Measurement of Higgs boson fiducial cross sections with the CMS detector and electromagnetic reconstruction with the high-granularity endcap calorimeter

Alessandro Tarabini

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAX101

Thèse de doctorat

ÉCOLE POLYTECHNIQUE

IP PARIS

Measurement of Higgs boson fiducial cross sections with the CMS detector and electromagnetic reconstruction with the high-granularity endcap calorimeter

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat : Physique des particules

Thèse présentée et soutenue à Palaiseau, le 24/11/2023, par

ALESSANDRO TARABINI

Composition du Jury :

Chiara Mariotti Directrice de recherche, INFN (Torino)	Président
Giovanni Marchiori Chargé de recherche, APC (Paris)	Rapporteur
Christos Anastopoulos Chercheur, University of Sheffield (Sheffield)	Rapporteur
Marie-Hélène Genest Directrice de recherche, LPSC (Grenoble)	Examinatrice
Massimiliano Grazzini Professeur, University of Zurich (Zurich)	Examinateur
Adinda de Wit Chargée de recherche, LLR (Palaiseau)	Examinatrice
Roberto Salerno Directeur de recherche, LLR (Palaiseau)	Directeur de thèse

"I am on the edge of mysteries and the veil is getting thinner and thinner"

Louis Pasteur

"The effort to understand the universe is one of the very few things which lifts human life a little above the level of farce and gives it some of the grace of tragedy"

Steven Weinberg

Abstract

This PhD thesis presents a comprehensive characterisation of the Higgs boson in the H $\rightarrow$ ZZ $\rightarrow$ 4ℓ decay channel, commonly known as the four-lepton channel, using proton-proton collisions at a centre-of-mass energy of 13 TeV recorded with the CMS experiment during the Run 2 of the CERN LHC. The four-lepton channel is regarded as the golden channel of Higgs physics due to its clear peak over an almost flat background, the large signal-to-background ratio, a fully reconstructible final state, and the synergy with the highly performing lepton reconstruction of the CMS detector. Cross section measurements are one of the best methods to study the Higgs boson properties, probing its couplings to other particles and testing theoretical predictions. In this thesis, cross sections are obtained by removing detector effects from data and measured in a fiducial phase space defined to match the experimental acceptance closely. This methodology ensures maximal model independence and reinterpretability for the results. Cross sections are measured both inclusively and in bins of 32 single- and double-differential observables, providing insights into the production and decay of the Higgs boson, the tensor structure of the vertex between the Higgs boson and the Z bosons, as well as QCD effects. The constraint from data of the non-resonant ZZ$\rightarrow$4ℓ background is also included. The set of results is completed by the constraints of the trilinear self-coupling of the Higgs boson and its couplings to bottom and charm quarks. All results are consistent with the theoretical predictions of the standard model of particle physics. Looking ahead, with the Run 2 concluded and Run 3 underway, this thesis introduces a novel method for estimating the reducible background of the four-lepton channel that will be used in forthcoming analyses. This method can also be extended to all channels with leptonic final states. The reducible background comprises non-prompt leptons and other particles misidentified as leptons, which can mimic the signal signature. The proposed strategy explores the possibility of modelling this source using a novel approach and reducing the sizeable systematic uncertainty typical of the current methods, which will be a limiting factor during the next phase of the LHC. The High-Luminosity LHC (HL-LHC) aims to increase the integrated luminosity by a factor of 10 beyond the LHC's design value, opening new horizons for discoveries and precision physics. In order to cope with the large number of simultaneous collisions per bunch crossing, known as pileup, and sustain the high radiation dose, the CMS experiment foresees the complete replacement of the endcap calorimeter. The new High-Granularity endcap CALorimeter (HGCAL) will be a silicon-based sampling calorimeter, offering the possibility of performing calorimetry with tracker-like granularity. This thesis contributes to the development and reassessment of the electron and photon offline reconstruction for the HGCAL. The first contribution regards cleaning electromagnetic showers from spurious contaminations resulting from the high-pileup environment, which degrades the properties of reconstructed electromagnetic objects. The second contribution focuses on electron reconstruction. An electron can start showering while traversing the inner tracker before reaching the calorimeter. This effect, combined with the 3.8 Tesla of the CMS magnet, leads the electron energy to be spread in several clusters around the primary one. These contributions should be clustered together to reconstruct the original electron. A purely geometrical algorithm currently performs this procedure, and its performance is assessed in the HGCAL for the first time. Additionally, this thesis proposes a new, dedicated algorithm based on Deep Neural Networks explicitly tailored for the new calorimeter.

Résumé

Cette thèse de doctorat présente une caractérisation complète du boson de Higgs dans le canal de désintégration H→ZZ→4ℓ, généralement appelé canal à quatre leptons, en utilisant des collisions proton-proton à une énergie de 13 TeV dans le centre de masse et collectées avec l'expérience CMS lors du Run 2 du LHC au CERN. Le canal à quatre leptons est considéré comme le canal optimal de la physique du boson de Higgs en raison de son pic clairement visible au-dessus d'un bruit de fond presque plat, de l'excellent rapport signal sur bruit, d'un état final entièrement reconstructible qui bénéficie en outre de la reconstruction très performante des leptons du détecteur CMS. Les mesures de section efficace sont l'une des meilleures méthodes pour étudier les propriétés du Boson de Higgs, en sondant les couplages du boson de Higgs et en testant les prédictions théoriques. Cette thèse présente les sections efficaces obtenues en éliminant les effets du détecteur et mesurées dans un espace de phase fiduciaire défini pour correspondre étroitement à l'acceptance expérimentale. Les sections efficaces sont mesurées dans des bins de 32 observables simples et doublement différentielles, fournissant des informations sur la production et la désintégration du boson de Higgs, la structure tensorielle du vertex entre le boson de Higgs et les bosons Z, et les effets liés à l'interaction forte. La contrainte issue des données du fond non résonant ZZ→4ℓ est également incluse. L'ensemble des résultats est complété par la contrainte de l'auto-couplage trilinéaire du boson de Higgs et des couplages aux quarks bottom et charm. L'ensemble des résultats est cohérent avec les prédictions théoriques du modèle standard de la physique des particules. En regardant vers l'avenir, alors que le Run 2 est terminé et que le Run 3 est en cours, cette thèse présente une nouvelle méthode pour estimer le bruit de fond réductible du canal des quatre leptons qui sera utilisée dans les analyses avec les nouvelles données. Le bruit de fond réductible comprend des leptons non prompts et d'autres particules mal identifiées comme des leptons, qui peuvent imiter la signature du signal. La stratégie proposée explore la possibilité de modéliser cette source en utilisant une approche nouvelle et de réduire l'incertitude systématique considérable typique des méthodes actuelles, ce qui sera également un facteur limitant lors de la prochaine phase du LHC. Le LHC à haute luminosité (HL-LHC) vise à augmenter la luminosité intégrée par un facteur de 10 par rapport à la valeur nominale du LHC, ouvrant de nouvelles perspectives pour les découvertes et la physique de précision. Pour faire face au grand nombre de collisions simultanées par croisement de paquets, appelé pileup, et supporter la forte dose de rayonnement, l'expérience CMS prévoit le remplacement complet du calorimètre dans les bouchons. Le nouveau calorimètre à haute granularité (HGCAL) offrira la possibilité de réaliser une calorimétrie avec une granularité similaire à celle du trajectographe. Cette thèse contribue au développement et à la réévaluation de la reconstruction électromagnétique hors ligne pour le HGCAL. La première contribution concerne le nettoyage des gerbes électromagnétiques des contaminations parasites résultant de l'environnement à grand pileup. La deuxième contribution concerne la reconstruction des électrons. Un électron peut commencer à créer une gerbe en traversant le trajectographe avant d'atteindre le calorimètre. Cet effet, combiné aux 3.8 Tesla de l'aimant de CMS, fait que l'énergie de l'électron est étalée en plusieurs amas. Ces contributions devraient être regroupées pour reconstruire l'électron tel que produit lors de la collision dure. Un algorithme géométrique effectue actuellement cette procédure, et sa performance est évaluée dans le HGCAL pour la première fois. De plus, cette thèse propose un nouvel algorithme dédié basé sur des Deep Neural Networks spécialement conçus pour le nouveau calorimètre.

Acknowledgments

To mum and my sister Odetta

At the end of this long journey, which seems to have gone by too quickly, I look back on my steps and feel obliged to give due thanks to some people.

In the first place, I wish to thank the referees, Christos Anastopoulos and Giovanni Marchiori for their meticulous review of the manuscript, as well as the other members of the jury, Adinda De Wit, Marie-Hélène Genest, Massimiliano Grazzini, and Chiara Mariotti.

My deepest gratitude goes to my supervisor, Roberto. Four years ago, he chose to believe in me and granted me a remarkable opportunity. His vision of physics has been inspirational, and his unwavering support has been invaluable. I owe an immeasurable debt of thanks to Roberto for allowing me to work with Matteo. I have found an exceptional mentor, an enthusiastic colleague, and, most importantly, a cherished friend. Without him, this thesis would have never come to end, and probably would not have even started. He has always been there to support me in my existential crisis and share laughter to cheer my days up.

I am particularly grateful to the entire CMS group at LLR, whose vibrant and collaborative atmosphere has been instrumental in the success of my PhD. I am particularly grateful to Claude; talking to him is always enlightening and delightful. I also extend heartfelt thanks to Florian and Shamik for the incredible adventures we shared in TICLand. Your guidance helped me navigate and stay on the right course.

Recognizing that behind every great outcome lies a collective effort, I extend my thanks to the entire CJLST group for their support in my work with the four-lepton channel, and to the TICL working group for their guidance in my work with the HGCAL reconstruction. I have found stimulating, friendly groups instrumental in delivering the results in this thesis.

During my time at LLR, I had the privilege of meeting exceptional individuals. An old girl who communicates in a language seemingly without vowels proved to be the most incredible office mate one could hope for. An amazing "Réginaburgiennes" who is a prodigy in her job and should start to be aware of that. The best PhD student in the world who privileged me with his friendship and taught me that a book should never be judged by its cover. A French-not-so-French guy who has a profound passion for music but a very questionable grasp of the concept of rhyme. An ardent protector of the pure Portuguese language that is a delightful conversationalist full of talent, albeit not in multitasking. A brilliant Indian gentleman with the uncommon quality for wonder. A girl that think of being born in a real city with a peculiar honey for clinical cases (maybe I am one of those). A young man desperately seeking visibility and a daily intake of calories near zero. With my loved ones to the other side of the Alps, you have been my family in Paris, cheering up my days inside and outside the office.

This Parisian adventure would have not been the same without my friend Agatino. I met you right after my arrival in Paris and you have been my mate since then. Thank you for always being there for a light-hearted chat, a weekend at the Louvre during a global pandemic, and for a walk under a snowstorm in an empty German village.

I must thank my *Amici di serie A+++*. Although we are scattered across the four corners of Europe, our connection remained unbroken throughout the years. Whenever we can see each other, it is like feeling at home for me. The physicist I am today is indebted to the studying sessions we shared in the *auletta*.

A mia zia Emilia, mio zio Luigi, e i miei cugini Ilenia e Andrea. Ci siete sempre stati e posso contare che ci sarete sampre. Grazie.

Un ringraziamento speciale a Ester, compagna di vita e avventure. Sono consapevole che certe volte, anche se sono molto poche, posso essere complicato. Per questo, ti ringrazio ancora di più

per essere sempre e costantemente al mio fianco. Le nostre lunghe chiamate sono state il sole che mancava nelle grigie giornate parigine. Se siamo anche resisiti anche ai 700 Km di distanza, forse questo dovrà pur significare qualcosa. Questo ringraziamento deve essere esteso anche alla sua famiglia che mi ha sempre fatto sentire a casa e mi ha sempre fatto sentire il loro affetto e supporto.

A conclusione di questi ringraziamenti, il grazie più grande a mia mamma Rita e mia sorella Odetta. Questa tesi è dedicata a voi perchè è soprattutto grazie a voi se ho raggiunto questo traguardo. Odi, sei stata molto di più di una semplice sorella. La persona che sono oggi lo devo in gran parte a te e ti ringrazio per il tuo sostegno, senza se e senza ma, a tutte le mie scelte. Mamma, da sola hai fatto l'impossibile. Nonostante alcune mie decisioni non ti siano state congeniali, hai capito che mi rendevano felice e ti ringrazio per non avermi mai fatto mancare il tuo supporto e conforto.

At the end of these acknowledgments, I find myself with tears in my eyes, recognising the immense fortune I have been granted. Grazie!

Alessandro

Contents

Introduction

> *"While it is never safe to affirm that the future of Physical Science has no marvels in store even more astonishing than those of the past, it seems probable that most of the grand underlying principles have been firmly established and that further advances are to be sought chiefly in the rigorous application of these principles to all the phenomena which come under our notice. It is here that the science of measurement shows its importance — where quantitative work is more to be desired than qualitative work. An eminent physicist remarked that the future truths of physical science are to be looked for in the sixth place of decimals."*

Albert Abraham Michelson pronounced these prophetic words in 1894 [1]. At the dawn of the *short century*, a part of the scientific community believed that Newtonian mechanics, thermodynamics, and the Maxwell theory provided an exhaustive description of the universe, and the unknown phenomena were expected to find an explanation within these theories. However, a few years later, during a Friday Evening Lecture at the Royal Institution in London on April 27, 1900 [2], Lord Kelvin perceived that two *clouds* were on the horizon of physics: the failure to account for the radiation spectrum emitted by a blackbody and the results of the Michelson-Morley experiment [3]. Lord Kelvin was probably unaware that the explanation of these two phenomena would lead to two profound revolutions in physics a few years later: the advent of quantum physics and the special theory of relativity. Their combination with classical field theory would lead to establishing a theoretical frame for particle physics, known as Quantum Field Theory (QFT).

The birth of particle physics can be traced back to the discovery of the electron by Sir Joseph John Thomson in 1897 [4]. This discovery established that the atom was not homogenous and indivisible but that, in fact, it is a more complex structure. This finding laid the foundation for the exploration of the fundamental constituents of matter. Throughout the 20th century and into the present days, QFT and particle physics have been intertwined, the former as the theoretical frame and the latter as the field of experimental research.

The first great success of this interplay between theory and experiments was represented by Quantum Electrodynamics, developed in the 1930s and 1940s to describe the interaction between photons and matter, providing a theoretical framework for the electromagnetic force. It represented the first successful QFT. This success was made possible by the development of the renormalisation procedure, which allowed for the treatment of infinities in calculations and rendered the theory predictive. Following this milestone, QFTs suffered a period of crisis, not being able to find a theoretical framework for the weak and strong interactions. However, according to Steven Weinberg [5], three brilliant ideas emerged in the 1950s and 1960s that, over the following years, allowed to bring QFTs back to the forefront and led to the formulation of the Standard Model (SM) of particle physics. The first idea was the quark model proposed in 1964 by Murray Gell-Mann and George Zweig [6, 7]. The idea that the large plethora of hadrons being discovered in those years were composed of more fundamental constituents - quarks and antiquarks - tidied

up amidst the chaotic *particle zoo*. The quark model paved the way to the development of Quantum ChromoDynamics (QCD), the QFT describing the strong interactions. The second concept was that local symmetries dictate the form of interactions [8], proposed by Chen Ning Yang and Robert Mills in 1954. This is the principle of gauge invariance, which stands as a pillar of the current SM. The third idea was the concept of spontaneously broken symmetries: symmetries of the SM Lagrangian that are not symmetries of the vacuum states. This concept was brought to particle physics by Yoichiro Nambu at the beginning of the 1960s. Many symmetries were known at the time, but they posed a puzzle because they were only approximate and affected some interactions but not others. The idea that these symmetries could be broken began to emerge, but adding explicit symmetry-breaking terms in the theory would violate renormalization. The solution could therefore lie in assuming that the symmetry was spontaneously broken. However, there was one problem with this idea, the Goldstone theorem: for every spontaneously broken symmetry there must be a massless spin-0 bosons; but there were no experimental signs of such particles. This impasse was overcome due to the work of Peter Higgs, François Englert, and Robert Brout in 1964 [9–11]. They understood not only that the Goldstone theorem does not apply in the case of gauge symmetries as opposed to global symmetries, but that spontaneous gauge symmetry breaking leads to the formation of massive bosons. All these theoretical successes had one problem: they were wrongly applied to the strong interactions. It was not until 1967 that Steven Weinberg and Abdus Salam [12, 13] applied the concept of the Brout-Englert-Higgs mechanism to the electromagnetic and weak interactions, leading to the development of the unified electroweak theory. This theory considers the electromagnetic and weak interactions as a manifestation of the same electroweak force. The spontaneous breaking of the gauge symmetry of this unified theory gives mass to the mediators of the weak force - the W and Z vector bosons - while preventing the mediator of the electromagnetic force - the photon - to be massive, and provides a description of the interactions of these particles. The electroweak symmetry breaking is made possible by a massive scalar field - the Higgs field - that should give raise to a massive scalar boson - the Higgs boson. The introduction of this new field also allows the inclusion of mass terms for fermions via Yukawa interactions [14]. Each fermion interacts with the Higgs field with a different strength (coupling), and the stronger the interaction, the larger the resulting mass for the particle.

These theoretical developments started one of the biggest experimental programmes in the annals of science. Finding and measuring the properties of the W, Z, and Higgs bosons took over 40 years and spanned four main projects all based at the CERN in Geneva: Gargamelle, the Super Proton Synchrotron, the Large Electron Positron collider (LEP), and the Large Hadron Collider (LHC). Starting from the discovery of weak neutral currents in 1973 [15], the electroweak theory of the SM managed to withstand any experimental evidence and was found to describe the weak and electromagnetic interactions adequately. An enormous success for a theory that Weinberg himself said to have *"too many arbitrary features for these predictions to be taken very seriously"* [5].

All the ingredients of the SM, including the ones predicted by the QCD sector, were discovered over the years and their characteristics found to be compatible with the SM predictions. After the observation of the top quark in 1995 at the Tevatron accelerator at the Chicago Fermilab [16, 17], only a single ingredient was still to be found, and arguably the most important to close the circle and prove the solidity of the SM: the Higgs boson. The challenge in finding the Higgs boson is that its mass is a free parameter of the theory, leaving experimentalists without any indication of where to search. An extract from a 1975 paper written by the theoreticians John Ellis, Mary Gaillard, and Dimitri Nanopoulos [18] clearly shows the sentiment at that time:

> *"We apologise to experimentalists for having no idea what is the mass of*

> *the Higgs boson [...] and for not being sure of its couplings to other particles, except that they are probably all very small. For these reasons we do not want to encourage big experimental searches for the Higgs boson [...].”*

However, the search for the Higgs boson became one of the primary goals of experimental particle physics, giving support to theoreticians who claim that experimentalists do not listen to their advice. The sole way to find the Higgs boson was basically to search in the largest phase space possible by imposing ever-increasing stringent limits to the search area. These searches were mainly carried out at the Tevatron and the LEP, unfortunately without finding any sign of the particle. At the start of the LHC operations, the limits were pointing to a relatively light Higgs boson in the mass range of 114-150 GeV. Finally, in 2012, nearly half a century after its postulation, a new particle was found by the ATLAS and CMS collaboration at a mass of approximately 125 GeV that resulted in being highly compatible with the long-sought Higgs boson [19, 20].

After this historical discovery, one of the primary goals of the LHC became to determine whether the newfound particle was indeed the SM Higgs boson, with no significant deviations from the SM predictions emerging thus far. Therefore, ten years after the discovery of the Higgs boson, sixty years after the theoretical formulation of the SM, and more than a century after the birth of particle physics, we stand at a juncture reminiscent of the late 19th century when Michelson pronounced his (wrong) prophetic words. All the ingredients of the known theory are now in place, all experimental pieces of evidence are in accordance with it, and we are moving in ever more precise measurements of physics quantities. However, like at the beginning of the 20th century, some clouds are on the horizon that could revolutionise physics again in the next years. Despite its tremendous success, the SM has a level of arbitrariness that leaves scientists somewhat dissatisfied, cannot incorporate in a unique theory the gravitational interaction, and some plausible explanations for cosmological evidence lie outside the grasp of the current theory. The physics community is now almost unanimously convinced that the SM, as we know it, is not the final theory. While Michelson foresaw a precision physics as an end in itself, it has taken on paramount importance in the present era for pushing the boundaries of our knowledge. Finding a discrepancy from the SM expectations would unavoidably indicate the direction for extending the SM for a more general description of the universe.

The work presented in this thesis is situated in this research context. The manuscript is structured into three parts. Part I serves as the foundation, providing essential contextualization for my research work. Chapter 1 introduces the theoretical framework of the SM, with particular emphasis on the electroweak theory - the sector under study in this thesis. It also underscores the importance of studying the Higgs boson in the quest for physics Beyond the SM (BSM), presenting and contextualising the current methodologies employed for this pursuit. Chapter 2 offers an introduction to the LHC and the CMS experiment. This overview includes the CMS offline reconstruction and the prospects for the forthcoming years.

Part II focuses on the characterisation of the Higgs boson properties. A meticulous characterisation of its properties would reveal whether the particle discovered in 2012 is indeed the Higgs boson responsible for the electroweak symmetry breaking as prescribed in the SM. Should this be the case, then the Higgs boson, coupling with all massive particles, could represent the bridge to the realm of BSM physics.

Since the lifetime of the Higgs boson is too short for direct detection, its identification relies on the study of its decay products. This thesis focuses on the decay into two Z bosons, which subsequently decay into four leptons ($\mathrm{H} \to \mathrm{ZZ} \to 4\ell$). The clear peak over an almost flat background, the large signal-to-background ratio, a fully reconstructible

final state, and the synergy with the highly performing lepton reconstruction of the CMS detector have put this channel to the forefront during the discovery, alongside the $H \to \gamma\gamma$ channel, and the subsequent characterisation. The discovery and the first characterisation was made possible by analysing the data of the Run 1 (2009-2013) of the CERN LHC. The study of the Higgs profile pursued later with the following data-taking period. Due to the increased centre-of-mass energy and luminosity, the Run 2 (2015-2018) of the LHC produced a larger amount of Higgs boson, essential for a precise, granular, extensive study of the Higgs sector. The second part of this manuscript provides a comprehensive characterisation of the $H \to ZZ \to 4\ell$ channel through measurement of cross sections exploiting the data collected during the Run 2. I was the main analyser and co-editor of the paper, steering the analysis from conception to publication in the Journal of High Energy Physics [21].

Chapter 3 presents the definition and validation of the physics objects needed to reconstruct the decay of the Higgs boson in four leptons. The objective of Ch. 4 is to use the physics objects defined in the preceding chapter to set up an analysis strategy to measure cross sections. The measurement targets results with enduring validity, maximal model independence, and an extensive characterisation of the decay channel. The first objective is attained by removing experimental effects from the data through an unfolding procedure. The second objective is achieved by measuring cross sections in a fiducial phase space, defined to match closely the experimental acceptance. Finally, the third objective is made possible by the differential measurement in 32 one- and two-dimensional observables, of which 28 implemented for the first time in this thesis. These observables provide insights into the production and decay of the Higgs boson, the tensor structure of the vertex between the Higgs boson and the Z bosons, and effects due to the strong interactions. The results are then presented in Ch. 5, where they are also interpreted to constrain the couplings of the Higgs boson with itself and the bottom and charm quarks.

This PhD thesis covers the years between 2020 and 2023, thus embracing the beginning of the Run 3 of the LHC (2022-2025). This new run is continuing the effort to provide an ever-increasing precise characterisation of the Higgs sector. At the end of the Run 3, the statistical precision is not expected to improve significantly compared to Run 2 (improvement of the order $\sqrt{2}$). Consequently, this is a suitable period to reassess, rethink, and improve the current analysis strategies to boost the precision from the systematic side. Chapter 6 is situated in this context. An arduous challenge in the $H \to ZZ \to 4\ell$ channel is modelling the background associated with non-prompt and misidentified leptons, which was estimated in Run 2 with an uncertainty equal to 40%. This chapter presents a novel method for estimating this source of background and explores the possibility of reducing the sizeable systematic uncertainty of the current method.

While Part II is about the past and the present of the CMS collaboration, Part III looks to the future of the detector. In order to increase the potential for discoveries and boost the reachable precision, the LHC will undergo a major upgrade to increase its luminosity by a factor of 10 beyond the LHC's design value. The High-Luminosity LHC (HL-LHC) operations will be starting in 2029. In order to cope with the large number of simultaneous collisions per bunch crossing, known as Pile Up (PU), and sustain the high radiation dose, the CMS experiment foresees the complete replacement of the endcap calorimeter. The new High Granularity CALorimeter (HGCAL) will be the first large-scale silicon-based calorimeter ever employed in experimental particle physics, opening a new era in calorimetry. Such a revolution on the hardware side must accompany an equal revolution on the reconstruction algorithms. The HGCAL offline reconstruction should be reassessed to cope with the harsh environment and the innovative technology of the new detector.

Chapter 6 presents how the reconstruction will be performed with the HGCAL and characterises the reconstruction of unconverted photons. Due to the high-PU environment, the reconstructed electromagnetic objects will be inevitably contaminated by spurious energy contributions, lowering the quality of the object. This chapter presents a PU reduction algorithm to improve the quality of the reconstructed object. These results are public in the CMS Detector Performance note in Ref. [22].

Chapter 7 focuses on a typical issue of electron reconstruction: the superclustering. An electron can start showering while traversing the inner tracker before reaching the calorimeter. This effect, combined with the 3.8 Tesla of the CMS magnet, leads the electron energy to be spread in several clusters around the primary one. These contributions should be clustered together to reconstruct the original electron. A purely geometrical algorithm currently performs this procedure, and its performance is assessed in the HGCAL for the first time. Additionally, the chapter presents a new, dedicated algorithm based on Deep Neural Networks explicitly tailored for the new calorimeter.

Part I

Setting the scene

Chapter 1

The Standard Model of particle physics and the Higgs boson

The Standard Model (SM) of particle physics is a theory that embeds in a unique theoretical framework three of the four fundamental forces in nature, the electromagnetic, the weak, and the strong interaction, while the gravity has not yet found a place within the framework. The theoretical skeleton of the SM was developed throughout the second half of the XX century and has so far withstood every experimental test.

Section 1.1 provides a theoretical introduction to the SM. The core of the mathematical structure of the SM is presented in Sec. 1.1.1, with the introduction of the Yang-Mills Lagrangian, while the fundamental particles are introduced in Sec 1.1.2, serving as the bricks for the model. The main focus of this chapter is the electroweak sector, which is the main target of the physics analysis presented in this thesis. The unification of the electromagnetic and weak interactions is accomplished through the Glashow-Weinberg-Salam theory, presented in Sec. 1.1.3. The cornerstone of the SM is the introduction of a scalar field, the Higgs field, which allows the Brout-Englert-Higgs mechanism to happen (Sec. 1.1.4) and permit the introduction of Yukawa couplings (Sec. 1.1.5). With their specificities, these two theories are responsible for the mass of all fundamental particles. The presentation of the SM is concluded in Sec. 1.1.6 with a discussion of its current limitations, acknowledging the boundaries and open questions within the theory. Following this theoretical introduction, Sec. 1.2 is dedicated to exploring how the Higgs boson is produced at a hadron collider and the methods employed to characterise its properties.

1.1 The Standard Model of Particle Physics

The SM is a renormalisable gauge quantum field theory based on the $SU(2)_L \times U(1)_Y \times SU(3)_C$ symmetry group. The $SU(2)_L \times U(1)_Y$ part describes the electroweak sector, the unification of the electromagnetic and weak interactions, and incorporates the interaction with the Higgs field. The $SU(3)_C$ part instead describes the Quantum ChromoDynamics (QCD), the theory governing the strong interactions. Being renormalisable is crucial for a quantum field theory to be physically meaningful. Renormalisability allows the theory to handle effectively and resolve infinities arising from calculations, providing finite and measurable predictions. In addition to being renormalisable, the SM is also a gauge theory: the Lagrangian is invariant under a local transformation, i.e., depending on the spacetime coordinates, according to a specific symmetry group. The mathematical framework that allows the construction of a gauge theory is the Yang-Mills Langrangian, which is explained in the following section.

1.1.1 Yang-Mills Lagrangian

A Yang-Mills theory is based on a Lagrangian that is required to be invariant under a local transformation, i.e., dependent on the spacetime point, according to a specific symmetry group. Two groups are relevant for the SM, the group of the unitary transformation U(N) and its subgroup SU(N):

$$\mathrm{U(N)} = \{\text{Matrices N} \times \text{N} \;\; \mathrm{U} \mid \mathrm{U^\dagger U} = \mathrm{UU^\dagger} = \mathbb{1}\}$$

$$\mathrm{SU(N)} = \{\text{Matrices N} \times \text{N} \;\; \mathrm{U} \mid \mathrm{U^\dagger U} = \mathrm{UU^\dagger} = \mathbb{1}, \det(\mathrm{U}) = 1\}$$

The dimension of the groups is given by the number of independent parameters in the matrix satisfying the conditions of the groups. For U(N) is $2\mathrm{N}^2 - \mathrm{N}^2 = \mathrm{N}^2$, where $2\mathrm{N}^2$ is the number of parameters in a N × N matrix and N^2 are the number of equations represented by $\mathrm{U^\dagger U}$, whereas for SU(N) is $2\mathrm{N}^2 - 1 - \mathrm{N}^2 = \mathrm{N}^2 - 1$, where the -1 is added due to the condition on the determinant. Each element of U(N) and SU(N) can be written as $\mathrm{U} = e^{ig\theta^a T^a}$, where a runs from 1 to the dimension of the group, θ^a are real parameters, T^a are matrices that form a basis for the group, and g is a number that assumes the meaning of the coupling constant.

Assume we have N particle fields described by N Dirac spinors [1] $\psi(x) = (\psi_1(x), ..., \psi_N(x))$ and restrict the discussion to the SU(N) group. The simplest Lagrangian that can be defined is:

$$\mathscr{L} = i\overline{\psi}(x)\partial\!\!\!/\psi(x) \tag{1.1}$$

This is the Dirac Lagrangian density corresponding to a massless particle. This Lagrangian is invariant under a global SU(N) transformation, i.e., a transformation that does not depend on the spacetime point, but it is not invariant under a local[2] SU(N) transformation $\psi(x) \to \mathrm{U}(x)\psi(x) = e^{ig\theta^a(x)T^a}\psi(x)$. The invariance fails due to the derivative:

$$\partial_\mu\psi(x) \to \partial_\mu(e^{i\theta^a(x)T^a}\psi(x)) = e^{i\theta^a(x)T^a}(\partial_\mu + iT^a\partial_\mu\theta^a(x))\psi(x) \tag{1.2}$$

To make the Lagrangian invariant under a local transformation of SU(N), the procedure is to substitute the canonical derivative with another object named *covariant derivative* D_μ, which is defined as:

$$D_\mu = \partial_\mu - igT^aA^a_\mu(x) = \partial_\mu - igA_\mu(x) \tag{1.3}$$

where $A^a_\mu(x)$ is a generic function of x, whose expression should be found in order to satisfy the invariance. These objects are called *gauge fields*, and for the group SU(N) there are $\mathrm{N}^2 - 1$ of them, equal to the dimension of the group. The gauge invariance is satisfied if the fields A_μ transform as follows:

$$A'_\mu = UA_\mu U^{-1} - \frac{i}{g}\partial_\mu UU^{-1} \tag{1.4}$$

From now on, the x dependence is implied to lighten the notation. The global symmetry of the Lagrangian shown in Eq. 1.1 is promoted to be a local symmetry at the price of including N^2-1 fields A_μ that interacts with the particle fields $\psi(x)$. Thus, the Lagrangian density becomes:

$$\mathscr{L} = i\overline{\psi}(x)D\!\!\!/\psi(x) \tag{1.5}$$

A kinetic term for the A_μ fields can be defined from the field strength tensor $F_{\mu\nu}$:

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - ig[A_\mu, A_\nu] \tag{1.6}$$

[1] A Dirac spinor is a solution of the Dirac equation and describes all fermions, with the possible exception of neutrinos.

[2] Local transformation and gauge transformation are synonyms in this context.

In order to be included in the Lagrangian, the term should be gauge invariant and a scalar. It can be proven that a term satisfying these conditions is:

$$\mathscr{L}_{\text{kin}} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} \tag{1.7}$$

The final Yang-Mills Lagrangian is given by the sum of Eq. 1.5 and Eq. 1.7:

$$\mathscr{L}_{\text{Yang-Mills}} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + i\overline{\psi} \not{D} \psi \tag{1.8}$$

This Lagrangian density is invariant under the gauge symmetry SU(N). The imposition of the gauge invariance forced the introduction of gauge fields that couple with the particle fields. Notably, it is not possible to introduce directly a mass term à-la Dirac in Eq. 1.8, for both the gauge fields ($m^2 A^a_\mu A^{\mu,a}$) and for the particle fields ($m^2 \overline{\psi}\psi$). Such terms are not allowed since they do not preserve the gauge invariance of the Lagrangian. This issue will be addressed by introducing the Higgs field, allowing the spontaneous symmetry breaking of the electroweak theory and the introduction of Yukawa terms in the Lagrangian.

1.1.2 Particles

While the Yang-Mills theory is the tool that allows the mathematical development of the SM, the ingredients of the theory are the elementary particles. There are two classes of particles in the SM that play different roles. One is composed of spin-1 bosons that are the mediators of the forces. Eight gluons mediate the strong force, while the $W^{\pm}$, Z, and γ bosons mediate the electroweak force. The other family comprises fermions, which interact among themselves through the bosons earlier introduced. The Higgs boson completes the SM - a scalar boson (spin-0 particle) that is essential to describe the world as we know it, as it will become clear in the following sections. All the actors of the SM are represented in Fig. 1.1.

Fermions are divided into six *quarks* and six *leptons*, together with the corresponding anti-fermions. Quarks and leptons are then further classified into three generations, and the elements of each generation are arranged in SU(2) doublets. The three leptons are the electron (e), the muon (μ), and the tau (τ), all with electric charge $Q = -1$. The lepton doublet is made with the corresponding neutrino:

$$\begin{pmatrix} \nu_e \\ e \end{pmatrix} \quad \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix} \quad \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}$$

Neutrinos are neutral particles and interact only through the weak force. When the SM was formulated in the second half of the XX century, the neutrino was supposed to be massless. However, the discovery in 2016 of the neutrino flavour oscillations [23] proved that neutrinos are actually massive. Despite having a non-null mass, its value(s) is tiny, making its measurement extremely challenging. The KATRIN experiment provided the best limit of $m_\nu < 0.8$ eV at 90% CL in 2022 [24]. Similarly to leptons, the three generations of quarks are composed of the up (u) and down (d) quarks, the charm (c) and the strange (s) quarks, and the top (t) and bottom (b) quarks:

$$\begin{pmatrix} u \\ d \end{pmatrix} \quad \begin{pmatrix} c \\ s \end{pmatrix} \quad \begin{pmatrix} t \\ b \end{pmatrix}$$

The charge of the upper quark in the doublet is $Q = 2/3$, while for the other quark is $Q = -1/3$. In addition to the electroweak force, the quarks are the only particles subject to the strong force. Hence, they also have a *colour* charge, where the *colour* is the charge of the strong force.

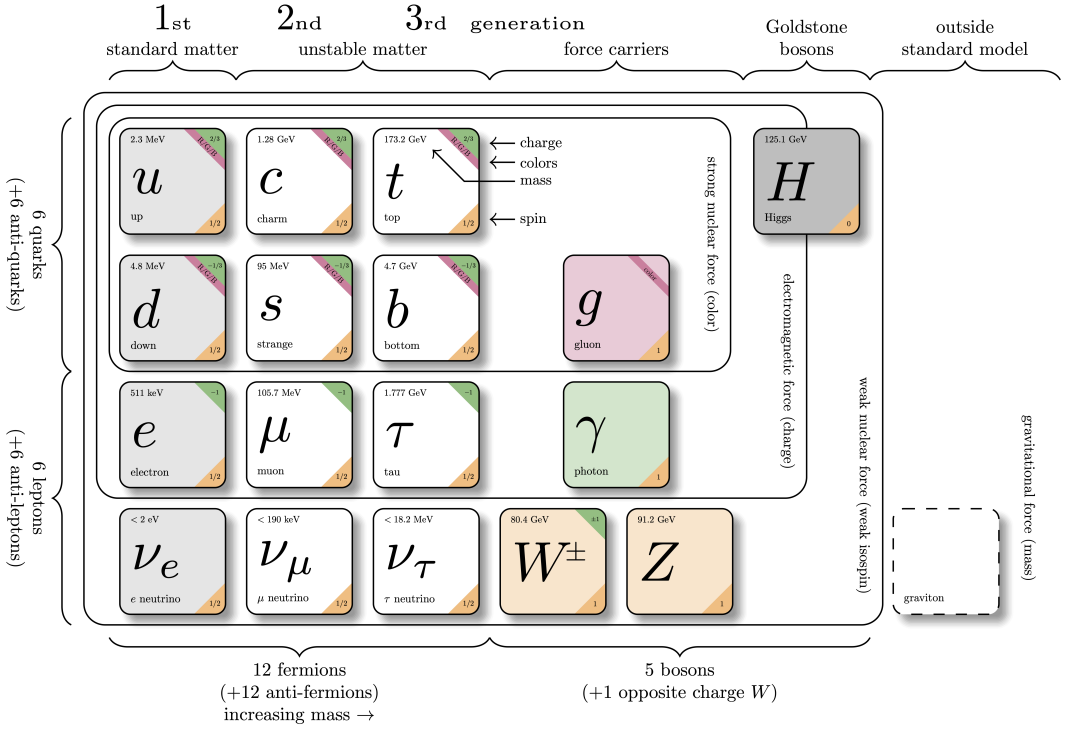

Figure 1.1: Illustration of the SM particles and their properties. The diagram also includes the graviton as an *outside standard model* particle. It is an unobserved boson that is a candidate to be the mediator of the gravitational force. Figure taken from [25].

1.1.3 Glashow-Weinberg-Salam theory

The Glashow-Weinberg-Salam (GWS) theory, also known as the electroweak theory, is a key component of the SM. It was independently proposed by Sheldon Glashow, Abdus Salam, and Steven Weinberg in the 1960s[3]. The GWS theory unifies two of the fundamental forces in nature: the electromagnetic and weak nuclear force, which were previously considered separate phenomena. The theory postulates that the electromagnetic force and the weak force are different aspects of a single electroweak force that can be described by a Yang-Mills theory with the gauge symmetry of $\mathrm{SU(2)_L \times U(1)_Y}$.

One of the most fundamental, and astonishing at the time of the discovery, feature of the weak force is the parity violation. This symmetry violation can be theoretically described by a chiral theory, where the left-handed and right-handed chiral components of the particles fields behave differently. Any Dirac spinor can be decomposed into the left- and right-handed chiral components:

$$\psi = \psi_L + \psi_R = \frac{1}{2}(1+\gamma^5)\psi + \frac{1}{2}(1-\gamma^5)\psi \tag{1.9}$$

where γ^5 is the matrix defined as:

$$\gamma^5 = \begin{pmatrix} 0_{2\times2} & \mathbb{1}_{2\times2} \\ \mathbb{1}_{2\times2} & 0_{2\times2} \end{pmatrix}$$

[3]Glashow proposed his model in 1961, three years before the work of Salam and Weinberg. Although he assumed the right symmetry $\mathrm{SU(2)_L \times U(1)_Y}$, he did not consider the possibility of spontaneous symmetry breaking. He was therefore forced to introduce *by hand* the mass terms for the gauge bosons, spoiling the renormalizability of the theory.

The eigenstates of the γ^5-matrix represent the left- and right-handed chiral states. In the limit of massless particles, the chirality is equivalent to the helicity, i.e., the projection of the spin vector over the momentum vector. The parity violation indicates that the weak force does not treat left-handed and right-handed particles symmetrically; it only interacts with left-handed particles. Therefore, in the $\mathrm{SU(2)_L}$ group, left-handed fermions are represented as doublets, whereas right-handed particles are $\mathrm{SU(2)_L}$ singlets, meaning they do not transform under the SU(2) gauge symmetry. For simplicity, only one generation of leptons will be considered in what follows without loss of generality.

$$L_L = \frac{1}{2}(1-\gamma^5)\begin{pmatrix}\nu_\ell\\ \ell\end{pmatrix} = \begin{pmatrix}\nu_{\ell_L}\\ \ell_L\end{pmatrix} \tag{1.10}$$

$$\nu_{\ell_R} = \frac{1}{2}(1+\gamma^5)\nu_\ell \tag{1.11}$$

$$\ell_R = \frac{1}{2}(1+\gamma^5)\ell \tag{1.12}$$

The $\mathrm{SU(2)_L}$ group is usually called the *weak isospin group*. Due to the chiral nature of the gauge symmetry, the form of the generators of this group depends on which chiral state they act. When acting on doublets, the generators are $T^a = \sigma^a/2$, where σ^a are the three Pauli matrices; otherwise, if the group acts on singlets, the unique generator is $T^a = 0$. The Pauli matrices are defined as:

$$\sigma_1 = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix} \quad \sigma_2 = \begin{pmatrix}0 & -i\\ i & 0\end{pmatrix} \quad \sigma_3 = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}$$

As discussed in Sec. 1.1.1, the gauge invariance introduces three gauge bosons W^1_μ, W^2_μ, and W^3_μ, and the gauge coupling is denoted as g. The $\mathrm{U(1)_Y}$ group has only one generator $Y(\psi)$, whose value depends on the fermionic field. For this group, there is only one gauge boson B_μ with the gauge coupling denoted as g'. Thus, the Lagrangian can be written as:

$$\mathscr{L} = -\frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}W^i_{\mu\nu}W_i^{\mu\nu} + i\overline{L}_L \not{D} L_L + i\overline{\nu}_{\ell_R}\not{D}\nu_{\ell_R} + i\overline{\ell}_R\not{D}\ell_R \tag{1.13}$$

where the covariant derivative is:

$$D_\mu = \partial_\mu - igW^i_\mu T^i - ig'\frac{Y(\psi)}{2}B_\mu \tag{1.14}$$

where the generator of $\mathrm{U(1)_Y}$ is divided by 2 for convention. We focus on the part with the covariant derivative first. Three components can be identified. The first one is the kinetic term for the fermionic fields, describing their propagation in spacetime:

$$\mathscr{L}_{\text{kin}} = i\overline{L}_L\not{\partial} L_L + i\overline{\nu}_{\ell_R}\not{\partial}\nu_{\ell_R} + i\overline{\ell}_R\not{\partial}\ell_R \tag{1.15}$$

The second component is related to the Charged Currents (CC):

$$\begin{aligned}\mathscr{L}_{CC} &= gW^1_\mu\overline{L}_L\gamma^\mu\frac{\sigma_1}{2}L_L + gW^2_\mu\overline{L}_L\gamma^\mu\frac{\sigma_2}{2}L_L\\ &= \frac{g}{\sqrt{2}}W^+_\mu\overline{\nu}_L\gamma^\mu\ell_L + \frac{g}{\sqrt{2}}W^-_\mu\overline{\ell}_L\gamma^\mu\nu_L\end{aligned} \tag{1.16}$$

The CC can be seen as the sum of two currents[4] if the W^1_μ and W^2_μ bosons are linearly combined to define two new bosons $W^\pm_\mu$, which will turn out to be the physical W bosons, mediators of the weak force. The relation between the two sets of gauge fields is:

$$W^\pm_\mu \equiv \frac{1}{\sqrt{2}}(W^1_\mu \mp W^2_\mu) \tag{1.17}$$

[4]The currents are crucial components in the mathematical formulation of the SM. They are related to the Noether's theorem, and they describe the interactions between bosons and fermions.

The third component is related to the Neutral Currents (NC):

$$\begin{aligned}\mathscr{L}_{\text{NC}} =& \frac{g}{2} W_\mu^3 (\overline{\nu}_{\ell_L} \gamma^\mu \nu_{\ell_L} - \overline{\ell}_L \gamma^\mu \ell_L) \\ &+ \frac{g'}{2} B_\mu [Y(L)(\overline{\nu}_{\ell_L} \gamma^\mu \nu_{\ell_L} + \overline{\ell}_L \gamma^\mu \ell_L) + Y(\nu_{\ell_R}) \overline{\nu}_{\ell_R} \gamma^\mu \nu_{\ell_R} \\ &+ Y(\ell_R) \overline{\ell}_R \gamma^\mu \ell_R]\end{aligned} \tag{1.18}$$

Unlike the CC Lagrangian, in the NC Lagrangian neither W_μ^3 nor B_μ is a good candidate to be the photon; they both couple to the neutrinos that, being neutral, do not interact electromagnetically. In order to write a more compact version of Eq. 1.18, three objects are introduced: $\Psi \equiv (\nu_{\ell_L}, \ell_L, \nu_{\ell_R}, \ell_R)$, $\mathcal{T}_3 \equiv \text{diag}(1/2, -1/2, 0, 0)$, and $\mathcal{Y} \equiv \text{diag}(Y(L), Y(L), Y(\nu_{\ell_R}), Y(\ell_R))$:

$$\mathscr{L}_{\text{NC}} = g \overline{\Psi} \gamma^\mu \mathcal{T}_3 \Psi W_\mu^3 + g' \overline{\Psi} \gamma^\mu \frac{\mathcal{Y}}{2} \Psi B_\mu \tag{1.19}$$

The GWS theory dictates a rotation in the space of the two neutral gauge fields W_μ^3 and B_μ to find another gauge field that can be interpreted as the photon field. The rotation is defined by the matrix:

$$\begin{pmatrix} B_\mu \\ W_\mu^3 \end{pmatrix} = \begin{pmatrix} \cos\theta_w & -\sin\theta_w \\ \sin\theta_w & \cos\theta_w \end{pmatrix} \begin{pmatrix} A_\mu \\ Z_\mu \end{pmatrix} \tag{1.20}$$

where θ_w is named *Weinberg angle*. The names of the new fields are not random since they will be identified as the physical photon field A_μ and the physical Z boson field Z_μ. The Lagrangian can then be written as:

$$\mathscr{L}_{\text{NC}} = \overline{\Psi} \gamma^\mu \left(g \sin\theta_w \mathcal{T}_3 + g' \cos\theta_w \frac{\mathcal{Y}}{2} \right) \Psi A_\mu + \overline{\Psi} \gamma^\mu \left(g \cos\theta_w \mathcal{T}_3 - g' \sin\theta_w \frac{\mathcal{Y}}{2} \right) \Psi Z_\mu \tag{1.21}$$

The first term can be interpreted as the current that couples the photon field A_μ with fermions Ψ by setting the electric charge as:

$$g \sin\theta_w \mathcal{T}_3 + g' \cos\theta_w \frac{\mathcal{Y}}{2} = eQ \tag{1.22}$$

On the other hand, the hypercharge $\mathcal{Y}$ is always a multiplicative factor in front of g', hence the arbitrariness in the choice of its value. This freedom is used to set $Y(L) = -1$. The doublet of left-handed fermions gives:

$$\begin{cases} 0 = \frac{g}{2} \sin\theta_w - \frac{g'}{2} \cos\theta_w \\ -e = -\frac{g}{2} \sin\theta_w - \frac{g'}{2} \cos\theta_w \end{cases} . \tag{1.23}$$

so that:

$$g \sin\theta_w = g' \cos\theta_w = e \tag{1.24}$$

Substituting in Eq. 1.22 gives the *Gell-Mann-Nishijima* relation:

$$Q = \mathcal{T}_3 + \frac{\mathcal{Y}}{2} \tag{1.25}$$

From this relation, the other hypercharges can be computed, resulting in $Y(\nu_{\ell_R}) = 0$ and $Y(\ell_R) = -2$. The ν_{ℓ_R} has null hypercharge, null electric charge, and it is a singlet of

$SU(2)_L$; it can be discarded since it does not participate in any interaction[5]. Combining all relations obtained above, the $\mathscr{L}_{NC}$ can be rewritten as:

$$\mathscr{L}_{NC} = e\overline{\Psi}\gamma^{\mu}Q\Psi A_{\mu} + \overline{\Psi}\gamma^{\mu}Q_Z\Psi Z_{\mu} \tag{1.26}$$

where:

$$Q_Z = \frac{e}{\cos\theta_w \sin\theta_w}(\mathcal{T}_3 - Q\sin^2\theta_w) \tag{1.27}$$

It was shown that the covariant derivative describes naturally the interactions of the $W^{\pm}$, Z, and γ bosons with the fermions and describe accurately the behaviour of the electroweak force. All this procedure can be iterated with the three lepton generations and the three quark generations, *mutatis mutandis*.

The field strength tensors component of Eq. 1.13 still need to be discussed. The Weinberg rotation should be propagated to these terms as well. However, being an orthogonal transformation, the rotation preserves the form of these terms in the Lagrangian:

$$-\frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}W^i_{\mu\nu}B^{\mu\nu}_i = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} - \frac{1}{2}W^+_{\mu\nu}W^{\mu\nu}_- - \frac{1}{4}Z_{\mu\nu}Z^{\mu\nu} \tag{1.28}$$

where:

$$\begin{cases} F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \\ Z_{\mu\nu} = \partial_\mu Z_\nu - \partial_\nu Z_\mu \\ W^{\pm}_{\mu\nu} = \partial_\mu W^{\pm}_\nu - \partial_\nu W^{\pm}_\mu \end{cases} . \tag{1.29}$$

This piece of the Lagrangian is responsible for the trilinear and quadrilinear couplings between the gauge bosons.

The GWS theory successfully describes both weak and electromagnetic interactions as different aspects of the unified electroweak force. It accurately predicts the interactions among gauge bosons and their interactions with fermionic fields. However, one piece is still missing, the mass term. Experimental evidence indicates that the $W^{\pm}$ and Z bosons do possess mass, yet introducing a simple Dirac mass term would lead to a violation of the crucial gauge symmetry in the theory (Sec. 1.1.1). This is where the Higgs field comes to our aid.

1.1.4 Brout-Englert-Higgs mechanism

The $W^{\pm}$ and Z bosons are proven experimentally to be massive, but the gauge symmetry prevents adding any mass term in the Lagrangian. To address this issue, the SM incorporates the Brout-Englert-Higgs (BEH) mechanism, which elegantly reconciles the need for mass terms for the gauge bosons with the gauge symmetry. The main protagonist of the BEH mechanism is the scalar field named *Higgs field* that acquires a non-zero vacuum expectation value, leading to the spontaneous breaking of the electroweak symmetry. Suppose we have a Lagrangian that is invariant under a specific transformation; therefore, the set of the fundamental states is invariant under this symmetry. Spontaneous symmetry breaking is the phenomenon whereby nature chooses one specific state among the many possible options. The BEH mechanism is based on the introduction of 4 scalar real fields arranged in a $SU(2)_L$ doublet:

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} \phi^0_1 + i\phi^0_2 \\ \phi^+_1 + i\phi^+_2 \end{pmatrix} \tag{1.30}$$

[5]Actually the right-handed neutrino is coming back to life to include massive neutrinos in the SM. However, its role is strictly related to the nature of neutrinos, depending on whether they are Fermi or Majorana fermions. This is still an open issue in particle physics and an active research area.

The Higgs term of the Lagrangian is given by:

$$\mathscr{L}_{\text{BEH}} = (D_\mu \phi)^\dagger D^\mu \phi - V(\phi^\dagger \phi) \tag{1.31}$$

where the covariant derivative D_μ of ϕ is given by Eq. 1.14. Its insertion assures the $\mathscr{L}_{\text{BEH}}$ to be invariant under the $\text{SU(2)}_\text{L}\times\text{U(1)}_\text{Y}$ gauge symmetry. To ensure renormalisability and $\text{SU(2)}_\text{L}\times\text{U(1)}_\text{Y}$ gauge invariance, the BEH potential $V(\phi^+\phi)$ is of the form:

$$V(\phi^\dagger \phi) = -\mu^2 \phi^\dagger \phi + \lambda(\phi^\dagger \phi)^2 \tag{1.32}$$

where μ^2 and λ are positive real numbers. Although they look alike, $-\mu^2\phi^\dagger\phi$ is not a proper mass term. A mass term would require a positive sign, while here, it is negative. The minima of the potential are given by the points:

$$\phi^\dagger \phi = \frac{\mu^2}{2\lambda} \equiv \frac{v^2}{2} \tag{1.33}$$

Equation 1.30 can be rewritten as:

$$\phi(x) = \begin{pmatrix} \phi^+ \\ \frac{1}{\sqrt{2}}(v + H(x) + i\chi(x)) \end{pmatrix} = \frac{1}{\sqrt{2}} e^{i\frac{\sigma_i \theta^i(x)}{v}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix} \tag{1.34}$$

where $H(x)$ and $\chi(X)$ are real fields parametrising ϕ_0 as deviations from the minimum, while $\theta_i(x)$ are three real fields called *Goldstone bosons*. The freedom to perform a SU(2)_L gauge transformation on ϕ can be used to absorb the three Goldstone bosons:

$$\phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix} \tag{1.35}$$

This gauge is usually called *unitary*, and it has the advantage of manifesting explicitly the physical quantities. The $H(x)$ will be associated with the Higgs field. It can be easily verified that the vacuum state breaks the $\text{SU(2)}_\text{L}\times\text{U(1)}_\text{Y}$ gauge symmetry. After choosing one of the infinite possible minima, all connected by the gauge transformation, the new configuration is no longer invariant: this is the spontaneous symmetry breaking. Nevertheless, it can be proven that the new configuration is now invariant under the group U(1)_em, which is the gauge symmetry of the quantum electrodynamics. The covariant derivative expanded around the vacuum state becomes:

$$D_\mu \phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \partial_\mu H \end{pmatrix} - \frac{i}{2}\left(1 + \frac{H}{v}\right) \begin{pmatrix} g v W_\mu^+ \\ \frac{v}{\sqrt{2}} A_\mu e(Y(\phi) - 1) - \frac{v}{\sqrt{2}} Z_\mu (g \cos\theta_w + g' Y(\phi) \sin\theta_w) \end{pmatrix} \tag{1.36}$$

This is $D_\mu\phi$, but it is $(D_\mu\phi)^\dagger(D_\mu\phi)$ that enters the Lagrangian. When computing this contribution, the term with A_μ would lead to a mass term for the photon, which should be avoided. Thus, the hypercharge of the ϕ field is set to $Y(\phi) = 0$ in order to suppress the possibility of a massive photon. After the electroweak symmetry breaking, the photon is still massless by our imposition. Eventually, the covariant derivative generates a term given by:

$$(D_\mu \phi)^\dagger D^\mu \phi = \frac{1}{2}\partial_\mu H \partial_\mu H + \left[\left(\frac{gv}{2}\right)^2 W_\mu^+ W_-^\mu + \frac{1}{2}(g + g')\frac{v^2}{4} Z\mu Z^\mu\right]\left(1 + \frac{H}{v}\right)^2 \tag{1.37}$$

This is the crucial point of the BEH mechanism and its greatest success. The Lagrangian now contains two mass terms for the $\text{W}^\pm$ and Z bosons, where the masses are given by:

$$m_W^2 = \frac{g^2 v^2}{4} \qquad m_Z^2 = \frac{(g^2 + g'^2)v^2}{4} \tag{1.38}$$

The HWW and HZZ couplings arise from the $2H/v$ term, while the HHWW and HHZZ couplings from H^2/v^2.

The covariant derivative is one half of the BEH Lagrangian, the other half is given by the scalar potential that, when expanded around the chosen minimum, becomes:

$$V = \frac{1}{2}(2\lambda v^2)H^2 + \lambda v H^3 + \frac{\lambda}{4}H^4 - \frac{\lambda}{4}v^4 \tag{1.39}$$

Important considerations can be derived from this term. Firstly, the Higgs field is massive:

$$m_H^2 = 2\lambda v^2 \tag{1.40}$$

The contributions proportional to H^3 and H^4 are responsible for the trilinear and quadrilinear self-couplings, respectively. These two fundamental features of the Higgs boson make it unique since no other particle in the SM interacts with itself without changing its nature. In addition, they also give direct access to the λ parameter of the Higgs potential. Last but not least, the constant term proportional to v^4. Since it is constant, it could be removed by the Lagrangian without any change in the physics[6].

1.1.5 Yukawa couplings

The next major missing piece in the SM is the generation of the fermion masses. Due to the introduction of the Higgs scalar, the mass term for fermions is introduced in the SM via the Yukawa Lagrangian density:

$$\begin{aligned}\mathscr{L}_{\text{yukawa}} = &- \Gamma_d^{ij}\overline{Q}_L'^i \phi d_R'^j + \text{h.c.} (= -\Gamma_d^{ij*}\overline{d}_R'^i \phi^\dagger Q_L'^j) \\ &- \Gamma_u^{ij}\overline{Q}_L'^i \phi_c u_R'^j + \text{h.c.} \\ &- \Gamma_e^{ij}\overline{L}_L'^i \phi e_R'^j + \text{h.c.}\end{aligned} \tag{1.41}$$

This term satisfies all the conditions in order to be included in the SM Lagrangian: dimension-four, Lorentz-invariance, $\mathrm{SU(2)_L \times U(1)_Y}$ gauge invariance, and renormalizability. Q', u', and d' (L'_L and e'_R) are quark (lepton) fields that are generic linear combinations of the mass eigenstates u and d (e) – this choice will become clear at the end of this section. Γ_u, Γ_d, and Γ_e are 3×3 complex matrices that represent the arbitrary Yukawa couplings. The indices i and j are generation labels such that:

$$Q_L^i = \begin{pmatrix} u_L^i \\ d_L^i \end{pmatrix} \to \begin{pmatrix} u_L \\ d_L \end{pmatrix}, \begin{pmatrix} c_L \\ s_L \end{pmatrix}, \begin{pmatrix} t_L \\ b_L \end{pmatrix} \qquad L_L^i = \begin{pmatrix} \nu_{e_L}^i \\ e_L^i \end{pmatrix} \to \begin{pmatrix} \nu_{e_L}^i \\ e_L \end{pmatrix}, \begin{pmatrix} \nu_{\mu_L}^i \\ \mu_L \end{pmatrix}, \begin{pmatrix} \nu_{\tau_L}^i \\ \tau_L \end{pmatrix}$$

$$u_R^i \to u_R, c_R, t_R \qquad d_R^i \to d_R, s_R, b_R \qquad e_R^i \to e_R, \mu_R, \tau_R$$

The ϕ_c field is introduced to give mass also to the lower part of the fermion $\mathrm{SU(2)_L}$ doublets, and it is defined as:

$$\phi_c = i\sigma_2\phi^* = \frac{1}{\sqrt{2}}\begin{pmatrix} v + H \\ 0 \end{pmatrix} \tag{1.42}$$

Equation 1.41 could also accommodate another term to account for the neutrino mass. The inclusion of neutrino masses is still an open topic in theoretical physics and goes

[6] If gravity is considered, the constant term cannot be neglected.

beyond the scope of this introductory chapter. In the following, the neutrino will be considered massless. Each term should be evaluated in the unitary gauge:

$$\overline{Q}'^i_L \phi d'^j_R = (\overline{u}'^i_L \;\; \overline{d}'^i_L)\begin{pmatrix} 0 \\ \frac{v+H}{\sqrt{2}} \end{pmatrix} d'^j_R = \frac{v+H}{\sqrt{2}} \overline{d}'^i_L d'^j_R \tag{1.43}$$

$$\overline{Q}'^i_L \phi_c u'^j_R = (\overline{u}'^i_L \;\; \overline{d}'^i_L)\begin{pmatrix} \frac{v+H}{\sqrt{2}} \\ 0 \end{pmatrix} u'^j_R = \frac{v+H}{\sqrt{2}} \overline{u}'^i_L u'^j_R \tag{1.44}$$

Equation 1.43 shows why the ϕ_c is necessary to provide mass to the up-type quarks[7]. In order to give a mass dimension to the Yukawa matrix, the $M^{ij} = \Gamma^{ij}(v/\sqrt{2})$ matrix is introduced:

$$\mathscr{L}_{\text{yukawa}} = -\left(1 + \frac{H}{v}\right)(M^{ij}_u \overline{u}'^i_L u'^j_R + M^{ij}_d \overline{d}'^i_L d'^j_R + M^{ij}_e \overline{e}'^i_L e'^j_R + \text{h.c.}) \tag{1.45}$$

The terms in the Yukawa Lagrangian can be potentially mass terms if $i = j$, which is achieved only if the matrices M are diagonals. At this point, a little help from linear algebra is needed. For any generic complex square matrix C, two unitary matrices K and V exist so that $D = K^\dagger C V$ is real and diagonal. Therefore, the M_f matrices, with $f = (u, d, e)$, can be diagonalised with two unitary matrices K_f and V_f. The diagonalization leads to a redefinition of the fields as:

$$f'_{Li} = (K_f)_{ij} f_{Lj} \qquad f'_{Ri} = (V_f)_{ij} f_{Rj} \tag{1.46}$$

Combining all together, the Yukawa Lagrangian manifests mass terms for fermions:

$$\mathscr{L}_{\text{yukawa}} = -\sum_f \overline{f}'^i_L M^{ij}_f f'^j_R \left(1 + \frac{H}{v}\right) + \text{h.c.} \tag{1.47}$$

$$= -\sum_f \overline{f}'^i_L [(K_f)^\dagger M_f V_f]^{ij} f'^j_R \left(1 + \frac{H}{v}\right) + \text{h.c.} \tag{1.48}$$

$$= -\sum_f m_f (\overline{f}_L f_R + \overline{f}_R f_L)\left(1 + \frac{H}{v}\right) \tag{1.49}$$

where m_f is the mass of the fermion f that results to be totally arbitrary in the Yukawa theory. When multiplying by 1 in Eq. 1.47, one gets the mass term for the fermions, while when multiplying by H/v one gets the couplings of the Higgs boson with all fermions, and this coupling is proportional to the mass itself.

Cabibbo-Kobayashi-Maskawa matrix

Incorporating mass terms for fermions required a redefinition of the fermion fields, a change that should be propagated throughout the SM Lagrangian. The only part of the Lagrangian that undergoes physical changes is the charged current interaction described by Eq. 1.16. The quark component can be rewritten as:

$$\mathscr{L}_{\text{CC}} = \frac{e}{\sqrt{2}\sin\theta_w} \overline{u}'^i_L \gamma^\mu W^+_\mu d'^i_L + \text{h.c.} \tag{1.50}$$

After the mass diagonalization, the modified form of this interaction is given as:

$$\mathscr{L}_{\text{CC}} = \frac{e}{\sqrt{2}\sin\theta_w} \overline{u}^i_L [(K_u)^\dagger V_d]_{ij} \gamma^\mu W^+_\mu d^i_L + \text{h.c.} \tag{1.51}$$

[7] And neutrinos

The matrix $V_{\mathrm{CKM}} \equiv (K_u)^\dagger V_d$ is defined as the Cabibbo-Kobayashi-Maskawa matrix. It is a complex unitary matrix that describes the mixing of quark flavours in weak interactions. In the SM, there are not, at tree-level, neutral currents that change the lepton flavours; there are, however, due to the Yukawa couplings, charged currents that do it. Each entry (V_{CKM}^{ij}) represents the probability amplitude for a quark of flavour i to transform into a quark of flavour j through weak interactions. The theory cannot predict the elements of the matrix; instead, they must be determined experimentally.

$$V_{\mathrm{CKM}} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} = \begin{pmatrix} 0.97373 \pm 0.00031 & 0.2243 \pm 0.0008 & 0.00382 \pm 0.00020 \\ 0.221 \pm 0.004 & 0.975 \pm 0.006 & 0.0408 \pm 0.0014 \\ 0.0086 \pm 0.0002 & 0.0415 \pm 0.0009 & 1.014 \pm 0.029 \end{pmatrix}$$

The experimental values of the CKM matrix are taken from [26]. There is no theoretical explanation for the hierarchy of the CKM elements. However, it can be noted that the transitions within the same generation are more likely to happen than the ones between quarks from two different generations: the larger the difference in generation between the quarks, the smaller the CKM element, hence the coupling.

1.1.6 Limitations of the Standard Model

The SM is the state-of-the-art knowledge of particle physics. It is a mathematical, solid theory that has been tremendously successful in describing all experimental evidence. After the discovery of the Higgs boson in 2012 [19, 20], the SM is now complete. However, beneath its successes lie glaring unanswered questions that leave us unsatisfied and point out that a more general theory must be out there. The SM arises from the gauge symmetry group $\mathrm{SU(2)_L \times U(1)_Y \times U(3)_C}$, yet the reasons behind this specific symmetry group remain elusive. Moreover, nineteen parameters in the SM can only be constrained by experimental data, leaving us to wonder why such a seemingly arbitrary set of values governs our universe. The presence of three generations of quarks and leptons and the exclusion of gravity from the SM beg for deeper insights into the underlying principles that shape our universe. The original formulation of the theory envisaged massless neutrinos. However, the discovery of neutrino oscillations [23] was a clear proof that neutrinos are massive, although their mass is tiny. This could be seen as a hint of physics Beyond the Standard Model (BSM). Nevertheless, neutrino masses can be included by easily extending the current SM. This leads to uncomfortable solutions like increasing the free parameters and introducing back the right-handed neutrino, a sterile particle that interacts with nothing and is impossible to detect. These are just some of the profound questions that the SM is unable to answer that limit our understanding of the universe. Additionally, experimental evidence from other fields sheds light on some limitations of the SM. For instance, on a cosmological scale, the enigmatic, hypothetical dark matter and dark energy, which together make 95% of the universe, lie beyond the grasp of the SM.

These are just some limitations that motivate the quest for BSM physics. Theoreticians have proposed a large plethora of BSM models, such as supersymmetry, grand unified theories, and string theories, to address the SM limitations. Many of these proposals await experimental validation. The search for BSM physics follows two main paths: the hunt for new resonances and deviations from the SM expectations. If the mass scale of new physics is beyond the reach of the current experimental apparatus, the evidence of new resonances would be impossible. If this is the case, the experimental results can be parametrised in an effective Lagrangian with higher dimensional operators to characterise potential BSM effects. This approach assumes that the SM is only an Effective Field Theory (EFT), i.e., a low-energy approximation of a more general theory. The EFT approach remains agnostic to the specific BSM model, offering a tool to explore the unknown realm of BSM

physics. These lines of research in experimental particle physics, namely the quest for new resonances and precision physics, hold the promise of shedding light on the mysteries that beckon us beyond the boundaries of the Standard Model.

1.2 The Higgs boson at a proton-proton collider

The Higgs boson can be produced at a proton-proton collider through different processes. The main Feynman diagrams of the five leading production mechanisms are displayed in Fig. 1.2: gluon fusion (ggH); vector boson fusion (VBFH); associated production with vector bosons (VH) or two top quarks (ttH). The mechanism with the largest cross section is the ggH process mediated by the exchange of a virtual top quark in the loop. The processes mediated by lighter quarks are suppressed proportionally to m_q^2. The second-most prominent production mode is the VBFH process, which is characterised by the scattering of two (anti-)quarks, mediated by the t- or u-channel exchange of a W or Z boson, with the Higgs boson radiated off the weak boson propagator. The two outgoing quarks produce two back-to-back hard jets in the forward and backward regions of the detector. Since the electroweak force mediates the process, the gluon radiation from the central rapidity regions is strongly suppressed. Next in line, the associated production with a W or Z boson is usually called *Higgsstrahlung* since the Higgs boson is radiated off the vector boson. The Higgs boson production with a top quark pair is about 100 times rarer than ggH. Still, it is the only one giving direct access to the Yukawa H–t coupling. The associated production with bottom quarks (bbH) has a comparable cross section to the ttH process, but it is not considered in the analysis presented in this thesis as it is expected to provide a negligible contribution. The cross section of every production mode depends on the centre-of-mass energy of the pp collision and the value of the Higgs boson mass. In a 10 GeV window around the measured value of $m_{\mathrm{H}} \simeq 125$ GeV, the cross sections are almost flat, while the centre-of-mass energy dependence is reported in the left plot of Fig. 1.3.

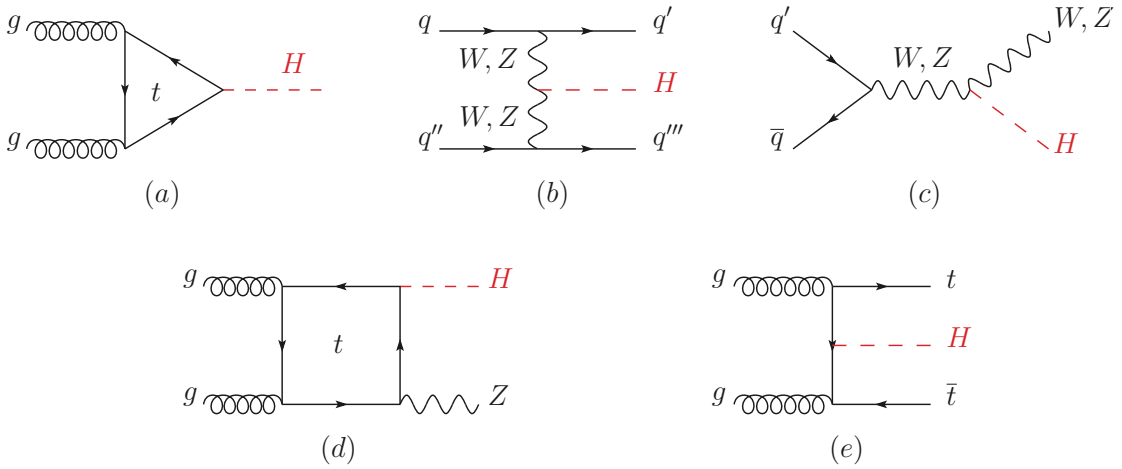

Figure 1.2: Main leading-order Feynman diagrams contributing to the Higgs boson production: (a) gluon fusion; (b) vector boson fusion; (c) associated production with a gauge boson at tree level from a quark-quark interaction; (d) associated production with a gauge boson at loop level from a gluon-gluon interaction; (e) associated production with a pair of top quarks. Figure from [26].

The study of the Higgs boson properties hinges on investigating its various decay channels, and nature has positioned the Higgs boson at a mass value that grants access to a large variety of couplings to the SM particles (right plot of Fig. 1.3). The couplings

of the Higgs boson are proportional to the mass of the coupled particle, endowing heavier particles with larger Branching Ratios (BRs). Consequently, the experimental landscape is dominated by the decays $H \to b\bar{b}$ and $H \to WW^*$, followed by $H \to gg$, $H \to \tau^+\tau^-$, $H \to c\bar{c}$, and $H \to ZZ^*$. When the Higgs boson decays to a W or Z boson, at least one of the two vector bosons is forced to be off-shell, as the mass of the Higgs boson is insufficient to produce two on-shell W or Z. With a BR one or two orders of magnitudes smaller, the Higgs boson decays to $H \to \gamma\gamma$, $H \to Z\gamma$, and $H \to \mu\mu$. Although the Higgs boson couples directly only to massive particles, the decay to massless particles $\gamma\gamma$, $Z\gamma$, and gg, are accessible since the process occurs through loops of heavy particles. As a result, these decay modes play a crucial role, as, through their loops, they provide a means to probe scales well above the Higgs boson mass: a particle too heavy to be produced directly could participate in the loop. One noteworthy channel is missing in the picture, the coupling of the Higgs boson to the most massive fermion in the SM, $H \to t\bar{t}$. This channel only opens at approximately $2m_t$, well above the mass of the discovered Higgs boson.

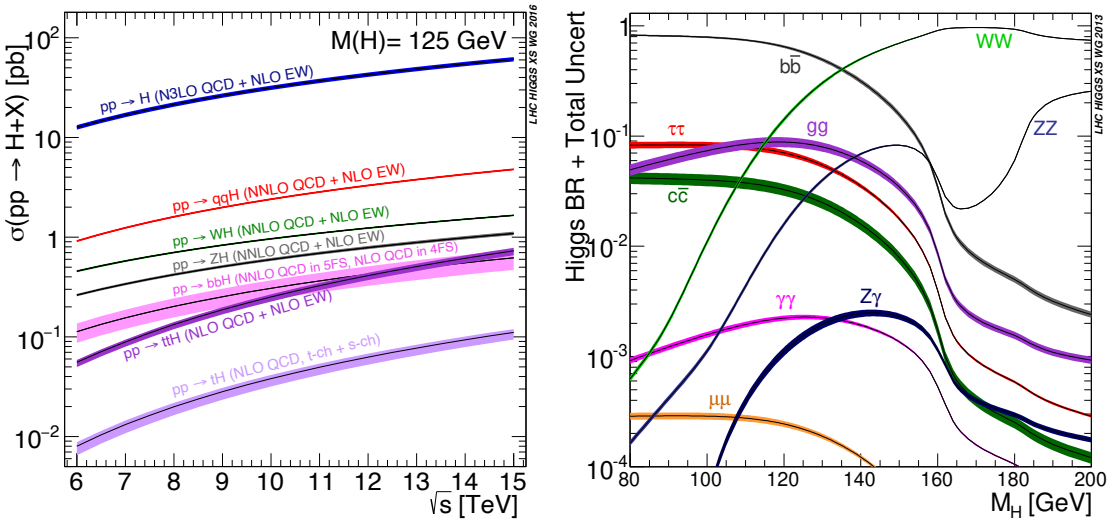

Figure 1.3: (Left) The SM Higgs boson production cross section as a function of the centre-of-mass energy, $\sqrt{s}$, for pp collisions. (Right) The branching ratios for the main decays of the SM Higgs boson. In both plots the theoretical uncertainties are indicated as bands. Figure from [27].

1.2.1 Why study the Higgs boson?

The study of the Higgs boson has been a central objective of the physics programmes of the CMS and ATLAS collaborations as it could be tied to potential solutions to the limitations outlined in Sec. 1.1.6. The *hierarchy problem* is one of the major puzzles of the SM. The weak and Higgs interactions are much stronger than the gravitational interaction by a factor of about 10^{32}. If one wishes to find a theory that unifies all the fundamental forces of nature, this scale difference poses some challenges. One of the possible solutions implies that the Higgs boson is not fundamental as prescribed by the SM but rather composed of yet undiscovered fundamental constituents [28–33]. The compositeness of the Higgs boson would imply deviations of the SM Higgs boson couplings, including the self-couplings, and the presence of partners of the top and bottom SM quarks.

The *baryonic asymmetry*, namely the abundance of matter over antimatter, is another problem that cannot be explained within the SM. A possible solution could be found in extended Higgs sectors predicting the presence of additional Higgs bosons. One of the simplest of these extensions is the so-called Two Higgs Doublet Model (2HDM) [34]. This

model prescribes the presence of the observed 125 GeV CP-even neutral scalar Higgs boson, an additional CP-even neutral scalar boson, a CP-odd Higgs boson, and a pair of charged Higgs bosons. These additional bosons can be either searched directly or probed through the precision measurements of the Higgs couplings.

The transition from a universe where matter and antimatter are present in equal quantities to a universe where the former prevails over the latter, like the one where we live, can be explained through a process called *baryogenesis* that might have happened in the early universe. Baryogenesis can proceed through two different processes: leptogenesis or electroweak baryogenesis. The leptogenesis is more difficult to probe and is related to neutrino physics [35]. On the other hand, the electroweak baryogenesis can be tested at the LHC as it prescribes deviations of the SM Higgs potential and additional scalars at the electroweak scale that couple to the Higgs boson [36]. These particles can be searched directly if kinematically allowed or indirectly through their impact on the Higgs boson couplings. In particular, these new scalars affect the coupling of the Higgs boson to the Z boson, which is extensively investigated in this thesis.

For the baryogenesis to happen, the Sakharov conditions should be satisfied [37]. One of these conditions requires a sizeable CP violation. In the SM, the only source of CP violation lies in the electroweak sector, and it is not enough to explain the baryon asymmetry. The Higgs sector could provide an additional source of CP violation. Although the experimental evidence points to a CP-even resonance, it leaves room for small CP-violating Higgs boson decays. Precious information can be extracted from the angular analysis of the decay of the Higgs boson into vector bosons and the difference in pseudorapidity between the hard jets generated in the VBFH mechanism.

As the Higgs boson is responsible for the mass of SM particles, it could also be responsible for the dark matter mass. The dark matter could be constituted by a series of particles that could couple to the SM Higgs boson, providing a portal to the unexplored *dark world*. One possible way to probe the presence of dark matter at a collider is to study the decay of the Higgs boson into invisible, i.e., non-detectable, particles. Production of such particles in the collisions would create an energy imbalance with visible particles, which can be measured.

The pattern of the masses of the SM particles and their coupling values to the SM Higgs boson represents an intriguing mystery to decipher. This configuration could arise naturally in some models predicting a different structure of the Higgs sector [38–40]. Experimentally, these models could be proven by finding non-SM Higgs boson interactions or decays into a pair of quarks or leptons with different flavours, such as $H \to \tau^{\pm}\mu^{\mp}$.

In summary, this non-exhaustive section shows how the Higgs boson represents a potential portal to provide answers to many of the open questions in particle physics. Many of the proposed solutions predict the existence of new particles that could couple with the SM Higgs boson. Even if producing these particles directly were impossible at the LHC due to insufficient energy, they could influence the Higgs boson production and decay rates. Hence, the pivotal importance of measuring cross sections.

1.2.2 Measuring the Higgs boson cross section

One of the best methods to decipher the properties of the observed Higgs-like boson is to measure cross sections. They represent a way to test theoretical computations, access the value of the couplings, and infer the presence of particles in quantum loops. An ideal measurement should extract as much information as possible from data while, at the same time, being model-independent. These requirements are however in contradiction. Maximising one reduces the other. The high-energy physics community has standardised the cross section measurements to have different levels of compromise between model independence and experimental sensitivity. The three approaches are summarised graphically

in Fig. 1.4.

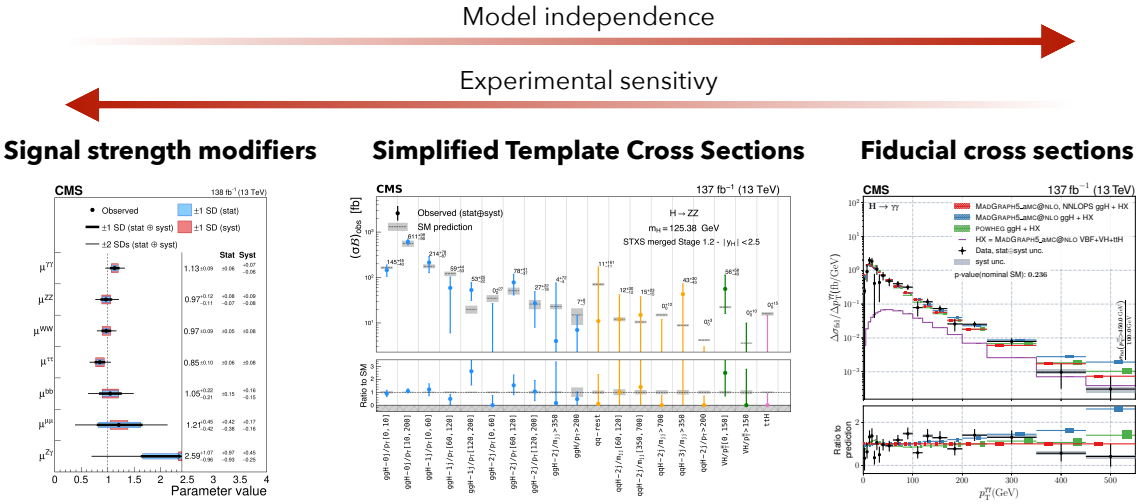

Figure 1.4: Signal strength modifiers, simplified template cross sections, and fiducial cross sections represent three benchmarks for measuring the Higgs boson cross section, providing different levels of compromise between model independence and experimental sensitivity. Below each approach, an example is reported and they are taken, from left to right, from Refs. [41], [42], and [43], respectively.

Formal introduction to the unfolding problem

Before going into the details of the various approaches, a formal introduction to the unfolding is indispensable. Let us consider a physical variable, denoted as x, characterised by a probability density function (pdf) $f(x)$. Due to limited acceptance and resolution of the detector, as well as background processes, the observed distribution becomes $g(y)$, which differs from the true distribution $f(x)$. It is worth emphasising that both x and y represent the same underlying physical quantity, such as the transverse momentum of a particle or the number of jets in an event. The different names merely stress that one is the *true* variable (x), while the other is the *measured* variable (y). Simulations rely on the forward process, from an assumption $f(x)^{\text{model}}$ on the true distribution $f(x)$ to the expected measured distribution $g(y)^{\text{expected}}$. The unfolding process, on the other hand, is the inverse process, aiming to deduce the true distribution $f(x)$ from the measured distribution $g(y)$. Mathematically, the unfolding is formulated as an integral of the type:

$$\int k(y,x)f(x)\mathrm{d}x = g(y) \tag{1.52}$$

where $f(x)$ is the term the unfolding aims to extract, $g(y)$ is measured, and the kernel function $k(y,x)$ is usually known from simulations generated from $f(x)^{\text{model}}$. Finding the solution of the inverse of Eq. 1.52 is not trivial. Thankfully, real analysis relies on a finite number of bins. Events are detected and classified depending on their properties in bins of some kinematic variables, hence Eq. 1.52 can be discretised and a solution can be found. The true distribution $f(x)$ is replaced by a vector $\mathbf{x}$ of dimension n with entries corresponding to the expected number of events in a bin j at *true level*; the measured distribution $g(y)$ is replaced by a vector $\mathbf{y}$ of dimension m with entries corresponding to the number of measured events at *detector level*; the kernel function is replaced by a matrix $\boldsymbol{\mathcal{E}}$ of dimension $m \times n$:

$$\boldsymbol{\mathcal{E}}\mathbf{x} + \mathbf{b} = \mathbf{y} \tag{1.53}$$

where **b** represents the background. The matrix $\boldsymbol{\mathcal{E}}$ usually assumes various names, such as *unfolding matrix*, *response matrix*, or *efficiency matrix*. This matrix is the core of the unfolding. It encodes the probability for an event observed in the i-th bin at detector level to be originated from the j-th bin at true level:

$$\varepsilon_{ij} = P(\text{observed in bin } i \mid \text{true value in bin } j) \tag{1.54}$$

Summing over all bins i gives the reconstruction efficiency of the j-th bin: $\sum_i \varepsilon_{ij} = \varepsilon_j$, which is generally different from one since some events can go undetected. The result of the unfolding procedure is an estimator $\hat{\mathbf{x}}$ of the true distribution $\mathbf{x}$ and the corresponding covariance matrix $C_{ij} = \text{cov}[\hat{x}_i, \hat{x}_j]$. The unfolding can proceed through very different algorithms, some of them are summarised in Ref. [44].

One of the main challenges of the unfolding is when the response matrix is ill-conditioned. If the matrix $\boldsymbol{\mathcal{E}}$ allows for significant migration of events between bins, then the estimators will be very sensitive to the Poisson fluctuation in $\mathbf{y}$, leading to large variances of these estimators, sometimes to the point that $\hat{\mathbf{x}}$ bear no resemblance to the true $\mathbf{x}$. Suppose the number of bins in the true space exceeds the number of bins in the detector phase space ($n > m$). In that case, the problem is ill-posed by definition since there will be an infinite number of solutions. It is thus good practice to set $n = m$. This is often not enough to guarantee the stability of the unfolding procedure. The mathematical figure of merit used to determine how ill-conditioned the response matrix is the *condition number*. The condition number of a matrix is a measure of how sensitive the solution of linear equations is to changes in the input data. A higher condition number indicates a greater sensitivity to perturbations in the inputs, which can lead to large errors or instability in the solution. Conversely, a lower condition number implies more robustness in solving linear equations. In practical terms, when the condition number of a matrix is large, it suggests that small changes in the input data or coefficients of the system can lead to significant variations in the solution. On the other hand, when the condition number is small, the system is considered well-conditioned, indicating that it is more stable and less prone to large errors. If the resolution effects are small and $\boldsymbol{\mathcal{E}}$ is almost diagonal, the condition number results to be small. If this is not the case, some forms of regularisation are necessary to damp down these oscillations and to produce a smoother solution. Two of the most common methods are based on the Tikhonov regularisation [45, 46] and the D'Agostini iteration with early stopping [47]. The computation of the condition number is given by $\text{cond}(\boldsymbol{\mathcal{E}}) = s_{\max}/s_{\min}$, where $s_{\max}$ and $s_{\min}$ represent the largest and the smallest singular values of $\boldsymbol{\mathcal{E}}$. The singular value decomposition will be discussed in Sec. 7.4.1.

Signal strength modifier

The maximum amount of information, at the price of maximal model dependence, can be extracted by measuring signal strength modifiers μ. In order to measure signal strength modifiers, the number of events that an analysis measures in a given category N can be parametrised as:

$$N = \mu \cdot \sum_{ij} [(\epsilon \cdot A^{\text{SM}})_{ij} \cdot \sigma_i^{\text{SM}} \cdot \text{BR}_j^{\text{SM}} \cdot \mathcal{L}] + b \tag{1.55}$$

where the index i spans over the production modes, the index j over the decay channels, and $\mathcal{L}$ is the integrated luminosity (cfr Sec. 2.1.1). The fitted parameter of interest is the signal strength modifier μ. The definition of cross section implies that the product $\sigma_i^{\text{SM}} \cdot \text{BR}_{\text{j}}^{\text{SM}} \cdot \mathcal{L}$ gives the number of events detected by an ideal experiment. This number should be then adjusted by the effects of the real world: ϵ is the reconstruction efficiency, A is the geometrical acceptance, and b is the background contamination. This represents the SM expectation that is then scaled by the signal strength modifier μ. Naively, the

measurement of signal strength modifier can be regarded as a check of the agreement between the observed signal yields and the SM predictions. This strong reliance on the prediction of the theory reduces the model independence and reinterpretability of the extracted result. The μ parameter can also be split to measure the signal strength in production and decay or for production mode.

Simplified Template Cross Sections

The signal strength modifiers were initially used as one of the standard approaches following the discovery of the Higgs boson, aimed at characterizing its profile. However, this method has a strong model dependence. The Simplified Template Cross Section (STXS) framework was developed to provide a natural extension, finding a compromise between model independence and experimental sensitivity.

Examining Eq. 1.55, the theoretical uncertainties enter in two places when measuring signal strength modifiers. The first one is related to the SM cross section predictions σ^{SM}, while the second is related to the experimental acceptance A^{SM}. The STXS approach, in contrast, directly measures physical cross sections, rather than signal strengths, thereby moving the former source of theoretical uncertainty to the interpretation stage, making the measurements more long-term useful. The uncertainties related to the event kinematics, which influence the analysis acceptance, are still taken into account in the unfolding of the experimental event categories to the STXS bins. The various $A^{\mathrm{SM}}_{\mathrm{bin}}$ only depend on the SM kinematics inside a given bin and serve as kinematic templates - hence the name of the STXS. It is worth noting that, in principle, other models besides the SM can be utilized as templates. From these considerations, Eq. 1.55 is modified as follows:

$$N = \sum_{\mathrm{bin}} [\sigma_{\mathrm{bin}} \cdot (\epsilon \cdot A^{\mathrm{SM}})_{\mathrm{bin}} \cdot \mathrm{BR}^{\mathrm{SM}} \cdot \mathcal{L}] + b \tag{1.56}$$

where σ_{bin} is the parameter of interest in the fitting procedure. In the STXS framework, cross sections are measured in mutually exclusive phase space regions known as *STXS bins*, which are defined centrally by the LHCHWG[8]. In the STXS, the measured analysis categories are unfolded to the STXS bins, which are common for all analyses, allowing for a subsequent global combination of different decay channels (the STXS is inclusive in the various decay modes) as well as measurements from ATLAS and CMS. The STXS bins are defined for specific production modes. This separation into production modes is essential to reduce the model dependence, i.e., to eliminate the dependence of the measurements on the relative fractions of the production modes in the SM. The choice of the bin boundaries also targets the reduction of the residual theory uncertainty due to the experimental acceptance for each STXS bin. Moreover, the STXS bins are crafted to isolate BSM effects, typically occurring at large kinematic scales where they could potentially manifest distinctly from the SM background. This separation not only enhances the sensitivity to BSM physics but also minimizes reliance on assumptions about the SM kinematic distributions.

Fiducial cross sections

The maximal model independence is ensured by measuring fiducial cross sections. This approach is based on removing detector effects by unfolding data to a carefully defined fiducial phase space. The restriction of the unfolding to this fiducial region, which closely matches the detector level selections, avoids large extrapolations that would introduce model dependence in the final results. Fiducial cross sections are usually measured differentially in bins of some kinematic variables. The number of events in a kinematic bin j at

[8]The LHC Higgs Working Group (LHCHWG) provides recommendations on theoretical predictions and methodologies on Higgs boson(s) studies to the LHC experiments.

detector level can be expressed as:

$$N_j = \sum_i (\sigma^i_{\text{fid}} \cdot \varepsilon_{ij} \cdot \mathcal{L}) + b^{\text{out}}_j + b_j \tag{1.57}$$

where σ^i_{fid} is the fiducial cross section, which already accounts for the BR and the acceptance effects, and $\epsilon_{i,j}$ is the unfolding matrix. A pictorial representation of the fiducial approach is depicted in Fig. 1.5. The diagonal elements of the response matrix represent the number of events generated in the j-th bin at fiducial level and successfully reconstructed in the corresponding bin at detector level, while the off-diagonal elements represent the bin migrations. In the fiducial measurements, a typical background source is given by reconstructed events generated outside the fiducial region b^{out}_j (non-fiducial background). A proper definition of the fiducial phase space ensures a small contribution of this term and it is usually set to the SM expectations.

In contrast to the STXS approach, fiducial cross sections are almost completely insensitive to the production mode. This translates into a maximal sensitivity to the ggH production mechanisms. The sensitivity to the other production modes can be enhanced by exploiting the binning in kinematic observables or the definition of the phase space.

Measuring fiducial cross sections also maximise the results re-interpretability and longevity. The fiducial phase space should be defined by means of simple cut-based selections that can be easily replicated. The comparison with theoretical predictions, either SM or BSM, is obtained by correcting the theoretical cross section by the fiducial acceptance, which denotes the fraction of signal events at true level falling into the fiducial phase space. A key advantage is the possibility of reinterpreting data with models that may not yet have been developed at the time of the measurement.

1.2.3 Higgs boson trilinear self-coupling

The Higgs boson self-couplings originate from the Higgs potential, represented by Eq. 1.39, which can be expressed as:

$$V(H) = \frac{m^2_{\text{H}}}{2} H^2 + \lambda_3 v H^3 + \lambda_4 H^4 \tag{1.58}$$

where $m^2_{\text{H}} = 2\lambda v^2$, $\lambda^{\text{SM}}_3 = \lambda$, and $\lambda^{\text{SM}}_4 = \lambda/4$. The value of λ is completely determined in the SM once the values of v and m_{H} are known:

$$\lambda = \frac{m^2_{\text{H}}}{2v^2} \tag{1.59}$$

The vacuum expectation value v can be computed from the Fermi constant G_F as $v = (\sqrt{2}\text{G}_\text{F})^{-1/2} \simeq 246$ GeV. The Fermi constant is measured to 0.6 part-per-million precision from the muon lifetime [48]. Once the Higgs boson mass is measured, the Higgs potential is then fully defined. However, several BSM models predict a modification of both the trilinear and quadrilinear self-couplings (cfr Sec. 1.2.1), hence the need for direct access to these couplings. While the direct probe of λ_4 is out of reach at the current and upcoming accelerators, the estimation of λ_3 is one of the main long-term goals of the physics programme at the LHC. The most natural approach to probe this parameter is to exploit the double-Higgs (HH) production, where the cross section depends on λ_3 at Leading Order (LO). Nevertheless, the HH production poses several challenges. Its cross section is predicted to be significantly smaller than that of single-Higgs production by a factor of 1000. The most stringent constraint on λ_3 from the CMS Collaboration comes from the combination of the most sensitive HH decay channels [41], and it is ascertained at 95% CL that is in the range $-1.24\ \lambda^{\text{SM}}_3 < \lambda_3 < 6.49\ \lambda^{\text{SM}}_3$.

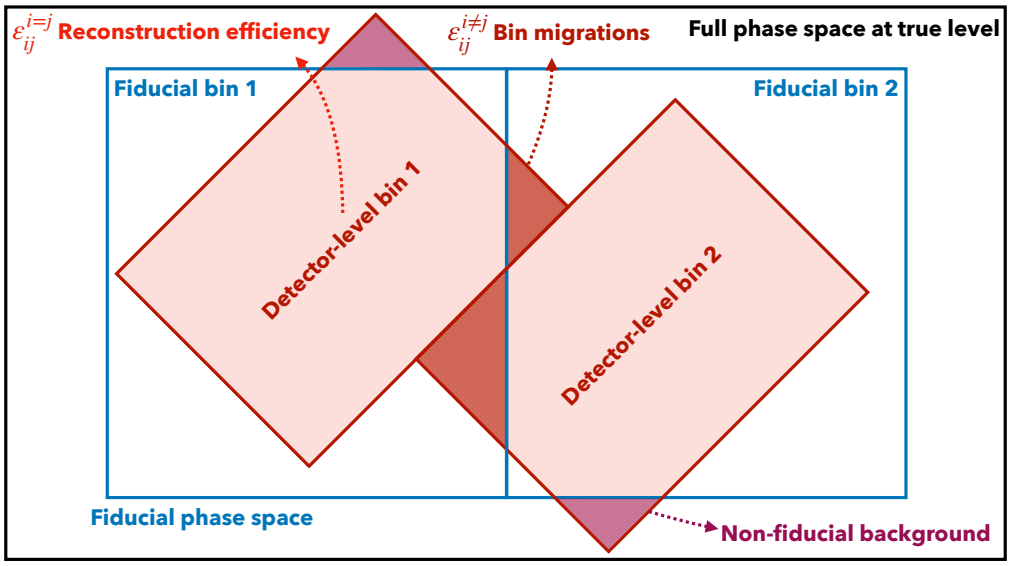

Figure 1.5: Pictorial representation of the fiducial cross section approach. The fiducial phase space is defined as a subset of the full phase space at true level, defined to match as closely as possible the detector level selections. In this representation, differences are emphasised to enhance readability. Events generated, for instance, in "fiducial bin 1" and reconstructed in "detector-level bin 1" are accounted for in the diagonal elements of the unfolding matrix. Conversely, events generated, e.g., in "fiducial bin 1" and reconstructed in "detector-level bin 2" are accounted for in the off-diagonal elements of the unfolding matrix, representing bin migrations. Due to the non-perfect alignment between fiducial and detector-level, some events outside the fiducial phase space can be reconstructed at detector-level. This last contribution is described by the non-fiducial background.

A complementary approach to probe the trilinear self-coupling is to exploit the single Higgs production [49–51]. Although single-Higgs production does not depend on λ_3 at Leading Order (LO) or higher orders in QCD, it does depend on λ_3 at Next-to-Leading Order (NLO) in the electroweak corrections. This approach can benefit from the larger cross section of the single-Higgs production and will be investigated in this thesis in Ch. 5.

1.2.4 Kappa framework

After the discovery of the Higgs boson, the central focus became the characterisation of the new resonance. The kappa (κ) framework [52] provides a way to explore the coupling structure of the Higgs boson based on three assumptions.

- The observed signal with a mass near 125 GeV originates from a single particle. The case of several, possibly overlapping, resonances in this mass region is not considered.

- The narrow-width approximation holds for this resonance, hence production and decay can be factorised for each channel as:

$$(\sigma \cdot \mathrm{BR})(i \to \mathrm{H} \to f) = \frac{\sigma_i \cdot \Gamma_f}{\Gamma_\mathrm{H}} \tag{1.60}$$

  where σ_i is the cross section through the production mode i, Γ_f is the partial decay width into the final state f, and Γ_H is the total width of the Higgs boson. The ratio $\Gamma_f/\Gamma_\mathrm{H}$ corresponds to the definition of BR.

- Only modifications to the coupling strengths are considered, while the tensor structure of the couplings is assumed to be the same as in the SM. This implies that the observed particle is considered CP-even as the SM Higgs boson.

The κ framework is based on introducing scaling factors - the κ factors - to scale the production and decay process at LO. Equation 1.60 is modified as follows:

$$(\sigma \cdot \mathrm{BR})(i \to \mathrm{H} \to f) = \frac{\sigma_i^{\mathrm{SM}} \kappa_i^2 \cdot \Gamma_f^{\mathrm{SM}} \kappa_f^2}{\Gamma_{\mathrm{H}}^{\mathrm{SM}} \kappa_{\mathrm{H}}^2} \quad \to \quad \mu_i^f \equiv \frac{\sigma \cdot \mathrm{BR}}{\sigma_{\mathrm{SM}} \cdot \mathrm{BR}_{\mathrm{SM}}} = \frac{\kappa_i^2 \cdot \kappa_f^2}{\kappa_{\mathrm{H}}^2} \tag{1.61}$$

The SM scenario is recovered by setting $k_{i,f} = 1 \ \forall \ i, f$. The expression for κ_{H}^2 adjusts the SM Higgs boson width to take into account κ_i modifications of the SM Higgs boson coupling strengths:

$$\kappa_{\mathrm{H}}^2 = \sum_j \frac{\kappa_j^2 \Gamma_j^{\mathrm{SM}}}{\Gamma_{\mathrm{H}}^{\mathrm{SM}}} \tag{1.62}$$

where the index j runs over all the decay modes. The scaling defined in Eq. 1.61 is valid only at LO because SM couplings are rescaled without accounting for higher-order contributions to these couplings. Nonetheless, NLO QCD corrections factorise with respect to coupling rescaling. These corrections are incorporated wherever possible to ensure that any deviations from the SM Higgs boson hypothesis are not artificially caused by ignored NLO corrections. For specific cases where the relation between the κ scale factor and cross sections (or decay widths) is non-trivial, dedicated computations have been carried out [52, 53]. This category includes loop-induced processes, such as ggH or $\mathrm{H} \to \gamma\gamma$, where the framework offers the flexibility to either resolve the loop or use an effective coupling strength parameter. For instance, in the case of the ggH cross section, it can be rescaled by a unique $\kappa_g^2 = \sigma_{\mathrm{ggH}}/\sigma_{\mathrm{ggH}}^{\mathrm{SM}}$ parameter or through a more complex expression involving the individual contributions from top quark (tt), bottom quark (bb), and top-bottom interference (tb):

$$\kappa_g^2(\kappa_{\mathrm{b}}, \kappa_{\mathrm{t}}, m_{\mathrm{H}}) = \frac{\kappa_{\mathrm{t}}^2 \sigma_{\mathrm{ggH}}^{\mathrm{tt}}(m_{\mathrm{H}}) + \kappa_{\mathrm{b}}^2 \sigma_{\mathrm{ggH}}^{\mathrm{bb}}(m_{\mathrm{H}}) + \kappa_{\mathrm{t}}\kappa_{\mathrm{b}} \sigma_{\mathrm{ggH}}^{\mathrm{tb}}(m_{\mathrm{H}})}{\sigma_{\mathrm{ggH}}^{\mathrm{tt}}(m_{\mathrm{H}}) + \sigma_{\mathrm{ggH}}^{\mathrm{bb}}(m_{\mathrm{H}}) + \sigma_{\mathrm{ggH}}^{\mathrm{tb}}(m_{\mathrm{H}})} \tag{1.63}$$

The κ framework thus provides a simple parametrisation to probe the Higgs boson couplings in a model-independent way. For this reason, it has been extensively used by the high-energy physics community. It compares the experimental measurements to their best SM predictions and does not require any BSM computations. It fully captures the leading effects in single Higgs processes, and the constraints on the κ parameters can be readily used to constraint parameters of specific BSM models [54]. The κ framework will be used in this thesis to explore the Higgs boson couplings to bottom and charm quarks in Ch. 5.

1.3 Experimental status of Higgs boson physics

Since the Higgs boson mass is a free parameter of the SM, its discovery was particularly challenging as there was not a clear indication of where to search for it. Major search efforts were carried out at the LEP (cfr introduction Ch. 2) and Tevatron colliders. The LEP search for the Higgs boson provided a lower bound to the the mass of 114.1 GeV at the 95% confidence level [55]. In hindsight, LEP could not have achieved the energy required to discover the resonance that would later be discovered. In contrast, the Tevatron was kinematically capable of *seeing* it, but the discovery was hindered by limited statistics. The breakthrough finally occurred after 60 years its postulation, the Higgs boson was discovered by the CMS and ATLAS collaborations at CERN in 2012 [19, 20]. Actually,

as also stated in the titles of the discovery papers, it was not found the Higgs boson, but simply a new boson. Starting from the discovery, an immense effort has been underway to understand whether the particle discovered in 2012 is indeed the SM Higgs boson.

The five leading production modes have now been observed (ggH, VBFH, VH, and ttH) as well as the $H \to \gamma\gamma$, $H \to ZZ$, $H \to WW$, $H \to bb$, $H \to \tau\tau$ decay channels. Measurements of the Higgs boson couplings to gauge bosons have reached a precision of 6-8%, while those to the third-generation fermions are at 7-11% precision. The spin-parity properties were extensively studied right after the discovery, proving its compatibility with the $J^{CP} = 0^{++}$ hypothesis and ruling out various alternative scenarios at more than 99% confidence level [56]. The mass has been measured with exceptional precision by both CMS and ATLAS, the former having as best result the value of $m_H = 125.08 \pm 0.12$ GeV [57] using the $H \to ZZ$ channel, and the latter the value of $m_H = 125.11 \pm 0.11$ GeV [58] from the combination of the $H \to ZZ$ and $H \to \gamma\gamma$. One of the latest remarks of the CMS collaboration has been the first evidence of the Higgs boson off-shell contribution, where the Higgs boson mass is far away from its nominal value, and the first measurement of its width to $\Gamma_H = 3.2^{+2.4}_{-1.7}$ GeV [59].

Alongside the study of the single Higgs boson production, extensive efforts are ongoing for the observation of the double Higgs production, which will give direct access to the shape of the Higgs potential. The double-Higgs cross section is now found to be less than $3.4 \times \sigma^{SM}_{HH}$ from the CMS collaboration [41] and less than $2.4 \times \sigma^{SM}_{HH}$ from the ATLAS collaboration [60]. The most sensitive channels for these searches are represented by final states where at least one of the Higgs bosons decays to $b\bar{b}$, to benefit from the sizeable $BR(H \to b\bar{b}) \simeq 58\%$, with the second Higgs boson decaying to final states assuring a trade-off between statistics and background contamination. These *silver bullets* are $HH \to b\bar{b}b\bar{b}$, $HH \to b\bar{b}\gamma\gamma$, and $HH \to b\bar{b}\tau^+\tau^-$.

Additionally, the exploration of the Higgs sector is now focusing on searches for rare decays. The combination of CMS and ATLAS data enabled the first evidence of the $H \to Z\gamma$ channel [61]. This search is extended with the so-called Dalitz Higgs boson decay $H \to \gamma^*\gamma \to \ell^+\ell^-\gamma$ in the low mass γ^* range of approximately 20-30 GeV, where the decay though γ^* completely dominates. The first evidence of this process was found by the ATLAS collaboration [62].

Ever-increasing datasets and the advancements in analysis techniques have led to promising signs in exploring Higgs boson decays to second-generation fermions. The CMS collaboration presented the first evidence for $H \to \mu\mu$ [63], whereas the $H \to cc$ channel poses some difficulties due to the overwhelming background. Currently, this search is carried out by both CMS and ATLAS looking at the VH associated production. In the search for this process, the first direct constraints of the Higgs-charm coupling are also now possible [64, 65].

As outlined in Sec. 1.2.1, the lepton flavour violating Higgs boson decays are suppressed in the SM but could be enhanced in many BSM theories such as the 2HDM. However, the most recent results from both ATLAS and CMS in the $H \to e\tau$, $H \to \mu\tau$, and $H \to e\mu$ channels do not show any significant excess [66–69].

Extensive searches have been conducted to explore the potential of the Higgs boson as a gateway to Beyond the Standard Model (BSM) physics, investigating decay channels beyond those expected in the Standard Model (SM). Of particular interest are the Higgs boson decays into stable particles, such as candidates for dark matter, that escape the detector without detection. Detecting these so-called invisible decays requires the Higgs boson to be produced alongside other particles. Therefore, the search should be carried out relying on associated productions, mainly VBFH and VH, with large amount of missing energy. These searches can also be extended to other exotic particles coupling with the Higgs boson and leaving signs of their passage in the detector. Numerous analyses have

been undertaken to set constraints on various processes, yet as of now, no definitive signs or evidence have been found.

Chapter 2

The CMS experiment at the LHC

The European Organisation for Nuclear Research, known as CERN (*Conseil Européen pour la Recherche Nucléaire*) is the largest and most important particle physics laboratory in the world. As reported in its founding convention [70]:

> *The Organization shall provide for collaboration among European States in nuclear research of a pure scientific and fundamental character, and in research essentially related thereto. The Organization shall have no concern with work for military requirements and the results of its experimental and theoretical work shall be published or otherwise made generally available.*

The foundation dates back to 1954 in the northern suburbs of Geneva, on the border between France and Switzerland. Since then, the CERN has been an absolute protagonist in the particle physics landscape, making several groundbreaking discoveries and pursuing cutting-edge research. The CERN's first accelerator was the 600-MeV SynchroCyclotron (SC) started up in 1957. It was built to observe and study the rare decay of pions into an electron and a neutrino. The discovery was made with the dataset accumulated in only a few hours [71]. The SC focused on nuclear physics in the following years, leaving particle physics to the CERN's second accelerator, the Proton Synchrotron (PS). The PS started up on 24 November 1959, and in 1965, it allowed the first observation of antinuclei [72]. An epoch-making event happened on 27 January 1971 when the PS was used to feed the Intersecting Storage Rings (ISR), the world's first hadron collider. For the first time in history, two proton beams were accelerated to collide head-on. The ISR was crucial to gaining valuable knowledge and expertise for subsequent colliding-beam projects. In 1971, the CERN's council approved the construction of a new accelerator, the Super Proton Synchrotron (SPS). With a circumference of seven kilometres, the SPS became the CERN's first accelerator to cross the Franco-Swiss border. Initially functioning as a proton-proton collider, in 1979, CERN decided to convert it into a proton-antiproton collider. The world's first proton-antiproton collision occurred on 4 April 1981 at the ISR, paving the path to the activities in the SPS. The proton-antiproton collisions in the SPS started immediately after, and in 1983, CERN announced the pivotal discovery of the mediators of the weak force, the W and Z bosons. Carlo Rubbia and Simon van der Meer received the Nobel Prize in physics in 1984 *for their decisive contributions to the large project, which led to the discovery of the field particles W and Z, communicators of weak interaction.* After discovering these two fundamental actors, the electroweak theory had to be studied in detail, and the last missing component, the Higgs boson, should be sought. For this purpose, in 1989, the first beam circulated in the new CERN accelerator, the Large Electron Positron (LEP) collider, with its four experiments, ALEPH, DELPHI, L3, and OPAL. With its 27-kilometre circumference, LEP was the largest accelerator ever built in history. It operated for 11 years, producing a large set of valuable results [73, 74],

until its final shutdown at 8 am on 2 November 2000. Following the decommissioning of LEP facilities, the existing tunnel and caverns became available to host CERN's current main accelerator, the LHC, with its four main experiments: A Toroidal LHC Apparatus (ATLAS), A Large Ion Collider Experiment (ALICE), the Compact Muon Solenoid (CMS), and the LHC beauty (LHCb).

This chapter begins with a description of the LHC in Sec. 2.1. It then delves into the CMS experiment in Sec. 2.2, offering insights into the detector's structure with all its sub-detectors and components. The offline reconstruction is discussed in Sec. 2.3, with a particular focus on the physics objects that will be utilised in the analysis presented in the next part of this thesis. As the CMS experiment approaches the conclusion of its first phase of data-taking and prepares for the next one, Sec 2.4 presents a general account of the major upgrades being undertaken. A more detailed description is reserved for the new endcap calorimeter in Sec 2.4.1, which will be the object of the last part of this thesis. Finally, the chapter concludes with an introduction to the Monte Carlo simulation within the CMS collaboration, discussed in Sec 2.5.

2.1 The Large Hadron Collider

The LHC is the world's largest and most powerful accelerator. It was built inside the 27-kilometre circular tunnel used for the LEP accelerator. The main goals of the LHC physics programme were to study the electroweak symmetry breaking via the search of the Higgs boson, observed by the CMS and ATLAS collaborations in 2012 [19, 20], and exploring the TeV scale. The LHC is a proton-proton collider, however, for part of each year, the LHC provides collisions between lead ions to create quark-gluon plasma, which is a state of matter that is believed to have existed in the very early universe, just moments after the Big Bang.

2.1.1 Design

The LHC [75] is one of the several accelerators currently in operation at CERN. The CERN accelerator complex, illustrated in Fig. 2.1, constitutes a sophisticated acceleration chain, where each machine sequentially boosts a beam of particles to a specific energy before transferring it to the subsequent machine in the chain. The life of proton beams for the LHC starts from a bottle of hydrogen gas, serving as a source of negative hydrogen ions. The initial step in the accelerator chain is the Linac4, responsible for accelerating the ions up to 160 MeV before entering the PS Booster. The ions undergo a stripping process during injection from the Linac4 into the PS Booster, leaving only the desired protons. Then, the PS Booster further accelerates these protons to an energy of 2 GeV. The protons advance along their path to the PS, where their energy increases to 26 GeV. The SPS accomplishes the final acceleration stage, which operates at up to 450 GeV. Ultimately, the beams are injected into the LHC, where they reach their final operational energy. The reason for having an acceleration chain, instead of directly injecting protons into the LHC and accelerating them to the desired energy, is driven by engineering and practical limitations. Once the proton beams attain the desired energy inside the LHC, they are ready for collisions in correspondence with the *Fab Four*. Four experiments are located along the LHC circumference: ATLAS, ALICE, CMS, and LHCb. ATLAS and CMS are the two multi-purpose experiments with a broad physics programme, while LHCb and ALICE are more specialised in flavour physics and heavy-ion physics, respectively.

The two proton beams inside the LHC travel in opposite directions in separate beam pipes kept at ultra-high vacuum. Thousands of superconducting NbTi magnets are placed

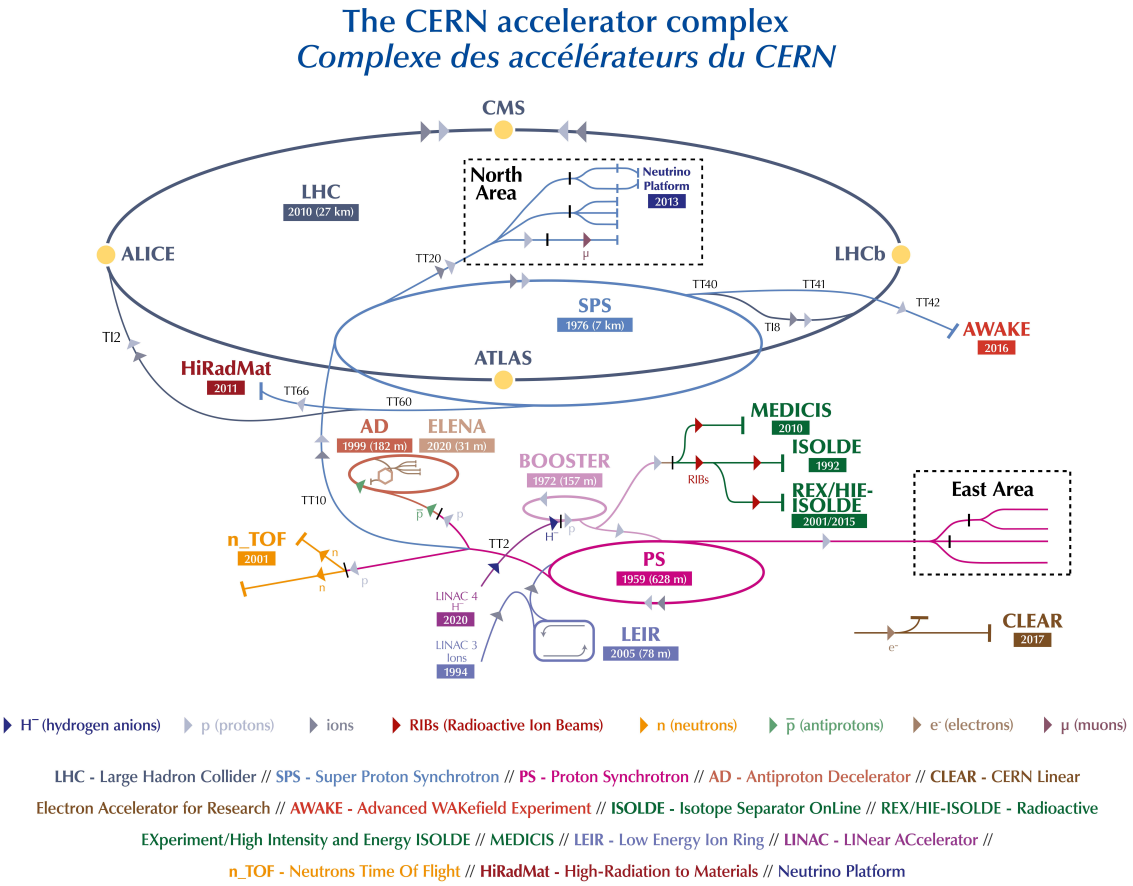

Figure 2.1: Representation of the CERN accelerator complex.

along the collider to guide the particle beams. These include 1232 dipole magnets, which bend the beam trajectory, and 392 quadrupole magnets, which focus the beams ensuring a narrow size of the bunches. When beams arrive at the interaction point, the insertion magnets, made by a combination of three quadrupoles, squeeze the beams closer together to increase the probability of interaction. The number of events per unit of time generated in a collider is given by:

$$\frac{\partial N}{\partial t} = L\sigma \tag{2.1}$$

where σ is the cross section (times the BR) of the physics process and L is the instantaneous luminosity. While the former is a parameter related to the physics under study, the latter is a parameter of the machine. For the LHC, the design value is $L = 10^{34}$ cm^{-2}s^{-1}. The time integral of Eq. 2.1 introduces an important parameter qualifying a period of data taking, the *integrated luminosity* L_{int}:

$$N = L_{\text{int}}\sigma \tag{2.2}$$

The luminosity depends only on the beam parameters and, for a Gaussian beam distribution, can be written as:

$$L = \frac{N_b^2 n_b f_{\text{rev}} \gamma_r}{4\pi\epsilon_n \beta^*} F \tag{2.3}$$

where N_b is the number of particles per bunch, n_b the number of bunches present in the beam, f_{rev} is the beam revolution frequency, γ_r is the corresponding relativistic gamma factor, and the product $\epsilon_n\beta^*$ is related to the beam optics. The F term is introduced to account for the reduction of luminosity due to the crossing angle at the interaction point.

The exploration of rare events in the LHC thus requires a high luminosity. However, this inevitably leads to the occurrence of multiple interactions for bunch crossing. This phenomenon is referred to as *PileUp* (PU). The average number of collisions per bunch crossing is given by:

$$< \mu >= \frac{L\sigma_{pp}^{\text{inelastic}}}{n_b f_{\text{rev}}} \tag{2.4}$$

where $\sigma_{pp}^{\text{inelastic}}$ is the inelastic proton-proton cross section. Figure 2.2 shows the recorded PU distributions during the different data-taking periods.

2.1.2 Operations

The first proton beam was successfully injected into the LHC at 10:28 am on 10 September 2008, marking the dawn of a new era in particle physics. However, there was an initial abrupt stop on 19 September 2008. During some tests, a faulty electrical connection between two magnets caused a large liquid helium loss. After one year of technical stop, where 53 magnets had to be refurbished or completely replaced, the first low-energy beams circulated again in the LHC in November 2009. After few days, the LHC set a new world record when the beam energy was ramped up to 1.18 TeV per beam, becoming the world's highest-energy particle accelerator, beating the Tevatron's previous record of 0.98 TeV. In 2010, energy constantly increased until reaching 3.5 TeV per beam, allowing the first proton-proton collision with a centre-of-mass energy ($\sqrt{s}$) of 7 TeV. This marked the beginning of the main physics programme of the LHC. The dataset collected in 2011 at $\sqrt{s} = 7$ TeV and in 2012 at $\sqrt{s} = 8$ TeV constitutes the so-called Run 1 of the LHC. This is the dataset used by the ATLAS and CMS collaboration for the discovery of the Higgs boson. The LHC was then shut down on 13 February 2013 for its first Long Shutdown (LS). After two years of maintenance and upgrades, the Run 2 of the LHC started at a beam energy of 6.5 TeV, corresponding to $\sqrt{s} = 13$ TeV, a new world record. During the Run 2, the LHC delivered 163.55 fb^{-1} of proton-proton data that have been used to carry out precision measurements, especially of the newly discovered Higgs boson. On 10 December 2018, the second LS of the LHC started. The LHC, its experiments, and the whole CERN accelerator complex were maintained and upgraded. After more than three years of LS, the Run 3 of the LHC started in 2022. One day after the 10th anniversary of the announcement of the Higgs boson discovery, 5 July 2022, the LHC delivered proton beams at $\sqrt{s} = 13.6$ TeV, setting another world record. This will be the operational energy for the ongoing run[1], which will last until 2025. Starting from the first data-taking periods, the LHC has continuously increased the instantaneous luminosity, going above its nominal value. This continuous growth led to a greater collection of data while, at the same time, a larger value of PU, as reported in Fig. 2.2. Run 3 will also mark the end of the LHC as we know it, closing its *phase-1*. After the third LS, where both the accelerator and the experiments will undergo a profound upgrade, the High-Luminosity LHC (HL-LHC) is expected to start in 2029. The HL-LHC project aims to boost the performance of the LHC in order to increase the potential for discoveries and allow precise measurements. The LHC foresees a peak instantaneous luminosity of 5×10^{34} cm^{-2}s^{-1} with the goal of integrating about 3000 fb^{-1} by the end of operations. The complete timeline of the LHC and HL-LHC projects is shown in Fig. 2.3.

2.2 The CMS detector

The CMS detector [76] is one of the four large experiments at the LHC. It is located at the interaction point number 5, near the French village of Cessy. Together with the ATLAS

[1]Thesis written in summer 2023

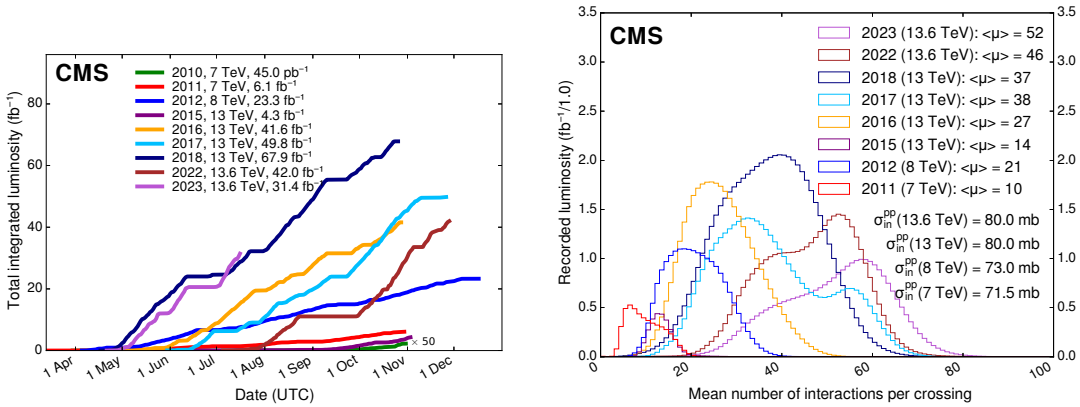

Figure 2.2: (Left) Delivered luminosity as a function of the time for all data-taking periods. (Right) Average number of interactions per bunch crossing for all data-taking periods. The overall mean values and the inelastic pp cross sections are also shown. For both plots, since the 2023 data taking is still ongoing at the time of writing of this manuscript, the related information is partial.

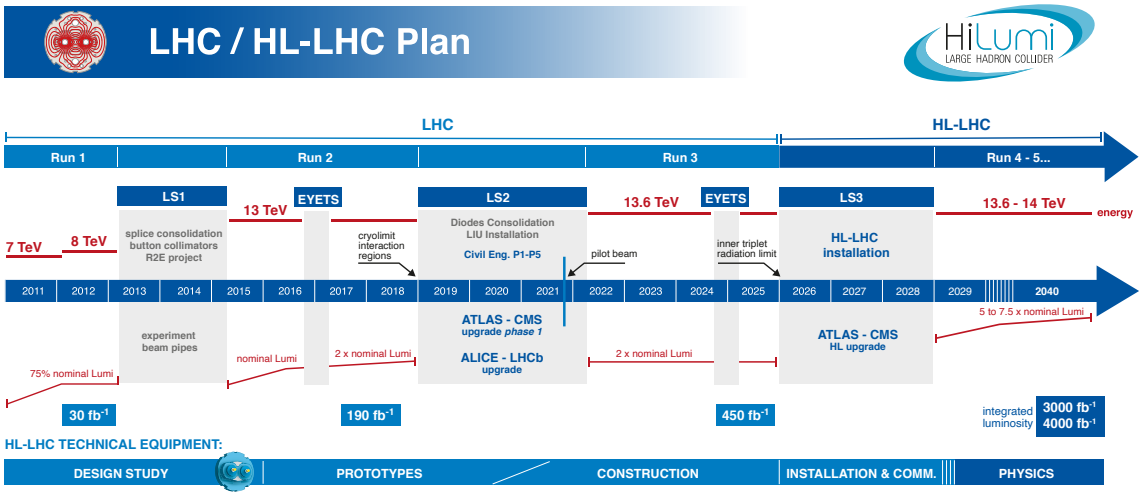

Figure 2.3: LHC and HL-LHC timeline.

detector, it is one of the two multi-purpose experiments at the LHC. The CMS detector was conceived with the primary objectives of discovering the Higgs boson, characterising its properties and those of the electroweak sector, searching for signs of physics BSM physics, and studying the high-energy heavy-ion collisions. The CMS detector's name itself reveals some distinctive features. Despite a weight of 14,000 tons, it is just 15 metres high and 21 metres long, earning the adjective of *compact*[2]. A strong emphasis is on *muons*, which serve as an excellent signature for interesting physics. Muons can be easily identified, and high-energy muons, unlike other particles, can only originate from the decay of heavy resonances. The beating heart of CMS is its 4 Tesla superconducting *solenoid* magnet, which is the most powerful ever built. Inside the solenoid and immersed in the large magnetic field, there are the tracker and the calorimeters, while the dedicated muon detection system is in the solenoid iron return yoke. This configuration allows an excellent performance of the inner tracker, due to the combination of a silicon-based tracker and a high magnetic field, while keeping the budget material in front of the calorimeters low. However, the trade-off is a relatively limited space for the hadronic calorimeter, which has

[2]The 7,000 t ATLAS detector is 25 metres high and 46 metres long

to fit within the solenoid. As described later in this section, the CMS collaboration has addressed the hadron calorimeter issue by placing an additional section of the hadronic calorimeter outside the solenoid and, more importantly, fully leveraging the particle-flow approach to boost its performance (cfr Sec. 2.3). After explaining the coordinate system in Sec. 2.2.1, the different sub-detectors will be detailed in Sec. 2.2.2.

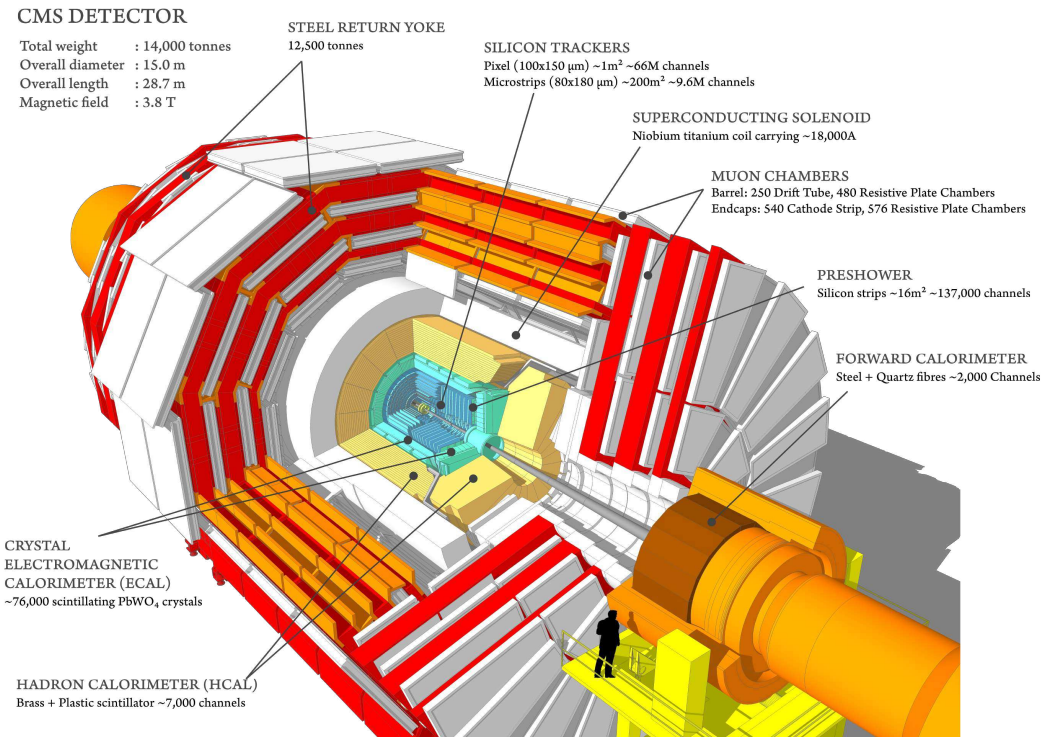

Figure 2.4: Pictorial representation of the CMS detector.

2.2.1 Coordinate system

The CMS experiment adopts a right-handed coordinate system, with the origin placed at the nominal collision point. The x-axis points toward the centre of the LHC ring, the y-axis points vertically upward, and the z-axis thus points towards the Jura mountains. The structure of the detector makes it natural to use a polar coordinate system. The azimuthal angle ϕ is measured in the $x-y$ plane from the x-axis, the radial coordinate is given by r, and the polar angle θ is measured from the z-axis in the $y-z$ plane. A schematic representation of the CMS coordinate system is shown in Fig. 2.5. The relation between the Cartesian and the CMS coordinates is given by:

$$\begin{cases} x = r \sin\theta \cos\phi \\ y = r \sin\theta \sin\phi \\ z = r \cos\theta \end{cases} \tag{2.5}$$

While this coordinate system is well suited for expressing the orientation of the detector and macroscopic observables, it is not ideal for describing proton-proton collisions. A proton-proton collision is actually a parton-parton collision. Protons, not being fundamental particles, are made of different constituents, named *partons*. Consider a scenario where two protons collide head-on. The four-momentum of the two partons can be written as:

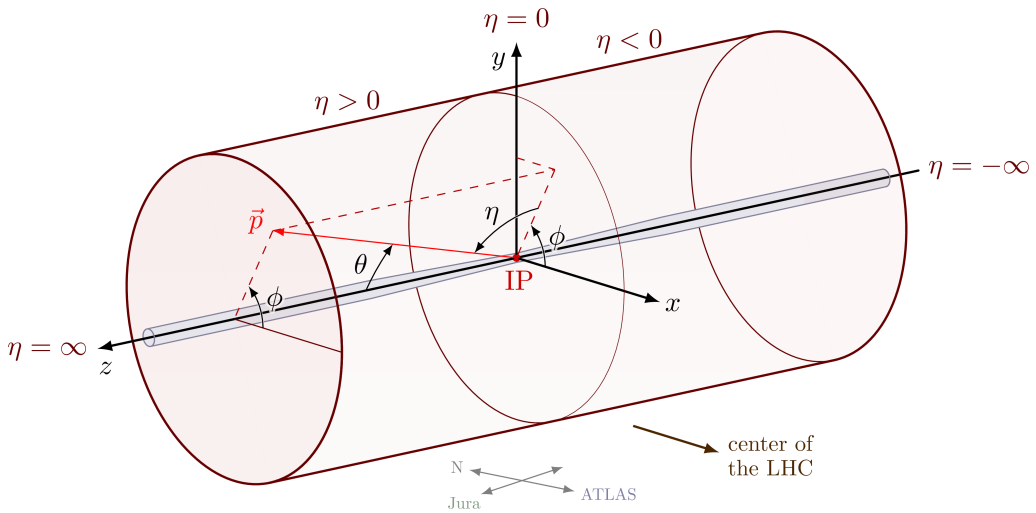

Figure 2.5: CMS coordinate system.

$$p_1^\mu = (x_1 E, 0, 0, x_1 p) \tag{2.6}$$
$$p_2^\mu = (x_2 E, 0, 0, -x_2 p) \tag{2.7}$$

In these four-vectors, E and p denote the energy and momentum of the proton, respectively. The x and y components are null as protons are primarily accelerated longitudinally at high energies, leading to negligible transverse momentum for the partons inside them. Consequently, the x_1 and x_2 variables can be interpreted as the fractions of energy that the partons carry away from the original protons. After the collision, the resulting momentum is given by:

$$(p_1 + p_2)^\mu = ((x_1 + x_2)E, 0, 0, (x_1 - x_2)p) \tag{2.8}$$

The CMS frame is not the centre-of-mass frame of the collision. The relativistic β factor, in the high-energy limit where $E \simeq p$, results to be:

$$\beta = \frac{x_1 - x_2}{x_1 + x_2} \tag{2.9}$$

The x_1 and x_2 quantities are unknown and change from event to event, consequently also the related Lorentz-boost is unknown[3]. It is thus beneficial to use variables that are Lorentz-invariant for boosts along the longitudinal direction. The *par excellence* observables are the transverse momentum p_T and transverse mass m_T:

$$p_\mathrm{T}^2 = p_x^2 + p_y^2 \tag{2.10}$$
$$m_\mathrm{T}^2 = m^2 + p_x^2 + p_y^2 = E^2 - p_z^2 \tag{2.11}$$

From these variables, the transverse energy is defined as:

$$E_\mathrm{T}^2 = m^2 + p_\mathrm{T}^2 \tag{2.12}$$

which is equal to the transverse momentum for massless particles. Another variable is the rapidity y:

$$y = \frac{1}{2} \ln\left(\frac{E + p_z}{E - p_z}\right) \tag{2.13}$$

[3] *A priori*, x_1 and x_2 could be known by measuring the momenta of all particles produced in the collision, but this is not feasible.

A Lorentz-boost along the z-axis shifts the rapidity by a constant term that depends on the boost itself. The rapidity itself is thus not Lorentz-invariant; its difference is. One of the advantages of using the rapidity is that particle production is roughly constant as a function of y. For ultra-relativistic particles, the rapidity is usually converted to the pseudorapidity η:

$$y \simeq \frac{1}{2}\ln\left[\frac{E(1+\cos\theta)}{E(1-\cos\theta)}\right] = -\frac{1}{2}\ln\left[\tan\left(\frac{\theta}{2}\right)\right] \equiv \eta \tag{2.14}$$

The pseudorapidity is a pure geometrical variable, depending only on the angle θ. It can be also used to define a Lorentz-invariant spatial separation between two particles:

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} \tag{2.15}$$

Based on the η coordinate, the detector is divided into a central part called *barrel*, and two forward parts called *endcaps*. The exact η value marking the transition between the two regions depends on the specific sub-detector.

2.2.2 Detector structure

The CMS experiment, like ATLAS, is a multipurpose detector, i.e., designed to be versatile and capable of effectively detecting a wide range of physics processes across a wide range of energies. The physics requirements and fundamental principles guiding the CMS design were outlined in the letter of intent [77] in 1992 when the experiment was officially proposed. At the time, the primary goal was the discovery of the Higgs boson, leading to constraints to meet this objective. Among the crucial channels for discovering a light Higgs boson with $m_{\mathrm{H}} < 150$ GeV were the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ and $\mathrm{H} \to \gamma\gamma$ channels. Both of these channels need an excellent, hermetic electromagnetic calorimeter to detect electrons and photons effectively. It was required a constant term of the energy resolution in the electromagnetic calorimeter equal or less than 0.5%. In addition, the four-lepton channel also needs a performing muon reconstruction. Muon reconstruction became a focal point of the new detector, entering the name itself. The experience gained at the SPS for discovering the W boson pointed to the necessity for a robust and redundant muon system. Although the limits from LEP and Tevatron were indicating a low-mass Higgs boson, the CMS detector was optimised for searches in a mass range from 90 GeV to 1 TeV. This flexibility allows the exploration of new physics beyond the mere Higgs boson using various signatures. At high-mass, it is vital to have a good muon momentum resolution, as muons can be used as signature of BSM resonances. The requirement was of a resolution less than 10% at 1 TeV and order of 1% or less at low momenta. Another fundamental requirement was a high-efficient central tracking system. The tracker plays a central role in momentum measurement, particle identification (linking a track to calorimeter hits to distinguish between an electron and a photon), and vertex reconstruction. Achieving a precise momentum measurement unavoidably requires a large bending power from the magnet, which led to the exploration of superconducting technology. The following sections show how the CMS detector successfully meets these requirements by showcasing the structure and components of the detector.

Superconducting solenoid magnet

The CMS magnet [78] is the central and pivotal component of the detector, playing a defining role in the design of the experiment. The CMS detector adopts a solenoidal magnetic field generated by a superconducting magnet made of Niobium-Titanium, with an internal diameter of 6 m and a length of 12.5 m. The CMS collaboration originally designed the magnet to operate at a nominal magnetic field strength of 4 Tesla. However,

to maintain a balance between pushing the boundaries of physics research and safeguarding the operational stability of the magnet, especially regarding the uncertainty related to the long-term ageing of the magnet, the operational field strength is conservatively set at 3.8 Tesla. The magnet is enclosed in a vacuum vessel made of two stainless steel cylinders to isolate it from the exterior, as the solenoid has to operate at liquid helium temperature of -269 °C. To control the magnetic field lines outside the solenoid, a *return yoke* is employed. The magnetic flux is returned via a 1.5 m thick saturated iron yoke placed in the detector's outer part, interspersed with the muon detection system. Hence, the muon chambers are immersed in the return field of 2 Tesla, which bends particles in the opposite direction compared to the particle travelling in the inner part. Figure 2.6 shows the CMS detector's magnetic field and field lines. In summary, the CMS magnet's robust and precisely engineered design, coupled with its ability to create a powerful magnetic field, significantly helps the momentum measurement capabilities, the ability to separate nearby particles, and enables the CMS experiment to conduct frontier research in particle physics.

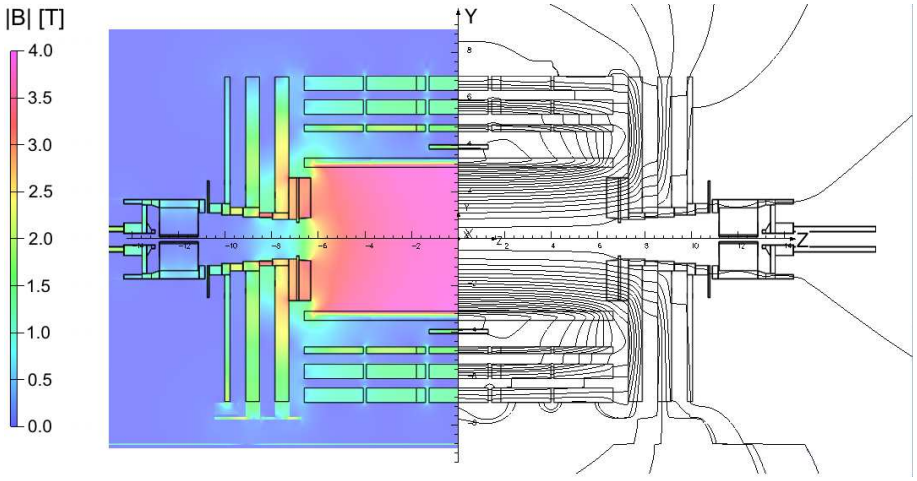

Figure 2.6: Longitudinal view of the CMS detector showing the predicted value of the magnetic field (left) and field lines (right), at a central magnetic flux density of 3.8 T. Each field line represents a magnetic flux increment of 6 Wb. Figure taken from [79].

Inner tracking system

The inner tracking system [80] is the closest component to the interaction point in the CMS detector. The purpose of the tracker is twofold. Firstly, in synergy with the large magnetic field produced by the solenoid, it enables the precise measurement of charged particles' momenta. As charged particles traverse the inner tracker, the strong magnetic field causes them to follow curved trajectories. The tracker, segmented into strips or pixels, captures detailed information from multiple layers as the particles pass through. By combining this multi-layer information, the tracker reconstructs the curved tracks of the particles, and the curvature of these tracks provides a precise estimation of the particle's momentum. The second purpose of the tracker is the reconstruction of vertices, both the primary proton-proton interactions and the secondary vertices produced by long-lived particle decays.

The CMS tracking system is entirely made of silicon detection elements and comprises two subcomponents: the pixels, at the very core of the detector dealing with the highest intensity of particles, and the microstrip detectors surrounding it. Both systems cover the region $|\eta| < 2.5$ and a volume of 5.6 m in length and 2.4 m in diameter.

The CMS pixel tracking detector comprises 65 million silicon pixels of 100×150 μm^2. The original pixel detector, mounted on CMS at the beginning of its operations, consisted of three barrel layers at radii 44, 73, and 102 mm, and two endcap disks on each end at distances of 345 and 465 mm from the interaction point. This configuration was designed for a maximum instantaneous luminosity of 1×10^{34} cm^{-2}s^{-1} and a maximum PU of 25. During the LS1 upgrade, the LHC underwent major upgrades, and it became available to provide more than double these values. In order to cope with the harsher experimental environment, the pixel detector was replaced during the extended year-end technical stop of the LHC between 2016 and 2017 [81]. The new CMS pixel detector is provided with four concentric barrel layers at radii 29, 68, 109, and 160 mm, and three disks on each endcap at distances of 291, 396, and 516 mm from the interaction point. Additionally, the implementation of advanced carbon-fibre material for the mechanical support, the use of a new lower mass CO_2 cooling system, and the displacement of electronic boards outside the tracking volume significantly reduced the material budget. Figure 2.7 shows the pixel tracker and compares the budget material before and after the upgrade.

The CMS strip tracking detector comprises silicon microstrip sensors and covers the radial region between 20 cm and 116 cm. It is subdivided into four components, as shown in Fig. 2.8. The Tracker Inner Barrel (TIB) and Tracker Inner Disks (TID) are made of 4 concentric layers and 3 disks, respectively, covering the area just outside the pixel tracker. The TIB and TID are surrounded by the Tracker Outer Barrel (TOB), which comprises 6 concentric layers forming a shell structure. All barrel components extend in z between $\pm$118 cm. Beyond this region, the Tracker EndCap (TEC) cover the region up to 282 cm, and each TEC comprises nine disks.

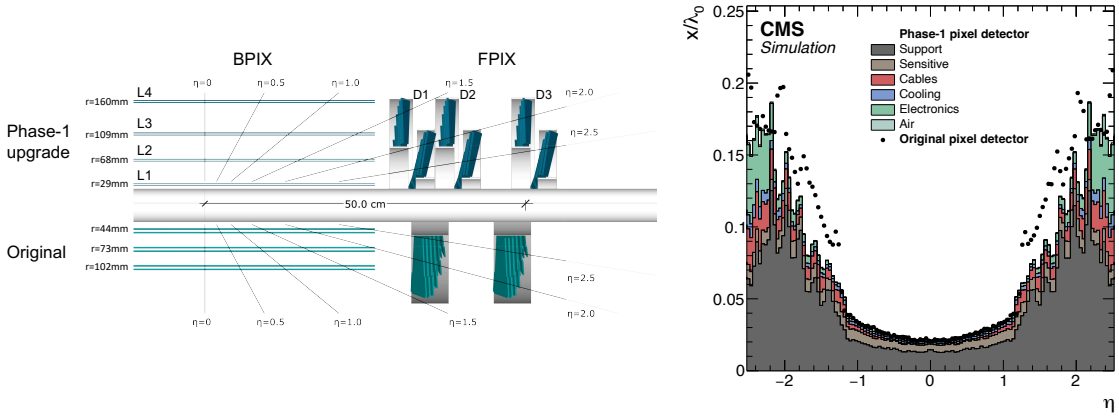

Figure 2.7: (Left) Longitudinal view of the original and new (Phase-1 upgrade) CMS pixel tracker system. BPIX and FPIX indicates the barrel and endcap components. (Right) Comparison of the material budget between the original pixel detector and the upgraded one, in units of hadronic interaction length λ_0, as a function of pseudorapidity η. Figures taken from [81].

Electromagnetic calorimeter

The CMS Electromagnetic CALorimeter (ECAL) [82] surrounds the tracker and is arranged in a central barrel section (EB), extending up to $|\eta| = 1.479$, and two endcap sections (EE), extending up to $|\eta| = 3$. It is a scintillating, hermetic, homogenous calorimeter made of around 75,848 crystals of lead tungstate ($PbWO_4$), a material that is 98% metal but completely transparent. This material features a high density of 8.28 g cm^{-3} and a

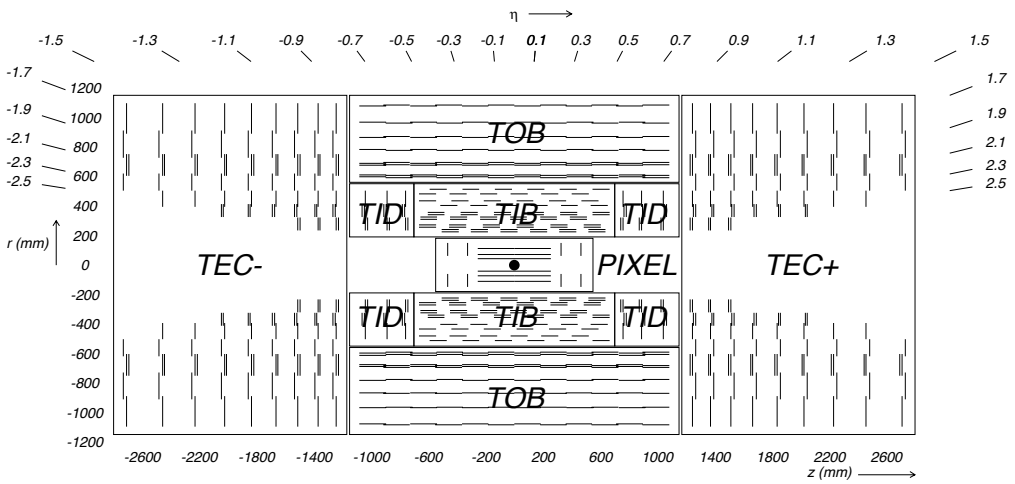

Figure 2.8: Cross section of the CMS tracker, showing both the pixel detector and the four components of the microstrip detector. Figures taken from [76].

high effective atomic number, resulting in a short radiation length[4] of $X_0 = 0.89$ cm and a small Molière radius[5] of $R_M = 2.19$ cm. These characteristics enable the complete containment of electromagnetic showers. In addition, the fast scintillation decay time allows for the collection of approximately 80% of the total light emission within 25 ns, which matches the LHC bunch crossing rate. In the EB, each crystal is 23 cm long, corresponding to a total radiation length of $25.8X_0$, while in the EE, each crystal is 22 cm long, corresponding to a total radiation length of $24.7X_0$. Being a homogenous calorimeter, the crystals both generate the electromagnetic shower and produce the scintillation light. Avalanche photodiodes, specifically designed to withstand the immense magnetic field and radiation levels, are attached to the back of each crystal to collect the scintillation light and convert it into an electric signal.

To enhance the spatial resolution, particularly for π^0 rejection, a preshower (ES) is situated in the endcaps before the EE in the $1.65 < |\eta| < 2.6$ region. The ES is a sampling calorimeter employing lead absorbers and silicon strip sensors. Its primary purpose is to measure the transverse profile of electromagnetic showers, distinguishing between single high-energy photons and close pairs of photons resulting from π^0 decays. The total radiation length of the ES is equal to $3X_0$. The cross-section of the CMS ECAL is shown in Fig. 2.9.

Hadronic calorimeter

The CMS Hadronic CALorimeter (HCAL) [84] is a sampling calorimeter employing alternating layers of brass absorbers and plastic scintillator tiles. The absorber thus generates the hadron shower, while the scintillating material produces the scintillation light, which is collected by optic fibres. The HCAL is composed of three sections: the barrel (HB) and endcap (HE) calorimeters located inside the magnet, and the outer (HO) and forward (HF) calorimeters placed outside the magnet coils.

The HB covers the $|\eta| < 1.392$ region and is radially constrained between the outer radius of the ECAL ($R = 1.77$ m) and the inner radius of the magnet ($R = 2.95$ m). The combined material budget of the HB and the preceding detector components results in a

[4]The radiation length X_0 is defined as the average distances that an electron needs to travel to reduce its energy to 1-e^{-1} $\simeq$ 63.1% of its original energy. It depends on the material of the calorimeter.

[5]The Molière radius is defined as the radius of a hypothetical cylinder around the shower's main axis in which 90% of the shower energy is deposited. It depends on the material of the calorimeter

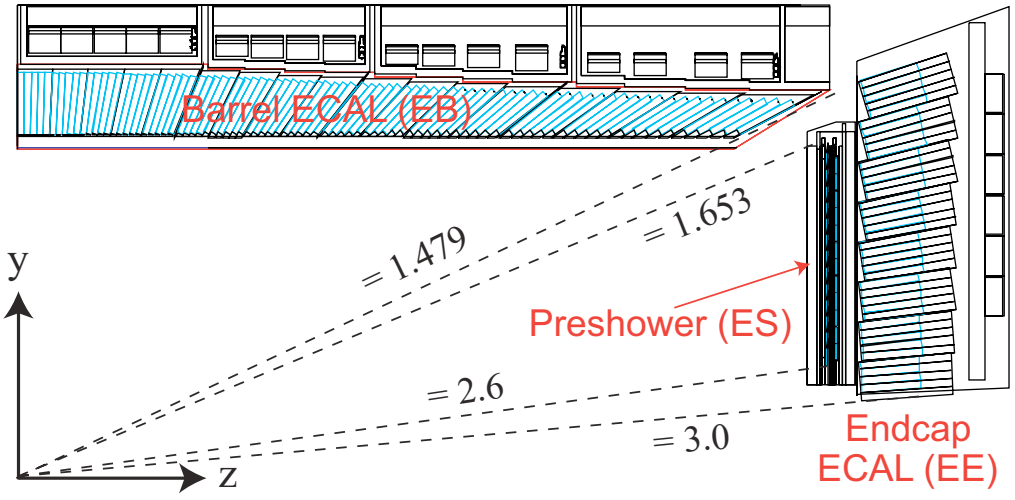

Figure 2.9: Cross section of the CMS electromagnetic calorimeter. Figure taken from [83].

total interaction length[6] of 7.8λ at $\eta = 0$. Extensive Monte Carlo studies and beam test analyses [85] proved that this radiation length is insufficient to ensure complete containment of hadronic showers. As a result, the HB leaks approximately 3% of 300 GeV-pions, leading to missing energy of up to 100 GeV. A few events with large lost energy leads to a significant degradation of the energy resolution that is hard to correct offline. To address this issue, the CMS collaboration implemented another scintillating layer outside the cryostat and within the magnetic flux return yoke, exploiting the magnet as an absorber. This additional component, the HO, acts as a tail-catcher to improve energy measurements. The HO is placed as the first sensitive layer in the iron yoke rings. As the region around $\eta = 0$ is the one with the lowest λ, the corresponding ring has two layers of scintillators placed on either side of a 19.5 cm thick piece of iron. All other rings have a single layer. The HO provides an additional calorimeter coverage of about 3λ and substantially enhances the energy resolution, as reported in Fig. 2.11.

On the other hand, the HE is placed in the endcaps of the detector and covers the $1.305 < |\eta| < 3.0$ region. The HE provides an average constant value of around 10λ, not requiring a solution similar to the HO in the barrel region. Completing the HCAL system are the two forward components, the HF, placed at the extreme ends of the CMS detector at $z = \pm 11.2$ m from the interaction point, providing coverage up to $|\eta| = 5.2$. Due to the small angle relative to the beam pipe, the radiation flux in the HF is significantly higher compared to other HCAL regions. To handle these conditions, the HF employs plastic-clad quartz fibres as active material due to their superior radiation hardness. The HF relies on the Cherenkov effect to produce scintillating light that is effectively collected by photomultiplier tubes placed just behind the absorber.

Muon chambers

The CMS muon tracker system is the outermost component of the detector, covering the $|\eta| < 2.4$ region. It is based on gaseous detectors sandwiched among the layers of the steel flux-return yoke. As a result, it is immersed in a 2 Tesla magnetic field, which causes muons to bend, allowing for a precise momentum measurement. The bending direction of particles in the muon tracker is opposite to that of the inner tracker. The decision to use gaseous

[6]The interaction length λ is defined as the mean free path of a hadron before it undergoes a nuclear interaction. It depends on the material of the calorimeter.

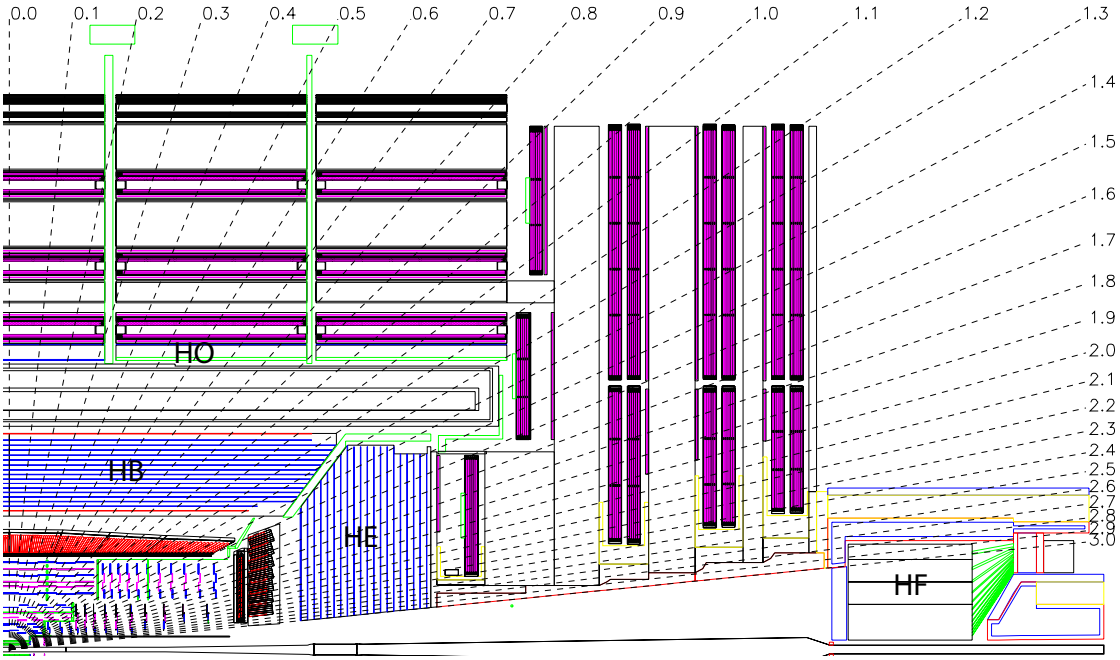

Figure 2.10: Cross section of the CMS hadronic calorimeter. Figure taken from [76].

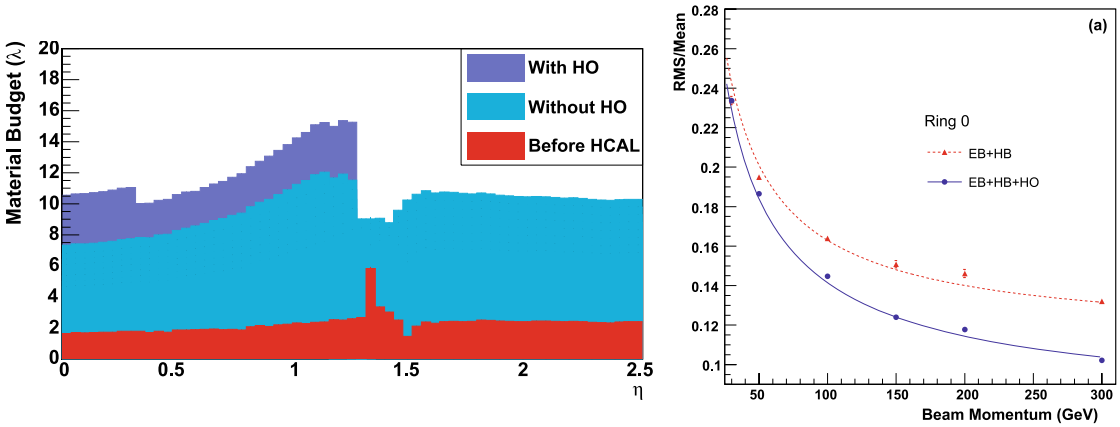

Figure 2.11: (Left) Material budget, expressed in units of hadronic interaction length λ, as a function of the pseudorapidity η. The figure compares the profiles with and without the outer hadronic calorimeter (HO). The "With HO" profile exhibits a drop before $\eta = 0.5$ due to the presence of a piece of iron in the corresponding ring, as explained in the text. (Right) Improvement in the energy resolution for pions fired at $\eta = 0.22$ as a function of the beam energy. Figures taken from [85].

detectors was driven by the need to cover a large area while keeping costs manageable. The muon system is composed of three (plus one) components: the Drift Tube chambers (DTs), the Cathode Strip Chambers (CSCs), the Resistive Plate Chambers (RPCs), and the new Gas Electron Multiplier (GEMs). The layout of the muon system is shown in Fig. 2.12.

The DTs consist of 250 chambers arranged in four barrel stations placed in the $|\eta| < 1.3$ region. The fundamental element of the DT system is the drift cell, which has a transverse dimension of 42×13 mm^2 with a 50-μm-diameter gold-plated stainless-steel anode wire at the centre, maintained at a voltage of +3600 V. The drift cell is filled with a gas mixture composed of 85% argon and 15% CO_2. When a muon passes through the cell,

it ionizes the gas, and the resulting electrons drift towards the anode wire, generating a signal that enables precise measurement of the crossing point of the muon. Multiple DT layers, arranged perpendicular to each other, allow the extraction of time and position information.

Due to the challenges posed by a higher muon rate in the endcap region, as well as a stronger and non-uniform magnetic field, the CMS collaboration opted for a different type of muon chamber in this area. Specifically, the CSCs are installed in the $0.9 < |\eta| < 2.4$ region due to their advantageous features, including fast response time, finely segmented readout, and tolerance to the non-uniform magnetic field. The CSCs function as standard multi-wire proportional counters, with a cathode strip readout that precisely measures the position at which a muon (or any other charged particle) crosses the gas volume. The CSCs have various dimensions tailored to their respective regions. The gas mixture used in all chambers is composed of 50% CO_2, 40% argon, and 10% CF_4. The combination of DTs and CSCs ensures complete coverage of the CMS pseudorapidity range without any gaps, thereby ensuring robust muon identification.

In addition to these tracking detectors, the CMS muon system is completed by the RPCs, which are a complementary, dedicated triggering detector system with excellent time resolution. They are located both in the barrel and in the endcap up to $|\eta| < 1.9$, and they are primarily designed to provide a fast, independent muon trigger. Each RPC is made by a double-gap chamber, and each gap consists of two 2-mm-thick resistive Bakelite plates separated by a 2-mm-thick gas gap. The outer plate is at a voltage of 9.6 kV. The gas mixture is made of 95.2% Freon, 4.5% isobutane, and 0.3% sulphur hexafluoride. The RPC is operated in avalanche mode, i.e., the electrons generated by the passage of muons undergo avalanche multiplication, generating a rapid electrical signal collected by the readout strips.

The CMS muon system is now welcoming a new component, the GEMs. They will cover the forward $1.6 < |\eta| < 2.4$ range, and they are designed to improve muon triggering capabilities in regions close to the beam pipe, where large radiation doses and high event rates will be expected. A first GEM station was installed during LS2 and will contribute to the Run 3 data-taking, while the entire GEM system will be installed during LS3, prior to the beginning of the HL-LHC.

2.2.3 Trigger and data acquisition

With a proton-proton interaction rate of about 40 MHz, the CMS detector produces a large amount of data that should be stored for offline analyses. This leads to an overwhelming volume of data that cannot be feasibly stored, as the full detector information amounts to approximately 1 Mb per event, and there is no technology nowadays able to read out and store such vast volumes of data. However, most of the collisions occurring at the LHC are not of interest to the LHC physics programme. The CMS experiment therefore demands the task of identifying events worthy of saving to the Trigger and Data Acquisition System (TriDAS) [87, 88]. The trigger system is organised into two layers, the *Level-1 trigger*, which reduces the rate from 40 MHz to 100 kHz with a latency time, i.e., time available for data processing, of 3.8 μs, and the *High-Level Trigger* (HLT), which reduces further the rate down to 1 kHz with a latency time of 300 ms.

The L1 trigger takes the raw data from the front-end readout electronic and has to take in few microseconds a decision about the event. It resembles a reader swiftly scanning newspaper headlines to spot captivating stories. Given its need for quick decision-making, the L1 trigger is positioned close to the detector, and it is mounted on custom hardware, such as Field Programmable Gate Arrays (FPGAs) and Application Specific Integrated Circuits (ASICs), for direct and rapid access to the detector information. The small amount of time allocated to the L1 trigger prevents a detailed event reconstruction. In-

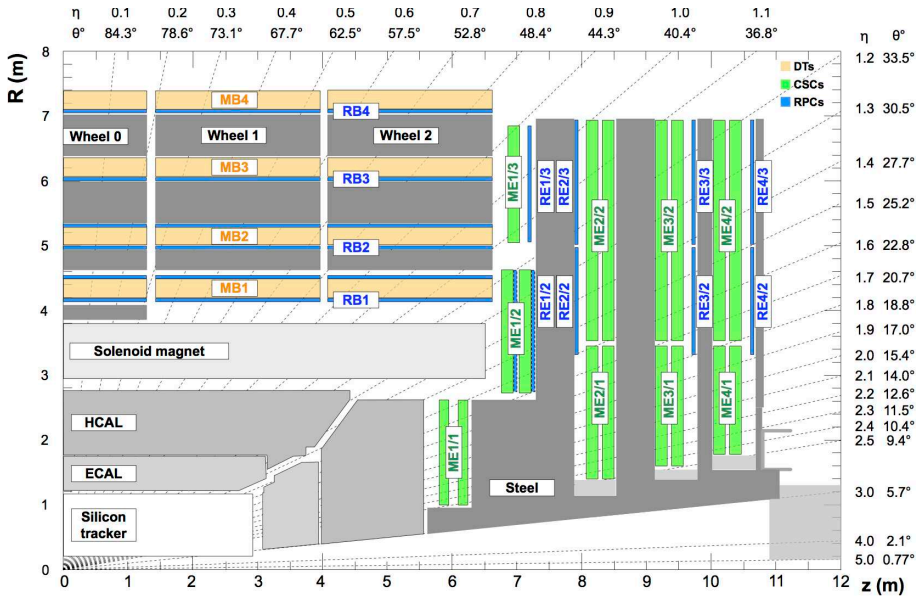

Figure 2.12: Longitudinal cross section of a quadrant of the CMS detector where the components of the muon detection system are highlighted. The drift tube stations (DTs) are labeled MB (Muon Barrel) and the cathode strip chambers (CSCs) are labeled ME (Muon Endcap). Resistive plate chambers (RPCs) are mounted in both the barrel and endcaps of CMS, where they are labeled RB and RE, respectively. The GEMs detectors are not displayed as they are a recent introduction in the muon system. Figure taken from [86].

stead, the L1 trigger produces the so-called *L1 candidates*, based on low-granular detector information and reconstruction of low-resolution physics objects. These L1 candidates rely only on the calorimeters and muon chambers, with the tracker excluded from the L1 reconstruction. The ECAL and HCAL information is combined in *trigger towers* to reconstruct jets, hadronic-taus, e/γ objects, and evaluate the energy sums. Since the tracker information cannot be used, electrons and photons are reconstructed as the same object. On the one hand, the calorimeter information forms the *Calo Trigger*. On the other hand, the redundancy of the muon system is used to define the *Global Muon Trigger*, combining data from DTs, CSCs, and RPCs. The Calo and Global Muon Trigger are then merged into the *Global Trigger*, used by the L1 trigger system to accept or discard the event.

Events meeting the L1 trigger requirements proceed to the HLT trigger, which operates within a software computing farm with 32,000 CPU cores. The HLT works with a dataset that has already been enriched with interesting physics events, thanks to the previous L1 trigger selection, thereby reducing the number of events that should be discarded. Unlike the L1 trigger, the HLT can work with the same raw data as offline reconstruction and include the tracker information previously excluded at the L1 stage. It also employs more sophisticated algorithms akin to those used offline. To meet time constraints, the HLT reconstruction is carried out only around an L1 seed, significantly reducing the computational time. The output of the HLT is then streamed to the Tier-0 at CERN, where it is prepared for offline reconstruction and organised into *primary datasets* (PDs). These PDs constitute collections of HLT paths, representing the selected events that the CMS collaboration will further process and analyse.

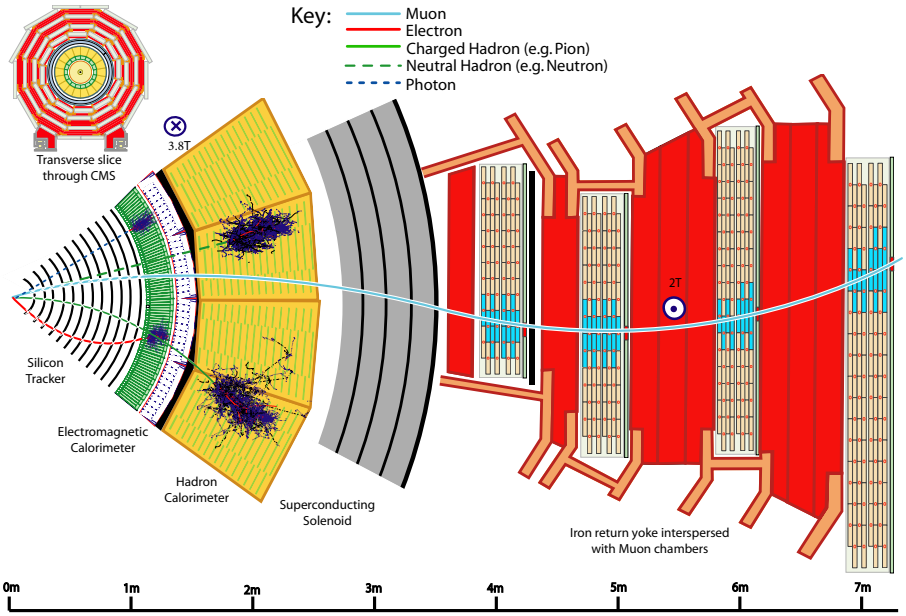

Figure 2.13: Transverse slice of the CMS detector, showing the different sub-detectors and how different particles interact. Figure taken from [89].

2.3 Offline reconstruction of physics objects at the CMS

Essentially, six types of particles can be observed in a detector at a collider, and each of them has a characteristic signature based on few sub-detectors, as sketched in Fig. 2.13. Photons, being neutral, do not leave any signal in the tracker and go straight to the ECAL, where they are completely absorbed and deposit all their energy. Electrons (or positrons) are somehow photons' cousins, with a similar signature. The difference is in being negatively (positively) charged; hence, their trajectory bends in the magnetic field, and they leave hits in the tracker. Charged hadrons, such as pions and kaons, leave a signal in the tracker and can initiate a shower in the ECAL, but they are then completely absorbed in the HCAL. Neutral hadrons feature the same signature of their charged brothers, without leaving hits in the tracker. Muons and neutrinos cross the detector with little or no interactions. While neutrinos escape undetected, and their presence could only be inferred from missing energy, muons produce hits both in the inner tracker and in the outer muon chambers and deposit small energy in the calorimeters. This apparent simplicity could lead to the implementation of a reconstruction framework that only relies on a few sub-detectors per particle, and indeed this has been the approach in many hadron colliders. For instance, one can reconstruct jets only relying on their deposits in the calorimeters. As the energy resolution of the CMS HCAL is $100\%/\sqrt{E}$, this will lead to a poor reconstruction. An improved energy measurement could come from trying to separate the individual jet particles and asking for help from the tracker, which has a better energy resolution. This effect can be seen in Fig. 2.14 and will be discussed later. Hence, event reconstruction can be significantly improved by correlating the information from all detectors to identify each final-state particle. This is the *particle-flow* (PF) approach [89]. This approach had already been used successfully at LEP, but CMS is the first experiment at a hadron collider employing this strategy. The CMS detector was not conceived with PF in mind, but it turned out to be well-suited for this purpose. It has a large magnetic field that effectively separates energy deposits of charged and neutral particles in jets. The fine granular tracker can efficiently reconstruct tracks, and the highly segmented ECAL allows for distinguishing energy deposits from nearby particles. The hermetic HCAL, while less

segmented than the ECAL, can still separate deposits from charged and neutral hadrons. The low-material budget in front of the calorimeters reduces the likelihood of particles initiating showers before reaching them. Lastly, the excellent muon system efficiently and reliably reconstructs muons with high purity.

In the Particle-Flow (PF) approach, the successful reconstruction of particle trajectories and energy measurement depends on two essential building blocks: *PF tracks* and *PF clusters*.

PF tracks The reconstruction of charged-particle tracks is accomplished through a combinatorial track finder algorithm based on the Kalman Filtering (KF) technique [90]. To achieve high tracking efficiency while maintaining a low fake rate, the tracking algorithm is applied in several successive iterations, each targeting different types of particles. The first iteration targets prompt, high-p_{T} particles, passing through iterations dedicated to particles with displaced vertices, to the last iterations designed to increase the muon-tracking efficiency. As the iterations proceed, more complex and time-consuming seeding, filtering, and tracking algorithms are employed, masking the hits already associated with a track. Globally, this iterative tracking is more efficient, purer, and even less CPU demanding than a single-iteration tracking, which was in place at the beginning of Run 1.

PF clusters Energy deposits in neighbouring cells in the ECAL and HCAL are grouped to form energy clusters. The clustering is carried out independently in each component of the calorimeter system: EE, EB, HE, HB, and the two ES. In the HF, the energy deposits are not clustered together, but each cell gives rise to a cluster. The clustering parameters are different in each component, but the algorithm itself is always the same. Starting from a cluster seed, identified as an isolated cell with an energy above a given threshold, a topological cluster is formed by aggregating neighbouring cells with an energy above a threshold set to twice the noise level of the cell. PF clusters are then reconstructed within a topological cluster using an algorithm based on a Gaussian-mixture model. This model postulates that the energy deposits in the M individual cells of the topological cluster arise from N Gaussian energy deposits, where N is the number of seeds. This procedure helps disentangle contributions from overlapping showers.

A particle crossing the CMS detector is expected to produce several PF elements in various CMS sub-detectors. For instance, an electron is expected to form at least a PF track in the inner tracker and a PF cluster in the ECAL. Once the PF elements are defined, a linking algorithm is necessary to connect them through the entire detector. *A priori*, the linking algorithm could test any combination of PF elements. However, to reduce the time needed by the linking and be computationally efficient, pairs of elements are considered only in an η, ϕ neighbourhood. The conditions for linking a pair of elements depend on the nature of the elements themselves. In the case of electrons, the linking can occur between a PF track and a PF cluster. It can also be a cluster-cluster linking, like between a PF cluster in the ECAL and another in the preshower. Finally, two PF tracks can be linked together, like for muons, where a match between a track in the inner tracker and a track in the muon chambers is strongly envisaged. The output of the linking is a *PF block*. Particle reconstruction and identification are carried out sequentially in each PF block. After each step, the corresponding PF elements are removed from the PF block.

- Muons are the first particles that are identified and reconstructed from the corresponding PF tracks.
- Electrons and energetic, isolated photons (both converted and unconverted) follow.

- The remaining elements in the block are then subject to a cross-identification of charged hadrons, neutral hadrons, and photons arising from parton fragmentation, hadronization, and decays in jets.
- Hadrons experiencing a nuclear interaction in the tracker and creating secondary particles are identified and reconstructed at this stage.
- The last step is a global post-processing to reduce the probability of particle misidentification and misreconstruction.

At this point, the PF provides a collection of particles, named *PF candidates* that are then used for higher-level operations: reconstruction of jets, as explained in Sec. 2.3.3, determination of the missing energy, reconstruction of tau leptons from their decay products, and definition of isolation variables.

The main protagonists of the decay channel analysed in this thesis, $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$, are electrons and muons. The following sections will be devoted to a detailed description of these two objects within the PF approach.

2.3.1 Electrons

The electron reconstruction in the PF algorithm [91] starts with the formation of PF clusters in the ECAL. The procedure is outlined in the previous section, which is tuned in the ECAL for the definition of the seed as follows. Cluster seeds are identified as cells with an energy larger than 230 MeV in the barrel and 600 MeV in the endcaps and larger than the energy of the eight neighbouring cells, including cells that only share a corner with the seed candidate. In the endcaps, because the noise level increases as a function of θ, seeds are additionally required to satisfy a threshold requirement on $E_{\mathrm{T}}^{\mathrm{seed}} > 150$ MeV.

While crossing the tracker, an electron can emit bremsstrahlung radiation, which can then undergo pair production. As a result, by the time the electron reaches the face of ECAL, the original electron could give rise to a complicated pattern that results in several PF clusters in the calorimeter. These clusters must be combined into a single *SuperCluster* to reconstruct the original electron. The superclustering will be extensively discussed in Ch. 8.

The standard KF tracks used in the PF approach are not optimal for electrons. The KF track algorithm cannot well describe the sudden and significant energy losses along the trajectory. To address this issue, another approach based on a Gaussian-Sum Filter (GSF) algorithm [92] is employed for electron track fitting. However, the GSF track fitting algorithm is CPU intensive, making running on all hits in the tracker infeasible. Electron tracks must thus be seeded, i.e., a hit pattern likely to lie on an electron trajectory should be identified. The electron track seeding can be either *ECAL-driven* or *tracker-driven*. The former approach is more suitable for high-p_{T} , isolated electrons, while the latter is designed to recover efficiency for low-p_{T} or non-isolated electrons.

The ECAL-driven approach starts from a supercluster with $E_{\mathrm{SC,T}} > 4$ GeV and $H/E_{\mathrm{SC}} < 0.15$, where $E_{\mathrm{SC,T}}$ is the supercluster transverse energy, E_{SC} is the supercluster energy, and H is the sum of the energy deposited in the HCAL towers within a cone of $\Delta R = 0.15$ centred on the supercluster position, defined as its barycentre. The trajectory is then extrapolated back to the interaction point considering the position of the supercluster, its transverse energy, and the intensity of the magnetic field. The accuracy of the extrapolation is directly related to the correct clustering of the energy deposits into the supercluster. Consider an ideal case of an electron that does not emit any photon and is reconstructed in the ECAL as a single cluster. Hence, the supercluster corresponds to the ECAL cluster. The position of this supercluster is denoted as the energy-weighted average of its constituents. Consider now the same electron that instead

undergoes Bremsstrahlung emissions. The corresponding supercluster is the cluster of the single ECAL clusters originated by the different particles of the shower. It is proven that the position of this supercluster, assuming all the energy of the original electron is adequately recovered, is the same as the ideal case. This evidence can be summarised as: for a given electron momentum at the vertex, the barycentre of its energy deposits does not depend on its interactions in the tracker. After computing the extrapolated trajectory, if the first two hits of a tracker seed are matched to the predicted trajectory, this seed is selected for seeding a GSF track.

The tracker-driven approach iterates over all KF tracks. If a track is found to be compatible with an ECAL cluster, the associated seed is passed to the GSF algorithm for refitting. The compatibility is assessed through a logical OR of some cut-based selections and a selection based on a Boosted Decision Tree (BDT). In both cases, variables related to the track quality and track-cluster matching are used.

The collection of ECAL-driven and tracker-driven seeds are used to fit the GSF tracks, which are then extrapolated to the ECAL to perform track-cluster association. A BDT is employed to decide whether to associate a GSF track to an ECAL PF cluster. The BDT combines track information, supercluster observables, and track-cluster matching variables. For tracker-driven tracks, only the BDT is used for the linking, while for ECAL-driven tracks they can either pass the BDT requirements or a cut-based selection based on a geometrical matching:

$$|\Delta\eta| = |\eta_{\mathrm{SC}} - \eta_{\mathrm{trk}}^{\mathrm{extrap}}| < 0.02 \qquad |\Delta\phi| = |\phi_{\mathrm{SC}} - \phi_{\mathrm{trk}}^{\mathrm{extrap}}| < 0.15$$

where $\eta_{\mathrm{trk}}^{\mathrm{extrap}}$ and $\phi_{\mathrm{trk}}^{\mathrm{extrap}}$ indicate the closest position obtained by extrapolating the innermost track position.

Up to this point, we have only talked about electrons, but this chain is actually valid also for (isolated) photons. At this stage, there is no distinction between photons and electrons. While photons do not emit bremsstrahlung emission like electrons, they can undergo pair production in the tracker, thus leading to a similar reconstruction challenge as encountered for electrons. In a given PF block, an electron candidate is seeded from a GSF track if the corresponding ECAL cluster is not linked to three or more additional tracks. A photon candidate is seeded from an ECAL supercluster with $E_{\mathrm{T}} > 10$ GeV, without any link to a GSF track. Electron candidates must satisfy additional IDentification (ID) criteria, which will be explained in the next chapter, particularly for the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ analysis.

2.3.2 Muons

Muons are the first particles to be reconstructed in CMS due to their clean signatures. The location of the muon chambers as the final layer ensures a high purity in the reconstruction. All other particles are absorbed in the previous sub-detectors, meaning that if a hit is seen in the muon chambers, it is likely due to a real muon, while the inner tracker precisely measures its momentum. The muon PF tracks can be reconstructed in three different ways. Hits from all detectors in the muon chambers can be used to fit a *standalone-muon* track. Then, each standalone muon can be matched to a track in the inner tracker to form a *global-muon* track. The global fit improves the momentum resolution compared to the tracker-only fit for $p_{\mathrm{T}} \gtrsim 200$ GeV. The *tracker-muon* track goes in the opposite direction. Starting from a track in the inner tracker, if it has $p_{\mathrm{T}} > 0.5$ GeV and $p > 2.5$ GeV, the track is extrapolated to the muon chambers. If at least one muon segment matches the extrapolated track, the muon is promoted to be a tracker muon.

An efficient global-muon reconstruction needs segments in at least two muon stations. However, for low-momentum muons, the larger multiple scattering in the steel of the return

yoke often causes this requirement to fail. In such cases, the tracker-muon reconstruction, typically for momenta below $p < 5$ GeV, ensures higher efficiency as it only needs a single muon segment in the muon chambers.

The muon ID relies on a set of variables that, when combined in different ways, provide different compromises between efficiency and purity, and then it is demanded to each analysis to choose their desired type of reconstructed muon. There are five primary IDs used in CMS physics analyses. The starting point for all identifications is a *loose muon* (also known as *PF muon*), which is either a tracker or a global muon. It targets prompt muons, i.e., originating from the Primary Vertex (PV), and muons from hadron decays. The PV is defined as the reconstructed vertex with the largest value of summed physics-objects p_{T}^2. There are three types of IDs that apply tighter requirements on top of a loose muon, *medium ID*, designed to be highly efficient for prompt muons and for muons from heavy quark decays, *tight ID*, aiming to suppress muons from decay in flight and hadronic remnants, and *High-p_{T} ID*, optimised for muons with $p_{\mathrm{T}} > 200$ GeV. The last standard ID is the *Soft ID*, which starts from a tracker muon and is optimised for low-p_{T} muons for B-physics and quarkonia analyses. As a general remark, charged hadrons can cause background for muon reconstruction when some of their shower remnants reach the muon system. To improve muon identification performance, specific identification criteria are applied. These criteria are often based on matching tracks in the muon chambers with energy deposits in the ECAL, HCAL, and especially the HO. A detailed description of the muon IDs can be found in [86].

2.3.3 Jets

When a proton-proton collision occurs at the LHC, a significant fraction of interactions involves quarks and gluons. These quarks and gluons are not directly observed as free particles due to a phenomenon called confinement, which confines them within composite particles known as hadrons. The process leading to the formation of hadrons is named *hadronisation* and occurs immediately after quarks and gluons are produced. Quarks and gluons fragment during hadronisation, leading to a collimated spray of particles moving in similar directions. These collimated sprays are the experimental signatures of quarks and gluons, called jets.

Jets can be reconstructed using different techniques. One approach is to reconstruct jets from energy deposits in the calorimeter towers, and this is what is usually called a *Calo jet*. Consequently, the jet energy will be determined by the energy resolution of the calorimeters. However, the CMS experiment profits from the PF algorithm, which reconstructs all particles in the event by clustering the four-momenta of PF candidates in *PF jets*. The PF algorithm's use of the inner tracker and the high granularity of the ECAL enables better energy measurement of charged hadrons and photons inside a jet, which together constitute around 85% of the jet energy. As a result, the PF jets provide improved energy resolution and spatial resolution compared to Calo jets, as reported in Fig. 2.14.

Jets are not physics objects *per se*, they need to be defined. A set of rules are needed on how to group particles into jets. Jet algorithms usually rely on a metric, i.e., one or more parameters indicating how close two particles must be to belong to the same jet. Additionally, jet algorithms are usually associated with a recombination scheme, indicating how the momenta of individual particles are combined to determine the momentum of the final jet. For Calo jets, the clustering algorithm is applied to the energy deposits in the calorimeter towers[7], while for PF jets, the clustering algorithm is applied to the PF candidates. The common conditions to be satisfied for a jet algorithm were commonly

[7] A calorimeter tower is composed of an HCAL tower and the 25 underlying ECAL crystals

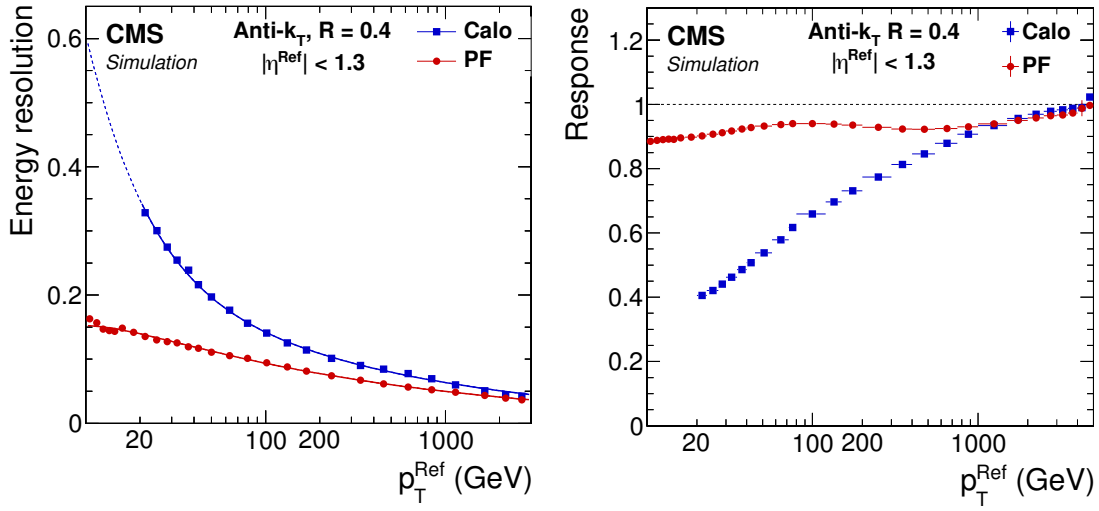

Figure 2.14: Jet energy resolution and response of Calo and PF jets, as a function of the transverse momentum of the reference jet $p_{\mathrm{T}}^{\mathrm{Ref}}$. The reference jet is defined as the result of the jet algorithm applied to all stable particles produced by the event generator excluding neutrinos. Figures taken from [89].

established in 1990 [93]:

- simple to implement in an experimental analysis;
- simple to implement in the theoretical calculations;
- defined at any order of perturbation theory;
- yields finite cross sections at any order of perturbation theory;
- yields a cross section that is relatively insensitive to hadronisation, which is a non-perturbative process that is hard to compute.

To these conditions, it should be added another important requirement, the infrared and collinear safety. This implies that if one modifies an event by a collinear splitting or the addition of a soft emission, the set of hard jets that are found by the algorithm should be unchanged. The algorithm that is now extensively used by the CMS collaboration is the anti-k$_{\mathrm{T}}$ [94], defined by the metric:

$$\begin{cases} d_{ij} = \min(p_{\mathrm{T},i}^{-2}, p_{\mathrm{T},j}^{-2}) \frac{\Delta R_{ij}^2}{R^2} \\ d_{i\mathrm{B}} = p_{\mathrm{T},i}^{-2} \end{cases} . \tag{2.16}$$

where $\Delta R_{ij}^2 = \Delta y_{ij}^2 + \Delta \phi_{ij}^2$, d_{ij} is the metric evaluating the distance between two particles, $d_{i\mathrm{B}}$ is defined as the particle-beam distance, and R is the cut-off distance parameter. This algorithm falls under the so-called *sequential recombination algorithms*, i.e., a family of iterative algorithms following a bottom-up approach by combing particles starting from closest ones:

1. for each pair of particle i and j, compute the distances d_{ij} and $d_{i\mathrm{B}}$;
2. find the smallest between d_{ij} and $d_{i\mathrm{B}}$ for all particles;
3. if the minimum distance is d_{ij}, combine the particles i and j together, otherwise if it is $d_{i\mathrm{B}}$, call i a jet and remove it from the list of particles;

4. this procedure is repeated starting from the first step until no particles are left.

The anti-k_T satisfies all the conditions mentioned above and it only uses variables that are Lorentz-invariant under longitudinal boosts (cfr Sec. 2.2.1). The only counterpart is the complexity of the algorithm, which scales as $N^{3/2}$, where N is the number of particles.

The energy measurement of jets can significantly differ from the true energy of the hadron they are supposed to represent. Many effects, such as the non-linear response of calorimeters, detector noise, and additional contributions from PU, cause this discrepancy. To address these issues, the Jet Energy Corrections (JEC) have been developed as a set of tools to properly map the measured jet energy deposition to the particle-level jet energy [95, 96]. The CMS experiment has adopted a factorised solution to the problem, where each correction level takes care of a different effect. Each correction is essentially a scaling of the jet's four-momentum with a scale factor, and the corrections are applied sequentially in a pre-determined fixed order (Fig. 2.15). The correction chain begins with a PU correction, which accounts for the spurious energy contribution from PU interactions. It is an offset energy that is subtracted from the jet energy. Next, simulation-truth-based corrections are applied to tackle the non-uniformity of the detector response as a function of the jet p_T and η. Finally, minor residual corrections are applied to the data to account for remaining pileup effects that are dependent on the jet's η and p_T]. The PF approach, with the definition of PF jets, combined with the chain of energy corrections outlined above, provides a final jet energy resolution for Run 2 of less than 20% for jets with $p_T > 30$ GeV in the barrel and less than 30% in the endcap. The resolution across the three data-taking years is similar and does not show significant deviations [96].

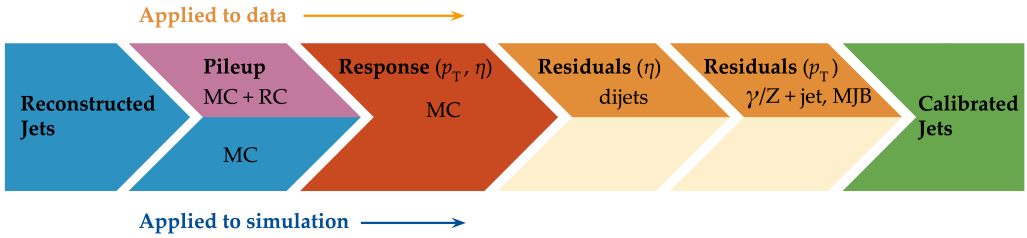

Figure 2.15: Sequential jet energy corrections to apply to obtain a calibrated jet. This is the configuration used for Run 2. Figure taken from [96].

In addition to the Jet Energy Scale (JES), also the Jet Energy Resolution (JER) should be corrected as it is worse in data than in the simulation. Hence, jets in simulations need to be smeared to describe the data. The CMS collaboration utilises the combination of two different methods. The first one rescales the four-momentum of a reconstruct jet with the factor:

$$c_{\mathrm{JER}} = 1 + (s_{\mathrm{JER}} - 1)\frac{p_T - p_T^{\mathrm{gen}}}{p_T} \tag{2.17}$$

where p_T is the transverse momentum of the reconstructed jet, p_T^{gen} is the transverse momentum of the corresponding jet obtained from clustering generator-level particles, and s_{JER} is the data/simulation core resolution scale factor. This method can be applied only if there is a well-matched generator-level jet. When this is not the case, a stochastic smearing is used, which rescales the jet four-momentum with the factor:

$$c_{\mathrm{JER}} = 1 + \mathcal{N}(0, \sigma_{\mathrm{JER}})\sqrt{\max(s_{\mathrm{JER}}^2 - 1, 0)} \tag{2.18}$$

where, σ_{JER} is the relative p_T resolution in simulation, $\mathcal{N}(0, \sigma_{\mathrm{JER}})$ denotes a random number sampled from a normal distribution with the mean equal to zero and variance equal to σ_{JER}.

2.4 The phase-2 of the CMS detector

As of the time of writing this thesis[8], the phase-1 of the CMS detector is to end, with the second year of Run 3 currently underway. The phase-1 has already seen a vast quantity of valuable physics results, summarised in more than 1,000 published papers. The main highlight was undoubtedly the discovery of a new resonance in 2012 that has matched very closely all the properties to be identified as the SM Higgs boson. However, CMS is not only Higgs physics. The study of the electroweak symmetry breaking is complemented with vector-boson scattering processes, allowing a precise test of the gauge structure of electroweak interactions through quartic and their interplay with trilinear couplings. During Run 2, there have been the first observations of some processes and evidence of many others [97–99]. B-physics has also played an important role, with notable achievements like the observation of the rare $B_s^0 \to \mu^+\mu^-$ decay and the evidence of $B^0 \to \mu^+\mu^-$ [100]. These processes are highly suppressed in the SM but could receive contributions from BSM physics, enhancing their production rate. Their study allows strict constraints on models of new physics. Although direct searches for exotic processes, dark matter, and supersymmetric particles have not yet produced any evidence, they have been a precious laboratory to sharpen our experimental tools and shape the theory landscape of BSM models.

The CMS experiment will continue in its quest for BSM physics and push the boundaries of our knowledge of fundamental physics during the HL-LHC with its phase-2. The HL-LHC will allow the CMS detector to collect a significantly larger amount of data, with a planned accumulation of 3 ab^{-1} by the end of the HL-LHC. The dataset from phase-1 will account for only 10% of this total, indicating the vast increase in statistics available during phase-2, opening many physics opportunities. The study of the Higgs boson will remain at the forefront of the CMS physics programme. The precision of Higgs boson couplings will be improved 3/4 times compared to today, with nearly all couplings measured to a precision of 2% [41]. The observation of the $H \to \mu\mu$ channel will be already feasible by the end of Run 3, while $H \to Z\gamma$ will be observed for the first time during phase-2. The increased sensitivity will also enable the establishment of the existence of the SM HH production [41], allowing a direct probe of the shape of the Higgs potential. The study of the triple-gauge coupling and quartic-gauge coupling will continue during phase-2 via vector boson processes, and the larger dataset may open the possibility of finding BSM contributions. In general, more statistics will allow more sensitivity to discover rarer processes or with more challenging experimental signatures. In support of the whole CMS physics programme, the higher statistics will provide a deeper insight into topics that will help many other analyses, such as SM backgrounds and Parton Distribution Functions (PDFs) of protons, which will be limiting sources of uncertainty in many analyses without significant progress in that regard. The search for new physics builds on our knowledge of SM physics.

The larger luminosity will open new horizons but also new experimental challenges for the detector. The main challenge for the CMS detector will be to withstand the radiation damage and progressive degradation of the physics performance due to the higher radiation dose. Figure 2.16 shows the simulation of the absorbed dose at the end of phase-2. The absorbed dose during one year of data-taking at the HL-LHC will correspond to the absorbed dose during the entire phase-1. The second main challenge will be the high level of PU. Increasing luminosity will allow us to pursue precision physics and access rarer phenomena at the price of increasing the average number of interactions in a single crossing. During Run 2, the average PU was 35; during phase-2 this number will increase to 140, with the possibility of going up to 200. This huge leap in PU will increase the amount

[8] Summer 2023

of data to be read, the lepton isolation definition will be stress tested, and trigger and offline reconstruction should be thoroughly revised. Consequently, the CMS collaboration foresees a significant upgrade of its detector to maintain and possibly improve the physics performance achieved during phase-1. The increased radiation level will require improved radiation hardness from detectors and front-end electronics, while the more considerable particle flux from PU will require higher detector granularity, increased bandwidth to accommodate higher data rates, and improved trigger capabilities to keep the trigger rate at an acceptable level.

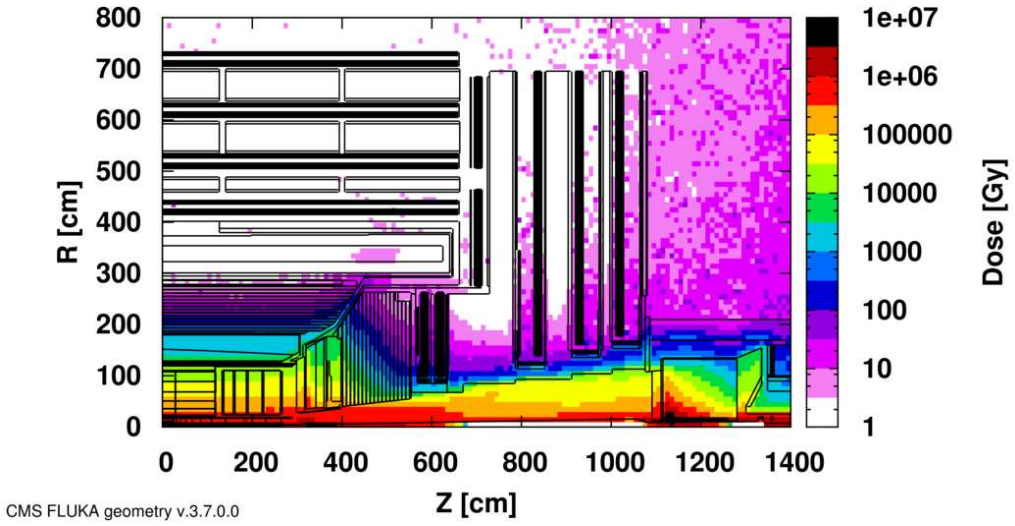

Figure 2.16: Distribution of absorbed dose over the CMS detector after an integrated luminosity of 3 ab^{-1}. Figure taken from [101].

The region of the detector that will suffer the most of the changed beam conditions will be the one closer to the interaction point, i.e., the inner tracker, and the forward region, i.e., the ECAL and HCAL endcap calorimeters. The current tracker will be severely damaged by radiation by the end of phase-1 and will not be able to sustain the data-taking periods during phase-2. Hence, it will undergo a complete replacement [102]. Its granularity will be increased by a factor of 4, the material budget will be reduced, lowering the probability of early showers, and the forward acceptance will be increased to $|\eta| \simeq 4$. On the other hand, the upgrade foreseen for the endcap calorimeters is one of the topics of this thesis and will be discussed in detail in the next section.

The trigger and data acquisition system will also undergo major changes [103, 104], particularly in relation to the L1 trigger. One of the main changes in the L1 trigger is the incorporation of tracking information, which aims to maintain a sustainable event rate without compromising the physics performance. However, this modification will require an increase in the latency time from the current 3.8 μs to 12.5 μs. The L1 output rate will also increase from 100 kHz during phase-1 to 750 kHz. Due to the increased L1 output rate, the HLT will also need to be upgraded to achieve the same rate reduction factor of 100. It has been determined that the maximum acceptable rate for storage and offline processing is 7.5 kHz.

Conversely, the muon chambers are expected to sustain the HL-LHC harsh experimental environment without drastic changes [105]. In order to increase radiation tolerance and readout speed the electronics of the DTs and CSCs will be replaced. The main detector improvement concerns the very forward region, where improved RPCs and the new GEM system will be installed. This will add redundancy, improve trigger and reconstruction performance, and increase the forward acceptance to $|\eta| \simeq 3$.

The barrel calorimeters, both ECAL and HCAL [106], will change the front-end elec-

tronics and back-end readouts to cope with the new L1 requirements. Particularly, the ECAL upgrade will allow the usage of single crystal information at L1, while now the calorimeter information is gathered in 5×5 groups of crystals.

As stated before, the CMS experiment will rely on timing information to mitigate PU. It has been recently decided to include a new MIP Timing Detectors (MTD) for phase-2 [107] to be placed in front of the barrel and endcap calorimeters. A summary of the main upgrades foreseen for the phase-2 of the CMS detector are reported in Fig. 2.17.

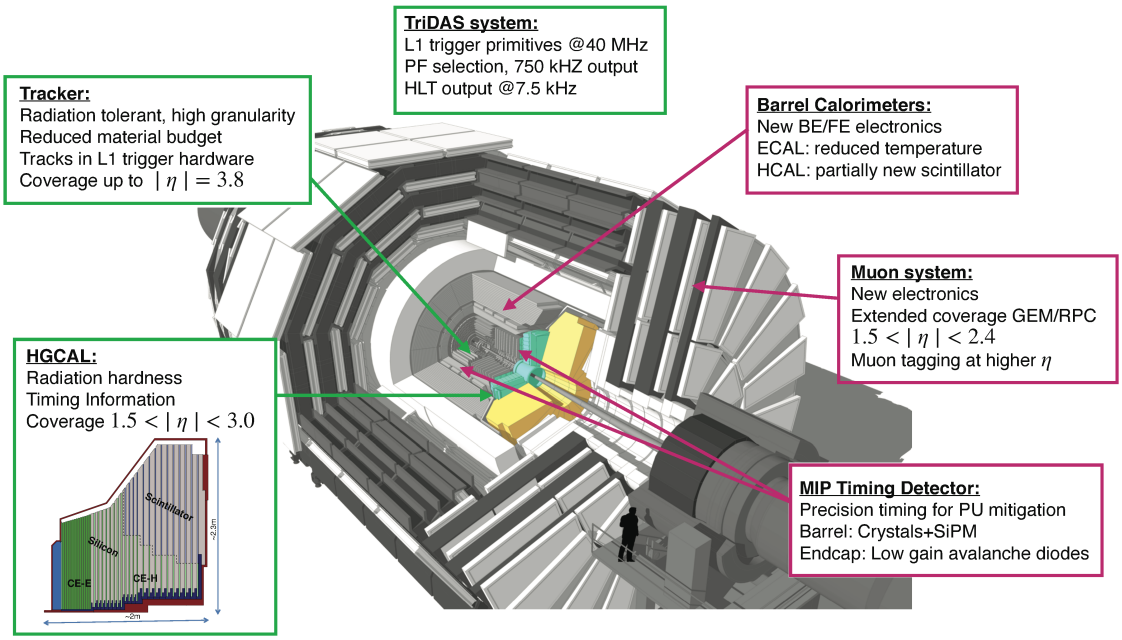

Figure 2.17: Pictorial representation of the CMS detector with the main upgrades foreseen for the HL-LHC. The green boxes represent detectors and system that will be completely replaced, while purple boxes indicates systems that will undergo minor upgrades. Figure taken from [108].

2.4.1 The High Granularity CALorimeter

The existing ECAL and HCAL forward calorimeters were designed for an integrated luminosity of 500 fb^{-1}, which is expected to be exceeded shortly after the beginning of the HL-LHC. Beyond this point, the physics performance will degrade to an unacceptable level [101]. The CMS experiment thus foresees the complete replacement of the endcap calorimeters with a profoundly different calorimeter. It is clear from simulations that the new sub-detector will have to withstand a fluence of 10^{16} n_{eq}/cm^2 and a dose of 2 MGy (cfr Fig. 2.16). R&D activities have proven that the best material to meet these requirements is silicon, which can cope with fluences up to 1.5×10^{16} n_{eq}/cm^2, 50% higher than the one expected during phase-2. Hence, silicon was selected to be the active material of the new detector. In addition to radiation hardness, the new calorimeter must satisfy other requirements outlined below.

- A dense calorimeter to ensure lateral containment of showers.
- A fine lateral granularity to allow the separation of close-by showers and the observation of narrow jets. The consequent small cell size will reduce the energy equivalent of electronics noise increasing the S/N ratio.

- A fine longitudinal granularity in order to sample the longitudinal development of showers for good energy resolution, implementing pattern recognition algorithms, and improving PU rejection.
- A precise timing measurement that will mainly help in PU rejection and identification of vertices.
- The ability to effectively contribute to the L1 decision.

The result of all these requirements is the new High Granularity endcap CALorimeter (HGCAL) [109], a sampling calorimeter composed of an electromagnetic section (CE-E) and a hadronic section (CE-H), covering the $1.5 < |\eta| < 3.0$ region, and weighing 215 tonnes per endcap. The active material will be hexagonal silicon sensors in the more demanding radiation regions, i.e., the entire CE-E compartment and a large fraction of the CE-H sector. The choice of the hexagonal shape is to cover the entire area more efficiently. Instead, in the more outer region of the CE-H, where the dose and fluence will be lowered (dose less than 3 kGy and fluence limited to $8 \cdot 10^{13}$ n_{eq}/cm^2), the active material will be replaced by cheaper highly-segmented plastic scintillator tile boards. The CE-E will extend for 26 layers, with a sequence of CuW, Cu, stainless steel, and Pb absorbers, for a total radiation length of $27.7X_0$ and a nuclear length of 1.5λ. On the other hand, the CE-H will extend for 21 layers, with stainless steel as absorber, for a total interaction length of 8.5λ. Everything will be enclosed in a thermally shielded volume at $-35°$C, to ensure the proper functioning of the silicon sensors. A summary of the properties of the HGCAL is reported in Fig. 2.18.

The 8-inch hexagonal silicon sensors will be deployed with three different thicknesses of 300, 200, and 120 μm, in regions of increasing fluence. In order to optimise the charge collection and reduce the leakage current, it is advantageous to use thinner sensors in the regions of higher fluence. Each silicon sensor will be made of different cells for the readout with two different active areas: 0.52 cm^2 for the 120 μm active thickness sensors, and 1.18 cm^2 for the 300 and 200 μm active thickness sensors. This will define two regions in the detector, namely a *high-density* and *low-density* region, depending on the size of the single readout diode. The transition region will be at a radius of 70 cm^2, corresponding to $|\eta| \simeq 2.15$. The high-density, i.e., more granular region, is located at higher pseudorapidity, where it is expected a larger number of tracks entering to the HGCAL.

The silicon sensors will be placed inside *modules*, mounted on one side to a baseplate, and on the other side to the hexaboard containing the front-end electronics and the printed circuit board. The baseplate is composed of CuW in the CE-E, contributing to the CE-E absorber, while in the CE-H the baseplate material is carbon fibre, with a negligible contribution to the CE-H absorber material. These modules are mounted on either side of a 6 mm thick Cu cooling plate that forms, combined with the CuW baseplate, one absorber layer. At a distance of 1.5 mm from the hexaboard, the motherboard groups the hexaboards in larger physical and logical units. A sequence of motherboard-silicon module-motherboard is sandwiched between two 2.1 mm thick lead planes clad with 0.3 mm stainless steel (SS) sheets, forming an alternative absorber layer. This composition leads to an alternate sequence of SS + Pb and CuW + Cu absorber layers, hence a different amount of absorbing material in front of an active layer depending on whether it is odd or even, as shown in Fig. 2.19. This structure has visible consequences in the longitudinal development of a shower, resulting in a different amount of energy released in the odd and even layers (cfr Sec. 7.3). The HGCAL will have a total of 6 million silicon channels read out independently, organised in 30,000 modules. These modules will be assembled and mounted into 60° self-supporting units called *cassettes*.

Table 2.1: Features of the silicon sensors in the layers deploying only silicon sensors. The silicon cell size defines two regions, namely the high-density and low-density region.

Region	Low-density		High-density
Active thickness (μm)	300	200	120
Cell size (cm^2)	1.18	1.18	0.52
Expected range of fluence ($\times 10^{15}$ n$_{eq}$/cm^2)	0.1-0.5	0.5-2.5	2-7
Largest outer radius (cm)	$\sim$ 180	$\sim$ 100	$\sim$ 70
Smallest inner radius (cm)	$\sim$100	$\sim$ 70	$\sim$ 35

Where the dose permits in the CE-H, the silicon sensors will be replaced by plastic scintillators. Consequently, the CE-H is subdivided into two sections: the first 7 layers, where only silicon sensors are deployed; the remaining layers, where the inner part is composed of silicon sensors and the outer part is composed of scintillators. This configuration will result in the $|\eta| > 2.4$ region of the HGCAL that will be covered exclusively by silicon sensors. The scintillating cells will have a variable size from 4 cm^2 in the inner region to 30 cm^2 in the outer region. The scintillation light will be read out directly by on-tile silicon photo-multipliers. The absorber in the CE-H consists of 10 planes of 41.5 mm thick SS plates, followed by another 10 planes with a thickness of 60.7 mm. The first absorber layer, dividing the CE-E from the CE-H, is instead 45 mm thick, also serving as a structural support of the entire CE-E. In total, there will be 240,000 scintillator channels organised in 4,000 boards. For layers featuring both types of active material, the inner silicon component and the outer scintillator component will be assembled into cassettes with an angular width of 30°, that are later joined together to form a 60° unit.

In conclusion, the new endcap calorimeter will be the first large-scale silicon-based imaging calorimeter employed in a high-energy experiment. This detector will offer the unique capability of performing calorimetry with tracker-like granular information, enabling unprecedented accuracy using position, energy, and timing information. This will open a new era in calorimetry. Such a revolution on the hardware side must be accompanied by another similar revolution on the reconstruction side, both online and offline. The development and optimisation of the offline reconstruction is one of the topics of this thesis and will be discussed in Ch. 7 and Ch. 8.

2.5 Monte Carlo and data processing in the CMS collaboration

Monte Carlo[9] (MC) samples are a powerful and indispensable tool in particle physics. They model complex physics processes, predict experimental outcomes, and help in understanding the behaviour of particles in detectors, simulating a real-life scenario. In this thesis, MC samples are extensively utilised, both in the physics analysis presented in Part II and the development of the HGCAL reconstruction showcased in Part III. In particular, the so-called *full simulation* is used. The concept of full simulation refers to a comprehensive simulation of the entire experimental setup, including the interaction of particles with the detector material, the propagation of particles through the detector components,

[9]The choice of the name dates back to the dawn of the Monte Carlo method during World War II. It reflects the intrinsic random nature of the method, similar to the roulette game one can play at the Monte Carlo Casino

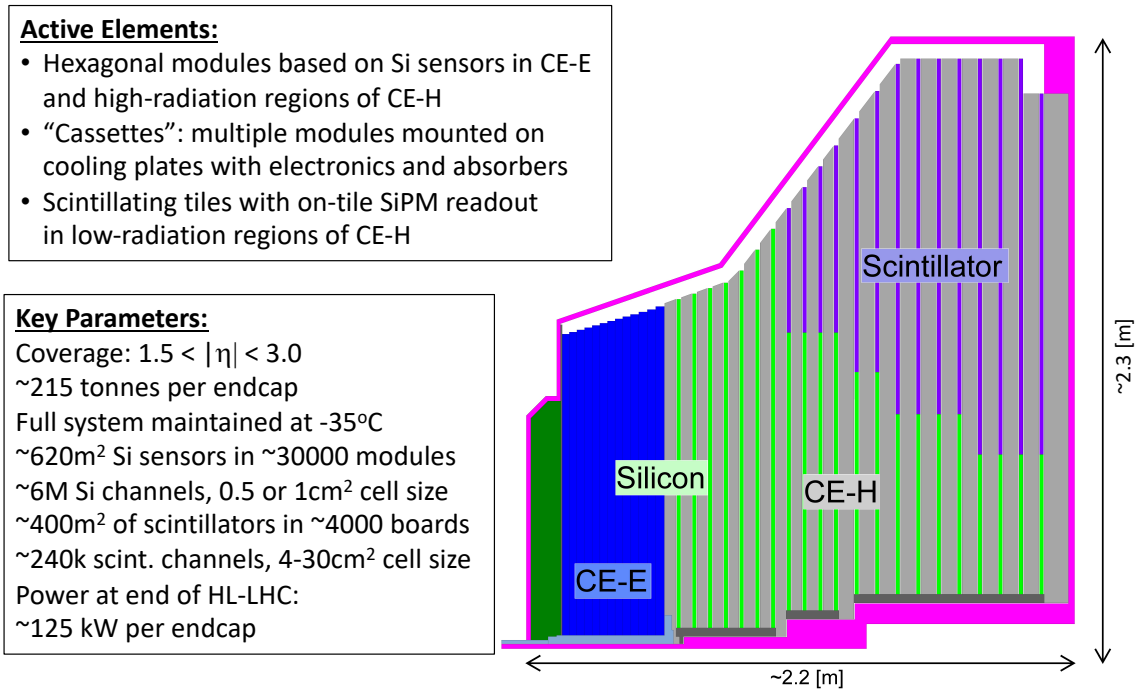

Figure 2.18: Overview of the features of the HGCAL and cross section view of the calorimeter. The CE-E and first layers of the CE-H sections will be made entirely of silicon sensors, while the last layers will be a mixture of silicon sensors and plastic scintillators. The electromagnetic calorimeter (CE-E) comprises 26 layers ($27.7X_0$, 1.5λ), whereas the hadronic calorimeter (CE-H) comprises 7 silicon layers and others 14 layers made of silicon and scintillators ($\sim 8.5\lambda$).

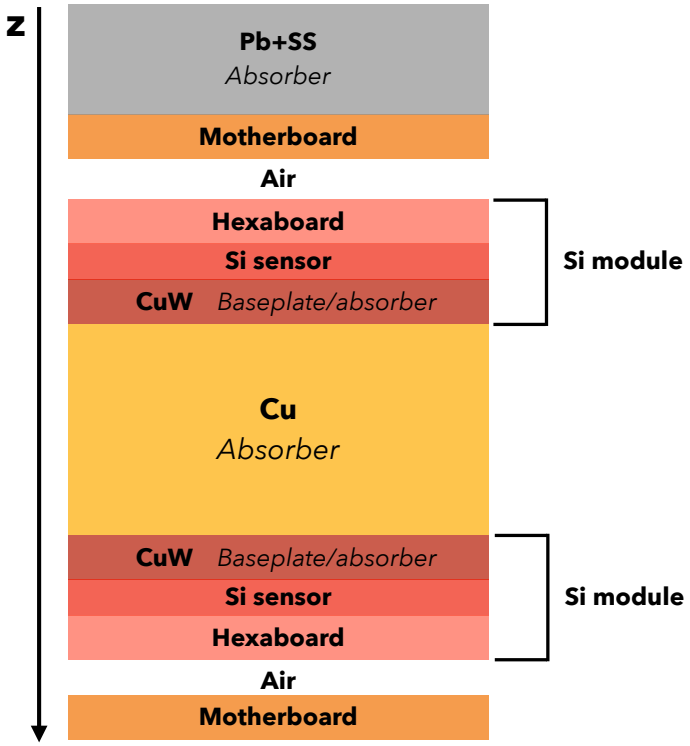

Figure 2.19: Longitudinal structure of a fundamental unit of the CE-E. Each unit comprises two sampling layers.

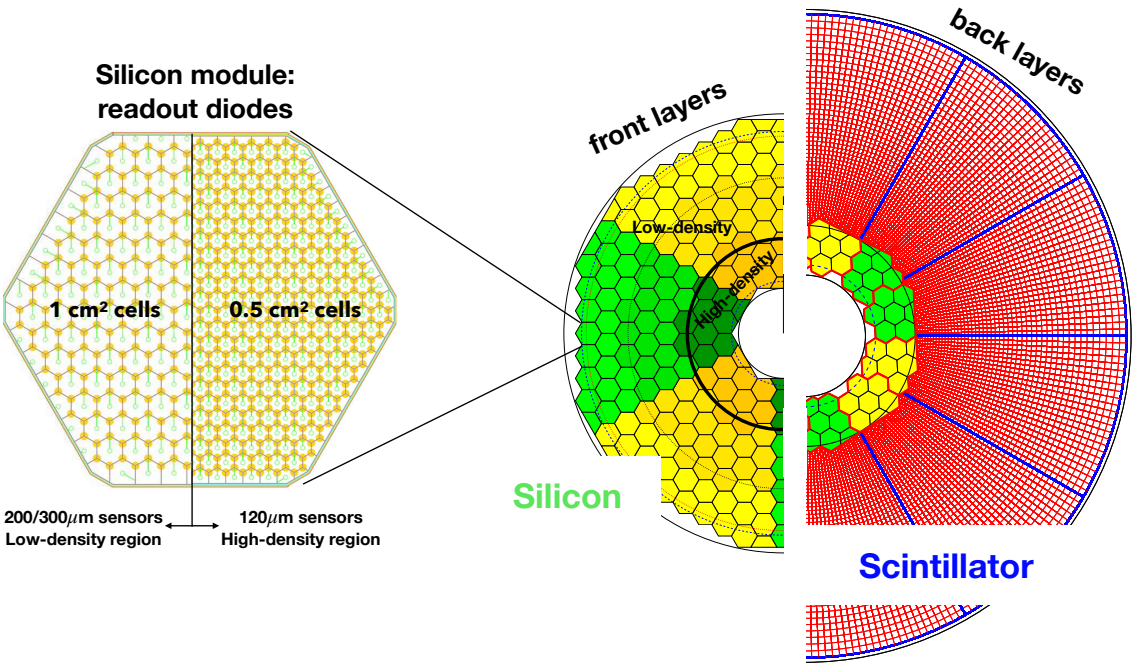

Figure 2.20: (Left) Representation of the silicon sensors with two possible cell sizes. (Right) The left half-circle shows the layout of a layer where only silicon sensors are present. The radial changes in darkness of colour indicate the different silicon thickness: 300, 200, and 120 μm. The solid black line marks the boundary between the high-density and low-density region. The succession of green and yellow colours delimit the 60° cassettes. The right half-circle shows the layout of a layer where both silicon sensors and scintillators are present. The blue lines in the scintillator part and the red lines in the silicon part delimit the 30° cassettes. Figure adapted from [108].

and the response of the various sub-detectors. The CMS collaboration has implemented a standardised workflow for full MC simulations. Once MC events are generated and real data are collected, both undergo the same data processing, ultimately producing the data formats that the different CMS analysis teams will use. All the steps are illustrated in Fig. 2.21.

GEN The initial step involves the GENeration (GEN) of the physics process to be studied, which can range from simulating a single particle to more complex scenarios involving particle production and decay in perturbation theory, along with consideration of the underlying event. The underlying event represents all activity originating from a single particle-particle interaction occurring on top of the process of interest. This includes both initial and final state radiation as well as the interaction between the beam remnants. Unlike the PU, the underlying event is characterised by having the same vertex as the hard scatter, resulting in being tied to the process of interest.

SIM Once the physics process to study is generated, the output of the GEN step is merely a list of stable particles. This list of particles is then fed to the SIMulation (SIM) of the detector, which is carried out by `GEANT4` [110, 111]. `GEANT4` is a widely used software toolkit for the simulation of the passage of particles through matter. It propagates the particles produced in the GEN step in very little steps in the volume of the detector. At each step, it simulates the interactions with materials and the external electromagnetic field. These very granular iterations cause the simulation with `GEANT4` to be very time-consuming. One important feature of this step is the

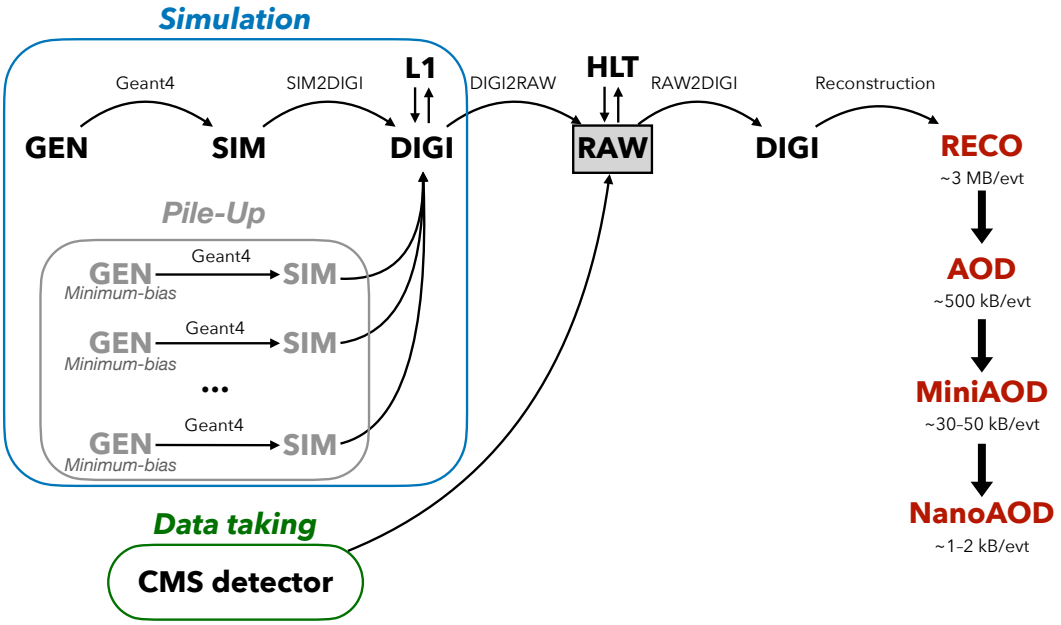

Figure 2.21: Workflow for MC events generation and data processing within the CMS collaboration.

geometry loaded in the simulation. A detailed geometry of the CMS detector is embedded in `GEANT4`, and the one for phase-2 is continuously updated and fixed following the changes in the design of the new detectors to simulate the environment foreseen during the HL-LHC faithfully. The output of this step is a collection of *hits*, representing energy deposits in different volumes of the detector.

DIGI The output of the SIM step is then DIGItalised (DIGI) into electric signals, commonly referred to as *digis*. During this process, the hits generated by `GEANT4` are converted into the same digital format produced by the actual detector electronics. This includes simulating noise, readout logic, and shaping and digitalising the pulse to faithfully replicate the output of the detector electronics. Since the DIGI level faithfully replicates the output of the detector electronics, the Level-1 (L1) is emulated at this stage, ensuring consistency with the actual data processing. Additionally, if foreseen by the simulation, the simulation of PU is merged at the DIGI level. The generation of PU relies on the same GEN and SIM steps outlined earlier, requiring the production of a number of *minimum bias* events. This number is determined by a random extraction from the PU distribution. Minimum bias events model inelastic proton-proton collisions and are named after the trigger used to select and study them. This trigger is very loose, hence introducing a minimal bias. In CMS, the minimum-bias trigger is solely based on the Hadronic Forward (HF) calorimeter, requiring at least one trigger tower to register a signal above a certain threshold.

RAW The digis are then formatted and packed into the RAW data format, which is the same format provided by the CMS detector. This is also the right format to emulate the HLT. The RAW step is the *trait d'union* between simulation and real data, from this point onward, both types of data follow the same path.

The data processing chain begins with the RAW data that are unpacked back into digis. This marks the starting point of data processing for real data. These digis and the previous ones must be identical for simulations to ensure consistency between simulated

and real data. Once this step is completed, the data is ready for offline reconstruction, as outlined in Section 2.3. However, in most cases, the digis are not directly utilised for reconstruction; instead, they are clustered into *recHits* (reconstructed hits). The output of the offline reconstruction is a data format usually called RECO, which contains detailed information on the reconstructed physics objects. Due to its high computational intensity, the RECO data format is produced a few times per period of collected data. This typically occurs within 48 hours of data collection (prompt-reco), at the end of the yearly data-taking period (ReReco), and once during the LS periods (Legacy). An improved calibration of the detector characterizes each reprocessing. Run 2 has also seen a second legacy reprocessing, named ultra-legacy. Since RECO files contain the most detailed and comprehensive information, they are quite large ($\sim$3 MB/event) and unsuitable for efficient physics analyses. The AOD format prioritizes the physics object collections used in analyses, retaining only essential hits and a few detector-level details. The space needed by each event drops to 500 kB per event. During Run 1, this format served as the standard reference for CMS analysts. However, Run 2 collected significantly more data than Run 1, making AOD files impractically large for the related analyses. To address this challenge, the CMS collaboration introduced a condensed data format known as MiniAOD [112]. MiniAODs are only 10% of the size of AODs and can be produced from the AOD dataset in 1–2 days. The size reduction was achieved by reducing numerical precision when not necessary, using lightweight formats for high-level physics objects, storing only those with transverse momentum above a certain threshold and really necessary for physics analyses. MiniAOD has been the reference data format for Run 2, effectively meeting the needs of most analyses. To further decrease the data size, the CMS collaboration has now introduced the NanoAOD format [113]. The content has been chosen based on Run 2 analysis experience and interactions within the collaboration: a large set of analyses use the same high-level information, requiring fewer lower-level details. A NanoAOD resembles the typical structure and size of private ntuples, with new features to make it more universal and compatible with central processing tools.

Part II

The Higgs boson production in the four-lepton decay channel

Chapter 3

Definition and validation of the Higgs-to-four-lepton objects with the Ultra-Legacy reconstruction

The Higgs boson decay into four charged leptons ($H \to ZZ \to 4\ell$, where ℓ denotes either an electron e or a muon μ henceforth) is widely acknowledged as the *golden channel* of Higgs physics. This reputation owes to several virtues, including a clear mass peak over a nearly flat background, a large signal-to-background ratio, and a fully reconstructible final state, which benefits from the highly performing lepton reconstruction of the CMS detector. These properties render the four-lepton channel one of the two channels, alongside $H \to \gamma\gamma$, that are suited for the precise measurement of the Higgs boson mass. However, the low branching fraction is the drawback of this channel. For a SM Higgs boson with $m_H = 125$ GeV, the BR is expected to be $BR(H \to 4\ell) = 1.24 \times 10^{-4}$ [27]. Considering detector acceptance and analysis selections, it is anticipated to collect approximately 2 four-lepton events per fb^{-1} of integrated luminosity at a centre-of-mass energy of $\sqrt{s} = 13$ TeV. This corresponds to around 250 events when considering the 138 fb^{-1} of integrated luminosity accumulated during Run 2.

The four-lepton channel has been one of the pillars for the search [114, 115], discovery [19, 20, 116], and characterisation of the Higgs boson since the inception of the CMS and ATLAS physics programmes. Over the years, the properties of the Higgs boson were measured in this decay channel with the Run 1 dataset at centre-of-mass energy of 7 and 8 TeV and with the Run 2 dataset at 13 TeV. These measurements include the determination of its mass, spin and parity [56, 117–121], width [122–125], and tensor structure for interactions with a pair of gauge bosons [56, 121, 123, 126–128]. All the results based on the Run 1 dataset supported that the observed particle was the SM Higgs boson responsible for the electroweak symmetry breaking. This was later further confirmed more robustly with the larger Run 2 dataset. The most precise value of the Higgs boson mass to date, measured by the ATLAS collaboration, is $m_H = 125.11 \pm 0.11$ GeV, obtained from the combination of the $H \to ZZ \to 4\ell$ and $H \to \gamma\gamma$ decay channels from the analysis of the Run 1 and Run 2 datasets [58]. From the analysis of the entire Run 2 dataset, the CMS Collaboration reported the first evidence for the off-shell Higgs boson production in events with a final state of two Z bosons decaying into either four charged leptons or two charged leptons and two neutrinos, with a measured value of the Higgs boson width of $\Gamma = 3.2^{+2.4}_{-1.7}$ MeV [59].

In conjunction with the characterisation outlined above, a comprehensive set of cross sections have been measured. Signal strength modifiers, both inclusive and for each production mode, were among the initial measurements after the discovery of the resonance. They were then measured with higher precision with the various datasets collected over

the years [117, 129, 130]. With the Run 2 dataset, both CMS and ATLAS collaborations provided the first characterisation of the Higgs boson via the STXS framework in the four-lepton channel [130, 131], which was later combined with all other decay channels [132].

The first measurements of fiducial cross sections were made using the Run 1 dataset [133, 134]. Despite the low number of events in the golden channel, Run 1 data enabled the differential measurement of a few variables with coarse binning. The ATLAS collaboration also provided the first combination with the $H \to \gamma\gamma$ channel [135]. All these results were strongly limited by the low number of events and primarily dominated by statistical uncertainty. The larger dataset accumulated during Run 2 allowed the study of the decay channel with higher precision. The first Run 2 data collected in 2016 were exploited for the first measurements at $\sqrt{s} = 13$ TeV [120, 136], later combined with the $H \to \gamma\gamma$ and $H \to b\bar{b}$ channels by the CMS collaboration [137], and with the $H \to \gamma\gamma$ channel by the ATLAS collaboration [138]. The true turning point for the $H \to ZZ \to 4\ell$ channel came with the availability of the entire Run 2 dataset for analysis. The statistical uncertainty was significantly reduced and, for inclusive results, approached a level similar to the systematic uncertainty. The number of differential observables increased, as did their granularity, allowing for an examination of the four-lepton channel with unprecedented precision, marking its entry into a new era of precision measurement. The ATLAS collaboration published the measurement with the entire Run 2 dataset [139] and its combination with the $H \to \gamma\gamma$ channel [135]. This thesis presents the CMS counterpart and the combination is currently ongoing.

Chapters 3, 4, and 5 present the measurement of inclusive and differential fiducial cross sections with the data collected in 2016, 2017, and 2018 of the Run 2 of the LHC with the CMS detector. The result of this analysis is now public in the Journal of High Energy Physics [21]. This initial chapter is dedicated to defining and validating the physics objects necessary for reconstructing the decay of the Higgs boson into four leptons. This analysis makes use of the latest data processing campaign known as Ultra-Legacy (UL), which supersedes the previous campaign, known as End-Of-Year (EOY). The UL campaign represents the final and most refined calibration of objects for Run 2 data. The improvements concern the realignment and recalibration of all sub-detectors, and the re-derivation of corrections for the related physics objects. The simulation was also improved ensuring a more accurate description of the data in terms of dynamic effects and status of the detector. All these actions were undertaken with the aim of enhancing performance, resolution, and the agreement between data and MC. Thus, it is essential to assess the properties of the objects reconstructed with UL and determine whether adjustments to the EOY setup are necessary. This chapter begins with introducing the MC samples and data (Sec. 3.1) that will be used throughout the presentation of the analysis. Electron and muon selections, in order to identify prompt and isolated leptons, are discussed in Sec. 3.2. Leptons can be accompanied by final-state-radiation that must be recovered in order to maintain a high resolution and careful description of the Higgs boson resonance, as discussed in Sec. 3.3. Despite not being part of the distinctive signature of the four-lepton channel, jets play a central role in the analysis, and their selection will be discussed in Sec. 3.4.

3.1 Monte Carlo samples and dataset

Before delving into the specifics of the objects of the $H \to ZZ \to 4\ell$ analysis, it is essential to introduce two fundamentals ingredients without which making an analysis would be impossible: Monte Carlo (MC) samples and the dataset to analyse.

3.1.1 Monte Carlo samples

As explained in Sec. 2.5, MC event generations follow a precise workflow in the CMS collaboration. The part that requires customisation is the first one, the GEN step. For modelling the physics processes needed to describe the $H \to ZZ \to 4\ell$ channel, the GEN step comprises two different sub-steps.

The first one is the simulation of the hard physics process between partons at a specific order in perturbative-QCD[1] (pQCD) and QED. Tables 3.1 and 3.2 show the list of signal and background processes used in the analysis with the corresponding MC generators. The description of the SM Higgs boson production is obtained at NLO using the `POWHEG 2.0` generator [140–142] for all five main production modes: ggH [143], VBFH [144], WH, ZH, and ttH [145]. For the associated production with W and Z bosons, the `MiNLO` extension of `POWHEG` is used [146]. In order to profit from the latest theoretical computations, the ggH sample is reweighed to match the NNLOPS predictions [147]. A dedicated weight is computed in bins of transverse momentum and number of associated jets, by taking the ratio of the NNLOPS and POWHEG cross sections.

For signals, `POWHEG` only handles the production of the Higgs boson; its decay is simulated equally for all production modes using `JHUGen 7.0.2` [148–152]. As far as backgrounds are concerned, the $q\bar{q} \to ZZ$ contribution is simulated at NLO using `POWHEG 2.0` [153], while the $gg \to ZZ$ process is generated at LO using `MCFM 7.0.1` [154]. Despite not being used directly to model the signal and background, three MC samples are extensively employed to cross-check, validate, and study the reducible background: Drell-Yan, WZ, and $t\bar{t}$; the first two generated with `MadGraph5_aMC@NLO`[155], and the latter with `POWHEG 2.0`. These three samples are generated at NLO, and an additional Drell-Yan sample is generated at LO.

Table 3.1: List of signal MC samples. $\ell = e, \mu, \tau$, and X stands for other particles that can be produced in the final state, e.g., neutrinos or quarks from the decay of the associated particles. The last column quotes the values at $m_H = 125$ GeV and $\sqrt{s} = 13$ TeV.

Process	MC generator	$\sigma \times$ BR [pb]
$gg \to H \to ZZ \to 4\ell$ **(ggH)**	`POWHEG 2.0 + JHUGen V7`	1.34×10^{-2}
$qq' \to Hqq' \to ZZqq' \to 4\ell qq'$ **(VBFH)**	`POWHEG 2.0 + JHUGen V7`	1.04×10^{-3}
$q\bar{q} \to \mathbf{ZH} \to ZZZ \to 4\ell + X$	`POWHEG 2.0 (MiNLO HZJ) + JHUGen V7`	0.24×10^{-3}
$q\bar{q} \to \mathbf{W^+H} \to W^+ZZ \to 4\ell + X$	`POWHEG 2.0 (MiNLO HWJ) + JHUGen V7`	0.23×10^{-3}
$q\bar{q} \to \mathbf{W^-H} \to W^-ZZ \to 4\ell + X$	`POWHEG 2.0 (MiNLO HWJ) + JHUGen V7`	0.15×10^{-3}
$gg \to \mathbf{t\bar{t}H} \to t\bar{t}ZZ \to 4\ell + X$	`POWHEG 2.0 + JHUGen V7`	0.14×10^{-2}

After the generation of the hard scattering process, the following step is the simulation of the parton shower, hadronisation of quarks and gluons, hadrons decays, and the underlying event. This step is accomplished by interfacing the previously generated MC samples with `PYTHIA`. Version 8.230 [156] is used for the three data-taking years with the CUETP8M1 tune [157] for 2016 and the CP5 tune [158] for 2017 and 2018. PDFs are taken from the NNPDF3.0 set [159] for the three data taking periods.

The response of the CMS detector, i.e., the SIM step introduced in Sec. 2.5, is modelled using a detailed simulation of the detector based on `GEANT4` [160, 161].

[1] The strong coupling constant α_s is relatively large at low energies, meaning that the usual approach of making perturbative computations in quantum field theory is not valid anymore. For instance, this is the regime of the hadronisation process, which cannot be easily theoretically treated because of that. However, in the high-energy limit (or short-distance interaction limit) the running of α_s predicts a sufficiently small value of the coupling constant, making it possible for a reliable computation using perturbation theory. The stress on *perturbative* is usually done only for QCD, as the QED coupling constant is small enough at all accessible energies.

Table 3.2: List of background MC samples. $\ell = e, \mu, \tau$. The last column quotes the values at $m_{\mathrm{H}} = 125$ GeV and $\sqrt{s} = 13$ TeV. The processes in the upper section are used to model the irreducible background, whereas the ones in the lower section are used to study the reducible background. The process "$\mathrm{Z}/\gamma^* \to 2\ell$ + jets" is widely known as *Drell-Yan*.

Process	MC generator	$\sigma \times$ BR [pb]
$\mathrm{q\bar{q}} \to \mathrm{ZZ} \to 4\ell$	`POWHEG 2.0`	1.26
$\mathrm{gg} \to \mathrm{ZZ} \to 4e$	`MCFM`	1.6×10^{-3}
$\mathrm{gg} \to \mathrm{ZZ} \to 4\mu$	`MCFM`	1.6×10^{-3}
$\mathrm{gg} \to \mathrm{ZZ} \to 4\tau$	`MCFM`	1.6×10^{-3}
$\mathrm{gg} \to \mathrm{ZZ} \to 2e2\mu$	`MCFM`	3.2×10^{-3}
$\mathrm{gg} \to \mathrm{ZZ} \to 2e2\tau$	`MCFM`	3.2×10^{-3}
$\mathrm{gg} \to \mathrm{ZZ} \to 2\mu2\tau$	`MCFM`	3.2×10^{-3}
$\mathrm{Z}/\gamma^* \to 2\ell$ + jets	`MadGraph5_aMC@NLO`	6225.4
$\mathrm{WZ} \to 3\ell + \nu$	`MadGraph5_aMC@NLO`	5.3
$\mathrm{t\bar{t}} \to 2\ell + 2\nu + 2\mathrm{b}$	`POWHEG 2.0`	87.3

MC reweighing

MC samples need to be reweighed and normalised to be compared with data and to model the expected signal and background accurately. The fundamental unit of the per-event normalisation factor is the following:

$$w_{\mathrm{gen}} \cdot \frac{L_{\mathrm{int}} \cdot (\sigma \cdot \mathrm{BR})_{\mathrm{th}}}{\sum_{\mathrm{events}} w_{\mathrm{gen}}} \tag{3.1}$$

Here, w_{gen} is the event weight of the MC generator, L_{int} is the integrated luminosity of the data-taking period, $(\sigma \cdot \mathrm{BR})_{\mathrm{th}}$ denotes the product between the theoretical cross section and branching ratio at which the sample is to be normalised, and the denominator is the sum over all MC weights of all events. This term represents the core of every MC weight and normalises the sample to a given luminosity and cross section.

Additional weights may be introduced on top of Eq. 3.1 to account for residual differences between data and MC. While some additional weights can vary in origin and may depend on the specific sample or data-taking period, others are general and always included in CMS analyses. The general structure of the MC event weight, combined with Eq. 3.2, can be written as follows:

$$w_{\mathrm{event}} = w_{\mathrm{gen}} \cdot \frac{L_{\mathrm{int}} \cdot (\sigma \cdot \mathrm{BR})_{\mathrm{th}}}{\sum_{\mathrm{events}} w_{\mathrm{gen}}} \cdot \mathrm{SF}_{4\ell} \cdot w_{\mathrm{PU}} \tag{3.2}$$

where $\mathrm{SF}_{4\ell}$ is the per-event scale factor, which will be explained in Sec. 3.2.4, and w_{PU} is the PU weight. Despite being simulated as explained in Sec. 2.5, the simulated PU profile does not match the one observed in the data. Hence, the correction is applied to account for this discrepancy. The simulated PU is based on the expected PU profile, and this estimation is done before the end of the yearly data-taking period.

Equation 3.2 can be expanded with additional weights, tackling specific issues. For instance, the NNLOPS weight introduced previously, which is only applied to the ggH sample, and the *L1 prefiring weight*, which is applied to 2016 and 2017 MC samples to correct for an issue in the interface of the ECAL and the L1 trigger. Due to the improper propagation to the L1 of the gradual timing shift of the ECAL, a significant fraction of high-η events were mistakenly associated with the previous bunch crossing. As the L1

trigger prevents two consecutive bunch crossings to fire, the consequence of this issue was the self-veto of the event itself. A dedicated corrective weight was computed by parametrising the probability for an event to cause prefiring as a function of its transverse momentum and pseudorapidity.

3.1.2 Dataset

The dataset employed in this $H \to ZZ \to 4\ell$ analysis was collected during the Run 2 of the LHC, spanning the years 2016, 2017, and 2018, corresponding to a validated integrated luminosity of 138 fb^{-1}. As detailed in Sec. 2.5, the data is stored in PDs, each associated with a unique set of HLT paths. Inclusion in a particular PD requires an event to have triggered at least one of the HLT paths associated with that specific PD. While an HLT path can only belong to a single PD, an event can be present in multiple PDs, as it may have triggered different paths. The PDs utilised in the $H \to ZZ \to 4\ell$ analyses includes the *DoubleMuon*, *MuonEGamma*, *SingleMuon*, *EGamma* (only for 2018), *DoubleEGamma* (not for 2018), and *SingleElectron* (not for 2018) datasets. The names of these datasets are self-explanatory, providing an insight into the set of HLT paths included in each PD. The *EGamma* sample is introduced for the 2018 period and merges the previous *DoubleEGamma* and *SingleElectron* in the same PD. For the sake of the efficiency of the reconstruction algorithm, the analyses are usually limited to a relevant subset of the HLT paths. A summary of the HLT paths and PDs can be found in Tab. 3.3, 3.4, and 3.5.

In order to avoid duplications of events originating from different PDs, an event is taken:

- from the EGamma or DoubleEGamma PD if the event passes DiElectron or TripleElectron triggers;
- from the DoubleMuon PD if the event passes DiMuon or TripleMuon trigger and fails the DiElectron and TripleElectron triggers;
- from the MuonEGamma PD if the event passes MuonElectron or MuonDiElectron or DiMuonElectron triggers and fails DiElectron, TripleElectron, SingleElectron, DiMuon, and TripleMuon triggers;
- from the EGamma or SingleElectron PD if the event passes SingleElectron triggers and fails MuonElectron, MuonDiElectron, DiMuonElectron, DiElectron, TripleElectron, SingleElectron, DiMuon, and TripleMuon triggers;
- from the SingleMuon PD if the event passes SingleMuon triggers and fails all aforementioned triggers.

The minimal lepton p_T thresholds required in these triggers are reported in Tab. 3.6.

3.2 Leptons

The distinctive feature of the $H \to ZZ \to 4\ell$ channel is the presence of four prompt and isolated leptons. A prompt lepton originates directly from the PV of the collision. As the Higgs boson lifetime is too short to see a significant displacement of the decay vertex, the leptons should be associated with the PV of the event. Instead, an isolated lepton is formally defined as one well-separated from other particles in the event. Searching for isolated leptons is one of the most powerful handles for selecting electrons and muons produced in the electroweak decay of massive particles, like the Z boson in our case, and rejecting leptons produced within jets or from the in-flight decay of charged light-flavoured mesons, such as pions and kaons. Section 2.3 outlined the PF reconstruction of electrons

Table 3.3: Primary datasets and corresponding HLT paths used in 2016 data.

HLT path	primary dataset
HLT_Ele17_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	DoubleEG
HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	DoubleEG
HLT_DoubleEle33_CaloIdL_GsfTrkIdVL	DoubleEG
HLT_Ele16_Ele12_Ele8_CaloIdL_TrackIdL	DoubleEG
HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL	DoubleMuon
HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL	DoubleMuon
HLT_TripleMu_12_10_5	DoubleMuon
HLT_Mu8_TrkIsoVVL_Ele17_CaloIdL_TrackIdL_IsoVL	MuonEG
HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL	MuonEG
HLT_Mu17_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL	MuonEG
HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL	MuonEG
HLT_Mu23_TrkIsoVVL_Ele8_CaloIdL_TrackIdL_IsoVL	MuonEG
HLT_Mu8_DiEle12_CaloIdL_TrackIdL	MuonEG
HLT_DiMu9_Ele9_CaloIdL_TrackIdL	MuonEG
HLT_Ele25_eta2p1_WPTight	SingleElectron
HLT_Ele27_WPTight	SingleElectron
HLT_Ele27_eta2p1_WPLoose_Gsf	SingleElectron
HLT_IsoMu20 OR HLT_IsoTkMu20	SingleMuon
HLT_IsoMu22 OR HLT_IsoTkMu22	SingleMuon

Table 3.4: Primary datasets and corresponding HLT paths used in 2017 data.

HLT path	primary dataset
HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_*	DoubleEG
HLT_DoubleEle33_CaloIdL_GsfTrkIdVL	DoubleEG
HLT_Ele16_Ele12_Ele8_CaloIdL_TrackIdL	DoubleEG
HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8	DoubleMuon
HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass8	DoubleMuon
HLT_TripleMu_12_10_5	DoubleMuon
HLT_TripleMu_10_5_5_D2	DoubleMuon
HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL	MuonEG
HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ	MuonEG
HLT_Mu12_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ	MuonEG
HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_DZ	MuonEG
HLT_DiMu9_Ele9_CaloIdL_TrackIdL_DZ	MuonEG
HLT_Mu8_DiEle12_CaloIdL_TrackIdL	MuonEG
HLT_Mu8_DiEle12_CaloIdL_TrackIdL_DZ	MuonEG
HLT_Ele35_WPTight_Gsf_v*	SingleElectron
HLT_Ele38_WPTight_Gsf_v*	SingleElectron
HLT_Ele40_WPTight_Gsf_v*	SingleElectron
HLT_IsoMu27	SingleMuon

Table 3.5: Primary datasets and corresponding HLT paths used in 2018 data.

HLT path	primary dataset
HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_v*	DoubleEG
HLT_DoubleEle25_CaloIdL_MW_v*	DoubleEG
HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8_v*	DoubleMuon
HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_v*	MuonEG
HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ_v*	MuonEG
HLT_Mu12_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ_v*	MuonEG
HLT_DiMu9_Ele9_CaloIdL_TrackIdL_DZ_v*	MuonEG
HLT_Ele32_WPTight_Gsf_v*	SingleElectron
HLT_IsoMu24_v*	SingleMuon

and muons in a very general way. This section, picking up these objects as they were left at the end of the reconstruction process, will focus on their identification and isolation, which is less general and strictly related to the specific physics analysis. With identification, one usually refers to a series of requirements to enhance the purity of the selected objects.

Leptons in the $H \to ZZ \to 4\ell$ channel present unique peculiarities and challenges.

Table 3.6: Thresholds applied on the p_{T} of the leading/subleading leptons in each data-taking period for the main dielectron (e/e), dimuon (μ/μ), and electron-muon ($e/\mu, \mu/e$) HLT algorithms.

	e/e (GeV)	μ/μ (GeV)	e/μ, μ/e (GeV)
2016	17/12	17/8	17/8, 8/23
2017	23/12	17/8	23/8, 12/23
2018	23/12	17/8	23/8, 12/23

For a Higgs boson with $m_{\mathrm{H}} \simeq 125$ GeV, at least one of the Z bosons is off-shell, giving rise to relatively low-p_{T} leptons. Considering experimental reconstruction efficiencies, the typical invariant mass of these off-shell Z bosons is only about 30 GeV, and the leptons originating from them have a median p_{T} of 15 GeV, in contrast to the median p_{T} of 40 GeV for leptons stemming from the on-shell Z boson. Additionally, given the low BR and the fact that the per-lepton efficiency propagates to the event selection with the power of four, it is imperative to maximize the lepton reconstruction efficiency. The CMS detector was designed with these challenges in mind. However, a customized selection procedure can further enhance the purity of objects for four-lepton analyses. This is accomplished through tailored selection criteria that often implement more relaxed requirements than other analyses employing leptons. The following sections present how this task is effectively carried out.

3.2.1 Lepton preselection

Some primary selections are applied starting from the electrons and muons candidates provided by the PF reconstruction. The extension of the silicon tracker and the outer muon chambers dictates the pseudorapidity range to be $|\eta^e| < 2.5$ and $|\eta^\mu| < 2.4$. As the introduction explains, the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ analysis needs to work with very low-p_{T} leptons. However, going down in p_{T} poses some difficulties, such as a large number of fake leptons and difficulties in reconstructing tracks, measuring momentum, and computing efficiencies. For these reasons, electrons and muons are reconstructed only if $p_{\mathrm{T}}^e > 7$ GeV and $p_{\mathrm{T}}^\mu > 5$ GeV, respectively.

Non-prompt leptons, i.e., not compatible with the PV, are discarded by checking their compatibility with the PV of the event. They can originate from long-lived hadrons, muons from cosmic rays, or electrons from photon conversion in the detector and beam pipe. The conditions to be satisfied are $d_{xy} < 0.5$ cm and $d_z < 1$ cm, which denote the impact parameter in the transverse and longitudinal plane, respectively. Leptons satisfying all these conditions are defined as *soft leptons*.

An additional requirement to suppress even further the presence of non-prompt leptons is introduced through the Significance of the Impact Parameter (SIP):

$$\mathrm{SIP} = \frac{\mathrm{IP}_{3\mathrm{D}}}{\sigma_{\mathrm{IP}_{3\mathrm{D}}}} < 4 \tag{3.3}$$

It is formally defined as the 3D impact parameter between the lepton track and the PV divided by its uncertainty. Soft leptons satisfying the SIP cut are defined as *loose leptons*.

These selections constitute minimal requirements for leptons to be considered either in the signal or control regions. However, leptons must undergo further tighter selections to be considered as the signature of the golden channel, and, even though the pattern is basically the same, there are some differences depending on the flavour of the lepton.

3.2.2 Lepton identification and isolation

Numerous objects in the detector can mimic the electron and muon signatures, resulting in a larger fake rate and a reduced purity of the selected objects. Misreconstructed muons can stem from remnants of hadron showers that penetrate through the calorimeters and reach the outer muon system, or from random matches between a hadron track in the inner tracker and a segment in the muon system. Differently, misreconstructed electrons can arise from light-flavour hadrons, photon conversion, and jet activity. It is thus evident that lepton candidates must satisfy a set of criteria to minimise the presence of such objects. This procedure is referred to as *lepton identification*. As anticipated in the introduction of this section, in addition to genuine electrons and muons, the $H \to ZZ \to 4\ell$ analysis requires prompt leptons that can be associated to the decay of a Z boson. The main tool for attaining this involves checking whether the leptons are isolated. A lepton passing the identification and isolation requirements is referred to as *tight lepton*.

Electron identification and isolation

Electrons are identified and isolated in one single step using an eXtreme Gradient Boosting (XGBoost) [162]. It implements a series of BDTs and uses 24 observables sensitive to the properties of electrons to identify prompt and isolated electrons, especially in the low-p_T regime. The inputs include observables related to the SC shape in the ECAL, the reconstruction of electron tracks in the tracker, the geometrical and kinematic matching between the SC in the ECAL and the track in the tracker, the electron isolation, and variables to increase the PU resilience. The list of variables is detailed in Tab. 3.7.

The model is trained to distinguish between prompt and isolated electrons linked to the decay of a Z boson and fake leptons stemming from jets and hadronic decays. For the BDT training process, only the Drell-Yan sample is used, serving as both the source of signal and background leptons. Signal leptons are identified with a geometrical matching of the reconstructed leptons and the generator information. A priori, the signal leptons should be sourced from a signal sample, such as ggH. It has been demonstrated that training the BDT with the combination of ggH/Drell-Yan is equivalent to just using Drell-Yan. In both cases, signal leptons come from the decay of a Z boson.

The training is performed independently for each of the three data-taking periods to account for the different detector conditions and material budgets. Additionally, the training is done in three pseudorapidity bins, $|\eta| < 0.8$, $0.8 < |\eta| < 0.1.479$, $|\eta| > 1.479$. To further enhance the performance for low-p_T electrons, a specific low-p_T bin ($5 < p_T < 10$ GeV) is introduced. Figure 3.1 shows the BDT score distributions for signal and background, offering insight into the classifier's performance. The Receiver Operating Characteristic (ROC) curves for the low-p_T bin in 2017, comparing UL and EOY performance, are reported in Fig. 3.2.

The BDT working point is chosen to attain the same signal efficiency as EOY. Hence, a value between 96-98% for bins with $p_T > 10$ GeV for a background efficiency of 3-4% at low/central pseudorapidity and 7-8% at large pseudorapidity. However, in the low-p_T bins, achieving the same performance becomes challenging. Here, in order to keep the background efficiency at an acceptable value of roughly 4%, the signal efficiency is forced to decrease to 75-80%. In 2018 and 2017, UL demonstrates a general improvement compared to EOY, with a lower background efficiency that decreases up to 9%, while maintaining the same signal efficiency. Conversely, in 2016, the performance of the UL BDT slightly deteriorates. This effect is more pronounced for $p_T^e > 10$ GeV and can be attributed to the UL recalibration, which introduces a larger overlap between signal and background for certain BDT variables. Despite this, the BDT performance at the selected working point is found to be in agreement between EOY and UL and provides a satisfactory performance.

Table 3.7: List of input variables for the BDT employed in the analysis for electron identification and isolation. When defining the observables related to the shape of the SC in the ECAL, it is useful to introduce the quantity $E_{x\times y}$. It is defined as the energy in the $x \times y$ block of crystals centred on the crystal seed of the SC, where x is defined along η and y along ϕ.

	Shape of the SC in the ECAL
$\sigma_{i\eta i\eta}$, $\sigma_{i\phi i\phi}$	Variance of the log-energy-weighed distribution of crystal energies along η and ϕ
Δ_η, Δ_ϕ	Width of the SC along η and ϕ
H/E	H is the energy deposited in the HCAL in a cone of radius $\Delta R = 0.15$ around the SC direction, and E is the energy of the SC
Circularity	$(E_{5\times5} - E_{5\times1})/E_{5\times5}$
R_9	$E_{3\times3}/E_{\mathrm{SC}}$
$E_{\mathrm{ES}}/E_{\mathrm{raw}}$	Contribution of the energy deposited in the ECAL preshower (E_{ES}) to the uncorrected SC energy (E_{raw})
	Reconstruction of electron tracks
$\mathrm{f}_{\mathrm{brem}}$	Fraction of bremsstrahlung emission measured by the momentum loss of the GSF track
N_{KF}, N_{GSF}	Number of hits of the KF and GSF tracks
χ^2_{KF}, χ^2_{GSF}	Reduced χ^2 of the KF and GSF fit
$N_{missHits}$	Number of missing hits in the inner tracker, for a prompt electron no missing hits are expected
P_{conv}	Conversion probability obtained from the χ^2 of the conversion fit
	Geometrical and kinematic matching between the SC in the ECAL and the track in the tracker
$E_{\mathrm{SC}}/p_{\mathrm{in}}$	Ratio of SC energy (E_{SC}) and inner track momentum (p_{in})
$E_{\mathrm{PFclus}}/p_{\mathrm{out}}$	Ratio of the energy of the closest ECAL PF cluster to the extrapolated track at the surface of the calorimeter (E_{PFclus}) and outer track momentum (p_{out})
$1/E - 1/p$	Agreement between the SC energy (E) and the electron track momentum (p)
$\Delta\eta_{\mathrm{in}}$, $\Delta\phi_{\mathrm{in}}$	Distance in η and ϕ between the energy-weighted SC barycentre and the inner track extrapolation: $\Delta\eta_{\mathrm{in}} = \lvert\eta_{\mathrm{SC}} - \eta_{in}\rvert$, $\Delta\phi_{\mathrm{in}} = \lvert\phi_{\mathrm{SC}} - \phi_{in}\rvert$
$\Delta\eta_{\mathrm{seed}}$	Distance between seed cluster and outer track extrapolation to calorimeter $\Delta\eta_{\mathrm{seed}} = \lvert\eta_{\mathrm{seed}} - \eta_{\mathrm{out}}\rvert$
	Electron isolation
$\sum_{\mathrm{h}^\pm} p_{\mathrm{T}}$	p_{T} sum of charged PF hadrons from primary vertex in a cone of radius $\Delta R = 0.3$
$\sum_{\mathrm{h}^0} p_{\mathrm{T}}$	p_{T} sum of neutral PF hadrons in a cone of radius $\Delta R = 0.3$
$\sum_{\gamma} p_{\mathrm{T}}$	p_{T} sum of PF photons in a cone of radius $\Delta R = 0.3$
	PU resilience
ρ	Mean energy density of the event

A summary of the BDT performance is outlined in Tab. 3.8.

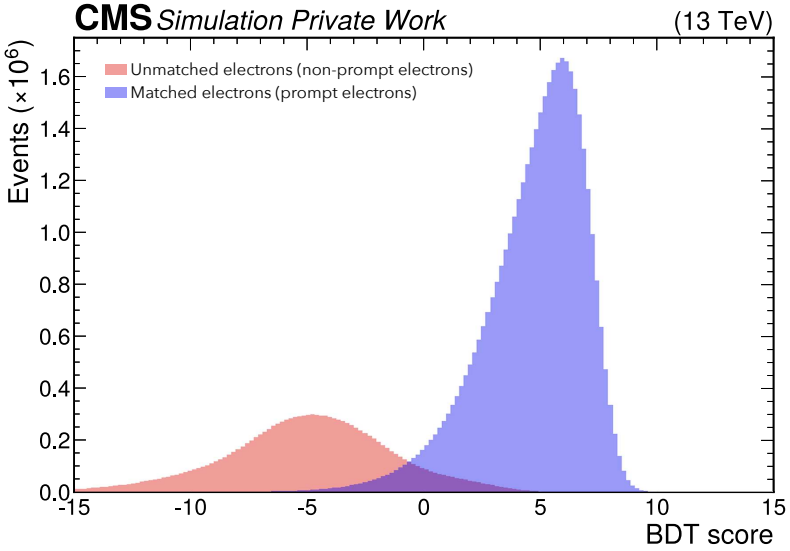

Figure 3.1: BDT score for 2018 data inclusively in p_{T} and η. The blue histogram shows the distribution for signal-like events, i.e. matched prompt electrons, whereas the red histogram shows the distribution for background-like events, i.e. unmatched and non-prompt leptons.

Table 3.8: Performance of the BDT for electron identification and isolation. The signal and background efficiency is reported for every $|\eta^e|$ and p_{T}^e bin. The choice of the working point for UL is done by requiring the same signal efficiency as EOY. The background efficiency is reported for UL and in parentheses for EOY.

$\lvert\eta^e\rvert$ bin	p_{T}^e bin	Signal efficiency	Background efficiency		
			2016	2017	2018
< 0.8	$[5, 10]$	81.64%	3.68% (3.93%)	5.26% (5.66%)	5.45% (5.66%)
	> 10	97.45%	2.47% (2.17%)	2.96% (3.26%)	2.92% (3.28%)
$[0.8, 1.479]$	$[5, 10]$	80.31%	3.64% (3.63%)	4.86% (4.74%)	4.75% (4.69%)
	> 10	96.68%	3.35% (2.75%)	4.04% (4.05%)	3.81% (4.12%)
> 1.479	$[5, 10]$	74.37%	3.56% (3.06%)	3.29% (3.59%)	3.42% (3.26%)
	> 10	96.62%	8.67% (7.66%)	7.56% (8.10%)	7.29% (8.06%)

Muon identification

As seen in Sec. 2.3.2, several muon IDs can be defined starting from the reconstructed standalone, global, and tracker muons within the PF approach. In the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ analyses, the identification criteria that are applied are the following:

- if $p_{\mathrm{T}}^{\mu} < 200$ GeV, the *loose ID* is applied;
- if $p_{\mathrm{T}}^{\mu} > 200$ GeV, either *loose ID* or the *tracker high-p_{T} ID* must be satisfied.

The difference in ID between low-p_{T} and high-p_{T} muons is dictated by the need to increase the signal efficiency for high-mass searches. If one thinks about the decay of a heavy resonance $\mathrm{X} \to \mathrm{ZZ} \to 4\ell$, the leptons in the laboratory frame will be nearly collinear and

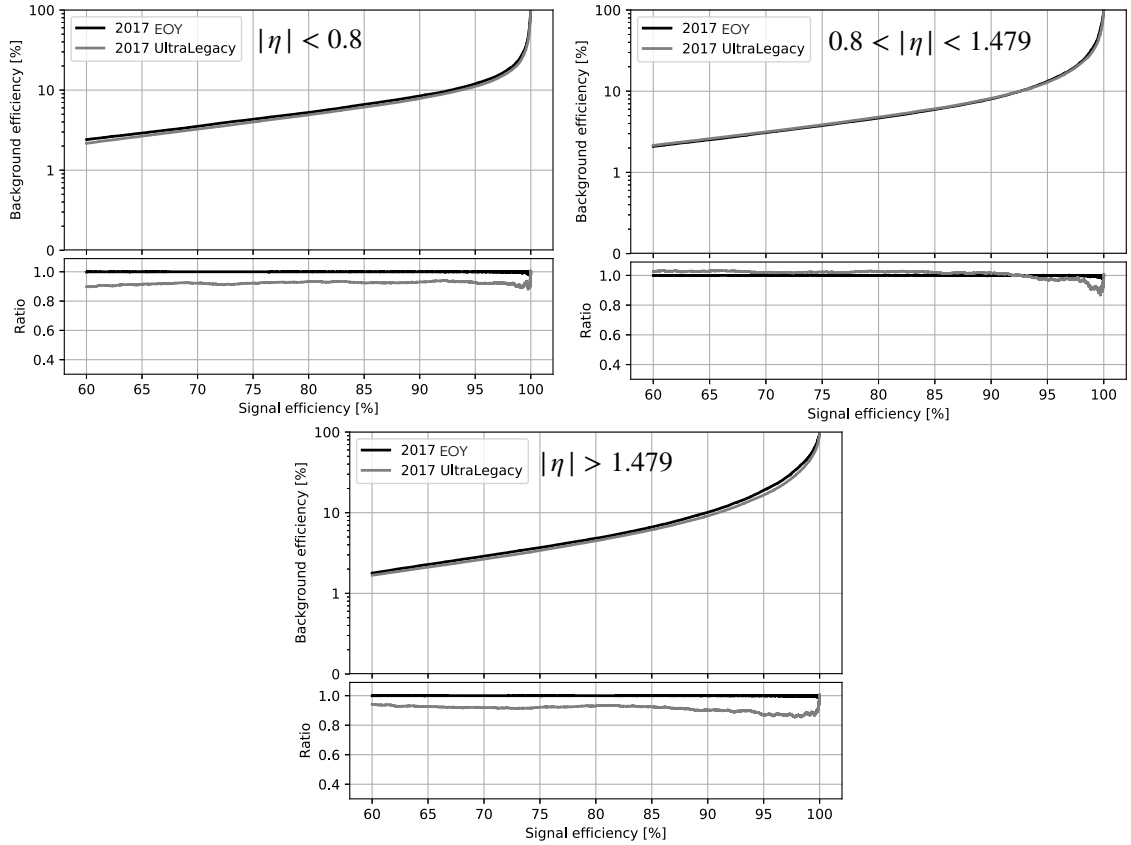

Figure 3.2: ROC curves for the BDT used for the electron identification and isolation in 2018 in the low-p_{T} bin $5 < p_{\mathrm{T}}^{e} < 10$ GeV: $|\eta^{e}| < 0.8$ (left), $0.8 < |\eta^{e}| < 1.479$ (centre), $|\eta| > 1.479$ (right). The black lines represent the curve for EOY calibration and the grey lines for UL calibration. The bottom panels show the ratio to the EOY curve.

the loose muon ID features a loss in efficiency when the separation between the two muons is $\Delta R < 0.4$. The tracker high-p_{T} ID is specifically optimised for the reconstruction of boosted Z bosons, where two close-by muons may appear. The requirements for this ID are summarised as follows:

- the muon candidate is reconstructed as a tracker muon;
- there must be muon segments in at least two muon chambers in order to suppress the accidental track-to-segment matches and the background from leakages from the HCAL;
- good momentum resolution $\sigma_{p_{\mathrm{T}}}/p_{\mathrm{T}} < 0.3$;
- hits are required in at least 6 layers in the inner tracker in order to guarantee a good p_{T} measurement, for which a minimal number of points in the tracker is needed;
- at least one hit in the pixel tracker;
- vertex compatibility of the tracker track in the transverse plane $d_{xy} < 2$ mm;
- vertex compatibility of the tracker track in the longitudinal plane $d_{z} < 2$ mm;

The conditions on the pixel hits and the vertex compatibility are mainly to reduce the contributions from cosmic muons, muons from in-flight decays, and tracks from PU.

An additional *ghost-cleaning* step is introduced for muons to fix the scenario where a single muon is incorrectly reconstructed as two or more muons. There are mainly two classes of ghost muons: the *split tracks*, where the tracker track of a muon is broken in two segments and both are identified as muons; the *mismatched tracks*, where the track of another particle in the event is also found to be compatible with the same muon hits in the outer chambers. To cope with this issue, tracker muons that are not global muons are discarded, and if two muons share more than 50% of their track segments, the one with the lowest track quality is removed.

Muon isolation

While for electrons, the isolation is included directly in the BDT for their identification, for muons, a cut-based selection is introduced based on the PF relative isolation. This selection requires the energy flow in the vicinity of a lepton to be below a certain threshold:

$$\mathcal{I}_\ell(\Delta R) = \frac{1}{p_{\mathrm{T}}^{\ell}} \left(\sum_{h^\pm} p_{\mathrm{T}} + \sum_{h^0} p_{\mathrm{T}} + \sum_{\gamma} p_{\mathrm{T}} \right) \tag{3.4}$$

The lepton isolation is based on the p_{T} sum of charged hadrons ($h^\pm$), neutral hadrons (h^0), and photons around the direction of the lepton in a cone with radius ΔR. However, the presence of PU can spoil this evaluation of the lepton isolation due to the presence of additional particles inside the cone. To mitigate this effect, Eq. 3.4 is modified to remove spurious contributions as follows:

$$\mathcal{I}_\ell(\Delta R) = \frac{1}{p_{\mathrm{T}}^{\mu}} \left(\sum_{h^\pm,\mathrm{PV}} p_{\mathrm{T}} + \max\left[0, \sum_{h^0} p_{\mathrm{T}} + \sum_{\gamma} p_{\mathrm{T}} - 0.5 \sum_{h^\pm,\mathrm{PU}} p_{\mathrm{T}}\right] \right) \tag{3.5}$$

Here, only the charged hadrons linked to the PV vertex are considered. While compatibility with the PV can be reliably established for charged hadrons, neutral particles lack clear PV association due to the absence of hits in the tracker. It is estimated from simulation that the contribution of neutral particles from PU to the isolation cone is approximately half that of charged particles. Therefore, the p_{T} sum of charged hadrons identified as PU is subtracted from the neutral components, with a scaling factor of 0.5 applied.

In the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ analyses, muons are considered isolated if $\mathcal{I}_\mu(0.3) < 0.35$, where the sum over photons excludes the ones associated with the final-state-radiation (Sec 3.3). The value of the radius of the cone and the isolation threshold were optimised during Run 1. It provides performance that is comparable with the BDT isolation of electrons.

3.2.3 Lepton calibration

In the PF approach, the electron momentum is determined by combining information from the ECAL and the inner tracker, while for muons, it is obtained by combining information from the inner tracker and the outer muon chambers. MC samples are generated passing as inputs the best knowledge of the detector conditions and supposedly reproduce faithfully the output of the detector. However, many effects are difficult to model, such as imperfections in the detector materials, effects due to the strong magnetic field, and the time evolution of the transparency of ECAL crystals, pedestals, and noise. As a result, a mismatch when comparing data and MC is unavoidable and should be removed. The lepton energy scale and resolution are corrected using a well-known dilepton resonance and computing corrections as a function of certain kinematic variables and time.

Electrons

A multistep, multivariate procedure is implemented to correct the electron energy scale in data, and to determine the parameters of Gaussian smearing to be applied to showers in simulated events so as to reproduce the energy resolution seen in data [91]. The electron energy scale and resolution are obtained from a fit with a Crystal Ball function to the m_{ee} distribution, where the former is given by the central value and the latter by the width of the fitted function. The calibration is carried out using a pseudo-random Gaussian smearing procedure to the MC electron energies as a function of p_{T}^{e} and $|\eta^{e}|$. The data spectrum and the MC distribution after the smearing procedure, both for EOY and UL, are shown in Fig. 3.3 for the three data-taking periods considered in this analysis. The improved UL calibration can be noted by the shift of the electron energy scale (400 MeV in 2018, 300 MeV in 2016), getting closer to the world-average Z boson mass [26], and the narrower width of the resonance (5−7% improvement). This is due to many improvements, like better noise mitigation of ECAL using proper PF recHit thresholds, re-derivation of PF ECAL cluster calibration, re-derivation of the energy regression, and so forth. Although there are some discrepancies in the tails, which are tamed in the high-mass region by the UL reconstruction, the agreement is overall satisfactory. The residual discrepancies are likely due to the ageing effects of the crystals, and it has a marginal impact on the analysis and will be covered by the related systematic uncertainties that will be included.

Muons

Muon corrections are computed using the combination of the $J/\Psi \to \mu^{+}\mu^{-}$ and $\mathrm{Z} \to \mu^{+}\mu^{-}$ resonances. The muon energy scale and resolution are corrected with a dedicated method, presented in [163] and usually referred to as *Rochester method*. The procedure to measure the muon scale and resolution is the same as for electrons, and the resulting distributions are shown in Fig. 3.4. Unlike for electrons, the UL reconstruction does not bring improvements for muons, as the EOY reconstruction was already satisfactory.

3.2.4 Lepton efficiency

The previous section showed how the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ analysis tries to obtain the highest lepton efficiency possible. However, maximising the lepton efficiency is not the only concern. It is imperative to meticulously evaluate the lepton efficiency both in data and simulation with the same procedure. This procedure is indispensable for accounting for any discrepancies that might otherwise introduce bias into the final results. Once the lepton efficiency is computed with data ($\varepsilon_{\mathrm{data}}$) and MC ($\varepsilon_{\mathrm{MC}}$), the ratio of the two defines the per-lepton *scale factor*:

$$\mathrm{SF}_{\ell}(p_{\mathrm{T}}^{\ell}, \eta^{\ell}) = \frac{\varepsilon_{\mathrm{data}}(p_{\mathrm{T}}^{\ell}, \eta^{\ell})}{\varepsilon_{\mathrm{MC}}(p_{\mathrm{T}}^{\ell}, \eta^{\ell})} \tag{3.6}$$

The estimation of the lepton efficiency is commonly carried out in bins of p_{T} and η to account for possible kinematic dependencies. From the per-lepton scale factor, the per-event scale factor is computed as:

$$\mathrm{SF}_{4\ell} = \prod_{\ell=1}^{4} \mathrm{SF}_{\ell}(p_{\mathrm{T}}^{\ell}, \eta^{\ell}) \tag{3.7}$$

This is the scale factor already encountered in Sec. 3.1.1, which is used to reweigh simulated samples to match with data.

Efficiencies are commonly evaluated with the Tag-and-Probe method (TnP) that allows the measurement directly from data. It exploits a well-known di-lepton resonance, such as

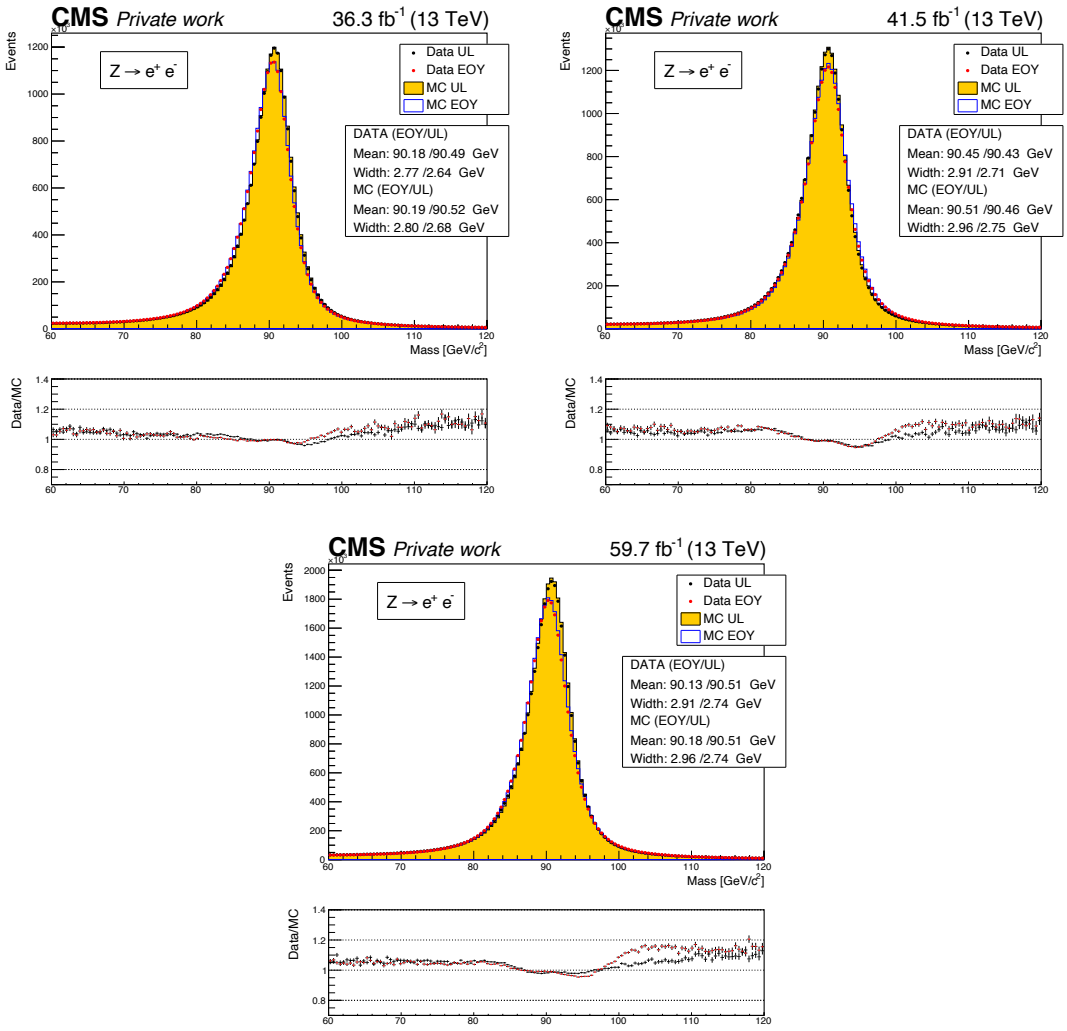

Figure 3.3: Comparison in data and MC of the invariant Z boson mass in the e^+e^- final state between EOY and UL datasets. The MC simulations include Drell-Yan and $t\bar{t}$. The distributions are fitted with a double-sided Crystal Ball to evaluate the peak position (electron scale) and the distribution width (electron resolution). The bottom panels display the data/MC ratio. The distributions are shown for the three data-taking periods: 2016 (top left), 2017(top right), and 2018 (bottom).

the Z or the J/Ψ peak, which provides an unbiased source of leptons. As can be deduced from the name, two objects are involved, a *tag* and a *probe*. The tag is a lepton passing tight selections to ensure a very low fake rate and high sample purity. It is used to select the type of particle to investigate, electrons or muons in our specific case. Once the tag is selected, the next step is to *probe* the efficiency by searching for another type of particle within the same event, passing as loose selections as possible. As it should be paired with the tag, the probe is sought to be of the same flavour and opposite sign in a mass window around the resonance under study. The efficiency is then estimated as:

$$\varepsilon = \frac{N_P}{N_P + N_F} \tag{3.8}$$

where N_P and N_F are the number of probes passing and failing the selection to investigate. The collection of probes is not a pure source of leptons, it is generally polluted by the presence of QCD background, especially at low-p_{T}. This implies the need to fit the

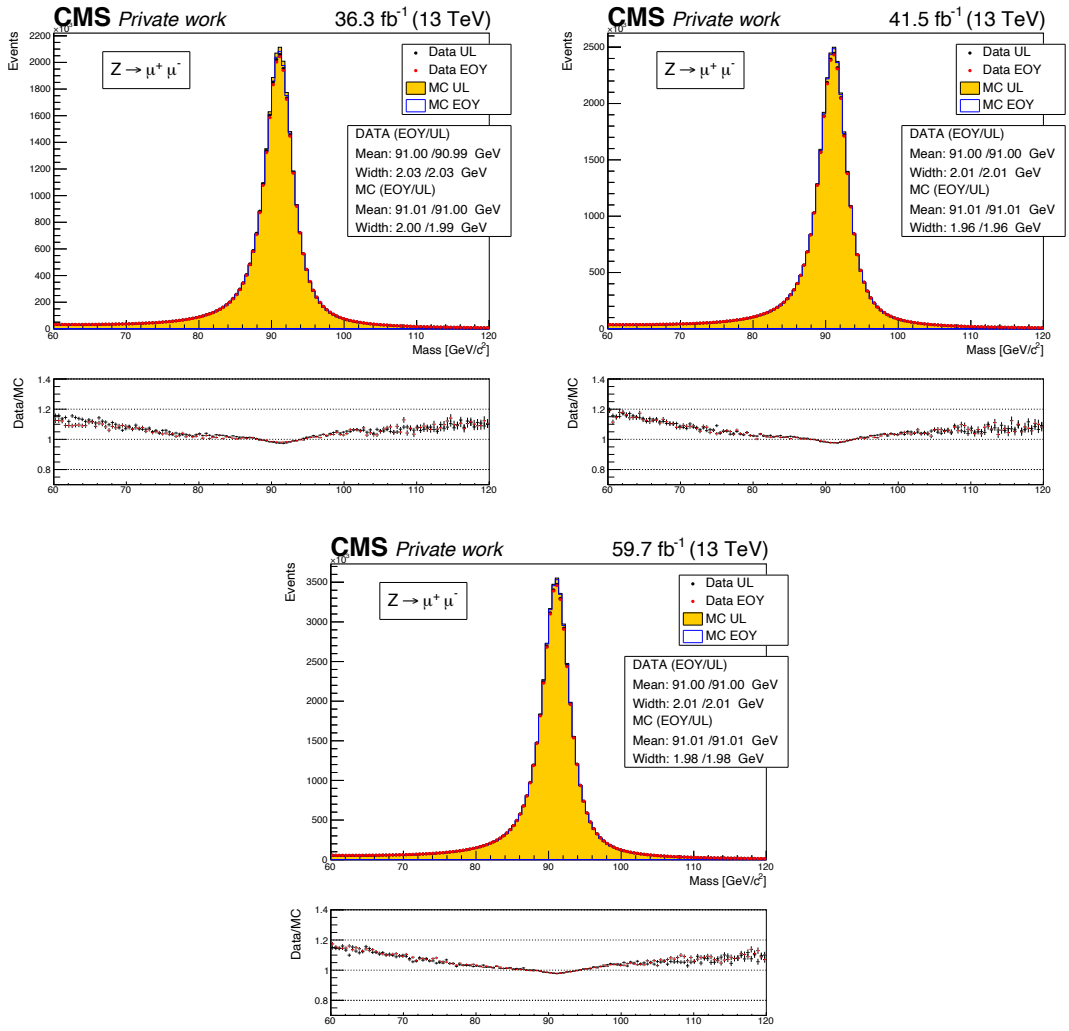

Figure 3.4: Comparison in data and MC of the invariant Z boson mass in the $\mu^+\mu^-$ final state between EOY and UL datasets. The MC simulations include Drell-Yan and $t\bar{t}$. The distributions are fitted with a double-sided Crystal Ball to evaluate the peak position (muon scale) and the distribution width (muon resolution). The bottom panels display the data/MC ratio. The distributions are shown for the three data-taking periods: 2016 (top left), 2017(top right), and 2018 (bottom).

line shape of the resonance for passing and failing probes and removing the background contribution. A simple cut-and-count approach is not feasible in many cases. It is well-established that the selection efficiency measured with the TnP method outlined above is almost insensitive to the selection criteria applied to the tag.

The assessment of the lepton efficiency is usually factorised into distinct components, each targetting a different selection in the reconstruction and identification chain. Hence, the general form of the efficiency can be written as:

$$\varepsilon_\ell = \varepsilon_{\text{sel1}} \cdot \varepsilon_{\text{sel2}} \cdot \varepsilon_{\text{sel3}} \cdot \ldots \tag{3.9}$$

Each term represents the efficiency of the probes to pass a given selection, contingent upon meeting the selections of the preceding steps. Consequently, the order of the product is of significance. For instance, ϵ_{sel2} is the efficiency of passing sel2, given that the probes have already met the requirements of sel1.

Electrons

The electron efficiency is broken down into two components:

$$\varepsilon_e = \varepsilon_{\text{reco}} \cdot \varepsilon_{\text{ID}} \tag{3.10}$$

where $\varepsilon_{\text{reco}}$ is the efficiency related to the electron reconstruction explained in Sec. 2.3, while ε_{ID} is related to the identification explained in Sec. 3.2.2. The main component of the former efficiency is the track reconstruction, since it was shown during Run 1 that the cluster reconstruction efficiency is very close to 100%. Thus, the reconstruction efficiency comes down to the only GSF tracking efficiency. This term is computed centrally and will not be detailed here, an extensive description can be found in Ref. [91].

On the other hand, the identification efficiency, being the ID itself analysis-dependent, is specific to the $\text{H} \to \text{ZZ} \to 4\ell$ channel. Its computation relies on the TnP technique applied to the Z resonance, where the tag is required to pass the tight cut-based ID [91], to have a geometrical matching to the leg of a single electron trigger object, transverse momentum $p_{\text{T}}^{e} > 30$ GeV, and a supercluster rapidity $|\eta_{\text{SC}}^{e}| < 2.17$ (ECAL cracks are masked[2]). The probes are instead selected among GSF candidates with an opposite charge to that of the tag. In the case of simulation, an additional condition is added to the tag requirement of being geometrically linked to a generated electron from the Z boson.

The main challenge of the TnP method for electrons is the large QCD background in the *failing probe* distribution, which makes it difficult to distinguish the signal contribution from the background. This feature is even more evident in the low-p_{T} bins ($p_{\text{T}}^{e} < 20$ GeV), where the Z resonance naturally does not produce many electrons in this phase space region. In order to reduce the background and make the fit more stable, the tag requirements are tightened. The transverse momentum cut is increased to 50 GeV (only if $p_{\text{T}}^{e} < 15$ GeV), the electron is required to have an MVA ID [91] greater than 0.92, and satisfy the following kinematic cut to remove contributions from W decays:

$$\sqrt{2 \cdot \text{MET}_{\text{PF}} \cdot p_{\text{T}}^{\text{tag}} \cdot (1 - \cos(\phi_{\text{MET}} - \phi_{\text{tag}}))} < 45 \text{ GeV} \tag{3.11}$$

The numbers of passing and failing probes are extracted from a template fit to the invariant mass of the two electrons. The resonant signal shape is taken from MC and convoluted with a Gaussian function. The data distribution is then fitted with the sum of the template mentioned before and an error function with a one-sided exponential tail. For $p_{\text{T}}^{e} > 30$ GeV, the selection efficiency results to be very high, ranging from 85% to 95%. While for low-p_{T} electrons, it can decrease to 70%.

After the evaluation of the efficiency, the related uncertainty should be computed. The uncertainty of the selection efficiency is crucial for the $\text{H} \to \text{ZZ} \to 4\ell$ channel as the per-lepton efficiency propagates with the power of four. As will be discussed in Ch. 4, the electron efficiency is the leading systematic uncertainty in this analysis. The canonical approach to evaluate the uncertainty with the TnP method is to rerun the fit by considering variations to the nominal setting. These variations include the change of the signal and background models, commonly with a double-sided Crystal Ball and an exponential, respectively; the tag selections are changed to consider the bias introduced by the tag requirements; only for simulations, a change in the perturbative order of the DY simulation to take into account the uncertainty on the overall description of the event.

The approach used until the analysis presented in this thesis involved summing in quadrature the uncertainties related to these variations. However, this method was found

[2]The ECAL cracks are the transition regions between EE and EB, i.e., in the pseudorapidity region $1.4442 < |\eta_{\text{SC}}| < 1.566$, where due to the position of the crystals and modules, the electron reconstruction is degraded

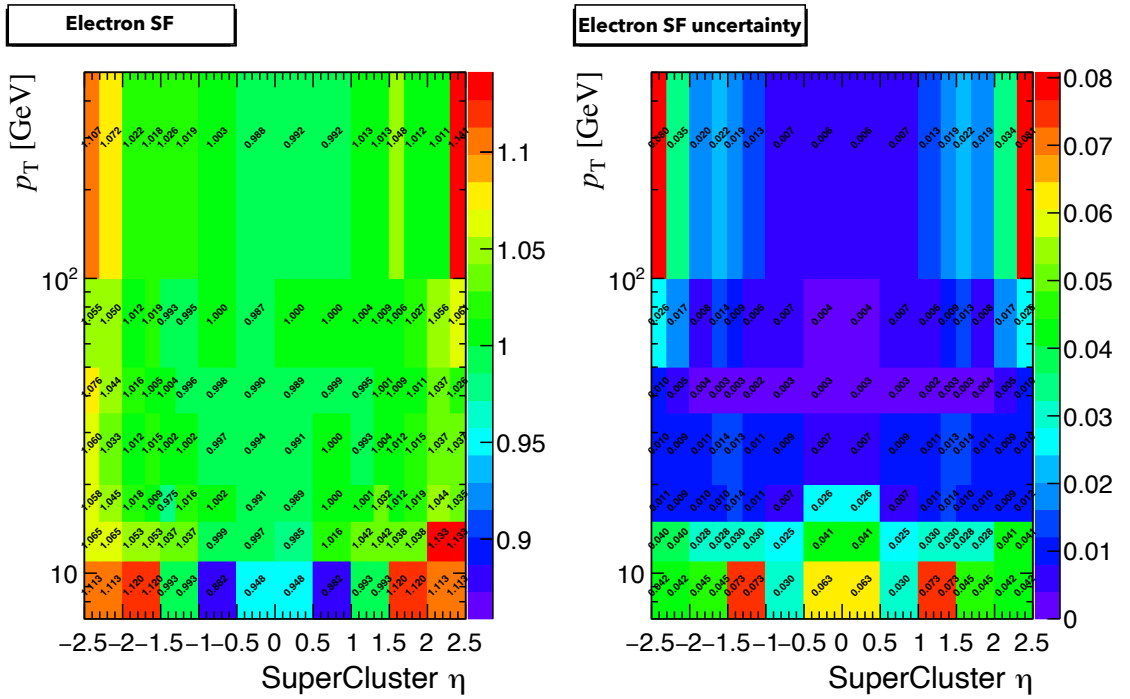

Figure 3.5: Electron scale factors (left) and related uncertainties (right) in bins of the transverse momentum p_T and supercluster pseudorapidity η of the electron. These values are related to the 2017 data reconstructed with the UL calibration.

to overestimate the uncertainty. The UL reprocessing provided an opportunity to revise the statistical method and implement a root-mean-square (RMS) approach [164]. The main conceptual difference coming from the RMS approach is that the alternative measurements are treated as different fitting models addressing the same source of systematic. Therefore, doing different fits is just doing different measurement on the same sample and, if we consider each of them equally valid, the RMS gives a good estimate of the spread. Following the RMS idea, the mean value of all variations is taken to be the central SF value, as we are not considering any measurements to be better motivated. The nominal setting is also now included among the variations. The uncertainty is given as an error of the mean RMS/$\sqrt{N}$, where N is the number of variations. The improvement is more pronounced in the lowest p_T bins ($p_T^e > 20$ GeV), where the previous uncertainty was larger. In these kinematic regions, the RMS method can reduce the uncertainty by a factor ranging from 30% to 40%. The impact on the final systematic uncertainty will be discussed in Sec. 4.8. The electron SF and their uncertainties are reported in Fig. 3.5 for 2017 data.

Despite the introduction of the new root-mean-square method, the uncertainties in the low-p_T bins are still quite large, exhibiting an increase of up to 60% with respect to the corresponding η bin at higher p_T. A potential enhancement could be achieved by leveraging the J/Ψ resonance in the low-p_T bins, as will be discussed in the context of muons later. For electrons, the use of this resonance has been hampered by the absence of suitable triggers in the CMS L1 and HLT menus. Studies are ongoing to assess the feasibility of implementing the J/Ψ resonance for the forthcoming runs of the LHC.

Muons

The muon efficiency is factorised into three components:

$$\varepsilon_\mu = \varepsilon_{\text{reco/ID}} \cdot \varepsilon_{\text{PV}} \cdot \varepsilon_{\text{ISO}} \tag{3.12}$$

where $\varepsilon_{\text{reco/ID}}$ is the reconstruction and identification efficiency, ε_{PV} is the efficiency on the impact parameter requirements, and ε_{ISO} refers to the isolation requirement. The efficiency to reconstruct a muon in the inner tracker is missing from the equation above. It was measured and found to be 99% or higher within the whole tracker acceptance and in good agreement with the expectation from simulations. Its computation relies on considering probe muons reconstructed just in the muon chambers, i.e., standalone muon, and finding a matching in the inner tracker. More details on this computation can be found in Ref. [165].

The computation of the first reconstruction and identification efficiency $\varepsilon_{\text{reco/ID}}$ relies on the Z resonance, if $p_{\text{T}}^{\mu} > 20$ GeV, and the J/Ψ resonance, if $p_{\text{T}}^{\mu} < 20$ GeV. The J/Ψ resonance is useful at low-p_{T} to increase the number of signal muons and improve the stability of the fit. Events are selected if passing a single muon trigger, and the tag is required to have $p_{\text{T}}^{\mu} > 8$ GeV, $|\eta^{\mu}| \leq 2.4$, pass the tight ID, and be geometrically matched with the HLT object. The probes are tracks reconstructed in the inner tracker, and the passing probes are those that also pass the loose muon ID. The resulting efficiency is close to one; almost every bin is above 98%. Even though it always remains high, the only significant decrease in efficiency is in the barrel below $p_{\text{T}}^{\mu} < 8$ GeV, where it ranges between 92% and 97%.

The efficiency computation of the impact parameter requirements ε_{PV} relies only on the Z resonance. The tag is required to have $p_{\text{T}}^{\mu} > 28$ GeV, SIP < 4, and be isolated $I_{\mu}(0.4) < 0.35$. For this measurement, the probe is a muon passing the loose ID, and it is considered a passing probe if it satisfies the SIP, d_{xy}, and d_z cuts presented in Sec. 3.2.1. The efficiency is measured to range between 94% and 100%, with the lowest values in the low-p_{T} bins (the efficiency plateau is reached around 50 Gev) and in the high-η bins.

The last component is the efficiency of the isolation requirement ε_{ISO}. Similarly to the impact parameter case, the isolation efficiency is measured using events from the Z resonance for any p_{T} and employs the same tag definition. The isolation efficiency ranges between 88% and 100%, where the lowest values are attained for low-p_{T} muons (the efficiency plateau is reached around 60 GeV) and in the high-η bins. The decrease in efficiency at larger pseudorapidity is steeper than ε_{PV}.

The systematic uncertainties are evaluated with the same approach for all components. The analytical signal and background shape models are varied with other functions; the number of bins in the dimuon mass range and the dimuon mass range are varied. In addition, variations on the tag requirements are also considered, especially on the tag isolation and the SIP cut.

Figure 3.6 shows the overall muon scale factors and uncertainties for 2017, given by combining the three components outlined above. The scale factors do not show any significant p_{T} dependence; they are only slightly higher for larger pseudorapidity, and the related uncertainties are almost constant in all the phase space with a value less than 1%. These considerations can be extended to all three years of data-taking.

3.3 FSR photons

Photons do not belong to the signature of the final state investigated in the context of the $\text{H} \to \text{ZZ} \to 4\ell$ channel. However, after the decay of the Z boson, leptons can potentially emit a photon, a phenomenon known as *Final State Radiation* (FSR). Since FSR

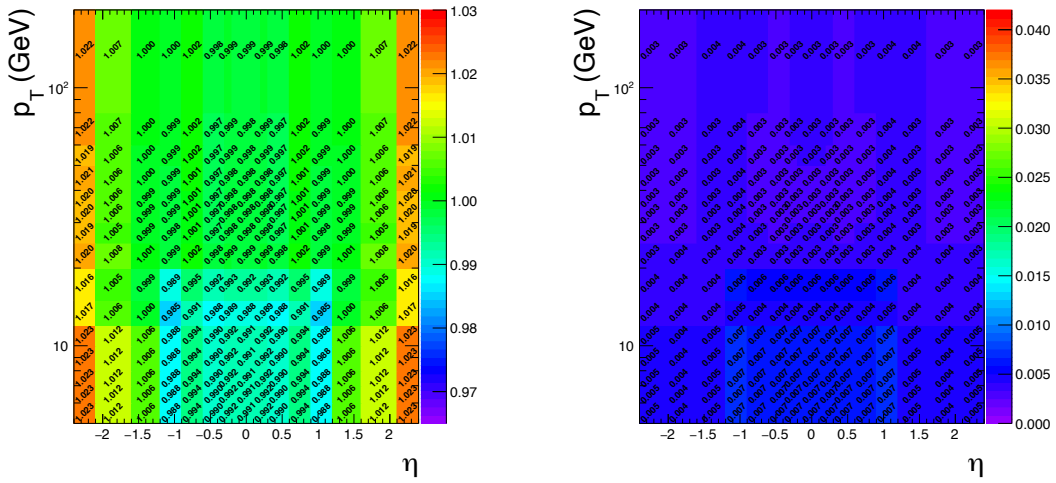

Figure 3.6: Muon scale factors (left) and related uncertainties (right) in bins of the transverse momentum p_{T} and pseudorapidity η of the muon. These values are related to the 2017 data reconstructed with the UL calibration.

radiations carry away a fraction of the lepton energy, it is vital to recover such photons to properly reconstruct the kinematic of the four-lepton final state and not to degrade the resolution, as it can be seen in Fig. 3.7. The challenge when establishing an FSR recovery procedure is to ensure the collection of genuine FSR photons. This is achieved by exploiting the characteristics of the FSR radiation, which tends to be soft and nearly collinear to the lepton.

Starting from the collection of PF photons (cfr Sec. 2.3), the identification of FSR and their associated leptons starts with basic kinematic requirements:

$$p_{\mathrm{T}}^{\gamma} > 2 \text{ GeV}, \quad |\eta^{\gamma}| < 2.5$$

FSR candidates are then isolated using Eq. 3.4 and required to have $\mathcal{I}_{\gamma}(0.3) < 1.8$. In the computation, a threshold of 0.2 GeV and a veto cone of radius 0.0001 is applied on charged hadrons, as well as a threshold of 0.5 GeV and a veto cone of radius 0.01 on neutral hadrons and photons.

Next, the *supercluster veto* is applied. Any PF photons previously linked to their respective electron candidates at the PF level during the superclustering process are excluded. This step is executed for all loose leptons.

The surviving photons are linked to the closest loose lepton in the event and, to suppress photons from initial-state radiation, PU interactions, and π^0 decays, the physics properties of the FSR radiation are exploited by imposing:

$$\Delta R(\ell, \gamma) < 0.5, \quad \Delta R(\ell, \gamma)/(p_{\mathrm{T}}^{\gamma})^2 < 0.012 \text{ GeV}^{-2}$$

If more than a photon can be associated to the same lepton, only the one with the lowest $\Delta R(\ell, \gamma)/(p_{\mathrm{T}}^{\gamma})^2$ is retained. The emission of FSR photons affects a few percent of events. In the Run 2 dataset, the FSR recovery is applied to 6.5% of events satisfying the selections detailed above (5.3% in $4e$, 7.2% in 4μ, and 6.5% in $2e2\mu$).

As discussed in Sec. 3.2.2, FSR photons are not included in the isolation sum since they usually lie within the isolation cone, causing the isolation condition to fail. All the FSR photons in the event, no matter they are linked to the lepton under study or any other loose lepton in the event, are removed from isolation sum.

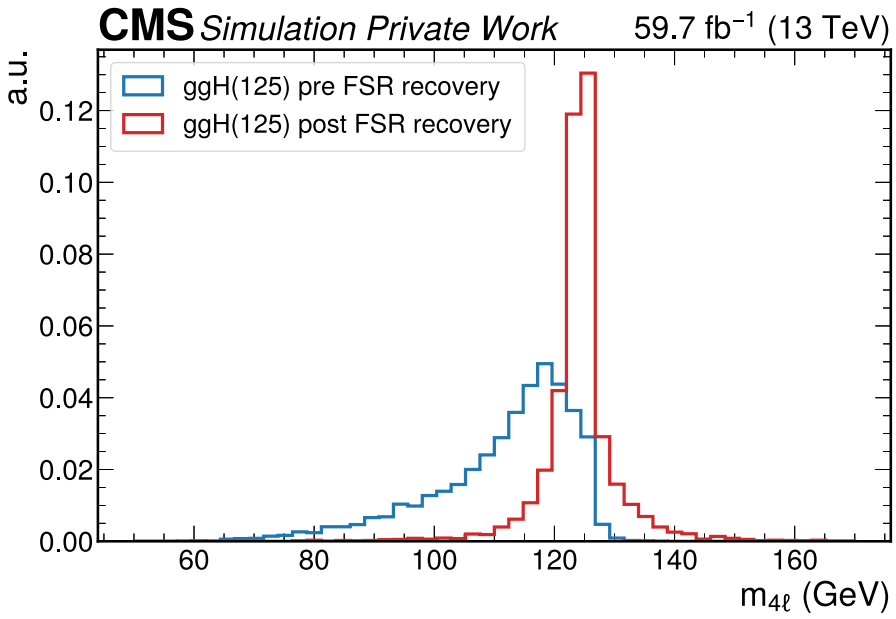

Figure 3.7: Invariant mass of the four-lepton system for events affected by FSR radiation, both before (blue) and after (red) its recovery. The distributions are computed using the 2018 MC simulation of the ggH production mode.

3.4 Jets

Jets are not among the distinctive features of the $H \to ZZ \to 4\ell$ channel. Nevertheless, they can be a precious handle to target specific production modes, such as VBFH, and measuring cross sections in bins of jet observables, such as the transverse momentum of jets or their multiplicity, can provide insights in the QCD computations. The four-lepton analyses employ PF jets (cfr Sec. 2.3.3) reconstructed with the anti-kt algorithm with a radius of $\Delta R = 0.4$. The JEC and JES corrections are applied as explained in Sec. 2.3.3. Starting from these jets, some identification conditions are applied to improve the purity of the jet collection.

Central recommendations from the CMS collaboration suggest employing a cut-based jet ID relying on some variables such as the fraction of neutral hadrons, charged hadrons, and electromagnetic objects, the number of the jet constituents, and the charged multiplicity. These cuts are tuned to reject fake, badly reconstructed and noisy jets, and attain 98-99% of efficiency in all pseudorapidity regions and 99% of background rejection for $|\eta^{\text{jet}}| < 3$. The list of the cuts is reported in Tab. 3.9.

In addition to the jet ID outlined above, another ID has been developed to cope with the precise issue of jets from PU interactions. It is applied only for jets with $p_{\text{T}}^{\text{jet}} < 50$ GeV,

Table 3.9: Cuts on the variables used to define the jet ID employed in the $H \to ZZ \to 4\ell$ analyses.

	$\eta \leq 2.4$	$2.4 < \|\eta\| \leq 2.7$	$2.7 < \|\eta\| \leq 3.0$	$3.0 < \|\eta\| \leq 5.0$
Neutral hadron fraction	< 0.90	< 0.90	< 0.90	> 0.2
Neutral electromagnetic fraction	< 0.90	< 0.99	$\in (0, 0.99)$	< 0.9
Number of constituents	> 1	–	–	–
Charged hadron fraction	> 0	–	–	–
Charged multiplicity	> 0	–	–	–
Number of neutral particles	–	–	> 1	> 10

as it is the region where the highest contribution of such jets is expected, and relies on three distinctive properties:

- inside the tracker acceptance, the trajectories of tracks associated to the jets can be used to establish the compatibility of the jet with the PV;
- the topology of the jet shape can be used to disentangle jets arising from the overlap of multiple interactions from truly hard jets;
- the object multiplicity can be used as an additional handle.

This information is integrated in a BDT that is able to reject 89% of PU jets while maintaining 96% of hard jets [166].

A jet passing all the identification criteria outlined above is considered in the analysis if $p_{\mathrm{T}}^{\mathrm{jet}} > 30$ GeV, $|\eta^{\mathrm{jet}}| < 4.7$, and if it is cleaned from the presence of any tight lepton and FSR photon, i.e., $\Delta R(\mathrm{jet}, \ell/\gamma) > 0.4$. The extension of the pseudorapidity range is another beneficial effect of the UL reprocessing. Due to an increase of the ECAL noise related to the ageing of crystals, previous jet reconstructions showed a larger multiplicity, particularly enhanced for 2017. The main effect was the formation of *horns* in the regions between $2.5 < |\eta_{\mathrm{jet}}| < 3.0$ and a worse data/MC agreement in the endcaps. As a result, previous $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ fiducial analyses employed a pseduorapidity veto for jets below 2.5. This issue has been successfully addressed and tamed by the UL calibration, allowing the extension of the $|\eta^{\mathrm{jet}}|$ up to 4.7. The control plot comparing the $|\eta_{\mathrm{jet}}|$ distributions of the leading jet for EOY and UL datasets is reported in Fig. 3.8.

Summary

This chapter introduced the necessary physics objects for reconstructing and describing the decay of the Higgs boson into four leptons. The UL recalibration does not bring significant changes for muons. Conversely, the enhanced ECAL reconstruction visibly influences electrons, thereby directly impacting associated systematic uncertainties. Furthermore, the improvements in the jet reconstruction and correction allows the extension of the jet phase up to $|\eta| = 4.7$, including the more forward region. A summary is reported in Fig. 3.9.

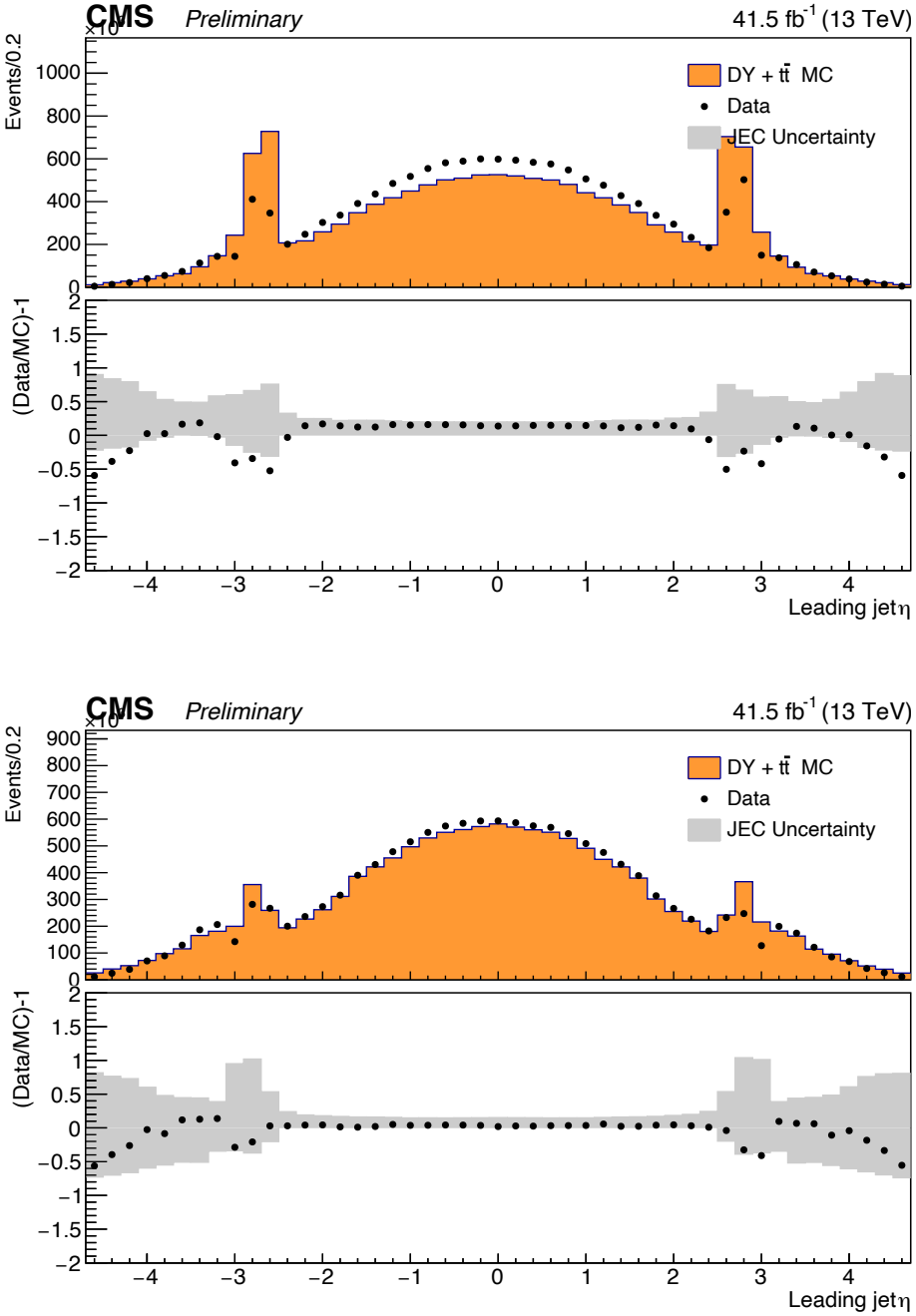

Figure 3.8: Comparison between data and MC for the leading jet η for 2017, for EOY (top) and UL (bottom). Z$\rightarrow \ell\ell$ +jets events are used. Jet ID and Jet PU ID are applied. MC samples include DY and $t\bar{t}$. Data/MC ratio plot is shown in the lower pad together with the uncertainties (shaded histograms) from Jet Energy Corrections (JEC).

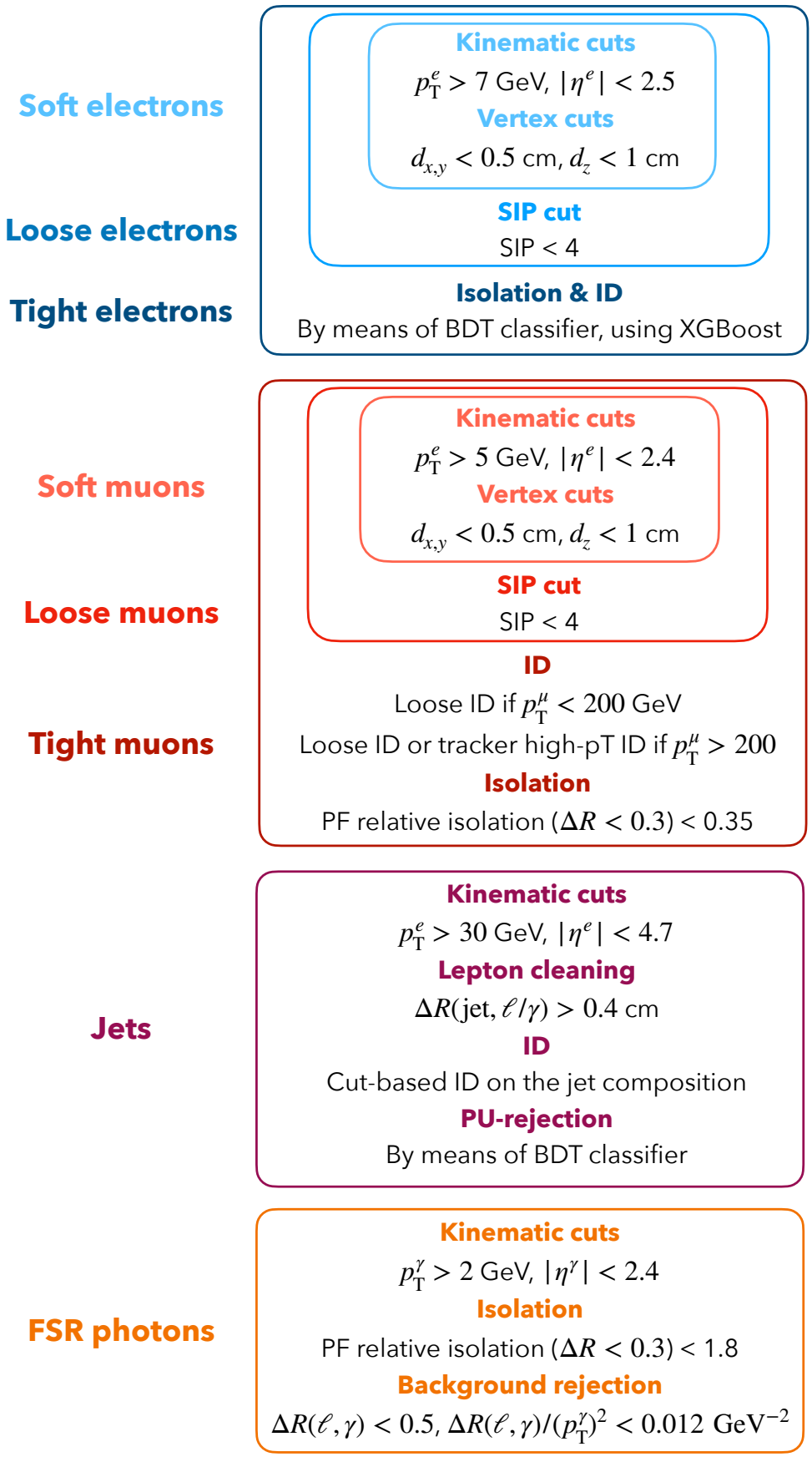

Figure 3.9: Summary of the selection requirements applied to electrons, muons, FSR photons, and jets that enter the $H \to ZZ \to 4\ell$ analysis.

Chapter 4

Development of the strategy to measure fiducial cross sections of Higgs boson production in the four-lepton final state

This chapter is devoted to developing a strategy for measuring the Higgs boson fiducial cross sections in the four-lepton channel. Such measurements offer a model independent approach to characterise the Higgs boson and yield results with enduring validity. These results can be directly compared to theoretical models and other decay channels.

Building upon the physics objects defined in the previous chapter, the first step of the analysis strategy is to select $H \to ZZ \to 4\ell$ events from all the processes occurring in proton-proton collisions. This procedure starts with reconstructing the Z boson candidates, which are subsequently combined to form ZZ candidates. Specific criteria will be established among these candidates to select the most likely Higgs boson candidate. This event selection procedure will be elaborated in Sec. 4.1. However sophisticated the event selection procedure may be, there will invariably be a fraction of background events in addition to the Higgs boson signal. The signal and background modelling will be explained in Sec. 4.2 and 4.3, respectively. After these considerations, which are similar to all $H \to ZZ \to 4\ell$ analyses, Sec. 4.4 and 4.5 will delve into the core of the fiducial analysis. This involves clearly defining the unfolding algorithm and the fiducial phase space. This thesis will not only present the fiducial cross section measured inclusively, but also in 32 differential observables, thereby providing a more detailed description of the decay channel. These variables will be introduced in Sec. 4.6. The observed events must be adequately parametrised to measure fiducial cross sections, as detailed in Sec. 4.7. The ultimate missing element required to determine the final results accurately is the correct evaluation and implementation of all systematic uncertainties, which will be addressed in Sec. 4.8.

4.1 Event selection

An event selection algorithm consists of a series of criteria that an event must satisfy to be considered in the Signal Region (SR) of the analysis.

The preliminary step, as already outlined in Sec. 3.1.2, is the trigger selection. This selection step aims to select events that likely correspond to the $H \to ZZ \to 4\ell$ decay channel from the extensive pool of events recorded by the CMS detector. The trigger efficiency is evaluated to have a plateau at 99% and in agreement between data and simulations. The

adopted strategy is not to apply any SF but to include a systematic uncertainty, ranging 1-3% depending on the final state and the kinematics of the leptons, to cover the small discrepancy. This uncertainty is included in the lepton efficiency uncertainty, as outlined in Sec. 4.8.

The event is then required to have at least one PV with a radius smaller than 2 cm, a number of degrees of freedom[1] greater than 4, and collisions limited to the z-axis.

After these general requirements, the physics objects defined in the previous chapter come into play for the reconstruction of the Higgs boson candidate. The Z boson candidates are formed by combining opposite-sign tight leptons, as defined in Sec. 3.2, with invariant mass in the range $12 < m_{\ell^+\ell^-(\gamma)} < 120$ GeV, where the invariant mass includes the FSR photons.

The Z boson candidates are then combined to form ZZ candidates. In the doublet, the candidate with the mass closest to the PDG value is denoted as Z_1, while the other is labelled as Z_2. If two Z candidates share at least one lepton, the ZZ candidate is discarded. To check the overlap, it is required that $\Delta R(\ell\ell) > 0.02$ for all pairs of leptons from the two Z candidates.

The final Higgs boson candidate is selected among all reconstructed ZZ candidates following the requirements below.

- The invariant mass of the Z_1 must be above 40 GeV. This cut filters out background contributions from events without a Z boson, such as $\mathrm{t\bar{t}}$ and $\mathrm{W}\gamma$ + jets, while still allowing for the possibility of the decay to two off-shell Z bosons $\mathrm{H} \to \mathrm{Z}^*\mathrm{Z}^* \to 4\ell$.

- The leading and sub-leading leptons are required to have $p_\mathrm{T} > 20$ GeV and $p_\mathrm{T} > 10$ GeV, respectively, to ensure a high trigger efficiency.

- The invariant mass of all opposite-sign lepton pairs that can be defined with the four selected leptons must have $m_{\ell^+\ell'^-} > 4$ GeV, without including the FSR recovery. This cut is designed to reduce contributions from decays of bound quark states.

- For the $4e$ and 4μ final states, the Z_a and Z_b candidates are defined as the mass-sorted alternative pairing of leptons forming Z boson candidates, i.e., Z_a and Z_b are formed with one lepton from the Z_1 and the other from the Z_2. The event is discarded if $|m_{\mathrm{Z}_a} - m_{\mathrm{Z}_0}| < |m_{\mathrm{Z}_1} - m_{\mathrm{Z}_0}|$ and $m_{\mathrm{Z}_2} < 12$ GeV, where m_{Z_0} is the world-average Z boson mass [26]. The FSR recovery is included in the invariant mass computations. This cut removes events where the alternative pairing looks like an on-shell Z boson together with a low-mass $\ell^+\ell^-$ resonance.

- The invariant mass of the four leptons must be larger than 70 GeV.

This selection chain usually ensures the presence of a single ZZ candidate that is promoted to a Higgs boson candidate. However, in case of associated productions (VH and ttH), more than one ZZ candidate can be identified. In such cases, the one with the highest transverse momentum is retained.

The outcome of the event selection procedure is reported as a function of the reconstructed four-lepton invariant mass in Fig. 4.1.

4.2 Signal modelling

The distinctive feature of the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ channel is the clear peak of the Higgs boson resonance standing out over the almost flat background. This characteristic allows treating

[1] This parameter is related to the track fitting stage, and it is linked to the number of tracks forming the vertex.

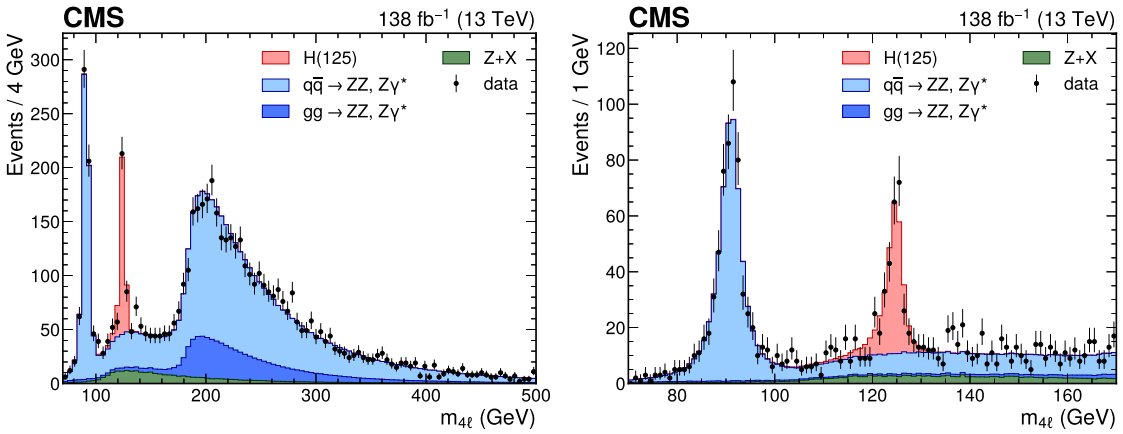

Figure 4.1: Distribution of the reconstructed four-lepton invariant mass $m_{4\ell}$ from 70 GeV to 500 GeV (left) and to 170 GeV (right), with Run 2 data. The SM Higgs boson signal with $m_{\mathrm{H}} = 125$ GeV, denoted as H(125), and the ZZ backgrounds are normalised to the SM expectation (cfr Sec. 4.3.1), the Z+X background to the estimation from data (cfr Sec. 4.3.2).

this analysis with a parametric shape-based approach, where an analytical function is used to model the signal resonant shape. This approach ensures a good description of the signal across the entire mass range considered in the analysis, without having to deal with issues related to the limited statistics available for defining histogram templates for the signal. In order to attain an optimal signal parametrisation, MC simulations at various m_{H} points are used to include in the analytical model an explicit dependence on m_{H}, thereby reducing the model-dependence of the analysis. Consequently, m_{H} can be either fixed to a certain value or profiled during the fitting process.

The SM predicts the width of the Higgs boson to be $\Gamma_{\mathrm{H}} = 4$ Mev at a mass of $m_{\mathrm{H}} = 125$ GeV. While this value would still permit the use of the narrow-width approximation and the description of the Higgs resonance with the typical Breit-Wigner function, real-world experimental factors come into play. Experimental effects such as the finite resolution of the detector, energy losses in the passive materials, bremsstrahlung effects, and FSR emissions, contribute to the observation of a reconstructed Higgs bosons resonance with a width well above the $\Gamma_{\mathrm{H}} = 4$ Mev predicted by the SM. This condition is one of the main limitations to the direct measurement of Γ_{H} using on-shell data, but this goes beyond the scope of this thesis. The relevant point here is that the overall resonant shape of the Higgs boson peak at the LHC experiments has to be described by a function different from the Breit-Wigner, to take into account the increased width of the resonance as well as the asymmetric tails that arise from the experimental effects. A double-sided Crystal Ball function $\mathcal{P}_{res}(m_{4\ell}|m_{\mathrm{H}})$ is found to well describe the probability density function (pdf) of the Higgs resonance:

$$\mathcal{P}_{res}(m_{4\ell}|m_{\mathrm{H}}) = N \cdot \begin{cases} A_L \cdot (B_L + |\xi|)^{-n_L} & \text{if } \xi < \alpha_L \\ \exp(-\xi^2/2) & \text{if } \alpha_L \leq \xi \leq \alpha_R \\ A_R \cdot (B_R + |\xi|)^{-n_R} & \text{if } \xi > \alpha_R \end{cases} \tag{4.1}$$

where:

$$A_i = \left(\frac{n_i}{|\alpha_i|}\right)^{n_i} \cdot \exp\left(\frac{-|\alpha_i|^2}{2}\right) \tag{4.2}$$

$$B_i = \frac{n_i}{|\alpha_i|} - |\alpha_i| \tag{4.3}$$

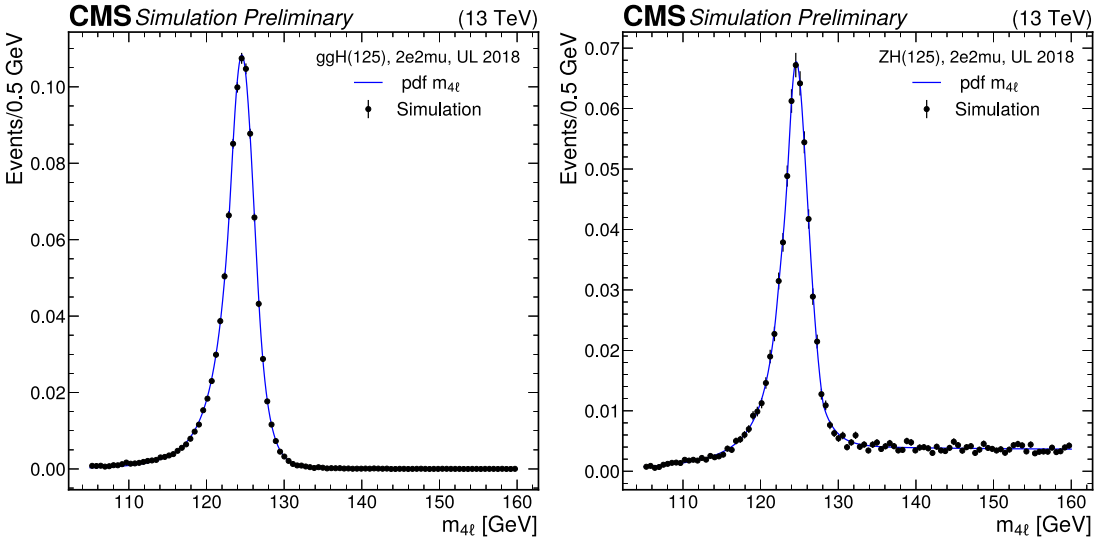

Figure 4.2: Signal modelling for the ggH (left) and ZH (right) production mechanisms in the $2e2\mu$ final state for the 2018 data-taking period. The black points represent the $m_{4\ell}$ invariant mass reconstructed in simulation, while the error bars correspond to the statistical uncertainty. The blue solid lines depict the analytical parametrisation obtained from the simultaneous fit over five mass points, as described in the text. For the associated production ZH, the final pdf is described by the sum of a double-sided Crystal Ball function and a Landau distribution to take into account possible wrong-lepton assignments in the reconstruction.

In the expression of the double-sided Crystal Ball, $\xi = (m_{4\ell} - m_{\mathrm{H}} - \Delta m_{\mathrm{H}})/\sigma_m$ is the independent variable of the pdf. There are six independent parameters that can be more generally denoted as p_i: the variance of the Gaussian core σ_m that reflects the width of the four-lepton resonance; the systematic mass shift of the peak Δm_{H}; α_R and α_L that define the boundary between the Gaussian core and the two non-Gaussian tails; n_L and n_R that control the prominence of the tail. The coefficient N is the normalisation factor of the pdf. Each parameter p_i of the double-sided Crystal Ball function is given by a linear dependence on m_{H} for a total of 12 free parameters:

$$p_i(m_{\mathrm{H}}) = a_i^0 + a_i^1 \cdot (m_{\mathrm{H}} - 125\ \mathrm{GeV}) \tag{4.4}$$

An initial guess of the best fit values $p_i^0(m_{\mathrm{H}})$ is obtained, for all the parameters, from a fit of the sample at $m_{\mathrm{H}} = 125$ GeV. These values are then used as inputs for the simultaneous fit over the 5 mass points (120, 124, 125, 126, and 130 GeV) in the mass range $105 \leq m_{4\ell} \leq 160$ GeV, which allows finding the best-fit values of the $p_i(m_{\mathrm{H}})$ coefficients. The normalisation coefficient N is the only parameter that is not subject to this procedure, as it is proportional to the fiducial cross section, which is the parameter of interest of the analysis and therefore left floating in the final fits. More details about N and the statistical inference procedure will be given in Sec. 4.7.

The parameters of Eq. (4.1) are derived from a fit of the ggH production mechanism, as it is the dominant one and the results are found to describe well also the other production modes. Closure tests have been performed by deriving the parameterisation for each production mode independently and showed an excellent agreement with the parametrisation obtained only with ggH. For the associated productions, namely VH and ttH, the signal line shape can be distorted by the presence of non-resonant events. This contribution arises from one of the four reconstructed leptons originating from the additional vector boson (or top quark) being misidentified as originating from the Higgs boson decay. In

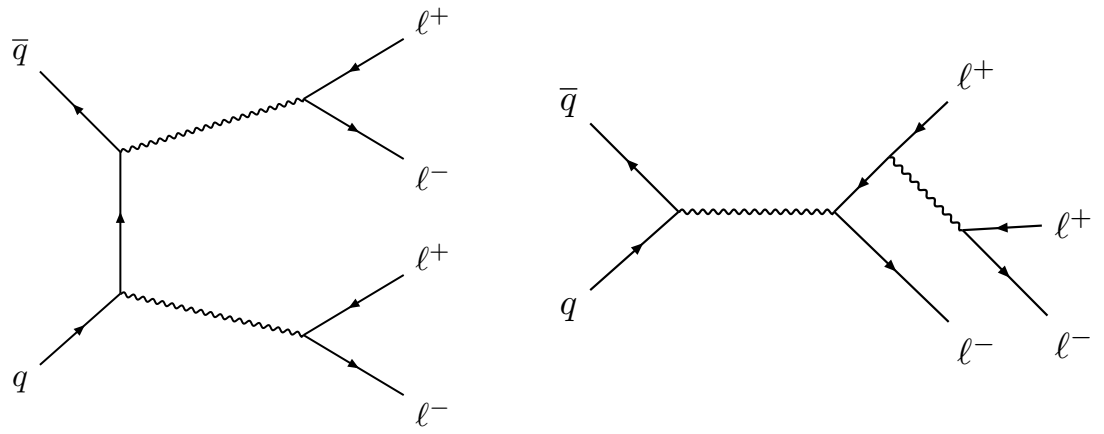

Figure 4.3: The LO Feynman diagrams for the quark-initiated background production of 4ℓ: the t-channel production (left); the s-channel production of $q\bar{q} \to Z^{(*)}/\gamma^* \to \ell^+\ell^-$ with associated radiative decay to an additional lepton pair.

order to account for this effect, a Landau distribution is added to the double-sided Crystal Ball function of Eq. 4.1 when fitting the associated production modes.

The projections of the simultaneous fit with $\mathcal{P}_{res}(m_{4\ell}|m_{\mathrm{H}} = 125\ \mathrm{GeV})$ is shown in Fig. 4.2 for the $2e2\mu$ final state and the 2018 year of data-taking.

4.3 Background modelling

The backgrounds in the four-lepton channel can be classified as either *reducible*, when particles are wrongly classified as signal leptons, or *irreducible*, when the leptons are associated to the non-resonant production of four-leptons. The four leptons in the irreducible component are produced by the non-resonant production of two Z bosons either via quark-antiquark annihilation ($q\bar{q} \to ZZ$) or gluon fusion ($gg \to ZZ$). The reducible background is made of events where at least one of the leptons is not a genuine, prompt lepton and this contribution is mainly related to the jet activity. While MC simulations can be used to model the irreducible component, they do not model properly the reducible processes, hence the need to rely on data-driven methods.

4.3.1 Irreducible background

Quark-antiquark annihilation

The non-resonant production of two Z bosons via quark-antiquark annihilation is mediated by both the t-channel and the s-channel (Fig. 4.3), and its $m_{4\ell}$ spectrum presents various sub-structures as can be seen in the left plot of Fig. 4.4. While the s-channel is responsible for the resonant peak at the Z boson mass, the t-channel contributes with a step-like behaviour. The evolution of the t-channel contribution unfolds gradually. Initially, neither of the two propagators can be on-shell. With increasing energy, the first step occurs at around $m_{4\ell} \sim 100$ GeV, when one of the two propagators can be an on-shell Z boson. Subsequently, around $m_{4\ell} \sim 200$ GeV, a second step takes place, marking the point where two on-shell Z bosons can be produced leading to a sharp rise of the cross section.

The process is simulated at NLO in pQCD using `POWHEG 2.0`, as detailed in Sec. 3.1.1. Although the partonic events are generated at NLO, the fully differential cross section has been computed at NNLO in pQCD [167]. Therefore, a *k-factor* is introduced to profit of the higher-order theoretical computation. This corrective factor is calculated as a function of m_{ZZ} by taking the ratio of the NNLO and NLO cross sections. The k-factor is then

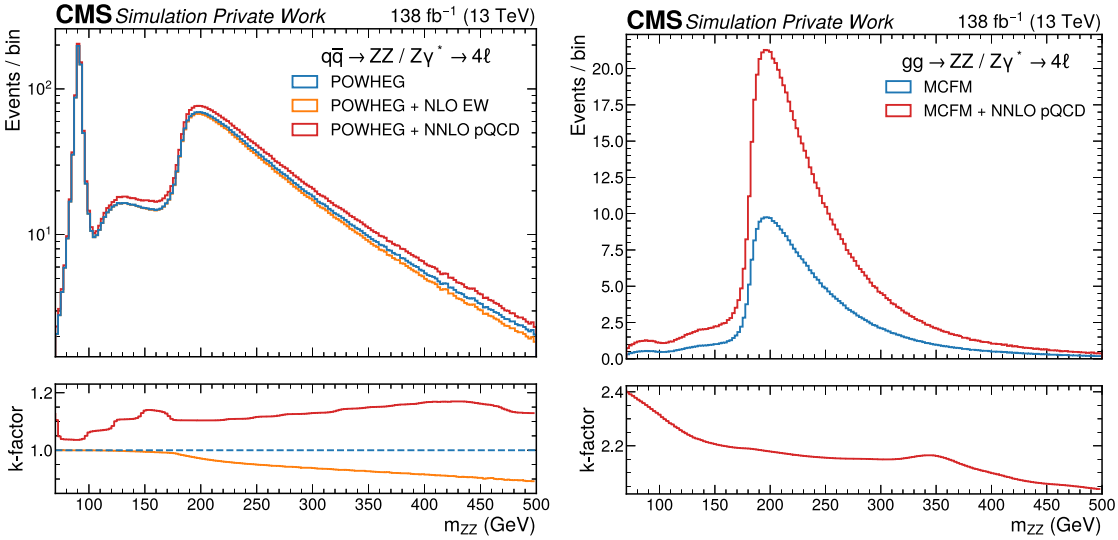

Figure 4.4: (Left) Distribution of m_{ZZ} for the $q\bar{q} \to 4\ell$ as generated by `POWHEG 2.0` at NLO pQCD (blue line), after the application of the NLO EW k-factor (orange line), and after the application of the NNLO pQCD k-factor (red line). (Right) Distribution of m_{ZZ} for the $gg \to 4\ell$ process as generated by `MCFM 7.0.1` at LO pQCD (blue line) and at NNLO pQCD (red line) after the application of the corresponding k-factors.

incorporated as an additional weight to the MC-event weight introduced in Sec. 3.1.1, and it has an effect of increasing the cross section by a factor ranging from 1.05 to 1.2.

An additional k-factor is introduced to account for NLO ElectroWeak (EW) corrections, following the computation presented in [168]. Similarly to the pQCD counterpart, it is computed as a function of m_{ZZ} and incorporated in the MC-event weight. The impact of the EW corrections is negligible below $2m_Z$, but it gradually reduces the predicted yield by a factor that increases as m_{ZZ} increases.

The differential effects as a function of $m_{4\ell}$ of both k-factors is reported in Fig. 4.4.

Gluon-fusion

The non-resonant production of two Z boson through gluon fusion is generated at LO using `MCFM 7.0.1`, as detailed in Sec. 3.1.1. Similarly to the approach adopted for $q\bar{q} \to ZZ$, also for $gg \to ZZ$ a k-factor is introduced to account for the difference between the state-of-the-art theoretical computation and the perturbative order of the MC generator. The soft collinear approximation has demonstrated to describe accurately the cross section and the interference term for the gluon fusion ZZ production at NNLO in pQCD [169]. Additional calculations have further revealed that the k-factors are very similar at NLO for signal and background [170], and at NNLO for the signal and interference terms [171]. Hence, the k-factor is computed using the HNNLO v2 program [172–174] as a function of m_{ZZ} from the ratio of the NNLO and LO $gg \to H \to 2\ell 2\ell'$ cross sections for the predicted SM Higgs boson decay width of 4.07 MeV [26]. This NNLO/LO k-factor, computed for the signal, is then applied to the background [175]. It varies from approximately 2.0 to 2.4, and it is 2.27 at $m_{ZZ} = 125$ GeV. The effect can be seen in Fig. 4.4.

4.3.2 Reducible background

Among the different sources of background, the reducible component is the most challenging to model. This background receives contributions mainly from three sources: the misidentification of light hadrons as leptons; the semi-leptonic decays of heavy mesons;

and a sub-leading contribution from photon conversion in the detector volume. Although all these leptons are widely called *fake leptons*, with an abuse of language, the only real *fake* contribution comes from the first component, in the other cases we are dealing with real leptons that, unlike the signal leptons, are non-prompt. The main physics processes contributing to this signature are the Drell-Yan (Z + jets) production, $t\bar{t}$, WW + jets, Zγ + jets, and WZ + jets. Since the leading contribution is associated to the Z + jets process, this background is known in the CMS H $\rightarrow$ ZZ analyses as *Z+X background*. The signature of these processes is the production of two prompt leptons and two non-prompt/misidentified leptons. There is also a minor contribution from backgrounds with three prompt leptons and one non-prompt/misidentified leptons, such as for the WZ + jets component.

In contrast to the irreducible background, the reducible component cannot be estimated with MC simulations. State-of-the-art simulations do not model properly the Z+X background in the signal region, hence the need of using what one could claim to be the best MC samples we have at our disposal, data itself. The underlying idea of data-driven methods for the estimation of backgrounds is using Control Regions (CRs), i.e., regions of the phase of space that are orthogonal to the SR and enriched of the background events the method targets to estimate.

The data-driven algorithm used in the H $\rightarrow$ ZZ $\rightarrow 4\ell$ channel is usually referred to as *fake rate method* and it is actually the combination of two sub-methods: the Opposite-Sign (OS) and the Same-Sign (SS) method. The philosophy of the method and the structure of the CRs is the same for both method, what changes is how the CRs are populated. The underlying idea of the method is to reweigh events from the CR to the SR by the *fake rate*, i.e., the probability of a non-prompt/misidentified lepton to satisfy the tight lepton selections (cfr. Sec. 3.2), hence selected in the SR. The computation of the fake rate is also done with data, but in a different CR.

Workflow of the fake rate method

The CRs for the evaluation of the fake rates consists of a good Z candidate, i.e., satisfying all conditions outlined in Sec. 4.1 for a Z_1 candidate, and one additional loose lepton. These CRs will be denoted as Z+L. The additional loose lepton is used to probe the fake rate by assessing the frequency of passing the tight requirements. In order to mitigate the contamination due to prompt leptons from WZ and $t\bar{t}$ processes, each event must pass the condition on the missing transverse momentum of $p_{\mathrm{T}}^{\mathrm{miss}} < 25$ GeV. Moreover, the additional lepton and the opposite-sign lepton in the Z candidate must have an invariant mass greater than 4 GeV. This is the prototype of the Z+L CRs that will be used in the OS and SS methods with distinct supplementary requirements depending on the method. In both cases, the fake rates are computed separately in the endcap and in the barrel, separately for electrons and muons, in bins of the transverse momentum of the additional lepton, and independently for each data-taking period.

Once the fake rate is computed, it is applied to the Z + LL CR to obtain the estimation of the background in the SR. The definition of these CRs depends on the specific method and will be outlined later. The Z+LL CR is obtained by events passing the chain defined in Sec. 4.1 either without applying or requiring the failure of some selections on the pair of leptons forming the Z_2 candidate. In both CRs, the FSR recovery is performed with the same method of the SR (cfr. Sec. 3.3).

This is the common workflow of the OS and SS methods. The specifics of each approach will be detailed as follows.

OS method

The Z+L CR for the computation of the fake rates in the OS method has an additional requirement to the skeleton described above. The $|m_{Z_1} - m_{Z_0}| < 7$ GeV cut is introduced to reduce the bias in the fake rates from asymmetric photon conversion in the detector material, which predominantly populates the low-$m_{\ell\ell}$ region. The WZ contribution surviving to the p_T^{miss} cut is subtracted by using the corresponding MC sample. The fake rates obtained with 2018 data are reported in Fig. 4.6 both for muons and electrons. The *corrected* and *uncorrected* labels refer to the fake rate before and after the WZ subtraction.

The OS method employs two Z+LL CRs. The first one is the 2P2F region, which is populated by events with two leptons passing (2P) the tight requirements and forming a Z boson candidate, and two additional leptons failing (2F) the tight selections. This sample is expected to be populated by events with intrinsically two prompt leptons, like Drell-Yan and $t\bar{t}$. On the other hand, the second CR is the 3P1F region, which is populated by events with three leptons passing (3P) the tight requirements and two of them forming a Z boson candidate, and one lepton failing the tight requirements (1F). This region has inevitably an overlap with the 2P2F CR, and it contains events that have intrinsically three prompt leptons, like WZ and $Z\gamma^*$.

To compute the contribution to the SR from the 2P2F CR ($N^{Z+X}_{\text{SR from 2P2F}}$), each event is weighed by the product of the fake rates of the two additional leptons (f_3, f_4). In the following, to lighten the notation, the kinematic dependence of the fake rates will be omitted, but they do depend on the p_T and η of the lepton. Hence:

$$N^{Z+X}_{\text{SR from 2P2F}} = \sum_i^{N_{2P2F}} \left(\frac{f_3^i}{1-f_3^i} \frac{f_4^i}{1-f_4^i} \right) \tag{4.5}$$

where N_{2P2F} denotes the number of events in the 2P2F CR. The computation for the 3P1F CR follows the same logic:

$$N^{Z+X}_{\text{SR from 3P1F}} = \sum_i^{N_{3P1F}} \left(\frac{f_4^i}{1-f_4^i} \right) \tag{4.6}$$

While combining the two contributions, some precautions should be taken. The double-counting of 2P2F events in the 3P1F region ($N_{3P1F \cap 2P2F}$) should be avoided, and the contribution from the ZZ reducible background in the 3P1F region should be taken into account. The final estimation of the background with the OS method can be therefore written as:

$$N^{Z+X}_{\text{SR from OS}} = \sum_i^{N_{3P1F}} \frac{f_4^i}{1-f_4^i} - N_{3P1F \cap 2P2F} - N^{ZZ}_{3P1F} + \sum_j^{N_{2P2F}} \left(\frac{f_3^j}{1-f_3^j} \frac{f_4^j}{1-f_4^j} \right) \tag{4.7}$$

where the N^{ZZ}_{3P1F} term is computed with MC simulation, and the $N_{3P1F \cap 2P2F}$ factor is calculated as:

$$N_{3P1F \cap 2P2F} = \sum_i^{N_{2P2F}} \frac{f_3^i}{1-f_3^i} + \sum_i^{N_{2P2F}} \frac{f_4^i}{1-f_4^i} \tag{4.8}$$

SS method

In the SS method, comparing to the OS method, the Z+L CR utilised for calculating the fake rates implements a more relaxed criterion for the invariant mass of the Z boson, in alignment with the dilepton invariant mass requirement of the event selection algorithm.

Precisely, it is required that $40 < m_{Z_1} < 120$ GeV. Similarly to the OS method, the residual WZ contribution is subtracted using MC simulations.

However, the wider Z mass window in the SS method results in more events with asymmetric photon conversion contributing to the Z+e sample. An asymmetric photon conversion occurs when one low-p_T leg is not identified, hence the corresponding events will end up in the Z+e CR. To account for this enhancement and the resulting mismatch between the composition of the Z+L and the Z+LL CRs, a further correction is applied to the Z+e CR.

The strategy develops in two steps. The first step involves the Z+e CR itself. Four categories are defined, each characterised by different cuts on $|m_{Z_1} - m_{Z_0}|$ and $|m_{eZ_1} - m_{Z_0}|$ to account for varying contributions from conversion. The minimal amount of conversion is obtained by requiring $|m_{Z_1} - m_{Z_0}| < 7$ GeV, similarly to what is done in the OS method, whereas the maximal enrichment of electrons from conversion is obtained with $|m_{eZ_1} - m_{Z_0}| < 5$ GeV. Each event category is then binned in four (p_T, η) bins of the additional loose lepton, and in each bin, the fake rate and the $< N_{\text{missingHits}} >$ are computed. The $< N_{\text{missingHits}} >$ denotes the mean number of missing hits in the pixel detector, which is a sensitive observable to conversion. Since $< N_{\text{missingHits}} >$ is expected to grow linearly with the fraction of conversion, a linear dependence of the fake rate with respect to $< N_{\text{missingHits}} >$ is anticipated, as shown in Fig. 4.5. The second step involves measuring the $< N_{\text{missingHits}} >$ in the Z+ee CR for the additional leptons in each (p_T, η) bin. The corrected fake rate is then determined from the linear relation derived in the first step. The determination of these corrected fake rates mostly suffers from the limited statistics of the Z+e CRs, which translates into a large uncertainty on $< N_{\text{missingHits}} >$. This source will be one of the leading uncertianty in the SS estimation. The fake rates are reported in Fig. 4.6. The computation for muon fake rates follows the same logic of the OS method.

In the SS method, the Z+LL CR is defined as 2P2Lss, which is populated by events with two leptons passing (2P) the tight selections and two additional same-sign loose leptons (2Lss). While the OS method is forced to require failing leptons in the Z+LL CR in order to be orthogonal to the SR, the SS method, by considering same-sign leptons, can afford to simply asking loose additional leptons without the risk of collecting signal events. The main benefit is that the SS CR is much more populated to that of the OS method. The computation of events in the SR from the SS method is performed by weighing each event in the 2P2Lss CR as follows:

$$N^{\text{Z+X}}_{\text{SR from SS}} = \left(\frac{\text{OS}}{\text{SS}}\right) \sum_{i}^{N_{\text{2P2Lss}}} f_3^i \cdot f_4^i \tag{4.9}$$

where $\frac{\text{OS}}{\text{SS}}$ is the correction factor that accounts for the difference between the SS and OS phase space, and it is computed as the ratio between the yields of the 2P2Lss and the 2P2Los CRs.

Uncertainties

The estimation of the normalisation of the Z+X background with the OS and SS method explained above is affected by three different sources of uncertainties, which are summed in quadrature.

- A statistical component due to the number of events in both the Z+L and Z+LL CR. In the OS method, it ranges 7-10% in the muon channels and 4-6% in the electron channels. In the SS method, it ranges 1.4-2% in the muon channels and 0.5-0.9% in the electron channels. The larger number of events populating the SS CR is clearly visible in the lower statistical uncertainty.

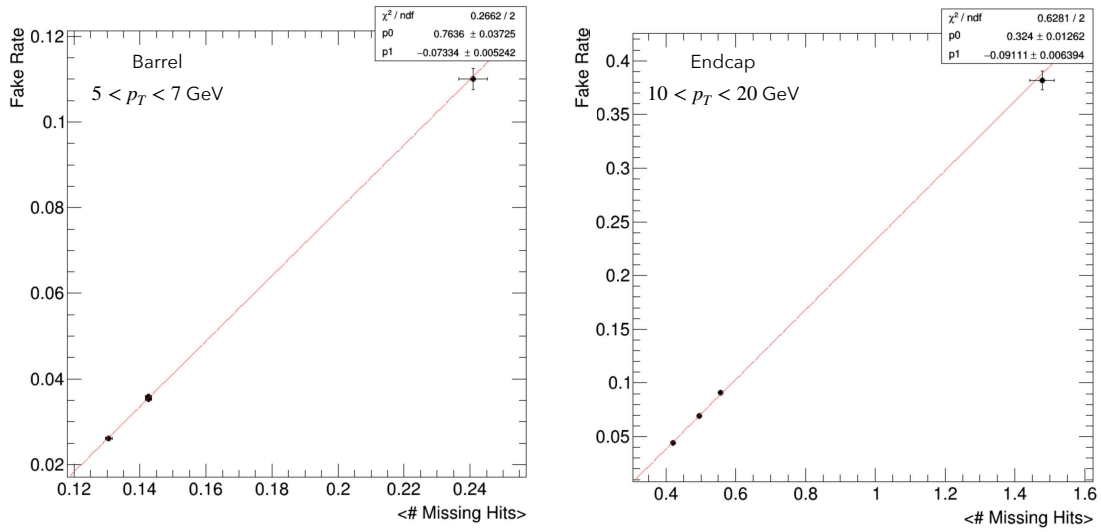

Figure 4.5: Linear relation between the fake rate and the mean number of missing hits $< N_{\text{missingHits}} >$ in the Z+L CR of the SS method. Each point represents one Z+L sub-CR, which is identified by a different fraction of photon conversion.

- A dominant systematic component due to the different composition of the CR where the fake rates are computed (Z+L) and the CRs where the fake rates are applied (Z+LL). Both the OS and SS implements strategies to reduce the differences, but a residual bias is inevitable. This contribution is computed with MC samples and estimated at 30%.

- A sub-dominant systematic component given by the variation of the expected yield considering up and down variation of the FR. When combined with the previous component, the total systematic uncertainty ranges 30-40%, depending on the final state and the data-taking period.

Combination and shape

The OS and SS methods provide two independent estimations for the reducible background. The numbers are reported in Tab. 4.1 for the entire $H \to ZZ \to 4\ell$ phase space. These two estimates are found to be in agreement within their uncertainties and eligible for combination. The combined value is obtained by means of a weighted average, taking into account their respective uncertainties.

However, the measurement of differential fiducial cross sections required the assessment of both the yield of the background and its shape across different kinematic bins. In practice, only the SS method is used for this purpose, with a correction to match the combined estimation when summing up over all bins. This choice is primarily motivated by the issue of lower number of events in the OS CRs. When evaluating the Z+X background in poorly populated kinematic bins, such as those at large transverse momentum, there is the potential risk for Eq. 4.7 to provide negative values. Statistical fluctuation may make the values to be subtracted larger than the others.

Practically, to estimate the reducible background within a specific kinematic bin of the SR, the same kinematic region is selected in the 2P2Lss CR, and each event is weighed by the factor:

$$w_{\text{evt}} = \left(\frac{\text{Combined yield}}{\text{SS yield}}\right)\Bigg|_{m_{4\ell}>70\text{ GeV}} \cdot \left(\frac{\text{OS}}{\text{SS}}\right) \cdot f_3^{SS} \cdot f_4^{SS}. \tag{4.10}$$

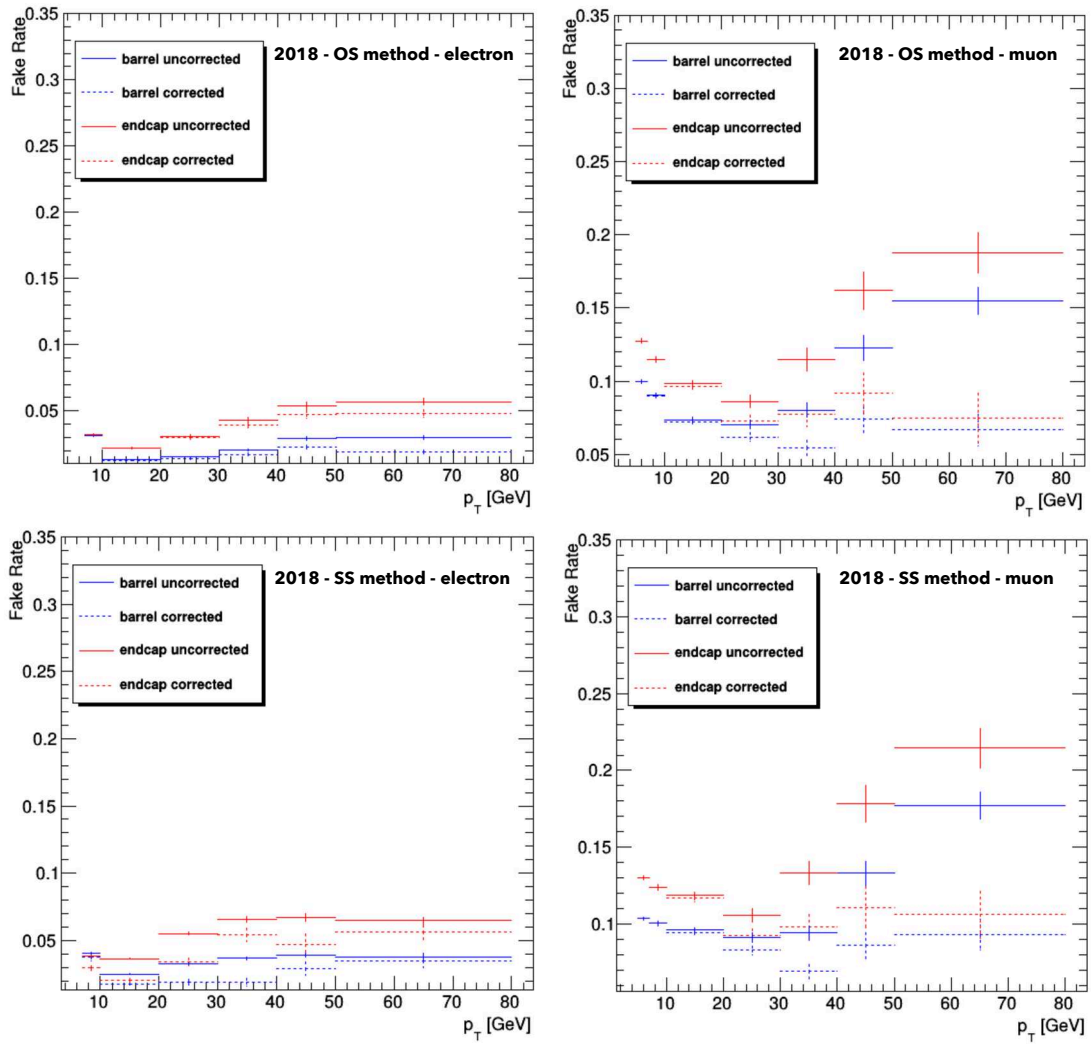

Figure 4.6: (Top) Fake rates for the OS method for electrons (left) and muons (right) computed with 2018 data. The corrected fake rates refer to the values after the subtraction of the WZ contribution. (Bottom) Fake rates for the SS method for electrons (left) and muons (right) computed with 2018 data. The corrected fake rates refer to the values after the subtraction of the WZ contribution, both for muons and electrons, and the additional correction to reduce the contamination from asymmetric photon conversion, only for electrons.

The last three terms correspond to Eq. 4.9, whereas the first ratio corrects the SS yield in order to account for the combination between the OS and SS methods. The final yield is then computed as the summation of all event weights, while the shape is determined by the normalised binned template. The computation is performed independently in each kinematic bin, final state, and data-taking period. Figure 4.7 shows the Z+X shapes in different bins of the transverse momentum of the Higgs boson in the $4e$ final states with 2017 data, along with the shapes for the irreducible backgrounds extracted from MC samples.

4.4 Maximum likelihood unfolding

The physics analysis presented in this thesis aims to measure fiducial cross sections. The general approach has already been introduced and contextualised in Sec. 1.5. In this

Table 4.1: Estimate of the normalisation of the Z+X background in the $m_{4\ell} > 70$ GeV mass range obtained with both the SS and OS method in the three data taking periods and four possible final states.

		2016	2017	2018
4μ	OS	30.6 ± 9.8	31.5 ± 10.0	47.1 ± 14.6
	SS	32.2 ± 9.8	34.8 ± 10.7	49.6 ± 15.1
4e	OS	18.2 ± 5.5	13.3 ± 4.1	20.6 ± 6.3
	SS	13.3 ± 5.7	11.1 ± 4.0	14.9 ± 5.3
$2e2\mu$	OS	29.6 ± 9.3	23.8 ± 7.7	34.1 ± 10.7
	SS	26.4 ± 8.1	27.4 ± 8.4	39.0 ± 11.9
$2\mu2e$	OS	21.6 ± 6.6	19.5 ± 5.9	26.1 ± 7.9
	SS	16.8 ± 6.7	14.7 ± 5.4	19.6 ± 7.0

section, it is explained how that general framework is implemented in the H → ZZ → 4ℓ channel.

The core of fiducial cross sections is the unfolding. Figure 4.8 defines the two levels that are bridged by the unfolding procedure. To define these levels, it is instructive to trace the main physics stages of a MC simulation. The first step is the generation of the partonic process, which is then interfaced to the parton shower and hadronisation. This constitutes the generator-level (*gen-level* for short), representing what one would measure under ideal experimental conditions, characterised by an ideal detector and a perfect reconstruction process. Beyond this, particles produced at gen-level travel into the detector, release energy deposits in the active material that are then converted to electric signals. These signals, passing though a reconstruction algorithm, are identified as originating from one type of particle or an other. This is what one actually measures in a real-life experiment. This is usually referred to as reconstruction-level (*reco-level* for short). The unfolding is exactly the inverse of this chain of steps. Starting from experimental data, which are by definition reco-level, the aim is to remove detector and reconstruction effects to obtain the gen-level information.

Different unfolding methods have been developed and utilised in high-energy physics. The H → ZZ → 4ℓ channel, as all other Higgs analyses in CMS, employs a *maximum likelihood unfolding*. As the name itself suggests, it is necessary to introduce the likelihood function. The output of an experiment is a set of observations $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$. For simplicity, suppose that this dataset populates one single channel, and this channel receives contributions from one signal process, which is described by the $f_S(y)$ model, and one background process, which is described by the $f_B(y)$ model. The total number of expected signal events is denoted by S, while for background is denoted by B. It is common to introduce a parameter μ such that for $\mu = 0$ the background-only hypothesis is found, while for $\mu = 1$ the nominal signal plus background is retrieved. The probability for obtaining n events is composed of two components. The first one is the Poisson probability of obtaining n events with an expectation of $\mu S + B$. The second component is the probability density of obtaining a certain value y_i based on the relative mixture of the signal and background contributions for a given value of μ. The combination of these two components gives the so-called *marked Poisson model*:

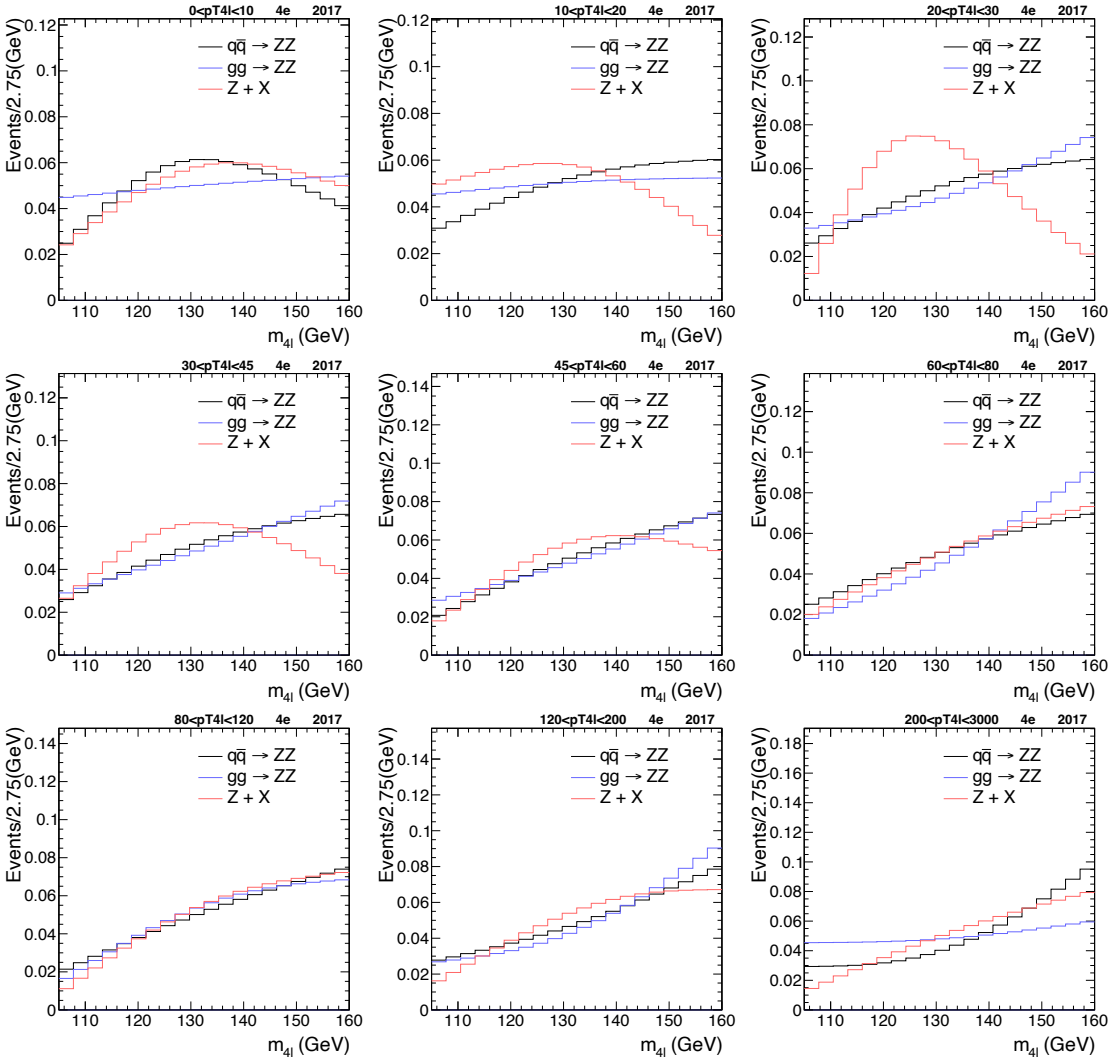

Figure 4.7: Background shapes in different bins of the transverse momentum of the Higgs boson in the 4e final state for the 2017 period. The shape is estimated in the mass range $105 < m_{4\ell} < 160$ GeV. The irreducible background, i.e., $q\bar{q} \to ZZ$ and $gg \to ZZ$ are estimated with MC, whereas the reducible background Z+X is estimated with data as described in the text.

$$
\begin{aligned}
\mathcal{P}(\boldsymbol{y}|\mu) &= \mathrm{Pois}(n|\mu S + B) \prod_{i=1}^{n} \left[\frac{\mu S f_S(y_i) + B f_B(y_i)}{\mu S + B} \right] \\
&= \frac{(\mu S + B)^n \mathrm{e}^{-(\mu S + B)}}{n!} \prod_{i=1}^{n} \left[\frac{\mu S f_S(y_i) + B f_B(y_i)}{\mu S + B} \right] \\
&= \frac{\mathrm{e}^{-(\mu S + B)}}{n!} \prod_{i=1}^{n} [\mu S f_S(y_i) + B f_B(y_i)] \\
&\equiv \prod_{i=1}^{n} \mathrm{Pois}(n|\mu S f_S(y_i) + B f_B(y_i))
\end{aligned}
\tag{4.11}
$$

In a real-life experiment, data are measured and hence fixed. The statistical inference aims to estimate μ, which is usually denoted with the name of *parameter of interest*. The likelihood function is numerically equivalent to the above-mentioned probability, with $\boldsymbol{y}$

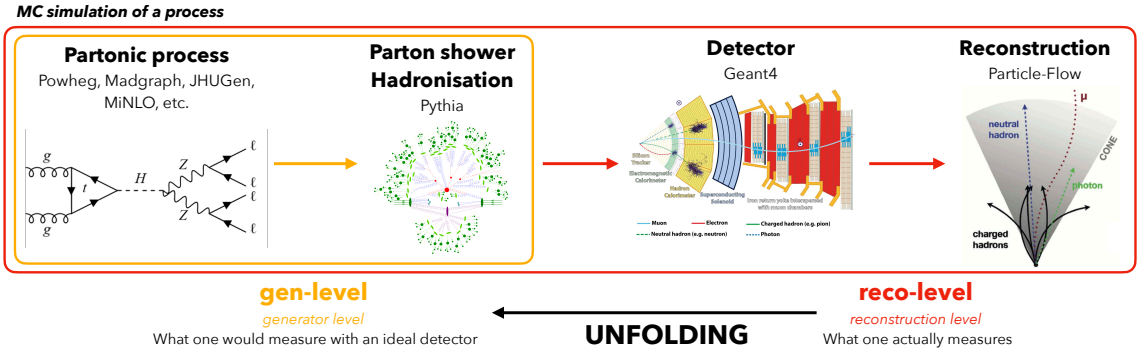

Figure 4.8: The process proceeding from the left to the right shows the most important physics steps that an event undergo in a high-energy physics experiment. The generator-level (gen-level) is composed of the partonic process and the following parton shower and hadronisation. The reconstruction level (reco-level) is the output of an experiment, after the interaction with the detectors and the reconstruction algorithm. The unfolding is the opposite process. Starting from experimental data, which are intrinsically reco-level, the unfolding aims to estimate the gen-level distribution.

fixed. Consequently, the likelihood for a single channel c, it may be a final state, kinematic bin, category, and so forth, can be written as:

$$\mathcal{L}_c(\mu) = \prod_{i=1}^{n} \text{Pois}(n|\mu S f_S(y_i) + B f_B(y_i)) \tag{4.12}$$

For the specific case of the analysis presented in this thesis, the form of the term $\mu S f_S(y_i) + B f_B(y_i)$ will be detailed in Sec. 4.7.

Equation 4.12 represents a simplified model. Real-life scenarios are more complicated. There are usually many channels, and the description of the signal and background can depend on many other parameters. In the case under study here, the parameters of interest are fiducial cross sections ($\boldsymbol{\mu}_{\text{fid}} = \boldsymbol{\sigma}_{\text{fid}}/\boldsymbol{\sigma}_{\text{fid}}^{\text{SM}}$) and the other parameters, usually named *nuisance parameter* $\boldsymbol{\theta}$, affect both signal and background contributions.

$$\mathcal{L}(\boldsymbol{\mu}_{\text{fid}}, \boldsymbol{\theta}) = \prod_{c=1}^{n_c} \mathcal{L}_c(\boldsymbol{\mu}_{\text{fid}}, \boldsymbol{\theta}) \prod_{k=1}^{n_k} p_k(\tilde{\theta_k}|\theta_k) \tag{4.13}$$

Here, $p_k(\tilde{\theta_k}|\theta_k)$ represents the pdf of the uncertainty on the constrained value $\tilde{\theta_k}$ of the nuisance parameter θ_k, which is usually obtained with auxiliaries measurements. A detailed discussion of the nuisances entering this analysis is delayed to Sec. 4.8.

Once the likelihood function is formed, the fiducial cross sections are extracted using an asymptotic approach [176] with a *profile likelihood ratio test statistic* [177], defined as:

$$q(\boldsymbol{\mu}_{\text{fid}}) = -2\Delta \ln \mathcal{L}(\boldsymbol{\mu}_{\text{fid}}) = -2 \ln \frac{\mathcal{L}(\boldsymbol{\mu}_{\text{fid}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}_{\text{fid}}})}{\mathcal{L}(\hat{\boldsymbol{\mu}}_{\text{fid}}, \hat{\boldsymbol{\theta}})} \tag{4.14}$$

In this equation, the $\hat{\boldsymbol{\mu}_{\text{fid}}}$ and $\hat{\boldsymbol{\theta}}$ represents the maximum likelihood estimators of the parameters of interest $\boldsymbol{\mu}_{\text{fid}}$ and nuisance parameters $\boldsymbol{\theta}$. The numerator denotes the *conditional maximum likelihood*, where $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}_{\text{fid}}}$ corresponds to the values of the nuisance parameters that maximise the likelihood function for the specific $\boldsymbol{\mu}_{\text{fid}}$. The Wilks theorem [178] ensures that the distribution of $q(\boldsymbol{\mu}_{\text{fid}})$ can be approximated with a χ^2 distribution with m degrees of freedom, where m is the number of parameters of interest in the model. Therefore, the Confidence Level (CL) intervals on the measured values of $\boldsymbol{\mu}_{\text{fid}}$ can be extracted

directly from the scan of projections the test statistic. The 68% CL interval is obtained for $-2\Delta \ln \mathcal{L}(\mu^i_{\text{fid}}) < 1$, while the 95% CL interval is obtained for $-2\Delta \ln \mathcal{L}(\mu^i_{\text{fid}}) < 3.84$.

The maximum likelihood-based unfolding has two main advantages compared to other methods. First, the unfolding is carried out simultaneously with the extraction of the fiducial cross sections. In contrast, all the other methods first unfold data and then measure the cross section. Second, this approach permits a consistent treatment of the systematic uncertainties by implementing corresponding nuisance parameters in the likelihood that are profiled in the fitting procedure. To estimate the systematic uncertainties, other methods usually rely on performing the unfolding process many times by computing new response matrices varying the quantities by one standard deviation up and down.

4.5 Definition of the fiducial phase space

The unfolding method presented in the previous section represents just one component of the fiducial cross section framework. The second key element to introduce is in the name itself. In order to minimise the extrapolation corrections, which would inevitably rely on a specific model, the data are not unfolded to the entire gen-level phase space, but to a subset of it. Cross sections are measured in a *fiducial phase space* defined to match closely the experimental acceptance of the reco-level selections. This fiducial phase space is defined by imposing criteria on lepton kinematics, isolation, and event topology.

The selection flow retraces the one at reco-level: the Z boson candidates are defined starting from muons and electrons; the Z candidates are combined to form ZZ candidates; after the application of some quality conditions, the Higgs boson candidate is identified. Leptons at fiducial level are considered as *dressed*, i.e., after the recovery of FSR photons. The FSR algorithm recovers all photons within a cone of radius 0.3 around the lepton, if the mother of the photon has the same flavour of the lepton. Leptons are also considered as *isolated*. Lepton isolation is ensured by requiring the scalar sum of the p_{T} of all stable particles, i.e., those particles not decaying in the detector volume, within a cone of radius $\Delta R = 0.3$ to be less than 0.35 times the p_{T} of the lepton. Neutrinos, FSR photons, and leptons (electrons and muons) are not included in the computation of the isolation sum to bring the definition of the fiducial phase space closer to the reconstruction level selection and enhance the model independence of the measurements, following the findings of Ref. [179]. The inclusion of the isolation in the fiducial phase space definition reduces the differences in signal selection efficiency between different models. The event is retained if the leading (sub-leading) lepton has $p_{\text{T}} > 20\,(10)$ GeV. Additional electrons (muons) that may be present in the event are required to have $p_{\text{T}} > 7\,(5)$ GeV and $|\eta| < 2.5\,(2.4)$.

Events passing these requirements are retained if they have at least two same-flavour opposite-sign lepton pairs. The pair with invariant mass closest to the world-average Z boson mass [26] is labelled as Z_1 and it must have $40 < m_{\text{Z}_1} < 120$ GeV. The second Z boson candidate is referred to as Z_2 and it must have $12 < m_{\text{Z}_2} < 120$ GeV. Each lepton pair ℓ_i, ℓ_j must be separated by $\Delta R(\ell_i, \ell_j) > 0.02$, while any opposite-sign lepton pair must satisfy $m_{\ell^+\ell'^-} > 4$ GeV, reflecting the selection criteria used at reconstruction level.

Jets do not enter in the definition of the fiducial phase space, but they should be defined at gen-level when dealing with jet observables. Jets at fiducial level are defined with the anti-kt clustering algorithm with a distance parameter of 0.4 out of stable particles, excluding neutrinos. They are retained if they satisfy $p_{\text{T}}^{\text{jet}} > 30$ GeV and $|\eta^{\text{jet}}| < 4.7$, similarly to the condition used at reconstruction level. Only jets with no leptons inside a cone of radius 0.4 are kept.

The definition of the fiducial phase space is summarized in Table 4.2.

Table 4.2: Requirements for the $H \to ZZ \to 4\ell$ fiducial phase space

Lepton kinematics and isolation	
Leading lepton p_T	$p_T > 20$ GeV
Sub-leading lepton p_T	$p_T > 10$ GeV
Additional electrons (muons) p_T	$p_T > 7(5)$ GeV
Pseudorapidity of electrons (muons)	$\lvert\eta\rvert < 2.5$ (2.4)
p_T sum of all stable particles within $\Delta R < 0.3$ from lepton	$< 0.35 \cdot p_T$
Event topology	
Existence of at least two same-flavor OS lepton pairs, where leptons satisfy criteria above	
Inv. mass of the Z_1 candidate	$40 < m_{Z_1} < 120$ GeV
Inv. mass of the Z_2 candidate	$12 < m_{Z_2} < 120$ GeV
Distance between selected four leptons	$\Delta R(\ell_i, \ell_j) > 0.02$ for any $i \neq j$
Inv. mass of any opposite sign lepton pair	$m_{\ell^+\ell'^-} > 4$ GeV
Inv. mass of the selected four leptons	$105 < m_{4\ell} < 160$ GeV

4.6 Kinematic observables

The fiducial cross sections analysis presented in this thesis aims to a complete characterisation of the $H \to ZZ \to 4\ell$ decay channel. In addition to the inclusive fiducial cross section, 32 differential kinematic variables are implemented describing the production, the decay of the Higgs boson, and the associated jet activity. A dedicated binning optimisation is performed and is described in the following. In addition to one dimensional (1D) observables, some of the kinematic observables are combined in double-differential (2D) observables in order to probe specific phase space regions and nail down possible BSM effects.

4.6.1 Production observables

The two quintessential differential observables to describe the Higgs boson production are the transverse momentum (p_T^H) and rapidity ($|y_H|$). The transverse momentum is one of the most precious observables for theoreticians, as it is extensively used as a benchmark to test theoretical calculations. The ggH production modes, being a loop-induced process, is sensitive to particles inside the loop, and their effects will be mainly visible in the p_T spectrum. For instance, the low-p_T region is sensitive to variations of the Higgs boson couplings to light quarks [180, 181]. In addition, this region of the spectrum exhibits a Sudakov peak due to initial state radiation, it is therefore a proper region to test the resummation. On the other hand, in the region of the spectrum above $p_T^H > m_t$, the approximation of considering infinite the top mass fails, and its effect should be taken into account. The high-p_T regime is also a sensitive region to BSM effects due to the reduced presence of SM events.

The main interest for the rapidity distribution is related to its sensitivity to the Parton Distribution Functions (PDFs) of the proton. This can be seen easily with a simple example. In a proton-proton collision at a centre-of-mass energy equal to $\sqrt{s}$, each proton brings an energy equal to $\sqrt{s}/2$. By denoting with x_1 and x_2 the fraction of energy of the original proton that each parton brings to the interaction, the following relations are valid:

$$\begin{cases} E = (x_1 + x_2)\frac{\sqrt{s}}{2} \\ p_z = (x_1 - x_2)\frac{\sqrt{s}}{2} \end{cases}$$

where p_z is the net longitudinal momentum of the colliding parton-parton system. These relations were already used in Sec. 2.2.1. From the above system of equations, it follows

that:

$$\frac{x_1}{x_2} = \frac{E + p_z}{E - p_z} = e^{2y}$$

where the last equality follows from the definition of rapidity (cfr Eq. 2.13). Therefore, the rapidity distribution gives insights in the momentum of the partons participating in the hard collision, hence in the proton PDF.

The production of the Higgs boson can be associated to the presence of jet activity. The presence and the number of jets is a powerful handle to target different production mechanisms. In the SM, events without jet activity are dominated by the ggH production mode. The production of the Higgs boson via VBFH is distinctly characterised by the presence of two hard jets, well separated in pseudorapidity and with large invariant mass. Looking at the N_{jets} distribution, the VBFH events may also be accompanied by the presence of one single jet, as one of the two can be lost in the jet reconstruction. The contribution from Higgs boson production in association with top quarks becomes more important in the largest jet multiplicity. An interesting feature of the VBFH production mode can be noted in the $\Delta\eta_{jj}$ distribution. As said before, the VBFH topology is characterised by the emission of two jets with a large separation in η, and this feature is usually exploited to isolate such events. However, the $\Delta\eta_{jj}$ distribution from the VBFH simulation features a bump around zero, due to events with low-m_{jj}. This is due to the hard gluon radiation radiated from the leading jet, sufficiently separated from the parent jet to be reconstructed as a separate jet. This phenomenon is extensively discussed in [182]. Additionally, jet-observables can also provide information about the CP properties of the Higgs boson though the azimuthal angle between the leading and the subleading jet $\Delta\phi_{jj}$ defined as:

$$\Delta\phi_{jj} = \frac{(\hat{j}_{T_1} \times \hat{j}_{T_2}) \cdot \hat{z}}{|(\hat{j}_{T_1} \times \hat{j}_{T_2}) \cdot \hat{z}|} \cdot \frac{(\vec{j}_1 - \vec{j}_2) \cdot \hat{z}}{|(\vec{j}_1 - \vec{j}_2) \cdot \hat{z}|} \cdot \cos^{-1}(\hat{j}_{T_1} \cdot \hat{j}_{T_2}). \tag{4.15}$$

This definition is invariant under the exchange of the two jets. The vectors $\vec{j}_{1,2}$ represent the direction of the leading and subleading jet in the laboratory frame, while the unit vectors $\hat{j}_{T_{1,2}}$ represent the corresponding transverse component. This definition is also independent of the choice of the positive z axis direction, $\hat{z}$. The corresponding distribution is expected to be symmetric in the SM, while it becomes asymmetric in the case of CP violation. Other meaningful jet-observables are the transverse momentum of the leading and sub-leading jets, as they probe the theoretical modelling of hard quark and gluon radiation. Some of the jet variables can be seen in Fig. 4.9.

Additionally, the jet activity can be investigated by using rapidity-weighed jet-observables $\mathcal{T}_C^{max}$ and $\mathcal{T}_B^{max}$. These are defined, following Ref. [183], as:

$$\mathcal{T}_C^{max} = \max_j \left(\frac{\sqrt{E_j^2 - p_{z,j}^2}}{2\cosh(y_j - y_H)} \right), \tag{4.16}$$

$$\mathcal{T}_B^{max} = \max_j \left(m_T^j e^{-|y_j - y_H|} \right), \tag{4.17}$$

where y_j and m_T^j denote the rapidity and transverse mass of the jet, defined from its mass m and momentum p as $m_T^j = \sqrt{m^2 + p_x^2 + p_y^2}$, while y_H is the rapidity of the Higgs boson. The value of each observable is computed for each jet in the event and the maximum value is taken for each event. Since their resummation structure is different from the canonical

p_{T}^{j}, they give complementary information on the properties of jets in an event and can be used as a test of QCD resummation. The 0-jet phase space can be redefined using these observables. The events with no jets are defined as the ones with $\mathcal{T}_{\mathrm{C}}^{\mathrm{max}} < 15$ GeV and $\mathcal{T}_{\mathrm{B}}^{\mathrm{max}} < 30$ GeV, where the values of these cuts are chosen accordingly to the findings of Ref. [183]. In the following, these events will be denoted as 0-jet$|\mathcal{T}_{\mathrm{C}}^{\mathrm{max}}$ and 0-jet$|\mathcal{T}_{\mathrm{B}}^{\mathrm{max}}$, respectively.

The properties of the H+jet(s) system are also studied by measuring differential cross sections in bins of the transverse momentum and invariant mass of the four-lepton system plus either the leading jet (m_{Hj}, $p_{\mathrm{T}}^{\mathrm{Hj}}$) or the leading and subleading jet (m_{Hjj}, $p_{\mathrm{T}}^{\mathrm{Hjj}}$). These variables can only be defined for events with at least one or two jets, respectively. In all other cases, an underflow bin is introduced to consider all events for which the observable is undefined.

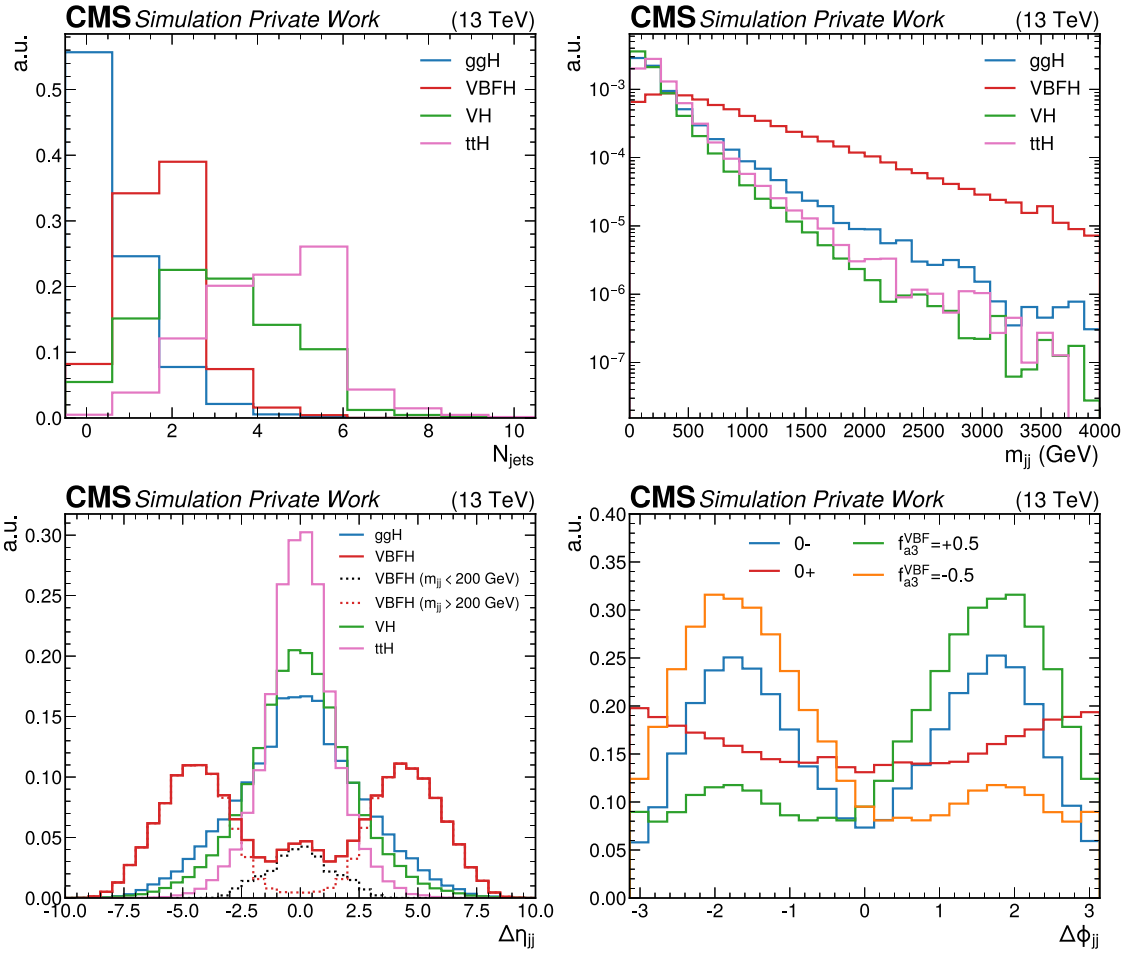

Figure 4.9: Distributions for the jet multiplicity $\mathrm{N_{jets}}$(top left), the invariant mass m_{jj} (top right), the difference in pseudorapidity $\Delta\eta_{\mathrm{jj}}$ between the leading and subleading jet (bottom left), and the difference in azimuthal angle $\Delta\phi_{\mathrm{jj}}$ between the leading and subleading jet. The shape is extracted in the mass range $105 < m_{4\ell} < 160$ GeV. All distributions are normalised to unity. The top left, top right, and bottom left plots compare the shape of the main production modes of the Higgs boson, where VH = $\mathrm{W^+H}$ + $\mathrm{W^-H}$ + ZH. The $\Delta\phi_{\mathrm{jj}}$ frame shows the distribution only for the VBFH production mode under different CP hypothesis for the Higgs boson. The notation $f_{\mathrm{a3}}^{\mathrm{VBF}}$ is introduced in Eq. 4.22. In this case, the green and orange lines indicate a mixed CP-state.

4.6.2 Decay observables

After the inspection of the production of the Higgs boson, the decay side should be investigated. The study of the decay products is a portal to obtain information about the Higgs boson couplings and its spin-parity properties. In addition to the invariant mass of the four leptons, $m_{4\ell}$, and the invariant masses of the two Z bosons, m_{Z_1} and m_{Z_2}, the decay of the Higgs boson to four leptons is fully described by five angles, three of them related to fermion kinematics (ϕ, θ_1, θ_2), and two of them connecting the production and the decay (ϕ_1, θ^*). Their schematic representation can be found in Fig. 4.10 and they are defined as follows.

- θ^* is defined in the Higgs boson rest frame and represents the angle between the incoming partons and the axis of the $H \to ZZ$ decay.
- ϕ_1 is defined in the Higgs boson rest frame and denotes the angle between the plane containing the incoming partons and the axis of the $H \to ZZ$ decay, and the plane of the Z_1 decay.
- θ_1 and θ_2 are defined in the Z_1 and Z_2 rest frames, respectively, as the angles between the Z boson direction in the Higgs boson rest frame and the direction of the negative decayed lepton.
- ϕ is defined in the Higgs boson rest frame as the angle between the decay planes of the two Z bosons.

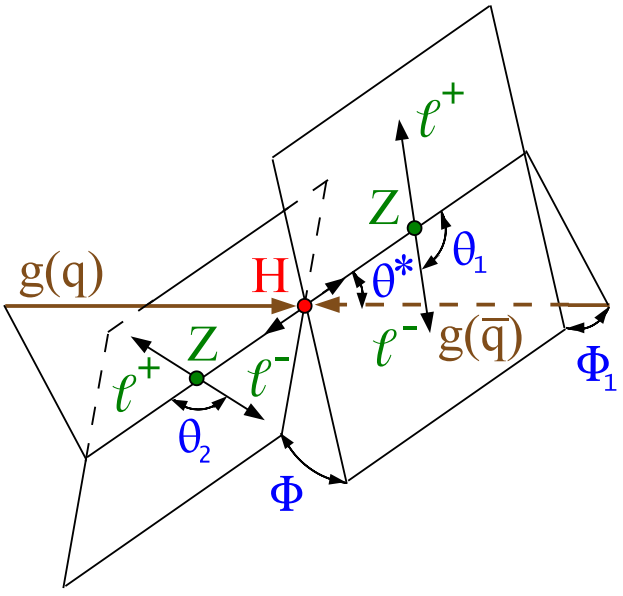

Figure 4.10: Schematic representation of the $gg/q\bar{q} \to H \to ZZ \to 4\ell$ process. The incoming partons are shown in brown and the intermediate or final-state particles are shown in green. The Higgs boson is represented in red. The five angles depicted in blue are considered in the analysis and detailed in the text.

The angles θ^*, θ_1, and θ_2 are defined in the $[0,\pi]$ range, while the angles ϕ_1 and ϕ are defined in the $[-\pi, \pi]$ range. More details about the definition of the angles can be found in Ref. [148]. For consistency with other CMS publications, and since they are defined as the scalar product of unit vectors, instead of measuring θ^*, θ_1, and θ_2, their cosine will be quoted. The decay variables defined above are hereafter referred to as

$\mathbf{\Omega}(\theta^*, \theta_1, \theta_2, \phi, \phi_1, m_{Z_1}, m_{Z_2}|m_{4\ell})$, defining the seven degrees of freedom that completely characterise the $H \to ZZ \to 4\ell$ process.

While the distribution of $\cos\theta_1$, $\cos\theta_2$, and ϕ are expected to be symmetric in the SM, they become asymmetric in the presence of CP violation. When the phase of a CP-violating term in the Lagrangian is purely real relative to the SM coupling, the distribution of the angle ϕ is expected to be asymmetric. Conversely, if this phase is purely imaginary, it results in asymmetry in the distributions of $\cos\theta_1$ and $\cos\theta_2$. In addition to it being advantageous to increase the number of events in each bin, these considerations justify the need of not using the absolute value for such angular observables. For a spin-zero particle, such as the SM Higgs boson, the distributions of $\cos\theta^*$ and ϕ_1 are trivially flat, but this does not hold true for higher-spin states [149] or when accounting for detector effects.

Matrix-Element discriminants

A special class of variables targetting the decay of the Higgs boson are the so-called matrix-element discriminants. MC generators in high-energy physics commonly rely on calculating matrix elements, which provide the probability for a quantum event to happen. The Matrix Element Likelihood Approach (MELA) package [148–152] can be employed to extract the LO matrix element probabilities from the `JHUGen` and `MCFM` generators - already used for the MC generation of signal and background events. Thus, MELA takes as input the kinematic information of an event and allows the computation of the matrix element probability $\mathcal{P}_i$ for an event to arise from a process i, given the value of the reconstructed invariant mass of the four-lepton system $m_{4\ell}$, as a function of $\mathbf{\Omega}$, defined in Sec. 4.5. These probabilities are computed exploiting all degrees of freedom of an event and therefore retain the maximal information for the description of the underlying physics contained therein. The Neyman-Pearson lemma [184] asserts that the most powerful way to distinguish between two hypothesis is by using the ratio, or a function thereof, of the probabilities. Consequently, the $\mathcal{P}_i(\mathbf{\Omega})$ probabilities computed with MELA can be used to construct likelihood-ratio-like matrix element discriminants to discriminate between two physics processes a and b, may them be two different production mechanisms of the SM Higgs boson or the test of a BSM hypothesis against the SM scenario. From these considerations, the discriminant confronting alternative (alt) hypotheses can be defined as:

$$\mathcal{D}_{\text{alt}}\left(\mathbf{\Omega}\right) = \frac{\mathcal{P}_{\text{a}}\left(\mathbf{\Omega}\right)}{\mathcal{P}_{\text{a}}\left(\mathbf{\Omega}\right) + \mathcal{P}_{\text{b}}\left(\mathbf{\Omega}\right)}. \tag{4.18}$$

This definition of the discriminant ensures that its value is always bounded between 0 and 1. Similarly, if the two hypotheses interfere (int) with each other, a second discriminant can be defined as:

$$\mathcal{D}_{\text{int}}\left(\mathbf{\Omega}\right) = \frac{\mathcal{P}_{\text{int}}\left(\mathbf{\Omega}\right)}{2\,\sqrt{\mathcal{P}_{\text{a}}\left(\mathbf{\Omega}\right)\;\mathcal{P}_{\text{b}}\left(\mathbf{\Omega}\right)}}, \tag{4.19}$$

which result to be always bounded between -1 and 1. The denominator is chosen in that form to reduce the correlation with $\mathcal{D}_{\text{alt}}$. The interference probability $\mathcal{P}_{\text{int}}\left(\mathbf{\Omega}\right)$ cannot be directly derived from the generator. Instead, it should be computed by generating the probability under the hypothesis a, the probability under the hypothesis b, and the combined probability under both hypotheses. The combined probability is given by:

$$\mathcal{P}_{a+b} = |\mathcal{A}_a + \mathcal{A}_b|^2 = |\mathcal{A}_a|^2 + |\mathcal{A}_b|^2 + 2\Re(\mathcal{A}_a^*\mathcal{A}_b) = \mathcal{P}_a + \mathcal{P}_b + \mathcal{P}_{\text{int}}, \tag{4.20}$$

where $\mathcal{A}$ represent the matrix element. Therefore, $\mathcal{P}_{\text{int}}$ is obtained by subtracting $\mathcal{P}_a$ and $\mathcal{P}_b$ from $\mathcal{P}_{a+b}$.

These kinds of matrix element discriminants have been widely used in the context of the $H \to ZZ \to 4\ell$ analyses, from the measurement of the Higgs boson properties Ref. [42]

up to the constraints on its anomalous couplings Ref. [128].

A total of six matrix element discriminants are included among the differential fiducial observables. These discriminants are sensitive to different anomalous couplings of the Higgs boson to vector bosons, where the SM is always the reference hypothesis. The most general scattering amplitude describing the interaction between a spin-zero (H) boson and two spin-one gauge bosons V_1 and V_2 can be written, following the conventions of Ref. [128], as:

$$A(\mathrm{HV_1V_2}) = \frac{1}{v}\left[a_1^{\mathrm{VV}} + \frac{\kappa_1^{\mathrm{VV}} q_{\mathrm{V1}}^2 + \kappa_2^{\mathrm{VV}} q_{\mathrm{V2}}^2}{\left(\Lambda_1^{\mathrm{VV}}\right)^2} + \frac{\kappa_3^{\mathrm{VV}}(q_{\mathrm{V1}}+q_{\mathrm{V2}})^2}{\left(\Lambda_Q^{\mathrm{VV}}\right)^2}\right] m_{\mathrm{V1}}^2 \epsilon_{\mathrm{V1}}^* \epsilon_{\mathrm{V2}}^* + \\ + \frac{1}{v} a_2^{\mathrm{VV}} f_{\mu\nu}^{*(1)} f^{*(2),\mu\nu} + \frac{1}{v} a_3^{\mathrm{VV}} f_{\mu\nu}^{*(1)} \tilde{f}^{*(2),\mu\nu}, \quad (4.21)$$

where v is the vacuum expectation value of the Higgs potential, $f^{(i)\mu\nu} = \epsilon_{\mathrm{V}i}^{\mu} q_{\mathrm{V}i}^{\nu} - \epsilon_{\mathrm{V}i}^{\nu} q_{\mathrm{V}i}^{\mu}$, $\tilde{f}_{\mu\nu}^{(i)} = \frac{1}{2}\epsilon_{\mu\nu\rho\sigma} f^{(i),\rho\sigma}$, and $\epsilon_{\mathrm{V}i}$, $q_{\mathrm{V}i}$, and $m_{\mathrm{V}i}$ are the polarization vector, four-momentum, and pole mass of the i-th gauge boson, respectively. The constants Λ_1 and Λ_Q are the scales of BSM physics. The a_i and κ_i terms correspond to the strengths of vector boson couplings.

In the above equation, the only SM tree-level contribution is given by $a_1^{\mathrm{ZZ}} \neq 0$ and $a_1^{\mathrm{WW}} \neq 0$. The rest of the ZZ and WW couplings are considered as anomalous contributions, which are either small contributions arising in the SM due to loop corrections or new BSM contributions. Loop induced SM processes contribute effectively via a_2^{VV}, and SM processes at three-loop level contribute via a_3^{VV}. However, these tiny SM contributions are not yet distinguishable experimentally from zero with the available sensitivity, and they will be considered as zero in the SM.

Among the various couplings, the a_3 stands out as the only CP-odd term, making it sensitive to possible BSM effects that would result in CP violation. The a_2 term corresponds to a CP-even contribution to the HVV vertex and is sensitive to BSM contributions from heavy Higgs bosons. The $\kappa_{1,2}/(\Lambda_1)^2$ and $\kappa_3/(\Lambda_Q)^2$ terms are sensitive to possible physics at new energy scales represented by the denominator. The $\kappa_3/(\Lambda_Q)^2$ coupling allows for scenarios that violate the gauge symmetries of the SM and is not considered in this thesis. Symmetries and gauge invariance force $\kappa_1^{ZZ} = \kappa_2^{ZZ}$, leading to the single coupling $\kappa_1^{ZZ}/(\Lambda_1^{ZZ})^2$, which will be denoted simply as Λ_1 in what follows. Gauge invariance imposes $\kappa_1^{\mathrm{Z}\gamma} = 0$, making it impossible to measure the $\Lambda_1^{\mathrm{Z}\gamma}$ coupling in any process involving an on-shell photon. However, the four-lepton channel contains events featuring an off-shell photon, i.e. $\mathrm{H} \to \mathrm{Z}\gamma^* \to 4\ell$, that can be exploited to study the $\Lambda_1^{Z\gamma}$ coupling.

After all these considerations, four couplings parametrising the HZZ vertex are left to investigate: a_3, a_2, Λ_1, and $\Lambda_1^{\mathrm{Z}\gamma}$. Before introducing the six discriminants that will be used to probe the tensor structure of the HZZ vertex, it is necessary to introduce a common notation to refer to the coupling values. Following the conventions adopted in Ref. [128], rather than using the value of the coupling itself to identify the anomalous coupling scenario, the cross section fractions f_{ai} are used:

$$f_{ai} = \frac{|a_i|^2 \sigma_i}{\sum_j |a_j|^2 \sigma_j} \cdot \mathrm{sign}\left(\frac{a_i}{a_1}\right), \quad (4.22)$$

where σ_i is the cross section for the process corresponding to $a_i = 1, a_{j\neq i} = 0$ in Eq. (4.21). The term for Λ_1 is $\tilde{\sigma}_{\Lambda 1}/(\Lambda_1)^4$ instead of $|a_i|^2\sigma_i$, where $\tilde{\sigma}_{\Lambda 1}$ is the effective cross section

Table 4.3: Matrix element kinematic discriminants considered in the analysis. Some discriminants have a special label to identify the targeted Higgs boson property rather than the name of the coupling. $\mathcal{D}_{0^-}^{\text{dec}}$ is sensitive to a CP-odd Higgs boson, $\mathcal{D}_{\text{CP}}^{\text{dec}}$ is the observable sensitive to the CP-mixing, and $\mathcal{D}_{0h+}^{\text{dec}}$ is sensitive to heavy CP-even Higgs boson.

	$\mathcal{D}_{\text{alt}}$				$\mathcal{D}_{\text{int}}$	
	Coupling					
	a_3	a_2	Λ_1	$\Lambda_1^{Z\gamma}$	a_3	a_2
Discriminant	$\mathcal{D}_{0^-}^{\text{dec}}$	$\mathcal{D}_{0h+}^{\text{dec}}$	$\mathcal{D}_{\Lambda 1}^{\text{dec}}$	$\mathcal{D}_{\Lambda 1}^{Z\gamma,\text{dec}}$	$\mathcal{D}_{\text{CP}}^{\text{dec}}$	$\mathcal{D}_{\text{int}}^{\text{dec}}$

for the process corresponding to $\Lambda_1 = 1$ TeV, given in units of fb·TeV^4. The cross section fractions reflect kinematic features in a direct way, they are invariant with respect to the coupling conventions, and they are naturally bounded between -1 and $+1$.

The coupling a_3, being the only CP-odd contribution to the scattering amplitude, is probably the most interesting coupling to investigate. When defining matrix-element discriminants, the two hypothesis should be well-established. For all discriminants employed in this analysis, one of the hypothesis is always the SM scenario $f_{a1} = 1$. In this specific case, the alternative hypothesis is the one implying sizeable contributions to the a_3 coupling resulting in $f_{a3} \neq 0$. The cross section is then proportional to the terms reported in Eq. 4.20, where the squared amplitudes are CP-even, since they are squared, and the interference term is CP-odd. Consequently, the related alternative discriminant, denoted as $\mathcal{D}_{0^-}^{\text{dec}}$, is also CP-even. The $\mathcal{D}_{0^-}^{\text{dec}}$ discriminant is the optimal variable to separate the scalar and pseudoscalar amplitude, providing an insight in their relative strength. On the other hand, the interference discriminant is CP-odd and incorporates the full information about the CP violation in the Higgs boson decay. The distribution is symmetric in absence of CP violation, but it exhibits forward-backward asymmetry in case of $f_{a3} \neq 0$, indicating sizeable CP-violating terms in the HZZ vertex. This explains why this discriminant is referred to as $\mathcal{D}_{\text{CP}}^{\text{dec}}$. In principle, all this information is also in the angular variables introduced above. However, in those variables the information is hidden in the multidimensional space, instead here it is contained in just two observables.

The same identical approach can be used to probe the a_2 coupling by defining the corresponding alternative discriminant $\mathcal{D}_{0h+}^{\text{dec}}$ and interference discriminant $\mathcal{D}_{\text{int}}^{\text{dec}}$. For the Λ_1 and $\Lambda_1^{Z\gamma}$ couplings, the approach is the same, but the related interference discriminants are not used. They are less appealing for theoretical reinterpretations and found to provide little additional information due to the large correlation with the alternative discriminants, which can be used effectively to probe also the intermediate scenarios. Table 4.3 details the set of kinematic discriminants considered and the couplings to which they are sensitive. The index "dec" indicates that only decay information is used to build these discriminants. The distributions for $\mathcal{D}_{0^-}^{\text{dec}}$, $\mathcal{D}_{\text{CP}}^{\text{dec}}$, $\mathcal{D}_{0h+}^{\text{dec}}$, and $\mathcal{D}_{\Lambda 1}^{\text{dec}}$ are shown in Fig. 4.11.

4.6.3 Double-differential observables

Due to large dataset accumulated during Run 2 of the LHC, it is now feasible to measure fiducial cross sections in bins defined by pairs of observables. This analysis measures the cross section in bins of m_{Z_1} vs. m_{Z_2}, $|y_H|$ vs. p_T^H, number of associated jets vs. p_T^H, p_T of the leading vs. p_T of the subleading jet, p_T^{Hj} vs. p_T^H, and $\mathcal{T}_C^{\text{max}}$ vs. p_T^H. This approach permits to probe with higher-granularity specific regions of the phase space. As an example, measuring the p_T^H spectrum in bins of N_{jets} offers the opportunity to examine the transverse

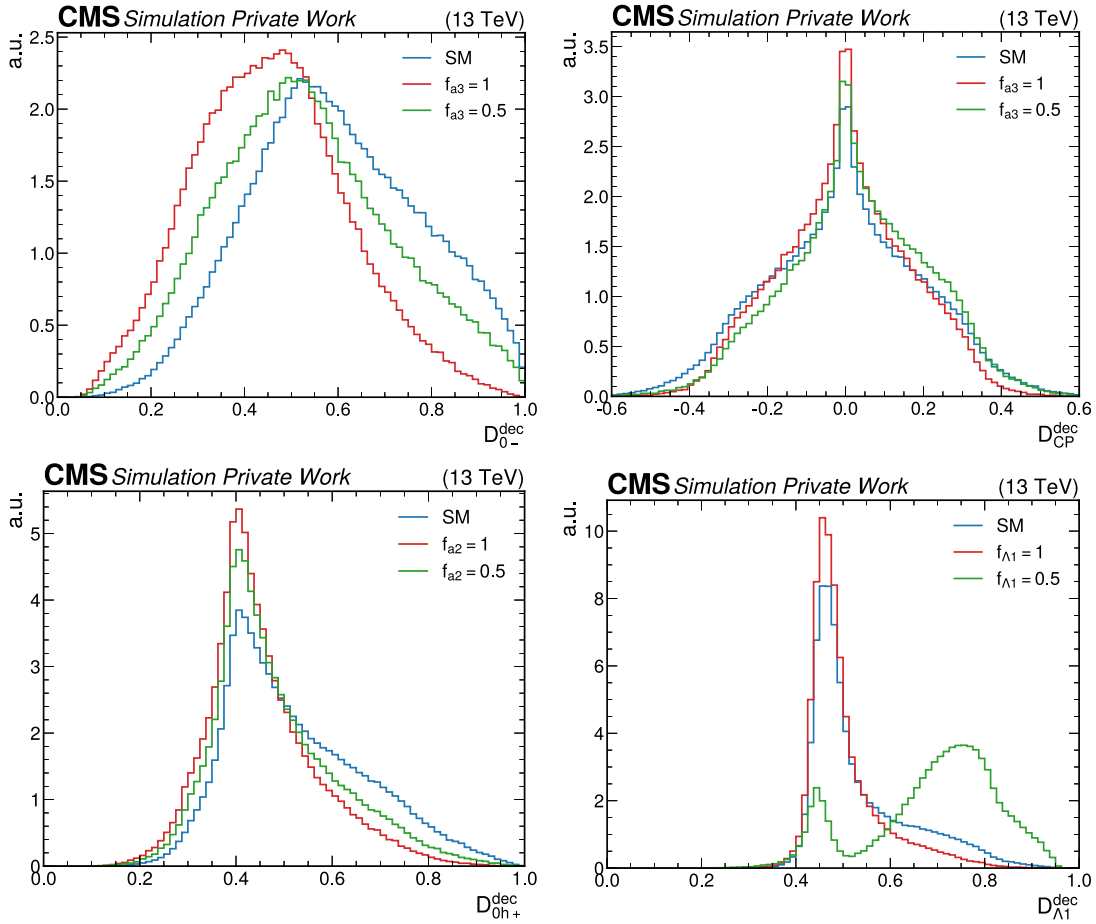

Figure 4.11: Distributions for $\mathcal{D}^{\text{dec}}_{0^-}$ (top left), $\mathcal{D}^{\text{dec}}_{\text{CP}}$ (top right), $\mathcal{D}^{\text{dec}}_{0\text{h}+}$ (bottom left), and $\mathcal{D}^{\text{dec}}_{\Lambda 1}$ (bottom right). The shape is extracted in the mass range $105 < m_{4\ell} < 160$ GeV. All distributions are normalised to unity. Each plot compare the SM distribution with the corresponding $f_{ai} = 1$ and $f_{ai} = 0.5$ distributions.

momentum distribution with varying relative contributions from the production modes. In cases without jets, the p_{T}^{H} for Higgs boson produced through ggH is isolated, providing the chance to amplify potential effects in the gluon loop. The purity of this bin is very high, the ggH contribution accounts for 98% of events. As the number of jets increases, the ggH contribution decreases, thereby highlighting the features of the p_{T}^{H} spectrum associated with the sub-leading production mechanisms. In the large jet multiplicity case, $\text{N}_{\text{jets}} > 5$, the ggH contribution is suppressed to less than 30%, while the ttH contribution dominates with 65% of events. Unfortunately, the $\text{H} \to \text{ZZ} \to 4\ell$ channel does not have yet enough statistics to access this interesting regime.

4.6.4 Binning optimisation

Once the differential observables are defined, the aim of fiducial analyses is to measure cross sections in bins of such observables. The choice of the bin boundaries and the number of bins are two important aspects to optimise. In comparison to the Run 1 analysis [179], this analysis can benefit of the larger dataset accumulated during Run 2 to increase the granularity of the results. The binning optimisation is based on ensuring a high enough experimental sensitivity to the SM Higgs boson in each distinct bin. This is usually done by computing the p-value of the background-only hypothesis, which express how likely is to make a discovery if the signal is present. However, in the context of this analysis,

this quantity is not used for trying to rediscover the Higgs boson in each bin, it is rather used to ensure a high signal-to-background ratio, leading to a more stable fit and smaller uncertainty on the final cross sections.

The p-value is usually related to the *significance* Z, defined as an equivalent number of standard deviations of a Gaussian distributed random variable:

$$Z = \Phi^{-1}(1-p) \tag{4.23}$$

where Φ^{-1} is the standard Gaussian quantile. The discovery threshold of $p \leq 2.9 \times 10^{-7}$ corresponds to $Z \geq 5$. Both p and Z, being a function of data, are random variables characterised by a certain distribution, and they also depend on the assumed value of the signal s. The expected significance is evaluated with the median of these distributions assuming a fixed value of s. The choice of the median, rather than the mean, ensures the validity of Eq. 4.23. If the expected number of background events b has an uncertainty σ_b, the most general significance can be written using the Asimov Median Significance (AMS) [185]:

$$\text{AMS} = \sqrt{2(s+b)\log\left(\frac{(s+b)(b+\sigma_b^2)}{b^2+(s+b)\sigma_b^2)}\right) - \left(\frac{b}{\sigma_b}\right)^2 \log\left(1+\frac{s\sigma_b^2}{b(b+\sigma_b^2)}\right)} \tag{4.24}$$

where:

$$\sigma_b = \sqrt{\frac{(\sigma_{\text{Z+X}}/\text{Z}+\text{X})^2 + \sum_i (\sigma_{\text{bkg}_i}/\text{bkg}_i)^2}{\text{Z}+\text{X}^2+\sum_i \text{bkg}_i^2}} \tag{4.25}$$

and i=$\text{q}\overline{\text{q}} \rightarrow \text{ZZ}$, $\text{gg} \rightarrow \text{ZZ}$. At first order, Eq. 4.24 corresponds to the widely used discovery significance typical of high-energy physics experiments:

$$\text{AMS} = \frac{s}{\sqrt{b+\sigma_b^2}}(1+\mathcal{O}(s/b)+\mathcal{O}(\sigma_b^2/b)) \tag{4.26}$$

which is valid only if $s \ll b$ and $\sigma_b^2 \ll b$.

The analysis, whenever possible, follows the bin boundaries recommendations provided by the LHCHWG, in order to facilitate the comparison and combination between different Higgs boson decay channels. When this is not possible, either because such recommendations are not provided or because the observable is specific of the decay side, bin boundaries are optimized by maximising the quantity defined in Eq. 4.24 in each kinematic bin. A desirable target of AMS $\sim$ 3.5 is used as a threshold value in the binning optimisation procedure. This value is chosen as it is the one obtained in the p_{T}^{H} distribution implementing the central recommended binning. However, this value is not always attainable, especially in the 1-jet and 2-jet phase space. In such scenarios, a lower threshold of AMS $\sim$ 2 is used. The resultant bin boundaries are reported in Tab. 4.4.

The binning optimisation for 2D observables adds a further degree of arbitrariness. 2D bin boundaries can be either defined as mutually exclusive phase space regions for the two observables or as multiple bins of the first observable in bins of the second one. The former approach is applied to m_{Z_1} vs. m_{Z_2}, p_{T}^{j1} vs. p_{T}^{j2}, and p_{T}^{H} vs. $p_{\text{T}}^{\text{H}j}$, the latter approach is applied to $\mathcal{T}_{\text{C}}^{\text{max}}$ vs. p_{T}^{H}, $|y_{\text{H}}|$ vs. p_{T}^{H}, and N_{jets} vs. p_{T}^{H}. An example per type is reported in Fig. 4.12 and the bin boundaries are listed in Tab. 4.5.

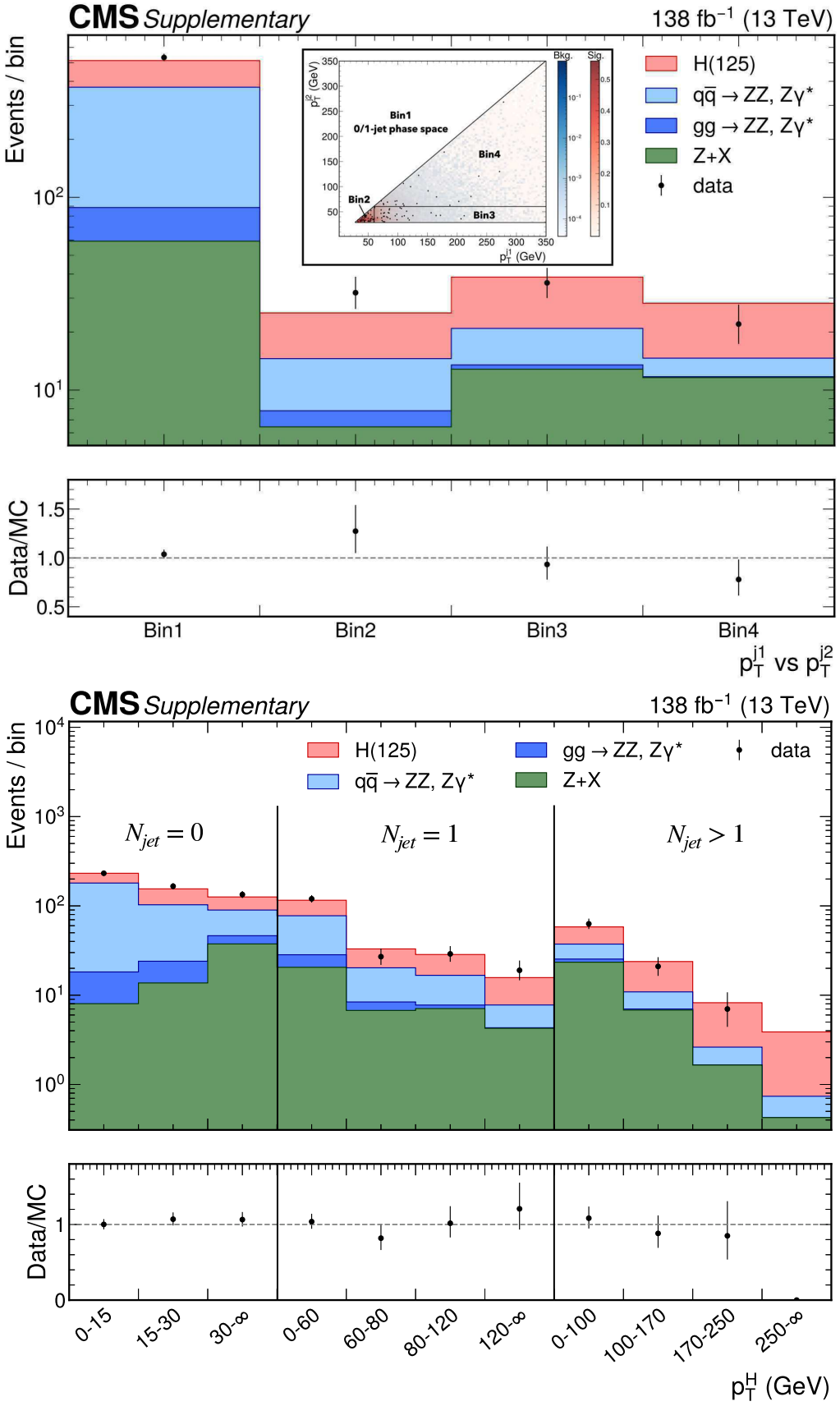

Figure 4.12: Distribution of the double-differential variable p_{T}^{j1} vs. p_{T}^{j2} (top) and N_{jets} vs. $p_{\mathrm{T}}^{\mathrm{H}}$ (bottom). Points with error bars represent the data, solid histograms the predictions from simulation. The bottom panel shows the ratio of the measured values to the expectations from simulation.

Table 4.4: One-dimensional observables with their corresponding bin boundaries.

Observable	Definition	Bin boundaries	Target
$m_{4\ell}$	Invariant mass of the 4ℓ system	[105,160] GeV	Inclusive
$p_{\mathrm{T}}^{\mathrm{H}}$	Transverse momentum of the 4ℓ system	[0,10,20,30,45,60,80,120,200,∞[GeV	Production
$\lvert y_{\mathrm{H}}\rvert$	Rapidity of the 4ℓ system	[0,0.15,0.3,0.45,0.6,0.75,0.9,1.2,1.6,2.5]	Production
$\cos\theta^*$	Cosine of the decay angle of the leading lepton pair in the 4ℓ rest frame	[-1.0,-0.75,-0.50,-0.25,0.0,0.25,0.50,0.75,1.0]	Decay
$\cos\theta_1$, $\cos\theta_2$	Cosine of the production angle, relative to the Z vector, of the antileptons from the two Z bosons	[-1.0,-0.75,-0.50,-0.25,0.0,0.25,0.50,0.75,1.0]	Decay
Φ, Φ_1	Azimuthal angles between the decay planes	[$-\pi$, $-3\pi/4$, $-\pi/2$, $-\pi/4$, 0, $\pi/4$, $\pi/2$, $3\pi/4$, π]	Decay
m_{Z_1}	Invariant mass of the two leading leptons	[40,65,75,85,92,120] GeV	Decay
m_{Z_2}	Invariant mass of the two subleading leptons	[12,20,25,28,32,40,50,65] GeV	Decay
$p_{\mathrm{T}}^{\mathrm{j}_1}$	Transverse momentum of the leading jet	[0-jet,30,55,95,200,∞[GeV	Production
$p_{\mathrm{T}}^{\mathrm{j}_2}$	Transverse momentum of the subleading jet	[0/1-jet,30,40,65,90,∞[GeV	Production
N_{jets}	Number of associated jets in the event	=0,=1,=2,=3,$\geq$4	Event level
$\mathcal{T}_{\mathrm{C}}^{\mathrm{max}}$	Rapidity-weighted jet veto	[0-jet|$\mathcal{T}_{\mathrm{C}}^{\mathrm{max}}$,15,20,30,50,80,$\infty$[GeV	Production
$\mathcal{T}_{\mathrm{B}}^{\mathrm{max}}$	Rapidity-weighted jet veto	[0-jet|$\mathcal{T}_{\mathrm{B}}^{\mathrm{max}}$,30,45,75,150,$\infty$[GeV	Production
m_{jj}	Invariant mass of the leading and subleading jets system	[0/1-jet,0,120,300,∞[GeV	Production
$\lvert\Delta\eta_{\mathrm{jj}}\rvert$	Difference in pseudorapidities of the leading and subleading jets	[0/1-jet,0.0,1.6,3.0,10.0]	Production
$\Delta\phi_{\mathrm{jj}}$	Azimuthal angle difference between the leading and subleading jets	[0/1-jet,$-\pi$, $-\pi/2$, 0, $\pi/2$, π]	Production
$p_{\mathrm{T}}^{\mathrm{Hj}}$	Transverse momentum of the 4ℓ and leading jet system	[0-jet,0,30,50,110,∞[GeV	Production
m_{Hj}	Invariant mass of the 4ℓ and leading jet system	[0-jet,110,180,220,300,400,600,∞[GeV	Production
$p_{\mathrm{T}}^{\mathrm{Hjj}}$	Transverse momentum of the 4ℓ, leading and subleading jets system	[0/1-jet,0,20,60,∞[GeV	Production
$\mathcal{D}_{0^-}^{\mathrm{dec}}$	Matrix element discriminant targeting a_3 coupling	[0.0,0.4,0.5,0.6,0.7,0.8,0.9,1.0]	Decay
$\mathcal{D}_{0\mathrm{h}+}^{\mathrm{dec}}$	Matrix element discriminant targeting a_2 coupling	[0.0,0.35,0.4,0.45,0.55,0.65,0.75,1.0]	Decay
$\mathcal{D}_{\Lambda 1}^{\mathrm{dec}}$	Matrix element discriminant targeting k_1 coupling	[0.0,0.45,0.5,0.6,0.7,1.0]	Decay
$\mathcal{D}_{\Lambda 1}^{\mathrm{Z}\gamma,\mathrm{dec}}$	Matrix element discriminant targeting $k_2^{Z\gamma}$ coupling	[0.0,0.35,0.45,0.5,0.55,0.65,1.0]	Decay
$\mathcal{D}_{\mathrm{CP}}^{\mathrm{dec}}$	Interference matrix element discriminant targeting a_3 coupling	[-0.75,-0.25,-0.1,0.0,0.1,0.25,0.75]	Decay
$\mathcal{D}_{\mathrm{int}}^{\mathrm{dec}}$	Interference matrix element discriminant targeting a_2 coupling	[0.0,0.7,0.8,0.9,0.95,1.0]	Decay

Table 4.5: Double-differential observables with their corresponding bin boundaries.

Observable	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6	Bin 7	Bin 8	Bin 9	Bin 10	Bin 11	Bin 12
$m_{\mathrm{Z_1}}$(GeV)	[40,85]	[40,70]	[70,120]	[85,120]	[85,120]	[85,120]						
$m_{\mathrm{Z_2}}$(GeV)	[12,35]	[35,65]	[35,65]	[30,35]	[24,30]	[12,24]						
$\lvert y_{\mathrm{H}}\rvert$	[0,0.5]	[0,0.5]	[0,0.5]	[0,0.5]	[0.5,1.0]	[0.5,1.0]	[0.5,1.0]	[1.0,2.5]	[1.0,2.5]	[1.0,2.5]		
$p_{\mathrm{T}}^{\mathrm{H}}$(GeV)	[0,40]	[40,80]	[80,150]	[150,∞[	[0,45]	[45,120]	[120,∞[	[0,45]	[45,120]	[120,∞[		
N_{jets}	0	0	0	1	1	1	1	>= 2	>= 2	>= 2	>= 2	
$p_{\mathrm{T}}^{\mathrm{H}}$(GeV)	[0,15]	[15,30]	[30,∞[	[0,60]	[60,80]	[80,120]	[120,∞[	[0,100]	[100,170]	[170,250]	[250,∞[	
$p_{\mathrm{T}}^{\mathrm{j_1}}$(GeV)	$N_{jets} < 2$	[30,60]	[60,350]	[60,350]								
$p_{\mathrm{T}}^{\mathrm{j_2}}$(GeV)		[30,60]	[30,60]	[60,350]								
$p_{\mathrm{T}}^{\mathrm{Hj}}$(GeV)	$N_{jets} < 1$	[0,30]	[0,45]	[30,350]	[45,350]							
$p_{\mathrm{T}}^{\mathrm{H}}$(GeV)		[0,85]	[85,350]	[0,85]	[85,350]							
$\mathcal{T}_{\mathrm{C}}^{\mathrm{max}}$(GeV)	0-jet$\lvert\mathcal{T}_{\mathrm{C}}^{\mathrm{max}}$	0-jet$\lvert\mathcal{T}_{\mathrm{C}}^{\mathrm{max}}$	0-jet$\lvert\mathcal{T}_{\mathrm{C}}^{\mathrm{max}}$	0-jet$\lvert\mathcal{T}_{\mathrm{C}}^{\mathrm{max}}$	0-jet$\lvert\mathcal{T}_{\mathrm{C}}^{\mathrm{max}}$	0-jet$\lvert\mathcal{T}_{\mathrm{C}}^{\mathrm{max}}$	[15,25]	[15,25]	[25,40]	[25,40]	[40,∞[	[40,∞[
$p_{\mathrm{T}}^{\mathrm{H}}$(GeV)	[0,15]	[15,30]	[30,45]	[45,70]	[70,120]	[120,∞[	[0,120]	[120,∞[	[0,120]	[120,∞[	[0,200]	[200,∞[

4.7 Physics model

As explained in Sec. 4.4, this analysis relies on a maximum likelihood unfolding. One of the fundamental steps of this type of unfolding is to implement a proper parametrisation of the observed events ($\mathcal{N}_{\text{obs}}$) enabling to unfold data and measure fiducial cross sections. This parametrisation is then embedded in the likelihood function for the maximum likelihood fit, as explained in Sec. 4.4.

Following the models for signal and background contributions described in Sec. 4.2 and Sec. 4.3, the number of events in each final state f, in each kinematic bin i of a given observable, and year of data taking y, is parametrised as a function of $m_{4\ell}$ as:

$$\begin{aligned}\mathcal{N}_{\text{obs}}^{f,i,y}(m_{4\ell}) &= \mathcal{N}_{\text{fid}}^{f,i,y}(m_{4\ell}) + \mathcal{N}_{\text{nonfid}}^{f,i,y}(m_{4\ell}) + \mathcal{N}_{\text{nonres}}^{f,i,y}(m_{4\ell}) + \mathcal{N}_{\text{bkg}}^{f,i,y}(m_{4\ell}) \\ &= \sum_j \epsilon_{i,j}^{f,y} \cdot \left(1 + f_{\text{nonfid}}^{f,i,y}\right) \cdot \sigma_{\text{fid}}^{f,j,y} \cdot L_{\text{int}}^{y} \cdot \mathcal{P}_{\text{res}}^{f,y}(m_{4\ell}) \\ &\quad + N_{\text{nonres}}^{f,i,y} \cdot \mathcal{P}_{\text{nonres}}^{f,y}(m_{4\ell}) + N_{\text{bkg}}^{f,i,y} \cdot \mathcal{P}_{\text{bkg}}^{f,i,y}(m_{4\ell})\end{aligned} \tag{4.27}$$

The various contributions are explained below.

$\mathcal{N}_{\text{fid}}^{f,i,y}(m_{4\ell})$: resonant fiducial component This contribution represents resonant events, i.e., the four leptons are associated to the decay of the Higgs boson, originating within the fiducial phase space. The shape of the resonant signal contribution, $\mathcal{P}_{\text{res}}(m_{4\ell})$, is described by a double sided Crystal Ball function, as described in Sec. 4.2. The term entering in Eq. 4.27 is given by:

$$\mathcal{N}_{\text{fid}}^{f,i,y}(m_{4\ell}) = \sum_j \epsilon_{i,j}^{f,y} \cdot \sigma_{\text{fid}}^{f,j,y} \cdot L_{\text{int}}^{y} \cdot \mathcal{P}_{\text{res}}^{f,y}(m_{4\ell}) \tag{4.28}$$

where σ_{fid} is the fiducial cross section, L_{int} is the integrated luminosity, and $\varepsilon_{i,j}^{f,y}$ is the element of the response matrix (cfr Sec. 1.2.2). The response matrix provides the probability for an event in the j-th bin at fiducial level to be reconstructed in the i-th bin at reco-level. The elements of this matrix are computed using the true information in MC samples as:

$$\varepsilon_{i,j} = \frac{N(\text{fid}_j \rightarrow \text{reco}_i)}{N(\text{fid}_j)} \tag{4.29}$$

The numerator corresponds to the number of events generated in the j-th bin at fiducial level and reconstructed in the i-th bin at reco-level; the denominator corresponds to the total number of events generated in the j-th bin at fiducial level, including events that are not reconstructed in any of the reco-level bins.

$\mathcal{N}_{\text{nonfid}}^{f,i,y}(m_{4\ell})$: resonant non-fiducial component An additional resonant contribution arises from events that are selected at reco-level but do not originate from the fiducial phase space. These events are due to detector effects which cause differences between the quantities used at fiducial level and the analogous quantities at reco-level. The $m_{4\ell}$ shape distribution of these events is verified using simulations to be identical to the shape of the fiducial signal. As the pdf is the same, the normalisation of this contribution is fixed to be a fraction of the fiducial resonant component. The value of this fraction, which is denote by $f_{\text{nonfid}}^{f,i}$, is determined from simulations as:

$$f_{\text{nonfid}}^{i} = \frac{N(\text{nonfid}_j \rightarrow \text{reco}_i)}{N(\text{fid}_j \rightarrow \text{reco}_i)} \tag{4.30}$$

The f_{nonfid}^{i} fraction describes the ratio of the non-fiducial and fiducial signal contribution in i-th bin at the reconstruction level.

$\mathcal{N}^{f,i,y}_{\text{nonres}}(m_{4\ell})$: non-resonant signal-induced background Non-resonant events, also known as combinatorial events, arise when the Higgs boson is produced in association with either a vector boson or a pair of top quarks, i.e., through the WH, ZH, and ttH production mechanisms. In these cases, the event selection algorithm may select 4ℓ candidates where at least one reconstructed lepton is not from the decay of the Higgs boson. The contribution of such events from WH, ZH, and ttH, in the mass range $m_{4\ell} \in [105, 160]$ GeV, is approximately 5%, 22%, and 17%, respectively. These events are treated as a signal-induced background and the pdf $\mathcal{P}_{\text{nonres}}(m_{4\ell})$ is empirically modelled using a Landau distribution. The normalisation of the distribution is extracted from MC simulations, whereas the parameters of the Landau are extracted with the procedure outlined in Sec. 4.2.

$\mathcal{N}^{f,i,y}_{\text{bkg}}(m_{4\ell})$: reducible and irreducible backgrounds The last term of Eq. 4.27 refers to the contribution from the irreducible and reducible backgrounds. Details about the normalisation and the shape of this contributions are given in Sec. 4.3. The corresponding templates $\mathcal{P}_{\text{bkg}}(m_{4\ell})$ are shown in Fig. 4.7.

An important coefficient, which does not enter directly in the parametrisation but it will be used thoroughly this chapter and the next one, is the *fiducial acceptance* $\mathcal{A}_{\text{fid}}$. It is evaluated with MC samples, and it is defined as the ratio between the number of events passing the fiducial selections and the total number of generated events. It is implicitly included within σ_{fid}, as it is given by the product of the cross section, branching ratio, and fiducial acceptance.

Extrapolation and combination of the coefficients

All the coefficients mentioned above, namely $\mathcal{A}_{\text{fid}}$, $\varepsilon_{i,j}$, and f^{i}_{nonfid}, are computed using simulations. As the mass of the Higgs boson will be fixed in the likelihood fit at the best-known value (at the time of the analysis) of $m_{\text{H}} = 125.38$ GeV [186], all the inputs of the likelihood should be evaluated at this value for consistency. Central MC samples provided by the CMS collaboration are generated at integer values of the Higgs boson mass, and the point $m_{\text{H}} = 125.38$ GeV is not directly generated. Consequently, the coefficients are computed at $m_{\text{H}} = 124$ GeV, $m_{\text{H}} = 125$ GeV, and $m_{\text{H}} = 126$ GeV and extrapolated at $m_{\text{H}} = 125.38$ GeV using a spline function. It was investigated the possibility of using more mass points and a linear fit instead of a spline, but it turned out that the difference of the extrapolated value was minimal and without effect on the final result, at the price of a much higher computation time. An example of the extrapolation procedure is reported in Fig. 4.13. For the inclusive measurement, i.e., one single bin, these arrays of coefficients are actually single numbers. The combined numbers on the three final states are reported in Tab. 4.6.

In addition, the coefficients are evaluated independently for each production mode. However, the measurement carried out in this analysis is inclusive in production modes, therefore in the likelihood function the inclusive numbers, i.e., the SM numbers, should be included. The combined value is obtained by a weighed mean, where the weights are given by the relative fiducial cross sections. For instance, the elements of the response matrices are combined as:

$$\varepsilon^{\text{SM}}_{i,j} = \sum_{p}^{\text{Prod. modes}} \frac{\sigma^{\text{SM},p}_{\text{fid},j}}{\sum_p \sigma^{\text{SM},p}_{\text{fid},j}} \cdot \epsilon^{p}_{i,j} \tag{4.31}$$

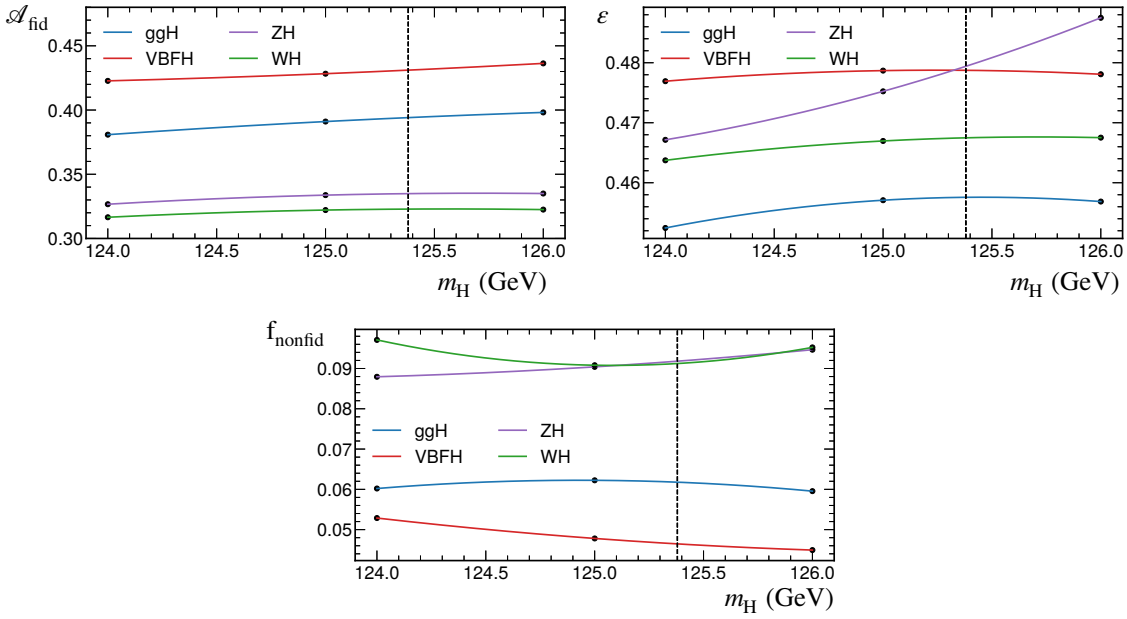

Figure 4.13: Extrapolation of the fraction of signal events within the fiducial phase space ($\mathcal{A}_{\text{fid}}$), the reconstruction efficiency (ϵ) in the fiducial phase space, and the ratio of the number of reconstructed events outside the fiducial phase space to that of the ones inside the fiducial phase space (f_{nonfid}) at 125.38 GeV. The black dots represent the computation of the coefficients with the available MC samples. The coloured lines show the spline function used to obtain the extrapolated value at 125.38 GeV represented with a black dotted vertical line. These numbers are related to the $4e$ final state for the 2018 data-taking period.

Table 4.6: Summary of the inputs to the maximum likelihood based unfolding. The fraction of signal events within the fiducial phase space (acceptance $\mathcal{A}_{\text{fid}}$), the reconstruction efficiency (ϵ) in the fiducial phase space, and the ratio of the number of reconstructed events outside the fiducial phase space to that of the ones inside the fiducial phase space (f_{nonfid}) are quoted for each production mechanism for $m_{\text{H}} = 125.38$ GeV. The last column shows the value of $(1 + f_{\text{nonfid}})\epsilon$, which regulates the signal yield for a given fiducial cross section. All values are shown with their statistical uncertainty.

Signal process	$\mathcal{A}_{\text{fid}}$	ϵ	f_{nonfid}	$(1 + f_{\text{nonfid}})\epsilon$
ggH	0.408 ± 0.001	0.619 ± 0.001	0.053 ± 0.001	0.652 ± 0.001
VBFH	0.448 ± 0.001	0.632 ± 0.002	0.043 ± 0.001	0.659 ± 0.002
WH	0.332 ± 0.001	0.616 ± 0.002	0.077 ± 0.001	0.664 ± 0.002
ZH	0.344 ± 0.002	0.626 ± 0.003	0.083 ± 0.002	0.678 ± 0.003
ttH	0.320 ± 0.002	0.614 ± 0.003	0.179 ± 0.003	0.725 ± 0.005

4.7.1 Do we need regularisation?

The inspection of the response matrices is an important step when performing an unfolding. As explained in Sec. 1.2.2, a response matrix with large off-diagonal elements is an ill-posed matrix for the unfolding problem. This implies that the unfolded result will be very sensitive to the Poisson fluctuations of data, leading to unphysical oscillation that should be tamed by implementing some regularisation term. The usual method to assess whether a matrix is well-posed is to compute the condition number. The condition numbers of all response matrices of this analysis are found to be smaller than 3, far from the

conventional threshold of 10, where a regularisation should be taken into account. Figure 4.14 reports the response matrices for four representative cases: $|y_{\mathrm{H}}|$, $m_{\mathrm{Z_2}}$, p_{T}^{j1}, and m_{Hj}. In general, the response matrix for production and decay observables are diagonal with very small off-diagonal elements. Their condition number is usually around 1.5. This is given by the high momentum resolution of the CMS detector that prevents from large bin-by-bin migrations from fiducial- to reco-level. In particular, the diagonal elements of the rapidity matrix feature a decreasing trend, from 81% of efficiency in the first bin down to 67% in the last bin. This reflects the experimental environment, where at high rapidity, i.e., closer to the beam pipe, the environment is harsher and the reconstruction efficiency goes down. On the other hand, the jet resolution is not as high as that of leptons. The evidence of this effect can be found in the response matrices for jet-observables, which present larger off-diagonal elements. Fortunately, the bin-by-bin migration is any case limited, preventing the implementation of a regularisation procedure. The less diagonal matrix is represented by the m_{Hj} observable, where the condition number is equal to 2.82.

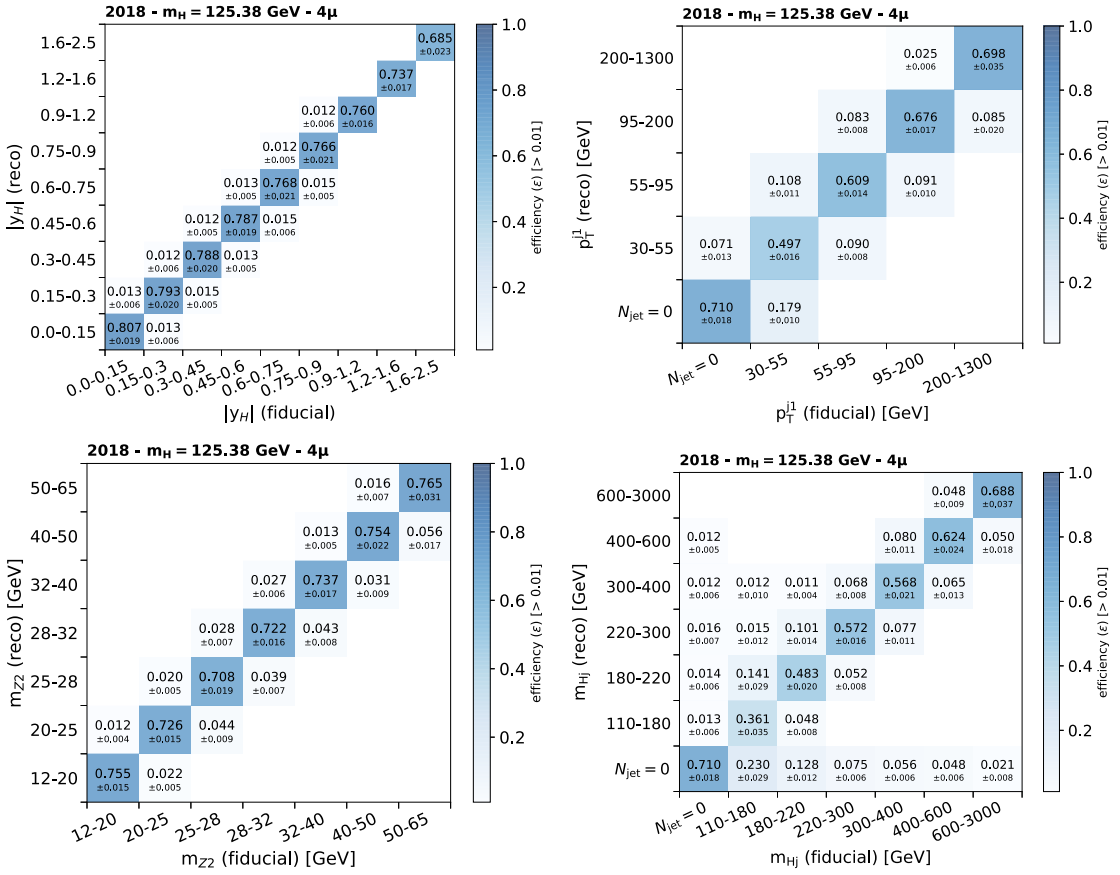

Figure 4.14: Response matrices for the rapidity of the Higgs boson (top left), the transverse momentum of the leading jet (top left), the invariant mass of the $\mathrm{Z_2}$ boson (bottom left), and the invariant mass of the H + leading jets system (bottom right). The matrices are related to the 2018 period, the 4μ final state, and for $m_{\mathrm{H}} = 125.38$ GeV. The entry of the matrix is reported only if it is larger than 0.01. All values are shown with their statistical uncertainty. The condition number of the top left matrix is 1.19, of the top right matrix is 1.21, of the bottom left matrix is 2.0, and of the bottom right matrix is 2.82. The m_{Hj} matrices is one of the less diagonal matrix in the entire set of this analysis.

4.7.2 Fiducial cross section

The parameter of interest of Eq. 4.27 is $\sigma_{\text{fid}}^{f,j}$, which denotes the fiducial cross section of the j^{th} bin in the final state f. Depending on the type of observable, the fiducial cross section is evaluated in three different ways.

$4e/4\mu/2e2\mu$ The first approach is to quote the cross section in the three different final states separately, and it is only implemented for the inclusive cross section. The parametrisation of Eq. 4.27 does not need to be modified.

$4e + 4\mu + 2e2\mu$ The second approach is to report the cross section inclusively across the three final states. This is the method that is used for all observables. In order to increase the model independence, the BRs of the Higgs boson in $2e2\mu$, $4e$, and 4μ are allowed to float in the fit procedure. Thus, the parameter representing the fiducial cross section should be redefined as:

$$\sigma_{\text{fid}}^{f,j} = \sigma_{\text{fid}}^{j} \cdot f(\text{K1}^j,\ \text{K2}^j,\ \text{frac(4e)}^j,\ \text{frac}(4\mu)^j) \tag{4.32}$$

where σ_{fid}^{j} is the new parameter of interest. The function f represents a particular combination of four elements, which are combined differently depending on the final state. The $\text{frac}(4\mu)^j$ and $\text{frac}(4e)^j$ terms are defined as the ratio of the fiducial cross section in the j^{th} bin for the 4μ or $4e$ channel, respectively, and the total fiducial cross section in the j^{th} bin. Their values are calculated theoretically and they are set as fixed parameters.

$$\text{frac(4e)}_j = \frac{\sigma(\text{4e})_j^{\text{fid}}}{\sigma_j^{\text{fid}}} \qquad \text{frac(4mu)}_j = \frac{\sigma(\text{4mu})_j^{\text{fid}}}{\sigma_j^{\text{fid}}} \tag{4.33}$$

Instead, K1^j and K2^j are two additional floating parameters in the range:

$$K1^j \rightarrow \left[0, \frac{1}{\text{frac(4e)}^j}\right] \qquad K2^j \rightarrow \left[0, \frac{1-\text{frac(4e)}^j}{\text{frac}(4\mu)^j}\right] \tag{4.34}$$

In order to let the branching ratios float and to account that the fractions of events in the three final states are correlated, $\sigma_{\text{fid}}^{f,j}$ is defined as follows:

$$\begin{aligned} \sigma_{\text{fid}}^{\text{4e},j} &= \sigma_{\text{fid}}^{j} \cdot \text{frac(4e)}^j \cdot K1^j \\ \sigma_{\text{fid}}^{4\mu,j} &= \sigma_{fid}^{j} \cdot \frac{[1-\text{frac(4e)}^j \cdot K1^j] \cdot K2^j \cdot \text{frac}(4\mu)^j}{1-\text{frac(4e)}^j} \\ \sigma_{\text{fid}}^{\text{2e2}\mu,j} &= \sigma_{fid}^{j} \cdot [1-\text{frac(4e)}^j \cdot K1^j] \cdot \left[1 - \frac{\text{frac}(4\mu)^j \cdot K2^j}{1-\text{frac(4e)}^j}\right] \end{aligned} \tag{4.35}$$

The SM scenario is retrieved when K1^j and K2^j are equal to 1.

$4e + 4\mu/2e2\mu$ The third approach is to report the fiducial cross section in the opposite-flavour $2e2\mu$ and same-flavour $4e$+4μ final states. This approach is only implemented for decay observables. The final states with same-flavour leptons have different physics than the final state with opposite-flavour leptons, when decay observables are considered. This is a consequence of the destructive interference between the two alternative ways to construct the $\text{H} \rightarrow \text{ZZ} \rightarrow 4f$ diagrams in the same-helicity cases [187]. The effect of this interference can be seen in Fig 4.15 for some decay observables. Additionally, data are subject to statistical fluctuations, especially after events are divided into several bins. As a result, not all K1^j and K2^j fractions would

be the same across all bins, and therefore, the final distribution would not be a simple linear combination of distributions from the $2e2\mu$, $4e$, and 4μ final states. Given that the distributions in the $2e2\mu$ final state are not expected to be the same as those in the $4e + 4\mu$ state, there would not be a clear way to separate statistical fluctuations in the $2e2\mu$:$4e$:4μ composition with real, physical differences. It was also investigated the possibility to provide a further result by splitting the $4e$ and 4μ, which is justified by the possibility of BSM physics affecting the two final states differently, but the low number of events in the $4e$ channel made this approach unfeasible for differential observables. Consequently, in order to perform the measurement of the cross section in the $4e + 4\mu$ and $2e2\mu$ final states, the definition of $\sigma_{\text{fid}}^{f,j}$ is modified as:

$$\sigma_{\text{fid}}^{4e} = \sigma_{\text{fid}}^{4e+4\mu} \cdot \text{frac}(4\text{e}) \qquad \sigma_{\text{fid}}^{4\mu} = \sigma_{\text{fid}}^{4e+4\mu} \cdot \text{frac}(4\mu) \tag{4.36}$$

The new parameters of interest become $\sigma_{\text{fid}}^{4e+4\mu}$ and $\sigma_{\text{fid}}^{2e2\mu}$ where:

$$\text{frac}(4\text{e}) = \frac{\sigma(4\text{e})_{\text{fid}}}{\sigma_{\text{fid}}^{4e+4\mu}} \qquad \text{frac}(4\mu) = \frac{\sigma(4\mu)_{\text{fid}}}{\sigma_{\text{fid}}^{4e+4\mu}} \tag{4.37}$$

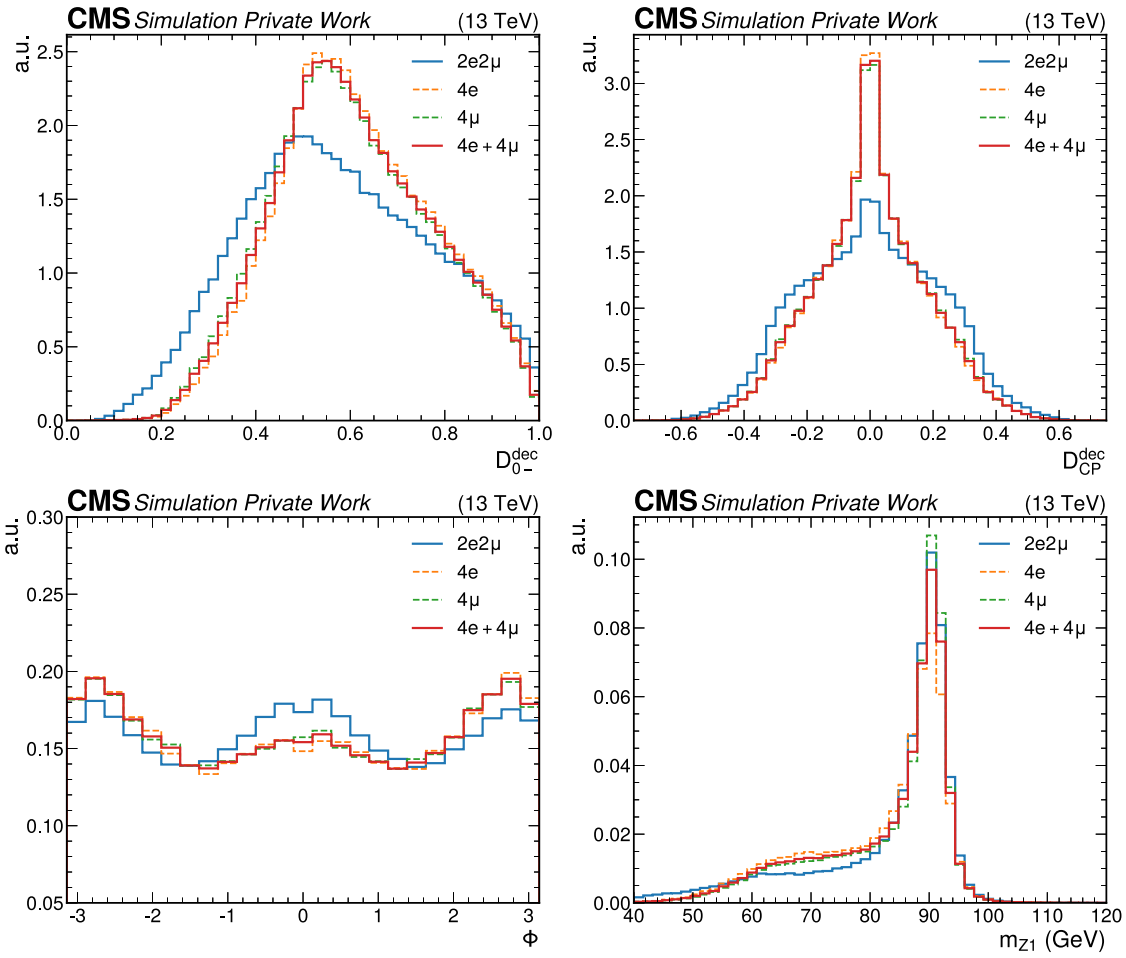

Figure 4.15: Distributions of four decay observables in the $2e2\mu$, $4e$, 4μ, and $4e + 4\mu$. All distributions are normalised to unity. All production modes (ggH, VBFH, WH, ZH, and ttH) are summed together as well as the three periods (2016, 2017, and 2018). The distributions are shown for events in the mass range $m_{4\ell} \in [105, 160]$ GeV.

Table 4.7: Summary of the experimental systematic uncertainties.

Common experimental uncertainties			
	2016	2017	2018
Luminosity uncorrelated	1%	2%	1.5%
Luminosity correlated 2016–2017–2018	0.6%	0.9%	2%
Luminosity correlated 2017–2018	–	0.6%	0.2%
Electron ID/reco efficiency	6.2–10.9%	4.3–8.5%	5.2–9.5%
Muon ID/reco efficiency	0.7–1.9%	0.6–1.9%	0.6–1.9%
Jet energy scale	0.1–27%	0.1–33%	0.1–33%
Background related uncertainties			
Reducible background (Z+X)	25–43%	23–36%	24–36%
Signal related uncertainties			
Electron energy scale	0.06%	0.06%	0.06%
Muon energy scale	0.01%	0.01%	0.01%
Electron energy resolution	10%	10%	10%
Muon energy resolution	3%	3%	3%

4.8 Systematic uncertainties

The evaluation of systematic uncertainties is a crucial step in every physics analysis. Although the four-lepton channel suffers from a low number of events, and in the past, precision was limited by the statistical uncertainty, the large integrated luminosity accumulated during the Run 2 of the LHC is expected to reduce the statistical component, especially in the context of inclusive measurements. As discussed in Ch. 3, this analysis employs the Ultra-Legacy (UL) calibration for physics objects. The nuisance parameters included in this analysis has been re-evaluated with the new reconstructed objects.

Systematic uncertainties in an analysis are typically divided into two categories: experimental and theoretical. Experimental uncertainties are related to effects originating from detectors imperfections and performance of the object reconstruction algorithm, while theoretical uncertainties relate to the precision of theoretical computations. In the context of maximum likelihood unfolding, as already discussed in Sec. 4.4, systematic uncertainties are introduced as nuisance parameters and are profiled in the fitting procedure. The initial constraints on these nuisance parameters correspond to the associated systematic uncertainties and, unless noted otherwise, are implemented as *log-normal* pdf:

$$p(\tilde{\theta}|\theta) = \frac{1}{\sqrt{2\pi}\ln(\kappa)} \exp\left(-\frac{\ln^2(\theta/\tilde{\theta})}{2\ln^2\kappa}\right)\frac{1}{\theta} \tag{4.38}$$

where the parameter κ controls the width of the distribution. The log-normal is the recommended distribution for all multiplicative corrections and implies a change in the normalisation of the corresponding process.

Table 4.7 gives an overview of the different sources of experimental systematic uncertainties considered in the analysis and they are detailed below.

Luminosity The integrated luminosities of the 2016, 2017, and 2018 data-taking periods are individually known with uncertainties in the range 1.2–2.5% [188–190], while the total Run 2 (2016–2018) integrated luminosity has an uncertainty of 1.6%. The partial correlation scheme across the three years is reported in Tab. 4.7. This uncertainty affects all the channels and processes, except for the reducible background as it is evaluated with a data-driven method.

Lepton efficiency This nuisance parameter includes the uncertainties related to the lepton trigger, reconstruction, and identification efficiency (cfr Sec. 3.2.4). It is estimated from data for the three different final states using the TnP method. For electron, it ranges 4.3–10.9 %, while for muons it ranges 0.6–1.9 %, depending on the final state and data-taking year. This uncertainty affects all the channels and processes, except for the reducible background since it is evaluated with a data-driven method. The big difference between the electron and muon values is due to the large uncertainty from the TnP method in the low-p_{T} region for electrons (7-20 GeV). Since the resonance that is used for the TnP for electron is the Z peak, the low-p_{T} region is sparsely populated, making it difficult to distinguish the signal from the large QCD background, which is particularly large in this region. For muons, this problem is solved by using a low-mass resonance, namely the J/ψ peak. This solution cannot be implemented for electrons due to the absence of specific triggers for J/$\psi \to ee$. One has to dig into the CMS data to find the J/ψ, and this is an ongoing effort within the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ analysis team. Despite still being large, the values for the electron efficiency benefit of the new root-mean-square method for the combination of the various sources of systematic uncertainties for the identification part of the efficiency (cfr Sec. 3.2.4). Compared to the previous values, these are $\sim$40% lower.

Lepton energy scale and resolution The systematic uncertainties on the lepton momentum scale and resolution are estimated from dedicated studies on the $Z \to \ell^+\ell^-$ mass distribution in data and simulation. The scale uncertainty is found to be 0.01% in the 4μ channel and 0.06% in the $4e$ channel, while the resolution uncertainty is 10% in the $4e$ channel and 3% in the 4μ channel. These uncertainties impact the signal line shape and they are not implemented as log-normal nuisances. The parameters of the double-sided Crystal ball function are instead to be assigned a Gaussian uncertainty corresponding to the value of the systematic uncertainty. The scale impacts on the mean of the function, while the resolution impacts on its width. The better data/MC agreement and the re-derived corrections for UL lowered these uncertainties of approximately 50% compared to the previous calibration.

Jet energy scale Systematic uncertainties on the jet energy scale are considered when dealing with measurements involving jet-related observables, as they can introduce migrations across bin boundaries. These uncertainties are considered to act on the normalisation of the processes and are modelled with a set of 11 nuisance parameters following the central CMS recommendations, taking into account partial correlations among the different sources of uncertainty that impact on the jet energy scale. They are computed by evaluating the difference in yield in each bin when considering the $+1\sigma$ and -1σ variation of the JES source.

Reducible background The estimation of the uncertainty on the reducible background is explained in Sec. 4.3. The overall effect of these three sources is found to be between 23 and 43%, depending on the final state.

The theoretical uncertainties are listed and explained below.

PDF, α_s, and QCD scales The choice of the PDF, the value of the strong coupling constant α_s, and the QCD scales for the MC generation affects the rates of the irreducible backgrounds. The QCD uncertainty due to the choice of the renormalization and factorization scales is determined by varying these scales between 0.5 and 2 times their nominal value, while keeping their ratio between 0.5 and 2, for an overall 4% effect. The uncertainty on the choice of the PDF set is determined

following the PDF4LHC recommendations by taking the root-mean-square of the variation of the results when using different replicas of the default NNPDF set [191]. The α_s uncertainty is evaluated following the PDF4LHC recommendations, which prescribe to vary the values similarly to the QCD scales. These theoretical uncertainties are considered also for the signal processes, even though in this case they do not enter the measurement as nuisance parameters but are taken into account only in the comparison of the results with the theoretical predictions (cfr Sec. 5). The QCD scale uncertainties range between 0.3% and 6% across the different production mechanisms, while the PDF uncertainties are found to be between 1.6% and 3.6%.

Branching ratio Similarly to the PDF and QCD scales for signal processes, the uncertainty on the $H \to 4\ell$ branching ratio is not included in the measurement as a nuisance parameter, but it is included as explained in Sec. 5. Its value is 2%.

k-factor The k-factor introduced in Sec. 4.3 to rescale the gg→ZZ to the best-known cross section has an associated uncertainty of 10%. In principle, also the k-factors for the $q\bar{q} \to ZZ$ should be considered, but their value of 0.1% was found to be completely negligible and therefore not considered in the analysis.

Chapter 5

Characterisation of the Higgs sector with fiducial cross sections in the four-lepton channel

This chapter marks the culmination of the analysis dedicated to the measurement of Higgs boson fiducial cross sections in the four-lepton channel. The presentation of the results begins with the inclusive fiducial cross sections in Sec. 5.1. Particularly noteworthy is Sec. 5.1.1, which shows the novel approach of measuring the inclusive fiducial cross section while leaving the irreducible background normalisation unconstrained in the fit. While the inclusive results provide a holistic overview of the decay channel, in order to have a deeper insight, it is necessary to go differential. Section 5.2 presents the measurement of fiducial cross sections in bins of 32 kinematic observables, targeting the Higgs boson production, decay, and the associated jet activity. Additionally, the fiducial framework put in place extends beyond mere cross section measurement purposes. It can also be used to constrain parameters of the SM, i.e., the fiducial cross sections can be interpreted in terms of such parameters. Section 5.3 presents the first constraint on the Higgs boson trilinear self-coupling via fiducial cross sections in single-Higgs production, while Sec. 5.4 quotes the Run 2 constraints of the Higgs boson couplings to the charm and bottom quarks.

5.1 Inclusive fiducial cross section

The measurement of the fiducial cross sections rely on a maximum likelihood fit, which is performed simultaneously in all final states ($4e$, 4μ, $2e2\mu$), years (2016, 2017, 2018), and bins of each variable, to the observed $m_{4\ell}$ distribution. In the fit procedure, the mass of the Higgs boson is fixed to $m_{\mathrm{H}} = 125.38$ GeV. Compared to the previous fiducial analyses [130, 134], the fit range is extended from $105 < m_{4\ell} < 140$ GeV to $105 < m_{4\ell} < 160$ GeV This change, influenced by studying the possibility to constrain the irreducible ZZ background directly from data (cfr Sec. 5.1.1), leads to a slight reduction in the final systematic uncertainty. An Asimov dataset [177] has been used to gauge the sensitivity of the analysis and test the behaviour of the systematic uncertainty in the two different mass ranges. This artificial dataset is generated to match perfectly the expected distributions while suppressing statistical fluctuations. During the generation, the nuisance parameters are frozen to their prefit values. When estimating parameters using an Asimov dataset, one obtains the true parameter values. The inclusive $\mu = \sigma/\sigma_{\mathrm{SM}}$ parameter is measured to be:

$$\mu^{\mathrm{Asimov}}_{[105,140]} = 1.00 \pm 0.085\ (\mathrm{stat})^{+0.089}_{-0.068}\ (\mathrm{syst}) \qquad \mu^{\mathrm{Asimov}}_{[105,160]} = 1.00 \pm 0.085\ (\mathrm{stat})^{+0.082}_{-0.060}\ (\mathrm{syst})$$

Extending the mass region reduces of approximately 10% the systematic component of the uncertainty. The mass range $105 < m_{4\ell} < 160$ GeV is the largest region around the Higgs boson peak where the background can be considered with good approximation as flat. The extended region offers a handle to better constrain the background nuisances, especially the leading electron efficiency uncertainty.

The inclusive fiducial cross section for Higgs boson production in the four-lepton final state is:

$$\begin{aligned}\sigma_{\text{fid}} &= 2.73 \pm 0.22\ (\text{stat}) \pm 0.15\ (\text{syst})\ \text{fb} \\ &= 2.73 \pm 0.22\ (\text{stat}) \pm 0.12\ (e) \pm 0.05\ (\text{lumi}) \pm 0.05\ (\text{bkg th}) \pm 0.03\ (\mu)\ \text{fb}\end{aligned} \tag{5.1}$$

for a Higgs boson mass of $m_{\text{H}} = 125.38$ GeV. The corresponding scan of the test statistic is reported in Fig. 5.1. The four-lepton mass distributions and the shapes resulting from the maximum likelihood fit are reported in Fig. 5.2. The cross section is measured with a total precision of 10%, where the statistical uncertainty is at the level of 8%, whereas the systematic component is at the level of 5%. The systematic uncertainty is dominated by the component related to electrons, especially the electron efficiency, which is the leading systematic of this analysis with a 4% impact on the final result. On the other hand, the uncertainties related to muons, backgrounds predictions, and the luminosity measurement, play a minor role on the overall systematic uncertainty on σ_{fid}.

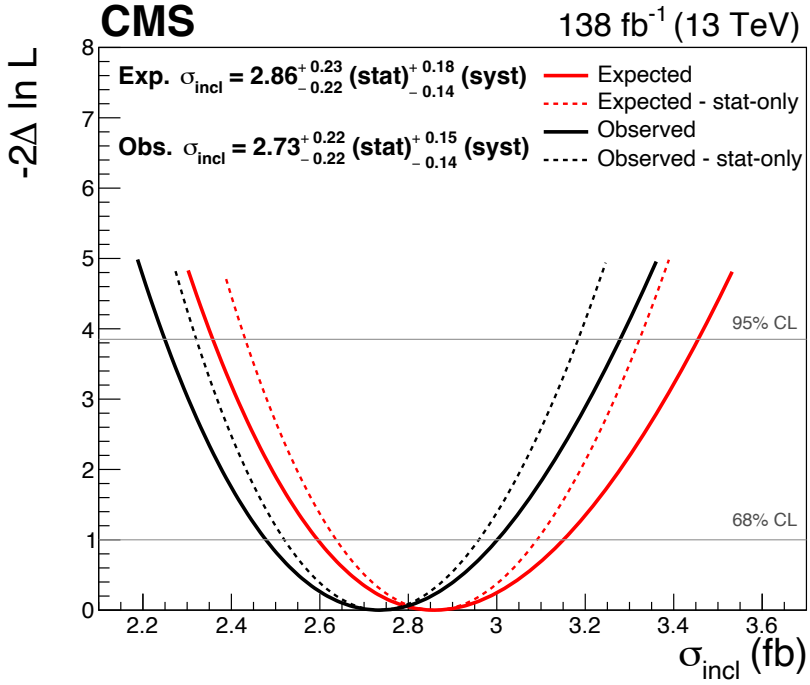

Figure 5.1: Scan of the test statistic for the measured inclusive fiducial cross section. The scan is shown with (solid line) and without (dashed line) systematic uncertainties profiled in the fit.

The result is in agreement with the SM expectation of $\sigma_{\text{fid}}^{\text{SM}} = 2.86 \pm 0.15$ fb, which is computed as:

$$\sigma_{\text{fid}}^{\text{SM}} = \sum_{i}^{\text{Prod.modes}} \sigma_i \times \text{BR}(\text{H}\to 4\ell) \times \mathcal{A}_{\text{fid}}^{i} \tag{5.2}$$

In this equation, the cross section and the BR are taken from [27], representing the best-known values from theoretical calculations. The fiducial acceptance ($\mathcal{A}_{\text{fid}}$), defined in

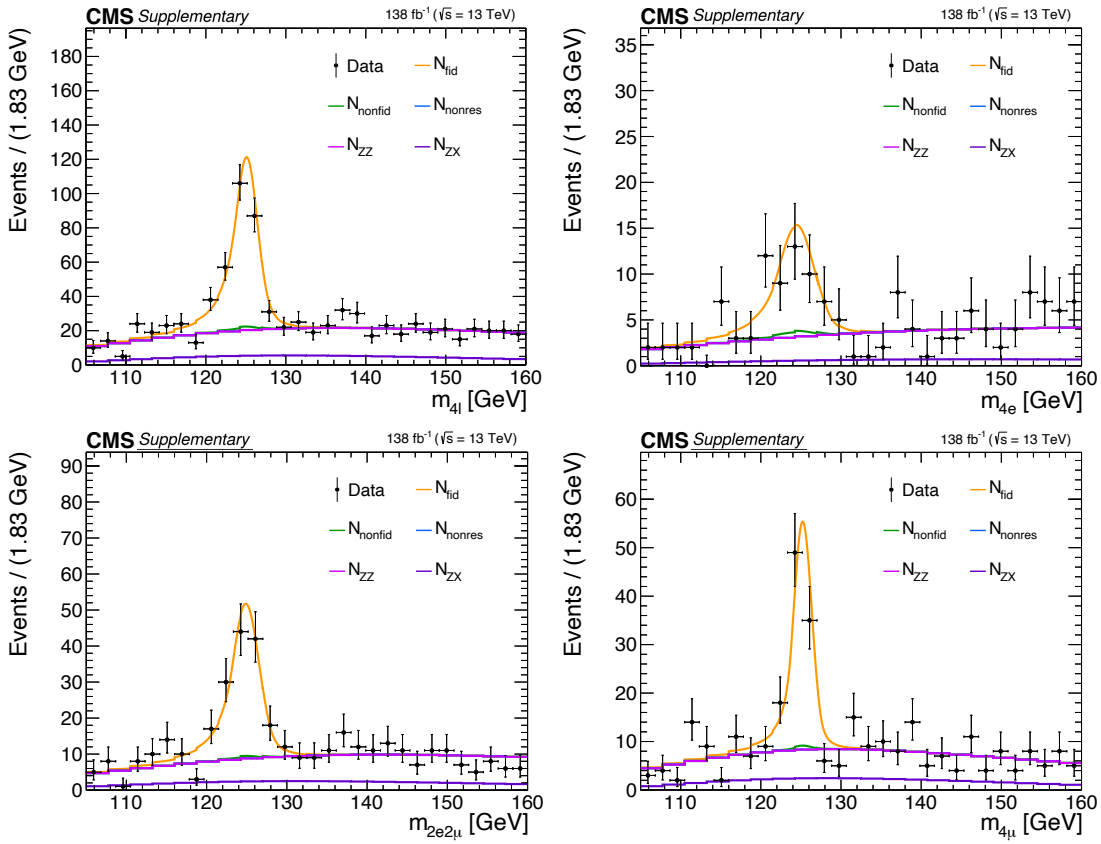

Figure 5.2: Four-lepton mass distributions for the inclusive final state (4ℓ) and three final states considered in the analysis ($4e$, 4μ, $2e2\mu$). The black points with error bars represent the Data in the $105 < m_{4\ell} < 160$ GeV mass range. The coloured lines represent the signal and background contributions as described Sec. 4.7. The yellow and green lines represent the fiducial ($N_{\text{fid}}(m_{4\ell})$) and non-fiducial ($N_{\text{nonfid}}(m_{4\ell})$) distributions originating from within and outside the fiducial volume, respectively. The blue line represents the non-resonant contribution ($N_{\text{nonres}}(m_{4\ell})$) to the signal arising from the WH, ZH, and ttH production mechanisms. Compared to other contributions, N_{nonres} is small (in the given mass interval) and therefore not visible in the figure. The pink and purple lines represent the background templates for the reducible and irreducible backgrounds, respectively. The sum of these two contributions is considered in the analysis and referred to as N_{bkg}.

Sec. 4.7, is computed using the `POWHEG` samples. The uncertainty in this prediction is the combination of the uncertainties in the fiducial acceptance (2%), the BR (2%), and variations of the PDF replicas, α_s value, and QCD scales. These theoretical uncertainties are taken from [27] and depend on the specific production mechanisms: the QCD scale uncertainties range between 0.3% and 6%; while the parametric (PDF+α_s) uncertainties range between 1.6% and 3.6%.

The result of Eq. 5.1 represents the last step of a very long journey, started almost 10 years ago. Fiducial cross sections have been measured since the beginning of the CMS physics programme, and Fig. 5.3 shows the evolution of the precision on the inclusive cross sections at different stages of this journey. Starting from the first result at $\sqrt{s} = 7$ TeV with only 5 fb^{-1} of data and an uncertainty of more than 100%, the precision is now at the level of 10%. As the CMS detector accumulated more statistics, there was a continuous decrease in the statistical uncertainty, while at the same time a similar decrease of the systematic component due to the increased knowledge of the experimental effects. In the

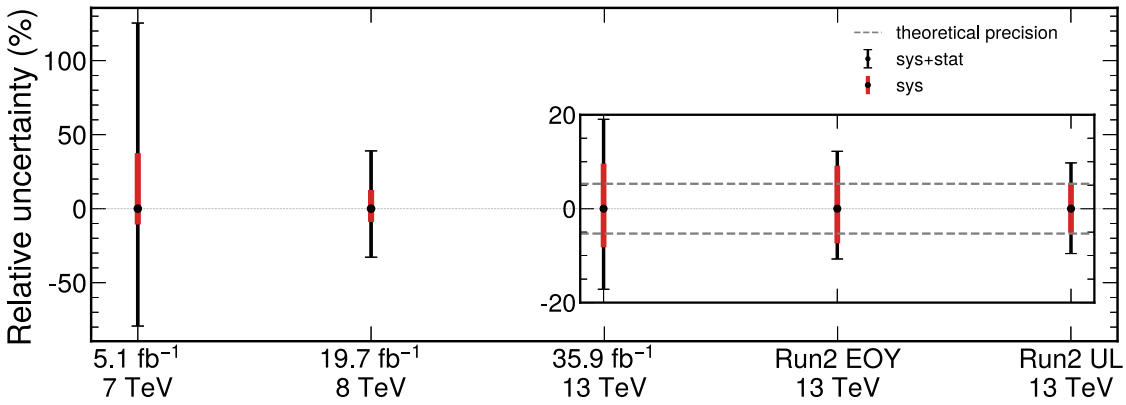

Figure 5.3: Evolution of the precision of the inclusive fiducial cross sections. The results at 7 TeV and 8 TeV are taken from [134]. The result at 13 TeV with 35.9 fb^{-1}, corresponding to the 2016 data-taking year, is taken from [137]. The Run 2 result obtained with EOY calibration and without implementing the RMS method for the electron efficiency systematic is taken from [42]. The legacy result presented in this thesis (Run2 UL) is taken from [21].

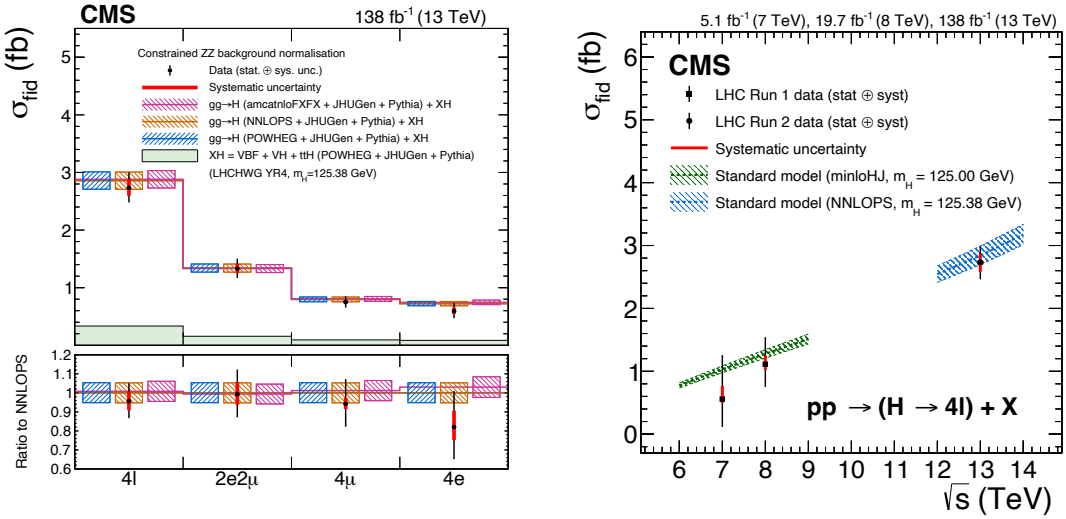

Figure 5.4: Measured inclusive fiducial cross section for the various final states (left); and as a function of the center-of-mass energy $\sqrt{s}$ (right). In the left panel the acceptance and theoretical uncertainties are calculated using POWHEG (blue), NNLOPS (orange), and MadGraph5_aMC@NLO (pink). The subdominant component of the signal (VBFH + VH + ttH) is denoted as XH and is fixed to the SM prediction. In the right panel the acceptance is calculated using POWHEG at $\sqrt{s} = 13$ TeV and HRes [174, 192] at $\sqrt{s} = 7$ and 8 TeV.

plot, two distinct Run 2 results may be noticed. The first one is the result computed with EOY samples, while the second one shows the result presented in this thesis and represents the Run 2 legacy value for the $H \to ZZ \to 4\ell$ channel. Compared to the preliminary Run 2 result, there is a 15% reduction of the uncertainty, driven by a 40% decrease of the systematic component. This significant reduction is due to the revised method for the estimation of the electron scale factors uncertainties as explained in Sec. 4.8, the improved UL objects calibration, and the extended mass range for the fit. In addition, the systematic uncertainty is now the same level of the current theoretical precision, proving

that the Higgs physics is indeed entering into the precision era, where the properties of the Higgs boson will be probed with unprecedented precision.

As explained in Sec. 4.7, the inclusive cross section is also measured separately in the three final states $2e2\mu$, 4μ, and $4e$. The corresponding results are shown in the left plot of Fig. 5.4. The largest discrepancy from the SM occurs in the $4e$ channel, which, however, suffers from low statistics and large fluctuations. As the dominant systematic uncertainty of the analysis is the electron efficiency, the 4μ channel is measured with the slightest systematic uncertainty of 4% and a total precision of 13%. In contrast, the $4e$ channel exhibits a more substantial systematic component at the level of 10%, resulting in an overall precision of 22%. The mixed $2e2\mu$ channel is measured with a precision equivalent to that of the four-muon channel, standing at 13%. This equivalence arises due to a trade-off: while the decay of one of the Z bosons into electrons increases the systematic uncertainty, the channel benefits from a higher event count. The right plot of Fig. 5.4 also reports the evolution of the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ fiducial cross section as a function of the centre-of-mass energy.

5.1.1 Unconstrained irreducible background normalisation

In addition to the results presented above, the inclusive cross section is also measured with a different treatment of the irreducible ZZ background ($\mathrm{q\bar{q}} \to \mathrm{ZZ} + \mathrm{gg} \to \mathrm{ZZ}$). The $\mathrm{H} \to \mathrm{ZZ}$ analyses have traditionally relied on a simulation-based approach to estimate this component of the background, fixing both the shape and the normalisation to the MC expectations. However, in the context of this thesis, the feasibility of constraining the ZZ normalisation directly from data, still taking shapes from simulations, is investigated. This approach is motivated by both theoretical and experimental considerations, profiting of the large dataset accumulated during the Run 2 of the LHC. From a theoretical perspective, the aim is to be sensitive to potential BSM contributions and effects that might not be adequately modelled by the available MC simulations. On the experimental side, this approach may offer the advantage of improving the precision of the result as theoretical and luminosity uncertainties would no longer contribute to the ZZ processes.

The core of the statistical procedure remains unaltered, the only change lies in considering the normalisation of the irreducible background as an unconstrained parameter in the fit, in addition to the signal fiducial cross section. As anticipated at the beginning of this section, the mass range is set to $105 < m_{4\ell} < 160$ GeV to include sidebands.

The measurement of the inclusive cross section is then repeated with this new approach. The results are presented in the left panel of Fig. 5.5, both for the inclusive final state and the three distinct final states. The correlation coefficient (ρ) between the measured inclusive fiducial cross section and the ZZ normalisation in the 4ℓ final state is found to be $\rho = -0.03$, while the correlations between the ZZ normalisation in each final state and their respective fiducial cross sections are shown in the right panel of Fig. 5.5. All the results are summarised in Tab. 5.1.

In general, no loss of sensitivity is observed in the measurement of signal cross sections across the two different methods, and the measured values are well in agreement with each other. The uncertainty on the measured background normalisation of $\sim 10\%$ is larger than the theoretical uncertainty in its prediction, which results to be approximately 6%. As of now, the Run 2 statistics is not yet sufficient to have a better estimation of the ZZ processes from data rather than from MC.

An intriguing systematic trend is evident across all final states: the data consistently favour a slightly higher ZZ background, deviating at the 1σ level from predictions. On average, the post-fit value is 7% higher. This effect is also visible when the ZZ normalisation is constrained to the SM predictions. In this case, the post-fit value is only 3%

higher than MC expectations. The effect is milder as the nuisance constraining terms tame the variation. It is thus a general trend of data to push the ZZ normalisation up, more conspicuously when the parameter is unconstrained in the fit. When comparing the two approaches, although one might naively anticipate a lower cross section due to a higher underlying background, the outcome of the measurement contradicts this expectation. The impact on the cross section is instead almost negligible, with certain cases ($2e2\mu$ and inclusive) even exhibiting a slight increase. This apparent counter-intuitive outcome is caused by a twofold interplay. First, the great $m_{4\ell}$ resolution reduces the impact of a shift of the background. Second, the interplay with nuisance parameters effectively counterbalances this effect. In particular, the leading electron efficiency nuisance is less pulled to larger values when the ZZ normalisation is allowed to float (cfr Appendix B).

Despite providing a more comprehensive description of the H $\rightarrow$ ZZ $\rightarrow$ 4ℓ channel relying directly on data and being strongly requested by theoreticians, this approach is not extended to the measurement of differential observables. When binning the dataset, the ZZ background may suffer for low yields in most of the bins, thus introducing the need of merging most of them to avoid unstable fit and results completely dominated by the uncertainties. In addition, merging bins would also introduce additional model dependence, which is something that is desired to keep as low as possible in fiducial analyses.

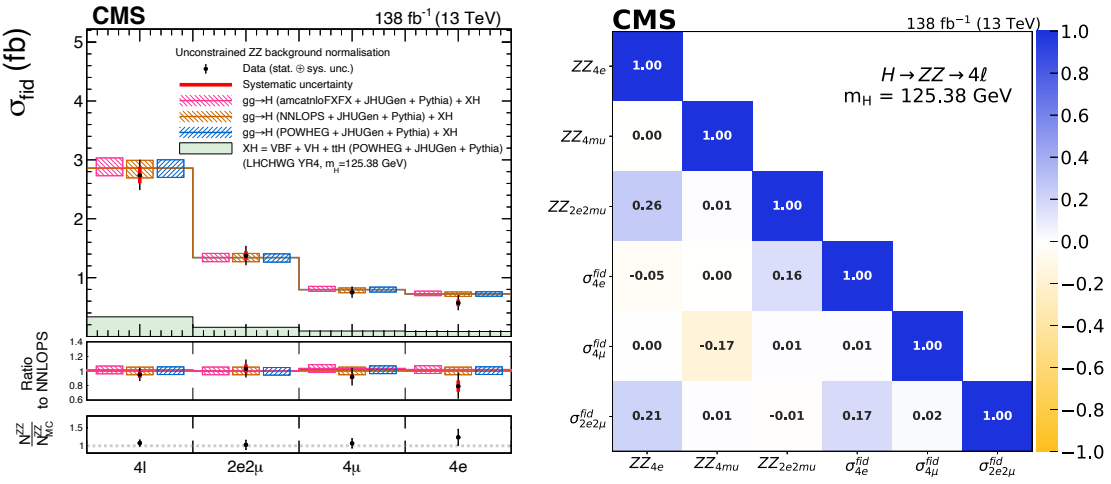

Figure 5.5: Inclusive fiducial cross section measured for the various final states with the irreducible backgrounds normalisation ZZ unconstrained in the fit (left) and the corresponding correlation matrix (right). The acceptance and theoretical uncertainties in the differential bins are calculated using POWHEG (blue), NNLOPS (orange), and MadGraph5_aMC@NLO (pink). The subdominant component of the signal (VBFH + VH + ttH) is denoted as XH and it is fixed to the SM prediction. The ratio of the measured cross section to the theoretical prediction obtained from each generator is shown in the central panel, while the lower panel shows the ratio between the values derived from the measured ZZ normalisation and the MC prediction.

5.2 Differential fiducial cross sections

Following the inclusive results, this section presents the differential fiducial cross sections. The measured values are compared to the SM expectations following the strategy outlined in the previous section, where the fiducial acceptances are computed with `POWHEG`, `NNLOPS`,

Table 5.1: Measured inclusive fiducial cross section and ±1 standard deviation uncertainties for the various final states at $m_{\mathrm{H}} = 125.38$ GeV. The upper row summarises the results obtained when the irreducible background normalisation is constrained to the SM expectation and theoretical uncertainty, while the lower section present the results from a fit with the ZZ normalisation treated as an unconstrained parameter. The first row presents the fiducial cross section, the middle row shows the ZZ background normalisation extracted from the fit, and the bottom row the ZZ estimation from MC. The uncertainties on $N_{\mathrm{MC}}^{\mathrm{ZZ}}$ are the pre-fit uncertainties summing the statistical and systematic uncertainty.

	$4e$	4μ	$2e2\mu$	Inclusive
		Constrained ZZ background		
σ_{fid}	$0.59^{+0.13}_{-0.12}$ fb	$0.75^{+0.10}_{-0.09}$ fb	$1.33^{+0.17}_{-0.16}$ fb	$2.73^{+0.22}_{-0.22}$ (stat)$^{+0.15}_{-0.14}$ syst fb
		Unconstrained ZZ background		
σ_{fid}	$0.57^{+0.15}_{-0.12}$ fb	$0.75^{+0.10}_{-0.09}$ fb	$1.37^{+0.17}_{-0.16}$ fb	$2.74^{+0.24}_{-0.23}$ (stat)$^{+0.14}_{-0.11}$ (syst) fb
N^{ZZ}	92^{+16}_{-13}	162^{+19}_{-18}	193^{+23}_{-21}	445^{+27}_{-26} (stat)$^{+21}_{-19}$ (syst)
$N_{\mathrm{MC}}^{\mathrm{ZZ}}$	74^{+7}_{-8}	152^{+7}_{-8}	188^{+13}_{-14}	414^{+24}_{-28}

and MADGRAPH5_aMC@NLO. The theoretical uncertainties related to the choice of the QCD scales, α_s value, and PDF replica are evaluated in each differential bin for the ggH process, while for the other production modes the inclusive value is used. As ggH dominates, the differential computation is carried out only for this production mode. The compatibility between the observed results and the corresponding SM prediction is assessed by computing the p-value of each measurement. As in the asymptotic approximation the test statistic $-2\Delta \ln \mathscr{L}$ is approximated by a χ^2 distribution (cfr Sec. 4.4), the p-value can be defined as:

$$p = \int_{-2\Delta\mathcal{L}(\mathrm{SM})}^{\infty} \chi^2_{\mathrm{n_{bins}}}(x)\mathrm{d}x \tag{5.3}$$

where $-2\Delta\mathcal{L}(\mathrm{SM})$ corresponds to the difference between the likelihood function evaluated at the SM point and at the best-fit point. The numbers will be reported in the corresponding differential plot. Overall, a good compatibility between measured values and the SM predictions is observed for all the cross sections considered in the analysis, with p-values ranging from 0.05 to 0.99.

The fiducial cross sections measured in bins targeting the production of the Higgs boson are reported in Fig. 5.6. Figure 5.7 shows the measurements related to the associated jet activity, i.e., number of jets and p_{T} of the leading and sub-leading jet, whereas Fig. 5.8 focuses on variables targeting the di-jet system, hence enhancing the sensitivity to phase space regions where VBFH and ttH dominate. Figure 5.9 reports the cross section measured in bins of observables of the $\mathrm{H} + j$ and $\mathrm{H} + jj$ system. The cross sections measured in differential bins of the rapidity-weighed jet vetoes are reported in Fig. 5.10.

The differential fiducial cross sections in bins of observables targeting the decay of the Higgs boson in four leptons are shown in Figs. 5.11–5.17. Cross sections are also measured in differential bins of the six decay matrix-element discriminants introduced in Sec. 4.6. The results of these measurements, shown in Figs. 5.18–5.23, are also compared to distributions corresponding to various anomalous coupling hypotheses. Each discriminant is compared to the anomalous coupling scenario for which it is optimised. The alternative discriminants $\mathcal{D}_{0^-}^{\mathrm{dec}}$ and $\mathcal{D}_{0\mathrm{h}+}^{\mathrm{dec}}$ are compared to the distributions obtained for pure anomalous coupling scenarios corresponding to $f_{a3} = 1$ and $f_{a2} = 1$, respectively. The

corresponding interference discriminants, $\mathcal{D}_{\text{CP}}^{\text{dec}}$ and $\mathcal{D}_{\text{int}}^{\text{dec}}$, are compared to the interference scenario where $f_{a3} = 0.5$ and $f_{a2} = 0.5$, respectively. A value of $f_{ai} = 0.5$ corresponds to a maximal mixing between the SM and the BSM hypothesis. To inspect the Λ_1 and $\Lambda_1^{\text{Z}\gamma}$ couplings, the interference discriminant is not measured as it does not provide additional information and the corresponding $\mathcal{D}_{\text{alt}}$ can also be used to study the interference. For this reason, the measurements of $\mathcal{D}_{\Lambda 1}^{\text{dec}}$ and $\mathcal{D}_{\Lambda 1}^{\text{Z}\gamma,\text{dec}}$ are compared with the pure anomalous coupling scenario $f_{\Lambda 1} = 1$ and $f_{\Lambda 1}^{\text{Z}\gamma} = 1$, as well as to the interference hypotheses $f_{\Lambda 1} = 0.5$ and $f_{\Lambda 1}^{\text{Z}\gamma} = 0.5$. These values of f_{ai}, $f_{\Lambda 1}^{\text{Z}\gamma}$, and $f_{\Lambda 1}$ are illustrative extreme scenarios chosen for a qualitative representation of the corresponding kinematic discriminants. The best constraints on these parameters are much stricter, as reported in Ref. [128].

For all the decay observables, since the final state is sensitive to interference effects in the case of identical particles, the results are also quoted for same- and different-flavour final states. The largest deviations with respect to the expected values are observed in the central bins of $\cos\theta_2$ and ϕ and are compatible with statistical fluctuations in the observed data. The p-values of these two measurements are found to be 0.23 and 0.24, respectively, thus corroborating the compatibility with the SM predictions.

The results for the double-differential observables are shown in Fig. 5.24 and show consistency with the SM expectations, with the largest difference observed in the p_{T}^{H} bins in the $\text{N}_{\text{jets}} = 1$ phase space region. The deficit in the low-p_{T}^{H} bins for $\text{N}_{\text{jets}} = 1$ is explained by large correlations with the high-p_{T}^{H} bin of $\text{N}_{\text{jets}} = 1$ and the first p_{T}^{H} bin in the $\text{N}_{\text{jets}} > 1$ phase space regions, where the fit to the data shows an excess with respect to the SM prediction.

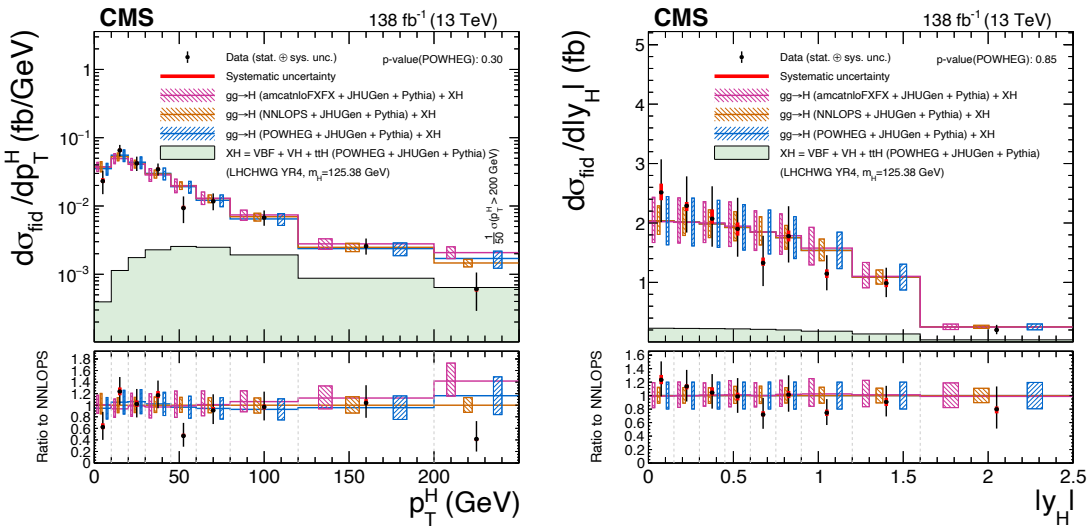

Figure 5.6: Differential cross sections as functions of the transverse momentum of the Higgs boson $p_{\mathrm{T}}^{\mathrm{H}}$ (left) and of the rapidity of the Higgs boson $|y_{\mathrm{H}}|$ (right). The fiducial cross section in the last bin (left) is measured for events with $p_{\mathrm{T}}^{\mathrm{H}} > 200$ GeV and normalized to a bin width of 50 GeV. The acceptance and theoretical uncertainties in the differential bins are calculated using the ggH predictions from the POWHEG generator (blue) normalized to next-to-next-to-next-to-leading order (N^{3}LO). The subdominant component of the signal (VBFH + VH + ttH) is denoted as XH and is fixed to the SM prediction. The measured cross sections are also compared with the ggH predictions from NNLOPS (orange) and MadGraph5_aMC@NLO (pink). The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratios of the measured cross sections and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical predictions.

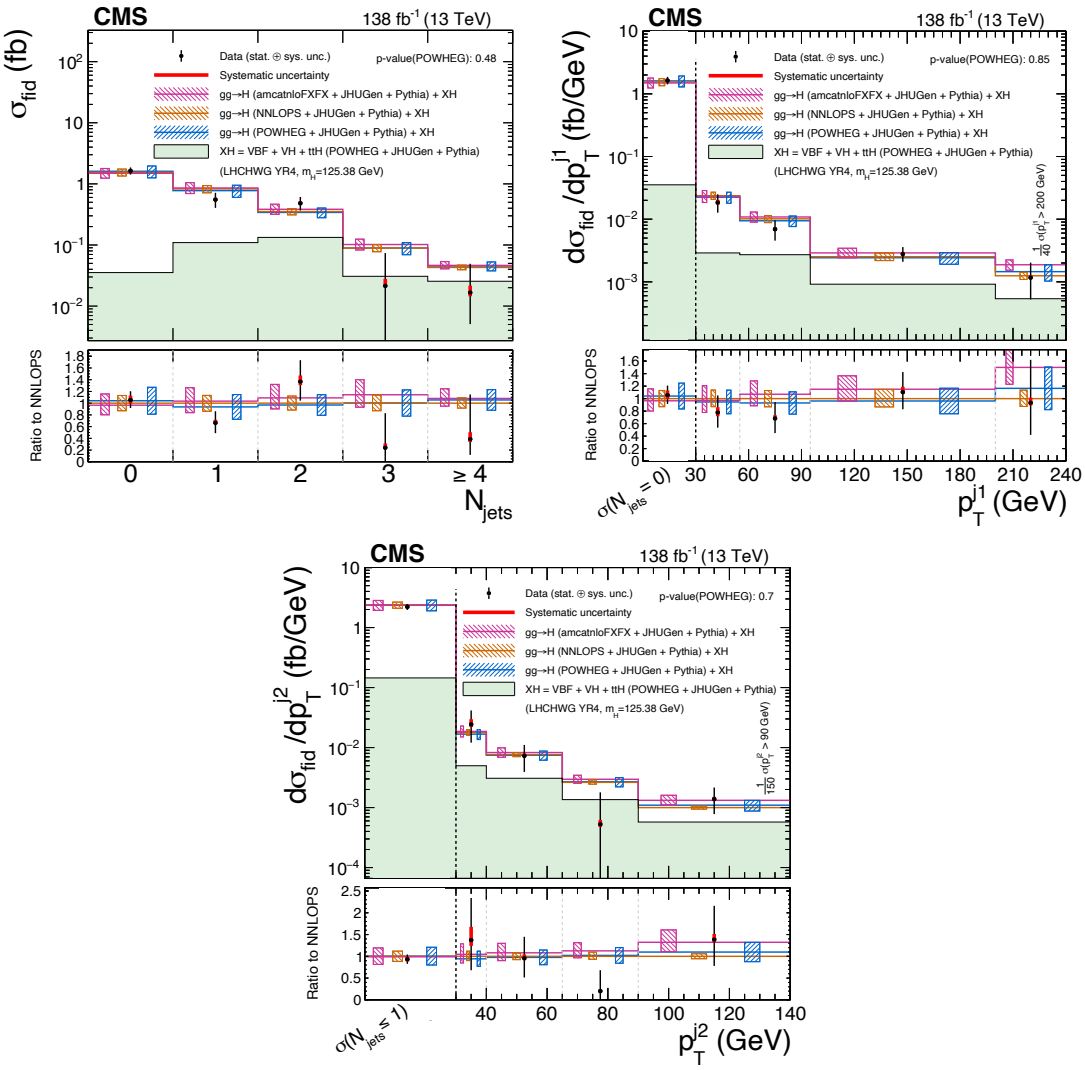

Figure 5.7: Differential cross sections as functions of the number of jets in the event (upper left) and of the p_{T} of the leading (upper right) and subleading (lower) jet. Upper right: the fiducial cross section in the last bin is measured for events with $p_{\mathrm{T}}^{\mathrm{j1}} > 200$ GeV and normalized to a bin width of 40 GeV. The first bin comprises all events with less than one jet, for which $p_{\mathrm{T}}^{\mathrm{j1}}$ is undefined. Lower: the fiducial cross section in the last bin is measured for events with $p_{\mathrm{T}}^{\mathrm{j2}} > 90$ GeV and normalized to a bin width of 150 GeV. The first bin comprises all events with less than two jet, for which $p_{\mathrm{T}}^{\mathrm{j2}}$ is undefined. The acceptance and theoretical uncertainties in the differential bins are calculated using the ggH predictions from the POWHEG generator (blue) normalized to N^{3}LO. The subdominant component of the signal (VBFH + VH + ttH) is denoted as XH and is fixed to the SM prediction. The measured cross sections are also compared with the ggH predictions from NNLOPS (orange) and MadGraph5_aMC@NLO (pink). The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratios of the measured cross sections and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical predictions.

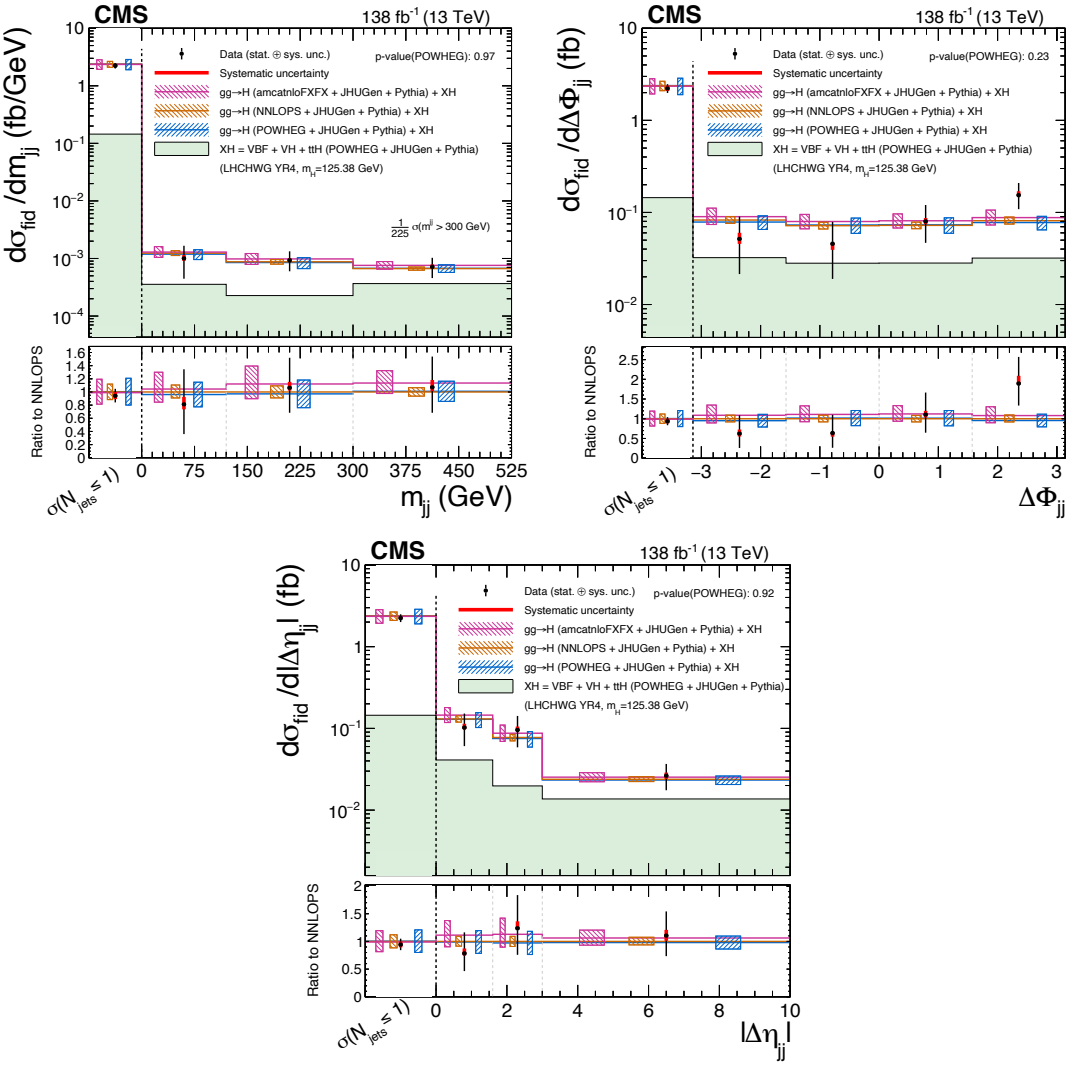

Figure 5.8: Differential cross sections as functions of the invariant mass m_{jj} (upper left), the difference in azimuthal angle $\Delta\phi_{jj}$ (upper right) the difference in pseudorapidity $|\Delta\eta_{jj}|$ (lower) of the dijet system. Upper Left: the fiducial cross section in the last bin is measured for events with $m_{jj} > 300$ GeV and normalized to a bin width of 225 GeV. The first bin comprises all events with less than two jets, for which m_{jj} is undefined. Upper right: the first bin comprises all events with less than two jet, for which $|\Delta\phi_{jj}|$ is undefined. Lower: the first bin comprises all events with less than two jet, for which $|\Delta\eta_{jj}|$ is undefined. The acceptance and theoretical uncertainties in the differential bins are calculated using the ggH predictions from the POWHEG generator (blue) normalized to N^{3}LO. The subdominant component of the signal (VBFH + VH + ttH) is denoted as XH and is fixed to the SM prediction. The measured cross sections are also compared with the ggH predictions from NNLOPS (orange) and MadGraph5_aMC@NLO (pink). The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratios of the measured cross sections and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical predictions.

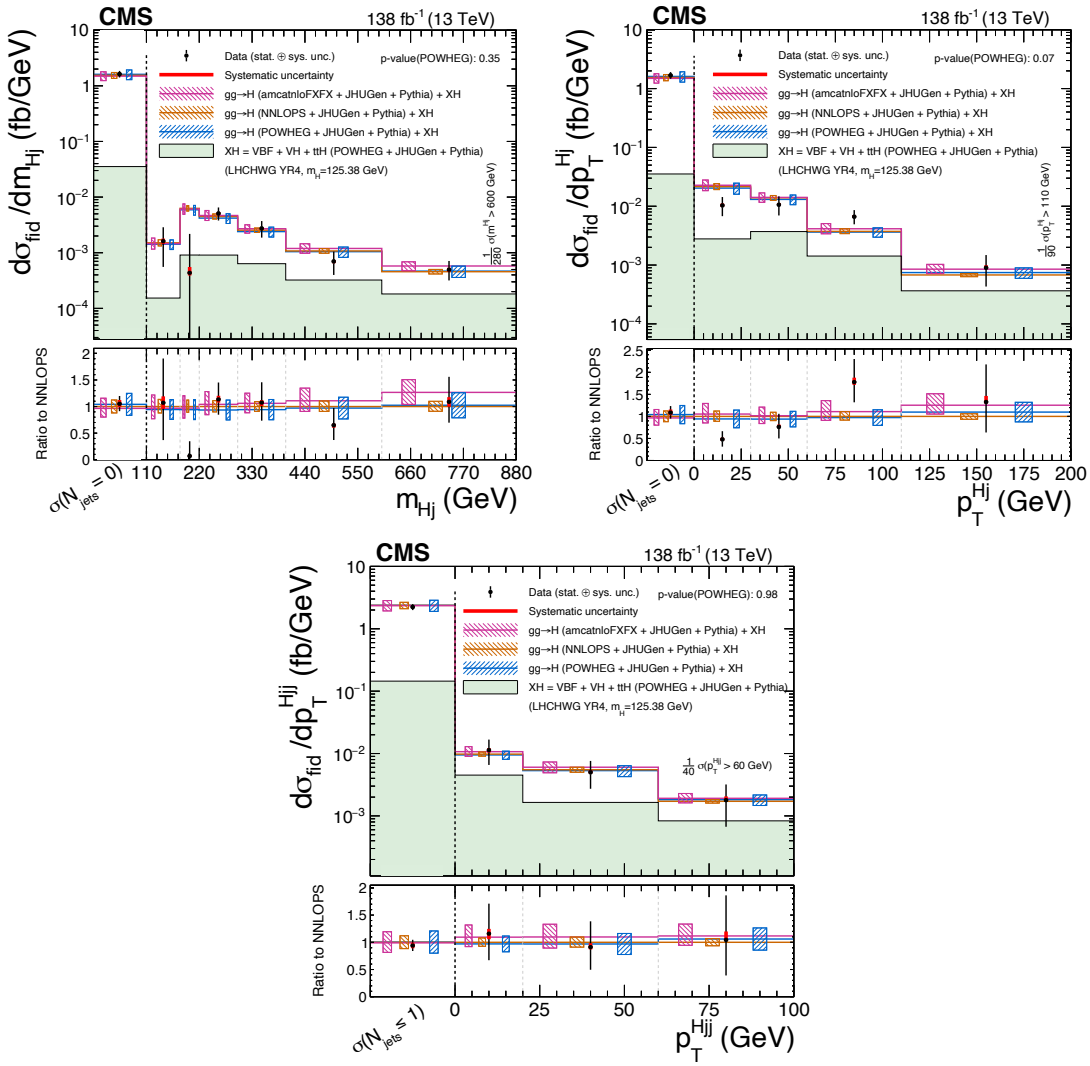

Figure 5.9: Upper left: differential cross sections as functions of the invariant mass of the $\mathrm{H}+j$ system $m_{\mathrm{H}j}$, where j is the leading jet in the event. The fiducial cross section in the last bin is measured for events with $m_{\mathrm{H}j} > 600$ GeV and normalized to a bin width of 280 GeV. The first bin comprises all events with less than one jet, for which $m_{\mathrm{H}j}$ is undefined. Upper right: differential cross sections as functions of the transverse momentum of the $\mathrm{H}+j$ system $p_{\mathrm{T}}^{\mathrm{H}j}$. The fiducial cross section in the last bin is measured for events with $p_{\mathrm{T}}^{\mathrm{H}j} > 110$ GeV and normalized to a bin width of 90 GeV. The first bin comprises all events with less than one jet, for which $p_{\mathrm{T}}^{\mathrm{H}j}$ is undefined. Lower: differential cross sections as functions of the transverse momentum of the $\mathrm{H}+jj$ system $p_{\mathrm{T}}^{\mathrm{H}jj}$. The fiducial cross section in the last bin is measured for events with $p_{\mathrm{T}}^{\mathrm{H}jj} > 60$ GeV and normalized to a bin width of 40 GeV. The first bin comprises all events with less than two jet, for which $p_{\mathrm{T}}^{\mathrm{H}jj}$ is undefined. The acceptance and theoretical uncertainties in the differential bins are calculated using the ggH predictions from the POWHEG generator (blue) normalized to N^{3}LO. The subdominant component of the signal (VBFH + VH + ttH) is denoted as XH and is fixed to the SM prediction. The measured cross sections are also compared with the ggH predictions from NNLOPS (orange) and MadGraph5_aMC@NLO (pink). The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratios of the measured cross sections and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical predictions.

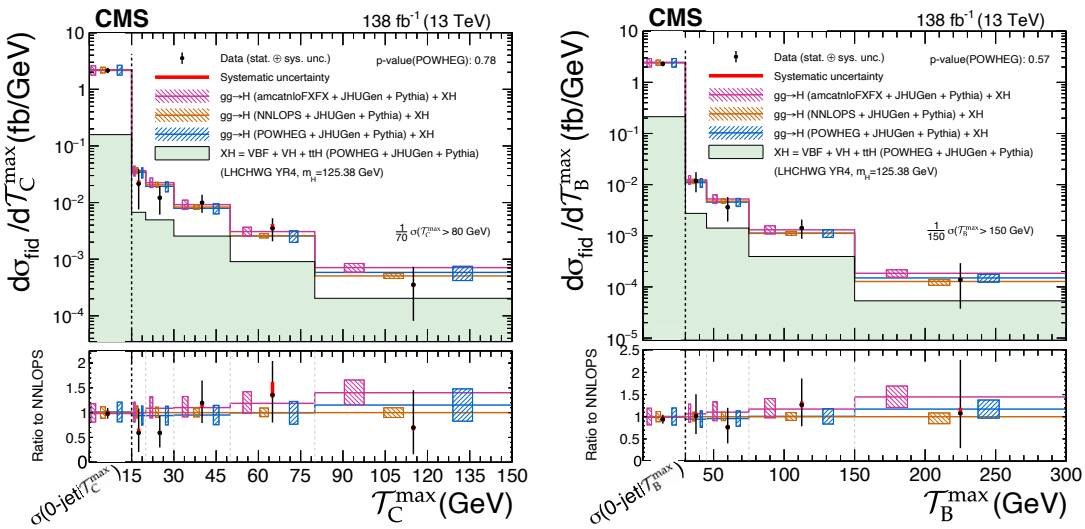

Figure 5.10: Left: differential cross sections as functions of the rapidity-weighed jet veto $\mathcal{T}_{\mathrm{C}}^{\mathrm{max}}$. The fiducial cross section in the last bin is measured for events with $\mathcal{T}_{\mathrm{C}}^{\mathrm{max}} > 80$ GeV and normalized to a bin width of 70 GeV. The first bin comprises all events in the 0-jet phase space region redefined as a function of $\mathcal{T}_{\mathrm{C}}^{\mathrm{max}}$, i.e., events with less than one jet, for which $\mathcal{T}_{\mathrm{C}}^{\mathrm{max}}$ is undefined, and events with $\mathcal{T}_{\mathrm{C}}^{\mathrm{max}} < 15$ GeV. Right: differential cross sections as functions of the rapidity-weighed jet veto $\mathcal{T}_{\mathrm{B}}^{\mathrm{max}}$. The fiducial cross section in the last bin is measured for events with $\mathcal{T}_{\mathrm{B}}^{\mathrm{max}} > 150$ GeV and normalized to a bin width of 150 GeV. The first bin comprises all events in the 0-jet phase space region redefined as a function of $\mathcal{T}_{\mathrm{B}}^{\mathrm{max}}$, i.e., events with less than one jet, for which $\mathcal{T}_{\mathrm{B}}^{\mathrm{max}}$ is undefined, and events with $\mathcal{T}_{\mathrm{B}}^{\mathrm{max}} < 30$ GeV. The acceptance and theoretical uncertainties in the differential bins are calculated using the ggH predictions from the POWHEG generator (blue) normalized to N^{3}LO. The subdominant component of the signal (VBFH + VH + ttH) is denoted as XH and is fixed to the SM prediction. The measured cross sections are also compared with the ggH predictions from NNLOPS (orange) and MadGraph5_aMC@NLO (pink). The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratios of the measured cross sections and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical predictions.

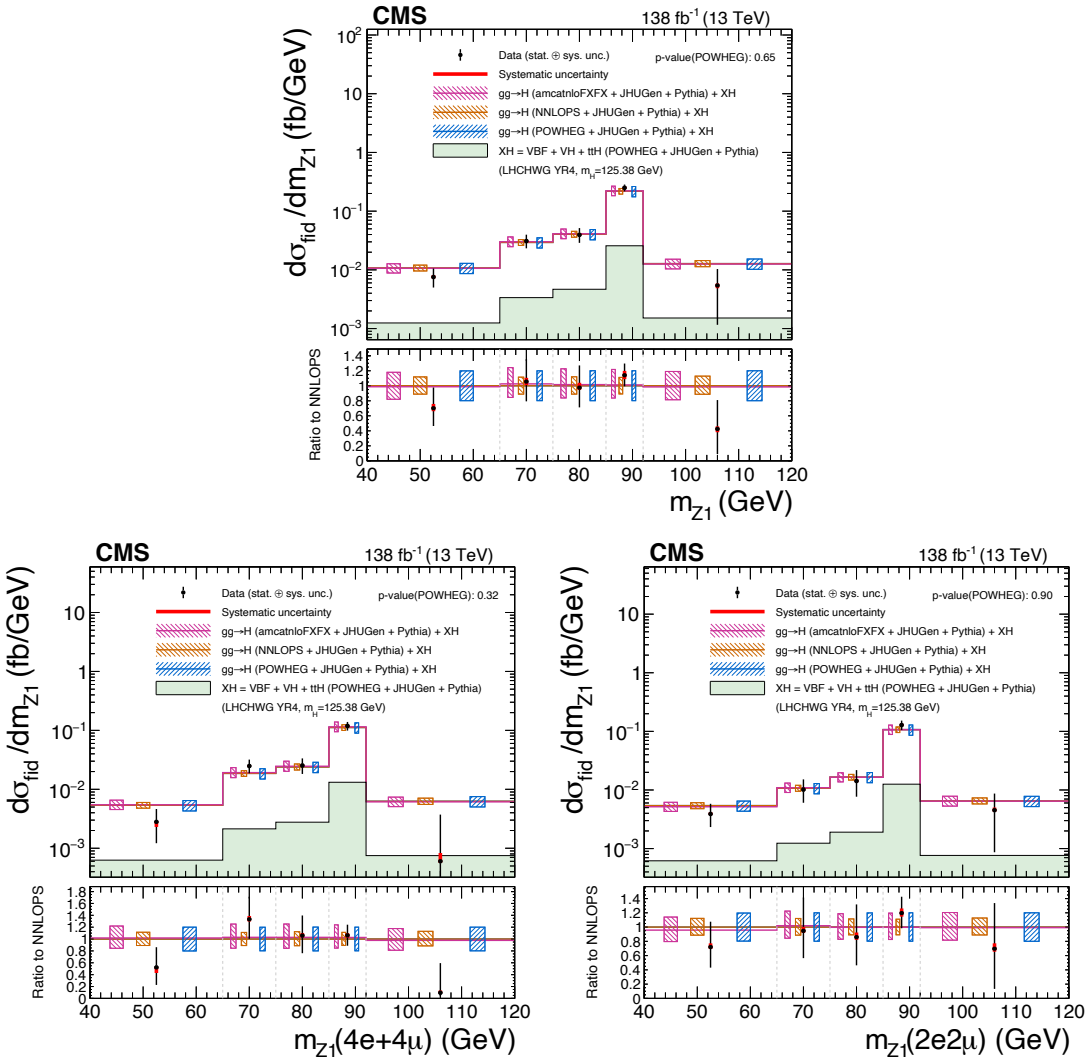

Figure 5.11: Differential cross sections as functions of the invariant mass of the leading dilepton pair m_{Z_1} in the 4ℓ (upper) and in the same-flavor (lower left) and different-flavor (lower right) final states. The acceptance and theoretical uncertainties in the differential bins are calculated using the ggH predictions from the POWHEG generator (blue) normalized to N^{3}LO. The subdominant component of the signal (VBFH + VH + ttH) is denoted as XH and is fixed to the SM prediction. The measured cross sections are also compared with the ggH predictions from NNLOPS (orange) and MadGraph5_aMC@NLO (pink). The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratios of the measured cross sections and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical predictions.

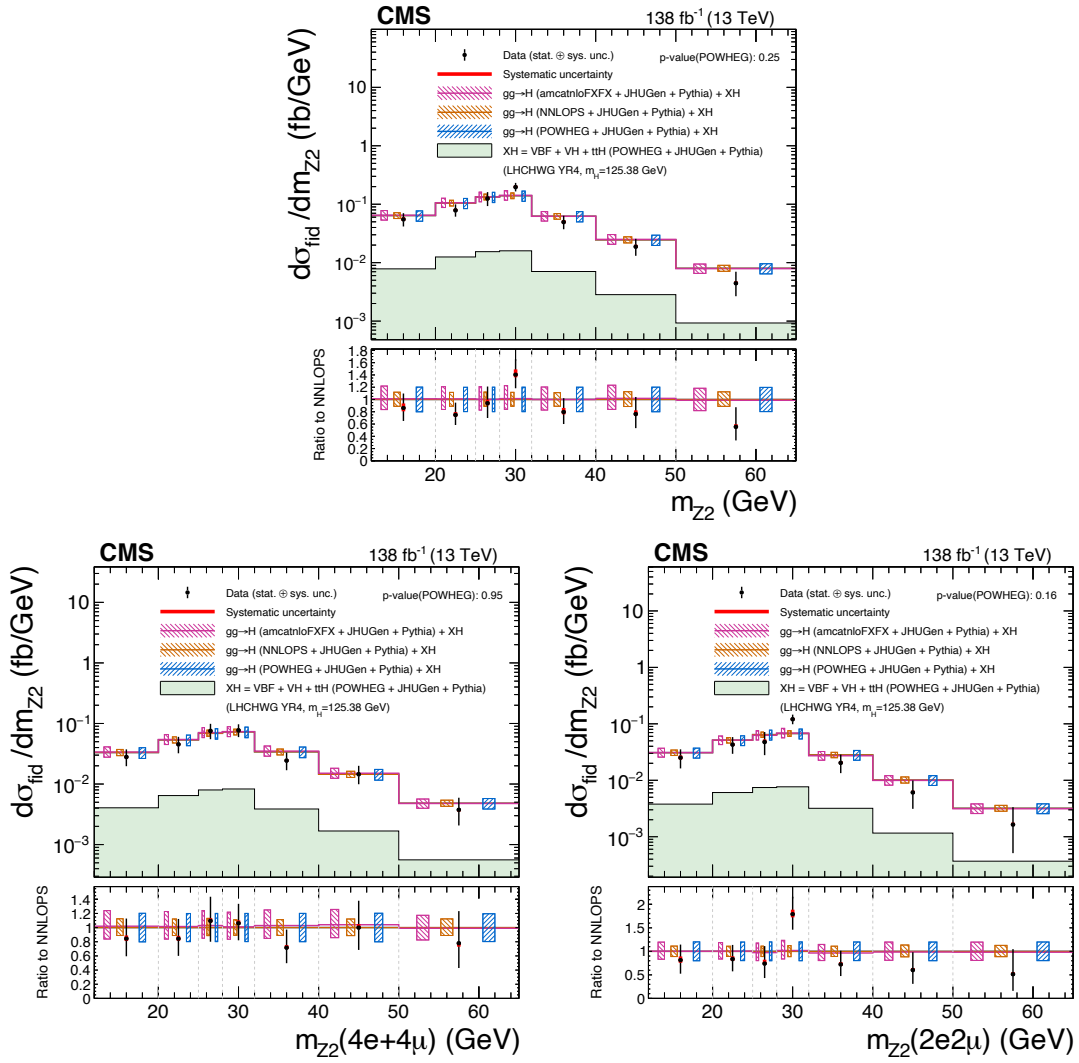

Figure 5.12: Differential cross sections as functions of the invariant mass of the subleading dilepton pair m_{Z_2} in the 4ℓ (upper) and in the same-flavor (lower left) and different-flavor (lower right) final states. The acceptance and theoretical uncertainties in the differential bins are calculated using the ggH predictions from the POWHEG generator (blue) normalized to N^{3}LO. The subdominant component of the signal (VBFH + VH + ttH) is denoted as XH and is fixed to the SM prediction. The measured cross sections are also compared with the ggH predictions from NNLOPS (orange) and MadGraph5_aMC@NLO (pink). The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratios of the measured cross sections and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical predictions.

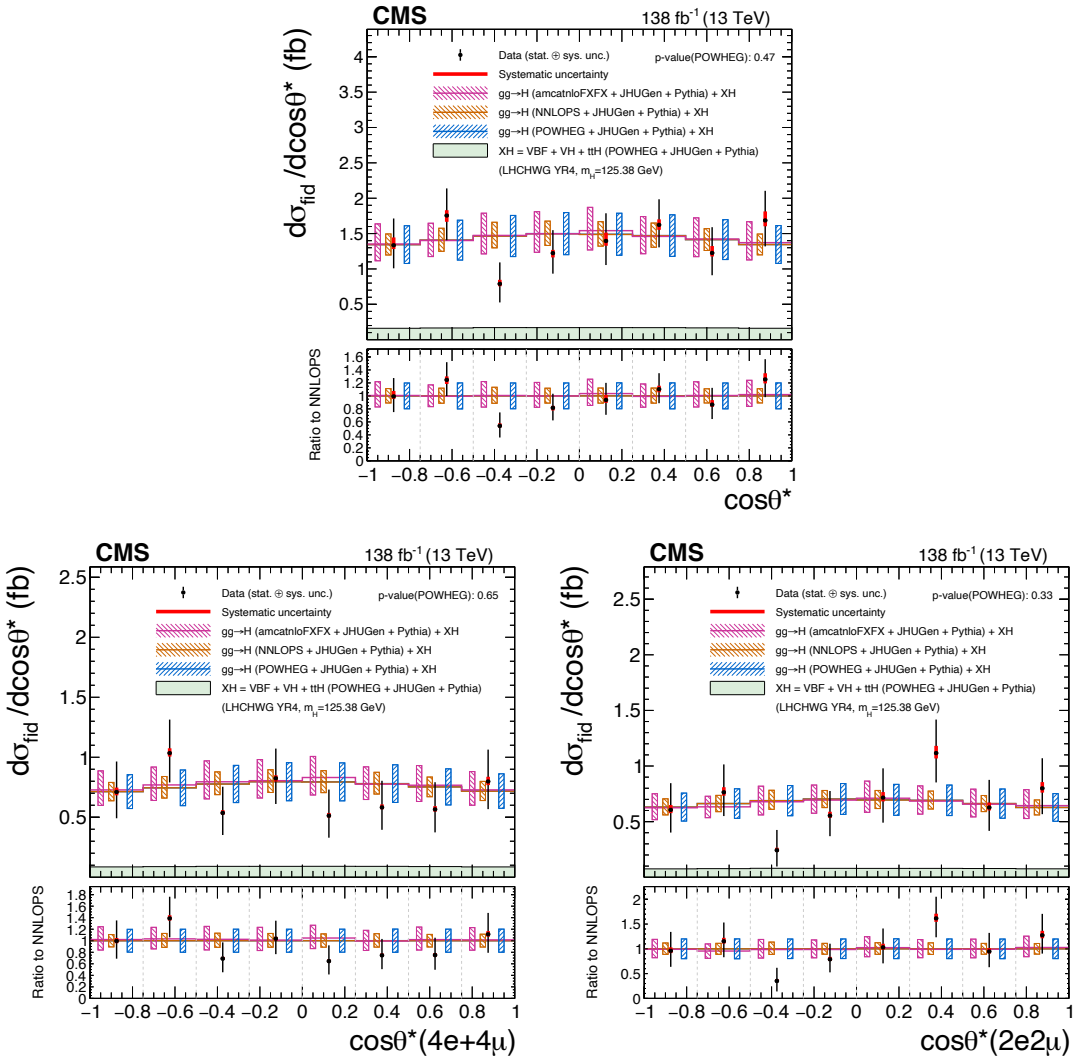

Figure 5.13: Differential cross sections as functions of $\cos\theta^*$ in the 4ℓ (upper) and in the same-flavor (lower left) and different-flavor (lower right) final states. The acceptance and theoretical uncertainties in the differential bins are calculated using the ggH predictions from the POWHEG generator (blue) normalized to N^{3}LO. The subdominant component of the signal (VBFH + VH + ttH) is denoted as XH and is fixed to the SM prediction. The measured cross sections are also compared with the ggH predictions from NNLOPS (orange) and MadGraph5_aMC@NLO (pink). The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratios of the measured cross sections and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical predictions.

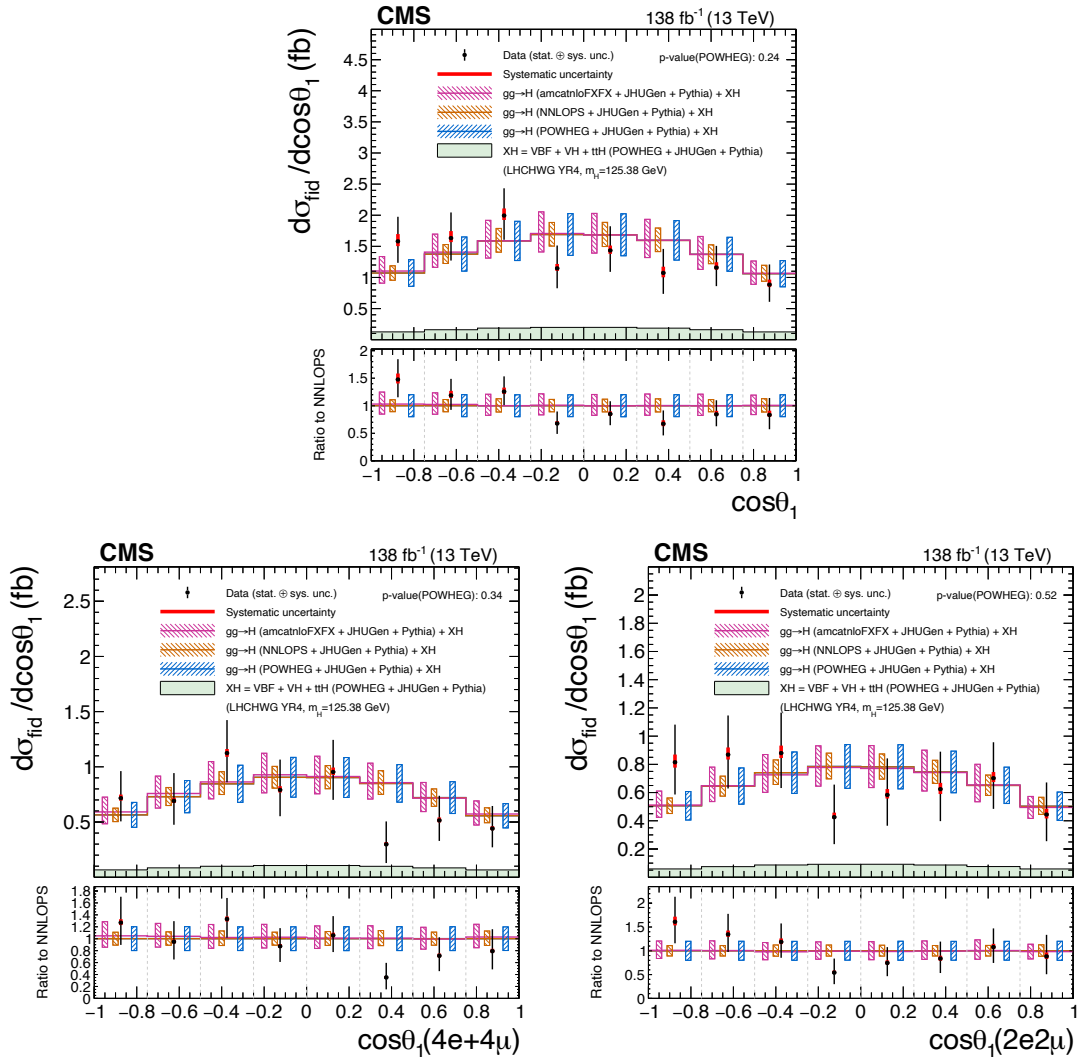

Figure 5.14: Differential cross sections as functions of $\cos\theta_1$ in the 4ℓ (upper) and in the same-flavor (lower left) and different-flavor (lower right) final states. The acceptance and theoretical uncertainties in the differential bins are calculated using the ggH predictions from the POWHEG generator (blue) normalized to N^{3}LO. The subdominant component of the signal (VBFH + VH + ttH) is denoted as XH and is fixed to the SM prediction. The measured cross sections are also compared with the ggH predictions from NNLOPS (orange) and MadGraph5_aMC@NLO (pink). The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratios of the measured cross sections and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical predictions.

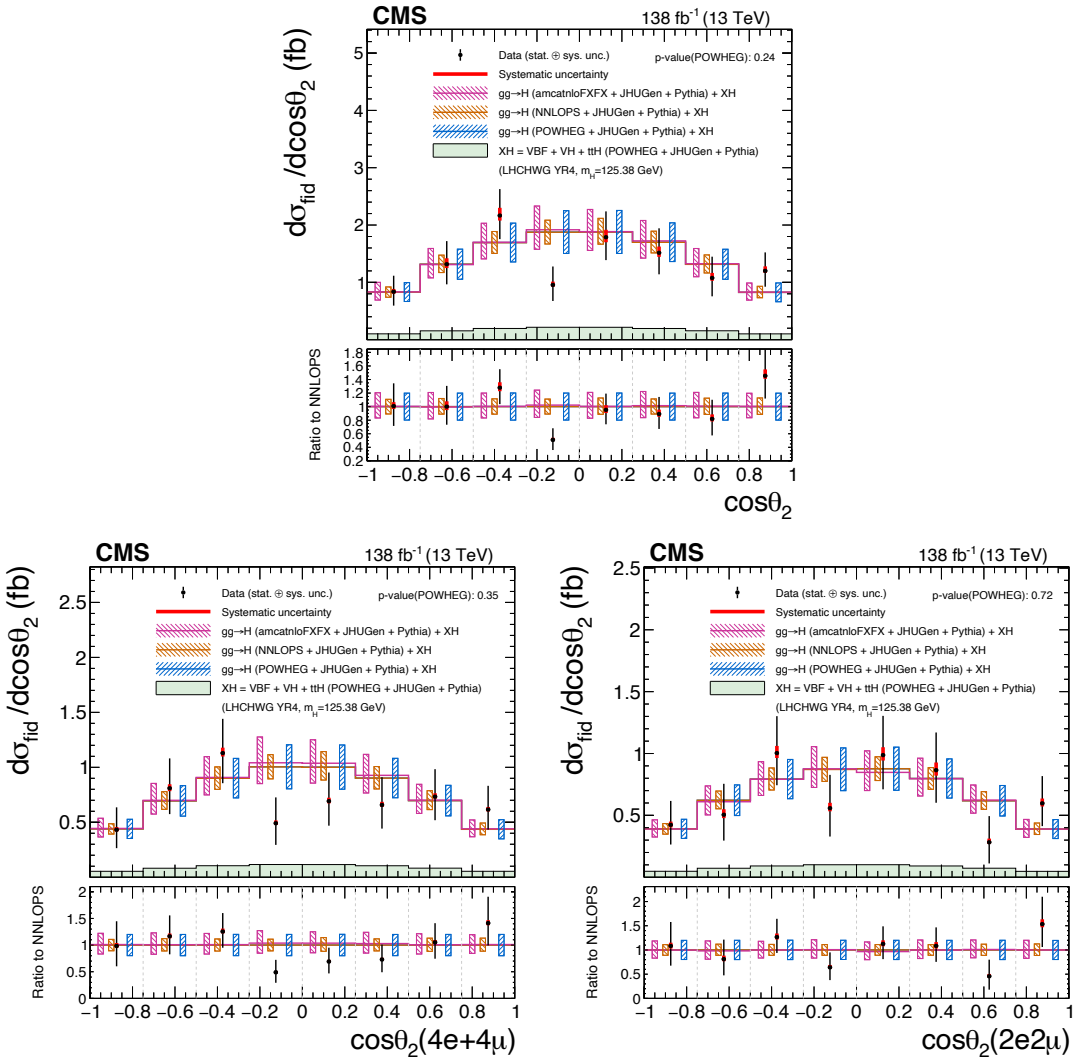

Figure 5.15: Differential cross sections as functions of $\cos\theta_2$ in the 4ℓ (upper) and in the same-flavor (lower left) and different-flavor (lower right) final states. The acceptance and theoretical uncertainties in the differential bins are calculated using the ggH predictions from the POWHEG generator (blue) normalized to N^{3}LO. The subdominant component of the signal (VBFH + VH + ttH) is denoted as XH and is fixed to the SM prediction. The measured cross sections are also compared with the ggH predictions from NNLOPS (orange) and MadGraph5_aMC@NLO (pink). The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratios of the measured cross sections and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical predictions.

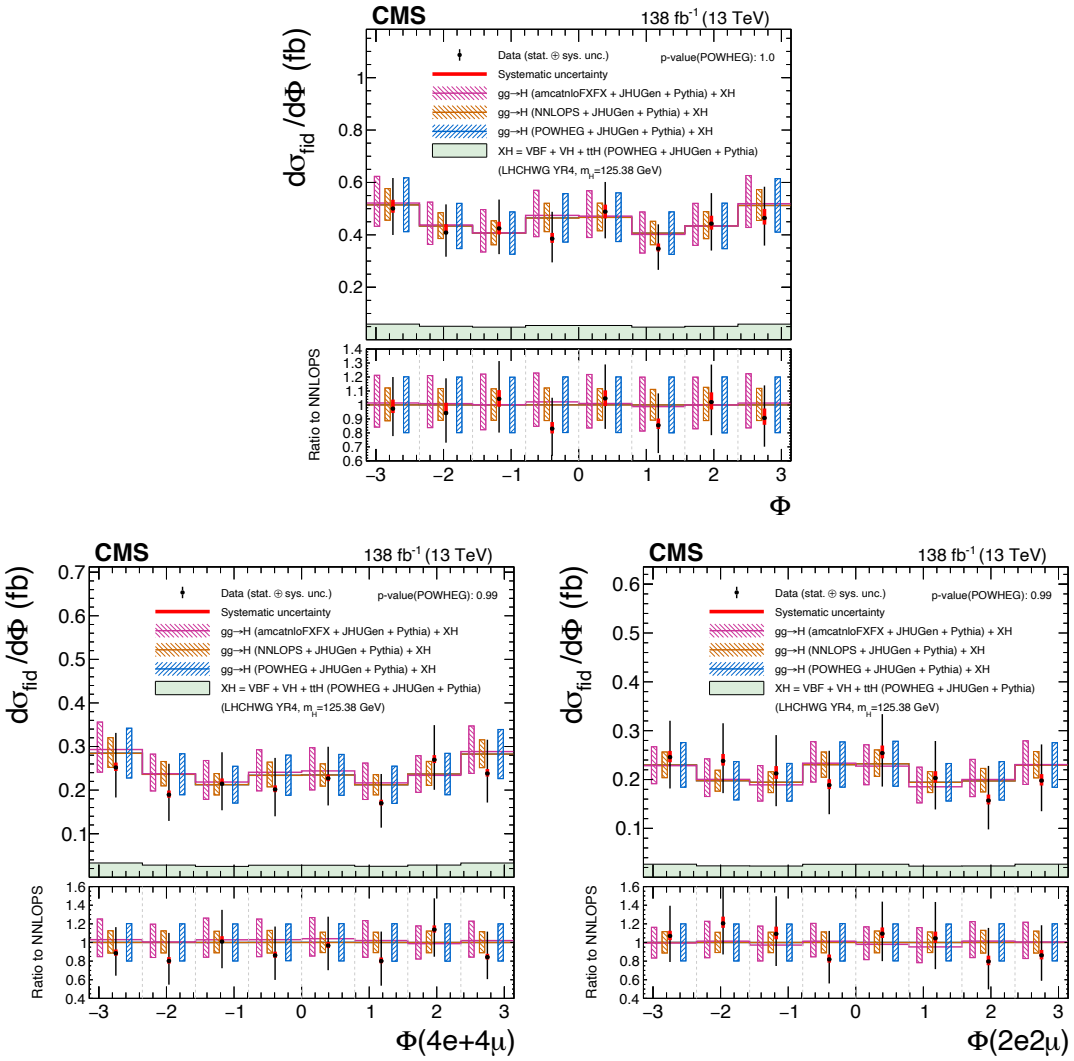

Figure 5.16: Differential cross sections as functions of the Φ angle in the 4ℓ (upper) and in the same-flavor (lower left) and different-flavor (lower right) final states. The acceptance and theoretical uncertainties in the differential bins are calculated using the ggH predictions from the POWHEG generator (blue) normalized to N^{3}LO. The subdominant component of the signal (VBFH + VH + ttH) is denoted as XH and is fixed to the SM prediction. The measured cross sections are also compared with the ggH predictions from NNLOPS (orange) and MadGraph5_aMC@NLO (pink). The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratios of the measured cross sections and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical predictions.

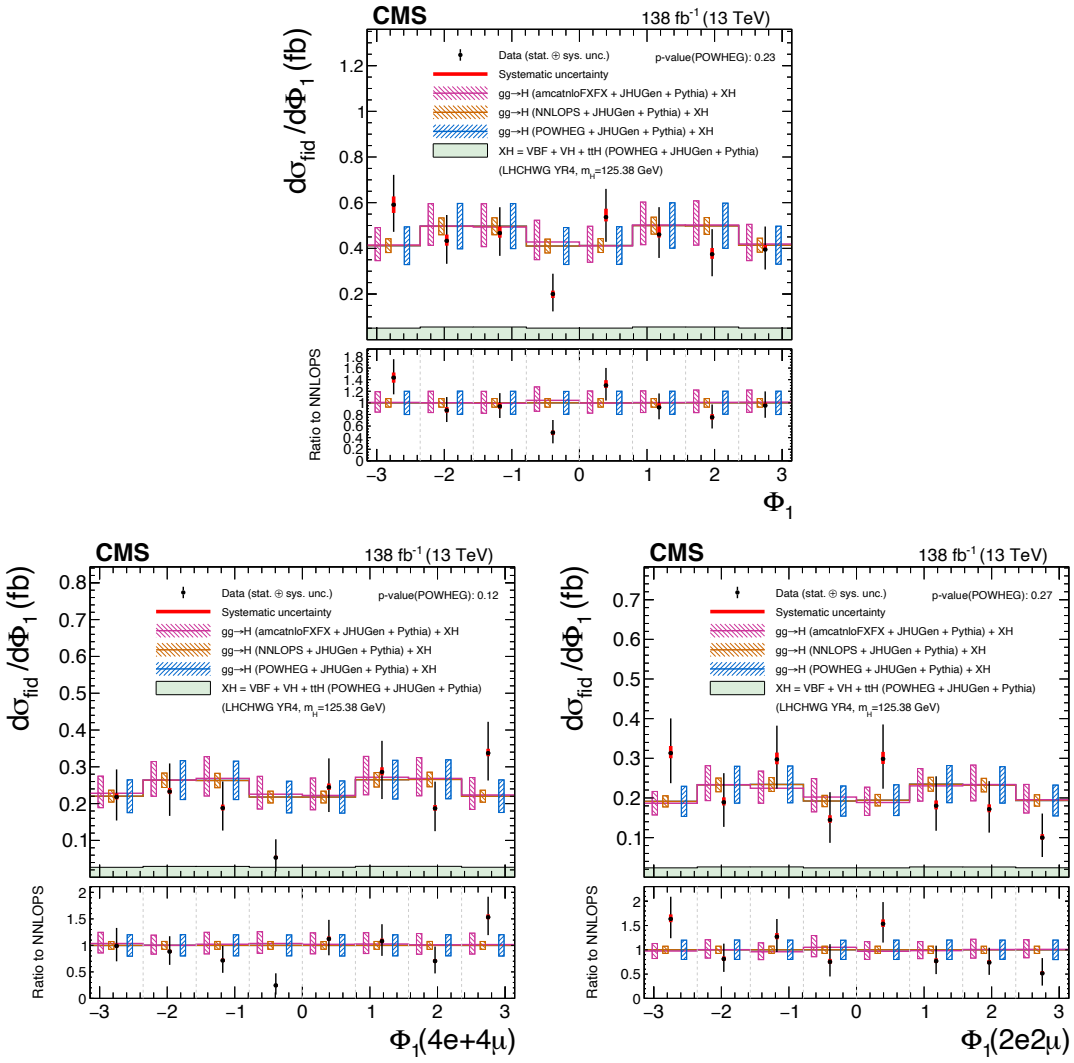

Figure 5.17: Differential cross sections as functions of the Φ_1 angle in the 4ℓ (upper) and in the same-flavor (lower left) and different-flavor (lower right) final states. The acceptance and theoretical uncertainties in the differential bins are calculated using the ggH predictions from the POWHEG generator (blue) normalized to N^{3}LO. The subdominant component of the signal (VBFH + VH + ttH) is denoted as XH and is fixed to the SM prediction. The measured cross sections are also compared with the ggH predictions from NNLOPS (orange) and MadGraph5_aMC@NLO (pink). The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratios of the measured cross sections and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical predictions.

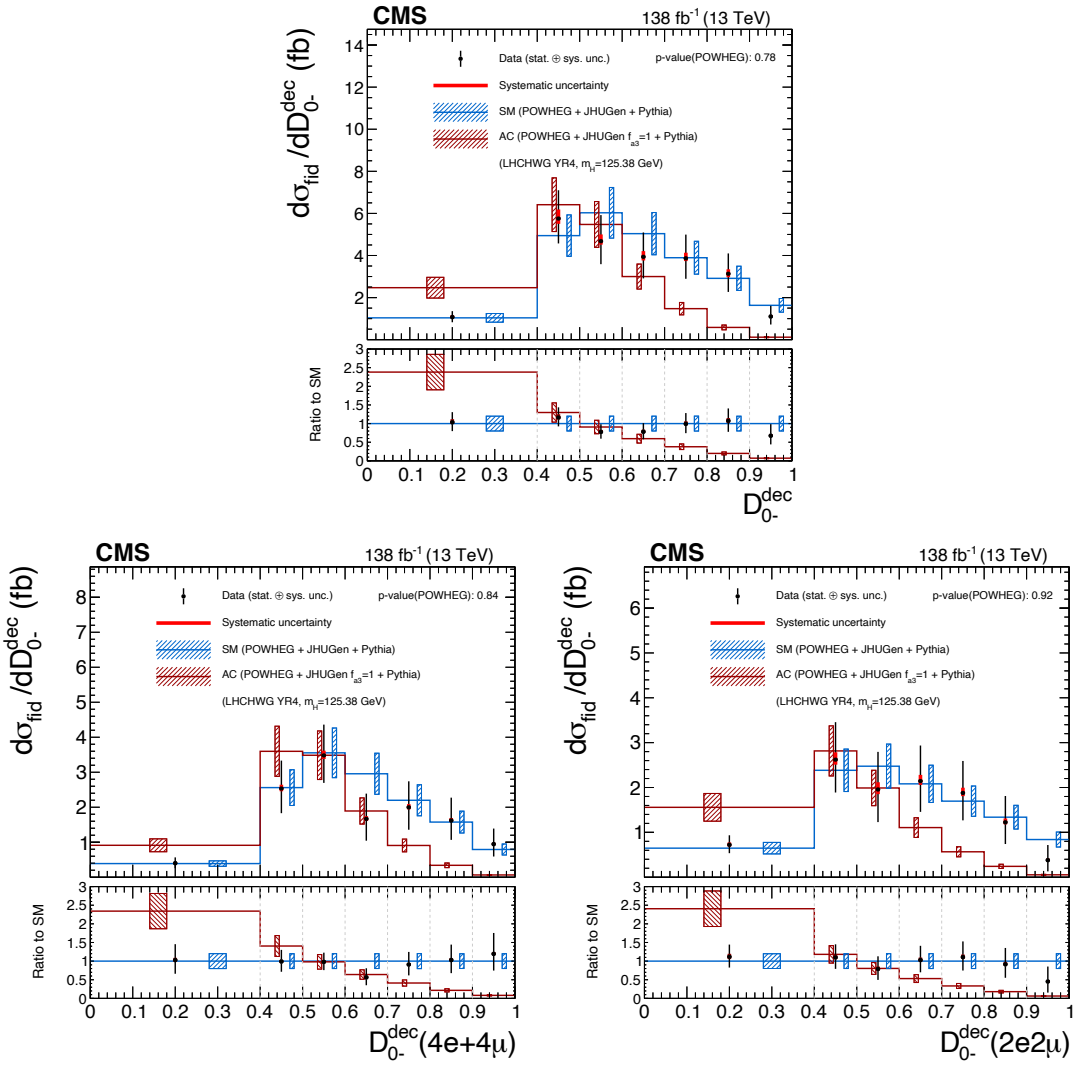

Figure 5.18: Differential cross sections as functions of the matrix element kinematic discriminant $\mathcal{D}_{0^-}^{\text{dec}}$ in the 4ℓ (upper) and in the same-flavor (lower left) and different-flavor (lower right) final states. The brown histograms show the distribution of the matrix element discriminant for the HVV anomalous coupling scenario corresponding to $f_{a3} = 1$. The subdominant component of the signal (VBFH + VH + ttH) is fixed to the SM prediction. The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratios of the measured cross sections and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical predictions.

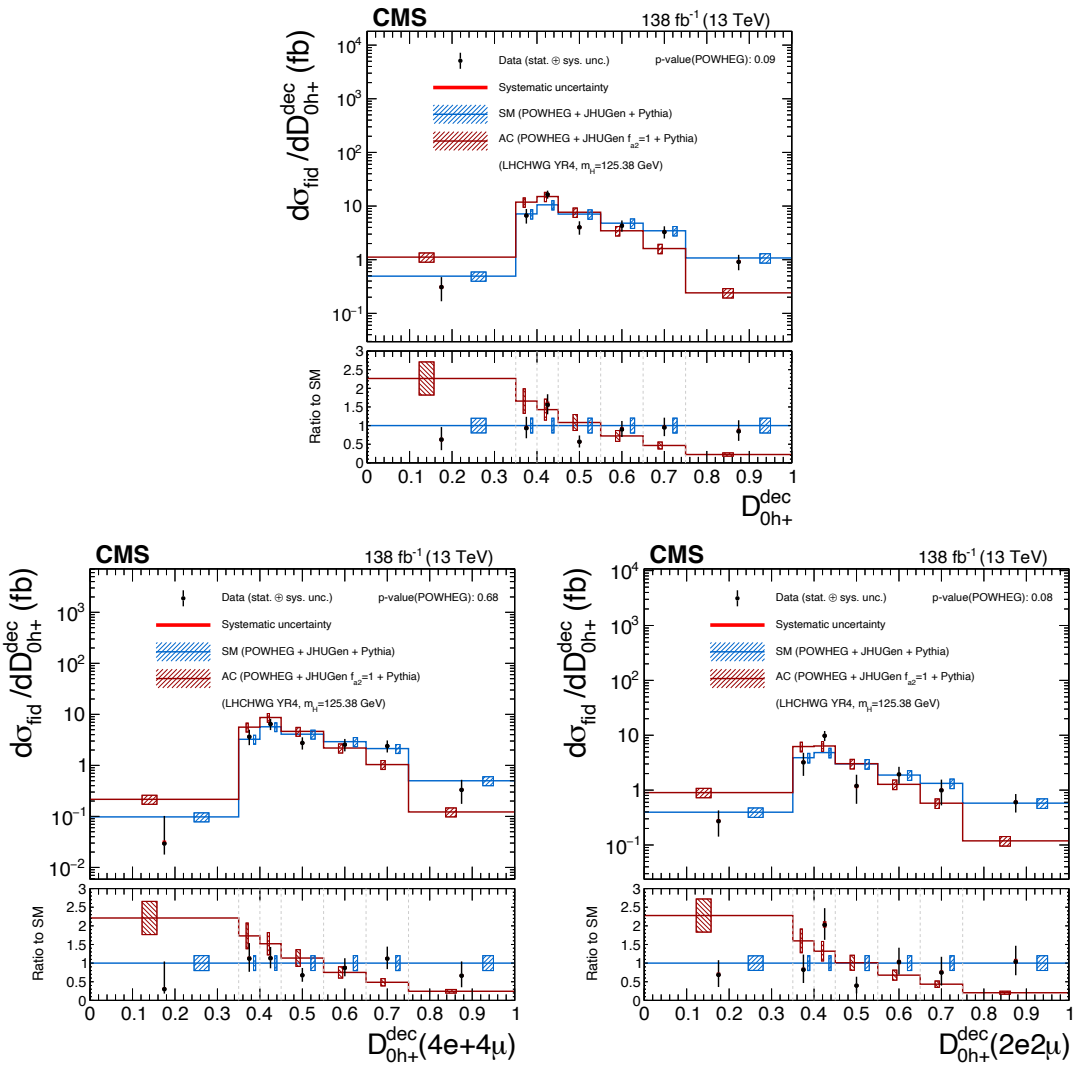

Figure 5.19: Differential cross sections as functions of the matrix element kinematic discriminant $\mathcal{D}^{\text{dec}}_{0\text{h}+}$ in the 4ℓ (upper) and in the same-flavor (lower left) and different-flavor (lower right) final states. The brown histograms show the distribution of the matrix element discriminant for the HVV anomalous coupling scenario corresponding to $f_{a2} = 1$. The subdominant component of the signal (VBFH + VH + ttH) is fixed to the SM prediction. The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratios of the measured cross sections and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical predictions.

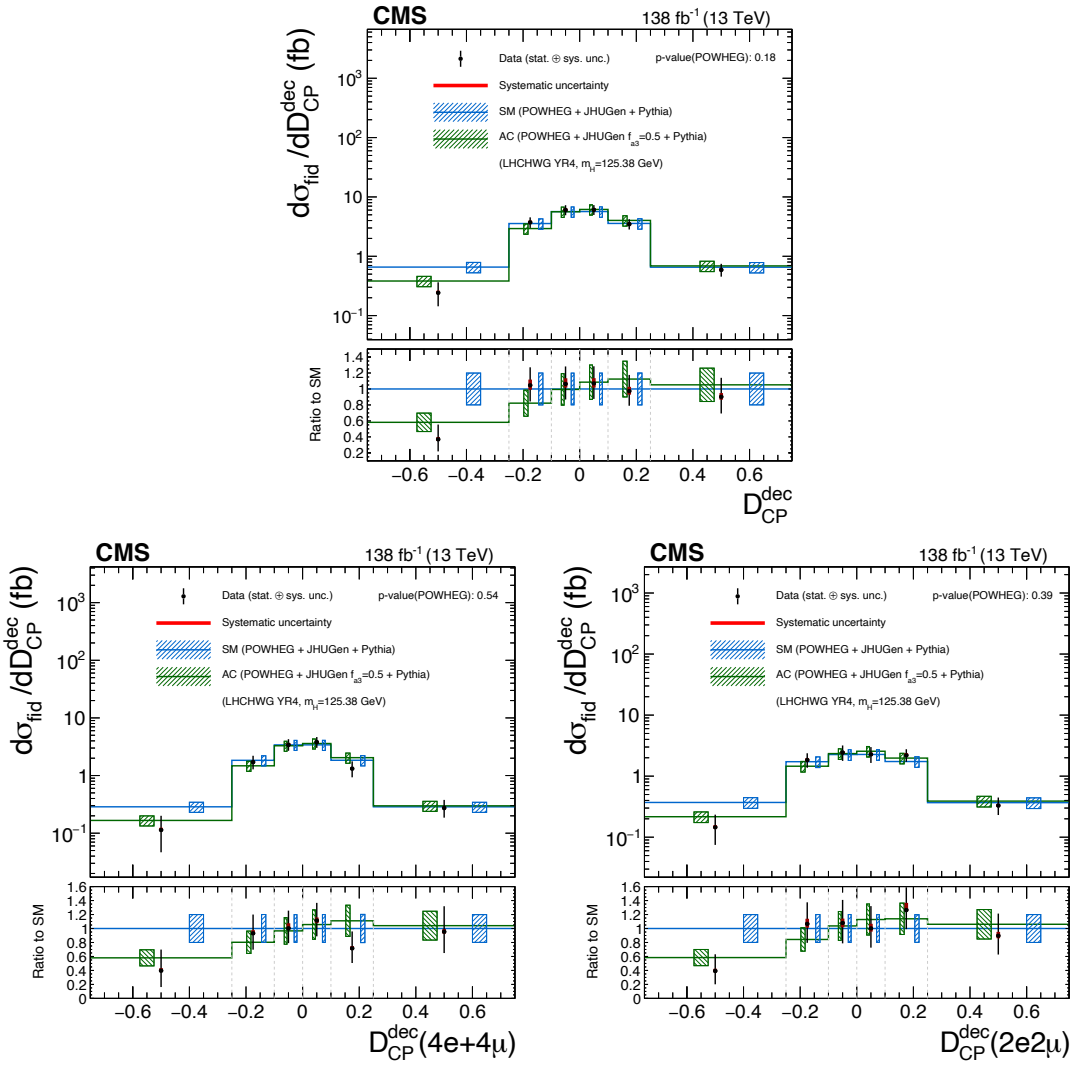

Figure 5.20: Differential cross sections as functions of the matrix element kinematic discriminant $\mathcal{D}_{\mathrm{CP}}^{\mathrm{dec}}$ in the 4ℓ (upper) and in the same-flavor (lower left) and different-flavor (lower right) final states. The green histogram shows the distribution of the discriminant for the HVV anomalous coupling scenario corresponding to $f_{a3} = 0.5$. The subdominant component of the signal (VBFH + VH + ttH) is fixed to the SM prediction. The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratios of the measured cross sections and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical predictions.

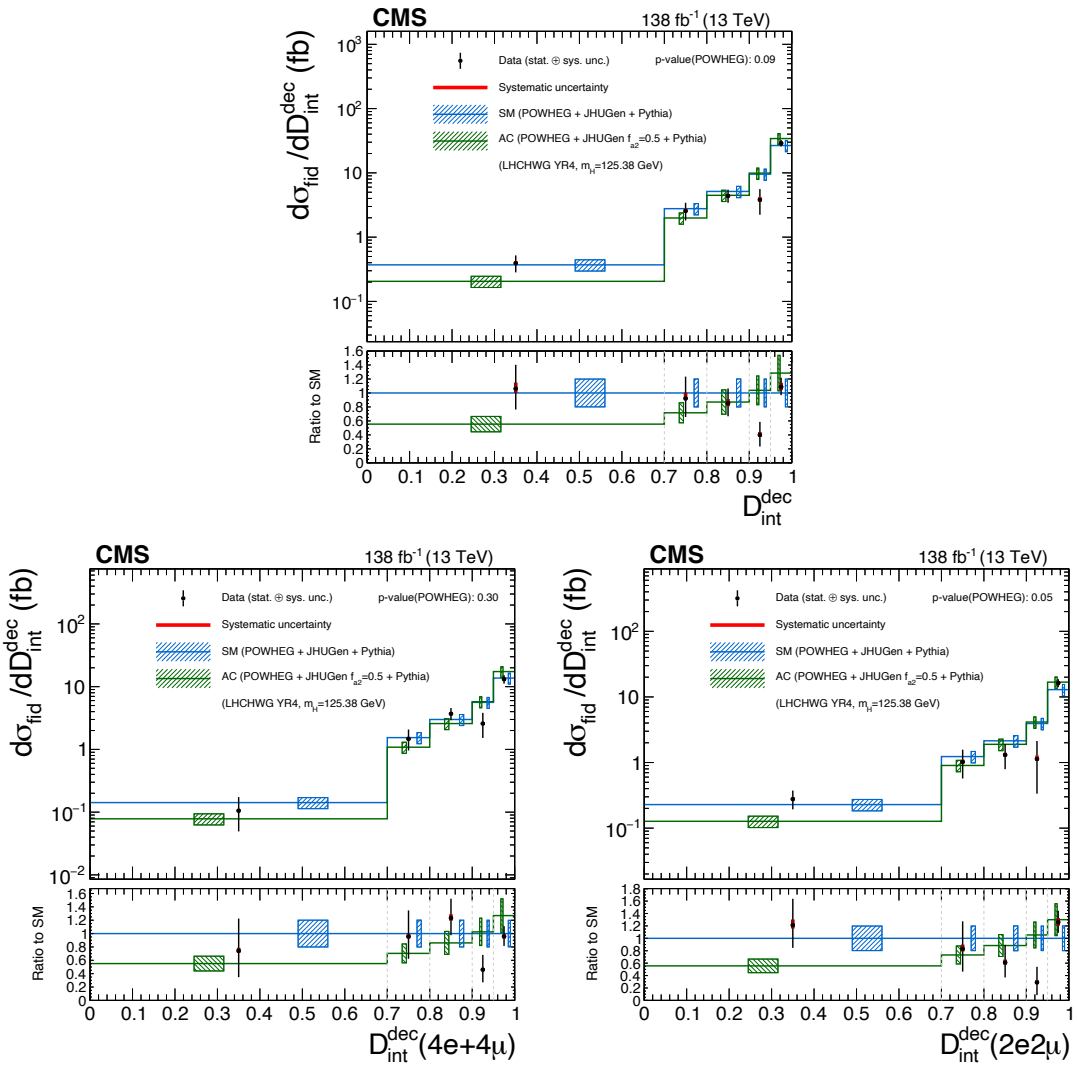

Figure 5.21: Differential cross sections as functions of the matrix element kinematic discriminant $\mathcal{D}_{\text{int}}^{\text{dec}}$ in the 4ℓ (upper) and in the same-flavor (lower left) and different-flavor (lower right) final states. The green histogram shows the distribution of the discriminant for the HVV anomalous coupling scenario corresponding to $f_{a2} = 0.5$. The subdominant component of the signal (VBFH + VH + ttH) is fixed to the SM prediction. The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratios of the measured cross sections and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical predictions.

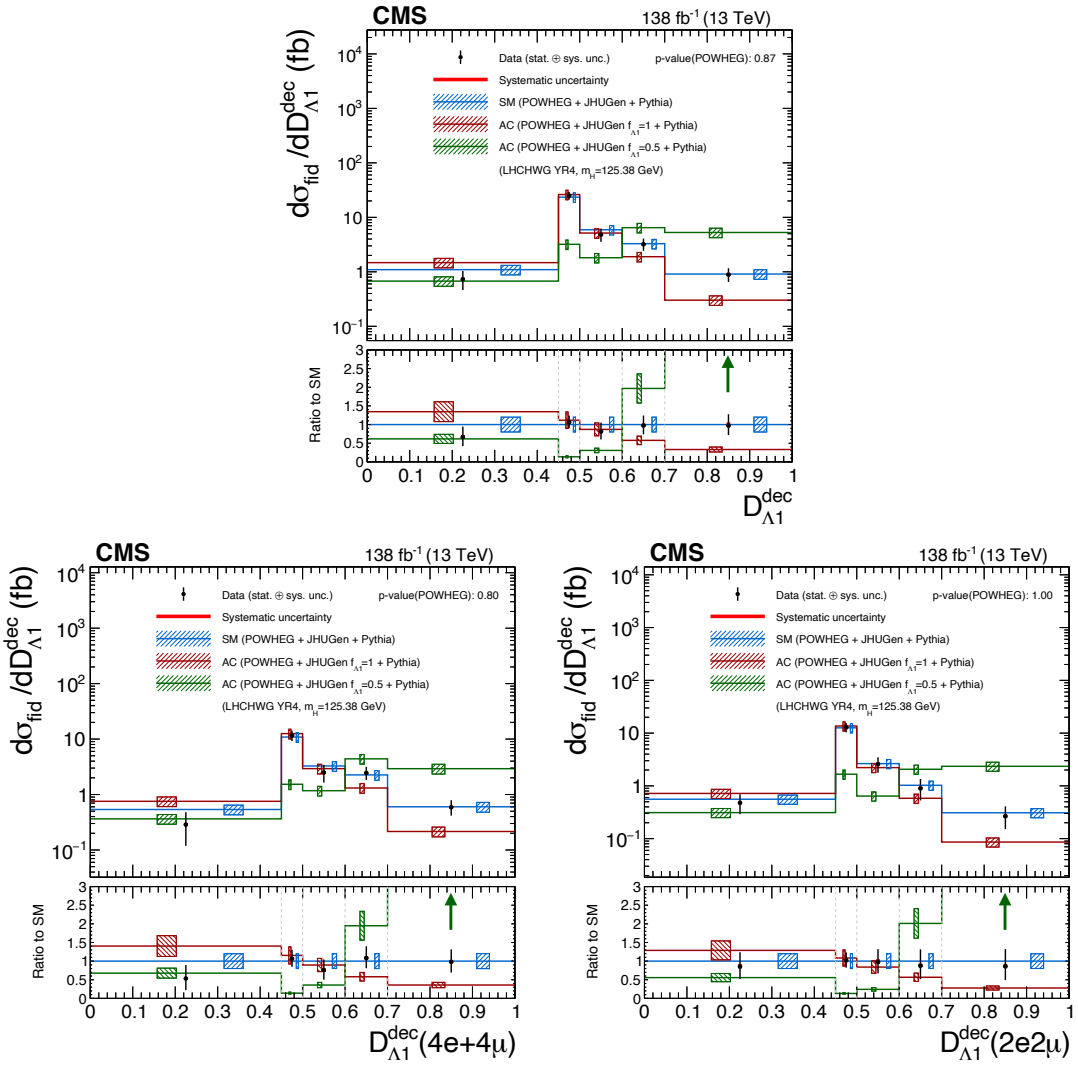

Figure 5.22: Differential cross sections as functions of the matrix element kinematic discriminant $\mathcal{D}_{\Lambda 1}^{\text{dec}}$ in the 4ℓ (upper) and in the same-flavor (lower left) and different-flavor (lower right) final states. The brown and green histograms show the distributions of the discriminant for the HVV anomalous coupling scenarios corresponding to $f_{\Lambda 1} = 1$ and $f_{\Lambda 1} = 0.5$. The subdominant component of the signal (VBFH + VH + ttH) is fixed to the SM prediction. The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratio of the measured cross section and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical expectation.

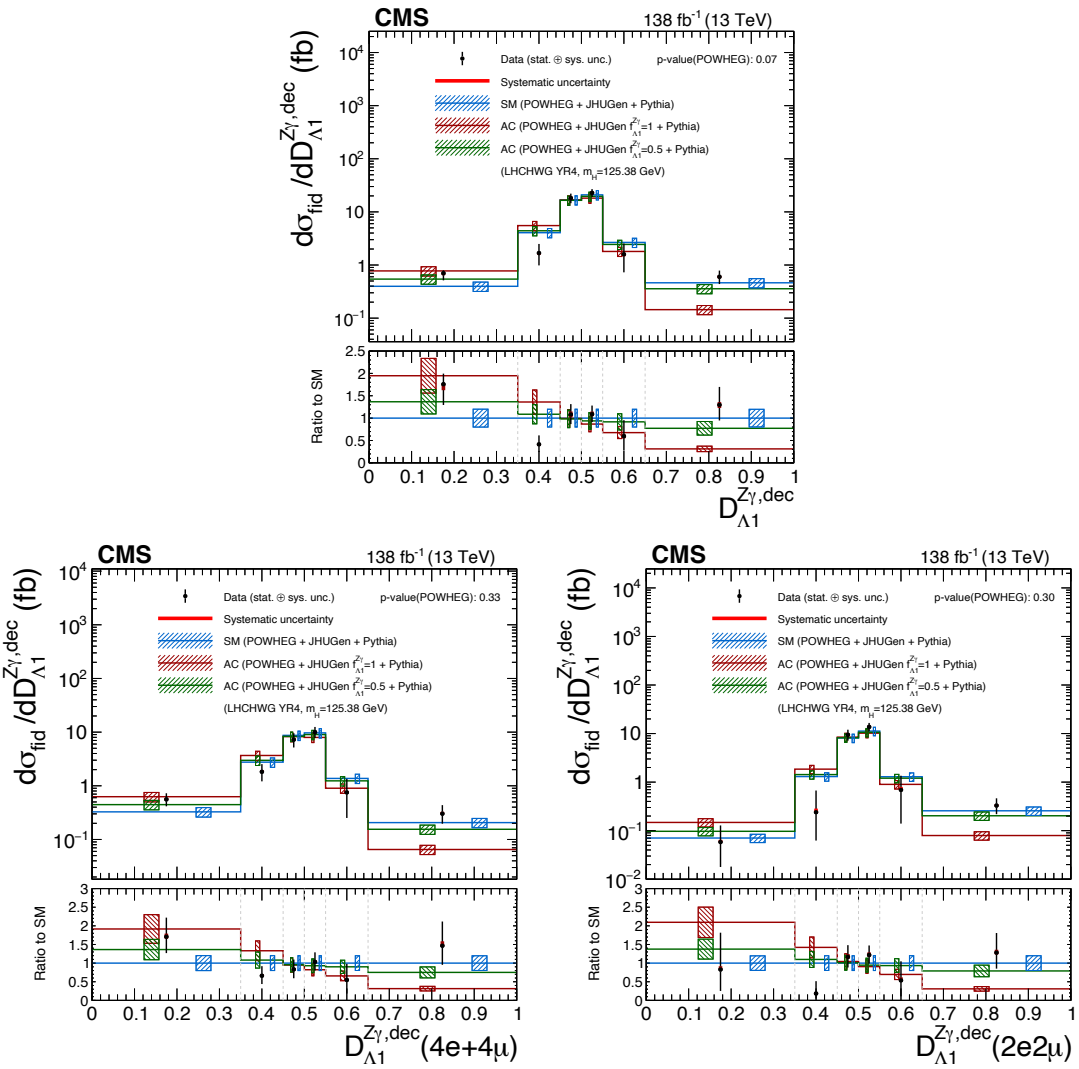

Figure 5.23: Differential cross sections as functions of the matrix element kinematic discriminant $\mathcal{D}_{\Lambda 1}^{Z\gamma,\mathrm{dec}}$ in the 4ℓ (upper) and in the same-flavor (lower left) and different-flavor (lower right) final states. The brown and green histograms show the distributions of the discriminant for the HVV anomalous coupling scenarios corresponding to $f_{\Lambda 1}^{Z\gamma} = 1$ and $f_{\Lambda 1}^{Z\gamma} = 0.5$. The subdominant component of the signal (VBFH + VH + ttH) is fixed to the SM prediction. The hatched areas correspond to the systematic uncertainties in the theoretical predictions. Black points represent the measured fiducial cross sections in each bin, black error bars the total uncertainty in each measurement, red boxes the systematic uncertainties. The lower panels display the ratios of the measured cross sections and of the predictions from POWHEG and MadGraph5_aMC@NLO to the NNLOPS theoretical predictions.

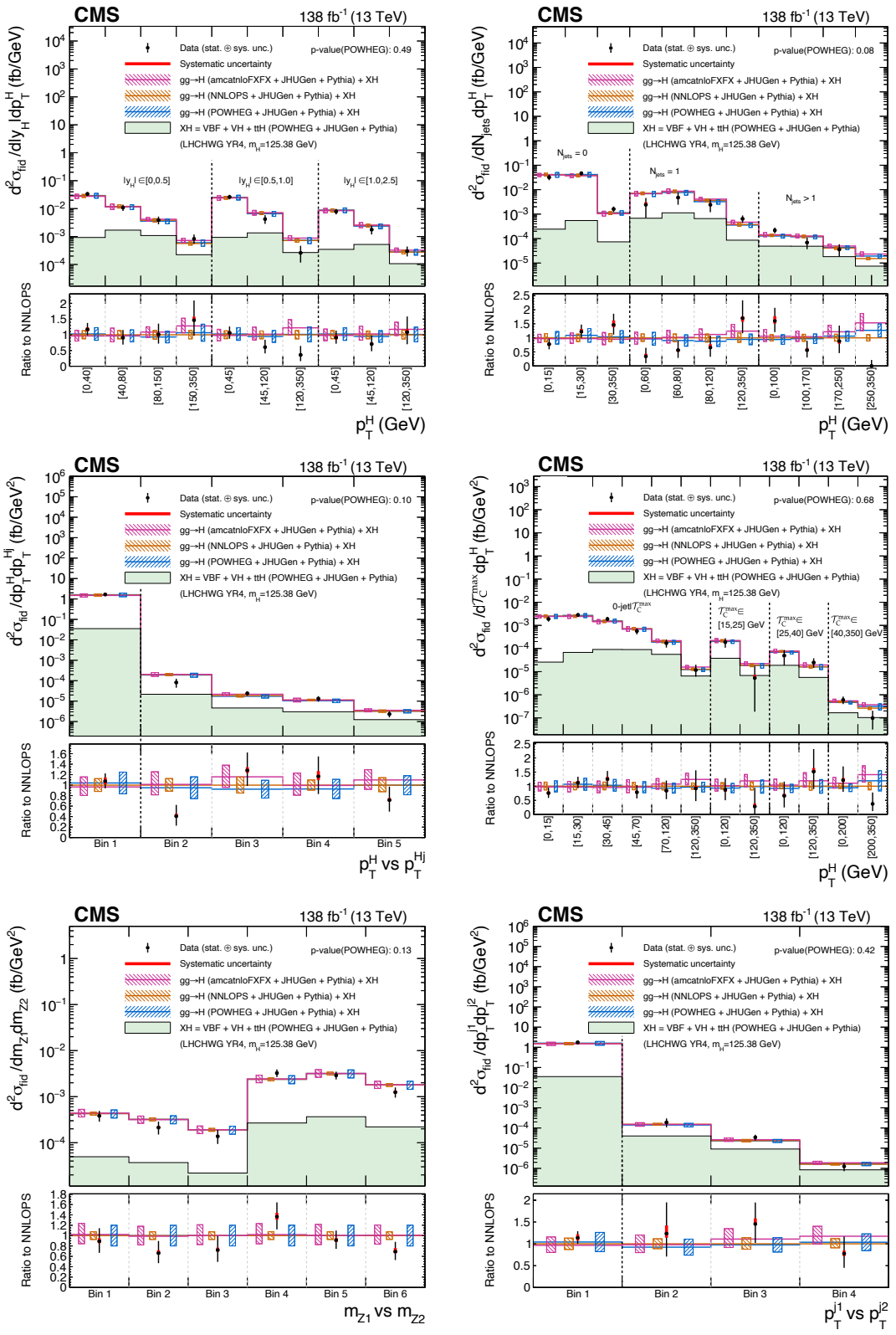

Figure 5.24: Double differential cross sections in bins of $|y_{\mathrm{H}}|$ vs. $p_{\mathrm{T}}^{\mathrm{H}}$ (top left), number of associated jets vs. $p_{\mathrm{T}}^{\mathrm{H}}$ (top right), $p_{\mathrm{T}}^{\mathrm{Hj}}$ vs. $p_{\mathrm{T}}^{\mathrm{H}}$ (centre left), $\mathcal{T}_{\mathrm{C}}^{\mathrm{max}}$ vs. $p_{\mathrm{T}}^{\mathrm{H}}$ (centre right), m_{Z_1} vs. m_{Z_2} (lower left), and p_{T} of the leading vs. subleading jet (lower right). The content of each plot is described in the caption of Fig. 5.6.

5.3 Higgs boson trilinear self-coupling

The measurement of the trilinear self-interaction of the Higgs boson λ_3 is of utmost priority in the LHC physics programme (cfr Sec. 1.2.3). At LO, the production modes and decay channels of the Higgs boson are insensitive to λ_3. The direct constraint at LO is only feasible by studying double-Higgs production, but this suffers from a significantly low cross section ($\mathcal{O}(1000)$ smaller than single-Higgs production). A complementary approach to probe λ_3 can be found in single-Higgs production and decay [49–51], which benefits from a larger cross section. Although the cross section and decay widths for single-Higgs production do not depend on λ_3 at LO EW or at any higher order in pQCD, they receive sizeable contributions at NLO EW due to radiative corrections. Some representative Feynman diagrams are shown in Fig. 5.25.

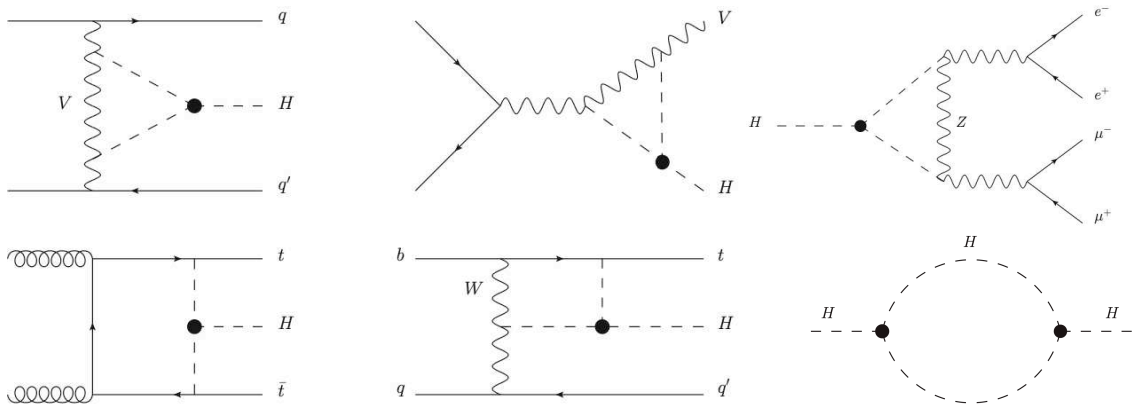

Figure 5.25: Representative NLO EW Feynman diagrams with Higgs trilinear coupling, representing the single-Higgs production (left and centre), the decay into four leptons (top right), and the one-loop λ_3-dependent diagram in the Higgs self-energy. Figures taken from [49, 50].

The parametrisation of λ_3-induced one-loop effects to single-Higgs processes relies on the assumption that BSM physics only affects the Higgs potential, modifying the Higgs boson self-coupling, and leaving the other couplings to the SM particles unchanged. This is equivalent to assuming that the BSM modifications of the Higgs boson couplings to other SM particles are negligible, which is justified for some cases by the tight experimental constraints of some of these couplings, and do not interfere to the NLO effects considered here. This is also the same assumption employed by the majority of constraints from the double-Higgs physics. It should also be noted that in a variety of specific BSM models, new physics appears at low energy only as modification of the λ_3, such as the model explained in Ref. [193]. With this assumption, the BSM effects can be parametrised by scaling λ_3 by a single parameter κ_λ:

$$\lambda_3 = \kappa_\lambda \lambda_3^{\mathrm{SM}} \tag{5.4}$$

The effects of an anomalous λ_3 impact both the shape and the normalisation of kinematic distribution. In this thesis, the transverse momentum of the Higgs boson is employed as it allows disentangling the effects of modified Higgs boson self-coupling values from other effects, such as deviations of other couplings. In order to exploit the fiducial cross sections to constrain κ_λ, the physics model defined in Sec. 4.7 should be modified accordingly. The fiducial cross section and BR are fixed to their SM expectation values and a scaling function μ_{ij} is introduced in each kinematic bin i and for each production mode j. This scaling function is then split into the production and decay component:

$$\mu_{ij} \to \mu_{ij}^{\mathrm{prod}}(\kappa_\lambda) \times \mu^{\mathrm{dec}}(\kappa_\lambda) \tag{5.5}$$

The production side is parametrised as:

$$\mu_{ij}^{\text{prod}}(\kappa_\lambda) = \frac{1 + \kappa_\lambda C_{1,ij} + \delta Z_{\text{H}}}{(1 - (\kappa_\lambda^2 - 1)\delta Z_{\text{H}})(1 + C_{1,ij} + \delta Z_{\text{H}})} \tag{5.6}$$

In this equation, δZ_{H} is a universal term, same for all processes, which is related to the corrections to the Higgs boson self-energy. Its value can be computed as:

$$\delta Z_{\text{H}} = -\frac{9}{16\sqrt{2}\pi^2}\left(\frac{2\pi}{3\sqrt{3}} - 1\right) G_\mu m_{\text{H}}^2 = -1.536 \times 10^{-3} \tag{5.7}$$

On the other hand, the $C_{1,ij}$ coefficients embed the process and kinematic dependence. The differential parametrisation for the leading ggH production mode is not yet available [49–51]. For this production mode, the inclusive value is therefore used in each kinematic bin. Conversely, in order to compute the scaling functions $\mu_{ij}^{\text{prod}}(\kappa_\lambda)$ for the other production modes, LO parton-level events are generated using `MADGRAPH5_aMC@NLO 2.5.5` and are reweighed on an event-by-event basis using a dedicated EW reweighing tool, which computes the corresponding NLO λ_3-corrections ($\mathcal{O}(\lambda_3)$). The ratio of the $\mathcal{O}(\lambda_3)$ to the LO distributions in bins of p_{T}^{H} is used to derive the scaling functions $\mu_{ij}^{\text{prod}}(\kappa_\lambda)$, as detailed in Ref. [50]:

$$C_{1,ij} = \frac{2\Re(\mathcal{A}_{\text{LO}}^* \mathcal{A}_{\text{NLO}}^{\lambda_3^{\text{SM}}})}{|\mathcal{A}_{\text{LO}}|^2} \tag{5.8}$$

where $\mathcal{A}$ denotes the corresponding matrix-element. The $C_{1,ij}$ values can be interpreted as the relative strength of the $\mathcal{O}(\lambda_3)$ corrections with respect to the LO. Since enough sensitivity in each bin is required, the p_{T}^{H} binning that is used for this interpretation is coarser than the corresponding fiducial spectrum measured before, and the bin boundaries, together with the corresponding $C_{1,ij}$, can be found in Tab. 5.2. The values of the $C_{1,ij}$ coefficients give an insight about the sensitivity of each production mode to κ_λ. The least sensitive mechanisms are the gluon and vector boson fusions, while the associated productions feature both a sizeable normalisation and shape effect, where ttH leads among all production modes.

The decay side is parametrised as:

$$\mu^{\text{dec}}(\kappa_\lambda) = 1 + \frac{(\kappa_\lambda - 1)(C_1^{\Gamma_{\text{ZZ}}} - C_1^{\Gamma_{\text{tot}}})}{1 + (\kappa_\lambda - 1)C_1^{\Gamma_{\text{tot}}}} \tag{5.9}$$

where $C_1^{\Gamma_{\text{ZZ}}} = 0.0083$ and $C_1^{\Gamma_{\text{tot}}} = 2.5 \times 10^{-3}$. The C_1 coefficients for the decay to four-lepton do not have any kinematic dependence. In summary, the scaling functions for both production and decay are reported in Fig. 5.26.

The constraint on κ_λ is extracted using the same test statistic employed for the measurement of the fiducial cross sections, with the difference that now the parameter of interest becomes κ_λ. As the parametrisation outlined above assumes that BSM physics manifests itself only as deviation of the Higgs trilinear coupling at the LHC energy, all the other couplings are set to the SM value of $\kappa_i = 1$. It was tested the possibility of extracting κ_λ while removing the SM hypothesis on the other couplings: the single $\text{H} \to \text{ZZ} \to 4\ell$ channel does not have enough power to constrain κ_λ within the validity range of the model. A scan of the profile negative log-likelihood as a function of κ_λ is shown in Fig. 5.27. The scan is performed in the range $-10 \leq \kappa_\lambda \leq 20$, as these are the limits of theoretical validity of the model. The minimum of the profiled likelihood corresponds to a measured value of:

$$\kappa_\lambda = 4.1^{+6.4}_{-5.9} = 4.1^{+6.1}_{-5.8}(\text{stat})^{+2.0}_{-1.2}(\text{syst}) \tag{5.10}$$

Table 5.2: Values of the $C_{1,ij}$ for the different production mode in the 5 $p_{\mathrm{T}}^{\mathrm{H}}$ kinematic bins. The values are reported in percentage. The ggH value is the same for all bins given the absence of the differential theoretical computation. The bin boundaries are reported in GeV.

$p_{\mathrm{T}}^{\mathrm{H}}$ bin	ggH	VBFH	VH	ttH
0 - 45	0.66	0.66	1.66	5.31
45 - 80	0.66	0.66	1.23	4.73
80 - 120	0.66	0.64	0.77	3.92
120 - 200	0.66	0.58	0.35	2.79
200 - ∞	0.66	0.55	0.02	1.42

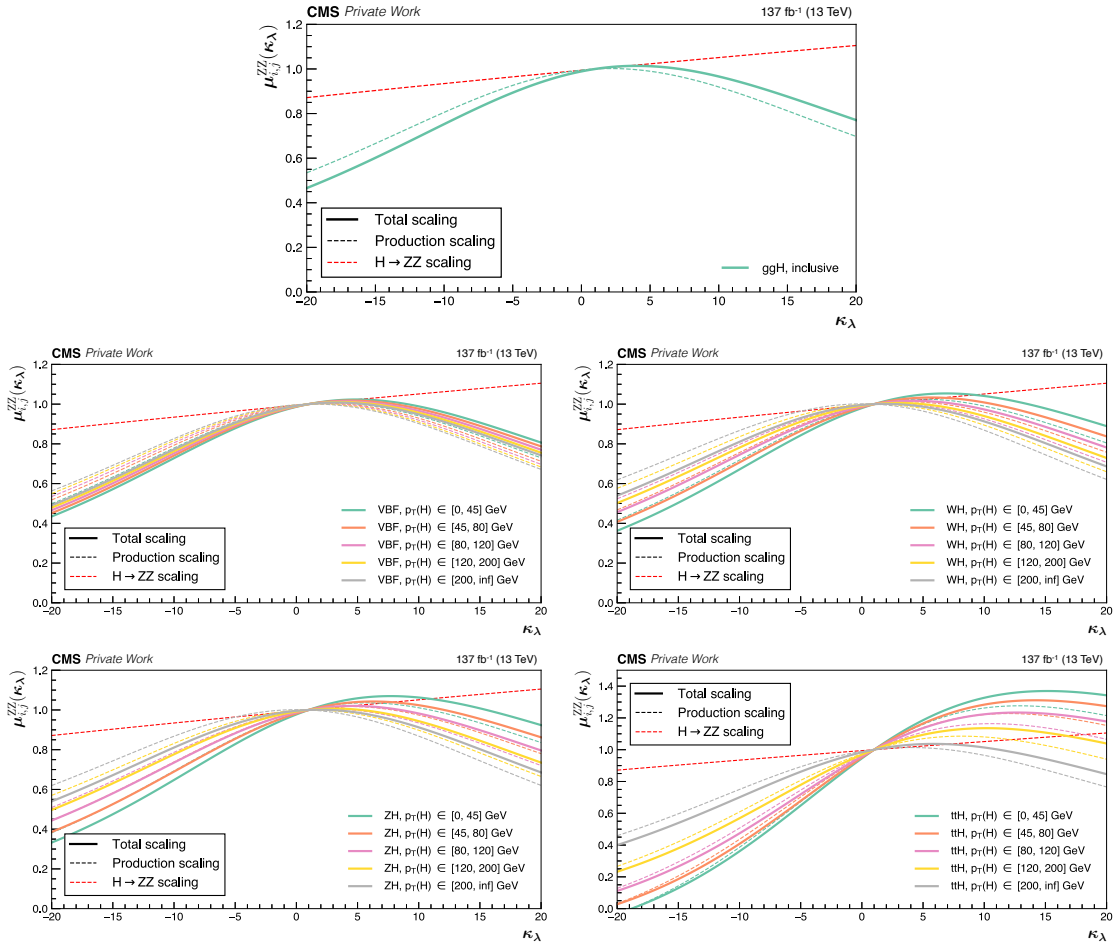

Figure 5.26: Scaling functions $\mu_{ij}(\kappa_\lambda)$ for the different Higgs boson production modes: ggH (top), VBFH (centre left), WH (centre right), ZH (bottom left), and ttH (bottom right). The coloured dashed lines represents the signal strength parametrisation in each $p_{\mathrm{T}}^{\mathrm{H}}$ bin. The red dashed line in each plot represents the scaling of the Higgs boson decay to two Z bosons, which is independent on the $p_{\mathrm{T}}^{\mathrm{H}}$ bin. The coloured continuous line show the combination of the production and decay scaling in each $p_{\mathrm{T}}^{\mathrm{H}}$ bin.

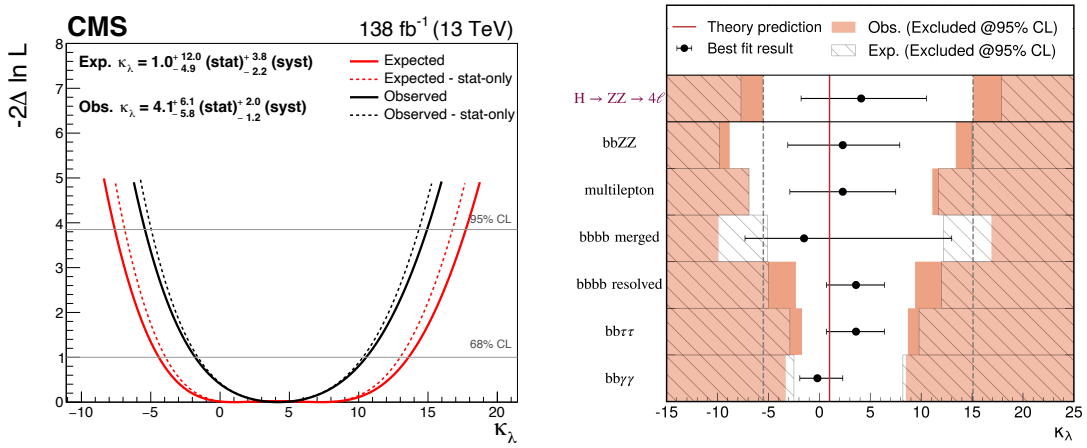

Figure 5.27: (Left) Likelihood scan as a function of κ_λ. The scan is shown with (solid line) and without (dashed line) systematic uncertainties profiled in the fit. (Right) Comparison of the H $\to$ ZZ $\to 4\ell$ result with the direct constraints from the most sensitive double-Higgs searches in the CMS collaboration. The value for the Higgs boson pair production in the four leptons plus two b jets final states (bbZZ) is taken from Ref. [194]. The value for Higgs boson pairs decaying to WWWW, WW$\tau\tau$, and $\tau\tau\tau\tau$ (multilepton) is taken from Ref. [195]. The value for pair production of highly energetic Higgs bosons decaying to bottom quarks (bbb merged) is taken from Ref. [196]. The value for Higgs boson pair production in the four bottom quark final state (bbbb resolved) is taken from Ref. [197]. The value for Higgs boson pair production in final states with two bottom quarks and two tau leptons (bb$\tau\tau$) is taken from Ref. [198]. The value for Higgs boson pair production in final states with two bottom quarks and two photons (bb$\gamma\gamma$) is taken from Ref. [199].

for an expected value of:

$$\kappa_\lambda = 1.0^{+12.6}_{-5.4} = 1.0^{+12.0}_{-4.9}(\text{stat})^{+3.8}_{-2.2}(\text{syst}). \tag{5.11}$$

where this value is extracted with an Asimov dataset generated with $\kappa_\lambda = 1$. The corresponding observed (expected) excluded κ_λ range at the 95% confidence level is:

$$-5.4(-7.6) < \kappa_\lambda < 14.9(17.7). \tag{5.12}$$

Figure 5.28 shows the likelihood scan for the different production modes. This plot is obtained by setting to 1 the κ_λ scaling functions of the production mechanisms not under investigation. Since the leading production mode in the H $\to$ ZZ $\to 4\ell$ channel is the gluon fusion, even though it is not the most sensitive to the κ_λ effects, it drives the sensitivity. Despite being the rarest production mode considered in this analysis, the constraint power of ttH is the second in importance in the regions for $\kappa_\lambda > 1$. With the available statistics, it is not yet feasible a categorisation to isolate the ttH contribution that could bring an improvement in the final constraint. Figure 5.28 also shows the likelihood scan for the production and decay independently. For κ_λ values greater than 1, the constraint from production-only is tighter than when the full information is used. The explanation can be found in Fig. 5.26. For $\kappa_\lambda > 1$ the scaling for production and decay has opposite effects, it is increasing for the decay, and decreasing for the production.

The current best available constraints on κ_λ are obtained from the combination of measurements of Higgs boson pair production performed with the Run 2 dataset. The limits set by the ATLAS and CMS Collaborations correspond to observed limits at the 95% confidence level of $-0.6 < \kappa_\lambda < 6.6$ [60] and $-1.24 < \kappa_\lambda < 6.49$ [41], respectively.

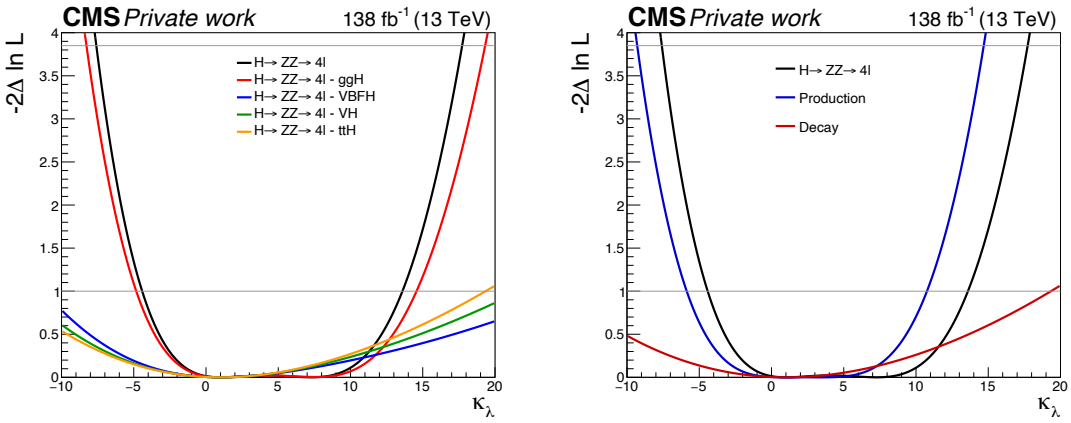

Figure 5.28: Expected likelihood scan as a function of κ_λ obtained for each production mode (left) and for production with an Asimov dataset generated in the $\kappa_\lambda = 1$.

The limit presented in this chapter is obviously much looser, but it is the first time that κ_λ is constrained with fiducial cross sections of single-Higgs production. The right plot in Fig. 5.27 reports the comparison of the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ result with the most sensitive double-Higgs analysis in the CMS collaboration. Despite the weaker constraint with respect to the silver bullets of the double-Higgs physics, namely $\mathrm{bb}\gamma\gamma$, $\mathrm{bb}\tau\tau$, and bbbb, the four-lepton result is competitive with other direct searches, like bbZZ and multileptonic final states. This demonstrates that the single-Higgs physics can help the double-Higgs physics in the effort to constrain the Higgs trilinear coupling. However, the path to measuring this fundamental parameter of the SM does not lie in relying solely on either double-Higgs researches or indirect limits from single-Higgs physics; rather, the keyword is *combination*. By combining different decay channels as well as HH and H results, the maximum amount of information can be extracted from data. Each analysis may possess varying sensitivities, and through this combination, the utmost knowledge can be extracted. On one side, the H channels are more sensitive where the destructive interference in HH production is more pronounced. On the other side, the combination plays a key role when more parameters are allowed to float, as highlighted in [60].

5.4 Higgs boson couplings to charm and bottom quarks

Analogously to what is done in the previous section, the $p_{\mathrm{T}}^{\mathrm{H}}$ spectrum can be interpreted to constrain the couplings of the Higgs boson to charm and bottom quarks. The measurement is performed in the κ-framework (cfr Sec. 1.2.4) by measuring $\kappa_b = y_b/y_b^{SM}$ and $\kappa_c = y_c/y_c^{SM}$, where y_i is the corresponding Yukawa coupling. The $p_{\mathrm{T}}^{\mathrm{H}}$ spectrum of the ggH production mode has been calculated for simultaneous variations of κ_b and κ_c [180], providing the theoretical inputs needed for this experimental measurement.

In this context, since the Higgs boson production cross section is parametrised as a function of its coupling to the bottom quarks, the related associated production (bbH) can play a significant role. In the SM, the bbH cross section is computed to be $\sigma_{\mathrm{bbH}} = 4.88^{+20.2\%}_{-23.9\%}$ for a Higgs boson mass of $m_{\mathrm{H}} = 125$ GeV, which is far smaller than the ggH cross section of $\sigma_{\mathrm{ggH}} = 48.58^{+4.6\%}_{-6.7\%}$. However, the σ_{bbH} scales quadratically with the value of κ_b, and for certain values of the coupling it may become the leading production mode. The theoretical inputs accounts properly for this effect.

Following the parametrisation in the κ-framework, the ggH cross section can be expressed as a quadratic polynomial in terms of the couplings modifiers entering the gluon loop. Supposing that the sizeable contributions are given by the top, bottom, and charm

quarks, the cross section can be parametrised as:

$$\sigma_{\mathrm{ggH}} = A\kappa_b^2 + B\kappa_c^2 + C\kappa_b\kappa_c + D\kappa_t\kappa_b + F\kappa_t\kappa_c + F\kappa_t^2. \tag{5.13}$$

In the context of this interpretation the Higgs boson coupling to the top quark is considered SM-like ($\kappa_t = 1$):

$$\sigma_{\mathrm{ggH}} = A\kappa_b^2 + B\kappa_c^2 + C\kappa_b\kappa_c + D\kappa_b + F\kappa_c + F. \tag{5.14}$$

If the cross section is known for at least six points in the (κ_b,κ_c) phase space, the six coefficients parametrising the cross section can be computed by a simple matrix inversion:

$$\begin{bmatrix} A \\ B \\ C \\ D \\ E \\ F \end{bmatrix} = \begin{bmatrix} \kappa_{b,1}^2 & \kappa_{c,1}^2 & \kappa_{b,1}\kappa_{c,1} & \kappa_{b,1} & \kappa_{c,1} \\ \kappa_{b,2}^2 & \kappa_{c,2}^2 & \kappa_{b,2}\kappa_{c,2} & \kappa_{b,2} & \kappa_{c,2} \\ \kappa_{b,3}^2 & \kappa_{c,3}^2 & \kappa_{b,3}\kappa_{c,3} & \kappa_{b,3} & \kappa_{c,3} \\ \kappa_{b,4}^2 & \kappa_{c,4}^2 & \kappa_{b,4}\kappa_{c,4} & \kappa_{b,4} & \kappa_{c,4} \\ \kappa_{b,5}^2 & \kappa_{c,5}^2 & \kappa_{b,5}\kappa_{c,5} & \kappa_{b,5} & \kappa_{c,5} \\ \kappa_{b,6}^2 & \kappa_{c,6}^2 & \kappa_{b,6}\kappa_{c,6} & \kappa_{b,6} & \kappa_{c,6} \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{\mathrm{ggH}}^1 \\ \sigma_{\mathrm{ggH}}^2 \\ \sigma_{\mathrm{ggH}}^3 \\ \sigma_{\mathrm{ggH}}^4 \\ \sigma_{\mathrm{ggH}}^5 \\ \sigma_{\mathrm{ggH}}^6 \end{bmatrix}$$

Therefore, using six points from the theoretical inputs, the cross section is parametrised a function of κ_c and κ_b in each kinematic bin. The effects of these variations are represented in Fig. 5.29.

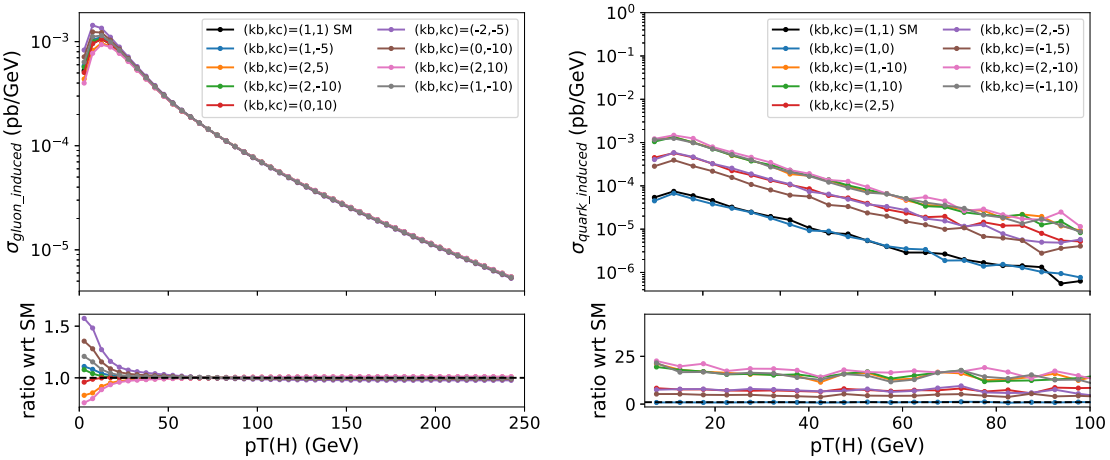

Figure 5.29: Spectrum of the $p_{\mathrm{T}}^{\mathrm{H}}$ for representative values of κ_b and κ_c. The dots depict the theoretical computations and the line are simply segments connecting the points corresponding to the same values of coupling modifiers. The theoretical cross sections are computed separately for the gluon-induced (left) and quark-induced (right) contributions.

The decay side also plays a role in the constrain of κ_b and κ_c. The result strongly depends on the assumption one makes on the BR. The scenario where the BR is fixed to the SM is not taken into account. Assuming only coupling dependence on the production side, and leaving unaltered the decay side, is not physically motivated. Two other scenarios are left.

- The BR is dependent on κ_b and κ_c, implementing the parametrisation of the κ framework, without assuming BSM contributions. The constraining power in this scenario mainly comes from the normalisation, as the contribution of some decay channels can quickly saturate the total width of the Higgs boson.

- The BR is a complete freely floating parameter in the fit. The constraint from the total width and the overall normalisation is removed, and what remains is purely the constraint obtained from the shape of the $p_{\mathrm{T}}^{\mathrm{H}}$ spectrum.

The statistical approach is always based on the test statistic of Sec. 4.4, where now the parameters of interest are κ_b and κ_c. As the parametrisation is provided only for ggH, the other production modes are fixed to the SM expectations and treated as a background. Confidence intervals on κ_b and κ_c are obtained from the scan of the profile likelihood leaving one of the two parameters unconstrained in the fit and scanning the other. The observed (expected) exclusion limits at the 95% CL are:

$$-1.1\,(-1.3) < \kappa_b < 1.1\,(1.2) \tag{5.15}$$
$$-5.3\,(-5.7) < \kappa_c < 5.2\,(5.7), \tag{5.16}$$

assuming a dependence of the BR on κ_b and κ_c, and:

$$-5.6\,(-5.5) < \kappa_b < 8.9\,(7.4) \tag{5.17}$$
$$-20\,(-19) < \kappa_c < 23\,(20)\,, \tag{5.18}$$

treating the BR as an unconstrained parameter in the fit. A simultaneous constraint on κ_b and κ_c in the two scenarios is shown in Fig. 5.30. These results can be compared to the recent direct constraint of κ_c from the H → cc search [200], which sets the observed (expected) 95% CL interval to $1.1 < |\kappa_c| < 5.5$ ($|\kappa_c| < 3.4$).

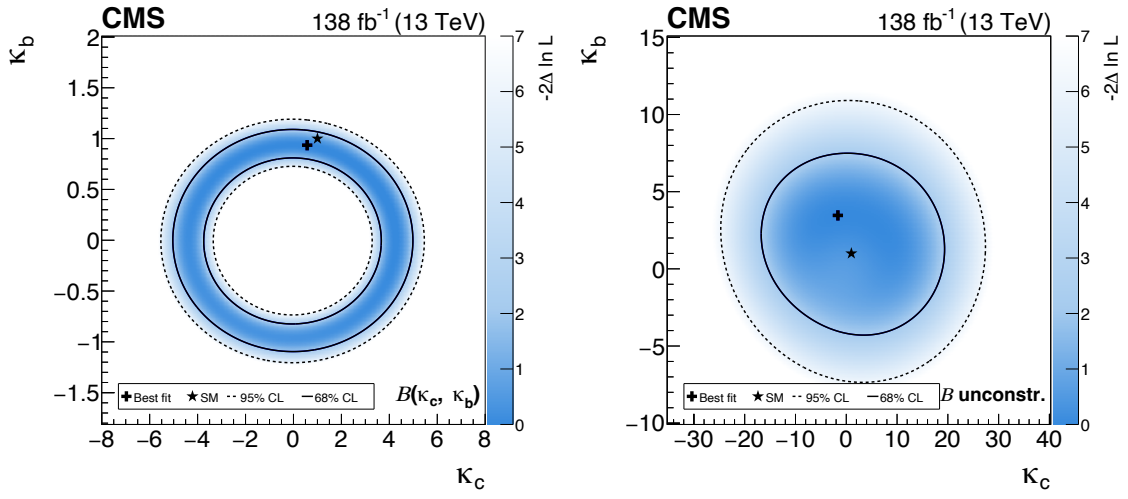

Figure 5.30: Simultaneous fit of κ_b and κ_c, assuming a coupling dependence of the branching ratio $\mathcal{B}$ (left) and treating it as an unconstrained parameter in the fit (right).

Summary

A comprehensive characterisation of the H → ZZ → 4ℓ decay channel via the measurements of fiducial cross sections has been presented. This analysis is performed exploiting proton-proton collisions at $\sqrt{s} = 13$ TeV, corresponding to an integrated and validated luminosity of 138 fb^{-1} collected by the CMS experiment during the 2016, 2017, and 2018 data-taking periods of the Run 2 of the LHC. With 60 measured fiducial cross sections of 32 differential observables (28 of them never shown in previous H → ZZ → 4ℓ CMS analyses), which give a deeper insight in the properties of the Higgs boson, and 3 interpretations of the $p_{\mathrm{T}}^{\mathrm{H}}$ spectrum, this is one of the most extensive fiducial analyses ever performed. The

Higgs boson production is characterized via measurements of differential cross sections in bins of $p_{\mathrm{T}}^{\mathrm{H}}$ and $|y_{\mathrm{H}}|$, the p_{T} of the leading and subleading jets and observables of the di-jet system, when associated with jets. For the first time in CMS, fiducial cross sections are measured in bins of the seven kinematic observables that completely define the four-lepton decay: the invariant mass of the two Z bosons and the five angles that describe the fermions kinematic properties and the production and decay planes. Differential cross sections are also measured for the first time in bins of six matrix element kinematic discriminants sensitive to various anomalous couplings of the Higgs boson to vector bosons, probing the tensor structure of the scattering amplitude. The dynamical evolution of the renormalisation and factorization scales, and resummation effects are probed by measuring cross sections in bins of rapidity-weighted jet vetoes, and in bins of observables of the Higgs boson plus jets system. An extensive set of double-differential measurements is presented, providing a complete coverage of the phase space under study. This analysis profit of an improved object calibration and corresponding corrections, leading to a 40% reduction of the systematic uncertainty. The $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ inclusive fiducial cross section is $\sigma_{\mathrm{fid}} = 2.73 \pm 0.26$ fb $= 2.73 \pm 0.22$ (stat) ± 0.15 (syst) fb, in agreement with the SM expectation of 2.86 ± 0.15 fb. The measurement of the fiducial cross section in differential bins of $p_{\mathrm{T}}^{\mathrm{H}}$ is used to set constraints on the trilinear self-coupling of the Higgs boson, with an observed (expected) limit of $-5.4\,(-7.6) < \kappa_\lambda < 14.9\,(17.7)$ at the 95% CL. This is the first time that this type of constraint is carried out in a single-Higgs fiducial analysis. Finally, constraints on the modifiers of Higgs boson couplings to c and b quarks (κ_b and κ_c) are also determined with an observed (expected) limit of $-1.1\,(-1.3) < \kappa_b < 1.1\,(1.2)$ and $-5.3\,(-5.7) < \kappa_c < 5.2\,(5.7)$ at the 95% CL, obtained assuming a dependence of the BR on κ_b and κ_c. All results are consistent with the SM predictions for the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ decay channel in the considered fiducial phase space.

Chapter 6

Development of a novel method for the estimation of the reducible background for Run 3 and beyond

The Run 1 of the LHC marked a historic milestone with the discovery of the long-sought Higgs boson in 2012 and its first characterisation. With the large dataset of 138 fb^{-1} collected at $\sqrt{s} = 13$ TeV, the subsequent Run 2 opened the possibility to study the newly discovered particle in greater depth. The main five production mechanisms have been observed, the properties and the compatibility with the SM hypothesis have been tested with high precision and granularity, and the couplings to the third-generation fermions and gauge bosons have been firmly established. However, as we eagerly anticipate the advent of the HL-LHC, already in the Run 2, the dawning of a new era is in sight - the precision era of Higgs physics. Remarkably, with just 5% of the final HL-LHC dataset, the Run 2 data gave access to rare decay channels and allowed to probe the production and decay phase space of the Higgs boson with unprecedented granularity. The analysis presented in the previous chapters is a clear example of this achievement. Before the start of the CMS phase-2, another low-luminosity run (Run 3) is underway until 2025. The LHC Run 3 will not bring a significant statistical improvement, but it will give the opportunity to reassess, rethink, and improve the analysis strategies in view of the high-luminosity phase.

There are several directions worth investigating in the $\text{H} \to \text{ZZ} \to 4\ell$ channel. Firstly, studies are ongoing to tackle the leading systematic uncertainty related to the electron efficiency, which makes CMS less competitive than ATLAS and will be a major limitation with the growth of statistics. As already discussed in previous chapters, the way to go is very clear in that regard: the TnP method should employ the J/ψ resonance for low-p_{T} electrons. Another improvement can come by enlarging the phase space by lowering the p_{T} cut on electrons and/or the minimal cut on the m_{Z_2}. The effects of these changes will be mainly an increase in the background rate, which could be healed by using either matrix-element discriminants or machine-learning techniques to retain the maximal amount of signal events while reducing the background contamination.

The study of the reducible background is another appealing area of improvement. This chapter presents a novel method for the estimation of this challenging source of background. There are mainly two motivations for this effort. The first one is that the current fake rate method estimates the reducible background with an uncertainty of about 40% (cfr Sec. 4.3.1). This is mainly a systematic uncertainty intrinsic to the method itself, which will be a strong limitation in the high luminosity phase. The second motivation has a broader reach. The reducible background in the four-lepton channel is made of misidentified and non-prompt leptons, a major challenge in many other CMS analyses. Having a new method for estimating this kind of background could be helpful for other analyses to

have a deeper understanding of this topology of events.

The chapter begins in Sec. 6.1 with the explanation of the underlying rationale on which the method is based. Following this introductory overview, the ingredients for estimating the reducible background are introduced: the computation of the normalisation is detailed in Sec. 6.2, while the assessment of the transfer function is explained in Sec. 6.2. These two blocks should be then combined as discussed in Sec. 6.4. The method is then completed by the extraction of the shape presented in Sec. 6.5.

This new method will be implemented for the first time in the upcoming Run 3 $H \to ZZ \to 4\ell$ analyses. At the time of these studies, the new data were not yet available; therefore, the method was developed using 2017 data and then cross-checked with 2018 data. Since the conclusions are the same in both years, and the only difference is in the numerical values, all the plots in this chapter are reported with 2017 data. Only in the context of the transfer function, the 2018 results are shown, as they help strengthen the validity of some considerations, discarding the possibility that they are only valid due to fortunate coincidences in 2017.

6.1 The rationale of the method

The reducible background is already presented in Sec. 4.3.2, highlighting that the two main contributions to this background are from misidentified and non-prompt leptons. Among the physics processes contributing to these two topologies, the leading contributions are from the Drell-Yan process and $t\bar{t}$ production.

Non-prompt leptons typically arise from the decay of B mesons within jets originating from the hadronisation of b quarks. Inside these b-jets, there is generally the formation of B hadrons that subsequently decay semileptonically in lighter mesons. The spectator model can be used to describe the decay of heavy mesons. In the vast majority of hadrons, the heavy quark is paired with light quarks (u, d, s). In the decay of the hadron, in the first approximation, the role of the light quark can be neglected. It follows that when analysing the semileptonic decay of B-hadrons, one can simply look at the decay of the single b-quark. This approximation will be used throughout the chapter to provide physics insights. The first application of this approximation is to understand the displaced vertex for non-prompt leptons. These genuine leptons are originated by the semileptonic decay of the b-quark inside jets. Since the b-quark is lighter than its partner in the doublet (the t-quark), its decay can only go through processes described by off-diagonal elements in the CKM matrix, leading to a large lifetime. The typical lifetime of a B hadron is around $1.5 \cdot 10^{-12}$ s. Since $c\tau_B = 462$ μm, B hadrons fly for a few hundred micrometres before decaying. This implies that the decaying vertex of a B hadron is displaced from the PV of the collision, and the impact parameter of leptons coming from these decays is thus large. The same considerations apply to C-hadrons, *mutatis mutandis*. In what follows, this component will be referred to as *Z+b jets topology*.

On the other hand, misidentified leptons mainly arise from light jets where light mesons, such as pions, can be misreconstructed as leptons. Since these *fake* leptons originate from light jets, their vertex is very close to the PV, and the impact parameter is therefore small. In what follows, this component is referred to as *Z+light jets topology*.

From these considerations, one of the most powerful observables to discriminate between these two contributions is the impact parameter, particularly the SIP variable introduced in Sec. 3.2. For the Z+light jets topology, the SIP distribution is centred at small values and peaks at around 1. For the Z+b jets topology, the SIP distribution presents a longer tail at larger values instead. The novel method presented in this chapter relies on these features to estimate the normalisation of the reducible background in the SR of the

$H \to ZZ \to 4\ell$ channel. The SIP variable is thus the cornerstone of the method, hence aptly named *SIP method*.

Similar to the current fake rate approach (cfr Sec. 4.3.2), the SIP method is also data-driven. Consequently, the initial stage involves defining the CR:

- a Z_1 candidate satisfying the conditions outlined in Sec. 4.1;
- a pair of additional same-sign same-flavour soft leptons passing the identification and isolation requirements, as outlined in Ch. 3;
- invariant mass of the two additional soft leptons in the range $12 < m_{\ell\ell} < 120$ GeV;
- similarly to the selections for the SR: $p_T > 20$ GeV and $p_T > 10$ GeV for the leading and sub-leading lepton; $m_{\ell^+\ell'^-} > 4$ GeV, the event is discarded in the $4e$ and 4μ final states if $|m_{Z_a} - m_{Z_0}| < |m_{Z_1} - m_{Z_0}|$ and $m_{Z_2} < 12$ GeV;
- invariant mass of the four leptons greater than $m_{4\ell} > 70$ GeV.

This CR is similar to the Z+LL CR used in the SS fake rate method. The requirement on the same sign of the additional leptons make the CR orthogonal to the SR. The only missing requirement for the additional leptons to satisfy the tight requirements is the SIP cut. This CR does not target a particular component of the background, but it is enriched of all the topologies of the Z+X background. From these considerations, it follows that the estimation of the Z+X background in the SR can be computed as:

$$N_{SR}^{Z+X} = N_{CR}^{Z+X} \cdot r_{OS/SS} \tag{6.1}$$

where N_{CR}^{Z+X} is the number of Z+X events in the CR, i.e., the number of events with additional leptons passing the tight requirements, and $r_{OS/SS}$ is the transfer function from the SS to the OS phase space. Here, SS and OS refers to the sign of the two additional leptons. The first term will be discussed in Sec. 6.2, while the second in Sec. 6.3.

6.2 The normalisation

The normalisation of the SIP method N_{CR}^{Z+X} is formally defined as the number of events in the CR with four leptons passing the tight selections. As discussed in the previous section, the sole remaining requirement for considering an event in the CR as signal-like is the SIP cut for the additional leptons. The estimation of the normalisation can be obtained by assessing the probability of a single additional lepton satisfying the missing SIP cut. By denoting the SIP values of the two additional leptons as SIP_1 and SIP_2:

$$\begin{aligned} N_{CR}^{Z+X} &= N_{CR} \cdot \mathcal{P}(SIP_1 < 4 \cap SIP_2 < 4) \\ &= N_{CR} \cdot \mathcal{P}(SIP_1 < 4) \cdot \mathcal{P}(SIP_2 < 4) \\ &= N_{CR} \cdot \mathcal{P}(SIP_{either} < 4)^2 \\ &= N_{CR} \cdot \left(\int_0^4 f(SIP) \right)^2 \end{aligned} \tag{6.2}$$

where N_{CR} are the total number of events in the CR and $\mathcal{P}$ denotes the probability. The first step in the equation chain is valid as long as SIP_1 and SIP_2 are independent random variables. A priori, there is not any physical reason for the two variables to be correlated since the two denominations are arbitrary and not related to physics. SIP_{either} represents the probability for either the first or the second lepton to satisfy the SIP cut, which can

be computed as the integral between 0 and 4 of the pdf of the SIP, $f(\text{SIP})$, of the two additional leptons in the CR. Practically, the pdf is the normalised distribution of the SIP of the two additional leptons. To compute the integral, the pdf is fitted with the sum of two Landau functions: one describing the Z + light jets topology and the other for the Z + b jets topology. The functional form is given by:

$$f(\text{SIP}) = s \cdot \text{Landau}_1 + (1 - s) \cdot \text{Landau}_2 \tag{6.3}$$

where s denotes the strength parameter of the first Landau. The result of the fit is shown in Fig. 6.1.

The uncertainty on $N_{\text{CR}}^{\text{Z+X}}$ is obtained by propagating the post-fit uncertainties of the Landau parameters, considering the respective correlations encoded into the covariance matrix.

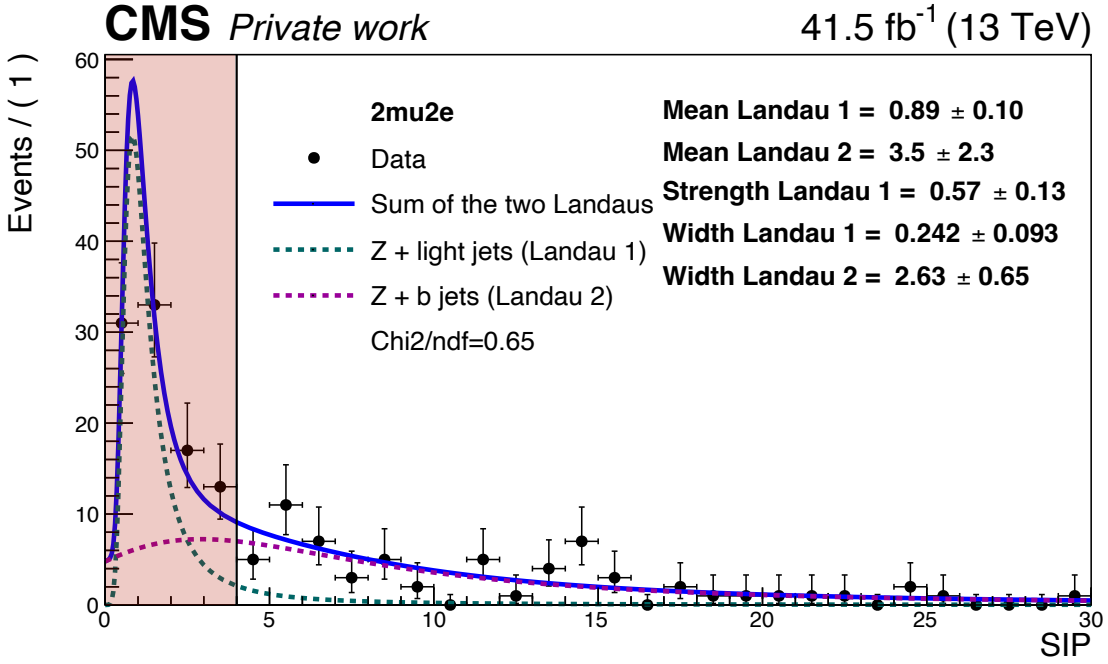

Figure 6.1: Fit of the SIP distribution of the two additional leptons in the CR with 2017 data in the $2\mu 2e$ final state. The red-coloured area represents the range where SIP< 4, i.e., the region up to the fitted function should be integrated.

6.2.1 Combination of same-flavour channels

In principle, the SIP fit should be done independently in each of the four final states, i.e., $2e2\mu$, $2\mu 2e$, $4e$, and 4μ. It turns out that, in some cases, the number of events populating the CR is low, leading to large statistical fluctuations and poor fit quality. In addition, considering the evaluation of the Z+X in categories and/or kinematic bins, where the CR can become significantly depleted, a strategy to increase the statistics should be employed.

The SIP method is based on the fit of the SIP variable of the two additional leptons in the CR. It means that in the $2e2\mu$ and 4μ final states, the muon SIP is fitted, while in the $2\mu 2e$ and $4e$ final states, the electron SIP is fitted. The flavour of the Z_1 does not play a role in the shape of the SIP distributions. It only affects the normalisation. A trick that can be implemented to increase the statistics and make the fit more stable is combining the final states where the flavour of the additional leptons is the same. Consequently, in the $2\mu 2e$ and $4e$ final states, the SIP distribution is fitted in the combined 2X2e channel

f_{2X2e}, whereas in the 2e2μ and 4μ final states, the SIP distribution is fitted in the combined 2X2μ channel $f_{2X2\mu}$:

$$N^{\text{Z+X}}_{\text{CR},4\mu} = N_{\text{CR},4\mu} \cdot \left(\int_0^4 f_{2X2\mu}(\text{SIP}) \right)^2$$
$$N^{\text{Z+X}}_{\text{CR},2e2\mu} = N_{\text{CR},2e2\mu} \cdot \left(\int_0^4 f_{2X2\mu}(\text{SIP}) \right)^2$$

$$N^{\text{Z+X}}_{\text{CR},4e} = N_{\text{CR},4e} \cdot \left(\int_0^4 f_{2X2e}(\text{SIP}) \right)^2$$
$$N^{\text{Z+X}}_{\text{CR},2\mu2e} = N_{\text{CR},2\mu2e} \cdot \left(\int_0^4 f_{2X2e}(\text{SIP}) \right)^2$$

As a cross-check of the validity of this assumption, the comparison between the SIP distributions in the four distinct final states is shown in Fig. 6.2. When comparing final states where the flavour of the additional leptons is not the same, e.g., 4e and 2e2μ, the SIP distributions are systematically in disagreement. In particular, the distribution for muons is broader than the electron's. This is explainable by looking at the definition of SIP, which is the ratio of the impact parameter and its uncertainty. Since larger uncertainties are associated with electrons, the SIP values will be smaller for the same impact parameter. On the other hand, when the comparison is made between final states where the flavour of the additional leptons is the same, e.g., 4e and 2μ2e, the SIP distributions are closely in agreement, and their ratio is compatible with one.

The fits in the combined 2X2μ and 2X2e channels are shown in Fig. 6.3, and the comparison between the normalisations obtained using independent and combined channels is reported in Tab. 6.1. The numerical values are in perfect accordance between the two approaches. The combination of the channels allows a reduction of the uncertainty from 27%, in the 2μ2e channel, to 39%, in the 2e2μ channel.

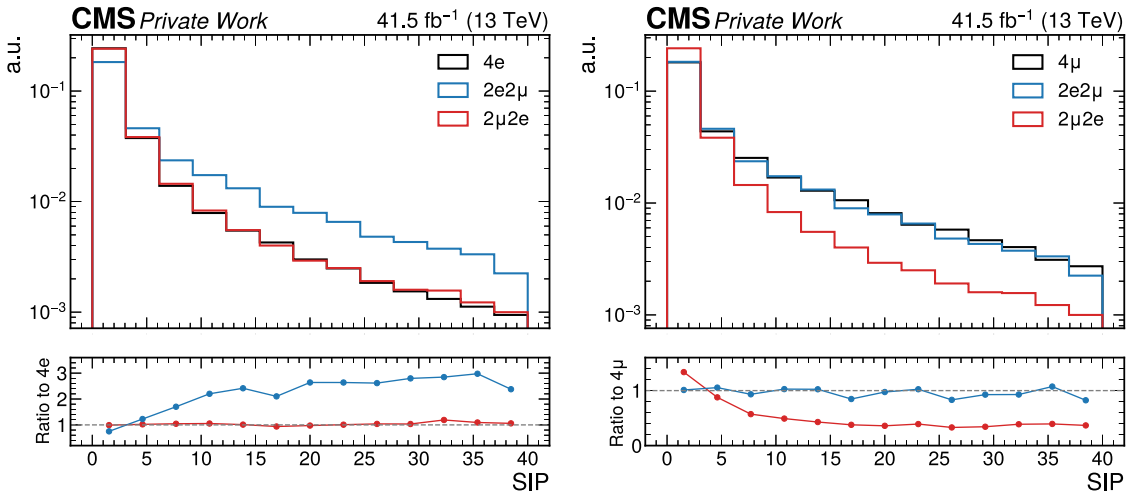

Figure 6.2: Distribution of the SIP variable of either additional leptons in the various channels with 2017 data in the CR. The additional leptons are considered as soft: cut on the transverse momentum, the pseudorapidity, and distance from the PV. (Left) Comparison of the distributions for the 4e, 2e2μ, and 2μ2e final states. The bottom pad shows the ratio to the 4e distribution. (Right) Comparison of the distributions for the 4μ, 2e2μ, and 2μ2e final states. The bottom pad shows the ratio to the 4μ distribution.

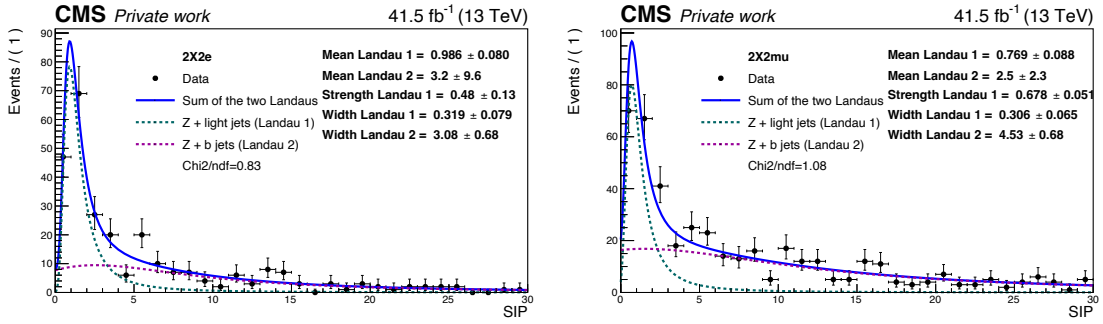

Figure 6.3: Fit of the SIP distribution of the two additional leptons in the CR with 2017 data in the combined channels: (Left) 2X2e = 4e + 2μ2e; (Right) 2X2μ = 4μ + 2e2μ.

Table 6.1: Estimation of the Z+X normalisation ($N_{\mathrm{CR}}^{\mathrm{Z+X}}$) with the SIP method using 2017 data. The values obtained by fitting the SIP distribution independently in each final state are reported in the column *independent channels*. The values computed by fitting the SIP distribution in the final states obtained by merging the channels with same flavour for the two additional leptons are reported in the column *combined channels*. The uncertainty in the second case decreases by 32%, 37%, 39%, and 27% in the 4e, 4μ, 2e2μ, and 2μ2e channels, respectively.

$N_{\mathrm{CR}}^{\mathrm{Z+X}}$	Indipendent channels	Combined channels
4e	20.15 ± 3.21	18.29 ± 1.97
4μ	25.15 ± 4.76	24.88 ± 2.94
2e2μ	17.95 ± 3.46	17.92 ± 2.11
2μ2e	25.37 ± 3.73	27.28 ± 2.94

6.2.2 Inversion of the isolation cut

An alternative approach for computing $N_{\mathrm{CR}}^{\mathrm{Z+X}}$ relies on inverting the isolation cut. This different strategy is based on doing two fits. First, fitting and integrating the SIP distribution removing the isolation cut on the additional leptons in the CR, $N_{\mathrm{CR}}^{\mathrm{ISO}>0}$. Second, fitting and integrating the SIP distribution considering events where the isolation cut on the additional leptons is reversed, $N_{\mathrm{CR}}^{\mathrm{ISO}>0.35}$. Then, by taking the difference between these two quantities, one gets the number of events with the proper cut on the isolation:

$$N_{\mathrm{CR}}^{\mathrm{Z+X}} = N_{\mathrm{CR}}^{\mathrm{ISO}<0.35} = N_{\mathrm{CR}}^{\mathrm{ISO}>0} - N_{\mathrm{CR}}^{\mathrm{ISO}>0.35} \tag{6.4}$$

The main advantage of this approach is the larger number of entries in the distributions to fit. When the isolation cut is inverted, the number of events increases significantly. For instance, in the 2X2μ channel with 2017 data, it goes from 222 to more than 3,000.

Since the isolation and identification criteria for electrons are jointly incorporated into a BDT, inverting selectively only the isolation cut is delicate. In this section, just as a showcase scenario, the focus is placed on the muon channels 2e2μ and 4μ. Adopting the same strategy delineated in the previous section, the SIP distribution for the 2X2μ final state is fitted and integrated up to 4. The resulting distributions are shown in Fig. 6.4, and the results of the integrals are reported in Tab. 6.2. The uncertainty propagates as:

$$\sigma_{N_{\mathrm{CR}}^{\mathrm{Z+X}}} = \sqrt{\sigma^2_{N_{\mathrm{CR}}^{ISO>0}} + \sigma^2_{N_{\mathrm{CR}}^{ISO>0.35}} - 2 \cdot \mathrm{Cov}(N_{\mathrm{CR}}^{ISO>0}, N_{\mathrm{CR}}^{ISO>0.35})} \tag{6.5}$$

As $N_{\mathrm{CR}}^{ISO>0.35}$ is obtained from a subset of $N_{\mathrm{CR}}^{ISO>0}$, the two terms are likely to be correlated. While the fitting procedure prevents an immediate computation of the correlation

between these two quantities, it is plausible to assume that $\mathrm{Cov}(N_{\mathrm{CR}}^{ISO>0}, N_{\mathrm{CR}}^{ISO>0.35}) \geq 0$. By setting the covariance to zero, this provides an upper bound to the estimation of the uncertainty. The rationale behind this assumption is that increasing one term would conceivably lead to an analogous increment in the other. This approach is denoted as *conservative* in what follows. However, an estimation of the covariance can be attempted assuming that it would be the same as for a simple counting experiment, without doing any fit:

$$\begin{aligned}\mathrm{Cov}(\sigma_{N_{\mathrm{CR}}^{ISO>0}}, \sigma_{N_{\mathrm{CR}}^{ISO>0.35}}) &= \mathrm{Cov}(\sigma_{N_{\mathrm{CR}}^{ISO<0.35}} + \sigma_{N_{\mathrm{CR}}^{ISO>0.35}}, \sigma_{N_{\mathrm{CR}}^{ISO>0.35}}) \\ &= \mathrm{Cov}(\sigma_{N_{\mathrm{CR}}^{ISO>0.35}}, \sigma_{N_{\mathrm{CR}}^{ISO>0.35}}) \\ &= \sigma^2_{N_{\mathrm{CR}}^{ISO>0.35}}\end{aligned} \tag{6.6}$$

Substituting in Eq. 6.5, it follows that:

$$\sigma_{N_{\mathrm{CR}}^{\mathrm{Z+X}}} = \sqrt{\sigma^2_{N_{\mathrm{CR}}^{ISO>0}} - \sigma^2_{N_{\mathrm{CR}}^{ISO>0.35}}} \tag{6.7}$$

The results are summarised in Tab. 6.2. The values of $\sigma_{N_{\mathrm{CR}}^{Z+X}}$ are perfectly in agreement with those quoted in Tab. 6.1, even though the approach explained in Sec. 6.2.1 relies on fitting a much less populated SIP distribution. This agreement underscores the robustness of the fit of the SIP distribution, even when dealing with fewer events, which will be helpful for estimating the background in categories. The larger number of events available using this approach has an evident effect on the precision of the integral, allowing for an approximate precision of 3%, compared to the 10% of the standard approach. However, this improved precision does not propagate to the final estimation of the normalisation. The subtraction between the integrals evaluated without isolation and with inverted isolation leads to a deterioration of the precision. This holds true when both the conservative approach and the counting experiment approximation are used to compute the uncertainty. Assuming no correlation provides a normalisation with 50% precision, unlike the roughly 10% quoted in Tab. 6.1. In summary, the approach presented in this section has been abandoned due to the lower attained precision, the difficulties in computing the correlation between the subtracted terms, and the not straightforward implementation for electrons. Nevertheless, this alternative remains relevant as a closure test and can be revived when necessary, especially for sparsely populated categories.

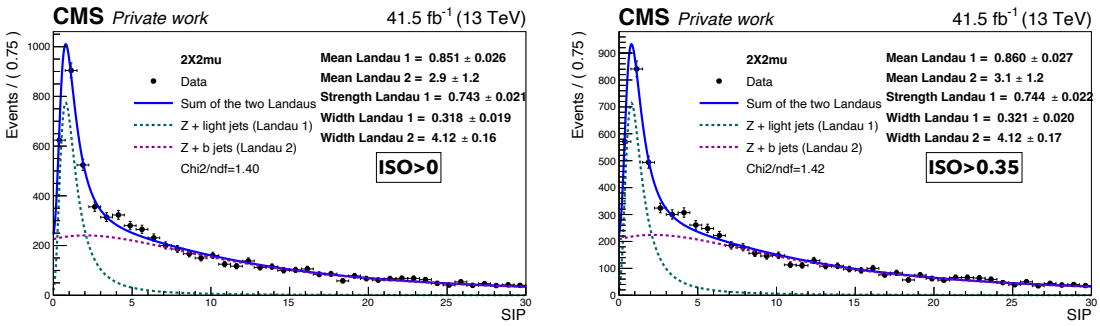

Figure 6.4: Fit of the SIP distribution of the two additional leptons with 2017 data in the combined channel 2X2μ = 4μ + 2e2μ. (Left) The definition of the CR is modified by removing the isolation cut. (Right) The definition of the CR is modified by inverting the isolation cut.

Table 6.2: Result of the integral of the SIP distribution removing the isolation cut on the additional leptons ($N_{\mathrm{CR}}^{ISO>0}$) and inverting the isolation cut on the additional lepton ($N_{\mathrm{CR}}^{ISO>0.35}$). The difference between these two quantities provides the estimation of the normalisation of the SIP method ($N_{\mathrm{CR}}^{\mathrm{Z+X}}$). The uncertainty on the difference is quoted both assuming no correlations and in parenthesis assuming the same correlation of a counting experiment. These numbers are computed for 2017 data.

	4μ	$2e2\mu$
$N_{\mathrm{CR}}^{ISO>0}$	282.19 ± 9.74	220.37 ± 7.61
$N_{\mathrm{CR}}^{ISO>0.35}$	257.18 ± 9.08	202.07 ± 7.13
$N_{\mathrm{CR}}^{\mathrm{Z+X}}$	25.01 ± 13.3 (3.5)	18.3 ± 10.4 (2.7)

6.3 The transfer function

In the previous section the first ingredient of the SIP method was computed. However, the $N_{\mathrm{CR}}^{\mathrm{Z+X}}$ term is just half of the method. This number cannot be applied directly to the SR, it should be corrected by a transfer function, connecting the CR to the SR, formally defined as:

$$r_{\mathrm{OS/SS}} = \frac{N_{\mathrm{SR}}^{\mathrm{Z+X}}}{N_{\mathrm{CR}}^{\mathrm{Z+X}}} \tag{6.8}$$

This ratio is usually referred to as *opposite-sign same-sign ratio*, since the CR is defined with same-sign additional leptons, while in the SR the additional leptons must be with opposite sign, forming a real Z boson candidate. Evaluating this ratio is the real challenge of the SIP method.

A MC-driven approach could be used to estimate this ratio. As already mentioned at the beginning of the chapter, the two leading physics processes contributing to this background are Drell-Yan and $\mathrm{t\bar{t}}$, with a strong predominance of the former over the latter. The related MC samples could be used to estimate the different compositions between the CR and the SR. Unfortunately, this approach cannot be implemented at the moment due to the inadequacy of the currently available MC samples. The Drell-Yan and $\mathrm{t\bar{t}}$ samples generated for Run 2 were not tailored for studies on the Z+X background: the number of events ending up in the CR and SR of the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ channel are not enough to extract meaningful information on the transfer function. These samples cannot be used for the purpose of evaluating the transfer function. However, this path will be investigated during the preparation of the Run 3 analyses when specific MC samples will be generated.

Another approach is to estimate the ratio relying directly on data, as follows:

$$r_{\mathrm{OS/SS}} = \frac{N_{\mathrm{SR}}^{\mathrm{Z+X}}}{N_{\mathrm{CR}}^{\mathrm{Z+X}}} \simeq \left.\frac{N_{\mathrm{SR}} - N_{\mathrm{ZZ}}^{\mathrm{MC}}}{N_{\mathrm{CR}}^{\mathrm{tight}}}\right|_{m_{4\ell}\notin[105,140]} \tag{6.9}$$

In this equation, N_{SR} are the events in the SR, from which the irreducible background component $N_{\mathrm{ZZ}}^{\mathrm{MC}}$ is subtracted using the corresponding MC samples, and $N_{\mathrm{CR}}^{\mathrm{tight}}$ are the number of events in the CR with two additional tight leptons. Both at the numerator and the denominator, the region of the Higgs peak $105 < m_{4\ell} < 140$ GeV is masked. The main issue with this approach is the magnitude of the numerator that results in being small, especially if the ratio is estimated independently in each final state. When considering the sum of the four final states together, the ratio is estimated to be:

Table 6.3: Values of the transfer functions $r_{\mathrm{OS/SS}}$ with the tight cuts applied to the additional leptons in the four final states. The numbers are reported for both 2017 and 2018 and, given the large uncertainty, the results are in agreement.

	$4e$	4μ	$2e2\mu$	$2\mu2e$
2017	0.6 ± 1.4	2.1 ± 1.5	1.3 ± 1.0	2.0 ± 1.0
2018	0.7 ± 0.8	4.0 ± 2.1	2.2 ± 1.5	1.5 ± 0.9

$$r^{\mathrm{tight}}_{\mathrm{OS/SS,2017}} = 1.57 \pm 0.60$$

$$r^{\mathrm{tight}}_{\mathrm{OS/SS,2018}} = 1.80 \pm 0.59$$

These two estimates come with an uncertainty of approximately 35%. This ratio will be addressed to as *tight* ratio for reasons that will become clear later in this section. This uncertainty should then be combined with the uncertainty of the normalisation, leading to a poor precision of the Z+X estimation in the SR. To give an insight into the level of precision for the single final states, the per-channel values of the ratio are reported in Tab. 6.3.

To mitigate the significant uncertainty associated with the ratio, a potential solution can be to loosen the constraints on cuts for the two additional leptons. Loosening the cuts increases the number of events both at the numerator and the denominator, providing a more precise estimation of the ratio. The underlying assumption of this approach is that the ratio is independent of the cuts applied to the additional leptons. This assumption will be scrutinised later in this section; for the time being, it is presumed to be valid. Consequently, it follows that:

$$r_{\mathrm{OS/SS}} = \frac{N^{\mathrm{Z+X}}_{\mathrm{SR}}}{N^{\mathrm{Z+X}}_{\mathrm{CR}}} \simeq \frac{N^{\mathrm{soft}}_{\mathrm{SR}} - N^{\mathrm{MC,soft}}_{\mathrm{ZZ}}}{N^{\mathrm{soft}}_{\mathrm{CR}}}\Bigg|_{m_{4\ell}\notin[105,140]} \tag{6.10}$$

The most radical approach is to remove all tight cuts, and stick to the soft cuts only, i.e., minimal requirements on the kinematics and the distance from the PV. By doing so, both the CR and the SR become much more populated, leading to a ratio for the inclusive final state equal to:

$$r^{\mathrm{soft}}_{\mathrm{OS/SS,2017}} = 1.025 \pm 0.007$$

$$r^{\mathrm{soft}}_{\mathrm{OS/SS,2018}} = 1.026 \pm 0.006$$

The uncertainty drops from 35% down to less than 1%. These values are in agreement with the $r^{\mathrm{tight}}_{\mathrm{OS/SS}}$, given also the large uncertainty of the tight ratio. The same effect is seen in the single final states.

If this were the end of the story, the method would provide a very precise estimation of Z+X and would be much more competitive than the fake rate method. There is however one strong assumption to test, which is the independence of the ratio from the cuts on the additional leptons. To investigate this assumption, the ratio is evaluated by progressively loosening the cuts at different stages. Figure 6.6 illustrates the observed trend for the 2017 and 2018 periods, focusing solely on inclusive results, as the low statistics in the four final states prevent the extraction of meaningful results.

Starting with the soft selections, successive combinations of cuts are applied on top, up to the tight lepton cuts (soft+identification+isolation+SIP). To ensure consistency between muons and electrons, the ID and ISO cuts are always considered together. As

explained in Ch. 3, the electron identification and isolation requirement are integrated together into a single BDT, whereas for muons, these selections are cut-based and independent. As it is not feasible to separate the two cuts for electrons, the same approach is adopted also for muons. Despite relatively large uncertainties associated with the *tight* and *soft+identification+isolation* ratio, a systematic trend emerges, proving that the ratio is indeed influenced by the applied cuts. Notably, a distinct increase is observed upon implementing the identification+isolation cut. From a physics perspective, this effect could be attributed to the Z+b topology and the isolation cut.

The Z+b jets topology comprises real leptons originating from the semileptonic decay of heavy hadrons inside heavy-flavour jets. The same-sign contribution to the ratio arises from two main processes. The Z+b jets topology is usually characterised by the presence of two b-jets. One additional lepton can be picked up from one of the b-jets. The second additional lepton can be selected from the other b-jet, where a double decay chain can occur, e.g., involving a B-meson decaying into a C-meson followed by another semileptonic decay. In such a scenario, the two chosen leptons exhibit the same sign. This process requires a double decay within one of the two b-jets, making it less likely than the corresponding opposite-sign contribution. A pictorial representation of this scenario is given in Fig. 6.5. Another source of same-sign events is associated with the oscillation of neutral B-mesons. If a B-hadron undergoes oscillation before decaying, the selected leptons may also exhibit the same sign. Similarly to the double decay, this scenario is less probable than the corresponding opposite-sign process, as the probability for a B-hadron to oscillate is approximately 10% [201, 202].

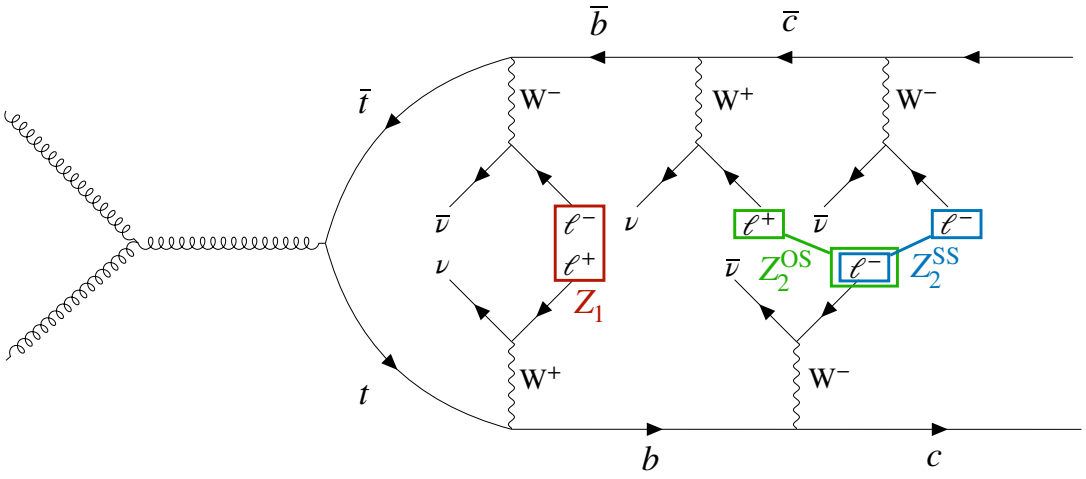

Figure 6.5: Illustration of the $t\bar{t}$ process, showing potential sources of opposite-sign and same-sign contributions. Upon the decay of the top quark, the b quarks hadronise. The subsequent decay chain is to be considered to happen inside hadrons, using the spectator model. The Z_1 candidate is formed from the leptons of the semileptonic decay of the $t\bar{t}$ pair. If the two additional leptons are selected as the following decay of the b quarks, they form an opposite-sign Z_2. On the other hand, if one of the two additional leptons is selected from the decay of a c-quark, the pair of additional leptons form a same-sign Z_2.

Since the transverse momentum of a particle is proportional to $m_{\text{mother}}/2$, these leptons originating from heavy-flavour hadrons can be considered as semi-isolated, meaning they are not neither entirely inside the jet cone nor outside of it. Consequently, the requirement for lepton isolation significantly suppresses these contributions, resulting in an excess of opposite-sign leptons over same-sign contributions. As a result, the $r_{\text{OS/SS}}$ becomes greater than one.

On the other hand, it can be inferred, given the current precision, that the relax of

the sole SIP does not have a significant effect on the value of the ratio, but it allows to have more statistics and reducing the uncertainty. The final approach that is used for the evaluation of the ratio is:

- for its value, the value of the ratio by simply removing the SIP cut, i.e., soft + identification + isolation, in the combined channels 2X2μ and 2X2e;
  - $r_{\mathrm{OS/SS}}^{2017,2\mathrm{X}2e} = 1.29\ (\pm 0.32)$
  - $r_{\mathrm{OS/SS}}^{2017,2\mathrm{X}2\mu} = 1.44\ (\pm 0.24)$
  - $r_{\mathrm{OS/SS}}^{2018,2\mathrm{X}2e} = 1.55\ (\pm 0.26)$
  - $r_{\mathrm{OS/SS}}^{2018,2\mathrm{X}2\mu} = 1.83\ (\pm 0.21)$
- for its uncertainty, the relative uncertainty on the tight ratio in the inclusive final states.
  - 2017 $\rightarrow$ 40%
  - 2018 $\rightarrow$ 33%

The value of the ratio is computed in the combined same-flavour channel following the same reasoning as in Sec. 6.1. The choice on the uncertainty allows to be more conservative against the relax of the cuts. The large uncertainty on the tight ratio could hide a systematic bias due to the relax of the SIP cut.

This is evidently not the best choice for the ratio; it is a conservative, physics-motivated approximation that is used to gauge the method. The optimal approach would be to either use MC samples, but this is not currently feasible with the available Run 2 MC samples, or exploiting the larger datasets that will be collected in the following years.

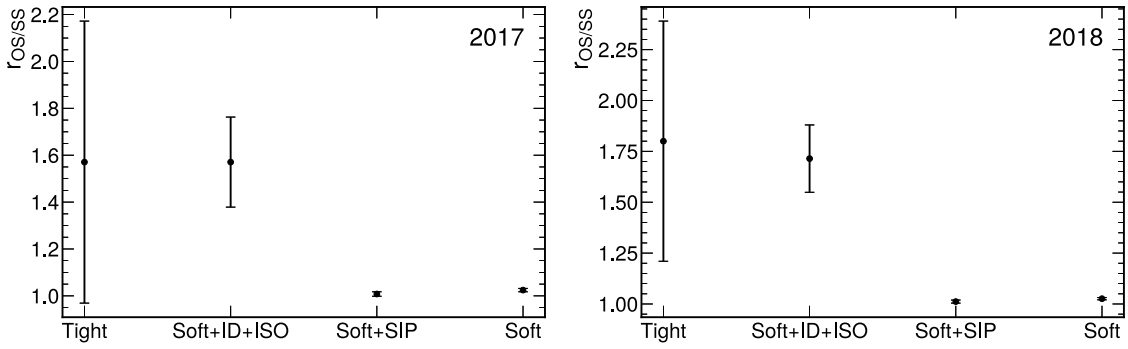

Figure 6.6: Evaluation of the transfer function $r_{\mathrm{OS/SS}}$ with different cuts applied to the additional leptons for 2017 (left) and 2018 (right). A soft lepton passes the kinematic requirements on its transverse momentum, rapidity, and distance from the PV. The acronyms ID and ISO refer to the identification and isolation requirements. A lepton passing soft selections ad the SIP cut is denoted in Ch. 3 as loose lepton. The trends in 2017 and 2018 are in agreement.

What about the transfer function in the SS fake-rate method?

The current method for estimating the reducible background is explained in Sec. 4.3.2. It is based on combining two sub-methods, the OS and SS fake rates computation. The SS method is based on a similar CR to the one used in the SIP method, and it also relies on a SS-OS transfer function, which is computed assuming that the ratio does not depend on the cuts. That means the transfer function in the SS method should also be revised

following the studies outlined in this chapter. Given the large uncertainty on the fake rate method and the small fraction of this background in the SR, the correction of the $r_{\mathrm{OS/SS}}$ transfer function does not invalidate the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ analyses exploiting this method. In the quest to provide an increasingly precise estimation of Z+X, these considerations should be taken into account in the upcoming $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ analyses.

6.4 The Z+X estimation

The SIP method has two ingredients, the normalisation $N_{\mathrm{CR}}^{\mathrm{Z+X}}$ and the transfer function $r_{\mathrm{OS/SS}}$, which then should be combined together (cfr Eq. 6.1) to obtain the estimation of the reducible background in the SR. For the final uncertainty, some considerations about the correlation are necessary. Equation 6.1 can be rewritten as:

$$N_{\mathrm{SR}}^{\mathrm{Z+X}} = N_{\mathrm{CR}}^{\mathrm{Z+X}} \cdot \frac{n_r}{d_r}, \tag{6.11}$$

where n_r and d_r represent the numerator and the denominator of the transfer function. The propagation of uncertainty results in:

$$\sigma_{N_{\mathrm{SR}}^{\mathrm{Z+X}}} = N_{\mathrm{SR}}^{\mathrm{Z+X}} \cdot \sqrt{\left(\frac{\sigma_{n_r}}{n_r}\right)^2 + \left(\frac{\sigma_{d_r}}{d_r}\right)^2 + \left(\frac{\sigma_{N_{\mathrm{CR}}^{\mathrm{Z+X}}}}{N_{\mathrm{CR}}^{\mathrm{Z+X}}}\right)^2 - 2\frac{\mathrm{Cov}(d_r, N_{\mathrm{CR}}^{\mathrm{Z+X}})}{d_r \cdot N_{\mathrm{CR}}^{\mathrm{Z+X}}}}. \tag{6.12}$$

The only possible source of correlation is between the denominator d and the result of the fit $N_{\mathrm{CR}}^{\mathrm{Z+X}}$, since they are both the outcome of an operation on a subset of the CR. Similarly to what was done in Sec. 6.2.2, certain logical assumptions can be made. If the number of events in the CR increases, it is reasonable to expect that both d and N_{CR}^{Z+X} will experience a similar increase. Consequently, the assumption that $\mathrm{Cov}(d, N_{\mathrm{CR}}^{Z+X}) \geq 0$ appears to be logically justified. Setting the covariance between the two terms simply equal to zero results in a more conservative approach by obtaining an upper bound on the uncertainty. Therefore, for the computation of the uncertainty, all the terms are considered as uncorrelated.

The final results for 2017 are reported in Fig. 6.7. The fake-rate method is quoted in its two OS and SS sub-methods (cfr Sec. 4.3.1). For the SS approach, the estimation with the same transfer function used in the SIP method is also reported. The final estimation of the SIP method is also quoted considering the transfer function evaluated with soft leptons. This number should be taken just as a reference and represents the precision that can could be achieved in the limit of a transfer function estimated with an uncertainty less than 1%. This target is not out of reach. As previously discussed, the usage of proper MC samples with enough statistics or the larger datasets that will be collected in the following years, will allow a more precise estimation of the ratio. The SIP method will provide a more competitive approach for the Z+X background compared to the fake-rate method.

6.5 The shape

The last missing ingredient to complete the SIP method is the shape extraction. The strategy is based on fitting the $m_{4\ell}$ spectrum in the CR with a Landau function for events with additional loose leptons. It is assumed that the shape in the CR is the same as that in the SR. The fit is performed in each final state independently. This procedure has been validated thoroughly in the past as it has been used in previous $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ publications [42, 128, 134]. Two examples of the Landau fit are shown in Fig. 6.8.

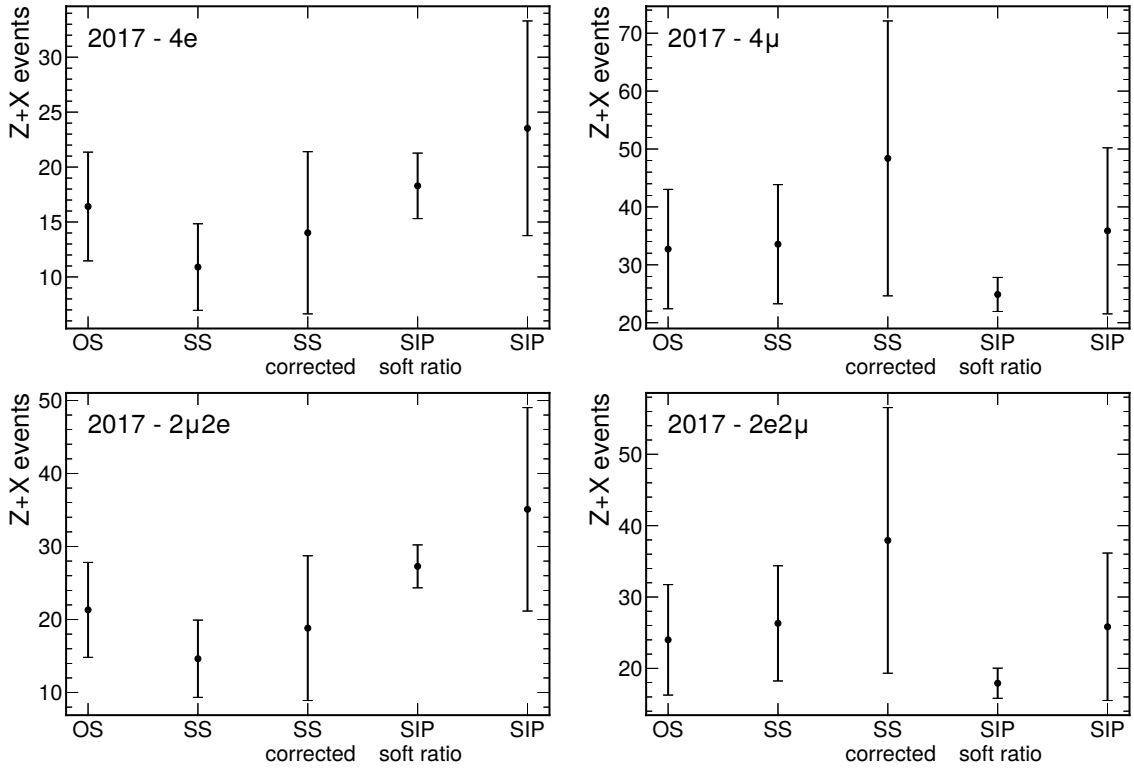

Figure 6.7: Estimation of the Z+X background in the SR of the H → ZZ → 4ℓ analysis with 2017 data. The results from the same-sign (SS) and opposite-sign (OS) fake rate methods are displayed as reference. The SS result is also reevaluated using the same transfer function used in the SIP method (SS corrected). The result of the SIP method is shown both with the application of the soft transfer function defined in Eq. 6.10 (SIP soft ratio) and as explained in the main text (SIP).

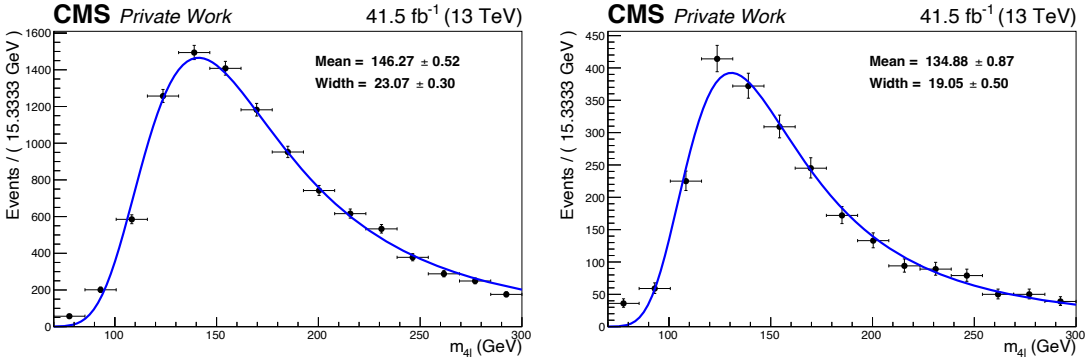

Figure 6.8: Fit of the $m_{4\ell}$ distribution for events in the CR with additional loose leptons with 2017 data in the 4e (left) and 4μ (right) channel. The distribution is fitted with a Landau function.

Summary

This chapter presented a novel data-driven method for estimating the reducible background in the H → ZZ → 4ℓ channel that will be employed for the first time in the upcoming Run 3 analyses. Although initially designed for the H → ZZ → 4ℓ channel, the method could be extended to all other analyses involving leptons with similar background topology.

To ensure orthogonality to the SR, events in the CR are required to have an exotic Z_2 boson candidate composed of two same-sign leptons, with the removal of the SIP cut. The proposed method is based on fitting the SIP distribution of these same-sign leptons to compute the number of events in the CR with additional leptons meeting the tight requirements. The primary limitation at present lies in the transfer function from the CR to the SR, which is derived with data and its precision is currently limited suffer due to low statistics. This bottleneck will be addressed in the future exploiting, either through proper MC samples or by utilising larger datasets that will be collected in the following years. The latter option holds promise for significantly reducing associated systematic uncertainties, relying more on data and less on simulation.

Upon achieving a desirable reduction in uncertainty for the transfer function, the SIP method demonstrates a potential precision of approximately 10%, outperforming the current 40% uncertainty associated with the fake-rate approach. Besides being simpler and easier to implement, the SIP method offers a notable advantage: its uncertainty is primarily driven by statistics. In contrast, the fake-rate uncertainty largely stems from systematic effects due to the different composition of the CR where fake rates are computed and where they are applied. This source of uncertainty will be a major limitation during the HL-LHC. The availability of larger datasets will substantially reduce the uncertainty associated with the SIP method, in contrast to the fake-rate method. Beyond a purely accuracy argument, misidentified and non-prompt leptons pose and will continue to pose a substantial challenge as background. Having a method that employs a different concept for their estimation will afford better control over this specific background, along with the potential to combine with the existing fake-rate method.

Part III

Offline reconstruction in the new High-Granularity endcap CALorimeter for the CMS phase-2

Chapter 7

Study of the reconstruction of unconverted photons and development of an algorithm to reduce pileup contamination

The previous part of this thesis was dedicated to the past and present of the CMS physics programme, with the analysis of the Run 2 data and the preparation for the upcoming Run 3 analyses. This part is instead devoted to the future of the experiment. The HL-LHC represents a new era in high-energy physics. It aims to boost the potential for discoveries of the LHC by increasing the integrated luminosity by a factor of 10 beyond the LHC's design value. This goal will be attained by raising the instantaneous luminosity, leading to an average number of simultaneous collisions per bunch crossing up to 200. A major upgrade is being prepared to maintain and even improve the physics reach of the CMS detector in these harsh conditions. The current endcap calorimeters will not be able to withstand the immense radiation flux expected in the forward region and will be replaced by a new HGCAL. The new endcap calorimeter will be the first large-scale silicon-based imaging calorimeter in a high-energy physics experiment. This detector offers the unique capability of performing calorimetry with tracker-like granular information, enabling unprecedented accuracy using position and timing information as well as superior particle identification capabilities with a PF approach. Such a revolution on the hardware side must accompany a paradigm shift on the software side as well. The offline reconstruction should be reassessed to cope with the dense environment and the innovative technology of the new detector. The Iterative CLustering (TICL) is a framework being developed within the CMS reconstruction software to carry out an optimal PF reconstruction together with all other subdetectors. This thesis focuses on the electromagnetic reconstruction, particularly in this chapter on studying unconverted photons.

After introducing the MC samples used in the chapter in Sec. 7.1, the TICL framework is presented in Sec. 7.2. The latest version TICLv4 was released during the timeframe covered by this thesis. In the same section, the novelties introduced in the new version are discussed, as well as comparing the performance compared to the previous TICLv3. The reconstruction of electromagnetic objects is scrutinised in Sec. 7.3, using MC samples without PU. In the presence of PU, TICL may pick up spurious PU-induced energy contributions, spoiling the resulting reconstructed electromagnetic object. A dedicated denoising procedure has been developed and is presented in Sec. 7.4.

7.1 Monte Carlo samples for photon studies

The studies presented in this chapter are all based on MC simulations, particularly on the GEANT4-based full simulations described in Sec. 2.5. For the samples needed here, the GEN step is represented by a *particle gun*. Particle guns are a group of generators implemented in the CMS software that allow the generation of a particle, according to the specified PDG identification number, in a predefined kinematic region. The particles are *shot* from the front face of the HGCAL to generate unconverted photons. Despite being generated in front of the calorimeter, the generated particle points towards the interaction point: its momentum direction corresponds to the segment connecting the interaction point and the point where the particle gun is placed. This alignment mimics the production of the particle at the interaction point. The HGCAL geometry implemented in GEANT4 for the SIM step is simplified compared to what will be the real calorimeter. The alternating active and passive layers are properly implemented, while all other additional components (such as readout chips, screws, supporting materials, etc.) are not simulated and replaced by layers of air. A graphical representation is reported in Fig. 7.1. Two sets of MC samples are used, one in the presence of PU and another in the absence of PU.

- **PU samples:** Each event has two unconverted photons generated uniformly in the energetic range $E = [10, 1000]$ GeV between pseudorapidity 1.65 and 2.75 with the two photons well separated so that their showers are not overlapping ($\Delta R =$ 100 cm). The choice of generating two photons is driven by the need to increase the statistics without increasing the number of samples and, consequently, improving the computational efficiency. A PU scenario at 14 TeV with the number of true interactions distributed following a Poisson distribution with a mean of 200 is used. The expected mean number of collisions per bunch crossing during the HL-LHC will be 140. However, the LHC can deliver 50% higher values for both the instantaneous and integrated luminosity, hence the simulation with PU = 200.

- **no-PU samples:** Each event has a single unconverted photon generated uniformly in the energetic range $E = [10, 1000]$ GeV between pseudorapidity 1.65 and 2.75.

During the course of this thesis, an engineering change was made to the HGCAL project. Specifically, the number of layers for the electromagnetic section CE-E was reduced from the original configuration of 28 ($26.3X_0$, 1.73λ) to the current design of 26 ($27.7X_0$, 1.5λ). This decision was mainly motivated by the pragmatic goal of cost reduction. A positive side effect of this adjustment is that each layer has more space to fit the boards and their connections, which is an essential consideration in the project development. Some results presented in this chapter were first derived using the 28-layer geometry, then re-validated with the 26-layer geometry without finding any significant discrepancy. If it is not explicitly indicated, it is implied that the new 26-layer geometry was used for the corresponding simulation. It is worth noting that, despite removing one layer, the total radiation length increases in the new geometry. The mere removal of one layer indeed reduces the radiation length to $25.4X_0$. However, the lead absorber in the last four cassettes has been increased such that the overall CE-E thickness is $27.7X_0$.

When working with samples generated in the presence of PU, it is essential to define a matching condition that effectively disentangle reconstructed photons from other reconstructed particles stemming from PU interactions. The criteria adopted are as follows. A reconstructed object in the HGCAL is matched to the generated photon if it shares at least 80% of its energy with the simulated photon, and satisfies a geometrical requirement between the direction of the generated photon and the position of the reconstructed object corresponding to $\Delta R < 0.2$. If a simulated photon has more than a matched trackster,

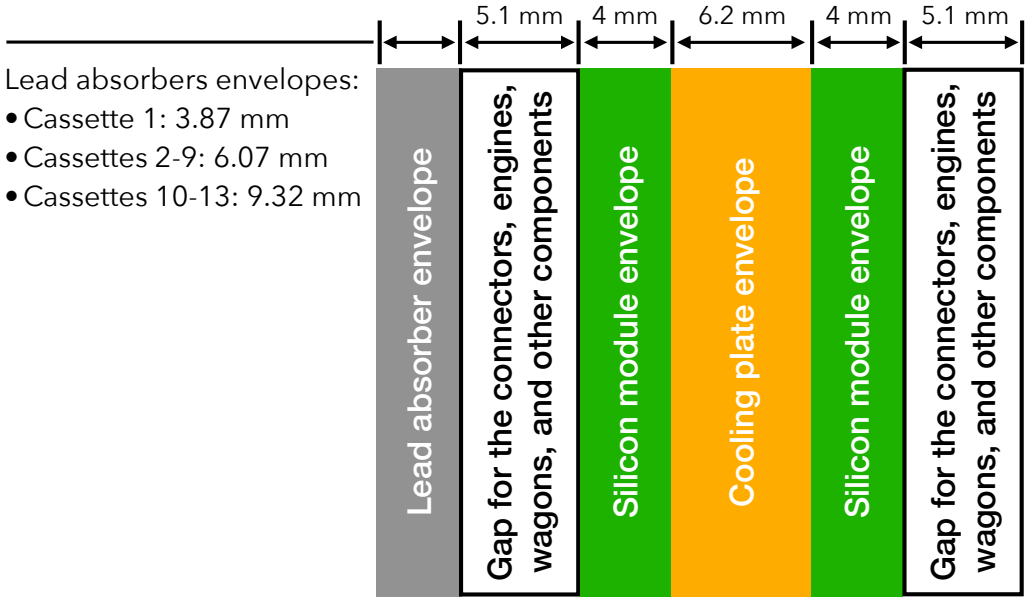

Figure 7.1: Representation of the simulated CE-E cassette. Connectors, engines, wagons, and other components are not simulated and replaced by a layer of air. The different thickness of the lead absorber layers is driven to ensure the needed total radiation length while having a finer sampling where the largest release of energy is expected for electromagnetic objects.

the one with the lowest energy difference to the generated photon is considered. This procedure is usually referred to as *gen-matching*.

7.2 The Iterative CLustering

The HGCAL offline reconstruction is developed within the TICL framework, which is the candidate to be the new PF framework for phase-2. It is a fast, fully parallelisable, modular framework that reconstructs particle showers and returns particle properties and identification probabilities. The advantage of the modularity is that new algorithms and techniques can be developed as plugins and then be quickly integrated with the framework. It eases the localisation and the identification of failures in the flow while also facilitating the comparison between different algorithms in each part of the chain. It naturally provides a fertile ground for continuous development with the most up-to-date technologies. The current data structure and algorithms are GPU-friendly to profit from heterogeneous computing. In addition, the framework is designed to be fast; the full HGCAL reconstruction using TICL takes about 4% of the total phase-2 offline reconstruction on a traditional single CPU core at 200 PU. The flowchart of the framework is reported in Fig. 7.2.

TICL is continuously being developed and improved, with new versions being periodically released within the CMS software. At the start of this thesis, TICLv3 was in release, while preparations were underway for the release of TICLv4. As this chapter unfolds and delves into the details of TICL, the main differences between these two versions will be highlighted.

7.2.1 recHits

As already introduced in Sec. 2.5, the inputs to the reconstruction framework are the so-called *recHits*, representing the reconstructed energy deposits in the volume corresponding

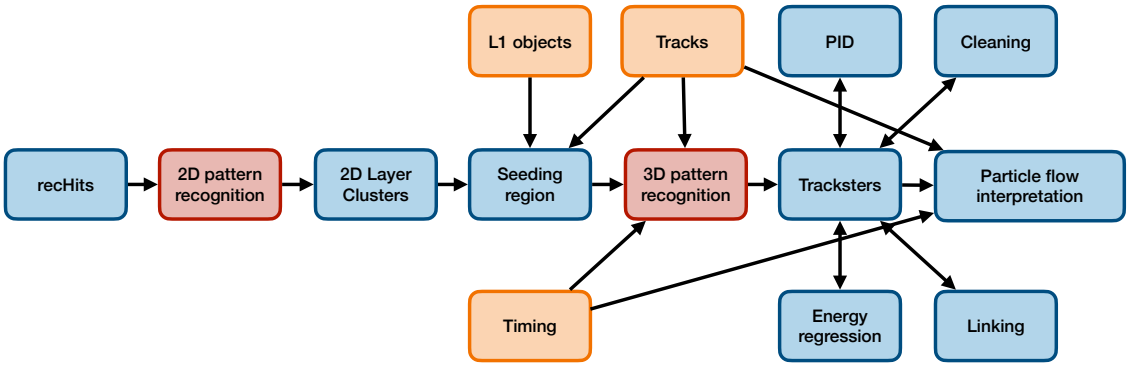

Figure 7.2: Flowchart of the TICL framework. The sequential modules are represented with blue boxes. External information to the HGCAL, which can be integrated into the TICL, is denoted by orange boxes. The pattern recognition algorithms are represented with red boxes.

to the cells of the HGCAL. Outputs from the detector electronics have a magnitude measured in femtoCoulomb. In general, the same amount of induced charge in the sensors may translate into different reconstructed signal amplitudes. This is why a calibration procedure is necessary, producing calibrated recHits that can be used for the subsequent clustering steps of TICL. For HGCAL Rechits, the calibration procedure is commonly referred to as *MIP calibration*, where MIP stands for Minimum Ionising Particle[1]. The first step of the procedure is to convert the charge to the average number of MIPs. MIPs have a well-defined energy loss per unit distance as they pass through a material. This energy loss is relatively constant and can be accurately modelled, making MIPs a convenient reference for calibration. Then, the equivalent number of MIPs is converted to energy, i.e., GeV, using the integrated energy loss for a MIP in an absorber layer. This procedure is valid since ionisation is the dominant energy absorption process in an electromagnetic shower. The final step involves an overall rescaling of the recHits magnitudes to match the sum of the energies of all recHits with the energy of the particle that initiated the shower. The recHits represent the real starting point of the TICL framework.

7.2.2 2D pattern recognition: CLUE

The TICL framework is expected to operate in an exceptionally harsh and challenging environment. The anticipated high level of PU during the HL-LHC phase will lead to a significant number of recHits per layer, and overlapping showers will be a common occurrence. These are the challenges that should be addressed by the reconstruction chain in the TICL framework. Fundamentally, a reconstruction algorithm aims to group individual recHits originating from the same particle shower effectively. This grouping is accomplished by pattern recognition algorithms. *A priori*, one single pattern recognition could be executed, starting from recHits and obtaining the final reconstructed shower as output. However, due to the large number of expected recHits (of the order of 10^5, i.e., a 10% occupancy), it is computationally advantageous to first collect recHits in 2D Layer Clusters (LCs) on a layer-by-layer basis, and then link LCs together to create 3D clusters representing particle showers. As a result, TICL implements two distinct pattern recognition algorithms, one designed for clustering in 2D and the other tailored for clustering in 3D.

[1]In practical cases, most relativistic particles have mean energy loss rates close to the minimum, and are said to be Minimum Ionizing Particles

The 2D pattern recognition algorithm is the first pivotal point of TICL. Naive algorithms exploring many combinations among all possible paths are anticipated to fail, given the likelihood of surpassing memory and timing constraints. What is implemented in TICL is a fast, fully-parallelisable density-based algorithm named CLUstering of Energy (CLUE) [203]. The CLUE algorithm can be summarised in five steps.

Setup of the data structure Each sensor cell on an HGCAL layer is treated as a point in a 2D space, with its energy as the key information of interest. As a density-based algorithm, it is paramount for CLUE to query the neighbourhood points of a given point to compute distances. The CLUE algorithm employs a fixed grid as a spatial index to do it efficiently. In essence, a grid is defined for each layer of the HGCAL. When querying a point with coordinates $(x_i,\ y_i)$, instead of scanning through all points in the layer, CLUE only loops over points residing within the bins touched by the square window $(x_i \pm d,\ y_i \pm d)$, where d is the distance to probe. This procedure is shown in Fig. 7.3. It is important to emphasise that all distance assessments in the later steps are computed using this approach and this does not represent an approximation.

Local energy density The local energy density ρ_i for a point i is computed as follows:

$$\rho_i = \sum_{j:d_{ij}<d_c} \chi(d_{ij}) E_j \tag{7.1}$$

where d_c is the cut-off distance in the calculation of the local density, and $\chi(d_{ij})$ is a convolution kernel implemented to give more importance to the point itself than its neighbours:

$$\chi(d_{ij}) = \begin{cases} 1 \text{ if } d_{ij} = 0 \\ 0.5 \text{ if } d_{ij} > 0 \end{cases}$$

Nearest-higher Once the local energy density is computed, the nearest point possessing a higher energy density, i.e., the nearest-higher, is identified for each given point. At this stage, two parameters, δ_o and δ_c, come into play. They are formally defined as the distance for outlier demotion and seed promotion, respectively. They will be explained better in the next step. Here, they are just used to define the distance within which the search for the nearest-higher is conducted: $d_m = \max(\delta_o, \delta_c)$. Consequently, the nearest-higher is searched in points with $d_{ij} < d_m$ and local energy density greater than the point itself $N_{d_m}(i) = \{j : d_{ij} < d_m, \rho_j > \rho_i\}$:

$$\text{nearest-higher}_i = \begin{cases} \underset{j \in N_{d_m}(i)}{\arg\min}\, d_{ij}, \text{ if size}(N_{d_m}(i)) \neq 0 \\ -1 \text{ otherwise} \end{cases}$$

The distance between the point i and its nearest-higher, if there is one, is denoted as δ_i.

Seeds, outliers, and followers The last parameter to introduce is ρ_c, the minimal local energy density for a point to be promoted as seed. As all the ingredients are now computed, and all tunable parameters are introduced, everything is ready for the crucial step of the CLUE algorithm. The points, i.e., the recHits, are labelled:

- as seed if $\rho_i > \rho_c$ and $\delta_i > \delta_c$;
- as outlier if $\rho_i < \rho_c$ and $\delta_i > \delta_0$;

After having defined seeds and outliers, CLUE labels each remaining point as a follower to its nearest-higher:

$$F_i = \{j : \text{nearest-higher}_j = i\}, \tag{7.2}$$

Clustering The final step of CLUE is the creation of the cluster, in our case of the 2D LCs. The LCs are formed starting from seeds and adding recHits through their chains of followers. Outliers are rejected from the final LCs.

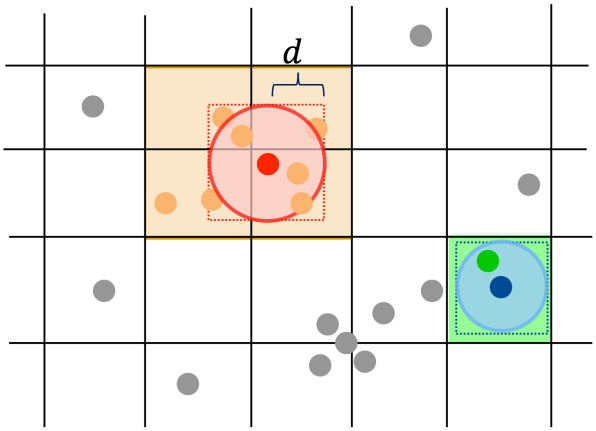

Figure 7.3: Visualisation of the setup of the data structure for CLUE and method for probing the neighbourhood of a point. Each grey point is a recHit. When querying the red (blue) point, only the points within the bins touched by the square window ($x_i \pm d$, $y_i \pm d$) are considered, i.e., the orange (green) points. Figure taken from [203].

Navigating through the CLUE algorithm, one encounters four tunable parameters: ρ_c, d_c, δ_c, and δ_0. Respectively, these parameters denote the minimum local energy density for a recHit to be promoted as a seed, the radius for the computation of the local energy density, and the distances of the nearest-higher for seed promotion and outlier demotion. Their values in the TICL's CLUE for the silicon layers are [204]:

- $\rho_c = 9 \times \sigma_{\text{noise}}$ (where σ_{noise} is defined as the energy-equivalent-noise and depends on the cell and its value is usually around 1-2 MeV);
- $d_c = \delta_c = 1.3$ cm;
- $\delta_0 = 2 \cdot \delta_c = 2.6$ cm.

The CLUE algorithm is run on a filtered set of recHits that have an energy of more than three times the corresponding σ_{noise}. With this choice of parameters, the algorithm is robust against noise and is able to cluster almost all the energy deposited by particles.

7.2.3 3D pattern recognition

The CLUE algorithm produces LCs, which subsequentially serve as input to the following pattern recognition, which aims to interlink LCs and reconstruct the 3D shower. Prior to running the algorithm, a seeding region can be delimited. The concept behind the seeding region is to restrict the number of LCs provided as input to the pattern recognition, thereby speeding up the iteration. The default configuration, employed for the results presented in this thesis, runs globally in the full HGCAL. However, a subregion of the detector could be potentially identified either through track extrapolation from the tracker to the HGCAL

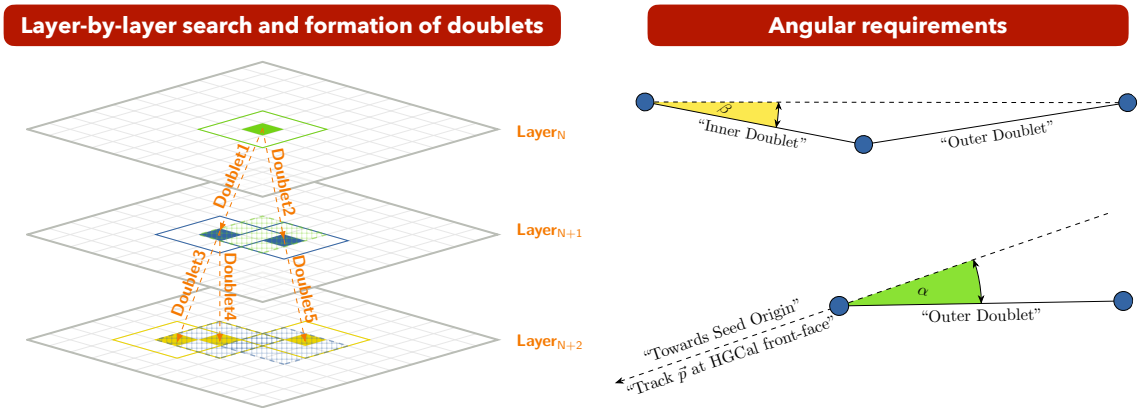

Figure 7.4: Visual representation of the Cellular Automaton (CA) algorithm. (Left) Layer-by-layer search of Layer Clusters to form doublets. (Right) Angular requirements between doublets to establish a linking.

surface or by a L1 trigger seed. For instance, the latter approach could help execute the algorithm at the HLT, which needs fast algorithms.

The output of the 3D pattern recognition represents the final reconstructed shower, commonly referred to as *trackster*. This name stems from the granular properties of the HGCAL that enable the reconstruction of calorimeter clusters with properties akin to that of a tracker. Formally, a trackster is defined as a Direct Acyclic Graph, where nodes are the LCs and the edges represent the connections identified by the pattern recognition among the LCs.

The main difference between TICLv3 and TICLv4 lies in the 3D pattern recognition: TICLv3 employs the Cellular Automaton (CA) algorithm, while TICLv4 utilizes the CLUE3D algorithm. Conceptually, these two algorithms differ in their fundamental principle: CA is a purely geometrical algorithm, while CLUE3D, like its 2D counterpart, follows the energy flow in the reconstruction of the shower. TICLv3 employed four CA-based iterations, each targeting a different type of particle. The first was a track-seeded iteration for electrons and charged hadrons, followed by an electromagnetic iteration for photons, a hadronic iteration for neutral hadrons, and a MIP iteration for muons. During the transition between two iterations, LCs successfully clustered in the preceding iteration were masked before starting the new one. On the other hand, TICLv4 currently employs an unseeded iteration with CLUE3D, with shower identification being delegated to specific algorithms in subsequent reconstruction steps.

Cellular Automaton

In the TICL implementation of CA, the algorithm can be summarised in the following steps:

- consider a LC in layer N;
- open a window around the LC under study in the (η, ϕ) space and project this window onto the following layer N+1;
  - To define the projection window, a 2D grid in η and ϕ is defined for each layer. The grid is defined with a series of squares with size set to 0.05, i.e., ∼70 mm at $\eta = 1.6$ and 20 mm at $\eta = 2.8$. The window is made of $3 \times 3 (5 \times 5)$ of these bins for $\eta < 2.1 (\eta \geq 2.1)$, and it is centred on the bin where the LC is.
- identify all LCs in layer N+1 that fall within the projected region and establish a connection with the LC in layer N, these connections are called *doublets*;

- iterate this procedure for all LCs within layer N, and then proceed to all subsequent layers.

Once the doublets are formed, they are linked together if two angular requirements are satisfied. The first is a requirement on the direction of each doublet with respect to the origin of the seeding region, or the origin of the coordinate system if the iteration is unseeded. The second is a requirement on the angle between a pair of doublets. The specific values of these angles depend on the particular iteration.

CLUE3D

The CLUE3D pattern recognition shares the same underlying logic as CLUE, yet it should not be considered as a 3D extension of CLUE - an algorithm that directly forms tracksters from recHits. Instead, CLUE3D clusters the LCs generated by CLUE to reconstruct the 3D shower. The workflow closely mirrors that of CLUE. For each LC i on layer j, the local energy density ρ_i is computed by considering all LCs in layers $j \pm k$ (within the HGCAL layer boundaries), whose projected distance onto layer j falls within Δ from i. The adjustable parameters k and Δ are set in TICL to $k = 3$ and $\Delta = 3.24$ cm. This computation does not consider the layer j on which the LC i resides. This choice relies on CLUE's ability to separate energy from two objects into two different LCs, thereby assuming that two distinct LCs belong to two physics objects. The kernel function is defined as:

$$\chi(\mathrm{LC_p}) = \begin{cases} 1 \text{ if p = i} \\ 0.2 \text{ if p} \neq \text{i} \end{cases}$$

The next step is the identification of the nearest-higher, which is again done by masking the layer j. Seeds, followers, and outliers are labelled following the same approach implemented in CLUE.

- Seed:
  - $\rho_i > \rho_c$;
  - distance on the transverse plane to the nearest-higher greater than d_c^{xy};
  - layer separation along z greater than n_c;
  - $E_i/\rho_i > \rho_c^{rel}$
- Outlier:
  - $\rho_i < \rho_c$;
  - distance on the transverse plane to the nearest-higher greater than d_o^{xy};
- Follower: all the LCs that are left.

As a final skimming step, tracksters composed of less than 2 LCs are removed. The five tunable parameters are set to:

$$\rho_c = 0.6 \text{ GeV} \quad \rho_c^{rel} = 0.2 \quad d_c^{xy} = 3.24 \text{ cm} \quad n_c = 5 \quad d_o^{xy} = 2 \cdot d_c^{xy}$$

As highlighted in the explanation above, CLUE3D is configured to favour the linking of LCs along the z direction rather than in the transverse plane. However, this does not preclude CLUE3D from forming tracksters with more than one LC per layer. If such linking occurs, it would be indirect, tracing a connection via another LC on another layer.

7.2.4 Performance of CA and CLUE3D with unconverted photons in the presence of pileup

In order to conduct a fair comparison of the performance and features of the two pattern recognition algorithms, CA and CLUE3D, the same set of events is processed using the two algorithms. The primary figure of merit is to evaluate the reconstruction efficiency, defined as the ratio of gen-matched tracksters to the number of generated photons. The results, reported in Fig. 7.5, show the dependence of the reconstruction efficiency on the energy, pseudorapidity, transverse momentum, and azimuthal angle of the generated photons. Despite the relatively high efficiency achieved by the CA (on average 93%), the CLUE3D algorithm outperforms it across all kinematic bins, consistently maintaining a value close to unity. In the known critical kinematic regions, i.e., low p_{T} and high pseudorapidity, CLUE3D demonstrates a significantly superior efficiency compared to CA, exhibiting a less significant drop to the plateau of the distributions. The lower efficiency of CA can be attributed to situations where the algorithm clusters photon energy in multiple sub-tracksters, comprising a subset of LCs from the photon shower and others from PU. This occurrence is often observed when PU LCs are closely aligned along the same direction of the photon shower. This effect is due to the geometrical intrinsic nature of CA. Besides, CLUE3D also exhibits greater PU resilience. Its reconstruction efficiency is unaffected by varying levels of PU, while CA suffers from a 10% reduction in efficiency at 200 PU compared to 0 PU.

One of the requirements for the HGCAL is the capacity to separate between closely spaced showers, which drove the necessity for a fine lateral granularity in the design of the calorimeter. Also in that domain, CLUE3D outperforms CA. The CLUE3D algorithm is able to separate showers down to 2.5 cm in the high-density region (cell size of 0.52 cm^2), while CA requires at least 3 cm to resolve two showers effectively. In the low-density region (cell size of 1.18 cm^2), the minimum separation is 3 cm in both cases. From a computational standpoint, CLUE3D has better timing performance. It scales better than CA, being almost 1.8 times faster when operating at 200 PU, and it has been developed to be easily ported on GPU, which will improve even more the run time of the algorithm [205].

The performance of CLUE3D unquestionably surpasses that of CA. Nonetheless, it is also crucial to study the quality of the output objects. Figure 7.6 shows the energy response, defined as the ratio of the trackster energy to the energy of the generated photon. It is important to note that no energy corrections or regressions are applied to the tracksters; the study focuses on the raw output of the pattern recognition algorithms. At higher energies ($E_{\mathrm{GEN}} > 400$ GeV), a perfect agreement is observed between the CA and CLUE3D distributions. However, as the energy decreases, the two distributions begin to diverge. The CLUE3D distribution tends to shift upwards toward zero, while the CA distribution shifts downwards. Although it may seem that CLUE3D delivers a superior response compared to CA at low energy, this effect is illusory and can be attributed to PU. Both pattern recognition algorithms unavoidably capture a fraction of PU-induced LCs. Given that the energy of such LCs is typically low, their impact becomes more evident when the algorithm attempts to reconstruct a shower triggered by a low-energy photon. The observed divergence in response between CA and CLUE3D indicates that the latter is a greedier algorithm. This effect can be attributed to the intrinsic nature of the algorithm itself. When the electromagnetic shower originates from a low-energy photon, the energy of the associated LCs is also low. Since CLUE3D is an energy-based algorithm, it may find pathways in the energy flow driven by PU rather than by the electromagnetic shower. An example is shown in Fig. 7.7.

In summary, this section outlined the main differences between CA and CLUE3D,

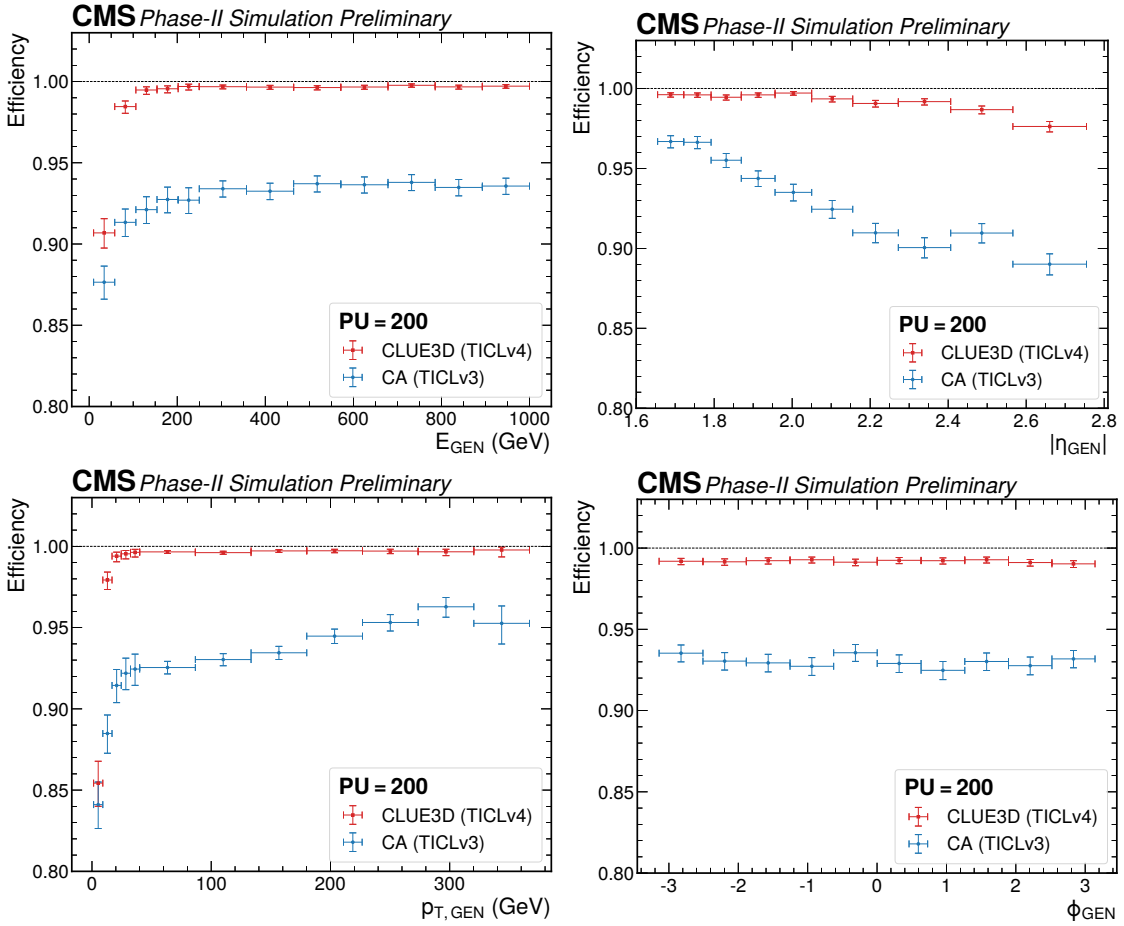

Figure 7.5: Trackster reconstruction efficiency as a function of energy, pseudorapidity η, transverse momentum p_{T} , and azimuthal angle ϕ of the generated photon. The same set of events is reconstructed with TICLv3, implementing the CA pattern recogntion, and with TICLv4, implementing the CLUE3D pattern recognition. The 28-layer geometry is used. The reconstruction efficiency is defined as the number of gen-matched tracksters divided by the number of generated photons. A photon trackster is gen-matched if it shares at least 80% of the energy with a simulated photon and satisfies a geometrical requirement between the gen-direction and the trackster position equal to $\Delta R < 0.2$. If a simulated photon has more than a gen-matched trackster, the one with the lowest energy difference with respect to the generated photon is considered.

showing that the latter is a more efficient, more reliable, faster pattern recognition algorithm. These considerations drove the decision to migrate from CA to CLUE3D as the standard 3D pattern recognition algorithm in TICLv4. In addition, CA always needs a seed, and the choice of the seed impacts the resulting trackster. The CLUE3D algorithm, being unseeded, removes this arbitrariness. However, although the collection of spurious LCs in the electromagnetic showers is a feature of both algorithms, CLUE3D is more aggressive than CA. This issue will be addressed in the dedicated *cleaning* module of TICL described in Sec. 7.4.

7.2.5 Life of a trackster

The output of the 3D pattern recognition is the trackster, which represents the reconstructed shower in the HGCAL. Once the trackster is reconstructed, it may undergo a sequence of operations to refine it and to obtain, at the end of the PF chain, the final

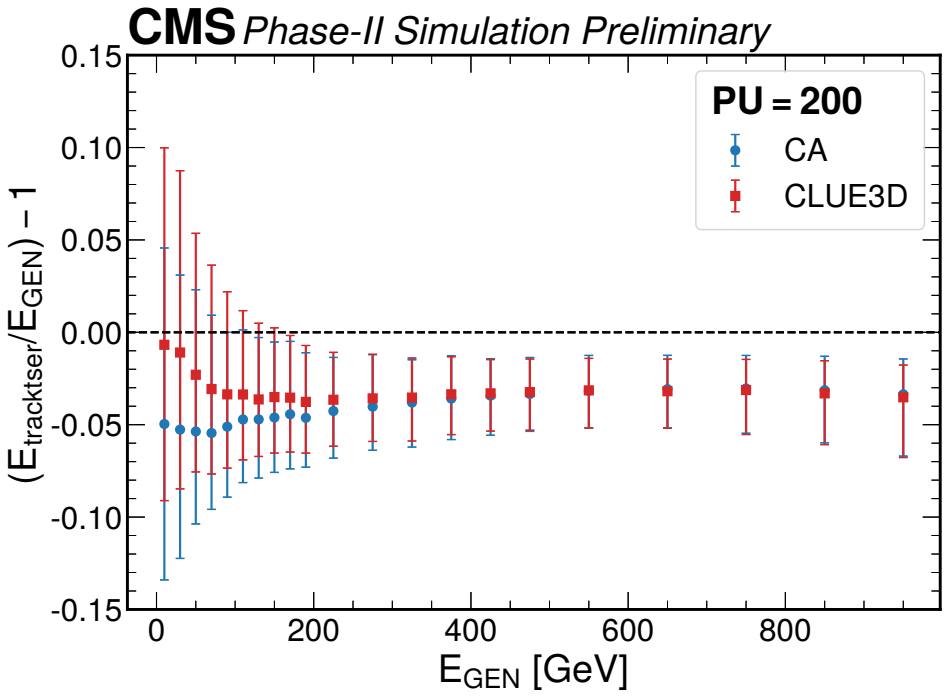

Figure 7.6: (top) Energy response ($E_{\text{trackster}}/E_{\text{GEN}}$) as a function of the energy of the simulated photon E_{GEN}. No energy corrections are applied. The points represent the median of the distributions. The error bars are the difference between the 15.9th and 84.1th percentile: this gives the 68.3% coverage around the median. The 28-layer geometry is used.

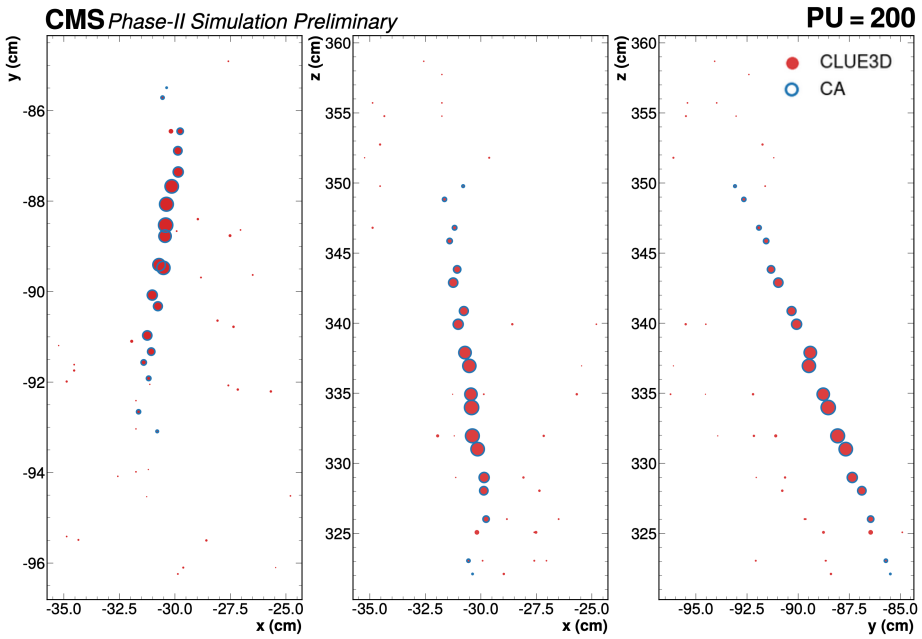

Figure 7.7: Event display of the same unconverted photon reconstructed with CA and CLUE3D to show that the latter tends to pick up more PU than the former. The additional LCs in CLUE3D are related to PU. The radius of the 2D LCs is proportional to their energy. Generated photon: $\eta_{\text{GEN}} = 1.98$, $\phi_{\text{GEN}} = -1.90$, $E_{\text{GEN}} = 80.15$ GeV, $p_{T,\text{GEN}} = 21.65$ GeV. The 28-layer geometry is used.

reconstructed particles to be used in physics analyses. Different modules have been, are being, and will be developed to cope with different steps of the reconstruction chain. Currently, a big fraction of such steps is still to be built, and the work presented in this thesis contributes to this effort.

A crucial step after the reconstruction of the shower is the energy determination, nowadays often carried out through a multivariate approach, namely *energy regression*. Due to factors such as instrumental effects, reconstruction limitations, and energy losses, the measured energy may not directly correspond to the actual energy of the particles. The energy regression aims to correct these effects and provide a more accurate estimation of the true energy.

While unconverted photons presented here are a reference candle to study the performance of a calorimeter, a real scenario can be much more complicated. Photons originating from the PV may undergo pair production along their path to the HGCAL, electrons crossing the tracker can emit bremsstrahlung radiation, and hadrons can undergo nuclear interactions. Despite the peculiarities of each scenario, the common denominator of all these situations is the presence of many tracksters in the HGCAL associated to the same original object. The *linking* modules are being developed to cope with such situations in order to combine different tracksters to reconstruct the original particle. In particular, the next chapter will present one of the member of this family, the so-called *electron superclustering*.

The CLUE3D algorithm is implemented as a unique iteration, reconstructing all tracksters without targetting any particular type of shower. The identification is left to a series of Particle IDentification (PID) modules, which determine the type of particle responsible for initiating the shower. PID techniques rely on various shower species' unique characteristics and properties, often employing machine learning techniques. All the other steps of the reconstruction chain can benefit from this module. For instance, passing as input to the electron superclustering only tracksters likely to originate from electromagnetic objects improves the quality of the reconstructed electron, reducing the probability of collecting spurious tracksters.

These are the most relevant steps that can be implemented in TICL to improve the quality of the reconstruction. Particularly, Sec. 7.4 will focus on the cleaning module that tackles the issue highlighted in Sec. 7.2.4: the tendency of both CA and, more markedly, CLUE3D to collect PU-induced LCs, which result in noisier tracksters and a degradation of the energy response.

7.3 Characterisation of the photon reconstruction in the absence of pileup with CLUE3D

This section is devoted to study the fundamental characteristics of photon reconstruction with CLUE3D. To do so, the single-photon sample without PU (cfr Sec. 7.1) is a valuable tool to assess the basic features of the reconstruction, nailing down how an electromagnetic shower looks like in the HGCAL after being reconstructed with the combination of CLUE and CLUE3D. A typical event, with the corresponding energy profile, is depicted in Fig. 7.8. One specific feature of showers reconstructed in the HGCAL is the wiggling energy profile, which is a direct consequence of the layer structure of the calorimeter itself. As shown in Fig. 7.1, the odd and even layers are preceded and followed by different amounts of passive layers, leading to such fluctuations in the energy profile. More details on this effect can be found in [108].

Naively, as we are testing the reconstruction with single-photon samples in the absence of PU, one would expect to find a single trackster per event: one generated photon, one

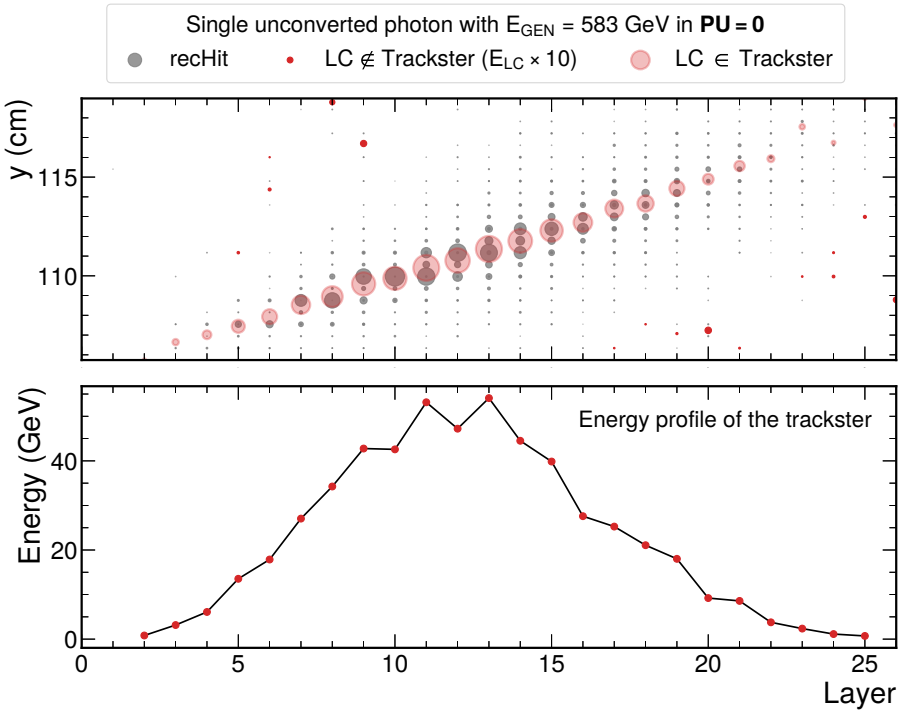

Figure 7.8: Reconstruction of an unconverted photon in the absence of PU with TICLv4. (Above) Event display in the layer-y plane. The dimension of the dots is proportional to the energy. The grey dots represent the recHits. The red dots represent the LCs that are not clustered into the trackster. Their energy is multiplied by 10 to be more visible. The shaded red dots represent the LCs that form the trackster. (Bottom) Energy profile of the reconstructed trackster as a function of the layer number.

electromagnetic shower, and one reconstructed trackster. However, it is not the case in the majority of events. In approximately 70% of cases, more than one trackster is reconstructed. The number of reconstructed trackster can even reach as high as 20, albeit with a rapidly decreasing probability as the count increases. The probability of having more than 4 tracksters is 30%, while for more than 8 tracksters, it is 17%, and for more than 10 tracksters, a mere 3%. A typical situation is reported in Fig. 7.9. While the main shower is usually well reconstructed, some low-energy, spurious, satellite tracksters are present around the main one. In other words, CLUE3D has a rather strong tendency to split showers. This splitting phenomenon exhibits a clear kinematic pattern, as reported in Fig. 7.10. It occurs more frequently as the energy and pseudorapidity of the photon increases. The growth as a function of the energy is linear. The number of particles produced in an electromagnetic shower depends on the energy of the photon. As the number of particles involved in the development of the shower increases, fluctuations at the edges of the shower will be more likely and, given the high granularity of the HGCAL, could be identified as independent tracksters by CLUE3D. On the other hand, the η dependence reveals two well-defined regimes at small and large pseudorapidity. The transition region is placed at approximately $|\eta| \simeq 2.15$, which is aligned with the boundary of the high-density and low-density region (cfr Sec. 2.4.1). In the low-density region, the mean number of trackster is around 2, whereas, after crossing the boundary, it rises abruptly to an average value of 5. This feature is related to the cells size in the two regions of the calorimeter. The forward region is more granular and the chances of having more satellites is higher.

More crucial than their number is the study of the kinematic properties of these sec-

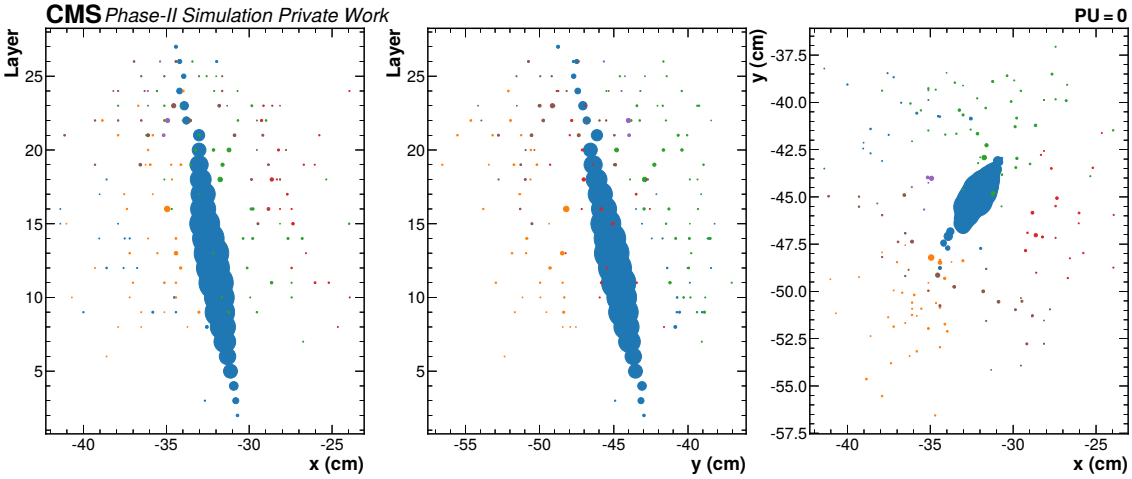

Figure 7.9: Event display of an unconverted photon in the absence of PU. The two longitudinal projections (x, Layer) and (y, Layer), and the transversal projection (x, y) are shown. All points in the plots correspond to LCs belonging to reconstructed tracksters, where each trackster is identified by a different color. The kinematics of the generated photon is: $E_{\text{GEN}} = 737$ GeV, $\eta_{\text{GEN}} = 2.5$, $\phi_{\text{GEN}} = -2.2$. In this event there are six reconstructed tracksters with energy $E_{\text{blue}} = 677$ GeV, $E_{\text{orange}} = 7$ GeV, $E_{\text{green}} = 9$ GeV, $E_{\text{red}} = 4$ GeV, $E_{\text{purple}} = 2$ GeV, and $E_{\text{brown}} = 6$ GeV.

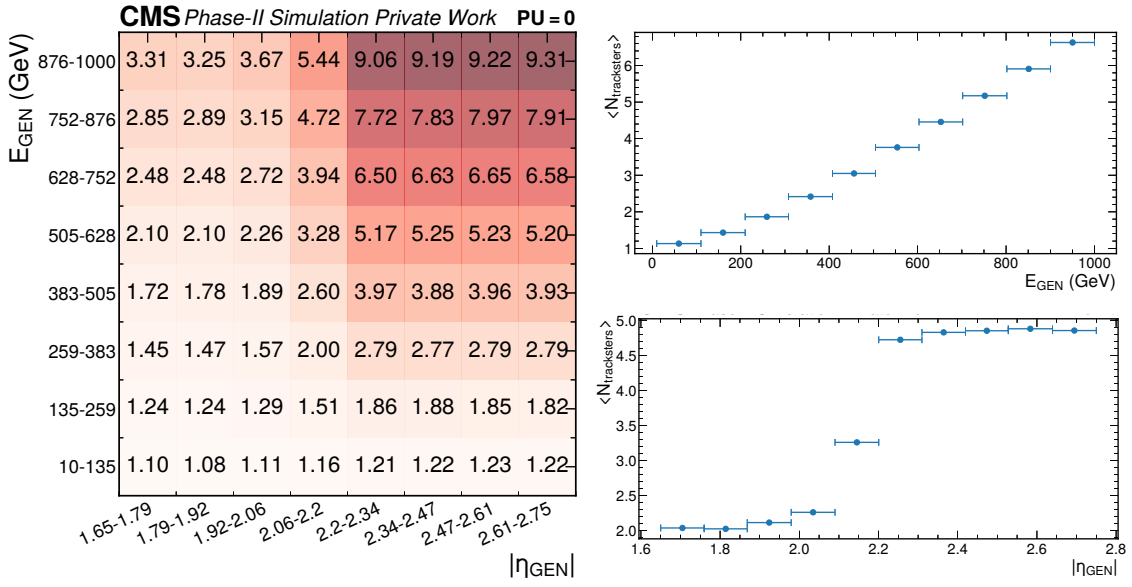

Figure 7.10: (Left) Mean number of tracksters in an event in 2D bins of energy and pseudorapidity of the generated photon. (Right) Mean number of tracksters in an event in bins of energy (top) and in bins of pseudorapidity (bottom) of the generated photon.

ondary tracksters and their influence on the reconstruction of the leading shower. Figure 7.11 illustrates the energy ratio between the two most energetic tracksters in an event. The distribution is squeezed towards low values, with its 95% quantile at 1.8%. This is a positive indication, pointing to scenarios where the secondary tracksters take away only a small fraction of energy from the main trackster. In addition, the bulk of the distribution, concentrated at low ratio values, exhibits a precise bimodal shape due to the differences in the low-density and high-density regions of the HGCAL. In the high-density region, the

subleading tracksters are likelier to carry a more significant fraction of energy. Although the large-η distribution is shifted to larger values, the magnitude of this shift is small. Specifically, the 95% quantile in the high-density region is 1.9%, compared to 1.1% in the low-density region. The kinematics of these satellite tracksters can be observed in Fig. 7.12. Their kinematic dependencies mirror their number: a continuous linear growth as a function of the energy and a clearly visible different behaviour depending on the η-region of the detector.

These features of the CLUE3D reconstruction are not worrisome. The kinematic impact of these satellites on the event is minimal, and if needed, they could be linked back together to the main trackster in a dedicated linking plugin of TICL.

More concerning is the presence of a very long tail in the distribution shown in Fig. 7.11, which can extend up to 100%. Despite being sparsely populated (0.36% of events for value of $E^{\text{subMaxE}}_{\text{trackster}}/E^{\text{subMaxE}}_{\text{maxE}}$ greater than 5%), this tail may represent rare events where CLUE and/or CLUE3D face significant problems in correctly reconstructing the electromagnetic shower. Since few events are falling in this region, it is challenging to extract explicit kinematic dependencies. It can be inferred that events are mainly concentrated at low energy and low transverse momentum. While there is no pronounced eta dependence, a mild tendency favouring higher eta values could be inferred.

The distinguishing feature of these events is the presence of a hard trackster splitting, as reported in Fig. 7.13. What can be clearly identified as a unique electromagnetic shower is, in fact, reconstructed by CLUE3D as two separate and independent tracksters. The energy profile provides insight into the origin of this splitting. The shower develops as expected in the initial layers and reaches its peak at layer 9, consistent with typical electromagnetic showers in the HGCAL. However, instead of a continuous descent, around layer 15, the energy stops declining and begins to increase again, forming a secondary peak at layer 19. Given this energy profile, CLUE3D does its job correctly. There are two well-separated local energy maxima that are promoted as seeds during the CLUE3D clustering; hence, the algorithm can do no more than build two distinct tracksters. The same pattern is also found at recHit level, excluding issues in the CLUE pattern recognition.

After ruling out concerns related to the pattern recognition algorithms, two plausible explanations are left: either a bug in the simulation chain or a real physics feature of the shower. Since these events have been noticed for the first time in the context of this thesis, the logic would lead to the former explanation. However, as highlighted before, these kinds of events are extremely rare and require specific scrutiny to be identified, often remaining obscured within the overwhelming number of *standard* showers. It has been asked to search explicitly for this kind of events in data and simulations analysed by the test-beam study group [206], as well as in an independent simulation of a sampling calorimeter set up by the GEANT4 collaboration. Remarkably, the split tracksters were observed in all three cases. To paraphrase Agatha Christie, one coincidence is just a coincidence, two coincidences are a clue, three coincidences are a proof. These findings strongly point to an underlying physics process in electromagnetic showers.

Debugging a shower in a calorimeter to understand the cause of the secondary peak is a real challenge. The number of processes and particles involved is huge, and tracking them all down is a formidable job. In collaboration with the GEANT4 experts, intensive efforts are underway to pinpoint the source of these unique events. The most likely explanation is that a hard bremsstrahlung photon is emitted during the shower unfolding, which triggers a secondary displaced shower and is responsible for the secondary peak. For the time being, and the studies presented in this thesis, setting aside the physics explanation, these events do not currently pose a major issue. Firstly, they are extremely rare, and indeed they went unnoticed for years. Secondly, they can be linked easily in a linking module that

can be included in the TICL workflow. The alignment between the two split tracksters is almost perfect, and some geometrical matching conditions can be used to merge them together.

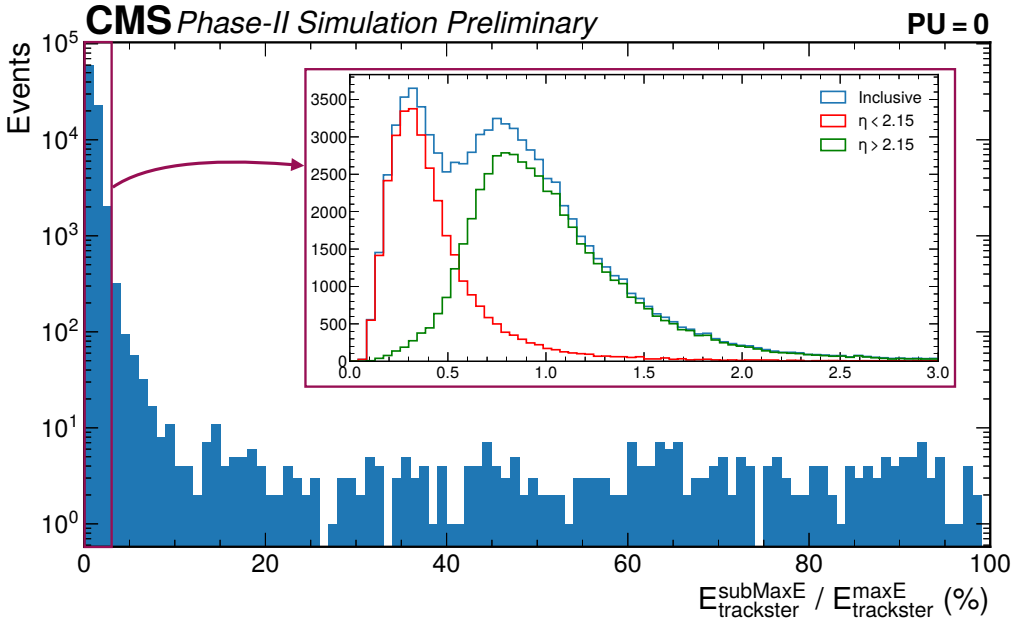

Figure 7.11: Distribution of the ratio between the second most energetic trackster ($E^{\text{subMaxE}}_{\text{trackster}}$) and the most energetic trackster ($E^{\text{subMaxE}}_{\text{maxE}}$) in an event. The values are reported in percentage (%). The insert shows a zoom of the region between 0 and 3, highlighting the different distributions in the high-density region ($|\eta| \gtrsim 2.15$) and low-density region ($|\eta| \lesssim 2.15$) of the HGCAL.

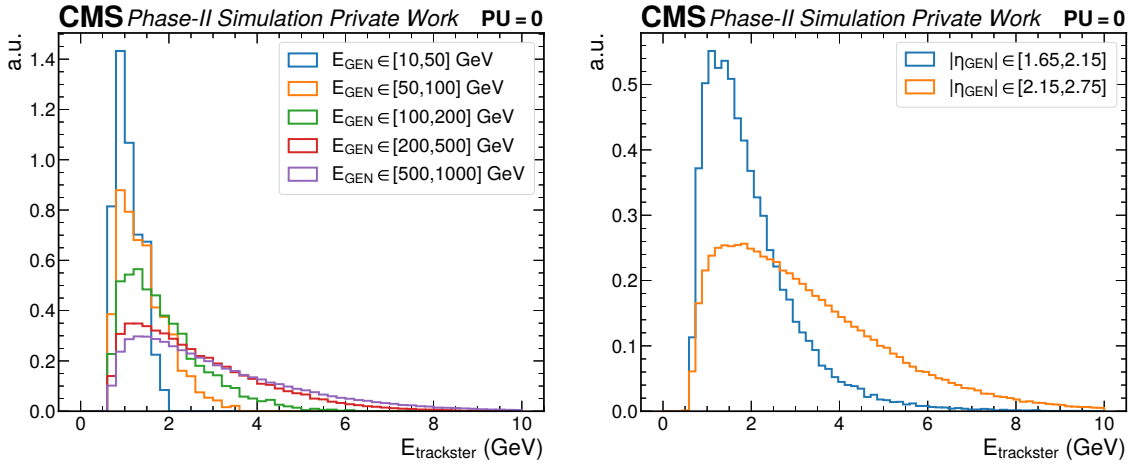

Figure 7.12: Energy of the satellite tracksters in an event in bins of energy (left) and pseudorapidity (right) of the generated photon.

7.4 Cleaning the tracksters

The HL-LHC will increase the luminosity level to unprecedented levels to access rare phenomena and do precision physics, causing the PU level to increase to three times its current value. This increase will directly affect the reconstruction algorithm that has to cope with

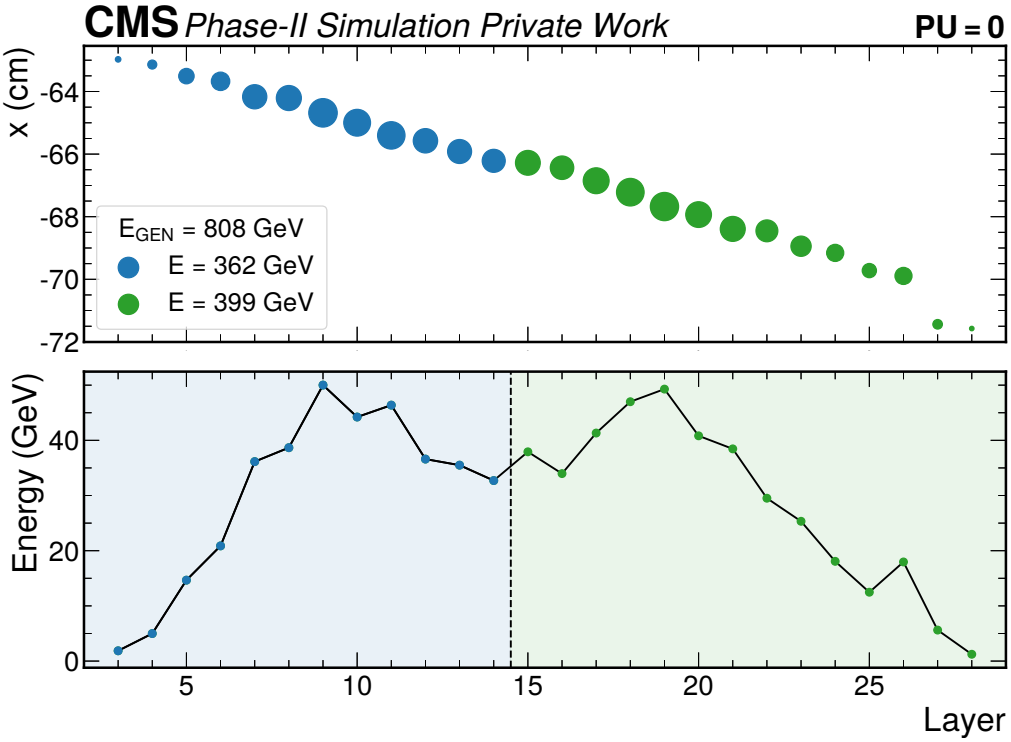

Figure 7.13: Reconstruction of an unconverted photon without PU in TICLv4. (Above) Event display in the x-layer plane. Each dot represents a LC of the reconstructed trackster. The radius of the dots is proportional to the related energy. Each color depicts a different trackster. (Bottom) Energy profiles of the two reconstructed tracksters as a function of the layer number.

a very dense environment. The previous sections show that the combination of CLUE and CLUE3D provides an efficient and reliable reconstruction for showers in the HGCAL. However, CLUE3D (but in general any pattern recognition algorithm) is naturally subject to the clustering of a certain amount of PU. This spurious contribution mainly affects the extremities of the shower, especially for low-energy particles. In such cases, the presence of low-energy LCs in the electromagnetic shower could mislead CLUE3D, collecting LCs associated with PU interactions and spoiling the resulting trackster. A dedicated PU reduction algorithm, referred to as *cleaning* algorithm, has been developed in the context of this thesis and implemented as a plug-in in the TICL framework. This algorithm revolves around the Principal Component Analysis (PCA), explained in the following section.

7.4.1 Principal Component Analysis

The PCA is a dimension reduction technique that aims to retain as much information as possible. In essence, it involves transforming high-dimensional data into a lower-dimensional space that maintains the fundamental characteristics of the original data. The first step is to define what *information of a dataset* means. From a mathematical standpoint, the variance is an objective variable to quantify the amount of information in a dataset. According to the definition, the variance expresses the deviation of the data points from the mean of the dataset. Consider a scenario where a dataset is perfectly aligned along a direction. If the data points are projected onto an orthogonal axis, they will all collapse to the same value, losing all the information in the dataset, resulting in a variance equal to zero. Conversely, the axis aligned with the dataset conveys the maximum amount of information, leading to the highest possible variance. Consequently, the computation

of the principal components relies on the covariance matrix of the dataset. The objects of this chapter are the electromagnetic showers, and each shower can be represented as a matrix:

$$\mathbf{X} = \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ & . & \\ & . & \\ x_n & y_n & z_n \end{bmatrix} \tag{7.3}$$

where each row represents the (x,y,z) coordinates of a LC. A common procedure when computing the PCA is to center the data to obtain the so-called *mean centred dataset*. The mean $\overline{x}$, $\overline{y}$, and $\overline{z}$ are computed for the dataset and subtracted to each component. For simplicity, $\mathbf{X}$ will also represent the centred dataset in what follows. The next step involves the computation of the covariance matrix $\mathbf{C}$. Since the dataset $\mathbf{X}$ is now centred, $\mathbf{C}$ can be computed as:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} \tag{7.4}$$

By definition, $\mathbf{C}$ is a symmetric matrix, therefore it can be diagonalised:

$$\mathbf{C} = \mathbf{E}\mathbf{L}\mathbf{E}^{\mathrm{T}} \tag{7.5}$$

where $\mathbf{E}$ is the eigenvector matrix and $\mathbf{L}$ is the eigenvalue matrix, which is a diagonal matrix with the eigenvalues λ_i along the diagonal in descending values. The eigenvalues represent the amount of variance explained by each principal component, while the eigenvectors define the directions in the original data space along which the variance is maximised. Larger eigenvalues indicate that the corresponding principal components capture more significant variations in the data. The principal components can be then computed from the eigendecomposition of the covariance matrix. Then the PCA could proceed further with the projection of the dataset onto the PCA axes and carrying out the real dimension reduction, but this goes beyond the needs for the cleaning algorithm.

From a practical point of view, the PCA usually relies on the Singular Value Decomposition (SVD) for the computation. The SVD ensures that a complex matrix, that will be called $\mathbf{X}$ and considered real for direct application to the case under study, can always be decomposed as:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}} \tag{7.6}$$

where $\mathbf{U}$ is an $m \times m$ unitary matrix, $\mathbf{S}$ is an $m \times n$ diagonal matrix with non-negative real numbers on the diagonal, $\mathbf{V}$ is an $n \times n$ unitary matrix. The diagonal elements of $\mathbf{S}$ are called *singular values*. From Eqs. 7.4 and 7.6:

$$\begin{aligned} \mathbf{C} &= \frac{1}{n-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} \\ &= \frac{1}{n-1}(\mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}})^{\mathrm{T}}(\mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}) \\ &= \frac{1}{n-1}\mathbf{V}\mathbf{S}\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}} \\ &= \frac{1}{n-1}\mathbf{V}(\mathbf{S}\mathbf{S}^{\mathrm{T}})\mathbf{V}^{\mathrm{T}} \end{aligned} \tag{7.7}$$

The last equality is a consequence of the unitarity of the matrix $\mathbf{U}$, i.e., $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbb{1}$, where $\mathbb{1}$ is the unit matrix. The product $\mathbf{S}\mathbf{S}^{\mathrm{T}}$ is a square matrix $m \times m$, where the diagonal entries are the squared elements of the entries on the diagonal of the $\mathbf{S}$ (or the $\mathbf{S}^{\mathrm{T}}$) matrix.

Comparing Eq. 7.7 with Eq. 7.5, the singular values are related to the eigenvalues of the covariance matrix and the matrix $\mathbf{V}$ is related to the eigenvectors. From the computation

of the SVD, one calculates automatically the eigenvectors and eigenvalue of the covariance matrix. The computation of the PCA components via the SVD is the most widespread approach and the advantage is mainly related to the numerical stability.

The PCA computation can be extended to take into account a weighed dataset [207]. In the case of the electromagnetic showers, the weight of each LC will be its energy. The procedure outlined above is extended by considering a weighed covariance matrix:

$$\mathbf{C}_w = \frac{1}{1 - \sum_{i=1}^{n} w_i} \mathbf{X}^{\mathrm{T}} \mathbf{W} \mathbf{X} \tag{7.8}$$

where $\mathbf{W}$ is the diagonal matrix with the weights w_i as its diagonal elements.

The output of the PCA is a set of orthogonal axes, i.e., the principal components, where the first axis accounts for the largest amount of variance in the dataset. This axis is usually called the *main PCA axis*. Subsequent axes are chosen to maximise the remaining variance while being orthogonal to the previous components. For the purpose of the cleaning algorithm, only the main PCA axis will be used as an estimate of the shower axis.

7.4.2 The algorithm

Starting from the reconstructed photon trackster, the cleaning algorithm can be summarised as follows:

1. The trackster generated by CLUE3D may contain more than a single LC per layer. Only the most energetic LC per layer, which is more likely to belong to the proper photon shower, is retained.

2. The *centre* of the shower is defined as the position of the LC with the highest energy. The search for this maximum value is limited to the CE-E compartment of the HGCAL. In cases where an electromagnetic shower starts showering later and/or exhibits a deeper extension, the energy released in the first CE-H layer is typically much larger than the energy released in the last CE-E layer, creating a secondary peak in the energy profile (Fig. 7.14). This effect is due to the 45 mm thick stainless steel back disk between the electromagnetic and hadronic sections. It serves as an engineering support for the CE-E and as absorber before the hadronic calorimeter. However, for the subsequent steps of the cleaning procedure, it is necessary to identify the actual peak of the shower development, which lies in the CE-E.

3. With the subset of selected LCs from step 1, the shower axis is defined as the main energy-weighed PCA axis (cfr Sec. 7.4.1). The computation utilises the (x, y, z) coordinates of the LCs, focussing on the central part of the shower within (+15, -10) layers from the centre of the shower identified during step 2. This reduced range reduces instabilities due to fluctuations at the extremities of the shower.

4. After the computation of the shower axis, all LCs come back into play, and in each layer, only the LCs with the smallest distance to the axis are retained.

5. The trackster is eventually cleaned by keeping (+15, -12) layers from the centre of the shower. The choice of the number of layers is a trade-off to keep most of the energy, as shown in Fig. 7.15, while reducing the contamination from PU.

The cleaning algorithm can be seen in action in Fig. 7.16. The shower axis that is computed in the cleaning algorithm is promoted to be the axis of the reconstructed object, which gives the direction of propagation of the shower. It was investigated the possibility to

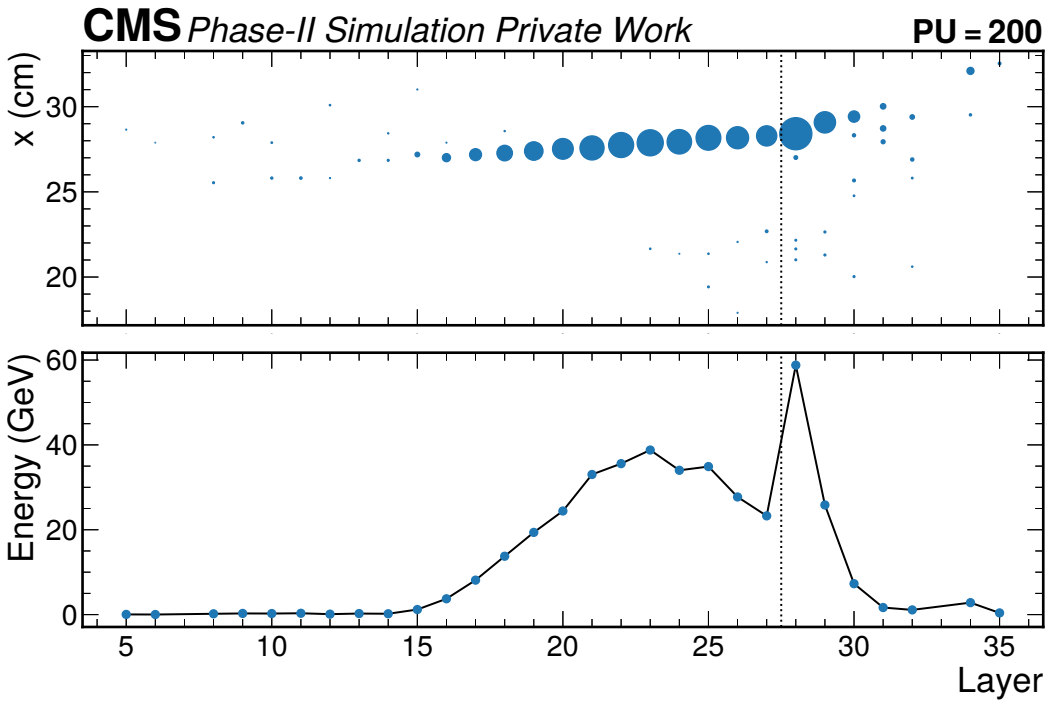

Figure 7.14: Event display of an unconverted photon (top) and the energy profile of the most energetic LCs per layer (bottom). The trackster is reconstructed with CLUE3D. All dots in the plots correspond to LCs belonging to the reconstructed trackster. The radii of the 2D LCs are proportional to their energy. The vertical dotted black line shows the boundary between CE-E and CE-H. The knee of the electromagnetic shower after entering the CE-H is due to the greater intra-layers distance in the hadronic section, which produces this illusory effect when looking at the layer numbers.

recompute the shower axis by using only the LCs belonging to the cleaned tracksters, but the difference between the two axes was minimal and did not justify a second round of PCA.

The cleaning algorithm is able to provide higher-quality reconstructed tracksters by rejecting PU contributions that are mistakenly associated to the electromagnetic shower. The most striking effect is the improvement of the energy resolution, as shown in Fig. 7.17. The energy resolution and response is evaluated by fitting the $E_{\text{trackster}}/E_{\text{GEN}}$ distribution with a Cruijff function. The Cruijff function is a centred Gaussian with different left-right resolutions and non-Gaussian tails:

$$f(x) = \exp\left(\frac{(x-\mu)^2}{2\sigma_{L,R}^2 + \alpha_{L,R}(x-\mu)^2}\right) \tag{7.9}$$

where σ_L and α_L are used when $x - \mu < 0$, and σ_R and α_R when $x - \mu > 0$. The response is identified as μ, while the resolution $\sigma = (\sigma_L + \sigma_R)/2$. The effect of PU is to degrade the energy resolution, with asymmetric distributions and prominent rightward tails. After the cleaning procedure, the distributions are more symmetric, narrower, and more Gaussian. These features will be essential for a robust energy regression and application of scale factors. Such improvements are mainly evident at low energy, for photons below 250 GeV. On average, below 100 GeV, the resolution improves up to 15%. On the other hand, at higher energies the effects of the cleaning are negligible.

An additional beneficial outcome of the cleaning algorithm is its positive impact on the direction estimation, as shown in Fig. 7.18. The PCA provides a powerful mathematical

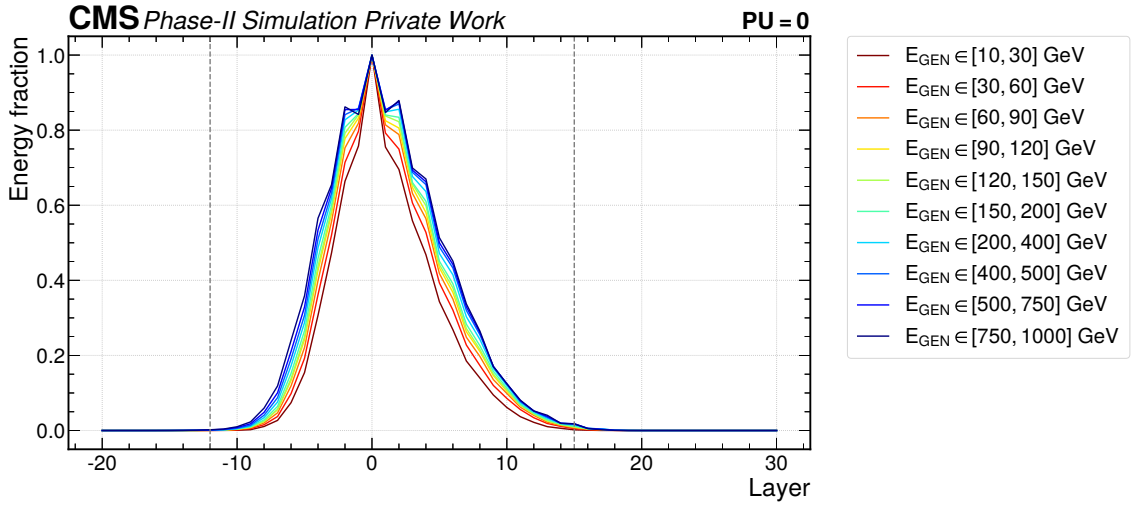

Figure 7.15: Longitudinal energy extension of an unconverted photon as a function of the energy of the generated photon. The points are averaged among all tracksters in each energy bin. The number of layers is from/to the most energetic one, i.e., layer number 0 corresponds to the peak of the electromagnetic shower. The vertical grey dotted lines are drawn at Layer 15 and Layer 12, which is the range used in the cleaning algorithm.

tool for computing the axis of a shower in a sampling calorimeter. Notably, the axis determined following the cleaning algorithm aligns more closely with the line defined by propagating the incident photon. This improvement is more pronounced in regions where PU is more aggressive, such as the high-eta region and for low-energy photons. The possibility of relying on a robust estimation of the shower direction will be a revolution in CMS. The current ECAL is a homogenous calorimeter and lacks the capability to estimate the direction of the original particle's propagation. This added quality of the HGCAL can be exploited to improve some reconstruction steps (cfr Ch. 8) or as a handle to access exotic physics signatures - for instance, the search for long-lived particles decaying to a pair of photons. A hallmark of this process is the presence of two photons in the electromagnetic calorimeter that do not point to the PV.

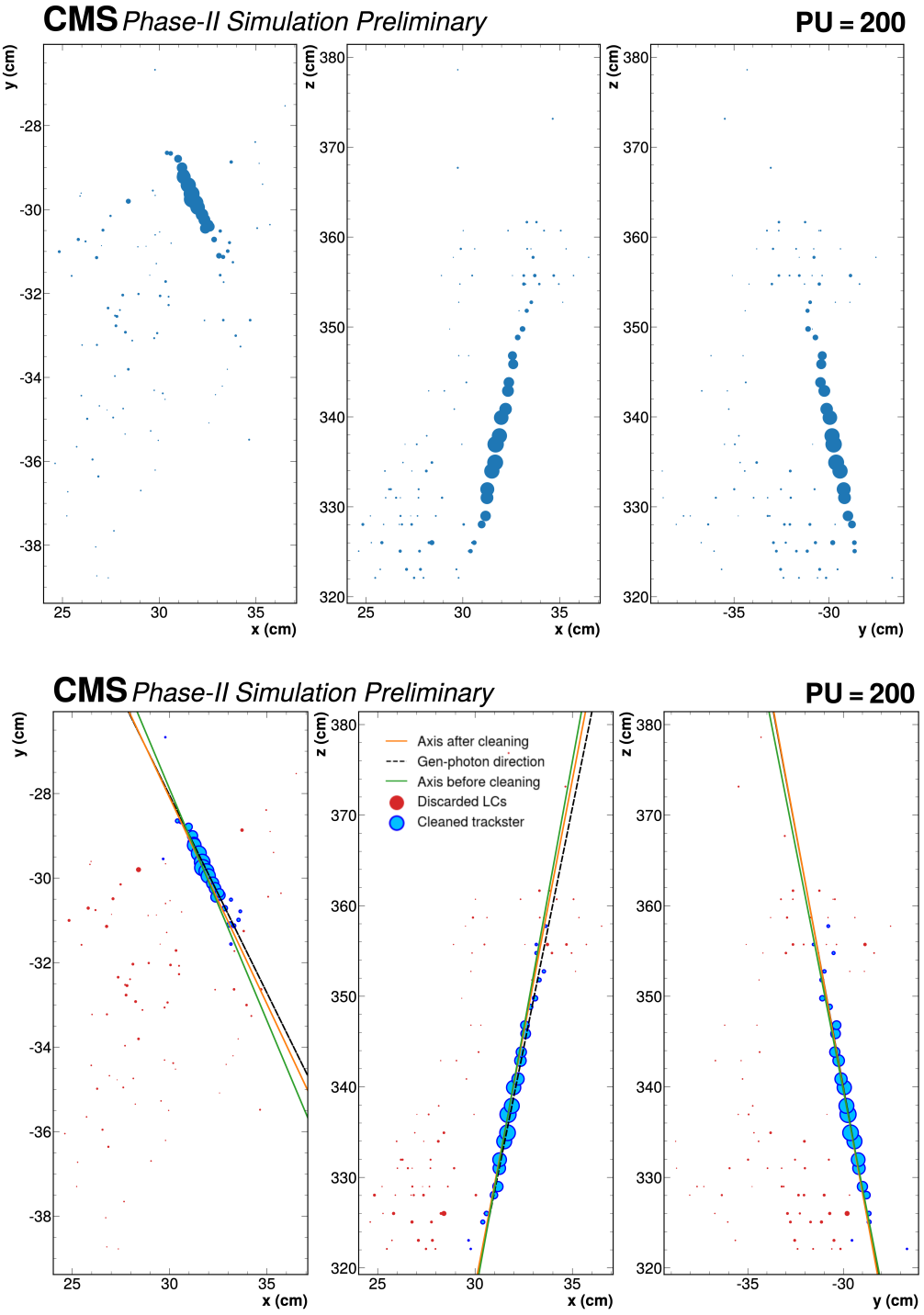

Figure 7.16: Event display of an unconverted photon reconstructed with CLUE3D before (top) and after (bottom) the cleaning algorithm. All circles in the plots correspond to LCs belonging to the reconstructed trackster and the radii are proportional to their energy. The reconstructed energy ($E_{\text{RECO}} = 361$ GeV) provides an overestimate of around 20% compared to the generated energy ($E_{\text{GEN}} = 293$ GeV). The energy after the cleaning ($E_{\text{CLEAN}} = 288$ GeV) is much closer to the simulated one. The PCA axis that is used in the cleaning algorithm is displayed (orange) as well as the main energy weighed PCA axis computed before the cleaning with all 2D LCs (green). The direction of the generated photons is displayed with a dotted black line. The angle between the generated photon and the estimated shower axis is 0.60° before the cleaning and 0.28° after the cleaning.

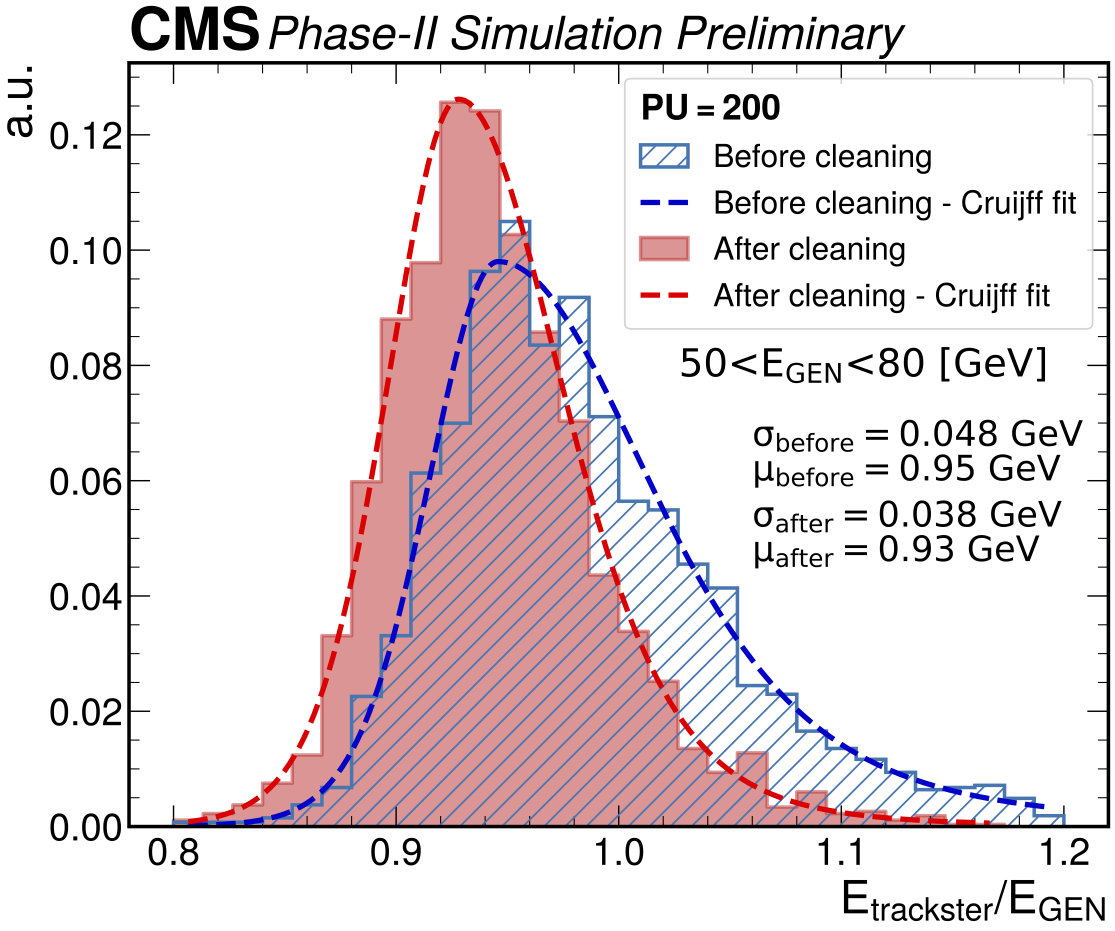

Figure 7.17: Energy response in the energy bin $E_{\text{GEN}} \in [50, 80]$ GeV before and after the cleaning with CLUE3D tracksters. No energy corrections are applied. Before the cleaning procedure the CLUE3D distribution is broader and features a more prominent rightward tail. This effect is more evident at low energy. After the cleaning procedure the distribution becomes more symmetric and narrower.

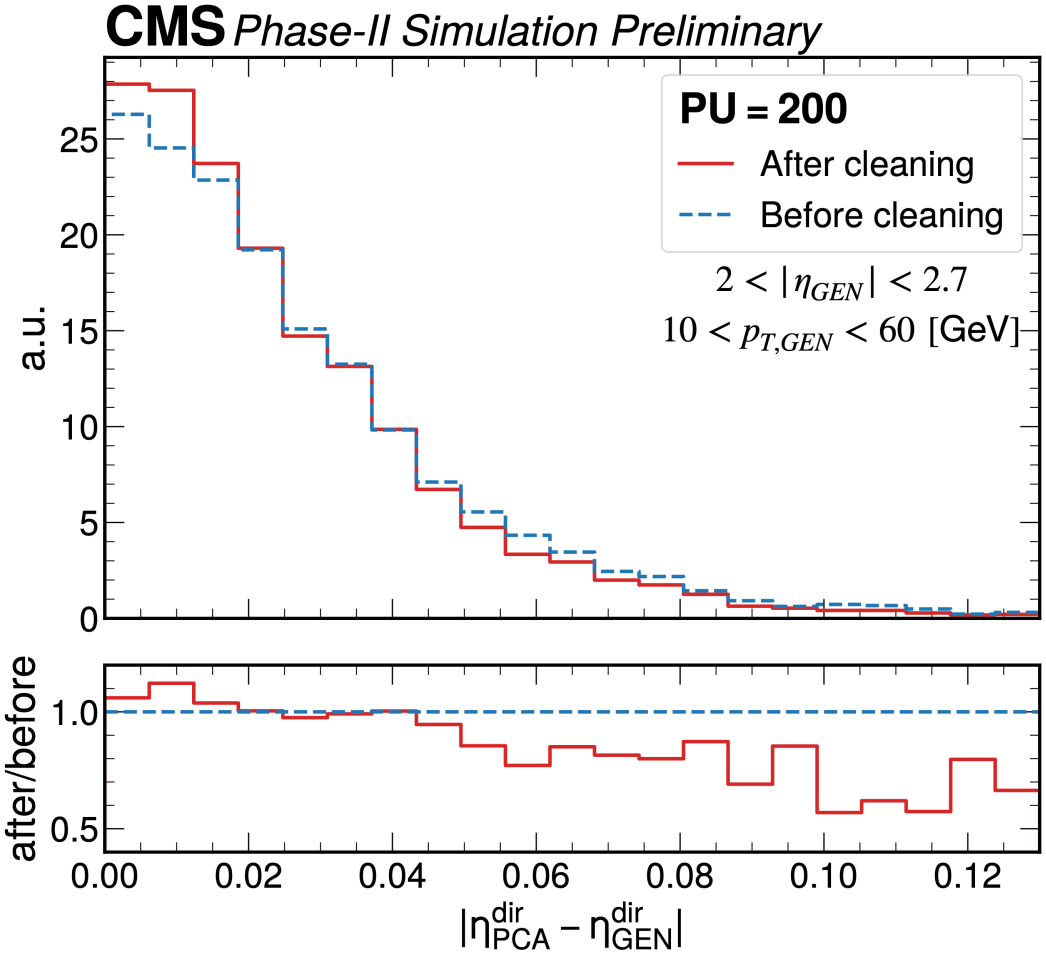

Figure 7.18: Difference in the trackster direction estimate ($\eta^{\mathrm{PCA}}_{\mathrm{dir}}$) from the generated photon direction ($\eta^{\mathrm{GEN}}_{\mathrm{dir}}$) before and after the cleaning in the range $2 < |\eta_{\mathrm{GEN}}| < 2.7$ and $10 < p_{\mathrm{T,GEN}} < 60$ [GeV]. Trackster direction estimates are obtained with energy-weighed PCA using the LCs inside the trackster. The estimation of the direction is closer to the generated particle after the cleaning, which shows that cleaning can improve the information on the trackster direction. Tracksters are formed by the CLUE3D algorithm.

Summary

This chapter assessed the shower reconstruction of electromagnetic objects in the HGCAL using the TICL framework through unconverted photons. The energy-based CLUE3D pattern recognition algorithm outperforms the previously used geometrical CA algorithm for shower reconstruction, supporting the decision to implement the former in the new releases of TICL.

The in-depth analysis of the reconstructed showers revealed the presence of low-energy, irregularly shaped satellite tracksters surrounding well-reconstructed main showers. Their presence is more prevalent in the high-density region of the calorimeter. However, their minor kinematic impact does not jeopardise significantly the reconstruction of the primary electromagnetic shower. A dedicated TICL linking plug-in could be developed to recover these energy contributions. Additionally, these studies unveiled for the first time the presence of split showers. These events are characterised by an electromagnetic shower reconstructed as two independent objects due to the presence of a secondary peak deeper within the HGCAL. Investigating the origins of this phenomenon is ongoing, but its extreme rarity does not pose a setback for the development of TICL.

One of the challenges encountered in the shower reconstruction in the expected dense, high-PU environment at the HL-LHC is the contamination from PU-induced contributions. This contamination can lead to degradation in energy resolution and overall object quality, particularly for photons with energy below 250 GeV. To mitigate this effect, a PU reduction algorithm was developed. This algorithm provides higher-quality showers with a beneficial effect on the energy resolution. In addition, it supplies a robust estimation of the shower axis through PCA.

This shower-cleaning algorithm was designed to be versatile, making it easily applicable in diverse situations. It exploits the unique features of the HGCAL to reduce PU contamination effectively and enhance the quality of reconstructed objects.

Chapter 8

Development of the electron superclustering in the HGCAL

This chapter is devoted to assessing and developing the electron superclustering in the HGCAL. Prior to reaching the surface of the electromagnetic calorimeter, either the current ECAL or the future HGCAL, an electron propagates through the material preceding the calorimeter. During this propagation, it may interact with the material, leading to the emission of Bremsstrahlung photons. These photons can subsequently convert into an electron-positron pair, initiating a cascade of such events analogously to what happens inside an electromagnetic shower. Thus, when the electron reaches the calorimeter, it may no longer be a single particle but could consist of a shower of electrons and photons giving raise to different clusters in the calorimeter. These clusters must be combined in one single *supercluster* in order to reconstruct the energy and direction of the original electron.

Section 8.1 presents the MC samples used throughout this chapter. Section 8.2 introduces the physics features of electron emission, illustrating how the CMS collaboration has coped with the superclustering during the various data-taking periods. The poor PU resilience of the current geometrical algorithm and the fact that it does not exploit the full potential of the HGCAL will be a limitation during phase-2. This thesis proposes a novel superclustering algorithm based on Deep Neural Networks (DNNs) designed explicitly for the HGCAL. An overview of DNNs, tailored for the objectives of this chapter, is given in Sec. 8.3. The DNN-based superclustering is then presented in Sec. 8.4.

8.1 Monte Carlo samples for electron studies

Similarly to the previous chapter, the studies presented here are all based on full `GEANT4` MC simulations, following the steps outlined in Sec. 2.5. The GEN step is represented by a particle gun generating either an electron or a positron per endcap. The number of electrons and positrons in each endcap is the same to avoid bias due to charge imbalances. The generation is uniform in pseudorapidity between 1.65 and 2.75, in transverse momentum between 2 and 100 GeV, and uniform over the entire range of azimuthal angles ϕ. The electron gun is positioned in a way mimicking the expected distribution of primary vertices, simulating a real-life scenario. In `GEANT4`, the HGCAL is simulated using the 26-layer geometry for the CE-E. The electron showers are reconstructed using TICLv4, i.e., with CLUE3D as 3D pattern recognition algorithm. In the following, an event will be considered as either one electron or positron in one endcap. The simulation is repeated twice: first, in the absence of PU, and second, in the presence of PU. In the latter generation, a PU scenario at a centre-of-mass energy of 14 TeV is used, with the number of true interactions distributed following a Poisson distribution with a mean equal to 200.

8.2 From rectangles to moustache

This section presents a historical overview of the superclustering algorithms employed by the CMS collaboration. Starting with the description of the phenomenology of electron emission, it is then explained how these features have been employed over the years for the purpose of the superclustering.

It is worth emphasising that any algorithm must only be based on information from the calorimeter. As explained in Sec. 2.3.1, the superclusters are needed to seed the GSF tracks. Starting from Run 2, the tracker information has been used in a subsequent step to refine the supercluster [91]. However, for the needs of the CMS PF framework, the superclustering algorithm can only use information from the calorimeter.

8.2.1 Phenomenology

As already outlined in the introduction, electrons have a significant probability of showering when traversing the CMS inner tracker. The interplay between Bremsstrahlung emissions and pair productions gives rise to intricate patterns, as shown in Fig. 8.1. The original single electron arrives at the face of the calorimeter as a combination of photons, electrons, and positrons. In addition, inelastic nuclear interactions may also occur during the propagation, leading to the release of hadrons in the showers. However, this hadronic process is extremely rare and does not enter the scope of the superclustering process.

The inner region of the CMS detector is immersed in the solenoid magnetic field of 3.8 Tesla generated by the CMS magnet. As a result, electron emissions are mainly spread along the ϕ direction, as electrons and positrons travel in helicoidal trajectories following the magnetic field lines. The bending in the η direction is less important and becomes more pronounced as the electron energy diminishes.

Superclustering can be considered the art of collecting (relatively) low-energy contributions. As illustrated in Fig. 8.2, the probability of emitting a Bremsstrahlung photon with a sizeable fraction of the original electron energy is very small. For instance, the probability of having an emission with more than 60% of the initial electron energy is only 5%. The typical pattern observed at the surface of the HGCAL consists of a collection of particles, where one stands out due to its higher energy compared to the others, which possess lower energy. When not accounting for the most energetic particle, in our MC samples, 68% (95%) of particles crossing the HGCAL boundary have a transverse energy below 0.8 GeV (18 GeV). Although Bremsstrahlung emissions are more likely to take away a small fraction of energy from the electron, there may be many subsequent emissions along the trajectory, leaving the original electron with a small fraction of its original energy. During phase-2, it is anticipated that nearly half of the electrons will lose more than 75% of their energy before reaching the HGCAL. Therefore, a robust superclustering algorithm is essential to prevent degradation in the electron energy resolution.

8.2.2 Run 1: Rectangular algorithm

The features of electron emissions outlined above have been utilised for developing various superclustering algorithms in the history of CMS. In the Run 1, the formation of superclusters relied on geometrical algorithms that took advantage of the characteristic narrow width along η and broader spread in ϕ [208, 209]. Throughout the whole Run 1, the algorithm underwent slight modifications and improvements, yet the fundamental concept remained unchanged. The strategy implemented during the data collection at 8 TeV is presented below.

In the barrel, the *hybrid algorithm* was employed. The first step was the definition of the seed crystal, defined as the one with the largest transverse energy above 1 GeV.

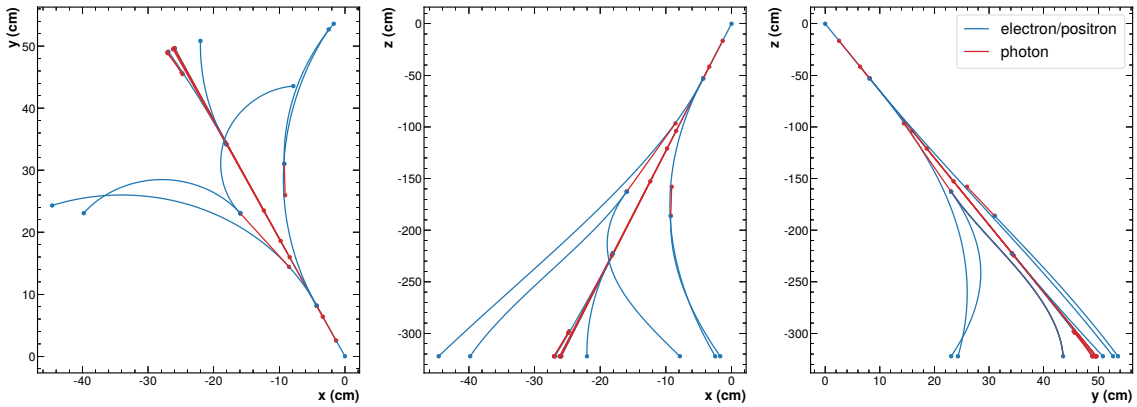

Figure 8.1: Representation of the trajectories of electron emissions before reaching the HGCAL in the transverse plane (left), horizontal plane (centre), and vertical plane (right). The electron is generated at $(x, y, z) = (0, 0, 0)$. Each point represents either a Bremsstrahlung emission or a pair production. The final points that are not connected to any other lines represent the entrance into the HGCAL, located at $|z| \simeq 321$ cm.

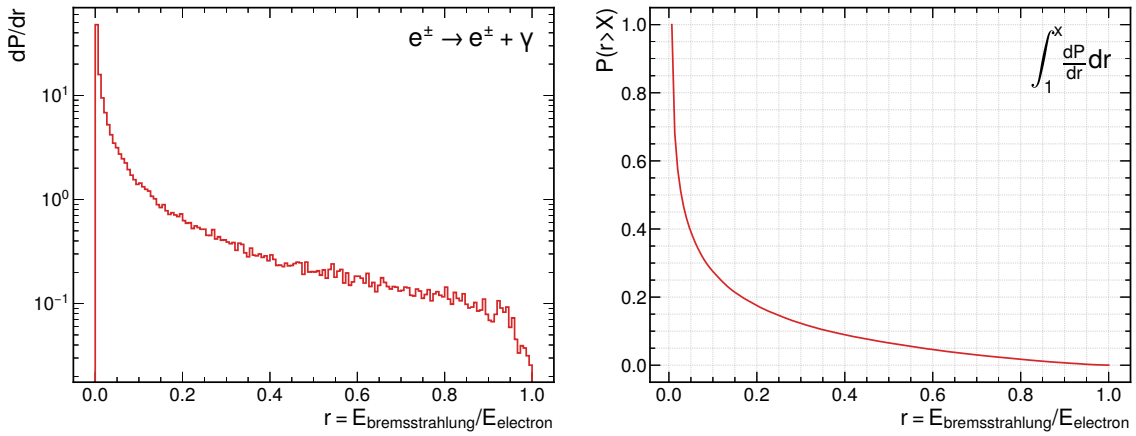

Figure 8.2: (Right) Probability of emission of a Bremsstrahlung photon with an energy equal to a fraction r of the energy of the original electron. (Left) The corresponding cumulative distribution function.

This seed was designated as the supercluster seed. Energy clusters were searched within a fixed rectangular window around this seed. The search window consisted of a sequence of 5×1 arrays of crystals spanning $\eta \times \phi$ with a range of 17 crystals in both directions of ϕ. Arrays meeting the energy criterion above 0.1 GeV were retained. Subsequently, the contiguous arrays were grouped into separate clusters and linked to the supercluster seed if the energy of their seeds, defined as the crystal with the largest transverse energy, was above 0.35 GeV.

In the endcap, the *multi-5×5 algorithm* was used. The seed crystals were defined as those exhibiting local energy maxima surpassing 0.18 GeV in transverse energy. Starting from the seed with the largest E_{T}, clusters of 5×5 crystals were defined. These clusters were then combined if their transverse energy exceeded 1 GeV and fell within an $\eta \times \phi$ region of 0.07×0.3 rad around each seed crystal. Subsequently, the position of each cluster was then extrapolated to the ECAL preshower in order to recover the energy deposited there.

8.2.3 Run 2 and Run 3: Moustache algorithm

In the transition from Run 1 and Run 2, the CMS collaboration decided to switch to a dynamic superclustering to mitigate PU contamination, which had increased significantly over time due to the continuous increase in luminosity. This algorithm leverages the kinematic dependence of the geometrical features of electron emissions. The new procedure is applied to already-formed energy clusters, namely PF clusters in the ECAL, thus forming superclusters as clusters of clusters. This algorithm is still in use in the ongoing Run 3 and, without the commissioning of new algorithms, is the candidate for the CMS phase-2. This thesis presents the first implementation and evaluation in the HGCAL.

Although the primary bending occurs along the ϕ direction, electrons also exhibit a curvature along η, which becomes more pronounced at higher pseudorapidity. Moreover, the curvature radius in the magnetic field depends on the particle's momentum, resulting in a more pronounced bending for electrons with lower energy. Hence, the usage of a fixed rectangular window was limiting the performance, as it suffered from PU contamination in the low-η region, missed some Bremsstrahlung emissions in the high-η region, and not exploiting the anticipated curvature magnitude for electrons with specific momentum. These kinematic features are clearly visible in the distributions of $\Delta\eta = |\eta| - |\eta_{\text{seed}}|$ versus $\Delta\phi = \phi - \phi_{\text{seed}}$, as reported in Fig. 8.3 for the HGCAL. The definition of seed is extended from the algorithms used in Run 1, representing the cluster with the highest transverse energy. The ϕ and η coordinates of a cluster are identified by computing the energy-weighed barycentre of the shower. The influence of the magnetic field in both the η and ϕ directions results in the positioning of electron emissions in a moustache-like shape, which is why this algorithm was named *Moustache*.

The Moustache algorithm employs a dynamic parabolic-shaped window in $\Delta\eta$, which emulates the bending of electrons in the $\eta-\phi$ plane, combined with a dynamic $\Delta\phi$ window, which accounts for the varying electron curvatures based on their respective momenta.

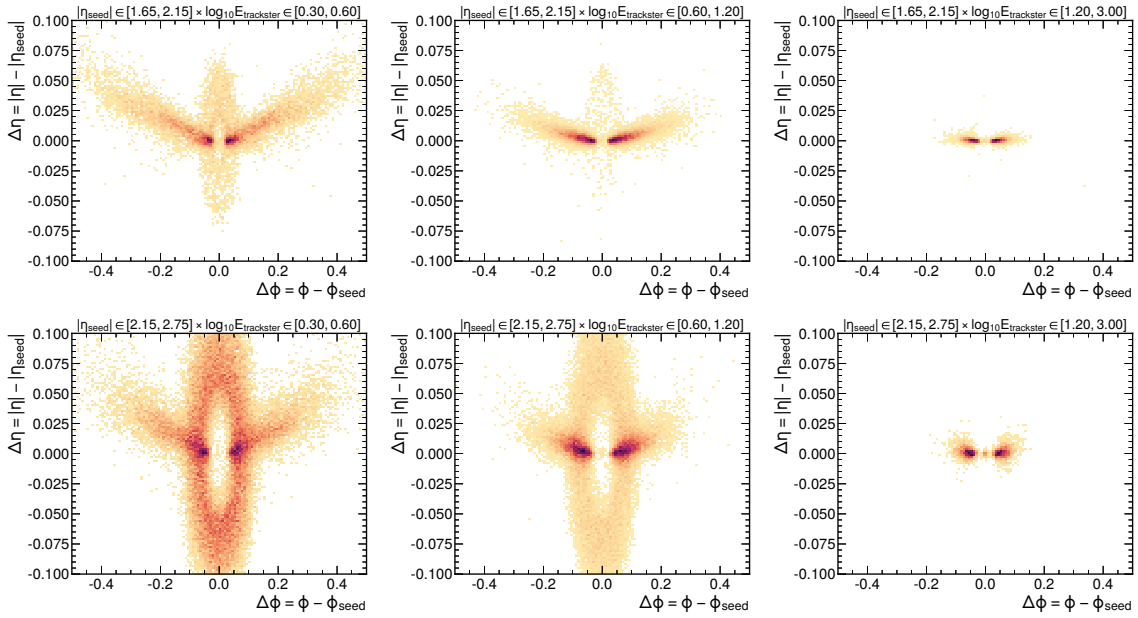

Figure 8.3: Distribution of $\Delta\eta = |\eta| - |\eta_{\text{seed}}|$ versus $\Delta\phi = \phi - \phi_{\text{seed}}$ for electrons in the absence of PU in bins of pseudorapidity of the seed (η_{seed}) and energy of the tracksters ($E_{\text{trackster}}$). The intensity of the red color indicates the occupancy of the bin. The upper plots show the distributions in the low-density region of the HGCAL, while the bottom plots in the high-density region.

Seed definition The seed is defined as the cluster with the largest transverse energy among the ones with $E_{\mathrm{T}} > 1$ GeV. Once the seed is defined, the algorithm starts looping over all clusters in the event.

$\Delta\eta$ parabolas The $\Delta\eta$ region considered in the superclustering is defined by two parabolas:

$$\Delta\eta_{\mathrm{dn}}(\Delta\phi; \eta_{\mathrm{seed}}, E_{\mathrm{cluster}}) < \Delta\eta < \Delta\eta_{\mathrm{up}}(\Delta\phi; \eta_{\mathrm{seed}}, E_{\mathrm{cluster}}) \tag{8.1}$$

Each parabola is a function of $\Delta\phi$ and depends on the pseudorapidity of the seed (η_{seed}) and the energy of the cluster under study (E_{cluster}):

$$\Delta\eta_{\mathrm{dn,up}}(\eta_{\mathrm{seed}}, E_{\mathrm{cluster}}) = A_{\mathrm{dn,up}}(\eta_{\mathrm{seed}}, E_{\mathrm{cluster}}) \cdot (\Delta\phi)^2 + B_{\mathrm{dn,up}}(\eta_{\mathrm{seed}}, E_{\mathrm{cluster}}) \tag{8.2}$$

The $A_{\mathrm{dn,up}}$ and $B_{\mathrm{dn,up}}$ parameters are empirically parametrised as a function of η_{seed} and $\log_{10}(E_{\mathrm{cluster}})$.

$\Delta\phi$ window The $\Delta\phi$ window delimits a region in the transverse plane around the seed. As the dispersion along ϕ depends only on the momentum of the particle, the requirement can be written as:

$$|\Delta\phi| < \Delta\phi_{\mathrm{window}}(E_{\mathrm{T,cluster}}) \tag{8.3}$$

Here, $\Delta\phi_{\mathrm{window}}$ is defined as:

$$\Delta\phi_{\mathrm{window}} = y_{\mathrm{offset}} + \frac{s}{1 + \exp[(\log_{10} E_{\mathrm{T,cluster}} - x_{\mathrm{offset}}) \cdot w]} \tag{8.4}$$

In this equation, y_{offset}, s, x_{offset}, and w denote four parameters that are optimised in bins of η_{seed}. The $\Delta\phi_{\mathrm{window}}$ is not left to float completely unconstrained as a function of the transverse momentum of the cluster, but a cutoff and saturation scale are introduced to define the maximum and minimum extension of the window:

$$\Delta\phi_{\mathrm{window}} = \min(\Delta\phi_{\mathrm{window}}, \mathrm{cutoff}) \tag{8.5}$$

$$\Delta\phi_{\mathrm{window}} = \max(\Delta\phi_{\mathrm{window}}, \mathrm{saturation}) \tag{8.6}$$

Both the cutoff and saturation scales are also optimised in bins of η_{seed}. All the parameters are optimised with an envelope fit of the $|\Delta\phi| - \log_{10} E_{\mathrm{T}}$ distribution using Eq. 8.4 in order to contain 98% of the $\log_{10} E_{\mathrm{T}}$ of the original electrons in each $|\Delta\phi|$ bin.

Superclustering If the cluster being examined falls within the regions delineated by the $\Delta\eta$ parabolas and $\Delta\phi$ window, it is combined with the seed. After the iteration over all clusters is completed, if there remains at least one cluster with $E_{\mathrm{T}} > 1$ GeV, this cluster is promoted as seed, and the iteration is repeated to form another supercluster.

An example of the Moustache algorithm on a simple case without considering PU can be seen in Fig. 8.4.

A noticeable, unexpected feature of the distributions reported in Fig. 8.3 is the presence of a circular halo surrounding the seed. Actually, in the figures, the halo appears as an elongated oval along the $\Delta\eta$ direction. However, this visual distortion is solely due to the different scales of the x and y axes of the plots. The halo is, in fact, a perfect circle.

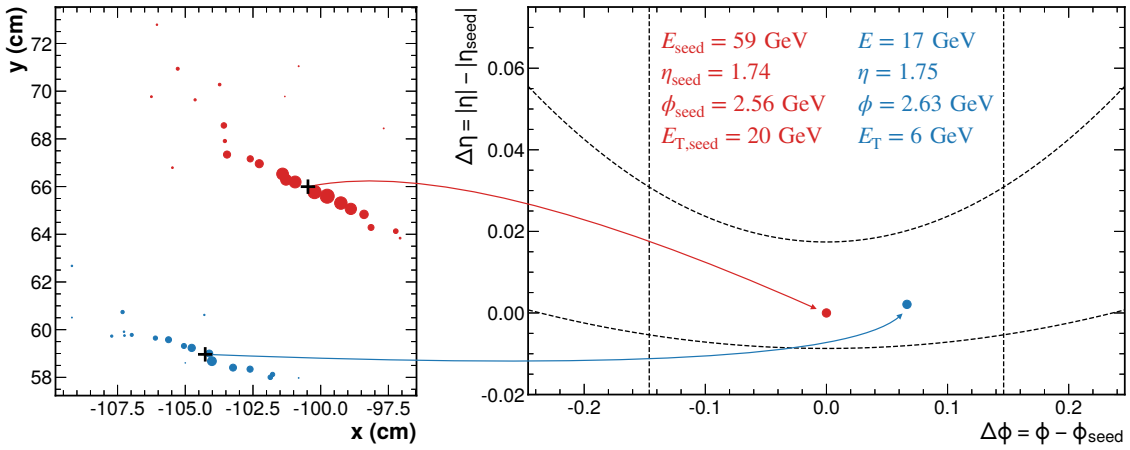

Figure 8.4: (Left) Transverse projection of an electron event in the HGCAL in the absence of PU. Each dot represents a LC, and the size is proportional to its energy. Diverse tracksters are distinguished by unique colors. The black *plus* symbols pinpoint the barycentres of the trackters. (Right) Representation of the same event in the $\Delta\eta = |\eta| - |\eta_{\text{seed}}|$ versus $\Delta\phi = \phi - \phi_{\text{seed}}$ plane. The black dotted lines represents the $\Delta\eta$ parabolas and $\Delta\phi$ window resulting from the Moustache algorithm. Since the trackster falls within the intersection between the two regions, it is combined with the seed, forming a supercluster.

This halo becomes significantly more prominent in the high-density region of the HGCAL, with the halo predominately populated by low-energy tracksters. These features are better summarised in Fig. 8.5. Events that contain at least one trackster in the halo tend to have seeds located in the high-density region and with larger E_{T}. Tracksters forming the halo are characterised by very low energy and a deeper starting point for the shower in the calorimeter. A cut at $E_{\text{trackster}} = 15$ GeV completely eliminates the halo, with 95% of the events with energy below 5.8 GeV. These characteristics are distinctive of the satellite tracksters described in Sec. 7.3. Consequently, the halo is a direct consequence of energy that fails to get *trackstered* into the seed but is part of the seed particle shower.

8.3 Deep Neural Networks in a nutshell

Deep learning is a subfield of both machine learning[1] and artificial intelligence[2] that focuses on building and training neural networks to perform a wide variety of tasks. The structural building block of neural networks is the *perceptron*, also known as *neuron*, illustrated graphically in Fig. 8.6. The perceptron takes m inputs $(x_1, x_2, \ldots, x_m)$, each of which is multiplied by a weight w_i. The results of these multiplications are then summed up, with the inclusion of a bias term. This bias term can be seen as an additional input with a value equal to 1 and a weight equal to w_0. The outcome of the summation is then fed into a non-linear activation function g. The purpose of activation functions is to introduce non-linearity into the network, which are essential to approximate arbitrarily complex functions. Mathematically, this can be written as:

$$\hat{y} = g\left(w_0 + \sum_{i=1}^{m} x_i w_i\right) \tag{8.7}$$

where $\hat{y}$ represents the output of the perceptron.

[1] Ability to learn without explicitly being programmed
[2] Any technique that enables computers to mimic human behaviour

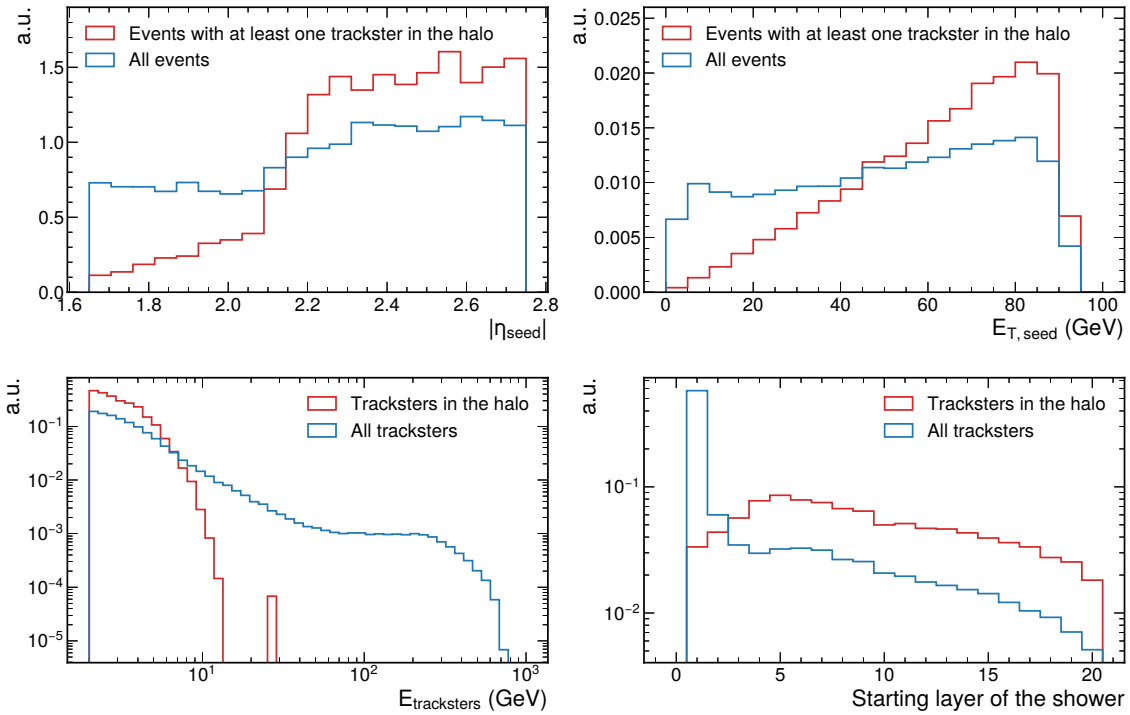

Figure 8.5: (Top left) Absolute value of the pseudorapidity of the seeds for events with at least one trackster in the halo. (Top right) Transverse energy of the seeds for events with at least one trackster in the halo. (Bottom left) Energy of the tracksters populating the halo. (Bottom right) Starting layer for showers populating the halo. To isolate the contribution of the halo, a rectangular window is used, defined as $|\Delta\eta| < 0.025$ and $|\Delta\phi| < 0.01$. This window ensures a good isolation of the halo without incurring contamination from the *moustache*.

A Deep Neural Network (DNN) comprises multiple layers of interconnected perceptrons, where each neuron in a layer takes the outputs of the preceding layer as its inputs. A generic example of feedforward fully-connected DNN architecture is reported in Fig. 8.7. This architecture is categorised as a feedforward DNN because it does not include loops, information flows from the inputs, though the hidden layers, and finally to the outputs. The output of a perceptron does not influence its own input. It is also a fully connected DNN because every neuron receives the output of every node of the preceding layer as input. Consequently, the generic output of a perceptron i at layer k can be expressed as:

$$z_{k,i} = w_{0,i}^{(k)} + \sum_{j=1}^{n_{k-1}} g(z_{k-1,j}) w_{j,i}^{(k)} \tag{8.8}$$

At this stage, the previously defined DNN is simply a sequence of neurons that take a set of m variables as inputs, execute some mathematical operations, and generate an output. The idea of neural networks is to use the inputs, i.e., the *features* of a dataset, to predict the values of the *target* variables. This is achieved by first training the DNN on a dataset where the values of the target variables are already known. In order to perform a training, there needs to be a way to determine if the DNN is making mistakes and how large this mistake is. The training aims to tune the weights of the DNN in order to minimise this mistake. In deep learning jargon, these mistakes are usually called *losses* and evaluated with a *loss function*, which takes as input the output of the DNN and the true prediction:

$$\text{Loss}(f(\boldsymbol{x}^{(i)}; \mathbf{W}), \boldsymbol{y}^{(i)}) \tag{8.9}$$

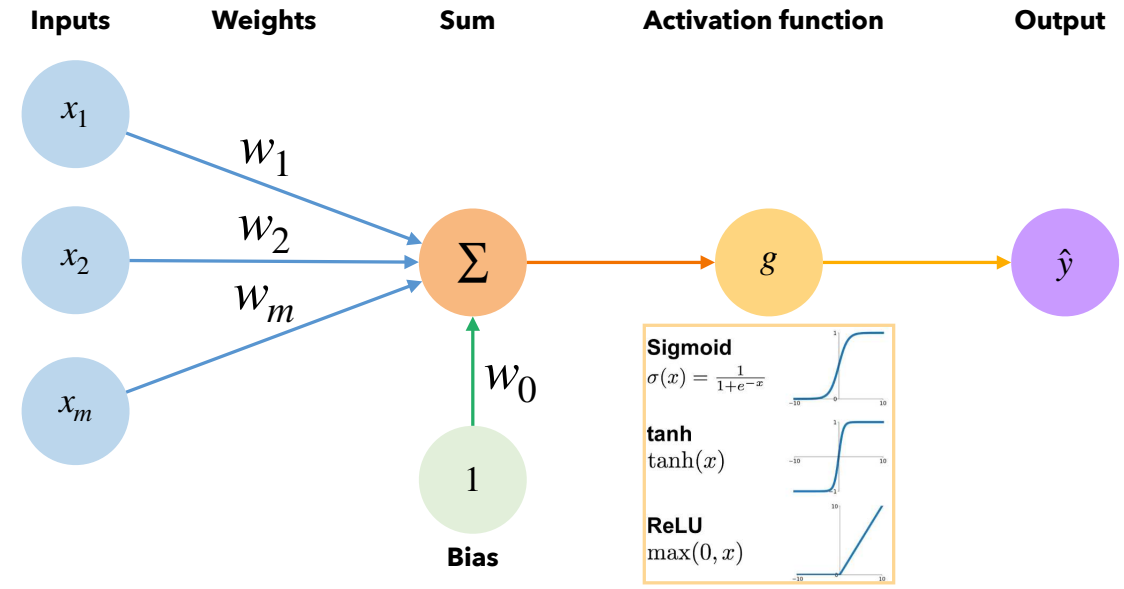

Figure 8.6: Illustration of a perceptron. The input features are multiplied by a weight and then summed together. The bias can be regarded as an additional input feature with value equal to one. The result of the sum is passed to a non-linear activation function (g). The yellow box reports some of the most used functions.

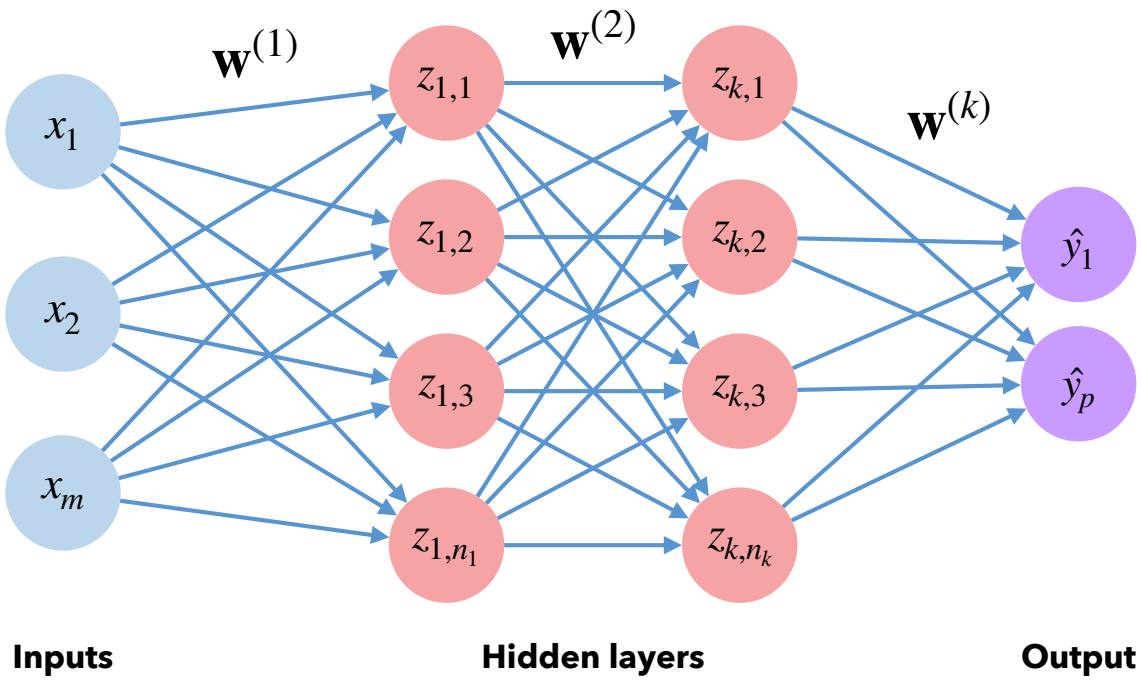

Figure 8.7: Generic illustration of a feedforward fully-connected Deep Neural Network with m inputs, k hidden layers, n_k nodes per layer, and p outputs. The output of each perceptron does not influence its own input, and every neuron receives the output of every node of the preceding layer as input.

The loss function defined above is valid for the i-th point of the dataset. However, the ultimate target of the training process is to minimise, on an average basis, the mistakes that the neural network makes over the entire dataset. This broader aim is encapsulated by the *empirical loss*, which quantifies the cumulative loss spanning the entirety of the dataset:

$$\boldsymbol{J}(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n} \text{Loss}(f(\boldsymbol{x}^{(i)}; \mathbf{W}), \boldsymbol{y}^{(i)}) \tag{8.10}$$

where n represents the number of points in the dataset. During the training procedure the network weights are tuned in order to achieve the lowest empirical loss:

$$\mathbf{W}_{\text{optimal}} = \arg\min_{\mathbf{W}} \boldsymbol{J}(\mathbf{W}) \tag{8.11}$$

The minimisation of the empirical loss is done using the *gradient descent* procedure, summarised as follows.

1. Weights are initialised randomly.
2. Evaluate the loss function at the current weight values.
3. Calculate the gradient of the loss function with respect to the set of weights $\frac{\partial \boldsymbol{J}(\mathbf{W})}{\partial \mathbf{W}}$. The gradient points in the direction of the fastest increase in the loss function.
4. Update the weights at the new point $\mathbf{W} - \eta \frac{\partial \boldsymbol{J}(\mathbf{W})}{\partial \mathbf{W}}$, where η is the *learning rate*.
5. Iterate over steps 2 and 4 until convergence.

By repeatedly computing gradients and adjusting weights, this procedure gradually guides the DNN's weights to values that minimise the loss function. The learning rate η expresses the size of the step to take in each training iteration. While small learning rate converges slowly and may get stuck in local minima, large learning rates may overshoot the optimal weights, become unstable and diverge. The right learning rate is the one that allows a smooth convergence, avoiding local minima. There are now a series of adaptive learning rate algorithms that tune the learning rate at each step depending on various aspects, such as how large the gradient is, how fast learning is happening, or the size of particular weights.

The evaluation of the gradient $\partial \boldsymbol{J}(\mathbf{W})/\partial \mathbf{W}$ can be computationally intensive because it is calculated as a sum over all the points in the dataset, which can be very large. The *Mini-batch gradient descent* addresses this computational challenge by dividing the training dataset into smaller subsets called *mini-batches*. In each iteration, the algorithm computes the gradient of the loss function using only the data points in the selected mini-batch.

After one training iteration over a mini-batch dataset, the loss function is evaluated on an independent dataset, known as *validation dataset*, to validate the neural network performance with the updated weights. The anticipated outcome is for the training and validation losses to exhibit a similar descending trend. This pattern indicates that the neural network is learning from the training dataset to describe also brand-new, unseen data. A common issue during neural network training is the phenomenon of *overfitting*. Overfitting arises when a model learns to describe extremely well the training dataset but fails to generalise to other data. In other words, an overfitted model fits the training data too closely, capturing not only the underlying patterns but also the noise or random fluctuations present in the data. An indicator of overfitting is the divergence between the validation loss and the training loss. An increase of only the validation loss suggests a loss of generality in the predictions of the neural network. Overfitting is avoided by

implementing regularisation techniques. One of the most common is the *early-stopping*. The basic idea behind early stopping is that the risk of overfitting increases as the model continues to be trained. Therefore, training is halted after a predetermined number of iterations, usually set by the user; if there is no further improvement in the validation loss function.

The performance of DNN models in the case of classification problems can be comprehensively evaluated using four quantities: True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP). The first word refers to the matching between the actual and predicted label, and the second word refers to the predicted label. These quantities are usually summarised in the *confusion matrix*, as shown in Fig. 8.8. The matrix entries are commonly normalised by the sum of elements in each row, providing the correlated rates.

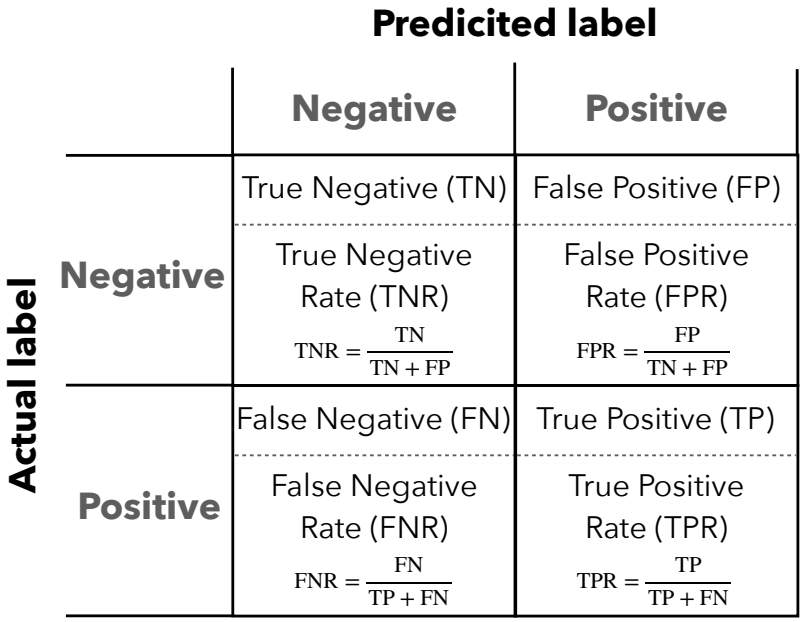

Figure 8.8: Prototype of a confusion matrix for a classification problem. When the entries of the matrix are normalised by the sum of elements in each row, one obtains the corresponding rates, reported in the bottom section of each entry block.

8.4 DNNsuperclustering

The Moustache algorithm has been used successfully for the past Run 2 and the ongoing Run 3. Being Moustache a purely geometrical algorithm, it features a high signal efficiency, but it is naturally subject to the cluster of PU contributions and, to a lesser degree, of electronic noise. Since during phase-2 the PU will increase to unprecedented levels, up to 200, this moderate PU resilience could degrade significantly the resolution. In addition, Moustache will not use the full potential of the HGCAL. As being conceived for the current CMS ECAL, it only exploits the relative positions of the barycentres of the showers. However, HGCAL will provide much more information that could be exploited to improve the superclustering algorithm. With these premises, this thesis investigates a new superclustering algorithm based on DNN, explicitly tailored for the new calorimeter, which will be referred to in the following as *DNNsuperclustering*.

The methodology employed in the DNNsuperclustering resembles the logic of the Moustache approach. The algorithm initiates with selecting a seed trackster, defined as the one

with the largest E_{T}. Subsequently, the score of the DNN is evaluated for each trackster, and those tracksters with a DNN score above the predefined working point are associated with the seed to progressively form a supercluster.

Similarly to the previous superclustering approaches, also the potential algorithms for the HGCAL are required to use only the information associated with the calorimeter. The strategy of seeding the GSF tracks with the supercluster will be unaltered during the CMS phase-2.

8.4.1 Target variable

In order to train the DNN, it is essential to define the target variable of the model carefully. In this case, it is the set of tracksters matching the energy deposits induced by the original genuine electron. The gen-matching for electrons cannot rely on simple geometrical conditions as for unconverted photons of the previous chapter. Electrons bend in the magnetic field and create several spatially separated tracksters. Thus, the gen-matching condition evolves by taking into account the energy contributions in each trackster. The fraction of the trackster energy associated with the generated electron can be evaluated using the following score, denoted as *gen-score*:

$$\text{gen-score (\%)} = \frac{\sum_i^{\text{recHits}} E_i \cdot f_i^{\text{ele}}}{E_{\text{trackster}}} \cdot 100 \tag{8.12}$$

In this equation, the sum is over all recHits forming the trackster under study, and f_i^{ele} is the fraction of the recHit energy associated with one of the particles originating from the generated electron. The threshold for the gen-matching is chosen to be 60%, as it is the value that maximises the energy resolution over the entire energy range, as shown in Fig. 8.9. Lower thresholds degrade the resolution at high-E_{T}, while higher thresholds degrade the resolution at low-E_{T}. At high-E_{T}, there is the significant formation of satellite tracksters which can receive sizeable PU contributions. By loosing the gen-score threshold, these tracksters are collected and the resolution degrades. On the other hand, low-E_{T} tracksters are more contaminated by PU; therefore, increasing the gen-score threshold suppresses many tracksters, and the resolution worsens.

8.4.2 Skimming procedure

TICLv4 implements a single CLUE3D iteration, reconstructing all tracksters in the event without any attempt to identify the type of shower. This operation will be performed by one or more algorithms that are currently under development, which are usually named Particle IDentification (PID) algorithms. In the context of superclustering, these kinds of algorithms will be beneficial to select only tracksters likely to have originated from electromagnetic objects. As this PID is not yet available, a cut-based, loose skimming procedure is implemented for the studies presented in this chapter. This procedure consists of a series of quality requirements on the tracksters:

- the trackster must have at least 3 LCs in the CE-E;
- $E_{\text{trackster}} > 2$ GeV;
- explained variance ratio of the trackster greater than 95% if $E_{\text{trackster}} > 50$ GeV, otherwise it must be greater than 92%.

The first condition regarding the number of layers is a basic requirement for an electromagnetic shower. While late showers are possible, it is improbable that a photon or an electron does not release any energy in the electromagnetic compartment. The number is

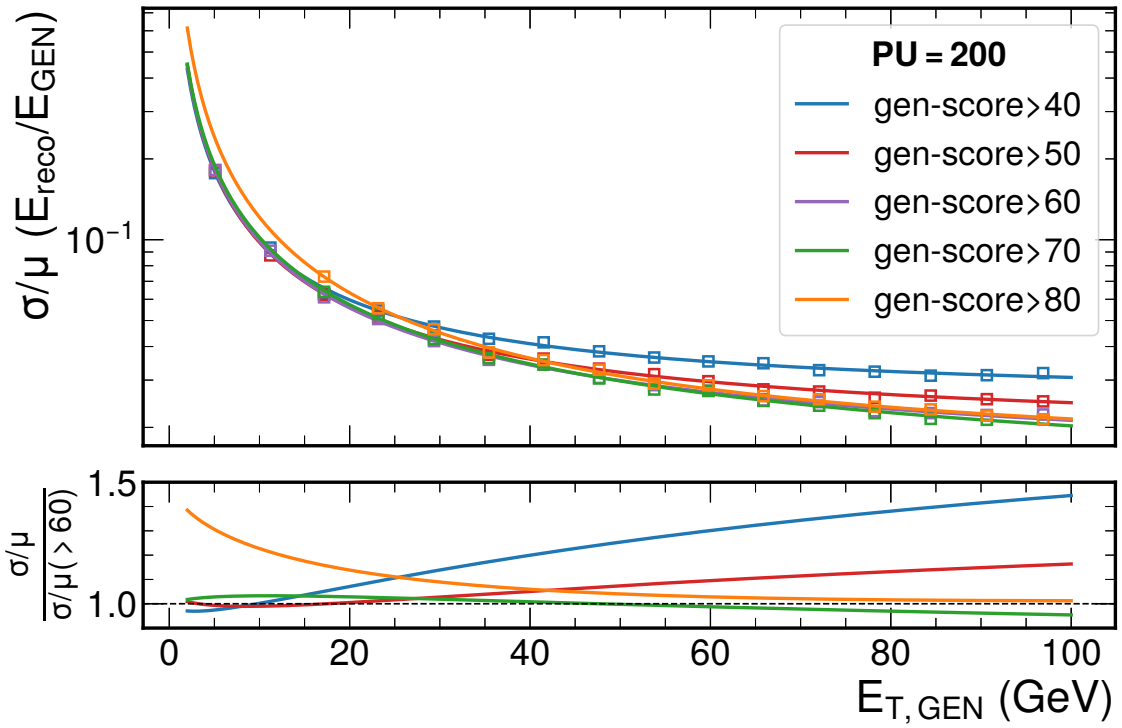

Figure 8.9: Energy resolutions for electrons generated in the presence of PU for different threshold values of the gen-score. The energy scale (μ) and resolution (σ) are computed by fitting the $E_{\text{reco}}/E_{\text{GEN}}$ distribution with a Cruijff function, and the values are reported in the plot as empty squares. The continuous lines represent the fit of the relative resolutions with the function $f(E_{\text{T}}) = a/\sqrt{E_{\text{T}}} + b/E_{\text{T}} + c$. The bottom plot shows their ratio with respect to the curve corresponding to a gen-score greater than 60%.

set to 3, representing the minimum number of points necessary to compute the PCA in three dimensions.

The second condition regarding the energy of the trackster aims to reduce the contribution of the most problematic satellite tracksters described in Sec. 7.3. Their recovery is not the main target of the superclustering and should be addressed at CLUE3D-level. This specific aspect is more extensively addressed in the subsequent sections. By setting an energy threshold of 2 GeV, equivalent to the energy released by a MIP traversing the entire HGCAL, a significant fraction of the lowest-quality of such tracksters is effectively eliminated.

The last condition pertains to the explained variance ratio. The explained variance is a quantity associated with the PCA computation (cfr Sec. 7.4.1). The PCA axes are ordered by their corresponding variance. Mathematically, PCA axes are identified as the eigenvectors of the covariance matrix, and the corresponding eigenvalues are identified as the explained variances. Geometrically, this quantity can be seen as the sum of the projected distances on the PCA axis from the centre of the dataset, as shown in Fig. 8.10. It is evident from the figure that a rigid translation of the points in a direction perpendicular to the axis does not alter the explained variance. The situation is however different for the other PCA axes as their corresponding variance varies during this translation. This information is therefore encapsulated in the explained variance ratio, defined as the explained variance of an axis divided by the sum of the explained variances of all other axes. For our purposes, the explained variance ratio can serve as a metric to estimate the dispersion of LCs around the shower axis. A higher explained variance ratio indicates LCs are more aligned to the axis. A typical electromagnetic shower is expected to have

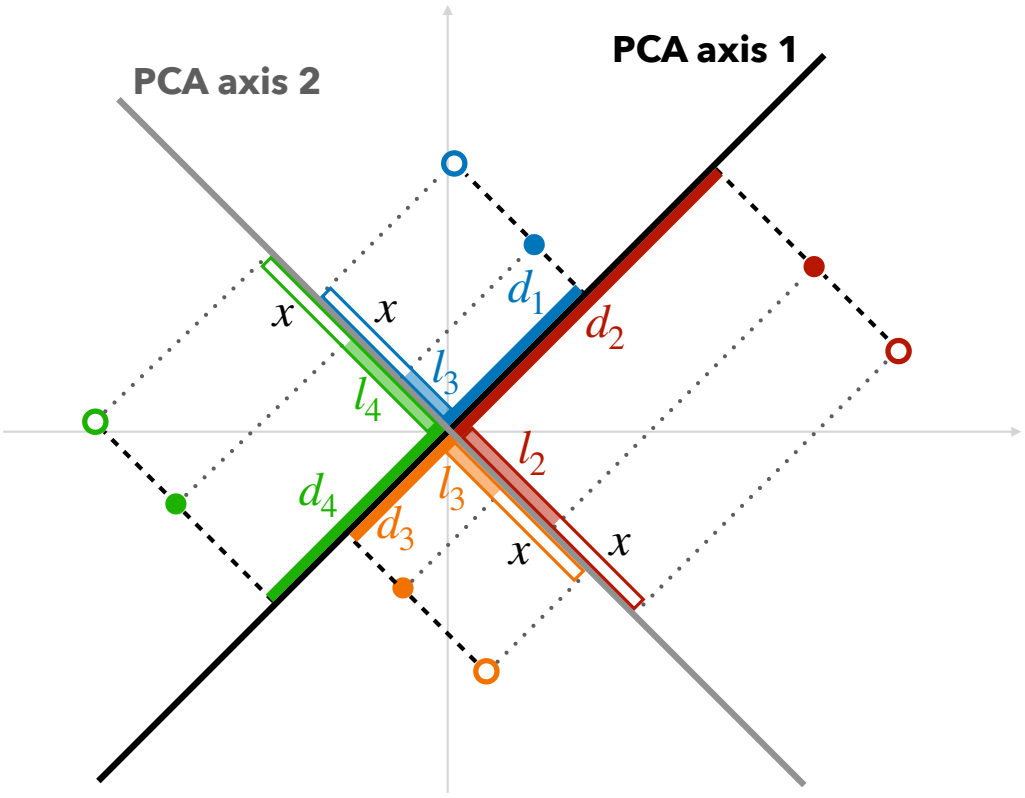

Figure 8.10: Geometrical representation of the explained variance in a two-dimensional case. The black and grey lines represent the PCA axes. The coloured, filled dots depict the points in a generic dataset (in our case, the 2D Layer Clusters). Each point is projected onto PCA axis 1. The quantities d_i are the distances between the projected points and the centre of the dataset. The sum of all d_i corresponds to the eigenvalue of the respective PCA eigenvector. The variance can be calculated as $\sum_i d_i^2/(n-1)$. Similarly, for the PCA axis 2, the variance is calculated as $\sum_i l_i^2/(n-1)$. The coloured, empty dots represent the points after being translated by a fixed amount x in a direction perpendicular to PCA axis 1. The variance with respect to PCA axis 1 remains unchanged since the distances d_is do not change. In contrast, the variance with respect to PCA axis 2 increases to the value $\sum_i (l_i + x)^2/(n-1)$.

energy deposits well aligned along the shower axis. A larger dispersion of LCs, hence a lower explained variance ratio, is a symptom of non-electromagnetic tracksters.

8.4.3 Input variables

Two distinct DNN models are defined using separate sets of input variables. The DNNv1 model only exploits the kinematic observables already employed in the Moustache algorithm. These observables include the difference in pseudorapidity $\Delta\eta$ and azimuthal angle $\Delta\phi$ between the barycentres of the tracksters and the seed, as well as the kinematic quantities of the seed and the other tracksters. Essentially, this model can be regarded as a direct DNN-based extension of the Moustache algorithm. Instead, the DNNv2 model, in addition to the variables used in the DNNv1, integrates novel variables leveraging the properties of the HGCAL.

In Sec. 7.4, it was shown that the cleaning algorithm for electromagnetic tracksters provides a shower axis, computed as the main energy-weighed PCA axis, that is able to estimate the direction of propagation of the shower. This quantity, made available by the HGCAL characteristics, can be exploited to find additional geometrical patterns that can

help reject PU contributions. When dealing with axes, the most straightforward operation to do is to compute angles using the scalar product between the corresponding unit vectors $\hat{v_1}$ and $\hat{v_2}$:

$$\theta = \arccos(\hat{v_1} \cdot \hat{v_2}) \tag{8.13}$$

In addition to the angle, also its projections onto the various planes can be exploited. Since the bending is primarily along ϕ, the projection onto the transverse $x - y$ plane will give a more important discrimination power than the angle itself.

The DNNsuperclustering revolves around the seed, trying to find patterns relative to this object. The computation of the shower axis also allows defining a new reference frame around the seed itself to bring to light specific patterns. The *seed frame* is identified by the unit vectors ($\hat{x}_{\text{seed}}$, $\hat{y}_{\text{seed}}$, $\hat{z}_{\text{seed}}$) defined as follows:

- $\hat{z}_{\text{seed}}$: the shower axis of the seed;
- $\hat{x}_{\text{seed}} = \hat{z}_{\text{seed}} \times \hat{z}$;
- $\hat{y}_{\text{seed}} = \hat{x}_{\text{seed}} \times \hat{z}_{\text{seed}}$.

With this definition of the seed frame, the $z_{\text{seed}} - y_{\text{seed}}$ plane contains the beam-pipe, hence the z-axis of the CMS frame (defined by the unit vector $\hat{z}$). The angle computed in Eq. 8.13, being a scalar, is invariant in the new frame, but its projections onto the planes are not. In the seed frame, the projections on the $x_{\text{seed}} - y_{\text{seed}}$ plane cannot be computed as the seed axis boils down to a point.

In addition to these angular observables, two other quantities related to the shower axis are included. As outlined in Sec. 8.4.2, the explained variance and the explained variance ratio are robust observables to select electromagnetic showers and discard PU contributions. In order to increase the PU resilience of the DNNv2 model, these two variables are also included.

Since the DNNsuperclustering is based on the same seed definition of the Moustache algorithm, it is not necessary to search for tracksters in the entire volume of the HGCAL. Profiting from the expected geometrical distribution of tracksters around the seed (cfr Fig. 8.3), the search area is restricted to a rectangle defined as $|\Delta\eta| < 0.1$ and $|\Delta\phi| < 0.5$. This region is anticipated to contain all tracksters associated with an electron. The plots for all the variables inside this geometrical region are reported in Fig. 8.11.

It is good practice when training neural networks to scale each future to a given range [210]. The advantages of this procedure are multifold and some of them are reported in the following. The first advantage is a faster convergence of the loss function. Neural networks, especially those using gradient-based optimisation techniques, are sensitive to the scale of input features. When features are on vastly different scales, it can lead to slow convergence or the network getting stuck in local minima. Scaling ensures that the optimisation process is more stable and efficient. Large input values can also lead to numerical instability during computation, especially in the intermediate layers of a DNN. This can result in issues like vanishing/exploding gradients or NaN (not-a-number) values during training. Scaling mitigates these problems, making the training process more robust. In addition, scaling can improve the generalization of a model. When input features are scaled to a similar range, the model can more easily identify relevant patterns in the data, as it does not have to learn different scaling factors for each feature independently. Hence, scaling can be regarded as a regularisation procedure preventing overfitting. For our purposes, the input features are scaled in the range between 0 and 1 as follows:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{8.14}$$

where x and x' denote the original and scaled value, respectively. This procedure simply scales the values without changing the shape and preserves zero entries in sparse data.

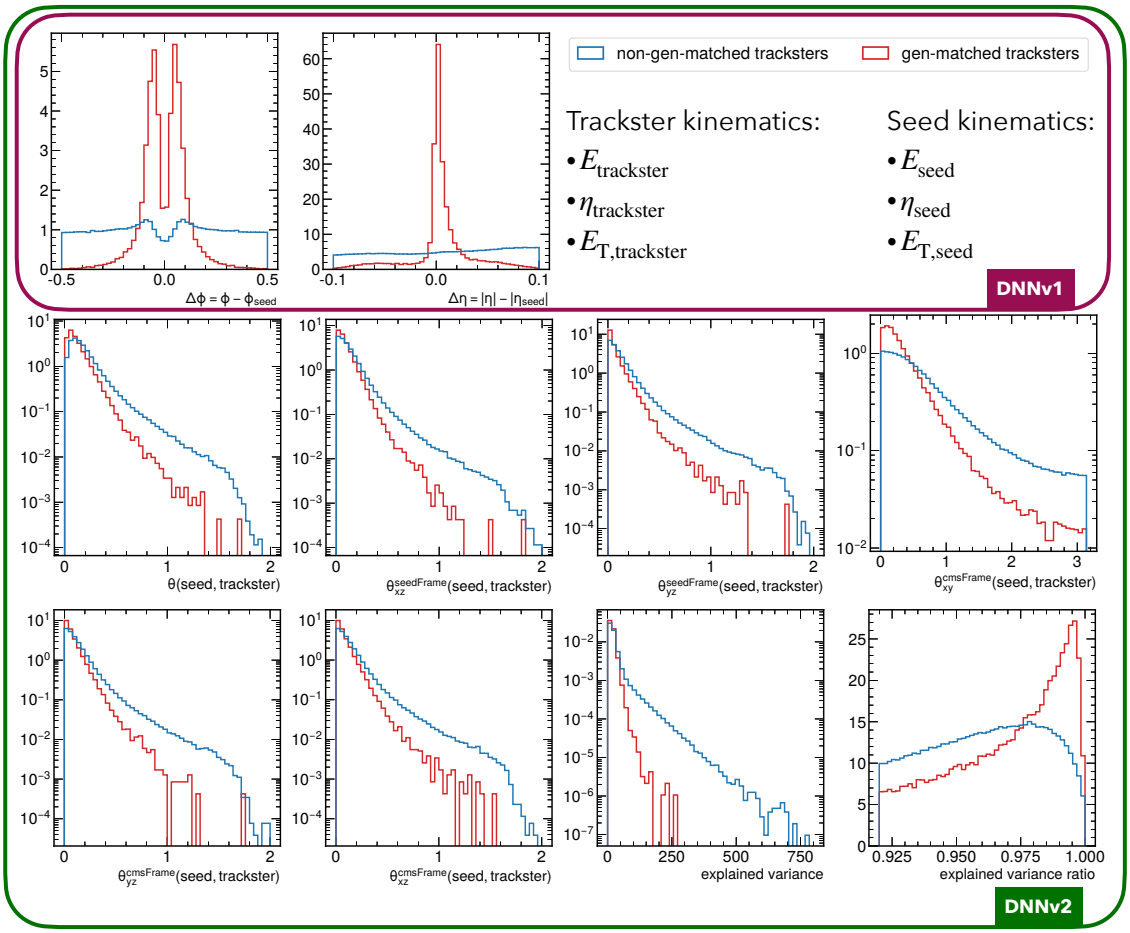

Figure 8.11: Input variables for the DNNv1 and DNNv2 models. The distributions are computed in a rectangular region around the seed defined as $|\Delta\eta| < 0.1$ and $|\Delta\phi| < 0.5$.

8.4.4 DNN architecture, training, and performance

The DNN architecture is the same for both models: a feedforward fully-connected DNN with 3 hidden layers with 100, 200, and 100 neurons per layer, respectively. Since the superclustering is a binary classification problem, the output layer is composed of a single neuron. This configuration results in 41,301 weights to optimise for the DNNv1, and 42,101 weights for the DNNv2[3]. In case of forward networks, the most common activation function for hidden layers is the Rectified Linear Unit (ReLU), which is zero for negative values and increases linearly for positive values:

$$\text{ReLU}(x) = \max(0, x) \tag{8.15}$$

The output of the ReLU is in the range of $[0, +\infty)$. The last layer of a network for binary classification is characterised by a sigmoid activation function:

$$\text{sigmoid}(x) = \frac{1}{1 + \text{e}^{x}} \tag{8.16}$$

with the output in the range of $[0, 1]$. For binary classification, where the output is a probability between 0 and 1, the empirical loss function (cfr Eq. 8.10) is commonly chosen

[3]If N is the number of input features, the number of weights is given by $(N \cdot 100 + 100) + (200 \cdot 100 + 200) + (200 \cdot 100 + 100) + (100 + 1)$. For each layer, the number of inputs is multiplied by the number of neurons and added to the number of biases. The number of bias entries is equal to the number of neurons.

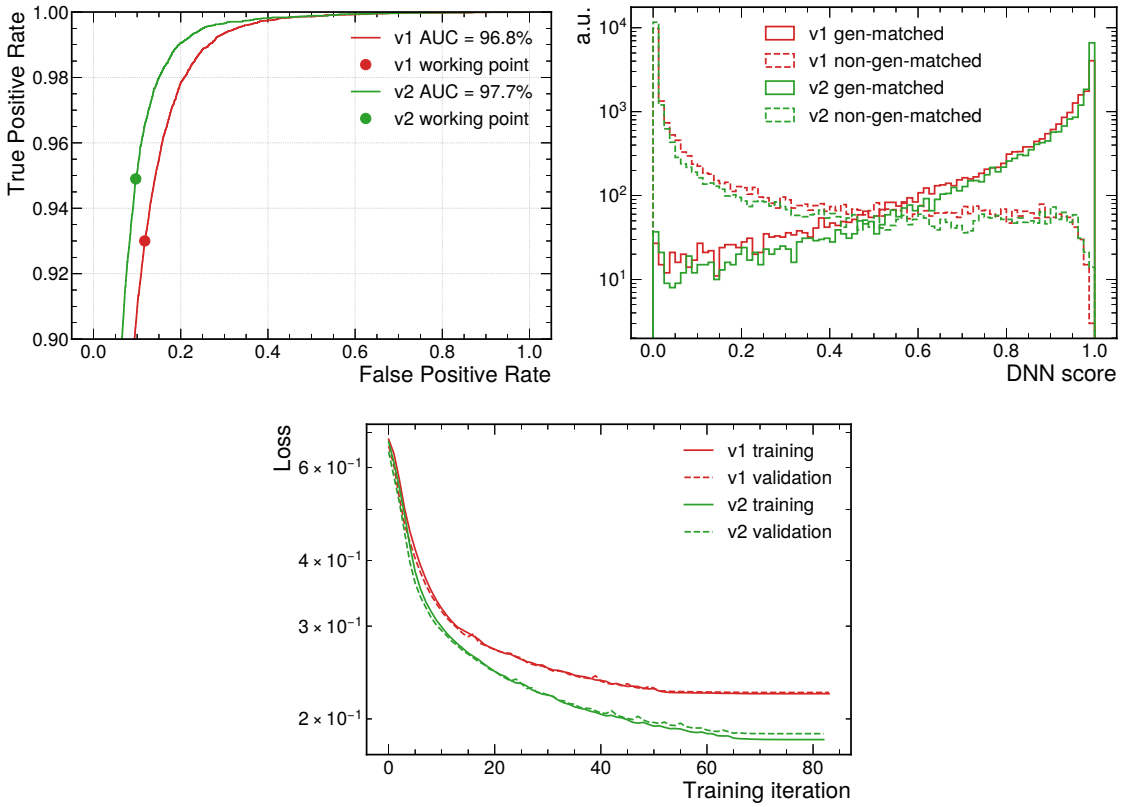

Figure 8.12: Results of the training and validation process for the DNNv1 and DNNv2 models. (Top Right) ROC curves. The dots represent the working point, chosen to maximise the difference between the true positive rate (signal efficiency) and false positive rate (PU contamination). The area under the ROC curve is reported in the legend. (Top Left) DNN score for gen-matched, drawn with solid lines, and not gen-matched tracksters, drawn with dotted lines. (Bottom) Training and validation empirical losses as a function of the training iteration.

to be the *binary cross entropy*:

$$\boldsymbol{J}(\mathbf{W}) = -\frac{1}{n}\sum_{i=1}^{n} y^{(i)} \log(f(x^{(i)};\mathbf{W})) + (1-y^{(i)}) \log(1-f(x^{(i)};\mathbf{W})) \tag{8.17}$$

The handcrafted variables defined in Sec. 8.4.3 allows the implementation of a simple, lightweight architecture, as the one outlined above. Being faithful to the principle of the *Occam's razor*[4], this setup makes these DNN models a better candidate to be integrated in the CMS software.

The learning rate is optimised during the training procedure using A Method for Stochastic Optimisation (ADAM) [211]. The early-stopping regularisation method is implemented to prevent overtraining: the training is halted after 10 training iterations with no improvement.

For the definition of the training and validation dataset, all tracksters are combined together, the seeds are removed, and the training and validation dataset are extracted randomly from this collection. In order to prevent any bias in the training procedure, the same amount of gen-matched and not gen-matched tracksters are selected in the datasets.

[4] *Pluralitas non est ponenda sine necessitate* (plurality should not be posited without necessity). The principle gives precedence to simplicity: of two competing theories, the simpler explanation of an entity is to be preferred.

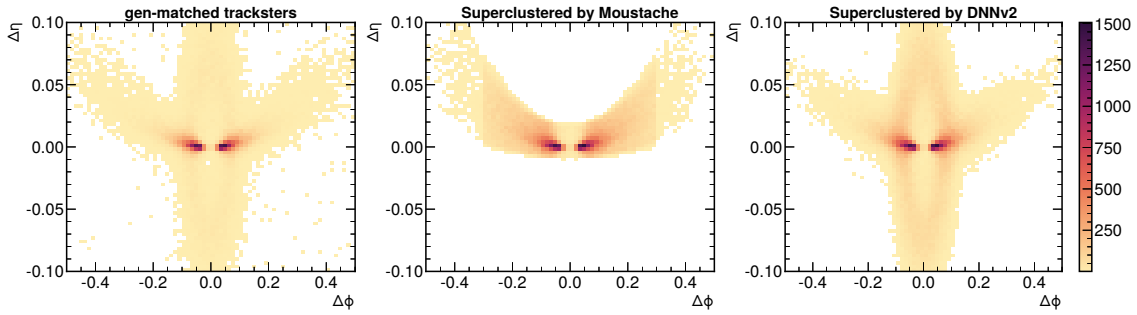

Figure 8.13: Distributions of the $\Delta\eta = |\eta| - |\eta_{\text{seed}}|$ versus $\Delta\phi = \phi - \phi_{\text{seed}}$ variables for gen-matched trackster (left), tracksters forming superclusters using the Moustache algorithm (centre), and tracksters forming superclusters using the DNNv2 algorithm (right). The seed is defined as the trackster with the largest transverse energy.

The training is carried out with 80,000 trackster, while the validation dataset is composed of 40,000 tracksters.

The results of the training and validation procedures are reported in Fig. 8.12. Both the training and validation loss functions exhibit a smooth descending trajectory, remaining in close agreement with each other, which rules out the occurrence of overtraining. The working point is chosen to maximise the number of correctly clustered gen-matched tracksters, referred to as the *trackster collection efficiency* or *signal efficiency*, while minimising the number of erroneously collected not gen-matched tracksters, indicating the *PU contamination*. Using ML terminology, the objective is to maximise the difference between the TPR and FPR (cfg Fig. 8.8). By sheer coincidence, this working point corresponds to a DNN score of 0.52 for both models. The performance attained by both DNN models is high, with an area under the ROC curve exceeding 95%, and a good discrimination power between the two type of tracksters. The already satisfying performance of DNNv1 is further increased by including the HGCAL angular variables in the DNNv2, as will be discussed later.

A visual comparison between the performance of the DNNv2 model and the existing Moustache algorithm, serving as a benchmark, is made through the examination of the $\Delta\eta - \Delta\phi$ distributions reported in Fig. 8.13. The distinctive geometric nature of the Moustache is apparent, as well as its difficulty in catching the nuances of the pattern of electron emissions. Notably, the Moustache algorithm completely fails to cluster contributions in the circular halo surrounding the seed. The geometrical window of the Moustache hardly cut out this contribution. In contrast, the DNNsuperclustering effectively identifies electron emissions within the anticipated region and properly recovers contributions from the halo. Regarding the halo, the DNNsuperclustering is slightly aggressive in this region, which can be populated by poorly-formed tracksters, prone to being misclassified as originating from PU.

Being the superclustering a binary classification problem, the confusion matrices shown in Fig. 8.14 embed all the information to characterise the different models. As a general trend, the matrices become more diagonal, going from Moustache to DNNv1 and finally to DNNv2, proving the better performance of the latter over the other models. Among the matrix entries, the most meaningful quantities for the superclustering are the TPR (signal efficiency) and FPR (PU contamination). While the migration from the geometrical Moustache to the related DNN-based algorithm, represented by the DNNv1, does not significantly improve the FPR, its true strength lies in boosting the TPR by a noteworthy 26%. The PU contamination is instead reduced by introducing the HGCAL angular observables in the DNNv2, causing an 18% reduction in the FPR. The interplay between

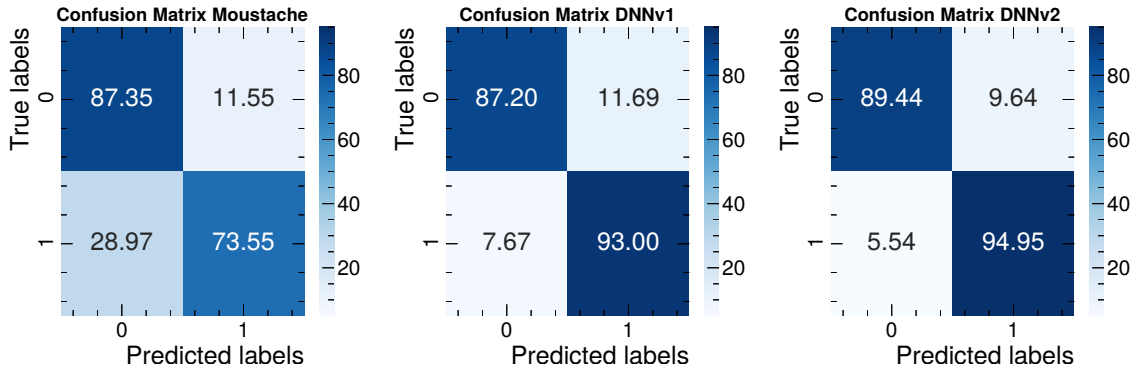

Figure 8.14: Confusion matrices for the Moustache algorithm (left), the DNNv1 model (centre), and the DNNv2 model (right). The label "1" indicates gen-matched tracksters, while the label "0" indicated not gen-matched tracksters. The True Positive Rate (bottom right entry) denotes the signal efficiency. The False Positive Rate (top right entry) estimates the PU contamination. The True Negative Rate (top left entry) denotes the PU rejection. The False Negative Rate (bottom left) is also referred to as background efficiency.

ML techniques and features of the new HGCAL offers the possibility to improve the superclustering in view of the CMS phase-2.

Since these quantities are essential to judge the performance of the superclustering algorithm, Fig. 8.15 presents the TPR and FPR across different bins of the transverse energy of the seed and within the two distinct density regions of the HGCAL. Across all kinematic regimes and detector regions, the utilisation of DNN-based algorithms yields an overall enhancement in signal efficiency. Particularly noteworthy are the results achieved with the DNNv2 model, where the signal efficiency reaches an average of 90% in the high-density region and 98% in the low-density region. The signal efficiency and the presence of PU contamination directly influence the energy resolution of the supercluster, which is one of the most important figures of merit to evaluate the quality of reconstructed objects. Figure 8.16 compares the energy resolution and scale between the DNNv2 and the Moustache superclusters (the comparison with the DNNv1 is reported in Appendix A). While the energy scale remains stable across the entire energy range and calorimeter regions, the energy resolution for the DNNv2 model either shows an improvement or remains at the same level of Moustache. The features in different kinematic domains are detailed below.

High-density region, $E_{\mathrm{T,seed}} \lesssim 60$ GeV This region exhibits the largest enhancement in energy resolution, characterised by an average reduction of approximately 15%. In cases where the transverse energy of the seed particle falls within the range of 10 to 25 GeV, this improvement can be even more pronounced, reaching up to 20%. This improvement can be primarily attributed to the heightened PU resilience of both DNN models since, within this kinematic domain, the Moustache superclustering is found to be particularly aggressive, aggregating a sizeable fraction of PU. The reduction in PU contamination can be inferred from the drop of the FPR by 60% for $E_{\mathrm{T,seed}} < 30$ GeV and 30% for $30 < E_{\mathrm{T,seed}} < 60$ GeV, ultimately reaching a value of approximately 10%.

High-density region, $E_{\mathrm{T,seed}} \gtrsim 60$ GeV This region stands out for the most significant increase of signal efficiency, with the TPR more than doubling when forming superclusters using either the DNNv1 or DNNv2 model. Since electron emissions

typically yield low-energy tracksters, this increase in the TPR for high-energy electrons is not manifest as a substantial improvement in energy resolution, as their contribution to the supercluster energy becomes negligible. This noteworthy enhancement in supercluster efficiency stems from the recovery of the circular halo around the seed that is more populated for this topology of events (cfr Sec. 7.3). The presence of the halo is also responsible for the corresponding increase of the FPR from ∼12% to ∼18% when employing the DNNv1 model. The DNN model is well-trained to search for tracksters in the halo but, since some of the tracksters populating this region can be very ill-posed, distinguishing them from PU contributions becomes arduous. The DNNv2 model mitigates this effect by 15%, reducing the PU contamination to ∼15%. This milder PU rejection does not substantially impact the resolution as the energy of the seed is high, and the resolution is found to be compatible with the one obtained with the Moustache algorithm.

Low-density region In contrast to the high-density region, the outcomes of the superclustering in this region do not exhibit a pronounced kinematic dependence. The utilisation of the DNNv1 model leads to a uniformly improved signal efficiency across $E_{\mathrm{T,seed}}$, with even more substantial gains observed when employing the superclustering based on the DNNv2 model. However, within this region of the detector, the DNNv1 exhibits larger FPR values compared to the Moustache algorithm, although consistently remaining below 10%. This effect is mitigated by the DNNv2 model, which reduces PU contamination to slightly lower levels achieved by the Moustache algorithm. By excluding events with $E_{\mathrm{T,seed}} < 20$ GeV, where a 5% improvement in resolution is observed, the supercluster resolution obtained from the DNNv2 model is found to be comparable to that achieved by the Moustache for higher values of the transverse energy of the seed.

The impact of a correct and robust superclustering algorithm extends beyond the energy resolution, as it also affects the lepton isolation. Electron emissions often result in tracksters located in close proximity to the seed. Failure to cluster these contributions or inadvertent collection of energy from other particles may bias the computation of the lepton isolation. This is vital for many analyses, like the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ channel presented in Part II of this thesis. The accuracy is the correct figure of merit to investigate this concern. The accuracy is defined as the fraction of correct predictions:

$$\text{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \tag{8.18}$$

On average, the accuracy for the Moustache is approximately 80%, while for the DNNv1 and DNNv2, it raises to 91% and 93%, respectively.

Continuing along the same logical thread regarding what is included in superclusters and what is not, the investigation can be broadened by relaxing the gen-matching requirement. Section 8.4.1 set a 60% gen-score threshold for a trackster to qualify as an electron emission. Although this threshold was optimised to maximise the energy resolution, it is, to some extent, an arbitrary choice. It is interesting to study the gen-score of the tracksters composing the superclusters formed by the different algorithms as well as the ones of the rejected tracksters. The distributions are shown in Fig. 8.17. The results further underscore the superior PU resilience of the DNNsuperclustering, as it incorporates fewer low-gen-score tracksters and a higher proportion of tracksters with elevated gen-scores. While there is a *middle-earth* represented by tracksters with a score that is neither too high nor too low on which one could argue about the usefulness of their inclusion in the supercluster, surely everyone will agree in aiming to discard trackster with gen-score = 0. Such tracksters are purely spurious PU-tracksters, and the actual contribution to suppress

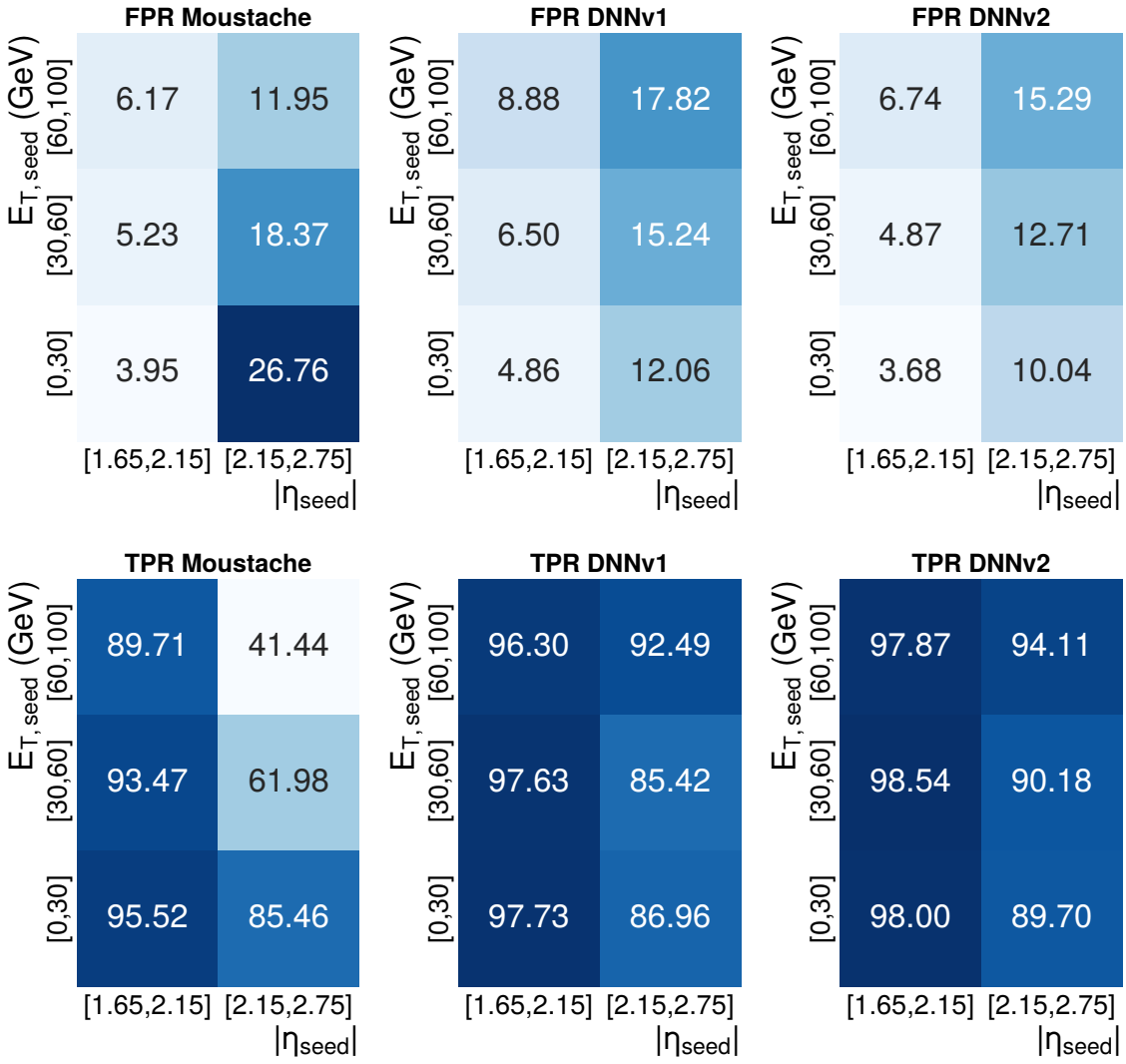

Figure 8.15: False Positive Rate (top) and True Positive Rate (bottom) values for Moustache (left), DNNv1 (centre), and DNNv2 (right), in bins of transverse energy ($E_{\mathrm{T,seed}}$) and pseudorapidity (η_{seed}) of the seed. The True Positive Rate (TPR) indicates the signal efficiency, while the False Positive Rate (FPR) denotes the PU contamination. The two pseudorapidity bins encloses the low-density ($1.65 < |\eta| < 2.15$) and high-density region ($2.15 < |\eta| < 2.75$) of the HGCAL. The False Positive Rate estimates the PU contamination of the resulting superclusters. The True Positive Rate denotes the supercluster efficiency.

in the superclustering. The fraction of this contribution in the Moustache superclusters is around 21%, while for DNNv1 and DNNv2, it reduces down to 6% and 5%, respectively.

A comparison between the superclustering carried out with the Moustache and the DNNv2, on a specific event, can be seen in Fig. 8.18. Additional information to evaluate the performance of the DNN-based superclustering can be found in Appendix A.

Summary

Electron superclustering represents a pivotal step in achieving accurate electron reconstruction, and this chapter presented the first assessment of this procedure in the HGCAL. The chapter examined the characteristics of the existing CMS ECAL Moustache algorithm

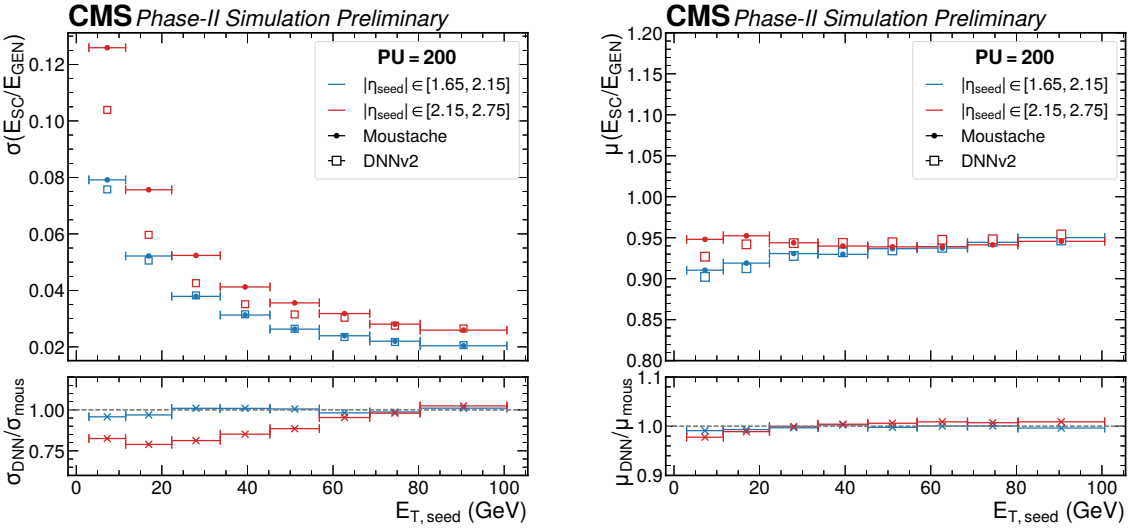

Figure 8.16: Resolution (left) and scale (right) of the supercluster energy (E_{SC}) divided by the energy of the generated electron (E_{GEN}), as a function of the transverse energy of the seed ($E_{\mathrm{T,seed}}$). Events are divided in two bins of pseudorapidity of the seed (η_{seed}) corresponding to the high-density and low-density region of the HGCAL. No energy corrections are applied. The resolution and scale are computed by fitting the $E_{\mathrm{SC}}/E_{\mathrm{GEN}}$ with a Cruijff function. The lower panel shows the ratio of the resolution and scale of the DNNv2 and the Moustache algorithm.

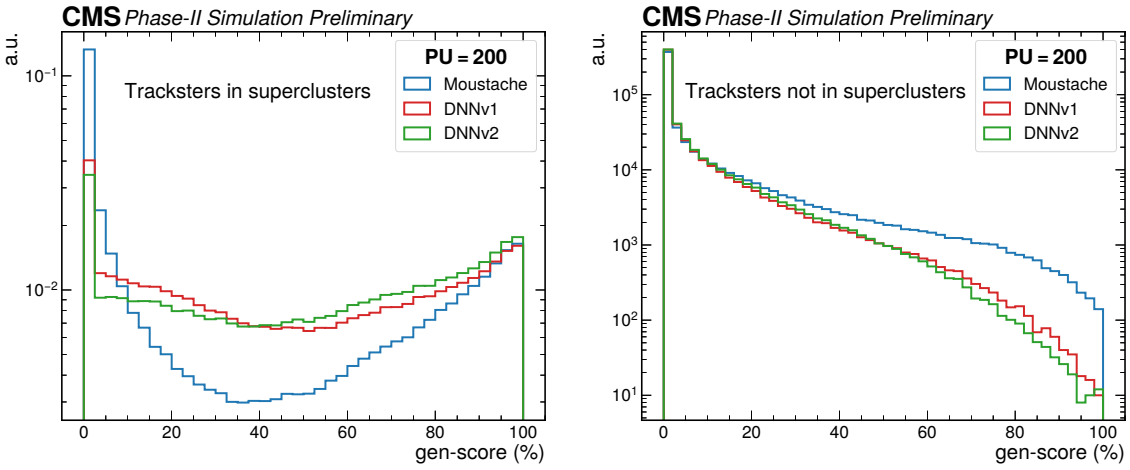

Figure 8.17: Distribution of the gen-score for tracksters forming superclusters (left) and tracksters rejected by the superclustering algorithms (right). The threshold for the association of a trackster to an energy deposit from the generated electron is set to 60%.

when adapted for use in the HGCAL and introduces a promising DNN-based alternative. Notably, it was found that the utilisation of DNNs significantly increases the signal efficiency of the superclustering algorithm, while the implementation of variables based on the possibility to compute the shower axis in the HGCAL reduces the PU contamination. These improvements have a direct impact on the energy resolution, which is particularly enhanced in the high-density region of the calorimeter and for low-energy seeds. Unlike the geometrical Moustache approach, the DNN algorithms have the capability to capture nuances in the electron emission pattern that hard geometrical cuts miss.

One of the features of the HGCAL reconstruction, as it is now, is the formation of satellite tracksters around a reconstructed shower due to energy that is not clustered

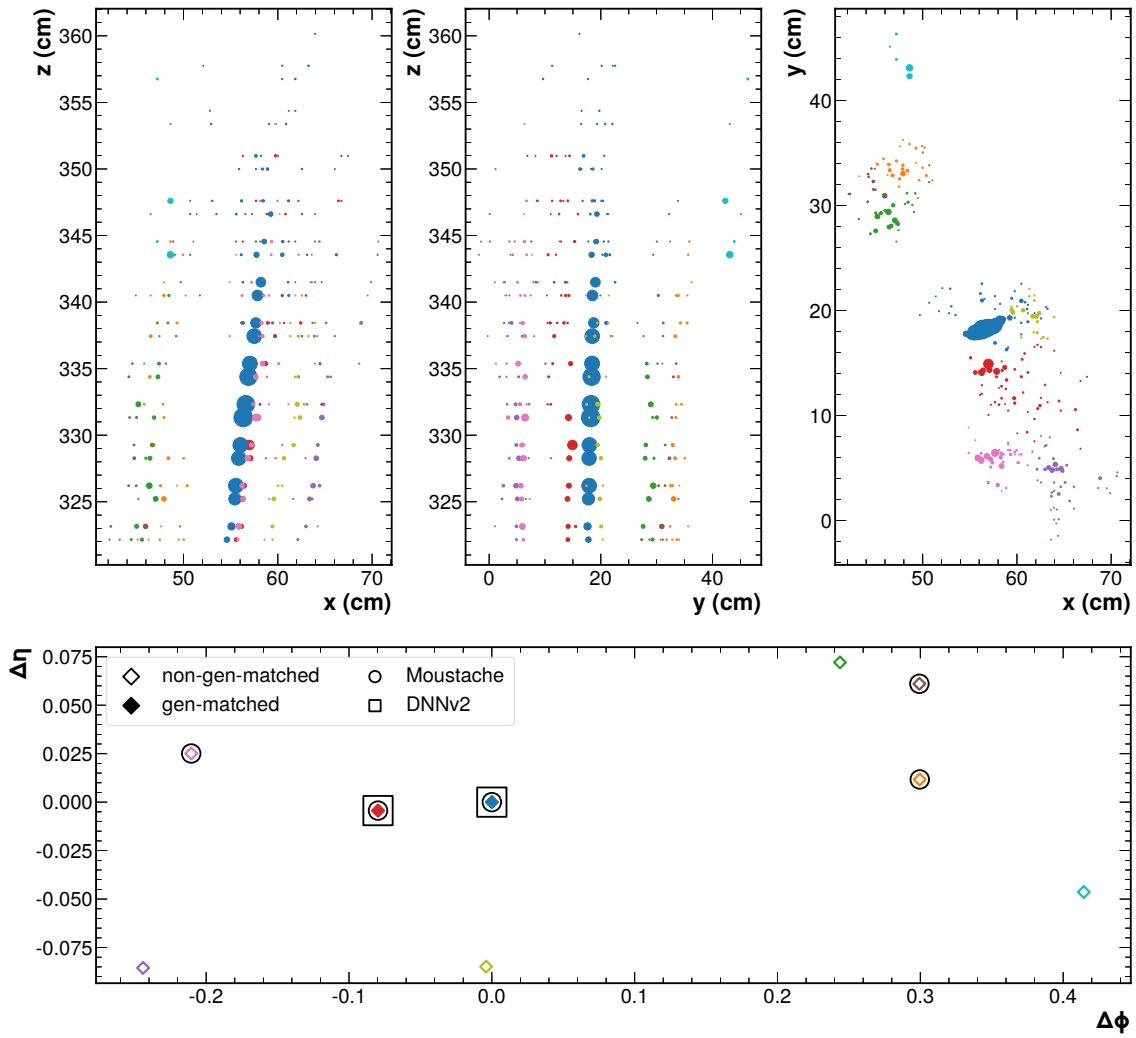

Figure 8.18: (Top) Event display in the HGCAL of an electron. All tracksters passing the skimming procedure presented in Sec. 8.4.2 and inside the region of interest defined in Sec. 8.4.3 are displayed. Each circle represents a LC and its radius is proportional to its energy. The LCs forming a trackster are identified with the same colour. (Bottom) Representation of the same event in the $\Delta\eta = |\eta| - |\eta_{\text{seed}}|$ versus $\Delta\phi = \phi - \phi_{\text{seed}}$ plane. The gen-matched tracksters are identified with a filled diamond, while non-gen-matched tracksters are empty. If the trackster is part of the supercluster formed by Moustache, the corresponding diamond is surrounded by a black circle. If the trackster is part of the supercluster formed by the DNNv2, the corresponding diamond is surrounded by a black square. In this specific case, the gen-matched tracksters are captured by both the Moustache and the DNNv2 algorithms. Three not gen-matched tracksters are clustered by Moustache but effectively rejected by the DNNv2. Kinematic information: $E_{\text{GEN}} = 119$ GeV, $E^{\text{Moustache}}_{\text{supercluster}} = 142$ GeV, $E^{\text{DNNv2}}_{\text{supercluster}} = 122$ GeV.

within the same object. This phenomenon leads to the formation of a circular halo in the $\Delta\eta - \Delta\phi$ distributions. While the primary objective of superclustering is not to recover these contributions, the proposed DNN algorithm is effective in recovering these tracksters as long as they are reasonably well-formed. In cases where these peculiar contributions are poorly formed, typically suppressed by the skimming procedure, their recovery should be addressed earlier, at the patter recognition level.

The path designated in this chapter is the right one to go for the superclustering for the phase-2 of the CMS experiment: abandon geometrical greedy algorithms in favour of ML models able to leverage the great amount of information provided by the new HGCAL.

There is one aspect of superclustering that has been omitted in this chapter: the photon superclustering. Similarly to electrons, also photons have a significant probability to shower in the path towards the calorimeter. As the features of the photon emission are akin to those of electron emission, their superclustering is currently based on the same identical Moustache algorithm employed with electrons. Although the DNNsuperclustering has not been yet tested with converted photons, its extension does not cause particular concern. In the case of poor performance, the model could be either trained with a mixture of electrons and photons or a dedicated model only for photons could be trained. The underlying physics and philosophy will not change when dealing with photons.

Conclusions

The discovery of the Higgs boson in 2012 marked a watershed moment for particle physics. It culminated a decades-long search and put the last elusive particle in the Standard Model (SM) frame. However, is this particle the Higgs boson as prescribed by the current SM? This is a question we will never be able to answer. Instead, we can rather prove that this particle is not the SM Higgs boson with precision measurements and, from the theory side, accurate theoretical predictions. The 138 fb^{-1} of data collected by the Large Hadron Collider (LHC) during its Run 2 opened the possibility for an extensive, granular characterisation of the Higgs boson profile. The fiducial cross section measurements in the $\text{H} \to \text{ZZ} \to 4\ell$ channel presented in this thesis are part of this extensive effort from the CMS collaboration. The inclusive cross section is found to be $\sigma_{\text{fid}} = 2.73 \pm 0.26$ fb, in agreement with the SM expectation of 2.86 ± 0.15 fb. The fiducial cross sections are also measured in bins of 32 single- and double-differential observables to probe the production and decay of the Higgs boson, as well as the associated jet activity. The measurement of the fiducial cross section in differential bins of transverse momentum is used to set constraints on the trilinear self-coupling of the Higgs boson, with an observed (expected) limit of $-5.4\,(-7.6) < \kappa_\lambda < 14.9\,(17.7)$ at the 95% confidence level. Finally, constraints on the modifiers of Higgs boson couplings to c and b quarks are also determined with an observed (expected) limit of $-1.1\,(-1.3) < \kappa_b < 1.1\,(1.2)$ and $-5.3\,(-5.7) < \kappa_c < 5.2\,(5.7)$ at the 95% confidence level, obtained assuming a dependence of the branching ratio on κ_b and κ_c. All results are consistent with the theoretical predictions of the SM.

The CMS physics programme is continuing with the Run 3 of the LHC. Alongside continuous detector upgrades and the collection of more number of events, the analysis strategies must evolve to enhance the precision of final results. The most challenging background source in the $\text{H} \to \text{ZZ} \to 4\ell$ channel is the so-called reducible background, consisting of non-prompt and misidentified leptons. During the Run 2, this contribution was estimated though a data-driven method, yielding a final uncertainty of 40%, largely dominated by the systematic component. This thesis presented a novel method that will be employed in the forthcoming analyses. This approach provides an alternative means of estimating the reducible background, offering the potential to substantially reduce the associated systematic uncertainty in future measurements, which will be performed with ever increasing datasets.

The next era in high-energy physics will be represented by the High-Luminosity LHC (HL-LHC), where the potential for discoveries and precision physics will be boosted due to an increase in the instantaneous luminosity. A brand-new High-Granularity CALorimeter (HGCAL) will replace the current electromagnetic and hadronic endcap calorimeters. The HGCAL will be the first large-scale silicon-based sampling calorimeter employed in high-energy physics, opening a new era in calorimetry. Such a revolution on the hardware side must accompany a paradigm shift on the software side as well. The HGCAL offline reconstruction is currently being developed to fully exploit the novel features of the detector and address the challenges posed by the dense environment resulting from the large Pile Up (PU). It has been assessed that the implementation of energy-based pattern recognition algorithms is able to provide a robust reconstruction of electromagnetic show-

ers even in such a harsh environment. However, the reconstruction process may pick up spurious PU-induced energy contributions, compromising the quality of the reconstructed shower. This thesis proposed a cleaning algorithm that effectively removes PU contributions mistakenly associated with the electromagnetic shower, providing better-quality objects. Particularly, at low energy, where the impact of PU is more pronounced, the energy resolution improves by an average of 15%. Additionally, this thesis assessed for the first time the electron superclustering in the HGCAL, one of the main challenges in electron reconstruction. The strong solenoid magnetic field and the possibility of an electron initiating showering before reaching the surface of the HGCAL can give rise to multiple energy clusters spread throughout the calorimeter that should be combined to reconstruct the original electron. Historically, the CMS collaboration has tackled this problem in the ECAL using purely geometrical algorithms. The Moustache algorithm represents the last evolution of such techniques, which has been used throughout the Run 2 and the underway Run 3. Looking ahead to phase-2 of the CMS detector and the installation of the new HGCAL, this thesis proposed a novel way to approach the problem. Geometrical algorithms are intrinsically prone to collect PU contributions within their acceptance. In addition, compared to the current ECAL, the HGCAL will give access to more information that can be wisely leveraged. The proposed superclustering algorithm is based on Deep Neural Networks, incorporating new variables defined using the possibility of properly estimating the shower axis though Principal Component Analysis. The interplay between machine learning techniques and features of the HGCAL is able to both increase the signal efficiency by 30% and reduce the PU contamination by 20%, leading to significant improvements in the energy resolution.

The exploration of the TeV scale is just at the beginning. At the end of the 2023 data-taking period, the LHC delivered a total of 266.42 fb^{-1}, of which 245.54 fb^{-1} was recorded by the CMS detector. This represents less than 10% of the luminosity expected to be collected by the end of the HL-LHC, and the last data of the Run 3 are still to be analysed. All results thus far do not show significant discrepancies from the SM, but we have only started to scratch the surface. Quoting the suggestion of Nima Arkhani-Hamed [212]: *"While we continue to scratch our heads as theorists, the most important path forward for experimentalists is clear: measure the hell out of these crazy phenomena! [...] when theorists are more confused, it's the time for more, not less experiments."* In this huge effort, the $\text{H} \to \text{ZZ} \to 4\ell$ channel will continue to play a leading role and be at the forefront of precision Higgs physics due to its many virtues. Furthermore, the HGCAL, with a solid offline (and online) reconstruction, will be a precious tool to disclose the mysteries ahead of us.

"Now, this is not the end.
It is is not even the beginning of the end.
But it is, perhaps, the end of the beginning."

Winston Churchill

Appendix A

DNNsuperclustering: additional information

This appendix shows additional information to evaluate the performance of the DNN-based superclustering algorithms compared to the Moustache. From the confusion matrix, many figure of merits can be extracted. Figures A.1 and A.2 shows the False Discovery Rate (FDR) and the Positive Predictive Value (PPV), respectively. Their definition is the following:

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}} = 1 - \mathrm{PPV} \tag{A.1}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{TP}} = 1 - \mathrm{FDR} \tag{A.2}$$

The sum of these two quantities is always equal to one. The FDR expresses the fraction of not gen-matched tracksters forming the superclusters. On the other hand, the PPV expresses the fraction of gen-matched tracksters forming the superclsuters, that is the purity (or precision) of the reconstructed electron.

Figures A.3 and A.4 shows the $\Delta\eta - \Delta\phi$ distributions to have a visual representation of what is clustered by each superclustering algorithm.

Figure A.5 shows the comparison of the energy scale and resolution between the DNNv1 model and the Moustache algorithm. The corresponding plot for the DNNv2 is reported in Fig. 8.16.

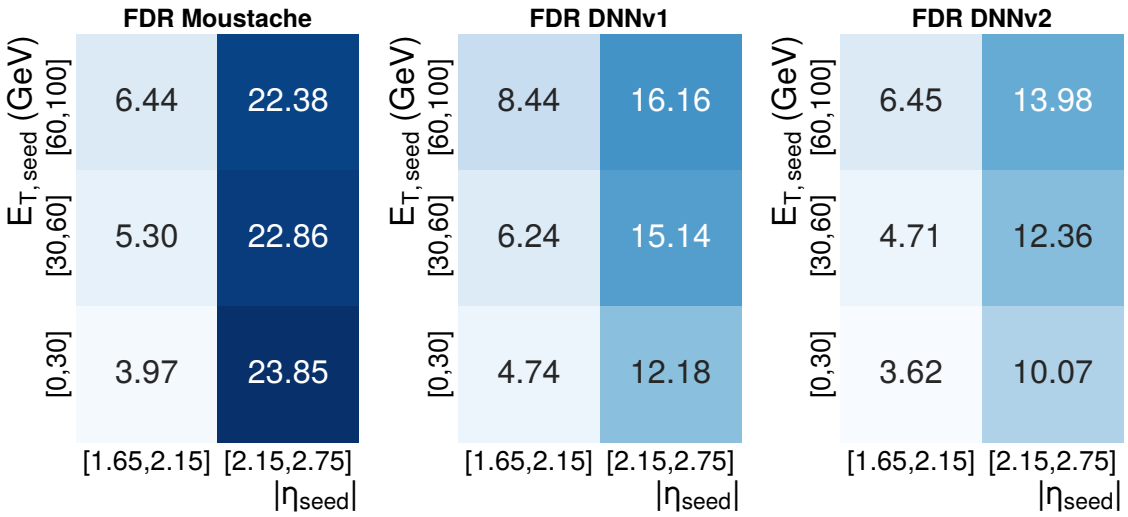

Figure A.1: False Discovery Rate (FDR) values for Moustache (left), DNNv1 (centre), and DNNv2 (right), in bins of transverse energy ($E_{\mathrm{T,seed}}$) and pseudorapidity (η_{seed}) of the seed. The two pseudorapidity bins encloses the low-density and high-density region of the HGCAL. The FDR is defined as FDR = FP/(FP+TP).

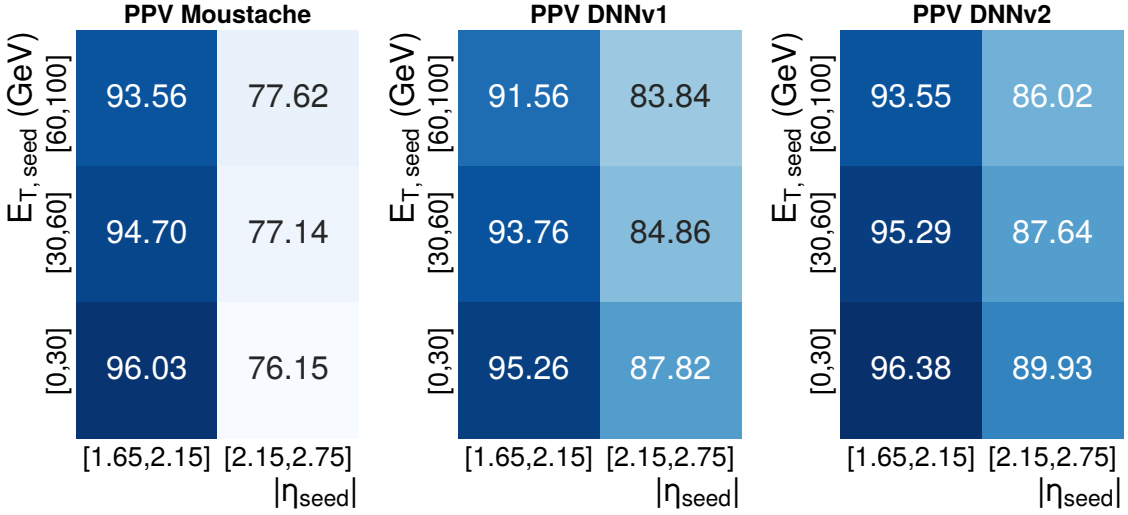

Figure A.2: Positive Predictive Value (PDV), also known as precision, values for Moustache (left), DNNv1 (centre), and DNNv2 (right), in bins of transverse energy ($E_{\mathrm{T,seed}}$) and pseudorapidity (η_{seed}) of the seed. The two pseudorapidity bins encloses the low-density and high-density region of the HGCAL. The PPV is defined as PPV = TP/(FP+TP).

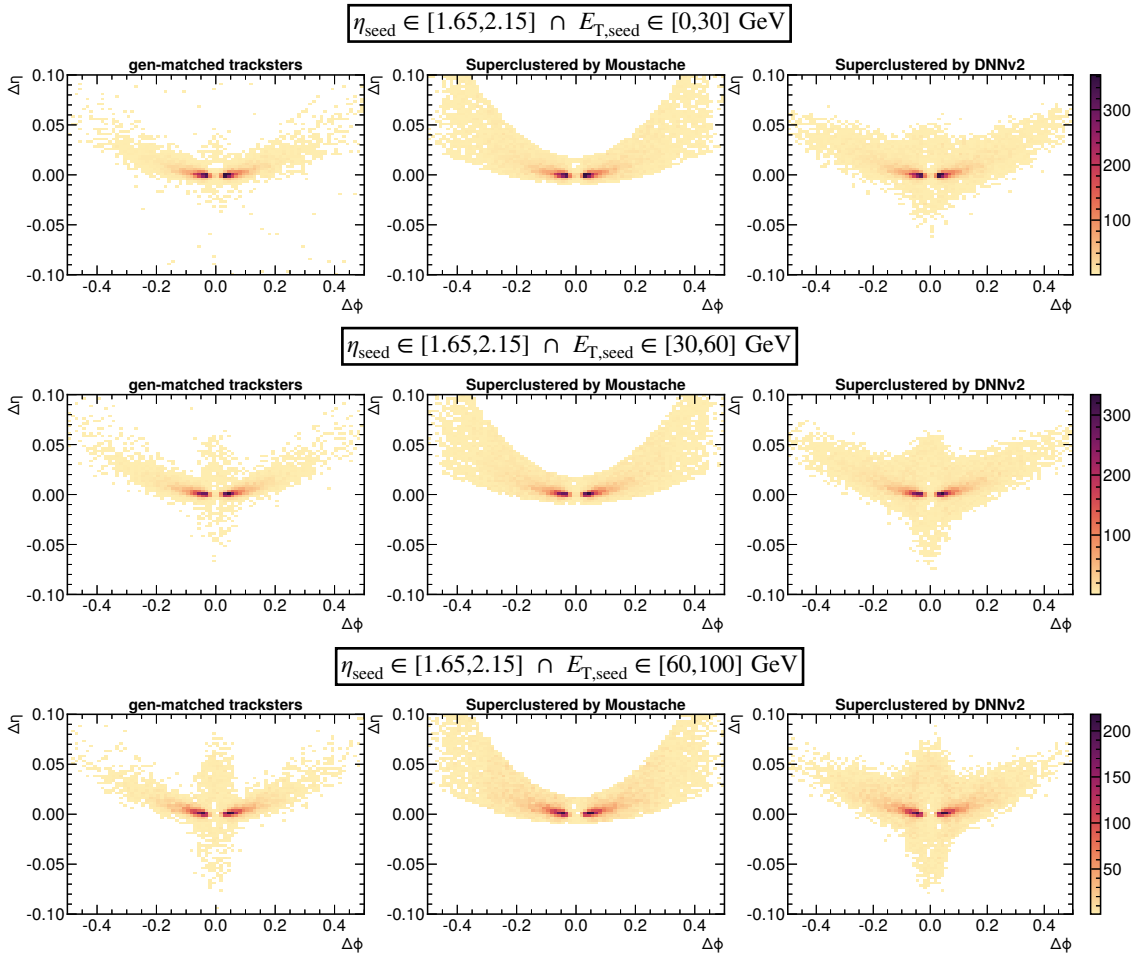

Figure A.3: Distributions of the $\Delta\eta = |\eta| - |\eta_{\text{seed}}|$ versus $\Delta\phi = \phi - \phi_{\text{seed}}$ variables for gen-matched trackster (left column), tracksters forming superclusters using the Moustache algorithm (central column), and tracksters forming superclusters using the DNNv2 algorithm (right column). The seed is defined as the trackster with the largest transverse energy. The distribution are plotted for events with the seed in the low-density region of the HGCAL ($|\eta| \in [1.60, 2.15]$), and each row represents a different bin of transverse energy of the seed.

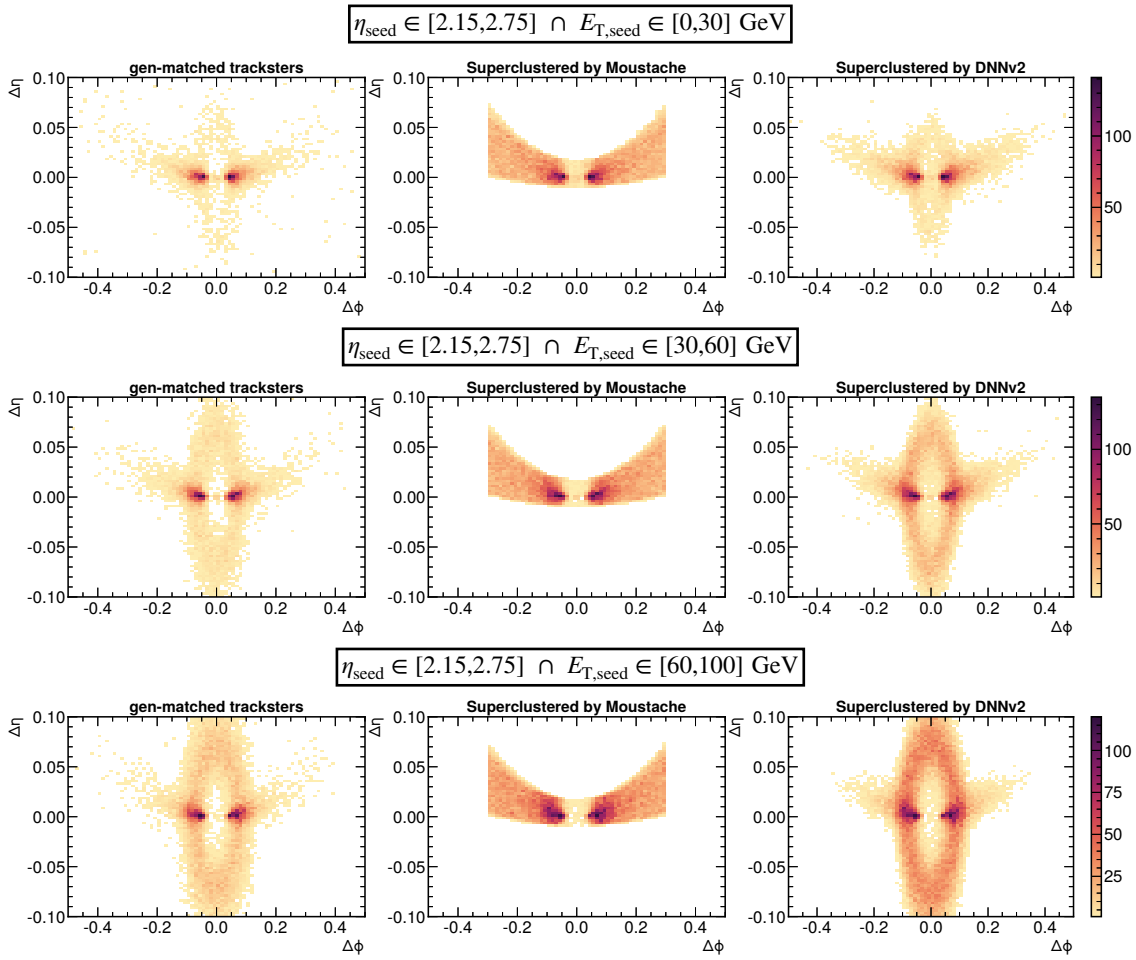

Figure A.4: Distributions of the $\Delta\eta = |\eta| - |\eta_{\text{seed}}|$ versus $\Delta\phi = \phi - \phi_{\text{seed}}$ variables for gen-matched trackster (left), tracksters forming superclusters using the Moustache algorithm (centre), and tracksters forming superclusters using the DNNv2 algorithm (right). The seed is defined as the trackster with the largest transverse energy. The distribution are plotted for events with the seed in the high-density region of the HGCAL ($|\eta| \in [2.15, 3.10]$), and each row represents a different bin of transverse energy of the seed.

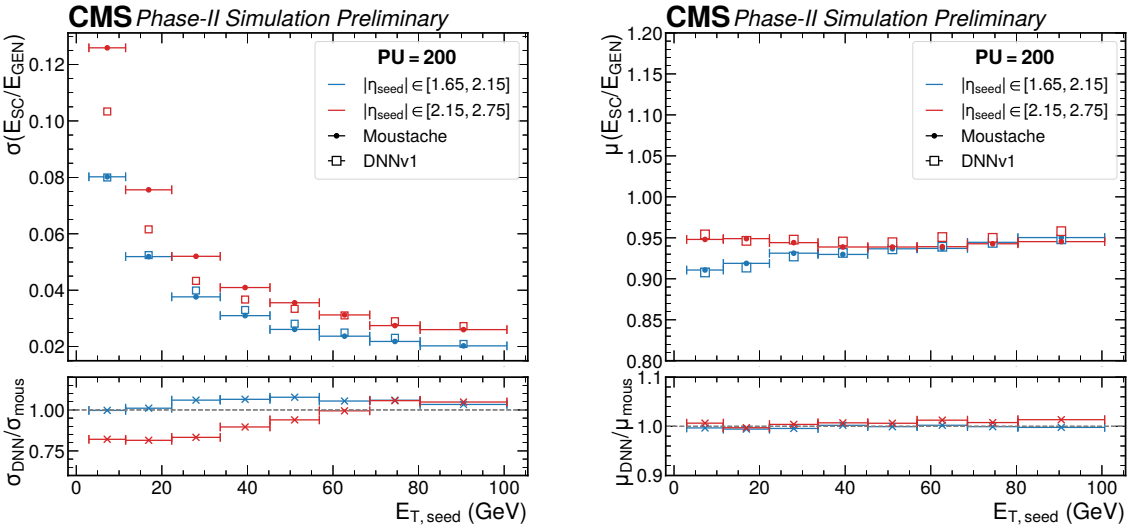

Figure A.5: Resolution (left) and scale (right) of the supercluster energy (E_{SC}) divided by the energy of the generated electron (E_{GEN}), as a function of the transverse energy of the seed ($E_{\mathrm{T,seed}}$). Events are divided in two bins of pseudorapidity of the seed (η_{seed}) corresponding to the high-density and low-density region of the HGCAL. No energy corrections are applied. The resolution and scale are computed by fitting the $E_{\mathrm{SC}}/E_{\mathrm{GEN}}$ with a Cruijff function. The lower panel shows the ratio of the resolution and scale of the DNNv1 and the Moustache algorithm.

Appendix B

Impacts and pulls plots

This appendix shows the impact and pull plots for the inclusive cross sections when the irreducible background normalisation is either constrained to the SM expectation (Fig. B.1) and left as an unconstrained parameter in the fit procedure (Fig. B.2). The pull is reported as the difference between the pre-fit and post-fit values divided by the post-fit uncertainty of the nuisance parameter: $(\hat{\theta} - \theta_0)/\Delta\theta$. The impact of a nuisance parameter θ on the parameter of interest μ is defined as the shift $\Delta\mu$ that is induced as θ is fixed and brought to its $+1\sigma$ or -1σ post-fit values, with all other parameters profiled as normal.

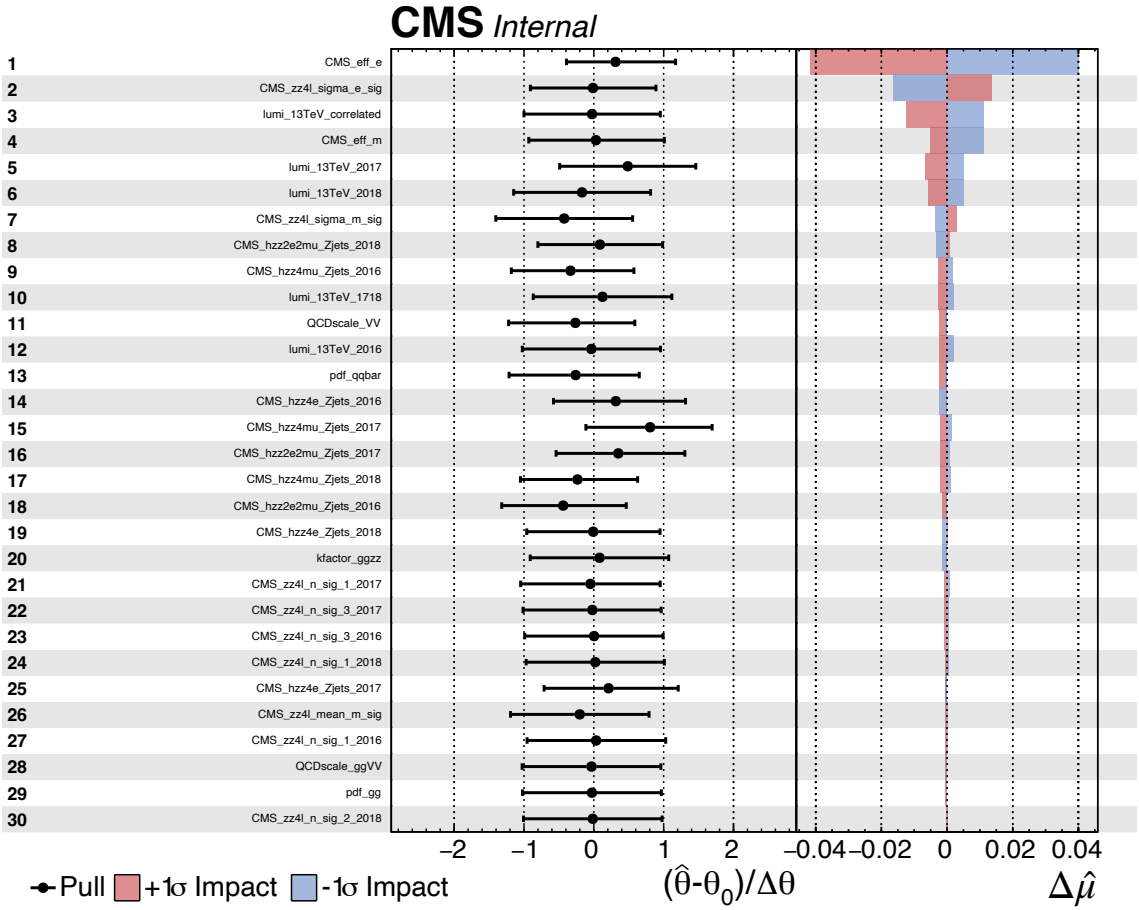

Figure B.1: Difference between the pre-fit and post-fit values of the 30 leading nuisance parameters $(\hat{\theta}-\theta_0)/\Delta\theta$ considered in the measurement of the inclusive fiducial cross section when constraining the irreducible background normalisation to the SM expectations, and their impact on the corresponding parameter μ.

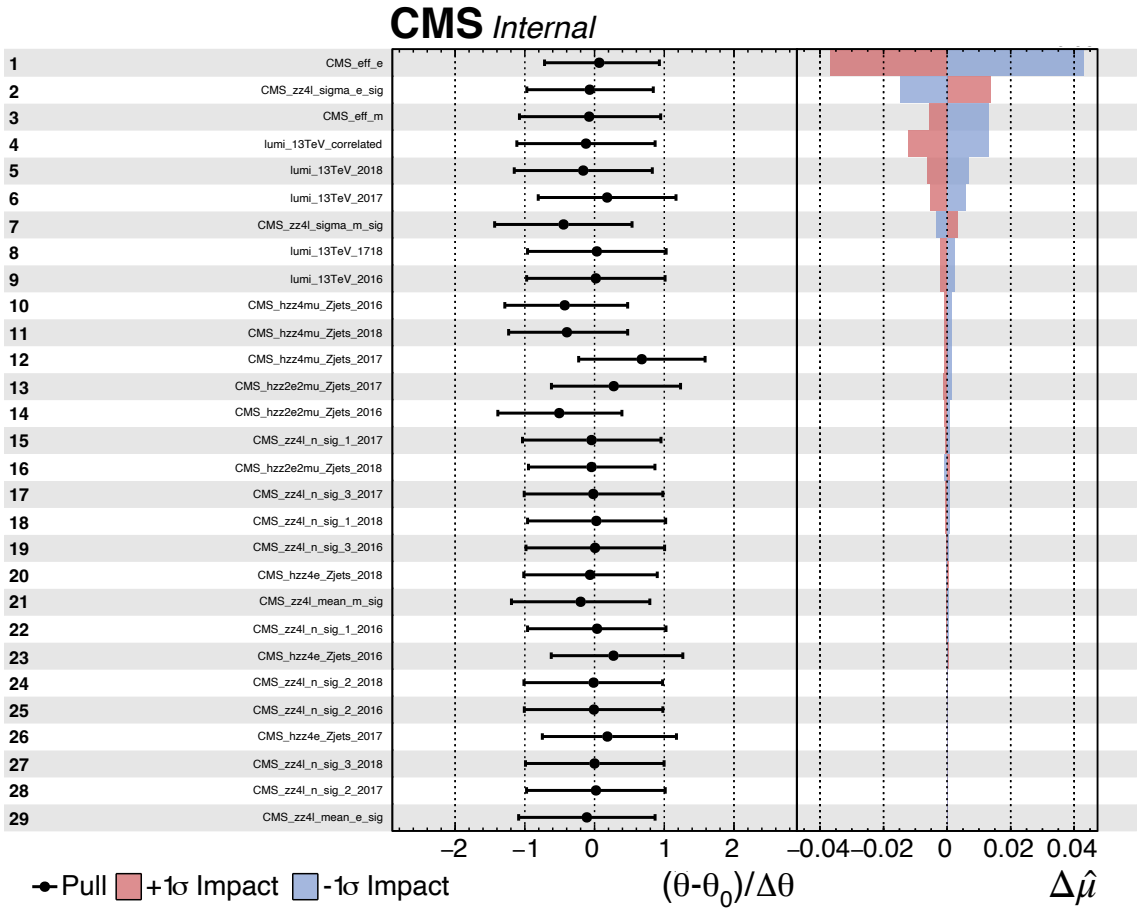

Figure B.2: Difference between the pre-fit and post-fit values of the 30 leading nuisance parameters $(\hat{\theta}-\theta_0)/\Delta\theta$ considered in the measurement of the inclusive fiducial cross section when leaving the irreducible background normalisation unconstrained in the fit procedure, and their impact on the corresponding parameter μ.

Appendix C

Detector-level distributions and correlation matrices

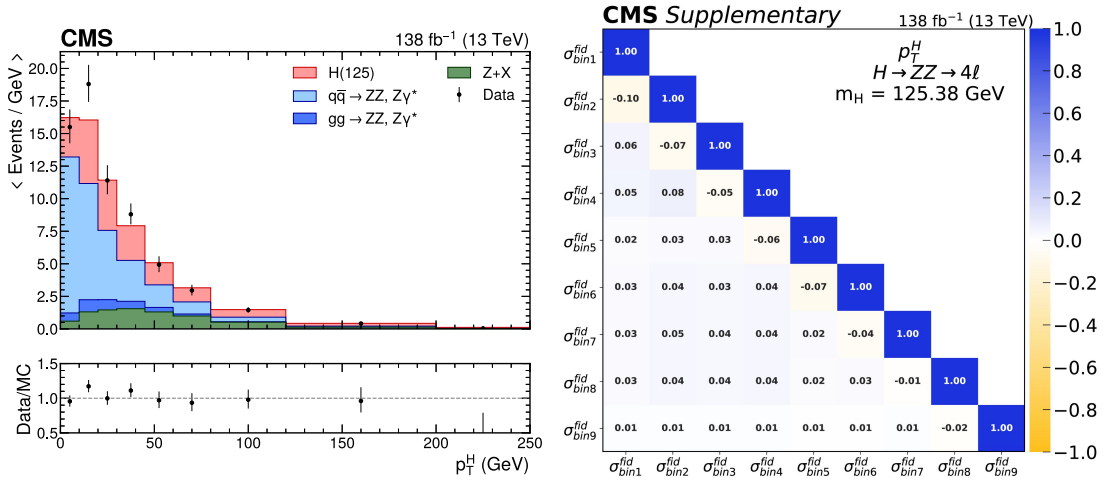

Figure C.1: Distribution of the transverse momentum of the Higgs boson (left) and the correlation matrix of the corresponding measured fiducial cross sections (right). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The bottom panel shows the ratio of the measured values to the expectations from simulation.

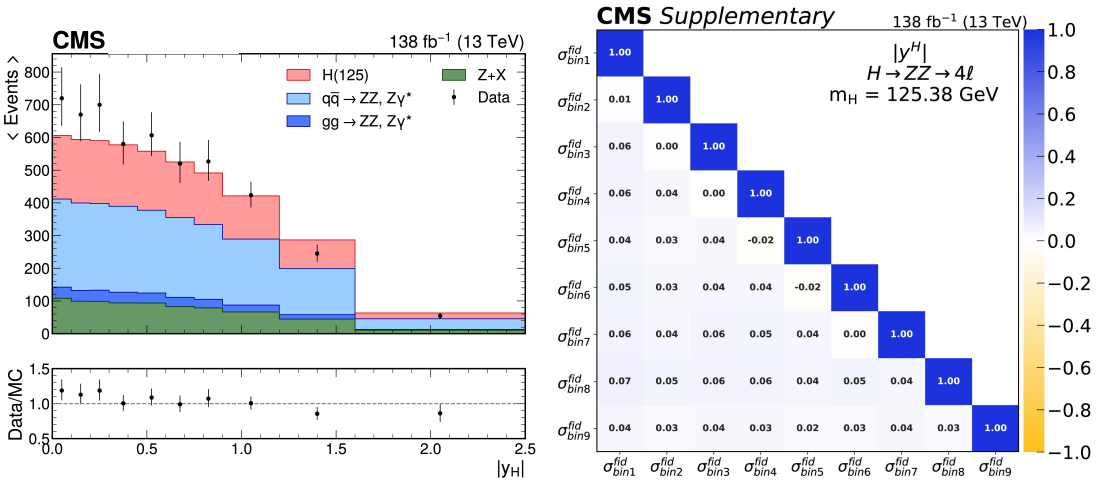

Figure C.2: Distribution of the rapidity of the Higgs boson (left) and the correlation matrix of the corresponding measured fiducial cross sections (right). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The bottom panel shows the ratio of the measured values to the expectations from simulation.

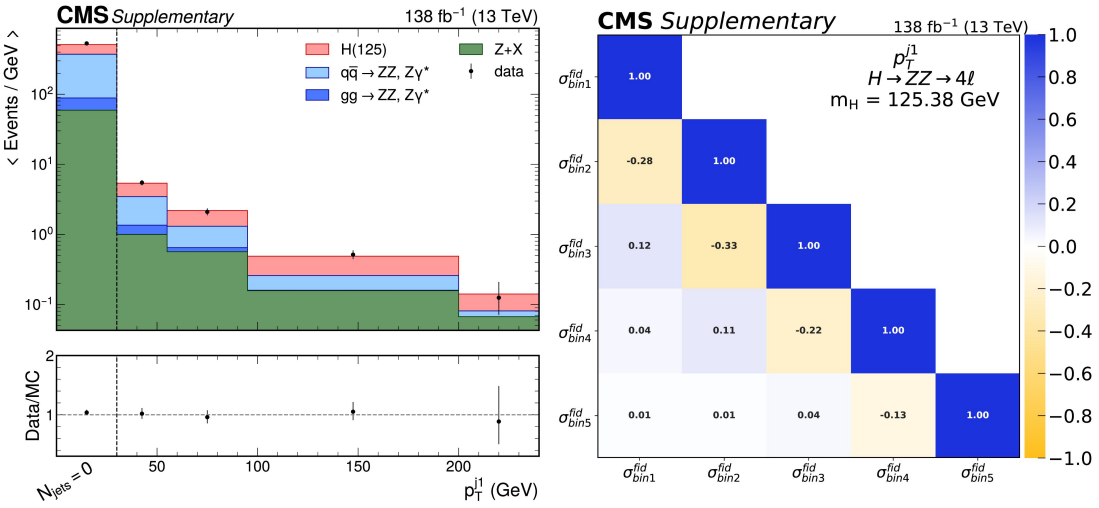

Figure C.3: Distribution of the transverse momentum of the leading jet (left) and the correlation matrix for the corresponding measured fiducial cross sections (right). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The first bin shows the number of events for which the variable cannot be defined, i.e., events with 0 jets. The bottom panel shows the ratio of the measured values to the expectations from simulation.

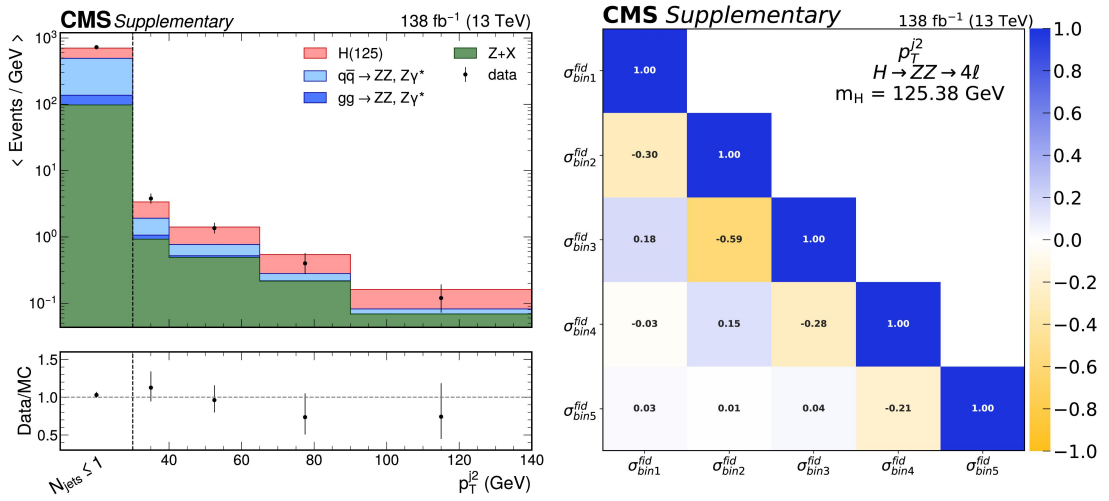

Figure C.4: Distribution of the transverse momentum of the subleading jet (left) and the correlation matrix of the corresponding measured fiducial cross sections (right). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The first bin shows the number of events for which the variable cannot be defined, i.e., events with less than 2 jets. The bottom panel shows the ratio of the measured values to the expectations from simulation.

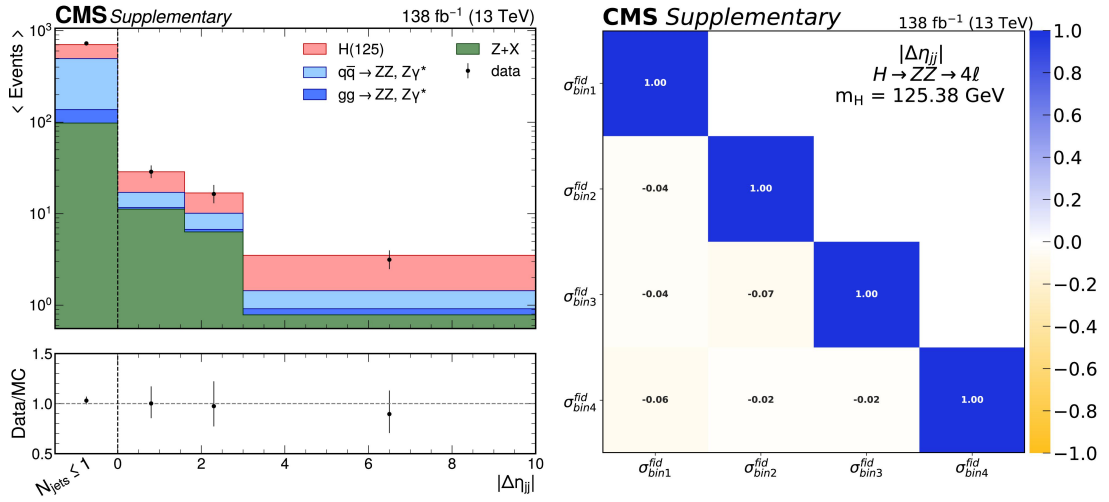

Figure C.5: Distribution of the difference in pseudorapidity between the leading and subleading jet (left) and the correlation matrix of the corresponding measured fiducial cross sections (right). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The first bin shows the number of events for which the variable cannot be defined, i.e., events with less than 2 jets. The bottom panel shows the ratio of the measured values to the expectations from simulation.

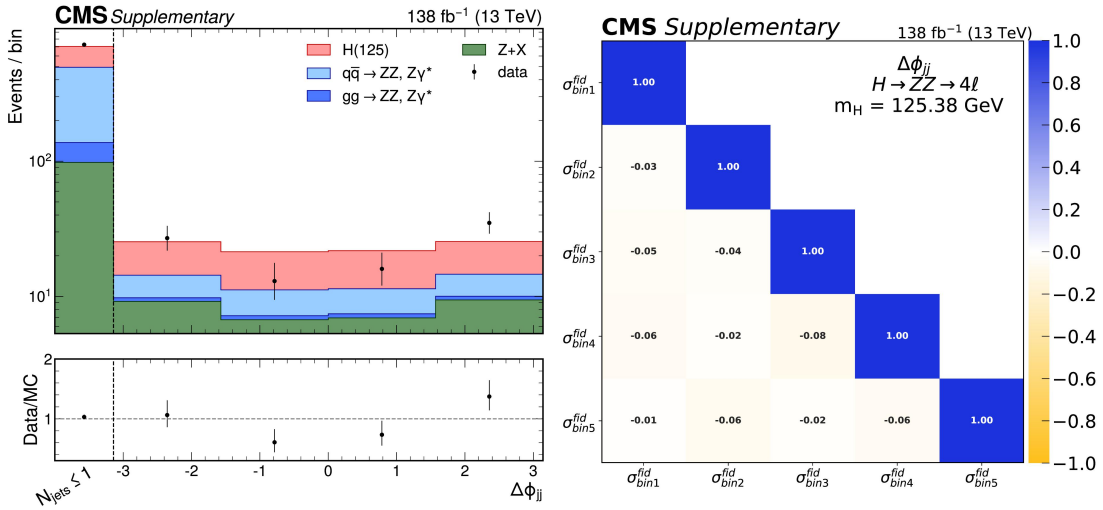

Figure C.6: Distribution of the difference in azimuthal angle between the leading and subleading jet (left) and the correlation matrix of the corresponding measured fiducial cross sections (right). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The first bin shows the number of events for which the variable cannot be defined, i.e., events with less than 2 jets. The bottom panel shows the ratio of the measured values to the expectations from simulation.

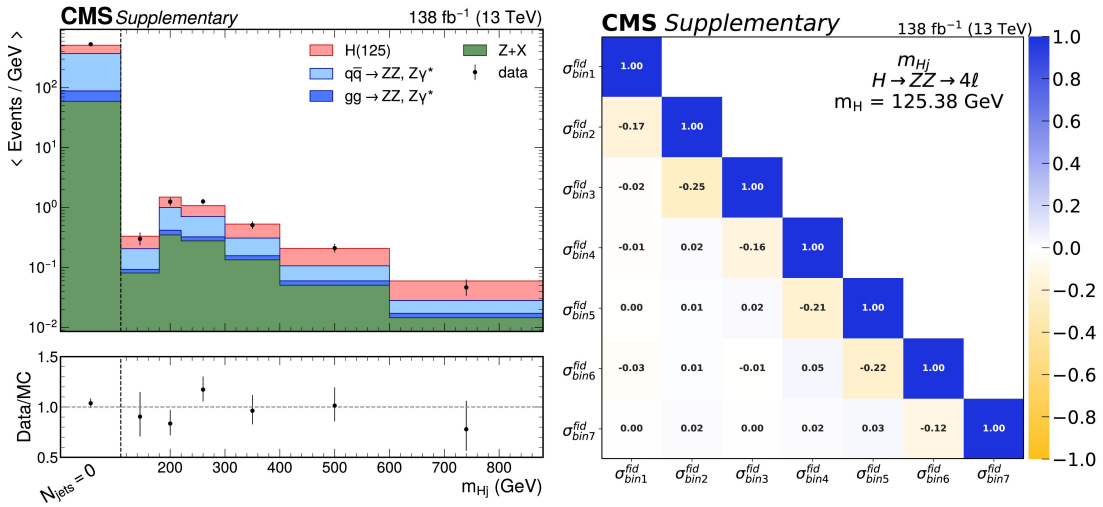

Figure C.7: Distribution of the invariant mass of the H + j system (left), where j is the leading jet in the event, and the correlation matrix of the corresponding measured fiducial cross sections (right). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The first bin shows the number of events for which the variable cannot be defined, i.e., events with 0 jets. The bottom panel shows the ratio of the measured values to the expectations from simulation.

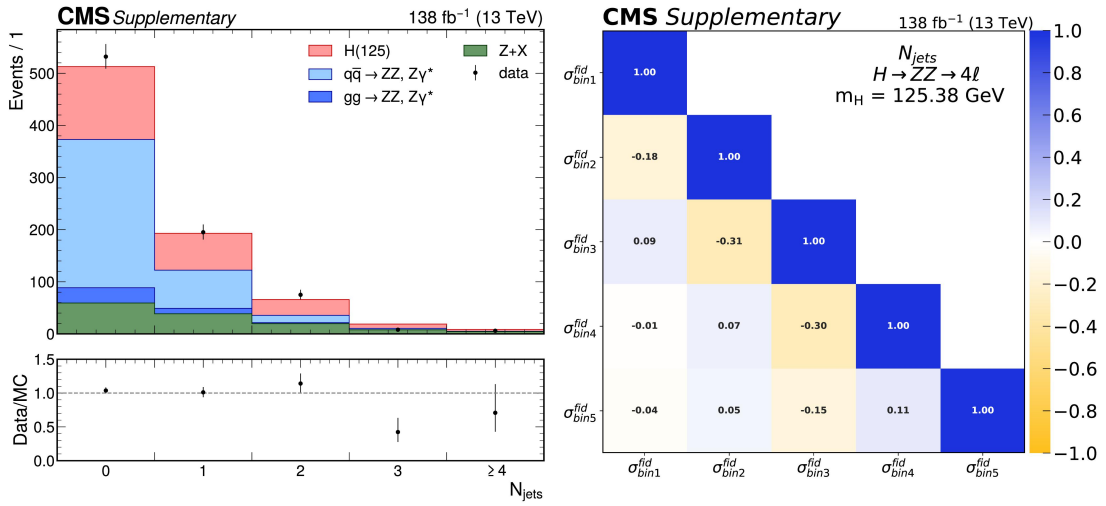

Figure C.8: Distribution of the number of associated jets (left) and the correlation matrix of the corresponding measured fiducial cross sections (right). Points with error bars represent the data, solid histograms the predictions from simulation. The bottom panel shows the ratio of the measured values to the expectations from simulation.

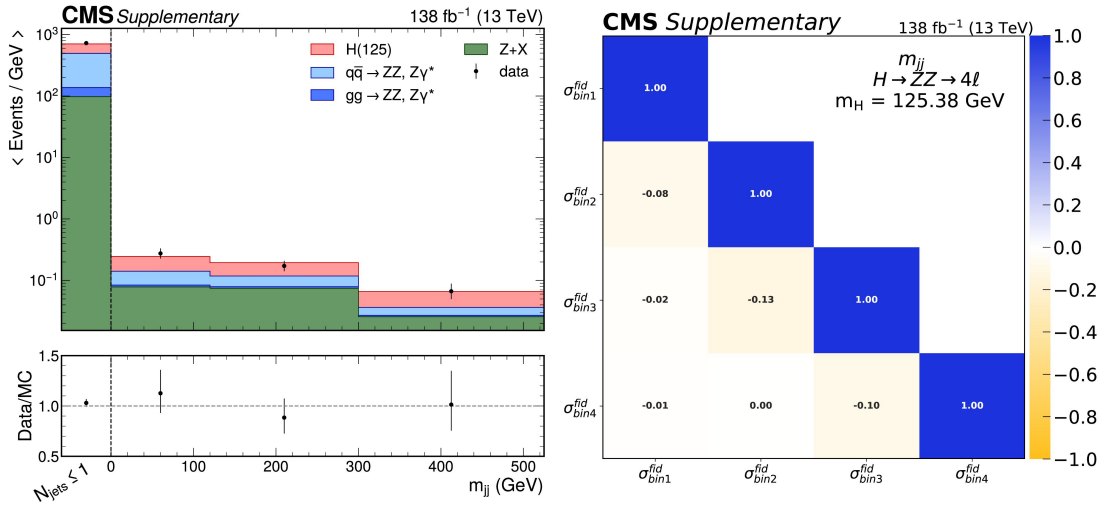

Figure C.9: Distribution of the invariant mass of the leading jet + subleading jet system (left) and the correlation matrix of the corresponding measured fiducial cross sections (right). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The first bin shows the number of events for which the variable cannot be defined, i.e., events with less than 2 jets. The bottom panel shows the ratio of the measured values to the expectations from simulation.

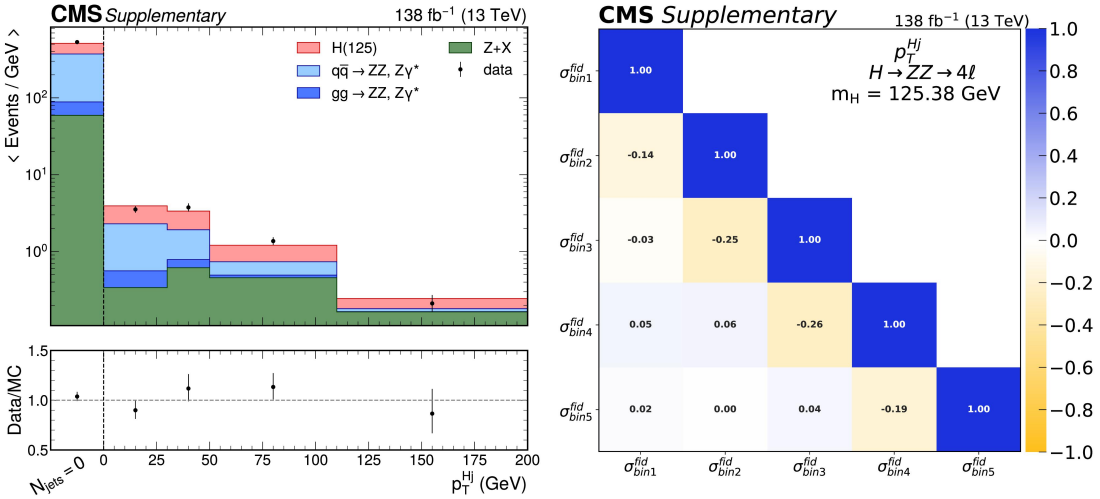

Figure C.10: Distribution of the transverse momentum of the H+j system (left), where j is the leading jet in the event, and the correlation matrix of the corresponding measured fiducial cross sections (right). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The first bin shows the number of events for which the variable cannot be defined, i.e., events with 0 jets. The bottom panel shows the ratio of the measured values to the expectations from simulation.

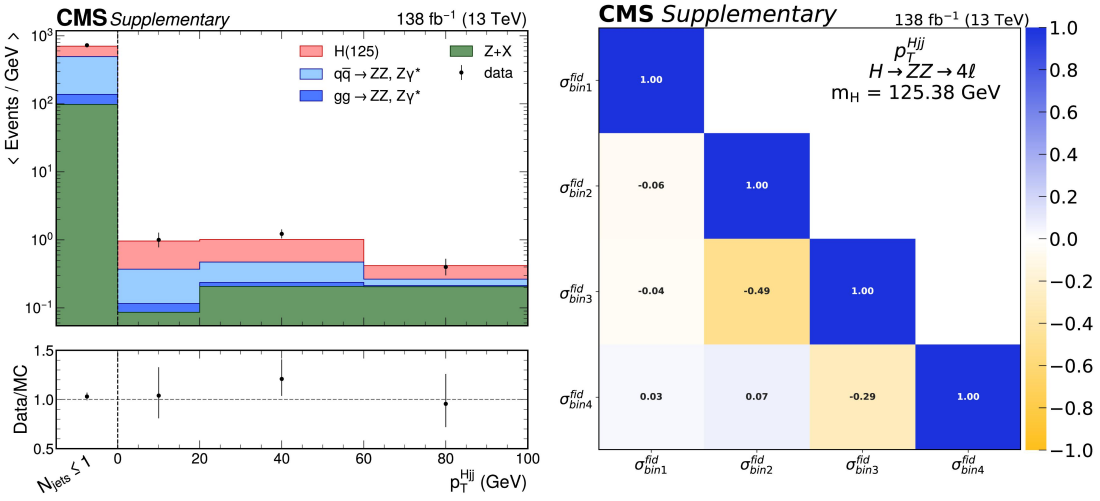

Figure C.11: Distribution of the transverse momentum of the H+jj system (left), where jj represents the leading and the subleading jet in the event, and the correlation matrix of the corresponding measured fiducial cross sections (right). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The first bin shows the number of events for which the variable cannot be defined, i.e., events with less than 2 jets. The bottom panel shows the ratio of the measured values to the expectations from simulation.

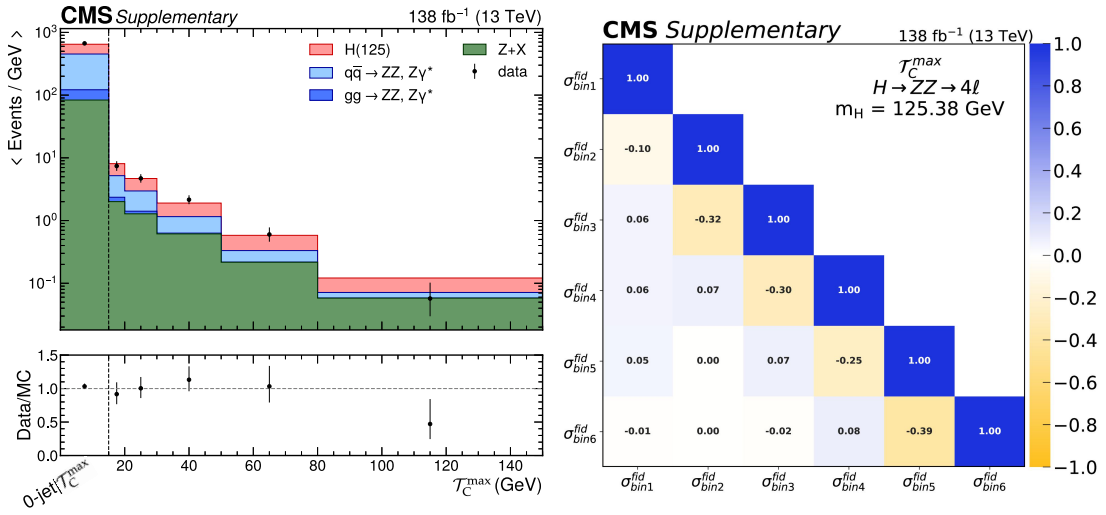

Figure C.12: Distribution of the $\mathcal{T}_C^{\max}$ (left) and the correlation matrix of the corresponding measured fiducial cross sections (right). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The first bin shows the number of events in the 0-jet|$\mathcal{T}_C^{\max}$ phase space. The bottom panel shows the ratio of the measured values to the expectations from simulation.

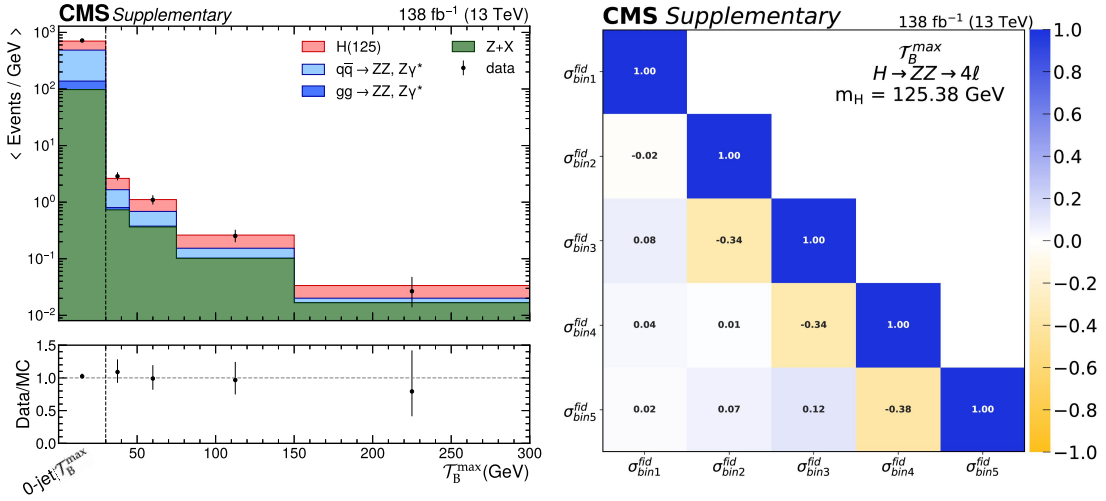

Figure C.13: Distribution of the $\mathcal{T}_B^{\max}$ (left) and the correlation matrix of the corresponding measured fiducial cross sections (right). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The first bin shows the number of events in the 0-jet|$\mathcal{T}_B^{\max}$ phase space. The bottom panel shows the ratio of the measured values to the expectations from simulation.

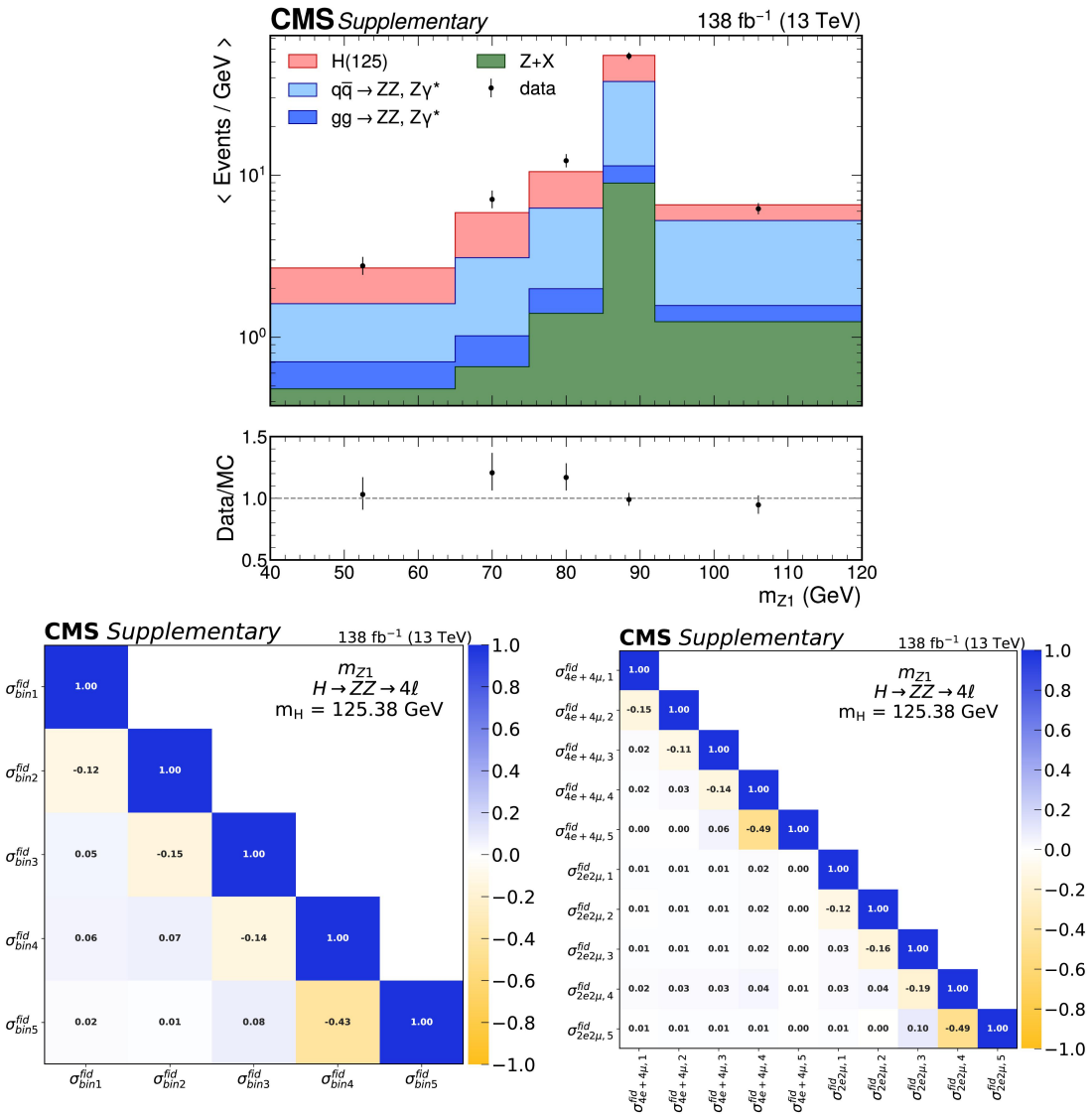

Figure C.14: Distribution of the m_{Z1} (top) and the correlation matrices of the corresponding measured fiducial cross sections (bottom). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The bottom panel shows the ratio of the measured values to the expectations from simulation.

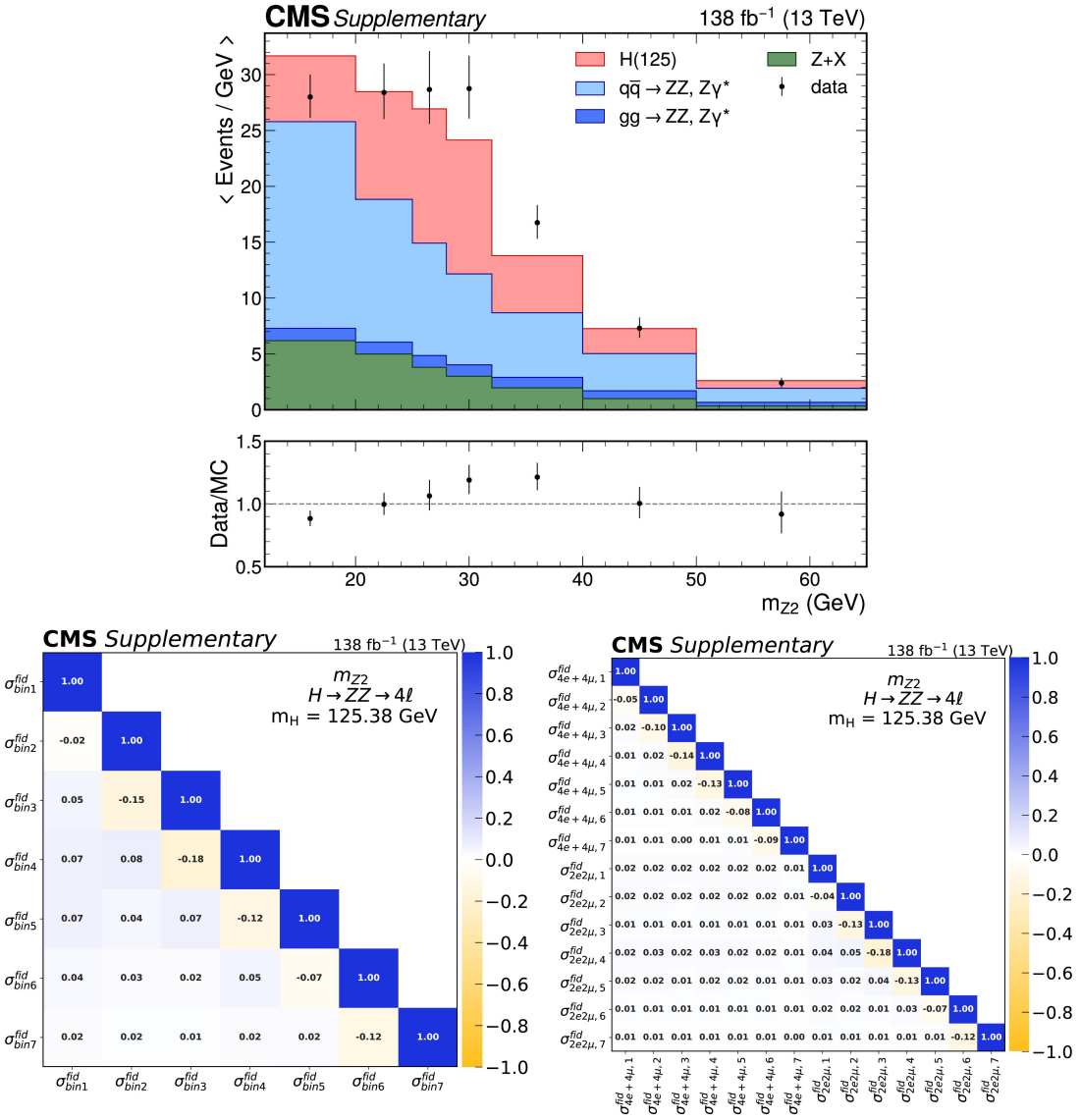

Figure C.15: Distribution of the m_{Z2} (top) and the correlation matrices of the corresponding measured fiducial cross sections (bottom). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The bottom panel shows the ratio of the measured values to the expectations from simulation.

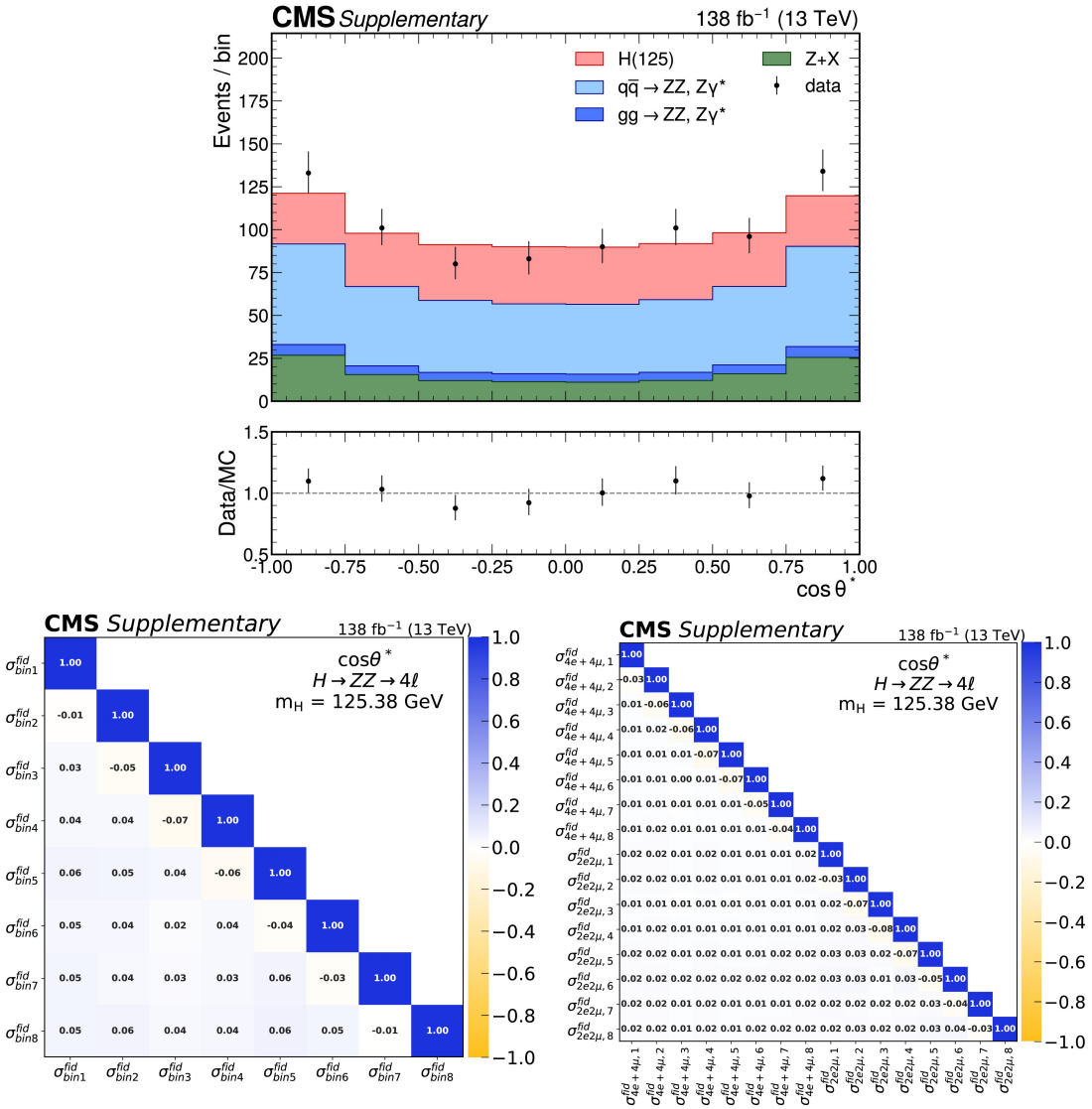

Figure C.16: Distribution of $\cos\theta^*$ (top) and the correlation matrices of the corresponding measured fiducial cross sections (bottom). Points with error bars represent the data, solid histograms the predictions from simulation. The bottom panel shows the ratio of the measured values to the expectations from simulation.

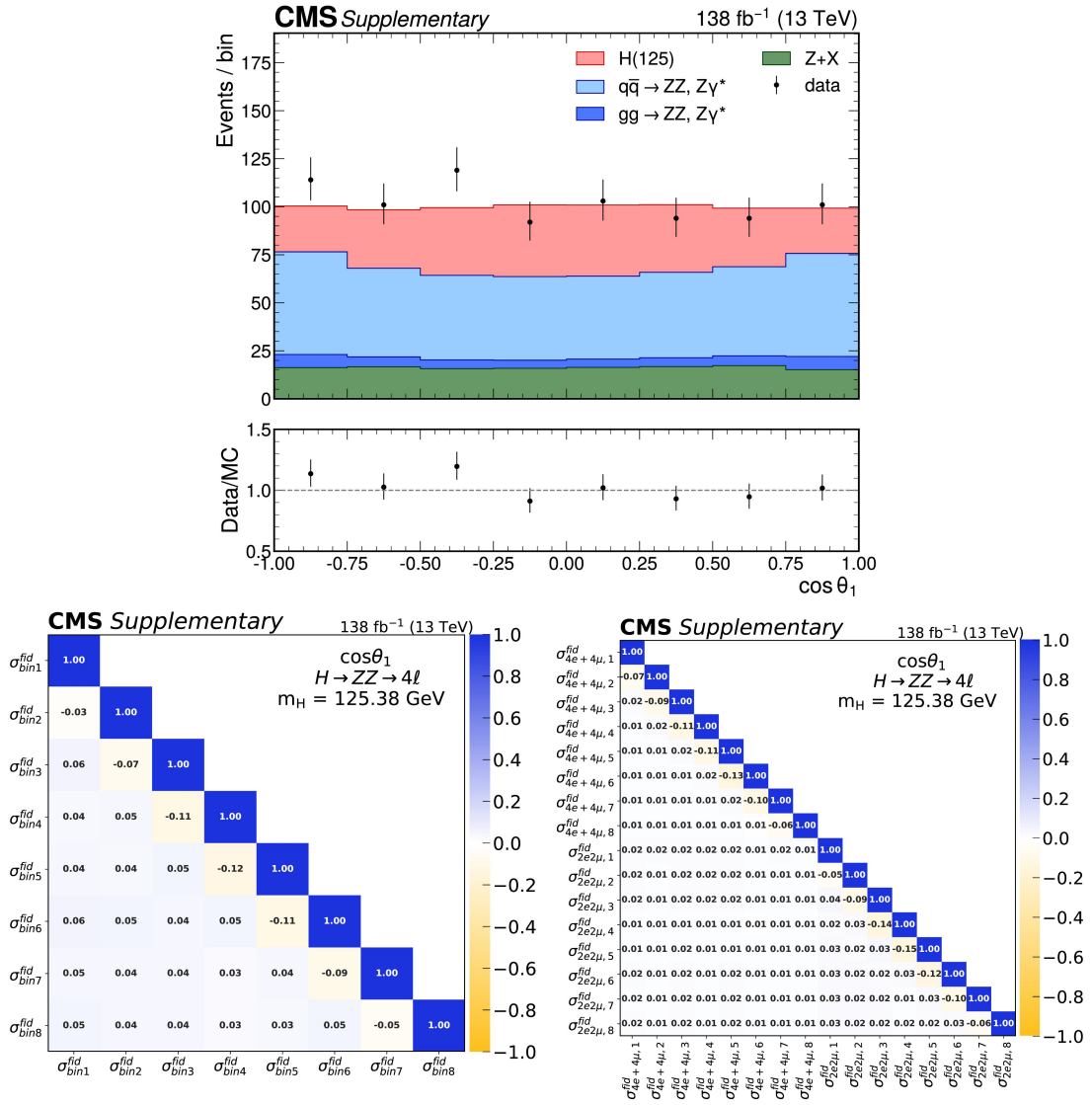

Figure C.17: Distribution of $\cos\theta_1$ (top) and the correlation matrices of the corresponding measured fiducial cross sections (bottom). Points with error bars represent the data, solid histograms the predictions from simulation. The bottom panel shows the ratio of the measured values to the expectations from simulation.

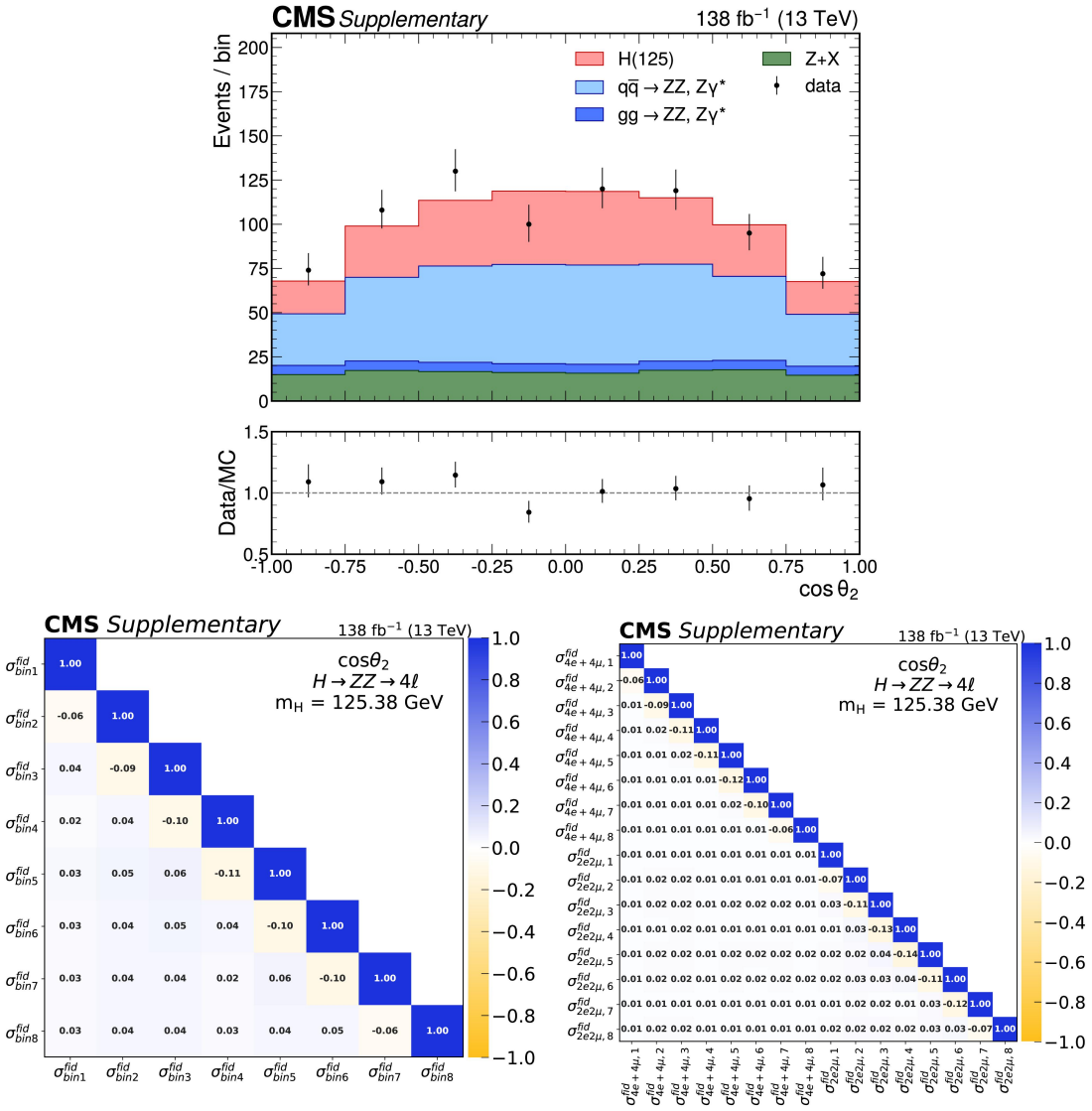

Figure C.18: Distribution of $\cos\theta_2$ (top) and the correlation matrices of the corresponding measured fiducial cross sections (bottom). Points with error bars represent the data, solid histograms the predictions from simulation. The bottom panel shows the ratio of the measured values to the expectations from simulation.

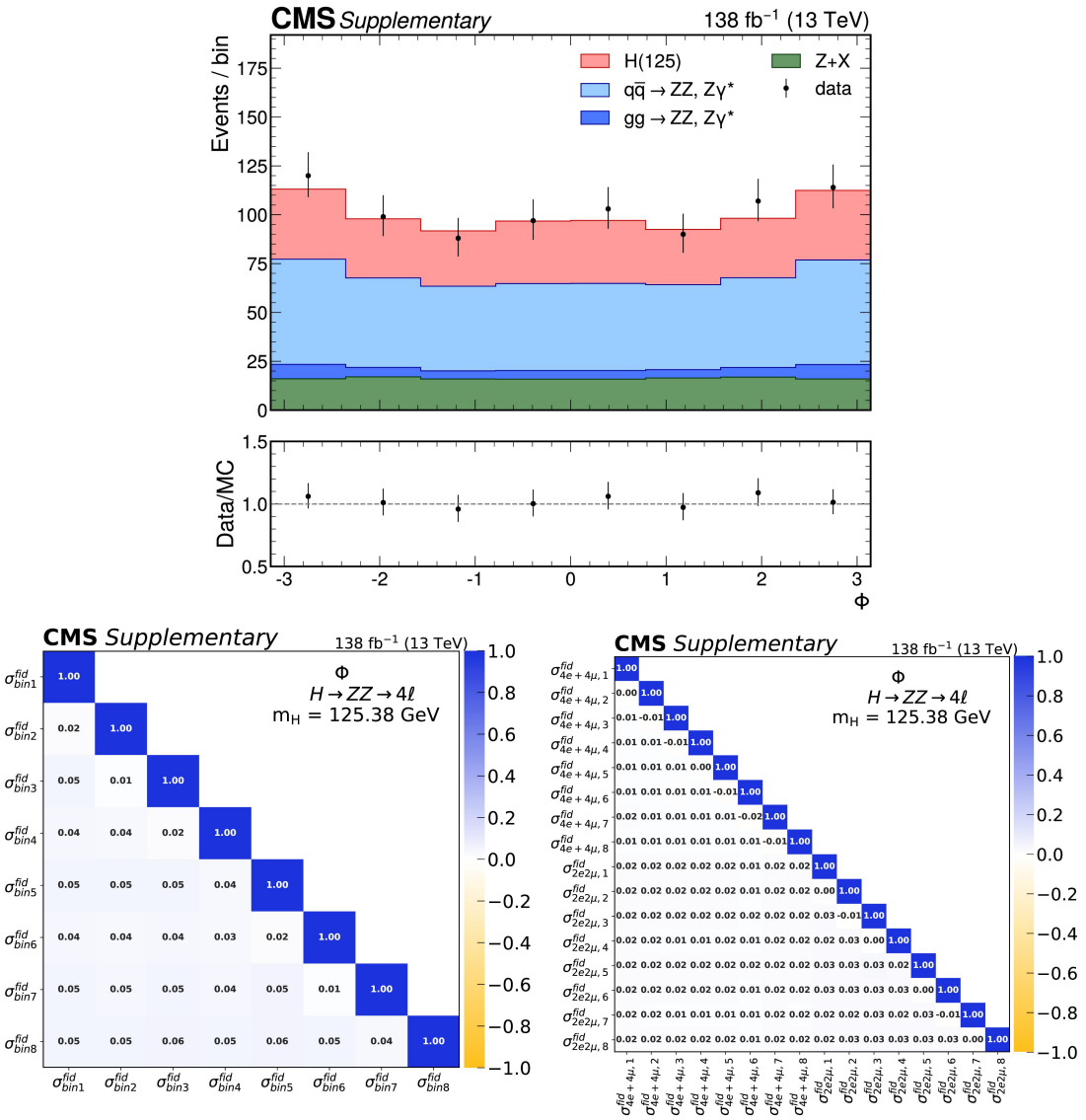

Figure C.19: Distribution of the ϕ angle (top) and the correlation matrices of the corresponding measured fiducial cross sections (bottom). Points with error bars represent the data, solid histograms the predictions from simulation. The bottom panel shows the ratio of the measured values to the expectations from simulation.

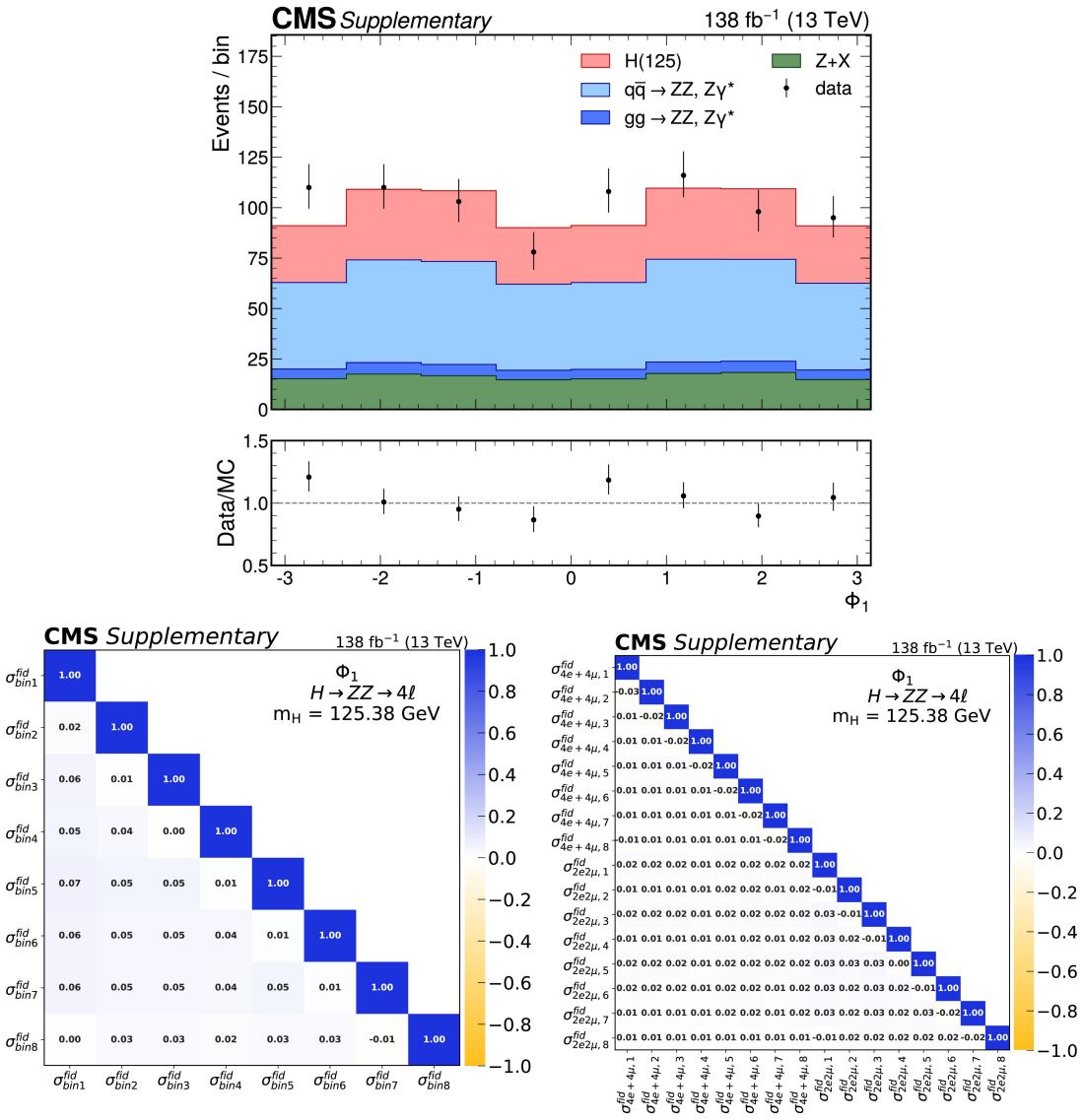

Figure C.20: Distribution of the ϕ_1 angle (top) and the correlation matrices of the corresponding measured fiducial cross sections (bottom). Points with error bars represent the data, solid histograms the predictions from simulation. The bottom panel shows the ratio of the measured values to the expectations from simulation.

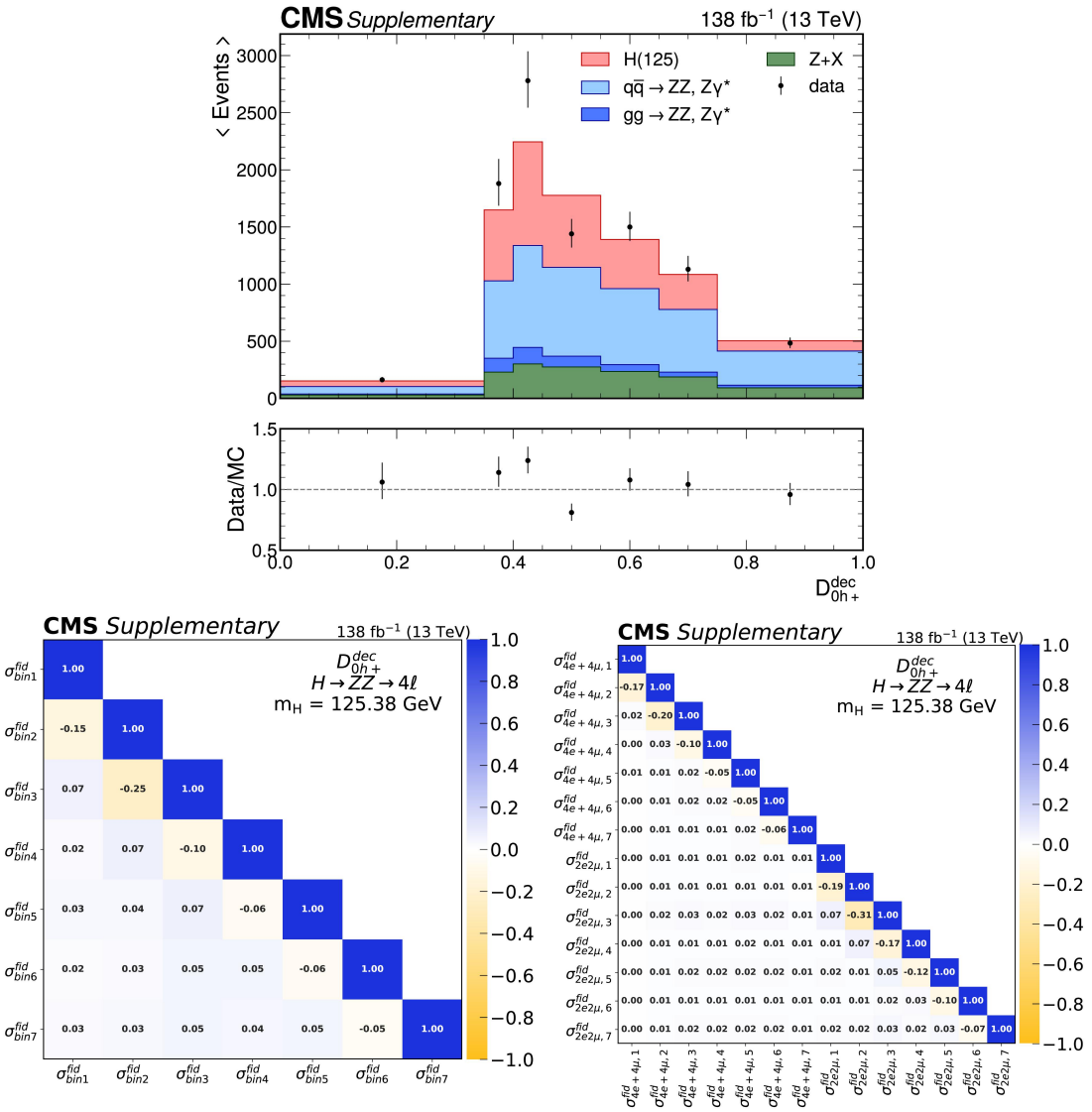

Figure C.21: Distribution of the $D^{\text{dec}}_{0\text{h}+}$ discriminant targeting the a_2 coupling (top) and the correlation matrices of the corresponding measured fiducial cross sections (bottom). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The bottom panel shows the ratio of the measured values to the expectations from simulation.

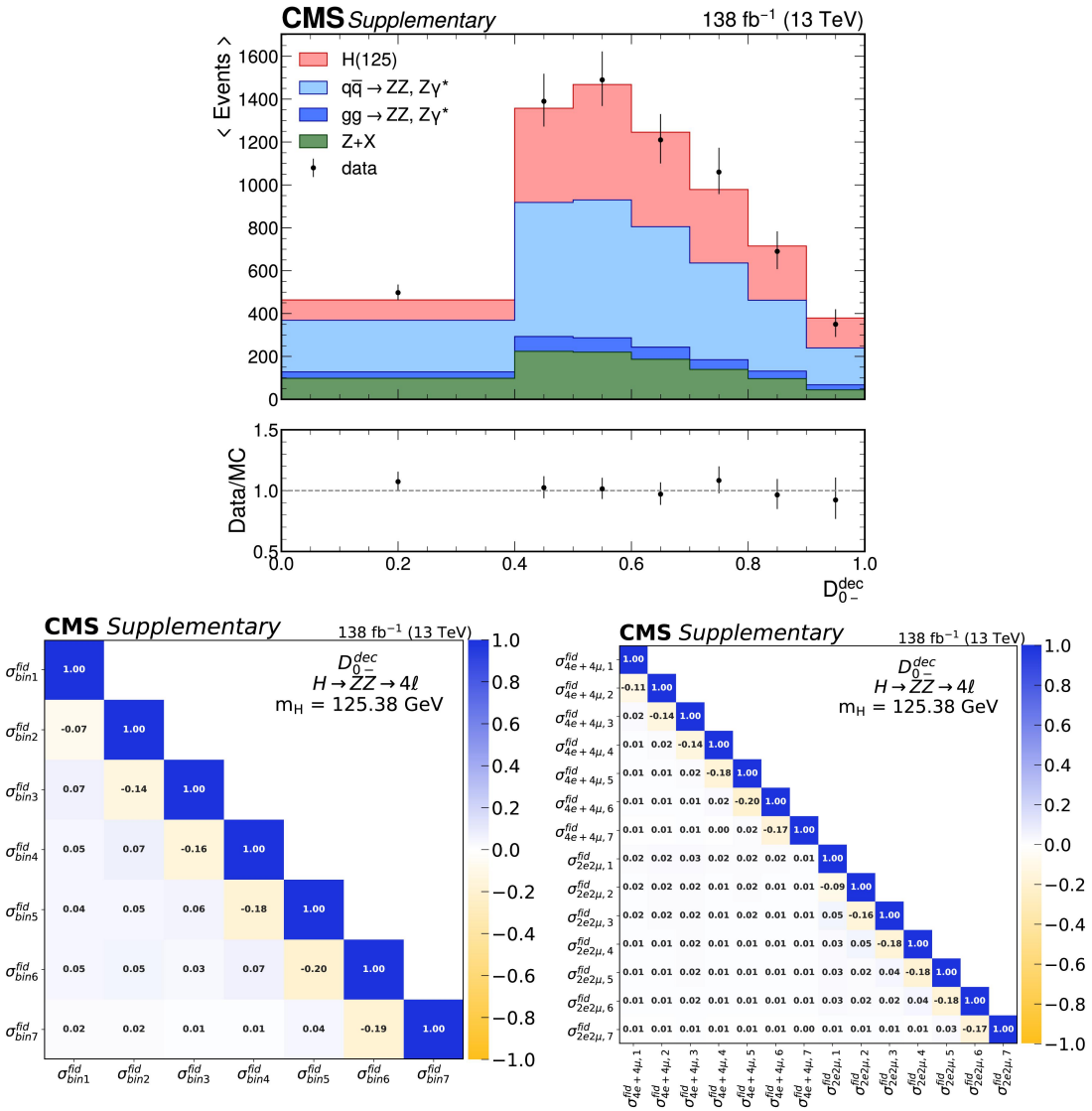

Figure C.22: Distribution of the D_{0-}^{dec} discriminant targeting the a_3 coupling (top) and the correlation matrices of the corresponding measured fiducial cross sections (bottom). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The bottom panel shows the ratio of the measured values to the expectations from simulation.

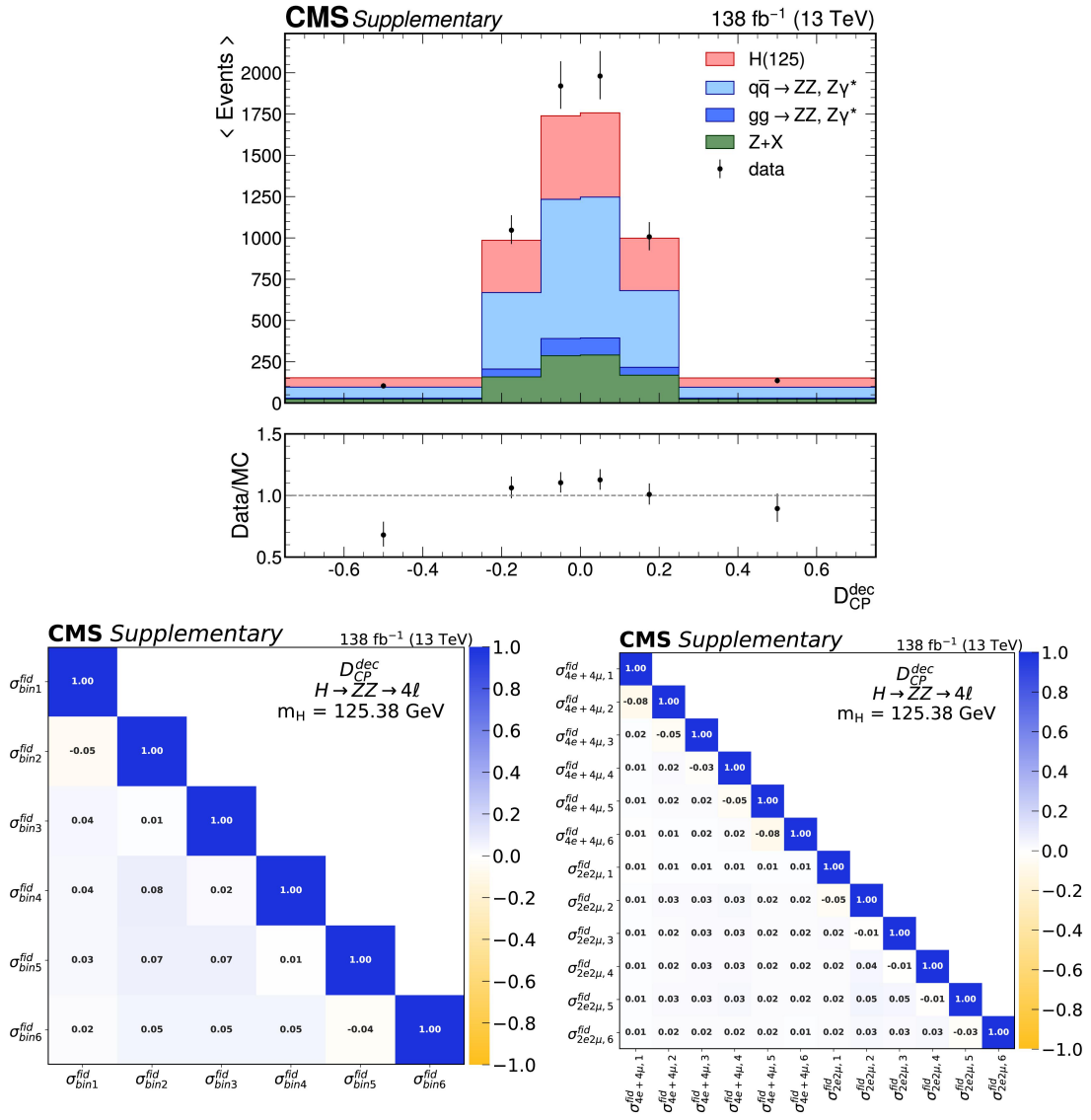

Figure C.23: Distribution of the interference discriminant $D_{\mathrm{CP}}^{\mathrm{dec}}$ targeting the a_3 coupling (top) and the correlation matrices of the corresponding measured fiducial cross sections (bottom). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The bottom panel shows the ratio of the measured values to the expectations from simulation.

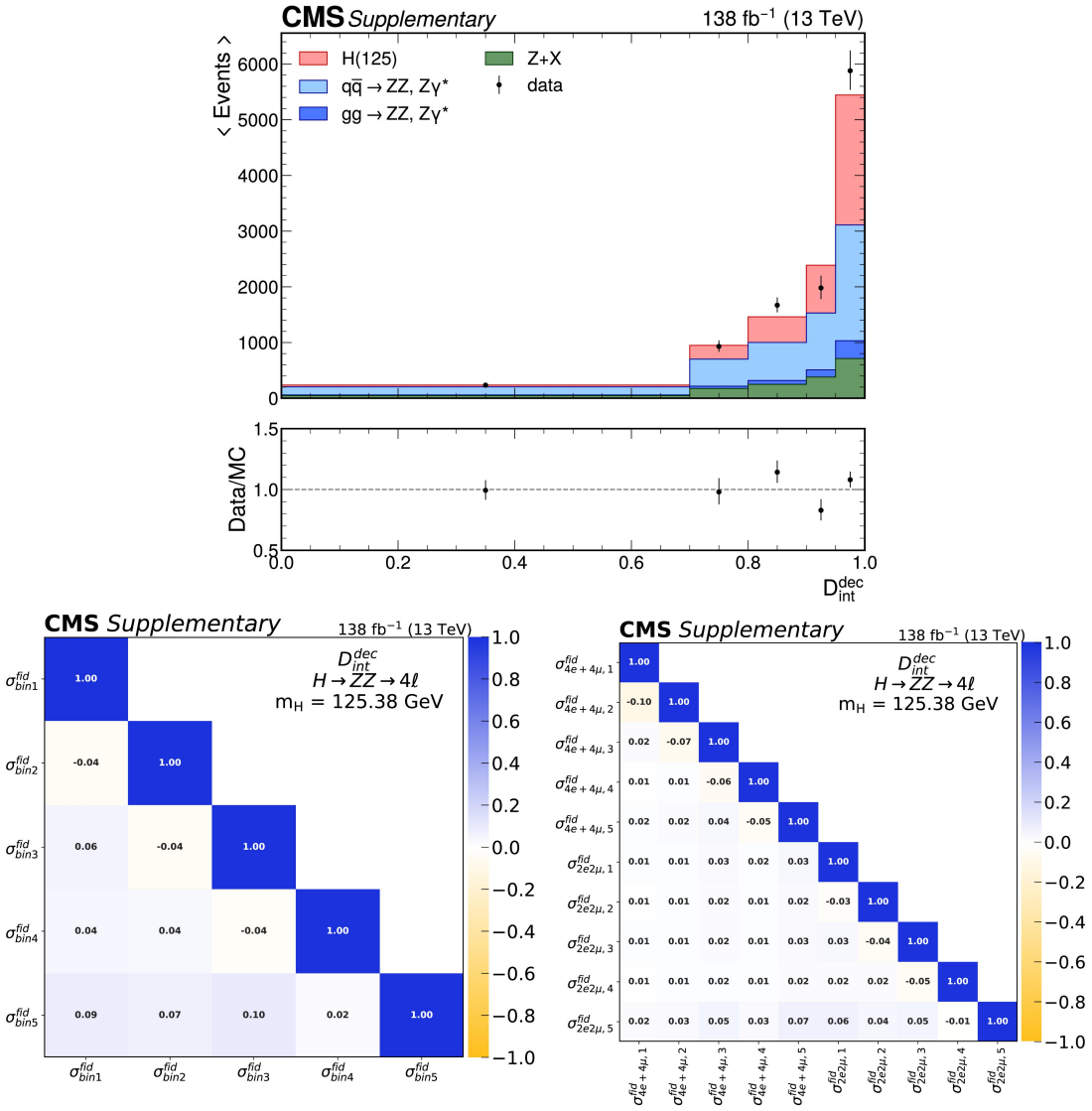

Figure C.24: Distribution of the interference discriminant $D_{\text{int}}^{\text{dec}}$ targeting the a_2 coupling (top) and the correlation matrices of the corresponding measured fiducial cross sections (bottom). Points with error bars represent the data, solid histograms the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The bottom panel shows the ratio of the measured values to the expectations from simulation.

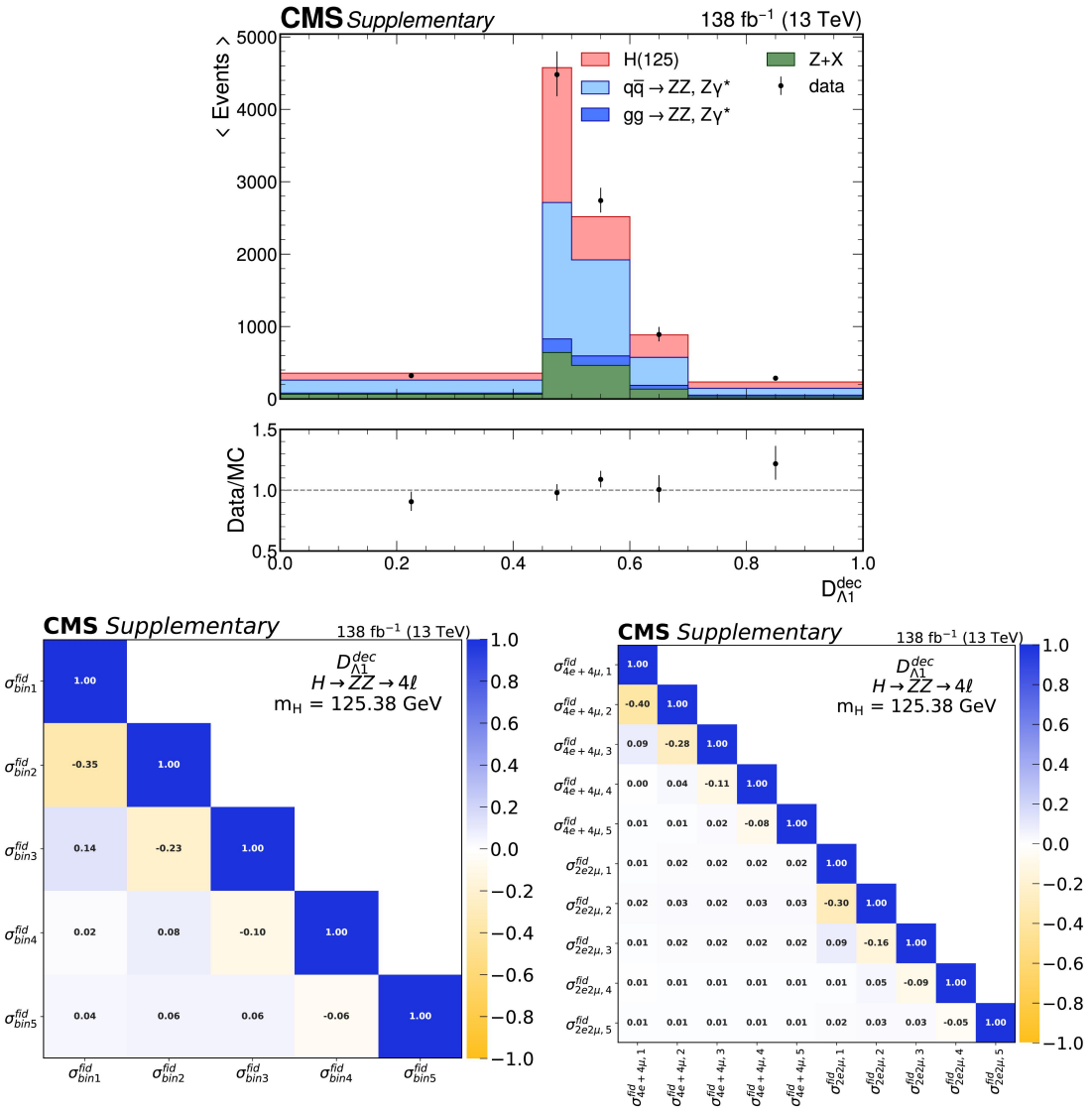

Figure C.25: Distribution of the interference discriminant $D_{\Lambda 1}^{\text{dec}}$ targeting the Λ_1 coupling (top) and the correlation matrices of the corresponding measured fiducial cross sections (bottom). Points with error bars represent the data, solid histograms represent the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The bottom panel shows the ratio of the measured values to the expectations from simulation.

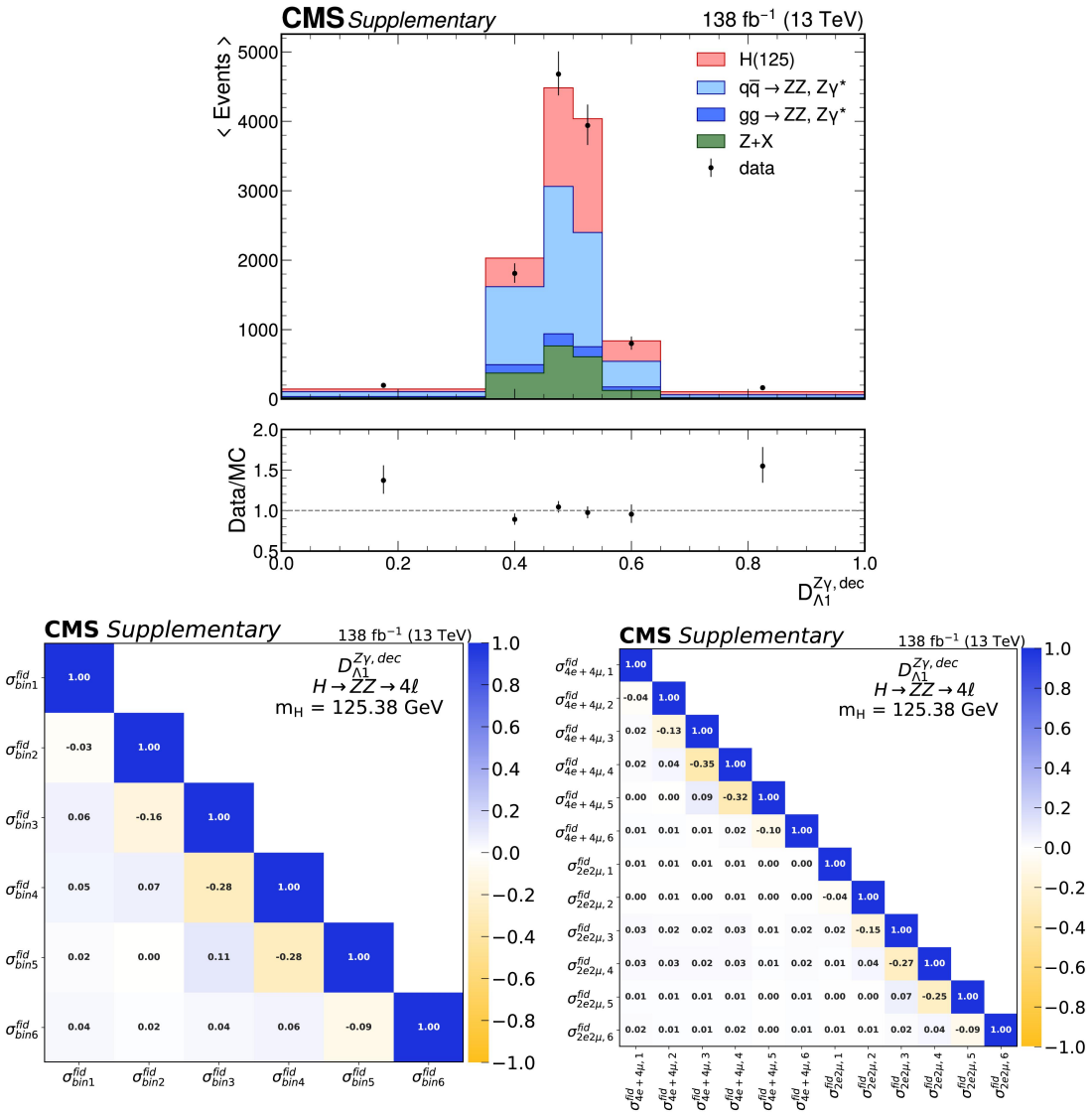

Figure C.26: Distribution of the interference discriminant $D_{\Lambda 1}^{Z\gamma,\mathrm{dec}}$ targeting the $\Lambda_1^{Z\gamma}$ coupling (top) and the correlation matrices of the corresponding measured fiducial cross sections (bottom). Points with error bars represent the data, solid histograms represent the predictions from simulation. The y-axis of the top panel has been rescaled to display the number of events per bin divided by the width of each bin. The bottom panel shows the ratio of the measured values to the expectations from simulation.

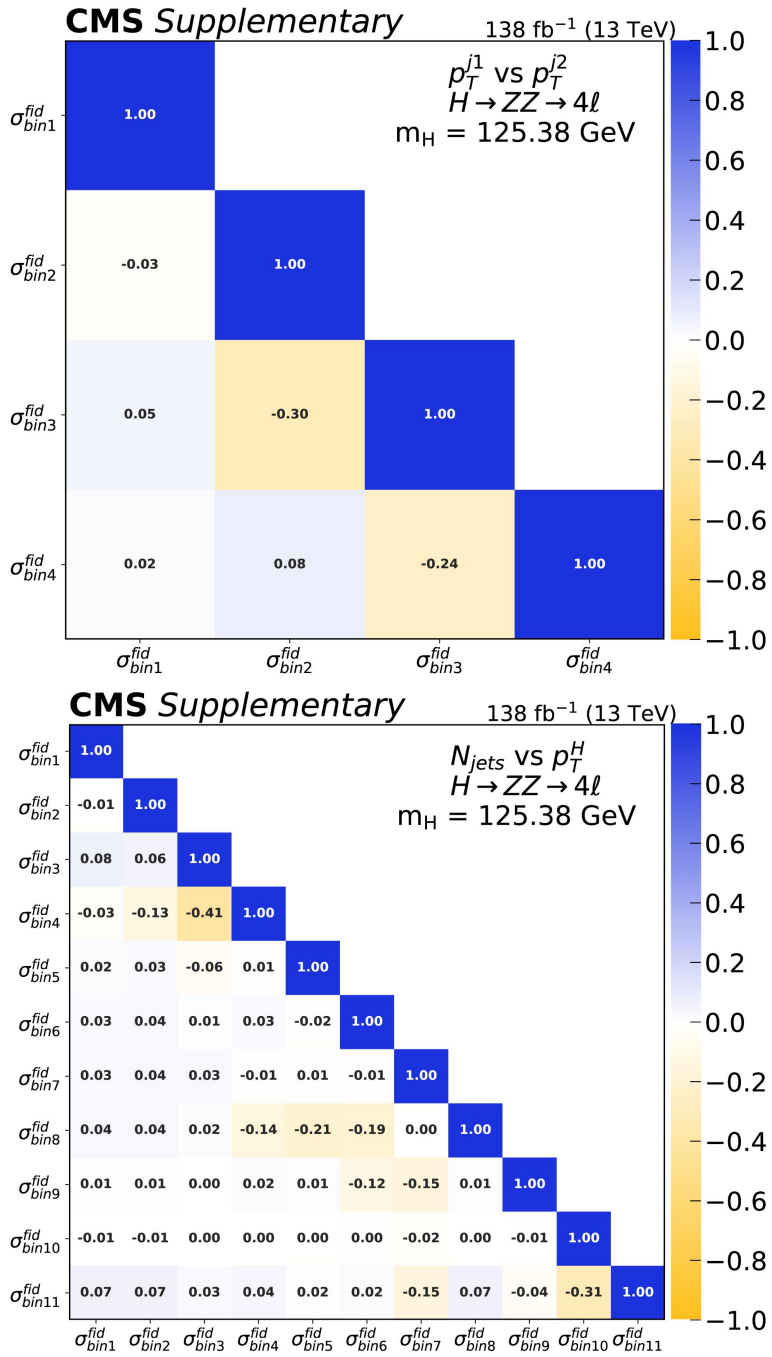

Figure C.27: Correlation matrices of the fiducial cross sections measured in bins of p_T^{j1} vs p_T^{j2} (top) and p_T^H vs N_{jets} (bottom)

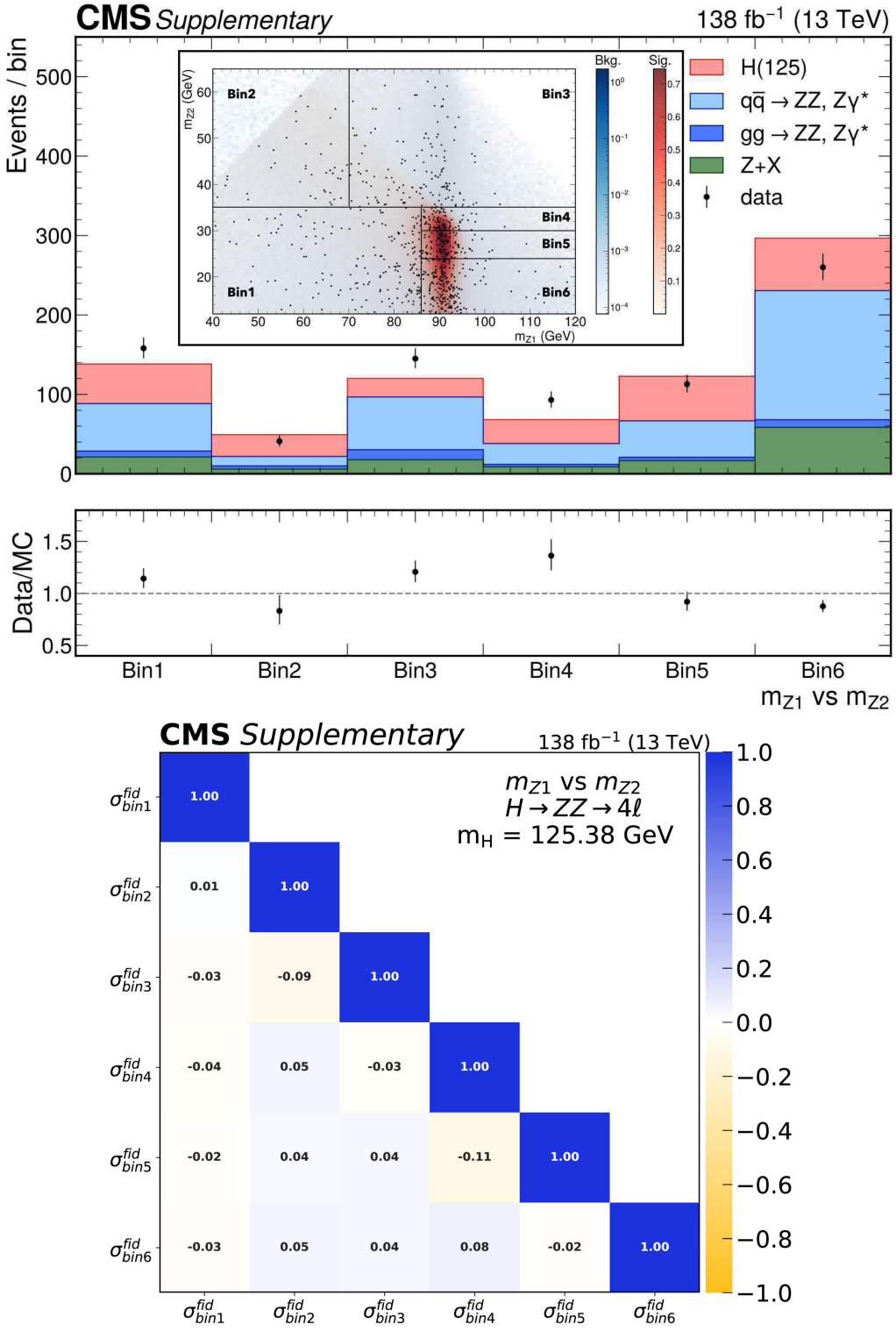

Figure C.28: Distribution of the double-differential variable m_{Z1} vs m_{Z2} (top) and the correlation matrices of the corresponding measured fiducial cross sections (bottom). Points with error bars represent the data, solid histograms the predictions from simulation. The bottom panel shows the ratio of the measured values to the expectations from simulation.

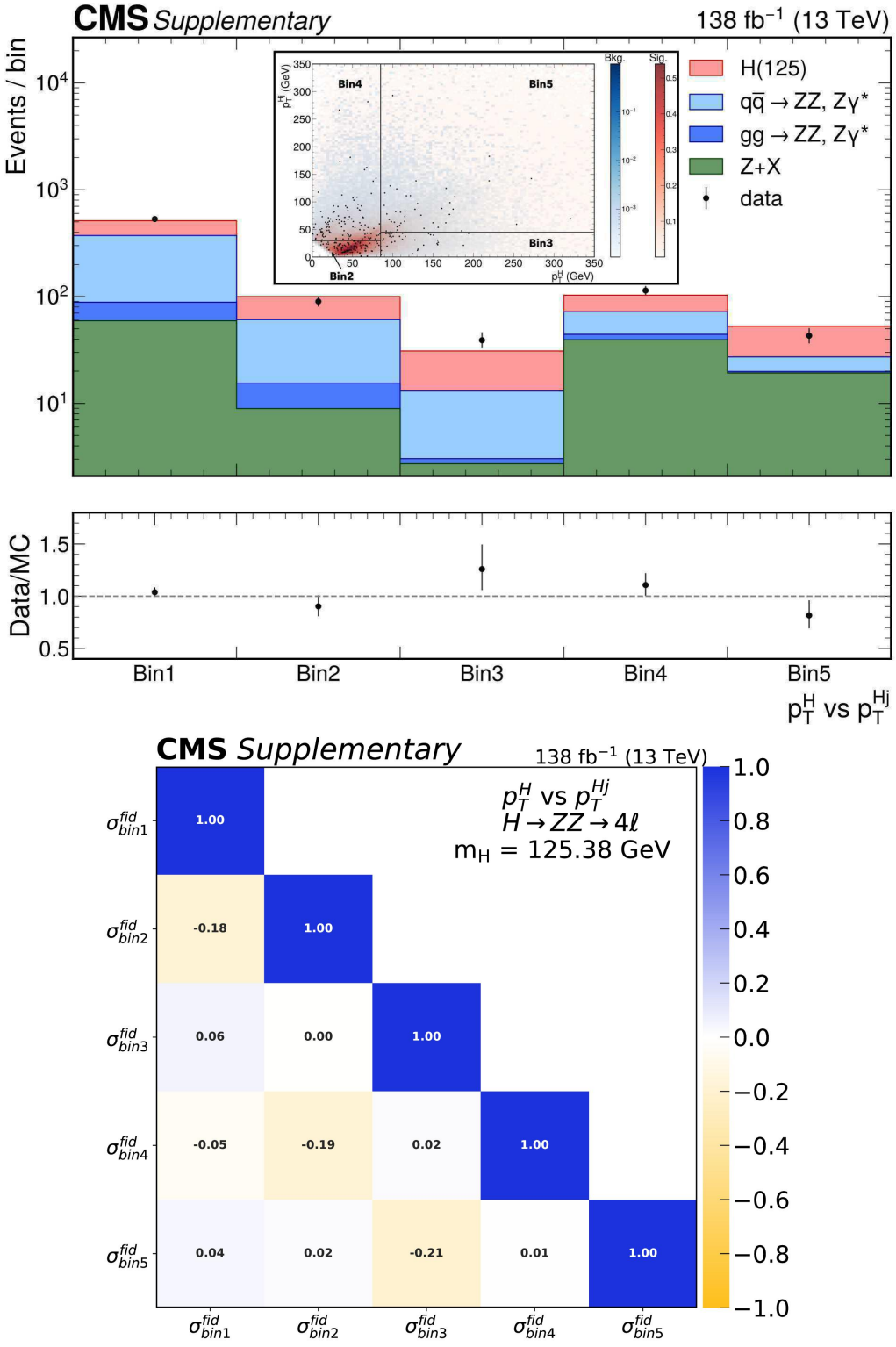

Figure C.29: Distribution of the double-differential variable p_T^H vs p_T^{Hj} (top) and the correlation matrix of the corresponding measured fiducial cross sections (bottom). Points with error bars represent the data, solid histograms the predictions from simulation. The bottom panel shows the ratio of the measured values to the expectations from simulation. On the left panel, the first bin (bin1), which is not displayed in the insert, contains events for which the p_T^{Hj} variable cannot be defined, i.e., events with 0 jets.

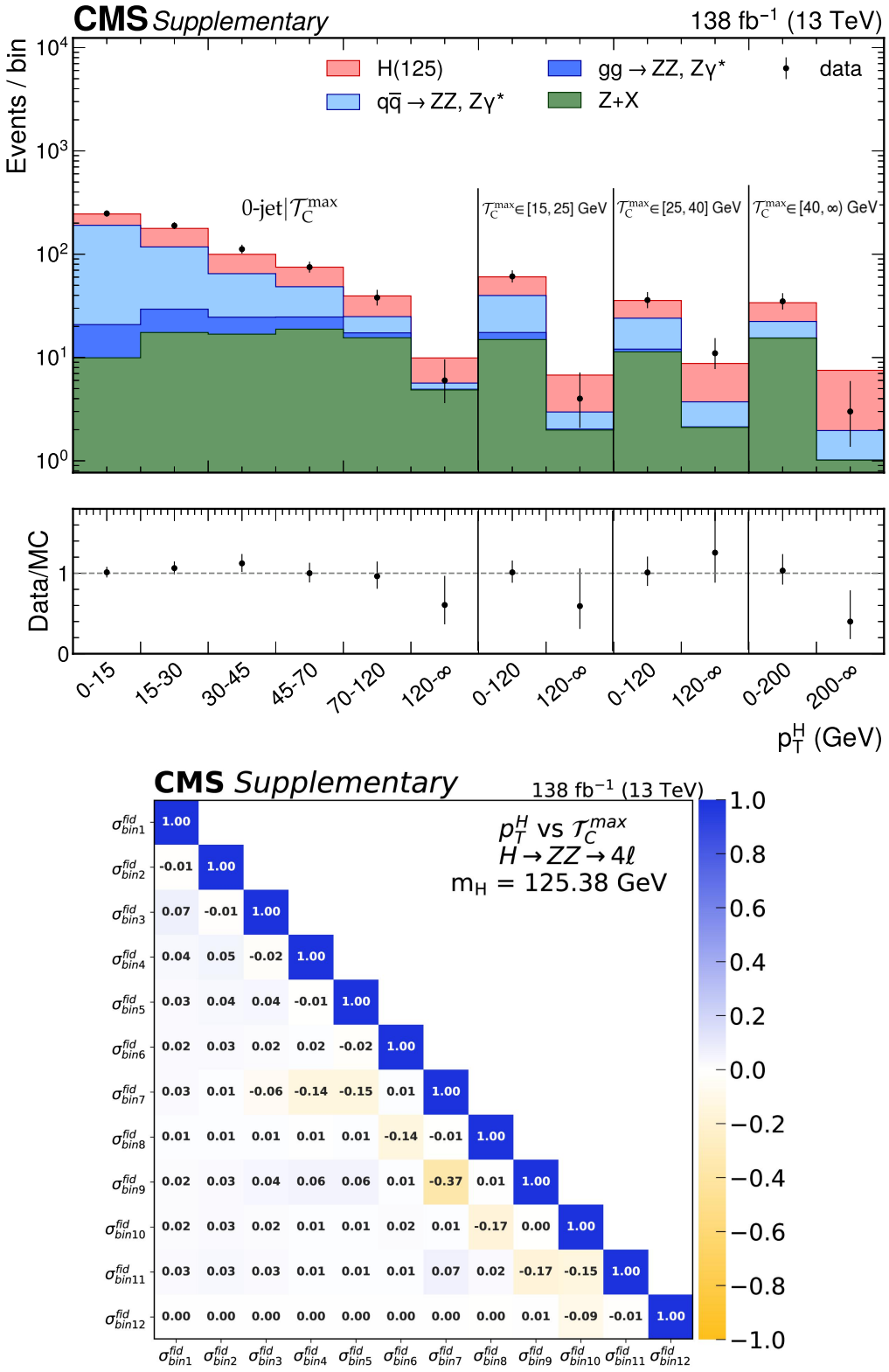

Figure C.30: Distribution of the double-differential variable p_T^H vs $\mathcal{T}_C^{\max}$ (top) and the correlation matrix of the corresponding measured fiducial cross sections (bottom). Points with error bars represent the data, solid histograms represent the predictions from simulation. The bottom panel shows the ratio of the measured values to the expectations from simulation.

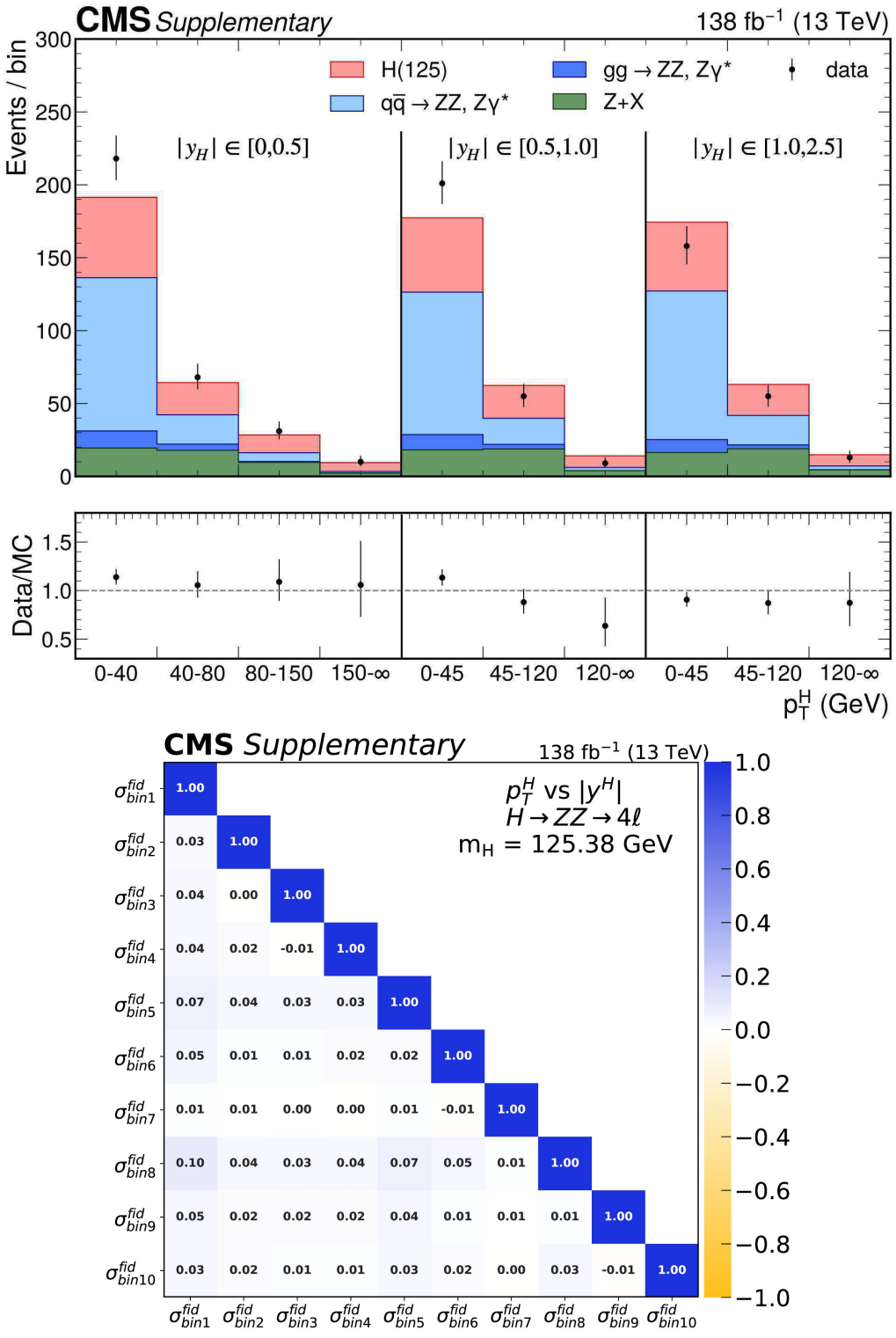

Figure C.31: Distribution of the double-differential variable p_T^H vs $|y_H|$ (top) and the correlation matrix of the corresponding measured fiducial cross sections (bottom). Points with error bars represent the data, solid histograms the predictions from simulation. The bottom panel shows the ratio of the measured values to the expectations from simulation.

Acronyms

2HDM Two Higgs Doublet Model

ALICE A Large Ion Collider Experiment

AMS Asimov Median Significance

ATLAS A Toroidal LHC Apparatus

BDT Boosted Decision Tree

BEH Brout-Englert-Higgs

BSM Beyond the Standard Model

CA Cellular Automaton

CC Charged Currents

CERN Conseil Européen pour la Recherche Nucléaire

CLUE CLUstering of Energy

CKM Cabibbo-Kobayashi-Maskawa

CMS Compact Muons Solenoid

CR Control Region

DNN Deep Neural Network

ECAL Electromagnetic CALorimeter

EFT Effective Field Theory

EOY End-Of-Year

FPR False Positive Rateption

FSR Final State Radiation

GSF Gaussian-Sum Filter

GWS Glashow-Weinberg-Salam

HCAL Hadronic CALorimeter

HGCAL High Granularity CALorimeter

HL-LHC High-Luminosity Large Hadron Collider

ID IDentification

ISR Intersecting Storage Rings

JEC Jet Energy Corrections

JER Jet Energy Resolution

JES Jet Energy Scale

LEP Large Electron Positron collider

LC Layer Cluster

LHC Large Hadron Collider

LHCb LHC beauty

LHCHWG LHC Higgs Working Group

LO Leading Order

LS Long Shutdown

HLT High-Level Trigger

MC Monte Carlo

MELA Matrix Element Likelihood Approach

MET Missing transverse energy

MIP Minimum Ionising Particle

NC Neutral Currents

NLO Next-to-Leading Order

NNLO Next-to-Next-to-Leading Order

N^{3}LO Next-to-Next-to-Next-to-Leading Order

OS Opposite Sign

PCA Principal Component Analysis

PD Primary Dataset

PDF Parton Density Function

pdf probability density function

PDG Particle Data Group

PF Particle Flow

PID Particle IDentification

pQCD perturbative Quantum ChromoDynamics

PS Proton Synchrotron

PU PileUp

PV Primary Vertex

QCD Quantum ChromoDynamics

QFT Quantum Filed Theory

RMS Root-Mean-Square

SC SynchroCyclotron

SIP Significance of the Impact Parameter

SM Standard Model

SPS Super Proton Synchrotron

SR Signal Region

SS Same Sign

STXS Simplified Template Cross Section

TICL The Iteratice CLustering

TnP Tag and Probe

TPR True Positive Rate

UL Ultra Legacy

Bibliography

[1] *Annual Register 1896, p. 159*. URL: https://books.google.fr/books?id=HysXAAAAYAAJ&pg=PA159&redir_esc=y#v=onepage&q&f=false (cit. on p. 1).

[2] Right. Hon. Lord Kelvin G.C.V.O. D.C.L. LL.D. F.R.S. M.R.I. "I. Nineteenth century clouds over the dynamical theory of heat and light". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.7 (1901), pp. 1–40. DOI: 10.1080/14786440109462664 (cit. on p. 1).

[3] A. A. Michelson and E. W. Morley. "On the relative motion of the Earth and the luminiferous ether". *American Journal of Science* s3-34.203 (1887), pp. 333–345. ISSN: 0002-9599. DOI: 10.2475/ajs.s3-34.203.333. eprint: https://www.ajsonline.org/content/s3-34/203/333.full.pdf. URL: https://www.ajsonline.org/content/s3-34/203/333 (cit. on p. 1).

[4] J. J. Thomson M.A. F.R.S. "XL. Cathode Rays". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 44.269 (1897), pp. 293–316. DOI: 10.1080/14786449708621070 (cit. on p. 1).

[5] Steven Weinberg. "The Making of the standard model". *Eur. Phys. J. C* 34 (2004), pp. 5–13. DOI: 10.1140/epjc/s2004-01761-1. arXiv: hep-ph/0401010 (cit. on pp. 1, 2).

[6] Murray Gell-Mann. "A Schematic Model of Baryons and Mesons". *Phys. Lett.* 8 (1964), pp. 214–215. DOI: 10.1016/S0031-9163(64)92001-3 (cit. on p. 1).

[7] G. Zweig. "An SU(3) model for strong interaction symmetry and its breaking. Version 2". *DEVELOPMENTS IN THE QUARK THEORY OF HADRONS. VOL. 1. 1964 - 1978*. Ed. by D. B. Lichtenberg and Simon Peter Rosen. Feb. 1964, pp. 22–101 (cit. on p. 1).

[8] Chen-Ning Yang and Robert L. Mills. "Conservation of Isotopic Spin and Isotopic Gauge Invariance". *Phys. Rev.* 96 (1954). Ed. by Jong-Ping Hsu and D. Fine, pp. 191–195. DOI: 10.1103/PhysRev.96.191 (cit. on p. 2).

[9] P.W. Higgs. "Broken symmetries, massless particles and gauge fields". *Physics Letters* 12.2 (1964), pp. 132–133. ISSN: 0031-9163. DOI: https://doi.org/10.1016/0031-9163(64)91136-9 (cit. on p. 2).

[10] Peter W. Higgs. "Spontaneous Symmetry Breakdown without Massless Bosons". *Phys. Rev.* 145 (1966), pp. 1156–1163. DOI: 10.1103/PhysRev.145.1156 (cit. on p. 2).

[11] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons". *Phys. Rev. Lett.* 13 (1964). Ed. by J. C. Taylor, pp. 321–323. DOI: 10.1103/PhysRevLett.13.321 (cit. on p. 2).

[12] Steven Weinberg. "A Model of Leptons". *Phys. Rev. Lett.* 19 (1967), pp. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264 (cit. on p. 2).

[13] Abdus Salam. "Weak and Electromagnetic Interactions". *Conf. Proc. C* 680519 (1968), pp. 367–377. DOI: 10.1142/9789812795915_0034 (cit. on p. 2).

[14] Hideki Yukawa. "On the Interaction of Elementary Particles I". *Proc. Phys. Math. Soc. Jap.* 17 (1935), pp. 48–57. DOI: 10.1143/PTPS.1.1 (cit. on p. 2).

[15] F. J. Hasert et al. "Observation of Neutrino Like Interactions Without Muon Or Electron in the Gargamelle Neutrino Experiment". *Phys. Lett. B* 46 (1973), pp. 138–140. DOI: 10.1016/0370-2693(73)90499-1 (cit. on p. 2).

[16] F. Abe et al. "Observation of top quark production in $\bar{p}p$ collisions". *Phys. Rev. Lett.* 74 (1995), pp. 2626–2631. DOI: 10.1103/PhysRevLett.74.2626. arXiv: hep-ex/9503002 (cit. on p. 2).

[17] S. Abachi et al. "Observation of the top quark". *Phys. Rev. Lett.* 74 (1995), pp. 2632–2637. DOI: 10.1103/PhysRevLett.74.2632. arXiv: hep-ex/9503003 (cit. on p. 2).

[18] John R. Ellis, Mary K. Gaillard, and Dimitri V. Nanopoulos. "A Phenomenological Profile of the Higgs Boson". *Nucl. Phys. B* 106 (1976), p. 292. DOI: 10.1016/0550-3213(76)90382-5 (cit. on p. 2).

[19] 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), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex] (cit. on pp. 3, 19, 28, 32, 65).

[20] 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), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex] (cit. on pp. 3, 19, 28, 32, 65).

[21] CMS Collaboration. "Measurements of inclusive and differential cross sections for the Higgs boson production and decay to four-leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV". *JHEP* 08 (2023), p. 040. DOI: 10.1007/JHEP08(2023)040. arXiv: 2305.07532 [hep-ex] (cit. on pp. 4, 66, 128).

[22] CMS Collaboration. "Reconstruction of unconverted photons in the presence of high pileup at the CMS Phase-2 High Granularity Endcap Calorimeter" (2022). URL: https://cds.cern.ch/record/2841536 (cit. on p. 5).

[23] Takaaki Kajita. "Nobel Lecture: Discovery of atmospheric neutrino oscillations". *Rev. Mod. Phys.* 88.3 (2016), p. 030501. DOI: 10.1103/RevModPhys.88.030501 (cit. on pp. 11, 19).

[24] KATRIN Collaboration. "Direct neutrino-mass measurement with sub-electronvolt sensitivity". *Nature Phys.* 18.2 (2022), pp. 160–166. DOI: 10.1038/s41567-021-01463-1. arXiv: 2105.08533 [hep-ex] (cit. on p. 11).

[25] Carsten Burgard. URL: https://texample.net/tikz/examples/model-physics/ (cit. on p. 12).

[26] Particle Data Group, P. A. Zyla, et al. "Review of particle physics". *Prog. Theor. Exp. Phys.* 2022 (2022), p. 083C01. DOI: 10.1093/ptep/ptac097 (cit. on pp. 19, 20, 77, 90, 94, 103).

[27] LHC Higgs Cross Section Working Group. "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector". 2/2017 (Oct. 2016). DOI: 10.23731/CYRM-2017-002. arXiv: 1610.07922 [hep-ph] (cit. on pp. 21, 65, 126, 127).

[28] David B. Kaplan and Howard Georgi. "SU(2) x U(1) Breaking by Vacuum Misalignment". *Phys. Lett. B* 136 (1984), pp. 183–186. DOI: 10.1016/0370-2693(84)91177-8 (cit. on p. 21).

[29] Roberto Contino, Leandro Da Rold, and Alex Pomarol. "Light custodians in natural composite Higgs models". *Phys. Rev. D* 75 (2007), p. 055014. DOI: 10.1103/PhysRevD.75.055014. arXiv: hep-ph/0612048 (cit. on p. 21).

[30] Roberto Contino et al. "Warped/composite phenomenology simplified". *JHEP* 05 (2007), p. 074. DOI: 10.1088/1126-6708/2007/05/074. arXiv: hep-ph/0612180 (cit. on p. 21).

[31] David B. Kaplan. "Flavor at SSC energies: A New mechanism for dynamically generated fermion masses". *Nucl. Phys. B* 365 (1991), pp. 259–278. DOI: 10.1016/S0550-3213(05)80021-5 (cit. on p. 21).

[32] Michael J. Dugan, Howard Georgi, and David B. Kaplan. "Anatomy of a composite Higgs model". *Nuclear Physics B* 254 (1985), pp. 299–326. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(85)90221-4. URL: https://www.sciencedirect.com/science/article/pii/0550321385902214 (cit. on p. 21).

[33] Simone Blasi and Florian Goertz. "Softened Symmetry Breaking in Composite Higgs Models". *Phys. Rev. Lett.* 123.22 (2019), p. 221801. DOI: 10.1103/PhysRevLett.123.221801. arXiv: 1903.06146 [hep-ph] (cit. on p. 21).

[34] G. C. Branco et al. "Theory and phenomenology of two-Higgs-doublet models". *Phys. Rept.* 516 (2012), pp. 1–102. DOI: 10.1016/j.physrep.2012.02.002. arXiv: 1106.0034 [hep-ph] (cit. on p. 21).

[35] Sacha Davidson, Enrico Nardi, and Yosef Nir. "Leptogenesis". *Phys. Rept.* 466 (2008), pp. 105–177. DOI: 10.1016/j.physrep.2008.06.002. arXiv: 0802.2962 [hep-ph] (cit. on p. 22).

[36] Peisi Huang, Andrew J. Long, and Lian-Tao Wang. "Probing the Electroweak Phase Transition with Higgs Factories and Gravitational Waves". *Phys. Rev. D* 94.7 (2016), p. 075008. DOI: 10.1103/PhysRevD.94.075008. arXiv: 1608.06619 [hep-ph] (cit. on p. 22).

[37] A. D. Sakharov. "Violation of CP Invariance, C asymmetry, and baryon asymmetry of the universe". *Pisma Zh. Eksp. Teor. Fiz.* 5 (1967), pp. 32–35. DOI: 10.1070/PU1991v034n05ABEH002497 (cit. on p. 22).

[38] Stephen M. Barr and A. Zee. "A New Approach to the electron-Muon Mass Ratio". *Phys. Rev. D* 15 (1977), p. 2652. DOI: 10.1103/PhysRevD.15.2652 (cit. on p. 22).

[39] J. D. Bjorken and Steven Weinberg. "A Mechanism for Nonconservation of Muon Number". *Phys. Rev. Lett.* 38 (1977), p. 622. DOI: 10.1103/PhysRevLett.38.622 (cit. on p. 22).

[40] K. S. Babu and S. Nandi. "Natural fermion mass hierarchy and new signals for the Higgs boson". *Phys. Rev. D* 62 (2000), p. 033002. DOI: 10.1103/PhysRevD.62.033002. arXiv: hep-ph/9907213 (cit. on p. 22).

[41] CMS Collaboration. "A portrait of the Higgs boson by the CMS experiment ten years after the discovery". *Nature* 607.7917 (2022), pp. 60–68. DOI: 10.1038/s41586-022-04892-x. arXiv: 2207.00043 [hep-ex] (cit. on pp. 23, 26, 29, 53, 155).

[42] CMS Collaboration. "Measurements of production cross sections of the Higgs boson in the four-lepton final state in proton–proton collisions at $\sqrt{s} = 13\,\text{TeV}$". *Eur. Phys. J. C* 81.6 (2021), p. 488. DOI: 10.1140/epjc/s10052-021-09200-x. arXiv: 2103.04956 [hep-ex] (cit. on pp. 23, 108, 128, 172).

[43] CMS Collaboration. "Measurement of the Higgs boson inclusive and differential fiducial production cross sections in the diphoton decay channel with pp collisions at $\sqrt{s} = 13$ TeV". *JHEP* 07 (2023), p. 091. DOI: 10.1007/JHEP07(2023)091. arXiv: 2208.12279 [hep-ex] (cit. on p. 23).

[44] Stefan Schmitt. "Data Unfolding Methods in High Energy Physics". *EPJ Web Conf.* 137 (2017). Ed. by Y. Foka, N. Brambilla, and V. Kovalenko, p. 11008. DOI: 10.1051/epjconf/201713711008. arXiv: 1611.01927 [physics.data-an] (cit. on p. 24).

[45] Andreas Hocker and Vakhtang Kartvelishvili. "SVD approach to data unfolding". *Nucl. Instrum. Meth. A* 372 (1996), pp. 469–481. DOI: 10.1016/0168-9002(95)01478-0. arXiv: hep-ph/9509307 (cit. on p. 24).

[46] Tim Adye. "Unfolding algorithms and tests using RooUnfold" (2011), pp. 313–318. DOI: 10.5170/CERN-2011-006.313. arXiv: 1105.1160 [physics.data-an] (cit. on p. 24).

[47] G. D'Agostini. "A multidimensional unfolding method based on Bayes' theorem". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 362.2 (1995), pp. 487–498. DOI: 10.1016/0168-9002(95)00274-X (cit. on p. 24).

[48] D. M. Webber et al. "Measurement of the Positive Muon Lifetime and Determination of the Fermi Constant to Part-per-Million Precision". *Phys. Rev. Lett.* 106 (2011), p. 041803. DOI: 10.1103/PhysRevLett.106.079901. arXiv: 1010.0991 [hep-ex] (cit. on p. 26).

[49] Giuseppe Degrassi et al. "Probing the Higgs self coupling via single Higgs production at the LHC". *JHEP* 12 (2016), p. 080. DOI: 10.1007/JHEP12(2016)080. arXiv: 1607.04251 [hep-ph] (cit. on pp. 27, 152, 153).

[50] Fabio Maltoni et al. "Trilinear Higgs coupling determination via single-Higgs differential measurements at the LHC". *Eur. Phys. J. C* 77 (2017), p. 887. DOI: 10.1140/epjc/s10052-017-5410-8. arXiv: 1709.08649 [hep-ph] (cit. on pp. 27, 152, 153).

[51] Stefano Di Vita et al. "A global view on the Higgs self-coupling". *JHEP* 09 (2017), p. 069. DOI: 10.1007/JHEP09(2017)069. arXiv: 1704.01953 [hep-ph] (cit. on pp. 27, 152, 153).

[52] J R Andersen et al. "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties" (July 2013). Ed. by S Heinemeyer et al. DOI: 10.5170/CERN-2013-004. arXiv: 1307.1347 [hep-ph] (cit. on pp. 27, 28).

[53] LHC Higgs Cross Section Working Group. "LHC HXSWG interim recommendations to explore the coupling structure of a Higgs-like particle" (Sept. 2012). arXiv: 1209.0040 [hep-ph] (cit. on p. 28).

[54] J. de Blas et al. "Higgs Boson Studies at Future Particle Colliders". *JHEP* 01 (2020), p. 139. DOI: 10.1007/JHEP01(2020)139. arXiv: 1905.03764 [hep-ph] (cit. on p. 28).

[55] "Search for the standard model Higgs boson at LEP". *2001 Europhysics Conference on High Energy Physics*. July 2001. arXiv: hep-ex/0107029 (cit. on p. 28).

[56] CMS Collaboration. "Constraints on the spin-parity and anomalous HVV couplings of the Higgs boson in proton collisions at 7 and 8 TeV". *Phys. Rev. D* 92.1 (2015), p. 012004. DOI: 10.1103/PhysRevD.92.012004. arXiv: 1411.3441 [hep-ex] (cit. on pp. 29, 65).

[57] CMS Collaboration. *Measurement of the Higgs boson mass and width using the four leptons final state*. Tech. rep. Geneva: CERN, 2023. URL: https://cds.cern.ch/record/2871702 (cit. on p. 29).

[58] ATLAS Collaboration. "Combined measurement of the Higgs boson mass from the $H \to \gamma\gamma$ and $H \to ZZ^* \to 4\ell$ decay channels with the ATLAS detector using $\sqrt{s}$ = 7, 8 and 13 TeV *pp* collision data" (Aug. 2023). arXiv: 2308.04775 [hep-ex] (cit. on pp. 29, 65).

[59] CMS Collaboration. "Measurement of the Higgs boson width and evidence of its off-shell contributions to ZZ production". *Nature Phys.* 18.11 (2022), pp. 1329–1334. DOI: 10.1038/s41567-022-01682-0. arXiv: 2202.06923 [hep-ex] (cit. on pp. 29, 65).

[60] ATLAS Collaboration. "Constraining the Higgs boson self-coupling from single- and double-Higgs production with the ATLAS detector using *pp* collisions at $\sqrt{s} = 13$ TeV" (Nov. 2022). arXiv: 2211.01216 [hep-ex] (cit. on pp. 29, 155, 156).

[61] CMS and ATLAS Collaboration. *Evidence for the Higgs boson decay to a Z boson and a photon at the LHC*. Tech. rep. Geneva: CERN, 2023. URL: https://cds.cern.ch/record/2859713 (cit. on p. 29).

[62] ATLAS Collaboration. "Evidence for Higgs boson decays to a low-mass dilepton system and a photon in pp collisions at s=13 TeV with the ATLAS detector". *Phys. Lett. B* 819 (2021), p. 136412. DOI: 10.1016/j.physletb.2021.136412. arXiv: 2103.10322 [hep-ex] (cit. on p. 29).

[63] CMS Collaboration. "Evidence for Higgs boson decay to a pair of muons". *JHEP* 01 (2021), p. 148. DOI: 10.1007/JHEP01(2021)148. arXiv: 2009.04363 [hep-ex] (cit. on p. 29).

[64] CMS Collaboration. "Search for Higgs boson decay to a charm quark-antiquark pair in proton-proton collisions at $\sqrt{s} = 13$ TeV" (May 2022). arXiv: 2205.05550 [hep-ex] (cit. on p. 29).

[65] ATLAS Collaboration. "Direct constraint on the Higgs-charm coupling from a search for Higgs boson decays into charm quarks with the ATLAS detector". *Eur. Phys. J. C* 82 (2022), p. 717. DOI: 10.1140/epjc/s10052-022-10588-3. arXiv: 2201.11428 [hep-ex] (cit. on p. 29).

[66] CMS Collaboration. "Search for lepton-flavor violating decays of the Higgs boson in the $\mu\tau$ and eτ final states in proton-proton collisions at $\sqrt{s} = 13$ TeV". *Phys. Rev. D* 104.3 (2021), p. 032013. DOI: 10.1103/PhysRevD.104.032013. arXiv: 2105.03007 [hep-ex] (cit. on p. 29).

[67] CMS Collaboration. "Search for the lepton-flavor violating decay of the Higgs boson and additional Higgs bosons in the eμ final state in proton-proton collisions at $\sqrt{s}$ = 13 TeV". *Phys. Rev. D* 108.7 (2023), p. 072004. DOI: 10.1103/PhysRevD.108.072004. arXiv: 2305.18106 [hep-ex] (cit. on p. 29).

[68] ATLAS Collaboration. "Searches for lepton-flavour-violating decays of the Higgs boson into $e\tau$ and $\mu\tau$ in $\sqrt{s} = 13$ TeV *pp* collisions with the ATLAS detector". *JHEP* 07 (2023), p. 166. DOI: 10.1007/JHEP07(2023)166. arXiv: 2302.05225 [hep-ex] (cit. on p. 29).

[69] ATLAS Collaboration. "Search for the Higgs boson decays $H \to ee$ and $H \to e\mu$ in *pp* collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". *Phys. Lett. B* 801 (2020), p. 135148. DOI: 10.1016/j.physletb.2019.135148. arXiv: 1909.10235 [hep-ex] (cit. on p. 29).

[70] *Convention for the establishment of a European organization for nuclear research: Paris, 1st July, 1953 : as amended. Convention pour l'établissement d'une Organisation européenne pour la Recherche nucléaire. Paris, le 1er juillet 1953 : telle qu'elle a été modifiée*. CERN, 1971. URL: https://cds.cern.ch/record/330625 (cit. on p. 31).

[71] Giuseppe Fidecaro. "The Discoveries of Rare Pion Decays at the CERN Synchrocyclotron". *Adv. Ser. Dir. High Energy Phys.* 23 (2015), pp. 397–414. DOI: 10.1142/9789814644150_0016. URL: https://cds.cern.ch/record/2103305 (cit. on p. 31).

[72] Thomas Massam et al. "Experimental observation of antideuteron production". *Nuovo Cimento* 39 (1965), pp. 10–14. DOI: 10.1007/BF02814251. URL: https://cds.cern.ch/record/345976 (cit. on p. 31).

[73] ALEPH DELPHI L3 OPAL SLD Collaborations et al. "Precision electroweak measurements on the Z resonance". *Phys. Rept.* 427 (2006), pp. 257–454. DOI: 10.1016/j.physrep.2005.12.006. arXiv: hep-ex/0509008 (cit. on p. 31).

[74] ALEPH DELPHI L3 OPAL Collaborations and LEP Electroweak Working Group. "Electroweak Measurements in Electron-Positron Collisions at W-Boson-Pair Energies at LEP". *Phys. Rept.* 532 (2013), pp. 119–244. DOI: 10.1016/j.physrep.2013.07.004. arXiv: 1302.3415 [hep-ex] (cit. on p. 31).

[75] "LHC Machine". *JINST* 3 (2008). Ed. by Lyndon Evans and Philip Bryant, S08001. DOI: 10.1088/1748-0221/3/08/S08001 (cit. on p. 32).

[76] CMS Collaboration. "The CMS Experiment at the CERN LHC". *JINST* 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004 (cit. on pp. 34, 41, 43).

[77] CMS Collaboration. "CMS: The Compact Muon Solenoid: Letter of intent for a general purpose detector at the LHC" (Oct. 1992) (cit. on p. 38).

[78] CMS Collaboration. "CMS, the magnet project: Technical design report" (May 1997) (cit. on p. 38).

[79] CMS Collaboration. "Precise Mapping of the Magnetic Field in the CMS Barrel Yoke using Cosmic Rays". *JINST* 5 (2010), T03021. DOI: 10.1088/1748-0221/5/03/T03021. arXiv: 0910.5530 [physics.ins-det] (cit. on p. 39).

[80] CMS Collaboration. "The CMS tracker system project: Technical Design Report" (1997) (cit. on p. 39).

[81] CMS Tracker Group. "The CMS Phase-1 Pixel Detector Upgrade". *JINST* 16.02 (2021), P02027. DOI: 10.1088/1748-0221/16/02/P02027. arXiv: 2012.14304 [physics.ins-det] (cit. on p. 40).

[82] CMS Collaboration. "The CMS electromagnetic calorimeter project: Technical Design Report" (1997) (cit. on p. 40).

[83] CMS Collaboration. "CMS Physics: Technical Design Report Volume 1: Detector Performance and Software" (2006) (cit. on p. 42).

[84] CMS Collaboration. "The CMS hadron calorimeter project: Technical Design Report" (1997) (cit. on p. 41).

[85] CMS HCAL Collaboration. "Design, performance, and calibration of the CMS Hadron-outer calorimeter". *Eur. Phys. J. C* 57 (2008), pp. 653–663. DOI: 10.1140/epjc/s10052-008-0756-6 (cit. on pp. 42, 43).

[86] CMS Collaboration. "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV". *JINST* 13.06 (2018), P06015. DOI: 10.1088/1748-0221/13/06/P06015. arXiv: 1804.04528 [physics.ins-det] (cit. on pp. 45, 50).

[87] S. Dasu et al. "CMS. The TriDAS project. Technical design report, vol. 1: The trigger systems" (Dec. 2000) (cit. on p. 44).

[88] "CMS: The TriDAS project. Technical design report, Vol. 2: Data acquisition and high-level trigger" (Dec. 2002). Ed. by P. Sphicas (cit. on p. 44).

[89] CMS Collaboration. "Particle-flow reconstruction and global event description with the CMS detector". *JINST* 12.10 (2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003. arXiv: 1706.04965 [physics.ins-det] (cit. on pp. 46, 51).

[90] W. Adam et al. "Track reconstruction in the CMS tracker" (Dec. 2005) (cit. on p. 47).

[91] CMS Collaboration. "Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC". *JINST* 16.05 (2021), P05014. DOI: 10.1088/1748-0221/16/05/P05014. arXiv: 2012.06888 [hep-ex] (cit. on pp. 48, 77, 80, 204).

[92] W Adam et al. "Reconstruction of electrons with the Gaussian-sum filter in the CMS tracker at the LHC". *Journal of Physics G: Nuclear and Particle Physics* 31.9 (July 2005), N9–N20. DOI: 10.1088/0954-3899/31/9/n01 (cit. on p. 48).

[93] John E. Huth et al. "Toward a standardization of jet definitions". *1990 DPF Summer Study on High-energy Physics: Research Directions for the Decade (Snowmass 90)*. Dec. 1990, pp. 0134–136 (cit. on p. 51).

[94] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The anti-k_t jet clustering algorithm". *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ph] (cit. on p. 51).

[95] CMS Collaboration. "Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS". *JINST* 6 (2011), P11002. DOI: 10.1088/1748-0221/6/11/P11002. arXiv: 1107.4277 [physics.ins-det] (cit. on p. 52).

[96] CMS Collaboration. "Jet energy scale and resolution measurement with Run 2 Legacy Data Collected by CMS at 13 TeV" (2021). URL: https://cds.cern.ch/record/2792322 (cit. on p. 52).

[97] Roberto Covarelli. "Vector-boson fusion and scattering measurements". *PoS* LHCP2021 (2021), p. 126. DOI: 10.22323/1.397.0126 (cit. on p. 53).

[98] Diogo Buarque Franzosi et al. "Vector boson scattering processes: Status and prospects". *Rev. Phys.* 8 (2022), p. 100071. DOI: 10.1016/j.revip.2022.100071. arXiv: 2106.01393 [hep-ph] (cit. on p. 53).

[99] Riccardo Bellan et al. "VBSCan Thessaloniki 2018 Workshop Summary". *2nd Vector Boson Scattering Coordination and Action Network Annual Meeting*. June 2019. arXiv: 1906.11332 [hep-ph] (cit. on p. 53).

[100] CMS and LHCb Collaboration. "Observation of the rare $B_s^0 \to \mu^+\mu^-$ decay from the combined analysis of CMS and LHCb data". *Nature* 522 (2015), pp. 68–72. DOI: 10.1038/nature14474. arXiv: 1411.4413 [hep-ex] (cit. on p. 53).

[101] "Technical Proposal for the Phase-II Upgrade of the CMS Detector" (June 2015). Ed. by D. Contardo et al. DOI: 10.17181/CERN.VU8I.D59J (cit. on pp. 54, 55).

[102] CMS Collaboration. "The Phase-2 Upgrade of the CMS Tracker" (June 2017). DOI: 10.17181/CERN.QZ28.FLHW (cit. on p. 54).

[103] CMS Collaboration. "The Phase-2 Upgrade of the CMS Level-1 Trigger" (2020) (cit. on p. 54).

[104] CMS Collaboraton. "The Phase-2 Upgrade of the CMS Data Acquisition and High Level Trigger" (2021). URL: https://cds.cern.ch/record/2759072 (cit. on p. 54).

[105] CMS Collaboration. "The Phase-2 Upgrade of the CMS Muon Detectors" (Sept. 2017) (cit. on p. 54).

[106] CMS Collaboration. "The Phase-2 Upgrade of the CMS Barrel Calorimeters" (2017). URL: https://cds.cern.ch/record/2283187 (cit. on p. 54).

[107] CMS Collaboration. "A MIP Timing Detector for the CMS Phase-2 Upgrade" (2019) (cit. on p. 55).

[108] Matteo Bonanomi. "Response of the High Granularity Calorimeter HGCAL and Characterisation of the Higgs Boson: With the CMS Experiment at the LHC". PhD thesis. DOI: 10.1007/978-3-031-26833-5 (cit. on pp. 55, 59, 188).

[109] CMS Collaboration. "The Phase-2 Upgrade of the CMS Endcap Calorimeter" (2017) (cit. on p. 56).

[110] S. Agostinelli et al. "GEANT4–a simulation toolkit". *Nucl. Instrum. Meth. A* 506 (2003), pp. 250–303. DOI: 10.1016/S0168-9002(03)01368-8 (cit. on p. 59).

[111] John Allison et al. "Geant4 developments and applications". *IEEE Trans. Nucl. Sci.* 53 (2006), p. 270. DOI: 10.1109/TNS.2006.869826 (cit. on p. 59).

[112] CMS Collaboration. "Mini-AOD: A New Analysis Data Format for CMS". *J. Phys. Conf. Ser.* 664.7 (2015), p. 7. DOI: 10.1088/1742-6596/664/7/072052. arXiv: 1702.04685 [physics.ins-det] (cit. on p. 61).

[113] CMS Collaboration. "A further reduction in CMS event data for analysis: the NANOAOD format". *EPJ Web Conf.* 214 (2019), p. 06021. DOI: 10.1051/epjconf/201921406021 (cit. on p. 61).

[114] CMS Collaboration. "Search for the Standard Model Higgs Boson in the Decay Channel H to ZZ to 4 Leptons in pp Collisions at $\sqrt{s} = 7$ TeV". *Phys. Rev. Lett.* 108 (2012), p. 111804. DOI: 10.1103/PhysRevLett.108.111804. arXiv: 1202.1997 [hep-ex] (cit. on p. 65).

[115] ATLAS Collaboration. "Combined search for the Standard Model Higgs boson using up to 4.9 fb^{-1} of pp collision data at $\sqrt{s} = 7$ TeV with the ATLAS detector at the LHC". *Phys. Lett. B* 710 (2012), pp. 49–66. DOI: 10.1016/j.physletb.2012.02.044. arXiv: 1202.1408 [hep-ex] (cit. on p. 65).

[116] CMS Collaboration. "Observation of a New Boson with Mass Near 125 GeV in pp Collisions at $\sqrt{s} = 7$ and 8 TeV". *JHEP* 06 (2013), p. 081. DOI: 10.1007/JHEP06(2013)081. arXiv: 1303.4571 [hep-ex] (cit. on p. 65).

[117] ATLAS Collaboration. "Measurements of Higgs boson production and couplings in the four-lepton channel in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector". *Phys. Rev. D* 91.1 (2015), p. 012006. DOI: 10.1103/PhysRevD.91.012006. arXiv: 1408.5191 [hep-ex] (cit. on pp. 65, 66).

[118] CMS Collaboration. "Measurement of the Properties of a Higgs Boson in the Four-Lepton Final State". *Phys. Rev. D* 89.9 (2014), p. 092007. DOI: 10.1103/PhysRevD.89.092007. arXiv: 1312.5353 [hep-ex] (cit. on p. 65).

[119] CMS Collaboration. "Study of the Mass and Spin-Parity of the Higgs Boson Candidate Via Its Decays to Z Boson Pairs". *Phys. Rev. Lett.* 110.8 (2013), p. 081803. DOI: 10.1103/PhysRevLett.110.081803. arXiv: 1212.6639 [hep-ex] (cit. on p. 65).

[120] CMS Collaboration. "Measurements of properties of the Higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s} = 13$ TeV". *JHEP* 11 (2017), p. 047. DOI: 10.1007/JHEP11(2017)047. arXiv: 1706.09936 [hep-ex] (cit. on pp. 65, 66).

[121] ATLAS Collaboration. "Measurement of the Higgs boson coupling properties in the $H \to ZZ^* \to 4\ell$ decay channel at $\sqrt{s} = 13$ TeV with the ATLAS detector". *JHEP* 03 (2018), p. 095. DOI: 10.1007/JHEP03(2018)095. arXiv: 1712.02304 [hep-ex] (cit. on p. 65).

[122] CMS Collaboration. "Constraints on the Higgs boson width from off-shell production and decay to Z-boson pairs". *Phys. Lett. B* 736 (2014), pp. 64–85. DOI: 10.1016/j.physletb.2014.06.077. arXiv: 1405.3455 [hep-ex] (cit. on p. 65).

[123] CMS Collaboration. "Limits on the Higgs boson lifetime and width from its decay to four charged leptons". *Phys. Rev. D* 92.7 (2015), p. 072010. DOI: 10.1103/PhysRevD.92.072010. arXiv: 1507.06656 [hep-ex] (cit. on p. 65).

[124] ATLAS Collaboration. "Constraints on the off-shell Higgs boson signal strength in the high-mass ZZ and WW final states with the ATLAS detector". *Eur. Phys. J. C* 75.7 (2015), p. 335. DOI: 10.1140/epjc/s10052-015-3542-2. arXiv: 1503.01060 [hep-ex] (cit. on p. 65).

[125] ATLAS Collaboration. "Constraints on off-shell Higgs boson production and the Higgs boson total width in $ZZ \to 4\ell$ and $ZZ \to 2\ell 2\nu$ final states with the ATLAS detector". *Phys. Lett. B* 786 (2018), pp. 223–244. DOI: 10.1016/j.physletb.2018.09.048. arXiv: 1808.01191 [hep-ex] (cit. on p. 65).

[126] CMS Collaboration. "Constraints on anomalous Higgs boson couplings using production and decay information in the four-lepton final state". *Phys. Lett. B* 775 (2017), pp. 1–24. DOI: 10.1016/j.physletb.2017.10.021. arXiv: 1707.00541 [hep-ex] (cit. on p. 65).

[127] CMS Collaboration. "Measurements of the Higgs boson width and anomalous HVV couplings from on-shell and off-shell production in the four-lepton final state". *Phys. Rev. D* 99.11 (2019), p. 112003. DOI: 10.1103/PhysRevD.99.112003. arXiv: 1901.00174 [hep-ex] (cit. on p. 65).

[128] CMS Collaboration. "Constraints on anomalous Higgs boson couplings to vector bosons and fermions in its production and decay using the four-lepton final state". *Phys. Rev. D* 104.5 (2021), p. 052004. DOI: 10.1103/PhysRevD.104.052004. arXiv: 2104.12152 [hep-ex] (cit. on pp. 65, 109, 132, 172).

[129] CMS Collaboration. "Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV". *Eur. Phys. J. C* 75.5 (2015), p. 212. DOI: 10.1140/epjc/s10052-015-3351-7. arXiv: 1412.8662 [hep-ex] (cit. on p. 66).

[130] CMS Collaboration. "Measurements of production cross sections of the Higgs boson in the four-lepton final state in proton–proton collisions at $\sqrt{s} = 13\,\text{TeV}$". *Eur. Phys. J. C* 81.6 (2021), p. 488. DOI: 10.1140/epjc/s10052-021-09200-x. arXiv: 2103.04956 [hep-ex] (cit. on pp. 66, 125).

[131] ATLAS Collaboration. "Higgs boson production cross-section measurements and their EFT interpretation in the 4ℓ decay channel at $\sqrt{s}$ =13 TeV with the ATLAS detector". *Eur. Phys. J. C* 80.10 (2020). [Erratum: Eur.Phys.J.C 81, 29 (2021), Erratum: Eur.Phys.J.C 81, 398 (2021)], p. 957. DOI: 10.1140/epjc/s10052-020-8227-9. arXiv: 2004.03447 [hep-ex] (cit. on p. 66).

[132] ATLAS Collaboration. "A detailed map of Higgs boson interactions by the ATLAS experiment ten years after the discovery". *Nature* 607.7917 (2022). [Erratum: Nature 612, E24 (2022)], pp. 52–59. DOI: 10.1038/s41586-022-04893-w. arXiv: 2207.00092 [hep-ex] (cit. on p. 66).

[133] ATLAS Collaboration. "Fiducial and differential cross sections of Higgs boson production measured in the four-lepton decay channel in *pp* collisions at $\sqrt{s}$=8 TeV with the ATLAS detector". *Phys. Lett. B* 738 (2014), pp. 234–253. DOI: 10.1016/j.physletb.2014.09.054. arXiv: 1408.3226 [hep-ex] (cit. on p. 66).

[134] CMS Collaboration. "Measurement of differential and integrated fiducial cross sections for Higgs boson production in the four-lepton decay channel in pp collisions at $\sqrt{s} = 7$ and 8 TeV". *JHEP* 04 (2016), p. 005. DOI: 10.1007/JHEP04(2016)005. arXiv: 1512.08377 [hep-ex] (cit. on pp. 66, 125, 128, 172).

[135] ATLAS Collaboration. "Measurement of the total and differential Higgs boson production cross-sections at $\sqrt{s} = 13$ TeV with the ATLAS detector by combining the H $\rightarrow$ ZZ*$\rightarrow$ 4ℓ and H $\rightarrow \gamma\gamma$ decay channels". *JHEP* 05 (2023), p. 028. DOI: 10.1007/JHEP05(2023)028. arXiv: 2207.08615 [hep-ex] (cit. on p. 66).

[136] ATLAS Collaboration. "Measurement of inclusive and differential cross sections in the $H \rightarrow ZZ^* \rightarrow 4\ell$ decay channel in *pp* collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". *JHEP* 10 (2017), p. 132. DOI: 10.1007/JHEP10(2017)132. arXiv: 1708.02810 [hep-ex] (cit. on p. 66).

[137] CMS Collaboration. "Measurement and interpretation of differential cross sections for Higgs boson production at $\sqrt{s} = 13$ TeV". *Phys. Lett. B* 792 (2019), pp. 369–396. DOI: 10.1016/j.physletb.2019.03.059. arXiv: 1812.06504 [hep-ex] (cit. on pp. 66, 128).

[138] ATLAS Collaboration. "Combined measurement of differential and total cross sections in the $H \rightarrow \gamma\gamma$ and the $H \rightarrow ZZ^* \rightarrow 4\ell$ decay channels at $\sqrt{s} = 13$ TeV with the ATLAS detector". *Phys. Lett. B* 786 (2018), pp. 114–133. DOI: 10.1016/j.physletb.2018.09.019. arXiv: 1805.10197 [hep-ex] (cit. on p. 66).

[139] ATLAS Collaboration. "Measurements of the Higgs boson inclusive and differential fiducial cross sections in the 4ℓ decay channel at $\sqrt{s} = 13$ TeV". *Eur. Phys. J. C* 80.10 (2020), p. 942. DOI: 10.1140/epjc/s10052-020-8223-0. arXiv: 2004.03969 [hep-ex] (cit. on p. 66).

[140] Simone Alioli et al. "NLO vector-boson production matched with shower in POWHEG". *JHEP* 07 (2008), p. 060. DOI: 10.1088/1126-6708/2008/07/060. arXiv: 0805.4802 [hep-ph] (cit. on p. 67).

[141] Paolo Nason. "A New method for combining NLO QCD with shower Monte Carlo algorithms". *JHEP* 11 (2004), p. 040. DOI: 10.1088/1126-6708/2004/11/040. arXiv: hep-ph/0409146 (cit. on p. 67).

[142] Stefano Frixione, Paolo Nason, and Carlo Oleari. "Matching NLO QCD computations with Parton Shower simulations: the POWHEG method". *JHEP* 11 (2007), p. 070. DOI: 10.1088/1126-6708/2007/11/070. arXiv: 0709.2092 [hep-ph] (cit. on p. 67).

[143] E. Bagnaschi et al. "Higgs production via gluon fusion in the POWHEG approach in the SM and in the MSSM". *JHEP* 02 (2012), p. 088. DOI: 10.1007/JHEP02(2012)088. arXiv: 1111.2854 [hep-ph] (cit. on p. 67).

[144] Paolo Nason and Carlo Oleari. "NLO Higgs boson production via vector-boson fusion matched with shower in POWHEG". *JHEP* 02 (2010), p. 037. DOI: 10.1007/JHEP02(2010)037. arXiv: 0911.5299 [hep-ph] (cit. on p. 67).

[145] Heribertus B. Hartanto et al. "Higgs boson production in association with top quarks in the POWHEG BOX". *Phys. Rev. D* 91.9 (2015), p. 094003. DOI: 10.1103/PhysRevD.91.094003. arXiv: 1501.04498 [hep-ph] (cit. on p. 67).

[146] Gionata Luisoni et al. "$HW^{\pm}$/HZ + 0 and 1 jet at NLO with the POWHEG BOX interfaced to GoSam and their merging within MiNLO". *JHEP* 10 (2013), p. 083. DOI: 10.1007/JHEP10(2013)083. arXiv: 1306.2542 [hep-ph] (cit. on p. 67).

[147] Keith Hamilton et al. "NNLOPS simulation of Higgs boson production". *JHEP* 10 (2013), p. 222. DOI: 10.1007/JHEP10(2013)222. arXiv: 1309.0017 [hep-ph] (cit. on p. 67).

[148] Yanyan Gao et al. "Spin determination of single-produced resonances at hadron colliders". *Phys. Rev. D* 81 (2010), p. 075022. DOI: 10.1103/PhysRevD.81.079905. arXiv: 1001.3396 [hep-ph] (cit. on pp. 67, 107, 108).

[149] Sara Bolognesi et al. "On the spin and parity of a single-produced resonance at the LHC". *Phys. Rev. D* 86 (2012), p. 095031. DOI: 10.1103/PhysRevD.86.095031. arXiv: 1208.4018 [hep-ph] (cit. on pp. 67, 108).

[150] Ian Anderson et al. "Constraining anomalous HVV interactions at proton and lepton colliders". *Phys. Rev. D* 89 (2014), p. 035007. DOI: 10.1103/PhysRevD.89.035007. arXiv: 1309.4819 [hep-ph] (cit. on pp. 67, 108).

[151] Andrei V. Gritsan et al. "Constraining anomalous Higgs boson couplings to the heavy flavor fermions using matrix element techniques". *Phys. Rev. D* 94 (2016), p. 055023. DOI: 10.1103/PhysRevD.94.055023. arXiv: 1606.03107 [hep-ph] (cit. on pp. 67, 108).

[152] Andrei V. Gritsan et al. "New features in the JHU generator framework: constraining Higgs boson properties from on-shell and off-shell production". *Phys. Rev. D* 102 (2020), p. 056022. DOI: 10.1103/PhysRevD.102.056022. arXiv: 2002.09888 [hep-ph] (cit. on pp. 67, 108).

[153] Tom Melia et al. "W+W-, WZ and ZZ production in the POWHEG BOX". *JHEP* 11 (2011), p. 078. DOI: 10.1007/JHEP11(2011)078. arXiv: 1107.5051 [hep-ph] (cit. on p. 67).

[154] John M. Campbell and R. K. Ellis. "MCFM for the Tevatron and the LHC". *Nucl. Phys. B Proc. Suppl.* 205-206 (2010). Ed. by Johannes Blümlein, Sven-Olaf Moch, and Tord Riemann, pp. 10–15. DOI: 10.1016/j.nuclphysbps.2010.08.011. arXiv: 1007.3492 [hep-ph] (cit. on p. 67).

[155] 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. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph] (cit. on p. 67).

[156] Torbjörn Sjöstrand et al. "An introduction to PYTHIA 8.2". *Comput. Phys. Commun.* 191 (2015), pp. 159–177. DOI: 10.1016/j.cpc.2015.01.024. arXiv: 1410.3012 [hep-ph] (cit. on p. 67).

[157] CMS Collaboration. "Event generator tunes obtained from underlying event and multiparton scattering measurements". *Eur. Phys. J. C* 76.3 (2016), p. 155. DOI: 10.1140/epjc/s10052-016-3988-x. arXiv: 1512.00815 [hep-ex] (cit. on p. 67).

[158] CMS Collaboration. "Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements". *Eur. Phys. J. C* 80.1 (2020), p. 4. DOI: 10.1140/epjc/s10052-019-7499-4. arXiv: 1903.12179 [hep-ex] (cit. on p. 67).

[159] NNPDF Collaboration. "Parton distributions for the LHC Run II". *JHEP* 04 (2015), p. 040. DOI: 10.1007/JHEP04(2015)040. arXiv: 1410.8849 [hep-ph] (cit. on p. 67).

[160] GEANT4 Collaboration. "GEANT4–a simulation toolkit". *Nucl. Instrum. Meth. A* 506 (2003), pp. 250–303. DOI: 10.1016/S0168-9002(03)01368-8 (cit. on p. 67).

[161] John Allison et al. "Geant4 developments and applications". *IEEE Trans. Nucl. Sci.* 53 (2006), p. 270. DOI: 10.1109/TNS.2006.869826 (cit. on p. 67).

[162] Tianqi Chen and Carlos Guestrin. "XGBoost". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Aug. 2016. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145%2F2939672.2939785 (cit. on p. 72).

[163] A. Bodek et al. "Extracting Muon Momentum Scale Corrections for Hadron Collider Experiments". *Eur. Phys. J. C* 72 (2012), p. 2194. DOI: 10.1140/epjc/s10052-012-2194-8. arXiv: 1208.3710 [hep-ex] (cit. on p. 77).

[164] ATLAS Collaboration. "Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton-proton collision data at $\sqrt{s} =$ 13 TeV". *Eur. Phys. J. C* 79.8 (2019), p. 639. DOI: 10.1140/epjc/s10052-019-7140-6. arXiv: 1902.04655 [physics.ins-det] (cit. on p. 81).

[165] CMS Collaboration. "Description and performance of track and primary-vertex reconstruction with the CMS tracker". *JINST* 9.10 (2014), P10009. DOI: 10.1088/1748-0221/9/10/P10009. arXiv: 1405.6569 [physics.ins-det] (cit. on p. 82).

[166] CMS Collaboration. "Jet algorithms performance in 13 TeV data" (2017) (cit. on p. 85).

[167] Massimiliano Grazzini, Stefan Kallweit, and Dirk Rathlev. "ZZ production at the LHC: Fiducial cross sections and distributions in NNLO QCD". *Physics Letters B* 750 (2015), pp. 407–410. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2015.09.055 (cit. on p. 93).

[168] Anastasiya Bierweiler, Tobias Kasprzik, and Johann H. Kühn. "Vector-boson pair production at the LHC to $\mathcal{O}(\alpha^3)$ accuracy". *JHEP* 12 (2013), p. 071. DOI: 10.1007/JHEP12(2013)071. arXiv: 1305.5402 [hep-ph] (cit. on p. 94).

[169] Marco Bonvini et al. "Signal-background interference effects in $gg \to H \to WW$ beyond leading order". *Phys. Rev. D* 88 (2013), p. 034032. DOI: 10.1103/PhysRevD.88.034032. arXiv: 1304.3053 [hep-ph] (cit. on p. 94).

[170] Kirill Melnikov and Matthew Dowling. "Production of two Z-bosons in gluon fusion in the heavy top quark approximation". *Phys. Lett. B* 744 (2015), p. 43. DOI: 10.1016/j.physletb.2015.03.030. arXiv: 1503.01274 [hep-ph] (cit. on p. 94).

[171] Chong Sheng Li et al. "Soft gluon resummation in the signal-background interference process of gg($\to$ h*) $\to$ ZZ". *JHEP* 08 (2015), p. 065. DOI: 10.1007/JHEP08(2015)065. arXiv: 1504.02388 [hep-ph] (cit. on p. 94).

[172] Stefano Catani and Massimiliano Grazzini. "An NNLO subtraction formalism in hadron collisions and its application to Higgs boson production at the LHC". *Phys. Rev. Lett.* 98 (2007), p. 222002. DOI: 10.1103/PhysRevLett.98.222002. arXiv: hep-ph/0703012 [hep-ph] (cit. on p. 94).

[173] Massimiliano Grazzini. "NNLO predictions for the Higgs boson signal in the H → WW → $\ell\nu\ell\nu$ and H→ ZZ → 4ℓ decay channels". *JHEP* 02 (2008), p. 043. DOI: 10.1088/1126-6708/2008/02/043. arXiv: 0801.3232 [hep-ph] (cit. on p. 94).

[174] Massimiliano Grazzini and Hayk Sargsyan. "Heavy-quark mass effects in Higgs boson production at the LHC". *JHEP* 09 (2013), p. 129. DOI: 10.1007/JHEP09(2013)129. arXiv: 1306.4581 [hep-ph] (cit. on pp. 94, 128).

[175] Giampiero Passarino. "Higgs CAT". *Eur. Phys. J. C* 74 (2014), p. 2866. DOI: 10.1140/epjc/s10052-014-2866-7. arXiv: 1312.2397 [hep-ph] (cit. on p. 94).

[176] LHC Higgs Combination Group. "Procedure for the LHC Higgs boson search combination in Summer 2011" (Aug. 2011) (cit. on p. 102).

[177] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". *Eur. Phys. J. C* 71 (2011). [Erratum: Eur.Phys.J.C 73, 2501 (2013)], p. 1554. DOI: 10.1140/epjc/s10052-011-1554-0. arXiv: 1007.1727 [physics.data-an] (cit. on pp. 102, 125).

[178] S. S. Wilks. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses". *Annals Math. Statist.* 9.1 (1938), pp. 60–62. DOI: 10.1214/aoms/1177732360 (cit. on p. 102).

[179] CMS Collaboration. "Measurement of differential and integrated fiducial cross sections for Higgs boson production in the four-lepton decay channel in pp collisions at $\sqrt{s} = 7$ and 8 TeV". *JHEP* 04 (2016), p. 005. DOI: 10.1007/JHEP04(2016)005. arXiv: 1512.08377 [hep-ex] (cit. on pp. 103, 111).

[180] Fady Bishara et al. "Constraining Light-Quark Yukawa Couplings from Higgs Distributions". *Phys. Rev. Lett.* 118.12 (2017), p. 121801. DOI: 10.1103/PhysRevLett.118.121801. arXiv: 1606.09253 [hep-ph] (cit. on pp. 104, 156).

[181] Yotam Soreq, Hua Xing Zhu, and Jure Zupan. "Light quark Yukawa couplings from Higgs kinematics". *JHEP* 12 (2016), p. 045. DOI: 10.1007/JHEP12(2016)045. arXiv: 1606.09621 [hep-ph] (cit. on p. 104).

[182] A. Buckley et al. "A comparative study of Higgs boson production from vector-boson fusion". *JHEP* 11 (2021), p. 108. DOI: 10.1007/JHEP11(2021)108. arXiv: 2105.11399 [hep-ph] (cit. on p. 105).

[183] Shireen Gangal, Maximilian Stahlhofen, and Frank J. Tackmann. "Rapidity-Dependent Jet Vetoes". *Phys. Rev. D* 91.5 (2015), p. 054023. DOI: 10.1103/PhysRevD.91.054023. arXiv: 1412.4792 [hep-ph] (cit. on pp. 105, 106).

[184] Jerzy Neyman and Egon Sharpe Pearson. "On the Problem of the Most Efficient Tests of Statistical Hypotheses". *Phil. Trans. Roy. Soc. Lond. A* 231.694-706 (1933), pp. 289–337. DOI: 10.1098/rsta.1933.0009 (cit. on p. 108).

[185] G. Cowan. "Discovery sensitivity for a counting experiment with background uncertainty". *Tech. rep., Royal Holloway, London* (2012). URL: http://www.pp.rhul.ac.uk/~cowan/stat/%20medsig/medsigNote.pdf (cit. on p. 112).

[186] CMS Collaboration. "A measurement of the Higgs boson mass in the diphoton decay channel". *Phys. Lett. B* 805 (2020), p. 135425. DOI: 10.1016/j.physletb.2020.135425. arXiv: 2002.06398 [hep-ex] (cit. on p. 117).

[187] A. Bredenstein et al. "Precise predictions for the Higgs-boson decay H —> WW/ZZ —> 4 leptons". *Phys. Rev. D* 74 (2006), p. 013004. DOI: 10.1103/PhysRevD.74.013004. arXiv: hep-ph/0604011 (cit. on p. 120).

[188] CMS Collaboration. "Precision luminosity measurement in proton-proton collisions at $\sqrt{s} = 13$ TeV in 2015 and 2016 at CMS". *Eur. Phys. J. C* 81.9 (2021), p. 800. DOI: 10.1140/epjc/s10052-021-09538-2. arXiv: 2104.01927 [hep-ex] (cit. on p. 122).

[189] CMS Collaboration. "CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV" (2018) (cit. on p. 122).

[190] CMS Collaboration. "CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV" (2019) (cit. on p. 122).

[191] Jon Butterworth et al. "PDF4LHC recommendations for LHC Run II". *J. Phys. G* 43 (2016), p. 023001. DOI: 10.1088/0954-3899/43/2/023001. arXiv: 1510.03865 [hep-ph] (cit. on p. 124).

[192] D. de Florian et al. "Higgs boson production at the LHC: transverse momentum resummation effects in the $H \to \gamma\gamma$, $H \to WW \to \ell\nu\ell\nu$ and $H \to ZZ \to 4\ell$ decay modes". *JHEP* 06 (2012), p. 132. DOI: 10.1007/JHEP06(2012)132. arXiv: 1203.6321 [hep-ph] (cit. on p. 128).

[193] Marcela Carena et al. "Alignment limit of the NMSSM Higgs sector". *Phys. Rev. D* 93 (3 Feb. 2016), p. 035013. DOI: 10.1103/PhysRevD.93.035013. URL: https://link.aps.org/doi/10.1103/PhysRevD.93.035013 (cit. on p. 152).

[194] CMS Collaboration. "Search for nonresonant Higgs boson pair production in the four leptons plus twob jets final state in proton-proton collisions at $\sqrt{s} = 13$ TeV". *JHEP* 06 (2023), p. 130. DOI: 10.1007/JHEP06(2023)130. arXiv: 2206.10657 [hep-ex] (cit. on p. 155).

[195] CMS Collaboration. "Search for Higgs boson pairs decaying to WW*WW*, WW*$\tau\tau$, and $\tau\tau\tau\tau$ in proton-proton collisions at $\sqrt{s} = 13$ TeV". *JHEP* 07 (2023), p. 095. DOI: 10.1007/JHEP07(2023)095. arXiv: 2206.10268 [hep-ex] (cit. on p. 155).

[196] CMS Collaboration. "Search for Nonresonant Pair Production of Highly Energetic Higgs Bosons Decaying to Bottom Quarks". *Phys. Rev. Lett.* 131.4 (2023), p. 041803. DOI: 10.1103/PhysRevLett.131.041803. arXiv: 2205.06667 [hep-ex] (cit. on p. 155).

[197] CMS Collaboration. "Search for Higgs Boson Pair Production in the Four b Quark Final State in Proton-Proton Collisions at s=13 TeV". *Phys. Rev. Lett.* 129.8 (2022), p. 081802. DOI: 10.1103/PhysRevLett.129.081802. arXiv: 2202.09617 [hep-ex] (cit. on p. 155).

[198] CMS Collaboration. "Search for nonresonant Higgs boson pair production in final state with two bottom quarks and two tau leptons in proton-proton collisions at s=13 TeV". *Phys. Lett. B* 842 (2023), p. 137531. DOI: 10.1016/j.physletb.2022.137531. arXiv: 2206.09401 [hep-ex] (cit. on p. 155).

[199] CMS Collaboration. "Search for nonresonant Higgs boson pair production in final states with two bottom quarks and two photons in proton-proton collisions at $\sqrt{s}$ = 13 TeV". *JHEP* 03 (2021), p. 257. DOI: 10.1007/JHEP03(2021)257. arXiv: 2011.12373 [hep-ex] (cit. on p. 155).

[200] CMS collaboration. *Search for Higgs boson decay to a charm quark-antiquark pair in proton-proton collisions at $\sqrt{s} = 13$ TeV*. Tech. rep. Geneva: CERN, 2022. arXiv: 2205.05550. URL: https://cds.cern.ch/record/2809290 (cit. on p. 158).

[201] DELPHI collaboration. "Measurement of the B^0 - $\bar{B}^0$ mixing parameter in DELPHI". *Phys. Lett. B* 332 (1994), pp. 488–500. DOI: 10.1016/0370-2693(94)91285-8 (cit. on p. 170).

[202] DELPHI collaboration. "Measurement of the B(d)0 oscillation frequency using kaons, leptons and jet charge". *Z. Phys. C* 72 (1996), pp. 17–30. DOI: 10.1007/s002880050219 (cit. on p. 170).

[203] Marco Rovere et al. "CLUE: A Fast Parallel Clustering Algorithm for High Granularity Calorimeters in High-Energy Physics". *Front. Big Data* 3 (2020), p. 591315. DOI: 10.3389/fdata.2020.591315. arXiv: 2001.09761 [physics.ins-det] (cit. on pp. 181, 182).

[204] E. Brondolin. "CLUE: a clustering algorithm for current and future experiments". *J. Phys. Conf. Ser.* 2438.1 (2023), p. 012074. DOI: 10.1088/1742-6596/2438/1/012074 (cit. on p. 182).

[205] CMS collaboration. "The TICL (v4) reconstruction at the CMS Phase-2 High Granularity Calorimeter Endcap" (2022). URL: https://cds.cern.ch/record/2839740 (cit. on p. 185).

[206] CMS HGCAL Collaboration. "Response of a CMS HGCAL silicon-pad electromagnetic calorimeter prototype to 20–300 GeV positrons". *JINST* 17.05 (2022), P05022. DOI: 10.1088/1748-0221/17/05/P05022. arXiv: 2111.06855 [physics.ins-det] (cit. on p. 191).

[207] L. Delchambre. "Weighted principal component analysis: a weighted covariance eigendecomposition approach". *Monthly Notices of the Royal Astronomical Society* 446.4 (Dec. 2014), pp. 3545–3555. DOI: 10.1093/mnras/stu2219. URL: https://doi.org/10.1093%2Fmnras%2Fstu2219 (cit. on p. 195).

[208] CMS Collaboration. "Electron reconstruction and identification at sqrt(s) = 7 TeV" (2010) (cit. on p. 204).

[209] CMS Collaboration. "Performance of Electron Reconstruction and Selection with the CMS Detector in Proton-Proton Collisions at $\sqrt{s} = 8$ TeV". *JINST* 10.06 (2015), P06005. DOI: 10.1088/1748-0221/10/06/P06005. arXiv: 1502.02701 [physics.ins-det] (cit. on p. 204).

[210] Lei Huang et al. "Normalization techniques in training dnns: Methodology, analysis and application". *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023) (cit. on p. 216).

[211] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG] (cit. on p. 218).

[212] *In it for the long haul*. URL: https://cerncourier.com/a/in-it-for-the-long-haul/ (cit. on p. 228).

ECOLE
DOCTORALE

Titre : Mesure des sections efficaces fiduciaires du boson de Higgs avec le détecteur CMS et reconstruction électromagnétique avec le nouveau calorimètre à haute granularité.

Mots clés : LHC, CMS, Physique des Hautes Energies, Higgs boson

Résumé : Cette thèse présente la mesure des sections efficaces du boson de Higgs dans le canal de désintégration $H \to ZZ \to 4\ell$ en utilisant des collisions de protons-protons à une énergie de 13 TeV dans le centre de masse et collectées avec l'expérience CMS lors du Run 2 du LHC au CERN. Ce canal est considéré le canal optimal de la physique du Higgs en raison de son pic clairement visible au-dessus d'un bruit de fond presque plat, de l'excellent rapport signal sur bruit, et d'un état final entièrement reconstructible qui bénéficie en outre de la reconstruction très performante des leptons du détecteur CMS. Les sections efficaces sont mesurées dans des bins de 32 observables simples et doublement différentielles, fournissant des informations sur les différents aspects de la physique du Higgs. La première contrainte issue des données du fond non résonant $ZZ \to 4\ell$ est également incluse. L'ensemble des résultats est complété par la contrainte de l'auto-couplage trilinéaire du boson de Higgs et des couplages aux quarks bottom et charm. Tous les résultats sont cohérents avec les prédictions théoriques du modèle standard de la physique des particules. En regardant vers l'avenir, alors que le Run 2 est terminée et que le Run 3 est en cours, cette thèse introduit une nouvelle méthode pour estimer bruit de fond réductible du canal à quatre leptons qui sera utilisée dans les analyses avec les nouvelles données. La stratégie proposée explore la possibilité de modéliser cette source en utilisant une approche novatrice et de réduire l'incertitude systématique considérable typique des méthodes actuelles, ce qui sera également un facteur limitant lors de la prochaine phase du LHC. Le LHC à haute luminosité (HL-LHC) vise à augmenter la luminosité intégrée d'un facteur de 10 par rapport à la valeur nominale du LHC, ouvrant de nouvelles perspectives pour les découvertes et la physique de précision. Afin de faire face à le haute niveau de pileup et de supporter une dose de radiation élevée, l'expérience CMS prévoit l'installation d'un nouveau calorimètre à haute granularité (HGCAL) dans les bouchons, qui offrira la possibilité de réaliser une calorimétrie avec une granularité similaire à celle du trajectographe. Cette thèse contribue au développement et à la réévaluation de la reconstruction hors ligne électromagnétique pour le HGCAL. La première contribution concerne le nettoyage des gerbes électromagnétiques des contaminations parasites résultant de l'environnement à grande pileup. La deuxième contribution concerne la reconstruction des électrons. Un électron peut commencer à produire une gerbe en traversant le trajectographe avant d'atteindre le calorimètre, répartissant ainsi l'énergie de l'électron en plusieurs amas. Ces contributions doivent être combinées pour reconstruire l'électron tel que produit lors de la collision dure. Actuellement, un algorithme purement géométrique effectue cette procédure, et sa performance est évaluée pour la première fois dans le HGCAL. De plus, cette thèse propose un nouvel algorithme dédié basé sur des Deep Neural Network spécialement conçus pour le nouveau calorimètre.

Title : Measurement of Higgs boson fiducial cross sections with the CMS detector and electromagnetic reconstruction with the high-granularity endcap calorimeter

Keywords : LHC, CMS, High energy physics, Higgs boson

Abstract : This thesis presents the measurement of the Higgs boson cross sections in the $H \to ZZ \to 4\ell$ decay channel using proton-proton collisions at a centre-of-mass energy of 13 TeV recorded with the CMS experiment during the Run 2 of the CERN LHC. This channel is regarded as the golden channel of Higgs physics due to its clear peak over an almost flat background, the large signal-to-background ratio, and a fully reconstructible final state that also benefits from the highly performant lepton reconstruction of the CMS detector. Cross sections are measured inclusively and in bins of 32 single- and double-differential observables, providing insights into the different aspects of Higgs physics. The first constraint from data of the non-resonant $ZZ \to 4\ell$ background is also included. The set of results is completed by the constraint of the trilinear self-coupling of the Higgs boson and the couplings to bottom and charm quarks. All results are consistent with the theoretical predictions of the standard model of particle physics. Looking forward, as the Run 2 is over and Run 3 is underway, this thesis introduces a novel method for estimating the reducible background of the four-lepton channel that will be used in the analyses with the new dataset. The proposed strategy explores the possibility of modelling this source using a novel approach and reducing the sizeable systematic uncertainty typical of the current methods, which will also be a limiting factor during the next phase of the LHC. The High-Luminosity LHC (HL-LHC) aims to increase the integrated luminosity by a factor of 10 beyond the LHC's design value, opening new horizons for discoveries and precision physics. In order to cope with the large pileup and sustain the high radiation dose, the CMS experiment foresees the installation of a new High-Granularity endcap CALorimeter (HGCAL), which will offer the possibility of performing calorimetry with tracker-like granularity. This thesis contributes to developing and reassessing the electromagnetic offline reconstruction for the HGCAL. The first contribution regards cleaning electromagnetic showers from spurious contaminations resulting from the high-pileup environment. The second contribution pertains to electron reconstruction. An electron can start showering while traversing the tracker before reaching the calorimeter, spreading the electron energy in several clusters. These contributions should be combined to reconstruct the original electron. A purely geometrical algorithm currently performs this procedure, and its performance is assessed in the HGCAL for the first time. Additionally, this thesis proposes a new, dedicated algorithm based on Deep Neural Networks explicitly tailored for the new calorimeter.

Institut Polytechnique de Paris
91120 Palaiseau, France