Recherche de la diffusion de boson vecteur dans le canal semi-leptonique avec le détecteur CMS et études sur la classification de gerbes électromagnétiques avec HGCAL

Alexandre Hakimi

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAX132

Thèse de doctorat

ÉCOLE POLYTECHNIQUE

IP PARIS

Search for Vector Boson Scattering in semileptonic decay at CMS and studies on HGCal trigger electromagnetic shower classification

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à 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 14/12/2022, par

Hakimi Alexandre

Composition du Jury :

Emmanuel Monnier Directeur de recherche, Centre de Physique des Particules de Marseille	Président
Lucia Di Ciaccio Professeure, Université de Savoie Mont-Blanc)	Rapportrice
David Rousseau Directeur de recherche, Laboratoire de physique des deux infinis Irène Joliot-Curie	Rapporteur
Pietro Govoni Associate professor, Università degli studi di Milano Bicocca	Examinateur
Jean Baptiste Sauvan Chargé de recherche, Ecole polytechnique (Laboratoire Leprince Ringuet)	Directeur de thèse

Résumé

Cette thèse présente la recherche du processus de diffusion de bosons vecteur (VBS) dans le canal ZV semi-leptonique par l'exploitation des données de collisions proton-proton à une énergie du centre de masse de 13 TeV récoltées par le détecteur CMS au LHC du CERN entre 2016 et 2018. Ce processus est particulièrement intéressant à étudier car il est intimement lié au mécanisme de brisure spontanée de la symétrie électrofaible. Il offre également un accès direct aux couplages quadratiques entre bosons vecteurs. D'éventuels écarts aux prédictions du modèle standard (SM) pourraient indiquer la présence de nouvelle physique.

Les phénomènes de VBS sont très rares et particulièrement complexes à étudier, même à l'aide du très grand nombre d'événements accumulés au LHC. Cependant, des stratégies d'extraction du signal sophistiquées ont permis l'observation des cannaux de désintégration leptoniques, plus facilement reconstruits dans les détecteurs, mais aussi de désintégrations semi-leptoniques, avec le canal WV où un boson W se désintègre en deux leptons et un boson Z ou W se désintègre hadroniquement. Le canal semi-leptonique ZV étudié dans ce travail, est plus ardu à reconstruire que le canal leptonique de par la présence d'une désintégration hadronique, mais possède en contrepartie un rapport d'embranchement plus important et fait appel au couplage quadratique entre bosons vecteurs. Il offre également une bonne sensibilité aux eventuelles modifications du SM par de la nouvelle physique telle que formalisée sous la forme de théorie effective des champs.

Le travail d'analyse décrit dans ce document comprend la définition d'un espace des phases enrichi en signal et minimisant le bruit de fond, en exploitant la signature caractéristique des événements VBS au LHC. Les quarks initiaux provoquent la présence d'une paire de jets de particules produits avec une masse invariante typiquement large et une séparation en pseudorapidité élevée. Pour maximiser la sensibilité à ce processus rare, différentes catégories sont établies en fonction de la topologie des jets issus de la désintégration hadronique du boson Z ou W, ainsi que de la présence ou non de jets issus de quarks b dans l'événement.

L'extraction du signal requiert une bonne connaissance des bruits de fond pouvant reproduire le même état final, ainsi que leur modélisation précise. Des corrections sont appliquées pour tenir compte des divergences entre les simulations et les données observées, et une procédure de correction des distributions basée sur les données est mise en place.

Les variables les plus discriminantes pour isoler le signal VBS, notamment la masse invariante de la paire de jets VBS et sa séparation en pseudorapidité, sont combinées avec un modèle d'apprentissage automatique afin de rejeter les sources de bruits de fond. Après une comparaison avec des algorithmes de type *arbres de décisions boostés* (BDT), l'utilisation de *réseaux de neurones artificiels* a été choisie et leur architecture optimisée pour maximiser les performances. Le jeu de variables d'entrées a également été réduit afin de ne conserver que les variables les plus pertinentes. Une approche statistique utilisant un *maximum likelihood fit*, réalisée simultanément dans la région enrichie en signal et dans des régions enrichies en bruits de fond, est employée pour la mesure du signal VBS, en tenant compte des différentes sources d'incertitudes. La puissance statistique attendue est estimée à 1.8 sigmas

Les données récoltées à ce jour par CMS ne sont donc pas suffisantes pour confirmer l'observation de phénomènes rares comme certains des canaux VBS. C'est une des raisons pour lesquelles le CERN prévoit, pour la fin de la décennie, une phase à plus haute luminosité (HL-LHC), fournissant en une dizaine d'années un jeu de données dix fois plus important que celui accumulé à ce jour. Cette amélioration permettra une sensibilité accrue aux phénomènes de physique rares, ainsi qu'une meilleure précision des mesures réalisées, au prix d'une aug-

mentation des radiations reçues par les détecteurs et d'un taux d'empilement plus élevé. Une mise à niveau des équipements est donc nécessaire pour faire face à ces nouveaux défis. La collaboration CMS prévoit notamment de remplacer les calorimètres des bouchons par un calorimètre hautement granulaire (HGCAL). Basé sur une technologie de détecteurs au silicium connus pour leur résistance aux radiations et de scintillateurs plastiques, il offrira la plus fine segmentation jamais atteinte pour un calorimètre, permettant ainsi de rejeter efficacement les interactions dues à l'empilement. Le système de déclenchement de niveau 1 (L1T) de CMS sera lui aussi mis à jour, avec des latences plus élevées, afin d'améliorer la sensibilité à des phénomènes de physique rare. Il est prévu que la nouvelle configuration du L1T ait accès aux informations du HGCAL, sous la forme de primitives de déclenchement, afin de caractériser et identifier les *clusters* d'énergies déposés dans le calorimètre.

Une partie du travail de cette thèse porte sur l'optimisation de ces primitives de déclenchement pour l'identification de gerbes électromagnétiques. Les contraintes dues aux architectures du détecteur et du système de déclenchement sont prises en compte dans une procédure d'optimisation multi-objectifs. En particulier les algorithmes d'apprentissage machine ne doivent pas monopoliser une trop grande part des ressources des cartes de type FPGA sur lesquelles ils sont implémentés. De plus, la taille des données communiquées par le HGCAL est limitée par la bande passante disponible entre le détecteur et le L1T. Les résultats obtenus montrent la possibilité de conserver des performances d'identification adéquates tout en satisfaisant ces limites imposées sur le système, fournissant des informations cruciales pour l'optimisation du *design* du système de génération de primitives de déclenchement du HGCAL.

Acknowledgments

Those three years of work have been a great journey, sometimes hard, but overall very fulfilling. The many people around me are most of the reasons it made it through to the end, and as such I would like to offer my deepest gratitude to all of them.

First of all, I thank the members of the jury for the precious time and attention they dedicated to my work, and in particular my rapporteurs that helped improve this work with their thoughtful comments and sound advice.

I thank all the friends I made at CERN or during summer schools, and with whom I shared very fun times. Meeting such nice people from all around the world was not only very enriching, but also gave me a sense that I was part of something bigger, a community of bright minds working towards better understanding of our world.

Then to the members of the LLR, I feel very fortunate to have been able to work alongside you. Though the sanitary crisis didn't allow to enjoy it to the fullest, it has been a blessing to meet such inspiring people, both intellectually and personally. A special thanks to all the young and bright scientist with whom I shared my office, for always maintaining a nice atmosphere and the sometimes silly but more often than not very enriching discussions.

A special thanks to Matteo, with whom I closely collaborated for the VBS analysis, it has been a pleasure working with you and though we didn't have the chance to meet in person as of now, I really hope we'll be able to in the near future.

Of course, this work would not have been possible without the guidance of Jean-Baptiste, my supervisor, to whom I owe my deepest gratitude. Though it has not always been easy, he has always been very understanding, and managed to keep me motivated during those long years. Thanks to him I have never felt left alone, even at times where everyone was confined at home. His knowledge and dedication to research are very awe-inspiring, and feel very lucky to have been able to worke with you.

My thanks go also to my friends and family, which allowed me to maintain a good work-life balance. Though they might not always have really known or understood exactly what I was doing, I have always felt supported by everyone.

I need to also thank my cats for always be there, and I really do mean *always*, and for warming my lap while I was working from home (not thanks for waking me up early in the morning though). Their role into maintaining my moral, alongside all my other pets cannot be understated.

And finally, nothing could have been achieved without the love and support of my future wife, Camille. You have always seen the best in me and for that (amongst many things), I owe you my greatest thanks.

Contents

Introduction

Throughout history, mankind has striven to understand the universe in which we live, down to what matter is composed of. Numerous experimental observations in the 19^{th} century have shown that the infinitely small world does not adhere to the same physics law that are used to describe macroscopic phenomena. Only the advent of quantum mechanics and special relativity during the following century could begin to explain the subatomic world. In the late 1960's, particle physics theories were gathered in a common framework to form the Standard Model (SM), based on quantum field theory. In parallel to those theoretical advances, technological progress opened the door to the conception of particle accelerators, and later colliders, providing access to powerful probes in the high energy realm. The CERN Large Hadron Collider (LHC) is currently the largest and most energetic representative of these experimental tools. With the discovery of the Higgs boson in 2012 by two of the LHC experiments, CMS and ATLAS, all the fundamental particles predicted by the SM have been observed experimentally. However, the SM description is in tension with some observations and theoretical considerations, suggesting that the current formulation is incomplete. Not all phenomena can be explain solely using the SM, as it does not account for gravitation for example, and is unable to explain the existence of dark matter or the matter/antimatter asymmetry. It is thus believed that the SM could be the expression at the currently reachable energy scale of a larger, more fundamental theory valid at higher energy scales. As such, the field of particle physics as now entered era of precise measurements, looking for the smallest weaknesses in the SM predictions.

One of the most promising sectors to investigate is known as the Vector Boson Scattering (VBS), a class of rare processes solely mediated by weak interaction where a quark from each incoming proton emit an electroweak gauge boson that interacts with each others. This class of process is intricately linked to the spontaneous symmetry breaking of the electroweak theory, permitting to probe its nature at different energy scales from the Higgs boson mass. In addition, interactions between multiple bosons in the tree diagrams offer a direct access not only to the well studied triple gauge boson coupling, but also to the quartic gauge boson couplings. Only the precise cancellation of diagrams involving triple couplings, quartic couplings and Higgs-vector bosons couplings allow the process to conserve unitary.

With their very low cross-sections, the first evidences of VBS processes have only been permitted with the statistics accumulated during the Run 2 (2016-2018) of the CERN's LHC. The first results concern the more easily reconstructed fully leptonic final state, where both bosons decays to leptons, but analyses concerning the semi-leptonic final states, where one of the two bosons decays hadronically and produce jets in the detector, such as the one presented in this thesis, are starting to become available with sophisticated analysis strategies. They rely on the very characteristic signature of VBS at the LHC, which produces two jets with a large invariant mass and pseudorapidity separation, to reject the very important background noise.

Even with the extensive dataset of the Run 2 and the addition of the currently ongoing Run 3 data sample, the sensitivity of the LHC is not foreseen to provide enough sensitivity for the observation and precise measurement of such rare processes. Thus, after the end of the current operations, the CERN plans to enter a new phase of exploitation of the LHC with an increased instantaneous luminosity of 4-5 times the nominal design during a decade, with

up to 200 simultaneous interactions per bunch crossing in the detectors. The unprecedented amount of data that will be delivered will permit an ambitious physics program, in particular in the physics beyond the standard model sector. Unfortunately, these benefits come at the cost of serious challenges for the detectors, both in terms of radiation and of occupancy, and for the data acquisition systems. To face those, the CMS collaboration will have to perform several upgrades to its detectors, amongst those the replacement of the calorimeter endcaps by the Highly Granular Calorimeter (HGCAL), on which I had the opportunity to work, and the upgrade of the trigger system.

The HGCAL is an ambitious detector, not only the largest ever silicon-based detector with the finest segmentation achieved for this class of calorimeter, but also providing precise timing information. Those features are required to reject the additional interactions in the detector that would degrade the sensitivity to interesting hard scattering events. The use of the information from the enormous number of channels of the HGCAL at the level-1 trigger level is a daunting task that CMS aims to achieve. The usage of FPGA boards providing fixed latency and reconfigurability for future improvements make possible the use of information computed in the HGCAL trigger primitive generation (TPG) for the identification of the energy cluster deposited in the calorimeter. Due to the requirements of the trigger system architecture, which must operate in real time during data-taking to decide which events to record, the design of the whole system is an extremely challenging task. Part of the work presented in this thesis was dedicated to the optimization of the HGCAL TPG for the identification of electromagnetic events. Of course, the discrimination performance must be maximized, but at the same time the size of the TPG data must be limited due to the available bandwidth constraints, and even the size of the machine learning algorithms performing the classification must be minimized due to resource constraints.

The first two chapters of this thesis provide some insight into the theoretical and experimental framework in which this work is situated. Chapter 1 sets the theoretical framework, with a brief history of particle physics, and the description of the Standard Model and its limitations. Particular attention is given to the vector boson scattering process and the current state of the related experimental searches. In Chapter 2, the experimental set up of the CMS detector at LHC is detailed. The three remaining chapters present a description of the work realized during this thesis. Chapters 3 describes the search for the VBS ZV semi leptonic production to which I was one of the main contributor. The whole analysis process is described, from the definition of the physics objects to the statistical extraction measurement of the process. The selection of data events is detailed as well as the production and correction of Monte Carlo simulations used to compute the expected VBS production according to the standard model. Particulars attention is given to the discrimination of the signal from the various backgrounds with neural networks. Chapters 4 and 5 are dedicated to the HL-LHC phase. In Chapter 4 the plans for the LHC upgrade are detailed, with a focus on the HGCAL and the upgraded level-1 trigger. In chapters 5 are presented the studies I realised towards the optimisation of the HGCAL TPG for the identification of electromagnetic showers at L1T.

Chapter 1

Theoretical context and motivations

Understanding the world around us at its fundamental level is an ambitious goal, but it is the undertaking of the particle physics field. Particle physicists strive to identify the basic set of building blocks from which all observed phenomena can be explained. The Standard Model (SM) is the theory that was developed after spectacular progresses during the past centuries. It endeavors to provide a set of particles and fundamental interactions to describe in the simplest way possible all observed physical phenomena in the universe. Simplicity has indeed been one of the driving paradigms towards the building of this model; and one that has led to many theoretical and experimental successes. Nonetheless, holes and limitations still exist and many paths towards improvement are currently being studied, with particle colliders experiments at the forefront of research.

1.1 The standard Model of particle physics

1.1.1 A brief history of particles

The current Standard Model of particle physics is a beautiful theory that describes all elementary particles known in the universe as well as the fundamental forces that govern all their interactions, with the exception of gravitation. It is a relatively recent achievement in the history of science, with its current formulation dating from the mid-1970's. The idea of an indivisible particle can be traced back to Greek antiquity and the idea of the atom, meaning *an entity that cannot be cut*, was at that time more of a philosophical notion than scientific knowledge. The idea will not begin to be treated as a scientific notion until the 18^{th} century and the development of chemistry, and then of modern physics. A. Lavoisier enunciates his law of conservation of mass ("*Rien ne se perd, rien ne se crée, tout se transforme*" [1]) and crystallography is developed. The idea of elements rises again, A. Avogadro distinguishes the molecule from the atom and the chemical elements are being classified, up to the achievement in 1869 by D. Mendeleev of the *Periodic Table of elements* [2]. The elements then amount to more than a hundred, with properties that seem to appear in a periodic way once ranked by their mass, which also seems to be a multiple of the mass of the lightest element, the hydrogen. This is

enough for some scientists as W. Prout to speculate the existence of substructures of those elements, which would then be composed of simpler, more fundamental particles.

The field of subatomic physics really blossoms towards the end of the 19^{th} century, with the discovery in 1897 by J.J. Thomson of the first of such elementary particle: the electron [3]. By studying cathodic rays, Thomson discovers that when directing the rays towards two metallic plates with opposite charge, the flux would move towards the positively charged one. By varying the magnetic field, he manages to calculate the ratio of the mass to the charge for those rays, which remained identical even when changing the metal that the plates are made of. This led Thomson to postulate the existence of a negatively charged particle governing the nature of electricity: the electron, which is still considered elementary to this day.

It is also known at the time that the atoms are neutral, which led the physicists toconclude that a positively charged partner to the electron should exist. Thomson postulates that the negatively charged electrons, 1800 times lighter than the lightest element, the hydrogen, are distributed in a sea of positive charges. However, a few years later, E. Rutherford who, with the help of H. Geiger and E. Marsden, was building an experiment for alpha particles, discovers the existence of a tiny nucleus inside the atom. When bombarding a metal foil with alpha particles the *soup* model predicted that they should pass through without scattering, since the positive charge concentration should not be enough to deflect a very fast alpha particle. However, the experiment showed that scattering did happen, leading Rutherford to surmise the existence of a solid core, the nucleus, inside all elements and concentrating the positive charge of the atom. Rutherford would name it the *proton*, from the Greek word for "first".

In the 1920's, the existence of new types of particles is postulated. P. Dirac theorises the *antimatter*, composed of particles similar to the already known ones in matter but with opposite charge. They offered an elegant answer to some of the questions that the recent *quantum mechanics*, a framework for describing the universe at its smallest scale, was beginning to raise. The discovery of the beta decay required the existence of a neutral particle with very low mass in order to satisfy the energy conservation paradigm. This particle, called neutrino, would only be experimentally discovered in 1956. The neutron, a neutral counterpart to the proton, is discovered by one of Rutherford's own students, J. Chadwick, after having been predicted many years before in order to solve some shortages of the proton-electron model.

At the beginning of the 1930's, the atom is thus modeled as a positively charged nucleus, composed of protons and neutrons of similar masses and collectively known as *nucleons*, with smaller electrons orbiting around it: this is known as the Bohr model of the atom. The rest of the decade sees the rise of the field of cosmic rays, particles produced outside of our planet and that decay when entering the atmosphere. Their study would lead to the discovery of several new particles:

- the positron, the antimatter counterpart of the electron.
- the pions, that had been predicted by the strong interaction theory developed later in Sec. 1.1.2 .
- the muon, a heavier version of the electron.
- particles similar to the proton but with higher masses that were described as *strange*.

This multiplicity of new particles, that seem to go against the desire for simplicity that had been guiding the scientist up until now with success, questioned the elementarity of those particles. This, in addition to other theoretical considerations and the development of particle accelerators where a entire *zoo* of particles is discovered, led to the creation in 1964 by M. Gell-Mann of the *quark* model. The quarks are a new subcomponents that combine to form composite particles such as the nucleons. At the time of its creation, the model contains two *flavors* of quarks: the up quark, with a positive charge Q_u=2/3 e where e is the charge of the electron (which remains as a fundamental unit of the electrical charge), and the down quark of Q_d = -1/3 e. Through the years flavors will be added to form a 6-quark system accompanied by their antiquarks as developed in Sec. 1.1.2.

The mathematical description of those particles, and the three interactions that govern their behavior (strong, weak, and electromagnetic interactions, the gravitation being negligible at the particle scale) form the Standard Model. Based on quantum field theory, it bore numerous predictions that would be experimentally observed at particles colliders through the second half of the century and beginning of the 21^{st} century, culminating in 2012 with the discovery of the Higgs boson that had been postulated back in 1964.

This is a brief overview of the rich and fascinating history of the particle physics fields, and more in depth account can be found in Ref. [4].

1.1.2 Current state of the Standard Model

The entire work described in this thesis is based on the SM, of which a brief historical account was given in Sec. 1.1.1. This very successful theoretical framework has led to many predictions and discoveries and the discovery in July 2012 of the Higgs boson by the ATLAS and CMS experiments at the CERN LHC [5, 6] offered the last fundamental block to the model.

For its description of the universe, the SM differentiates two types of fundamental elements: the particles from which matter is made of are known as *fermions* of spin 1/2, and the particles that act as mediator of the fundamental interactions are called *bosons* of spin 1, which are exchanged by interacting fermions. In the mathematical framework of the SM, the Quantum Field theory, all fermions and bosons are described as excited states of fields. In a system, the evolution of those fields is governed by the Lagrangian based on underlying symmetries. Three of the four fundamental interactions stem from local gauge symmetry groups, the strong interaction corresponding to $SU(3)_C$, and the weak and electromagnetic interactions represented by $SU(2)_L \times U(1)_Y$ as unified as the electroweak (EW) interaction. A particle quantum number, or *charge*, is associated to each of those interactions, the electrical charge for the electromagnetic interaction, the *color* charge for strong interaction and *flavor* for the weak interaction. The gravitational interaction is not included in this formalism of the SM, which is one of the weaknesses of the SM, but can be neglected at the high energy scales of particle physics where it is several order of magnitudes weaker than the weak interaction. An additional field, giving rise to the Higgs boson, explains the mass of the bosons based on a mechanism called *spontaneous symmetry breaking*. An overview of the particle content of SM is shown in Fig. 1.1 and the properties of the different types of particles, as well as a more in-depth discussion of each force is given in the following paragraphs.

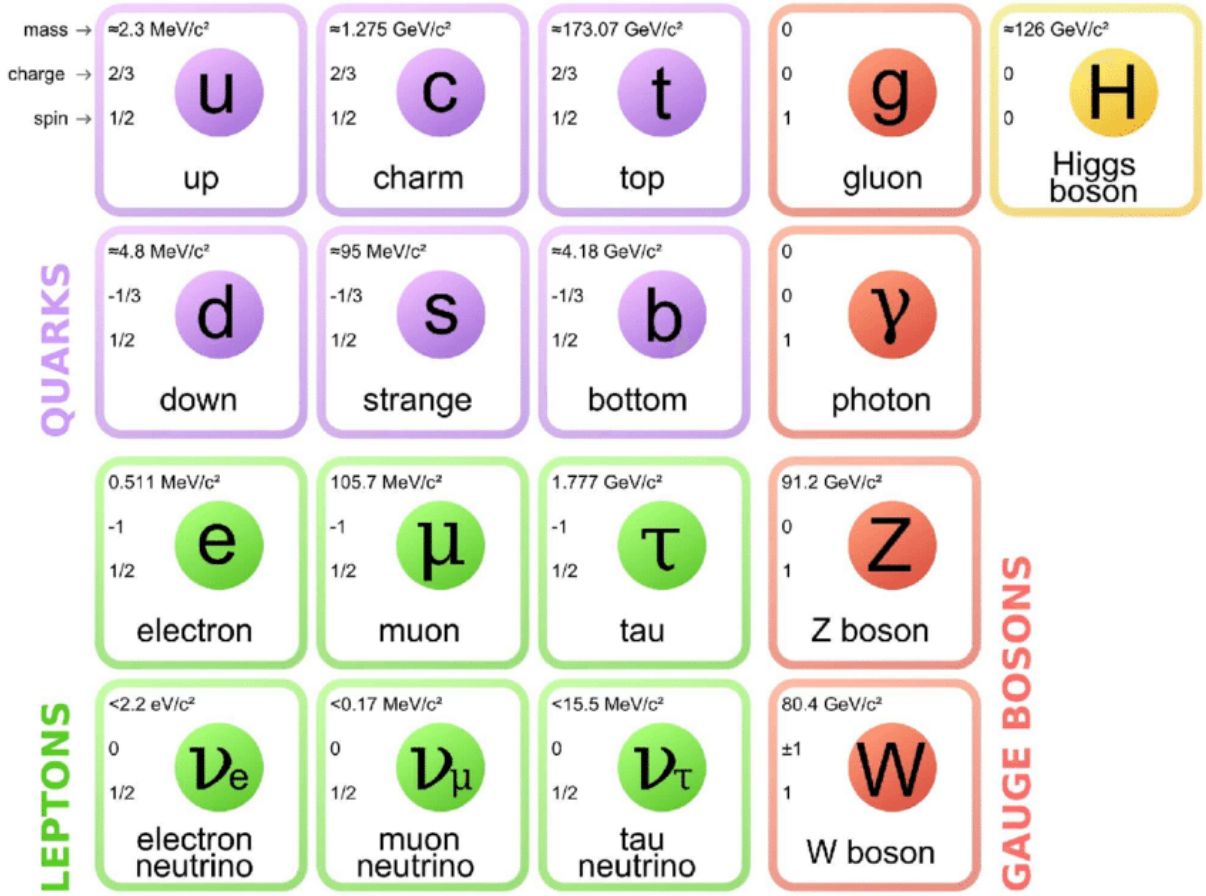


Figure 1.1: Standard model of elementary particles: the 12 fundamental fermions and five fundamental bosons. The masses of the particles corresponds to the 2019 data and have been reevaluated since, see Ref. [7] for the latest measurements. Figure taken from Ref. [8].

Fermions

The building blocks of matter are the fermions, named by their compliance to the Fermi-Dirac statistics, with a spin $s = 1/2$. They satisfy the Pauli exclusion principle, which states that two or more identical fermions cannot share the same quantum state within a quantum system simultaneously. There exist two categories of fermions: the *quarks* and the *leptons*. For each category, six members have been observed experimentally and are divided into three *generations*, each generation composed of a pair of up-like quark and down-like quark, with electrical charges +2/3 and -1/3 respectively, and a lepton pair of electrical charges -1 and 0 (in units of *e*, the absolute value of the charge of the electron). Those 12 fermions are completed by the same number of antiparticles, each antiparticle identical to one of the fermion but with opposite quantum numbers.

The quarks are particles subjected to the three forces of SM, the electromagnetic, weak and strong force and possess a quantum number associated to each one of those forces. Their electrical charge is $Q = +2/3$ or $Q = -1/3$, the *color* charge corresponds can take three values (red, blue or green), and the *flavor* depends on the type of quark. They can not be observed individually in experiments but only as bound states called hadrons due to the color con-

finement predicted by Quantum Chromodynamics (QCD), the quantum theory of the strong interaction. Those hadrons can be composed of a quark-antiquark pair, in which case they are called *mesons* or in aggregates of three quarks (or antiquarks) forming *baryons*, such as the proton. The first family of quark is composed of the up (u) and down (d) quarks with masses of 2.2 MeV [7] and 4.7 MeV [7] respectively, that form the ordinary matter such as nucleons. Heavier quarks form the second generation, with the charm (c) at 1.3 GeV [7], and strange quark (s) at 93 MeV [7]. The third generation is made of the heaviest quark pair, with the top (t) quark at 172.9 GeV [7] and the beauty or bottom (b) quark at 4.18 GeV [7]. The top quark the heaviest fundamental particle in the SM and has a lifetime so short ($\tau_t = 0.5 \times 10^{-24}$ s [7] that it decays before forming any bound state.

The other category of fermions are the leptons, that interact only via electromagnetic and weak forces. The three charged leptons are the electron (e), the muon (μ) and the tau (τ), with Q=-1 electrical charge and masses of 511 KeV, 106 MeV and 1.8 GeV [7] respectively. While not considered stable as the electron, the large lifetime of the muon allows it to be detected before decaying in particle colliders. The tau lepton however has a very short lifetime of $\tau_\tau = 2.9 \times 10^{-13}$ s[7] and can thus only be detected in colliders by reconstructing it from its decay products. To each of the charged leptons is associated a neutral particle called neutrino (ν_e, ν_μ and ν_τ) that only interacts via weak force. The neutrinos have long been considered massless but they were observed to oscillate between flavor eigenstates in 2015[9], proving that they are in fact massive. Experimentally, they do not interact in the particle collider detectors and only appear in the form of missing energy in events.

Bosons

The gauge bosons are the *force carriers* of the SM, giving rise to the fundamental interactions between other particles. They possess integer spin $s = 1$ and follow the Bose-Einstein statistics. The *gauge bosons* comes from the fact that, in terms of quantum fields theory, their existence arises from local gauge invariance.

The *gluons* (g) are the mediators of the strong interaction. They are massless particles, of neutral electric charge but carry a color and anti-color quantum number and come in eight color states. They interact with every colored particle such as the quarks but also with themselves. The electromagnetic gauge boson is the photon (γ), also massless, electrically neutral and not colored. It does not possess *weak flavor* and does not self-interact. The mediators of the weak interaction are the weak bosons, $W^{\pm}$ and Z bosons, with masses of respectively 80.4 GeV[7] and 91.2 GeV [7] and which can self-interact. The limited range of the weak interaction is directly correlated to its gauge boson not being massless as opposed to the electromagnetic force (the strong force limited range stems from color confinement and not mediator mass).

The Higgs boson is a special kind of boson, not mediating any fundamental interaction and with a spin $s = 0$. It is named after the scientist who postulated its existence 60 years before. Its role in the SM is unique as it does not mediate a fundamental interaction and does not arise from local gauge invariance. It originates in the spontaneous symmetry breaking mechanism which gives an explanation to the gauge bosons masses without making the theory non-renormalizable. It also couples to the fermions via Yukawa interaction and confers them their masses. It is electrically neutral, has no color or flavor and its mass has been ex-

perimentally measured at 125 GeV [10].

Strong interaction

In the SM, the strong interaction is described by the Quantum Chromodynamics theory that governs the interactions of *colored* particles such as quarks and gluons. The QCD gauge theory is non-abelian and originates from the SU(3)$_C$ local symmetry group, with C standing for color. The color is the charge quantum number associated to the strong interaction giving its name to the theory.

The Dirac Lagrangian of a quark free field q corresponding to a $s = 1/2$ spin is written as:

$$\mathcal{L} = \bar{q}(i\gamma^\mu \partial_\mu - m)q, \tag{1.1}$$

where m is the mass of the fermionic field, γ^μ the Dirac matrices and ∂_μ denotes the space-time partial derivative. This Lagrangian is invariant under the $SU(3)$ transformation of the form

$$q \to q' = e^{-ig_s \frac{\lambda^\alpha}{2} \alpha_\alpha} q, \tag{1.2}$$

with g_s a constant and $\lambda^\alpha/2$ the 8 generators of the $SU(3)$ group, 3×3 traceless hermitian matrices called Gell-Mann matrices. The local invariant is obtained from this global invariant by introducing gauge fields in the Lagrangian expression of Eq. 1.1 in the form of a covariant derivative D_μ:

$$D_\mu = \partial_\mu + ig_s \frac{\lambda^\alpha}{2} G^\alpha_\mu, \tag{1.3}$$

where G^α_μ is the gauge vector field corresponding to the eight gluons. In order to maintain the local invariance and compensate the additional term in Eq. 1.3, the gluonic fields need to transform under $SU(3)$ as:

$$G^\alpha_\mu \to G'^\alpha_\mu = G^\alpha_\mu + \partial_\mu \alpha^\alpha + g_s f_{abc} \alpha^b G^c_\mu, \tag{1.4}$$

with f^{abc} the structure constants of the $SU(3)_C$ group that must satisfy the commutation rules $\left[\frac{\lambda^a}{2}, \frac{\lambda^b}{2}\right] = i f_{abc} \frac{\lambda^c}{2}$.

The Lagrangian for this fermionic free field must be completed with terms describing the propagation of the gluons. The kinetic energy is written as $-\frac{1}{4} G^{\mu\nu}_\alpha G^\alpha_{\mu\nu}$ where $G^\alpha_{\mu\nu}$ is the strength of the gluon field defined as

$$G^\alpha_{\mu\nu} = \partial_\mu G^\alpha_\nu - \partial_\nu G^\alpha_\mu - g_s f_{abc} G^b_\mu G^c_\nu. \tag{1.5}$$

The last term in Eq. 1.5 gives rise to the cubic and quartic self-interactions of the gluons.

The total Lagrangian for the QCD can then be written as

$$\mathcal{L}_{QCD} = i\bar{q}\gamma^\mu \partial_\mu q - m\bar{q}q - g_s \bar{q}\gamma^\mu \frac{\lambda^\alpha}{2} q G^{\mu\nu}_\alpha - \frac{1}{4} G^{\mu\nu}_\alpha G^\alpha_{\mu\nu}. \tag{1.6}$$

The constant g_s determines the strength of the interaction. It is also commonly used as the strong coupling constant $\alpha_s = g_s^2/4\pi$.

Electroweak interaction

In the SM, the weak and electromagnetic interactions are unified under a local gauge invariance based on the $SU(2)_L \times U(1)_Y$ symmetry.

The electromagnetic interaction is described by the Quantum Electrodynamics (QED) theory developed by Glashow, Salam and Weinberg [11, 12, 13], that governs phenomena involving the electrical charge of particles mediated by the exchange of photons. QED is an abelian gauge theory based on the $U(1)_{em}$ symmetry group. Similarly to the QCD, the Lagrangian of QED is expressed as:

$$\mathcal{L}_{QED} = i\bar{\psi}\gamma^\mu\partial_\mu\psi - m\bar{\psi}\psi - eQ\bar{\psi}\gamma^\mu\psi A_\mu - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}, \tag{1.7}$$

where ψ is the fermionic field, which can be a quark or charged lepton. The first term is the free field Lagrangian for a $s = 1/2$ spin particle and the second one is the mass term. The third term arises from the covariant derivative of $U(1)_{em}$:

$$D_\mu = \partial_\mu + ieQA_\mu, \tag{1.8}$$

describing the interaction between a photon represented by the gauge potential A_μ and a fermion of electrical charge Q. The strength of the interaction is proportional to e, the charge of the electron, and Q. The last term describes the propagation of the photon with $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ the Maxwell tensor. There is no mass term for the photon and the abelian nature of QED prevents self-interaction.

The weak interaction is more complex to describe in terms of quantum field theory. It is based on the non-abelian $SU(2)_{weak}$ symmetry group. It has been experimentally observed to violate parity, as opposed to the strong and electromagnetic interactions. This fact is accounted for theoretically by introducing a *chirality* property to the fermionic field. This chirality is represented by a Lorentz-invariant operator $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$ which eigenvalues, 1 or -1, give rise to *left* and *right* chirality fields with the projectors:

$$P_L = \frac{1-\gamma^5}{2} \text{ and } P_R = \frac{1+\gamma^5}{2} \tag{1.9}$$

The fermionic fields are then represented as left chirality doublets (Ψ_L) and two right chirality singlets (ψ_R, ψ'_R). The weak Lagrangian can then we expressed as

$$\mathcal{L}_{weak} = i\bar{\Psi}_L\gamma^\mu D_\mu\Psi_L + i\bar{\psi}_R\gamma^\mu D_\mu\psi_R + i\bar{\psi}'_R\gamma^\mu D_\mu\psi'_R - \frac{1}{4}W^i_{\mu\nu}W^{\mu\nu}_i, \tag{1.10}$$

where the mass term is omitted. The covariant derivative of the $SU(2)_{weak}$ group is written as

$$D_\mu = \partial_\mu + ig_w T^i_\mu W^i_\mu, \tag{1.11}$$

with $W^i_{\mu\nu}$ the tensor field defined as:

$$W^i_{\mu\nu} = \partial_\mu W^i_\nu - \partial_\nu W^i_\mu - g_w\epsilon_{ijk}W^j_\mu W^k_\nu, \tag{1.12}$$

where $i = 1, 2, 3$. The W^i_μ are the three gauge fields corresponding to the $W^\pm$ and Z boson and $T^i_\mu = \sigma_i/2$ are the generators of the $SU(2)_{weak}$ group based on the Pauli matrices σ_i. The ϵ_{ijk} are the structure constants of the group following the commutation rules

$[\sigma_i/2, \sigma_j/2] = i\epsilon_{ijk}\sigma_k/2$ and g_w is the coupling constant of the interaction. As for QCD, the last term in Eq. 1.12 contains trilinear and quadrilinear self interactions terms. The quantum number associated to the $SU(2)_{weak}$ group is called the *weak isospin* $I_{1,2,3}$. The I_3 component in particular is equal to 0 for the ψ_R and ψ'_R as they are $SU(2)_{weak}$ singlets, while the left chirality field Ψ_L doublet possess $I_3 = +1/2$ and $I_3 = -1/2$ for the upper and lower members respectively. The $SU(2)_{weak}$ only acts on the left-handed component of the fermionic fields and is thus commonly referred to as $SU(2)_L$.

The gauge field described in Eq. 1.12 does not directly describes the physical $W^{\pm}_{\mu}$ and E_{μ} fields associated to the weak bosons. The $W^{\pm}$ bosons can be expressed as the linear combinations

$$W^{\pm}_{\mu} = \frac{1}{\sqrt{2}}(W^1_{\mu} \mp iW^2_{\mu}). \tag{1.13}$$

In order to describe the Z boson field, the $U(1)_Y$ abelian group associated to a gauge field B_{μ} described by a similar Lagrangian to $\mathcal{L}_{QED}$ with a *weak hypercharge* Y quantum number must be introduced. The corresponding Lagrangian is written as

$$\mathcal{L}_Y = i\bar{\psi}\gamma^{\mu}D_{\mu}\psi - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}, \tag{1.14}$$

without the mass term. The associated covariant derivative for $U(1)_Y$ is

$$D_{\mu} = \partial_{\mu} + ig_Y B_{\mu}, \tag{1.15}$$

with a coupling constant g_Y. The $B_{\mu\nu}$ tensor field is defined as

$$B_{\mu\nu} = \partial_{\mu}D_{\nu} - \partial_{\nu}D_{\mu}. \tag{1.16}$$

The action of the $U(1)_Y$ group on the fermionic field ψ couples to both the left-handed and right-handed terms in an interaction term $-g_Y\bar{\psi}\gamma^{\mu}(Y/2)\psi B_{\mu}$.

The Lagrangians of the electromagnetic and weak interactions can be combined in a unified electroweak theory represented by the $SU(2)_L \times U(1)_Y$ symmetry group. The electromagnetic field A_{μ} and Z_{μ} boson field can then be obtained by mixing the B_{μ} and W^3_{μ} fields by the rotation

$$\begin{pmatrix} A_{\mu} \\ Z_{\mu} \end{pmatrix} = \begin{pmatrix} \cos\theta_W & \sin\theta_W \\ -\sin\theta_W & \cos\theta_W \end{pmatrix} \begin{pmatrix} B_{\mu} \\ W^3_{\mu} \end{pmatrix} \tag{1.17}$$

with the Weinberg angle θ_W. The charge Q, hypercharge Y and Isospin I are related following the Gell-Mann-Nishijima equation $Y = 2(Q - I_3)$ and the Weinberg angle can be expressed in terms of the e, g_w and g_Y constants as $e = g_w\cos\theta_W = g_Y\sin\theta_W$.

The complete electroweak Lagrangian can be written as

$$\mathcal{L}_{EW} = i\bar{\Psi}_L\gamma^{\mu}D_{\mu}\Psi_L + i\bar{\psi}_R\gamma^{\mu}D_{\mu}\psi_R + i\bar{\psi}'_R\gamma^{\mu}D_{\mu}\psi'_R - \frac{1}{4}W^i_{\mu\nu}W^{\mu\nu}_i - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}. \tag{1.18}$$

This unification of electromagnetic and weak interactions gives rise to terms allowing the trilinear and quadrilinear interactions between the weak bosons and the photon.

It is possible to write a global Lagrangian comprising both $\mathcal{L}_{QCD}$ and $\mathcal{L}_{EW}$ terms, resulting in an unified gauge theory based on $SU(3)_C \times SU(2)_L \times U(1)_Y$ symmetry. This unified

Lagrangian however is lacking explicit terms for the boson and fermion masses, which can not be added without breaking the gauge invariance, contradicting the observations of the masses of the W and Z bosons. In order to explain this phenomena, a breaking of the electroweak symmetry must be introduced, described in the SM by the Brout-Englert-Higgs (BEH) mechanism.

Electroweak symmetry breaking

As explained in the previous section, the masses of the weak gauge bosons couldn't be accounted in the electroweak Lagrangian without breaking local gauge invariance. A solution to this problem was postulated in 1964 independently by F. Englert and R. Brout[14], P. Higgs[15] and G. Guralnik, C.R. Hagen and T. Kibble [16] in the form of the *electroweak symmetry breaking* (EWSB). It is based on the introduction of two complex scalar fields (ϕ^+, ϕ^0) forming a weak isospin doublet ϕ called the Higgs field. This field must have a potential $V(\phi)$ that is invariant under the symmetry of the system but that *spontaneously break* this symmetry when a ground state is chosen. This potential is expressed as

$$V(\phi) = \mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2, \tag{1.19}$$

where μ and λ are constants. The BEH field can be written in terms of real scalar fields ϕ_i (i=1,2,3,4):

$$\phi = \begin{pmatrix}\phi^+\\ \phi^0\end{pmatrix} = \begin{pmatrix}\phi_1 + i\phi_2\\ \phi_3 + i\phi_4\end{pmatrix}. \tag{1.20}$$

The associated Lagrangian describing the interaction of the field is

$$\mathcal{L}_{BEH} = (D_\mu\phi)^\dagger(D_\mu\phi) - V(\phi), \tag{1.21}$$

where D_μ is the covariant derivative.

This Higgs potential presents an unstable local maximum at $\phi = 0$ but a continuum of ground states is found at

$$|\phi^\dagger\phi| = \frac{\mu^2}{2\lambda} = \frac{v^2}{2} \tag{1.22}$$

where v is known as the vacuum expectation value (VEV). Choosing a ground state does not break the Lagrangian gauge invariance but *spontaneously* breaks the symmetry. The ground state breaks the electroweak $SU(2)_L \times U(1)_Y$ symmetry but is invariant under the $U(1)_{em}$ symmetry group. By selecting a particular gauge that is parallel to one of the doublet field components and expanding this term around the VEV the field becomes

$$\phi = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\ v + h(x)\end{pmatrix} \tag{1.23}$$

By inserting this formula into the Lagrangian expression, new couplings appear between the

h field and the vector bosons:

$$\begin{aligned}\mathcal{L}_{BEH} =& \frac{1}{2}\partial_\mu h \partial^\mu h + \mu^2 h^2 \\ &+ \frac{g_w^2 v^2}{4} W_\mu^+ W^{-\mu} + \frac{(g_w^2 + g'^2)v^2}{8} Z_\mu Z^\mu \\ &+ \frac{g_w^2 v}{2} h W_\mu^+ W^{-\mu} + \frac{g'^2 v}{2} h Z_\mu Z^\mu \\ &+ \frac{g'^2}{4} h^2 Z_\mu Z^\mu + \frac{\mu^2}{v} h^3 + \frac{\mu^2}{4v^2} h^4,\end{aligned} \tag{1.24}$$

where $g' = g_w \text{tan}\theta_W$. The first term yields the mass of the Higgs field $m_H^2 = 2\lambda v^2 = 2\mu^2$ with its value a free parameter of the SM. A mass term also appears for the weak vector bosons:

$$\begin{aligned}m_Z =& \frac{\sqrt{g^2 + g'^2}}{2} v \\ m_{W^\pm} =& \frac{gv}{2} = m_Z \text{cos}\theta_W\end{aligned} \tag{1.25}$$

The mass of the vector boson arises from the explicit choice of a ground state breaking the symmetry. The A_μ field does not appear in Eq. 1.24 so the photon remains massless in accordance with the observations. Trilinear and quadrilinear couplings to the vector boson appears (HVV and $HHVV$, where $V = W, Z$), as well as trilinear and quadrilinear self couplings.

In a similar fashion, the fermions acquire their mass by the introduction of the so-called Yukawa interaction of the Higgs field with the left- and right-handed fields s described in the Lagrangian

$$\mathcal{L}_{Yukawa} = -y_{f'}\big(\bar{\Psi}_L \phi \psi'_R + \bar{\psi}'_R \phi^\dagger \Psi_L\big) - i y_f \big(\bar{\Psi}_L \sigma_2 \phi^* \psi_R + \bar{\psi}_R \phi^\dagger \sigma_2 \Psi_L\big), \tag{1.26}$$

where y_f and y'_f are the Yukawa couplings for the up-type fermions and down-type fermions respectively. After choosing the ground state and breaking the symmetry, this Lagrangian becomes

$$\mathcal{L}_{Yukawa} = \sum_f -m_f \bar{\psi}\psi - \frac{m_f}{v} h \bar{\psi}\psi \tag{1.27}$$

where $m_f = v y_f / \sqrt{2}$ is the mass of the fermions that is proportional to the strength of the Yukawa interaction between a given fermion and the Higgs field. In particular, the top quark has the highest mass amongst known fermions and should then have the biggest coupling to the Higgs boson.

1.2 Beyond standard model

Despite its impressive degree of precision in depicting most phenomena, and the amount of discoveries resulting from theoretical predictions, this formulation of the SM suffers from limitations and fails to adequately explain some observations. New physics theories Beyond Standard Model (BSM) have been invoked to try and address several of the SM shortcomings.

1.2.1 Standard model limitations

Amongst such challenges for the SM to explain, one can cite:

The dark matter and dark energy: The SM is currently inconsistent with the main paradigm of cosmology, the Λ-CDM model. Observations of the cosmic microwave background and the galaxies rotation speed show that only 5% of the matter in the universe can be explained by the baryonic matter described in the SM. The SM fails to offer a candidate for the Cold Dark Matter that should amount to 25% of the universe density and can neither explain the remaining 70% attributed to the dark energy.

The matter-antimatter asymmetry: According to the SM, particles and antiparticles should have been created in the same quantity. An explication to the fact that the visible universe is uniquely composed of matter is required.

Inclusion of gravitation: While the SM aims to formulate a unified description of all particles and their interactions, it fails to provide a common framework that could formalize the gravitation alongside the other three forces. The existence of an associated boson, the *graviton* has been postulated and could help bridge the gap, but no observation of it has yet been realized. A quantum field theory of gravitation at Planck scale, where the effects of the gravitational force are expected to become important, is necessary to build a reliable theory of the early universe.

The masses of the neutrino: Originally described as massless particles, recent observations of neutrino oscillation have proven that the neutrino do have masses. Those small masses can not be derived from the Higgs interaction and could require additional parameters to the SM.

The hierarchy problem: If a new physics coupling to the Higgs is present at a high energy scale, extremely fine tuning of the parameter is required to get the correct physical mass parameters [17].

Complexity: One of the driving motivation for the scientists building of a *theory of everything* has been to keep it simple. With its 19 arbitrary parameters (and maybe more if accounting for the neutrino masses), seemingly arbitrary number of generations of the fermions and order of magnitudes between the lightest and heaviest particles, the SM can be considered as too complicated to be a fundamental description.

1.2.2 Beyond Standard Model theories

The previous section underlines some of the challenges a successful theory of everything has to address. Nonetheless, the SM is the most successful theory of particles to date that correctly explains most of the observations realized and has made many successful predictions. Most efforts are towards using the SM as a basis and completing it with new physics extensions accounting for its deficiencies. Such theories are collectively known as Beyond Standard Model and the current section aims to deliver a quick overview of some of them that have gained the most traction in the community without providing specific details.

Supersymmetry

Supersymmetry (SUSY) is an extension of the SM that adds a new class of symmetries to the Lagrangian, giving rise to a new supersymmetric partner to each SM particle with a spin differing by 1/2, associating fermions with bosons. Those particles would have much higher mass than their ordinary counterpart due to a supersymmetry breaking, as light candidates would have already have been observed experimentally.

Grand Unification theories

The SM is based on three gauge symmetries, the $SU(3)_C$ group for the strong force and the $SU(2)_L$ and $U(1)_Y$ groups for the electroweak interaction. Grand unified theories (GUT) speculates that those three groups could result from a single unified gauge theory with a unique coupling constant at high energy scale. The gauge symmetries of the SM would then just result from this symmetry being spontaneously broken at observed energy scales. This is supported by the fact that the couplings of each one of the three symmetries become similar around 10^{16} GeV. Popular choices for this unified gauge symmetry are the $SU(5)$ and $SO(10)$ groups. Some of the predictions of most GUT are the existence of magnetic monopoles and the instability of the proton, that have yet to be observed, setting strong limits on the possible GUT.

Quantum gravitation theories

Several attempts towards including the gravitation in the same framework as the three SM interactions have been developed. In particular, the loop quantum gravity theory is one such candidate to the mathematical unification of the four forces by postulating that the space-time structure is composed of finite loops woven into a fine network. It is based on quantum field theory and general relativity, and thus requires less drastic changes to the current theories compared to other candidates. However, recent studies on the effects of quantum gravity on the speed of light are in tension with several models of quantum gravitation [18]. Another popular group of theories are String theories, that aim to revise the SM by postulating that the current particles are replaced by one-dimensional objects called vibrating strings. The postulates of those theories are drastic, amongst the many variants the M-theory requiring the existence of 11 dimensions for example. Many criticism of those theories exists, amongst them the multiplicity of solutions that could accommodate any observations, thus being irrefutable.

1.2.3 Effective Field Theories

The Effective Field Theory (EFT) framework [19, 20] is an approach to describe effects at low energy scales of BSM physics happening at an energy scale larger than can be probed in current experiments. An EFT only includes the degrees of freedom relevant to a particular energy scale while ignoring the substructures at lower scale or higher energy scale. For example, at an energy scale lower than the mass of a heavy particle, the lagrangian can ignore the corresponding degree of freedom and still be valid. By imposing the known symmetries

$SU(3)_C \times SU(2)_L \times U(1)_Y$, the lagrangian valid at the EW scale can be expanded as

$$\mathcal{L}_{\text{eff}} = \mathcal{L}_0 + \frac{1}{\Lambda}\mathcal{L}_1 + \frac{1}{\Lambda^2}\mathcal{L}_2 + \dots, \tag{1.28}$$

where $\mathcal{L}_0$ is the SM lagragian of dimension four, the $\mathcal{L}_m(m = 1, 2, ...)$ are dimension $(4 + m)$ terms and Λ an energy scale. Those higher dimension terms, allowed by the lift of the renormalizability restriction of EFTs, can be expressed in terms of BSM operators as

$$\mathcal{L}_m = \sum_i f_i \mathcal{O}_i^m \tag{1.29}$$

where the f_i are the coupling strength to these operators called Wilson coefficients. The baryon and lepton number conservation forbids the presence of operators of odd dimensions so the lower order terms of EFTs are the operators of dimension 6 and 8. The dimension 8 operators in particular describe anomalous quartic gauge couplings (aQGC) that can appear in VBS processes such as the subject of this thesis.

1.3 Vector Boson scattering

1.3.1 Particle scattering

The equations of motion of the SM fields can be derived from the Lagrangian described in Sec. 1.1.2. Simple solutions exist for the kinetic term by treating the particles as free fields, but the complex interaction terms must also be computed. When the interaction couplings are small, this can be done by treating the interactions as *perturbations* of the free-field propagation. This is represented by the interaction Hamiltonian, and the number of times the Hamiltonian acts in the equation is the order of the perturbation expansion.

The protons colliding in the LHC are not fundamental particles but composite. While at low energies they can be described as bound states of uud quarks, the description at high energy is more complex, with the quarks radiating gluons that can themselves create $q\bar{q}$ pairs. Those gluons and quarks are collectively named *partons*, which can only be described through the Parton Distribution Function (PDF) giving the probability density of finding a parton with a given fraction of the composite particle momentum. A correct description of the scattering events in the LHC needs to be performed at the parton level.

Those partons can be treated as free fields that scatter against each other in the colliders, creating several particles that are then measured in the detectors. The mathematical expression for the behavior of such particles colliding can be complex and difficult to compute. The Feynman diagrams offer a simple graphical depiction of the perturbative expansion terms and correspond to the elements of the scattering matrix. The Feynman representation for the possible interactions allowed in the SM are illustrated in Fig. 1.2. The lines represent the particles propagation, with a different style depending on their nature: full lines for the fermions, wavy lines for photons and weak bosons, loops for gluons and dashed line for the Higgs boson. The vertices represent interaction points and are a representation of the Lagrangian interaction terms. The strength of those interaction is proportional to the coupling constant of the interaction at work for a particular vertex.

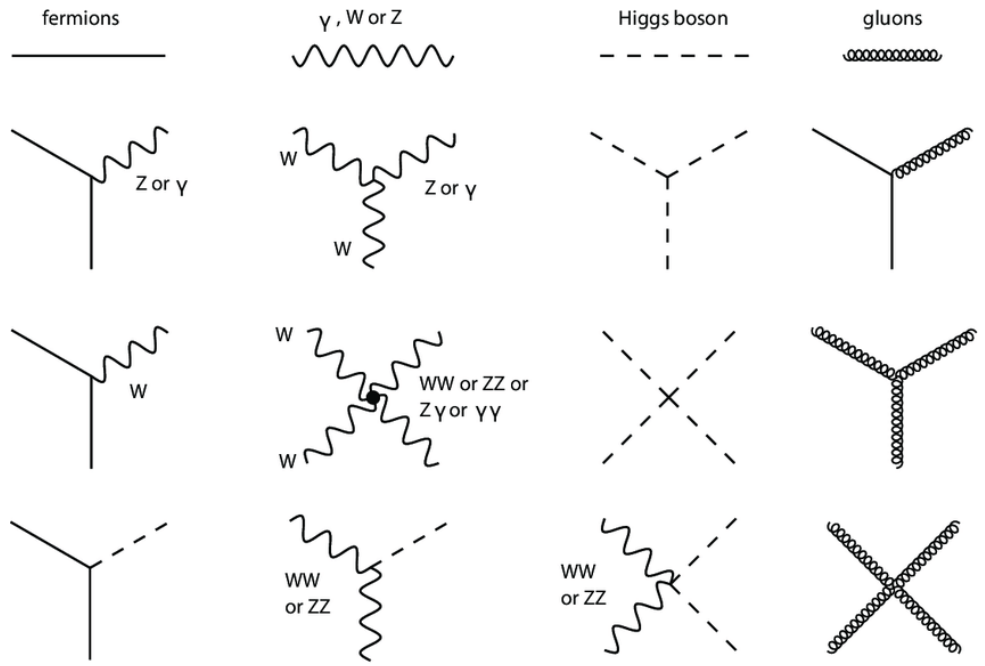


Figure 1.2: Feynman diagrams for the different interactions allowed in the SM. Taken from Ref. [21]

.

The Feynman rules allow the navigation between the diagram representations to the formal mathematical expression of the matrix elements, allowing the computation of the probability amplitude. This amplitude translates directly into the rate of production of a given final states for a particular initial state.

1.3.2 Electroweak diboson production

In the SM, the electroweak (EW) bosons can exhibit two transverse polarizations as all spin-1 massless particles:

$$\epsilon^{\mu}_{\pm} = \mp\frac{1}{\sqrt{2}}(0, 1, \pm i, 0), \tag{1.30}$$

but also a longitudinal polarization due to the mass acquired from EWSB:

$$\epsilon^{\mu}_{L} = \frac{1}{m}(p_z, 0, 0, E). \tag{1.31}$$

While the transverse polarizations are constant terms, the longitudinal contribution scales as E/m. From this fact follows that at high energies, the relative contribution of the longitudinal polarization increases and will dominate the transverse components. Without bounds on this behavior, the cross-section would grow to infinity and violate unitarity at an energy scale of around 1.2 TeV [22].

At the high-energy limit, the longitudinal vector bosons scattering amplitude can be written as

$$\mathcal{A} \approx -i\frac{m_H^2}{v^2}\left[2 + \frac{m_H^2}{s - m_H^2} + \frac{m_H^2}{t - m_H^2}\right], \tag{1.32}$$

where s and t are the Mandelstam variables. When $s, t >> m_H$, the amplitude becomes a constant and the cross-section decreases linearly with the scattering energy, $\sigma \propto 1/s$. The unitarity-breaking behavior is removed by cancellation between the Higgs diagrams and the Goldstone diagrams. For the unitarity to be conserved, the existence of a Higgs boson is necessary and its mass can not be high. The discovery of such a boson with $m_H = 125$ GeV provides a satisfying explanation of the absence of unitarity-violation.

This cancellation depends on the strength of the couplings between the Higgs boson and the vector bosons HVV, and as such the precise measurement of the scattering provides a complementary measure of the value of those couplings in addition to the direct measurement of the Higgs production. Deviation from the coupling strength predicted by the standard model could provide critical information on BSM effects. While the Higgs measurements require an on-shell boson to be produced, diboson processes allow the probing of higher energy scales. The VBS processes also provide a direct access to the quadrilinear couplings between vector bosons described in the SM Lagrangian: $WWZZ, WWZ\gamma, WW\gamma\gamma$ and $WWWW$ (no quartic coupling between only neutral vector bosons is predicted). The measurement of these couplings can put strong constraints on dimension-8 EFT operators, with potential anomalous couplings hinting towards higher energy scale effects.

1.3.3 Vector Boson Scattering at the LHC

At the LHC, the VBS processes stem from a quark from each incoming proton radiating a vector boson that interacts together. The top row of Fig. 1.3 shows the leading order Feynman diagram corresponding to the $pp \to ZV \to \ell\ell jjjj$ process. The bottom row details several processes contributing to this interaction, highlighting cubic and quartic vector boson couplings (on the left and center) and trilinear couplings with the Higgs boson (on the right).

The VBS diagram includes six electroweak vertices when considering the two additional jets and are thus proportional to the order six in the weak coupling constant (α_{EW}^6). Diagrams involving QCD vertices $\mathcal{O}(\alpha_s^2\alpha^4)$ can produce a similar final state. The key to distinguishing VBS events is the presence of two *VBS* or *tag* jets originating from the initial quarks. The scattering amplitude is proportional to

$$|\mathcal{A}|^2 \propto \frac{p_1 \cdot p_2 p_3 \cdot p_4}{(q_1^2 - M_V^2)^2 (q_2^2 - M_Z^2)^2}, \tag{1.33}$$

where $p_{1,2}$ are the incoming quark quadrimomenta, $p_{3,4}$ the momenta of the outgoing quarks and $q_{1,2}$ the momenta of the vector bosons. At a given scattering energy $\sqrt{s} = \sqrt{p_1 \cdot p_2}$, the amplitude increases when the boson momentum becomes small. This momentum q_1 can be expressed as a function of the scattering angle θ and the incoming and outgoing energies E_1 and E_3 (q_2 is expressed similarly as a function of the two others quarks):

$$q_1^2 = -2p_1 \cdot p_3 = -2E_1E_3(1 - \cos\theta_1) = -\frac{2}{1 + \cos\theta_1}\frac{E_1}{E_3}p_{T,3}^2 \tag{1.34}$$

This expression is minimal when the scattering angle is small or when the outgoing quark transverse momentum is small. The recoil of the quark against the vector boson they radiate

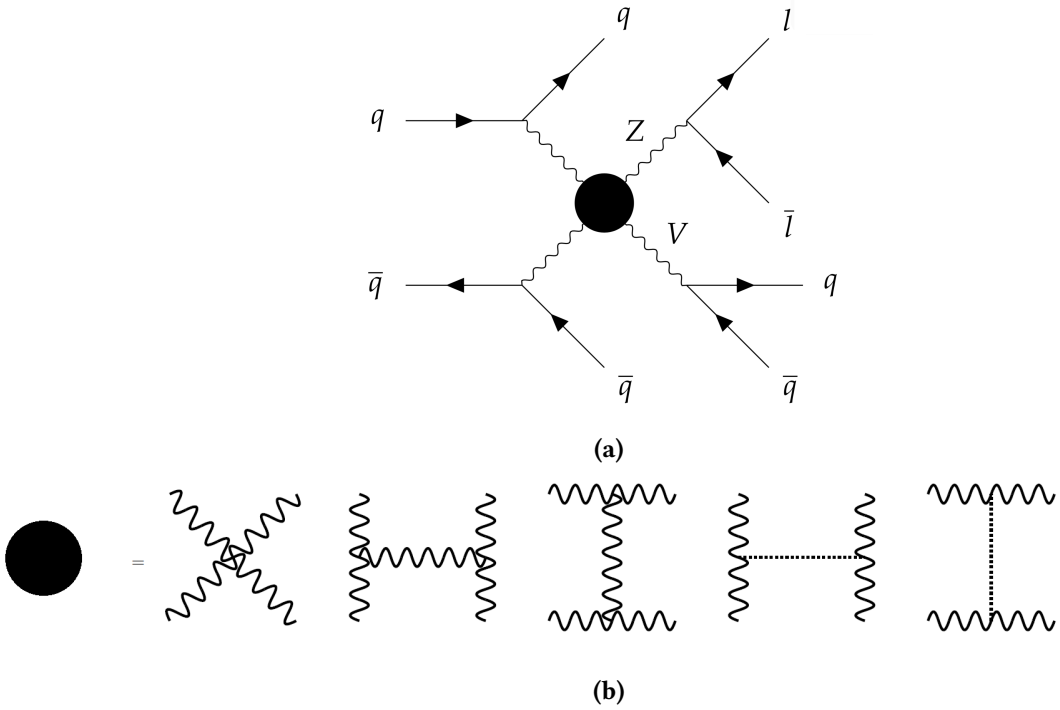


Figure 1.3: (a) : The Feynman diagram of the VBS process with an hadronically decaying gauge boson V and a leptonically decaying Z. (b) : The possible VBS interaction diagrams. The dashed line represent a H boson.

means that nonetheless the outgoing quark must possess sufficient energy to produce an on-shell V boson and so $p_{\mathrm{T}} \approx M_V$. The amplitude expressed in Eq. 1.33 also increases when the numerator expression $p_3 \cdot p_4$ becomes large. This corresponds to the mass of the dijet system composed of the two tag jets $j_{1,2}$:

$$m_{jj}^2 = 2{p_{\mathrm{T}}}^{j_1}{p_{\mathrm{T}}}^{j_2}\big(\cosh(\eta_{j_1} - \eta_{j_2}) - \cos(\phi_{j_1} - \phi_{j_2})\big). \tag{1.35}$$

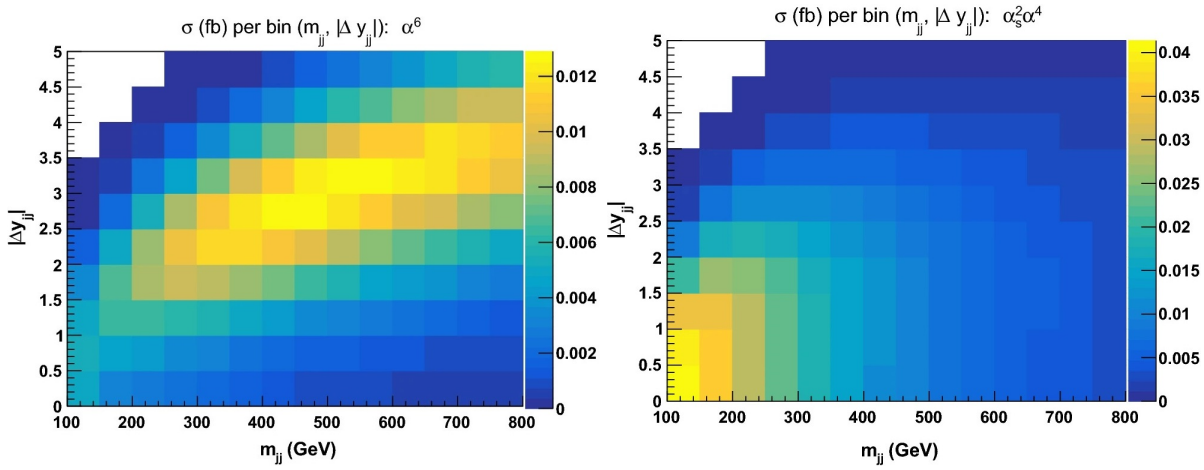
where the η are the pseudorapidities of the jets and ϕ their azimuthal angles. This expression, and thus the VBS amplitude, is the largest when the pseudorapidity gap between the tag jets is large and they are produced back-to-back $(\phi_{j_1} - \phi_{j_2}) \approx \pi$.

The VBS signature of high dijet mass and η separation is illustrated in Fig. 1.4 which shows the two-dimensional distributions of the VBS processes and of the QCD production mimicking the VBS final state. Cuts on those two variables can be performed to isolate the VBS signature and suppress the QCD background contribution.

The centrality of the gauge bosons emission is also a feature of VBS processes, as they tend to be inside the pseudorapidity gap of the tag jets. The Zeppenfeld [24] variable can quantify this centrality. It is defined as:

$$Z_X = \frac{\eta_X - \bar{\eta}_{VBS}}{\Delta\eta_{jj}} \tag{1.36}$$

where $\bar{\eta}_{VBS}$ the average pseudorapidity of the VBS tag jets and $\Delta\eta_{jj}$ the pseudorapidity gap. This variable can be used to discriminate the VBS production from the QCD production of a Z boson associated with jets, as illustrated in Fig. 1.5.



(a) (b)

Figure 1.4: Two-dimensional distributions in the dijet mass m_{jj} and rapidity separation $|\Delta y_{jj}|$ for (a) the VBS $\mathcal{O}(\alpha^6)$ process and (b) the QCD $\mathcal{O}(\alpha_s^2\alpha^4)$ process producing the same final state. The VBS events exhibit a distinctive high dijet mass and high rapidity separation signature used to define a VBS enriched fiducial space. Figure taken from Ref. [23].

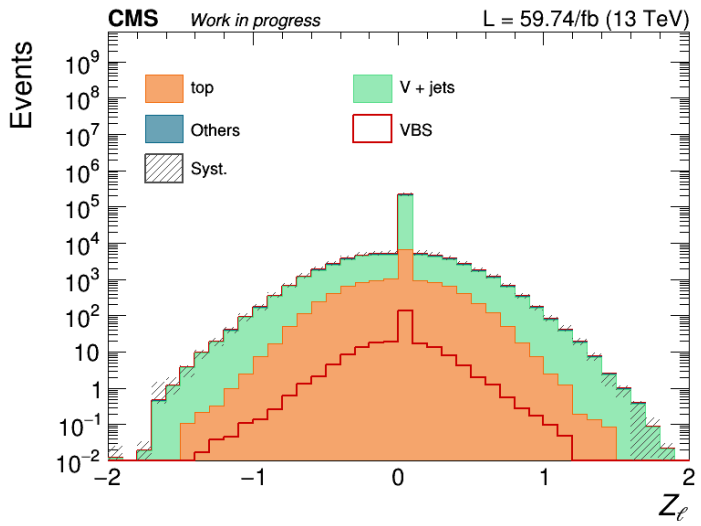


Figure 1.5: Distribution of the leptons Zeppenfeld variable in simulated data scaled to the 2018 luminosity after VBS ZV semileptonic preselections have been applied, requiring two leptons, and at least two jets forming a pair of invariant mass $m_{jj} > 500$ GeV and pseudorapidity separation $\Delta\eta_{jj} > 2.5$. The VBS production appears to be more central ($Z_\ell = 0$) than the QCD production of a Z boson associated with jets.

The VBS processes can be classified according to the decay modes of the vector bosons. The hadronic channel has the highest cross-section but suffers from large background due to the hadronic activity, and no LHC analysis has yet be able to employ this final state. Despite the low branching ratios of $W^{\pm} \to \ell^{\pm}\nu_{\ell}$ of 20% and $Z \to \ell^{+}\ell^{-}$ of 6.7%, the leptonic final states are the cleanest to reconstruct and are the source of the first VBS observations. The semi leptonic channel, where one boson decays to two leptons and the other to jets is a middle ground between leptonic and hadronic channels. While the branching ratio is higher than the leptonic channel due to the hadronic decay of one V boson, it suffers from overwhelming background originating in the production of a single V boson associated with jets. The analysis of semi-leptonic channels usually address mixed W and Z final states because the resolution on the hadronic boson mass is not enough to discriminate between the W and Z bosons masses. The measurement of the very boosted jets produced by the hadronic decay allow the exploration of a phase space very sensitive to EFT effects, hence semi-leptonic final states yield the most stringent limits on Wilson coefficients.

To summarize, in addition to the decay products from the vector bosons, the VBS signature in proton-proton colliders is the presence of two back-to-back tag jets, with high dijet mass and pseudorapidity gap; and the central emission of the gauge bosons characterized by the Zeppenfeld variables.

1.3.4 Status of experimental VBS searches

The search for VBS processes is one of the main focus of the physics program for both the ATLAS and CMS experiments. As such, many experimental results have been published using the data collected during Run 2 of the LHC. The first observation of several channels have been claimed by the two experiments and for the cleanest channels like $W^{\pm}W^{\pm}$ precision measurements are being conducted to assess the compatibility with the SM or BSM interpretations such as EFT. A recent review (2021) of the status of VBS experimental searches is presented in Ref. [25].

VBS $W^{\pm}W^{\pm}$ fully leptonic channel. The VBS same sign WW production in the leptonic channel is considered as the *golden channel* for VBS, with the EW contribution can become much larger than the QCD one after preselection. This process has been observed at both the CMS and ATLAS experiments using a partial Run 2 dataset and the analyses are now trying to measure the cross-section and compare it to the available NLO QCD and EW calculations. ATLAS published a search for this channel using a partial dataset [26] while CMS performed a differential cross section measurement using the full Run 2 dataset in combination with the $W^{\pm}Z$ channel [27]. Both measurements are in agreement with the SM predictions, though limited by the available statistics. CMS also published a measurement of the W boson polarization in this final state [28] using the same dataset, putting limits on the longitudinal scattering $W_L^{\pm}W_L^{\pm} < 1.17$ fb at 95% confidence level, not in tension with SM predictions.

$W^{\pm}Z$ fully leptonic channel. While the WZ channel has a larger cross section than same sign WW, the lower cross section for the leptonic Z decay causes the cross section for this channel to be comparable to the precedent. It is also affected by larger background from WZ QCD production, resulting in a more challenging observation. CMS performed the analysis in

combination with the WW final state [27] while ATLAS used multivariate techniques to extract signal from the full run 2 dataset [29]. Both observed the decay channel with a significance over 5 σ and reported higher cross section than expected from SM calculations.

ZZ fully leptonic channel. The ZZ channel has the smallest leptonic cross section and is amongst the rarest observed processes at LHC. Its signature is very clean however, with the full reconstruction of all final states particle possible. The first observation have been claimed by the ATLAS collaboration [30] with more than 5 σ, with CMS only providing strong evidence of the channel at 4 σ [31], but strongly constrained some EFT operators (mainly T8 and T9).

W^+W^- fully leptonic channel. The same sign WW channel is dominated by top pair background with dileptonic decay requiring a highly efficient b-jet tagger to isolate the signal. CMS has recently published a first observation [32] at above 5 σ with a cross section consistent with the SM predictions.

WV and ZV semi leptonic final channels. The semileptonic channels have larger branching ratio than the fully leptonic ones but suffer from higher background contamination from QCD production of single vector boson associated with jets. The generation of events with two leptons and four jets is extremely complex and only calculations at LO are available. Both ATLAS and CMS published analysis of a partial run 2 dataset, reaching a significance of 2.7 σ with 36 fb^{-1} for ATLAS [33] for all semi leptonic final states combined, while CMS [34] obtained competitive limits on anomalous quartic gauge couplings. The first evidence of the WV semi leptonic as been reported by CMS [35] with a significance of 4.4 σ. The search for the ZV semi leptonic channel is the focus of the work presented in this thesis, and a future combination of the ZV and WV semi leptonic channels using the CMS full run 2 dataset is expected.

Fully hadronic channels. The fully hadronic channels are currently under investigation by the two collaboration but no results have been published. It is expected they could have increased sensitivities to EFT operators, but their search present important challenges, both for the generation of MC and the complex signal extraction.

Final states including a photon.

The VBS production of $W^{\pm}\gamma jj$ and $Z\gamma jj$ is not directly connected to the EWSB mechanism but can be useful for verifying SM calculations. It shares many features with heavy gauge bosons gauge scattering, and can thus also be interesting for BSM searches. Only the leptonic decays of the W or Z are considered, and the fiducial cross sections are particularly large due to the direct production of the high energy photon. ATLAS published an evidence of VBS $Z\gamma$ cross section [36] at 4.1 σ consistent with SM predictions using a partial Run 2 dataset, while CMS observed the $Z\gamma$ channel [37] at 9.4 σ significance with good agreement to the SM predictions. The CMS experiment observed the $W\gamma$ channel [38] at 5.3 σ, with a fiducial cross section in agreement with SM.

Chapter 2

The physicists tools

2.1 Introduction

From the scattering experiments of the early 1900s such as Rutherford's, it was clear that a mean to accelerate particles at higher energies was needed to probe the nucleus structure. This lead to the creation of first particle accelerators, which would became high energy physicists' favoured tool to unravel the universe at its smallest scale. The first accelerators used simple static high voltage to accelerate charged particles through a vacuum tube. This electrostatic technology allowed J. Cockcroft and E. Walton to split Lithium into Helium atoms in 1932 [39]. At the same time, E. Lawrence invented the cyclotron [40] to accelerate charged particles, this time in circular orbits, by using a large dipole magnet to provide a constant magnetic field. This technology, that would become the basis to recent accelerators, suffered from limitations in the maximal energy it could provide. Due to relativistic effects, the accelerated particles would acquire higher effective masses and fall out of synchronisation with the electric field. Protons could therefore only be accelerated up to 15 MeV. Synchrocyclotrons and later synchrotrons would be developed to mitigate this issue and increase the available energies. In synchrotrons, bunches of particles are accelerated in a ring of constant radius and bent by a magnetic field that increases during the particle acceleration. The most powerful accelerators available to date are synchrotron type, the largest of the kind being the Large Hadron Collider (LHC) at the European Organization for Nuclear Research (CERN) in Geneva.

The CERN, from the french *Conseil Européen pour la Recherche Nucléaire*, was founded in 1954 at the Franco-Swiss border. It is one of the most important particle physics institute in the world, with more than 10000 researchers from many different countries, and hosts the largest particle accelerator to date, the LHC. In this collider, two beams of protons or heavy ions are collided up to a design center-of-mass energy of 14 TeV. The particles run in a beam pipe housed in a circular underground tunnel of 27 km of circumference, making it the largest facility of the kind. It was built between 1998 and 2008 in the same tunnel that was previously used by its predecessor, the Large Electron-Positron (LEP) collider. The two vacuum beam pipes are located in this tunnel at depths ranging from 45 m to 170 m and the beams are made to collide at four interaction points (IP), where the four main LHC experiments are located to

record the collisions. Each of these experiments relies on a detector built for a specific physics program. The LHCb detector is dedicated to the study of b-quarks physics while the ALICE (A Large Ion Collider Experiment) investigates heavy ions physics such as quark-gluon plasma. The two biggest experiments, ATLAS (A Toroidal LHC Apparatus) and CMS (Compact Muon Solenoid) are general-purpose detectors, originally designed to search for the Higgs boson, a feat achieved successfully in 2012, and to probe physics Beyond Standard Model at the TeV scale.

This thesis presents the analysis of the data recorded by CMS in proton-proton collisions at a center of mass $\sqrt{s} = 13$ TeV from 2016 to 2018 in the so-called Run 2. Studies for the development of Machine Learning algorithms for the future trigger system of the CMS experiment are also discussed.

In this chapter, an overview of the LHC facility, with specifics on its design, operation and the main experiments is given in Sec. 2.2. The CMS detector design, as well as details about its data acquisition and trigger systems, and event reconstruction is discussed in Sec 2.3. The extremely high amount of complex data recorded by the detectors makes it necessary to use sophisticated analysis strategies. In this thesis, multiple Machine Learning (ML) techniques were employed for such tasks, and are discussed in Sec. 2.6.

2.2 The Large Hadron Collider at CERN

The LHC was designed to perform proton-proton (pp) collisions at a nominal center-of-mass energy of $\sqrt{s} = 14$ TeV, with an instantaneous luminosity of $\mathcal{L} = 10^{34}$ cm^{-2}s^{-1} [41]. The design was driven by the intention to probe the EWSB and search for the Higgs boson. Another objective of the LHC was to explore BSM scenarios at the TeV scale, searching for discrepancies with the SM and testing postulated scenarios. In addition to the pp runs, an heavy ion program based on lead-lead (Pb-Pb) and proton-lead collisions is carried out to study QCD, and in particular the behavior of quark and gluons in their deconfined plasma state. Since its inauguration in 2008, the LHC has already delivered two eras of data recording, the Run 1 that lasted from 2009 to 2013, and the Run 2 from 2015 to 2018. The Run 3 has recently started on the 5^{th} of July 2022 and is expected to last for around four years. At the end of this data-taking period, a shutdown of the LHC operations is foreseen during which the facilities will undergo profound upgrades to prepare for the next phase of the LHC program: the High Luminosity LHC (HL-LHC).

2.2.1 Acceleration complex

As illustrated in Fig. 2.1, the LHC is built as the last step of a chain of older accelerator facilities. The proton undergo several steps, from production, acceleration, and injection in the final LHC ring where they are made to collide at the four dedicated interaction points surrounded by the detectors. For the Run 2 of operations, the initial step of the acceleration complex was the production of protons from hydrogen by strong electric fields that ionize the atoms. A magnetic field generated with a Radio Frequency Quadrupole (RFQ) is used to form bunches of protons

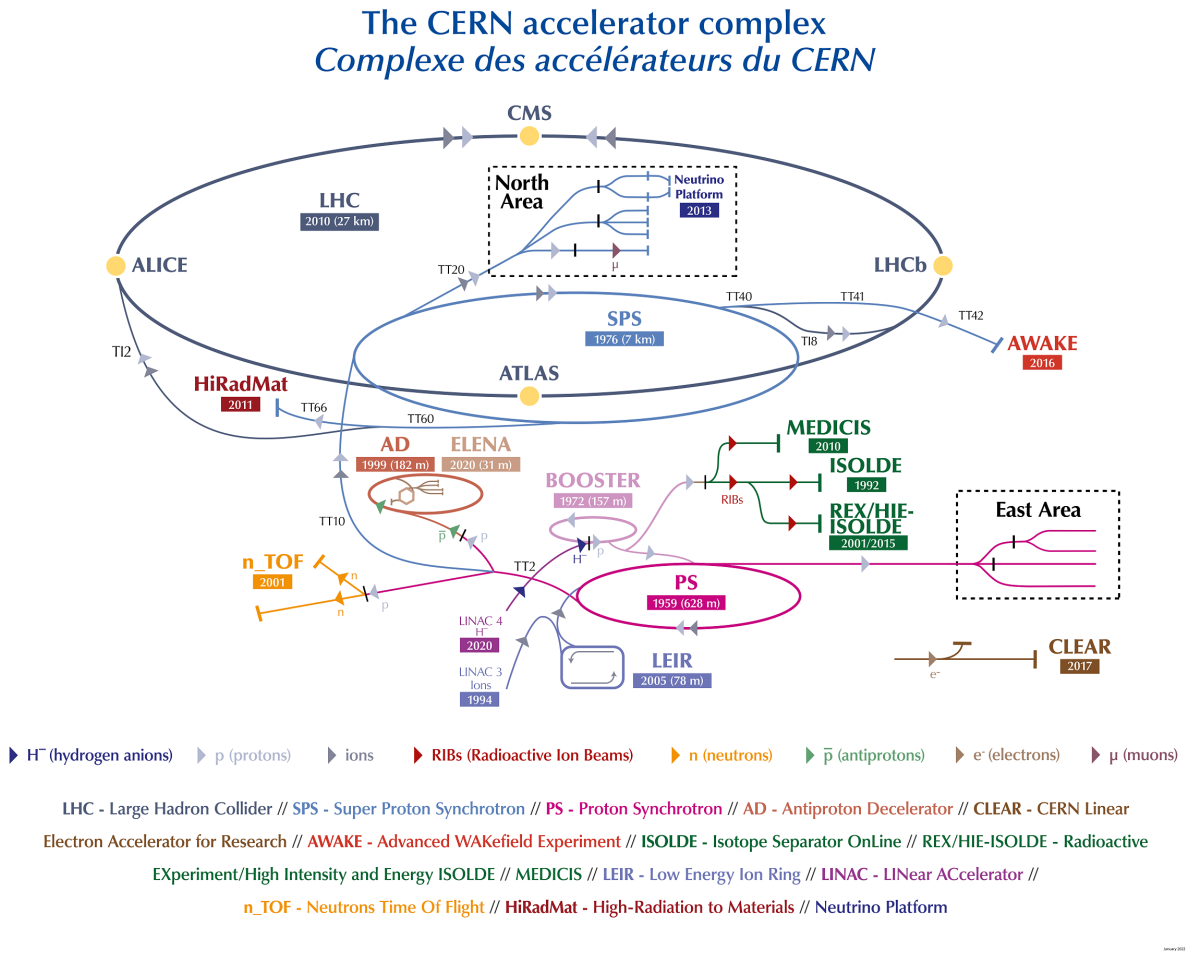


Figure 2.1: Layout of the CERN accelerator complex in 2022. Protons are produced in the LINAC 4, and then accelerated to increasing energies through the Booster, PS and SPS before being injected in the LHC for their final acceleration. The two parallel beams are collided at the four interaction points occupied by the CMS, ATLAS, LHCb and ALICE experiments. Figure taken from Ref. [42].

and accelerate them to 750 keV. The bunches are then injected in to the Linear Accelerator 2 (LINAC 2) where they reach an energy of 50 MeV before being sent to the Proton-Synchrotron Booster (PSB). In this circular accelerator of 150 m, the beam is further accelerated to reach up to 1.4 GeV. The particles are then accelerated further in two additional circular accelerators, the Proton Synchrotron (PS) with a ring of 620 m, and the Super Proton Synchrotron (SPS) of 6912 m of circumference. In those rings, they acquire up to an energy of respectively 25 and 450 GeV before being ultimately sent to the LHC ring. Since 2020, the acceleration system has changed, and the Linear accelerator 4 (LINAC 4) has become the primary source of proton beams. Hydrogen ions H^- are accelerated to 160 MeV before being sent to the PSB, losing their two electrons during the injection process. They subsequently travel through the PSB, PS and SPS before injection in the LHC.

The beam is split in two parallel beamlines by fast kicker magnets, and injected in the two beam pipes of the LHC. One of the beams circulates in a clockwise manner and the other anticlockwise. They are then accelerated by *high frequencies accelerating cavities* to reach their

maximal energy. The cavities are placed in 545 m long straight sections along the ring. The beams are bent with 1232 dipoles of superconducting NbTi magnets, disposed into eight arcs of 2.45 km in length. Those magnets need to be operated at a temperature of 1.9 K which is achieved through superfluid Helium-4 and provide a 8.3 T magnetic field. The proton bunches are kept collimated in the accelerator by quadrupoles, with additional ones used to collide the two beams at the four interaction points.

Two of these IPs are occupied by the two general purpose detectors, ATLAS and CMS, which can record both pp and Pb-Pb collisions. They are situated at opposite points of the ring where the maximum luminosity is achieved. The LHCb is a mono-directional detector dedicated to the studies of b-quarks physics, exploring charge-parity (CP) violation and rare decays. The ALICE experiment is dedicated to the study of heavy ions physics in Pb-Pb collisions, and was designed for very high particles multiplicity in order to probe the quark-gluon plasma state.

2.2.2 Design parameters and luminosity

The LHC is a hadron accelerator that mainly collides proton, which are charged and stable baryons. Proton-proton collisions suffer from the more complex collisions than could be offered by fundamental particles as electron, that were collided by its predecessor, the LEP. However, the very light electron lose a large portion of their energy in circular colliders through synchrotron radiation, radiating energy when accelerated radially. Since the intensity of this process decreases with the mass, the much heavier protons can sustain higher energies in circular orbits. At the LHC, the nominal center-of-mass energy is of $\sqrt{s} = 14$ TeV, meaning that the two countercirculating proton beams can be accelerated up to 7 TeV. The LHC performance is also characterized by the collision rate it offers and which is expressed as the *instantaneous luminosity*. The luminosity $\mathcal{L}$ is a way to measure the frequency of observation of an event $\partial N/\partial t$ for a given process with a probability σ, also known as the cross section of the process, with formula

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

Though higher instantaneous luminosity makes the observation of rare processes such as VBS possible, increases are limited by technological constraints, such as the strong pressure it places on the data acquisition systems. This instantaneous luminosity, expressed in units of $\mathrm{cm}^{-2}\mathrm{s}^{-1}$, can be integrated over time to express the amount of data recorded over a given period. This *integrated luminosity* is expressed as

$$L = \int \mathcal{L} dt, \tag{2.2}$$

usually in units of pico- or femtobarn (pb^{-1} or fb^{-1}).

The instantaneous luminosity can be computed from the beam parameters as a function of the number of particles N_p in each of the n_b bunches, the revolution frequency f_{rev} and the beam overlap area A_{eff} with the formula

$$\mathcal{L} = \frac{N_p^2 n_b f_{\text{rev}}}{A_{\text{eff}}}. \tag{2.3}$$

This overlap can be expressed, considering Gaussian profiles for the beams, as

$$A_{\text{eff}} = \frac{4\pi\beta^\star\epsilon_n}{\gamma_r F}, \tag{2.4}$$

where $\beta^\star$ is the Beta function characterizing the beam focus at the interaction point, ϵ_n is the emittance that represents the confinement of the beam and γ_r the relativistic factor. Finally, the F factor characterizes the geometry of the collision, encoding the reduction of the luminosity due to the crossing angle θ_c of the beams and the longitudinal and transverse RMS widths of the bunches (σ_z and $\sigma^\star$)

$$F = \left(1 + \frac{\theta_c\sigma_z}{2\sigma^\star}\right)^{-1/2} \tag{2.5}$$

The nominal values of those parameters are reported in Tab. 2.1.

Parameter	Description	Nominal value
$\sqrt{s}$	Center-of-mass energy	14 TeV
Δt	Time between bunches	25 ns
n_b	Number of bunches	2808
N_p	Number of protons per bunches	1.15×10^{11}
f_{rev}	Revolution frequency	11245 Hz
$\sigma^\star$	Transverse RMS for the bunches at the IP	16.7 µm
σ_z	Longitudinal RMS of the bunches	7.55 cm
$\beta^\star$	Beta function at the IP	0.55 m
θ_c	Crossing angle at the IP	285 µrad
ϵ_n	Transverse emittance	3.75 µm

Table 2.1: Nominal parameters of pp collisions at the LHC.

A high instantaneous luminosity increases the LHC sensitivity to processes with low cross sections, but it also increases the multiplicity of interactions happening in the same bunch crossing, called pileup (PU). The amount of PU is driven by the instantaneous luminosity and the cross section for inelastic pp interaction $\sigma_{pp}^{\text{inel}}$

$$\langle N_{\text{PU}} \rangle = \frac{\mathcal{L}\sigma_{pp}^{\text{inel}}}{n_b f_{\text{rev}}}. \tag{2.6}$$

At the LHC nominal values of $\sqrt{s} = 14$ TeV, the resulting average PU rate is of around 22 interactions [43]. As described in the following section, the LHC has been increasing its operational values during the years, resulting in an average PU rate of about 50 in 2018. In the HL phase of the LHC, the increased luminosity is expected to be the source of an increased PU rate in the range of 140 to 200 additional interactions per event. A high PU value results

in high detector occupancy, thus degrading the particles reconstruction's efficiency and resolution. The mitigation of this loss of performance due to PU is one of the main motivations for the new subdetectors that are foreseen to be installed during the long shutdown period before the HL phase. In particular, the CMS endcap calorimeters upgrade, the High Granularity Calorimeter (HGCAL), has been designed to be extremely radiation tolerant and able to better identify objects originating from the hard scattering event and reject the ones caused by PU, as reported in Sec. 4.3.

2.2.3 Operational history

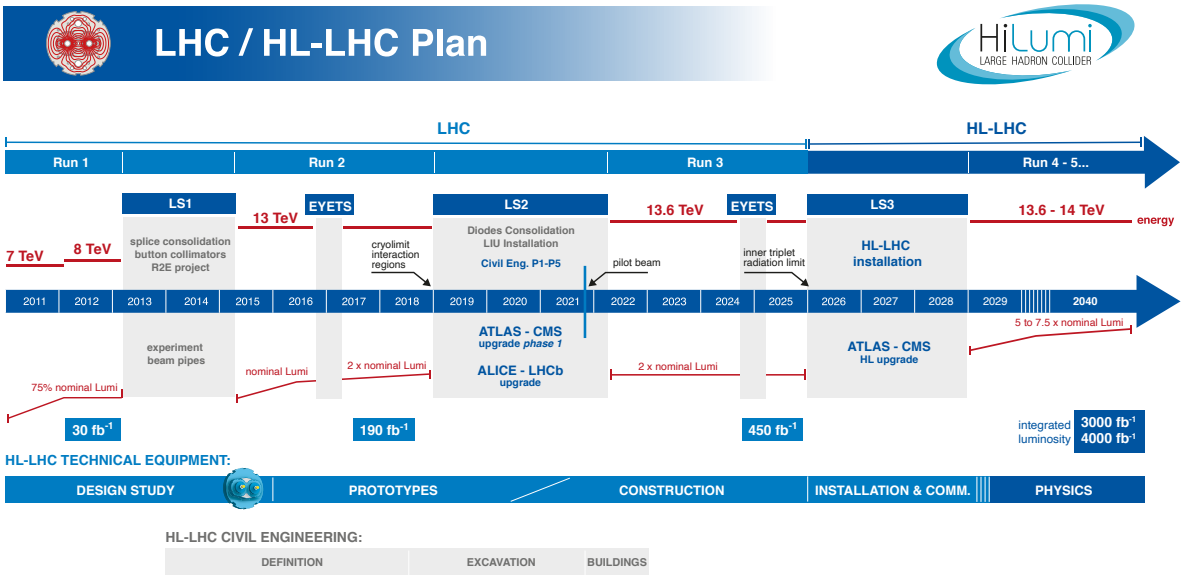
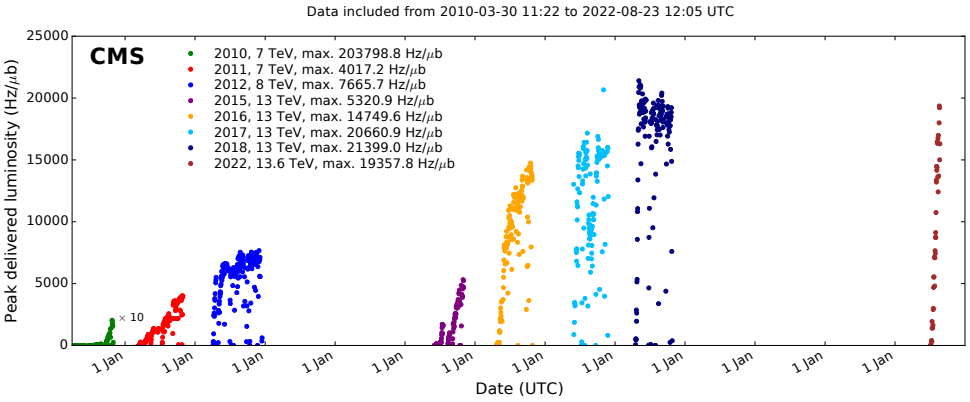


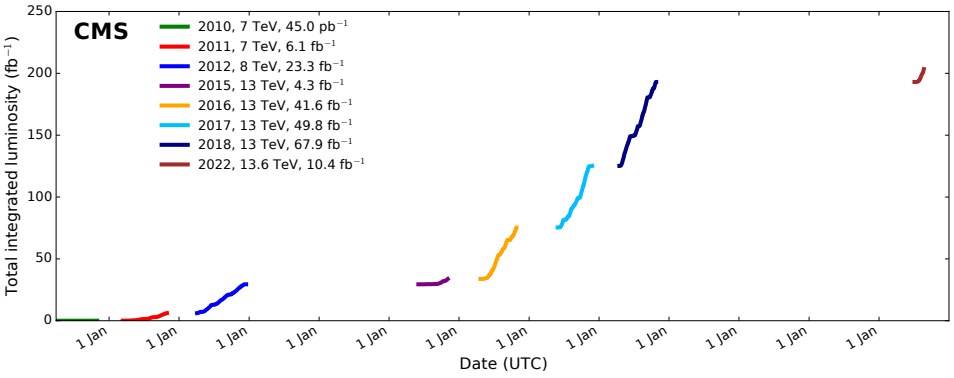
Figure 2.2: Schedule of the LHC and HL-LHC operations as of February 2022.

The timeline for the past and nominal future operational runs of the LHC is shown in Fig. 2.2. Across an operation period of around 30 years, the physics program at LHC is divided in two main operational phases: Phase I (2015-2025) and Phase II expected to record data from 2029 up to the end of the 2030's. The Phase I, currently in progress, is expected to see the center-of-mass energy gradually increase from 7 TeV to 13.6 TeV, with the instantaneous luminosity reaching twice the nominal value by the end of the phase. The luminosity achieved during the Run 1 and Run 2 are presented in Fig. 2.3 for a peak luminosity of 2.1×10^{34} cm^{-2}s^{-1} attained in 2018. The Phase II, also known as the High Luminosity phase, will keep the nominal center-of-mass energy but ramp up the instantaneous luminosity up to at least five times the design value.

After ten years of construction between 1998 and 2008, the first protons beams where injected in the LHC for an inaugural run on the 10^{th} of September 2008. The run had to be stopped shortly after however, and initial testing was delayed for more than a year due to an intervention needed to repair a magnet quench incident that damaged over 50 superconducting



(a)



(b)

Figure 2.3: The (a) peak luminosity and (b) integrated luminosity delivered per day to the CMS experiment during LHC Run 1, Run 2 and beginning of Run 3 [44].

magnets. In November 2009, 450 GeV beams were injected in the tunnel for the first time after the incident, and subsequently ramping up to 1.18 TeV in the following weeks, becoming the world's highest energy accelerator.

The energy was constantly increased during the course of 2010 reaching $\sqrt{s} = 7$ TeV in march 2010. The LHC then proceeded to record data for the so-called Run 1. By the end of 2011, 6.1 fb^{-1} of integrated luminosity had been delivered to CMS and in 2012 the center-of-mass energy was further increased to reach 8 TeV, for an integrated luminosity of 23.3 fb^{-1}. This dataset lead the ATLAS and CMS collaborations to conjointly claim the discovery of a new boson compatible with the SM Higgs boson in July 2012. The LHC was shut down early 2013 to allow for a two years upgrade and maintenance program during the so-called Long Shutdown 1 period. Many renovations and enhancements of the facilities took place during the time in order to enable collisions at a higher center-of-mass energy of $\sqrt{s} = 14$ TeV.

The Run 2 of the LHC operations started on the 5^{th} April 2015 and recorded data until the end of 2018. A record-breaking center-of-mass energy of $\sqrt{s} = 13$ TeV was reached one month after the restart, and the 2015 year was dedicated to the production the first data at such a high energy. During the course of 2016, the operations focused on increasing the

instantaneous luminosity, past the initial design values. Integrated luminosities of 4.2 fb^{-1}, 41.0 fb^{-1}, 49.8 fb^{-1} and 67.9 fb^{-1} were delivered to CMS from 2015 to 2018. The analysis of this extensive dataset, allowed to learn more about the Higgs boson, improve the precision of several measurements, and claim first observations of rare processes such as Vector Boson Scattering. This work presented in this thesis is part of this still ongoing work.

A second period of maintenance and upgrades called the Long Shutdown 2 started on the 10^{th} of December 2018. One of the main purpose of the upgrades was to prepare for the Run 3 but also the future High Luminosity phase. It ended on the 22^{nd} of April 2022 when the Run 3 started. A record collision energy of 13.6 TeV was achieved a few days later. The official data recording are expected to take place until 2026 for an integrated luminosity of around 300 fb^{-1}.

The end of Run 3 will also be the end of the LHC Phase I, and will be followed by the Long Shutdown 3. During this period, the LHC and the main experiments will undergo profound upgrades toward the High Luminosity phase, where it is expected to reach an instantaneous luminosity of 5×10^{34} $cm^{-2}s^{-1}$. After running for about a decade, this will yield an integrated luminosity of around 3000 fb^{-1}, improving significantly the sensitivity to rare phenomena and the exploration of the BSM boundary. With this record collision rate, the PU will also increase significantly, for an average number of additional interactions per event of 140. The LHC is ultimately foreseen to be pushed to 7.5×10^{34} $cm^{-2}s^{-1}$ for an average PU of 200. To face the new challenges posed by these conditions, several detectors of CMS will be enhanced or replaced, and the data acquisition system will be upgraded.

2.3 The Compact Muon Solenoid

CMS is one of the two general-purpose experiments of the LHC (the other one being ATLAS), designed to study particle production and interactions at the TeV scale. One of the main focus of CMS design was the study of the electroweak symmetry breaking, and in particular search for the Higgs boson at masses that could not have been reached by previous machines, up to a mass of 1 TeV. This was successfully achieved in 2012 with the conjoint discovery with ATLAS of a particle consistent with the SM Higgs boson at a mass of 125 GeV in July 2012. To achieve this ambitious goal, CMS and the ATLAS experiment adopted a different and complementary design philosophy. The key features of this design, as defined in Ref. [45] are:

- Very good muon identification capabilities, with high momentum resolution. The dimuon mass resolution should be of the order of $\approx 1\%$ at 100 GeV, and the muon charge should be correctly assessed up to 1 TeV momenta;
- Good performance of the inner tracker for charged particles. In particular, the efficient triggering and tagging of τ leptons and b-jets is essential, requiring good pixel detectors near the interaction point;
- Good resolution on the energy of electromagnetic objects, in particular the mass resolution of diphotons and dielectrons should be of the order of $\approx 1\%$ up to momenta of 100 GeV. The detector should possess wide acceptance, high π^0 rejection, and efficient photon and electron isolation even at high luminosities;

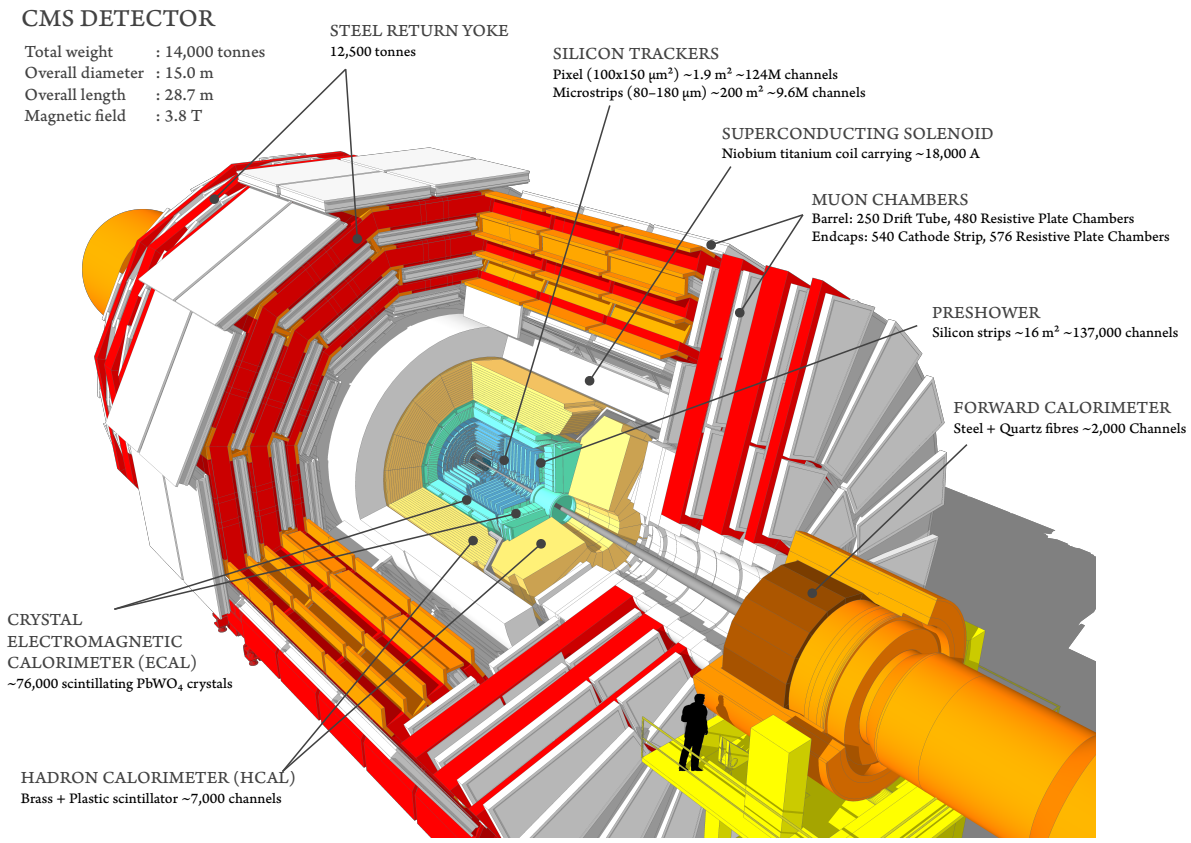


Figure 2.4: Schematic cutaway view of the CMS detector. The various subdetectors are placed in concentric fashion around the beam pipe and enclosed in a powerful magnetic field.

- The missing transverse energy (such as coming from the presence of neutrinos in the event) should be correctly identified with good resolution. The dijets mass should also be accurately measured, requiring hadron calorimeters with large geometric coverage and fine lateral segmentation.

Situated in a underground cavern 100 meters deep under the ground at the Interaction Point 5 (IP5), CMS is a cylindrical detector structured in several concentric subdetectors which work in harmony to achieve the goals previously mentioned. With a total weight of around 14000 t for only half the size of ATLAS, CMS is a *compact* detector as hinted by its acronym. At the center of CMS design is the intense 3.8 T magnetic field induced by the superconducting solenoid magnet that surrounds the subdetectors (except the muon system). This magnetic field bends the charged particles passing through the detector, allowing their interaction vertices, tracks and momenta to be precisely measured in the pixel and strip trackers placed nearest to the interaction point. Around this inner tracking system, the electromagnetic and hadronic calorimeters are designed to reconstruct electrons, photons and hadrons. Finally, the muons, that are not stopped by the inner detectors, are measured in the muon tracking systems situated on the outer edge of the detector. An overview of the different subdetectors can be found in the illustration in Fig. 2.4, and each component is described in more details in Sec. 2.3.2.

2.3.1 Coordinate system

The CMS experiments uses a right-handed coordinate system centered at the interaction point in the collider. The x-axis points towards the center of the LHC, the y-axis points opposite to the center of the earth, and the z-axis is parallel to the beam axis, with its direction given by the right-handedness of the system as the anticlockwise beam direction. A polar coordinate system is also used to provide a more appropriate description of the cylindrical structure of CMS. The azimuthal angle ϕ is defined as the angle in the (x, y) transverse plane with the x-axis, the polar angle θ is measured in the (y, z) plane with respect to the z-axis, and r is the radial coordinate. Those two coordinate systems are illustrated in Fig. 2.5.

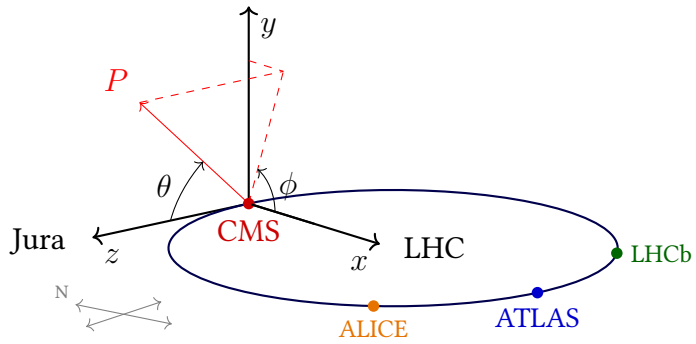


Figure 2.5: Schematic representation of the CMS coordinate systems.

To describe the pp collisions at the parton level, the geometrical coordinates of the experiment are not suitable. The momentum fraction carried by each constituent of the proton along the beam axis and the boost in the experiment rest-frame are unknown. It is thus preferable to express the events in terms of Lorentz boost-invariant coordinates. Quantities measured as projections on the transverse plane, such as the transverse momentum p_{T} and transverse mass m_{T}, are introduced to describe particles kinematics as

$$\begin{aligned} p_{\mathrm{T}}^2 =& p_x^2 + p_y^2 \\ m_{\mathrm{T}}^2 =& m^2 + p_x^2 + p_y^2 = E^2 - p_z^2 \end{aligned} \tag{2.7}$$

Additionally, the pseudorapidity of a particle is defined as

$$\eta = -\ln(\tan(\theta/2)), \tag{2.8}$$

which replaces the polar angle for ultra relativistic particles. When a particle is emitted perpendicular to the beam axis, its pseudorapidity is equal to 0, and it goes to infinity when the emission is parallel to the beam line. The forward regions of the detectors are defined as the regions with high η. In CMS, they are covered by the endcap calorimeters, which will be replaced by the HGCAL for the LHC Phase II, and are of great significance for the study of VBS events that exhibit highly boosted jets. Another important quantity is the angular distance between two particles

$$\Delta R^2 = \Delta\phi^2 + \Delta\eta^2, \tag{2.9}$$

which is used to characterize the isolation between two particles or jets.

2.3.2 Subdetectors

The CMS detector is structured in multiple subdetectors organized in a concentric fashion, immersed in the magnetic field produced by the superconducting solenoid magnet that bends the charged particles through the detector. Each layer of subdetector employs different technologies to measure some aspect of the particles issued from the collisions. Starting from the component closest to the interaction point, those subdetectors are:

- The **Inner tracking system** that reconstructs the tracks left by the charged particles near the primary vertex (PV) of interaction. It is subdivided into an inner pixel tracker and an outer strip tracker, with the highest granularity closest to the IP;
- The **calorimeters**, both electromagnetic, where electrons and photons deposit their energy, and hadronic, which is primarily used to measure the energy of hadronic showers such as those originating in jets;
- The **superconducting magnet** that provides an 3.8 T magnetic field to the whole detector, allowing properties from charged particles to be infered from the curvature of their trajectories;
- The **muon chambers**, situated at the outer end of the detector to measure the energy and momentum of the muons. The muon chambers use multiple technologies depending on the η region covered, with drift tubes, cathode strip chambers and resistive plate chambers.

Superconducting magnet

The defining component of CMS is the niobium-titanium (NbTi) superconducting solenoid magnet [46]. Providing a constant magnetic field of about 3.8 T, it encloses the inner tracker and the calorimeters so that the properties of charged particles can be inferred from the curved trajectories they follow under the effect of the magnetic field. It is a cylindrical coil of superconducting cable with a diameter of 6 m and a length of 12.5 m that requires a very low temperature of 4.5 K to be operated. The magnetic field is confined with a steel return yoke that, with a weight of 12500 t, accounts for most of the detector's mass. The muon chambers are embedded in the yoke and are only immersed in a 2 T field. While possessing a nominal capacity of 4 T, CMS operates the magnet at a lower value of 3.8 T, keeping a safety margin notably for the unknown aging effects it could exhibit. The magnitude of the magnetic field through the CMS detector is illustrated in Fig. 2.6.

When immersed in a magnetic field $\vec{B}$, particles of charge q and velocity $\vec{v}$ are submitted to a Lorentz force

$$\vec{F} = q\vec{B} \times \vec{v}. \tag{2.10}$$

In the case of the solenoid of CMS, as the magnetic field is aligned with the z-axis, the force is only applied in the transverse direction, causing the charged particles to travel in an helical trajectory of radius $R = p_{\mathrm{T}}|q|B$. The transverse momentum can thus be inferred from the measure of the bending radius of a particle of known charge. It also appears from Eq. 2.10 that

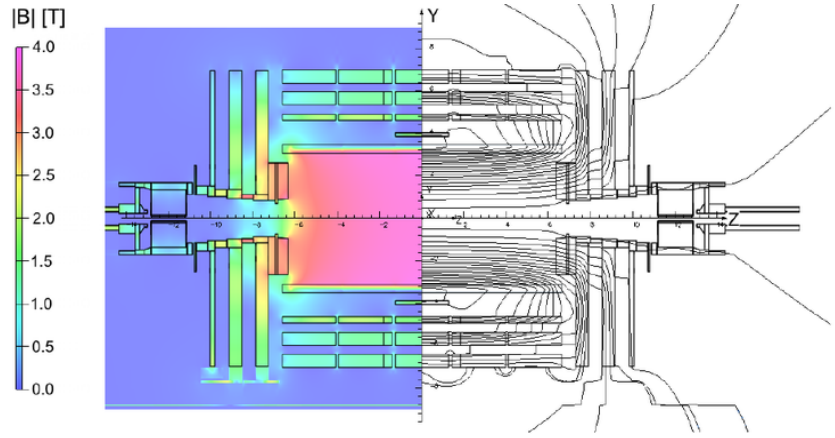


Figure 2.6: Value of the predicted magnitude of the magnetic field |B| (left) and field lines (right) alongside a longitudinal view of the CMS detector. Figure taken from Ref. [47].

a strong magnet is critical for accurate momentum measurement. With a bending power BR of nearly 12 T.m, the magnetic field of CMS is instrumental to achieve the required very high resolution.

Silicon tracking system

The innermost subdetector of the CMS experiment is the inner tracking system, located directly around the interaction point. It is composed of silicon detectors, encompassed in the 3.8 T field, arranged in a cylindrical shape of 5.8 m length and 2.5 m diameter and cooled to around -20°C to withstand the high level of radiations. The high pseudorapidity forward regions up to $|\eta| < 2.5$, called the *endcaps*, are covered with disks of silicon sensors. To maximize the granularity near the interaction point, the innermost part of the tracker detectors cells consists in silicon pixel detectors, while the outer parts are covered with silicon strips, as illustrated in Fig. 2.7.

The tracking system's role is to measure the helical paths of the charged particles coming from the interaction point and that are bent by the magnetic field. When a charged particle interacts with one of the silicon modules, an ionization current is produced and can be measured to accurately infer the position of the *hit*. From the very high spatial resolution achieved by the 200 m^2 of active silicon area, the primary vertex (PV) can be identified and additional PU interactions can be discriminated. Short-lived particles such as τ leptons and heavy quarks can also be identified through their characteristics displaced vertices a few millimeters away from the PV.

The CMS pixel tracker surrounding the IP is made of 65 million silicon pixel of 100×150 μm^2. This technology is able to provide accurate spatial and time resolution even under the very high particle flux the regions is subjected to. The pixel tracker is able to provide precise identification of the PV and serves as a basis for the reconstruction algorithms described in Sec. 2.5. Originally composed of three layers in the barrel and two layers in the disks, it was upgraded in March 2017 to cope with the improved instantaneous luminosity [48, 49]. In the new configuration, the pixel tracker features an additional layer in both the barrel and the

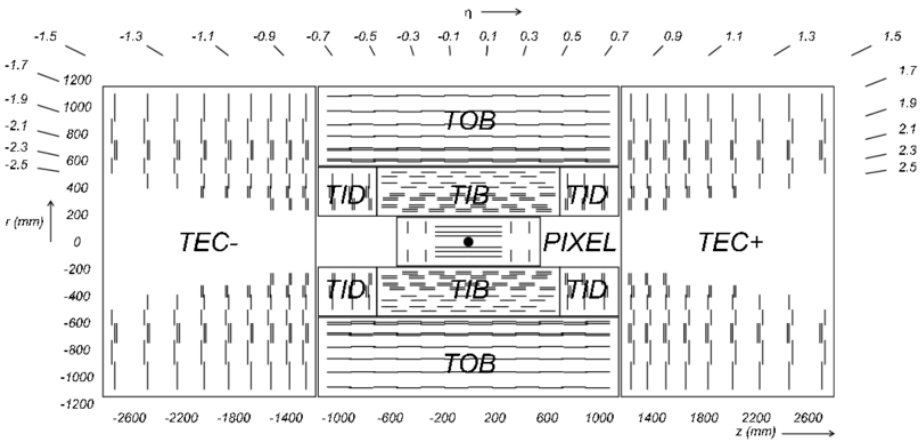


Figure 2.7: Schematic view of the CMS tracking system in the (r, z) plane. The pixel detector is located around the interaction point, and surrounded by the tracker inner (TIB) and outer (TOB) barrels. The forward regions are covered with the tracker inner disk (TID) and the tracker endcap (TEC) around the beam pipe. Figure taken from Ref. [45].

forward region, as well as faster readout electronics. The inner part of the tracker was also moved closer to the IP and the outer part moved further. With an upgraded total of 125 million silicon pixels, the pixel tracker achieves a tracking efficiency of 99.95% and a spatial resolution of 5-10 μm [50].

The second part of the inner tracking system is covered with silicon strips sensors. The lower flux of particles compared to the innermost region made the usage of micro-strips detectors possible. This portion of the tracker is composed of around 9.3 million silicon strips arranged in a way that depends on the distance to the IP. The size of the strips is inversely proportional to the flux of particles expected in the region, with a size ranging from 205 μm in the innermost part to 20 cm in the outer part. The barrel is subdivided into an inner part, the TIB, covering the $20 < r < 55$ cm region and an outer part, the TOB, up to 116 cm. The forward region is similarly divided in the TID in the $50 < |z| < 124$ cm region and the TEC up to $|z| < 282$ cm. The strip tracker possesses a hit efficiency of 99.8%[51] for a transverse spatial resolution of 23-34 μm in the TIB and of 35-52 μm in the TEC, and a ten times larger longitudinal resolution.

The CMS inner tracker design is the result of a trade-of between the best tracking performance and the smallest possible amount of inactive material. The minimization of inactive material is critical as it limits the amount of unwanted interaction that would degrade the detector performance. In the original design, the total passive material in the tracker was of 1.6 interaction lengths, and the pixel tracker upgrade reduced this value by 40% in the forward regions and 10% in the barrel.

Electromagnetic calorimeter

Surrounding the tracking system is the Electromagnetic CALorimeter (ECAL), whose primary role is to provide an accurate measurement of the electrons and photons energies. The ECAL is composed of ≈ 70000 lead-tungstate ($PbWO_4$) crystals possessing high density ($\rho = 8.29$ g.cm^{-3}) . It is highly transparent, with a small radiation length of $X_0 = 0.89$ cm, and scintillates when interacting with particles. The short Molière radius of the material ($R_M \approx 2.2$ cm) is critical to ensure that electromagnetic showers are contained in the crystals. This technology allows for a fast response of the detector, as required to contend with the passing of a bunch crossing every 25 ns. Because of the low light-yield of the ECAL ($\sim 30\gamma/$ MeV), the usage of photodetectors with internal amplification factors was required, and silicon avalanche photodiodes and vacuum phototriodes are used to that end in the barrel and endcaps respectively.

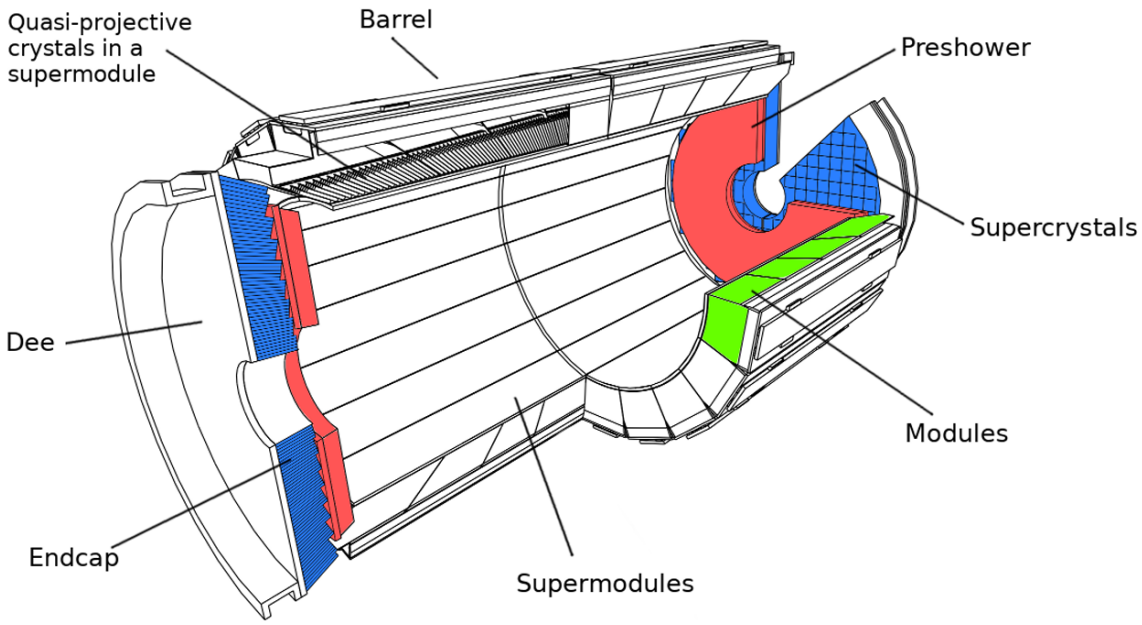


Figure 2.8: Schematic illustration of CMS ECAL. The barrel (green) is made of 36 super-modules. The endcaps (blue) are placed after a preshower module (red) and are composed of two half disks (Dees) on each side.

As illustrated on Fig. 2.8, the ECAL is subdivided into a barrel (EB) consisting in two half cylinders, and the two endcap disks (EE) covering the forward regions. Each half of the EB is made is 18 super-modules, 20 crystals wide in ϕ and 85 crystals wide in η for a total of around 1700 crystals and 1.5 t. The barrels cover the $|\eta| < 1.479$ region for a total of 62000 crystals of 22×22 mm^2 surface. They are completed by the EE disks covering the $1.479 < |\eta| < 2.6$ region with groups of 5×5 crystals known as *supercrystals*. The ECAL is not fully hermetic and gaps exist mainly at the transition between the barrels and endcap and in the $\eta = 0$ region.

The ECAL is complemented by an electromagnetic preshower detector (ES), placed in front of the endcaps in the $1.65 < |\eta| < 2.6$ region. The ES is a sampling calorimeter composed of

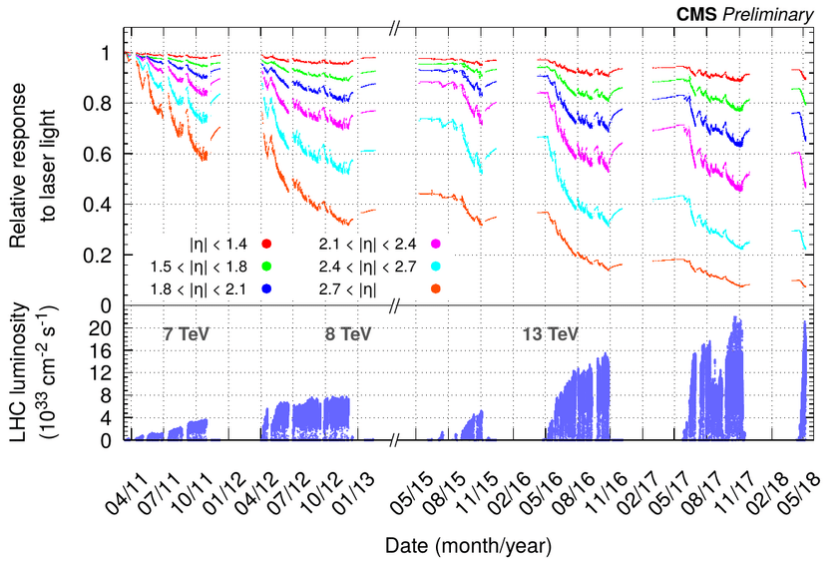


Figure 2.9: The relative response of the ECAL to laser light in bins of pseudorapidity (top) and the instantaneous luminosity of the LHC (bottom). The response of the detector is decreasing due to the loss of transparency caused by the increasing luminosity, in particular in the forward high η regions. Figure taken from Ref. [52].

two layers of lead absorber followed by a layer of 22 mm silicon strips operated between -10°C and -15°C. As was the case for the tracker, the choice of silicon for the active material was motivated by the higher radiation delivered to this part of the calorimeter. The addition of the 1 X_0 depth of the ES material to the 2 X_0 of the tracker ensures that most of the incident photons will start showering inside the sensors. This additional layer also helps the discrimination of diphotons originating from $\pi^0 \rightarrow \gamma\gamma$ decay from the single high energy incident photons. While around 6% to 8% of electromagnetic showers energy is deposited in the ES, it suffers from high parasitic noise originating in the pions decays in the tracker material and is not widely used for reconstruction.

The measurement of the particle energy in the calorimeter is based on the proportionality relation between the light emitted by the crystal and the energy deposited by the incident particle. For the ECAL, the intrinsic energy resolution σ_E has been measured in electron test beams studies [53] as a function of the energy E to be

$$\left(\frac{\sigma_E}{E}\right)^2 = \left(\frac{2.8\%}{\sqrt{E(GeV)}}\right)^2 + \left(\frac{12\%}{E(GeV)}\right)^2 + (0.3\%)^2. \tag{2.11}$$

The first term accounts for the event-per-event stochastic fluctuations in the measurement of the energy, and is small due to the homogeneous nature of the ECAL, which allows for precise energy measurement since all the energy is deposited in active material. The second term is coming from the noise due to the electronics in the data acquisition chain and to the pileup, and becomes dominant at low energy due to the dependency to the shower energy. The last term is constant represents the non-uniformities in the response, as well as miscalibrations and energy leakage.

When submitted to high doses of radiation, the ECAL crystals lose transparency, in partic-

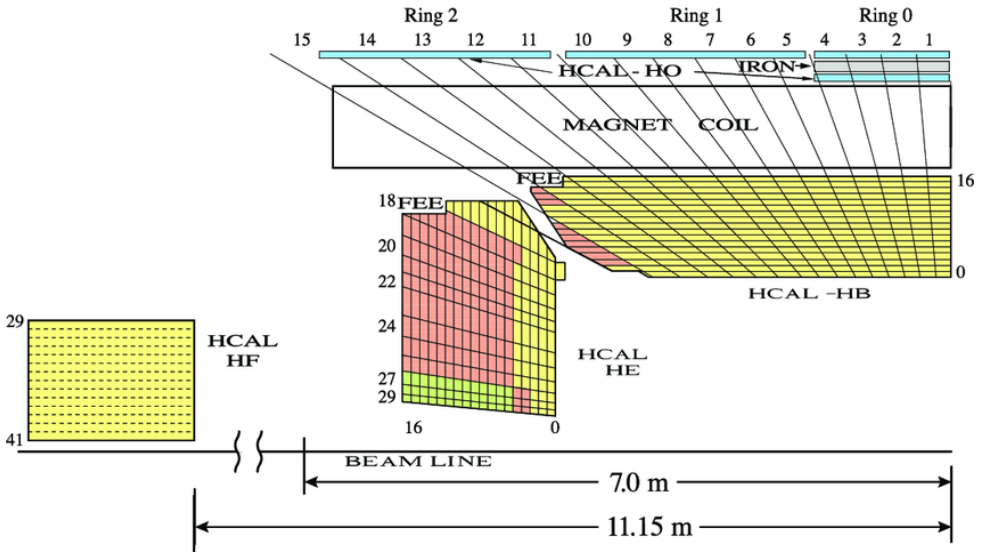


Figure 2.10: Illustrated view of a quadrant of the HCAL of the CMS experiment. It consists in a barrel calorimeter (HB) completed by an outer hadronic calorimeter (HO) in the low η regions, and the endcap (HE) and forward hadronic calorimeters (HF) in the high pseudorapidity regions of the detector. Figure taken from Ref. [54].

ular in the forward regions, leading to a deterioration of the detector's response. This degradation is monitored at each re-fill of the LHC and is illustrated in Fig 2.9. The detector response is significantly degraded with the luminosity increase. To cope with the even higher luminosities foreseen for the Phase II of the LHC, the endcap calorimeter, that receive the highest radiation dose, will be replaced by the High Granularity Calorimeter (HGCAL) based on radiation hard materials such as silicon sensors, as described in Sec. 4.3.

Hadronic calorimeter

The measurement of hadronic showers in CMS is performed by the hadronic calorimeter (HCAL) [55]. The HCAL is a sampling calorimeter that alternates layers of brass absorber and layers of active material consisting in plastic scintillating tiles. The barrel of the hadronic calorimeter (HB) covers the $|\eta| < 1.4$ region with a thickness of around $7\lambda_i$ while the endcap (HE) that covers the $1.3 < |\eta| < 3.0$ region has an increased thickness of 10 λ_i to absorb the more boosted showers in high η regions. As some particles might be too energetic to be fully absorbed by the HB, a outer hadronic calorimeter (HO) complements the detector from outside the solenoid, extending the interaction depth to $11\lambda_i$ with only scintillating material. Finally, the very forward region $2.9 < |\eta| < 5.2$ is covered by the forward hadronic calorimeter (HF) situated at 11.2 m from the IP. The HF uses radiation hard Cerenkov quartz fibers with steel absorber to contend with the heavy particle flux received in that region. This structure of the HCAL is illustrated in Fig. 2.10.

The HCAL was designed to measure the energy of hadrons interacting via strong force and that only deposit $\approx 30\%$ of their energy while passing through the ECAL. The HCAL possesses a much larger interaction length than the ECAL to completely absorb the hadronic

showers. It is a crucial component, the only subdetector in CMS able to measure the energy of neutral hadrons, allowing to infer the presence of missing energy as coming from non-detectable particles such as neutrinos. It design was however constrained by the limited space available inside the CMS solenoid.

The energy resolution of the ECAL and HCAL combined has been measured to be [56]

$$\left(\frac{\sigma_E}{E}\right)^2 = \left(\frac{84.7\%}{\sqrt{E(GeV)}}\right)^2 + (7.4\%)^2. \tag{2.12}$$

This limited resolution is due to the complex nature of hadronic showers, that can present non-detectable and electromagnetic components, such as the $\pi^0 \to \gamma\gamma$ for example. The detector performance can nevertheless be enhanced by combining the HCAL information with those measured in other part of the CMS detector, with the Particle Flow (PF) algorithm detailed in Sec. 2.5.

The HCAL underwent upgrades by the end of 2017 in order to maintain its level of performance [57]. In particular, the hybrid photodiodes of the HB, HE and HF were replaced by silicon photomultipliers, enhancing the longitudinal granularity of the detector. The readout electronics were also upgraded to provide precise timing measurement to the PF algorithm.

Muon detection system

As its name suggests, the precise identification and measurement of the muons is a core feature of the CMS detector. Indeed, the muons are a critical product in many events where W, Z or Higgs bosons are involved. In the LHC, they are produced with high energy that allow them to pass through the tracker and calorimeters of CMS without significant interaction. A muon detection system [58] was embedded in the yoke as outermost layer of the CMS detector, outside of the solenoid, to identify the muons and perform precise charge and momentum measurement.

It is composed of four *wheels* that measure particles momenta from the curved trajectory caused by the 2 T return magnetic field. The muon detection system exploits three different types of gaseous detectors in a total of 1400 chambers. Drift tubes (DTs) are used in the central region $|\eta| < 1.2$, Cathode Strip Chambers (CSCs) cover the endcaps at $0.9 < |\eta| < 2.4$, and Resistive Plate Chambers (RPCs) add redundancy up to $|\eta| < 1.6$. In total, the detection surface reaches up to 2500 m^2 for a radiation depth of up to 20 λ_i in the low pseudorapidity sections. The placement of the different technologies of muon chambers, detailed in Fig. 2.11, is motivated by the expected background rates as well as the intensity of the magnetic field.

Drift Tubes: The central $|\eta| < 1.2$ region of the muon detector, immersed in a low magnetic field (generally bellow 0.2T) and were low background rates are expected, is composed of drift tubes (DTs). They are composed of 2.4 m long rectangular aluminum tubes with a cross section of 1.3×4.2 cm^2, filled with a mixture of Ar (85%) and CO_2 (15%). Electrodes on the top- and bottom-end of the cells provide a constant field, ensuring a uniform drift velocity of about 55 µm/s. The position of the muons can thus be inferred from the drift time of the knocked electrons towards the anode of the cell. Layers of DTs are stacked upon each others, providing

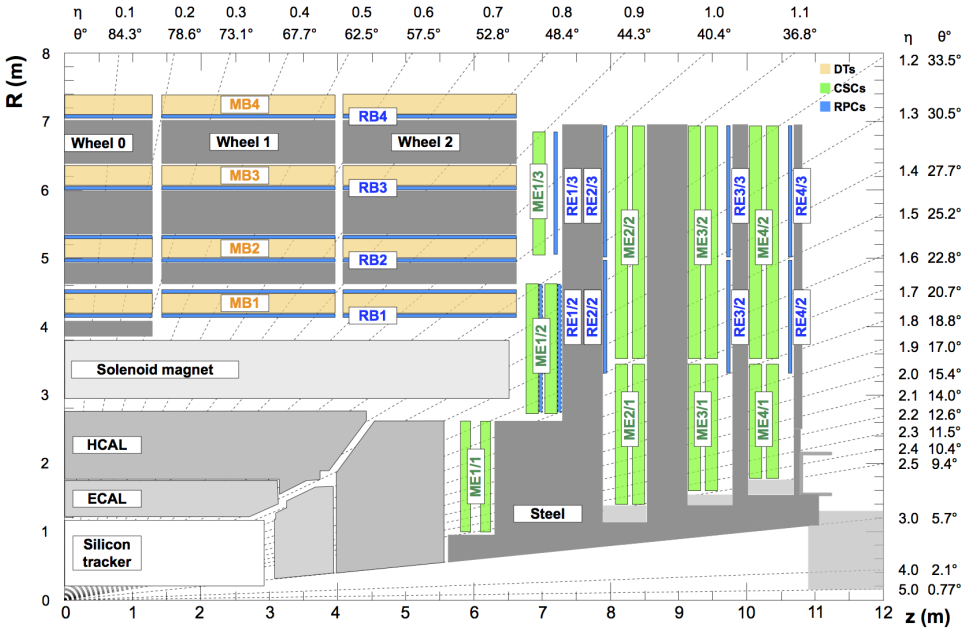


Figure 2.11: Illustrated view of a quadrant of the muon detector system of the CMS experiment. Drift Tubed (DTs), Cathode Strip Chambers (CSCs) and Resistive Plate Chambers (RPCs) are used in different regions of the system. Figure taken from Ref. [58].

a time resolution of less than 3 ns and a spatial resolution of around 180 µm, corresponding to a total position resolution of 80-120 µm.

Cathode Strip Chambers: For the regions at $0.9 < |\eta| < 2.4$, the expected background rate is more important and the magnetic field is stronger and less uniform. For those reasons, the detector employs CSCs of trapezoidal shape in the endcaps, alternating layers of anode wires and cathode strips filled with a mixture of Ar (45%), CO_2 (50%) and CF_4 (10%). The muon passage is detected by the ionization of the gas, that causes an avalanche of electrons in each layer. The technology has many advantages, with a fine granularity that provides up to 40-150 µm spatial resolution, precise timing with a resolution of around 3 ns, and a fast detector response useful for triggering purposes.

Resistive Plate Chambers: In addition to those two types of chambers, RPCs are installed up to $|\eta| < 1.9$ to provide a standalone triggering system and add the redundancy needed to ensure the correct association of events. Those double gaps chambers are constructed from thin layers of readout chips placed between high voltage electrodes and filled with mostly $C_2H_2F_4$ gas. They are operated in avalanche mode and provide mild spatial resolution of around 1 cm but with very fast response with a time resolution of less than 3 ns, complementing the DTs and CSCs and providing a fundamental handle to triggering events even at high pileup.

In addition to this original design of the muons chambers, Gas Electrons Multipliers (GEMs) were introduced at the end of 2017 [59] to increase the system's redundancy in the endcaps. The choice of GEMs was motivated by this technology's characteristic high rate capabilities even in harsh radiation environment such as the forward region of CMS.

2.4 CMS Trigger and Data Acquisition

The LHC collides bunches at a rate of 40 MHz. With a storage size of around 1 Mb per event for recording the full detector information, the reading and writing of this amount of data is not achievable with current technologies. Moreover, most of the interaction originate from low energy pp collisions which are not part of the focus of CMS's physics program. The Fig. 2.12 shows the cross sections for the main SM processes explored by CMS, with the most frequent one ($\sim 10^5$ pb) being 6 order of magnitude lower than the total pp interaction cross section of $\sim 10^{11}$ pb.

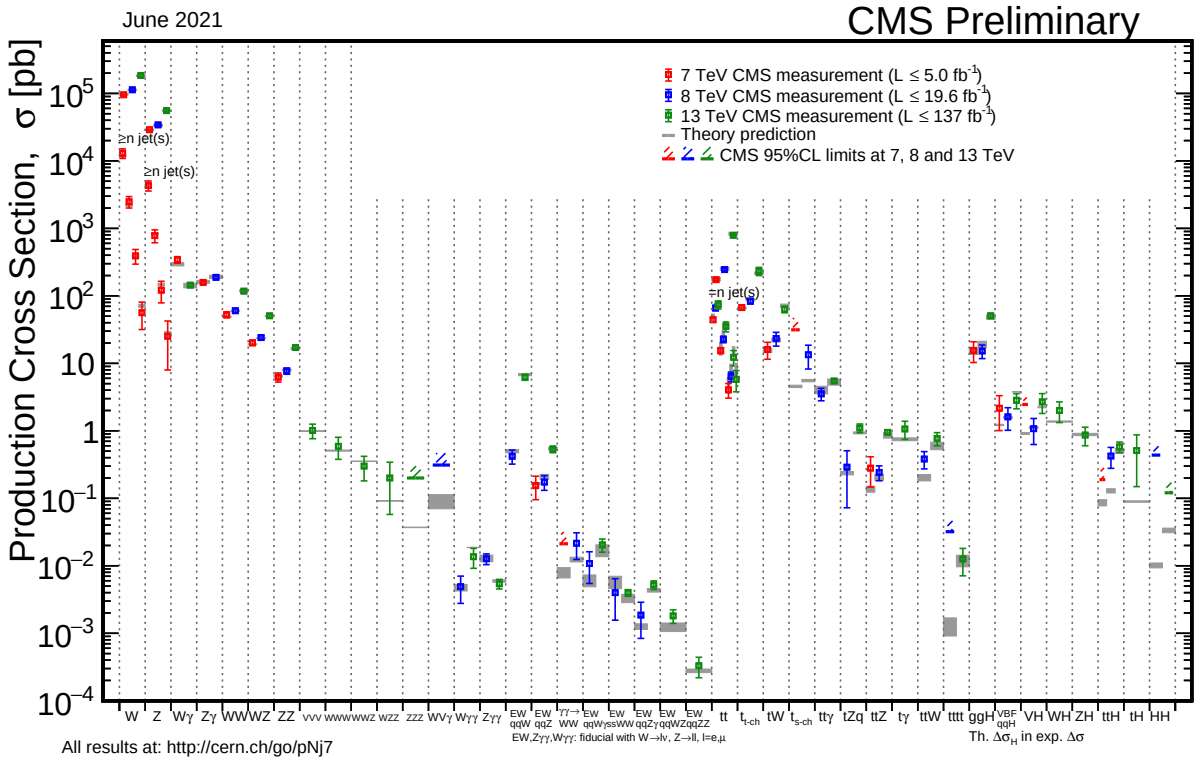


Figure 2.12: Cross sections of the main SM processes measured by CMS at $\sqrt{s} = 7, 8, 13$ TeV and the corresponding theoretical predictions [60].

Hence, selections must be implemented in order to identify and record only the relevant interactions. That is the role of CMS Trigger and Data Acquisition System (TriDAS). Designed to reduce the acquisition rate by five orders of magnitude, the TriDAS acts as the interface between the real-time or *online* data-taking, and the offline analyses. The design of this system must face complex challenges: satisfy the technical requirements of a fast selection while maintaining a high efficiency and background rejection, even in the harsh environment caused by the ever-increasing luminosity and pileup rate of the LHC.

The trigger selects events based on variables representing the kinematical properties of the different physics objects produced in the interaction, according to two consecutive steps. The first step, called *Level-1 trigger* (L1T) is executed by custom hardware processors that reduce the input rate from 40 MHz to up to 100 kHz of interesting events. The time available for the L1T to perform this selection, called *latency*, is of only 3.8 µs. As such, not all information from

the detector can be used and only a rough reconstruction can be performed. The L1T is then completed by a more detailed investigation of the selected events in the *High-Level Trigger* (HLT). Based on an out-detector computing farm, it further reduces the rate down to 1 kHz with a latency at the order of ≈ 200 ms. The events selected by the HLT are then recorded on permanent tapes at the CERN Tier-0 and can be fully reconstructed as needed. The following sections detail each step of the TriDAS system.

2.4.1 Level-1 Trigger

The first skimming of the data is realized by the L1T which reduces the input rate to only 100 kHz, with only 3.8 µs latency. With such a low amount of time to perform the selection, the L1T can not perform a full reconstruction of the event. It uses the raw data from the front-end electronics of the calorimeter and muon detector to create coarsely segmented and low resolution physics objects and produce L1 *candidates*. The available latency does not allow for the costly reconstruction of particle tracks, thus the information from the tracking system is not included at this step.

Initially designed for the nominal operation of the LHC, the L1T was upgraded in order to maintain the physics performance during Run 2 [61]. Between 2015 and 2016, the electronics boards were replaced with new advanced mezzanine cards (AMC) on which the Field Programmable Gate Arrays (FPGAs) are installed. The L1T algorithms are implemented of those circuits that can be configured through hardware description language, allowing for flexibility and future improvements on a fast hardware with fixed latency.

The architecture of this upgraded L1T is detailed in Fig. 2.13. A muon trigger aggregates the hits information from the various types of muon chambers while in parallel a calorimeter trigger collects the information of the energy deposited in the ECAL and HCAL. Those low-level information constitute the Trigger Primitives (TPs) which are used to creates candidates to represent the physics objects present in the event. The outputs of those two subtriggers are then processed by the *global trigger* which decides whether an event is accepted or rejected.

The global trigger performs its selection based on a list of algorithms made of thresholds applied to the kinematics ($E_T, p_{\mathrm{T}}, \eta$) of the L1 objects ($e/\gamma$, hadronic tau τ_h, jets, missing energy transverse E_T^{miss}, muons). The objects passing those cutoffs and additional isolation criteria form the L1 *seeds*.

2.4.2 High-Level Trigger

The events selected by the L1T are forwarded to the 32000 CPU cores of the HLT computing farm, where the rate is further reduced down to 1 kHz. This rate is limited by the speed of prompt reconstruction algorithm used before the tape recording of the events. Full detector information, this time including the tracking system, is used to perform finer reconstruction than at the L1T. With a maximum allowed latency of 320 ms per event, the reconstruction by the HLT is only a simplified version of the offline reconstruction detailed in Sec. 2.5.

The HLT reconstructs objects only locally around the selected L1 seeds in a sequence of

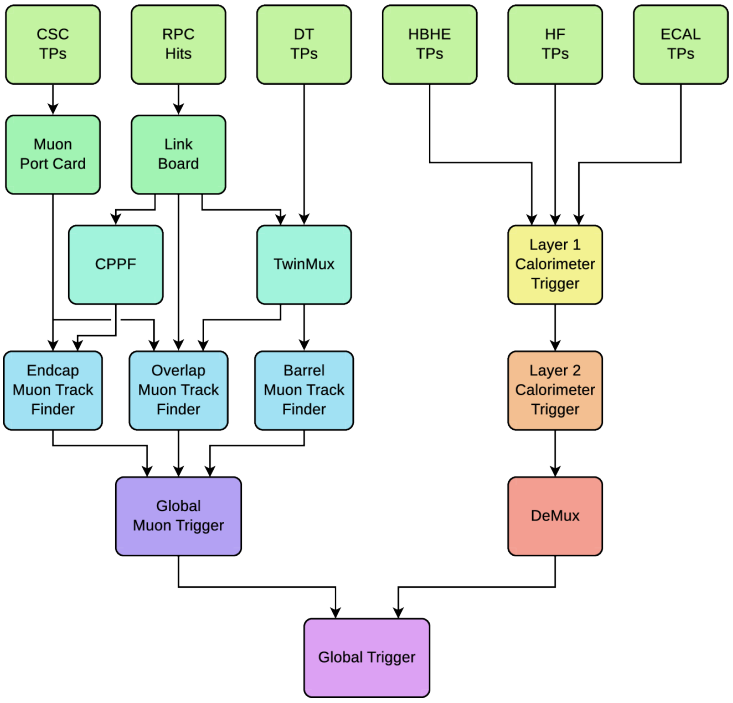


Figure 2.13: Architecture of the upgraded Level-1 trigger system of CMS. The muon trigger processes the hits information of the muon chambers while the calorimeter trigger constructs γ, e, hadronic τ and jets candidates. The outputs are then processed by the global trigger which decides to accept or reject the event based on the objects kinematics. Figure taken from Ref. [61].

reconstruction steps. Events are further filtered alongside those reconstructions, and the more complex and computationally expensive reconstructions such as tracks are only performed in the final stages on the events that have not yet been discarded. The HLT then classifies the remaining events in *paths* according to the topology of the event and archives them on the Tier 0 tapes at CERN where they remain available for full offline reconstruction.

The algorithms of HLT are constantly being updated in order to keep up with the changes in the conditions of the LHC operation and upgrades of the detector. In particular, new strategies have been implemented at the start of Run 2 to satisfy the exigences of the ever more sophisticated analyses. The *scouting* technique reduces the amount of recorded data based on the similarities with the offline reconstruction, thus increasing the available rates. A second important technique named *parking* consist in the recording of additional data on tape without reconstruction, and that could be mobilised for an *a posteriori* reconstruction if improved sensitivity was required in order to explore suspected deviation from the SM.

2.5 Event reconstruction

The particles produced in pp collisions in CMS leave different signatures in the various components of the detector depending on their nature, as illustrated in Fig. 2.14. Charged particles are bent from the solenoid's magnetic field and leave *hits* in the tracking system's active layers before continuing toward the calorimeter. Electrons and photon are then absorbed by the

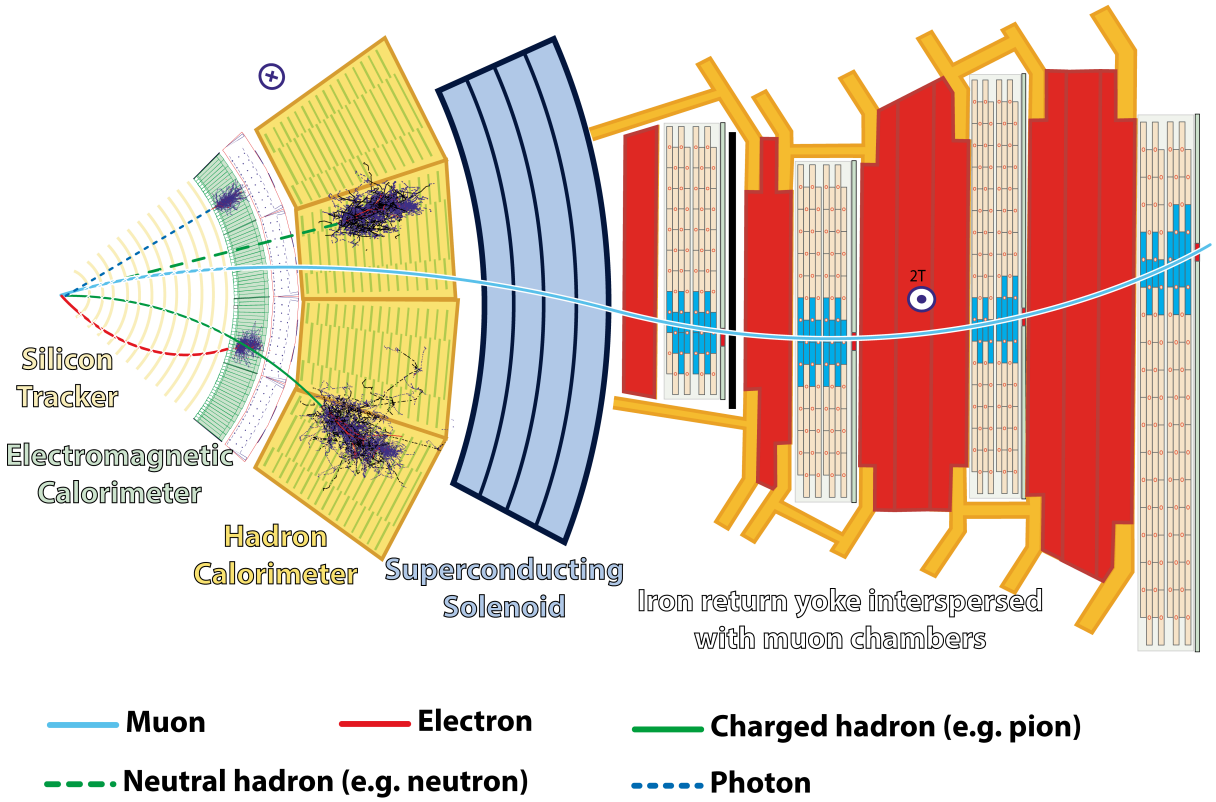


Figure 2.14: Schematic view of the typical signature left by various kind of particles in a transverse slice of the CMS detector. Figure taken from Ref. [62].

ECAL where they leave their energy in *clusters* of cells. The charged and neutral hadrons shower in a more complex fashion in the calorimeters but can similarly be *clustered* in the ECAL and HCAL. The muons mostly continue their path towards the muon chambers where they are measured while the neutrino leave the detector's volume without interacting.

Most of the events are skimmed by the L1T and HLT selections and only events with properties making them relevant to CMS physics program are recorded on the Tier 0 tapes in the RAW format, taking around 2 Mb per event. The increase from the 1 Mb space required for the full detector information originates from the trigger instruction that are recorded additionally. This raw information can then be mobilized for full offline reconstruction to identify and measure the physics objects with the maximal efficiency available by using sophisticated algorithms.

The basic algorithm used in CMS for reconstitution is the so-called Particle-Flow (PF) algorithm [63]. PF starts from the hits in the silicon tracker and energy deposited in the calorimeters' cells to build tracks and clusters, that are further used for reconstructing the different physics objects. Muon are identified first by matching tracks in the muon chambers to PF tracks. The electrons can then be reconstructed by associating the tracks to ECAL clusters while accounting for bremsstrahlung effects. The remaining clusters that can be linked to a track are identified as charged hadrons. The calorimeters clusters that can not be associated with tracks are then classified as photons or neutral hadrons depending on the position of the clusters. Those basic objects can then be combined to form more complex objects such as jets

and used to determine the transverse energy missing from the events hinting to the presence of neutrinos.

2.5.1 Particle Flow building blocks

The building blocks for the PF algorithm are the tracking system tracks and the calorimeters clusters.

The tracks left by charged particles are reconstructed from the hits they left in the active layers of the silicon tracker with the Combinatorial Track Finder [64] (CTF). This algorithm, based on Kalman filters, is an iterative approach that provides a low fake rate with high efficiency. The first step is to identify tracks with an origin consistent with the Primary Vertex (PV), which are then selected and ignored in the following steps. The subsequent stages perform the reconstruction of more complex tracks such as those coming from B-hadrons in 12 iterations. Tracks produced less than 60 cm away from the PV and with transverse momenta as low as 0.1 GeV can be reconstructed by this technique.

The clusters are reconstructed from the energy deposited in neighboring cells of the ECAL and HCAL. Local maxima of deposited energy called *cluster seeds* are identified, and neighbouring energy is *clustered* around those seeds if they are above a 2σ threshold higher than the electronic noise. PF clusters are then built by smoothing the cell clusters with a Gaussian mixture model expectation-maximization algorithm. In order to enhance the ability to disentangle overlapping showers, the cluster reconstruction is ran independently in the preshower, ECAL and HCAL.

From those basic elements, the PF algorithm can then form *candidates* by *linking* tracks and clusters, and categorize them as a particular kind of object (charged or neutral hadron, electron, photon or muon) depending on mostly topological considerations. Corrections to the detector response depending on the candidate category can then be applied.

To link the tracks and clusters, the linking algorithm extrapolates from the last hit of each inner track to a depth in the ECAL corresponding to the typical maximum of energy deposited for an electromagnetic shower, the plane of the ECAL preshower layers, and a depth corresponding to one interaction length in the HCAL. If a cluster is found in a calorimeter cell, it is linked to the track. In case of ambiguity between several clusters, the closest one to the track is selected. Clusters in the ECAL overlapping with clusters in HCAL are also linked together. Those linked tracks and cluster from the PF *blocks*. The blocks are then further analysed and categorized as a specific particle candidate.

2.5.2 Muons

The first step towards identifying the PF blocks as a particular particle type is to check if they are consistent with a muon, due to the very clean signature provided by the muon chambers. The hits in the DTs and CSC are combined to form *muon seeds*, which combined with RPCs hits form the standalone muon tracks. *Global muon tracks* are then formed by linking such muon tracks with a PF track. So-called tracker muon tracks can also be formed by linking a PF track

of p_{T}> 0.5 GeV and total momentum p > 2.5 GeV with muon hits at a compatible position. Most of the muons produced within the muon system acceptance are identified simultaneously by those two approaches, in which case they are fused, and the overall efficiency for at least one of the types of tracks to be reconstructed is of 99%.

A set of quality criteria is then applied to the reconstructed muon candidates, such as the goodness of the track fit, the number of hits per track, and the matching level between tracker and standalone tracks. The choice of a trade-off between efficiency and purity of the selection can be performed by the selection of a working point (loose, medium or tight). The isolation of a muon relative to its p_{T} is defined to distinguish prompt muons from those originating from weak decays in jets, computed by summing the energy in $DeltaR$ cones around the muon. Corrections are applied to mitigate the contribution from PU. For charged particles, this amount to only selecting the particles associated to the PV in the isolation computation, while for neutral particles a correction is applied by subtracting the energy deposited in the isolation cone by particles that are not associated with the PV multiplied by a factor 0.5:

$$\begin{aligned} ISO = \Bigg(& \Sigma p_{\mathrm{T}}\left(h^{\pm}_{PV}\right) + \max\Big(0, \Sigma E_{\mathrm{T}}\left(h^{0}\right) \\ & + \Sigma E_{\mathrm{T}}\left(\gamma\right) - 0.5 \times \Sigma p_{\mathrm{T}}\left(h^{\pm}_{PU}\right)\Big)\Bigg) / p_{\mathrm{T}}\left(\mu\right), \end{aligned} \tag{2.13}$$

where $h^{\pm}$ is a charged hadron coming from the PV or PU, h^0 a neutral hadron, μ the muon and γ a photon. A tight and a loose working point are defined for the isolation criterion, with respectively 95% and 98% efficiencies.

The muon momentum is determined with the Tune-P algorithm [65]. Several fit strategies are implemented, and Tune-p selects the most relevant one based on goodness-of-fit and $\sigma(p_{\mathrm{T}})/p_{\mathrm{T}}$ criteria.

- *Inner-Track fit* uses only the information from the inner tracker to perform the fit. It is favored for low p_{T} muons for which the contribution of the muon subsystem is negligible.
- *Tracker-plus-First-Muon-Station fit* compute the momentum from the hits from the global muon track, and refits using the information from the inner tracker and innermost muon station, which provides the best momentum measurement amongst the muon subsystem.
- *Picky fit* is used to determine the momentum of muons that shower within a muon chamber. Starting from the global muon tracks, it performs a refit in high-occupancy chambers using only the hits compatible with the extrapolated trajectory.
- *Dynamic-Truncation fit* is used for muons that lose a significant part of their energy due to bending. The tracker tracks are propagated to the innermost station and a refit is performed using the hits from the closest segment to this extrapolated trajectory.

The reconstruction, identification and isolation efficiencies are measured with the tag-and-probe method [66] on dimuons coming from Z boson decays and $J/\psi \to \mu^+\mu^-$ decays for the

low p_{T} muons. The *tag* muon uses strict requirements while the *probe* muon uses more relaxed selections. The efficiencies can then be measured from the fraction of probe muons that satisfy the studied selection. The efficiency of the tracker track reconstruction, the reconstruction and identification efficiency, the efficiency of the muon isolation and the efficiency of the trigger are determined independently. The systematic uncertainties in data/simulation scale factors are estimated by varying the tag-and-probe conditions, in particular the choice of the signal and background modeling. The uncertainties are estimated to be at the 1% level for reconstruction and 0.5% level for isolation [67].

The muon momentum scale and resolution performance are determined from cosmic rays and collisions. For low to intermediate p_{T} muons, data from Run 2 collisions are used to correct the momentum scale and estimate the resolution, either from the measurement of the distribution of $\langle 1/p_{\mathrm{T}} \rangle$ for tight muons coming from Z boson decays, or using Kalman filters on tight muons from J/ψ and Υ(1S) decays. The momentum scale corrections are of about 0.2% in the barrel and 0.3% in the endcap, while the resolution for muons with $p_{\mathrm{T}} \approx 100$GeV is 1% in the barrel and 3% in the endcap. For higher p_{T} muons, the momentum resolution is estimated with cosmic rays, by comparing the momentum measured in the upper half of the detector with the momentum measured in the lower half. The momentum scale of high p_{T} muons is determined by looking at distortion in the shape of the q/p_{T} spectrum

2.5.3 Electrons and photons

Once eventual muons have been identified and the corresponding blocks removed from further identification, electrons (e) are reconstructed.

Electrons start depositing energy in the passive material of the tracking system mostly by bremmstrahlung effect before depositing the remaining energy in the ECAL. This early showering complexifies the reconstruction as they are deviated from their original path. To account for these challenges, sophisticated algorithms are employed to identify electrons tracks and clusters.

Independently from the PF iterative tracking procedure, a Gaussian-Sum Filter [68, 69] (GSF) algorithm is applied to reconstruct the electrons tracts. Compared to the Kalman filters based on single Gaussian used in the PF algorithm, the GSF relies on weighted sum of Gaussian PDFs to model the energy loss while accounting for the bremsstrahlung effect. The GSF is computationally expensive and is only used on seeds likely to represent electrons.

In the calorimeter, *seeds* are identified as the most energetic clusters, and neighboring clusters collected around a *Mustache* road [70] with $E_T > 4$ GeV are regrouped around seeds to create *superclusters*. The superclusters are then associated back to the tracks. For electrons with low p_{T}, superclusters can be too small to contain all the bremsstrahlung radiation. To reconstruct those softer electrons, a complementary *track-based* approach is used by projecting GSF tracks with $p_{\mathrm{T}} > 2$ GeV as seeds for the clustering. If the matching quality between the GSF track and the supercluster is good enough, they are combined to form an electron candidate. The electron reconstruction efficiency is estimated from the ratio between the number of reconstructed superclusters matched to reconstructed electrons and the number of all reconstructed superclusters measured through a tag-and-probe method using $Z \to ee$ events.

This efficiency is measured to be higher than 95% with less than 2% difference between data and simulations.

Electron candidates are then submitted to additional identification criteria to reject fakes such as misidentified jets. The identification is based on a Boosted Decision Tree (BDT) relying on shower shape, track and matching quality, and isolation variables. Different cutoffs are used on the BDT output, creating three working points (loose, medium, tight) with increasing purity and decreasing efficiency.

Compared to electrons, photons (γ) interact less during their passing through CMS tracker and deposit almost all their energy ($\sim$ 97%) in the ECAL where they shower. Thus, the reconstruction of photons in CMS [71] is based on ECAL superclusters seeds not associated to GSF tracks. To be selected as photon candidates, those clusters must be isolated from other clusters in the calorimeter, and the distribution of energy in the cluster must be consistent with the development of a shower originating in a photon. Those *ECAL photons* are the main sources of candidates, but in some events photons do interact in the tracker by pair production of an electron and a positron. Those leptons can easily be identified in the tracker due to the inverted curved trajectories and parallel momenta. The photons interacting as such are seeded whether in an *inside-out* or *outside-in* fashions. In the former case, the tracks are seeded by the displaced secondary vertices and extrapolated to the ECAL, while in the latter case the superclusters are extrapolated back to the tracker.

Once the photon candidates are identified, an isolation parameter is used to discriminate against misidentified photons, mostly coming from neutral mesons decays that emits collimated photons. This isolation is computed by successive cuts on discriminating variables or by multivariate techniques described in. The selection efficiency and sclae factors are measured in $Z \rightarrow ee$data using the tag-and-probe method. The ratios between data and simulations efficiencies only deviate from unity by less than 5%.

The energy of electrons and photons is extracted by using the energy deposited in the ECAL and the tracker measurements only for electrons. A precise calibration of the calorimeter's crystals was performed through $\pi^0 \rightarrow \gamma\gamma$ pairs to ensure the reconstruction quality of electromagnetic objects. In order to compensate effects of the transparency loss of some crystal or imperfect calibration, a correction is applied by measuring the mass of the Z boson. A sample of e^+e^- is generated for each data run and the correction is computed in bins of p_{T} and $|\eta|$ by fitting the data and simulation of the m_{ee} distribution around the Z peak with a Crystal Ball function.

A series of multivariate regressions are also applied to correct the energy of the e/γ, providing corrections to the supercluster energy and an estimation of the supercluster energy resolution in real detector conditions. Those regressions are based on Boosted Decision Trees trained on simulated samples events with two electrons or photons. After those corrections, the energy resolution is slightly higher in simulations compared to data. The energy scales are thus corrected by varying the scale in the date to match the one observed in simulated events, which amounts to a total uncertainty smaller than 0.1(0.3)% in the barrel(endcap).

2.5.4 Jets

Due to color confinement, quarks and gluons produced in events cannot be observed directly in the detector but hadronize, with the exception of the top quark that decays before that, giving rise to several collimated particles that are reconstructed together as *jets*. They are a powerful handle to identify the Vector Boson Scattering events studied in this thesis as those processes have a typical signature in two jets with high mass and pseudorapidity separation.

The jets are narrow cones of particles that deposit energy in the ECAL and HCAL. Hence, the first step of the jets reconstruction is to link clusters in the ECAL and the HCAL. Depending on the presence of energy in the ECAL and of a track, the jets can be further categorized: neutral hadrons do not leave tracks, while charged hadrons HCAL clusters can be linked to tracks and possibly to ECAL clusters.

The reconstruction is then performed by the anti-k_{T} algorithm [72]. The anti-k_{T} is an algorithm that clusters objects into jets based on their transverse momentum k_{T} and their distance parameter R based on the rapidity y and ϕ separations. The clustering is performed recursively by grouping pairs of PF candidates based on a distance metric defined to express the jets conic shape. The tightness of the cone is determined by the R parameter, that is usually set to $R = 0.4$ or $R = 0.8$. The cone is built around the highest p_{T} cluster by pairing the clusters iteratively in decreasing order of momentum.

A variety of techniques to mitigate the PU are employed in CMS. The Charged Hadron Substraction (CHS) [73] algorithm uses the tracker information to remove charged particles associated with a PU vertex from the clustering procedure. An alternative technique is the PileUp Per Particle Identification (PUPPI) [74] which calculates on a per-event basis, the probability for each particle to be originating from the primary vertex, and scales their energy based on this probability. On top of those algorithms, a PU jet ID [75] based on a BDT is employed to identify low p_{T} jets coming from PU.

The jet energy is calibrated in a multiple steps approach. The first step consists of a simulation-based correction that subtracts from the jet energy the average energy offset coming from PU. The second step corresponds to the calibration of the jet energy response to account for detector effects that change the reconstructed jet energy compared to the particle-level. To that end, several standard candles such as $Z \to ee, \mu\mu$ or $\gamma\gamma$ are used in a global fit, resulting in a precision at the order of a few %. The Jet energy resolution is then corrected to match the jet resolution observed in real data by applying scale factors. They are particularly important in the $2.5 < |\eta| < 3$ region where the calibration is degraded due to the overlap between multiple sub-detectors.

The momentum of the reconstructed jet can then be computed from the vector sum of all clustered PF candidates, corrected for detector effects. While jets are complex objects, the typical CMS energy resolution remains under 15% across all p_{T} ranges with a detector response ~ 0.9.

The reconstructed jets must then satisfy a set of requirements to mitigate the contribution of instrumental noise and bad quality reconstructions. Depending on the strength of the requirements, *loose* and *tight* jets are defined. The criteria for these identifications are based on

the charged hadron multiplicity, the fraction of neutral hadrons in the jet and the fraction of energy deposited in the ECAL.

It is critical for many analysis, such as the one presented in this thesis, to be able to identify jets coming from b-quarks. In particular, this can allow to reject the background processes where a top quark is produced and reconstructed as b-jets. B-hadrons from the b-jets hadronization possess a signature allowing the discrimination between b-jets and lighter quarks jets. Their high mass and long lifetime gives rise to a displaced secondary vertex with high jet multiplicity. Two b-taggers are used in CMS to exploit these characteristics with Deep Neural Networks. The one used in this analysis is the DeepCSV ('Combined Secondary Vertex') [76] tagger that offers three stringency levels (loose, medium, tight) corresponding to respectively 10%, 1% and 0.1% of background misidentification and 83%, 69%, 49% of b-jet efficiency (those values are only indications since the efficiency depends on the jet p_{T} and η). Scale factors are computed to correct the efficiency in simulations based on b-jets enriched events.

Another algorithm, the Quark Gluon Likelihood [77] (QGL) is used to discriminate jets originating in quarks from those caused by gluons. The jets induced by gluons are generally less collimated than quark-jets, with higher jet multiplicity and softer energy fragmentation. This tagger is a likelihood discriminant based on three variables based on the PF candidates aggregated in the jet:

- The jet **energy sharing variable** p_{T}D defined as the sum over PF candidates $p_{\mathrm{T}}D = \sqrt{\sum {p_{\mathrm{T}}}^2} / \sum {p_{\mathrm{T}}}^2$
- The jet **multiplicity** defined as the number of candidates inside the clustered jet
- The minor axis RMS in the (η, ϕ) plane of the PF candidates

Those variables are computed with CHS and only neutral hadron candidates above 1 GeV are considered. The QGL can only be computed for jets clustered with at least three candidates. The likelihood discriminant is binned in p_{T} and PU in order to account for the strong dependence of the means and shape of the variables. The QGL is constructed for jets in the central $|\eta| < 2.4$ region and the forward $2.4 < |\eta| < 4.7$ region.

2.6 Machine Learning

With the very high amount of data collected by detectors in particle colliders and the complex analysis tasks that need to be performed on them, the high energy physics community has been widely relying on Machine Learning (ML) algorithms in the recent years. Machine Learning refers to methods that *learn*, meaning they infer from the data, the best way to perform a given task. Such uses of ML are ubiquitous in the current state of the high-energy physics, from their use in triggers to the discrimination of rare decay channels in analysis. The work presented in this thesis also widely relies on two particular kind of ML models, the Boosted Decision Trees (BDTs) and the Deep Neural Networks (DNNs), which are detailed in the following sections.

2.6.1 Supervised learning

A subset of machine learning algorithms,*supervised learning*, allows particle physicists to *train* models on perfectly known simulated data, notably to identify or classify physics objects such as particles. If the simulation is an accurate reproduction of the observed data, the trained models can be applied on the collected data to perform the same kind of discrimination.

In supervised learning, a training data sample of i points is used, possessing several features, also called input variables, x_i and a known target variable y_i. In the case of a binary classification between signal and background for example, the target can be set as $y_i = 0$ or $y_i = 1$.

A ML model is a mathematical object that performs a prediction of the target variable y_i based on the values of the features x_i. The most simple and common examples are *linear models* where the prediction is given by a linear combination of the weighted inputs

$$\hat{y}_i = \sum_j \theta_j x_{ij}. \tag{2.14}$$

The models possess a set of parameters, corresponding to the θ_j weights in the linear case, that need to be optimized for the model to perform accurate prediction.

This optimization process is called the *training* of the model. In order to find the best values of θ, an *objective* function is defined, describing how well the target y is predicted from the inputs x. Such objective functions are composed of two parts, first the loss of the model L that measures how good are the predictions; and a regularization term Ω that constrains the model's complexity.

A common choice for the loss function is the logistic loss, that has been widely used in this thesis, defined as

$$L(\theta) = \sum_i \left[y_i \ln\left(1 + e^{-\hat{y}_i}\right) + (1 - y_i) \ln\left(1 + e^{\hat{y}_i}\right) \right]. \tag{2.15}$$

The regularization term is used to avoid overfitting. In supervised learning, the model is optimized to predict the target of a given training sample, but with the goal of generalizing to other data (predicting already known data is not very useful). Without regularization, the model can become too complex and represent the training data very well, but give inaccurate predictions on a slightly different dataset. The most well-known regularization techniques are the L1 and L2 regularizations. The L1 regression, also known as Lasso regression, adds a penalty to the loss corresponding to the absolute value of magnitude of the model coefficient while the L2 regularization, sometimes called Ridge regression, adds the square magnitude of the coefficients as a loss penalty, thus forcing them to be small. The relative importance of the loss and regularization term is a trade-off between the predictivity of the model and its ability to generalise well.

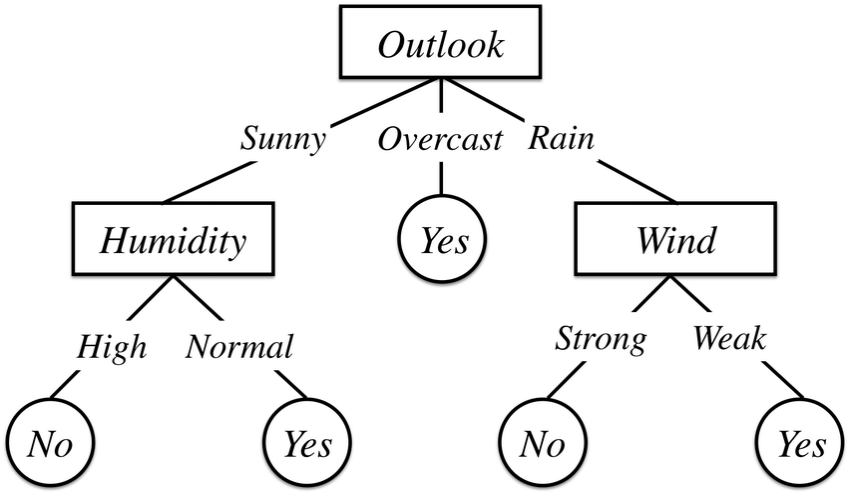


Figure 2.15: Graphical illustration of a decision tree for the question "Should I play tennis outside". The evaluation is performed starting from the top node and descending through the branches. The value on the bottom *leaf* determines the output. Figure taken from Ref. [78].

2.6.2 Boosted Decision Trees

Decision Trees

Decision trees are a decision tool based on a tree-like model to perform classification and regression.

Supervised learning is used to create models that can predict the value of a variable by applying subsequent binary decision rules. The data is classified by a series of nodes that perform a decision based on the value of a particular variable. From each node two branches corresponding to the possible decision are created and lead to subsequent nodes, until a termination criterion is reached.

This criterion can be the maximum depth of a tree, the number of final nodes called *leaves*, if the classes are perfectly split or if nodes are not populated enough. Each leaf then corresponds to a class and any event can be classified by following the path through the tree depending on the nodes splits. An example of such a tree is illustrated in Fig 2.15 for answering the question "should I go play tennis outside?". Depending on the weather the left or right *branch* is selected, and then the level of humidity or strength of the wind lead us to a leaf and the corresponding decision.

Advantages of the decisions trees is that they are a light and easily understandable model, with graphical visualization. They don't need a high amount of data to perform well and human knowledge can be used to predetermine the model. They can be used in combination with other techniques, but also combined with other decision trees.

In facts, a single decision tree is usually not strong enough to be used for complex cases. A common approach to enhance the performance is to sum multiple trees together.

Gradient boosting

When all the trees are trained on different random subsamples, the ensemble is called a *random forest*. But there can be redundancy in such a training method as most trees might be correlated. A more efficient method of training a tree ensemble is called *boosting*. The idea is to train a first tree, evaluate it, identify the data points incorrectly predicted and train a new tree that will focus on correcting those errors. In this way, the model is progressively made less biased with every new tree in the ensemble.The process can then be reproduced iteratively until a termination criterion, a given number of trees for example, is reached.

A BDT is an additive model in the sense that the final model $F_M(x)$ is a weighted sum of M weak learners (decision trees) $F_m(x)$, where m is the index of the tree in the ensemble. For gradient boosting, the model is first initialized with a constant $F_0 = \gamma$ that is fit to the actual target values y by minimizing a loss function $L(y_i, \gamma)$ where the summation on the i points is implicit

$$F_0(x) = \gamma_{opt} = \min_{\gamma} L(y_i, \gamma). \tag{2.16}$$

Unless the problem is particularly simple, the model will exhibit a high loss at this stage. Once this first model has been fitted and for all successive tree, the *pseudo-residuals* r are computed for each training point i as

$$r_{im} = -\frac{\partial L\big(y, F_{m-1}(x)\big)}{\partial F_{m-1}(x)}, \tag{2.17}$$

At every new step m, the pseudo-residual are thus extracted from the weighted average of the $m-1$ previous steps. The new tree $h_m(x)$ is then trained on a modified dataset D_{mod} defined as $D_{mod} = \{x_i, r_{im}\}$ and added to the total model as

$$F_m(x) = F_{m-1}(x) + \lambda h_m(x). \tag{2.18}$$

The λ parameter determines how much the model must be updated. It can once again be optimized by minimizing the loss function $L(y_i, F_m(x_i))$:

$$\lambda_{opt} = arg\min_{\lambda} L(y_i, F_m(x_i)) = arg\min_{\lambda} L\big(y_i, F_{m-1}(x_i) + \lambda h_m(x_i)\big) \tag{2.19}$$

Overall, the λ parameter must be optimized at each step on the original training dataset while each new tree parameters are optimized with respect to the modified dataset.

The eXTreme Gradient Boosting [79] algorithm is a particular implementation of this tree boosting concept that adds a regularization term based on L1 and L2 regularizations to improve the model generalization capabilities. It delivers higher performance than standard Gradient Boosting and is very fast.

2.6.3 Artificial Neural Networks

Another very popular class of algorithms in ML are the artificial neural networks (NN), based on the emulation of our brain architecture. A basic artificial neuron is the application of a non-linear function to a weighted linear combination of entries. Mathematically, it can be expressed as

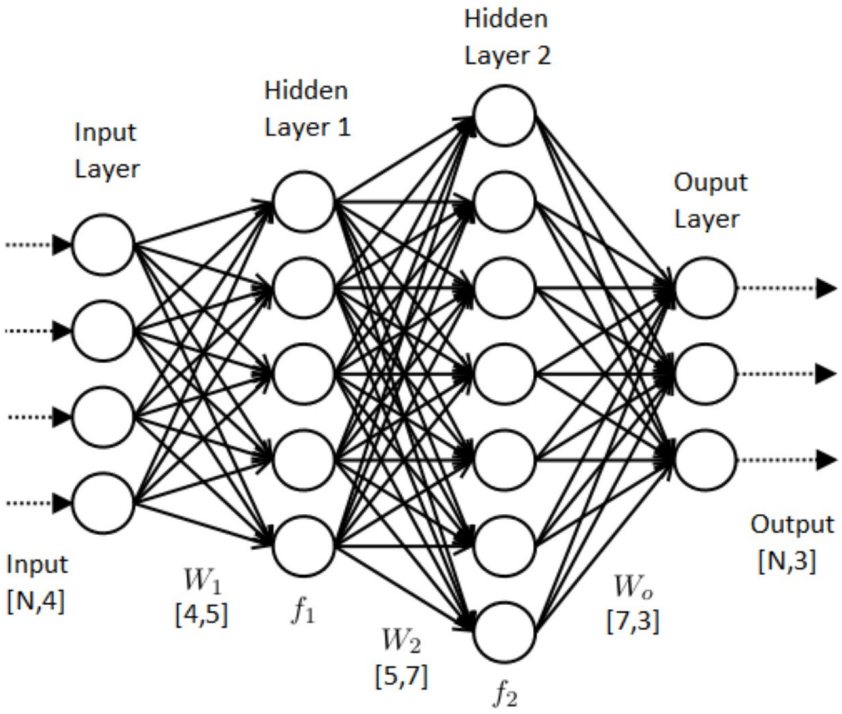


Figure 2.16: Graph of a multilayer perceptron with four features, two hidden layer and three outputs.

$$y = f\big(\sum_i (x_i \times w_i) + b\big), \tag{2.20}$$

where x_i are the inputs, w_i the associated weights, b the neuron bias term and f is a non-linear function.

As the name suggest, neural networks are a combination of neurons. A simple example, called *multilayer perceptron* (MLP) or *fully connected* NN is composed of successive layers of neurons where each of the neurons in a layer is connected to all the neuron of the previous layer and the following layer. The first layer, called input layer, is composed of one neuron per feature in the data and is followed by a given number of *hidden* layers. The last layer, called the output layer, can consist of multiple neurons if several values are expected, like classes probabilities or a single neuron for regression tasks for example. The idea is to *feed forward* the inputs through each layer's neuron up to the output layer. An illustration of such a MLP is shown in Fig. 2.16.

The output value of a neuron can be compressed from Eq. 2.20 by using the dot product

$$y = f(\vec{x} \cdot \vec{w} + b), \tag{2.21}$$

where $\vec{x}$ and $\vec{w}$ are the vector of the inputs and weights respectively. The *training* of a NN consists in optimizing the parameters of the network, the weights and bias, to perform a particular task. In the case of supervised learning, this is performed by optimizing a loss function $L(\hat{y}_i, y_i)$, and in particular the *cost function* C defined as the average of the loss across all data points.

The optimal weights and bias can be automatically obtained by the means of gradients and the chain rule:

$$\frac{\partial C}{\partial w_i} = \frac{\partial C}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z} \times \frac{\partial z}{\partial w_i}, \tag{2.22}$$

for the weight, where $z = x \cdot w + b$, and with a similar expression for the bias. As long as the loss function and non-linear terms are differentiable, the gradients $\frac{\partial C}{\partial w_i}$ and $\frac{\partial C}{\partial b}$ can be computed and the weight and bias can be updated by gradient descent as

$$\begin{aligned} w_i =& w_i - (\alpha \times \frac{\partial C}{\partial w_i}) \\ b =& b - (\alpha \times \frac{\partial C}{\partial b}), \end{aligned} \tag{2.23}$$

where α is called the *learning rate* and dictates how fast the parameters are updated.

This simple gradient descent can be generalized to multiple layers, and the gradient can be *backpropagated* up to the input layer.

The training of the network can be summarized as the initialization of the parameters followed by multiple iterations called *epochs* where:

- the data points are fed forward in the network;
- the output is compared to the target by means of the loss function and averaged in the cost function;
- the gradient of this cost function is backpropagated to update the weights and biases of all neurons.

While several activation functions can be used for the neurons, only non-linear functions allow the networks to perform non trivial tasks. It has been showed that any problem can be approximated by a fully connected network with enough non-linear neurons [80], and that increasing the number of hidden layers is more efficient than the number of neuron per layer. Such MLP with more than two hidden layers are called Deep Neural Networks or DNNs.

Amongst the most popular activation functions, one can cite the sigmoid function that provides a S-shape output and the Rectified Linear Unit (ReLU). The ReLU function is equal to 0 if the input is negative and to identity above 0. This fast to compute function is one of the sources of a revolution in the field of NNs as it allowed fast training of deep networks.

In addition to the simple gradient descent presented above, more sophisticated optimizers have been developed to improve the networks training. In particular, the Adam optimizer [81] provides independent learning rates for each parameters that are adapted while accounting for the recent changes (called *moments*). It improves the convergence capabilities of the neural networks and the neural networks used in this thesis rely on this algorithm.

2.6.4 Hyperparameter optimization

While ML model are great in that the best values for their internal parameters are automatically determined by the optimizer algorithm, they also introduce new *hyperparameters* that

also need to be tuned to maximize the performance of the models. Examples of such parameters are the maximum depth of the trees in a BDT and the number of neurons in a layer of a NN. Those hyperparameters can not be *trained* from the data via the ML optimizers. However, methods exist to optimize those hyperparameters. To that end, an objective function is defined that takes the hyperparameters values as inputs and return the score that has to be optimized.

The most basic algorithm to optimize this function is to perform a grid search. An instance of the model is build for every particular combination of hyperparameters to sample the phase space. However, it can be computationally expensive in particular when the hyperparameter space has many dimensions. A random search that selects values of each hyperparameter according to a statistical distribution can converge faster.

An advantage of the method previously presented is that the different evaluations can be performed in parallel, but that comes with the drawback that nothing is learned from precedent iterations. A more sophisticated approach is the so-called bayesian optimization. The basic concept is to draw information from already sampled models to predict the most promising region of the hyperparameters space to sample next, reducing the number of iterations needed to converge.

To that end, a surrogate function is built to represent the stochastic objective function in terms of probability in a continuous space. In practice, this is often based on Gaussian Processes. An acquisition function can then be used to select the values to be tested in the next iteration depending on the surrogate model probabilities. Those acquisition (or utility) functions are defined such as high acquisition corresponds to portions of the parameter space where the expected value of the model objective function is high. The next point at which the model must be evaluated is chosen by maximizing the function. A good acquisition function must offer a good trade-off between exploration and exploitation. The former means sampling portions of the space with few information, while the later consists in relying on the previous results to infer the most promising values. The Upper Confidence Bound [82] and Expected Improvement [83] can be cited amongst the most popular choices. An illustration of the bayesian optimisation is shown in Fig. 2.17. The Gaussian processes is constrained by the points in the hyperparameter space already sampled, and the utility function selects the next iteration that maximizes the probability of improvement.

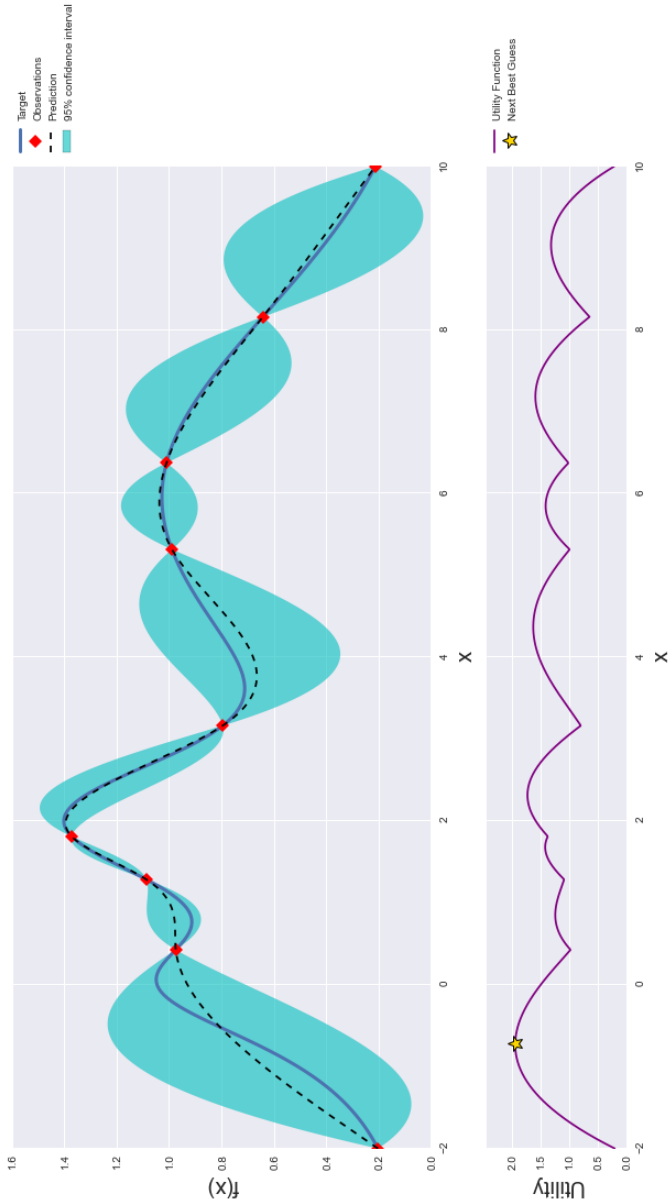


Figure 2.17: Illustration of the Gaussian process (top) and acquisition function (bottom) after nine steps of bayesian optimisation. Figure taken from Ref. [84].

Chapter 3

Search for Vector Boson Scattering production of a Z boson decaying to two leptons and a V boson decaying to jets

In this chapter is detailed the search for the electroweak production of a Z and V (V=Z,W) boson pair in the semileptonic decay channel in proton-proton collisions at 13 TeV with the 137 fb^{-1} data taken by CMS during the LHC Run 2 (2016-2018).

VBS processes such as this one are ideal to study the spontaneous symmetry breaking mechanism in hadron colliders like the LHC, as detailed in Sec. 1.3. They give a primary access to the non-Abelian gauge structure of the electroweak interactions through triple and quartic gauge couplings. At the LHC, both CMS and ATLAS have reported first observations of several VBS processes, as detailed in Sec. 1.3.4. Concerning semileptonic decay channels, CMS has recently published the first evidence of the VBS production of WV in the semileptonic channel [85] but no evidence has yet been shown for the VBS semileptonic ZV channel.

3.1 Analysis strategy

This analysis is a search for the rare ZV VBS production in the semileptonic decay channel. This channel is characterized by the presence of two leptons of same flavor (e or μ) with opposite charges, and a total of four jets: a pair of VBS tag jets and a pair of jets originating from the hadronic decay of the V boson. The analysis sensitivity is maximized by targeting two possible regimes for the emission of the hadronically decaying boson. In the case of a highly boosted boson, the two jets can not be disentangled and are reconstructed as a single large radius jet in the detector. On the other hand, bosons of moderate p_{T} are reconstructed as a pair of resolved jets with an invariant mass close to the W or Z boson mass.

Large background contributions affect the VBS ZV semileptonic final state due to the hadronic decay of the V boson. The dominant process is the production of a Z boson associated with jets, often named DY in this work after the Drell-Yan $q\bar{q} \rightarrow Z \rightarrow \ell\ell$ process,

followed by top quarks production and QCD multijet backgrounds. The choice of a signal region (SR) isolating the VBS signature requiring high dijet mass and pseudorapidity separation permits the reduction of the contamination from most backgrounds. The additional requirement on the invariant mass of the V boson to fall in a mass window centered around the gauge bosons masses allows to mitigate the Z + jets background, and the inversion of this criterion defines a control region dedicated to this process. The top background contribution is reduced by requiring lepton flavors to be identical, and a dedicated control region is created by inverting this requirement. In order to maximize the sensitivity, the regions of the phase space containing a b-tagged jet are treated separately from those without b-tagged jets.

The significance of the signal observation is extracted by performing a likelihood profiling across all those regions, in a simultaneous fit that uses the backgrounds normalizations extracted from the control regions to constrain the signal region. This likelihood is constructed based on background and signal simulation corrected for possible mismodelings. The top and Z + jets contributions are further corrected with a data-driven approach. In particular, the top-related distributions are taken from simulations but the normalization is left floating in the fit and estimated from the dedicated control region. The Z + jets normalization is also left free to float in the fit, but in a set of two dimensional bins (see Sec. 3.4.2) to correct observed discrepancies in the description of the kinematics between the MC simulation and data. Backgrounds entering the signal region via fake leptons are also estimated with a data-driven approach with jet-to-lepton fake rates.

In order to discriminate the rare VBS signal from the large background, the most discriminating variables are regrouped in a Deep Neural Network discriminator. Several models are trained depending on the category and year of data taking, as detailed in Sec. 3.6. The importance of each variable is estimated and the less discriminative are removed from the input set. Amongst them, the quark-gluon likelihood presents significant data-simulation discrepancies and is corrected with a morphing approach. The DNN provides an output whose last bins are particularly enriched in signal and the final analysis fit is performed on the binned shape of this output.

3.2 Physics objects

In this section are described the selection of the objects found in this final state, namely the leptons and jets. The selection of tight leptons is detailed before discussing the reconstruction of the different jets populating the final state. The event categorization, and definition of all the analysis regions is also described.

Muons

The leptons resulting from the Z boson decay are selected as tight electrons and muons. The muons are selected in this analysis following the selections used in the HWW analysis [86]. Particle Flow[63] *tight* [65]muons, as described in Sec. 2.5.2, are selected if they follow the following requirements:

- the muon has a pseudorapidity $|\eta| < 2.4$,
- the impact parameter of the muon track with respect to the primary vertex (PV) in in the (x, y) plane must satisfy $|d_{xy}| < 0.01$ cm for $p_{\mathrm{T}} < 20$ GeV and $|d_{xy}| < 0.02$ cm for $p_{\mathrm{T}} > 20$ GeV,
- the longitudinal distance between the track the muon belongs to and the primary vertex must satisfy $|d_z| < 0.1$ cm
- a muon isolation criterion based on PF combined relative isolation is used to reject the muons coming from hadrons in jets. A tight working point is defined with the requirement that $ISO_{\text{tight}} < 0.15$ which corresponds to an efficiency of 95%. Those selections where found in the HWW analysis to maximize the $S/\sqrt{S+B}$ ratio where S and B are respectively the signal and background yields.

Electrons

Electrons are primordial for many analysis and several strategies are used in the CMS collaboration to ensure that they are well reconstructed. The main background sources for isolated electrons are from photon conversions, misidentified jets and b or c quarks semileptonic decay. In this analysis, the electron selection is performed according to the method of the HWW analysis [86]. A MVA discriminator combining discriminating variables with a Boosted Decision Tree algorithm is used, with a working point (WP) of 90% signal efficiency. The set of variables used for the MVA training is detailed in the HWW paper in Tab. 18. A further requirement on the relative isolation is also applied: $ISO < 0.0588(0.0571)$ in the barrel (endcaps).

Jets

The quarks are observed in the detector as clusters of hadronized particles called jets. Those jets are confined in a small cone due to the large momentum magnitude of the parent particle compared to the small transverse momentum gained during fragmentation.

In this analysis, the PF candidates are clustered following the anti-k_T algorithm[72]. The standard jets are clustered with a parameter $\Delta R = 0.4$, called AK4. The pileup contamination is mitigated by using the Charged Hadron Subtraction technique and a PU jet ID is performed on jets with $p_{\mathrm{T}} < 50$ GeV based on a BDT discriminator with a loose WP. Additionally, jets candidates falling under $\Delta R_{j,\ell} < 0.4$ of a lepton are removed in order to keep leptons to be wrongly reconstructed as jets.

In some cases, the hadronically decaying V boson can be strongly boosted and thus the jets it produces can not be reconstructed as two resolved AK4 jets. They are considered in this analysis as a single large "Fat" jet reconstructed with the same anti-k_T algorithm but with a distance parameter of $\Delta R = 0.8$ and are therefore referred to as "AK8" jets. The pileup-per-particle (PUPPI) [74] algorithm is applied on the objects clustered in AK8 jets to remove PU tracks at the reconstructed particle level. The AK8 jets mass is computed via the softdrop [87] algorithm to remove the soft and wide-angle radiations from the large jet

clustering, thus improving the jet mass reconstruction. A further cut on the N-subjettiness variable $\tau_{21} < 0.45$ [88] is used to identify the jets coming from a V boson decay. This variable quantifies how well the jet can be divided into N subjets. The observable $\tau_{21} = \tau_2/\tau_1$ is employed to discriminate 2-prong objects arising from hadronic decays of or bosons from those from light quarks or gluons. AK4 jets falling within $\Delta R < 0.8$ from the boosted jets are also excluded from the event.

The present analysis requires to be able to identify jets coming from the decays of top quarks instead of vector bosons or VBS jets. The top quarks practically always decay in b quarks, which can be identified using b-tagging algorithms. Several b-taggers are available in CMS, based on different MVA methods giving an output corresponding to the likeliness that a jet originated from a b quark. The one used in this analysis is the DeepCSV ('Combined Secondary Vertex') [76] tagger with a loose WP (10% misidentification rate) to veto b-jets and a tighter WP (0.1% misidentification rate) to tag them.

3.3 Dataset and selections

As detailed in Section 1.3, Vector Boson Scattering (VBS) are purely electroweak processes where a quark (or anti-quark) scatters against another quark (or anti-quark) via space-like exchange of electroweak gauge boson and emits two electroweak bosons. The interacting quarks leave a very characteristic signature in the detector with two forward jets with large dijet mass and pseudorapidity separation as schematically represented in Fig. 3.1. These jets are often called *VBS* or *tag* jets and are a powerful way to identify VBS events, in conjunction with the more central emission of the two vector bosons and the resulting decays.

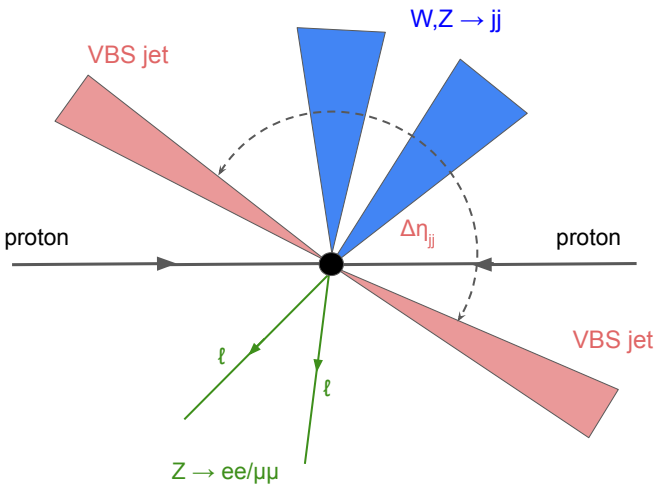


Figure 3.1: Schematic representation of a VBS ZV semileptonic event from proton-proton collision. The VBS jets resulting from the original quarks scattering exhibit the large pseudorapidity separation characteristic of VBS processes.

In the ZV semileptonic channel, the Z boson decays into two charged leptons of same

flavor and opposite charge, while the V boson decays into two quarks that are reconstructed as jets after parton showering and hadronization. The semileptonic channel benefits from the larger branching ratio of the hadronic decay compared to the fully leptonic channels, but at the same time suffers from bigger irreducible background, namely the QCD production of a Z boson associated with jets.

3.3.1 Data and triggers

This search aims to analyse the data taken during the full Run 2 of the LHC proton-proton runs, which corresponds to the data taking years of 2016, 2017 and 2018 at $\sqrt{s} = 13$ TeV. Only events corresponding to portions of registered data where all CMS detectors were confirmed to perform correctly were selected. Collision events are further selected by trigger paths requiring the presence of a single or double lepton in order to remove irrelevant portions of the data. The values for the minimum trigger threshold applied on the p_{T} of the leading and subleading leptons are summarized in Tab. 3.1.

Lepton p_{T} threshold (GeV)						
Year	2016		2017		2018	
Trigger	Leading	Subleading	Leading	Subleading	Leading	Subleading
eμ	23	8	23	12	23	12
μe	23	12	23	12	23	12
$\mu\mu$	17	8	17	8	17	8
μ	24	-	27	-	24	-
ee	23	12	23	12	23	12
e	27 (25)[1]	-	35	-	32	-

[1]: Electrons with $|\eta| < 2.1$ use a 25 GeV threshold instead of 27 GeV.

Table 3.1: Values of p_{T} thresholds for the leading and subleading leptons in the single and double lepton triggers used in this analysis.

All events firing at least one of those triggers are included in the analysis, with the full sample corresponding to an integrated luminosity of 137 fb^{-1} (35.87 fb^{-1} for 2016, 41.53 fb^{-1} for 2017, 59.74 fb^{-1} for 2018).

3.3.2 Background composition

The semileptonic VBS ZV final state can be mimicked by other processes that contaminate the analysis signal region described in Sec. 3.3.3. Those backgrounds are called "prompt" if the leptons are produced in the prompt decays following the hard scattering event. These includes the Z+jets, $t\bar{t}$, single top production tV and tZq processes, as well as minor backgrounds resulting from single or multiboson processes (VV, VVV, Vγ, VBF-V and $\gamma\gamma$WW) and W+jets production. Since only the electroweak production of ZV is of interest for this analysis, the QCD production of ZV is also considered as a background. The interference between QCD and electroweak is neglected as its contribution was found to be very small.

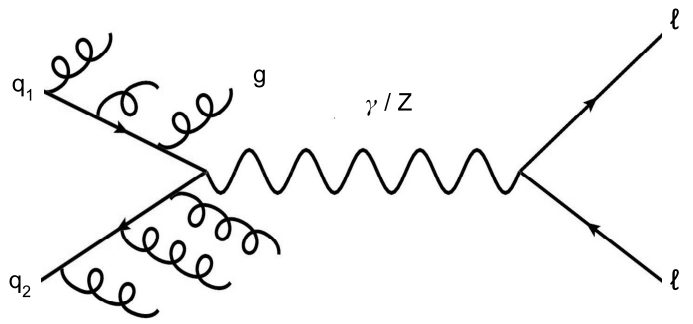


Figure 3.2: Diagram for the Drell-Yann production of a Z boson decaying leptonically with additional jets.

In particular, the Z + jets production is the dominant prompt background contribution, corresponding to between 55% and 80% of the total background contribution in the analysis signal-enriched subregions as defined in Sec. 3.3.3. This production can be mediated by the Drell-Yan (DY) production of a Z boson through the annihilation of two quarks as represented in Fig.3.2 .

The backgrounds processes involving a top have a very small contribution in regions of the phase space without b-tagged jet but amount to 10 - 15% of the portions where a b-jet is identified. A "non-prompt" background contribution arises from the incorrect identification as *prompt* leptons, in particular by the decays of hadron states in jets and photon conversions to leptons in the detector.

3.3.3 Event selection

To observe the very rare VBS process, a signal enriched region (SR) is designed, as well as two control regions (CR). Those CR are used to validate the simulation of the most important backgrounds, as well as correct their normalization (and shape in the case of the Z+jets background, as detailed in Sec. 3.4.2).

To be selected for the signal region, an event is required to have exactly two light leptons $\ell^+\ell^-$ where ℓ = e, μ. While the decay of the Z boson into τ leptons is part of the signal, $\tau leptons are not selected in the analysis. Their reconstruction is more challenging compared to the lighter fla$ 35 GeV and the other one $p_{\mathrm{T}}^{\ell 2} > 15$ GeV. To reduce the contribution from $t\bar{t}$ events, the invariant mass of this dilepton system is required to fall into a window centered around the nominal Z boson mass : $M_Z \in [76, 106]$ GeV. The leptons are additionally required to have a pseudorapidity $|\eta_\ell| < 2.5$. Events with additional leptons are rejected.

Events are also required to have four jets: the pair of *VBS* or *tag* jets, originating from the hard scattering process, and the pair of "V-jets" produced in the hadronic decay of the V boson.

Depending on the boson momentum, the two jets can be either reconstructed as two resolved jets or one very boosted large radius jet (AK8) as shown in Fig. 3.3.

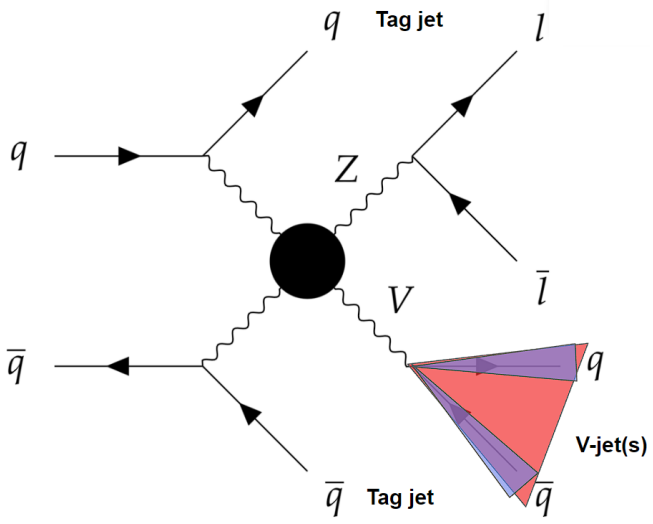


Figure 3.3: The diagram of the VBS ZV semileptonic process showing the the pair of VBS or Tag jets and the pair of V-jets. The V-jets can be reconstructed either as two resolved AK4 jets (blue) or one boosted AK8 jet (red).

The process to match the jets present in the event with the VBS and V-jets is as follows:

- First the VBS jets pair is chosen as the pair of jets with the highest invariant mass in the event with $m_{jj} > 500$ GeV and a large pseudorapidity gap $\Delta\eta_{jj} > 2.5$ corresponding to the typical VBS signature. Those thresholds are the result of an optimization of the $S/\sqrt{B}$ in the SR.

- Then the event is scanned for the presence of an AK8 jet with a softdrop mass corresponding to an on-shell vector boson $M_V \in [65, 105]$ GeV. If such a jet is found, the event is selected for the boosted topology.

- If no AK8 jet is found, we look for the presence of at least two more AK4 jets. If more than two AK4 jets are found, the V-jet pair is selected as the one with the invariant mass closest to the V boson nominal mass. This mass is computed as the average mass of a Z and W boson: $M_V = (M_Z + M_W)/2 = 85.78$ GeV. Events following these requirements are selected for the resolved topology.

The usage of this algorithm is motivated by studies in jet matching performed in Ref. [89] in which different matching algorithms were compared. The performance of the jet pairing strategies in the SR are reported in Tab. 3.2, highlighting the very good efficiency of the selected algorithm.

The complete set of requirements on the leptons and jets for the events to be selected as part of the SR are summarized in Tab. 3.3.

VBS Selection	Algorithm	Efficiency
Max m_{jj}	$\max\left(\sqrt{p^{\mu}p_{\mu}}\right)$	$64,4\%$
Max $\Delta\eta$	$\max_{i,j}\left(\lvert\eta_i-\eta_j\rvert\right)$	$49,5\%$
Max ΔR	$\max_{i,j}\left(\sqrt{\left(\eta_i-\eta_j\right)^2+\left(\phi_i-\phi_j\right)^2}\right)$	$47,9\%$
V Selection	**Algorithm**	**Efficiency**
Min $\Delta\eta$ / Nearest W	$\min_{i,j}\left(\frac{\lvert 80-\sqrt{p^{\mu}p_{\mu}}\rvert}{80}+\lvert\eta_i-\eta_j\rvert\right)$	$76,7\%$
Nearest W	$\min_{i,j}\left(\lvert 80-\sqrt{p^{\mu}p_{\mu}}\rvert\right)$	$76,5\%$
Min $\Delta\eta$/ Nearest W or Z	$\min_{i,j}\left(\min\Delta\eta+\min_{i,j}\left(\frac{\lvert 80-\sqrt{p^{\mu}p_{\mu}}\rvert}{80},\frac{\lvert 91-\sqrt{p^{\mu}p_{\mu}}\rvert}{91}\right)\right)$	$76,3\%$
Nearest W or Z	$\min_{i,j}($ Nearest W, Nearest Z$)$	$73,7\%$
Nearest Z	$\min_{i,j}\left(\lvert 91-\sqrt{p^{\mu}p_{\mu}}\rvert\right)$	$65,6\%$
Min $\Delta\eta$	$\min_{i,j}\left(\lvert\eta_i-\eta_j\rvert\right)$	$50,2\%$
Min ΔR	$\min_{i,j}\left(\sqrt{\left(\eta_i-\eta_j\right)^2+\left(\phi_i-\phi_j\right)^2}\right)$	$44,6\%$
Max $\mathrm{p_T}$	$\max_{i,j}\left(\sqrt{p_x^2+p_y^2}\right)$	$14,3\%$

Table 3.2: Summary of the Generator-Level studies on the efficiency of several selection algorithms for V-jets and VBS-jets selection presented in Ref. [89]. The efficiency for each individual algorithm is measured in the signal region as the ratio between the number of correct parton-jet matching normalized to the total number of events in the signal sample.

Object	Variables	Requirement
Leptons	N_ℓ	= 2
	$m_{\ell\ell}$	$[76,106]$ GeV
	${p_{\mathrm{T}}}^{\ell_1}$	>25 GeV
	${p_{\mathrm{T}}}^{\ell_2}$	>15 GeV
	$\lvert\eta_{\ell_{1,2}}\rvert$	<2.5
VBS jets	${p_{\mathrm{T}}}^{j_{1,2}}$	>30 GeV
	m_{jj}	$>500\,GeV$
	$\lvert\eta_{j_{1,2}}\rvert$	<5
	$\Delta\eta_{jj}$	>2.5
V-jet (boosted)	${p_{\mathrm{T}}}^{J}$	>200 GeV
	$\lvert\eta_J\rvert$	<2.5
	m_J	$[65,105]$ GeV
V-jets (resolved)	${p_{\mathrm{T}}}^{j_{1,2}}$	>30 GeV
	m_{jj}	$[65,105]$ GeV

Table 3.3: Summary of the jets and leptons requirements for an event to be selected for the signal region.

In addition to this signal region, two control regions (CR) close to the signal region phase space but enriched in the most important backgrounds are used. The goal of those CR are

two-fold:

- Ensuring the correct description of the data by the simulation in a phase space close to the signal region before unblinding the analysis.
- Performing a data-driven correction of those important backgrounds. For the top backgrounds, only the normalization is corrected while the Z+jets background shape is also modified since the normalization is corrected in bins, as described in Sec. 3.4.2.

The first control region is dedicated to the Z + jets production is called Drell-Yann Control Region (DYCR). This CR is defined by using similar requirements as for the SR but inverting the on-shell requirement on the reconstructed V boson mass : $m_V \notin [65, 105]$ GeV.

The second control region is enriched in processes involving the production of at least a top quark: $t\bar{t}$ and tV processes. This top CR is created by using the same requirement as for the signal region except the inversion of the leptons same flavor requirement, since most of those processes are not flavor symmetric, as opposed to the VBS signal.

The workflow for the event categorization, with the three regions (SR, DYCR and TOPCR) and two topologies (Boosted and Resolved) is schematically summarized in Fig. 3.4.

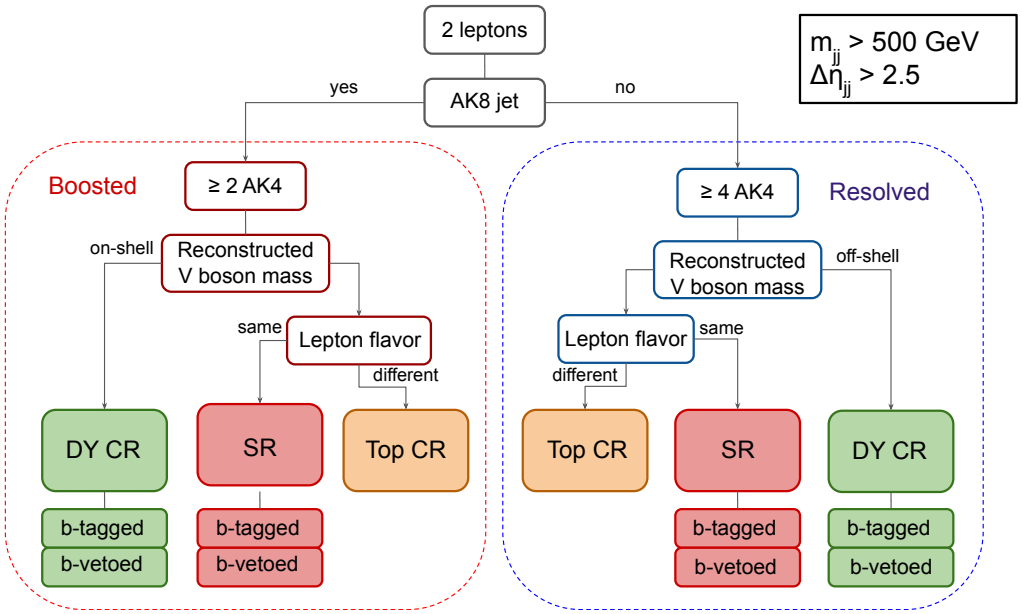


Figure 3.4: Workflow for the event categorization of the events. At least one pair of AK4 jets must correspond to the VBS signature high m_{jj} and large Δ_{jj}. The boosted topology is identified by the presence of one AK8 jets as opposed to two V-jets in the resolved topology. The DYCR is then separated by requiring an offshell V boson and the top CR is separated from the SR by requiring opposite flavor for the leptons. A final category subdivision is performed depending on the presence of a tagged b-jet.

Furthermore, the treatment of events with at least one of the VBS or V-jets identified as coming from a b quark by the DeepCSV tagger (at a tight WP) is differentiated from events with no identified b-jet. Each region of the analysis phase space except the TOPCR is thus split into a b-tagged category and a b-vetoed category.

3.4 Monte Carlo simulations

Since deviations from the predicted VBS ZV signal strength could hint towards BSM physics, the signal and background yields in the analysis phase space according to SM predictions must be estimated. The generation of those simulations through Monte Carlo (MC) generators is detailed in this section, as well as the corrections performed to ensure that the data are adequately emulated.

3.4.1 Signal and prompt backgrounds simulation

The VBS ZV signal is simulated at the leading order (LO) with six EW and zero QCD vertices with the MADGRAPH5_aMC@NLO[90] v2.6.5 generator. The pair of vector bosons in the intermediate state is produced using the narrow width approximation and the bosons are then decayed by MADSPIN[91], partially accounting for finite-width effects and spin correlations. The QCD mediated production of a ZV boson pair with the same final state as the VBS process is considered as a background process. The sample corresponding to this production is simulated with the MadGraph5_aMC@NLO v2.6.5 generator at LO in perturbative QCD. In the SR described in the previous section, the QCD contamination amounts to around 5% (14%) in the resolved (boosted) topology. Because of the difficulty to reconstruct τ leptons, the leptonic decays to two τ are not directly considered in the simulations, but can enter the simulations when their decays lead to the reconstruction fo two electrons or muons.

The diboson processes WW, WZ and ZZ are produced with pythia[92] v8.2 and normalized to next-to-leading order (NLO). The simulation of $t\bar{t}$ and single top productions are performed respectively at next-to-next-leading-order (NNLO) and NLO (NNLO for tW channel) with powheg [93, 94, 95] 2.0 while the Z+jets, W+jets, Vg, VBF-V, ttZ, tZq and triboson VVV productions are generated with MadGraph5_aMC@NLO v2.4.2 at LO accuracy, with the V+jets and ttZ using the MLM matching scheme [96]. The Z+jets and W+jets are also renormalized to NNLO. To ensure good statistics across the whole phase space, the Z+jets and W+jets samples were generated in bins of the scalar sum of the jets p_{T} (HT). Correction factors are applied during the stitching of HT bins cross-sections in order to ensure consistency.

The parton shower, hadronization and simulation of the underlying event of all simulated samples except VBF-V are handled by pythia v8.226 for 2016 simulated samples and v8.230 for 2017 and 2018. The underlying event modelling is performed with the CUETP8M1[97] and CP5[98] tune respectively for the simulation of 2016 on one hand and 2017 and 2018 on the other hand. For 2016, the NNPDF 3.0 NLO[99] parton distribution function (PDF) is used while the 3.1[100] version is used for 2017 and 2018. The VBS signal parton shower is simulated by pythia using the dipole recoil scheme to improve the description of additional jets emission in VBS processes [101]. For the VBF-V background, the herwig7.0[102, 103] program was used for similar reasons.

After generation, the signal and background samples are processed through a full CMS simulation based on the geant4 package[104] to emulate the detector response. Additional interactions simulated through minimum bias events with pythia are overlaid to reproduce the effect of PU, with a distribution corresponding to the one observed in data, for an average

of 23 per event in 2016 and 32 in 2017-2018.

3.4.2 Simulated samples corrections

The simulated samples produced by the procedure described in the previous section are not always a perfectly accurate reproduction of the state of the art calculations and cross section computations, and similarly the detector simulation may not always be a perfect description of the reality. To minimize the impacts of these discrepancies, several corrections to the MC samples are implemented as detailed further.

- **Pileup reweighting** : The MC samples are simulated with a distribution for the number of PU interactions designed to emulate the condition expected for the corresponding data-taking period, but some deviations can still be anticipated. In particular, the distribution for the number of reconstructed vertices can be biased by the event selection and trigger. In order to correct those effects, an event-per event reweighting based on the ratio between the normalized distribution for the number of inelastic collisions for each data-taking period and the MC PU profile is applied. The number of collisions distribution is derived from the proton inelastic cross section value of (69.2 $\pm$ 4.6%) mb and the MC PU profile is a function of the true number of generated inelastic collision in the event.

- **Lepton identification efficiency scale factor (SF)**: A scale factor accounting for the differences between the data and MC samples lepton selections efficiencies is applied. The values of the scale factors and details about their computation can be found in Ref. [86]. The SF is applied in bins of the lepton transverse momentum and pseudo-rapidity.

- **Heavy-flavor tagging efficiency scale factor** : A SF designed to make the distribution of the b-tagger in the simulation close to the one in the data is applied to each event where the b-tagging discriminator is evaluated. The SF depends on the jet momentum, η, flavor and b-tagging discriminator value.

- **Pileup jet ID scale factor**: As for the b-tagger, the PU jet identifier is not 100% efficient on real PU jets and must be corrected via a similar procedure. A SF parameterized in bins of p_{T} and η is extracted as the ratio between the probability of a real PU jet to be tagged in data and the same probability in MC samples. The total weight per event is the product of the individual SF for all jets with $p_{\mathrm{T}} > 30$ GeV, $|\eta| < 4.7$ in the event.

- **Level 1 prefiring correction**: During 2016 and 2017 data taking, the loss of transparency of the ECAL endcap led to a timing shift of the trigger primitives. This causes errors where a portion of the trigger primitives are associated with a previous bunch crossing, making the global trigger incorrectly veto events. The trigger inefficiency due to this *prefiring* effect has been measured as a function of the pseudorapidity of high momentum objects and is used to correct the MC simulated samples. The procedure is as follows:

  - Prefiring probability maps for electrons, photons and jets are taken from Ref. [105].

  - All photons and jets with $20 < p_{\mathrm{T}} < 2500$ GeV, $2 < |\eta| < 3$ found in the events are considered.
  - A prefiring probability $p(\eta, p_{\mathrm{T}})$ is associated to each jet or photon object according to the probability maps.
  - If a photon and a jet are overlapping, the object with the higher associated probability is selected.
  - A 20% globally correlated uncertainty is given to the prefiring probability of each object in the event.
  - The overall trigger efficiency correction factor ω i computed according to the formula:

$$\omega_v = \prod_{i=\text{photons,jets}} (1 - p_{v,i}(\eta, p_{\mathrm{T}})) \qquad (v = \text{nom, up, down}), \tag{3.1}$$

    where the subscripts nom, up, down corresponds to the nominal, upward-variation, and downward-variation probabilities, respectively.

- **DY p_{T} reweighting**: For the signal and control regions, a precise prediction of the main background contamination coming from Z+jets production is necessary. The MC samples simulated at LO are reweighted using higher order corrections corresponding to NLO QCD and NLO EWK terms. Additionally, the Z+jets background distribution is corrected in a data-driven way to reproduce the kinematics observed in the data with the procedure detailed in Sec. 3.5.1.

- **Top p_{T} reweighting**: During Run 1 and Run 2 of the LHC, the transverse momentum spectrum of the top quarks in $t\bar{t}$ data was significantly softer than those expected by the simulations based on LO or NLO matrix elements interfaced with parton showers. The latest NNLO (+NLO EWK) calculations including state-of-the-art knowledge of the SM $t\bar{t}$ productions are used to correct the top quarks p_{T} spectra in the MC samples. Additionally, the top backgrounds normalization is corrected in a data-driven way to reproduce the yields observed in the data with the procedure detailed in Sec. 3.5.2.

- **Boosted V-jets τ_{21} tagging SF**: The difference between the data and the simulation in the efficiency for the AK8 jets tagging based on the τ_{21} selection is accounted for with a scale factor.

- **Quark-gluon likelihood (QGL) correction**: One of the variables used for building a signal discriminator described in Sec. 3.6, the QGL, suffers from bad modeling. A procedure called *qgl morphing* is employed to reproduce the data behavior and is detailed in Sec. 3.5.4

3.5 Background estimation

3.5.1 Data driven Z+jets background corrections

The most significant background is coming from DY production of a Z boson associated with jets. To ensure the best agreement of the simulation with the observed data, we use a control

region enriched in DY events as described in Sec. 3.3.3. The composition of this control region is detailed in Table. 3.4.

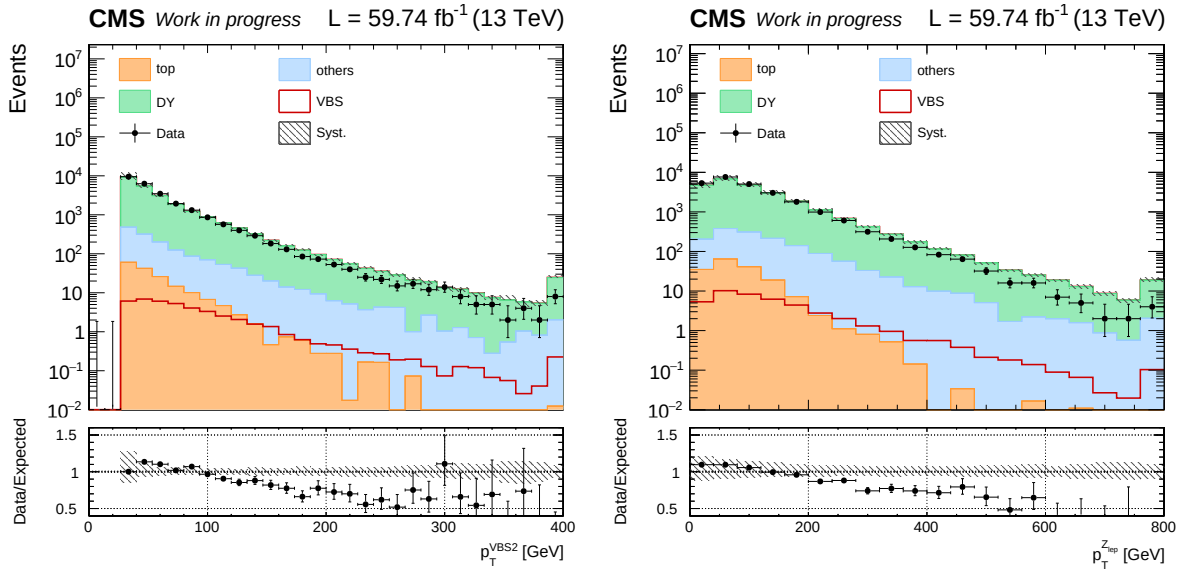
Sample	Resolved		Boosted	
	b-vetoed	b-tagged	b-vetoed	b-tagged
Z+jets	91%	76%	86%	72%
Tops	<1%	13%	<1%	10%
VBS QCD	2%	3%	6%	8%
Non prompt	1%	2%	<1%	1%
Multiboson	3%	3%	5%	5%
VBF-V	2%	2%	3%	3%
Signal	<1%	<1%	<1%	1%

Table 3.4: The composition of the DY CR in each category for the different signal and background samples computed in 2018 simulation data. In each category, the control region is heavily dominated by the Z+jets process.

The distributions of two relevant variables for the analysis, the second VBS jet transverse momentum ($p_{\mathrm{T}}^{VBSj_2}$) and leptonically decaying Z transverse momentum ($p_{\mathrm{T}}^{Z_{lep}}$) shown in Fig. 3.5 present significant bad modelling of data by the simulation of the DY sample. The issue is present mainly in the resolved topology for 2017 and 2018 with all topology of 2016 simulations being less affected. It is due to a MadGraph setting that causes different boson p_{T} spectra compared to the one observed in the data. In order to more accurately describe the observations, a data-driven approach is employed based on the strategy applied in Ref. [35]. The DY samples for the resolved topology in 2017 and 2018 are binned following two of the most problematic distributions, the $p_{\mathrm{T}}^{Z_{lep}}$ and the $p_{\mathrm{T}}^{\mathrm{VBS2}}$. The normalization of each bin is estimated independently during the global simultaneous fit. For the less affected boosted topology, and both topologies in 2016, a simpler one-dimensional binning according to the $p_{\mathrm{T}}^{Z_{lep}}$ is used.

The performance of this shape correction is strongly dependent on the choice of binning. In order to optimize this choice, the MC and data distributions are finely binned in a 2D histograms of 256 bins following the aforementioned variables. Each bin is made to have an identical population of data in order to avoid bias coming from statistical fluctuations in low population bins. This is achieved with the KD-trees method [106], that sequentially applies binary partitioning cuts in order to create k (in this case $k = 256$) same-population subspaces.

Each bin is affected with a rate parameter in the simultaneous fit, so the number must be kept small for the fit to converge properly. The 256 bins were reduced to twelve bins, as reported in Tab. 3.5, chosen in a way that maximizes the dependency to the ratio between data and expected MC as represented in Fig. 3.6. During the fit to the data the normalization of each bin is corrected independently. The results of this procedure in the postfit DY control regions are shown in Fig. 3.7 and Fig. 3.8 for the Resolved b-vetoed and Boosted b-vetoed categories respectively. The postfit distributions are adequately corrected and the effect is propagated to the distributions of other kinematic variables. The distributions of two of the most poorly modeled distributions are shown for 2018 DY CR Resolved b-vetoed category in Fig. 3.9. One draw back to this correction however is the introduction of a new source of uncertainty that



(a) **(b)**

Figure 3.5: Binned distributions for (a) the second VBS jet transverse momentum and (b) the leptonically decaying Z boson transverse momentum in the resolved b-vetoed DY control region for 2018 before fitting the simulation to the data. The last bin contains the overflow. The ratio between the observations in the data and the expected events from MC shows a bad modelisation of the data by the simulation, with an underestimation at low p_{T} and overestimation at high p_{T}. The trend is similar in the b-tagged subcategory and in the 2017 samples.

affects the results significance.

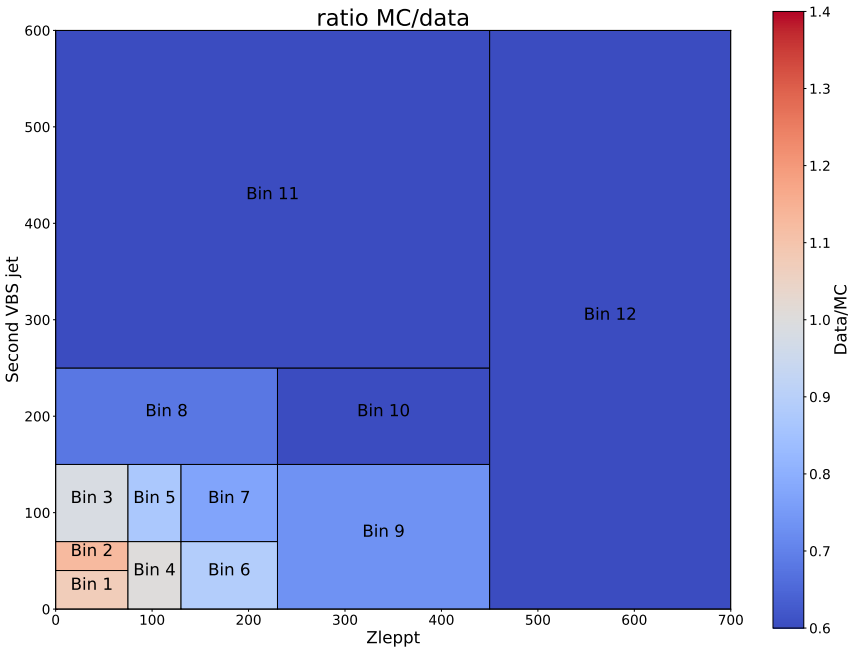
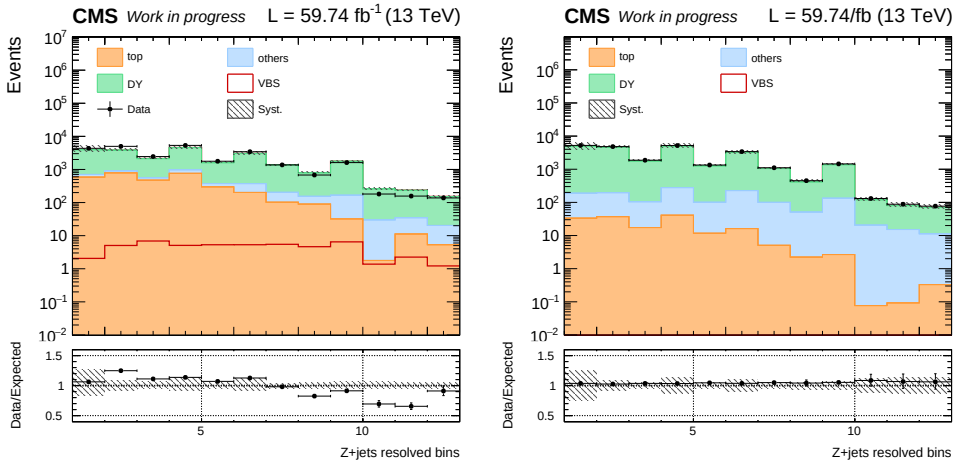


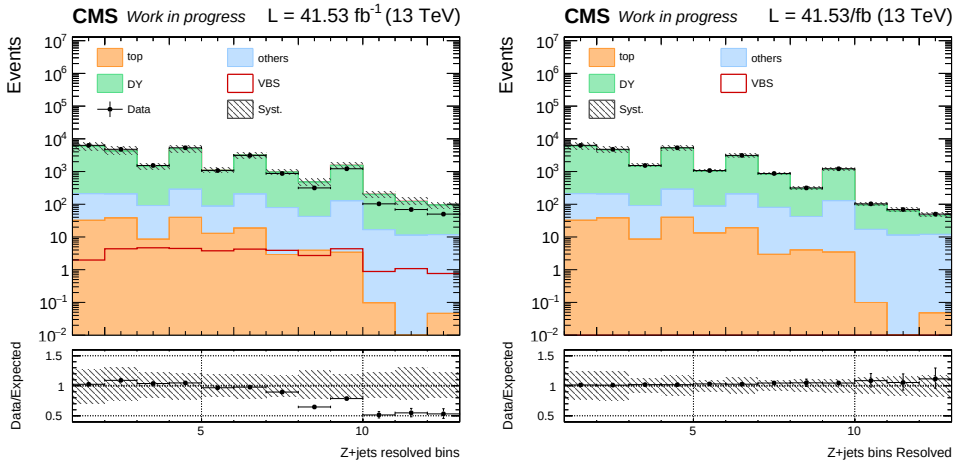
Figure 3.6: The 12 bins used for the DY normalization in the fit and the corresponding data/simulation ratio. They were chosen in a way to maximize the sensitivity to the data/MC ratio from the 256 same-population bins created with the KD-trees method.

Bin	2017 and 2018 Resolved		All years Boosted and 2016 resolved
	$p_{\mathrm{T}}^{Z_{lep}}$(GeV)	$p_{\mathrm{T}}^{\mathrm{VBS2}}$ (GeV)	$p_{\mathrm{T}}^{Z_{lep}}$(GeV)
Bin 1	[0,75]	[0,40]	[0,75]
Bin 2	[0,75]	[40,70]	[75,150]
Bin 3	[0,75]	[70,150]	[150,250]
Bin 4	[75,130]	[0,70]	[250,400]
Bin 5	[75,130]	[70,150]	[400, +∞]
Bin 6	[130,230]	[0,70]	-
Bin 7	[130,230]	[70,150]	-
Bin 8	[0,230]	[150,250]	-
Bin 9	[230,450]	[0,150]	-
Bin 10	[230, 450]	[150,250]	-
Bin 11	[0,450]	[250, +∞]	-
Bin 12	[450, +∞]	[0, +∞]	-

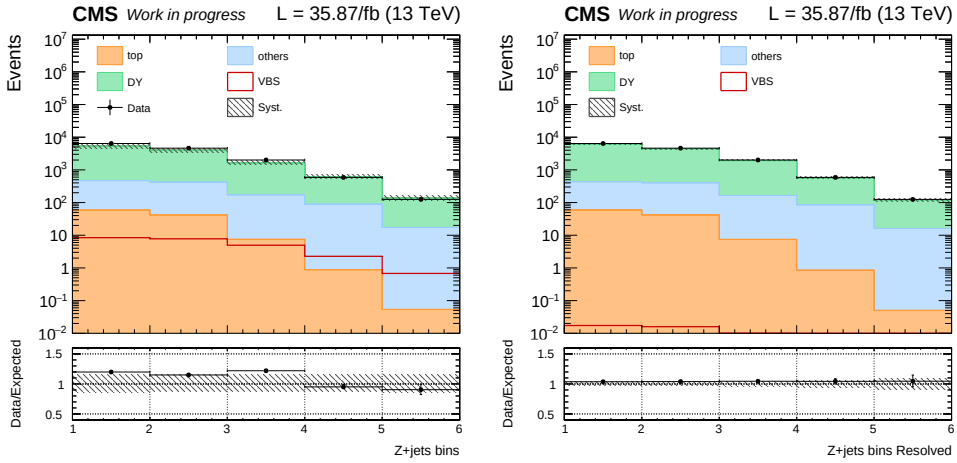
Table 3.5: Binning of the DY sample for the correction procedure. The resolved topology for 2017 and 2018 uses 12 bins of $p_{\mathrm{T}}{}^{Z_{lep}}$ and $p_{\mathrm{T}}^{\mathrm{VBS2}}$ while the less affected 2016 resolved topology and all years boosted topologies use only 5 bins of $p_{\mathrm{T}}^{Z_{lep}}$.



(a)

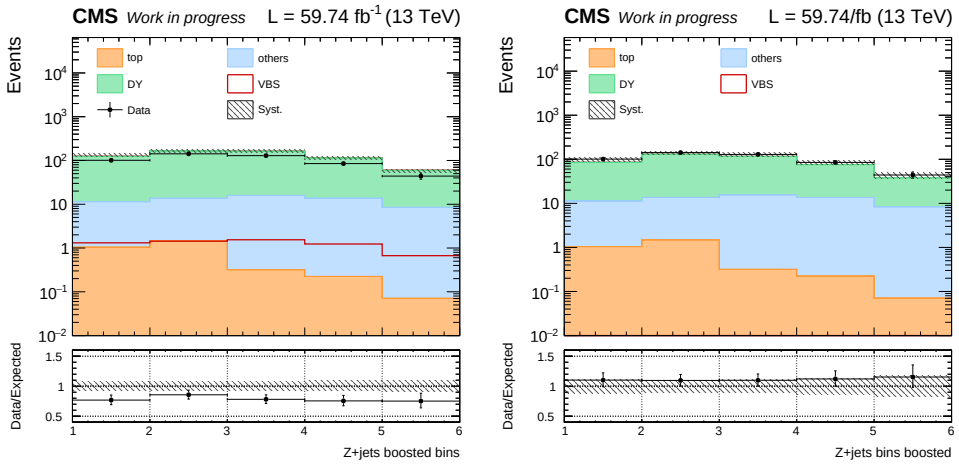


(b)

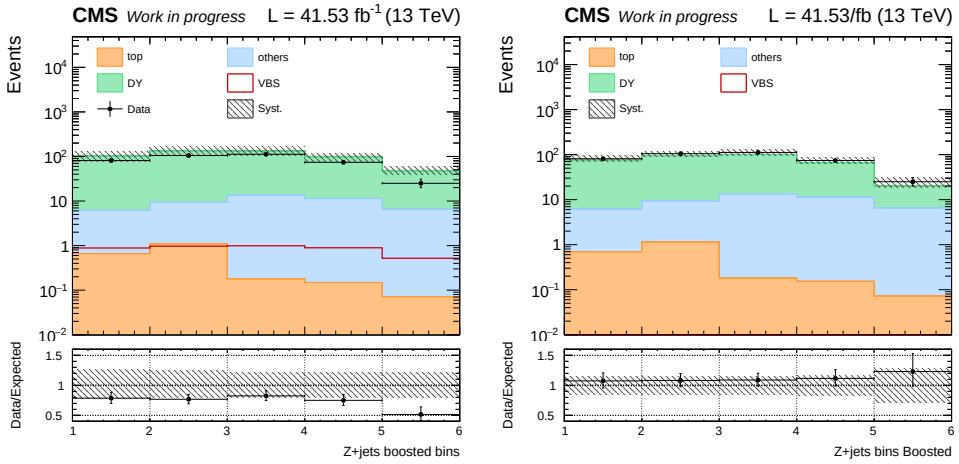


(c)

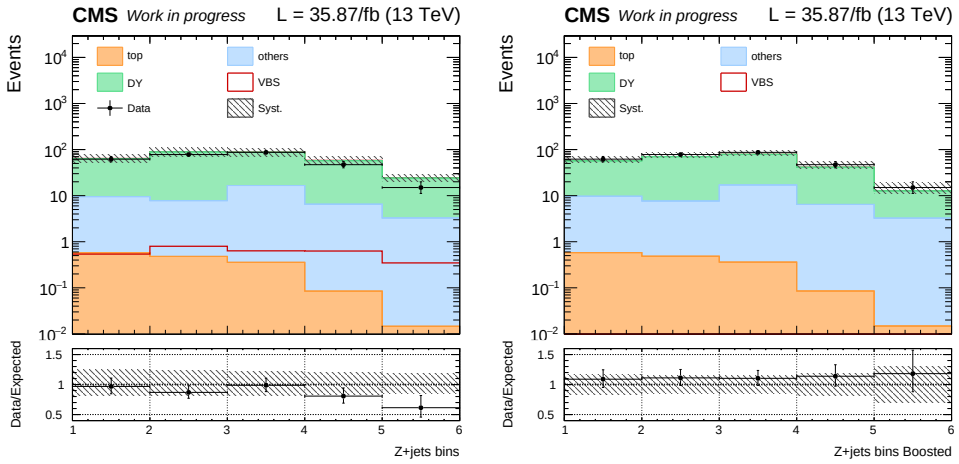
Figure 3.7: The prefit (left) and postfit (right) Z+jets bins in the (a) 2018, (b) 2017 and (c) 2016 Resolved DY CR b-vetoed region (c). 2017 and 2018 samples are divided into bins of ${p_{\mathrm{T}}}^{Z_{lep}}$ and $p_{\mathrm{T}}^{\mathrm{VBS2}}$ while for 2016, the samples are only binned according to ${p_{\mathrm{T}}}^{Z_{lep}}$.



(a)

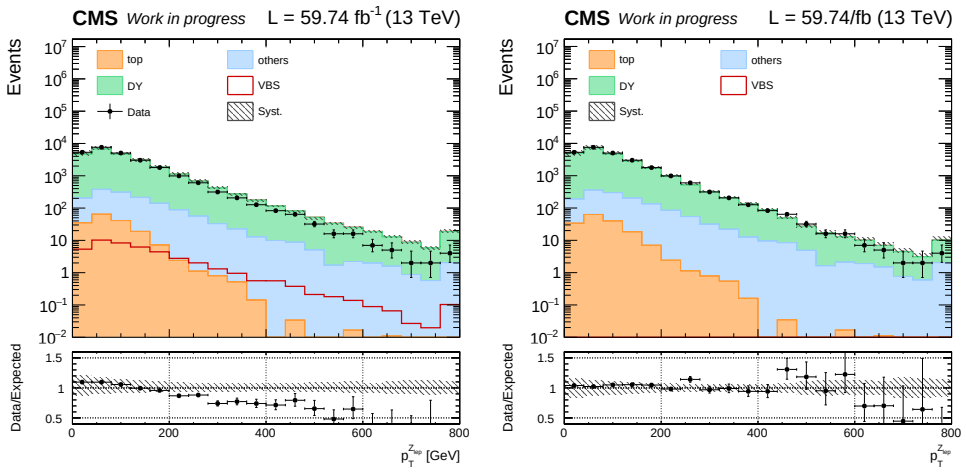


(b)

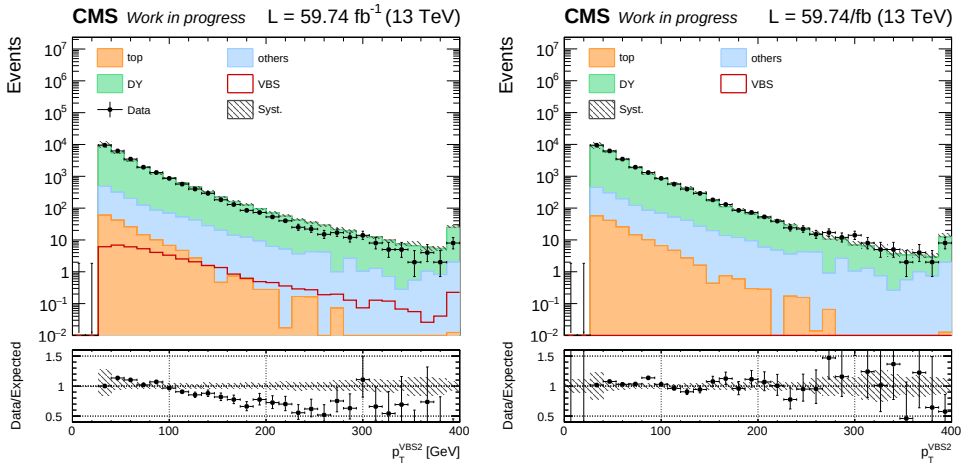


(c)

Figure 3.8: The prefit (left) and postfit (right) Z+jets bins in the (a) 2018, (b) 2017 and (c) 2016 Boosted DY CR b-vetoed region (c). For all years, the samples are binned according to ${p_{\mathrm{T}}}^{Z_{lep}}$.



(a)



(b)

Figure 3.9: The prefit (left) and postfit (right) distributions of (a) the $p_{\mathrm{T}}{}^{Z_{lep}}$ and (b) the $p_{\mathrm{T}}^{\mathrm{VBS2}}$ in the 2018 Resolved DY CR b-vetoed region. The two kinematics distributions are also well corrected by the procedure.

3.5.2 Data driven top backgrounds estimation

The expected contribution of $t\bar{t}$ and single-top production in the signal region is estimated with MC simulations, except for its normalization that is measured in the top quark enriched control region in the final fit to the data. As described in Sec. 3.3.3, this control region is created by requiring events (Table 3.3) with opposite flavor in the two leptons and in the on-shell region $65 < m_V < 105$ GeV, as the branching ratio of the processes contributing to those backgrounds to a pair of leptons of a different flavor, $e\mu$, is twice as large as the branching ratio to ee and $\mu\mu$. The composition of the control region is detailed in Table 3.6 for 2018 simulated data.

Sample	Resolved	Boosted
Single top	2%	5%
$t\bar{t}$	90%	82%
tZq	<1%	<1%
Z+jets	<1%	<1%
Multiboson	2%	3%
Non-prompt	6%	10%
Signal	<1%	<1%

Table 3.6: The composition of the top CR for the different signal and background samples computed in 2018 simulation data. In both categories, the control region is heavily dominated by the $t\bar{t}$ processes.

The VBS dijet mass m_{jj}, which counts amongst the most discriminating variables for discriminating the signal, is plotted in Fig. 3.10 for the data and simulations. The overall shape of the simulations provides an adequate description of the eμ data so only the normalization of the top processes is estimated in a data driven way by fitting the top CR simultaneously with the SR.

3.5.3 Estimation of the non-prompt background

The rates of background processes with non-prompt leptons is estimated using a data-driven technique, depending on the lepton p_{T} and η following the 'fakeable object' method described in Ref. [86]. A control sample of events dominated by dijet QCD production (which has a high rate of non-prompt leptons and jets misreconstructed as leptons) is created with leptons identification and isolation requirements loosened compared to the definition of the analysis SR. However prompt leptons can enter this sample through W or Z boson decays. In order to suppress this prompt contribution, the events are required to have $E_{\mathrm{T}}^{\mathrm{miss}} < 20$ GeV and $m_{\mathrm{T}}^{W} < 20$ GeV to reject lepton originating from W boson decays and a veto around the Z boson mass peak $60 > m_{\ell\ell} < 120$ GeV is used to suppress the Z boson decay leptons.

The lepton fake rate, the ratio of tight leptons over loose leptons is extracted in bins of p_{T} and η. The uncertainty is estimated by varying the p_{T} threshold of the sample. Once this fake rate has been estimated as a function of the lepton kinematics, it can be propagated to the different regions of the analysis.

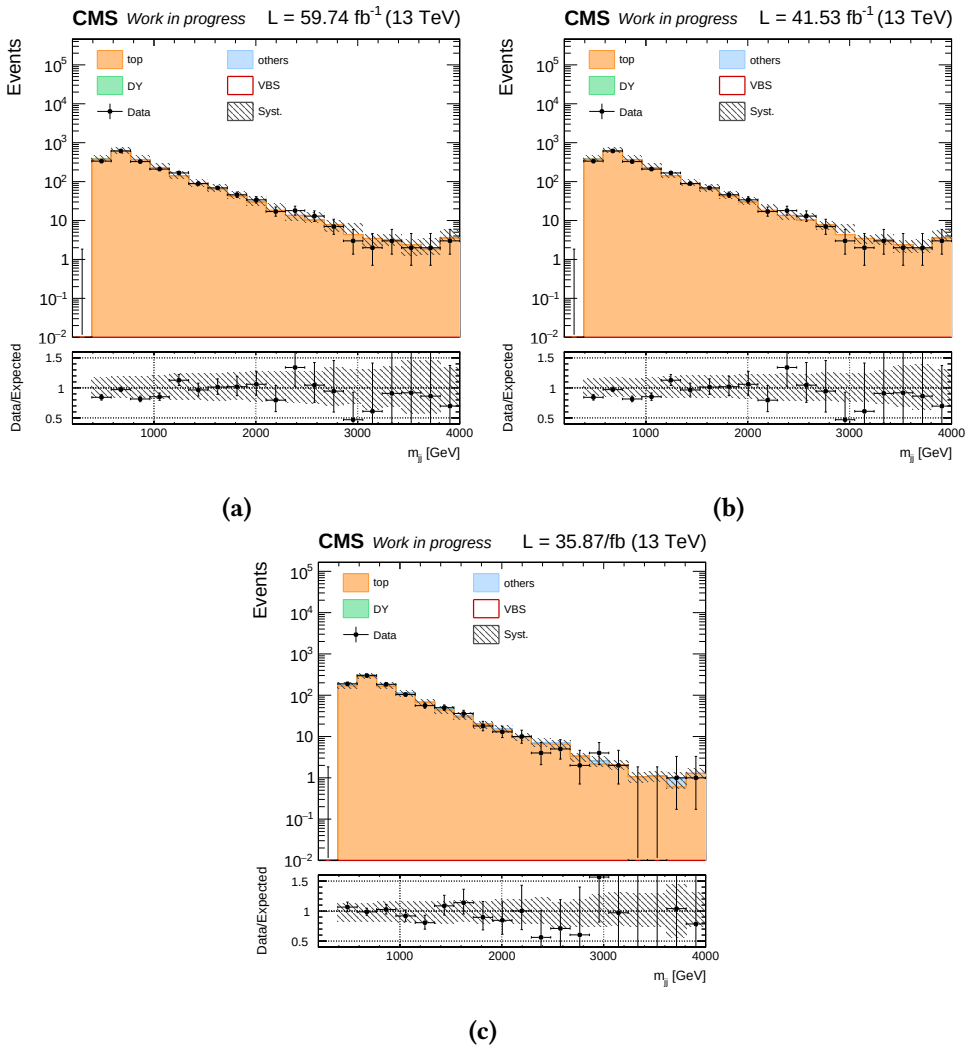


Figure 3.10: Prefit distributions for the VBS dijet system mass in the Resolved top control region for (a) 2018, (b) 2017 and (c) 2017. The top processes normalization is extracted from these regions to reproduce the data yields.

3.5.4 Quark-gluon likelihood variable

The quark gluon likelihood (QGL) variable is a tagger made to discriminate (light) quarks jets from gluon originating jets and is an important variable for the DNN in the resolved category as shown in Table 3.11. The gluon and quarks components are not perfectly described by PYTHIA8 parton shower however as showcased in Fig. 3.13. A correction procedure, called *QGL morphing*, already applied in Ref. [35] was used to correct the QGL shape without modifying the normalization.

This method consists in the morphing of the MC QGL cumulative distribution function (CDF) to reproduce the one observed in data, as illustrated in Fig. 3.11. First a global correction depending on the jet origin is extracted from each years resolved category. The jet is tagged as a gluon- or quark-jet from the flavor of the partons from which the jet originates. A second order correction is then computed, accounting for the residual differences of the QGL behavior in different regions of the jet η/ jet p_{T} regions.

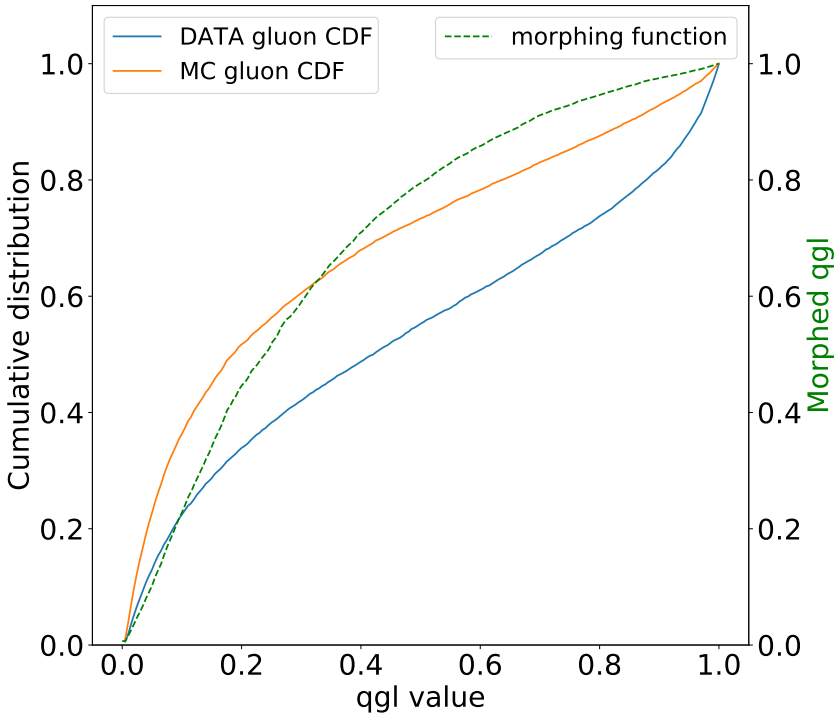


Figure 3.11: The cumulative distribution functions for the data (blue) in a gluon-dominated region (62% of jets are issued from gluons) and the gluon-originating jets in the MC samples (orange). The objective is to build a function that will transform the MC gluon/quark CDF to reproduce the distribution observed in the data in regions dominated by gluons/quarks. The resulting morphing function is represented as the green dashed-line curve.

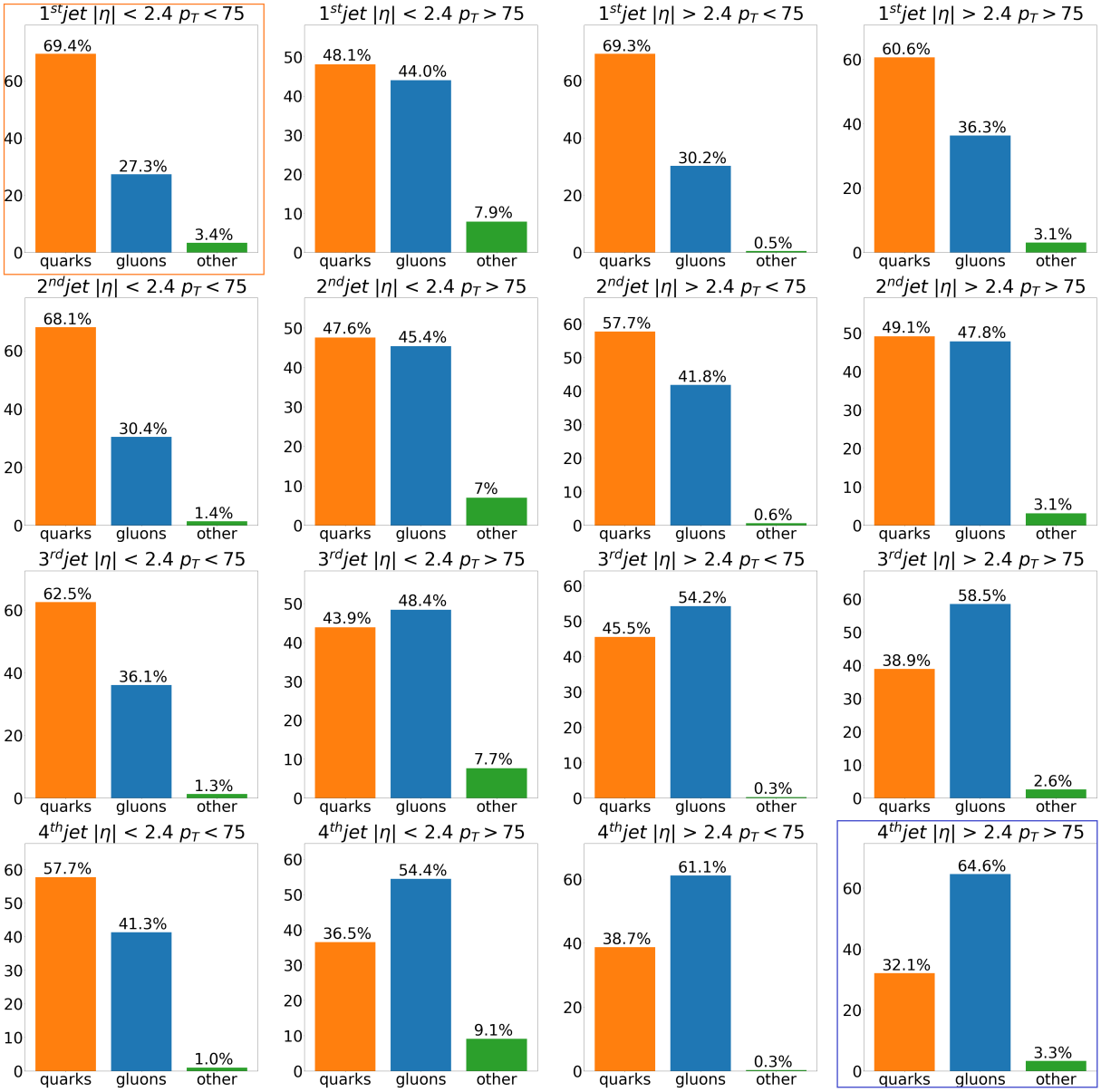
For the global jet type-specific correction, the purest regions in quarks/gluon need to be identified: the four highest p_{T} jets are separated in high/low p_{T} (at a 75 GeV threshold) and

high/low η (with a 2.4 threshold) and then ordered from highest to lowest p_{T}. The regions fractions of gluon and quark jets are shown in Fig. 3.12 for the 2018 data, as well as the purest regions for 2017 and 2016. For 2017, the region at the interface between the endcap and the barrel ($2.3 < |\eta| < 3.5$) is known to have issues: an increased jet multiplicity was reported, creating *horns* in the jet pseudorapidity distribution, leading to degradation of the QGL algorithm performance. This problem originates from an increase of the ECAL noise, dependent on PU and bunch-crossing. In order to avoid being biased by the faulty behavior in this η range, the region was excluded during the morphing functions extraction for 2017. The CDF for gluon and quark are extracted in the corresponding highest purity regions, and a function morphing the MC CDF to the data CDF in this region is computed as

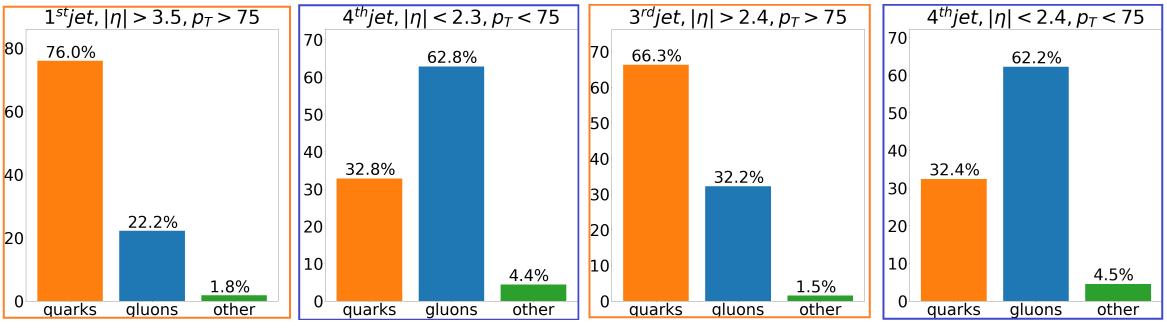
$$f(QGL)^{\text{gluon/quark}} = CDF^{-1}_{\text{data}}(CDF^{\text{gluon/quark}}_{\text{MC}}(QGL)) \tag{3.2}$$

This correction is then applied to all of the corresponding quark/gluon jets.

A second order correction is extracted in a similar way in the same bins of η/p_{T}, but this time without distinguishing the quark and gluon-jets. The final morphing function is computed by composing the first and second order corrections and is extracted independently for each year. The corrected V-jets QGL distributions are shown in Figures 3.13. The morphing is applied before the DNN training so that the model can discriminate according to the corrected distributions.

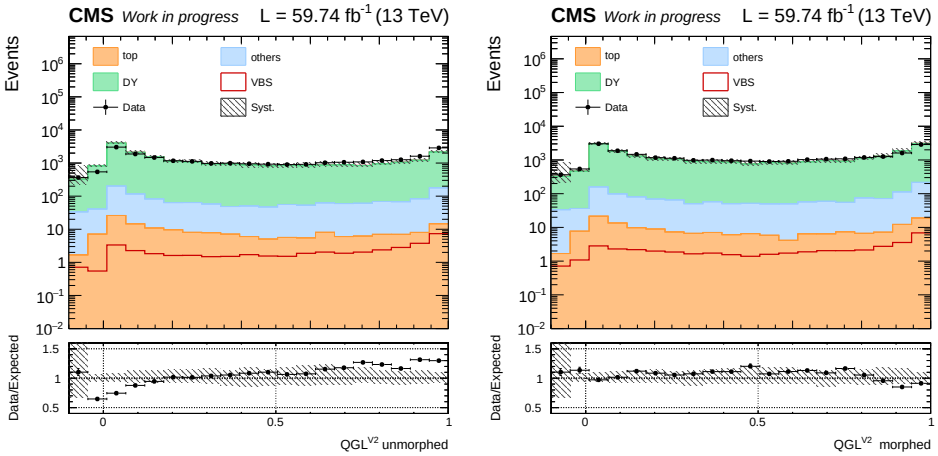


(a) 2018

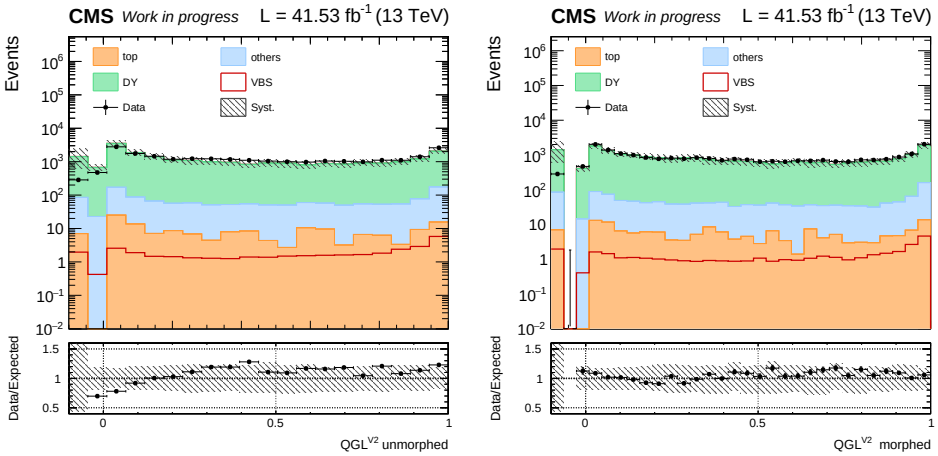


(b) 2017 **(c)** 2016

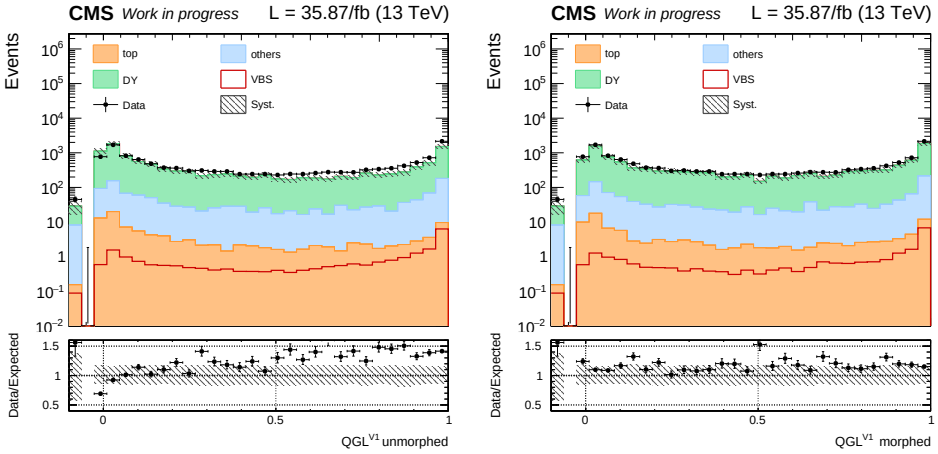
Figure 3.12: (a) : Fractions of the jet population coming from quarks , gluons, or unidentified objects for different bins in jets η and p_{T}, taking 2018 simulated samples as reference. The jets are ranked according to their p_{T}. The purest region for quarks (gluons) is highlighted in orange (blue).
(b) and (c): Bar plots of jet origin in the quark purest region on the left and gluon purest region on the right for respectively 2017 and 2016. The definition of the high and low η regions in 2017 is changed to avoid the $2.3 < |\eta| < 3.5$ region. The purities are of the order of 60-70%



(a)



(b)



(c)

Figure 3.13: Quark-gluon likelihoods for the V jet showing the worst data/simulation behavior before morphing (left) and after the morphing procedure has been applied (right) for the b-vetoed DY control region of (a) 2018 , (b) 2017 and (c) 2016 data. The negative values correspond to jets where the tagger can not be computed.

3.6 Signal extraction

The VBS production of ZV is rare process and its observation at the LHC is a complex challenge. It requires sophisticated techniques to discriminate signal events from the background events that contaminate the SR. A cut-based strategy aiming to improve the $R = S/\sqrt{B+S}$ ratio (where S is the number of signal events and B the number of background events) is insufficient to isolate VBS signal events. The very small cross section of the VBS signal compared to the irreducible backgrounds and their similarity make it extremely hard to obtain statistically significant results that way, as illustrated in Fig. 3.14.

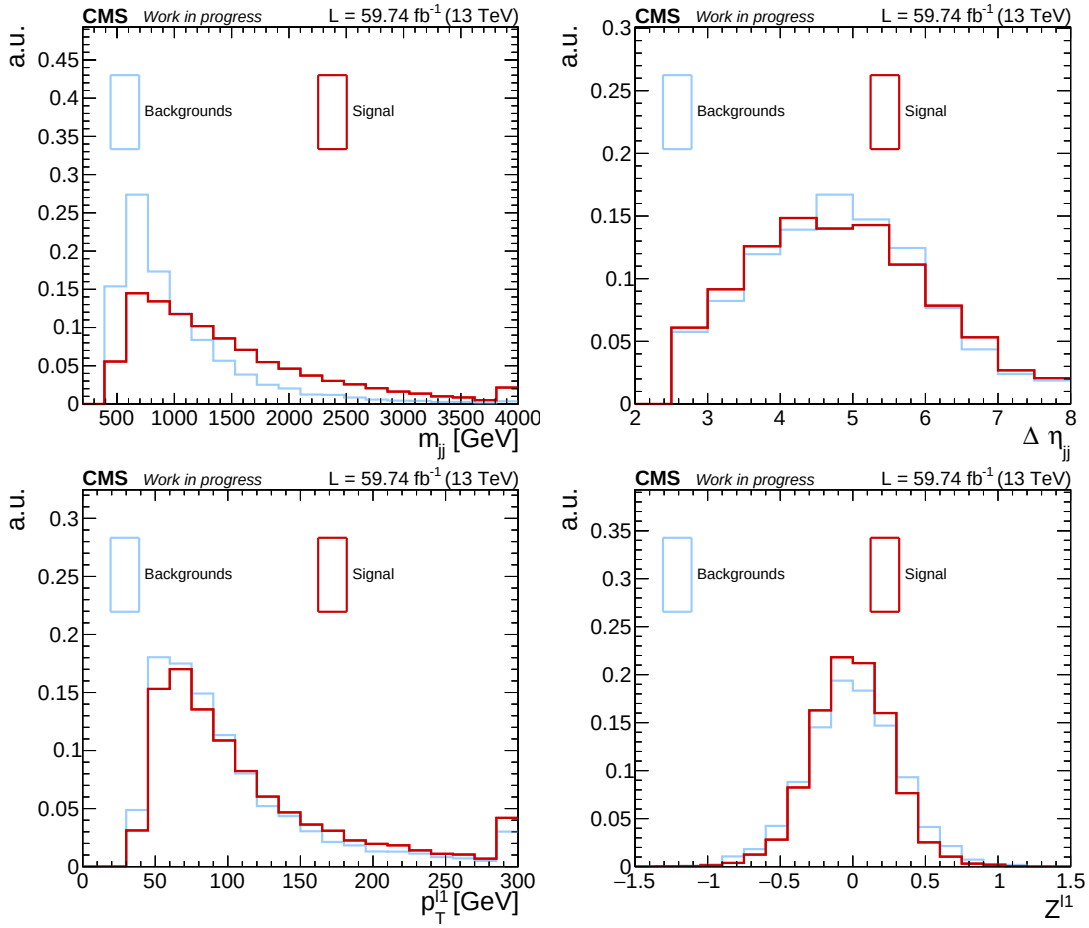


Figure 3.14: Shape comparison of the major backgrounds and signal processes in the SR for 2018 resolved b-vetoed category. Some differences in the distributions for signal and backgrounds are present, but none is sufficient to extract significant results.

A multivariate analysis technique (MVA), based on the aggregation of the most discriminating variables into one single discriminator is thus implemented. Two different types of machine learning (ML) algorithms were evaluated: Boosted Decision Trees (BDT) and artificial Deep Neural Networks (DNN). According to the studies described in this section, the DNN strategy was chosen and further refined.

3.6.1 Model training

A binary classifier is trained per category (Boosted/Resolved $\times$ b-tagged/b-vetoed) to discriminate the signal against all backgrounds at the same time. The training is performed in the same Signal Region described in Section 3.3.

In the resolved category where the amount of simulation data is sufficient, the training is performed separately for each year, allowing the algorithm to be tuned to each year data taking conditions. In the boosted topology however, the MVA model training is performed inclusively on all years data due to the more limited amount of simulation data.

The dataset is made of simulation samples corresponding to the signals and all backgrounds described in Sec. 3.4. A weight coefficient is attributed to each event by taking into account the cross section of the samples and the corrections detailed in Sec. 3.4.2. An overall event reweighting is applied on top to avoid unbalance between the number of signal and background events while conserving the backgrounds relative importance. 80% of the dataset is used for training the models while the rest is used as a testing dataset to monitor that the model generalizes well on unseen events. A common set of variables, showcased in Tab. 3.7, was established and used to compare the efficiencies of BDTs and DNN before being refined.

For both architectures, BDTs and DNNs, a binary logistic objective function (also known as the cross-entropy of the model) is used to evaluate and optimize the model. To minimize the risks of overfitting the model, regularization techniques are employed and two metrics are monitored during the training:

- The **loss** of the model, which compares the predicted probability to actual truth value of each event. The loss of the model on the training samples shouldn't continue to decrease if the loss on the testing samples is stable or increases. Such a case would show that the model is becoming too finely tuned to the train sample and won't generalize well.
- **ROC AUC**. The Receiver Operating Characteristic (ROC) curve is created by plotting for various threshold settings the number of signal events correctly labeled, also called true positives or signal efficiency, against the number of background events incorrectly labelled as signal, also known as false positive or background efficiency. Different interpretations of this ROC curve can also be obtained by using the background reduction computed as $1 -$ background efficiency instead of the false positive rate. The Area Under the Curve (AUC) measures the area under this curve that should be maximized by a well-performing model. Similarly to the loss, in the absence of overtraining, the AUC for the training and testing samples should not diverge.

A last overfitting check is performed after the optimization procedure by using a Kolmogorov-Smirnov (KS) test [107]. This statistical test measures the similarity of the model output between the training and testing datasets. The trained model is validated only if the KS test p-value is over 5%, which indicates that there is no significant differences between the distributions. If a model fails this requirement, the optimization procedure is performed once again with increased regularizations.

An optimization of the model's hyperparameters is performed to maximize a scoring func-

Input variable	Description	Resolved	Boosted
Leptons			
$p_{\mathrm{T}}^{\ell_1}$	Leading lepton transverse momentum	✓	✓
$p_{\mathrm{T}}^{\ell_2}$	Subleading lepton transverse momentum	✓	✓
η_{ℓ_1}	Leading lepton pseudorapidity	✓	✓
η_{ℓ_2}	Subleading lepton pseudorapidity	✓	✓
$m_{\ell\ell}$	Leptonically decaying Z invariant mass	✓	✓
Z_{ℓ_1}	Zeppenfeld $Z_{\ell_1} = (\eta_{\ell_1} - \bar{\eta})/\Delta\eta_{jj}$ [1]	✓	✓
Z_{ℓ_2}	Zeppenfeld $Z_{\ell_2} = (\eta_{\ell_2} - \bar{\eta})/\Delta\eta_{jj}$ [1]	✓	✓
VBS jets			
p_{T}^{VBS1}	Leading VBS jet transverse momentum	✓	✓
p_{T}^{VBS2}	Subleading VBS jet transverse momentum	✓	✓
η^{VBS1}	Leading VBS jet pseudorapidity	✓	✓
η^{VBS2}	Subleading VBS jet pseudorapidity	✓	✓
QGL^{VBS1}	Leading VBS jet quark-gluon likelihood	✓	✓
QGL^{VBS2}	Subleading VBS jet quark-gluon likelihood	✓	✓
m_{jj}	VBS dijet invariant mass	✓	✓
$\Delta\eta_{jj}$	VBS dijet η separation	✓	✓
$\Delta\phi_{jj}$	VBS dijet ϕ separation	✓	✓
V-jets			
p_{T}^{V1}	Leading V jet transverse momentum	✓	
p_{T}^{V2}	Subleading V jet transverse momentum	✓	
η^{V1}	Leading V jet pseudorapidity	✓	
η^{V2}	Subleading V jet pseudorapidity	✓	
QGL^{V1}	Leading V jet quark-gluon likelihood	✓	
QGL^{V2}	Subleading V jet quark-gluon likelihood	✓	
m_V	Reconstructed V jet invariant mass	✓	✓
$p_{\mathrm{T}J}$	AK8 jet momentum		✓
η_J	AK8 jet pseudorapidity		✓
Z_J	Zeppenfeld of the AK8 jet $Z_J = (\eta_J - \bar{\eta})/\Delta\eta_{jj}$ [1]		
All jets			
n_j^{30}	Number of jets with $p_{\mathrm{T}} > 30$ GeV	✓	✓
n_j^{btag}	Number of b-tagged jet	✓	✓

[1] $\bar{\eta} = (\eta^{\mathrm{VBS1}} + \eta^{\mathrm{VBS2}})/2$

Table 3.7: The list of input variables used for the MVA model training. A checkmark in the Resolved or Boosted column indicates that the variable is used in the corresponding topology. Lowercase j indices denote AK4 jets while uppercase J refer to AK8 jets. The AK8 jet Zeppenfeld was not used for the comparison of BDTs and DNNs

tion defined as the testing sample ROC AUC penalized by the gap between test and train AUC:

$$S = AUC_{\text{test}} - \alpha \times |AUC_{\text{test}} - AUC_{\text{train}}| \tag{3.3}$$

Where α is an arbitrary coefficient which value can be increased to improve the overtraining control. A value of $\alpha = 1$ was found to be enough for the trained model to not exhibit characteristic signs of overtraining described above. The details of the hyperparameters optimization procedures differ between the two architectures and are precised in the next section.

3.6.2 Architectures

For the BDT approach, the EXtremGradientBoosting (XGBoost) library[79] is selected as an efficient implementation of the gradient boosting algorithm. The overtraining is controlled by L1 and L2 regularizations and by using only 80% of randomly chosen training events for each tree and 80% of the columns. The hyperparameters of the BDT, namely the maximal tree depth, learning rate and both regularizations parameters, are optimized via grid search. During this optimization, the maximum number of boosting rounds is set to 400 and an early stopping is set to stop the training if the loss doesn't improve for 5 consecutive boosting steps. To allow the full usage of the training dataset for the training of the model with the optimised hyperparameters, each iteration of the grid search was performed with a 5-fold cross-validation. The cross validation (CV) is a technique that iteratively uses different portions of the training samples to train and validate a model. In the case of a 5-fold CV, the training sample is divided into five subsamples, and the model is trained five times by using four of the subsamples and testing on the remaining one. The CV then averages the performance of the five models and the optimal maximum tree depth,learning rate and regularization parameters are identified. A new 5-fold CV is performed with those parameters and is used to extract the optimal number of boosting rounds as the number of rounds maximizing the scoring function described in Eq 3.3. The evolution of the metrics during one cross validation is showcased in Fig. 3.15 and the values of the optimized hyperparameters are reported in Tab. 3.8.

Parameter	Optimization range	Optimum (Resolved)	Optimum (Boosted)
Max tree depth	4-10	4	3
Learning rate	[0.01,0.05,0.1,0.2,0.3]	0.1	0.1
Subsample ratio of the training instances	not optimized	0.8	0.8
Subsample ratio of columns	not optimized	0.8	0.8
L1 regularization	[0.01, 0.1,1,10,15]	10	10
L2 regularization	[0.01, 0.1,1,10,15]	10	15
Number of boosting rounds	[0-800]	186	129

Table 3.8: Optimized values for BDT hyperparameters

In Fig. 3.16 are presented the performance obtained by the final training on 2018 simulation data and in b-vetoed Resolved and Boosted categories. The ROC AUC is used as the figure of merit. The histogram showing the BDT output for the signal and backgrounds is also shown.

Artificial Deep Neural Networks (DNN) are also evaluated to discriminate the signal. Since the problem is a simple classification task, the architecture chosen was a MultiLayer Percep-

tron (MLP), also known as fully connected network. It is trained with the Keras [108] framework and TensorFlow2[109] backend.

The model is trained with Adam [81], a gradient descent optimizer that dynamically updates the learning rate on a per-parameter basis. The dense layers are built with the Rectified Linear Unit [110] (ReLU) activation function. To avoid overfitting, L1 and L2 kernel regularizations are applied. A dropout[111] function is added to each dense layer, making the NN forget a given fraction of weights at every epoch to avoid overtraining. Additionally, *early stopping* with a patience $P = 5$ was used, meaning that the network training is stopped when the loss of the test model does not decrease for $P = 5$ consecutive generations. The output layer is a single neuron activated by a sigmoid function giving an output in the [0,1] range. The closer the output is to 1, the more *confident* is the network that the event is part of the signal.

The NN has a larger number of hyperparameters to optimize compared to the BDT, namely the regularizations parameters (L1, L2 and dropout rate), the initial learning rate and numbers of neurons and layers, and can be computationally more expensive to train. For those reasons, the optimization of the hyperparameters with a grid search would be ineffective and a Bayesian Optimisation[112] (BO) based on Gaussian Processes (GP), known to converge faster, is employed. The Matern [113] kernel was chosen for the GP and the Upper Confidence Bounds [82] (UCB) method was chosen for the acquisition function with a parameter $\kappa = 2.5$. The acquisition function determines how much the BO will explore, i.e. look in region of the phase space not sampled before, and exploit, i.e. relies on the results of previously sampled points.

In Fig. 3.17 are presented the performance obtained by the final training on 2018 simulation data and in b-vetoed Resolved and Boosted categories. The ROC AUC is used as the figure of merit. The histogram showing the DNN output for the signal and backgrounds is also shown.

Category	BDT	DNN
2018 Resolved b-vetoed (test)	85.0%	85.3%
2018 Resolved b-vetoed (train)	86.2%	86.6%
2018 Boosted b-vetoed (test)	81.0%	81.9%
2018 Boosted b-vetoed (train)	84.1%	83.8%

Table 3.9: ROC AUC values for BDT and DNN trained on 2018 data after optimisation of both models.

The results of the comparisons between the BDT and DNN performance on the 2018 Resolved and Boosted b-vetoed categories in the simulated data are reported in Tab.3.9. The values of the AUC are very similar between the two techniques, but the DNN was chosen due to the better overfitting regulation it offered. The size of the model can be more finely controlled (number of neurons per layer and number of layers versus number of trees with a given maximum depth) and the dropout function tuning offers additional control.

3.6.3 Neural network refinement

The DNN training was further refined in order to improve its performance. The training was split into eight different models : six models corresponding to the resolved category for each

years in b-vetoed/b-tagged sub-regions, and two models corresponding the boosted b-vetoed and b-tagged regions with all years included.

The baseline set of features described in Tab. 3.7 is comprehensive, but a high number of features can increase the DNN sensitivity to discrepancies between the simulated and real data. A pruning strategy is implemented to remove the two less impactful variables by ranking them according to their importance. The feature importance is evaluated in terms of SHAP values [114], extracted with a model explainability method based on game theory. Those values are a representation of how much each feature contributes to steer the DNN output towards signal or background on an event-by-event basis. An example of this SHAP values ranking is shown in Fig. 3.18.

The two least important features are removed from the input set before retraining the model. This is done iteratively, and the evolution of the AUC depending on the number of features in the input set is shown in Fig. 3.19. The optimal pruned input set is selected as a trade-off between minimal number of features and minimal loss of performance. The full list of features and their importance in the pruned input set are reported in Tab. 3.11 for each model and the five most important variables are detailed in Table. 3.10. In the resolved topology, the same set of variables is used for the training of the model on each year samples.

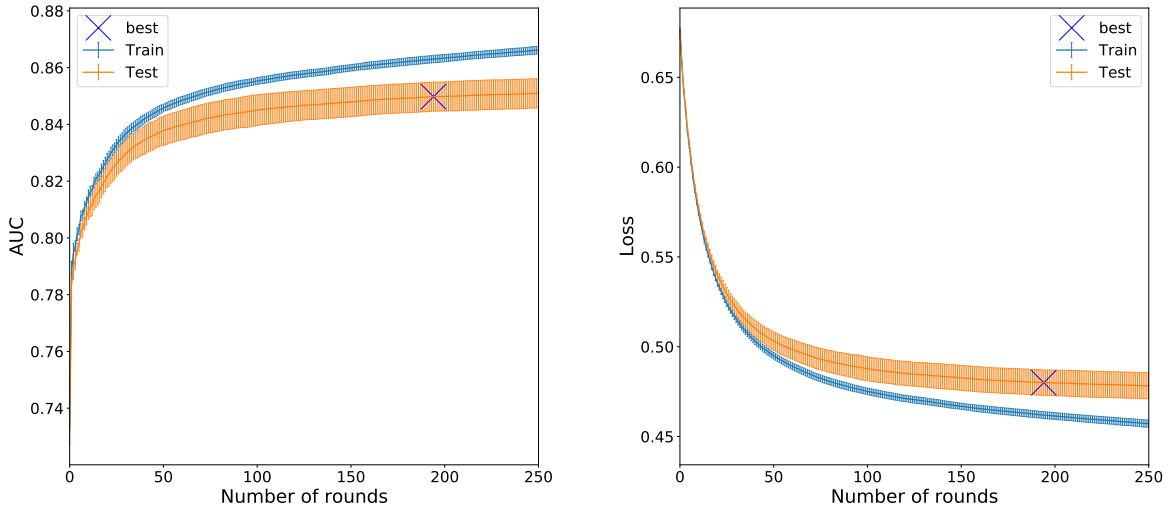
Category		Feature rank				
		1	2	3	4	5
Resolved	b-vetoed	m_{jj}	n_j^{30}	$\Delta\eta_{jj}$	Z_{ℓ_1}	QGL^{V1}
	b-tagged	n_j^{30}	m_{jj}	QGL^{V1}	Z_{ℓ_1}	$\Delta\phi_{jj}$
Boosted	b-vetoed	m_{jj}	n_j^{30}	$p_{\mathrm{T}J}$	Z_{ℓ_1}	$p_{\mathrm{T}}^{\mathrm{VBS2}}$
	b-tagged	n_j^{30}	m_{jj}	Z_{ℓ_1}	$p_{\mathrm{T}}^{\mathrm{VBS2}}$	n_j^{btag}

Table 3.10: The five most important variables for each model, ranked according to their SHAP values.

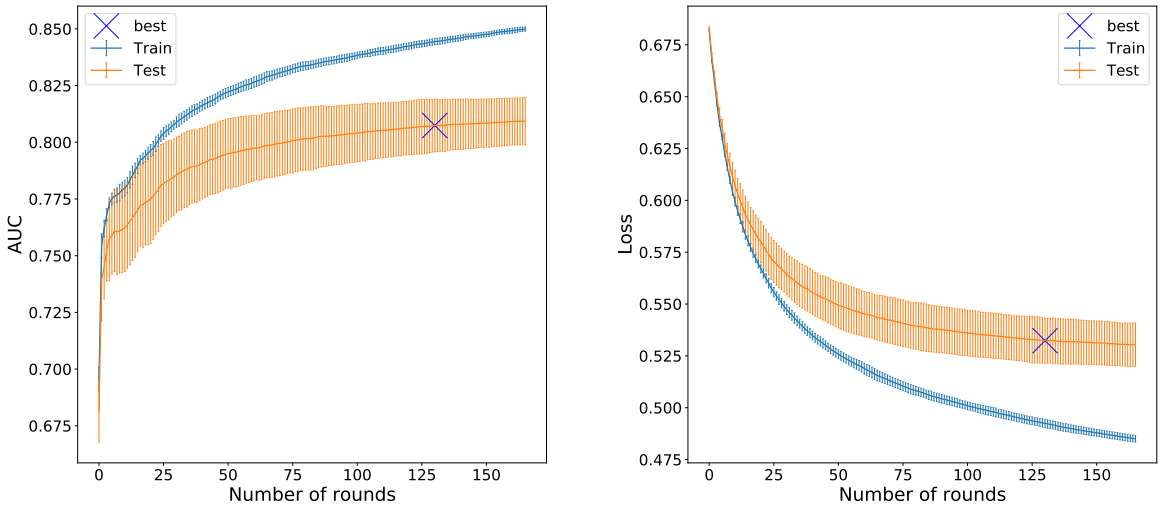
The hyperparameters for the pruned input set are optimized according to the same method as detailed in the previous section, and their value are reported in Tab 3.12. The resulting ROC curves and output shown in Fig. 3.20 and Fig. 3.21.

Feature	Category			
	Resolved		Boosted	
	b-vetoed	b-tagged	b-vetoed	b-tagged
$p_{\mathrm{T}}^{\ell_1}$	-	✓	✓	✓
η_{ℓ_1}	-	-	✓	-
η_{ℓ_2}	-	-	✓	-
$m_{\ell\ell}$	-	✓	-	-
Z_{ℓ_1}	✓	✓	✓	✓
Z_{ℓ_2}	✓	-	✓	✓
$p_{\mathrm{T}}^{\mathrm{VBS2}}$	✓	✓	✓	✓
η^{VBS1}	-	-	-	✓
η^{VBS2}	-	-	✓	✓
m_{jj}	✓	✓	✓	✓
$\Delta\eta_{jj}$	✓	✓	✓	✓
$\Delta\phi_{jj}$	✓	✓	✓	-
n_j^{30}	✓	✓	✓	✓
p_{T}^{V1}	✓	✓	-	-
p_{T}^{V2}	✓	✓	-	-
η^{V1}	✓	✓	-	-
η^{V2}	✓	✓	-	-
QGL^{V1}	✓	✓	-	-
QGL^{V2}	✓	-	-	-
$p_{\mathrm{T}J}$	-	-	✓	✓
η_J	-	-	✓	✓
Z_J	-	-	✓	✓
m_V	✓	✓	-	-
n_j^{btag}	-	-	-	✓

Table 3.11: The list of input features used for each of the models. Only the variables used for the training of at least one model are shown.

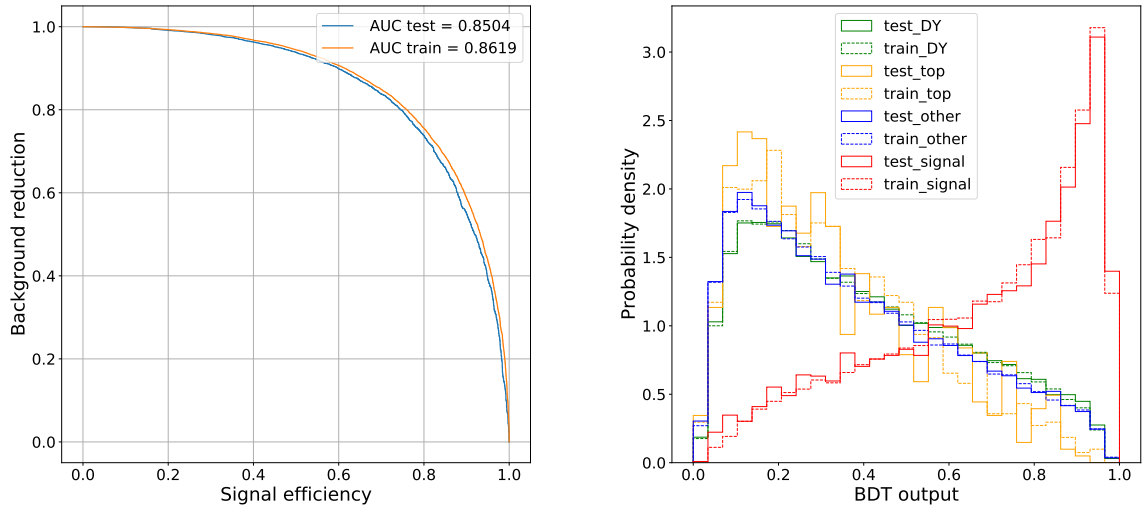


(a)

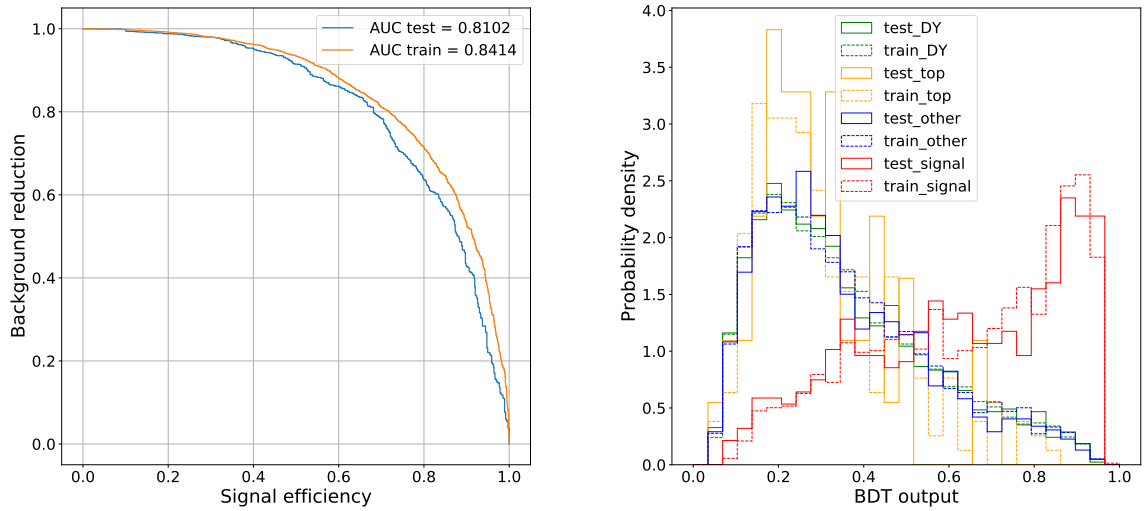


(b)

Figure 3.15: Evolution of the BDT metrics during cross validation for (a) Resolved and (b) Boosted categories. Left: the Area Under the ROC Curve. Right: the logarithmic loss of the model. The "best" point (black cross) corresponds to the optimum of the scoring function and gives the optimal number of boosting rounds. The error bars correspond to the variability inside of the five-fold cross-validation.

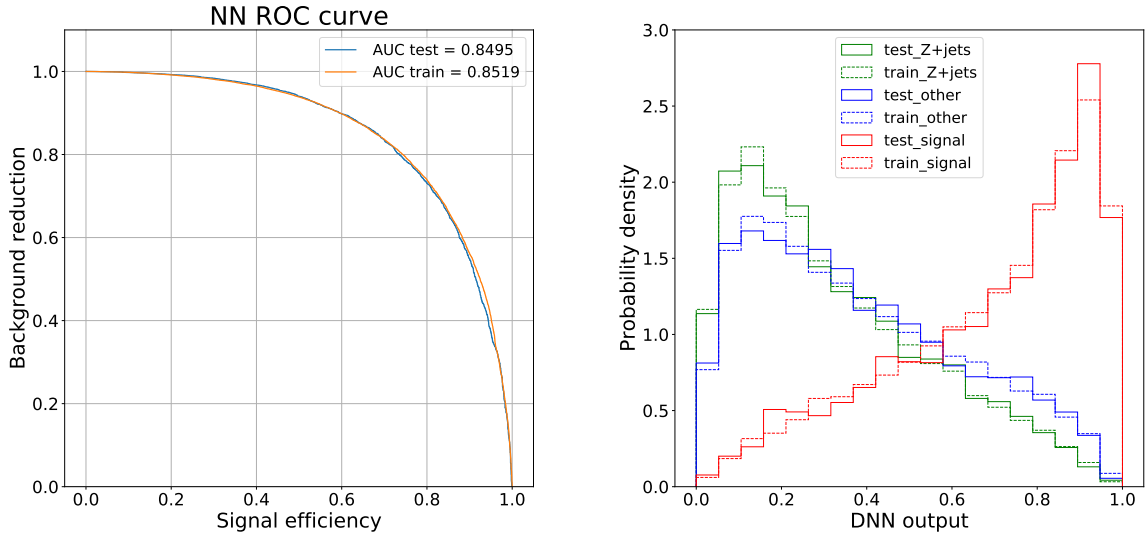


(a)

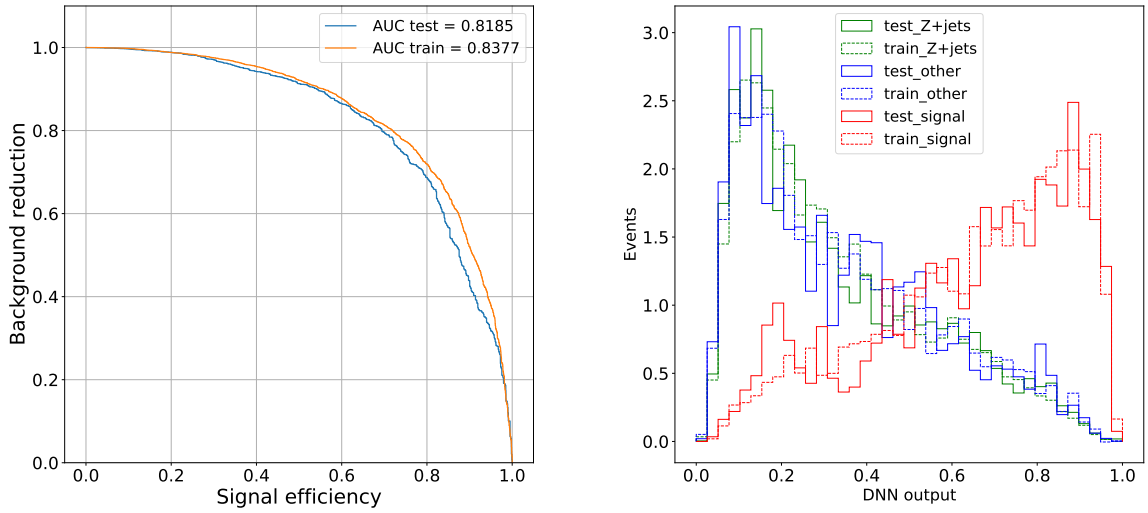


(b)

Figure 3.16: The BDTs results for 2018 (a) Resolved category and (b) Boosted category. The ROC curve is shown on the left and the BDT output on the right, with training output is shown in full lines and the test set output is in dashed line.



(a)



(b)

Figure 3.17: The DNNs results for 2018 (a) Resolved category and (b) Boosted category. The ROC curve is shown on the left and the DNN output on the right, with the training output is shown in full lines and the test set output is in dashed line.

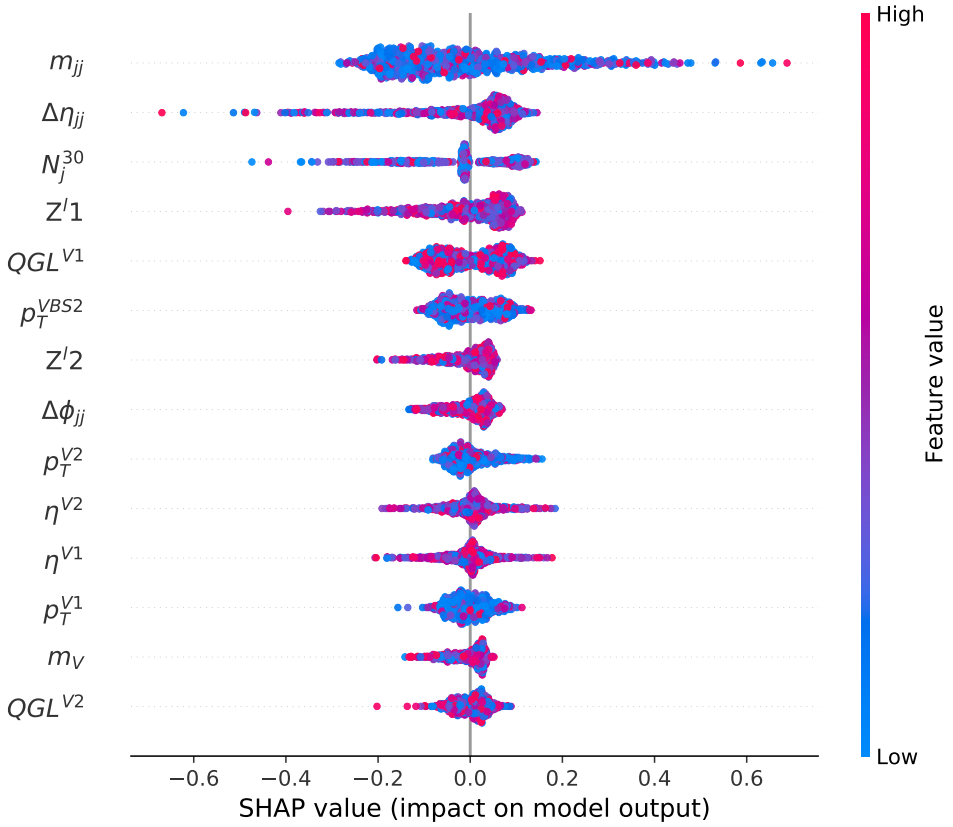


Figure 3.18: Example of SHAP values ranking for a trained model (here resolved b-vetoed on 2018 data). Each row corresponds to an input variable, ordered by average importance, with each point corresponding to one event. The higher a value the more positive the impact is on the output, meaning the event classification will get closer to 1. The color of the point corresponds to the value of the feature. There is no clear correlation pattern between a single feature value and the model output, which emphasizes the usefulness of multivariate methods.

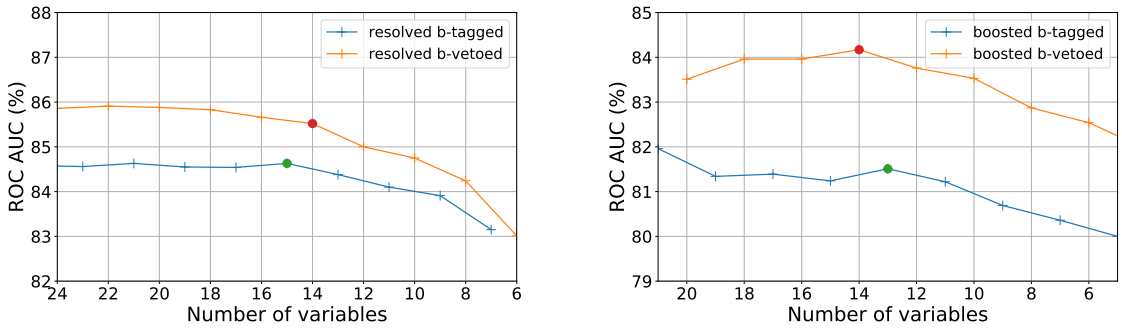
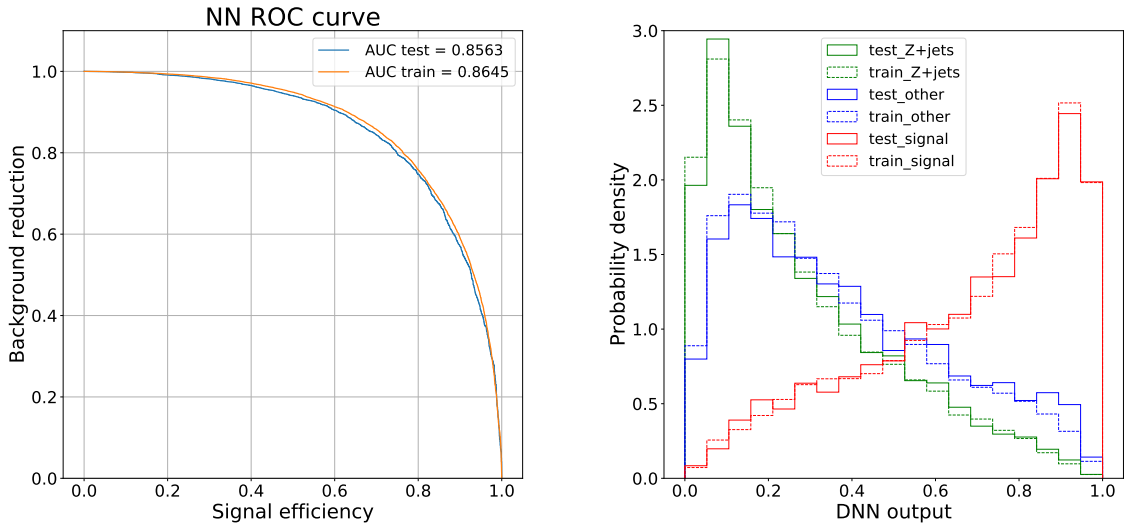
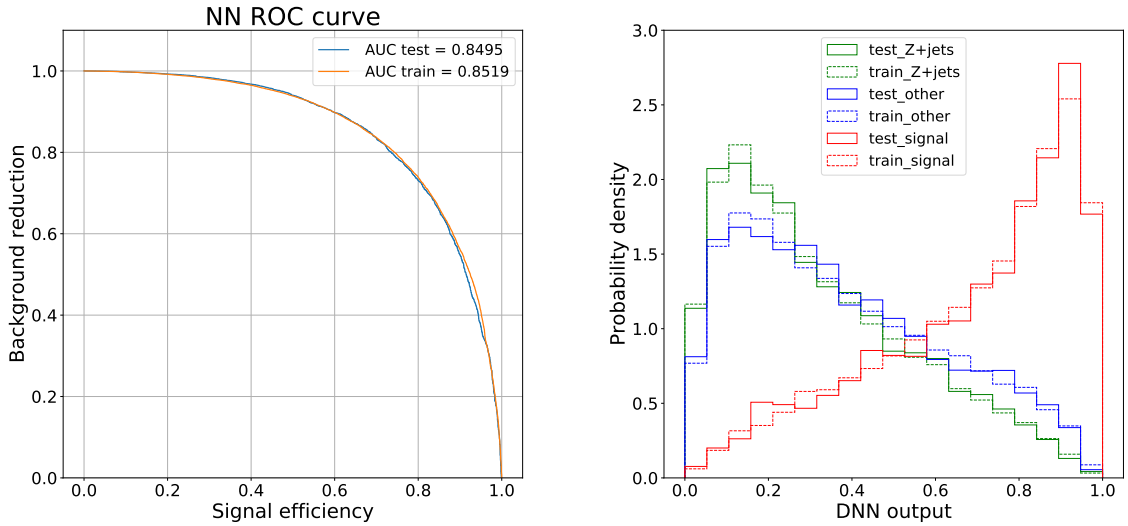


Figure 3.19: Pruning curves for Resolved (Boosted) b-vetoed category and b-tagged on the left (right). The points highlighted in red (green) show the compromise chosen as the smallest network before significant loss of performance.

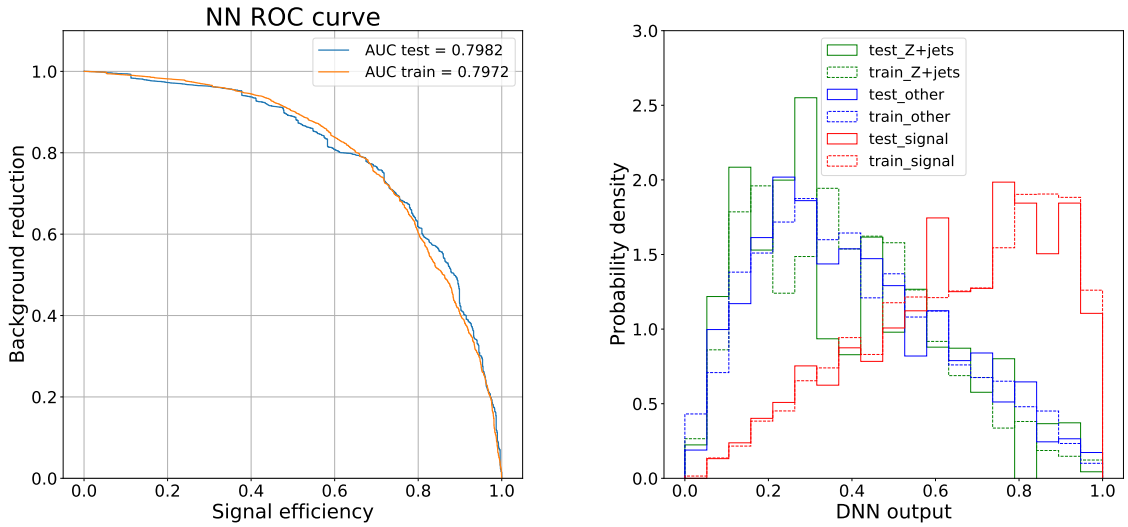


(a)

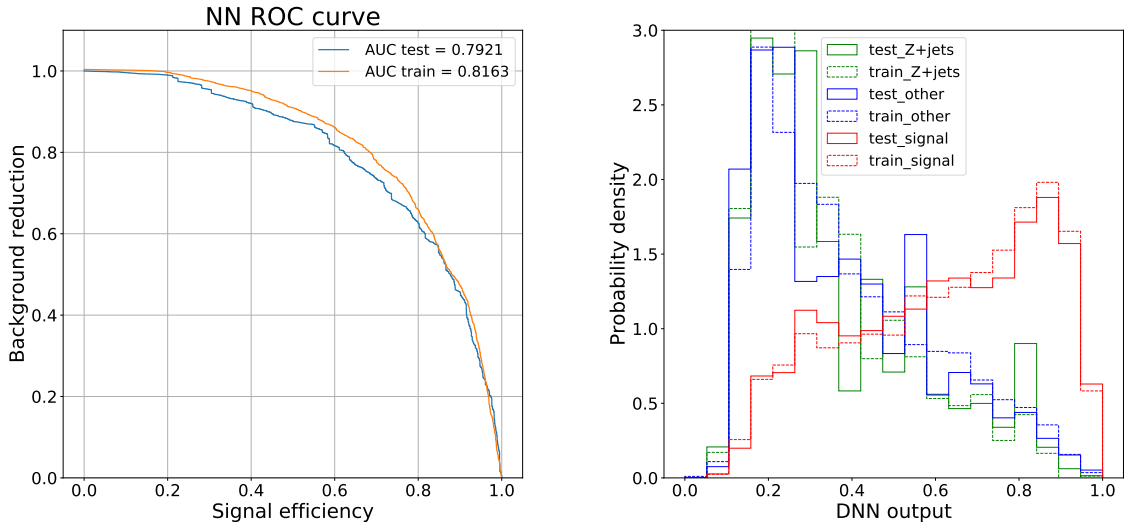


(b)

Figure 3.20: The ROC curves (left) and output (right) for the DNNs trained on 2018 Resolved SR (a) b-vetoed and (b) b-tagged simulations using the pruned input set. Both signal and backgrounds KS test are above $p = 0.05$.



(a)



(b)

Figure 3.21: The ROC curves (left) and output (right) for the DNNs trained on all years combined Boosted SR (a) b-vetoed and (b) b-tagged simulations using the pruned input set. Both signal and backgrounds KS test are above $p = 0.05$.

Parameter	Optimization range (step)	Resolved						Boosted	
		2018		2017		2016		All years	
		b-veto	b-tag	b-veto	b-tag	b-veto	b-tag	b-veto	b-tag
Number of hidden layers	[1,5] (1)	3	2	2	2	2	2	1	2
Neuron per hidden layer	[20, 400] (10)	280-200-150	270-100	270-100	270-100	200 - 100	190 - 110	200	80 - 80
Dropout rate	[0, 50]% (5%)	30%	45%	30%	45%	30%	40%	10%	20%
Initial learning rate	$[10^{-4}, 10^{-3}]$ (10^{-4})	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}	6.10^{-4}	1e-4
L1 kernel regularization	$[0, 10^{-3}]$ (10^{-4})	0	0	0	0	0	0	0	0
L2 kernel regularization	$[10^{-4}, 10^{-2}]$ (10^{-4})	10^{-4}	5.10^{-3}	8.10^{-3}	5.10^{-3}	5.10^{-3}	5.10^{-3}	9.10^{-2}	7.10^{-2}
Max number of epochs	not optimized	400							
Early stopping patience	not optimized	256							
Batchsize	not optimized	5							

Table 3.12: Results of the bayesian optimization for all DNN models hyperparameters. For each hyperparameter the optimization range and the size of the steps are indicated.

3.7 Systematic uncertainties

The measurement of the significance of the VBS production of ZV in the semi-leptonic channel is affected by multiple sources of uncertainties, represented by nuisance parameters in the signal extraction fit. The most obvious is the statistical one coming from fluctuations depending on the amount of data taken; but other sources are also taken into account. The list of those systematics considered in the analysis are as follow:

3.7.1 Uncertainties affecting all simulations

All those uncertainties are considered uncorrelated between the three years of the Run 2 unless stated otherwise:

- **Luminosity uncertainty** : The uncertainty determined by the CMS luminosity monitoring is 1.2% [115], 2.3% [116], and 2.5% [117] for 2016, 2017, and 2018 data sets, respectively.
- **Lepton uncertainties**: uncertainties on the lepton reconstruction and identification efficiency, as well as the momentum and its scale. The Lepton reconstruction and identification efficiencies effect is estimated by varying the weights by $\pm 1\sigma$. The lepton momentum scale and resolution systematic is propagated to the analysis phase space by varying the lepton momenta and scale resolution factors and smearing by $\pm 1\sigma$.
- **Jet uncertainties**: this includes uncertainties on the Jet Energy scale (JES) and resolution (JER) , as well as on the b-tagging efficiencies and pileup identification. For the AK8 jets, the uncertainty on the mass scale and resolution is also taken into account.
  - JES: the effect is estimated by shifting the JES by $\pm 1\sigma$. The JES uncertainty is composed of several components with most of them correlated between years. The effect on the SR rates are of the order of 10%.
  - JER: The effect of the JER uncertainty is estimated by smearing the jet energy scale by $\pm 1\sigma$ variations, with a fully correlated pattern between the different years.
  - b-tagging SF: the uncertainties on the b-tagging SF are evaluated by shifting the SF on a per-jet basis by $\pm 1\sigma$. There are 10 components corresponding to different sources of the SF measurement uncertainty: four variations concerning the statistical uncertainties if the measurement samples that are uncorrelated between the years; and six correlated variations related to the lack of knowledge on the jet flavor compositions of the samples.
  - Jet pileup identification (PU ID) SF: The uncertainty on the jet PU ID SF is applied on all the jets with $p_{\mathrm{T}} > 30$ GeV and is uncorrelated between the years.
  - AK8 jet JES and JER: similar estimation as for the regular AK4 jets. The inclusive set of sources is used. This uncertainty is only applied in the boosted category.
  - AK8 mass scale and resolution: the soft-drop PUPPI mass is scaled to $\pm 1\sigma$ variations . The effect is of the order of less than 1%.

- **Trigger efficiency**: uncertainty of the order of less than 1%. These uncertainties are extracted by varying the tag selection and Z window in the tag and probe method employed to compute the scale factors. Their effect is propagated by shifting the scale factors by $\pm 1\sigma$.

- **Pileup reweighting**: the uncertainties on the amount of PU present in the event are estimated by varying the minimum bias cross section used to generate the PU distributions by one standard deviation. This uncertainty is considered uncorrelated between the years.

- **Prefiring**: Uncertainty associated with the prefiring corrections in 2016 and 2017. It is calculated as one standard deviation up and down of the event weight. The uncertainty is uncorrelated between 2016 and 2017.

3.7.2 Background estimation related uncertainties

- **DY background** : the DY shape is corrected by the implementation of bin-dependent normalization parameters as described in Section 3.5.1. Other sources of uncertainties such as the QCD scale can have large effects on the normalization. As such, their variation for each bin of the Z + jets sample is renormalized to give the overall event yield precision identical to the nominal prediction. This is done, for a given category, by summing the contribution in control region and the signal region. This way, both the shape corrections and migration effects between the regions are modelled, but the overall normalization can be freely fitted.

- **Top background** : The data-driven normalization on the top processes is implemented by treating the $t\bar{t}$ and single top processes as a single background. An uncertainty on the shape and normalization is assessed in a per-event basis with weights corresponding to the up and down variations of the renormalization and factorization scales by a factor two and are correlated between the three years. The p_{T} dependant reweightings described in Section 3.4.2 are another source of uncertainty which is treated as uncorrelated. A last nuisance parameter relative to the data-driven correction of the overall normalization of the top backgrounds is also taken into account in a similar way as for the DY background.

- **non-prompt background** : Non-prompt background is modeled by applying weights (fake factors) to signal candidate events identified by leptons passing the loose selection criteria, as described in Section **??**. Systematic uncertainties in fake factors arise from the limited size of the samples used to measure them and the difference in the flavor composition of the jets faking the leptons between the measurement sample and the signal region. The maximum deviation of the fake factors from the nominal values are of order 5 to 10% for both sources. There is, however, a limitation in capturing the full effect of jet flavor composition difference with the fake factors. Therefore, a conservative 30% normalization uncertainty is additionally assigned to the fake background prediction, independently for each source of non-prompt background. Uncertainties on the fake factors are uncorrelated among the three data sets.

3.7.3 Theoretical uncertainties

Since the analysis makes extensive usage of MC simulated data, several uncertainties arise from the the modelling of physical processes used for event generation. The sources of theoretical uncertainties considered are detailed bellow, with all of them considered correlated between years.

- **Parton distribution functions (PDF)**: The uncertainty on the PDF used for MC modelling is estimated by reweighting the simulated data with variations of the default PDF set.

- **Higher-order corrections**: The generation of the MC is tuned with several parameters including the renormalization and factorization scale. An uncertainty resulting from the choice of those scales is estimated by considering simultaneous variations of those parameters by factors 0.5 and 2.

- **Parton shower modeling**: the uncertainty on parton shower modeling affects mainly the jet multiplicity. Its computed by using weights corresponding to per-event variations of the initial and final states radiations scales between 0.5 and 2. The envelopes of the distributions obtained by varying those scales are used as up and down uncertainties.

- **Underlying event modeling**: uncertainty in the UE is evaluated by using variations of the nominal UE tunes of PYTHIA8 (CUETP8M1 for 2016 and CP5 for 2017 and 2018).

All the nuisance parameters corresponding to the different sources of systematics are reported in Tab. 3.13.

3.7.4 Impact plots

A way to visualize how each uncertainty affects the fit is through the so-called *impact plots*. The impacts are evaluated by taking the best-fit value of each nuisance parameter $\hat{\theta}$ and its 68% confidence level interval, and then performing a new maximum likelihood fit while fixing the parameter value to $\theta_{\pm} = \hat{\theta} \pm \sigma_{\theta}$. The impact on the signal strength is then evaluated by computing the change in the best fit value compared to the global minimum $\hat{\mu}$: $I_{\mu}^{\pm} = \hat{\mu}_{\pm} - \hat{\mu}$. The best-fit and impact values of the 60 most important nuisances are summarized in Fig. 3.22. The parameters are ordered by their impacts.

The systematics that have the largest impact on the significance are the QCD scale for the QCD diboson production associated with jets, the prefiring effect in 2017, the QCD scale for the vector boson fusion production of a V boson, the Parton Distribution Function for 2016 and the QCD scale for the VBS signal. Those uncertainties are mostly theoretical and can not be lowered without improvement on the models used for MC generation. The uncertainty introduced by the rate parameter used for the normalization of the main background also has a significant impact on the signal strength uncertainty.

Uncertainty	process	type	correlation
Integrated luminosity	All MC except Z+jets and tops	rate	partially correlated
Trigger efficiency	all MC	shape	uncorrelated
Lepton efficiency	all MC	shape	uncorrelated
Lepton momentum scale	all MC	shape	uncorrelated
Prefiring	all MC (only 2016 and 2017)	shape	uncorrelated
Fake rate	all MC	shape	correlated
Pileup reweigthing	all MC	shape	uncorrelated
Jet pileup ID	all MC	shape	uncorrelated
AK4 jet energy scale	all MC	shape	partially correlated
AK4 jet energy resolution	all MC	shape	uncorrelated
AK8 jet energy scale	all MC	shape	uncorrelated
AK8 jet energy resolution	all MC	shape	uncorrelated
AK8 jet mass scale	all MC	shape	uncorrelated
AK8 jet mass resolution	all MC	shape	uncorrelated
b-tagging scale factor	all MC	shape	partially correlated
Single top $t\bar{t}$ composition	tops	shape	correlated
top p_{T} reweighting	tops	shape	correlated
Parton shower ISR and FSR	all MC	shape	correlated
Underlying event	all MC except Z+jets, tops	rate	correlated
QCD scale	all MC	shape	correlated
Parton distribution function	all MC	shape	2017-2018 correlated

Table 3.13: Summary of the nuisance parameters used to model the uncertainties and correlation between years.

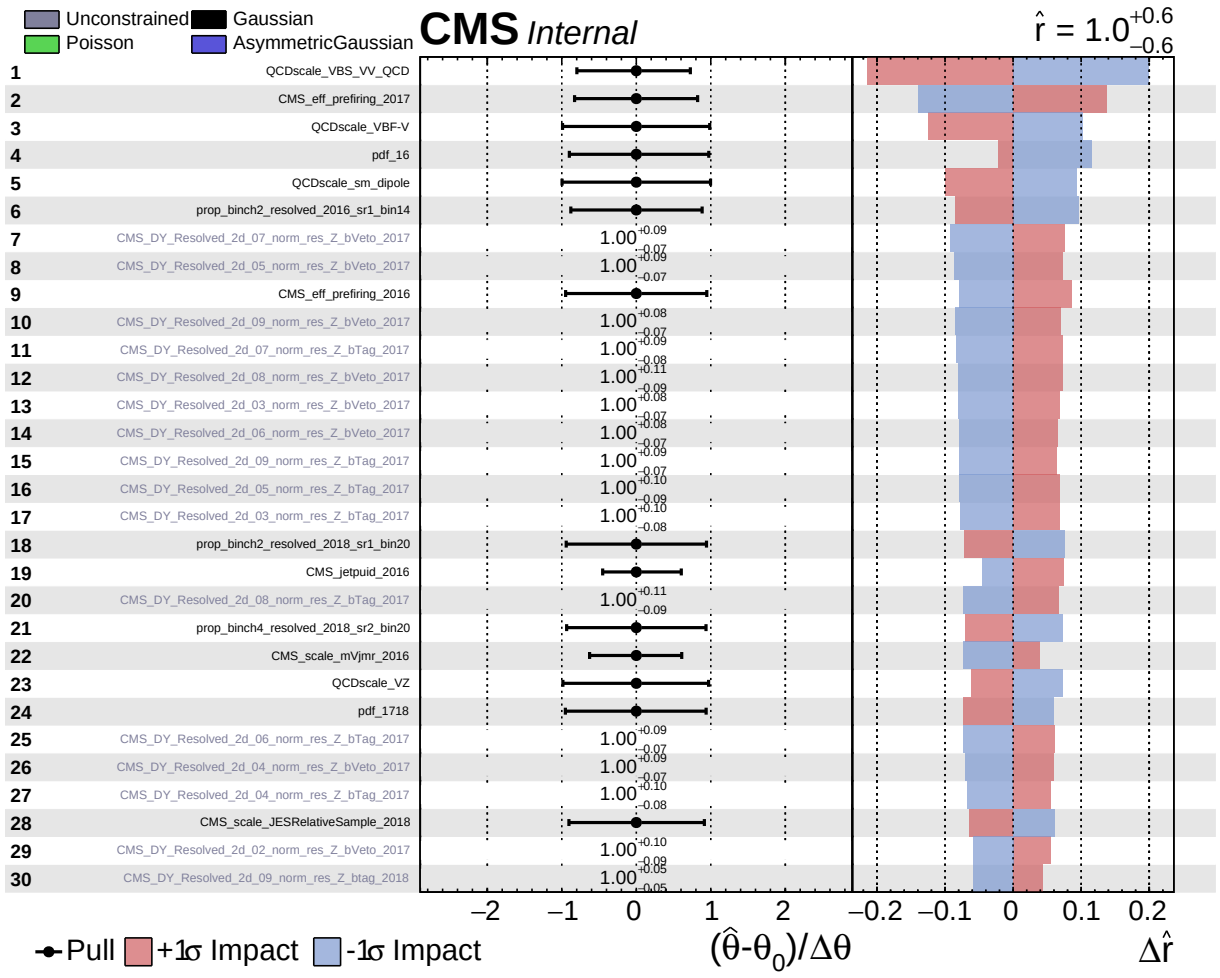


Figure 3.22: Impact plots for the statistical analysis of the Full Run 2 dataset with signal + background hypothesis for the 30 most important systematics. The plots are obtained using the Combine tool[118]. The parameter $\hat{r}$ corresponds to the signal strength $\hat{\mu}$. Though the normalization of the Z+jets and top backgrounds are corrected, the impact show a rate parameter of 1.0 due to a limitation of the framework when computing the impacts on blinded data.

3.8 Results

3.8.1 Statistical approach

The observation of a signal in a given dataset requires that the hypothesis where only the backgrounds are present can be statistically rejected. Statistics test are performed to measure the inconsistency of the data with this background-only hypothesis also known as null hypothesis H_0. The *p-value* measuring this probability is computed as

$$p = \int_t g(t|H_0)dt, \tag{3.4}$$

where $g(t|H_0)$ is the PDF of the statistic test t under the null hypothesis. The *significance* of a result is often expressed as the number of standard deviations corresponding to the area under the tail of a normal distribution at a given level of p-value.

This analysis uses a likelihood ratio test comparing the H_1 hypothesis where both signal and background are present, with expected event yields $v = \mu s + b$, with the background only H_0 hypothesis with yields $v = b$. The μ coefficient is called the *signal strength* and measures the compatibility of the observation with the signal predictions s. The ratio is computed according to the Neyman-Pearson lemma as:

$$\lambda(\mu, \vec{\theta}) = \frac{\mathcal{L}_{s+b}(\vec{x}_1, \ldots, \vec{x}_n|\mu, \vec{\theta}}{\mathcal{L}_b(\vec{x}_1, \ldots, \vec{x}_n|\vec{\theta}}, \tag{3.5}$$

where θ is the vector of the nuisances parameters applied on a data sample $(\vec{x}_1, \ldots, \vec{x}_n)$.

The value of the signal strength is obtained by minimizing $-2\log(\lambda(\mu))$. The extraction of the significance of the test statistic under the background only hypothesis must be computed. This is achieved with the *asymptotic approximation* [119] for the distribution of the test statistics.

According to the Wilk's theorem [120], the distribution of $2\log(\lambda(\mu))$ can be approximated by a χ^2 distribution with a number of degrees of freedom equal to the number of nuisances parameters. The significance can then be computed as the square root of this quantity.

In this statistical approach, the uncertainties on the predictions of both background and signals yields are handled under the form of nuisance parameters $\vec{\theta}$. They follow probabilistic laws and are modelled with probability distribution functions $\rho(\theta|\bar{\theta})$ where $\bar{\theta}$ is the default value of the given nuisance parameter. This distribution can be interpreted as the posterior probability of some real measurements of $\bar{\theta}$:

$$\rho(\theta|\bar{\theta}) \approx p(\bar{\theta}|\theta) \cdot \pi(\theta), \tag{3.6}$$

where $\pi(\theta)$ is the prior distribution of the measurement, that we consider as flat.

3.8.2 Likelihood fit inputs

In the analysis, the final fit is performed simultaneously on all the signal and control regions. The DNN output is used for the signal regions separately for each year, resolved or boosted topology and b-tagged or b-vetoed subcategory. The top background normalization is also constrained in the fit by using the DNN score in the dedicated top control regions. The shape of the Z + jets background is also corrected for the data-MC discrepancies, by using the events yields as an input to the fit in several bins as detailed in Sec. 3.4.2.

The likelihood can thus be written as

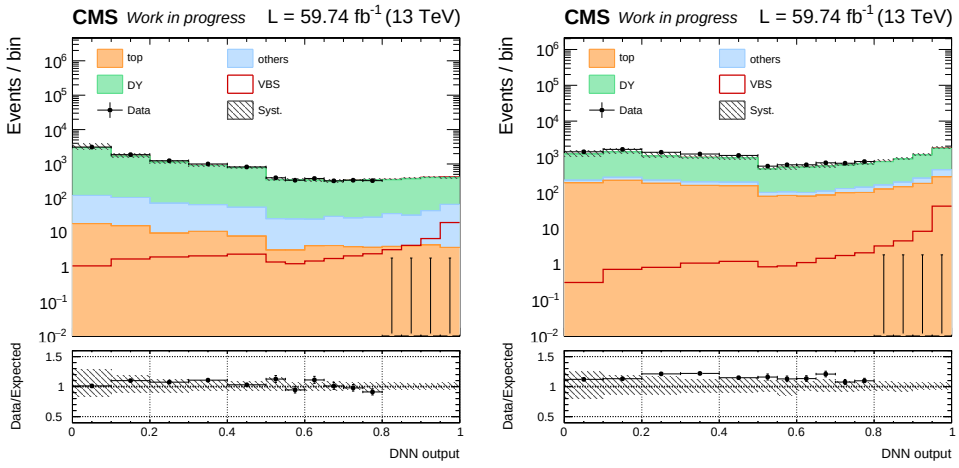
$$\mathcal{L}(\text{data}|\mu,\vec{\theta}) = \text{Poisson}(\text{data}|\ \mu\text{s}(\vec{\theta}) + b(\vec{\theta})) \cdot p(\vec{\theta}) \tag{3.7}$$

where the Poisson(data$|\mu s(\vec{\theta}) + b(\vec{\theta}))$ is the product of Poisson probabilities to observe a yield n_i in the bin i. The likelihood can then be expressed for m data points n

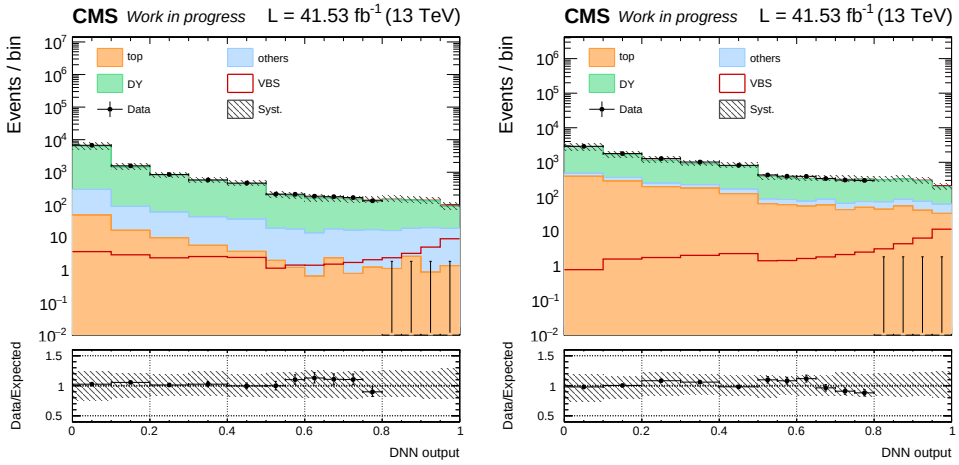
$$\mathcal{L}(n_1,\ldots,n_m;\mu,\vec{\theta}) = p(\vec{\theta}) \cdot \prod_i^m \frac{(\mu s_i(\vec{\theta}) + b_i(\vec{\theta}))^{n_i}}{n_i!} e^{-\mu s_i(\vec{\theta}) - b_i(\vec{\theta})}, \tag{3.8}$$

where s_i and b_i are respectively the signal and background yields in the bin i. In this section are presented the results of the statistical analysis for the measurement of the VBS ZV semileptonic process significance.

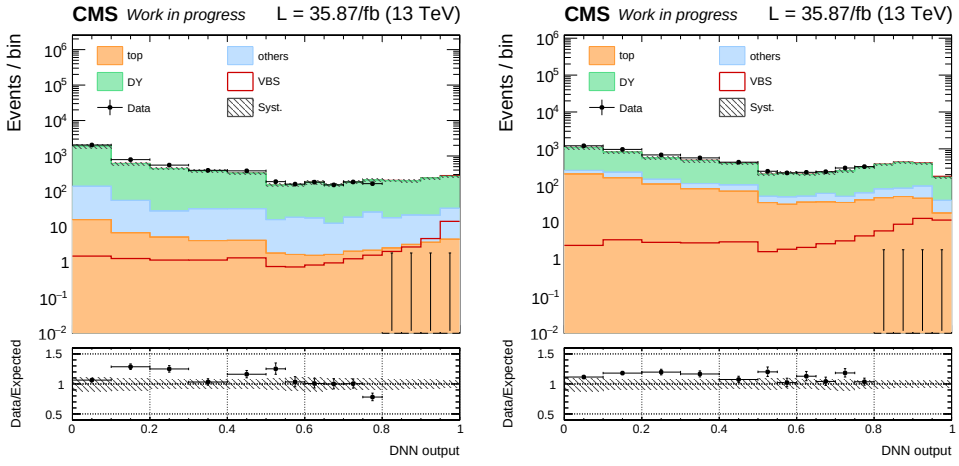
Fig. 3.23 and Fig. 3.24 show the prefit distributions for the DNN output used for the fit in the signal regions, Fig. 3.27 shows the prefit normalization in the top control regions and the number of events in the bins of the Z + jets control region are shown in Fig. 3.25 and Fig. 3.26. The analysis is not yet unblinded, so the data yields are set to zero in the signal region at high DNN output.



(a)

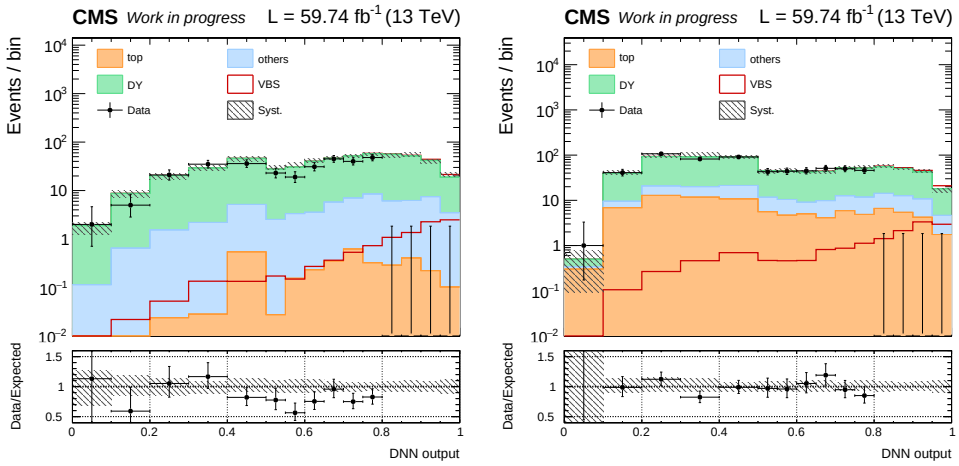


(b)

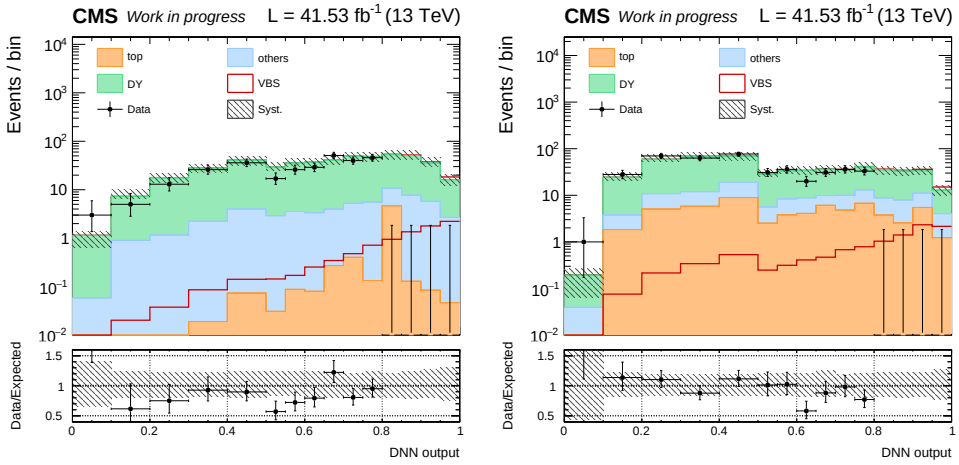


(c)

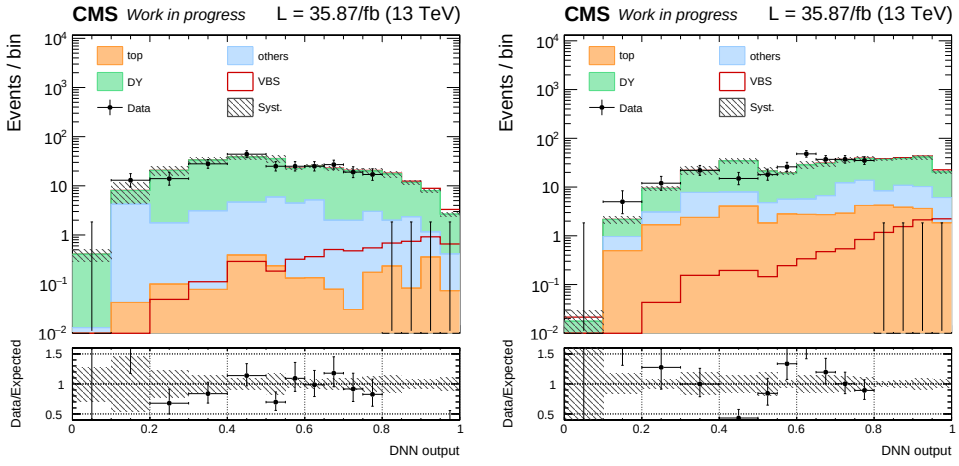
Figure 3.23: The distributions of the DNN used in the fit in the SR in the Resolved b-vetoed (left) and b-tagged (right) category for the data taken in (a) 2018, (b) 2017 and (c) 2016. The distribution of the data is blinded in [0.8,1].



(a)

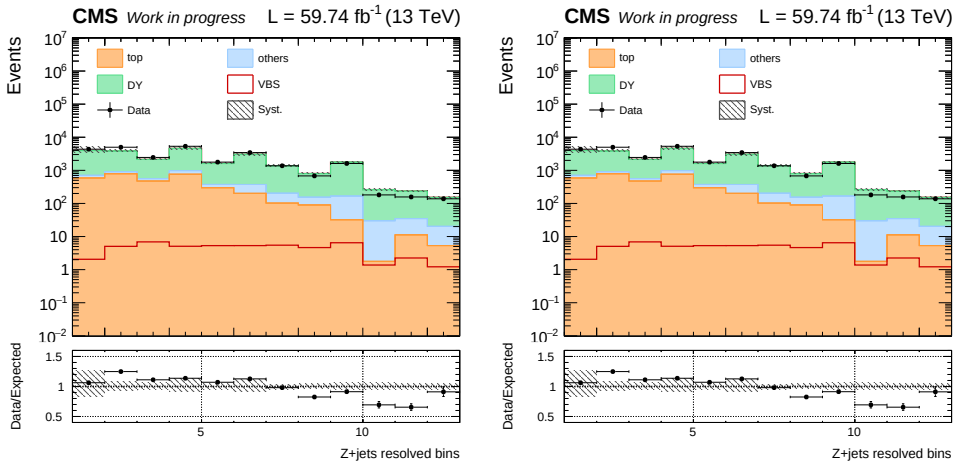


(b)

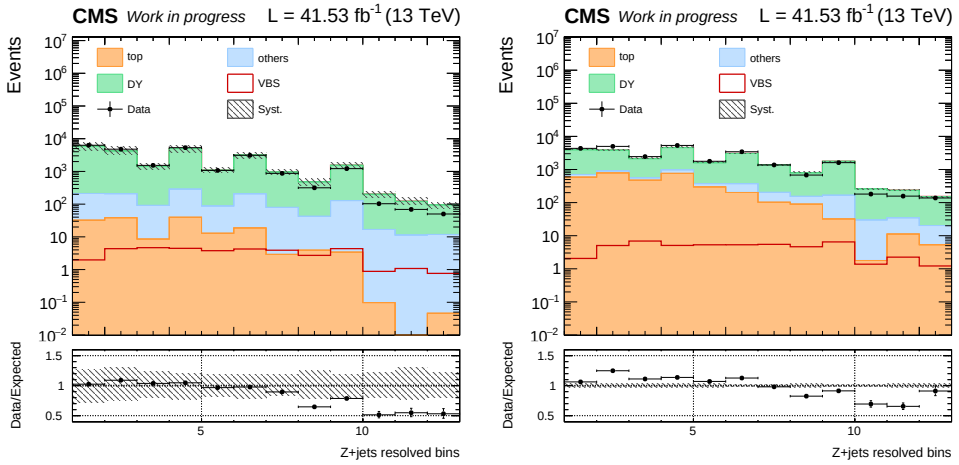


(c)

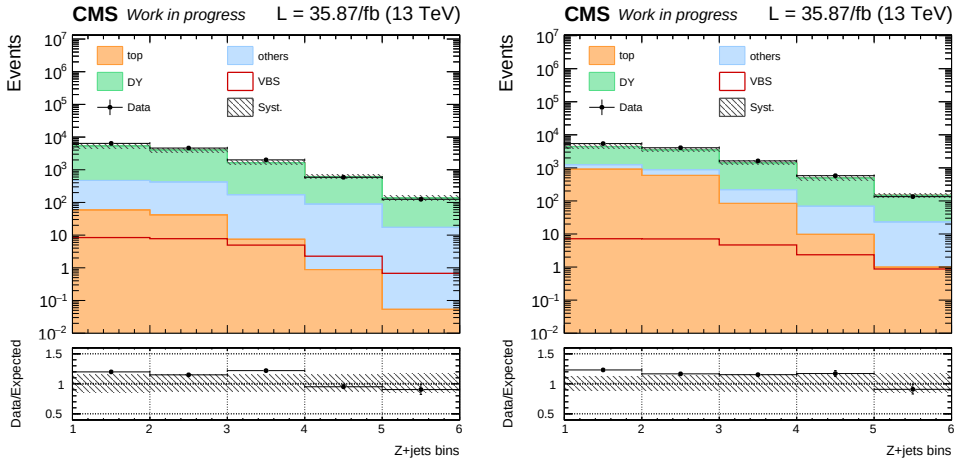
Figure 3.24: The distributions of the DNN used in the fit in the SR Boosted b-vetoed (left) and b-tagged (right) category for the data taken in (a) 2018, (b) 2017 and (c) 2016. The distribution of the data is blinded in [0.8,1].



(a)

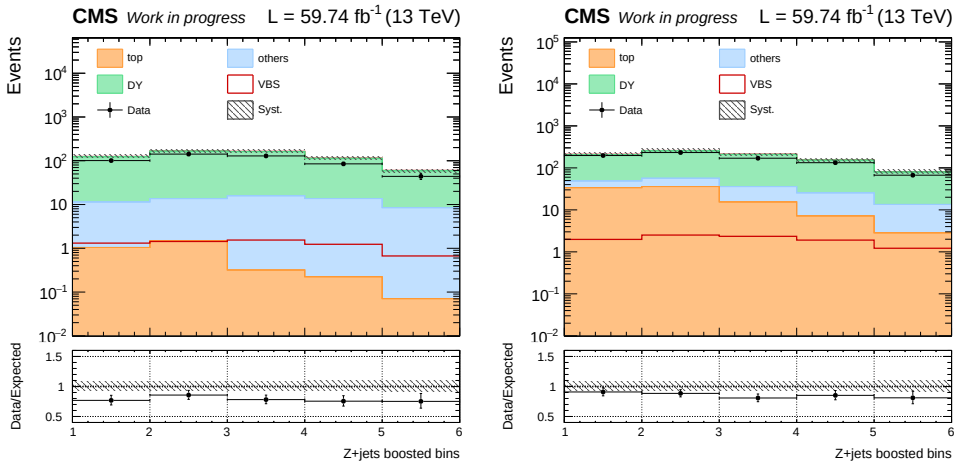


(b)

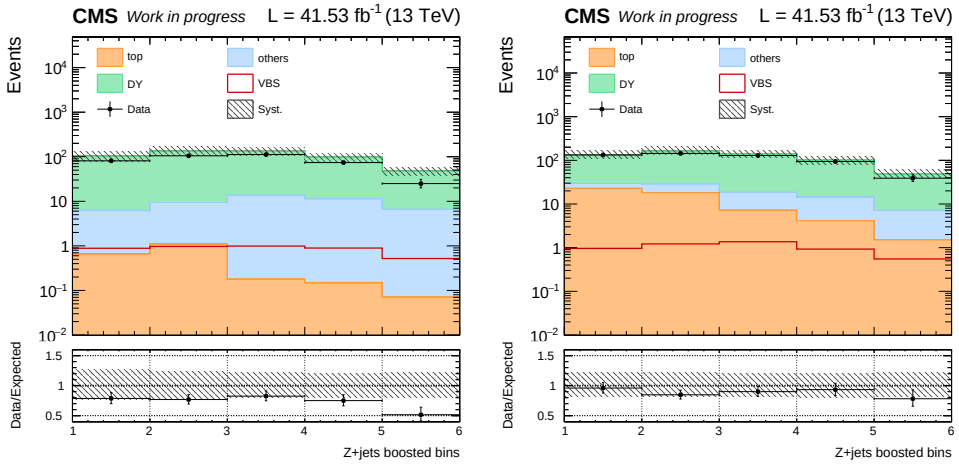


(c)

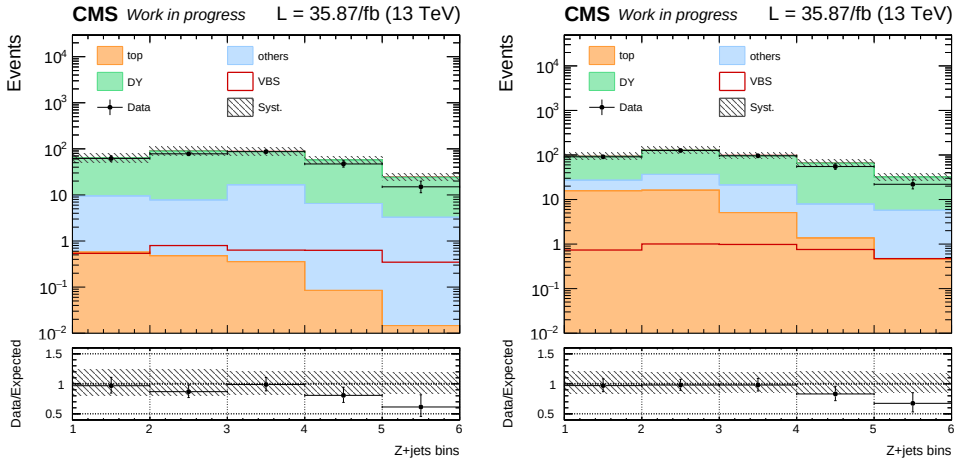
Figure 3.25: The number of events in the bins used in the fit in the DY CR Resolved b-vetoed (left) and b-tagged (right) region for the data taken in (a) 2018, (b) 2017 and (c) 2016.



(a)

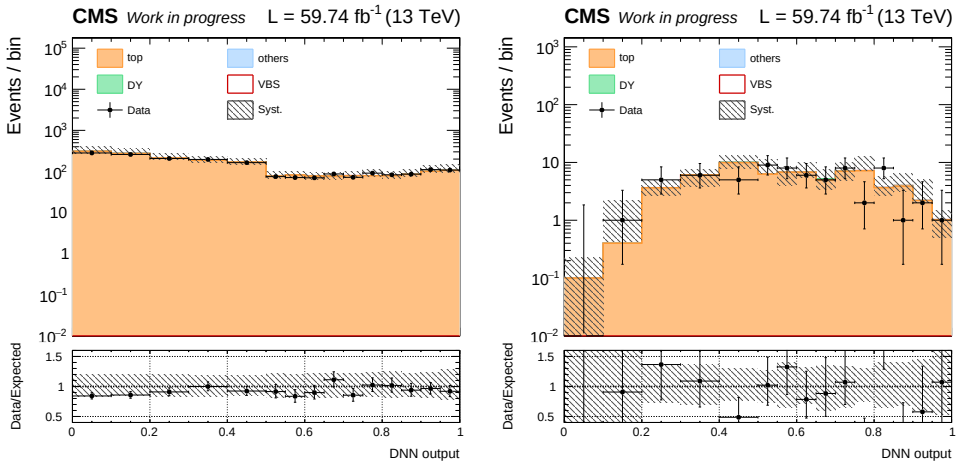


(b)

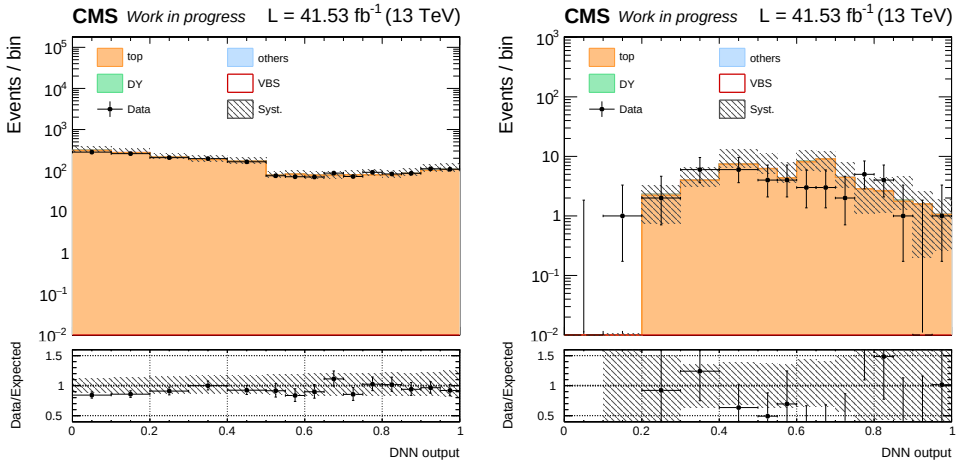


(c)

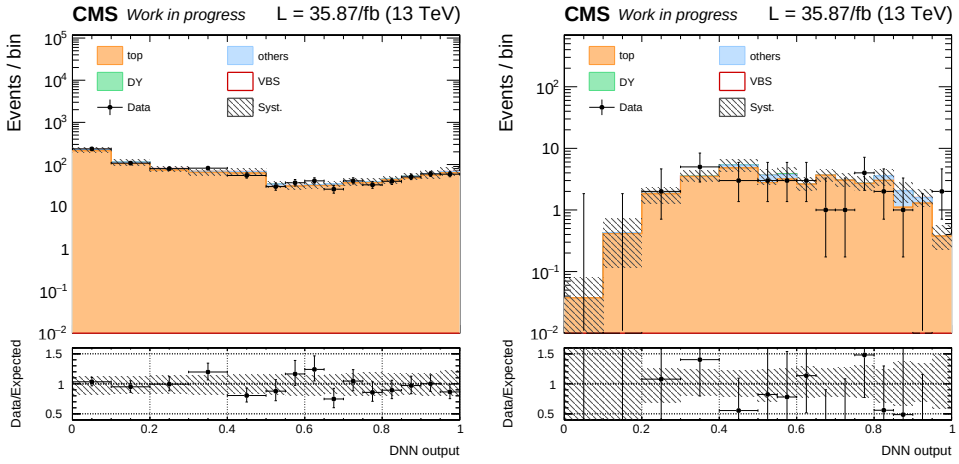
Figure 3.26: The number of events in the bins used in the fit in the DY CR Boosted b-vetoed (left) and b-tagged (right) region for the data taken in (a) 2018, (b) 2017 and (c) 2016.



(a)



(b)



(c)

Figure 3.27: The distributions of the DNN used in the fit in the top CR regions for the Resolved top CR (left) and Boosted top CR (right) category for the data taken in (a) 2018, (b) 2017 and (c) 2016.

3.8.3 Expected significance

With the procedure detailed in the previous sections, the fit is able to extract the expected signal strength from the SM simulations, and the observed signal strength from the data. The rate parameters controlling the normalization of the Z+jets and top backgrounds are also simultaneously extracted from the dedicated control regions.

As of today, the analysis is still blinded, meaning that the data yields in the signal region are kept unseen, so the results presented in Table 3.14 are the expected statistical significance corresponding to the signal strength of the SM. The total Run 2 expected significance is of 1.8 σ.

Significance(σ)			
2016	Boosted	Resolved	Combined
b-veto	0.32	0.54	
b-tag	0.38	0.38	
combined	0.47	0.62	0.72
2017	Boosted	Resolved	Combined
b-veto	0.46	0.77	
b-tag	0.48	0.58	
combined	0.61	0.83	0.99
2018	Boosted	Resolved	Combined
b-veto	0.49	0.75	
b-tag	0.63	0.64	
combined	0.76	0.92	1.15
Full Run 2	Boosted	Resolved	Combined
b-veto	0.71	1.12	1.28
b-tag	0.84	0.84	1.12
combined	1.02	1.35	1.8

Table 3.14: The expected significance for the different years and categories.

3.9 Future prospects

The analysis, though not yet unblinded, is well advanced and currently being reviewed internally, more complete results will be available soon. Because of the limited statistics and the discrepancies in the data modeling, even with such complex corrections and signal extraction the Run 2 dataset is not enough to observe significantly the rare VBS ZV semileptonic process. The limits extraction for anomalous quartic gauge couplings in the EFT framework is the next step of the analysis. The Boosted topology in particular is sensitive to quartic couplings, and non-zero aQGC should enhance the production cross section at large invariant masses of the boson pair. A likelihood scan based on the mass of the ZV system in the signal regions will

be performed, extending the results obtained in Ref. [34] on a partial Run 2 dataset. A combination with the complementary analysis involving the production of WV [35] is foreseen, increasing the sensitivity.

In the future, the larger amount of data collected during Run 3 and even more so during the HL phase of the LHC will improve tremendously the statistical uncertainty. This fact, in addition to improvements on the theory modeling and more sophisticated signal extraction techniques should allow for the observation of such rare processes, and allow deeper probing of the EWSB sector of the SM, and provide more stringent limits on BSM effects. The VBS studies is a very active field, and the sensitivity to these rare processes is part of the considerations taken for the design of the LHC detectors upgrades.

Chapter 4

CMS High Granularity Calorimeter

4.1 Introduction: the High-luminosity LHC upgrade

The LHC has recently started the Run 3 of its Phase I of exploitation, at an increased center of mass energy $\sqrt{s} = 13.6$ TeV, with the objective of improving the precision measurements and new physics searches through hardware innovations and sophisticated analysis techniques, as well as higher energy and luminosity. This data-taking phase should go on for three years, at the end of which the LHC will enter the Long Shutdown 3 (LS3). During this time, the LHC will prepare for the Phase II of its exploitation program: the so-called High-Luminosity LHC (HL-LHC). The timeline is detailed in Fig. 4.1.

This upgrade is designed to reach an instantaneous luminosity of 5.10^{34} cm^{-2}.s^{-1}, compared to the Run 3 luminosity of 3.10^{34} cm^{-2}.s^{-1}, for an expected integrated luminosity of 3 ab^{-1} after around 10 years of exploitation. These extremely high values mean new challenges for the detectors, in term of radiation levels and amount of pileup (PU) - the multiplicity of additional pp interactions in addition to the hard scattering event. At the HL-LHC luminosity, the expected average amount of PU per bunch crossing is of 140, and could reach up to an average of 200 by taking into account the LHC capacity to still deliver 50% higher luminosity. A typical event with 200 PU interactions is shown in Fig. 4.2.

The increased amount of data collected will allow for the search and precision measurement of rare processes. In particular, several processes related to the Higgs boson or VBS, as described in Sec 1.3.3, involve highly boosted jets in the forward regions of the detector, which are incidentally the regions exposed to the maximum radiations. One of the main objectives of the upgrade plan is thus to ensure that the performance of the detectors is not degraded in those crucial regions. In the case of CMS, the endcaps of the current calorimeter were built to cope with a maximum of 500 fb^{-1} and will suffer from a loss of transparency of the crystals inducing unacceptable degradation of the detector's performance before the end of the HL program.

To contend with those new challenges, the LS3 will be a period of time dedicated to the replacement and upgrade of several LHC detectors parts. For CMS, an overview of the upgrade

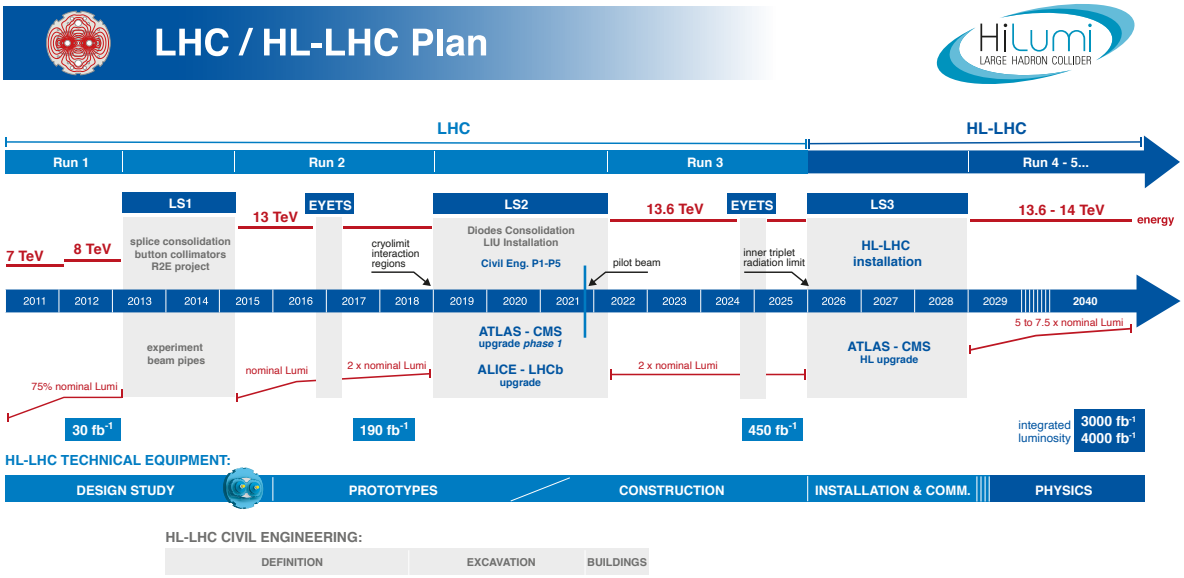


Figure 4.1: The timeline for the planned LHC and HL-LHC upgrade and operation. The LS3 is expected to start in 2026 for a duration of three years during which the detectors will receive several upgrades to prepare for the high luminosity upgrade.

plan is given in Sec. 4.2. Part of this upgrade plan for CMS is to replace the endcap calorimeters with the so-called High Granularity Calorimeter (HGCAL), whose design is detailed in Sec. 4.3. The HGCAL will have to withstand up to ten times the integrated radiation level of the initial CMS endcaps design, placing strong constraints on the materials used for its construction. Silicons detectors are known to perform well in high radiation environments [121] and were thus chosen as the active material in the parts of HGCAL submitted to the highest fluences and accumulated radiation. The regions submitted to less radiation will be covered by less expensive plastic scintillator tiles connected to on-tile silicon photomultipliers (SiPM). The very high transverse and longitudinal segmentations of HGCAL translate into an unprecedented granularity (which gave its name to the detector) that will be needed to mitigate the strong PU. This granularity and the high PU environment of the HL-LHC also lead to new challenges in terms of data processing and triggering at very high rates, motivating the upgrade of the CMS triggering system presented in Sec. 4.4.

4.2 CMS upgrade plans

During the HL phase, CMS is foreseen to undergo multiple upgrades to cope with the harsh environment while keeping, and even improving, the current physics performance. The increased luminosity is a source of major challenges for the detector. First of all, the radiation doses the equipment will have to face are unprecedented, and substantially higher than what was expected during the original design. This means the tracker and endcap calorimeters, where the radiations are the most important, will need to be replaced, and the barrel calorime-

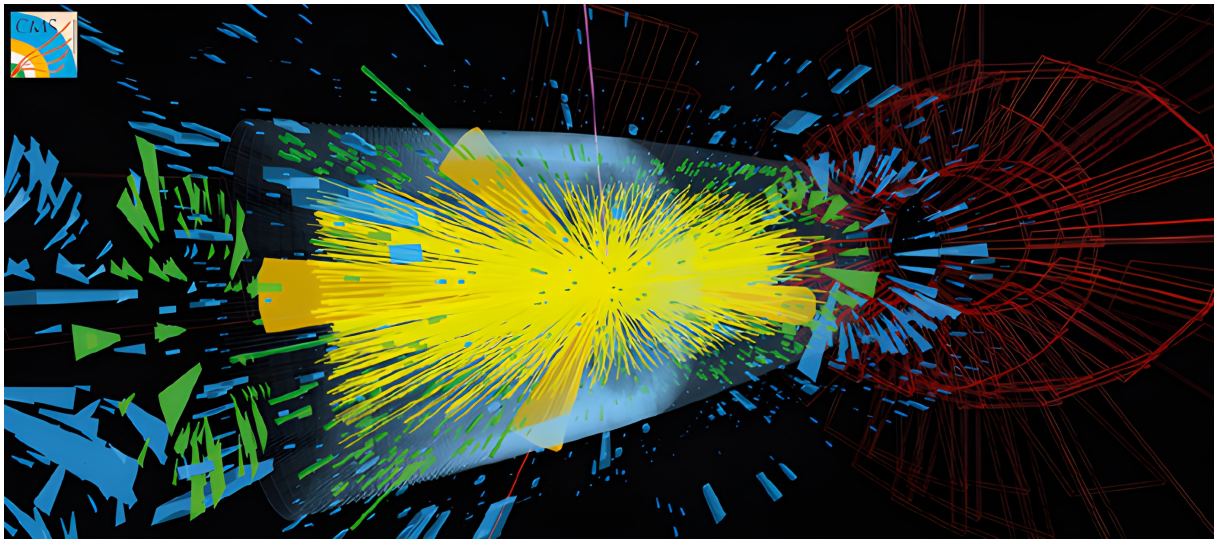


Figure 4.2: Event display of a collision in a 200 PU environment.

ter and muon detectors will require substantial improvements. The increase in instantaneous luminosity dictates the usage of highly granular detectors and advanced technologies to successfully mitigate the higher PU rate. To fully exploit those detector upgrades, an updated and improved trigger system will need to be implemented. A schematic view of CMS upgrades is represented in Fig. 4.3.

The tracker will be completely replaced with a new inner tracker consisting of pixel detectors of smaller size compared to the current ones, and outer tracking stations that use strips and macro pixel sensors up to $|\eta| = 3.8$, thus improving the granularity of the detector[122]. Both longitudinal and transverse resolution will increase, enhancing the reconstruction performance and lowering the misidentification rate.

The endcap calorimeters, already significantly degraded by the Phase I operation, will be replaced by the HGCAL whose design is detailed in the Sec. 4.3. The unprecedented granularity of this calorimeter will increase the shower separation and thus improve particle identification, while also providing precise timing information [123]. The electronic readouts for the barrel calorimeter and muon subsystem will be upgraded [124]. The forward region of the muon chambers will also receive an upgrade with the addition of RPC and GEM detectors that will extend the coverage and overall redundancy [125]. New MIP timing detectors will be installed in front of the barrel and endcap calorimeters for additional precise timing information [126].

The CMS trigger and data acquisition system is also planned to be totally replaced. The upgrade of the L1 trigger system, detailed in Sec. 4.4, is designed to maintain the current efficiencies while at the same time enhancing sensibility to new physics. The algorithms used for reconstruction and identification will be implemented on FPGAs and make use of the tracker information for the first time, allowing a complete implementation of the Particle Flow [63]. The high level trigger (HLT) will use the information from all subdetectors at an increased maximum input rate of 750 kHz and reduce the input rate by a factor 100 to 7.5 kHz [61].

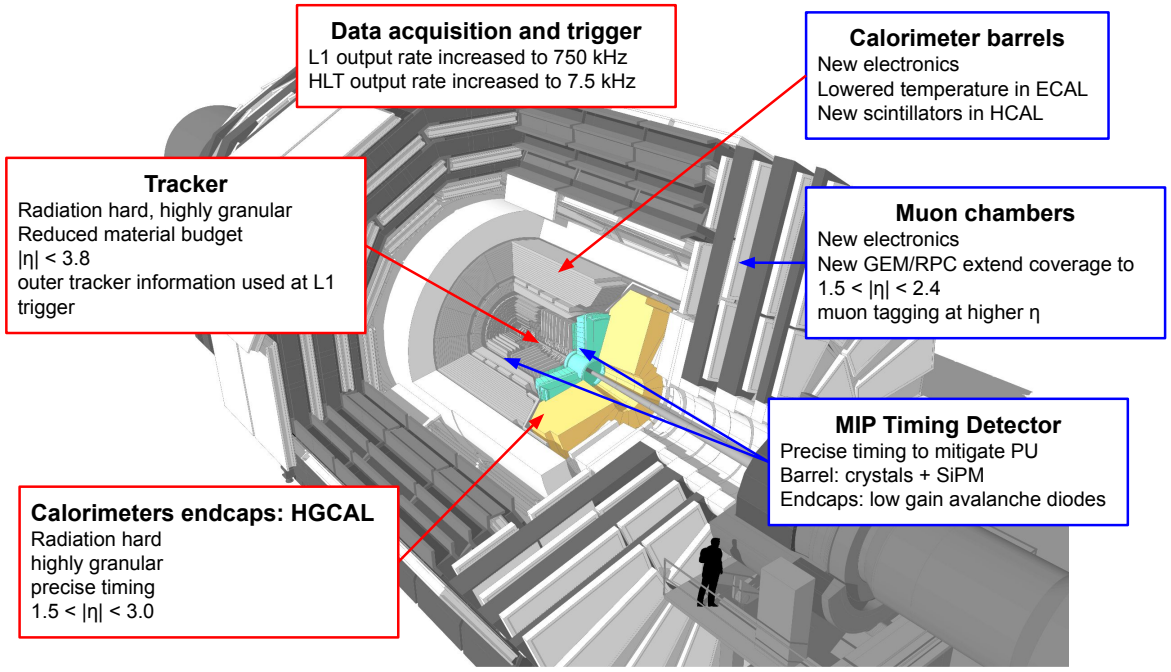


Figure 4.3: A schematic view of the CMS detector with the foreseen upgrades for HL-LHC phase. The upgrades highlighted in red show elements that will be completely replaced while the elements highlighted in blue correspond to new modules or extensions of the Phase 1 detector.

4.3 The new HL-LHC CMS endcap calorimeter HGCAL

By the end of the LHC Phase I operation, CMS endcap calorimeters will have suffered irrecoverable degradation of their performance. The exposition to high amount of radiation causes the lead-tungstate crystals to lose their transparency, at a level that will be critical after 500 fb^{-1} of integrated luminosity. CMS plans for their replacement is the construction of the High Granularity Calorimeter (HGCAL).

The future HGCAL is one of the most ambitious calorimeter to date. With its exceptional granularity and high amount of channels, it is an imaging calorimeter capable of providing three-dimensional images of the particle showers. A first in detector calorimetry, it additionally provides precise timing information with $\mathcal{O}(10\,\text{ps})$ resolution for EM showers. While those characteristics are necessary to maintain and even improve performance compared to Phase I despite the more complicated environment, they also entail demanding constraints on the design of the HGCAL.

4.3.1 Structure of the calorimeter

The HGCAL is a sampling calorimeter which alternates distinct absorber layers and active materials layers. The incident particles interact with the dense material of the absorber layers, initiating showers in the detector. The interaction with the active layers provides a signal proportional to the energy deposited allowing the reconstruction of the particle energy.

This sampling design of HGCAL provides degraded energy resolution for EM showers compared to the current CMS homogeneous ECAL, but the fine segmentation allows much better particle separation, a critical feature at HL-LHC where one of the main challenges is the PU mitigation. The HGCAL is also significantly more radiation-hard compared to the current ECAL technology based on lead-tungstate crystal, a necessary feature to contend with the higher doses of radiation provided during the Phase II.

The details of the HGCAL design are reported in the HGCAL Technical Design Report [123] (TDR). The calorimeter is divided into an electromagnetic part (CE-E) and a hadronic part (CE-H) as illustrated in Fig. 4.4. In the geometry described in the TDR, it is envisioned that the electromagnetic compartment will feature 28 active layers and CuW, Cu and Pb absorbers, while the hadronic compartment is foreseen to possess 24 active layers and stainless steel absorber. The 26.3 radiations length (X_0) and 1.73 nuclear radiation lengths (λ_n) of the CE-E are extended to a total detector length of 10.75 λ_n when adding the CE-H.

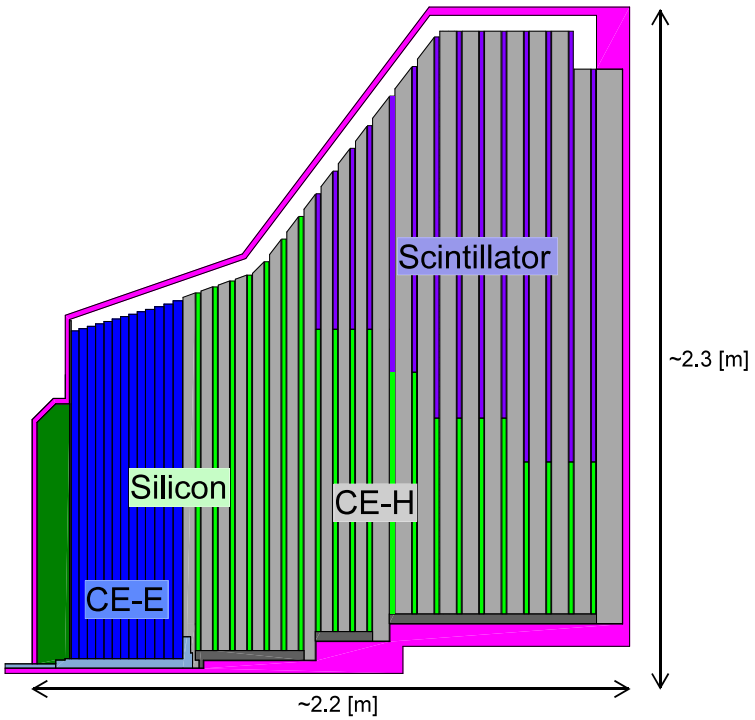


Figure 4.4: Overview of the HGCAL cross-section. The CE-E and inner part of the CE-H use silicon modules while the outer CE-H uses plastic scintillators. The expected pseudorapidity coverage is 1.5 < $|\eta|$ <3. Updated design as of April 2022.

The design of the active layers was dictated by the need for radiation-hard materials. As such the CE-E and central part of the CE-H, that are the regions exposed to the highest radiation doses, will employ around 600 m^2 of hexagonal silicon sensors. Those sensors are divided into cells of sizes ranging from 0.5 cm^2 to 1.2 cm^2 depending on the pseudorapidity, a choice motivated by the cell capacitance limitations and by physics performance. In particular, they were optimized to maximize the expected S/N ratio for a Minimum Ionization Particle (MIP) and ensure that a shower can not be completely contained in a single cell by keeping their size under the Moliere radius. Those cells are assembled in 8" sensor modules (SM) with an active

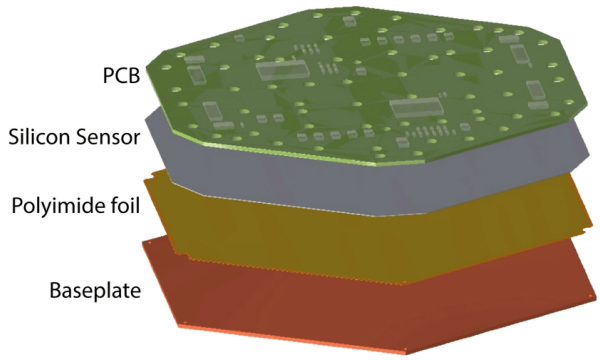


Figure 4.5: Schematic view of the silicon modules showing the different layers: the baseplate of WCu (carbon fiber) in the CE-E (CE-H), kapton-sheet foil, sensor, and front-end PCB.

thickness of 120, 200 or 300 µm depending on the detector region and expected fluences, in order to optimize the charge collection during the whole operation of the detector. The modules are composed of stacked layers including a baseplate, a kapton-gold sheet, the silicon sensor and a printed circuit board (PCB) on which the front-end electronics is situated, as illustrated in Fig. 4.5. The total number of silicon cells adds up to about six million cells that are read individually during operation.

The regions of the CE-H further away from the beam, which are exposed to significantly lower radiations as evidenced in Fig. 4.6, use instead plastic scintillator tiles in order to optimize the costs. The size of the square tiles depends on the pseudorapidity and range from 2×2 to 5.5×5.5 cm^2, for a total of about 400 m^2 active area. The light emitted by the scintillators is read by photomultipliers that produce the electric signal readable by the detector electronics.

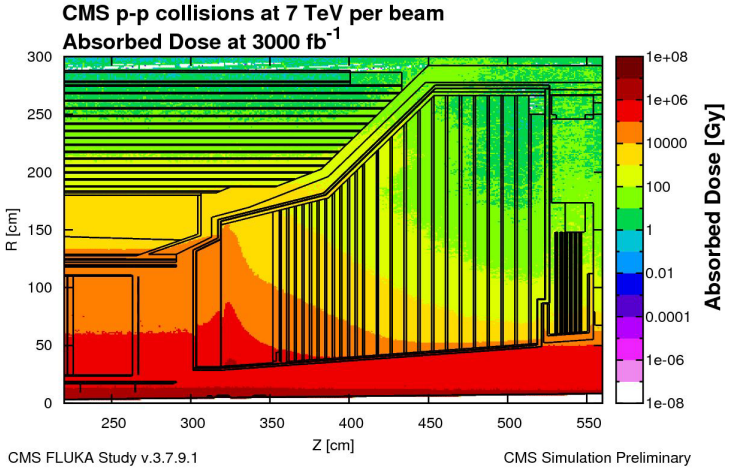


Figure 4.6: Dose of radiation absorbed by the HGCAL after 3000 fb^{-1} of collisions in a cross section view as simulated with FLUKA[127]. The sections of the detector exposed to the largest dose are covered with radiation-hard silicon modules while in the outer parts the plastic scintillators are submitted to lower irradiation.

The nominal parameters such as the number of layers presented here correspond to the TDR design and have since evolved. In particular, the detector simulation used in the trigger

studies described in Sec. 5 uses a slightly updated geometry, with only 22 layers in the CE-H. The current foreseen geometry is composed of 26 layers in the CE-E and 21 in the CE-H.

The high precision timing information measured by the readout ASICs, providing a EM shower timing resolution of $\mathcal{O}(10\text{ ps})$, is added to the fine lateral and longitudinal segmentation, making the HGCAL a 5D (3D position, energy and timing) calorimeter. This innovative feature for a calorimeter will allow better pileup mitigation by rejecting interactions recorded outside of a certain time frame and provide additional information for the Particle Flow reconstruction. With around six million channels, it improves the granularity by a factor 500 compared the the Phase I endcaps calorimeters. The readouts of the HGCAL are used to build highly granular trigger primitives, used for particle identification in the Level 1 trigger, as described in the following sections.

4.4 CMS trigger system upgrade

The CMS experiment uses a two-level triggering system composed of a hardware-based Level-1 trigger and a software-based High-Level Trigger (HLT). The current L1 trigger receives the information from the calorimeter and muon subsystems and has a maximum latency of 4 μs to make a decision, with an average event accept rate of 100 kHz. If the L1 trigger identifies an event as interesting, the HLT will then reconstruct this event with information from all detector subsystems and perform a selection for an average output rate of 1 kHz. Both trigger subsystems are planned to be replaced for the Phase II of the LHC, with an increased throughput. The new HLT is foreseen to operate at an input rate of 750 kHz and output rate of 7.5 kHz.

The upgrade of the L1 trigger system is designed to retain the signal efficiency of the Phase I trigger in a more complicated environment and at higher luminosity, and even significantly improve the sensitivity to rare physics manifestations. The coverage of the trigger will be improved in the forward region of detector and, for the first time, the L1 trigger will have access to the outer tracker information. The new L1 trigger will feature an implementation of the Particle Flow (PF) algorithm for the first time at this level, since it has to date only been employed in the HLT and offline reconstruction. Using this new information and with the improved granularity of the subsystems, the PF will allow a full exploitation of the collisions even in a challenging setting. Those trigger algorithms will be implemented on Field Programmable Gate Arrays (FPGA) providing fixed latencies, and use optical links to propagate information between the components at high speed. The output rate of the L1 trigger will be increased to 750 kHz with an associated enhanced latency of 12.5 μs.

4.4.1 HGCAL trigger primitive generation

The input of the HGCAL TPG consists in the trigger cells (TCs) summed energies. These TCs correspond to an area of $\approx$4 cm^2 in the silicon regions and 4×4 to 10×10 cm^2 for the regions covered by scintillators. The charges deposited in these TCs are compressed to 7 bits with a floating point format, without timing information because of the enormous bandwidth it would require. Since the bandwidth is strongly limited, only every odd layer of the CE-

E serves as a trigger layer and not all TCs can be processed, hence further data reduction strategies are implemented.

To that end, three different strategies are considered and are implemented in the fronted electronics. The first one, which is employed for producing the data used in the studies described in Sec. 5 is based on the selection of TCs above a certain energy threshold of 1 to 2 MIP_T[1] depending on the TCs occupancy. The second one, called *Best choice* (BC) consist in the selection of a fixed number of the highest MIP_T TCs. The last one consists in the creation of *Super Trigger Cells* (STC) by grouping TCs to form blocks of reduced granularity, while keeping the information on the position of the most energetic TCs. The threshold strategy has the overall best performance across the whole cluster energy range. However, it suffers from the variability of its output size, since the number of selected TCs can not be predicted. This requires additional processing steps in the later steps of the TPG to unpack the data, in particular to resynchronize the input streams. The BC strategy is competitive for compact EM objects and τ leptons that are usually depositing their energy in a small number of TCs that can all be selected, but loses efficiency for larger clusters. The STC strategy on the other hand is competitive for jets. A combination of BC in the CE-E and STC in the CE-H is thus considered as the current baseline data reduction strategy, providing a fixed format output and good performance in both regions of the detector.

Part of the loss of energy resulting from these selections is compensated by sending in parallel the sum of the energy deposited in each module (a HGCAL module is composed of 48 trigger cells in the silicon region and a similar number for the scintillators). This first step of the TPG consisting of the TC summation, energy compression and selection, and the module sums creation, is assured by two on-detector Application Specific Integrated Circuits (ASIC), the High Granularity Calorimeter Readout Chip or HGCROC, followed by the Endcap CONcentrator Trigger or ECON-T.

The output are then communicated to off-detector FPGAs, by thousands of lpGBT [128] (low power GigaBit Transceiver) optical links providing a transfer rate of 10 Gb/s, where the TPG is implemented in two stages as illustrated in Fig. 4.7. The first stage is composed of 42 boards per endcap, each of them equipped with a VU13P FPGA receiving the input from the ECON-T via multiple lpGBT links. It performs data repacking and calibration before sending the information to the second stage with 16 Gb/s links in a time-multiplexing manner: the FPGAs of the second stage cover a 120° portion of an endcap and process one over 18 bunch crossings each. The data located close to the boundary of each Stage 2 quadrant is shared with the neighbouring FPGA. This way, showers depositing energy near the border between two FPGA regions can be fully reconstructed. The large regions in depth covered by the Stage 2 FPGAs allow the implementation of 3D clustering algorithms.

The reconstruction algorithms implemented in the second stage are divided in two steps: the *seeding* of clusters and the *clustering* around the identified seeds. The seeding consists in a projection of the TCs into histograms bins in the (r/z, ϕ) plane. The raw histograms are smoothed to reduce fluctuations and the seeds are defined as local maxima above a 10 MIPT seeding threshold. Their coordinates are computed from the weighted barycenter of the TCs contained in the seeded bin. With those seedings parameters, the risk of reconstructing

[1]The transverse energy deposited by a particle at the minimum energy loss rate.

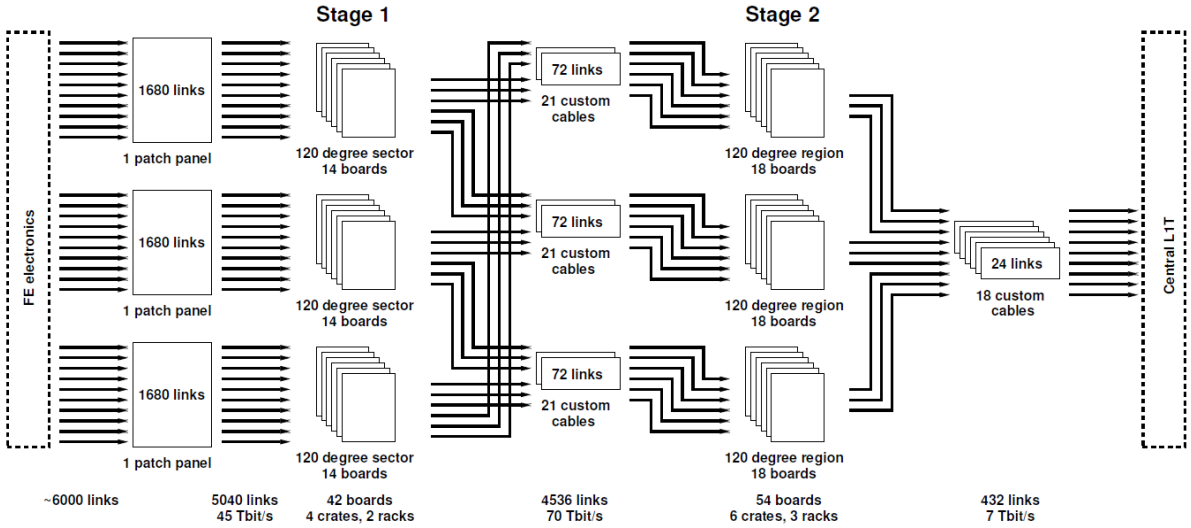


Figure 4.7: Illustration of the two stages of the HGCAL TPG backend for one endcap. The first stage receives the trigger data from the frontend via lpGBT links and sends repacked and calibrated data to the second stage in a time-multiplexing fashion. The output of the Stage 2 is sent to the central L1T.

multiple seeds for a single shower is alleviated, though not completely erased. When building the trigger objects in the later stages, nearby clusters are merged to limit the impact of energy splitting on the trigger performance. In the clustering step, the TCs are aggregated into 3D clusters around the closest seed with a distance parameter in the $(x/z,\ y/z)$ plane varying from 0.015 in the first layers to 0.050 in the last layers. The cluster size was selected to attain a good shower containment while limiting the contamination from PU.

The position of the reconstructed clusters is defined as the barycenter of the position of the TCS contained in it, weighted by their energy. To characterize those clusters, several quantities describing the longitudinal and transverse profiles of the shower, collectively known as *shower shapes* variables, are computed, in addition to information on the possibility of the cluster being reconstructed from overlapping showers. The available bandwidth for communicating those variables to the central LT is limited, and as such it is critical to select the cluster shapes that have the most discriminative power. A study on this variable selection for the identification of electromagnetic showers is detailed in Sec. 5.3. In parallel with the cluster reconstruction, "energy towers" are created to recover the energy not accounted for in any cluster. The module sums are aggregated in projective towers of $\pi/36$ size in the $(\eta - \phi)$ plane, matching the geometry of the barrel calorimeter towers. The TPG architecture, covering both the frontend and backend, is summarized in Fig. 4.8.

4.4.2 CMS Phase 2 Level 1 trigger

To fully exploit the capabilities of the upgraded subdetectors of CMS and maximize the physics performance in the harsh environment of the HL-LHC, the L1 trigger will also undergo an upgrade. Data from the new tracker and from the new endcap calorimeter, the HGCAL, are used in addition to the more granular data the muon system and calorimeter. In order to

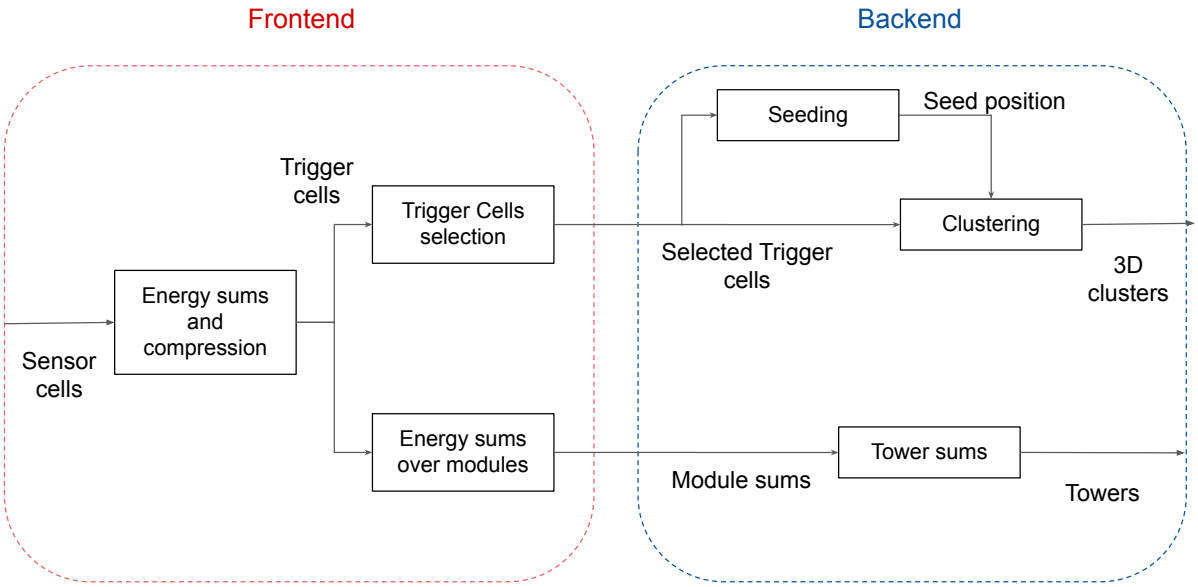


Figure 4.8: Summary of the different processing steps in the frontend and backend of the HGCAL TPG system. 3D clusters are built in the backend from trigger cells selected in the frontend and shower shape variables are computed. Towers with coarser energy sums are sent in parallel to conserve the information of the non clustered energy.

process the complex tracking information and increased calorimeter granularity, the latency is increased from 3.8 µs to 12.5 µs. Similarly, the previous output rate of 100 kHz is upped to 750 kHz to maintain the performance.

The foreseen updated L1 trigger architecture is shown in Fig. 4.9. The global trigger uses the information from the calorimeters, tracker and muon chambers to perform its selection. For the first time, an implementation of the Particle Flow algorithm is also included at the L1 trigger to reconstruct high level objects and pass the information to the global trigger.

The calorimeter trigger receives the trigger primitives from the HGCAL, described in Sec. 4.4, the barrel calorimeter (BC) and hadron forward calorimeter (HF). It benefits from the high granularities provided by the HGCAL and barrel compared to the Phase I trigger and the upgraded readout. A global calorimeter trigger (GCT) receives these information and compute the L1 e/γ, L1 τ_h, L1 jets and L1 sums.

The inputs from the various muon chambers are collected in the muon trigger, with an enhanced coverage up to $|\eta| < 2.8$ provided by the upgrade of the muon spectrometer. The muon tracks are reconstructed in three different pseudorapidity regions (the BMTF, EMTF and OMTF) and sent to the global muon trigger (GMT) where they can be combined with the tracker data.

A novelty of the Phase 2 L1 trigger is the inclusion of tracking information, providing an additional handle for triggering purpose. The tracks from the upgraded outer tracker are reconstructed in the track finder (TF) and the primary vertices are reconstructed in the global

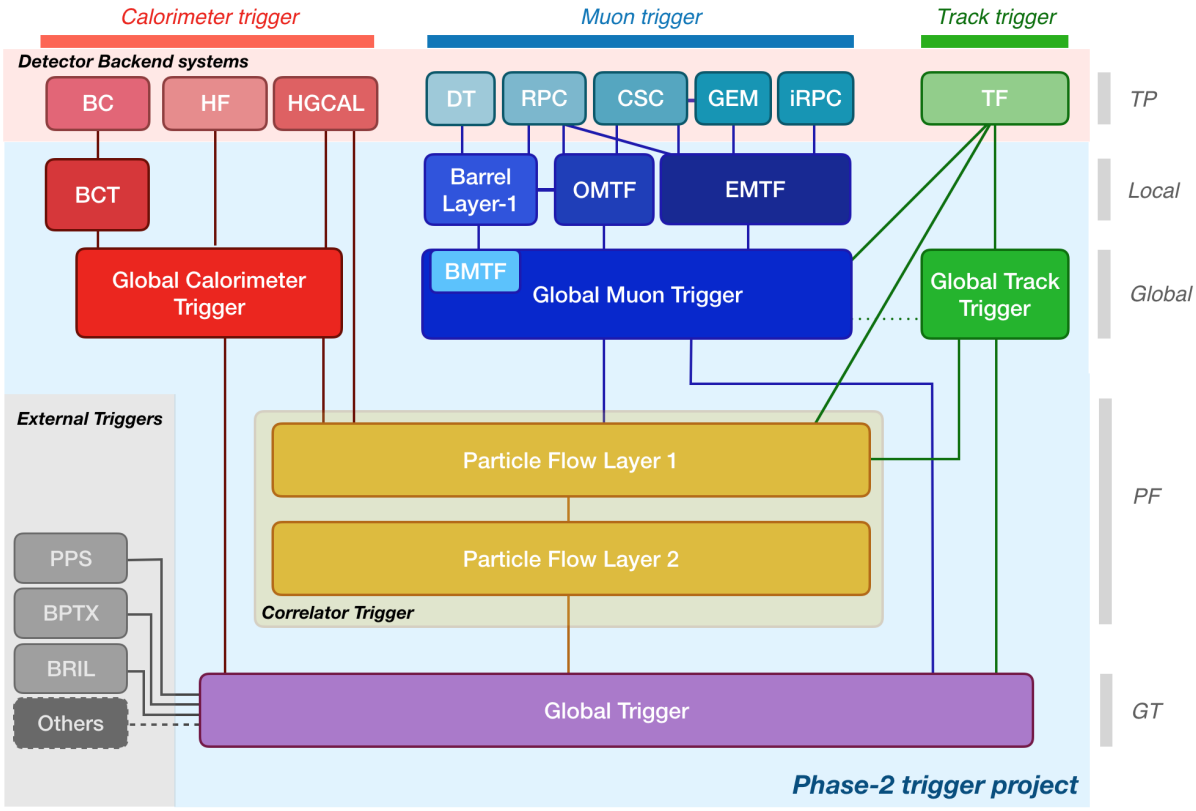


Figure 4.9: Summary of the Phase 2 Level 1 trigger of CMS. The global calorimeter trigger receives the trigger primitives from the new endcap HGCAL in addition to the barrel calorimeter (BC) and hadron forward calorimeter (HF). In parallel, the global muon trigger receives the primitives from the muon chambers track finders (BMTF, OMTF and EMTF) and for the first time at L1 tracker information are received in the global track trigger from the track finder (TF). The information from those three subtriggers are treated by the Particle Flow algorithm in the correlator trigger. The global trigger then collects the output from all the subsystems to perform the trigger decision [61].

track trigger (GTT).

Another feature of the upgraded L1T is the first online implementation of the Particle Flow algorithm to reconstruct high level objects in the correlator trigger (CT). The system works in two steps, the first layer matches the calorimeter clusters to tracks and form trigger candidates, while the second layer performs additional identifications. A simplified version of the Pileup Per Particle Identification [74] (PUPPI) is also implemented to mitigate the pileup contribution during the reconstruction based on the vertexing information.

Finally, the global trigger uses the output from the GCT, GMT, GTT and CT to perform the trigger decision based on several algorithms in a similar way as in Phase 1. The consistency of candidates with the different types of physics objects is evaluated to classify them. The communication between the complementary modules, ensured by high speed optical links, allow for the global view of the detector critical for pileup mitigation and global quantities evaluation such as missing energy. The new available information allow sophisticated event selection strategies that could not have been implemented during the Phase I. Additional trigger signals are thus designed with the purpose of identifying rare and exotic processes, notably

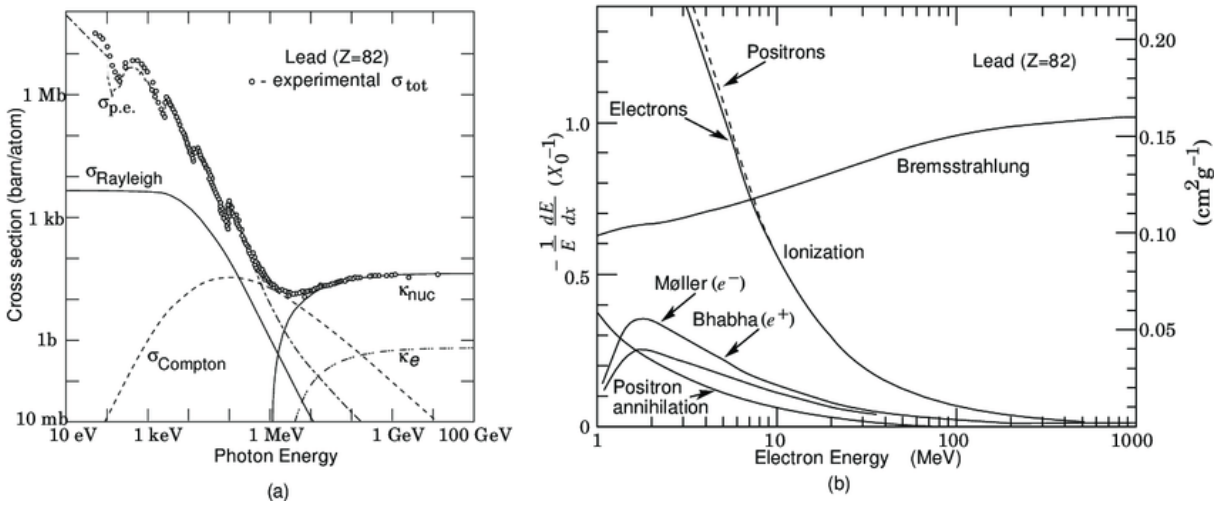


Figure 4.10: (a) Photon cross section as function of the energy in a dense absorber (Pb) and (b) the fractional energy loss of electrons per radiation length as a function of the energy, showing the contribution of different processes. Both figures have been taken form Ref. [7].

by exploiting soft and correlated muons, light mesons and jets.

4.5 Shower phenomenology in the HGCAL

4.5.1 Electromagnetic showers

The electromagnetic (EM) showers are created by the interaction of photons or electrons in the detector. The dominant interaction processes for EM objects depend on the energy of the objects, as illustrated in Fig. 4.10. At the energy scale of the LHC, the development of electromagnetic showers is dominated by the pair creation of an electron and a positron from photons and by photon emission via bremsstrahlung for the electrons. These two types of interactions alternate with each other, causing cascades of electrons an photons of lower energy in the detector as represented in Fig. 4.11 for a photon. Once the energy reaches the critical energy E_C, other processes become more important and the secondary particles are slowly stopped (electrons) or absorbed (photons), transmitting their energy to the calorimeter where it can then be measured.

Electromagnetic showers usually start in the early layers of the electromagnetic compartment and deposit most of their energy before reaching the hadronic compartment. They possess a narrow core corresponding to the earlier stages of the shower, where particles duplicates approximately every X_0 until the critical energy is reached, corresponding to the shower maximum with the highest number of secondary particles. Around 90% of the shower energy is contained in the Moliere radius parametrized as

$$R_M = \frac{21\text{MeV} \times X_0}{E_C}; \tag{4.1}$$

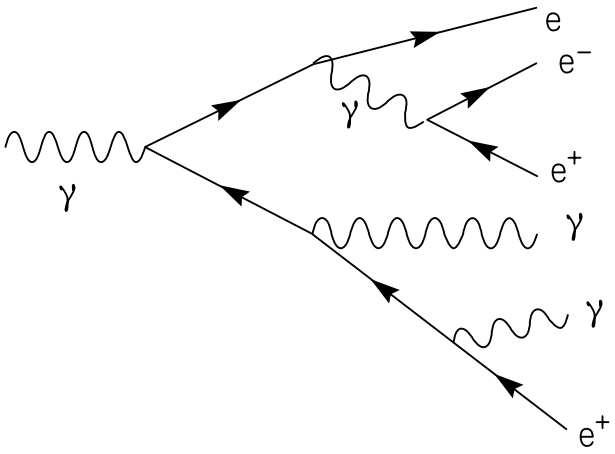


Figure 4.11: Schematic representation of the beginning of an electromagnetic shower initiated by a photon (γ). The photons create electron-positron pairs and the e^-e^+ pair radiate photons by bremsstrahlung.

where X_0 is the radiation length that governs the longitudinal development of the shower. At larger angles, the shower present tails due to low-energy isotropic processes like Compton scattering.

4.5.2 Hadronic showers

Hadronic showers can also develop in the HGCAL. They originate in completely different processes than EM showers and are more complex to describe. The high mass of the charged hadrons compared to that of the electron suppresses the bremsstrahlung process by $\propto 1/m^4$, and ionization becomes the most important process for energy loss, alongside the EM interaction. As illustrated in Fig. 4.12, hadronic showers generally present an electromagnetic component and an hadronic component. The EM component originates from the production of neutral particles such as π^0 that can decay to two photons. The typical EM fraction of the shower energy is of around 30%, depending on the material, and increases with the energy of the primary particle. Interactions due to spallation, evaporation and fission processes account for the most part of the hadronic component resulting from strong interaction. Spallation is the dominant process and corresponds to the destruction of a nucleus by high-energy particles through inelastic nuclear reactions and resulting in several secondary particles. During this process, invisible energy can be lost for the calorimeter measurement due to the binding energy of the nucleus. Amongst the secondary particles that can be produced, neutrons are invisible to the calorimeter and scatter until most of their kinetic energy is lost (thermalization). The capture of these low energy neutrons by nuclei can give rise to the emission of γ rays long after the beginning of the shower.

The hadronic showers are often characterized by the nuclear interaction length λ_n that corresponds to the mean path a hadron can travel before undergoing an inelastic interaction. For most materials, λ_n is much larger than X_0 (e.g. by nearly a factor 10 for iron). They often present an early core followed by a component developing at a later time. They can start in the CE-E but develop mostly in the CE-H. Since they possess fractions of their energy that is

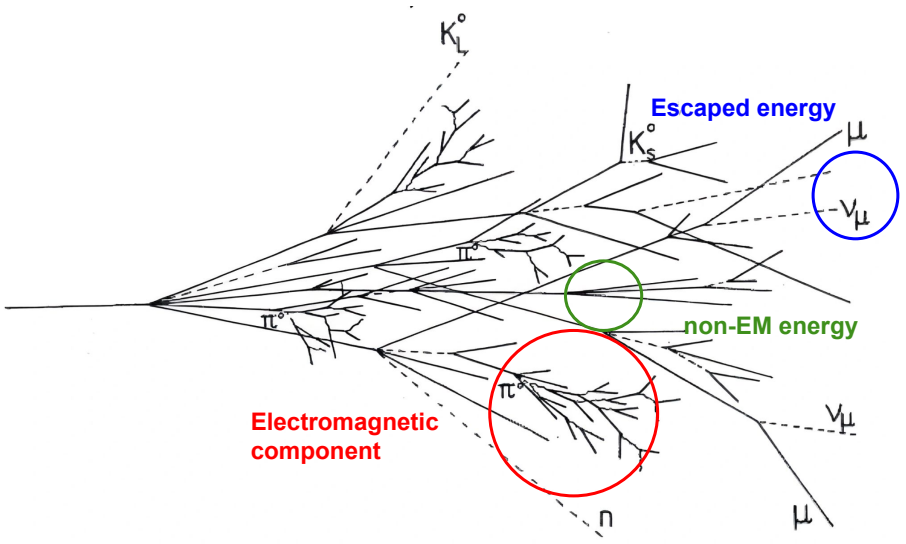


Figure 4.12: Schematic representation on an hadronic shower. They typically present an electromagnetic component from the decay of neutral hadrons in photons and an hadronic component originating from the strong interaction of charged hadrons. Escaped energy arises from the presence of neutrino that don't interact in the detector and invisible energy results from the nuclear binding energy, neutron scattering and capture. Figure taken from Ref. [129].

invisible to the detector, the calorimeter response to an EM shower and an hadronic shower of identical energy will differ.

4.5.3 Pileup

With the high luminosity of LHC Phase 2, the number of pileup interactions is expected to reach up to an average of 200 by bunch crossing. The clusters reconstructed in the HGCAL can originate from such interactions and must be rejected when identifying the physics objects. The PU consists in softer interactions compared to the hard scattering event and are thus mostly reconstructed in low-p_{T} clusters. Compared to the electromagnetic showers, they are more longitudinally spread out

4.5.4 Cluster shape variables

In the second stage of the TPG, several shape variables are computed to characterize the reconstructed 3D clusters. The computation of these variables is strongly limited by the firmware resource and latency constraints. The number of bits available for each cluster's information is limited, and in particular the shape variables are currently foreseen to be afforded only 128 bits. Thus the choice of the set of shape variables that can be computed for each cluster must be motivated by the performance they can provide to the trigger. To that end, part of this thesis work has been focused on the optimization of this set of variables for the identification of clusters reconstructed from electromagnetic showers. The baseline list of variables is detailed in Table 4.1.

Variable	Description
η	The pseudorapidity of the cluster
σ_{rr}	Standard deviation of the r-coordinates of the TCs[1]
$\sigma_{\phi\phi}$	Standard deviation of the ϕ-coordinates of the TCs[1]
$\sigma_{\eta\eta}$	Standard deviation of the η-coordinates of the TCs[1]
σ_{zz}	Standard deviation of the z-coordinates of the TCs[1]
$\langle z \rangle$	Mean of the z-coordinates of the TCs[2]
First layer	The index of the first layer in the cluster
Max layer	The index of the layer with the maximum energy deposited in the cluster
Core shower length	The number of consecutive layers with deposited energy in the cluster
Shower length	The number of layers between the first and last layers where energy was deposited
$E_{\text{max}}/E_{\text{tot}}$	The energy deposited in max layer over the total energy in the cluster
H/E	The energy of the cluster deposited in the CE-H over the energy in the CE-E
Layer $X\%$	The minimal number of layers containing $X\%$ of the cluster energy, X = [10,50,90]
N$_{TC}$X%	The minimal number of trigger cells containing $X\%$ of the cluster energy, X= [67,90]

[1] $\sigma_{xx} = \sum E_{TC}(x_{TC} - \langle x \rangle_c)^2 / \sum E_{TC}$, summing over TCs inside the cluster c.
[2] $\langle x \rangle = \sum E_{TC} x_{TC} / \sum E_{TC}$

Table 4.1: The detailed list and description of the baseline set of shape variables implemented in the simulation software.

Chapter 5

Studies on electromagnetic showers classification at the CMS L1 trigger using HGCAL trigger primitives

Amongst the physics objects that need to be identified at the CMS L1 trigger, the electromagnetic showers resulting from photons or electrons interacting in the detector are critical since leptons and photons are used in many analyses as they provide very clear signatures. Part of this thesis work has been dedicated to studies on the identification of these electromagnetic showers at the CMS Phase 2 L1 trigger by exploiting the HGCAL trigger primitives described in the previous chapter.

The identification of EM objects is foreseen to be performed by machine learning discriminators implemented on the FPGA boards of the L1 trigger. The choice of FPGA for the firmware allows the algorithms to be upgraded during the exploitation period of the detector, as they are fully re-programmable, while also providing fixed latencies compatible with the requirements of the trigger. However, the high number of operations performed by sophisticated ML algorithms can require more resources, notably logic blocks, than offered by FPGA boards. The purpose of the studies presented in this section is to answer the following questions:

- What are the most important shape variables? The performance of the ML algorithm depends on the discriminating power of the input variables and an optimal set must be identified.
- How many bits must be allocated to each of those variables? The trigger algorithms are implemented on FPGA boards that use fixed point operations. As such, the shape variables must be encoded with a fixed precision that doesn't need to be the same for all variables. Identifying how to distribute the *bit budget* between the input variables is a key handle to extract the best trigger performance.
- How does this change when accounting for the limited resources available for the model implementation? The optimal bit allocation might depend on the size of the model. It is possible that a given shape variable can be very discriminative at low precision only

when using large algorithms. Hence, a full optimization of the input set and the corresponding bit allocation must be performed while accounting for the resource limitations.

5.1 Training samples

5.1.1 Signal and backgrounds

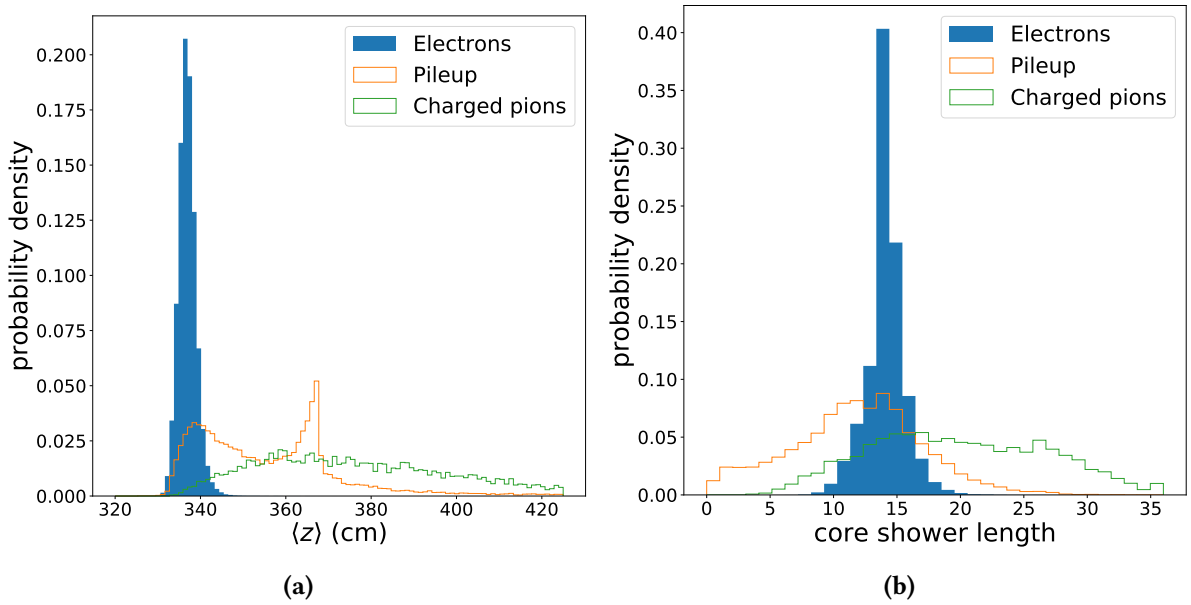
In order to reconstruct electromagnetic objects such as electrons and photons, the L1 trigger must be able to discriminate clusters originating from EM showers from those produced from hadronic showers or due to pileup. Since the HGCAL is not yet constructed at the time of this thesis, the studies presented in this section are realized with data simulated by the CMS software (CMSSW). CMSSW is a modular software that provides a full GEANT4 simulation of the CMS detector and electronics response, as well as an emulation of the trigger, resulting in an accurate event reconstruction. Modules dedicated to the Phase II of CMS have already been implemented, in particular the simulation of the HGCAL and of the upgraded L1 trigger and TPG. A ML discriminator is trained on those simulated data to discriminate the electromagnetic signal from several sources of background.

The signal sample is composed of clusters reconstructed from electrons generated with a flat p_{T} distribution between 2 and 200 GeV and a flat η distribution, on which are overlaid an average of 200 additional minimum bias interactions. The different backgrounds considered are clusters reconstructed from hadronic showers initiated by charged pions, and clusters resulting from pileup. The charged pions are simulated with a similar flat p_{T} and η distributions as the electrons and 200 PU. The PU samples is composed of clusters reconstructed from an average of 200 minimum bias interactions per event. A $E_{\mathrm{T}} > 20\,\mathrm{GeV}$ threshold is applied to this sample in order to avoid biasing the discrimination towards the low energy PU clusters.

The clusters originating from the electron showering, pion showering or jets need to be matched to the generated particle. For both pions and electrons, this is achieved by finding the candidate clusters in a cone of $\Delta R < 0.05$ around the extrapolated position of the generated particle in the HGCAL and selecting the matched cluster as the candidate with the highest p_{T}.

A selection of $1.6 < |\eta| < 2.9$ on both the reconstructed cluster and the generated particle (when applicable) pseudorapidity is performed in order to eliminate clusters outside of the HGCAL acceptance or on the edges. Additionally, only the clusters with $p_{\mathrm{T}} > 5$ GeV and generated particles $p_{\mathrm{T}} > 20$ GeV are selected. For the PU samples, the criterion on the cluster momentum is raised to $p_{\mathrm{T}} > 20$ GeV to emulate the selection on the generated particles in the other samples. This removal of the low p_{T} cluster is motivated by the tendency of the discriminator to be otherwise biased towards the low momentum range where PU dominates to the detriment high p_{T} clusters discrimination.

The differences between the distributions of the $\langle z \rangle$ and core shower length shape variables for those different samples are illustrated in Fig. 5.1. The electron clusters exhibit a narrower core and have a more compact longitudinal distribution.



(a) (b)

Figure 5.1: The distribution of (a) the $\langle z \rangle$ and (b) the core shower length (in number of layers) for electrons and the relevant backgrounds.

5.1.2 Sample pre-processing

The population in each of these samples after those selections is severely unbalanced, as reported in Table 5.1, which can impede the training of ML models. In particular, there is a factor 30 of difference between the signal sample size and the pions background sample size. To mitigate this effect, a balancing procedure based on the imbalanced-learn [130] python library was implemented.

Sample	Number of events
Electrons	224441
Charged pions	7220
Pileup	29317

Table 5.1: Number of events in the electron signal sample and in the pions and PU background samples. The population is severely unbalanced, with for example ≈30 times less pions than electrons.

The baseline technique consisting in reweighting the events such that the Sum of Weights (SOW) of the signal is equal to the background SOW was compared to other approaches :

- **Oversampling**: two techniques for increasing the size of the minority samples were evaluated:
  - The **Synthetic Minority Oversampling Technique** [131] (SMOTE) that consists in adding new points that are similar to existing points in the dataset. To that end, the process consists in building a vector between a point and one of its nearest neighbors in the minority class, and multiplying it by a random $0 < \alpha < 1$ coefficient to create a new point. The number of new events being computed that way

is tunable by selecting how many vectors are built for each point in order to reach a given amount of oversampling.

- **Adaptative Synthetic** [132] (ADASYN) which is a similar technique to SMOTE but that gives more importance to events close to the boundary between classes. The fraction of majority class points r_i amongst the nearest neighbors of a minority class point is computed. The neighborhoods with high values of r_i are more difficult to learn since they are more ambiguous for the model and are given more *weight* when building new events. For example, a region of k=8 nearest neighbors with four majority class points will be used to build twice as many new events as neighborhoods with only two majority events. This way, more data is generated for hard to learn region of the feature space.

- **Undersampling** consisting in randomly removing events in the majority class towards a given ratio with the minority class.

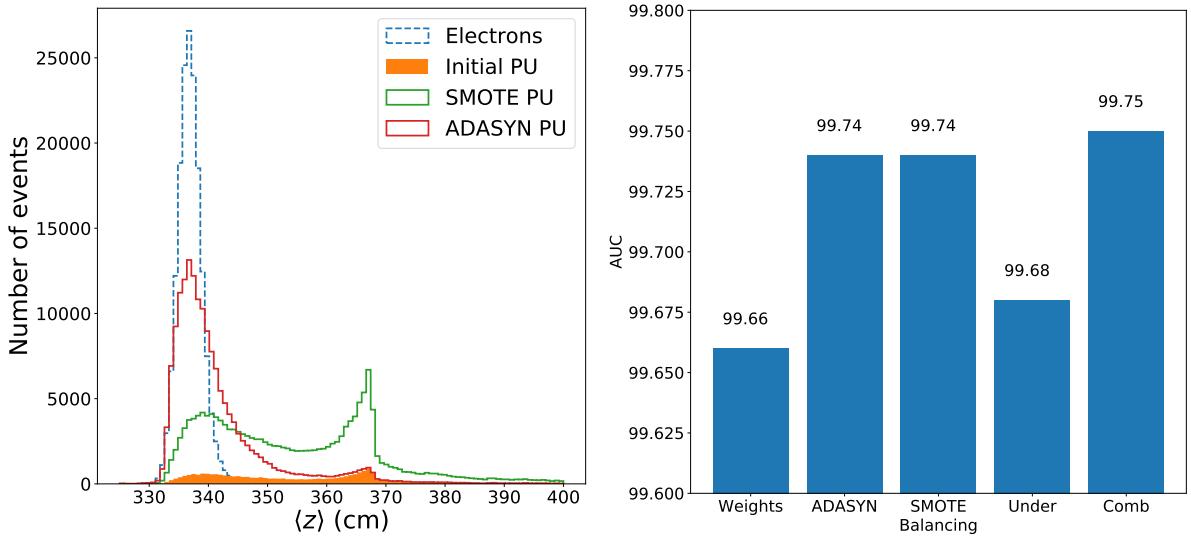
A comparison of the initial normalized pions distribution of the $\langle z \rangle$ variable and the oversampled distribution is shown in Fig. 5.2 as well as a comparison on the performance obtained for different balancing approaches. The SMOTE and ADASYN overbalancing were tuned for the minority class sample to reach 80% of the majority class. A combination of SMOTE oversampling up to 20% of the majority class followed by an undersampling of the electron sample down to twice the number of events in the background was also considered. Both oversampling approaches and the combination of oversampling and undersampling perform similarly and the combined method was selected.

5.2 Choice of Machine Learning model

The optimization of the HGCAL TPG requires to understand how the cluster information will be exploited in the central CMS L1 trigger. Multivariate algorithms are foreseen to be employed in order to extract and combine the information from multiple shape variables and perform various tasks such as discrimination. In particular, Boosted Decision Trees (BDTs) are currently foreseen to be employed for the identification of EM e/γ clusters. They are known to have successfully been implemented on FPGA and can be relatively small. While BDTs are known to exploit very well the information of the shape variables, it was thought that Neural Networks could improve performance using lower level information of the detector while also being known to be implementable on FPGA. In order to understand if the combination of low level variables and neural network could perform better than BDTs using the high level cluster shape variables, the two architectures were evaluated for the discrimination of electron clusters and charged pion clusters with two input sets:

- The information of the cluster p_{T} deposited in each layer of the HGCAL
- The per-layer p_{T} values used in the previous set and the cluster shape variables presented in Table 4.1.

The neural networks are trained with Keras [108] using TensorFlow 2 [109] backend on GPU. A fully connected architecture was selected as it is particularly well adapted for discrim-



(a) (b)

Figure 5.2: (a) Comparison of the initial distribution of $\langle z \rangle$ for the PU (orange), electrons (blue) and the same PU distribution after SMOTE (green) and ADASYN (red) oversampling. The new events created by the SMOTE algorithm maintain the shape of the distribution while ADASYN mainly creates new events in the hard-to-learn region of the feature space. (b) Comparison of the area under the ROC curves for BDTs trained on sample balanced according to different techniques. The combination of SMOTE oversampling and random undersampling yields the best AUC.

ination tasks using a relatively small number of high level variables. The number of hidden layer was optimized as well as the number of neurons per layer, with a limitation on the total number of neurons in the model that should not exceed 10000 to limit the resource usage for the model implementation on FPGA. The output layer is composed of a single neuron with sigmoid activation. The hyperparameters were optimised by Bayesian optimization (see Sec.2.6.4) and the overtraining carefully constrained by dropout layers and regularization techniques (both L1 and L2). It has to be noted that the optimization of the NN was hindered by the intrinsic non-deterministic nature of the GPU training. The numerous local minima in the cost function make the optimization very sensitive to starting conditions and different parameters could be selected if changing initialization. The discrimination power is however stable as those local maxima yield very similar performance.The BDTs are trained using eXtreme Gradient Boosting [79] (XGBoost) with the binary logistic objective function. The parameters of the BDT were optimized by grid search. The details of the optimized architecture for both types of models are given in Table 5.2.

A good trigger algorithm must be able to reject most background events while still retain good signal efficiency. Thus, the performance of the models is measured with the area under the ROC curve, and by assessing the background efficiency at a signal efficiency working point of 0.99. Those results are shown in Fig. 5.3 and the resulting background efficiencies are reported in Table 5.3. The resulting background efficiencies for 99% signal efficiencies are between 1% and 2%. While the BDTs background rejection rate improves (-60% of background event being misidentified) with the addition of the shape variables, it is not the same for the DNN where the performance are even degraded (+15% background efficiency). The informa-

Hyperparameter	Optimised value
Boosted Decision Tree	
Learning rate	0.2
Maximum tree depth	4
Events subsampling by tree	0.8
Input variable subsampling by tree	0.8
Number of boosting rounds (number of trees)	81
Multi layer perceptron (layer p_{T} values only)	
Optimizer	Adam[81]
Learning rate	0.001
Dropout rate	0.3
Neurons in hidden layers	[2500, 750]
Multi layer perceptron (shape variables and layer p_{T} values)	
Optimizer	Adam
Learning rate	0.002
Dropout rate	0.2
Neurons in hidden layers	[2350,74,74,74,74,74,74,74,74,74,74]

Table 5.2: The optimized values of the hyperparameters for the different discriminators. For the neural network, the optimization was performed independently for the two sets of input features: the layer p_{T} values only and shape variables + layer p_{T} values since NN are expected to be able to exploit low level information. The maximum number of neurons in the NN was fixed at 10000 to limit the total size, and the size of the first layer and subsequent layers (that all use the same number of neurons) are allowed to vary under this constraint. Each dense layer use the leaky ReLu activation function and the output neuron uses sigmoid activation function. While it is expected that the second set of inputs would require a deeper DNN, the widely different number of layers is likely to be due to different local maxima found during the optimization.

tion of the shape variables does appear to be critical for extracting the best discrimination power from BDTs. The degradation of the DNNs performance however was not expected. While the DNNs performs better than the BDT when using the low level information, as predicted, it appears that not only the additional high level information does not improve the performance but even degrade them. Since most of the information in the shape variables is coming from the layer p_{T} values, this means that the DNN had a harder time learning when exposed to larger and redundant input. It is well possible that the choice of architecture, in particular the constrained size of the model, is not optimal to fully exploit this input set. However, since the resources used for the implementation of the discriminator for the L1 trigger are indeed constrained, this points to the fact that fully connected networks are not the optimal choice of algorithm for this task.

In accordance to those results, the usage of DNNs and low level information does not seem justified, and the studies detailed in the following sections only use BDTs for e/γ identification, as planned in the L1 trigger TDR [61]. It does not indicate, however, that the DNNs must be completely abandoned, and adequate choices of training inputs and architecture could yield

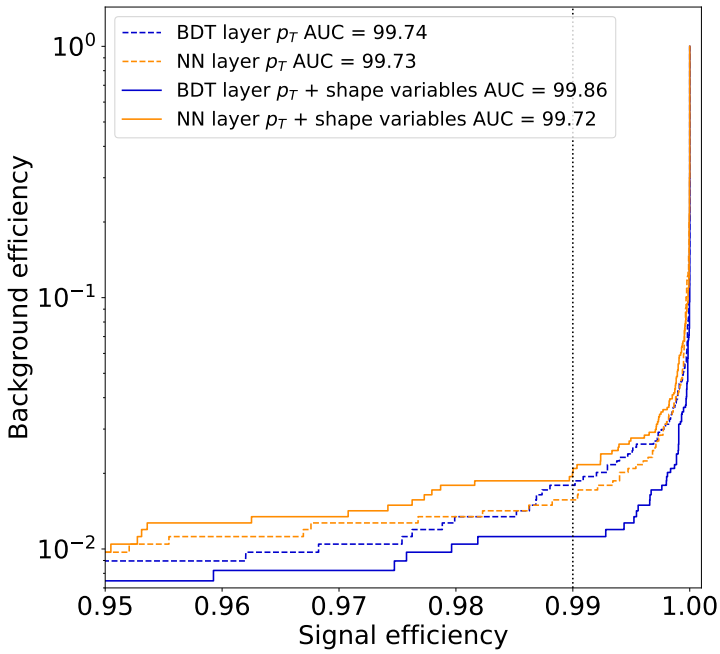


Figure 5.3: Comparison of the ROC curves for the BDT and DNN trained on the layer p_{T} values only (dashed lines) and layer p_{T} values + shape variables (full lines). While the BDT efficiency improves (-60% background efficiency) with the inclusion of the shape variables, the neural network doesn't profit as much from the high level information that even degrades slightly (+15% background efficiency) its performance. The blue dotted line corresponds to the chosen working point of 99% signal efficiency.

Model	Background efficiency at 0.99 WP	
	Layer p_{T} values only	Cluster shapes + layer p_{T}values
BDT	1.79%	1.12%
DNN	1.56%	1.79%

Table 5.3: The background efficiencies at a signal (electrons) efficiency working point of 99%. The best background rejection is obtained for the BDT trained with layer p_{T} and cluster shape variables.

to performance improvements. Several attempts using more sophisticated NN architectures are currently being studied, in particular with graph neural networks [133] that could fully exploit the image-like structure of the HGCAL, but could pose significant challenges towards their implementation on the trigger FPGAs due to their resource utilization and latency.

5.3 Selection of inputs variables implementable in HGCAL TPG

As seen in the previous section, the choice of the input features used to train the discriminators can heavily impact the performance of the algorithm. In a context where the bandwidth for sending the input variables to the trigger is limited (currently foreseen to use 128 bits per cluster for the shape variables), it is critical to assess the impact of those variables on the performance and select the relevant ones. The shape variables that can be computed in the TPG are also limited by the architecture and latency of the TPG system. The primitives are computed from cluster quantities called fields, for example the number of TCs in the cluster, the total sum of energy in the cluster or in the electromagnetic part of the calorimeter. With the current set of fields, some of the shape variables described in Table 4.1 can not be computed. Removing them from the input set however causes a significant loss of performance with 18% and 35% higher background efficiency. Alternative variables using the available fields have been designed to mitigate this degradation of the performance and were implemented in CMSSW.

Those alternative variables are as follow:

- **The Variances Var**$_{xx} = \sigma^2_{xx}$ were used to replace the standard deviations. The variances are one of the primitive fields and their usage in place of the σ_{xx} saves a square root operation.

- **E/(E+H)** : this variable is implemented to replace the H/E that could be ill-defined for clusters with no energy deposited in the CE-E.

- **Layer fractions** : the summation of the cluster energy in the $N \in [1, 5]$ first layers of the CE-E, $N \in [1, 5]$ first layers of the CE-H, and in the $N \in [1, 5]$ last layers of the detector. The sums are then normalized to the cluster total energy.

- **E**$_{\max}$ **sums**: The cluster energy fraction in the $N \in [1, 5]$ layers around shower maximum of electromagnetic showers. This maximum was determined from the longitudinal profile of the electrons between 2 and 200 GeV. In the case of even values of N, the summation can not be symmetric around the maximum, and as such a *left* and a *right* variations are defined.

- **Energy bitmaps**: bitmaps encoding the presence or absence of energy in each layer of the CE-E or the CE-H with the process illustrated in Fig. 5.4. The CE-E and CE-H compartments are treated independently so that the bitmaps can be encoded into a single 32-bits integer. Since in typical clusters some low amount of energy is deposited in most layers of the CE-E, a variation of the bitmap considering only the layers above a 1 GeV threshold is defined. By construction, the bitmap value is more strongly affected by the last layers, for example, the presence of energy in the 10^{th} layer changes the output by 512 while in the second layer it only shifts the output by 2. A variation of the bitmap considering the last layer as the first bit as been considered as a "reverse" bitmap, potentially highlighting more the variations in the first layers of the detector.

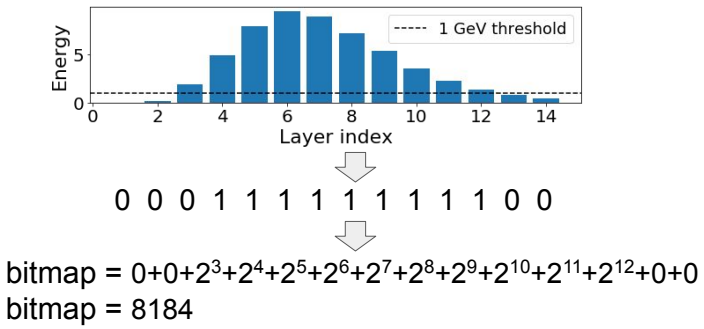


Figure 5.4: Illustration of the computation of the bitmap variable, here for the CE-E bitmap with 1 GeV threshold. For each layer N where the cluster has deposited more than the threshold value (0 or 1 GeV) of energy, 2^N is added to the bitmap.

The performance of the algorithms trained on different inputs set is compared independently for each background:

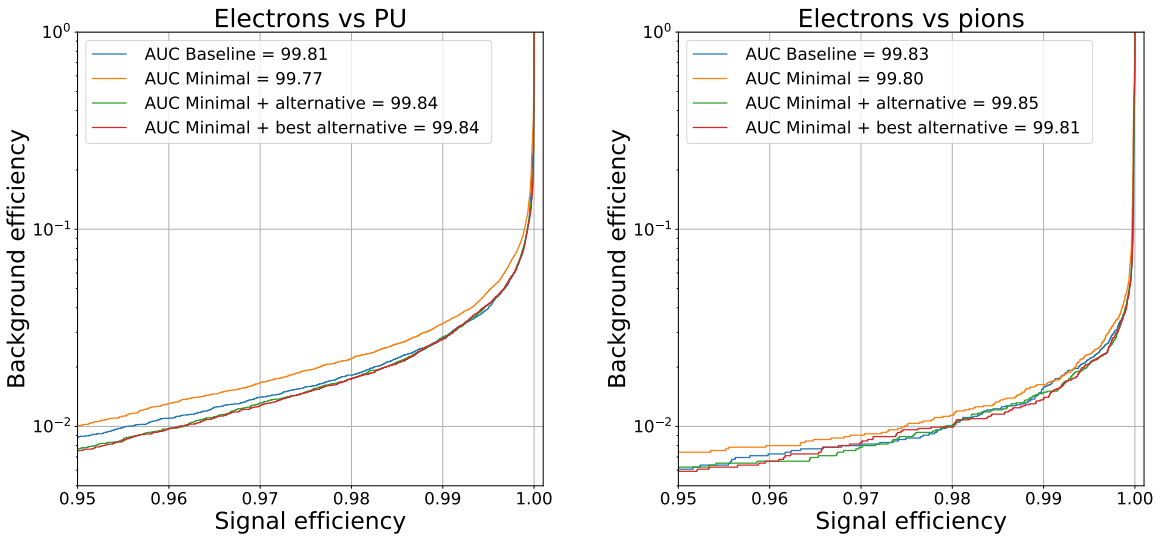
- the 17 baseline shape variables
- a minimal set of 10 inputs retaining only the baseline variables that could be computed from the available cluster fields. It is composed of the variances in the r, ϕ, η and z coordinates, the index of the first layer, the length of the shower and its core, the fraction of energy in the CE-E, the absolute value of the cluster pseudorapidity and the mean of the z-coordinate.
- the minimal set completed with the new alternative variables for a total of 42 variables
- the minimal set and the six best variables amongst the alternatives.

The resulting ROC curves are shown in Fig. 5.5. The inclusion of the alternative variables allow to recover the baseline performance, but add up to a large input set with redundant variables. The final set of input variables is selected by adding the most important features amongst the alternative variables to the minimal set. The importance is measured in terms of SHAP (SHapley Additive exPlanations) values, a game theoretic approach to explain the output of any machine learning model [114]. It computes on a per-event basis the marginal effect of each feature, how much it is driving the output towards background-like (a negative shap value) or signal-like (a positive shap value). The importance ranking is determined by taking the mean of the absolute shap values on a subset of events. The detailed list of these input sets and the variable ranking is detailed in Table 5.4 for each of the backgrounds.

Variable	Background	
	Pileup	Charged pions
$\|\eta\|$	6	4
Var_{rr}	1	5
$\mathrm{Var}_{\phi\phi}$	5	11
$\mathrm{Var}_{\eta\eta}$	11	16
Var_{zz}	16	14
$\langle z \rangle$	15	1
First layer	12	13
Core shower length	8	6
Shower length	14	15
E/(E+H)	2	2
Energy fraction in CE-H first layers	13	9
Energy fraction in CE-H first 5 layers	10	-
$E_{\mathrm{max}}^{2,R}$ [1]	7	-
$E_{\mathrm{max}}^{4,R}$ [2]	4	12
E_{max}^{5} [3]	9	7
CE-E bitmap 1 GeV threshold	3	3
"reverse" CE-E bitmap 1 GeV threshold	-	10
CE-H bitmap 0 GeV threshold	-	8

[1] : The fraction of the energy deposited in the layers [6,7].
[2] : The fraction of the energy deposited in the layers [5,8].
[3] : The fraction of the energy deposited in the layers [4,8].

Table 5.4: The list of input variables for the training of BDTs to discriminate electrons from the various backgrounds. The input set is composed of the minimal feature set completed with the best alternative variables against that particular background. Each variable is shown with its ranking in terms of SHAP values and the variables that counts in the five most important variables for at least one background are highlighted in blue. An hyphen (-) indicates the variable is not used in the input set for this background.



(a) **(b)**

Figure 5.5: The ROC curves for BDTs trained with different input variables for (a) PU discrimination and (b) charged pions discrimination. Only the region at high signal efficiency (> 0.95) is shown. The blue curves correspond to the baseline feature set, the orange one is the minimal set with all variables that can not be easily computed from the cluster fields removed from baseline, the green one is the minimal set with the addition of all alternatives variables, and the red one is the minimal set with the addition of the most important variables amongst the new alternative ones.

5.4 Optimization of the shape variables precision

The relative importance of the cluster variables presented in the previous section does not account for the limitation due to the design of the CMS L1 trigger. In particular the computation of those variables is made with a floating point format while FPGA only handle fixed point operations. Thus a *quantization* of those shapes variables must be performed and might impact their discrimination power. Since the shape variables must be encoded on a maximum of 128 bits per cluster, an optimization of the bit budget attributed to the variables must be performed with respect to the performance. Because BDTs are structured in series of comparisons between the value of a variable and a threshold, further called *splits* in the tree ensemble, the number of bits used for each variable can be optimized independently.

The number of bits attributed to each input can be interpreted as an external parameter of the model and thus optimized alongside other hyperparameters such as the maximum depths of the trees. This optimization however must not only focus on reducing the loss of the model as in classical optimization problems, but also minimizing a second objective defined as the total number of bits used for encoding the model's inputs. Such optimization task with more than one objective can be handled by multi-objective optimization (MOO) methods as described in the next section.

5.4.1 Multi-objective optimization

While ML models automatically optimize their internal parameters in order to perform a task, their performance also depend on external parameters that also need to be optimized. Several methods have been developed to maximize the efficiency, such as the Bayesian optimization, and perform very well for optimizing a single objective function such as the background rejection. However, it is often the case that real life problems are subjected to constraints that can take the form of additional objectives, for example minimizing the costs or resource usage. Those different objectives can often be competing with each other, and the solution of handcrafting a function combining the different objectives can be arbitrary and thus unsatisfying. Multi-objective optimization are a class of method that aim to solve this problem by finding a set of solutions, i.e. external parameters values, that define the best trade-off between different objectives.

When no single solution exists that can simultaneously maximize all of the objectives of a problem, MOO aims to determine the set of non-dominated – i.e. optimal – solutions. A set of parameters is considered optimal when improving one of the objective functions requires to degrade an other one, as illustrated in Fig. 5.6. In general there exists an infinite number of non-dominated solutions, where none is superior to the others without an additional subjective criterion. The set of all those optimal solutions is called *Pareto's front.* MOO methods are a class of algorithms whose goal is to find solutions as close as possible to this Pareto boundary, allowing for the decision maker to make an informed choice.

Genetic or evolutionary algorithms are a type of algorithms especially suited to this kind of optimization. They draw inspiration in the natural selection mechanism to evolve towards optimal solution. The Non-dominated sorting algorithm II[134] (NSGA-II) is a prime candidate

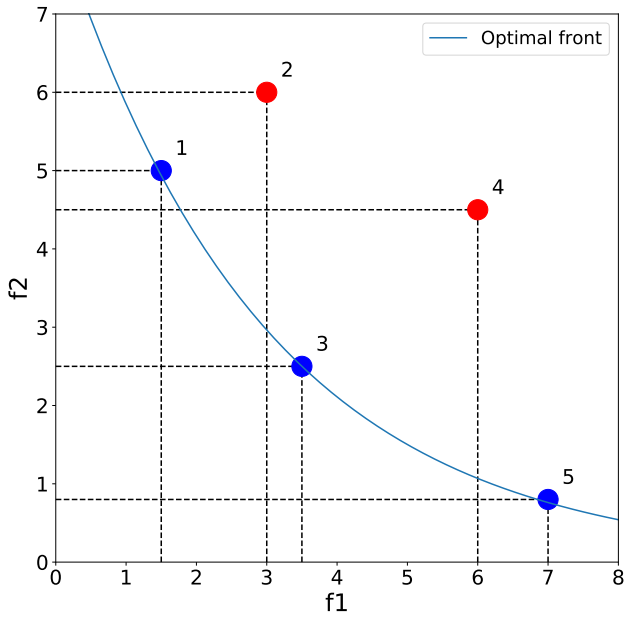


Figure 5.6: An illustration of domination in a multi-objective problem with two objective functions f1 and f2 that must be minimized. The blue curve corresponds to the optimal or Pareto's front, which represents the best trade-offs possible. In this case, amongst the five points drawn here, point 2 is dominated by point 1, and point 4 is dominated by point 3. The set of non-dominated solutions corresponds to the blue points.

for such a task that can roughly be summarized in the following steps:

- **Step 1 : Initializing the population.**
  Creating an initial population according to the problem phase space and constraints.

- **Step 2 : Population sorting**
  Non-dominated sorting of the individuals in the initial population.

- **Step 3 : Selection**
  Individuals are selected by tournament selection: randomly pick groups of individuals and select the best pairs as *parents*. The selection is based on domination criterion and crowding distance. The crowding distance provides an estimate of the density of solutions surrounding that solution and is defined as the Manhattan distance in the objective space.

- **Step 4 : Reproduction**
  The selected individuals are combined two by two by applying crossover and mutation operators. The crossover mixes the parameter values of the two *parents* and mutation gives a chance for the parameter to take a different value than that of the parents, allowing for better exploration of the parameter space.

- **Step 5: New generation**
  The *offsprings* form a new *generation* that is once again sorted and reproduced until a termination criteria (e.g. number of generations) is reached.

5.4.2 Impact of the inputs precision level on the HGCAL e/γ performance

In addition to maximizing the e/γ identification performance obtained with BDTs trained on the HGCAL TPG output, MOO can simultaneously optimize the precision used for each variable. Two objective functions are defined and the NSGA-II algorithm is employed to find the best trade-offs between them.

The first objective is related to the discrimination performance. The efficiency of a trigger algorithm is in general measured by the trigger rate. Without direct access to these trigger rates in this study, the figure of merit used to evaluate the performance of the algorithm is the area under the ROC curve above a 80% signal efficiency threshold. This choice allows to integrate the background rejection rates for all high signal efficiency threshold, while not considering effects at low signal efficiency.

The second objective function, that needs to be minimized, accounts for the limited bandwidth between the TPG and the central L1 trigger. It is computed as the sum of the number of bits attributed to each of the cluster shape variables. Since the algorithm will be implemented on FPGAs that use fixed point quantities, the variables are *quantized* so that a quantized-aware training of the XGBoost model can be performed. Precisions in the [0,16] bits are considered, with a value of 0 bits for a particular variable meaning that it is removed from the input set. The quantization is performed by casting the feature in 2^N bins, constructed by selecting uniform intervals between the second and 99^{th} percentiles of the feature range, and values outside the range are put in the first or last bin. The first and last percentiles are removed in order to lessen the impact of outliers on the bin definition. Some of the shape variables are discrete and computed with an integer format, like the bitmaps or layer indices. Their range can therefore be smaller than 16 bits by construction, in which case the maximum quantization precision is set to the feature range. An example of the effects of quantization is given on the E/(E+H) variable in Fig. 5.7. Because of the way the CE-E and CE-H bitmaps are implemented, such a quantization would heavily degrade the information they contain. Instead of applying the quantization process described above, the precision for those variables is decreased by fusing a certain number of consecutive layers before computing the bitmap. For example, to reduce the number of bits by a factor two, pairs of neighboring layers are fused. As such, not all levels of precision can be used for those variables.

The parameter optimization is performed with the NSGA-II algorithm implemented by the pymoo [135] python package. The initial population is composed of an hyperdiagonal sampling (16 individuals, one for each possible bits precision) completed by random sampling for a total of 60 individuals. Each successive generation is also composed of 60 individuals, until the 40^{th} generation is reached. Duplicated individuals in a generation are eliminated. The reproduction is performed by crossover and mutation operators:

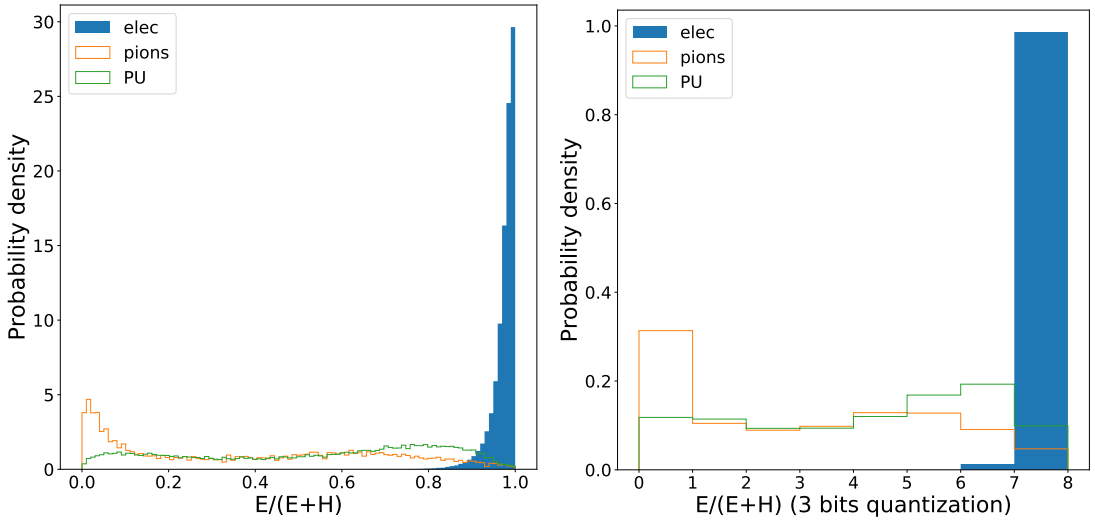


Figure 5.7: A comparison of the binned distributions for the fraction of the cluster energy left in the CE-E (a) unquantized and (b) quantized with 3 bits, for the signal and the different backgrounds.

- The **crossover**, sometimes called recombination, is used to combine the parameters of two parent solutions to generate a new solution. For this optimization, simulated binary crossover [136] (SBX) with a crossover index $\eta = 3$ is employed. In many applications of genetic algorithms, the parents parameters are binary encoded, meaning that for each parameter the value of parent A is 0 and the value of parent B is 1. Such binary-coded genetic algorithms often employ single-point crossover, meaning that the parameter vectors from both parents are split at a random crossing point and the left part of one parent vector is mixed with the right part from the other parent. The position of this crossing point can be described by a probability distribution. The SBX is used to simulate this probability distribution, generalized to variables that are not binary encoded. As such, each children parameter value follow a probability distribution which peaks around the values of this parameter for each of the parents, as illustrated in Fig. 5.8. The η parameters dictates how much the distribution are peaked around the parents values, with higher η meaning higher chances for children values that are close to one of the parent value. Since the shape variables are encoded in integer format, a rounding of the children value is applied.

- The **mutation** is employed to ensure that the parameter space is well explored by creating for new values of child parameters than that of the parents. The polynomial mutation is employed for this optimization after the crossover has been applied. It follows the same probability distribution as SBX, around a single pole defined as the parameter value obtained for a child solution with the crossover operator. The η parameter is set to $\eta = 3$

The algorithm converges well, as illustrated in Fig. 5.9 for the training against PU background. The last generation is composed of different sets of parameters providing the best trade-offs between algorithm performance and number of bits used for the inputs.

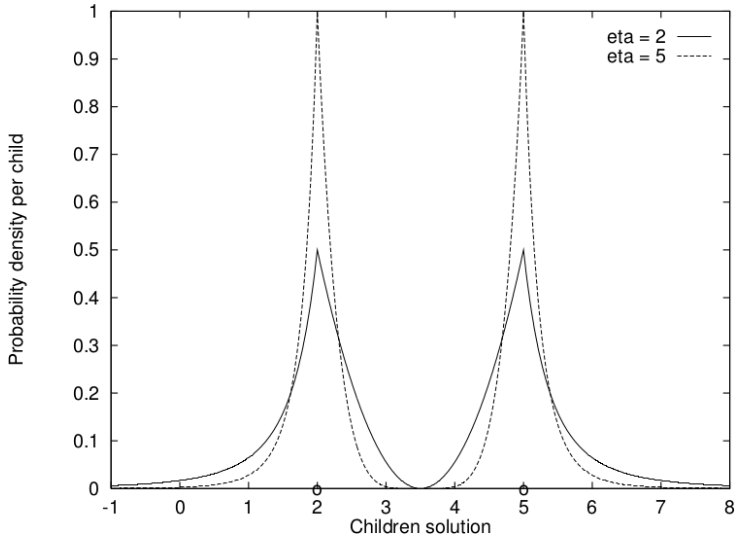


Figure 5.8: The probability density function for setting the value of an child parameter with SBX. The circles at 2 and 5 correspond to the parent values. Higher values of η improve the probability of values close to that of the parents. Figure taken from Ref. [137].

The solutions that can potentially be selected are those that perform well for both objectives at the same time. The members of the last generation that simultaneously perform well in both objectives are selected by applying cuts on AUC > 99.6% and N <50 . Those solutions are considered *optimal* and are detailed in Fig 5.10 for the training against PU, with the number of bits per variable and resulting performance. Several conclusions can be drawn from their scrutiny. First of all, the total number of bits ranges from 19 to 44, well under the 128 bits budget foreseen for the cluster shapes by CMS. A reduction of the total number of bits from 256 (if all variables used 16 bits precision) to 44 barely impacts the resulting performance, that only decreases from 99.83% to 99.82%. Moreover, while the total number of bits is correlated to the number of features that are kept, the number of bits per variable is not distributed evenly amongst the shape variables. In particular, it seems that the CE-E bitmap is given 11 bits in all solutions (between 25% and 50% of the total number of bits) and is always used. This means that though this variables requires a large number of bits, removing it from the inputs impacts the performance too much to be considered in optimal solutions. There is also a correlation between the variables that are attributed a high number of bits and the variables that were identified as important in terms of SHAP values. The highest precision variables are the CE-E bitmap, the Var_{rr} and E/(E+H) which were respectively the third, first and second most important variables in terms of SHAP values.

On the other hand, some variables like the shower length can be used with a very low number of bits, sometimes only one, and even be dropped in several solutions. Overall, a high number of inputs can be discarded, most optimal solutions keeping only around eight to ten shape variables out of 16. The E/(E+H) fraction, that was one of the most important variable before minimizing the number of bits, is dropped in several solutions. This shows that the variables is less *bits-effective* than other variables in the input set can provide the same information. After the bitmap, such as the Var_{rr}, E/(E+H), $|\eta|$ and $\text{Var}_{\phi\phi}$ are afforded the highest number of bits, but tend to be dropped in solutions with lower total of bits. In particular, it seems that when the precision for those is decreased under a critical threshold,

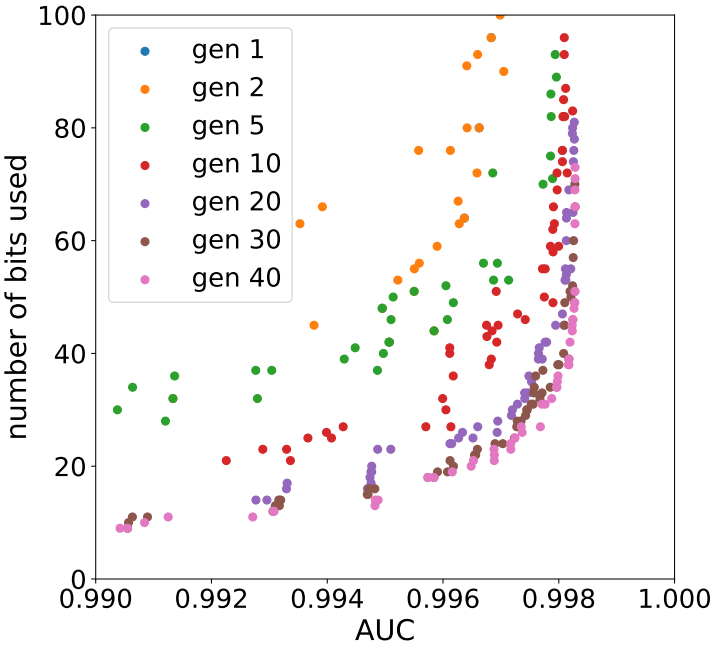
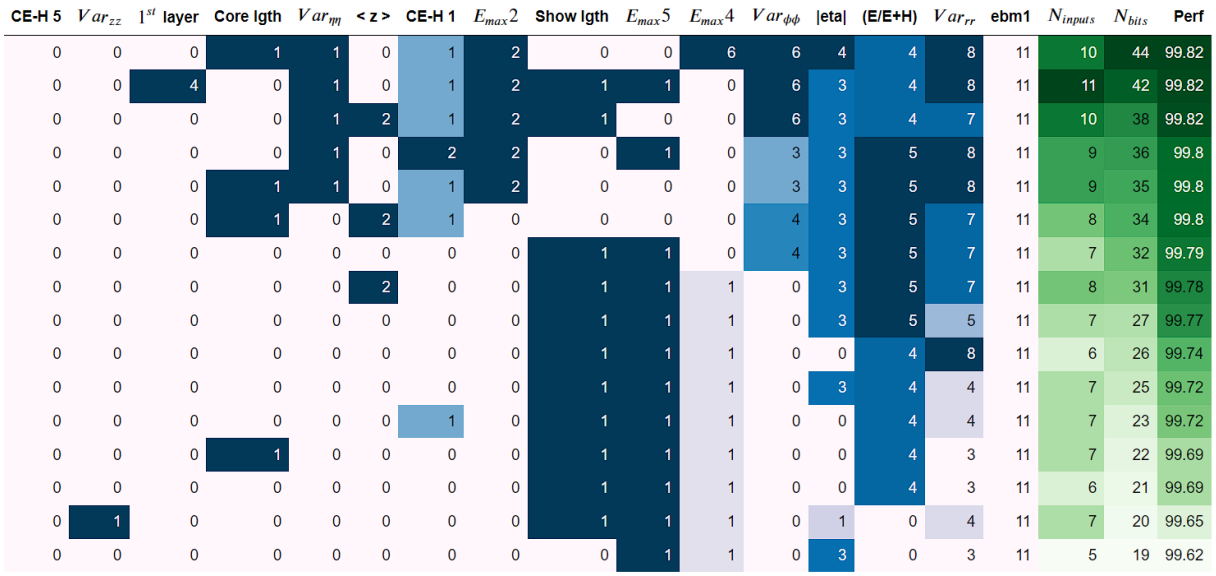


Figure 5.9: Performance of the algorithm (measured by the AUC of the ROC curve above a 80% signal efficiency) against the total number of bits used to train a BDT discriminating between electrons and PU. The MOO converges well as each generation's population gets closer to the optimal bottom right corner.

like 3 bits for $\text{Var}_{\phi\phi}$, affording at least one bit in different variables such as the shower length is more efficient. It has to be noted also that all optimal solutions remove a least a few variables.

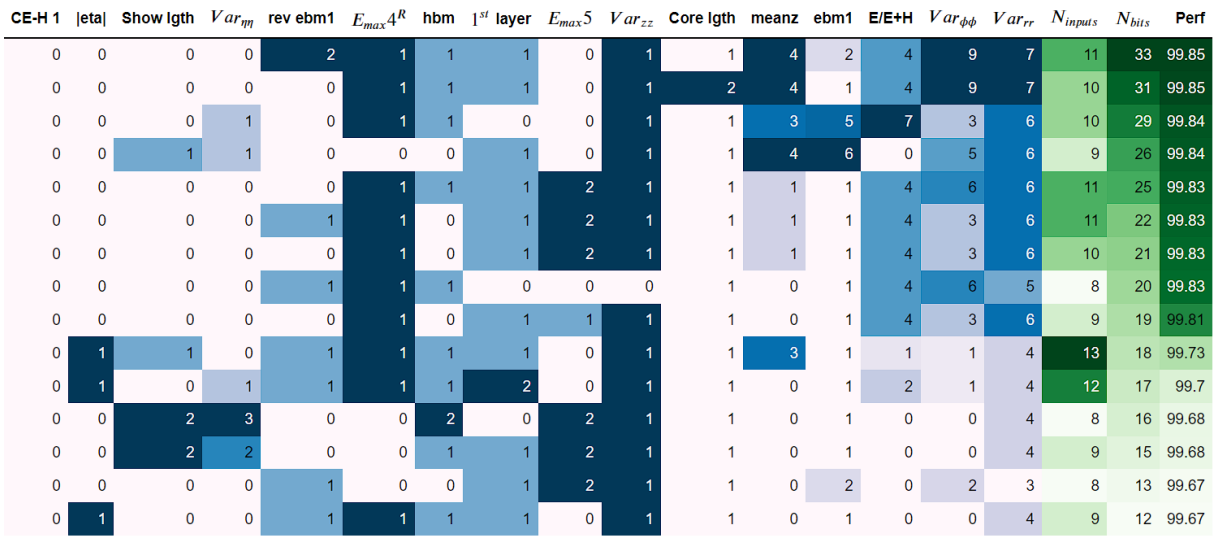
Those conclusions are drawn from the discrimination of electron clusters versus PU clusters which are the main source of trigger rates. Similar conclusions can be drawn from the results based and pion background, shown in Fig 5.11. The highest number of bits are also concentrated on a few variables, amongst which the $\text{Var}_{\phi\phi}$, Var_{rr} and E/(E+H) . The CE-H 1 variable is never used in optimal solutions, and the $|\eta|$ and shower length are seldom used, and with only very small precision when it is the case. The $\langle z \rangle$ of the cluster is now dropped in several optimal solutions while it was the most important variable in terms of SHAP values for discriminating electrons and pions. The CE-E bitmap requires lower precision for the discrimination of pions cluster. That can be explained by the fact that pions leave the major part of their energy in the CE-H, so a high granularity in the CE-E is not required.

The optimization against the pion sample manages to reduce more drastically the total number of bits which ranges from 12 bits to 33 for the optimal solutions, with similar AUCs as for the PU sample. As expected, this shows that the clusters reconstructed from charged pions are more dissimilar to electromagnetic clusters than PU clusters, and can be easily discriminated with few variables such as the energy fraction in the CE-E and the cluster variances in ϕ and r.



CE-H 5	Var_{zz}	1^{st} layer	Core lgth	$Var_{\eta\eta}$	< z >	CE-H 1	$E_{max}2$	Show lgth	$E_{max}5$	$E_{max}4$	$Var_{\phi\phi}$	|eta|	(E/E+H)	Var_{rr}	ebm1	N_{inputs}	N_{bits}	Perf
0	0	0	1	1	0	1	2	0	0	6	6	4	4	8	11	10	44	99.82
0	0	4	0	1	0	1	2	1	1	0	6	3	4	8	11	11	42	99.82
0	0	0	0	1	2	1	2	1	0	0	6	3	4	7	11	10	38	99.82
0	0	0	0	1	0	2	2	0	1	0	3	3	5	8	11	9	36	99.8
0	0	0	1	1	0	1	2	0	0	0	3	3	5	8	11	9	35	99.8
0	0	0	1	0	2	1	0	0	0	0	4	3	5	7	11	8	34	99.8
0	0	0	0	0	0	0	0	1	1	0	4	3	5	7	11	7	32	99.79
0	0	0	0	0	2	0	0	1	1	1	0	3	5	7	11	8	31	99.78
0	0	0	0	0	0	0	0	1	1	1	0	3	5	5	11	7	27	99.77
0	0	0	0	0	0	0	0	1	1	1	0	0	4	8	11	6	26	99.74
0	0	0	0	0	0	0	0	1	1	1	0	3	4	4	11	7	25	99.72
0	0	0	0	0	0	1	0	1	1	1	0	0	4	4	11	7	23	99.72
0	0	0	1	0	0	0	0	1	1	1	0	0	4	3	11	7	22	99.69
0	0	0	0	0	0	0	0	1	1	1	0	0	4	3	11	6	21	99.69
0	1	0	0	0	0	0	0	1	1	1	0	1	0	4	11	7	20	99.65
0	0	0	0	0	0	0	0	0	1	1	0	3	0	3	11	5	19	99.62

Figure 5.10: The list of solutions that optimize both objectives for the discrimination of electrons versus PU clusters, ordered by performance. The 16 first columns correspond to the number of bits used for each variable, sorted by their average number of bits. The individual values are of darker color when the value is high relative to the range of precision used for this variables.



CE-H 1	|eta|	Show lgth	$Var_{\eta\eta}$	rev ebm1	$E_{max}4^R$	hbm	1^{st} layer	$E_{max}5$	Var_{zz}	Core lgth	meanz	ebm1	E/E+H	$Var_{\phi\phi}$	Var_{rr}	N_{inputs}	N_{bits}	Perf
0	0	0	0	2	1	1	1	0	1	1	4	2	4	9	7	11	33	99.85
0	0	0	0	0	1	1	1	0	1	2	4	1	4	9	7	10	31	99.85
0	0	0	1	0	1	1	0	0	1	1	3	5	7	3	6	10	29	99.84
0	0	1	1	0	0	0	1	0	1	1	4	6	0	5	6	9	26	99.84
0	0	0	0	0	1	1	1	2	1	1	1	1	4	6	6	11	25	99.83
0	0	0	0	1	1	0	1	2	1	1	1	1	4	3	6	11	22	99.83
0	0	0	0	0	1	0	1	2	1	1	1	1	4	3	6	10	21	99.83
0	0	0	0	1	1	1	0	0	0	1	0	1	4	6	5	8	20	99.83
0	0	0	0	0	1	0	1	1	1	1	0	1	4	3	6	9	19	99.81
0	1	1	0	1	1	1	1	0	1	1	3	1	1	1	4	13	18	99.73
0	1	0	1	1	1	1	2	0	1	1	0	1	2	1	4	12	17	99.7
0	0	2	3	0	0	2	0	2	1	1	0	1	0	0	4	8	16	99.68
0	0	2	2	0	0	1	1	2	1	1	0	1	0	0	4	9	15	99.68
0	0	0	0	1	0	0	1	2	1	1	0	2	0	2	3	8	13	99.67
0	1	0	0	1	1	1	1	0	1	1	0	1	0	0	4	9	12	99.67

Figure 5.11: The list of solutions that optimize both objectives for the discrimination of electrons versus pion clusters, ordered by efficiency. The 16 first columns correspond to the number of bits used for each variable, sorted by their average number of bits. The individual values are of darker color when the value is high relative to the range of precision used for this variables.

5.4.3 Impact of the model complexity on the optimization

While this optimization gives an outlook on the possible bit budget required by each variable, and their relative importance when the total number of bits is constrained, it is possible that some of the conclusions may be only true for very complex models that could not be implemented given the resources available in the trigger FPGAs. A more complete optimization must account for these limitations. The complexity of the model, or more precisely the amount of resources needed for its implementation in the trigger FPGAs, can be added as a third objective function to the MOO.

To that end, an estimation of the resources used for the BDT implementation must be performed. Conifer [138] is a software that can translate a python-trained model, like XGBoost BDTs into FPGA firmware implementation via conversion into Very High-Speed Integrated Circuit Hardware Description Language (VHDL). The synthesized model reproduces very well the performance from the python model as illustrated in Fig. 5.12. The resources needed for the firmware implementation of the VHDL model can then be estimated with software such as Xilinx Vivado suite [139] to provide a report on the utilization estimates for every type of FPGA resources. For BDTs, the limiting factor is the Look Up Table (LUT) utilization, which will therefore be minimized during the optimization.

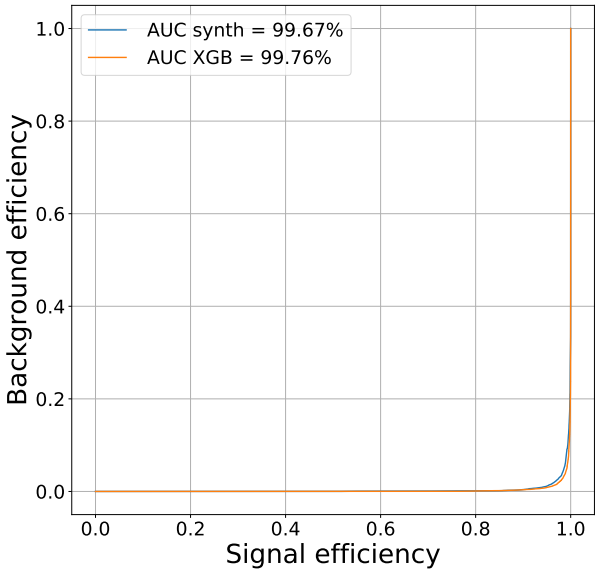
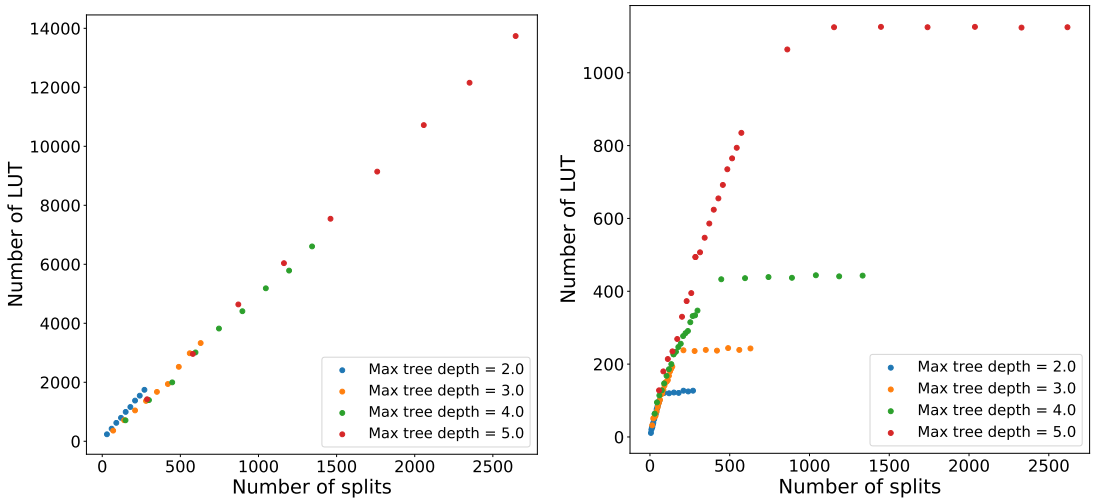


Figure 5.12: Comparison of the ROC curves between the XGBoost model and the model synthesized by conifer. (a) shows a XGBoost model trained on unquantized data and its synthesis performed with 16 bits precision. The synthesized model performance are very similar to the XGBoost model.

However, the conifer software currently uses only a single global level of precision for all the quantities in the FPGA implementation. This means the precision of all the features, the precision used for the nodes thresholds and that of the coefficient multiplying each following trees are the same. This limitation forbids the correct synthesis of BDTs with different precisions for each input. To circumvent this problem, a *proxy* was constructed in order to evaluate the resource utilization of such a BDT by estimating the number of LUT required for each *splits* in the BDT at a given precision level.

To that end, BDTs with different sizes and precision levels are trained and synthesized with conifer to estimate the total number of LUTs used and extract the ratio $R = N_{\text{LUT}}/N_{\text{splits}}$. The ratio is treated separately for each maximum tree depth value to account for the dependence of the gradient on this parameter. While it was expected that the number of LUT used for the model implementation should grow roughly linearly with the number of splits -and it was verified for quantization with high precision- low quantization results in a plateauing effect evidenced in Fig.5.13. Since the coefficient that multiplies each tree score in the BDT also use the same precision, the coefficient of trees with low weight can become 0 at low precision, causing the tree to be dropped from synthesis. This is an optimization of the software to avoid using resources for parts of the model that are not used. This is the source of this plateauing effect which stops larger models from using more resources and can affect the computation of R.



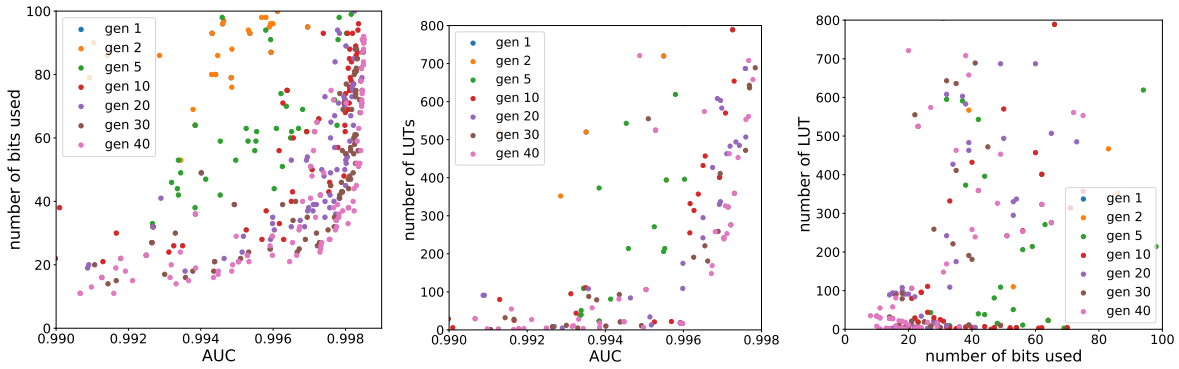
(a) 12 bits precision

(b) 5 bits precision

Figure 5.13: The number of LUTs used for the BDT implementation at a precision of (a) 12 bits and (b) 5 bits. The different colors correspond to different values of the tree maximum depth, on which the ratio has a slight dependence and are thus treated independently while estimating the resource usage. While for a high precision level the ratio is constant, at low precision the number of LUTs used stagnates due to the suppression of trees for which the multiplying coefficient vanishes after quantization. In such cases, only the linear part is considered to compute the proxy.

To limit the effect of tree score coefficient suppression at low precisions, only the linear part was considered while extracting R. For precisions of one and two bits, no linear behavior can be extracted, so the same ratio as for three bits precision was used as a conservative estimate. Under a precision of five bits the prediction is unreliable and the resulting RMS reach between 50% and 100%. At higher precision, the error is $\mathcal{O}(1 - 10\%)$.

The MOO is then performed to simultaneously optimize the model performance (measured as previously with the ROC AUC above 80% signal efficiency), the total number of bits and the model size. This size is computed by summing for each split in the BDT at a given precision and maximum depth the corresponding value of R. To account for the additional dimension, the population per generation is increased to 100. The remaining NSGA-II parameters are kept identical to the one used in Sec.5.4.2. The optimization converges well across all dimensions,



(a) **(b)** **(c)**

Figure 5.14: Performance of the algorithm (measured by the AUC of the ROC curve above a 80% signal efficiency threshold) against (a) the total number of bits used for the inputs and (b) the estimation of the number of LUTs used for the implementation of a BDT discriminating electrons and PU clusters. The number of bits versus the number of LUTs used for the model is shown in (c). The MOO converges well across all dimensions as each new generation gets closer to minimizing the number of bits and of LUTs while maximizing the AUC.

as illustrated in Fig.5.14.

As in the two-dimensional case, a selection of the optimal solutions that perform well for all objectives at the same time is applied and shown in Fig. 5.15 for the discrimination against PU. While another type of constraint was added with the minimization of the model size, the optimal solutions reach similar AUCs as before. The largest models, in terms of estimated number of LUTs, are not necessarily the ones that maximize the performance, which still seems to be correlated with the total number of bits. The optimal models seem to use between 20 and 40 bits (similar to the two-dimensional case) and all use less than 10000 LUTs, which corresponds to around 0.5% of the LUT available on VU13P FPGAs. The average size is around 2500 LUTs or 0.15% of the VU13P resources. This modest usage of LUTs, even when considering the high uncertainty on the low precision nodes resource usage, allows the implementation several instances of the model to process in parallel the shape variables from multiple clusters.

More variables are kept in the 3D optimization case, with from 11 up to the full set of 16 variables. This means that the reduction of the number of variables shown in the 2D case must have required the utilization of complex BDTs requiring many resources for their implementation. In particular, the sum of energy in the first layers of the CE-H and the index of the first layer in the cluster, which were removed from the input set in most 2D optimization solutions, are conserved in most 3D solutions. The distribution of the number of bits between the variables is similar, with the CE-E bitmap still requiring the highest number of bits. The variance in the r coordinate has still the second highest bit utilization but it is reduced compared to the 2D optimization. Concerning the optimization of the BDT hyperparameters, the best performance are obtained either by BDTs with high number of shallow trees or with a lower number of deeper trees. For the learning rate a value of 0.1 is favoured for all optimal solutions.

The discrimination of clusters reconstructed from charged pions favors smaller BDTS and

lower precisions overall, as illustrated in Fig.5.16. In particular, shallow trees with a maximum depth of 2 seem to be favored. The number of LUTs required for the BDT implementation is thus drastically reduced, for an average of around 500 LUTs representing 0.02% of the FPGA resources. The few variables that are used in the pions input set but not in the PU input set, such as the CE-H bitmap, do not require a high amount of bits. As such, conserving them to broaden the discriminative capabilities of the BDTs does not drastically impacts the total number of bits required.

$Var_{\eta\eta}$	CE-H 1	Show lgth	Core lgth	CE-H 5	meanz	Var_{zz}	$E_{max}2$	$E_{max}5$	1^{st} layer	(E/E+H)	$E_{max}4$	|eta|	$Var_{\phi\phi}$	Var_{rr}	ebm1	max boost rounds	max depth	lr	N_{inputs}	N_{bits}	N_{LUTs}	Perf
0	1	0	0	1	2	0	1	4	0	3	1	4	5	5	10	292	4	0.1	11	37	4933	99.82
2	1	1	1	1	1	1	2	1	2	1	4	4	3	3	10	235	5	0.1	16	38	4497	99.81
0	1	1	0	1	1	1	0	1	2	1	4	4	3	3	10	235	4	0.1	13	33	2562	99.81
2	1	1	0	1	1	1	2	1	2	1	4	4	3	3	10	235	2	0.1	15	37	821	99.79
0	0	0	2	1	0	1	4	0	3	4	3	3	2	3	10	372	2	0.1	11	36	1403	99.78
0	0	0	1	1	0	1	4	0	3	4	3	3	2	3	10	420	2	0.1	11	35	1679	99.78
2	1	0	1	1	0	4	2	0	1	4	3	2	4	3	11	116	3	0.1	13	39	658	99.78
0	1	2	0	1	0	0	0	4	2	1	1	1	5	3	11	280	5	0.1	11	32	9225	99.78
0	0	0	1	1	0	1	4	0	3	4	3	3	1	3	10	382	2	0.1	11	34	1628	99.77
0	1	0	1	1	0	4	1	0	0	1	3	2	4	3	10	166	3	0.1	11	31	952	99.77
0	1	1	6	0	1	1	2	0	2	4	0	4	3	3	10	235	2	0.1	12	38	708	99.77
0	1	2	0	1	0	1	0	4	1	1	1	1	2	3	10	280	4	0.1	12	28	3569	99.76
0	1	1	0	1	1	1	0	1	2	1	4	0	3	3	9	151	4	0.1	12	28	1637	99.76
0	0	2	1	1	2	1	0	2	0	0	1	3	1	3	10	254	2	0.1	11	27	1252	99.73
0	1	1	1	1	2	1	0	2	0	1	1	1	1	3	10	142	5	0.1	13	26	3108	99.73
0	0	2	1	1	2	1	0	2	0	0	1	1	1	3	10	142	4	0.1	11	25	1874	99.73
0	1	1	0	1	1	1	0	0	0	1	2	1	2	3	10	375	4	0.1	11	24	5326	99.73
0	1	2	1	1	5	0	1	1	3	4	0	1	2	3	10	123	2	0.1	13	35	463	99.72

Figure 5.15: The list of solutions that optimize all objectives for the discrimination of electrons versus PU clusters, ordered by their performance. In gradients of blue are the number of bits used for each variable, ordered by average precision. In gradients of red are the three hyperparameters controlling the size of the model and in gradients of green are the objective functions and number of feature remaining in the input set.

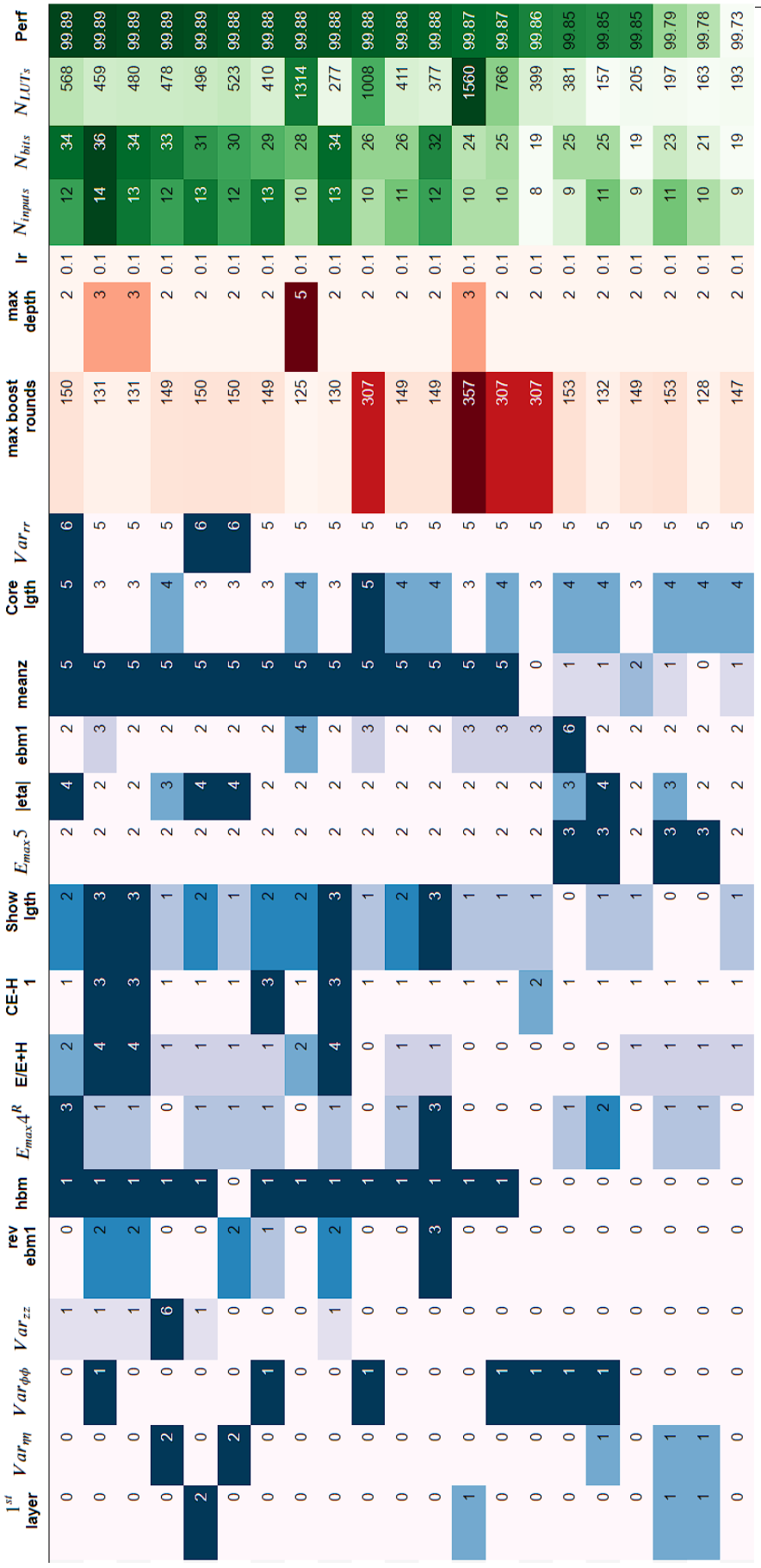
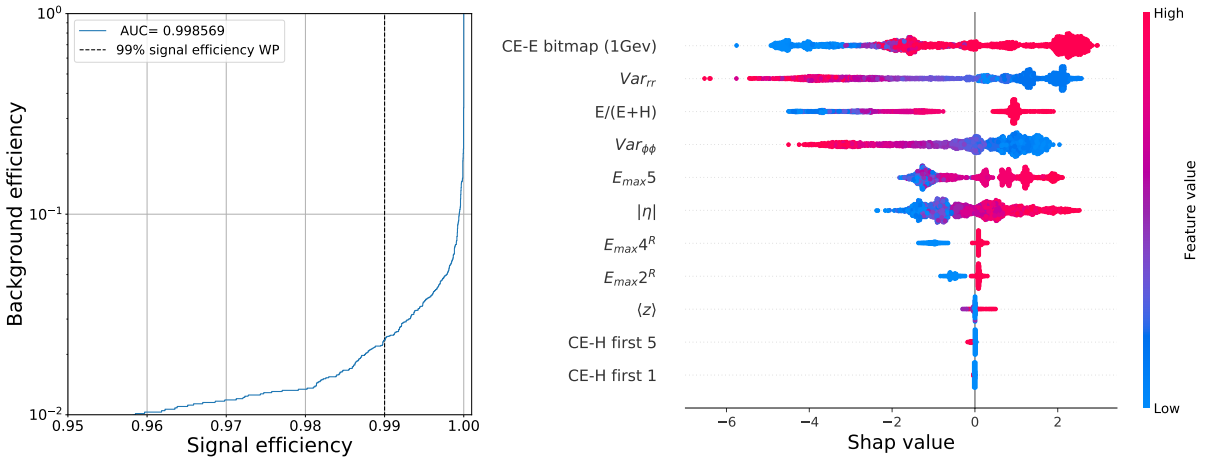


Figure 5.16: The list of solutions that optimize all objectives for the discrimination of electrons versus pions clusters, ordered by their performance. In gradients of blue are the number of bits used for each variable, ordered by average precision. In gradients of red are the three hyperparameters controlling the size of the model and in gradients of green are the objective functions and number of feature remaining in the input set.



(a) **(b)**

Figure 5.17: (a) The ROC curve for one of the optimal solutions in the three-dimensional PU MOO and (b) the importance of the variables in terms of SHAP values. The area under the ROC curve is indicated for the full curve and not only above a 80% signal efficiency threshold. The background rate corresponding to the 99% signal efficiency working point (WP) is of 2.3%.

To estimate the typical performance obtained by those solutions, the first set of parameters in Fig. 5.15, giving the highest AUC above 80% signal efficiency was scrutinized. It uses 37 bits over 11 variables, for an estimated number of LUTs of 4933 ± 480 when accounting for the RMS uncertainty and corresponding $0.29\% \pm 0.03\%$ occupation of the FPGA LUTs. The ROC curve and each feature importance in terms of SHAP values are shown in Fig. 5.17. The corresponding background rate at a 99% signal efficiency is of 2.3%.

5.5 Conclusions

The CMS L1 trigger needs to identify electromagnetic 3D clusters reconstructed in the future High Granularity Calorimeter, while rejecting backgrounds sources such as clusters induced by pileup or hadronic showers. This identification is foreseen to be performed by Boosted Decision Trees implemented on Field Programmable Gate Arrays, though different type of Machine Learning models, such as neural networks are considered. This identification will use inputs from the HGCAL trigger primitive generation in the form of variables describing the shape of the cluster. The computation of those variables will use limited resources, and must in particular be encoded on less than 128 bits per cluster. Similarly, the BDT model is constrained in size (and thus complexity) as it must not occupy too large a portion of the FPGA.

A Multi-Objective optimization was performed to train BDTs with optimal performance while minimizing the number of bits used for encoding the shape variables and the number of LUTs used for implementing the model. The constraints were satisfied, with model using only up to 40 bits, thus leaving space for additional variables that could be required for other tasks, and around 0.15% of the FPGA LUTs, so multiple instances of the model could be implemented to run in parallel. The estimation of the resource usage however is limited by the method employed to measure it, and could be refined with improved frameworks, in particular with BDT synthesis allowing different precision levels for each variable, the nodes thresholds and tree scores.

These studies prove however the feasibility of using the HGCAL cluster shape variables at the L1 trigger for cluster identification with low background rate and accounting for the strong limitations arising from the detector and trigger system architecture. Complementary studies for identification of different objects that need to be reconstructed at the L1 trigger, such as τ leptons, must be pursued to settle on the set of shape variables that must be computed to describe the 3D clusters.

Chapter 6

General conclusion

The search for the semi-leptonic VBS ZV production was presented in this thesis with an expected statistical significance of 1.8 σ. It uses the 137 fb^{-1} of data recorded during the Run 2 of the LHC by the CMS experiment at a center of mass energy of $\sqrt{s} = 13TeV$. This process is of particular interest because of the access it provides to the quartic gauge boson coupling and the electroweak symmetry breaking.

In order to search for this very rare electroweak process, a signal region was designed around the characteristic signature of VBS which produces two jets with a large invariant mass and pseudorapidity separation, and divided depending on the topology of the jets originating from the V boson decay. The signal was further isolated through with deep neural networks combining the most discriminating kinematical properties of the VBS events. To that end, the signal as well as the most relevant backgrounds were simulated with Monte Carlo generators and a full detector simulation. Due to the very large background contribution to the final state compared to that of the VBS signal, discrepancies in the simulations compared to the real data could have a large effect. Several corrections were thus employed to ensure adequate simulation, and a data-driven approach was further applied to adjust the most important backgrounds contributions, namely the Z+jets process and single top/$t\bar{t}$ productions. One of the important variable for signal discrimination

Even with this sophisticated approach, the signal strength extraction through a maximum likelihood fit yields only less than 2 σ expected significance mostly due to the limited statistics. In the close future, the analysis framework established for this analysis will be used in an EFT interpretation to extract stringent limits on several dimension-8 parameters. The combination with the semi-leptonic VBS WV channel will improve the sensitivity further.

Though the data recorded during the current Run 3 of the LHC will increase the dataset, the sensitivity to rare processes such as VBS will truly improve only with the upgrade of the LHC towards a high luminosity phase. The different LHC experiments have designed plans in order to fully exploit the new possibilities of the collider and deal with its new challenges. The improved level 1 trigger of CMS will employ the information from the new endcap highly granular calorimeter HGCAL in real time to identify physics objects. The optimization of the HGCAL trigger primitive generation must account for the limited size of the data that can

be communicated to the trigger and for the limited resources available for the identification algorithms implemented on the trigger FPGA boards. This work showed that the choice of the shower shape variables employed for the identification of electromagnetic showers and the number of bits used to encode them can be optimized with a multi-objective optimization technique based on genetic algorithms to meet these requirements.

The future of rare physics such as VBS is thus very bright, with several new results coming in the future. Sensitivity to semi-leptonic and even access to the hadronic final states will improve, probing critical sectors of the SM. Strong constraints on EFT operators will thus be determined and give more insight for possible new physics, pushing the boundaries of the known subatomic world.

Bibliography

[1] A.L. de Lavoisier and G.J. Cuchet. *Traité élémentaire de chimie: présenté dans un ordre nouveau et d'après les découvertes modernes ...* Traité élémentaire de chimie: présenté dans un ordre nouveau et d'après les découvertes modernes. Chez Cuchet, 1789. URL: https://books.google.fr/books?id=d2n7gY5SI2YC.

[2] E.R. Scerri. *The Periodic Table: Its Story and Its Significance*. Oxford University Press, Incorporated, 2019. ISBN: 9780190914363. URL: https://books.google.fr/books?id=IO4ExAEACAAJ.

[3] J. J. Thomson M.A. F.R.S. "XL. Cathode Rays". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 44.269 (1897), pp. 293–316. DOI: 10.1080/14786449708621070. eprint: https://doi.org/10.1080/14786449708621070. URL: https://doi.org/10.1080/14786449708621070.

[4] H. Kragh. *Quantum Generations: A History of Physics in the Twentieth Century*. Book collections on Project MUSE. Princeton University Press, 2002. ISBN: 9780691095523. URL: https://books.google.fr/books?id=ELrFDIldlawC.

[5] CMS collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Physics Letters B* 716.1 (2012), pp. 30–61. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2012.08.021. URL: https://www.sciencedirect.com/science/article/pii/S0370269312008581.

[6] ATLAS collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Physics Letters B* 716.1 (2012), pp. 1–29. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2012.08.020. URL: https://www.sciencedirect.com/science/article/pii/S037026931200857X.

[7] Particle Data Group and Workman. "Review of Particle Physics". In: *Progress of Theoretical and Experimental Physics* 2022.8 (Aug. 2022). 083C01. ISSN: 2050-3911. DOI: 10.1093/ptep/ptac097. eprint: https://academic.oup.com/ptep/article-pdf/2022/8/083C01/45434166/ptac097.pdf. URL: https://doi.org/10.1093/ptep/ptac097.

[8] *File:Standard Model of Elementary Particles.svg*. https://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg. Accessed: 2022-09-30.

[9] Takaaki Kajita. "Nobel Lecture: Discovery of atmospheric neutrino oscillations". In: *Rev. Mod. Phys.* 88 (3 2016), p. 030501. DOI: 10.1103/RevModPhys.88.030501. URL: https://link.aps.org/doi/10.1103/RevModPhys.88.030501.

[10] CMS collaboration ATLAS collaboration. "Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments". In: *Phys. Rev. Lett.* 114 (19 2015), p. 191803. DOI: 10.1103/PhysRevLett.114.191803. URL: https://link.aps.org/doi/10.1103/PhysRevLett.114.191803.

[11] S L Glashow. "PARTIAL-SYMMETRIES OF WEAK INTERACTIONS". In: *Nuclear Phys.* (). DOI: 10.1016/0029-5582(61)90469-2. URL: https://www.osti.gov/biblio/4082455.

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

[13] Steven Weinberg. "A Model of Leptons". In: *Phys. Rev. Lett.* 19 (1967), pp. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264.

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

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

[16] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. "Global Conservation Laws and Massless Particles". In: *Phys. Rev. Lett.* 13 (20 1964), pp. 585–587. DOI: 10.1103/PhysRevLett.13.585. URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.585.

[17] Csaba Csáki, Salvator Lombardo, and Ofri Telem. *TASI Lectures on Non-Supersymmetric BSM Models*. 2018. DOI: 10.48550/ARXIV.1811.04279. URL: https://arxiv.org/abs/1811.04279.

[18] Fermi GMT/LAT collaborations. "A limit on the variation of the speed of light arising from quantum gravity effects". In: *Nature* 462.7271 (2009), pp. 331–334. DOI: 10.1038/nature08574. URL: https://doi.org/10.1038%2Fnature08574.

[19] Steven Weinberg. "Phenomenological Lagrangians". In: *Physica A* 96.1-2 (1979). Ed. by S. Deser, pp. 327–340. DOI: 10.1016/0378-4371(79)90223-1.

[20] Céline Degrande et al. "Effective field theory: A modern approach to anomalous couplings". In: *Annals of Physics* 335 (2013), pp. 21–32. DOI: 10.1016/j.aop.2013.04.016. URL: https://doi.org/10.1016%2Fj.aop.2013.04.016.

[21] Christoph Schiller. "A Conjecture on Deducing General Relativity and the Standard Model with Its Fundamental Constants from Rational Tangles of Strands". In: *Physics of Particles and Nuclei* 50 (June 2019), pp. 259–299. DOI: 10.1134/S1063779619030055.

[22] Benjamin W. Lee, C. Quigg, and H. B. Thacker. "Strength of Weak Interactions at Very High Energies and the Higgs Boson Mass". In: *Phys. Rev. Lett.* 38 (16 1977), pp. 883–885. DOI: 10.1103/PhysRevLett.38.883. URL: https://link.aps.org/doi/10.1103/PhysRevLett.38.883.

[23] Alessandro Ballestrero et al. "Precise predictions for same-sign W-boson scattering at the LHC". In: *The European Physical Journal C* 78.8 (2018). DOI: 10.1140/epjc/s10052-018-6136-y. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-018-6136-y.

[24] David L. Rainwater, R. Szalapski, and D. Zeppenfeld. "Probing color singlet exchange in Z + two jet events at the CERN LHC". In: *Phys. Rev. D* 54 (1996), pp. 6680–6689. DOI: 10.1103/PhysRevD.54.6680. arXiv: hep-ph/9605444.

[25] Roberto Covarelli, Mathieu Pellen, and Marco Zaro. "Vector-Boson scattering at the LHC: Unraveling the electroweak sector". In: *International Journal of Modern Physics A* 36.16 (2021), p. 2130009. DOI: 10.1142/s0217751x2130009x. URL: https://doi.org/10.1142%2Fs0217751x2130009x.

[26] ATLAS collaboration. "Observation of Electroweak Production of a Same-Sign W Boson Pair in Association with Two Jets in pp Collisions at $\sqrt{s} = 13$ TeV with the ATLAS Detector". In: 123.16 (2019). DOI: 10.1103/physrevlett.123.161801. URL: https://doi.org/10.1103%2Fphysrevlett.123.161801.

[27] CMS collaboration. "Measurements of production cross sections of same-sign WW and WZ boson pairs in association with two jets in proton-proton collisions at $\sqrt{s}$ = 13 TeV". In: *Physics Letters B* 809 (2020), p. 135710. DOI: 10.1016/j.physletb.2020.135710. URL: https://doi.org/10.1016%2Fj.physletb.2020.135710.

[28] CMS collaboration. "Measurements of production cross sections of polarized same-sign W boson pairs in association with two jets in proton-proton collisions at $\sqrt{s} =$ 13 TeV". In: (Sept. 2020).

[29] ATLAS collaboration. "Observation of electroweak $W^{\pm}Z$ boson pair production in association with two jets in pp collisions at $\sqrt{s}$ = 13TeV with the ATLAS Detector". In: *Physics Letters B* 793 (2019), pp. 469–492. DOI: 10.1016/j.physletb.2019.05.012. URL: https://doi.org/10.1016%2Fj.physletb.2019.05.012.

[30] ATLAS Collaboration. "Observation of electroweak production of two jets and a Z-boson pair". In: (2020). DOI: 10.48550/ARXIV.2004.10612. URL: https://arxiv.org/abs/2004.10612.

[31] CMS collaboration. "Evidence for electroweak production of four charged leptons and two jets in proton-proton collisions at $\sqrt{s}$ = 13 TeV". In: *Phys.Lett.B* 812 (2021), p. 135992. DOI: 10.1016/j.physletb.2020.135992. URL: https://hal.archives-ouvertes.fr/hal-02934053.

[32] CMS collaboration. "Observation of electroweak W^+W^- pair production in association with two jets in proton-proton collisions at $\sqrt{s}$ = 13 TeV". In: (May 2022). arXiv: 2205.05711 [hep-ex].

[33] ATLAS collaboration. "Search for electroweak diboson production in association with a high-mass dijet system in semileptonic final states in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Physical Review D* 100.3 (2019). DOI: 10.1103/physrevd.100.032007. URL: https://doi.org/10.1103%2Fphysrevd.100.032007.

[34] CMS collaboration. "Search for anomalous electroweak production of vector boson pairs in association with two jets in proton-proton collisions at 13 TeV". In: *Physics Letters B* 798 (2019), p. 134985. DOI: 10.1016/j.physletb.2019.134985. URL: https://doi.org/10.1016%2Fj.physletb.2019.134985.

[35] CMS Collaboration. *Evidence for WW/WZ vector boson scattering in the decay channel ℓqq produced in association with two jets in proton-proton collisions at $\sqrt{s}$ = 13 TeV*. 2021. DOI: 10.48550/ARXIV.2112.05259. URL: https://arxiv.org/abs/2112.05259.

[36] ATLAS collaboration. "Evidence for electroweak production of two jets in association with a $Z\gamma$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Physics Letters B* 803 (2020), p. 135341. DOI: 10.1016/j.physletb.2020.135341. URL: https://doi.org/10.1016%2Fj.physletb.2020.135341.

[37] CMS collaboration. "Measurement of the electroweak production of $Z\gamma$ and two jets in proton-proton collisions at $\sqrt{s} = 13$ TeV and constraints on anomalous quartic gauge couplings". In: *Physical Review D* 104.7 (2021). DOI: 10.1103/physrevd.104.072001. URL: https://doi.org/10.1103%2Fphysrevd.104.072001.

[38] CMS collaboration. "Observation of electroweak production of $W\gamma$ with two jets in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Physics Letters B* 811 (2020), p. 135988. DOI: 10.1016/j.physletb.2020.135988. URL: https://doi.org/10.1016%2Fj.physletb.2020.135988.

[39] John Douglas Cockcroft and E. T. S. Walton. "Disintegration of Lithium by Swift Protons". In: *Nature* 129 (1932), pp. 649–649.

[40] Ernest O. Lawrence and M. Stanley Livingston. "The Production of High Speed Light Ions Without the Use of High Voltages". In: *Phys. Rev.* 40 (1 1932), pp. 19–35. DOI: 10.1103/PhysRev.40.19. URL: https://link.aps.org/doi/10.1103/PhysRev.40.19.

[41] Thomas Sven Pettersson and P Lefèvre. *The Large Hadron Collider: conceptual design*. Tech. rep. 1995. URL: https://cds.cern.ch/record/291782.

[42] Ewa Lopienska. "The CERN accelerator complex, layout in 2022. Complexe des accélérateurs du CERN en janvier 2022". In: (2022). General Photo. URL: https://cds.cern.ch/record/2800984.

[43] CMS collaboration. “Measurement of the inelastic proton-proton cross section at $$ \sqrt{s}=13 $$ TeV”. In: *Journal of High Energy Physics* 2018.7 (2018). DOI: 10.1007/jhep07(2018)161. URL: https://doi.org/10.1007%2Fjhep07%282018%29161.

[44] CMS collaboration. *CMS public luminosity results*. Accessed: 2022-09-14. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

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

[46] G. Acquistapace et al. “CMS, the magnet project: Technical design report”. In: (May 1997).

[47] Sergey Chatrchyan et al. “Precise Mapping of the Magnetic Field in the CMS Barrel Yoke using Cosmic Rays”. In: *Journal of Instrumentation* 5 (Mar. 2010), T03021.

[48] CMS collaboration. *CMS Technical Design Report for the Pixel Detector Upgrade*. Tech. rep. Additional contacts: Jeffrey Spalding, Fermilab, Jeffrey.Spalding@cern.ch Didier Contardo, Universite Claude Bernard-Lyon I, didier.claude.contardo@cern.ch. 2012. URL: https://cds.cern.ch/record/1481838.

[49] CMS collaboration. *The Phase-1 Upgrade of the CMS Pixel Detector*. Tech. rep. Geneva: CERN, 2017. DOI: 10.1088/1748-0221/12/07/C07009. URL: https://cds.cern.ch/record/2265423.

[50] CMS collaboration. “Test Beam Performance Measurements for the Phase I Upgrade of the CMS Pixel Detector”. In: *JINST* 12.05 (2017), P05022. DOI: 10.1088/1748-0221/12/05/P05022. arXiv: 1706.00222 [physics.ins-det].

[51] CMS collaboration. “Description and performance of track and primary-vertex reconstruction with the CMS tracker”. In: *JINST* 9 (2014). Comments: Replaced with published version. Added journal reference and DOI, P10009. 80 p. DOI: 10.1088/1748-0221/9/10/P10009. arXiv: 1405.6569. URL: https://cds.cern.ch/record/1704291.

[52] “CMS ECAL Response to Laser Light”. In: (2019). URL: http://cds.cern.ch/record/2668200.

[53] CMS collaboration. “Energy resolution of the barrel of the CMS electromagnetic calorimeter”. In: *JINST* 2 (2007), P04004. DOI: 10.1088/1748-0221/2/04/P04004.

[54] CMS collaboration. “Performance of the CMS hadron calorimeter with cosmic ray muons and LHC beam data”. In: *Journal of Instrumentation* 5 (Mar. 2010), T03012. DOI: 10.5167/uzh-45572.

[55] *The CMS hadron calorimeter project: Technical Design Report*. Technical design report. CMS. The following files are from http://uscms.fnal.gov/pub/hcal_tdr and may not be the version as printed, please check the printed version to be sure. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/357153.

[56] CMS collaboration. *The CMS Barrel Calorimeter Response to Particle Beams from 2 to 350 GeV/c*. Tech. rep. Geneva: CERN, 2008. URL: https://cds.cern.ch/record/1166316.

[57] CMS collaboration. *CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter*. Tech. rep. Additional contact persons: Jeffrey Spalding, Fermilab, spalding@cern.ch, Didier Contardo, Universite Claude Bernard-Lyon I, contardo@cern.ch. 2012. URL: https://cds.cern.ch/record/1481837.

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

[59] CMS collaboration. *CMS Technical Design Report for the Muon Endcap GEM Upgrade*. Tech. rep. 2015. URL: https://cds.cern.ch/record/2021453.

[60] CMS collaboration. *Summaries of CMS cross section measurements*. Accessed: 2022-09-18. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsCombined.

[61] *The Phase-2 Upgrade of the CMS Level-1 Trigger*. Tech. rep. Final version. Geneva: CERN, 2020. URL: https://cds.cern.ch/record/2714892.

[62] CMS collaboration. "Particle-flow reconstruction and global event description with the CMS detector. Particle-flow reconstruction and global event description with the CMS detector". In: *JINST* 12 (2017). Replaced with the published version. Added the journal reference and DOI. All the figures and tables can be found at http://cms-results.web.cern.ch/cms-results/public-results/publications/PRF-14-001 (CMS Public Pages), P10003. 82 p. DOI: 10.1088/1748-0221/12/10/P10003. arXiv: 1706.04965. URL: http://cds.cern.ch/record/2270046.

[63] Milos Dordevic. "The CMS Particle Flow Algorithm". In: *EPJ Web Conf.* 191 (2018), 02016. 7 p. DOI: 10.1051/epjconf/201819102016. URL: http://cds.cern.ch/record/2678077.

[64] *Application of Kalman filtering to track and vertex fitting*.

[65] The CMS collaboration. "Performance of CMS muon reconstruction in pp collision events atb$\sqrt{}$s = 7/bTeV". In: *Journal of Instrumentation* 7.10 (2012), P10002–P10002. DOI: 10.1088/1748-0221/7/10/p10002. URL: https://doi.org/10.1088%2F1748-0221%2F7%2F10%2Fp10002.

[66] CMS collaboration. *Tag and Probe method*. Accessed: 2022-09-14. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/TagAndProbe.

[67] CMS collaboration. "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *JINST* 13 (2018). Replaced with the published version. Added the journal reference and the DOI. All the figures and tables can be found at http://cms-results.web.cern.ch/cms-results/public-results/publications/MUO-16-001 (CMS Public Pages), P06015. 53 p. DOI: 10.1088/1748-0221/13/06/P06015. arXiv: 1804.04528. URL: https://cds.cern.ch/record/2313130.

[68] Wolfgang Adam et al. *Reconstruction of Electrons with the Gaussian-Sum Filter in the CMS Tracker at the LHC*. Tech. rep. Geneva: CERN, 2005. URL: https://cds.cern.ch/record/815410.

[69] CMS collaboration. "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s}$ = 8 TeV. Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at sqrt(s)=8 TeV". In: *JINST* 10 (2015). Replaced with published version. Added journal reference and DOI, P06005. 63 p. DOI: 10.1088/1748-0221/10/06/P06005. arXiv: 1502.02701. URL: https://cds.cern.ch/record/1988091.

[70] CMS collaboration. "Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC". In: *Journal of Instrumentation* 16.05 (2021), P05014. DOI: 10.1088/1748-0221/16/05/p05014. URL: https://doi.org/10.1088%2F1748-0221%2F16%2F05%2Fp05014.

[71] CMS collaboration. "Performance of photon reconstruction and identification with the CMS detector in proton-proton collisions at $\sqrt{s}$ =8 TeV". In: *JINST* 10 (2015). Comments: Submitted to JINST, P08010. 59 p. DOI: 10.1088/1748-0221/10/08/P08010. arXiv: 1502.02702. URL: https://cds.cern.ch/record/1988093.

[72] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The anti-k_t jet clustering algorithm". In: *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ph].

[73] *Pileup Removal Algorithms*. Tech. rep. Geneva: CERN, 2014. URL: https://cds.cern.ch/record/1751454.

[74] Daniele Bertolini et al. "Pileup per particle identification". In: *Journal of High Energy Physics* 2014.10 (2014). DOI: 10.1007/jhep10(2014)059. URL: https://doi.org/10.1007%2Fjhep10%282014%29059.

[75] "Performance of the pile up jet identification in CMS for Run 2". In: (2020). URL: https://cds.cern.ch/record/2715906.

[76] E. Bols et al. "Jet flavour classification using DeepJet". In: *Journal of Instrumentation* 15 (Dec. 2020), P12012–P12012. DOI: 10.1088/1748-0221/15/12/P12012.

[77] *Performance of quark/gluon discrimination in 8 TeV pp data*. Tech. rep. Geneva: CERN, 2013. URL: https://cds.cern.ch/record/1599732.

[78] Reyhane Hemmat and Abdelhakim Senhaji Hafid. "SLA Violation Prediction In Cloud Computing: A Machine Learning Perspective". In: (Nov. 2016).

[79] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.

[80] Kurt Hornik, Maxwell B. Stinchcombe, and Halbert L. White. "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2 (1989), pp. 359–366.

[81] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2014. DOI: 10.48550/ARXIV.1412.6980. URL: https://arxiv.org/abs/1412.6980.

[82] Niranjan Srinivas et al. "Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting". In: *IEEE Transactions on Information Theory* 58.5 (2012), pp. 3250–3265. DOI: 10.1109/tit.2011.2182033. URL: https://doi.org/10.1109%2Ftit.2011.2182033.

[83] Donald R. Jones, Matthias Schonlau, and William J. Welch. "Efficient Global Optimization of Expensive Black-Box Functions". In: *Journal of Global Optimization* 13 (1998), pp. 455–492.

[84] *Bayesian Optimization github*. Accessed: 2022-09-20. URL: https://github.com/fmfn/BayesianOptimization.

[85] CMS Collaboration. *Evidence for WW/WZ vector boson scattering in the decay channel ℓqq produced in association with two jets in proton-proton collisions at $\sqrt{s}$ = 13 TeV*. 2021. DOI: 10.48550/ARXIV.2112.05259. URL: https://arxiv.org/abs/2112.05259.

[86] Davide Di Croce. "Measurement of the Higgs boson decay to a W boson pair at 13 TeV with the CMS detector". 2020. URL: https://cds.cern.ch/record/2791110.

[87] Simone Marzani, Lais Schunk, and Gregory Soyez. "The jet mass distribution after Soft Drop". In: *Eur. Phys. J. C* 78.2 (2018), p. 96. DOI: 10.1140/epjc/s10052-018-5579-5. arXiv: 1712.05105 [hep-ph].

[88] Jesse Thaler and Ken Van Tilburg. "Identifying boosted objects with N-subjettiness". In: *Journal of High Energy Physics* 2011.3 (2011). DOI: 10.1007/jhep03(2011)015. URL: https://doi.org/10.1007%2Fjhep03%282011%29015.

[89] *Jet identification in semi-leptonic WV vector boson scattering at the LHC*. http://govoni.web.cern.ch/govoni/tesi/docs/Giacomo_Boldrini_Tesi.pdf.

[90] J. Alwall et al. “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations”. In: *Journal of High Energy Physics* 2014.7 (2014). DOI: 10.1007/jhep07(2014)079. URL: https://doi.org/10.1007%2Fjhep07%282014%29079.

[91] Pierre Artoisenet et al. “Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations”. In: *Journal of High Energy Physics* 2013.3 (2013). DOI: 10.1007/jhep03(2013)015. URL: https://doi.org/10.1007%2Fjhep03%282013%29015.

[92] Torbjörn Sjöstrand et al. “An introduction to PYTHIA 8.2”. In: *Computer Physics Communications* 191 (2015), pp. 159–177. DOI: 10.1016/j.cpc.2015.01.024. URL: https://doi.org/10.1016%2Fj.cpc.2015.01.024.

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

[94] Stefano Frixione, Paolo Nason, and Carlo Oleari. “Matching NLO QCD computations with Parton Shower simulations: the POWHEG method”. In: *JHEP* 11 (2007), p. 070. DOI: 10.1088/1126-6708/2007/11/070. arXiv: 0709.2092 [hep-ph].

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

[96] J. Alwall et al. “Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions”. In: *The European Physical Journal C* 53.3 (2007), 473–500. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-007-0490-5. URL: http://dx.doi.org/10.1140/epjc/s10052-007-0490-5.

[97] CMS collaboration. “Event generator tunes obtained from underlying event and multiparton scattering measurements”. In: *The European Physical Journal C* 76.3 (2016). DOI: 10.1140/epjc/s10052-016-3988-x.

[98] CMS collaboration. “Extraction and validation of a new set of CMS pythia8 tunes from underlying-event measurements”. In: *The European Physical Journal C* 80.1 (2020). DOI: 10.1140/epjc/s10052-019-7499-4. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-019-7499-4.

[99] Richard D. Ball et al. “Parton distributions for the LHC run II”. In: *Journal of High Energy Physics* 2015.4 (2015). DOI: 10.1007/jhep04(2015)040. URL: https://doi.org/10.1007%2Fjhep04%282015%29040.

[100] Richard D. Ball et al. “Parton distributions from high-precision collider data”. In: *The European Physical Journal C* 77.10 (2017). DOI: 10.1140/epjc/s10052-017-5199-5. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-017-5199-5.

[101] Barbara Jäger et al. "Parton-shower effects in Higgs production via vector-boson fusion". In: *The European Physical Journal C* 80.8 (2020). DOI: 10.1140/epjc/s10052-020-8326-7. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-020-8326-7.

[102] Manuel Bähr et al. "Herwig physics and manual". In: *The European Physical Journal C* 58.4 (2008), pp. 639–707. DOI: 10.1140/epjc/s10052-008-0798-9. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-008-0798-9.

[103] Johannes Bellm et al. "Herwig 7.0/Herwig 3.0 release note". In: *The European Physical Journal C* 76.4 (2016). DOI: 10.1140/epjc/s10052-016-4018-8. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-016-4018-8.

[104] S. Agostinelli et al. "GEANT4–a simulation toolkit". In: *Nucl. Instrum. Meth. A* 506 (2003), pp. 250–303. DOI: 10.1016/S0168-9002(03)01368-8.

[105] *Preliminary 2016/2017 L1 prefiring maps*. https://lathomas.web.cern.ch/lathomas/TSGStuff/L1Prefiring/PrefiringMaps_2016and2017/. Accessed: 2022-09-02.

[106] Songrit Maneewongvatana and David M. Mount. "Analysis of approximate nearest neighbor searching with clustered point sets". In: *CoRR* cs.CG/9901013 (1999). URL: https://arxiv.org/abs/cs/9901013.

[107] "Kolmogorov–Smirnov Test". In: *The Concise Encyclopedia of Statistics*. New York, NY: Springer New York, 2008, pp. 283–287. ISBN: 978-0-387-32833-1. DOI: 10.1007/978-0-387-32833-1_214. URL: https://doi.org/10.1007/978-0-387-32833-1_214.

[108] François Chollet et al. *Keras*. https://keras.io. 2015.

[109] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[110] Abien Fred Agarap. "Deep learning using rectified linear units (relu)". In: *arXiv preprint arXiv:1803.08375* (2018).

[111] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[112] Fernando Nogueira. *Bayesian Optimization: Open source constrained global optimization tool for Python*. 2014–. URL: https://github.com/fmfn/BayesianOptimization.

[113] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006, pp. I–XVIII, 1–248. ISBN: 026218253X.

[114] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[115] CMS Collaboration. *CMS Luminosity Measurements for the 2016 Data Taking Period*. CMS Physics Analysis Summary CMS-PAS-LUM-17-001. 2017. URL: https://cds.cern.ch/record/2257069.

[116] *CMS luminosity measurement for the 2017 data-taking period at* $\sqrt{s} = 13$ TeV. Tech. rep. CMS-PAS-LUM-17-004. Geneva: CERN, 2018. URL: https://cds.cern.ch/record/2621960.

[117] *CMS luminosity measurement for the 2018 data-taking period at* $\sqrt{s} = 13$ TeV. Tech. rep. CMS-PAS-LUM-18-002. Geneva: CERN, 2019. URL: https://cds.cern.ch/record/2676164.

[118] CMS collaboration. *Combine tool github*. https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/combine-tool. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/TagAndProbe.

[119] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". In: *The European Physical Journal C* 71.2 (2011). DOI: 10.1140/epjc/s10052-011-1554-0. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-011-1554-0.

[120] S. S. Wilks. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses". In: *Annals Math. Statist.* 9.1 (1938), pp. 60–62. DOI: 10.1214/aoms/1177732360.

[121] G. Lindstrom, M. Moll, and E. Fretwurst. "Radiation hardness of silicon detectors: A challenge from high-energy physics". In: *Nucl. Instrum. Meth. A* 426 (1999). Ed. by E. Borchi and M. Bruzzi, pp. 1–15. DOI: 10.1016/S0168-9002(98)01462-4.

[122] *The Phase-2 Upgrade of the CMS Tracker*. Tech. rep. Geneva: CERN, 2017. DOI: 10.17181/CERN.QZ28.FLHW. URL: https://cds.cern.ch/record/2272264.

[123] *The Phase-2 Upgrade of the CMS Endcap Calorimeter*. Tech. rep. Geneva: CERN, 2017. DOI: 10.17181/CERN.IV8M.1JY2. URL: https://cds.cern.ch/record/2293646.

[124] *The Phase-2 Upgrade of the CMS Barrel Calorimeters*. Tech. rep. This is the final version, approved by the LHCC. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2283187.

[125] *The Phase-2 Upgrade of the CMS Muon Detectors*. Tech. rep. This is the final version, approved by the LHCC. Geneva: CERN, 2017. URL: https://cds.cern.ch/record/2283189.

[126] *Technical proposal for a MIP timing detector in the CMS experiment Phase 2 upgrade*. Tech. rep. Geneva: CERN, 2017. DOI: 10.17181/CERN.2RSJ.UE8W. URL: http://cds.cern.ch/record/2296612.

[127] A Ferrari et al. *FLUKA: A multi-particle transport code (program version 2005)*. CERN Yellow Reports: Monographs. Geneva: CERN, 2005. DOI: 10.5170/CERN-2005-010. URL: https://cds.cern.ch/record/898301.

[128] CERN. *"LpGBT specification document*. Accessed: 2022-09-20. URL: https://espace.cern.ch/GBTProject/LpGBT/Specifications/LpGbtxSpecifications.pdf.

[129] Karl-Johan Grahn. "A Layer Correlation Technique for Pion Energy Calibration at the 2004 ATLAS Combined Beam Test". In: (2009). 7 pages, 12 figures. Submitted to the Conference Record of the 2009 IEEE Nuclear Science Symposium (Orlando, Florida, USA), 751–757. 7 p. DOI: 10.1109/NSSMIC.2009.5402211. arXiv: 0911.2639. URL: https://cds.cern.ch/record/1222464.

[130] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". In: *Journal of Machine Learning Research* 18.17 (2017), pp. 1–5. URL: http://jmlr.org/papers/v18/16-365.html.

[131] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[132] Haibo He et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning". In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 2008, pp. 1322–1328. DOI: 10.1109/IJCNN.2008.4633969.

[133] Zonghan Wu et al. "A Comprehensive Survey on Graph Neural Networks". In: *CoRR* abs/1901.00596 (2019). arXiv: 1901.00596. URL: http://arxiv.org/abs/1901.00596.

[134] K. Deb et al. "A fast and elitist multiobjective genetic algorithm: NSGA-II". In: *IEEE Transactions on Evolutionary Computation* 6.2 (2002), pp. 182–197. DOI: 10.1109/4235.996017.

[135] J. Blank and K. Deb. "pymoo: Multi-Objective Optimization in Python". In: *IEEE Access* 8 (2020), pp. 89497–89509.

[136] Kalyanmoy Deb, Karthik Sindhya, and Tatsuya Okabe. "Self-Adaptive Simulated Binary Crossover for Real-Parameter Optimization". In: *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation*. GECCO '07. London, England: Association for Computing Machinery, 2007, 1187–1194. ISBN: 9781595936974. DOI: 10.1145/1276958.1277190. URL: https://doi.org/10.1145/1276958.1277190.

[137] Kalyanmoy Deb, Karthik Sindhya, and Tatsuya Okabe. “Self-adaptive simulated binary crossover for real-parameter optimization”. In: *GECCO ’07*. 2007.

[138] S. Summers et al. “Fast inference of Boosted Decision Trees in FPGAs for particle physics”. In: *Journal of Instrumentation* 15.05 (2020), P05026–P05026. DOI: 10.1088/1748-0221/15/05/p05026. URL: https://doi.org/10.1088%2F1748-0221%2F15%2F05%2Fp05026.

[139] Declan O’Loughlin et al. “Xilinx Vivado High Level Synthesis: Case studies”. In: *25th IET Irish Signals Systems Conference 2014 and 2014 China-Ireland International Conference on Information and Communications Technologies (ISSC 2014/CIICT 2014)*. 2014, pp. 352–356. DOI: 10.1049/cp.2014.0713.

ECOLE
DOCTORALE

Titre: Recherche de la diffusion de boson vecteur dans le canal semi-leptonique avec le détecteur CMS et études sur la classification de gerbes électromagnétiques avec HGCAL

Mots clés: Modèle standard, CMS, HGCAL, diffusion de boson vecteurs, apprentissage machine, optimisation

Résumé: Parmi les phénomènes recherchés au LHC, la diffusion de bosons vecteurs est particulièrement intéressante de par son lien avec la brisure de symétrie électrofaible. Elle offre aussi un accès intéressant au couplage quadratique entre bosons vecteurs. Cette thèse exploite les données récoltées par CMS entre 2016 et 2018 pour rechercher le canal où deux bosons Z et V (Z ou W) se désintègrent respectivement en leptons et jets . Ce signal, très rare, est isolé à l'aide d'un réseau de neurones et la puissance statistique est estimée à 1.9 . Afin d'augmenter la sensibilité aux phénomènes rares, il est prévu d'augmenter la luminosité du LHC. Pour faire face aux défis que cela entraîne, CMS installera un calorimètre hautement granulaire (HGCAL) et mettra à jour son système de déclenchement. Une optimisation des données fournies au système de déclenchement pour l'identification d'électrons est réalisée en tenant compte des contraintes matérielles.

Title: Search for Vector Boson Scattering in semileptonic decay at CMS and studies on HGCal trigger electromagnetic shower classification

Keywords: Standard model, CMS, HGCAL, VBS, machine learning, optimization

Abstract:
Among the phenomena investigated at the LHC, the vector boson scattering is of particular interest because of its close relation with the electroweak symmetry breaking. It also offers an interesting access to the quadratic coupling between vector bosons. This thesis exploits the data collected by CMS between 2016 and 2018 to search for the decay channel where two bosons Z and V (Z or W) decay into leptons and jets respectively. This very rare signal is isolated using neural networks and the expected significance is estimated at 1.9 . In order to improve the sensitivity to rare phenomena, it is planned to increase the luminosity of the LHC. To meet the challenges this entails, CMS will install a highly granular calorimeter (HGCAL) and upgrade its trigger system. An optimisation of the data supplied to the trigger system for electron identification is performed while accounting for the hardware constraints.

Institut Polytechnique de Paris

91120 Palaiseau, France