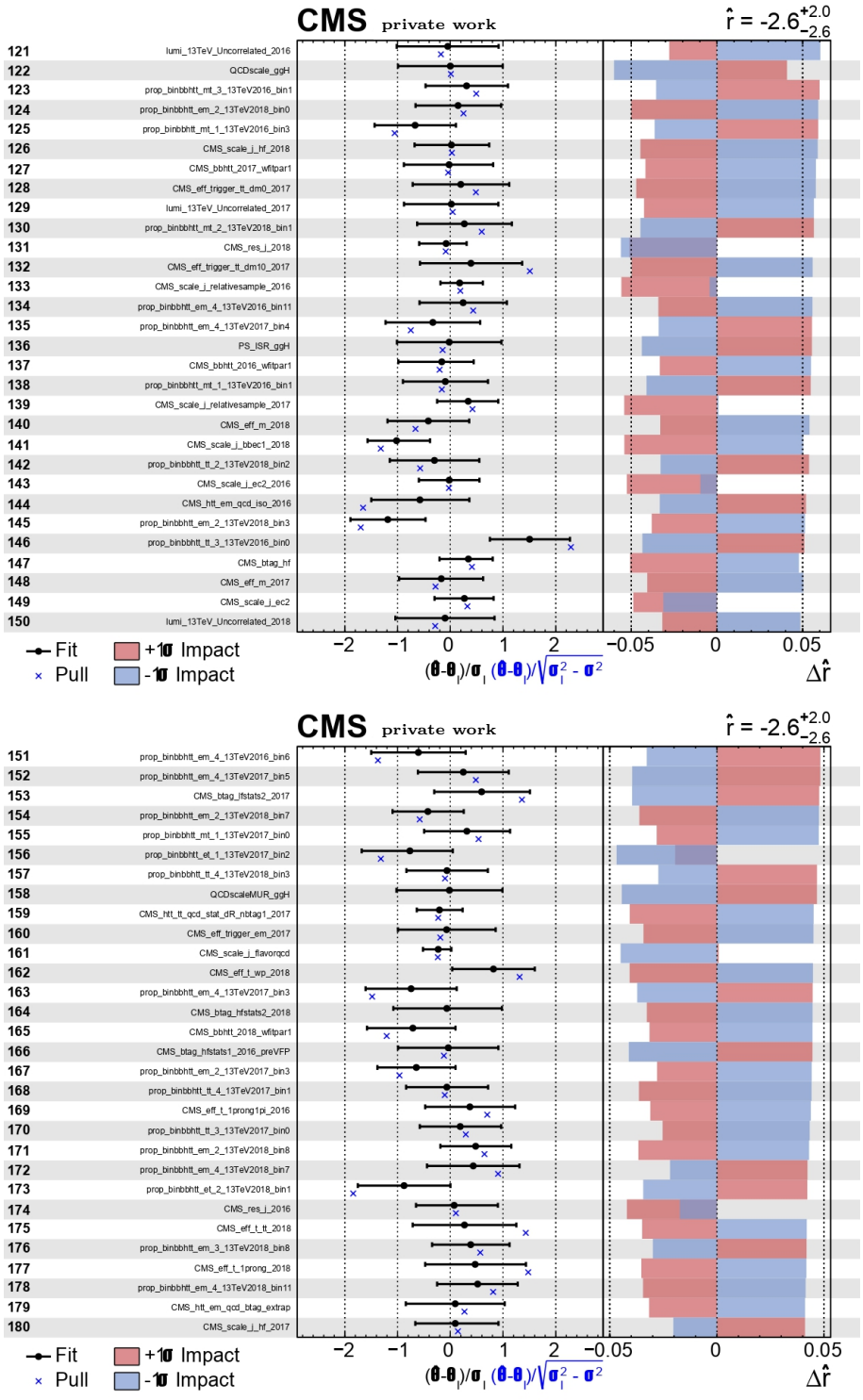


Figure A.12: The list of uncertainties with the largest impacts after combining all the semi- (fully-) leptonic and fully-hadronic channels for the whole Run 2 dataset.

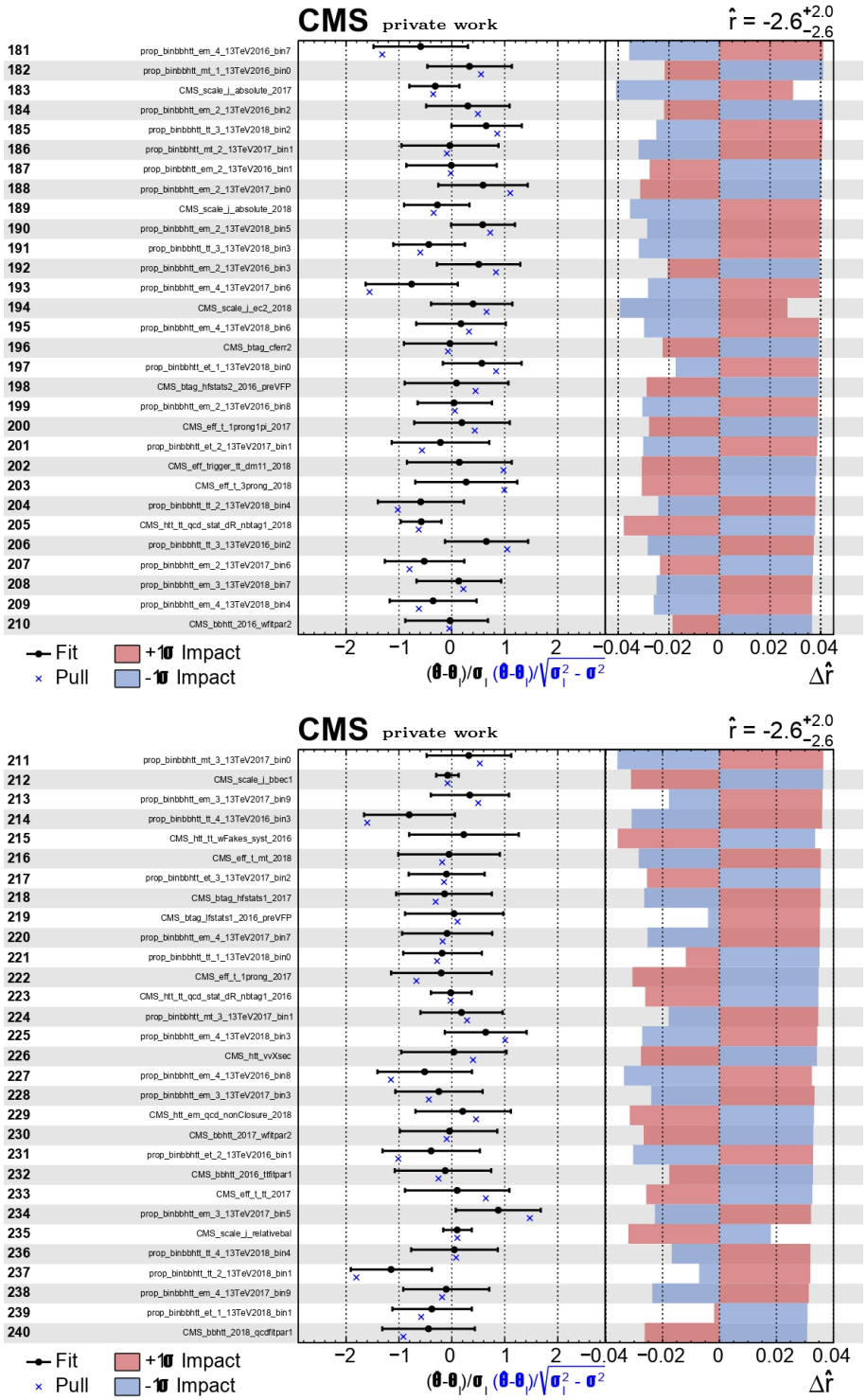


Figure A.13: The list of uncertainties with the largest impacts after combining all the semi- (fully-) leptonic and fully-hadronic channels for the whole Run 2 dataset.

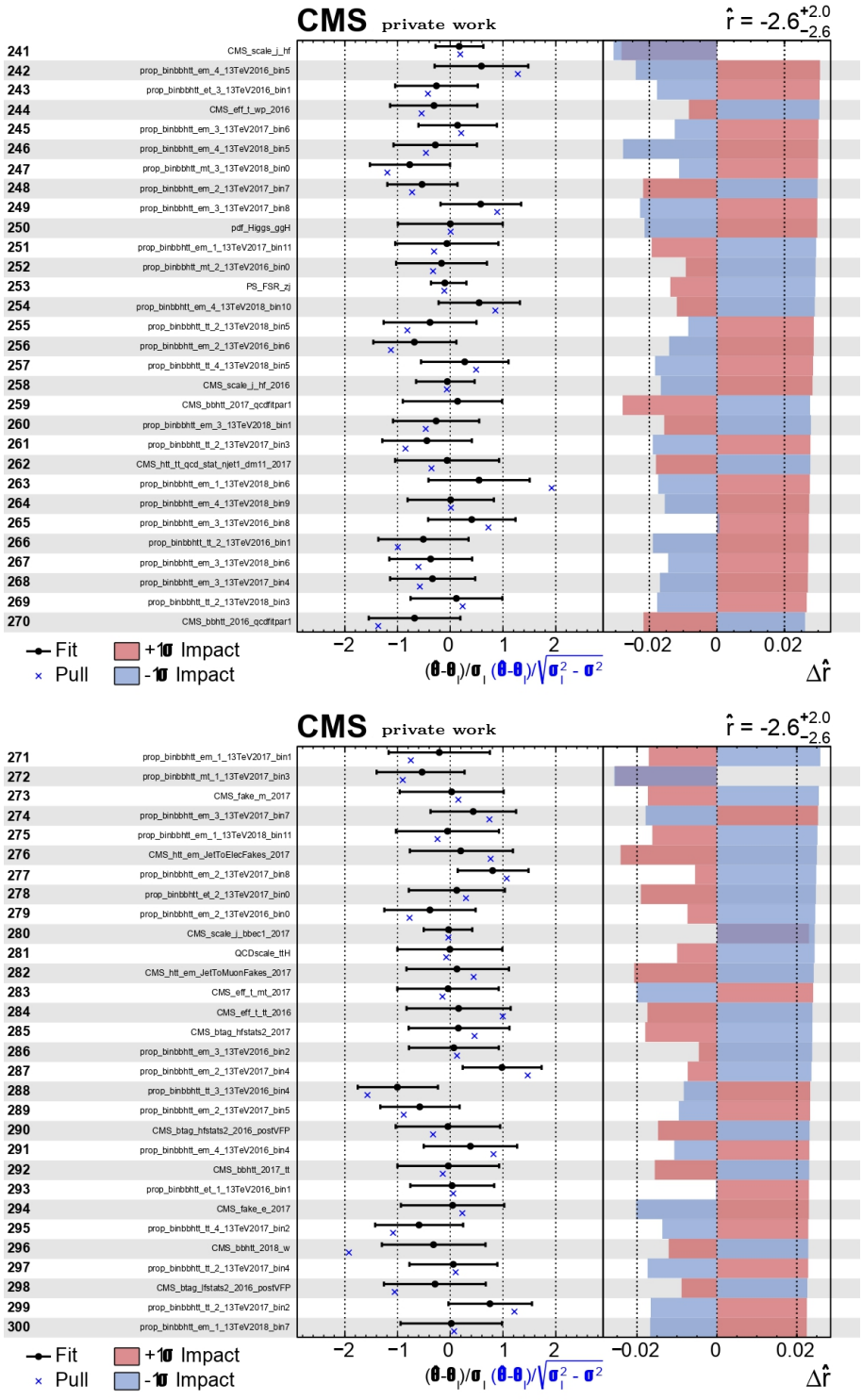


Figure A.14: The list of uncertainties with the largest impacts after combining all the semi- (fully-) leptonic and fully-hadronic channels for the whole Run 2 dataset.

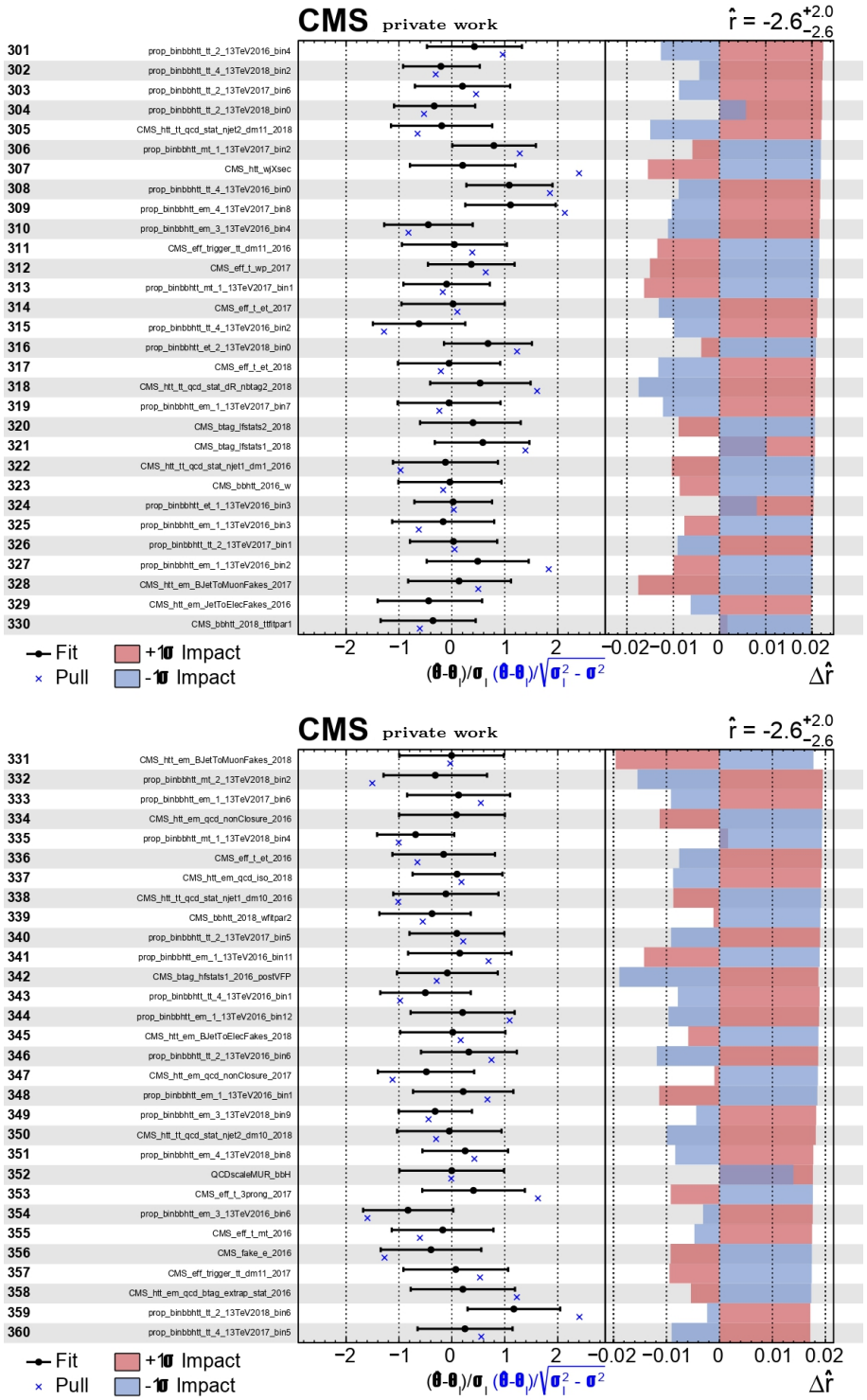


Figure A.15: The list of uncertainties with the largest impacts after combining all the semi- (fully-) leptonic and fully-hadronic channels for the whole Run 2 dataset.

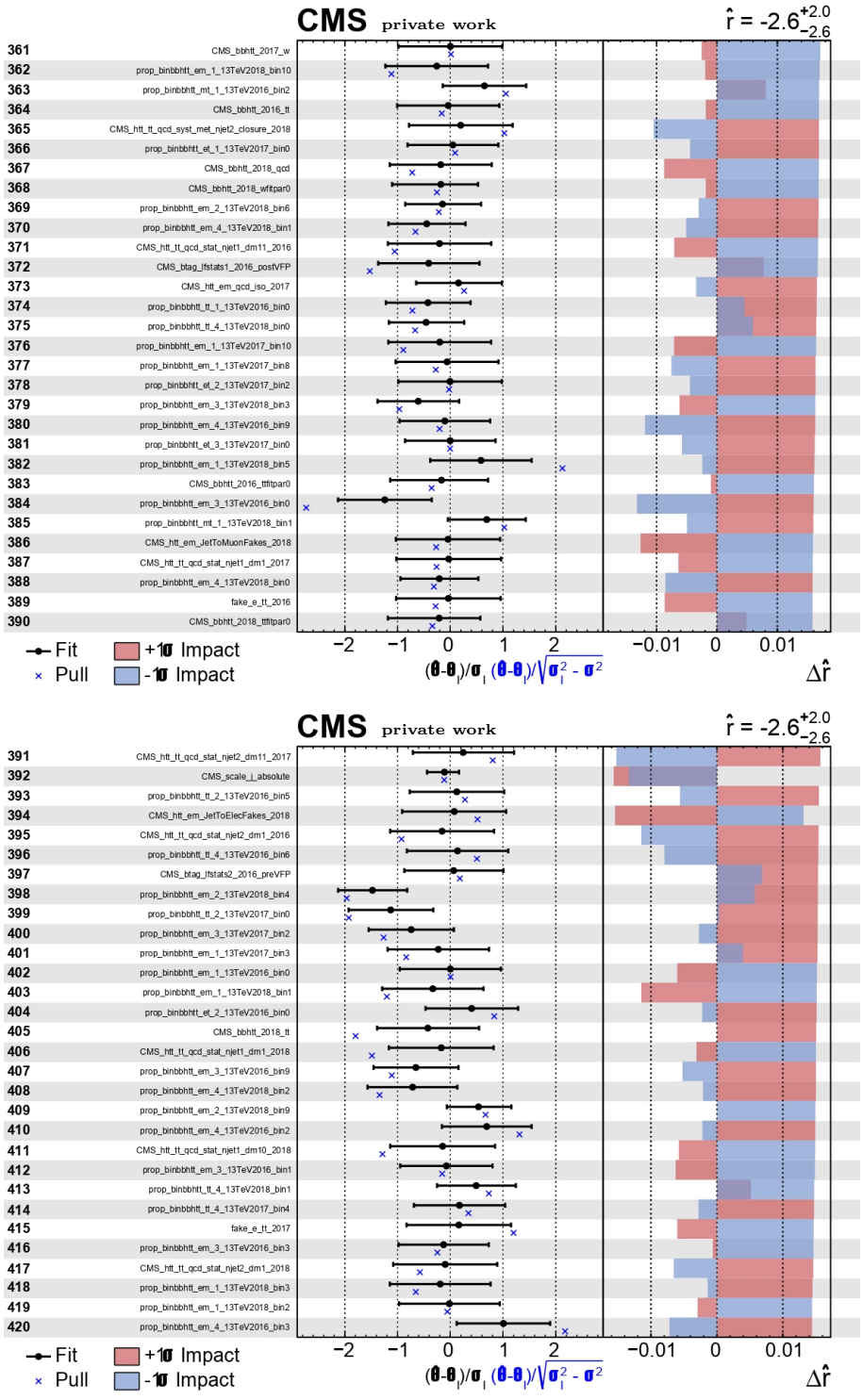


Figure A.16: The list of uncertainties with the largest impacts after combining all the semi- (fully-) leptonic and fully-hadronic channels for the whole Run 2 dataset.

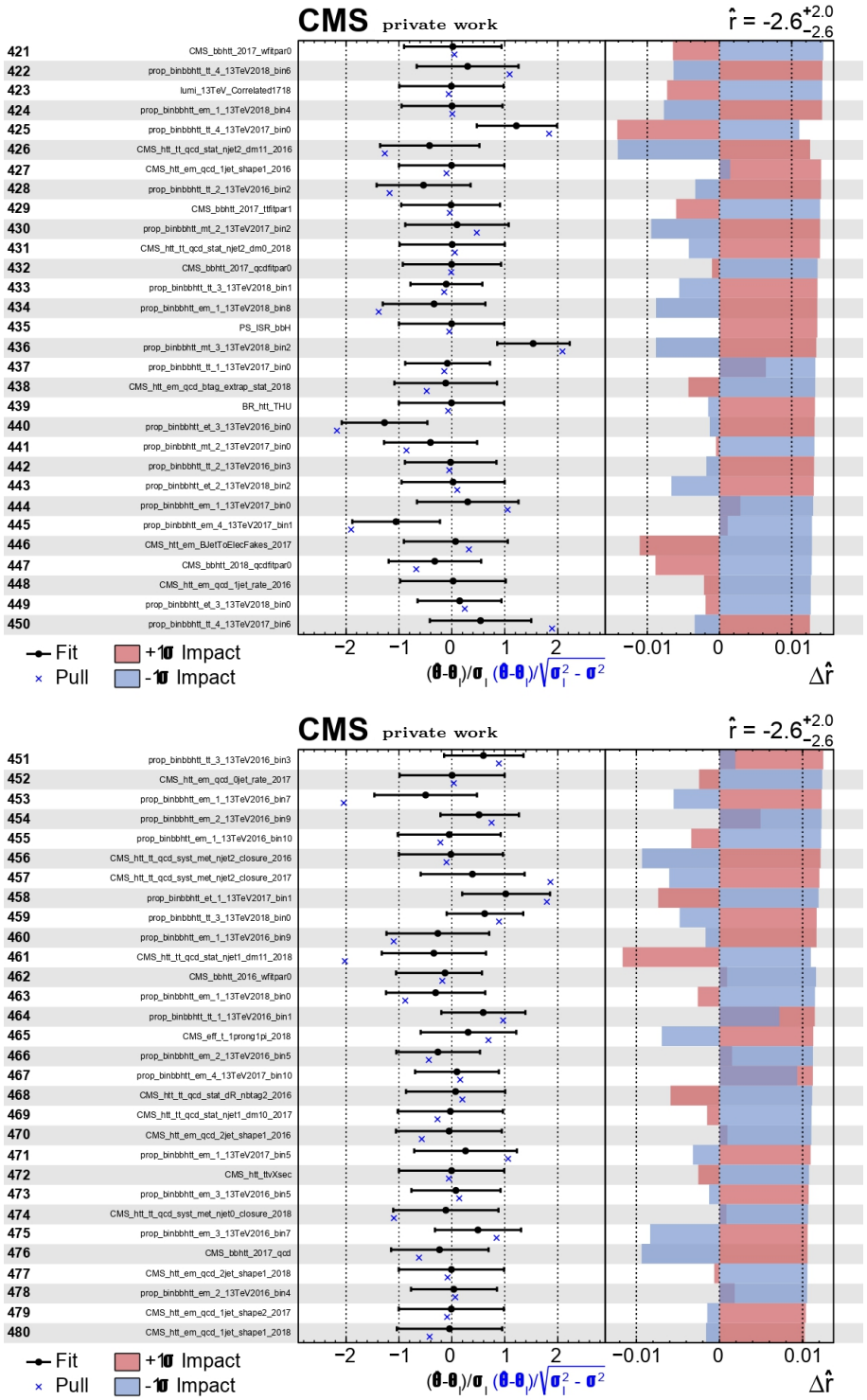


Figure A.17: The list of uncertainties with the largest impacts after combining all the semi- (fully-) leptonic and fully-hadronic channels for the whole Run 2 dataset.

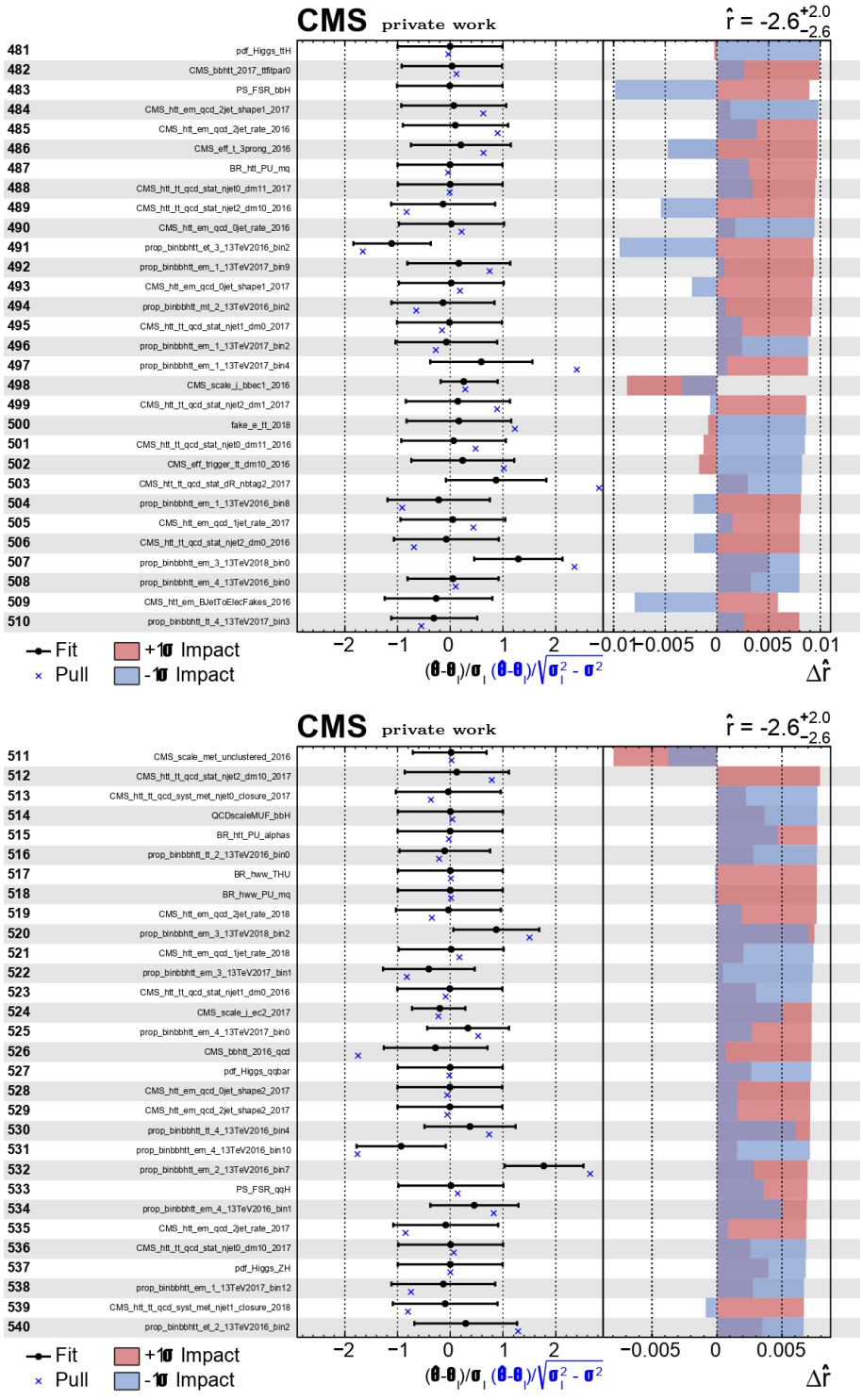


Figure A.18: The list of uncertainties with the largest impacts after combining all the semi- (fully-) leptonic and fully-hadronic channels for the whole Run 2 dataset.

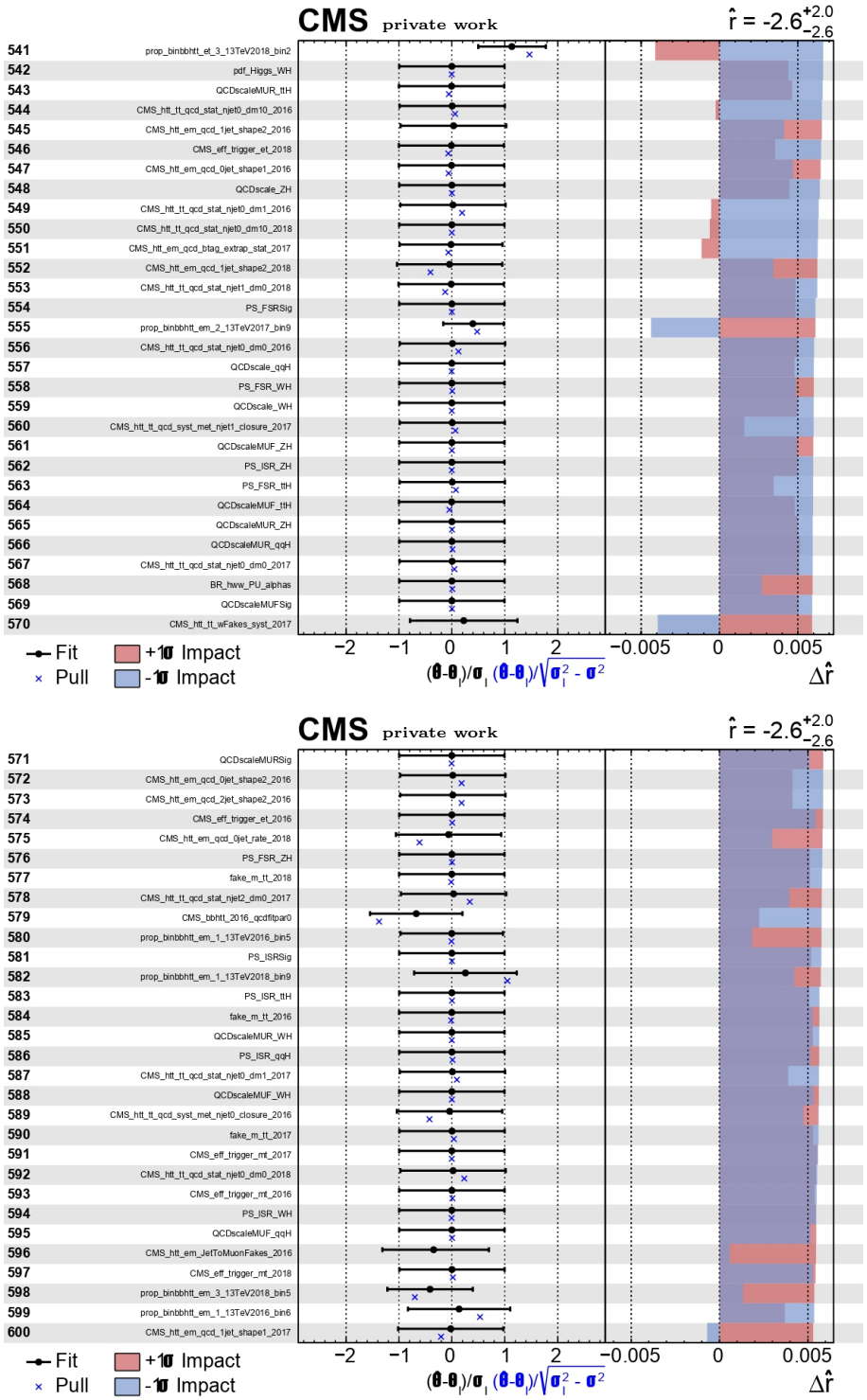


Figure A.19: The list of uncertainties with the largest impacts after combining all the semi- (fully-) leptonic and fully-hadronic channels for the whole Run 2 dataset.

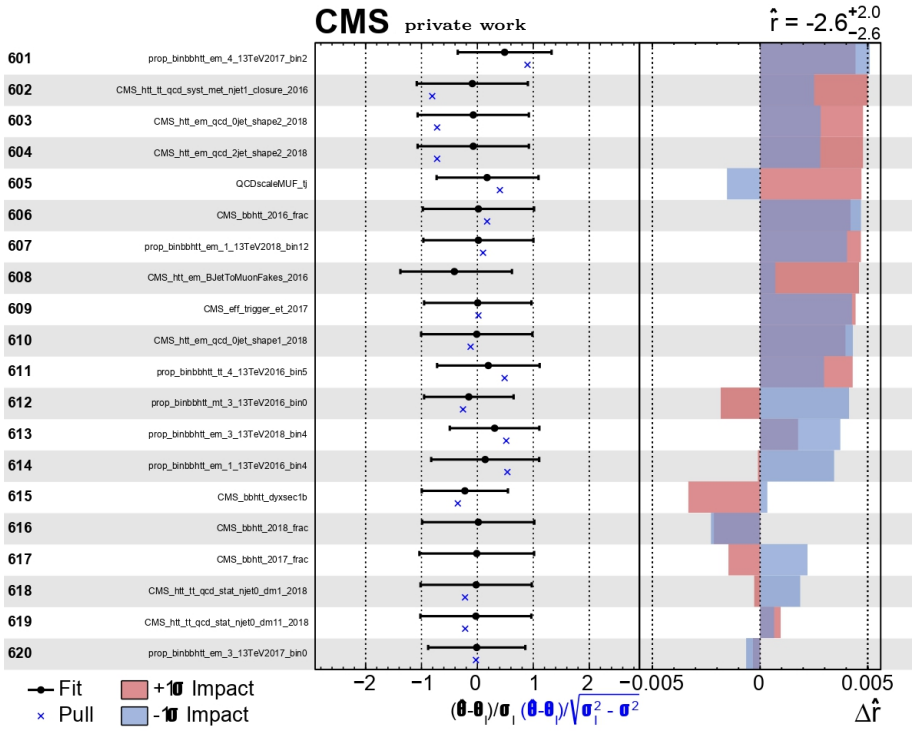


Figure A.20: The list of uncertainties with the largest impacts after combining all the semi- (fully-) leptonic and fully-hadronic channels for the whole Run 2 dataset.

APPENDIX B

# Supplementary material for ML

The classification task in this analysis was initially performed using Neural Networks, which was later substituted by a XGBoost model. The comparison between the performance of the preliminary NN model and the BDT is given in this appendix, followed by the additional training results performed during this thesis work.

## B.1 Neural Network

The multi-class Neural Network (NN) algorithm discriminates between signal and several background classes. Initially, due to the absence of the dedicated Standard Model samples for the present analysis, the NN classification was performed with the MSSM $H \rightarrow \tau\tau$ analysis samples as a preliminary test. The multi-classifier with NN was designed to separate the bbH signal from background classes $t\bar{t}$, Drell-Yan plus jets, ggH(Higgs decay to $\tau$) and a miscellaneous class, which includes W plus jets and single top combined. The classification task utilized a multilayer perceptron, specifically, a fully connected feed-forward neural network with two hidden layers comprising 50 and 100 nodes, respectively, where "Relu" served as the activation function. The softmax activation function was chosen for the output layer, with all event classes assigned equal weight during the training phase. Each batch of 100 events was used for training, employing cross-entropy as the loss function and Adam as the optimizer with a constant multiplicative learning rate of $10^{-4}$. The sample data was divided into three subsets, with the validation subset containing 25% of the training statistics. This subset was utilized to monitor the neural network's performance. The training is stopped if the loss evaluation shows no further improvement after 50 epochs. Since the samples were statistically limited, the NN performance was not optimal, and the training result was skewed towards the majority class $t\bar{t}$. Therefore, the SMOTE oversampling method was adopted for the two minority classes, bbH and ggH. The result of the training with NN before and after applying the SMOTE oversampling

technique is shown through the Confusion Matrix in Figure B.1.

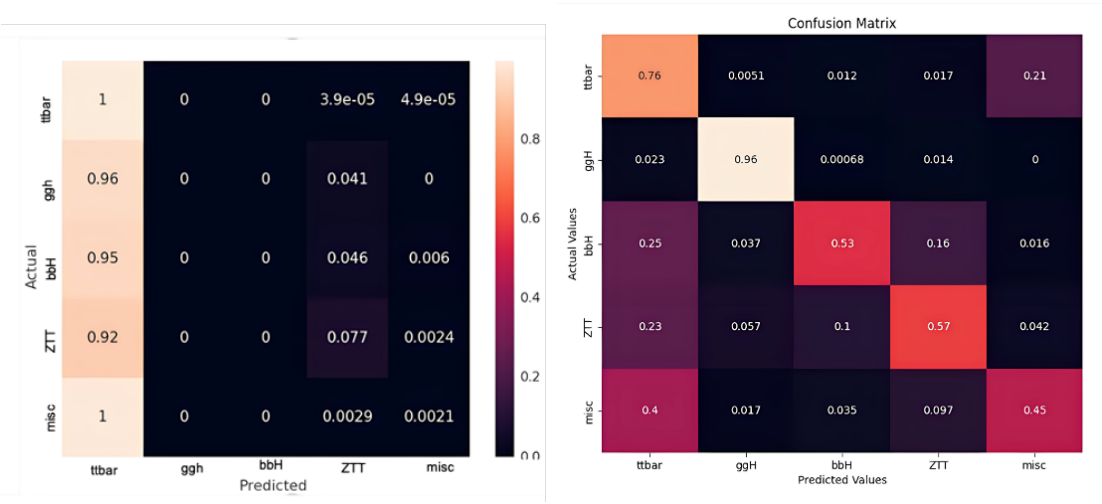


Figure B.1: Confusion matrix (CM) shows the actual value versus the predicted value by the NN model. High coefficient values on the diagonal axis indicate an optimal class separation. The initial training result is shown on the left, with no oversampling applied to the minority class bbH and ggH. The result is skewed towards the majority class $t\bar{t}$. On the right, the CM result after applying oversampling on the minority classes, which indicates a significant improvement in the classification power of the NN.

Furthermore, the result with the SMOTE oversampling technique was also tested with the XGBoost model and shown to perform significantly better than the NN oversampled result. This improvement is visible in the confusion matrix of the XGBoost model in Figure B.2, indicating reduced confusion between signal classes bbH and ggH and the miscellaneous background class.

As discussed in Chapter 5, XGBoost was later used to analyze this thesis work due to the performance improvement and lower optimization of the model hyperparameters concerning the multilayer perceptron. As soon as the dedicated simulated samples for this analysis were produced, the results were obtained without the oversampling technique and only by balancing the weight, resulting in an overall good performance.

## B.2 Confusion matrices

In this work, several class strategies and different combinations of input variables were tested to achieve the desired result. The confusion matrixes and the corresponding Receiver- Operating Characteristics (ROC) curves, a metric displaying the true versus the false positive rate, can be found in this section. The decision of which class and set

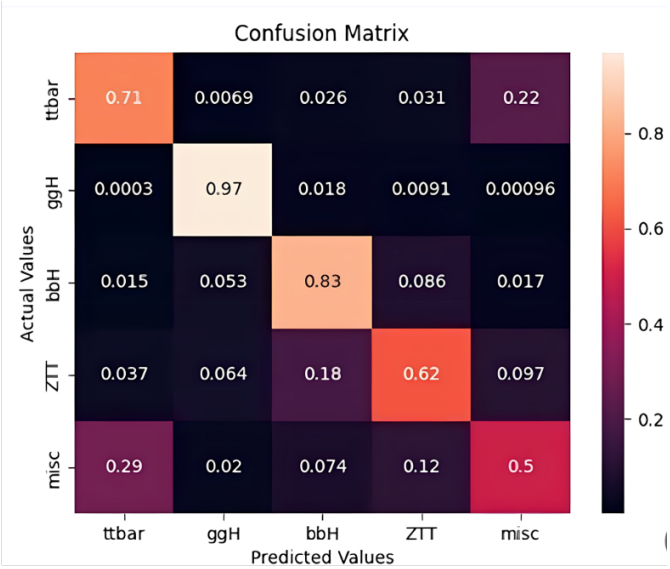


Figure B.2: Confusion matrix of the training results with the XGBoost model with oversampled signal classes.

of input variables to keep was not based solely only on the output of the training and classification power evinced from the confusion matrix but also on the trade-off between the latter, the pull distribution quantifying the impact of the systematic uncertainties, and the goodness-of-fittest derived after the likelihood fit. It is important to note that no obvious differences in classification performance for different data-taking periods have been observed, as the classification result is similar among all the years from 2016 to 2018.

## B.3 Shapley Additive Explanations

As previously discussed, the SHAP method was used to study the effect of input features on the model's output. The ranking shown in Figure B.5 is one of the many available plots the SHAP library provides to study the model's output. The so-called summary plot is another set of plots that can help analyze the output of the model. These are shown in Figure B.6. The summary plot combines feature importance with feature effects. Each point on the summary plot is a Shapley value for a feature and an instance. The position on the y-axis is determined by the feature, and on the x-axis, it is determined by the Shapley value. The colour scale represents the value of the feature from low (blue) to high (red), and the features are ordered according to their importance. In the summary plot, we see the first indications of the relationship between the value of a feature and the impact on the model prediction.


Figure B.3: Figure shows different confusion matrices trained for different class structures for different eras.

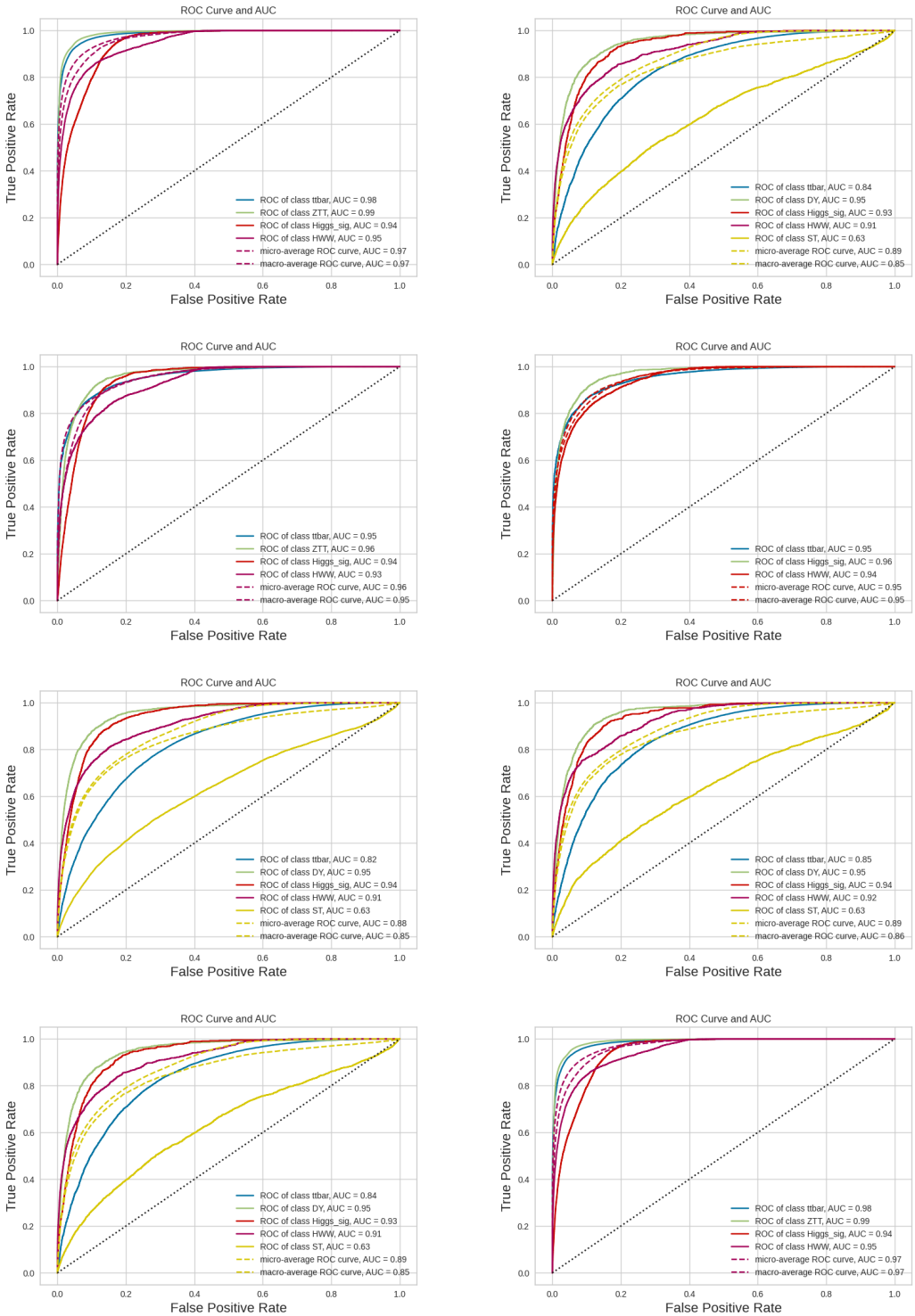


Figure B.4: Shown are ROC curves for different classification tasks and eras.

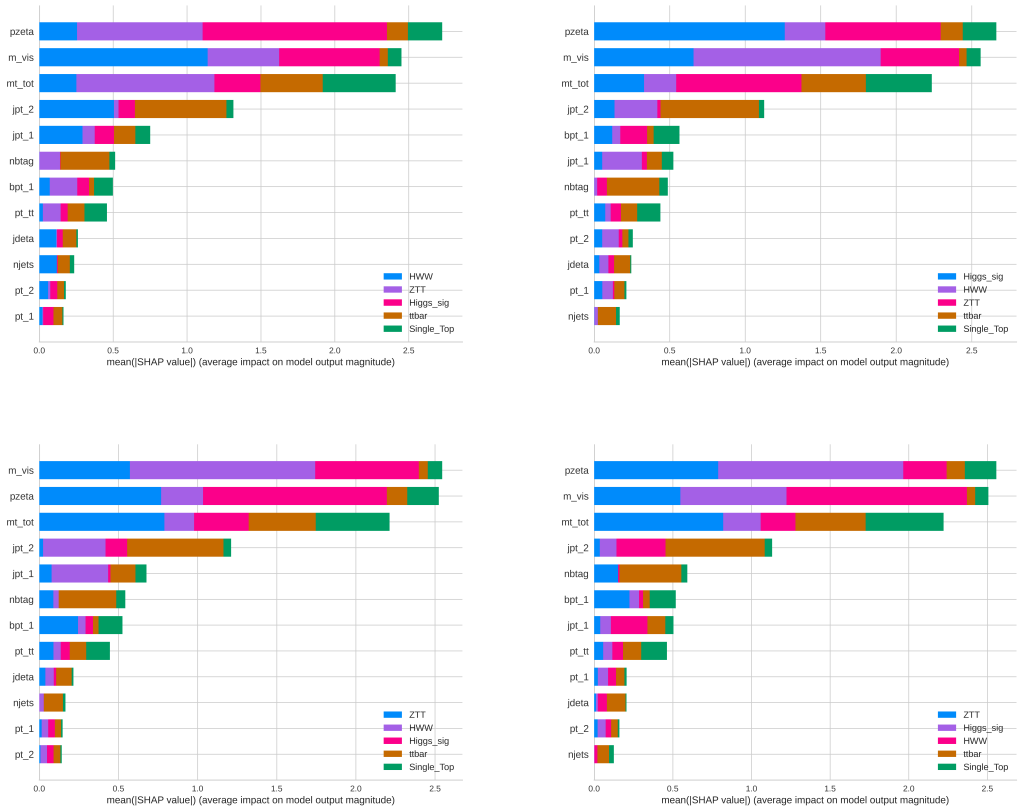


Figure B.5: SHAP plot showing the ranking of the variables used for the training task and the power of the variable employed for the classification task.

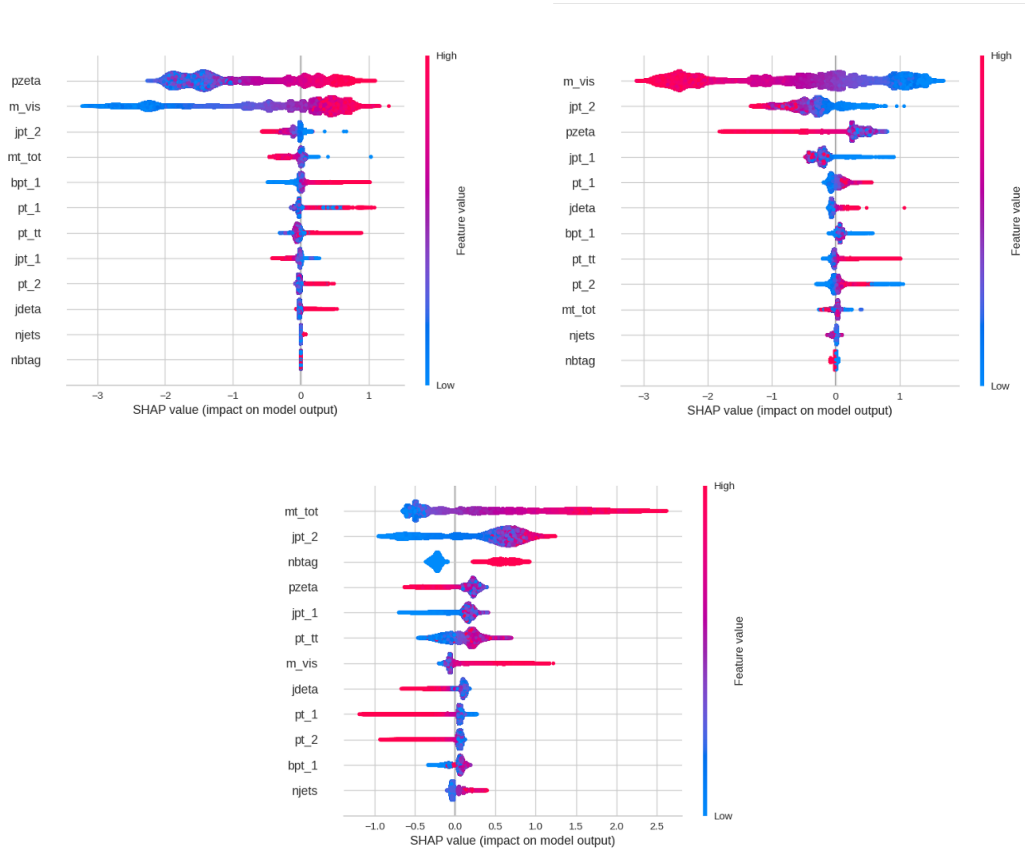


Figure B.6: The summary plots of three different classes: bbH signal class (top left), HWW signal class (top right) and $t\bar{t}$ background class (bottom).

## B.4 Results of hierarchical training

As discussed in chapters 5 and 6, hierarchical techniques were employed to optimize the analysis result in the $e\mu$ channel. The analysis of bbH takes into account the production of bbH as inclusive as possible; this means the bbH production signal process is considered to be the bbH ($y_b$) and the gluon splitting production of bbH ($y_t$). However, to constrain the bottom Yukawa coupling, one needs a better separation between these two processes and possibly only takes the bbH ($y_b$) process as the signal process. With the current statistics and theoretical uncertainties, as discussed in Chapter 7, considering the inclusive signal process is more meaningful. However, the hierarchical training was employed as a test to try the optimization for further separation of these two main bbH processes. The result has been studied only in the $e\mu$ channel.

Several hierarchies have been tested; the most efficient was the following structure. Firstly, the first level of training contains a binary classification which separates all the signal events from all the background events. The second-level model then uses the output of the first model as an additional input to separate all the subclasses from one another. The second level model is designed to separate the following classes from one another:

1. $t\bar{t}$ & Single top
2. Drell-Yan jets
3. Higgs decay to W bosons ($y_b$)
4. Higgs decay to W bosons ($y_t$)
5. Higgs decay to tau leptons ($y_t$)
6. Higgs decay to tau leptons ($y_b$)

The result of the training is shown for the second level training in the confusion matrix B.7.

The overall training results show an improvement regarding the nominal training used to retrieve the analysis result. Calculating the limits for all the years and the combination of Run 2 data gives the following expected (observed) limits listed in table B.1. These first set of limits are derived using the same signal model. as the initial analysis result. This means deriving the upper limits by including the $y_b$ and $y_t$ components as signal processes. This is done to have a better comparison with the original training. In this case, the derived limits are presented in table B.1.

Comparing these limits with the ones derived from the flat classification, a slight deterioration of the limits has been observed from the 19.1 median expected for emu to 26.1 for the Run2 data set. The upper limits are also obtained by considering only

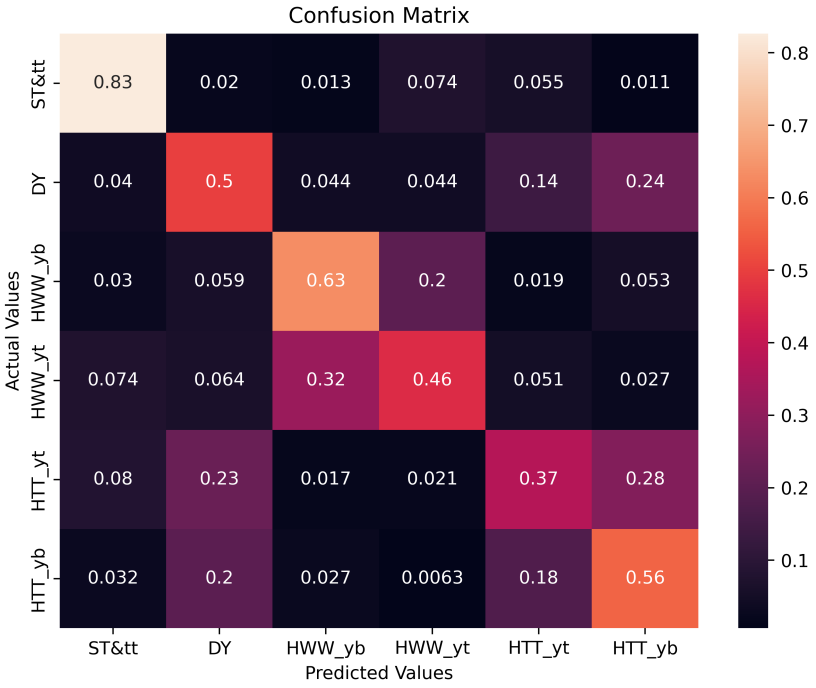


Figure B.7: Figure shows the confusion matrix for the training of the second level model to separate the bbH signal processes considering different amplitudes $y_b$ and $y_t$.

| $e\mu$ **channel** | | | | | | |
|---|---|---|---|---|---|---|
| | $-2\sigma$ | $-1\sigma$ | Median | $+1\sigma$ | $+2\sigma$ | **Observed** |
| 2016 | 22.1 | 31.4 | 48.8 | 81.2 | 136.5 | 61.7 |
| 2017 | 19.5 | 27.6 | 42.7 | 71 | 117.5 | 45.2 |
| 2018 | 16.2 | 23.04 | 35.9 | 60.2 | 100.4 | 34.6 |
| **Run 2** | 11.7 | 16.6 | 26.1 | 43.4 | 72.4 | 26.6 |

Table B.1: Expected (observed) upper limits derived on bbH production in $e\mu$ channel with the hierarchical training.

| $e\mu$ **channel** | | | | | | |
|---|---|---|---|---|---|---|
| | $-2\sigma$ | $-1\sigma$ | Median | $+1\sigma$ | $+2\sigma$ | **Observed** |
| 2016 | 22 | 31.4 | 48.7 | 81.3 | 136.2 | 61.7 |
| 2017 | 19.3 | 27.5 | 42.7 | 71 | 117.5 | 45.2 |
| 2018 | 16.1 | 23 | 35.9 | 60 | 100.4 | 34.5 |
| **Run 2** | 11.7 | 16.6 | 26 | 43.4 | 72.4 | 26.6 |

Table B.2: Expected (observed) upper limits derived on bbH production in $e\mu$ channel with the hierarchical training considering only the bbH($y_b$) component in signal model.

| $e\mu$ **channel** | | | | | | |
|---|---|---|---|---|---|---|
| | $-2\sigma$ | $-1\sigma$ | Median | $+1\sigma$ | $+2\sigma$ | **Observed** |
| **Run 2** | 22.4 | 31.4 | 47.7 | 75.2 | 118.2 | 44.5 |

Table B.3: Expected (observed) upper limits derived on bbH production in $e\mu$ channel with the flat training (nominal training used for the analysis result) considering only the bbH($y_b$) component in signal model.

the bbH ($y_b$) component as a signal and considering the gluon splitting (bbH($y_t$)) as a background while performing the fit. These are presented in the table B.2.

As a result from B.2 indicates the sensitivity to bbH($y_b$), only production can be increased by using the hierarchical training. A combination of this training with more data can lead to a significant improvement in the result. Testing this method on other channels can also enhance the final result. The provided results in B.1 and B.2 show that the sensitivity is coming from the bbH($y_b$) component since the comparison of the limits with two different signal models does not show any deterioration or a significant improvement.

Furthermore, for a better comparison and to ensure that the hierarchical method can improve the sensitivity to bbH($y_b$), the limits with the flat classification are also calculated with the signal model that only considers the bbH($y_b$) component and ggH($y_t$) as a background, these results are shown in B.3.

APPENDIX C

# Alignment of the CMS tracker

In this section, an overview of the performance of the CMS Tracker during Run3 is given [131].

PROCEEDINGS OF SCIENCE

# Performance of the CMS Tracker during Run 3

**Daina Leyva Pernia**[a] **and Maryam Bayat Makou**[a,*] **for the CMS collaboration**

[a] *Deutsches Elektronen Synchrotron (DESY),*
*Notkestraße 85, Hamburg, Germany*

*E-mail:* daina.leyva.pernia@desy.de, maryam.bayat.makou@desy.de

The current CMS silicon tracker consists of two tracking devices: the inner pixel and the outer strip detectors. The tracker occupies the region around the center of CMS, where the LHC beams collide, and therefore, operates in a high-occupancy and high-radiation environment produced by the particle collisions within the LHC tunnel.

This article provides an overview of the excellent performance of the CMS silicon tracker during the ongoing Run 3 data-taking period. It discusses the behavior of local observables, such as hit reconstruction efficiency, their response to the accumulated integrated luminosity, and the precision achieved in aligning the detector components.

*European Physical Society Conference on High Energy Physics (EPS-HEP)*
*20-25 August 2023*
*Hamburg , Germany*

*Speaker



https://pos.sissa.it/

PoS(EPS-HEP2023)591

## 1. The CMS tracker detector

Comprising 1856 silicon pixel detector modules and 15 148 silicon strip detector modules, the CMS tracker [1, 2] plays a crucial role in physics research. The Pixel detector, located closest to the interaction point, is particularly susceptible to radiation damage. Its modules are arranged in four cylindrical layers around the beampipe and three endcap disks on each side of the detector. It is surrounded by the Silicon Strip detector, which features ten cylindrical layers and twelve endcap disks. Together, they deliver robust tracking and contribute with a pivotal role in CMS vertex reconstruction. This article highlights the remarkable performance attained in the face of challenging conditions during the Run 3 LHC data-taking period, including managing up to 62 interactions per beam crossing.

## 2. Tracker detector performance during Run 3

### 2.1 Pixel detector performance

Being the closest component to the interaction point, the Pixel detector is much more likely to suffer from radiation damage effects. These can lead to inefficiencies or instabilities, impacting the data quality. Consequently, during the second LHC Long Shutdown (LS2) period, from 2018 to 2022, the Pixel detector was extracted for a series of improvements and refurbishments [3]. This includes the installation of a whole new pixel barrel layer to replace the one nearest to the interaction point and the repair of modules and electronics in the other layers and disks. A measure of its performance during the present data-taking period is shown in Fig. 1 (left), showing the hit efficiency with instantaneous luminosity during Run 3 [4]. The distribution exhibits rather stable performance, which slightly deteriorates towards larger instantaneous luminosity for all layers, with the layer one efficiency being the most affected. This is mostly caused by the saturation of the readout buffer in the chips [5]. The improvements in preparation for Run 3 allowed for a hit efficiency higher than 96% at $22 \times 10^{33}$ cm$^{-2}$s$^{-1}$. A summary of the hit efficiency in the barrel layers for the delivered integrated luminosity in Run 3 is shown in Fig. 1 (right). Here again, Layer 1 efficiency decreases rather rapidly with accumulated radiation. The effect can be partially recovered by increasing the application voltage for the sensors and through continuous calibrations [6], which can be seen as the discontinuities where the efficiency increases rapidly in the figure.

### 2.2 Silicon Strip detector performance

The integrity of the strip detector is essential for data-taking. The stability during Run 3 can be seen in the fraction of bad module components trend with the integrated luminosity, reflecting the integrity of its components to maintain an excellent tracking performance. This trend is shown in Fig. 2 for 2022 and 2023 proton-proton collisions [7], showing a rather stable trend, with a fraction of active channels of about 96%. The jumps at 205 fb$^{-1}$ are caused by the recovery of a cooling loop on the endcap region. Furthermore, some of the module power supplies in the Tracker Inner Disks were turned off during 2023 because of technical issues with the Front-End Drivers after 245 fb$^{-1}$[7]. As can be seen, the trend returned to usual values after the power to the modules was restored. Overall, no major issues have affected data quality.

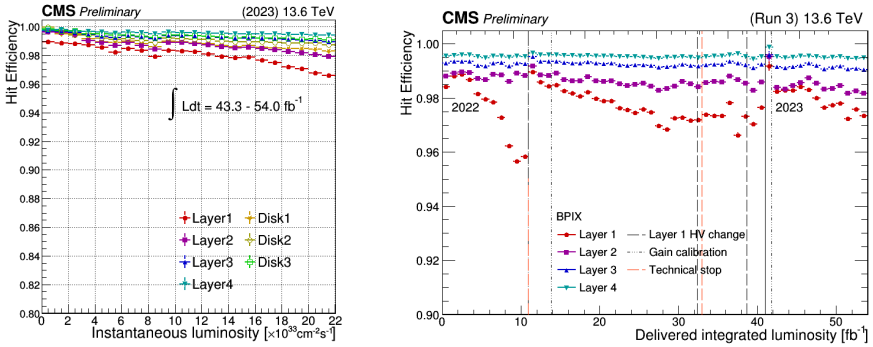


**Figure 1:** Pixel detector Hit Efficiency vs Instantaneous Luminosity during data-taking runs in May and June 2023 (left) and vs delivered integrated luminosity during Run 3 (right).

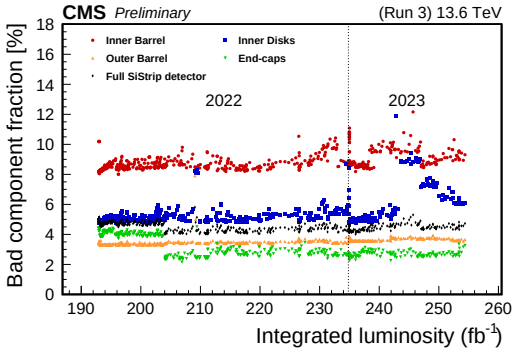


**Figure 2:** Evolution of fraction of modules flagged as bad vs delivered integrated luminosity during Run 3.

### 2.3 Tracker alignment performance

A feature of the CMS tracker detector is its outstanding hit resolution, of about 10 $\mu$m. However, after installation, a mechanical alignment can only yield a precision on the position and orientation of the modules of about 0.1 mm [8]. Furthermore, it has been observed that changes in the conditions, like magnet cycles and temperature changes, as well as the long-term exposure to a high-radiation environment, can cause real or apparent movements of the detectors [8, 9]. To improve the precision of the knowledge of the component's geometry, a track-based alignment approach, relying on the minimization of the sum of squares of normalized track-hit residuals, is performed. This process allows us to obtain changes to alignment parameters, which describe the geometrical location of the components.

After the technical stop at the end of 2022 and the beginning of 2023 (Year End Technical Stop or YETS), significant movements were expected as explained before. To overcome this, alignment geometries were iteratively derived using cosmic rays and proton-proton collision data at 900 GeV

and 13.6 TeV, as data became available [10]. The performance achieved, continuously improving the mean and reducing the width of the track hit residuals to guarantee the accuracy for data-taking, is shown in Fig. 3.

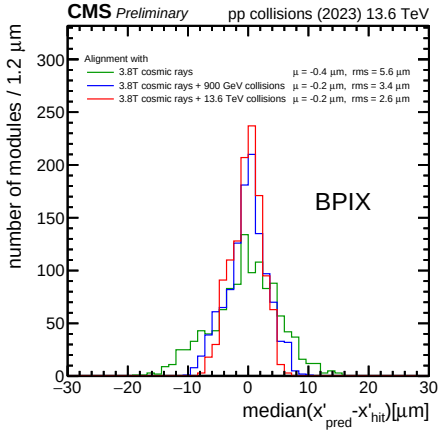


**Figure 3:** Distribution of median residuals in the local x coordinate on the barrel pixel detector according to alignment geometries derived iteratively in 2023.

## 3. Summary

The CMS tracker system plays a critical role in data-taking, enabling precise reconstruction of charged particle positions and momenta, even under the challenging conditions of Run 3, with a peak pileup of about 62 interactions per beam crossing. This article has discussed the performance of the Pixel and Silicon Strip detectors during Run 3, highlighting the continuous efforts to maintain exceptional performance and the role of the Tracker alignment in ensuring high-quality data.

## References

[1] CMS collaboration, "The CMS experiment at the CERN LHC", *Journal of Instrumentation* **3** (2008) S08004.

[2] CMS collaboration, "Development of the CMS detector for the CERN LHC Run 3", 2309.05466.

[3] L.O.S. Noehte and on behalf of the Tracker Group of the CMS collaboration, "CMS Phase-1 pixel detector refurbishment during LS2 and readiness towards the LHC Run 3", *Journal of Instrumentation* (2022) .

[4] CMS Collaboration, "Pixel Detector Performance in early 2023", CMS-DP-2023-041, https://cds.cern.ch/record/2865842, (2023).

[5] T.A. Vami for the CMS Collaboration, "Calibration and performance of the CMS pixel detector in LHC Run 2", Proceedings of 7th Annual Conference on Large Hadron Collider Physics — PoS(LHCP2019) https://doi.org/10.22323/1.350.0010, (2019).

[6] CMS Collaboration, "Pixel Detector Performance in Run 3", CMS-DP-2022-067, https://cds.cern.ch/record/2844889, (2022).

[7] CMS Collaboration, "CMS Silicon Strip Tracker Performance in 2023", CMS-DP-2023-040, https://cds.cern.ch/record/2865841, (2023).

[8] CMS Collaboration, "Strategies and performance of the CMS silicon tracker alignment during LHC Run 2", *Nucl. Inst and Meth. A* **1037** (2022) 166795.

[9] A.V. Barroso for the CMS Collaboration, "Tracker Alignment in CMS: Interplay with Pixel Local Reconstruction", Proceedings of Science — PoS(Pixel2022), arXiv:2303.16642, (2023). 10.22323/1.350.0010.

[10] CMS Collaboration, "Tracker alignment performance in early 2023", CMS-DP-2023-039, https://cds.cern.ch/record/2865840, (2023).

PoS(EPS-HEP2023)591

# Bibliography


[1] T. Kobayashi, "Experimental verification of the standard model of particle physics." *Proceedings of the Japan Academy, Series B* **97** (05, 2021) 211–235.

[2] H. Nilles, "Supersymmetry, supergravity and particle physics." *Phys. Rep.* **110** (1984), no. 1 1.

[3] CMS Collaboration, "Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc." *Physics Letters B* **716** (Sept., 2012) 30–61.

[4] ATLAS Collaboration, "Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc." *Physics Letters B* **716** (Sept., 2012) 1–29.

[5] F. Englert and R. Brout, "Broken Symmetry and the Mass of Gauge Vector Mesons." *Phys. Rev. Lett.* **13** (1964) 321. [157(1964)].

[6] P. W. Higgs, "Broken Symmetries and the Masses of Gauge Bosons." *Phys. Rev. Lett.* **13** (1964) 508. [,160(1964)].

[7] A. Bettini, *Introduction to Elementary Particle Physics* (Cambridge University Press, 1 edition, 2008).

[8] F. Wilczek, "Quantum field theory." *Reviews of Modern Physics* **71** (Mar., 1999) S85–S95.

[9] Particle Data Group Collaboration, R. L. Workman et al., "Review of Particle Physics." *PTEP* **2022** (2022) 083C01.

[10] S. Manzoni, *The Standard Model and the Higgs Boson*, pp. 9–38. Springer International Publishing, Cham, 2019.

[11] L. Husdal, "On effective degrees of freedom in the early universe." *Galaxies* **4** (09, 2016).

[12] E. Noether, "Invariante variationsprobleme." *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* **1918** (1918) 235.

[13] E. Fermi, "An attempt of a theory of beta radiation. 1.." *Z. Phys.* **88** (1934) 161–177.

[14] S. L. Glashow, "Partial Symmetries of Weak Interactions." *Nucl. Phys.* **22** (1961) 579–588.

[15] A. Salam, "Weak and Electromagnetic Interactions." *Conf. Proc.* **C680519** (1968) 367–377.

[16] S. Weinberg, "A Model of Leptons." *Phys. Rev. Lett.* **19** (1967) 1264–1266.

[17] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes, and R. P. Hudson, "Experimental test of parity conservation in beta decay." *Phys. Rev.* **105** (1957) 1413.

[18] N. Cabibbo, "Unitary Symmetry and Leptonic Decays." *Phys. Rev. Lett.* **10** (1963) 531. [,648(1963)].

[19] M. Kobayashi and T. Maskawa, "CP Violation in the Renormalizable Theory of Weak Interaction." *Prog. Theor. Phys.* **49** (1973) 652 KUNS-242.

[20] S. L. Glashow, "Partial-symmetries of weak interactions." *Nuclear Physics* **22** (1961), no. 4 579 – 588.

[21] J. Patera and H. Zassenhaus, "The pauli matrices in n dimensions and finest gradings of simple lie algebras of type an1." *Journal of Mathematical Physics* **29** (1988), no. 3 665–673.

[22] P. W. Higgs, "Spontaneous symmetry breakdown without massless bosons." *Phys. Rev.* **145** (May, 1966) 1156–1163.

[23] J. Ellis, "Higgs physics." (2013).

[24] N. Ky and N. Vân, "Latest results on the higgs boson discovery and investigation at the atlas-lhc." *Journal of Physics: Conference Series* **627** (06, 2015).

[25] D. E. Soper, "Parton distribution functions." *Nuclear Physics B - Proceedings Supplements* **53** (Feb., 1997) 69–80.

[26] CMS Collaboration, "Measurements of tth production and the cp structure of the yukawa interaction between the higgs boson and top quark in the diphoton decay channel." *Physical Review Letters* **125** (Aug., 2020).

[27] LHC Higgs Cross Section Working Group Collaboration, D. de Florian et al., "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector." `arXiv:1610.07922` FERMILAB-FN-1025-T, CERN-2017-002-M, [`arXiv:1610.07922`].

[28] L. Aaboud Zwalinski, "Observation of higgs decay to bottom quraks and vh production with the atlas detector." *Physics Letters B* **786** (Nov., 2018) 59–86.

[29] CMS Collaboration, N. Sirunyan, "Observation of higgs boson decay to bottom quarks." *Phys. Rev. Lett.* **121** (Sep, 2018) 121801.

[30] ATLAS Collaboration, "The atlas experiment at the cern large hadron collider." *Journal of Instrumentation* **3** (aug, 2008) S08003.

[31] CMS Collaboration, "The CMS experiment at the CERN LHC. The Compact Muon Solenoid experiment." *JINST* **3** (2008) S08004. Also published by CERN Geneva in 2010.

[32] CMS Collaboration, "The CMS Experiment at the CERN LHC." *J. Instrum.* **3** (2008) S08004.

[33] CMS Collaboration, "A portrait of the higgs boson by the cms experiment ten years after the discovery." *Nature* (2022).

[34] LHC Higgs Cross Section Working Group Collaboration, "Handbook of lhc higgs cross sections: 3. higgs properties: Report of the lhc higgs cross section working group." (2013).

[35] CMS Collaboration, C. Collaboration, "Evidence for higgs boson decay to a pair of muons." *Journal of High Energy Physics* **2021** (jan, 2021).

[36] CMS Collaboration, "A search for the standard model higgs boson decaying to charm quarks." *Journal of High Energy Physics* (2020).

[37] LHC Higgs Cross Section Working Group Collaboration, "LHC Higgs Cross Section Working Group." `https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWG` Accessed: 2019-10-02.

[38] C. Grojean, A. Paul, and Z. Qian, "Resurrecting $$ b\overline{b}h $$ with kinematic shapes." *Journal of High Energy Physics* **2021** (apr, 2021).

[39] CMS Collaboration, "Observation of higgs boson decay to bottom quarks." *Physical Review Letters* **121** (Sept., 2018).

[40] CERN, "Cern yellow reports: Monographs, vol 2 (2017): Handbook of lhc higgs cross sections: 4. deciphering the nature of the higgs sector." (2017).

[41] L. R. Evans and P. Bryant, "LHC Machine." *JINST* **3** (2008) S08001. This report is an abridged version of the LHC Design Report (CERN-2004-003).

[42] W. de Boer, *Precision Experiments at LEP*, p. 107–136. WORLD SCIENTIFIC, July, 2015.

[43] E. Mobs, "The CERN accelerator complex in 2019. Complexe des accélérateurs du CERN en 2019.". General Photo.

[44] P. Mouche, "Overall view of the LHC. Vue d'ensemble du LHC.". General Photo.

[45] LHCb Collaboration.

[46] T. A. Collaboration, "The alice detector at the lhc." *Journal of Instrumentation* **3** (aug, 2008) S08002.

[47] TOTEM Collaboration, G. Anelli et al., "The TOTEM experiment at the CERN Large Hadron Collider." *JINST* **3** (2008) S08007.

[48] LHCf Collaboration, A. Tiberio, "The LHCf experiment at the Large Hadron Collider: status and prospects." *PoS* **ICRC2023** (2023) 444.

[49] SND@LHC Collaboration, G. Acampora et al., "SND@LHC: The Scattering and Neutrino Detector at the LHC." `arXiv:2210.02784`.

[50] FASER Collaboration, A. Coccaro, "The FASER Experiment at the LHC." *Moscow Univ. Phys. Bull.* **77** (2022), no. 2 191–192.

[51] Moedal Collaboration, M. Staelens, "MoEDAL - Expanding the LHC's Discovery Frontier." *PoS* **LHCP2019** (2019) 031.

[52] CMS Collaboration, "Public cms luminosity information.".

[53] CMS Collaboration, "Illustration of the cms detector."

[54] W. P. S., "upgraded atlas and cms detectors and their physics capabilities."

[55] V. Blobel and C. Kleinwort, "A new method for the high-precision alignment of track detectors." (2002).

[56] V. Karimaki, T. Lampen, and F. P. Schilling, "The HIP algorithm for track based alignment and its application to the CMS pixel detector.".

[57] R. Brown, "The cms electromagnetic calorimeter." *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **572** (2007), no. 1 29–32. Frontier Detectors for Frontier Physics.

[58] C. Collaboration, "Calibration of the cms hadron calorimeters using proton-proton collision data at $\sqrt{s} = 13$ tev.".

[59] "Jet energy scale and resolution in the cms experiment in pp collisions at 8 tev." *Journal of Instrumentation* **12** (Feb., 2017) P02014–P02014.

[60] D. Abbaneo, M. Abbas, M. Abbrescia, A. Abdelalim, M. Akl, W. Ahmed, P. Altieri, R. Aly, C. Armaingaud, C. Asawatangtrakuldee, A. Ashfaq, P. Aspell, Y. Assran, I. Awan, S. Bally, Y. Ban, S. Banerjee, P. Barria, L. Benussi, and A. Zhang, "Quality control and beam test of gem detectors for future upgrades of the cms muon high rate region at the lhc." *Journal of Instrumentation* **10** (03, 2015) C03039–C03039.

[61] CMS Collaboration, "The Phase-2 Upgrade of the CMS Muon Detectors." tech. rep., CERN, Geneva, 2017. This is the final version, approved by the LHCC.

[62] CMS Collaboration, *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software.* Technical design report. CMS. CERN, Geneva (2006). There is an error on cover due to a technical problem for some items.

[63] G. Petrucciani, A. Rizzi, and C. Vuosalo, "Mini-aod: A new analysis data format for cms." *Journal of Physics: Conference Series* **664** (Dec., 2015) 072052.

[64] S. Höche, "Introduction to parton-shower event generators." (2015).

[65] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations." *Journal of High Energy Physics* **2014** (July, 2014).

[66] S. Frixione, P. Nason, and C. Oleari, "Matching nlo qcd computations with parton shower simulations: the powheg method." *Journal of High Energy Physics* **2007** (Nov., 2007) 070–070.

[67] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, "An introduction to pythia 8.2." *Computer Physics Communications* **191** (June, 2015) 159–177.

[68] M. L. Mangano, M. Moretti, F. Piccinini, and M. Treccani, "Matching matrix elements and shower evolution for top-pair production in hadronic collisions." *Journal of High Energy Physics* **2007** (Jan., 2007) 013–013.

[69] GEANT4 Collaboration, S. Agostinelli et al., "GEANT4–a simulation toolkit.".

[70] CMS Collaboration, "Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET." Tech. Rep. CMS-PAS-PFT-09-001, CERN, Geneva, 2009.

[71] C. Lippmann, "Particle identification." *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **666** (Feb., 2012) 148–172.

[72] "Strategies and performance of the cms silicon tracker alignment during lhc run 2." *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **1037** (Aug., 2022) 166795.

[73] P. Billoir and S. Qian, "Simultaneous pattern recognition and track fitting by the Kalman filtering method." *Nucl. Instrum. Methods Phys. Res. A* **294** (1990), no. 1 219.

[74] CMS Collaboration, "Particle-flow reconstruction and global event description with the cms detector." *Journal of Instrumentation* **12** (Oct., 2017) P10003–P10003.

[75] CMS Collaboration, "Performance of the cms muon detector and muon reconstruction with proton-proton collisions at s=13 tev." *Journal of Instrumentation* **13** (June, 2018) P06015–P06015.

[76] E. Widl and R. Frühwirth, "Application of the kalman alignment algorithm to the cms tracker." *Journal of Physics: Conference Series* **219** (apr, 2010) 032065.

[77] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems." *Proceedings of the IEEE* **86** (1998), no. 11 2210–2239.

[78] W. Waltenberger, "Adaptive vertex reconstruction." Tech. Rep. CERN-CMS-NOTE-2008-034, CERN, Geneva, 2008.

[79] CMS Collaboration, B. Sirunyan, and Wyslouch, "Performance of the cms muon detector and muon reconstruction with proton-proton collisions at s=13 tev." *Journal of Instrumentation* **13** (June, 2018) P06015–P06015.

[80] CMS Collaboration, "Electron and photon reconstruction and identification with the cms experiment at the cern lhc." *Journal of Instrumentation* **16** (May, 2021) P05014.

[81] W. Adam, R. Frühwirth, A. Strandlie, and T. Todorov, "Reconstruction of electrons with the gaussian-sum filter in the cms tracker at the lhc." *Journal of Physics G: Nuclear and Particle Physics* (2005).

[82] M. Cacciari, G. P. Salam, and G. Soyez, "The Anti-k(t) jet clustering algorithm." *J. High Energy Phys.* **04** (2008) 063 LPTHE-07-03, [arXiv:0802.1189].

[83] D. Bertolini, P. Harris, M. Low, and N. Tran, "Pileup per particle identification." *Journal of High Energy Physics* **2014** (Oct., 2014).

[84] "How cms weeds out particles that pile up." https://cms.cern/news/how-cms-weeds-out-particles-pile

[85] "Jet energy scale and resolution in the cms experiment in pp collisions at 8 tev." *Journal of Instrumentation* **12** (Feb., 2017) P02014–P02014.

[86] CMS Collaboration, "Identification of b-quark jets with the cms experiment." *Journal of Instrumentation* **8** (Apr., 2013) P04013–P04013.

[87] CMS Collaboration, "Cms b-tagging and vertexing results." https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsBTV Accessed: 2019-09-30.

[88] E. Bols, J. Kieseler, M. Verzetti, M. Stoye, and A. Stakia, "Jet flavour classification using deepjet." *Journal of Instrumentation* **15** (Dec., 2020) P12012–P12012.

[89] CMS Collaboration, "Performance of reconstruction and identification of $\tau$ leptons decaying to hadrons and $\nu_\tau$ in pp collisions at $\sqrt{s} = 13$ TeV." *JINST* **13** (2018), no. 10 P10005, [arXiv:1809.02816].

[90] CMS Collaboration, K. Androsov, "Identification of tau leptons using Deep Learning techniques at CMS." Tech. Rep. CMS-CR-2019-272, CERN, Geneva, Nov, 2019.

[91] CMS Collaboration, "Performance of missing transverse momentum reconstruction in proton-proton collisions at s = 13 tev using the cms detector." *Journal of Instrumentation* **14** (July, 2019) P07004–P07004.

[92] D. Bertolini, P. Harris, M. Low, and N. Tran, "Pileup per particle identification." *Journal of High Energy Physics* **2014** (Oct., 2014).

[93] CMS Collaboration, "Pileup mitigation at cms in 13 tev data." *Journal of Instrumentation* **15** (Sept., 2020) P09018–P09018.

[94] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press (2016). http://www.deeplearningbook.org.

[95] K. Albertsson, "Machine learning in high energy physics community white paper." (2019).

[96] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research* **16** (June, 2002) 321–357.

[97] A. Hallin, G. Kasieczka, T. Quadfasel, D. Shih, and M. Sommerhalder, "Resonant anomaly detection without background sculpting." *Physical Review D* **107** (June, 2023).

[98] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system." in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Aug., 2016.

[99] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?." (2022).

[100] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need." (2021).

[101] B. R. Muller and W. Smith, "A hierarchical loss for semantic segmentation." in *VISIGRAPP*, 2020.

[102] M. Romero, O. Ramírez, J. Finke, and C. Rocha, "Feature extraction using spectral clustering for gene function prediction." (03, 2022).

[103] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions." (2017).

[104] CMS Collaboration, "Observation of $t\bar{t}h$ production." *Phys. Rev. Lett.* **120** (Jun, 2018) 231801.

[105] N. Deutschmann, F. Maltoni, M. Wiesemann, and M. Zaro, "Top-yukawa contributions to bbh production at the lhc." *Journal of High Energy Physics* **2019** (July, 2019).

[106] D. Pagani, H.-S. Shao, and M. Zaro, "RIP $Hb\bar{b}$: how other higgs production modes conspire to kill a rare signal at the LHC." *Journal of High Energy Physics* **2020** (nov, 2020).

[107] R. Harlander, M. Kramer, and M. Schumacher, "Bottom-quark associated Higgs-boson production: reconciling the four- and five-flavour scheme approach." `arXiv:1112.3478`.

[108] S. Forte, D. Napoletano, and M. Ubiali, "Higgs production in bottom-quark fusion in a matched scheme." *Physics Letters B* **751** (Dec., 2015) 331–337.

[109] J. Alwall, S. Höche, F. Krauss, N. Lavesson, L. Lönnblad, F. Maltoni, M. Mangano, M. Moretti, C. Papadopoulos, F. Piccinini, S. Schumann, M. Treccani, J. Winter, and M. Worek, "Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions." *The European Physical Journal C* **53** (Dec., 2007) 473–500.

[110] R. D. Ball, V. Bertone, S. Carrazza, L. D. Debbio, S. Forte, P. Groth-Merrild, A. Guffanti, N. P. Hartland, Z. Kassabov, J. I. Latorre, E. R. Nocera, J. Rojo, L. Rottoli, E. Slade, and M. Ubiali, "Parton distributions from high-precision collider data: Nnpdf collaboration." *The European Physical Journal C* **77** (Oct., 2017).

[111] CMS Collaboration, "Extraction and validation of a new set of cms pythia8 tunes from underlying-event measurements." *The European Physical Journal C* **80** (Jan., 2020).

[112] L. Bianchini, J. Conway, E. K. Friis, and C. Veelken, "Reconstruction of the Higgs mass in $H \to \tau\tau$ Events by Dynamical Likelihood techniques." *J. Phys. Conf. Ser.* **513** (2014) 022035.

[113] CDF Collaboration, A. Abulencia, D. Acosta, J. Adelman, A. Affolder, T. Akimoto, M. Albrow, D. Ambrose, S. Amerio, D. Amidei, A. Anastassov, K. Anikeev, A. Annovi, J. Antos, M. Aoki, G. Apollinari, J.-F. Arguin, T. Arisawa, A. Artikov, W. Ashmanskas, and S. Zucchelli, "Search for neutral higgs bosons of the minimal supersymmetric standard model decaying to $\tau$ pairs in p p collisions at s = 1.96 tev." *Physics Research Publications* **96** (01, 2006).

[114] CMS Collaboration, "Observation of the higgs boson decay to a pair of  leptons with the cms detector." *Physics Letters B* **779** (Apr., 2018) 283–316.

[115] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree." in *Neural Information Processing Systems*, 2017.

[116] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. G. andand Jaques Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project." in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

[117] CMS Collaboration, "Precision luminosity measurement in proton-proton collisions at $\sqrt{s} = 13$ TeV in 2015 and 2016 at CMS." *Eur. Phys. J. C* **81** (2021), no. 9 800, [arXiv:2104.01927].

[118] CMS Collaboration, "CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV." tech. rep., CERN, Geneva, 2018.

[119] CMS Collaboration, "CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV." tech. rep., CERN, Geneva, 2019.

[120] CMS Collaboration, "Tau id recommendations for run-2: 2016, 2017, and 2018."

[121] CMS Collaboration, ""egammaul2016to2018."

[122] CMS Collaboration, "Jet energy scale and resolution measurement with Run 2 Legacy Data Collected by CMS at 13 TeV.".

[123] CMS Collaboration, "Measurement of the Higgs boson production rate in association with top quarks in final states with electrons, muons, and hadronically decaying tau leptons at $\sqrt{s} = 13$ TeV." *Eur. Phys. J. C* **81** (2021), no. 4 378, [arXiv:2011.03652].

[124] LHC Higgs Cross Section Working Group Collaboration, "Handbook of lhc higgs cross sections: 3. higgs properties: Report of the lhc higgs cross section working group." (2013).

[125] TheATLAS, TheCMS, TheLHCHiggsCombinationGroup Collaboration, "Procedure for the LHC Higgs boson search combination in Summer 2011." tech. rep., CERN, Geneva, 2011.

[126] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics." *The European Physical Journal C* **71** (feb, 2011).

[127] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics." *The European Physical Journal C* **71** (Feb., 2011).

[128] CMS Collaboration, "Search for the SM Higgs boson produced in association with bottom quarks in final states with leptons." tech. rep., CERN, Geneva, 2024.

[129] CMS Collaboration, "Measurements of Higgs boson production in the decay channel with a pair of $\tau$ leptons in proton–proton collisions at $\sqrt{s} = 13$ TeV." *Eur. Phys. J. C* **83** (2023), no. 7 562, [arXiv:2204.12957].

[130] CMS Collaboration, "An embedding technique to determine $\tau\tau$ backgrounds in proton-proton collision data." *Journal of Instrumentation* **14** (Jun, 2019) P06032–P06032.

[131] D. Leyva Pernia and M. Makou, "CMS tracker performance in Run3." *PoS* **EPS-HEP2023** (2023) 591.

# Acknowledgements

Finishing this thesis in three years was a big challenge, and I could not have done it without the invaluable support of many colleagues and friends. I am extremely thankful for that.

Foremost, I extend my genuine thankfulness to my supervisor, Prof. Dr. Elisabetta Gallo, for her guidance, support, and profound expertise, which have been crucial throughout this research journey. Her mentorship has enabled my academic and personal growth, and her careful attention to detail on every occasion, such as proofreading my work, and her continuous encouragement have been very influential.

I deeply appreciate Prof. Dr. Peer Stelldinger's support. His extensive knowledge of computer science and willingness to engage in countless discussions have enriched my learning experience greatly. His patience in addressing my questions and his approachable and humble manners have taught me valuable lessons beyond the realms of machine learning.

My co-supervisor, Dr. Rainer Mankel, deserves special mention for his detailed comments, feedback, and support throughout these years. His contributions, from research insights to proofreading of the thesis and presentation advice, have been priceless.

Dr. Roberval Walsh's technical insight, especially concerning the tracker and trigger subsystems, has been crucial when facing challenging problems. His guidance during my first tracker offline DQM shift was a significant learning opportunity for which I am very grateful.

I am also extremely grateful to Dr. Alexei Raspereza, who has provided a huge amount of help and support at every step of this thesis work. Finalizing the analysis would not have been possible without his suggestions and great knowledge of physics, which he provides unconditionally. He is a great scientist and an amazing human with a positive attitude, which often encouraged me when experiencing technical difficulties.

Furthermore, I would also like to thank Andrea Cardini, the postdoctoral of the project, for all the assets he provided, from setting up the analysis chain to being the contact person for the analysis to taking care of all the formalities and meetings with the review committee and also taking care of the paper.

I would also like to thank DASHH graduate school for providing full-time positions for PhDs. This was a life changer for me, especially as an immigrant to Germany, and it helped me solve several of my bureaucratic problems.

I would also like to thank Daina, a great friend and colleague who was always available to help and answer many questions. She suggested attending the Latin American school in Chile, and we also jointly presented the Tracker Alignment poster during the EPS-HEP conference. Last, I want to mention her warm attitude, many deep discussions about life, and the ambrosial Cuban food she prepared.

I am also thankful for my friend Valentina for always being present with a positive mindset. The gym and swimming pool sessions with her have been very fun and enjoyable.

I also want to thank Federico, Alessia, and Alberto for being great friends and for all the amazing moments we cherished.

I am also grateful to all the DESY-CMS group colleagues, PhDs, and postdocs for creating a positive atmosphere and enjoyable lunch breaks, enhanced by the food quality.

I would also like to thank my previous office mates Oleg, Aliya, and Andrea, as well as the current ones Stepan and Jacopo, for making the office so lively and fun. Thanks especially to Stepan for redecorating the office and always providing much-appreciated coffee and sweets.

I would also like to thank my family, my brother Ali and my parents, to whom my thoughts go constantly. It would not have been possible to come this far without their support. Thanks for constantly encouraging me and always supporting my decisions. It wasn't easy to be thousands of kilometres away from you, but I am grateful for your love and support regardless of the distance. The close proximity to my brother Ali made this journey more pleasant. Thanks for always being there to listen to me and giving me your valuable advice.

I want to thank my partner for all his unconditional love and support and for believing in me. I thank you for always being there during the hardships and for bringing a humorous, happy-go-lucky attitude, and for providing exquisite Italian meals. Thank you especially for making "la vita frizzante", even though I still do not enjoy sparkling water.

I'd also like to extend my deepest thanks to my best friend—or rather, my sister—Neda, for all the incredible days we've shared. Thank you for always being there for me, no matter what. I feel incredibly fortunate to have found such an amazing friend at DESY.

I'm incredibly thankful for my friend Emanuela. Having her in my life is such a gift. Our endless chats about everything and anything mean the world to me. I also want to thank my friend Amir for all the great conversations we've had and for sharing the challenges we faced while living abroad.

I also want to extend my heartfelt thanks to my incredible friends from university days in Tehran - Parnia, Parastoo, Kiana, Niloofar, Farnaz, Yeganeh, Nanor, Saina and Zahra. Your support and the wonderful times we shared over those four years were truly unforgettable and always bring a smile to my face when I reminisce. Studying and being friend with all of you was a joyful experience.

I also want to thank specially my bestie, Tara "Schatzi FBI" with whom I've shared not only great days but also stressful ones as we planned our future in Germany together. I am incredibly happy and grateful to have you in my life, and even more thrilled that we've been able to achieve our dreams together while building an amazing friendship.

Also, a huge shout out to my best friends in my hometown, Tina and Saeed. Even though we're miles apart, chatting with them fills me with joy and happiness. Their friendship is something I cherish deeply, and it just goes to show how true connections can make the distance feel a little less.