VRIJE
UNIVERSITEIT
BRUSSEL

Dissertation submitted in fulfilment of the requirements for the Master of Science in Physics and Astronomy

BACKGROUND ESTIMATIONS IN THE MEASUREMENT OF THE CHARM-HIGGS YUKAWA COUPLING AT THE CMS EXPERIMENT

Stef Duponcheel

May 2025

Under the guidance of Prof. Dr. Michael Tytgat and Dr. Gerrit Van Onsem

Faculty of Science and Bio-Engineering Sciences

Abstract

In this thesis, the charm-Higgs Yukawa coupling y_c is investigated via the associated production of a Higgs boson and a charm quark (H+c), where the Higgs boson decays via the four-muon channel ($\mathrm{H} \to \mathrm{ZZ}^* \to 4\mu$). The analysis characterises the signal signature, estimates irreducible backgrounds using simulation, and determines the reducible background through a data-driven method based on proton-proton collisions recorded in 2018 by the CMS detector at the LHC. By combining jet flavour-tagging information related to the identification of jets initiated by charm quarks in a two-dimensional approach, an expected upper limit on the H+c production signal strength of $\mu_{\mathrm{H+c}} = 33.12$ at 95% confidence level is reported. This marks the first time data recorded by the CMS detector is employed to study H+c production in the four-muon final state of the Higgs boson. The result demonstrates the potential of this production mode to constrain the charm-Higgs Yukawa coupling, and the methods developed lay the groundwork for future studies.

Acknowledgements

First, I would like to thank Prof. Dr. Michael Tytgat, my promoter, for allowing me to work on this research project. Although he has since parted ways with the VUB, I also wish to thank Prof. Dr. Jorgen D'Hondt, who originally established this thesis subject at our institute. Both of them provided me opportunities in my research career that I am genuinely thankful for.

Next, I would like to thank the two people who were most closely involved in the making of this project: Gerrit and Felix. Thank you both for your time and patience throughout the year, especially for all the meetings, and for putting up with my silly questions. Gerrit, thank you for your guidance. Every time I had a physics-related question, you were there to help. Felix, thank you for keeping me grounded. Whenever I ran into a problem, whether coding-related or just what the next steps were, you were always there to steer me in the right direction. I would also like to thank Nordin. I didn't have time to make any acknowledgments for my bachelor's thesis, so I want to extend my thanks here. You indirectly shaped this thesis through the things I learned from you. To everyone else in the office, thank you for the company.

Next, I would like to thank my dearest friends: Maxime, Tobi, and Joren. If I were to list the reasons why I should thank each of you, it would amount to a second thesis. You all hold a special place in my heart. Thank you for the countless memories made over the 5 years of studying together, and here's to many more. To everyone else in my year, thank you as well. Ono, thank you for listening to me complaining about literally anything. Liesbeth, thanks for the infamous 20h15 breaks, which were almost my only form of social interaction in the last week before the deadline. Even to my friends outside of physics, you have all shaped me into the person I am today. A special shout-out goes to Manon. We must keep our tradition of the *spaghetti-avonden*!

Lastly, I want to thank my family, especially my parents. You have both always believed in me, even in the moments I did not believe in myself. Everything that I have achieved and will ever achieve is thanks to you. From the deepest of my heart: I love you and thank you for your never-ending support.

Contents

1 INTRODUCTION

As complex as the universe may seem, it appears to be composed of only a finite set of indivisible constituents at the smallest scales. Particle physics aims to understand what these building blocks, known as the elementary particles, are and how they interact via the fundamental forces. Our current understanding is encoded in one of the greatest scientific frameworks: The Standard Model of particle physics (SM). In Section 1.1, we will introduce the SM and familiarise ourselves with its mathematical framework, Quantum Field Theory (QFT). As we will see in the last part of this section, one of the key predictions made by the SM is the existence of the Yukawa couplings, which characterise the interactions between the Higgs boson and massive particles. In Section 1.2, we will discuss state-of-the-art measurements of these parameters, where the emphasis will be placed on an unmeasured variable, the charm-Higgs Yukawa coupling. Afterwards, we will introduce the process of interest for the thesis, called the H+c process, and how it provides an alternative to measure this coupling. In the last part of this chapter, Section 1.3, other processes giving rise to similar signatures as H+c will be discussed, followed by a general overview of the work plan for the thesis.

1.1 The Standard Model of particle physics

1.1.1 The elementary particles and their fundamental interactions

The SM describes the elementary particles and their interactions governed by three fundamental forces. Throughout the field's history, our understanding of what these elementary particles are has changed significantly [1]. For example, protons and neutrons were once thought to be fundamental. However, inelastic electron-nucleon scattering experiments conducted between the 1960s and 1970s, together with theoretical advancements, revealed these particles are made up of point-like constituents known as quarks [2]. Based on our current understanding, the SM comprises 17 elementary particles. The different properties of these particles allow for a schematic classification of the Standard Model, as shown in Figure 1.1. One of these properties is the spin, dividing the SM particles into *fermions* and *bosons*, which will be discussed in more detail below. The following subsections find their inspiration from [3], [4], and [5]. Where applicable, other references will be cited.

Fermions

The SM fermions are the elementary particles with spin 1/2. Based on their interactions, fermions are divided into *leptons* and the aforementioned *quarks*, each grouped into three generations. Leptons either have an electromagnetic charge of -1 or are electrically neutral. The electrically charged leptons are the electron e, the muon μ, and the τ lepton. Each charged lepton has a corresponding electrically neutral particle, known as a neutrino. The electron and its neutrino form the first generation of leptons, while the second and third generations consist of the muon and muon neutrino, and the tau lepton and tau neutrino, respectively. Generally, a higher-generation particle corresponds to one with the same electric charge as the previous generation, but with a higher

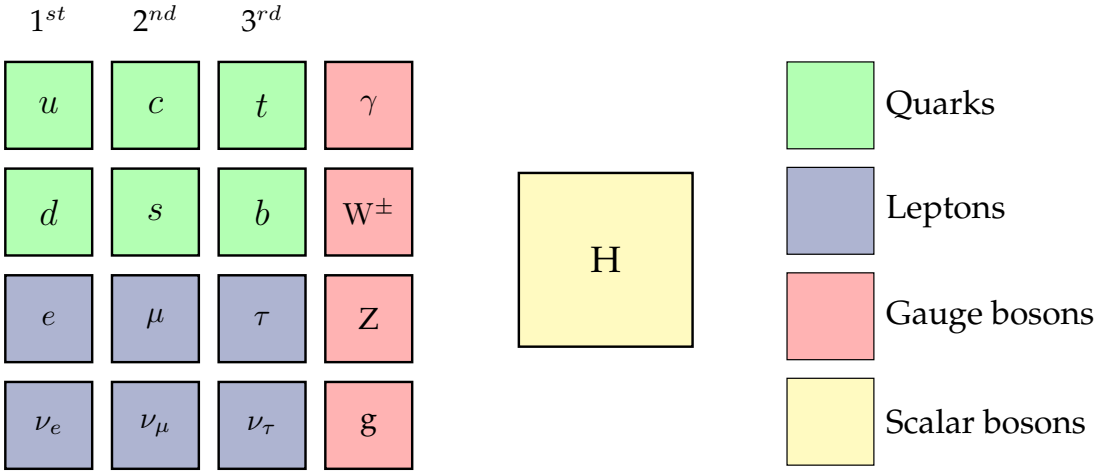

Figure 1.1: *The elementary particles of the Standard Model.*

mass. The only exceptions are neutrinos, which are considered to be massless in the SM[1]. Quarks, on the other hand, come in fractional electric charges of +2/3 and -1/3. The first generation of quarks consists of the up quark (u) with an electric charge of +2/3 and the down quark (d) with an electric charge of -1/3. The second generation of quarks consists of the charm quark (c) and the strange quark (s), with respective charges of +2/3 and -1/3. Finally, the third generation consists of the top quark (t) with a charge of +2/3 and the bottom quark (b) with a charge of -1/3. In addition to having an electric charge, quarks also possess a so-called colour charge, allowing them to interact via the strong nuclear force. The nature of this interaction has strong implications for the behaviour of quarks, particularly in how they are detected. Quarks have never been observed as isolated particles, but as colourless bound states, referred to as *hadrons*. This behaviour is hypothesized to result from the phenomenon of *colour confinement*, which prohibits the existence of an asymptotically free coloured particle. Only a handful of hadrons exist, the most common being mesons, consisting of a quark and an antiquark, such as pions, and baryons composed of three quarks. Perhaps the most familiar example of a baryon is a proton, consisting of two up and one down quark.

Bosons

The elementary bosons of the SM are the integer spin particles. These can be further categorized using their spin value. In particular, spin-1 particles are known as gauge bosons, and the spin-0 boson of the SM is the scalar Brout-Englert-Higgs boson. In the SM framework, an interaction between particles corresponds to the exchange of gauge bosons associated with the respective fundamental force. Therefore, describing each gauge boson also invites us to briefly describe the forces they mediate:

- **The electromagnetic force:** The electromagnetic force couples to all electrically charged particles. It is mediated by a massless and electrically neutral gauge boson, the photon (γ).
- **The weak force:** The weak force is the only interaction that affects all fermions and allows for flavour[2] change. It is mediated by the electrically charged $W^{\pm}$ bosons responsible for charged current interactions and the neutral Z boson responsible for neutral current interactions. Unlike other gauge bosons, the $W^{\pm}$ and Z bosons are massive.

[1] In contrast to what the SM assumes, the discovery of neutrino oscillations in 1998 by the Super-Kamiokande detector indicates that the neutrinos have a non-zero mass [6].

[2] The flavour of a particle refers to its species, i.e., electron flavour, muon flavour, ...

- **The strong force:** The strong force only couples to colour-charged particles. Interactions are mediated by gluons (g), which are massless and electrically neutral. Gluons carry a colour charge themselves, leading to self-interactions that characterise the unique nature of the interaction, as previously discussed in the context of fermions.
- **The gravitational force:** The most familiar fundamental force experienced daily is gravity, yet it is the only force not incorporated into the SM. Some theories extending the SM predict the existence of a spin-2 tensor boson, the graviton, as the mediator, but no experimental evidence supports this claim. This is not problematic at the energy scales probed by current experiments, as the gravitational interaction strength is negligible compared to the other forces. However, the lack of a unified description of the fundamental forces shows that the SM is an incomplete theory.

The final component of the SM is the Brout-Englert-Higgs boson, referred to as the Higgs boson in short. Being the only scalar particle of the SM, it is unlike anything else we previously described. As shall be explained in Section 1.1.3, the Higgs particle arises from the mechanism responsible for generating particle masses in a gauge-invariant way. Its existence was first proposed independently in 1964 by Brout and Englert [7], and Higgs [8]. The extensive search for SM Higgs concluded in 2012 with experimental confirmation by the ATLAS [9] and CMS [10] experiments at the Large Hadron Collider (LHC) at CERN (European Organization for Nuclear Research).

1.1.2 Quantum Field Theory

Equations of motion

The SM is mathematically formulated within the Quantum Field Theory (QFT) framework. This theory couples special relativity and quantum mechanics, allowing for a consistent description of subatomic processes at relativistic energies. In QFT, a particle is represented as an excitation of its associated quantum field $\phi(x)$ with x the spacetime coordinate. The dynamics of a quantum field, and thus its associated particle, is fully captured by the Lagrangian density $\mathcal{L} \coloneqq \mathcal{L}(\phi, \partial_\mu \phi)$, where $\partial_\mu \phi$ is the four-vector derivative to the spacetime coordinates $x^\mu = (ct, \mathbf{x})$. From this Lagrangian, the equations of motion are inferred from applying the principle of least action $\delta S = 0$, where the action is defined as $S = \int d^4x \, \mathcal{L}$, leading to the Euler-Lagrange equations for quantum fields:

$$\frac{\partial \mathcal{L}}{\partial \phi} - \partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \right) = 0. \tag{1.1}$$

To illustrate this formalism, we look at the fermions described by 4×4 Dirac spinors ψ. The dynamics of such a free spinor are encoded in the Dirac Lagrangian $\mathcal{L}_D$:

$$\mathcal{L}_D = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi, \tag{1.2}$$

where γ^μ are the 4×4 gamma matrices, and $\bar{\psi}$ the adjoint spinor defined as $\psi^\dagger \gamma^0$. Indeed, the equation of motion obtained from Equation (1.1) is the well-known Dirac equation $(i\gamma^\mu \partial_\mu - m)\psi = 0$. The negative energy solutions of this theory are interpreted as antiparticle states $\bar{f}$ with an opposite electric charge to their corresponding particle state f.

Gauge invariance

The interactions between particles arise from the *local gauge principle*. This generally comes down to requiring the free Lagrangian (e.g. $\mathcal{L}_D$ from Equation (1.2)) to be invariant under transformations implied by the chosen gauge group. The locality requirement of the transformations will necessarily lead to the introduction of new fields, the gauge fields, associated with the previously

mentioned gauge bosons. The most straightforward gauge theory to illustrate this principle is Quantum Electrodynamics (QED), which describes the electromagnetic interaction. The underlying symmetry gauge group of QED is U(1), the group of 1×1 hermitian matrices, simply complex phases. The requirement of U(1) local gauge invariance corresponds to invariance under the following local phase transformations of the spinor field:

$$\psi(x) \rightarrow \psi'(x) = e^{-igf(x)}\psi(x), \tag{1.3}$$

where $f(x)$ is the U(1) complex phase and g is a constant factor. Under these transformations, the free Dirac Lagrangian from Equation (1.2) is not invariant due to the four derivatives acting on the local phase. The invariance is restored by replacing the ordinary derivative ∂_μ by the covariant derivative D_μ defined as:

$$D_\mu = \partial_\mu + igA_\mu, \tag{1.4}$$

where we needed to introduce a new gauge field A_μ, the photon field, and where g can now be interpreted as a coupling strength that characterises the interaction between the spinor and gauge fields, more familiarly known as the electromagnetic charge e. Provided that this new field transforms as:

$$A_\mu(x) \rightarrow A'_\mu(x) = A_\mu(x) + \partial_\mu f(x), \tag{1.5}$$

the Dirac Lagrangian is now U(1) gauge invariant and is given by:

$$\mathcal{L}_{\mathrm{D}} = \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi - g\bar{\psi}\gamma^\mu A_\mu\psi, \tag{1.6}$$

where the last term characterised the interaction of the fermion and photon fields. To complete the QED framework, an additional kinetic term of the gauge boson, -$\frac{1}{4}F_{\mu\nu}F^{\mu\nu}$ where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is known as the field strength tensor, has to be added that connects the theory to the Maxwell equations. All of these ingredients lead to the QED Lagrangian, which fully captures the theory of electromagnetism:

$$\mathcal{L}_{\mathrm{QED}} = \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi - g\bar{\psi}\gamma^\mu A_\mu\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}. \tag{1.7}$$

The fundamental forces revisited

Having explained the local gauge principle, we can return to the fundamental forces embedded in the SM and explain the origins of their associated gauge bosons, introduced in Section 1.1.1:

- **The electroweak interaction:** It was previously explained that the U(1) gauge invariance of QED describes the theory of electromagnetism, however, there is more to this than meets the eye. In the SM, electromagnetism and the weak interaction are unified into the electroweak force, following the Glashow, Salam, and Weinberg model [11, 12, 13]. The underlying gauge group associated with this unified description is denoted as SU(2)$_L$ $\times$ U(1)$_Y$. The U(1)$_Y$ part refers to the U(1) of QED with a different but related quantity to the electromagnetic charge, the weak hypercharge Y. The electroweak theory's SU(2)$_L$ part, where SU(2) represents the group consisting of all 2×2 unitary matrices U with $\det(U) = 1$, reflects its chiral nature. Only left-handed particles and right-handed antiparticles participate in the electroweak interactions, meaning their associated fields will transform as doublets under SU(2)$_L$, explaining why we denote the group with a subscript L. Their chiral counterparts, the right-handed particles and left-handed anti-particles, will transform as singlets under this group. The left-handed doublets for the quarks Q_L, leptons L_L, and the right-handed

quark and lepton singlets are denoted by:

$$\begin{aligned} Q_L &: \begin{pmatrix} u \\ d \end{pmatrix}_L, \begin{pmatrix} c \\ s \end{pmatrix}_L, \begin{pmatrix} t \\ b \end{pmatrix}_L \\ L_L &: \begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_L, \begin{pmatrix} \nu_\mu \\ \mu^- \end{pmatrix}_L, \begin{pmatrix} \nu_\tau \\ \tau^- \end{pmatrix}_L \\ & u_R, d_R, c_R, s_R, t_R, b_R \\ & e_R, \mu_R, \tau_R \end{aligned} \tag{1.8}$$

The required SU(2)$_L$ × U(1)$_Y$ local gauge invariance introduces new gauge fields, W_μ^a with $a \in \{1, 2, 3\}$ and B_μ, and the associated covariant derivative is given by:

$$D^\mu = \partial^\mu - iW_k^\mu \frac{\sigma^k}{2} - ig' \frac{Y}{2} B^\mu, \tag{1.9}$$

where σ^k are the Pauli matrices, and g and g' the electroweak coupling constants. The physical gauge bosons fields of this theory: the photon field A_μ, the Z boson field Z_μ, and the charged W boson fields $W_\mu^\pm$ arise as linear combinations of the generated gauge fields under the SU(2)$_L$ × U(1)$_Y$ symmetry.

- **The strong interaction:** The strong interaction is described by quantum chromodynamics (QCD), which has SU(3) as its underlying gauge group. The charge associated with this group is the previously mentioned colour charge, which is why the group is often denoted as SU(3)$_C$. In group theory language, SU(3)$_C$ is a non-abelian group with eight generators. Requiring local SU(3)$_C$ gauge invariance introduces eight gauge fields G_μ^a for $a \in \{1, 2, .., 8\}$, identified as the eight gluon fields of QCD. The associated covariant derivative is given by:

$$D^\mu = \partial^\mu - ig_s \frac{\lambda^a}{2} G_a^\mu, \tag{1.10}$$

where g_s is the coupling constant of the strong interaction, and λ^a are the Gell-Mann matrices of the SU(3)$_C$ group.

In its full glory, the underlying gauge group of the SM G_{SM} can thus be written as:

$$G_{\text{SM}} = SU(3)_C \times SU(2)_L \times U(1)_Y,$$

which fully specifies how the particle fields transform and interact with each other.

Towards physical predictions

With the QFT formalism, physical observables of interaction processes can be calculated. These processes are formulated via time-dependent perturbation theory and involve calculating the quantum mechanical transition element $\mathcal{M}$ as one of the basic ingredients. While these calculations are possible via first principles, they are more commonly done with Feynman rules associated with the Feynman diagrams. A Feynman diagram gives a visual representation of how particles interact, starting from an initial state (left-hand side of the diagram) to a final state (right-hand side of the diagram), taking into account all the possible time orderings of the process. These calculations are expansions up to a certain order, more specifically in powers of the coupling constants of the interaction, such that each order in the expansion has associated diagrams. Higher-order processes thus involve more interaction vertices to get the same final state of the considered process. The diagrams with the minimal number of vertices are called tree-level or leading-order (LO), and higher-order diagrams are referred to as next-to-leading order (NLO), and after that,

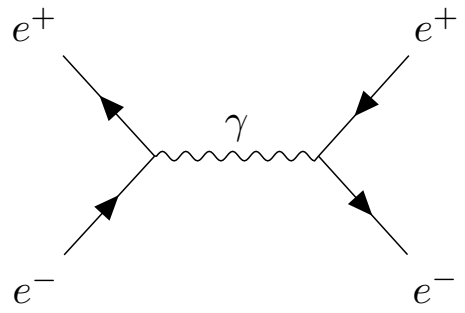

Figure 1.2: *Representative LO Feynman diagram showing the annihilation of an electron* (e^-) *and a positron* (e^+) *into a photon* (γ) *which gives rise to a new electron-positron pair.*

next-to-next-to-leading order (NNLO). An example of a Feynman diagram is shown in Figure 1.2. With the matrix element calculated, physical observables can be computed. Calculating these physical observables is essentially the end goal of any calculation in QFT, as they serve as the only theoretical predictions that can be compared to experimental data, providing us with direct tests of the SM. Among these variables is the cross section σ, which is inferred from the squared matrix element of the Feynman diagram $\mathcal{M}$. It can be interpreted as being proportional to the probability of an interaction, with higher cross section processes being more likely to occur. Following historical reasons, the cross section is often expressed in a somewhat peculiar unit, called a barn. In the SI unit system, one barn b corresponds to $10^{-24}\ \mathrm{cm}^2$.

1.1.3 Electroweak symmetry breaking

Although the local gauge principle allows for a gauge-invariant description of the fundamental interactions, it is only consistent if the introduced gauge bosons remain massless. Naively giving the photon a mass, for example, would correspond to extending the QED Lagrangian $\mathcal{L}_{QED}$ from Equation (1.7) with an additional term $m_\gamma^2 A_\mu A^\mu$, where m_γ^2 is interpreted as the mass of the gauge boson. However, under the U(1) gauge transformations, this extra term spoils the gauge invariance of the theory. Whereas the photon and the gluon are massless, so the previous reasoning is redundant, the $\mathrm{W}^\pm$ and Z bosons are massive, and thus another mechanism needs to be considered to generate the masses of the electroweak gauge bosons in a gauge-invariant way. In the Standard Model, this is provided by the Brout-Englert-Higgs (BEH) mechanism. The fundamental principle at the heart of mass generation is that of spontaneous symmetry breaking, which is the phenomenon of a theory being invariant under a symmetry while the ground state is not. The starting point of the BEH mechanism is introducing a scalar field Φ in the theory that transforms as a doublet under $SU(2)_L$ and is charged under the $U(1)_Y$ gauge group:

$$\Phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}, \tag{1.11}$$

with ϕ_i, $i \in \{1, 2, 3\}$ real scalar fields. The corresponding $SU(2)_L \times U(1)_Y$ invariant Lagrangian for this scalar field is given by:

$$\mathcal{L}_{BEH} = (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(\Phi) = (D_\mu \Phi)^\dagger (D^\mu \Phi) - \frac{\mu^2}{2} \Phi^\dagger \Phi - \frac{\lambda}{4} (\Phi^\dagger \Phi)^2, \tag{1.12}$$

where μ^2 is interpreted as a mass coefficient, $\lambda > 0$ characterises the self-interaction strength of the scalar field, and the covariant derivative is given by Equation (1.9). The shape of the Higgs potential $V(\Phi)$ and the physics it introduces depend on the sign of μ^2. In the case of $\mu^2 < 0$, the

ground state Φ_0 is degenerate with each minima satisfying $\Phi_0\Phi_0^\dagger = v^2/2$ where $v = \sqrt{\mu^2/\lambda}$ is the vacuum expectation value (vev) of 246 GeV [14]. When the field acquires the vev and sits in one of its many ground states, the theory spontaneously loses its $SU(2)_L \times U(1)_Y$ gauge invariance, hence the name spontaneous symmetry breaking. For each continuous symmetry that is broken, a massless boson is generated as a result of Goldstone's theorem [15]. These would-be Goldstone bosons are unphysical and can be 'rotated' away by an appropriate gauge transformation to the so-called unitary gauge.

For a particular ground state, an expansion can be made around the minimum. A common representation in the unitary gauge is as follows:

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix}, \tag{1.13}$$

where $H(x)$ represents the scalar Higgs field, for which its excitations correspond to the electrically neutral scalar Higgs boson with mass $\sqrt{2\lambda v^2}$. By evaluating $\mathcal{L}_{BEH}$ from Equation (1.12) with (1.13), one can identify mass terms for the electroweak bosons and interaction terms with the scalar doublet, resulting in the masses:

$$m_W = \frac{gv}{2} \quad \text{and} \quad m_Z = \frac{v\sqrt{g^2 + g'^2}}{2}. \tag{1.14}$$

1.1.4 The Yukawa interaction

The introduction of the scalar doublet also provides us with a mass generation mechanism for the electrically charged fermions. These masses are generated by interactions between the fermion fields and the Higgs doublet, known as Yukawa interactions. In the most general form, based on the notations from Equation (1.8), the SM Yukawa Lagrangian is given by:

$$\mathcal{L}_{Yukawa} = -y_d^{ij} \bar{Q}_L^i \Phi d_R^j - y_u^{ij} \bar{Q}_L^i i\sigma_2 \Phi^* u_R^j - y_e^{ij} \bar{L}_L^i \Phi e_R^j + (h.c.), \tag{1.15}$$

where y^{ij} are 3×3 non-diagonal complex matrices or Yukawa couplings that can be diagonalized. After spontaneous symmetry breaking, one can show that the Yukawa Lagrangian in the unitary gauge becomes:

$$\mathcal{L}_{Yukawa} = -\sum_f \frac{v y_f}{\sqrt{2}} \bar{f} f (1 + \frac{H}{v}) := -\sum_f m_f \bar{f} f (1 + \frac{H}{v}), \tag{1.16}$$

with f a fermion field. By introducing Yukawa terms into our theory, we not only generate mass terms for fermions but also interaction terms with the Higgs boson. The relation between the mass m_f and the Yukawa coupling y_f for an electrically charged fermion f following the SM framework can be identified as:

$$y_f = \sqrt{2}\frac{m_f}{v}. \tag{1.17}$$

It is worth noticing that the linearity of this relation implies that heavy fermions couple more strongly to the Higgs boson.

1.2 Probing the Yukawa couplings

1.2.1 State-of-the-art

Although the SM predicts a linear relation between a charged fermion's mass and the corresponding Yukawa coupling, the theory does not provide a prediction of any concrete values. Both the

masses and the couplings are free parameters that must be probed experimentally. The main framework for measuring the couplings is the *κ-framework*, where coupling modifiers are introduced to parametrize potential deviations from the SM predictions [16]. In particular, for each fermion F and massive gauge boson V, a coupling modifier is defined as the ratio of the measured and expected coupling with the Higgs boson:

$$\kappa_F = \frac{y_F}{(y_F)_{\text{SM}}} \quad \text{and} \quad \kappa_V = \frac{g_V}{(g_V)_{\text{SM}}}, \tag{1.18}$$

where y_F and g_V stand for the experimental values of the Yukawa coupling and the coupling between the Higgs and gauge boson V, and the subscript SM denotes their theoretical values following the SM. As a consequence of these definitions, the coupling modifiers are all equal to unity in the SM. Currently, the modifiers have been measured up to great precision for the muon, the massive vector bosons, and the third-generation fermions [17]. These measurements are conducted through studies of processes directly sensitive to the Higgs couplings, such as the decay channels of the SM Higgs boson. For example, the measured fermion Yukawa modifiers κ_f were studied via the $\mathrm{H} \to \mathrm{f\bar{f}}$ decay channel. The only exception is the top quark. Being the heaviest particle in the SM with a mass of 172.76 GeV [14], the Higgs boson kinematically cannot decay into a $\mathrm{t\bar{t}}$ pair. To probe the κ_t coupling modifier, Higgs boson production in association with top quarks was studied instead [18].

Returning to the $\mathrm{H} \to \mathrm{f\bar{f}}$ processes, not every decay channel has the same branching fraction. Furthermore, signal-insensitive processes giving rise to the same final state, commonly called backgrounds, must be considered as well. For example, the $\mathrm{H} \to \mathrm{b\bar{b}}$ is the most abundant decay channel of the SM Higgs boson with a branching fraction $\mathcal{B}(\mathrm{H} \to \mathrm{b\bar{b}}) = 58.2\%$ [14], but involves dealing with large QCD backgrounds. On the other hand, the $\mathrm{H} \to \mu^+\mu^-$ channel has a low branching fraction $\mathcal{B}(\mathrm{H} \to \mu^+\mu^-) = 2.18 \times 10^{-4}$ [14] but involves much less QCD background. Instead, it will experience leptonic backgrounds, which need to be properly accounted for given the expected low yields. To be general, backgrounds can pose significant challenges for the reconstruction of the signal, which in turn may result in a loss of sensitivity to the coupling parameter in question. Nevertheless, a general approach to maximise the sensitivity to a search is to take into account the various production modes of the SM Higgs boson as well. In the search for the $\mathrm{H} \to \mu^+\mu^-$ decay, for example, potential signal-like events are separated into orthogonal categories based on the signatures of the Higgs boson production processes: gluon-gluon fusion (ggH), produced in association with top quarks (ttH), produced via vector boson fusion (VBF), or produced in association with a massive vector boson (VH). These event categorization techniques make it possible, also bearing in mind that the CMS detector is an excellent device for muon-related physics, to still find evidence for this decay channel despite its staggering low branching fraction [19]. For the massive gauge bosons, the coupling modifiers were studied through SM Higgs boson decay channels as well. More specifically, the κ_Z modifier is studied through the $\mathrm{H} \to \mathrm{ZZ}^* \to 4\ell$ (ℓ = e or μ) decay channel[3], whereas the κ_W modifier involves the $\mathrm{H} \to \mathrm{WW}^*$ decay process.

The values of the coupling modifiers (κ_W, κ_Z, κ_t, κ_τ, κ_b, and κ_μ) are then extracted by a simultaneous maximum likelihood fit across the entire collection of event categorizations of all processes, where the coupling modifiers are handled as free-floating parameters. In the SM, the scaling of the coupling with the mass depends on the fermionic or bosonic nature of the particle. While Equation (1.17) tells us that the Yukawa coupling y_f scales linearly with the mass m_F of a fermion, the vector boson couplings are proportional to the square of the mass of the vector boson

[3]The Z^* stands for an off-shell Z boson, which is the general superscript notation for a particle that does not satisfy the Einstein energy-momentum relation. If a particle does satisfy this relation, it is called an on-shell particle.

m_V. To give an overview of the couplings, taking into account the nature of the particle, reduced coupling modifiers are introduced:

$$\kappa_F \frac{m_F}{v} \quad \text{and} \quad \sqrt{\kappa_V} \frac{m_V}{v}, \tag{1.19}$$

where v is the vacuum expectation value of the Higgs field. The reduced coupling strengths for the massive vector bosons (denoted with a subscript V), the third generation fermions, and the muon (denoted with a subscript F) with the 68% confidence level intervals from the simultaneous fit, are shown in Figure 1.3.

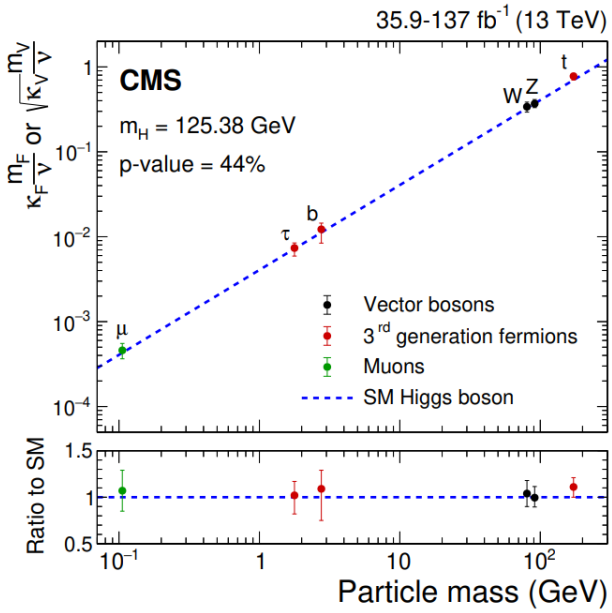

Figure 1.3: *The experimentally determined Yukawa couplings for the massive vector bosons (denoted with a subscript V), the third generation fermions, and the muon (denoted with a subscript F) with the 68% CL intervals as the error bars. The blue dotted line shows the SM prediction of the reduced coupling modifiers, where $\kappa_{V,F} = 1$, and the lower panel shows the ratios of the coupling values and their prediction. Image taken from [19].*

1.2.2 Measuring the charm-Higgs Yukawa coupling

While the Yukawa couplings of the third-generation fermions are statistically compatible with the SM prediction, the next frontier in the coupling measurements consists of validating the remaining second-generation fermions. Especially relevant for this thesis is the effort to measure the charm-Higgs Yukawa coupling y_c. Similar to the previous measurements, one possibility is to examine the decay channel, $\mathrm{H} \to \bar{\mathrm{c}}\mathrm{c}$, which has a branching fraction $\mathcal{B}(\mathrm{H} \to \bar{\mathrm{c}}\mathrm{c}) = 2.89\%$ [14]. However, this direct search does pose some significant challenges, mainly in terms of identification. The low branching fraction compared to $\mathrm{H} \to \bar{\mathrm{b}}\mathrm{b}$ leaves the charm signal almost completely overshadowed by backgrounds. Furthermore, distinguishing the objects resulting from a quark in the final state, known as jets, based on the flavour of the quark is a challenge on its own. This is handled by dedicated algorithms, known as jet tagging algorithms, which will be further explained in Chapter 3. These impracticalities lead to far less accurate measurements of y_c. At the time of writing, the most sensitive search in the direct decay channel comes from a recent analysis from the CMS collaboration, where events in which a Higgs boson is produced along with a W

or Z boson are targeted [20]. With the acquired dataset, a limit was set on the signal strength modifier μ, defined as $(\sigma \times \mathcal{B})/(\sigma \times \mathcal{B})_{\mathrm{SM}}$. In the previously introduced κ-framework, the signal strength modifier can be related to the charm-Higgs Yukawa coupling modifier κ_c:

$$\mu_{\mathrm{VH(H\to c\bar{c})}} = \frac{\kappa_c^2}{1 + \mathcal{B}_{\mathrm{SM}}(\mathrm{H} \to \mathrm{c\bar{c}})(\kappa_c^2 - 1)}, \tag{1.20}$$

from which the analysis reports an observed (expected) upper limit of $1.1 < |\kappa_c| < 5.5$ ($|\kappa_c| < 3.4$) at 95 % confidence level (CL). While this direct search approach is becoming more sensitive due to the continuous improvements made in the reconstruction and tagging of charm jets, it is not the only approach we have for measuring the charm-Higgs Yukawa coupling. Instead of investigating processes sensitive to the coupling through the decay of the Higgs boson, we turn our attention to a process sensitive at the production stage of the Higgs boson. More specifically, the process of interest for this thesis is the production of a Higgs boson associated with a charm quark, g+c $\to$ H+c , where the dependence on the coupling comes from the interaction of the charm quark in the initial state. The LO Feynman diagrams in the SM that contribute to the g+c $\to$ H+c process are shown in Figure 1.4, where a blue dot is placed to visualize where the coupling y_c comes into play.

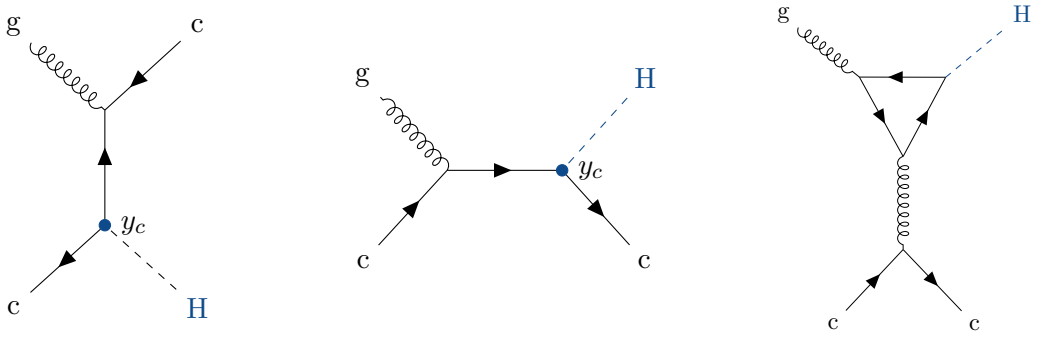

Figure 1.4: *The three LO Feynman diagrams which contribute to the g+c $\to$ H+c process in the SM. A blue dot is placed to visualize where the Yukawa coupling y_c enters. The rightmost Feynman diagram does not have a dependence on y_c, but instead shows a triangle fermion loop.*

If we compare the H+c process to the Higgs boson production process with the $\mathrm{H} \to \mathrm{c\bar{c}}$ decay, we immediately notice that instead of having to identify a pair of jets, now only one jet is required. Furthermore, being a process sensitive in the production stage, we have the freedom to reconstruct the Higgs boson in a channel that is most suitable for us [21]. However, there's no rose without a thorn. As shown in Figure 1.4, there is one contribution to the g+c$\to$ H+c process which is insensitive to y_c and instead a fermion loop is introduced. The main particle residing in these loops is the top quark, resulting from being the heaviest fermion in the SM. In terms of total contribution, this Yukawa-insensitive part will thus also dominate compared to the Yukawa-sensitive parts, making only a small fraction of the observed H+c final states signal-like.

Being a relatively novel procedure to measure the charm-Higgs Yukawa coupling, not many experimental searches involving the H+c final state have been conducted as of yet. One recent analysis, which is also the first published search from the CMS collaboration on this final state, focused on the decay channel of the Higgs to photons, $\mathrm{H} \to \gamma\gamma$, and reported an observed (expected) limit of $|\kappa_c| < 38.1$ ($|\kappa_c| < 72.5$) at 95% CL [22]. While the diphoton channel is promising, especially considering that the sensitivity will strongly improve with the acquisition of more data in the future, we will shift the focus to the $\mathrm{H} \to \mathrm{ZZ}^* \to 4\ell$ decay channel, visualized in Figure 1.5. Whenever we consider a Higgs final state drawn in a Feynman diagram, we will always

assume this decay channel. Despite being quite an elusive process, with a branching fraction of

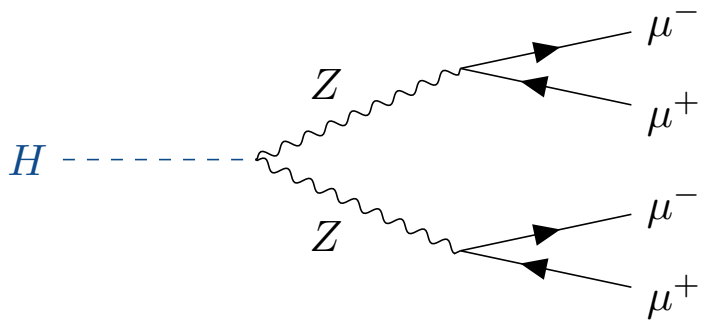

Figure 1.5: *The Feynman diagram showing the Higgs decay to two Z bosons, which each subsequently decay into a muon-antimuon pair.*

$\mathcal{B}(\mathrm{H} \to \mathrm{ZZ}^* \to 4\mu) = 0.03\%$ [14], it is regarded as one of the *golden* decay channels of the Higgs boson. The primary reason is the fully leptonic signature, which significantly reduces the amount of non-Higgs background compared to the almost overwhelming QCD background encountered in the traditional direct search method. Moreover, as we previously mentioned and will explain in the next chapter, the CMS detector excels in performance as a muon detector, further solidifying why the decay channel is so desirable. Nevertheless, specific backgrounds are still present, which is precisely the topic of the final section of this first chapter.

1.3 Backgrounds in the H+c final state

Suppose we want to extract meaningful information about the charm-Higgs Yukawa coupling through the H+c process. In that case, we must first characterise any other process that can mimic our signal by producing an identical final state. Without properly accounting for these backgrounds, we might mistake them for signal, biasing any conclusion we make. Furthermore, upon investigating collision data, we will observe the contributions from both the signal and various background processes. It is crucial to distinguish these components and quantify the amount of each present in the data. In our case, the final state of interest consists of four muons resulting from the $\mathrm{H} \to \mathrm{ZZ}^* \to 4\mu$ decay, associated with a jet produced by a charm quark. Here, the muons are often referred to as being *prompt*. This term describes muons that originate from the decay of short-lived particles (such as the Higgs or a Z boson), which are typically well isolated from other surrounding activity. Considering the other signature, the charm jet, it is an essential component of our H+c signal. Still, its presence will only serve as an additional requirement on top of the phase space defined by requiring four prompt muons. Only through them can we reconstruct a Higgs boson candidate, after which we have to associate it with a jet. Therefore, we consider any other process that can produce a prompt four-muon signature as a background process for our signal. They can generally be divided into two categories: *irreducible backgrounds* and *reducible backgrounds*, which we explain in the following paragraphs.

Irreducible backgrounds refer to the processes that produce a four-prompt muon final state. We refer to them as *irreducible* because, from the perspective of which final state particles are present, they are indistinguishable from our signal, meaning no selection or identification criteria would be able to completely remove them without also losing signal. However, these processes might still differ in production mechanisms or kinematics. Some may have additional leptons or jets present in addition to the four prompt muons, which can be exploited to separate them from

the signal. They can be further subdivided into two categories. The first consists of any process through which a Higgs boson is produced and decays via the $H \to ZZ^* \to 4\mu$ channel. We call these processes the *irreducible Higgs boson backgrounds*. The backgrounds belonging to this category that we will consider are the following: Yukawa-insensitive H+c production; Higgs boson production in association with a b quark or other light quark l (H+b, H+l); gluon-gluon fusion (ggH); Higgs boson production in association with top quarks (ttH); Higgs boson production via vector boson fusion (VBF H); Higgs boson production in association with a massive vector boson (VH); and Higgs boson production in association with a top quark and other flavoured quark (tqH). The LO Feynman diagram for Yukawa-insensitive H+c production is shown in the rightmost diagram of Figure 1.4. Considering the H+b and H+l processes, the diagrams are the same as for H+c, but with the charm quark substituted for a bottom or light quark. Therefore, we will not explicitly draw them. Representative diagrams of the other irreducible Higgs boson backgrounds are summarized in Table 1.1. The second subcategory consists of processes that produce four prompt muons but not through any Higgs boson decay. Consequently, we refer to them as *irreducible non-Higgs boson backgrounds*. Within this category, we consider the following processes: ZZ production through gluons (gg→ZZ), where each Z decays via the $Z \to \mu^+\mu^-$ channel; ZZ production through quarks ($q\bar{q} \to ZZ$), where each Z decays via the $Z \to \mu^+\mu^-$ channel; $Z \to \mu^+\mu^-$ production with initial state radiation that undergoes internal conversion $\gamma^* \to \mu^+\mu^-$ ($Z\gamma^* \to 4\mu$); and single resonant four-lepton production ($Z \to 4\mu$). Representative Feynman diagrams of these processes are shown in Table 1.2.

On the other hand, one or more non-prompt muons can appear in our prompt muon final state. This can occur when a particle is misreconstructed and misidentified as a prompt muon. For example, a jet giving rise to the signature of a prompt muon or a non-prompt muon passing the identification of prompt ones. These are known as *reducible backgrounds*. We call them *reducible* because in an ideal world with a perfect detector and selection efficiencies, these would vanish. Unfortunately, this is not realistic. Some detector-related or identification problems are inevitable, and we must account for these effects in our signature. In principle, any process that produces fewer than four prompt muons and includes jets can contribute to the reducible background. Since we will be analyzing proton-proton collisions, the requirement of jet presence is almost guaranteed. As will be explained in Chapter 3, these collisions are very complex, and jets follow naturally from them. Therefore, when referring to a reducible process X, we imply X+jets, which reflects that jets are expected in the final state regardless of whether they are part of process X or not. How large the contribution of a certain reducible background will be depends on its baseline signature (i.e., how many prompt muons it naturally produces) and on its production cross section. The dominant reducible backgrounds in our analysis will be Drell-Yan production (Z/γ^* + jets), in which Z/γ^* decays to a pair of oppositely charged muons, and $t\bar{t}$ production, where both top quarks decay to a bottom quark and a W boson that each decay leptonically. For these backgrounds to mimic the four-prompt muon final state, at least two additional objects (typically jets) must be misidentified as prompt muons. Other relevant processes include, for example, $W^{\pm}Z$ production, where the $W^{\pm}$ decays leptonically via $W^{\pm} \to \mu^{\pm}\nu_{\mu}$, and the Z boson via $Z \to \mu^+\mu^-$. In this case, only one extra object must be identified. However, the cross section will be significantly lower compared to the previously mentioned Drell-Yan and $t\bar{t}$ processes. While other processes exist, the main ones we will consider are: Drell-Yan production; $q\bar{q} \to ZZ$ where one of the Z bosons decays to a $q\bar{q}$ pair; $t\bar{t}$ production; $W^{\pm}Z$ production; and W^+W^- production. A representative Feynman diagram for the reducible ZZ background is already shown in Table 1.2, while diagrams for the other reducible backgrounds are shown in Table 1.3.

In this thesis, the charm-Higgs Yukawa coupling is statistically probed through the previously mentioned H+c process, where the Higgs decays via the $H \to ZZ^* \to 4\mu$ channel and the charm quark develops a jet. This will be done by setting a 95% CL upper limit on the signal strength μ_{H+c}, which quantifies the H+c production rate relative to the SM prediction. To derive this limit, both the irreducible (shown in Table 1.1 and Table 1.2) and reducible background processes (shown in Table 1.3) must be estimated, and the signal must be characterised. The development of the reconstruction techniques for events involving a Higgs boson and a jet is presented in Chapter 4. There, signal characterisation and estimation of irreducible backgrounds take place using simulated proton-proton collisions. The estimation of the reducible backgrounds is performed using a data-driven approach described in Chapter 5. Once the background components are accounted for, the expected upper limit is extracted and presented in Chapter 6. Before these analyses are carried out, Chapters 2 and 3 give the necessary experimental context. Chapter 2 introduces the Large Hadron Collider (LHC) and the Compact Muon Solenoid (CMS) detector, which are used to produce and detect proton-proton collisions, respectively. Chapter 3 then discusses the physics of these proton-proton collisions, explaining how collision events are simulated and reconstructed in general.

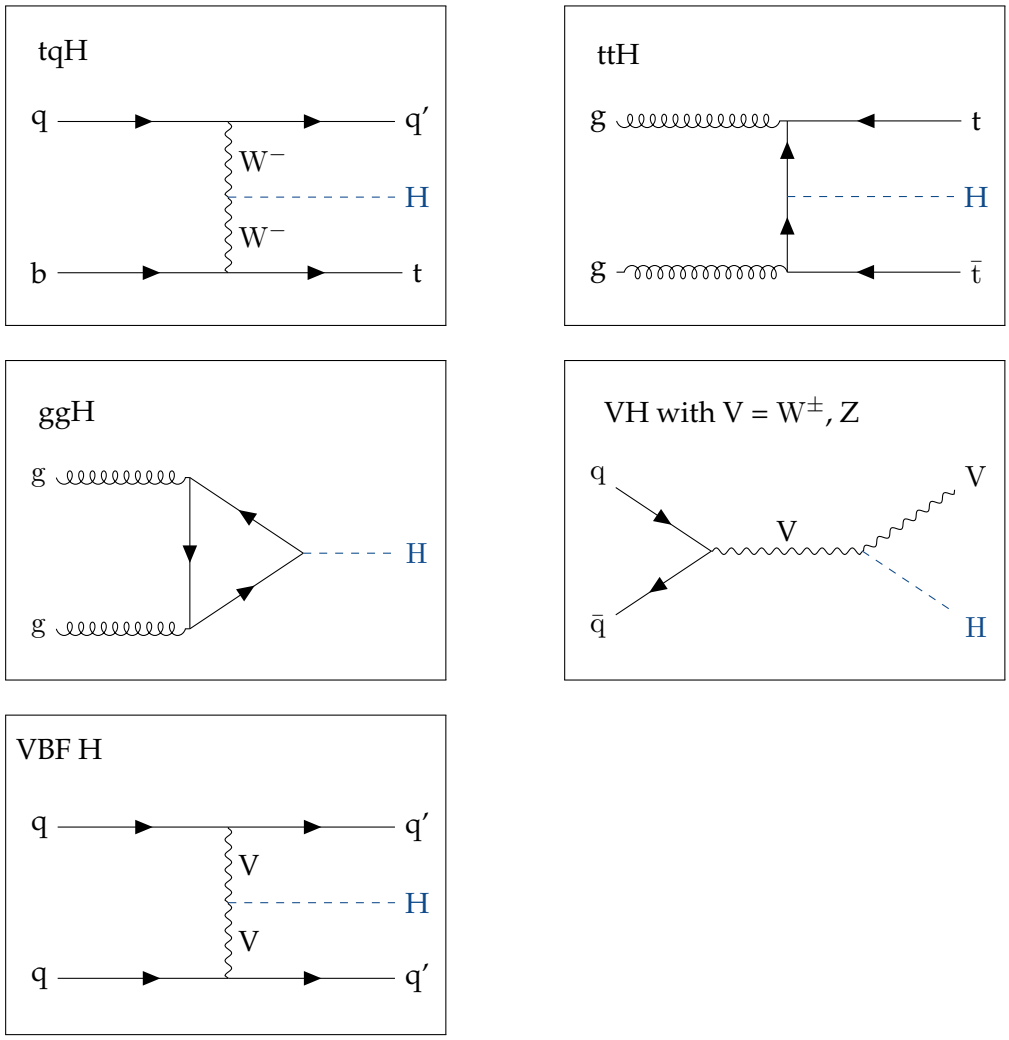

Table 1.1: *Representative Feynman diagrams of a selection of the irreducible Higgs boson backgrounds considered in this work, namely the* tqH, ttH, ggH, VH, *and* VBF H *processes.*

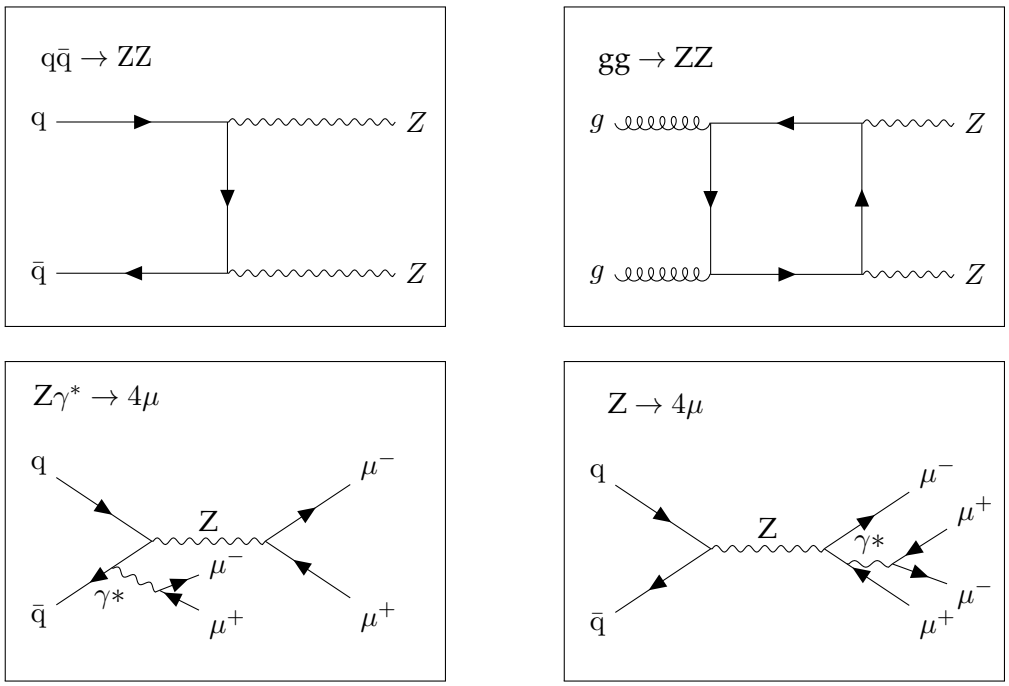

Table 1.2: *Representative Feynman diagrams of the irreducible non-Higgs backgrounds considered in this work.*

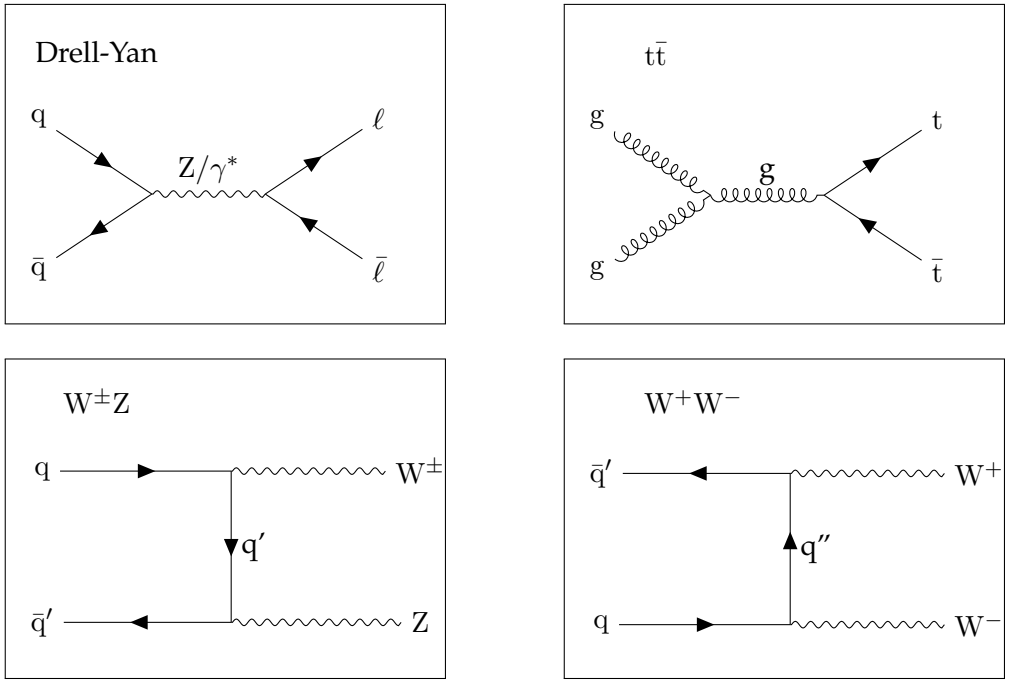

Table 1.3: *Representative Feynman diagrams for the* Drell-Yan, $t\bar{t}$, $W^{\pm}Z$, *and the* W^+W^- *reducible backgrounds considered in this work. Jet production is implicitly assumed.*

2 THE CMS EXPERIMENT AT THE LARGE HADRON COLLIDER

Theories are there to be tested, and the SM is no exception. The CERN (European Organization for Nuclear Research) laboratory in Geneva is one of the primary testing grounds of our theory, and will be the experimental setup for this thesis. It houses the most powerful particle accelerator in the world, known as the Large Hadron Collider (LHC). In Section 2.1, we will explain the design and operation of this machine, followed by a brief discussion of its successor, the High-Luminosity LHC (HL-LHC). To study the collisions provided by the LHC, several particle detectors have been constructed. One of these devices is known as the Compact Muon Solenoid (CMS) detector and is the subject of the second part of this chapter, Section 2.2. After a description of the CMS coordinate system, we will give an overview of how the CMS detector detects particles via dedicated subdetector systems. To end this chapter, we will briefly discuss how the CMS detector is preparing for the HL-LHC era, more precisely with its Phase-2 Upgrade.

2.1 The Large Hadron Collider

2.1.1 Design & operation

The Large Hadron Collider (LHC) at CERN (European Organization for Nuclear Research) is the worlds leading circular particle collider for high-energy physics research. Installed in a 26.7 km circumference underground tunnel, originally constructed for the Large Electron-Positron Collider (LEP), the LHC probes the SM at TeV scales by accelerating and colliding beams of protons. The design plans for the LHC were approved in 1994 by the CERN council, and construction started in 1998. A key factor in shaping the LHC was the search for the Brout-Englert-Higgs boson [23]. Direct searches for *Higgs-strahlung* (ZH) events at LEP had excluded the possibility of the Higgs boson having a mass below 114.4 GeV at a confidence level of 95 % [24]. Theoretically, it was also established that the Higgs boson must have had a mass below 1 TeV to ensure the unitarity of WW scattering processes [25]. As the only accelerator capable of exploring the energy range within these limits, the LHC and its surrounding particle detectors were placed in a unique position: if the Higgs boson existed, they were bound to find it.

One of the characterising concepts for any particle accelerator is the *luminosity* $\mathcal{L}$, which has two definitions. Firstly, it can refer to the number of collisions per $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, known as the *instantaneous* luminosity $\mathcal{L}_{\text{inst}}$. Together with the cross section σ, it provides us with a measurement of the event rate $\frac{dN}{dt}$ for a specific process:

$$\frac{dN}{dt} = \mathcal{L}_{\text{inst}}\sigma. \tag{2.1}$$

Simply integrating the instantaneous luminosity over time yields the second definition, the *integrated* luminosity $\mathcal{L}_{\text{int}}$. Instead of providing us with an event rate, this quantity allows us to directly determine the number of events N during the integrated time window [26]. Increasing the (instantaneous) luminosity corresponds to an increase in collisions, and potentially more data, but it has drawbacks. As will be explained in the following pages, protons are grouped into

bunches at the LHC before they collide. Therefore, multiple proton-proton collisions can occur during a crossing of two bunches, a phenomenon known as *pileup*. Not every pileup interaction will be interesting for us, but it will happen simultaneously with the ones that are. The experiments analysing the collisions, such as the CMS experiment, are thus confronted with the difficult task of correctly identifying the collision products of each interaction, without causing a mismatch between products of other pileup interactions. Needless to say, the instantaneous luminosity at the LHC is high, with the design value being $10^{34}\ \mathrm{cm}^{-2}\ \mathrm{s}^{-1}$ [27]. The data-taking periods of the LHC are divided into *runs*, with (Extended-)Year-End-Technical-Stops (EYETS) for maintenance, and a Long-Shutdown (LS) for installing further upgrades and replacements between runs. The total integrated luminosity $\mathcal{L}_{\mathrm{int}}$ provided by the LHC over the years, as well as the recorded $\mathcal{L}_{\mathrm{int}}$ by the CMS experiment, is shown in Figure 2.1. In the following paragraph, the different runs and the integrated luminosity obtained will be briefly discussed.

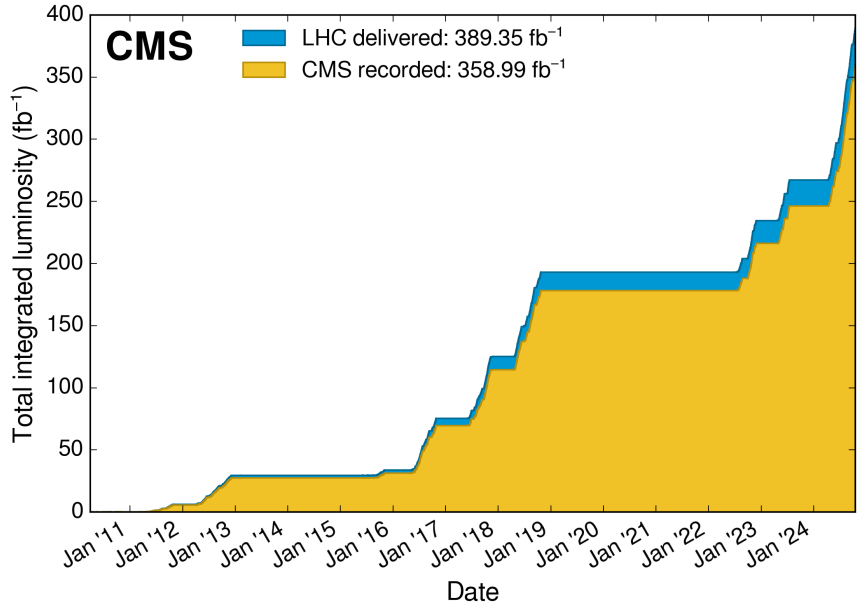

Figure 2.1: *The delivered and recorded luminosity from the LHC at the CMS experiment throughout the years. Image taken from [28].*

The first beams at the LHC were commissioned in 2008, with the first proton-proton collisions occurring in 2009. The first physics data-taking run, Run-1 (2010-2013), delivered collisions at centre-of-mass energies of 7 and 8 TeV, resulting in an integrated luminosity of about 30 fb^{-1}. The data used in this thesis were collected in 2018, the final year of LHC's second run (Run-2, 2015-2018), when the accelerator operated at a collision energy of 13 TeV (6.5 TeV per beam). In 2018 alone, the LHC delivered an integrated luminosity of 59.83 fb^{-1}. In total, Run-2 delivered an integrated luminosity of over 130 fb^{-1} [29]. The LHC is currently in Run-3, where collision energies are further increased to 13.6 TeV [30].

The protons required for the collisions are obtained by ionizing hydrogen atoms. Before reaching the LHC, they are injected into a sequence of smaller accelerators designed to progressively increase the energy of the particles as part of the CERN accelerator complex, illustrated in Figure 2.2. The first accelerator in the chain is the LINAC2, which accelerates the particles to an energy of 50 MeV. From there, they enter the Proton Synchrotron Booster (PSB), boosting the energy to 1.4 GeV, followed by the Proton Synchrotron (PS), where energies up to 26 GeV are reached. In the

PS, protons are grouped into 'bunches' by Radio-Frequency (RF) cavities. These metallic chambers house an electromagnetic field that oscillates at a specified frequency. The field oscillation accelerates or decelerates protons along the beamline based on their synchronization with the oscillation frequency, clumping them into bunches. Under nominal circumstances, there are 2808 proton bunches in a single beam, with each bunch containing $\mathcal{O}(10^{11})$ protons. After the PS, the Super Proton Synchrotron (SPS) accelerates the bunches to the LHC injection energy of 450 GeV. At this stage, the bunches are injected into the two beam pipes of the LHC, where they circulate in opposite directions, and are further accelerated by RF cavities until the current maximum energy of 6.8 TeV is reached. The circulation of the beams along the LHC is made possible by 1232 superconducting dipole magnets installed around the ring. By flowing currents of up to 11850 A through the coils, a magnetic field strength of 8.3 T is generated, which provides the necessary curvature to keep the particle beams on track. To minimize resistance and energy loss induced by these high currents, the magnets are cooled to 1.9 K using liquid helium, allowing them to operate in their superconductive region. Additionally, quadrupole magnets are used before the collision points that focus the beams (i.e., the size of the bunches is reduced) to increase the interaction rate [23, 27]. The LHC provides four points at which the beams can collide, indicated with yellow

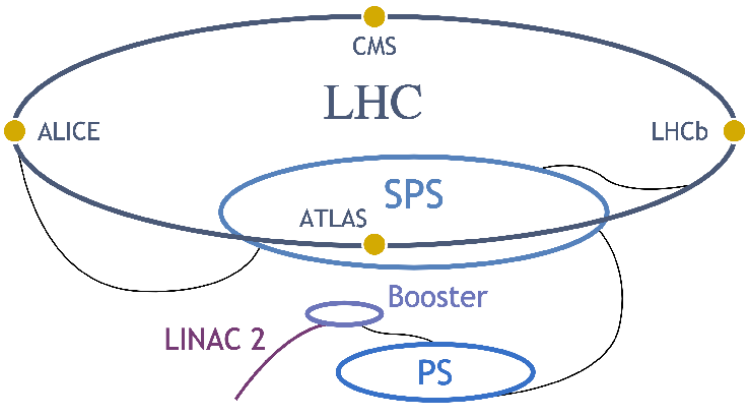

Figure 2.2: *The CERN accelerator complex. Image taken from [31].*

dots in Figure 2.2. At each of the collision points, a particle detector has been built. These are known as the CMS (Compact Muon Solenoid) detector, the ATLAS (A Toroidal LHC ApparatuS) detector, the LHCb (LHC Beauty) detector, and the ALICE (A Large Ion Collider Experiment) detector. The CMS and ATLAS detectors are considered general-purpose, while the other two detectors have more specific goals. The LHCb experiment mainly studies the interactions of b hadrons to contribute to understanding the matter-antimatter asymmetry in our universe. Dedicated to heavy-ion physics, the ALICE experiment investigates the lead-lead collisions that the LHC provides during certain periods of the physics runs.

2.1.2 The High-Luminosity LHC and onwards

The LHC has provided us with many new insights into particle physics. Nonetheless, every good story has an ending. For the LHC, this comes with the end of Run-3 in June 2026. Afterwards, the LHC will enter its third Long Shutdown (LS3) for extensive upgrades, transitioning into the High-Luminosity LHC (HL-LHC) era. A schematic overview of the LHC and HL-LHC timeline is given in Figure 2.3. Designed to be operational in 2030, the HL-LHC aims to achieve centre-of-mass energies of 14 TeV and an instantaneous luminosity five times the design value of the LHC, $5 \times 10^{34}\ \mathrm{cm}^{-2}\ \mathrm{s}^{-1}$. In terms of integrated luminosity, more than 3000 fb^{-1} is envisioned, coming close to 10 times beyond what the LHC will bring [32]. This upgrade will extend the sensitivity

to physics beyond the SM and allow for a deeper delve into the Higgs sector. Indeed, one of the main objectives of the physics program consists of measuring the Higgs self-coupling λ, for which the datasets gathered with the LHC are not sensitive enough.

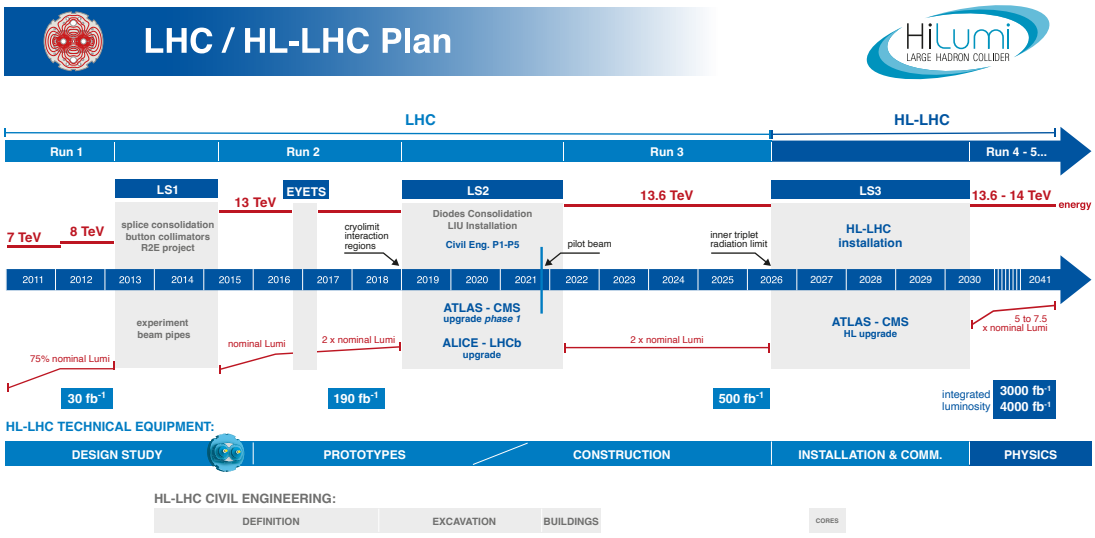

Figure 2.3: *The timeline of the LHC and the future HL-LHC up until 2041. Image taken from [33].*

While the HL-LHC will set the stage for a new generation of particle accelerators and detectors, what happens after is unclear. Whether or not CERN will stay the global centre for particle physics research will be revealed in the upcoming years. One possibility, which is also the direct follow-up of the LHC, is the Future Circular Collider (FCC). Currently in the stage of feasibility studies, the FCC would be a two-stage collider installed in a 91 km tunnel. In the beginning stage, it would house an electron-positron collider (FCC-ee), mainly aimed at electroweak and Higgs studies. Later on, it would transition into a hadron collider at centre-of-mass energies of 100 TeV (FCC-hh) [34]. However, other propositions, such as Japan's International Linear Collider (ILC), have also been made [35].

2.2 The Compact Muon Solenoid experiment

2.2.1 The CMS coordinate system

The collision of proton beams in the LHC can produce new particles, which may be stable or decay into other particles. To study and characterise the collision products, the CMS detector was constructed [36]. This cylindrical detector measures 26.1 m in length and 14 m in diameter, and the kinematics of the collisions are described with the cylindrical CMS coordinate system. In this subsection, we will further elaborate on this coordinate system and some important variables that can be calculated from it. The origin is the nominal collision point of the beams, with the x-axis pointing towards the centre of the LHC, the y-axis pointing upwards, and the z-axis aligned with the anticlockwise beam direction. The azimuthal angle ϕ is measured between the x-axis and the x-y plane, while the polar angle θ is measured from the z-axis. In a proton-proton collision, the centre-of-mass frame is not fixed, but rather boosted along the z-axis. The underlying reason is that actual collisions occur between the partons, which are the constituents of the proton that carry only a fraction of the proton's momentum. This fraction is not fixed, but varies from event to

event. Bearing this in mind, we would like to describe the collisions in a way that is independent of these unknown boosts by using kinematic variables that provide us with Lorentz-invariant quantities. One such variable is the rapidity y, defined as:

$$y = \ln\left(\frac{E + p_z}{E - p_z}\right), \tag{2.2}$$

where E and p_z are the energy and z-component of the momentum, respectively. One of the key properties of rapidity is that differences in rapidity are Lorentz-invariant. However, measuring y is difficult, as it requires knowledge of the boosted longitudinal momentum p_z. In the highly relativistic limit, however, one can show that the rapidity simplifies to the pseudorapidity η:

$$\eta = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right]. \tag{2.3}$$

Like with rapidity, differences in pseudorapidity are Lorentz-invariant [5]. The key difference lies in it being more easily measurable, needing only a measurement of the polar angle θ. For this reason, the pseudorapidity η has become the standard parameter to describe the angles of the particles relative to the beam axis.

For energy and momentum, then, the standard measurement quantities come from the components in the plane transverse to the beam direction, E_{T} and $\mathrm{p_T}$, which are derived from the x and y components. Apart from not being able to instrument close to the beam-line, one of the main reasons is again related to the partons colliding. As a result of the varying fraction of energy the partons carry, we don't have a good estimate of the overall initial energies and momenta, especially along the beam axis. In contrast, for the transverse plane, we do. As the collision mostly happens along the z-axis, the transverse components are initially zero, and conservation of momentum implies it to be zero after the collision as well. Not only are the transverse plane quantities well measurable, but they also provide the only experimental handle on detecting weakly interacting particles, such as neutrinos. While these particles are completely invisible to the CMS detector, their presence is inferred from the missing transverse momentum $\vec{\not{E}}_T$, defined as the negative vector sum of the transverse momenta $\vec{p}_T$ of all the particles in a collision event [37]:

$$\vec{\not{E}}_T = -\sum \vec{p}_T. \tag{2.4}$$

Calculating the missing transverse momentum necessarily implies that all the collision products in the event need to be identified. Consequently, CMS should consist of various subsystems, each designed to detect a specific particle type, which is precisely the topic of the next subsection.

2.2.2 The CMS detector

The CMS detector has an onion-shaped design structure, where each layer acts as an individual subdetector. A key feature of the CMS detector is its superconducting solenoid, which generates a 3.8 T magnetic field to bend the trajectories of charged particles, such as muons. By combining the data of the different layers, collision events can be reconstructed. A transverse slice of the CMS detector and how different particles interact with the detector is shown in Figure 2.4. In the following paragraphs, each subdetector will be briefly discussed.

Silicon tracker

The innermost layer of the CMS detector is the *silicon tracker system*. Being the subdetector closest to the interaction point, it is the first layer that the collision products encounter. The task of

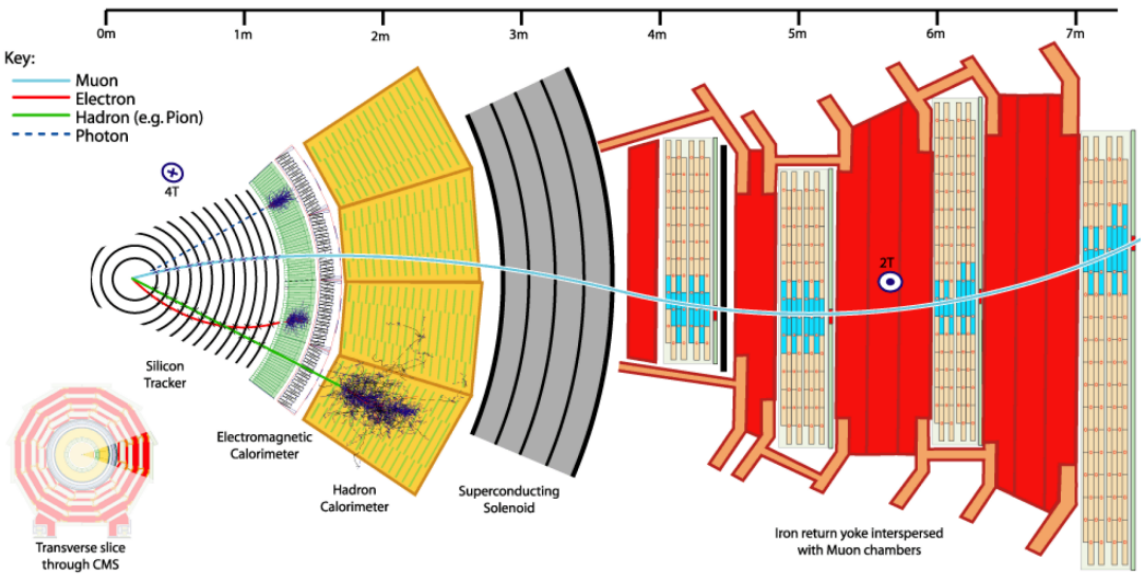

Figure 2.4: *A transversal slice through the CMS detector, showing the different layers and how the particles interact with them. Image taken from [38].*

the tracker is to measure the trajectory of charged particles, distinguish interaction vertices corresponding to multiple collisions in the same bunch crossing (=*pileup*), and reconstruct secondary vertices. The system has an acceptance of $|\eta| < 2.5$ and consists of a silicon pixel detector and a silicon strip tracker. A charged particle that passes through the pixels and strips can have sufficient energy to excite the silicon atoms, producing an electrical signal or *hit*. By combining the different hits across the system, the trajectory of the charged particle can be reconstructed. Even the momentum and charge of the particle can be inferred from the trajectory's bending due to the magnetic field and the direction of the curved path. These reconstructed trajectories, called *tracks*, can be extrapolated to one common point called a *vertex*. The vertex that has the highest summed p_t^2 formed by its tracks is known as the *primary vertex* (PV) of a collision event. It is referred to as being primary because we associate the most energetic interaction of a collision, the so-called *hard scattering*, with it [39]. More information on this will be presented in the next chapter.

Electromagnetic calorimeter

The layer following the tracker is the *electromagnetic calorimeter* (ECAL). This subdetector is designed to measure the energy of photons and electrons with the use of lead tungstate ($\mathrm{PbWO_4}$) crystals placed in the barrel ($|\eta| < 1.479$) and endcap ($1.479 < |\eta| < 3$) regions. The energy measurement is based on the *scintillation* principle. The photons and electrons that pass through the crystals induce electromagnetic showers. The secondary particles generated from these showers further interact with these crystals, producing an amount of light proportional to the particle's energy. A photodetector captures the emitted light, which is converted to an electrical signal that is further analyzed. The total amount of light emitted thus allows for an estimate of the energy of the impinging particle. An additional detector, the Endcap Preshower, has been installed in the endcap regions to distinguish closely spaced photons from π^0 decay and isolated single photons [40].

Hadronic calorimeter

The hadrons produced in the collisions generally fly through the ECAL, and mainly only interact with the surrounding layer, the *hadronic calorimeter* (HCAL). This subdetector is made out of alternating layers of thick absorbing and scintillator materials. Hadrons that pass through the absorbing layer initiate a hadronic shower, also known as a *cascade*. The particles originating from this cascade enter the scintillating layer, resulting in the production of scintillation light. The more energetic a particle is, the more layers it can penetrate, and thus the stronger the signal is. The HCAL must be able to measure the energy of all the interacting hadrons coming from the collision, especially relevant for the measurement of the jets and $\vec{\not{E}}_T$. For this reason, the HCAL extends close to the beampipe, having an acceptance of $|\eta| < 5.2$ [41].

Muon system

As its name suggests, muon detection and reconstruction are central to the CMS detector. Being minimally ionizing particles, muons typically fly through the whole set of previously described subdetectors, only leaving a trace primarily in the tracker system[1]. Their detection occurs in the outermost layer, the muon system, which consists of four stations, each with multiple detection planes, embedded within the superconducting solenoids steel return yoke. Each muon station contains different types of gas-ionization particle detectors, collectively referred to as muon chambers. There are four types of chambers used in the CMS detector: Drift Tubes (DT), Cathode Strip Chambers (CSC), Resistive Plate Chambers (RPC), and Gas Electron Multipliers (GEM). Each of them will be shortly explained in what follows:

- **DT:** The DTs are located in the barrel region, with an acceptance of $|\eta| < 1.2$. Each device comprises individual cells with a wire in the centre, filled with an Ar/CO_2 gas. A muon traversing through the tube will ionize the gaseous mixture, liberating electrons that drift towards the central wire along the electric field lines. The position of the crossing muon is determined by measuring the electron drift time to the wire. A muon track can be reconstructed by combining the positions measured across different DTs.

- **CSC:** The CSC cover the endcap regions, up to $|\eta| < 2.4$. The basic detection element in a CSC is a trapezoidal layer consisting of anode wires, placed perpendicular to the cathode strips inside a gaseous volume. Similar to the DT, a passing muon will ionize the medium, causing the freed electrons to move to the anode wires while the ions move to the cathode. A single CSC has six of these layers, allowing for the reconstruction of the muon's trajectory in the chamber.

- **RPC:** Both of the previous systems are complemented by the RPC. The design of an RPC is made up of two parallel plates separated by a small volume of gas. In this setup, ionization electrons coming from a muon are picked up by readout strips. Since the RPC is less dependent on the drift time of the electrons, the time resolution of its measurements is higher compared to the DT and CSC. However, its simpler design results in a less accurate spatial resolution [36].

- **GEM:** The newest additions to the muon system are the GEM muon chambers. The use of GEM technology is mainly motivated by its high rate capability. Hence, GEM chambers are installed in the endcaps to enhance the performance in these regions, with the first detectors installed during Long Shutdown 2 of the LHC. The central component of a GEM-based

[1]While they also pass through the calorimeters, the energy deposited is usually very low (a few GeV), making it easily misinterpreted as noise or a pileup contribution.

detector is a thin polymer foil, perforated with a high density of small holes and metal coatings on both surfaces. A muon GEM chamber consists of three foils between a drift cathode and a collection electrode inside an Ar/CO_2 medium. The electrons from a passing muon induce an avalanche within the holes of the GEM foils, which then drifts to the collection electrodes for readout [42].

Combining different muon chambers arranged over the stations results in a highly accurate muon spectrometer. Furthermore, with the additional information coming from the inner tracker, the CMS detector can achieve a very precise momentum resolution for the muons.

Trigger system

The amount of proton-proton collisions offered by the LHC is very high, providing the CMS detector with an event rate of 40 MHz. As data capacity is limited, storing all the data of these events is not feasible. By identifying collisions containing potentially interesting physics, the *CMS trigger system* reduces the event rate to manageable levels in a two-step process. This triggering process is irreversible, any rejected event is unrecoverable. The first filter in this system is the hardware-based *Level-1* (L1) *trigger*, which lowers the event rate to 100 kHz. The L1 Trigger is a pipelined system that operates within a 3.2 μs time window, during which it has to decide whether an event is accepted or not. Due to the small time allocated, only information from the calorimeters and the muon system is used. Each of these subsystems has a dedicated trigger that provides information about the identified objects. These trigger objects are combined in the Global Trigger (GT), which decides if an event can proceed further. The events that pass advance to the next step of the trigger system. The data of each accepted event is synchronised across the subdetectors and forwarded to the second filter, the *High-Level Trigger* (HLT). The HLT consists of a processor farm that uses software algorithms to further reduce the event rate to an order of 100 Hz. These algorithms, known as *HLT paths*, are simplified event reconstructions that require physics objects to meet specific conditions, such as thresholds on the $\mathrm{p_T}$. The remaining events are stored on the *LHC computing grid* after a full reconstruction, allowing CMS physicists from all over the world to access them [43].

2.2.3 The CMS Phase-2 Upgrade

To end this chapter, a glimpse into the future of the CMS detector is provided. It was mentioned that the LHC will transition into the HL-LHC. The increase in luminosity that HL-LHC will offer poses significant challenges for the existing CMS detector. These challenges include coping with higher levels of pileup, dealing with radiation damage due to intense particle flux, and handling complex event reconstruction at higher background levels. To address these issues, the CMS detector will be upgraded as part of the CMS Phase-2 upgrade. The existing silicon tracker has already sustained significant radiation damage and will require an upgrade. To handle the increasing pileup, the granularity of the inner tracker will also be enhanced. Similarly, current calorimeters, which have also endured radiation damage, will be replaced by the high-granularity calorimeter (HGCAL). The muon system will also undergo upgrades. The HL-LHC will demand reliable triggering in the forward regions. Moreover, the upgraded inner tracker will extend the coverage to $|\eta| = 4$. As the silicon tracker cannot exclusively identify muons, this increased range must be complemented by muon detectors. One of these upgrades consists of installing CMS ME0 GEM detectors close to the beamline, extending the acceptance of the muon system to $|\eta| < 2.8$. Many physics analyses that rely on the reconstruction of muons, such as for the $H \to ZZ^* \to 4\mu$ decay channel used in this thesis, will benefit from the increase in acceptance that the muon system will offer, as it will directly increase the signal acceptance. All of these upgrades are essential for the optimal performance of the CMS detector in the HL-LHC era [42].

3 Simulation and reconstruction of proton-proton collisions

The proton-proton collisions delivered by the LHC at CERN play a central role in experimentally validating the SM. By studying the results of these collisions with one of its four dedicated particle detectors, in this case, the CMS detector, the current theories about the fundamental particles and interactions are tested. However, experimental results only provide half of the picture. To test any theory, we must be able to make predictions. In our case, an underlying description of the proton-proton collisions and how they evolve into what a detector observes is needed. In the first part of this chapter, Section 3.1, a description of these collisions will be given. Afterwards, in Section 3.2, we discuss how they are simulated up to the detector level. For both simulation and data events, the next step consists of reconstructing physics objects from the output of the CMS detector. This will be described in Section 3.3. Particular attention will be drawn towards reconstructing the objects relevant for our upcoming analysis, namely muons and jets. Finally, in Section 3.4, we describe how one of the properties of these jets, the flavour, can be probed with the use of jet flavour identification algorithms.

3.1 The proton-proton collision

3.1.1 The stages of a collision

During our discussion of the CMS coordinate system in Section 2.2.1, it was mentioned that a proton-proton collision has to be viewed as a collision between its constituents. To further clarify this statement, the discussion of elementary particles in Section 1.1.1 is revisited. While the proton was introduced as a hadron consisting of two up and one down quark, often referred to as the *valence quarks* of the proton, its structure is more dynamic. In the proton's interior, the valence quarks interact via the exchange of gluons. However, given the nature of QCD, these gluons can interact with each other or give rise to new quark-antiquark pairs. These newly produced quarks are referred to as *sea quarks*, and can further interact with the valence or other sea quarks. Consequently, a proton consists of more than just its valence quarks. The overarching term for these constituents is called *partons*. An experimental description of these partons is provided by the *Parton Distribution Functions* (PDFs) $f(x, Q^2)$, which describe the probability for a parton to carry a fraction of the proton's momentum x at an energy transfer scale Q^2 [3]. Given this structural complexity and the unique nature of the strong interaction, a fundamental description of a proton-proton collision is highly involved and is typically broken down into different stages. Figure 3.1 shows a schematic overview of these stages. In the following paragraphs, each stage will be elaborated on [44].

The hard scattering

The process of interest in a proton-proton collision is known as *hard scattering*, which is characterised by the highest momentum transfer between two incoming partons, one coming from each proton. As a result of the hard scattering, outgoing partons and heavy particles such as our desired Higgs boson final state are produced. A correct description of this process is necessary, as

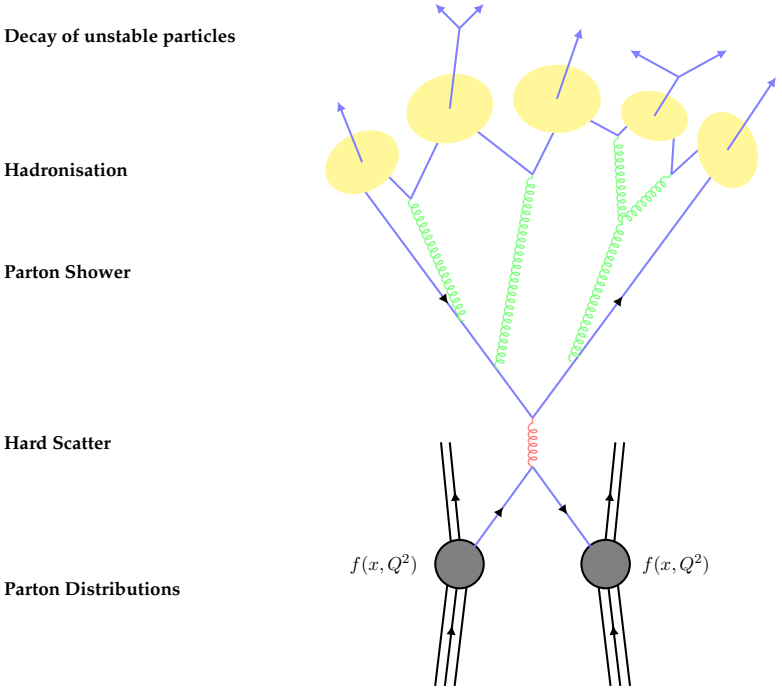

Figure 3.1: *The different stages of a proton-proton collision. Image inspired by [45].*

it is at the heart of what happens during proton-proton collisions. Here, the PDFs directly come into play, as the parton information is directly encoded inside them.

Parton showering

Being coloured particles, the partons can radiate gluons, which can create additional partons. Furthermore, quarks are also electromagnetically charged, meaning that photons, as a result of *Bremsstrahlung*, can be radiated as well. The radiation of additional particles following from one single parton is known as a *parton shower*. If the showering emerges from one of the initial colliding particles, referred to as *Initial State Radiation* (ISR). Similarly, a shower started by one of the outgoing partons of the hard scattering is referred to as *Final State Radiation* (FSR).

Hadronisation

The parton showering is not everlasting, but evolves to lower energy scales as a result of the experienced energy losses. After reaching the QCD energy scale, Λ_{QCD}, at which the theory loses its perturbative properties, the partons will undergo *hadronisation*. In this process, the produced partons combine to form colourless hadrons as a consequence of colour confinement. Most hadrons in the hadronisation are unstable and decay to stable particles, which are detectable by particle detectors such as CMS.

The underlying event

Up until this point, the parts of the proton that did not partake in the hard scattering were not mentioned. However, they can't be ignored, as the colour charge carried away by the hard-scattered parton makes them carry a non-zero colour charge too. These remnants can therefore hadronise themselves, possibly interacting with the hadronisation arising from the hard scattering. Furthermore, multiple parton collisions can also happen for one proton collision. These proton remnant hadronisation and multiple parton collisions processes are part of the *underly-*

ing event structure, which is the term for any contribution not resulting from the hard scattering process.

3.2 Simulation of collision events

3.2.1 Event generation

With the different stages of a proton-proton collision clarified, how they are simulated up to a detector level is explained. The correct simulation is the foundation of making comparisons with observations from the CMS detector, making it a crucial step. Essentially, any event sample used in the analysis for this work, whether centrally produced background event samples by the CMS collaboration or privately produced signal event samples, will have passed through the simulation chain. Every part of this chain has dedicated programs and algorithms dedicated to it. Most of this software relies on Monte Carlo (MC) techniques to account for the stochastic nature of the processes.

Simulation of the hard scattering

In the first step of the simulation chain, the process of interest emerging from the hard scattering needs to be described. All of the necessary physics of this process is encoded into its matrix element. Consequently, the software used at this stage, known as *event generators*, precisely allows for the calculation of these matrix elements. There are various event generators available, some calculating up to higher orders than others, with the most frequently used being `MadGraph5_aMC@NLO` [46] and `POWHEG` [47]. By supplying a theoretical model as input, such as the SM Lagrangian and the values of the masses and coupling constants, the software computes Feynman diagrams and their matrix elements. With these matrix elements at hand, events can be generated via sampling of the kinematic phase space.

Simulation of parton showering

The next step is to simulate the showering from initial or outgoing partons. The perturbative evolution of such a shower is governed by the DGLAP (Dokshitzer-Gribov-Lipatov-Altarelli-Parisi) equations [48]. Starting from a single parton, the mother parton, these equations describe how the probability of a splitting process to two daughter partons evolves with energy scale Q^2. The main contributions to the shower formation are the gluon radiation ($q \to qg$) and gluon splitting ($g \to gg$ and $g \to q\bar{q}$) processes for QCD, and Bremsstrahlung ($f \to \gamma f$) and pair creation ($\gamma \to f\bar{f}$) processes for QED. A further splitting of daughter particles is possible, but at a lower Q^2, because they receive only a portion of the mother particle's momentum. At some point, however, the energy of the partons is so low that the perturbativity of the model completely breaks down. Around the QCD energy scale, Λ_{QCD}, non-perturbative hadronisation models need to be considered as the next step of the simulation chain. Similar to event generators, multiple software programs exist to tackle the simulation of parton showering, with the most widely used being `PYTHIA` [49].

Simulation of hadronisation

Due to colour confinement, the coloured particles emerging from the hard scattering and subsequent parton showering have to be grouped into colourless hadrons. There is currently no model that can describe hadronisation from first principles, so phenomenological descriptions are considered instead. In the `PYTHIA` framework, the so-called *Lund string model* is used [50]. This model assumes that the colour field lines between a quark-antiquark pair, which are the QCD

equivalent of electromagnetic field lines, are confined to a tube (or string) structure as a result of the self-interacting nature of the gluons. The energy stored within this string-like configuration linearly increases with the distance r between the pair. At a large enough separation, sufficient energy becomes available to spontaneously break the string by the creation of a quark-antiquark pair. This leads to the formation of two new colourless quark-antiquark bound systems, which can further split if the invariant mass of the considered system is large enough. This way, a set of colourless on-shell hadrons is obtained after enough string breaks.

3.2.2 Detector simulation

The emerging final state particles from the proton-proton collisions pass through the various layers of material before any detection by one of the CMS subsystems happens. Consequently, the interactions between these final-state particles and the detector material must be simulated too. This is handled by the `GEANT4` toolkit, which includes a detailed description of the CMS detector's geometry and its magnetic field [51]. It implements both the electromagnetic and hadronic interactions that the particles may experience, and accounts for the magnetic bending of the charged particle trajectories. Additionally, the `GEANT4` software simulates the electronic response from each subdetector to the impinging particles. This process is known as *digitisation* and marks the end of the simulation chain. From this point onwards, a collision event generated via the simulation chain or recorded at the LHC can be viewed as equivalent from a software point of view. Indeed, any collision recorded at the LHC is initially stored in this digitised data format, referred to as the *raw* event. Starting from these raw events, algorithms must be developed to reconstruct the physics objects, which are explained in the following section.

3.3 Reconstruction of physics objects

3.3.1 The Particle-Flow (PF) algorithm

The raw event data resulting from a proton-proton collision consists of the digitised hits registered across the detector subsystems. This data format is considered rather analysis-unfriendly, because it does not provide us with manageable information on the physics objects (muons, electrons, ...) and their properties ($\eta, \mathrm{p_T}$, ...). Therefore, a dedicated algorithm must be employed that translates an a priori seemingly disordered collection of hits to an exhaustive list of final-state particles and their corresponding properties. To handle this complex task, the Particle-Flow (PF) algorithm was designed [52]. The basic elements are the tracks of charged particles and the energy deposits of the ECAL and HCAL, known as *clusters*. In the two paragraphs that follow, the construction of tracks and clusters is discussed. Afterwards, the PF algorithm is explained.

An interactive tracking procedure consisting of multiple stages is used to reconstruct tracks [53]. In a first stage, known as the *initial seed generation*, reconstructed hits (usually three) from the innermost pixel layers are grouped into what is referred to as a *seeds*. These seeds provide a first estimate on the trajectory of a charged particle and are used as input to the next stage, the *trajectory building*. Here, the initial seeds are extrapolated outwards to the subsequent tracker layers via a *Kalman Filtering* method, creating a track [54]. At each layer, the track parameters are updated with the information of the hits compatible with the trajectory provided by the initial seed. Following the trajectory building, it is possible that a single seed results in multiple tracks, or a single track can be reconstructed from different seeds. The removal of these ambiguities happens in the third stage, conveniently called the *ambiguity resolution*, by comparing the shared number of hits between two trajectories normalised to the minimum of total hits between the two tracks. If too many hits are shared, and the fraction exceeds a set threshold, only the trajectory with the

most hits is kept. In the case that the trajectories would have an equal number of hits, the track with the lowest χ^2 is kept. In the final stage, the *final track fit*, the track is refitted considering all the compatible hits found from the Kalman Filtering procedure.

To group up calorimeter energy deposits, a two-step clustering is used, which runs separately in both the ECAL and the HCAL. Like iterative tracking, the first step in the clustering algorithm deals with the identification of seeds. In this context, a *cluster seed* refers to a calorimeter cell whose energy deposit exceeds a predetermined threshold and is larger than the deposits registered in its neighbouring cells. If any of these neighbours still has a significant energy deposit, meaning it exceeds some looser energy threshold than that of a seed, the cells and the seed are joined together to form a *topological cluster*. In some cases, multiple seeds may end up in one singular topological cluster. An additional algorithm takes care of this by sharing the energy deposits of the seeds within such a cluster. At the end of the clustering, a set of clusters in both the ECAL and the HCAL is achieved. These are more generally named *PF clusters*. For a more in-depth discussion on the clustering algorithm, see Ref. [52].

The PF algorithm can begin the reconstruction procedure with the fundamental elements as input. For optimal performance, it fully exploits the distinct features of the interacting particles. For example, electrons will leave a charged particle track in the inner tracker and an energy cluster in the ECAL. This means that one single electron, and in general any charged particle, will give rise to multiple fundamental elements. The PF algorithm has to account for this by somehow being able to *link* the PF elements across the various subdetectors. This happens via a *linking algorithm* based on spatial proximity, where each pair of PF elements is linked and the distance in the (η,ϕ) plane quantifies the links' quality. Firstly, going back to the example of electrons, a link between an inner track and an ECAL cluster must be made. This is done by extrapolating the track into the ECAL, up to a depth of a typical electromagnetic shower[1]. Next up, there are links between calorimeter clusters. Such a link is made if the cluster position in the ECAL falls within the envelope of the cluster in the HCAL. Lastly, links between an inner track and a track in the muon system can be established as well. While charged particle tracks in the inner tracker were generally discussed, muon tracks specifically were not mentioned. How they are formed and how muons are reconstructed will be explained in Section 3.3.2.

The PF elements linked together are called *blocks*, and proceed to the next part of the reconstruction, where association to particle candidates (muons, electrons, photons, charged and neutral hadrons) happens. After reconstructing one of these candidates, the corresponding PF elements of the block are removed. Starting with muons, the algorithm checks for links between inner and muon tracks. After removal, the algorithm checks for ECAL clusters not linked to any HCAL cluster. A photon candidate has been found if the ECAL cluster is not linked to an inner track. An electron candidate has been found when there is a link, and the momentum associated with both elements of the link is comparable. What remains to identify are hadrons. For a neutral hadron, a HCAL cluster with an ECAL cluster link but no tracker links is sought after. When such a track link is present, a comparison between momenta is made. A single charged hadron is identified if the tracker and cluster momenta are comparable. When the tracker momentum is significantly smaller, the additional momentum in the cluster is interpreted to come from a neutral hadron or photon on top of the charged hadron. On the other hand, if the track momenta is larger, a charged hadron cannot be assigned. Instead, the algorithm will search for additional muons with less stringent criteria.

[1] In the case of a link between a track and an HCAL cluster, the track has to be extrapolated into the HCAL up to a depth of a typical hadronic shower length.

After all the blocks have been processed, the result is a collection of PF objects. With these objects, jets can be reconstructed, as will be explained in Section 3.3.3, and variables such as the previously introduced missing transverse momentum can be calculated.

3.3.2 Muon reconstruction

Muons are central physics objects throughout the analysis conducted in this thesis. Not only will they appear in reconstructing our H+c final state in Chapter 4, but they will also play a crucial role in the reducible background estimations presented in Chapter 5. In general, precise reconstruction and identification of muons are thus needed. During previous discussions, it was mentioned that, for CMS, this is indeed the case. In this subsection, more details are provided. First, the three different approaches to muon reconstruction are explained. Afterwards, the identification criteria for muons used in this thesis and how isolation can be quantified are elaborated. Finally, the concept of scale factors is introduced to correct for differences in selection efficiency observed between simulation and data.

Standalone, global and tracker muons

The muons traversing the CMS detector will leave clear signatures in the inner tracker and the muon system. Depending on how these signatures are exploited to form the muon object, three muon types are defined [55]:

- **Standalone muon:** The first muon type is reconstructed with the iterative fitting approach explained earlier, but now using only the information of the muon system. Hits are first clustered, forming track segments, which are then used as seeds for extrapolation with the Kalman Filter technique. The muon track reconstructed in this way is referred to as a *standalone muon track*, and associated with it is the *standalone muon* object.

- **Global muon:** The second type of muon is obtained via an *outside-in* approach, where standalone muon tracks are matched with inner tracker tracks. When a match is found, meaning that an extrapolation of the tracks onto a common surface yields comparable parameters, hits from both are combined and refitted with a Kalman Filter. The result of that fit is the *global-muon track*, and the associated object is the *global muon*.

- **Tracker muon:** The final muon type complements the second, with an *inside-out* approach. Here, the inner tracker tracks with $\mathrm{p_T} > 0.5\ \mathrm{GeV}$ and total momentum $p > 2.5\ \mathrm{GeV}$ are extrapolated to the muon system. If a match is found between at least one muon track segment and the extrapolated track, the inner track is considered a *tracker muon track*. Its associated object is a *tracker muon*. At low momenta, below $10\ \mathrm{GeV}$, the tracker reconstruction is more efficient than its global counterpart. The main reason is that only a segment match is required here, instead of a more stringent track match to fulfil the global muon requirement.

In general, the muon reconstruction is very efficient. Practically all muons identified within the muon system are global or tracker muons, or even both. In the latter case, if an inner track is shared, they are merged into the same candidate.

Muon identification and isolation

Picking up on our discussion of the PF algorithm, muon reconstruction does not directly happen inside it. Instead, the muon tracks are provided as input, and the associated objects are often called *reco* muons. To arrive at muons identified by the PF algorithm, the reco muons must pass one of three additional requirements [56]. The first one is called the *isolated* selection. Here,

a muon is considered isolated if the sum of the $\mathrm{p_T}$ of the tracks and the E_T of the hits in the calorimeter is no more than 10% of the muon $\mathrm{p_T}$, evaluated in a cone centred on the muon. To the remaining reco muons, two other selection criteria are applied: the *pf-tight* and *pf-loose* selections. For the pf-tight selection, reco muons must satisfy a minimum number of muon track hits and muon segment quality cuts, and they must have a calorimeter deposit matching a simulation template. On the other hand, staying true to its name, the pf-loose selection is more relaxed. When discussing the charged hadron identification, it was mentioned that muons were searched for with looser criteria if no candidate was found. This is done using the pf-loose selection. The number of hits required is reduced, and the calorimeter template criterion is substituted with a match between the track and hits in the muon system. Finally, after these three selection criteria have been applied to the reco muons, a reduced collection called the PF muons is obtained.

Our analysis will require the PF muons to pass additional identification requirements. These are known as the *Tight* and *Loose* ID working points. A description is found below [55]:

- **Loose ID:** In addition to being identified as a PF muon, the *Loose ID* requires them to be a global or tracker muon as well. Any muon selected later has to satisfy this working point, making it the basis for muon identification. It is designed to identify prompt muons and muons from secondary hadron decays.

- **Tight ID:** While the *Loose ID* already provides us with high efficiency, misidentifications can always occur. For example, highly energetic hadrons can lead to showers that leak out of the HCAL and end up in the muon system, which is called a *punch-through*. This is problematic, as muons may be reconstructed from this signature. Furthermore, muons that decay in flight are not always interesting. To suppress the occurrence of these processes, the muon *Tight ID* was developed. It selects only global muons and enforces the following additional quality cuts: (1) the track associated with the global muon should have a reduced χ^2 smaller than 10; (2) at least one muon chamber hit is included in the global fit; (3) the global muon should also be a tracker muon; (4) the number of hits recorded in the inner tracker is non-zero; (5) each tracker layer should have at least five hits; and (6) the transverse d_{xy} and longitudinal d_z impact parameter of the tracker track have to satisfy $|d_{xy}| < 2$ mm and $|d_z| < 5$ mm with respect to the primary vertex. The impact parameter is defined as the point of closest approach to the primary vertex of the event. To end, it is worth noting that by definition, any muon satisfying the Tight ID will also pass the Loose ID automatically.

The muons arising from prompt decays tend to be isolated from other signatures inside a collision event. To select these isolated prompt muons, one needs to quantify their relative isolation in the first place. The PF algorithm provides us with such information by defining the relative isolation $\mathcal{I}^\mu_\mathrm{rel}$ of a muon with transverse momentum p^μ_T as [52]:

$$\mathcal{I}^\mu_\mathrm{rel} = \frac{\sum \mathrm{p_T}^{\text{charged hadrons}} + max(0, \sum \mathrm{p_T}^{\text{neutral hadrons}} + \sum \mathrm{p_T}^{\text{photons}} - 0.5 \sum \mathrm{p_T}^{\text{PU}})}{\mathrm{p}^\mu_\mathrm{T}}. \tag{3.1}$$

The sums in Equation (3.1) run over all PF objects residing within a cone $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ < 0.3 around the muon. The last term in the numerator reflects the energy deposits arising from neutral (hadrons, photons) particles in pileup interactions and runs over all charged particles in the cone. The motivation behind this term is as follows. During the reconstruction of the PF charged hadrons, the pileup contribution from charged hadrons can already be estimated via a procedure known as *charged hadron subtraction* (CHS). In this approach, the track of the charged hadrons is identified, and the corresponding vertex is determined. If this vertex is not the primary vertex of the event collision, the charged hadron in question originates from a pileup interaction.

To arrive at the contribution from neutral particles, the contribution determined with CHS is rescaled by a factor of 0.5, which is the average ratio of neutral to charged particles produced in a proton-proton collision, which is determined from simulation.

Muon scale factors

A selection of muons relevant to an analysis typically requires them to pass criteria on both identification and isolation. However, the efficiency of selecting those muons may differ in simulation from what is observed in data. On the one hand, this can be related to the mismodelling of any step in the simulation of the proton-proton collision. On the other hand, this may be a reflection of the simulation not being able to fully capture the complex environment experienced in the data. Either way, any discrepancy in efficiency must be accounted for. This is done by applying *scale factors* to the simulation event samples after selection. Without such a correction, simulation is inherently biased because it would not accurately reflect the performance seen in data. A scale factor (SF) is usually measured in bins of pseudorapidity η and transverse momentum $\mathrm{p_T}$. It is defined as the ratio of the efficiency in data ϵ^{data} to the efficiency in simulation ϵ^{MC}:

$$\text{SF} = \frac{\epsilon^{\text{data}}}{\epsilon^{\text{MC}}}. \tag{3.2}$$

How scale factors are measured typically depends on the nature of the physical object itself. For muons, the common approach is the *tag-and-probe* method using a dimuon resonance (such as the Z boson or the J/ψ) [57]. In this technique, events containing two oppositely charged muons within a tight mass window around the resonance are selected. This mass requirement effectively means that only dimuon events originating from the resonance are considered. One muon, referred to as the *tag* muon, is required to pass strict and well-understood identification and isolation criteria to ensure it is a prompt muon. The other, the *probe* muon, is used to evaluate the efficiency of a specific criterion (e.g., the Tight ID) and is initially required to satisfy only a loose set of selection criteria. The efficiency is then defined as the fraction of probe muons that pass the selection criterion under study, relative to all probe muons. This procedure is done for data and simulation, after which Equation (3.2) is inferred to arrive at scale factors.

To compare data and MC, any analysis must have scale factors applied to be fully accurate. Currently, no scale factors have been derived for the muon selection criteria used in this thesis, and derivations are beyond the scope of this work. Consequently, no scale factors will be applied to muons throughout this thesis. However, due to the excellent muon reconstruction and identification performance of CMS, such scale factors typically correspond to only percent-level corrections. These corrections are not expected to significantly impact our results, since we are not conducting precision measurements. To conclude, it is worth noting that other corrections on muons, such as the momentum scale, also exist. These are minor compared to scale factors and will not be considered in this thesis either. For other objects, such as jets, such corrections will be more significant and explained in the following sections.

3.3.3 Jet reconstruction

The final state of the H+c process consists of more than muons. The signature from the c quark must also be detected, which is not as straightforward as it seems. Section 3.1 explained that partons cannot be detected individually. Indeed, the collection of colourless stable particles arising from the hadronisation of an emerging parton is more or less concentrated in a cone-like shape, called a *jet*. For any analysis involving final-state partons, these objects are of key importance. Especially for this analysis, we are only sensitive to our charm signal if the jet can be correctly reconstructed and identified. This subsection deals with the reconstruction of jets through cluster-

ing, followed by corrections to calibrate them. A discussion of how various jet types are identified and distinguished based on their flavour is presented in the next section.

Jet clustering

The reconstruction of a jet amounts to the clustering of the particles identified by the PF algorithm. From the various jet clustering algorithms that exist, the CMS collaboration employs the so-called *anti-k_t jet clustering algorithm* [58]. To start the clustering procedure, a distance d_{ij} between two cluster candidates i, j, and a distance d_{iB} between a cluster candidate i and the beam B are defined:

$$\begin{aligned} d_{ij} &= \min(k_{ti}^{2p}, k_{ti}^{2p}) \frac{\Delta^2}{R^2} \quad \text{with} \quad \Delta^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2, \\ d_{iB} &= k_{ti}^{2p}. \end{aligned} \tag{3.3}$$

In the above definitions, k_{ti}, ϕ_i, and y_i refer to the transverse momentum, azimuthal angle, and rapidity (Equation (2.2)) of candidate i respectively. The quantity p varies per algorithm, and for the anti-k_t case, $p = -1$. Finally, the quantity R is called the radius parameter, which reflects the angular size of a jet and typically has the value $R = 0.4$.

The clustering itself proceeds as follows. For each event, the PF clustering candidates are grouped into pairs for which the distances d_{ij} and d_{iB} are computed. If the minimum of these distances corresponds to d_{ij}, the cluster candidates are merged into a new object called the *pseudo-jet*. The clustering algorithm iteratively continues to measure the distances between pairs, now including pseudo-jets as possible cluster candidates as well. If at some point the minimum corresponds to the beam distance d_{iB}, the corresponding candidate i is considered to be a *reconstructed jet* and is removed from the candidate list. This way, a collection of reconstructed jets is obtained after the candidate list has been fully iterated through. In simulation, the clustering of all stable particles that have been generated from an event before any digitisation took place can also be considered. These jets are called *particle-level* jets of *GenJets* in short, and they have the advantage of not being affected by a non-linear response from the detector. This property is exploited to derive corrections for the reconstructed jet energy.

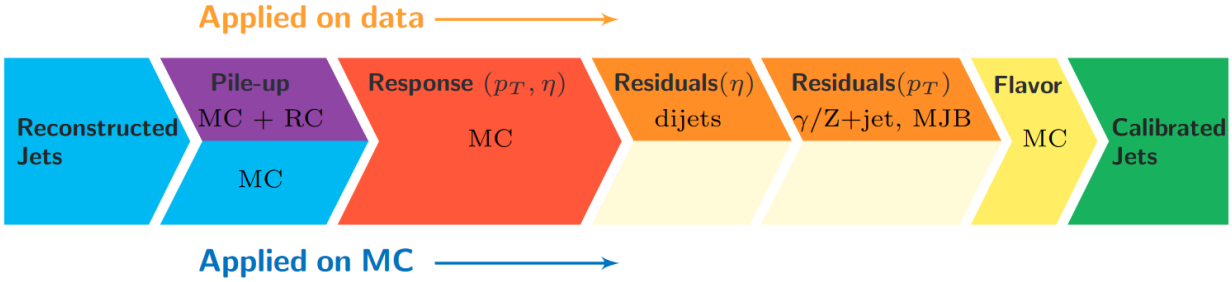

Figure 3.2: *The Jet Energy Corrections (JEC) flow scheme. The top flow represents the steps applied to data, while the bottom flow depicts the steps applied to MC. Image taken from [59].*

Jet Energy Scale (JES) corrections

In the ideal case, the energy of the reconstructed jet corresponds to the energy of the parton that produced it. However, circumstances are never ideal. Additional particles arising from pileup interactions may have been clustered into the jet, leading to an overestimation of the energy. On the other hand, particles with low $\mathrm{p_T}$ can be lost as they are bent too much in the magnetic field,

causing an underestimation. Furthermore, the electronic response of the detector is not linear, meaning that the determined energy of particles is not completely accurate. To mitigate against these effects and calibrate the energy of reconstructed jets in both MC and data, Jet Energy Scale (JES) corrections must be applied. This is done via a factorised approach, visualised in Figure 3.2, where each step considers correcting for a different effect. The output of the previous stage is used as input for the next. In what follows, each step is briefly elaborated on:

- **Pile-up correction (L1):** The first jet energy correction applied to jets deals with pileup. While pileup was introduced as additional proton-proton collisions occurring in the same bunch crossing, it can also manifest itself in another way. The electronic signals from the detector have a finite decay time, meaning that signals from the previous or the next bunch crossing can affect the present bunch crossings. This is known as *out-of-time pileup*, and the pileup first introduced is called *in-time pileup*. The effects of out-of-time pileup are estimated by changing the integration time of signals in the calorimeter and the spacing between subsequent bunch crossings. In-time pileup effects from charged hadrons are handled with the previously mentioned procedure of charged hadron subtraction (CHS). Any contribution arising from neutral particles and leftover out-of-time pileup is subtracted using the concept of *jet area* A, which reflects how sensitive a jet is to any form of contamination (e.g. from pileup) [60]. Lastly, an additional residual correction is applied to account for any difference in offset between the data and MC. This procedure is known as the *hybrid jet area method*. More details can be found in Ref. [59]. The result of this step is the pileup corrected transverse momentum p_{T}^{L1}.

- **Response correction (L2L3):** In the next step of the JES flow, a correction for the non-linear response of the calorimeters in the η-plane is applied to the pileup corrected jets. The method for this correction mainly relies on simulation and consists of matching reconstructed jets with particle-level jets. Such a unique match is found if a particle-level jet is within half of the jet radius parameter R of the closest reconstructed jet. The jet response $\mathcal{R}$ is then defined as the ratio between the means of transverse momenta of the matched reconstructed jet and particle-level jet. Using multi-jet event samples, calibration factors binned in $\mathrm{p_T}$ and η are defined as the inverse of the average response $\langle R(\mathrm{p_T}, \eta) \rangle$ in that bin, which are used to correct the momentum. More information can be found in Ref. [61].

- **Residual correction (L2L3Residual):** After applying pileup and response corrections, discrepancies between the energy scale in MC and data are still observed. Therefore, additional residual corrections are applied, but solely to the data. In total, there are two of these residuals. The first (L2Residual) focuses on η-dependent corrections by studying the transverse momentum imbalance of jets in dijet events. The second and final residual correction (L3Residual) handles any remaining $\mathrm{p_T}$ dependent effect in the data. More information can be found in Ref. [61].

- **Flavour correction:** The last correction in the JES flow accounts for any flavour dependencies on the jet energy. This effect is mostly subleading compared to the other corrections and will not be used in this thesis.

To summarise, after applying the corrections to account for pileup, non-linear response of the detector, and any remaining residual artefacts, *calibrated jets* are obtained in both data and MC that can be used for further analysis.

Jet Energy Resolution (JER) corrections

Although corrections are applied to the jet energy scale, discrepancies between data and simulation are still observed for the Jet Energy Resolution (JER). In terms of jet responses, the energy

resolution corresponds to the width of the response, while the jet energy scale can be interpreted as the mean of the distribution. To be precise, the definition of the resolution differs slightly for MC and data. In simulation, the jet response is computed as the ratio between the means of transverse momenta of the matched reconstructed jet and particle-level jet. The core of the jet response distribution can typically be described by a Gaussian distribution, with the JER defined as its width. However, the data does not provide access to particle-level jets. Instead, the jet response must be inferred indirectly, using transverse momentum imbalance techniques described before. While the resolution in both cases is defined as the width of the corresponding Gaussian, differences compared to MC are expected. Indeed, the energy resolution of calibrated jets is measured to be worse in data compared to MC. To resolve this discrepancy, jet momenta in simulation are artificially smeared with the use of scale factors to broaden the response distribution and match it to what is observed in the data [59].

JEC/JER uncertainties

To end this section, the focus is shifted to the uncertainties related to the JES and JER. Each energy correction or smearing factor has a corresponding uncertainty. The origin of these uncertainties is systematic, rather than statistical. In general, the impact of uncertainties is evaluated by considering "up" and "down" shifts to the associated parameters, which represent variations of ± 1 standard deviation. For the JES, an upward shift amounts to rescaling the $\mathrm{p_T}$ of the jet as $\mathrm{p_T} \rightarrow \mathrm{p_T} \times (1 + sys.unc.)$, while a down shift corresponds to a rescaling of $\mathrm{p_T} \rightarrow \mathrm{p_T} \times (1 - sys.unc.)$. Analogously, for the JER, the up and down variations correspond to additional or reduced smearing of the jet momentum. Each of these variations must then be propagated to estimate its corresponding effect on the final results. These uncertainties are beyond the scope of the thesis and can be investigated in future studies.

3.4 Jet flavour tagging

3.4.1 Jet flavour definition

The properties of jets are expected to differ based on the flavour of the originating parton. In some sense, we are blind to this information during the reconstruction, as we only care about the clustering of particles at this stage. Nevertheless, the identification and distinction of jets originating from different partons is crucial for many analyses. This is the aim of the *jet flavour taggers*: algorithms attempting to identify a jet's flavour based on properties of particles residing inside it. To construct such a tagger algorithm in the first place, a clear definition must be given of what it means for a jet to have a certain flavour in simulation, because this is where these algorithms are initially trained before being calibrated by data. In this subsection, the jet flavour definition is further explained. The commonly adopted approach looks at (generated) hadrons, using the *ghost association* technique to define jet flavours [60]. In ghost association, the generator-level b and c hadrons present in the event are included, and the entire collection of particles is reclustered. The momenta of these generated hadrons are set to negligible values, creating so-called *ghost hadrons*, which ensures that the jet kinematics is unaffected and only directional information is used. Once the reclustering is done, the newly reconstructed jets are matched with particle-level jets. For those jets where a match is found, the following jet flavours are defined:

- b **jet:** If the reconstructed jet carries at least one (ghost) b hadron, it is called a b jet.
- c **jet:** When the reconstructed jet in question is not a b jet but does carry at least one (ghost) c hadron, it is called a c jet.

- **light ($udsg$) jet:** The reconstructed jet is neither a b nor a c jet, implying that the emerging parton was either a gluon (g) or a light quark (u, d, or s). It is called a light ($udsg$) jet in that case.

It is possible that no match was found between a reconstructed and particle-level jet. Such reconstructed jets are called *pileup jets*, regardless of the presence of any (ghost) hadrons. The b and c jets are also called heavy-flavour jets, for which the identification is the subject of the next section.

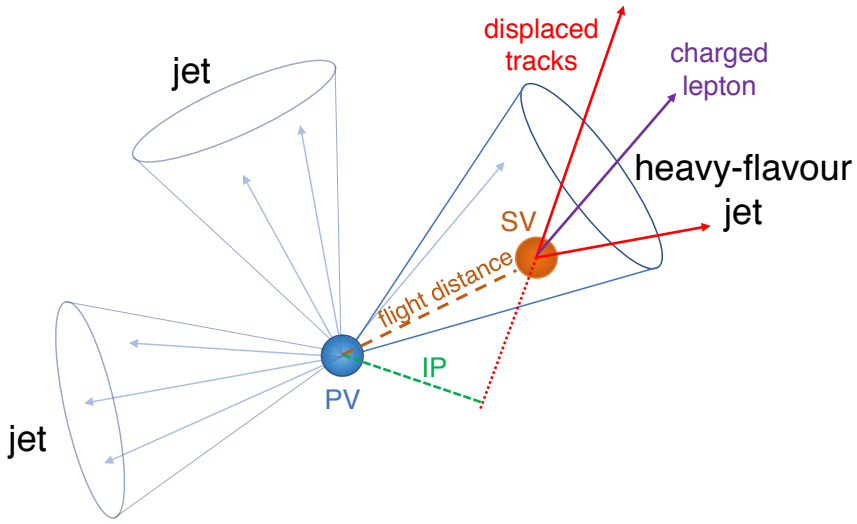

Figure 3.3: *A schematic overview of a heavy flavour jet. The lifetime of the residing b and c hadrons results in displaced tracks with respect to the primary vertex (PV). From them, a secondary vertex (SV) can be reconstructed in which charged leptons may be present. To characterise the SV, the flight distance is defined as the distance between the PV and the SV, and the impact parameter (IP) is defined as the distance between the PV and the displaced tracks at their point of closest approach [62].*

3.4.2 Identification of heavy-flavour jets

While the flavour definitions from the previous section are specific to the tagger utilised for this thesis and may seem somewhat artificial, the general idea behind heavy-flavour jet identification is quite straightforward: any accessible information about the flavour of the parton that initiated the jet is encoded in the hadrons produced through parton showering and hadronisation. To identify jets originating from b and c quarks, we must rely on identifying heavy-flavour jets, which contain b and/or c hadrons, and distinguish them from jets that do not. The tagger uses discriminating variables constructed from the properties of these hadrons to tackle this non-trivial task.

One of the main properties thoroughly exploited is the *lifetime*. In their respective rest frames, b hadrons have a lifetime of around 1.5 ps, while c hadrons live about 1 ps or less. Depending on their momentum, this allows them to travel a significant distance away from the primary vertex (PV) before decaying, ranging from a few millimetres up to about 1 cm. As a result, the tracks originating from these hadrons are displaced, and a secondary vertex (SV) can be reconstructed from them. This is illustrated in Figure 3.3. The distance between the primary and secondary vertices is called the flight distance. It is expected to be larger for b jets than for c jets and negligible for light-flavour jets. An SV's presence indicates a heavy-flavour jet, and additional discriminating variables can be derived from it. One such variable is the mass of the SV, which is closely

related to the mass of the hadron and is defined as $\sqrt{M_{\mathrm{SV}}^2 + p^2 \sin(\theta)} + p \sin(\theta)$. In this definition, M_{SV} denotes the invariant mass of the tracks associated with the SV, p their momentum, and θ the angle between the vector defined by the flight distance and the secondary vertex momentum. Another important variable is the flight distance significance in 2D, calculated as the flight distance divided by its uncertainty in the transverse plane. The distributions of these variables differ depending on the jet flavour, as explicitly shown in Figure 3.4.

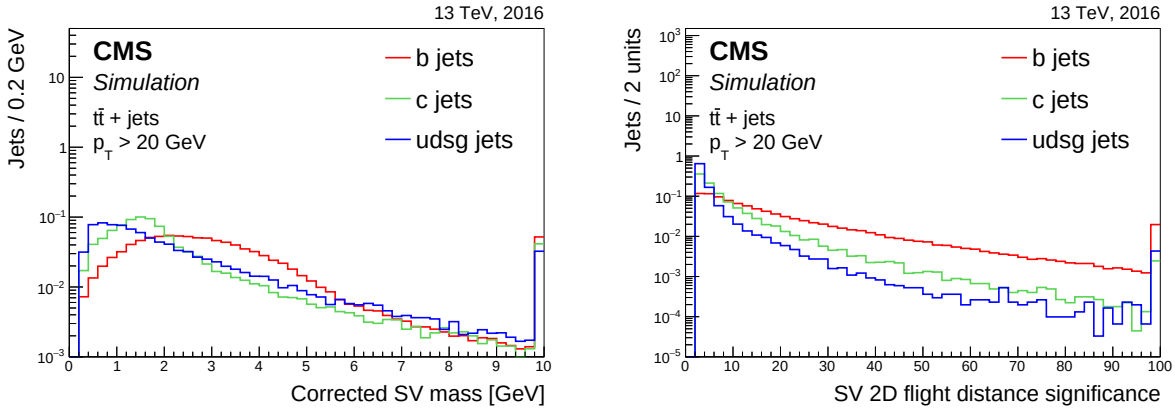

Figure 3.4: *(left) The distribution of the SV mass and (right) the distribution of the SV 2D flight distance significance for b jets (red), c jets (green), and light jets (blue) with* $\mathrm{p_T} > 20$ *GeV derived from simulated* $t\bar{t}$ *events [62].*

Regarding the SV mass distribution shown on the left in Figure 3.4, it is observed that the peak for b hadrons is shifted further to the right compared to c hadrons. Here, the mass difference between the hadrons comes into play. The peak for light jets appears even lower, as no SV is expected. Regarding the flight distance significance shown on the right, b jets dominate due to their increased lifetime. Together, these variables provide additional discrimination power to identify heavy-flavour jets from light jets. However, not every heavy-flavour jet will have an accompanying SV. This can occur if the associated heavy-flavour hadron decay occurs too close to the PV or if too few displaced tracks for reconstruction. Therefore, additional variables independent of the SV are considered, such as the *presence of soft charged leptons*. In the decay of a b (c) hadron, the probability of an electron or muon appearing is roughly 20 % (10 %). Their presence can thus further aid in distinguishing heavy-flavour jets from light jets, which are not expected to contain such leptons [62].

Despite the available tools, distinguishing a heavy-flavour c jet from b jets and other light jets remains challenging. This is the task of *c-taggers*. The main difficulty arises because the distributions of discriminating variables for c jets often lack prominent features, being in between those of b jets and light jets. As a result, misidentifications are more likely, and the efficiency of these taggers is generally lower compared to *b-taggers*, which aim to distinguish b jets from other flavours. Nevertheless, c-tagging techniques are continuously being improved. One of these taggers, DeepJet, is introduced in the next part.

3.4.3 The DeepJet tagger

The heavy-flavour tagger used in this thesis is the DeepJet algorithm [63] developed by the CMS collaboration. DeepJet is based on the technology of deep neural networks and takes combinations of over 600 event variables as input, including features from the primary vertex (PV), sec-

ondary vertices (SV), particle-flow (PF) candidates, event kinematics, and track information. As output, the tagger provides a classifier variable for each category of jet flavour, which can be interpreted as the probability that the jet belongs to that flavour: $P(c)$, $P(uds)$, $P(g)$, $P(b)$, $P(b_{lept})$, and $P(bb)$. Specifically, $P(\text{c})$ refers to the c jet category, light jets are split into the $P(uds)$ and $P(g)$ categories, and the b jet probability is subdivided into the $P(b), P(b_{lept})$, and $P(bb)$ classes. Here, the $P(\text{b})$ refers to b jets where the b hadrons decay hadronically, $P(b_{lept})$ to b jets with b hadrons decaying leptonically, and $P(bb)$ for jets with two b hadrons. To distinguish between the different categories, *discriminators* are defined, which combine the output of the various classifiers into one variable. For b-tagging, the discriminator is defined as $P(b) + P(b_{lept}) + P(bb)$. The main motivation behind this simple definition is that the properties of b jets are quite distinct from those of c jets and light jets. For c jets, the situation is more nuanced. As mentioned at the end of the previous subsection, the distribution of properties for c jets is typically situated in the middle of the b and light jets. This is why, for c-tagging, two discriminators are defined. The first is the CvsB discriminator, designed to distinguish c jets from b jets, and is defined as:

$$\text{CvsB} = \frac{P(c)}{P(c) + P(b) + P(b_{lept}) + P(bb)} \in [0, 1]. \tag{3.4}$$

The second discriminator, denoted as CvsL, aims to distinguish c jets from light jets and takes the form:

$$\text{CvsL} = \frac{P(c)}{P(c) + P(uds) + P(g)} \in [0, 1]. \tag{3.5}$$

A reconstructed jet that simultaneously has values close to unity for both discriminators is very likely to be a c jet, as this implies that the probabilities of the jet being any other flavour are small.

3.4.4 Calibration of jet taggers

To end this chapter, the calibration of heavy-flavour taggers is briefly discussed. Because these algorithms are trained on simulation, it is rather unlikely that they would perfectly match observations from real collision data. For example, mismodelling related to the detector response or event generation can lead to discrepancies between simulated and real events, which can be flavour-dependent. As previously discussed during the muon reconstruction in Section 3.3.2, these effects can be accounted for with scale factors (SFs). The calculation of such scale factors typically involves evaluating the tagging efficiency at a specific working point in both data and simulation (see Equation (3.2)). However, this analysis does not require jets to pass any working point in terms of c-tagging discriminators information. Instead, their distribution is of interest. This approach requires a different method for deriving scale factors, one that calibrates an entire shape. Furthermore, as c-tagging uses the CvsL and CvsB discriminators, both of their shapes must be calibrated simultaneously.

We will use the results of the novel method presented in Ref. [64], which uses an iterative approach. The idea is to define three event samples enriched in a particular flavour: a "c-enriched" event sample with $W+c$ events, a "b-enriched" event sample from $t\bar{t}$ events, and a "light-enriched" event sample that focuses on Z + jet processes. For each event sample, the number of jets belonging to each flavour is determined via simulation, resulting in flavour fractions per sample. These flavour fractions are varied iteratively until the differences between the data and simulation are minimised. Specifically, in each iteration, event samples are ranked by their flavour purity and SFs are derived by calculating the ratio of enriched jets in data to those in simulation. These SFs are applied in the next iteration until the ratio stabilises. After the final iteration, flavour-dependent SFs that vary across CvsL and CvsB values are obtained. These scale factors will be applied to all our selected reconstructed jets in simulation.

4 THE H+JET SELECTION FRAMEWORK

This chapter marks the beginning of the research conducted in this thesis, which has as its ultimate goal to extract an expected limit on the charm-Higgs Yukawa coupling via processes where a charm quark is produced in association with a Higgs boson. To achieve this, we must first carefully analyse the final state, where only the jet and muons are experimentally observable, and develop methods to reconstruct the process of interest. Next, we must carefully characterise and estimate the irreducible and reducible backgrounds. This general strategy is further laid out in Section 4.1. This chapter will focus on two fronts: (1) the reconstruction techniques for H+jet events via a dedicated framework and (2) the estimation and characterisation of irreducible backgrounds and H+jet signal events with simulated event samples. First, we describe the general data and simulation inputs in Section 4.2. The following two sections tackle the practical aspects of the reconstruction. Specifically, Section 4.3 handles the preliminary event and object selection before any reconstruction, while Section 4.4 deals with the reconstruction of the Higgs boson and selecting an accompanying jet. In the last section of this Chapter, Section 4.5, the irreducible backgrounds and the signal process are characterised by studying their event yields, kinematic and flavour discriminator distributions. In doing so, we lay the foundation for the studies conducted in later chapters.

4.1 Overall strategy

As mentioned in the preface, the overarching goal of this thesis is to probe the charm-Higgs Yukawa coupling y_c via the H+c process. To contribute towards this objective, the analysis focuses on key components necessary for such a measurement. In particular, accurately estimating backgrounds will be key. To achieve this, the study is divided into three parts. This chapter addresses the first and most foundational step: reconstructing and characterising the H+jet final state. The Higgs boson decays via the $H \to ZZ^* \to 4\mu$ decay channel, and the quark initiates a jet. Considering this topology, we will develop a general-purpose reconstruction framework to identify such events. We do not yet confront our techniques with data, but design our techniques as if we would. A key aspect of the reconstruction is that the jet accompanying the Higgs boson candidate will be selected based only on kinematic information, without any flavour tagging information. This design choice has two motivations. First, imposing flavour-based cuts could significantly reduce our statistics and sensitivity to the signal. Second, by retaining flavour information, we enable further studies to measure the flavour composition (i.e., the relative fractions of c, b, light quark or gluons in the spectrum) in other processes, such as Z+jet. These processes are more abundant than H+jet, and it could be explored how the similarity in flavour structure between these processes could help increase the significance of the y_c measurement. For this reason, we refer to the algorithms and software tools used for the reconstruction and characterisation of H+jet events as the H+jet framework. In addition to the signal, we also characterise the irreducible backgrounds, which are expected to have an identical leptonic final state, with these reconstruction techniques. By comparing their spectra, we illustrate the performance and learn how the signal differs from the irreducible backgrounds.

Apart from the irreducible backgrounds, there are also reducible backgrounds, which mimic our final state signature but are not expected to do so. The reducible background includes events where:

- A jet is misreconstructed as a prompt muon. Primarily dealing with light-flavour jets, where particles produced within the jet, such as decay products of in-flight mesons (e.g., pions or kaons) or other jet constituents, can carry a significant fraction of the jets transverse momentum and appear isolated. Note that, in general, any particle could be misidentified as a muon;

- A non-prompt muon within a heavy-flavour jet is identified as being prompt. As mentioned in Section 3.4, heavy-flavour jets may contain secondary charged leptons originating from the decay of heavy-flavour hadrons.

Although the individual probability of a misidentification is low, the large number of collision events occurring at the CMS detector may result in frequent lepton misidentification, leading to a significant amount of reducible background. Crucially, reducible backgrounds are not estimated using simulation. Practically speaking, generating a sufficient number of MC events for these rare misidentification scenarios is computationally heavy. More fundamentally, the complexity of the underlying physics of these fake signatures is difficult to simulate accurately. These effects include QCD effects at low energies where perturbativity may break down. In particular, a jet follows from a sequence of parton showers and hadronisation. While these parton showers can be described perturbatively, as explained in Section 3.2, many particle emissions occur at soft energy scales, making them harder to model. Additionally, hadronisation entirely relies upon phenomenological models. If uncertainties are present in the modelling, we also reduce our ability to model a misidentification. Moreover, simulating how these decay products in a jet pass the arbitrary detector criteria, such that they are reconstructed as a muon, introduces more complexity. Especially since detector noise and resolution could seriously affect this. Altogether, these challenges mean that relying on simulation alone to model reducible backgrounds will introduce significant and poorly constrained systematic uncertainties into our estimation. Instead, we will adopt a data-driven approach to estimate these backgrounds [65]. The methodology used for this data-driven estimation differs significantly from the H+jet reconstruction that we will present below, and will be presented in Chapter 5. There, we will apply the reconstruction techniques from this chapter to the data, compare the resulting distributions with those from simulations that account for the irreducible and reducible backgrounds, as well as the signal.

Finally, once irreducible and reducible backgrounds are accounted for, we extract information about the y_c coupling in Chapter 6 by deriving an expected upper limit on the H+c production signal strength $\mu_{\mathrm{H+c}}$.

4.2 Input datasets

4.2.1 Signal $\&$ irreducible backgrounds

After a proton-proton collision is reconstructed via the procedures described in Section 3.3, we are left with an event format containing the identified physics objects and relevant information such as the triggers fired by each event. In simulation, the generator-level particles, their properties,

and the weights[1]. This format is referred to as the *NanoAOD*[2] format, and it serves as the starting point for any reconstruction. We reconstruct the physics processes of interest from these input files by selecting and combining the relevant final state particles. We follow a two-tiered strategy to achieve this efficiently, as shown in Figure 4.1.

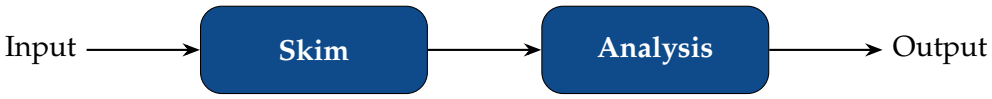

Figure 4.1: *Schematic overview of the two-tiered workflow employed in the reconstruction.*

First, any dataset passes through a technical preselection step, which we call the *skim* step. This serves as a preliminary screening of the collision events to select only those with signatures similar to the final state from the process of interest. Especially for data, *skimming* is essential. Unlike MC samples dedicated to the simulation of one specific process, we don't know the underlying processes that lead to the observed final state particles in data. By skimming, we retain only potentially interesting events, thereby significantly reducing the size of the dataset considered. The computational power needed for the next step is significantly reduced. This next step, called the *analysis*, takes any event that survived the skim as input. For these events, we try to reconstruct the underlying physics process by combining the information of the final state particles. The result is a reduced dataset focused on object candidates relevant to the H+jet process, ready for further analysis. However, we first need input event samples. This subsection introduces simulated samples used to model the H+jet signal and irreducible backgrounds. The following subsection will describe the data samples used to estimate reducible backgrounds and compare the data and simulation.

An overview of all the processes used in the signal and irreducible background processes is given in Table 4.1. Apart from the process name, we also denote the label provided to the process, which will be used later in the graphical figures, as well as the number of generated events and the production cross section in fb.

We begin with the H+jet event sample. This is a privately produced sample by the local research group and contains contributions to the Higgs production sensitive to the charm-Higgs Yukawa coupling and the bottom-Yukawa coupling. Consequently, this sample includes our H+c signal. The corresponding Feynman diagrams for this process are illustrated in the leftmost and middle panels of Figure 1.4. Later sections will explain the parton-based labelling at the object reconstruction stages.

Next, we consider the irreducible Higgs boson backgrounds: tqH, ttH, VH, VBF H, and ggH. These samples and any other sample referenced from this point onwards are centrally produced by the CMS collaboration. Representative Feynman diagrams of these processes are illustrated in Table 1.1. Among these, the ggH process represents the bottom- and charm-Yukawa-insensitive contribution to the Higgs production processes. A representative Feynman diagram for this process is also shown in the rightmost panel of Figure 1.4.

The final category consists of the irreducible non-Higgs backgrounds, for which representative

[1]These weights account for how the generator samples the phase space to produce predictions for differential cross sections. Due to the technical implementation, this means that sometimes events receive, e.g., a negative weight to ensure the spectra are correct.

[2]AOD refers to 'Analysis Object Data'. The 'Nano' prefix indicated that this format is minimal in storage, but still sufficient for any analysis.

Feynman diagrams are found in Table 1.2. Within this category, we first include the processes labelled as $\mathrm{qq} \to$ZZ, Zγ^*. As the label suggests, this label is given to the process in which ZZ is produced through quarks ($\mathrm{q\bar{q}} \to \mathrm{ZZ} \to 4\ell$), single resonant Z production with initial state radiation with internal conversion ($\mathrm{q\bar{q}} \to \mathrm{Z}\gamma^* \to 4\ell$). It also includes single resonant four-lepton production ($\mathrm{q\bar{q}} \to \mathrm{Z} \to 4\ell$). A caveat is needed here. When these backgrounds were first introduced in Section 1.3, we explicitly required the muon decay channels. However, we must make it clear that the event samples of these processes are inclusive, concerning the leptonic decay channels of the Z boson. That is, they include all charged leptons $\ell \in \{e, \mu, \tau\}$, which is why we generally denote the final state 4ℓ. This implies that, for example, one Z boson decays into an opposite-sign muon pair while the other may decay into an electron or tau pair. If, for any reason, these leptons are misidentified as prompt muons, such events are technically reducible backgrounds according to our definitions. However, we expect the contributions of these misidentified decays to be negligible compared to the prompt muon decays. Thus, this inclusive sample will suffice for our studies. This category's second group of processes is labelled as gg→ZZ, Zγ^*, referring to any process in which the Z bosons are produced via gluons. Implicitly, these processes also include off-shell photons γ^* being produced along the Z bosons, hence the label name. These samples also include all leptonic decay channels. In principle, only the muon decay channels will contribute sufficiently, but for completeness, they were included.

Category	Label	Process	σ [fb]	Generated events N_{MC}
Signal	H+c/b/l	H+jet	0.317	1 056 587
Irred. H. bkg.	tqH	tqH	0.086	969 000
	ttH	ttH	0.364	490 043
	ggH	ggH	13.3	940 000
	VBF	VBF H	1.04	477 000
	VH	$\mathrm{W^-H}$	0.146	193 853
		$\mathrm{W^+H}$	0.231	298 674
		ZH	0.532	544 550
Irred. non-H. bkg.	$\mathrm{qq} \to$ZZ, Zγ^*	$\mathrm{q\bar{q}} \to$ ZZ, Zγ^*, Z $\to 4\ell$	1256	98 488 000
	gg→ZZ, Zγ^*	gg $\to$ ZZ $\to 4e$	1.59	974 000
		gg $\to$ ZZ $\to 4\mu$		845 232
		gg $\to$ ZZ $\to 4\tau$		493 998
		gg $\to$ ZZ $\to 2e2\mu$	3.19	500 000
		gg $\to$ ZZ $\to 2e2\tau$		500 000
		gg $\to$ ZZ $\to 2\mu2\tau$		496 000

Table 4.1: *List of simulated event samples used in the nominal H+jet selection framework. The samples are grouped by their background category: Irreducible Higgs background (Irred. H. bkg.) and the irreducible non-Higgs background (Irred. non-H. bkg.). The signal category contains the H+c process and the H+b and H+l contributions, which are considered irreducible Higgs boson backgrounds. Each process is assigned a label, corresponding to those used later in figures, with some processes sharing a label for simplicity. The cross sections (σ) are shown in femtobarns [fb], and event counts indicate the number of generated events N_{MC}.*

4.2.2 Recorded data

With the simulation event samples characterised, we focus on the observed data event samples, used in the following chapters. In this thesis, we will use data collected by the CMS detector in the 2018 data-taking period of LHC Run-2. These collected datasets, called primary datasets, are

organised based on triggers that fired the collision events. Since we are interested in muon final states, we consider the datasets corresponding to triggers that only select muons. In particular, we will make use of two datasets: one linked to a single muon trigger, conveniently named the *SingleMuon* dataset, and one associated with a double muon trigger, known as the *DoubleMuon* dataset. More details on these triggers will be presented in the next section. Each dataset is further divided into data-taking periods, labelled Run A, B, C, or D, corresponding to different year periods. The number of recorded events for the primary dataset in each run is listed in Table 4.2.

Primary Dataset	Run	Recorded events
SingleMuon	A	241 596 817
	B	119 918 017
	C	110 032 072
	D	513 884 680
DoubleMuon	A	75 499 908
	B	35 057 758
	C	34 565 069
	D	168 620 231

Table 4.2: *List of Run-2 recorded data used in this analysis, grouped by primary dataset and run.*

Regarding the number of events, three considerations must be made:

1. **Dataset Overlap:** The shown amount may be misleading because of the overlap between the two datasets. Indeed, an event satisfying both triggers, a common occurrence, appears in both associated datasets. Unique events that fired at least one of the triggers must be stored to avoid double-counting.

2. **Blinding:** The decision was made to *blind* the data. This means we deliberately exclude events in regions where we expect the signal to appear. This is done to avoid bias. The specific details of the blinding strategy will be described in Section 4.4.

3. **Data quality:** Not all recorded events are considered suitable for analysis. During data-taking, some subdetectors may have experienced technical difficulties or downtime, in the worst case, resulting in incomplete information provided. To ensure high-quality inputs, only recorded data from subdetectors that were operational and all relevant information was available was used. The CMS collaboration provides this information. It is technically referred to as the *luminosity mask*.

4.3 Skim

4.3.1 Triggers

The purpose of the skim step is to apply a preselection that retains only events containing the types of final state particles expected from the signal process. The preselections consist of two main categories. The first deals with triggers. We ensure that events have been flagged by the appropriate HLT paths, with the recorded data event samples meeting the CMS data-quality criteria. The second category involves preliminary selections of the physics objects themselves. The triggering is handled in this subsection, while the next subsection deals with the object selections.

The 2018 dataset used is divided into two subsets based on the HLT paths during data taking:

the SingleMuon and DoubleMuon datasets. Each dataset contains multiple HLT paths, but a single representative trigger is selected from each. We explicitly require that an event pass at least one of the selected triggers. These are discussed below [66]:

- `HLT_IsoMu24`: This is the single muon trigger associated with the SingleMuon dataset. It requires the presence of a relatively isolated muon with a transverse momentum $\mathrm{p_T} > 24\ \mathrm{GeV}$.

- `HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8`: This is the double muon trigger associated with the DoubleMuon dataset. It required two muons with $\mathrm{p_T} > 17\ \mathrm{GeV}$ and $\mathrm{p_T} > 8\ \mathrm{GeV}$, respectively. Both muons must satisfy a loose track-based isolation cut (VVL = Very Very Loose), a cut on the longitudinal impact parameter $|d_z|$, and a cut on the invariant mass ($> 3.8\ \mathrm{GeV}$).

4.3.2 Object selections

In this analysis, the objects of interest for reconstructing the H+jet process are muons and jets, corresponding to the specific decay channel $\mathrm{H} \to \mathrm{ZZ}^* \to 4\mu$, which offers a clean final state, and the quark initiating a jet. Both have to pass additional object reconstruction quality criteria. Table 4.3 provides a general overview of the selection criteria. In what follows, we shortly describe the requirements for each object.

Muons

The muon requirements are inspired by the studies of the Higgs boson properties in the four-lepton final state, presented in Ref. [66]. We will consider two types of muon objects that satisfy different reconstruction quality criteria: *loose* and *tight* muons. Loose muons are PF muon objects with $\mathrm{p_T} > 5\ \mathrm{GeV}$ and $|\eta| < 2.4$, satisfying the Loose ID introduced in Section 3.3.2. Additionally, some compatibility cuts on the vertices are done to suppress cosmic muons. Specifically, we require the transverse d_{xy} and longitudinal d_z impact parameter of the associated track to satisfy $|d_{xy}| < 0.5\ \mathrm{cm}$ and $|d_z| < 1.0\ \mathrm{cm}$ with respect to the primary vertex. While the Loose ID was optimised to identify prompt muons and muons from secondary hadron decays, we want to select only prompt muons. Therefore, we take the muon selection a step further to define *tight* muons, a subset of loose muons. In terms of identification, these muons must satisfy the Tight ID. Furthermore, we expect the prompt muons to be isolated, meaning they are not surrounded by significant activity in the detector, such as nearby tracks or energy deposits, as would be typical for muons in jets. They must therefore satisfy a relative isolation requirement of $\mathcal{I}^{\mu}_{\mathrm{rel}} < 0.35$. To ensure a common primary vertex, the significance of the 3D impact parameter $\mathrm{SIP_{3D}}$ is required to satisfy $|\mathrm{SIP_{3D}}| < 4$, where $\mathrm{SIP_{3D}} = \frac{\mathrm{IP}}{\sigma_{\mathrm{IP}}}$. Here, IP is called the impact parameter and is defined as the point of closest approach to the primary vertex, and σ_{IP} is the uncertainty associated with the IP. In total, the presence of *exactly* four tight muons in the event is required.

Jets

Before any selection is applied to the reconstructed jets, the JES corrections are applied to both data and simulation. In contrast, the JER correction is only applied to simulation. The selection criteria for jets follow the recommendations described in Ref. [64]. These are followed so that the phase space of the jets used in this analysis matches that used in the derivation. Calibrated jets must satisfy $\mathrm{p_T} > 20\ \mathrm{GeV}$ and must be within the acceptance of the tracker, $|\eta| < 2.5$. In terms of object reconstruction quality, the jets must pass two criteria:

1. *Tight Jet ID*, designed to remove any jets initiated by detector noise or other signatures.

2. *Tight Pileup ID*, applied to jets with $\mathrm{p_T} < 50$ GeV to suppress jets arising from pileup. More information about these criteria can be found in Ref. [67].

To remove any jets that result from a clustering of pileup around a muon track, jets are required to be separated from the muons by imposing $\Delta R(\mathrm{jet}, \mu) < 0.4$. Finally, some regions of the detector, particularly the calorimeters, are known to have reduced performance due to suboptimal calibration. Jets reconstructed in these regions are vetoed. At least one jet passing all these criteria is required for each event.

Object	Selection Criteria
Loose muon	$\mathrm{p_T} > 5$ GeV, $\lvert\eta\rvert < 2.4$ $\lvert\mathrm{d}_{xy}\rvert < 0.5$ cm, $\lvert\mathrm{d}_z\rvert < 1.0$ cm *Loose ID*
Tight muon	Loose muon *Tight ID* $\mathcal{I}^{\mu}_{\mathrm{rel}} < 0.35$ $\lvert\mathrm{SIP}_{3D}\rvert < 4$
Jet	$\mathrm{p_T} > 20$ GeV, $\lvert\eta\rvert < 2.5$ *Tight Jet ID* & *Tight Pileup ID* (if $\mathrm{p_T} < 50$ GeV) $\Delta R(\mathrm{jet}, \mu) > 0.4$ Jets in veto regions are excluded

Table 4.3: *Summary of object selection criteria applied during the object preselection. The muon selection distinguishes between loose and tight muons. The latter category is used for the Higgs boson reconstruction.*

In summary, we require events to be triggered by either the single or double muon HLT path, have exactly four tight muons, and have at least one high-quality jet. These events subsequently progress to the next stage, technically referred to as analysis, where the H+jet process is reconstructed.

4.4 Analysis

4.4.1 Higgs boson reconstruction

After the object preselection, the Higgs boson is reconstructed and the accompanying jet is selected. This subsection focuses on the Higgs boson reconstruction. The subsequent subsection describes the jet selection. As with the muon selection criteria, the Higgs boson reconstruction is inspired by the measurement of the Higgs boson cross section in the four-lepton final state [66].

To begin, we require events to satisfy a **trigger compatibility** condition: at least one muon must have $\mathrm{p_T} > 20$ GeV and at least two muons must have $\mathrm{p_T} > 10$ GeV. For events passing this initial cut, we reconstruct the Higgs boson by reconstructing Z bosons and combining them into ZZ pairs. The logic is as follows:

1. **Z candidates:** We merge muon pairs into Z candidates if they form an opposite-sign pair.

2. **ZZ candidates:** Necessarily, events must contain two Z candidates to form a ZZ candidate. Among these, we label Z_1 as the Z candidate with an invariant mass closest to the nominal Z boson mass of $m_Z = 91.1876$ GeV [14], and the other candidate is labelled as Z_2. The ZZ candidate must pass an additional set of requirements, listed below:

- Z_1 **mass**: The Z_1 mass m_{Z_1} must satisfy $40\ \text{GeV} < m_{Z_1} < 120\ \text{GeV}$.
- Z_2 **mass**: The Z_2 mass m_{Z_2} must satisfy $12\ \text{GeV} < m_{Z_2} < 120\ \text{GeV}$.
- **Low-mass dilepton cleaning**: To avoid contamination of the low-mass J/ψ dilepton resonance, all opposite-sign muon pairs $(\mu, \mu^{'})$ in the event must satisfy $m_{\mu\mu'} > 4\ \text{GeV}$.
- **Ghost removal**: In rare cases, two muons may be reconstructed from a single track. These duplicated objects are called *ghost muons*. To ensure such duplicate muons are not selected in the final state reconstruction, we require each of the four muons to satisfy $\Delta R > 0.02$ between each other.

3. H **candidate:** If the ZZ candidate passed all the above criteria, the objects are merged into the Higgs boson candidate. The invariant mass $m_{4\mu}$ of the four-muon final state must satisfy $m_{4\mu} > 70\ \text{GeV}$. This invariant mass will also be called the invariant mass of the reconstructed Higgs boson candidate m(H).

We previously mentioned that the data is *blinded*. The blinding procedure is straightforward: data events are excluded from output if the reconstructed Higgs boson candidate has an invariant mass within the window $[115, 130]\ \text{GeV}$. This mass window reflects where the Higgs boson invariant mass distribution is expected to peak, as seen in simulation, and therefore almost fully removes potential signal from the data.

4.4.2 Jet selection

With the Higgs boson candidate reconstructed, the next step is selecting the accompanying jet candidate. As mentioned at the beginning of the chapter, the kinematic properties of the H+jet topology are the primary motivation behind the jet selection strategy.

Transverse momentum conservation implies that the Higgs boson and the jet are approximately back-to-back in the ϕ-plane, with their transverse momenta approximately balanced, since the initial state has no net transverse momentum before the collision. Using these variables, a likelihood-based selection is used, which evaluates the likelihood that a jet originated from the parton associated with the Higgs boson. The method consists of two main steps:

1. **Derivation of reference distributions from simulation:** Using the simulated H+jet sample, a parton-to-jet matching is performed to associate the generated jets in the event with the parton from the hard scattering process. A match is defined by ΔR(parton, jet) < 0.3. Matched jets are labelled 'good', and all other jets in the event are labelled 'bad'. Only the first matched jet is used if multiple 'good' jets are found. This procedure is repeated in bins of the reconstructed Higgs boson transverse momentum $\mathrm{p_T(H)}$. For each bin, we extract distributions of the following variables for both 'good' and 'bad' jets:

   - The azimuthal angle difference between the generated Higgs boson and the jet, labelled as $\Delta\phi(\text{jet})$. In the H+jet topology, this value is distributed around $\pm\pi$;
   - The transverse momentum ratio, labelled as $\frac{p_\mathrm{T}(\text{jet})}{p_\mathrm{T}(\mathrm{H})}$. In the H+jet topology, this value is distributed around 1.

   The resulting normalised distributions in the $30\ \text{GeV} < \mathrm{p_T(H)} < 50\ \text{GeV}$ bin are shown in Figure 4.2 as an example. As seen in Figure 4.2, the azimuthal difference distribution of 'good' jets peaks at $\pm\pi$, while 'bad' jets are distributed approximately uniformly. Similarly, as expected, the transverse momentum ratio shifts toward unity for 'good' jets than for 'bad' jets. These distributions are used as probability density functions, serving as our input for the likelihood evaluation.

2. **Likelihood evaluation:** With reference probability density functions defined, we can use them to evaluate the jets in any event. For each reconstructed jet in that event, the variables $\Delta\phi(\text{jet}, \text{H})$ and $\frac{p_\text{T}(\text{jet})}{p_\text{T}(\text{H})}$ are calculated, where H represents the reconstructed Higgs boson candidate in the event. Then, for each variable x, a likelihood ratio is calculated:

$$\mathcal{L}(x) = \frac{\mathcal{L}_\text{good}(x)}{\mathcal{L}_\text{bad}(x)} \quad \text{with} \quad x \in \left\{ \Delta\phi(\text{jet}, \text{H}),\ \frac{p_T(\text{jet})}{p_T(\text{H})} \right\}, \tag{4.1}$$

where $\mathcal{L}_\text{good}(x)$ and $\mathcal{L}_\text{bad}(x)$ are the values of the previously defined distributions evaluated for the jet in question. The final likelihood associated with a jet is the product of the two likelihood ratios:

$$\mathcal{L} = \mathcal{L}(\Delta\phi(\text{jet}, \text{H})) \cdot \mathcal{L}\left(\frac{p_\text{T}(\text{jet})}{p_\text{T}(\text{H})}\right). \tag{4.2}$$

The jet with the highest final likelihood in an event is selected as the candidate accompanying the reconstructed Higgs boson candidate.

Selecting a jet in its way, we inherently use the properties that a jet in the H+jet process is expected to have. Consequently, the chosen candidate is most likely to originate from the parton that is associated with the Higgs boson. We now briefly discuss the performance of the jet selection algorithm. The quantity used to characterise the performance is the *jet selection purity*, defined as the fraction of events where the selected jet is the jet that corresponds with the generator-level parton. This purity is only calculated for events where such a matched parton-jet pair exists and where the selected jets have a transverse momentum above $30\ \text{GeV}$. It is evaluated in bins of the transverse momentum of the generated parton. Furthermore, it is calculated separately for quark-initiated and gluon-initiated jets. A combined purity is also computed as a weighted average based on the relative abundance of the quarks and gluons in each bin. The results of this study are shown in Figure 4.3. Above $15 - 30\ \text{GeV}$, a combined purity of higher than $60\ \%$ is achieved, and above $30\ \text{GeV}$, purities higher than $80\ \%$ are achieved. This demonstrates that the likelihood method achieves a high degree of correctly identifying the jet in H+jet production.

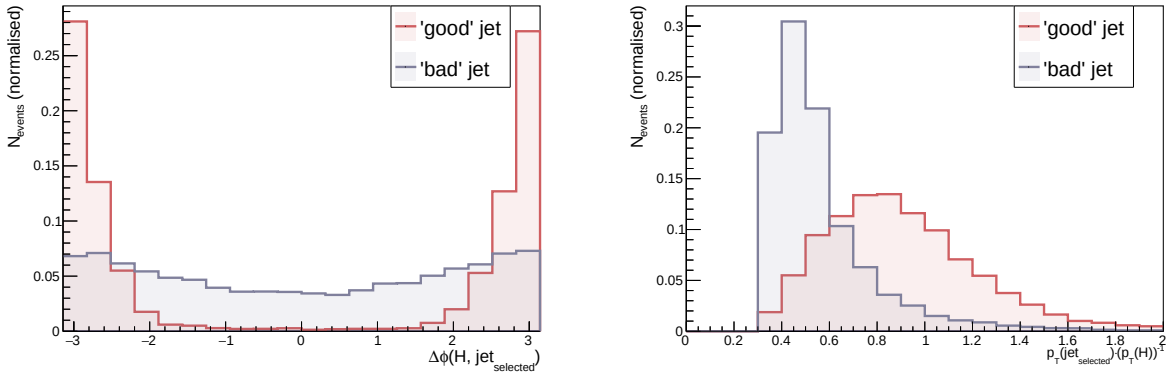

Figure 4.2: *The normalised distribution of the difference in the ϕ coordinate between the Higgs boson and jet $\Delta\phi(\text{jet}, \text{H})$ (left) and the ratio of the transverse momenta of the two objects $\frac{\text{p}_\text{T}(\text{jet})}{p_\text{T}(\text{H})}$ (right) for 'good' and 'bad' parton matched jets in the $30\ \text{GeV} < \text{p}_\text{T}(\text{H}) < 50\ \text{GeV}$ bin of transverse momentum of the Higgs boson in the event. These distributions are used as probability density functions in the likelihood-based jet selection [68].*

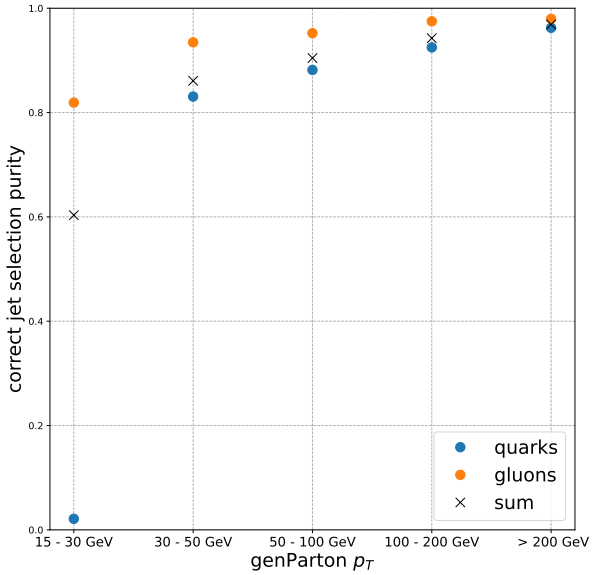

Figure 4.3: *The jet selection purity studied for the H+jet simulation sample in bins of the transverse momentum of the generated parton (genParton* $\mathrm{p_T}$*). Results are shown for quarks and gluons separately, as well as the average sum of the purities [68].*

In summary, the technical 'analysis' step of the H+jet framework consists of reconstructing a Higgs boson candidate using four tight muons and selecting an accompanying jet candidate. Based on this signature, we define two regions of events:

1. **Inclusive region (IR)**: Any event for which a H+jet candidate is reconstructed is considered in this region. Sometimes, we will refer to it as the full spectrum.

2. **Signal region (SR)**: A subset of the inclusive region, where the invariant mass of the reconstructed Higgs boson candidate lies within the range m(H)$\in [115, 130]$ GeV. This is the region used for deriving the upper limit on the production signal strength $\mu_{\mathrm{H+c}}$ in Chapter 6.

4.5 Output

4.5.1 Event yields

The previously described reconstruction algorithm is applied to both the signal and background simulated samples listed in Table 4.1.

In addition, the H+jet sample is split into different components by identifying the flavour of the generator-level parton that is associated with the Higgs boson:

- H+b, for events where the parton is a bottom quark;
- H+c, for events where the parton is a charm quark;
- H+l, for events where the parton is a light quark or gluon.

This categorisation allows us to isolate the subset of y_c-sensitive events that contain a charm quark in the final state, referred to here as the H+c process, which defines our signal process of interest. The H+jet contributions without a final-state charm quark, such as H+b and H+l, are treated as

background in the final measurement. Since all these contributions originate from the same simulation sample, we continue to refer to them collectively as the signal sample for simplicity.

Since an arbitrary number of events can be simulated, the number of selected simulated events N^{sel} must be scaled to match the expected yield corresponding to the quantity of recorded data, referred to as the *process yield* N_{exp}. This scaling is based on the cross section σ of the process, the integrated luminosity L of the recorded dataset ($59.83\ \text{fb}^{-1}$) and the total number of generated events N_{MC} of the process (see Table 4.1). The expected yield is then given by:

$$N_{exp} = L\sigma \frac{N^{\text{sel}}}{N_{\text{MC}}} \tag{4.3}$$

This procedure is applied uniformly across all simulated samples.

Table 4.4 summarizes the yields for each process in both the inclusive region (IR) and the signal region (SR). The total expected yield is defined as the sum of all irreducible background contributions, excluding those in the signal category (i.e., H+jet processes).

Category	Label	N_{exp} (IR)	N_{exp} (SR)
Signal	H+c	0.28	0.27
	H+b	1.61	1.53
	H+l	0.45	0.43
Irred. H. bkg.	tqH	0.17	0.16
	ttH	0.46	0.24
	ggH	15.59	14.84
	VBF	2.07	1.97
	VH	1.54	0.94
Irred. non-H. bkg.	qq →ZZ, Zγ^*	227.19	5.23
	gg→ZZ, Zγ^*	24.83	0.28
Total expected		271.85	23.66

Table 4.4: *Expected number of events yields in the inclusive region (IR), corresponding to all events for which a H+jet candidate is reconstructed, and the signal region (IR). The latter is defined as the subset of events for which the mass of the reconstructed Higgs boson candidate m(H) lies between* 115 *and* 130 *GeV. Labels correspond to those in Table 4.1. The matched parton flavour splits the signal event sample, and gg→ZZ contributions are combined. The total expected yield is defined as the sum of all irreducible background contributions, excluding those in the signal category (i.e., H+jet processes).*

In the inclusive region, over the full spectrum of events, the irreducible non-Higgs backgrounds dominate. In particular, the qq→ ZZ, Zγ^* process is the most prominent, with a yield of an order of magnitude higher than the second-highest process. Among the irreducible Higgs backgrounds, the ggH process has the highest yield. As a last remark, notice how the yields of the H+jet process are minimal compared to the total yield. Additionally, slight differences are observed based on the flavour of the parton. The H+b category has the highest yield due to the Yukawa coupling's relation to the mass. The H+c category has the lowest yield, while the H+l category lies in between. The latter consists of a combination of y_c and y_b sensitive events in which a gluon or light quark appears in the final state instead of a charm or bottom quark.

On the other hand, the observations reveal that the processes dominating changes vary signif-

icantly in the signal region. The non-Higgs backgrounds are substantially reduced. Instead, the ggH process is now the dominant one.

4.5.2 Kinematic distributions

Now, the kinematic distributions of the processes are investigated. The most informative distribution here is the invariant mass of the reconstructed Higgs boson candidate, $\mathrm{m(H)}$, which is expected to show the most differences between the irreducible backgrounds. The invariant mass

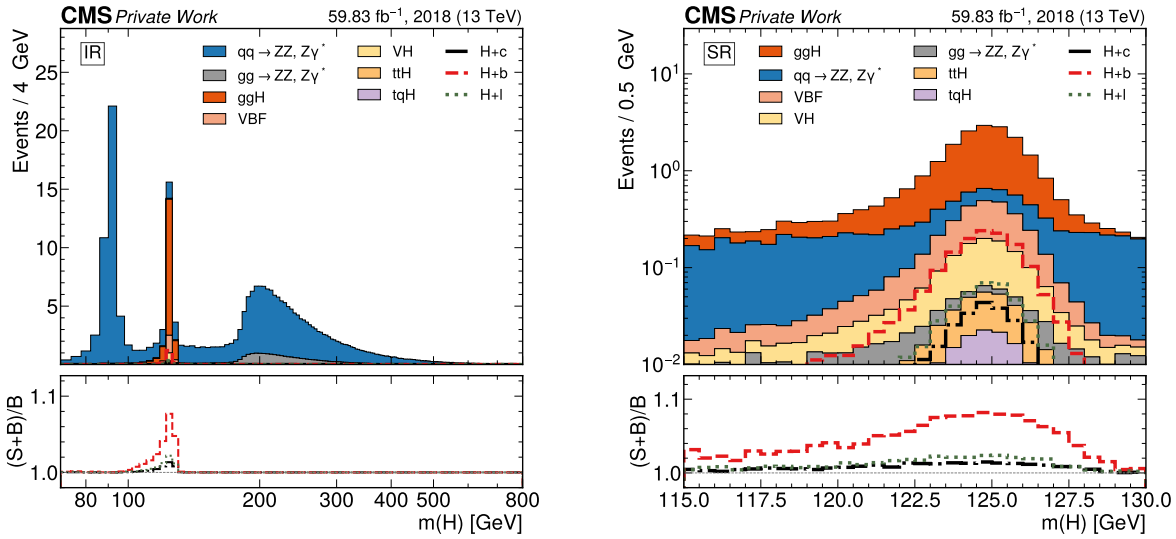

Figure 4.4: *Upper panels: Invariant mass of the reconstructed Higgs boson candidate* $\mathrm{m(H)}$ *in the inclusive region (left) and signal region (right) stacked for the irreducible backgrounds and overlaid with the H+jet signal sample. Lower panels: The (S+B/B) ratio for each signal category in the H+jet sample.*

distribution $\mathrm{m(H)}$ of the reconstructed Higgs candidate in the inclusive region is shown in the left plot of Figure 4.4. In the upper panel, the sum of histograms of the irreducible background processes is plotted, excluding the H+jet signal category. The H+jet contributions are overlaid separately and split by the flavour of the associated parton. In the lower panel, the relative impact of each signal category S on the overall spectrum is illustrated by plotting the ratio (S+B)/B, where B is the sum of the stacked background contributions. We observe that the qq$\rightarrow$ ZZ, Zγ^* processes cause a peak around 90 GeV, mainly due to single-resonant four-lepton production (Z $\rightarrow 4\mu$). Moving to higher invariant masses, around the Higgs boson mass of 125 GeV, the irreducible Higgs boson processes dominate, especially the ggH process. At around 180 GeV, both Z bosons can kinematically be produced on shell. This leads to a yield surge due to the qq$\rightarrow$ ZZ, Zγ^* and gg$\rightarrow$ZZ processes. The H+jet signal remains masked in the total spectrum in the upper panel and is almost invisible. The effect of the signal process is most clearly visible in the lower panel. The right plot of Figure 4.4 focuses more on the signal region. As discussed in the Section 4.5.1, the ggH background dominates here, but the signal contributions have a more pronounced effect. More clearly, we can see the individual distributions of the signal processes. Investigations into differences in the shape of the Higgs mass spectrum between signal and background are left for future work.

The transverse momentum distributions of the reconstructed Higgs boson candidate and the selected jet in the inclusive region are shown in Figure 4.5. All processes, including the signal, show

a steep fall for higher p_T and most events are concentrated in the low p_T region.

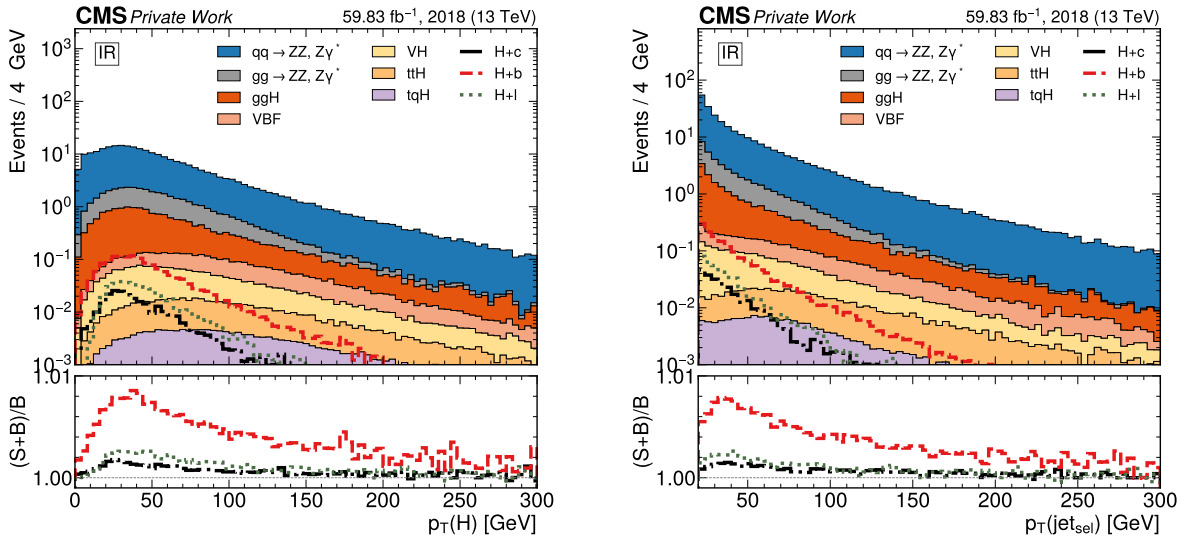

Figure 4.5: *Upper panels: Transverse momentum distributions of the reconstructed Higgs boson candidate* $p_T(H)$ *(left) and the selected jet* $p_T(jet_{sel})$ *(right) in the inclusive region (IR) stacked for the irreducible backgrounds and overlaid with the H+jet signal sample. Lower panels: The (S+B/B) ratio for each signal category in the H+jet sample.*

4.5.3 Jet discriminator spectra

Next to kinematic variables, we also have the DeepJet jet flavour discriminator variables CvsB and CvsL, introduced in Section 3.4.3, associated with each selected jet. These are particularly important for inferring the flavour origin of this selected jet in the H+jet sample. The CvsB output in both the inclusive and signal regions is shown in Figure 4.6, and for CvsL in Figure 4.7. We observe a significant contribution from H+b events. These features are even clearer in the signal region. This is expected, as the H+b process has a cross section approximately ten times larger than that of H+c alone.

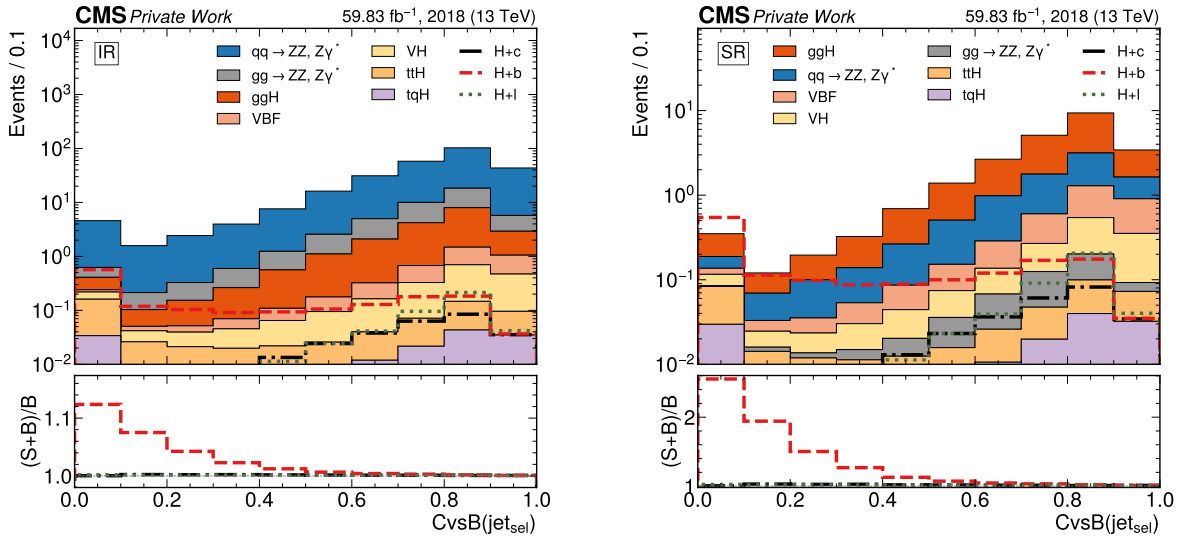

Figure 4.6: *Upper panels: CvsB jet flavour discriminator distributions for the selected jet in the inclusive region (left) and signal region (right) stacked for the irreducible backgrounds and overlaid with the H+jet signal sample. Lower panels: The (S+B/B) ratio for each signal category in the H+jet sample.*

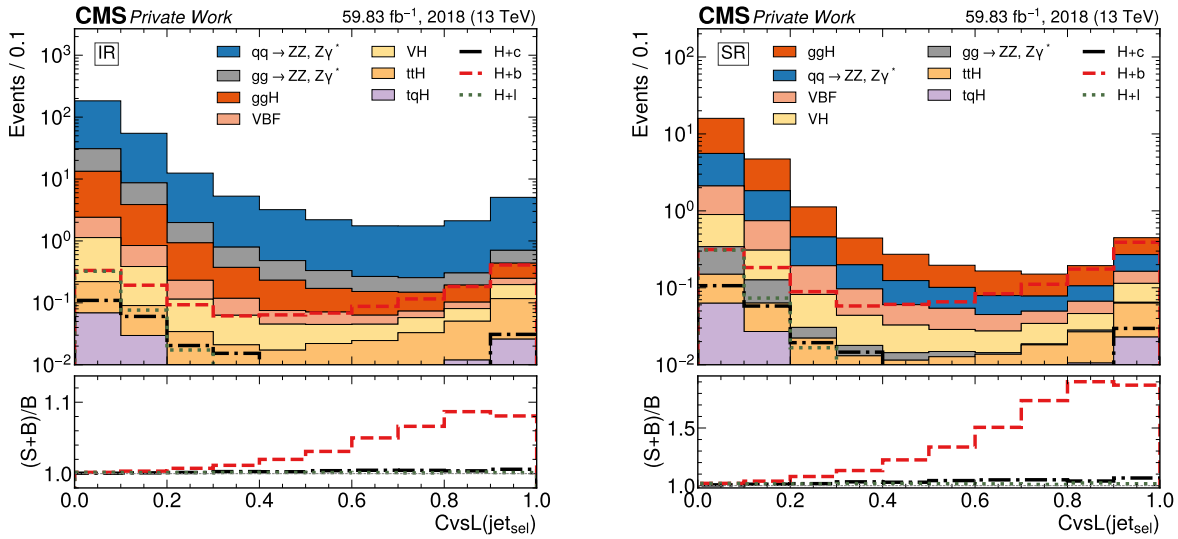

Figure 4.7: *Upper panels: the CvsL jet flavour discriminator distributions for the selected jet in the inclusive region (left) and signal region (right) stacked for the irreducible backgrounds and overlaid with the H+jet signal sample. Lower panels: The (S+B/B) ratio for each signal category in the H+jet sample.*

5 REDUCIBLE BACKGROUND ESTIMATION

In this chapter, the reducible backgrounds are estimated via a data-driven approach. Central to this method lies the misidentification rate, which characterises the probability of a non-prompt object satisfying the criteria of the prompt muons. Section 5.1 introduces the procedure for measuring the misidentification rate, along with the concept of control regions, which are defined as regions enriched by a background process. After that, Section 5.2 explains how the misidentification rate is applied to certain control regions dominated by reducible backgrounds to estimate their contribution to the inclusive and signal regions. As a reminder, the inclusive region includes all selected H+jet candidates via the procedures outlined in Chapter 4, while the signal region refers to the subset of events for which the mass of the reconstructed Higgs boson candidate m(H) lies between 115 and 130 GeV. After characterising the control regions, the overall yield of reducible background will be estimated. Additionally, the distributions of the reconstructed Higgs boson candidate and the jet flavour discriminators of the reducible background will be presented, along with a discussion of the associated uncertainties on the muon misidentification rate method. In the last section of this chapter, Section 5.3, the results from this and the previous chapter are combined: the distributions in the inclusive region are compared to data, and the signal region distributions are updated with the reducible background. The methods presented in this chapter closely follow the studies of the Higgs boson cross section measurement in the four-lepton final state, presented in Ref. [66]. For consistency and easy reference, the same notation is adopted.

5.1 Measuring the muon misidentification rate

5.1.1 Determination procedure

As repeatedly mentioned, reducible backgrounds are processes in which fewer than four prompt muons are produced. However, additional objects are misidentified, still resulting in the same leptonic signature as the signal, which is four prompt muons. In the literature, these misidentified objects, being non-prompt leptons or jets, are called *misidentified muons*[1]. Before any reducible background estimation can occur, we must establish the probability of a misidentified muon to satisfy the criteria we require prompt muons to do. This probability is referred to as the *misidentification rate*. To measure it, we must carefully design regions that are expected to be densely populated by these misidentified muons. Generally, a region in phase space expected to be dominated by specific background signatures or processes is called a *control region* (CR), and this naming will be used frequently throughout this chapter.

The misidentification rate is measured via a Z+ℓ control region in data. This phase space consists of three muon events, two of which originate from the Z boson and are therefore required to satisfy the prompt criteria of the analysis, which is the tight requirement. The third muon, which is not expected to be prompt, must satisfy only the minimal identification criteria and is therefore required to be at least a loose muon. The definitions of loose and tight muons are shown in Table

[1]Sometimes also referred to as the fake muons. However, we will try to avoid this terminology.

4.3. This third muon is referred to as the *probe muon*, because it serves as a probe for the misidentification rate measurement.

We define the misidentification rate f as the ratio of the number of probe muons that pass the tight selection N_{tight} to the total number of probe muons passing the loose selection N_{loose}:

$$f = \frac{N_{\text{tight}}}{N_{\text{loose}}}. \tag{5.1}$$

This ratio essentially captures what is necessary: the probability of a non-prompt muon satisfying the criteria that are required for a prompt muon. As muon reconstruction depends on the objects' pseudorapidity and transverse momentum, the misidentification rate is typically measured in bins of these variables.

5.1.2 Constructing the Z+ℓ control region

In this subsection, the construction of the Z+ℓ control region (CR) for measuring the misidentification rate in data is described. In general, we follow the same overall structure of event selection and analysis described in the previous chapter, visualized in Figure 4.1, but with differences in implementation. For one, the reconstruction in the Z+ℓ CR is significantly simpler. Here, we select events with exactly three muons, two of which we expect to originate from a Z$\rightarrow \mu^-\mu^+$ decay, and a third that serves as the probe for the misidentification rate measurement. For a data-driven determination of this rate, the measurement is performed using the 2018 data event samples listed in Table 4.2.

Care must be taken to ensure that only appropriate processes contribute to the Z+ℓ CR. The misidentification rate method estimates the probability of a misidentified muon passing the criteria required to be classified as a tight muon, as defined in Table 4. However, while this region is expected to be enriched in misidentified muons, it also receives contributions from processes with three prompt muons. Including such events would inflate the misidentification rate defined in Eq. (5.1), and their contribution must therefore be estimated using simulation and subtracted from the control region.

The dominant process with three prompt muons is WZ $\rightarrow 3\ell\nu$, in which the W boson decays leptonically via W $\rightarrow \mu\nu_\mu$ and the Z boson decays to a muon pair. All of these prompt muons are likely to satisfy the tight identification criteria and, as described above, their contribution is estimated using simulation. Processes with four prompt muons may also contribute to the Z+ℓ CR if one of the prompt muons fails to be reconstructed or identified. Such events are rare, but a non-negligible contribution is expected for the $\mathrm{qq} \rightarrow \mathrm{ZZ}, \mathrm{Z}\gamma^*$ processes. As shown in Table 4.1, these have a cross section that is much larger than that of any other process with four prompt muons. Its contribution to the control region is also estimated using simulation.

Apart from these WZ $\rightarrow 3\ell\nu$ and $\mathrm{qq} \rightarrow$ZZ, Zγ^* processes, several simulated event samples expected to contain misidentified muon signatures are analysed as well: Drell-Yan, $\mathrm{t\bar{t}}$, WZ $\rightarrow 2\mathrm{q}2\ell$, and $\mathrm{q\bar{q}} \rightarrow$ ZZ $\rightarrow 2\ell 2\mathrm{q}$. These are not relevant for the measurement of the data misidentification, but are used to investigate the composition in the Z+ℓ CR. Later on, they are also used to measure the misidentification rates in simulation, which will be used to determine the uncertainties on the method. These process samples are used throughout the chapter, albeit with some processes contributing more than others, depending on the CR being investigated. An overview of all the previously mentioned simulated event samples used is given in Table 5.1.

Label	Process	σ [fb]	Generated events N_{MC}
Drell-Yan	Drell-Yan	6424000	195 510 810
$t\bar{t}$	$t\bar{t}$	815960	304 895 029
WZ $\to$2q2ℓ	WZ $\to$2q2ℓ	55950	293 600
$q\bar{q} \to ZZ \to 2\ell 2q$	$q\bar{q} \to ZZ \to 2\ell 2q$	3220	200 000
$WZ \to 3\ell\nu$	$WZ \to 3\ell\nu$	4429.65	100 000
$qq \to ZZ, Z\gamma^*$	$q\bar{q} \to ZZ, Z\gamma^*, Z \to 4\ell$	1256	98 488 000

Table 5.1: *Simulated event samples used to validate the misidentification rate method. Cross sections (σ) are shown in femtobarns [fb], and N_{MC} indicates the number of generated events. The Label column shows the labels for the processes used in plots throughout this chapter.*

In what follows, the techniques for constructing the Z+ℓ control region are highlighted. These are taken from Ref. [66].

Object selection

Events must satisfy the following preselections:

- **Triggering:** Events must have triggered the relevant single- or double-muon trigger.
- **Muon multiplicity:** Exactly three loose muons must be present, with at least two being tight.
- **MET:** The event must have MET below 25 GeV, reducing the presence of the WZ$\to$ $3\ell\nu$ process, where the neutrino is responsible for most of the missing transverse momentum.

Event reconstruction

First, the event must pass the previously discussed trigger compatibility requirement in Section 4.4. The logic for the reconstruction and identification of Z+ℓ candidates is as follows:

1. **Z boson reconstruction:** Opposite-sign tight muon pairs are merged into Z boson candidates. If multiple Z candidates are present, the one whose invariant mass is closest to the nominal Z boson mass, $m_Z = 91.1876$ GeV [14], is selected as the candidate Z boson.
2. **Probe muon identification:** The muon not used for the Z reconstruction is considered to be the probe muon.
3. **Additional quality selections:** To ensure that events are consistent with a reconstructed Z+ℓ final state, the following selection criteria are applied:
    - **Mass**: The invariant mass of the Z candidate $m(Z_{\text{cand}})$ must be within the window $|m(Z_{\text{cand}}) - m(Z)| < 7$ GeV. By retaining only events within this mass window, we select a high-purity sample in which the muon pair is consistent with a Z boson decay.
    - **Low-mass dilepton removal**: The invariant mass of the probe muon μ and the opposite-sign tight muon from the reconstructed Z boson μ' must satisfy $m_{\mu\mu'} > 4$ GeV.
    - **Ghost removal:** $\Delta R > 0.02$ between each of the three muons.

As validation, the process composition in the Z+ℓ control region is investigated by considering the transverse momentum distribution of the probe muon. The comparison between data and simulation is shown in Figure 5.1 in the upper panel, while the bottom panel shows the Data / MC ratio. The binning of the transverse momentum is taken to be the same as in Ref. [66], with

the last bin containing overflow.

Regarding the process composition, it is seen that the Drell-Yan process mostly dominates, especially at lower $\mathrm{p_T}$ regions where most events are observed. However, in the high $\mathrm{p_T}$ regions, the contribution from the WZ $\rightarrow 3\ell\nu$ process starts to become significant as well. The contribution from the other considered processes is minimal. The Data / MC agreement is poor in the lower $\mathrm{p_T}$ regions. This is expected behaviour due to poor modelling of non-prompt leptons in Drell-Yan simulation, which dominates in this region.

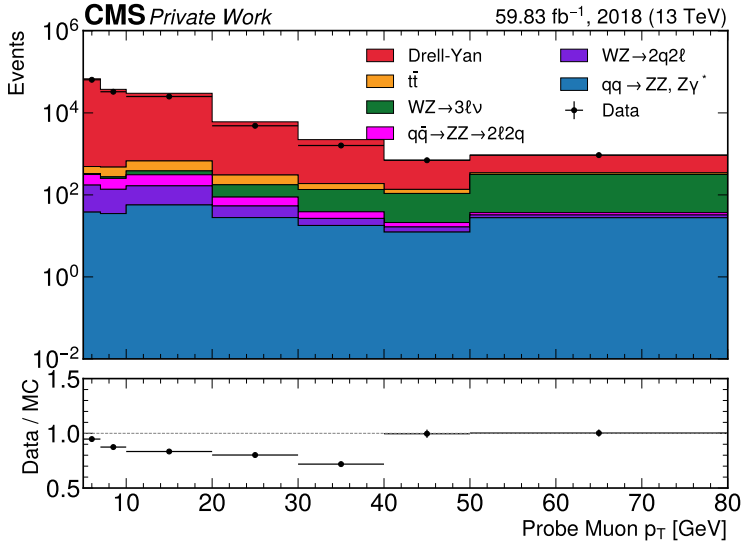

Figure 5.1: *Upper panel: The distribution of the probe muon transverse momentum ($\mathrm{p_T}$) in the Z+ℓ control region, comparing data to the stacked sum of simulated background processes given in Table 5.1. Lower panel: The data-to-simulation (Data / MC) ratio.*

5.1.3 Result

Having constructed a dedicated control region and studied the relative composition of processes inside it, the misidentification rate is calculated via Equation (5.1). This is done for both the barrel region ($|\eta| \leq 1.2$) and the endcap region ($|\eta| > 1.2$) with the same $\mathrm{p_T}$ binning as for the process composition. Specifically, the misidentification rate is calculated in two cases: for all the reconstructed Z+ℓ candidates in data (labelled as Uncorrected), and for the case where the expected yield in simulation of the WZ$\rightarrow 3\ell\nu$ and $\mathrm{qq} \rightarrow \mathrm{ZZ}, \mathrm{Z}\gamma^*$ processes is subtracted (labelled as Corrected). Considering the latter case, both simulation contributions are scaled to their expected yields, and the number of expected tight and loose events is subtracted from the observed number in each bin. Only after this operation is the misidentification rate calculated. The results of the procedure are shown in Figure 5.2. The uncertainties indicated here with the error bars are statistical. Systematic uncertainties related to the cross section of the WZ$\rightarrow 3\ell\nu$ or $\mathrm{qq} \rightarrow \mathrm{ZZ}, \mathrm{Z}\gamma^*$ processes are outside the scope of this work. These are expected to have only minimal effects on our final results.

From Figure 5.2, it is observed that the effect of removing the prompt lepton sources becomes more apparent for larger transverse momentum in both the barrel and endcap regions. This is expected, as observed from the process distribution in Figure 5.1, since the yields of WZ$\rightarrow 3\ell\nu$ increase for higher momenta. This effect is most drastic at the highest momenta, where a decrease in misidentification rate of over 50% is seen. Regarding the shape of the corrected misidentification

rate spectrum, first, a slight decline is seen in the misidentification rate as transverse momentum increases. One possible explanation for higher misidentification rates at low momenta is that the muon reconstruction may be more sensitive to detector noise, resulting in more misidentifications. For higher values of transverse momenta, > 40 GeV, an increase in misidentification rate is reported. Here, one hypothesis is that the selected probe tends to be isolated in the event and therefore more easily qualifies as tight. However, as mentioned at the beginning of the section, these misidentification rates strongly depend on which muon criteria are chosen. It is therefore hard to draw any conclusion without studying the effect of each criterion on its own, which is not done in this work. Instead, these corrected misidentification rates are used to determine the reducible backgrounds, as discussed in the next section.

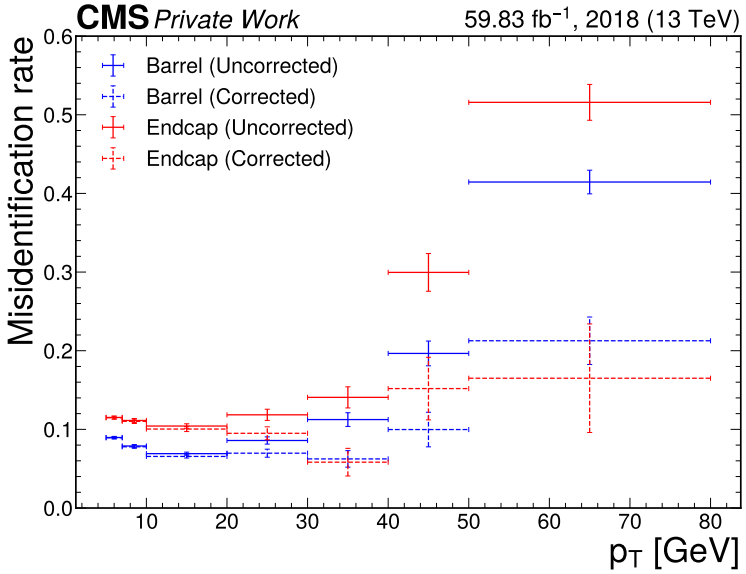

Figure 5.2: *The muon misidentification rate as a function of the transverse momentum* $\mathrm{p_T}$ *of the probe muon. The measurement is divided into two pseudorapidity regions: the barrel region* ($|\eta| \leq 1.2$)*, shown in blue, and the endcap region* ($|\eta| > 1.2$)*, shown in red. The solid lines represent the misidentification rate before subtraction of the* WZ$\rightarrow 3\ell\nu$ *and* $\mathrm{qq} \rightarrow \mathrm{ZZ}, \mathrm{Z}\gamma^*$ *processes (labelled as Uncorrected), while the dotted lines show those contributions have been removed (labelled as Corrected). Only the corrected misidentification rates are used in further analysis.*

5.2 Applying the muon misidentification rate

5.2.1 Application procedure

The previously calculated misidentification rate forms the basis of the reducible background estimations explained in this section. The general idea is as follows. The goal is to estimate the yield of the reducible background processes in the inclusive region. That is, we want to know the contribution of the reducible background in the region where we reconstruct the H+Jet process. By definition, the leptonic signature in this region is that of four muons identified as prompt, meaning that they all passed the tight identification requirement. This motivates us to refer to this region as 4P, where the P refers to a muon that passed the criteria. In contrast, reducible backgrounds are not expected to produce four prompt muons and be in the 4P region. Instead, they are expected to produce at least one muon that failed the tight criteria, denoted with an F, and is thus a loose-not-tight muon. To better understand these reducible contributions, four-muon

control regions need to be designed that are enriched by one or more of these failed muons F. Generally, these will be built by relaxing the muon criteria, matching the leptonic signature that the reducible contribution is expected to produce. Then, we must extrapolate these contributions from the control samples into the 4P region. The idea is to estimate, for each event, the probability that the loose-not-tight muons would have also passed the tight selection, and thus contributed to the 4P region. This is achieved by assigning weights to events based on the measured muon misidentification rate f in data, as calculated in Section 5.1. From the definition of the misidentification rate f given in Equation (5.1) we have:

$$N_{\text{tight}} = f \cdot N_{\text{loose}}. \tag{5.2}$$

However, in the control regions, we are interested in the contribution of the loose-not-tight muons, i.e., muons that passed the loose but failed the tight selection. This subset is given by:

$$N_{\text{loose-not-tight}} = (1 - f)N_{\text{loose}}, \tag{5.3}$$

Substituting into Equation (5.2), we find:

$$N_{\text{tight}} = \frac{f}{(1-f)} N_{\text{loose-not-tight}}. \tag{5.4}$$

This shows that for each loose-not-tight muon in the event, the weight is $\frac{f}{1-f}$ to estimate its contribution in the 4P region, where f is the misidentification rate associated with that muon, depending on its transverse momentum and pseudorapidity. In the analysis, this factor is applied per failing muon, and the total event weight is taken as the product over all such muons in that event.

Depending on the prompt and non-prompt signatures of the reducible background processes, two of these control regions are defined:

- **2P2F**: This first control region consists of four muon events where two prompt muons are produced (2P) and two additional muons failed the tight criteria, being loose-not-tight (2F). Hence, we refer to this region as the 2P2F region. The Drell-Yan and $\mathrm{t\bar{t}}$ processes are expected to dominate this region. To extrapolate events in this region into the 4P region, each event is weighted by the probabilities that the failing (i.e., loose-not-tight) muons would have passed the tight selection. If the total amount of events in this region is denoted by N_{2P2F}, and f_3^i and f_4^i represent the misidentification rate associated to the third and fourth muon that both failed the tight criteria, the total yield in the 4P region can be written as:

$$\sum_{i}^{N_{\text{2P2F}}} \frac{f_3^i}{(1-f_3^i)} \frac{f_4^i}{(1-f_4^i)}, \tag{5.5}$$

  where the weighting follows Equation (5.4) and where the sum runs over every event i in the control sample.

- **3P1F**: The second control region consists of events with three tight muons (3P) and one muon that failed the criterion, being loose-not-tight (1F), hence the name 3P1F. Consequently, reducible background processes producing three prompts should be accounted for here, such as the WZ$\rightarrow 3\ell\nu$ process. If the total amount of events in this region is denoted by N_{3P1F}, and f_4^i is the misidentification rate of the only failing muon in the event, then the total yield in the 4P region originating from this control region is written as:

$$\sum_{j}^{N_{\text{3P1F}}} \frac{f_4^j}{(1-f_4^j)}, \tag{5.6}$$

where the sum runs over each event j in the 3P1F control sample.

However, the total reducible background yield in the 4P region $N_{\mathrm{4P}}^{\mathrm{red.bkg.}}$ is not simply the sum of the yield contributions in the individual control regions, given by Equation (5.5) and (5.6). Two sources of contamination have to be accounted for. First, some events in the 3P1F regions might have one muon being misidentified as tight, meaning that the event only produced two prompt muons and should have been in the 2P2F region instead. This means that the 3P1F contribution is overestimated due to contamination of 2P2F reducible background processes. This contamination is estimated by calculating the expected amount of 3P1F events, denoted as $N_{\mathrm{3P1F}}^{\mathrm{exp.}}$, from the 2P2F region as:

$$N_{\mathrm{3P1F}}^{\mathrm{exp.}} = \sum_{i}^{N_{\mathrm{2P2F}}} \left(\frac{f_3^i}{(1-f_3^i)} + \frac{f_4^i}{(1-f_4^i)} \right). \tag{5.7}$$

The reasoning behind Equation (5.7) is that starting from an event in the 2P2F control region, we can weigh either failed muon in that event by its fake-rate-based weight to estimate its contribution in the 3P1F region, hence we take the sum of both. To evaluate the contribution of this $N_{\mathrm{3P1F}}^{\mathrm{exp.}}$ in the 4P region, the events must be additionally weighted by the fake-rate-based weight of the muon for which no weight has been applied yet. This means that we can write the 4P region contribution of this term as:

$$\sum_{i}^{N_{\mathrm{2P2F}}} \left(\frac{f_4^i}{(1-f_4^i)} \frac{f_3^i}{(1-f_3^i)} + \frac{f_3^i}{(1-f_3^i)} \frac{f_4^i}{(1-f_4^i)} \right) = 2 \sum_{i}^{N_{\mathrm{2P2F}}} \frac{f_3^i}{(1-f_3^i)} \frac{f_4^i}{(1-f_4^i)}, \tag{5.8}$$

from which it can be concluded that this term has twice the yield of the 2P2F region. This contribution must be subtracted from the total yield.

The second source of contamination is similar to the one encountered during the calculation of the misidentification rate. Processes expected to produce four prompt muons can end up in one of the control regions if one happens not to pass the tight criteria. Out of the two control regions, this is the most probable to occur in the 3P1F region. These contributions arise from irreducible processes, and since this estimation aims at reducible ones, they must be appropriately accounted for in simulation, and the yield must be subtracted. Again, any process producing four prompt muons needs to be considered. However, it is observed that only the $\mathrm{qq} \to \mathrm{ZZ}$, $\mathrm{Z}\gamma^*$ processes give rise to a non-negligible contamination yield. We denote the yield of this process in the 3P1F region as $N_{\mathrm{3P1F}}^{\mathrm{ZZ}}$, following the notations in the literature. To extrapolate its contribution into the 4P region, the events are weighted according to Equation (5.6).

With these considerations, the total reducible background yield in the 4P region $N_{\mathrm{SR}}^{\mathrm{red.bkg.}}$ is given by:

$$N_{\mathrm{4P}}^{\mathrm{red.bkg.}} = \sum_{j}^{N_{\mathrm{3P1F}}} \frac{f_4^j}{(1-f_4^j)} - \sum_{j}^{N_{\mathrm{3P1F}}^{\mathrm{ZZ}}} \frac{f_4^j}{(1-f_4^j)} - \sum_{i}^{N_{\mathrm{2P2F}}} \frac{f_3^i}{(1-f_3^i)} \frac{f_4^i}{(1-f_4^i)}, \tag{5.9}$$

where the first term reflects the 3P1F contribution (Equation (5.6)), the second term the $\mathrm{qq} \to \mathrm{ZZ}$, $\mathrm{Z}\gamma^*$ contamination that needs to be subtracted, and the third term the 2P2F contribution (Equation (5.5)) with the 3P1F contamination subtracted (Equation (5.7)). With the concepts explained, the construction of the control regions can be elaborated on.

5.2.2 Constructing 2P2F and 3P1F Control Regions

The design of these control regions is similar to that of the 4P region, as explained in Section 4.3 and Section 4.4, with key differences in muon identification. In what follows, we will briefly

describe the construction of these control regions, followed by the invariant mass distributions and the jet flavour spectra in these control regions.

2P2F

The only difference between the 2P2F control region and the 4P region's object selection is the muon selection: instead of requiring four tight muons, events must contain exactly two tight muons (2P) and two loose-not-tight muons (2F). To reconstruct a Higgs boson candidate in this region, the two tight muons are chosen to form the Z_1 candidate, while the remaining loose-not-tight muons form the Z_2 candidate. All kinematic criteria applied to these candidates remain identical to those in the 4P region (see Section 4.4.1). If a valid Higgs candidate is found, a jet is selected following the procedure described in Section 4.4.2. Events for which an H-jet pair is found with these selections are classified as belonging to the 2P2F control region.

3P1F

Similar to the 2P2F case, only the muon configuration differs from the 4P region. Here, three tight muons (3P) and one loose-not-tight muon (1F) are required. To reconstruct the Higgs boson candidate, the two tight muons that form an invariant mass closest to the nominal Z boson mass are selected to construct the Z_1 candidate. The remaining tight and loose-not-tight muon make up the Z_2 candidate. Again, all the kinematic selections remain the same. Events for which a Higgs boson candidate is reconstructed, a jet is selected as in the 4P region. Events for which an H-jet pair is found with these selections are classified as belonging to the 3P1F control region.

Distributions in the 2P2F and 3P1F control regions

The invariant mass of the reconstructed Higgs boson candidate in the 2P2F and the 3P1F control regions is shown in Figure 5.3. The CvsB and CvsL discriminator distributions are shown in Figures 5.4 and 5.5, respectively. Regarding the process composition, it is seen that the 2P2F region has large Drell-Yan and $\mathrm{t\bar{t}}$ contributions. Interestingly, the dominant contribution is the $\mathrm{t\bar{t}}$ process. On the contrary, in the studies in Ref. [66], the Drell-Yan process was the main reducible background in this region. The difference is likely due to different selection criteria. In particular, our analysis imposes explicit jet requirements on the final state, which were not required in the inclusive-four-muon decay channel. Fewer high p_{T} jets are expected at leading order in Drell-Yan events than in $\mathrm{t\bar{t}}$, as the latter is enriched by b jets from top quark decays. This way, more $\mathrm{t\bar{t}}$ events may pass the final selection. Considering the 3P1F region, shown in the right plot in Figure 5.3, the Drell-Yan and $\mathrm{t\bar{t}}$ processes are still the dominant, but the presence of $\mathrm{qq} \rightarrow$ ZZ, Zγ^*, and WZ$\rightarrow 3\ell\nu$ processes is increased compared to the 2P2F region. The Data / MC agreement in both control regions is reasonably consistent. Due to a lower overall yield, more statistical fluctuations are observed in the 3P1F region.

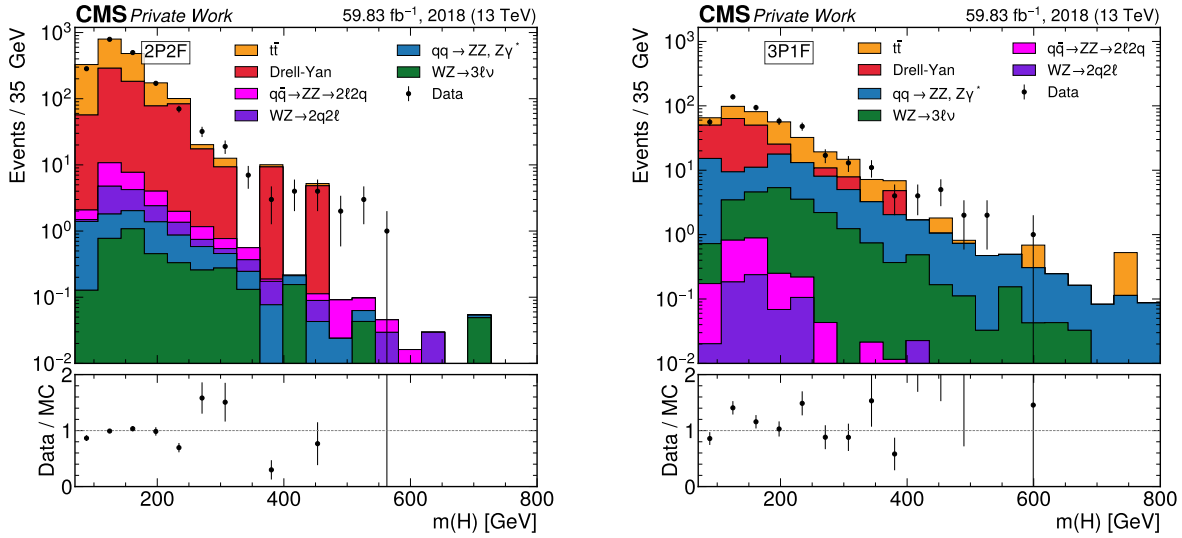

Figure 5.3: *Upper panels: The invariant mass of the reconstructed Higgs boson candidate m(H) for data and simulation in the 2P2F region (left) and the 3P1F region (right). Lower panels: The corresponding Data / MC ratio.*

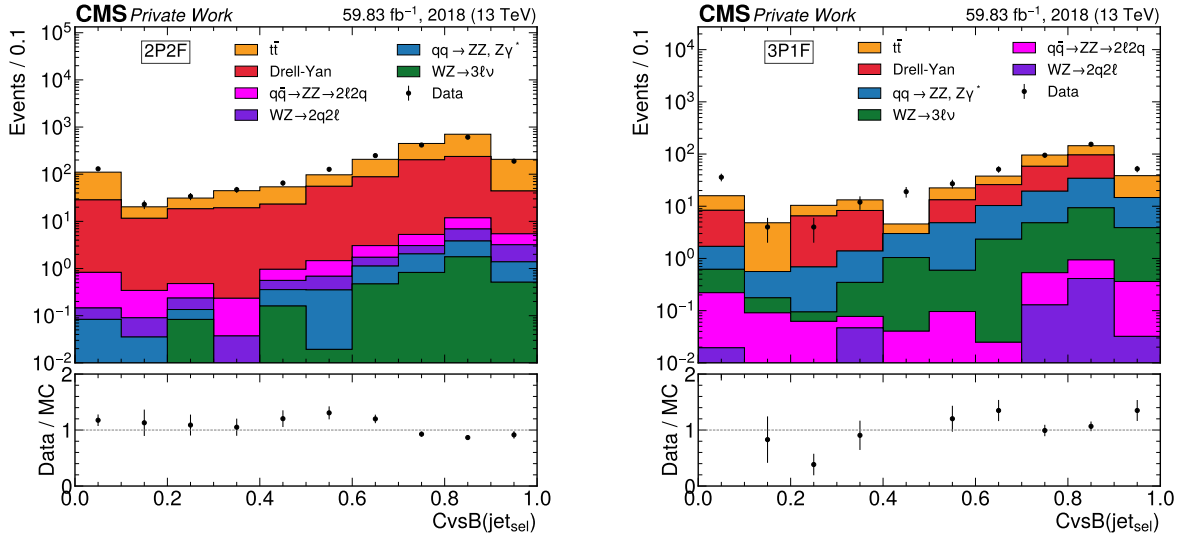

Figure 5.4: *Upper panels: CvsB jet flavour discriminator distributions for the selected jet in the 2P2F region (left) and 3P1F region (right). Lower panels: The corresponding Data / MC ratio.*

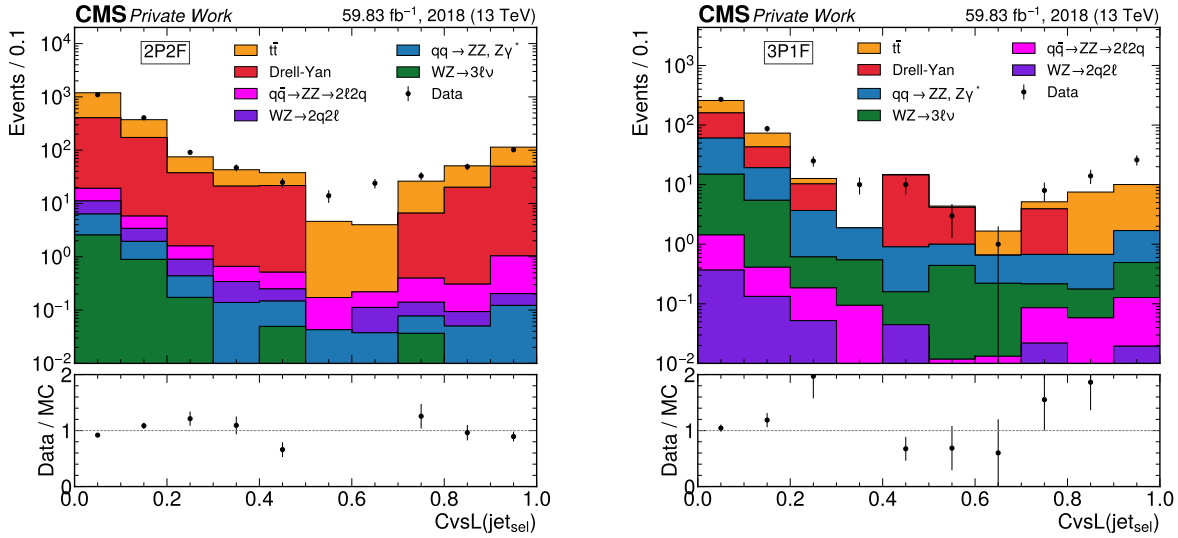

Figure 5.5: *Upper panels: CvsL jet flavour discriminator distributions for the selected jet in the 2P2F region (left) and 3P1F region (right). Lower panels: The corresponding Data / MC ratio.*

5.2.3 Reducible background yield and distributions

The goal of this subsection is to estimate the overall contribution from reducible processes and determine the distribution shapes needed for the final signal extraction. The yield is calculated with Equation (5.9), using the event counts in the 2P2F and 3P1F control regions. This is done for both the inclusive region, where no constraint is applied to the invariant mass of the reconstructed Higgs boson candidate, and for the signal region, defined as events where the invariant mass of the reconstructed Higgs candidate lies between 115 and 130 GeV. Beyond yields, we extract the shapes of key observables for the reducible background. These include the invariant mass of the reconstructed Higgs boson candidate and the outputs of the CvsL and CvsB jet flavour discriminators. To obtain these shapes, the binned control region histograms, shown in Figures 5.3, 5.4, and 5.5, are reweighted to get their contribution in the 4P region, and the final reducible background shapes are derived using Equation (5.9). The resulting histograms are used in the next section for comparison with data in the inclusive region. They are also added to the signal region results presented in Section 4.5. We will only show the CvsB and CvsL shapes for the inclusive region. Still, the same methods apply to the signal region, only with the invariant mass constraint imposed.

Reducible background yield

The reducible background yield in the inclusive is estimated to be 19.85 events. Within the signal region, defined as m(H) $\in [115, 130]$, the corresponding yield is 2.91 events. No uncertainties are presented on the final yields. However, individual sources of uncertainty will be investigated and discussed in the following section.

Invariant mass spectrum

The invariant mass spectrum of the reducible background is shown in Figure 5.6. The choice of binning is practical such that no negative yields are observed across the spectrum, which also motivates the binning used previously in the 2P2F and 3P1F control regions. Only invariant masses

ranging from 70 GeV to 400 GeV are shown since the yield beyond those values is negligible. Motivated by the studies in Ref. [66], a fit to the Landau function is performed on the data to extract a continuous shape from the binned distribution. This fitted shape is then used as the final invariant mass distribution for the reducible background, scaled to match the expected yield reported previously.

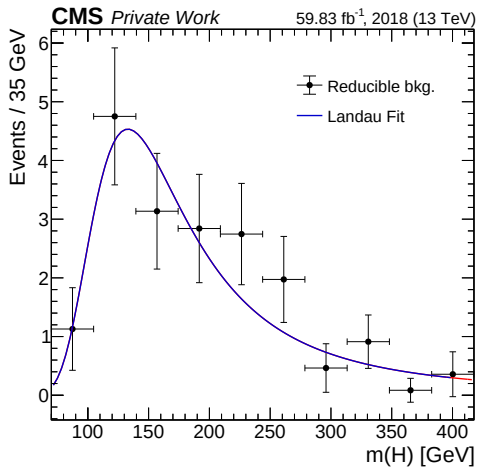

Figure 5.6: *The invariant mass distribution of the reconstructed Higgs boson candidate for the reducible background calculated according to Equation* (5.9)*. Only invariant masses ranging from* 70 GeV *to* 400 GeV *are shown since the yield beyond those values is negligible. The resulting distribution is fitted to a Landau distribution, for which the result is shown in blue.*

Jet Flavour Discriminator spectra

Figure 5.7 shows the jet flavour discriminator distributions for the reducible background. No analytical function is fitted to these distributions to extract a continuous shape. The outputs are derived from a complex neural network without a straightforward model. Investigating whether a reliable continuous parametrisation is possible can be explored in further studies.

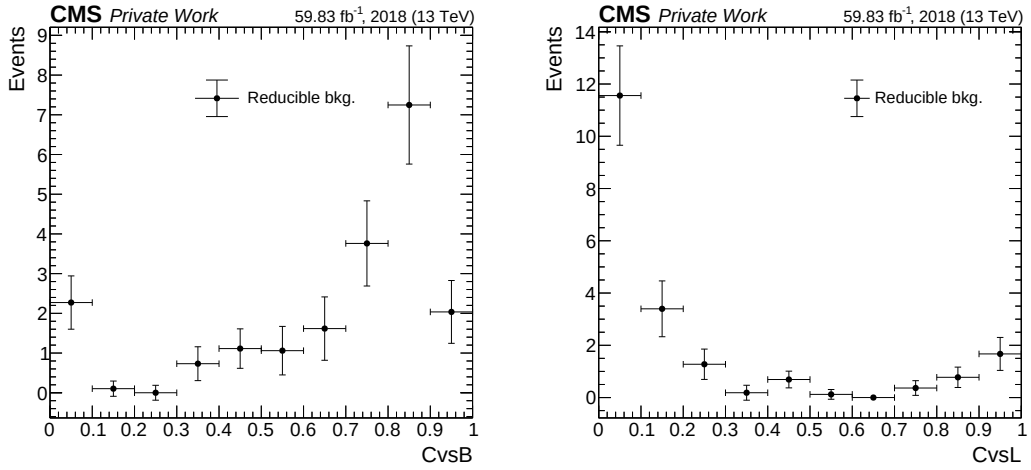

Figure 5.7: *The CvsB (left) and CvsL (right) jet flavour discriminator distributions of the reducible background in the inclusive region, calculated according to Equation* (5.9)*.*

5.2.4 Uncertainties on the reducible background estimation

While the method of reducible background estimation is data-driven to reduce systematic uncertainties related to modelling, it is still subject to other sources of uncertainty. Particularly, there are statistical uncertainties in the control region event samples, statistical uncertainties on the misidentification rate, and uncertainties related to the background composition. These are discussed in the following paragraphs.

Statistical uncertainties in the control regions

The total statistical uncertainty on the reducible background yield is approximately 12% in the inclusive region and 59% when restricting to contributions to the signal region. This reflects the limited number of events available in the control regions. In principle, statistical precision could be improved by loosening the muon selection criteria to increase event yields in the control regions. However, the subtractions performed during the method (Equation 5.9) do not guarantee a net gain in yield, especially since the misidentification rate also could increase. Optimising the object selection criteria to increase yields, especially for the signal region, can be investigated in future studies.

Misidentification rate variation uncertainty

The statistical uncertainties on the misidentification rate, visualised in Figure 5.2, represent a second source of uncertainty on the reducible background estimation. Any variation in the misidentification rate will directly affect the overall yield calculated via Equation (5.9). These statistical uncertainties are propagated as a systematic uncertainty to account for this. This involves varying the misidentification rate in each bin by its corresponding statistical uncertainty, shifting it up and down independently. For every such variation, the full reducible background yield is recalculated. With this procedure, the highest yield difference reported is 5%.

Background composition uncertainty

In the reducible background estimation, the implicit assumption was made that the background process composition is the same in both the region where the misidentification rate is measured (the Z $+\ell$ control region) and the regions where it is applied (the 2P2F and 3P1F control regions). However, this is incorrect. As previously shown (Figure 5.1), the Z$+\ell$ control region is dominated by the Drell-Yan process, while the $\mathrm{t\bar{t}}$ process was the dominating one in the 2P2F region (e.g., Figure 5.3). Different processes may yield different misidentification rates. For example, Drell-Yan processes produce fewer b jets than $\mathrm{t\bar{t}}$ processes, and b jets are more likely to fake leptons due to non-prompt leptons from semi-leptonic hadron decays. Applying a misidentification rate in a region where another process dominates may introduce a potential bias. This discrepancy in background composition is another source of uncertainty that needs to be accounted for.

To quantify this impact, a closer look is taken at the relative composition of background processes in the Z$+\ell$ control region and the 2P2F control region, as shown in Figures 5.8 (for Z + ℓ) and 5.9 (for 2P2F). Note that the prompt muon processes (WZ$\rightarrow 3\ell\nu$ and $\mathrm{qq} \rightarrow \mathrm{ZZ}, \mathrm{Z}\gamma^*$). These confirm that the composition of processes indeed shifts. Next, the effect of this change on the misidentification rate is estimated, via simulation only. To do this, misidentification rates are first measured separately for each background process. After that, an *average misidentification rate* is computed as the weighted sum based on the process fractions in the Z$+\ell$ region. To model the composition in the 2P2F region, each misidentification rate is reweighted according to the fractions observed in the 2P2F region, resulting in a *reweighted misidentification rate*. The results of the individual, averaged, and reweighted misidentification rates are shown in Figure 5.10.

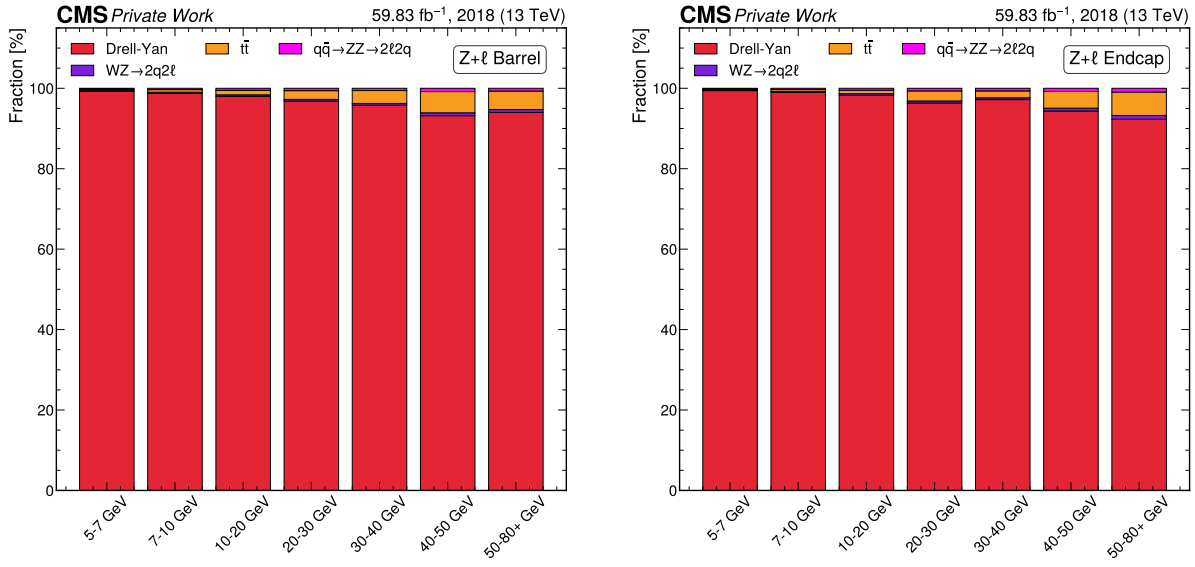

Figure 5.8: *The relative contribution of background processes in the Z+ℓ control region expressed in fractions. Results are shown separately for the barrel (left) and endcap (right) detector regions, using the same transverse momentum binning as for the calculation of the misidentification rate.*

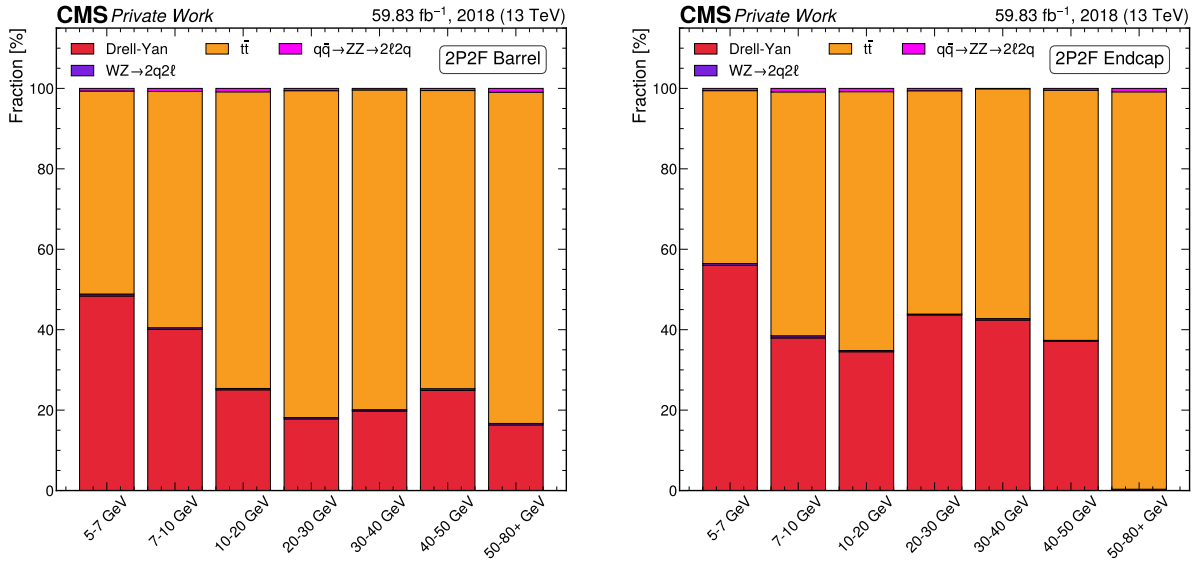

Figure 5.9: *The relative contribution of background processes in the 2P2F control region expressed in fractions. Results are shown separately for the barrel (left) and endcap (right) detector regions, using the same transverse momentum binning as for the calculation of the misidentification rate.*

The difference between the averaged and reweighted misidentification rate reflects the method's sensitivity to the relative background composition. Finally, this must be propagated as a yield uncertainty. This is done by shifting the misidentification rate measured in data by the relative difference, then recalculating the yield via Equation (5.9). This effect is reported to be much larger than the previous systematic uncertainty, resulting in a 36% increase.

To properly account for all these uncertainties, each source must be propagated into the final

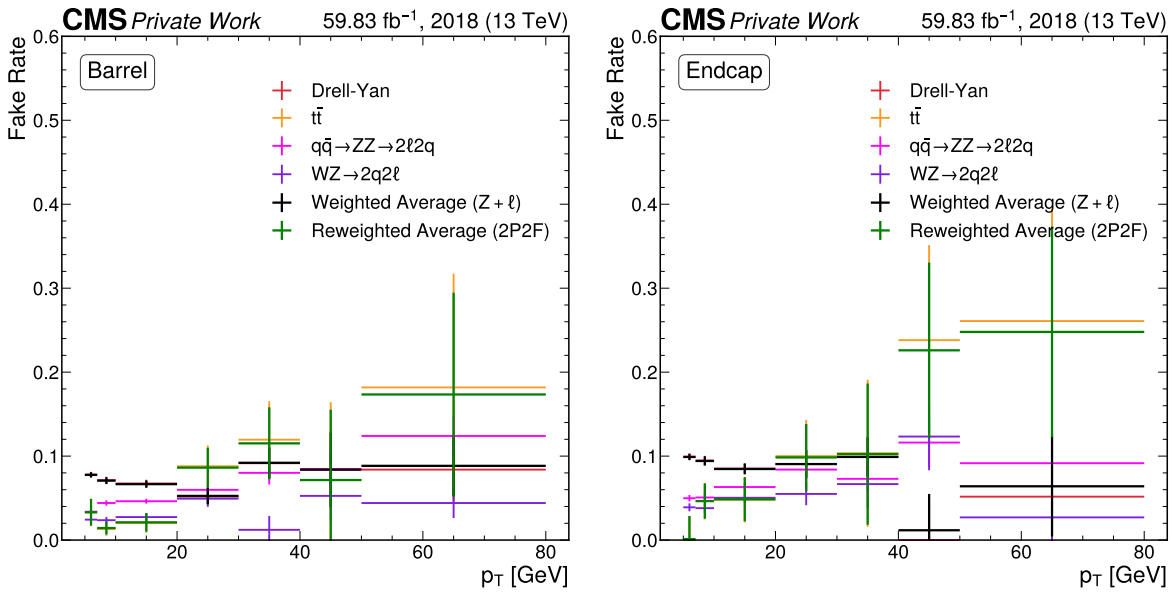

Figure 5.10: *Misidentification (fake) rates measured for each background process in simulation, along with the weighted average based on the composition in the Z+ℓ region and the reweighted average according to the background composition in the 2P2F region. Results are shown separately for the barrel (left) and endcap (right) detector regions.*

estimation. Only then is a complete characterisation of the total uncertainty achieved. However, such a detailed treatment lies beyond the scope of this thesis. Instead, the focus is on the nominal yields and the distribution shapes. These will be used in the following sections and in the final chapter, where the charm-Higgs Yukawa coupling is probed.

5.3 Inclusion of the reducible background

5.3.1 Comparison to data

This short section presents the combined result of the signal characterisation and both reducible and irreducible background estimations to conclude the chapter. The resulting distributions are given for the three representative variables throughout this and the previous chapter: the invariant mass of the reconstructed Higgs boson candidate m(H) and the jet flavour discriminator variables CvsB and CvsL. In this first subsection, the resulting distributions are compared to the data, which has been reconstructed via the procedures described in Chapter 4. In total, 287 events are observed.

The invariant mass spectrum of the reconstructed Higgs boson candidate is shown in the upper panel of Figure 5.11. Figure 5.12 shows the jet flavour discriminator distributions. The lower panel in each plot shows the Data / MC ratio. Considering the low statistics, an overall reasonable ratio is observed.

5.3.2 Signal region

In the signal region, where no data events are reconstructed, only the signal and background composition will be shown, accounting for the reducible backgrounds. The invariant mass spectrum of the reconstructed Higgs boson candidate is shown in the upper panel of Figure 5.13. In

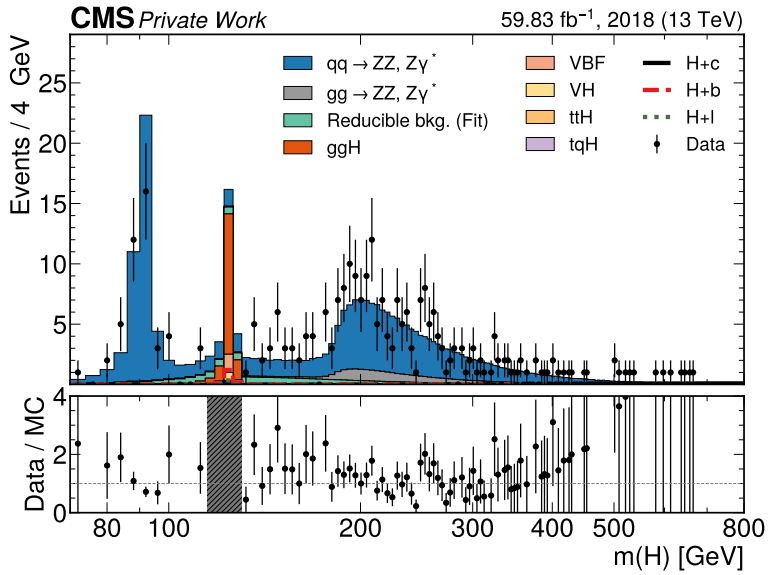

Figure 5.11: *Upper panel: The invariant mass of the reconstructed Higgs boson candidate m(H). The background prediction includes both irreducible and reducible components, and is compared to the data. Lower panel: The corresponding Data/MC ratio. The shaded region corresponds to the signal region, defined as events for which* $m(H) \in [115, 130]$ *GeV.*

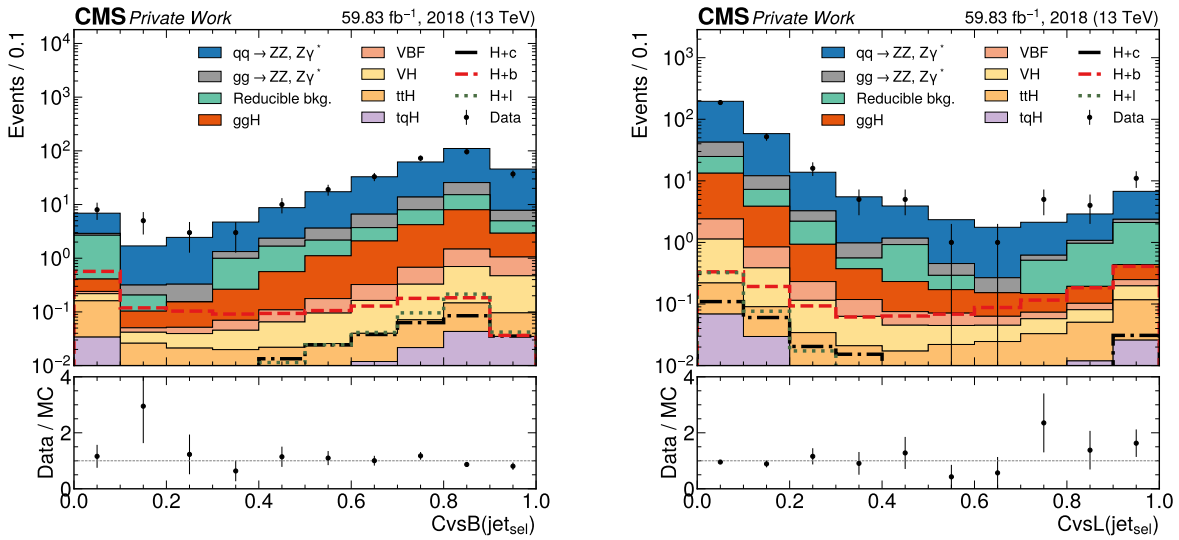

Figure 5.12: *Upper panels: The CvsB (left) and CvsL (right) jet flavour discriminator distribution for the selected jet. The background prediction includes both irreducible and reducible components, and is compared to the data. Lower panels: The corresponding Data / MC ratio.*

contrast, Figure 5.14 shows the jet flavour discriminator distributions. The lower panel in each plot shows the (S+B)/B ratio. These results effectively extend those from Section 4.5 and serve as a key input for the analysis in the final chapter.

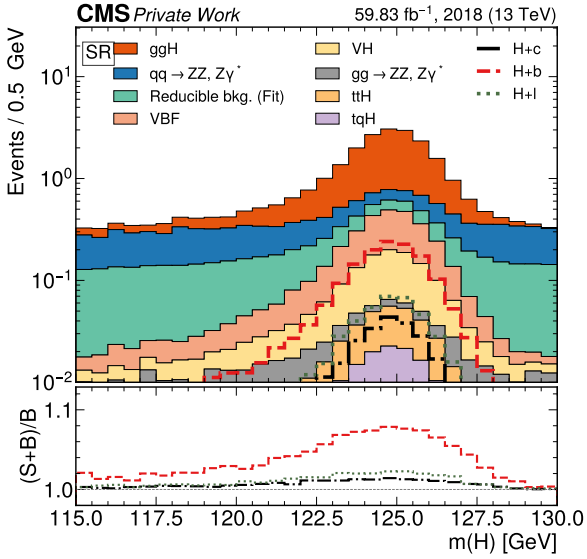

Figure 5.13: *Upper panel: The invariant mass of the reconstructed Higgs boson candidate m(H) in the signal region. The background prediction includes both irreducible and reducible components. Lower panel: The (S+B/B) ratio for each signal category in the H+jet sample.*

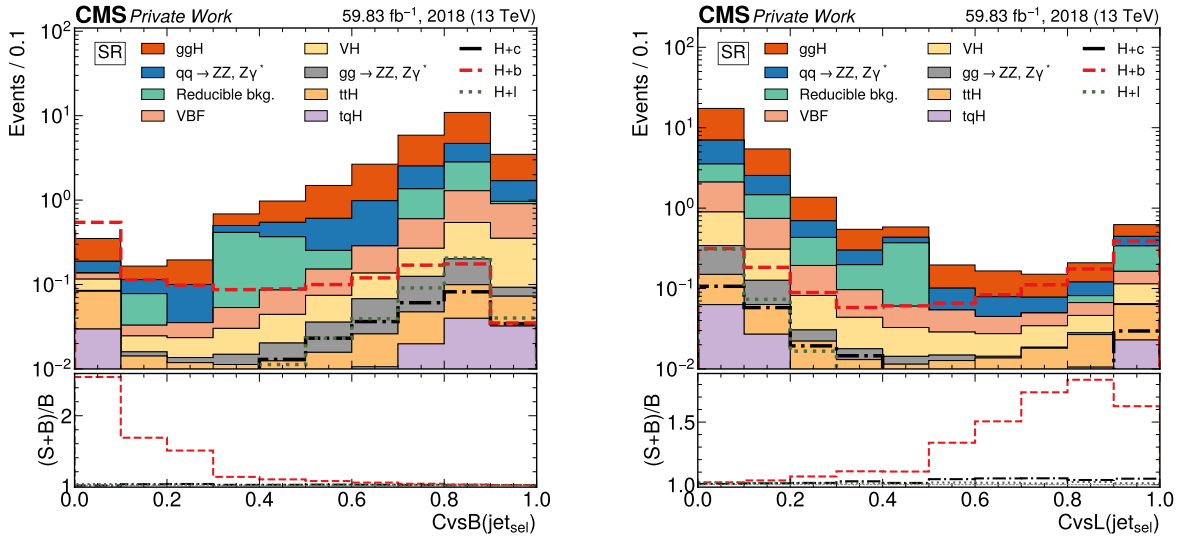

Figure 5.14: *Upper panel: The CvsB (left) and CvsL (right) jet flavour discriminator distribution for the selected jet in the signal region. The background prediction includes both irreducible and reducible components. Lower panel: The (S+B/B) ratio for each signal category in the H+jet sample.*

6 Measurement of the H+c production signal strength

In the closing chapter of the thesis, a simplified measurement of the H+c production signal strength is presented. Using the reconstructed signal and background distributions developed in earlier chapters, an expected upper limit on the H+c production strength $\mu_{\mathrm{H+c}}$ is extracted at 95% CL. In Section 6.1, the limit setting procedure is outlined, and the systematic uncertainties used in the extraction are listed. Last but not least, in Section 6.2, limits on the signal strength are presented using several observables individually, as well as for a two-dimensional unrolling technique that combines the distributions of both c-tagging discriminators.

6.1 Limit setting

6.1.1 Procedure

In this subsection, the general procedure for extracting an upper limit on the signal strength $\mu_{\mathrm{H+c}}$ is described. This signal strength acts as a scaling factor to what the SM predicts for the H+c production cross section. A value of $\mu_{\mathrm{H+c}} = 1$ corresponds to the SM expectation, while $\mu_{\mathrm{H+c}} = 0$ represents the background-only hypothesis. The inputs to this procedure are the binned distributions of relevant observables in the signal region, presented in Section 5.3. The goal is to set an upper limit on this signal strength, showing what the maximum contribution of the signal is to still be compatible with the observed (or expected) data at a certain confidence level under the assumption that no signal is present. This production cross section is directly sensitive to the charm-Higgs Yukawa coupling. At leading order, this dependence enters the cross section quadratically, meaning that any constraint on $\mu_{\mathrm{H+c}}$ can be translated into a constraint on the magnitude of y_c relative to the SM, which was previously defined as κ_c.

The central idea behind the upper limit extraction is statistically testing a background-only (b) hypothesis, H_0, against an alternative signal-plus-background hypothesis, H_1, that includes both the signal and background (s+b) contributions using a profile likelihood ratio test statistic via the $\mathrm{CL_s}$ method [69]. To achieve this, a statistical model must be considered that predicts a set of observations. Two kinds of parameters characterise any such model: the parameter of interest $\mu = \mu_{\mathrm{H+c}}$ and a set of nuisance parameters $\vec{\theta}$, which encode uncertainties in the model. In a binned analysis, such as this one, each bin i of an observable distribution is modelled as an independent Poisson counting experiment. Let n_i be the total number of events in the bin with $\nu_i(\mu, \vec{\theta}) = \mu s_i(\vec{\theta}) + b_i(\vec{\theta})$ as the number of expected events, where s_i and b_i represent the expected signal and background contributions, respectively. The likelihood of that bin can be written as:

$$L_i(\mu, \vec{\theta}) = \frac{(\nu_i(\mu, \vec{\theta}))^{n_i} e^{-\nu_i(\mu,\vec{\theta})}}{n_i!}. \tag{6.1}$$

By definition, the full likelihood $L(\mu, \vec{\theta})$ of the binned distribution is the product of Poisson likelihoods across all bins, combined with constraint terms $\pi_k(\theta_k)$ that result from external measure-

ments to constrain the nuisance parameters:

$$L(\mu, \vec{\theta}) = \prod_i L_i(\mu, \vec{\theta}) \cdot \prod_k \pi_k(\theta_k), \tag{6.2}$$

To compare between hypotheses, the *profile likelihood ratio* $\lambda(\mu)$ is defined:

$$\lambda(\mu) = \frac{L(\mu, \hat{\vec{\theta}}_\mu)}{L(\hat{\mu}, \hat{\vec{\theta}})}, \tag{6.3}$$

where $(\hat{\mu}, \hat{\vec{\theta}})$ are the values of the parameters that maximise the unconstrained likelihood, and $\hat{\vec{\theta}}_\mu$ maximises the likelihood under the constraint that the signal strength is fixed at μ. The latter procedure is known as *profiling*. The test statistic q_μ used for setting upper limits is:

$$q_\mu = \begin{cases} -2 \ln \lambda(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu, \end{cases} \tag{6.4}$$

The distribution of q_μ is used to derive p-values. Generating these distributions is typically computationally expensive, since no closed form exists in general. However, under the asymptotic approximation, analytical expressions do exist via Wilks' theorem [70]. From the observed test statistic, q_μ^{obs}, the p-values for the background-only and signal-plus-background hypotheses are given by:

$$p_b = \int_{q_\mu^{\text{obs}}}^{\infty} f(q_\mu|\mu = 0)\, dq \quad \text{and} \quad p_\mu = \int_{q_\mu^{\text{obs}}}^{\infty} f(q_\mu|\mu)\, dq, \tag{6.5}$$

Using these, the CL_s value is computed as:

$$\text{CL}_\text{s} = \frac{p_{\text{s+b}}}{1 - p_\text{b}} \tag{6.6}$$

An upper limit on the signal strength μ at 95% CL is calculated by scanning over μ until:

$$\text{CL}_\text{s} < 0.05. \tag{6.7}$$

This value is what will be quoted in the final result on the H+c production signal strength $\mu_{\text{H+c}}$. Since no data is used for this fit, an Asimov dataset is used, which allows for the calculation of expected upper limits assuming the SM expectation, and is constructed as the sum of all the processes. Practically, this limit extraction procedure is performed using the Combine tool [71].

6.1.2 Nuisance parameters

This subsection summarises the uncertainties incorporated into the statistical model as nuisance parameters. Each of these uncertainties describes an uncertainty on the yield of each specific process. Apart from statistical uncertainties, two sources of systematic uncertainties are also included. In what follows, a short description is given for each:

1. **Integrated luminosity:** The measurement of the integrated luminosity of the 2018 dataset has an uncertainty of 2.5% [72], and is applied to all MC processes.

2. **The $H \to ZZ^* \to 4\ell$ branching ratio:** The measurement of the $H \to ZZ^* \to 4\ell$ branching ratio has an uncertainty of 2% [73], which is applied to all processes involving a Higgs boson decay.

The above list of systematic uncertainties is not complete. The following non-exhaustive list presents other sources of systematic uncertainties that need to be accounted for in future analysis:

- **Reducible background:** The uncertainties regarding the reducible background estimation described in Section 5.2.4 must be applied.
- **Jet energy scale and resolution:** Since JES/JER corrections are applied to reconstructed jets, the uncertainties on these procedures must be evaluated according to the methods described in Section 3.3.3.
- **Scale factors:** In this analysis, only the c-tagging scale factors have been applied. However, the associated uncertainties have not been included. Other types of scale factors must also be applied. These include:
  - scale factors associated with muon identification and isolation criteria;
  - trigger efficiency scale factors.

Since not every systematic is accounted for, the extracted limit that will be presented can only be interpreted as an initial, order-of-magnitude indication of the sensitivity of the analysis.

6.2 Results

In this last section, results of the upper limit extraction on the H+c production signal strength $\mu_{\mathrm{H+c}}$ in the signal region are presented. First, a fit is performed using different individual variables to evaluate their sensitivity. These are:

- the invariant mass of the reconstructed Higgs boson candidate m(H), for which the distribution is shown in Figure 5.13,
- the jet flavour discriminator CvsB of the selected jet, for which the distribution is shown in Figure 5.14 (left),
- the jet flavour discriminator CvsL of the selected jet, for which the distribution is shown in Figure 5.14 (right).

For each of these distributions, only the H+c contribution is taken as signal, while the remaining processes are considered background. The resulting expected upper limits for each of these inputs are shown in Table 6.1. Among them, the CvsL variable provides the best sensitivity.

Since the flavour tagging information is critical and c-tagging relies on simultaneous rejection of udsg-jet and b-jet backgrounds via two discriminators, a two-dimensional distribution of these variables is expected to boost the sensitivity further. However, due to technical limitations, the Combine fitting framework that is used expects a one-dimensional distribution as input. As a workaround, the 2D histogram is transformed ("unrolled") into a 1D histogram. In this implementation, both the CvsB and CvsL axes are divided into N bins (Here, $N = 10$ is chosen). In each bin of CvsL, one considers the CvsB distribution. These CvsB distributions are then simply attached to each other, in order of increasing CvsL value, in a one-dimensional histogram. The distribution after unrolling the jet flavour discriminators is shown in Figure 6.1. The expected upper limit derived from this fit is:

$$\boxed{\mu_{\mathrm{H+c}} = 33.12 \text{ at } 95\% \text{ CL}}$$

This result demonstrates that an improvement in sensitivity on the signal strength is achieved when considering the information from both discriminators.

Fit variable(s)	Expected upper limit on $\mu_{\mathrm{H+c}}$
m(H)	42.62
CvsB	43.75
CvsL	39.63
Unrolled (CvsB $\oplus$ CvsL)	33.12

Table 6.1: *Expected 95% CL upper limits on the H+c signal strength $\mu_{\mathrm{H+c}}$.*

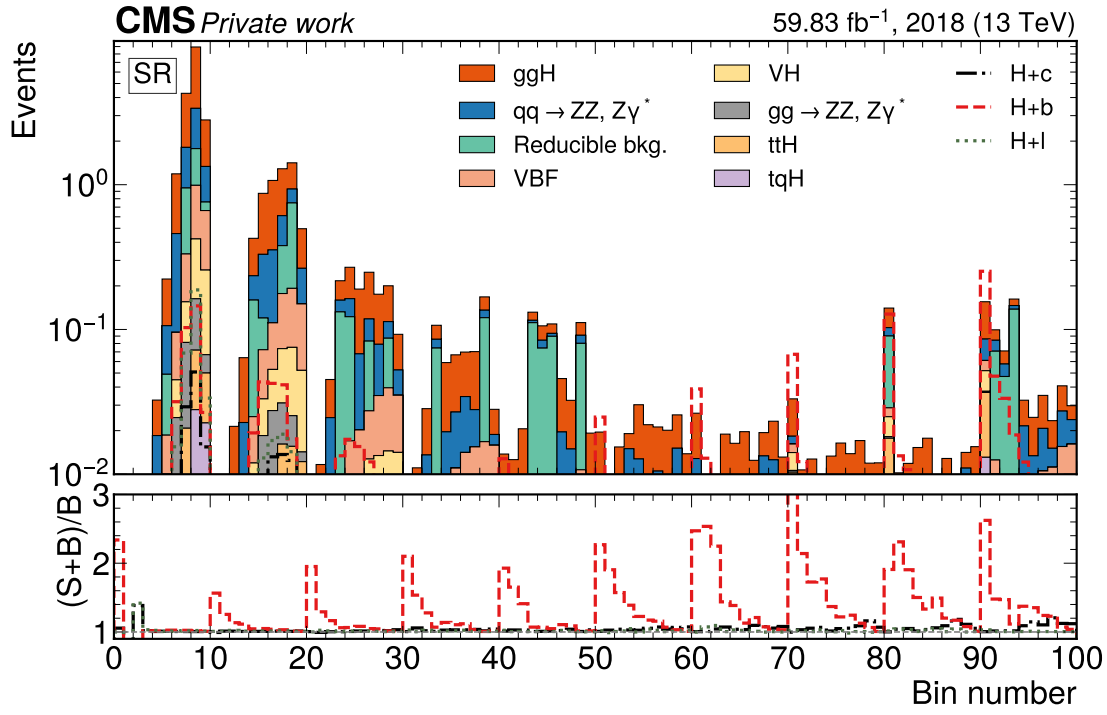

Figure 6.1: *Upper panel: The unrolled CvsB and CvsL histogram in the signal region. The first 10 bins represent the CvsB distribution in the first bin of the CvsL distribution. Bin numbers 11 to 20 represent the CvsB distribution in the second bin of the CvsL distribution. This trend continues. Lower panel: The (S+B/B) ratio for each signal category in the H+jet sample.*

Conclusion & Outlook

The Standard Model of particle physics (SM) encapsulates our understanding of the elementary particles and their fundamental interactions. A central test of the SM has been the validation of the couplings between the Higgs boson and the charged fermions, known as the Yukawa couplings. These parameters are expected to be proportional to fermion mass following the SM, and have been measured up to great precision for the muon and the third-generation fermions [17]. The next step in this program is the measurement of the charm-Higgs Yukawa coupling y_c. The most direct strategy for probing y_c is the $\mathrm{H} \to \mathrm{c\bar{c}}$ decay [20], but its sensitivity is mainly limited by the difficulty of distinguishing two charm jets. This motivates the exploration of other processes sensitive to the coupling.

One such process was investigated in this thesis: the associated production of a Higgs boson and a charm quark (H+c). This process benefits from being sensitive to the coupling at the production level, such that a decay channel can be freely chosen. Furthermore, only one jet needs to be identified. While some recent efforts have used this process with diphoton decays of the Higgs boson [22], this analysis offers a complementary approach by exploiting the signature of four isolated muons via the $\mathrm{H} \to \mathrm{ZZ}^* \to 4\mu$ decay channel.

The primary aim of this work was to derive a 95% CL upper limit on the signal strength $\mu_{\mathrm{H+c}}$, a parameter that quantifies the observed rate of H+c production relative to the SM expectation. A value of $\mu_{\mathrm{H+c}} = 1$ corresponds to the SM prediction, while $\mu_{\mathrm{H+c}} = 0$ denotes a background-only scenario. Since the H+c production cross section is sensitive to the charm-Higgs Yukawa coupling, this limit also constrains potential deviations from the SM production for this coupling. Achieving this goal, however, critically depends on the estimation of both the irreducible and reducible background processes and characterising the signal. These backgrounds refer to processes that mimic the signal by producing the same final state, particularly the four-muon signature. Irreducible backgrounds arise from processes that produce four prompt muons. On the other hand, reducible backgrounds originate from processes that do not produce four prompt muons, but where objects are misidentified as prompt muons, still leading to the signal signature. In particular, measuring the reducible backgrounds was a central focus of this thesis. The approach closely follows the reconstruction and background estimation techniques presented in Ref. [66], while adapting it specifically to the H+c topology. Notably, this thesis presents the first application of these techniques to proton-proton collision data recorded by the CMS detector in 2018, in the context of the H+c production process with the Higgs boson decaying via $\mathrm{H} \to \mathrm{ZZ}^* \to 4\mu$.

First, reconstruction techniques were developed to identify candidate H+c events, defined as those with a signature of four prompt muons accompanied by a jet. The H+c signal was characterised via a simulated event sample containing H+jet events. The irreducible background processes that mimic the signal signature were estimated using simulation. Second, the reducible backgrounds originating from misidentified and non-prompt muons were measured. This was achieved through a data-driven approach to mitigate the systematic uncertainties resulting from poor modelling in simulation. In particular, the method relied upon measuring the misidentification rate in a control region consisting of events with a Z boson and an additional probe muon, and subsequently applying it to control regions enriched in reducible background processes. This resulted in a total expected yield of 19.85 events, with 2.91 falling within the signal region, defined as the range where the invariant mass of the reconstructed Higgs boson candidate lies between

115 and 130 GeV. Additionally, the invariant mass of the reconstructed Higgs boson candidate, along with outputs of neural network-based algorithms that are optimised to discriminate charm jets from other flavour jets, was compared to data collected in 2018 by the CMS experiment. A good level of agreement was observed within statistical uncertainties.

Finally, using the signal and background distributions in the signal region, expected upper limits on $\mu_{\mathrm{H+c}}$ using the CL_s technique were derived. In a first step, fits were performed with different individual variables to evaluate their sensitivity. In a second step, the outputs of the neural network-based charm jet discriminators were combined into a two-dimensional unrolled histogram, achieving improved sensitivity compared to individual variables. The resulting expected upper limit was $\mu_{\mathrm{H+c}} = 33.12$ at 95% confidence level. This result served as proof of principle that the muon decay channel in the Higgs boson production associated with a charm quark can be used to constrain the H+c production process. While the sensitivity remains limited and not fully representable due to the low expected signal yields and the incomplete treatment of systematic uncertainties, the analysis does lay the groundwork for future studies.

Several improvements can be made to the four-muon final state analysis setup. The first is the incorporation of a complete set of systematic uncertainties. For example, individual sources of uncertainty related to the reducible background estimation were investigated, but their combined impact on the final upper limit was not included in the statistical model. Incorporating these uncertainties, as well as the others discussed in Chapter 6, would better reflect the true sensitivity of the analysis. Secondly, the background estimation strategy could be refined using multivariate techniques such as neural networks or boosted decision trees. With these tools, subtle differences in the kinematic properties of the signal and background can be explored. For example, VBF processes often produce jets highly forward in pseudorapidity, a characteristic that could be explored to design control regions.

Beyond improvements to the four-muon final state itself, the analysis could be expanded to include the other leptonic decay channels of the Higgs boson, namely the $2e2\mu$ and $4e$ final states. Additionally, the study can be extended to include data from other Run-2 years (2016 and 2017) and eventually Run-3. Whether these improvements will lead to a conclusive measurement of the charm-Higgs Yukawa coupling remains an open question. However, with the High-Luminosity LHC (HL-LHC) expected to begin operation in 2030 and planned upgrades to the CMS detector, only increased sensitivity is anticipated over the years.

Contributions by the Author

The following list summarises the authors contributions to the work:

- **Reconstruction techniques:** I contributed to the development of the H+jet reconstruction techniques outlined in Chapter 4, with a specific focus on implementing the selection for the 2018 dataset and handling the trigger requirements.
- **Reducible background estimation:** I designed and implemented the data-driven estimation of the reducible background presented in Chapter 5 following the methodology in [66], and extended the approach to the H+jet topology.
- **Limit derivation:** I performed the statistical analysis to derive the 95% confidence level upper limit on the signal strength $\mu_{\mathrm{H+c}}$, via the Combine tool [71] in Chapter 6. This included a first application of a two-dimensional unrolled histogram, combining the outputs of neural network-based algorithms optimised to discriminate charm jets from other flavour jets (CvsB and CvsL), enhancing the sensitivity to the measurement.

BIBLIOGRAPHY

[1] David J Griffiths. *Introduction to elementary particles; 2nd rev. version*. Physics textbook. New York, NY: Wiley, 2008. URL: https://cds.cern.ch/record/111880.

[2] Michael H. Riordan. "The Discovery of Quarks". In: *Science* 256.5061 (May 29, 1992), pp. 1287–1293. DOI: 10.1126/science.256.5061.1287.

[3] Mark Thomson. *Modern particle physics*. New York: Cambridge University Press, 2013. ISBN: 978-1-107-03426-6. DOI: 10.1017/CBO9781139525367.

[4] F. Mandl and G. Shaw. *Quantum field theory*. 2nd. Hoboken, N.J: Wiley, 2010. ISBN: 978-0-471-49683-0.

[5] *The Standard Model*. 2024. URL: https://www.damtp.cam.ac.uk/user/tong/sm/standardmodel.pdf.

[6] Y. Fukuda et al. "Evidence for Oscillation of Atmospheric Neutrinos". In: *Phys. Rev. Lett.* 81 (8 Aug. 1998), pp. 1562–1567. DOI: 10.1103/PhysRevLett.81.1562. URL: https://link.aps.org/doi/10.1103/PhysRevLett.81.1562.

[7] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons". In: *Phys. Rev. Lett.* 13 (9 Aug. 1964), pp. 321–323. DOI: 10.1103/PhysRevLett.13.321. URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.321.

[8] Peter W. Higgs. "Broken Symmetries and the Masses of Gauge Bosons". In: *Phys. Rev. Lett.* 13 (16 Oct. 1964), pp. 508–509. DOI: 10.1103/PhysRevLett.13.508. URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.508.

[9] G. Aad et al. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Physics Letters B* 716.1 (Sept. 2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. URL: https://doi.org/10.1016%2Fj.physletb.2012.08.020.

[10] S. Chatrchyan et al. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Physics Letters B* 716.1 (Sept. 2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. URL: https://doi.org/10.1016%2Fj.physletb.2012.08.021.

[11] Sheldon L. Glashow. "Partial-symmetries of weak interactions". In: *Nuclear Physics* 22.4 (1961), pp. 579–588. ISSN: 0029-5582. DOI: https://doi.org/10.1016/0029-5582(61)90469-2. URL: https://www.sciencedirect.com/science/article/pii/0029558261904692.

[12] Steven Weinberg. "A Model of Leptons". In: *Phys. Rev. Lett.* 19 (21 1967), pp. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264. URL: https://link.aps.org/doi/10.1103/PhysRevLett.19.1264.

[13] Abdus Salam. "Weak and Electromagnetic Interactions". In: *Conf. Proc. C* 680519 (1968), pp. 367–377. DOI: 10.1142/9789812795915_0034.

[14] S. Navas et al. "Review of particle physics". In: *Phys. Rev. D* 110.3 (2024), p. 030001. DOI: 10.1103/PhysRevD.110.030001.

[15] Jeffrey Goldstone, Abdus Salam, and Steven Weinberg. "Broken Symmetries". In: *Phys. Rev.* 127 (3 1962), pp. 965–970. DOI: 10.1103/PhysRev.127.965. URL: https://link.aps.org/doi/10.1103/PhysRev.127.965.

[16] C T Potter et al. *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties: Report of the LHC Higgs Cross Section Working Group*. en. 2013. DOI: 10.5170/CERN-2013-004. URL: http://cds.cern.ch/record/1559921.

[17] A. M. Sirunyan et al. "Combined measurements of Higgs boson couplings in proton - proton collisions at $\sqrt{s} = 13\,\text{TeV}$". In: *The European Physical Journal C* 79.5 (2019). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-019-6909-y. URL: http://dx.doi.org/10.1140/epjc/s10052-019-6909-y.

[18] A. M. Sirunyan et al. "Observation of $t\bar{t}H$ Production". In: *Phys. Rev. Lett.* 120 (23 2018), p. 231801. DOI: 10.1103/PhysRevLett.120.231801. URL: https://link.aps.org/doi/10.1103/PhysRevLett.120.231801.

[19] Albert M Sirunyan et al. "Evidence for Higgs boson decay to a pair of muons". In: *JHEP* 01 (2021), p. 148. DOI: 10.1007/JHEP01(2021)148. arXiv: 2009.04363 [hep-ex].

[20] Armen Tumasyan et al. "Search for Higgs Boson Decay to a Charm Quark-Antiquark Pair in Proton-Proton Collisions at s=13 TeV". In: *Phys. Rev. Lett.* 131.6 (2023), p. 061801. DOI: 10.1103/PhysRevLett.131.061801. arXiv: 2205.05550 [hep-ex].

[21] Ilaria Brivio, Florian Goertz, and Gino Isidori. "Probing the Charm Quark Yukawa Coupling in Higgs + Charm Production". In: *Physical Review Letters* 115.21 (Nov. 2015). ISSN: 1079-7114. DOI: 10.1103/physrevlett.115.211801. URL: http://dx.doi.org/10.1103/PhysRevLett.115.211801.

[22] Vladimir Chekhovsky et al. "Search for the associated production of a Higgs boson with a charm quark in the diphoton decay channel in pp collisions at $\sqrt{s}$ = 13 TeV". In: (Mar. 2025). arXiv: 2503.08797 [hep-ex].

[23] Lyndon Evans. *The large hadron collider : a marvel of technology*. Jan. 2009. URL: https://ci.nii.ac.jp/ncid/BA91712857.

[24] R. Barate et al. "Search for the standard model Higgs boson at LEP". In: *Phys. Lett. B* 565 (2003), pp. 61–75. DOI: 10.1016/S0370-2693(03)00614-2. arXiv: hep-ex/0306033.

[25] Benjamin W. Lee, C. Quigg, and H. B. Thacker. "Weak interactions at very high energies: The role of the Higgs-boson mass". In: *Phys. Rev. D* 16 (5 Sept. 1977), pp. 1519–1531. DOI: 10.1103/PhysRevD.16.1519. URL: https://link.aps.org/doi/10.1103/PhysRevD.16.1519.

[26] Werner Herr and B Muratori. "Concept of luminosity". In: (2006). DOI: 10.5170/CERN-2006-002.361. URL: https://cds.cern.ch/record/941318.

[27] Oliver Sim Brüning et al. *LHC Design Report*. CERN Yellow Reports: Monographs. Geneva: CERN, 2004. DOI: 10.5170/CERN-2004-003-V-1. URL: https://cds.cern.ch/record/782076.

[28] CMS Collaboration. *CMS Luminosity - Public Results*. Nov. 2024. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[29] J. T. Boyd. *LHC Run-2 and Future Prospects*. 2020. arXiv: 2001.04370 [hep-ex]. URL: https://arxiv.org/abs/2001.04370.

[30] Aram Hayrapetyan et al. "Development of the CMS detector for the CERN LHC Run 3". In: *JINST* 19.05 (2024), P05064. DOI: 10.1088/1748-0221/19/05/P05064. arXiv: 2309.05466 [physics.ins-det].

[31] Dhananjay Saikumar. "Future of computing at the Large Hadron Collider". In: (Sept. 2022). DOI: 10.48550/arXiv.2210.13213.

[32] O. Aberle et al. *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report*. CERN Yellow Reports: Monographs. Geneva: CERN, 2020. DOI: 10.23731/CYRM-2020-0010. URL: https://cds.cern.ch/record/2749422.

[33] *The HL-LHC project | High Luminosity LHC Project*. Jan. 2025. URL: https://hilumilhc.web.cern.ch/content/hl-lhc-project.

[34] Frank Zimmermann et al. "The Electron-Positron Future Circular Collider (FCC-ee)". In: *JACoW* NAPAC2022 (2022), pp. 315–320. DOI: 10.18429/JACoW-NAPAC2022-TUZD1.

[35] Philip Bambade et al. "The International Linear Collider: A Global Project". In: (Mar. 2019). arXiv: 1903.01629 [hep-ex].

[36] Sergey Chatrchyan et al. "The CMS experiment at the CERN LHC". In: *Journal of Instrumentation* 3 (Aug. 2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.

[37] Serguei Chatrchyan et al. "Missing transverse energy performance of the CMS detector". In: *JINST* 6 (2011), P09001. DOI: 10.1088/1748-0221/6/09/P09001. arXiv: 1106.5048 [physics.ins-det].

[38] David Barney. "Presentation for public - Introduction to CMS for CERN guides". In: (2013). URL: https://cds.cern.ch/record/2629323.

[39] V Karimäki et al. *The CMS tracker system project: Technical Design Report*. Technical design report. CMS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/368412.

[40] *The CMS electromagnetic calorimeter project: Technical Design Report*. Technical design report. CMS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/349375.

[41] *The CMS hadron calorimeter project: Technical Design Report*. Technical design report. CMS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/357153.

[42] *The Phase-2 Upgrade of the CMS Muon Detectors*. Tech. rep. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2283189.

[43] V. Khachatryan et al. "The CMS trigger system". In: *Journal of Instrumentation* 12.01 (Jan. 2017), P01020–P01020. ISSN: 1748-0221. DOI: 10.1088/1748-0221/12/01/p01020. URL: http://dx.doi.org/10.1088/1748-0221/12/01/P01020.

[44] Michael H. Seymour and Marilyn Marx. *Monte Carlo Event Generators*. 2013. arXiv: 1304.6677 [hep-ph]. URL: https://arxiv.org/abs/1304.6677.

[45] M. A. Dobbs et al. *Les Houches Guidebook to Monte Carlo Generators for Hadron Collider Physics*. 2004. arXiv: hep-ph/0403045 [hep-ph]. URL: https://arxiv.org/abs/hep-ph/0403045.

[46] J. Alwall et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". In: *Journal of High Energy Physics* 2014.7 (July 2014). ISSN: 1029-8479. DOI: 10.1007/jhep07(2014)079. URL: http://dx.doi.org/10.1007/JHEP07(2014)079.

[47] Stefano Frixione, Paolo Nason, and Carlo Oleari. "Matching NLO QCD computations with parton shower simulations: the POWHEG method". In: *Journal of High Energy Physics* 2007.11 (Nov. 2007), pp. 070–070. ISSN: 1029-8479. DOI: 10.1088/1126-6708/2007/11/070. URL: http://dx.doi.org/10.1088/1126-6708/2007/11/070.

[48] Yuri L. Dokshitzer. "Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics." In: *Sov. Phys. JETP* 46 (1977), pp. 641–653.

[49] Torbjörn Sjöstrand et al. "An introduction to PYTHIA 8.2". In: *Computer Physics Communications* 191 (June 2015), pp. 159–177. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2015.01.024. URL: http://dx.doi.org/10.1016/j.cpc.2015.01.024.

[50] Bo Andersson. *The Lund Model*. Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology. Cambridge University Press, 2023.

[51] S. Agostinelli et al. "Geant4 a simulation toolkit". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(03)01368-8. URL: https://www.sciencedirect.com/science/article/pii/S0168900203013688.

[52] A. M. Sirunyan et al. "Particle-flow reconstruction and global event description with the CMS detector". In: *JINST* 12.10 (2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003. arXiv: 1706.04965 [physics.ins-det].

[53] Wolfgang Adam et al. *Track Reconstruction in the CMS tracker*. Tech. rep. Geneva: CERN, 2006. URL: https://cds.cern.ch/record/934067.

[54] R. E. Kalman. "A New Approach to Linear Filtering and Prediction Problems". In: *Journal of Basic Engineering* 82.1 (Mar. 1960), pp. 35–45. ISSN: 0021-9223. DOI: 10.1115/1.3662552. URL: https://doi.org/10.1115/1.3662552.

[55] A. M. Sirunyan et al. "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *JINST* 13.06 (2018), P06015. DOI: 10.1088/1748-0221/13/06/P06015. arXiv: 1804.04528 [physics.ins-det].

[56] *Particle-flow commissioning with muons and electrons from J/Psi and W events at 7 TeV*. Tech. rep. Geneva: CERN, 2010. URL: https://cds.cern.ch/record/1279347.

[57] Serguei Chatrchyan et al. "Performance of CMS Muon Reconstruction in *pp* Collision Events at $\sqrt{s} = 7$ TeV". In: *JINST* 7 (2012), P10002. DOI: 10.1088/1748-0221/7/10/P10002. arXiv: 1206.4071 [physics.ins-det].

[58] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. "The anti-kt jet clustering algorithm". In: *Journal of High Energy Physics* 2008.04 (Apr. 2008), pp. 063–063. ISSN: 1029-8479. DOI: 10.1088/1126-6708/2008/04/063. URL: http://dx.doi.org/10.1088/1126-6708/2008/04/063.

[59] Vardan Khachatryan et al. "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV". In: *JINST* 12.02 (2017), P02014. DOI: 10.1088/1748-0221/12/02/P02014. arXiv: 1607.03663 [hep-ex].

[60] Matteo Cacciari and Gavin P. Salam. "Pileup subtraction using jet areas". In: *Phys. Lett. B* 659 (2008), pp. 119–126. DOI: 10.1016/j.physletb.2007.09.077. arXiv: 0707.1378 [hep-ph].

[61] The CMS collaboration. "Determination of jet energy calibration and transverse momentum resolution in CMS". In: *Journal of Instrumentation* 6.11 (Nov. 2011), P11002–P11002. ISSN: 1748-0221. DOI: 10.1088/1748-0221/6/11/p11002. URL: http://dx.doi.org/10.1088/1748-0221/6/11/P11002.

[62] A. M. Sirunyan et al. "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV". In: *JINST* 13.05 (2018), P05011. DOI: 10.1088/1748-0221/13/05/P05011. arXiv: 1712.07158 [physics.ins-det].

[63] E. Bols et al. "Jet flavour classification using DeepJet". In: *Journal of Instrumentation* 15.12 (Dec. 2020), P12012–P12012. ISSN: 1748-0221. DOI: 10.1088/1748-0221/15/12/p12012. URL: http://dx.doi.org/10.1088/1748-0221/15/12/P12012.

[64] Armen Tumasyan et al. "A new calibration method for charm jet identification validated with proton-proton collision events at $\sqrt{s}$ =13 TeV". In: *JINST* 17.03 (2022), P03014. DOI: 10.1088/1748-0221/17/03/P03014. arXiv: 2111.03027 [hep-ex].

[65] Konstantin Lehmann and Bernd Stelzer. "The Fake Factor Method and its relation to the Matrix Method". In: *Nucl. Instrum. Meth. A* 1054 (2023), p. 168376. DOI: 10.1016/j.nima.2023.168376.

[66] Albert M Sirunyan et al. *Measurements of production cross sections of the Higgs boson in the four-lepton final state in protonproton collisions at* $\sqrt{s} = 13\,TeV$. Tech. rep. 6. 2021, p. 488. DOI: 10.1140/epjc/s10052-021-09200-x. arXiv: 2103.04956 [hep-ex].

[67] Albert M Sirunyan et al. "Pileup mitigation at CMS in 13 TeV data". In: *JINST* 15.09 (2020), P09018. DOI: 10.1088/1748-0221/15/09/P09018. arXiv: 2003.00503 [hep-ex].

[68] F. Heyen. *Private communication*. 2025.

[69] A L Read. "Presentation of search results: the CLs technique". In: *Journal of Physics G: Nuclear and Particle Physics* 28.10 (Sept. 2002), p. 2693. DOI: 10.1088/0954-3899/28/10/313. URL: https://dx.doi.org/10.1088/0954-3899/28/10/313.

[70] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". In: *The European Physical Journal C* 71.2 (Feb. 2011). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-011-1554-0. URL: http://dx.doi.org/10.1140/epjc/s10052-011-1554-0.

[71] Aram Hayrapetyan et al. "The CMS Statistical Analysis and Combination Tool: Combine". In: *Comput. Softw. Big Sci.* 8.1 (2024), p. 19. DOI: 10.1007/s41781-024-00121-4. arXiv: 2404.06614 [physics.data-an].

[72] *CMS luminosity measurement for the 2018 data-taking period at* $\sqrt{s} = 13\,\mathrm{TeV}$. Tech. rep. Geneva: CERN, 2019. URL: https://cds.cern.ch/record/2676164.

[73] CERN. *CERN Yellow Reports: Monographs, Vol 2 (2017): Handbook of LHC Higgs cross sections: 4. Deciphering the nature of the Higgs sector*. en. 2017. DOI: 10.23731/CYRM-2017-002. URL: https://e-publishing.cern.ch/index.php/CYRM/issue/view/32.