UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI FISICA E ASTRONOMIA

GALILEO GALILEI

SCUOLA DI DOTTORATO DI RICERCA IN FISICA

CICLO XXXIV

DOCTORAL THESIS

Search for Bs decays to tau lepton pairs with the CMS experiment at the CERN LHC

PhD Candidate: Hevjin Yarar *Supervisor: Tommaso Dorigo*

Abstract

The high energy physics experiments at the LHC are designed to address many fundamental questions in modern physics. Extracting the relevant information from the collected data that can answer these questions is a difficult challenge due to the complexity and the high dimensionality. The emergence of deep learning algorithms have advanced the state of the data analysis methods by enabling the extraction of higher-level features and consequently reducing the dimensionality, which is a crucial improvement considering the vast size of collision data that is necessary to observe rare physics processes of interest. Within the scope of this thesis seveal machine learning techniques have been implemented to study the rare $B_s \to \tau\tau$ decay into tau leptons with the two tau decay modes $\tau \to \nu_\tau \mu \nu_\mu$ and $\tau \to \pi\pi\pi\nu_\tau$ respectively. To this purpose, B-Parking data containing a large number of B_s mesons, acquired by CMS during the Run 2 of the LHC and simulated Monte Carlo samples that include the decay channel of interest, have been used. The reconstructed events are filtered for the specific decay signature by a graph neural network that classifies triplets of charged particles as candidates for the 3-prong tau decay $\tau \to \pi\pi\pi\nu_\tau$ for events that are triggered by a muon, which is the candidate muon for the $\tau \to \nu_\tau \mu \nu_\mu$ decay. Identifying this decay channel is complicated by the escape of at least three neutrinos; two of which are produced in the 3-prong decay and the third in the semi-hadronic decay. Neural networks and gradient boosted decision trees have been explored as methodologies to recover the lost information from the measured momenta of the visible particles. Two supervised learning methods have been implemented; regressions to the four-momentum of the semi-hadronic and 3-prong decaying tau with the goal of estimating the four-momentum of the originating B_s meson and a classification between the signal and background events. Furthermore, a semi-supervised learning algorithm has been designed to complement the supervised classifier.

Contents

Preface **1**

1 Introduction **1**

1.1 Standard Model of Particle Physics 1

1.1.1 The Elementary Particles 2

1.1.2 The Fundamental Forces 4

1.1.3 Particle Decays and Interaction of Particles with Matter 5

1.2 Quantum Field Theory 12

1.2.1 QED 14

1.2.2 QCD 15

1.2.3 The weak interaction 16

1.3 Beyond The Standard Model 19

2 The Experimental Apparatus **21**

2.1 Large Hadron Collider 21

2.1.1 Injection and Acceleration of Protons 24

2.1.2 LHC Ring 25

2.1.3 Detectors 27

2.2 Compact Muon Solenoid 28

2.2.1 Interaction Point 30

2.2.2 Tracker 31

2.2.3 Electromagnetic Calorimeter 32

2.2.4 Hadronic Calorimeter 35

2.2.5 Magnet 37

2.2.6 Muon System 37

2.3 Trigger and Data Acquisition 38

2.4 Event Reconstruction . . . 40
2.4.1 Reconstruction of Charged Particle Tracks and Vertices . . . 41
2.4.2 Calorimeter Clustering . . . 42
2.4.3 Reconstruction of physics objects . . . 44
2.5 Detector Simulation . . . 47

3 Machine Learning Techniques **49**
3.0.1 Terminology . . . 49
3.1 Supervised Learning . . . 51
3.1.1 Decision Trees . . . 54
3.2 Neural Networks . . . 57
3.3 Anomaly Detection in Copula Space . . . 59
3.3.1 Problem Statement . . . 59
3.3.2 The idea of `RanBox` . . . 59
3.4 Algorithm Description . . . 61
3.4.1 Starting considerations . . . 61
3.4.2 Data preprocessing . . . 61
3.4.3 Dimensionality reduction . . . 63
3.4.4 Choices of a test statistic for the unsupervised learning task . . . 64
3.4.5 Box seeding . . . 68
3.4.6 Maximization of the test statistic . . . 71
3.5 Performance studies with synthetic data . . . 72
3.5.1 Event generation . . . 72
3.5.2 Power tests of the unsupervised `RanBox` . . . 75
3.6 Experiments . . . 79
3.6.1 Exotic signals in LHC data . . . 80

4 Search for B_s meson decays **89**
4.1 Introduction . . . 89
4.2 Datasets and Event Preprocessing . . . 93
4.2.1 B-Parking Data . . . 93
4.2.2 Monte Carlo Samples . . . 95
4.2.3 Trigger Strategy . . . 96

4.3 Preliminary Selection . . . 99
4.3.1 DNN selection of candidate tau leptons . . . 101
4.4 Regression to B_s Mass . . . 105
4.4.1 Initial considerations . . . 106
4.4.2 Neural Network regressors . . . 107
4.4.3 Semi-hadronic tau regression results . . . 108
4.4.4 Full B_s reconstruction . . . 109
4.4.5 B-Parking data from LHC Run 2 . . . 112
4.5 Classification . . . 113
4.5.1 Input variables . . . 114
4.5.2 Hyperparameters . . . 114
4.5.3 BDT performance . . . 114
4.6 Semi-supervised Search with Ranbox . . . 116
4.6.1 Algorithm description . . . 117
4.6.2 Sample results on the HEPMASS dataset . . . 123
4.6.3 Bias studies . . . 125

5 Conclusions and Outlook **131**

Bibliography **132**

Preface: INSIGHTS Project

This thesis reports the projects undertaken during the Physics PhD program of University of Padova in 2018-2022. The main work has been on the study of B_S decay channel into tau lepton pairs making use of state-of-the-art machine learning techniques, which are described in detail. The research activities have been performed within the collaboration of INFN Padova and the INSIGHTS MCSA-ITN, which is a Marie Sklodowska-Curie Innovative Training Networks project, focused on applying latest advances in statistics and in particular, machine learning to particle physics and funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2017, under Grant Agreement n. 765710. All projects in this work have been realized with collaborators from the Compact Muon Solenoid (CMS) experiment at the European Organisation for Nuclear Research (CERN).

Chapter 1

Introduction

1.1 Standard Model of Particle Physics

Particle physics aims to enhance our understanding of the laws of nature by providing a mathematical description to the constituents of the Universe, the *elementary particles* and *the fundamental forces* that govern the interactions between them. Our current understanding is embodied in the Standard Model (SM), which is a self-consistent scientific theory able to describe most observed phenomena successfully. The mathematical framework for SM is provided by the quantum field theory (QFT), where the dynamics of the system is described by a Lagrangian and every particle is manifested as a dynamical field which permeates the space-time. As with most field theories, QFT bases its framework on the set of symmetries of the system and formulates the Lagrangian from the particle (field) content in the system that follows these symmetries.

SM can describe three of the four fundamental forces (electromagnetic, weak and strong interactions) and all known elementary particles. Developed progressively based on the successively acquired empirical data, discoveries of the top quark [1], the tau neutrino [2] and the Higgs Boson [3] [4] have further strengthened the confidence in the SM. However, the model leaves many observed phenomena unexplained, such as the lack unification of the three forces with gravity, the baryon asymmetry or the neutrino oscillations and their non-zero masses.

1.1.1 The Elementary Particles

The visible matter around us seems to be made up of just a few types of elementary particles. Starting at the eV energy scale, atoms are bound states of electrons that are orbiting around nuclei composed of protons and neutrons. The electrostatic attraction between the opposite charges that binds the electrons is a low energy manifestation of Quantum Electrodynamics (QED), the fundamental theory of electromagnetism. At the same time protons and neutrons are bound together by the strong nuclear force, which is a manifestation of Quantum Chromodynamics (QCD), the fundamental theory of strong interactions.

The third fundamental force described by the SM is the weak force, responsible for the nuclear β decays of radioactive isotopes and the nuclear fusion processes. Both the β decay and nuclear fusion processes produce electron neutrinos ν_e. Together with electrons, protons, neutrons, they make up almost all observed phenomena at low energy scale. At GeV energy scale, protons and neutrons are found to be composite particles made up of quarks: protons consist of two up quarks and a down quark, whereas neutrons consist of two down and an up-quark. The up and down quarks, the electron and the electron neutrino make up the *first generation* of the elementary particles as depicted in Figure 1.1. For each of the four particles there are two sets of particles with higher masses that constitute the second and the third generations. Muon has the mass of $m_\mu \approx 200\ m_e$ and tau has $m_\tau \approx 3500\ m_e$. Based on the experimental evidence the Universe seems to be made up out of these twelve spin-half particles in Figure 1.1. Neutrinos ν_e, ν_τ, ν_μ are quantum mechanical mixtures of three fundamental neutrino states with well-defined masses. The dynamics of the twelve elementary particles (fermions) are described by the Dirac equation in the relativistic quantum mechanics, which specifies the existence of antiparticles for each fermion, particles with same mass but opposite charge.

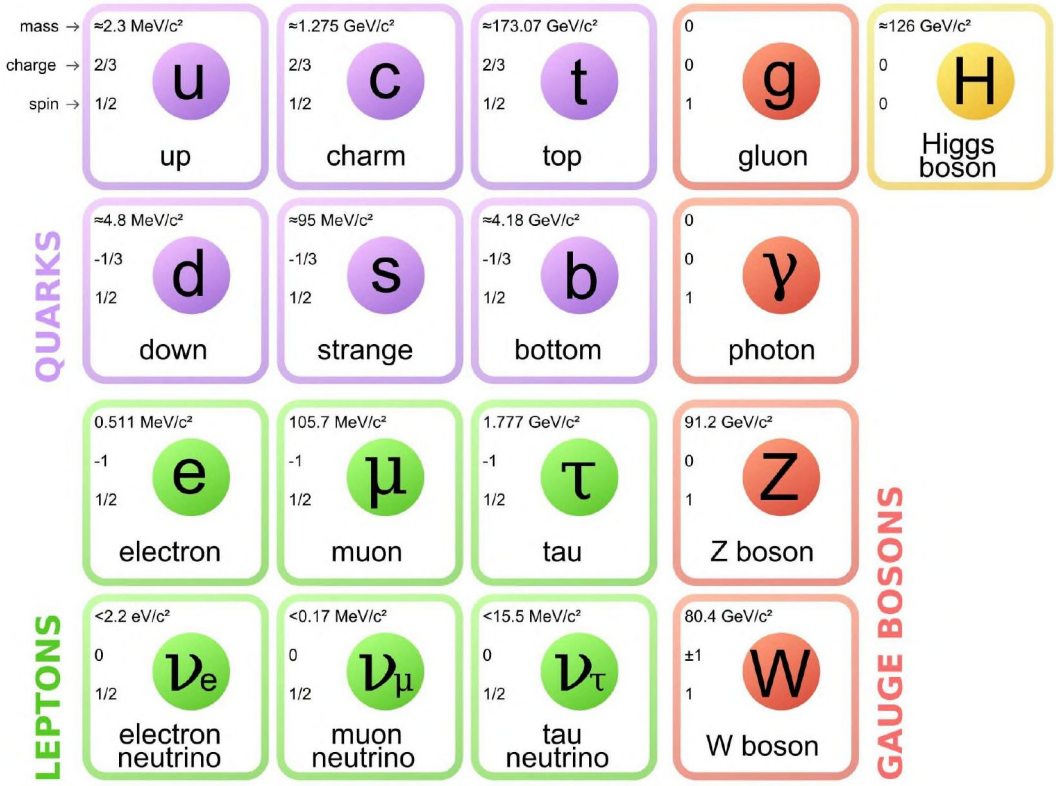


FIGURE 1.1: Standard model of elementary particles: the 12 fundamental fermions and 5 fundamental bosons [7].

1.1.2 The Fundamental Forces

The particles interact with each other through the four fundamental forces, in particle interactions gravity is neglected. All fermions experience weak interactions, while only charged fermions (neutrinos are electrically neutral) engage in the electromagnetic interaction of QED. The QCD equivalent of charge, referred as color charge, is carried only by quarks and thus they are the only ones partaking in the strong force interactions. Quarks are confined to bound states, called hadrons, such as in the proton or neutron due to nature of the QCD interaction. Quantum field theory describes each of the three interactions by the exchange of a spin-1 particle, referred as a gauge boson. In QED, the interactions between the particles are mediated by a virtual photon exchange (see Figure 1.1). In the case of the strong force, the force carrying particle is the massless gluon, whereas the mediator of the weak interaction are the charged W^+ and W^- bosons (for weak charged-current interaction) and neutral Z boson (for the weak neutral-current interaction). The final piece in SM elementary particles is the Higgs boson, discovered in 2012. It has a mass $m_H \approx 126\ GeV$ and unlike other gauge bosons, is a spin-0 scalar particle, the only elementary scalar particle known. In QFT, the Higgs boson is described as excitation of the Higgs field, which unlike the fields of fermions and other bosons, has a non-zero vacuum expectation value and through the interaction with this field, the initially massless particles gain their masses.

The coupling of the gauge bosons to fermions takes place at the so called interaction vertex, a three point vertex with the incoming and outgoing fermion and the gauge boson as depicted in Figure 1.2. The strength of the force exerted in the interaction is called the coupling constant g and is a measure of the probability that the spin-half fermion emits or absorbs the boson. Commonly, a dimensionless constant $\alpha \propto g^2$ is used instead. In the case of electromagnetism this is the fine-structure constant $\alpha = \frac{e^2}{4\pi_0 \hbar c} = 1/137$. The QCD interaction with $\alpha_S(M_Z) = 0.1175$ [6] is intrinsically stronger. The weak interaction with $\alpha_W \approx 1/30$ is intrinsically stronger than the QED, however at low energy scale (for example at particle decays), due to the large masses of the W bosons, the weak interaction is weaker than QED.

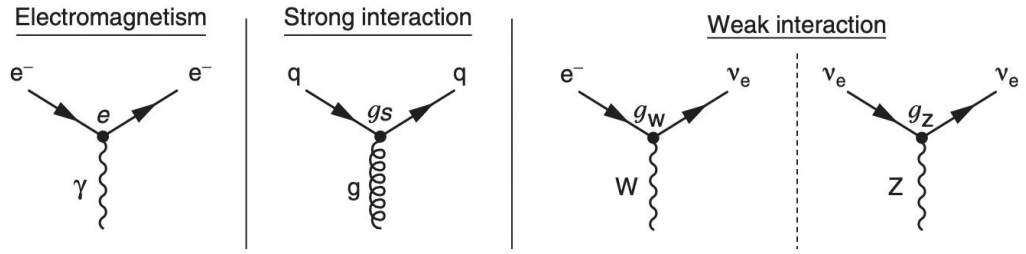


FIGURE 1.2: Standard Model interaction vertices [8].

1.1.3 Particle Decays and Interaction of Particles with Matter

Particle decays

Particle physics experiments can only detect long-lived particles since most particles decay with a short lifetime to a final state with a lower rest mass. The weak charged current is responsible for the particle decays with change in flavour such as the muon decay to electron and neutrinos. The only stable hadron is the proton, whereas free neutrons have a lifetime of 877.75 ± 0.28 s [5], after which they decay into a proton, electron and an electron neutrino via the weak interaction. Nuclei-bound neutrons act as stable particles due to the fact that the binding energy within nuclei is a lot larger than the neutron - proton mass difference. All other hadrons decay with very short lifetimes. The dominant decay mode for a specific particle is determined by the corresponding coupling strengths of the interactions. If a particle can decay via the strong interaction, this decay mode will almost always be the dominant one over the QED or the weak interactions. Additionally particles that only decay via the weak interaction are relatively long lived, such as taus and muons and thus can be detected at detectors.

Interaction of charged particles with matter

From all the elementary particles that are the building blocks of all matter, only electrons, photons, protons (and neutrons) are stable. All the rest decay at a distance of $\gamma v \tau$ (τ is the mean lifetime in the rest frame in this case) and $\gamma = 1/\sqrt{1 - v^2/c^2}$ is the Lorentz factor, where v is the particle velocity. This means, all particles at relativistic speeds with lifetimes longer than 10^{-10} s such as muons and charged pions, will reach distances of several meters before decaying. Particles with shorter

lifetimes decay earlier on, so that only the decay products can be detected at the experiments.

For a charged particle traversing at relativistic speed $v = \beta c$ through a medium with atomic number Z and number density n, the ionisation energy loss per unit length is given by the Bethe-Bloch equation:

$$-\frac{dE}{dx} = \frac{4\pi}{m_e c^2} \cdot \frac{n}{\beta^2} \cdot \left(\frac{e^2}{4\pi\varepsilon_0}\right)^2 \cdot \left[\ln\left(\frac{2m_e c^2\beta^2}{I \cdot (1-\beta^2)}\right) - \beta^2\right], \tag{1.1}$$

where I is the mean excitation potential and n is given as $n = \frac{N_A \cdot Z \cdot \rho}{A \cdot M_u}$ where ρ is the density of the material, A its relative atomic mass, N_A the Avogadro number and M_u the molar mass constant.

As can be deduced from the equation 1.1, the rate of energy loss depends on the material mainly via its density ρ, since approximately all nuclei have close numbers of protons and neutrons $Z/A \approx 1$ in the number density n, the ionization loss rate is proportional to the density of the material. Therefore highly dense materials are used in the particle experiments in order to stop the particles (via the absorption and the energy deposit, the energies of the particles can be estimated). Muons can travel large distances in dense materials like iron because for muons with energies below 100 GeV, ionisation is the dominant process of energy loss. For other charged particles several interaction processes are present. Thus detectors placed at a larger distance from the origin point of the particle, are likely to detect muons only.

Regardless of the dominant energy loss process, all charged particles leave a trail of freed electrons and the ionized atoms on their path through a medium, which allow the determination of the trajectories of the particles. The most commonly used detector technologies are gaseous detectors and semiconductor technology using silicon pixels and strips, which are described in detail in section 2 for the case of CMS. Commonly the tracking purposed detectors are placed in a large solenoid which produces a uniform magnetic field **B** in the direction perpendicular to the plane of particle interaction, in the z-axis. Due to the Lorentz force **L = v x B**, the charged particles move on a helix trajectory with a curvature radius R and a pitch angle λ. For a single charged particle q = $|e|$ the momentum of the particle is given as:

$$p = \frac{0.3BR}{cos\lambda}. \tag{1.2}$$

Charged particles emit Cherenkov radiation when they traverse through a dielectric medium and this feature is exploited for the purpose of detection at the particle physics experiments. As they traverse the medium with a refractive index n, they polarise the molecules which return to their unpolarized state once the particles have passed through, by emitting photons. If the particle velocity is greater than the speed of light in the medium $v > c/n$, Cherenkov radiation is emitted at an angle θ with respect to the particle track due to constructive interference of the emitted photons. The angle θ can be determined with:

$$cos\theta = \frac{1}{n\beta} \tag{1.3}$$

. **Interaction of electrons and photons with matter**

The dominating process for the energy loss of electrons at low energies below a critical energy E_c is ionisation. Above the E_c electrons lose their energy mainly via Brehmsstrahlung[1], where they emit photons as they get decelerated in the medium. The critical energy E_c is given as $E_c \approx \frac{800}{Z}$ MeV, where Z denotes the nucleus charge of the medium. Since the energy loss via Brehmsstrahlung is inversely proportional to the square of the particle's mass, it is dominant for the electrons rather than the muons at below 100 GeV scale. For low energy photons the dominating process is the photoelectric effect where the photon is absorbed by an electron that was ejected from an atom. At higher energies above 10 MeV, pair production plays the largest role in the energy loss of photons. An important parameter for the description of electromagnetic interactions of photons and electrons with matter is the *radiation length* X_0, which is the average distance the electron travels so that its energy is lowered down by a factor of 1/e. It can be approximated with:

$$X_0 \approx \frac{1}{4\alpha n Z^2 {r_e}^2 ln(287/Z^{1/2})} \tag{1.4}$$

[1]Production of electromagnetic radiation by the deceleration of a charged particle that gets deflected by another charged particle or by the atomic nucleus. The lost kinetic energy of the particle is converted into the radiation (photons) and the total energy is conserved.

where n is the number density in the material nuclei and r_e is the classical radius of electron, defined as $r_e = \frac{e^2}{4\pi_0 m_e c^2} = 2.8\times10^{-15}$ m. This indicates that for high Z materials the radiation length is short (for iron X_0 (Fe) = 1.76 cm). Alternating Brehmsstrahlung and pair production processes cause the production of a cascade of electrons, positrons and photons when the electron interacts with the material as it traverses it, referred as an electromagnetic (EM) shower. As the electromagnetic shower grows through the material, the number of generated particles approximately doubles at every X_0. Therefore the average energy of the particles at a distance of x radiation lengths can be estimated as $E_0/2^x$, where E_0 is the initial photon energy. The shower growth is stunted at the point where the cascading particles' energy drops to the critical energy E_c, below which they start losing the rest of their energy primarily via ionisation. The maximum number of radiation lengths an EM shower can grow, can therefore be determined:

$$x_{max} = \frac{ln(E_0/E_c)}{ln2}. \tag{1.5}$$

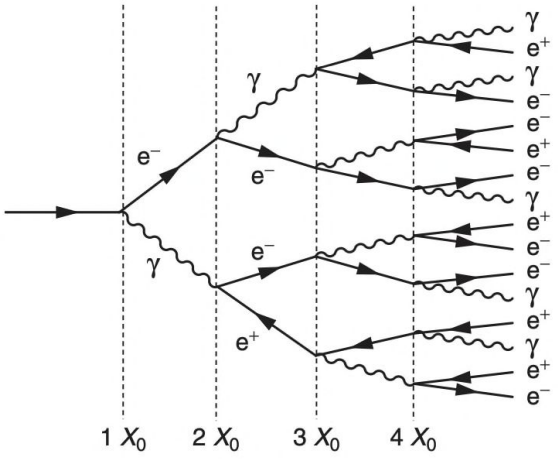


FIGURE 1.3: Electromagnetic shower growth. At each radiation length distance number of particles double (approximately) [9].

In high Z materials such as Pb, commonly used in particle detectors with $E_c \approx$ 10 MeV, an EM shower with an initial 100 GeV energy cascades until $x_{max} \approx 13\ X_0$ (less than 10 cm).

Interaction of hadrons with matter

When charged hadrons such as charged pions traverse a material, they lose their energy by ionisation as well as via the strong interaction with the nuclei, which produce other particles and consequently a cascade of particles, referred as a hadronic shower. Similar to EM shower, they are characterized by the mean distance between hadronic interactions within the shower, called *nuclear interaction length* λ_I. In comparison to the radiation length of EM shower, λ_I is much larger, for example for iron $\lambda_I \approx 17$ cm, whereas the X_0 is 1.8 cm. Hadronic showers grow in a less uniform manner unlike the EM showers, since many different final states can be generated including photons and electrons which give rise to further (EM) showers. On average 30% of the initial energy is lost due to nuclear excitations and cannot be detected.

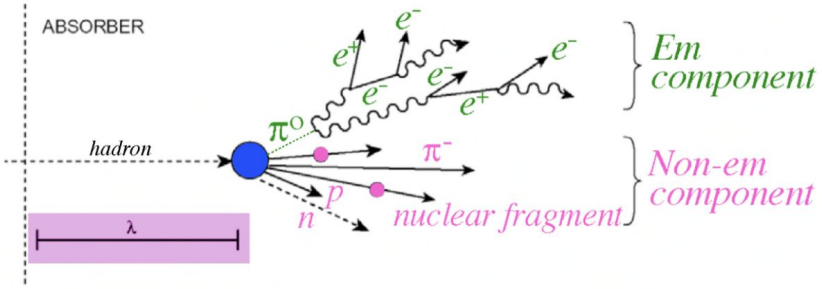


FIGURE 1.4: Hadronic shower development [10].

Neutrinos

Neutrinos do not leave any traces on the detector components in a collider environment, however evidence for their presence can be inferred from the *missing momentum*, defined as

$$\mathbf{p}_{miss} = -\sum_i \mathbf{p}_i. \tag{1.6}$$

where we sum over all measured momenta of the particles observed within an *event*, a recorded individual interaction. A non-zero total momentum would indicate the presence of undetected particles. Furthermore presence of neutrinos/missing momentum can be used as a key when identifying particles from their decay products. For example all main tau decay modes ($\tau^- \rightarrow \pi^-(n\pi^0)\nu_\tau(48\%)$, $\tau^- \rightarrow e^-\bar{\nu}_e\nu_\tau(17.8\%), \tau^- \rightarrow \mu^-\bar{\nu}_\mu\nu_\tau(17.4\%), \tau^- \rightarrow \pi^-\pi^+\pi^-(n\pi^0)\nu_\tau(15\%)$) have neutrinos produced in their final states, so the presence of missing momentum should accompany the other traces left in the detector components when one of the decay modes is observed.

In collider experiments the initial momentum of the colliding partons along the beam is unknown and the total missing energy cannot be measured. However, the initial momentum in the transverse direction is zero, therefore any net momentum in the transverse direction, so called missing ET (*MET*), indicates missing energy.

Particle accelerators

Particle accelerators collide beams of particles (or a single beam is fired at a stationary target), force them to interact and observe the collision outcome by many detector technologies with the goal to reconstruct the particles that were produced in the collision. The mechanism and operation principle of one such detector, CMS has been described in detail in section 2. One of the most important parameters that characterizes the performance of such an experiment is its *center-of-mass energy* $\sqrt{s}$ given in natural units with c = 1 for colliding beam accelerators with two initial-state particles:

$$s = (\sum_{i=1}^{2} E_i)^2 - (\sum_{i=1}^{2} \mathbf{p}_i)^2 \tag{1.7}$$

Since energy required to produce massive particles has to be provided by the center-of-mass energy available at the colliders, beam colliders have higher chance of producing massive particles such as the W, Z and Higgs bosons than the fixed target colliders.

One of the most important performance parameters for the detector alongside energy is called *luminosity* L, which is a measure the event rate. The total number of events N for a given process is the product of the process *cross section* σ, a measure of the quantum mechanical probability for the interaction, and the integrated luminosity over the period of time of the measurement:

$$N = \sigma \int L(t)dt. \tag{1.8}$$

In the case of LHC the beam particles are bunched together with a separation of 25 ns (collision frequency of 40 MHz). Assuming the beams have a Gaussian profile and collide head-on, the *instantaneous* luminosity at the LHC at a certain time at the interaction point is given by:

$$L = \frac{f \cdot N_1 \cdot N_2}{(4 \cdot \pi \cdot \sigma_x \sigma_y)}, \tag{1.9}$$

where N_i denote the number of protons per bunch per beam and σ_i the corresponding cross sections and f the crossing frequency (40 MHz).

Cross section measurements are performed by counting the number of observed events of the process of interest as well as of a process for which the cross section is already known:

$$\sigma = \sigma_{ref} \frac{N}{N_{ref}}. \tag{1.10}$$

1.2 Quantum Field Theory

As mentioned previously, SM is based on the QFT theoretical framework which unifies classical field theory with special relativity and quantum mechanics. Particles are treated as excited states (quanta) of the corresponding quantum fields. Particle interactions are described by the interaction terms in the Lagrangian for the corresponding quantum fields. Lagrangian can be defined in a similar manner as in classical dynamics, where the motion of a system is described with Newton's second law **F**= m$\ddot{\mathbf{x}}$ with the force F and the acceleration $\ddot{\mathbf{x}}$. Lagrangian L($q_i, \dot{q}_i$) for the generalized coordinates q_i and their (time) derivatives $\dot{q}_i$ is defined as the difference between the kinetic and potential energies of the system: L($q_i, \dot{q}_i$) = T - V. The equations of motion follows from the Euler-Lagrange equations:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = 0. \tag{1.11}$$

For a particle moving in one dimension, plugging in T and V into the equation 1.11 recovers the second law of motion. The treatment for a continuous system of particles is analogous, where the generalised coordinates q_i are replaced by the *fields* $\phi_i(t, x, y, z)$, time derivatives are replaced by the derivatives of the fields with respect to each of the four space-time coordinates $\partial_\mu \phi_i = \frac{\partial \phi_i}{\partial x^\mu}$ and the Lagrangian L($q_i, \dot{q}_i$) is replaced with the Lagrangian *density* $\mathcal{L}(\phi_i, \partial_\mu \phi_i)$. The Lagrangian L itself is given by:

$$L = \int \mathcal{L} d^3\mathbf{x}. \tag{1.12}$$

The equivalent Euler-Lagrange equations for a continuous system using the principle of least action follows:

$$\partial_\mu\left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi_i))}\right) - \frac{\partial \mathcal{L}}{\partial \phi_i} = 0, \tag{1.13}$$

where the field $\phi_i(x^\mu)$ is a continuous quantity with a value in each space-time point.

In QFT, spin-0, spin-half and spin-1 particles are described by their own Lagrangian densities. For spin-0 particles with scalar fields, the excitations satisfy the

Klein-Gordon equation, given as:

$$\mathcal{L}_S = \frac{1}{2}((\partial_\mu \phi_i)(\partial^\mu \phi_i)) - \frac{1}{2}m^2\phi^2, \tag{1.14}$$

for a free non-interacting scalar field. Plugging in the partial derivatives in the Euler-Lagrange equation recovers the Klein-Gordon equation for a free scalar field:

$$\partial_\mu \partial^\mu \phi + m^2 \phi = 0. \tag{1.15}$$

The Lagrangian density for spin-half particles (spinor field) $\psi(x)$ which satisfy the Dirac equation is given by:

$$\mathcal{L}_D = i\bar{\psi}\gamma^\mu \partial_\mu \psi - m\bar{\psi}\psi, \tag{1.16}$$

where $\psi(x)$ is a four dimensional complex spinor that can be expressed in terms of eight real fields $\psi_i(x) = \Psi_i(x) + i\Phi_i$ for i = 1,...,4, which in turn can be expressed as a linear combination of ψ and the adjoint spinor $\bar{\psi}$. Substituted in the Euler-Lagrange equation in 1.13 the Dirac equation for a spinor field follows:

$$i\gamma^\mu(\partial_\mu \psi) - m\psi = 0. \tag{1.17}$$

For an electromagnetic field $A^\mu = (\phi, A)$ expressed in the covariant form is given as:

$$\partial_\mu F^{\mu\nu} = j^\nu, \tag{1.18}$$

where $F^{\mu\nu}$ is the field-strength tensor:

$$F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu = \begin{bmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & -B_z & B_y \\ E_y & B_z & 0 & -B_x \\ E_z & -B_y & B_x & 0 \end{bmatrix}, \tag{1.19}$$

The Lagrangian for the photon field follows as :

$$\mathcal{L}_{EM} = -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} - j^{\mu}A_{\mu}, \tag{1.20}$$

where j = (ρ, **J**) is the four vector current with the charge density ρ and the current density **J**.

For spin-1 particles with mass this would be modified to:

$$\mathcal{L}_{EM} = -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} - j^{\mu}A_{\mu} + \frac{1}{2}m^2A^{\mu}A_{\mu}. \tag{1.21}$$

1.2.1 QED

In electromagnetism, the electric **E** (with the scalar potential ϕ) and the magnetic field **B** (with the vector potential **A**) do not change under the gauge transformation:

$$\phi \rightarrow \phi' = \phi - \frac{\partial\chi}{\partial t} \text{ and } \mathbf{A} \rightarrow \mathbf{A}' = \mathbf{A} + \boldsymbol{\nabla}\chi. \tag{1.22}$$

With the following operations A_{μ} = (ϕ, -**A**) and ∂_{μ} = (∂_0, $\boldsymbol{\nabla}$) the equation 1.22 can be written more compactly: $A_{\mu} \rightarrow A'_{\mu} = A_{\mu} - \partial_{\mu}\chi$.

Similarly, in relativistic quantum mechanics, the gauge invariance can be related to a local gauge invariance, where a fundamental symmetry requires the physics laws to stay invariant under a local phase transformation defined by:

$$\psi(x) \rightarrow \psi'(x) = U(x)\psi(x) = e^{iq\chi(x)}\psi(x), \tag{1.23}$$

where the phase $q\chi(x)$ can vary at all points in the space-time. If this is substituted in the free-particle Dirac equation of 1.17, we get:

$$i\gamma^{\mu}(\partial_{\mu} + iq\partial_{\mu}\chi)\psi - m\psi = 0, \tag{1.24}$$

which differs from the free-particle Dirac equation. A local phase invariance is not compatible with free-particle theory that does not include interactions between particles. Hence the Dirac equation is modified under the local phase invariance with a new term A_{μ} invariant under the local phase transformation $A_{\mu} \rightarrow A'_{\mu} = A_{\mu} - \partial_{\mu}\chi$ and becomes:

$$i\gamma^{\mu}(\partial_{\mu} + iqA_{\mu})\psi - m\psi = 0, \tag{1.25}$$

.

The new term $q\gamma^{\mu}A_{\mu}\psi$ is the interaction term of QED. The fact that physics is invariant under local U(1) phase transformations of the form U = $e^{iq\chi(x)}$ indicates that there exists a gauge field that couples to the particles in the same way as a photon.

1.2.2 QCD

Analogous to the U(1) local gauge symmetry of QED, the underlying symmetry of QCD is the invariance of SU(3) local phase transformations, given as:

$$\psi \rightarrow \psi(x)' = e^{[ig_S\boldsymbol{\alpha}(x)\cdot\boldsymbol{T}]]}\psi(x), \tag{1.26}$$

where **T** = $[T^{\alpha}]$ are the eight generators of the SU(3) symmetry group and $\alpha^{\alpha}(x)$ are eight functions of the space-time coordinate x. The generators of the SU(3) group are 3x3 matrices, which means the function ψ must have three additional degrees of freedom. The new 3D vector degree of freedom is called *color* and the three states are named *red*, *blue* and *green* respectively. Plugging the equation 1.26 in the Dirac equation of 1.17 we get:

$$i\gamma^{\mu}[(\partial_{\mu} + ig_S(\partial_{\mu}\boldsymbol{\alpha})\cdot T]\psi - m\psi = 0, \tag{1.27}$$

.

Similar to the procedure for QED in the previous section, eight new fields which transform as $G^k_{\mu} \rightarrow G^{k\prime}_{\mu} = G^k_{\mu} - \partial_{\mu}\alpha_k - g_S f_{ijk}\alpha_i G^j_{\mu}$ are introduced, so that the new Dirac equation becomes invariant under the SU(3) phase transformations:

$$i\gamma^{\mu}[(\partial_{\mu} + ig_S G^{\alpha}_{\mu}T^{\alpha}]\psi - m\psi = 0. \tag{1.28}$$

.

The eight new fields G^{α} are the gluons and the interaction term of QCD is $g_S T^{\alpha} \gamma^{\mu} G^{\alpha}_{\mu} \psi$. Additionally, since the SU(3) group generators do not commute, this new interaction term gives rise to self-interactions of gluons.

Comparison of the two different interaction types can help us see the bigger picture better. The QCD interaction is mediated by eight massless gluons, which correspond to the eight generators of the SU(3) local gauge symmetry, whereas QED is mediated by the massless photon, the generator of the U(1) local gauge symmetry. The single charge of QED corresponds to the three color charges r, b and g for QCD. Particles with non-zero color charge are the only ones that can couple to gluons, which is why leptons (all color neutral) do not experience the strong force. The color charge is carried by the quarks, which exist in three orthogonal color states. QCD interaction strength is thus independent of the color charge of the quark. Antiparticles have the opposite electric charge in QED, whereas antiquarks have the opposite color charges $\bar{r}, \bar{b}, \bar{g}$.

If free quarks existed, they would present themselves as fractionally charged particles. However, no such particles have been detected so far by any particle physics experiment. The hypothesis of *color confinement* states that coloured objects are confined to color singlet states and that objects with non-zero color charge cannot propagate as free particles. The main idea behind color confinement is following: similar to the QED interaction, a quark-quark interaction can be thought of as an exchange of a virtual gluon. Unlike the electrically neutral photon, a gluon carries the color charge. Therefore, the field lines surrounding the quark-quark system get squeezed into a tube shape due to the attraction between the gluons, whereas the electric field lines spread as the distance from an electron-positron system increases. The energy stored in the field is proportional to the distance between the quarks and it would take a very large amount of energy to separate the two quarks. Therefore quarks are confined to colorless hadrons.

1.2.3 The weak interaction

As discussed previously, QED and QCD have several common features; both are mediated by massless spin-1 bosons and the interaction terms both have the form

$\gamma^\mu \bar{u}(p')u(p)$. The charged-current weak interaction is quite different in many aspects, the first being the only interaction in SM where *parity* is not conserved. Parity operation is the spatial inversion at the origin defined as:

$$\psi(\boldsymbol{x},t) \rightarrow \psi'(\boldsymbol{x},t) = \hat{P}\psi(\boldsymbol{x},t) = \psi(-\boldsymbol{x},t) \tag{1.29}$$

and is conserved for both QED and QCD interactions. The non-conservation in the weak interaction was demonstrated by Wu and collaborators in 1957 by the study of β decay of polarised cobalt-60: $^{60}Co \rightarrow^{60} Ni^* + e^- + \bar{(}\nu)_e$ [11]. In the study, the ^{60}Co nuclei have a permanent nuclear magnetic moment $\boldsymbol{\mu}$ were aligned in a magnetic field $\mathbf{B}$ were emitting electrons via β decay. Both $\mathbf{B}$ and $\boldsymbol{\mu}$ do not change under the parity transformation. However, the sign of the momentum vector of the emitted electrons does change under the parity transformation. Thus, if parity were conserved, electrons emitted from the direction relative to the $\mathbf{B}$ field should be found on the opposite hemisphere as well, which was not observed.

The weak charged-current interaction term is given as:

$$\frac{-ig_W}{\sqrt{2}}\frac{1}{2}\gamma^\mu(1-\gamma^5), \tag{1.30}$$

where g_W is the weak coupling constant.

The chiral structure of the weak interaction is another important characterization of the interaction. The left and right handed chiral projection operators are given as:

$$P_R = \frac{1}{2}(1+\gamma^5) \text{ and } P_R = \frac{1}{2}(1-\gamma^5), \tag{1.31}$$

where any spinor can be decomposed into the left- and right-handed chiral components:

$$u = \frac{1}{2}(1+\gamma^5)u + \frac{1}{2}(1-\gamma^5)u = a_R u + a_L u, \tag{1.32}$$

,

with coefficient a_R and a_L. The weak interaction term 1.30 already includes the left handed chiral operator. The four vector weak-charged current given by:

$$j^{\mu} = \frac{g_W}{\sqrt{2}}\bar{u}(p')\frac{1}{2}\gamma^{\mu}(1-\gamma^5)u(p) \tag{1.33}$$

would thus become 0 for two right-handed chiral states. Therefore only left-handed chiral particle states (and right-handed antiparticle states) engage in the charged-current weak interaction. For massless particles and in the limit m << E chirality is equal to the helicity $H = (\boldsymbol{p}\cdot\boldsymbol{s})/|\boldsymbol{p}|$ which is the sign of the scalar product of momenta and spin. This means in the ultra-relativistic limit, only specific helicity combinations are allowed for weak interactions. This is the foundation of the phenomenon called *parity violation*, where the parity operation transforms an allowed weak interaction into one that is not allowed, violating the conservation of parity.

In the 1960s Glashow, Salam and Weinberg unified the theories of electromagnetic and weak interactions [12] [13], which postulates the mediation of a weak nuetral-current by the neutral Z boson. Similar to QED and QCD, the weak charged current interaction is associated with an underlying local gauge symmetry, $SU(2)_L$, which brings about the charged W bosons and a neutral gauge field $W^{(3)}$. In the unified electroweak theory, the neutral gauge field $W^{(3)}$ mixes with a photon like field of $U(1)_Y$ gauge symmetry to generate the photon and Z-boson fields:

$$A_{\mu} = +B_{\mu}cos\theta_W + W_{\mu}^{(3)}sin\theta_W \tag{1.34}$$

$$Z_{\mu} = -B_{\mu}sin\theta_W + W_{\mu}^{(3)}cos\theta_W \tag{1.35}$$

,

where θ_W is the weak mixing angle and B_{μ} is a new gauge field that couples to a charge called weak hypercharge Y. The couplings of photon, W and Z bosons are related with each other via: e = g_W sinθ_W = g_Z sin θ_W cosθ_W.

1.3 Beyond The Standard Model

SM is a model based on many theoretical ideas that were put together in order to reproduce the available experimental data. Despite its recent success of passing several precision tests and the discovery of the predicted Higgs boson, it is not the final theory of particle physics because of the many unanswered questions. A brief overview of some these topics will be given in this section.

Dark Matter

There are several direct evidences for the existence of dark matter, the most prominent being the velocity distributions of stars as they orbit the galactic center. Assuming that most of the mass is located at the center of the galaxy, the tangential star velocities should decrease as $r^{-1/2}$, however this is not observed. They decrease much slower which implies that the galaxy has a significant non-luminous mass component [14]. Further evidence comes from cosmological and astrophysical measurements, in particular from the precision measurements of the cosmic microwave background (CMB) [15] and gravitational lensing [16] . Within ΛCDM cosmological model only 5% of the energy-matter density of the Universe is in the form of baryonic matter, 23% of the energy-matter density is cold dark matter whereas the majority of the density is in the form of dark energy [17]. Many candidates have been proposed for the dark matter such as cold non-baryonic matter, specifically weakly-interacting massive particles (WIMPs)[18]. Neutrinos are the known WIMPS, however their masses are too small to account for the dark matter content of the universe.

Matter-antimatter symmetry

The Universe seems to be made out of mostly matter. However, SM predicts that matter and antimatter should exist in equal parts if the initial universe conditions did not have any disproportionate matter, but fails to explain the observed asymmetry. Asymmetric interaction processes such as CP-violating electroweak processes [19] can account for a fraction of the observed asymmetry, however this would indicate the existence of other unknown CP-violating processes that are responsible for the total asymmetry.

Neutrino masses

SM predicts neutrinos as massless particles, however neutrino oscillation experiments have shown that neutrinos do have small masses [20], and that they can mix together [21], just like the different types of quarks (with the same quantum numbers) can mix together. Thus highly energetic neutrinos can oscillate, change type from one flavor into another as they traverse matter. Furthermore, every observed neutrino is left-handed, and anti-neutrinos are right-handed. However, if neutrinos have mass, this would indicate that neutrinos are their own antiparticle, like a Majorana particle [22]. Experiments are therefore focusing on the possibility of neutrinoless double beta decay, which can only happen if neutrinos are Majorana particles.

Gravity

Gravity is not described by the SM due to the contradictory terms that arise when combining general relativity, theory of gravity and quantum mechanics. There are several theoretical studies on the unification of gravity with SM such as string theory [23], loop quantum gravity [24]. However at the energy scales of high energy physics (HEP) experiments, gravity is very weak and is therefore neglected.

Chapter 2

The Experimental Apparatus

In this section an overview of the LHC accelerator mechanism will be given, the experiments will be introduced and the Compact Muon Solenoid (CMS) experiment will be described in detail.

2.1 Large Hadron Collider

The Large Hadron Collider (LHC) is the highest energy particle accelerator and collider in the world. It is located within a tunnel of 26.7 km circumference and 50-175 m deep below the border between France and Switzerland (Figure 2.1), and operated by the European Organisation for Nuclear Research (CERN). Primarily collisions are generated by two beams of protons, accelerated to high energy velocities and intersected in the center of detection systems. Superconducting quadrupole magnets are used to keep the beams directed to and focused at the intersection points and dipole magnets make sure that the beams stay on the circular path. The proton beam is not continuous, but split into bunches so that the collisions happen at discrete intervals (25 ns) with a rate of 40 MHz.

Previously the same tunnel housed the Large Electron-Positron (LEP) Collider where electrons and positrons were being collided in similar manner. However, because electrons are less massive, the synchrotron radiation loss was very large and this resulted in the particles not being accelerated to the high velocities in an efficient manner. The maximum beam energy achieved by LEP was 209 GeV, whereas the beam energies reached by the LHC is at 6.5 TeV, giving a total 13 TeV collision energy, which is the current world record.

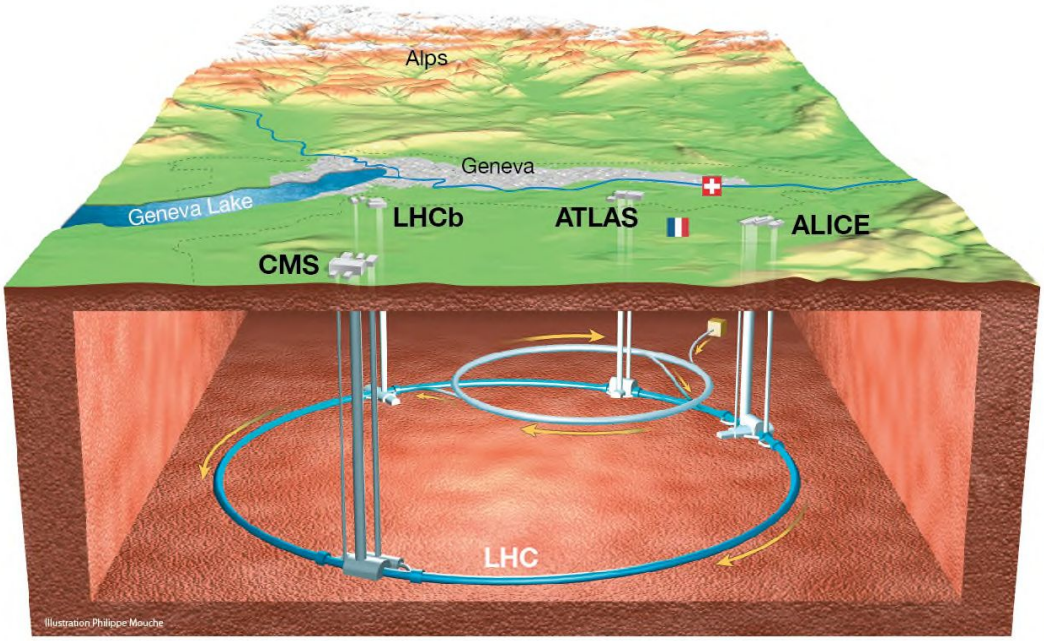


FIGURE 2.1: LHC tunnel below the French-Swiss border with the four experiments CMS, ATLAS, LHCb and ALICE at the interaction points [25] .

In 2009 LHC started circulating low energy beams and was able to achieve beam energies of TeV energy scale, beating the previous record held by Tevatron. By 2010 the total collision energy reached 7 TeV. This operation period called Run I, crowned with the discovery of the Higgs Boson in 2012, was concluded early 2013. After a shutdown of two years, which allowed many detector components to be upgraded, LHC started collecting data again in 2015. During Run II which lasted until 2018, the beam energy reached 6.5 TeV, resulting in a combined energy of 13 TeV. Data analysed within the scope of this work has been delivered during the Run II period.

One of the most important performance parameters for the detector alongside energy is luminosity, which is a measure of how many collisions take place in the detector. The higher the luminosity, the better chance does the LHC have of producing rare events of interest. Additionally, operating at higher luminosity results in data samples with large statistics that are necessary for more precise measurements. Luminosity is defined as the ratio of the number of detected events (dN) in a certain period of time (dt) to the cross-section (σ):

$$L = \frac{1}{\sigma}\frac{dN}{dt}, \tag{2.1}$$

with the dimensions of number of events per area per time $[cm^{-2} \cdot s^{-1}]$.

Starting from equation 1.9 the instantaneous luminosity at the LHC at a certain time at the interaction point can be estimated with the beam parameters ϵ denoting the transverse emittance and β^* denoting the amplitude function at the interaction point:

$$L = \frac{f \cdot N^2}{4 \cdot \epsilon \cdot \beta^*}, \tag{2.2}$$

where N denote the number of protons per bunch per beam and σ the corresponding cross section and f the crossing frequency (40 MHz). The emittance can be defined as the smallest opening the beam can squeeze through and is a measure of how parallel the beam is. The amplitude function $\beta = \pi \cdot \sigma^2/\epsilon$ is proportional to the width of the beam squared divided by emittance and is determined by the quadrupole magnet configuration. Low amplitude function and a low emittance give a narrow "squeezed" beam and per eq. 2.2 a higher luminosity.

The integrated luminosity is the integral of the luminosity with respect to time:

$$L_{int} = \int L dt, \tag{2.3}$$

with the dimensions of inverse cross section $[nb^{-1}]$. It is a measure of the size of the data collected over a certain time period, and therefore an important parameter for the performance of the accelerator (see Fig. 2.2 for the yearly integrated luminosity of CMS experiment).

The design luminosity of LHC is $10^{34}\ cm^{-2}\ s^{-1}$, which was first reached during Run II and is 100 times higher than the maximum luminosity achieved by LEP. During the shutdown between 2018 and 2022, the whole accelerator complex was maintained and upgraded for the preparation of the High Luminosity Large Hadron Collider (HL-LHC) phase, which will increase the luminosity by a factor of 10. As of writing, LHC is operational again for the Run III period (expected to end in 2026) with a beam energy of 6.8 TeV.

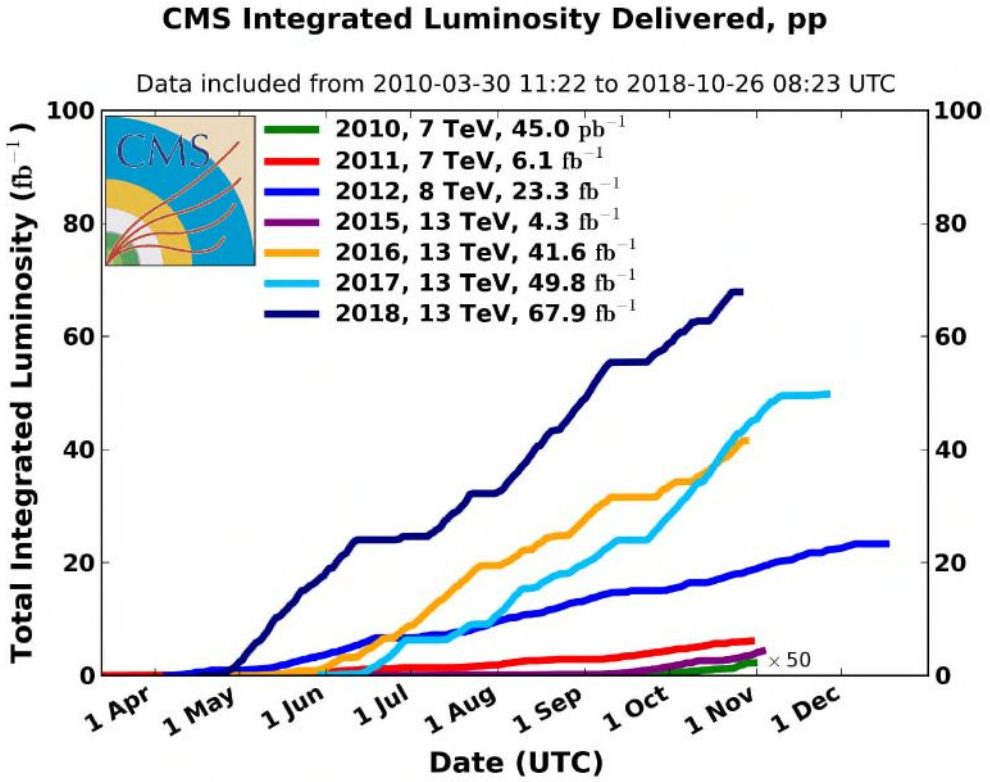


FIGURE 2.2: Integrated luminosity of CMS over yearly periods during Run I and Run II [26].

2.1.1 Injection and Acceleration of Protons

Protons in the LHC beam originate from a hydrogen tank and they go through a series of accelerating components where they increase their energy successively. The first step is the injection into a linear accelerator, called LINAC4 (see Fig. 2.3), which accelerates negative hydrogen ions up to the energy of 160 MeV. The ions are then injected into the Proton Synchrotron Booster (PSB), where they lose the electrons leaving the nucleus in the beam, which are then accelerated to the energy of 2 GeV. In this way by accelerating hydrogen ions instead of protons the beam loss during the injection is reduced. The next step is the injection into the Proton Synchrotron (PS) where they are accelerated to 26 GeV. The last accelerating component is the Super Proton Synchrotron which accelerates the protons to 450 GeV before injecting them into the main LHC ring in opposite directions. In order to minimize interaction of the particles with the gas molecules in the air, the beam pipes are kept at ultra high vacuum conditions ($10^{-10} - 10^{-11}$ mbar).

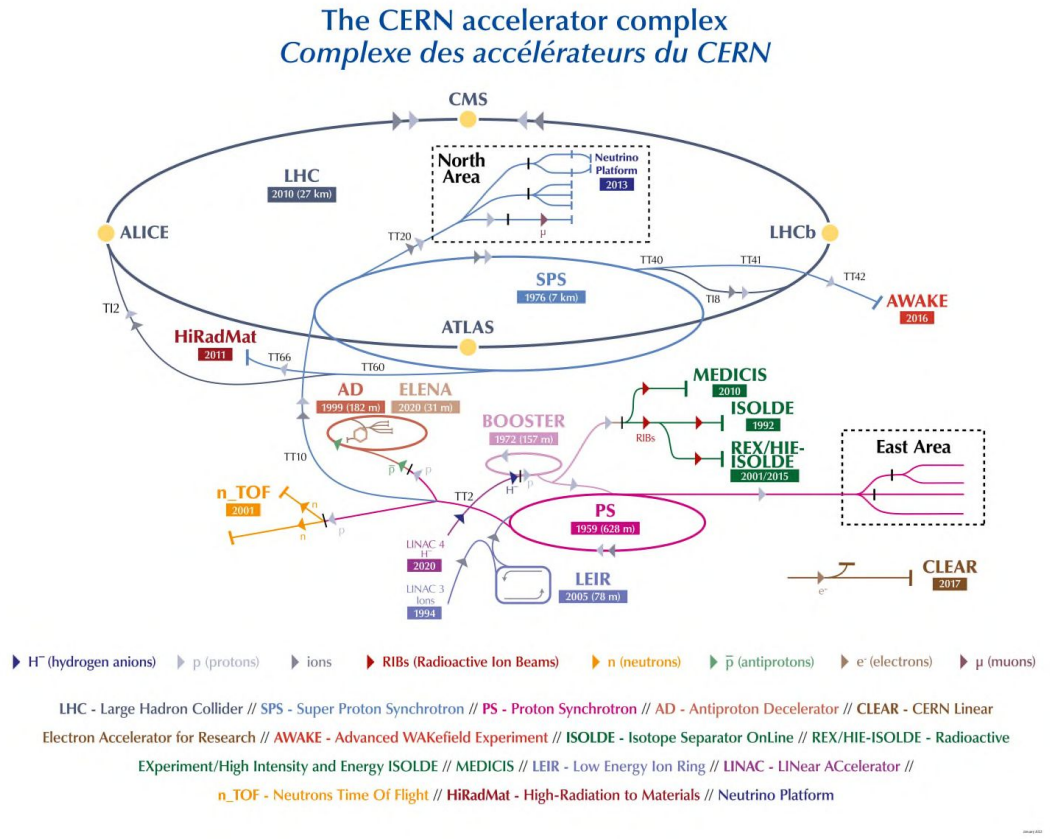


FIGURE 2.3: Overview of the CERN accelerator complex with the lINAC4, PSB, PS, SPS, LHC ring and the experiments [27].

2.1.2 LHC Ring

In order to keep the protons on the circular trajectory, strong magnetic fields are needed, which are provided by the 1232 dipole magnets. 392 quadrupole magnets force the beams further to close in at the intersection points so that the probability of the interaction of opposing protons is maximized. Additional higher multipole order magnets are used for minor field geometry corrections along the beam pipe. About 96 tonnes of superfluid helium is needed to cool the magnets and keep their temperature at 1.9 K (-271.25 °C).

Before the collisions can take place, the injected protons are accelerated from 450 GeV to 6.5 TeV via 16 radio-frequency (RF) cavities, metallic chambers containing electromagnetic field, where protons receive an electric impulse and get accelerated. Each RF cavity oscillates at a frequency of 400 MHz, so that a proton with the correct energy that reaches the cavity at the expected time will not be accelerated further,

while the slower protons arriving later will be accelerated (faster ones will be decelerated). In this way the proton beam is split into packs of protons called "bunches". Each cavity can reach a maximum voltage of 2 MV and they operate at a temperature of 4.5 K. At 2 MV voltage and 8 cavities per beam, each beam proton gets an 16 MeV energy. The collision energy of 6.5 TeV is reached in about 20 minutes after the bunches pass through the cavities more than 10 million times. During this acceleration time, the protons are steered away from the interaction points by dipole magnets.

After the proton bunches reach the desired energy levels, they circulate the ring and collide at the four interaction points. After some time (defined later as an LHC fill) the protons are dumped (they exit the ring and collide with graphite absorbers that are tangent to the beam pipes). The period between the initial injection and the dump is characterized as an *LHC fill*, and its duration depends on whether there is a technical issue or the beam properties degrade beyond correction or beam intensity drops too low due to losses in collisions and from beam-gas interactions.

In order to maximize the probability of observing rare particle interactions, LHC was designed to collide proton bunches each consisting of thousands of protons every 25 ns. However, this results in the collisions of interest being recorded together with a large number of unwanted additional proton collisions, so-called *pileup* (PU) interactions (see Figure 2.4). During Run II the number of PU interactions in the years 2017 and and 2018 were 32, and in short periods of time surpassed 50 (the average value is proportional to the instantaneous luminosity and the cross section of the process that takes place at the PU interaction, per equations 2.2 and 2.3). Most of the PU interactions end up in a large number of low energy particles around the interaction point (soft scattering interactions). While the probability of recording multiple hard collisions is very low due to the small cross section of such processes, selecting the interesting, high-energy hard interactions is a major challenge since the interaction outcome is contaminated by the surrounding soft interactions which are recorded at the same time. Many pileup mitigation techniques are employed to this purpose where the primary vertex of the interesting interaction is identified and the charged particles coming from pileup vertices are rejected.

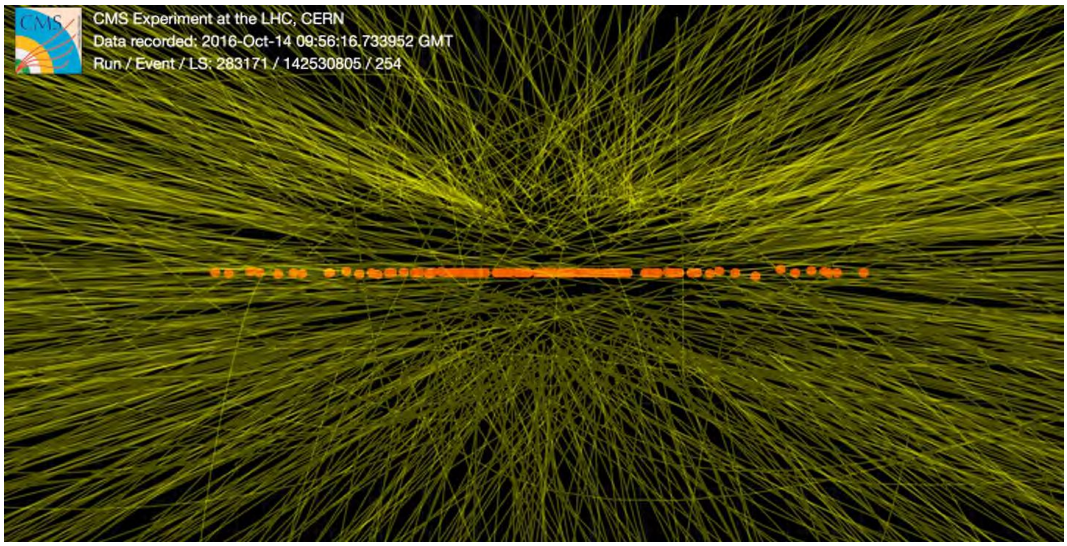


FIGURE 2.4: CMS event display with proton-proton collisions at a center-of-mass energy of 13 TeV, recorded during the high pile-up fill of Run 2 (October 2016). The events are from isolated bunches with average pileup roughly around 100. Yellow lines indicate the trajectories of the particles, while orange dots are the reconstructed primary vertices [28].

2.1.3 Detectors

Eight detectors are installed at the LHC intersection points to record and study the outcome of highly energetic particle collisions. The four largest are the following:

- **ATLAS** (A Toroidal LHC ApparatuS) [29]

  ATLAS is a general purpose detector and the largest experiment at the LHC. It was designed to primarily to study the highest-energy phenomena, which may involve the production of new high-mass particles. It has a broad study program ranging from Standard Model to search for evidence of beyond the Standard Model theories.

- **CMS** (Compact Muon Solenoid) [30]

  CMS is a general-purpose detector and has a similar research program to that of ATLAS. CMS and ATLAS performed independent studies that led to the discovery of the Higgs boson in 2012 [3] [4]. It has a different magnet system design with a magnetic field twice as intense (3.8 T) and is overall more compact than ATLAS. This detector recorded the data that was analysed in this work, and is described in detail later in this section.

- **LHCb** (Large Hadron Collider beauty) [31]

  LHCb is a single arm forward spectrometer and was designed to study primarily the CP violation in the interactions of b-hadrons, which are predominantly produced in the forward region. Such studies can shed light on the causes of the matter-antimatter asymmetry in the universe.

- **ALICE** (A Large Ion Collider Experiment) [32]

  ALICE records heavy ion (Pb-Pb) collisions at a center of mass energy up to 5 TeV, where the high energy density allows the existence of quark-gluon plasma, a state of matter where quarks and gluons that make up the hadrons are freed of their strong attraction for one another. Believed to be the primordial matter form, which existed within a fraction of a second after Big Bang, quark-gluon plasma will enhance our understanding of the nature of strong interaction as well as the mechanism that confines quarks and gluons.

The last four, TOTEM [33], MoEDAL [34], LHCf [35] and FASER [36] are smaller detectors and have specialized research goals. TOTEM is located at the same interaction point with CMS and positioned tangent to the beam pipe, focuses on measuring total cross section, elastic scattering and diffraction processes. MoEDAL shares the interaction point with LHCb and works primarily on the search for the magnetic monopoles and other highly ionizing stable massive particles. LHCf shares the interaction point with ATLAS and was designed to measure the energy and the number of neutral pions produced in the forward region of the collider with the purpose of understanding the origin of ultra-high-energy cosmic rays. FASER is installed at the interaction point used by ATLAS and will primarily search for new light and weakly coupled elementary particles, such as dark photons, axions and sterile neutrinos when it will be operational at Run III in 2022.

2.2 Compact Muon Solenoid

CMS is a general purpose detector that was installed at one of the four collision points at the LHC ring with the purpose of exploration of physics at TeV energies. This includes performing several studies on high energy physics phenomena such

as further investigation of the properties of the Higgs boson, discovered by CMS and ATLAS, search for evidence of physics beyond the Standard model such as supersymmetry, and scan the collision outcome for remnants of dark matter candidate production.

Photons, muons, electrons and other products of the collisions can be detected and identified by CMS via the signatures they leave on the sub-detectors that are designed to measure the energy and momentum of particles with high accuracy [38]. The innermost layer is a silicon tracker (as depicted on Figure 2.5), which is surrounded by a scintillating lead tungstate crystal electromagnetic calorimeter (ECAL). ECAL itself is encompassed by a scintillating brass hadronic calorimeter (HCAL). The subsystem of calorimeters and the tracker is fit inside the solenoid magnet with a diameter of 6 m capable of generating a homogeneous magnetic field of 3.8 T. A steel flux-return yoke encloses the solenoid and confines the magnetic field, wherein lie the large muon detectors, composed of cathode strip chambers, resistive plate chambers and drift tubes.

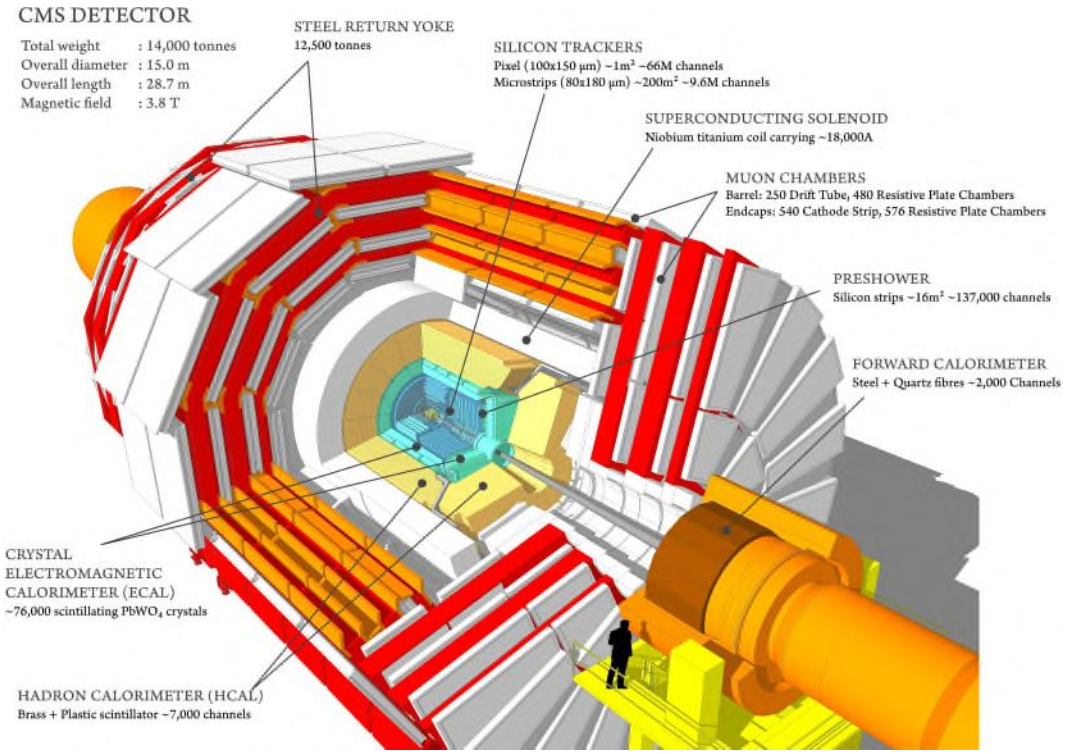


FIGURE 2.5: A cutaway diagram of CMS detector, with the sub-detector components indicated [37]

2.2.1 Interaction Point

The interaction point is where the beams are focused at by the quadrupole magnets and where the collision takes place. The CMS detector is accordingly centered around this point. At the time of collision the opposing beams each have a radius of 17 μm and an angle of 185 μrad. At full luminosity each beam contains 1.15×10^{11} protons per bunch at the start of an LHC fill and in total 2808 bunches per beam.

CMS uses a right-handed coordinate system with the origin at the interaction point, the x-axis pointing to the LHC ring center and the y-axis pointing upwards, perpendicular to the LHC plane, while the z-axis is tangent to the beam line and is along the anticlockwise beam direction. In polar coordinates, the azimuthal angle ϕ is measured from the positive x-axis in the x-y plane, while the polar angle θ is measured from the positive z-axis with respect to the LHC plane. The radius r is the distance from the z-axis.

Rapidity is a spatial coordinate defined as the angle between the particle momentum $\vec{p}$ and the beam axis:

$$y = -\frac{1}{2} ln(tan(\frac{E + p_z}{E - p_z})), \tag{2.4}$$

In the limit where particles are travelling close to the speed of light (m « $|\mathbf{p}|$), it can be approximated by the *pseudorapidity* given by:

$$\eta = -ln(tan(\tfrac{\theta}{2})), \tag{2.5}$$

which depends only on the polar angle θ. Particle trajectories with η = 0 (θ = $\pi/2$) are perpendicular to the beam, while particles with high η values generally escape along with the beam and are lost to the detector.

The pseudorapidity is commonly preferred over the polar angle θ, because the difference between the rapidities (or pseudorapidities) of two particles is invariant with respect to Lorentz boosts along the z-axis, whereas the difference in polar angle θ between two particles depend on the initial parton state Lorentz boost in the z direction, which is difficult to determine for different collisions in the laboratory frame of reference. Angular separation is defined using the pseudorapidity η and

the azimuthal angle ϕ:

$$\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}, \tag{2.6}$$

where $\Delta\phi$ is invariant under Lorentz boosts along the z-axis since it is measured on the x-y plane.

2.2.2 Tracker

Measurement of particle momentum is crucial for reconstructing the events taking place at the collision center. Therefore the tracking system is the inner most layer of the detector and the closest to the interaction point (Figure 2.5). The momentum is estimated by recording the trajectories of charged particles at a number of key points as they are passing through a magnetic field; the straighter the trajectory, the higher momentum the charged particle had, whereas particles with less momentum leave a curved path signature behind. Each key trajectory point measurement is accurate to 10 μm [39] for high momentum particles. The trajectories, called *tracks*, can be detected within a pseudorapidity range of $|\eta| \leq 2.5$ [40].

In order to minimize the disturbance to the particle trajectory, the tracker material has to be as lightweight as possible. Furthermore, since the tracker is the innermost layer, it receives the highest flux of charged particles from each bunch collision at a rate of 40 MHz. Therefore, high radiation tolerance (*radiation hardness*) is an important requirement for the tracker materials. To this purpose, the tracker is made completely out of silicon material and has two main parts: an inner silicon pixel layer and an outer silicon micro-strip layer surrounding the former layer. When the charged particles pass through the layers, the pixels and the micro-strips produce electric signals, which get amplified and detected.

In total, the tracker contains 1440 pixel modules with 66 million pixel cells covering about 1 m^2 total area, while the strip detector surrounding it, has 15148 strip modules with 10 million read-out channels covering a sensitive area over 200 m^2. The pixel detector is the innermost layer of the tracker in the barrel region set up cylindrically around the interaction region of the LHC beams (as seen in Figure 2.6). Each pixel cell has an are 100 x 150 m^2 and a depth of 285 m and gives a precise 2D

location of the hits that coincide with the cell. Together with the information of the cell position, the 3D coordinates of the hit are obtained. The strip detector consists of four sub-detectors: 4 layers of tracker inner barrel (TIB), 6 layers of tracker outer barrel (TOB), 3 tracker inner disks (TID) and 9 tracker endcap disks (TEC) covering up to 110 cm in radii and 280 cm in z direction (see Figure 2.6). In the barrel region the strips are set up parallel to the beam line, while in the endcap section they are perpendicular in a back-to-back arrangement, where each one of the neighboring module is rotated slightly. In this way, the location of the hit can be found in 2D (together with the module position, in 3D).

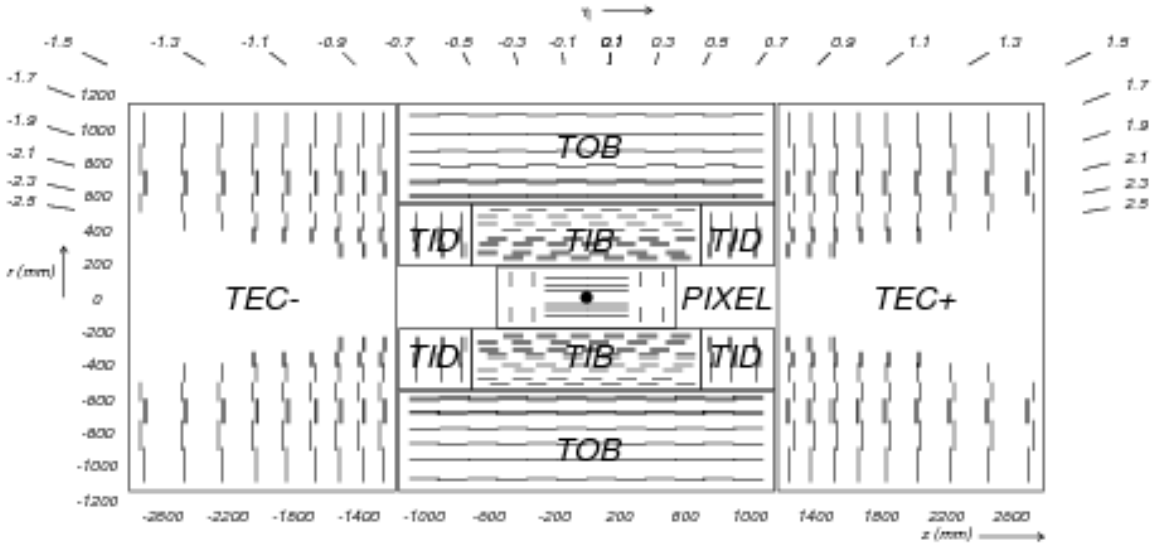


FIGURE 2.6: Schematic cross section through the CMS tracker in the r-z plane. [41].

2.2.3 Electromagnetic Calorimeter

The electromagnetic calorimeter (ECAL) measures the energies of electrons, positrons and photons by detecting the scintillation light, proportional to the particles' energy, which gets produced when they enter the calorimeter and cause electromagnetic showers. It is made of crystals of lead tungstate ($PbWO_4$), a dense but optically transparent material, which produces visible range light in the form of short, well-defined bursts of photons and thus can be detected with high precision [42]. The

crystals are positioned in carbon fiber so that they are optically isolated and surrounded by photo-detectors that detect the light and convert it into signal to be amplified and read out. The radiation length (see section 1.1.3) of the lead tungstate crystals is X_0 = 0.83 cm allowing the EM showers to stay in a compact region.

When high energy electrons or positrons enter the calorimeter, they get stopped by the dense lead tungstate crystals and during the deceleration, they emit photons via Bremsstrahlung radiation. For photons with high energy (MeV scale and higher; below this scale, photoelectric effect and Compton scattering are dominant), pair production of electron-positron is the dominant mode of interaction with matter, both of which in turn emit more photons via Brehmsstrahlung radiation as they decelerate. The two processes continue with decreasing energy, leading to a cascade of particles, referred as an electromagnetic shower, until the photon energy falls below the pair production energy threshold and other processes dominate the electron energy loss other than Brehmsstrahlung. Low energy photons in the final layer of the cascade produce light in the visible range in the crystals which gets detected and converted into electric signal by the photo-detectors.

ECAL is composed of two main parts: the barrel and the two endcaps (see Figure 2.7), and is positioned between the tracker and the HCAL (as depicted in Figure 2.5). The cylindrical barrel with an inner radius of 129 cm consists of 61200 blocks of crystals grouped together into 36 supermodules, each with 1700 crystals of 3 cm width that are placed around the collision region in a radial direction. The barrel has an inner radius of 129 cm and covers a pseudorapidity range of $|\eta| < 1.479$. The flat endcaps close off the barrel on either side at a distance of 314 cm from the collision point and carry 7324 further crystals each. It covers a pseudorapidity range of $1.479 < |\eta| < 3$.

Furthermore ECAL is equipped with preshower detectors (see Figure 2.7), which are made of two layers of lead interleaved with two layers of silicon strip detectors of 2 mm width, and thus have a higher granularity than the ECAL crystals. A single high energy photon is usually a sign of a rare physics process, however this signature can be missed by the ECAL when a neutral pion decays into two closely spaced lower energy photons that are identified as one photon. The preshower detectors are placed at the endcap sections, where the angle between the photons emitted

from the pion decay is small enough to be misidentified by the ECAL.

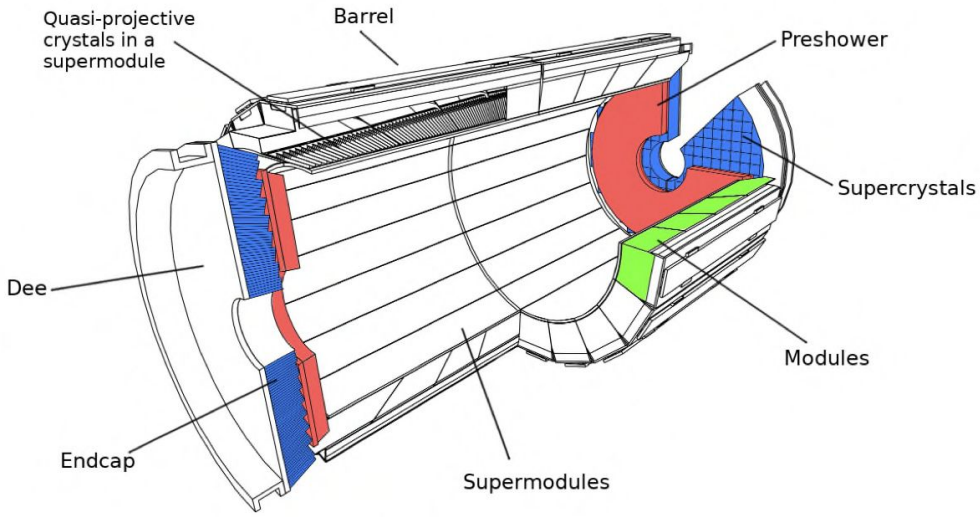


FIGURE 2.7: ECAL layout [43].

2.2.4 Hadronic Calorimeter

The hadronic calorimeter (HCAL) measures the energy and direction of hadrons, such as protons, neutrons, pions and kaons by detecting the scintillation light emitted when primary and secondary hadrons traverse the active material in the device. It consists of layers of dense material (brass and steel) acting as absorbers, where the incoming particles interact with the material nuclei and form secondary particles, interleaved with layers of plastic scintillator tiles, which convert the energy (deposited in the previous layer) into visible light to be collected and converted into electric signal by photodetectors.

The HCAL is a nearly *hermetic* (full solid angle) calorimeter, where to the extent of its capability, the presence of every particle that emerges from the interaction point is detected. If there is an imbalance in the measured momentum and energy (in the transverse direction with respect to the beam line), this would mean that there were "invisible" particles produced in the collision. However, this can only be deduced if the calorimeter does not let any particles through without detection. To this purpose HCAL is built with many alternating layers of absorbing material, as a so called sampling calorimeter. One quarter of HCAL layout is depicted in Figure 2.8 with its main components: the hadron barrel (HB), the two endcap sections (HE) at either side, the hadron forward (HF) located at a distance of 11.15 m from the interaction point at either side, and the hadron outer (HO) calorimeter as the outermost layer.

The barrel section HB sits between the radii 177.5 cm and 287.7 cm, covers a pseudorapidity range of $|\eta| < 1.39$, and is made of brass absorber plates with sampling layers of about 40000 plastic scintillators that are grouped into 16 parts according to the location in $|\eta|$ coordinate. The endcaps are made of the same material as HB and cover the pseudorapidity range of $1.30 < |\eta| < 3.00$. The endcap on each side is divided into 14 parts in $|\eta|$ as depicted in Figure 2.8. The outer barrel (HO) consists of scintilator layers, and is positioned outside the magnet coil to make sure no energy goes undetected outside the barrel region. The central ring (ring 0) has two layers of scintilators, each 10 mm thick (at radial distances 385 cm and 409.7 cm) and a stainless steel block in between, whereas all other rings have a single layer of scintillator. The HO calorimeter follows the HB $\eta - \phi$ phase-space segmentation closely

and covers the pseudorapidity range of $\eta < 1.26$.

Lastly, two hadronic forward calorimeters (HF) sit at either end of CMS to detect the particles coming from the interaction point at a low angle with respect to the beam line. Since they receive a high influx of particles, the calorimeter material has to be more resistant to radiation compared to the other components of HCAL. The two HF calorimeters cover the pseudorapidity range of $2.85 < |\eta| < 5.19$ and consist of 18 wedges made of steel with quartz fibers embedded along them. Secondary charged particles going through the quartz fibers emit Cherenkov light which get collected and converted to charge by the photomultiplier tubes. Long (164.9 cm) and short (142.6 cm) quartz fibers are placed alternately in a way that the long fibers reaches the calorimeter front and the short fibers start at 12.5 radiation lengths within the calorimeter. In this way the energy deposit difference between the fibers aids the differentiation between electromagnetic and hadronic showers, which are much longer than the former.

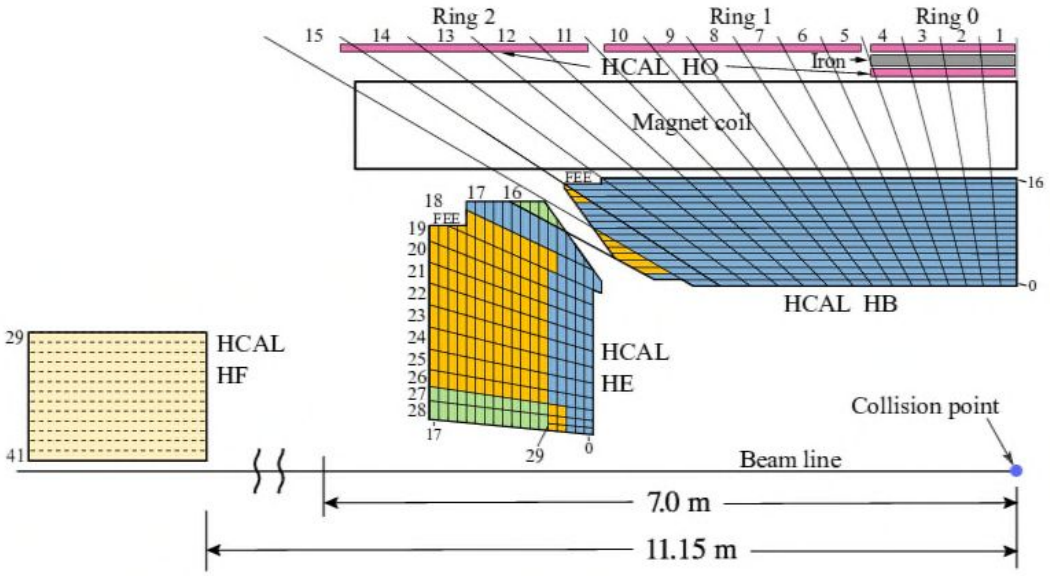


FIGURE 2.8: A schematic view of one quarter of the CMS HCAL with its four major components: the hadron barrel (HB), the hadron endcap (HE), the hadron outer (HO), and the hadron forward (HF) calorimeters [44].

2.2.5 Magnet

CMS has a large superconducting solenoid magnet, about 13 m long and 6 m in diameter, and provides an almost homogeneous magnetic field up to 3.8 T (at its center) in z direction. As mentioned before, its purpose is to bend the trajectories of charged particles, emitted from the interaction point, so that their charge can be determined and their transverse momenta p_T can be estimated (equation 1.2). It is the central component of the CMS experiment (as seen in Figure 2.5) and gives structural support to the whole detector. The inductance of the magnet is 15 H and for 3.8 T the nominal current is 18160 A. Its superconducting niobium-titanium coils are refrigerated at 4.5 K using a liquid helium cooling system.

2.2.6 Muon System

Muon detection is an important task for CMS, as the detection of those particles may indicate their origin in the decay of short-lived particles of interest, for example one of the decay channels that was studied extensively is the Higgs Boson decay into four muons. As muons pass through the tracker, their curved trajectories are detected just like other charged particles, however, they are not stopped by either of the calorimeter materials. Given that muons are much more massive than electrons, they do not suffer as much energy loss due to Bremsstrahlung radiation (electrons lose energy due to Bremsstrahlung at a rate $(m_\mu/m_e)^4 \approx 10^9$ times higher than muons). Since they do not deposit a significant amount of energy at ECAL or HCAL and are not detectable from the interaction point up until the magnet, the muon detection systems are set up as the outermost layer of the experiment, where the probability of detecting muons only, is high. The muon tracks once determined in the muon detector, are matched to the ones in the tracker.

To identify muons and estimate their momenta, three types of gaseous detectors are used: drift tubes (DT), cathode strip chambers (CSC) and resistive plate chambers (RPC). The drift tube chambers are placed at the barrel part of the detector (as depicted in Figure 2.9), where each DT chamber consists 12 aluminium layers, each with 60 tubes. Each tube holds a wire that is stretched in a volume of gas. When a charged particle passes through the gas volume, it knocks electrons off the gas

atoms, which move towards the positively charged wire due to the high electric field presence. Considering the positions where the electrons end up on the wire and the distance of the particle from the wire, DT is able to provide two coordinate points for the charged particle. DTs in the middle six layers measure the coordinate parallel to the beam and the outside six layers measure the coordinate in the perpendicular direction. Cathode strip chambers are placed in the endcap disks where the particle flux is high and the magnetic field is not uniform. They consist of arrays of positive charged wires (anode) that are crossed with negative charged copper strips (cathode) in a volume of gas. Similar to DT, when a charged particle passes through a CSC, it knocks off an electron from the gas atom, which moves to the anode, hitting other atoms and causing an avalanche of further electrons. On the other hand, positive ions move to the cathode and create an electric pulse in the strips. Since the wires are set up perpendicular and cover the whole 2D area, they provide precise time and location of the particle and thus are used for triggering purposes. Resistive plate chambers (RPC) are gaseous detectors with a similar mechanism as CSC and are placed at the endcap region (as depicted in Figure 2.9) of the detector. They consist of two parallel plates (anode and cathode) made of highly resistant plastic material which are placed in a gas volume.

Gas electron multiplier (GEM) detectors, once in operation, will complement the other detector components in the endcap section by offering additional coverage in the forward region. They consist of three layers of copper-cladded polyimide foil with mocroscopic holes placed in argon carbon dioxide gas mixture that gets ionised by the incident particles. The electron avalanche created by the ionization gets read out by strip detectors.

2.3 Trigger and Data Acquisition

In order to increase the probability of detecting interesting physics processes such as the production of a rare particle or a rare decay like the three prong decay of B_S, a large number of collisions take place, each of which generate approximately 1 MB of raw data. At the collision rate of 40 MHz the total rate of generated raw data becomes 40 TB per second, which totals an impractical size of data to store and

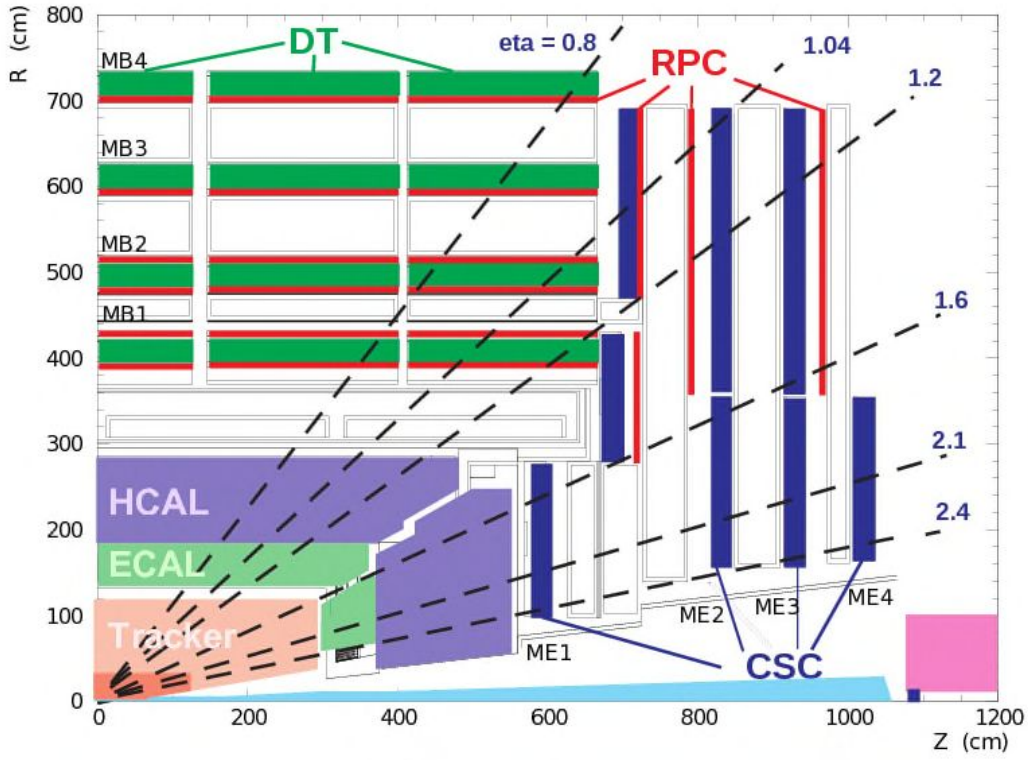


FIGURE 2.9: Layout of one quadrant of CMS [45].

analyze. For an efficient data acquisition a filtering technique is employed, where only the "interesting events" are stored. The condition that characterizes an event as interesting and to be stored or not is called a *trigger*.

CMS uses a two level trigger system. In the first stage *Level 1* (L1-Trigger) a small part of collision data is used to compute discriminatory features about the event, such as existence of muons, their properties or information about the missing transverse energy, using fast hardware devices called field-programmable gate arrays (FPGA) that can carry out logical operations. This key information is used to filter out events that fail the trigger condition. The computation takes about 1 μs and the event rate is reduced down to 50 kHz. In the second stage called *High Level Trigger* (HLT) all data that had passed the L1 trigger are sent to computer servers where software programs perform a more detailed test on the event content than that of the L1-trigger. HLT trigger lowers the event rate further down to 1000 events per second, which then get stored on tape.

2.4 Event Reconstruction

In a simplified view, the principle of the CMS experiment is as follows: particles coming from the collision point enter the tracker, where the charged particle tracks and vertices are reconstructed from the hits and the momenta and the electric charges are estimated from the curvature of the trajectories bent by the magnetic field. Electrons and photons that reach the ECAL get absorbed (see Figure 2.10), and their energy and direction can be determined by their electromagnetic showers which are detected and reconstructed as clusters of energy in the ECAL cells. Charged and neutral hadrons get fully absorbed in the HCAL, where they produce hadronic showers and the corresponding energy clusters are used for energy and direction estimation. Neutrinos escape the detector components undetected, while muons traverse the calorimeters with little interaction and get detected at the muon system placed outside the other detector components. Reconstruction of physics objects thus relies mainly on the specific detector component that is tasked with the object's detection; reconstruction of isolated photons and electrons is done primarily with ECAL data, jet reconstruction is done mainly with the calorimeter data (without separation of jet constituents), muon identification is primarily done with the information from the muon detectors, while the tagging of jets of hadronic decays and from b quark hadronization is performed with the tracker information.

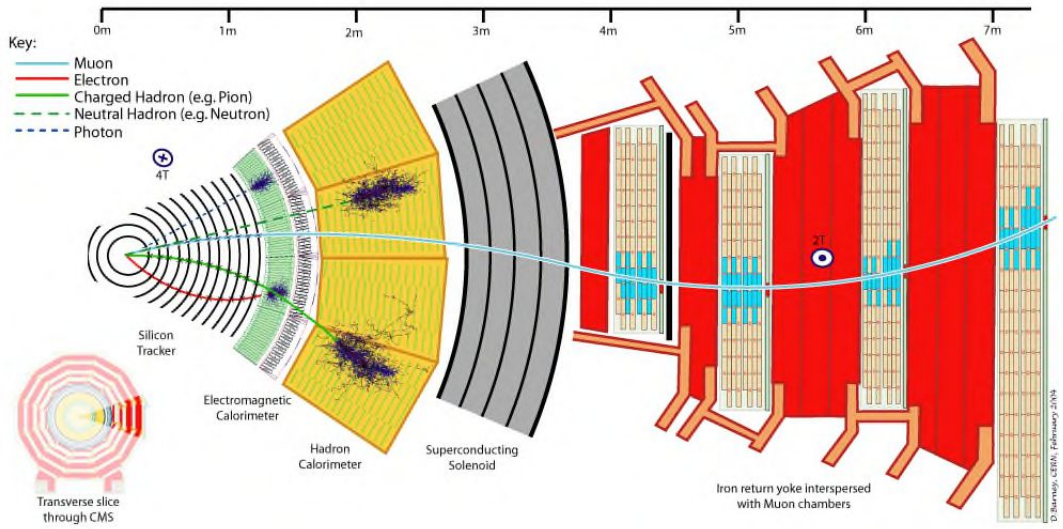


FIGURE 2.10: CMS detector slice [46].

The event reconstruction at CMS is heavily based on the *Particle Flow* (PF) algorithm [47], which uses the information coming from each sub-detector component to identify and reconstruct the final-state particles. Previously, traditional approaches were employed, where for each sub-detector a local reconstruction based on the tracks and hits would be carried out. Hadrons, muons, electrons and photons can be identified in a more efficient manner by combining all information on the initial position, energy and direction of the particles. A charged hadron is identified by a geometrical link in the $\eta - \phi$ space between an associated trajectory reconstructed in the tracker and one or more clusters in the calorimeter (and by the lack of signal in the muon system). The photons and neutral hadrons are identified by the clusters in the ECAL and HCAL without any geometrical link to tracks in the tracker. An electron is identified by a track linked to an ECAL cluster (no HCAL cluster), whereas muons are identified by a track in the inner tracker that can be connected to a track in the muon system. Furthermore, this approach allows an improved reconstruction of higher level objects such as jets, tau leptons and missing transverse energy.

2.4.1 Reconstruction of Charged Particle Tracks and Vertices

The information coming from the tracker is the most important input for the particle flow (PF) algorithm, given that on average a large portion of the particle jet energy is carried by charged particles. Most stable particles generated at the interaction point have a rather low transverse momentum p_T. Typically jets with a total p_T below 100 GeV are composed of particles with p_T of the order of a few GeV; on the other hand more energetic jets such as those belonging to heavy exotic particle decays, have constituents with p_T of the order of 10 GeV. Up to the energies of hundreds of GeV, the tracker is able to measure the momenta of the charged particles with a higher resolution than the calorimeters. Additionally, the tracker is able to provide a precise measurement of the direction of these particles at the production vertex, which is difficult to capture by the calorimeters due to the deviation of the trajectories (by the magnetic field) that takes place by the time they arrive at the calorimeters.

There are two constraints to take into consideration for the reconstruction of charged particle tracks. Firstly, as many tracks as possible have to be reconstructed

(high efficiency), since any track that was missed will be reconstructed using calorimeter data only with lower resolution. Secondly, the number of mis-reconstructed (fake) tracks need to be kept to a minimum (low fake rate). To this purpose, an iterative algorithm is adopted. In the first iteration the tracks are seeded (compatible pairs of hits in the detector layers have been detected) and reconstructed with a tight criteria, which results in a lower than desired tracking efficiency (as depicted in Figure 2.11) but a small fake track rate. Most of these tracks are easier to reconstruct because of their higher transverse momentum, smaller impact parameter or having a high number of hits in the tracker that belong to them. In the next iterations track seeding criteria is loosened systematically, while the hits that belong to the previously reconstructed tracks are removed, which reduces the combinatorial complexity at each step and allows to keep the fake rate at an acceptable level.

Primary vertex position is measured as the intersection of tracks by clustering tracks according to their origin vertex and fitting each vertex position corresponding to a cluster. The resolution of the vertex position is estimated as the difference between the positions of two vertices, which are found when all the associated tracks are split into two groups and a vertex position is estimated for each group. With iterative tracking a charged particle with at least three hits on the tracker, transverse momentum p_T higher than 150 MeV and primary vertices within a 50 cm distance from the beam axis can be reconstructed with a fake rate of the order of a percent [47].

2.4.2 Calorimeter Clustering

Calorimeters detect stable particles that deposit their energy in the calorimeter material as they traverse through them. An algorithm that has the calorimeter hits as input should differentiate the energy and direction of charged hadrons from the energy deposits of photons and neutral hadrons. Additionally, the charged hadrons (usually with high p_T) that were missed by the tracker need to be identified. To accomplish these tasks a clustering algorithm is employed for each sub-detector part: the ECAL barrel, the ECAL endcap, the HCAL barrel and the HCAL endcap, separately. In the HCAL HF no clustering is performed, each cell with a hit is registered as a potential cluster. The algorithm follows three steps: initially the energy values

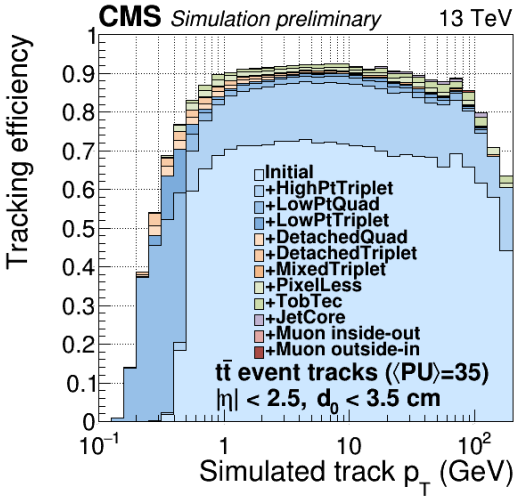


FIGURE 2.11: track reconstruction efficiency per tracking iteration as a function of simulated track transverse momentum p_T for the Phase I tracker (in operation during Run II) [48].

in the calorimeter cells are read out and the local maxima above a certain threshold are recorded as the cluster seeds, which then get aggregated with the adjacent cells that have deposits above a given energy and form *topological clusters* that in the final step are used as seeds for the resulting *particle flow clusters*. Each cluster center is determined in an iterative manner as the calorimeter's cells are read-out.

2.4.3 Reconstruction of physics objects

Reconstruction of muons

In the PF algorithm, the hits in the muon detector systems are put together to form tracks using the Kalman filter technique similar to the inner tracker. For the cases where more stringent criteria are needed, tracks in the muon system are linked to the ones in the inner tracker. The discrimination between the *prompt* muons, which are the products of hadron decays produced in the primary interaction from those that come from secondary decays, is a primary goal. For the so called *isolated* muons tighter criteria on the energy deposits in the calorimeter are applied such as a small distance of the deposit from the muon direction. Additionally, in order to reject the *punch-through hadrons*, very energetic hadrons that manage to reach the muon system, misidentified as muons, several filtering criteria are applied on the muon track quality and the associated energy deposit.

Since the muon system is the farthest away from the interaction region, where muons are the only particles expected to be observed, the density of particle traces are much lower than the other components. However, muon system has the challenge of discrimination of muons associated with the collisions from the muons produced by the cosmic rays that enter the atmosphere and go through the detector.

Reconstruction of electrons and isolated photons

Reconstruction of electrons is based on the combined information from the tracker and the ECAL. In the tracker electrons lose their energy via Brehmsstrahlung and produce photons, while photons convert to electron-positron pairs which in turn radiate Brehmsstrahlung, therefore the applied criteria for electrons and photons are similar. Low energy electrons are reconstructed more accurately in the tracker, whereas ECAL is more reliable for high energy electrons. Several tight criteria on the track quality and energy deposit variables are applied to distinguish between the electrons produced in the primary interaction and those from converted photons or secondary decays. High energy photons are reconstructed using the calorimeter energy deposit information by the detection of an electromagnetic shower compatible with a photon shower, characterized by energy isolation requirements.

Reconstruction of hadrons and non-isolated photons

Once electrons, isolated photons and muons are identified and removed from the PF blocks, the remaining particles to be identified are hadrons. The candidates are charged and neural hadrons and nonisolated photons, e.g. from π_0 decays. The energy deposits in the ECAL and HCAL without any tracks in the tracker indicate presence of photons and neutral hadrons. Within the tracker acceptance ($|\eta| < 2.5$), all these ECAL clusters are turned into photons and all these HCAL clusters are turned into neutral hadrons. The precedence given in the ECAL to photons over neutral hadrons is justified by the observation that, in hadronic jets, 25% of the jet energy is carried by photons, while neutral hadrons leave only 3% of the jet energy in the ECAL [47]. Outside the tracker acceptance ECAL clusters linked to an HCAL cluster are identified as belonging to the same hadron shower, as they leave 25 % of the jet energy in the ECAL.

Reconstruction of jets

Jets can be visualized as cones coming from the interaction region, with an angular area of ΔR. The constituents are hadrons and radiated photons that have the same direction as the initial parton. Jets are reconstructed with the anti-kT algorithm [49], where firstly the distances d_{ij} between entities i and j and distance d_{iB} between entity i and beam B are computed:

$$d_{ij} = min(k_{ti}^2, k_{tj}^2)\frac{\Delta_{ij}}{R^2}, \tag{2.7}$$

$$d_{iB} = k_{ti}^2, \tag{2.8}$$

$$\Delta_{ij}^2 = (\gamma_i - \gamma_j)^2 + (\phi_i - \phi_j)^2, \tag{2.9}$$

where k_{ti}, γ_i and ϕ_i are the transverse momentum, rapidity and azimuth of particle i, respectively. Clustering starts by finding the smallest distance and combining

the entities if the distance is between 2 entities d_{ij} or identifying the the entity i as a jet if the smallest distance is d_{iB}, in which case the entity i is removed from the list of entities. Subsequently, the distances are recalculated and the procedure is repeated until the list is empty.

The particle content of jets is described by the fragmentation functions and depends on the flavour of the parton that initiated the jet, which can only be inferred by the detected decay products of its constituents. For example, characteristically B and D hadrons decay into a large number of charged particles and produce a lepton in their decay chain. Furthermore, the lifetimes of most heavy flavor hadrons are long (1.638 ± 0.004 ps for B^+ [50]) which allow them to move several milimeters away from the primary vertex before decaying and can be identified via displaced tracks and the presence of secondary vertices. They can therefore be exploited when distinguishing between heavy flavor jets from other jets produced by light quarks and gluon hadronization.

Reconstruction of missing transverse momentum

The presence of "invisible" particles that do not interact with the detector material, e.g. neutrinos, can be inferred by the missing transverse momentum, which is defined as:

$$\vec{p}_{T,PF}^{\,miss} = -\sum_{i=1}^{N_{particles}} \vec{p}_{T,i} - \sum_{j=1}^{N_{PFjets}} (\vec{p}_{T,j}^{\,corr} - \vec{p}_{T,j}), \tag{2.10}$$

which includes the jet correction term, it replaces the raw momentum $\vec{p}_{T,j}$ of each PF jet with $\vec{p}_{T,j} > 10$ GeV with its corrected value.

2.5 Detector Simulation

The detector response modelling requires many complicated aspects to be taken into account such as the interaction between the beam particles, the produced particles and their decay products, the detector geometry and material, etc. This is achieved by a simulation, where the affect of the detector layout with the electric and magnetic fields on the particle trajectory is simulated, possible interactions and decay processes are considered during the passage of the particle through the material and the electrical response from detector components is determined. For CMS such a simulation is performed using the GEANT4 toolkit [51]. Additional packages are used for the simulation of the affect of the pile up interactions both for the simulation of the events as well as for the affect on the detector readouts. Such simulations are however, very time consuming. It can take several minutes of CPU time to generate one event and for a realistic modelling of the detector response, a large number of events is required. Alternative simulation packages exist, however the standard simulation package for CMS remains the GEANT4 toolkit due to its high accuracy.

Chapter 3

Machine Learning Techniques

Machine learning (ML) is a branch of Computer Science where computer programs are tasked with identifying patterns in sample data and making decisions or predictions based on what they "learn" without any human intervention during the prediction phase. The term itself was formulated by the IBM researcher Arthur Samuel. In his 1962 essay "Artificial Intelligence: A Frontier of Automation" [52] he described the idea behind the concept as following: "Suppose we arrange for some automatic means of testing the effectiveness of any current weight assignment in terms of actual performance and provide a mechanism for altering the weight assignment so as to maximize the performance. We need not go into details of such a procedure to see it could be made entirely automatic and to see that a machine so programmed would "learn" from its experience". The automatic nature of "assigning a weight", where each assignment has an "actual performance", which is tested by "automatic means" and the test resulting in a change in the weight assignment "to maximize the performance", is the main idea behind most ML algorithms. In this section the main ideas behind the ML techniques that were used within the scope of this work will be described.

3.0.1 Terminology

The modern terminology for an ML algorithm is a *model*, the weights are the model's *parameters* and the measure of performance is called *loss* with a suitable *loss function*, which is used for penalizing errors. Each model has an *architecture*, i.e. a blueprint of a mathematical function to which we pass the input data (*features*) and the model

parameters. The output of the model, the *predictions* are computed based on the independent variables, i.e. features that power the model to predict changes in the dependent variable (the *labels*). *Training/fitting* is the automation process where we update the model parameters at every pass through the input data (an *epoch*). *Hyperparameters* are high-level parameters of the model that are set before any training occurs, control the overall performance and need to be fine-tuned. One of the most important hyperparameter for tuning of such an algorithm is the *learning rate*. If the learning rate is too low, the model will train too slowly, if instead it is too high the learning might not converge to a minimum loss. For decision trees, it is a measure of modification per tree, which determines how fast the model learns. For neural networks, the learning rate is the step size at each iteration while moving toward a minimum of a loss function as described in Section 3.2. Another such hyperparameter for decision trees is the number of trees. If the number of trees is too high, the decision tree will start to *overfit*, where the prediction errors on the training data get minimized but the predictions on new data that the model has not seen, differ from the labels substantially.

The nature of the predictions depend on the learning task and can be *qualitative*, where there is no ordering in the output values and in most cases describe the data itself such as signal vs background, or *quantitative*, where some of the output values are higher than others such as the estimations of a particle mass. The naming convention for predicting qualitative outputs is *classification*, whereas predicting quantitative outputs is called *regression*.

Depending on the existence of the labels the learning tasks are divided in two groups. In *supervised learning* the model "learns by example" by minimizing the losses making use of the available labels, whereas *unsupervised learning* focuses on clustering the data into groups.

3.1 Supervised Learning

Supervised learning can be approximated as "function-fitting" at a first glance to be later expanded into the framework for creating ML models [53].

Regression

For the case of regression we have real valued random variables as features, their probability spaces and a quantitative output. Let $X \in \mathbb{R}^p$ be an input vector and let $Y \in \mathbb{R}$ be an output variable with a joint probability distribution *Pr(X,Y)*. The goal then is to find the function *f(X)* that predicts the Y values given X. We can choose a convenient loss function for this task such as the *squared error loss* : $L(Y, f(X)) = (Y - f(X))^2)$. Then the expected prediction error becomes:

$$E[L] = E[(Y - f(X))^2] = \int |y - f(x)|^2 Pr(dx, dy). \tag{3.1}$$

Since we condition on X, we can factor the joint probability density $Pr(X, Y) = Pr(Y|X)Pr(X)$ and split the integral in 3.1:

$$E[(Y - f(X))^2] = E_X E_{Y|X}[(Y - f(X))^2|X]. \tag{3.2}$$

.

We minimize the expected squared prediction error point wise by finding the values c that minimize the error given X:

$$f(x) = argmin_c E_{Y|X}[(Y - c)^2|X = x]. \tag{3.3}$$

.

The solution for f follows as

$$f(x) = E(Y|X = x), \tag{3.4}$$

and is called the *regression function*. Put another way, when the performance is measured by the squared error loss, the best prediction of Y is the conditional mean of Y given the knowledge of X.

If we want to to expand the idea of function-fitting to the framework of a model, we take the model parameters θ into account. For a linear model $f(x) = x^T\beta$ that would be $\theta = \beta$. Linear basis expansions can be used to approximate the function:

$$f_\theta(x) = \sum h_k(x)\theta_k, \tag{3.5}$$

where $h_k(x)$ are transformations of the input data vector x such as polynomial or trigonometric functions. Transformations such as sigmoid are applied to introduce further non-linearity:

$$h_k(x) = \sigma_k(x) = \frac{1}{(1 + exp(-x^T\beta_k))}, \tag{3.6}$$

Similar to 3.1 we can compute the squared error (sum of squares in this case) to estimate the model parameters θ:

$$\sum_i^N (y_i - f_\theta(x_i)), \tag{3.7}$$

however this would allow us to assume an additive model which can be expressed as $f(X) = \sum_j f_j(X_j)$, and is not the general case. *Maximum likelihood estimation* is the most commonly used estimation method, where we start with the assumption that each data point is generated independently from each other and thus the probability of observing the data is the product of the marginal probabilities. The log-probability for a sample $y_i, i = 1, ..., N$ from a probability distribution $Pr_\theta(y)$ is then:

$$L(\theta) = \sum_i^N log(Pr_\theta(y_i)), \tag{3.8}$$

The most probable model parameters are those that maximize the probability 3.8 of the observed sample.

Classification

For the case of classification we have a qualitative output (*categorical variable*). Let G denote the labels and $\hat{G}(X)$ an estimate of the output classes. The loss function $L(G, (\hat{G}(X)))$ can then be constructed as a matrix with zeroes on the diagonal and non-negative values on the non-diagonal, where $L_{i,j}$ is the penalization of misclassifying a data point of class i to be of class j. Commonly *0-1 loss* function is selected where every misprediction penalizes for a loss of 1 and 0 otherwise. Similar to 3.1 we condition on X and factor the the joint probability density $Pr(G, X) = Pr(X)Pr(G|X)$. The expected prediction error is then:

$$E(L(G, (\hat{G}(X)))) = E_X \sum L(G_k, \hat{G}(X))Pr(G_k|X), \tag{3.9}$$

Following the same procedure as for 3.3 we minimize the error point wise and find the estimate:

$$\hat{G}(X) = argmin_g \sum L(G_k, g)Pr(G_k|X = x), \tag{3.10}$$

For 0-1 loss this becomes:

$$\hat{G}(X) = G_k if Pr(G_k|X = x) = max Pr(g|X = x)), \tag{3.11}$$

This solution is called the *Bayes optimal classifier* and lets us classify the most probable class for new instances based on the conditional discrete probability distribution Pr(G | X).

If we consider a model for the classification task, we need to consider the conditional probability for each class (indexed by the model parameter vector θ) : $p_{k,\theta}(x) = Pr(G = G_k|X = x)$. The log-probability (or *cross entropy*) follows just as with 3.8:

$$L(\theta) = \sum_{i}^{N} log p_{g_i,\theta}(x_i). \tag{3.12}$$

The parameters for the model which best suit the observed data, are estimated by maximizing the cross-entropy function in 3.12.

3.1.1 Decision Trees

A decision tree algorithm is one that asks iterative questions and partitions the available feature subspace into sets of rectangles based on the answers [54]. There are two kinds corresponding to the two supervised learning algorithms.

Regression Trees

Let $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$ denote the input data vector of size p, $i = 1, ..., N$ and Y the continuous output variable for N observations. Suppose we have M partitions of the feature subspace with $R_1, .., R_M$ the corresponding regions. We can estimate the function that maps x_i into y_i as a constant value c_m in each region:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m). \tag{3.13}$$

As before, we can use squared error loss (sum of squares) as the loss function: $\sum(y_i - f(x_i))^2$. It follows that best $\hat{c}_m$ that minimizes the loss is the average of y_i in region R_m. However estimating the best $\hat{c}_m$ in each region is computationally expensive, so we turn to a recursive method, where the split that costs least is chosen. Because sub-spaces can be further split in the same recursive manner, the algorithm is referred as the *greedy algorithm*. The idea is the following: first we look at the half-planes constructed using the feature l for which we split the data and the point z where where subspace is split into two regions:

$$S_1(l, z) = \{X | X_l \leq s\} \text{ and } S_2(l, z) = \{X | X_l > s\}. \tag{3.14}$$

The goal is to find the splitting variable l and the point z that gives us the smallest loss:

$$min_{l,z}[min_{c_1}(\sum_{x_i \in S_1} (y_i - c_1)^2) + min_{c_2}(\sum_{x_i \in S_2} (y_i - c_2)^2)] \tag{3.15}$$

For any splitting variable l and point z, the solution to the equation I of 3.15 is:

$$\hat{c}_1 = average(y_i | x_i \in S_1) \text{ and } \hat{c}_2 = average(y_i | x_i \in S_2) \tag{3.16}$$

By computing the minimum loss for each pair (l,z), the best pair is found and the data is partitioned accordingly. The procedure is then repeated recursively.

One of the main concerns for the overall performance of a learning algorithm is its generalisation power, which can be tested by running unseen data through the data and checking the performance. The failure of predicting correctly on the unseen data due to memorization of the features of input data, is called *overfitting*. This is especially relevant for non-parametric learning algorithms such as decision trees. A large tree with many branches would be prone to this issue and would start memorizing the data. A small tree on the other hand might not be able to capture the underlying relationship between the input data and the output (*underfitting*). Therefore the size of the tree is one of the important hyperparameters of the model that needs to be tuned. One of the methods for tuning we employ is called *cost-complexity pruning*, where we prune a large tree T_0 to a smaller tree T by closing off some of its internal (non-final) nodes. Let n be the internal node index, R_m be the region of partition m, then with the following terms:

$$N_m = \text{ number of } \{x_i \in R_m\}, \tag{3.17}$$

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i, \tag{3.18}$$

$$\hat{Q}_m = \frac{1}{N_m} \sum_{x_i in R_m} (y_i - \hat{c}_m)^2, \tag{3.19}$$

we define the cost complexity criterion:

$$\hat{C}_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|. \tag{3.20}$$

The idea is then to find for each α the subtree $T_\alpha \subseteq T_0$ so that the $C_\alpha(T)$ is minimized.

Classification Trees

For the classification task the procedure is analogous, however the impurity measure $Q_m(T)$ has to be modified to allow discrete target values. For a tree node m,

representing the region R_m with N_m observations we define the proportion of class c as $\hat{p}_{mc} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = c)$. One of the following node impurity measure can be then used for the classification trees:

$$\text{Cross-entropy: } -\sum_{c=1}^{C} \hat{p}_{mc} log \hat{p}_{mc} \tag{3.21}$$

$$\text{Gini-index: } -\sum_{c=1}^{C} \hat{p}_{mc}(1 - \hat{p}_{mc}) \tag{3.22}$$

Gradient Descent

Gradient descent is an iterative optimization algorithm which finds a local minimum of a differentiable function, in this case, the loss function $L(f) = \sum_{i=1^N} L(y_i, f(x_i))$. Numerical optimization optimization procedures solve the minimization problem $\hat{f} = argmin_f L(f)$ as a sum of component vectors

$$f_m = \sum_{m=0}^{m} h_m, h_m \in \mathbb{R}^N, \tag{3.23}$$

where f_0 is an initial value and each successive value is computed based on the previous value and h_m is referred as *step*. For gradient descent algorithm the step $h_m = -\rho_m g_m$is proportional to the gradient of the loss function at f_{m-1} :

$$f_m = f_{m-1} - \rho_m g_m, \tag{3.24}$$

where ρ_m is a scalar and the gradient is defined as:

$$g_{im} = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i} \right]_{f(x_i)=f_{m-1}(x_i)} \tag{3.25}$$

3.2 Neural Networks

A neural network is a learning algorithm [53], a model with parameters θ that needs to be optimized for the available training data. For a K-class classification (or regression) there are K targets with a k-th unit of the neural network modeling the probability of class k. Newly derived features Z_m (*hidden units*) are generated from linear combinations of the inputs and then the targets Y_k is modeled as a function of linear combinations of Z_m:

$$\begin{aligned} Z_m &= \sigma(\alpha_{0m} + \alpha_m^T X), \\ T_k &= \beta_{0k} + \beta_k^T Z, \\ f_k(X) &= g_k(T), \end{aligned} \tag{3.26}$$

where $\sigma(x) = 1/(1+e^{-x})$ is the activation function. We can define the error function as a function of the parameters of the model θ consisting of $\alpha_0 m, \alpha_m$ and $\beta_0 k, \beta_k$:

$$\text{for regression: } R(\theta) = \sum_k \sum_i (y_{ik} - f_k(x_i))^2 = \sum_i R_i, \tag{3.27}$$

where $R(\theta)$ is minimized typically by the gradient descent, where the gradient can be defined using the chain rule:

$$\begin{aligned} &\text{with } z_{mi} = \sigma(\alpha_{0m} + \alpha_m^T x_i) : \\ &\frac{\partial R_i}{\partial \beta_{km}} = -2(y_{ik} - f_k(x_i)) g_k'(\beta_k^T z_i) z_{mi}, \\ &\frac{\partial R_i}{\partial \alpha_{ml}} = -\sum_k 2(y_{ik} - f_k(x_i)) g_k'(\beta_k^T z_i) \beta_{km} \sigma'(\alpha_m^T x_i) x_{il} \end{aligned} \tag{3.28}$$

Then the gradient descent update at the (r+1)-th iteration follows:

$$\begin{aligned} \beta_{km}^{r+1} &= \beta_{km}^r - \gamma_r \sum_i \frac{\partial R_i}{\partial \beta_{km}^r}, \\ \alpha_{ml}^{r+1} &= \alpha_{ml}^r - \gamma_r \sum_i \frac{\partial R_i}{\partial \alpha_{ml}^r}, \end{aligned} \tag{3.29}$$

where γ_r is the *learning rate*. With the error terms of the current model at the output unit δ_{ki} and at hidden units s_{mi} we can rewrite the derivatives of the error function:

$$\begin{aligned} \frac{\partial R_i}{\partial \beta_{km}} &= \delta_{ki} z_{mi} \\ \frac{\partial R_i}{\partial \alpha_{ml}} &= _{mi}\ x_{il} \end{aligned} \tag{3.30}$$

From their definitions we can relate the two error terms:

$$s_{mi} = \sigma'(\alpha_m^T x_i) \sum_k \beta_{km} \delta_{ki}. \tag{3.31}$$

The equations 3.31 are called the *back-propagation equations*. The gradient descent works as a two-pass algorithm, where in the *forward pass* the model weights are fixed and the predictions $\hat{f}_k(x_i)$ are calculated with the equation 3.31. In the *backward pass*, the errors δ_{ki} are computed and then back-propagated with the equation 3.31 to give the errors at the hidden units s_{mi}. These errors are then used to compute the updates in 3.29.

3.3 Anomaly Detection in Copula Space

3.3.1 Problem Statement

We consider a set of N data examples $x \in \mathcal{S} \subseteq \mathbf{R}^{\mathcal{D}}$ sampled from an unknown multivariate density function $p(x)$. In general, $p(x)$ can be written as the sum of a background component $p_b(x)$ and a possible signal contamination $p_s(x)$,

$$p(x) = (1 - f_s)p_b(x) + f_s p_s(x) \tag{3.32}$$

where f_s is the signal fraction. An anomaly detection problem may be defined as one of finding a localized region of the feature space $\mathcal{S}$ that contains a density of data examples significantly higher than that of its surroundings, as defined by some suitable metric. This problem may be cast as a semi-supervised or a unsupervised one, depending on whether the (by definition) non-anomalous density of the background component is assumed known or not, or even if we instead assume the signal component known, as is of interest for the $B_s \to \tau\tau$ search. In the case when the signal is unknown, a central issue is how to retain sensitivity to a wide variety of anomalous contaminations, which may produce distortions of the density in a subset of the $\mathcal{D}$ features; in the case when it is the background which is unknown, the method to extrapolate its density to a region of interest becomes the focus of attention.

3.3.2 The idea of `RanBox`

The unsupervised version of the problem [55] offers the benefit of avoidance of any model-related uncertainties and hence is relevant for new physics signals that characteristically produce localized, compact variations in the overall density of the feature space.

The algorithm searches the feature space $\mathcal{S}$ by considering a "box", *i.e.*, a multidimensional interval constructed in a subspace of $\mathcal{S}$. The random nature of the box lays not only in the endpoints x_i^{min}, x_i^{max} of its intervals in each marginal, $x_i \in [x_i^{min}, x_i^{max}]$, but also in the involved subspace $\mathcal{S}' \subseteq \mathbf{R}^{\mathcal{D}'}$ of $\mathcal{S}$ described by a subset of the x_i. Alternatively, one may think of the box as having restricted intervals in only a subset $\mathcal{D} - \mathcal{D}'$ of the dimensions of $\mathcal{S}$. If we consider for the time being the case

$f_s = 0$ and N data points in $\mathcal{S}$ sampled from a multi-dimensional uniform density $p_b(x) = \mathcal{U}(x)$, such a box will contain a predictable fraction of the total data: given the box volume V_{box} and the total volume V of the feature space $\mathcal{S}$, the expectation value of the number of events in the box is $N_{exp} = NV_{box}/V$. Conversely, if $f_s > 0$, the observed number of events captured within the box boundaries N_{obs} may yield an estimate of the density of the total sampling distribution in the corresponding region of $\mathcal{S}$, contributed by both $p_b(x)$ and $p_s(x)$:

$$\hat{p}(x) = \frac{N_{obs}}{VN_{exp}} = \frac{N_{obs}}{NV_{box}}. \tag{3.33}$$

The above estimate may be used to construct a test statistic sensitive to an anomalous local overdensity of the data; *e.g.* one may simply define the test statistic to equal the estimated excess of events in the box, $N_{obs} - N_{exp}$, or a significance measure of its non-null value. The maximization of such a test statistic will be appropriate for searches of anomalies that preferentially populate well-confined regions of the feature space, such as those of interest in collider searches for new physics, but also relevant to other branches of science as, *e.g.* astrophysical observations, or industrial applications such as process control, fraud detection, or spam filtering. Conversely, we expect little sensitivity to multi-modal signals, and (by construction) no sensitivity to broad deformations of a nearly-uniform background distribution $p_b(x)$. The locality of the signal to be detected, however, is the only assumption we allow ourselves to take in the construction of our anomaly detection procedure. The assumption of uniformity on which the estimate in Eq. 3.33 is based, can be loosened, if we work in the copula space[1] , as discussed more in detail in Sec. 3.4.

[1] Let $(X_1, X_2, ..., X_d)$ be a random vector with continuous marginals, i.e. the cumulative distribution functions $F_i(x) = \Pr[X_i \leq x]$ are continuous. By applying the probability integral transform to each component, the random vector $(U_1, U_2, ..., U_d) = (F_1(X_1), F_2(X_2), \ldots, F_d(X_d))$ has marginals that are uniformly distributed on the interval [0,1]. The copula space is then defined as the joint cumulative distribution of $C(u_1, u_2, ..., u_d) = \Pr[U_1 \leq u_1, U_2 \leq u_2, ..., U_d \leq u_d]$.

3.4 Algorithm Description

3.4.1 Starting considerations

The multitude of subnuclear particles resulting from proton-proton collisions recorded by LHC experiments, which we take as our target application in the construction of the algorithm, yield tens of millions of electronic signals in the detectors. This large body of information is summarized by a process called "event reconstruction" through the extraction of several tens of high-level features that describe the measurement of energy and direction of all observed particles (*e.g.* energetic electrons or muons) or sets of particles (hadronic jets) [2]. Even if we focus on specific interesting subsets of the available data, any energy-related feature of the observed particles will show a highly dis-uniform distribution, with a peak at low values and long tails extending to higher energy (see, *e.g.* Fig. 3.1). The variation in density between those peaks and tails may amount to orders of magnitude, and is due to the corresponding large variation in the probability that the collision is originated by quarks or gluons carrying a low or a high fraction of their parent's total momentum.

Because of the above, it seems natural to proceed by first pre-processing the data with an integral transform of all the features, such that each marginal becomes uniform by construction. The algorithm will then work in the copula space, examining the data structure with a metric unaffected, at least to first order, by the original strong density variations in the feature space.

3.4.2 Data preprocessing

The probability integral transform of a function $f(x)$ is defined by setting

$$F(x) = \int_{-\infty}^{x} f(t)dt, \tag{3.34}$$

which is such that $y = F(x)$ is uniform in $[0, 1]$:

[2] In HEP it is thus customary to call *events* the observed data examples, and we will stick to that convention in this work.

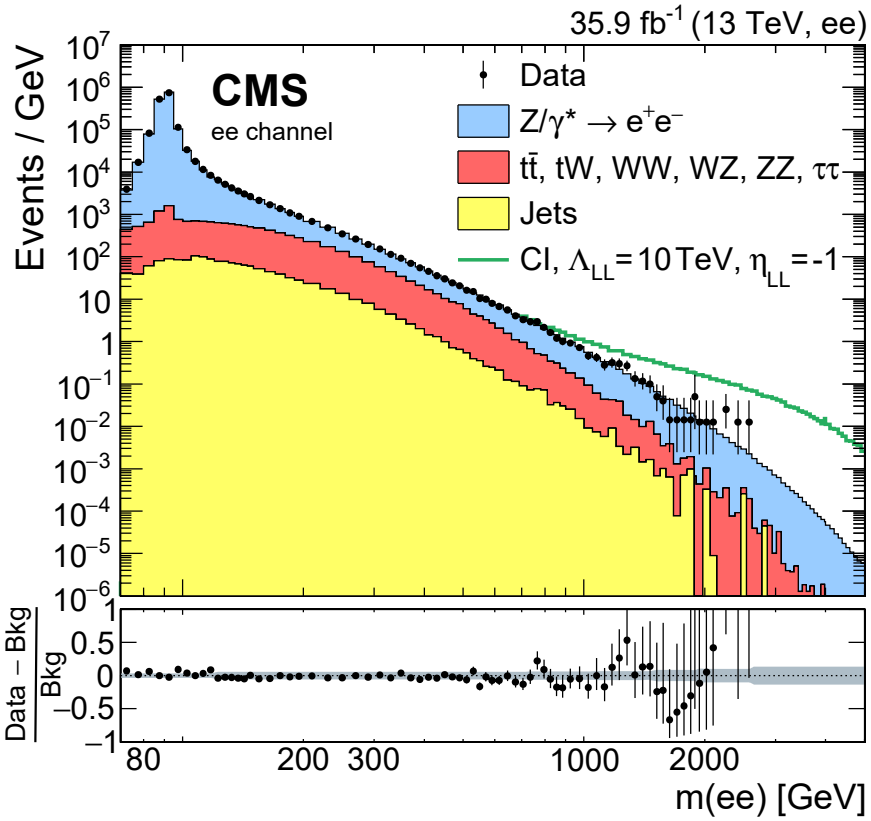


FIGURE 3.1: Distribution of the invariant mass of candidate electron-positron pairs observed by the CMS experiment in 36 fb^{-1} of Run 2 LHC collisions [56]. The data show a variation in density by several orders of magnitude as a function of mass. The cited reference reports on searches for a new physics contribution involving contact interactions, which could contribute to the distribution at its high-end tail (green curve).

$$\begin{aligned} F_y(y) &= P(Y \leq y) \\ &= P(F_X(X) \leq y) \\ &= P(X \leq F_x^{-1}(y)) \\ &= F_X(F_X^{-1}(y)) = y. \end{aligned} \tag{3.35}$$

Once each of the variables of the feature space x_i is transformed as above into the corresponding one in the set y_i, information once contained in the interdependence of the x_i is retained in the copula, which is the joint distribution function of variables with uniform marginals (Sklar's theorem) [57]. The advantage of the transformation is evident: a search for overdensities in the space spanned by y_i will not be spoiled by

uneven marginals, and will correctly concentrate on the regions of space which are dense because of interdependence of the features. An additional bonus of working with the y_i variable basis is that the feature space is now a unit hypercube, with volume $V = 1$.

3.4.3 Dimensionality reduction

The dreaded "curse of dimensionality" [58] affects any search in high-dimensional spaces populated by sparse data. In the typical applications considered in this work, the total data size N lays in the few thousands to few hundreds of thousands range; consequently, an investigation of subspaces $\mathcal{S}'$ of the feature space $\mathcal{S}$ quickly becomes meaningless as their dimensionality grows larger than about $\mathcal{D}' = 12 - 15$, when Poisson fluctuations prevent any reasonable multi-dimensional density estimate.

An additional optional preprocessing step, which may prove useful to reduce the dimensionality in cases when $\mathcal{D}$ is larger than a few tens, is the application of Principal Component Analysis (PCA) [59] to the feature space. PCA essentially consists in fitting a hyper-ellipsoid to the data, and remapping the feature space in a space spanned by the principal axes of the ellipsoid. One may then use the principal components, which are those on which the data exhibit the largest variance, and ignore the last few in the ordered list of components, which are likely to contain the least information. PCA can be useful for `RanBox` in cases when the search for subspaces of limited dimensionality $\mathcal{D}'$ of the feature space proves impractical because of the large binomial coefficient $\binom{\mathcal{D}}{\mathcal{D}'}$, which makes the exploration of a meaningful fraction of the possible $\mathcal{D}'$-dimensional subspaces too CPU-intensive. However, in our investigations we have found that PCA is generally liable to reduce the power of the search for overdense regions of the feature space when the data are composed of a large background component and a small signal contamination to which we wish to be sensitive. The typical reason of this effect is connected with the fact that a variable which exhibits little variance on the majority of the data, and is thus discarded by PCA, may still be very distinctive for a small signal.

A viable alternative to reduce the dimensionality of the problem, which may facilitate the identification of small signals, is to exploit the correlation matrix of

the features, by removing features which add little information. This is an attractive option when searching for small anomalous components in a background-rich dataset: by identifying and removing variables that are highly correlated with others on the majority component of the data, we reduce the possibility that such correlations affect negatively the chance of the algorithm to identify localized overdensities genuinely due to a clustering of multiple distinguishing features of a minority component. As a telling example, if in a $\mathcal{D} = 30$-dimensional feature space one of the variables were identically repeated 10 times, and `RanBox` performed a search in $\mathcal{D}' = 10$-dimensional subspaces, the algorithm would be very likely to end up focusing on the same narrow interval (any one would do) of each of those features: *e.g.* a 10-dimensional box of width 0.1 in each of the correlated features would have a volume of 10^{-10}; if there were $N = 10,000$ events in the space, such a box would be predicted to contain $N_{exp} = 10^{-6}$ events, while it would in fact contain exactly 1000 events!

Our correlated variable removal (CVR) procedure, which performs the identification of variables to be discarded, works as follows. We first compute the correlation coefficients ρ_{ij} among all pairs of variables ij, and order them in a list by decreasing absolute value $|\rho_{ij}|$. Then we choose the number of variables to be removed N_{void}; in order to identify these we consider that if the k-th variable is removed, all correlation coefficients that include k as one of the two indices will become irrelevant. We thus find the combination of N_{void} variables which, when removed, minimizes the value of the highest surviving correlation coefficient. A graphical example of the technique is shown in Fig. 3.2.

3.4.4 Choices of a test statistic for the unsupervised learning task

We consider two estimates of the expected number of events contained in a multi-dimensional region of the unit hypercube resulting from the standardization procedure, both corresponding to a binomial ratio. The first one is simply

$$N_{exp,V} = NV_{box}. \tag{3.36}$$

As the total copula space volume is $V = 1$, the above estimate is only driven by

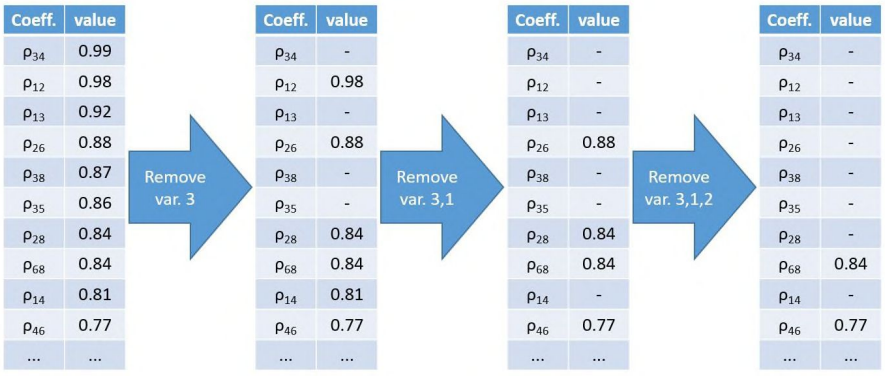


FIGURE 3.2: Graphical description of the CVR procedure available in the preprocessing stage of `RanBox`. The ordered list of absolute values of correlation coefficients among the variables defining the $\mathcal{D}-$dimensional feature space is scanned by searching for all possible combinations of N_{void} variables which, once removed, minimize the largest surviving correlation coefficient. In the figure, for $N_{void} = 3$ the removal of variables 3, 1, 2 (shown in succession for clarity) reduces the highest surviving correlation most effectively.

the extension of the box volume V_{box}. The expectation results from assuming that the data distribute in the feature space with a constant density, and is useful in cases when $p_b(x)$ contains little structure in its copula, as departures from that assumption can then easily be associated with anomalous contaminations. This measure is the default one for the studies of algorithmic performance presented in Sec. 3.5, which are performed on synthetic datasets where the assumption above is identically true in the limit $f_s = 0$.

A second estimate, affected by higher statistical uncertainty than the former but conversely much less affected by a non-uniform density $p_b(x)$ in the copula space, may be obtained by defining a sidebands (SB) region that surrounds the search box (see Fig. 3.3). In this case, no reliance is made on overall constancy of the density for non-anomalous events, and the estimate leverages the density of data in the immediate neighborhood of the search box. If $[x^i_{min}, x^i_{max}], i = 1....\mathcal{D}'$ are the boundaries of the search box, the SB region is defined by the following relations:

$$\delta_i = 0.5(x^i_{max} - x^i_{min})(2^{1/\mathcal{D}'} - 1), \tag{3.37}$$

$$x^i_{min,SB} = \max(0, x^i_{min} - \delta_i), \tag{3.38}$$

$$x^i_{max,SB} = \min(1, x^i_{max} + \delta_i), \tag{3.39}$$

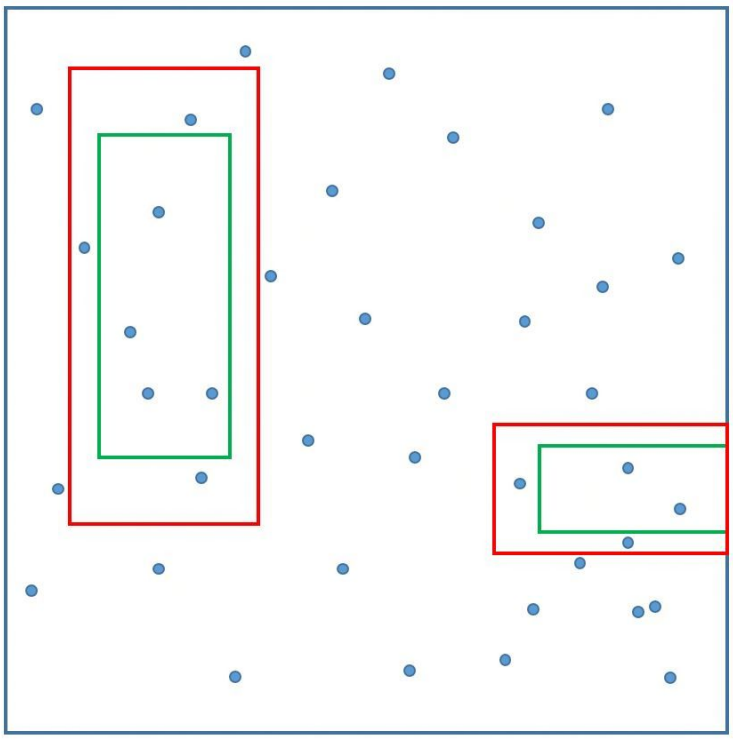


FIGURE 3.3: Representation of the sideband method for box density estimation. Two possible search boxes in a 2-dimensional space are shown in green; the relative sidebands, constructed according to the recipe of Eq. 3.39, are the regions between the red and the green rectangles. The sideband region on the lower right can only extend horizontally to the left, and the area it defines is thus smaller than that of the related search box.

with $x^i \notin [x^i_{min}, x^i_{max}]$ for at least one i, *i.e.* the SB volume does not include the search box volume. The SB then has a volume at most as big as the search box volume; it is in general smaller than that, as some of the intervals cannot extend on each side of the search box by the required length δ_i, due to the hard boundaries at 0 and 1 (see again Fig. 3.3). If one observes a number of events N_{out} in the sideband region, the expectation value of the number of events in the search box in the assumption of uniformity may be written as

$$N_{exp,\tau} = \tau N_{out}, \tag{3.40}$$

where

$$\tau = \frac{V_{box}}{V_{SB}} \tag{3.41}$$

is defined by the volumes of sideband region V_{SB} and search box V_{box}. A slight modification of the recipe for the expectation value above, which we have found to be effective, is operated when the number of observed sideband events N_{out} is zero. In that case, which is frequent for large dimensionality searches and small statistics of the data sample, it is useful to reset N_{exp} to the full-volume prediction, Eq. 3.36. We stick to this recipe in our applications of the sideband method in the studies described in this work.

To formulate the problem in its generality through the above definition of the extrapolation variable τ, we observe that the full-volume estimate in Eq. 3.36 corresponds to setting

$$\tau = \frac{V_{box}}{1 - V_{box}} \tag{3.42}$$

and $N_{out} = N - N_{in}$. In either case a likelihood-ratio-based test statistic may now be defined as follows:

$$Z_{PL} = \sqrt{2}\left\{ N_{in} \ln\left[(1+\tau)\left(\frac{N_{in}}{N_{in}+N_{out}}\right)\right] + N_{out} \ln\left[\frac{1+\tau}{\tau}\left(\frac{N_{out}}{N_{in}+N_{out}}\right)\right] \right\}^{0.5} \tag{3.43}$$

The above defined function has been shown [60] to be a good approximation of the Z-score corresponding to the binomial probability of observing an excess of events $N_{in} - N_{exp,\tau}$ in the box. It is to be noted, however, that Z_{PL} cannot be considered a genuine signal significance, because in real applications "non-anomalous" data contain structure in the copula due to interdependence of their features; as a result, the Z_{PL} test statistic for the null hypothesis has fatter tails at positive values than a Normal distribution. In addition, as discussed *infra* in more detail, `RanBox` effectively operates multiple testing on the dataset, hence Z_{PL} cannot be used as a significance measure in the absence of a Bonferroni or similar correction [61]. Despite the above caveats, the fact that Z_{PL} is a principled proxy to the significance of an excess in a Binomial counting experiment makes it a sound choice for a test statistic when the focus is the search for significant, anomalous signals.

We have observed that the Z_{PL} test statistic is especially useful when anomalies

are sought which may interest wide volumes of the feature space, with N_{exp} correspondingly being not very small —typically in the range of several tens to a hundred of events. Conversely, when the expectation N_{exp} in the overdense region amounts to only a few events or less, an attractive alternative is to use the function R_{reg} defined as

$$R_{reg} = \frac{N_{in}}{N_{exp} + N_{reg}}, \tag{3.44}$$

with, *e.g.*, the regularization term set to $N_{reg} = 1$. The maximization [3] of R_{reg} may identify more effectively small anomalies well confined in the search volume, in cases when the copula space of non-anomalous events has a rich structure, capable of producing high values of Z_{PL} in regions of large volume and thus diverting the algorithm's attention from small, well-confined anomalies.

3.4.5 Box seeding

The search for the most overdense multi-dimensional interval in a feature space populated by sparse data points is complicated by the presence of a large number of local extrema; hence, a careful initialization of the box location and dimensions may significantly improve the performance of the algorithm. Although we tried several recipes for this task, here we only describe three of them, which we found the most suitable for our applications.

The baseline method, "Algorithm 0", consists in a fixed initialization of the box to a multi-dimensional interval of total volume V_{box}, set to equal a given fraction of the unit volume of the full feature space hypercube. The box, which lives in a $\mathcal{D}'$-dimensional subspace of the copula, is constructed by defining intervals x^i_{min}, x^i_{max} (with $i = 1, ..., \mathcal{D}'$) as follows:

$$\Delta = \frac{1 - V_{box}^{1/\mathcal{D}'}}{2} \tag{3.45}$$

$$x^i_{min} = \Delta \tag{3.46}$$

$$x^i_{max} = 1 - \Delta \tag{3.47}$$

[3] In this work we stick to the setting $N_{reg} = 1$, and consequently address the test statistic as R_1.

An optimization of the initial value of V_{box} is of course impossible in a unsupervised search, where neither non-anomalous or anomalous data have a specified density. However, our tests suggest that setting $V_{box} = 0.1$ is a reasonable choice when, as is the case in several of our considered applications, $\mathcal{D}'$ lays in the 6-10 dimensions range. *E.g.*, with $\mathcal{D}' = 6$ one obtains starting intervals equal to $[0.16, 0.84]$, and with $\mathcal{D}' = 10$ intervals equal to $[0.10, 0.90]$. Note that this corresponds to a relatively large box, in terms of its extension along each marginal. When combined with a search algorithm that considers initial expansions or shrinkages in each of the box dimensions by amounts sufficient to extend all the way to the unit hypercube boundaries, the above initialization ensures that no overdensity laying close to the boundary of a coordinate will be overlooked by the search algorithm taking a step in the wrong direction at the start of the search.

The second method, "Algorithm 1", is instead based on clustering the data based on a specialized Nearest-Neighbour (NN) search. First, the nearest neighbour j is found for every event i in the data, by using as a distance the following function:

$$d_{ij} = \prod_{k=1}^{\mathcal{D}'/2} |x^i_{o_k(ij)} - x^j_{o_k(ij)}| \tag{3.48}$$

where $o_k(ij)$ are the $\mathcal{D}'/2$ indices identifying the spatial coordinates for which the intervals $|x^i - x^j|$ are the smallest. In other words, the map d_{ij} determines the minimum volume of a $\mathcal{D}'/2$-dimensional box that includes events i and j. Once d_{ij} is defined for all i and j, one may compute for every event i the number of neighbouring events $j = j_1 ... j_{N_{cl}}$ that have i as their closest event according to that metric. The event i_{maxNN} with the maximum number $N_{cl,maxNN}$ of such neighbours now allows to identify all $N_{2^{nd}\,order}$ events which have any of the $N_{cl,maxNN}$ events as their own nearest neighbours. The box can finally be initialized as the smallest $\mathcal{D}'$-dimensional interval that includes all the $N_{2^{nd}\,order}$ neighbours. A graphical description of the algorithm is provided in Fig. 3.4.

A third initialization method, "Algorithm 2", uses instead a kernel estimation of the density for the identification of starting box boundaries. The density is evaluated at the position of each of the N events as the sum of N $\mathcal{D}$-dimensional Gaussian distributions centered at the location of every event in the sample, and with equal

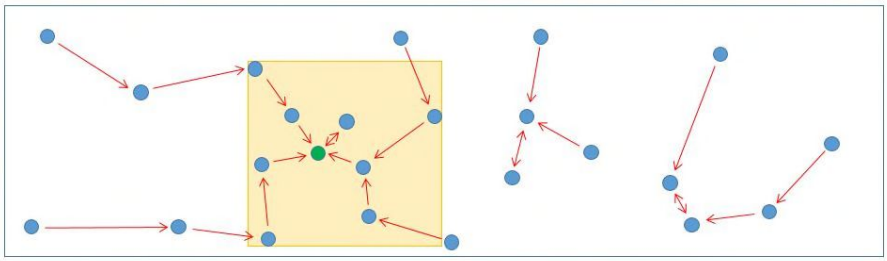


FIGURE 3.4: Graphical description of the clustering algorithm used for box initialization with Algorithm 1. Blue points indicate the position of events in the two shown variables of the feature space. Arrows pointing from an event to another indicate the location of the closest neighbour of the event originating the arrow (according to a metric described in the text). The green point is the closest to four others, and it provides the seed of the box: the collection of all events which point to those four events define the boundaries of the box.

diagonal covariance matrices $C = k^2 I_D$, with I_D the $\mathcal{D}$-dimensional identity matrix and k a tunable parameter which must be chosen according to the total dataset size and the dimensionality of the $\mathcal{D}'$ subspaces scanned by `RanBox`; its default value, used in the applications described in this work, is $k = 0.2$. Once the point of highest density x_{HD} is identified, the box is initialized as the multi-dimensional interval whose extension in each coordinate x is

$$[\max(x_{HD} - \delta_2, 0.), \min(x_{HD} + \delta_2, 1.)], \tag{3.49}$$

with the provision that if x_{HD} is at less than δ_2 distance from the boundary at 0 (1), the interval defaults to $[0., 2\delta_2]$ or $[1. - 2\delta_2, 1.]$, respectively. The default value of δ_2 is 0.2; *e.g.* this corresponds, for a 10-dimensional subspace search, to initial boxes of volume equal to or smaller than 0.0001: the expected number of events within a 10,000-event sample contained in a random box of that volume is 1.0, which is a suitable starting point for the background expectation in the test statistic maximization. Given that the initialization provided by Algorithm 2 offers a good candidate for an overdense region, the focusing on a small initial region of feature space has been observed to be effective in the tested applications of our interest: those are in fact cases when a small, overdense region exists in the first place.

3.4.6 Maximization of the test statistic

A search for the multi-dimensional interval providing the highest value of the chosen test statistic (either Z_{PL} or R_{reg} as defined in Sec. 3.4.4 above) in a $\mathcal{D}'$-dimensional subspace of the feature space can be performed as follows.

Step 1: The initialization of the box is performed with the algorithm of choice. A set of step parameters are set to the starting value $\lambda_i = 0.5\,(i = 1...\mathcal{D}')$. A loop counter N_{GD} is set to zero.

Step 2: Seven possible modifications are considered for each of the $\mathcal{D}'$ intervals defining the box:

$(x^i_{min})'$	$(x^i_{max})'$
$\max(x^i_{min} - \lambda_i, 0)$	x^i_{max}
$\min(x^i_{min} + \lambda_i, x^i_{max} - \epsilon)$	x^i_{max}
x^i_{min}	$\max(x^i_{max} - \lambda_i, x^i_{min} + \epsilon)$
x^i_{min}	$\min(x^i_{max} + \lambda_i, 1)$
$\max(x^i_{min} - \lambda_i, 0)$	$\max(x^i_{max} - \lambda_i, \epsilon)$
$\min(x^i_{min} + \lambda_i, 1 - \epsilon)$	$\min(x^i_{max} + \lambda_i, 1)$
$r^i_{min} = \min(r_1, r_2)$	$r^i_{max} = \max(r_1, r_2)$

where ϵ is a parameter determining the coarseness of the algorithmic scan in the feature space, fixed in applications described in this work to $\epsilon = 0.01$. In the last line the values r_1, r_2 determining a "random jump" in the i-th interval are random numbers sampled from a uniform distribution in $[0, 1]$. The values of $(x^i_{min})'$ and $(x^i_{max})'$ defined above are rounded off to two decimal places in all cases. For each of these $7\mathcal{D}'$ variations, an associated SB region is defined by the recipe described *supra*; this determines the numbers N_{in} and N_{out} and consequently the $7\mathcal{D}'$ values of the test statistic of choice.

Step 3: If the highest among the $7\mathcal{D}'$ values of the test statistic corresponding to the tentative box modifications is higher than the current maximum value, the box is modified to the corresponding new multi-dimensional interval, and all λ_i values for the coordinates not affected by the change are reduced as follows:

$$\lambda_i \to \max\left(f\lambda_i, \epsilon\right) \tag{3.50}$$

where the factor f is set to 0.9. In addition, if the box modification is chosen based on one of the $\mathcal{D}'$ random intervals $[r^i_{min}, r^i_{max}]$, a counter j_i is incremented by one; once a j_i reaches a maximum value (10 by default), no more random jumps are allowed for the intervals in variable i. This recipe allows to control the convergence of the algorithm as well as the trade-off between its CPU consumption and its freedom in exploring new box configurations in the considered feature space dimensions.

If, instead, the current value of the test statistic is higher than all of the $7\mathcal{D}'$ new values, no modifications to the box boundaries are applied, and λ_i values are reduced as in Eq. 3.50.

Step 4: The loop counter N_{GD} is incremented by one. If N_{GD} reaches a limiting value (set to 100 by default) the algorithm stops; the algorithm also stops if all values λ_i have reached the value ϵ. Otherwise, steps 2, 3, and 4 above are repeated.

Despite its simplicity, the procedure described above typically converges in 30 to 50 iterations for $\mathcal{D}' = 6 - 10$, which are typical values for the considered applications of fixed-subspace searches.

3.5 Performance studies with synthetic data

3.5.1 Event generation

A synthetic dataset sampled from a multi-dimensional Uniform distribution $p_b(x) = \mathcal{U}(x)$, with $x \in [0,1]^{\mathcal{D}}$, may be generated by repeated calls to the TRandom3→Uniform() routine [4] of the ROOT package [62], which we employ in our `C++` implementations of `RanBox`. Such a dataset may be considered the ideal background for an anomaly search: by lacking any internal structure in the copula, it constitutes a best-case scenario for performance evaluations of the algorithm in a controlled setting. The unknown signal may instead be generated by drawing samples from a multi-dimensional Gaussian distribution in a subset $x_g, g = 1, ..., N_g$ of the features, $x_g \in \mathbf{R}^{N_g}$, and the remaining ones $x_u, u = N_g + 1, ..., \mathcal{D}$ from a uniform density. While Gaussians have support on the real axis, the generation ensures that the drawn features are also contained in the $[0,1]$ interval, as detailed below.

[4] The random generation is based on the Mersenne primes, and has a periodicity of about 10^{6000}.

We define the following default set of parameters:

- (for background) $x_i = \mathcal{U}(0,1)$;
- (for signal) $x_u = \mathcal{U}(0,1)$;
- sigma $\sigma_{gg} = \mathcal{U}(0.01, 0.1)$;
- mean $\mu_g = \mathcal{U}(3\sigma_{gg}, 1 - 3\sigma_{gg})$;
- $r_{gh} = \mathcal{U}(-1,1)$ (with $g,h \in \{1,...,N_g\}, g \neq h$).

A random choice of σ_{gg} and r_{gh} values as defined above will not in general generate a positive-definite covariance matrix C with variances σ_{gg}^2 and $\sigma_{gh}^2 = r_{gh}\sigma_{gg}\sigma_{hh}$; hence the procedure of generating C is repeated until a Cholesky-Banachiewicz (CB) decomposition $LL^T = C$ into a lower-triangular matrix L [63] is found, which guarantees the positive-definite nature of C. Once successful, the CB decomposition allows to easily draw samples from the multi-dimensional Gaussian distribution by posing, for every g,

- (for signal) $x_g = \mu_g + \sum_{h=1..N_g} L_{gh} n_h$

with n_h sampled from a Normal distribution. During event generation, if a coordinate sampled from the multivariate Gaussian exceeds the range $[0,1]$, it is simply resampled. This truncation has the effect that Gaussians with μ_g values close to the boundaries have an up to twice higher local density than Gaussians closer the center of the $[0,1]$ interval. For this reason, in most tests we limit μ_g values to the range stated above, except when we explicitly study the performance at the edges (see *infra*). Although the background is already generated with flat marginals, after the inclusion of signal we of course re-standardize the dataset by using Eq. 3.35.

When performing power tests of the algorithm, we avoid the random effect of varying σ parameters, and use reference samples with a more narrowly defined signal component, by fixing all Gaussian sigmas to $\sigma_{gg} = 0.05$. In this case correlation coefficients r_{gh} are chosen at random within the discrete set $\{-max(r_{gh}), 0., max(r_{gh})\}$ by posing $max(r_{gh}) = 0.2$, and we allow means μ_g to vary at random in their default range, $[0.15, 0.85]$. The different signals that correspond to varied means and correlations have equal chance of being identified by the algorithm. For example, Fig. 3.5

shows average Z_{PL} values from runs of the algorithm with the following choice of parameters:

- $N_b = 4950$ background events
- $N_s = 50$ signal events
- $\mathcal{D} = 20$ active dimensions of feature space
- $N_G = 6$ Gaussian features in signal component
- $\mathcal{D}' = 6$ dimensions for box definition
- $N_{trials} = 1$ subspace sampled per dataset, of features coincident with the N_G in which signal component has a Gaussian distribution.
- $N_{rep} = 50$ datasets generated and searched
- Algorithm 0 (random box initialization) and 2 (kernel density) used
- No dimensionality reduction (PCA or correlated features removal) performed

By only considering, through the above choices, the subspace which yields the highest probability of locating a signal-rich box, we reduce the effect of randomness and allow for a more precise study of the impact of the tested parameters. In Fig. 3.5 the values of the test statistic Z_{PL} appear stable as a function of the sampled ranges $max(r_{gh})$ and $\Delta\mu_g = \mu_g^{max} - \mu_g^{min}$, indicating that the search algorithm is capable of locating overdensities regardless of their position in the space [5], and that the correlation between Gaussian-distributed variables does not affect the chance of identifying overdense multi-dimensional intervals. Similar results are obtained by initializing the box dimension with Algorithm 1 (kNN-seeded clustering), and/or by using R_1 as a test statistic.

[5] The observed slightly lower performance of searches initialized by Algorithm 0 for $\Delta\mu_g$ values close to 1 is an effect of the higher chance of central signals to be initially contained in randomly-initialized boxes. Instead, Algorithm 2 allows to exploit the slightly higher maximum density reached by signals with one or more features close to the boundaries of the space, due to the already mentioned truncation we operate outside the $[0, 1]$ range.

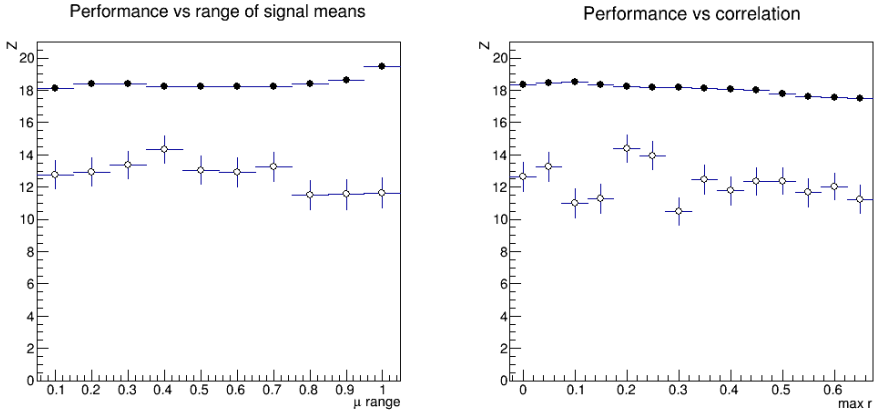


FIGURE 3.5: Mean values of the test statistic Z_{PL} as a function of the characteristics ($\Delta\mu_g$ for $r_{gh} = 0.2$ (left), and r_{gh} for $\Delta\mu_g = 0.7$ (right)) of the signal component, from 50 repetitions of searches in synthetic datasets each composed of 50 signal events and 4950 background events. Black points correspond to searches initialized with Algorithm 2, empty points correspond to searches initialized with Algorithm 0. For reference, the critical region (for $\alpha = 0.05$) corresponds to $Z_{PL} = 7.1(7.2)$ for Algorithm 0 (2, respectively). See the text for other detail.

3.5.2 Power tests of the unsupervised `RanBox`

While in a unsupervised search one cannot in general define a hypothesis test, given the absence of hypotheses for the sampling distributions, we are still interested in verifying the ability of `RanBox` to locate overdense regions of the feature space as a function of its free parameters for a set of different benchmark datasets. This will provide a scale of the algorithm sensitivity. Hence we construct a "flat" dataset containing events uniformly distributed in the feature space, and "signal" datasets where a fraction of the events are sampled from a PDF which includes, for some of the features, a multivariate Gaussian component (see *supra*). Once a type-I error rate α is defined, the tail integral of the test statistic distribution $f(TS|H_1)$, output by `RanBox` searches on alternative hypotheses H_1 corresponding to datasets contaminated with events having multivariate Gaussian features, allows to construct a power function $1 - \beta(\alpha)$ as

$$1 - \beta(\alpha) = \int_{x_{cr}(\alpha)}^{\infty} f(x|H_1)dx, \tag{3.51}$$

where $x_{cr}(\alpha)$ is defined by the relation

$$\alpha = \int_{x_{cr}(\alpha)}^{\infty} f(x|H_0)dx. \tag{3.52}$$

To check the performance of the algorithm in a controlled setting, we define signal parameters by fixing the Gaussian sigma values in signal events to $\sigma_{gg} = 0.05$, and allow means and correlations to vary in the range $\mu_g \in [0.15, 0.85]$ and $r_{gh} \in$ {-0.2,0.,0.2}, respectively. We consider again samples of 5000 events, and study the power $1 - \beta$ for the three choices $\alpha = 0.05, 0.01, 0.001$, using $\mathcal{D} = 20$ space dimensions. We also set the following algorithm hyperparameters:

- Algorithm = 0
- $N_{trials} = 1000$ subspaces scanned for each dataset
- test statistic used: Z_{PL}
- expectation value of events in the box: $N_{exp,V}$.

In a first test we fix the number of features where the signal component exhibits a Gaussian distribution to $N_g = 15$, and vary the number of signal events in the generated samples. The critical region is directly obtained for $\alpha = 0.05$ from the distribution $f(TS|H_0)$ obtained by repeating 500 times the procedure of generation and 1000-subspace-search of datasets including no signal. For the two smaller values of α (0.01, 0.001), we instead rely on the modeling of the distribution of $f(TS|H_0)$ with a Gamma function (see Fig. 3.6) to determine the corresponding x_{cr} values. For each studied value of the signal component we obtain 50 values of $f(TS|H_1)$, from which we extract the power as the fraction of values in the critical regions corresponding to the three chosen values of α. The results of this test are shown in Fig. 3.7 (top row). We observe that `RanBox` is fully capable of spotting localized accumulations due to a multivariate Gaussian signal, down to few-per-mille contaminations of the data sample.

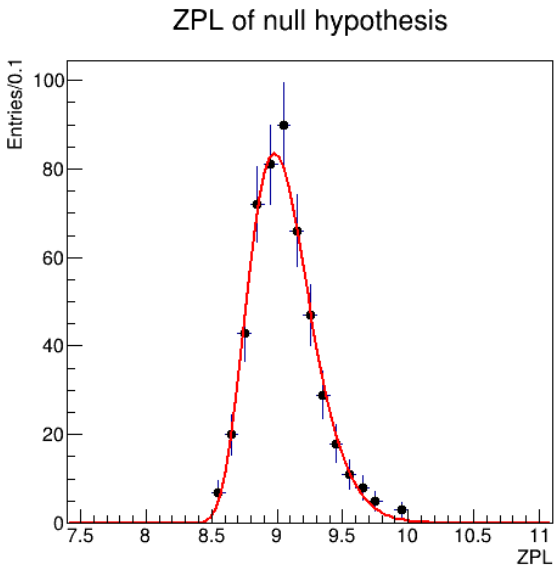


FIGURE 3.6: Distribution of the Z_{PL} test statistic for 500 repetitions of `RanBox` tests of the null hypothesis in 5000-event background-only samples; a fit to a Gamma function is overlaid. 1000 subspaces are scanned with Algorithm 0 for the box initialization. See the text for other details.

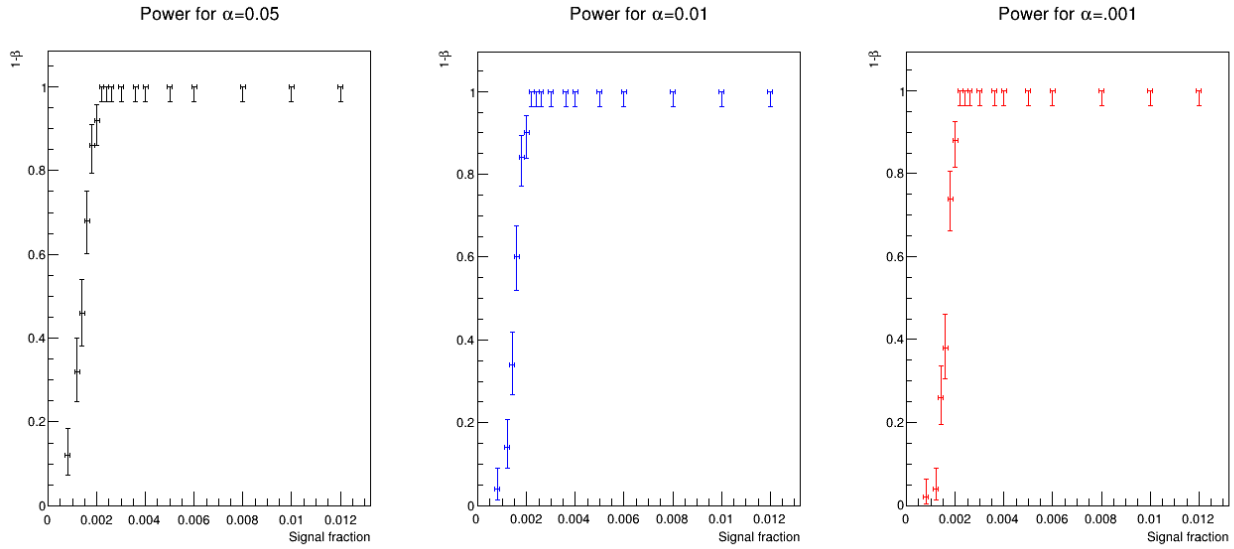


FIGURE 3.7: `RanBox` power curves for Z_{PL} as a function of the fraction of signal in 5000-event samples. The black points (left) correspond to $\alpha = 0.05$, the blue points (center) to $\alpha = 0.01$, and the red points (right) to $\alpha = 0.001$; the critical region for the latter two tests are obtained from extrapolated values of Z_{PL} for the null hypothesis. 68.3% intervals are computed with the Clopper-Pearson method for the Binomial ratio. See the text for other details.

In a second study we determine, with the same procedure described *supra*, the power

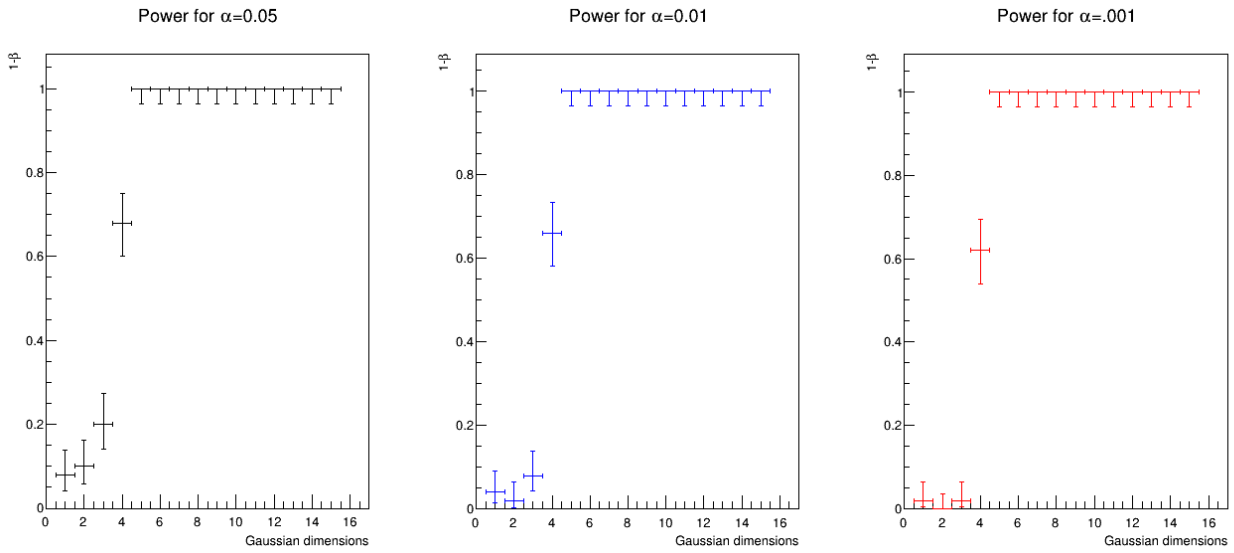


FIGURE 3.8: RanBox power curves for Z_{PL} as a function of the number of Gaussian features in signal events, in samples containing 50 signal events and 4950 flat-distributed events. The black points correspond to $\alpha = 0.05$, the green points to $\alpha = 0.01$, and the red points to $\alpha = 0.001$; the latter two are obtained from extrapolated values of the critical region. 68.3% intervals are computed with the Clopper-Pearson method for the Binomial ratio. See the text for other details.

of RanBox as a function of the number of Gaussian dimensions N_g of the signal component, by fixing the signal fraction to $f_s = 1\%$ (*i.e.*, 50 signal events and 4950 background events). We observe in Fig. 3.8 (top row) that there is sensitivity to multivariate Gaussian signals that involve even only few (4 and above) of the 20 dimensions of the feature space.

In Fig. 3.9 and Fig. 3.10 we provide a visualization of sample results of a RanBox run. The first figure shows marginal distributions of the six features where RanBox identifies an anomalous signal, in the copula space (where the total dataset has by definition uniform marginals before the selection). The subspace where the best box is found is one where the signal exhibits Gaussian distributions in all the features, and all the events in the box are in fact due to the signal component. The scatterplots of Fig. 3.10 show two-dimensional distributions of the full data sample and the data selected as the best box. This further demonstrates the correct working of the algorithm, which can effectively extract the overdense region from an apparently flat distribution. The conclusions we draw are that the algorithm performs as expected when run on a synthetic data sample and in controlled conditions.

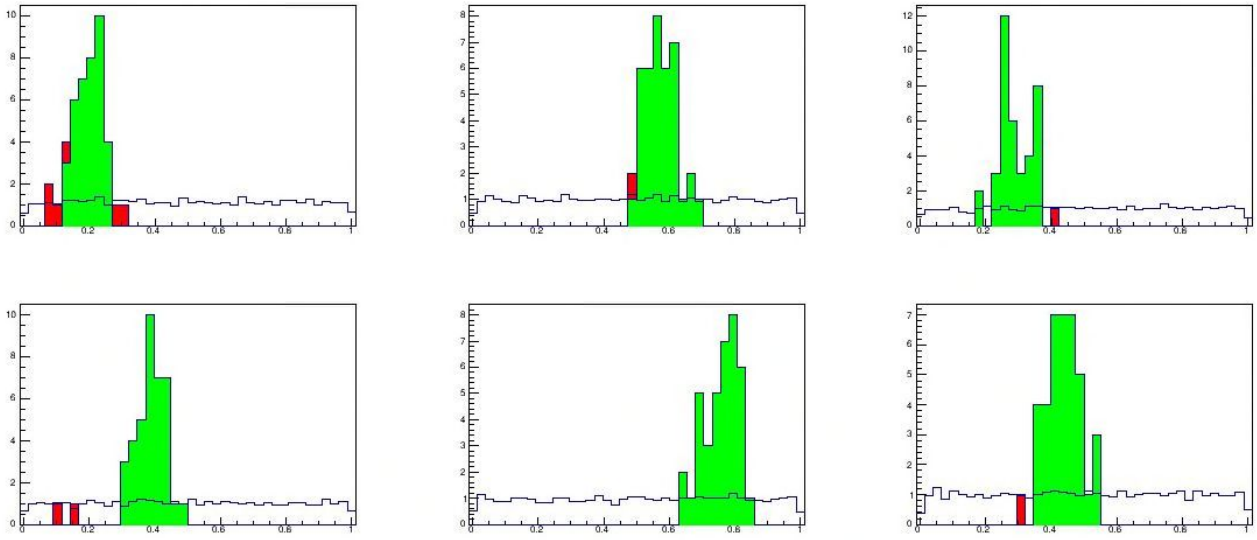


FIGURE 3.9: Distribution of the six features defining the subspace where RanBox finds the highest-Z_{PL} box in a run on 5000 synthetic events, 4950 of them generated from a $D = 20$-dimensional uniform distribution and the remaining 50 "signal" events generated with 11 features drawn from a multidimensional Gaussian distribution. The blue histograms show the totality of the data; the filled green histograms show the distribution of events contained in the highest-Z_{PL} box; the filled red histograms show the distribution of events that fail to be contained in the box only because of their value on the displayed variable. See the text for other details.

3.6 Experiments

The power tests described in Sec. 3.5 are about as far as one can go to characterize the performance of the unsupervised version of RanBox, since on any real-life dataset the specificities of the data structure and the lack of generalization power of the algorithm will make it pointless to investigate in a systematic way its optimal settings and resulting sensitivity. For this reason, in this Section we free ourselves of the need to assess confidence intervals on all the reported statistics, which would also entail a quite significant computing burden [6], and prefer to offer sets of results of single runs of the algorithm on samples of data taken from a dataset offered by particle physics research.

[6] The tests we report in this work overall cost several thousand hours of single-machine CPU by themselves.

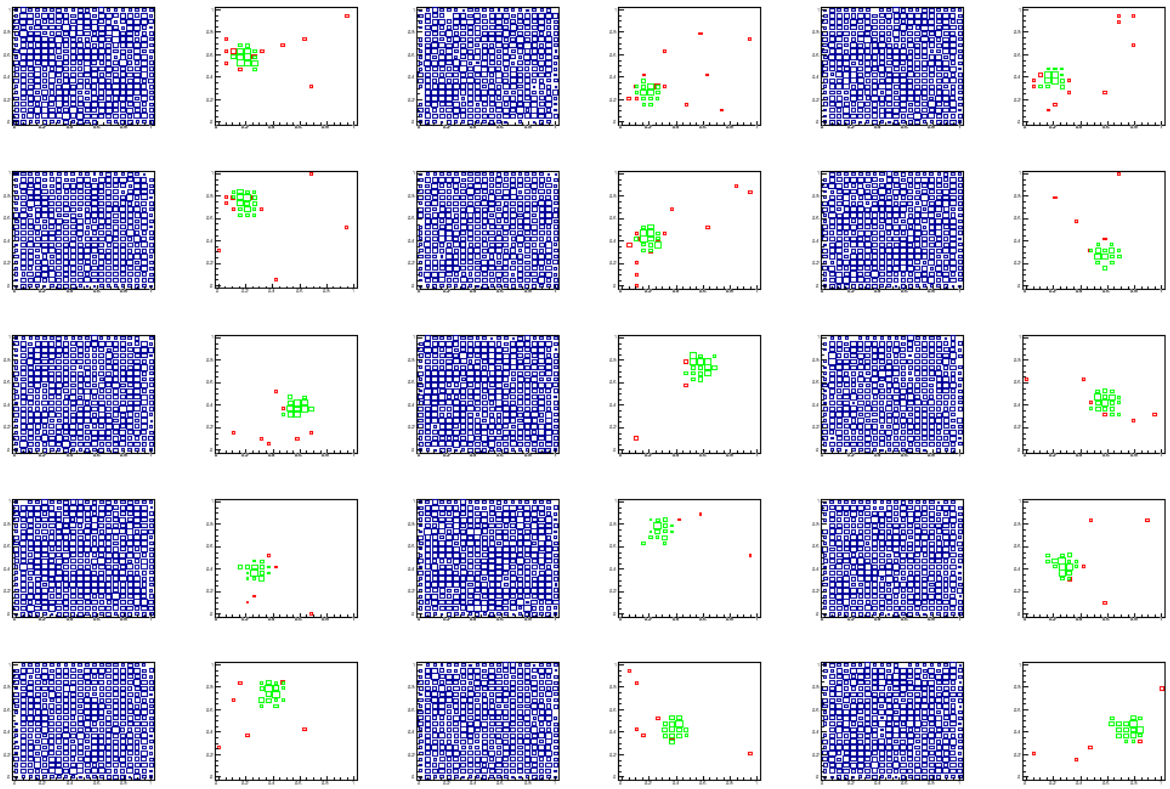


FIGURE 3.10: Scatterplots of the six features defining the subspace where RanBox finds the highest-Z_{PL} box in a run on 5000 synthetic events, 4950 of them generated from a $D = 20$-dimensional uniform distribution and the remaining 50 "signal" events generated with 11 features drawn from a multidimensional Gaussian distribution. The distribution of the totality of the data is shown in blue on the left of each pair of graphs, while the distribution of selected events (in green) is shown in green on the corresponding right graph; in red are events that fail to be included in the highest-Z_{PL} box only because of their value of the shown features. From top to bottom and left to right each pair of graph describes the spaces (v_1, v_2), (v_1, v_3), (v_1, v_4) (first row), (v_1, v_5), (v_1, v_6), (v_2, v_3) (second row), (v_2, v_4), (v_2, v_5), (v_2, v_6) (third row), (v_3, v_4), (v_3, v_5), (v_3, v_6) (fourth row), and (v_4, v_5), (v_4, v_6), (v_5, v_6) (fifth row). See the text for other details.

3.6.1 Exotic signals in LHC data

The search of new phenomena in LHC proton-proton collisions data is the very application that the unsupervised version of RanBox was designed to address. A signal of new physics may manifest itself as a localized increase in density in some of the features derived from particle interactions in the detector. A model-independent search should consider a complete set of kinematical features describing the observed particles in the final state of the collision events, and perform an unbiased

scan of their combined multi-dimensional distribution.

For a test of `RanBox` on the above use case we rely on the large dataset of simulated proton-proton collisions available in the University of Irvine's repository [64], a dataset known by its nickname "HEPMASS". This dataset was generated explicitly to test multivariate algorithms for classification and search of small signals in large background datasets. The generated signal is that of an exotic resonant particle X, with a mass of 1000 GeV, which decays to a pair of top quarks, $X \to t\bar{t}$, when the top quarks successively produce in their decay a single-lepton final state characterized by a high-energy electron or muon, a neutrino, and four hadronic jets. Background samples describe all Standard Model processes that produce a similar final-state signature. The ATLAS experiment is considered as the detector that performs the reconstruction of the produced particle signals; more detail on the generated dataset and the simulation are available in [65].

Number	Feature	Description
1-3	$P_T^{\ell}, \eta_{\ell}, \phi_{\ell}$	3-momentum of primary lepton
4	P_T^{miss}	Missing transverse momentum
5	$\phi_{P_T^{miss}}$	Missing transverse momentum azimuthal angle
6	N_{jets}	Number of additional jets
7-9	$P_T^{j_1}, \eta_{j_1}, \phi_{j_1}$	3-momentum of first jet
10	$P_{tag}^{j_1}$	First jet b-tag information
11-13	$P_T^{j_2}, \eta_{j_2}, \phi_{j_2}$	3-momentum of second jet
14	$P_{tag}^{j_2}$	Second jet b-tag information
15-17	$P_T^{j_3}, \eta_{j_3}, \phi_{j_3}$	3-momentum of third jet
18	$P_{tag}^{j_3}$	Third jet b-tag information
19-21	$P_T^{j_4}, \eta_{j_4}, \phi_{j_4}$	3-momentum of fourth jet
22	$P_{tag}^{j_4}$	Fourth jet b-tag information
23	$m_{\ell\nu}$	Mass of reconstructed lepton-neutrino system
24	m_{jj}	Mass of jets from $W \to qq'$ decay products
25	m_{jjj}	Mass of reconstructed $t \to Wb \to bqq'$ decay system
26	$m_{j\ell\nu}$	Mass of reconstructed $t \to Wb \to l\nu b$ decay system
27	m_{WWbb}	Mass of hypothetical X resonance

TABLE 3.1: List of the 27 features of signal and background events in the HEPMASS dataset. The first 22 are low-level features, the last 5 are higher-level ones produced by combining the low-level features into physics-motivated observables. See the text for more detail.

The data are characterized by reconstruction-level variables from a fast simulation. An idealized reconstruction of a proton-proton collisions yielding top quark

pairs is performed, identifying the observed jets, leptons, and b-jets [7]. From the reconstruction of the event, the low-level kinematic features obtained are particle momenta: the momentum of the leading lepton, the momentum of the four leading jets (in decreasing order of transverse momentum) and related b-tagging information, and magnitude and azimuthal angle of the so-called "missing transverse momentum" vector. The latter is defined as the opposite of the sum of the momentum vectors of all observed particles, calculated in the transverse plane of the particle beams [8].

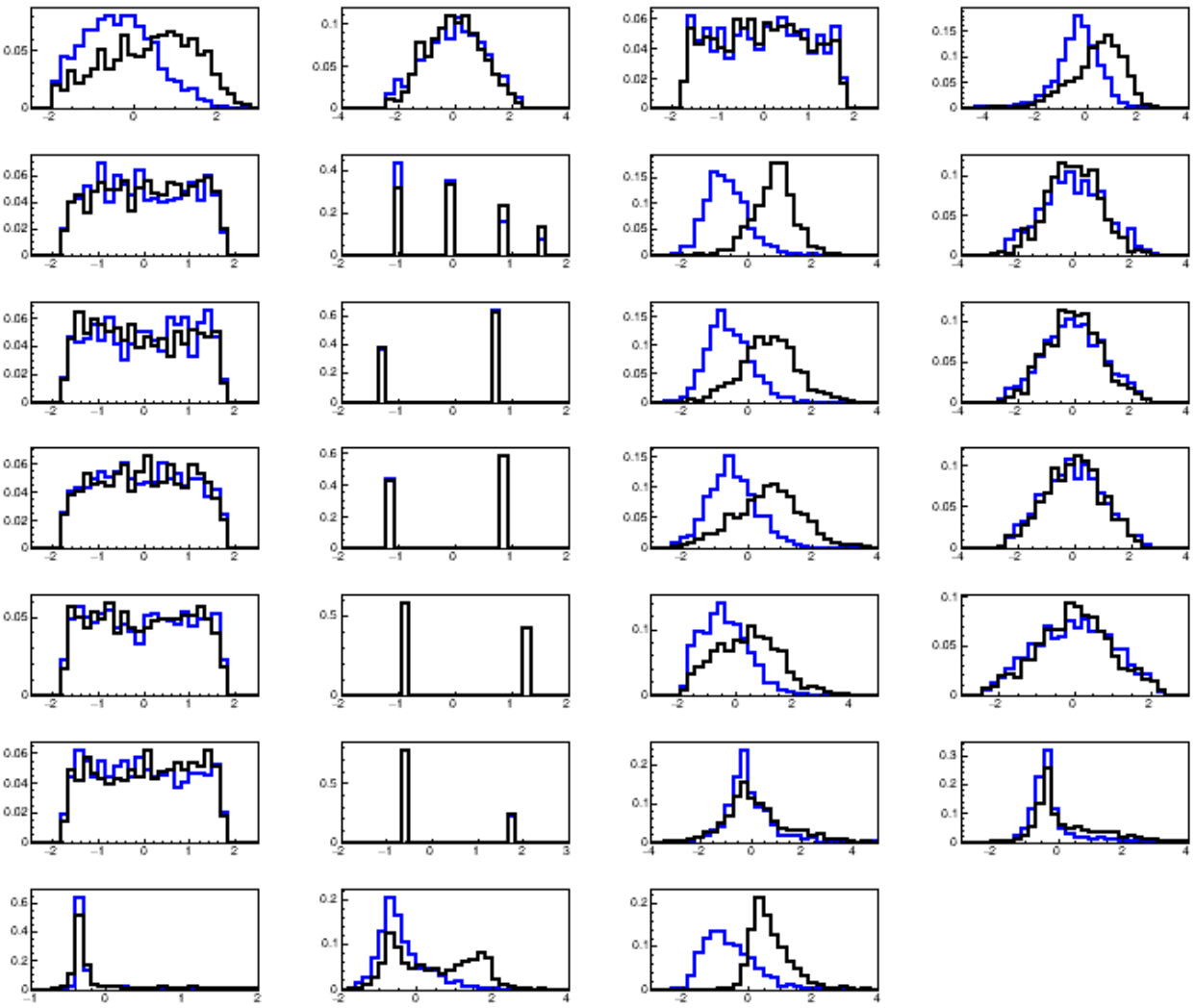


FIGURE 3.11: Normalized and standardized distributions of the 27 features of HEPMASS data for signal (black) and background (blue).

[7] We call "b-jet" a hadronic jet which has been originated from a b-quark. When classified as such by a software algorithm, the jet is said to be "b-tagged".

[8] Missing transverse momentum carries information on the momenta of neutrinos, particles typically produced in weak boson decays that do not leave a traceable signal in the detector but can still be inferred from the imbalance of the momenta of observed particles.

The high-level features of the set are the values of the invariant masses of the intermediate objects calculated using the low-level kinematic features, in the hypothesis that a correct identification of decay objects and assignment to final state particles has been obtained. These are: $m_{\ell\nu}$ from the decay process $W \to \ell\nu$, m_{jj} from the $W \to qq'$ process, m_{jjj} from the $t \to Wb \to bqq'$ process, $m_{j\ell\nu}$ from the $t \to Wb \to \ell\nu b$ process, and the combined m_{WWbb} mass of the decay products assumed for X. Table 3.1 lists identity and information of the 27 features.

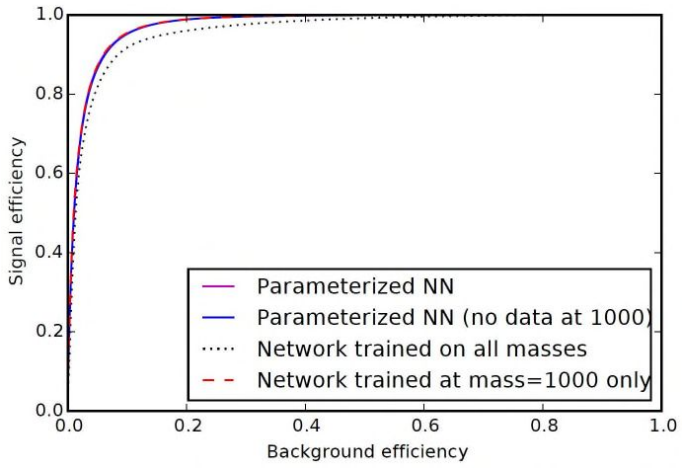


FIGURE 3.12: Comparison of signal and background efficiency curves for four classes of neural networks on the HEPMASS dataset. Of relevance here is the dashed red curve, which correspond to a non-parametrized network trained and tested on the sample of reference, with a resonance mass of 1000 GeV. Reprinted with permission from [66].

In [66] several ROC curves are presented to compare the performance of parametrized and non-parametrized neural networks on the HEPMASS signal discrimination problem. Those are the result of supervised classification, which duly exploits *a priori* knowledge of the signal density. As can be seen in Fig. 3.12, the non-mass-parametrized neural network achieves a background efficiency of about 3% for a signal efficiency of 80%. We will use these approximate values for a qualitative comparison to the performance of `RanBox`, bearing in mind all the caveats of any comparison of supervised and unsupervised classification methods.

In this section we use a mixture of signal and background events from the HEPMASS dataset to test under what conditions `RanBox` is capable of evidencing feature space regions with a dominant signal contamination. Since the feature space is rich

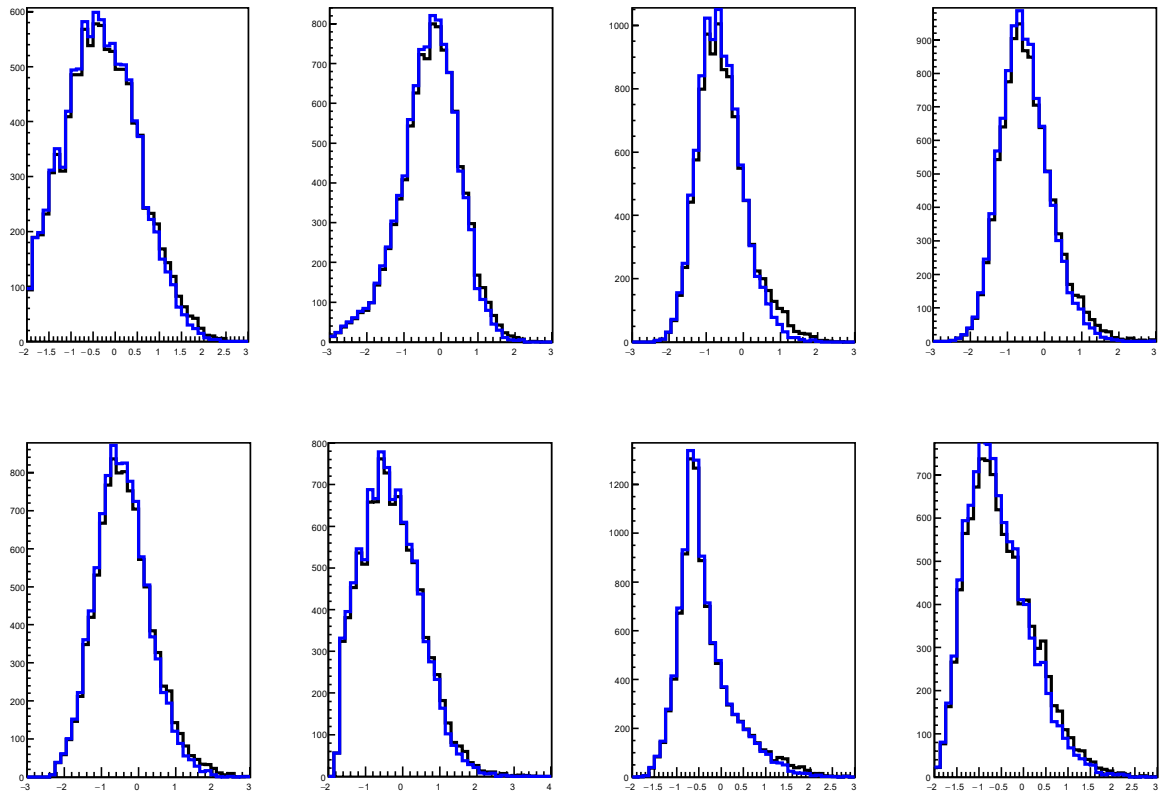


FIGURE 3.13: Comparison of the distribution of pure background (blue) and a mixture of 5% signal and background (black) in the most discriminating features in the HEPMASS dataset. Left to right, top to bottom: features 0, 3, 6, 10, 14, 18, 25, and 26.

with interdependencies among the features, the task of a unsupervised algorithm is considerably harder than in the case of the synthetic dataset studied in Sec. 3.5, as significant overdensities are expected to arise from the structure of background processes alone. Furthermore, in a real-life application of `RanBox`, the user would be unable to extract the distribution of the test statistic under the null hypothesis, as even slight differences between simulation and real data would distort the output. We consider therefore that in that case `RanBox` would be used by running it on real data as they come, without any pretense of assessing a significance level of the returned overdense regions or of studying the power of a selection criterion, but rather with the aim of focusing the attention of researchers on the combinations of features that exhibit interesting localized overdensities.

We proceed with exploratory runs of the `RanBox` algorithm on the HEPMASS dataset as we would perform them on real data. We construct a dataset comprised

Algorithm	Init.	T.S.	Extrap.	$\mathcal{D}'$	D red.	N_{iter}	N_{best}
RanBox	A2	R_1	SB	$\mathcal{D}' = 12$	no	10,000	N/A

TABLE 3.2: Run parameters of the RanBox algorithm for a test on the HEPMASS dataset with a 5% signal contamination. "Init." indicates the method defining the initial dimension of the search box; "Extrap." identifies the way by which a prediction of events in the box is computed; $\mathcal{D}'$ is the dimensionality of the subspaces scanned by RanBox; "Dim. red." indicates whether the dimensionality of the feature space was reduced with PCA or by discarding the most correlated variables; N_{iter} is the number of searched subspaces by RanBox.

of 250 signal and 4750 background events: the 5% signal fraction is small enough to make the signal indistinguishable in the marginal distributions of even the most discriminating variables, as shown in Fig. 3.13. We run RanBox with the parameters listed in Table 3.2. They constitute a reasonable choice for a run on HEPMASS. In particular, since we wish to be sensitive to a small signal contamination rather than having the algorithm get distracted by broader-scale background correlations, we initially consider that the R_1 test statistic might be more sensitive to a signal component. Also, we use the sidebands method to extrapolate the density in the search box, as this better factors out the local disuniformities in the data. The choice of dimensionality of the scanned subspaces is instead driven by preconceptions on the fact that a signal of new physics will most likely exhibit distinctive features only in a subset of the considered kinematical variables [9]; 12 is anyway close to the maximum meaningful choice for that parameter, as is clear if we consider that in a 12-dimensional space a box of sides equal to half the range of each feature will on average contain only $5000 \times 2^{-12} = 1.2$ events out of 5000. Finally, we do not apply any dimensionality reduction to the input data, as we observe that the maximum two-variable correlation coefficient (0.757) in the mixture dataset is not particularly high.

From the results listed in Table 3.3 we may draw a few interesting conclusions. First of all, the search of 10,000 subspaces performed by RanBox returns a good number of signal-rich regions, as five of the ten most significant boxes are dominated by

[9] Indeed, in the considered search for $X \to t\bar{t}$, apart from the resonant structure of the total invariant mass of the decay products, one expects only minor differences of the signal with respect to the non-resonant $t\bar{t}$ production predicted by the SM.

R_1	N_{in}	N_{exp}	N_s	ϵ_s	Gain	Active features
52.78	54	0.02	46	0.184	17.04	10101110011101000000110001
45.35	48	0.06	38	0.152	15.83	00011110001100101100010011
41.60	46	0.11	33	0.132	14.35	10001001011011011000100001
40.72	46	0.13	18	0.072	7.83	10100011011010100010010011
40.38	44	0.09	41	0.164	18.64	10010010010001000111011001
40.17	47	0.17	0	0.000	0.00	01100100010001011100110011
39.82	44	0.10	0	0.000	0.00	10000101010001100101010101
38.54	44	0.14	0	0.000	0.00	00100110110111000110110000
38.36	44	0.15	30	0.120	13.91	00011010111001000000110101
38.05	43	0.13	14	0.056	6.51	11000010011000111010001 1001

TABLE 3.3: Results of an exploratory `RanBox` search on the HEPMASS dataset with a 5% signal contamination; data for the 10 most significant boxes are reported. N_s indicates the number of signal events in the search box; ϵ_s is the efficiency of the box selection for the signal component; gain is computed as the increase in the signal fraction of the box over the initial dataset. For other detail see the text.

the signal component, and two more are also considerably signal-enriched, by factors above six [10]. Such an output, and in particular the most significant box alone, would certainly allow experimentalists to focus on the small signal now evident in the identified regions, hence we consider this output a success of the anomaly detection task. We also note that the scan of 10,000 12-dimensional subspaces costs nearly 10 hours of running on a single CPU; the scan of all 12-dimensional subspaces of the 27-dimensional feature space is instead not an easily viable option, as this would require $27!/(12!15!) = 1.738 \times 10^7$ iterations, or about two years of CPU on a single machine. Regardless, on the HEPMASS dataset a limited number of combinations of 12 features still allow to evidence a small signal.

If we now compute the signal and background efficiency of the regions returned by the algorithm in its exploration of the $f_s = 5\%$ dataset, we notice that the best box identified by `RanBox` contains 46 signal events out of 54, which corresponds to an 85% efficiency; the background efficiency is instead $8/4950 = 0.16\%$. These numbers compare quite favourably to those of the neural network results graphically displayed in Fig. 3.12. We stress again the modest value of this observation, given the improper nature of a comparison of this kind. In particular, the `RanBox` results have unknown generalization properties —they are obtained from a single dataset,

[10] In the following we take that factor as a threshold to count the number of signal-rich (SR) boxes among the first ten boxes, a number we report as $SR_{1:10}$.

on which multiple testing is performed: the performance would be less good on a different testing sample. On the other hand, the search algorithm was only shown a total data sample of 5000 events, a number over two orders of magnitude smaller than the training sample of the neural networks.

Test	N_s/N_b	T.S. max	N_{in}/N_{exp}	N_s	Gain	$SR_{1:10}$	$\overline{\epsilon_s^{1:10}}$
1	250/4750	$Z_{PL} = 28.06$	39/0.00	35	17.95	8	0.097
2	200/4800	$Z_{PL} = 27.46$	1003/133.00	0	0.00	6	0.079
3	150/4850	$Z_{PL} = 24.45$	24/0.00	21	29.17	7	0.140
4	100/4900	$Z_{PL} = 26.95$	45/0.01	0	0.00	1	0.014
5	80/4920	$Z_{PL} = 24.65$	43/0.05	14	20.35	3	0.044
6	70/4930	$Z_{PL} = 23.56$	41/0.01	0	0.00	1	0.010
7	250/4750	$R_1 = 52.78$	54/0.02	46	17.04	7	0.092
8	200/4800	$R_1 = 50.78$	60/0.18	33	13.75	6	0.097
9	150/4850	$R_1 = 43.58$	53/0.21	15	9.44	6	0.071
10	100/4900	$R_1 = 45.29$	49/0.08	0	0.00	3	0.038
11	80/4920	$R_1 = 51.22$	52/0.02	0	0.00	0	0.000
12	70/4930	$R_1 = 43.44$	48/0.10	0	0.00	0	0.000

TABLE 3.4: Sample results of `RanBox` runs on 5000 events from the HEPMASS dataset, with varying signal fraction and the two choices of test statistic. See the text for more details.

We perform a test using `RanBox`, as detailed in Table 3.4, using 10,000 trials for the subspace sampling and a subspace dimensionality of $\mathcal{D}' = 12$. This time we start (see test 1) by searching for a 5% signal in a set of 5000 events using the Z_{PL} test statistic. The algorithm returns as the most significant box one which is rich in signal component, and we observe that the three next-best-significance boxes (not reported in Table 3.4) are similarly enriched in signal events. We gradually reduce the signal fraction in tests 2-6 and observe that results are not uniform: `RanBox` in some cases identifies as the most significant box one devoid of signal. In general we observe that the number of boxes that are signal-enriched among the first 10 ($SR_{1:10}$) usually decreases as initial signal fraction is reduced; the average signal efficiency also becomes smaller. Yet the algorithm finds significantly signal-enriched boxes among the first 10 even for an initial signal fraction of 1.4% and 1.6%. We also observe that in test 2 the Z_{PL} maximization focuses on a very wide box, an indication of the existence of broad-scale multivariate density variations of the background component of this dataset.

Based on the above observation, in tests 7-12 we turn our attention to the R_1 test

statistic, which should give more importance to smaller feature-space regions. This indeed allows `RanBox` to converge on signal-rich regions when the signal fraction of the data sample is 3% or larger; for smaller signal fractions, however, `RanBox` becomes unable to evidence the signal component in the reported overdense regions.

The results also allow us to draw some conclusions on the most performing settings of `RanBox` to be used in the HEPMASS use case. Here, however, we stress one important point: by telling the tale of how these choices may be defined based on sample test results, we are implicitly declaring how the algorithm —but in general, we believe, any unsupervised search— requires an *ad hoc* tuning to perform its task most effectively. This is not to be taken as a demonstration that this kind of search is useless: quite on the contrary, the tool can be a very useful one in examining the properties of multi-dimensional data. It cannot, on the other hand, be employed as a catch-all machine ready to identify an anomaly in an arbitrary dataset: this is nothing else than a by-product of the well-known absence of a universal high-power test statistic, when the alternative hypothesis is not specified.

Chapter 4

Search for B_S meson decays

The search for rare decays of the B_s meson has been identified as a very promising avenue to detect small hints of deviations from the predictions of the Standard Model of particle physics. The CMS experiment at the CERN LHC has accumulated a large statistics data sample where a search for the $B_s \to \tau^+\tau^-$ decay may be carried out. As the signal is exceedingly rare, and buried in very large backgrounds, it is important to exploit as much as possible the observable features of the recorded final state. In this section we summarize the studies performed to reconstruct the B_s meson decay topology, using the observable particles generated by tau lepton pairs produced in the $B_s \to \tau^+\tau^-$ decay. The developed multi-stage, machine-learning-powered regression of the mass of the B_s meson as well as the classification algorithm for the 3-prong decay channel promises to enhance the sensitivity of the search we conduct with the parked data collected by CMS during LHC Run 2.

4.1 Introduction

With the 2012 discovery of the Higgs boson [4, 3], the Standard Model of particle physics (SM) [67] has consacrated itself as an extremely successful and predictive theory. The SM allows physicists to compute with high accuracy predictions for the probability of reactions among fundamental particles; hundreds of quantitative predictions have been verified by experimental tests performed at particle colliders in the past 50 years, and only a very small number of discrepancies exist at the time of writing.

Despite the general success of particle theory to explain subnuclear phenomena with the SM, the model cannot be the final theory of fundamental forces and matter

particles. There are a number of reasons for this; the most convincing are the fact that the SM does not include gravity, and the naturalness problem, arising from the observation that the Higgs boson has a physical mass which is abnormally small as the result of the addition and subtraction of quantum corrections each of orders of magnitude larger size than the mass itself. Because of these and other issues, physicists are searching for new phenomena that may provide clues on how to extend and make more coherent the SM.

The Large Hadron Collider (LHC) [68] is the facility producing the most energetic particle collisions. In operation since 2008, the LHC provides high-intensity beams of protons accelerated to energy of up to 7 TeV, and rotating in opposite verse inside superconducting magnets lining up a 27 km-long underground tunnel at the border between Switzerland and France. Collisions are generated between the proton beams in four experimental halls. One of them is home to the CMS experiment [30], a large multi-purpose detector that is capable of recording the signal deposited in hundred million electronic channels from all produced particles in the energetic collisions, which take place every 25 nanoseconds in its interior, and storing the most interesting of them to permanent media.

The collection and analysis of LHC collisions has been going on for over a decade, and will continue for at least as much time in the future, because new physics that extend beyond the SM may become observable only by accumulating data from large amounts of collisions. It is, in fact, a deeply-ingrained prejudice in particle theorists and experimentalists alike that amidst the rarest phenomena, which the SM allows to occur at the smallest rates, a signal of new physics might first pop up and become detectable. New massive particles which mediate as-of-yet-undiscovered forces of nature may in fact become evident in their contribution to variations in the rate of production in Figure 4.1 of those processes which the SM predicts to be the rarest.

The above line of research has brought the CMS collaboration to intensively search, and then observe with LHCb [31], the rare decay of the B_s meson (a particle made up by a bottom quark and an anti-strange quark) into pairs of muons: the rate of that process, which according to the SM takes place only three times in a billion B_s decays, is presently only slightly discrepant with theory predictions; more data will make this test more stringent in the near future. However, if new physics were

contributing to the decays of the B_s meson, a reasonable prediction is that it would also do so in the similar decay of the same particle into tau lepton pairs, and possibly to a higher effect. This might in fact be the result of the new physics particles contributing to the process preferentially through coupling with third-generation matter fields, such as the tau lepton. In the past, a search for the B_s decay to tau lepton pairs was performed by the LHCb experiment, with null results [69]. The sensitivity of that search was however insufficient to test the SM prediction for the involved branching fraction of the B_s meson.

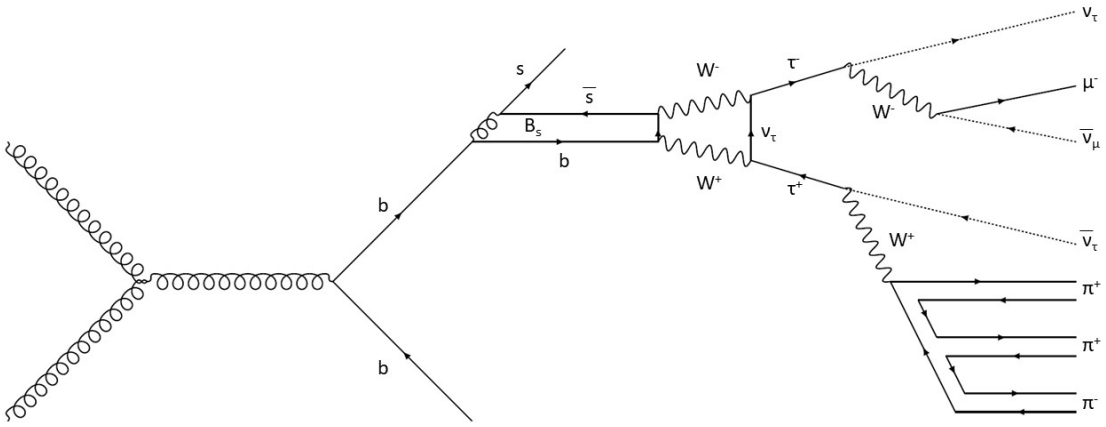


FIGURE 4.1: Feynman diagram showing the production of a B_s meson and its subsequent decay of interest to this work.

Detecting tau leptons is hard: unlike their lighter versions from the first and second generation (electrons and muons), tau leptons decay very quickly (in 2.9×10^{-13} s) both to leptons plus neutrinos, and to hadrons and neutrinos (which is thus called "semi-hadronic decay"). In the latter case the produced hadrons are mostly pions, which are among the most common particles produced in proton-proton collisions. For that reason the semi-hadronic decay is extremely hard to identify in the large background. Furthermore, the neutrino emitted in the decay escapes undetected, preventing a full kinematical reconstruction of the decay process. These difficulties have so far prevented the extraction of significant information from rare decays of hadrons involving tau leptons in the final state. On the other hand, precisely because

of those challenges, there are wide opportunities to progress in our understanding of nature by studying these processes. While it is foreseen that the data so far accumulated by the CMS experiment will not nearly be sufficient to put in evidence a decay of the B_s meson into tau leptons, it is necessary to fill the void by constructing a search strategy for that process, such that its future refinement and the larger datasets that the LHC will deliver in the course of the next few years will eventually allow the extraction of a measurement, and the comparison with SM predictions.

In this work we describe our studies to solve the difficult problem of reconstructing the decay of the B_s meson using the information available in the form of four-momenta of the four visible decay products from the combined leptonic and semi-hadronic decay of the pair: a muon and three charged pions. The chosen final state is advantageous because the muon provides an ideal triggering signature, with which data collection is significantly eased; on the other hand, the three-prong semi-hadronic decay also offers some advantages due to the possibility of reconstructing the originating tau lepton momentum. The mixed final state further benefits from a factor of two probability of occurrence due to the combinatorial factor, as we allow for positive and negative muons from the leptonic decay (and correspondingly, three pions of total charge -1 or $+1$, respectively). The use of machine learning tools may allow to extract more informative statistical summaries from the available information. The inferred kinematics of the hypothetical B_s particle producing the observed final state may then become the input to a performant classifier, capable of increasing as much as possible the signal to noise ratio in the analyzable data.

The structure of this document is as follows. In Section 4.2 we describe the data samples we have used for this work and how high-level information on the observable final state particles is reconstructed. In Section 4.3 we discuss the preliminary data selection, which employs a deep neural network for the identification of the most likely pions emitted by tau leptons decaying semi-hadronically. In Section 4.4 we describe the studies performed to assemble a regressor capable of producing an estimate of the B_s mass and momentum.

4.2 Datasets and Event Preprocessing

The decay channel of interest is the B_s meson decay into two tau leptons, where one tau decays into a muon and a neutrino and the other tau decays into three pions and two neutrinos. For this task we employ two learning algorithms: a regression to the most probable value of the invariant mass of the two tau leptons, using the four-momenta of observed particles in the events where two tau candidates may be the result of the B_s decay; and a classification algorithm of such events based on four-momenta of observed particles as well as higher level kinematic features.

In this section we discuss the data samples we worked with for these studies: "real" data, data collected in the Run 2 LHC collisions which are utilized to check the behaviour of background events passing a preliminary event selection targeting the B_s meson decay signature, and a Monte Carlo simulation of B_s meson decays including the final state of interest, with which the algorithms can learn to infer the B_s decay kinematics. We generated a Monte Carlo sample with a pile up profile[1] that corresponds better to the real data pile up conditions and simulate the B_s decay topology and the resulting reconstruction of observable features from detector readouts.

4.2.1 B-Parking Data

LHC collides proton bunches every 25 ns, hence with rates of 40 MHz. In order to reduce the resulting huge flow of information from its millions of detector channels, CMS uses a two level trigger system: Level 1 (L1) and the High Level Trigger (HLT). L1 is hardware-based and it relies on a fast read-out of the detector, without the tracker information. HLT is based on speed-optimized software which exploits a full read-out of the detector with the tracks which we use to reconstruct the physics objects. Events that are not selected by the triggers are not stored (for more details, see section 2.3).

For our studies of background processes, we used the so-called *B-Parked Data* acquired by CMS during the Run 2 of the LHC. This is a large sample collected at

[1] Pile-up (defined in section 2.1.2) is the collective signal of proton-proton collisions taking place in the same bunch crossing as the event of interest. Due to the high intensity of LHC beams, the typical number of simultaneous collisions in Run 2 data is O(30). Simulations need to account for these satellite collisions in order to improve the model of real data.

a center-of-mass energy $\sqrt{s} = 13\ TeV$ of proton-proton collisions, acquired in 2017-2018 through a loose selection enforcing the presence of a low-transverse-momentum muon candidate at trigger level. The large cross-section corresponding to the loose trigger requirement implies that the selection cuts were dynamically varied as a function of the instantaneous luminosity of the data-taking conditions, in order to not exceed the maximum bandwidth of accept rates. The large number of acquired events prevented CMS to directly process and reconstruct the data, which were thus "parked" to be processed later. The interest of this dataset is that the loose trigger requirements (see *infra*, Table 4.1) provide as unbiased a B-hadron-enriched dataset as they come. The large sample is very valuable for searches of rare phenomena in B-hadron decays, such as the one at the center of our research goals.

The triggering muon is required to pass the following criteria:

- A candidate muon identified at trigger level 2 with momentum above $p_T > 7 - 9\ GeV$; the actual threshold varied during Run 2 data taking, owing to the variable instantaneous luminosity conditions
- An impact parameter significance above 4-6, depending on instantaneous luminosity conditions.

The data sample comprises a total of about 12 billion events, of which roughly 80% are estimated to include processes involving b-quark production. In particular, it is estimated that the sample contains a total of 1.2×10^9 decays of B_s mesons [70].

Settings	Peak Lum. $(10^{34} cm^{-2} s^{-1})$	L1 μ p_T (GeV)	HLT μ p_T (GeV)	HLT μ IP SIG	Purity (%)	Peak rate (kHz)
1	1.7	> 12	> 12	> 6	92	1.5
2	1.5	> 10	> 9	> 6	87	2.8
3	1.3	> 9	> 9	> 5	86	3.0
4	1.1	> 8	> 8	> 5	83	3.7
5	0.9	> 7	> 7	> 4	59	5.4

TABLE 4.1: Settings[2] of the B-parking triggers employed by CMS in 2017-2018 [70]. See the text for details.

[2] HLT μ IP SIG: high level trigger muon impact parameter significance defined as the ratio between the transverse impact d_0 parameter and the measured uncertainty $\sigma(d_0)$.

4.2.2 Monte Carlo Samples

The Monte Carlo (MC) sample we employ in our signal studies is a simulation of proton-proton collisions yielding b-quarks in the final state. The MC is generated using PYTHIA8 [71] for proton-proton collisions, parton shower and hadronization. Pythia is a general purpose event generator, containing theory and models for a number of physics aspects, including hard and soft interactions, parton distributions, initial and final state parton showers, multiple interactions, and decays. It has been extensively used by the CMS collaboration for event generation, as well as for hadronization of the parton-level events. The B-meson decay is modeled with the EVTGEN [72] package. Events are then passed through the CMS detector simulation using the GEANT4 [51] package.

The MC generation includes a modeling of the trigger which selected the B-Parking dataset for Run 2, and the correct amount of pile-up proton-proton collisions in the same bunch crossing of the triggering collision expected for the Run 2 data-taking conditions. These are meant to model the details of the instantaneous luminosity of the machine during the acquiring of the real data, such that the correct number of simultaneous proton-proton collisions is generated along with the event of interest. The pile up profile of the overall MC sample was adjusted by additional samples we generated with comparable pile up profile to that of the parked dataset (see Fig. 4.2).

The data are simulated at generation level such that the hadronization of the b quarks includes at least one B_s meson; the latter is forced to decay into a pair of tau leptons. The simulation further filters events at generation level where at least one of the tau leptons decays into a muon, $\tau^+ \to \mu^+ \bar{\nu}_\mu \nu_\tau$ ($\tau^- \to \mu^- \bar{\nu}_\mu \nu_\tau$). When the event includes such a muon with a transverse momentum p_T above the triggering threshold (see *infra*, Table 4.1), it has all the features we try to select in real data in our search. With the reconstructed characteristics of the measured four-momenta of the muon and the other particles involved in the decay of the B_s, we may attempt a reconstruction of the B_s decay. This also means that events without the selected signature are not stored and can be neglected in our studies.

The simulated data are meant to reproduce the behaviour of events selected by

the B-parking muon trigger, which operated in CMS during Run 2. Table 4.1 includes detailed information on the applied thresholds of the muon triggers used to collect the B-parked dataset used in this study and for the B_s search.

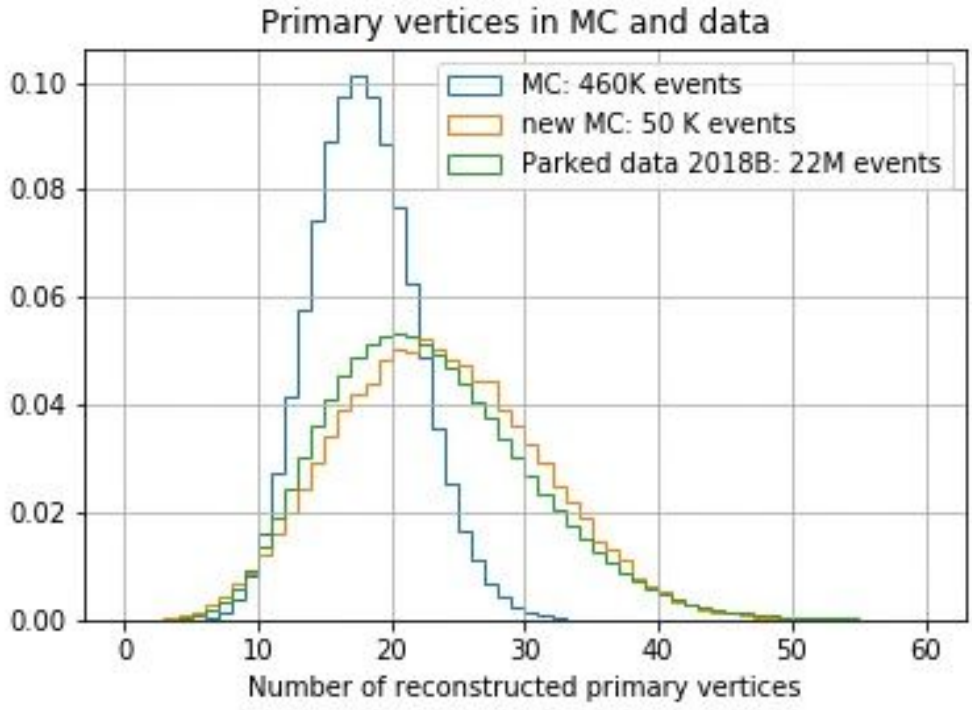


FIGURE 4.2: Number of reconstructed primary vertices in MC and data.

4.2.3 Trigger Strategy

In order to record as many B-hadrons as possible, many single muon triggers were turned on, distinguishable by the thresholds applied to the trigger muon transverse momentum p_T and to the impact parameter (IP) significance (Table 4.1) . The HLT trigger paths were turned off artificially and independently by portions of the LHC fill. During a store of protons and the colliding phase, the decay of the instantaneous luminosity allows to soften thresholds for data collection. Hence as rates allowed it, an active trigger path was turned off and a looser (with a looser cut on the p_T or IP significance) trigger path was turned on (see Fig. 4.3). This data collection mode leads to different integrated luminosities for each present trigger path.

Moreover the trigger efficiency in the MC samples and B-parked data are significantly different due to the fact that the artificial turning on/off of trigger paths is

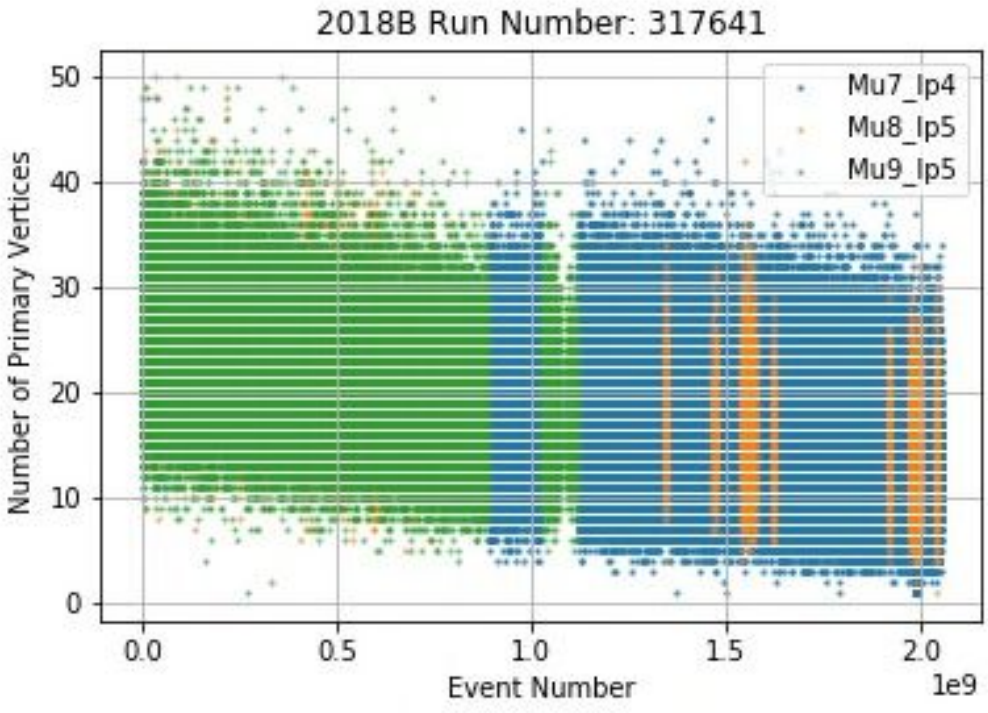


FIGURE 4.3: Number of vertices vs Event Number for different triggers.

not present in the MC simulations, as shown in Fig. 4.4 and 4.5. The list of available triggers in the menu are not the same as well as the proportions of the triggers.

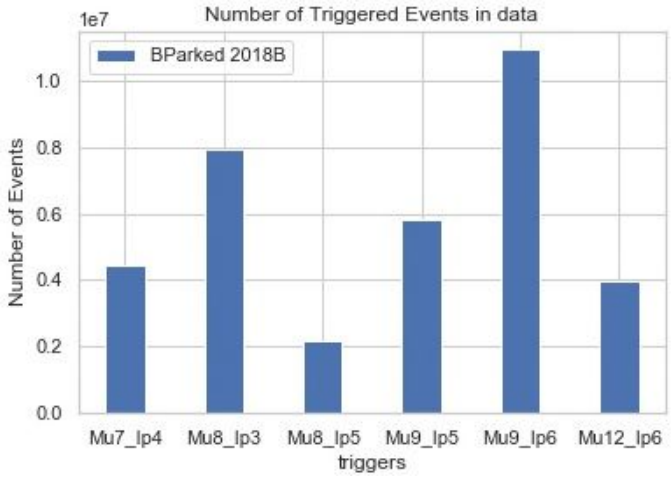


FIGURE 4.4: Triggers present in parked data.

Due to small statistics of some of the subsamples, as well as differences in the kinematics of each, it is too complicated to perform a measurement on individual

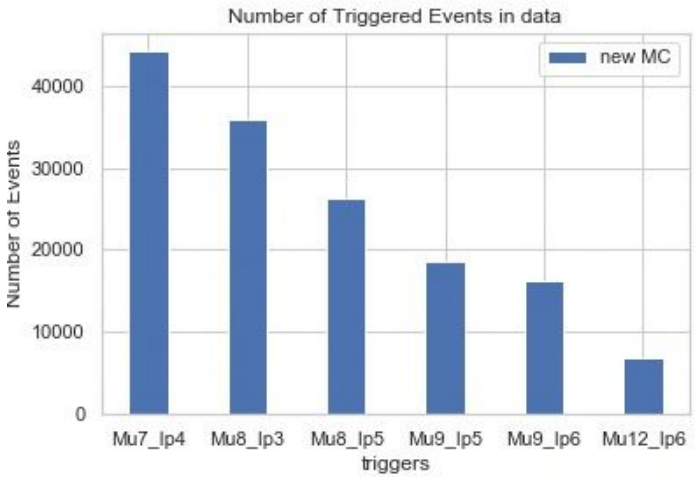


FIGURE 4.5: Triggers present in MC.

dataset sections that belong to a trigger path and then take an average of their results. Instead, we select the trigger paths that are present in both MC and B-parked data and that cover most of the events in the samples.

The general recipe is as follows: we start with the trigger path T1 that makes up for the largest fraction of events collected in the B-parked dataset; the corresponding MC simulated events passing trigger requirements of T1 are taken as a model of that portion of the dataset. We then turn to the remainder of B-parked data (which were not collected by the most common trigger T1 either because that trigger was not active, or because while active it did not fire), and identify the trigger which collected the largest part of those remaining events, T2. Now we need to distinguish among those events ones collected by T2 while T1 was also actively selecting data and ones collected by T2 while T1 was inactive. We turn to the MC events and identify the corresponding classes of events, assigning each to the modeling of the corresponding subclasses of data passing T2 and failing or not tested by T1. This procedure continues until all the main exclusive subsets of real data have been modeled. Since the MC events assigned to the modeling of each subclass have different relative numerosity, they are assigned a weight inversely proportional to the integrated luminosity that each subclass corresponds to. This procedure may not be the most effective in terms of the resulting global uncertainty due to MC statistics that we may end up assigning to derived quantities, but it correctly takes care of the combined effect of different pass/fail conditions on multiple trigger paths.

Monte Carlo Truth Matching

For signal MC events we can make use of the generator-level information and match the generated objects to the reconstructed ones, to study the properties of the reconstruction and to select specific topologies of the decay products without the worry of including events where the signature is mimicked by backgrounds or misreconstructed events. We have verified that for transverse momenta of the objects (muon, tau leptons, and charged pions from the hadronic decay of a tau lepton) higher than about 5 GeV this can be simply accomplished by constraining the azimuthal angle ϕ and pseudorapidity η [3] of the candidate object by requiring the condition $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} < 0.1$ where $\Delta\eta$ and $\Delta\phi$ are angular and pseudorapidity differences between generated and reconstructed level variables. At lower transverse momenta we instead rely specifically on our reconstruction algorithm, which is described *infra*.

4.3 Preliminary Selection

In order to extract an estimate of the mass of a particle decayed into two candidate tau leptons, it is necessary to filter the input data sample such that all selected events contain the minimum information required for the learning algorithms: the visible final-state objects produced in B_s decays. The preliminary selection also has the purpose of removing events that are most likely due to non-signal processes. In this section we describe the preliminary selection that was operated with those aims.

One important preliminary observation that can be made from the simulated signal dataset is that the energies of tau leptons from B meson decays in our data sample are typically within the range [1,15] GeV, with a long tail to higher energies; this is due to the combined effect of the production mechanisms of b-quarks and the trigger selection. Because of this, the decay products of the searched B_s meson –in particular, the charged pions– will be produced with low transverse momentum. This on one side means that it will be difficult to reconstruct them efficiently in some topologies, and on the other that they will spread within a wide solid angle, as opposed

[3] Pseudorapidity (defined in section 2.2.1) is a monotonous function of the polar angle θ of a particle (assumed massless) with respect to the beams direction, $\eta = -\log\tan(\theta/2)$. Since η transforms linearly upon boosts along the beam axis, it is an advantageous quantity for event interpretation in hadron collisions.

to what would happen in the decay of a much more boosted hadron. Hence, rather than searching for these particles within a jet of a given radius, we need to consider all charged particles we observe in a wide area around the triggering muon. Moreover, in the CMS high pile up scenario (see *supra*) we have too many reconstructed charged pions to consider as candidates when we are trying to identify the three ones emitted in the semi-hadronic tau decay. In order to overcome these challenges, following the work documented in [73] we proceed in the topology reconstruction of our event candidates with the following steps:

Vertexing: The first step is to determine the best candidate primary vertex (PV) where the b hadron is produced (where the $B_s \to \tau\tau$ decay takes place); we may then exclude from consideration the charged hadrons that clearly belong from different proton-proton collision vertices. For this aim we make use of the muon that triggers the event, by extrapolating it back to the beam line and selecting the reconstructed primary vertex that is closest to the extrapolated point along the z coordinate. This way we focus on the true event vertex, which allows us to get rid of all particles produced by pile-up vertices, thereby significantly reducing the number of considered charged hadrons.

Pre-filtering deep neural network: We use a deep neural network (DNN) to decide if a charged hadron is coming from a tau decay or not, such that we may identify triplets of charged pions with the correct charge combination to suit the hypothesis of being due to the decay of a tau lepton, when its partner produced the triggering muon (which must thus have the opposite charge to the sum of the three pions charges). We choose an operating point for the DNN selection which corresponds to a true positive rate of 80% for each pion; this way, about 50% of the chosen three-pion combinations from signal decays will be correctly identified. The architecture of the DNN that operates this selection is described in detail (see Sec. **??**), and the selection is further discussed there.

Post-filtering: Taking the following facts into account, we further reduce the number of charged hadron candidates after the DNN selection by the following means:

(i) Since the B_s meson has a comparatively long lifetime ($\tau_{B_s} = 1.515 \times 10^{-12}$ s [50]), we expect a displacement between the extrapolated primary vertex and

the secondary vertex, where the three charged pions are produced from the semi-hadronically decaying tau lepton. We reconstruct the three pions to a common vertex and we enforce the condition that the distance between primary and fitted secondary vertex is different from zero at three-sigma level, i.e. $d/\sigma_d > 3$, where $d \pm \sigma_d$ is the measured distance.

(ii) We also apply a constraint on the quality of the secondary vertex constructed with the three pion candidates, by estimating the compatibility of the three tracks with a single vertex. This is measured by the reduced χ^2 of the vertex fit, converted into the corresponding tail probability. We select pion triplets for which this probability is equal to at least 5%.

(iii) A third optional selection requirement which can be applied to further increase the purity of the selected three-pion combinations is to enforce the condition that the invariant mass of the three charged pions selected by the DNN is smaller than the nominal 1.77 GeV tau lepton mass, taking into account the loss of energy in the decay due to the escaping tau lepton neutrino. We may further exploit the fact that the three-pion final state often includes a ρ-meson resonance. As there are two possible opposite-charge combinations of two pions in a three-pion set of unit charge, we may require that at least one of the two two-pion combinations is consistent within resolutions with the mass of the ρ meson. In Fig. 4.6 we show how indeed one combination of pion pairs of null charge is in most cases within the range of the ρ particle mass (770 MeV), such that the correct pion combinations (yellow points) populates a cross in the two-dimensional graph.

After the filtering steps, if multiple candidate triplets are found, the one of combined highest transverse momentum is chosen.

4.3.1 DNN selection of candidate tau leptons

Identifying a hadronically-decaying tau lepton is a very complex task in the busy environment of LHC collisions. The investigated signature (three charged pions) is ridden by large backgrounds, which can be reduced by a detailed exploitation of a few distinctive characterics, due to the lifetime of the tau lepton and its mass. We use the classification power of a Graph Neural Network (GNN) in order to select tau candidates with three-prong decay.

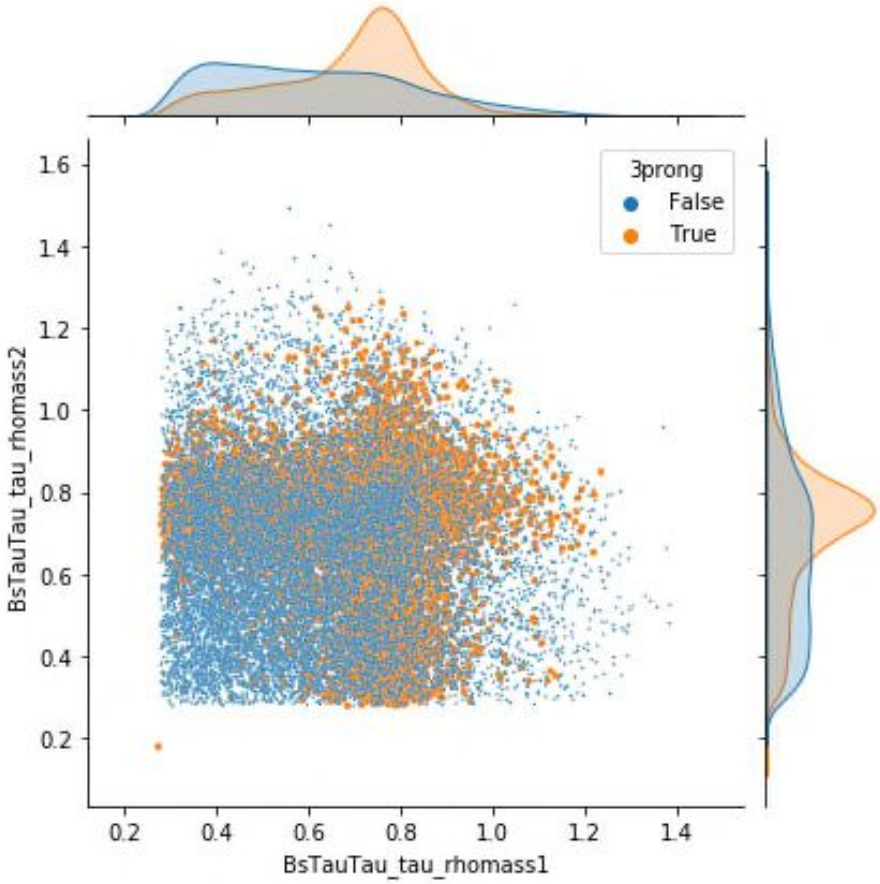


FIGURE 4.6: Scatterplot of the invariant mass [GeV] of oppositely charged two-pion combinations from pion triplets produced in the decay of tau leptons.

Generalities of the pion-selection DNN

The DNN is an implementation of an Attention Based Cloud Network (ABCNet) [74], a graph-based neural network (GNN). GNNs have the advantage of being permutation-invariant over sequential models; they allow us to work with a variable number of tracks. Attention layers added to the architecture further improve the feature extraction and the overall performance. Table 4.2 below lists the features exploited by the network to classify pions originating from semi-hadronic 3-prong tau lepton decays.

The elements of the GNN are graphs which are defined by nodes and edges. We can think of the nodes as the charged hadrons and the edges as the distances between them (connectors). The $\eta - \phi$ distance between reconstructed charged hadrons is used to construct a GAP layer [75], where each hadron is connected to 10 neighbors. The output of this layer is fed into fully-connected hidden layers, which are tasked

Variable	Description
η	Particle pseudorapidity
ϕ	Particle azimuthal angle
$\log p_T$	Logarithm of particle transverse momentum
Q	Particle charge in electron charge units
$\log E$	Logarithm of particle energy
d_z	Distance along z axis from particle vertex to event PV
d_{xy}	Distance in xy plane between particle vertex and event PV
isMuon	Boolean descriptor of whether the particle is flagged as a GlobalMuon
NMUONS	Number of global muons in the event
NPART	Total number of charged hadrons divided by 100

TABLE 4.2: Features used by the ABCNet classifier

with learning the features extracted by the first GAP layer and expand the dimensionality of the feature space. The newly created features are fed into another GAP layer, which accounts for the distance between hadrons in the full space. Other features that describe the event but are not specific of the hadrons, such as the number of muons in the event, are introduced in another fully-connected layer. Pooling and dropout are then operated to minimize over-fitting to the data. The final step of the architecture is constituted by a softmax operator [76], which provides an estimate of the probability that the charged hadron is a pion from tau lepton decay (see Fig. 4.7).

The model is trained with a small 20,000-event Monte Carlo sample, which features semi-hadronic tau decays into three charged pions, where all reconstructed pions are matched to generator-level ones.

For the evaluation of the model we make use of the efficiency of the DNN to correctly select pions from tau decays, and the rate of selected pions that do not originate from the decay. An optimization study suggests the working point of 80% efficiency of retaining the true pion and 10% fake rate, which is the result of a cut at the DNN output value of 0.1443 (see Table 4.3). The Receiver Operating Characteristic (ROC) curve, which describes the false positive rate (or more practically its inverse, for better clarity) as a function of the true positive rate, indeed displays that to an 80% efficiency for the signal corresponds a 10% fake rate in Table 4.3.

The regression task needs training and validation data consisting of true 3-prong decaying taus, which we get from the MC sample, where the pions reconstructed by the GNN are gen-matched and belong to a 3-prong decay of an existing tau particle.

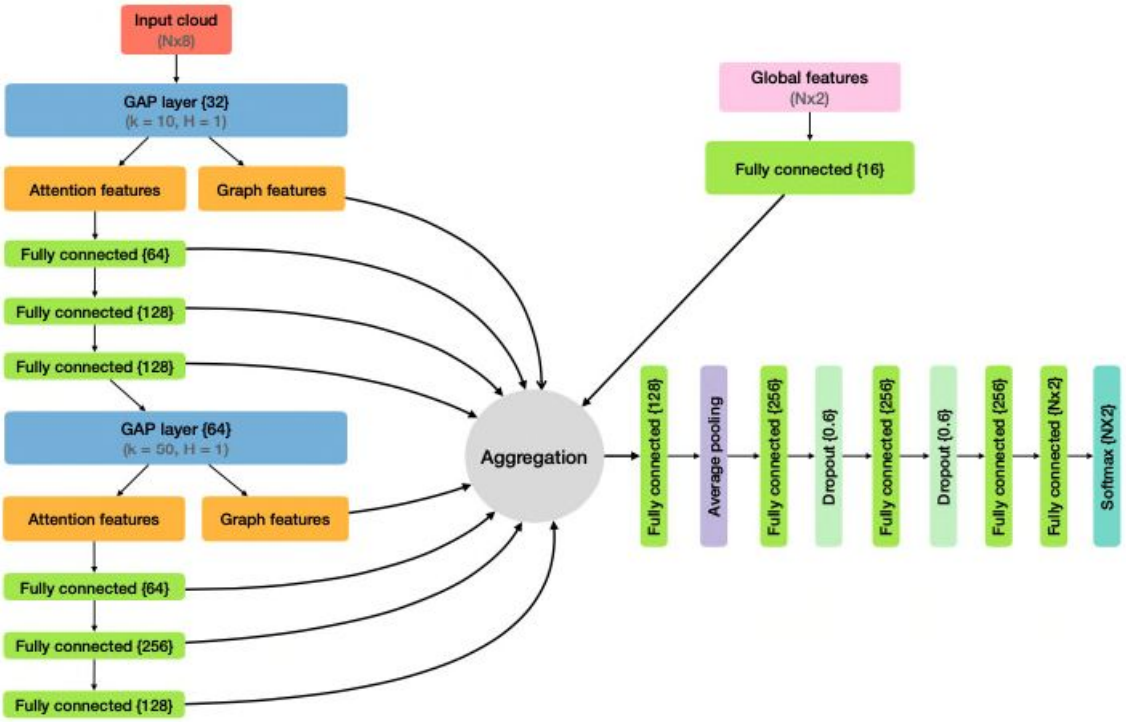


FIGURE 4.7: Scheme of the ABCNet architecture [74]. See the text for details of the blocks functionalities.

DNN cut value	efficiency of retaining true pion	fake rate
0.7113	0.60	0.029
0.6000	0.65	0.042
0.4617	0.70	0.058
0.2968	0.75	0.082
0.1443	0.80	0.116
0.0526	0.85	0.164
0.0131	0.90	0.237
0.0012	0.95	0.393

TABLE 4.3: GNN cut working points

Since the combined branching ratio of the muonic and 3-prong hadronic decay of tau leptons amounts to 1.62%, the expected number of such reconstructed taus is relatively small.

4.4 Regression to B_s Mass

The decay topology we are interested in includes a muon from a leptonic tau decay $\tau \to \nu_\tau \mu \nu_\mu$, whose presence among the reconstructed particles most of the times allows for the triggering of the event, and three charged pions from the semi-hadronic decay of the other tau, $\tau \to \pi\pi\pi\nu_\tau$. In our attempt to reconstruct the decay $B_s \to \tau^+\tau^-$ into this mixed final state, we are challenged by the escape of at least three neutrinos, two of which are produced in the $\tau \to \nu_\tau \mu \nu_\mu$ decay, and the third comes fom the semi-hadronic decay. A machine-learning algorithm may partially recover that lost information from the measured four-momenta of the visible particles. In this section we describe the construction of a regressor that addresses the reconstruction of the four-momentum of the semi-hadronic decaying tau, and the subsequent combination of the obtained information with the muon four-momentum in the higher-level task of estimating the four-momentum of the originating B_s meson.

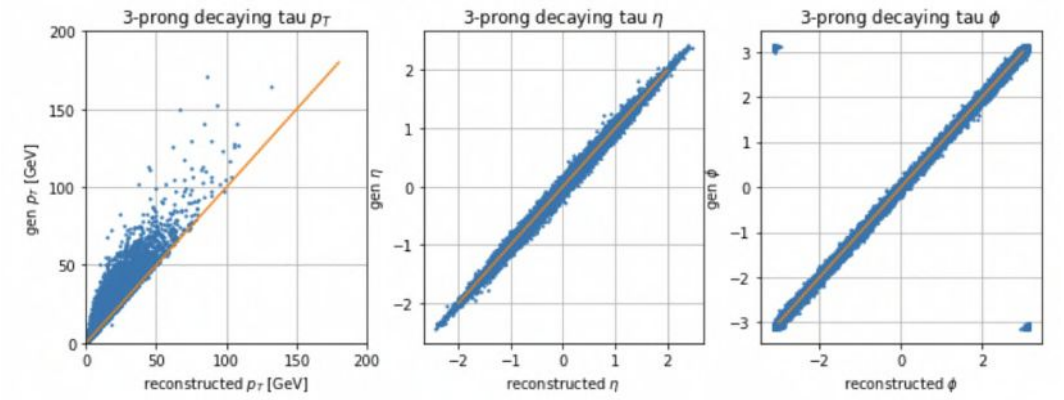


FIGURE 4.8: 2D scatterplots of generation level transverse momentum p_T, η and ϕ values vs the reconstructed quantities. The orange line in the left graph is meant to guide the eye and compare the two plotted quantities.

As depicted in Fig. 4.8, while for three-pion combinations matched to correct generator level particles we can reconstruct quite well the rapidity η and azimuth ϕ of the tau lepton, the same cannot be said of the tau transverse momentum p_T, which is affected by a negative bias due to the missing neutrino. A regression may recover the lost information in part, and nullify the bias.

4.4.1 Initial considerations

The reconstruction of the event triggered by a muon yields several variables belonging to the physics objects, as seen in Table 4.4. These are the features we use in the regression task. The targets are the 3-prong decaying tau momentum and the $\eta - \phi$ coordinates (see Table 4.5).

Variable	Description
$\pi_{\eta,i=1,2,3}$	pion pseudorapidity
$\pi_{\phi,i}$	pion azimuthal angle
$\pi_{p_{T},i}$	pion transverse momentum
$\pi_{Q,i}$	pion charge in electron charge units
τ_Q	tau charge in electron charge units
τ_η	tau pseudorapidity (visible)
τ_ϕ	tau azimuthal angle (visible)
τ_{p_T}	tau transverse momentum (visible)
τ_α	cosine of the opening angle between primary (PV) and secondary vertex (SV) line and the B_s momentum
$\tau_{dr,i}$	ΔR distance of tau from each pion
τ_{pvip}	primary vertex impact parameter
τ_{fl3d}	3d flight length distance
τ_{vprob}	vertex probability
$\tau_{rhomass1}$	invariant mass of 2 opposite charged pion combination
$\tau_{rhomass2}$	invariant mass of 2 opposite charged pion combination

TABLE 4.4: Reconstructed variables used for the regression task

Variable	Description
τ_η	tau pseudorapidity
$\tau_{\cos(\phi)}$	cosine of tau azimuthal angle
$\tau_{\sin(\phi)}$	sine of tau azimuthal angle
τ_{p_T}	tau transverse momentum

TABLE 4.5: Target variables for the regression task

The regression models need to be accurate as well as interpretable, so the first step is to understand which of the variables are important for the prediction. Once we see which features are significant, we can identify as redundant and remove the ones with least importance, in order to achieve a shorter training time without affecting the overall performance; this step needs to be re-done every time for new data samples to check the effect on the model performance. For this initial study the gradient

boosted decision tree algorithm (GBDT) [77] is employed, which is described in detail in section 3.1.1.

A decision tree algorithm is one that asks iterative questions to partition the data. On its own a decision tree is very prone to overfitting, so we combine the individual trees to get a better performance. One way of combining the trees is by the procedure called boosting, where we build a strong learner by stacking the individual trees (weak learners) sequentially. Each tree focuses on the minimization of the previous tree's error. For a gradient boosted decision tree algorithm this is done by fitting to the residual of the previous tree, where the residual is calculated via a loss function, mean squared error (MSE) or square root of mean squared error (RMSE) for the regression task. The most important hyperparameters for the tuning of such an algorithm are the learning rate (a measure of modification per tree, which determines how fast the model learns) and the number of trees. If the learning rate is too low, the model will train too slowly; if instead it is too high, the learning might not converge to a minimum loss. If the number of trees are too high, the GBDT model will instead start to overfit. Overfitting is not *per se* undesirable, but it may reduce the generalization properties of the learned model.

Following this study we can drop variables such as individial pion charges and pion-selection DNN scores since the information they contribute is redundant. Furthermore, we studied the effect of a coordinate change from (p_T, η, ϕ) space to cartesian coordinates (p_x, p_y, p_z for each particle) and we concluded that the performance of the two approaches is comparable.

4.4.2 Neural Network regressors

After the initial studies with the GBDT, which allowed us to identify the important features in the data, we use Neural Networks (NN) to perform the two separate regression tasks. We found that the overall performance of NN regressors is similar to the previously discussed GBDT. However, in a GBDT algorithm the weights are updated sequentially, whereas in an NN algorithm, each feature is fed to the model in parallel and each feature weight has a unique path in the back-propagation process. We study the effect of this implementation when combining the two regression

models. The baseline NN is a fully connected dense NN, with dropout and pooling operations to minimize overfitting.

We split the (shuffled) data in two ways for a training/validation/test cycle, with ratios of 50:40:10 and 80:10:10, where we have 50% (80%) of data for training, 40% (10%) of the data for validation, and the remaining 10% for testing our model in order to minimize the overfitting behaviour of the model.

The cyclic nature of the azimuthal angle is preserved by regressing to the cosine and sine of the variables (see Table 4.5). In this way the model is kept aware of the cyclicity even after rescaling the variable. Finally, the stability of the model is studied with a standard 5-fold cross-validation scheme, where we split the training data into five parts, train the model on the n-fold and check the performance on the validation dataset.

4.4.3 Semi-hadronic tau regression results

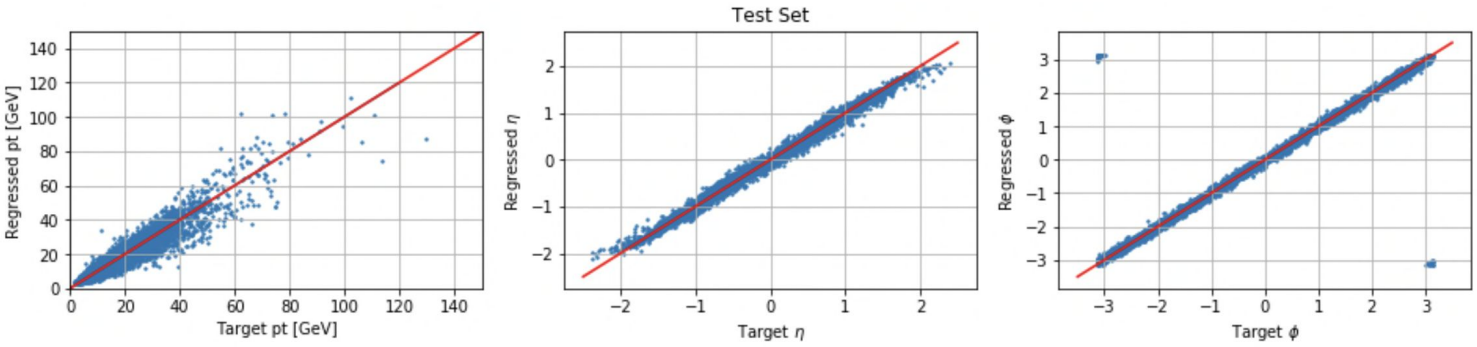


FIGURE 4.9: 2D scatterplots of regressed tau (y-axes) p_T , η and ϕ values vs the corresponding generated tau attributes (x-axes). The model is trained and validated on pions that are matched to the generator-level decay products of real taus.

The model regresses to the semi-hadronic tau transverse momentum p_T, pseudo-rapidity η and the azimuth angle ϕ. Figure 4.9 shows the predicted target variables as a function of the generator-level quantities with two different train/validation splits. A perfect regressor would place all the data points on the orange diagonal line. With a larger training data sample the model performs better and this is valid for all the regression models performing on the MC samples we have used. The model suffers from the missing knowledge about the three-prong decay due to the

non-observed neutrino and this is evident in the transverse momentum regression plot in Fig. 4.9.

B_s mass estimation	
τ_1	τ_2
gen-mass m_1	gen-mass m_2
regressed p_T, η, ϕ	gen. p_T, η, ϕ
energy $E_1 = \sqrt{m_1^2 + p_{regressed}^2}$	$E_2 = \sqrt{m_2^2 + p_{gen}^2}$
$M^2 = m_1^2 + m_2^2 + 2 \cdot (E_1 \cdot E_2 - \vec{p_{regressed}} \cdot \vec{p_{gen}})$	

TABLE 4.6: Generated or regressed tau attributes used in B_s mass estimation

Using the attributes of both taus listed in Table 4.6 we estimate the B_s mass (see Fig. 4.11) that is centered at a mean $\mu = 5.31$ GeV (with a gaussian fit $\sigma = 0.52$, purple curve); as reference, the true mass of the B_s meson is $M_{B_s} = 5.36$ GeV. Since at this step we are using generator level features for the tau which decays into a muon and two neutrinos in the calculation, this must be considered an upper limit to the precision of the B_s mass reconstruction.

4.4.4 Full B_s reconstruction

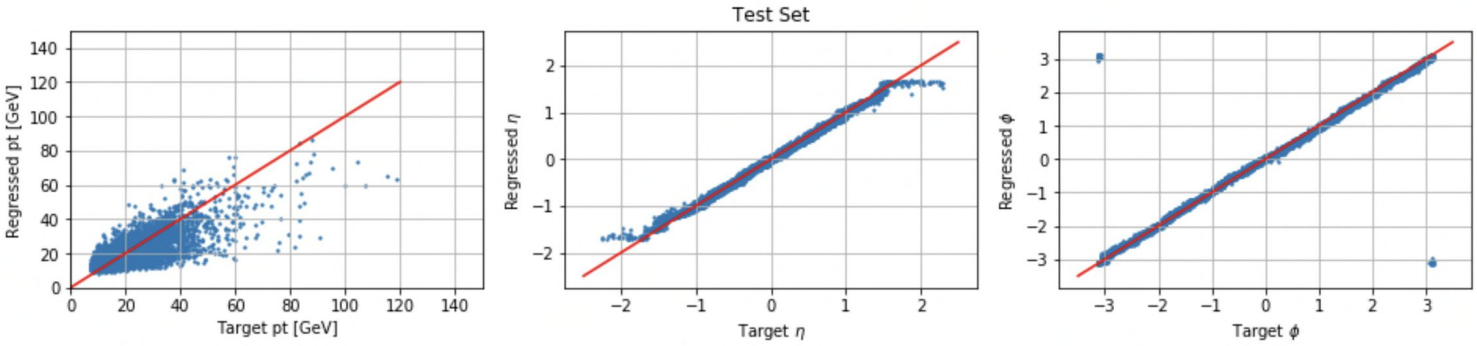


FIGURE 4.10: 2D scatter plots of the regressed tau (y-axes) p_T , η and ϕ values, which decays into a muon and two neutrinos vs the corresponding generated tau attributes (x-axes). The model is trained and validated on pions that are matched to the generator-level decay products of semi-hadronically decaying taus (40% validation split)

We now consider the other tau lepton, which decays into a muon and two neutrinos, and we attempt a regression to its attributes. This is a more challenging task for a learning algorithm than the semi-hadronic tau regression, since we have two neutrinos escaping our observation. The model performs poorer for p_T regression

(Figure 4.10), when compared to the previous result. However, we note that, most of the features are the "visible" attributes of the muon only for this regression task. Pseudorapidity η regression fails at around $|\eta| = 1.5$, this is due a cut on the trigger muon p_T at 7 GeV (the loosest trigger path for p_T has the cut at 7 GeV). Pseudorapidity η distribution of trigger muon (hence η distribution of the muon decaying tau) has slim shoulders at around $|\eta| = 1.5$. Due to the lack of statistics above $|\eta| = 1.5$, the regressor performs poorly.

Furthermore, we avoid using features from the semi-hadronic decay because this will lead to a B_s mass sculpting in the background events, where we will lose the discrimination power in the signal region around 5.36 GeV.

B_s mass estimation	
τ_1	τ_2
gen-mass m_1	gen-mass m_2
regressed p_T, η, ϕ	regressed p_T, η, ϕ
energy $E_1 = \sqrt{m_1^2 + p^2_{regressed,1}}$	$E_2 = \sqrt{m_2^2 + p^2_{regressed,2}}$
$M^2 = m_1^2 + m_2^2 + 2 \cdot (E_1 \cdot E_2 - \vec{p_{regressed,1}} \cdot \vec{p_{regressed,2}})$	

TABLE 4.7: Generated or regressed tau attributes used in B_s mass estimation

As previously done, we estimate the B_s mass using the quantities listed in Table 4.6 (switching the taus), where we have regressed transverse momentum p_T, pseudorapidity η and azimuth ϕ variables for the muon decaying tau and the gen attributes of the 3-prong decaying tau. Figure 4.11 shows the resulting B_s mass distribution, which is centered at a mean $\mu = 5.48$ GeV (with a gaussian fit $\sigma = 0.72$, red curve).

Finally we estimate the B_s mass using the quantities listed in Table 4.7, where the regressed attributes for both taus are plugged in. Figure 4.11 shows the resulting B_s mass distribution, which is centered at a mean $\mu = 5.40$ GeV (with a gaussian fit $\sigma = 0.78$, brown curve).

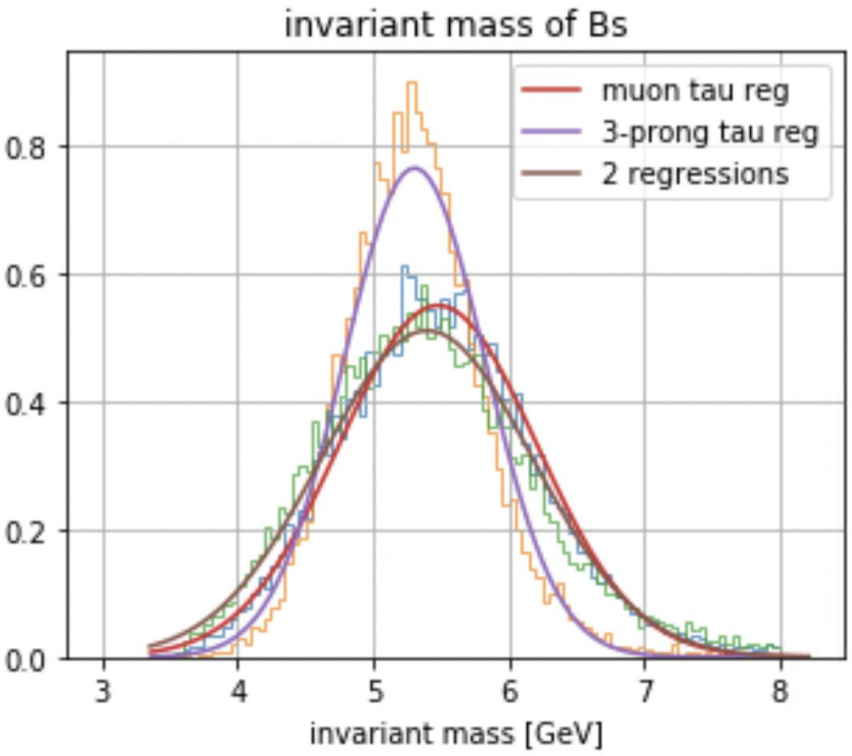


FIGURE 4.11: Histograms of the B_s mass estimated using the attributes predicted by the regressors on both tau candidates (yellow, blue and green lines). Red curve represents the gaussian fit to the distribution where we regress to the muon-decaying tau attributes and use the gen attributes of the 3-prong tau to estimate the B_s mass. Purple curve represents the opposite case, whereas the brown curve represents the gaussian fit to estimations computed with regressed attributes for both taus.

4.4.5 B-Parking data from LHC Run 2

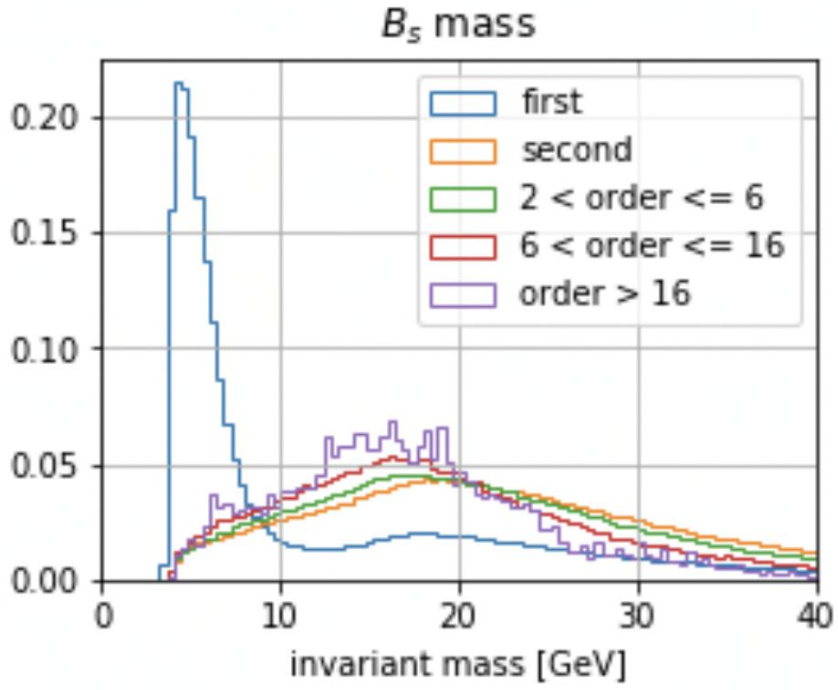


FIGURE 4.12: Histogram of the B_s mass estimated using the attributes predicted by the regressions on B-Parked data sample. In each event there are multiple tau candidates for the 3-prong decaying tau and are ordered by p_T. Blue curve shows the study with the first most energetic tau candidate, while the other curves represent the rest of the candidates in p_T order.

Given the expected rarity of the $B_s \to \tau\tau$ signal we may quite safely use real data as a representative sample of background events. This will not impair or bias our analysis choices until a stringent data selection is applied, targeting the highest increase in signal to noise. This is because, not only does the signal (possibly) present in the data does not affect our results, but also the possibility to perform a "blind analysis" on the final signal extraction remains possible. The observation of the behavior of data following a reconstruction of events according to the B_s hypothesis allows us to draw some conclusions on the effectiveness of the regression and to give us information to improve it for the subsequent stages of the analysis, still to be undertaken.

Figure 4.12 shows the distribution of the B_s mass estimated for the B-Parked data sample. After the reconstruction steps employed on the dataset (described in Sec. 4.3), which give us the pion candidates of a three-prong decaying tau from the

$B_s \to \tau\tau$ decay, we utilize the two regressors, which are trained and validated on the MC samples. The resulting B_s mass peak for the blue curve (most energetic tau candidates in the events which decays into three pions) is centered at a lower region than the signal region around 5.36 GeV, whereas the B_s mass estimated with the rest of the tau candidates in the event result in a higher invariant mass centered around 15-19 GeV. Since the phase space of the four considered particles is similar for both regressors (the muon distribution is driven by the trigger requirements, and pions from tau leptons are not very atypical in their momenta in the B-parked data sample) the discrimination power of the reconstructed B_s mass hypothesis is not as large as one could hope around the B_s mass of 5.36 GeV. Further studies of higher-level kinematic features that account for the similarity of background and signal may help provide further discrimination power from the kinematic characteristics of the observed particles.

4.5 Classification

For this study Boosted Decision Trees (BDT) are employed to distinguish between signal and background event candidates. For the training, $B_s \to \tau\tau$ MC simulation samples where one tau decays $\tau \to \nu_\tau \mu \nu_\mu$ and other tau decays to three charged pions ($\tau \to \pi\pi\pi\nu_\tau$) are used for the signal, and event candidates from B-parked data are used for the background. We expect negligible signal events in the real data, so we treat a part of the real data as background and study its discrimination from the signal. To avoid any selection bias, the signal and background events are randomly split into two sets, such that the training and testing of the BDT becomes independent to each other. We avoid the BDT to be sensitive to B_s mass, as we plan to use B_s mass as the final discriminating variable. However, some variables may be correlated to B_s mass, so we use a *planing* [78] approach to make BDT insensitive to B_s mass by re-weighting the background dataset in a way that B_s mass distribution is similar for both MC and the B-Parked data.

4.5.1 Input variables

For each BDT, a selection of variables is considered, out of which twenty variables are found to be effective and uncorrelated. Final input variables are shown in Table 4.8.

Variable	Description
τ_{mass}	reconstructed tau mass
$\tau_{rhomass1}$	invariant mass of 2 opposite charged pion combination
$\tau_{rhomass2}$	invariant mass of 2 opposite charged pion combination
τ_{α}	cosine of the opening angle between primary (PV) and secondary vertex (SV) line and the B_s momentum
$\tau_{dr_{\mu}}$	ΔR distance of tau from muon
τ_{vprob}	vertex probability of tau
τ_{alpha}	Cosine angle between 3 momentum vector of tau and position vector of tau w.r.t. vertex.
B_{alpha}	Cosine angle between 3 momentum vector of B_s and position vector of B_s w.r.t. vertex.
$\pi_{dnn,i=1,2,3}$	pion dnn
τ_{p_T}	tau transverse momentum (visible)
μ_{p_T}	muon transverse momentum (visible)
μ_{η}	muon pseudorapidity
E_{MET}	missing energy in the transverse direction
B_{iso}	isolation variable of B_s
$d\phi_{E_{\mu}}$	$\Delta\phi$ between transverse missing energy and muon

TABLE 4.8: Reconstructed variables used in training the BDTs

4.5.2 Hyperparameters

We tuned the hyperparameters of the BDT to get the maximum discrimination power; these include the number of trees set to 1000 with 2% minimum node size, maximum depth of 3, a boost type grad, the learning rate or shrinkage set to 0.005, the bagged sample fraction at 0.6 and the separation type Gini index.

4.5.3 BDT performance

We employ three different BDTs; for the training of each BDT $B_s \rightarrow \tau\tau$ MC samples where one tau decays $\tau \rightarrow \nu_{\tau}\mu\nu_{\mu}$ and other tau decays to three charged pions ($\tau \rightarrow \pi\pi\pi\nu_{\tau}$) are used as signal, whereas three different kinds of background event candidates are used per BDT. The output response curve for the BDTs are shown in the figure 4.14. Firstly, we use the combinatorial background and we get an area

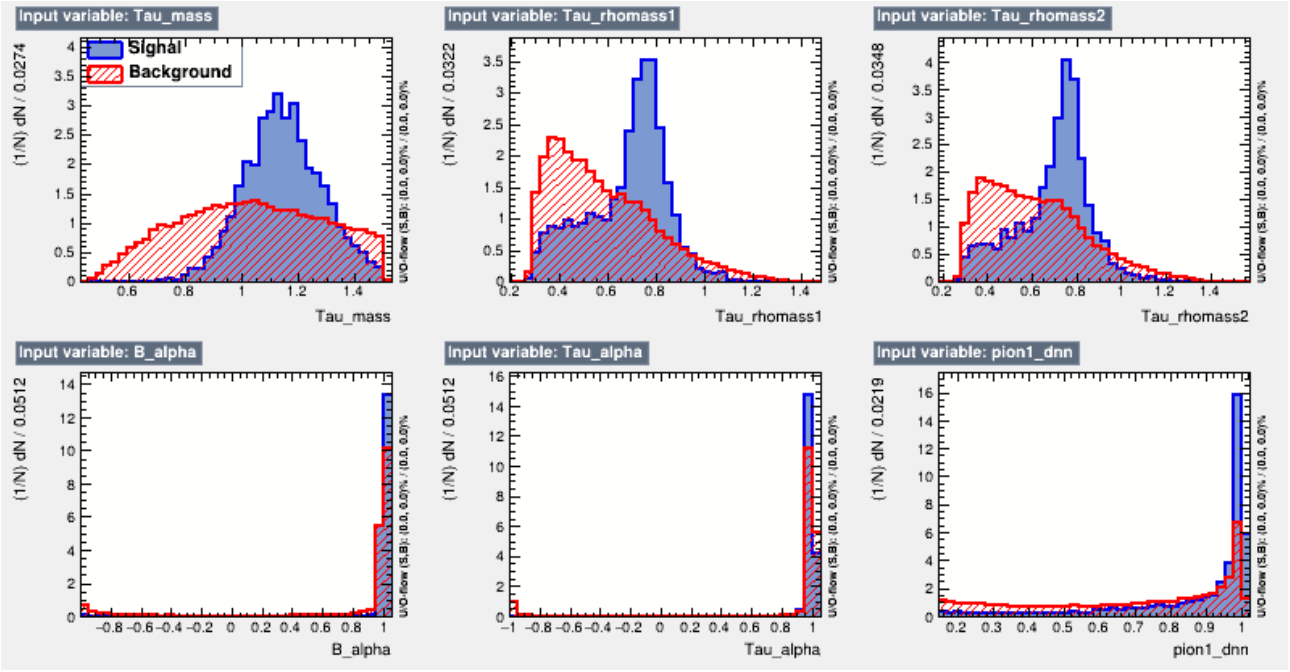
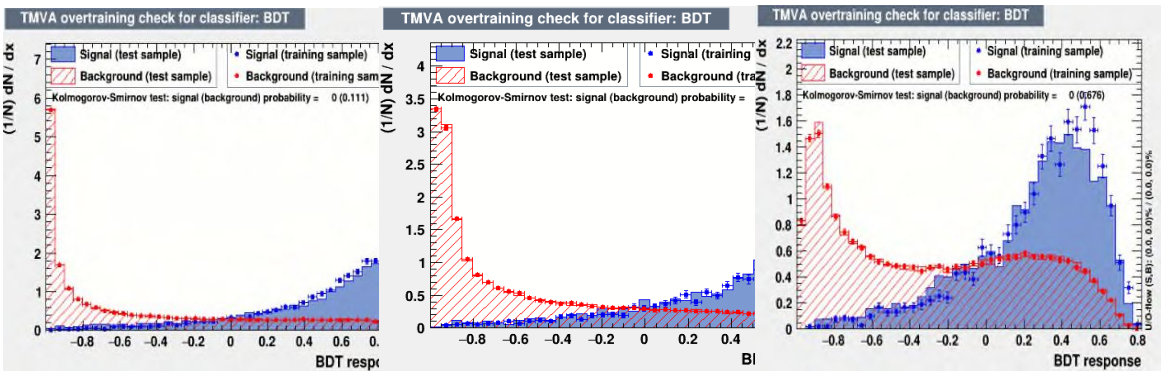


FIGURE 4.13: Selection of input variables for the BDT classifier

under ROC curve at 0.870, the BDT response curve for this is shown in figure 4.14a. Secondly, since we expect negligible signal events in data, we use a part of data as background. We get an area under ROC curve of 0.899, the BDT response curve for this is shown in figure 4.14b. Lastly, we implement a "planning" method where we re-weight the combinatorial background distribution to mimic the signal distribution such that B_s mass distribution becomes comparable for signal and background. In this case we get an area under ROC curve of 0.790, response curve is shown in figure 4.14c. After the implementation of planning, the BDT becomes insensitive to B_s mass.



(A) (B) (C)

FIGURE 4.14: Response curve for BDTs

4.6 Semi-supervised Search with Ranbox

When a signal of known kinematic characteristics is sought, but a precise model of the background is not available, the problem lays in middle ground between one of anomaly detection and one of a classical supervised learning discrimination. Several methods have been proposed to allow the construction of good discriminators in this situation; for a review see [78].

Here we discuss how `RanBox` can be adapted to this task, which is of interest to us because of its good match with the needs of the search for the rare $B_s \to \tau\tau$ decay in CMS proton-proton collision data. Indeed, the dataset where we wish to carry out the search is a very complex one, contributed by many different low-p_T processes that Monte Carlo simulations cannot reproduce in high detail. On the contrary, the B_s production and decay mechanisms are well understood and can reliably be simulated.

For the study of the semi-supervised version of `RanBox`, in order to avoid the risk of unconsciously biasing one's decisions and analysis choices, we refrain from performing the tests on real data, and rely instead on the HEPMASS dataset for the prototyping phase of the algorithm. By introducing weak supervision, we exploit our knowledge of the signal portion of the data, to identify a box in multi-dimensional space which is both rich in signal, and poor in background. We act as if the latter information is not available *a priori*, i.e. from a precise model of that process, and only rely on the local background density in sidebands of the search box in the construction of a useful test statistic.

The decision to not exploit the full labels of available data in the same phase space, but use label information for the two classes in different phase space regions, effectively corresponds to a semi-supervised task. It also allows one to remain blind to the real amount of data captured in the search box that the algorithm identifies at the end of the gradient descent procedure that maximizes the test statistic of choice. This enables a data-driven background prediction and avoids biases introduced unconsciously by the analysts, as the procedure is fully automated. However, as we well discuss below, a bias remains in the background estimation, due to the intrinsic correlations between the variables of the feature space, and the similarity of these

correlations between the signal and background processes.

4.6.1 Algorithm description

The semi-supervised version of `RanBox`, which we address in the following as `RanBox_SS`, works as follows.

1. Real data where the search is to be performed is read in. We will call this dataset D1 in the following. For the tests described in this report, D1 is composed of HEPMASS simulated events belonging to the background category, but for specific studies of the performance of the algorithm we include in D1 a fraction of signal simulated events.

2. Simulated signal events are read in a dataset called D2.

3. A pruning of the variable list removes ones that are categorical, as well as ones that have too high correlation with others, using the same procedure as that described in Sec. 2.

4. Dataset D1 is used to define a variable transformation that reduces the feature space into a copula space, as in the original version of `RanBox`.

5. Events from dataset D1 and dataset D2 are both subjected to the same variable transformation. This produces transformed datasets spanning the copula feature space, where the search for the box maximizing a suitable test statistic Z is performed.

6. A test statistic is defined to maximize the search power.

7. A number of dimensions of the subspaces scanned by the algorithm is chosen. This number, $\mathcal{D}'$, should be not larger than 10-12, because of the curse of dimensionality.

8. $\mathcal{D}'$ features are chosen at random from the list of active dimensions of the space.

9. The scan of the corresponding $\mathcal{D}'$-dimensional subspace of the transformed feature space is performed by maximizing the test statistic Z via gradient descent.

10. The previous two steps are repeated a large number of times. At the end of the iteration, the algorithm reports the boundaries of the box that maximizes Z across all subspaces, and related statistics.

A number of details need to be discussed in order to clarify the above procedure. We address them below.

Variable transformation

Because the variable transformation, based on the integral transform, is defined on dataset D1 but is then applied to both datasets D1 and D2, a prescription needs to be given for how to handle variable values which fall outside of the original range of their distribution in dataset D1. For example, it might happen that dataset D1 only contains events for which variable 1 is in the $[-100., 500.]$ range, but dataset D2 has a distribution of variable 1 which extends in the wider range $[-150., 600.]$. This might happen if the signal process produces observable features that are exceedingly rare in backgrounds. Since the range $[-100., 500.]$ is mapped by the integral transform into $[0, 1]$, one needs to define transformed values for variable 1 in the range $[-150., -100.]$ or $[500., 600.]$. This is simply done by assigning transformed value 0 to all values below $-100.$ and transformed value 1 to all values of variable 1 above 500.. This ensures that the copula space remains unaltered when the signal component is considered. However, one must keep note of the fact that such a situation complicates any extrapolation of the data density from sidebands to signal region, when the signal region includes the singular points at 0. or 1: a bias in the sidebands-driven background prediction can be expected in these situations.

Test statistic

Our focus in this work is to produce a suitable methodology that can be applied to the search for a rare process that we do not expect will be observable in the available data. In fact, current estimates for the branching fraction of the B_s meson decay to tau lepton pairs put this value below 10^{-7}, which corresponds to less than one

event in the available LHC luminosity once one accounts for the unavoidably small collection efficiency for identifiable signal events[4].

Because of the above, it makes little sense to maximize a test statistic which is monotonous with the significance of a signal component in the data. Rather, it appears reasonable to define a test statistic which minimizes the upper limit on the signal process under study. This is in fact the reachable scientific objective of the B_s search under way by the CMS collaboration in currently available data.

We therefore consider a counting experiment where a predicted background N_{exp} is compared to an observed event count N_{obs}, the former extracted from a sideband of equal feature space volume to the signal box, possibly after a bias correction (see *infra*). The observed event counts N_{obs} is expected to be sampled from the background prediction, i.e. from a Poisson distribution centered on N_{exp}. If we define a type-I error rate $\alpha = 0.05$, we can compute the expected upper limit at a confidence level $1 - \alpha$ on the number of signal events contributing to the signal box, when N_{exp} is predicted and N_{obs} is observed. An exact calculation involves extracting the tail integral of a Poisson distribution, and is sometimes impractical to perform (when N_{exp} is large). `RanBox_SS` uses a very good approximation given by the 95% C.L. upper limit on the signal, N^{up} :

$$N^{up} = \frac{1}{2} F^{-1}_{\chi^2}(1-\alpha) - N_{exp} \tag{4.1}$$

where F_{χ^2} is the cumulative distribution function of Chi-squared distribution with $2(N_{obs}+1)$ degrees of freedom. We may then convert the upper limit above into an absolute upper limit on the cross section times branching fraction of the studied process by the following formula:

$$\sigma B(B_s \to \tau\tau) = \frac{N^{up}}{L\epsilon_{tot}},\ 95\%\ C.L. \tag{4.2}$$

[4]The data where the search is performed is collected by a trigger which selects events with a muon candidate of transverse momentum above 7-9 GeV (depending on data taking period and running conditions). Once one accounts for the braching ratio of tau decays into muons, the small probability of this giving rise to a muon above the stated momentum threshold, and the additional requirements on the other tau lepton decay, which must include three charged pions, the signal efficiency becomes smaller than 10^{-3}; together with a 10^{-7} $B_s \to \tau\tau$ branching fraction, this means that one reconstructable event corresponds to over 10 billion B_s mesons, which correspond to an integrated luminosity much larger than what is available for the present search.

Above, the signal efficiency ϵ_{tot} is computed by multiplying a pre-selection efficiency ϵ_{presel} (obtained from the Monte Carlo simulation) that describes the probability that a $B_s \to \tau\tau$ signal event gets included in the analyzed sample, by the fraction of events captured in the signal box ϵ_{box}.

The procedure to define the test statistic that we wish to maximize therefore is to extract from a signal box the number of events N_{sig} in dataset D2 that are contained within it, and to assess the number of expected background events N_{exp} in the same signal box, by counting how many D1 events are collected in a properly defined sideband. Using N_{exp} we obtain, by the formula (4.1), the 95% C.L. upper limit on the signal, N^{up}; with N_{sig} we compute the signal efficiency ϵ_{box}; and with these inputs we may then define the test statistic to be maximized as

$$Z_{UL} = \frac{\epsilon_{box}}{N^{up}} \tag{4.3}$$

whose maximization explicitly minimizes the upper limit on the signal cross section times branching fraction.

It is clear what an algorithm tasked with maximizing Z will need to do: find a region of space which contains a large number of signal Monte Carlo events from dataset D2, while having as small as possible predicted contributions from dataset D1. The latter comes from a "non-local" estimate, one derived from a sideband constructed exactly as described for `RanBox` in Sec. 2.

Gradient Descent

The gradient descent procedure used by `RanBox_SS` when studying each subspace of the copula is the same of that of `RanBox`, and it has been already described in Sec. 3. However, due to the fact that the results of `RanBox_SS` are more affected by a biased background prediction, we implement for it a validation technique based on an early stopping criterion. Datasets D1 and D2 are both split evenly into a training and a validation subset, and only the training subset is used for the maximization of the test statistic by the gradient descent procedure. During the procedure, however, the algorithm keeps track of the value of the test statistic on the validation subset of datasets D1 and D2. At the end of the routine, the box which produces the highest

value of the test statistic on the validation sample is returned as the best one for the considered subspace.

Box boundaries and sidebands

It should be clear that the initialization of the search box boundaries must be driven by the density that can be assessed from the part of the data from which a local density estimate can be obtained, *i.e.* dataset D2, the signal Monte Carlo. Algorithm 2, described in Sec. 2.5 above, can fulfil that task effectively, and is used for `RanBox_SS`.

The choice we made for `RanBox_SS` of using a sideband of volume equal to the signal box to estimate the background in the signal region is dictated by the need of an estimate affected by low bias. Indeed, bias is a much worse enemy than variance in this particular application, as large correlations between the variables of the feature space have the potential of making any sideband-derived estimate completely unreliable. While non-local, a sideband estimate from events which lie very close to the signal box will suffer a manageable bias even in the presence of large correlations. However, the gradient descent procedure which maximizes the test statistic defined above explicitly tries to shrink the value of N_{exp}, by moving to regions where dataset D1 suffers negative fluctuations. Hence a strong negative bias on that number is anyway expected. We sidestep this problem by constructing a second sideband around the sideband used for the calculation of the test statistic. This second sideband is only used for the final estimate of N_{exp}, and should be unaffected by the gradient descent procedure.

At variance with `RanBox`, in `RanBox_SS` we enforce that sidebands (as well as the second sidebands described below) have a volume exactly equal to that of the signal box. This is a useful property when we need to characterize possible biases in the extrapolation procedure, as the extrapolation factor is always equal to 1.0 and thus is one less parameter to consider in such bias studies. In order to enforce that the sideband has a volume exactly equal to the signal box, we devise an iterative algorithm, described below.

1. The widening factor required to construct a box of volume twice larger than that of the signal box is computed as $f = 2^{1/\mathcal{D}'}$, such that if each side of the

signal box were widened by a factor f, the resulting box would have a volume equal to twice the signal box.

2. A loop on the subspace dimensions is performed, and for each dimension the signal box extension B_i and the available region A_i left in the $[0,1]$ interval once the signal box interval is excluded is stored. So, *e.g.*, if the signal box has an extension of $B_1 = 0.3$, having intervals $[0.2, 0.5]$ in variable 1, the available region on variable 1 is $A_1 = 1. - 0.3 = 0.7$.

3. Available region values are sorted in increasing order.

4. Starting with the smallest available region, the algorithm assigns sideband intervals to each variable i by comparing A_i to $f \times B_i$. If A_i is larger than $f \times B_i$, the interval defining the sideband in dimension i is simply defined by extending the box interval by a factor f (an attempt to construct a symmetric interval is made, and if there is not enough space in one of the sides of the signal box, all the extra space required to extend the interval to $f \times B_i$ is assigned to the sideband on the other side). If, on the other hand, A_i is smaller than $f \times B_i$, the sideband on direction i is defined as $[0,1]$, and the factor f required to each additional variable to obtain a sideband of volume twice larger than the signal box is recomputed as $f = (2B_i)^{1/(D'-1)}$. A similar rescaling is operated at each successive iteration until the considered A_i grows larger than the current f value.

5. The iteration on every variable continues, until all dimensions of the subspace have been included in the sideband definition. The procedure converges to a sideband of volume equal to twice the signal box (and thus a surrounding region of volume equal to the signal box, once the signal box is vetoed) unless the signal box has a volume larger than 0.5, which is however not allowed in any step of the program (the initialization algorithms, as well as all box extensions in the gradient descent routine, enforce that the signal box has a volume not larger than 0.25 in `RanBox_SS`).

In exact similarity to the algorithm described above, the second sideband is defined as a multi-dimensional interval in the considered $\mathcal{D}'$-dimensional subspace, of total volume exactly equal to three times the signal box, and thus also equal to 1.5 times the sidebands box. Once events in the second sidebands are vetoed if they are contained within the first sideband, the effective volume of the second sideband is equal to that of the signal box, hence the extrapolation factor required to predict the number of events in the signal box from the second sideband is equal to 1.0. A modification of this factor may be required if an estimate of bias is obtained, as discussed below.

4.6.2 Sample results on the HEPMASS dataset

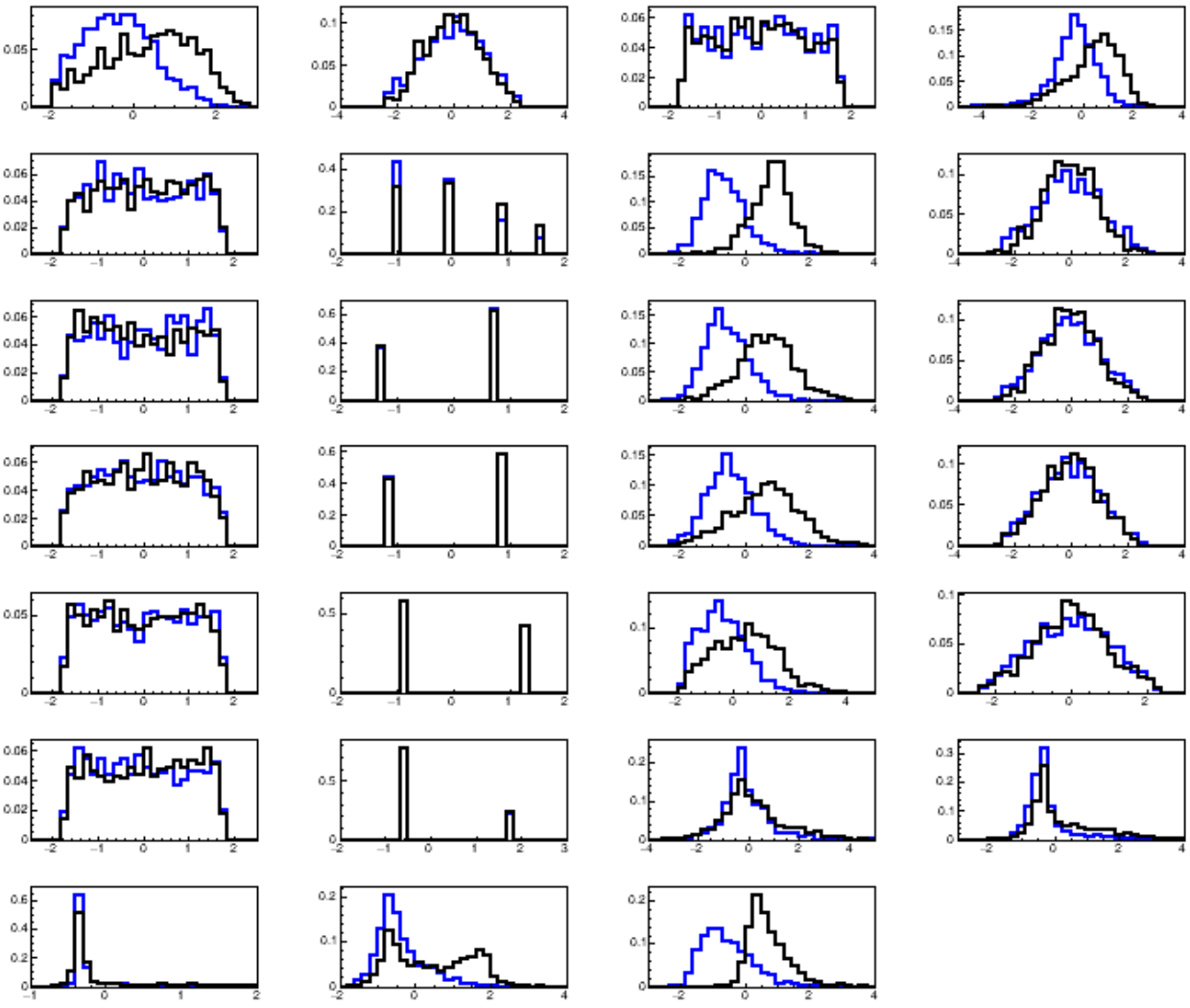


FIGURE 4.15: Normalized and standardized distributions of the 27 features of HEPMASS data for signal (black) and background (blue).

Below we detail results of running `RanBox_SS` on the continuous features of the HEPMASS dataset. As it was shown in the previous sections, the signal in that dataset is rather easy to isolate from backgrounds, due to its distinguishing mass-related features: multi-object invariant masses and energy of objects are all high for the signal component. This is quite different from what will be the case of the B_s search, which unfortunately features a signal very difficult to distinguish from backgrounds. We do not expect any variable to have nearly as strong a discrimination power as the most sensitive variables of the HEPMASS dataset. To make the HEPMASS data a better testbed of our algorithm, therefore, we remove from the list of 27 features not only the five categorical variables it contains (variables 6, 10, 14, 18, 22), which would complicate the preprocessing step of the data, but also six of the most discriminating features: variables 4, 7, 11, 15, 26, and 27 (refer to Fig. 4.15). We are thus left with a set of 16 features, which are in the same ballpark of the dimensionality of the space of discriminating variables we will use for the B_s search, and provide a discrimination problem of similar complexity to the one we are targeting.

To test the working of `RanBox_SS`, a run maximizing the test statistic Z_{UL} is performed on a dataset D1 composed of 5,000 background events, and a dataset D2 composed of 5000 signal events, by searching in 10,000 different subspaces. The results for the 10 best boxes are shown in Table 4.9 below.

Box	Z_{UL}	N_{obs}	N_{exp}	Volume	ϵ_{box}	Features
1	10.57	8	12	0.0024	0.044	0 1 2 4 5 6 8 9 10 11 12 14
2	10.39	17	8	0.0040	0.052	0 2 4 5 7 8 9 11 12 13 14 15
3	10.28	12	13	0.0044	0.059	0 1 2 4 5 6 8 9 10 11 12 15
4	10.13	15	8	0.0038	0.055	0 1 2 4 5 7 8 9 11 12 14 15
5	10.01	12	10	0.0034	0.052	0 1 4 5 6 7 9 10 11 12 14 15
6	9.97	18	5	0.0027	0.044	2 3 4 5 6 7 8 10 11 12 13 14
7	9.97	18	5	0.0027	0.044	1 2 4 5 6 8 10 11 12 13 14 15
8	9.97	18	5	0.0027	0.044	2 3 4 5 6 7 8 9 10 12 13 14
9	9.81	15	8	0.0041	0.056	0 2 3 4 7 8 9 10 11 12 13 14
10	9.77	16	9	0.0037	0.043	2 3 6 7 8 9 10 11 12 13 14 15

TABLE 4.9: *Results of a maximization scan of 5000 subspaces of the HEPMASS feature space, with a D1 dataset composed of 5000 background events, and a D2 dataset made of 5000 signal events. N_{obs} is the number of D1 events in the signal box. The best identified signal boxes are ordered by decreasing value of the Z_{UL} test statistic, whose value is inversely proportional to the estimated 95%CL upper limit on signal cross section achievable by a counting experiment.*

The Z_{UL} values listed in Table 4.9 correspond to upper limits on the signal cross section of $\sigma_{95\%CL} = 1000/(LZ_{UL}\epsilon_{presel})$, where ϵ_{presel} is the fraction of signal events that would be included in the original dataset before the `RanBox_SS` search, and L is the integrated luminosity corresponding to the analyzed data. For a preselection efficiency of $\epsilon_{presel} = 0.1$, *e.g.*, and an integrated luminosity $L = 1fb^{-1}$, the tabulated highest value of Z_{UL} corresponds to an upper limit of $\sigma_{95\%CL} = 946fb$.

A different test is performed by searching in 5,000 subspaces of a dataset D1 composed by 9500 background and 500 signal events, with a corresponding dataset D2 of 10,000 signal events. Having injected signal in D1, we can check the increase in signal purity of the returned boxes, and observe that the upper limit becomes higher, as the maximum value of Z_{UL} reached is smaller. The results for the five best boxes are shown in Table 4.10 below. One observes that the maximization of Z_{UL} corresponds to a significant increase in the signal over background fraction of the selected portion of dataset D1, as shown by the second-to-right column.

Box	Z_{UL}	N_{obs}	N_{exp}	$N_{s,in}$	Volume	ϵ_{box}	S/N gain	Features
1	7.13	59	37	22	0.0085862	0.08	7.45783	0 1 2 3 5 6 7 8 10 11 12 14
2	7.1	59	37	22	0.0085004	0.08	7.45783	0 1 2 3 5 6 7 8 10 11 12 15
3	7.01	61	41	22	0.00924	0.08	7.21331	0 1 2 3 4 5 6 7 8 9 13 15
4	6.97	58	32	21	0.0079897	0.08	7.24159	0 1 2 3 4 5 8 9 10 12 13 15
5	6.56	58	31	22	0.00858	0.08	7.58641	0 2 3 4 5 6 8 9 11 12 13 15

TABLE 4.10: *Results of a maximization scan of 5000 subspaces of the HEP-MASS feature space, with a D1 dataset composed of 9500 background events and 500 injected signal events, and a D2 dataset containing 10,000 signal events. N_{obs} is the number of D1 events in the signal box, and $N_{s,in}$ is the number of signal events in dataset D1 captured in the signal box. The best identified signal boxes are ordered by decreasing value of the Z_{UL} test statistic, whose value is inversely proportional to the estimated 95% C.L. upper limit on signal cross section achievable by a counting experiment.*

The results of a `RanBox_SS` scan can also be visualized graphically, as shown in Figs. 4.16 and 4.17.

4.6.3 Bias studies

As we discussed above, the estimate of events from dataset D1 in the signal box with events in the second sideband is expected to be negatively biased, due to the intrinsic correlations between kinematic variables defining the feature space. These

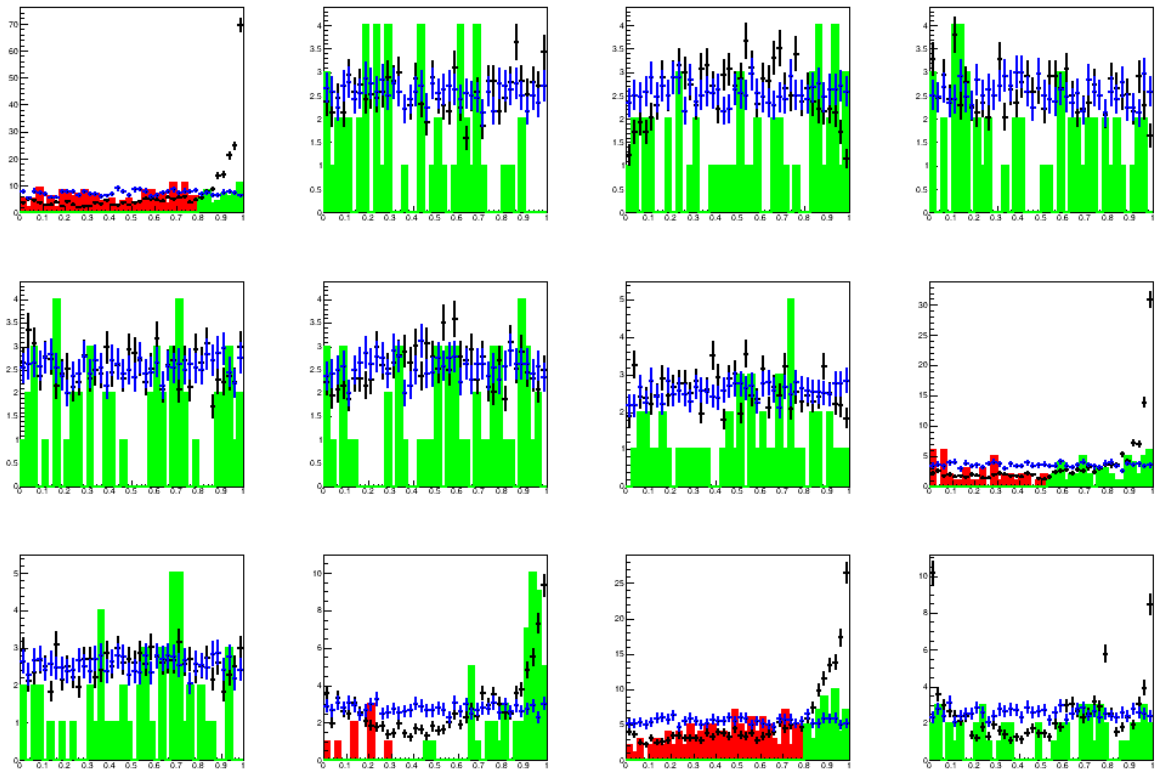


FIGURE 4.16: *Marginal distributions of the 12 features in the subspace where* `RanBox_SS` *finds the best box, for a large-statistics run. The green distributions show the selected data, the red distributions show data that are rejected only because of their value of the shown feature (they would otherwise be included in the signal box); the blue distributions show the original unselected dataset D1 (rescaled by an arbitrary factor to fit in the graph), and the black distributions show the unselected dataset D2 (also arbitrarily rescaled). See the text for more detail.*

correlations in many cases affect the signal, which drives the maximization of the numerator of Z_{UL}, and the background in a similar manner – both have to withstand to physical constraints between their kinematical features. We may define the bias as follows:

$$b = 2(N_{obs} - N_{exp})/(N_{obs} + N_{exp}) \tag{4.4}$$

A principled way to estimate the above bias in N_{exp} is to define a set of alternative signals, simulate their characteristics, obtain a set of alternative datasets $D2_1$, $D2_2$, $D2_3$,... and run `RanBox_SS` on each separately, maximizing Z_{UL} against the same background D1. The resulting mean and variance of the distribution of ratios between observed events in the signal box and expected events in the second sideband

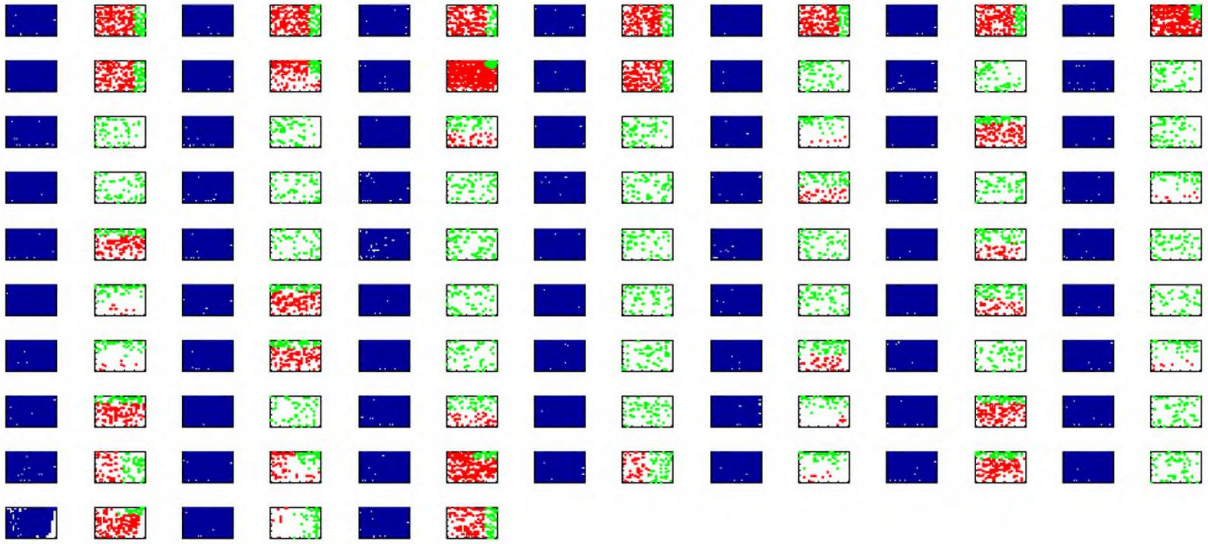


FIGURE 4.17: *Scatterplots of the 12 features in the copula space for a high-statistics run. Each pair of graphs shows a two-dimensional subspace of the 12-dimensional space where* `RanBox_SS` *finds the best box. The blue distribution shows data in the D1 dataset before any selection; the corresponding distribution on its right shows the selected data in the best box (in green) and the data that would have been included in the best box if it did not fail the selection on the two displayed features (in red). The 66 pairs of distributions show variables 1 vs 2, 1 vs 3, 1 vs 4, ... 11 vs 12.*

are then sound estimators of the bias and its variability, and they can be used to correct the prediction N_{exp} in the case of the original datasets D1 and D2. These alternative datasets might, *e.g.*, be constructed by artificially changing the mass of the tau leptons in the simulation, or the mass of the B_s meson, or even the mass of the ρ meson, as the latter is an almost certain intermediate state in the $\tau \to \rho\pi^- \to \pi^+\pi^-\pi^-$ decay.

Since we have been testing `RanBox_SS` on a different sample of data from the one which is our target, we study here a more general technique which is less dependent on the specific kinematic properties of the datasets. The technique consists in searching, for each signal box identified by `RanBox_SS` after gradient descent maximization of Z_{UL}, a corresponding alternative region of the considered subspace of the feature space, and correspondingly a sideband and a second sideband to it. The requirements of such an alternative region are the following:

- It must have the same extension in each subspace dimension as the original signal box;

- It must have no overlap with the original signal box;
- It must contain a number of D1 events in the signal box similar to the number of the original signal box.

The alternative box is sought for by random trials, by changing the location of the multi-dimensional interval while keeping its shape unaltered. This is a time-consuming procedure, as it may prove very difficult to fulfil the above criteria, especially if the definition of "similarity" in the observed event counts from dataset D1 in the signal boxes is too strict. We have observed that, with typical number of events and dimensionality of the subspaces in runs on the HEPMASS dataset, the identification of a box with the above characteristics typically requires less than 2000 trials; in few cases, when `RanBox_SS` has identified by gradient descent a unusually dense and small region of phase space, which cannot easily be replicated by random sampling, the required iterations diverge. A workable criterion of "similarity" is to impose that the difference between the observed event counts in the two signal boxes be smaller than 10% of their average value. We have observed that the bias estimates depend very little on the precise value of this criterion.

A run on 10,000 $\mathcal{D}'$-dimensional subspaces of the HEPMASS feature space, using 5000 events in dataset D1 (only composed of background events) and 5000 events in dataset D2, allows to verify the soundness of the above bias estimation procedure. The results are shown in Fig. 4.18 and reported in Table 4.11 below.

	Bias	SQM
Original box	0.2685 ± 0.0039	0.392
Alternative box	0.3452 ± 0.0038	0.378

TABLE 4.11: *Extrapolation bias and its estimate with random boxes. See the text for details.*

The estimated and real bias are different, but the difference in their means is not very large. A larger systematic effect on the extrapolation than the one due to the difference in mean biases above is potentially due to the variability of the bias, which is only partly explained by the statistical fluctuation of the observed and expected

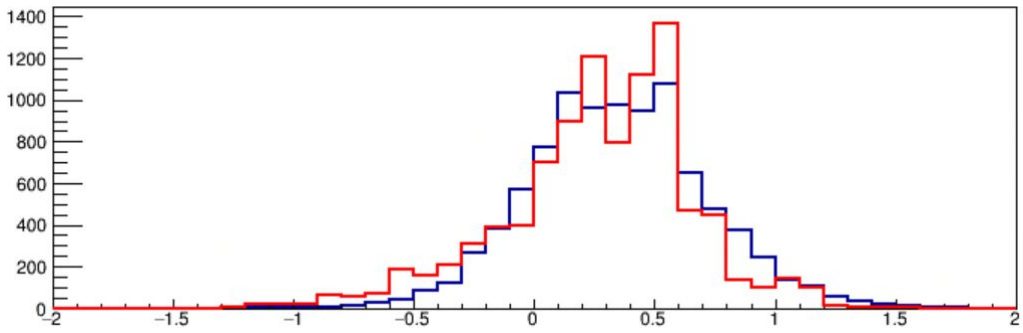


FIGURE 4.18: *Comparison of the distribution of bias (defined as in Eq. 4.4) in 10,000 signal boxes returned by the gradient descent search in a HEPMASS data sample of 5000 signal and 5000 background events, with* $N_{var} = 12$ *(red), with the bias estimated in the alternative boxes by the procedure described in the text (blue).*

event counts in the boxes [5]; however, by re-sampling multiple times random boxes of characteristics similar to the one obtained by gradient descent, it may be possible to reduce this effect.

The above study indicates that the procedure may in fact provide a viable correction to the background estimate provided by the second sideband. Of course, these results are strongly dependent on the characteristics of the studied problem (the inner correlations in the feature space), therefore a separate assessment needs to be carried out in every case. The procedure to handle large biases, which are however not expected in the case of the B_s search due to the less striking characteristics of the signal in that situation, is to run a Z_{UL} maximization to derive a bias estimate, and then to correct the calculation of Z_{UL} including it as a factor in the denominator of that test statistic, so that a second run may converge more precisely to the most advantageous signal region. More studies are needed to finalize this procedure.

[5] The typical number of events in the 10,000 signal boxes studied in this run is 37.3 (with a RMS of 36.5); in such conditions, and with the average bias of 0.27 mentioned above, this translates into a typical variability of bias estimates of about 0.21 due to Poisson statistics.

Chapter 5

Conclusions and Outlook

In the first chapter of this thesis, the elementary particles and the characteristics of their interactions with matter as well as their detection principles were discussed. Furthermore the theoretical models for the fundamental interactions between particles were briefly introduced. In the second chapter, components of the Compact Muon Solenoid (CMS) detector at the Large Hadron Collider (LHC) were described and the operation principles were reviewed. The inference of the parameters of the theoretical models occur by the aid of accurate simulations of the physics processes taking place in the collisions as well as the simulation of the response of the detector components. Simulated events can be used for training supervised learning models which are then employed on the successfully reconstructed events. The principle behind the supervised learning methods that were employed within the scope of thesis were briefly described in Chapter 3.

The search for anomalous regions of a complex feature space can be performed proficuously if the space is transformed into a standardized copula, where the marginal density of every feature is uniform. This allows to identify a multi-dimensional interval which captures unusual overdensities, possibly due to anomalous contaminations of the data sample. In chapter 3 and 4 we describe an algorithm that performs this search, `RanBox`, and demonstrated that it has considerable power in locating anomalous signals. We have shown how to customize `RanBox` to search for a specific, well-defined signal in data that are otherwise hard to model. In this semi-supervised version the algorithm, `RanBox_SS`, is designed to minimize the upper limit on the signal cross section extractable from the identified multi-dimensional interval by a counting experiment that uses as a background prediction the number of data events

captured in a suitable sideband in the multi-dimensional space.

The $B_s \to \tau\tau$ signal is a very difficult one to extract in LHC collisions data and has never been attempted by a CMS analysis, due to its extreme rarity and to the incompleteness of available information (missing neutrinos). Despite this, machine learning tools may partially recover the missing information and provide a means of improving the signal discrimination over the large backgrounds. Furthermore, the sheer size of the available B-Parked data, comprised of many triggers, complicates the data reduction and management. CMS does not have a simulation to model these data, which makes the search even more difficult. The absence of a simulation of real data and the complexity of the overlap of several prescaled triggers demands that the background be studied with a data-driven technique.We produced an initial separation of signal and background by exploiting a fraction of the data for the classification task.

Within the scope of this thesis, we have described the selection of a dataset where the signal is enhanced, the construction of an estimate of the B_s kinematics that allow to obtain a peak in the reconstructed mass distribution, and the definition of the feature space where the final search will be carried out, along with a mention of the semi-supervised algorithm we have developed for the signal extraction task. Final considerations have been discussed at the end of the chapter 4. The future steps of this analysis will involve the finalization of the search algorithm and its application to the selected data. We expect to set a competitive upper limit on the searched for process, and we believe this work will constitute a solid basis for future searches of this very rare but important Standard Model process. Tests of the `RanBox` algorithm show that it is a viable procedure for the search of the B_s meson in LHC collisions data. Future work will allow us to define in an optimal way a feature space where to run `RanBox_SS` and obtain a stringent upper limit on the cross section of that process, which is currently still beyond the observability with available LHC data. We believe that the characterization of rare decays of the B_s meson will have a chance to evidence deviations from the SM and pave the way to targeted searches for new physics.

Bibliography

[1] CDF Collaboration. "Observation of Top Quark Production in $Pbar - P$ Collisions". arXiv:hep-ex/9503002. doi:10.1103/PhysRevLett.74.2626

[2] DONUT Collaboration. "Observation of tau neutrino interactions". doi:10.1016/S0370-2693(01)00307-0.

[3] CMS Collaboration. "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC". In: Phys. Lett. B **716** (2012), 30-61. arXiv:1207.7235 [hep-ex]. doi:10.1016/j.physletb.2012.08.021

[4] ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: Phys. Lett. B **716** (2012), 1-29. arXiv:1207.7214 [hep-ex]. doi:10.1016/j.physletb.2012.08.020

[5] F. M. Gonzalez, et al. "Improved neutron lifetime measurement with UCN". arXiv:2106.10375v2 [nucl-ex]. doi.org/10.1103/PhysRevLett.127.162501

[6] CMS Collaboration. "Determination of the strong coupling constant S(mZ) from measurements of inclusive W and Z boson production cross sections in proton-proton collisions at $\sqrt{s}$ = 7 and 8 TeV". In: JHEP 06 (2020) 018. arXiv:1912.04387 [hep-ex]. doi:doi.org/10.48550/arXiv.1912.04387

[7] M, Cush. "Standard Model of Elementary Particles". PBS NOVA, Fermilab, Office of Science, United States Department of Energy, Particle Data Group. Wikipedia.

[8] M. Thomson. "Modern Particle Physics" (2013). ISBN:978-1-107-03426-6 pp. 7.

[9] M. Thomson. "Modern Particle Physics" (2013). ISBN:978-1-107-03426-6 pp. 19.

[10] L. Sehwook. "On the limits of the hadronic energy resolution of calorimeters". In: J. Phys.: Conf. Ser. 1162 012043 (2019). doi:10.1088/1742-6596/1162/1/012043

[11] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes, and R. P. Hudson. "Experimental Test of Parity Conservation in Beta Decay". In:Phys. Rev. 105, 1413 (1957). doi:10.1103/PhysRev.105.1413.

[12] Sheldon L Glashow. "Partial-symmetries of weak interactions". In: Nuclear Physics 22.4 (1961), pp. 579–588.

[13] A. Salam and J.C. Ward. "Partial-symmetries of weak interactions". In: Physics Letters 13.2 (1964), pp. 168–171. ISSN: 0031-9163. doi: org/10.1016/0031-9163(64)90711-5

[14] Edvige Corbelli and Paolo Salucci. 'The extended rotation curve and the dark matter halo of M33'. In: Monthly Notices of the Royal Astronomical Society 311.2 (2000), pp. 441–447.

[15] Planck Collaboration. 'Planck 2015 results. XIII. Cosmological parameters'. In: Astron. Astrophys. 594 (2016), A13. arXiv: 1502.01589 [astro-ph.CO]. doi:10.1051/0004-6361/201525830.

[16] Virginia Trimble. 'Existence and nature of dark matter in the universe'. In: Annual review of astronomy and astrophysics 25.1 (1987), pp. 425–472.

[17] Michael S. Turner. "ΛCDM: Much more than we expected, but now less than what we want". arXiv:2109.01760 [astro-ph.CO]. doi:10.48550/arXiv.2109.01760

[18] L.Roszkowski, E.M. Sessolo, S. Trojanowski. "WIMP dark matter candidates and searches - current status and future prospects". arXiv:1707.06277v2 [hep-ph]

[19] M. Carena, M. Quirós, Y. Zhang. "Electroweak Baryogenesis From Dark CP Violation". In: Phys. Rev. Lett. 122, 201802 (2019). arXiv:1811.09719 [hep-ph]

[20] J. Formaggio, A. Gouvêa, R. Robertson "Direct Measurements of Neutrino Mass" arXiv:2102.00594v2 [nucl-ex] . doi:10.1016/j.physrep.2021.02.002.

[21] Y. Fukuda et al. 'Evidence for oscillation of atmospheric neutrinos'. In: Physical Review Letters 81.8 (1998), p. 1562.

[22] H. Arodz. "Relativistic Quantum Mechanics of the Majorana Particle". arXiv:2002.07482 [quant-ph]

[23] Joseph Polchinski. String Theory. Cambridge monographs on mathematical physics. Cambridge: Cambridge Univ. Press, 1998.

[24] A. Ashtekar, E. Bianchi. "A Short Review of Loop Quantum Gravity". arXiv:2104.04394v1 [gr-qc]

[25] CERN Service graphique. 'Overall view of the LHC. Vue d'ensemble du LHC' (2014). General Photo. Url: https://cds.cern.ch/record/1708847

[26] CMS Public Luminosity Results

[27] CERN Service graphique.'The CERN accelerator complex, layout in 2022' (Feb, 2022). General Photo. Url: https://cds.cern.ch/record/2800984

[28] CERN Service graphique. 'Collisions recorded by the CMS detector on 14 Oct 2016 during the high pile-up fill' (Nov, 2016). Url: https://cds.cern.ch/record/2231915

[29] ATLAS Collaboration. 'The ATLAS Experiment at the CERN Large Hadron Collider'. In: JINST 3 (2008), S08003. doi: 10.1088/1748-0221/3/08/S08003 .

[30] CMS Collaboration. 'The CMS Experiment at the CERN LHC'. In: JINST 3 (2008), S08004. doi: 10.1088/1748-0221/3/08/S08005 .

[31] LHCb Collaboration. 'The LHCb Detector at the LHC'. In: JINST 3 (2008), S08005. doi: 10.1088/1748-0221/3/08/S08005 .

[32] ALICE Collaboration. 'The ALICE experiment at the CERN LHC'. In: JINST 3 (2008), S08002. doi: 10.1088/1748-0221/3/08/S08002 .

[33] TOTEM Collaboration. 'The TOTEM experiment at the CERN Large Hadron Collider'. In: JINST 3 (2008), S08007. doi: 10.1088/1748-0221/3/08/S08007 .

[34] MoEDAL Collaboration. 'The Physics Programme Of The MoEDAL Experiment At The LHC'. In: Int. J. Mod. Phys. A29 (2014), p. 1430050. doi: 10.1142/S0217751X14300506 .

[35] LHCf Collaboration. 'The LHCf detector at the CERN Large Hadron Collider'. In: JINST 3 (2008), S08006. doi: 10.1088/1748-0221/3/08/S08006.

[36] FASER Collaboration.'Technical Proposal for FASER: ForwArd Search ExpeRiment at the LHC'. (2018) arXiv:1812.09139 [physics.ins-det]

[37] CERN Service graphique. 'Cutaway diagram of CMS detector' (May, 2019). Url: https://cds.cern.ch/record/2665537

[38] CMS Collaboration. Technical Design Report Volume 1: Detector Performance and Software. Technical Design Report CMS (2006). https://cds.cern.ch/record/922757

[39] CMS Collaboration. "CMS Tracking Performance Results from Early LHC Operation". In: Eur.Phys.J.C70:1165-1192 (2010). arXiv: 1007.1988.

[40] CMS Collaboration. "The CMS Phase-1 Pixel Detector Upgrade". In: JINST 16, P02027. doi:10.1088/1748-0221/16/02/P02027.

[41] CMS Collaboration.'Description and performance of track and primary-vertex reconstruction with the CMS tracker'. In: 2014 JINST 9 (2018), P10009. doi:10.1088/1748-0221/9/10/P10009.

[42] The CMS electromagnetic calorimeter project: Technical Design Report (1997). https://cds.cern.ch/record/349375.

[43] P. Siddireddy for CMS Collaboration. 'The CMS ECAL Trigger and DAQ system: electronics auto-recovery and monitoring'. (2018) arXiv:1806.09136 [physics.ins-det].

[44] CMS Collaboration. 'Calibration of the CMS hadron calorimeters using proton-proton collision data at $\sqrt{s} = 13$ TeV'. In: JINST 15 (2020) P05002 arXiv:1910.00079 [physics.ins-det]. doi:10.1088/1748-0221/15/05/P05002.

[45] K. Minsuk for CMS Collaboration. 'CMS reconstruction improvement for the muon tracking by the RPC chambers'. In: JINST 8 (2013) T03001 arXiv:1209.2646 [physics.ins-det]. doi:10.1088/1748-0221/8/03/T03001.

[46] CERN Service graphique. 'CMS Detector Slice' (2016). CMS Collection. Url: https://cds.cern.ch/record/2120661

[47] CMS Collaboration. 'Particle-flow reconstruction and global event description with the CMS detector'. In: JINST 12.10 (2017), P10003. arXiv:1706.04965 [physics.ins-det]. doi: 10.1088/1748-0221/12/10/P10003.

[48] CMS Collaboration. CMS Tracking POG Performance Plots For 2017 with PhaseI pixel detector.

[49] M. Cacciari, G. P. Salam and G. Soyez. 'The anti-kt jet clus- tering algorithm'. In: JHEP 04 (2008), p. 063. arXiv: 0802.1189 [hep-ph]. doi: 10.1088/1126-6708/2008/04/063.

[50] Particle Data Group. "Review of Particle Physics" , In: Prog. Theor. Exp. Phys. 2020, 083C01 (2020). doi/10.1103/PhysRevD.98.030001

[51] S. Agostinelli et al. "Geant4 a simulation toolkit". In: Nucl. Instr. Meth. A 506 (2003) 250. doi:10.1016/S0168-9002(03)01368-8.

[52] Arthur L. Samuel. "Artificial Intelligence: A Frontier of Automation" (1962). doi:10.1177/000271626234000103

[53] T. Hastie, R. Tibshirani, J. Friedman "The elements of Statistical Learning". (2017) ISBN:978-0-387-84857-0.

[54] Leo Breiman. "Classification and regression trees". Routledge, 2017.

[55] T. Dorigo, M. Fumanelli, C. Maccani, M. Mojsovska, G.C. Strong, and B. Scarpa, "RanBox: Anomaly Detection in the Copula Space". (2021) arXiv:2106.05747 [physics.data-an].

[56] CMS Collaboration. " Search for contact interactions and large extra dimensions in the dilepton mass spectra from proton-proton collisions at $\sqrt{s} = 13$ TeV". In: JHEP 04 (2019) 114, arXiv:1812.10443 [hep-ex] doi:10.1007/JHEP04(2019)114.

[57] A. Sklar. "Fonctions de répartition à n dimensions et leurs marges", Publ. Inst. Statist. Univ. Paris, 8 (1959) 229.

[58] R.E. Bellman, Rand Corporation. "Dynamic programming". Princeton University Press (1957), p. ix. ISBN 978-0-691-07951-6.

[59] F. L. Gewers, et al. "Principal Component Analysis: A Natural Approach to Data Exploration". (2018) arXiv:1804.02502v2 doi:doi.org/10.1145/3447755.

[60] T.P. Li and Y.Q. Ma. "Analysis methods for results in gamma-ray astronomy". Astroph. Journ. 272 (1983) 317. doi:10.1086/161295.

[61] C.E. Bonferroni. "Teoria statistica delle classi e calcolo delle probabilità", Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 1936.

[62] ROOT

[63] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing* (second ed.) (1992). Cambridge University Press, ISBN 0-521-43108-5.

[64] UCI ML Repository

[65] P. Baldi, P. Sadowski, and D. Whiteson. Searching for Exotic Particles in High-Energy Physics with Deep Learning. Nature Comm. 5 (2014) 4308, arXiv:1402.4735v2 [hep-ph]. doi:10.1038/ncomms5308.

[66] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson, " Parameterized Machine Learning for High-Energy Physics". Eur. Phys. Journ. C76,5 (2016) 7. arXiv:1601.07913v1 [hep-ex]. doi:10.1140/epjc/s10052-016-4099-4.

[67] S. Weinberg, "A model of leptons", Phys. Rev. Lett. 19, 1264 (1967).

[68] L. Evans and P. Bryant, "LHC Machine", J. Instrum. 3 (2008) S08001, doi:10.1088/1748-0221/3/08/S08001.

[69] CMS and LHCb Collaborations, "Observation of the rare $B_s^0 \to \mu\mu$ decay from the combined analysis of CMS and LHCb data", Nature 522 (2015) 68-72, arXiv: 1411.4413 (2014), doi:10.1038/nature14474 .

[70] R. Bainbridge. "Recording and reconstructing 10 billion unbiased b hadron decays in CMS". In: EPJ Web of Conferences 245 (2020) 01025 . doi:10.1051/epjconf/202024501025.

[71] Torbjörn Sjöstrand, Stephen Mrenna, Peter Skands. "A Brief Introduction to PYTHIA 8.1". arXiv:0710.3820 [hep-ph] doi:10.1016/j.cpc.2008.01.036

[72] D. J. Lange. "The EvtGen particle decay simulation package". In: Nucl. Instrum. Meth. A 462 (2001) 152. doi:10.1016/S0168-9002(01)00089-4 .

[73] Y. Takahashi, V. M. Mikuni, B. Kilminster, and C. Galloni. "Low p_T τ_h reconstruction in 3 prong decay" [CMS AN-19-100]

[74] V. Mikuni, F. Canelli, "ABCNet: An attention-based method for particle tagging". arXiv:2001.05311v2 [physics.data-an]. doi:10.1140/epjp/s13360-020-00497-3

[75] C. Chen, L. Z. Fragonara, and A. Tsourdos. "GAPNet: Graph Attention based Point Neural Network for Exploiting Local Feature of Point Cloud". arXiv:1905.08705

[76] I. Goodfellow, Y. Bengio; A. Courville "6.2.2.3 Softmax Units for Multinoulli Output Distributions". Deep Learning, MIT Press (2016), pp. 180–184. ISBN 978-0-26203561-3

[77] Jerome H. Friedman, "Stochastic gradient boosting". Computational Statistics & Data Analysis **38** (2002), 367 - 378. doi:10.1016/S0167-9473(01)00065-2

[78] T. Dorigo and P. de Castro Manzano. " Dealing with Nuisance Parameters using Machine Learning in High Energy Physics: a Review". 2020. arXiv:2007.09121 [stat.ML]