UNIVERSITÄT HAMBURG

INSTITUT FÜR EXPERIMENTALPHYSIK

Calibration of the jet transverse momentum resolution and search for heavy resonances decaying into a Z and a Higgs boson with the CMS experiment

Dissertation
zur Erlangung des Doktorgrades
an der Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Physik
der Universität Hamburg

vorgelegt von

Andrea Malara

Hamburg, 2021

Gutachter/innen der Dissertation:	Prof. Dr. Johannes Haller
	Prof. Dr. Elisabetta Gallo
Zusammensetzung der Prüfungskommission:	Prof. Dr. Johannes Haller
	Prof. Dr. Elisabetta Gallo
	Prof. Dr. Dieter Horns
	Prof. Dr. Gudrid Moortgat-Pick
	Dr. Andreas Hinzmann
Vorsitzender der Prüfungskommission:	Prof. Dr. Dieter Horns
Datum der Disputation:	11.11.2021
Vorsitzender des Fach-Promotionsausschusses Physik:	Prof. Dr. Wolfgang Hansen
Leiter des Fachbereichs Physik:	Prof. Dr. Günter H. W. Sigl
Dekan der Fakultät MIN:	Prof. Dr. Heinrich Graener

“The world lies in the hands of those that have the courage to dream
and who take the risk of living out their dreams
- each according to his or her own talent.”
- Paulo Coelho

Abstract

The work presented in this thesis comprises two analyses performed using 13 TeV proton-proton collision data recorded in the years 2016 to 2018 with the CMS detector at the LHC. The dataset analysed corresponds to an integrated luminosity of about 137 fb^{-1}. First, the calibration of the jet transverse momentum resolution is described, and second, the search for heavy resonances decaying to a Z and Higgs boson is presented.

In the first part of this work, the technique for the calibration of the jet transverse momentum resolution adopted in the CMS Collaboration is described in detail. The method exploits QCD dijet events to calibrate the width of the jet response distribution in simulated events to match the one in data. A wide range in jet transverse momentum ranging from 100 GeV to 1 TeV is covered up to a pseudorapidity of $|\eta| = 5.2$. The uncertainties of the results obtained are improved by up to a factor of 3 compared to the previous measurement. In particular, a thorough statistical treatment of the systematic uncertainties leads to an enhanced calibration precision. The results derived with dijet events are combined for the first time with those obtained in Z+jet topologies, allowing the extension down to transverse momenta of 40 GeV.

In the second part of this work, a search for the resonant production of a hypothetical spin-1 massive particle decaying into a Z and a Higgs boson is presented. Predicted by a multitude of theories, such diboson resonances are promising particles to resolve several shortcomings of the Standard Model. The analysis is carried out in the final state with two electrons or muons and a large-radius jet, identified as originating from the hadronic decays of a Higgs boson. In particular, the 4-prong ($\mathrm{H} \to \mathrm{qqqq}$) and c flavour ($\mathrm{H} \to \mathrm{c\bar{c}}$) decays are targeted for the first time in this context. Recent advances in machine learning-based jet tagging algorithms are exploited to maximise the sensitivity of this search. A full statistical combination with an analysis targeting invisible Z boson decays is performed within the context of this thesis. No excess over the Standard Model expectation is observed and upper limits on the production cross section of the resonance are placed. Resonances decaying exclusively into a Z and a Higgs boson are excluded below masses of 2.45 and 2.72 TeV, depending on the theoretical model under consideration. The results obtained show a sensitivity to high resonance masses that exceeds that of the $\mathrm{H} \to \mathrm{b\bar{b}}$ channel despite its much larger branching fraction.

Zusammenfassung

Die vorgelegte Arbeit umfasst zwei Analysen, die auf dem Datensatz basieren, der mit dem CMS-Detektor am LHC in Proton-Proton-Kollisionen in den Jahren 2016 bis 2018 aufgenommenen wurde. Der Datensatz entspricht einer integrierten Luminosität von ungefähr 137 fb^{-1}. Zunächst wird die Kalibrierung der Auflösung des transversalen Jetimpulses beschrieben, danach wird die Suche nach schweren, in ein Z- und ein Higgs-Boson zerfallenden Resonanzen präsentiert.

Im ersten Teil dieser Arbeit wird die Methodik der Kalibration des transversalen Jetimpulses, die in der CMS-Kollaboration zur Anwendung kommt, detailliert beschrieben. QCD-Ereignisse mit zwei Jets werden genutzt, um die Breite der Verteilung des Ansprechverhaltens des Detektors in simulierten Ereignissen derjenigen in Daten anzugleichen. Ein großer Bereich des Transversalimpulses von Jets, von 100 GeV bis hin zu 1 TeV, wird auf diese Weise bis zu einer Pseudorapidität von $|\eta| = 5.2$ abgedeckt. Die Unsicherheiten der erhaltenen Ergebnisse sind um einen Faktor von bis zu 3 kleiner als die Unsicherheiten der vorherigen Messung. Insbesondere führt eine gründliche Behandlung der systematischen Unsicherheiten hierbei zu einer Verbesserung der Kalibrationsgenauigkeit. Die Ergebnisse aus Ereignissen mit zwei Jets werden zum ersten Mal mit denen aus Z+Jet-Ereignissen kombiniert, sodass die Kalibrierung auf Jets mit einem Transversalimpuls von mindestens 40 GeV ausgedehnt werden kann.

Im zweiten Teil dieser Arbeit wird die Suche nach der resonanten Produktion eines hypothetischen massiven Teilchens mit Spin 1, das in ein Z- und ein Higgs-Boson zerfällt, vorgestellt. Solche Resonanzen, die von einer Vielzahl an Theorien vorhergesagt werden, sind vielversprechende Kandidaten, um diverse Unzulänglichkeiten des Standardmodells zu beheben. Die Suche wird im Endzustand mit zwei Elektronen oder Myonen und einem Jet mit großem Radius, der als Produkt des hadronischen Zerfalls eines Higgs-Bosons identifiziert wurde, durchgeführt. Insbesondere werden Zerfälle in vier leichte Quarks ($\mathrm{H} \rightarrow \mathrm{qqqq}$) und solche in zwei Charm-Quarks ($\mathrm{H} \rightarrow \mathrm{c\bar{c}}$) zum ersten Mal in diesem Zusammenhang studiert. Kürzliche Fortschritte beim Jet-Tagging mithilfe maschinellen Lernens werden ausgenutzt, um die Sensitivität dieser Suche zu maximieren. Eine vollständige statistische Kombination dieser Suche mit einer Analyse, die unsichtbare Zerfälle von Z-Bosonen untersucht, wird im Rahmen dieser Arbeit durchgeführt. Es wird keine Abweichung von der Standardmodellvorhersage beobachtet, sodass obere Grenzen auf den Produktionswirkungsquerschnitt der neuen Resonanz gesetzt werden. Resonanzen, die ausschließlich in ein Z- und ein Higgs-Boson zerfallen, werden, je nach betrachtetem theoretischen Modell, unterhalb von Massen von 2.45 und 2.72 TeV ausgeschlossen. Die im Ergebnis erreichte Sensitivität auf hohe Resonanzmassen übertrifft diejenige des $\mathrm{H} \rightarrow \mathrm{b\bar{b}}$ Kanals trotz dessen deutlich größeren Verzweigungsverhältnisses.

List of own contributions

Jet energy calibration

I am responsible for the derivation of the η-dependent data-to-simulation scale factors (SFs) for the calibration of the jet transverse momentum resolution within the CMS Collaboration from March 2018 until present. The results of my work have been used by all CMS analyses published using the full Run 2 dataset. The main contributions include:

- Derivation of the SFs using the dijet method for data collected in the years 2017 and 2018 of the Run 2, corresponding to the pre-Legacy reconstruction, and the full Run 2 dataset, corresponding to the Legacy reconstruction.
- Reduction of the uncertainties associated to the calibration procedure with the application of a precise statistical treatment of the systematic errors.
- Extension of the calibration procedure to the low-p_{T} regime, obtained with the combination of the dijet and Z+jet results for p_{T}-dependent SFs.
- Documented the novel technique and the results in CMS-internal analysis note.
- Regular presentations in CMS working group meetings.

These results, together with the complete jet calibration results for Run 2, have been published as a *Detector Performance Summary* (DPS) [1]. My contributions are:

- Primary editor of the DPS.
- Regular presentations in CMS working group meetings.
- Coordination of all analysis groups that contributed to this publication.
- CMS-internal approval presentation for the publication of the results.

Moreover, I presented these results at the following international conferences:

- 40th International Conference on High Energy Physics (ICHEP) – Title: "CMS jet and missing transverse momentum performance at Run 2 and prospects for Run 3" – Presentation available at [2] – Proceedings published in Ref. [3].
- 13th International Workshop on Boosted Object Phenomenology, Reconstruction and Searches in HEP (BOOST), – Title: "Jet reconstruction and calibration for LHC Runs 2 and 3 in CMS" – Presentation available at [4].

The work was performed under the supervision of Dr. Anastasia Karavdina between March 2018 and November 2019. I supervised Alexander Paasch (PhD student) for additional studies on the jet transverse momentum resolution measurement.

Search for diboson resonance

I am the primary analyser of the CMS analysis searching for a heavy resonance decaying into a Z and a Higgs boson as presented in this thesis. The results are compared to existing analyses targeting different Higgs boson decay modes. My contributions include:

- Definition and optimisation of the analysis strategy based on the Z boson decays into a pair of charged leptons and neutrinos, as well as the 4-prong ($\mathrm{H} \to \mathrm{VV}^* \to \mathrm{qqqq}$) and c flavour ($\mathrm{H} \to \mathrm{c\bar{c}}$) decays of the Higgs boson.
- Study of machine learning-based approaches for jet identification and comparison with already existing algorithms provided by the CMS Collaboration.
- Study of optimised jet tagging selection criteria depending on the jet transverse momentum.
- Investigation of the impact of the c flavour composition of Higgs boson-initiated jets.
- Study and validation of the parametrisation of the background.
- Combination of results obtained with the different Z boson decay modes (electrons, muons and neutrinos).
- Comparison to existing analyses targeting different final states.
- Documented the analysis strategy in CMS-internal analysis note.
- Regular presentations in CMS working group meetings.
- Presentation for the endorsement of the results presented in this thesis. The publication of these results, not yet officially approved, is foreseen in the near future.

This work was performed under the supervision of Prof. Dr. Johannes Haller, Dr. Roman Kogler and Dr. Paolo Gunnellini. I supervised Tom Sokolinski (master student), whose work is published as Ref. [5] and focuses on the study of the neutrino channel and follows closely the analysis strategy and workflow defined by me.

Contents

Introduction

Consistently tested to high precision by a multitude of different experiments, the Standard Model (SM) of particle physics encloses our current best understanding of physics at the smallest scales. However, theoretical limitations and experimental evidence indicate the necessity of a more fundamental description of nature.

The Large Hadron Collider (LHC) is the world's largest and most powerful particle collider. It was designed for the search for the long-sought Higgs boson, which culminated with its discovery in 2012, as well as new physics phenomena. The data collected by the LHC experiments helped in consolidating the predictions of the SM in many of its aspects. The work presented in this thesis is based on 13 TeV proton-proton collision data recorded in the years 2016 to 2018 with the CMS detector at the LHC. The analysed data correspond to an integrated luminosity of about $137\,\mathrm{fb}^{-1}$.

Abundantly produced at hadron colliders, jets are the experimental signature of strongly interacting particles. Used to infer the properties of the initial particle, the reconstructed jets must be corrected for the detector response and differences between data and simulation. A miscalibration of the jet energy and resolution can lead to a momentum imbalance in the event and, consequently, also to a mismeasurement of the missing transverse momentum. Therefore, essentially any LHC physics analysis heavily relies on accurately calibrated jets for a detailed understanding of their properties.

The procedure adopted in the CMS Collaboration to calibrate the jet energy resolution, i.e. the width of the jet response distribution, in simulated events to match the one in data is described in detail. The main results are derived using the momentum conservation in the transverse plane of QCD dijet events. Two complementary methods exploit this topology to provide a wide coverage in pseudorapidity ($|\eta| < 5.2$). The dijet results, obtained for jets with high transverse momentum ($p_\mathrm{T} > 100\,\mathrm{GeV}$), are combined with those derived in the Z+jet topology, allowing for the first time the extension towards the low-p_T region. The combination of these orthogonal channels is performed entirely in the context of this thesis.

The large amount of data collected at the LHC necessitates the development of advanced analysis techniques to improve beyond the statistical precision only. Novel algorithms based on machine learning (ML) can provide improved performance, for example, in the field of jet tagging, which is of particular interest for searches for new physics and SM measurements.

The Higgs boson discovery marked the beginning of a new era in experimental particle physics. Not only does it represent the last missing piece of the SM, but it is also the first tangible portal to the vacuum. If on one side the greatest triumphs of the Standard Model (e.g. QED, QCD, flavour physics) are all consequences of the gauge principles, its mysteries (e.g. the origin of the masses, flavour mixing, dark energy and inflation, hierarchy problems), can be related to the vacuum. To this end, the study of the Higgs boson properties will help us answer the unresolved questions of the SM.

A plethora of theories extending the SM, often related to the Higgs boson sector, have been proposed and are currently being tested by experimental searches at the LHC. A search for the resonant production of a hypothetical spin-1 massive particle decaying into a Z and a Higgs boson is presented in this thesis. The final states with two electrons or muons and light-flavoured hadronic decays of the Higgs boson, reconstructed as a single large-radius jet, are studied. In particular, the 4-prong and the c flavour decays are investigated for the first time in this context. The event selection, heavily relying on ML-based jet tagging algorithms, is optimised to maximise the sensitivity while ensuring that the selected data are independent of those used in other analyses, particularly the one targeting the $\mathrm{H} \to \mathrm{b\overline{b}}$ final state, for a future statistical combination of the results. The combination of this search and the analysis targeting the invisible decays of the Z boson is performed within the context of this thesis. The results obtained show a sensitivity to high resonance masses that exceeds that of the $\mathrm{H} \to \mathrm{b\overline{b}}$ channel despite its much larger branching fraction.

This thesis is organised as follows. A description of the most relevant aspects of the SM is provided in chapter 1, together with an overview of beyond-the-Standard-Model theories and the relevant experimental results obtained by the CMS Collaboration. The experimental framework, both the LHC complex and the CMS detector, is outlined in chapter 2. The reconstruction of proton-proton collision events performed within the CMS Collaboration is described in chapter 3, where particular emphasis is given to the jet reconstruction and calibration. An overview of the state-of-the-art techniques for boosted jet tagging based on deep learning is given in chapter 4. A comprehensive discussion of the jet transverse momentum resolution measurement follows in chapter 5. The search for a diboson resonance is presented in chapter 6. The thesis concludes with a summary and prospects for future work.

1 Theoretical basis and motivations

The Standard Model of particle physics is the theory that explains the phenomenology of the microscopic world and describes its elementary constituents. Confirmed by numerous high energy physics experiments, it provides a predictive formulation of the electromagnetic, the weak and the strong forces, as well as the spontaneous symmetry breaking mechanism. To date, despite all the experimental data in agreement with the prediction by the Standard Model, there are several observational puzzles and structural issues that prevent the Standard Model from being a complete theory of fundamental interactions. For example, it does not include a description of the gravitational force, which is by several orders of magnitude weaker compared to the other forces. Therefore, there is the need to explore energies beyond the electroweak scale in search of symmetries with higher dimension than those that characterise the Standard Model. This chapter contains an overview of the Standard Model, a brief description of the Higgs mechanism and its phenomenology at the LHC. Finally, the shortcomings of the Standard Model are summarised, with particular emphasis on new theoretical models that are investigated in this thesis.

1.1 The Standard Model

The Standard Model (SM) is the theory that describes the fundamental components of the matter via fermionic fields and how they interact through the exchange of gauge boson, within the framework of quantum mechanics and special relativity. It includes 12 fermionic fields of spin-$\frac{1}{2}$, which are the constituents of matter and obey the Pauli exclusion principle, 4 vector boson fields of spin-1, which propagate the force fields, and an additional scalar boson field of spin-0, the recently discovered (light) Higgs boson, which is related to the mechanism that generates the masses for all fermions and bosons. A sketch of the SM constituents is shown in figure 1.1.1.

This model has been tested for many decades, and it has been able to predict and reproduce the experimental observations. Consequently, theoretical considerations and experimental data have led to the conclusion that the strong nuclear force, the weak nuclear force and the electromagnetic (EM) force are described by a renormalisable gauge-invariant quantum field theory (QFT) based on the local symmetry of the $\mathrm{SU(3)_C \times SU(2)_L \times U(1)_Y}$ group, with a partial breaking of the symmetry induced by the Brout-Englert-Higgs mechanism in the $\mathrm{SU(2)_L \times U(1)_Y}$ electroweak (EW) sector.

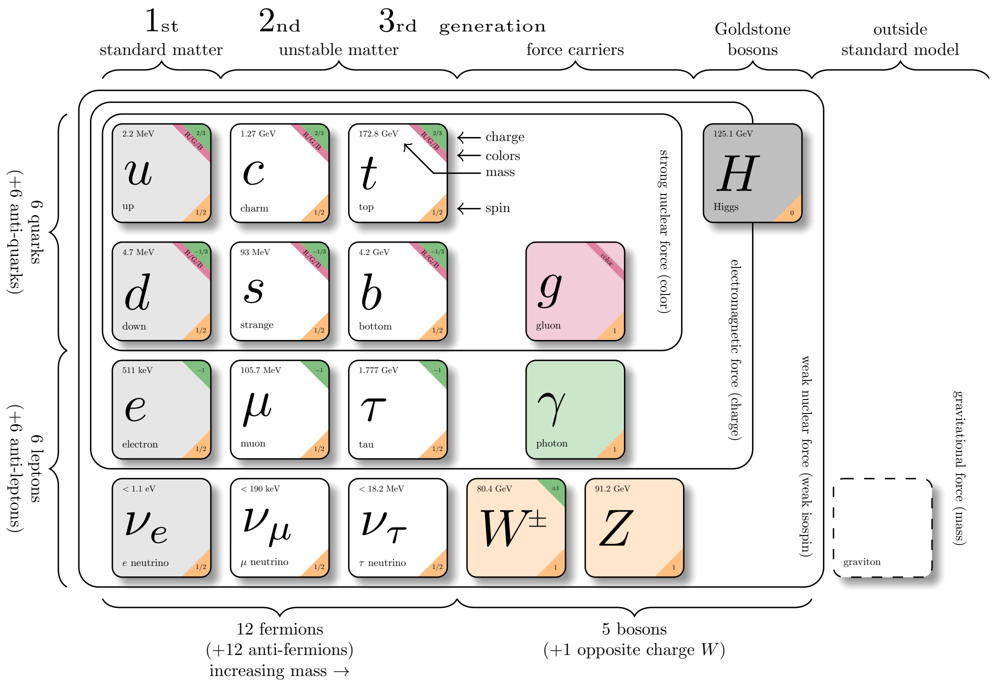


Figure 1.1.1: Field content of the SM, divided into interaction groups. Adapted from Ref. [6].

1.1.1 Fermions

Fermionic fields, or fermions, are classified into two types: quarks (q), which are a colour triplet (i.e. have a colour charge), carry EW charges and are subject to all SM interactions, and leptons, which are colourless but have EW charges. Charged leptons (ℓ) interact via all forces except the strong interaction, while the electromagnetically neutral leptons, called neutrinos (ν), are subject to only the weak force.

While leptons can exist as free particles, quarks seem not to do so. The explanation of this phenomenon is known as colour confinement, and it is a peculiarity of the strong force. Therefore, quarks cannot be found singularly but only in colour-singlet combinations, known as hadrons (see section 1.1.3).

Quarks and leptons are further grouped into 3 “families” or “generations” with equal quantum numbers but different masses. Further details on the coupling of each generation to the gauge bosons are provided in section 1.1.4. At present, there is no fundamental explanation for this subdivision.

A fermionic field ψ, which satisfies the Dirac equation, can be decomposed into its left-handed and right-handed components:

$$\psi = \psi_L + \psi_R \, . \tag{1.1}$$

The ψ_L and ψ_R parts identify the two irreducible representations of the restricted and orthochronous Lorentz group. It is worth mentioning that the Standard Model is a chiral theory, which is not invariant under parity transformations. Consequently, ψ_L and ψ_R behave differently under the gauge symmetries transformations.

In the following, quarks and leptons are represented in the following notation:

$$\psi_L^{\text{quarks}} = \begin{pmatrix} u \\ d \end{pmatrix}_L , \qquad \psi_L^{\text{leptons}} = \begin{pmatrix} \nu \\ l^- \end{pmatrix}_L ,$$
$$\psi_R^{\text{quarks}} = u_R, d_R , \qquad \psi_R^{\text{leptons}} = l_R^- ,$$

where all ψ_L are doublets and all ψ_R are singlets under $\mathrm{SU}(2)_\mathrm{L} \times \mathrm{U}(1)_\mathrm{Y}$. The former carry a third component of the weak isospin of $I_3 = +\frac{1}{2}$ and $I_3 = -\frac{1}{2}$ for the upper and lower row, respectively, while for the latter $I_3 = 0$. Moreover, the u-type quarks, carrying an electric charge of $Q = +\frac{2}{3}\,e$, are the *up* (u), *charm* (c) and *top* (t) quark. The d-type quarks, with $Q = -\frac{1}{3}\,e$, are the *down* (d), *strange* (s), and *bottom* (b) quark. Last, the charged leptons, with $Q = -e$ are the *electron* (e), *muon* (μ), and *tau* lepton (τ). The elementary positive charge e corresponds to $1.602\,176\,634 \times 10^{-19}$ C [7]. The *electron*-, *muon*-, and *tau*-neutrinos are electrically neutral and do not have a right-handed counterpart.

1.1.2 Gauge symmetries

A local gauge symmetry is an invariance under transformations that rotate the internal degrees of freedom with rotation angles that depend on the space-time point. The different types of fermions are distinguished by different quantum numbers, some of which correspond to a charge conserved under local transformations of the gauge invariance of the $\mathrm{SU}(3)_\mathrm{C} \times \mathrm{SU}(2)_\mathrm{L} \times \mathrm{U}(1)_\mathrm{Y}$ group: rotations in hypercharge space for $\mathrm{U}(1)_\mathrm{Y}$, in weak isospin space for $\mathrm{SU}(2)_\mathrm{L}$ and in colour space for $\mathrm{SU}(3)_\mathrm{C}$.

Before introducing the Lagrangian of the Standard Model, it is easier to start describing the gauge invariance of quantum electrodynamics (QED), the field theory describing EM interactions of electrically charged particles. The QED Lagrangian density is given by:

$$\mathcal{L}_{\text{QED}} = \overline{\psi}(i\gamma^\mu \mathcal{D}_\mu - m)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} , \tag{1.2}$$

where γ^μ are the Dirac matrices, $\mathcal{D}_\mu = \partial_\mu + iqA_\mu$ is the gauge covariant derivative, q is the coupling constant, that can be interpreted as the electric charge of the spinor field ψ, and m its mass, A_μ is the covariant four-potential of the EM field (photon) and $F_{\mu\nu}$ is the EM field tensor $F_{\mu\nu}(x) = \partial_\mu A_\nu(x) - \partial_\nu A_\mu(x)$. This Lagrangian density has three terms: a $\overline{\psi}\psi$ component that describes the kinematic of a free fermionic field, an A^2 part that describes the kinematics of a free photon, and a mixed $\overline{\psi}\gamma^\mu A_\mu \psi$ term that describes the interaction between the two fermion fields. Being an abelian U(1) symmetry, the QED does not include the self-interaction of the photon field; furthermore, a 4-point vertex (e.g. 2 fermions and 2 photons) does not exist due to renormalisability. Therefore, only a single elementary QED vertex exists, which is shown in figure 1.1.2.

The Lagrangian density in eq. (1.2) remains unchanged under the local transformation of the abelian unitary group $\mathrm{U}(1)_\mathrm{EM}$:

$$\begin{cases} \psi \longrightarrow \psi' = e^{i\mathrm{Q}\theta(x)}\psi \\ A_\mu \longrightarrow A'_\mu = A_\mu - \partial_\mu\theta(x) \end{cases} , \tag{1.3}$$

where Q is the electric charge operator of the $\mathrm{U}(1)_\mathrm{EM}$ group and $\theta(x)$ is the phase depending on the space-time coordinates. It is trivial to prove that a global transformation (i.e. not depending on x) of $\mathrm{U}(1)_\mathrm{EM}$ leads to the preservation of the electric

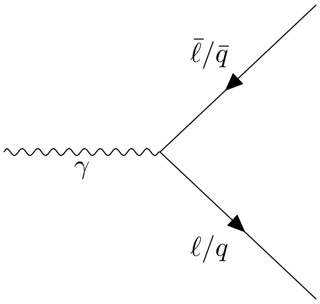


Figure 1.1.2: Elementary vertex of the EM interaction. Here γ represents the massless mediator of QED, the photon, while ℓ refers to charged leptons only. Adapted from Ref. [8].

charge. On the other hand, requiring invariance under a local $\mathrm{U(1)_{EM}}$ symmetry is essential to preserve the same description of nature everywhere in the universe. It is this requirement that fixes the transformation under $\mathrm{U(1)_{EM}}$ for $A_\mu(x)$, as reported in eq. (1.3).

It is to be noticed that a mass term ($m^2 A_\mu A^\mu$) for the A_μ field is not allowed by the gauge invariance, leaving the photon massless; this property has been confirmed several times by various experimental observations: the most recent limit on the photon mass is $m_\gamma < 10^{-18}\,\mathrm{eV}$ [7]. The addition of fermion mass terms of the form $m\overline{\psi}\psi$ into the Lagrangian is allowed only under $\mathrm{U(1)_{EM}}$ and forbidden otherwise, as these terms are not invariant under chiral gauge transformations. A more detailed theoretical explanation on the origin of the photon mass being null and the source of the mass of fermions in the complete SM theory is given in section 1.1.5.

The $\mathrm{U(1)_{EM}}$ is a particular case of the more general Lie group, but there are other cases with more complex symmetries, like the non-abelian compact Lie group $\mathrm{SU}(N)$. At this point, it is easy to extend the structure of the QED to the Yang-Mills (YM) gauge theory [9], with which it is possible to describe quantum chromodynamics (QCD), the theory of the strong force based on $\mathrm{SU(3)_C}$, as well as the unification of the EM and weak sectors (i.e. $\mathrm{SU(2)_L \times U(1)_Y}$). Then, it will be straightforward to apply these results to the SM Lagrangian.

Similarly to what was described before, it is possible to show that the YM Lagrangian remains unchanged under local $U(\alpha(x))$ transformations:

$$\begin{cases} \psi \longrightarrow \psi' = U(\alpha(x))\psi = e^{i\alpha^a(x)T^a}\psi \approx (1 + i\alpha^a(x)T^a + O(\alpha^2))\psi \\ A^a_\mu \longrightarrow A'^a_\mu = A^a_\mu - \dfrac{1}{\mathrm{g}}\partial_\mu\alpha^a(x) + f^{abc}A^b_\mu(x)\alpha^c(x) \end{cases} , \tag{1.4}$$

where T^a are the $N^2 - 1$ generators of the N-dimensional group satisfying the algebra $[T^a, T^b] = i\,f^{abc}\,T^c$ and f_{abc} represents the structure constant of the group of transformation. Starting from eq. (1.2) and extending to the case of the gauge vector fields $V_\mu(x)$, one obtains:

$$\mathcal{L}_{YM} = \overline{\psi}(i\gamma^\mu\mathcal{D}_\mu - m)\psi - \frac{1}{4}\sum_{a=1}^{N} F^a_{\mu\nu}F^{a\,\mu\nu}\,. \tag{1.5}$$

The gauge covariant derivative is now $\mathcal{D}_\mu = \partial_\mu - ig\sum_{a=1}^{N} V^a_\mu(x)T^a$, with g being the coupling constant and each component of the field tensor $F_{\mu\nu}$ is defined as:

$$F^a_{\mu\nu}(x) = \partial_\mu V^a_\nu(x) - \partial_\nu V^a_\mu(x) - g f^{abc} V^b_\mu(x) V^c_\nu(x)\,. \tag{1.6}$$

As already mentioned, a fermion mass term is forbidden since it does not respect the gauge invariance. Furthermore, as for the abelian case, the gauge bosons are massless. However, considering that the constant structure $f^{abc} \neq 0$, they carry a group charge and have self-interaction. This kind of symmetry is used to describe the strong and weak interaction, with $\mathrm{SU(3)_C}$ and $\mathrm{SU(2)_L}$, respectively.

1.1.3 QCD interactions

The QCD (quantum chromodynamics) theory describes the strong interaction between quarks and gluons. It is based on the more general YM theory with symmetry group $\mathrm{SU(3)_C}$. The Lagrangian density has as similar form as eq. (1.5), with the group's dimension being $N_C = 3$. and the coupling constant $\mathrm{g_s}$ is the only fundamental parameter of QCD.

It is usual to describe the quarks in the fundamental (F) representation of the group, as "colour-charged" fermions described by 3 degrees of freedom[1], namely red, green, and blue[2]:

$$\psi_\mathrm{q} = \begin{pmatrix} \psi_r \\ \psi_b \\ \psi_g \end{pmatrix}\,. \tag{1.7}$$

The colour group has 8 generators, hence 8 gluons, i.e. the fields mediating the strong interaction, which transform under the adjoint (A) representation of the symmetry group. Useful colour-algebra relations are the Casimir coefficients:

$$C_F = \sum T_a T_a = \frac{N_C^2 - 1}{2N_C} = \frac{4}{3} \qquad \text{and} \qquad C_A = \sum t_a t_a = N_C = 3\,, \tag{1.8}$$

where T_a and t_a are the generators of the fundamental and adjoint representations, respectively. The coefficients C_F and C_A are related to the gluon emission from a quark and a gluon, respectively (see Ref. [7]); this feature is used to distinguish their experimental signature, as described in section 3.3.2.

The interaction between quarks and gluons resembles the one for QED. Furthermore, since self-interaction is allowed by the symmetry group, both triple and quartic gauge couplings, which have no analogue in an abelian theory like QED, are present and of order $\mathrm{g_s}$ and $\mathrm{g_s}^2$, respectively. These elementary vertices are shown in figure 1.1.3.

The QCD theory has a simple structure but a very rich dynamic content. Among many other properties, it is worth mentioning the colour confinement, the asymptotic freedom and the lack of CP-violation. The latter is connected to new hypothetical particles, like axions, and it is discussed in more detail in section 1.3. A qualitative description of colour confinement and asymptotic freedom is reported in the following.

[1] The antiquarks have anticolour charges.

[2] Another peculiarity of $\mathrm{SU(3)_C}$ is that it is an exact symmetry; this means that quark masses and interactions do not change if one permutes the definition of colours.

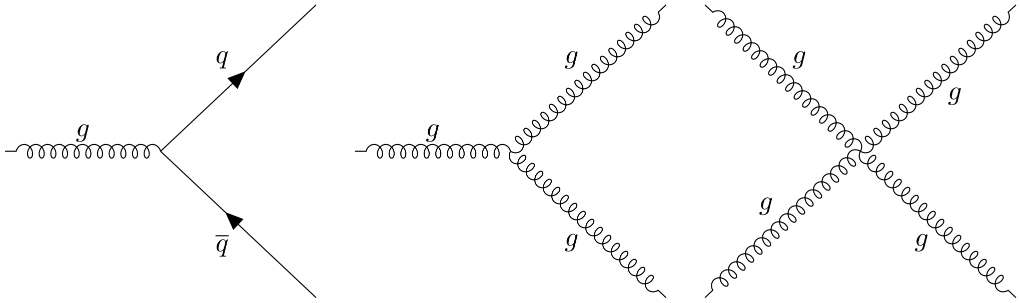


Figure 1.1.3: Elementary vertices of the QCD interaction. Adapted from Ref. [8].

A general property of QFT, and therefore any SM interaction, is that the coupling depends on the energy scale of the process considered; this behaviour is referred to as *running of the coupling.* Strongly related to the theory's renormalizability, this is a direct consequence of the ultraviolet (UV) divergence regularisation. Typically, one introduces a cut-off energy to eliminate divergent quantities that naturally arise from QFT, allowing the couplings to acquire a scale dependence by normalising them to a known (measured) value at a given scale. The origin of these divergences is often interpreted as the SM being a low-energy effective field theory of a more fundamental, yet unknown, theory (see section 1.3).

Qualitatively, the coupling's running can be seen as a contribution to the vacuum polarisation via loop corrections from fermions and gauge bosons; in fact, these interfering virtual particles cause a modification of the magnitude of the force. Due to their different symmetry groups, the electromagnetic, weak, and strong interactions behave differently. For example, the EM coupling is screened by a cloud of virtual electron-positron pairs, which gives, at a long distance (or small energy), the well-known value of the fine structure constant $\alpha^0_{EM} = \frac{e^2}{4\pi} \sim 1/137$ [10], corresponding to the electron mass scale. At higher energies, e.g. the Z mass scale at which the unification of the EW theory happens, the coupling value has been measured to be $\sim 1/129$ [11]. On the contrary, the gluon self-interactions in QCD lead to a cloud of virtual gluons, which, together with the quark contribution, makes the coupling strength $\alpha_s = \frac{\mathrm{g_s}^2}{4\pi}$ increase at large distances, or, equivalently, at low energies [12, 13].

The increase of the QCD coupling strength at large distances directly affects particle collisions, for example, in the hadronisation process. In fact, when a quark-antiquark pair is produced, increasingly high energy is required to separate the two quarks, which makes the creation of a new quark-antiquark pair energetically favourable. Eventually, quarks organise themselves in colourless bound states (hadrons) of either a quark-antiquark pair (mesons) or three quarks (baryons): this is a property of all non-abelian theories [14] and is referred to as “colour confinement”. Consequently, quarks and gluons produced in highly energetic particle collisions cannot be observed freely but form collimated sprays of particles, referred to as jets (see section 3.3).

Low-energy processes, like the hadronisation, cannot be calculated analytically but must be modelled phenomenologically, as at the GeV scale the QCD coupling cannot be treated perturbatively ($\alpha_s(m_\tau) \sim 0.3$). At higher energies, α_s decreases (e.g. $\alpha_s(100\,\mathrm{GeV}) \sim 0.118$ [15]), and allows the treatment of quarks as quasi-free particles, enabling perturbative calculations. This property is known as “asymptotic freedom”, and it is also a peculiarity of all the non-abelian gauge theories [16, 17].

1.1.4 Electroweak interactions

The theory originating from the combined symmetry group $\mathrm{SU(2)_L \times U(1)_Y}$, whose generators are the operator of weak isospin $\vec{I}$ and the operator of hypercharge Y, respectively, is known as the Weinberg-Salam electroweak theory[3] [19, 20]. It is derived from the unification of QED and weak force, with the aim of describing the two forces as different manifestations of the same interaction. In this case, the left-handed and right-handed components of the fermions transform under infinitesimal local gauges transformations as:

$$\begin{cases} \qquad SU(2)_L \\ \psi_L \longrightarrow \psi'_L = (1 + i\sum_{i=1}^{3} \alpha^a(x) I^a)\psi_L \\ \psi_R \longrightarrow \psi'_R = \psi_R \end{cases} \quad \begin{cases} \qquad U(1)_Y \\ \psi_L \longrightarrow \psi'_L = (1 + i\beta(x)Y)\psi_L \\ \psi_R \longrightarrow \psi'_R = (1 + i\beta(x)Y)\psi_R \end{cases} , \qquad (1.9)$$

where $\alpha(x)$ and $\beta(x)$ represent local gauge transformations, $I^i = \frac{\sigma^i}{2}$ for L-fields, and σ^i are the Pauli matrices. In the particular case of the SM, i.e. only including the observed particles and a single Higgs doublet, I^i is null for all R-fields, given that, for all known fermions, ψ_R is a singlet. In a more general theory, this term does not vanish for a non-singlet right-handed fermion.

The gauge invariance of this theory corresponds to having one gauge boson for $\mathrm{U(1)_Y}$ and three gauge bosons for $\mathrm{SU(2)_L}$, which leads to the following expression of the EW Lagrangian density:

$$\mathcal{L}_{EW} = i\overline{\psi_L}\gamma^\mu \mathcal{D}_\mu \psi_L + i\overline{\psi_R}\gamma^\mu \mathcal{D}_\mu \psi_R - \frac{1}{4}\sum_{a=1}^{3} W^a_{\mu\nu} W^{a\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} , \qquad (1.10)$$

where the left and right components of the spinors are split to emphasise the different transformations. The covariant derivative is defined as:

$$\mathcal{D}_\mu = \left(\partial_\mu + i\frac{\mathrm{g}'}{2} Y B_\mu + i\mathrm{g} \sum_{a=1}^{3} W^a_\mu I^a \right) . \qquad (1.11)$$

The tensor fields $B_{\mu\nu}$ and $W^a_{\mu\nu}$ have the same form of the EM and YM counterparts, respectively. In the case at hand, f^{abc} corresponds to the three-dimensional totally antisymmetric Levi-Civita symbol ε^{abc}. Furthermore, g' and g are the coupling constants introduced for $\mathrm{U(1)_Y}$ and $\mathrm{SU(2)_L}$ gauge groups, and their relative fields B_μ and W^a_μ respect the following infinitesimal transformation rules:

$$\begin{cases} \qquad SU(2)_L \\ W^a_\mu \longrightarrow W'^a_\mu = W^a_\mu + \partial_\mu \alpha^a(x) + \mathrm{g}\varepsilon^{abc}\alpha^b(x) W^c_\mu \\ B_\mu \longrightarrow B'_\mu = B_\mu \end{cases} \quad \begin{cases} \qquad U(1)_Y \\ W^a_\mu \longrightarrow W'^a_\mu = W^a_\mu \\ B_\mu \longrightarrow B'_\mu = B_\mu + \partial_\mu \beta(x) \end{cases} . \qquad (1.12)$$

[3] Earlier work on a similar model had been carried out by S. Glashow [18].

Couplings to fermions

All the interactions between gauge bosons and fermions can be derived from the equations above. Also, one can notice that by rotating W^1_μ and W^2_μ, it is possible to create eigenstates of the third component of the weak isospin operator I^3:

$$W^\pm_\mu = \frac{W^1_\mu \pm iW^2_\mu}{\sqrt{2}}, \qquad I^3 W^\pm_\mu = \pm W^\pm_\mu \,. \tag{1.13}$$

It is easy to associate the $W^\pm$ with the two charged W boson fields. In fact, a part of the EW Lagrangian can be written as:

$$W^1_\mu I^1 + W^2_\mu I^2 = \mathrm{W}^+_\mu I^- + W^-_\mu I^+, \qquad \text{with} \quad I^\pm = \frac{I^1 \pm iI^2}{\sqrt{2}}, \tag{1.14}$$

where $I^\pm$ are the raising and lowering operators of the isospin charge. As a consequence, the only possible combinations of couplings between charged vector bosons and fermions have the form:

$$\overline{d_L}\, W^+_\mu \, u_L \,, \qquad \overline{u_L}\, W^-_\mu \, d_L . \tag{1.15}$$

Here, u_L and d_L play the role of generic left-handed “up” or “down” states in the isospin space. These kinds of couplings are known as the charged current (CC) [21, 22] couplings and are responsible for flavour mixing. Moreover, these interactions change isospin by $\Delta I^3 = \pm 1$ and keep Y unchanged; as a consequence, the change in fermion charge is by $\Delta Q^3 = \pm 1$, as clarified later by eq. (1.19).

Furthermore, the fields W^3_μ and B_μ are mixed and neither of them couples exclusively to the EM charge. With a rotation of these two fields, it is possible to obtain two electrically neutral bosons:

$$\begin{pmatrix} A_\mu \\ Z_\mu \end{pmatrix} = \begin{pmatrix} \cos\theta_\mathrm{W} & \sin\theta_\mathrm{W} \\ -\sin\theta_\mathrm{W} & \cos\theta_\mathrm{W} \end{pmatrix} \begin{pmatrix} B_\mu \\ W^3_\mu \end{pmatrix} , \qquad \cos\theta_\mathrm{W} = \frac{\mathrm{g}}{\sqrt{\mathrm{g}^2 + \mathrm{g}'^2}} \,, \tag{1.16}$$

where θ_W defines the Weinberg electroweak mixing angle. To retrieve the EM interaction, the photon is required to couple to left- and right-handed fermions with a strength proportional to the electric charge:

$$e = \mathrm{g} \sin\theta_\mathrm{W} = \mathrm{g}' \cos\theta_\mathrm{W} \,. \tag{1.17}$$

Also in this case, a part of the EW Lagrangian can be rearranged as:

$$\mathrm{g} I^3 W^3_\mu + \frac{\mathrm{g}'}{2} Y B_\mu = eQA_\mu + \frac{e}{\sin\theta_\mathrm{W}\cos\theta_\mathrm{W}} \left(I^3 - Q\sin^2\theta_\mathrm{W} \right) Z_\mu \,, \tag{1.18}$$

where the electric charge operator Q is defined as:

$$Q = I^3 + \frac{1}{2} Y \,. \tag{1.19}$$

In the case of neutral gauge bosons, the coupling occurs only between fermions of the same type and family. Furthermore, while the CC involves only left-handed particles, the neutral current (NC) [23, 24] coupling is present also for right-handed fermions, as operator Q acts equally on both parts.

	Particle			spin	SU(3)$_\mathrm{C}$ dimension	SU(2)$_\mathrm{L}$ I	I^3	U(1)$_\mathrm{Y}$ Y	U(1)$_\mathrm{EM}$ Q
Leptons	$\nu_{\mathrm{e}L}$	$\nu_{\mu L}$	$\nu_{\tau L}$	$\frac{1}{2}$	1	$\frac{1}{2}$	$+\frac{1}{2}$	-1	0
	e_L	μ_L	τ_L				$-\frac{1}{2}$		-1
	e_R	μ_R	τ_R			0	0	-2	-1
Quarks	u_L	c_L	t_L		3	$\frac{1}{2}$	$+\frac{1}{2}$	$-\frac{1}{3}$	$+\frac{2}{3}$
	d_L	s_L	b_L				$-\frac{1}{2}$		$-\frac{1}{3}$
	u_R	c_R	t_R			0	0	$+\frac{4}{3}$	$+\frac{2}{3}$
	d_R	s_R	b_R					$-\frac{2}{3}$	$-\frac{1}{3}$
Gauge Bosons	g			1	8	0	0	0	0
	γ				1	not def.	0	0	0
	Z								
	W$^+$					1	$+1$	0	$+1$
	W$^-$						-1		-1
	H			0	1	$\frac{1}{2}$	$+\frac{1}{2}$	$+1$	0

Table 1.1.1: Quantum numbers for SM particles. Values taken from Ref. [7].

A summary of the quantum numbers derived so far is reported in table 1.1.1 for all SM particles. All quantities can be derived from the SM Lagrangian: for example, the values for Y are derived imposing the U(1) conservation in the Yukawa couplings (see section 1.1.5).

The branching fractions for the decays of W and Z bosons can be determined under the assumption of lepton flavour universality (LFU), i.e. the EW coupling of the gauge bosons to lepton families is independent of the lepton flavour. This property has been tested in several measurements, e.g. decays of tau leptons, light mesons, as well as in the Z boson's partial decay widths [25]. Any experimental evidence for LFU violation would be a clear sign of physics beyond the SM; an overview of the most recent results will be provided at the end of this chapter. The generic formula of the partial width for a W decay is:

$$\Gamma(\mathrm{W} \to \mathrm{f}\overline{\mathrm{f}'}) = N_f \cdot \frac{\mathrm{g}^2 m_\mathrm{W}}{48\pi} \,, \tag{1.20}$$

where N_f is a fermion-dependent factor. It is equal to 1 for leptons, while for quarks it is $N_f = N_C \cdot |V_{\mathrm{f}\overline{\mathrm{f}}}|^2 \cdot [1 + \alpha_s/\pi + \dots]$, where N_C is the number of colour flavours and the other factors account for the CKM elements (see section 1.1.5) and one-loop QCD corrections, respectively. Using the unitarity of the CKM matrix, it is straightforward to derive the leptonic branching ratio (BR):

$$\mathrm{BR}(\mathrm{W} \to \ell\nu_\ell) = \frac{1}{2 \cdot 3 \cdot [1 + \alpha_s/\pi] + 3} \sim 10.8\% \,, \tag{1.21}$$

which is in very good agreement with the experimental value (the average of the three leptonic modes) [7, 25]. As a consequence, the hadronic branching ratio consists of the remaining ~ 67%; it is dominated by the CKM-favored $\mathrm{u}\overline{\mathrm{d}}$ and $\mathrm{c}\overline{\mathrm{s}}$ final states (~ 31% each). In particular, the analysis presented in this thesis will make use of the results derived above (see chapter 6).

The formula above holds true for all quarks except the top quark, since the decay W → tb is kinematically forbidden. Additionally, the top quark is heavy enough that it decays before hadronising like the other quarks: its decay into a real bW pair is by far the most dominant decay channel (see section 1.1.5).

As already described, the Z boson decays into a fermion and its antiparticle, with a coupling strength that depends on $I^3 - Q\sin^2\theta_\mathrm{W}$. Since the third component of the weak isospin I^3 is different for left- and right-handed fermions, the coupling is different as well. The partial width of the decay of a Z boson to fermions, excluding the top quark for kinematic reasons, is given by:

$$\Gamma(\mathrm{Z} \to \mathrm{f\bar{f}}) = N_C \cdot \frac{\mathrm{g}^2 m_\mathrm{Z}}{192\pi\cos^2\theta_\mathrm{W}}\left[1 + \left(1 - 4|Q_f|\sin^2\theta_\mathrm{W}\right)^2\right] . \tag{1.22}$$

Table 1.1.2 reports the experimental values of the Z boson decay widths and BRs, which are compatible with the theoretical ones [26]. The charged and neutral leptonic decays of the Z boson are used as final states in this thesis, as reported in chapter 6.

	$\mathrm{Z} \to \ell^+\ell^-$	$\mathrm{Z} \to \mathrm{inv}$	$\mathrm{Z} \to \mathrm{q\bar{q}}$	total
$\Gamma_\mathrm{exp.}$ [MeV]	83.985(86)	499.0(15)	1744.4(20)	2495.2(23)
$\mathrm{BR}_\mathrm{exp.}$ [%]	3.3658(23)	20.000(55)	69.911(56)	-

Table 1.1.2: Experimental values for the Z boson decay widths. The decay to charged leptons is averaged over the three flavours, while the hadronic one is inclusive in quark flavours. Values taken from Ref. [7].

The measured value of the decay width of the Z boson to invisible particles leads to the determination of the number of neutrinos. The result of the world average value is $N_\nu = 2.9841 \pm 0.0083$ [27], and well compatible with the three known neutrino flavours. This was the first experimental proof, with important consequences also in astrophysics and cosmology, that there exist only the three generations of light neutrinos ($m_\nu < m_\mathrm{Z}/2$).

Self-coupling

The gauge boson self-interactions can be derived from eq. (1.6). Defined by the symmetry group itself, the only possible combinations of bosons for the triple and quartic gauge couplings are shown in figure 1.1.4 with the following coupling strengths:

$$\begin{aligned} \mathrm{g}_{\mathrm{W^+W^-}\gamma} &= \mathrm{g}\sin^2\theta_\mathrm{W} = e\,, & \mathrm{g}_{\mathrm{W^+W^-Z}} &= \mathrm{g}\cos^2\theta_\mathrm{W}\,, \\ \mathrm{g}_{\mathrm{W^+W^-}\gamma\gamma} &= -e^2\,, & \mathrm{g}_{\mathrm{W^+W^-W^+W^-}} &= \mathrm{g}^2\,, \\ \mathrm{g}_{\mathrm{W^+W^-}\gamma\mathrm{Z}} &= e\mathrm{g}\cos^2\theta_\mathrm{W}\,, & \mathrm{g}_{\mathrm{W^+W^-ZZ}} &= -\mathrm{g}^2\cos^2\theta_\mathrm{W}\,. \end{aligned} \tag{1.23}$$

The triple gauge vertices have been already tested at the LEP2 [28] and at the Tevatron [29]. The quartic coupling, being quadratic in g and hence small, could be directly measured only recently by the ATLAS and CMS collaborations [30], whose data were used to set new limits on anomalous triple and quartic gauge couplings not present in the SM.

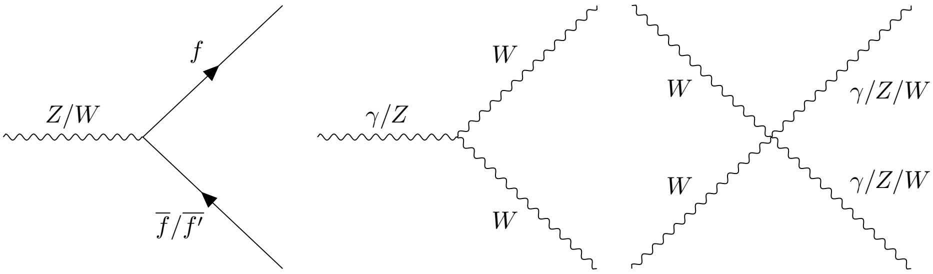


Figure 1.1.4: Elementary vertices of the EW interactions, including the self-interaction of gauge bosons. The symbol f ($\overline{f}$) represents any (anti)fermion. Adapted from Ref. [8].

1.1.5 The Higgs boson mechanism

To explain the presence of the experimentally measured non-zero masses of the fermions and gauge bosons, the SM introduces a single colourless $\mathrm{SU(2)_L}$-doublet scalar field, the Higgs (H).

$$\mathrm{H} \equiv \phi(x) = \begin{pmatrix} \phi^+ \\ \phi_0 \end{pmatrix} . \tag{1.24}$$

This all-pervasive field causes the spontaneous breaking of the $\mathrm{SU(2)_L \times U(1)_Y}$ gauge symmetry through the Brout-Englert-Higgs (BEH) mechanism[4] [34], providing mass to the particles while preserving the $\mathrm{U(1)_{EM}}$ invariance.

In order to describe the main idea of symmetry breaking, the Lagrangian for a complex scalar field ϕ with a quartic potential[5] is considered:

$$\mathcal{L}_\mathrm{H} = \frac{1}{2} \left(\mathcal{D}_\mu \phi\right)^\dagger \left(\mathcal{D}_\mu \phi\right) - V(\phi) = \frac{1}{2} \left(\mathcal{D}_\mu \phi\right)^\dagger \left(\mathcal{D}_\mu \phi\right) + \frac{1}{2}\mu^2\phi^2 - \frac{1}{4}\lambda\phi^4 , \tag{1.25}$$

where $\mathcal{D}_\mu$ is the same as in eq. (1.10). The potential $V(\phi)$ has different shapes depending on the sign of the parameters μ^2 and λ:

- λ must be positive to have bound states inside the potential as $\phi \to \infty$;
- if $\mu^2 > 0$, the minimum of the potential is at $\phi = 0$. In this case, the electroweak symmetry is unbroken in the vacuum because a gauge transformation acting on the ground state does not change it;
- if $\mu^2 < 0$, the minimum of the potential, the vacuum expectation value (VEV), is located on a spherical surface in four dimensions of radius $v = \sqrt{\frac{\mu^2}{\lambda}}$. In this case, the vacuum is not invariant under gauge transformations, and the symmetry is spontaneously broken.

Figure 1.1.5 shows the shape of the potential for $\lambda > 0$ and $\mu^2 < 0$. Expanding around the VEV with an opportune gauge fixing, the field ϕ becomes:

$$\phi(x) = \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix} . \tag{1.26}$$

[4] Although they were the first discussing a new massive gauge boson, to reach the complete renormalisable $\mathrm{SU(2)_L \times U(1)_Y}$ symmetry breaking theory some more work was needed. A more general name is, therefore, ABEGHHK'tH mechanism, for Anderson, Brout, Englert, Guralnik, Hagen, Higgs, Kibble, and 't Hooft, who extended the initial BEH mechanism [31–33].

[5] Higher powers of ϕ would lead to a not renormalisable theory [35].

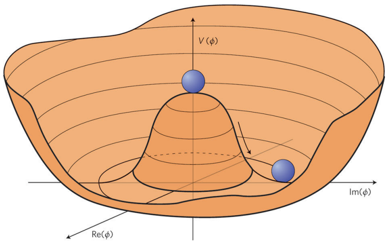


Figure 1.1.5: Simplified sketch of the quartic-potential shape that allows for spontaneous symmetry breaking. Taken from Ref. [36].

Substituting it into the Lagrangian and using the relations in equations (1.13) and (1.16), one obtains[6]

$$\begin{aligned}\mathcal{L}_{\mathrm{H}} = &\left[\frac{1}{2}\left(\partial_\mu h\right)^2 - \lambda v^2 h^2 + \lambda v h^3 + \frac{\lambda}{4} h^4\right] + \left[\frac{\mathrm{g}^2 v^2}{4}\mathrm{W}^+\mathrm{W}^- + \frac{\mathrm{g}^2 v^2}{8\cos\theta_{\mathrm{W}}}\mathrm{ZZ}\right] + \\ &+ \left[\frac{\mathrm{g}^2 v}{2} h\mathrm{W}^+\mathrm{W}^- + \frac{\mathrm{g}^2 v}{4\cos\theta_{\mathrm{W}}} h\mathrm{ZZ}\right] + \left[\frac{\mathrm{g}^2}{4} hh\mathrm{W}^+\mathrm{W}^- + \frac{\mathrm{g}^2}{8\cos\theta_{\mathrm{W}}} hh\mathrm{ZZ}\right] .\end{aligned} \tag{1.27}$$

It arises naturally that the H scalar field acquires mass via self-interaction. It can be seen that, when they interact with the vacuum, also the W^+, W^- and Z fields obtain a mass term, while the boson field associated to the photon remains massless. It means that despite the Higgs field breaks all the $\mathrm{SU(2)_L} \times \mathrm{U(1)_Y}$ symmetries, it maintains the $\mathrm{U(1)_{EM}}$ symmetry, leaving the vacuum electrically neutral[7]. The masses of the bosons depend on free parameters that are not predicted by the SM:

$$m_{\mathrm{H}} = \sqrt{2\lambda v^2}\,, \qquad m_{\mathrm{W}} = \frac{gv}{2}\,, \qquad m_{\mathrm{Z}} = \frac{gv}{2\cos\theta_W}\,. \tag{1.28}$$

Nevertheless, it is possible to determine the value of v, which appears in the Higgs field scaling definition, by using the experimental value of the Fermi constant G_F, determined by the decay of the muon [7], in the following way:

$$\frac{\mathrm{g}}{8m_W^2} = \frac{G_F}{\sqrt{2}} \qquad \Longrightarrow \qquad v = \frac{1}{\sqrt{G_F\sqrt{2}}} \sim 246\,\mathrm{GeV}\,. \tag{1.29}$$

Only with the relatively recent observation of the Higgs boson by the ATLAS [37] and CMS [38] collaborations, and the measurement of its mass ($m_{\mathrm{H}} \sim 125\,\mathrm{GeV}$), it was possible to give the first estimation of the other free parameter: $\lambda \sim 0.129$.

In eq. (1.27), the other terms correspond to the couplings with a massive gauge boson (hVV and $hhVV$) and the triple and quartic self-couplings (h^3 and h^4). These couplings are uniquely predicted by the SM once the boson masses and v are known. Furthermore, given the value of λ, it is possible to treat the Higgs boson self-coupling in a perturbative way. A sketch of the interaction vertices is shown in figure 1.1.6.

It is possible to use the Higgs doublet to also generate the masses of quarks and leptons, adding to the Lagrangian a Yukawa term that respects the gauge transformations. Rather easily for the lepton case, it is possible to obtain:

$$\mathcal{L}_Y^\ell = -g_\ell\left(\overline{\psi}_L\phi\psi_R + \overline{\psi}_R\phi^\dagger\psi_L\right) = -m_\ell\overline{\psi}\psi - \frac{m_\ell}{v}\overline{\psi}\psi H\,, \tag{1.30}$$

[6] For the sake of simplicity, the 4-dim indices are omitted when redundant.

[7] It is possible to reach the same conclusions with another gauge fixing (Goldstone theorem).

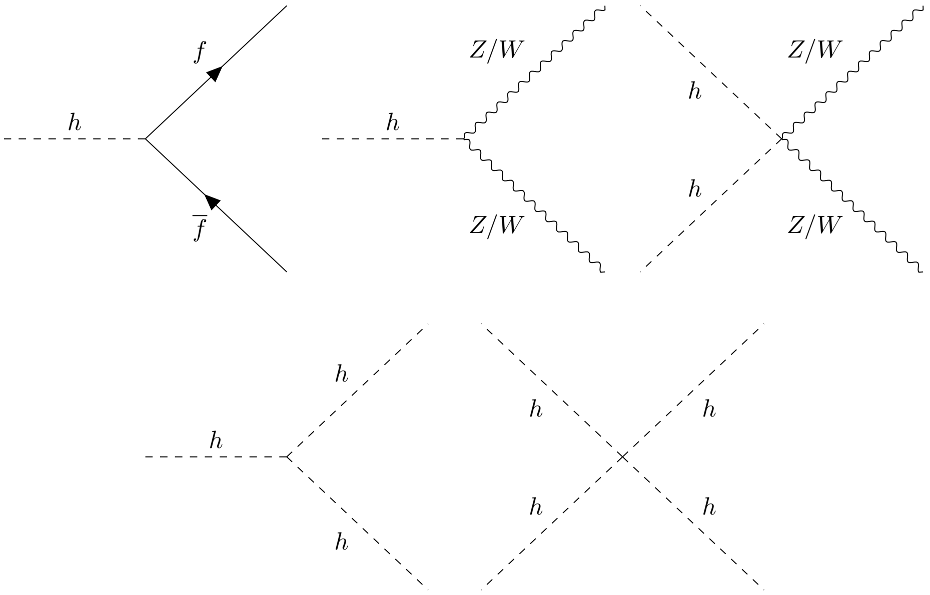


Figure 1.1.6: Elementary vertices of the Higgs boson coupling with fermions and vector bosons (upper row) and Higgs boson triple and quartic self-interactions (lower row), derived from the Lagrangian in equations (1.25), (1.30) and (1.31). Adapted from Ref. [8].

where g_ℓ is the coupling constant for leptons and $m_\ell = g_\ell \frac{v}{2}$. Both the mass term and the Higgs boson coupling with the fermion are present. Also in this case, the coupling parameters are arbitrary, and the masses have to be measured.

Theoretical hints and experimental proofs have led to the conclusion that the quarks as free particles are mass eigenstates, while they appear as $\mathrm{SU}(2)_\mathrm{L}$ eigenstates in the electroweak interaction: in other words, the latter is a mixture of the former and vice-versa. A possible explanation for the origin of this mixture can be attributed to some unknown hidden symmetry at higher energies [39].

The Yukawa coupling in eq. (1.30) can be generalised for the case of up and down fermions, using the Higgs doublet and its conjugate. Considering quarks, the corresponding term is:

$$\mathcal{L}_Y^{\mathrm{q}} = -\sum_{i=0}^{3}\sum_{j=0}^{3}\left(g^u_{ij}\overline{u}^i_L\phi^\dagger u^j_R + g^d_{ij}\overline{d}^i_L\phi d^j_R\right) , \tag{1.31}$$

where the sum runs over the lepton families and g^u_{ij} (g^d_{ij}) is the Yukawa coupling matrix for the $u(d)$-type quark in the isospin space. With an opportune rotation, it is possible to create mass eigenstates:

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}_{L,R} = U \begin{pmatrix} \mathrm{u} \\ \mathrm{c} \\ \mathrm{t} \end{pmatrix}_{L,R} , \qquad \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix}_{L,R} = D \begin{pmatrix} \mathrm{d} \\ \mathrm{s} \\ \mathrm{b} \end{pmatrix}_{L,R} , \tag{1.32}$$

such that the rotation matrices U and D are diagonalising the coupling matrices:

$$M^u \equiv U^{-1} g^u U = \begin{pmatrix} m_\mathrm{u} & 0 & 0 \\ 0 & m_\mathrm{c} & 0 \\ 0 & 0 & m_\mathrm{t} \end{pmatrix} \quad , \qquad M^d \equiv D^{-1} g^d D = \begin{pmatrix} m_\mathrm{d} & 0 & 0 \\ 0 & m_\mathrm{s} & 0 \\ 0 & 0 & m_\mathrm{b} \end{pmatrix} . \tag{1.33}$$

Since the NC and the interactions with gluons are diagonal in the quark fields, these terms remain diagonal also in a new basis. The only SM interactions that act between u- and d-type are the CC. For this case, it is worth to rewrite the notation in eq. (1.15) as:

$$\overline{\mathrm{d}}'_L \, W^+_\mu \, \mathrm{u}_L \quad , \qquad \overline{\mathrm{u}}_L \, W^+_\mu \, \mathrm{d}'_L \, , \tag{1.34}$$

where, this time, $\overline{\mathrm{u}}_L$ (d_L) is a generic up(down)-type quark in the mass basis, $\mathrm{d}'_L = V_\mathrm{CKM} \mathrm{d}_L$, and $V_\mathrm{CKM} = U^\dagger D$ is the Cabibbo–Kobayashi–Maskawa (CKM) unitary matrix. While the existence of this mixing is related to many physical examples (e.g. charge-parity (CP) violation, Glashow-Iliopoulos-Maiani (GIM) mechanism, flavour changing neutral currents (FCNC)), an analogue of this matrix for the leptons is not present in the SM, since no right-handed neutrinos are included. Only the observation of the neutrino oscillations [40] provided an experimental reason for a description of neutrinos as both flavour and mass eigenstates via the so-called Pontecorvo-Maki-Nakagawa-Sakata (PMNS) matrix, which emerges naturally as a consequence of the seesaw mechanism [41].

The CKM matrix properties are summarised briefly in the following. The CKM matrix is fully described by three mixing angles and one CP-violating phase. The magnitude of each component, obtained from the combination of the latest experimental results [7], is reported in the following:

$$V_\mathrm{CKM} = \begin{bmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{bmatrix} \approx \begin{bmatrix} 0.9737 & 0.225 & 0.004 \\ 0.221 & 0.987 & 0.041 \\ 0.008 & 0.039 & 1.013 \end{bmatrix} . \tag{1.35}$$

As the transition probability is proportional to $|V_{ij}|^2$, the transitions between the up and down types are favourite to be within the same generation, but other contributions, especially between the first and the second generations, are not negligible. Finally, the unitarity requirements is expressed as $\sum_k V_{ik} V^*_{jk} = 0$.

1.2 Higgs boson phenomenology at hadron colliders

From equations (1.25) and (1.30), it is possible to predict the partial width for each of the Higgs boson decays and the cross section (σ) for different Higgs boson production mechanisms in particle collisions. Both the BR and σ are of vital importance because they allow testing the hypothesis that the discovered boson is the SM Higgs boson. This section mostly focuses on proton-proton (pp) collisions at the Large Hadron Collider (LHC), and only briefly summarises the relevant details for Higgs boson studies at electron-positron colliders (e.g. at the proposed Future Circular Collider-ee (FCC-ee), International Linear Collider (ILC), Compact Linear Collider (CLIC) or Circular Electron Positron Collider (CEPC)).

1.2.1 Higgs boson decays

The Higgs boson couplings are proportional to the masses squared of the decay products, with a preference for heavier particles that are kinematically accessible.

As for any unstable particle, the branching ratios of the Higgs boson decays are determined by the partial widths of the decays into each final state (χ):

$$\mathrm{BR}(H \to \chi) = \frac{\Gamma(H \to \chi)}{\Gamma_H}\,, \tag{1.36}$$

where $\Gamma_H = \sum_\chi \Gamma(H \to \chi) = 4.1\,\mathrm{MeV}$ is predicted from the SM under the assumption of $m_\mathrm{H} = 125\,\mathrm{GeV}$ [42].

The Higgs boson decays into pairs of fermions through Yukawa-like interactions and the decay width at leading order is:

$$\Gamma(H \to ff) = \frac{N_C}{8\pi}\frac{m_f^2}{v^2} m_\mathrm{H}\left(1 - \frac{4m_f^2}{m_\mathrm{H}^2}\right)^{3/2}\,. \tag{1.37}$$

This expression is proportional to the square of the Yukawa coupling and linear in the Higgs boson mass. Given the measured Higgs boson mass, decays to $\mathrm{t\bar{t}}$ are negligible, and the most important fermionic final states are $\mathrm{b\bar{b}}$, $\tau\bar{\tau}$ and $\mathrm{c\bar{c}}$. In the case of decays to quarks, QCD corrections, known to the astonishing next-to-next-to-next-to-next-to-leading order (N4LO), are needed since the loop contributions for emission or exchange of a gluon in the final state are quite significant and reduce the partial width: this is why the decay width into a τ pair is larger by more than a factor of 2 with respect to the $\mathrm{c\bar{c}}$ channel, despite the colour factor and the similar mass.

The decay rate for a Higgs boson into a vector boson pair with at least one virtual boson (as $m_\mathrm{H} < 2m_\mathrm{V}$) is quite tedious to derive and a simple formula is not derivable[8]. Instead, one can calculate the rate under the assumption of a Higgs boson mass above the threshold production and derive some conclusions from it:

$$\Gamma(H \to VV) = \frac{m_\mathrm{V}^2}{32\pi v^2}\frac{m_\mathrm{H}^2}{m_\mathrm{V}^2} m_\mathrm{H}\delta_V\sqrt{1-4x}\left(1 - 4x + 12x^2\right)\,, \quad \text{with} \ \ x = \frac{m_V^2}{m_\mathrm{H}^2}\,, \tag{1.38}$$

where the subscript V can be either W or Z and $\delta_Z = 1, \delta_W = 2$. As in the fermionic case, the expression above contains the squared coupling term ($\frac{m_\mathrm{V}}{v}$), and the m_H proportionality. A polarisation term ($\frac{m_\mathrm{H}}{m_\mathrm{V}}$) is also present.

One can see that, due to its form, the rate becomes very large and dominates, even over the $\mathrm{t\bar{t}}$ term if m_H allows it. This relation is important in the case of beyond-the-Standard-Model (BSM) models with heavier Higgs boson partners. Below the diboson (VV) production threshold, the decays of the SM Higgs boson into virtual V bosons is also important, as even decays into two off-shell gauge bosons contribute. The branching ratios as a function of m_H are shown in figure 1.2.1.

The Higgs boson can directly interact only with massive particles, so that the decays $\mathrm{H} \to gg$, $\mathrm{H} \to \gamma\gamma$, and $\mathrm{H} \to \mathrm{Z}\gamma$ are absent at the tree level. These decay rates are generated by quantum loops, and the dominant contributions to the decay amplitude are given by massive particles, the top quark and W boson for the decays

[8] The formula can be found in Ref. [43] and the numerical calculation for $\mathrm{H} \to \mathrm{WW}^*$ in Ref. [44].

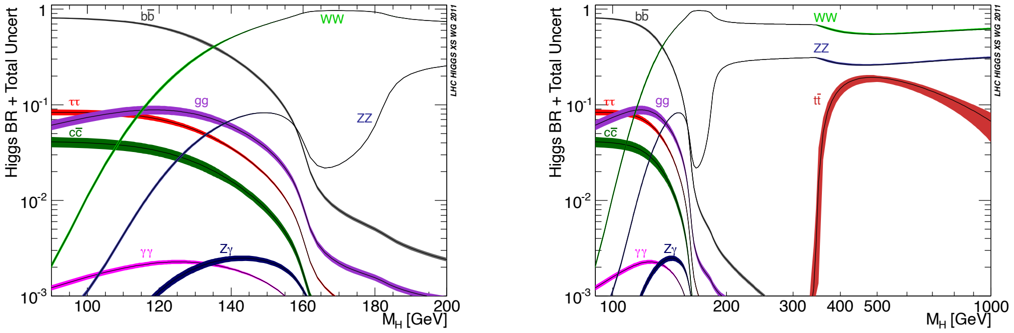


Figure 1.2.1: Higgs boson branching ratios and their uncertainties for the low mass range (left) and for the extended mass range (right). Adapted from Ref. [45].

into gluons and gauge bosons, respectively. The loop-induced Higgs boson decays are rare but experimentally important because of the photon's clean signature, and their rates are given by:

$$\Gamma(H \to GG) = \frac{\alpha_G^2 m_{\mathrm{H}}^3}{256\pi^3 v^2} I_G\,, \tag{1.39}$$

where G can be either a gluon, a γ, or a Z boson. I_G is a factor depending on m_{t} and m_{W} and α_G is α_s for the gluon case and α_{EM}^0 otherwise. The measurement of these couplings ruled out the possibility of having a heavier fourth generation of quarks since this would have a dominant contribution to these couplings.

The branching ratios for the dominant decay modes are listed in Table 1.2.1.

Decay mode	BR [%]
$\mathrm{b\bar{b}}$	58.1
$\tau^+\tau^-$	6.3
$\mathrm{c\bar{c}}$	2.9
$\mathrm{s\bar{s}}$	0.03
$\mu^+\mu^-$	0.02

Decay mode	BR [%]
WW*	21.5
gg	8.2
ZZ*	2.6
$\gamma\gamma$	0.2
Zγ	0.01

Table 1.2.1: Predicted Higgs boson decay BRs assuming $m_{\mathrm{H}} = 125\,\mathrm{GeV}$ [42].

1.2.2 Higgs boson production

In pp collisions at the centre-of-mass energy currently reached by the LHC (up to 13 TeV), the Higgs boson is expected to be produced mainly through four mechanisms: gluon-gluon fusion, vector boson fusion, vector boson and top associated productions. A brief description of the mechanisms is summarised in the following.

gluon-gluon fusion (ggF) This is the dominant production mode at the LHC. Due to the gluon being massless, there is no direct coupling between the gluons and the Higgs boson; the leading diagram involves a triangle quark loop: the dominant contribution to the SM amplitude arises from the top quark loops, supplemented by a smaller contribution of bottom quark loops. The NNLO QCD corrections increase the total cross section by about a factor of two to the LO prediction. This production mechanism is potentially sensitive to contributions from hypothetical new massive particles with non-zero colour charge.

vector boson fusion (VBF) The VBF process has a cross section of about a tenth of the ggF one. The leading diagrams involve $q\overline{q}$ scattering with a vector boson exchange and the emission of a real Higgs boson. Since incoming quarks tend to be scattered by a small angle, the momentum exchange is typically low, and the channel is experimentally characterised by two very energetic jets pointing close to the beamline in opposite hemispheres of the detector. These jets are referred to as "forward jets", and this high-rapidity topology allows for a sufficient background rejection in this production mode. This process is theoretically interesting because it allows the study of the Higgs boson couplings to vector bosons.

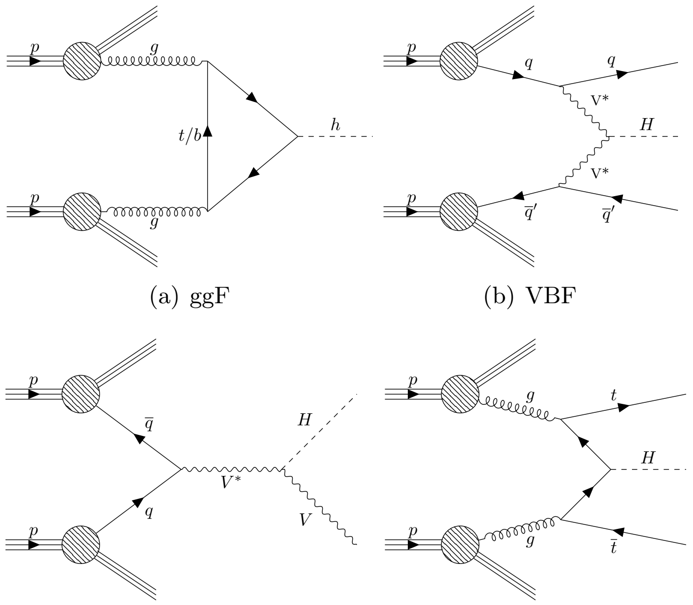
Higgs-Strahlung (VH) Also known as associated production, this process occurs when a virtual vector boson (V) decays to its on-shell state, radiating a Higgs boson. Both the W and Z bosons contribute to this process, and their combined cross section is about 60% of the VBF cross section. Vector boson leptonic decays provide a useful handle for background rejection in a hadronic environment.

top associated production ($t\overline{t}H$) A challenging but important process is the $t\overline{t}H$ associated production, in which the Higgs boson is mostly produced from the fusion of a $t\overline{t}$ pair or through radiation from a top quark. Despite the small production rate, this process plays an essential role in probing the top-Higgs Yukawa coupling via direct measurements [46, 47]. A relatively similar process is the $b\overline{b}H$ associated production, which is currently not object of direct searches; the Yukawa coupling with the bottom quarks was probed recently [48, 49] using the $b\overline{b}$ decay mode of the Higgs boson. A complementary process, but even more challenging experimentally, is the single-top associated production (tH). This process receives contributions from both the top-Higgs and W-Higgs couplings. Due to destructive interference between these two diagrams, the resulting cross section is very small; however, it provides an interesting additional test to the SM on the relative sign of this interference.

The hierarchy of the production cross sections is listed in table 1.2.2, and the leading-order Feynman diagrams are shown in figure 1.2.2.

	σ [pb]	σ/σ_{tot} [%]
ggF	48.58	87.3%
VBF	3.78	6.8%
WH	1.33	2.4%
ZH	0.88	1.6%
$t\overline{t}H$	0.50	0.9%
$b\overline{b}H$	0.49	0.9%
tH	0.08	0.1%

Table 1.2.2: Higgs boson production cross section at 13 TeV for each mode assuming $m_H = 125$ GeV, at NNLO+NNLL QCD accuracy [42]. The relative contribution of each production mechanism to the total cross section is reported in the third column.



(a) ggF

(b) VBF

(c) VH

(d) ttH

Figure 1.2.2: Examples of LO diagrams for different Higgs boson production modes.

Higgs boson production in e^+e^- colliders

The Higgs boson production rates change drastically in an e^+e^- collision, such as at the Large Electron-Positron Collider (LEP), or at the proposed International Linear Collider (ILC), Future Circular Collider-ee (FCC-ee), Compact Linear Collider (CLIC) or Circular Electron Positron Collider (CEPC). In this kind of environment, the most important Higgs boson production processes would be VH and VBF. The relative importance changes as a function of the centre-of-mass energy, as it is visible in figure 1.2.3 (left). In addition to direct and indirect BSM searches and precision top quark physics, these colliders provide rich potential for Higgs boson physics, such as the direct measurement of the Yukawa couplings, the Higgs boson self-coupling and the differential cross section with an expected precision at the percent-level. Compared to hadron colliders, like the LHC, these machines cannot provide a large dataset and at the same time reach very high energies. The expected luminosity for several e^+e^- machines is reported in figure 1.2.3 (right). Instead, the precision would come from a cleaner environment due to the lack of the omnipresent hadronic background and from the precision with which the collision energy is known.

1.2.3 Higgs boson measurements at the LHC

Due to the unpredicted value of its mass, an experimental confirmation of the Higgs boson's existence was eagerly needed to verify the last piece of the SM. Theoretical considerations were quite vague in constraining in which mass range to perform the search (from vanishing values up to several hundreds of GeV), making its discovery more elusive. From the experimental side, direct searches at LEP excluded masses up to 114 GeV [54], whereas the combined data from the CDF and D0 experiments at Tevatron was indicating some excess (global significance of 2.5σ) in the $[115 - 140]$ GeV range with a peak at 120 GeV [55].

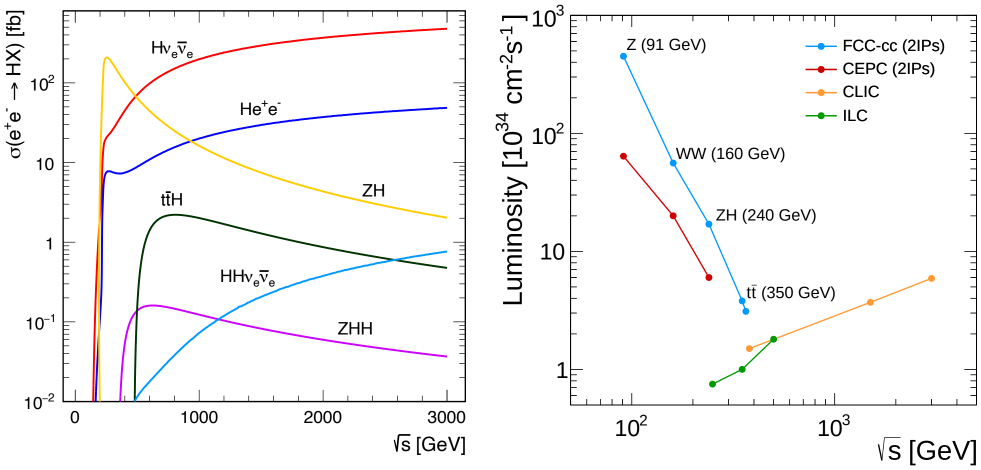


Figure 1.2.3: Left: Cross section as a function of $\sqrt{s}$ for several Higgs boson production processes at an e^+e^- collider. Taken from Ref. [50]. Right: Luminosity forecast for different e^+e^- colliders as a function of centre-of-mass energy. Adapted from Ref. [50–53].

The first experimental observation of the long-sought Higgs boson dates back to the LHC era. In 2012, both the ATLAS [37] and CMS [38] collaborations announced the discovery of a new Higgs-like particle that, with a mass of approximately 125 GeV, completed the today well-known Higgs boson phenomenology. Extensive studies have been made to characterise the newly discovered particle, whose properties are consistent with the SM Higgs boson according to the current experimental data.

The LHC Run 1 (cf. section 2.1) analyses focused on the discovery and the measurement of the basic properties. For example, the $\mathrm{H} \to \mathrm{ZZ}^* \to 4\ell$ and $\mathrm{H} \to \gamma\gamma$ channels allowed the measurement of the mass with excellent resolution ($\sim 1\,\mathrm{GeV}$), despite the small decay rates. The spin-parity properties have been studied exploiting the $\mathrm{H} \to \gamma\gamma$, $\mathrm{H} \to \mathrm{ZZ}^* \to 4\ell$, and the $\mathrm{H} \to \mathrm{WW}^* \to \ell\nu\ell\nu$ modes. The decay into a photon pair excluded the spin-1 hypothesis (Landau-Yang theorem). The observed data disfavoured the spin-2 hypotheses and, assuming that the boson has zero spin, was proven to be consistent with the pure scalar hypothesis, $\mathrm{J}^{\mathrm{P}} = 0^+$, as predicted by the SM, while rejecting the pure pseudoscalar hypothesis.

Furthermore, a combination of ATLAS and CMS measurements of the Higgs boson production and decay rates was performed at the end of Run 1, using five production processes and the six decay modes [56]. Under the assumption of a global signal strength (μ) that affects all processes and channels, the comparison with the SM predictions results in a best-fit value of:

$$\mu = 1.09^{+0.11}_{-0.10} = 1.09^{+0.07}_{-0.07}(\mathrm{stat.})^{+0.04}_{-0.04}(\mathrm{exp.})^{+0.08}_{-0.07}(\mathrm{theo.})\,. \tag{1.40}$$

The LHC analyses do not allow disentangle production and decay modes unambiguously; therefore, several parametrisations of these variables were used to extract their values. The different approaches were able to provide a direct comparison to the SM predictions, as well as a model-independent analysis, where the ratios with a reference σ or BR have the advantage of being free of theory uncertainties, and a BSM interpretation of the results to scan for possible deviations from the expectations. The results of these three approaches are shown in figure 1.2.4 and are compatible with the SM expectation.

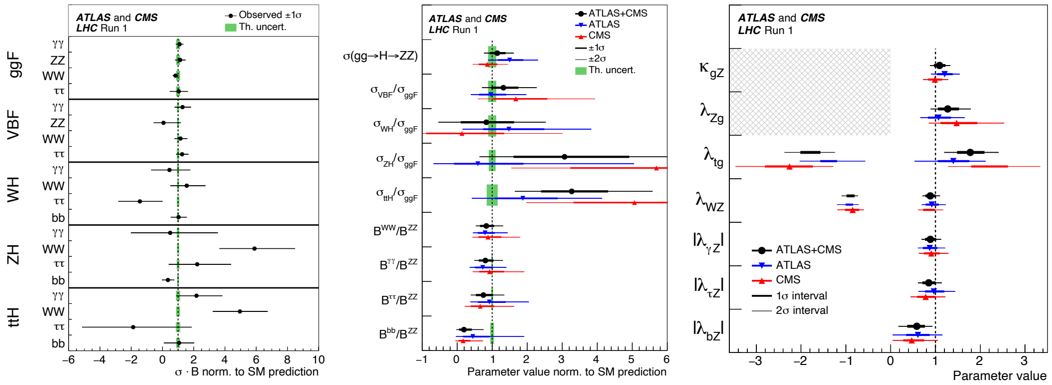


Figure 1.2.4: Left: Best fit values of $\sigma \cdot \mathrm{BR}$ for different channels. Center: Best fit values of the ratios of cross sections and branching ratios. Right: Best fit values of coupling modifiers for BSM intererpretation. Taken from Ref. [56].

Nonetheless, only the observation of decays into vector bosons had large enough significance to be announced as discovery modes. In contrast, none of the fermionic decays reached the 5σ discovery threshold during the Run 1 period: for example, the ATLAS and CMS experiments reported evidence for the $\tau\tau$ channel with an observed significance of 3.2 and 4.5 σ, respectively, and it was possible to reach a significance 2.6 σ for the bb decay in the VH mode only with their combined data.

During the Run 2 of LHC, a larger amount of data was collected and higher production cross sections were expected due to the increased centre-of-mass energy (e.g., the $\mathrm{t\bar{t}H}$ production rate at $\sqrt{s} = 13\,\mathrm{TeV}$ for Run 2 is 3.9 times larger than at $\sqrt{s} = 8\,\mathrm{TeV}$ for Run 1). This allowed each experiment to observe independently the $\mathrm{H} \to \tau^+\tau^-$ channel [57, 58], confirming the Yukawa coupling to charged leptons, and the $\mathrm{H} \to \mathrm{b\bar{b}}$ channel [48, 49], measuring the coupling to down-type quarks. Finally, the $\mathrm{t\bar{t}H}$ production mode was used to probe the coupling to up-type quarks [46, 47].

The coupling to the first and second generation of fermions is more complex and not yet observed. Despite the clean signature of the final states, the Higgs boson decay to a muon pair has an extremely small BR, and an even smaller value is predicted for the electron case. The measurement of the coupling to the light-flavour quarks (u, d and s) is experimentally challenging because of the enormous amount of irreducible QCD background; easier to distinguish and with higher BR is the $\mathrm{c\bar{c}}$ decay mode, although discriminate it from the $\mathrm{b\bar{b}}$ final state further reduces the selection efficiency, as highlighted in section 3.3.2. Consequently, it has only been possible to set upper limits on coupling strengths for the lighter generations, as they are not yet accessible for direct observation [59–62].

The data collected so far allowed for precision measurements of the parameters of the Higgs boson sector. Although the Higgs boson self-coupling is still beyond reach, measurements of differential production cross sections are already accessible [63–66]. Additionally, due to their extremely low rates, double and triple Higgs boson productions will be measurable only at the high luminosity era at LHC (HL-LHC). Current projections [67] foresee reaching only the 4σ significance with standard analysis techniques; therefore, new approaches, most likely based on machine learning (ML), are needed to increase the sensitivity, and new ways to constrain the systematic uncertainties of these analyses are required as well. An example of the possible improvements obtained using ML techniques for jet tagging is given in chapter 4.

1.3 Shortcomings of the SM

As it was shown in the previous chapters, the SM is a simple, powerful theory and, at the same time, it has very remarkable predictive power. Nonetheless, there are indications that it is neither complete nor final. Its most remarkable problems fall into two main categories: observational puzzles and structural limitations.

Observational puzzles

These are experimentally motived problems that cannot be explained with the current knowledge of the SM. Some examples are discussed below.

Matter-antimatter asymmetry An incontrovertible piece of evidence for the existence of physics beyond-the-Standard-Model is the excess of matter over antimatter in the universe. This imbalance implies a different behaviour between particles and antiparticles, which cannot be explained with the current Standard Model of particle physics or the Cosmological Model of inflation. Even with an initial asymmetry at the time of the Big Bang, the current prediction is extremely small compared to what is observed; this is a strong suggestion that the current theory is not complete.

Origin of neutrino masses The neutrino flavour oscillation experiments from astrophysical and atmospheric sources provide evidence that at least two out of three neutrinos have non-vanishing masses. Even introducing right-handed neutrinos, the extremely small values for the Yukawa coupling, together with the fact that these couplings are several orders of magnitude apart from the fermion ones, pose a puzzling question. Another option would be to consider Majorana neutrinos, i.e. the neutrino is its own antiparticle. This has been shown to lead to lepton flavour and unitary violation at high energies [68].

Presence of dark matter There is empirical evidence (rotation curves of galaxies [69], weak lensing measurements [70], microwave background experiments [71]) that the universe is full of an unknown type of matter, referred to as dark matter. This state of matter is neutral with respect to the electromagnetic and strong forces, but it is expected to be massive since it interacts with gravity.

Flavour anomalies A possible violation of lepton flavour universality (LFU) is generally referred to as flavour anomaly. The LFU is based on the assumption that the gauge couplings to leptons are flavour-independent. On the one hand, all LEP measurements confirmed this hypothesis [25]. On the other hand, recent measurements from the BaBar, LHCb and Belle experiments show tensions at the level of 2-3 sigma with the SM expectations [72–74].

Structural limitations

Based on theoretical considerations, these structural problems seem to point out that the SM lacks robustness at higher energies. As already pointed out, the SM could be a low-energy manifestation of a more fundamental, yet unknown, theory. A brief description of these issues, often associated to a fine-tuning of parameters, is reported in the following.

Origin of generations and mass hierarchy There is a pattern between the three generations hidden in the Yukawa coupling. Many theories try to derive an explanation from first principles, but there is currently no satisfactory solution for this question.

Lack of strong CP-violation No CP-violating process in the strong interaction is observed to date. This poses another structural problem in the SM, as it creates an unmotivated asymmetry among the gauge interactions. An additional U(1) symmetry can be introduced to account for this absence: this fine-tuning is known as Peccei-Quinn theory and linked to the potential existence of axions. The only necessary parameter must be very close to zero to explain the current CP symmetry. From the absence of an electric dipole moment of the neutron, it is possible to constrain this term's magnitude to be $< 10^{-10}$ [75].

High energy description of gravity A successful theory of gravity is already developed in the General Relativity framework, but a fully comprehended QFT version is still absent. Even though quantum computations of gravity are performed as an effective field theory, the violation of unitarity at the Plank scale ($\Lambda_P \sim 10^{19}$ GeV) is the major problem to make the unification of the SM and the gravitational force possible.

Hierarchy of fundamental scales Another long-standing structural problem, which can be seen from several points of view, is the vast difference (10^{16}) in magnitude between the gravitational force and the electroweak scales. The origin of this dissimilarity cannot be explained to date, and also, the unification of such far-apart scales is a challenging task for new theories.

1.3.1 BSM theories

Numerous BSM theories with variations and generalisations of the SM have been formulated in the attempt to resolve all, or part of, the tensions with experimental data mentioned above and be conceptually more satisfying from a theoretical point of view. These models often involve the introduction of new particles that, depending on the theory and the problem it addresses, can be bosonic or fermionic and whose masses vary within an extensive range. This section outlines some of the current theoretical ideas for BSM physics, giving priority to the multi- TeV scale theories.

One of the oldest and most appealing of these extensions is based on grand unified theories (GUTs), a set of theories described by a gauge group bigger than that of the SM. They are capable of affecting higher energy scales ($\Lambda_{GUT} \sim 10^{16}$ GeV) and influencing the cosmological models, and, at the same time, giving predictions at the collider energy scales [76]. The baseline of these theories is that it is possible to unify all the fundamental forces in a similar way as the electroweak interaction unifies the weak and the electromagnetic forces. A hypothetical theory that tries to include gravity is sometimes called a "theory of everything".

More centred around the unification of the spin-2 mediator of gravity (graviton) and the spin-1 gauge bosons within a unique algebra, there is the model based on the fermion-boson symmetry or supersymmetry (SUSY). By doing so, this theory can explain most of the shortcomings of the SM (e.g. hierarchy, the origin of masses, divergences, dark matter). It postulates the existence of supersymmetric partners for each SM particle. A quantity introduced by the theory is called R-parity and is defined as $R = -1^{2S+3B-L}$, where S, B and L are the spin and the baryon and lepton numbers, respectively. The conservation or violation of this quantity are both allowed, but the stringent experimental results on the absence of lepton and baryon number violation tip the balance towards R-parity conservation models, for which the lightest supersymmetric particle is supposed to be stable and a possible dark matter candidate.

Another group of theories that try to combine general relativity with gauge symmetries is known as string theory. The key feature is to redefine the core of the SM as QFT by including the existence of several (usually unobservable) extra-dimensions and adding vibrating filaments (strings) and membranes (branes) of energy as fundamental constituents of the universe. Among the plethora of possible models, some of the most commonly tested are the Kaluza-Klein (KK) and the Randall-Sundrum (RS) ones. The former aims at unifying the gravitational and the EM interactions, while the latter attempts to explain the relative weakness of gravity.

All the theories above assume coupling of the newly predicted particles with the SM Higgs boson or even the existence of an extended Higgs boson sector. Therefore, the Higgs boson itself is a unique tool to probe a large phase space in BSM searches, also thanks to the diversity of its final states; in this regard, the not yet fully explored $\mathrm{H} \rightarrow \mathrm{c\bar{c}}$ decay is investigated in detail in chapter 6.

Another common feature is a high-dimensional symmetry group, and many models also include a symmetry-breaking mechanism for their unification groups. When this happens, extra U(1) gauge symmetries appear naturally, and they are usually associated with massive charged and neutral vector particles, referred to as W', Z' and γ'. The expected ranges for their masses and couplings are very model-dependent, even though most of them assume electroweak scale couplings and masses around the TeV scale. In some scenarios, the assumption on the mass is arbitrary, leading to models with either vanishing or extremely high values for their masses.

The analysis performed in this thesis, and presented in chapter 6, is based on the search for a hypothetical spin-1 massive resonance decaying into a Z and a Higgs boson. A detailed overview of the theoretical framework and the experimental results in the context of heavy new resonances decaying into a pair of SM bosons is provided in the following section.

1.4 Diboson resonances

As detailed above, there is a multitude of theories that extend the SM, resulting in different phenomenological predictions, which can be tested by experimental searches at the LHC. In this work, emphasis is put on new phenomena that lead to the resonant production of new particles coupling to the gauge boson and Higgs boson sectors. The models that predict such resonances and the state-of-the-art of the LHC searches are discussed in this section.

1.4.1 HVT model

From an experimental perspective, searches for the new predicted particles are typically not sensitive to all the free parameters of the underlying model but mainly to those that affect mass, production, and decay rates. As a consequence, it is common to deploy simplified descriptions, in which the new particle is described as a resonance, whose peak shape is modelled well by a Breit-Wigner (BW) function.

The results of this work will be interpreted in one of these simplified models, namely the heavy-vector triplet (HVT) model [77]. It provides a simple but well-motivated example of electroweak-charged spin-1 resonances arising from different theories, such as weakly-coupled [78] or Composite Higgs [79, 80] models.

The newly introduced electroweak sector shows a phenomenology analogous to the SM vector bosons but with larger expected resonance masses. Also, both the charged (W′) and neutral (Z′) states are predicted to be degenerate in mass and to have comparable production rates. Two free parameters are introduced to describe the coupling to the SM Higgs and gauge bosons (c_{H}), and to fermions (c_F)[9]. These parameters are chosen to be dimensionless coefficients to parametrise the relative contribution to the typical interaction strength (g_V). In fact, the range of the g_V coupling can vary in different scenarios, from $\mathcal{O}(1)$ up to $\mathcal{O}(10)$ in weakly or strongly coupled models, respectively. The phenomenology of the model is entirely described, to a good approximation, in terms of the couplings and the mass m_V of the resonance. An example of the simplified formula of the decay widths is reported in equations (1.41) and (1.42), where g is the SM $\mathrm{SU(2)_L}$ gauge coupling.

$$\Gamma_{\mathrm{W'}\to ff} \simeq 2\,\Gamma_{\mathrm{Z'}\to ff} \simeq \frac{N_C \mathrm{g}^4 c_F^2}{g_V^2}\frac{m_V}{48\pi}\,, \tag{1.41}$$

$$\Gamma_{\mathrm{Z'}\to \mathrm{W^+W^-}} \simeq \Gamma_{\mathrm{Z'}\to \mathrm{ZH}} \simeq \Gamma_{\mathrm{W'}\to \mathrm{W^\pm Z}} \simeq \Gamma_{\mathrm{W'}\to \mathrm{W^\pm H}} \simeq \frac{g_V^2 c_H^2 m_V}{192\pi}\,. \tag{1.42}$$

It is a standard approach to consider a few benchmark models, inspired by weakly or strongly coupled extensions of the SM. One of the following scenarios is usually chosen:

- Model A: the coupling to SM bosons and fermions is of similar strength.
- Model B: the coupling to fermions is suppressed by several orders of magnitude.
- Model C: the coupling to fermions is forbidden. As a consequence, the only production mode is VBF.

For each model, the c_H and c_F parameters are fixed to specific values, while g_V and m_V are free and can be compared to experimental results. The chosen values for the HVT parameters in the different scenarios are reported in table 1.4.1. In particular, the analysis presented in this thesis will use signal models generated with fixed values of g_V (see table 1.4.1) and m_V (cf. chapter 6). The total widths and the BRs for these benchmarks are shown in figure 1.4.1.

Parameter	Model A	Model B	Model C
c_H	-0.556	-0.976	1
c_F	-1.316	1.024	0
g_V	1	3	1

Table 1.4.1: Chosen values for the HVT parameters to emulate different BSM models.

[9] The universality of lepton and quark couplings is assumed. Also, other free parameters are postulated to account for self- and quartic-couplings that are irrelevant for the LHC phenomenology.

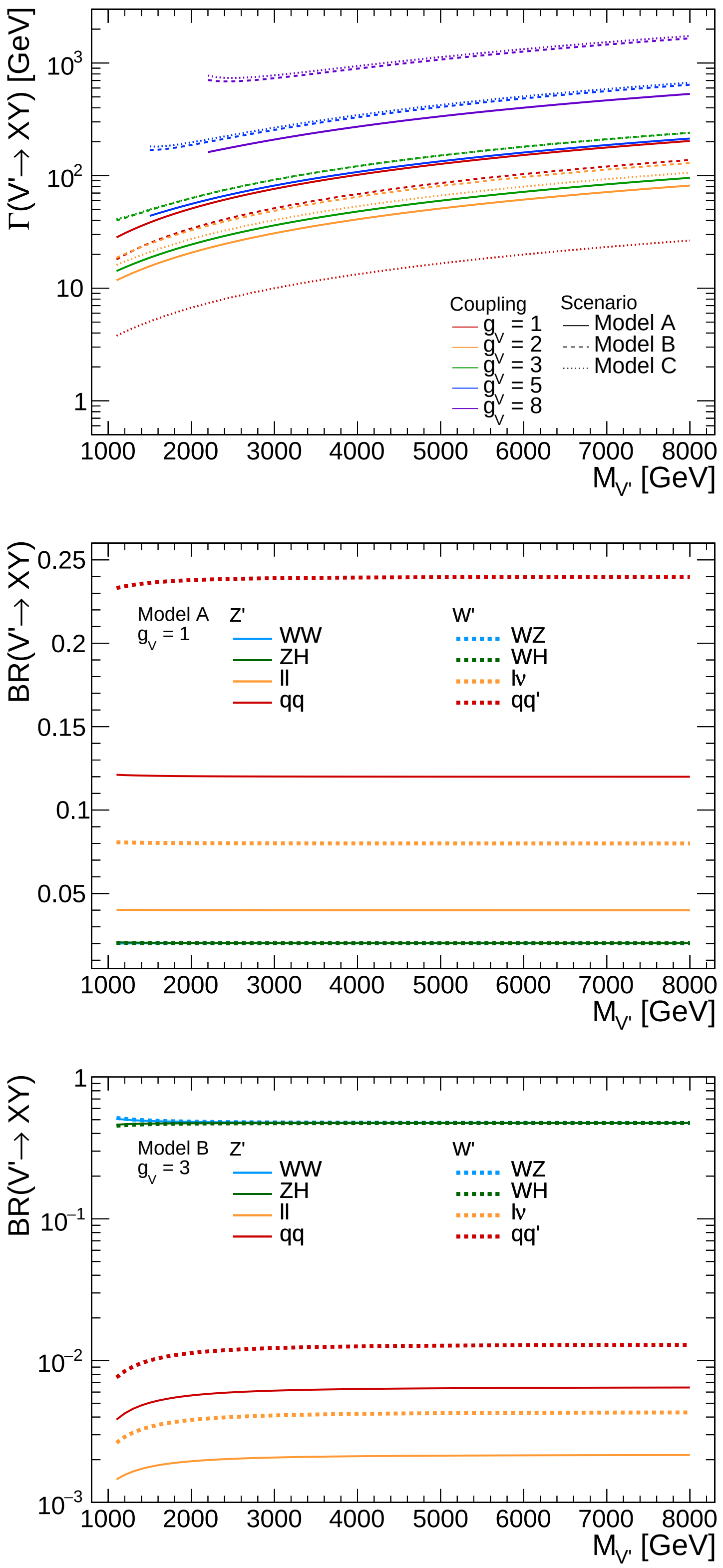


Figure 1.4.1: Total widths and branching ratios as a function of the resonance mass for different HVT benchmarks models. Calculated with [81].

1.4.2 Warped extra dimensions

Another class of theories that predicts diboson resonances is based on the existence of warped extra dimensions (WED) [82, 83]. To solve the hierarchy problem, such models propose a new higher-dimensional mechanism to connect the Planck and electroweak scales through an exponential hierarchy.

The extra dimensions are supposed to be compactified between two 4-dimensional boundaries, commonly called *branes.* In the simplest case of one spatial extra dimension [82], the following 5-dimensional metric is introduced:

$$ds^2 = e^{-2kr_c\phi}\eta_{\mu\nu}dx_\mu dx_\nu + r_c^2 d\phi^2 \,, \tag{1.43}$$

where $0 \leq \phi \leq \pi$ is the coordinate of the extra dimension, r_c its size and k its curvature. The brane where the extra dimension is localised ($\phi = 0$) is known as the Planck-brane. The other, where the typical SM energies are localised ($\phi = \pi$), is known as TeV-brane. The region between the branes is called *bulk*, and is controlled through an exponential "warp" factor ($e^{-kr_c\pi}$): $kr_c \approx 11$ is sufficient to explain the scale difference between the two branes. This factor generates two effective scales: on the one hand, the energy scales in the 4-dimensional space are the manifestation of their relative 5-dimensional counterparts through the warp factor; on the other hand, the 4-dimensional Planck scale barely depends on the wrap factor. This is a direct consequence of gravity being the only field that can propagate in the extra dimension, while the SM fields are confined to the TeV brane[10]. As a consequence, the hierarchy problem can be addressed by exploiting an additional dimension.

Perturbations of the space-time result in new spin-2 physical states (KK decomposition). The zero-mode of such oscillations corresponds to the massless mediator of gravity, the graviton, while the first massive excitation is the KK-Graviton. Similarly, fluctuations of the extra dimension produce the massive scalar Radion field and its related KK-states.

Depending on the scenario, these excitations are localised in the TeV-brane [82] or are allowed in the bulk as well [83]. In the former case, the KK states couple preferably to light quarks and gluons, while in the latter scenario, they couple preferentially with third-generation quarks and the Higgs and gauge bosons. Due to the spin of the new resonance, no ZH coupling is allowed; therefore, no comparison with the results of this thesis is possible. The production cross section of the KK-Graviton and the Radion in pp collisions for different $\sqrt{s}$ is shown in figure 1.4.2.

1.4.3 LHC results

As discussed in the previous section, many well-motivated models predict diboson resonances. Several searches for such new particles have been performed using data collected with the CMS and ATLAS experiments. Focus is given to the CMS analyses, although similar results are achieved by the ATLAS Collaboration.

The variety of Higgs and vector bosons final states makes it possible to adopt a multitude of techniques to investigate the different experimental signatures. These analyses are compared to several theoretical models and usually combined with similar searches to enhance the sensitivity.

[10] Other extensions [83] also allow the SM fields to propagate in the bulk to tackle the different scale of the fermion masses.

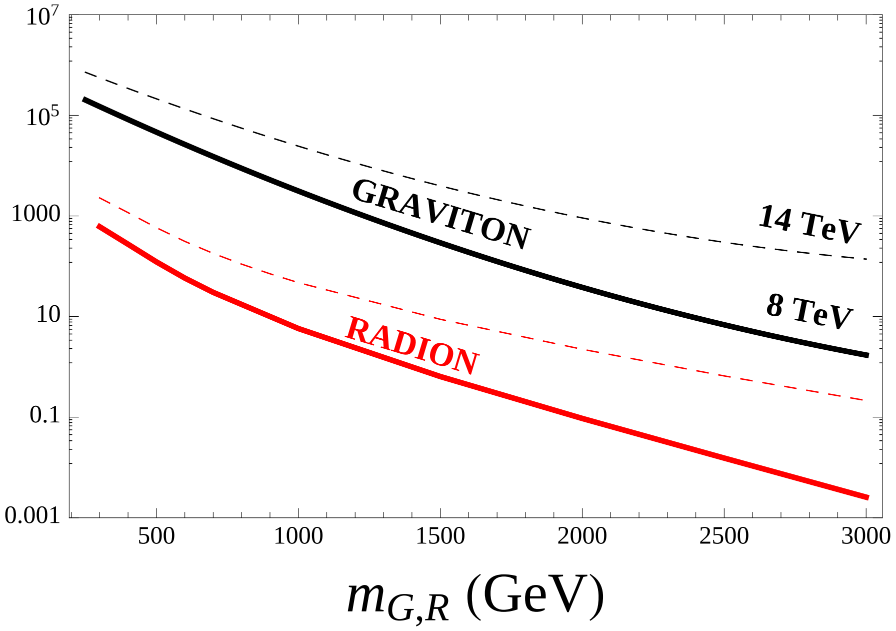


Figure 1.4.2: Production cross section in fb as a function of the KK-Graviton and Radion masses corresponding to 8 (solid) and 14 (dashed) TeV. Here the dependence on the interaction scale Γ and the coupling to gluons c is factored out. Taken from Ref. [84].

In this section, the latest published results, grouped by experimental signatures, are discussed, highlighting the experimental techniques used.

Searches in the $\ell\ell$+jet final states include the ZV [85] and ZH [86] channels, where the Higgs or vector bosons decay hadronically and the Z boson decays into a pair of oppositely charged leptons. Despite the relatively low BR, ranging from 4% to 10% depending on the channel under consideration, these searches take advantage of the clean signature provided by the presence of leptons, which results in a heavy suppression of the QCD multijet background. The presence of leptons plays an essential role in the online event selection, where the single lepton triggers are fully efficient, starting from very low transverse momentum (p_{T}) thresholds (20-35 GeV). These channels are, therefore, competitive despite the low BR, especially in the low mass range of the reconstructed resonance (see figure 1.4.3). In the higher mass range, the boosted topology of the final states is reflected in collimated leptons, which suffer from reconstruction and identification efficiencies due to the lack of isolation; for this purpose, dedicated strategies are often employed to improve the selection efficiency for boosted events (cf. sections 3.2.1 and 3.2.2). The major sources of background for this final state arise from Z+jet and diboson (VV) production. To further suppress the background and the pileup contribution, the jet coming from the boson decays is selected using jet tagging discriminators and algorithms (cf. section 3.3.2).

The searches based on ℓ+jet final states are characterised by a W boson decaying into a lepton and neutrino and a Higgs or vector boson decaying hadronically, with a BR ranging from 15% to 34%. The analysis strategies are similar to the one described for the $\ell\ell$+jet final states; additionally, a W mass constraint can be applied to estimate the z-component of the $p_{\mathrm{T}}^{\mathrm{miss}}$ (see section 3.5), which allows the reconstruction of the undetected neutrino. The main background in these searches is coming from $\mathrm{t\bar{t}}$ and W+jet production. This final state has a high sensitivity throughout the whole mass range provided by a 2D fit in the plane defined by the reconstructed diboson (m_{WV}) and jet (m_{jet}) masses [87].

The $E_{\mathrm{T}}^{\mathrm{miss}}$+jet searches involve the presence of zero charged leptons and a pair of neutrinos, coming from the invisible decays of the Z boson ($\mathrm{Z} \to \nu\nu$) and balanced by the hadronically decaying Higgs or vector bosons. In this case, the online trigger requirement is based on the missing transverse momentum, $p_{\mathrm{T}}^{\mathrm{miss}}$, with a high threshold ($\sim 200\,\mathrm{GeV}$) to ensure stable performance. The main backgrounds are coming from Z+jet and W+jet productions, and the BR is varying between 12% and 27%. Given that it is not possible to reconstruct the Z boson 4-momentum, the resonance's transverse mass (m_{T}) is used instead as a sensitive variable for this type of searches; the resulting broader distribution is reflected in reduced performance, placing the sensitivity of these analyses in between the ℓ+jet and $\ell\ell$+jet searches [86, 88].

The all-hadronic final states have the highest BR, ranging from 33-40% to 45-50% depending on the presence or absence of a Higgs boson. However, the QCD multijet background is overwhelming, and it complicates the analyses in both online and offline strategies [89]. Due to the high production rate of low-p_{T} multijet events, the online triggers are based either on H_{T}[11] thresholds above 700 GeV or boosted single-jet with groomed mass $m_{\mathrm{jet}} > 30\,\mathrm{GeV}$ and minimum p_{T} of 360 GeV. This combination allows a reasonable data-taking rate and a stable performance for dijet invariant masses starting from 1 TeV. To further reduce the background contamination, other offline selections are often used, including requirements on the angular separation between jets to reduce the t-channel QCD production, and the jet mass and substructure variables (see section 3.3.2), which can be used to discriminate quark- or gluon-initiated jets from those produced by the hadronic decays of the heavy bosons. The massive usage of jets makes the all-hadronic final states analyses, and in general all analyses using jets or $p_{\mathrm{T}}^{\mathrm{miss}}$, particularly sensitive to jets properties. In particular, precise calibration of the jet energy plays a crucial role in the accurate description of the p_{T} spectra and the reduction of the systematic uncertainties. An overview of the jet calibration procedure is reported in section 3.4, and a comprehensive description of the jet energy resolution is detailed in chapter 5.

All the types of searches presented above have in common the presence of jets, which leads to different analysis strategies depending on the resonance mass and final states under consideration. Analyses targeting the low-p_{T} regime consider the quarks in the final states to be sufficiently separated to be resolved into single small-radius jets. On the other hand, bosons originating from high-mass resonances have a large Lorentz boost, which is reflected in collimated decay products. In this case, the hadronic decays can be reconstructed as single jets with a larger radius than in the resolved categories. Consequently, jet substructure variables, b tagging and jet tagging algorithms play a crucial role in removing background contributions. Since the jet mass is often used to discriminate jets, great effort is put into developing these algorithms to make them as insensitive as possible to the jet mass. Further discussion on jet tagging and their mass decorrelation can be found in chapter 4.

Among the diboson searches, the VH and HH final states are of particular interest: the main difference is the presence of the Higgs boson, which increases the multiplicity of final states [90]. The $\mathrm{H} \to \mathrm{b\overline{b}}$ decay is commonly used as it comes with the highest BR (see table 1.2.1) and good background discrimination provided by the b tagging algorithms (see section 3.3.1). In particular, the HH searches in the $\mathrm{b\overline{b}b\overline{b}}$

[11] H_{T} is defined as the scalar sum of the reconstructed transverse jet momenta.

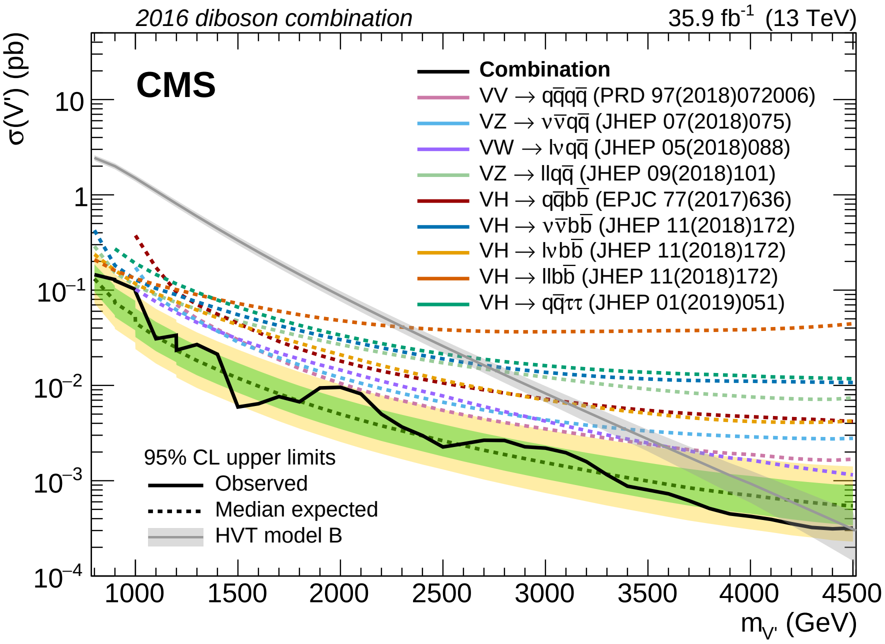


Figure 1.4.3: Observed and expected upper limits at 95% CL on the production cross section as a function of the HVT resonance mass. Taken from Ref. [92].

final states are expected to be quite competitive, despite the overwhelming QCD background [91]. Other final states, involving smaller BRs but cleaner signatures, are also considered: for example, H → $\gamma\gamma$, despite the tiny BR, achieves the best sensitivity at low masses thanks to the excellent resolution of the invariant mass of the reconstructed Higgs boson candidate and a very low QCD background contribution; the H → $\tau^+\tau^-$ decay can reach similar sensitivity as the H → b$\bar{\text{b}}$ analysis thanks to the improvements in the τ tagging algorithms [91]. The H → VV* → qqqq and H → c$\bar{\text{c}}$ decays have not been explicitly considered so far even though they are included in the all-hadronic channel [86].

Ultimate sensitivity can be achieved by combining the results obtained in the individual final states. Resonances involving VV and VH, and HH searches are combined separately to address different models [91, 92]; in particular, the HH analyses can be compared to BSM Higgs models predicting enhanced Higgs boson pair production. The most recent combinations by the CMS Collaboration are based on approximately 36 fb^{-1} of 13 TeV data, and an update of these results based on the entire 13 TeV dataset is expected in the near future. The expected and observed upper limits on the production cross section of the combination of the VV and VH resonant searches is shown in figure 1.4.3 for the HVT model.

The work presented in this thesis focuses on the yet unexplored H → c$\bar{\text{c}}$ and H → WW* → qqqq final states; in particular, a novel approach is investigated in chapter 6 that extends the usage of these decays, showing how they can play an essential role in BSM searches. Furthermore, a veto on jets originating from b quarks is required in order to ensure orthogonality with the H → b$\bar{\text{b}}$ channel; it will be shown how a further optimisation of this prerequisite is required in the future to maximise the contribution of each channel.

2
The experimental setup

The analysis presented in this thesis uses data recorded by the Compact Muon Solenoid (CMS) detector located at the Large Hadron Collider (LHC). This chapter provides a brief description of the experimental framework, both the accelerator system and the detector, focusing on the elements relevant in the context of this thesis.

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [93] is a two-ring particle collider operated by the European Organization for Nuclear Research (CERN), and it is built in the underground tunnel previously used for the Large Electron-Positron Collider (LEP).

The tunnel lies between 45 m and 170 m below the surface, and is 26.7 km long with eight straight sections and eight arcs; the bending dipole and focusing quadrupole magnets are located in each arc, while the straight sections host interaction points (IPs) with detectors or utilities, i.e. beam injectors and dump facilities, radio-frequency cavities, and collimation systems. The beams are guided around the accelerator ring by a strong magnetic field ($B_{max} = 8.33\,\mathrm{T}$) maintained by 1232 superconducting niobium-titanium (NbTi) dipole magnets. Additionally, a total of 392 quadrupole magnets are destined to focus the beam, while superconductive radio-frequency cavities, tuned at 400 MHz, are employed to increase the energy of the injected proton beams from 450 GeV up to 7 TeV. The magnets' design had to comply with the pre-existing LEP tunnel; as a consequence of the limited space, twin-bore magnets were adopted.

Four different experiments with different characteristics and purposes are located at the four IPs. The ATLAS (A Toroidal LHC ApparatuS) and CMS (Compact Muon Solenoid) experiments are designed to investigate a broad range of phenomena, focusing on the Higgs boson measurement and the exploration of the TeV energy frontier. The ALICE (A Large Ion Collider Experiment) experiment is a detector optimised for heavy-ion collisions to study the physics of the strong interaction at extremely high energy densities (cf. section 1.1.3). The LHCb (Large Hadron Collider beauty) experiment is specialised in in b quark physics and CP-violating measurements to address the matter-antimatter puzzle (cf. section 1.3).

As shown in figure 2.1.1, the LHC is the final part of the CERN accelerator complex, a chain of pre-accelerator utilised to increase the beam energy in stages.

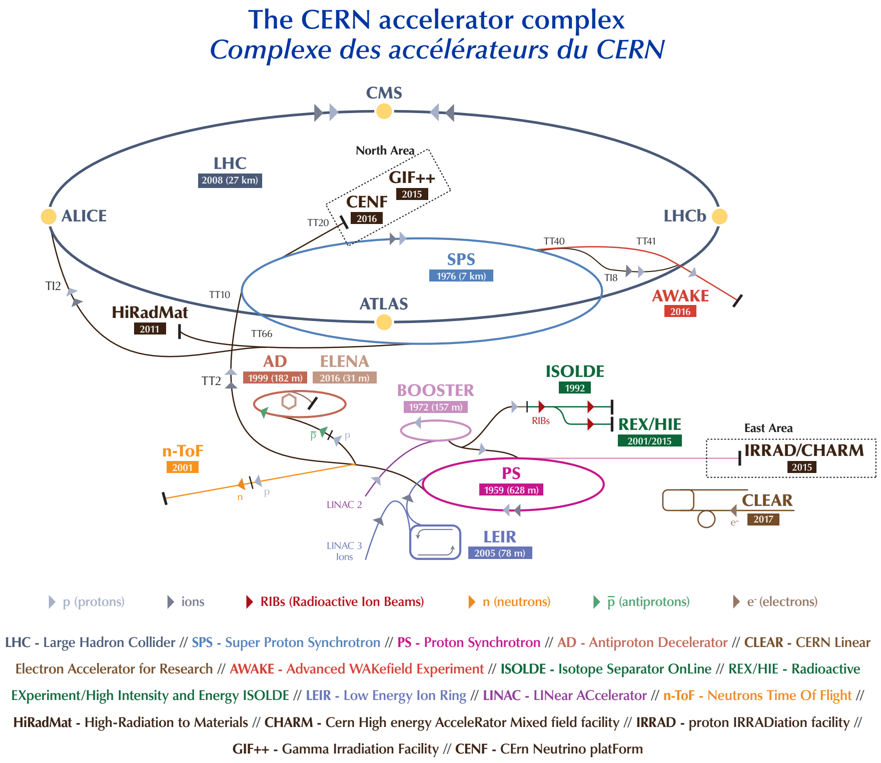

The CERN accelerator complex
Complexe des accélérateurs du CERN
CMS
LHC
2008 (27 km)
North Area
GIF++
2015
CENF
2016
ALICE
LHCb
TT20
SPS
1976 (7 km)
TT40
TT41
TI2
TI8
TT10
ATLAS
AWAKE
2016
HiRadMat
2011
TT66
AD
1999 (182 m)
ELENA
2016 (31 m)
TT2
BOOSTER
1972 (157 m)
ISOLDE
1992
RIBs
REX/HIE
2001/2015
East Area
IRRAD/CHARM
2015
n-ToF
2001
p̄
p
n
PS
1959 (628 m)
CLEAR
2017
e⁻
LINAC 2
LINAC 3
Ions
LEIR
2005 (78 m)
p (protons)
ions
RIBs (Radioactive Ion Beams)
n (neutrons)
p̄ (antiprotons)
e⁻ (electrons)
LHC - Large Hadron Collider // SPS - Super Proton Synchrotron // PS - Proton Synchrotron // AD - Antiproton Decelerator // CLEAR - CERN Linear Electron Accelerator for Research // AWAKE - Advanced WAKefield Experiment // ISOLDE - Isotope Separator OnLine // REX/HIE - Radioactive EXperiment/High Intensity and Energy ISOLDE // LEIR - Low Energy Ion Ring // LINAC - LINear ACcelerator // n-ToF - Neutrons Time Of Flight // HiRadMat - High-Radiation to Materials // CHARM - Cern High energy AcceleRator Mixed field facility // IRRAD - proton IRRADiation facility // GIF++ - Gamma Irradiation Facility // CENF - CErn Neutrino platForm

Figure 2.1.1: Sketch of the LHC accelerator complex. Taken from Ref. [94].

Protons are collected by ionisation of the hydrogen gas source and accelerated in steps by the Linear Accelerator (LINAC 2), followed by the Proton Synchrotron Booster (PSB), the Proton Synchrotron (PS) and, in the end, the Super Proton Synchrotron (SPS), where protons are accelerated until they reach the energy of 450 GeV and are finally grouped into beams. The last step consists of the injection of two counter-rotating beams into the LHC rings. Each beam is split into $n_b = 2808$ bunches with a nominal amount of protons per bunch of $N_p^b = 1.15 \times 10^{11}$. The bunches are separated by 25 ns in time, leading to a bunch crossing rate $r_{\mathrm{coll}} \sim 40\,\mathrm{MHz}$. The LHC design enables either proton beams with centre-of-mass energy ($\sqrt{s}$) up to 14 TeV or lead ion beams ($^{208}\mathrm{Pb}^{82+}$) with beam kinetic energy up to 2.76 TeV/nucleon, which corresponds to $\sqrt{s} = 1.15\,\mathrm{PeV}$ per ion-ion collision. This thesis uses data from proton-proton (pp) collisions collected by the CMS experiment between the years 2016 and 2018, defined as Run 2[1].

The number of events per second generated in the collisions at LHC is given by $N(t) = \mathcal{L}(t) \cdot \sigma$, where the cross section σ depends on the event under study. The instantaneous luminosity $\mathcal{L}$ in a collider, assuming two Gaussian bunches with similar properties, is given in the following:

$$\mathcal{L} = \frac{n_b N_p^{b_1} N_p^{b_2} r_{\mathrm{coll}}}{4\pi\sigma_x\sigma_y}\mathcal{F}\,. \tag{2.1}$$

[1] The data collected during 2015 is also part of the nominal Run 2 definition, but often not used for analysis given its little statistics. Additionally, the previous data-taking period, known as Run 1, was characterised by $\sqrt{s} = 7$ and 8 TeV for the years 2011 and 2012, respectively.

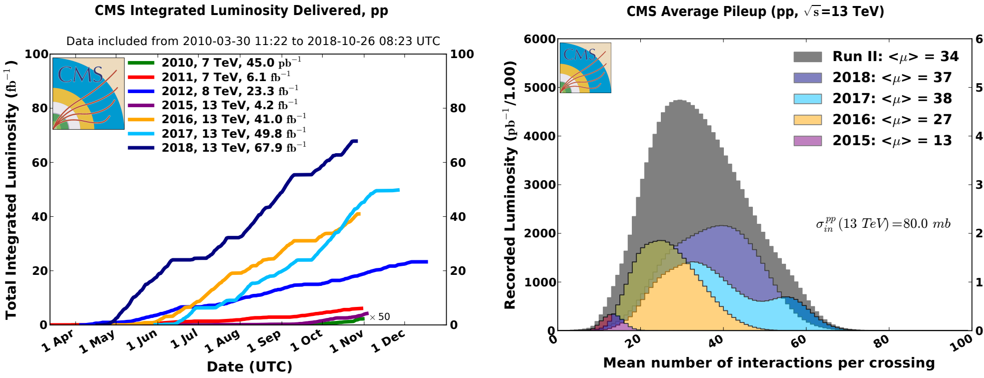


Figure 2.1.2: Left: Cumulative luminosity versus day delivered to CMS for pp collisions. The centre-of-mass energy and the total luminosity is reported for each year. Taken from Ref. [95]. Right: Distribution of the mean number of inelastic interactions per bunch crossing (pileup) in data for pp collisions. Taken from Ref. [96].

This quantity depends on beam-related parameters: $\sigma_{x,y}$ are the transverse beam sizes at the IP and $\mathcal{F}$ depends on the beam size and the crossing angle.

The LHC peak luminosity is expected at $\mathcal{L}(t) \sim 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, value that corresponds to the luminosity in the CMS and ATLAS IPs. Instead, the LHCb experiment adopted a different choice. The beam collision angle, as well as the detector geometry, is physics-motivated by the high density of b quarks originating from the colliding beams, which are emitted mainly at small angles along the beam direction. In order to keep the complexity of the event reconstruction at a manageable level, the LHCb experiment has operated at a luminosity circa 20 times smaller than the maximum luminosity provided by the LHC. Figure 2.1.2 (left) reports the cumulative luminosity as a function of time delivered to CMS.

Physics analyses benefit from high luminosities as the rates of rare processes are increased. The disadvantage comes from multiple pp collisions occurring during one proton-bunch crossing. The probability that more than one interaction produces an interesting process is negligible [97]. Therefore, at the analysis level, only the most energetic collision per event is selected, referred to as the primary hard interaction, while the other collisions in the event are called pileup (PU) interactions. A large amount of unavoidable PU events is an obstacle for the data taking and reconstruction since it produces additional overlapping particles throughout the detector that deteriorate the measurement accuracy. The pileup profiles of data collected by the CMS detector for different years are shown in figure 2.1.2 (right).

Coordinate system

The LHC coordinate system has its origin fixed at the nominal IP. The x-axis points towards the centre of the LHC ring, the y-axis points upwards, and the z-axis points along the counter-clockwise beam direction. Besides, the azimuthal angle ϕ is measured from the x-axis in the x-y plane, and the polar angle θ is measured from the positive z-axis. Experiments like CMS commonly utilise a cylindrical coordinate system (r, ϕ, η), where r is the distance from the z-axis and the pseudorapidity $\eta = -\ln\tan(\theta/2)$ is the relativistic limit of the rapidity of a particle, $y_z = \frac{1}{2}\ln\frac{E+p_zc}{E-p_zc}$, which depends on particle energy (E) and longitudinal momentum ($\mathrm{p_z}$).

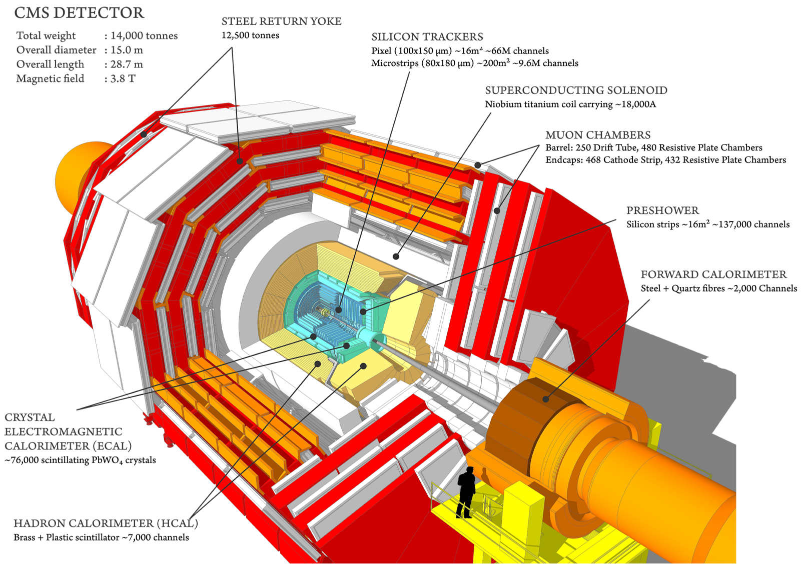

CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~16m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO$_4$ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Figure 2.2.1: Cut-away view of the CMS detector, showing the different layers around the LHC beam axis, with the collision point in the centre. Taken from Ref. [99].

In a typical pp collision, the centre of mass is boosted along the z-axis with respect to the laboratory frame. Therefore, the kinematics of the collision products are conveniently described by a set of 4 variables (p_{T}, y_z, ϕ, m). Here, m indicates the particle invariant mass, and $p_{\mathrm{T}} = p \sin\theta$ its transverse momentum. The transverse momentum, the azimuthal angle and the mass are invariant under boosts along the z-axis, while the rapidity changes only by an additive constant; therefore, the difference in rapidity between two particles is invariant under boosts along the z-axis.

2.2 The Compact Muon Solenoid detector

As one of the two multi-purpose detectors at the LHC, the Compact Muon Solenoid (CMS) detector [98] design was dictated by the quest for an excellent resolution of reconstructed leptons and photons, key components for the Higgs boson search in the $\mathrm{H} \to \gamma\gamma$ and $\mathrm{H} \to 4\mathrm{l}$ "golden" channels. Aiming at reconstructing each pp collision event in its entirety, this multi-purpose detector is based on a single high-field solenoid for detecting muons, together with pixel- and microstrip-based tracking system, an electromagnetic calorimeter comprising scintillating crystals for analysing electrons and photons, and a hadron calorimeter for jet energy measurement.

The most important feature of the experiment layout, shown in figure 2.2.1, is the state-of-the-art superconducting solenoid, which allows a compact cylinder-shaped design and provides a uniform magnetic field of 3.8 T. The solenoid itself is approximately 13 m long with a diameter of 6 m. It contains, from the inside out, the tracking system and the electromagnetic and hadron calorimeters. Outside the magnet coil, the iron return yoke of the magnet hosts the muon chambers. The overall length of the CMS detector is 21.6 m, the diameter 14.6 m and the total weight about 14500 t. The structure of the detector consists of two regions: the *barrel* with $|\eta| \leq 1.2$ made of sub-detectors positioned at increasing values of the cylinder radius, and the *endcaps* ($|\eta| \geq 1.2$) where sub-detectors are layered along the beam axis.

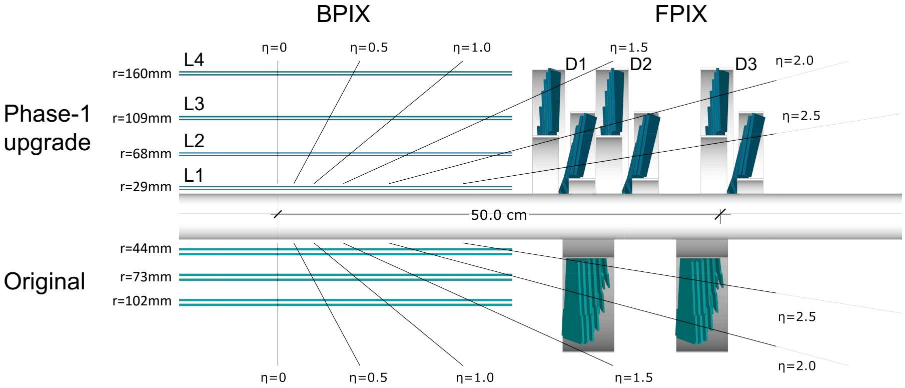


Figure 2.2.2: Comparison of layer and disk positions between the original and the Phase-1 pixel trackers. Taken from Ref. [101].

2.2.1 Tracking system

The tracking system [100], or tracker, constitutes the innermost part of the CMS detector that outgoing particles from the collisions encounter; it exploits the ionization process, which occurs when charged particles move through matter, to provide a precise measurement of the charged-particle trajectories and, as a consequence, an efficient reconstruction of the primary and secondary interaction vertices.

The tracker lies inside the almost uniform co-axial magnetic field provided by the CMS solenoid (see section 2.2.3). Its total length and diameter are of 5.8 m and 2.5 m, respectively, while the angular coverage reaches up to $|\eta| = 2.5$, for a total active surface of 210 m^2, which made it by far the largest silicon tracker ever built. A silicon pixel detector (PIXEL) is installed in the innermost region, closest to the IP, while silicon microstrip detectors are used in the outer region.

The PIXEL component of the original design is made of three co-axial barrel layers (BPIX) at radii of 4.4, 7.3 and 10.2 cm, and two pairs of endcap disks (FPIX) located at $|z| = 34.5$ cm and $|z| = 46.5$ cm, placed at a distance from the beam pipe of 6 to 15 cm, respectively, and covering the region of $|\eta| < 2.5$. This specific geometry was chosen to profit from the high pixel granularity to obtain a three-dimensional measurement of the hit position, a key component for a precise vertex reconstruction. The resulting hit position resolution of PIXEL is approximately 10 µm in the transverse coordinate and 20-40 µm in the longitudinal coordinate, depending on η.

During the 2016-2017 end-of-the-year shutdown, a new pixel detector has been installed to cope with the elevated level of exposure to radiation due to the large flux of particles and the high luminosity regime. The original sub-detector was replaced with a 4-layer barrel and 3-disk endcap system, known as the Phase-1 version. Figure 2.2.2 shows the layout for the Phase-1 upgrade pixel detector [101]. The reduced proximity of the first BPIX layer, the additional BPIX layer and FPIX disk, and the newer readout chips provided an improved track impact parameter resolution, redundancy in pattern recognition, higher tracking efficiencies[2] and reduction of track fake rates[3].

A strip silicon detector is used in the outer region between $20 < r < 110$ cm, where the flux of particles is smaller. The strip tracker surrounds the PIXEL de-

[2] Ratio between truth tracks matched to reconstructed tracks and truth tracks.

[3] Ratio between reconstructed tracks not matched to truth tracks and reconstructed tracks.

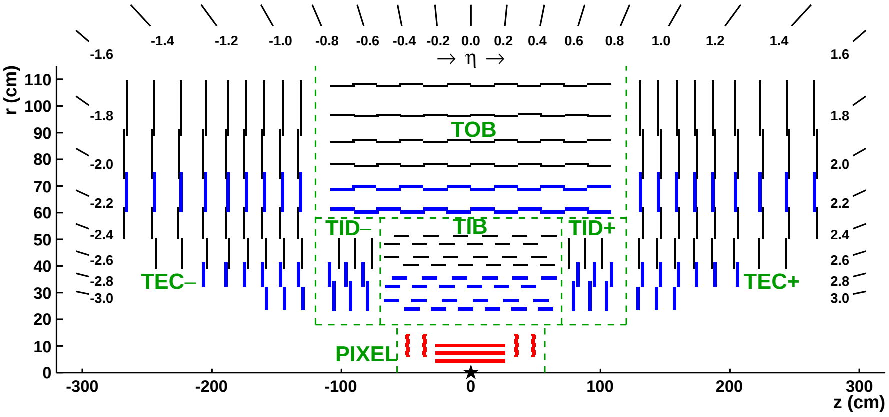


Figure 2.2.3: Schematic cross section through the CMS tracker in the r-z plane. The centre of the tracker, corresponding to the approximate position of the pp collision point, is indicated by a star. The pixel modules are shown by red lines. Back-to-back strip modules, which allow a 3D hit position reconstruction, are shown by blue lines, while the rest of the strip modules are shown by black lines. Taken from Ref. [100].

tector and is divided into four subsystems; the Tracker Inner Barrel (TIB) and Disks (TID) cover $r \leq 55\,\text{cm}$ and $|z| \leq 118\,\text{cm}$, and are composed of four-barrel layers and three disks on each side, respectively; the Tracker Outer Barrel (TOB) occupies the region of $55 \leq r \leq 110\,\text{cm}$ and $|z| \leq 118\,\text{cm}$ with six-barrel layers; the Tracker Endcap (TEC) covers the region $124 \leq |z| \leq 282\,\text{cm}$ using nine disks on each side. The strip configuration allows to simultaneously measure the transverse and longitudinal hit position in the pseudorapidity region up to $\eta = 2.5$, providing an r-ϕ resolution of approximately 10-50 µm. A schematic drawing of the CMS tracker is shown in figure 2.2.3.

2.2.2 Calorimeter system

The CMS experiment uses four calorimeters to efficiently measure the energies of different particles over an extensive η range. Mostly focused on heavy-ion and diffractive pp physics studies, the two forward calorimeters, CASTOR (Centauro And Strange Object Research) and ZDC (Zero Degree Calorimeter), have been designed to complement the CMS measurements in the very forward region ($|\eta| > 5.2$) [98]. The other two detectors, ECAL (Electromagnetic CALorimeter) and HCAL (Hadron CALorimeter), are used to achieve optimal measurement of electrons, photons and hadrons within a large pseudorapidity range ($|\eta| \leq 5.2$). This thesis uses data measured using only the last two detectors only.

Placed around the tracker, two calorimeter tiers measure the particle's energy and provide information complementary to the tracker. When a particle traverses a calorimeter, it initiates a shower of secondary particles that are used to measure the total energy of the initial particle. The shower grows until the particles' energy is sufficiently low for them to be captured and absorbed by the surrounding detector material. This is a destructive technique as the particles are destroyed in the process.

Electromagnetic showers in calorimeters are formed either through bremsstrahlung ($e^{\pm} \rightarrow \gamma e^{\pm}$), or photon pair-production ($\gamma \rightarrow e^{+}e^{-}$) processes. The average distance covered by an electron before its energy is reduced by a factor $\frac{1}{e}$ is known as radi-

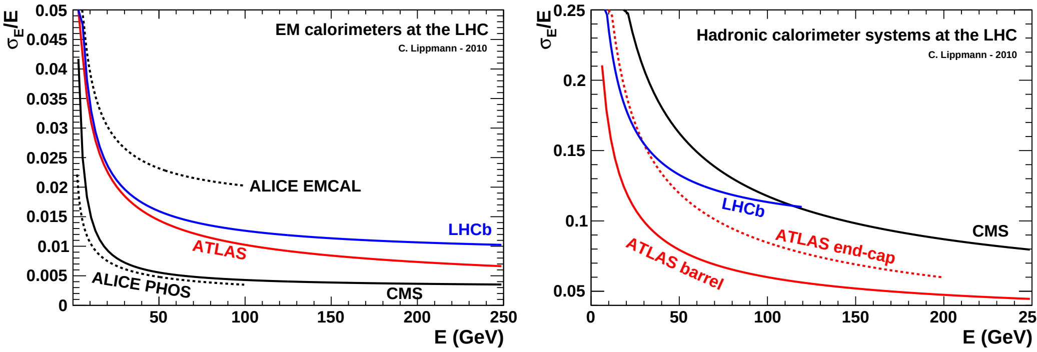


Figure 2.2.4: Comparison of the relative energy resolutions of EM (left) and hadronic (right) calorimeters for different LHC experiments derived from test-beam data. The hadron calorimeter resolution for the ATLAS and CMS includes the corresponding EM calorimeter contribution. Taken from Ref. [102].

ation length (X_0). It is equivalent to $\frac{7}{9}$ of the mean free path of the photon, the average distance covered before pair production occurs. In the transverse direction, the extent of a shower is characterised by the Molière radius R_M: a cylinder of radius R_M contains on average 90% of the shower's energy. Both X_0 and R_M are material dependent.

Hadrons shower through inelastic interactions, including multi-particle production (e.g. pion pair production) and nuclear decays. The characteristic length used to describe hadronic showers is the nuclear interaction length λ, the average distance crossed before undergoing an inelastic nuclear interaction. Also in this case, λ is material dependent, and generally larger than X_0. Part of the hadronic showers includes neutral pions, which decay electromagnetically; therefore, only a small fraction (approximately 10-15%) of the total energy can be registered by an electromagnetic calorimeter. Considering the larger characteristic length of hadronic showers and the presence of an EM component, electromagnetic calorimeters usually precede hadron calorimeters in the detector layout.

Technology-wise, two broad categories of calorimeters exist: homogeneous calorimeters, consisting entirely of active materials, and sampling calorimeters, alternating active and passive materials. Homogeneous calorimeters require a larger volume for shower containment compared to sampling calorimeters, which manage to contain particle showers with smaller material thickness, thanks to the absorber layers. Additionally, homogeneous calorimeters provide excellent energy resolution. In contrast, sampling calorimeters are more suited for particle identification because of the better spatial resolution. The material thickness requirement is often the main argument for sampling calorimeters. Besides, it reduces construction costs.

The design of the CMS electromagnetic calorimeter was based on reaching excellent energy and angular resolutions of photons and electrons. The main benchmark channel was H → $\gamma\gamma$, considered one of the "golden" channels for the Higgs boson discovery. The hadron calorimeter is designed to measure both the energy and direction of the hadronic component of jets precisely. The correct measurement of the energy in the event is a key component for an accurate derivation of the missing energy flow produced by neutrinos. In figure 2.2.4, the relative energy resolution of the CMS calorimeters is compared to other LHC experiments.

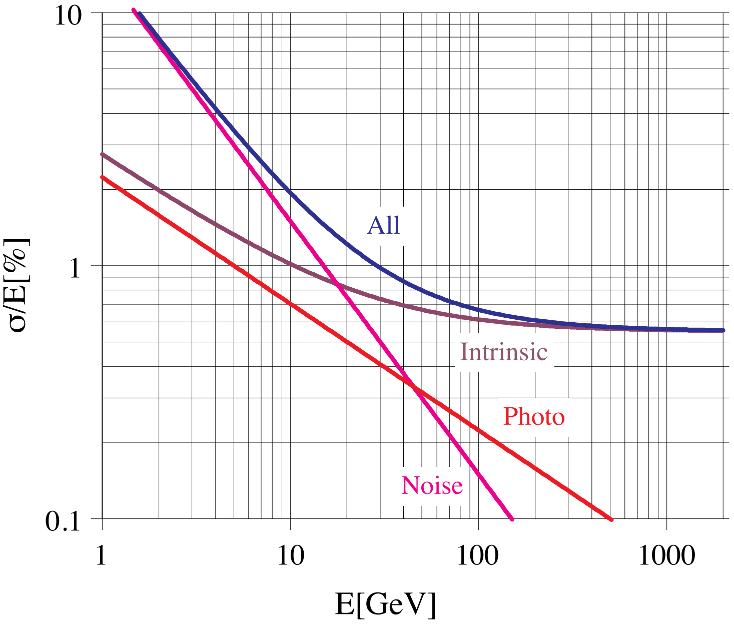


Figure 2.2.5: Different contributions to the energy resolution of the $PbWO_4$ calorimeter as a function of the energy. *Photo* and *intrinsic* refer to the stochastic and constant terms, respectively. Taken from Ref. [103].

The relative energy resolution of a calorimeter can be parametrised as

$$\left(\frac{\sigma}{E}\right) = \left(\frac{N}{E}\right) \oplus \left(\frac{S}{\sqrt{E}}\right) \oplus C \,. \tag{2.2}$$

The first term N is connected to the electronics noise and dominates at low energies. The second term S is a stochastic term accounting for fluctuations in the number of photons and electrons, and the last one C is a constant term related to the calibration of the calorimeter, which limits the calorimeter performance at high energies. The different contributions to the energy resolution are reported in figure 2.2.5, where the CMS electromagnetic calorimeter is taken as an example.

Electromagnetic calorimeter

The CMS electromagnetic calorimeter (ECAL) [103] is a hermetic, homogeneous scintillating crystal calorimeter, which offers great performance in terms of energy resolution since most of the energy from electrons or photons is deposited within the volume of the calorimeter. It is divided into barrel and endcaps; the ECAL barrel (EB) covers up to $|\eta| = 1.48$, extending from $r = 1.3\,\mathrm{m}$ to $r = 1.8\,\mathrm{m}$; the two ECAL endcaps (EE) cover the region $1.48 \leq |\eta| \leq 3.0$, extending from $|z| = 3\,\mathrm{m}$ to $|z| = 3.8\,\mathrm{m}$. A 20 cm thick pre-shower (ES) is placed in front of the EE covering the region $1.65 \leq |\eta| \leq 2.6.$, and consists of two active silicon strips and passive lead absorber layers, helping to distinguish energetic photons from a pair of a very close photons originating from a π^0 decay ($\pi^0 \to \gamma\gamma$).

The active material is lead tungstenate ($PbWO_4$) and has been chosen for its high density ($\rho = 8.28\ \mathrm{g\,cm^3}$), short radiation length ($X_0 = 0.89\,\mathrm{cm}$) and a small Molière radius ($R_M = 2.2\,\mathrm{cm}$). Thanks to these properties, the electromagnetic calorimeter

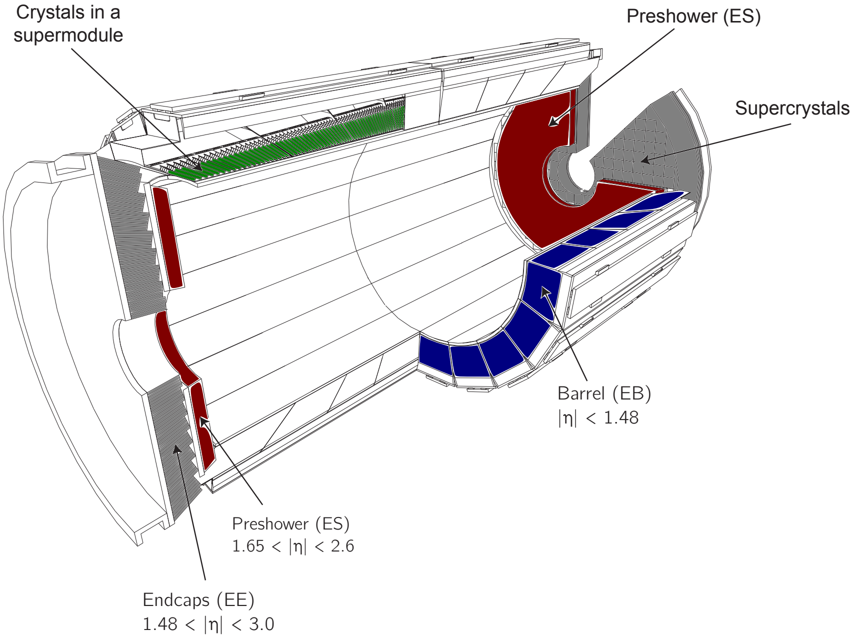


Figure 2.2.6: Longitudinal view of the CMS ECAL. Adapted from [104].

has a compact shape and a fine granularity; furthermore, as a result of the fast time-response of its crystals, 80% of the emitted signals are collected within 25 ns. The longitudinal section of the calorimeter is shown in figure 2.2.6.

The ECAL energy resolution is described by eq. (2.2), with the parameters reported in table 2.2.1; a graphic illustration of each contribution is shown in figure 2.2.5.

Contribution	EB ($\eta = 0$)	EE ($\eta = 2$)
Stochastic term	$2.7\%/\sqrt{E}$	$5.7\%/\sqrt{E}$
Constant term	0.55%	0.55%
Noise (high luminosity)	155 MeV	205 MeV
Noise (low luminosity)	210 MeV	245 MeV

Table 2.2.1: Contribution to the energy resolution of the ECAL design for barrel and endcap regions (E is expressed in GeV). Taken from Ref. [103].

Hadron calorimeter

Surrounding the ECAL, the hadron calorimeter (HCAL) [105] provides the energy measurements for hadrons in the event.

The HCAL is a sampling calorimeter with alternating layers of absorbers and scintillators, divided into four sub-detectors: the Hadron Barrel (HB) surrounds the electromagnetic calorimeter and covers the central pseudorapidity region up to $|\eta| = 1.3$; the Hadron Outer (HO), located outside the solenoid in the same η region of the HB, is used to avoid the misidentification of muons by catching the tails of the very energetic hadronic showers that escape the HB containment; the Hadron Endcap (HE) covers between $1.3 < |\eta| < 3.0$ on each side; the Hadron Forward (HF) extends the coverage up to $|\eta| = 5.2$ in the forward region.

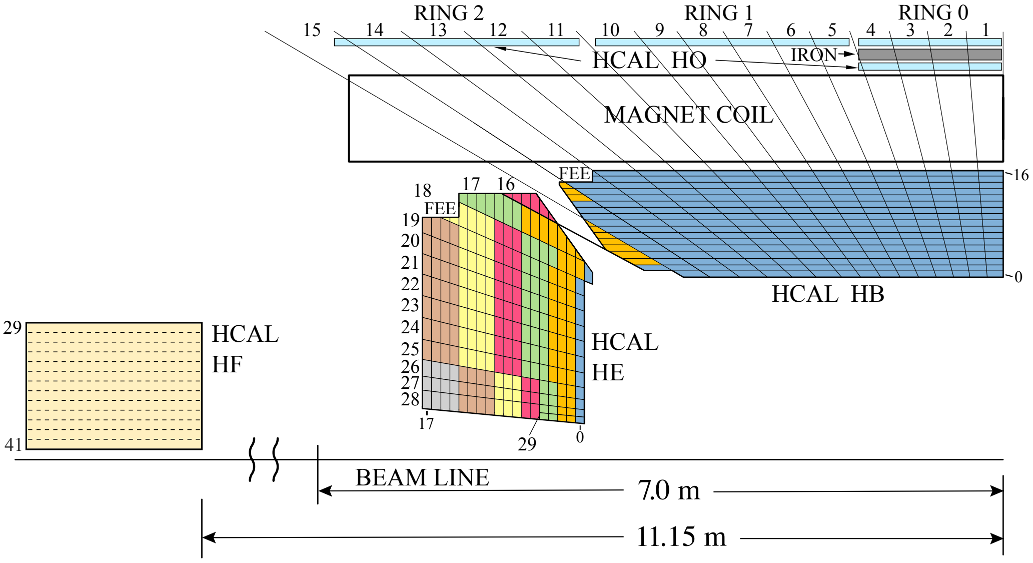


Figure 2.2.7: Longitudinal view of the HCAL detector. The calorimeter towers are separated by the black lines and indicated by the numbers. Each colour indicates the segmentation of scintillator layers available at the end of Run 2. Taken from Ref. [107].

The HB and HE are located inside the magnetic field within $1.77 \leq r \leq 2.95\,\mathrm{m}$, necessitating the use of materials that are not ferromagnetic; brass was chosen as absorber material, because of its properties, among which its relatively low nuclear interaction length ($\lambda = 16.42\,\mathrm{cm}$). Plastic scintillator tiles are used as active material in the HB and HE. The η-ϕ unit area in the HCAL, referred to as tower, has a granularity of 0.087×0.087, equivalent to the area of a 5×5 array of the ECAL crystals, but it can reach values of 0.175×0.175 in the HE, depending on η.

The HF is located in the forward part of the detector at $|z| = 11.2\,\mathrm{m}$. The absorber plates are made of steel to sustain very high fluxes of particles, while the quartz-based fibres collect the Cherenkov light produced by showers in the absorber. Since it provides good precision for energy measurements, it is used for the CMS luminosity measurements, as described in section 2.2.5.

The energy resolution of the HCAL can be described by eq. (2.2). Up to 300 GeV and including the ECAL contribution, the values of the stochastic and constant terms are $S = (111.5 \pm 2.1\%)\,\mathrm{GeV}$ and $C = (8.6 \pm 1.4\%)$, respectively, while the contribution from the noise N is negligible [106].

In order to improve upon the HCAL performance, several upgrades have been planned with a schedule over the whole Run 2. These changes, known as HCAL Phase-1 upgrades, include the replacement of the photodetectors and the front-end electronics, and the increment of the depth segmentation in the scintillators; the goal is to decrease the rate of anomalous signals and include new features such as signal timing information for improved calibration and a better pileup discrimination. Figure 2.2.7 shows the longitudinal view of the HCAL detector after the 2017-2018 end-of-the-year shutdown, during which the endcaps were upgraded to provide a depth read-out.

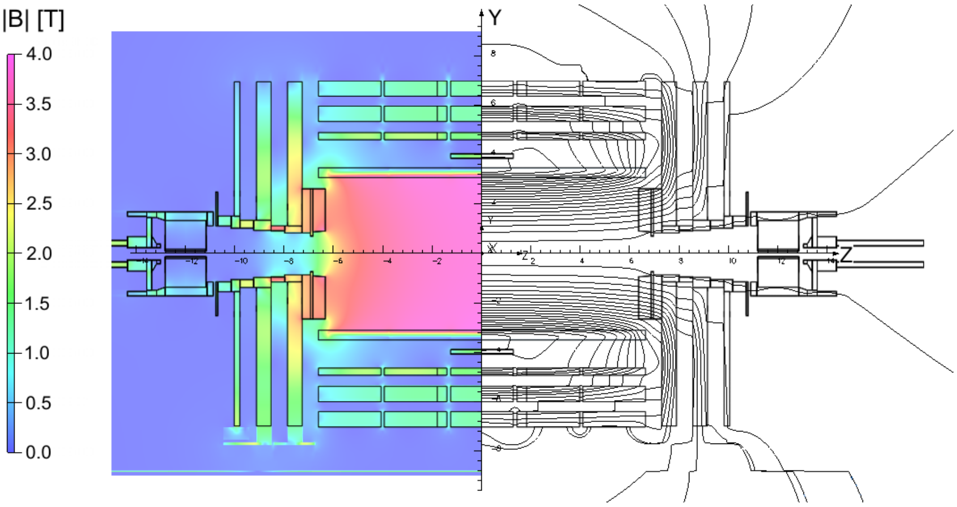


Figure 2.2.8: Intensity of the magnetic field $|B|$ and field lines in a longitudinal section of the CMS detector. Taken from Ref. [108].

2.2.3 Superconducting solenoid

One of the hallmarks of the CMS detector is its magnet. It provides a strong magnetic field of 3.8 T that bends the trajectories of charged particles coming from LHC collisions, allowing for a precise measurement of their transverse momenta.

The magnet is composed of a superconducting solenoid, made of 4 layers of NbTi, and is surrounded by an iron yoke, which closes the magnetic field loop and hosts the muon system. Figure 2.2.8 shows the intensity of the magnetic field inside the detector. The 2 T magnetic field in the iron yoke points in the direction opposite to the direction of the field inside the coil. Therefore, high momentum muons are bent in two opposite directions in the tracking and in the muon systems, improving the resolution of their momentum measurement.

2.2.4 Muon system

The CMS design was optimised also to reconstruct and identify muons, significantly contributing to the success of the Higgs boson discovery in the $\mathrm{H} \rightarrow 4\mathrm{l}$ final state, and measure their momentum over a wide range.

In the energy range from 1 GeV to 1 TeV, muons lose less than 3% of their energy due to ionisation processes occurring when traversing the entire detector volume. For this reason, the muon system is placed in the outermost part of the CMS detector.

As shown in figure 2.2.9, the muon system is divided into barrel and endcap sections, and it uses three different gas-based detector technologies to measure the muons: drift tubes (DT) in the barrel region, cathode strip chambers (CSC) in the endcap area, and resistive plate chambers (RPC) in both regions.

The barrel section covers the $|\eta| \leq 1.2$ region, where the DT chambers are used and organised into four stations. The first three stations, each composed of eight chambers, distributed to achieve the best angular resolution, are hosted inside the magnetic field return plates: four chambers provide measurements in the r-ϕ plane, and four are used for measurements in the r-z plane. The fourth station is placed outside the magnetic field, and it provides only r-ϕ coordinate measurements.

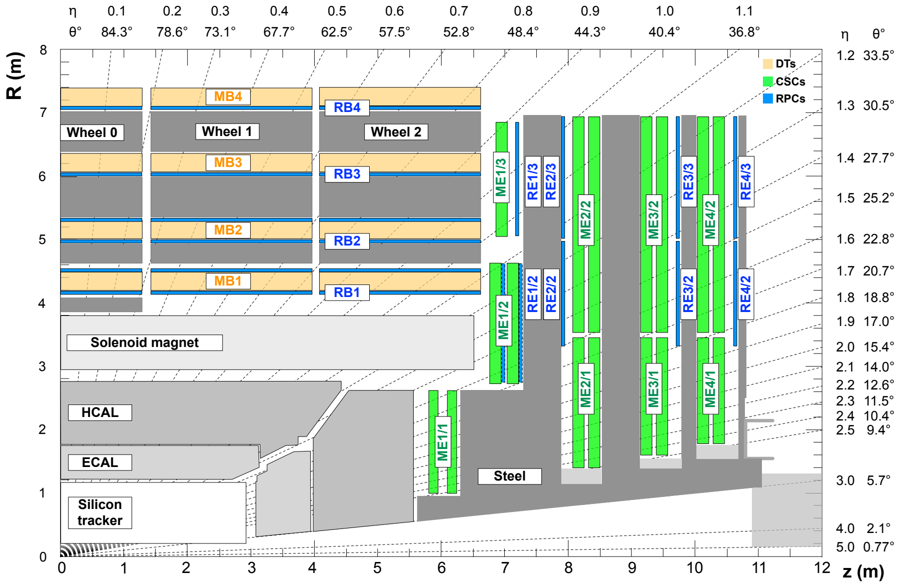


Figure 2.2.9: Quadrant of the CMS detector in the Run 2 configuration with the sub-detectors of the muon system in colour: DT in light orange, CSC in green, and RPC in blue. Taken from Ref. [109].

The two endcap sections allow to identify muons in $0.9 \leq |\eta| \leq 2.4$ region. In this region, where the background noise from radiation is increased with respect to the barrel, the CSC sub-detectors are used to ensure good performance in an environment with high LHC-beam-induced radiation and non-uniform magnetic field conditions. The advantages of the CSC are a fine segmentation, a fast response time and their radiation hardness. Each endcap section consists of four CSC stations inter-layered between the iron yoke plates. In each CSC station, six layers of approximately perpendicular anode wires and cathode strips provide efficient muon tracking and robust pattern recognition for high background rejection. Cathodes and anodes deliver r-ϕ and η measurements, respectively.

A crucial characteristic of the DT and CSC subsystems is their ability to trigger on the p_{T} of muons with good efficiency, high background rejection, and p_{T} resolution of about 15% and 25% in the barrel and endcap, respectively.

In addition to the DT and CSC, RPC sub-detectors are used in the region of $|\eta| \leq 1.6$. RPCs are gaseous parallel-plate detectors, operated in avalanche mode to ensure good operation at high rates with time resolution of about 2 ns. A total of 6 layers of RPCs are embedded in the barrel muon system, and 3 in the endcap stations. They provide a fast and independent trigger, but a coarser position resolution. For this reason, they can be used for redundancy for muon reconstruction.

After the end of Run 2, gas electron multiplier (GEM) detectors have been installed to complement the existing systems in the endcaps and extend the coverage in the very forward region. The first batch of chambers will be used during the forthcoming Run 3 while others will follow for Phase-2 of the LHC.

2.2.5 Luminosity measurement

The luminosity is a key parameter of each collider experiment since it provides the overall normalisation for the yields of all physical processes.

Furthermore, the luminosity uncertainty is often the dominant uncertainty for measurement of processes of high cross section, like W or Z boson production [110]. Besides, the bunch-by-bunch luminosity measurement is essential for real-time monitoring of the performance of the LHC [111].

For the online luminosity measurements, the CMS detector uses the HF calorimeter. Two algorithms are used, each reaching a statistical accuracy of 1% and providing results within 5% of each other. The first counting-based method uses the average fraction of empty HF towers to estimate the mean number of interactions per bunch crossing. The second method is based on the linear relation between the mean luminosity and the average transverse energy deposits.

In addition to the online methods, an offline algorithm based on pixel cluster counting is available for more precise luminosity estimates, given its excellent granularity. Starting from eq. (2.1), it is possible to measure the luminosity; in particular, the cross section σ is calibrated using the Van der Meer scan technique, while N is estimated from the average number of pixel clusters per event; in fact, the pixel response has a linear dependence with the total number of pp interactions.

The total integrated luminosity, calculated by the offline method, recorded by the CMS detector during Run 2 is reported in table 2.2.2.

Luminosity	2015	2016	2017	2018	2015-2018	2016-2018
Delivered (fb^{-1})	4.21	40.99	49.79	67.86	162.85	158.64
Recorded (fb^{-1})	2.26	35.92	41.53	59.74	139.45	137.19
Uncertainty (%)	2.3	2.5	2.3	2.5	1.8	1.8

Table 2.2.2: Run 2 luminosity measurements for pp collisions, based on Ref. [112–115].

2.2.6 Trigger system

At $\sqrt{s} = 13\,\mathrm{TeV}$, the design LHC luminosity is $\mathcal{L}(t) \sim 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, and the total inelastic pp cross-section is approximately 70 mb [116, 117]: this corresponds to a rate of 10^9 events per second from pp collisions. Additionally, the data size per event with zero suppression algorithms applied is of the order of 1 MB; as a consequence, also given the LHC bunch crossing frequency of 40 MHz (25 ns spaced beams), a reduction factor of at least 10^7 on the event rate is necessary to be able to write the information into permanent storage. Furthermore, not all the collisions are of interest for the LHC physics program; yet, the event selection criteria should be as inclusive as possible for unexpected new phenomena that may appear.

The Trigger and Data Acquisition System (TriDAS) [118, 119], a real-time selection and recording of the useful events, reduces the number of collected events for archiving and later offline analysis.

At the trigger level, events are selected based on the objects reconstructed with fast algorithms of limited precision compared to those used in the offline reconstruction (see chapter 3). The required rejection power is too large to be achieved in a single processing step. For this reason, the entire selection task is carried out in two

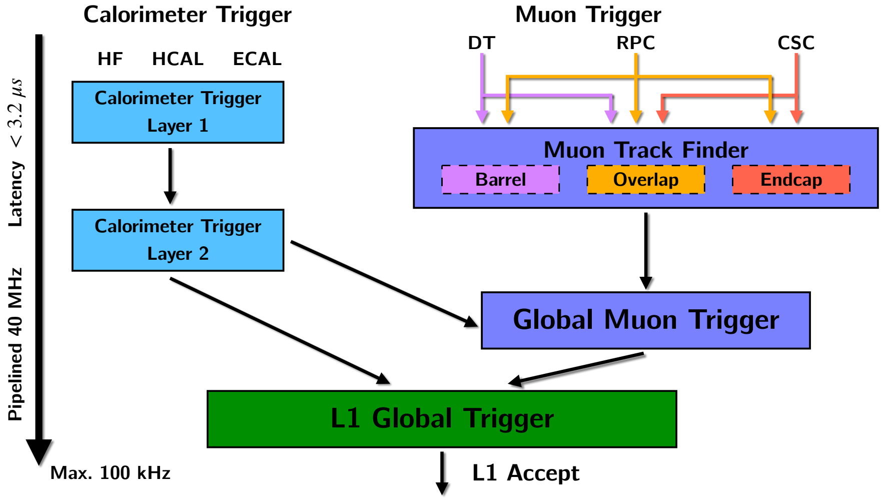


Figure 2.2.10: Schematic view of the L1 trigger system working flow. Adapted from [120].

stages: the Level-1 Trigger (L1), based on custom electronics, and the High-Level Trigger (HLT) system, relying upon commercial processors.

Level-1 Trigger

The Level-1 Trigger (L1) operates at the hardware level to perform an accept-reject decision that reduces the rate of information down to 100 kHz. Since the total incoming rate is too high, the data are memorised in local buffers for a maximum of 3.2 µs, a limit imposed by the storage capacity of the tracker and pre-shower front-end. Due to the latency of signal transmission, the actual data processing time is below 1 µs. This restriction means that the L1 trigger system can only process data from the calorimeters and muon chambers.

The L1 system is organised into three major subsystems: Calorimeter, Muon, and Global triggers. The original Run 1 design for the L1 trigger was built to read the information with coarse granularity, process it from each subdetector separately and combine the output only in a later stage. To improve the performance and sustain the increased PU rate, the Run 2 upgrade design [120] exploits a finer granularity and the natural synergy between subdetectors at an early stage. A graphic representation of the Run 2 design of the L1 trigger system is depicted in figure 2.2.10.

The Calorimeter trigger begins with the measured energy deposited in the ECAL, HCAL and HF towers, which are distributed by the Layer-1 to the Layer-2 of the Calorimeter trigger. The data from the calorimeters are used to find electrons and photons[4], τ leptons, and jets candidates. Subsequently, the Layer-2 calculates the total transverse energy and missing energy vector and provides an η-ϕ grid of calorimetry deposits to the Global Muon Trigger for muon isolation requirements.

In the Muon trigger, the information from the muon system is used to reconstruct muon tracks. The Muon Track Finder is split into three parts, each covering different η regions (cf. section 2.2.4): the barrel region up to $|\eta| < 0.83$ receives data from DTs and RPCs; the endcap region for $|\eta| > 1.24$ analyses data from RPCs and CSCs; the overlap region for $0.83 < |\eta| < 1.24$ combine the data from all the three subdetectors.

[4] Electrons and photon cannot be distinguished at L1 level due to the lack of tracking information.

All the reconstructed information is collected by the Global Muon Trigger, which determine the muon isolation and remove duplicates in case of ambiguity caused by two nearby muons.

A maximum of 4 of each type of reconstructed particles is selected from the Calorimeter and Muon triggers, based on their reconstruction quality and p_{T} values. Finally, at the third level, the Global Trigger (GT) selects the events that pass some predefined criteria. The decision to accept a given event is taken if the event satisfies all the requirements of at least one of the GT algorithms. The GT may execute in parallel up to 130 algorithms, from single variable thresholds to sophisticated algorithms. If the event is accepted, the entire detector information is read out and transferred to the HLT.

High-Level Trigger

The High-Level Trigger (HLT) is designed to reduce the data acquisition rate to match the capabilities of the mass storage and offline computing systems ($\sim 100\,\mathrm{Hz}$). Events accepted by the L1 trigger are read out using the complete detector information, including input from the tracker and the full granularity of the calorimeters, which is not available on the time scale of the L1 trigger decision. Furthermore, a series of filters are applied progressively to optimise the data flow and avoid the saturation of the system bandwidth.

The event selection requires a decision made on a more accurate (local) event reconstruction, with a complexity similar to the offline one. Therefore, a simplified particle flow algorithm (see section 3.2) is performed, including jet and τ lepton clustering, Monte Carlo-only-based jet energy corrections (see section 3.4) and b tagging algorithms (see section 3.3.2).

The data processing of the HLT is structured around the concept of a *Path* and *Menu*. An HLT Path is a set of algorithms and filters run in a predefined order and connected by logical *and*. An HLT Menu represents the sum of logical *or* of trigger paths that determines whether to reject or store an event. The accepted events are classified into a Primary Dataset that indicates the reason for their selection.

Although the entire detector readout is available at HLT, to minimise CPU time, the structure of a trigger path is such that the information that can be reconstructed quickly is produced first and used to reduce the data rate for the successive filter. For example, this allows the usage of the complete track reconstruction only for the events that cannot be rejected before using information from the calorimeters or the faster pixel-only track reconstruction. Not to supersaturate the trigger bandwidth, a potential random “prescale” factor can be used, already at the L1 trigger, to reduce the amount of stored data [121].

2.2.7 Computing system

In order to store, process and analyse the data recorded in pp collisions and the simulated ones, the CMS experiment takes advantage of the Worldwide LHC Computing Grid (WLCG), a distributed system of computing services and resources spread all over the world and linked with a high-speed network [122]. The WLCG is composed of four *Tiers*, each made up of several computer centres and providing a specific set of services. Furthermore, the CMS experiment uses a number of formats for data storage with diverse size, reprocessing frequency and content.

Tier system

Tier-0 (T0) is the CERN data centre. All of the LHC data passes through this central hub, which provides less than 20% of the total computing capacity. The T0 workflow is as follows:

- accept RAW data from the TriDAS;
- store the data into Primary Datasets based on trigger information;
- distribute RAW data among Tier-1 sites for backup purposes;
- perform the Prompt calibration and reconstruction and store the output into the RECO and AOD formats;
- distribute the RECO datasets among Tier-1 centres to match for each RAW the corresponding RECO;
- distribute AOD to all Tier-1 centres;

There is a set of 15 Tier-1 (T1) sites, which are round-the-clock support computer centres sited in CMS collaborating countries (large national labs, e.g. INFN, KIT and FNAL). Each T1 centre:

- supplies a secure second copy of a subset of the RAW data;
- provides CPU power for the reconstruction procedure and the data analysis;
- stores an entire copy of the AOD;
- provides storage and redistribution for RECO, AOD and simulated events generated by the Tier-2;

There are roughly fifty Tier-2 (T2) sites around the world. They are typically universities and other scientific institutes that can store sufficient data and provide adequate computing power for any specific analysis tasks for the whole collaboration. T2s also provide the generation of simulated events.

Individual scientists can access the stored data through local computing resources (Tier-3), consisting of local clusters in a university department or even individual computers, although there is no formal engagement between WLCG and Tier-3.

Data flow

The CMS data flow includes several steps as shown in figure 2.2.11. The data collected by the CMS detector and that fired the HLT is referred to as the RAW format (average size of 1 MB/event). The collision events to which reconstruction and identification algorithms are applied are known as *Reco* datasets (average size of 2 MB/event). This procedure usually happens shortly after the data is collected by the detector (*Prompt reco*), another time at the end of the yearly data-taking period (*Rereco*), and once more during the long shutdown periods[5] (*Legacy*). The latter constitutes the legacy of the collaboration in terms of the best possible calibration of each subdetector to reach the ultimate performance.

[5] Long shutdown periods coincides with the end of Run 1 and Run 2.

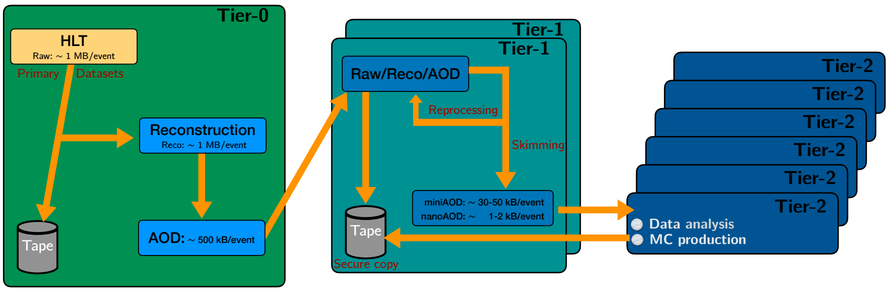


Figure 2.2.11: Data flow for CMS. Adapted from [132].

The datasets corresponding to the Monte Carlo (MC) simulation are GEN, SIM and DIGI (average size of 2 MB/event). GEN indicates the generation of physical processes using simulation tools like PYTHIA8 [123], POWHEG [124–126], HERWIG++ [127] and MADGRAPH5_AMC@NLO [128]. SIM indicates the simulation, where the interaction of the particles in all CMS detectors and their responses are described through the software GEANT4 [129–131]. Finally, DIGI indicates the digitisation of the electronic response. A detailed description of each step is given in section 3.6.

In order to provide event content in a convenient format usable directly by physics analyses, another format is derived for both real data and simulated samples. The Analysis Object Data (AOD) format contains high-level physics objects (such as reconstructed leptons and jets), together with a summary of the RECO information sufficient to support track refitting and calorimeter energy reclustering for analysis-specific needs. The removal of the detector hit information allows a size-reduction down to 400-500 kB/event.

The information needed by a large set of the analyses performed in CMS can be condensed on high-level quantities, like objects kinematic and identification properties, with little need for lower-level detail. Further reduction in size and increase in processing speed can therefore be achieved. Two new formats are designed by the CMS Collaboration to reduce the event size by one (miniAOD [133]) and two (nanoAOD [132]) orders of magnitude with respect to the AOD format, respectively.

The main difference between these two formats lies in the flexibility to recompute jet clustering, b tagging and substructure observables, together with the lepton and photon recalibration. The miniAOD format contains all individual particles rather than only the reconstructed physics objects. Furthermore, the nanoAOD format has been tailored to keep a simple structure and only a limited number of high-level physics objects with increased p_{T} thresholds. The average size of miniAOD and nanoAOD formats are about 30-50 kB/event and 1-2 kB/event, respectively.

3

Event reconstruction and simulation

The granular structure of the CMS apparatus provides a solid baseline to employ the particle reconstruction based on the Particle-Flow algorithm. This section briefly describes how the detector information can be used first to reconstruct high-level objects, like tracks, vertices and calorimeter clusters, and subsequently to build physical objects, namely leptons and jets. A detailed description of the jet calibration procedure is given, as closely related to the work presented in this thesis. Besides, high-energy physics data analysis relies heavily on a precise modelling of the SM prediction as well as hypothetical BSM signals via event simulation. A short description of the key features of the event simulation is given at the end of this chapter.

3.1 High-level object reconstruction

The event reconstruction in the CMS experiment uses a similar scheme for both the online and offline reconstruction, with some difference in the amount of information used and level of corrections applied to achieve fast performance during data-taking. This distinction arises from the online compelling time requirements, as described in section 2.2.6. Once the data are stored, the offline software can exploit the full detector readout. The information from each sub-detector is combined and progressively used to identify the particles produced in each event. This process is based on the Particle-Flow (PF) algorithm [134], which, starting from a local detector reconstruction, iteratively combines the signals from all subdetectors to reconstruct high-level objects; they include tracks, the trajectories of charged particles inside the tracker and muon systems, calorimeter clusters, the energy deposits in the ECAL and HCAL, and multiple pp interaction vertices. A brief description of the algorithms used to reconstruct these high-level objects is presented in the following.

3.1.1 Tracks

The first step of the reconstruction procedure consists of processing the inner tracking system readout to build hits. A tracker hit is the best estimate of a charged particle's impact position into a silicon pixel or strip, derived from the charge distribution released inside the material. From those hits, it is possible to estimate the trajectory and the momentum of a charged particle. Thanks to the approximately uniform magnetic field in the tracking region, a charged particle moves along a hel-

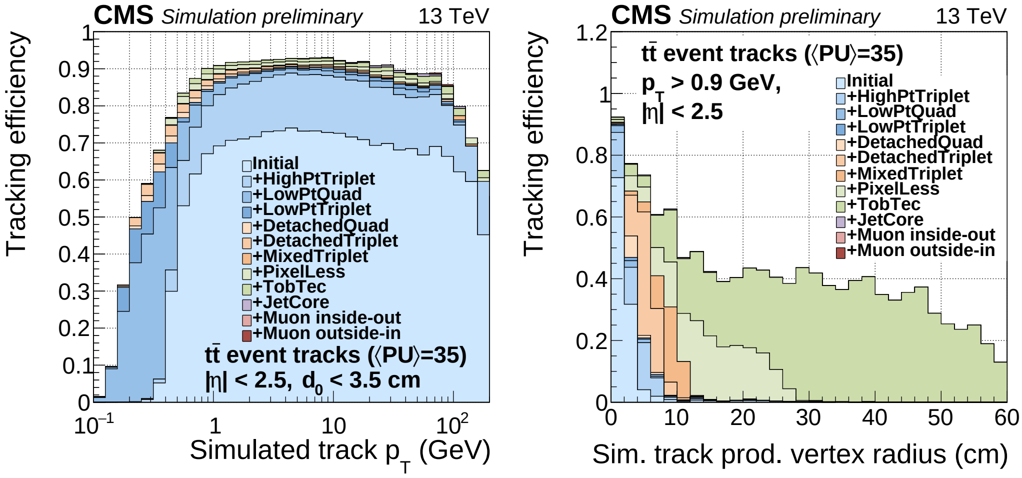


Figure 3.1.1: Tracking efficiency for simulated $\mathrm{t\bar{t}}$ events as a function of p_T (left) and vertex position (right) for the Phase-1 tracker. Different iterations are shown in different colours, as detailed in the text. Taken from Ref. [138].

ical path described by five parameters: the p_T of the particle, the polar angle θ with respect to the z-axis, and three additional parameters to define the particle's initial position, namely the distance from the beam spot along the z-axis (d_z), the distance from the beam-line in the transverse plane (d_{xy}), and the azimuthal angle (ϕ).

The CMS track reconstruction algorithm [135] is a pattern recognition algorithm based on the combinatorial Kalman filter [136] and improved by the “Cellular Automaton” (CA) technique [137]. It is an iterative procedure that starts from tracks that are easier to find, i.e. prompt and with relative high p_T, and progressively moves to more complex cases. Hits associated with high-quality reconstructed tracks are then masked to reduce the combinatorics and simplify the next iterations with a complex topology (low-p_T range or with displaced vertices). Figure 3.1.1 shows the improvement brought by each iteration represented by different colours: prompt tracks are reconstructed first in different p_T ranges (blue), followed by tracks belonging to displaced vertices (orange and green); ultimately, special iterations targeting environments with high-density tracks (jets, in purple) and using information from the muon subdetectors (red) are performed. The main cause of the tracking inefficiency is due to hadrons, which undergo elastic and inelastic nuclear reactions, and electrons, which lose energy through bremsstrahlung. Conversely, the tracking efficiency for isolated muons is much higher, and the efficiency measured for muons originating from $\mathrm{Z} \to \mu\mu$ events is above 99% [135].

A comparison of the tracking reconstruction performance in the Phase-0 (2016) and Phase-1 (2017) detectors is shown in figure 3.1.2 for various quantities. The misidentification rate (fraction of reconstructed tracks that are not associated to any simulated particle) and the p_T and position resolutions are chosen as an example; nonetheless, the tracking efficiency, the speed of the reconstruction algorithm and resolution of all the tracking parameters exhibit also a gain in performance. An overall improvement is observed across the entire η range, in particular in the transition region ($1.2 < |\eta| < 1.6$) and in the endcap region ($2.4 < |\eta| < 3.0$). Moreover, the improvement of the impact parameter resolution played a crucial role for the $\mathrm{H} \to \mathrm{b\bar{b}}$ analyses (see section 1.2.3) in improving the identification of b quark-initiated jets.

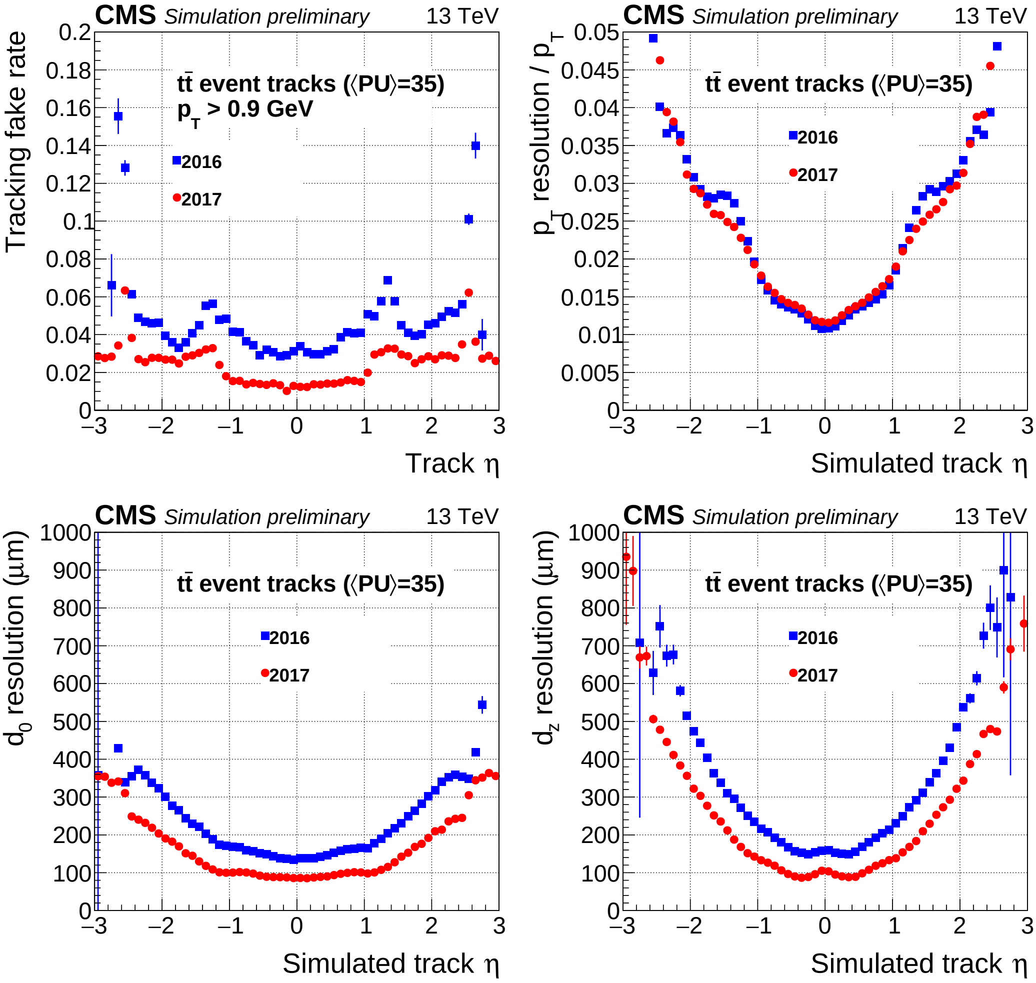


Figure 3.1.2: Fake rate (upper left), p_{T} resolution (upper right) and transverse (lower left) and longitudinal (lower right) position resolution for Phase-0 and Phase-1 trackers in simulated $t\bar{t}$ events as a function of the pseudorapidity of the track. Taken from Ref. [138].

3.1.2 Primary vertices

At the LHC, multiple pp collisions occur when proton bunches collide; nonetheless, most likely only one of these collisions, referred to as the primary hard interaction, produces a hard-scattering process [97]. A precise reconstruction of the positions of all pp interactions (primary vertices) in the event allows for an efficient identification of the primary hard interaction products from all the objects that originated from other collisions, usually called pileup (PU) interactions. Besides, the accurate position measurement of the primary vertex (PV) and secondary, displaced, vertices (SVs) plays an essential role in the correct identification of physics objects, e.g. jets initiated by heavy quarks (see section 3.3.2) or photon conversion (see section 3.2.2).

Reconstructed tracks, with high-purity quality criteria [135] and no significant displacement from the beam spot, are used to measure the position of the PVs [135]. The selected tracks are initially split into several clusters based on a possible common vertex of origin, using the Deterministic Annealing algorithm [139]. Afterwards, the Adaptive Vertex Fitter algorithm [140] is used to fit each cluster to compute

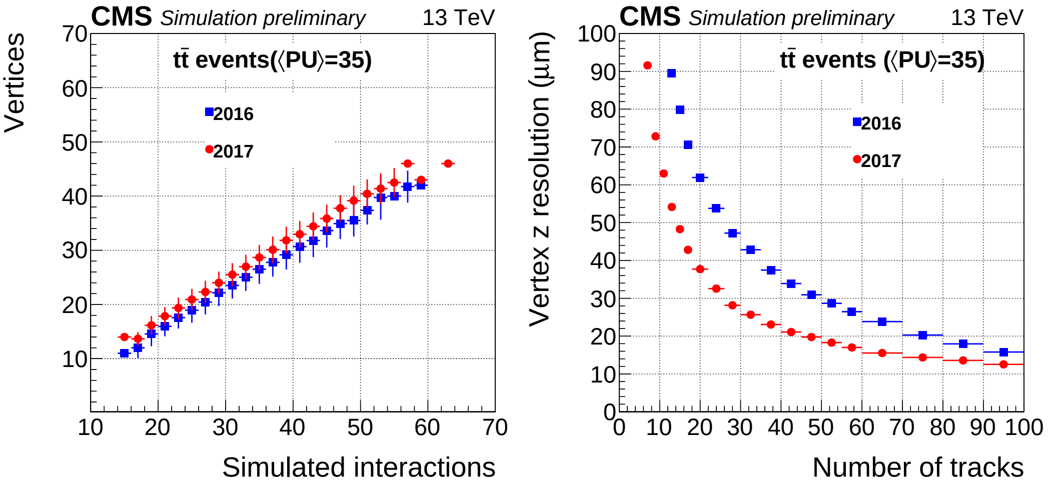


Figure 3.1.3: Comparison of PV reconstruction performace between Phase-0 and Phase-1 tracker in simulated $t\bar{t}$ events. Left: the number of reconstructed vertices as a function of the number of simulated interactions. Right: longitudinal resolution as a function of the number of tracks at the fitted vertex. Taken from Ref. [138].

the best estimate of vertex parameters (x, y and z positions) and their covariance matrix. In this process, each track is assigned a weight w_i between 0 and 1, which reflects the likelihood that it belongs to the given vertex. These probabilities are used to determine the fit quality given by the number of degrees of freedom, defined as $\mathrm{ndof} = -3 + 2 \sum_{\mathrm{tracks}} \mathrm{w_i}$. The fit is iterative, which means that the tracks are reweighted at each iteration so that the contribution of fake tracks gradually diminishes until the fit itself converges. Among the fitted PVs, the one with the highest p_{T}^2 sum of all the collision products is selected as a candidate for the leading PV[1]. The collision products are defined as the jets reconstructed with the PF algorithm (see section 3.3) and the resulting missing transverse energy (see section 3.5). The PV reconstruction efficiency and longitudinal resolution are reported in figure 3.1.3.

When a hadron interacts in the tracker material, many secondary charged and neutral secondary particles are often created. Therefore, SVs can be identified from such displaced tracks and the corresponding particles reconstructed by the PF algorithm, as explained in the following sections.

3.1.3 Energy clusters

As described in section 2.2.2, the energy of a particle is collected by the ECAL and HCAL cells. The cell energies are assembled and used as input to the jet clustering algorithms (see section 3.3). The energy clustering procedure allows for a better estimation of the energy and direction of stable neutral particles, the reconstruction of bremsstrahlung photons emitted by electrons, the separation of neutral from charged particles energy deposits, as detailed in the following.

The clustering is performed separately for each subdetector: ECAL and HCAL, barrel, endcaps and preshower. The first step consists of the identification of the cluster seeds, defined as the cells with energy larger than the neighbouring ones.

[1] During Run 1, the leading PV was identified using the sum of p_{T}^2 of its associated tracks.

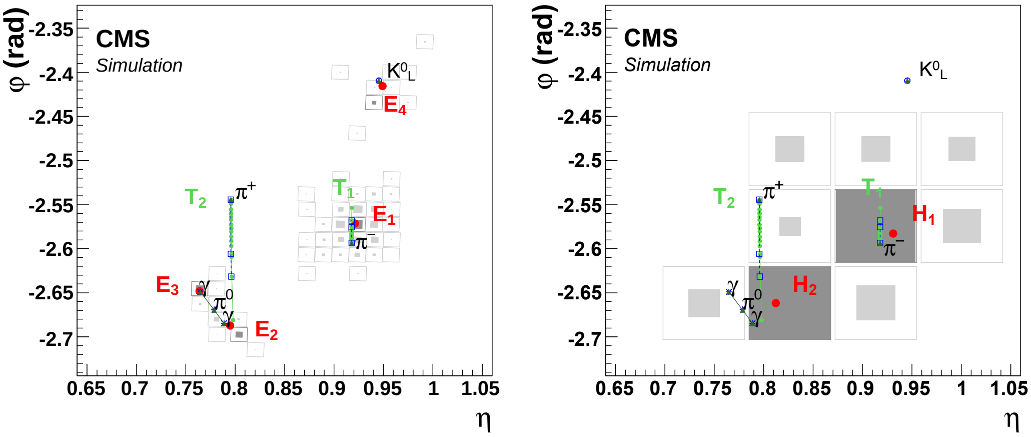


Figure 3.1.4: Event display in the η-ϕ plane for the ECAL (left) and HCAL (right). The calorimeter cells are represented as squares, with a grey-coloured area proportional to the logarithm of the cell energy and the dark grey symbolising the cluster seeds. Four well-separated ECAL clusters are reconstructed from the π^-, γ and K^0_L energy deposits (E_1, E_2, E_3, E_4), while the π^+ does not create a cluster in the ECAL. The two charged pions are reconstructed as tracks (T_1 and T_2), depicted as green lines and pointing towards the HCAL clusters (blue open markers). Taken from Ref. [134].

Afterwards, a topological clustering is performed around each seed by aggregating adjacent cells with energy exceeding the sum of the individual cell's threshold and twice the noise level. The selected cells are fitted with a multi-Gaussian profile to extract the best estimation of the η-ϕ position and the fraction of the energy deposit to be assigned to each cluster in case of overlap. The energy values are further corrected to account for threshold effects, misreconstruction and miscalibration, especially for calorimeter clusters that are not matched to any charged particle track and low-p_T particles [134].

An example of the energy cluster reconstruction process is shown in figure 3.1.4. Two cluster seeds are identified in the HCAL, and two red points indicate the fitted position of the topological clusters. Similarly, the ECAL topological clusters are shown with finer granularity, allowing the distinction of two close-by photons arising from the π^0 decay.

3.2 Particle-Flow reconstruction

A particle produced in a pp collision and crossing the CMS detector leaves traces in various subdetectors, depending on the nature of the particle. A schematic representation of the typical detector signature for different particles is shown in figure 3.2.1.

The CMS Particle-Flow (PF) [134] event reconstruction links the individual detector objects (charged-particle tracks, calorimeter clusters, and muon-system hits) to reconstruct all stable particles in the event and measure their properties.

First, all possible pairs of base elements, restricted to the nearest neighbours in the η-ϕ plane to reduce computing time, are connected by a geometrical link to form a block (PF block). The distance between any two elements is used to quantify the

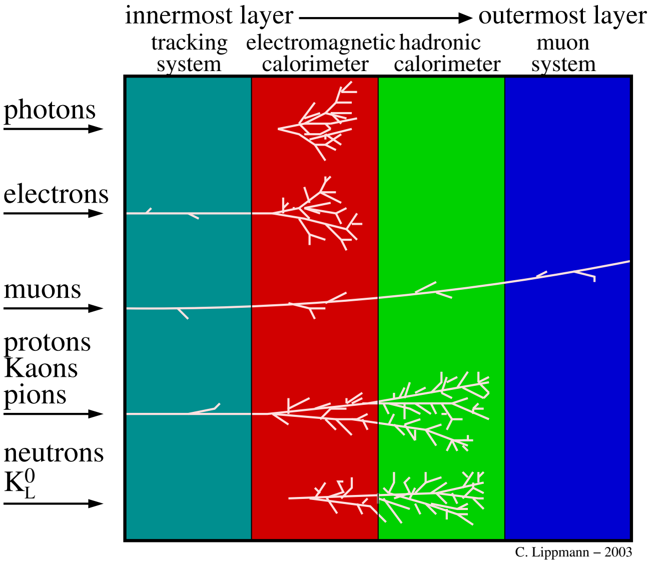


Figure 3.2.1: Typical detector signature for different particles. Taken from Ref. [102].

link quality, and each element can belong to multiple blocks. Eventually, all linked elements are combined to identify the particle candidate (PF candidate).

Several types of links between PF elements are possible. A track is first extrapolated from its last measured hit in the tracker to the calorimeter systems. The track is linked to a cluster if its extrapolated position is within the cluster area. Links between calorimeter clusters are sought between HCAL and ECAL, or between ECAL preshower, inside the preshower acceptance. Tracks can be linked together in case they are coming from a common SV; this happens if the SV has at least three tracks and at most one of them is associated to a PV, and the invariant mass of the outgoing tracks exceeds 0.2 GeV. Finally, links between tracks and hits in the muon system are made.

The identification proceeds in the following order for each PF block, and as soon as a PF candidate is identified, its corresponding PF elements are removed for the following steps. Muon candidates are identified first. Then, electrons and isolated photons are reconstructed at the same time. The remaining elements in the block are then categorised as either charged or neutral hadrons. Lastly, when all the objects are identified, a post-processing step is performed to account for particle misidentification and misreconstruction, especially in events with an artificially large missing transverse momentum: the high-p_T particles are reconstructed and identified with more quality criteria if the missing transverse momentum is drastically reduced [134].

3.2.1 Muons

The muon reconstruction is the first step in the PF algorithm. The muon spectrometer, which provides an almost unambiguous identification over the entire detector acceptance, and the HO calorimeter (see section 2.2.2), which is designed to absorb and contain the hadronic showers, guarantee a very efficient muon identification over a broad p_T range.

Muon tracks are reconstructed independently in the tracker and muon system and then used for the muon object reconstruction [141]; the muon collection comprises

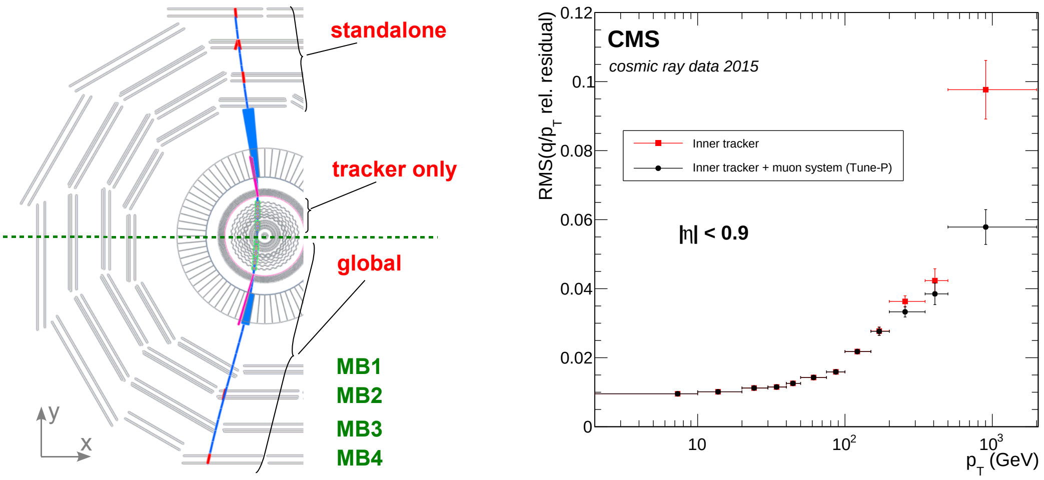


Figure 3.2.2: Left: Cosmic muon event reconstructed in the upper and lower side of the detector. Taken from Ref. [142]. Right: RMS of the relative difference of q/p_{T} as a function of p_{T} for cosmic rays recorded in 2015, using the inner tracker fit only (red squares) and including the muon system (black circles). The parameter q is the charge of the muon. The Tune-P algorithm refit the tracks to reach better performance. Taken from Ref. [141].

Standalone, Tracker, or Global muons, as illustrated in figure 3.2.2 (left). Standalone muons are reconstructed using the muon system tracks exclusively. Tracker muons are reconstructed with an inside-out approach; all tracker tracks with $p_T \geq 0.5\,\mathrm{GeV}$ and total momentum $|\vec{p}| \geq 2.5\,\mathrm{GeV}$ are extrapolated to the muon system, and, if at least one match with DT or CSC segments is present, the corresponding tracker track is qualified as a Tracker Muon. Conversely, the Global Muon reconstruction uses an outside-in perspective; for each Standalone muon, the two tracks are combined using the Kalman-filter technique if it is matched to a tracker track. The resulting global muon shows an improved momentum resolution compared to the tracker-only fit (see figure 3.2.2 (right)), mostly at large transverse momentum ($p_T \geq 200\,\mathrm{GeV}$).

About 99% of muons produced in pp collisions within the geometrical acceptance of the muon system and having sufficiently high momentum are reconstructed with at least one of the methods above. Muons reconstructed only as Standalone muon tracks have worse momentum resolution and higher admixture of cosmic-ray muons than the Global and Tracker Muons and are usually not used in physics analyses.

Muons with $p_{\mathrm{T}} \geq 100\,\mathrm{GeV}$ are reconstructed with a resolution of approximately 1% in the barrel and 3% in the endcap. The resolution is measured with cosmic ray data (see figure 3.2.2 (right)), using the relative difference in q/p_{T} between the upper and the lower halves of the detector, where q is the muon charge.

The reconstructed muon candidates are also required to pass an additional set of identification (ID) criteria. There are multiple IDs in use by the CMS Collaboration, with various efficiencies and misidentification rates. The efficiency is defined as the number of reconstructed muons divided by the number of expected muons, while the fake rate is the rate of particles misidentified as muons.

A "loose" ID defines a PF muon reconstructed either as a global or tracker muon. This ID has the highest efficiency ($\sim 99.7\%$) and targets all types of muons. The loose ID with additional track-quality and muon-quality requirements is defined as "medium" or "tight", depending on the criteria applied; both IDs have lower efficiencies, $\sim 98.5\%$ and $\sim 97\%$, respectively. The criteria of the "tight" ID are defined to suppress punch-through charged hadrons and muons produced in flight. Therefore, the "tight" muon identification is specialised in prompt muons, while the "medium" identification is also used to identify muons coming from heavy-flavour hadron decays. A dedicated ID ("soft" ID) is derived with the focus on low-p_{T} muons, not using PF muons but rather muon tracks with specific quality criteria.

Other specific IDs of particular interest for this analysis are developed for the high p_{T} scenario, the "high-p_{T}" and "tracker high-p_{T}" IDs, respectively. The "high-p_{T}" ID imposes the following quality requirements:

- The muon must be reconstructed as a Global Muon.
- The muon must have transverse and longitudinal distance from the PV of $d_{xy} < 0.2\,\mathrm{cm}$ and $d_z < 0.5\,\mathrm{cm}$, respectively. These requirements suppress cosmic muons while preserving the selection efficiency for muons from decays of b and c hadrons.
- The tracker muon must have at least 6 hits in the tracker, of which at least 1 in the PIXEL, to guarantee a good p_{T} measurement.
- The muon must have hits in at least two muon stations, to suppress punch-through hadrons and muons from decays in flight.
- If only one matched station is present, to exclude reconstruction failure due to the geometrical layout of the detector, at least one of the following conditions must be satisfied:
  - the single matched station must not be the first muon station.
  - the extrapolation of the inner track to the muon system must have at most 1 matched station.
  - at least two matched RPC layers must be present.
- The muon must have a relative error on the transverse momentum measurement ($\sigma_{p_{\mathrm{T}}}/p_{\mathrm{T}}$) of less than 30%.

A slight variation of the previous ID is the "tracker high-p_{T}" ID. It is designed to improve the reconstruction efficiency of muons originating from the decay of boosted Z boson, in which two close-by high-p_{T} muons appear. The requirements are similar to the ones for the "high-p_{T}" ID, but in this case, the muon is required to be a Tracker muon instead of being a Global Muon; furthermore, the conditions on the muon system are no longer a requirement; as a consequence, the momentum resolution is degraded with respect to the previous ID, with the benefit of increased efficiency for close-by muon pairs. Both the "high-p_{T}" and the "tracker high-p_{T}" are used in the context of the analysis presented in this thesis, as explained in chapter 6.

3.2.2 Electrons and photons

A distinctive characteristic of electrons and photons is that they deposit almost all of their energy in the ECAL. In contrast, only electrons leave traces in the tracker layers (see figure 3.2.1). Moreover, the interaction of an electron or a photon with the material in front of the ECAL results in bremsstrahlung photons ($\mathrm{e}^{\pm} \to \gamma\mathrm{e}^{\pm}$), or photon conversion ($\gamma \rightarrow e^+e^-$), respectively. In the attempt of recovering all the distinguishing features of the resulting multi-particle showers, the basic principle of the electron and photon reconstruction [143] relies on the combination of tracker tracks and ECAL deposits.

The first step of the reconstruction is the combination of multiple ECAL clusters into a single supercluster (SC) with an enlarged ϕ window to account for the azimuthal bending of an electron in the magnetic field. A dedicated tracking algorithm, based on the Gaussian Sum Filter (GSF) [144], is then used to re-estimate the electron track parameters to account for emitted bremsstrahlung photons and changes in the trajectory. All reconstructed tracks are tested for compatibility with the electron-only or the photon converting into e^+e^- pair hypotheses.

Finally, ECAL clusters, SCs, GSF tracks associated with electrons and conversion tracks are linked by the PF algorithm into blocks. These blocks are labelled as electrons or photons depending on their origin, a GSF track or a SC, respectively. Additional basic criteria are imposed to reduce the probability that hadrons are wrongly reconstructed as electrons or photons; these criteria comprise shower-shape and tracker-related variables as well as conditions on the energy deposits in the HCAL clusters close to the SC.

Unlike muons, whose charge is unambiguously measured, the electron charge measurement is more challenging due to the multitude of particles to reconstruct. The charge can be extracted from the curvature of the track fitted with the GSF algorithm or from the original Kalman filter track. An alternative method is based on the sign of the angle in the transverse plane between the relative position of the SC and the GSF track. With a misidentification probability of approximately 1.5%, the final electron charge is chosen as the one obtained from at least two of these three estimates.

The electron and photon energy is measured using both the calibrated energy of ECAL clusters and the momentum of the GSF track, which are combined with weights derived from a multivariate regression with a boosted decision tree (BDT) algorithm [145]. The final energy resolution, driven mainly by the energy resolution of the SC, is better than 2% in the barrel and 5% in the endcap for $p_T > 20\,\text{GeV}$ [143].

The photon reconstruction from a SCs, including the very loose selection criteria, is assumed to be 100% efficient. On the other hand, the electron reconstruction efficiency, reported in figure 3.2.3 (left), is above 95% and compatible between data and simulation within 2%.

Similarly to the muon case, the reconstructed electrons and photons are usually required to fulfil some identification criteria. Two different techniques are used by the CMS Collaboration; the first is based on one-dimensional cuts on several variables (cut-based), and the second is based on a BDT discriminant. Although the latter has slightly better performance, the former provides more flexibility for physics analyses to perform sidebands studies. Figure 3.2.3 (right) shows the performance of the electron BDT-based ID for different isolation requirements. As for the muon case, different working points (WPs) are defined: “loose”, “medium” and “tight” with an efficiency of about 90%, 80% and 70%, respectively. A modified version (HEEP) of the cut-based ID is provided for high-p_T electrons, where the main difference lies in using the subdetector-based isolation instead of the PF isolation. Dedicated studies on the performance of the electron IDs for boosted $Z \rightarrow ee$ decays in the context of the analysis presented in this thesis are reported in chapter 6.

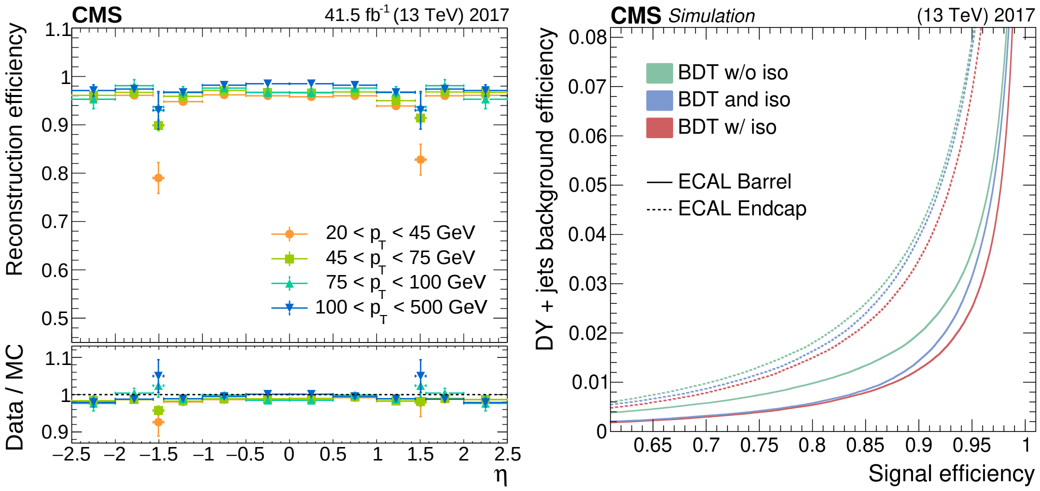


Figure 3.2.3: Left: Data-to-simulation (Data/MC) comparison of the electron reconstruction efficiency versus η for the 2017 data taking period. The drop in efficiency at $1.44 < |\eta| < 1.57$ is caused by the transition between the barrel and endcap in the ECAL. Right: Performance of the electron BDT-based identification algorithm with (red) and without (green) isolation variables; the blue curve refers to the same algorithm as in the green one with the isolation cuts applied on top. Taken from Ref. [143].

3.2.3 Hadrons and non-isolated photons

After the identification of muons, electrons and isolated photons, the remaining tracks and calorimeter clusters are used to reconstruct charged ($\pi^{\pm}$, $K^{\pm}$, or protons) and neutral (K^0, non-isolated γ, or neutrons) particles resulting from the hadronisation process.

The hadron reconstruction and calibration operate as described in the following. The calorimeter clusters that are not linked to any track are reconstructed as photons or neutral hadrons; otherwise, they are classified as charged hadrons. Inside the tracker acceptance and with no associated tracks, ECAL clusters that are not matched to any HCAL are reconstructed as photons; for all the other cases, the resulting PF candidate gives rise to a neutral hadron. Photons are given precedence over neutral hadrons because they carry 25% of the jet energy, compared to the 10% brought by neutral hadrons; this assumption will be clarified in more detail in section 3.4.5. Beyond the tracker acceptance, however, charged and neutral hadrons cannot be distinguished, and, therefore, precedence is given to the hadrons. In the HF, a distinction between hadron and electromagnetic deposits is possible from the shower shape information.

The energy of neutral particles is measured directly from the calorimeter clusters. For charged hadrons, the sum of the momentum of all the associated tracks is compared to the calorimetric energy to infer the potential presence of additional neutral particles. If the calibrated calorimetric energy exceeds the sum of the track momenta, the excess is attributed to additional PF candidates, identified as photons or neutral hadrons in case of an ECAL or HCAL excess, respectively; otherwise, no other neutral particle is reconstructed.

3.2.4 Tau leptons

The τ lepton is the heaviest of the SM leptons; it has a short lifetime of approximately 2.9×10^{-4} ns [7]; thus, unlike the other leptons, it decays before reaching the active volume of the detector. The leptonic decay modes have a BR of approximately 35% and consists of two ν and either e or μ in the final state; in these cases, only the electron or the muon can be reconstructed, making it difficult to distinguish them from the other electron or muon production processes. In the other ~ 65 % of the cases, which contain one ν_τ and several charged and neutral (mostly π^0) hadrons, the τ is reconstructed as a hadronic tau jet (τ_h) using the Hadron Plus Strips (HPS) algorithm [146].

The τ_h reconstruction starts using a PF jet (see section 3.3) candidate as seed. The first step of the HPS algorithm consists of the reconstruction of the π^0 candidates from PF photons and electrons; they are clustered in a p_T-dependent area in the η-ϕ plane, called "strip". Strips are combined with the remaining charged particles inside the original PF jet, and, finally, their compatibility with the expected signatures from different decay modes is tested.

Several discriminators have been developed within the CMS Collaboration to suppress misidentified τ leptons [146]. The discriminant against muons is based on vetoing τ_h candidates if signals in the muon detector are found in its direction. A BDT discriminator is trained to separate τ_h decays from electrons, aiming at distinguishing the different showers produced. To provide the best possible discrimination between τ_h decays and quark or gluon jets, a cut-based and BDT-based approaches are available. The former uses an isolation variable formed from the p_T of surrounding particles; the latter combines the isolation variable with track-related variables and information of the PF candidates to improve the performance. The identification efficiency for τ leptons with $p_\mathrm{T} > 30\,\mathrm{GeV}$ are between 40% and 60%, depending of the WP.

3.3 Jet reconstruction

QCD processes represent a dominant part of all the processes taking place in pp collisions at the LHC. As mentioned in section 1.1.3, particles carrying a colour charge cannot be observed freely and almost immediately ($\tau_\mathrm{QCD} = 5 \times 10^{-24}$s) undergo fragmentation and hadronisation [7], with the only exception being the top quark; in fact, its sufficiently short lifetime ($\tau_t = 0.3 \times 10^{-24}$s) causes it to decay predominantly into a W boson and a b quark before the hadronisation takes place [7].

Due to the hadronisation process, quarks and gluons produced in pp collisions create collimated showers of hadrons. One of the main challenges of physics analyses is to infer the initial energy, momentum and, possibly, the nature of the parton (gluon or quark) produced in the original interaction. For this reason, the spray of particles is clustered together into what is generally called "jet". There is a large variety of jet clustering algorithms, although they all have in common the infrared and collinear (IRC) safety requirements [147]; this condition allows for a comparison with theory and robustness for soft and collinear emissions.

It is possible to cluster several types of jets, depending on the objects in use. Particle-level (ptcl) jets are clustered from all stable ($c\tau > 1$ cm) and visible particles (excluding neutrinos) in simulated events before the detector simulation takes place.

The exclusion of neutrinos is a convention adopted by the CMS Collaboration, which significantly reduces differences between heavy-flavour (c and b) and light-flavour (u, d and s) and gluon jets [148].

The calorimeter (CALO) jets are reconstructed from energy deposits in the calorimeter towers alone. This relatively simple yet robust approach was the method of choice for CMS analyses using data at 7 TeV; however, with the improved understanding of the detector, the PF reconstruction algorithm proved to be reliable and with better performance. Almost all the CMS analyses of data recorded during Run 2 use PF jets, which are reconstructed by clustering the PF candidates. This definition takes advantage of the tracking information and high granularity of the ECAL to improve performance. An example of the performance of PF jets with respect to CALO jets is given in chapter 5.

3.3.1 Clustering algorithms

In CMS experiment, jets are clustered from PF objects using the anti-k_T clustering algorithm [149], implemented in the FastJet package [150]. The anti-k_T algorithm is a sequential recombination algorithm similar to the k_T and Cambridge/Aachen (CA) algorithms [151, 152].

The three algorithms can be described in the following way. The first step consists of calculating for each PF element i the "distances" d_{iB} and d_{ij}, from the beam (B) and the j-th PF candidate, respectively. These distances are defined as:

$$d_{ij} = \min\left(p_{\mathrm{T},i}^{2n}, p_{\mathrm{T},j}^{2n}\right) \frac{\Delta R_{ij}^2}{R^2}\,, \tag{3.1}$$

$$d_{iB} = p_{\mathrm{T},i}^{2n}\,, \tag{3.2}$$

where $\Delta R_{ij}^2 = (\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2$ and R is a distance parameter that defines the cone radius, typically called "jet radius". The value of n characterises each algorithm: $n = 0$ for the CA algorithm and $n = \pm 1$ for the k_T and anti-k_T algorithms, respectively. The historical approach of the k_T algorithm was based on the idea of inverting the QCD splitting process by combining objects with soft and collinear momenta. The CA algorithm was introduced to improve the performance of jet substructure (see section 3.3.2). Finally, the anti-k_T has been proven to be characterised by circular cone-shaped jets and to be less sensitive to extra or soft radiation.

The clustering proceeds by identifying the smallest of such distances; if d_{iB} is the smallest, the element i is promoted to a reconstructed jet and excluded from further iterations. Otherwise, elements i and j' (defined as the element that minimises d_{ij}) are merged into a new element i', and the algorithm proceeds to the next iteration. The algorithm stops as soon as all jet candidates have been promoted to reconstructed jets.

The functionality of the anti-k_T algorithm can be understood by considering an event with two separated high-energetic (hard) particles and many less energetic (soft) particles. The distance between one of the hard particles and the soft particles is exclusively determined by the transverse momentum of the hard particle and their angular separation. The distance between soft particles will instead be much larger due to their low transverse momentum. As a consequence, soft particles are more likely to be combined with a hard particle instead of being clustered with another

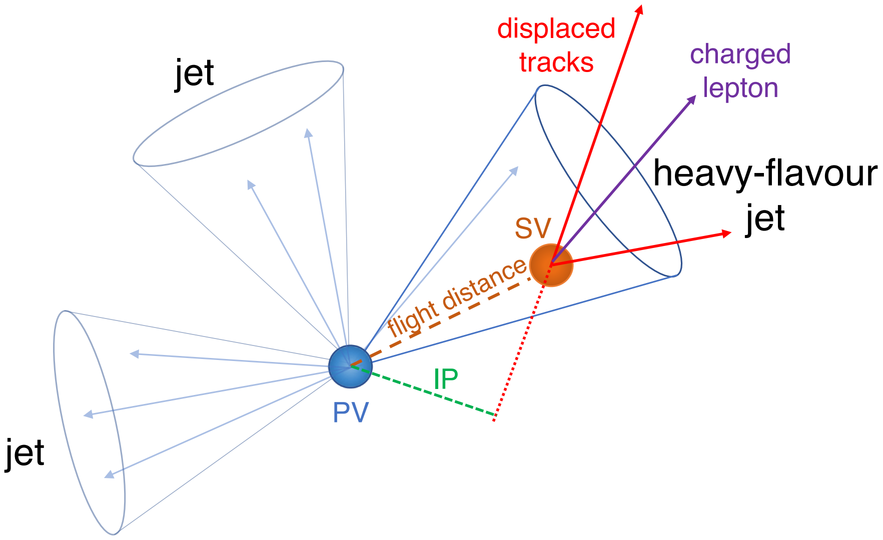


Figure 3.3.1: Simplified representation of a heavy-flavour jet. Taken from Ref. [153].

close-by soft particle, resulting in a conical jet of radius R. On the other hand, if the two hard particles are close but still separated enough to be reconstructed as two jets, none of the two reconstructed jets will have a perfectly conical shape.

A distance parameter of 0.4 has been chosen for CMS standard jets in Run 2, while 0.8 is employed for large-radius jets when looking for boosted heavy particles decaying to hadrons (cf. section 3.3.2). With the anti-k_{T} being the most widely-used jet algorithm in CMS analyses, the small (large-) radius jets are referred to as AK4 (AK8) jets.

3.3.2 Jet tagging and substructure

It is crucial for many CMS analyses to reliably identify the type of particle that initiated a given jet. An accurate classification (tagging) of the jet origin enhances the selection efficiency of events relevant for the final states under consideration. Several algorithms have been developed to discriminate between quarks and gluons and (semi-)hadronic decays of heavy particles (top quark or SM bosons).

Flavour tagging

Heavy-flavour jet identification techniques analyse the properties of the hadrons inside the jet to distinguish between jets originating from light-flavour quarks or gluons (light-flavour jets) and those arising from c or b quarks (heavy-flavour jets).

One of the features most frequently analysed for heavy-flavour discrimination is the relatively large lifetime of heavy quarks (approximately 1.5 ps for b quarks and up to 1 ps for c quarks [7]). These lifetimes give rise to displaced tracks from which a SV may be reconstructed, as illustrated in figure 3.3.1.

Although it is not always possible, the reconstruction of SVs plays an essential role since powerful discriminating variables can be derived from it. For example, the SV mass, which is directly related to the mass of the heavy-flavour hadron, and the distance from the PV, have been found to be particularly effective for flavour tag-

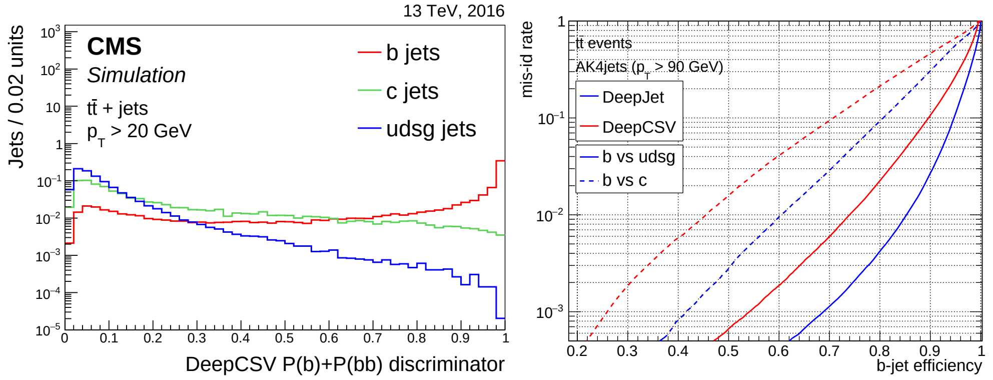


Figure 3.3.2: Left: Distribution of the $P(b) + P(bb)$ probability of the DeepCSV discriminator in $\mathrm{t\bar{t}}$ simulated events. The distributions are normalised to unit area. Taken from Ref. [153]. Right: Performance comparison between the DeepCSV and DeepFlavor, referred to as DeepJet, discriminators for different jet flavours in $\mathrm{t\bar{t}}$ simulated events. Taken from Ref. [154].

ging. For this reason, dedicated algorithms, e.g. the inclusive vertex finding (IVF) algorithm [153], have been developed to reconstruct b hadron decays starting from tracking information. The properties of the reconstructed SV, the IP of charged-particle tracks, and the presence of a soft lepton, or absence thereof, are the main ingredients to several b tagging algorithms, which has been developed and used by the CMS Collaboration during Run 1 and the early Run 2. Modern and more sophisticated approaches combine these variables in complex ML approaches to exploit correlations between variables and create more powerful discriminators.

The discriminator used in many Run 2 analyses is the DeepCSV algorithm [153], developed using a deep neural network (DNN) approach and feed-forward layers (cf. chapter 4). It has a multi-class output structure, which allows to tackle light-, c- and b-flavoured jets separately. An example of the output variables of the DeepCSV algorithm is shown in figure 3.3.2 (left), where $P(x)$ represents the probability of a given jet to be classified as a x-flavoured jet. In particular, the $P(b) + P(bb)$ variable is used in the $\mathrm{Z}' \to \mathrm{ZH}$ analysis presented in this thesis, as further detailed in chapter 6.

Other discriminators are also available and used, for example, the DeepFlavour algorithm [154]; it adds more low-level features from the jet constituents, which are passed into a more complex architecture involving recurrent and convolutional layers. The performance of these two taggers is shown in figure 3.3.2 (right).

Jet substructure

The boosted decays of heavy SM particles, like the top quark or the Higgs boson, result in collimated decay products, which are often clustered into a single large-radius jet. Consequently, it is a challenging task to identify the underlying process, distinguishing QCD-initiated processed, like gluon splitting (g→qq, g→gg) or gluon radiation from a quark (q→gq), from the hadronic decays of SM bosons (e.g. $\mathrm{V} \to \mathrm{qq}$, $\mathrm{H} \to \mathrm{b\bar{b}}$). On the other hand, the reconstructed jet has a distinct substructure, which can be exploited to identify its origin and to create a tool for jet classification. The main algorithms (or taggers), based on different jet substructure techniques and used by CMS analyses to identify specific final states, are described in the following.

The hadronisation processes of quarks and gluons result in different kinematic distributions, which depend on the C_F and C_A factors (see section 1.1.3). Jets originating from gluons (gluon jets) show higher hadron multiplicity, broader angular development, and a softer momentum spectrum than jets originating from quarks (quark jets). These variables are used by the CMS Collaboration to build a likelihood discriminator, used in many analyses; for example, in VBF events, forward jets are most likely quark jets while the background is composed mainly of gluon jets.

A good approximation for boosted two-body decays of a massive particle is that their angular separation is approximately $\Delta R \sim 2M/p_{\mathrm{T}}$, where M and p_{T} are the particle's mass and transverse momentum, respectively. For example, the decay products of a W boson or a top quark can be reconstructed in a single large-radius jet with $R = 0.8$ starting from above transverse momenta of 200 GeV and 400 GeV, respectively. It is common practice to use jets with a large distance parameter (AK8) to encapsulate all the decay products of boosted objects. The drawback of this choice is an increase in extra radiation collected in the jet, which originates from underlying event (UE) and PU. This unwanted contribution can deteriorate reconstructed jet kinematic variables and worsen the resolution of substructure quantities. Jet grooming and PU mitigation techniques have been developed to mitigate these effects. The former is conceived to remove soft and wide-angle radiation [155], and the latter, described in more detail in the following section, is designed to remove the contribution of uncorrelated radiation originating from PU vertices.

The approach adopted by the CMS Collaboration for jet grooming is based on the soft-drop (SD) technique [156]. This algorithm starts from a jet with radius R_0 (typically 0.8 for AK8 jets), whose constituents are reclustered with the CA algorithm; as mentioned in section 3.3.1, the CA algorithm reconstructs jets based only on their angular separation. The SD procedure requires to undo the last stage of the CA clustering to break the jet j into two subjets, $j_{1,2}$. Then, the SD requirement is checked:

$$\frac{\min(p_{\mathrm{T}}(j_1), p_{\mathrm{T}}(j_2))}{p_{\mathrm{T}}(j_1) + p_{\mathrm{T}}(j_2)} > z_{\mathrm{cut}} \left(\frac{\Delta R(j_1, j_2)}{R_0} \right)^{\beta} , \tag{3.3}$$

where z_{cut} and β are free parameters that control the threshold and the angular separation importance in the grooming procedure, respectively. This equation is constructed to reject wide-angle soft radiation. If the subjets pass this condition, then j is the final SD jet, composed by the subjets j_1 and j_2. Otherwise, the subjet with the largest p_{T} is promoted to a jet, and the other subjet is removed. If the jet cannot be declustered any further, then either it can be removed ("tagging mode") or can be left as the final SD jet ("grooming mode"). The CMS Collaboration uses $\beta = 0$ and $z_{\mathrm{cut}} = 0.1$ as parameter values and the grooming mode as default configuration.

A natural choice to distinguish a jet is to use the reconstructed jet mass. In particular, the distribution of the SD mass, calculate from the two subjets returned by the SD algorithm, provides a clear distinction between QCD-initiated jets and jets coming from the hadronic decays of the top quark, Higgs and vector bosons. A complementary approach is provided by other jet substructure variables, which are used to study the angular and energy distributions of jets. Examples of these observables are the energy correlation functions [157] and the N-subjettiness variables [158]. In particular, the ratio between two N-subjettiness variables is used by

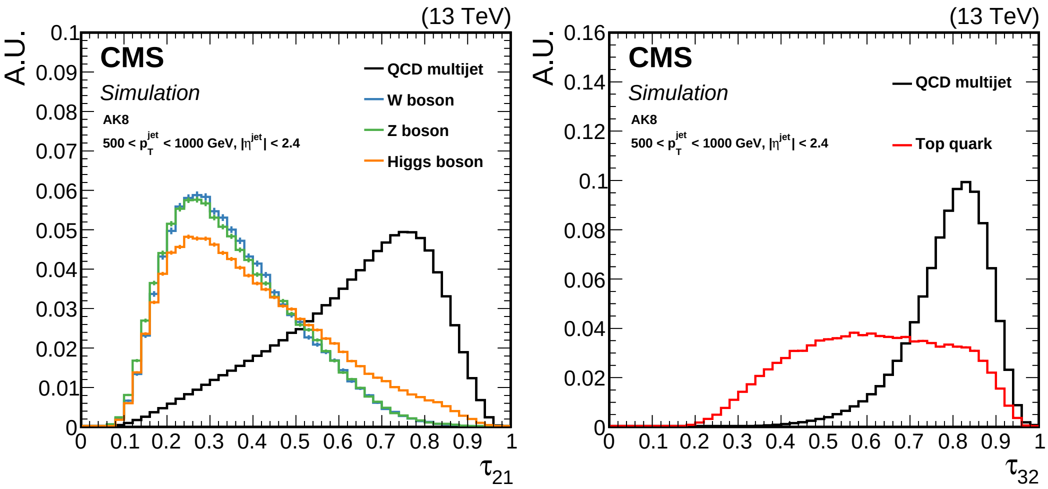


Figure 3.3.3: Shape comparison of the τ_{21} (left) and τ_{32} (right) for AK8 jets for different simulated samples. Taken from Ref. [159].

the CMS Collaboration within the context of dedicated vector boson and top taggers.

The N-subjettiness is a measure for the likelihood of a jet having N or fewer subjets, and it is defined as:

$$\tau_N = \frac{1}{d_0}\sum_k p_{\mathrm{T},k}\min_i\left(\Delta R_{i,k}\right)\,, \qquad \text{with} \quad d_0 = \sum_k p_{\mathrm{T},k} R_0\,. \tag{3.4}$$

Here, k runs over the PF jet constituents and i over the N candidate subjets, which are calculated, forcing the anti-k_T algorithm to return exactly N subjets. If the jet constituents are aligned along N axes, the τ_N is close to 0, and it is more likely for the jet under consideration to have N or fewer subjets. On the contrary, when τ_N is closer to 1, then the jet energy is more widely distributed, making the hypothesis of the jet having more than N subjets more likely.

The ratio of two N-subjettiness variables, which has a high discrimination power and direct theoretical calculability, can be used to test different hypotheses. In particular, τ_{21}, defined as the ratio between τ_2 and τ_1, is used to separate jets originating from boosted bosons, which tend to have two distinct subjets with similar momenta, from quark and gluon jets, which have either a single prong or two prongs with the second being considerably less energetic than the first. Similarly, τ_{32}, defined as the ratio between τ_3 and τ_2, is used for the identification of the three-prong structure typical for the hadronic decay of the top quark. The τ_{21} and τ_{32} variables are shown in figure 3.3.3 for different simulated samples; QCD-initiated jets show a clear separation from all other jets. The N-subjettiness ratios, together with requirement on the SD mass, are often employed in CMS analyses for jet identification [159].

The recent developments in ML allow for many different approaches for jet classification. In fact, ML-based taggers are getting increasingly more used in recent CMS analyses. A more detailed description of these taggers is given in chapter 4.

3.3.3 Pileup mitigation techniques

The instantaneous luminosities reached during data-taking imply multiple pp collisions to occur in the same bunch crossing. Additional particles coming from pileup

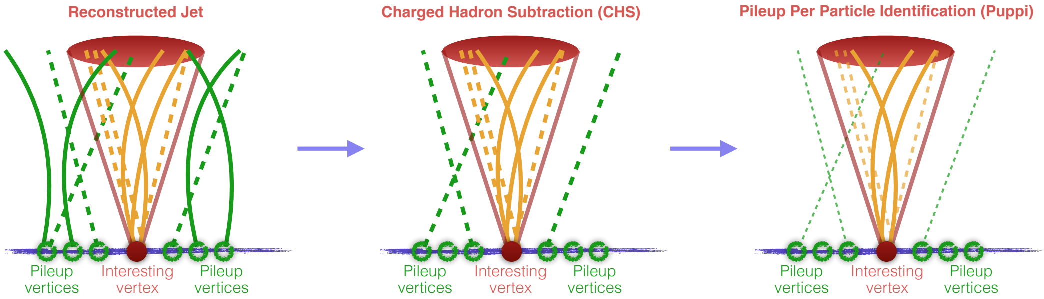


Figure 3.3.4: Sketch of PU suppression techniques. Solid (dashed) lines depict charged (neutral) PF candidates. Thin lines represent down-weighted 4-momenta.

vertices alter jet quantities since they may be clustered inside the jets. Hence, the identification of interesting collisions has become an ever-growing challenge at the LHC. During Run 2, the CMS detector collected data with up to 60 interactions per bunch crossing, with an average pileup of 30 interactions (see figure 2.1.2 (right)). Moreover, the expected average pileup for Run 3 is even higher, making PU rejection an even more challenging task.

The CMS Collaboration uses different techniques for PU mitigation. One example is the charged hadron subtraction (CHS) algorithm [134], widely used in Run 2 analyses. It uses information from the tracker to remove the charged particles that are associated with a PU vertex from the jet clustering procedure. Due to its limited coverage in η, outside the tracker acceptance, no information on the origin of a particle is available; consequently, dedicated jet energy corrections are applied to account for the impact of charged PU outside the tracker coverage, and of neutral PU everywhere (cf. section 3.4). Based on the measured average energy deposited per unit area, this approach has a limited impact since the additional corrections act on the four-momentum and not on the jet shape or substructure. To overcome this limitation, the pileup per particle identification (PUPPI) technique was introduced as an alternative approach for PU mitigation [160]. It calculates, event by event and for each particle, the probability that a given particle originates from the leading PV and scales down the energy of these particles based on that probability. The sketch in figure 3.3.4 depicts the CHS and PUPPI algorithms.

The PUPPI rescaling procedure works in the following way. Prior to the clustering procedure, the algorithm assigns each PF particle a weight $w \in [0, 1]$. Charged particles are treated similarly as in the CHS algorithm; the tracking information is used to decide whether the particle belongs to the leading or to a PU vertex, and, consequently, whether to remove ($w = 0$) or keep ($w = 1$) the PF candidate. Moreover, if a particle is not associated to any vertex, the weight is calculated based on its p_{T}, η and distance from the leading PV [161]. For neutral particles, the computation of the weight exploits the information on the surrounding particles in the following way. As a first step, the variable α is calculated for each particle i as:

$$\alpha_i = \log \sum_{\substack{j \neq i \\ \Delta R_{i,j} < R_0}} \left(\frac{p_{\mathrm{T},j}}{\Delta R_{i,j}} \right)^2 , \tag{3.5}$$

where the sum runs over all particles within a distance of $R_0 = 0.4$ from the particle

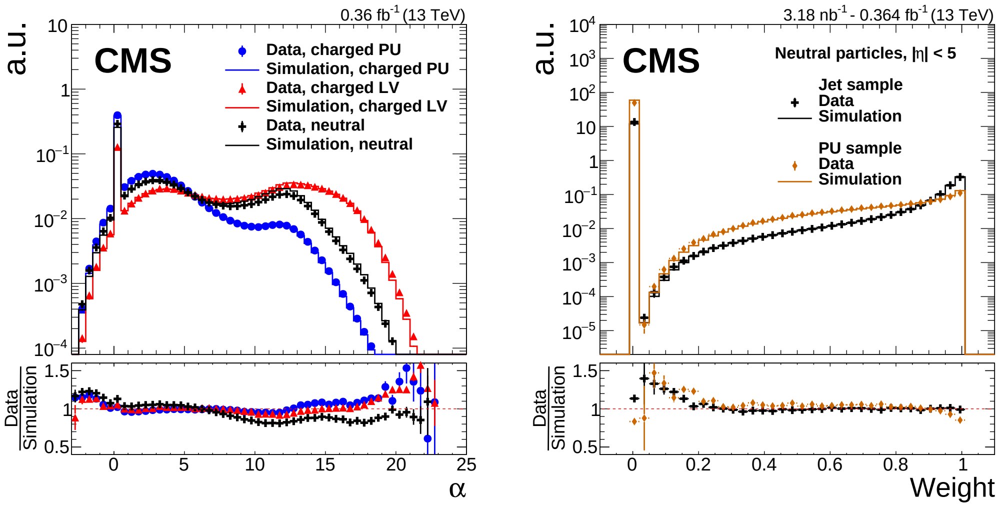


Figure 3.3.5: Data-to-simulation comparison of the α variable (left) and weight for neutral particles (right) for a subset of the 2016 data and QCD multijet events. Taken from Ref. [96].

under investigation, and for $|\eta| < 2.5$ only charged particles are considered.

Under the assumption that charged and neutral particles behave similarly, the α value of neutral particles is compared to the expected value for charged pileup particles. Significant deviations from the expected values correspond to particles from the hard-scattering, while small deviations correspond to pileup particles. The comparison is made using the following metric:

$$\text{signed } \chi_i^2 = \frac{|\alpha_i - \alpha_{\text{PU}}^{\text{mean}}|(\alpha_i - \alpha_{\text{PU}}^{\text{mean}})}{(\alpha_{\text{PU}}^{\text{RMS}})^2}\,, \tag{3.6}$$

where $\alpha_{\text{PU}}^{\text{mean}}$ and $\alpha_{\text{PU}}^{\text{RMS}}$ are extracted, event by event, from the α distribution of charged particles associated with a PU vertex. Given that $\alpha_{\text{PU}}^{\text{mean}}$ and $\alpha_{\text{PU}}^{\text{RMS}}$ are defined only for $|\eta| < 2.5$, for particles outside this region these values are multiplied with transfer factors derived from simulation. The weight for neutral particles is then calculated from the cumulative distribution function of the signed χ^2. The distribution of the α variable and the weight for neutral particles are shown in figure 3.3.5.

To further reduce the noise and the dependence on the number of vertices, weights with low values ($w_i < 0.01$) and weights that satisfy the $w_i p_{\text{T},i} < A + B \cdot N_{\text{vertices}}$ condition are set to 0, where N_{vertices} is the number of vertices in the event, and A and B are tunable parameters.

Another technique specifically targeting low-p_{T} PU jets uses a multivariate approach to reject clustered jets [96, 162]. This PU jet ID is able to reject 95% of PU jets with a minimal efficiency loss. These techniques can be used in a complementary way. This is showcased in figure 3.3.6, which shows efficiency and purity distributions for several algorithms. The efficiency (purity) is defined as the fraction of particle-level (PF) jets with $p_{\text{T}} > 30\,\text{GeV}$ that are matched with a PF (particle-level) jet with $p_{\text{T}} > 20\,\text{GeV}$ within $\Delta R < 0.4$. Inside the tracker acceptance, PUPPI has a good performance in both efficiency and purity. In contrast, for CHS, even though the efficiency is close to 100%, the purity is significantly reduced for a larger number of PU interactions. The combination of CHS+PU jet ID improves the purity, especially for low-p_{T} jets, but with a reduction of the efficiency. Similar behaviour is also observed

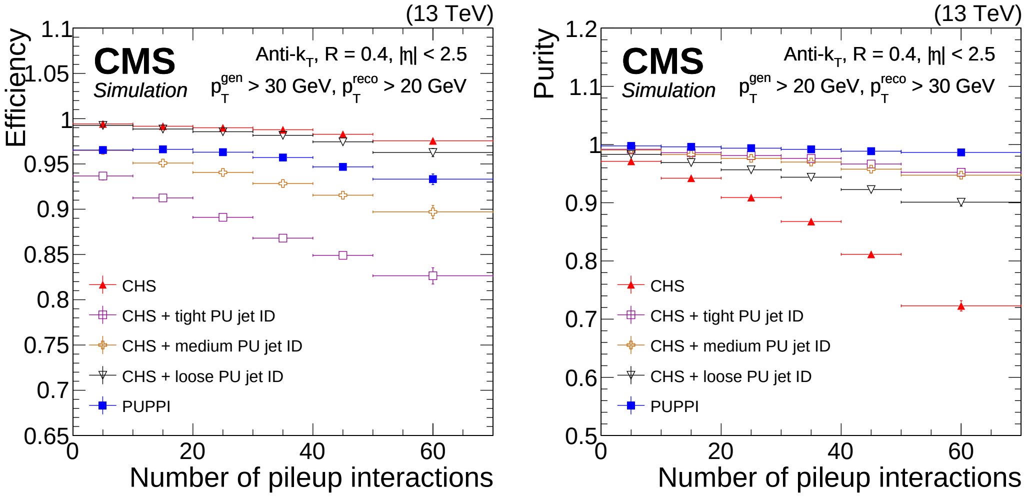


Figure 3.3.6: Efficiency (left) and purity (right) as a function of the number of PU interactions for PUPPI, CHS and CHS+PU jet ID algorithms. Taken from Ref. [96].

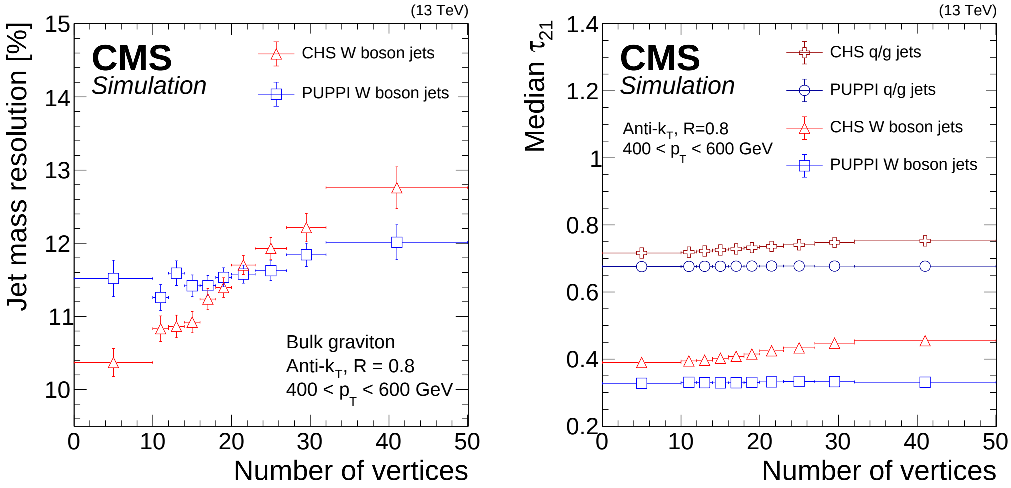


Figure 3.3.7: SD mass resolution (left) and τ_{21} (right) distributions for AK8 jets as a function of the number of vertices for PUPPI and CHS jets. Taken from Ref. [96].

at high values of $|\eta|$. Especially in the forward region, different combinations of the pileup mitigation techniques allow an analysis-dependent optimisation.

A distinctive feature of the PUPPI algorithm is the stability against PU, which is observed for jet-related variables, as well as for objects whose performance depends on the p_T of the PF particles, like the missing transverse momentum (cf. section 3.5) and lepton isolation. As an example, the PU dependence for CHS and PUPPI jets is shown in figure 3.3.7 for two substructure variables, τ_{21} and the SD mass resolution, defined as the ratio between the width and the mean of the distribution of the jet mass of PF jets.

Thanks to its advantages, the PUPPI algorithms is widely used in CMS analyses, especially for studies on substructure variables, and will be the method of choice in Run 3. All these PU mitigation techniques are used in the context of the results presented in this thesis. Further details are provided in chapters 5 and 6.

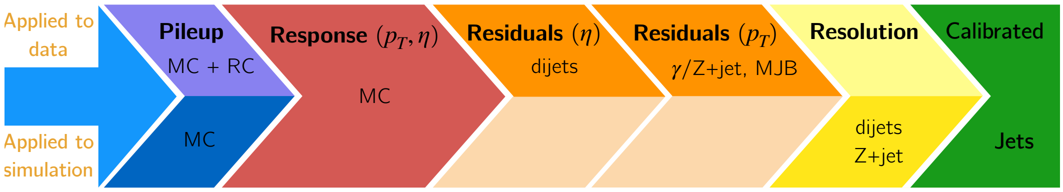


Figure 3.4.1: Simplified representation of the factorised approach used for jet calibration in CMS. More information on each step is given in the text. Adapted from [148].

3.4 Jet calibration

Due to non-linearities in the detector response, imperfect detector modelling, noise and PU effects, the 4-momentum of the clustered PF jet (p^{raw}) can differ from the true 4-momentum of the corresponding particle-level jet (p^{true}).

A correction is applied as a multiplicative factor C, which can be factorised into a set of sequential corrections [148]. The first step removes the average energy offset coming from PU. The core of the calibration procedure is the second step, whose primary goal is to correct the discrepancy in the jet energy introduced by detector non-uniformity in η and the non-linearity in p_{T}. Residual differences between data and simulation are then corrected in two steps: an $|\eta|$-dependent correction, to calibrate the different response of each sub-detector, and a p_{T}-dependent correction to adjust the absolute energy scale. Both residual corrections are smaller than 5% everywhere but in the transition regions between sub-detectors, where they can become sizeable. After the jet energy scale (JES) has been corrected, the jet transverse momentum resolution (JER) in simulation needs to be adjusted to match the jet resolution in data. Therefore, scale factors (SFs) are applied in simulation to broaden the detector response distribution. A more detailed description of the JER SFs measurement is given in chapter 5.

3.4.1 Pileup offset corrections

As mentioned above, the PU contribution due to multiple pp collisions is clustered into the jets, and its additional contribution to the jet energy and momentum is referred to as the "pileup offset". The average value per interaction and jet is approximately 0.5 GeV. Inside the tracker acceptance, the CHS algorithm removes approximately only 50% of the charged PU contribution. The remaining charged component comes from particles that are not associated to any vertex due to vertex reconstruction inefficiency. It is evident that a correction, referred to as "L1 offset correction", is needed to account for the remaining contribution [148]. In contrast, the PUPPI algorithm tackles the PU differently, and it shows a reduced pileup offset contribution; therefore, the L1 correction is unnecessary.

The amount of pileup present in the event can be estimated from three different quantities: the number of reconstructed PVs (N_{PV}), the offset energy density (ρ), defined as the median of the measured energy calculated in η-ϕ bins, and the average number of PU interactions per bunch crossing (μ), obtained by multiplying the

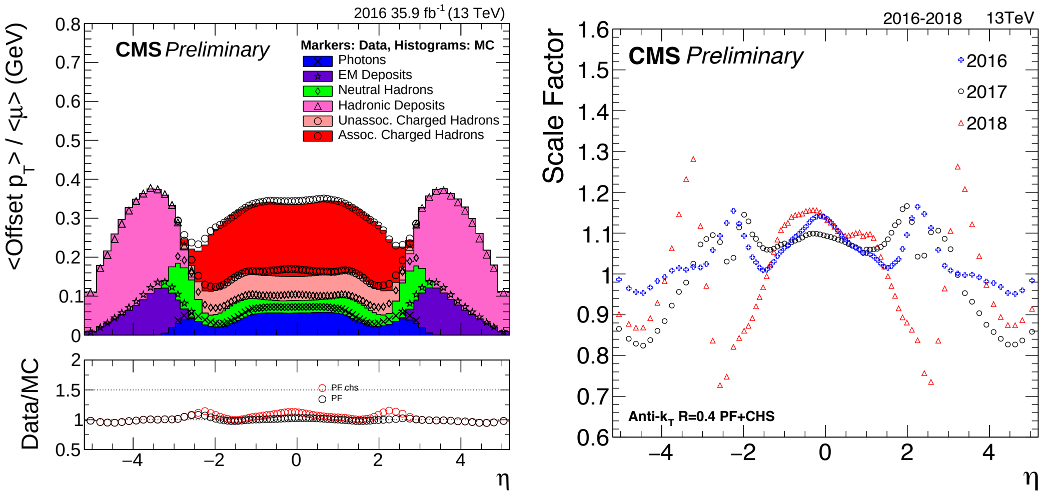


Figure 3.4.2: Left: Data-to-simulation (Data/MC) comparison for the average energy offset per pileup interaction, calculated for each type of PF candidates: the "unassociated charged hadrons" are not removed by CHS algorithm. Right: Evolution of the data-to-simulation SFs during Run 2. Taken from Ref. [1].

instantaneous luminosity with the minimum bias cross section.

The derivation of the correction factor is the following. The hybrid jet area method [148] is used to estimate the average energy offset to be subtracted from the jets. The average offset energy in an event is parametrised as:

$$\langle p_{\mathrm{T}}^{\mathrm{offset}} \rangle = [\alpha(\eta) + \beta(\eta) \cdot \rho] \cdot [1 + \gamma(\eta) \cdot \log(p_{\mathrm{T}}^{\mathrm{raw}})] \cdot A_j \,, \tag{3.7}$$

where A_j is the jet area and $p_{\mathrm{T}}^{\mathrm{raw}}$ is the transverse momentum of the reconstructed jet. The first term captures the PU dependence as a function of ρ, where the α and β parameters correct for non-uniformity versus η. The second term is a minor additional correction accounting for detector and reconstruction inefficiencies at high p_{T}, and it is assumed to have a logarithmic dependence. The parameters in the correction factor are obtained from simulation, where the energy offset is estimated from the average p_{T} difference between matched jets in QCD multijet samples simulated with and without PU. The average offset per pileup interaction is monitored for each type of PF candidate, as shown in figure 3.4.2 (left).

The correction factor, applied to each jet in data and simulation, is given by:

$$C_{\mathrm{hybrid}} = 1 - \langle p_{\mathrm{T}}^{\mathrm{offset}} \rangle / p_{\mathrm{T}}^{\mathrm{raw}} \,. \tag{3.8}$$

The usage of the L1 offset correction is consistent with the absence of additional pileup energy. As the last step, differences between data and simulation (MC) are corrected with a SF, determined with the Random Cone (RC) method from zero-bias data and simulation [148]. The method consists of clustering particles in randomly placed cones, assuming that, in events with no contribution from hard scattering, the main contributions to the jet energies come from the pileup. The L1RC SFs for the Run 2 dataset are reported in figure 3.4.2 (right). The change of the MC tune after 2016 resulted in greater energy flow in the HF. Larger SFs in 2018 are due to additional changes in HF simulation and PF calibration.

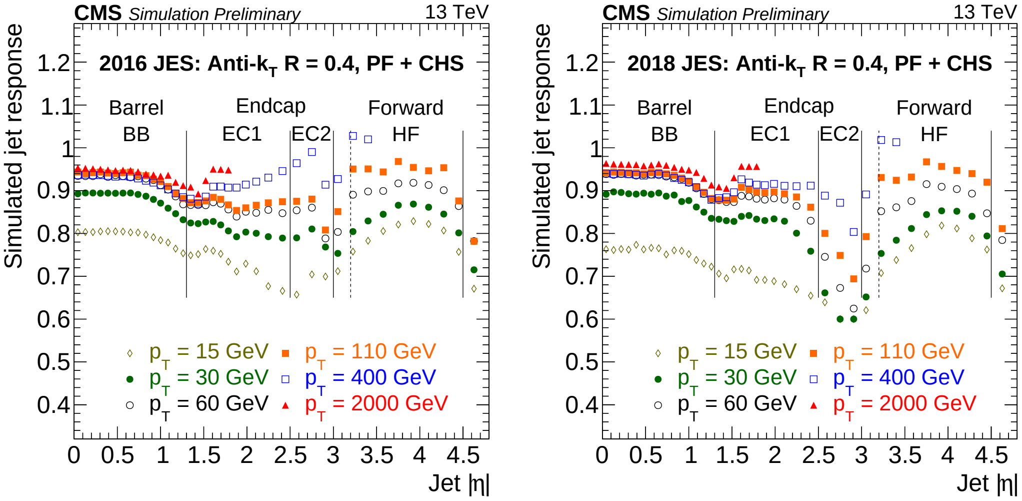


Figure 3.4.3: Simulated jet response in 2016 (left) and 2018 (right) as a function of $|\eta|$ for different p_{T} values. Taken from Ref. [1].

3.4.2 Simulated jet energy response corrections

These corrections are the bulk of the jet energy correction (JEC) scheme; they are derived to ensure that the energy scale of reconstructed PF jets is on average equal to the one of particle-level jets [148].

The CMS detector simulation contains a detailed model of the detector geometry, data-based alignment and calibration of the detector elements, and emulation of the readout electronics, to describe the evolution of particles and their interaction with the detector material. The corrections address the non-uniformity of the detector response as a function of η and p_{T} and are applied to jets that have been corrected for the pileup offset. Evaluated on simulated QCD multijet events, they have the advantage of covering a phase space that is not easily accessible in data, i.e. very small ($p_{\mathrm{T}} < 30\,\mathrm{GeV}$) and very large ($p_{\mathrm{T}} > 1\,\mathrm{TeV}$) momenta, as well as particularly low ($\mu < 5$) and high (μ> 40) number of pileup collisions.

The jet response R is defined as the ratio of the transverse momentum of PF and particle-level jets:

$$R = \frac{p_{\mathrm{T}}}{p_{\mathrm{T}}^{\mathrm{true}}}\,. \tag{3.9}$$

The response is calculated for matched jets in bins of η and $p_{\mathrm{T}}^{\mathrm{true}}$, as shown in figure 3.4.3. The barrel region ($|\eta|$ <1.3) exhibits a stable response at approximately 95%; it is due to a lower response of the detector to the neutral hadrons ($\sim$ 60%), which correspond to approximately 10% of the total energy of the jet (cf. figure 3.4.9). Moreover, the lower response for $p_{\mathrm{T}} < 30\,\mathrm{GeV}$ is caused by the HCAL acceptance. The jump of the response value between $|\eta| = 3$ and $|\eta| = 3.2$ is due to the transition between subdetectors, and the drop for $|\eta| > 4.5$ is because of detector acceptance. Finally, the lower response in 2018 compared to 2016 in the "EC2" region is due to the calorimeter degradation over time.

The jet energy correction based on simulation, referred to as "L2L3Response", is defined as the inverse of the response and, after its application, the response agrees with unity within 1% [148].

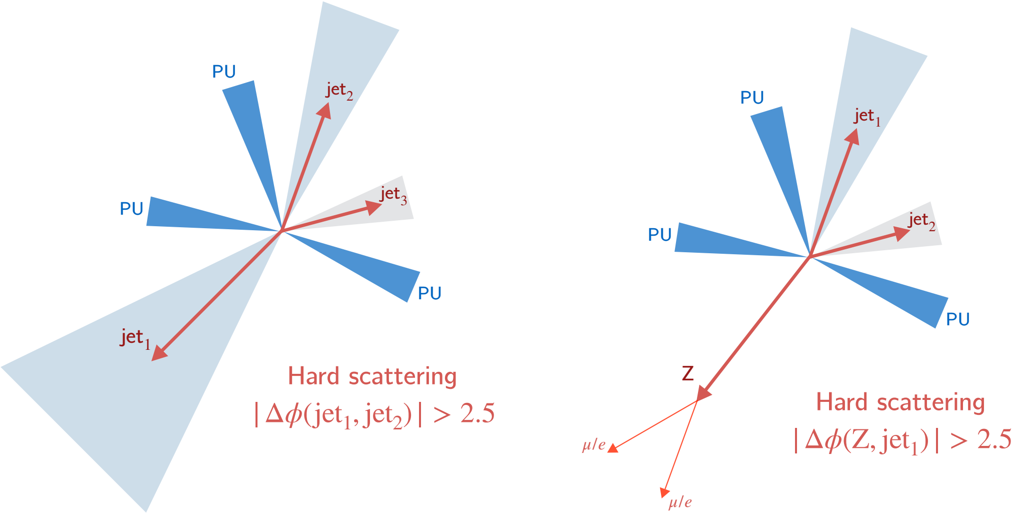

PU
jet$_2$
PU
jet$_3$
PU
jet$_1$
Hard scattering
$|\Delta\phi(\mathrm{jet}_1, \mathrm{jet}_2)| > 2.5$
PU
jet$_1$
PU
jet$_2$
PU
Z
μ/e
μ/e
Hard scattering
$|\Delta\phi(\mathrm{Z}, \mathrm{jet}_1)| > 2.5$

Figure 3.4.4: Sketch of dijet (left) and Z+jet (right) events used in the p_T-balance and MPF methods.

3.4.3 Residual corrections

At this stage, the jet energy scale is well calibrated in simulation. Therefore, residual differences between data and simulation can be corrected into two steps: an η-depentent part, referred to as “L2Residual” and with the purpose of correcting the different response of each subdetector with respect to the central, better-calibrated part, and a p_T-depentent component, referred to as “L3Absolute” and used to adjust the absolute scale difference in the central region of the detector.

These corrections are derived using precisely calibrated objects as a reference and applied to data as simulation-to-data (MC/data) SF. A sample of dijet events is used to derive the η-dependent corrections, while the p_T response is corrected using a combination of Z+jet, γ+jet and multijet events. A schematic representation of such events is shown in figure 3.4.4. The experimental techniques employed and the residual corrections derived are discussed in the following.

Experimental techniques

The jet energy response is studied using the p_T-balance and MPF (missing transverse momentum projection fraction) methods. In the former, the jet response is evaluated by comparing the reconstructed jet momentum ($p_\mathrm{T}^\mathrm{probe}$) directly to the momentum of a reference object ($p_\mathrm{T}^\mathrm{ref}$); in the latter, the response of the whole hadronic activity in the event recoiling against the reference object is considered.

The p_T-balance response, centred at 1 for perfectly calibrated jets, is defined as:

$$R_\mathrm{Bal} = \frac{p_\mathrm{T}^\mathrm{probe}}{p_\mathrm{T}^\mathrm{ref}} \,. \tag{3.10}$$

The MPF method is based on the missing transverse momentum ($\vec{p}_\mathrm{T}^\mathrm{\,miss}$), defined as the negative transverse vector sum of all particles. The Type-I definition is used, described in more detail in section 3.5.

The missing transverse momentum can be seen as:

$$\vec{p}_{\mathrm{T}}^{\,\mathrm{ref}} + \vec{p}_{\mathrm{T}}^{\,\mathrm{recoil}} = -\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}\,, \tag{3.11}$$

where the recoil includes the leading and subleading jets and all the other unclustered particles. Projecting onto the axis of the reference object, the MPF response is:

$$R_{\mathrm{MPF}} = 1 + \frac{\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}} \cdot \vec{p}_{\mathrm{T}}^{\,\mathrm{ref}}}{(p_{\mathrm{T}}^{\mathrm{ref}})^2}\,. \tag{3.12}$$

As no genuine missing transverse energy is expected in these kinds of events, but it originates only from miscalibration, the second term is expected to be small.

Both methods are affected by unavoidable biases. Considering dijet or multijet events, the measured response is biased towards the object with the worse resolution. This can be easily seen if one considers the measurement is a specific p_{T} bin. Reconstructed jets can migrate from adjacent p_{T} bins because of their finite JER. Due to the steeply falling p_{T} spectrum, jets with lower $p_{\mathrm{T}}^{\mathrm{true}}$ fluctuate more often than jets with higher $p_{\mathrm{T}}^{\mathrm{true}}$; as a consequence, the measured response is systematically higher. By performing a measurement in bins of $p_{\mathrm{T}}^{\mathrm{ave}} = (p_{\mathrm{T}}^{\mathrm{ref}} + p_{\mathrm{T}}^{\mathrm{probe}})/2$, this effect can be reduced as the bias is cancelled out on average. Both the MPF and p_{T}-balance methods are sensitive to the JER. This bias is expected to cancel out for the ratio of the relative responses when the jets in the simulation are smeared to match the measured resolution in data. More details about the resolution smearing procedure is reported in chapter 5, together with the interplay between JER and L2Residual corrections.

Another source of bias in the relative response between two objects (reference and probe) arises from radiation and can be shown in the following way. The response of each object is defined following eq. (3.9). In the presence of additional jets in the event, an imbalance exists already at the particle level (Δp_{T}). From the combination of equations (3.9) and (3.10), the relative response can be expressed as:

$$R_{rel} = R^{\mathrm{probe}}/R^{\mathrm{ref}} \cdot \left(1 - \Delta p_{\mathrm{T}}/p_{\mathrm{T,true}}^{\mathrm{ref}}\right)\,. \tag{3.13}$$

This bias can be as large as 5% and is corrected in different ways, as discussed below for the residual corrections and in chapter 5 for the jet transverse momentum resolution measurement.

Relative η-dependent corrections

Residual η-dependent corrections to the jet response are obtained using dijet events, illustrated in figure 3.4.4 (left). The reference jet is required to be in the barrel ($|\eta| < 1.3$), and the probe jet is free to scan the whole η range. The barrel region is chosen because of the uniformity and the small variation of the jet response, and because it provides the highest p_{T}-reach. The two jets are also required to exhibit a back-to-back topology ($\Delta\phi > 2.7$). Moreover, the presence of extra radiation in the event is parametrised with the variable α, defined as the ratio between the p_{T} of the most energetic jet that does not belong to the dijet topology and $p_{\mathrm{T}}^{\mathrm{ave}}$. Events are selected with a maximum of $\alpha < 0.3$.

The response is studied in bins of $p_{\mathrm{T}}^{\mathrm{ave}}$ to reduce the impact of the bias due to the jet p_{T} resolution. Both the p_{T}-balance and the MPF residual responses can be rewritten to account for this change as:

$$R_{\mathrm{Bal}} = \frac{1+\langle\mathcal{A}\rangle}{1-\langle\mathcal{A}\rangle}\,, \qquad \text{with} \qquad \mathcal{A} = \frac{p_{\mathrm{T}}^{\mathrm{probe}} - p_{\mathrm{T}}^{\mathrm{barrel}}}{2p_{\mathrm{T}}^{\mathrm{ave}}}\,. \tag{3.14}$$

$$R_{\mathrm{MPF}} = \frac{1+\langle\mathcal{B}\rangle}{1-\langle\mathcal{B}\rangle}\,, \qquad \text{with} \qquad \mathcal{B} = \frac{\vec{p}_{\mathrm{T}}^{\mathrm{\,miss}} \cdot (\vec{p}_{\mathrm{T}}^{\mathrm{\,barrel}}/p_{\mathrm{T}}^{\mathrm{barrel}})}{2p_{\mathrm{T}}^{\mathrm{ave}}}\,. \tag{3.15}$$

Under the assumption of sufficiently small $p_{\mathrm{T}}^{\mathrm{ave}}$ bins, these two equations are equivalent to equations (3.10) and (3.12).

The presence of ISR and FSR prevent from having an ideal dijet topology. To account for the additional jet radiation, a correction factor (k_{FSR}) is derived[2]. This multiplicative factor is defined as:

$$k_{\mathrm{FSR}} = \lim_{\alpha\to 0}\left[\left(\frac{R^{\alpha}_{\mathrm{MC}}}{R^{\alpha}_{\mathrm{data}}}\right) / \left(\frac{R^{\alpha<0.3}_{\mathrm{MC}}}{R^{\alpha<0.3}_{\mathrm{data}}}\right)\right]\,, \tag{3.16}$$

where R^{α} is the response for a given value of α. The k_{FSR} factor is derived separately for the two methods. It is close to unity for the MPF method, proving that this method is less sensitive to the extra radiation since it exploits the entire hadronic recoil; on the other hand, it can reach values up to a few percent for the p_{T}-balance method, especially in the endcap region.

After correcting for ISR and FSR effects, both methods agree with each other, and the relative response is a good proxy of the ratio between the reference and probe jet responses. Hence, it is possible to derive simulation-to-data (MC/data) SF to correct data for residual differences in the response in different η bins. The residual η-dependent corrections are based on results obtained with the MPF method, while the p_{T}-balance results are used as a cross-check.

The final L2Residual corrections are shown in figure 3.4.5; it is evident that jets in the barrel ($|\eta| <1.3$) are better calibrated, as the magnitude of the residual correction is smaller than 1%. Time-dependent corrections address the evolution and ageing of the detector in different data-taking periods and years; this is particularly important for the endcap region outside the tracker coverage ($2.5 < |\eta| < 3$) and in the HF region due to the exposure to a higher level of radiation. The data during RunD in 2018 was reconstructed as Prompt, while all the others are Rereco (cf. section 2.2.7). This difference is reflected in the large discrepancy in the transition region between subdetectors ($2.5 < |\eta| < 3$). Consequently, the luminosity-averaged correction for 2018 in the same region shows a different behaviour compared to the other years. This is also an indication that the method is able to correct for miscalibration of the reconstruction process and that the Legacy calibration is needed to achieve the best possible results.

Absolute p_{T}-dependent corrections

The absolute p_{T} dependence of the jet response in $|\eta| < 1.3$ is corrected using Z+jet, both in the electron ($\mathrm{Z} \to \mathrm{ee}$) and muon ($\mathrm{Z} \to \mu\mu$) channels as illustrated in figure 3.4.4 (right), γ+jet and multijet events. Each channel uses both the MPF and the p_{T}-balance methods, which are then combined to constrain the relative biases. The L3Absolute corrections for 2018 are reported in figure 3.4.6 as an example; the response smaller than 1 is due to the bias coming from FSR and ISR effects.

[2] The subscript FSR is used instead of ISR and FSR for brevity.

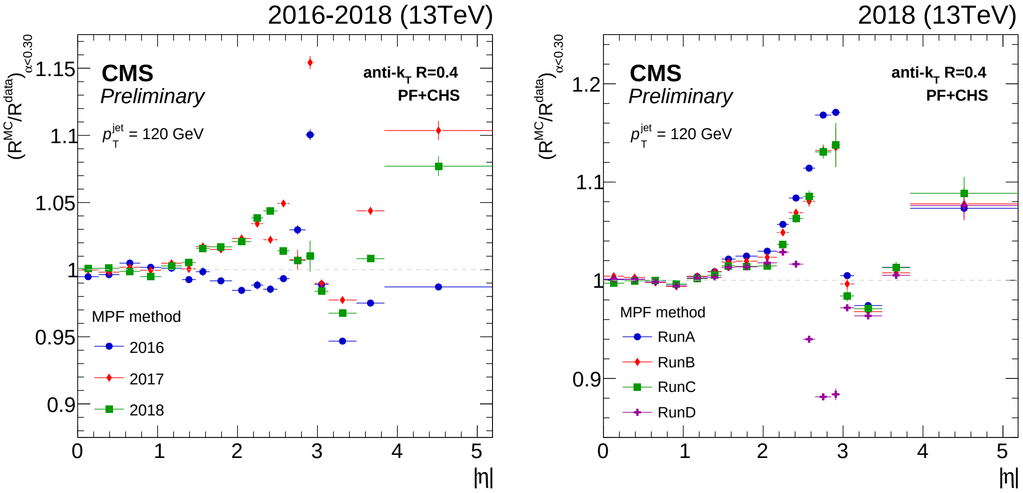


Figure 3.4.5: L2Residual corrections as a function of $|\eta|$ derived using dijet events with the MPF method in different data-taking periods. Left: Luminosity-averaged corrections for each year of Run 2. Right: Run-dependent corrections for 2018. Taken from Ref. [1].

An extensive p_T-spectrum is investigated, where each channel covers a specific jet p_T-range; the multijet results provide the dominant contribution at high p_T (up to 3 TeV), the Z+jet results allow to probe the low-p_T region (down to 30 GeV), and the γ+jet results are used in the intermediate range, having overlap with both of the other channels.

For the Z+jet and γ+jet channels, the methods used are similar to the ones discussed above for the dijet channel. As opposed to the L2Residual corrections, the absolute jet response is measured relative to the more precisely calibrated γ or Z boson. These channels exhibit a similar bias coming from the ISR and FSR; therefore, a k_FSR correction is defined as:

$$k_\mathrm{FSR} = \frac{\lim\limits_{\alpha\to 0} R^{\alpha}_\mathrm{jet}}{R^{\alpha=0.3}_\mathrm{jet}}\,, \tag{3.17}$$

where α, in this case, is the ratio between the p_T of the second most energetic jet and $p_\mathrm{T}^\mathrm{ref}$. The jet response is linearly dependent on α in both methods, although the MPF method is significantly less sensitive to radiation than the p_T-balance method, as shown in figure 3.4.7 (left). The k_FSR is used to correct both MPF and p_T-balance results before they are used as input for the next step of the procedure. The k_FSR is also extrapolated as a function of p_T and smoothed with a log-quadratic fit, used in a later stage. The parametrisations of the k_FSR correction for different channels are shown in figure 3.4.7 (right), together with the pre- and post-fit values and uncertainties.

A similar procedure is used for the multijet channel, with an additional complication from the correlation of the JES of two p_T regimes, the one of the leading jet and the one of the recoil system. In fact, the response for high-p_T jets obtained using multijet events is relative to the response of the lower-p_T jets; for this reason, the multijet analysis can only constrain the JES p_T-dependance at high-p_T, while the Z+jet and γ+jet analyses are also sensitive to the absolute scales with a precision dominated by the Z → $\mu\mu$ channel.

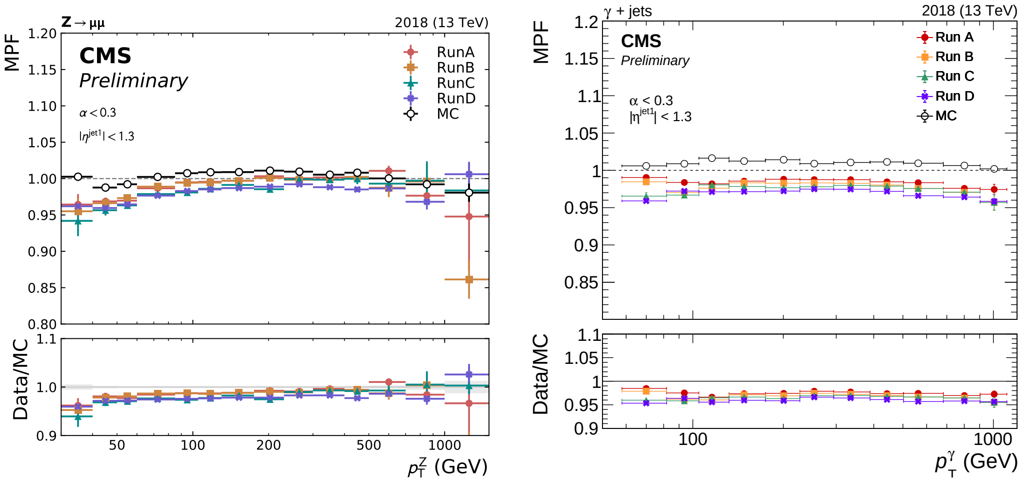


Figure 3.4.6: Data to simulation comparison of the jet response as a function of $p_{\mathrm{T}}^{\mathrm{Z}}$ in Z → $\mu\mu$ +jet events (left) and p_{T}^{γ} in γ+jet events (right) derived using the MPF method. Taken from Ref. [1].

The responses from the different channels are combined in a global fit to extract the p_{T}-dependence of the response. The global fit allows for a reduction of fluctuations between different channels, caused by slight shifts in the lepton and photon energy scales. It is also possible to further constrain the uncertainties by combining the results obtained with the p_{T}-balance and the MPF methods, as they are statistically more precise in the low-p_{T} and high-p_{T} regimes, respectively. This is related to the JER bias present in both methods. The smearing of ISR and FSR jets can change the balance between the two leading objects, which influences the p_{T}-balance more than the MPF method. On the other hand, JES miscalibration of low-p_{T} jets (mostly from PU) affects the resolution of $p_{\mathrm{T}}^{\mathrm{miss}}$; hence the MPF method is more susceptible to noise compared to p_{T}-balance.

The data-to-simulation ratio of the jet response obtained from all channels and methods, after the k_{FSR} correction, is given as input to the global fit, together with the following nuisance parameters:

- **Lepton/photon scale uncertainties:** The scale parameters are motivated by residual miscalibrations of the lepton and photon scale between data and simulation. Their effect is of the order of 0.1-0.5%, and it is assumed to be uncorrelated among the channels and independent of p_{T}[3].

- **EM footprint uncertainty:** The EM footprint removal algorithm is applied to photons but not to electrons [148]. Two uncorrelated p_{T}-independent nuisance parameters are included for the Z → ee and γ+jet channels to account for potential biases with respect to the Z → $\mu\mu$ channel. This is considered for the MPF method only and has an impact of about 0.2%.

- **ISR+FSR correction uncertainty:** The parameters of the k_{FSR} corrections are treated as uncorrelated nuisance parameters for each channel and method[4].

[3] The p_{T}-independence assumption had been revisited for the Legacy calibration of Run 2.

[4] This method been revisited for the Legacy calibration of Run 2, where specific corrections are applied to extra jets and unclustered energy individually.

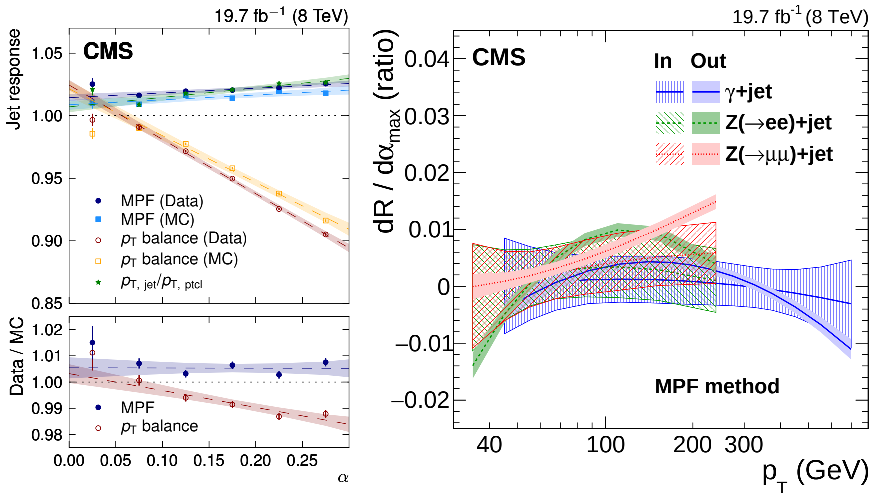


Figure 3.4.7: Left: Data-to-simulation comparison of the jet response as a function of $\alpha = p_{\mathrm{T}}^{\mathrm{jet},2}/p_{\mathrm{T}}^{\mathrm{Z}}$ for MPF and p_{T}-balance in the $\mathrm{Z} \to \mu\mu$ channel. Right: Central value and uncertainties for the k_{FSR} correction as a function of p_{T} for the MPF methods and different channels. The shadowed regions show the input distributions to the global fit, while the filled coloured regions show the post-fit distributions. Taken from Ref. [148].

- **Fit parameters:** They are the degrees of freedom used to model the p_{T}-dependence. During Run 1 only two parameters were used. This choice was driven by studies on the single-pion response in HCAL and ECAL. During Run 2, separate parametrisations for HCAL and ECAL scales are used, and other detector effects have been included.

The result of the global fit (solid black line) and its uncertainty (black dotted lines) are shown in figure 3.4.8. The $\mathrm{Z} \to \mu\mu$ and $\mathrm{Z} \to \mathrm{ee}$ results are combined into a single Z+jet channel after the scale corrections are applied. The procedure is validated on a Z boson enriched sample, where the two channels are found to be in agreement. The data points in figure 3.4.8 are shifted by the nuisance parameter values taken at their post-fit values. The deviation of the absolute scale uncertainty (yellow band) from the fit is used in the “Time stability” component of the total JES uncertainty, as explained in section 3.4.5.

3.4.4 Jet resolution smearing

Jets generally have a wider energy resolution compared to other physics objects, like electrons, muons and photons. Furthermore, the JER in data systematically exceeds that in simulation. To correct for this effect, data/MC SFs are applied to smear the jet energy resolution in simulation to match the one in data.

The JER is computed from fully-calibrated jets, whose JES has been corrected with the entire chain detailed above. The measurement is an extension of the methods used for estimating the JEC, with the difference that the width of the response distribution is the variable under consideration.

A comprehensive description of the determination of the JER in simulation and its SF is presented in chapter 5, while the impact of the JER smearing has already been described in the previous sections.

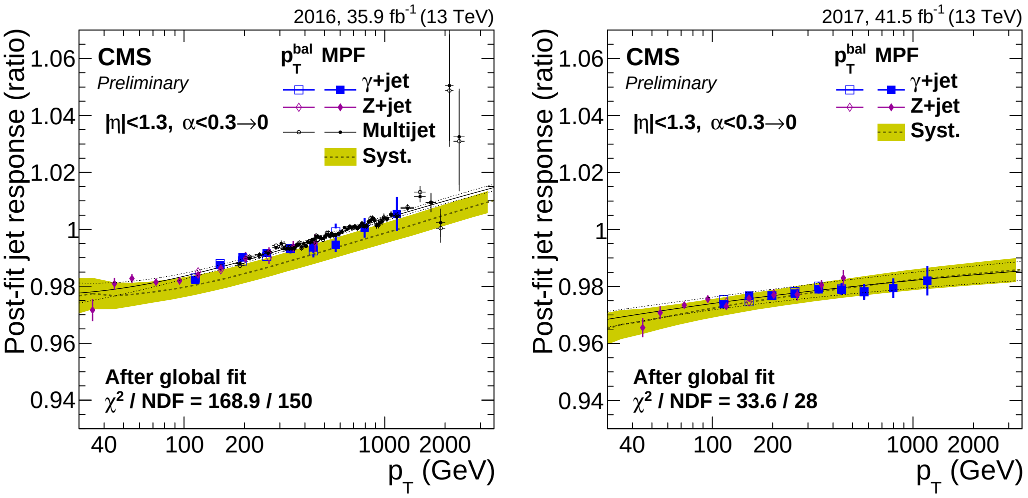


Figure 3.4.8: Data-to-simulation comparison for the jet response dependence on $p_{\mathrm{T}}^{\mathrm{jet}}$ for 2017 and 2018. Different channels and methods are combined into a global fit, shown as solid black line with its post-fit nuisance parameters. Yellow band indicates the per-run luminosity-weighted average of the absolute scale uncertainty. Taken from Ref. [1].

3.4.5 Calibration uncertainties and jet composition

The outcome of the jet energy calibration procedure (see figure 3.4.1) is an overall correction for the JES with a precision at the level of a few percent.

The uncertainties from each step are propagated and correlated across the whole phase space. A total uncertainty is provided, together with the individual sources of uncertainties, which can be further constrained in analyses particularly susceptible to the jet energy scale [148].

The correlation across η is based on a division provided by the sub-detector composition, used already in each step of the calibration procedure to ensure consistency. The main regions are the barrel region ($|\eta| < 1.3$), the two endcap regions within ($1.3 < |\eta| < 2.5$) and outside ($2.5 < |\eta| < 3.0$) the tracker coverage, and the hadron forward region ($3.0 < |\eta| < 5.2$).

All the sources of systematic uncertainties are grouped into six categories:

- **Pileup:** It is extracted from the MC-truth-based offset corrections, contributing the most at low-p_{T}.
- **Relative η-dependent:** It is calculated from the difference of the L2Residual corrections when varying the JER SFs within their uncertainty and from the simulation-based closure-tests using PYTHIA8 and HERWIG++ samples to assign an uncertainty due to the different modelling of ISR and FSR.
- **Absolute p_{T}-dependent:** It is obtained from the global fit as described in the previous section. It accounts for the differences between the MPF and p_{T}-balance results, from Z+jet, γ+jet, dijet events.
- **Method and sample:** It accounts for differences in the JEC corrections derived using the p_{T}-balance and MPF methods and all the available channels.

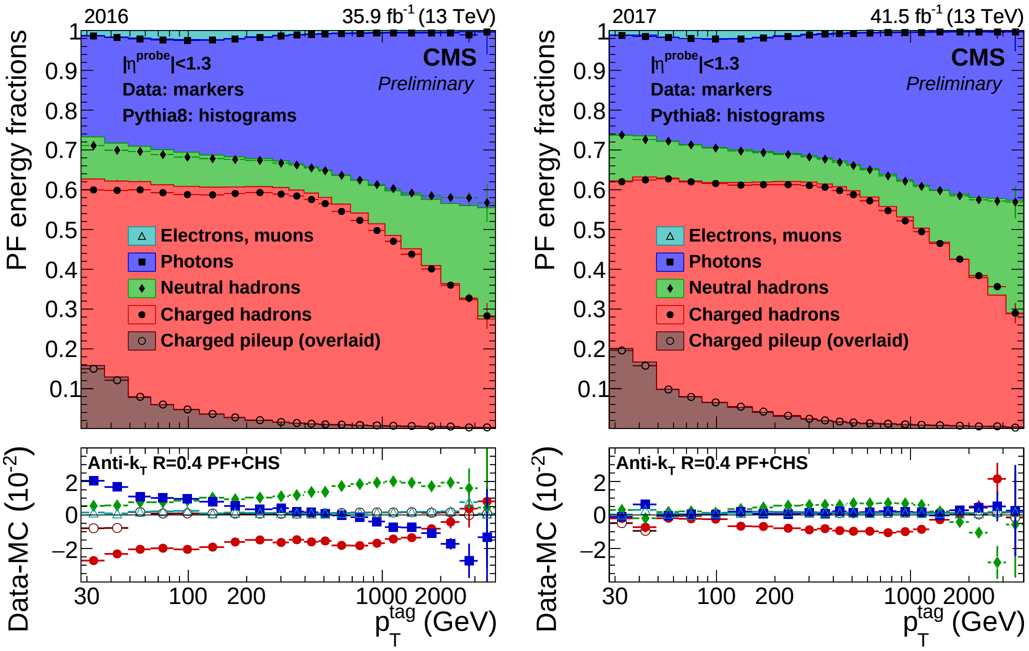


Figure 3.4.9: Jet PF composition using fully corrected jets in 2016 (left) and 2017 (right). The energy of the charged PU contribution, removed by the CHS algorithm, is overlaid. Taken from Ref. [1].

- **Time stability:** It accounts for residual miscalibration after the correction for detector effects, like radiation damage to ECAL and HCAL. It is evaluated from differences between the L2Residual corrections per data-taking period per year and extracted from the global fit stability.

- **Flavour response:** It is determined from differences between HERWIG++ and PYTHIA8 responses for different flavour jets.

A summary of these uncertainties is shown in figure 3.4.10. The final uncertainties are below 3% in the barrel region and in the p_T region considered by most analyses ($p_T > 30\,\text{GeV}$). The Run 1 uncertainty without flavour and time sources is shown for comparison. Further reduction of the uncertainties associated to the jet energy calibration procedure is expected for the Legacy reconstruction to reach a calibration precision below 1%.

A comparison between data and simulation for monitoring the stability of JES can be carried out by studying the jet energy fraction of different PF candidates: photons, leptons, neutral and charged hadrons. The jet's momentum is carried on average by its constituents is the following: 65% for charged hadrons, 25% for photons, and 10% for neutral hadrons. The PF jet composition is determined using a dijet sample with the tag-and-probe method. The measured PF energy fractions are shown in figure 3.4.9 as a function of the jet transverse momentum; the agreement between data and simulation is at the level of 1-2% in the barrel, consistent within the JES uncertainties.

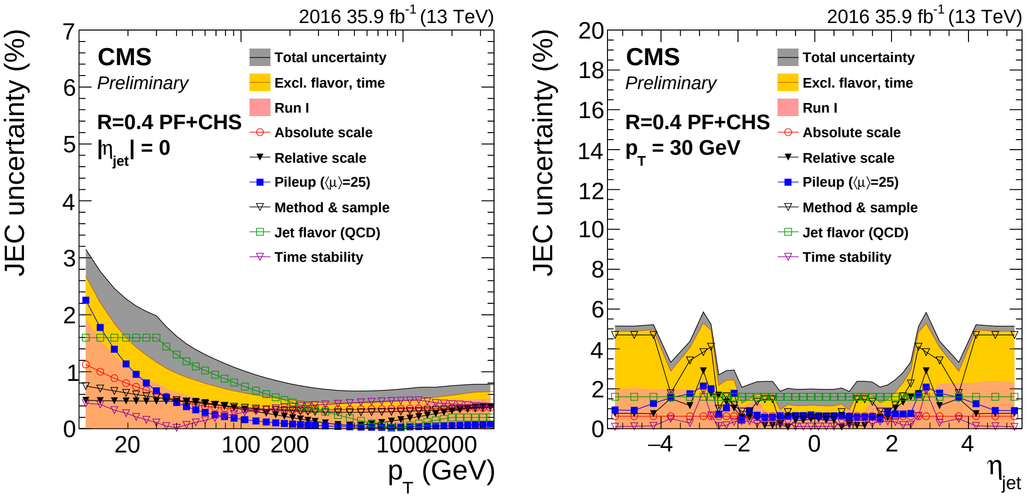


Figure 3.4.10: JES uncertainty sources and total uncertainty as a function of $p_{\mathrm{T}}^{\mathrm{jet}}$ (left) and η_{jet} (right). The results for the 2016 dataset are shown as a showcase. The Run 1 uncertainty is shown for comparison. Taken from Ref. [1].

3.5 Missing transverse energy

Particles that undergo only weak interactions have an extremely small interaction cross section with matter and cannot be directly traced by the detector. As a consequence, the presence of such particles, like neutrinos or new particles arising from BSM effects, results in undetected energy in an event. A correct evaluation of the missing energy plays an essential role for many CMS analyses. Its use in the $\mathrm{Z}' \to \mathrm{ZH}$ analysis presented in this thesis and targeting the invisible decays of the Z boson will be described in chapter 6.

The missing transverse momentum ($\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$) can be inferred from an imbalance in the momentum in the transverse plane using the visible particles of the event. In CMS analyses, $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ is calculated as the negative sum of the $\vec{p}_{\mathrm{T}}$ of all PF candidates. The CHS algorithm is not applied, as it acts only on charged particles inside the tracker volume and would create an imbalance. However, the PUPPI algorithm can be utilised to reduce the presence of PU, as it does not act preferentially on any particular region. For the $p_{\mathrm{T}}^{\mathrm{miss}}$ computation, the PUPPI metric described in section 3.3.3 is used, where all leptons[5] and photons reconstructed in the tracker region ($|\eta| < 2.5$) with $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ are considered as prompt and assigned a weight of 1.

Anomalous high-$p_{\mathrm{T}}^{\mathrm{miss}}$ events can appear from malfunctioning detector components and reconstruction failures. Special noise-rejection techniques are used to reject such events with fake $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$. These techniques tackle different origins: machine-induced backgrounds, especially beam halo, noisy sensors in calorimeter cells, identified from their pulse shape and timing information, and badly reconstructed muons and punch-through hadrons. Figure 3.5.1 showcases the filter effectiveness to suppress these spurious events in data.

Furthermore, the accuracy of the $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ reconstruction depends also on the calibration of PF objects, as a miscalibration of the visible part of the event leads to an

[5] It has been studied that treating the lepton as charged particles would artificially create a PU dependence by giving PU particles around the prompt lepton a higher weight [96].

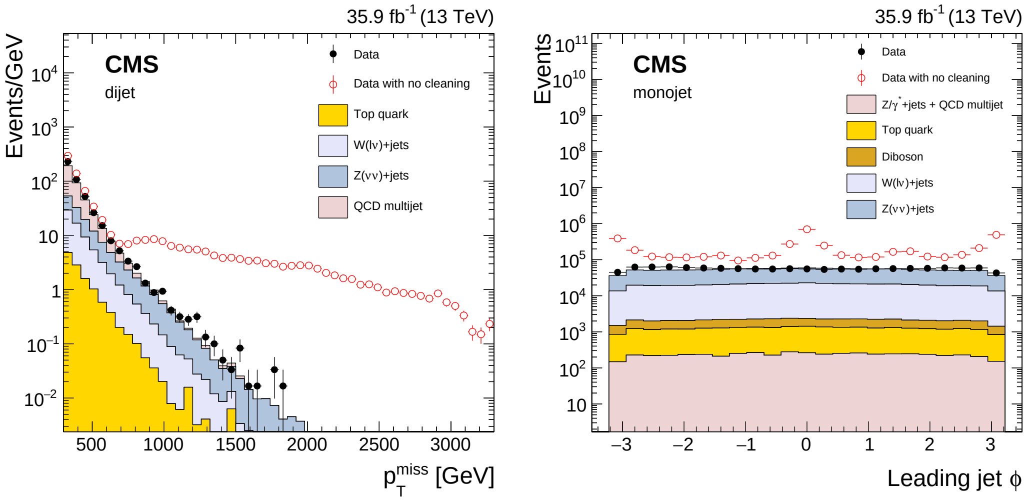


Figure 3.5.1: $p_{\mathrm{T}}^{\mathrm{miss}}$ (left) and jet ϕ distributions with the event filtering algorithms applied on a dijet selection. The excess at $\phi \sim 0$ and $\phi \sim \pi$ are due to the shape of the beam halo in the LHC tunnel. Taken from Ref. [163]. Run 2 performance reported in Ref. [164].

inaccurate estimation of the genuine $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ due to invisible particles. The largest impact is induced by the jet energy corrections. The calibration procedure considerably alters the jet energy. Therefore, these corrections are propagated onto the missing transverse energy [148]. The resulting $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$, referred to as “Type-I”-corrected $p_{\mathrm{T}}^{\mathrm{miss}}$, is calculated as:

$$\vec{p}_{\mathrm{T}}^{\,\mathrm{miss,corr}} = \vec{p}_{\mathrm{T}}^{\,\mathrm{miss,raw}} + \sum_{\mathrm{jet}} \left(\vec{p}_{\mathrm{T}}^{\,\mathrm{jet,raw}} - \vec{p}_{\mathrm{T}}^{\,\mathrm{jet,corr}} - \mathcal{O}^{\mathrm{RC}} \right) . \tag{3.18}$$

The $\vec{p}_{\mathrm{T}}^{\,\mathrm{jet,corr}}$ is the fully corrected jet p_{T}, and $\mathcal{O}^{\mathrm{RC}}$ is the average PU offset obtained with the RC method, as described in section 3.4. The pileup offset correction is not propagated to ensure that no bias is introduced as the pileup offset is by definition isotropic. The L1 offset correction are not derived for PUPPI jets, for which $\mathcal{O}^{\mathrm{RC}}_{\mathrm{PUPPI}} = 0$.

The performance of missing transverse energy is measured in Z+jet and γ+jet events, in which the boson defines a reference scale and axis. No genuine $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ is expected in such events, but it arises from the miscalibration of the other objects. The momentum conservation in the transverse plane is defined in eq. (3.11). The parallel and perpendicular projections of the hadronic recoil onto the boson axis are denoted by $u_{||}$ and $u_{\perp}$, respectively, and used to study the response and resolution of the magnitude of the missing transverse momentum ($p_{\mathrm{T}}^{\mathrm{miss}}$). The results are shown in figure 3.5.2 (left), where the agreement between the different channels is demonstrated within a few percent. A response of unity is achieved starting from p_{T} >100 GeV, while the turn-on at low-p_{T} is a consequence of the imperfect calibration of jets with $p_{\mathrm{T}} < 15$ GeV, and unclustered particles, for which no dedicated corrections are available. The $p_{\mathrm{T}}^{\mathrm{miss}}$ resolution is dominated by the resolution of the hadronic activity, since the momentum resolution for leptons and photons is approximately 1% compared to 5–20% for the jet momentum resolution (cf. chapter 5).

The influence of the PUPPI algorithm on $p_{\mathrm{T}}^{\mathrm{miss}}$ is shown in figure 3.5.2 (right). A direct improvement in the resolution of the transverse mass in W+jet events,

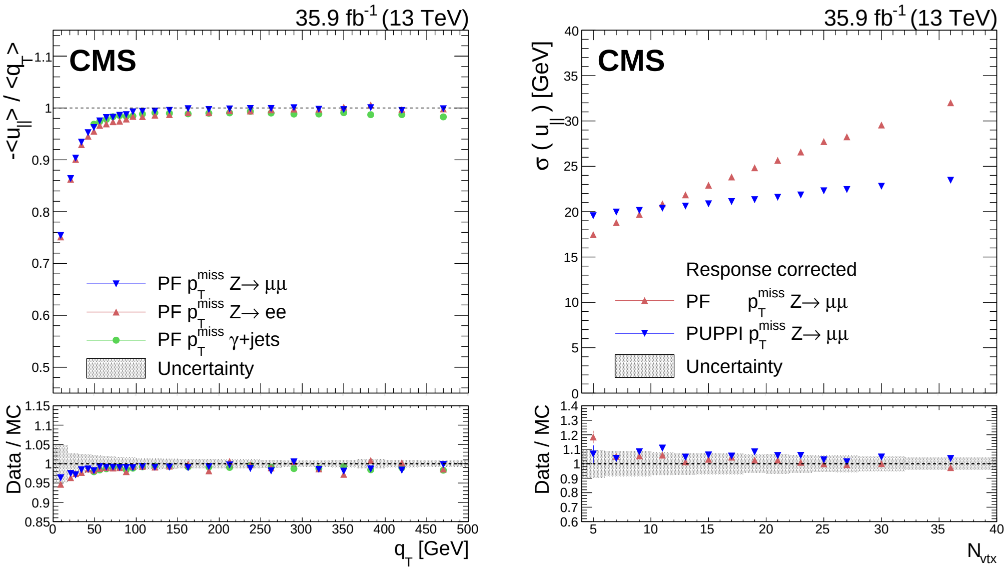


Figure 3.5.2: Left: Response of $p_{\mathrm{T}}^{\mathrm{miss}}$ in Z+jet and γ+jet events. Right: Parallel resolution of the hadronic recoil in Z+jet events as a function of the number of vertices for different PU suppression algorithms. Taken from Ref. [163].

where genuine $p_{\mathrm{T}}^{\mathrm{miss}}$ is expected, is observed [163]. As for other jet-related variables presented in section 3.3.3, the PUPPI algorithm shows better stability against PU, and the gain at the expected PU for Run 3 is significant.

3.6 Event simulation in proton-proton collisions

The ability to predict SM processes to high precision and model potential BSM interactions is crucial for most LHC analyses. Being manifestations of quantum mechanics, processes in pp collisions cannot be precisely calculated on an event-by-event basis. Therefore, Monte Carlo (MC) event generators are employed to simulate such complex events, as summarised in the following.

In simulations of pp collisions, the composite proton structure is taken into account in order to accurately describe the *hard scattering* process, i.e. the parton interaction with the highest momentum transfer in a collision. The cross section of a given process $ij \to \mathrm{X}$, $\hat{\sigma}_{ij\to\mathrm{X}}$, for two partons i and j is calculated perturbatively. The first and second orders in the perturbation series are referred to as leading order (LO) and next-to-leading order (NLO), respectively. In the context of simulation of pp collisions, NLO usually refers to the perturbation order in QCD; however, also EW corrections are often applied at the analysis level to further improve the precision of the prediction (see section 6.3).

The cross section of each interaction between partons must be convolved with the parton distribution function (PDF) $f(x, q^2)$ of the proton constituents. The PDF describes the probability to find a given parton with a fraction x of the total momentum of the initial proton and at the energy scale q^2 probed by the hard interaction. The form of PDFs is not predicted by QCD and, therefore, extracted from experimental data [165]. Summing over the possible parton flavours i and j in a given process, the cross section of the process pp → X can be factorised as [166–168]:

$$\sigma_{\mathrm{pp}\to\mathrm{X}} = \sum_{i,j} \int \int f_i(x_i, q^2)\; f_j(x_j, q^2)\; \hat{\sigma}_{ij\to\mathrm{X}}(x_i, x_j, q^2)\; \mathrm{d}x_i\, \mathrm{d}x_j\,. \tag{3.19}$$

The hard scattering process can be simulated by several MC event generators, for example, MadGraph5_amc@nlo [128], powheg [124–126], herwig++ [127] and pythia8 [123].

After the generation of the hard interaction, the emission of initial state radiation (ISR) and final state radiation (FSR) from the partons of the hard scattering event is simulated. The shower induced by the coloured final state particles, known as parton shower, is simulated, and its evolution is modelled until the energy scale is too small to perform perturbative calculations. Afterwards, the hadronisation of the individual particles into stable colourless hadrons is simulated. In the case of events simulated with MC generators that do not provide a description of the parton showering, the generator is interfaced to other programs like pythia8 or herwig++ for this purpose.

Additional interactions between the remaining partons that do not directly take part in the hard scattering process are known as underlying event (UE). Such interactions, mostly soft non-perturbative scatterings, are treated phenomenologically by the MC generators, which strongly rely on experimental measurements for their correct modelling. Finally, the additional collisions (PU) that occur in the same or adjacent bunch crossing are simulated.

Last, all particles produced in the previous steps are interfaced to a simulation of the CMS detector, based on the Geant4 toolkit [129–131], to account for their interaction with the detector components. At this stage, the event reconstruction algorithms described in the previous sections are applied and the recorded data can be compared to the simulated prediction.

4 Deep learning approaches for jet tagging in CMS

The recent developments in machine learning and its success in several fields prompted the usage of such techniques also in high energy physics [169–171]. The data to be analysed is rapidly increasing both in complexity and size, and traditional approaches are being replaced by modern tools. These higher-dimensional problems make machine learning algorithms, and in particular the deep learning approach, excellent candidates for the task. This chapter gives an overview of the machine learning application to jet flavour classification in the CMS Collaboration.

4.1 Introduction

Machine learning ML is a field of computer science based on data analysis that automates the creation of analytical models. The task of the ML algorithm, applied in classification problems, is to approximate, as good as possible, a function with potentially very large dimensionality (N) by reducing it to contain the critical information necessary to perform the classification. However, for very high dimensional spaces ($N > 50$), the task remains complicated, and until the recent advent of deep learning, it appeared to be overwhelming.

An overview of the jet flavour tagging algorithms used in the CMS Collaboration is presented in the following. The more traditional approaches, as well as the state-of-the-art techniques based on ML, are discussed. Particular emphasis is put on the DeepAK8 and ParticleNet approaches. Their application in the context of this thesis is discussed in chapter 6.

4.2 Jet flavour tagging

The hadronic decay products of highly energetic, heavy particles, like the top quark or SM bosons, are likely to be reconstructed in a single large-radius jet, if the initial particle is strongly Lorentz-boosted. Such jets are characterised by a distinctive radiation pattern ("substructure") and, potentially, flavour content, which can be exploited in order to distinguish the underlying physics processes. Therefore, the nature of the jet-initiating particle can be inferred from these properties, providing discrimination power between signal and background processes and potentially increasing the sensitivity of both SM measurements and searches for new physics.

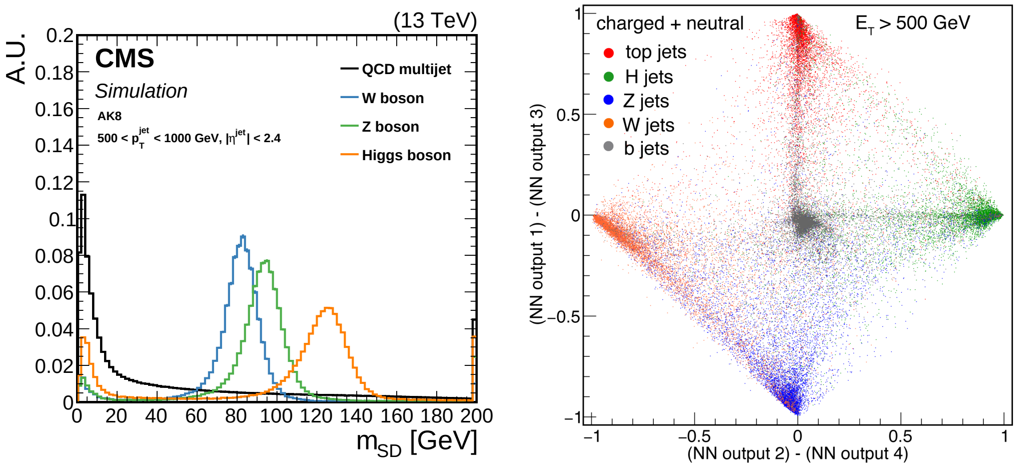


Figure 4.2.1: Left: SD mass distribution for large-radius jets originating from different physics processes. Taken from Ref. [159]. Right: Distributions of the BEST tagger output differences for different particle initiating the jet. Taken from Ref. [172].

4.2.1 Traditional approach

The first attempts to the identification of boosted jets arising from heavy particles were based on one-dimensional criteria applied to the jet mass or several substructure variables (cf. section 3.3.2). Such approaches provide powerful discriminators against jets produced from the hadronisation of quarks and gluons, thanks to the relatively low mass values and the soft, collimated radiation typical for QCD processes.

However, such observables change substantially in the presence of additional particles arising from ISR, UE and PU. The SD algorithm is the most frequently used technique in the CMS Collaboration to remove soft and uncorrelated radiation from jets, resulting in a weaker correlation between the jet's mass and p_T. Figure 4.2.1 (left) shows the distribution of the SD-corrected invariant mass (SD mass) of large-radius jets originating from different physics processes. The jet mass spectrum of heavy boson or top quark decays allows for clear separation from QCD-initiated jets. Jet classification relying on those properties have become a standard in CMS analyses and are described in more detail in Ref. [159].

4.2.2 Machine learning approach

Despite the variety of taggers already available in the CMS Collaboration, more sophisticated algorithms are being developed to cope with the increasing complexity of the classification task. Recent developments in ML, in particular deep neural networks (DNNs), allow for the usage of a larger number of low-level features, giving the ability to improve the jet categorisation further. In fact, the algorithms described above rely only on high-level observables, i.e. quantities calculated analytically. The natural supposition is that the high-level, low-dimensional information sacrifices information needed for classification, which can be recovered to improve the performance. DNNs are capable of learning complex, non-linear correlations when trained on a sufficiently large data sample [169] directly from low-level inputs such as the four-momenta of the jet constituents.

Another advantage of DNNs is the possibility to easily perform a multi-class classification instead of distinguishing only between signal and background. This feature allows, for example, targeting different backgrounds individually.

Finally, there exists a multitude of representations of the jet properties that can be used to analyse the low-level information and perform the classification task. A more detailed overview of deep learning and deep neural networks can be found, for example, in Ref. [173]. In the following, examples of different DNN-based algorithms for jet classification used in the CMS Collaboration are presented.

Boosted event shape tagger

The boosted event shape tagger (BEST) [159, 172, 174] is a multi-class classification algorithm designed to identify high-p_{T} jets originating from the hadronic decays of top quarks and SM bosons. The tagger exploits the substructure and b tagging information using a fully connected neural network. For each jet, substructure-related variables are calculated under the assumption of different initiating particles: top quark, W, Z and Higgs boson. For each of these four hypotheses, a Lorentz transformation of the jet constituents into the rest frame of the respective hypothetical original particle is performed. Should a jet have been initiated by one of the four candidate particles, its constituents are expected to be isotropically distributed, with balanced momenta and an N-prong topology only in the corresponding rest frame.

Figure 4.2.1 (right) shows the distributions of the difference between the tagger outputs in the two-dimensional plane for different jet categories; a good discrimination power is observed for all categories, including W and Z decays. This approach outperforms classical tagging algorithms, which are based on only a few substructure observables. A comparison of the performance of this tagger to other algorithms is given in section 4.4.

ImageTop

The ImageTop tagger [159] is based on the assumption that the PF candidates used to reconstruct the jet provide additional information to improve the classification performance. This algorithm uses an image recognition technique based on two-dimensional convolutional neutral networks (CNNs) to discriminate jets originating from top quark decays and QCD processes.

The jet is described as a series of pixelated representations (channels) in the η-ϕ plane. The intensity of each pixel encodes the p_{T} sum of all particles reconstructed in a given angular region, while the information on the PF candidate flavours, namely charged and neutral hadrons, photons, electrons, and muons, is encoded in different channels.

In order to remove unnecessary rotational and translational degrees of freedom, each PF particle undergoes the following preprocessing before the pixelation process. First, a shift is performed such that the jet axis is centred in the picture frame; then, a rotation is applied, such that the major jet axis always points in the vertical direction; finally, the image is flipped, such that the lower-right quadrant contains the maximum intensity of the variable considered. An example of the two-dimensional representations used by the ImageTop neural network is shown in figure 4.2.2. The 3-prong substructure is visible for the jet initiated from the hadronic top quark decays, while a more isotropic distribution is observed for QCD-initiated jets.

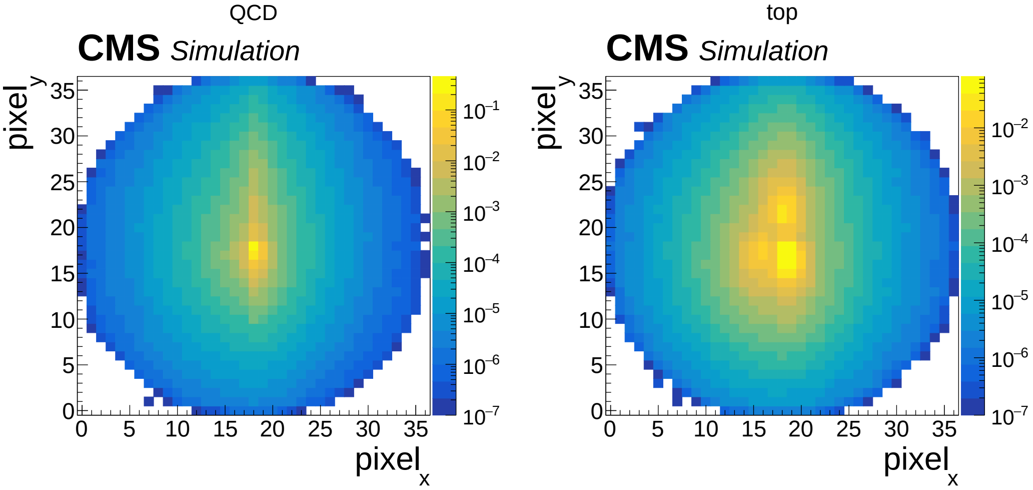


Figure 4.2.2: Examples of two-dimensional representations of jets originating from QCD precesses (left) and top quark decays (right), obtained from the overlaid images of several jets after post processing. Taken from Ref. [159].

DeepAK8

An alternative approach to analyse the information carried by the jet constituents involves a customised DNN architecture known as DeepAK8 [159]. This multi-class classifier aims at identifying the hadronic decays of top quark, SM vector and Higgs bosons, and QCD-initiated jets. Furthermore, each main category is subdivided into minor classes, corresponding to the primary decay modes of each particle (e.g. $H \rightarrow b\bar{b}$, $H \rightarrow c\bar{c}$ and $H \rightarrow qqqq$). The output classes for the DeepAK8 classifier are summarised in table 4.2.1.

Category	Label
Higgs	H(bb)
	H(cc)
	H(VV* → qqqq)
Z	Z(bb)
	Z(cc)
	Z(qq)
W	W(cq)
	W(qq)

Category	Label
Top	t(bcq)
	t(bqq)
	t(bc)
	t(bq)
QCD	QCD(bb)
	QCD(cc)
	QCD(b)
	QCD(c)
	QCD(others)

Table 4.2.1: DeepAK8 output classes.

The DeepAK8 tagger exploits the information of up to 100 PF jet constituents, sorted decreasing in p_T, and up to 7 SVs, sorted by the two-dimensional impact parameter significance. A total of 42 variables, based on kinematic and angular properties, for each particle are considered. For each SV, 15 features based on the displacement and tracks of charged particles are included to allow the extraction of features related to the presence of c or b quarks. Two one-dimensional CNNs are used to analyse the variables related to PF candidates and SVs, respectively. The CNN outputs are processed with a fully connected neural network to perform the jet classification. A schematic representation of the DeepAK8 architecture is shown in figure 4.2.3.

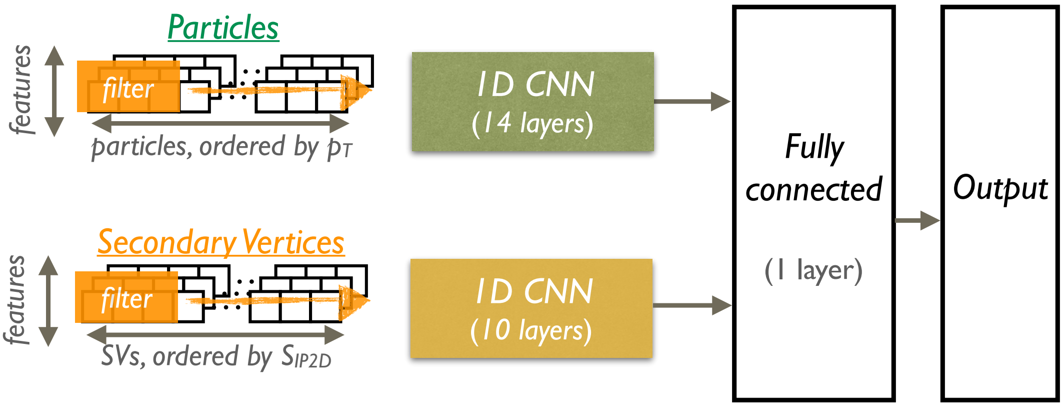


Figure 4.2.3: Architecture of the DeepAK8 neural network. Taken from Ref. [159].

ParticleNet

The DNN architectures presented before use inputs with a fixed dimension, which might result in a sparsely populated representation for a given jet under study. Moreover, the jet constituents are often sorted in an arbitrarily chosen order to remove unnecessary degrees of freedom.

An innovative and more natural way to represent a jet is via an unordered set of particles, also known as the "point-cloud" [175]. This approach comes with the advantage of a more flexible representation of any kind of features and a variable number of particles associated to each jet.

The ParticleNet tagger [176] uses a graph neutral network (GNN) to process the jets represented as point-clouds. Moreover, the *EdgeConv* operation [175] is used to extract and better exploit the correlation between the particles. The key idea of this approach is to capture the common structures between a point and its neighbours in the geometric and feature spaces; a more detailed description of this approach is given in Ref. [175, 176].

4.3 Mass decorrelation

It has been observed that the taggers described above feature a correlation of their output variables with the jet mass [159]. Since the mass of the jet is a powerful variable to distinguish jets originating from different physical processes, it is hardly surprising that these taggers tend to reconstruct the jet mass and use it as a distinctive feature. As a result, the jet mass distribution of background processes becomes more similar to that of the process under consideration after a selection made with any of these taggers.

This effect, also known as "mass sculpting", does not represent a problem per se unless the jet mass distribution is explicitly used in the analysis, e.g. to separate signal from background. As shown in figure 4.3.1, the BEST and the DeepAK8 algorithms lead to significant sculpting in the jet mass distribution after a selection on the tagger's output variables. A similar effect, but less pronounced, is present also for the algorithms using one-dimensional selection criteria based on the N-subjettiness ratios, e.g. τ_{21}, and energy correlation functions, e.g. N_2, (see section 3.3.2).

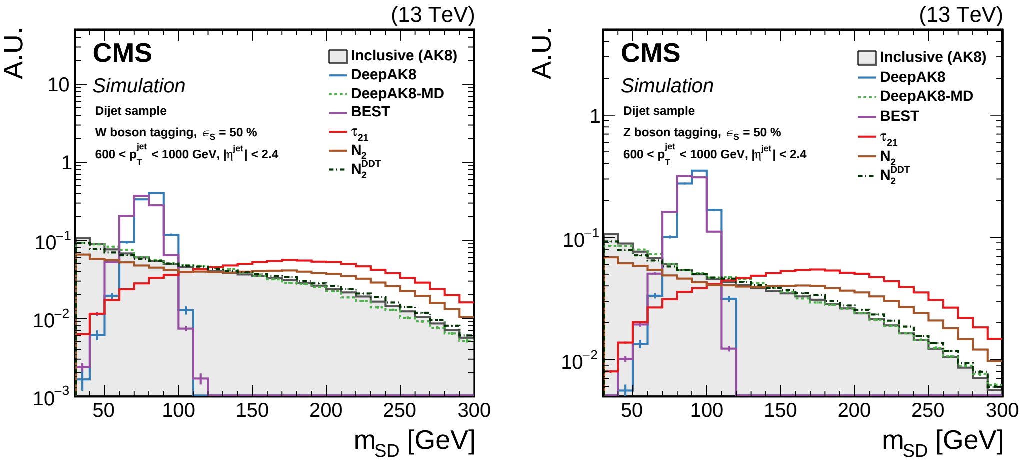


Figure 4.3.1: The normalised SD mass distribution for QCD-initiated jets before and after the selection using different algorithms for W (left) and Z (right) boson tagging. The different algorithms are descibed in the text. Taken from Ref. [159].

Mass-independent taggers are often more desirable, and various methods are adopted by the CMS Collaboration to reduce the correlation of the tagger output with the jet mass. For example, the designed decorrelated tagger (DDT) method [159, 177, 178] consists of the transformation of the tagging variable under consideration to ensure a selection with a constant background efficiency across the entire phase space considered. The tagger observables are usually transformed as a function of the jet p_{T} and the variable $\rho = \ln(m_{\mathrm{SD}}^2/p_{\mathrm{T}}^2)$ to ensure a fixed QCD rejection efficiency. This approach is powerful, although limited to the WP chosen and the phase space considered.

An alternative approach is based on the usage of an adversarial neutral network [179]. This method consists of a simultaneous training of the nominal DNN and a *mass-prediction* network; the latter is used to predict the jet mass only. The accuracy of the predicted mass is included as a penalty term in the classification loss function [173] of the nominal DNN to prevent a mass correlation. The DeepAK8 tagger exploits this approach to reduce the effect of the mass sculpting. Furthermore, events from different processes are also weighted to yield flat distributions in both p_{T} and SD mass in the training of the DNN.

Finally, the training of the ParticleNet tagger uses simulated events generated according to a uniform mass distribution in the $[15, 250]$ GeV range. As for the DeepAK8 algorithm, the events are additionally weighted to obtain a flat distribution in jet p_{T} and SD mass.

Examples of the mass sculpting for QCD-initiated jets for the mass-decorrelated version of the DeepAK8 and ParticleNet taggers are shown in figure 4.3.2. Both approaches provide a significant improvement in mass decorrelation. Furthermore, ParticleNet presents a jet mass distribution smoother than that of DeepAK8 and the adversarial training. Similar results are obtained performing the decorrelation of the DeepAK8 tagger with the DDT method [159]. These mass-decorrelated approaches usually come with a loss in tagging performance, as discussed in the next section.

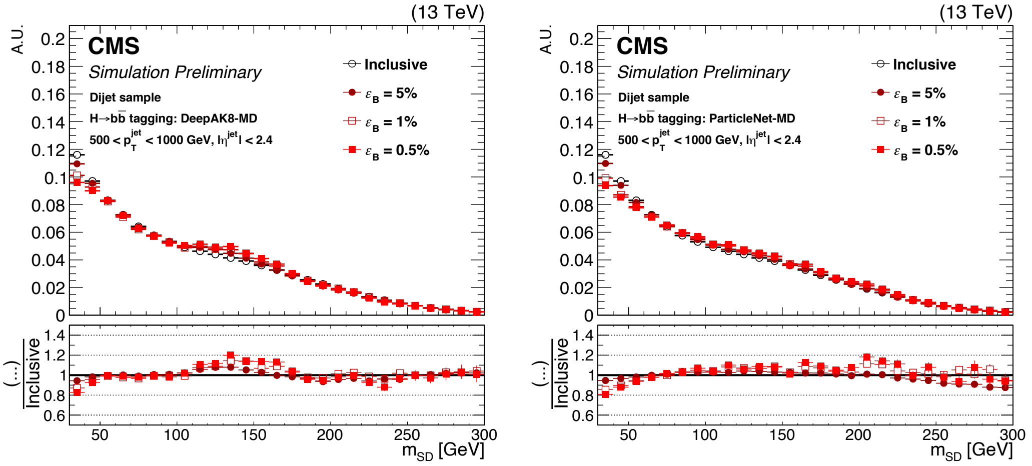


Figure 4.3.2: Shape of the SD mass distribution for QCD-initiated jets, inclusively and after different selections for the H → c$\bar{\text{c}}$ identification using the DeepAK8 (left) and the ParticleNet (right) taggers. Taken from Ref. [180].

4.4 Performance in simulation

The tagging performance of the different algorithms has been studied in simulated events [159]. The hadronic decays of heavy bosons and top quarks are considered as *signal*, while jets originating from gluons, and light-flavoured and b quarks in QCD multijet processes are treated as *background*. The receiver operating characteristic (ROC) curves are used as a figure of merit to evaluate and compare the different performances. These curves show the background efficiency (ϵ_B) as a function of the signal efficiency (ϵ_S). These efficiencies are defined as:

$$\epsilon_{\mathrm{X}} = \frac{\mathrm{N}_{\mathrm{X}}^{\mathrm{tagged}}}{\mathrm{N}_{\mathrm{X}}^{\mathrm{total}}}\,, \tag{4.1}$$

where N_X^{total} and N_X^{tagged} are the total number of jets of a given origin and the number of jets that satisfy a given selection criterion.

An example of the ROC curves for the top quark and H → c$\bar{\text{c}}$ tagging is shown in figure 4.4.1. Similar results are obtained for all decay modes considered. For the top tagging, several traditional algorithms are shown as a comparison. The new DNN-based algorithms outperform the previous methods, and the best discrimination is achieved with algorithms based on lower-level features. ImageTop and the mass-decorrelated version of DeepAK8 yield comparable performance, while the nominal version of DeepAK8 shows the highest tagging performance. Moreover, similar results are obtained for both the nominal and the mass decorrelated version of the ParticleNet tagger [180].

Similar conclusions can be derived also for the H → c$\bar{\text{c}}$ tagging. A slight reduction in performance is observed for the mass-decorrelated version of ParticleNet. Both versions of the ParticleNet tagger show similar or better performance than the nominal version of DeepAK8. The DDT method applied to the DeepAK8 tagger also shows good performance, improving with respect to the adversarial training, at the expense of less flexibility, as explained above.

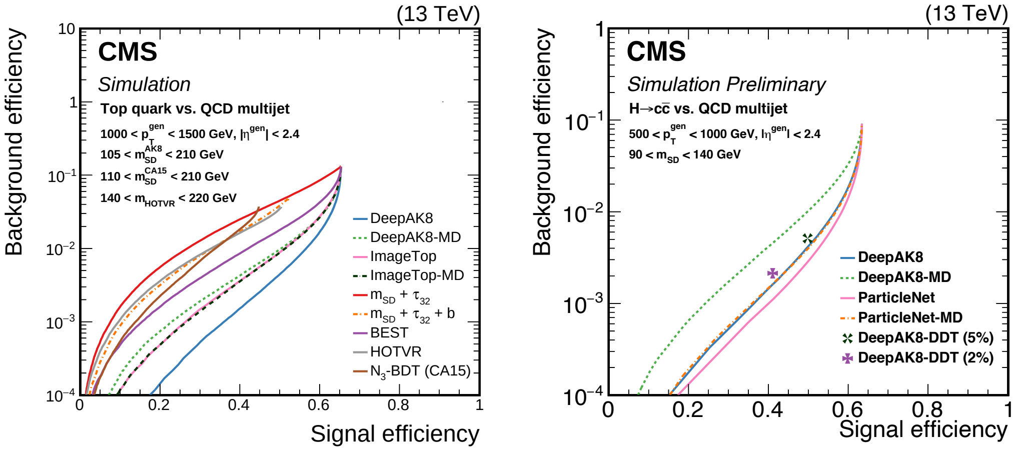


Figure 4.4.1: Comparison between the ROC curves for several algorithms for jet initiated by top quark (left) and $H \to c\bar{c}$ (right) decays. Taken from Ref. [159, 180].

4.5 Calibration in data

To properly use the taggers presented above in physics analyses, it is crucial to validate their performance and calibrate the efficiency in simulated events using real data. Potential differences between the performance in data and simulated events may arise since the jet taggers are trained on simulated samples only. Such discrepancies, due to the imperfect MC simulation of the jet (sub)structure, are corrected with data-to-simulation scale factors. The calibration strategy employed in the CMS Collaboration relies on the isolation of a dedicated phase space that matches or mimics the jet flavour under consideration.

The measurement of the top quark and W boson tagging efficiencies in data are derived with the "tag-and-probe" method [181] using events containing at least one muon, enriched in semi-leptonic $t\bar{t}$ events. Such events are selected by requiring the presence of a close-by muon and b-tagged jet. This system, used as a *tag*, is balanced in the opposite hemisphere by a large-radius jet, considered as a *probe*.

Moreover, different categories are defined for the SM $t\bar{t}$ simulated samples, depending on the angular separation of the parton-level hadronic decays of the top quark with respect to the large-radius jet. In particular, the *Merged t quark* and *Merged W boson* categories contain events in which the three partons from the top quark and only the two partons from the W boson decay are inside the jet, respectively. All other topologies are included in the *Nonmerged* category. Alternative matching categories, requiring only the angular separation between the jet and the parton-level initiating particle, are often used.

For each tagger, a simultaneous fit of the jet mass distributions in all categories is performed in the "pass" and "fail" categories, defined for a given WP for the tagger under consideration. Additional contributions from other SM background processes are considered as well. Several systematic uncertainties, discussed in more detail in Ref. [159], are considered as nuisance parameters in the fit. An example of the pre- and post-fit distributions for the mass-decorrelated version of the DeepAK8 tagger is shown in figure 4.5.1. Finally, the post-fit efficiencies in data and simulation are used to derive the data-to-simulation SF, which are typically consistent with unity

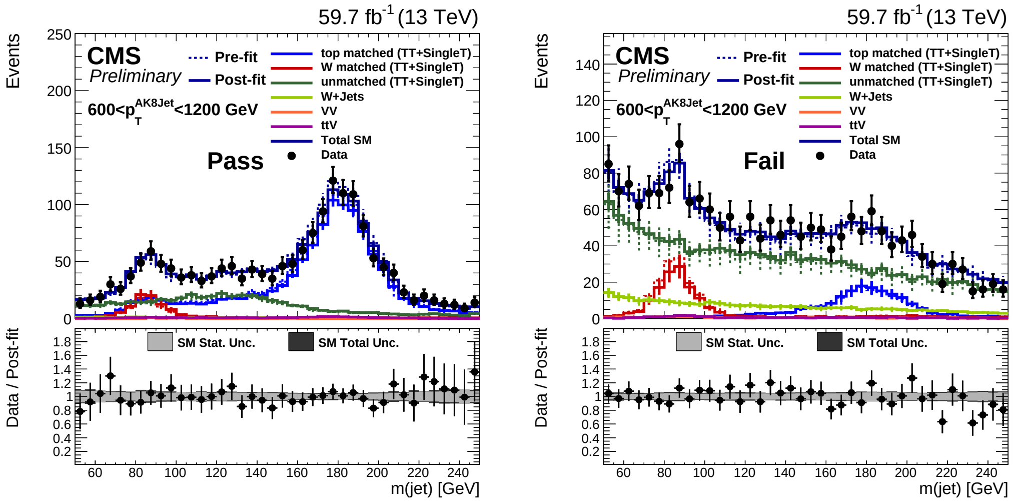


Figure 4.5.1: Jet mass distribution for data and simulated events in the pass (left) and fail (right) categories for the mass-decorrelated version of the DeepAK8 tagger. The WP corresponding to a misidentification rate of 1% is shown. Taken from Ref. [182].

with an uncertainty of 10-20%. The SFs for different WPs and p_T ranges are shown in figure 4.5.2 as an example.

A different method is used to calibrate the taggers involving the Z and Higgs bosons decays into a $b\bar{b}$ or $c\bar{c}$ pair, as it is challenging to isolate a pure sample of such final states in data. For this reason, the calibration of jets reconstructed from the clustering of such final states relies on the use of *proxy* jets, i.e. QCD-initiated jets resulting from the gluon splitting into a $b\bar{b}$ or $c\bar{c}$ pair. The approach exploits the large statistical precision available in QCD events. Therefore, a high-purity sample with characteristics similar to targeted jets can be selected, even for sub-dominant processes like $g \to c\bar{c}$. This technique is widely used within the CMS Collaboration [153, 183]. Moreover, it is crucial that the tagger under consideration does not strongly depend on the jet mass. For this reason, only the mass decorrelated versions of the DeepAK8 and ParticleNet taggers are considered for this approach.

A similar procedure compared to the W boson and top quark tagging case is used. The SM simulated samples are classified into three exclusive high-purity categories, defined in table 4.5.1. Events are selected from QCD dijet events using the *tag-and-probe* method. In particular, the *probe* jet has to be in a back-to-back topology with the *tag* jet and pass the selection criteria defined by the WP and the tagger under consideration. Finally, a simultaneous fit of the different categories is performed. The resulting data-to-simulation SFs for the $X \to c\bar{c}$ final state using the DeepAK8 tagger are shown in figure 4.5.3.

Category	Definition
b	Jet matched with at least 1 c-flavoured hadron
c	Jet matched with at least 1 c-flavoured and no b-flavoured hadrons
light	Jet not matched with any b- or c-flavoured hadrons

Table 4.5.1: Definition of the exclusive categories used for the calibration of the DeepAK8 and ParticleNet taggers.

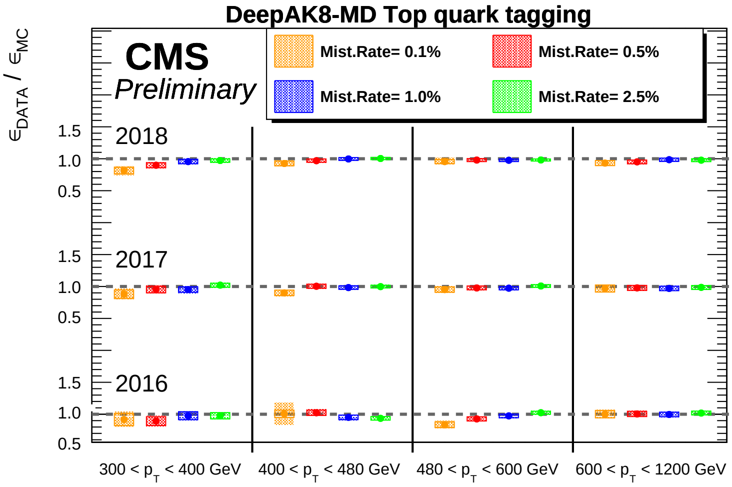


Figure 4.5.2: Data-to-simulation SFs for the mass-decorrelated version of the DeepAK8 tagger. The corrections, relative to the identification of top quark-initiated jets, are shown for different years of the Run 2 dataset and for different misidentification rates. Taken from Ref. [182].

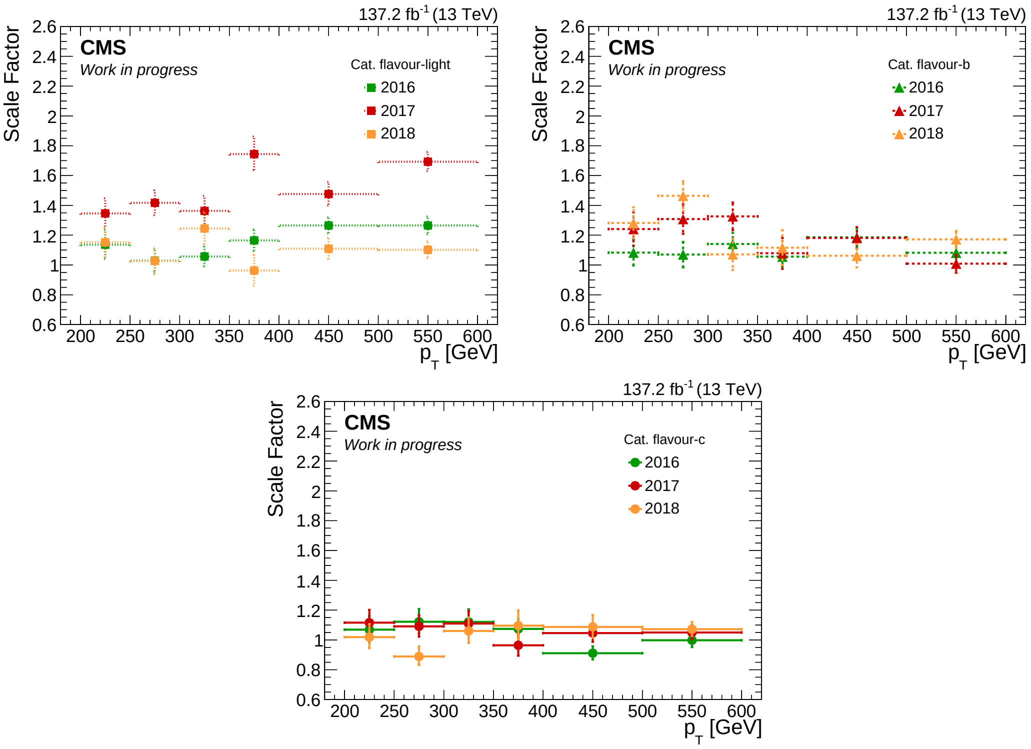


Figure 4.5.3: Data-to-simulation SFs for the mass-decorrelated version of the DeepAK8 tagger. The corrections, relative to the identification of Higgs boson-initiated jets, are shown for different years of the Run 2 dataset and for different decay modes. The flavours categories are defined in the text. Values taken from Ref. [184].

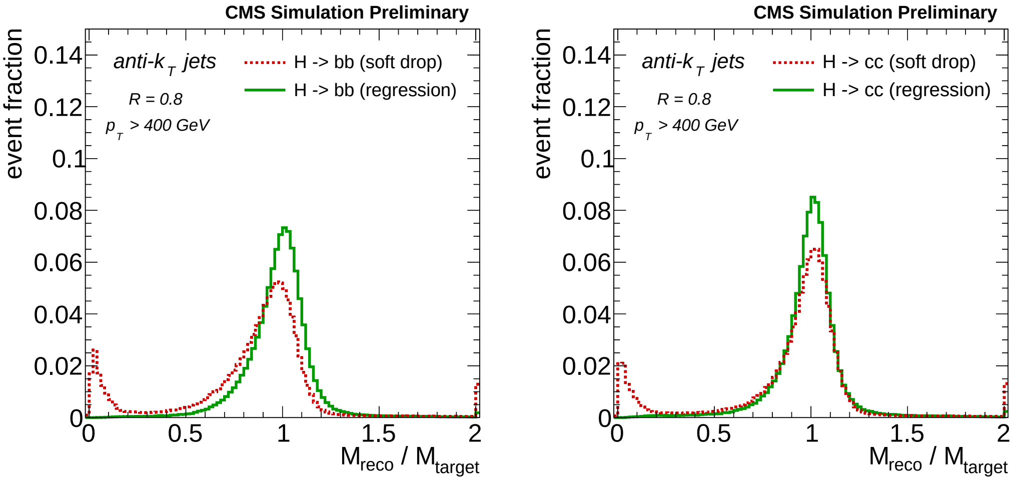


Figure 4.6.1: Performance of the ParticleNet regression (solid green) and the SD algorithm (dashed red) for $H \rightarrow b\bar{b}$ (left) and $H \rightarrow c\bar{c}$ (right) decays. Taken from Ref. [185].

4.6 Mass regression

As described above, the jet mass constitutes a very powerful observable for boosted jet tagging. It gives rise to a characteristic mass peak centred around the nominal mass of the initiating particle (top quark or heavy boson), while it shows a continuous spectrum for QCD-initiated jets. Moreover, the jet mass is sensitive to a potential contamination due to additional radiation, UE, or PU. Several grooming techniques, e.g. the SD algorithm, can be employed in the calculation of the jet mass. Despite the good performance achieved, as shown in figure 4.2.1 (left), this algorithm presents some limitations. One example is the loss of efficiency due to the small secondary peak around values of 0. This peak, caused by the too harsh grooming, is visible for all resonances and it is particularly important for $H \rightarrow b\bar{b}$ decays.

A novel technique based on ML has recently been developed to reconstruct the jet mass with the best possible resolution [185]. This mass regression algorithm uses the same inputs and a similar architecture employed for the training of the ParticleNet tagger. The focus is placed on the 2-prong hadronic decays of boosted heavy particles, e.g. $H \rightarrow b\bar{b}$ and $H \rightarrow c\bar{c}$.

This mass regression technique improves significantly both the jet mass scale and resolution, compared to the SD algorithm. This improvement is shown in figure 4.6.1 as the ratio of the reconstructed mass and the target mass. The target mass is the SD mass of the corresponding particle-level jet for QCD-initiated jets, and the generated particle mass otherwise. The results obtained with the ML-based regression provide a sharper peak and a more centred distribution. Moreover, the peak at low mass values is not present anymore. These improvements are consistent for all jet flavours. Finally, this method shows robust performance in terms of mass sculpting and stable results for a wide jet mass range [185].

The possible application of this mass regression algorithm is discussed at the end of chapter 6.

5 Jet transverse momentum resolution measurement

The detailed understanding of jet properties represents a crucial point under many aspects. Any Standard Model precision measurement or search for new physics relies on accurately calibrated jets. Besides, a miscalibration of the jet energy and resolution can lead to a momentum imbalance in the event and, consequently, to a mismeasurement of the missing transverse momentum. Therefore, a good understanding of jet properties, among these the jet transverse momentum resolution, is of significant importance. In this chapter, the strategy adopted by the CMS Collaboration for the measurement of the jet transverse momentum resolution is discussed in detail; particular emphasis is given to the derivation of the data-to-simulation scale factors. A similar strategy as that used in Ref. [148] is presented, highlighting the changes and improvements in the technique; an outlook to future challenges and possible developments is given at the end of the chapter. The results presented in this chapter have been derived entirely within the context of this thesis and have been published in Ref. [1]. They have been used by all CMS analyses published using the full Run 2 dataset. The results relative to the 2018 data-taking are shown as a showcase.

5.1 Introduction

Jets are the primary reconstructed physics objects used to infer properties of coloured particles produced in pp collisions. For a precise description of the energy and momentum of the initial particle, the reconstructed jets must be corrected for both the detector response and potential differences between data and simulation.

The transverse momentum of a reconstructed jet is not necessarily equal to that of the original particle, as already discussed in section 3.4. This effect is quantified by the jet transverse momentum response $\mathcal{R}$, which is defined in eq. (3.9) and reported here for simplicity:

$$\mathcal{R} = \frac{p_{\mathrm{T}}}{p_{\mathrm{T}}^{\mathrm{true}}}\,. \tag{5.1}$$

Here, p_{T} and $p_{\mathrm{T}}^{\mathrm{true}}$ denote the transverse momentum of the reconstructed and the particle-level jet, respectively. The average response is referred to as jet energy scale (JES), while its width is denoted as jet transverse momentum resolution (JER). Examples of the jet response distribution are shown in figure 5.1.1 for jets whose JES has already been calibrated.

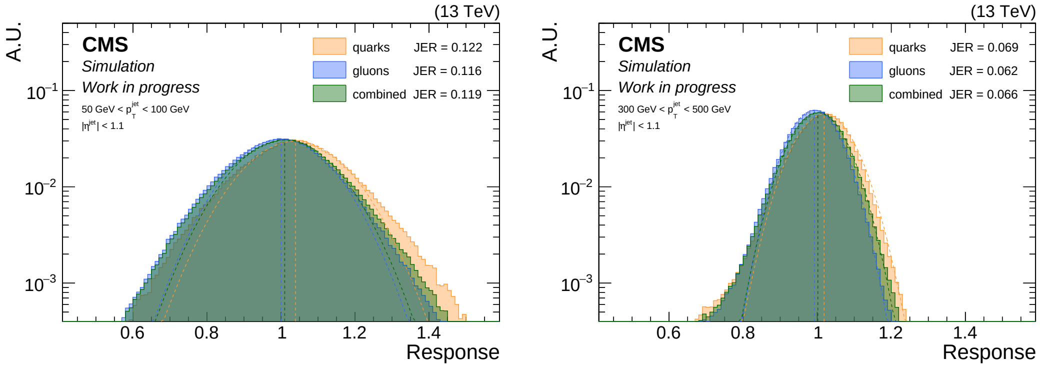


Figure 5.1.1: Normalised jet response distribution for low-p_T (left) and high-p_T (right) jets obtained from QCD multijet samples. The reconstructed fully-calibrated jet is matched with a particle-level jet with a $\Delta R < 0.2$ and split between quark- and gluon-initiated jets.

The response distribution has a Gaussian core, resulting from the intrinsic resolution of the various sub-detector components and the performance of the PF algorithm. The jet momentum is obtained from PF constituents, which are reconstructed by combining the tracking and calorimeter information. Jets reconstructed with the PF algorithm have better energy resolution than jets reconstructed purely relying on calorimetric information, whose energy is derived only from the calorimeter information. This improvement, shown in figure 5.1.2, is related to the evolution of the energy resolution of the subdetectors. The calorimeter resolution improves with increasing momentum; on the contrary, it is dominated by electronic noise and PU at low momentum. In contrast, the limitation in measuring the track curvature for energetic particles restrains the tracker resolution at high-p_T. Consequently, the p_T-dependence of jet resolution is the convolution of these effects. Moreover, the reconstruction performance of each subdetector is affected by the amount of PU. Besides, each component covers a specific η-range. Therefore, the jet resolution depends also on the pseudorapidity and the number of simultaneous interactions.

The response distributions in figure 5.1.1 also show non-Gaussian tails, which are caused by severe energy mismeasurements, due to noise effects, detector leakages from dead regions, or punch-through hadrons. At low-p_T, symmetric tails appear due to combinatorics where two generator-level jets produce a single reconstructed jet, and vice-versa.

The JER in simulated events is defined as the width of a Gaussian function obtained from the fit of the jet response distribution. The fitting range is chosen such that only the central 95% of the response are used in order to reduce the influence of the non-Gaussian tails. Also, the JER is parameterised with the NSC functional form expected for calorimeter-based resolutions (cf. eq. (2.2)). A slight modification of the original formula is used to account for the combined resolution of the PF algorithm:

$$\frac{\sigma_{p_\mathrm{T}}}{p_\mathrm{T}}(N,S,C,d) = \sqrt{\frac{N|N|}{p_\mathrm{T}^2} + \frac{S^2}{p_\mathrm{T}^d} + C^2}\,. \tag{5.2}$$

As evident from figure 5.1.2, the contribution from the calorimeter noise is reduced for low-p_T PF jets compared to the calorimeter-only-based jets. For this reason, a possible shape difference of the p_T-dependent contribution, compared to the nominal

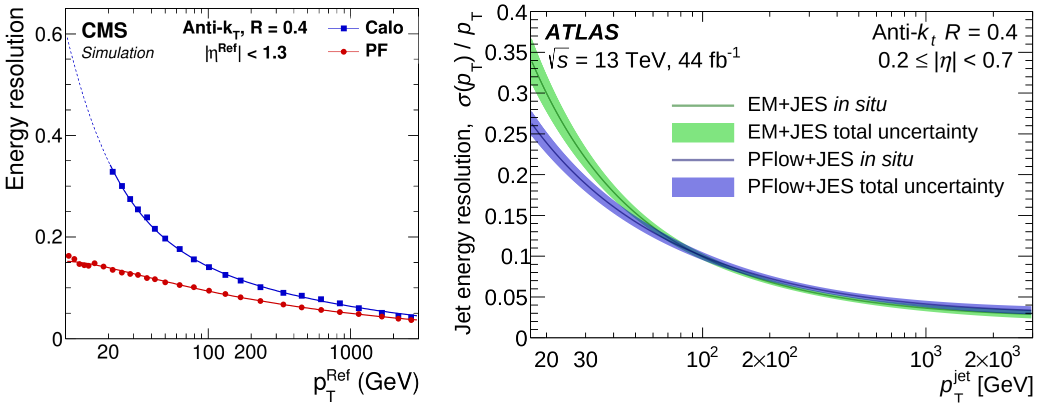


Figure 5.1.2: Jet resolution for CMS (left) and ATLAS (right) as a function of the p_{T} of reconstructed and calibrated jets. The PF algorithm [134, 187] shows improvements with respect to the calorimeter-only-based algorithm, especially in the low-p_{T} region. Thanks to the stronger magnetic field in the CMS detector, the tracking information improves the performance over a wider p_{T} range. Taken from Refs. [134, 188].

calorimeter-only case, is accounted for by allowing the parameter d to assume values different than 1 and the noise term N to have negative values to compensate for tracker effects.

Furthermore, the response depends on the jet flavour. Typically, quark jets consist of fewer and more energetic particles than gluon jets (see section 3.3.2) and thus have different distributions due to the non-linearity of the calorimeter response. In addition, the definition of particle-level jets does not include the energy carried by the neutrinos produced in semileptonic decays of heavy hadrons (see section 3.3); as a consequence, the heavy-flavoured jets show a similar resolution to the light-flavoured ones [148].

There is a multitude of cases in which the precise knowledge of the jet energy resolution is relevant. Primarily, it is used in the last step of the jet energy correction procedure adopted by the CMS Collaboration. In this context, the JER for simulated events is smeared to match the resolution in data, as described in section 5.4.2. Moreover, the JER represents an essential input to many physics analyses; for example, it is used to describe migration effects in differential jet spectra unfolding for jet cross section measurements. Another example is the usage in the prediction of background contributions to new physics searches in the $E_{\mathrm{T}}^{\mathrm{miss}}$+jet final states [186], where a mismeasurement of the jet energy leads to spurious missing transverse momentum. The JER is also used, together with the jet angular resolution, in the calculation of the missing transverse momentum significance [163]. Besides, it sets a benchmark criterion for the comparison of the performance of various jet reconstruction algorithms [96].

In the following, a detailed description of the smearing procedure and the derivation of the data-to-simulation SFs is provided. The primary method, used by the CMS Collaboration in Run 2, involves the p_{T}-balance in dijet events. This method covers a large phase space in p_{T} and η with high statistical precision. A complementary approach utilises the p_{T}-balance method in Z or γ+jet events [148]; the combination of these results with those obtained with the dijet method is detailed in section 5.8.

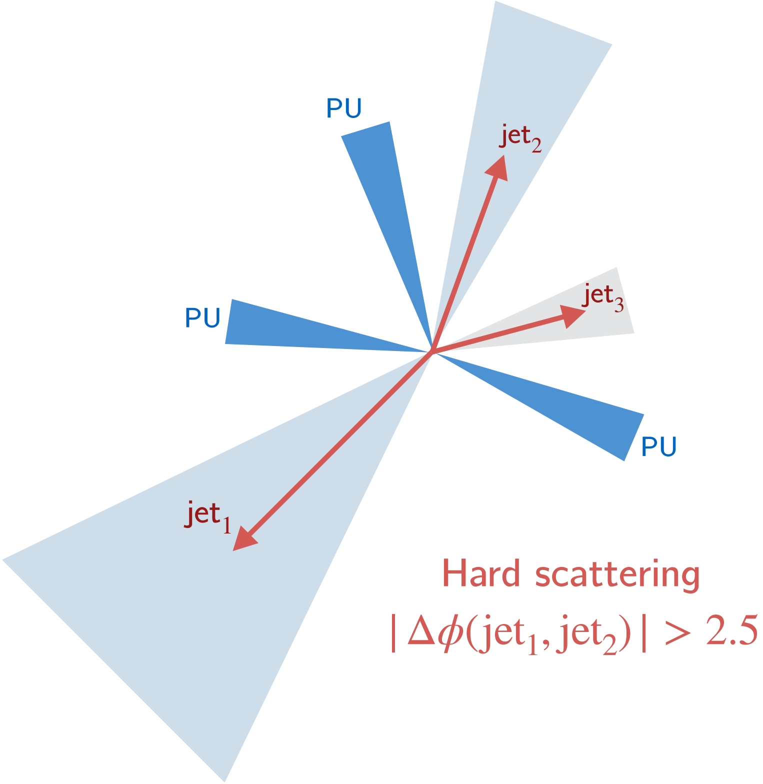


Figure 5.1.3: Sketch of dijet events used in the p_T-balance method.

5.1.1 Asymmetry method

The asymmetry method exploits the momentum conservation in the transverse plane of the dijet system; a sketch of a typical dijet event is shown in figure 5.1.3. It is an extension of the p_T-balance method used for the jet energy correction derivation, with the difference that the width of the response distribution is now the variable under consideration. Similarly to eq. (3.14), the asymmetry for events with at least two jets is defined as:

$$\mathcal{A} = \frac{p_{\mathrm{T},1} - p_{\mathrm{T},2}}{2\, p_\mathrm{T}^\mathrm{ave}}, \qquad \text{with} \qquad p_\mathrm{T}^\mathrm{ave} = \frac{p_{\mathrm{T},1} + p_{\mathrm{T},2}}{2}, \tag{5.3}$$

where $p_{\mathrm{T},1}$ and $p_{\mathrm{T},2}$ correspond to the transverse momenta of the two leading jets.

By definition, the asymmetry distribution is always positive and, ideally, symmetric around zero. However, as shown in section 5.9, this assumption is not always guaranteed due to miscalibration. Therefore, a new procedure with respect to Ref. [148] has been adopted, in which a random ordering for the two leading jets in p_T is chosen. As a consequence, the asymmetry distribution is double-sided.

Neglecting the non-Gaussian tails and for a sufficient number of events, the asymmetry is normally distributed with a standard deviation of:

$$\sigma_\mathcal{A} = \left(\left| \frac{\partial \mathcal{A}}{\partial p_{\mathrm{T},1}} \right| \cdot \sigma(p_{\mathrm{T},1}) \right)^2 \oplus \left(\left| \frac{\partial \mathcal{A}}{\partial p_{\mathrm{T},2}} \right| \cdot \sigma(p_{\mathrm{T},2}) \right)^2 . \tag{5.4}$$

In an ideal dijet topology, the two jets are exactly balanced at the particle level. Moreover, if they belong to the same $|\eta|$ region, it can be assumed that $\langle p_{\mathrm{T},1} \rangle \sim \langle p_{\mathrm{T},2} \rangle$

and $\sigma(p_{\mathrm{T},1}) \sim \sigma(p_{\mathrm{T},2})$. This assumption allows the simplification of eq. (5.4) and provides the following important relation between the width $\sigma_{\mathcal{A}}$ of the asymmetry and the jet p_{T} resolution $\sigma(p_{\mathrm{T}})$:

$$\sigma_{\mathrm{JER}} \equiv \frac{\sigma(p_{\mathrm{T}})}{\langle p_{\mathrm{T}} \rangle} = \sqrt{2} \cdot \sigma_{\mathcal{A}} \,. \tag{5.5}$$

This formula was already used at the Tevatron experiments [189, 190], the ATLAS experiment [191] and in previous CMS publications [148] to measure the jet resolution from dijet events.

Forward extension method

The requirement that the two leading jets belong to the same $|\eta|$ region reduces the number of events available in the forward region of the detector significantly. In fact, the jets in this region have low p_{T} and, therefore, are triggered by the highly prescaled triggers (see section 2.2.6). As a consequence, the statistical precision is already limited for pseudorapidities beyond $|\eta| = 2.0$. For this reason, the data-to-simulation ratio could be determined with the standard approach only in wide intervals of $|\eta|$ and with large uncertainties.

In order to extend the analysis to the detector's forward region, a forward extension (FE) of the asymmetry method is used instead, which allows the two leading jets to have different $|\eta|$. In this case, the relation expressed in eq. (5.5) is no longer valid, since $\sigma(p_{T,1}) \neq \sigma(p_{T,2})$ if $|\eta_1| \neq |\eta_2|$. Nonetheless, the jet resolution $\sigma(p_{\mathrm{T}}^{\mathrm{probe}})$ in a given region under study, $|\eta_{\mathrm{probe}}|$, can be determined if the resolution $\sigma(p_{\mathrm{T}}^{\mathrm{ref}})$ in a reference region $|\eta_{\mathrm{ref}}|$ is known. Therefore, a slightly modified definition of the asymmetry is used:

$$\mathcal{A} = \frac{p_{\mathrm{T}}^{\mathrm{probe}} - p_{\mathrm{T}}^{\mathrm{ref}}}{p_{\mathrm{T}}^{\mathrm{probe}} + p_{\mathrm{T}}^{\mathrm{ref}}} \,. \tag{5.6}$$

Therefore, the probe jet resolution is then given by:

$$\sigma_{\mathrm{JER}} \equiv \frac{\sigma(p_{\mathrm{T}}^{\mathrm{probe}})}{\langle p_{\mathrm{T}}^{\mathrm{probe}} \rangle} = \sqrt{4 \cdot \sigma_{A} - \left(\frac{\sigma(p_{\mathrm{T}}^{\mathrm{ref}})}{\langle p_{\mathrm{T}}^{\mathrm{ref}} \rangle} \right)^2} \,. \tag{5.7}$$

The resolution of the reference jet is assumed to be known from the standard method derivation; moreover, the reference jet is chosen to be within the barrel region ($|\eta|$ <1.3), where the uncertainty is smaller.

5.1.2 Realistic dijet events

Similar to the jet energy response and resolution for simulated events, the asymmetry is derived in different p_{T} and $|\eta|$ intervals. The trigger thresholds dictate the p_{T}-binning, and the exact values are discussed in section 5.2. The η-binning is derived based on the detector geometry and uniformity in the subdetector response [192].

As already discussed in section 3.4.3, there is an intrinsic bias caused by the finite jet transverse momentum resolution and by the monotonically decreasing p_{T} spectrum. In fact, a binning based on single jet momentum is not suitable due to the migration effects causing jets with lower $p_{\mathrm{T}}^{\mathrm{true}}$ to fluctuate more often than jets with higher $p_{\mathrm{T}}^{\mathrm{true}}$. This effect can be reduced using bins of $p_{\mathrm{T}}^{\mathrm{ave}}$, in which case the bias is cancelled out on average.

In realistic collision events, part of the momentum of the hard scattering process is transferred to soft particles or jets arising from initial or final state radiation, which can lead to a momentum imbalance in the dijet system. This additional jet activity can be described in good approximation by the variable α, which is defined as the ratio of the transverse momentum of the third most energetic jet[1] to the average transverse momentum of the first two leading jets:

$$\alpha = \frac{p_{\mathrm{T},3}}{p_{\mathrm{T}}^{\mathrm{ave}}}\,. \tag{5.8}$$

This additional jet activity causes a broadening of the asymmetry distribution; therefore, this effect must be corrected in order to determine the intrinsic resolution.

Another difference between the particle-level and the detector-level jets is arising from out-of-cone showering effects. Typically, some particles might be too soft to be included in the clustered jet or even be wrongly associated to a jet. Such effects lead to an overall momentum imbalance in an event and need to be corrected for, as further discussed in section 5.4.

5.2 Data and simulated events

This analysis uses events from pp collisions, which have been recorded during the Run 2 with the CMS detector at $\sqrt{s} = 13\,\mathrm{TeV}$. Multijet events are recorded with a set of triggers based on the jet transverse momenta. In particular, the p_{T} average of the two leading jets is used for the online requirement on the jet momentum. Moreover, two sets of triggers are designed to target specifically the central and forward region of the detector. Triggers requiring central jets are used up to $|\eta| \leq 2.853$, while forward triggers are used otherwise. Triggers based on the dijet system are available only for small-radius jets; therefore, triggers requiring a single jet with p_{T} above a given threshold are used when repeating the measurement for large-radius jets. Moreover, dijet triggers were not active at the beginning of the 2017 data-taking period; therefore, the single jet triggers were used instead. The online and offline p_{T} threshold for the triggers used in this analysis are reported in table 5.2.1; the offline requirement corresponds to the value at which each trigger reaches the 99% efficiency plateau. The 2018 thresholds are indicated for dijet and single AK8 jet triggers as a showcase, while the 2017 values are listed for the single AK4 jet triggers. This p_{T}-binning is used later in the analysis.

The QCD simulated samples used in this analysis are generated with MadGraph5_amc@nlo [128], while the parton-shower and hadronisation processes modelled with pythia8 [123]. These samples are generated at NLO precision and provide high statistical precision in the whole p_{T}-η phase space. The pythia8 tune used in 2016 was CUETP8M1 [193], while the CP5 tune [194] was used for all other years, including the 2016 Legacy version.

Simulated events are reweighted in order to match the number of additional pp collisions to that in data, assuming a minimum bias cross section of $69.2 \pm 4.6\%$ mb.

[1] The presence of jets beyond the third jet is neglected in the parametrization of the additional activity in the event as these jets bring an even smaller contribution.

Region	Dijet		Single AK8 jet		Single AK4 jet 2017	
	Online	Offline	Online	Offline	Online	Offline
	$p_{\mathrm{T}}^{\mathrm{ave}}$ thr.	$p_{\mathrm{T}}^{\mathrm{ave}}$ thr.	$p_{\mathrm{T}}^{\mathrm{jet}}$ thr.	$p_{\mathrm{T}}^{\mathrm{ave}}$ thr.	$p_{\mathrm{T}}^{\mathrm{jet}}$ thr.	$p_{\mathrm{T}}^{\mathrm{ave}}$ thr.
Central	40	66	40	70	40	70
	60	93	60	78	60	87
	80	118	80	96	80	111
	140	189	140	119	140	180
	200	257	200	193	200	247
	260	325	260	262	260	310
	320	391	320	328	320	373
	400	478	400	393	400	457
	500	585	450	481	450	510
			500	534	500	562
			550	588		
Forward	60	93	60	62		
	80	116	80	95		
	100	142	140	110		
	160	210	200	182		
	220	279	260	260		
	300	379	320	339		
			400	420		
			500	508		

Table 5.2.1: Single jet and dijet HLT trigger path with their online and offline p_{T} threshold. The threshold values for dijet and single AK8 jet triggers are reported for the year 2018, while the values for the central single AK4 jet triggers refer to the year 2017.

5.3 Event selection

Data and simulated events are reconstructed with the PF algorithm. The selection starts by requiring the event to have fired one of the triggers described above and to have at least one well-reconstructed PV.

An offline event selection is applied to select QCD dijet events. This analysis uses jets reconstructed with the anti-k_{T} algorithm and a radius parameter of $R = 0.4$.

The PF candidates associated to pileup vertices are removed from the jet constituents using the CHS algorithm [134]. The results presented in the following refers to such jets, although similar results and conclusions can also be applied to other jet radii and PU subtraction algorithms.

The jet energy scale of all jets in the event, in both data and simulation, is calibrated using the L1L2L3 corrections. The L2L3Residual corrections are of particular importance in this measurement, as as they are strongly correlated to the JER. A more detailed description of their correlation is discussed in section 5.7.

Dijet events are then selected by requiring the presence of at least two back-to-back jets ($\Delta\phi > 2.7$). Reconstructed jets with p_{T} >15 GeV are considered; at the particle level, jets with transverse momentum above 10 GeV are considered.

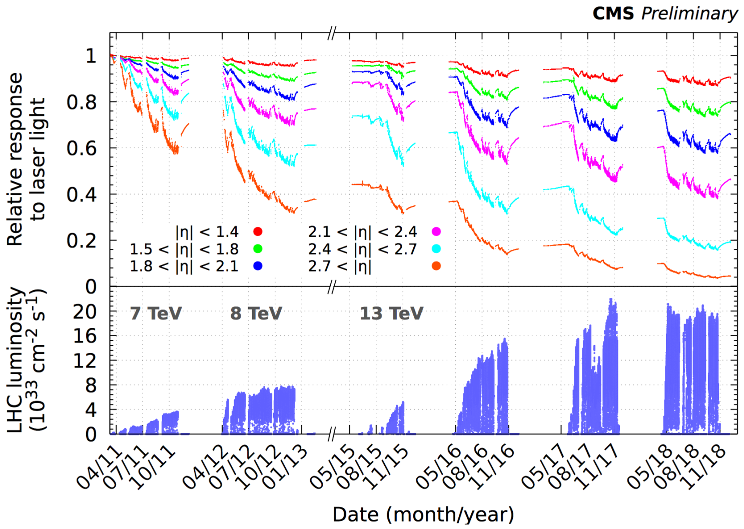


Figure 5.3.1: Evolution of the ECAL response versus time for different ranges of $|\eta|$. The delivered instantaneous luminosity is shown in the bottom panel. Taken from Ref. [195].

Additional cleaning

Because this analysis targets QCD dijet events, a veto on isolated leptons that pass loose identification criteria is imposed. The cut-based and the BDT-based IDs are used for the muons [141] and electrons [143], respectively. Additionally, the tight ID is applied to each jet and the PU ID for jets with $p_T < 50\,\text{GeV}$. Moreover, a series of selection criteria on the missing transverse momentum are applied to both data and simulation in order to suppress events with spurious ill-reconstructed quantities.

The ultimate goal of the resolution smearing procedure is to correct the mismodelling of the detector response in simulation. A crucial point is to mask locally malfunctioning detector areas that might occur during data-taking. Events that are affected by such cases are discarded, or their impact is mitigated, for the nominal SF derivation procedure. These effects can be grouped by source, namely radiation damage and electronic failure. In the following, the issues that had the most significant impact are quickly described as a showcase.

In general, any detector exposed to a high level of radiation manifests a decrease in performance over time due to radiation damage causing, for instance, the ECAL crystals to darken. As shown in figure 5.3.1, the response of the crystal decreases over time. The transparency loss is moderate in the central part of the detector and more critical for the crystals in the forward region, where a higher flux of ionising particles occurs.

One of the most dramatic consequences of the transparency loss of the ECAL crystals is a slowly developing time shift in the ECAL pulses. This effect, which was not entirely corrected at the trigger level until 2018, produces a gradually increasing fraction of ECAL-triggered events that had been wrongly associated to an energy deposit of the previous bunch crossing. The L1 trigger system was therefore caused to “prefire” [120], i.e. to accept the earlier collision instead of the collision of interest, while the subsequent trigger chain and offline selections might reject such events. The main consequence of prefiring is, therefore, an inefficiency in recording potentially interesting events. This effect affects primarily the region at $|\eta| > 2.5$, and it is

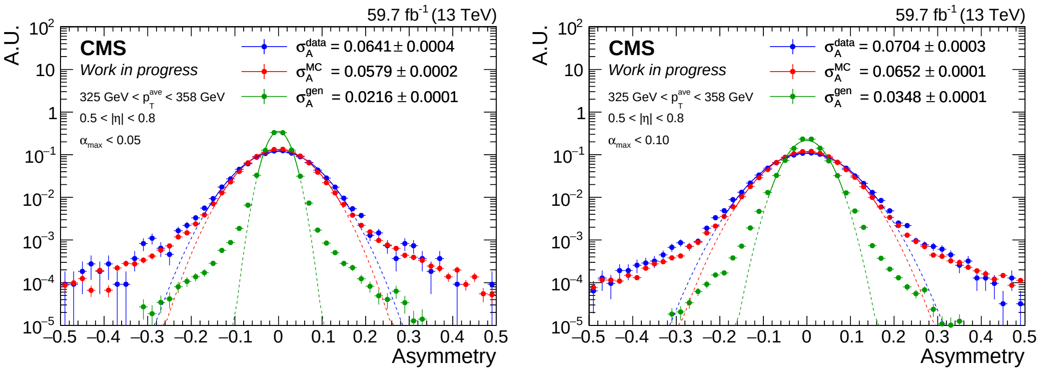


Figure 5.4.1: Normalised asymmetry distribution for two values of $alpha_{\max}$ in data (blue), simulated events (red) and particle level (green) jets. The Gaussian fit to the core of the distribution is shown as a solid line excluding 1.5% of the total area on both sides. The width, calculated as the RMS, and its error are reported for each distribution.

corrected at analysis-level with ad-hoc corrections, where the simulated events are scaled down to emulate the efficiency loss.

Moreover, during the 2018 data-taking, two endcap sectors of the HCAL were not functional. The region of $\eta \in [-2.96, -1.31]$ and $\phi \in [-1.57, -0.87]$ is affected by this outage. There are two major implications for the reconstructed objects in the event, given the unmeasured energy in that region. A jet passing through this region will either have a heavily mismeasured energy or even be completely misidentified as an electron. As a consequence, spurious missing transverse momentum is created. Events with jets and electrons in the affected region are not considered to avoid any bias in the event.

5.4 Scale factor derivation

The JER scale factors are derived in bins of $|\eta|$ for both the standard and the FE methods utilising a similar procedure. The standard method requires both leading jets to be in the same $|\eta|$ bin, while for the FE method, the reference jet is always inside the barrel ($|\eta|$ <1.3). For the latter case, the SFs are derived as a function of $|\eta|$ of the probe jet.

Asymmetry width

The width of the asymmetry distribution is estimated from its Gaussian core to avoid biasing the measurement through the non-Gaussian tails. Hence, the asymmetry width is defined as the root mean square (RMS) of the distribution truncated at a certain percentage of the tail regions. The outer 1.5% of the distribution are removed on each side.

This procedure is applied to data as well as simulated events; both reconstructed and particle-level jets are analysed for the latter. An example of the asymmetry distribution is shown in figure 5.4.1. Evidently, the width of the asymmetry is larger in data compared to simulated events. Moreover, all distributions are wider as the extra jet becomes more energetic compared to the p_T^{ave}. This effect is discussed in the next section.

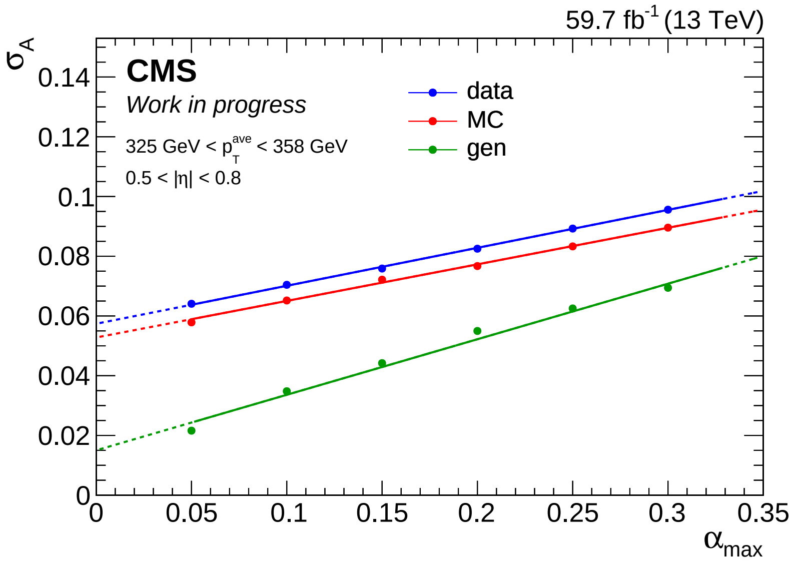


Figure 5.4.2: Extrapolation to zero additional jet activity for data (blue), simulated events (red) and particle level (green) jets.

Correction for additional jet activity

The relations between the width of the asymmetry distribution and the jet energy resolution, as expressed in equations (5.5) and (5.7), hold only for the case of an ideal dijet topology. However, as shown in figure 5.4.1, the contribution from additional jets increases the measured asymmetry width. In order to determine the intrinsic asymmetry width, this imbalance contribution is removed with the following extrapolation procedure. The asymmetry distribution is calculated for each interval in $|\eta|$ and $p_{\mathrm{T}}^{\mathrm{ave}}$ with different thresholds on the maximum value of α (α_{max}). The measured values for $\sigma_{\mathcal{A}}$ empirically show a linear behaviour as a function of α_{max}. Therefore, the value of $\sigma_{\mathcal{A}}$ is extrapolated to the case without any jet activity in the event ($\alpha_{\mathrm{max}} \to 0$). A fit to the measured values is performed:

$$\sigma_{\mathcal{A}}(\alpha_{\mathrm{max}}) = a + b \cdot \alpha_{\mathrm{max}} \,, \tag{5.9}$$

where a is assumed to be the best estimate for $\sigma_{\mathcal{A}}(\alpha_{\mathrm{max}} \to 0)$. The bins in α_{max} are inclusive by constructions, and their correlation is accounted for in the extrapolation fit to estimate the statistical uncertainty correctly. A detailed description of the correlated fit can be found in Ref. [196]. An example of the extrapolation procedure is shown in figure 5.4.2.

Particle level imbalance

Another effect that causes an imbalance in dijet events can arise from out-of-cone showering and the underlying event. Some particles created during the fragmentation and hadronisation processes might be produced with particularly low momentum or an overly wide angular separation, and therefore are not clustered into the jet. At the same time, other particles, originating from other physical processes, e.g. PU and the underlying event that are unrelated to the original parton, might be clustered inside

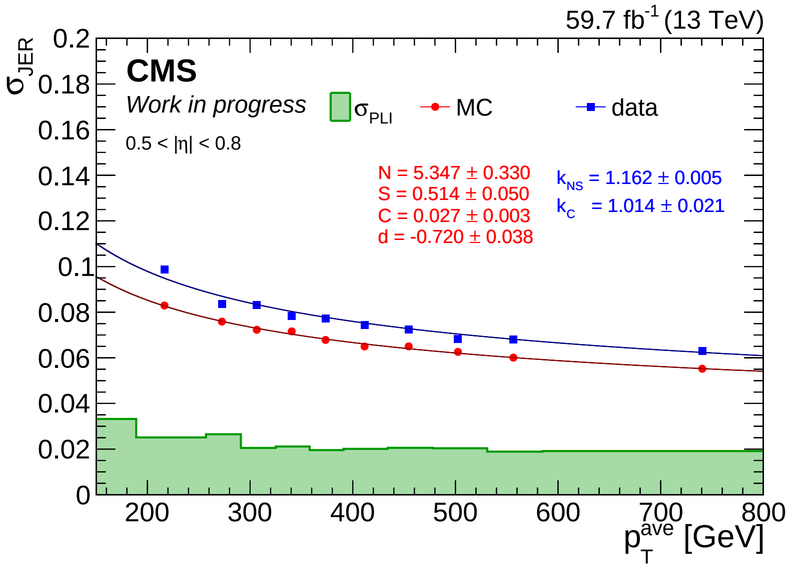


Figure 5.4.3: Jet transverse momentum resolution as a function of jet p_T after the PLI (green) subtraction. More details can be found in the text.

the jet. As a consequence, a broadening of the asymmetry distribution is observed, caused by the additional p_T imbalance.

This effect is visible already at the particle level and is present also in the derivation of the L2L3Residual corrections, where is corrected for in the global fit (see section 3.4.3). For the JER SF derivation, this additional imbalance is referred to as particle-level imbalance (PLI) and its contribution is estimated from the total asymmetry width of particle-level jets as:

$$\sigma_{\text{PLI}} = \sqrt{2}\sigma_{\mathcal{A}}^{\text{gen}}(\alpha_{\text{max}} \to 0)\,. \tag{5.10}$$

The PLI, assumed to be equal in data and simulation, is subtracted in quadrature from the asymmetry width $\sigma_{\mathcal{A}}$ obtained from the α-extrapolation. The impact on the PLI due to the simulation of the hadronisation process of different MC generators is considered as a systematic uncertainty and discussed in section 5.5.

An example of σ_{JER} for data and simulated events after the PLI (green) subtraction is shown in figure 5.4.3. Moreover, the resulting σ_{JER} distribution is well-described with eq. (5.2). These fits, discussed in more detail in section 5.5, are used to derive a systematic uncertainty.

5.4.1 Data-to-simulation ratio

After correcting for the effects discussed above, the data-to-simulation ratio is calculated as:

$$c_{\text{res}} = \frac{\sigma_{\text{JER}}^{\text{DATA}}}{\sigma_{\text{JER}}^{\text{MC}}}\,. \tag{5.11}$$

The resulting distribution is evaluated in each $|\eta|$ bin as a function of p_T^{ave}, as shown in figure 5.4.4. The ratio is independent of p_T^{ave} to good approximation and parametrised with a constant. Similar results, used as cross-check, are obtained when

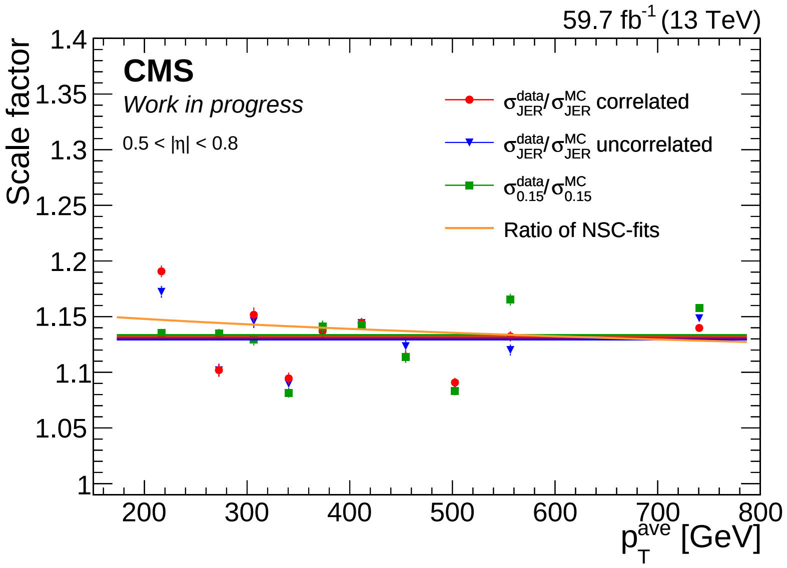


Figure 5.4.4: Data-to-simulation ratio of σ_{JER} as a function of $p_{\mathrm{T}}^{\mathrm{ave}}$. The points obtained under the assumption of (no) correlation between $\alpha_{\max}$ bins are shown in red (blue). The green points are obtained using $\alpha_{\max} = 0.15$, and no extrapolation has been considered. The red fit corresponds to the nominal results, while the other are used as a cross-check. The orange curve is the data-to-simulation ratio of the NSC-fits.

deriving the SF under the assumption of uncorrelated $\alpha_{\max}$ bins and when the value of $\alpha_{\max} = 0.15$ is used with no extrapolation procedure. Moreover, a potentially neglected p_{T}-dependance is accounted for as a systematic uncertainty, derived from the data-to-simulation ratio of the NSC-fits shown in figure 5.4.2.

In each $|\eta|$ region and p_{T} range, the ratio is above unity, confirming that the resolution in data is systematically broader than in simulation. The convolution of the limited parton-shower model of the MC generators, the imperfect simulation of the detector response, and a residual miscalibration of the CMS calorimeters are possible causes of this imperfect modelling.

5.4.2 Smearing procedure

The simulated jet resolution is now broadened ("smeared") to accurately describe the data. The scaling and smearing methods are described and discussed in the following.

Stochastic method

The *stochastic* method constitutes the most straightforward approach, easily applicable to every jet. The method is constructed upon the idea of adjusting the jet response in simulation to match the one observed for data by convolving it with a Gaussian function of width σ_c to emulate the detector effects. The resolution of the convolved distribution is known to be:

$$\sigma_{\mathrm{JER}}^{\mathrm{MC,\,corr}} = \sigma_{\mathrm{JER}}^{\mathrm{MC}} \oplus \sigma_c\,, \tag{5.12}$$

where $\sigma_{\mathrm{JER}}^{\mathrm{MC}}$ represents the width of the Gaussian core for simulated events.

Under the assumption that the response has a Gaussian core also in data, it is possible to require that the corrected resolution in simulation is equal to the resolution in data, i.e. $\sigma_{\mathrm{JER}}^{\mathrm{MC,\,corr}} = \sigma_{\mathrm{JER}}^{\mathrm{DATA}}$, obtaining:

$$\sigma_c = \sigma_{\mathrm{JER}}^{\mathrm{MC}} \sqrt{c_{\mathrm{res}}^2 - 1}\,. \tag{5.13}$$

Finally, a random number is sampled from a Gaussian function $\mathcal{N}(0, \sigma_c)$ and it is used to smear the simulated jet momentum.

The limitations of this method are threefold. First, the method is limited to have $c_{\mathrm{res}} > 1$, allowing only to degrade the resolution. Second, the application of the smearing procedure can be not reproducible unless the random seed is kept fixed. Last, the randomness of the method can alter the jet balance in the event, artificially creating spurious $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$. In order to mitigate these effects, the scaling method is introduced.

Scaling method

In case the match between particle-level and reconstructed jet is possible, the *scaling* method can be applied to achieve a more accurate description of the data. In this case, the reconstructed jet p_{T} is shifted even further from the true value of the particle-level jet ($p_{\mathrm{T}}^{\mathrm{true}}$). Hence, the corrected jet p_{T} in simulation is obtained:

$$p_{\mathrm{T}}^{\mathrm{MC,\,corr}} = p_{\mathrm{T}}^{\mathrm{true}} + c_{\mathrm{res}} \left(p_{\mathrm{T}}^{\mathrm{MC}} - p_{\mathrm{T}}^{\mathrm{true}} \right)\,, \tag{5.14}$$

This method can be used only in the presence of a well-matched particle-level jet, leading to a large shift of the response otherwise. Therefore, the following requirements are usually imposed for the matching:

$$\Delta R < R_{\mathrm{cone}}/2 \qquad \text{and} \qquad \frac{|p_{\mathrm{T}}^{\mathrm{MC}} - p_{\mathrm{T}}^{\mathrm{true}}|}{p_{\mathrm{T}}^{\mathrm{MC}}} < 3\,\sigma_{\mathrm{JER}}^{\mathrm{MC}}\,, \tag{5.15}$$

where ΔR represents the angular separation between the reconstructed and particle-level jets, and R_{cone} is the jet cone size (e.g. 0.4 for AK4 jets).

This method solves all the limitations encountered with the stochastic method; in particular, the p_{T} balance in multijet events is preserved and non-spurious $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ is additionally created.

In case one of the two conditions is not met, i.e. for a PU jet or if the energy of the reconstructed jet is far beyond the Gaussian approximation, the stochastic method can be applied. The combination of these two approaches has been verified to provide the most accurate and robust calibration of the jet energy resolution.

5.5 Systematic uncertainties

The steps of the JER SF derivation procedure are subject to systematic uncertainties related to the assumptions made or procedures employed. Each source is described in the following and their impact is evaluated as the shift (δc) of the data-to-simulation ratio related to a certain variation (Δ) in the measurement procedure:

$$\delta c_{\mathrm{res}} = c_{\mathrm{res}}^{\Delta} - c_{\mathrm{res}}^{\mathrm{nominal}}\,. \tag{5.16}$$

Pileup

A pileup reweighting procedure is used in simulated samples to match the observed pileup distribution in data. In order to quantify the impact of the pileup modelling, the SF derivation is repeated by varying the minimum bias cross section within its uncertainty of ±4.6% mb.

Jet energy scale uncertainty

The JES of all jets used in the analysis has been corrected to the particle level. All jet momenta in the simulated samples are shifted up and down by the JEC uncertainty as a function of p_{T} and η of each jet. Afterwards, the data-to-simulation ratio is re-determined based on the altered jet momenta.

Non-Gaussian tails

The width of the asymmetry distribution is calculated as a truncated RMS to reject the contributions of the non-Gaussian tails. For the nominal scale factor derivation, 1.5% are excluded symmetrically from each side of the distribution. In general, the tail contributions can differ between data and simulation and, consequently, do not necessarily cancel out in the ratio. In order to estimate a systematic uncertainty for this effect, the data-to-simulation ratio is re-evaluated by excluding 2.5% of the tails.

Correction for additional jet activity

The contribution of additional jet activity in the event is removed from the measured asymmetry widths via the extrapolation procedure. A linear behaviour of the widths as a function of $\alpha_{\max}$ is assumed. This choice implies that the linear behaviour also holds at small values of α, which, however, cannot be tested explicitly. A systematic uncertainty is estimated by lowering the minimum jet p_{T} threshold from 15 GeV to 10 GeV.

Particle-level imbalance

The measured resolution is corrected for the imbalance at the particle level. It has been measured that the PLI contribution changes from different generators [148]; differences between simulated samples generated with PYTHIA8 and HERWIG++ are up to 25%. Therefore, the PLI correction factor is shifted by ±25%. The obtained uncertainty is asymmetric by construction, as the PLI correction is subtracted in quadrature. This procedure is applied simultaneously to data and simulated events.

p_{T}-dependence

The $\sigma_{\mathrm{JER}}^{\mathrm{MC}}$ is generally described well by the NSC function expressed in eq. (5.2). Under the assumption that the JER SFs do not have a p_{T}-dependence, there should be one common multiplicative scale factor k_{NSC} for the N, S and C parameters, which allow to describe the $\sigma_{\mathrm{JER}}^{\mathrm{DATA}}$ points with the same function. On the other hand, differences between data and simulation can result in different shapes of the

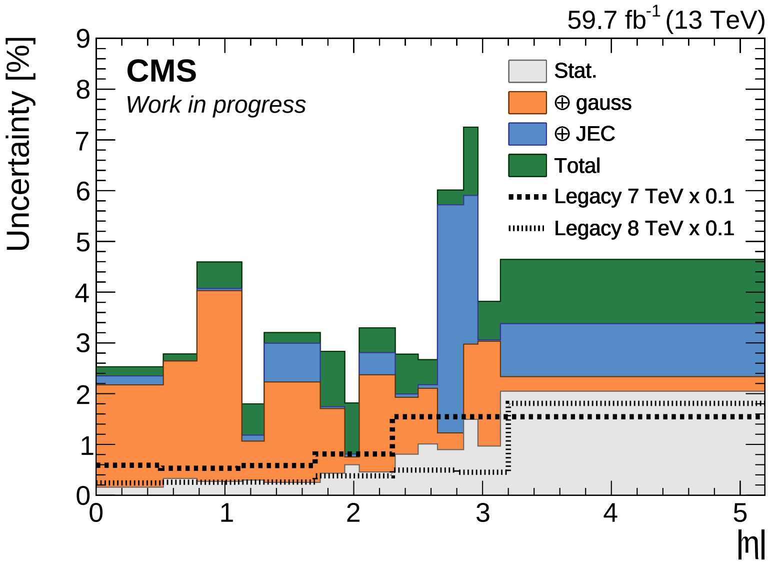


Figure 5.5.1: Relative uncertainties as a function of $|\eta|$ for the Run 2 Legacy dataset. The total uncertainty relative to Run 1 Legacy dataset are shown as comparison.

resolution. This hypothesis is tested with a fit of the data points with the following function:

$$f(p_{\mathrm{T}}) = \frac{k_{NS} \cdot N}{p_{\mathrm{T}}} \oplus \frac{k_{NS} \cdot S}{p_{\mathrm{T}}^{d}} \oplus k_C \cdot C\,, \tag{5.17}$$

where the N, S and C parameters are fixed to the results of the fit to simulation, while k_{NS} and k_C are free parameters. One common scale factor for the N and S parameters is used due to the limited statistical precision at low-p_{T}.

An example of the fit to data and simulation is shown in figure 5.4.2, where the best-fit values for the k parameters are found to be different, indicating a p_{T}-dependance. Typically, a weak p_{T}-dependance is present and no significant deviations from the linear behaviour have been observed so far. Therefore, p_{T} independent SFs are usually derived and a systematic uncertainty is obtained as the maximum deviation between the ratio between the two fitted functions, as illustrated in figure 5.4.4. Further discussion on the p_{T}-dependance of the resolution ratio can be found in section 5.8.

5.5.1 Total uncertainty

The individual sources of uncertainty are combined to provide the total uncertainty. A conservative assumption consists of symmetrising the uncertainty by considering the largest absolute deviation of each source of systematic uncertainty. This approach was used for the previous iteration of this measurement during Run 1 and was justified by one particular source of uncertainty dominating over the others. During Run 2, several systematic uncertainties become comparable in size. Based on the statistical consideration discussed in Ref. [197], a more precise way of combining the systematic errors is employed. A comparison of the total uncertainty between Run 1 and Run 2 results is shown in figure 5.5.1, where the largest components are highlighted; the derivation of the total uncertainty is discussed in the following.

First, each source that has an up ($\sigma^{\uparrow}$) and down ($\sigma^{\downarrow}$) variation, namely the ones related to the PU, PLI and JES, are symmetrised according to:

$$\sigma_{\text{sys}}^{\uparrow\downarrow} = \sqrt{\left(\frac{\sigma^{\uparrow} + \sigma^{\downarrow}}{2}\right)^2 + 2\left(\frac{\sigma^{\uparrow} - \sigma^{\downarrow}}{2}\right)^2} \,. \tag{5.18}$$

A direct consequence of combining asymmetric errors is to have a shift of the central value [197]. This bias is correct for each source of uncertainty as:

$$c_{\text{res}}^{\text{corrected}} = c_{\text{res}}^{\text{nominal}} + \frac{\sigma^{\uparrow} - \sigma^{\downarrow}}{2} \,. \tag{5.19}$$

All other uncertainty sources are assumed to be symmetric. Finally, the total uncertainty is derived as the quadratic sum of the statistical and the individual systematic uncertainties:

$$\sigma_{\text{tot}} = \sqrt{\sigma_{\text{stat}}^2 + \sum_i \sigma_{\text{sys,i}}^2} \,, \tag{5.20}$$

where σ_{stat} is the statistical component of the constant fit, as described in section 5.4.1, and the $\sigma_{\text{sys,i}}$ are the symmetrised systematic uncertainties.

Figure 5.5.1 shows the relative uncertainties as a function of $|\eta|$ for the Run 2 Legacy dataset. The uncertainty related to the non-Gaussian tails is one of the most dominant across the entire η range; the uncertainty related to the JES becomes relevant outside the barrel region ($|\eta| > 1.3$), in particular in the transition region ($2.8 < |\eta| < 3.0$). Finally, the JER total uncertainties relative to the 2018 data-taking are shown. An overall reduction is observed with respect to the Run 1 uncertainty. Despite the radiation damage of the CMS detector, a similar level of precision is observed in the central region, while a 50% reduction of the uncertainty is achieved in the endcap region. Further improvement (approximately a factor 3) is expected from the combination of the Run 2 Legacy dataset.

5.6 Results

The jet transverse momentum resolution scale factors for the Run 2 dataset corresponding the $137\,\text{fb}^{-1}$ at the $\sqrt{s} = 13\,\text{TeV}$ are shown in figure 5.6.1. The SFs, used in the resolution smearing procedure, are derived as a function of $|\eta|$ using the p_{T}-balance method in dijet events. These results have been used by all CMS analyses published using the full Run 2 dataset.

This measurement shows SFs systematically above unity, which implies that the jet energy resolution in data is broader than in simulation. The SFs correspond to a correction of 10-20%, consistent among the years and relatively stable across the η range. In the endcap region, a higher correction caused by the radiation damage is needed and increases over time.

The preliminary results for the Legacy calibration are discussed in section 5.9, where a significant improvement is observed.

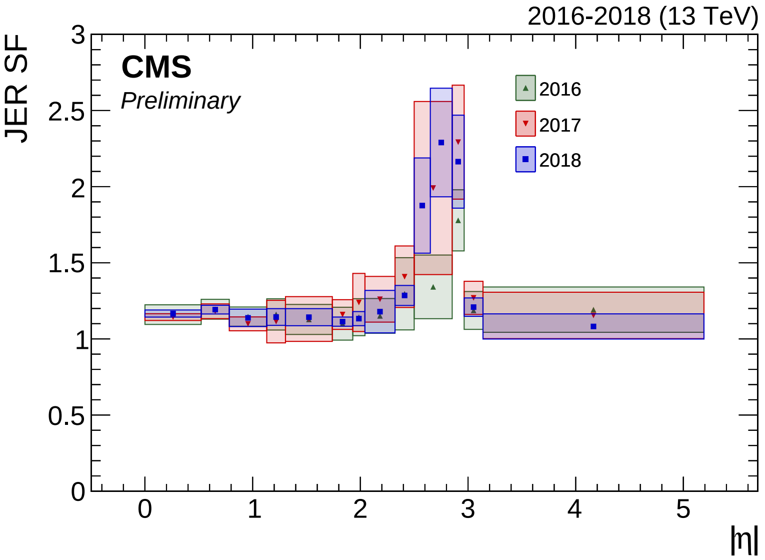


Figure 5.6.1: Year-dependent JER SF as a function of the pseudorapidity derived during Run 2 with the p_T-balance method using dijet events. Published in Ref. [1].

5.7 Correlation with the L2Residual corrections

The calibration of the jet transverse momentum resolution in simulation and the L2Residual corrections for the jet energy scale (cf. section 3.4.3) in data are strictly connected. The accurate calibration of the JES in data is needed in order to derive precise JER SFs. At the same time, reconstructed jets can migrate from adjacent p_T bins due to their finite JER, resulting in broader and shifted response distributions. Therefore, the incorrect smearing of the jet transverse momentum resolution in simulation can result in an imperfect L2L3Residual correction in data and vice-versa.

An example of this correlation is shown in figure 5.7.1. In the barrel region, where the JES is calibrated more accurately, the application of the resolution smearing corrections or the variation of the JES within the JEC uncertainty affects neither the mean nor the width of the asymmetry distribution significantly. On the other hand, the shift of the mean and the change of the width is more evident in the endcap region, where the JES and JER corrections are generally larger compared to the barrel regions and have more considerable uncertainties. An iterative approach between the two analyses is taken to reach stable and accurate corrections and reduce uncertainties.

Not only the central JES and JER corrections are correlated, but also their uncertainties. As described in section 3.4.5, the “Relative η-dependent” component of the JEC uncertainty is evaluated from the difference of the L2Residual corrections with the JER SFs varied within their uncertainty. As a consequence, the uncertainties in the JES and JER are correlated. Since the largest effect is present only in the endcap region, these two uncertainties are usually considered completely uncorrelated at the analysis level. Nonetheless, they can be further constrained in analyses that are particularly susceptible to the jet energy scale and resolution calibration. This procedure has been made possible by the statistically more accurate treatment of the JER sources of uncertainty.

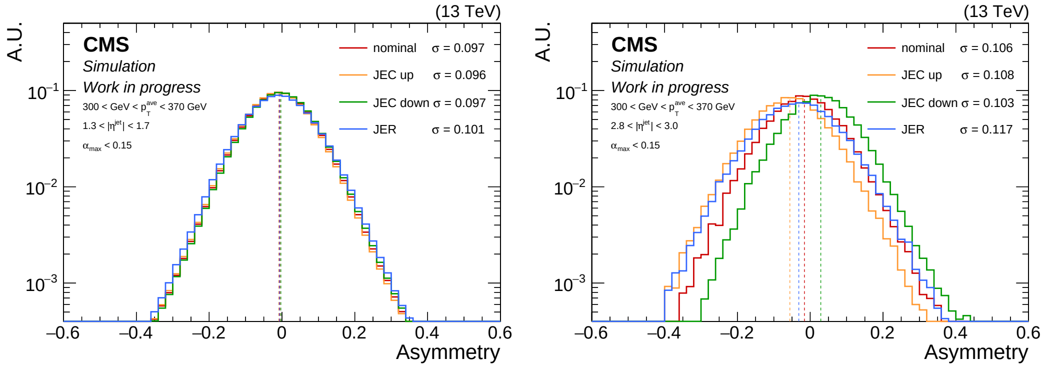


Figure 5.7.1: Normalised asymmetry distribution for simulated events in the barrel (left) and endcap (right) regions. The nominal (red) distribution refers to jets whose JES has been calibrated. The orange and green curve refer to the up and down variation of the JEC, respectively. The blue curve refers to the jets after the JER smearing procedure.

5.8 Complementary methods

The p_T-balance method using dijet events allows for the derivation of the data-to-simulation ratios over a wide range in both pseudorapidity ($|\eta| < 5.2$) and transverse momentum ($p_T > 100\,\text{GeV}$) of the jet to be calibrated.

The precise measure of the energy scale and resolution of low-p_T jets constitutes an important aspect for many physics analyses, especially those involving an accurate calibration of the p_T-balance in the event for the unbiased estimation of the p_T^{miss}.

The possibility of extending the dijet analysis towards low-p_T values is linked to the extrapolation procedure described in section 5.4. The requirement of an additional third jet for which $\alpha < 0.3$ (or $p_{T,3} < 30\,\text{GeV}$) drastically reduces the number of dijet events with $p_T^{ave} < 100\,\text{GeV}$, and also increases the challenge to separate low-p_T jets from PU jets.

Alternative methods to the dijet analysis can be exploited to extend the coverage towards low-p_T values. Some examples of such complementary approaches, involving Z/γ+jet events and the Random Cone method, are discussed in the following.

5.8.1 Z/γ+jet events

Similarly to the L2L3Residual derivation described in section 3.4.3, the γ+jet and Z+jet events can provide complementary measurements of the jet transverse momentum resolution. In such events, the jet is balanced in the transverse plane by a photon or a Z boson, and no additional third jet is required. Besides, the usage of the photon and leptons from the Z boson decay allows for a more precisely calibrated reference object utilising the superior performance of the ECAL and muon system.

The γ+jet analysis provides an independent and complementary cross-check of the results obtained with the dijet asymmetry method and was used during Run 1 and at the beginning of Run 2. The results obtained from the γ+jet analysis were found to be in agreement with the dijet results [148]. On the other hand, it was observed that only a minimal extension towards lower p_T could be achieved with this approach, which is limited to the central region of the detector ($|\eta| < 2.5$); outside this region, the number of events available is low and the uncertainty exceeded that of the dijet

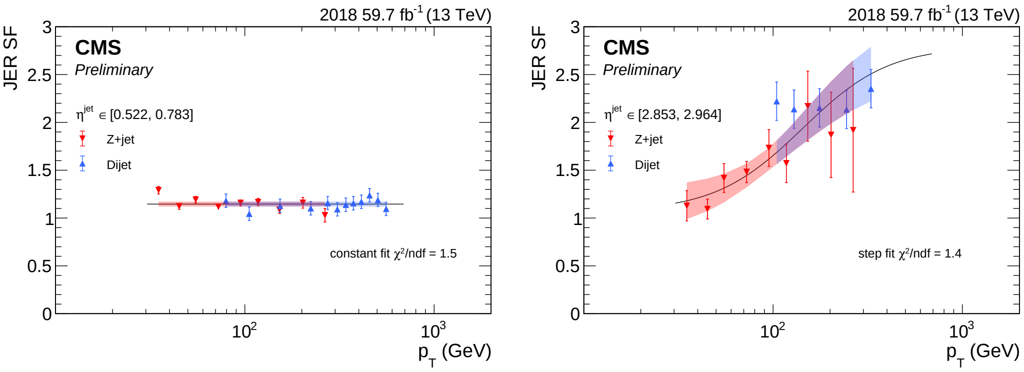


Figure 5.8.1: Combination of the JER SF obtained from the Z+jet and dijet analyses in the barrel (left) and in the endcap (right). Published in Ref. [1].

channel. Nonetheless, the γ+jet analysis can be used to reduce the uncertainties by performing a multi-channel combination.

In contrast, the Z+jet analysis allows for a SF measurement down to 40 GeV, despite the limited statistical precision for $p_{\mathrm{T}} > 150\,\mathrm{GeV}$. This channel has been used during Run 2 in combination with the dijet approach. Due to the radiation damage in the region of $|\eta| > 2.5$ (see section 5.3), an increase over time of the dijet-based SF is observed in the transition region between the endcap and the HF calorimeter ($|\eta| \in [2.5, 3]$). In the same region, a strong p_{T}-dependence of the SF is measured by combining the results of the dijet and Z+jet events. This combination is shown in figure 5.8.1, where it is visible that the SF increases towards high-p_{T} in the transition region. No p_{T}-dependence is observed elsewhere.

Besides the extension of the p_{T} coverage and the cross-check of the dijet results, the Z/γ+jet channels offer the possibility to improve the JER SFs derivation. Similarly to the global fit procedure for the L2L3Residual derivation, the combination of these channels should be explored to achieve higher precision across the entire phase space.

5.8.2 Noise measurement using the Random Cone method

The dominant contribution at low-p_{T} to the jet transverse momentum resolution comes from the noise term N of eq. (5.2), which is the result of the electronic noise and the pileup contribution:

$$N = N^{\mathrm{PU}} \oplus N^{\mu=0}\,. \tag{5.21}$$

The N term is only partially constrained in the analysis using dijet events. The electronic noise term $N^{\mu=0}$ can be estimated from dedicated MC simulated samples with $\mu = 0$. An indirect derivation of the noise term due to PU, N^{PU}, can be obtained from the difference of resolution derived with and without simulated pileup. However, the *random cone* method offers a way to estimate this contribution explicitly. This method, already used to subtract the PU contribution for the JES (see section 3.4.1), relies on the measurement of the fluctuations in the energy deposits due to PU. In fact, the difference in p_{T} between two randomly placed cones has been observed by the ATLAS Collaboration [188] to be a good estimation of the fluctuations of the energy deposits due to pileup.

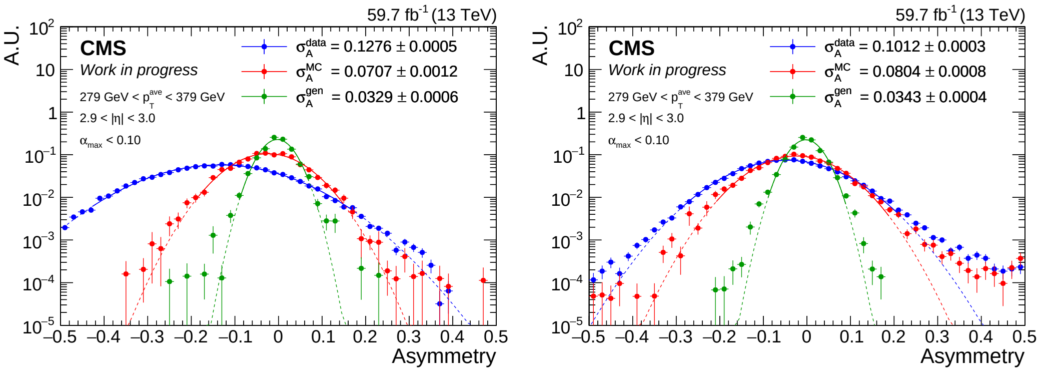


Figure 5.9.1: Normalised asymmetry distribution for data (blue), simulated events (red) and particle level (green) jets. The Gaussian fit to the core of the distribution is shown as a solid line excluding 1.5% of the total area on both sides. The width, calculated as the RMS, and its error are reported for each distribution. The pre-(left) and post-(right) Legacy calibration in 2018 is shown.

This innovative method has the advantage of being directly applicable to data by selecting events collected with unbiased triggers. Furthermore, it allows extending the measurement of the jet transverse momentum resolution to very low p_{T} values, a region that was not accessible before.

5.9 Legacy dataset

The data collected by the CMS detector are processed several times to improve the calibration (cf. section 2.2.7). The data reconstructed with the best possible calibration is referred to as “Legacy” dataset and allows to achieve excellent performance in terms of energy scale and resolution of several reconstructed physical objects, like jets. In addition to the recalibration, MC simulation of the data-taking conditions for each year of Run 2 is used to obtain an improved description of the detector response. The JER SFs are an excellent showcase of this improvement.

Before the Legacy calibration, the transition region in $2.8 < |\eta| < 3.0$ represents a critical area, in which the corrections for both the jet energy scale and resolution were more significant and associated with larger uncertainties.

An example of the improvements due to the Legacy reconstruction is shown in figure 5.9.1. The asymmetry distribution in data before the Legacy calibration is wider and its mean shifted compared to the Legacy result. Moreover, also the asymmetry width in simulation is closer to the one in data for the Legacy dataset, which results in smaller SFs, thanks to the better-calibrated data and more accurate simulations.

The preliminary results of the JER SF measurement for the Run 2 dataset using the Legacy calibration are shown in figure 5.9.2. The improvement consists of a large reduction of the systematic uncertainty (approximately a factor 3), shown in more detail in figure 5.5.1, and smaller corrections across the entire η range. The SFs are consistent in the barrel region ($|\eta| < 1.3$) for all years, while a strong time dependence is visible elsewhere. In particular, the large SF for 2017 are related to the “prefiring” issue. Furthermore, the p_{T}-dependence is strongly reduced, and an even better precision of the order of 1% or less is expected when performing a combination with the Z/γ+jet results.

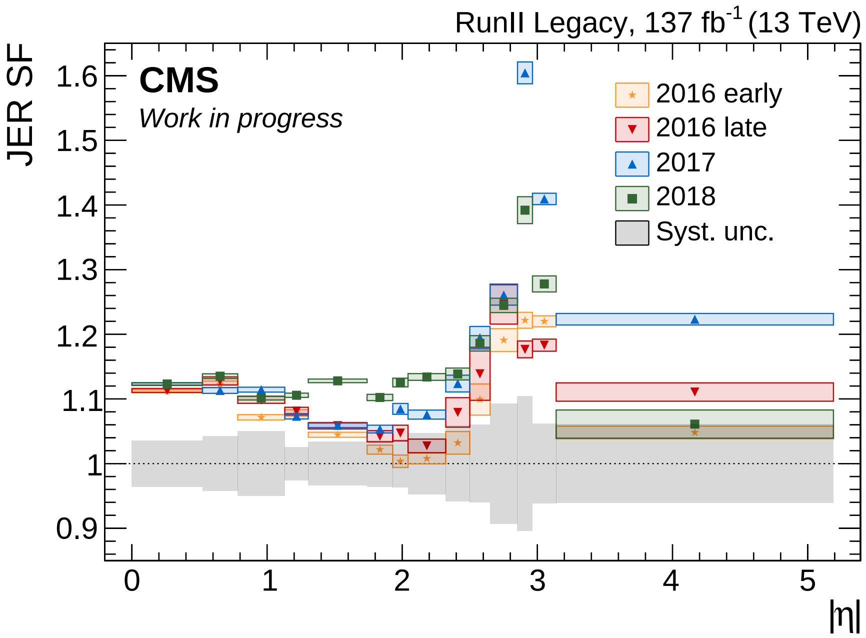


Figure 5.9.2: Year-dependent JER SF as a function of the pseudorapidity derived with the p_{T}-balance method using dijet events. The results refers to the Legacy calibration. The statistical uncertainty is shown for each point, while the total systematic uncertainty (gray band) is shown around 1.

5.10 Summary

The comprehensive knowledge of the jet transverse momentum resolution is important for Standard Model precision measurement as well as searches for search for new physics. The accurate calibration of jets and the implications to the reconstructed missing transverse momentum play an essential role in this kind of searches.

The p_{T}-balance method based on dijet events has been successfully used since Run 1 [148] to derive corrections for the jet transverse momentum resolution. The increased understanding of the CMS detector during Run 2 data-taking and the evolution of analysis strategies made it possible to reach an unprecedented level of precision in the calibration of the jet at hadron colliders, both in energy scale and resolution.

The studies presented in this thesis for the Run 2 Legacy datasets show a precision of the order of 1%, obtained using dijet events. This channel alone makes it possible to derive data-to-simulation scale factors over a wide range in both pseudorapidity ($|\eta| < 5.2$) and transverse momentum ($p_{\mathrm{T}} > 100\,\mathrm{GeV}$). In the future, a multi-channel combination and the extension towards low-p_{T} are the key parameters to reach higher precision across the entire phase space.

6 Search for diboson resonances

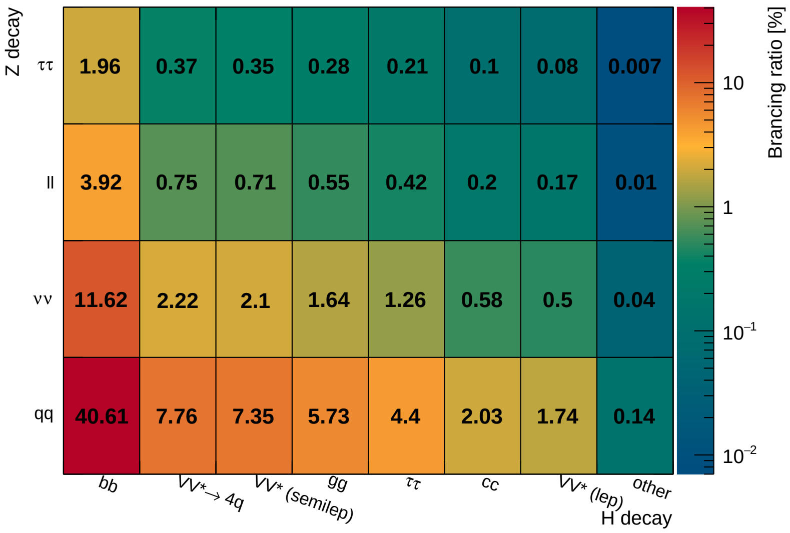
This chapter describes a model-independent search for a heavy resonance decaying into a Z *and a Higgs boson. The analysis is performed using the* pp *collision data recorded by the CMS detector at the LHC in* 2016-2018 *with a centre-of-mass energy of 13* TeV*. The search focuses on resonances with masses in the range between 1* TeV *and 5* TeV *and natural widths small compared to the detector resolution. The final states investigated include the light-flavoured hadronic decays of the Higgs boson, with particular emphasis on the 4-prong* ($\mathrm{H} \rightarrow \mathrm{VV}^* \rightarrow \mathrm{qqqq}$) *and the c flavour* ($\mathrm{H} \rightarrow \mathrm{c\overline{c}}$) *decays, which have not been directly explored in the context of BSM searches before, and a pair of oppositely charged leptons, arising from the* Z *boson decay. The analysis strategy and the selection criteria are discussed in the following, together with the usage of state-of-the-art techniques for Higgs boson tagging. The combination with the analysis based on the* Z *boson decaying into a pair of neutrinos [5] is discussed afterwards. The chapter closes with a comparison with existing analyses in different final states and an outlook for future improvements.*

6.1 Introduction

Predicted by a multitude of BSM theories, diboson resonances remain one of the most promising discovery channels for new particles. In this work, emphasis is put on the resonant production of a hypothetical spin-1 massive particle decaying into a Z and a Higgs boson. The results of this analysis will be interpreted in theoretical models using the HVT framework [77] introduced in section 1.4.

A number of experimental techniques can be used to explore the various signatures of the different final states of the Z and Higgs bosons. An overview of the most recent CMS results related to diboson resonance searches is reported in section 1.4. The analysis presented in this chapter is based on the light-flavoured hadronic decays of the Higgs boson, with particular emphasis on the 4-prong ($\mathrm{H} \rightarrow \mathrm{VV}^* \rightarrow \mathrm{qqqq}$) and the c flavour ($\mathrm{H} \rightarrow \mathrm{c\overline{c}}$) decays. These decay channels have not been explicitly considered before in the context of BSM searches, although together they constitute the second-largest BR (14%) after the $\mathrm{H} \rightarrow \mathrm{b\overline{b}}$ decay mode (58%).

Despite the significantly smaller BR, the newly explored final states can reach a sensitivity for high resonance masses that exceeds that of the $\mathrm{H} \rightarrow \mathrm{b\overline{b}}$ channel [86], as will be shown at the end of this chapter. Moreover, the diboson resonance search profits from the analysis and combination of all available channels.



Z decay \ H decay	bb	VV*→ 4q	VV* (semilep)	gg	ττ	cc	VV* (lep)	other
ττ	1.96	0.37	0.35	0.28	0.21	0.1	0.08	0.007
ll	3.92	0.75	0.71	0.55	0.42	0.2	0.17	0.01
νν	11.62	2.22	2.1	1.64	1.26	0.58	0.5	0.04
qq	40.61	7.76	7.35	5.73	4.4	2.03	1.74	0.14

Figure 6.1.1: Branching fractions of the combined Z and Higgs boson decay modes.

In order to suppress the overwhelming QCD background typical for pp collisions, the Z boson decays into a pair of oppositely charged leptons ($\ell\ell$), here and in the following referring to electrons and muons, are investigated. Thus, regardless of their low BR (approximately 3.3% each), the clean signature of these charged leptons allows for an efficient selection of potential signal events. Furthermore, the Z boson decay into a pair of neutrinos constitutes another interesting channel, in which the 6-times larger BR and the higher selection efficiencies compensate for the larger amount of background expected. An overview of the combined ZH decay modes and their respective BR is shown in figure 6.1.1.

In the following section, the overview of the analysis strategy for the search involving the charged lepton and neutrino decay modes is given. Then, the data and simulated samples, as well as the event selection, are described in detail; the selection criteria have been studied to maximise the sensitivity of this analysis while ensuring orthogonality with other analyses targeting different final states, particularly the $\mathrm{H} \to \mathrm{b\bar{b}}$ decay, for a future statistical combination of both results.

6.2 Analysis strategy

This analysis searches for potential signals of a heavy resonance decaying into a Z and a Higgs boson. A pair of oppositely charged leptons or the missing transverse momentum are used to identify the Z boson, while a large-radius jet is used as a candidate to reconstruct the Higgs boson. An illustration of the signal process considered in this analysis is shown in figure 6.2.1.

New, heavy resonances with masses of a few TeV and natural widths small compared to the detector resolution are considered in this analysis. The TeV scale range entails the decay products of the new resonance to be highly energetic and collimated. On the one hand, this boosted topology provides a distinctive way to distinguish between SM and BSM signatures. On the other hand, it makes it harder to reconstruct the individual objects and, therefore, dedicated methods and techniques are required. In particular, the collimated charged leptons originating from

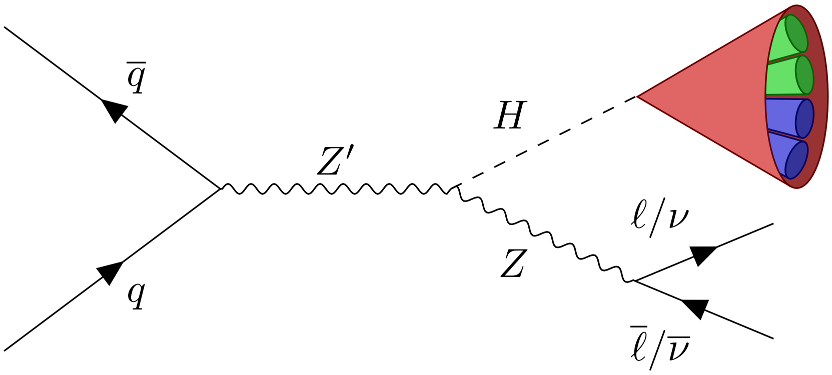


Figure 6.2.1: Schematic representation of the production and decay chain of a Z′ boson.

the Z boson decay suffer from reconstruction and identification inefficiencies due to the small angular separation of these particles and their straight trajectory. This effect becomes increasingly relevant for higher resonance masses. In order to mitigate these inefficiencies, dedicated strategies are employed, as described in section 6.4.

Similar to the leptons from the Z boson decay, also the hadronic decay products of the Higgs boson are closely collimated and therefore likely to be clustered into a single large-radius jet. The substructure and the flavour of the resulting jet provide powerful tools to discriminate Higgs-boson-initiated jets from the background ones. Jet tagging plays an essential role in this analysis and state-of-the-art techniques are employed to enhance the sensitivity to the newly explored Higgs boson final states.

The search is performed in a signal-enriched region (SR) by examining the distribution of the invariant mass of the reconstructed Z and Higgs boson candidates for a localised excess, due to the potential signal, over a monotonically decreasing background distribution. The smooth description of the SM background prediction entirely relies on data. This approach comes with the advantage of reducing systematic uncertainties and smoothing statistical fluctuations. Finally, the background modelling strategy is validated in a background-enriched validation region (VR), with kinematic properties similar to the SR.

6.3 Datasets and simulated events

The analysis is performed using the pp collision data recorded in 2016-2018 with the CMS detector at the LHC at $\sqrt{s} = 13\,\text{TeV}$, corresponding to an integrated luminosity of about $137\,\text{fb}^{-1}$. The integrated luminosities of each data-taking period and the associated uncertainties are reported in table 2.2.2 per each year.

The data analysed in the presented search have been recorded by triggers requiring the presence of a single lepton or missing transverse momentum for the charged lepton and neutrino channels, respectively. A combination of the isolated with a low-p_T threshold (above 27 GeV) and non-isolated with a high-p_T threshold (above 115 GeV) electron triggers is used to achieve optimal efficiency in the whole p_T range. Moreover, photon triggers are used to recover efficiency at high-p_T (above 300 GeV). Single muon triggers ($p_\mathrm{T} > 50\,\text{GeV}$) without isolation are chosen to avoid losses in case of boosted $\mathrm{Z} \to \mu\mu$ events. The triggers based on missing transverse momentum (p_T^miss above 100 GeV) achieve an efficiency above 99% in the phase space considered.

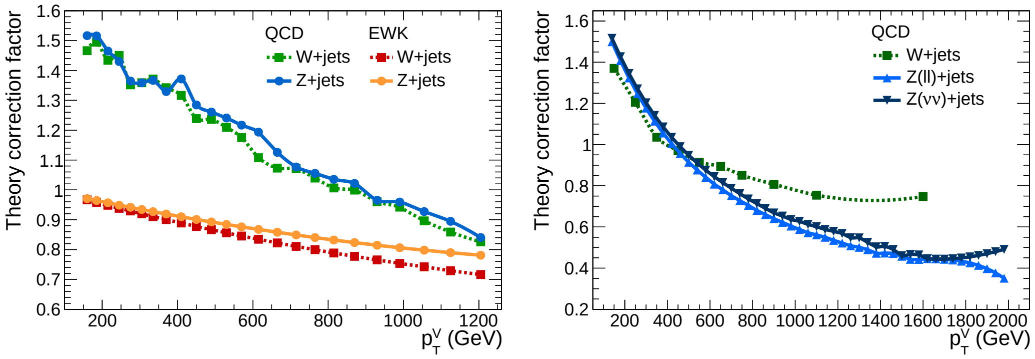


Figure 6.3.1: Left: NLO QCD and EW correction factors for V+jet processes as a function of the vector boson p_{T}. Right: NLO QCD correction factors, including the corrections accounting for differences in the generated samples, for V+jet processes as a function of the vector boson p_{T}. Derived from Ref. [198, 199].

The signal samples based on the HVT framework have been generated at LO using MadGraph5_amc@nlo [128], while the subsequent parton showering and hadronisation processes are simulated with pythia8 [123]. The Z′ resonance exclusively decays to a Z and a Higgs boson in these samples and is simulated with a decay width of 0.1% of the resonance mass to satisfy the narrow-width approximation. The Z′ → ZH decay is simulated with only the Z boson decays to leptons[1], while all decays modes of the Higgs boson are simulated. Two different sets of signal samples have been generated to account for Z boson decays to charged leptons and neutrinos, respectively.

The final states of signal events consist of a large-radius jet coming from the hadronic decay of the Higgs boson, and a pair of leptons or missing transverse momentum for the charged lepton and neutrino channels, respectively. The primary source of background that can mimic the signal signature is arising from Z+jet production. For such events, the Z boson is produced in association with an ISR jet, which is misidentified as a Higgs-boson-initiated jet. For the neutrino channel, the W+jet events constitute the second dominant background process in the case of a lost lepton from the W → $\ell\nu$ decay. Simulated samples of V+jet events have been generated with MadGraph5_amc@nlo at LO, while the parton showering and hadronisation processes are simulated with pythia8. Subdominant background processes include diboson production (WW, WZ and ZZ) and $\mathrm{t\bar{t}}$ events, which have been generated with pythia8 at LO and powheg at NLO, respectively. A summary of the simulated samples used in this analysis is given in tables 6.3.1 and 6.3.2. The production cross section of diboson and $\mathrm{t\bar{t}}$ samples include the NLO and NNLO accuracy, respectively. The values for all other samples are given with the LO accuracy. Moreover, the V+jet samples are further corrected to account for NLO QCD and EW contributions and improve the modelling of highly energetic events. Figure 6.3.1 (left) shows the correction as a function of the V boson p_{T}, derived from the studies in Ref. [198]. Furthermore, the production cross section of signal samples is assumed to be 1 pb for illustration.

The UE is simulated for all samples using the CUETP8M1 tune [193] for 2016 and CP5 tune [194] for 2017 and 2018. For the $\mathrm{t\bar{t}}$ sample in 2016, the CUETP8M2T4

[1] The generated samples also include the Z → $\tau\tau$ decays, which are not considered in this analysis.

tune [200] is used instead. Furthermore, all simulated events are reweighted such that the pileup distribution matches the one in data. A minimum bias cross section of $69.2 \pm 4.6\%$ mb is assumed.

Dataset	$\sigma \times$ BR	Weighted number of generated events		
	[pb]	2016	2017	2018
$\mathrm{Z}(\rightarrow \ell\ell)$+jet, $H_\mathrm{T} \in [100, 200]$ GeV	$1.47 \cdot 10^{2}$	10977326	11180126	11530510
$\mathrm{Z}(\rightarrow \ell\ell)$+jet, $H_\mathrm{T} \in [200, 400]$ GeV	$4.10 \cdot 10^{1}$	9589193	10675441	11225887
$\mathrm{Z}(\rightarrow \ell\ell)$+jet, $H_\mathrm{T} \in [400, 600]$ GeV	$5.68 \cdot 10^{0}$	9725661	10174800	9643184
$\mathrm{Z}(\rightarrow \ell\ell)$+jet, $H_\mathrm{T} \in [600, 800]$ GeV	$1.37 \cdot 10^{0}$	8253178	8691608	8862104
$\mathrm{Z}(\rightarrow \ell\ell)$+jet, $H_\mathrm{T} \in [800, 1200]$ GeV	$6.30 \cdot 10^{-1}$	2673066	3089712	3138129
$\mathrm{Z}(\rightarrow \ell\ell)$+jet, $H_\mathrm{T} \in [1200, 2500]$ GeV	$1.51 \cdot 10^{-1}$	596079	616923	536416
$\mathrm{Z}(\rightarrow \ell\ell)$+jet, $H_\mathrm{T} \in [2500, \infty)$ GeV	$3.57 \cdot 10^{-3}$	399492	401334	427051
$\mathrm{Z}(\rightarrow \nu\nu)$+jet, $H_\mathrm{T} \in [100, 200]$ GeV	$3.03 \cdot 10^{2}$	19026540	22737266	23702894
$\mathrm{Z}(\rightarrow \nu\nu)$+jet, $H_\mathrm{T} \in [200, 400]$ GeV	$9.26 \cdot 10^{1}$	5136083	21675916	23276346
$\mathrm{Z}(\rightarrow \nu\nu)$+jet, $H_\mathrm{T} \in [400, 600]$ GeV	$1.33 \cdot 10^{1}$	8771480	9134120	10928927
$\mathrm{Z}(\rightarrow \nu\nu)$+jet, $H_\mathrm{T} \in [600, 800]$ GeV	$3.26 \cdot 10^{0}$	5766322	5697594	5748975
$\mathrm{Z}(\rightarrow \nu\nu)$+jet, $H_\mathrm{T} \in [800, 1200]$ GeV	$1.49 \cdot 10^{0}$	2170137	2058077	2066798
$\mathrm{Z}(\rightarrow \nu\nu)$+jet, $H_\mathrm{T} \in [1200, 2500]$ GeV	$3.43 \cdot 10^{-1}$	143957	334332	343198
$\mathrm{Z}(\rightarrow \nu\nu)$+jet, $H_\mathrm{T} \in [2500, \infty)$ GeV	$6.95 \cdot 10^{-3}$	405030	6446	350181
$\mathrm{W}(\rightarrow \ell\nu)$+jet, $H_\mathrm{T} \in [100, 200]$ GeV	$1.40 \cdot 10^{3}$	38593839	32948954	29611903
$\mathrm{W}(\rightarrow \ell\nu)$+jet, $H_\mathrm{T} \in [200, 400]$ GeV	$4.08 \cdot 10^{2}$	19069732	18463508	25468933
$\mathrm{W}(\rightarrow \ell\nu)$+jet, $H_\mathrm{T} \in [400, 600]$ GeV	$5.75 \cdot 10^{1}$	7759701	14313274	5932701
$\mathrm{W}(\rightarrow \ell\nu)$+jet, $H_\mathrm{T} \in [600, 800]$ GeV	$1.29 \cdot 10^{1}$	18687480	21709087	19771294
$\mathrm{W}(\rightarrow \ell\nu)$+jet, $H_\mathrm{T} \in [800, 1200]$ GeV	$5.37 \cdot 10^{0}$	7830536	11261008	8192251
$\mathrm{W}(\rightarrow \ell\nu)$+jet, $H_\mathrm{T} \in [1200, 2500]$ GeV	$1.33 \cdot 10^{0}$	6872441	39070488	7542264
$\mathrm{W}(\rightarrow \ell\nu)$+jet, $H_\mathrm{T} \in [2500, \infty)$ GeV	$3.22 \cdot 10^{-2}$	2637821	20467960	3119311
$\mathrm{t}\bar{\mathrm{t}}$ +jet	$8.32 \cdot 10^{2}$	76738314	13543218350	77080828940
WW	$1.15 \cdot 10^{2}$	7982180	7765891	7846136
WZ	$4.71 \cdot 10^{1}$	3997571	3901180	3884167
ZZ	$1.65 \cdot 10^{1}$	1988098	1928489	1978777

Table 6.3.1: Simulated background samples. The σ ×BR is shown in pb. The last three columns give the weighted number of generated events for each of data-taking periods.

Z′ mass [GeV]	$\mathrm{Z} \rightarrow \ell\ell$ channel			$\mathrm{Z} \rightarrow \nu\nu$ channel		
	2016	2017	2018	2016	2017	2018
1400	99600	100000	99950	107152	107128	107137
1600	100000	100000	99945	108312	108245	101772
1800	97200	100000	99949	109150	109365	109219
2000	99800	100000	99946	43200	110244	110167
2500	95800	100000	99934	111937	112019	111996
3000	99700	100000	99919	111664	59801	112482
3500	100000	100000	99885	106858	109494	73826
4000	85000	100000	99830	107927	107968	107542
4500	94000	97000	99787	99629	99241	99766
5000	100000	100000	99643	88940	44009	86411

Table 6.3.2: Weighted number of generated events of simulated signal samples for each data-taking periods.

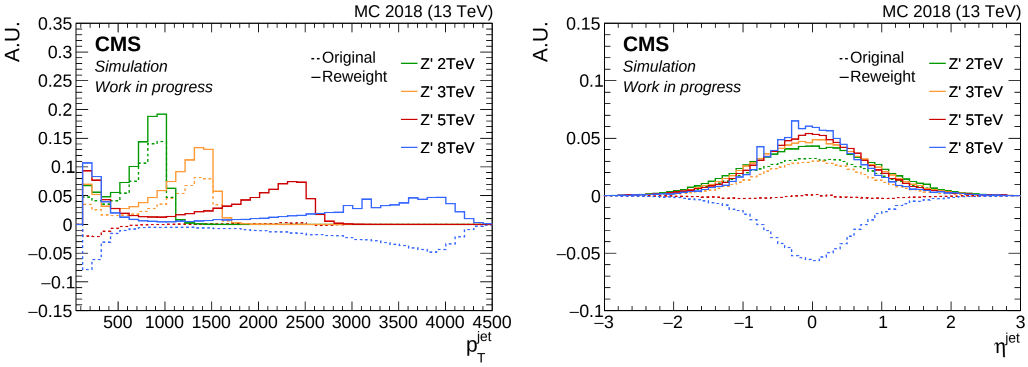


Figure 6.3.2: Jet p_T (left) and η (right) distributions before (dashed) and after (solid) the PDF reweighting procedure for different signal samples of the neutrino channel in 2018. The distributions are normalised to the same absolute area. This plot is made by the author of this work; an adapted version was already shown in Ref. [5].

Additional reweight

The simulated V+jet sample production campaign in 2017 and 2018 used slightly different MC generator parameters ("pdfwgt" [128]) than that in 2016, resulting in significantly different boson p_T spectra and worse agreement between data and simulation; therefore, an additional correction is applied to these samples. The correction factor, multiplied by the NLO QCD correction to the production cross section described above, is shown in figure 6.3.1 (right).

The parton distribution functions (PDFs) used to produce all signal samples are taken from the NNPDF3.1 set [201]. Several different PDF versions are provided, corresponding, for example, to different values of parameters and levels of precision in perturbation theory. The signal samples for the $Z \to \ell\ell$ and $Z \to \nu\nu$ decays were produced with different PDF version. In order to have a simplified treatment of the correlation of the PDF uncertainties when performing the combination of these channels, a common set is preferred instead. Moreover, the PDF version at next-to-next-to-leading order (NNLO) used for the $Z \to \nu\nu$ channel presented large negative weights resulting in unphysical distributions, mostly evident at high resonance masses. Therefore, these samples are reweighted to the PDF version at LO used to generate the $Z \to \ell\ell$ sample. Figure 6.3.2 shows the distribution of jet related variables before and after the reweighting procedure for illustration.

6.4 Event selection

Data and simulated events are reconstructed with the PF algorithm. All events are required to be recorded by one of the triggers described above and to have at least one well-reconstructed PV. The kinematic and identification requirements applied to each PF candidate are summarised in the following.

Charged leptons

Leptons are considered only if they have $p_T > 52\,\text{GeV}$ and $|\eta| < 2.4$. Furthermore, each lepton is required to fulfil a set of ID criteria. As already described in section 3.2, dedicated algorithms for boosted leptons are available to optimise the

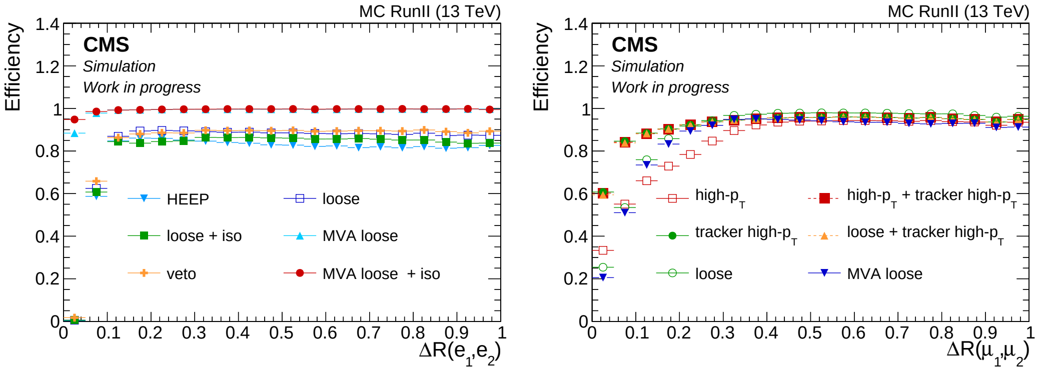


Figure 6.4.1: Electron (left) and muon (right) identification efficiency for different IDs as a function of the ΔR between the two leptons matched to a respective generated lepton at the truth-level. The events from all signal samples are considered.

selection efficiency in case of high-p_T and small angular separation. The selection efficiency for different IDs is shown in figure 6.4.1 for electrons (left) and muons (right). The loose WP for the MVA-based electron ID provides a selection efficiency close to unity, outperforming all other IDs in the whole phase space; since the electron isolation is already included in the ID definitions, no additional requirement is imposed.

The muon IDs show similar performance for well-separated ($\Delta R(\mu_1, \mu_2) > 0.45$) muons, while the usage of the "tracker high-p_T" ID provides a clear gain in efficiency for close-by muons. A muon pair is selected such that one of them fulfils the "high-p_T" ID and the other one at least the "tracker high-p_T" ID; this choice is also used in the search in the $H \to b\overline{b}$ channel [86]. Besides, similar performance are observed when substituting the "high-p_T" ID requirement with the loose or the "tracker high-p_T" IDs, but with no significant difference in the final results. Additionally, the muons are required to have a relative isolation $I_{rel} < 0.15$ [134]. Finally, muons are corrected for the bias in their momentum scale due to the mismeasurement of the curvature of high-p_T tracks. These corrections are derived using the generalized-endpoint method [109], and their effect on the muon energy scale is approximately 1%.

Data-to-simulation corrections for the reconstruction, identification, isolation and trigger efficiencies have been measured by the CMS Collaboration and are applied for each muon and electron in the event.

Jets and missing transverse momentum

The large-radius jets used in this analysis are clustered with the anti-k_T algorithm with a radius parameter of $R = 0.8$, and the SD algorithm is used to identify the subjets. Moreover, the PUPPI algorithm is used to suppress the PU contribution. The jet energy scale and resolution are corrected in both data and simulated events, as described in section 3.4. The corrected jets are further considered in this analysis only if they have $p_T > 200\,\text{GeV}$ and $|\eta| < 2.4$, and fulfil the tight jet ID. Moreover, small-radius jets ($R = 0.4$) are used for the calculation of the missing transverse momentum. Therefore, the JES and JER of these jets are calibrated as well; the JES corrections for small-radius jets with $p_T > 15\,\text{GeV}$ are then propagated to the Type-1 $\vec{p}_T^{\,\text{miss}}$ (cf. section 3.5).

Additional cleaning

As already discussed in section 5.3, several detector malfunctioning occurred during data-taking periods. Therefore, dedicated corrections are applied to mitigate their impact on the analysis whenever possible. Otherwise, affected events are discarded.

In particular, the data collected in 2016 and 2017 was affected by a trigger inefficiency in the region of $|\eta| > 2.0$ [120]; dedicated data-to-simulation corrections are applied for simulated events to emulate the efficiency loss.

Moreover, during the 2018 data-taking period, two endcap sectors of the HCAL were not functional. Therefore, a jet in this region can either be reconstructed as less energetic jet or misidentified as an electron. Moreover, a misreconstruction of the $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ caused by the unmeasured energy in the event is observed in both scenarios.

Depending on the final states under consideration, each analysis is affected differently. For example, in the charged lepton channel of this analysis, the requirement of two close-by leptons and the back-to-back topology with a highly boosted jet reduces the number of misidentified electrons in the affected region. Events that contain a jet or electron in the region of $\eta \in [-2.96, -1.31]$ and $\phi \in [-1.57, -0.87]$ are discarded. Since less than 0.5% of the events otherwise passing the entire event selection are discarded by this criterion, this requirement has negligible impact.

Examples of the η-ϕ distribution of jets for the electron and neutrino channels before and after the event veto are shown in figure 6.4.2. The muon channel shows similar results to the electron one.

The impact is most evident for the neutrino channel, where the mismeasured energy due to this detector malfunctioning creates spurious $p_{\mathrm{T}}^{\mathrm{miss}}$. Moreover, an enhance of jets in the opposite region in ϕ is observed. Those events are real QCD dijet events for which one jet in the affected area is not reconstructed as such. Such events are not effectively removed by the selection criteria, described in the previous and following sections, and are characterised by an enhancement of $p_{\mathrm{T}}^{\mathrm{miss}}$ in the affected region. These events are suppressed by the high threshold on $p_{\mathrm{T}}^{\mathrm{miss}}$.

Preselection

The boosted topology of signal events can be further exploited to suppress SM background processes. After the kinematic and ID requirements mentioned above for the physics objects used, the following selection criteria, referred to as *preselection*, have been applied to increase the sensitivity of this analysis.

In the charged lepton channels, each event must have 2 opposite-sign leptons of the respective flavour (e or μ) and no additional leptons of the opposite flavour. Moreover, the defined dilepton system is required to be boosted ($p_{\mathrm{T}}(\ell\ell) > 200\,\mathrm{GeV}$), consistent with the Z boson mass hypothesis ($81\,\mathrm{GeV} \leq m(\ell\ell) \leq 101\,\mathrm{GeV}$) and back-to-back to one boosted, large-radius jet ($\Delta\phi(\ell\ell, \mathrm{jet}) \geq 2$). These selection requirements efficiently reject events arising from background processes and retain a large fraction of signal events. In the rare case of multiple lepton pairs or jets that fulfil these criteria, those whose invariant mass is closer to the nominal values of the Z and Higgs bosons [7], respectively, are chosen.

Furthermore, as shown in figure 6.4.3, leptons in background events have large angular separation, while they are expected to be collimated in signal events due to the boosted Z boson decay; therefore, the requirement of two close-by leptons

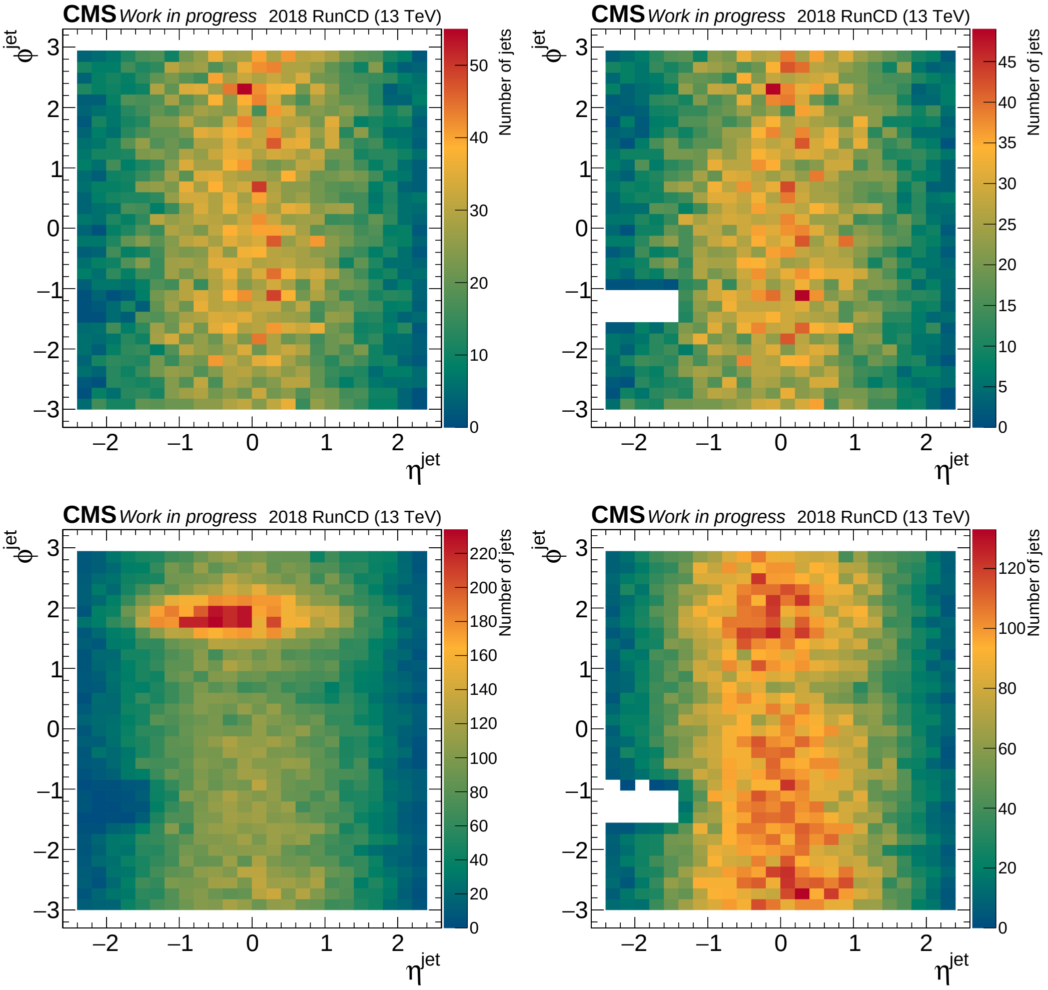


Figure 6.4.2: Number of jets in the η-ϕ plane shown in data collected in 2018 when during the detector malfunctioning before (left) and after (right) the cleaning procedure described in the text for the electron (upper) and neutrino (lower) channels. The lower plots are made by the author of this work; an adapted version was already shown in Ref. [5].

($\Delta R(\ell\ell) \leq 0.45$) improves the signal-to-background ratio for all generated signal masses further.

The large-radius jets that fulfil the requirements described above are further investigated for their flavour content. Jets are rejected if all their subjets fulfil the loose b tagging working point of the DeepCSV tagger (cf. section 3.3.2). After this selection, the signal yield is reduced with almost no effect on the background. Although it does not improve the sensitivity of this analysis, this requirement is applied in order to ensure orthogonality of this analysis and the corresponding search in the $\mathrm{H} \to \mathrm{b\bar{b}}$ channel [86]. Therefore, it has the advantage of allowing the combination of the two analyses. Furthermore, the DeepAK8 tagger (cf. chapter 4) is employed to identify the jets coming from a Higgs boson decay. The usage of this tagger is explained in greater detail in the following section.

In the neutrino channel, events containing muons or electrons that fulfil the selection criteria described in the previous section are excluded to ensure an event selection orthogonal to that of the charged lepton channels. In the absence of leptons, similar selection criteria are applied to the missing transverse momentum. Events in the neutrino channel are required to have $p_{\mathrm{T}}^{\mathrm{miss}} > 250\,\mathrm{GeV}$ and $\Delta\phi(p_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{jet}) \geq 2$ to exploit the boosted back-to-back topology of the expected signal. Moreover, an additional angular requirement of $\Delta\phi(p_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{jet}) \geq 0.5$ is applied to all small-radius jets to suppress QCD events.

Figure 6.4.3 shows the distributions of the transverse momentum of the dilepton pair and of the large-radius jet in events that fulfil the preselection criteria described above; generally, a good data-to-simulation agreement is observed for each variable and across channels. Moreover, the distributions for three signal samples with different masses are shown as a reference.

The number of events in data and simulated SM background passing the preselection is reported in table 6.4.1. The dominant background contributions come from Z+jet and W+jet production, while all other backgrounds contribute at the level of a few percent.

Sample	2016			2017			2018		
	muon	electron	neutrino	muon	electron	neutrino	muon	electron	neutrino
Z+jet	18280.69	14533.37	242139.92	19489.9	16243.25	281157.0	28495.67	24418.47	406784.34
W+jet	-	-	203839.34	-	-	224926.42	-	-	321918.13
$\mathrm{t\bar{t}}$	18.77	15.73	28363.9	27.12	18.24	31403.47	33.69	35.32	44983.23
WW	1.37	2.03	3248.43	2.46	1.9	3448.8	3.67	4.91	5054.0
WZ	246.96	211.16	3318.82	249.28	232.76	3489.85	363.98	313.72	4940.2
ZZ	133.41	103.97	1531.99	134.83	116.67	1616.45	207.15	182.77	2331.16
Tot. exp.	18681.2	14866.26	482442.39	19903.59	16612.83	546041.99	29104.15	24955.19	786011.06
Data	18327	15250	471484	22188	17655	512918	29504	24212	799949

Table 6.4.1: Number of events for data and different simulated SM background processes after the preselection criteria in each channel and year. The values for the neutrino channel are taken from Ref. [5].

Finally, the selection of the Higgs boson-initiated jet is performed using the DeepAK8 tagger, and described in detail in the following section. A summary of the final selection criteria used in this analysis, comprising the preselection and the requirement on the Higgs boson tagging, is reported in table 6.4.2 for each channel. The signal and background efficiencies for consecutive selection requirements are shown in figure 6.4.4. Each of the selection steps based on the DeepAK8 tagger (H4qvsQCD and ZHccvsQCD) is applied relative to the final step of the preselection, referred to as "b tag veto" in the figure.

Selection	Channel	
	Charged lepton	Neutrino
boosted Higgs boson	≥ 1 AK8 jet with $p_{\mathrm{T}} > 200\,\mathrm{GeV}$	
b tag veto	$<$ 2 b-tagged subjets (DeepCSV, loose WP)	
Higgs boson tagging	ZHccvsQCD > 0.8	
boosted Z boson	$p_{\mathrm{T}}(\ell\ell) > 200\,\mathrm{GeV}$ $\Delta R(\ell\ell) \leq 0.45$	$p_{\mathrm{T}}^{\mathrm{miss}} > 200\,\mathrm{GeV}$
Z boson mass	$81\,\mathrm{GeV} \leq m(\ell\ell) \leq 101\,\mathrm{GeV}$	
back-to-back topology	$\Delta\phi(\ell\ell, \mathrm{AK8jet}) \geq 2$	$\Delta\phi(p_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{AK8jet}) \geq 2$
QCD rejection		$\Delta\phi(p_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{AK4jet}) > 0.5$

Table 6.4.2: Summary of the final selection criteria for each channel.

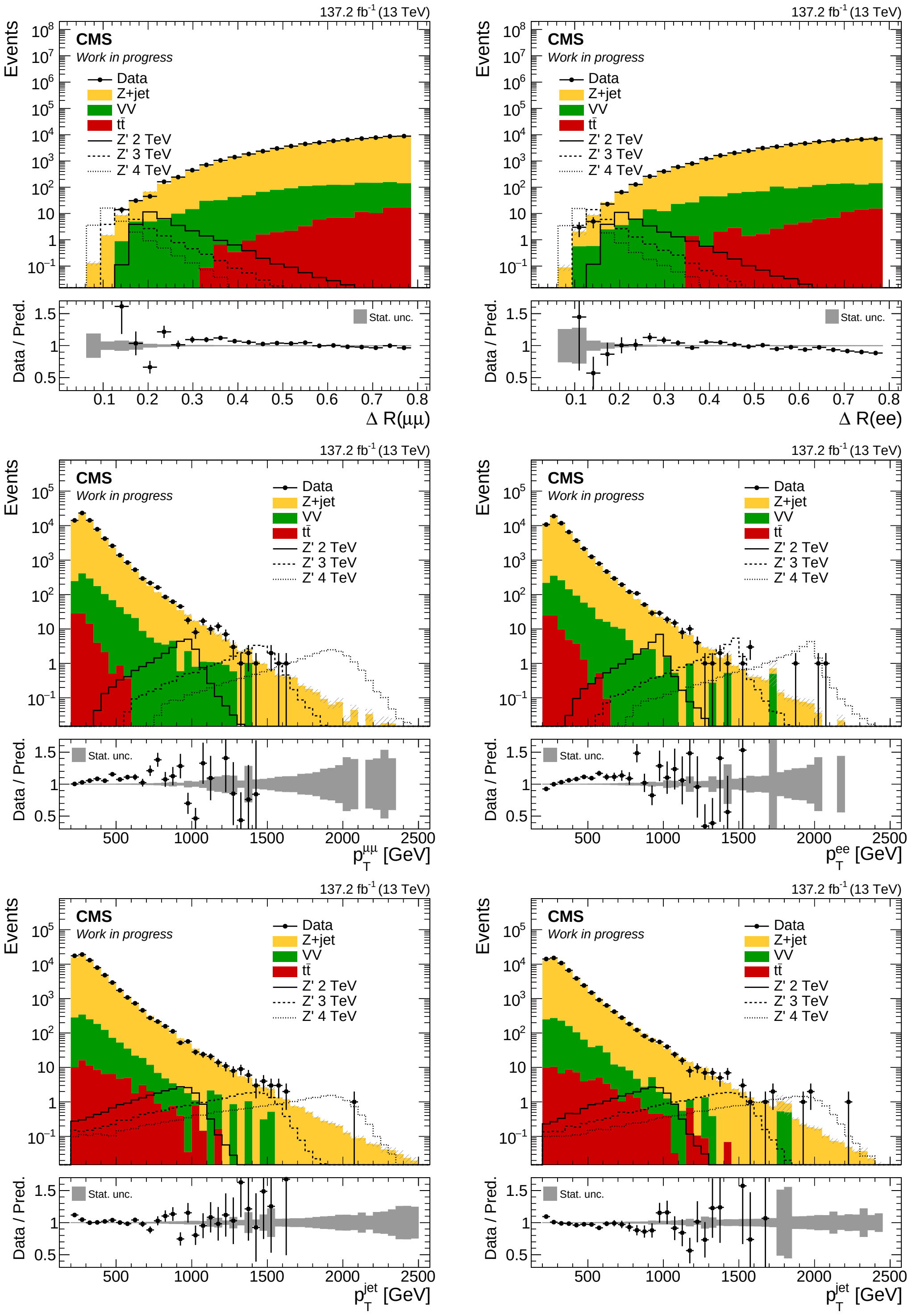


Figure 6.4.3: $\Delta R(\ell\ell)$ (upper), dilepton p_{T} (centre) and jet p_{T} (lower) distributions for the muon (left) and electron (right) channels after the preselection criteria for the full Run 2 dataset.

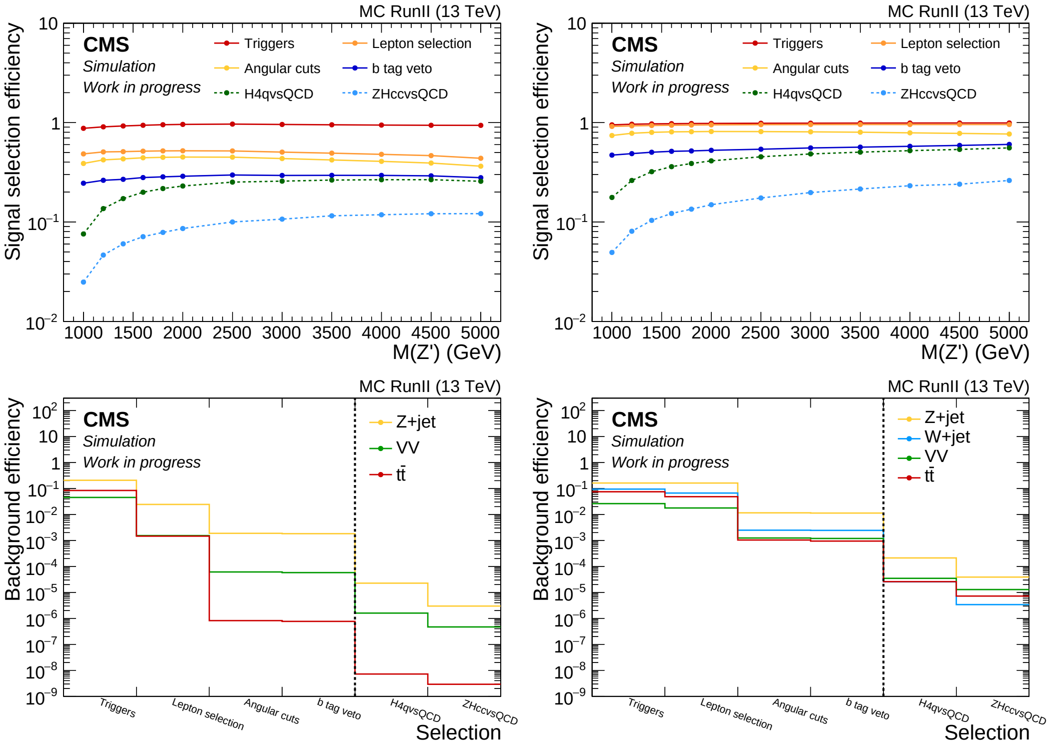


Figure 6.4.4: Selection efficiency for signal events as a function of the generated Z′ mass (upper) and for SM background events (lower) for the charged lepton (left) and neutrino (right) channels after different selection criteria. The H4qvsQCD and ZHccvsQCD selections are relative to the b tag veto selection. Upper right plot adapted from Ref. [5].

6.5 Higgs boson tagging

The identification of the jet originating from the hadronic decay of the Higgs boson constitutes a crucial step in this analysis. As described in the previous section, large-radius jets associated with a $b\bar{b}$ pair are excluded to ensure orthogonality with the analysis targeting the $H \to b\bar{b}$ final state [86]. This b tagging requirement constitutes an excellent method to suppress background events for the $H \to b\bar{b}$ search, although a large fraction of true $H \to b\bar{b}$ signal events are rejected due to inefficiencies in the b tagging algorithm. The analysis presented here is sensitive to such events, which can be used to increase the sensitivity. Furthermore, without the strong discrimination power provided by the b tagging, additional differences between signal and background events regarding jet properties need to be investigated to increase the sensitivity of this analysis. This section outlines alternative approaches for background rejection employing the jet substructure and flavour component.

After the b tagging veto requirement, approximately 46% of the selected signal events still contain true $H \to b\bar{b}$ decays, given the large BR and the b tagging inefficiency. With a share of 31%, the second largest fraction of events contains $H \to VV^*$ decays, which are equally distributed between the full-hadronic ($H \to VV^* \to qqqq$) and the semi-leptonic ($H \to WW^* \to \ell\nu qq$ and $H \to ZZ^* \to \ell\ell qq$) decay modes. In the following, these events are split into the "$H \to VV^* \to qqqq$ (merged)" and the "$H \to VV^*$ (other)" categories. The first one contains events matched to the full-hadronic decay mode where all the decay products are fully clustered inside the

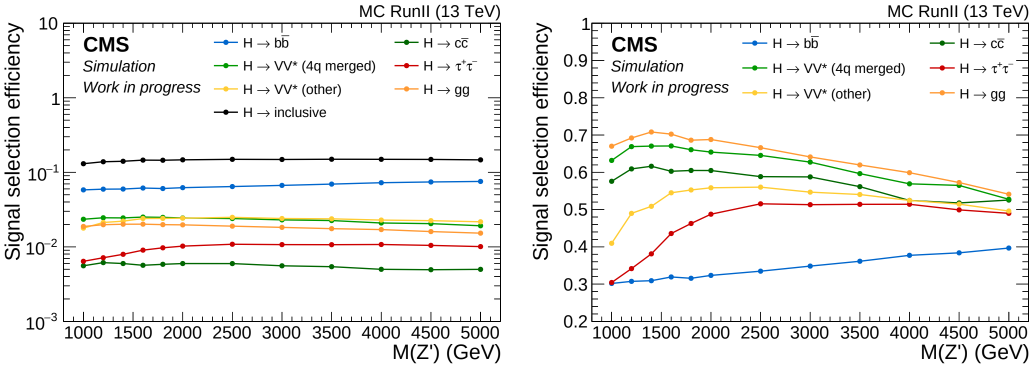


Figure 6.5.1: Selection efficiency for each Higgs boson decay mode as a function of the generated Z′ mass after the full event selection for the full Run 2 in the muon channel. The selection efficiencies with respect to the total number of events (left) and the number of events of the decay mode under consideration (right) are shown.

jet, which happens in approximately 95% of the cases. All the remaining events are included in the second category. Finally, smaller contributions arise from H → gg (12%), H → $\tau^+\tau^-$ (7%) and H → c$\bar{c}$ (4%) decays. The selection efficiency for each Higgs boson decay mode with respect to the total number of generated events after the selection requirements described in section 6.4 is shown in figure 6.5.1 (left).

Similarly, the selection efficiency with respect to the number of generated events of the decay mode under consideration is shown in figure 6.5.1 (right). An efficiency between 50% and 70% is reached for the H → qqqq, H → c$\bar{c}$ and H → gg decay modes, while only between 30% and 40% of the H → b$\bar{b}$ decays are selected.

The substructure and the flavour component of Higgs boson-initiated jets can be exploited to further increase the analysis sensitivity. In fact, jets in background events, primarily arising from V+jet processes, are mostly due to initial or final state radiation. Such jets are expected to be identified as a QCD jet, for which a substantially different substructure is expected compared to the Higgs boson decays, in particular for the 4-prong structure of the H → qqqq mode.

Another interesting feature of the final states considered in this analysis is that, besides the H → c$\bar{c}$ final state, also the H → VV* → qqqq final state contains a significant c-flavour component (cf. section 1.1.4). On the other hand, jets originating from QCD processes are mainly arising from gluons or light-flavoured quarks; this difference can be exploited to suppress background events.

Higgs-boson-initiated jets are identified using the output scores of the DeepAK8 tagging algorithm. As described in more detail in chapter 4, DeepAK8 [159] is a DNN-based multi-class tagger used to identify jets arising from a variety of boosted particles, for example, top quarks, vector or Higgs bosons. For the latter, a number of final states can be targeted, including the H → c$\bar{c}$ and H → qqqq decays. Moreover, for QCD-initiated jets, different categories are defined based on the number of b- or c-hadrons inside the jet. To maximise the discrimination power of the tagger, the scores of two output classes (X and Y) can be combined as:

$$\mathrm{XvsY} = \frac{P(\mathrm{X})}{P(\mathrm{X}) + P(\mathrm{Y})}\,, \tag{6.1}$$

where P represents the DeepAK8 score of a given output class.

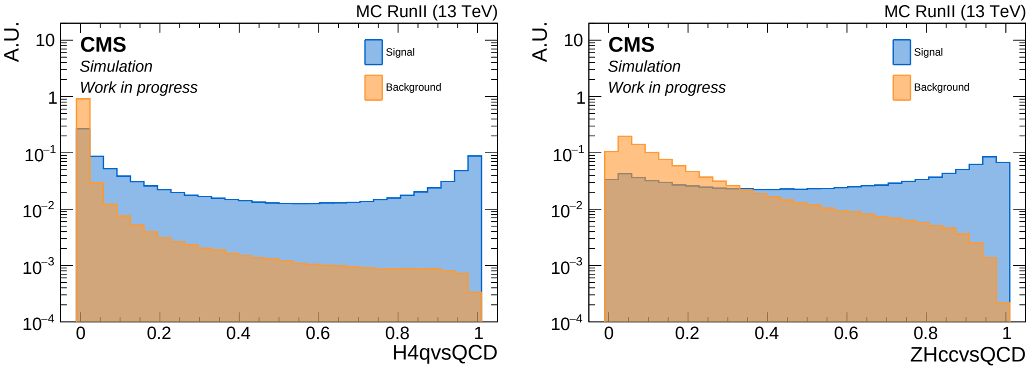


Figure 6.5.2: Distributions for the H4qvsQCD (left) and ZHccvsQCD (right) scores of the Higgs boson candidates after the preselection criteria in the muon channel. This plot is made by the author of this work; an adapted version was already shown in Ref. [5].

By comparing one Higgs boson decay class to the QCD class, defined as the sum of all QCD-related scores, it is possible to increase the separation between signal and background events, as shown in figure 6.5.2. The 4-prong (H4qvsQCD) and double-c (ZHccvsQCD[2]) scores are the most promising variables. Jets coming from backgrounds processes likely take extremely low values for these variables, whereas jets from all Higgs boson decays assume higher values. The performance of these two variables has been studied for both signal and background events. The two methods are compared and the one that maximises the overall analysis sensitivity is chosen.

4-prong tagger

It is evident that the H4qvsQCD score provides a clear separation between signal and background events. Nonetheless, a simple one-dimensional selection criterion on the H4qvsQCD score would unnecessarily hurt the signal efficiency for events with high-p_{T} jets. As illustrated in figure 6.4.3, the background is mainly composed of low-p_{T} jets, while the signal events are characterised by highly boosted jets. Therefore, a more stringent cut at low p_{T} can be used to suppress the background, while this requirement can be loosened for high-p_{T} jets. This approach ensures an optimal sensitivity across the whole phase space under consideration.

The algorithm described in the following has been developed to find the optimal p_{T}-dependent threshold that maximises the sensitivity of the analysis. For each simulated resonance mass, the optimal threshold is chosen as the DeepAK8 score corresponding to the highest significance. Using the profile likelihood ratio test [202], the significance is defined as:

$$\sqrt{q_0} = \sqrt{2 \cdot \left((s+b) \cdot \log\left(1 + \frac{s}{b}\right) - s\right)}\,, \tag{6.2}$$

where s and b are the numbers of events with the H4qvsQCD score of the selected large-radius jet higher than a given threshold for signal and background, respectively. Events are selected such that their ZH reconstructed invariant mass, described in section 6.6, lies in a range containing 95% of the signal sample under consideration.

[2] Both the Z → $c\bar{c}$ and H → $c\bar{c}$ classes are considered; the mass decorrelated version is used to derive the data-to-simulation correction, as described in chapter 4.

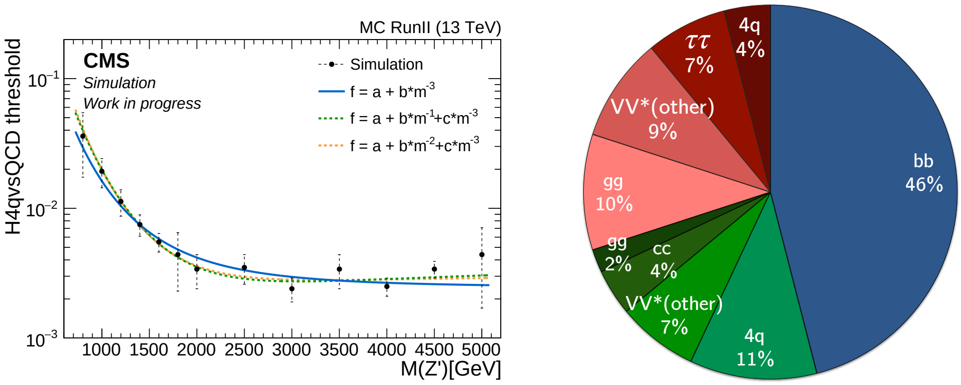


Figure 6.5.3: Left: Fit of the H4qvsQCD threshold as a function of the generated Z′ mass. The comparison between the nominal (blue) and alternative parametrizations is shown. Right: Signal composition after the full event selection for the full Run 2. The colour scheme is given in the text.

The procedure is repeated for each simulated resonance mass in each channel and year, where similar results are obtained. Therefore, events from all channels and years are combined to reduce statistical fluctuations. The result of this procedure is illustrated in figure 6.5.3 (left), demonstrating that a tighter selection is needed only for less boosted regimes. A smooth description of the tagger threshold as a function of the generated Z′ mass can be obtained by performing a fit to the outcome of the procedure described above. Several parametrisations have been tested, all with a very similar fit quality. The nominal functional form is chosen to be the one with the smallest number of parameters and with good fit results:

$$f(m(\mathrm{Z}')) = a + b \cdot m(\mathrm{Z}')^{-3}\,, \tag{6.3}$$

where m is the generated Z′ mass. The alternative parametrisations are shown in figure 6.5.3 (left) for comparison.

Double-c tagger

The flavour content of the reconstructed jet provides another powerful discriminator for background suppression. Background events are most likely to arise from gluon or light-flavour jets. Conversely, jets originating from Higgs boson decays are characterised by a relatively large c-flavour component, as illustrated in figure 6.5.3 (right). The colour scheme reflects the exclusive matching categories introduced in section 4.5. Jets containing at least one b-flavoured hadron belong to the "b-category" (blue), while jets with at least one c-flavoured hadron and no b-flavoured hadrons correspond to the "c-category" (green); last, the "light-category" (red) includes jets with no b- or c-flavoured hadrons.

Furthermore, the selected jet in each event can be matched at the parton-level with the Higgs boson decay products; the jet is matched if the decay products of the Higgs boson are found inside the jet area; the matching efficiency exceeds 99.5%. Jets matched with the $\mathrm{H} \to \mathrm{b\bar{b}}$ and $\mathrm{H} \to \mathrm{c\bar{c}}$ decay modes are virtually always associated to the b- and c-category, respectively. Moreover, approximately 75% and only 2% of the events matched to the $\mathrm{H} \to \mathrm{VV}^* \to \mathrm{qqqq}$ (merged) and $\mathrm{H} \to \mathrm{VV}^*$ (other) categories contains two parton-level c-quarks, respectively.

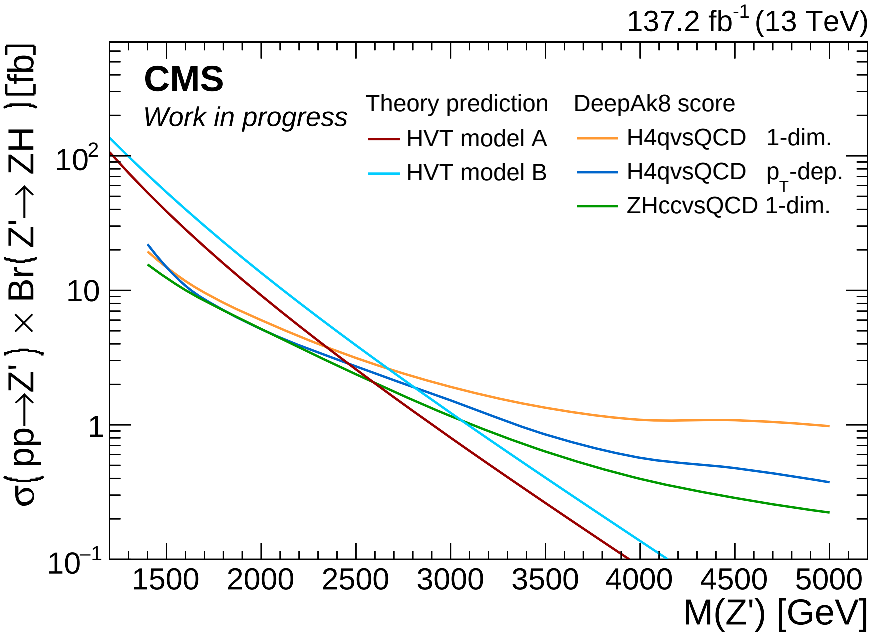


Figure 6.5.4: Expected upper limits at the 95% CL on the product of the production cross section $\sigma\,(\mathrm{pp} \to \mathrm{Z}')$ and the branching ratio $\mathrm{BR}\,(\mathrm{Z}' \to \mathrm{ZH})$ as a function of the generated Z′ mass for the combination of all channels and different selections on the DeepAK8 scores.

This high fraction of jets with a c-flavoured component motivates the usage of a double-c tagger to enhance the sensitivity of this analysis. Contrary to the H4qvsQCD case, no p_T-dependant threshold is observed, and a simple one-dimensional cut on the ZHccvsQCD score provides the optimal sensitivity in the whole mass range. The best sensitivity is achieved when ZHccvsQCD $>$ 0.8.

Choice of the tagger

The signal and background efficiencies after requiring either the p_T-dependent cut on the H4qvsQCD score, under the assumption of $p_\mathrm{T}^\mathrm{jet} \sim m(\mathrm{Z}')/2$, or the one-dimensional cut on the ZHccvsQCD score are shown in figure 6.4.4. Both selections strongly suppress the SM background events. However, a more efficient rejection is achieved in the latter case. The requirement on the H4qvsQCD variable has a minor effect on the number of selected signal events; on the contrary, requiring a minimum on the ZHccvsQCD score more notably reduces the signal selection efficiency. Taking these effects into account, the usage of the ZHccvsQCD score provides the best sensitivity. The comparison of the expected cross section upper limits is shown in figure 6.5.4; the results obtained with a one-dimensional cut on the H4qvsQCD score show the improvement obtained with the p_T-dependent cut. A detailed description of the limit-setting procedure is given in section 6.10.

The signal-enriched region (SR) of this analysis includes events that pass all selection criteria, including the Higgs boson tagging performed by selecting a jet with ZHccvsQCD $>$0.8; the VR, used to validate the background strategy, is defined by inverting the ZHccvsQCD requirement.

The selection efficiency in the SR for each Higgs boson decay mode is shown in figure 6.5.5 with respect to the total number of generated events (left) and the number

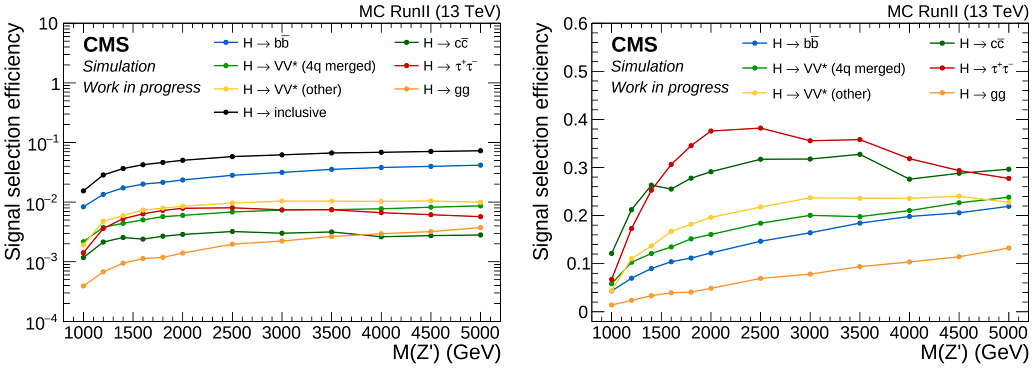


Figure 6.5.5: Selection efficiency for each Higgs boson decay modes as a function of the generated Z′ mass after the cut of the ZHccvsQCD score for the full Run 2 in the muon channel. The selection efficiencies with respect to the total number of events (left) and the number of events of the decay mode under consideration (right) are shown.

of generated events of the decay mode under consideration (right). The selection efficiency is higher towards high values of the generated Z′ mass. Events matched to the H → $c\bar{c}$ decay mode show a higher selection efficiency than those matched to the H → VV* and H → $b\bar{b}$ decay modes, although the latter constitute the largest fraction (approximately 85%) of the total signal events. Events matched to the H → gg decay mode are largely suppressed, being relatively similar to QCD-like events. Furthermore, the double-c tagger is able to select several final states besides the H → $c\bar{c}$ decay. Therefore, its usage makes the analysis presented here sensitive to several hadronic decays of the Higgs boson.

6.6 Z′ candidate reconstruction

This search is performed by examining the distribution of the invariant mass of the reconstructed Z and Higgs boson candidates in the SR, where a potential signal would result in a localised excess over a monotonically decreasing background.

A natural and commonly used approach is to reconstruct the whole Z′ boson decay chain using a "bottom-up" approach, from the individual reconstructed objects, through the Z and Higgs bosons, to the Z′ boson itself. In particular, a large-radius jet is used as a candidate for the H boson, while the Z boson is identified via the charged lepton pair in the muon and electron channel. Then, the 4-momenta of the H and Z boson candidates are combined to reconstruct the Z′ boson candidate. The invariant mass of the reconstructed ZH system is expected to produce a sharp distribution localised around the generated mass value of a given signal sample; in contrast, background processes tend to produce smooth non-resonant distributions, which allow for a clear identification of potential signals as a peak on top of the expected SM background.

In the neutrino channel, the Z boson can be reconstructed only from the $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$. As a consequence, only the transverse mass of the Z′ candidate is calculated using the formula:

$$m_{\mathrm{T}} = \sqrt{2 \cdot p_{\mathrm{T}}^{\mathrm{jet}} p_{\mathrm{T}}^{\mathrm{miss}} \cdot \left(1 - \cos \Delta\phi(\mathrm{jet}, p_{\mathrm{T}}^{\mathrm{miss}})\right)}\,. \tag{6.4}$$

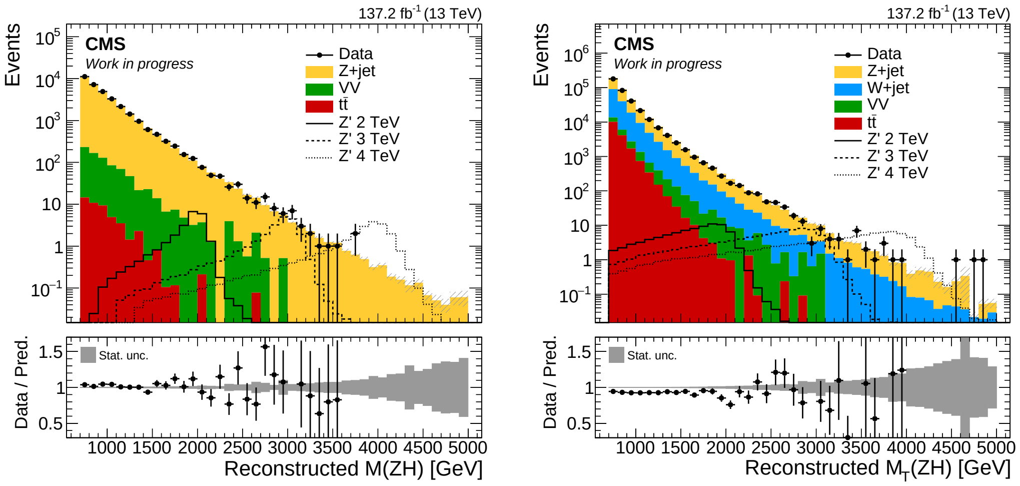


Figure 6.6.1: Reconstructed Z′ invariant mass (left) and transverse mass (right) distributions for the muon and neutrino channels, respectively. Right plot adapted from Ref. [5].

The resulting distributions of the Z′ invariant mass and transverse mass in the VR are shown in figure 6.6.1, where the different behaviour of signal and background events is evident. Good data-to-simulation agreement is observed in the charged lepton channel; despite the well-modelled shape, a normalisation offset is present in the neutrino channel. However, the background estimation, fully relying on data and discussed in more detail in the following section, is insensitive to this effect.

Examples of the reconstructed invariant mass and transverse mass distributions after the Higgs boson tagging selection for signal events in the muon and neutrino channels, respectively, are shown in figure 6.6.2, where the Higgs boson decay modes are shown in different colours. For events in which all the decay products are reconstructed inside the large-radius jet, a sharper distribution around the value of the generated Z′ mass is observed. The tails towards lower masses arise from the $\mathrm{H} \to \tau^+\tau^-$ and the "$\mathrm{H} \to \mathrm{VV}^*$ (other)" category, which includes events with at most three quarks matched with a jet and the semi-leptonic decays. In those cases, part of the energy is lost due to the undetected neutrinos in the final states or not clustered inside the jets; therefore, the total 4-momentum of the Z′ is not well reconstructed.

Moreover, a wider distribution is also observed for the $\mathrm{H} \to \mathrm{b\bar{b}}$ decays, for which soft hadronisation products are groomed away by the SD algorithm. After the Higgs boson tagging, a larger relative contribution of the $\mathrm{H} \to \mathrm{c\bar{c}}$ decays is observed. Similar conclusions hold for the neutrino channel, in which the presence of large tails in the transverse mass distribution is expected. A good reconstruction of the Z′ invariant mass distribution is obtained in all channels; the distribution shows a peak at the generated Z′ mass, with a resolution of approximately 4% and 6% for the charged lepton and neutrino channels, respectively.

6.7 Background modelling

The steeply falling ZH invariant mass distribution of the background presented in figure 6.6.1 can be modelled with a monotonically decreasing functional form. This approach for the background estimation comes with the advantage of a reduction of

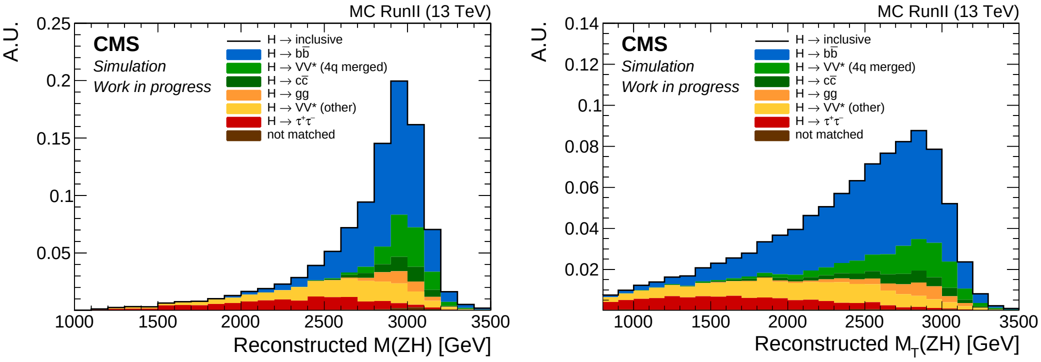


Figure 6.6.2: Reconstructed Z′ invariant mass (left) and transverse mass (right) distributions for samples generated with Z′ mass of 3 TeV for the muon and neutrino channels, respectively, after the Higgs boson tagging selection.

statistical fluctuations. Moreover, a fit to the distribution in data rather than relying on simulated events allows for a substantial reduction of systematic uncertainties associated to the simulated background expectation.

In order to reduce experimental biases, a "blind" procedure is performed. Therefore, the background modelling strategy is first validated in the VR, under the assumption that any functional form that describes the background well in the SR can also be used to fit the background in the VR. This hypothesis is tested using SM simulated background events in the signal and validation regions and the distribution of data in the VR. The functional forms used to fit the Z′ mass distribution of the background are given by:

$$f_N(x) = \exp(\sum_{i=0}^{N} p_i \cdot x^i)\,, \tag{6.5}$$

where N represents the degree of the polynomial function in the exponent. Examples of the fits of the background for simulated events and for data in the VR are shown in figure 6.7.1. The fitting range excludes the turn-on created by the kinematic selection criteria. Different values of N are examined, and a good description is generally obtained in all cases for mass values up to 6 TeV. In order to identify the functional form that describes the background distribution well with the least number of parameters, the F-test [203] is used. The F-statistic is given by:

$$F(a,b) = \frac{\chi_a^2 - \chi_b^2}{p_b - p_a} \Big/ \frac{\chi_b^2}{n - p_b - 1}\,, \tag{6.6}$$

where p_x and χ_x^2 are the number of free parameters and the χ^2, respectively, of the two hypotheses to be tested, and n is the number of fitted points. The two models are compared to determine if the higher number of parameters ($p_a < p_b$) provides a significantly better fit. The model with more freedom is rejected if the probability of F being distributed as an F-distribution with ($p_a - p_b$, $n - p_b$) degrees of freedom is larger than 0.05. The outcome of the F-test, performed for different channels in both SR and VR, suggests $N = 2$ in approximately 90% of the cases. Therefore, this functional form is used for the final statistical interpretation of the results.

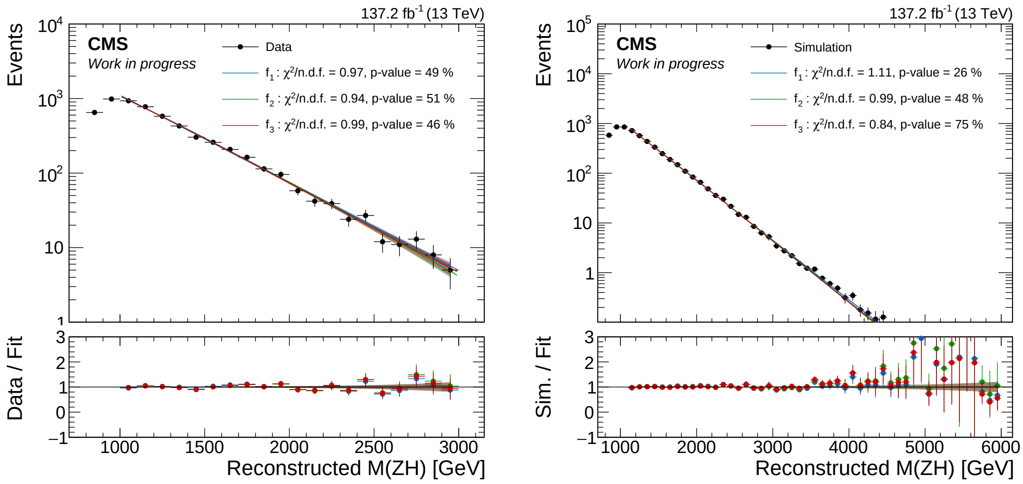


Figure 6.7.1: Background modelling of the reconstructed Z′ invariant mass distributions for data (left) and simulated events (right) in the VR for the muon channel in Run 2 using different functional forms.

6.8 Signal modelling

As shown in figure 6.6.2, the reconstructed invariant mass of signal samples consists of a Gaussian core, centred around the generated sample mass, and asymmetric tails, resulting from the partial loss or misreconstruction of the energies and momenta of the decay products.

Therefore, the Z′ mass distribution for the signal samples is modelled in the SR with an empirical function. The functional form chosen is the Crystal Ball function [204, 205]:

$$f(x;\bar{x},\sigma,\alpha,n) = \begin{cases} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma}\right)^2}, & \text{for } \dfrac{x-\bar{x}}{\sigma} > -\alpha \\ \left(\dfrac{n}{|\alpha|}\right)^n e^{-\frac{|\alpha|^2}{2}} \left(\dfrac{n}{|\alpha|} - |\alpha| - \dfrac{x-\bar{x}}{\sigma}\right)^{-n}, & \text{for } \dfrac{x-\bar{x}}{\sigma} \leq -\alpha \end{cases}, \tag{6.7}$$

which consists of a Gaussian core and a power-law tail. Examples of the fits to the reconstructed Z′ invariant mass and transverse mass distributions for signal events in the muon and neutrino channels, respectively, are shown in figure 6.8.1. The events correspond to samples generated with Z′ mass of 3 TeV.

Despite being a typical choice for modelling the resonance mass, the Crystal Ball function often leads to numerically unstable fits due to the power-law parameter. This effect is observed in particular for the transverse mass distributions in the neutrino channel; despite the relatively good modelling of the distribution, the parameter uncertainties, especially for the n parameter, can be significant, producing unstable results. To reduce this effect, the *ExpGaussExp* function [206] is used as an alternative parametrisation:

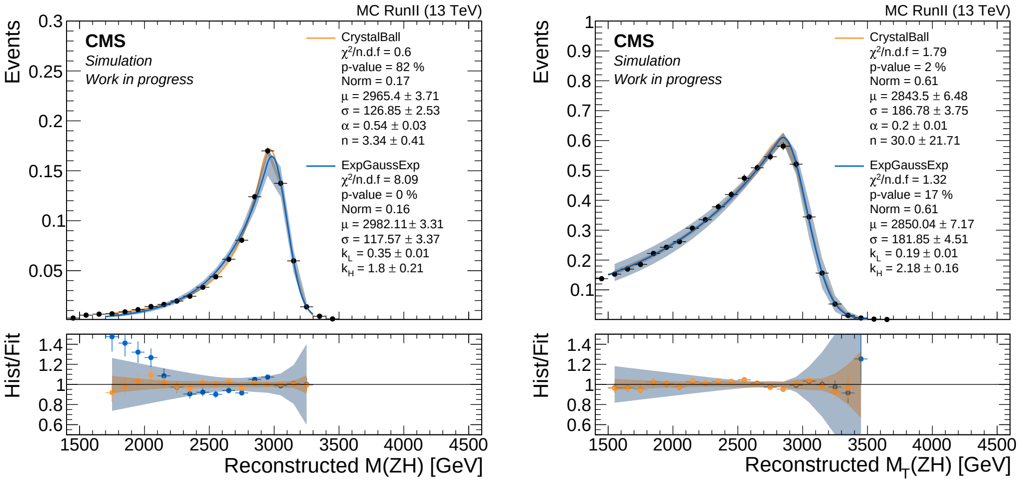


Figure 6.8.1: Signal modelling of the reconstructed Z′ invariant mass and transverse mass distributions in the muon (left) and neutrino (right) channels, respectively. The events correspond to samples generated with Z′ mass of 3 TeV. Right plot taken from Ref. [5].

$$f(x;\bar{x},\sigma,k_L,k_H) = \begin{cases} e^{\frac{k_L^2}{2}+k_L\left(\frac{x-\bar{x}}{\sigma}\right)}, & \text{for} \qquad \dfrac{x-\bar{x}}{\sigma} \leq -k_L \\ e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma}\right)^2}, & \text{for} \; -k_L < \dfrac{x-\bar{x}}{\sigma} \leq k_H \\ e^{\frac{k_H^2}{2}-k_H\left(\frac{x-\bar{x}}{\sigma}\right)}, & \text{for} \quad k_H < \dfrac{x-\bar{x}}{\sigma} \end{cases} . \tag{6.8}$$

This alternative function presents a Gaussian core and two independent exponential tails, which result in a numerically more stable fit. The opposite behaviour is observed for the charged lepton channels, where the Crystal Ball function is able to describe the tails better. The comparison of these fits is shown in figure 6.8.1. The signal samples considered are well described in the charged lepton and neutrino channels by the Crystal Ball and *ExpGaussExp* functions, respectively.

6.9 Systematic uncertainties

Several systematic uncertainties can affect the Z′ mass distribution for signal samples and, thus, the final results. The modified Z′ mass distributions corresponding to the up and down variation by one standard deviation σ of each source of uncertainty are derived and modelled as described in the previous section. The primary sources of experimental and theoretical uncertainties, treated as uncorrelated among themselves, are discussed in the following.

Figure 6.9.1 shows the results of the signal modelling fits of the nominal distributions, as well as the shapes relative to each systematic variation, for several generated Z′ masses. The most significant variation to the fit results is in the overall yield, while minor deviations from the nominal values are observed for the other parameters, as shown in figure 6.9.2. Moreover, the theory uncertainty related to the choice of the PDF constitutes the largest variation for high-mass samples. The

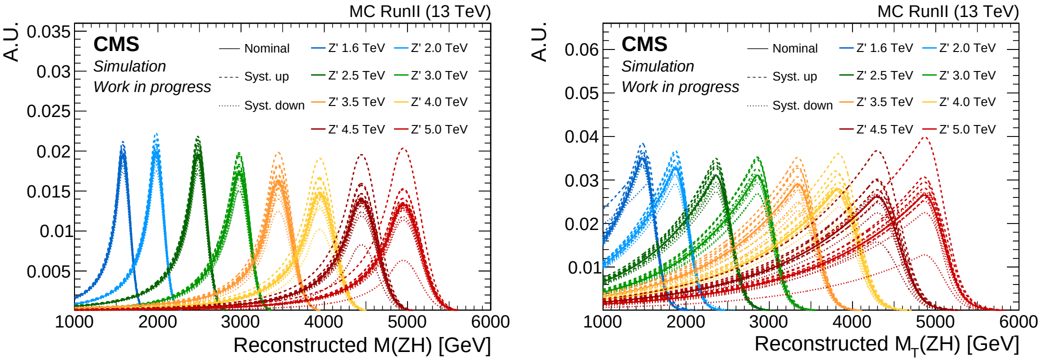


Figure 6.9.1: Shapes of the reconstructed Z′ mass for signal samples in the muon (left) and neutrino (right) channels. Systematic variations shown as dashed and dotted lines. These plots are made by the author of this work; an adapted version was already shown in Ref. [5].

other sources show a constant effect for all masses; in particular, the uncertainty associated to jet tagging are dominant compared to all other sources.

Luminosity

The integrated luminosity of the data recorded by the CMS Collaboration is measured with an uncertainty of 1.8%, as detailed in section 2.2.5. As the luminosity value defines the normalisation for simulated signal processes, this uncertainty affects the expected yield of the Z′ mass distributions.

Pileup

As described in section section 6.3, a pileup reweighting procedure is used in simulated samples to match the observed pileup distribution in data. In order to quantify the impact of the pileup modelling, the reweighting is repeated by varying the minimum bias cross section within its uncertainty ($69.2 \pm 4.6\%\,\text{mb}$). This uncertainty affects both the normalisation and shape of the simulated signal samples.

Jet energy corrections

The jet energy scale and resolution have been corrected to the particle level. These corrections are varied within 1 σ of their corresponding uncertainties and the analysis is repeated with the varied jet energy. Moreover, the modified JES is propagated to the missing transverse momentum (Type-I correction). These uncertainties affect both the normalisation and shape of the simulated signal samples.

Jet tagging

The DeepCSV and DeepAK8 taggers are calibrated using data-to-simulation SFs (cf. section 4.5). These corrections, derived for b-, c- and light-flavoured jets, are varied within 1 σ of their corresponding uncertainties. The different-flavour variations are considered correlated, while the DeepCSV and DeepAK8 variations are considered uncorrelated. These uncertainties affect both the normalisation and shape of the simulated signal samples.

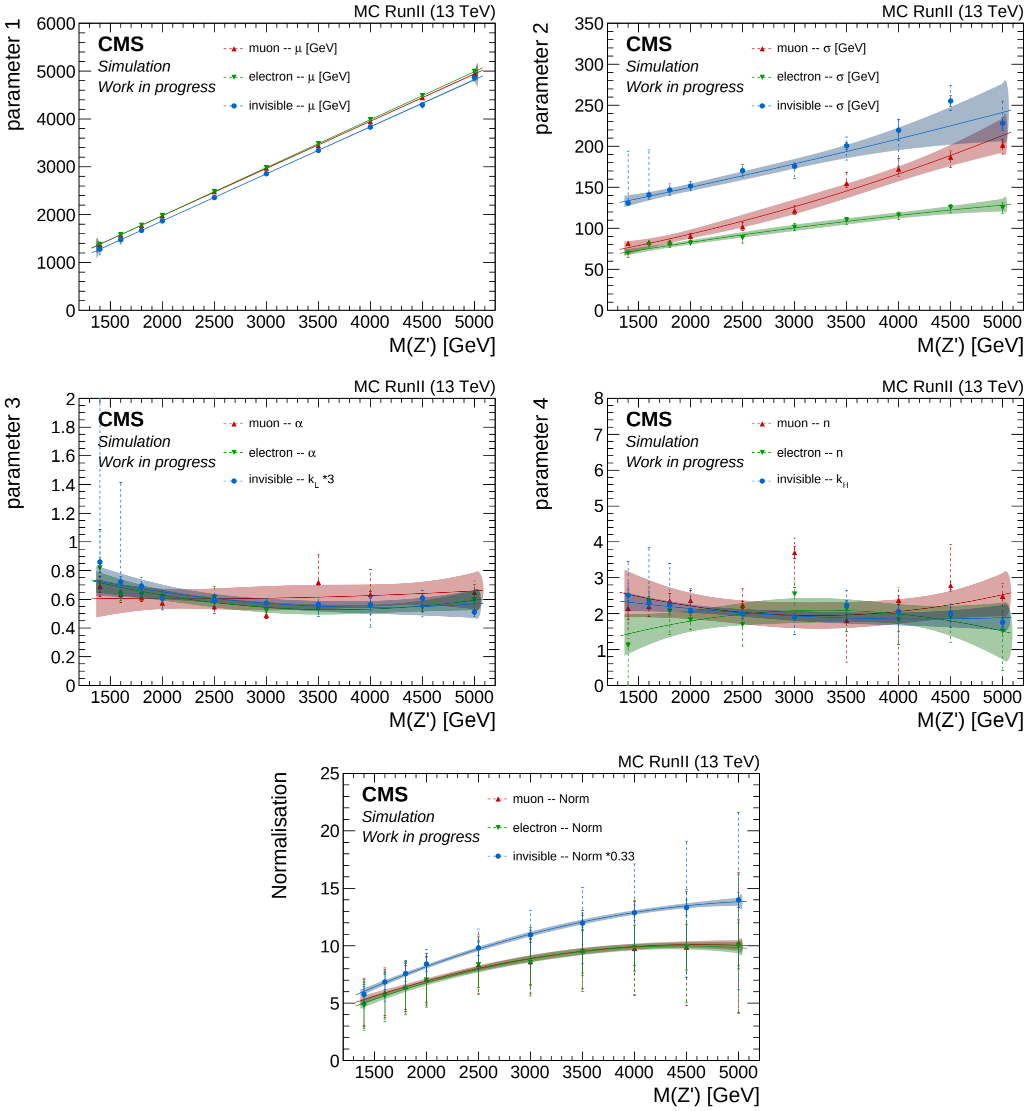


Figure 6.9.2: Fit parameters as a function of the generated Z′ mass for different channels. The points indicate the parameter values resulting from the nominal Z′ mass distribution fit in signal events. The solid error bars correspond to the statistical uncertainty of the fit. The dashed error bars correspond to variations with respect to the nominal fit value obtained from the fit of the Z′ mass distributions relative to all the systematic variations. These plots are made by the author of this work; an adapted version was already shown in Ref. [5].

Lepton efficiencies

The uncertainties of the data-to-simulation SF for the efficiency of lepton reconstruction, ID, and trigger selections are varied within 1 σ. These uncertainties affect both the normalisation and shape of the simulated signal samples.

Prefiring

The uncertainty in the prefiring weights (see section 6.4) are applied for 2016 and 2017 samples and affects only the expected yield.

Theory uncertainties

Since the Z′ resonance is assumed to be produced via the strong interaction, differences are to be expected for different QCD scale values and different choices of the PDF used to generate the signal events. Two sets of uncertainties related to theoretical uncertainties are used in this analysis and described below.

The uncertainty in the choice of the perturbative QCD renormalisation (μ_R) and factorisation (μ_F) scales is taken into account with the following procedure based on Ref. [165]. First, both μ_R and μ_F are independently varied up and down with respect to their nominal values by a factor 1/2 and 2, respectively, resulting in eight different configurations of the Z′ mass distribution. The envelope of all the distributions, including the nominal one and excluding the two unphysical configurations with variations in opposite directions, is taken as systematic uncertainty in the choice of μ_R and μ_F. This uncertainty affects both the normalisation and shape of the simulated signal samples.

Furthermore, the systematic uncertainty in the choice of the PDF set used to generate the signal samples (see section 6.3) is taken into account by considering 100 replicas of the NNPDF set to construct 100 varied distributions of the reconstructed Z′ mass. For each bin, the RMS of all the variations defines the uncertainty associated to the PDF set used, and it affects both the normalisation and shape of the simulated signal samples.

6.10 Statistical interpretation

In order to give a quantitative statement on the possible presence of a signal, a statistical inference, based on the $\mathrm{CL_s}$ method [207], is performed. This method consists of a likelihood fit of the signal strength modifier μ, where the *background-only* (b) and *signal + background* ($s+b$) hypotheses are tested. The modified frequentist approach [208] adopted in this analysis has been developed at the LHC by the ATLAS and CMS collaborations in the context of the Higgs boson search results, and it uses the profile likelihood ratio $\tilde{q}_\mu$ [209] as a test statistics:

$$\tilde{q}_\mu = -2\ln\frac{\mathcal{L}(n\,|\,\mu,\hat{\theta}_\mu)}{\mathcal{L}(n\,|\,\hat{\mu},\hat{\theta})} \qquad \text{with} \quad 0 \le \mu \le \hat{\mu}\,. \tag{6.9}$$

Here, $\hat{\theta}_\mu$ represents the profiled maximum-likelihood estimator of the nuisance parameters θ, while $\hat{\mu}$ and $\hat{\theta}$ correspond to the values of θ and μ that globally maximise the likelihood. The likelihood function $\mathcal{L}$ is defined as:

$$\mathcal{L}(n\,|\,\mu,\theta) = \mathcal{P}\left(\mathrm{n}|\mu\cdot s(\theta)+b(\theta)\right)\cdot\Pi(\theta)\,, \tag{6.10}$$

where the first term is the Poisson probability of observing n events in data, given the number of expected background (b) and signal (s) events, and the second factor is the product of the priors of all the nuisance parameters. The $\mathrm{CL_s}$ statistic is defined as the ratio of the p-value of the *signal + background* and *background-only* hypotheses given the actual experimental observation (obs):

$$\mathrm{CL_s}(\mu) = \frac{\mathrm{CL_{s+b}}(\mu)}{\mathrm{CL_b}(\mu)} = \frac{P(\tilde{q}_\mu \ge \tilde{q}_\mu^{\mathrm{obs}}\,|\,s+b)}{P(\tilde{q}_\mu \ge \tilde{q}_\mu^{\mathrm{obs}}\,|\,b)}\,. \tag{6.11}$$

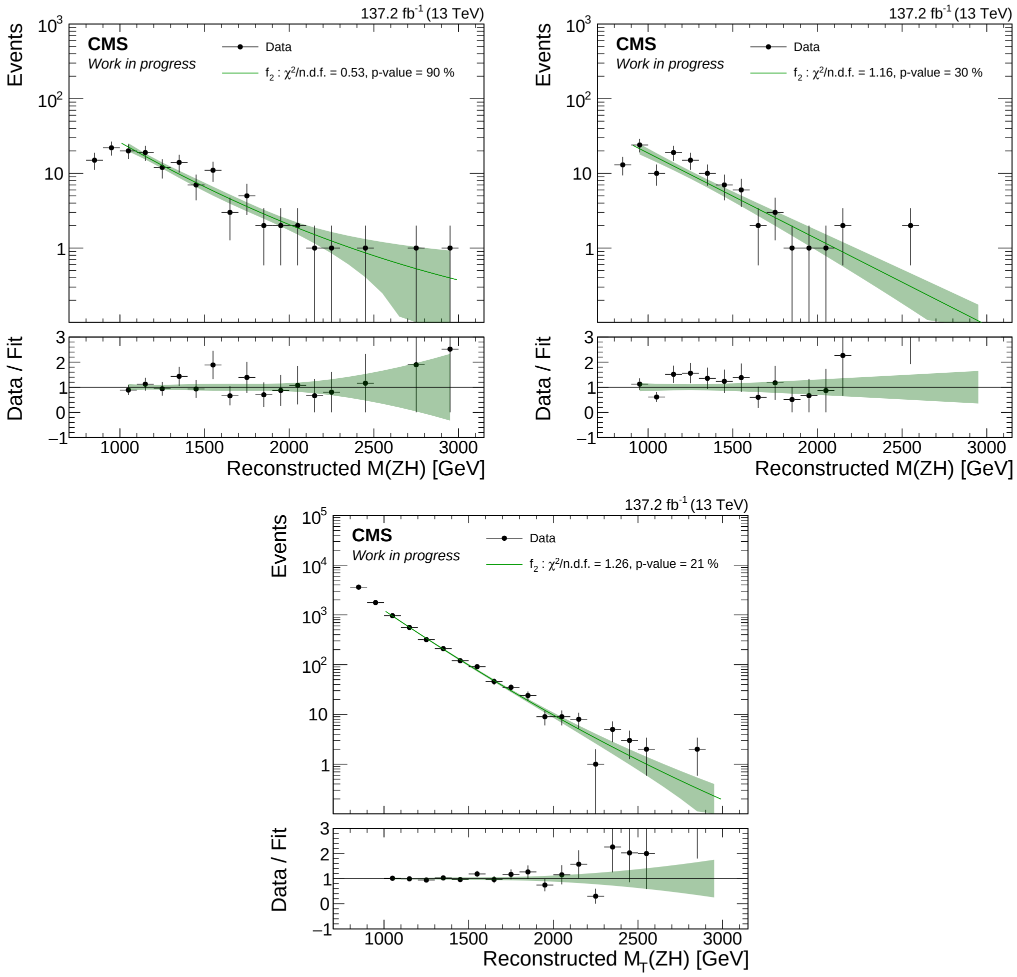


Figure 6.11.1: Fit of the reconstructed Z′ invariant mass distributions in data in the SR under the background-only hypothesis for the muon (upper left), electron (upper right), and neutrino (lower) channels.

The signal strength $\mu_{95\%}$ corresponding to $\mathrm{CL_s}(\mu) = 0.05$ is used to extract the 95% confidence level (CL) upper limit on the product of the production cross section $\sigma\,(\mathrm{pp} \to \mathrm{Z}')$ and the branching ratio $\mathrm{BR}\,(\mathrm{Z}' \to \mathrm{ZH})$.

The *Combine* software package [210] is used to perform the statistical procedure described above. In particular, the background and signal functional forms are passed as inputs, as well as the nuisance parameters with a log-normal prior of each of the systematic uncertainties described in section 6.9.

6.11 Results

Expected exclusion limits on the product of the production cross section and the branching ratio are derived under the background-only hypothesis to evaluate the sensitivity of the analysis in the absence of signal. The results obtained from the fit of the reconstructed Z′ mass distributions in data in the SR for the different channels using the background-only hypothesis are shown in figure 6.11.1. No deviation from the expectation is observed.

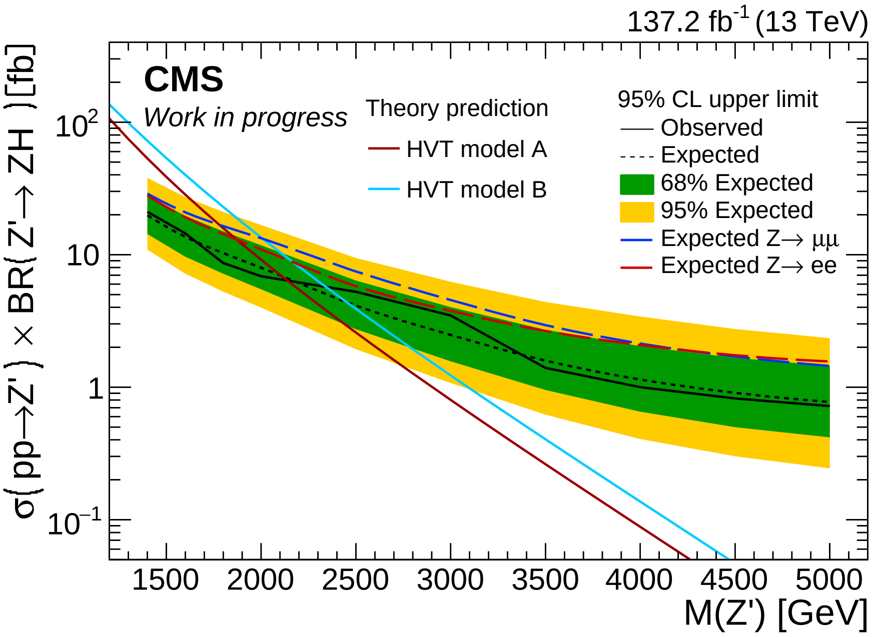


Figure 6.11.2: Expected and observed upper limits at the 95% CL on the product of the production cross section $\sigma\,(\mathrm{pp} \to \mathrm{Z}')$ and the branching ratio $\mathrm{BR}\,(\mathrm{Z}' \to \mathrm{ZH})$ for the combination of the $\mathrm{Z} \to \mu\mu$ and $\mathrm{Z} \to \mathrm{ee}$ channels as a function of the generated Z' mass. The expected upper limits for the individual channels are shown with coloured dashed lines as a comparison. The coloured solid lines show the production cross section predicted by the HVT model.

Therefore, upper limits at the 95% CL are placed on the product of the production cross section and the branching ratio as a function of the Z' mass. The statistical procedure described in the previous section is adopted for this purpose. The results obtained from the profile likelihood fit for the Z' mass distributions of signal and background for the electron and muon channels are shown in figure 6.11.2.

These channels have similar sensitivity and, given the orthogonal event selections of the two decay modes, the signal regions employed in both channels are combined statistically in a simultaneous maximum-likelihood fit to improve the results. The systematic uncertainties detailed in the previous sections are treated as fully correlated across channels. The resulting expected exclusion limits for the charged lepton channel is shown in figure 6.11.2.

The combination of the two individual channels improves the expected upper limits by approximately 40% at low masses and up to a factor of 2 at high masses compared to the individual channels alone. Furthermore, observed exclusion limits are derived from the recorded data. Good agreement between the observed and expected exclusion limits is observed across the entire mass range considered. Lower limits on the resonance mass are placed at 2.14 and 2.34 TeV, depending on the theoretical model under consideration, for resonances decaying exclusively into a Z and a Higgs boson.

The sensitivity of the charged lepton channel is compared with that of the neutrino channel [5]. Similar sensitivities are expected at low masses; the neutrino channel dominates towards higher masses, thanks to the larger BR and the higher selection efficiencies. These two channels are also combined to maximise the sensitivity.

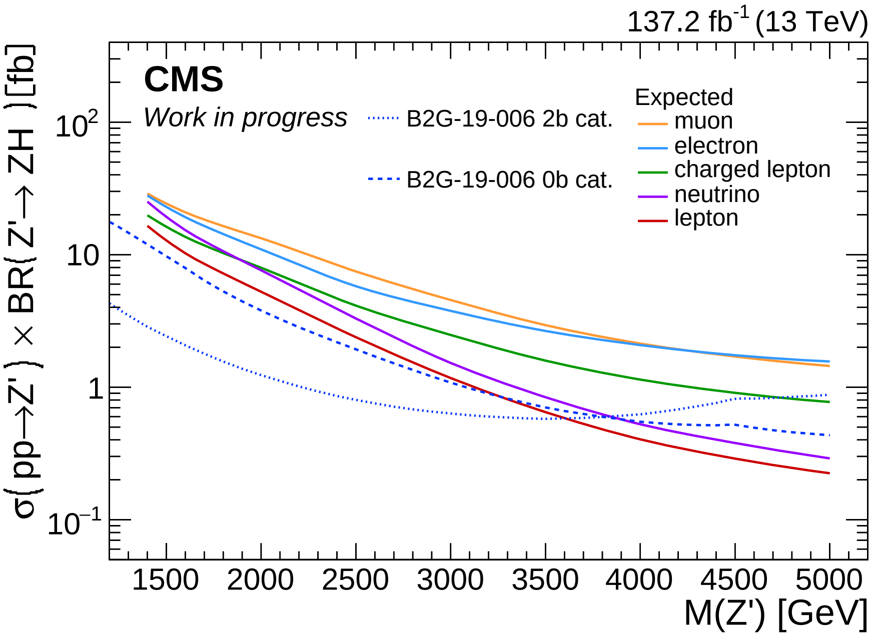


Figure 6.11.3: Expected upper limits at the 95% CL on the product of the production cross section $\sigma\,(\mathrm{pp} \to \mathrm{Z}')$ and the branching ratio $\mathrm{BR}\,(\mathrm{Z}' \to \mathrm{ZH})$ for different channels as a function of the generated Z′ mass. The expected upper limits relative to the analysis described in Ref. [86], referred to as "B2G-19-006", are reported as a comparison. This plot is made by the author of this work; an adapted version was already shown in Ref. [5].

Similar to the procedure described above, the statistical combination of all leptonic decay modes of the Z boson ($\mathrm{Z} \to \ell\ell$ and $\mathrm{Z} \to \nu\nu$) is performed. The expected exclusion limits for each channel and their combination are shown in figure 6.11.3. A further improvement between 30% and 50% from the combination of the charged lepton and neutrino channels is found. The final expected and observed exclusion limits resulting from the combination of the muon, electron and neutrino channels are compared in figure 6.11.4. The observed limits agree with the background-only expectation within about one standard deviation over the full mass range considered. An improvement on the mass exclusion limit is obtained with respect to the charged lepton channel, setting the new lower limit on the Z′ mass at 2.45 TeV and 2.72 TeV, depending on the theoretical model considered.

6.12 Sensitivity compared to other results

The light-flavoured hadronic decays of the Higgs boson, with particular attention to the 4-prong ($\mathrm{H} \to \mathrm{VV}^* \to \mathrm{qqqq}$) and the c flavour ($\mathrm{H} \to \mathrm{c\bar{c}}$) decays, had not been considered so far in the context of diboson resonance searches by the CMS Collaboration (cf. section 1.4). The results presented here are, therefore, unprecedented. Nonetheless, the results presented in the previous section can be compared to existing analyses targeting different final states. In particular, the results are compared to the analysis reported in Ref. [86]. The main focus of the analysis is placed on the $\mathrm{H} \to \mathrm{b\bar{b}}$ final states (2b cat.). Thanks to the larger BR and the efficient background suppression obtained with the b tagging requirement, this final state has the

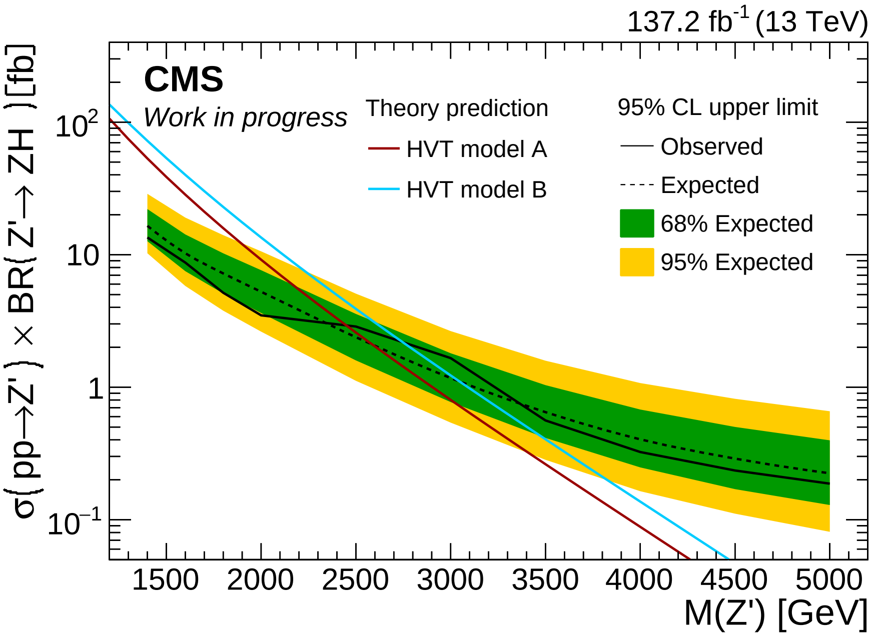


Figure 6.11.4: Expected and observed upper limits at the 95% CL on the product of the production cross section $\sigma\,(\mathrm{pp} \to \mathrm{Z}')$ and the branching ratio $\mathrm{BR}\,(\mathrm{Z}' \to \mathrm{ZH})$ for the combination of the $\mathrm{Z} \to \mu\mu$, $\mathrm{Z} \to \mathrm{ee}$ and $\mathrm{Z} \to \nu\nu$ channels as a function of the generated Z' mass. The coloured solid lines show the production cross section predicted by the HVT model.

highest sensitivity in the low mass range. The inverted b tagging selection is used for the analysis presented in this thesis; consequently, both analyses can be combined to improve the overall sensitivity. This combination is foreseen in the view of the Run 2 Legacy combination of all searches involving diboson resonances within the CMS Collaboration.

Furthermore, an additional category (0b cat.) is considered in Ref. [86] to increase the sensitivity of the search at high masses. This category utilises the same b tagging veto requirement adopted in this work; therefore, the two analyses cannot be considered orthogonal. This second category includes all hadronic Higgs boson final states, without particular emphasis on any specific final state. As shown in figure 6.11.3, the search sensitivity is improved by up to a factor of 2 at the highest resonance masses considered by focusing on the c flavour component of the hadronic Higgs boson final states with a dedicated event selection.

A similar search has been performed by the ATLAS Collaboration [211]. This recent result, using the full Run 2 dataset, focuses on the $\mathrm{H} \to \mathrm{b\bar{b}}$ decay mode and includes the $\mathrm{Z} \to \mu\mu$, $\mathrm{Z} \to \mathrm{ee}$, and $\mathrm{Z} \to \nu\nu$ final states. The ATLAS result can be directly compared to that obtained by the CMS Collaboration. In particular, the expected upper limits of Ref. [86] exceed the ATLAS ones by approximately 20% in the low mass region and by up to a factor of 2 in the high mass region. The significant improvement in the CMS result is achieved by combining the 2b and 0b categories and analysing all the hadronic Higgs boson final states. Besides, no dedicated analysis for the c flavoured final states of the Higgs boson has been performed by the ATLAS Collaboration so far [212].

6.13 Summary and outlook

A novel search for the resonant production of a hypothetical spin-1 massive particle decaying into a Z and a Higgs boson is presented. The analysis is performed using the pp collision data recorded in the years 2016 to 2018 by the CMS detector at the LHC at $\sqrt{s} = 13\,\text{TeV}$, corresponding to an integrated luminosity of about $137\,\text{fb}^{-1}$.

The final states with two electrons or muons arising from the Z boson decay and a large-radius jet, used to reconstruct the light-flavoured hadronic decays of the Higgs boson, are studied. The charged leptons in the final state provide a clean experimental signature, which is used to suppress events from SM background processes. The search targets the 4-prong ($\text{H} \to \text{qqqq}$) and c flavour ($\text{H} \to \text{c}\bar{\text{c}}$) decays, which are explicitly investigated for the first time in the context of diboson resonances in the CMS Collaboration. State-of-the-art techniques for identifying Higgs boson-initiated jets are employed to enhance the sensitivity of these final states. The event selection is optimised to ensure orthogonality with other analyses, particularly the one targeting the $\text{H} \to \text{b}\bar{\text{b}}$ final state, for a future statistical combination of the results. A statistical combination of the channels defined by the Z boson decay into a pair of electrons, muons and neutrinos is performed within the context of this thesis.

The results obtained show a sensitivity for high resonance masses that exceeds that of the $\text{H} \to \text{b}\bar{\text{b}}$ channel despite its much larger branching fraction. Moreover, an improvement of a factor of 2 compared to existing results is achieved in the high mass regime.

The sensitivity of the search presented in this chapter is currently limited by the statistical precision available from the data. Advanced analysis techniques can improve the current prediction of the SM background, and therefore the results of this analysis. Furthermore, the application of even more powerful jet tagging algorithms will enhance the signal selection efficiency while maintaining a similar background rejection power. In particular, this analysis can profit from the newer ParticleNet tagger, presented in chapter 4. Not only the c flavoured final states but also other Higgs boson decay modes will benefit from this tagger. Therefore, thorough studies are needed in the future to maximise the selection efficiency for the different final states and obtain the best sensitivity in the whole mass range.

The successful Run 2 data acquisition period at the LHC resulted in a remarkable dataset of approximately $137\,\text{fb}^{-1}$. However, the dataset is only expected to double by the end of Run 3, which is planned to start in 2022 and continue until the end of 2024. On the one hand, the sensitivity of searches for beyond-the-Standard-Model effects will only slightly improve given the relatively small increase in the statistical precision. On the other hand, new channels might finally become accessible for both SM measurements and searches for new physics. In both scenarios, innovative analysis strategies are crucial to improve beyond statistical precision only.

To this end, the final states considered in this analysis are found to provide improved sensitivity for the investigation of highly energetic events. Together with the development and application of novel ML algorithms based on the jet substructure, these channels are a powerful tool for future searches for new physics, as well as SM measurements in boosted topologies.

Conclusions

The Standard Model (SM) of particle physics has been remarkably successful in explaining the multitude of experimental measurements in the last decades to very high precision. Nevertheless, several observed phenomena, e.g. gravitational force, masses of the neutrinos and dark matter, are not yet coherently included nor explained in its formulation. Therefore, an extended description of nature is needed, especially at higher energies.

Particularly puzzling, and strongly interconnected to the vacuum and its properties, is the Higgs sector. Given its extensive and varied phenomenology, the Higgs boson is an excellent tool in the quest for new physics and will constitute a crucial component of the high-energy physics program at the LHC in the following years.

Various extensions of the SM, predicting hypothetical new particles coupling to the Higgs boson, have been proposed in the last decades. In this thesis, a search for the resonant production of a hypothetical spin-1 massive particle decaying into a Z and a Higgs boson was presented. The analysis was performed using the pp collision data recorded in the years 2016 to 2018 by the CMS detector at the LHC at $\sqrt{s} = 13\,\text{TeV}$, corresponding to an integrated luminosity of about $137\,\text{fb}^{-1}$. The final states with two electrons or muons arising from the Z boson decay, providing a clean experimental signature for the background suppression, were analysed. The combination of these decay modes with the channel involving the Z boson decay into a pair of neutrinos was performed within the context of this thesis.

Moreover, the search targeted for the first time the 4-prong ($\mathrm{H} \to \mathrm{qqqq}$) and c flavour ($\mathrm{H} \to \mathrm{c\bar{c}}$) decays of the Higgs boson. The increase of the dataset size available for physics analysis and the more advanced analysis techniques based on ML have only recently made these decay modes accessible. The event selection of this search was optimised to ensure that the selected data are independent from those used in other analyses, particularly the one targeting the $\mathrm{H} \to \mathrm{b\bar{b}}$ final state, for a future statistical combination of the results. The sensitivity of the $\mathrm{H} \to \mathrm{b\bar{b}}$ channel is highest at low values of the resonance mass. On the other hand, the light-flavoured hadronic channel holds higher sensitivity for high resonance masses exceeding that of the $\mathrm{H} \to \mathrm{b\bar{b}}$ channel despite its much larger branching fraction. Moreover, the results obtained in this thesis prove that a dedicated analysis strategy involving the c-flavoured final states increases the sensitivity at high masses by a factor of 2 compared to existing results.

In conclusion, the final states considered in this analysis demonstrate how novel ML algorithms based on jet substructure can improve the sensitivity of searches at the LHC, making these new and powerful channels very promising for future searches for new physics and SM measurements.

In the coming years, the size of the dataset available for analysis will steadily increase, culminating with the High-Luminosity LHC in a sample one order of magnitude larger than what has been recorded so far. This enormous amount of data necessitates the development of more sophisticated analyses techniques to improve beyond statistical precision only for the current suite of analyses and, at the same time, extend the reach towards unexplored final states. For example, the recent development in ML algorithms for jet identification with substructure and flavour composition are of vital importance in probing the Yukawa coupling to second-generation fermions, and in particular to the c quarks. The measurement of the Higgs boson properties is one of the highest priorities of the LHC program. Direct measurements of several properties of the Higgs boson, like its self-coupling and width, are still beyond reach. Nonetheless, differential production cross section measurements and indirect measurements are already accessible, which can have profound theoretical implications on the actual structure of the Higgs potential. In this regard, the analysis techniques presented in this thesis constitute a baseline for probing the boosted regime in SM measurements involving the Higgs boson, where a precise characterisation of its production cross section in the high-p_{T} regime is crucial to unravel BSM effects.

Precise jet calibration is essential for the success of the LHC physics program and becomes even more crucial with the increased integrated luminosity. The method used in the CMS Collaboration in Run 2 for the measurement and calibration of the jet transverse momentum resolution was discussed comprehensively. The technique exploits the momentum conservation in the transverse plane in QCD dijet events. Two complementary methods are used to cover a wide range in pseudorapidity up to $|\eta| = 5.2$. A high level of precision is reached thanks to the thorough statistical treatment of the systematic uncertainties leading to an improvement in the calibration precision of approximately a factor of 3. These results, obtained with the dijet topology in a wide range in jet transverse momentum ranging from 100 GeV to 1 TeV, are combined for the first time with those derived in the Z +jet topology, allowing the extension down to transverse momenta of 40 GeV. The imminent Run 3 at the LHC paves the way to an era of unprecedented precision. The knowledge acquired during the previous data-taking periods, both in the understanding of the detector and in the evolution of analysis strategies, will allow for a calibration accuracy below the per mille level. As a consequence, all CMS physics analyses will profit from the substantial reduction of uncertainties.

Bibliography

[1] CMS Collaboration, "Jet energy scale and resolution performance with 13 TeV data collected by CMS in 2016-2018", Technical Report CMS-DP-2020-019, CERN, Geneva, 2020.

[2] A. Malara, "CMS jet and missing transverse momentum performance at Run 2 and prospects for Run 3", 2020. 40th International Conference on High Energy Physics (ICHEP).

[3] CMS Collaboration, "Reconstruction of jets and missing transverse momentum at the CMS experiment: Run 2 and perspective for Run 3", *PoS* **ICHEP2020** (2021) p. 752, `arXiv:2012.06271`.

[4] A. Malara, "Jet reconstruction and calibration for LHC Runs 2 and 3 in CMS", 2021. 13th International Workshop on Boosted Object Phenomenology, Reconstruction and Searches in HEP (BOOST).

[5] T. Sokolinsky, "Search for heavy resonances in the $p_{\mathrm{T}}^{\mathrm{miss}}$ + jet final state with the CMS experiment", Master's thesis, U. Hamburg, Dept. Phys., 2021.

[6] C. Burgard, "Standard model of physics". `https://texample.net/tikz/examples/model-physics/`, last accessed: 03.09.2021.

[7] Particle Data Group, "Review of Particle Physics", *PTEP* **2020** (2020) p. 083C01.

[8] M. E. Peskin and D. V. Schroeder, "An introduction to quantum field theory". Westview, Boulder, CO, 1995.

[9] C. N. Yang and R. L. Mills, "Conservation of Isotopic Spin and Isotopic Gauge Invariance", *Phys. Rev.* **96** (1954) p. 191.

[10] L. Morel, Z. Yao, P. Cladé et al., "Determination of the fine-structure constant with an accuracy of 81 parts per trillion", *Nature* **588** (2020) p. 61.

[11] H. Burkhardt and B. Pietrzyk, "Recent BES measurements and the hadronic contribution to the QED vacuum polarization", *Phys. Rev. D* **84** (2011) p. 037502, `arXiv:1106.2991`.

[12] A. Deur, S. J. Brodsky, and G. F. de Téramond, "The QCD running coupling", *Progress in Particle and Nuclear Physics* **90** (2016) p. 1–74, `arXiv:1604.08082`.

[13] OPAL Collaboration, "Measurement of the strong coupling constant alpha(s) and the vector and axial vector spectral functions in hadronic tau decays", *Eur. Phys. J. C* **7** (1999) p. 571, `arXiv:hep-ex/9808019`.

[14] J. R. Yablon, "QCD Theory of the Hadrons and Filling the Yang–Mills Mass Gap", *Symmetry* **12** (2020) p. 1887.

[15] CMS Collaboration, "Determination of the strong coupling constant $\alpha_S(m_\mathrm{Z})$ from measurements of inclusive $\mathrm{W}^\pm$ and Z boson production cross sections in proton-proton collisions at $\sqrt{\mathrm{s}} = 7$ and 8 TeV", *JHEP* **06** (2020) p. 018, `arXiv:1912.04387`.

[16] D. J. Gross and F. Wilczek, "Ultraviolet Behavior of Non-Abelian Gauge Theories", *Phys. Rev. Lett.* **30** (1973) p. 1343.

[17] H. D. Politzer, "Reliable Perturbative Results for Strong Interactions?", *Phys. Rev. Lett.* **30** (1973) p. 1346.

[18] S. L. Glashow, "Partial-symmetries of weak interactions", *Nuclear Physics* **22** (1961) p. 579 .

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

[20] A. Salam, "Weak and Electromagnetic Interactions", *Conf. Proc.* **C680519** (1968) p. 367.

[21] UA1 Collaboration, "Experimental observation of isolated large transverse energy electrons with associated missing energy at s = 540 GeV", *Physics Letters B* **122** (1983) p. 103 .

[22] UA2 Collaboration, "Observation of single isolated electrons of high transverse momentum in events with missing transverse energy at the CERN pp collider", *Physics Letters B* **122** (1983) p. 476 .

[23] Gargamelle Collaboration, "Search for elastic muon-neutrino electron scattering", *Physics Letters B* **46** (1973) p. 121 .

[24] Gargamelle Collaboration, "Observation of neutrino-like interactions without muon or electron in the gargamelle neutrino experiment", *Physics Letters B* **46** (1973) p. 138 .

[25] The LEP Electroweak Working Group, "Combination of measurements of the ALEPH, DELPHI, L3 and OPAL experiments on electroweak observables". `http://lepewwg.web.cern.ch/LEPEWWG`, last accessed: 03.09.2021.

[26] V. Novikov, L. B. Okun, A. N. Rozanov et al., "Theory of Z boson decays", *Rept. Prog. Phys.* **62** (1999) p. 1275, `arXiv:hep-ph/9906465`.

[27] ALEPH, DELPHI, L3, OPAL, SLD, LEP Electroweak Working Group, SLD Electroweak Group and SLD Heavy Flavour Group collaborations, "Precision electroweak measurements on the Z resonance", *Phys. Rept.* **427** (2006) p. 257, `arXiv:hep-ex/0509008`.

[28] ALEPH, DELPHI, L3, OPAL and LEP Electroweak collaborations, "Electroweak Measurements in Electron-Positron Collisions at W-Boson-Pair Energies at LEP", *Phys. Rept.* **532** (2013) p. 119, `arXiv:1302.3415`.

[29] P. Pétroff, "W mass and Triple Gauge Couplings at Tevatron", *EPJ Web of Conferences* **49** (2013) p. 14002.

[30] A. Kupco, "Triple and quartic gauge boson couplings at the LHC", *Nuovo Cim. C* **40** (2018) p. 202.

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

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

[33] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, "Global Conservation Laws and Massless Particles", *Phys. Rev. Lett.* **13** (1964) p. 585.

[34] P. W. Higgs, "Spontaneous Symmetry Breakdown without Massless Bosons", *Phys. Rev.* **145** (1966) p. 1156.

[35] C. Delaunay, C. Grojean, and J. D. Wells, "Dynamics of Non-renormalizable Electroweak Symmetry Breaking", *JHEP* **04** (2008) p. 029, `arXiv:0711.2511`.

[36] J. Ellis, "Higgs Physics", in *2013 European School of High-Energy Physics*, pp. 117–168. 2015. `arXiv:1312.5672`.

[37] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Phys. Lett. B* **716** (2012) p. 1, `arXiv:1207.7214`.

[38] CMS Collaboration, "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC", *Phys. Lett. B* **716** (2012) p. 30, `arXiv:1207.7235`.

[39] T. Matsuoka, "The CKM matrix and its origin", *Prog. Theor. Phys.* **100** (1998) p. 107, `arXiv:hep-ph/9804329`.

[40] Super-Kamiokande Collaboration, "Evidence for oscillation of atmospheric neutrinos", *Phys. Rev. Lett.* **81** (1998) p. 1562, `arXiv:hep-ex/9807003`.

[41] R. N. Mohapatra and G. Senjanovic, "Neutrino Mass and Spontaneous Parity Nonconservation", *Phys. Rev. Lett.* **44** (1980) p. 912.

[42] LHC Higgs Cross Section Working Group, “Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector”, *CERN Yellow Reports* **Vol.2/2017** (2016) `arXiv:1610.07922`.

[43] R. N. Cahn, “The Higgs boson”, *Reports on Progress in Physics* **52** (1989) p. 389.

[44] A. Denner, S. Dittmaier, and A. Mück, “PROPHECY4F 3.0: A Monte Carlo program for Higgs-boson decays into four-fermion final states in and beyond the Standard Model”, *Comput. Phys. Commun.* **254** (2020) p. 107336, `arXiv:1912.02010`.

[45] S. Dittmaier et al., “Handbook of LHC Higgs Cross Sections: 2. Differential Distributions”, *CERN Yellow Reports* (2012) `arXiv:1201.3084`.

[46] ATLAS Collaboration, “Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector”, *Phys. Lett. B* **784** (2018) p. 173, `arXiv:1806.00425`.

[47] CMS Collaboration, “Observation of $\mathrm{t\bar{t}H}$ production”, *Phys. Rev. Lett.* **120** (2018) p. 231801, `arXiv:1804.02610`.

[48] ATLAS Collaboration, “Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector”, *Phys. Lett. B* **786** (2018) p. 59, `arXiv:1808.08238`.

[49] CMS Collaboration, “Observation of Higgs boson decay to bottom quarks”, *Phys. Rev. Lett.* **121** (2018) p. 121801, `arXiv:1808.08242`.

[50] CLICdp and CLIC collaborations, “The Compact Linear Collider (CLIC) - 2018 Summary Report”, *CERN Yellow Reports* **2/2018** (2018) `arXiv:1812.06018`.

[51] A. Abada et al., “FCC-ee: The Lepton Collider: Future Circular Collider Conceptual Design Report Volume 2”, *Eur. Phys. J. ST* **228** (2019) p. 261.

[52] T. Behnke, J. E. Brau, B. Foster et al., “The International Linear Collider Technical Design Report - Volume 1: Executive Summary”, *CERN Yellow Reports* (2013) `arXiv:1306.6327`.

[53] CEPC Study Group, “CEPC Conceptual Design Report: Volume 1 - Accelerator”, *CERN Yellow Reports* (2018) `arXiv:1809.00285`.

[54] The LEP Working Group for Higgs Boson Searches, “Search for the Standard Model Higgs boson at LEP”, *Physics Letters B* **565** (2003) p. 61 .

[55] Tevatron New Physics Higgs Working Group, “Updated Combination of CDF and D0 Searches for Standard Model Higgs Boson Production with up to $10.0\,\mathrm{fb}^{-1}$ of Data”, *Phys. Rev. Lett.* (2012).

[56] ATLAS and CMS collaborations, “Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and $8\,\mathrm{TeV}$”, *JHEP* **08** (2016) p. 045, `arXiv:1606.02266`.

[57] ATLAS Collaboration, "Cross-section measurements of the Higgs boson decaying into a pair of τ-leptons in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$ with the ATLAS detector", *Phys. Rev. D* **99** (2019) p. 072001, arXiv:1811.08856.

[58] CMS Collaboration, "Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector", *Phys. Lett. B* **779** (2018) p. 283, arXiv:1708.00373.

[59] ATLAS Collaboration, "Search for the Decay of the Higgs Boson to Charm Quarks with the ATLAS Experiment", *Phys. Rev. Lett.* **120** (2018) p. 211802, arXiv:1802.04329.

[60] CMS Collaboration, "A search for the standard model Higgs boson decaying to charm quarks", *JHEP* **03** (2020) p. 131, arXiv:1912.01662.

[61] ATLAS Collaboration, "A search for the dimuon decay of the Standard Model Higgs boson with the ATLAS detector", *Phys. Lett. B* **812** (2021) p. 135980, arXiv:2007.07830.

[62] CMS Collaboration, "Evidence for Higgs boson decay to a pair of muons", *JHEP* **01** (2021) p. 148, arXiv:2009.04363.

[63] CMS Collaboration, "Measurement of the inclusive and differential Higgs boson production cross sections in the leptonic WW decay mode at $\sqrt{s} = 13\,\text{TeV}$", *JHEP* **03** (2021) p. 003, arXiv:2007.01984.

[64] ATLAS Collaboration, "Measurements of gluon-gluon fusion and vector-boson fusion Higgs boson production cross-sections in the $H \to WW^* \to e\nu\mu\nu$ decay channel in pp collisions at $\sqrt{s} = 13\,\text{TeV}$ with the ATLAS detector", *Phys. Lett. B* **789** (2019) p. 508, arXiv:1808.09054.

[65] ATLAS Collaboration, "Measurements of Higgs boson properties in the diphoton decay channel with $36\,\text{fb}^{-1}$ of pp collision data at $\sqrt{s} = 13\,\text{TeV}$ with the ATLAS detector", *Phys. Rev. D* **98** (2018) p. 052005, arXiv:1802.04146.

[66] CMS Collaboration, "Measurement and interpretation of differential cross sections for Higgs boson production at $\sqrt{s} = 13\,\text{TeV}$", *Phys. Lett. B* **792** (2019) p. 369, arXiv:1812.06504.

[67] A. Dainese, M. Mangano, A. B. Meyer et al., "Report from Working Group 2: Higgs Physics at the HL-LHC and HE-LHC", *CERN Yellow Rep. Monogr.* **7** (2019) p. 221, arXiv:1902.00134.

[68] G. Hernández-Tomé, J. I. Illana, M. Masip et al., "Effects of heavy Majorana neutrinos on lepton flavor violating processes", *Phys. Rev. D* **101** (2020) p. 075020, arXiv:1912.13327.

[69] G. Bertone and D. Hooper, "History of dark matter", *Rev. Mod. Phys.* **90** (2018) p. 045002, arXiv:1605.04909.

[70] A. Refregier, “Weak gravitational lensing by large scale structure”, *Ann. Rev. Astron. Astrophys.* **41** (2003) p. 645, `arXiv:astro-ph/0307212`.

[71] Planck Collaboration, “Planck 2018 results. VI. Cosmological parameters”, *Astron. Astrophys.* **641** (2020) p. A6, `arXiv:1807.06209`.

[72] BaBar Collaboration, “Measurement of an Excess of $\bar{B} \to D^{(*)}\tau^-\bar{\nu}_\tau$ Decays and Implications for Charged Higgs Bosons”, *Phys. Rev. D* **88** (2013) p. 072012, `arXiv:1303.0571`.

[73] LHCb Collaboration, “Test of Lepton Flavor Universality by the measurement of the $B^0 \to D^{*-}\tau^+\nu_\tau$ branching fraction using three-prong τ decays”, *Phys. Rev. D* **97** (2018) p. 072013, `arXiv:1711.02505`.

[74] Belle Collaboration, “Measurement of $\mathcal{R}(D)$ and $\mathcal{R}(D^*)$ with a semileptonic tagging method”, *Phys. Rev. Lett.* **124** (2020) p. 161803, `arXiv:1910.05864`.

[75] C. Abel et al., “Measurement of the Permanent Electric Dipole Moment of the Neutron”, *Phys. Rev. Lett.* **124** (2020) p. 081803.

[76] D. Croon, T. E. Gonzalo, L. Graf et al., “GUT Physics in the era of the LHC”, *Front. in Phys.* **7** (2019) p. 76, `arXiv:1903.04977`.

[77] D. Pappadopulo, A. Thamm, R. Torre et al., “Heavy Vector Triplets: Bridging Theory and Data”, *JHEP* **09** (2014) p. 060, `arXiv:1402.4431`.

[78] V. D. Barger, W. Keung, and E. Ma, “A Gauge Model With Light W and Z Bosons”, *Phys. Rev. D* **22** (1980) p. 727.

[79] B. Bellazzini, C. Csáki, and J. Serra, “Composite Higgses”, *Eur. Phys. J. C* **74** (2014) p. 2766, `arXiv:1401.2457`.

[80] R. Contino, D. Marzocca, D. Pappadopulo et al., “On the effect of resonances in composite Higgs phenomenology”, *JHEP* **10** (2011) p. 081, `arXiv:1109.1570`.

[81] D. Pappadopulo, A. Thamm, R. Torre et al., “Tools for the study of heavy vector triplets.”. `http://rtorre.web.cern.ch/rtorre/Riccardotorre/vector_triplet_t.html`, last accessed: 03.09.2021.

[82] L. Randall and R. Sundrum, “A Large mass hierarchy from a small extra dimension”, *Phys. Rev. Lett.* **83** (1999) p. 3370, `arXiv:hep-ph/9905221`.

[83] T. Gherghetta and A. Pomarol, “Bulk fields and supersymmetry in a slice of AdS”, *Nucl. Phys. B* **586** (2000) p. 141, `arXiv:hep-ph/0003129`.

[84] Gouzevitch et al., “Scale-invariant resonance tagging in multijet events and new physics in Higgs pair production”, *JHEP* **07** (2013) p. 148, `arXiv:1303.6636`.

[85] CMS Collaboration, "Search for heavy resonances and nonresonant axion-like particles in semileptonic ZZ, ZW, or ZH final states at $\sqrt{s} = 13\,\text{TeV}$", Technical Report CMS-PAS-B2G-20-013, CERN, Geneva, 2021.

[86] CMS Collaboration, "Search for a heavy vector resonance decaying to a Z boson and a Higgs boson in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$", *Eur. Phys. J. C* **81** (2021) p. 688, arXiv:2102.08198.

[87] CMS Collaboration, "Search for heavy resonances decaying to WW, WZ, or WH boson pairs in the lepton plus merged jet final state at $\sqrt{s} = 13\,\text{TeV}$", Technical Report CMS-PAS-B2G-19-002, CERN, Geneva, 2021.

[88] CMS Collaboration, "Search for weak vector boson and gluon-gluon fusion production of heavy resonances decaying to $\text{Z}(\nu\bar{\nu})\text{V}(\text{qq})$", Technical Report CMS-PAS-B2G-20-008, CERN, Geneva, 2021.

[89] CMS Collaboration, "A multi-dimensional search for new heavy resonances decaying to boosted WW, WZ, or ZZ boson pairs in the dijet final state at 13 TeV", *Eur. Phys. J. C* **80** (2020) p. 237, arXiv:1906.05977.

[90] CMS Collaboration, "Search for resonant Higgs boson pair production in four b quark final state using large-area jets in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$", Technical Report CMS-PAS-B2G-20-004, CERN, Geneva, 2021.

[91] CMS Collaboration, "Combination of searches for Higgs boson pair production in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$", *Phys. Rev. Lett.* **122** (2019) p. 121803, arXiv:1811.09689.

[92] CMS Collaboration, "Combination of CMS searches for heavy resonances decaying to pairs of bosons or leptons", *Phys. Lett. B* **798** (2019) p. 134952, arXiv:1906.00057.

[93] E. Lyndon and B. Philip, "LHC Machine", *Journal of Instrumentation* **3** (2008) p. S08001.

[94] E. Mobs, "The CERN accelerator complex - August 2018. Complexe des accélérateurs du CERN - Août 2018". https://cds.cern.ch/record/2636343, last accessed: 03.09.2021.

[95] CMS Collaboration, "Public Luminosity Information". https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults, last accessed: 03.09.2021.

[96] CMS Collaboration, "Pileup mitigation at CMS in 13 TeV data", *JINST* **15** (2020) p. P09018, arXiv:2003.00503.

[97] ATLAS Collaboration, "Study of multiple hard-scatter processes from different pp interactions in the same ATLAS event", Technical Report ATL-PHYS-PUB-2018-007, CERN, Geneva, 2018.

[98] CMS Collaboration, "The CMS experiment at the CERN LHC", *JINST* **3** (2008) p. S08004.

[99] T. Sakuma and T. McCauley, “Detector and Event Visualization with SketchUp at the CMS Experiment”, *J. Phys. Conf. Ser.* **513** (2014) p. 022032, `arXiv:1311.4942`.

[100] CMS Collaboration, “Description and performance of track and primary-vertex reconstruction with the CMS tracker”, *JINST* **9** (2014) p. P10009, `arXiv:1405.6569`.

[101] CMS Tracker Group, “The CMS Phase-1 Pixel Detector Upgrade”, (2020). `arXiv:2012.14304`. Submitted to JINST.

[102] C. Lippmann, “Particle identification”, *Nucl. Instrum. Meth. A* **666** (2012) p. 148, `arXiv:1101.3276`.

[103] CMS Collaboration, “The CMS electromagnetic calorimeter project: Technical Design Report”, *Technical Design Report CMS* (1997).

[104] CMS Collaboration, “The CMS ECAL data acquisition system and its performance at LHC Run 2”, *PoS* **LHCP2018** (2018) p. 069.

[105] CMS Collaboration, “The CMS hadron calorimeter project: Technical Design Report”, Technical Report CERN-LHCC-97-031 ; CMS-TDR-2, CERN, Geneva, 1997.

[106] CMS ECAL/HCAL Collaboration, “The CMS barrel calorimeter response to particle beams from 2-GeV/c to 350-GeV/c”, *Eur. Phys. J. C* **60** (2009) p. 359. [Erratum: Eur.Phys.J.C 61, 353–356 (2009)].

[107] CMS HCAL depth segmentation. `https://home.fnal.gov/~chlebana/CMS/Phase1/depthSegmentation.html`, last accessed: 28.01.2021.

[108] CMS Collaboration, “Precise Mapping of the Magnetic Field in the CMS Barrel Yoke using Cosmic Rays”, *JINST* **5** (2010) p. T03021, `arXiv:0910.5530`.

[109] CMS Collaboration, “Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13\,\mathrm{TeV}$”, *JINST* **13** (2018) p. P06015, `arXiv:1804.04528`.

[110] CMS Collaboration, “Measurement of inclusive W and Z boson production cross sections in pp collisions at $\sqrt{s} = 13\,\mathrm{TeV}$”, Technical Report CMS-PAS-SMP-15-004, CERN, Geneva, 2015.

[111] CMS Collaboration, “Measurement of CMS Luminosity”, Technical Report CMS-PAS-EWK-10-004, CERN, Geneva, 1900.

[112] CMS Collaboration, “CMS Luminosity Measurement for the 2015 Data Taking Period”, Technical Report CMS-PAS-LUM-15-001, CERN, Geneva, 2016.

[113] CMS Collaboration, “CMS Luminosity Measurements for the 2016 Data Taking Period”, Technical Report CMS-PAS-LUM-17-001, CERN, Geneva, 2017.

[114] CMS Collaboration, “CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13\,\text{TeV}$”, Technical Report CMS-PAS-LUM-17-004, CERN, Geneva, 2018.

[115] CMS Collaboration, “CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13\,\text{TeV}$”, Technical Report CMS-PAS-LUM-18-002, CERN, Geneva, 2019.

[116] CMS Collaboration, “Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 13\,\text{TeV}$”, *JHEP* **07** (2018) p. 161, `arXiv:1802.02613`.

[117] ATLAS Collaboration, “Measurement of the Inelastic Proton-Proton Cross Section at $\sqrt{s} = 13\,\text{TeV}$ with the ATLAS Detector at the LHC”, *Phys. Rev. Lett.* **117** (2016) p. 182002, `arXiv:1606.02625`.

[118] CMS Collaboration, “CMS. The TriDAS project. Technical design report, vol. 1: The trigger systems”, Technical Report CERN-LHCC-2000-038, CERN, Geneva, 2000.

[119] CMS Collaboration, “CMS: The TriDAS project. Technical design report, Vol. 2: Data acquisition and high-level trigger”, Technical Report CERN-LHCC-2002-026, CERN, Geneva, 2002.

[120] CMS Collaboration, “Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$”, *JINST* **15** (2020) p. P10017, `arXiv:2006.10165`.

[121] CMS Collaboration, “The CMS trigger system”, *JINST* **12** (2017) p. P01020, `arXiv:1609.02366`.

[122] CMS Collaboration, “CMS: The computing project. Technical design report”, Technical Report CERN-LHCC-2005-023, CERN, Geneva, 2005.

[123] T. Sjöstrand et al., “An introduction to PYTHIA 8.2”, *Comput. Phys. Commun.* **191** (2015) p. 159, `arXiv:1410.3012`.

[124] P. Nason, “A New method for combining NLO QCD with shower Monte Carlo algorithms”, *JHEP* **11** (2004) p. 040, `arXiv:hep-ph/0409146`.

[125] S. Frixione, P. Nason, and C. Oleari, “Matching NLO QCD computations with Parton Shower simulations: the POWHEG method”, *JHEP* **11** (2007) p. 070, `arXiv:0709.2092`.

[126] S. Alioli, P. Nason, C. Oleari et al., “A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX”, *JHEP* **06** (2010) p. 043, `arXiv:1002.2581`.

[127] M. Bahr et al., “Herwig++ Physics and Manual”, *Eur. Phys. J. C* **58** (2008) p. 639, `arXiv:0803.0883`.

[128] J. Alwall et al., “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations”, *JHEP* **07** (2014) p. 079, `arXiv:1405.0301`.

[129] J. Allison et al., “Geant4 developments and applications”, *IEEE Transactions on Nuclear Science* **53** (2006) p. 270.

[130] S. Agostinelli et al., “Geant4—a simulation toolkit”, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **506** (2003) p. 250.

[131] J. Allison et al., “Recent developments in Geant4”, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **835** (2016) p. 186.

[132] CMS Collaboration, “A further reduction in CMS event data for analysis: the NANOAOD format”, *EPJ Web Conf.* **214** (2019) p. 06021.

[133] CMS Collaboration, “Mini-AOD: A New Analysis Data Format for CMS”, *J. Phys. Conf. Ser.* **664** (2015) p. 7, `arXiv:1702.04685`.

[134] CMS Collaboration, “Particle-flow reconstruction and global event description with the CMS detector”, *JINST* **12** (2017) p. P10003, `arXiv:1706.04965`.

[135] CMS Collaboration, “Description and performance of track and primary-vertex reconstruction with the CMS tracker”, *JINST* **9** (2014) p. P10009, `arXiv:1405.6569`.

[136] R. Fruhwirth, “Application of Kalman filtering to track and vertex fitting”, *Nucl. Instrum. Meth. A* **262** (1987) p. 444.

[137] F. Pantaleo, “New Track Seeding Techniques for the CMS Experiment”. PhD thesis, U. Hamburg, Dept. Phys., 2017.

[138] CMS Collaboration, “2017 tracking performance plots”, Technical Report CMS-DP-2017-015, CERN, Geneva, 2017.

[139] K. Rose, “Deterministic annealing for clustering, compression, classification, regression, and related optimization problems”, *IEEE Proc.* **86** (1998) p. 2210.

[140] CMS Collaboration, “Adaptive Vertex Reconstruction”, Technical Report CMS-NOTE-2008-033, CERN, Geneva, 2008.

[141] CMS Collaboration, “Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13\,\mathrm{TeV}$”, *JINST* **13** (2018) p. P06015, `arXiv:1804.04528`.

[142] CMS Collaboration, “Performance of CMS muon reconstruction in cosmic-ray events”, *Journal of Instrumentation* **5** (2010) p. T03022–T03022.

[143] CMS Collaboration, “Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC”, *JINST* **16** (2021) p. P05014, `arXiv:2012.06888`.

[144] CMS Collaboration, “Reconstruction of Electrons with the Gaussian-Sum Filter in the CMS Tracker at the LHC”, Technical Report CMS-NOTE-2005-001, CERN, Geneva, 2005.

[145] A. Hocker et al., “TMVA - Toolkit for Multivariate Data Analysis”, Technical Report CERN-OPEN-2007-007, CERN, Geneva, 2007.

[146] CMS Collaboration, “Performance of reconstruction and identification of τ leptons decaying to hadrons and ν_τ in pp collisions at $\sqrt{s} = 13\,\text{TeV}$”, *JINST* **13** (2018) p. P10005, `arXiv:1809.02816`.

[147] CTEQ Collaboration, “Handbook of perturbative QCD: Version 1.0”, *Rev. Mod. Phys.* **67** (1995) p. 157.

[148] CMS Collaboration, “Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV”, *JINST* **12** (2017) p. P02014, `arXiv:1607.03663`.

[149] M. Cacciari, G. P. Salam, and G. Soyez, “The anti-k_t jet clustering algorithm”, *JHEP* **04** (2008) p. 063, `arXiv:0802.1189`.

[150] M. Cacciari, G. P. Salam, and G. Soyez, “FastJet user manual”, *Eur. Phys. J. C* **72** (2012) p. 1896, `arXiv:1111.6097`.

[151] S. D. Ellis and D. E. Soper, “Successive combination jet algorithm for hadron collisions”, *Phys. Rev.* **D48** (1993) p. 3160, `arXiv:hep-ph/9305266`.

[152] M. Wobisch and T. Wengler, “Hadronization corrections to jet cross-sections in deep inelastic scattering”, in *Monte Carlo generators for HERA physics. Proceedings, Workshop, Hamburg, Germany, 1998-1999*, pp. 270–279. 1998. `arXiv:hep-ph/9907280`.

[153] CMS Collaboration, “Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV”, *JINST* **13** (2018) p. P05011, `arXiv:1712.07158`.

[154] E. Bols, J. Kieseler, M. Verzetti et al., “Jet Flavour Classification Using DeepJet”, *JINST* **15** (2020) p. P12012, `arXiv:2008.10519`.

[155] R. Kogler et al., “Jet Substructure at the Large Hadron Collider: Experimental Review”, *Rev. Mod. Phys.* **91** (2019) p. 045003, `arXiv:1803.06991`.

[156] A. J. Larkoski, S. Marzani, G. Soyez et al., “Soft Drop”, *JHEP* **05** (2014) p. 146, `arXiv:1402.2657`.

[157] A. J. Larkoski, G. P. Salam, and J. Thaler, “Energy Correlation Functions for Jet Substructure”, *JHEP* **06** (2013) p. 108, `arXiv:1305.0007`.

[158] J. Thaler and K. Van Tilburg, “Identifying Boosted Objects with N-subjettiness”, *JHEP* **03** (2011) p. 015, `arXiv:1011.2268`.

[159] CMS Collaboration, “Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques”, *JINST* **15** (2020) p. P06005, `arXiv:2004.08262`.

[160] D. Bertolini, P. Harris, M. Low et al., “Pileup Per Particle Identification”, *JHEP* **10** (2014) p. 059, `arXiv:1407.6013`.

[161] CMS Collaboration, “Pileup-per-particle identification: optimisation for Run 2 Legacy and beyond”, Technical Report CMS-DP-2021-001, CERN, Geneva, 2021.

[162] CMS Collaboration, “Performance of the pile up jet identification in CMS for Run 2”, Technical Report CMS-DP-2020-020, CERN, Geneva, 2020.

[163] CMS Collaboration, “Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13\,\mathrm{TeV}$ using the CMS detector”, *JINST* **14** (2019) p. P07004, `arXiv:1903.06078`.

[164] CMS Collaboration, “Mitigation of anomalous missing transverse momentum measurements in data collected by CMS at $\sqrt{s} = 13\,\mathrm{TeV}$ during the LHC Run 2”, Technical Report CMS-DP-2020-018, CERN, Geneva, 2020.

[165] J. Butterworth et al., “PDF4LHC recommendations for LHC Run II”, *J. Phys. G* **43** (2016) p. 023001, `arXiv:1510.03865`.

[166] J. C. Collins and D. E. Soper, “The Theorems of Perturbative QCD”, *Ann. Rev. Nucl. Part. Sci.* **37** (1987) p. 383.

[167] J. C. Collins, D. E. Soper, and G. F. Sterman, “Factorization of Hard Processes in QCD”, *Adv. Ser. Direct. High Energy Phys.* **5** (1989) p. 1, `arXiv:hep-ph/0409313`.

[168] J. C. Collins, D. E. Soper, and G. F. Sterman, “Factorization for Short Distance Hadron - Hadron Scattering”, *Nucl. Phys. B* **261** (1985) p. 104.

[169] P. Baldi, P. Sadowski, and D. Whiteson, “Searching for Exotic Particles in High-Energy Physics with Deep Learning”, *Nature Commun.* **5** (2014) p. 4308, `arXiv:1402.4735`.

[170] P. Baldi, P. Sadowski, and D. Whiteson, “Enhanced Higgs Boson to $\tau^+\tau^-$ Search with Deep Learning”, *Phys. Rev. Lett.* **114** (2015) p. 111801, `arXiv:1410.3469`.

[171] P. Baldi, K. Bauer, C. Eng et al., “Jet Substructure Classification in High-Energy Physics with Deep Neural Networks”, *Phys. Rev. D* **93** (2016) p. 094034, `arXiv:1603.09349`.

[172] J. S. Conway, R. Bhaskar, R. D. Erbacher et al., “Identification of High-Momentum Top Quarks, Higgs Bosons, and W and Z Bosons Using Boosted Event Shapes”, *Phys. Rev. D* **94** (2016) p. 094027, `arXiv:1606.06859`.

[173] I. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning”. MIT Press, 2016. `http://www.deeplearningbook.org`.

[174] CMS Collaboration, “New Developments for Jet Substructure Reconstruction in CMS”, Technical Report CMS-DP-2017-027, CERN, Geneva, 2017.

[175] Y. Wang, Y. Sun, Z. Liu et al., “Dynamic Graph CNN for Learning on Point Clouds”, (2019). `arXiv:1801.07829`. Computer Vision and Pattern Recognition.

[176] H. Qu and L. Gouskos, “ParticleNet: Jet Tagging via Particle Clouds”, *Phys. Rev. D* **101** (2020) p. 056019, `arXiv:1902.08570`.

[177] J. Dolen, P. Harris, S. Marzani et al., “Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure”, *JHEP* **05** (2016) p. 156, `arXiv:1603.00027`.

[178] CMS Collaboration, “Search for low mass vector resonances decaying into quark-antiquark pairs in proton-proton collisions at $\sqrt{s} = 13\,\mathrm{TeV}$”, *JHEP* **01** (2018) p. 097, `arXiv:1710.00159`.

[179] G. Louppe, M. Kagan, and K. Cranmer, “Learning to Pivot with Adversarial Networks”, *NIPS’2017* (2016) `arXiv:1611.01046`.

[180] CMS Collaboration, “Identification of highly Lorentz-boosted heavy particles using graph neural networks and new mass decorrelation techniques”, Technical Report CMS-DP-2020-002, CERN, Geneva, 2020.

[181] CMS Collaboration, “Measurements of Inclusive W and Z Cross Sections in pp Collisions at $\sqrt{s} = 7\,\mathrm{TeV}$”, *JHEP* **01** (2011) p. 080, `arXiv:1012.2466`.

[182] CMS Collaboration, “W and top tagging scale factors for Run 2 data”, Technical Report CMS-DP-2020-025, CERN, Geneva, 2020.

[183] CMS Collaboration, “A search for the standard model Higgs boson decaying to charm quarks”, *JHEP* **03** (2020) p. 131, `arXiv:1912.01662`.

[184] L. Gouskos, “Private communication”, 2021-01-29.

[185] CMS Collaboration, “Mass regression of highly-boosted jets using graph neural networks”, Technical Report CMS-DP-2021-017, CERN, Geneva, 2021.

[186] CMS Collaboration, “Search for New Physics with Jets and Missing Transverse Momentum in pp collisions at $\sqrt{s} = 7\,\mathrm{TeV}$”, *JHEP* **08** (2011) p. 155, `arXiv:1106.4503`.

[187] ATLAS Collaboration, “Jet reconstruction and performance using particle flow with the ATLAS Detector”, *Eur. Phys. J. C* **77** (2017) p. 466, `arXiv:1703.10485`.

[188] ATLAS Collaboration, “Jet energy scale and resolution measured in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector”, *Eur. Phys. J. C* **81** (2021) p. 689, `arXiv:2007.02645`.

[189] D0 Collaboration, “High-p_T jets in $\bar{p}p$ collisions at $\sqrt{s} = 630$ GeV and 1800 GeV”, *Phys. Rev. D* **64** (2001) p. 032003, `arXiv:hep-ex/0012046`.

[190] D0 Collaboration, “Measurement of the inclusive jet cross section in $p\bar{p}$ collisions at $\sqrt{s} = 1.96\,\mathrm{TeV}$”, *Phys. Rev. D* **85** (2012) p. 052006, `arXiv:1110.3771`.

[191] ATLAS Collaboration, “Jet energy resolution in proton-proton collisions at $\sqrt{s} = 7\,\mathrm{TeV}$ recorded in 2010 with the ATLAS detector”, *Eur. Phys. J. C* **73** (2013) p. 2306, `arXiv:1210.6210`.

[192] CMS Collaboration, “CMS Physics: Technical Design Report Volume 1: Detector Performance and Software”, Technical Report CERN-LHCC-2006-001, CMS-TDR-8-1, CERN, Geneva, 2006.

[193] CMS Collaboration, “Event generator tunes obtained from underlying event and multiparton scattering measurements”, *Eur. Phys. J. C* **76** (2016) p. 155, `arXiv:1512.00815`.

[194] CMS Collaboration, “Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements”, *Eur. Phys. J. C* **80** (2020) p. 4, `arXiv:1903.12179`.

[195] F. Cavallari and C. Rovelli, “Calibration and Performance of the CMS Electromagnetic Calorimeter in LHC Run2”, *EPJ Web Conf.* **245** (2020) p. 02027.

[196] K. Goebel, “Probing supersymmetry based on precise jet measurements at the CMS experiment”. PhD thesis, U. Hamburg, Dept. Phys., 2015.

[197] R. Barlow, “Asymmetric errors”, *eConf* **C030908** (2003) p. WEMT002, `arXiv:physics/0401042`.

[198] J. M. Lindert et al., “Precise predictions for V + jets dark matter backgrounds”, *The European Physical Journal C* **77** (2017).

[199] A. Albert, “Private communication”, 2020-06-10.

[200] CMS Collaboration, “Investigations of the impact of the parton shower tuning in Pythia 8 in the modelling of $\mathrm{t\bar{t}}$ at $\sqrt{s} = 8$ and 13 TeV”, Technical Report CMS-PAS-TOP-16-021, CERN, Geneva, 2016.

[201] NNPDF Collaboration, “Parton distributions from high-precision collider data”, *Eur. Phys. J. C* **77** (2017) p. 663, `arXiv:1706.00428`.

[202] G. Cowan, K. Cranmer, E. Gross et al., “Asymptotic formulae for likelihood-based tests of new physics”, *The European Physical Journal C* **71** (2011).

[203] A. M. Mood, F. A. Graybill, and D. C. Boes, “Introduction to the theory of statistics”. McGraw-Hill, 1973.

[204] M. Oreglia, “A Study of the Reactions $\psi' \to \gamma\gamma\psi$”. PhD thesis, SLAC, 1980.

[205] T. Skwarnicki, “A study of the radiative CASCADE transitions between the Upsilon-Prime and Upsilon resonances”. PhD thesis, Cracow, INP and DESY, 1986.

[206] S. Das, “A simple alternative to the Crystal Ball function”, `arXiv:1603.08591`.

[207] A. L. Read, “Presentation of search results: The CL(s) technique”, *J. Phys. G* **28** (2002) p. 2693.

[208] ATLAS and CMS collaborations and LHC Higgs Combination Group, "Procedure for the LHC Higgs boson search combination in Summer 2011",.

[209] G. Cowan, K. Cranmer, E. Gross et al., "Asymptotic formulae for likelihood-based tests of new physics", *Eur. Phys. J. C* **71** (2011) p. 1554, `arXiv:1007.1727`. [Erratum: Eur.Phys.J.C 73, 2501 (2013)].

[210] CMS Higgs Physics Analysis Group, "Combine - a software tools used for statistical analysis". `http://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/`, last accessed: 03.09.2021.

[211] ATLAS Collaboration, "Search for heavy resonances decaying into a Z boson and a Higgs boson in final states with leptons and b-jets in $139\,\mathrm{fb}^{-1}$ of pp collisions at $\sqrt{s} = 13\,\mathrm{TeV}$ with the ATLAS detector", Technical Report ATLAS-CONF-2020-043, CERN, Geneva, 2020.

[212] ATLAS Collaboration, "Summary of ATLAS diboson searches". `https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2021-018/fig_01.png`, last accessed: 03.09.2021.

Eidesstattliche Versicherung / Declaration on oath

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

I hereby declare upon oath that I have written the present dissertation independently and have not used further resources and aids than those stated.

Ort, Datum | City, date

Unterschrift | Signature

Acknowledgements

This is the end of my PhD, a journey that most definitely had a great impact on my life and future. This is why I want to thank many people, without whom all of this would not have been possible.

First of all, I would like to thank my supervisor Johannes Haller for making it possible to obtain my PhD in Hamburg and for opening a whole new world of possibilities for me. Joining his group was one of the best things that could have happened to me.

I would also like to express my warmest thanks to my parents and my sister for their unceasing and unconditional support in this 2000 km far away adventure. A big thanks to the rest of my family for welcoming me every time with open arms like I had never left.

I would also like to express my gratitude to all the people who have merged their path with mine, in one way or the other.

To Roman, for setting high standards as a physicist. Thanks to all the encouragement you gave me and all the valuable comments over the last years.

To Anastasia, for introducing me to the jet calibration world. I could not see it at first, but it turned out to offer a wide range of possibilities.

To Matthias, for helping me in the final sprint. We did not spend much time together, but I hope we will meet again in the future.

To my friend Paolo, my role model for a good and happy physicist. A special thanks goes to you for showing support and believing in me every day.

To Robin, my British friend, who pushed me to improve my English and my coding skills. Thanks for unravelling the mystery of *it-eat-heat-hit*. I will keep our language discussions in my heart forever.

To Dennis and Karla, for our physics discussions and cheerful coffee breaks, for the dancing moments and the shared laughs, and for being great friends inside and outside the office.

To Anna, for believing in my abilities, for our chats, for slowly (and patiently) speaking German to me, and for being a worthy opponent in board games. I look forward to working with you for the next two years.

To Irene, Alex and Arne, for making our office the best on the right side of the corridor. It has been a pleasure to share my PhD with you.

To my sister, my alter ego who cannot help but bring me back to reality when I need it and who has kept me motivated more than she can imagine.

To my *other* sisters, Ylenia and Roberta, for our long phone calls and for your songs, for your sincere and unbounded friendship in life's up and down moments.

To my long-time friends, Francesca and Valentina, for the epic moments together. Although, I am still waiting for you to come to visit.

To Flavio, for showing me that there is always another way, and for teaching me that "Nella vita si diventa grandi *nonostante*!".

To my friends, Giulia and Veronica, for their presence, now a bit more *social*, despite the distance and years, for always being there for a laugh.

My deepest gratitude goes to Arne, to whom I dedicate this thesis. Thanks for reading my thesis, for pushing me to improve, for our infinite physics discussions, for our shared interests, for the gastronomical and alcoholic exchange, for the happy moments and for the tough ones, for your endless optimism, for our gym challenge, for our maniacal precision that led to the best plots ever, for our office-tip list, for the "you should pay attention!" with which everything has started. The list is still pretty long, so simply "Grazie di tutto".

Thanks to all of you for sharing a piece of your life with me!