UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

Dipartimento Interateneo di Fisica "M. Merlin"

Dottorato di Ricerca in Fisica
XXXVII Ciclo
Settore Scientifico Disciplinare Fis/01

Search for $H \to c\bar{c}$ at CMS in VBF production with Run-3 data

Supervisors:
Prof. Anna Colaleo
Prof. Rosamaria Venditti

Candidate:
Angela Zaza

Coordinator:
Prof. Domenico Di Bari

Contents

List of Figures

List of Tables

Introduction

The discovery of the Higgs boson in 2012 by the ATLAS [1] and CMS [2] experiments at the Large Hadron Collider (LHC) [3] represents a milestone in the history of high energy physics. It confirmed the spontaneous symmetry breaking mechanism predicted by the Standard Model (SM), which is currently the best theory to describe fundamental particles and their interactions [4].
Over a decade since the discovery, many properties of the Higgs boson-its mass, its width, its coupling with the weak bosons and third generation fermions-have been measured with very good precision. Quite recently, also evidence for the Higgs decay into a pair of muons has been achieved. All these measurements align with the SM theory.
However, the SM has known limitations and many beyond standard model (BSM) theories have been formulated in order to address these. In this context, it is extremely important to continue probing the Higgs boson properties with increasing precision. Any discrepancy between experimental results and SM predictions could reveal signs of new physics, offering valuable insights into the validity of specific beyond standard model theories.
The next step in this probing campaign will be the measurement of the Higgs boson couplings to the second generation quarks, and thus to the charm quarks. This is currently one of the highest priority goals of the CMS Collaboration.

For long time, it has been thought that the observation of the direct decay of the Higgs boson into a charm quark-antiquark pair is out of reach at the current experiments because of the small branching ratio ($\sim$3%) and the large impact of the quantum chromodynamics (QCD) multi-jet background. However, new sophisticated analysis methods developed very recently have enhanced the sensitivity to this search, making it crucial to explore this channel already within the physics program of the LHC proton-proton collisions. Moreover, some BSM models allow deviations of the Higgs coupling to charm quarks with respect to the SM prediction, at a scale that is observable at the present day LHC experiments.

A first search for $H \to c\bar{c}$ with the Higgs boson produced in association with a vector boson (VH) has been conducted by the CMS and ATLAS collaborations with Run-2 data, leading to set an observed (expected) upper limit of respectively 11.5 (10.5) [5] and 14 (7.6) [6] on the signal strength.
In this work, for the first time the vector boson fusion (VBF) production mechanism, which is the one with the second highest cross section at the center of mass energy of LHC, is investigated for the first time.
The final state topology of the VBF production is a powerful handle to suppress the QCD multi-jet background, as demonstrated by the results reached in the $H \to b\bar{b}$ and $H \to \mu\mu$ searches.

Before the Run-3 startup, the feasibility of the $VBF\ H \to c\bar{c}$ search with new incoming CMS data was affected by two challenges. First of all, there was no trigger suitable for this search. Secondly, the amount of QCD background due to the misidentification of light jet with c jets seems to be too much high to make this search competitive. I addressed the first point by developing a dedicated trigger that exploits the VBF event signature. The second point was addressed by the development reached in the context of the CMS Collaboration for the jet flavour tagging thanks to the extension of the ParticleNet algorithm to the AK4 jets. The search for $VBF\ H \to c\bar{c}$ presented in this work is performed on the data collected in 2023 with this novel trigger. The search includes a multivariate analysis (MVA) algorithm for the reduction of the QCD background. The signal and the peaking backgrounds are modelled from Monte Carlo simulations, while the QCD background is extracted with a data driven technique. Finally, an expected upper limit on the signal strength is estimated at 95% confidence level by performing a simultaneous binned maximum likelihood fit to the reconstructed mass of the Higgs boson candidate.
This search is very new in CMS and the whole analysis chain, starting from the trigger itself, is an original contribution of the author.

Since the analysis is still under review by the CMS Collaboration, data in the mass range close to the Higgs boson nominal mass are blinded, a standard practice adopted in order to prevent the anlyzers from introducing a bias in the signal extraction. For this reason, the upper limit quoted in this thesis is the one expected from MC simulation of SM processes.

This thesis is structured in four chapters:

- **Chapter 1: The Standard Model**
  In the first chapter, an overview of the Standard Model is provided, with

a focus on the spontaneous symmetry breaking and the Higgs boson physics. The state-of-the-art results of the Higgs boson properties are summarized and the status of the search for $H \to c\bar{c}$ is described. Then, an introduction to the Standard Model Effective Field Theory (SMEFT) is provided, being one of the most reliable methods adopted to extend the SM while searching for discrepancies. Finally, an example of BSM theory that allows for a Higgs boson coupling larger than the one predicted by the SM is provided.

- **Chapter 2: The LHC and the CMS experiment**
  The second chapter is focused on the LHC accelerator and the CMS experiment. The main characteristics of the detector apparatus are described together with an overview of the Run-3 data taking conditions relevant for this analysis.

- **Chapter 3: Object reconstruction at the CMS experiment**
  In this chapter, the algorithms used for the reconstruction and identification of the main objects used in this search are described (tracks, vertices, jets). A particular emphasis is dedicated to the latest developments in the jet heavy flavour identification algorithms.

- **Chapter 4: Search for $VBF\ H \to c\bar{c}$**
  The last chapter provides a detailed description of the whole analysis chain developed for the signal search. The first part is focused on the trigger development and the study of its performance during the data taking. Then the offline selection, which includes a MVA algorithm for signal versus background discrimination, is described. Finally, the statistical analysis implemented for the extraction of the signal and the consequent estimation of the expected upper limit on the signal strength is presented.

Chapter 1

The Standard Model

The Standard Model [4] of Particle Physics is a relativistic quantum field theory that explains how the building blocks of matter interact under the effect of three Fundamental Forces: Electromagnetic, Weak and Strong Interactions. A significant milestone in confirming the Standard Model, finalized in the mid-1970s, was the discovery of the Higgs Boson at the Large Hadron Collider in 2012. Despite its successes, the Standard Model has limitations, notably its exclusion of gravitational interaction. Consequently, new theories are needed to address phenomena that the Standard Model does not yet explain. This chapter will delve into the fundamentals of the Standard Model.

1.1 Elementary Particles

The Standard Model provides a systematic classification of elementary particles, the fundamental constituents of matter. They are categorized, according to their spin, into two main groups: fermions and bosons. Fermions, with half-integer spin, obey to Fermi-Dirac statistics, while bosons, with integer spin, follow Bose-Einstein statistics.

- Fermions: semi-integer spin particles.
  Fermions can be further classified in two groups: quarks and leptons, each comprising six particles and their corresponding antiparticles. Fermions show to be related in pairs, called "families" or "generations":
  - Leptons: $(e, \nu_e), (\mu, \nu_\mu), (\tau, \nu_\tau)$
  - Quarks: $(u, d), (c, s), (t, b)$

  Only fermions from the first generations, which are the lightest ones, compose ordinary matter.

- Bosons: integer spin particles.
  Nature is governed by four fundamental forces: electromagnetic, weak, strong and gravitational interactions. Bosons are the carriers of fundamental interactions and particles interact with each other by exchanging bosons:

  – the photon (γ) is the carrier of the electromagnetic force,

  – the Z and $W^{\pm}$ bosons are the carriers of the weak interaction,

  – gluons g are the carriers of the strong interaction.

  The gravitational force is hypotesized to be mediated by a yet-to-be-discovered boson called the graviton.
  Additionally, the Higgs Boson, which is not associated with any fundamental force, is the carrier of the Higgs Field, that permeates the universe, providing mass to all the particles interacting with it.

Figure 1.1 provides a comprehensive scheme of the fundamental particles of the SM.

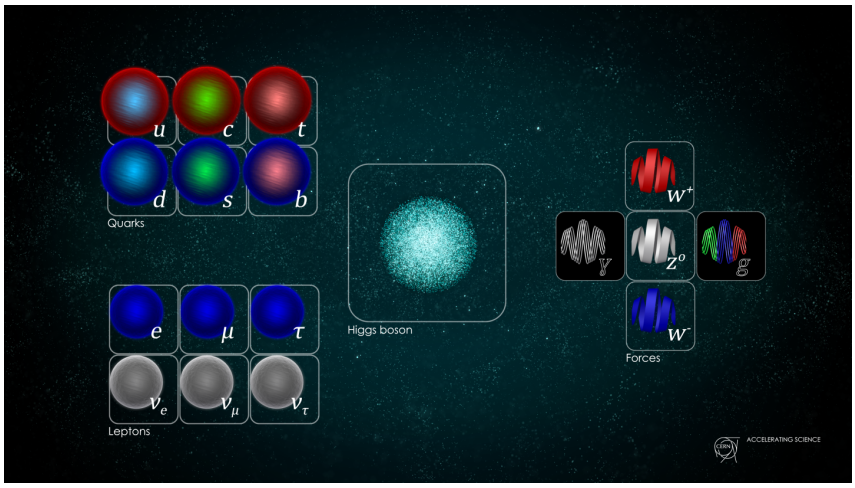


Figure 1.1: SM classification of Elementary Particles [7]. On the left side of the figure, quarks (top) and leptons (bottom) are organized in generations (columns). At the center of the picture, the Higgs Boson is represented. On the right side, the other bosons, carriers of fundamental forces, are shown.

1.2 Gauge Symmetries

In addition to being a relativistic quantum field theory, the Standard Model is a gauge theory, meaning that its Lagrangian density is invariant under gauge

transformations [8]. This condition is strictly related to the experimental observation of certain symmetries in Particle Physics. According to the Noether's theorem, the invariance of the action under a certain transformation is associated to a conservation law. A simple example of this theorem is given by the energy conservation law: the energy of a system is conserved if there is time-translation symmetry.
The Lagrangian density of the Standard Model is constructed in order to be invariant under the fundamental symmetry group:

$$\mathscr{F} = SU(3) \times SU(2)_L \times U(1)_Y \tag{1.1}$$

$SU(3)$ is the symmetry group of Quantum Chromodynamics, while $SU(2)_L \times U(1)_Y$ is the symmetry group of Electroweak Theory.

1.2.1 Quantum Chromodynamics (QCD)

Quantum Chromodynamics (QCD) is the gauge field theory that describes the strong interaction between quarks and gluons. The equivalent of the electric charge for strong interactions is the *color*. Quarks occur in three different colors (green, red or blue) and combine together to form hadrons. There are two types of hadrons: mesons, made of two quarks, and baryons, made of three quarks.
Both mesons and baryons are colorless, or, equivalently, they are color singlet states. In mathematical language, this means that they are invariant under $SU(3)$ rotations in color space. As a consequence, the symmetry group of QCD is $SU(3)$, called color group.

In order to obtain the Lagrangian density for QCD, the Lagrangian density for free quarks is introduced:

$$\mathscr{L}_{free} = \sum_{j=1}^{f} \bar{q}^j (i\gamma^\lambda \partial_\lambda - m_j) q^j, \tag{1.2}$$

where j is the index over the possible flavours, and $q^j = \begin{pmatrix} q_1^j \\ q_2^j \\ q_3^j \end{pmatrix}$ is the field describing a quark of flavour j and accounts for the three colors.
Given U, a transformation of $SU(3)$, it can be written as a function of the group generators T_k with $k = 1, .., 8$:

$$U = 1 + i \sum_{k=1}^{8} \delta\varphi^k T_k, \tag{1.3}$$

where $\delta\varphi^k$ are infinitesimal parameters and T_k is defined as:

$$T_k = \frac{\lambda^k}{2} \tag{1.4}$$

In the previous equation, λ^k are the Gell-Mann matrices.
In general, the number of generators of a symmetry group $SU(N)$ is $N^2 - 1$ and the group algebra is given by:

$$[T_a, T_b] = if^{abc}T^c, \tag{1.5}$$

where f^{abc} (called structure constants of the $SU(3)$ group) are totally antisymmetric.
The Lagrangian density introduced for free quarks 1.2 results to be invariant under a global transformation (constant matrix) U of $SU(3)$:

$$q^j(x) \to U \cdot q^j(x) \tag{1.6}$$

By constrast, it is not invariant under a local transformation $U(x)$ of $SU(3)$:

$$q^j(x) \to U(x) \cdot q^j(x) \tag{1.7}$$

$$\begin{aligned} \mathscr{L}_{free} &= \sum_{j=1}^{f} \bar{q}^j(i\gamma^\lambda\partial_\lambda - m_j)q^j \\ &\to \sum_{j=1}^{f} \bar{q}^j(i\gamma^\lambda\partial_\lambda - m_j + i\gamma^\lambda U^\dagger(x)\partial_\lambda U(x))q^j \end{aligned} \tag{1.8}$$

The gauge principle of chromodynamics, and thus the invariance under local transformations in color space, requires the introduction of 8 $(N^2 - 1)$ vector fields $G_\lambda(x)$, called gluons:

$$\begin{aligned} G_\lambda(x) =& G^a_\lambda(x)\frac{\lambda_a}{2} = G^\dagger_\lambda(x), \;\; a = 1, .., 8, \\ & TrG_\lambda(x) = 0 \end{aligned} \tag{1.9}$$

The Lagrangian has to be modified in order to take into account the interaction between quarks and gluons. At this scope, on the basis of the "minimal coupling" in electrodynamics, all ordinary derivatives ∂_μ are replaced by:

$$\partial_\mu \to D_\mu = \partial_\mu + ig_sG_\lambda(x) \tag{1.10}$$

where D_μ is called covariant derivative and g_s is a dimensionless coupling constant. Consequently, the Lagrangian density becomes:

$$\mathscr{L}_q(x) = \sum_{j=1}^{f} \bar{q}^j(x)(i\gamma^\lambda D_\lambda - m_j)q^j(x) \tag{1.11}$$

This Lagrangian is invariant under local transformations.
However, it is not complete, since it does not contain the gluon dynamics term. In order to construct this term, again on the basis of electrodynamics, a gluon field-strength tensor $G_{\lambda\rho}(x)$ is introduced:

$$G_{\lambda\rho}(x) = \partial_\lambda G_\rho(x) - \partial_\rho G_\lambda(x) + ig_s[G_\lambda(x), G_\rho(x)] \tag{1.12}$$

Its components $G^a_{\lambda\rho}$ are defined by:

$$G_{\lambda\rho}(x) = G^a_{\lambda\rho}\frac{\lambda_a}{2} \tag{1.13}$$

The quadratic term in 1.12 is typical of non-Abelian groups as $SU(3)$.
At this point, the Lagrangian density for QCD can be written:

$$\mathscr{L}_{QCD}(x) = -\frac{1}{4}G^a_{\lambda\rho}(x)G^{\lambda\rho a}(x) + \sum_{j=1}^{f}\bar{q}^j(x)(i\gamma^\lambda D_\lambda - m_j)q^j(x) \tag{1.14}$$

In conclusion, thanks to the application of the gauge principle, the QCD theory can be written as a function of a single parameter g_s and predicts the existence of 8 gluons.

1.2.2 Electroweak theory

The Electroweak theory describes electromagnetic and weak interactions in a single unified model. The gauge group of the electroweak interaction is: $SU(2)_L \times U(1)_Y$, where $SU(2)_L$ is the *weak isospin group* and $U(1)_Y$ is the *weak hypercharge group*.

The three generators of $SU(2)_L$ are:

$$T^a = \frac{\tau^a}{2}, \tag{1.15}$$

where τ^a , a=1, ..., 3, are the Pauli matrices.

It is assumed, according to experimental observations, that the left-handed components of a leptonic family form a doublet of $SU(2)_L$, while the right-handed component of the charged lepton behaves like a singlet of $SU(2)_L$, meaning that it is invariant under a $SU(2)_L$ transformation. For example, for the electronic family, we have: $\begin{pmatrix}\nu_e\\ e\end{pmatrix}_L$ and e_R.

They can be arranged into a unique spinor: $\psi = \begin{pmatrix}\nu_{eL}\\ e_L\\ e_R\end{pmatrix}$.

Under a local transformation of $U(1)_Y$ the spinor transforms as follows:

$$\psi \to e^{i\chi(x)Y}\psi \tag{1.16}$$

where Y is the matrix:

$$Y = \begin{pmatrix} y_L & 0 & 0 \\ 0 & y_L & 0 \\ 0 & 0 & y_R \end{pmatrix} \tag{1.17}$$

The parameter y_L is conventionally set to $-\frac{1}{2}$ and the value of y_R is set to -1 in order to have the correct form of the electromagnetic coupling.

As in the case of QCD, the Lagrangian density is constructed in order to be invariant under local transformations of $SU(2)_L \times U(1)_Y$:

$$\mathscr{L} = -\frac{1}{2}Tr(W_{\lambda\rho}(x)W^{\lambda\rho}(x)) - \frac{1}{4}B_{\lambda\rho}(x)B^{\lambda\rho}(x) + \bar{\psi}i\gamma^{\lambda}D_{\lambda}\psi(x) \tag{1.18}$$

D_λ is the covariant derivative, defined as:

$$D_\lambda\psi(x) = (\partial_\lambda + igW^a_\lambda(x)T_a + ig'B_\lambda(x)Y)\psi(x), \tag{1.19}$$

where g and g' are the gauge coupling constants, respectively for $SU(2)_L$ and $U(1)_Y$. $W^a_\lambda(x)$, with a=1, ..., 3, are the vector fields corresponding to $SU(2)_L$ and B_λ is the vector field corresponding to $U(1)_Y$. The matrices T_a (a=1,...,3) have the form

$$T_a = \begin{pmatrix} \frac{1}{2}\tau_a & 0 \\ 0 & 0 \end{pmatrix} \tag{1.20}$$

and they form, together with the hypercharge matrix Y, a representation of the generators of the group $SU(2)_L \times U(1)_Y$.

Finally, the strength tensors $W_{\lambda\rho}$ and $B_{\lambda\rho}$ are defined as follows:

$$W_{\lambda\rho}(x) = \partial_\lambda W_\rho(x) - \partial_\rho W_\lambda(x) + ig[W_\lambda(x), W_\rho(x)] \tag{1.21}$$

$$B_{\lambda\rho}(x) = \partial_\lambda B_\rho - \partial_\rho B_\lambda \tag{1.22}$$

The fields that describe the vector bosons $W^\pm_\lambda$ and Z_λ, carriers of the weak interaction, and the photon A_λ, carrier of the electromagnetic interaction, can be derived from the vector fields W^1_λ, W^2_λ, W^3_λ and B_λ, obtained from the application of the gauge principle to the electroweak interaction.
The charged boson fields are defined as:

$$W^\pm_\lambda = \frac{1}{\sqrt{2}}(W^1_\lambda \mp iW^2_\lambda) \tag{1.23}$$

while, the neutral boson fields are given by:

$$\begin{aligned} Z_\lambda &= cos\theta_W W_\lambda^3 - sin\theta_W B_\lambda \\ A_\lambda &= sin\theta_W W_\lambda^3 + cos\theta_W B_\lambda \end{aligned} \tag{1.24}$$

where θ_W is the *weak angle* (Glashow, 1961), function of the coupling constants g and g':

$$\begin{aligned} sin\theta_W &= \frac{g'}{\sqrt{g^2 + g'^2}} \\ cos\theta_W &= \frac{g}{\sqrt{g^2 + g'^2}} \end{aligned} \tag{1.25}$$

1.2.3 Spontaneous Symmetry Breaking

The lagrangian density introduced by Equation 1.18 does not contain a term accounting for the mass of the vector bosons, which would have the form $m_W^2 W_\lambda^+ W^{\lambda -}$, and it is not possible to include such a term, because it would break the gauge invariance. However, according to experimental observations, the vector bosons $W^\pm$ and Z, carriers of the weak interaction, have a non zero mass.

Within the SM, the mechanism that gives mass not only to the massive gauge bosons, but also to fermions, is called "Higgs mechanism" and was hypotesized independently by Higgs [9],[10], Englert and Brout [11]. The Higgs mechanism incorporates the *Spontaneous Symmetry Breaking* (SSB), discussed by Weinberg and Salam, in a gauge invariant field theory.

In general, a SSB occurs when the Lagrangian of a system is symmetric under a certain transformation, while the ground state of the system is not. In order to make a simple example, let us consider a point particle moving along an axis ρ in a one-dimensional potential $V(\rho)$ given by:

$$V(\rho) = -\frac{1}{2}\mu^2\rho^2 + \frac{1}{4}\lambda\rho^4 \tag{1.26}$$

where $\mu^2 > 0$ and $\lambda > 0$ are fixed constants. This potential, represented in Figure 1.2, is symmetric under the transformation: $\rho \to -\rho$. The equilibrium positions of the system, obtained by requiring $\frac{\partial V}{\partial \rho} = 0$, correspond to $\rho_0 = \pm\sqrt{\frac{\mu^2}{\lambda}}$ and are not symmetric under the exchange of ρ and $-\rho$. As a consequence, the symmetry is spontaneously broken.

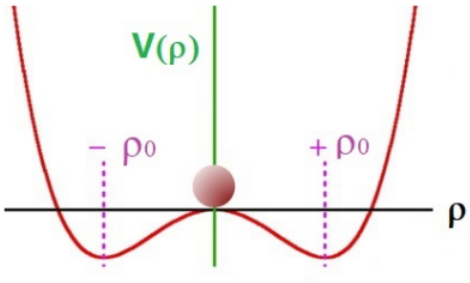


Figure 1.2: An example of one-dimensional potential $V(\rho)$ that shows spontaneous symmetry breaking.

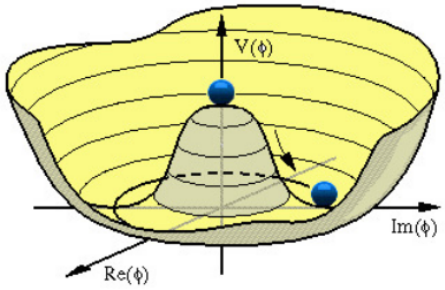


Figure 1.3: Representation of the Higgs potential as function of the real and imaginary parts of the field ϕ.

The mechanism suggested by Weinberg and Salam requires additional scalar fields, called *Higgs fields*.
Let us introduce two complex scalar fields ϕ_1 and ϕ_2, which form a doublet of $SU(2)_L$:

$$\phi(x) = \begin{pmatrix} \phi_1(x) \\ \phi_2(x) \end{pmatrix} = \begin{pmatrix} \Re\phi_1(x) + i\Im\phi_1(x) \\ \Re\phi_2(x) + i\Im\phi_2(x) \end{pmatrix} \tag{1.27}$$

Under a transformation of $SU(2)_L$, $\phi(x)$ transforms as follows:

$$\phi(x) \to U(x)\phi(x), \tag{1.28}$$

while, under a transformation of $U(1)_Y$, it becomes:

$$\phi(x) \to e^{iy_H\chi(x)}\phi(x), \tag{1.29}$$

where y_H is the weak hypercharge of the doublet and is equal to $\frac{1}{2}$. Again, the Lagrangian is constructed in order to be invariant under gauge transformations:

$$\mathscr{L}_H = (D_\lambda\phi)^\dagger(D^\lambda\phi) - V(\phi^\dagger\phi), \tag{1.30}$$

where D_λ is the covariant derivative, defined as:

$$D_\lambda = \partial_\lambda + igW_\lambda + ig'B_\lambda y_H \tag{1.31}$$

and $V(\phi^\dagger\phi)$ must be chosen in order to lead to a spontaneous symmetry breaking.
At this purpose, let us consider

$$V(\phi^\dagger\phi) = \frac{k}{2}\rho^2 + \frac{\lambda}{4}\rho^4 \tag{1.32}$$

whith $k < 0$ and $\lambda > 0$. This potential, represented in Figure 1.3, has two stable equilibrium positions, corresponding to $\rho_0 = \pm\sqrt{\frac{-k}{\lambda}}$, and an unstable position: $\rho = 0$. The ground state is given by:

$$\phi = e^{i\alpha^a T^a} \begin{pmatrix} 0 \\ \frac{\rho_0}{\sqrt{2}} \end{pmatrix}, \quad a = 1, ..., 3 \tag{1.33}$$

where T^a are the generators of $SU(2)_L$. Since the coefficients α_a vary with continuity, there is an infinite number of ground states. The gauge symmetry is broken when one of the infinite configurations of the ground state is chosen, for example: $\phi_0 = \begin{pmatrix} 0 \\ \frac{\rho_0}{\sqrt{2}} \end{pmatrix}$.
If ϕ_0 is substituted to ϕ in the first term of Lagrangian equation 1.30, we obtain the term:

$$(D_\lambda \phi_0)(D^\lambda \phi_0) = \frac{g^2 \rho_0^2}{4} W_\lambda^- W^{\lambda +} + \frac{(g^2 + g'^2)}{4} \rho_0^2 \frac{Z_\lambda Z^\lambda}{2} \tag{1.34}$$

from which the vector boson masses can be derived:

$$m_W^2 = \frac{g^2 \rho_0^2}{4} = \frac{e^2 \rho_0^2}{4\, sin^2\theta_W}, \qquad m_Z^2 = \frac{(g^2 + g'^2)\rho_0^2}{4} = \frac{e^2 \rho_0^2}{4\, sin^2\theta_W\, cos^2\theta_W} \tag{1.35}$$

The masses m_W and m_Z have been measured at LEP 2[12], Tevatron[13] and by the ATLAS experiment at LHC[14] and the world average values are here reported [4]:

$$m_W = 80.3692 \pm 0.0133 \text{ GeV}, \qquad m_Z = 91.1880 \pm 0.0020 \text{ GeV}, \tag{1.36}$$

The combination of the m_W measurements reported above does not include the Collider Detector at Fermilab (CDF) Collaboration result: $m_W = 80.433 \pm 0.0094$ GeV [15], which differs by 7σ from the SM prediction. Very recently, the CMS Collaboration published the result of the m_W measurement with data collected in 2016 [16]:

$$m_W = 80.3602 \pm 0.0099 \text{ GeV} \tag{1.37}$$

This result is particularly important, since it is characterized by a much better precision than the others used for the world average estimation (excluding CDF) and it is consistent with the SM prediction.

Let us consider a state $\phi'(x)$ that corresponds to a fluctuation around the the ground state:

$$\phi'(x) = \begin{pmatrix} 0 \\ \frac{\rho_0}{\sqrt{2}}(1 + \frac{\rho'(x)}{\rho_0}) \end{pmatrix} \tag{1.38}$$

By evaluating the Lagrangian density in this state, we obtain the term describing the interaction between the Higgs boson (H), carrier of the Higgs field $\rho'(x)$, and the weak vector bosons:

$$\begin{aligned}(D_\lambda\phi'(x))(D^\lambda\phi'(x)) =& m_W^2(1+2\frac{\rho'(x)}{\rho_0}+\frac{\rho'(x)^2}{\rho_0^2})W_\lambda^+W^{\lambda-} \\ &+\frac{1}{2}m_Z^2(1+2\frac{\rho'(x)}{\rho_0}+\frac{\rho'(x)^2}{\rho_0^2})Z_\lambda Z^\lambda\end{aligned} \tag{1.39}$$

From equation 1.39, the Higgs boson couplings to W^+W^- and ZZ bosons result to be:

$$g_{HZZ}=\frac{2}{\rho_0}m_Z^2, \qquad g_{HWW}=\frac{2}{\rho_0}m_W^2 \tag{1.40}$$

It can be seen that the Higgs boson couplings to the gauge bosons are proportional to the square of the boson masses. Moreover, by substituting $\phi'(x)$ in the expression for $V(\phi(x)^\dagger\phi(x))$, it results:

$$V(\phi'(x)^\dagger\phi'(x))=\lambda\rho_0^2\rho'^2+\frac{1}{2}\lambda\rho_0\rho'^3+\frac{1}{4}\lambda\rho'^4 \tag{1.41}$$

From the first term of this sum, the expression for the mass of the Higgs boson can be obtained:

$$m_H^2=2\lambda\rho_0^2 \tag{1.42}$$

λ corresponds to the Higgs self coupling constant and is a free parameter of the theory. Therefore, there is no a priori prediction for the Higgs mass.
The second and the third terms of the sum provide the expressions respectively for the trilinear (λ_3) and quadrilinear (λ_4) self couplings. Within the SM, we have:

$$\lambda_3=\lambda_4=\frac{m_H^2}{2\rho_0^2} \tag{1.43}$$

The value of ρ_0 (vacuum expectation value), which is fixed by the Fermi coupling constant G_F, is:

$$\rho_0=(\sqrt{2}G_F)^{-\frac{1}{2}}=246\ \text{GeV} \tag{1.44}$$

So far, it has been shown as, through the spontaneous symmetry breaking mechanism, the weak bosons acquire mass, while the photon remains massless. Following a similar development, also the fermion mass terms can be introduced. For this aim, let us introduce the Yukawa Lagrangian that describes the interaction between the Higgs boson and fermions. In the case of the electron, it has the form:

$$\mathscr{L}_{Yukawa}=-c_e\bar{e}_R\phi^\dagger\begin{pmatrix}\nu_{eL}\\ e_L\end{pmatrix}+h.c. \tag{1.45}$$

where c_e is a coupling constant (Yukawa coupling). The mass of the electron results to be:

$$m_e = -c_e \frac{\rho_0}{\sqrt{2}}, \tag{1.46}$$

while the neutrino is massless. For the other leptonic families, the Yukawa Lagrangian has an analogous expression. For what concerns the quarks, the weak isospin eigenstates (referred to with the corresponding primed letter) are different from the mass eigenstates, and the relations between them are defined by the $Cabibbo - Kobayashi - Maskawa\, matrix$ (V_{CKM}):

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = V_{CKM} \begin{pmatrix} d \\ s \\ b \end{pmatrix} \tag{1.47}$$

In the case of the first quark generation, the Yukawa Lagrangian is defined as:

$$\mathscr{L}_{Yukawa} = -c'_q \bar{u}_R \phi^T \epsilon \begin{pmatrix} u \\ d' \end{pmatrix}_L - c_q \bar{d'}_R \phi^\dagger \begin{pmatrix} u \\ d' \end{pmatrix}_L + h.c. \tag{1.48}$$

where $\epsilon = i\tau^2$. In a similar way, it can be derived for the other quark generations. From equations 1.45 and 1.48, it can be observed that the Higgs boson couplings to fundamental fermions are linearly proportional to the fermion masses. Moreover, it is important to highlight that the SSB mechanism does not explain the large variety of mass values of the fermions (flavour hierarchy) and the fermion masses, which represent a large number of the free parameters of the SM, are translated into Yukawa couplings (equation 1.46) [4]. Therefore, it is extremely important to investigate with the highest accuracy the Higgs boson coupling with fermions.
The field $\phi'(x)$ is constructed in order to be invariant under $SU(3)$ transformations. Indeed, the Higgs boson does not interact with gluons, according to their massless nature.

Finally, the final expression of the Standard Model Lagrangian can be written:

$$\begin{aligned} \mathscr{L}_{SM} = &-\frac{1}{4} G_{\lambda\rho} G^{\lambda\rho} - \frac{1}{4} W_{\lambda\rho} W^{\lambda\rho} - \frac{1}{4} B_{\lambda\rho} B^{\lambda\rho} + \bar{\psi} i \gamma_\lambda D^\lambda \psi \\ &+ (D_\lambda \phi)^\dagger (D^\lambda \phi) - V(\phi) + \mathscr{L}_{Yukawa} + h.c. \end{aligned} \tag{1.49}$$

where the covariant derivative D_λ has the form:

$$D_\lambda = \partial_\lambda + i g_s G_\lambda + i g W_\lambda + i g' B_\lambda Y \tag{1.50}$$

1.3 Status of Higgs Boson Physics at LHC

In 2012, the ATLAS[1] and CMS[2] experiments at LHC discovered a new resonance with a mass of approximately 125 GeV, and the subsequent studies of its properties proved the compatibility of the measured resonance with the Higgs boson theorized by the the Standard Model[4].
The SM Higgs boson is a CP-even scalar of spin 0. The most precise mass measurement has been obtained recently by the ATLAS collaboration, by combining the results in the diphoton and the four-lepton channels from Run-1 and Run-2 [17] [4]:

$$m_H = 125.11 \pm 0.09\ (stat) \pm 0.06\ (syst)\ \text{GeV} \tag{1.51}$$

By measuring the Higgs boson mass, it is possible to predict the value of the Higgs self coupling constant λ, according to equation 1.42: $\lambda \approx 0.13$.

As it has been observed in the previous Section (Section 1.2.3), the Higgs boson couplings to fundamental particles are set by their masses. As a consequence, the dominant Higgs production and decay processes involve the coupling of H to the weak gauge bosons and to the third generation of quarks and leptons.

Production mechanisms at LHC

The dominant Higgs production mechanisms at LHC, and in general at hadron colliders, are gluon fusion (ggF), weak-boson fusion (VBF), associated production with a gauge boson (VH), and associated production with a pair of $t\bar{t}$ quarks ($t\bar{t}H$) or with a single top quark (tHq). The corresponding Feynman diagrams are shown in Figure 1.4.
Figure 1.5 (Left) shows the cross sections for the production of a SM Higgs boson as a function of the center of mass energy ($\sqrt{s}$), for proton-proton (pp) collisions.
At the center of mass energy between 13 and 14 TeV, that is the configuration of the data analyzed in this work, the Higgs boson is mostly produced through the gluon-fusion mechanism, which is mediated by the exchange of a virtual top quark (Figure 1.4 (a)). The second-largest cross section production mechanism is the vector boson fusion, which is mediated by t- or u-channel exchange of a W or a Z boson, with the Higgs boson radiated off the weak-boson propagator (Figure 1.4 (b)). Then, following the decrasing order of the cross section, the most dominant production processes are associated production with W and Z gauge bosons (Figure 1.4 (c)) and the associated production with $t\bar{t}$ (Figure 1.4 (e)) and $b\bar{b}$. Finally, the lowest cross section Higgs production mechanism

is the production in association with a single top quark (Figure 1.4 (f-g)), wich can be studied in order to obtain important information, e.g. the sign of the top Yukawa coupling.

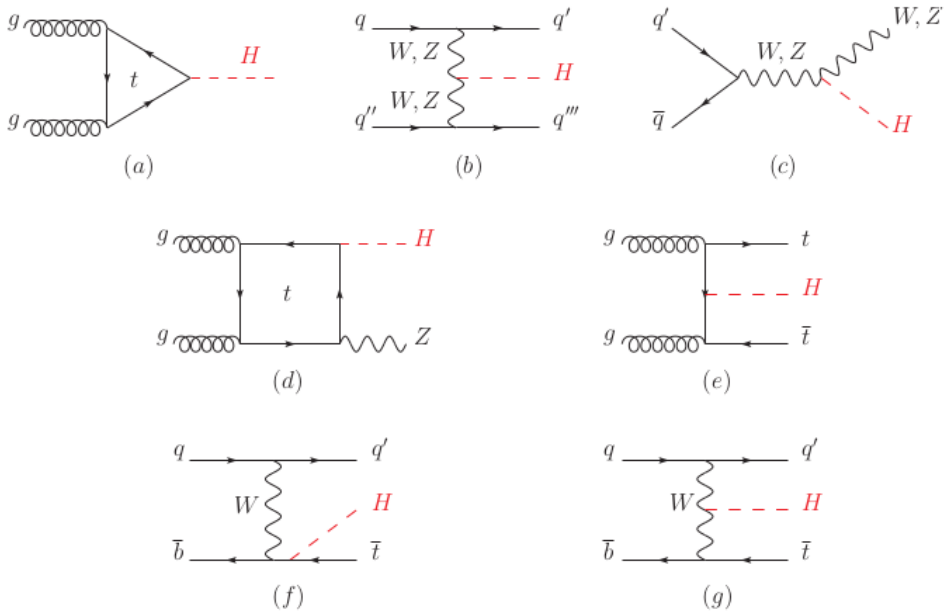


Figure 1.4: Main leading order Feynman diagrams contributing to the Higgs boson production in (s) gluon fusion, (b) Vector-boson fusion,(c) Higgs-strahlung (or associated production with a gauge boson at tree level from a quark-quark interaction), (d) associated production with a gauge boson (at loop level from a gluon-gluon interaction), (e) associated production with a pair of top quarks, (f-g) production in association with a single top quark [4].

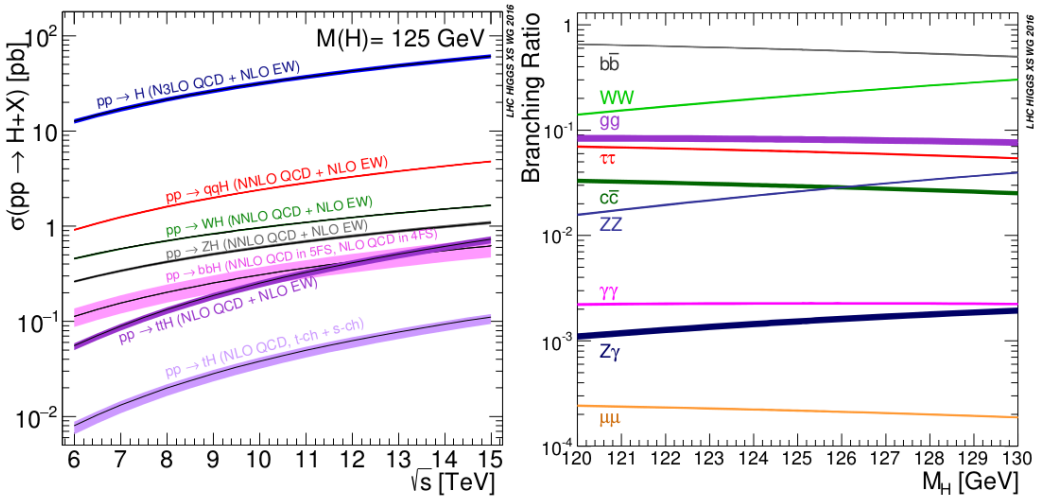


Figure 1.5: (Left) The SM Higgs boson production cross sections as function of the center of mass energy ($\sqrt{s}$). (Right) The branching ratios for the main decays of the SM Higgs boson near $m_H = 125$ GeV. The theoretical uncertainties are indicated as bands [4].

Table 1.1 summarizes the theoretical cross sections of the most dominant Higgs boson production processes in pp collisions as function of $\sqrt{s}$ for $m_H = 125$ GeV. The relative uncertainties, arising for instance from higher-order perturbative QCD corrections, theoretical uncertainties on Parton Distribution Functions (PDF) and α_S, EWK corrections, are also reported.

Table 1.1: The SM Higgs production cross sections for $m_H = 125$ GeV in pp collisions as function of $\sqrt{s}$ [4]. The uncertainties are estimated assuming no correlation between α_S and PDF uncertainties.

$\sqrt{s}$ (TeV)	Production cross section (in pb) for $m_H = 125$ GeV					
	ggF	VBF	WH	ZH	$t\bar{t}H$	total
1.96	$0.95^{+17\%}_{-17\%}$	$0.065^{+8\%}_{-7\%}$	$0.13^{+8\%}_{-8\%}$	$0.079^{+8\%}_{-8\%}$	$0.004^{+10\%}_{-10\%}$	$1.23^{+15\%}_{-15\%}$
7	$16.9^{+5.5\%}_{-7.6\%}$	$1.24^{+2.2\%}_{-2.2\%}$	$0.58^{+2.2\%}_{-2.3\%}$	$0.34^{+3.1\%}_{-3.0\%}$	$0.09^{+5.6\%}_{-10.2\%}$	$19.1^{+5\%}_{-7\%}$
8	$21.4^{+5.4\%}_{-7.6\%}$	$1.60^{+2.1\%}_{-2.1\%}$	$0.70^{+2.1\%}_{-2.2\%}$	$0.42^{+3.4\%}_{-2.9\%}$	$0.13^{+5.9\%}_{-10.1\%}$	$24.2^{+5\%}_{-7\%}$
13	$48.6^{+5.6\%}_{-7.4\%}$	$3.78^{+2.1\%}_{-2.1\%}$	$1.37^{+2.0\%}_{-2.0\%}$	$0.88^{+4.1\%}_{-3.5\%}$	$0.50^{+6.8\%}_{-9.9\%}$	$55.1^{+5\%}_{-7\%}$
13.6	$52.2^{+5.6\%}_{-7.4\%}$	$4.1^{+2.1\%}_{-1.5\%}$	$1.46^{+1.8\%}_{-1.9\%}$	$0.95^{+4.0\%}_{-3.6\%}$	$0.57^{+6.9\%}_{-9.9\%}$	$59.2^{+5\%}_{-7\%}$
14	$54.7^{+5.6\%}_{-7.4\%}$	$4.28^{+2.1\%}_{-2.1\%}$	$1.51^{+1.8\%}_{-1.9\%}$	$0.99^{+4.1\%}_{-3.7\%}$	$0.61^{+6.9\%}_{-9.8\%}$	$62.1^{+5\%}_{-7\%}$

Principal decay channels

For a fixed value of the Higgs boson mass, the sensitivity of a decay channel depends on the production cross section of the Higgs boson, its decay branching ratio, the reconstructed mass resolution, the selection efficiency and the level of background in the final state [4]. The BR of the Higgs decays to SM particles are governed by equations 1.39 and 1.45, showing that the Higgs coupling to bosons and fermions are proportional respectively to the squared mass for bosons and to the mass for fermions. For a mass of the Higgs boson of 125 GeV, the most studied decay channels at LHC, ordered according to their branching ratios, are: $H \to b\bar{b}$, $H \to WW^*$, $H \to \tau^+\tau^-$, $H \to c\bar{c}$, $H \to ZZ^*$ and $H \to \gamma\gamma$.

Figure 1.5 (Right) shows the branching ratios for these decay channels as function of the Higgs boson mass in the range 120 GeV $< m_H <$ 130 GeV. Finally, in figure 1.6, the combined measurements, performed by ATLAS [18] and CMS [19], of the product of the Higgs boson production cross section and decay branching ratio, for the five main production and five main decay modes, are displayed. An overall good agreement with the SM prediction is registered.

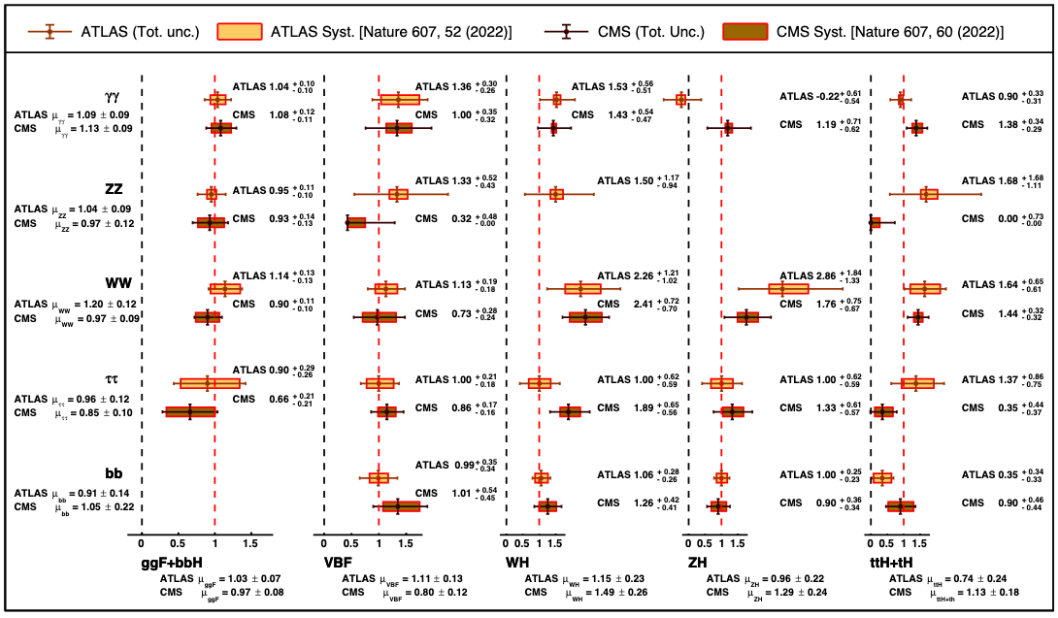


Figure 1.6: Combined measurements by ATLAS [18] and CMS [19] of the products $\sigma \cdot BR$, normalised to the SM predictions, for the five main production and five main decay modes [4].

Measurement of the Higgs Boson mass

The decay channels that allow us to measure the Higgs boson mass with the highest resolution are: $H \rightarrow \gamma\gamma$ and $H \rightarrow ZZ^* \rightarrow 4l$, since all final state particles can be measured very precisely. Figure 1.7 shows, on the left side, the invariant mass distribution of diphoton candidates obtained by ATLAS and, on the right side, the CMS four lepton invariant mass distribution. The best mass resolution is achieved, by both the experiments ATLAS and CMS, in the diphoton channel for central diphoton pairs.

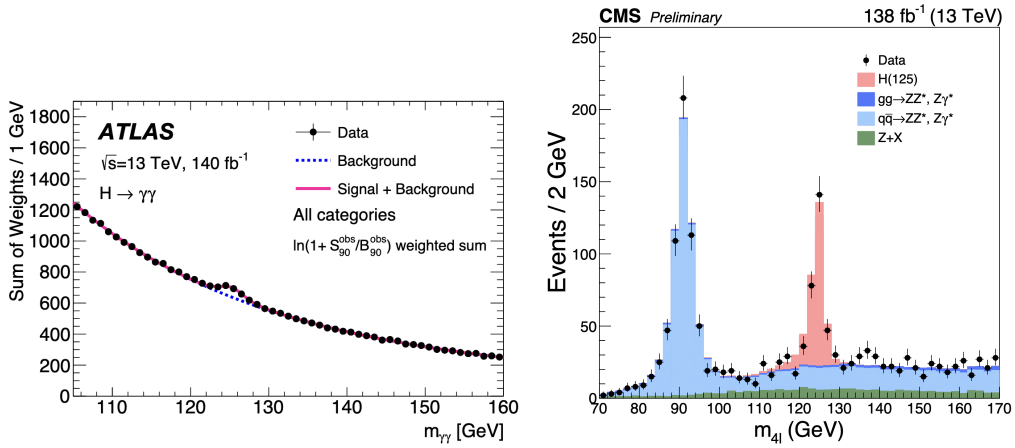


Figure 1.7: (Left) The invariant mass distribution of diphoton candidates, with each event weighted by a factor $ln(1 + \frac{S_{90}^{obs}}{B_{90}^{obs}})$, where S_{90}^{obs} and B_{90}^{obs} are the fitted signal and background yields in the smallest $m_{\gamma\gamma}$ interval containing 90% of the expected signal, observed by ATLAS [20]. (Right) The four lepton invariant mass from CMS [21].

Higgs boson couplings

According to SM electroweak symmetry breaking mechanism, fermions acquire their mass by interacting with the Higgs field. Probing the Higgs boson couplings is extremely important in order to confirm the SM predictions and simultaneously search for new physics evidence. The Higgs boson was discovered essentially by investigating the bosonic final states. However, since the predominant Higgs boson production mode is the gluon fusion, which occurs only through fermionic loops, the observation of the Higgs boson in the two photons or two gluons decay modes is also an indirect evidence for the coupling of the Higgs boson to fermions. Nevertheless, it is strictly necessary to observe directly the Higgs boson couplings to fermions, either in production or decay mechanisms. One of the highest priority goals of the LHC Run-2 (2016-2018) physics program was the direct observation of the Yukawa cou-

pling of the Higgs boson to fermions of the third generation (bottom quarks, tau leptons and top quarks). This goal has been reached independently by both ATLAS and CMS and with only partial Run-2 datasets [4]. It follows that one of the predominant goals of the Run-3 program (2022-present) is to assess the Higgs boson coupling with second generation fermions. Already by analyzing the Run-2 dataset, the CMS Collaboration reported [22] the first evidence for $H \to \mu\mu$ at 3σ level with a signal strength of $\mu = 1.19 \pm 0.40$ (stat.) $\pm$ 0.15 (syst.) and an expected analysis sensitivity of 2.5σ, leading to the first direct evidence for the $H \to \mu\mu$ decay and therefore to the measurement of the Yukawa coupling of the Higgs boson to second generation leptons.

The properties of the Higgs boson couplings are expressed, within the κ framework [23], in terms of a series of Higgs coupling strength modifier parameters k_i, defined as the ratios between the couplings of the Higgs bosons to particles i and their corresponding SM values.
Within the κ framework, the cross section is decomposed as a product of two factors describing the production and the decay of the Higgs boson:

$$(\sigma \cdot BR)(i \to H \to f) = \frac{\sigma_i \Gamma_f}{\Gamma_H}, \tag{1.52}$$

where σ_i is the cross section of the Higgs boson production starting from the initial state i, Γ_f is the partial decay width into the final state f and Γ_H is the total width of the Higgs boson.
These factors can be expressed as the product of their SM expectation times the square of a coupling strength modifier parameter k_i, specific of the process. Then, the rate relative to the SM expectation μ_i^f results to be:

$$\mu_i^f \equiv \frac{\sigma \cdot BR}{\sigma_{SM} \cdot BR_{SM}} = \frac{k_i^2 \cdot k_f^2}{k_H^2} \tag{1.53}$$

where k_H^2 is an expression that adjusts the SM Higgs width to take into account the modifications induced by the deformed Higgs boson couplings.
For $k_i = 1$, the SM is reproduced. Figure 1.8 shows the Higgs couplings measured by ATLAS (left) and CMS (right) at $\sqrt{s} = 13$ TeV, in agreement within uncertainties with the SM predictions.

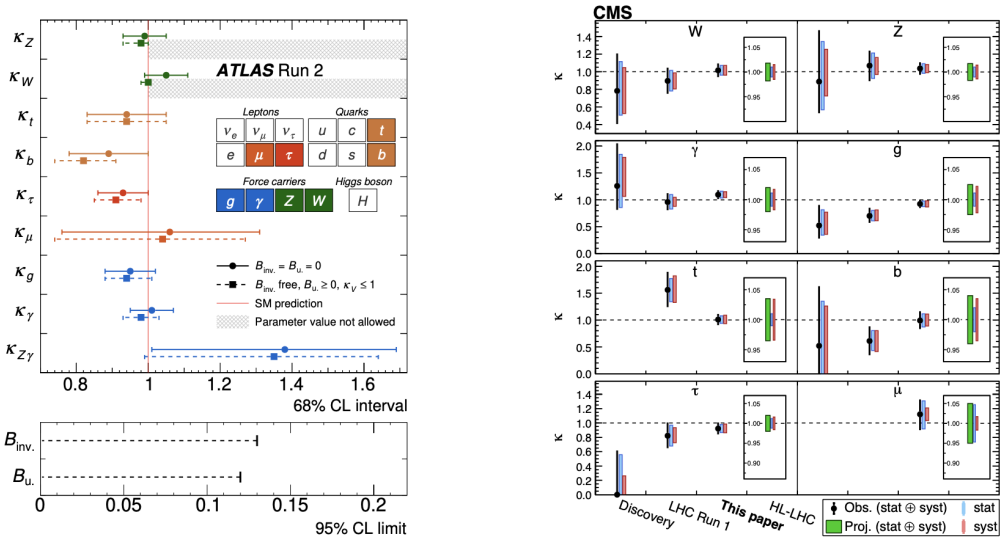


Figure 1.8: ATLAS [18] (left) and CMS [19] (right) combined measurements of coupling modifiers with various assumptions [4]. [24]

1.4 Standard model $H \to c\bar{c}$ state-of-the-art

The observed Higgs boson decay rates are in general in agreement with the SM predictions within a 10% level of precision, and global fits to the Higgs precision measurements show no evidence of new physics. However, it is fundamental to continue probing the Higgs boson decay rates with better and better precision. Indeed, there could be deviations from the SM not yet probed at the LHC, revealing the presence of new physics at the weak scale. In addition, deviations from the SM predictions of the Higgs boson couplings with first and second generation fermions, still out of reach, may be an indication of a more complex mechanism of mass generation than the one theorized by the SM. Then, there may be decays of the Higgs bosons into exotic particles not yet detected by the LHC, or, there could also be hidden correlations between the Higgs couplings with rates in agreement with the SM predictions. After observing the Higgs boson decay to muons, the next step toward completing the Higgs boson picture is to measure the Higgs boson coupling with charm quarks. Large deviations of k_c from one would affect the Higgs width and consequently its decay branching ratios. In such a case, the couplings of the Higgs boson to gauge bosons and third-generation fermions must also be modified in order to preserve the agreement with experimental observations [25].

For a long time, the observation of the Higgs boson decay to charm quarks has been considered out of reach at the LHC, due to the very small branching

ratio (∼3%), the overwhelming QCD multi-jet background, and, most important, the complexity of discriminating c jets from b jets and light flavour jets. However, novel data analyses techniques developed in particular by the CMS Collaboration, as c-tagging algorithms based on the state-of-the-art machine learning architectures [26], have greatly increased the sensitivity to this search.

The most sensitive Higgs boson production mechanism for the investigation of the $H \to c\bar{c}$ decay is the VH channel: by targeting the the leptonic decays of the vector boson produced in association with the Higgs boson, the QCD multijet background is highly suppressed. A crucial step of the search for $H \to c\bar{c}$ is the c-tagging, i.e., the discrimination of jets originated by the hadronization of charm quarks (c jets) against those originated by beauty quarks (b jets) or light quarks and gluons (udsg jets). This goal is particularly challenging as charmed jets have intermediate characteristics between b- and udsg-jets. As anticipated, after the development of these new powerful c-tagging techniques, which will be discussed in details in Chapter 3, the CMS Collaboration has achieved outstanding improvements in this search, leading to the first observation of Z boson decays to charm quarks at a hadron collider with a significance of 5.7 standard deviations [6] and an expected sensitivity to the exclusion of a Higgs boson signal decaying to a pair of charm quarks of 7.6 times the Standard Model rate. The CMS Collaboration set an upper limit of 14 on the signal strength and limits on the Yukawa coupling of the Higgs boson to charm quarks: the observed (expected) 95% CL interval is $1.1 < |k_c| < 5.5$ ($|k_c| < 3.4$).
Very recently, the ATLAS Collaboration published the result of the simultaneous investigation of $H \to b\bar{b}$ and $H \to c\bar{c}$ in VH prodcution with Run-2 data. An observed (expected) upper limit of 11.5 (10.6) at 95% CL was set on the $VH\,H \to c\bar{c}$ signal strength and the corresponding limit on the charm Yukawa coupling modifier is $|k_c| < 4.2$ [5].

Additionally, a search for the Higgs boson produced with $p_T > 450$ GeV and decaying to a charm quark-antiquark pair was performed by the CMS Collaboration, targeting the ggF production mode. The observed (expected) 95% CL upper limit on the signal strength $\mu_{H\to c\bar{c}}$ is 47. This search is validated by measuring the $Z \to c\bar{c}$ process, observed in association with high-pT jets for the first time in this production channel, with signal strength $\mu = 1.00^{+0.17}_{-0.14}$ (syst) ± 0.08 (theo) ± 0.06 (stat) [27].

Recently, the CMS Collaboration reported for the first time the search for

the Higgs boson produced in association with a charm quark in the diphoton decay channel, setting an observed (expected) upper limit on k_c at 95% CL equal to $|k_c| < 38.1$ (72.5) [28].
All these searches aim to enhance the sensitivity to the Higgs boson coupling to charm quarks, with its measurement being one of the highest priority goals of the CMS Collaboration.
In this work, for the first time the search for $H \to c\bar{c}$ in the VBF Higgs boson production channel is investigated and the analysis strategy and results will be discussed in detail in Chapter 4.

A projection study demonstrated that at the High Luminosity phase (Phase-2) of LHC (HL-LHC), designed to increase the integrated luminosity by a factor of 10 beyond the LHC's original value, the expected sensitivity on the Hcc coupling could approach the SM value, by extracting simultaneously the constraints on $H \to c\bar{c}$ and $H \to b\bar{b}$ [29]. Figure 1.9 shows the HL-LHC projections of the expected constraints on the $H \to b\bar{b}$ and $H \to c\bar{c}$ signal strengths (left) and k_b and k_c (right).

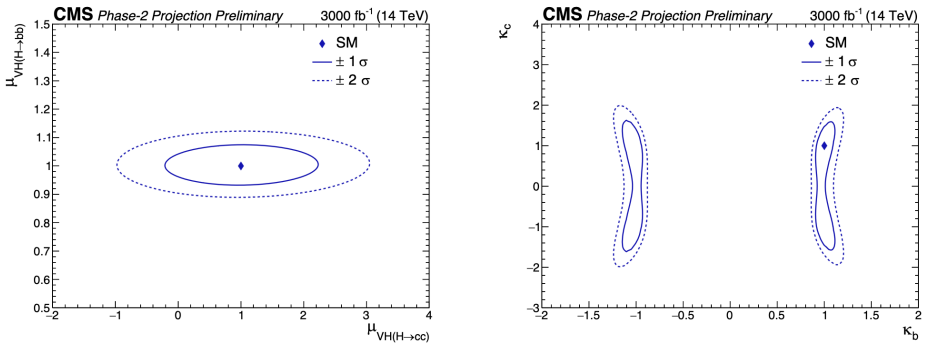


Figure 1.9: HL-LHC projection of the expected constraints on the $H \to b\bar{b}$ and $H \to c\bar{c}$ signal strengths (left) and k_b and k_c (right). [29]

1.5 SM Effective Field Theory

Although the SM is currently our best theory for describing subatomic physics, it is incomplete and it fails to address several critical phenomena. First, the SM does not explain how gravity is mediated. There is no experimental evidence supporting the existence of hypothesized gravitons, making the SM incompatible with the Theory of General Relativity. Moreover, the SM does not include any fundamental particles that can be a suitable candidate for dark matter, which constitutes the $\sim$ 26% of the universe. Neutrinos are predicted by the SM to have null mass. However, neutrinos oscillation, which are directly re-

lated to non-null mass values, have been observed experimentally. In addition, the SM is unable to explain the matter-antimatter asymmetry observed in the universe. According to the SM, matter and antimatter should have been produced in nearly equal amounts during the Big Bang. However, the universe is predominantly composed of matter, and the SM does not contain a mechanism to adequately explain this discrepancy. In other to address these and other open questions, several beyond SM (BSM) theories have been formulated, such as the supersymmetry and the string theory.
Two different approaches are usually adopted to look for new physics: model dependent and model independent. The former approach searches for new particles predicted by the BSM model under investigation. According to the model independent one, instead, one should measure interactions between SM particles with extremely high precision and look for discrepancies. This second approach is followed by the SM Effective Field Theory (SMEFT) application.

The assumption on the basis of EFT is that the dynamics at low energies does not depend on the details of the dynamics at high energies. It follows that low energy physics can be described using an effective Lagrangian that contains only a few degrees of freedom, ignoring additional degrees of freedom present at higher energy [30]. In d spacetime dimensions, the lagrangian has dimension d and it can be written as a linear combination of local, gauge invariant and Lorentz invariant operators O_i, through the coefficients c_i [31]:

$$\mathscr{L}_{EFT}(x) = \sum_i c_i O_i(x) \tag{1.54}$$

The operator dimension is denoted by $\mathscr{D}$, and its coefficient has dimension $d - \mathscr{D}$. Equation 1.54 can be written as follows:

$$\mathscr{L}_{EFT}(x) = \sum_{\mathscr{D}\geq 0, i} \frac{c_i^{(\mathscr{D})} O_i^{(\mathscr{D})}}{\Lambda^{\mathscr{D}-d}} = \sum_{\mathscr{D}\geq 0} \frac{\mathscr{L}_{\mathscr{D}}}{\Lambda^{\mathscr{D}-d}} \tag{1.55}$$

where Λ is a scale introduced in order to make the coefficients $c_i^{(\mathscr{D})}$ dimensionless and represents the scale for new physics. The EFT lagrangian can be written as an infinite series of terms of increasing operator dimension:

$$\mathscr{L}_{EFT} = \mathscr{L}_{\mathscr{D}\leq 4} + \frac{\mathscr{L}_5}{\Lambda} + \frac{\mathscr{L}_6}{\Lambda^2} + ... \tag{1.56}$$

In this picture, SMEFT is an EFT of the SM degrees of freedom. The SMEFT lagrangian contains an infinite set of higher-dimensional gauge-invariant interaction terms, in addition to the SM one:

$$\mathscr{L}_{SMEFT} = \mathscr{L}_{SM} + \sum_i \frac{c_i^{(5)} O_i^{(5)}}{\Lambda} + \sum_i \frac{c_i^{(6)} O_i^{(6)}}{\Lambda^2} + \sum_i \frac{c_i^{(7)} O_i^{(7)}}{\Lambda^3} + \sum_i \frac{c_i^{(8)} O_i^{(8)}}{\Lambda^4} + ..., \tag{1.57}$$

Bosonic CP-even		Bosonic CP-odd	
O_H	$(H^\dagger H)^3$		
$O_{H\Box}$	$(H^\dagger H)\Box(H^\dagger H)$		
O_{HD}	$\left\|H^\dagger D_\mu H\right\|^2$		
O_{HG}	$H^\dagger H\, G^a_{\mu\nu} G^a_{\mu\nu}$	$O_{H\widetilde{G}}$	$H^\dagger H\, \widetilde{G}^a_{\mu\nu} G^a_{\mu\nu}$
O_{HW}	$H^\dagger H\, W^i_{\mu\nu} W^i_{\mu\nu}$	$O_{H\widetilde{W}}$	$H^\dagger H\, \widetilde{W}^i_{\mu\nu} W^i_{\mu\nu}$
O_{HB}	$H^\dagger H\, B_{\mu\nu} B_{\mu\nu}$	$O_{H\widetilde{B}}$	$H^\dagger H\, \widetilde{B}_{\mu\nu} B_{\mu\nu}$
O_{HWB}	$H^\dagger \sigma^i H\, W^i_{\mu\nu} B_{\mu\nu}$	$O_{H\widetilde{W}B}$	$H^\dagger \sigma^i H\, \widetilde{W}^i_{\mu\nu} B_{\mu\nu}$
O_W	$\epsilon^{ijk} W^i_{\mu\nu} W^j_{\nu\rho} W^k_{\rho\mu}$	$O_{\widetilde{W}}$	$\epsilon^{ijk} \widetilde{W}^i_{\mu\nu} W^j_{\nu\rho} W^k_{\rho\mu}$
O_G	$f^{abc} G^a_{\mu\nu} G^b_{\nu\rho} G^c_{\rho\mu}$	$O_{\widetilde{G}}$	$f^{abc} \widetilde{G}^a_{\mu\nu} G^b_{\nu\rho} G^c_{\rho\mu}$

Table 1.2: Bosonic D=6 operators in the Warsaw basis [32].

where the $c_i^{\mathscr{D}}$ are called Wilson coefficients and the operators of dimension $\mathscr{D}$ are suppressed by $\Lambda^{\mathscr{D}-4}$.
As the SM, the SMEFT theory is invariant under the local $SU(3) \times SU(2) \times U(1)$ symmetry. This EFT is intended to parameterize observable effects of a large class of BSM theories where new particles, with mass of order Λ, are much heavier than the SM ones and much heavier than the energy scale at which the experiment is performed [32].
Constraints on the parameters of the Effective Field Theory can subsequently be reinterpreted as constraints on several BSM model predictions.
Dimension 5 operators violate the lepton number and their most important effect is the appearance of Majorana-type neutrino masses after electroweak symmetry breaking. From the observations of neutrino oscillations, it results $\frac{\Lambda}{c_5} \geq 10^{15}$ GeV, making the effect of dimension-5 operators practically unobservable outside of the neutrino oscillation experiments.

It is useful to define a basis, i.e. a complete and non-redundant set of operators. One of the most commonly used one is the Warsaw basis and its operators are summarized in Tables 1.2, 1.3 and 1.4.

It results convenient to switch from the k-parameterization to one given in terms of Wilson coefficients of a non-redundant EFT basis. In fact, it is possible to identify a one way mapping between the k_i and C_i. For each k_i, we can write:

$$k_i = 1 + \Delta k_i, \tag{1.58}$$

where Δk_i is a linear combination of EFT parameters [33]. Let us consider the

Yukawa	
$[O_{eH}^\dagger]_{IJ}$	$H^\dagger H e_I^c H^\dagger \ell_J$
$[O_{uH}^\dagger]_{IJ}$	$H^\dagger H u_I^c \widetilde{H}^\dagger q_J$
$[O_{dH}^\dagger]_{IJ}$	$H^\dagger H d_I^c H^\dagger q_J$

Vertex	
$[O_{H\ell}]_{IJ}$	$i\bar{\ell}_I \bar{\sigma}_\mu \ell_J H^\dagger \overleftrightarrow{D_\mu} H$
$[O_{H\ell}^{(3)}]_{IJ}$	$i\bar{\ell}_I \sigma^i \bar{\sigma}_\mu \ell_J H^\dagger \sigma^i \overleftrightarrow{D_\mu} H$
$[O_{He}]_{IJ}$	$ie_I^c \sigma_\mu \bar{e}_J^c H^\dagger \overleftrightarrow{D_\mu} H$
$[O_{Hq}]_{IJ}$	$i\bar{q}_I \bar{\sigma}_\mu q_J H^\dagger \overleftrightarrow{D_\mu} H$
$[O_{Hq}^{(3)}]_{IJ}$	$i\bar{q}_I \sigma^i \bar{\sigma}_\mu q_J H^\dagger \sigma^i \overleftrightarrow{D_\mu} H$
$[O_{Hu}]_{IJ}$	$iu_I^c \sigma_\mu \bar{u}_J^c H^\dagger \overleftrightarrow{D_\mu} H$
$[O_{Hd}]_{IJ}$	$id_I^c \sigma_\mu \bar{d}_J^c H^\dagger \overleftrightarrow{D_\mu} H$
$[O_{Hud}]_{IJ}$	$iu_I^c \sigma_\mu \bar{d}_J^c \tilde{H}^\dagger D_\mu H$

Dipole	
$[O_{eW}^\dagger]_{IJ}$	$e_I^c \sigma_{\mu\nu} H^\dagger \sigma^i \ell_J W_{\mu\nu}^i$
$[O_{eB}^\dagger]_{IJ}$	$e_I^c \sigma_{\mu\nu} H^\dagger \ell_J B_{\mu\nu}$
$[O_{uG}^\dagger]_{IJ}$	$u_I^c \sigma_{\mu\nu} T^a \widetilde{H}^\dagger q_J \, G_{\mu\nu}^a$
$[O_{uW}^\dagger]_{IJ}$	$u_I^c \sigma_{\mu\nu} \widetilde{H}^\dagger \sigma^i q_J \, W_{\mu\nu}^i$
$[O_{uB}^\dagger]_{IJ}$	$u_I^c \sigma_{\mu\nu} \widetilde{H}^\dagger q_J \, B_{\mu\nu}$
$[O_{dG}^\dagger]_{IJ}$	$d_I^c \sigma_{\mu\nu} T^a H^\dagger q_J \, G_{\mu\nu}^a$
$[O_{dW}^\dagger]_{IJ}$	$d_I^c \sigma_{\mu\nu} \bar{H}^\dagger \sigma^i q_J \, W_{\mu\nu}^i$
$[O_{dB}^\dagger]_{IJ}$	$d_I^c \sigma_{\mu\nu} H^\dagger q_J \, B_{\mu\nu}$

Table 1.3: Two-fermion D=6 operators in the Warsaw basis [32].

$(\bar{R}R)(\bar{R}R)$	
O_{ee}	$\eta(e^c\sigma_\mu\bar{e}^c)(e^c\sigma_\mu\bar{e}^c)$
O_{uu}	$\eta(u^c\sigma_\mu\bar{u}^c)(u^c\sigma_\mu\bar{u}^c)$
O_{dd}	$\eta(d^c\sigma_\mu\bar{d}^c)(d^c\sigma_\mu\bar{d}^c)$
O_{eu}	$(e^c\sigma_\mu\bar{e}^c)(u^c\sigma_\mu\bar{u}^c)$
O_{ed}	$(e^c\sigma_\mu\bar{e}^c)(d^c\sigma_\mu\bar{d}^c)$
O_{ud}	$(u^c\sigma_\mu\bar{u}^c)(d^c\sigma_\mu\bar{d}^c)$
$O_{ud}^{(8)}$	$(u^c\sigma_\mu T^a\bar{u}^c)(d^c\sigma_\mu T^a\bar{d}^c)$

$(\bar{L}L)(\bar{R}R)$	
$O_{\ell e}$	$(\bar{\ell}\bar{\sigma}_\mu\ell)(e^c\sigma_\mu\bar{e}^c)$
$O_{\ell u}$	$(\bar{\ell}\bar{\sigma}_\mu\ell)(u^c\sigma_\mu\bar{u}^c)$
$O_{\ell d}$	$(\bar{\ell}\bar{\sigma}_\mu\ell)(d^c\sigma_\mu\bar{d}^c)$
O_{eq}	$(e^c\sigma_\mu\bar{e}^c)(\bar{q}\bar{\sigma}_\mu q)$
O_{qu}	$(\bar{q}\bar{\sigma}_\mu q)(u^c\sigma_\mu\bar{u}^c)$
$O_{qu}^{(8)}$	$(\bar{q}\bar{\sigma}_\mu T^a q)(u^c\sigma_\mu T^a\bar{u}^c)$
O_{qd}	$(\bar{q}\bar{\sigma}_\mu q)(d^c\sigma_\mu\bar{d}^c)$
$O_{qd}^{(8)}$	$(\bar{q}\bar{\sigma}_\mu T^a q)(d^c\sigma_\mu T^a\bar{d}^c)$

$(\bar{L}L)(\bar{L}L)$	
$O_{\ell\ell}$	$\eta(\bar{\ell}\bar{\sigma}_\mu\ell)(\bar{\ell}\bar{\sigma}_\mu\ell)$
O_{qq}	$\eta(\bar{q}\bar{\sigma}_\mu q)(\bar{q}\bar{\sigma}_\mu q)$
O'_{qq}	$\eta(\bar{q}\bar{\sigma}_\mu\sigma^i q)(\bar{q}\bar{\sigma}_\mu\sigma^i q)$
$O_{\ell q}$	$(\bar{\ell}\bar{\sigma}_\mu\ell)(\bar{q}\bar{\sigma}_\mu q)$
$O'_{\ell q}$	$(\bar{\ell}\bar{\sigma}_\mu\sigma^i\ell)(\bar{q}\bar{\sigma}_\mu\sigma^i q)$

$(\bar{L}R)(\bar{L}R)$	
O_{quqd}	$(u^c q^j)\epsilon_{jk}(d^c q^k)$
$O_{quqd}^{(8)}$	$(u^c T^a q^j)\epsilon_{jk}(d^c T^a q^k)$
$O_{\ell equ}$	$(\bar{\ell}^j\bar{e}^c)\epsilon_{jk}(\bar{q}^k\bar{u}^c)$
$O_{\ell equ}^{(3)}$	$(\bar{\ell}^j\bar{\sigma}_{\mu\nu}\bar{e}^c)\epsilon_{jk}(\bar{q}^k\bar{\sigma}^{\mu\nu}u^c)$
$O_{\ell edq}$	$(\bar{\ell}\bar{e}^c)(d^c q)$

Table 1.4: Four-fermion D=6 operators in the Warsaw basis [32].

$H \to b\bar{b}$ decay; in the Warsaw basis we get:

$$\kappa_b^2 = \frac{\mathcal{A}^2(h \to \bar{b}b)_{\text{SMEFT}}}{\mathcal{A}^2(h \to \bar{b}b)_{\text{SM}}} = 1 + \Delta\kappa_b \,,$$
$$\Delta\kappa_b = 2\,\bar{v}_T^2 \left(C_{H\Box} - \frac{C_{HD}}{4} - C_{Hl}^{(3)} + \frac{C'_{ll}}{2} - \frac{C_{dH}}{[Y_d]_{33}} \right) . \tag{1.59}$$

where C_{dH} represents the only direct d = 6 contribution, which is due to the operator O_{dH}, that perturbs the Yukawa coupling. This way, by setting constraints on the k_i coupling modifiers, it is possible to obtain constraints on the SMEFT operators.

The Warsaw basis contains 59 non redundant dimension 6 operators, assuming minimal flavour violation, and, among these, 20 are relevant for the Higgs, diboson and electroweak precision observables. A global fit to precision electroweak data, W^+W^- measurements at LEP, and Higgs and diboson data from Runs 1 and 2 of the LHC has been performed in the framework of the SMEFT [34]. The results in the Warsaw basis are displayed in Figure 1.10: in the upper plot all the operators are included simultaneously, while in the lower one they are switched individually. The results obtained before the inclusion of Run-2 data are also superimposed in blue. So far, these fits provide no sign or evidence of any physics beyond the Standard Model. However, the constraints obtained on the dimension-6 operator coefficients can be applied to several BSM scenarios. Tighter constraints will be set in the near future, as new data from the CMS and ATLAS Run-2 and Run-3 analyses will become available.

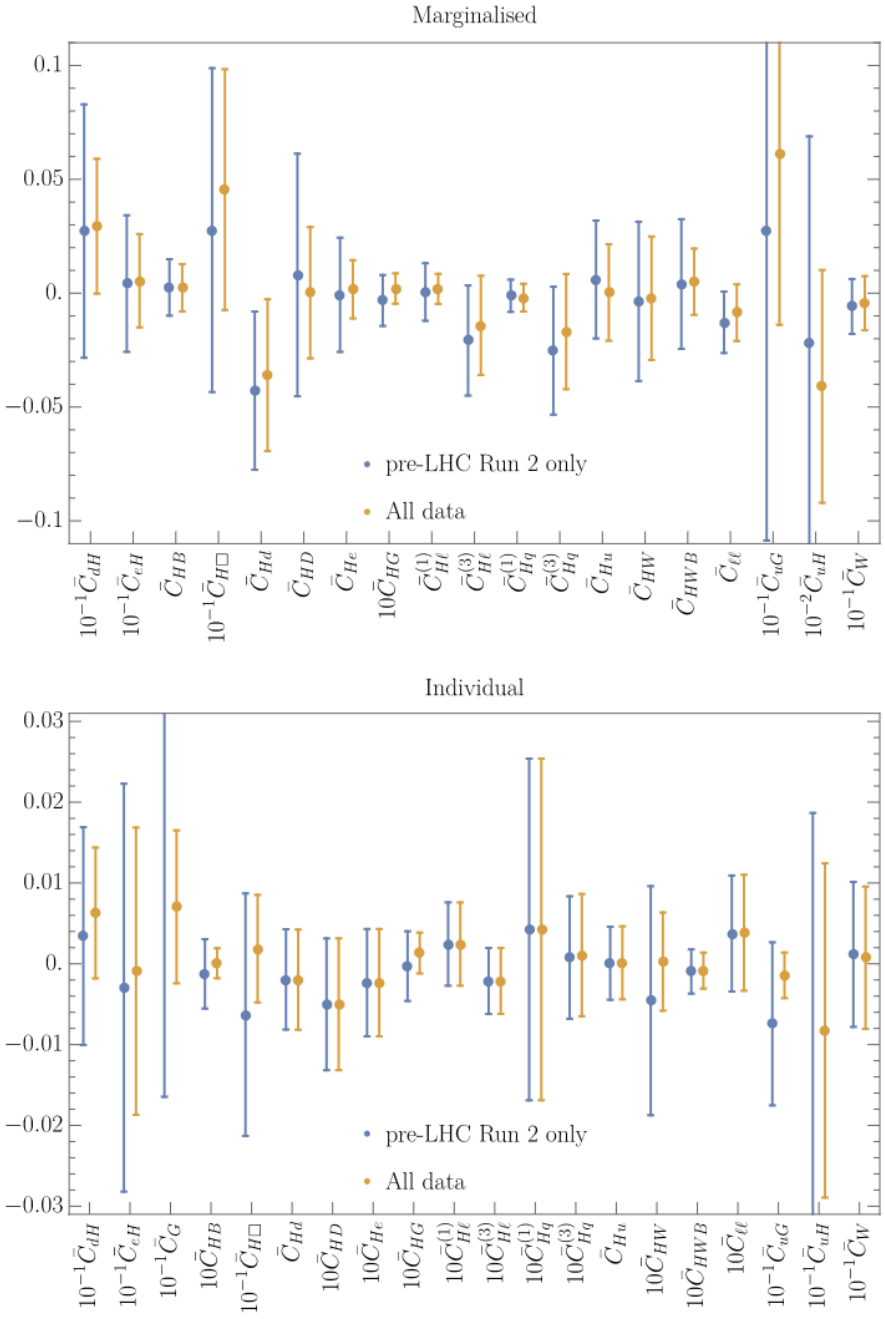


Figure 1.10: Results from global fits in the Warsaw basis (orange) including all operators simultaneously (upper panel) and switching each operator on individually (lower panel). Also shown are fits omitting the LHC Run 2 data (blue). We display the best-fit values and 95% CL ranges [34].

1.6 Higgs coupling to charm quarks in a Two Higgs Doublet Model

An interesting study conducted in [35] shows that using a Two Higgs Doublet Model (2HDM) as an example, within the framework of Spontaneous Flavour Violation (SFV), it is possible to get an enhanced charm Yukawa coupling, that could be constrained or even directly measured at the LHC.
In typical models of flavour BSM, it is challenging to obtain a larger coupling of the Higgs boson to charm quarks, without violating other experimental constraints. A valid alternative is provided by the SFV, that has been previously studied for down-type quark Yukawa couplings. In particular, the authors of [35] focus on a Two Higgs Doublet Model, with the Higgs sector extended with an additional doublet. The results suggest that, in a particular configuration (non-zero alignment parameter) of the SVF 2HDM, experimental constraints can allow for a second Higgs with Yukawa couplings of $O(10^{-1})$ to any up-type quark without having to respect any of the hierarchies of the SM Yukawas. For example one could have a new Higgs that couples at this strength only to the charm quark. When this new Higgs has a non-zero mixing with the SM Higgs, this allows for large deviations of O(10) to $O(10^4)$ to the SM charm and up Yukawa couplings consistent with measurements of flavour changing neutral currents (FCNC).
An important finding of this study is that, although charm tagging is currently a weaker bound than normalization, in the large mass regime it could be more effective, providing valuable new information compared to other constraints. This provides strong incentive to continue improving charm tagging algorithms.

Chapter 2

The LHC and CMS experiment

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) stands as the world's most powerful particle accelerator, engineered to collide protons or lead ions at unprecedented energies. It occupies a 27-kilometer circular tunnel, 100 meters underground, located at CERN in Geneva. This tunnel, initially built between 1983 and 1988 for the Large Electron-Positron Collider (LEP), was later repurposed for the LHC.
Construction of the LHC spanned from 1998 to 2008, with the primary goal of verifying the Standard Model (SM) by detecting the elusive Higgs boson. The collider also explores the TeV-energy range, searching for new physics beyond the Standard Model, such as supersymmetric particles, which were widely considered potential extensions of the SM at the time.

The accelerator is designed to collide proton beams at a center-of-mass energy of $\sqrt{s} = 14$ TeV, achieving an instantaneous luminosity of 1×10^{34} cm^{-2}s^{-1}. Additionally, it can collide lead ions with an energy of 5.0 TeV per nucleon, reaching peak luminosities of 10^{27} cm^{-2}s^{-1}. Proton and heavy ion beams are accelerated in bunches, with 40 million bunch crossings occurring per second. These bunches are directed to collide at four key interaction points along the tunnel, where the detector systems of the LHC four major experiments are located, as illustrated in Figure 2.1.
These experiments are: *ALICE* (A Large Ion Collider Experiment), *ATLAS* (A Toroidal LHC ApparatuS), *CMS* (Compact Muon Solenoid), and *LHCb* (Large Hadron Collider beauty). ATLAS and CMS are multipurpose detectors, designed to discover the Higgs boson and investigate beyond Standard Model scenarios. ALICE focuses on studying quark-gluon plasma created in

heavy ion collisions, while LHCb examines particles produced in the forward region, particularly focused on the study of heavy-flavoured mesons.

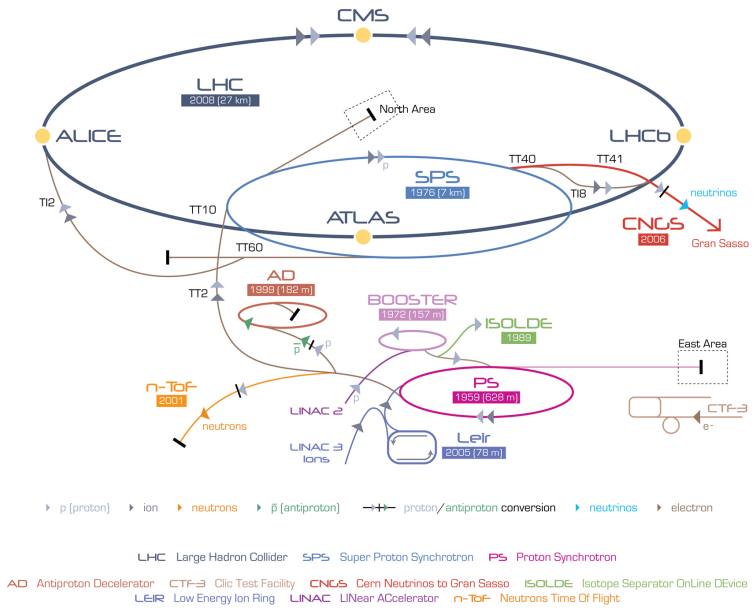


Figure 2.1: Illustration of the CERN accelerator complex (Image: CERN).

Proton beams are generated from hydrogen gas, undergoing several stages of acceleration and focusing. Initially, negative hydrogen ions are produced at 160 MeV by LINAC2 and then passed through the Proton Synchrotron Booster (PSB), where electrons are stripped, leaving only protons. These protons are further accelerated by the Proton Synchrotron (PS) and Super Proton Synchrotron (SPS) to an energy of 450 GeV before being injected into the LHC rings for final acceleration.

Within the LHC, superconducting dipole magnets, operating at an 8.3 T magnetic field, guide the proton beams along their circular path. These magnets are cooled to 1.9 K using superfluid helium. The beams are focused by quadrupole magnets, while eight radiofrequency cavities per beam accelerate them.

A critical parameter in collider experiments is luminosity, which measures the number of particle collisions per unit time relative to the interaction cross section:

$$\mathcal{L} = \frac{1}{\sigma}\frac{\mathrm{d}N}{\mathrm{d}t}$$

Luminosity depends on various machine parameters, including beam size, bunch content, and revolution frequency. In the LHC, each beam is made of about 2500 bunches, separated by 25 ns intervals, and each proton bunch typically

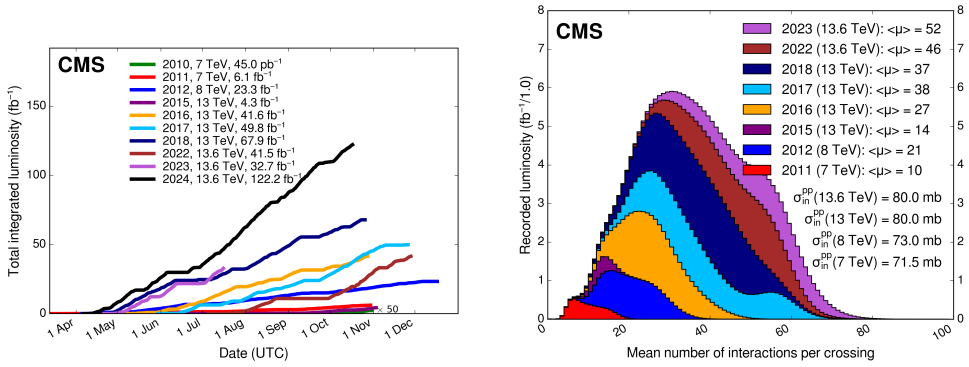


Figure 2.2: Left: luminosity delivered to the CMS experiment during stable beams for proton-proton collisions at nominal center-of-mass energy, in 2010-2012 (Run-1), 2015-2018 (Run-2) and 2022-2024 (Run-3) data taking periods, separately for each year [36]. Right: distribution of the average number of interactions per crossing (pileup) for pp collisions in Run-1, Run-2 and beginning of Run-3. The overall mean values and the minimum bias cross sections are also shown [36].

contains around 10^{11} protons.

Figure 2.2 (left) illustrates the LHC's integrated luminosity achievements during Run-1 (2010-2012), Run-2 (2015-2018) and beginning of Run-3 (2022-2024). During Run-1, proton-proton collisions were first achieved at a center-of-mass energy of 7 TeV in 2011, later increased to 8 TeV in 2012. This period culminated in the discovery of the Higgs boson by the CMS and ATLAS collaborations. Following upgrades, the LHC resumed operations in 2015, reaching 13 TeV collisions until 2018. By the end of Run 2, the LHC exceeded its design luminosity by a factor of two. After further upgrades, proton-proton collisions restarted in 2022 with Run-3, which is currently ongoing.

Given the high beam intensities needed for such luminosity, multiple proton-proton collisions occur with each bunch crossing, a phenomenon known as *pileup* (PU). In physics analyses, only the highest energy collision is typically considered, with the others classified as pileup events. The LHC was designed with an average PU of 25, as shown in the distribution measured by CMS during Run-1, Run-2 and Run-3 in Figure 2.2 (right). It is evident that Run-3 proton-proton collisions are characterized by significantly higher instantaneous luminosity, resulting in a much more intense pileup. Operating at this level of pileup represents an unprecedented challenge for both the detector apparatus, which was originally designed for lower radiation intensity, and the software reconstruction chain.

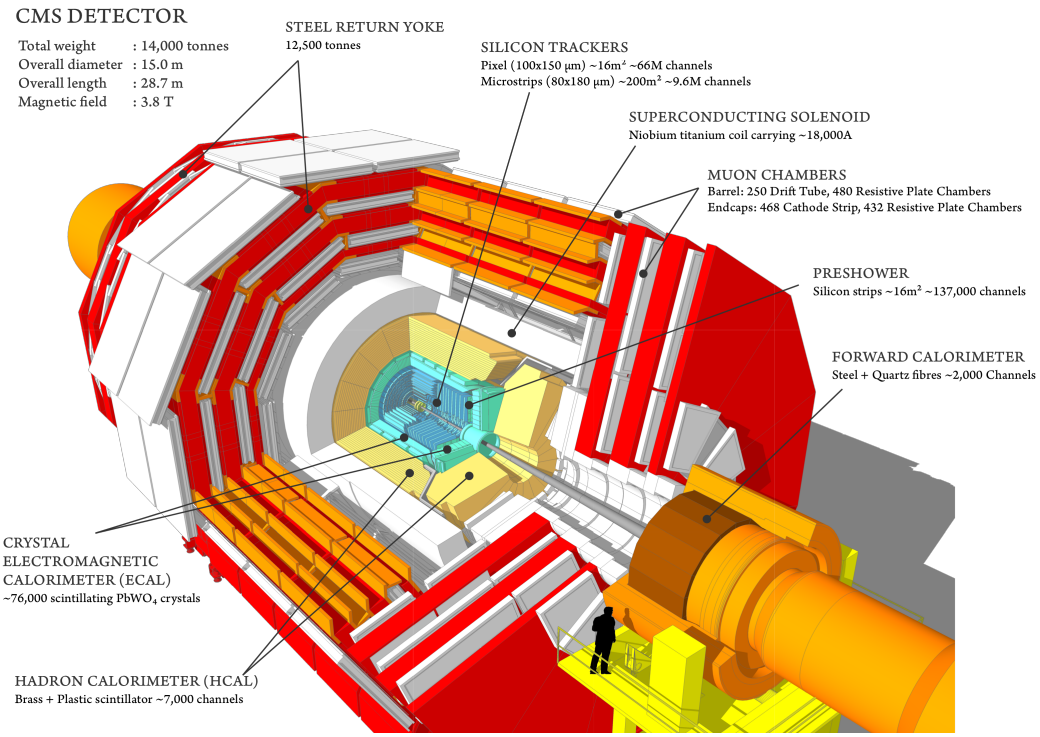


Figure 2.3: Layout of the CMS detector [37].

A significant upgrade, known as the High Luminosity LHC (HL-LHC), is planned for 2026-2029. This upgrade will considerably boost the luminosity delivered to the experiments, achieving an integrated luminosity of 250 fb^{-1} per year, compared to the 68 fb^{-1} delivered to CMS in 2018. The enhanced luminosity will increase sensitivity to rare decays, allowing us to probe with higher precision the Standard Model and explore for new physics.

2.2 The Compact Muon Solenoid experiment

The CMS experiment is one of the four major experiments at the LHC [38]. The detector has a cylindrical structure typical of collider experiments, with an approximate length of 30 m and a diameter of 15 m. As its name implies, a key component of the CMS apparatus is a superconducting solenoid with an internal diameter of 6 m, generating a magnetic field of 3.8 T.

Starting from the interaction point and moving outward, the detector consists of a silicon pixel and strip tracker, a lead tungstate crystal electromagnetic calorimeter (ECAL), and a brass and scintillator hadron calorimeter (HCAL), all housed within the solenoid. Having both the calorimeters within the magnetic field is a key feature of the CMS detector system, distinguishing it from the ATLAS system, which has the hadronic calorimeter outside the magnetic

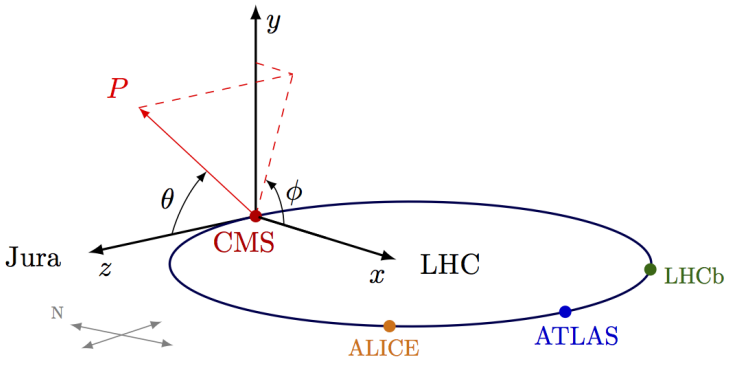


Figure 2.4: CMS coordinate system [39].

field.
Outside the solenoid, the muon reconstruction system is composed of gaseous detectors embedded in the steel flux-return yoke, which is essential for guiding the magnetic field lines. Figure 2.3 illustrates the structure of the CMS detector.

2.2.1 Coordinate system and kinematics of proton-proton collisions

The CMS detector is symmetrically structured around the proton beam line, with the collision point at its center. Consequently, the coordinate system used is as depicted in Figure 2.4, where the z-axis aligns with the beam line, and proton collisions occur at $z = 0$. In this system, the x-y plane is transverse to the beam, with the y-axis pointing upwards towards the surface and the x-axis pointing towards the LHC ring's center (right-handed coordinate system).

Spherical coordinates provide a convenient solution for describing the positions of outgoing particles, given the symmetrical design of the apparatus. The radial coordinate r is the distance from the z-axis on the x-y plane, the azimuthal angle ϕ is measured in this plane starting from the x-axis, and the polar angle θ is measured relative to the z-axis.

In each bunch crossing, multiple proton-proton interactions occur. Quantum Chromodynamics (QCD) governs these interactions, with their description and kinematics varying significantly between *hard* and *soft* processes. In soft interactions, the momentum transfer between particles is minimal, causing most of the resulting particles to escape detection, as they travel almost parallel to the beam pipe. These processes are dominated by non-perturbative QCD effects. Hard scattering processes, on the other hand, involve collisions between the

proton constituents, quarks and gluons (partons), and the momentum transfer Q^2 is large compared to the QCD scale ($Q^2 \gg 1GeV^2$). The cross-sections of these processes can be described as a convolution of the parton distribution functions (PDFs) of the incoming protons and the hard-scattering cross-section at the parton level, calculated using perturbative expansions in the strong coupling constant α_s. The results of the hard scattering of partons are color-charged particles, which cannot exist freely due to QCD confinement, and generate, with other *colored* objects around them, color-neutral hadrons. This process, known as hadronization, is inherently non-perturbative and is modeled phenomenologically [40].

At high energy hadron colliders, the colored fragments from hard scattering have typically large momenta in the x-y plane and tend to move in the same direction, causing the products of hadronization to form narrow "jets" of particles. Additionally, after the hard interaction, the remnants of the two protons must also hadronize, forming jets that travel nearly parallel to the beam axis. Since the energy of the partons involved in the primary interaction is unknown and many products are scattered at small angles, escaping detection, it is impossible to measure the total energy of the event. Therefore, if invisible particles (e.g., neutrinos) are produced, their net momentum can only be constrained in the plane transverse to the beam direction. Moreover, the center of mass can be boosted along the beam direction, making it useful to exploit the transverse component of momentum and energy, which are Lorentz-invariant quantities, to describe the kinematics of proton-proton collisions.

The three-momenta of the particles are often described using two components: the longitudinal momentum p_z and the transverse momentum p_T, defined as:

$$|\vec{p_T}| = |p| \sin\theta = \sqrt{p_x^2 + p_y^2}. \tag{2.1}$$

Similarly, the transverse energy is defined as $|\vec{E_T}| = |\vec{E}| \sin\theta$. A commonly used spatial coordinate to describe a particle's angle relative to the z-axis is the pseudorapidity η, defined as:

$$\eta \equiv -\ln\left[\tan\frac{\theta}{2}\right] = \operatorname{arctanh}\left(\frac{p_z}{|p|}\right) \tag{2.2}$$

where θ is the polar angle. As θ approaches zero, η tends towards infinity. Particles produced at high η are referred to as being produced in the "forward" direction. In the limit of relativistic particles, pseudorapidity converges to the definition of rapidity y

$$y \equiv \frac{1}{2}\ln\left(\frac{E + p_z}{E - p_z}\right) \tag{2.3}$$

and differences in pseudorapidity are Lorentz invariant under boosts along the z-axis.
The angular separation between particles is often measured using:

$$\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} \tag{2.4}$$

which is also Lorentz invariant under boosts along the z-axis is the relativistic limit.

2.2.2 Magnet

The superconducting solenoid of the CMS apparatus provides a uniform magnetic field in the axial direction which is essential to bend charged particles and thus determine the particles charge/mass ratio from the track curvature. The magnet is made of refrigerated superconducting nobium-titanium coils, for a total length of 13 m and a diameter of 6 m. It was originally designed to generate a magnetic field of 4 T, but it was later lowered to 3.8 T in order to ensure better longevity [41]. The tracker and the calorimeter detectors are housed within the solenoid, which is surrounded by the return yoke, a 12-sided iron structure that contains and guides the field lines. This return yoke extends up to 14 m of diameter and is composed of three layers. The muon detectors are embedded in the return yoke, interleaved with its layers.
Charged particle trajectories are bent by the magnetic field and the curvature is exploited to determine the electric charge sign and the transverse momentum. In particular, the best momentum resolution σ_{p_T}/p_T achievable for a solenoidal magnetic field, is given by:

$$\frac{\sigma_{p_T}}{p_T} = S \times \frac{8\ p_T}{0.3 \cdot B \cdot R^2} \tag{2.5}$$

where B is the magnetic field intensity in Tesla, S is the sagitta of the particle trajectory and R is the solenoid radius, both measured in meters. In order to maximize the p_T resolution, it is necessary to have large magnetic field intensities along with large sized magnets.
In case of the CMS experiment, the presence of the return yokes allows to establish a strong magnetic field, provided by the large solenoid magnet, while keeping the overall size of the apparatus limited and compact, as the name indicates. Figure 2.5 shows the magnetic field distribution within the CMS apparatus: the nominal intensity of 3.8 T is established in the volume of the inner tracker and calorimeters, while it is reduced to 1.5 - 2 T in the volume defined by the return yoke.

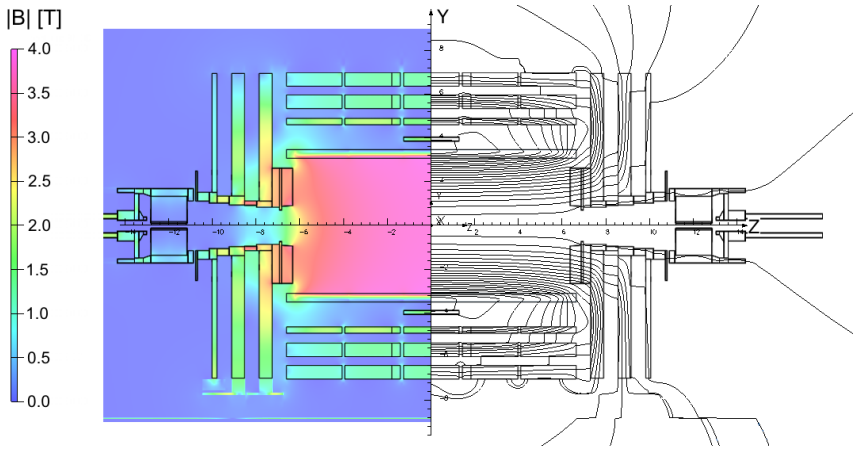


Figure 2.5: Illustration of the longitudinal section of the CMS detector displaying the distribution of the magnetic field intensity (left) and lines (right) [41].

2.2.3 Silicon Tracker

The tracker detector of the CMS experiment is optimized to reconstruct with great precision the trajectories of charged particle within the 3.8 T magnetic field provided by the solenoid. Being the innermost detector of the apparatus, it is positioned closest to the interaction point. The tracker has a cylindrical structure, measuring 5.8 m in length and 2.5 m in diameter. It employs silicon detector technology, chosen for its high radiation hardness, fine granularity and large hit redundancy, which are crucial to perform a good pattern recognition. The CMS tracker structure is optimized to be as lightweight as possible, ensuring that it minimally alters the trajectory of the particles crossing it. A sketch of the tracker in the $r - z$ plane is displayed in Figure 2.6.

The innermost part of the tracker consists of layers of silicon pixel detectors. In the original design, the pixel detectors covers the pseudorapidity range of $\eta < 2.5$. The cylindrical part, called Tracker Pixel Barrel (TPB). is composed of three layers, which have diameters ranging from 8.8 cm to 20.4 cm. In addition, two disks, called Tracker Pixel Endcap (TPE), are located at each end of the TPB. The pixel detector provides a very good resolution on the impact parameter, which is crucial for the reconstruction of secondary vertices, such as those created bu the decay of heavy-flavour hadrons.
Each pixel cells has a size of 100×150 μm^2, offering high granularity. This design provides spatial resolutions of 10 μm in the $r - \phi$ plane (perpendicular to the beam axis) and 20 μm along the z-axis (parallel to beamline).

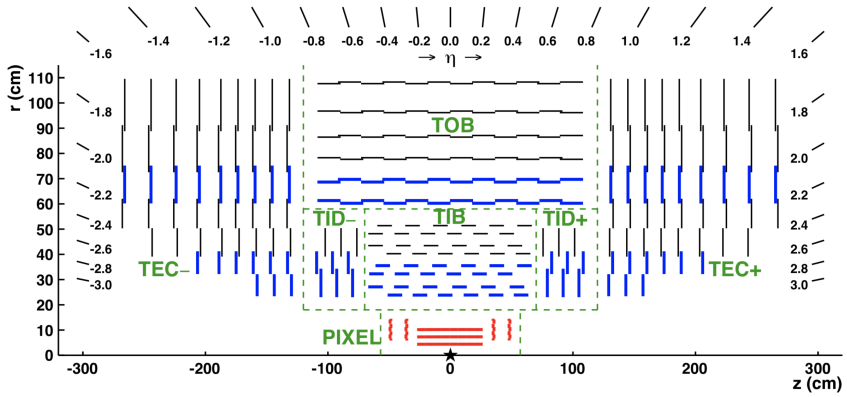


Figure 2.6: Schematic cross section through the CMS tracker in the r-z plane. Due to the tracker symmetric around the horizontal line r = 0, only the top half is displayed [42].

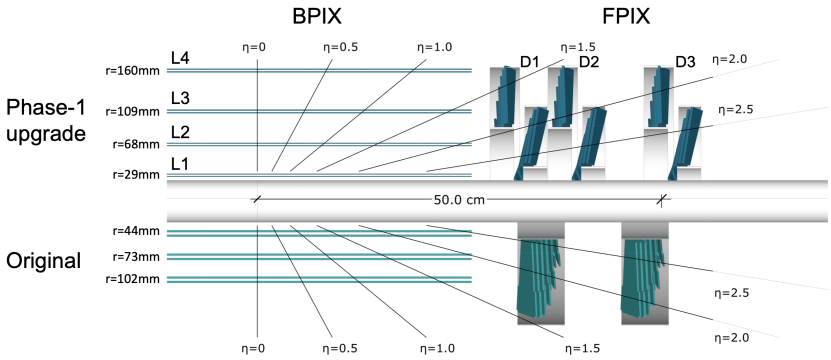


Figure 2.7: Layout of the CMS Phase-1 pixel detector compared to the original detector layout, in longitudinal view [43].

The original CMS pixel tracker was upgraded for Run-2 in order to handle the increased luminosity and center-of-mass energy of the LHC, as well as the higher PU environment. During the extended year-end technical stop of the LHC in 2016/2017 the original CMS pixel detector has been therefore replaced with the *CMS Phase-1 pixel detector*. As shown in Figure 2.7, the Phase-1 detector layout has several improvements over the original design. The CMS Phase-1 pixel detector consists of 4 barrel layers (instead of three from the original design) and 3 endcap disks on each side (instead of two). These additional layers and disks extend the hit coverage in regions of high pseudorapidity. The Phase-1 pixel upgrade also introduced a digital readout chip capable of handling higher rates and a new cooling system, necessary to mantain the detector's performance at high luminosities.

The outermost part of the CMS silicon tracker is composed of silicon microstrip detectors, which are structured with different geometries, each optimized for a

different region of the detector volume. The barrel is organized in two parts:

- the Tracker Inner Barrel (TIB) consists of 4 layers and is characterized by a single-hit resolution of 13-38 μm in the $r - \phi$ direction [42].
- the Tracker Outer Barrel (TOB) consists of 6 layers and ensures a resolution of 18-47 μm in the $r - \phi$ direction and 47 μm in the longitudinal direction, similar to the TIB.

The endcap disks are equipped with concentric rings of silicon strip modules: the Tracker Inner Disks (TID) comprise three disks, while the Tracker Endcaps (TEC) comprise nine disks. The TID and TEC disks are characterized by the same spatial resolution as the TIB and TOB layers, respectively.

2.2.4 Calorimeter

The CMS calorimeter system is a hermetic detector with the goal of measuring the energy of outgoing particles in collision, crucial for reconstructing the total energy of an event and for calculating the missing energy associated with undetected particles like neutrinos.
The energy measurement is based on detecting the energy loss of particles passing through the calorimeter material, where they produce cascades of secondary particles referred to as showers. The showers initiated by particles that interact electromagnetically, such as electrons and photons, are called "electromagnetic showers"; the ones produced by hadrons that interact via the strong force are called "hadronic showers". The CMS calorimeter comprises two subsystems optimized to contain and detect the two types of showers: the electromagnetic (HCAL) and the hadronic (HCAL) calorimeters.

Electromagnetic Calorimeter

The CMS ECAL is designed to detect photons and electrons with good energy resolution. The barrel and the endcap are equipped with 75848 lead tungstate ($PbWO_4$) crystals and cover a large range of pseudorapidity up to $|\eta| < 3$ [45]. Figure 2.8 displays a layout of the ECAL barrel, endcap and preshower subsystems:

- The ECAL Barrel (EB) is made of 61200 crystals which are 23 cm long and have a frontal surface of $\sim 2.2 \times 2.2$ cm^2. They are arranged in modules and provide a coverage of $\Delta\phi \times \Delta\eta = 0.0175 \times 0.0175$.
- The two ECAL Endcaps (EE) are composed by 7324 crystals each and cover the pseudorapidity range $1.48 < |\eta| < 3.0$ and . The single EE

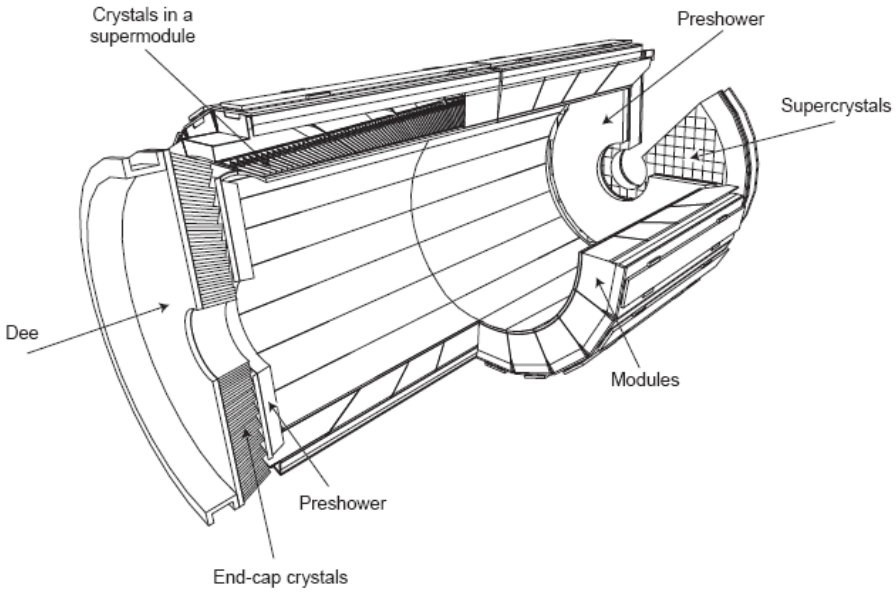


Figure 2.8: Layout of the the CMS ECAL, showing the crystal barrel and endcap detectors, as well as the silicon preshower detector [44].

crystals are characterized by a $\Delta\phi \times \Delta\eta$ that ranges from 0.0175×0.0175 to 0.05×0.05.

- Two pre-shower detector (ES), consisting of two lead radiators and two planes of silicon strip detectors, are positioned in front of the two EE subsystems, respectively. The ES covers the pseudorapidity range $1.65 < |\eta| < 2.6$. It is aimed to identify two close-by photons from neutral pion decay, allowing for $\pi^0 - \gamma$ separation, and to improve the estimation of the direction of photons.

When an electron enters the calorimeter material, it emits photons by bremsstrahlung. These photons, if they have enough energy, can in turn create an electron-positron pair. These secondary electrons and positrons then repeat the process. This way an electromagnetic shower develops and continues until the energy of the photons fall below the threshold for further electron-positron pair production, roughly twice the mass of the electron. To effectively measure the energy of a primary electron or photon, the calorimeter must be large enough to contain most of the electromagnetic shower. The size of this shower depends on two parameters related to the calorimeter material: the radiation length X_0 and the Moliére radius. The longitudinal extension of the shower can be described in terms of the interaction length X_0, defined as the mean distance over which a high-energy electron loses all but $1/e$ of its energy by bremsstrahlung. In practical terms, usually 25-30 X_0 are needed to contain

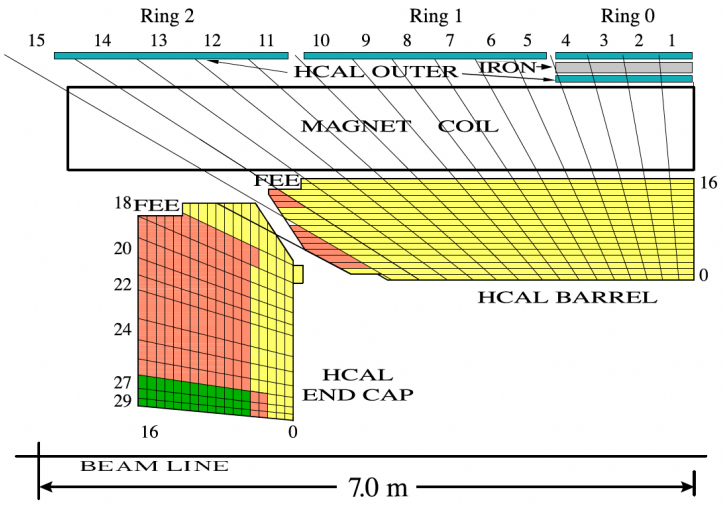


Figure 2.9: A quarter slice of the CMS HCAL detectors. The right end of the beam line is the interaction point. FEE denotes the location of the Front End Electronics for the barrel and the endcap [46].

the shower. The radial spread of the shower is described by the Moliére radius, which is the radius of an ideal cone that would contain the 90% of the energy released by the shower. The $PbWO_4$ crystals are characterized by a short radiation length (0.89 cm) and a small Moliére radius (2.19 cm), ensuring a good shower containment within a limited volume.

The $PbWO_4$ crystals are scintillators, characterized by a scintillation decay time comparable with the 25 ns time interval between two consecutive bunch crossings and an emission peak at 425 nm. Each crystal is equipped with two avalanche photomultipliers (PMT) in the barrel and a single vacuum phototriode in the endcap, for scintillation light detection.

In general the energy resolution of an electromagnetic shower is parametrized as a function of the incident electron/photon energy, E, expressed in GeV:

$$\frac{\sigma_E}{E} = \frac{a}{\sqrt{E}} + \frac{b}{E} + c \tag{2.6}$$

where a, the stochastic term, depends on event to event fluctuations in lateral shower containment, photo-statistics and photodetector gain; b, the noise term, depends on the electronic noise and event pile-up; and c, which is the constant term, depends on non-uniformity of the longitudinal light collection, leakage of energy from the rear face of the crystal and the accuracy of the detector inter-calibration constants. For the CMS ECAL, the stochastic and constant terms are found to be $2.8\%/\sqrt{E(\text{GeV})}$ and 0.3%, respectively [44].

Hadron Calorimeter

The CMS HCAL [47] is a sampling calorimeter which utilizes alternating layers of brass as absorber and plastic scintillator as active material. It comprises, as usual, a barrel (HB) and two endcap (HE) subsystems, as schematized in Figure 2.9. It can be observed that a minor part of the HCAL, HCAL Outer (HO), is placed outside the magnet solenoid: this way, the magnet itself and the return yoke serve as absorbers, enhancing the detection of high energy hadrons. A forward calorimeter (HF), instead, increases the geometrical acceptance in the range of pseudorapidity $2.9 < |\eta| < 4.1$.

Hadronic showers are the result of many different processes and are in general much more complex than electromagnetic showers. When a hadron with an energy above around 5 GeV interacts with the calorimeter material, both inelastic and elastic scattering processes occur between the incoming particle and the nucleons in the atomic nuclei of the material. The high-momentum transfer in these interactions leads to the production of a large number of secondary hadrons, which in turn continue to interact, creating a cascade of further interactions and energy loss. The hadronic shower gradually ceases as the energy of the secondary hadrons decreases, eventually stopping through ionization energy loss or nuclear absorption. Neutron pions produced in the cascade, decay in two photons, which initiate electromagnetic showers. For this reason, a hadronic shower usually contains an electromagnetic component, which accounts for the 20-30% of its total energy.

In hadronic cascades, the lateral spread of the shower is primarily determined by large transverse momentum transfers in nuclear interactions, as opposed to the multiple scattering of charged particles that drives the lateral development of electromagnetic showers. This distinction leads to hadronic showers generally being broader and more complex in structure than electromagnetic ones. Similar to electromagnetic showers, the longitudinal extension of a hadronic shower can be described in terms of the interaction length, which, in the case of hadrons, is defined as the mean distance a hadron travels in a material before undergoing an inelastic interaction with a nucleus. The hadronic interaction length depends on the atomic number (A) of the detector material and is typically much larger than the electromagnetic interaction length. Because of that, hadronic showers penetrate much deeper into the calorimeter than electromagnetic showers. To effectively confine hadronic showers, dense, high-A materials are typically used in hadronic calorimeters. In the case of

the CMS HCAL, brass is chosen as the absorber material, since it has a short interaction length (∼15 cm) and is non-magnetic. The energy is measured by plastic scintillators equipped with wavelength shifting fibres and photodiodes.

The HB is structured as 36 identical azimuthal wedges, each made up of brass absorber plates that are positioned parallel to the beam axis, interleaved with plastic scintillators segmented in both ϕ and η directions with a granularity of $\Delta\phi \times \Delta\eta = 0.087 \times 0.087$.
The HO complements the HB in the central region: their combination offers nearly 11 hadronic interaction lengths in total.
The HE, which covers the forward region, is similarly structured with 79-mm-thick brass plates. These plates are separated by 9-mm gaps that accommodate scintillators, and the HE granularity is $\Delta\phi \times \Delta\eta = 0.17 \times 0.17$.
Finally, the Hadronic Forward (HF) calorimeter extends the pseudorapidity coverage up to $|\eta| < 5$. It is constructed using steel and quartz fibers aligned parallel to the beam. Charged particles produce Cherenkov light in the quartz fibers, which is then collected by PMT tubes. The HF detector is housed within a radiation shielding which consists of layers made up of layers of 40 cm thick steel, 40 cm of concrete, and 5 cm of polyethylene for neutron shielding.
The signals from photodiodes or PMTs are processed by being integrated and digitized through a custom-designed chip, which is located directly on the detector.

The hadronic energy resolution follows a similar parametrization to that used for the electromagnetic calorimeter:

$$\frac{\sigma_E}{E} = \frac{a}{\sqrt{E}} + b \tag{2.7}$$

where E is in GeV, a is a stochastic term and b is the constant term, which becomes dominant at high energies. For the combined HCAL and ECAL system, the hadronic energy resolution is measured to be $a = 84.7\% \pm 1.6\% \sqrt{\text{GeV}}$ and $b = 7.4\% \pm 0.8\%$. In the HF, the resolution parameters are found to be $a = 198\% \sqrt{\text{GeV}}$ and $b = 9\%$ [46].

2.2.5 Muon System

The muon system is the outermost detector of the CMS apparatus and it is housed within the three layers of the return yoke. This system relies on gaseous detectors of four different technology: Drift Tubes (DT), Cathode Strip Chambers (CSC), Resistive Plate Chambers (RPC) and Gaseous Electron Multipli-

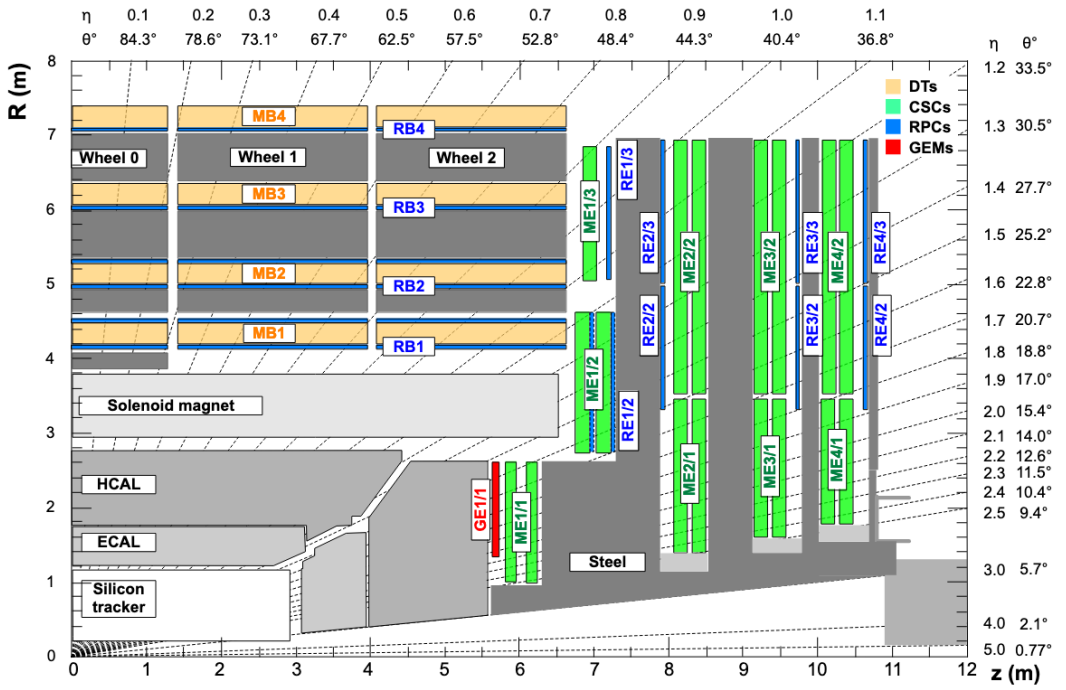


Figure 2.10: One quadrant of the CMS detector in its Run 3 configuration, with the Muon detectors in colour.

ers (GEM). In the layout of Figure 2.10, which displays the transverse section of a CMS quadrant, the confguration of the muon chambers is shown, with different colors for the different technologies employed. The barrel system extends up to $|\eta| < 1.2$ and the endcap up to $|\eta| < 2.4$, which corresponds to the geometrical acceptance of the muon system.

In contrast to calorimeters, where particles are destructed in order to be measured, muons can pass through the entire CMS detector volume and reach the muon chambers. High-energy muons lose energy primarily through repeated elastic scattering with charged particles and nuclei, allowing them to travel long distances even in dense materials. In the muon chambers, each hit position is recorded, and the full muon track is reconstructed by combining hits from multiple detector layers. The return yokes provide a magnetic field of ~ 2 T, enabling precise momentum measurement in a compact setup. For this reason, the combined set of muon chambers is often referred to as the Muon Spectrometer.

However, in addition to muons, also other particles can reach the muon system, such as punch-through hadrons from the inner detectors and neutrons from particle showers or gaps in the HCAL's forward shielding. Neutron exposure, in particular, can be a risk to the gaseous detector's durability. Moreover,

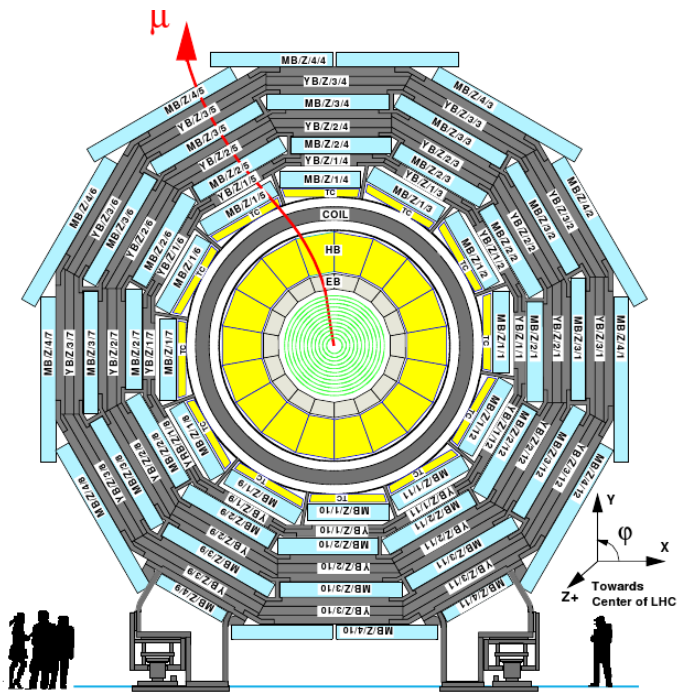


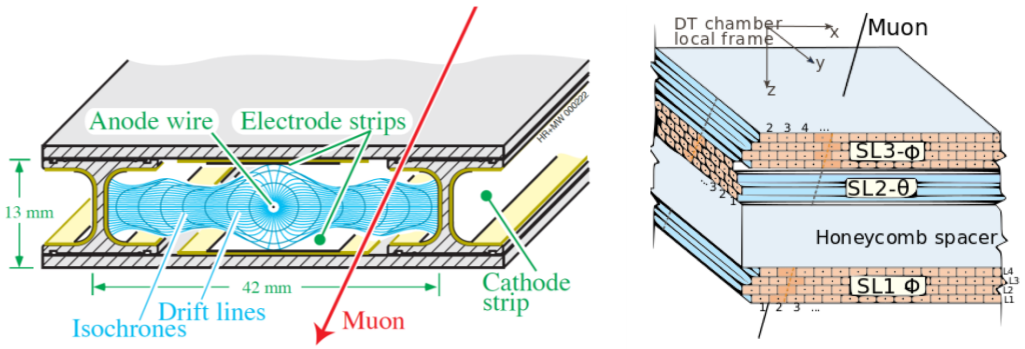
Figure 2.11: Layout of the CMS barrel muon DT chambers in one of the 5 wheels. The chambers in each wheel are identical with the exception of wheels -1 and +1 where the presence of cryogenic chimneys for the magnet shortens the chambers in 2 sectors.

low energy neutrons can activate surrounding materials, leading to radiative de-excitations of the nuclei and photon emissions that generate background signals.
Therefore, accurate measurement of the muon track, with good spatial resolution and ample hit redundancy, is crucial for background rejection and precise muon final-states identification. A good time resolution is required too, as the muon system serves also as trigger and a fast response time is needed in order to unambiguously assign a muon trigger candidate to the right bunch crossing. The various detector technologies employed in the muon system are described in the next sections.

Drift Tubes (DT)

The DTs are installed within each of the four concentric cylindrical stations forming the barrel section of the CMS muon system. This detector technology was selected for the barrel area due to the low rates expected and the relatively weak magnetic field in this region, as illustrated in Figure 2.5.

The 250 DT chambers are arranged across five wheels, covering a pseudo-



(a) (b)

Figure 2.12: Left: Section of a drift cell of a Drift Tube detector, showing the anode wire and the cathode strips, as well as the drift lines and the isochrones. Right: Structure of a DT Chamber with three *superlayers* composed of 4 layers each [48].

rapidity range of up to $|\eta| < 2.1$. Figure 2.11 illustrates the configuration of one of these barrel wheels. The fundamental element of the DT system is depicted in Figure 2.12a. It is essentially a rectangular drift cell bounded by cathode aluminum strips on the sides and aluminum plates on the top and bottom, with an anode wire centered along the cell's symmetry axis. The cell is filled with an Ar-CO_2(85% − 15%) gas mixture, which ionizes as charged particles pass through, generating electron-ion pairs that then drift toward the electrodes. The cathode electrodes are designed to shape the electric field within the cell, ensuring a uniform drift velocity across its width. This enables accurate conversion of drift time into distance.
The maximum drift distance is $\sim$ 21 mm, which translates in a maximum drift time of about 380 ns in Ar-CO_2(85% − 15%) gas mixture, resulting in a single-wire resolution of approximately 200 μm.

Each DT chamber is composed of 2 or 3 superlayers, with each superlayer consisting of 4 layers of DT cells that are staggered by half a cell width, as illustrated in Figure 2.12b. The superlayers are conFigured with wires oriented either parallel or transverse to the beam line, allowing for precise measurements of the muon position in both the $r - \phi$ plane and along the z-axis. The spatial resolution achieved in each DT chamber, by combining all layer measurements, is approximately 80-120 μm in the $r - \phi$ plane and 130-390 μm along the $r - z$ plane [48]. The design specifications for the DT subdetector set a time resolution requirement of 5 ns. However, tests have shown that the chambers and their associated electronics surpass this target, achieving a time

resolution of under 3 ns for high-p_T muons [48].

Cathode Strip Chambers (CSC)

The CSCs are installed on four disks (stations) within each endcap section, between the iron disks of the return yoke, which also function as shielding. The choice of CSC technology for the endcap, as opposed to DTs, is driven by the the higher muon rates and the intense and inhomogeneous magnetic field in that region. The muon chambers in the endcap feature a trapezoidal design and are arranged to ensure complete coverage in the $r - \phi$ plane. The CSC is a type of multi-wire proportional chamber where the cathode plane is divided into strips that run perpendicular to the direction of the wires. When an ionizing particle crosses the gas volume, it generated electron-ion pairs. The electrons then drift toward the anode wires and, thanks to the strong electric field established within the chamber, they further ionize the gas, leading to an avalanche effect. This avalanche induces a distributed charge on the cathode plane, which is segmented into strips. By interpolating the fractions of charge collected by the strips, the position of the particle track along the wire can be accurately reconstructed, as illustrated in Figure 2.14.

Each CSC trapezoidal chamber is constructed from six wire planes, with the cathode strips arranged radially to facilitate the measurement of the $r - \phi$ coordinates. Each plane consists of 80 cathode strips, with a pitch that varies between 2.2 and 4.7 mrad in the ϕ direction. The orthogonal anode wires are spaced from 2.5 to 3.16 mm apart. These chambers are filled with a gas mixture of CO_2, Argon and CF_4 in the proportion of 40%-50%-10%.
In each endcap muon station (ME1-ME4 in Figure 2.10), two rings of chambers are installed, with the first station (ME1) containing three rings. Each ring is composed of either 18 or 36 trapezoidal chambers.
The spatial resolution achieved by the CSC subsystem varies from approximately 70 μm in the ME1/1 chamber to about 210 mum in the ME4/1 chamber, while the time resolution is around 3 ns [48].

Resistive Plate Chambers (RPC)

The RPCs are installed in both the barrel and endcap CMS sections, and they complement the DTs and CSCs with a very fast response time, crucial for trigger purposes. The RPC is a type of gaseous detector with planar geometry, composed of two bakelite planes coated with a thin film of graphite, a positively-charged anode and a negatively-charged cathode, separated by a 2

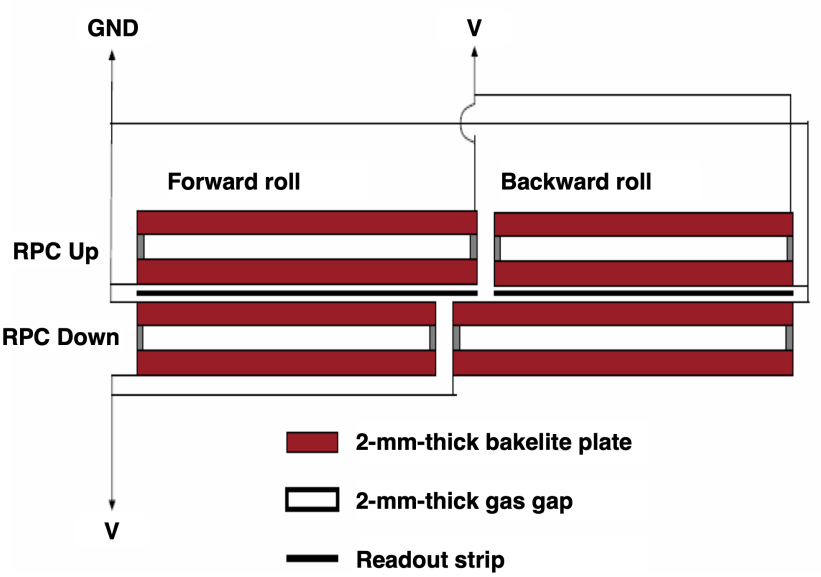


Figure 2.13: Schematic view of a RPC [48].

mm wide gas gap (a mixture of freon, isobutane, sulphur hexafluoride and water vapor). Each CMS RPC chamber is composed of two such gas gaps facing a shared layer of readout strips, as illustrated in Figure 2.13. A high voltage of approximately 9.6 kV is applied on the outer graphite coated surfaces of the bakelite plates. When a charged particle crosses the gas volume, it initiates an ionization cascade, resulting in an avalanche that induces a signal on the readout strips.

In the barrel, RPCs chambers follow the same segmentation as the DT superlayers, as shown in Figures 2.10 and 2.11. Six layers of them are embedded within the barrel iron yoke: four of these layers situated in the inner and outer sides of the first two stations of the DT chambers, while the other two are located on the inner side of the third and fourth stations of the DTs. Each barrel RPC chamber is equipped with 96 readout strips running parallel to the beam line, with pitch ranging from 2.1 cm (in RB1) to 4.1 cm (in RB4).
In the endcap, the RPCs are organized across four stations, similarly to the CSCs. The chambers have a trapezoidal shape and are equipped with readout strips arranged radially characterized by a length spanning from 25 cm in the most forward RE1 chamber to 80 cm in the RE4 chamber and a strip pitch ranging The strip pitch also ranges from 0.7 up to 3 cm. The RPC performance benefits from a spatial resolutions of approximately 1 cm and a time resolution around 2 ns.

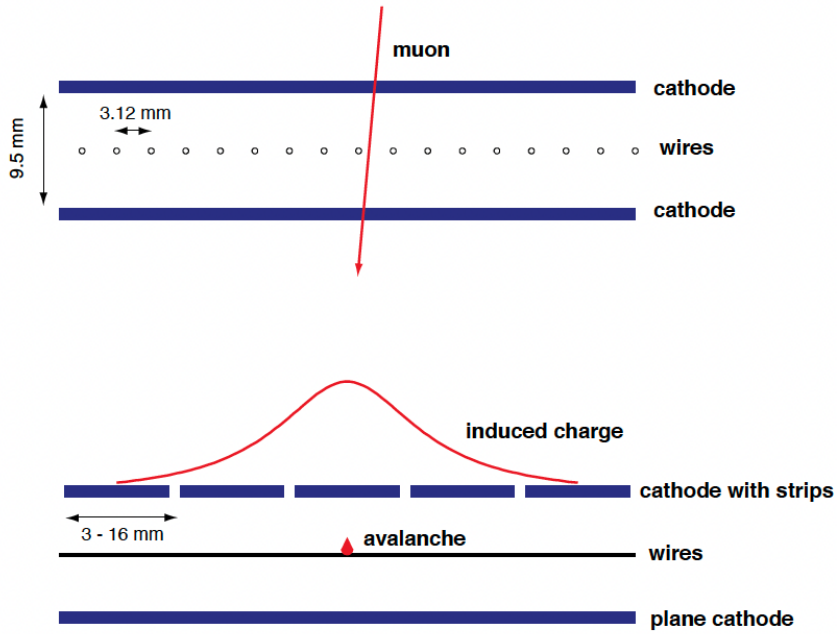


Figure 2.14: Principle of coordinate measurement with a cathode strip chamber. Top: crosssection across wires. Bottom: across cathode strips. Close wire spacing allows for fast chamber response, while a track coordinate along the wires can be measured by interpolating the signals induced on the strips [49].

2.2.6 Gaseous Electron Multiplier (GEM)

During the future high-luminosity phase of the LHC, the muon hit rate in the forward region is expected to reach a value of 5 kHz/cm^2 in the first muon layer. Such a high rate necessitates, in the forward region, of a detector resistant to radiation, with high rate capability and capable of minimize the number of misidentified tracks in order to keep the trigger rate under control [50].
Therefore, in order to enhance track reconstruction and trigger capabilities of the endcap muon system, large-area triple layer GEM detectors were installed in the CMS endcap before the start of Run-3. This single station, called GE1/1, covers a pseudorapidity range $1.55 < |\eta| < 2.18$ and is the first of three endcap rings that are foreseen for the HL-LHC upgrade. A sketch of the GE1/1 station is displayed in Figure 2.15.

Each GE1/1 endcap ring contains two layers of 36 triple GEM chambers positioned just in front of the first CSC station, ME1/1, with each chamber spanning a 10° sector in azimuth. The chambers are manufactured in two different sizes: the odd-numbered chambers in GE1/1 are slightly longer to

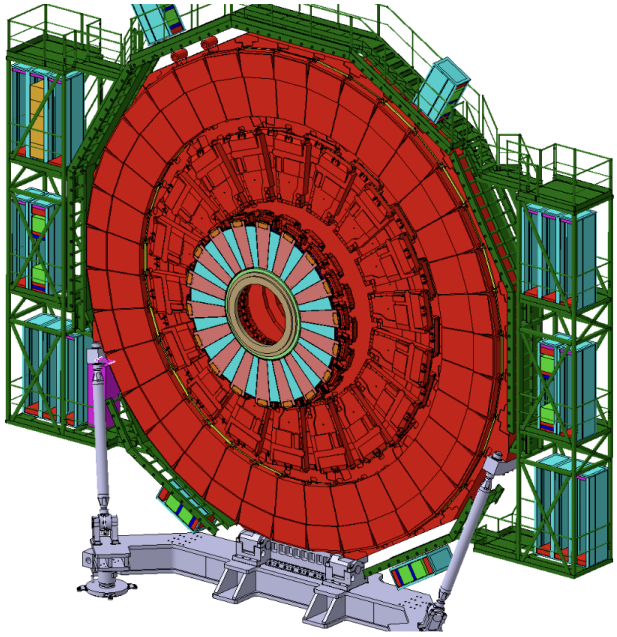


Figure 2.15: Sketch of GE1/1 system of one endcap [50].

optimize pseudorapidity coverage while fitting within the spatial limitations set by the support structure, as illustrated in 2.16. The GE1/1 subsystem is located between 566 and 574 cm in z and spans a radial range of 145 to 230 cm.

The CMS triple GEM detector, a micro-pattern gas detector, consists of four gas layers separated by three GEM foils. At the base of this GEM assembly is a printed circuit board that holds the drift electrode. The top layer, the readout board, has radially oriented strips along the chamber's long side, with strip pitch ranging from 0.6 to 1.2 mm. The readout board is segmented into up to 8×3 $\eta - \phi$ partitions, each equipped with 128 strips.

The basic unit of a GEM detector is the GEM foil, a thin polyimide sheet coated on both sides with copper and etched with a uniform array of fine holes. The polyimide layer is typically 50 μm thick, with a 5 μm copper coating on each side. The hexagonal hole pattern has a pitch of about 140 μm, with holes generally taking on a biconical shape—inner diameters around 50 μm and outer diameters about 70 μm, though specific dimensions vary by etching method. When a voltage difference is applied across the GEM foil, charge multiplication takes place within the holes. In a single GEM detector, most electrons are driven toward the anode by the induction field, while some are collected at the bottom of the GEM. By cascading multiple GEM foils, each operating at a lower gain, the avalanche spreads over multiple holes,

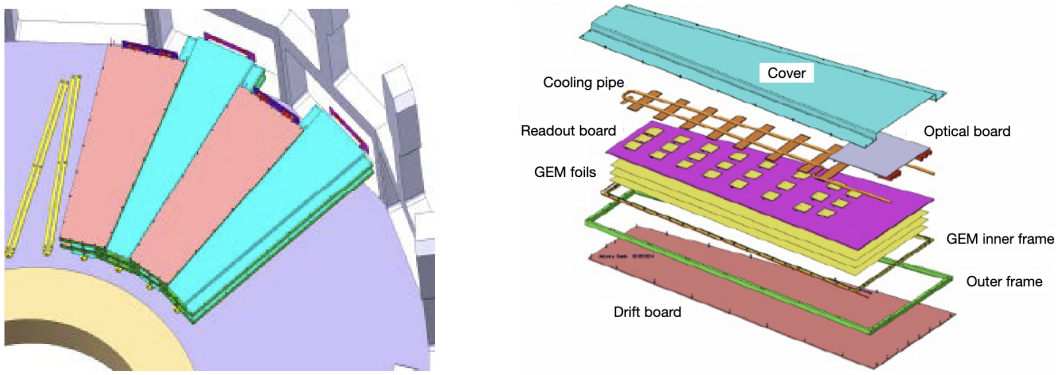


Figure 2.16: Left: layout of the GE1/1 chambers along the endcap ring, indicating how the short and long chambers fit in the existing volume. Right: blowup of the trapezoidal detector, GEM foils, and readout planes, indicating the geometry and main elements of the GEM detectors [51] [50].

thus limiting charge density and reducing discharge risks. In CMS, the triple GEM foils operate at stable gains between 10^4 and 10^5. The GEM detectors in the GE1/1 station operate with an Ar/CO_2 gas mixture in a 70:30 ratio. In GE1/1, pair of chambers are matched to form a "super-chamber", providing two measurement planes and maximizing the detection efficiency for the station.
GEM technology is well-suited for the high particle rates of the forward region, tolerating rates up to hundreds of kHz/cm^2 while providing precise spatial and timing resolution, approximately 250–500 μm and under 10 ns per layer, respectively. When combined in the GE1/1 station, these layers achieve a spatial resolution of about 100 μm.

2.2.7 The CMS Trigger System

At the LHC proton bunches cross at a rate of 40 MHz and each bunch contains 10^{11} protons, resulting in a proton collision rate of 10^7-10^9 Hz. It is not possible to store the result of each collision to disk: considering a 2 MB size per event and a collision rate of 1 GHz, a storage capacity of approximately 1 PB/s would be required. With the current technologies, CERN is capable of storing a few tens of PB per year from the LHC.
For this reason, the CMS experiment is equipped with a two-level trigger system, responsible for selecting events that are interesting from the physics point of view, within a very broad research program. The first level trigger (L1T) is a hardware based system, designed to make very fast decisions (4 μs of latency) by exploiting only information reconstructed by the calorimeters and

the muon detectors [52]. The L1T reduces the rate from 1 GHz to 110 kHz, which is further handled, in the next stage, by the high-level trigger (HLT), a software-based system which runs the full event reconstruction customized for fast processing. In Run-3, the HLT has reached a final output rate of 2.6 kHz. Additionally, the HLT stores extra samples, under the name of "parking" datasets, which are reconstructed with a certain delay and only when the resources are not needed for the main reconstruction, at a rate of 3 kHz. Another strategy in the trigger reconstruction, called "scouting", allows us to store 30 kHz of HLT-reconstructed data and 40 MHz of L1T-reconstructed data, without undergoing offline reconstruction.

L1 Trigger

The L1 trigger selects events on the basis of information reconstructed by the calorimeters and the muon system, as schematized in Figure 2.17. In particular, calorimeters and muon detectors reconstruct the so-called trigger primitives (TPs), which are basically energy and position measurements. The L1T calorimeter and muon systems then reconstruct jets, electrons, photons, hadronically decaying τ leptons, and muons, and the calorimeter trigger computes energy sums. Finally, the L1T global trigger (GT) makes the final decision, called "L1 Accept" (L1A), by applying selections on the multiplicity and kinematic quantities of these objects, as well as the proximity to each other, timing information and beam presence. A complete set of L1 selections dedicated to a specific region of the phase space is referred to as "L1 seed" and the so called "L1 menu" gathers together all the L1 seeds to be used during p-p collisions.
The L1 latency, which is the time needed to perform this fast reconstruction and to decide whether to read out the full-event information, is 4 μs for CMS and it includes the needed latency for taking the data out of the detector to the counting room and sending the L1A back to the on-detector electronics.

High Level Trigger

The L1T output rate is further reduced to 2.6 kHz by the HLT, a software-based trigger system which exploits the full event reconstruction to select events considered interesting for the wide CMS physics program. Since Run-3, the HLT has started to make use of graphical processing units (GPUs) in the farm. Many reconstruction algorithms were implemented to run on both central processing units (CPUs) and GPUs: they are executed on a GPU if available, otherwise on a CPU.

Both the L1 and the HLT output rates can be tuned by defining a prescale factor: in this case, not all the events passing the specific selection criteria are selected, but a random subset. A complete sequence of L1 and HLT selection criteria, including the prescale, is called "trigger path".
The HLT selection consists of two sequential steps: first, the full information from calorimeters and muon detectors is exploited reducing the event rate by approximately one order of magnitude; then, also the information from the silicon tracker is reconstructed and further selections are applied.
HLT algorithms are very versatile, and therefore customized HLT paths dedicated to specific signals can be implemented. The event reconstruction performed at the HLT is mainly based on the same algorithms used for offline object reconstruction, but optimized for fast processing.
After the HLT makes the final decision, the data is initially stored locally on disk in raw data format and then transferred to the Tier-0 computing centre, which perform offline reconstruction and store the data permanently.
Offline reconstruction consists of a large number of complex and sophisticated algorithms, described in Chapter 3, aimed at identifying the different particles created in the proton-proton collisions.
For storage, different datasets are defined on the basis of different classes of trigger paths: for instance, the "JETMET" dataset collects data selected by trigger paths applying primarily requirements on jets and missing energy.

Run-3 trigger strategies: Scouting and Parking

In order to increase the rate of stored events, the CMS Collaboration introduced two trigger strategies, called "scouting" and "data parking", which were optimized and extended for the Run-3 physics program [53]. The total number of events that can be selected and reconstructed is limited by several factors, including the finite bandwidth available to store data permanently and the affordable rate of the offline event reconstruction.
The idea under the scouting strategy is that it is possible to store events at a higher rate by reducing the event size. Two different scouting strategies were implemented in order to store events reconstructed respectively at the L1T and at the HLT.
The data parking strategy, instead, is a different approach that allows us to increase the storage rate to disk by delaying the offline data reconstruction.

- **L1 Scouting System**
  For Run-3, a new hardware system was added to the L1T in order to record, without the trigger filter, all the events as they are reconstructed

by the L1T, at a rate of 40 MHz. These objects are not submitted to the offline reconstruction, having thus a poor resolution. Such a large event rate is a great opportunity for studying rare physics processes.

- **HLT data scouting**
  The HLT data scouting was originally implemented in Run-1 and then extended for Run-2 and Run-3. In Run-3, events reconstructed at HLT for the scouting dataset have a size of 7 kB, to be compared to the full raw event size of 1 MB, allowing for recording events at a rate of 30 (22) kHz in 2022 (2023).

- **Data parking**
  Data parking consists of sending events selected by the HLT directly to tape for storage and submitting them to offline reconstruction only when resources are available. This allows to store additional data at a rate which is not constrained by the limited capacity of the prompt reconstruction system, but only by the bandwidth of the CMS DAQ and by the amount of tape storage space. Data parking was originally introduced in Run-1, with a total rate of 300-350 Hz, and it included trigger paths dedicated to VBF topology and Higgs boson measurements, SUSY and dark matter searches and some dimuon triggers for B physics measurements.
  In preparation of 2018, the last year of Run-2 data taking, the so-called "B parking" strategy was introduced. It included paths dedicated to the search for B flavour anomalies and collected data at an average rate of 2-3 kHz.
  In Run-3, the HH and VBF parking strategies were developed. The HH parking includes paths dedicated to the search for $HH \rightarrow 4b$ and $HH \rightarrow 2b2\tau$ (from 2024). The VBF parking dataset, deployed online during the 2023 data taking, includes inclusive and exclusive trigger paths dedicated to the VBF topology, for a total rate of approximately 1 kHz.
  As it will be illustrated in details in Chapter 4, the new trigger I developed specifically for the search for $VBF\ H \rightarrow c\bar{c}$, has been included in the VBF parking dataset and, therefore, part of the data analysed in this work are collected by exploiting the CMS parking strategy.

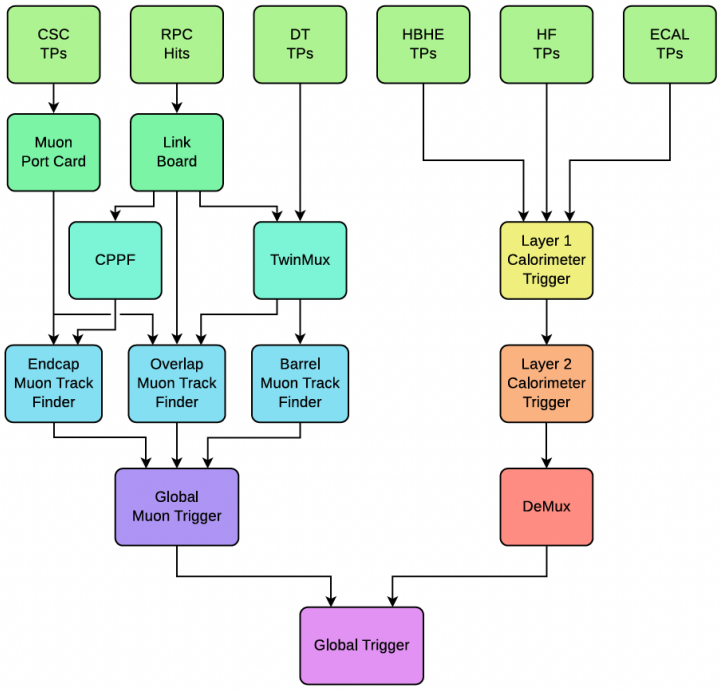


Figure 2.17: Overview of the CMS L1 trigger system. Trigger primitives (TP) from the forward (HF) and barrel (HCAL) hadronic calorimeters, and from the electromagnetic calorimeter (ECAL), are processed by the Calorimeter Trigger System and sent to a demultiplexing card (DeMux). Energy deposits (hits) from the resistive-plate chambers (RPC), cathode strip chambers (CSC), and drift tubes (DT) are processed either via a pattern comparator or via a system of segment- and track-finders and sent onwards to a global muon trigger (GMT). The information from the DeMux and GMT is combined in a global trigger (GT), which makes the final trigger decision. This decision is sent to the tracker, ECAL, HCAL or muon systems via the trigger, timing and control (TTC) system. The data acquisition system (DAQ) reads data from various subsystems for offline storage [52].

Chapter 3

Object reconstruction at the CMS experiment

The final state of the VBF $H \to c\bar{c}$ process involves two jets from the hadronization of charm quarks originating from Higgs boson decay (charmed jets), along with two jets resulting from the hadronization of quarks emitted in the VBF process. At the CMS experiment, jets are reconstructed using the anti-kt algorithm [54], which clusters Particle Flow [55] objects. A critical step in this search is distinguishing charmed jets from beauty and light-flavour jets. To achieve this, the state-of-the-art jet flavour tagging algorithm, ParticleNet [26], is utilized. This sophisticated machine learning algorithm exploits information related to tracks, secondary vertices and Particle Flow objects associated with the jets to predict the flavour of the originating quark.
This chapter will provide a thorough description of the algorithms and techniques essential for reconstructing the main final state objects in this search. Finally, a brief description of the specific acquisition and reconstruction of the 2023 data, which are used in this work, is provided.

3.1 Tracks and primary vertex

At the CMS experiment, track reconstruction relies primarily on the Combinatorial Kalman filter [56].
A track seed is initially built from either two hits and an estimated beam-spot or three hits, in order to reduce the number of possible hit combinations to be processed (seed generation). The track seeds allow to obtain a first estimate of the track parameters, essential for the track building process. Then, the Kalman filter extrapolates the seed trajectories along the expected flight path of a charged particle and assign additional hits to the track candidate (track

finding) [57]. After the track finding step, the track parameters are fitted by taking into account all the hits associated to the track (track fitting). Tracks are then selected according to the normalized χ^2 value (track selection). This helps to discard tracks incorrectly reconstructed from hits of different particles. This tracking sequence-seed generation, track finding, track fitting, and track selection-is repeated iteratively. Prompt tracks with high p_T, which are the easiest to find, are reconstructed at the first iterations, while later iterations target tracks more difficult to reconstruct. At the end of each iteration, hits already associated to a reconstructed track are discarded (masked) to reduce the combinatorial complexity in the next iterations [58]. Additionally, the Cellular Automata (CA) track seeding algorithm developed for parallel architectures is employed to further enhance the efficiency of track reconstruction [58]. The CA algorithm creates hit doublets (cells) for each pair of layers and then compute the compatibility between adjacent cells as schematized in the sketch of Figure 3.1.

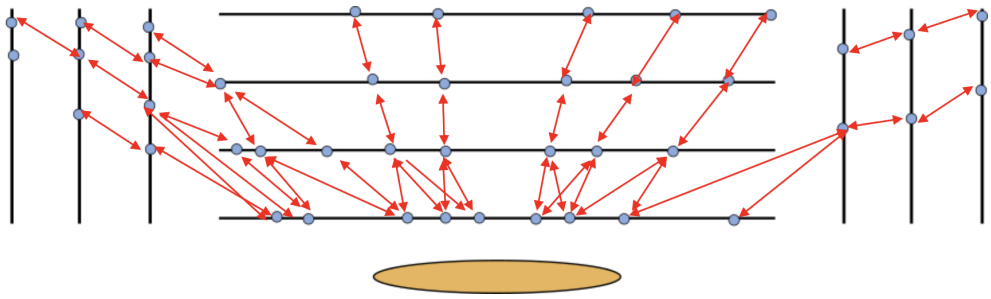


Figure 3.1: sketch of the Cellular Automata track seeding. [55].

This approach facilitates parallel processing of track candidates, thereby improving both the speed and robustness of pattern recognition. The integration of the Combinatorial Kalman filter and Cellular Automata provides a comprehensive framework for achieving high-quality track reconstruction in the CMS detector.

For each event, a large number of proton-proton interaction vertices is originated, which includes the "signal" vertex, called primary vertex (PV), and the vertices from pileup collisions. The PV is reconstructed from the track collection following three main steps: selection of the tracks, clustering of the tracks compatible with the same interaction vertex and fitting for determining the vertex position. For each vertex reconstruction, a weight is assigned to the tracks used for that vertex, based on their compatibility with the vertex. At the end, the vertex with the highest summed track weights is selected as the

PV. The resolutions in x and z are estimated to be, in minimum-bias [1] events, less than 20 and 25 μm respectively, for primary vertices reconstructed with at least 50 tracks [57], as it can be observed in Figure 3.2 (red). Overall, a better resolution is observed in a jet-enriched sample (black), produced by requiring each event to have a reconstructed jet with transverse energy $E_T > 20$ GeV. Having significantly higher mean p_T, these tracks benefit from a better resolution.

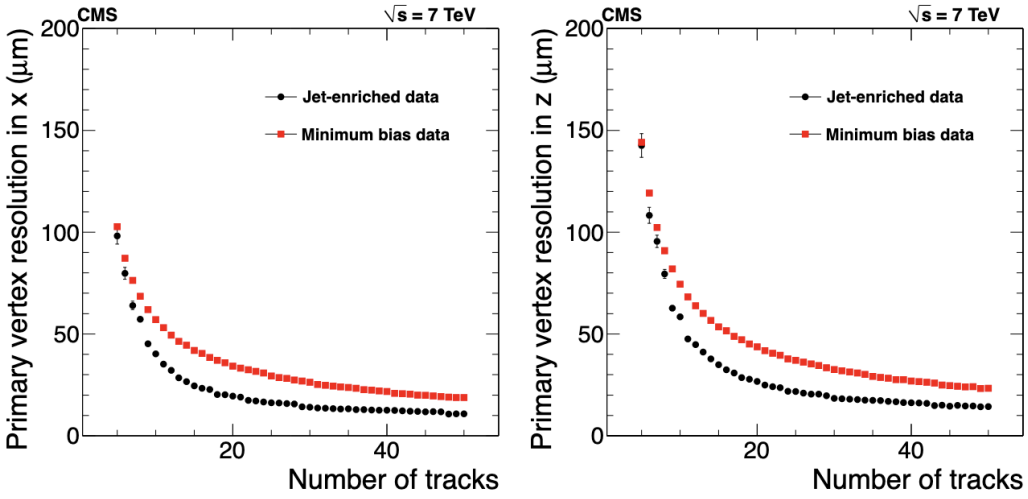


Figure 3.2: Primary-vertex resolution in x (left) and in z (right) as a function of the number of tracks at the fitted vertex, for two kinds of events with different average track p_T values [57].

Figure 3.3 shows the resolution, as a funtion of p_T, of d_0 and z_0, defined as the coordinates of the impact point in the radial and z directions. These results are evaluated on simulations of isolated muons with with p_T = 1, 10, and 100 GeV, in different η partitions.

[1] The Minimum-bias (MB) triggers collect events from non-single diffractive inelastic interactions by requiring a minimum number of hit or tracks in the pixel detectors or towers in the Hadron Forward calorimeter.

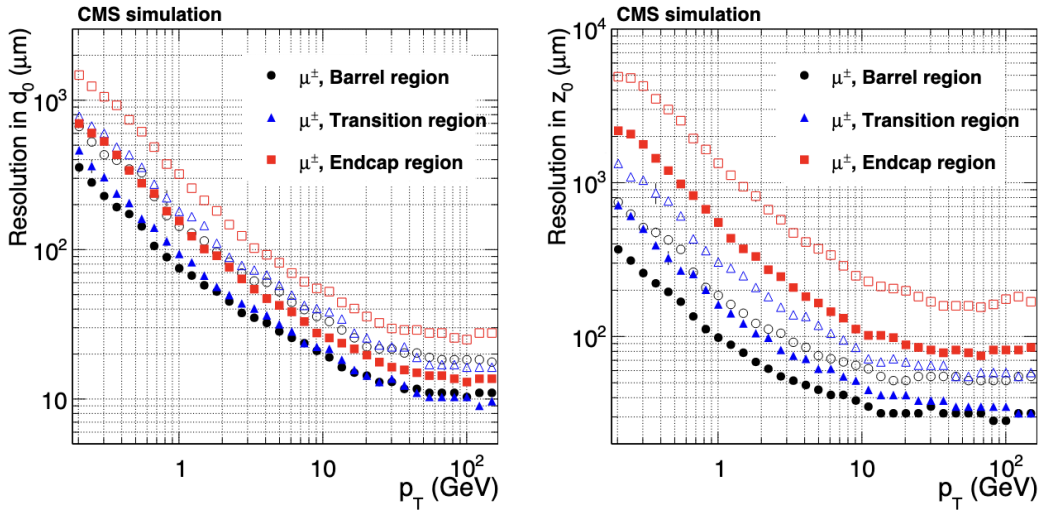


Figure 3.3: Resolution, as a function of p_T, of d_0 (left) and z_0 (right), for single isolated muons in the barrel, transition, and endcap regions, defined by η intervals of 0–0.9, 0.9–1.4 and 1.4–2.5, respectively. For each bin in p_T, the solid (open) symbols correspond to the half-width for 68% (90%) intervals centered on the mode of the distribution in residuals [57].

The performance of tracking in 2023 data has been studied using the ZeroBias [2] dataset, by selecting tracks with high quality and p_T larger than 1 GeV and compared with MC simulation. For instance, Figure 3.4 shows the distribution of the significance of the 3D impact parameter with respect to the PV (3DSIP). A good agreement between data and MC simulation, at a level of 10%, is achieved.

[2]The Zero-bias trigger collects randomly events using only the information on the beam-beam coincidence.

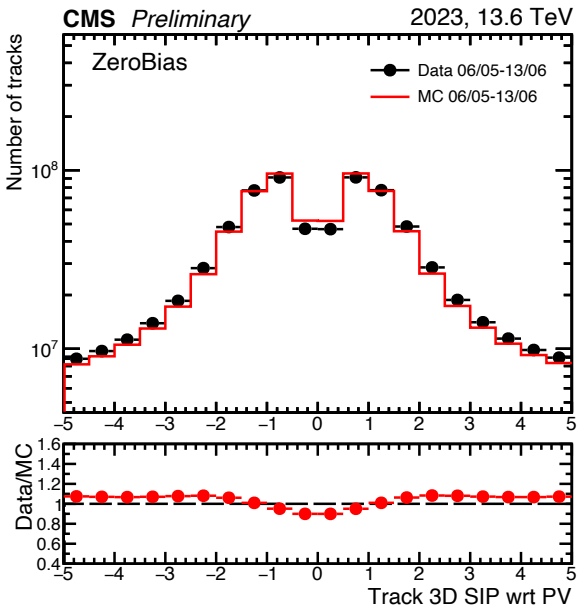


Figure 3.4: Distributions of the significance of 3D impact parameter, abbreviated as 3DSIP, with respect to the primary vertex for tracks with high quality and p_T larger than 1 GeV [59].

3.2 Particle Flow

Particle Flow (PF) reconstruction is a sophisticated technique used in the CMS experiment to achieve high-quality identification and reconstruction of particles produced in p-p collisions. Figure 3.5 shows a sketch of different particle interactions with the subsystems of the CMS detector. In principle, before the introduction of the Particle Flow algorithm, jets can be reconstructed by exploiting only the information collected by the hadron calorimeter, while electrons and photons can be reconstructed primarily by the electromagnetic calorimeter, and muons can be reconstructed solely by the muon chambers. However, a significant improvement in event description can be achieved by correlating the basic elements from all the sub-detectors and combining all the information collected to identify each final-state particle and reconstruct their properties. This integrated approach, known as Particle Flow, leverages the strengths of each sub-detector to provide a more accurate and comprehensive picture of the collision events [55].

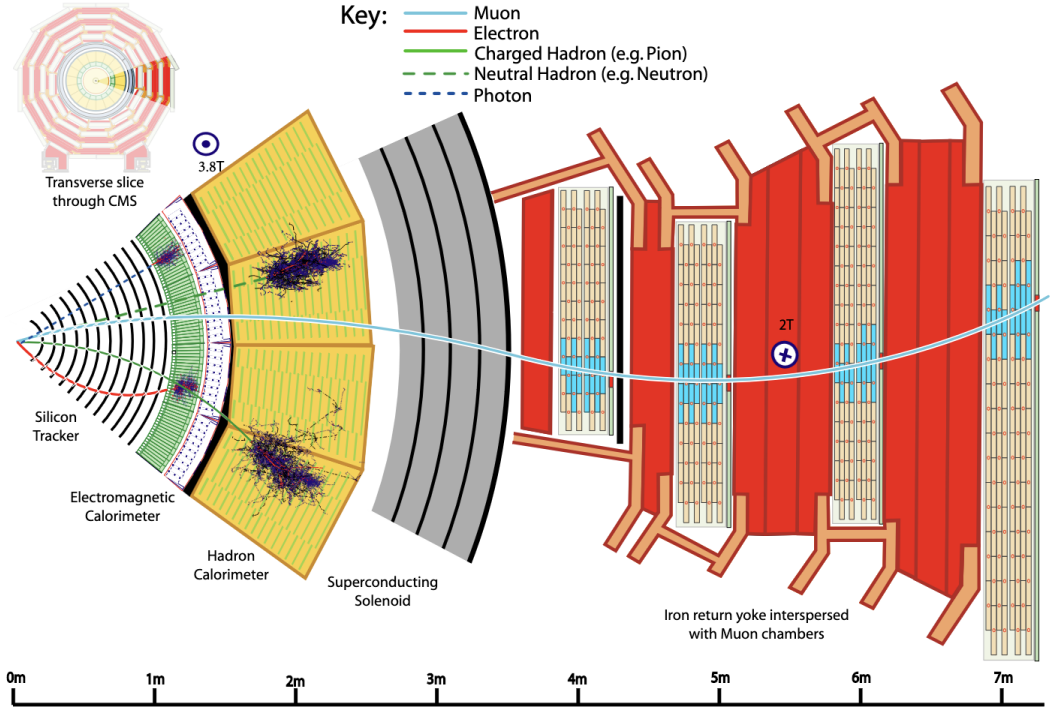


Figure 3.5: Sketch of the specific particle interactions in a transverse slice of the CMS detector, from the beam interaction region to the muon detector. The muon and the charged pion are positively charged, and the electron is negatively charged [55].

The PF approach is based on the following steps:

1. Track reconstruction.
   Tracks are reconstructed as described in 3.1. Originally, track reconstruction was aimed at measuring the momentum of energetic and isolated muons, at identifying energetic and isolated hadronic τ decays, and at tagging heavy flavour jets. It was then optimized for the PF algorithm with the iterative approach, which increases the tracking efficiency, while keeping the background rate under control. It is very important to reach a tracking efficiency as higher as possible: a charged hadron without an associated track would be only detected by the calorimeter, which has a degraded energy resolution. It should be noticed that the tracking efficiency at hight p_T is instead limited, but it does not affect the jet energy resolution, as the calorimeter resolutions are already excellent at these energies.

2. Calorimeter cluster finding.
   The clustering algorithms in the calorimeters have the multiple purposes

of detecting and measuring the energy and direction of stable neutral particles (photons and neutral hadrons), separating them from charged hadron energy deposits, reconstructing and identifying electrons and bremsstrahlung photons and retreiving energy measurents for charged hadrons with low quality or high-p_T tracks. The clustering is carried out independently in each subdetector. It starts from the identification of cluster seeds, i.e. cells that have an energy larger than the neighbouring ones besides overcoming a given threshold. Then, *topologial clusters* are formed by taking into account the neighbouring cells contributions under certain conditions. Final clusters are reconstructed from the topological clusters by using a dedicated algorithm and are accurately calibrated.

3. PF element association.
   While crossing the CMS subdetectors, each particle originates different PF elements that are connected together by a *link algorithm*. For instance, a track and a calorimeter cluster are linked if the extrapolated track position in the calorimeter is geometrically compatible with the cluster area. If more than one cluster is associated to the same track, the link which minimizes the spatial distance is kept.

4. Particle identification.
   Particles are identified in the following order: first, muons are reconstructed from silicon tracks linked with tracks in the muon system (standalone muon tracks) and their calorimeter deposits are subtracted from the Particle Flow hit collection. Then, it is the turn of electrons and isolated photons: electrons are reconstructed by combining tracks and ECAL clusters, while photons are reconstructed from ECAL clusters not connected to a track. Finally, the last particles to be identified are hadrons from jet fragmentation and hadronization, which can be reconstructed as charged or neutral hadrons, nonisolated photons or additional muons. ECAL and HCAL clusters not linked to any track are reconstructed as photons or neutral hadrons: within the tracker acceptance, all these ECAL clusters are reconstructed as photons and the HCAL ones as neutral hadrons. Outside the tracker geometrical acceptance, it is not possible to distinguish charged and neutral hadrons and the following strategy is adopted: ECAL clusters linked to a given HCAL cluster are assumed to arise from the same (charged- or neutral-) hadron shower, while ECAL clusters without such a link are classified as photons.
   At each step, all the tracks and clusters already associated to a particle are masked.

3.3 Pileup per particle identification (PUPPI)

In 2023, as showed in Figure 2.2 (right), the average number of pileup (PU) collisions was 52. It is thereby absolutely necessary to isolate as much as possible the products of a specific p-p interaction from the PU.
Throughout the years, the CMS Collaboration has developed various PU mitigation techniques. The current technique employed, pileup per particle identification (PUPPI) [60], is discussed in this section. PUPPI was developed to overcome the limitations of the previous standard mitigation technique, charged-hadron subtraction (CHS) [55]. The CHS method removes charged particles associated with PU vertices before the jet clustering step. For neutral PU particles, which lack a track that can be associated with a vertex, an event-by-event jet-area-based correction is applied to the jet four-momenta. Additionally, jets predominantly composed of PU particles are discarded through PU jet identification (ID). While this mitigation technique corrects the jet's four-momentum, it does not affect the jet shape or substructure observables.

The PUPPI technique, instead, assigns a probability of originating from PU to each particle and scales its energy accordingly, resulting in a more robust event reconstruction [61]. Specifically, the PUPPI algorithm computes a weight from 0 to 1 for each particle. A weight of 1 is assigned to charged particles used in the fit of the PV and to those not associated with any vertex but having a distance of closest approach to the PV along the z-axis smaller than 0.3 cm. In all other cases, charged particles receive a weight of 0. For neutral particles, the weight is determined based on a discriminating variable α. For a particle i, α is defined as:

$$\alpha_i = \log \sum_{j \neq i, \Delta R_{ij} < R_0} \left(\frac{p_{T,j}}{\Delta R_{i,j}} \right)^2 \begin{cases} \text{for } |\eta_i| < 2.5, \text{ j are charged particles from PV} \\ \text{for } |\eta_i| > 2.5, \text{ j are all kinds of reconstruced particles} \end{cases} \tag{3.1}$$

where j are the other particles, $\Delta R_{i,j}$ is the ΔR between the i and j particles and R_0 is chosen to be 0.4. If there are no particles in the cone of R_0 aperture around the particle i, α_i is set to 0. For $\eta < 2.5$ (geometrical acceptance of the tracker) it is possible to consider only charged particles associated to the PV, thanks to the availability of track information. Outside this region, all the reconstructed particles are considered. According to this definition, α is larger for particles close to others associated with the PV, as particles originating from the same hadronic shower tend to be clustered. This α value is then translated into a probability by using the distribution of α values from charged particles associated with PU vertices to define the expected PU distribution.

A signed χ_i^2 is calculated by comparing the α_i of the neutral particle i with the median ($\bar{\alpha}_{PU}$) and RMS (α_{PU}^{RMS}) of the α distribution obtained from charged PU particles:

$$\text{signed } \chi_i^2 = \frac{(\alpha_i - \bar{\alpha}_{PU})|\alpha_i - \bar{\alpha}_{PU}|}{(\alpha_{PU}^{RMS})^2} \tag{3.2}$$

A large signed χ_i^2 value indicates a higher probability of originating from a PV.

This weight is then used to scale the particle's four-momentum. The p_T weight is evaluated as:

$$w_i = F_{\chi^2,\text{NDF}=1}(\text{signed } \chi_i^2), \tag{3.3}$$

where $F_{\chi^2,\text{NDF}=1}$ is the cumulative distribution function of the χ^2 distribution with one degree of freedom. Finally, particles with $w_i < 0.01$ (probability of being originated from PU vertices larger than 99%) are rejected, along with neutral particles which have $w_i\, p_{T,i} < (A + BN_{\text{vertices}})$ GeV, where N_{vertices} is the number of vertices in the event and A and B are tunable parameters.

Figure 3.6 shows the comparison between data and simulation of the distributions of the three different variables computed and used by the PUPPI algorithm: α, signed χ^2 and weight. The jet sample corresponds to data collected with trigger paths based on high H_T values requirements, where H_T is defined as the scalar sum of the p_T of jets with $p_T > 30$ GeV and $|\eta| < 3$ and an offline threshold of 1500 GeV on H_T. This dataset is compared to a QCD multijet simulated sample. In addition, a dataset enriched in PU particles (PU sample), obtained by using zero-bias trigger that randomly selects collision events, is compared with PU-only simulated sample. It can be observed that these variables are highly discriminating between particles originating from the PV and PU particles.

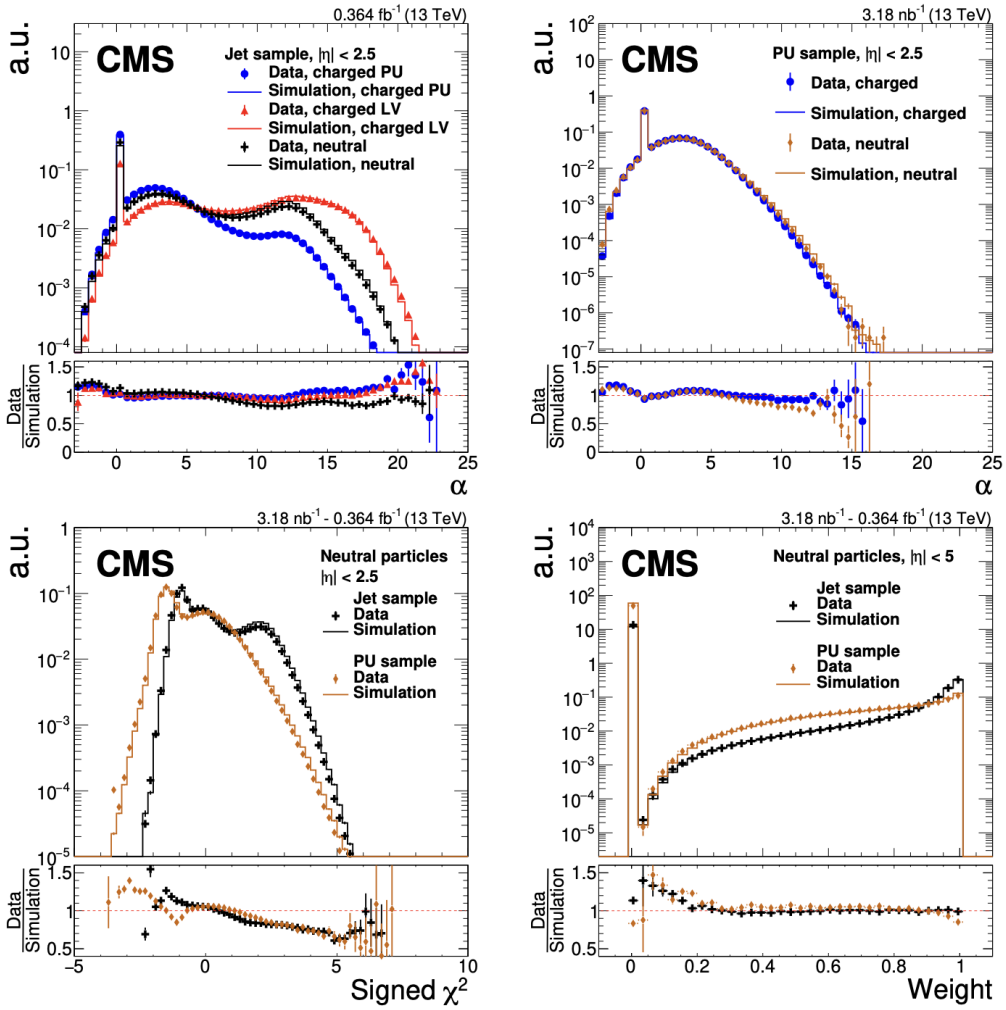


Figure 3.6: Data-to-simulation comparison for three different variables of the PUPPI algorithm. Markers refer to data, solid lines to simulations. The upper left plot shows the α distribution in the jet sample for charged particles associated with the PV (red triangles), charged particles associated with PU vertices (blue circles), and neutral particles (black crosses) for $|\eta| < 2.5$. The upper right plot shows the α distribution in the PU sample for charged (blue circles) and neutral (orange diamond) particles. The lower left plot shows the signed χ^2 distribution for neutral particles with $|\eta| < 2.5$ in the jet sample (black crosses) and in the PU sample (orange diamonds). The lower right plot shows the PUPPI weight distribution for neutral particles in the jet sample (black crosses) and the PU sample (orange diamonds). Each lower plot shows the data to simulation ratio [61]. LV stays for leading vertex and is analogous to PV.

3.4 Jets

Quarks and gluons produced in p-p collisions generate a parton shower and hadronize, resulting in jets of collimated particles. At the CMS experiment, jets are reconstructed by clustering the PF objects with the anti-k_T algorithm [54] within the FastJet software package [62]. The jets used in this search are clustered with a distance parameter ΔR of 0.4 (AK4 jets). This value is optimized for the reconstruction of jets produced by the hadronization of quarks and gluons, while larger ΔR values (0.8 or 1.5) are used for the recontruction of Lorentz-boosted W, Z and Higgs bosons and for top quarks. Jets are clustered after the PUPPI algorithm for PU mitigation is applied to the PF particles. The jet momentum is computed as the vectorial sum of all particle momenta in the jet, and from simulation it is found to be, on average, within 5 to 20% of the true momentum over the whole p_T spectrum and detector acceptance [61].
Jet energy corrections, determined from simulations, are applied in order to bring the measured response of jets to that of the true originating partons and to reduce discrepancies between real data and simulations [60] [63]. Jet energy corrections (JECs) are composed of jet energy scale (JES) and jet energy resolution (JER) smearing corrections. JES corrections, which are used to correct any residual differences in jet energy scale between data and simulation, are computed from in situ measurements of the momentum balance in dijet, photon+jets, Z+jets, and multijet events. The JER is defined as the spread in the jet p_T response distribution and is well approximated by a Gaussian. Since the JER is usually better in simulations than in data, JER corrections are applied to simulated jets in order to guarantee a more realistic description of real data.

Additionally, a list of selections should be applied to jets in order to reject spurious jets reconstructed primarily from detector noise (JetID). The JetID imposes thresholds on variables describing the jet's particle composition: neutral and charged hadron and electromagnetic fraction, muon fraction, number of constituents, charged and neutral multiplicity. These thresholds are tuned for different pseudorapidity regions.

In 2023, certain regions of the calorimeters produced anomalously high or low jet rates due to sub-optimal calorimeter calibration, and, additionally, inefficiencies in some zones of the tracking system affected jet reconstruction. Therefore, the CMS JETMET group released "Jet Veto maps" in the η-ϕ space, which are used to veto events containing jets in problematic zones [64].

3.5 Missing Transverse Energy (MET)

In addition to the visible objects, like electrons, muons, hadrons, etc., in p-p collisions at the LHC, several weakly interacting particles such as neutrinos are produced but not reconstructed, since they cross the detectors without interacting. It is possible to retrieve information on their energy by computing the missing transverse energy (MET). From the energy conservation, we know that the sum of the momenta of all the particles produced in a p-p collision along the transversal plane should be null. Therefore, the transverse momentum of undetected particles ($\vec{E}_T^{miss}$ or $\vec{p}_T^{miss}$) can be estimated as the negative vector sum of the transverse momenta of detected particles [65]:

$$\vec{E}_T^{miss} = -\sum_{i,\ \text{reco particles}} \vec{p}_{T,i} \tag{3.4}$$

The impact of PU on the MET resolution is more difficult to mitigate with respect to jets. The pileup component of events has a natural tendency to have near zero $\vec{p}_T^{miss}$.
Figure 3.7 shows the resolution on $\sum E_T$ (left) and E_x^{miss} (right) achieved with the PUPPI algorithm (pink) in comparison with the previous methods, PF (blue) and CHS (red) [60]. The PUPPI algorithm showed better performance and therefore is adopted as a standard in CMS also for MET reconstruction.

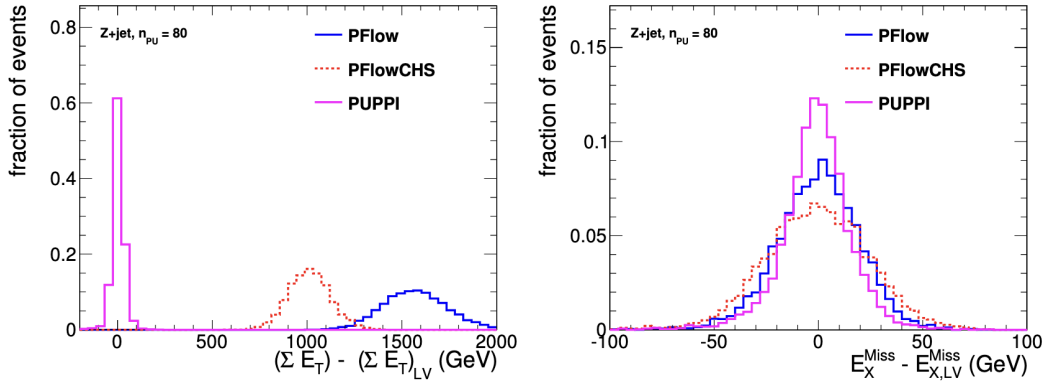


Figure 3.7: (Left) the resolution of $\sum E_T$ and (right) the resolution of E_x^{miss} in Z+jets events with $n_{PU} = 80$ [60].

3.6 Heavy-flavour tagging

A crucial step in this search is the discrimination of c jets from b jets and light-flavour jets. To achieve this, the CMS Collaboration has developed several heavy-flavour tagging algorithms over the years. Heavy-flavour jets, which originate from b or c quarks, are characterized by unique properties that can

be exploited for identification.
Firstly, b and c hadrons, produced through the hadronization of b and c quarks, have relatively long lifetimes of approximately 1.5 ps and 1 ps, respectively. As a result, they can travel a few millimeters (Lorentz-boost) before decaying, creating tracks that are displaced from the primary vertex (PV). These displaced tracks often allow for the reconstruction of a secondary vertex (SV).

Additionally, due to the larger mass and harder fragmentation of b and c quarks compared to light quarks and gluons, the decay products of heavy hadrons typically have higher transverse momentum (p_T) than other jet constituents [66].

Lastly, in approximately 20% of b jets and 10% of c jets, a soft muon or electron is present within the jet, which further helps in distinguishing heavy-flavour jets from light-flavour jets.
Figure 3.8 illustrates some of these properties, as the presence of a SV reconstructed from the displaced tracks within a heavy-flavour jet, the non-negligible flight distance and the impact parameter.

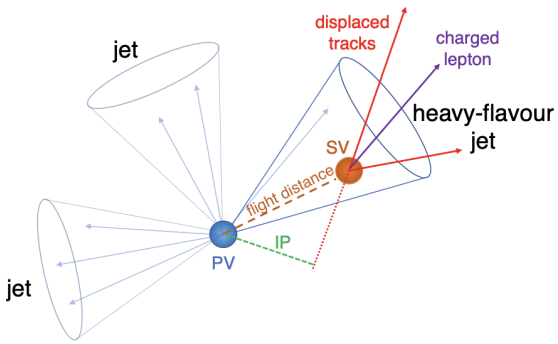


Figure 3.8: Sketch of a heavy-flavour jet: tracks from the decay of a b or c hadron give rise to displaced tracks with respect to the PV and a secondary vertex (SV) can be reconstructed from them [66].

Heavy-flavour tagging algorithms are based on multivariate analysis (MVA) techniques which combine a large number of variables related to the distinctive properties of heavy-flavour jets.
Given that c hadrons have shorter lifetimes and lower masses compared to b hadrons, the distributions of discriminating variables for c jets are intermediate between those of b jets and light-flavour jets. This makes c tagging more challenging than b tagging, necessitating the introduction of two separate discriminators: one to distinguish c jets from light-flavour jets (CvsL) and another to differentiate c jets from b jets (CvsB) [67].

Secondary vertices, which are crucial for heavy-flavour tagging, are reconstructed using the inclusive vertex finding (IVF) algorithm [68]. Independently from the jet reconstruction, IVF takes as input all reconstructed tracks with $p_T > 0.8$ GeV and a longitudinal impact parameter IP< 0.3 cm. These tracks are then clustered and fitted with the adaptive vertex fitter [69], and only high-quality SVs are retained. If a track is found to be more compatible with the PV, it is removed from the SV, and a refit is performed to avoid ambiguity.

The tagger that has been the standard in CMS in recent years is DeepJet [70] [71], which has an architecture based on a convolutional recurrent neural network.
This advanced network takes as input 613 features, categorized into four groups: global variables, charged PF candidate variables, neutral PF candidate variables and SVs variables.

- Global features: these include per-jet variables such as jet kinematics, the number of tracks in the jet, and the number of SVs associated with the jet, as well as per-event variables like the number of PVs reconstructed in the event.

- Charged PF Candidates: for the first 25 PF candidates, ordered by their displacement significance, 16 features are considered. These features include track kinematics, track fit quality, and displacement parameters with respect to the PV.

- Neutral PF Candidates: Six features of the first 25 neutral PF candidates are used as input

- SVs: 12 variables, such as the flight distance significance, of the firts 4 SVs are exploited.

To summarize, DeepJet utilizes low-level features from a large number of jet constituents. To avoid dependencies on the jet p_T and η, data are pre-processed before being fed into the neural network.
The DeepJet architecture employs three 1×1 convolutional layers, one for each category of jet constituents (charged candidates, neutral candidates, and SVs), which perform an automatic feature engineering . This is followed by three recurrent layers of the Long Short-Term Memory (LSTM) type, which combine the information separately for each category. At the end, the whole information is combined by fully connected layers and the network provides as output six probabilities, P(b), P(bb), P(lepb), P(c), P(g), P(uds), which corrspond

to the probability that a jet originates from a b quark, two b quarks, a leptonic b hadron decay, one or more c quarks, a gluon, or a light-flavour quark respectively. The DeepJet output scores are defined in Table 3.1.

Tagger	BvsC/L	CvsB	CvsL
DeepJet	$P(b) + P(bb) + P(lepb)$	$\frac{P(c)}{P(c)+P(b)+P(bb)+P(lepb)}$	$\frac{P(c)}{P(c)+P(uds)+P(g)}$

Table 3.1: DeepJet tagger definition for both b and c tagging.

The CMS b tagging and vertexing (BTV) POG defines thresholds for the tagger output scores corresponding to fixed working points (WP): loose (L), medium (M) and tight (T). These thresholds are set on the basis of specific mistag probability values [72].

3.6.1 State-of-the-art: ParticleNet

Recently, the CMS Collaboration developed a powerful heavy-flavour tagging algorithm called ParticleNet [26]. This algorithm is a customized dynamic graph convolutional neural network (DGCNN) [73], originally designed specifically for the identification of hadronic decays of highly Lorentz-boosted heavy particles such as top quark, W, Z and Higgs boson, and for classifying different decay modes (e.g., $Z \to b\bar{b}$, $Z \to c\bar{c}$, $Z \to q\bar{q}$) [74].
The novelty of ParticleNet is that each jet is represented as an unordered, permutation-invariant set of particles, referred to as *particle cloud*. This approach is advantageous over sequence or tree representation, which require particles to be sorted in some way, as the constituent particles in a jet have no intrinsic order.

Figure 3.9 illustrates the architecture of ParticleNet, which contains three EdgeConv [73] blocks. For each particle, the EdgeConv block identifies the k nearest neighboring particles. These neighboring particles, along with the particle features, are given as input to the EdgeConv operation, which constructs the "edge features". The EdgeConv operation itself consists of a three-layer MLP (multi-layer perceptron). After passing through the EdgeConv blocks, the learned feature from all the particles are combined by a channel-wise global average pooling operation, followed by a fully connected layer. Finally, a fully connected layer with two units and a softmax function provide the output for a binary classification task, as the ParticleNet architecture was first evaluated

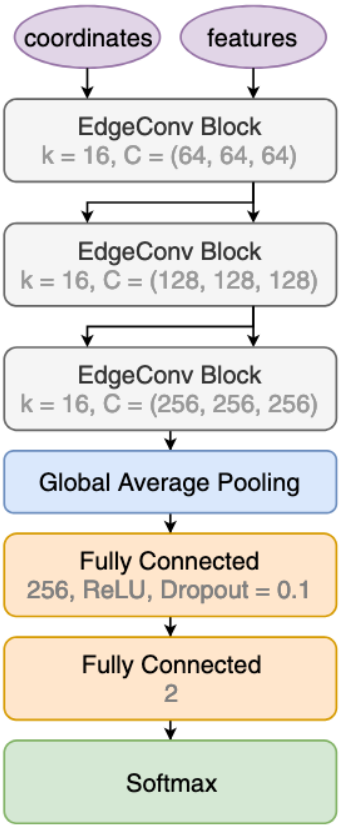


Figure 3.9: ParticleNet algorithm architecture [26].

on two jet tagging benchmarks [26]. Figure 3.10 shows the ParticleNet performance (pink solid line) in identifying boosted $H \to b\bar{b}$ (left) and $H \to c\bar{c}$ (left) decays, compared to DeepAK8 (blue solid line), which has been the standard for AK8 jet tagging in recent years. A mass-decorrelated version of both algorithms was also developed, with their corresponding performance displayed [74]. ParticleNet significantly outperforms its predecessors, such as DeepAK8, demonstrating the strength of this novel approach.
Given the great success of ParticleNet for AK8 jet tagging, its architecture was then customized for AK4 jet classification, commonly referred to as ParticleNetAK4. I evaluated the performance of ParticleNet and DeepJet in tagging the AK4 jets, by using jets with $p_T > 30$ GeV and $|\eta| < 2.5$ contained in a MC sample of TTbar events simulated with 2023 data-taking conditions. Figure 3.11 shows the BvsL and BvsC (left) and CvsL and CvsB (right) discriminating power of ParticleNet (violet) and DeepJet (red). Also in this case, ParticleNet demonstrates significantly better performance than that of its predecessor, DeepJet.

In preparation for Run-3, ParticleNetAK4 was extended in order to perform simultaneously multiple tasks: jet flavour classification, τ identification and jet

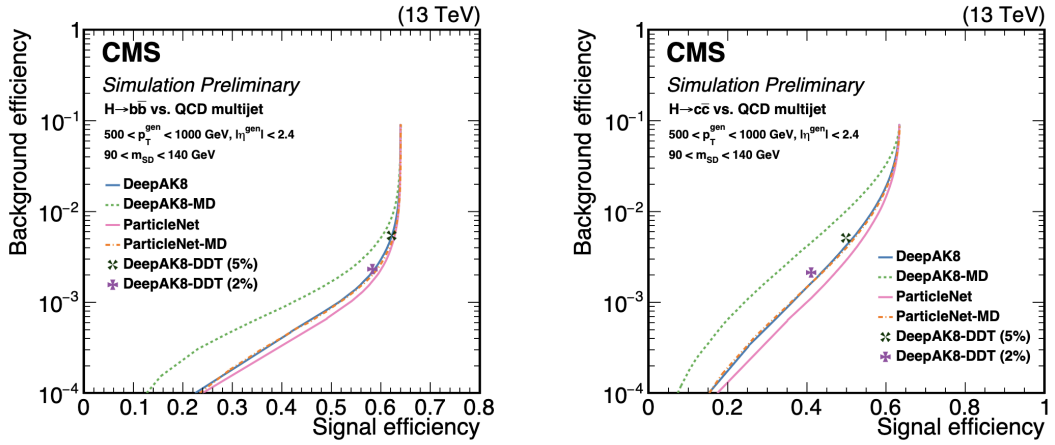


Figure 3.10: Performance of ParticleNet and DeepAK8 at identifying hadronically decaying Higgs bosons: (left) $H \to b\bar{b}$ and (right) $H \to c\bar{c}$ [74].

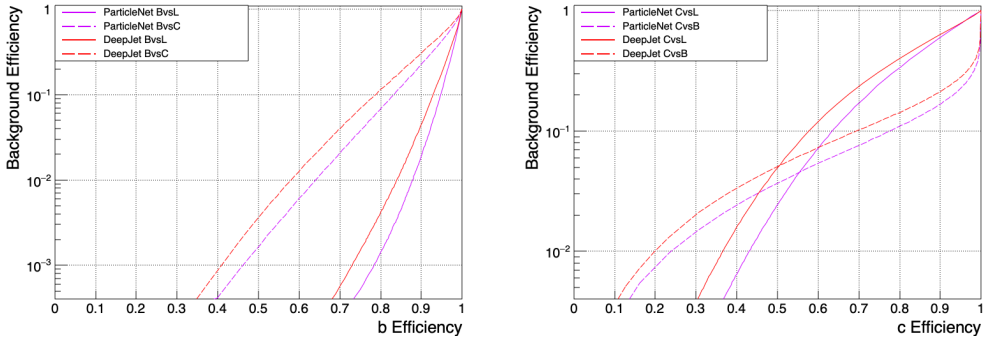


Figure 3.11: (Left) comparison between DeepJet (red) and ParticleNet (violet) in discriminating b jets against light flavour jets (solid line) and c jets (dashed line). (Right) comparison between DeepJet (red) and ParticleNet (violet) in discriminating c jets against light flavour jets (solid line) and b jets (dashed line).

energy regression.

3.6.2 Heavy-flavour tagging at HLT

Great improvements have been achieved in recent years in heavy-flavour tagging at the HLT level. In Run-2, trigger paths with b tagging selection used the DeepCSV model trained on offline-reconstructed variables. Given that the variable reconstruction performed at the HLT level is less accurate than the offline one, this approach was suboptimal.

In prepration for Run-3, the DeepJet model, which was originally intended to be the default tagger in the HLT menu, was trained using online-reconstructed

features. Additionally, a dedicated training of ParticleNet with HLT features was performed and deployed online. It was employed for the first time in the trigger path dedicated to the search for the double Higgs boson. It has been observed that the heavy-flavour tagging performance online strongly benefits from retraining the taggers with HLT-level input variables. Moreover, ParticleNet shows greatly improved performance with respect to the previous taggers.

During the initial phase of Run-3, the first trigger paths utilizing ParticleNet demonstrated both reliability and strong performance, leading to the development of additional trigger paths. Notably, I implemented a dedicated trigger path for the search for VBF $H \to c\bar{c}$, incorporating a ParticleNet c tagging selection. This path was deployed online just before the Run-3 2023 data collection and it has been accumulating statistics since then. It marks the first trigger path with a c tagging selection applied directly at the HLT level and is among the first to use the ParticleNet tagger. The performance of this novel HLT path is studied in detail in Chapter 4.
Due to the success of the ParticleNet tagger, a decision was made to migrate all paths with b tagging selections from DeepJet to ParticleNet for the 2024 data collection, establishing it as the standard tagger at the HLT for Run-3.

3.6.3 Data-to-simulation flavour tagging corrections

Heavy-flavour tagging algorithms are trained on simulated samples enriched with b, c, and light-flavour jets. However, since simulations do not perfectly replicate real data, scale factors (SFs) must be calculated and applied to correct for mismodelling effects. These scale factors are defined as the ratio between tagging efficiency in data and simulation:

$$SF_f = \frac{\epsilon_f^{data}(p_T, \eta)}{\epsilon_f^{MC}(p_T, \eta)} \tag{3.5}$$

where $\epsilon_f^{data}(p_T, \eta)$ and $\epsilon_f^{MC}(p_T, \eta)$ represent the tagging efficiencies for a jet with flavour f in data and simulation, respectively, computed as a function of the jet p_T and η. The b/c tagging efficiency probability is evaluated as the ratio of tagged b/c (according to a fixed WP) to the total number of b/c jets. In simulations, jet flavour is easily determined by geometrically matching the jet with generated hadrons. In data, the tagging efficiency is measured by using pure samples of jets with a certain flavour [66]:

- c jet tagging efficiency is assessed using a sample enriched in c jets obtained from events where a W boson produced in association with a c

quark ($W + c$). These events are characterized by a final state with the c quark and the W boson of opposite-sign (OS) electric charge, as shown in the first two Feynman diagrams of Figure 3.12. The dominant background, represented by the third Feynamn diagram, contributes equally to OS and same-sign (SS) events. Hence, for each variable distribution, the SS contribution is subtracted from the OS contribution ("OS-SS" method).

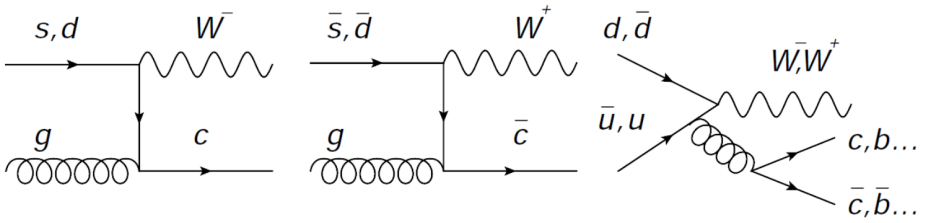


Figure 3.12: Leading order production of $W + c$ with opposite-sign electric charges (left and midle), and of $W + q\bar{q}$ through gluon splitting (right) [66].

- b jet tagging efficiency is calculated using samples of multijet events and top pair events ($t\bar{t}$) with various methods to select events enriched in true b jets.

In addition to providing b and c tagging scale factors for fixed working points, the BTV POG derives SFs to correct the tagger discriminant shape. These SFs are crucial when the full distribution of the tagging score is used in the analyses, for instance as an input variable in an MVA algorithm. The shape correction scale factors are computed through an iterative fit that minimizes a χ^2-based metric, which quantifies the discrepancy between data and simulation.
The c tagging calibration and scale factor derivation for newly collected Run-3 2023 can be inferred from Figure 3.13, where the distribution of the tagger scores CvsL and CvsB is plotted for Data and MC simulation. Overall, a good agreement between data and MC is achieved, already before the application of scale factors. However, the SF evaluation is still ongoing.

3.6.4 Object reconstruction in 2023 data-taking

Data considered in this work were collected by the CMS experiment from April to July 2023. The 2023 data taking stopped in July, before than expected, because of a serious LHC incident and an integrated luminosity of 27 fb^{-1} was registered.
Moreover, in June 2023, after Technical Stop 1, 27 modules in the Barrel Pixel

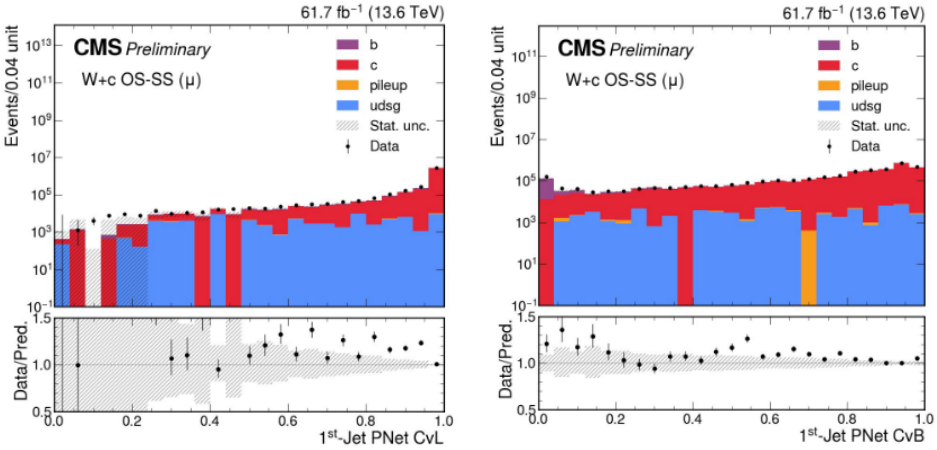


Figure 3.13: Comparison between data (color filled histogram) and MC (black dots) of the CvsL (left) and CvsB (right) score distributions evaluated in the W+c phase space for the combination of 2022 and 2023 data-taking periods [75].

Layer 3 and 4 (BPix3 and BPix4) became inoperable due to a problem distributing the LHC clock to the modules (BPix issue). These modules have been turned off since this incident. They span a region of approximately 0.4 radians (∼23 degrees) in ϕ at negative pseudo-rapidity. Since the regions covered by these modules are fully overlapping in eta and phi across the two detector layers, a full gap in acceptance is produced when attempting to seed tracks with traditional "high purity" pixel-hit combinations (triplets and quadruplets). The BPix issue caused a reduction in the track reconstruction performance both online and offline, around η ∼-1, affecting jet reconstruction and heavy flavour tagging as well.
Overall a 10% reduction of the heavy flavour tagging efficiency was registered inside the hole.

The 2023 datasets for physics studies are splitted in two eras, gathering respectively data collected before and after the BPix issue: 2023C and 2023D.

Chapter 4

Search for $VBF\ H \to c\bar{c}$

The search for Higgs boson decay in charm quark pair with the Higgs produced in vector boson fusion ($VBF\,H \to c\bar{c}$) at the CMS experiment is performed for the first time in this thesis. As discussed in Chapter 2, at the CMS experiment, data acquisition relies on a two level trigger, due to the impossibility to store the results of each proton-proton interaction to disk.
This means that events discarded by the trigger are irretrievably lost. The design of the trigger paths is therefore a crucial step for the success of the CMS physics program.
At the start of my PhD program, no dedicated trigger existed for the $VBF\,H \to c\bar{c}$ search. Existing triggers, such as those designed for similar analyses like $VBF\,H \to b\bar{b}$, were not suitable due to their limited signal efficiency for this process.
The first step of my work was to develop a dedicated trigger path for this search. This path was deployed online before the start of 2023 data-taking and has been collecting data since. It is the first trigger ever deployed with a c tagging selection at the HLT and one of the earliest to utilize the state-of-the-art ParticleNet tagger, providing crucial insights into the performance of this new tagger at the HLT level. Concurrently, the CMS Collaboration introduced the VBF parking trigger strategy, a set of paths tailored to the VBF topology, collecting data within the parking dataset at a rate of 1 kHz. From the EraD of 2023, the trigger path I developed was also moved to parking.

Following the trigger implementation, I developed a strategy for analyzing the data collected in 2023 using the newly deployed trigger. The main goals to achieve with the analysis consist of minimizing the contribution of the QCD multijet events, which represents the major background in this search, and discriminating the signal against the resonant $Z \to q\bar{q}$, $W \to q\bar{q}$ and $H \to b\bar{b}$

backgrounds. I trained a boosted decision tree (BDT) algorithm to distinguish the signal from QCD background events. A crucial aspect of the analysis is the use of ParticleNet taggers, specifically CvsB and CvsL (defined in Chapter 3), which allow for discrimination of c jets from b and light-flavour jets, respectively. The Higgs boson candidate is reconstructed indeed from the two most c tagged jet in an event.

Since this analysis is still blinded, only data events outside the Higgs boson nominal mass region are accessible. Blinding the analysis is a common method used to prevent the analysers from being biased while developing the analysis strategy.
For the final step, I performed a parametric shape analysis to estimate the expected upper limit on the signal strength at 95% CL, incorporating major systematic uncertainties that could impact the results. The signal strength is extracted from a combined fit to the Higgs candidate mass distribution. Since the MC simulation of the QCD multijet process is not highly accurate, its mass distribution is modeled with a data-driven method.

4.1 Data and Monte Carlo samples

The analysis described in this chapter is carried out on proton-proton collision data collected and certified as good for physics during the 2023 at the CMS experiment, which correspond to an integrated luminosity of 27 fb^{-1} out of 32.7 fb^{-1} delivered (see Figure 2.2).
In order to study the signal and suppress major background contributions, simulated Monte Carlo (MC) samples are used.
Table 4.1 lists the signal and background MC processes used in this analysis along with their cross section.

The $VBF\,H \to c\bar{c}$ is the main process contributing to the signal in this analysis. It is generated with POWHEG-BOX framework [76] [77] with next-to-leading order accuracy interfaced with PYTHIA [78] for fragmentation and hadronization, where a dipole correction [79] that takes into account the color connection between the incoming and outgoing partons of the VBF process is used (VBF dipole recoil).
The $ggF\,H \to c\bar{c}$ process, which is difficult to disentangle from the signal, is treated as a signal process, although the selection is specifically designed to enhance the sensitivity to the $VBF\,H \to c\bar{c}$ signal. The $ggF\,H \to c\bar{c}$ process

is generated with POWHEG-BOX with multi-scale improved next-to-leading order (MiNLO) accuracy [80] [81] interfaced with PYTHIA. Figure 4.1 shows the Feynamn diagrams of the $VBF,\ H \to c\bar{c}$ (left) and $ggF\,H \to c\bar{c}$ (right) processes.

The dominant background for this search is the QCD mulitijet process, produced by MADGRAPH5_aMC@NLO [82] at leading order accuracy (LO) of the coupling constant α_S. The jets from the matrix element calculations are matched to the parton shower produced by PYTHIA using MLM [83] prescription. The QCD multijet samples are divided in bins of H_T, the scalar sum of jet transverse momenta.

The $VBF\,H \to b\bar{b}$ and $ggF\,H \to b\bar{b}$ processes, which constitute the main resonant background, are generated in the same way as the signal ones.

For the $Z \to q\bar{q}$ and $W \to q\bar{q}$ processes, which also peak in the signal region of the mass spectrum, both the Drell-Yan (DY) and electroweak (EWK) production mechanisms are considered. The latter is characterized by a smaller cross section, but has the same final state topology as the signal process. The DY production is generated with MADGRAPH5_aMC@NLO interfaced with PYTHIA. The corresponding samples are divided in bin of transverse momentum of the quarks originating from the Z decay. The EWK production is generated with MADGRAPH5_aMC@NLO at NLO accuracy and the FxFx [84] prescription for matrix element and softer parton showers matching, finally interfaced with PYTHIA for parton showering and hadronization.

The last important contribution to the background are given by the inclusive single top ($t/\bar{t}$ + X) and top pair ($t\bar{t}$ + X) processes, all generated with POWHEG-BOX framework interfaced with PYTHIA.

Although the MC samples used in this search were produced centrally by the CMS Collaboration, I implemented the first stages of the analysis by using MC samples that I generated privately. This was necessary because the central production was not complete until March 2024. Moreover, I contributed to the central CMS MC production process by providing some of the MC generator configuration files that were missing from the database and were requested specifically for this search.

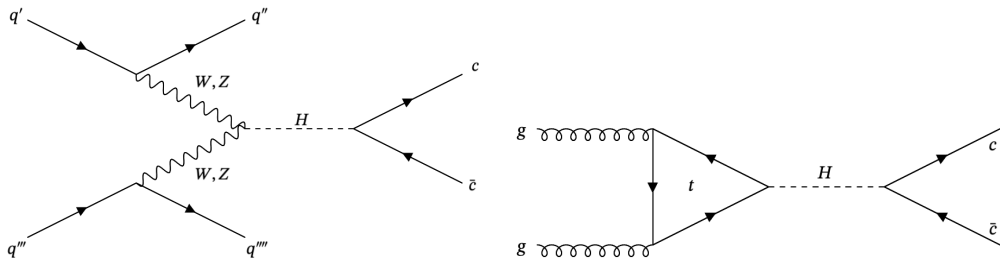


Figure 4.1: Feynman diagram of the $VBF\,H \rightarrow c\bar{c}$ (left) and $ggF\,H \rightarrow c\bar{c}$ (right) processes.

4.2 Trigger development

As described in Section 2.2.7, the events collected by the second level of CMS trigger are selected by a suite of selections known as HLT path. Each step of this selection is processed by a module. The first module in the sequence contains a list of L1 seeds, which are essentially the initial filters. The HLT selections are applied only to events that pass these L1 seeds.
Therefore, the first step in developing the HLT path was to identify a set of L1 seeds capable of selecting events within the phase space of the signal region.

The final state of the $VBF\,H \rightarrow c\bar{c}$ process consists of two c jets produced by the Higgs decay, emitted centrally, and two light jets (VBF jets) produced in the forward and backward regions due to the scattered quarks (q'' and q''' in Figure 4.1(left)). This specific signature is exploited to develop the trigger.

4.2.1 L1 seeds

In order to define a suitable set of L1 seeds, I studied the distributions of key kinematic observables of jets reconstructed at L1, considering as signal the ones matched to generator-level Monte Carlo information [1]. The background jets are studied in a neutrino gun sample, accounting for jets that are clustered from PU interactions. Figure 4.2 shows pseudorapidity (η) distributions for L1 jets produced by the Higgs decay in blue, L1 jets generated by VBF emission in purple and L1 background jets in orange. It is also interesting to look at the distance in η ($\Delta\eta$) for pair of jets, reported in Figure 4.3 (left). It can be seen that jets from Higgs decay tend to be emitted centrally and close to one another, while VBF jets are more likely produced in the forward-backward directions. Background jets are mostly emitted centrally, but they also have

[1] L1 jets are matched geometrically with generator-level c quarks found to be the daughter particles of the Higgs boson

	process	sample	$\sigma \cdot BR$ (fb)
	$VBF\,H \to c\bar{c}$		$1.21{\cdot}10^{2}$
signal	$ggF\,H \to c\bar{c}$		$8.73{\cdot}10^{2}$
	$VBF\,H \to b\bar{b}$		$2.43{\cdot}10^{3}$
background	$ggF\,H \to b\bar{b}$		$1.76{\cdot}10^{4}$
	QCD multijet	100 GeV $<$ H_T $<$ 200 GeV	$2.53{\cdot}10^{10}$
		200 GeV $<$ H_T $<$ 400 GeV	$1.96{\cdot}10^{9}$
		400 GeV $<$ H_T $<$ 600 GeV	$9.68{\cdot}10^{7}$
		600 GeV $<$ H_T $<$ 800 GeV	$1.37{\cdot}10^{7}$
		800 GeV $<$ H_T $<$ 1000 GeV	$3.03{\cdot}10^{6}$
		1000 GeV $<$ H_T $<$ 1200 GeV	$8.81{\cdot}10^{5}$
		1200 GeV $<$ H_T $<$ 1500 GeV	$3.87{\cdot}10^{5}$
		1500 GeV $<$ H_T $<$ 2000 GeV	$1.27{\cdot}10^{5}$
		H_T $>$ 2000 GeV	$2.66{\cdot}10^{4}$
	DY $Z \to qq$	100 GeV $<$ $p_T(qq)$ $<$ 200 GeV	$3.44{\cdot}10^{5}$
		200 GeV $<$ $p_T(qq)$ $<$ 400 GeV	$4.84{\cdot}10^{4}$
		400 GeV $<$ $p_T(qq)$ $<$ 600 GeV	$2.68{\cdot}10^{3}$
		$p_T(qq)$ $>$ 600 GeV	$4.46{\cdot}10^{2}$
	EWK $Z \to qq$		$1.37{\cdot}10^{4}$
	DY $W \to qq$	100 GeV $<$ $p_T(qq)$ $<$ 200 GeV	$1.76{\cdot}10^{6}$
		200 GeV $<$ $p_T(qq)$ $<$ 400 GeV	$2.27{\cdot}10^{5}$
		400 GeV $<$ $p_T(qq)$ $<$ 600 GeV	$1.28{\cdot}10^{4}$
		$p_T(qq)$ $>$ 600 GeV	$2.13{\cdot}10^{3}$
	EWK $W \to q\bar{q}$		$9.56{\cdot}10^{4}$
	$t\bar{t}$ + X	$t\bar{t} \to 2l2\nu$	$8.54{\cdot}10^{4}$
		$t\bar{t} \to 4q$	$3.52{\cdot}10^{5}$
		$t\bar{t} \to l\nu 2q$	$3.64{\cdot}10^{5}$
	single top	$\bar{t}W^{+} \to 4q$	$1.66{\cdot}10^{4}$
		$tW^{-} \to 4q$	$1.66{\cdot}10^{4}$
		$\bar{t}W^{+} \to l\nu 2q$	$1.61{\cdot}10^{4}$
		$tW^{-} \to l\nu 2q$	$1.60{\cdot}10^{4}$

Table 4.1: List of simulated signal and background MC processes used in this search with their cross section.

a large forward component, while the $\Delta\eta$ between jets from each possible combination peaks at small values with a long tail. Similarly, Figure 4.3 (right) shows the transverse momentum (p_T) distribution for the same jet categories. Jets from the signal process have a p_T distribution peaking at approximately

50 GeV, while background jets are mainly characterized by small values of p_T.

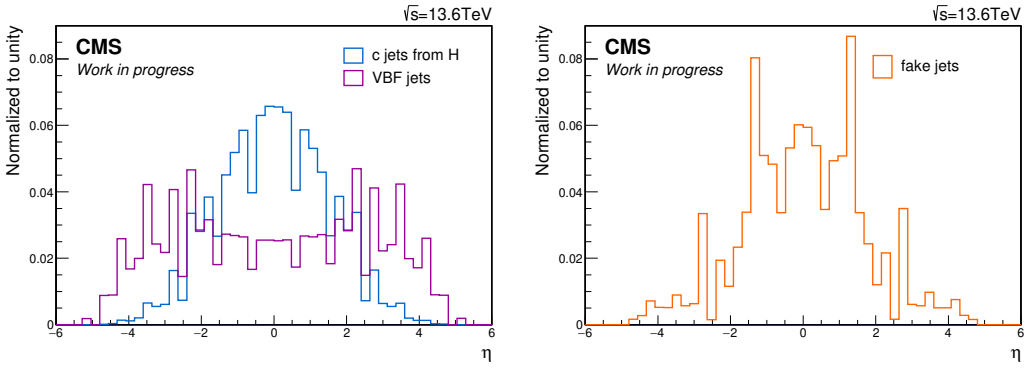


Figure 4.2: η distribution for L1 jets from Higgs decay (left plot, blue), VBF jets (left plot, purple) and background jets (right plot, orange).

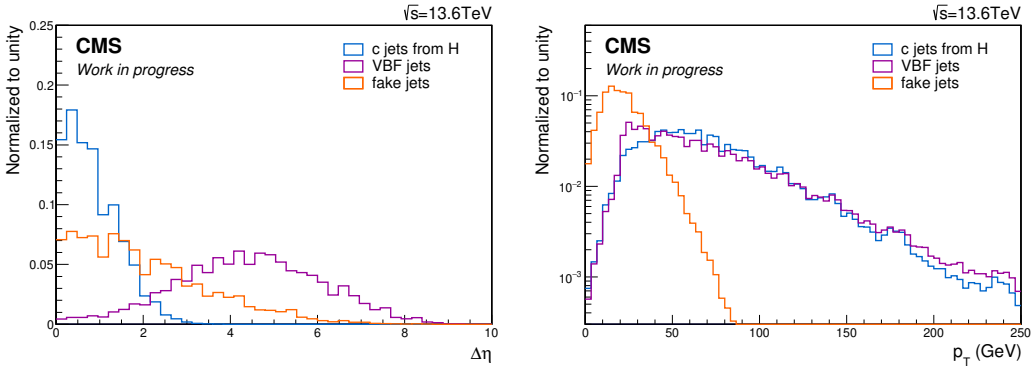


Figure 4.3: Left: $\Delta\eta$ distribution for pair of charm jets from Higgs decay (blue), VBF jets (purple) and background jets (orange). Right: p_T distribution for charm jets from Higgs decay (blue), VBF jets (purple) and background jets (orange).

In order to discriminate signal jets from background ones, it is important also to consider the invariant mass of pair of jets, shown in Figure 4.4. A big fraction of background jets can be discarded by cutting at large values of the invariant mass, but in order to preserve the structure of the Higgs boson candidate, we cannot set a cut larger than 60 GeV.

Finally, the last variable taken into account for this study is the scalar sum of jet transverse momenta (referred to as HTT in the L1 menu), which distribution is shown in Figure 4.5 for VBF $H \to c\bar{c}$ events in blue and background events in orange. This variable is highly discriminating and therefore very useful for trigger purposes.
Finally, by examining all these distributions, I selected the L1 seeds listed in Table 4.2, along with their description.

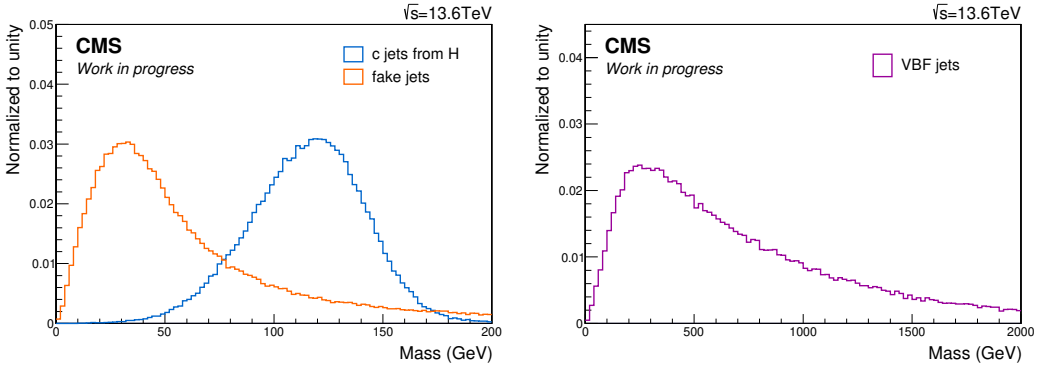


Figure 4.4: Invariant mass distribution for pair of L1 jets from Higgs in blue (left), VBF jets in purple (right) and background jets in orange (left).

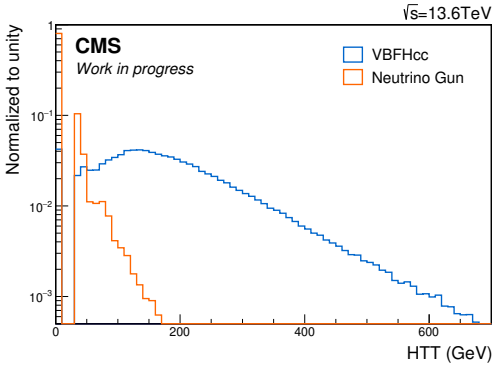


Figure 4.5: HTT distribution for signal (blue) and background (orange) events.

4.2.2 HLT path

In general, for the development of a new HLT path, it is necessary to maximize the signal acceptance, while keeping the rate under acceptable amounts. For each path, we define as "pure" the contribution to the total rate delivered only by that path. The pure rate is evaluated by running the HLT path on Zero-bias events. For the integration of a new path dedicated to the search for VBF $H \to c\bar{c}$, a pure rate of approximately 10 Hz could be allocated, following the prescriptions of the Trigger Study Group of the CMS Collaboration.
The logic of the trigger selection I implemented by studying the phase-space of the signal is summarized as follows:

- four jets with high p_T thresholds
- among the four selected jets, two of them are required to be in the central region of CMS apparatus
- one central jet tagged as charmed

L1 seed	Selection
L1_TripleJet_95_75_65_*e* _DoubleJet_75_65_er2p5	3 jets with p_T>95,75,65 GeV 2 jets with p_T>75,65 GeV and $\|\eta\|$<2.5
L1_TripleJet_100_80_70_ _DoubleJet_80_70_er2p5	3 jets with p_T>100,80,70 GeV 2 jets with p_T>80,70 GeV and $\|\eta\|$<2.5
L1_SingleJet180	1 jet with p_T>180 GeV
L1_SingleJet200	1 jet with p_T>200 GeV
L1_DoubleJet_110_35_ DoubleJet35_Mass_Min620	2 jets with p_T>110, 35 GeV 2 jets with p_T>110 GeV and mass > 620 GeV
L1_QuadJet_95_75_65_20_ _DoubleJet_75_65_er2p5_Jet20_ _FWD3p0	4 jets with p_T>95,75,65,20 GeV 2 jets with p_T>75,65 GeV and $\|\eta\|$<2.5 1 jet with p_T>20 GeV and $\|\eta\|$>3.0
L1_HTT360er	HTT>360 GeV
L1_HTT280er	HTT>280 GeV

Table 4.2: List of L1 seeds included in the VBF $H \to c\bar{c}$ HLT path.

- the remaining two jets tagged as VBF with large invariant mass and a large pseudorapidity gap

It is important to consider that it is not possible to tune each threshold of the trigger selection independently, since the effect on the rate increase can be estimated only for the whole HLT path. Therefore, I examined the distributions of the most important variables in order to define approximately a range for applying cuts and then, I tuned the thresholds by evaluating the pure rates of several HLT paths based on different combinations of these cuts. For example, in order to define the p_T thresholds, I took into account the p_T distributions of the first four jets, sorted by decreasing p_T, for simulated MC $VBF\,H \to c\bar{c}$ events selected by the L1 seeds described in Table 4.2. These distributions are shown in Figure 4.6.

A crucial step of the trigger selection consists of identifying c jets. At this purpose, the state-of-the-art ParticleNet tagger, integrated for the first time in the CMS trigger for Run-3 data taking, was used. The tagger associates to each jet the probability to be originated by a gluon (*probg*), by a light (*probuds*), by a b (*probb*) or by a c (*probc*) quark. Three tagger output scores

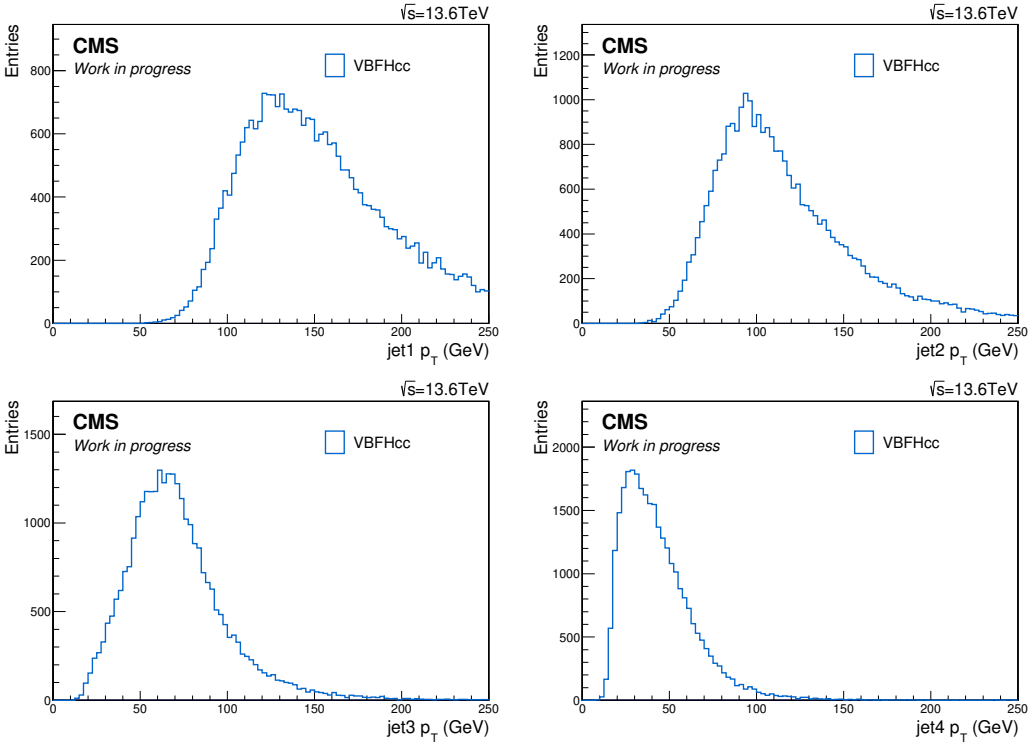


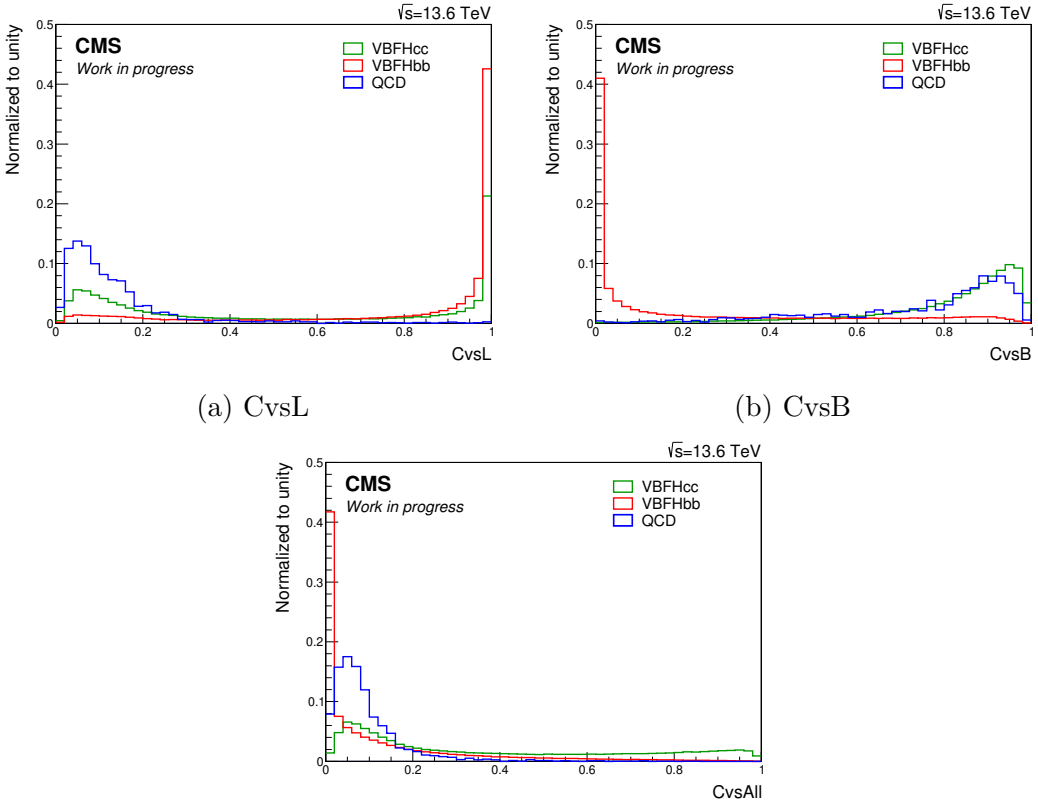
Figure 4.6: p_T distribution of the first four jets reconstructed at HLT and sorted by decreasing p_T, for simulated MC VBF $H \to c\bar{c}$ events passing the L1 selection summarized in Table 4.2.

are defined:

$$\begin{aligned} CvsAll &= \frac{probc}{probc + probb + probg + probuds} \\ CvsL &= \frac{probc}{probc + probuds + probg} \\ CvsB &= \frac{probc}{probc + probb} \end{aligned} \tag{4.1}$$

Figure 4.7 shows the ParticleNet CvsL (a), CvsB (b) and CvsAll (c) scores for c jets in green, b jets in red and light jets in blue, which true flavour is identified through a geometrical match with quarks from the MC truth. I used three different MC simulated samples: VBF $H \to c\bar{c}$, VBF $H \to b\bar{b}$ and QCD, since each of them is enriched with a particular jet flavour. In the attempt of maximizing the signal acceptance while reducing the pure rate, I chose to use the CvsAll score to tag and select c jets in the trigger path. The final version of the HLT path I developed, named
HLT_QuadPFJet100_88_70_30_PNet1CvsAll0p5_VBF3Tight,
is schematized in Figure 4.8 and described below:

- at least four jets with $p_T > 100, 88, 70, 30$ GeV



(a) CvsL

(b) CvsB

(c) CvsAll

Figure 4.7: ParticleNet scores for c jets in green, b jets in red and light jets in blue.

- at least two jets with $p_T > 30$ GeV and $|\eta| < 2.5$
- at least one jet with $p_T > 30$ GeV, $|\eta| < 2.5$ and CvsAll > 0.5
- among the first four jets sorted by decreasing p_T, the two jets with the largest CvsAll score are tagged as c jets, while the remaining ones are considered VBF jets. The VBF jets are required to have a $\Delta\eta$ greater than 3.5 and an invariant mass larger than 460 GeV.

Before choosing this as the final configuration of the HLT path, I implemented and tested several options. For each of them, I estimated the total and pure rates on Zero-bias events and the efficiency on the MC signal. At that time, I used a private MC sample of $VBF\ H \to c\bar{c}$ produced with a preliminary set of data taking conditions and a limited statistics. It is important to underline that each time a new HLT path option was implemented, it was necessary to re-produce the signal MC sample incorporating the new trigger menu.

The efficiency is evaluated as the ratio between the number of simulated signal events firing the trigger and the total number of generated events, without

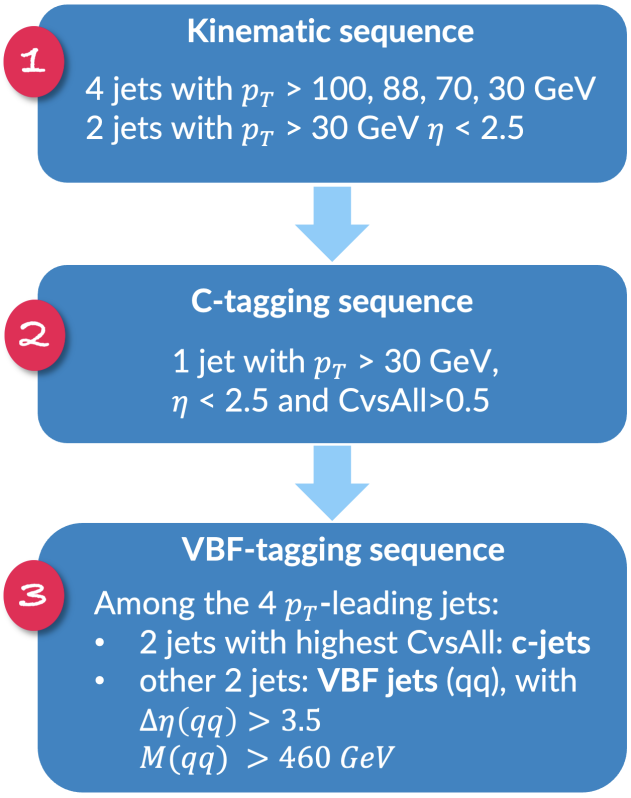


Figure 4.8: Scheme of the HLT path dedicated to the $VBF\,H \to c\bar{c}$ search.

selecting a particular phase space.
Table 4.3 gathers some of the most relevant HLT paths studied and shows their rate and efficiency on the signal. For the c-tagging sequence, in addition to the cut on the CvsAll score, selections on the CvsL and BCvsAll (probability to be a c or b jet) scores were also considered.

The final trigger path dedicated to the $VBF\ H \to c\bar{c}$ has an efficiency on the MC signal of 1.8% and a pure rate of ∼10 Hz. For reference, it can be considered that the Run-2 trigger path dedicated to the $VBF\,H \to b\bar{b}$ search, has an efficiency on the $VBF\,H \to c\bar{c}$ process of 0.3%.

Figure 4.9 shows the total rate of this path as a function of the integrated luminosity. The rate starts from a value of 13 Hz and increases with a step up to ∼18 Hz and after a certain time it decreases again, as a result of the deployment online of new calibrations for the HCAL system.

HLT path description	Rate	Pure rate	Efficiency
4 jets with $p_T > 100, 75, 50, 30$ GeV 2 jets with mean CvsL > 0.5 $M(qq) > 300$ GeV - $\Delta\eta(qq) > 3.5$	56.18 Hz	37.57 Hz	2.6%
4 jets with $p_T > 100, 75, 50, 30$ GeV 2 jets with CvsL > 0.5 $M(qq) > 300$ GeV - $\Delta\eta(qq) > 3.5$	16.22 Hz	8.22 Hz	1.0%
4 jets with $p_T > 100, 75, 50, 30$ GeV 2 jets with CvsL > 0.4 and 0.2 $M(qq) > 460$ GeV - $\Delta\eta(qq) > 3.5$	33.56 Hz	20.86 Hz	-
4 jets with $p_T > 100, 75, 50, 30$ GeV 2 jets with CvsL > 0.5 and 0.3 $M(qq) > 460$ GeV - $\Delta\eta(qq) > 3.5$	21.7 Hz	12.08 Hz	1.4%
4 jets with $p_T > 100, 75, 50, 30$ GeV 2 jets with BCvsAll > 0.4 and 0.2 $M(qq) > 460$ GeV - $\Delta\eta(qq) > 3.5$	46.7 Hz	30.48 Hz	1.3%

Table 4.3: Rate and efficiency on the $VBF\ H \rightarrow c\bar{c}$ process of the most relevant HLT path options tested in this study.

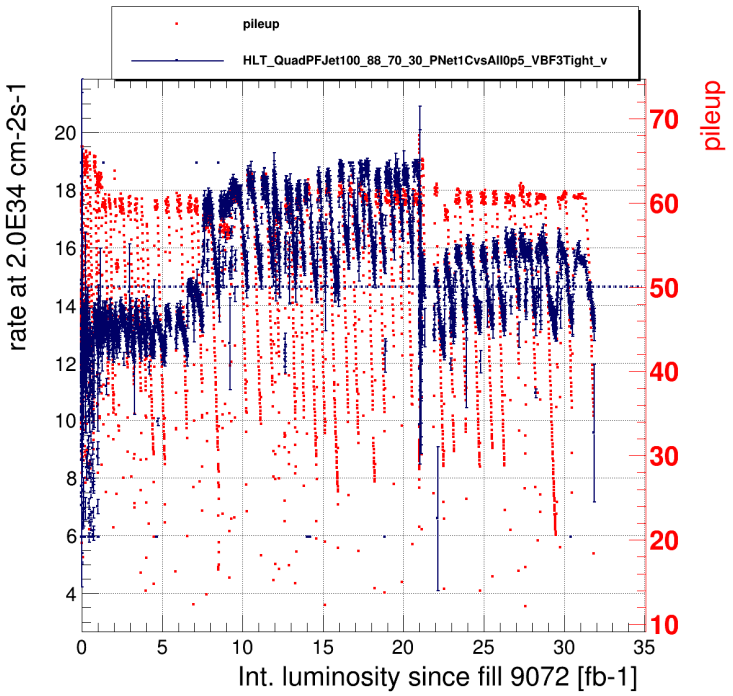


Figure 4.9: Rate of the HLT path *HLT_QuadPFJet100_88_70_30_ _PNet1CvsAll0p5_VBF3Tight* as a function of the integrated luminosity (blue). The number of PU interactions is also shown (red).

4.2.3 Checks on VBF parking dataset

As discussed in Chapter 2, the CMS Collaboration introduced, during the 2023 data taking, the VBF parking strategy, consisting of a set of inclusive and exclusive triggers specifically tailored for the VBF topology which collect data at an overall rate of 1 kHz.
The inclusive trigger paths require two jets with high p_T, separated in η and with a large invariant mass, without making assumptions about the Higgs boson decay.
In contrast, the exclusive trigger paths target specific decay processes.
These parking paths were deployed online after the path I developed specifically for the $VBF\ H \to c\bar{c}$ process and therefore they collected data with a smaller integrated luminosity.
To explore their potential, I assessed the suitability of the VBF parking paths for the $VBF\ H \to c\bar{c}$ search. As, at that time, these new triggers were not yet integrated into the MC simulations, I generated private MC simulations of the signal process based on the new trigger menu.

Among the many VBF parking trigger paths deployed, I tested the following, which were the most suitable for the signal topology:

- *HLT_VBF_DiPFJet105_40_Mjj1000_Detajj3p5*
  This inclusive path selects events containing:
  - at least one jet with p_T greater than 105 GeV;
  - at least an additional jet with p_T greater than 40 GeV;
  - invariant mass of the two jets exceeding 1000 GeV and $\Delta\eta$ larger than 3.5.
- *HLT_VBF_DiPFJet70_40_Mjj600_Detajj2p5_DiPFJet60_ _JetMatchingQuadJet*
  This exclusive path collects event with:
  - at least two jets with p_T greater than 70 and 40 GeV, respectively, invariant mass greater than 600 GeV and $\Delta\eta$ larger than 2.5;
  - at least two additional jets with p_T greater than 60 GeV;
  - the four jets of the previous steps matching with the L1 objects.

I applied an offline pre-selection similar to the trigger one, with the inclusion of a requirement on the c tagging score (CvsAll > 0.5). This way, the phase-space selected by the VBF parking triggers with the pre-selection on top is

similar to the one selected by the $VBF\ H \to c\bar{c}$ trigger and allows to make a fair comparison.
Table 4.4 shows the efficiency of the VBF parking triggers alone in the second column and the overall efficiency of the trigger and offline pre-selection in the third column.

trigger path	trigger efficiency	pre-selection efficiency
inclusive VBF parking	6.1%	1.4%
exclusive VBF parking	3.8%	1.2%

Table 4.4: Efficiency of the inclusive and exclusive VBF parking trigger paths selection. The efficiency of the offline pre-selection applied on events selected by each trigger is also reported.

Figure 4.10 shows, respectively for the inclusive (left) and exclusive (right) trigger paths, the distribution, after the offline pre-selection, of the Higgs boson candidate mass, reconstructed from the two most c-tagged jets in the event.

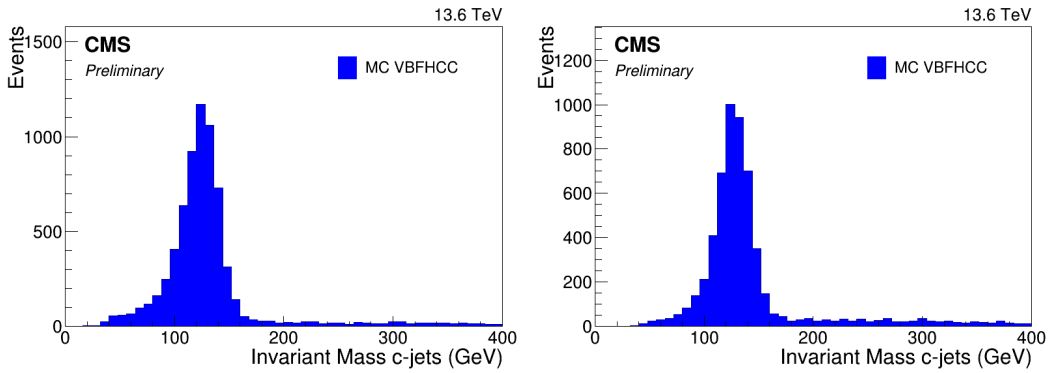


Figure 4.10: Distribution of the Higgs boson candidate mass obtained for signal MC events passing the offline pre-selection on top of the inclusive (left) and exclusive (right) VBF parking trigger paths.

Overall, these new triggers show a better efficiency than the one dedicated to the $VBFH \to c\bar{c}$ search, due to the much more relaxed requirements applied because of the higher rate allocated for the parking strategy. From this preliminary study, they appear to offer a further opportunity to enhance the sensitivity of this search. Further studies will be necessary to fully assess the impact of the new triggers on final results, particularly considering background contributions.
The trigger path dedicated to the $VBFH \to c\bar{c}$ search remains part of the menu for 2024 data-taking and will likely be maintained in the future runs, as it extends the acceptance of the VBF parking triggers.

4.3 Trigger performance and scale factors

The performance of the trigger quoted in the previous sections is evaluated on MC simulated samples. Despite being very precise, the MC simulations do not perfectly replicate real data, necessitating the use of scale factors to correct for mismodelling effects. These scale factors ensure that the efficiency of the each trigger selection is consistent between data and MC simulation. The HLT path I developed for the $VBF\,H \to c\bar{c}$ search has a complex structure and thus requires an accurate study for the evaluation of the scale factors. As illustrated in Figure 4.8, the HLT path is composed by three main sequences: the kinematic sequence, primarily based on large p_T cuts, the c tagging sequence and the VBF sequence. The overall trigger scale factor is an event-based weight computed as the product of p_T (SF_{p_T}), c tagging ($\mathrm{SF}_{\mathrm{ctag}}$) and VBF ($\mathrm{SF}_{\mathrm{VBF}}$) scale factors:

$$\mathrm{SF}_{\mathrm{trigger}} = \mathrm{SF}_{p_T} \cdot \mathrm{SF}_{\mathrm{ctag}} \cdot \mathrm{SF}_{\mathrm{VBF}} \tag{4.2}$$

For the evaluation and validation of the scale factors, a HLT path identical to the signal one but excluding the c tagging and VBF sequences, referred to as the control path, was included in the trigger menu:
HLT_QuadPFJet100_88_70_30.
To summarize, the selections applied by the control trigger path are given by the first sequence of Figure 4.8.

4.3.1 Trigger p_T scale factors

The kinematic sequence of the HLT path applies large p_T thresholds (100, 88, 70 and 30 GeV) on the first four p_T-leading jets. For each of these thresholds, the scale factors are evaluated as the ratio between the efficiency of that cut in data and its efficiency in the MC simulation. To compute this scale factors, two prescaled single-jet HLT paths, *HLT_PFJet80* and *HLT_PFJet60*, are utilized, which require at least one jet with p_T larger than 80 and 60 GeV, respectively. Even though it sets a tighter threshold, the *HLT_PFJet80* collects data with a smaller prescale and it is therefore used in addition to *HLT_PFJet60* in order to gather more statistics. However, it is only exploited for the first two thresholds (100 and 88 GeV) as it cannot be used to estimate the efficiency of p_T thresholds smaller than 80 GeV.

For each p_T threshold (denoted as p_T-thr), I used a *tag-and-probe* method to evaluate the selection efficiency. I selected events passing the HLT path *HLT_SinglePFJet60* (*HLT_SinglePFJet80*) and fulfilling the following condi-

tions:

- p_T-leading offline jet (tag jet) has $p_T > 100$ (130) GeV and $\eta < 2.2$ and matches a jet reconstructed at HLT (HLT jet)
- p_T-subleading offline jet (probe jet) is distant in ϕ from the tag jet: $\Delta\phi_{tag-probe} > 2.5$
- 3rd jet by p_T-order has p_T smaller than 30% of the tag and probe jets mean p_T

The efficiency for each p_T-thr is evaluated on both data and simulated QCD multijet events as a function of p_T in four different pseudorapidity ($|\eta|$) intervals: $0-1.4$, $1.4-2.4$, $2.4-3.0$, and $3.0-4.7$. This is done by calculating the ratio of two histograms: the denominator histogram is filled with the p_T of the probe jet for events that pass the selection; for the numerator histogram, an additional condition requires that the probe jet matches an HLT jet with p_T greater than p_T-thr.
Figures 4.11-4.13 show the efficiency of the p_T trigger thresholds on data (blue) and MC simulation (red) as a function of p_T in the four pseudorapidity intervals considered. In the bottom panel of each figure, the ratio of data to MC efficiency, which correspond to the scale factors, is displayed. Overall, the scale factors rapidly reach 1 in the barrel, while in the endcap a smoother turn-on curve is observed.

The scale factors corresponding to the fourth p_T threshold are not evaluated, as there is no HLT trigger path with a threshold below 30 GeV in the trigger menu that would allow for a similar calculation. However, in the offline event selection a 5 GeV larger cut than the one applied by the trigger path is used, ensuring for a data-MC efficiency ratio close to one.
The final p_T trigger SF, which is applied to the entire event, is calculated as the product of the p_T trigger scale factors corresponding to the first three p_T-leading jets:

$$\mathrm{SF}_{p_T} = \mathrm{SF}_{p_T}^{\mathrm{jet1}} \cdot \mathrm{SF}_{p_T}^{\mathrm{jet2}} \cdot \mathrm{SF}_{p_T}^{\mathrm{jet3}}, \tag{4.3}$$

where $\mathrm{SF}_{p_T}^{\mathrm{jet1}}$ is the SF taken from the plots in Figure 4.11 in correspondence of the p_T and η of the p_T-leading jet, $\mathrm{SF}_{p_T}^{\mathrm{jet2}}$ is the SF taken from the plots in Figure 4.12 in correspondence of the p_T and η of the p_T-subleading jet and $\mathrm{SF}_{p_T}^{\mathrm{jet3}}$ is the SF taken from the plots in Figure 4.13 in correspondence of the p_T and η of the third p_T-leading jet.

To validate the p_T trigger scale factors, I compared the p_T and η distributions of the four p_T-leading jets for data and MC simulated QCD multijet events selected using the control HLT path (*HLT_QuadPFJet100_88_70_30*), both before and after applying the scale factors. These distributions are shown in Figures 4.14–4.21. The application of the p_T trigger scale factors significantly improves the agreement between data and simulation, particularly in the low p_T region, where the differences were most pronounced.

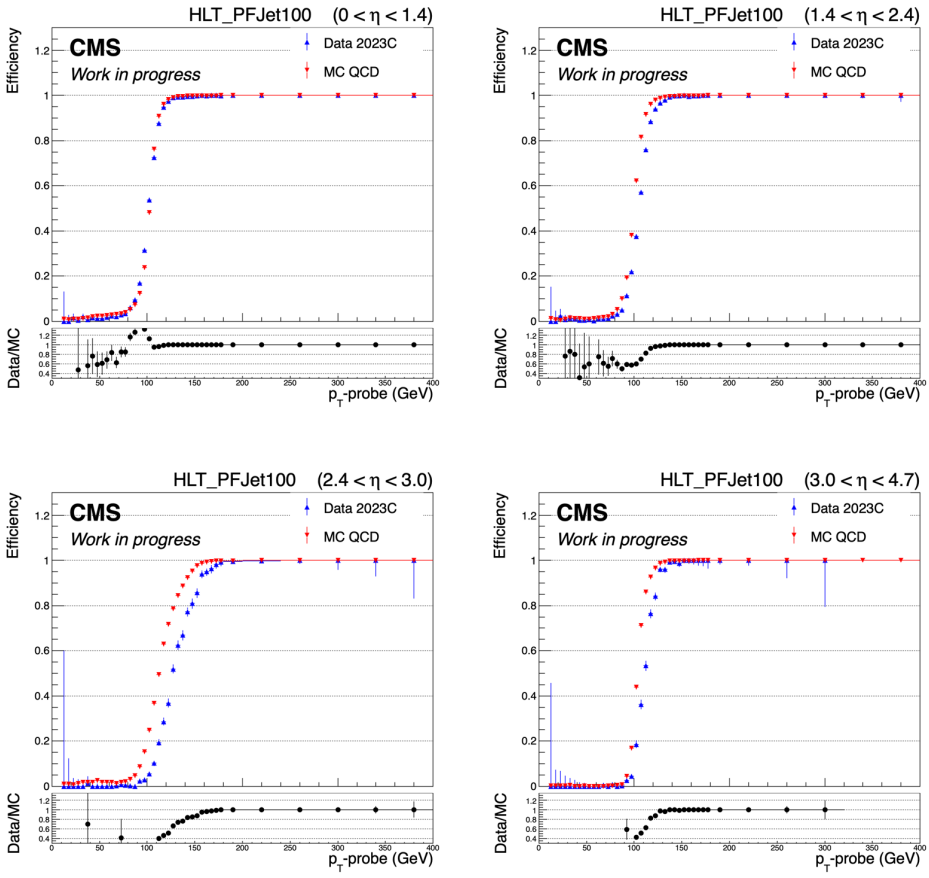


Figure 4.11: p_T trigger scale factors for the 100 GeV p_T threshold in the pseudorapidity regions $0<\eta<1.4$ (left top), $1.4<\eta<2.4$ (right top), $2.4<\eta<3.0$ (left bottom) and $3.0<\eta<4.7$ (right bottom). In the upper panel the comparison between MC simulated QCD multijet events (red) and data (blue) efficiency is plotted. In the lower panel the the ratio of these efficiencies are displayed as corresponding scale factors.

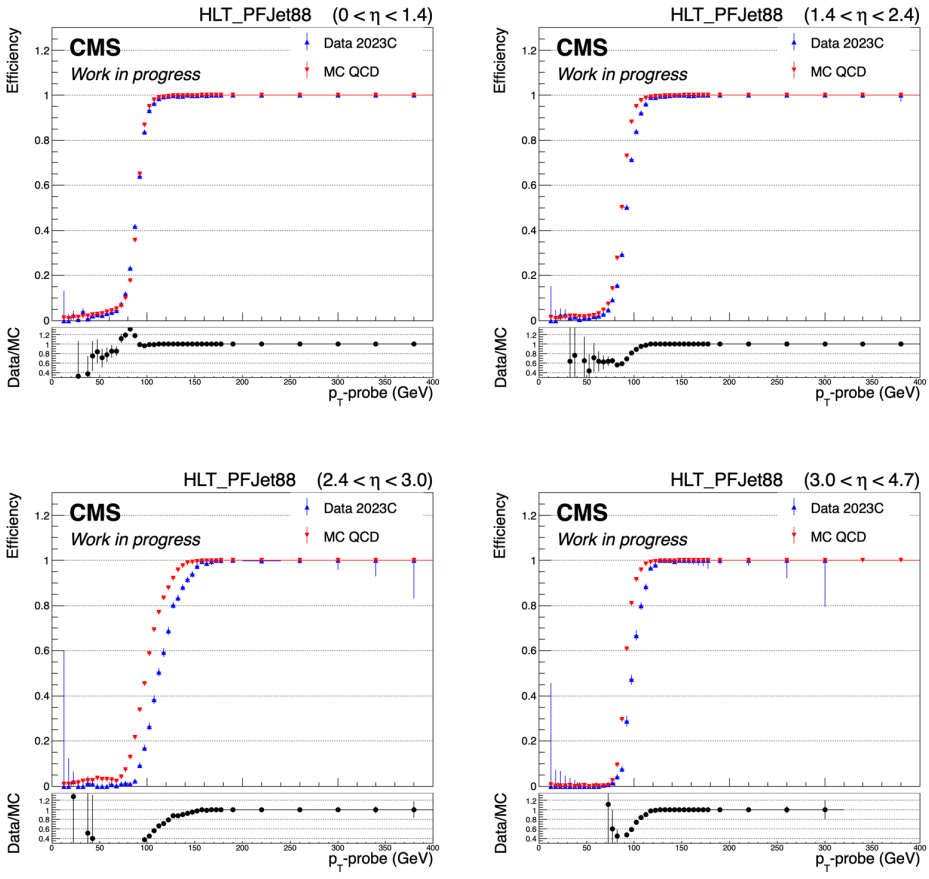


Figure 4.12: p_T trigger scale factors for the 88 GeV p_T threshold in the pseudorapidity regions $0<\eta<1.4$ (left top), $1.4<\eta<2.4$ (right top), $2.4<\eta<3.0$ (left bottom) and $3.0<\eta<4.7$ (right bottom). In the upper panel the comparison between QCD MC and data efficiency is plotted. In the lower panel the the ratio of these efficiencies are displayed as corresponding scale factors.

4.3.2 Trigger VBF scale factors

After applying the trigger p_T SFs, it is necessary to estimate the product of the c tagging and VBF scale factors:

$$\mathrm{SF}_{\mathrm{ctag}}(\mathrm{ctag}) \cdot \mathrm{SF}_{\mathrm{VBF}}(\Delta\eta, \mathrm{mass}) \tag{4.4}$$

In the signal HLT path, the VBF requirements, specifically $\Delta\eta$ larger than 3.5 and an invariant mass greater than 460 GeV, are applied to the two jets which are the ones with the smallest c tagging score among the four p_T-leading jet. As a result, it is not straightforward to decouple the contribution of the VBF requirements from the c tagging requirement when quantifying mismodelling effects between data and MC simulation. Ideally, this would require "switching off" the c tagging condition.

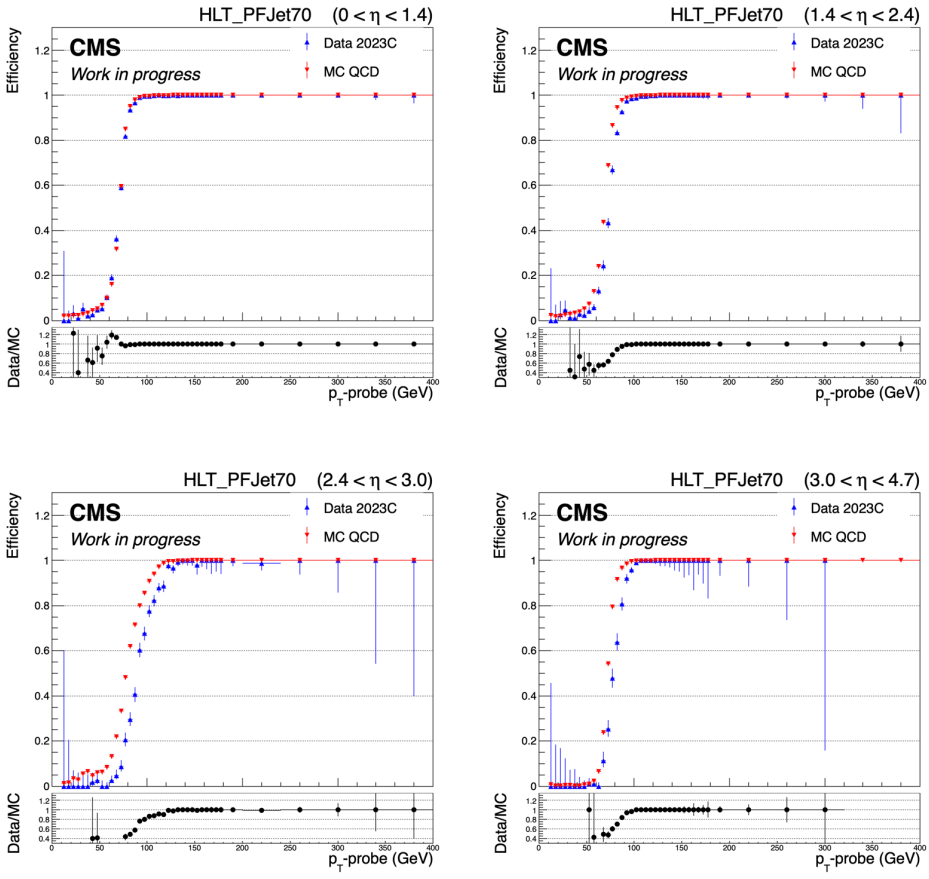


Figure 4.13: p_T trigger scale factors for the 70 GeV p_T threshold in the pseudorapidity regions $0<\eta<1.4$ (left top), $1.4<\eta<2.4$ (right top), $2.4<\eta<3.0$ (left bottom) and $3.0<\eta<4.7$ (right bottom). In the upper panel the comparison between QCD MC and data efficiency is plotted. In the lower panel the the ratio of these efficiencies are displayed as corresponding scale factors.

A similar approach was used in the CMS search for $VBF\,H \to b\bar{b}$ [85], where a trigger path with a b tagging selection was employed. In that case, the effect of the VBF requirements alone was studied by using an emulation of the control HLT path, which, in addition to the kinematic sequence, applied the HLT VBF requirements to the two jets with the largest η-opening among the four p_T-leading jets. This study showed no significant disagreement between data and MC simulation attributable to the HLT VBF requirements, allowing equation 4.4 to be simplified as follows:

$$\mathrm{SF}_{\mathrm{ctag}}(\mathrm{ctag}) \cdot \mathrm{SF}_{\mathrm{VBF}}(\Delta\eta, \mathrm{mass}) = \mathrm{SF}_{\mathrm{ctag}}(\mathrm{ctag}) \cdot const = \mathrm{SF}_{\mathrm{ctag}}(\mathrm{ctag}) \quad (4.5)$$

where *const* is a constant factor that can be absorbed into the HLT c tagging scale factors.

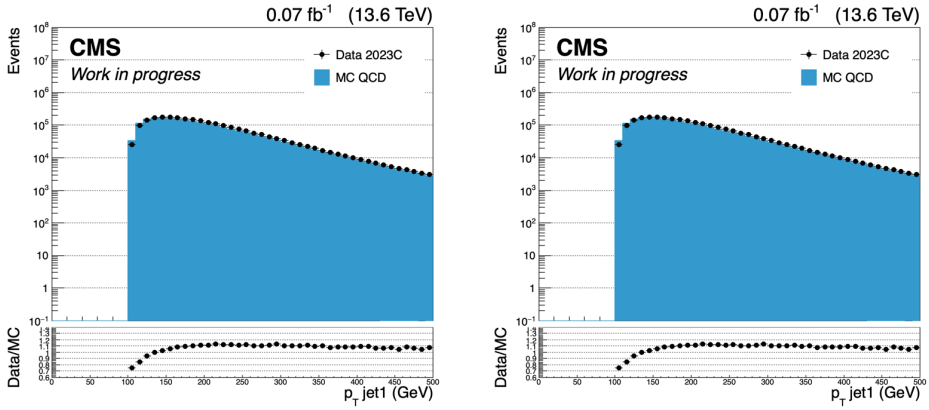


Figure 4.14: p_T distribution of the p_T-leading jet for data (black dots) and MC simulated QCD multijet (blue) events selected with the control HLT path before (left) and after (right) the application of the trigger p_T SFs.

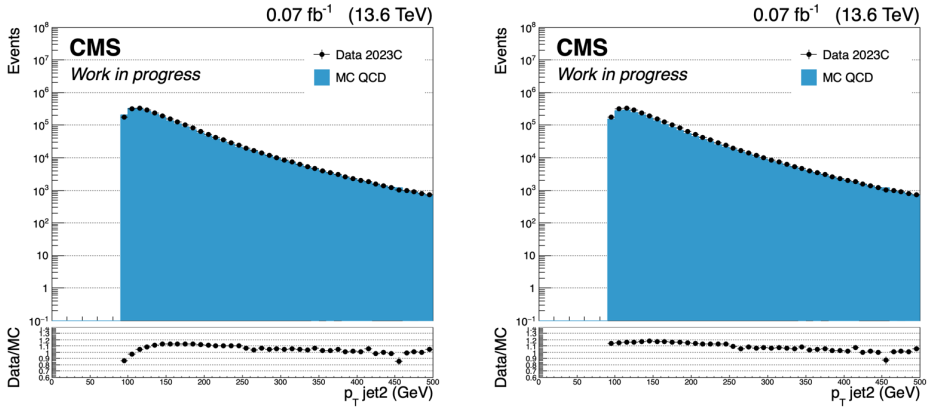


Figure 4.15: p_T distribution of the p_T-subleading jet for data (black dots) and MC simulated QCD multijet (blue) events selected with the control HLT path before (left) and after (right) the application of the trigger p_T SFs.

Given the structural similarity between the trigger paths for $VBFH \to c\bar{c}$ and $VBFH \to b\bar{b}$, it is reasonable to assume that the same behavior holds in the case of the HLT path for $VBFH \to c\bar{c}$.

4.3.3 Trigger c tagging scale factors

To compute the scale factors that correct for mismodelling effects due to the c tagging cut, I began by selecting events using the control HLT path, which has the same kinematic sequence as the signal path. I then applied the following offline event selection criteria:

- at least four offline reconstructed jets with p_T larger than 105, 90, 75

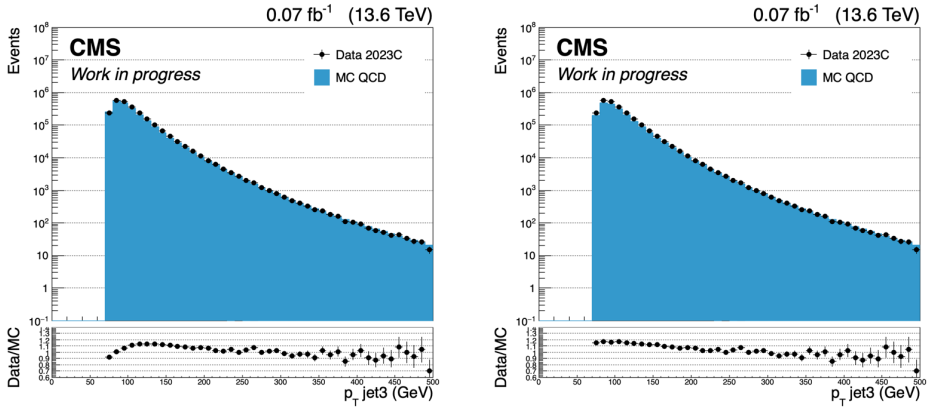


Figure 4.16: p_T distribution of the third p_T-leading jet for data (black dots) and MC simulated QCD multijet (blue) events selected with the control HLT path before (left) and after (right) the application of the trigger p_T SFs.

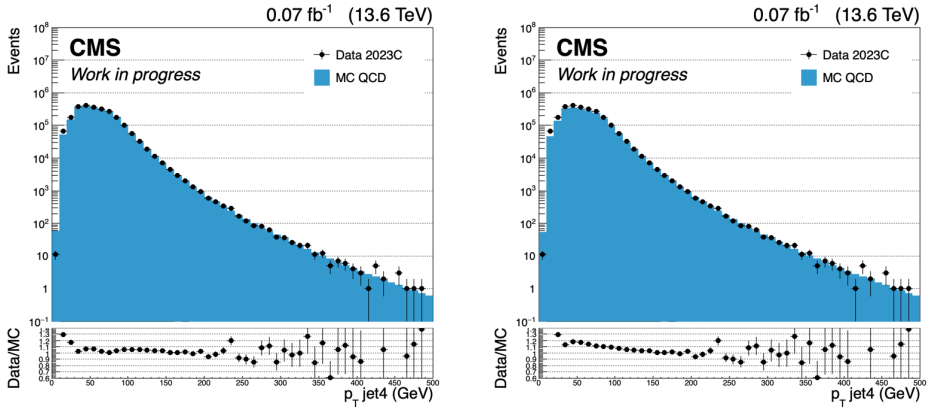


Figure 4.17: p_T distribution of the fourth p_T-leading jet for data (black dots) and MC simulated QCD multijet (blue) events selected with the control HLT path before (left) and after (right) the application of the trigger p_T SFs.

and 35 GeV

- the first three p_T-leading jets must match an HLT jet with p_T larger than 100, 88 and 70 GeV respectively
- Among the four p_T-leading jets:
  - the two jets with the highest ParticleNet CvsAll score are identified as c jets,
  - the other two jets are identified as VBF jets and are required to have $\Delta\eta$ larger than 3.8 and an invariant mass exceeding 500 GeV

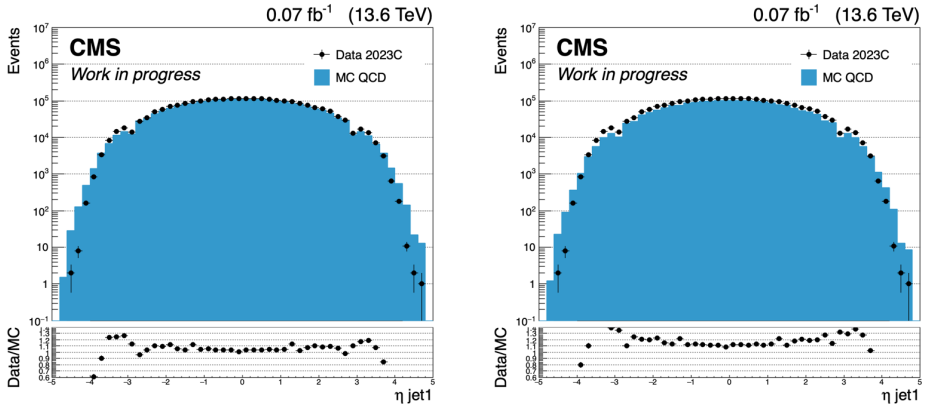


Figure 4.18: η distribution of the p_T-leading jet for data (black dots) and MC simulated QCD multijet (blue) events selected with the control HLT path before (left) and after (right) the application of the trigger p_T SFs.

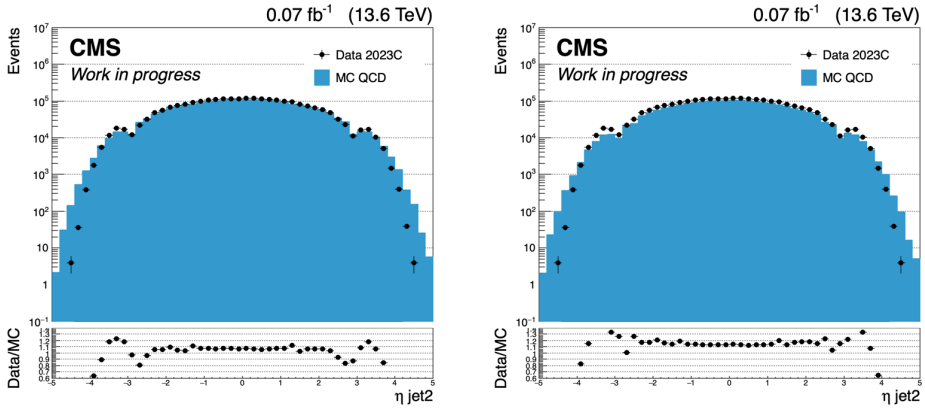


Figure 4.19: η distribution of the p_T-subleading jet for data (black dots) and MC simulated QCD multijet (blue) events selected with the control HLT path before (left) and after (right) the application of the trigger p_T SFs.

To compute the efficiency of the c tagging trigger cut, I estimated the ratio between two histograms:

- the numerator is filled with the offline CvsAll ParticleNet score of the leading c jet for events passing both the control and the signal HLT paths
- the denominator is filled with the offline CvsAll ParticleNet score of the leading c jet for events passing at least the control HLT path

Figure 4.22 (top) shows the trigger c tagging efficiency for data (blue) and MC simulated QCD multijet events (red). The corresponding SFs are calculated as usual as the ratio between these two efficiencies and are plotted in the bottom

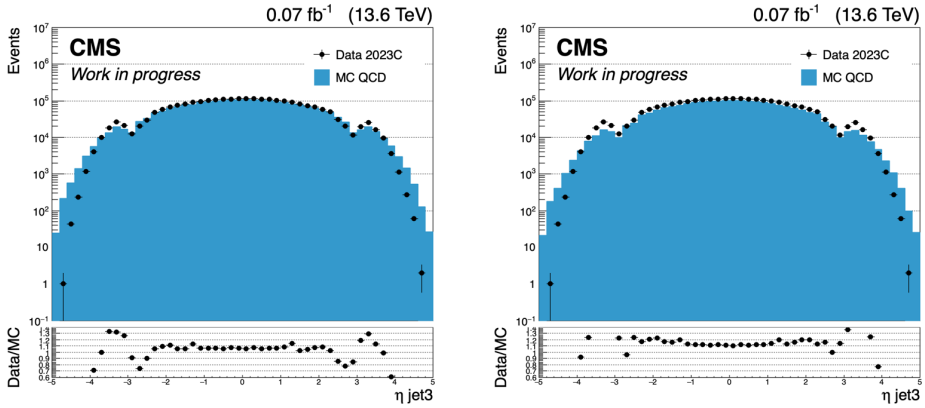


Figure 4.20: η distribution of the third p_T-leading jet for data (black dots) and MC simulated QCD multijet (blue) events selected with the control HLT path before (left) and after (right) the application of the trigger p_T SFs.

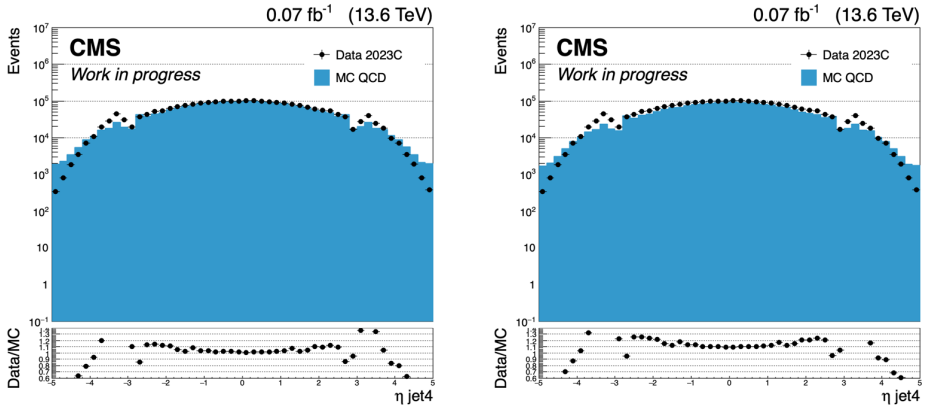


Figure 4.21: η distribution of the fourth p_T-leading jet for data (black dots) and MC simulated QCD multijet (blue) events selected with the control HLt path before (left) and after (right) the application of the trigger p_T SFs.

panel of the figure. Additionally, the efficiency from the MC signal sample is plotted in green. This efficiency is higher compared to that of the QCD sample, as the jet tagged as the leading c jet is more likely to be a true c jet in the signal events.

A future improvement for the c-tagging HLT scale factor calculation is the application of the offline data-to-MC scale factor correction, discussed at the end of Chapter 3, to the value of the offline c-tagging score.

As discussed in Section 3.6.3, the impact of such further correction is expected to be below 10%.

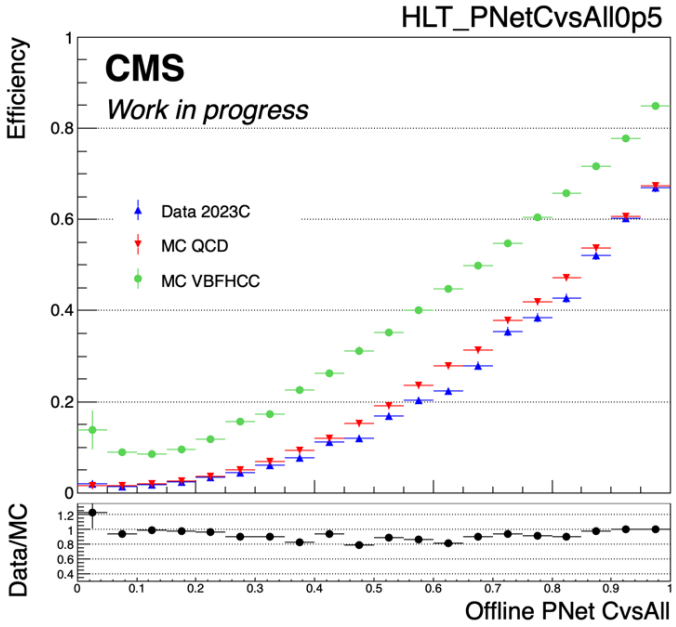


Figure 4.22: (Top) efficiency of the trigger c tagging selection estimated on data (blue) and MC simulated QCD multijet events (red) as a function of the offline CvsAll ParticleNet score of the first c jet. The same efficiency computed on MC simulated signal events is superimposed in green. (Bottom) trigger c tagging SFs, computed as the ratio between the data and MC QCD efficiencies, are plotted.

4.4 Event selection

The number of $VBF\ H \to c\bar{c}$ signal events expected to be produced with an integrated luminosity of 27 fb^{-1} is around 3300. By considering the effect of the trigger selection, this number is reduced to ~ 60.
To enhance the sensitivity to the $VBF\ H \to c\bar{c}$ signal process, I implemented a robust event selection strategy. The first step involves a pre-selection of offline-reconstructed events, following the logic of the HLT path but with slightly higher thresholds.

Subsequently, I developed a Boosted Decision Tree (BDT) algorithm, specifically designed in order to distinguish between the signal and the QCD multijet process, which is the dominant background.

4.4.1 Offline pre-selection

I applied the following event conditions to data collected during the 2023 and the MC simulated samples listed in Table 4.1:

- the event must pass the HLT trigger path *HLT_QuadPFJet100_88_70_30_PNet1CvsAll0p5_VBF3Tight*
- events containing an isolated electron or muon are vetoed to maintain orthogonality with the $VHH \to c\bar{c}$ analysis;
- the MET p_T must be smaller than 170 GeV: this condition is required since, given the signal final state, we expect no or negligible MET and the specific threshold is chosen in order to maintain orthogonality with the $VHH \to c\bar{c}$ analysis;
- the event contains at least four reconstructed jets with p_T larger than 105, 90, 75 and 30 GeV, which are matched geometrically with jets reconstructed at HLT that fired the analysis trigger path;
- among the four p_T leading jets:
  - the two with the highest CvsL score and $\eta < 2.4$ are selected as the c jet candidates from which the Higgs boson is reconstructed. These two jets must satisfy the medium and loose CvsL and CvsB c tagging working point (WP) thresholds, respectively. These thresholds have been chosen as a compromise between background rejection and preserving the possibility to use these variables as an input for an MVA analysis.
  - the other two jets are assumed to originate from the VBF process. They are required to have a pseudorapidity separation $\Delta\eta > 3.8$ and and invariant mass exceeding 500 GeV.

In order to check the agreement between data and MC simulation after this preliminary selection, I plotted the distributions of the most relevant variables describing the c tagged and VBF jets.
Figure 4.23 shows the p_T (top) and η (bottom) distributions of the leading (left) and subleading (bottom) c tagged jets for MC simulation (colored histograms) and data (black dots) collected in 2023 before the BPix issue discussed in Chapter 2.

Figure 4.24 shows the ParticleNet c tagging scores, specifically CvsL (top) and CvsB (bottom), for both the leading (left) and subleading (bottom) c

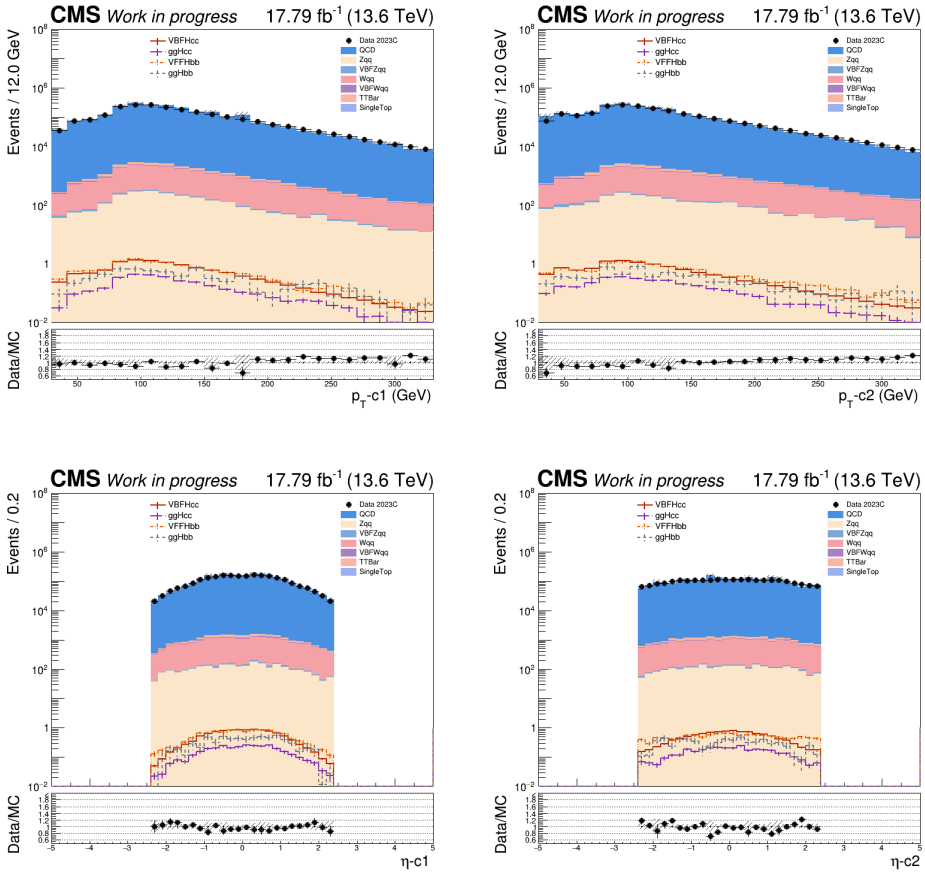


Figure 4.23: p_T and η distribution of the leading (left) and subleading (right) c tagged jets in data collected in 2023 before the BPix issue and MC simulation. The ratio between data and MC distributions is displayed in the bottom panel of each plot.

tagged jets. These are plotted for the same MC and 2023 data as mentioned above, prior to the BPix issue. Analogous plots for the data collected in 2023 after the BPix issue are displayed in Figure 4.27 and Figure 4.28.
Figures 4.25 and 4.29 show the p_T and η distributions of the leading (left) and subleading (right) VBF jets, respectively, for the period before and after the BPix issue.

One important variable for identifying the VBF topology is the quark-gluon discriminator (QvsG) for jets. In this analysis, I used the ParticleNet QvsG score. The QvsG score is defined as the ratio of the ParticleNet probability that a jet originates from a light quark to the probability that it originates from a gluon. This score is particularly useful for distinguishing between quark-initiated jets (which are more common in signal processes) and gluon-initiated jets (which dominate in QCD background processes).
Figure 4.26 and 4.30 (top) show the ParticleNet QvsG score distribution of

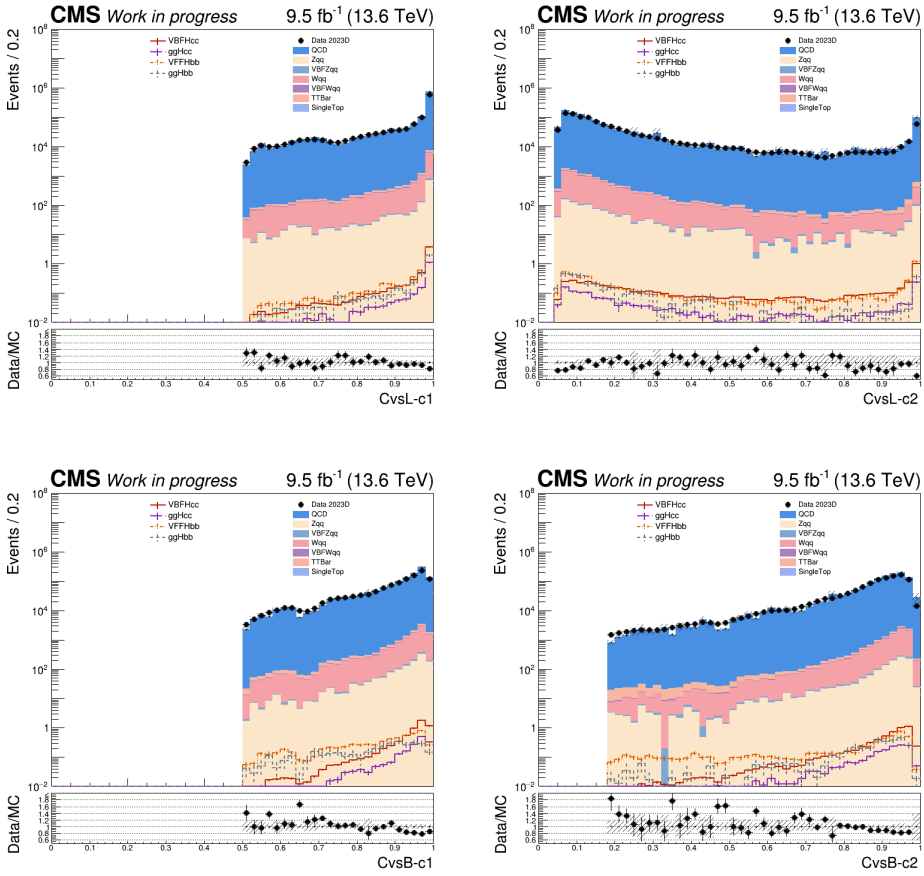


Figure 4.24: CvsL and CvsB distribution of the leading (left) and subleading (right) c tagged jets in data collected in 2023 before the BPix issue and MC simulation. The ratio between data and MC distributions is displayed in the bottom panel of each plot.

the p_T leading (left) and subleading (right) VBF jets for MC simulation and data collected respectively before and after the BPix issue.
Finally, the distance in eta $\Delta\eta$ (left) and the invariant mass (right) of the VBF jets are plotted in the bottom side of Figures 4.26 and 4.30.

It is necessary to check the data and MC agreement separately before and after the BPix issue, beacuse this issue affects track reconstruction and consequently, it impacts jet reconstruction and the performance of heavy-flavour tagging.
In all the plots described earlier, the MC simulation distributions are scaled to the integrated luminosity of data and an additional normalization factor for QCD, $k = 1.8$, is used. This factor is introduced to account for the imprecise description of the QCD MC simulation. It is derived by comparing data and MC in QCD enriched control region.
Several correction factors are also used to ensure a more accurate comparison

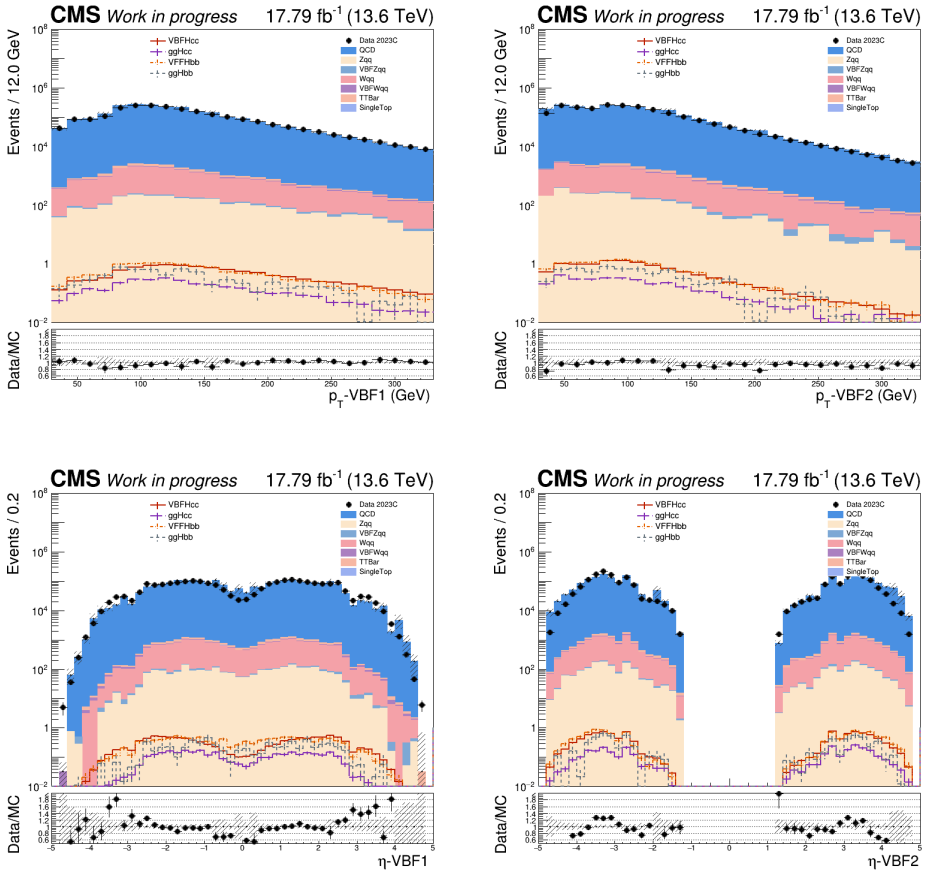


Figure 4.25: p_T and η distribution of the leading (left) and subleading (right) VBF jets in data collected in 2023 before the BPix issue and MC simulation. The ratio between data and MC distributions is displayed in the bottom panel of each plot.

between data and MC simulation: generator weights, JECs, trigger p_T and c tagging scale factors and PU reweighting, which takes into account any PU mismodelling effect in simulation.

A good overall agreement between data and MC simulation is observed in all the distributions reported here, as depicted in the bottom panel of Figures 4.23-4.30, which show the bottom panel of each plot shows the ratio of data to MC distributions. The statistical uncertainty is displayed by error bars centered on 1, which is the SM expectation value for the ratio.

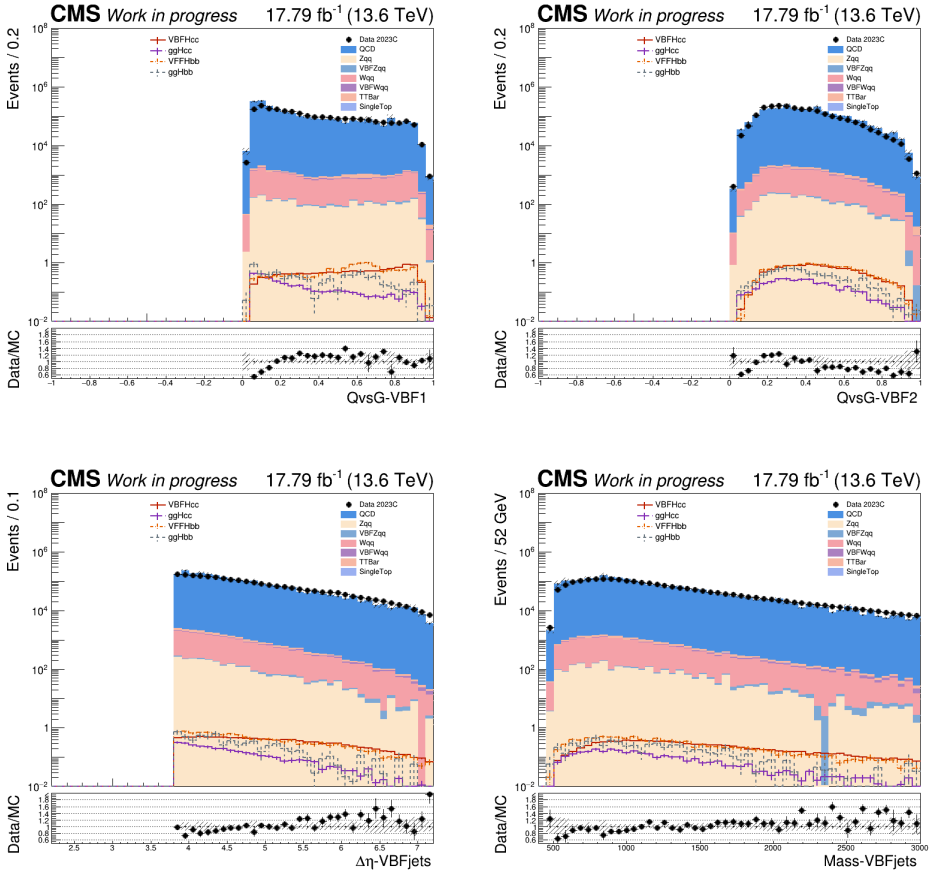


Figure 4.26: (Top) ParticleNet QvsG distribution of the leading (left) and subleading (right) VBF jets in data collected in 2023 before the BPix issue and MC simulation. (Bottom) $\Delta\eta$ (left) and invariant mass (right) distribution of the VBF jets. The ratio between data and MC distributions is displayed in the bottom panel of each plot.

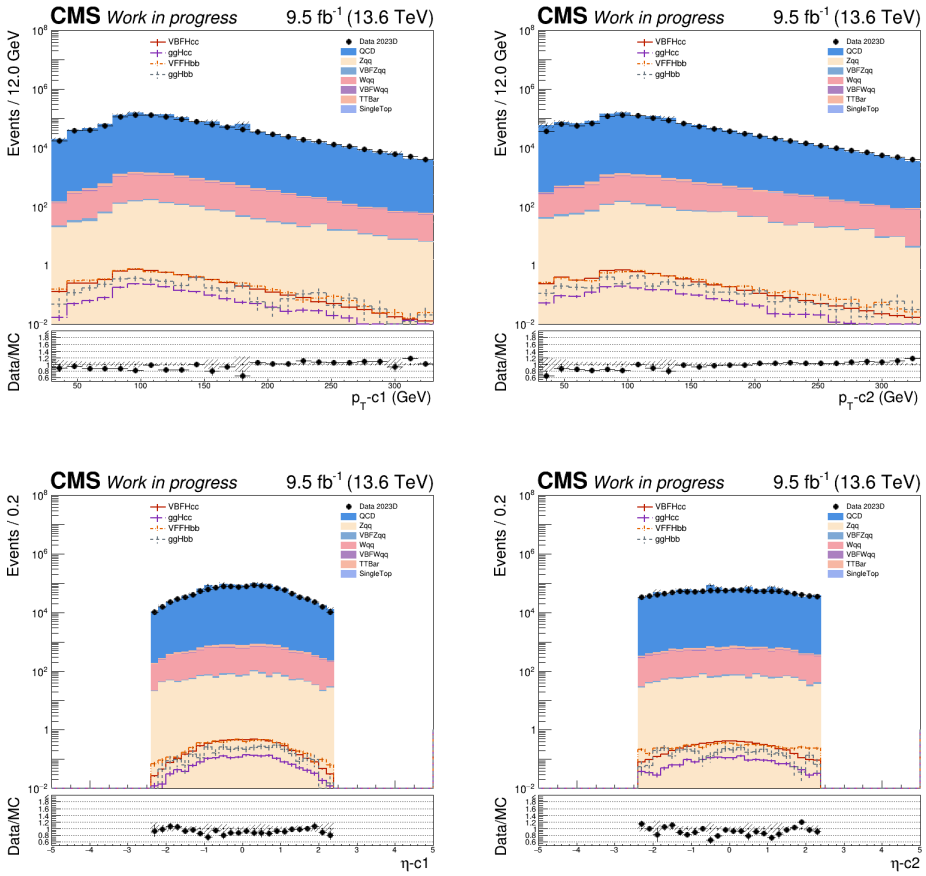


Figure 4.27: p_T and η distribution of the leading (left) and subleading (right) c tagged jets in data collected in 2023 after the BPix issue and MC simulation. The ratio between data and MC distributions is displayed in the bottom panel of each plot.

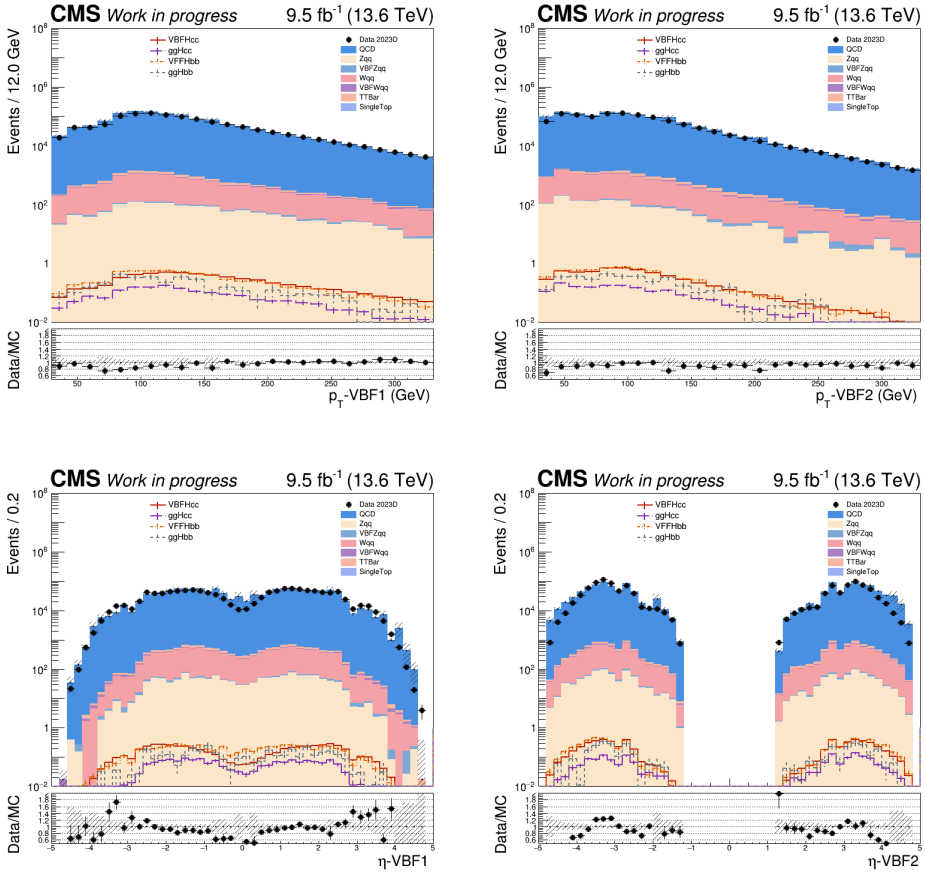


Figure 4.28: CvsL and CvsB distribution of the leading (left) and subleading (right) c tagged jets in data collected in 2023 after the BPix issue and MC simulation. The ratio between data and MC distributions is displayed in the bottom panel of each plot.

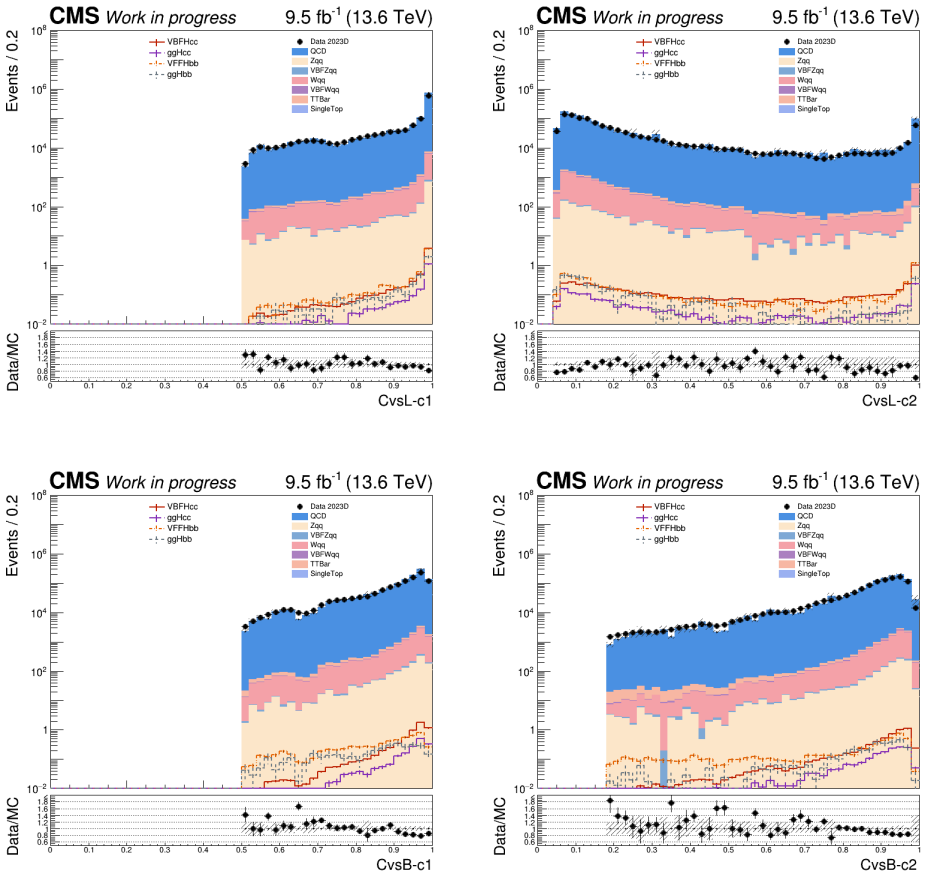


Figure 4.29: p_T and η distribution of the leading (left) and subleading (right) VBF jets in data collected in 2023 after the BPix issue and MC simulation. The ratio between data and MC distributions is displayed in the bottom panel of each plot.

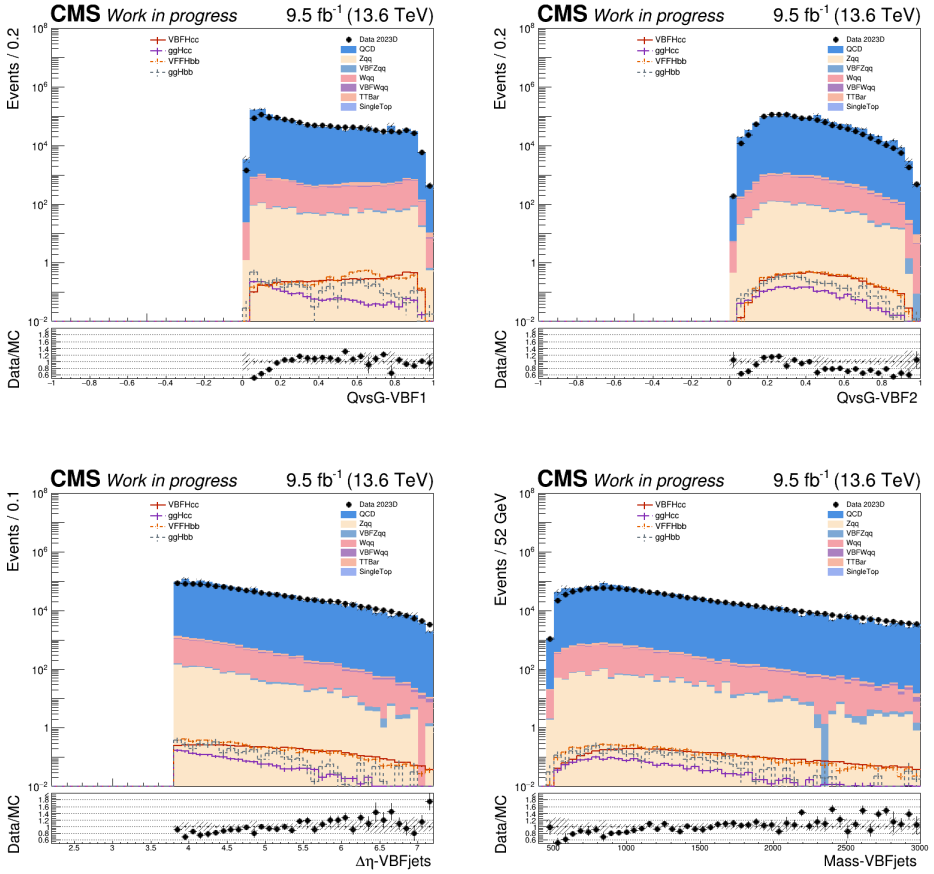


Figure 4.30: (Top) ParticleNet QvsG distribution of the leading (left) and subleading (right) VBF jets in data collected in 2023 after the BPix issue and MC simulation. (Bottom) $\Delta\eta$ (left) and invariant mass (right) distribution of the VBF jets. The ratio between data and MC distributions is displayed in the bottom panel of each plot.

4.4.2 Multivariate analysis

In order to discriminate the signal from the QCD background, I implemented a gradient boosted decision tree (BDTG) algorithm with the Root TMVA package [86]. After applying the pre-selection described in the previous section, I trained the multivariate discriminant by using the MC simulated sample of VBF $H \to c\bar{c}$ as signal and data events falling in the sidebands of the Higgs boson nominal mass region as background. It is not feasible to train the model on MC simulated QCD multijet events because the of the limited number of generated events available and because of the imprecise description affecting the QCD MC simulation. Both the signal and the background samples are equally and randomly splitted in two subsamples: one is used for the training (training sample) of the algorithm, the other for the evaluation of its performance (test sample). In general, the logic of BDT algorithms is displayed in Appendix A.
The input variables used are listed below:

- VBF related variables:
  - m_{qq}: invariant mass of the two VBF jets.
  - $|\Delta\eta_{qq}|$: absolute pseudorapidity difference of the two VBF jets.
  - $\Delta\phi_{qq}$: absolute azimuthal angle difference of the two VBF jets.
  - α_{qq}: Min$(\alpha_{q1}, \alpha_{q2})$, where $\alpha_{q1/q2}$ is the angle between lead/sublead VBF-jet and the boosted system of the VBF jet pair.
  - QvsG: ParticleNet QvsG score of the two VBF jets.
- Higgs candidate related variables:
  - c tagging: CvsL and CvsB Particle Net scores of the two c jets.
  - $p_z^{q1} + p_z^{q2} + p_z^{c1} + p_z^{c2}$: total longitudinal momentum of the selected four jets.
  - $\sum \vec{p}_T^{\,i} / \sum p_T^i$, where i=q1, q2, c1, c2: Normalized sum of the transverse momentum.
  - Angular distance: ΔR between the Higgs boson candidate and the leading and subleading VBF jets.
  - $\Delta(\phi_{qq} - \phi_{cc})$: difference of the azimuthal angle of the VBF jet pair system and the c jet pair system (H candidate).
- Event related variables:

- The jet multiplicity in the region of $|\eta| < 2.4$ above 20 GeV.
- Sum of the energy and the transverse momentum of all the jets above $p_T > 30$ GeV and $|\eta| < 2.4$ excluding the selected four jets.

These variables are chosen because they are expected to exibit a different trend in signal and QCD background.

Figure 4.31 shows the comparison between the signal and the background distributions of these input variables. Most of them demonstrate indeed a great discrimination power.
In the previous section, the data to MC comparison has been studied for most of the training input variables, finding a reasonable agreement, and therefore excluding the possibility that the discrimination power is due to mismodelling effects.

The BDT algorithm provides a score as output which can range between -1 and 1. Events that are more likely identified as signal ones, receive an output score close to 1.
Figure 4.32 shows the distribution of the BDT output score for the signal in blue and the background in red. Additionally, it displays the comparison between the distribution of the score observed in the training (dots) and in the test subsamples (color-filled). It is important to check that these two distributions are in a good agreement for both the signal and the background, in order to exclude the possibility that the model is overtrained. In order to quantify the overtraining, described in Appendix A.0.2, the distribution of the BDT score obtained independently in test and training samples are compared with a $Kolmogorov - Smirnov$ test (KS). In general, the KS test is used to decide if a sample comes from a population with a specific parent distribution. Before starting the test, it should be fixed the significance level at which we want to test the hypothesis that the sample follows the specified parent distribuiton. The test consists of measuring the maximum distance between the data and the parent distribution and comparing it with a value that depends on the fixed significance level. If the computed distance exceeds this critical value, the hypothesis is rejected.
Even though it only applies to unbinned data, the framework ROOT [87] provides a version of this test that can be performed on binned data. This function has been used to compare the BDT score distributions for training and test samples, shown in Figure 4.32. The test returns the probability that the two distributions come from the same population, which is reported in the same Figure. As it can be seen, for both the signal and the background this probability is very high ($\sim$70%), excluding the overtraining hypothesis.

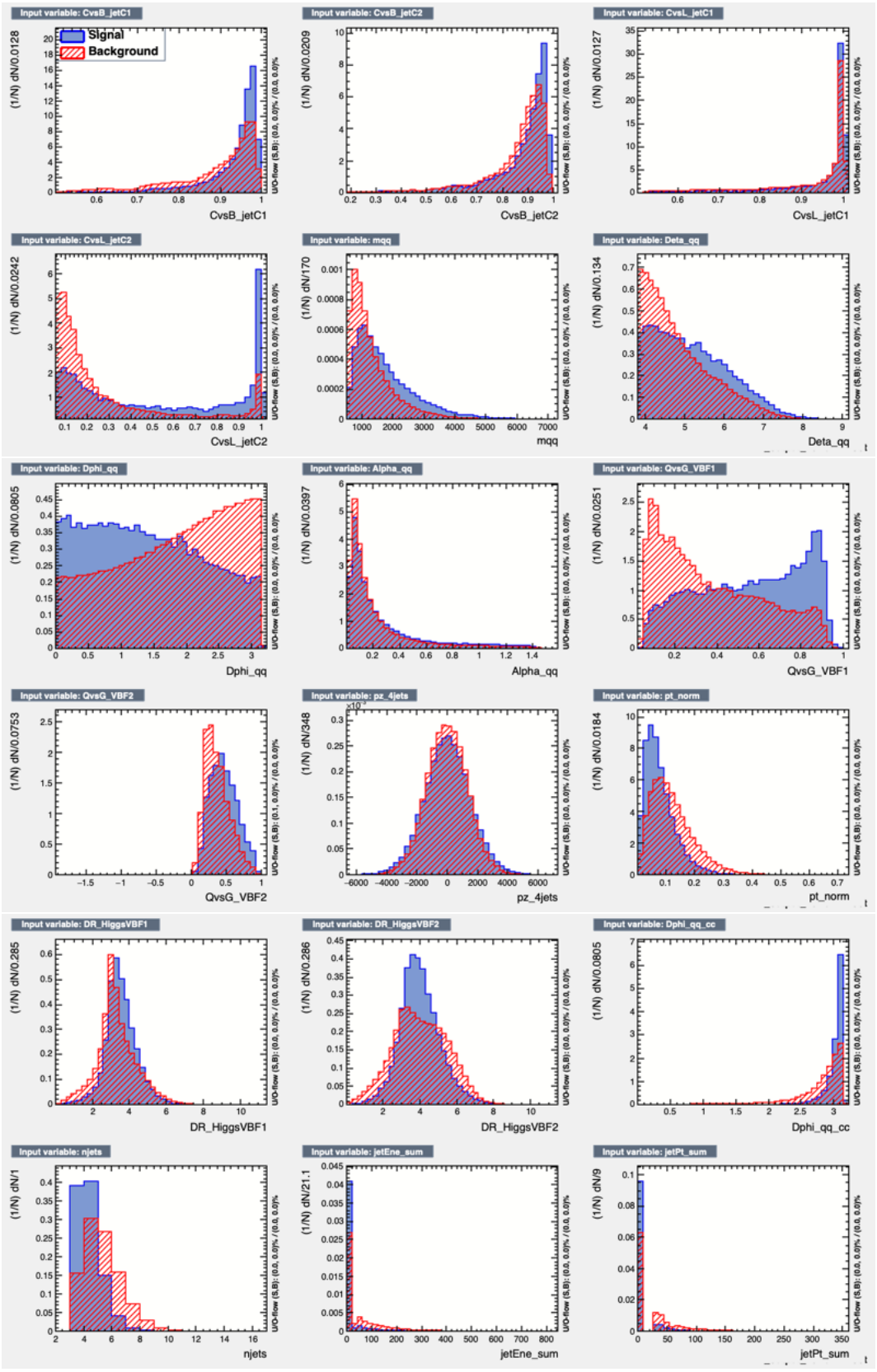


Figure 4.31: BDT input variable distributions in signal and background.

In order to estimate the performance of a MVA model, usually the receiving operating characteristic (ROC) curve is used. It corresponds to the background

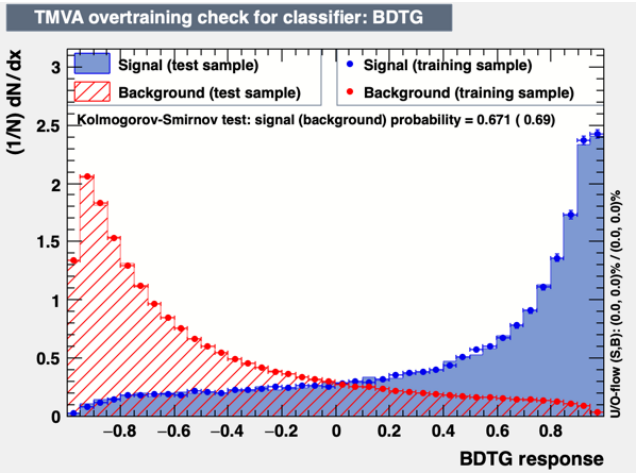


Figure 4.32: BDT output score distribution for signal (blue) and background (red). The comparison between the test (color-filled) and training (dots) distributions is shown.

rejection as a function of the signal efficiency. The better the model performs, the larger is the area under this curve: ideally we would have a background rejection probability and a signal efficiency simultaneously equal to one, meaning that we are able to classify correctly all the signal events without misidentifying any background event as signal. The ROC curve for the BDT model I trained is shown in Figure 4.33 (left).

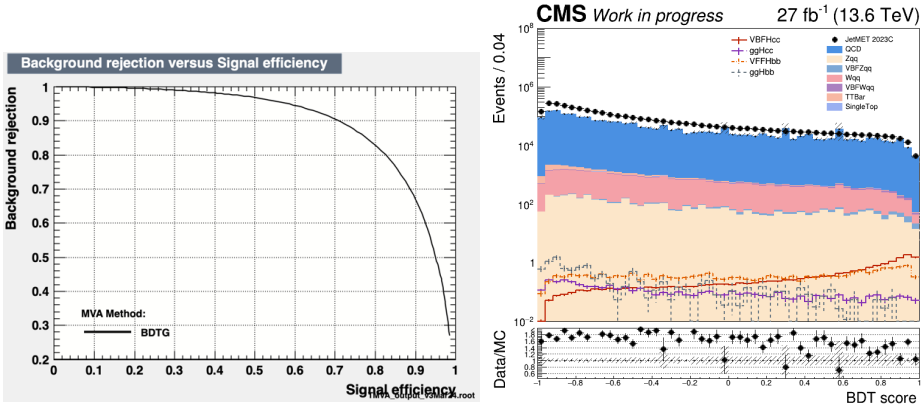


Figure 4.33: (Left) ROC curve: background . (Right) BDT output score distribution for data (black dots) collected in 2023 and MC simulation (colored). The ratio between data and MC simulation is shown in the bottom panel.

Figure 4.33 (right) shows the distribution of the BDT output score for data collected in 2023 (black dots) and MC simulation (colored). All the MC distributions are scaled to the integrated luminosity and no additional QCD normalization factor is applied. Three analysis categories are defined

on the basis of the BDT output score value, as reported in Table 4.5. For the statistical analysis described in Section 4.5 only events falling in these categories are considered and all the others are discarded. More specifically, all the events with a BDT score smaller than 0.8 are discarded and not taken into account in the next steps of this analysis. The threshold on the BDT score at 0.8 has an efficiency of $\sim$ 39% on the signal and 1.8% on the background. The final selection efficiency on signal obtained by combining the preselection with the BDT cut is roughly 0.2%.

analysis category	BDT score
cat-0	0.8 - 0.9
cat-1	0.9 - 0.95
cat-2	0.95 - 1.0

Table 4.5: Definition of the analysis categories on the basis of the BDT output score.

4.5 Statistical analysis and results

The $H \to c\bar{c}$ decay is very difficult to observe because of the small branching ratio ($\sim$3%) and the overwhelming QCD multijet background and it is statistically out of reach at the current experiments. For this reason, the goal of this analysis is to set an upper limit (UL) on the signal strength, μ, which represents the largest value of μ that would not be rejected by the likelihood-based hypothesis test at a CL of 95%. The UL on μ is determined by using the CLs method, described in Section 4.5.1, within the Combine framework [88]. Since this is a blinded analysis, the upper limit is computed not on real data but on MC samples generated accordingly to the SM (expected upper limit).

To extract the upper limit on μ, the profile likelihood ratio is used as the test statistic. This is constructed based on the modelling of the signal and background distributions of the Higgs boson candidate mass, m_{cc}, under the assumption that the data follows a Poisson distribution. For a binned analysis with N bins, the likelihood function is defined as:

$$L(\mu, \theta) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \tag{4.6}$$

where θ denote all the nuisance parameters, s_j and b_j are the expected signal and background events , respectively, in bin j of the m_{cc} spectrum and n_j is

the number of observed events in bin j.
This likelihood function takes into account both the signal and background models, which are described in more detail in Section 4.5.2.

The impact of systematic uncertainties on the yields and shapes of the signal and background distributions is incorporated in this analysis through the nuisance parameters. These uncertainties can affect both the signal and background predictions and thus must be accounted for in the likelihood function. Section 4.5.3 provides a deteiled discussion of the primary sources of systematic uncertainties considered in this analysis.

Finally, the expected upper limit on the signal strength computed with Combine is presented in Section 4.5.4. Moreover, a projection of the full Run-3 expected upper limit is estimated. This provides an indication of the sensitivity that could be achieved with the complete data collected during Run-3.

4.5.1 The CLs method for upper limits

The final goal of this search is the estimation of the upper limit on the signal strength μ. In order to explain the CLs method used for the estimation, it is useful to introduce a proper formalism in the context of likelihood-based tests for new physics [89].
First, we introduce the test statistic q_μ, which quantifies the level of incompatibility between the data and the hypothesis under test that the true signal strength is μ. The test statistic q_μ is usually defined as:

$$q_\mu = -2 \ln \lambda(\mu) \tag{4.7}$$

where $\lambda(\mu)$ is the likelihood ratio:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})} \tag{4.8}$$

In this equation, θ represents all the nuisance paramerers, e.g. the parameters that characterize the probability density functions (pdfs) used to fit the signal and the background. More specifically, $\hat{\hat{\theta}}$ is the value of θ that maximizes the likelihood evaluated at a fixed μ; $\hat{\mu}$ and $\hat{\theta}$ are the values of μ and θ that absolutely maximize the likelihood and therefore the denominator of equation 4.8 represents the maximized likelihood function (ML).
The profile likelihood as a function of μ is broadened by the presence of the nuisance parameters. This indeed reflects the loss of information about μ caused

by the systematic uncertainties, which are treated as nuisance parameters.

The value of the likelihood ratio $\lambda(\mu)$ can range between 0 and 1, approaching 1 in case of good agreement between data and the hypothesized value of μ. The test statistics q_μ, instead, increases as the incompatibility between the data and μ increases. The corresponding p-value, which quantifies the level of disagreement, is defined as the probability of obtaining a value of q_μ as large or larger than $q_{\mu,obs}$ under the assumption that the true signal strength is μ:

$$p_\mu = \int_{q_{\mu,\text{obs}}}^{\infty} f(q_\mu, \mu) dq_\mu \tag{4.9}$$

where $q_{\mu,\text{obs}}$ is the value of the test statistic q_μ observed from the data and $f(q_\mu, \mu)$ represents the pdf of q_μ under the assumption of the signal strength μ.

When calculating the upper limit, the so-called "modified test statistic for upper limit" $\tilde{q_\mu}$ is used:

$$\tilde{q_\mu} = \begin{cases} -2\ln \frac{L(\mu,\hat{\hat{\theta}})}{L(0,\hat{\hat{\theta}}(0))} & \text{if } \hat{\mu} < 0, \\ -2\ln \frac{L(\mu,\hat{\hat{\theta}})}{L(\hat{\mu},\hat{\theta})} & \text{if } 0 \leq \hat{\mu} \leq \mu, \\ 0 & \text{if } \hat{\mu} > \mu. \end{cases} \tag{4.10}$$

In this definition, by setting $\tilde{q_\mu}$ to 0 for $\hat{\mu} > \mu$, we assume the data to show lack of agreement with the hypothesized μ only if $\hat{\mu} < \mu$. Moreover, in case $\hat{\mu} < 0$, which means that statistical fluctuations result in a number of data events smaller than the one expected from the background-only hypothesis, we set the ML value of μ to 0.

When taking into account $f(q_\mu, \mu)$, the subscript of q refers to the hypothesis being tested and the second argument of f gives the value of μ assumed in the distribution of the data. The upper limit on the signal strength is the largest value of μ (μ_{UL}) such that the the p-value is larger than or equal to a fixed threshold $1 - \text{CL}$, that in the case of this analysis is set to 0.05 (95% CL):

$$p_\mu = \int_{q_{\mu,\text{obs}}}^{\infty} f(q_\mu, \mu) dq_\mu < 1 - \text{CL} \tag{4.11}$$

Any value of μ larger than μ_{UL} would result in a smaller p-value and is considered excluded at the given confidence level.

In this context, the CLs method [90] [91] is introduced to prevent overly aggressive exclusion of signal hypotheses, especially in cases where the data shows fewer events than expected from the background alone. According to the CLs method, equation 4.11 is modified as follows:

$$\frac{p_\mu}{1 - p_b} < 1 - CL, \text{with } p_b = \int_{q_0}^{\infty} f(q_0|\mu = 0) dq_0 \tag{4.12}$$

The CLs method ensures more conservative limits by accounting for downward fluctuations in the background. When the observed data is lower than expected under the background-only hypothesis, $1 - p_b$ becomes small. Dividing p_μ by $1 - p_b$ prevents the exclusion of signal hypotheses based purely on low-background fluctuations, which could otherwise lead to overly stringent upper limits.

In the case of a blinded analysis, since it is not possible to access data in the region of the mass spectrum around the Higgs boson nominal mass, the expected upper limit is quoted, which is the upper limit computed under the assumption that the data are generated accordingly to the SM.

4.5.2 Signal and Background modelling

In this analysis, the signal is extracted from a simultaneous binned maximum likelihood fit to the reconstructed mass of the Higgs boson candidate (m_{cc}) in the three categories. Five parametric analytical functions are used to describe the m_{cc} distribution in:

- signal;
- resonant $H \to b\bar{b}$, Z+jets and W+jets backgrounds;
- continuum background, which is dominated by the QCD multijet contribution and includes $t\bar{c} + X$ and single-top contributions.

While the signal and resonant backgrounds shape and yields are estimated from MC simulation, the QCD background contribution, which is poorly modelled by simulation, is estimated from data.

Signal and $H \to b\bar{b}$ background modelling

The signal is modelled by using the $VBF\, H \to c\bar{c}$ and $ggH \to c\bar{c}$ MC simulations. In particular, I fitted the m_{cc} distribution, accounting for both the two production modes, with a combination of two probability density functions:

- a Crystal Ball (CB) [92] at the peak due to its suitability to model a Gaussian core with a power-law tail caused by detector resolution effects;
- a Bernstein polynomial of 2^{nd} order for the tail, which is populated by events with the Higgs boson candidate reconstructed from the wrong pair of jets.

Figure 4.34 (left) shows the signal m_{cc} distribution scaled to the integrated luminosity of data collected in 2023 and its modelling. The $H \to b\bar{b}$ background, which has the same final state as the signal, except for the flavour of the Higgs decay products (b quarks instead of c quarks), is modelled in the exact same way, as shown in Figure 4.34 (right). In this case the shape of the peak around the Higgs boson mass is broader than the one obtained from the $H \to c\bar{c}$ sample. This is due to fact that the Higgs boson is reconstructed from two c tagged jets and not b tagged ones.

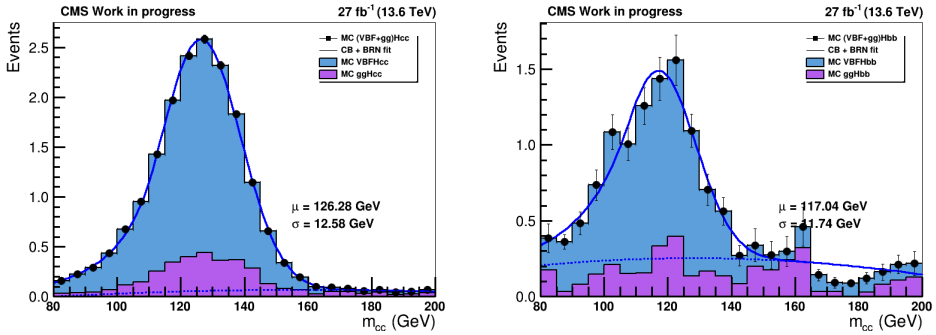


Figure 4.34: Modelling of the signal (left) and $H \to b\bar{b}$ background (right) m_{cc} distribution. The VBF contribution is plotted in blue, the ggH contribution in purple and their combination in black dots with statistical uncertainties. The distribution is fitted with a CB and a Bernstein polynomial: the overall result of the fit is plotted with a blue solid line, while the polynomial contribution is shown by the dotted blue line.

Z and W background modelling

The proximity of the Z and W bosons masses to the Higgs boson mass makes it crucial to accurately model the $Z \to q\bar{q}$ and $W \to q\bar{q}$ backgrounds. Both QCD and EWK production modes contribute to these backgrounds, and therefore dedicated MC simulations are used. Similarly to the signal modelling, I fitted the m_{cc} distribution with a CB to model the peak around 90 GeV for the Z boson and 80 GeV for the W boson, and a 2^{nd} order Bernstein polynomial for the tail. Figure 4.35 shows the modelling of $Z \to q\bar{q}$ (left) and $W \to q\bar{q}$ (right) backgrounds.

Continuum background modelling

The continuum background, mainly dominated by the QCD multijet process, is estimated by fitting the data m_{cc} distribution in the sidebands of the Higgs boson nominal mass region, i.e. $80 < m_{cc} < 104$ GeV and $146 < m_{cc} < 200$

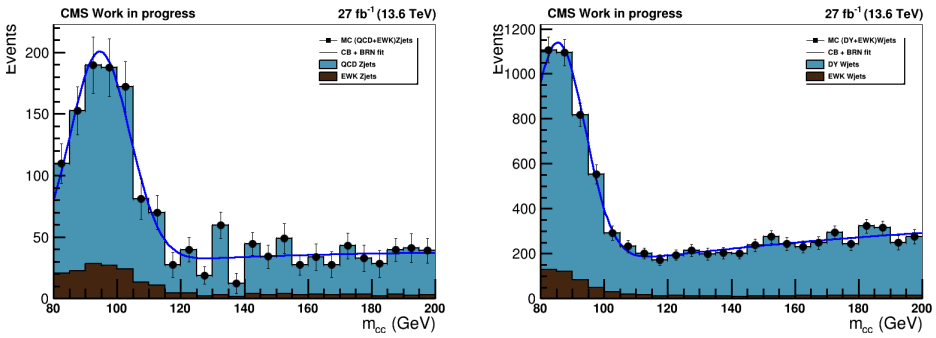


Figure 4.35: Modelling of the $Z \to q\bar{q}$ (left) and $W \to q\bar{q}$ (right) background m_{cc} distribution. The QCD production mode contribution is plotted in blue, the EWK contribution in brown and their combination in black dots with statistical uncertainties. The distribution is fitted with a CB and Bernstein polynomial: the overall result of the fit is plotted with a blue solid line, while the polynomial contribution is shown by the dotted blue line.

GeV. More specifically, in each category i, the shape of the continuum background is modelled individually by a product of exponential and polynomial functions:

$$F_i^{QCD} = exp(-b_i \cdot m_{cc}) \cdot (1 + \sum_{j=0}^{n} a_i^j \cdot m_{cc}^j) \tag{4.13}$$

where the a_i and b_i coefficients are left as free parameters of the fit. Figure 4.36 shows the modelling of the continuum background in the three analysis categories. The final fit function includes the Z/W+jets contributions described in the previous Section.

For the fit, I used a polynomial function of order zero. In order to check the bias introduced by the choice of the function for the fit, I additionally fitted these distributions with a polynomial function of 1^{st} order and checked that the final results do not change more than 10%.

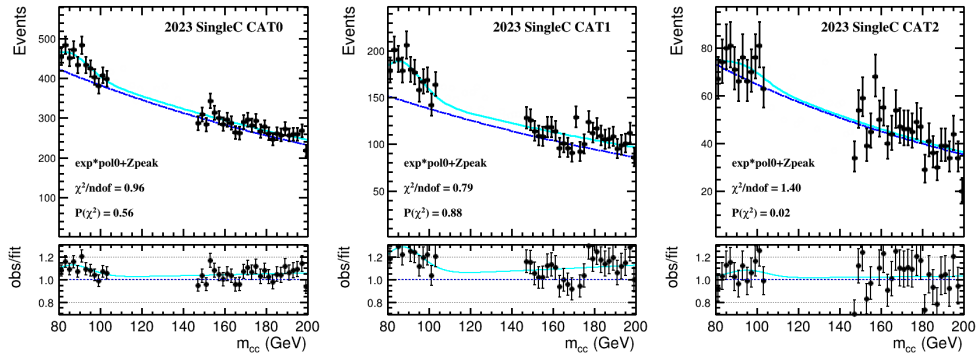


Figure 4.36: Continuum background m_{cc} distribution modelling from sideband data. Black dots represent data, the blue dashed curves represent the exponential fit for the continuum and the cyan curves represent the Z and W peaks on top of the continuum.

Yields

The yields of the signal and background processes are listed in Table 4.6 along with their statistical uncertainties. As discussed previously, for the signal and peaking backgrounds, they are computed from the MC simulation distributions, taking into account the cross section and integrated luminosity. For the continuum background, they are estimated from the fit to data.
The QCD contribution decreases in increasing BDT-score category, while the signal yield is approximately constant. The contribution coming from $gg\ H \to c\bar{c}$ is negligible, while the $H \to b\bar{b}$ contamination, even if smaller than signal, is relevant, as expected. The yields of Z/W+jets are smaller than QCD, but the m_{cc} distributions peak in the signal region.

4.5.3 Systematic uncertainties

This section provides an overview of the main systematic uncertainties that are accounted for in this analysis. These uncertainties affect mostly the rate and the shape of the signal and resonant backgrounds, since the continuum background is modelled with a data-driven method.

- **Luminosity:** the 2023 integrated luminosity measured by the CMS experiment is affected by a 1.4% systematic uncertainties [93]. This uncertainty impacts the yields of signal and all MC estimated backgrounds.

- **Parton showering and hadronization model for VBF production:** in order to evaluate the systematic uncertainty arising from the choice of the model for parton showering and hadronization in the VBF production process, I compared the signal acceptance obtained with two

Process	Yield		
	Cat0	Cat1	Cat2
$VBF\ H \to c\bar{c}$	2.66 ± 0.02	2.13 ± 0.02	2.01 ± 0.02
$ggH \to c\bar{c}$	0.14 ± 0.01	0.09 ± 0.01	0.06 ± 0.01
$VBF\ H \to b\bar{b}$	1.39 ± 0.07	0.98 ± 0.06	0.51 ± 0.04
$ggH \to b\bar{b}$	0.06 ± 0.04	0	0
Z+jets	92.02 ± 1.61	45.80 ± 1.41	24.71 ± 1.21
W+jets	241.06 ± 1.98	52.89 ± 3.42	24.21 ± 1.74
QCD	19491.4 ± 153.6	8130.8 ± 99.2	3119.7 ± 61.9

Table 4.6: Yields of the signal and background processes with statistical uncertainties.

MC generators: PYTHIA (nominal model) and HERWIG [94] [95] (alternative model).

- **Jet Energy Scale (JES):** the uncertainties on the JES correction factors affect not only the signal and background yields, but also the shape of the m_{cc} spectra. In order to study their impact, the whole analysis workflow is repeated by applying up and down uncertainty variations to the JES correction factors. By varying the four-momentum of the jets, the shape of the reconstructed Higgs boson candidate mass m_{cc} changes, as shown in Figure 4.37 (left). For this reason, JES is treated as a source of shape systematic uncertainty.

- **Jet Energy Resolution (JER):** analogously to JES, uncertainties on the JER correction factors affect both the shape and the yields of signal and backgrounds and are therefore treated as shape systematic uncertainties. The effect of JER up and down uncertainty variations on the shape of the reconstructed Higgs boson candidate mass m_{cc} is shown in Figure 4.37 (right).

- **Trigger:** the uncertainties on the trigger SFs, described in Section 2.2.7, impact on the signal and MC estimated background yields. This effect is estimated by reprocessing the whole analysis with trigger SFs up and down uncertainty variations.

- **Bias uncertainty:** the choice of the continuum background fitting function introduces a bias which is taken into account by setting a conservative systematic uncertainty of 20% on the signal, according to the spurious signal method [96].

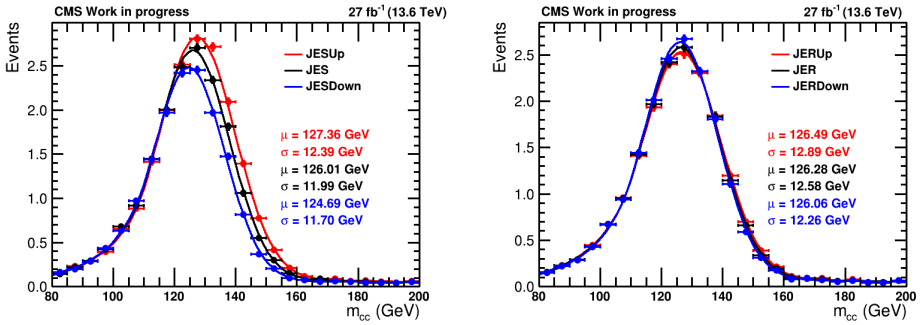


Figure 4.37: Parametric fit of m_{cc} distribution of signal process with the uncertainty variations of the JES (left) and JER (right) correction factors. The fit result of the distribution obtained with the nominal correction factors is plotted in black, the ones obtained with up and down uncertainty variations are plotted respectively in red and blue.

- **c tagging:** in the search for VBF $H \to c\bar{c}$ an MVA discriminant for multijet background suppression is exploited. This discriminant uses, among other variables, the CvsB and CvsL ParticleNet scores of the two c jet candidates. This means that shape correction c tagging scale factors must be applied to the simulation. However, as discussed in Section 3.6.3, the calibration and scale factor derivation for newly collected Run-3 2023 data is still on-going.
  However, a conservative 10% uncertainty on the signal and MC estimated background yields is considered, by looking at the overall data-MC discrepancy in W+jets control region described in Section 3.6.3.

- **Theoretical uncertainties on Higgs production cross section:** the uncertainty on the theoretical prediction of the Higgs boson production cross section arises mostly from [97]:

  - approximations used in perturbative calculations of QCD, carried out as a series expansion in the strong coupling constant (α_s)
  - uncertainty on the parton distribution functions (PDFs)
  - uncertainty on α_s

  The uncertainty on the VBF production cross section is:
  $\Delta\sigma(\text{VBF}) = ^{+0.4\%}_{-0.3\%}$ (QCD) $\pm\, 2.1\%$ (PDF $+\, \alpha_s$)
  The uncertainty on the ggF production cross section is:
  $\Delta\sigma(\text{QCD}) = ^{+4.6\%}_{-6.7\%}$ (QCD) $\pm\, 3.2\%$ (PDF $+\, \alpha_s$)

- **Theoretical uncertainty on $H \to c\bar{c}$ decay BR:** the main systematic uncertainties affecting the theoretical prediction of the $H \to c\bar{c}$ decay

branching ratio are the higher order QCD and electroweak corrections considered in the theoretical calculation, the uncertainty on the c quark mass (m_c) and α_s:
$BR(H \to c\bar{c}) = 2.891 \cdot 10^{-2} \pm 1.20\%\,(theo)\,^{+5.26}_{-0.98}\,(m_c)\,\pm 1.25\%\,(\alpha_s)$.

A summary of the systematic uncertainties discussed in this section impacting on the yield estimate of the signal and MC estimated backgrounds is provided in Table 4.7. Apart from the bias, all the systematic uncertainties are considered fully correlated across the categories.

Source of uncertainty	$VBF\,H \to c\bar{c}$	$gg\,H \to c\bar{c}$	$VBF\,H \to b\bar{b}$	$gg\,H \to b\bar{b}$	Z/W+jets
Luminosity	1.4%	1.4%	1.4%	1.4%	1.4%
VBF model	8%	-	- 8%	-	-
Trigger	3%	3%	4%	1%	3%
c tagging	10%	10%	10%	10%	10%
Higgs XS QCD scale	$^{+0.4\%}_{-0.3\%}$	$^{+4.6\%}_{-6.7\%}$	$^{+0.4\%}_{-0.3\%}$	$^{+4.6\%}_{-6.7\%}$	-
Higgs XS PDF + α_s	2.1%	3.2%	2.1%	3.2%	-
Higgs decay BR	$^{+5\%}_{-3\%}$	$^{+5\%}_{-3\%}$	-	-	-

Table 4.7: Main systematic uncertainties affecting the yield estimation of the signal and peaking backgrounds.

4.5.4 Final result: expected upper limit

The final result of this analysis is the expected upper limit on the signal strength μ. This is estimated with Combine, which takes as input the parametric shapes modelling the signal and the backgrounds along with the expected yields, which are estimated for the signal and peaking backgrounds from MC simulation and for the QCD background from data, and incorporates the most relevant systematic uncertainties as nuisance parameters of the likelihood function. Table 4.8 shows the expected upper limits established separately in the three categories and the combined value. The category with the highest BDT score shows the best sensitivity.

To summarize, the expected upper limit at 95% CL for an integrated luminosity of 27 fb^{-1}, which corresponds to p-p data collected in 2023 by the CMS

Category	Upper Limit
Cat-0	75.88
Cat-1	60.00
Cat-2	40.25
Combination	30.87

Table 4.8: Expected upper limits at 95% CL estimated for each category and their combination.

experiment, is found to be:

$$\mu = \frac{XS(VBF) \cdot BR(H \to c\bar{c})}{XS(VBF)^{SM} \cdot BR(H \to c\bar{c})^{SM}} \leq 30.87 \quad @\,95\%\,CL \tag{4.14}$$

Projection full Run-3

The expected upper limit presented above is based solely on the 2023 dataset. The trigger path I implemented for this analysis was not online during the 2022 data-taking period, so that data cannot be included. However, a projection can be made for the expected upper limit by the end of Run-3, based on the anticipated luminosity of approximately ∼360 fb^{-1} (excluding 2022).
By scaling the result from equation 4.14 to this full Run-3 luminosity, the expected upper limit would improve to approximately 8, which is expected to be comparable with the results from the $VH\,H \to c\bar{c}$ analysis in Run-2.

Projection HL-LHC

During the HL-LHC, CMS is expected to collect an integrated luminosity of approximately 3000 fb^{-1}, as illustrated in Figure 4.38.

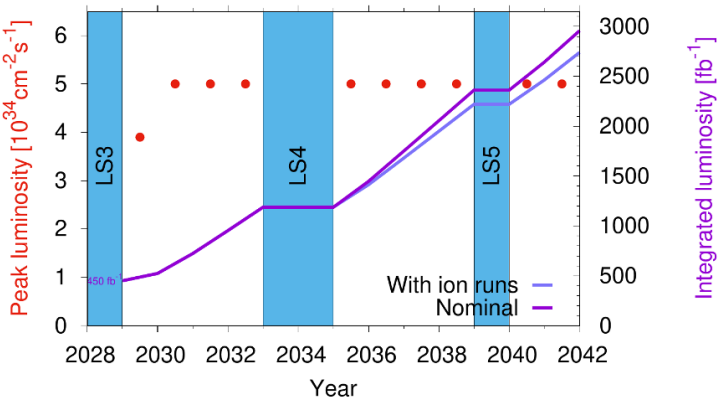


Figure 4.38: HL-LHC proton-proton luminosity expected to be collected until 2041 [98].

This corresponds to an enormous amount of data with respect to the statistics which will be available at the end of Run-3. A rough estimate of the improvement achievable in the search for $H \to c\bar{c}$ can be obtained by scaling the expected upper limit evaluated in this thesis for the 27 fb^{-1} collected in 2023 to the integrated luminosity of 3000 fb^{-1} quoted for HL-LHC.
Approximately, the expected UL on the signal strength of the $VBFH \to c\bar{c}$ process would improve to a value of 3.
Then, assuming that a similar sensitivity will be achieved in the ttH production mechanism, and including the VH channel, it is reasonable to divide this result by a factor $\sqrt{3}$, which corresponds to an upper limit of approximately 2.
This rough estimation assumes the same analysis technique performance as the ones developed up until now. It is realistic to consider that new heavy-flavour taggers and machine learning algorithms will be developed in the next future, providing a significant further boost to this important search.

Conclusion

The measurement of the Yukawa coupling of the Higgs boson to charm quarks is paramount in the context of the LHC physics program for the coming decade. On one hand, the large dataset foreseen in future LHC runs, coupled with advanced analysis techniques, offers the possibility of achieving sub-percent precision. On the other hand, even probing this coupling with the current Run-3 dataset could provide insights into BSM theories that predict enhancements to the Higgs–charm coupling beyond the SM value.

While it remains statistically out of reach with the data collected so far, the measurement of the Higgs coupling to charm quarks is crucial. Using the full Run-2 dataset, corresponding to an integrated luminosity of approximately 140 fb^{-1}, both ATLAS and CMS have set upper limits on the production cross section times branching ratio of, respectively, 11 and 14 times the SM prediction. These results have benefited significantly from state-of-the-art heavy-flavour tagging algorithms developed during Run-2. At the time of writing this thesis, these analyses have focused exclusively on the associated production mechanism (VH), where the Higgs boson is produced alongside a vector boson (W or Z).

In this thesis, the Vector Boson Fusion (VBF) production mechanism is explored for the first time. The VBF process, characterized by two forward jets with large pseudorapidity separation and minimal additional hadronic activity, offers a clean signature that has already proven effective in enhancing the sensitivity of $H \to b\bar{b}$ and $H \to \mu\mu$ measurements.
Given the novelty of this search within CMS, no dedicated trigger path for $VBF\,H \to c\bar{c}$ was available in the Run-3 trigger menu, which was finalized in 2021, at the startup of my PhD. Similarly, the Run-2 dataset, lacking the necessary trigger paths, was unsuitable for this analysis.

Therefore, a significant part of my activity was dedicated to designing and implementing a High Level Trigger (HLT) optimized for this search during

the Run-3 data-taking period. The trigger path I developed exploits the distinct VBF topology and incorporates ParticleNet−a state-of-the-art jet flavor tagging algorithm based on graph neural networks. ParticleNet, which demonstrated exceptional performance during offline jet identification in Run-2, was ported to HLT at the start of Run-3. The trigger, designed to meet the stringent CMS requirements on rate and timing constraints, was approved and deployed during the 2023 data-taking, successfully collecting a dataset of 27 fb^{-1}. Its performance and data-to-simulation agreement were validated in a QCD-dominated control region using the tag-and-probe method.

The analysis workflow presented in this thesis was entirely designed and implemented by the author. A data selection optimized for enhancing the acceptance of the events in the signal region of the phase-space is presented. This includes a MVA algorithm to discriminate the signal from the dominant QCD multijet background, that largely exploits both the VBF final state topology and the ParticleNet scores for the VBF-tagged jets and jets originating from Higgs decays. Reasonable agreement between data and Monte Carlo simulations was observed, considering that only the most critical correction factors were applied.

The analysis workflow achieved an efficiency of approximately 0.2%, yielding $\sim$7 expected signal events.

The final result, extracted using a binned maximum likelihood fit across three categories in the high BDT score region, sets an expected 95% CL upper limit on $\sigma(\text{VBF } H) \times BR(H \to c\bar{c})$. With an integrated luminosity of 27 fb^{-1}, this limit is approximately $\sim$30. Extrapolating to the full Run-2 luminosity, the upper limit would be comparable to that obtained in the VH Run-2 analyses.

Looking ahead, the full Run-3 dataset, expected to total 360 fb^{-1}, is projected to improve the upper limit to approximately $\sim$8. Further enhancements could come from exploiting advanced jet flavor tagging algorithms like Particle Transformer [99] and optimizing the MVA discriminator using deep-learning techniques tailored to the larger dataset. Moreover, the new triggers of the VBF parking strategy could increase the signal acceptance.

A dedicated study is needed to deal with the irreducible $H \to b\bar{b}$ background, that is treated as a background in this study. Further investigations are needed to reduce its contribution (e.g. implementing a dedicated MVA discriminator).

Another possibility is to simultaneously extract Hbb and Hcc couplings as done in ATLAS Run-2 analysis [100].

Given the continuous improvements in analysis techniques and the growing LHC dataset, it is plausible that the observation of the Higgs–charm coupling could become feasible before the conclusion of the HL-LHC.

Appendices

Appendix A

Multivariate analysis and Boosted Decision Trees

A.0.1 Introduction to multivariate analysis

The multivariate analysis includes several techniques able to separate two classes of events on the basis of a large number of observables that characterize each event [101]. For this reason, it has gained a great interest in High Energy Physics.

In general, an MVA is performed in order to discriminate signal from background events. For the purpose of illustrating how a multivariate analysis may achieve this goal, a generic set of signal and background events is considered. Figure A.1 shows the scatter plots that describe the distributions of two variables, x_1 and x_2. Blue circles are used for signal events, while red triangles represent background events.

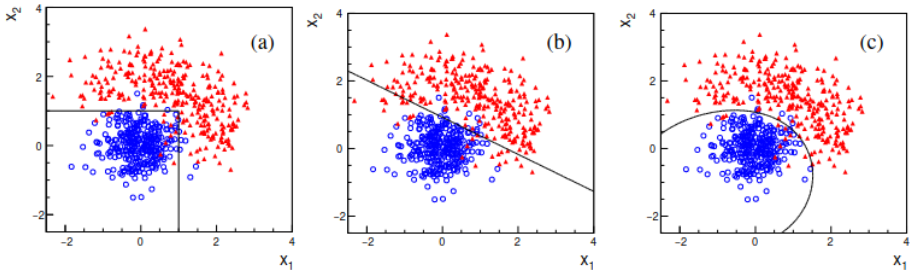


Figure A.1: Scatter plot of two variables corresponding to two hypotheses: signal and background. Event selection could be based, e.g. on (a) cuts, (b) a linear boundary, (c) a nonlinear boundary [101].

Three different decision boundaries have been drawn in an attempt to sep-

arate the two classes of events. In this case, because of the curved nature of the distribution, the best solution is given by the nonlinear boundary. The decision boundary is a surface in the n-dimensional space of input variables, described by the equation $y(x) = y_{cut}$, where y_{cut} is a constant. Once the decision boundary has been chosen, it is used to build the acceptance region. The side of the boundary that contains the largest fraction of signal events becomes the acceptance region, while the other side is the rejection one. It follows that if an event falls in the acceptance region, it is classified as a signal event, otherwise it is registered as a background.

The MVA classifier that I used for this study is the Boosted Decision Tree. Its implementation, together with the most important features, will be presented in the following section.

A.0.2 Boosted Decision Tree

Decision Tree

Decision tree is one of the most consolidated tools for supervised learning [102]. Its structure is represented in Figure A.2. Let us consider a training set of N observations belonging to two classes (signal and background), each described by D input variables x_i. The entire set of events form the root node of the tree. The node is then splitted in two subsets, L (left) and R (right), by applying the first cut. Each of the two children nodes is splitted again by following the same logic and this procedure is repeated recursively until a certain condition is verified (stopping criterion).

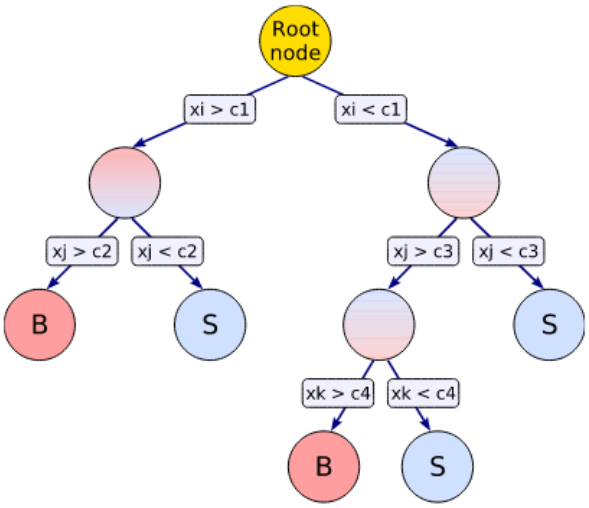


Figure A.2: Schematic view of a decision tree. Starting from the root node, a sequence of binary splits using the discriminating variables x_i is applied to the data. Each split uses the variable that at this node gives the best separation between signal and background when being cut on. The same variable may thus be used at several nodes, while others might not be used at all. The leaf nodes at the bottom end of the tree are labeled "S" for signal and "B" for background depending on the majority of events that end up in the respective nodes. [86]

A probability $P(t) = N_t/N$, where N_t is the number of observations in the node t, is assigned to each node. Also the posterior probability for each class can be evaluated: for the signal it is given by $P(S|t) = N_S/N_t$, being N_S the number of signal events in the node, and it is similarly defined for the background. A node predicts into the class with the largest posterior probability. It follows that the complementary probability, i.e. the posterior probability of the other class, represents the training error of that node: $\epsilon(t) = min_{\gamma\in\{S,B\}}P(\gamma|t)$. In order to have a criterion to choose the best cut to apply on each node, the *node impurity* is introduced. It is defined by Equation A.1.

$$i(t) = \phi(P(S|t), P(B|t)) \qquad 0 \leq \phi(p,q) \leq \frac{1}{2} \tag{A.1}$$

where $p = P(S|t)$ and $q = P(B|t)$.
The cut should be chosen in order to reduce the weighted impurity when passing from the parent to the children nodes. The weighted impurity of a node $I(t)$ corresponds to the impurity multiplied by the probability assigned to that node. It follows that the best cut is the one that maximizes the *impurity gain*:

$$\Delta I = I(t_{parent}) - I(t_{children}) \tag{A.2}$$

If there is no gain in splitting a node, it becomes a *leaf*. In the specific case of this study, the function chosen to evaluate the node impurity (*Boost Type*) is the *Gini diversity index*, defined by:

$$\phi(p, q) = 1 - p^2 - q^2 \tag{A.3}$$

Boosted Decision Tree

Boosting consists of combining weak classifiers into a new more stable one, with smaller error [103]. In a Boosted Decision Tree, several weak decision trees are built in a loop, such that each tree gives more importance to the observables misclassified by the previous tree, and at the end they are combined in one final strong tree. The boosting algorithm employed in the current study is the Adaptive Boost (*AdaBoost*), and its description is provided in the following.

At first, a weight w_n, normalized to the total number of events in the set, is assigned to each observable x_n in the training set. Then, the observables are given as input to a loop. In each step of this loop a decision tree is trained in order to get an hypothesis function that attributes $+1$ to the observables falling in the acceptance region, -1 otherwise:

$$h_t : x \to \{-1; 1\} \tag{A.4}$$

The training error is calculated as the sum of the weights of the observables misclassified by this hypothesis:

$$\epsilon_t = \sum_{n=1}^{N} w_n^t I(yn \neq h_t(x_n)) \tag{A.5}$$

where I assumes value 1 if the condition it is applied to is verified, 0 if it is not. Then, a weight α_t is assigned to the hypothesis t:

$$\alpha_t = \beta \; log \frac{1 - \epsilon_t}{\epsilon_t}, \tag{A.6}$$

where β is a fixed parameter (*Adaboost Beta*). Finally, the observable weights are updated:

$$w_n^{t+1} = \frac{w_n^t \, e^{-\alpha_t \, y_n \, h_t(x_n)}}{\sum_{n=1}^{N} w_n^t \, e^{-\alpha_t \, y_n \, h_t(n)}} \tag{A.7}$$

At the end of the loop, the final output function is constructed by summing all the hypothesis, each multiplied by its weight:

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{A.8}$$

where T is the total number of trained trees. Therefore, $f(x)$ is the BDT score assigned to each event in the training set.

In order to avoid the uncontrolled growth of these trees a stopping criterion has been imposed. It is based on two conditions:

- the size of each node cannot go under a fixed minimal threshold *Minimum node size*;
- the tree depth cannot exceed a fixed maximal value *Maximum depth*.

Bagging

The AdaBoost algorithm suffers from overtraining, meaning that it is sensitive to statistical fluctuations of the training sample. In order to avoid this issue, the Bagging technique is employed. Each tree is provided with a training sample that contains only a fraction of events randomly picked up from the original set *Bagged Boost sample fraction*. The not picked events form the "out of the bag" validation sample.

Bibliography

[1] ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". *Physics Letters B* 716.1 (2012), 1–29. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2012.08.020. URL: http://dx.doi.org/10.1016/j.physletb.2012.08.020.

[2] CMS Collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". *Physics Letters B* 716.1 (2012), 30–61. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2012.08.021. URL: http://dx.doi.org/10.1016/j.physletb.2012.08.021.

[3] 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.

[4] S. Navas et al. (Particle Data Group). "Review of Particle Physics". *To be published in Phys. Rev. D* 110, 030001 (2024).

[5] ATLAS Collaboration. *Measurements of WH and ZH production with Higgs boson decays into bottom quarks and direct constraints on the charm Yukawa coupling in* 13 TeV *pp collisions with the ATLAS detector*. 2024. arXiv: 2410.19611 [hep-ex]. URL: https://arxiv.org/abs/2410.19611.

[6] CMS Collaboration. "Search for Higgs Boson Decay to a Charm Quark-Antiquark Pair in Proton-Proton Collisions at $\sqrt{s}$=13 TeV ". *Physical Review Letters* 131.6 (Aug. 2023). ISSN: 1079-7114. DOI: 10.1103/physrevlett.131.061801. URL: http://dx.doi.org/10.1103/PhysRevLett.131.061801.

[7] Daniel Dominguez/CERN. *Particles of the Standard Model of particle physics*. URL: https://www.home.cern/science/physics/standard-model.

[8] Otto Nachtmann. *Elementary particle physics: concepts and phenomena*. Springer Science & Business Media, 2012.

[9] Peter Ware Higgs. "Broken symmetries, massless particles and gauge fields". *Phys. Lett.* 12 (1964), pp. 132–133.

[10] Peter W Higgs. "Broken symmetries and the masses of gauge bosons". *Physical Review Letters* 13.16 (1964), p. 508.

[11] François Englert and Robert Brout. "Broken symmetry and the mass of gauge vector mesons". *Physical Review Letters* 13.9 (1964), p. 321.

[12] ALEPH Collaboration et al. "Electroweak measurements in electron–positron collisions at W-boson-pair energies at LEP". *Physics reports* 532.4 (2013), pp. 119–244.

[13] Terhi Aaltonen et al. "Combination of CDF and D0 W-boson mass measurements". *Physical Review D* 88.5 (2013), p. 052018.

[14] ATLAS Collaboration. "Measurement of the W-boson mass in pp collisions at $\sqrt{s} = 7$TeV with the ATLAS detector". *The European Physical Journal C* 78.2 (2018). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-017-5475-4. URL: http://dx.doi.org/10.1140/epjc/s10052-017-5475-4.

[15] CDF Collaboration†‡ et al. "High-precision measurement of the W boson mass with the CDF II detector". *Science* 376.6589 (2022), pp. 170–176. DOI: 10.1126/science.abk1781. eprint: https://www.science.org/doi/pdf/10.1126/science.abk1781. URL: https://www.science.org/doi/abs/10.1126/science.abk1781.

[16] CMS Collaboration. *Measurement of the W boson mass in proton-proton collisions at sqrts = 13 TeV*. Tech. rep. Geneva: CERN, 2024. URL: https://cds.cern.ch/record/2910372.

[17] ATLAS Collaboration. "Combined Measurement of the Higgs Boson Mass from the $H \to \gamma\gamma$ and $H \to ZZ^* \to 4l$ decay channels with the ATLAS detector using $\sqrt{s} = 7$, 8 and 13 TeV *pp* collision data". *Physical Review Letters* 131.25 (Dec. 2023). ISSN: 1079-7114. DOI: 10.1103/physrevlett.131.251802. URL: http://dx.doi.org/10.1103/PhysRevLett.131.251802.

[18] ATLAS Collaboration. "A detailed map of Higgs boson interactions by the ATLAS experiment ten years after the discovery". *Nature* 607.7917 (July 2022), 52–59. ISSN: 1476-4687. DOI: 10.1038/s41586-022-04893-w. URL: http://dx.doi.org/10.1038/s41586-022-04893-w.

[19] CMS Collaboration. "A portrait of the Higgs boson by the CMS experiment ten years after the discovery". *Nature* 607.7917 (July 2022), 60–68. ISSN: 1476-4687. DOI: 10.1038/s41586-022-04892-x. URL: http://dx.doi.org/10.1038/s41586-022-04892-x.

[20] ATLAS Collaboration. "Measurement of the Higgs boson mass with $H \to \gamma\gamma$ decays in 140 fb1 of $\sqrt{s}$ TeV pp collisions with the ATLAS detector". *Physics Letters B* 847 (Dec. 2023), p. 138315. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2023.138315. URL: http://dx.doi.org/10.1016/j.physletb.2023.138315.

[21] CMS Collaboration. *Measurement of the Higgs boson mass and width using the four leptons final state*. Tech. rep. Geneva: CERN, 2023. URL: https://cds.cern.ch/record/2871702.

[22] CMS Collaboration. "Evidence for Higgs boson decay to a pair of muons". *Journal of High Energy Physics* 2021.1 (Jan. 2021). ISSN: 1029-8479. DOI: 10.1007/jhep01(2021)148. URL: http://dx.doi.org/10.1007/JHEP01(2021)148.

[23] S Heinemeyer et al. "Handbook of LHC Higgs cross sections: 3. Higgs properties" (2013). arXiv: 1307.1347.

[24] Albert M Sirunyan et al. "Evidence for Higgs boson decay to a pair of muons". *Journal of High Energy Physics* 2021.1 (2021), pp. 1–68.

[25] Nina M. Coyle, Carlos E. M. Wagner, and Viska Wei. "Bounding the charm Yukawa coupling". *Phys. Rev. D* 100 (7 2019), p. 073013. DOI: 10.1103/PhysRevD.100.073013. URL: https://link.aps.org/doi/10.1103/PhysRevD.100.073013.

[26] Huilin Qu and Loukas Gouskos. "Jet tagging via particle clouds". *Phys. Rev. D* 101 (5 2020), p. 056019. DOI: 10.1103/PhysRevD.101.056019. URL: https://link.aps.org/doi/10.1103/PhysRevD.101.056019.

[27] CMS collaboration. "Search for Higgs Boson and Observation of Z Boson through Their Decay into a Charm Quark-Antiquark Pair in Boosted Topologies in Proton-Proton Collisions at $\sqrt{s} = 13$ Tev". *Physical Review Letters* 131.4 (July 2023). ISSN: 1079-7114. DOI: 10.1103/physrevlett.131.041801. URL: http://dx.doi.org/10.1103/PhysRevLett.131.041801.

[28] *Search for Higgs boson production in association with a charm quark in the diphoton decay channel*. Tech. rep. Geneva: CERN, 2024. URL: http://cds.cern.ch/record/2905239.

[29] Huilin Qu and Luca Mastrolorenzo. "Constraining the Higgs-charm coupling at CMS. Constraining the Higgs-charm coupling at CMS" (2022). URL: http://cds.cern.ch/record/2802770.

[30] Aneesh V Manohar. "Effective field theories". *Perturbative and nonperturbative aspects of quantum field theory: Proceedings of the 35. Internationale Universitätswochen für Kern-und Teilchenphysik, Schladming, Austria, March 2–9, 1996*. Springer, 2007, pp. 311–362.

[31] Aneesh V. Manohar. *Introduction to Effective Field Theories*. 2018. arXiv: 1804.05863 [hep-ph]. URL: https://arxiv.org/abs/1804.05863.

[32] Adam Falkowski. "Lectures on effective field theories". *Lecture notes for Saclay* (2017).

[33] Ilaria Brivio and Michael Trott. "The standard model as an effective field theory". *Physics Reports* 793 (Feb. 2019), 1–98. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2018.11.002. URL: http://dx.doi.org/10.1016/j.physrep.2018.11.002.

[34] John Ellis et al. "Updated global SMEFT fit to Higgs, diboson and electroweak data". *Journal of High Energy Physics* 2018.6 (June 2018). ISSN: 1029-8479. DOI: 10.1007/jhep06(2018)146. URL: http://dx.doi.org/10.1007/JHEP06(2018)146.

[35] Artemis Sofia Giannakopoulou, Patrick Meade, and Mauro Valli. *How charming can the Higgs be?* 2024. arXiv: 2410.05236 [hep-ph]. URL: https://arxiv.org/abs/2410.05236.

[36] CMS collaboration. *CMS Luminosity measurements*. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[37] Tai Sakuma. *Cutaway diagrams of CMS detector*. 2019. URL: https://cds.cern.ch/record/2665537.

[38] CMS Collaboration. "The CMS experiment at the CERN LHC". *Journal of Instrumentation* 3.08 (2008), S08004–S08004. DOI: 10.1088/1748-0221/3/08/s08004. URL: https://doi.org/10.1088/1748-0221/3/08/s08004.

[39] UZH CMS wiki contributors. *CMS coordinate system — UZH CMS wiki*. [Online; accessed 3-October-2021]. 2017. URL: https://wiki.physik.uzh.ch/cms/_detail/latex:cms_coordinate_system.png?id=latex%3Aexample_spherical_coordinates.

[40] John C. Collins, Davison E. Soper, and George Sterman. "Heavy particle production in high-energy hadron collisions". *Nuclear Physics B* 263.1 (1986), pp. 37–60. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(86)90026-X. URL: https://www.sciencedirect.com/science/article/pii/055032138690026X.

[41] CMS Collaboration. "Precise mapping of the magnetic field in the CMS barrel yoke using cosmic rays". *Journal of Instrumentation* 5.03 (2010), T03021–T03021. DOI: 10.1088/1748-0221/5/03/t03021. URL: https://doi.org/10.1088/1748-0221/5/03/t03021.

[42] CMS Collaboration. "Description and performance of track and primary-vertex reconstruction with the CMS tracker". *Journal of Instrumentation* 9.10 (2014), P10009–P10009. DOI: 10.1088/1748-0221/9/10/p10009. URL: https://doi.org/10.1088/1748-0221/9/10/p10009.

[43] CMS Collaboration. "The CMS Phase-1 Pixel Detector Upgrade". *JINST* 16 (2020), P02027. 84 p. DOI: 10.1088/1748-0221/16/02/P02027. arXiv: 2012.14304. URL: https://cds.cern.ch/record/2748381.

[44] CMS Collaboration. "Performance and operation of the CMS electromagnetic calorimeter". *Journal of Instrumentation* 5.03 (2010), T03010–T03010. ISSN: 1748-0221. DOI: 10.1088/1748-0221/5/03/t03010. URL: http://dx.doi.org/10.1088/1748-0221/5/03/T03010.

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

[46] CMS Collaboration. "Performance of CMS hadron calorimeter timing and synchronization using test beam, cosmic ray, and LHC beam data". *Journal of Instrumentation* 5.03 (2010), T03013–T03013. ISSN: 1748-0221. DOI: 10.1088/1748-0221/5/03/t03013. URL: http://dx.doi.org/10.1088/1748-0221/5/03/T03013.

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

[48] CMS Collaboration. "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV". *Journal of Instrumentation* 7.10 (2012), P10002–P10002. ISSN: 1748-0221. DOI: 10.1088/1748-0221/7/10/p10002. URL: http://dx.doi.org/10.1088/1748-0221/7/10/P10002.

[49] CMS Collaboration. “The CMS muon project: Technical Design Report” (1997).

[50] CMS Collaboration. “Development of the CMS detector for the CERN LHC Run 3. Development of the CMS detector for the CERN LHC Run 3”. *JINST* 19.05 (2024), P05064. DOI: 10.1088/1748-0221/19/05/P05064. arXiv: 2309.05466. URL: https://cds.cern.ch/record/2870088.

[51] M. Abbas et al. “Benchmarking LHC background particle simulation with the CMS triple-GEM detector”. *Journal of Instrumentation* 16.12 (Dec. 2021), P12026. ISSN: 1748-0221. DOI: 10.1088/1748-0221/16/12/p12026. URL: http://dx.doi.org/10.1088/1748-0221/16/12/P12026.

[52] CMS Collaboration. “Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$”. *JINST* 15 (2020), P10017. DOI: 10.1088/1748-0221/15/10/P10017. arXiv: 2006.10165 [hep-ex].

[53] CMS Collaboration. *Enriching the physics program of the CMS experiment via data scouting and data parking*. 2024. arXiv: 2403.16134 [hep-ex]. URL: https://arxiv.org/abs/2403.16134.

[54] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. “The anti-kt jet clustering algorithm”. *Journal of High Energy Physics* 2008.04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. URL: https://dx.doi.org/10.1088/1126-6708/2008/04/063.

[55] CMS collaboration. “Particle-flow reconstruction and global event description with the CMS detector”. *Journal of Instrumentation* 12.10 (2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003. URL: https://dx.doi.org/10.1088/1748-0221/12/10/P10003.

[56] R. Frühwirth. “Application of Kalman filtering to track and vertex fitting”. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 262.2 (1987), pp. 444–450. ISSN: 0168-9002. DOI: https://doi.org/10.1016/0168-9002(87)90887-4. URL: https://www.sciencedirect.com/science/article/pii/0168900287908874.

[57] The CMS Collaboration. “Description and performance of track and primary-vertex reconstruction with the CMS tracker”. *Journal of Instrumentation* 9.10 (Oct. 2014), P10009–P10009. ISSN: 1748-0221. DOI: 10.1088/1748-0221/9/10/p10009. URL: http://dx.doi.org/10.1088/1748-0221/9/10/P10009.

[58] Felice Pantaleo. "New Track Seeding Techniques for the CMS Experiment". CERN, 2017. URL: http://cds.cern.ch/record/2293435.

[59] CMS Collaboration. "CMS tracking performance in 2023" (2023). URL: http://cds.cern.ch/record/2882249.

[60] Daniele Bertolini et al. "Pileup per particle identification". *Journal of High Energy Physics* 2014.10 (Oct. 2014). ISSN: 1029-8479. DOI: 10.1007/jhep10(2014)059. URL: http://dx.doi.org/10.1007/JHEP10(2014)059.

[61] CMS collaboration. "Pileup mitigation at CMS in 13 TeV data". *Journal of Instrumentation* 15.09 (2020), P09018. DOI: 10.1088/1748-0221/15/09/P09018. URL: https://dx.doi.org/10.1088/1748-0221/15/09/P09018.

[62] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "FastJet user manual: (for version 3.0.2)". *The European Physical Journal C* 72.3 (Mar. 2012). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-012-1896-2. URL: http://dx.doi.org/10.1140/epjc/s10052-012-1896-2.

[63] CMS Collaboration. "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV". *Journal of Instrumentation* 12.02 (Feb. 2017), P02014–P02014. ISSN: 1748-0221. DOI: 10.1088/1748-0221/12/02/p02014. URL: http://dx.doi.org/10.1088/1748-0221/12/02/P02014.

[64] CMS Collaboration. "CMS JETMET public results" (). URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsJME.

[65] CMS collaboration. "Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector". *Journal of Instrumentation* 14.07 (2019), P07004. DOI: 10.1088/1748-0221/14/07/P07004. URL: https://dx.doi.org/10.1088/1748-0221/14/07/P07004.

[66] CMS collaboration. "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV". *Journal of Instrumentation* 13.05 (2018), P05011. DOI: 10.1088/1748-0221/13/05/P05011. URL: https://dx.doi.org/10.1088/1748-0221/13/05/P05011.

[67] The CMS collaboration. "A new calibration method for charm jet identification validated with proton-proton collision events at $\sqrt{s} = 13$ TeV". *Journal of Instrumentation* 17.03 (2022), P03014. DOI: 10.1088/1748-0221/17/03/P03014. URL: https://dx.doi.org/10.1088/1748-0221/17/03/P03014.

[68] CMS collaboration. "Measurement of $B\overline{B}$ angular correlations based on secondary vertex reconstruction at $\sqrt{s} = 7\,\text{TeV}$". *Journal of High Energy Physics* 2011.3 (2011). ISSN: 1029-8479. DOI: 10.1007/jhep03(2011)136. URL: http://dx.doi.org/10.1007/JHEP03(2011)136.

[69] Wolfgang Waltenberger, Rudolf Frühwirth, and Pascal Vanlaer. "Adaptive vertex fitting". *Journal of Physics G: Nuclear and Particle Physics* 34.12 (2007), N343. DOI: 10.1088/0954-3899/34/12/N01. URL: https://dx.doi.org/10.1088/0954-3899/34/12/N01.

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

[71] CMS collaboration. "Performance of the DeepJet b tagging algorithm using 41.9/fb of data from proton-proton collisions at 13TeV with Phase 1 CMS detector". *CMS-DP-2018-058* (2018). URL: http://cds.cern.ch/record/2646773.

[72] CMS Collaboration. "CMS B-tagging and vertexing public results" (). URL: https://cms-results.web.cern.ch/cms-results/public-results/publications/BTV/index.html.

[73] Yue Wang et al. *Dynamic Graph CNN for Learning on Point Clouds*. 2019. arXiv: 1801.07829 [cs.CV]. URL: https://arxiv.org/abs/1801.07829.

[74] "Identification of highly Lorentz-boosted heavy particles using graph neural networks and new mass decorrelation techniques" (2020). URL: https://cds.cern.ch/record/2707946.

[75] "Run 3 commissioning results of heavy-flavor jet tagging at $\sqrt{s}$ =13.6 TeV with CMS data using a modern framework for data processing" (2024). URL: https://cds.cern.ch/record/2898463.

[76] Paolo Nason. "A New method for combining NLO QCD with shower Monte Carlo algorithms". *JHEP* 11 (2004), p. 040. DOI: 10.1088/1126-6708/2004/11/040. arXiv: hep-ph/0409146 [hep-ph].

[77] Simone Alioli et al. "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX". *JHEP* 06 (2010), p. 043. DOI: 10.1007/JHEP06(2010)043. arXiv: 1002.2581 [hep-ph].

[78] Christian Bierlich et al. “A comprehensive guide to the physics and usage of PYTHIA 8.3” (2022). arXiv: 2203.11601 [hep-ph]. URL: https://arxiv.org/abs/2203.11601.

[79] Baptiste Cabouat and Torbjörn Sjöstrand. “Some dipole shower studies”. *The European Physical Journal C* 78.3 (Mar. 2018). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-018-5645-z. URL: http://dx.doi.org/10.1140/epjc/s10052-018-5645-z.

[80] Keith Hamilton et al. “NLO Higgs boson production via gluon fusion matched with shower in POWHEG”. *JHEP* 10 (2012), p. 155. DOI: 10.1007/JHEP10(2012)155. arXiv: 1206.3572 [hep-ph].

[81] Keith Hamilton et al. “Improving NLO-parton shower matched simulations with higher order matrix elements”. *JHEP* 05 (2013), p. 082. DOI: 10.1007/JHEP05(2013)082. arXiv: 1212.4504 [hep-ph].

[82] Johan Alwall et al. “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations”. *JHEP* 07 (2014), p. 079. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph].

[83] M Mangano. “The so-called MLM prescription for ME/PS matching”. *Fermilab ME/MC Tuning Workshop*. Vol. 4. 2002.

[84] Rikkert Frederix and Stefano Frixione. “Merging meets matching in MC@NLO”. *Journal of High Energy Physics* 2012.12 (Dec. 2012). ISSN: 1029-8479. DOI: 10.1007/jhep12(2012)061. URL: http://dx.doi.org/10.1007/JHEP12(2012)061.

[85] CMS Collaboration. “Measurement of the Higgs boson production via vector boson fusion and its decay into bottom quarks in proton-proton collisions at $\sqrt{s} = 13$ TeV”. *Journal of High Energy Physics* 2024.1 (Jan. 2024). ISSN: 1029-8479. DOI: 10.1007/jhep01(2024)173. URL: http://dx.doi.org/10.1007/JHEP01(2024)173.

[86] Andreas Hoecker et al. “TMVA-toolkit for multivariate data analysis”. *arXiv preprint physics/0703039* (2007).

[87] Rene Brun and Fons Rademakers. “ROOT—an object oriented data analysis framework”. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 389.1-2 (1997), pp. 81–86.

[88] CMS Collaboration. *The CMS statistical analysis and combination tool: COMBINE*. 2024. arXiv: 2404.06614 [physics.data-an]. URL: https://arxiv.org/abs/2404.06614.

[89] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". *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.

[90] Thomas Junk. "Confidence level computation for combining searches with small statistics". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 434.2 (1999), pp. 435–443. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(99)00498-2. URL: https://www.sciencedirect.com/science/article/pii/S0168900299004982.

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

[92] M. Oreglia. "A Study of the Reactions $\psi' \to \gamma\gamma\psi$". Other thesis. Dec. 1980.

[93] CMS Collaboration. "Measurement of the offline integrated luminosity for the CMS proton-proton collision dataset recorded in 2023" (2024). URL: https://cds.cern.ch/record/2904808.

[94] Johannes Bellm et al. "Herwig 7.0/Herwig++ 3.0 release note". *Eur. Phys. J. C* 76.4 (2016), p. 196. DOI: 10.1140/epjc/s10052-016-4018-8. arXiv: 1512.01178 [hep-ph].

[95] Manuel Bähr et al. "Herwig++ physics and manual". *The European Physical Journal C* 58.4 (Nov. 2008), 639–707. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-008-0798-9. URL: http://dx.doi.org/10.1140/epjc/s10052-008-0798-9.

[96] *Recommendations for the Modeling of Smooth Backgrounds*. Tech. rep. Geneva: CERN, 2020. URL: https://cds.cern.ch/record/2743717.

[97] LHC Higgs working group. "LHC Higgs working group public results" (). URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHWG.

[98] *LHC commissioning web page*. URL: https://lhc-commissioning.web.cern.ch/.

[99] Huilin Qu, Congqiao Li, and Sitian Qian. *Particle Transformer for Jet Tagging*. 2024. arXiv: 2202.03772 [hep-ph]. URL: https://arxiv.org/abs/2202.03772.

[100] ATLAS Collaboration. "Direct constraint on the Higgs–charm coupling from a search for Higgs boson decays into charm quarks with the ATLAS detector". *The European Physical Journal C* 82.8 (Aug. 2022). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-022-10588-3. URL: http://dx.doi.org/10.1140/epjc/s10052-022-10588-3.

[101] G. Cowan. *Topics in statistical data analysis for high-energy physics*. 2010. arXiv: 1012.3589 [physics.data-an].

[102] Ilya Narsky and Frank C Porter. *Statistical analysis techniques in particle physics*. Wiley Online Library, 2013.

[103] Yann Coadou. "Boosted decision trees". *IN2P3 School of Statistics 2016*. 2016.