DIPLOMARBEIT

Background Studies and Machine Learning Methods Applied to the Analysis of Central Exclusive Production Events in ALICE

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Technische Physik

eingereicht von

Sebastian Ratzenböck, BSc
Matrikelnummer 01027270

ausgeführt am Atominstitut
der Technischen Universität Wien
(in Zusammenarbeit mit dem Stefan-Meyer-Institut für subatomare Physik)

Betreuung
Betreuer: Prof. Dr. Eberhard Widmann
Mitwirkung: Dr. Paul Bühler

Wien, 20.08.2018

(Unterschrift Verfasser) (Unterschrift Betreuer)

Abstract

This master thesis is a report of a survey aimed at understanding and reducing background sources in central exclusive production events measured at the ALICE experiment, located at CERN–LHC. The ALICE experiment consists of a central barrel and a forward muon spectrometer. Additional smaller detectors for global event characterization and triggering are located at small angles outside of the central barrel. Such a geometry allows the investigation of many properties of diffractive reactions at hadron colliders, for example the measurement of single and double diffractive dissociation cross-sections and the study of central exclusive production (CEP). Central diffractive events are defined experimentally by hits in the central barrel and no activity outside of it, creating an activity gap in the observed rapidity of measured particles. The study of *Pythia-8* simulations of these processes show a drastic reduction of non-diffractive events (background) by enforcing the rapidity gap condition. The remaining background is largely composed of partially reconstructed CEP events, so called feed-down events. Often feed-down events are accompanied by neutral particles, which are not detected. This missing mass and momentum leads to a shift of the invariant mass spectrum to lower masses. This thesis aims at understanding and suppressing background sources in the two pion invariant mass spectrum in $X \to \pi^+\pi^-$ decays of the centrally produced system X. This is done in two ways: First, a feed-down template is constructed by using background events marked by a detected gamma in the main calorimeter of ALICE, and by using events with more than two detected charged tracks. Despite facing possibly tedious efficiency corrections for the sake of complete feed-down descriptions, this method yields promising results. Second, machine learning methods for background suppression of CEP events are employed. The measured variables *e.g.* the four-momentum of particles, energy loss in the detectors, deduced kinematic quantities, and global event characteristics are generally correlated. To obtain a maximal separation of signal and background it is necessary to treat these observables in a fully multivariate way. Although achieving good results, *i. e.* the signal purity can be increased by 30% while maintaining a nearly constant signal efficiency, the trained classifiers tend to obtain a strong mass bias which results in a cut-like behavior of the trained model. It can be concluded that multivariate techniques trained on Pythia-8 generated CEP simulations generally suffer from incomplete Monte Carlo descriptions, including only high mass continuum production. However, promising new packages are currently being developed which provide interesting prospects for further studies.

Zusammenfassung

Die vorliegende Arbeit versucht das Verständnis von Untergrundereignissen in speziellen inelastischen Streuvorgängen zu fördern. Das Charakteristikum dieser Streuvorgänge ist, dass beide Streupartner erhalten bleiben und ein Teilchen X bei einer zentralen Rapidität (um null) erzeugt wird, während die ausgehenden Streupartner mit hohen Rapiditäten den Streuvorgang verlassen. Diese Prozesse werden in der Literatur als "central exclusive production events" (CEP Events) bezeichnet. Diese Studie bezieht sich auf die Analyse von CEP Events, die im Zuge des ALICE Experiments am CERN–LHC gemessen wurden. Das ALICE Experiment besteht aus einem zentralen Detektor-Bereich und einem Vorwärts-Detektorsystem zum Nachweis von Myonen. Außerhalb des zentralen Detektorsystems befinden sich verschiedene kleinere Detektorsysteme, welche für die Bestimmung von globalen Event-Eigenschaften unter kleinen Winkeln zur Strahlachse platziert sind. Um CEP Events in ALICE zu messen, wird verlangt, dass Teilchenspuren im zentralen Bereich von ALICE detektiert werden, während die Vorwärts-Detektorsysteme frei von jeglicher Signalaktivität bleiben: *d.h.* ein Rapiditäten-Doppelspalt-Filter wird implementiert. Die Analyse von *Pythia-8* Simulationen zeigt, dass dieser Rapiditäten-Doppelspalt-Filters eine drastische Reduktion von nicht-CEP Events hervorruft. Der bestehende Untergrund besteht aus CEP Events selbst, in denen nur ein Teil des vollständigen Zerfallskanals von X gemessen wird. Dadurch geht ein Teil der gesamten Energie und Masse des Ursprungssystems verloren, was zu einer Verschiebung des invarianten Massenspektrums zu kleineren Massen führt. Dieser Untergrund wird als "feed-down" bezeichnet. Die vorliegende Arbeit beschäftigt sich mit Methoden, deren Ziel es ist feed-down Untergrund im folgenden beobachteten Zerfallskanal: $X \to \pi^+\pi^-$, (1) zu *beschreiben* und (2) zu reduzieren. Um den feed-down Untergrund zu *beschreiben* wird eine Schablone des Untergrunds erzeugt. Dafür wird die Massenverteilung von Events mit einem detektierten Photon im Kalorimeter (EMCal) und von Events mit mehr als drei detektierten geladenen Teilchen gemessen. Obwohl diese Methode vor möglicherweise schwierigen Effizienz-Korrekturen steht, ähneln die Ergebnisse stark der originalen feed-down Gestalt. Des Weiteren werden multivariate Analysetechniken verwendet, um den Untergrund in CEP Events zu *reduzieren*. Da im Allgemeinen die gemessenen Variablen, wie *e. g.* der Viererimpuls eines Teilchens oder der Energieverlust in den Detektoren, korreliert sind, bietet eine multivariate Analyse einen vielversprechenden Ansatz. Trotz einer Verbesserung des Signal-Untergrund-Verhältnisses von bis zu 30% während die Signaleffizienz beinahe gleich bleibt, erhalten die trainierten Modelle eine starke Massen-Verzerrung. *D.h.* die multivariate Methode verhält sich sehr ähnlich zu einer eindimensionalen Entscheidung, welche die Massen-Variable betrifft. Es kann die Schlussfolgerung gezogen werden, dass es multivariaten Analysemethoden, welche auf herkömmlichen Trainingsmethoden basieren, *d.h.* "fully supervised learning" auf Pythia-8 simulierten Daten, an vollständigen Beschreibungen von CEP Prozessen mangelt. Allerdings werden zurzeit neue, vielversprechende Simulationspakete entwickelt, welche eine große Relevanz für zukünftige Studien aufweisen könnten.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to use this opportunity to express my gratitude to everyone who supported me throughout the course of this project.

Foremost, my sincere thanks goes to my supervisor Eberhard Widmann for his support during my master thesis. I am deeply indebted for his continuous presence, especially in the last few weeks of the writing and correction process. His calm and immediate replies to my countless e-mails enforced me in my scientific endeavors and lifted me up, when I was seemingly facing a dead end. Not least his role as director of the Stefan Meyer Institute is one of the main factors contributing to the extremely familiar atmosphere at the institute, where I felt welcome from the very first day.

Most importantly, I would like to express my gratitude to my co-supervisor Paul Bühler. His continuous support, and insightful comments are the cornerstones of this thesis. Besides teaching me a great deal about scientific research, Paul supported me in every way possible and made working for my master project a real pleasure.

Additionally, I have to thank Michael Weber, Sebastian Lehner, and Aaron Capon, who have with their expertise in data analysis at ALICE and machine learning always provided me with excellent answers. Moreover, they, and all the other office colleagues were extremely fun and pleasant to be around, which made this project all the more enjoyable.

I would like to express my warmest thanks to my family, especially my parents for their love and continuous support in all of my choices. A special thanks goes to my sister Johanna and my friend Theresa who spent countless hours proof reading this thesis. Finally, my wonderful friends should not go unmentioned - thank you.

This work was supported by the Austrian Academy of Sciences.

Chapter 1

Motivation

Particle physics is the study of the smallest, irreducible building blocks of nature and the interaction between them. The governing theory - the Standard Model - describes the known elementary particles and the dynamics via the four fundamental forces in astonishing detail. A central idea is the concept of reductionism, which states that physical phenomena can be described by breaking the problem down into smaller, fundamental constituents. Despite hints for physics beyond the Standard Model (*e. g.* dark matter and quantum gravity), phenomena within the theory, like the strong force, can become equally challenging due to completely distinct behavior at different energy and momentum scales. The framework of the strong force is quantum chromodynamics (QCD), which was developed according to the principles of quantum electrodynamics, following its vast successes in describing fundamental interactions between electrically charged particles. At high energy levels QCD is experimentally well established where it can be described with perturbative methods achieving great precision. In this energy regime the interaction is characterized in terms of basic quark and gluon exchanges. However, at lower energy levels accurate descriptions become increasingly difficult to outright impossible - even if knowledge of the higher energy dynamics is considered - as complex (high-order) interactions become the dominating processes. *Diffraction physics* at LHC energies lies in-between these two energy scales describing strong interactions outside the QCD framework via *Regge theory*. To resolve the issue of the rising total cross section at high energy levels a *Pomeron* ansatz is used to describe the mediation of the strong force. This Pomeron-state carries vacuum quantum numbers. However, the Pomeron is not represented by a single physical particle but instead it is associated with a superposition of multiple particles with constitute the so-called Pomeron trajectory. In the current set of known compound particles none can be attributed to this Pomeron trajectory. The glueball, however, a hypothetical particle consisting only of gluons would classify as a candidate since the simplest exchange of vacuum quantum number is via a pair of gluons in a color singlet state [1].

This study describes a special diffractive process called *central exclusive production* (CEP) in proton-proton collisions with a center of mass energy $\sqrt{s}$ of 13 TeV. measured in ALICE at LHC–CERN. CEP events are defined as processes in which the two interacting protons stay intact but exchange sufficient energy and momentum to create a particle X. According to Regge theory theses states (at LHC energies) are produced by a fusion of two Pomerons, which are emitted by the interacting protons. The production of X via double Pomeron fusion is a col-

orless mechanism which results in a clear experimental signature with large voids of particles between the outgoing protons and the centrally produced system in the pseudorapidity variable η. This is referred to as a rapidity gap. Measuring the decay products of X (*i. e.* $\pi^+\pi^-$ in this thesis) allows for a detailed study of the Pomeron. However, the analysis of the $X \to \pi^+\pi^-$ invariant mass spectrum is prone to background sources from high mass states that decay into $X \to \pi^+\pi^- + N$, two charged pions and N additional unobserved particles. In order to reduce this background component a general background study is carried out followed by a multivariate approach to reduce its contribution to the total mass spectrum.

This thesis is structured as follows: First, Chap. 2 features an introduction to the Standard Model (Sec. 2.1) and diffractive physics (Sec. 2.2) and subsequently its role in the ALICE experiment (Sec.2.3). Second, Chap. 3 aims at understanding the background components by studying simulated data as well as creating a background template (Sec. 3.3) which can be subtracted from the data yielding signal events. Third, Chap. 4 describes a multivariate approach using neural networks to reduce background components. Finally, Chap. 5 summarizes results obtained in the two previous chapters and gives an outlook for future CEP studies.

Chapter 2

Theoretical background

2.1 The standard model

In this section a short introduction to the standard model and QCD is presented. It summarizes the ideas found in [2–5] where also further information is available.

The Standard Model of particle physics describes the known *fundamental* particles and the interactions between them. It is based on the principle of symmetry under space-time[1] and gauge transformations. Since space-time transformations represent a change in coordinate system, it follows that the laws of physics must be invariant under space-time transformations. To kinematically describe a relativistic theory which is required to be invariant under local gauge transformations one uses a Lagrangian density. The set of gauge transformations under which the Standard Model Lagrangian is invariant is given by the following symmetry group

$$SU(3)_C \times SU(2)_L \times U(1) \tag{2.1.1}$$

$U(n)$ is the group of all $n \times n$ unitary matrices and $SU(n)$ are a subgroup of the $U(n)$ with unit determinant, called the *special* unitary matrices. The symmetry group $SU(2)_L \times U(1)$ is associated with the electro-weak interaction. It represents a unification of electromagnetic and weak interactions. The subscript L indicates coupling only to left handed particles. The $SU(3)_C$ part gives rise to the strong interactions where the subscript C denotes that particles with color charge transform under the $SU(3)$ group. The electromagnetic force is described by a subgroup of the $U(1)$ part of the electro-weak symmetry. The weak interactions are described by the rest of the $SU(2)_L \times U(1)$ part. The Higgs mechanism spontaneously breaks the electro-weak symmetry $SU(2)_L \times U(1) \rightarrow U(1)_{EM}$. The group $U(1)_{EM}$ is associated with the electromagnetic interaction and acts only on electrically charged particles. The fact that the weak and strong force are described by a higher symmetry as the electromagnetic force gives rise to more than one boson mediating the weak and strong forces.

Additionally, there exist multiplicative discrete symmetries that are conserved in some particle interactions called the charge (C), parity (P) and time (T) reversal symmetry. When these symmetry operations act on particles a transformation of (affected) particles takes place. The charge symmetry operation flips all charge quantum numbers which include electric charge, baryon and lepton numbers as

[1] Space-time transformations are translations, rotations and Lorentz boosts.

well as the quark numbers, isospin, strangeness, charm, bottomness and topness. As a consequence, the particle is transformed into its anti-particle. The parity operation affects only the spatial coordinates flipping their sign while leaving the sign of the time component unaffected. The T-symmetry reverses the sign of a particle's time component. A time symmetric process requires an interaction to be also possible in the other direction[2]. Although the individual symmetries can be violated by *e. g.* the weak interaction, the standard model requires that the combined CPT symmetry is conserved. *I. e.* the Lagrangian has to be invariant under the simultaneous application of all three operations. The pure electromagnetic and strong interaction conserve also C, P and T operations separately.

The particles currently considered elementary are fermionic particles which make up all the standard matter and the bosons that mediate the electromagnetic, weak and strong forces. The fermionic matter consists of six quarks and six leptons. The quark sector is made up of the *up*, *down*, *charm*, *strange*, *top* and *bottom* quarks. The leptons are split into the charged leptons, the electron, the muon, and tau as well as the three neutrinos.

2.1.1 QCD

To describe the processes that are relevant to this analysis we have to focus on the strong interaction which is described by Quantum chromodynamics (QCD). The Lagrangian density $\mathcal{L}$ of QCD yields the kinematics of quarks and gluons. It is defined by the following formula[3] [5]

$$\mathcal{L}_{QCD} = -\frac{1}{4} F^a_{\mu\nu} F^{\mu\nu\ a} + \sum_q \overline{q}(i\gamma^\mu D_\mu - m_q)q \tag{2.1.2}$$

Here q describes the quark fields for $q = \{u, d, s, c, b, t\}$ with their associated mass m_q and γ^μ denotes the gamma matrices. μ represents the four-dimensional space-time indices. The spinor indices of γ_μ and q a have been suppressed in the interest of readability. The gluon field strength tensor $F^a_{\mu\nu}$ is given by

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - g_s f_{abc} A^b_\mu A^c_\nu \tag{2.1.3}$$

A^a_μ are the gluon vector fields with color $a = \{1, 2, \ldots, 8\}$ and g_s is the strong coupling constant which measures the strength of the interaction. The functions f^{abc} are the structure constants of the $SU(3)$ group and the indices a, b, c run over the eight color degrees of freedom. The reason the gluons are massless is explained by the impossibility of including a mass term such as $m^2 A^\mu A_\mu$ to $\mathcal{L}_{QCD}$ whilst maintaining gauge invariance [5]. D_μ denotes the covariant derivative of the quark fields. It is given by the following relationship

$$D_\mu q = (\partial_\mu + i\frac{g_s}{2} A^a_\nu \lambda_a)q \tag{2.1.4}$$

λ_a are the *Gell-Mann* matrices which make up the linearly independent generators of the $SU(3)$. QCD is a so-called *non-Abelian* gauge theory, *i. e.* the generators of

[2]However, the time reversed process is less likely due to phase space arguments like mass and energy conservation.

[3]In general, the Einstein notation applies, *i. e.* indices appearing twice in a single term (and not otherwise defined) are to be summed over their respective ranges.

the $SU(3)$ group do not commute. Instead, they obey a the following relationship

$$[\lambda_a, \lambda_b] = i f^c_{ab} \lambda_c \tag{2.1.5}$$

Here, the structure constants f^{abc} of the $SU(3)$ appear again defining the commutation relationship between the generators of the group. The formula Eq. 2.1.5 is the origin of the gluon self-interaction term $-g_s f_{abc} A^b_\mu A^c$ in the Lagrangian, enabling three or four pure gluon vertices. This is contrasted with the $U(1)$ symmetry group of electrodynamics which contains no self interaction terms, *i. e.* the photon field does not carry an electric charge and, therefore, cannot interact with itself.

2.1.2 The running coupling constant

The interaction terms in the Lagrangian can be used to calculate the probability of a particular scattering process. At leading order the coupling between particles with a color charge is described by the coupling strength g_s. The actual coupling, however, does not correspond solely to the leading order term, but includes higher order, or so called loop corrections as well. This results in a *running* coupling *i. e.* $g_s \rightarrow g_s(|q^2|)$ where the strength of the interaction is dependent on the momentum scale $|q^2|$. By convention, the measured parameter is the strong coupling $\alpha_s = \frac{g_s}{4\pi}$. For a one-loop approximation the strong coupling is given by

$$\alpha_s(|q^2|) = \frac{12\pi}{(11n_c - 2n_f)} \frac{1}{\ln(|q^2|/\Lambda^2_{QCD})} \tag{2.1.6}$$

n_c is the number of color charge states and n_f is the number of quark flavors at the momentum scale $|q^2|$. Λ_{QCD} is a free parameter and must be measured experimentally. According to the standard model $n_c = 3$ and $n_f = 6$, we can conclude that

(i) as $|q^2| \rightarrow \infty$ the coupling strength α_s approaches 0 (*asymptotic freedom*),

(ii) as $|q^2| \rightarrow 0$ the interaction becomes so strong that the colored objects are *confined* into color neutral states.

The property of asymptotic freedom makes QCD calculations tractable at high momenta by allowing the interaction to be treated as a perturbation of free fields. However, at low momentum transfer non-negligible long-range correlation and multiparticle interactions in higher-order loops make quantitative application of QCD impracticable.

2.2 Diffractive physics

One way to test the theory of QCD and the Standard Model in general is via high energy particle collider experiments. To describe the different ways a collision can occur the total cross section is used. It describes the probability that two particles will collide and react in a certain way, and generally depends on the energy of the colliding particles. The total cross section σ_{tot} is divided into two major parts: the elastic and the inelastic cross section whereas the inelastic one can be further divided into the diffractive (D) and non-diffractive (ND) cross section:

$$\sigma_{tot} = \sigma_{el} + \sigma_{inel} \qquad \text{with:} \quad \sigma_{inel} = \sigma_{ND} + \sigma_D. \tag{2.2.1}$$

Elastic processes are defined as events where the initial particles emerge from the interaction without any exchange of quantum numbers. The final state particles remain unchanged. Hence, only the kinematics of the process changed. This can be formulated as

$$1 \ + \ 2 \ \rightarrow \ 1' \ + 2' \tag{2.2.2}$$

In inelastic scattering the result are multi-particle final states X. This is expressed by

$$1 \ + \ 2 \ \rightarrow \ X \tag{2.2.3}$$

X describes the whole system of emerging particles which differ from the initial states 1 and 2. In contrast to non-diffractive events, diffractive scattering is defined as a vacuum quantum number exchange between the two initial protons (see Sec. 2.2.1). To distinguish non-diffractive events from diffractive ones at the experimental level the rapidity distribution of particles emerging in both event categories has to be considered. The rapidity relative to the beam axis of a particle is a measure of its forward momentum, which is defined as follows

$$y = \frac{1}{2} \ln \frac{E + p_z}{E - p_z} \tag{2.2.4}$$

E is the particle's energy and p_z the momentum in the initial z-direction along the beamline. In collider experiments where $m \ll p \ \Rightarrow \ E \approx p$ the rapidity is approximately equal to the *pseudorapidity* η

$$y \ \approx \ \eta = -\ln \left[\tan \left(\frac{\theta}{2} \right) \right] = \frac{1}{2} \ln \frac{|\mathbf{p}| + p_z}{|\mathbf{p}| - p_z} \tag{2.2.5}$$

θ denotes the angle enclosed between the particle and the beamline. $|\mathbf{p}|$ is the particles spatial momentum magnitude.

In contrast to diffractive events, non-diffractive collisions are characterized through a color quantum number exchange between the two protons. As neither of the systems is color neutral they are pulling a color field between them as they are moving apart. The energy of this field grows with the separation displacement. As soon as the distance reaches the order one unit in rapidity new particles are created filling the rapidity gap. Therefore, large rapidity gaps between final state particles are exponentially suppressed [6]:

$$\frac{dN_{ND}}{d\Delta\eta} \ \sim \ e^{-\Delta\eta} \tag{2.2.6}$$

To ensure two final states, where the quantum numbers equal those of the incoming protons, a large non-exponentially suppressed rapidity gap distribution of the final states is required:

$$\frac{dN_D}{d\Delta\eta} \ \sim \ \text{const.} \tag{2.2.7}$$

This difference in rapidity gap distribution is illustrated in Fig. 2.1.

Thus, we can summarize the two dependent conditions that characterize a diffractive event as follows:

(i) No color exchange resulting in (ii),

(ii) Constant rapidity gap distribution of the final state.

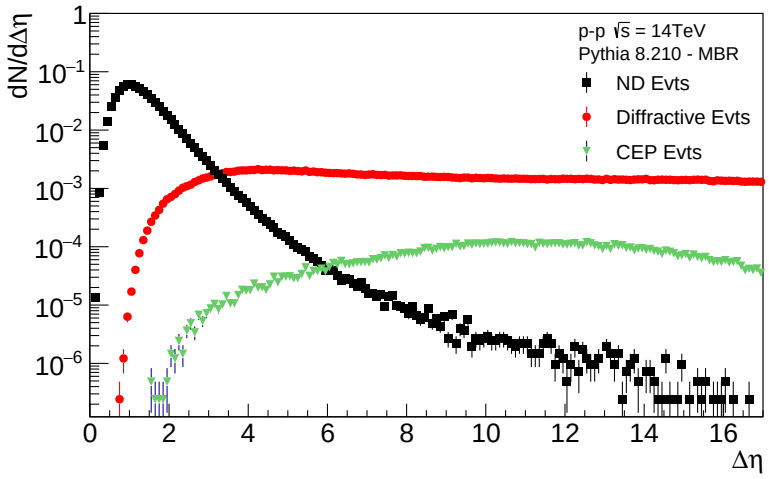


FIGURE 2.1: *Maximum rapidity gap distribution in non-diffractive and diffractive events simulated with Pythia-8 (MBR). In contrast to diffractive events, non-diffractive ones experience an exponentially suppressed rapidity gap. This can be explained by the color exchange in ND events. As ND events exchange color, the energy of the field grows with the separation displacement of the scattered particles. As the distance reaches the order of unit rapidity new particles are created filling the rapidity gap.*

Among the diffraction events we distinguish three kinds of event classes: single (SD), double (DD), and central diffraction (CD), also called central exclusive production (CEP). In single and double diffraction, either one (SD) or both incoming protons (DD) break apart after the interaction. As the exchange is still mediated by particles with vacuum quantum numbers, the decay products of the dissociated protons have net quantum numbers identical to those of the initial s ate proton. Thus, the central rapidity gap remains intact [7]. In central exclusive production both protons emerge unchanged and a single object at central rapidity - *i. e.* with a rapidity much closer to zero as the outgoing protons - is created. These processes can be schematically viewed as

$$\begin{aligned} 1 + 2 &\overset{SD}{\rightarrow} 1' + X \\ 1 + 2 &\overset{DD}{\rightarrow} X_1 + X_2 \\ 1 + 2 &\overset{CEP}{\rightarrow} 1' + X + 2' \end{aligned} \tag{2.2.8}$$

In this thesis we will focus on central exclusive production.

The energy scale at which diffractive processes with large rapidity gaps are happening is relatively low. This process is classified as belonging to the *soft* energy regime. Consequently, the running coupling constant α_s is large limiting the application of perturbative QCD as higher order terms can no longer be neglected. Therefore, an alternative description for diffraction in the soft regime is needed. A formalism describing diffractive exchanges is *Regge theory*.

2.2.1 Regge theory and the pomeron

The following section is a brief discussion of Regge theory and the pomeron approach will be held very briefly. For a more holistic approach see [6, 7]. Historically, Regge theory was introduced to describe the strong interaction. This approach was later succeeded by QCD.

At its core, Regge theory studies the properties of scattering as a function of angular momentum, which is not quantized in multiples of $\hbar$ but treated as a complex variable. Initial free particles at the time $-\infty$ interact via the unitary scattering matrix at time 0 producing final free particle states at time $+\infty$ [8, 9]. Hadronic interactions at large energies are described by assuming the exchange of an object called *Reggeons*. These Reggeons carry angular momentum $\alpha(t)$ which have a functional dependence on the four-momentum transfer $t = (p_1 - p_{1'})^2 = (p_2 - p_{2'})^2$. A consequence of its $t-$dependence the Reggeon is not represented by a single physical particle but instead it is associated with a superposition of multiple particles (mesons) that all contribute simultaneously to the total cross section. This superposition of particles all follow the following function [10]

$$\alpha(t) \;=\; \alpha(0) \;+\; \alpha' t \tag{2.2.9}$$

This linear function is called a Regge trajectory. The slope α' and intercept $\alpha(0)$ can be measured by fitting the angular momentum against the squared mass of light mesons [11] as shown in Fig. 2.2. This results in an intercept $\alpha(0) \sim 0.5$. For large values of s, the center of mass energy squared, the total cross section shows the following relationship

$$\sigma_{tot} \;\propto\; s^{\alpha(0)-1} \tag{2.2.10}$$

As the intercept $\alpha(0)$ is roughly $\alpha(0) \simeq 0.5$, the dependence is assumed to be $1/\sqrt{s}$. Therefore, as the center of mass energy increases, the total cross section is expected to vanish asymptotically.

However, as illustrated in Fig. 2.3, at around 20 GeV the total cross section is increasingly defying the predicted trend. To explain the observed energy dependence of σ_{tot} a Regge trajectory with an intercept $\alpha(0) > 1$ is introduced, which results in a positive exponent (see Eq. 2.2.10) and leads to a rising total cross section with an increase in s.

This trajectory called the *Pomeron* is the dominant exchange propagator in diffractive processes. Contrary to the mesonic Reggeon, the Pomeron is not expected to be based on quarks. In the current set of known compound particles none can be attributed to lie on the Pomeron trajectory. The glueball, however, a hypothetical particle consisting only of gluons[4], would classify as a candidate since the simplest exchange of vacuum quantum number is via a pair of gluons in a color singlet state [1].

2.2.2 Central exclusive production

Central exclusive production is defined as a diffractive process in which the two incoming particles do not disintegrate and a single object is produced by a color-less

[4]According to QCD, gluons carry color charge and, thus, interact with themselves enabling such a state to exist.

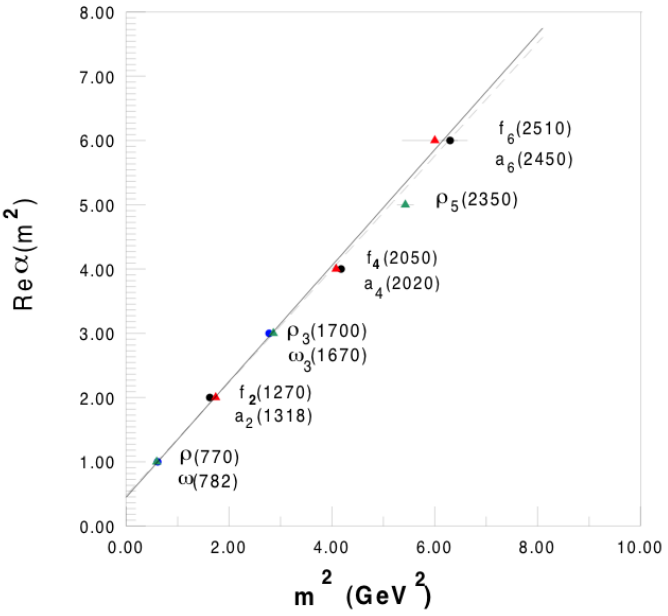


FIGURE 2.2: *Example of a Regge trajectory: The Regge trajectory describes a quadratic relationship between the mass of mesons and resonances and their angular momentum. The parameters of the Regge trajectory are extracted by finding the best fit [12].*

exchange at central rapidity. In proton-proton interactions it can be schematically written as

$$pp \rightarrow p + X + p \tag{2.2.11}$$

X is the system produced at mid-rapidity, which is separated from the outgoing protons by large rapidity gaps. There are three processes that contribute to the t-channel exchange of a color-singlet object: diphoton fusion, photoproduction, and double Pomeron exchange (DPE). Diphoton fusion is a pure QED process where both protons radiate a photon that fuse to produce the central system like $\gamma\gamma \rightarrow X$. Photoproduction and DPE are both processes that involve the emission of a Pomeron. In double Pomeron emission, as the name suggests, both protons emit a Pomeron, which then join together and create the central system. Photoproduction can be seen as the intermediate process between diphoton fusion and DPE. It describes the emission of both a photon and a Pomeron fusing together to create the X system. At high energies Pomeron mediated processes are expected to predominantly contribute to the CEP cross section since the Pomeron is strongly interacting [14].

The characteristics of CEP entail some interesting features of the centrally produced system X. Since the protons stay intact and the Pomeron carries vacuum quantum numbers, X must be a color singlet. Such a state is even under charge conjugation as well as parity transformation[5]. Collectively they represent the selection rule for the centrally produced particle: $J^{PC} = (even)^{++}$ [15, 16]. Additionally, the protons in CEP scatter in the very forward region. As the scattering angle between the in- and outgoing protons decreases asymptotically to zero, so does $J \rightarrow 0$ by conservation of angular momentum. For small non-zero scattering angles this

[5]This means that the charge conjugation and parity operation are C= +1 and P= +1 respectively.

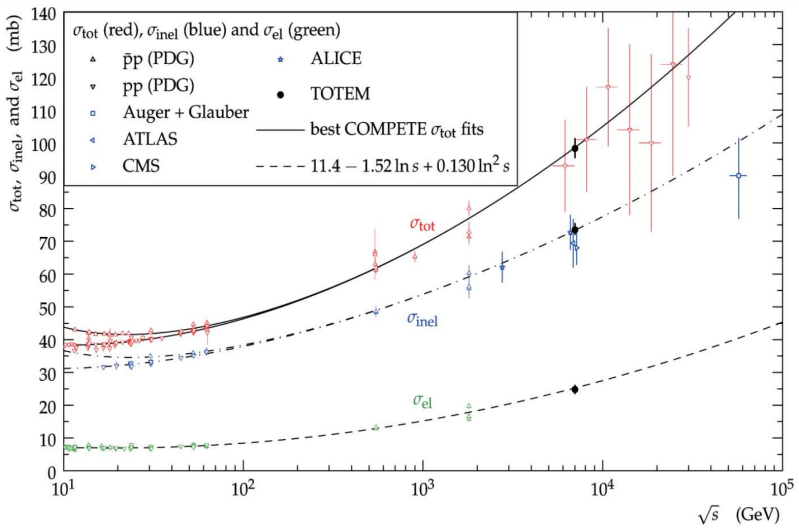


FIGURE 2.3: *Total cross section over center of mass energy $\sqrt{s}$: At $\sim$ 20 GeV the cross section increases contrary to predictions from Regge theory with pure Reggeons [13].*

rule still approximately holds true and large values of J are greatly suppressed [17]. Thus, a large contribution to the experimental signature of the CEP system is expected to consist of scalar mesons[6] and potentially glueballs. The importance of the study of these states is mentioned in the 2010 Particle Data Group Note on scalar mesons [18]: *"The scalar mesons are especially important to understand because they have the same quantum numbers as the vacuum ($J^{PC} = 0^{++}$). Therefore they can condense into the vacuum and break a symmetry such as a global chiral $U(N_f) \times U(N_f)$. The details of how this symmetry breaking is implemented in Nature is one of the most profound problems in particle physics."*

2.3 The ALICE experiment

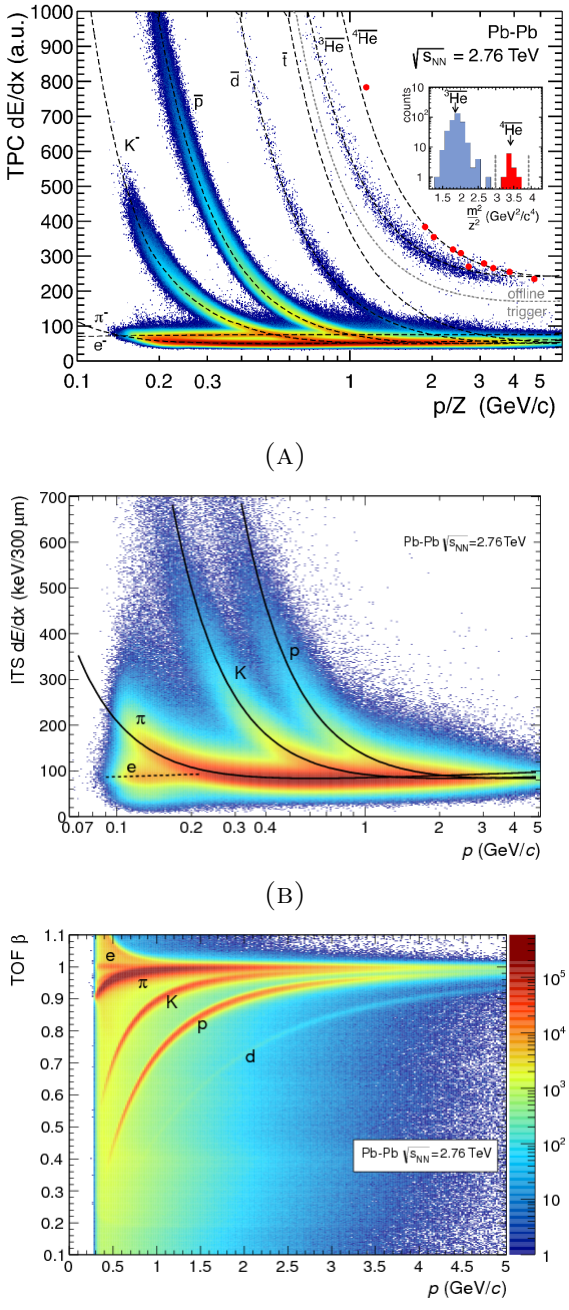
ALICE[7] is a general-purpose heavy-ion experiment at the CERN large hadron collider (LHC) that aims to study the physics of strongly interacting matter in heavy-ion, proton-ion, as well as proton-proton collisions [19]. The detector is built around the interaction point at *Point 2* where the decay products of the colliding particles are measured by its 18 sub-detector systems [20]. The sub-detectors can be categorized into the central barrel, the forward muon spectrometer, and additional detectors for event characterization and trigger purposes lying outside of the central barrel.

The central barrel is made up of the Inner Tracking System (ITS) [21], Time Projection Chamber (TPC) [22], Time-of-Flight detector (TOF) [23], High Momentum Particle Identification Detector (HMPID) [24], Transition Radiation Detector (TRD) [25], Electromagnetic Calorimeter (EMCal) [26, 27], and the Photon Spectrometer (PHOS) [28]. This set of detectors provide excellent particle tracking and identification capabilities in the mid-rapidity region around $|\eta| < 0.9$ in the whole 2π azimuthal (ϕ) range. Additionally, the inner tracking system which is located

[6]Scalar mesons have quantum numbers $J^{PC} = 0^{++}$

[7]A large ion collider experiment

very close to the beamline covers an extended rapidity range of $|\eta| < 2$ and $|\eta| < 1.4$ with its inner and outer layers of silicon pixel detectors respectively. The particle identification (PID) performance is illustrated in Fig. 2.4. It shows a very good ability to identify even low-momentum particles which is a key feature for studying CEP events [29].



(A)

(B)

(C)

FIGURE 2.4: *PID performances of barrel sub-detector systems (see [30]): (A) TPC momentum versus deposited energy per unit length (dE/dx). (B) ITS, energy loss (dE/dx) versus momentum. (C) TOF signal versus momentum.*

Among the forward detectors outside of the central region are the V0 [31], the T0 [31], the Forward Multiplicity Detector (FMD) [31] and the ALICE diffractive de-

tector (AD) [32]. The V0 detector system consists of two scintillator arrays situated on both sides of the central barrel in the pseudorapidity regions of $-3.7 < \eta < -1.7$ and $2.7 < \eta < 5.1$, respectively. The FMD detector consists of five silicon semiconductor sub-detectors which are used to estimate the charged particle multiplicity in events. Its acceptance overlaps with the V0 detector with a pseudorapidity range of $-3.4 < \eta < -1.7$ and $1.7 < \eta < 5.0$. During the long shutdown 1 of LHC the AD system, an additional forward detector was installed by the ALICE Collaboration. The AD consists of two modules one on each side of the interaction point made of two layers of scintillator pads in the pseudorapidity region of $-7.0 < \eta < -4.9$ and $4.8 < \eta < 6.3$, respectively. This upgrade has therefore considerably increased the forward coverage of the ALICE detector to over 12 units in pseudorapidity. This makes, combined with the excellent low-momentum, tracking in the central barrel ALICE well suited for diffractive studies. In general, all detectors cover the full azimuthal range except for HMPID, PHOS, and EMCal+DCal. In Fig. 2.5a a schematic view of the detector is shown and Fig. 2.5b illustrates the pseudorapidity coverage of the individual sub-detectors.

Central exclusive production events are defined experimentally by activity in the mid rapidity region, *i. e.* in the central barrel and an activity gap in the forward direction. At ALICE, this topology can be implemented at trigger level zero (L0) by requiring hits in the ITS or TOF systems [33]. The double-gap condition is realized by the absence of V0 signal. In the offline analysis additional information from FMD, TPC, and AD detectors extend the activity gap region to over 12 units in pseudorapidity.

Since the protons remain intact, CEP events provide a distinct low multiplicity signature to search for in the detectors. Therefore, the mass of the centrally produced particle X can be measured with a high degree of accuracy if the momenta of the outgoing protons are detected. Additionally, the knowledge of the outgoing proton momenta can be used to reduce background sources from partially reconstructed X masses where one or multiple decay products of the central system were not detected. However, despite the wide η-coverage of the ALICE detector system the scattered protons do not enter the detector acceptance due to the very small scattering angle. Consequently, a significant amount of background arises from partially reconstructed events.

Although not available at ALICE, these proton momenta measurements can be achieved using ultra forward detectors typically housed in so called roman pots in the beamline itself [34].

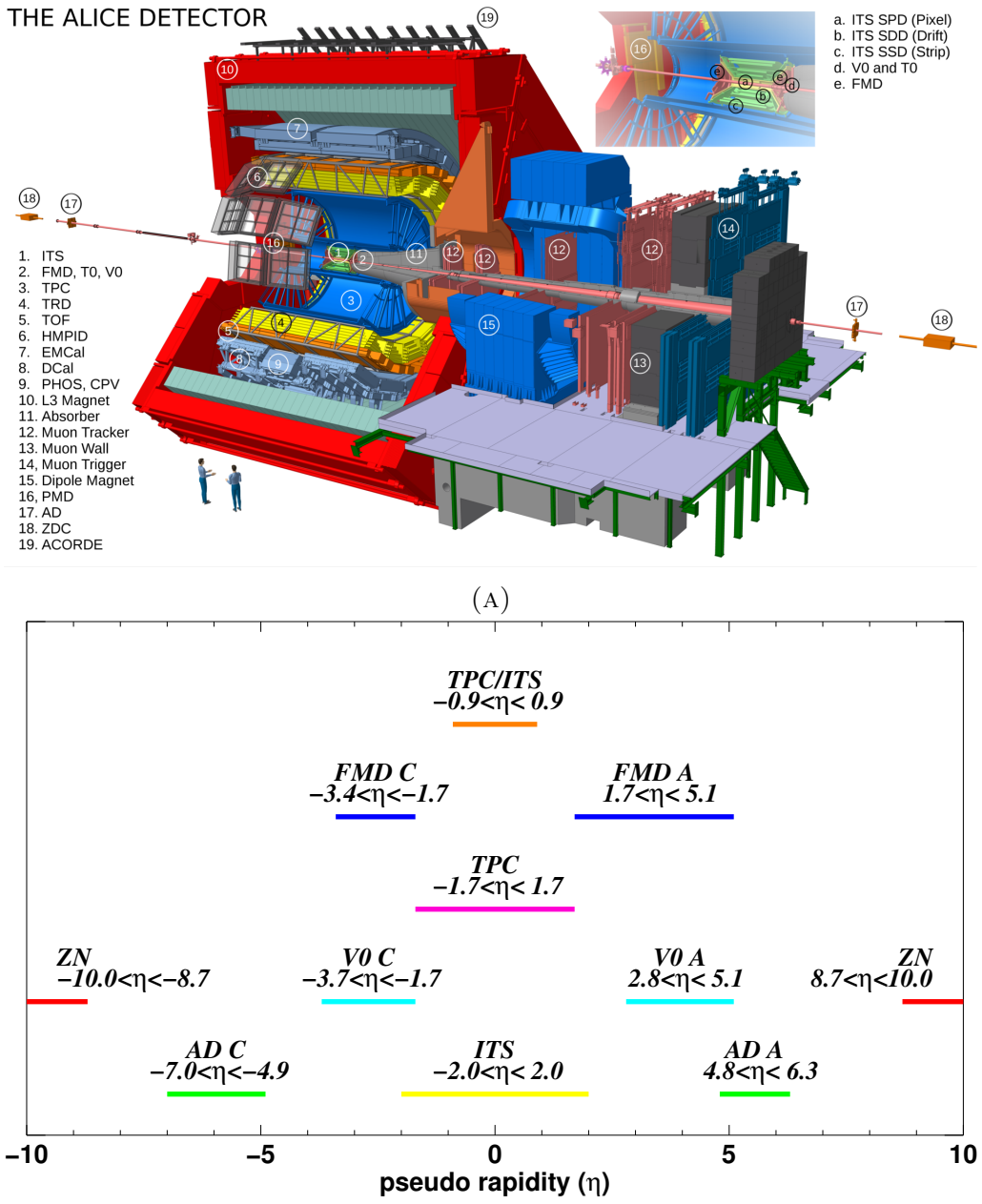


FIGURE 2.5: *Two representations of the ALICE detector components: On top (A) the detector components are displayed in schematic drawing for LHC Run 2. Below that, the pseudorapidity coverage of the sub-detectors is shown (B) illustrating the ability to perform diffractive studies over a large rapidity region.*

Chapter 3

Background studies

This thesis aims at understanding and subsequently reducing background (BG) sources in the analysis of CEP events at ALICE. For this purpose Monte Carlo (MC) simulations are used to generate a data set $\{\boldsymbol{x}_k\}_{1\leq k\leq N}$. This data set consists of N independent and identically distributed samples drawn from an underlying distribution $p(\boldsymbol{x}|\theta)$, where the parameter θ corresponds to the setting of the simulator [35]. These settings θ are formulated in theoretical frameworks, *e. g.* QCD or Regge Theory, describing production, decay, or annihilation mechanisms, which have to be established beforehand. Monte Carlo methods are used to approximate the probability $p(\boldsymbol{x}|\theta)$ by sampling from a large space of unobserved processes: $p(\boldsymbol{x}|\theta) = \int p(\boldsymbol{x}, \boldsymbol{z}|\theta)d\boldsymbol{z}$ [35]. The variable $\boldsymbol{z}$ is generally regarded as the *MC truth*. $\boldsymbol{z}$ describes the realm of possible event configurations where a fixed value entirely predefines all event characteristics: *i. e.* its kinematics, the initial particles created from the scattering process as well as individual particle-detector interactions. Standard reconstruction algorithms make estimates on a subset of $\boldsymbol{z}$ components such as particle momenta, energies and particle identification (PID) given the observed data $\boldsymbol{x}$.

MC simulations rely heavily on good theoretical models and have to be constantly compared and tuned to real data. The major advantage of using a simulated set of data over a real one is the precise knowledge of the "observed" data $\boldsymbol{x}$ via $\boldsymbol{z}$. Therefore, MC simulations provide a tool to study the resulting background mass spectrum in any degree of detail. This opens up the possibility to find intrinsic mechanisms to reduce it.

3.1 Used framework & data sets

A widely used high energy pp MC simulation package that includes diffractive physics is called *PYTHIA-8* [36]. It is used in this thesis to generate the event sample $\{\boldsymbol{x}_k\}_{1\leq k\leq N}$. To simulate the diffractive processes the MBR (Minimum Bias Rockefeller) model [37] is used. The MBR description is the most recently implemented diffractive model, which generates events following a renormalized-Regge-theory approach. The default numeric values of the cross sections of each subprocess, *i. e.* non-, single-, double-, and central-diffractive processes, have been used. This includes parameters describing the Pomeron trajectory (see Sec 2.2.1) labeled internally as ε and α, which describe the intercept above 1 and the slope of the curve, respectively. Quite often the parameter ε is varied from the default param-

eter of $\varepsilon = 0.104$. However, in this thesis the default configuration are used. This is symbolized by ($\varepsilon = 0.104$) in the following plots. Although MBR is quite successfully tested on data it describes only non-resonant continuum production of the centrally produced particle (X) in CEP events at masses ≥ 1.5 GeV/c^2. A comparison of continuum and resonant production of a 2π final state can be seen in Fig. 3.1. The mediator particles shown (*i. e.* $\gamma, \mathbb{P}, \mathbb{R}$) are the photon, the Pomeron, and the Reggeon. For a summary on the possible production mechanisms see Sec. 2.2.2. The generated mass distribution (MBR) up to 4 GeV/c^2 is plotted in Fig. 3.2.

Additional imprecisions may arise as described by Lebiedowicz *et al.* [38] due to absorption effects which may favor the cross section of photoproduction processes over DPE. This would lead to an increased number of ρ final states in the $\pi^+\pi^-$ spectrum. This effect is also not included in the MBR simulation model. Despite some limitations the benefits of using a carefully tested general purpose MC simulation providing precise knowledge of all kinematic and PID information and a good understanding of a large portion of the CEP decay channels (made up of continuum events) is crucial for this study. Thus, the background reduction results discussed in the following sections lack information of the background contributions of photoproduction and resonantly produced CEP particles. Nevertheless, continuum produced CEP events are assumed to make up a large portion of the general CEP spectrum [38] making the study of their contribution to the background a vital task.

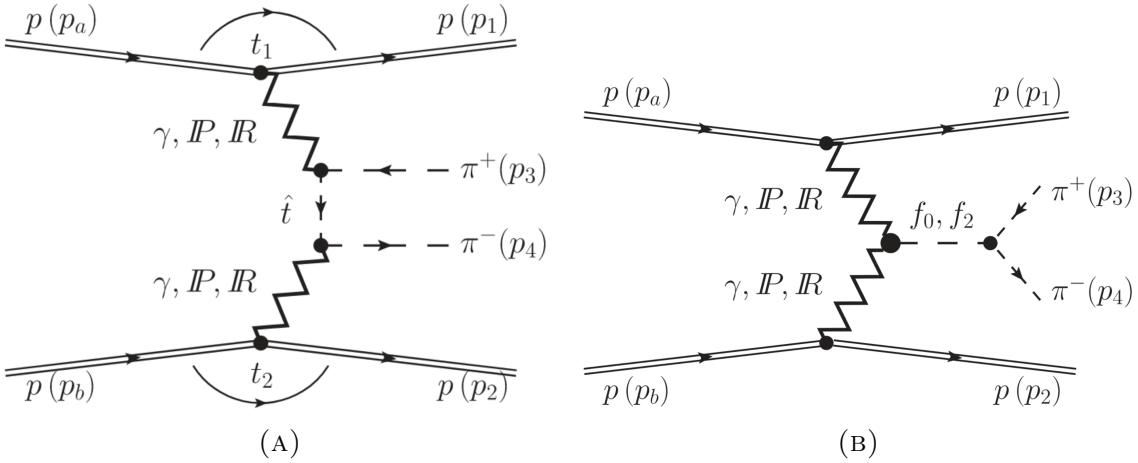


FIGURE 3.1: *Feynman graphs of continuum (A) and resonant (B) 2π production in central exclusive events [38].*

The tracking and interaction of the generated particles with the material of the detector is simulated with GEANT [39] and happens within the ALICE software framework. Altogether, a data set comprised of approximately three million CEP events has been simulated[8]. In order to study the background the data set is prefiltered by enforcing a double gap selection as well as track quality cuts summarized in Tab. 3.1. It should be noted that this thesis studies the background contribution in the $\pi^+\pi^-$ data. This decay channel is chosen as it provides the largest amount of data. However, the following methods apply also to different final states such as: K^+K^- or their respective four particle final states $2\pi^+2\pi^-$ and $2K^+\,2K^-$. To summarize, the data set is obtained in the following way. First, event selection according

[8]The data is internally classified as belonging to the run-period of 2016k

to Tab. 3.1 is applied yielding two accepted pion tracks. For this study perfect particle identification is assumed, as the focus lies on general background characteristics and not on background introduced by false PID estimates. PID studies is a separate field of study which goes beyond the constraints of this thesis. The application of the prefilter results in a data set of roughly 2×10^4 events (see Sec. 3.2 for more details) which entail a reduction factor of $\sim 10^2$. Second, the invariant mass of the particle X is calculated via the measured energy E_i and momentum $\vec{p}_i$ of the detected pions ($i = 1, 2$ and $c = 1$) as follows

$$M = \sqrt{\left(\sum_i E_i\right)^2 - \left(\sum_i \vec{p}_i\right)^2} \tag{3.1.1}$$

To process and analyze the data the AliRoot framework [40] is used which is an extension to ROOT v5.34/30 [41] a scientific software framework.

Cuts applied to the simulated data

1. Double gap:
   !V0, !FMD, !AD

2. Track cuts:
   - *Two (2) tracks reconstructed in the TPC & ITS*
   - *Good tracks quality:* $\chi^2/dof < 4$
   - $DCA_z < 0.5$ *cm*
   - *Tracks require at least 70 pad hit clusters in the TPC*
   - *Tracks have to be in* $|\eta| < 0.9$ *due to bad tracking outside*
   - *The number of SPD fired chips has to be* $\leq N_{tracks}$

3.2 Double gap selection & the invariant mass spectrum

CEP events at ALICE are selected via a double gap condition requiring activity in the central barrel and the absence of signal in the forward region (discussed in detail in Sec. 2.3). However, this trigger mechanism does not explicitly specify the size of the η-gap. In order to decrease the non-diffractive background component the rapidity gap condition outside the barrel is maximized. While the rapidity gap distribution for non-diffractive events decreases exponentially, the diffractive η-gap distribution stays constant (as discussed in Sec. 2.2, see Fig. 2.1). This shape difference between non-diffractive and diffractive events can be exploited. A larger $\Delta\eta$ eventually means a better signal to background ratio[9]. In Fig. 3.3 the influence of different rapidity gap sizes can be seen. The three sub-plots each feature the 2π invariant mass spectrum measured in the central barrel region over the relative count. Here, the specific detector simulation and tracking of the individual particles is reduced to a simple detector acceptance cut in ϕ and η. *I. e.* a charged particle

[9] *I. e.* a better ratio of diffractive to non-diffractive events.

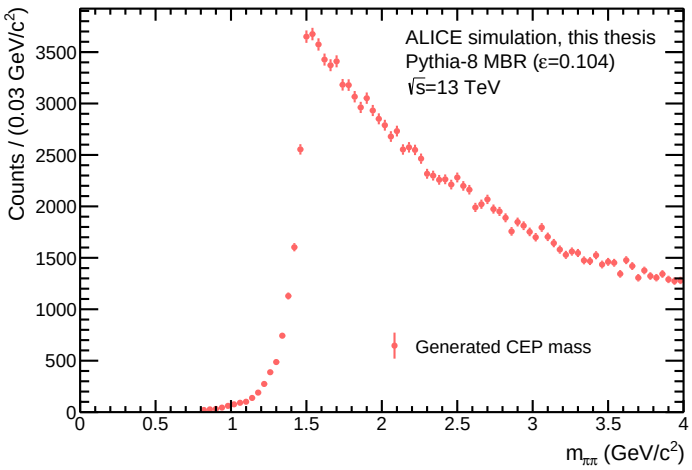


FIGURE 3.2: *Generated mass distribution of the centrally produced X particle in Pythia-8 MBR CEP simulations. ε refers to an important parameter of the MBR simulation scheme which describes the intercept of the Pomeron trajectory above* 1. *This spectrum shows no structures as the MBR model simulates only non-resonant continuum DPE production of the X mass ≥ 1.5 Gev/c^2.*

is detected if its trajectory coincides with a detectors spatial coverage. This assumes perfect detection efficiency of charged pions, which is justified by the aim of highlighting the effects of a variable rapidity gap imposed on different background components (non-, single-, and double-diffractive BG). The left plot illustrates the invariant mass spectrum in absence of a double gap condition: here, all events with exactly two detected pions ($\pi^+\pi^-$) in the central barrel are plotted. As the imposed rapidity gap $\Delta\eta$ increases from the left panel to the right one, the background contribution from ND, SD, and DD events decreases and approach zero. The increasing η-gap is accomplished by successively requiring no signal in the FMD and V0 (in the middle panel) and no signal in FMD, V0, and the AD detector system in the right most panel. This double gap condition eliminates nearly 100% of the ND, SD, and DD background components. The remaining sample is defined as *feed-down* (FD) background. The source of FD are CEP events themselves. These events are only partially reconstructed, *i. e.* the detected pions come from $n > 2$ final state events, where n is the total number of final state particles[10]. This means that at least one final particle generated in the X decay remains undetected, resulting in a loss of mass and energy. As a consequence, the reconstructed invariant mass of the X particle is understated which induces a shift of the FD invariant mass spectrum towards to lower masses. Since feed-down represents the majority of the background portion, it is crucial to understand its origin and composition in order to describe and ideally eliminate it. One option to reduce feed-down events on an experimental level is to introduce detectors measuring the scattered protons in ALICE. By combining the four momenta of the scattered protons with the two measured pion tracks one can check for a deviation from the initial center of mass energy of $\sqrt{s} = 13$ TeV (and zero three-momentum), which concludes a background event. As this is no viable

[10]This information is available by inspecting the MC truth $\boldsymbol{z}$ of the event.

option, two alternative methods are considered here: FD *description* and *reduction* via multivariate methods.

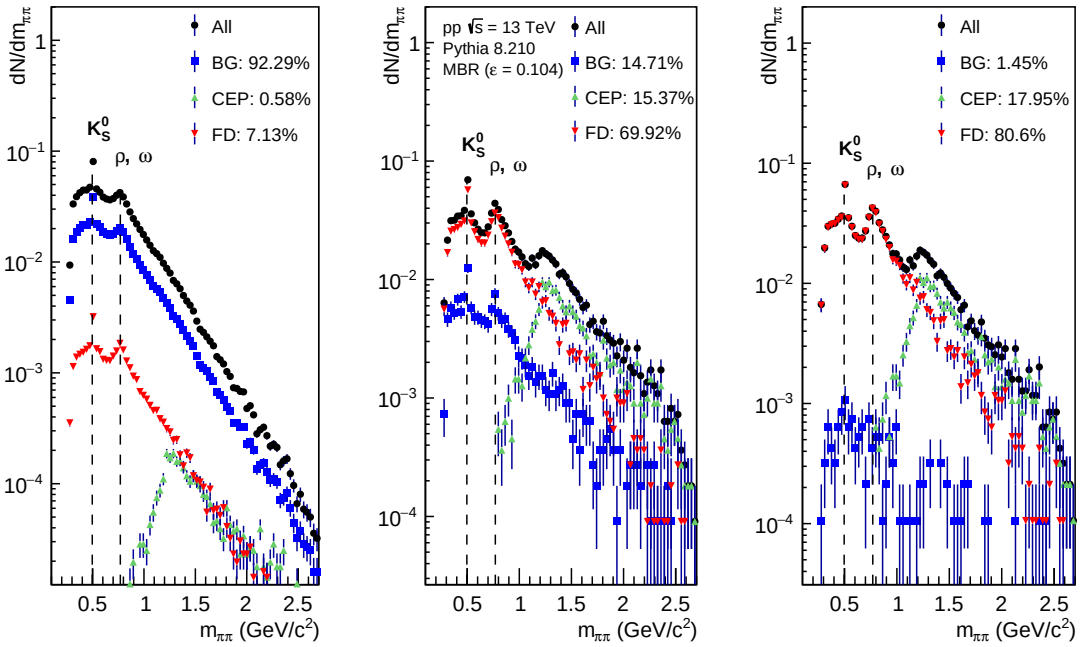


FIGURE 3.3: *Invariant mass distribution for different rapidity gap filters. The non-diffractive, SD, and DD background can be reduced by selecting a large rapidity gap as is done by using the FMD, V0 and AD detector systems. See Sec. 3.2 for more details.*

3.3 Background estimation

The goal of a background estimation study is to construct a representative template of the background shape which can then be subtracted from the whole data, yielding the excessive data as signal. To extract the signal yield, a background as well as a signal shape are used to fit the data. A common approach to describe the combinatorial background, *i. e.* the background which arises from pairs of particles originating from different mother particles[11], is to employ the *like-sign* (LS) method. The like-sign method constructs a combinatorial BG estimation from pairs of two positive or two negative pions, respectively. These pairs of identical charge cannot be the only two particles originating from X (due to charge conservation, since X has vacuum quantum numbers) and consequently their mass spectrum is expected to have a similar shape as the contribution from uncorrelated opposite-sign pairs [42]. The probability of measuring a pair of opposite-sign pions is determined by the number of available positive N_+ and negative N_- pions, respectively. The total

[11] *I. e.* the detected tracks are *combined* in the wrong way. Therefore, the combinatorial background consists of totally uncorrelated particles.

number of available pions is $N = N_+ + N_-$

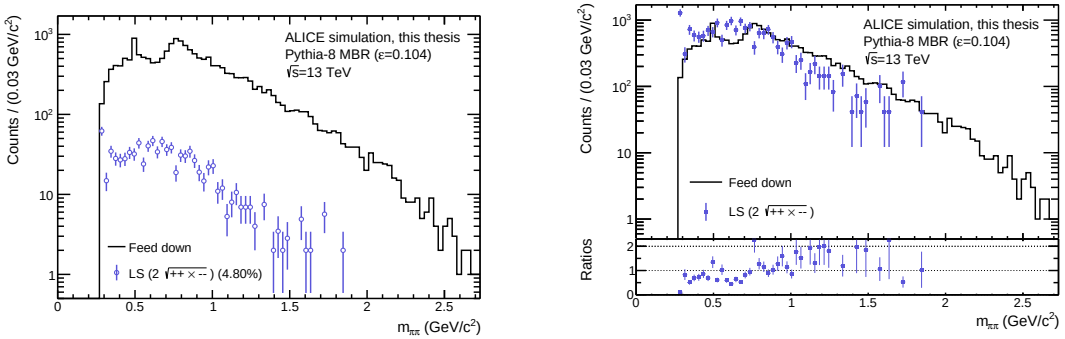
$$\begin{aligned} P(+-,-+) &= P(+)P(-) + P(-)P(+) = \\ &= \frac{N_+}{N}\frac{N_-}{N-1} + \frac{N_-}{N}\frac{N_+}{N-1} = 2P(+)P(-) \end{aligned} \tag{3.3.1}$$

The like-sign ansatz is formulated in the following way

$$\begin{aligned} P(+-,-+) = 2P(+)P(-) &\simeq 2\sqrt{P(+)P(+)P(-)P(-)} \\ &= 2\sqrt{P(++)P(--)} \end{aligned} \tag{3.3.2}$$

The probabilities are defined as $P(++) = \frac{N_+}{N}\frac{N_+-1}{N-1}$ (the same goes for $P(--)$ with N_-). Consequently, the background (with $N > 2$) can be estimated by combining the measurement of positive and negative like-sign pairs such as $P(+-,-+) \simeq 2\sqrt{P(++)P(--)}$.

However, the comparison of the LS background estimation in Fig. 3.4a shows that the like-sign hypothesis underestimates the feed-down substantially. In addition to underestimating the feed-down background, the like-sign distribution in Fig. 3.4b yields a poor approximation for the feed-down shape. A hint for the unrelated results



(A) (B)

FIGURE 3.4: *Like-sign approximation of the combinatorial background: Left: Total like-sign approximation using positive and negative like-sign pairs with* $2\sqrt{(++)(--)}$. *The like-sign hypothesis states that these curves should match. Right: A shape comparison between feed-down background (black) and the like-sign (blue) modeling of the combinatorial background yields a rather unrelated description of the data.*

of the like-sign background estimation can be found by studying the uncertainty in Eq. 3.3.2: *i. e.* when $P(+-,-+)$ equals $2\sqrt{P(++)P(--)}$.

$$\begin{aligned} P(+-,-+) &\stackrel{?}{=} 2\sqrt{P(++)P(--)} \\ 2\frac{N_+N_-}{N(N-1)} &= 2\sqrt{\frac{N_+(N_+-1)\ N_-(N_--1)}{N^2(N-1)^2}} \\ N_+N_- &= \sqrt{N_+(N_+-1)\ N_-(N_--1)} \end{aligned} \tag{3.3.3}$$

The equal sign in Eq. 3.3.3 applies for $N_{+/-} = N_{+/-} - 1$. Hence, the error introduced by the like-sign estimation decreases for large $N_{+/-}$. In extreme cases, such as $N = 3$

either $N_+ - 1$, or $N_- - 1$ becomes zero, the right hand side of Eq. 3.3.3 vanishes, which causes a drastic underestimation of the like-sign assumption (in Eq. 3.3.2). In fact, cases with low N (*i. e.* few available pions) tend to be the norm rather than the exception when studying the feed-down composition. The main decay channels are listed in Tab. 3.1 (an extended table is listed in Tab. B.1 in the appendix).

Decay	Occurrence[%]	Cumulative [%]
X → π^+, ρ^- → π^0 → $\gamma\gamma$, π^-	21.82	21.82
X → $\pi^+\pi^-$	19.66	41.48
X → π^0 → $\gamma\gamma$, ρ^0 → $\pi^+\pi^-$	7.75	49.23
X → π^0 → $\gamma\gamma$, π^-, ρ^+ → π^0 → $\gamma\gamma$, π^+	5.37	54.60
X → π^0 → $\gamma\gamma$, π^+, π^-	4.20	58.80
X → π^0 → $\gamma\gamma$, ω → π^0 → $\gamma\gamma$, π^+, π^-	3.64	62.44

X ρ^+ π^0 $\gamma\gamma$ π^+ ρ^- π^0 $\gamma\gamma$ π^-	3.52	65.96
X π^0 $\gamma\gamma$ ρ^+ π^0 $\gamma\gamma$ π^+ ρ^- π^0 $\gamma\gamma$ π^-	1.23	67.19
X π^0 $\gamma\gamma$ π^0 $\gamma\gamma$ ρ^0 $\pi^+\pi^-$	1.20	68.40

TABLE 3.1: *Decay channels in the feed-down background listed by highest relative occurrence. The nine most frequent decay modes make up over 2/3 of all detected events. The second most frequent event is what we consider signal. Here the central system X decays into two pions, which both get measured up in the detector. All depicted background events - highlighted in red - show X decaying into two pions accompanied by two additional final state gammas. Their energy is not reconstructed in the ALICE detector and therefore is missing when constructing the invariant mass of X.*

The table features decay channels found in the data, sorted by highest relative occurrence. The left column schematically illustrates the decay chain in a hierarchical way featuring all intermediate and final state particles[12] Additionally, the cumulative occurrence of all decay modes is reported. Red colored rows refer to

[12]Final state particles are defined by a life time which is long enough to reach the main detector components in ALICE.

feed-down events with two charged pions accompanied by additional final state photons (γ background). The table features the nine most common decay channels in the data which together make up roughly 68 %. The second most frequent event is what we consider signal, *i. e.* the central system X decays into two pions which get measured in the detector. The other decay channels listed are gamma accompanied background processes. The photon energies are not reconstructed[13] in the ALICE detector and therefore go missing. The feed-down contributions can be categorized into three groups (see Fig. 3.5) depending on the background composition (*i. e.* the additional undetected particles: $\pi^+\pi^- + N_{undet}$). First, the largest group consists of events with two pions accompanied only by additional final state gammas, *i. e.* $\pi^+\pi^- + N_\gamma$. This "gamma background" accounts for a little over 83% of all feed-down events. Second, with $\sim 12\%$, events with additional charged particles, *i. e.* $\pi^+\pi^- + N_{charged} + (N_{neutral})$, are considered. These decay channels possibly also include neutral particles like photons but most importantly have more than two detectable[14] charged particles. This contribution is referred to (in this work) as 3+ background. Third, the least frequent decay channels feature neutral particles besides photons, *e. g.* neutrons, and neutral kaons such as $K^0_{L/S}$, which make up around 4% of the total feed-down.

Furthermore, the knowledge about the feed-down composition (illustrated in Tab. 3.1) helps explain the structures present in the dipion invariant mass spectrum (plotted in Fig. 3.5). At roughly 0.77 GeV/c^2 a dominant ρ^0 peak is present. Many decay channels (*e. g.* the third most frequent in Tab. 3.1) feature a neutral ρ-meson which decays into two charged pions, which finally get measured in the central barrel. Alongside the ρ^0 additional particles (*e. g.* neutral pions π^0) are produced whose decay products (primarily gammas) go undetected. Therefore, the measured charged pions produce an invariant mass contribution at the ρ^0 mass at $\simeq$ 0.77 GeV/c^2. The same is true for neutral kaons, *i. e.* K^0_S, which also decay into two charged pions creating a peak at $\simeq$ 0.49 GeV/c^2. Additional less prominent peaks also exist in the data, however, their contribution is rather small, compared to decay channels which involve processes such as $\rho^0 \to \pi^+\pi^-$ and $K^0_S \to \pi^+\pi^-$. The remaining feed-down decay channels can be considered as combinatorial background contributions.

Since these structures arise from correlated particles, the like-sign method falls short to describe such events. The LS approximation can, therefore, only model the 3+ FD contribution, amounting to merely 12%. Within the 3+ background contribution many decay channels exist, which feature other extra charged particle types besides pions, such as kaons. This background contribution is also not describable via the like-sign approximation. Therefore, the like-sign method even under-represents the 3+ background contribution, plotted in Fig. 3.6.

Consequently, the like-sign estimation fails to capture the essence of a large potion of the background. In an effort to better approximate the background shape, two alternative approaches are attempted: the γ-hit and the 3+ track estimation describing a majority of feed-down events.

[13] *I. e.* in the current state of the analysis. See the next sections for more details.

[14] Detectable in a sense that the detection efficiency of a charged track entering the detector acceptance is - within a certain momentum range - approximately equal to one. Therefore, charged particles entering the detector acceptance are significantly more likely to be detected than photons in the the calorimeter acceptance.

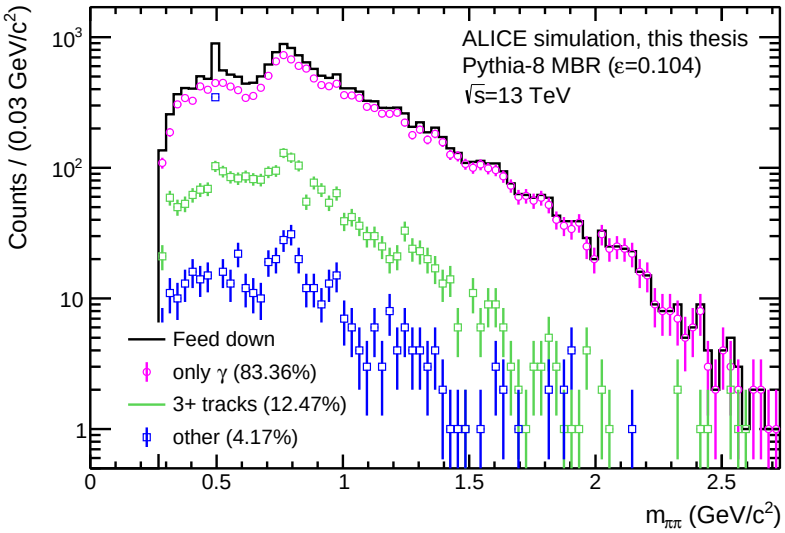
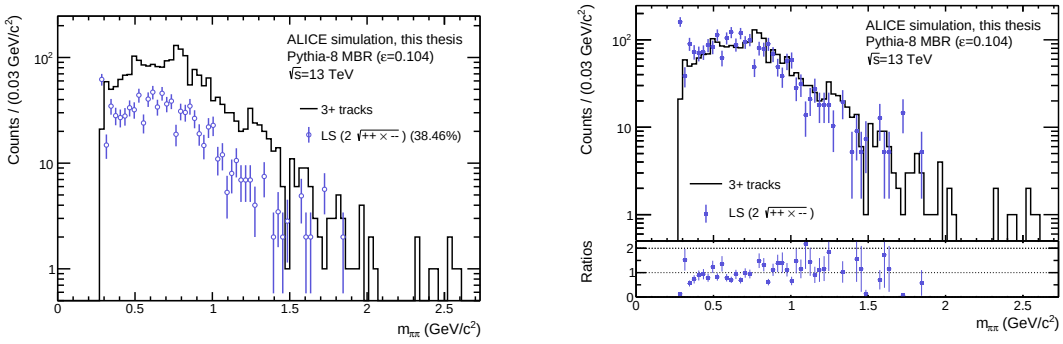


FIGURE 3.5: *Invariant mass distribution of various feed-down contributions: The feed-down background can be categorized into three groups. First, gamma-component which consists of $\pi^+\pi^-$ events which are accompanied only by additional photons. Second, the 3+ contribution consisting of decay channels with more than two charged particles. Third, the least frequent decay channels feature neutral particles besides photons.*



(A) (B)

FIGURE 3.6: *Like-sign approximation of the 3+ background component: Left: Total like-sign approximation using positive and negative like-sign pairs with $2\sqrt{(++)(--)}$. The like-sign hypothesis states that these curves should match. Right: A shape comparison between the 3+ background component (black) and the like-sign (blue) approximation of the combinatorial background yields a rather unrelated description of the 3+ FD part.*

3.3.1 The γ-hit background estimation

As seen in Fig. 3.5 a large fraction of the feed-down can be attributed to events containing final states photons which are invisible to the ALICE detector systems cur-

rently in use. The invariant mass distributions of all gamma-accompanied feed-down events is plotted in Fig. 3.7. The total FD invariant mass spectrum (black) is compared to decay channels with at least one final state gamma (in color). Altogether, they account for $\sim 95\%$ of the total FD. Furthermore, over 83% of all FD events only

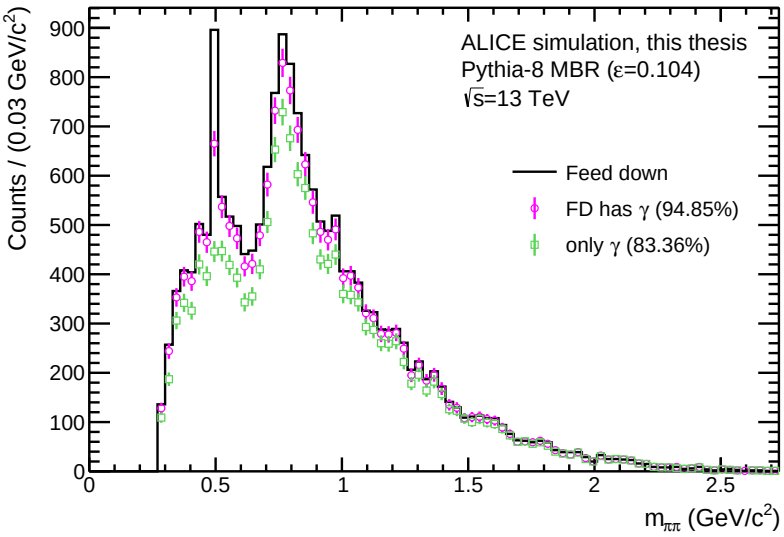
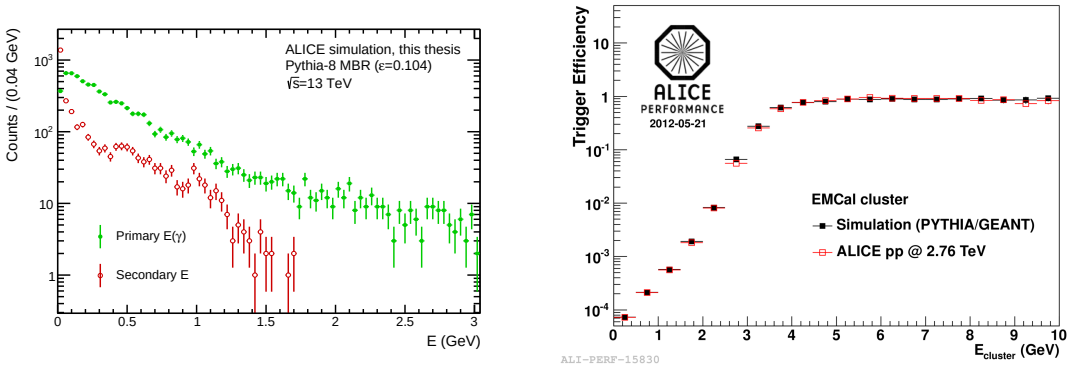


FIGURE 3.7: *Invariant mass distribution of feed-down event accompanied by gammas. The total feed-down mass spectrum (black) contains contributions which have at least an additional γ (pink) account for $\sim 95\%$. Over 83% of all FD events are only accompanied by gammas (green). The prominent peaks in the dipion spectrum originate from decays such as $\rho^0 \to \pi^+\pi^-$ and $K^0_S \to \pi^+\pi^-$ where additional final state particles go undetected (see Sec. 3.3 for more details).*

consist of two pions with N additional final state photons, *i. e.* $\pi^+\pi^- + N\gamma$. Consequently, a large portion of the feed-down could be reduced by vetoing events with gamma signals in the detector (similar to the double gap veto detectors V0, FMD & AD). The detector system capable of measuring photons is the EMCal+DCal [26, 27] (see Fig. 2.5a for its integration in ALICE). The two opposing calorimeters cover the pseudorapidity region between $|\eta| < 0.7$, and 110° and 60° in azimuthal angle, respectively. To assess the feasibility of using the EMCal to detect gammas related to FD events the energy deposited in the EMCal is studied. In Fig. 3.8a the primary photon energy as well as the secondary particle energy reaching the EMCal is plotted. As the calorimeters are placed quite far away from the beam pipe the primary photons may interact with the material of the detector, *i. e.* gaseous and solid matter from various detector systems between the interaction point and the calorimeters. Thus, in general the photons lose energy on their way to the calorimeter *e. g.* by producing secondary particles. This secondary particle energy is the maximum available energy to produce a signal in the calorimeter, which provides an approximation of the energy scale of particles entering the EMCal. Both primary and secondary energies have a peak occurrence near zero and then decrease exponentially. The actual measurable energy distribution, *i. e.* the secondary particle



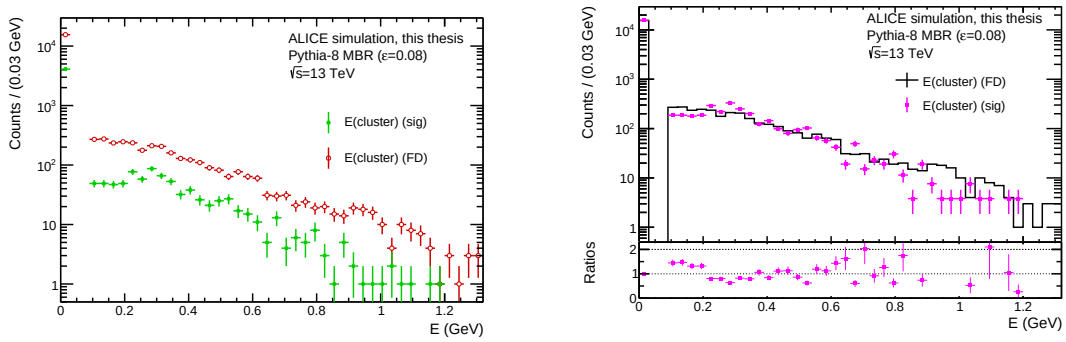
(A) (B)

FIGURE 3.8: *Energy distribution of primary gammas and secondary particles reaching the EMCal (A). Energy dependent trigger efficiency of the EMCal (B). A comparison of the maximum energy depositable in the EMCal (A) with the energy dependent efficiency of the EMCal in (B) a small percentage - roughly* $0.1\% - 0.01\%$ *of all gammas - are assumed to be detected.*

energies, goes to zero at roughly 1.5 GeV. A comparison of the energy dependent trigger efficiency of the EMCal (see Fig. 3.8b) yields an estimated percentage of $0.1\% - 0.01\%$ of detectable gammas. Therefore, the background reducing capability of the EMCal is limited by the low-energy range of feed-down gammas.

Similarly to the like-sign background estimation, actual detected photons provide an indication of the background shape. In the like-sign case, the fact that both pions have the same charge identifies the event unambiguously as background. Correspondingly, an energy deposition in the EMCal, called a cluster, also indicates a partially reconstructed event. Thus, the mass spectrum constructed from $\pi^+\pi^-$ with at least one additional γ-hit in the calorimeter should result in an excellent feed-down approximation. The measured energy deposited in the calorimeter is plotted in Fig. 3.9. The energy distributions of signal and background events are nearly identical. This similar shape is caused by the fact that clusters most dominantly originate from the charged pions and actually not from gammas (as discussed earlier) entering the calorimeter. Consequently, a cluster in the EMCal provides no direct indication for a background event: *i. e.* a registered EMCal response does not suffice to distinguish signal from background events. The total EMCal response EMC_{tot} consists (mainly[15]) of two parts: $EMC_{tot} = EMC_{\pi} + EMC_{\gamma}$. The goal is to find a variable in which we can discriminate the pion from gamma induced clusters, thus obtaining $EMC_{\gamma} = EMC_{tot} - EMC_{\pi}$. A promising observable is the distance between a track and the spatial position of the measured energy deposition in the calorimeter. This distance is obtained via the following method: an algorithm prolongates the track measured in the ITS and TPC with the knowledge of its four-momentum and the magnetic field present in the detector to the EMCal surface.

[15]Small contaminations arise from other decay particles as well. As we assume perfect PID, other charged particles types get rejected by the TPC, which lies in front of the EMCal. Therefore, these additional contributions are limited to neutral particles which are, like the gammas, part of feed-down events which we aim to reduce. Hence, the following description is also valid in the case of contaminations.



(A) (B)

FIGURE 3.9: *Comparison of EMCal energy distributions of pion vs. gamma induced calorimeter showers (A) and a direct shape comparison (B): The clusters most dominantly originate from the charged pions entering the calorimeter. Therefore, the energy distribution of signal and background clusters are nearly identical.*

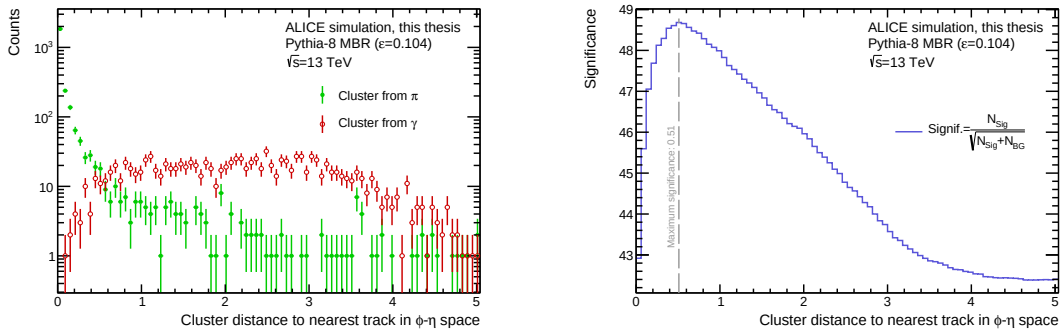
This results in a point of impact on the EMCal for every detected track. Combined with the spatial information of registered calorimeter responses a distance between clusters and tracks d_{C-T} can be calculated. Due to the curvature of the EMCal surface the displacement is calculated in the 2D rapidity azimuthal plane between the two points. This measure is also referred to as the R distance in cylindrical coordinates (measured in radians). It is given by the following relation

$$d_{C-T} = min_i(\sqrt{(\phi_{cluster} - \phi_{track,i})^2 + (\eta_{cluster} - \eta_{track,i})^2}) \tag{3.3.4}$$

The minimum function ensures that d_{C-T} is the distance to the *nearest* track (see algorithm A in the appendix). This variable seems suitable, as the distance d_{C-T} is expected to be small for pion induced clustered compared to photon induced ones. In Fig. 3.10a a direct comparison is plotted, where a clear trend can be seen. Where gamma induced clusters are almost uniformly distributed between $0.5 < d_{C-T} < 4.0$ rad, clusters created by charged pions tend to be very close to the pion track itself (which seems intuitive). Additionally, the probability that a gamma cluster is close to a pion track is small. This means a cut can be introduced increasing the chance of separating pion from gamma signals. The optimal cut value in d_{C-T} is obtained via a significance cut determination illustrated in Fig. 3.10b. The significance is defined via

$$S = \frac{N_{Sig}}{\sqrt{N_{Sig} + N_{BG}}} \tag{3.3.5}$$

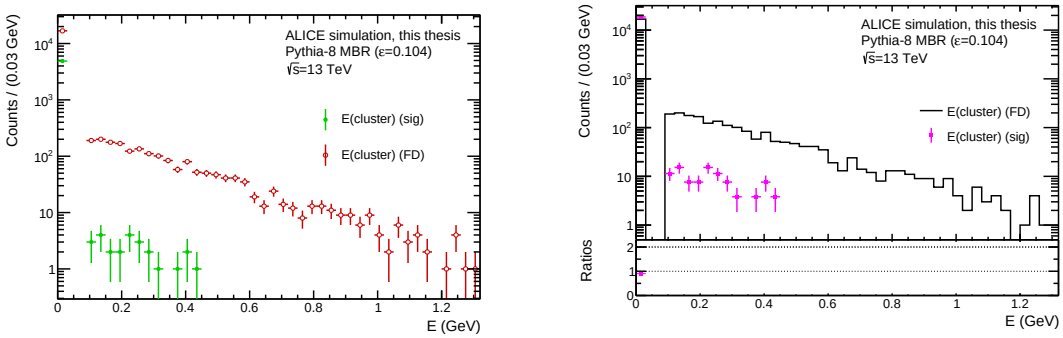
N_{Sig} represents the number of signal and N_{BG} the number of background samples on the left hand side of the cut (less than the cut value). The maximum significance presents a trade-off between an optimal signal to background ratio while a relatively large amount of signal data remains. This results in a cut value of $d^{cut}_{C-T} = 0.51$ rad. At this point the purity P and signal efficiency ε_S are $P = 98.05\%$, and $\varepsilon_S = 93.18\%$, respectively. This cut is used to distinguish clusters in the EMCal originating from gammas against clusters produced by charged pions. Thus, the EMC_γ response used to discriminate signal from background events is obtained.



(A) (B)

FIGURE 3.10: *Comparison of the minimum cluster-track distance in the $\eta - \phi$-space d_{C-T} between γ and π^+/π^- induced calorimeter showers (A). A clear difference between γ caused and π^+/π^- caused clusters can be seen. To separate them an optimal cut value, with a high signal amount and lowest possible background contamination, is searched for via the maximum significance plotted in (B). It lies at $d_{C-T} = 0.51$ rad. At this point the purity P and signal efficiency ε_S are $P = 98.05\%$, and $\varepsilon_S = 93.18\%$, respectively. This cut is used to distinguish clusters in the EMCal coming from gammas to clusters produced by charged pions.*

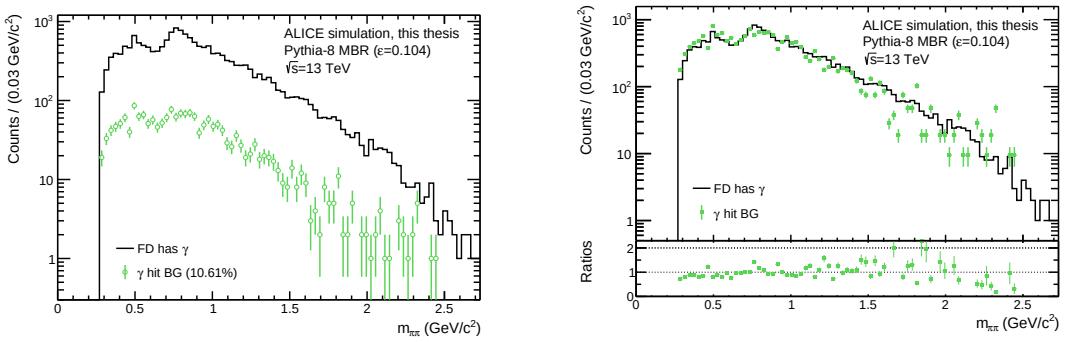
Since the EMCal does not cover the same region as the tracking detectors (TPC, ITS), tracks may lie outside of the calorimeter acceptance. In case no track can be prolongated to the EMCal surface, a measured energy deposition in the calorimeter is assumed to originate from a photon (EMC_γ). With these requirements implemented, the energy deposited in the EMCal is plotted in Fig. 3.11. Contrary to Fig. 3.9, the signal from charged pions EMC_π is nearly eliminated yielding only a small contamination of π^+/π^- clusters of $\sim 2\%$ (see purity). After obtaining a



(A) (B)

FIGURE 3.11: *Comparison of EMCal energy distributions of pion vs. gamma induced calorimeter showers after a cluster-track distance cut (A) and direct shape comparison (B): Contrary to Fig. 3.9 the signal from charged pions is nearly eliminated yielding mostly clusters produced by gammas in background events.*

rather clean sample of γ-hits in the EMCal, an estimation of the background mass spectrum can be constructed. If a measured cluster hit has $d_{C-T} > 0.51$ rad or if neither of the pion tracks can be propagated to the EMCal surface, the invariant mass of the detected, opposite-sign pions is calculated. This is referred to (in this work) as the γ-hit background. In Fig. 3.12b the γ-hit BG template is compared to the feed-down proportion of events with at least one final state gamma yielding a reasonable agreement. In Fig. 3.12b a comparison of the total background with



(A) (B)

FIGURE 3.12: *Comparison of the γ-hit background approximation (green) with feed-down events with at least one final state gamma (black). Left: A comparison of the total background with gammas and the γ-hit approximation reveals the limitations of the method. In contrast to the like-sign assumption the γ-hit model faces tedious efficiency-corrections in order to truly estimate the total contribution. Right: A direct comparison between the constructed template and the true background yields a reasonable agreement between the model and the actual shape.*

gammas and the γ-hit approximation reveals the limitations of the method. In contrast to the like-sign assumption the γ-hit model faces tedious efficiency-corrections in order to truly estimate the total contribution. Despite yielding good shape agreements, this approach is somewhat limited by the relatively low amount of obtainable statistics due to the low EMCal-efficiency at the expected γ-energies.

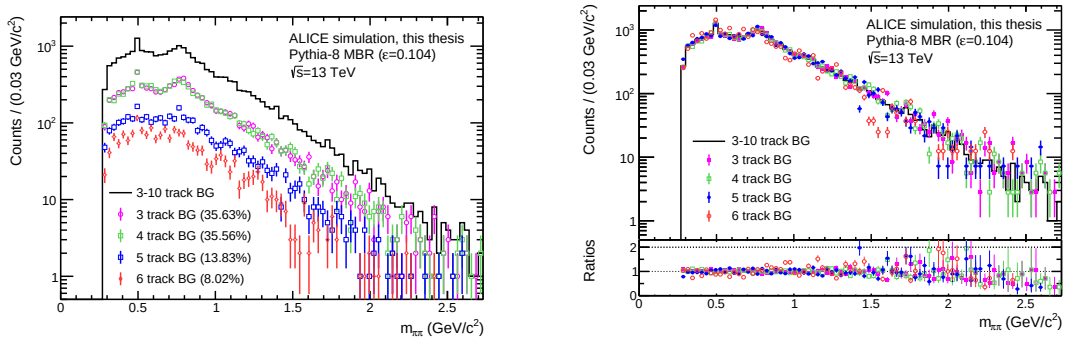
3.3.2 The 3+ background estimation

The second largest feed-down contribution comes from decay channels which feature more than two charged tracks ($n_{charged} > 2$) in the final state, hence 3+ detectable charged tracks. By analyzing events in which more than two tracks can be found in the TPC, rendering it a certain BG event, the 3+ background shape can be estimated. More specifically, the BG-template is constructed by making combinations of two opposite-sign pions with all detected tracks. *E. g.* in the following case of three detected pions $\pi^+\pi^-\pi^+$ two pairs – $(\pi^+\pi^-)$ π^+ and π^+ $(\pi^-\pi^+)$ – yield two invariant masses[16]. This procedure provides three advantages. First, although only

[16]The generated data set has to contain at least a $\pi^+\pi^-$ pair. Beyond that, any kind of particle detected is allowed. *E. g.* in the case of two detected pions and an additional kaon only one pair, *i. e.* one invariant mass can constructed.

12% of all feed-down events are in the category of 3+ events, charged particles, unlike photons, are detected very efficiently. Second, as seen in the example above, one event can account for more than one combined invariant mass, resulting in an increased statistic. Third, by combining opposite-sign instead of like-sign pairs the structure present in the FD is expected to be preserved.

Despite these benefits, the impact of the number of tracks N on the background template shape has to be studied. As N grows the likelihood of such events to constitute to the 2π invariant mass spectrum (of the feed-down) shrinks drastically. In general, the ratio of high N events constituting to the 2π invariant mass spectrum (of the feed-down) is expected to be exponentially suppressed, as events with a high number of charged tracks N are more likely to be identified as background. Therefore, in order to study the different background shapes as a function of varying N, a maximum of 10 detected tracks $N = 3, 4, \ldots, 10$ is chosen. In Fig. 3.13 the invariant mass distribution of opposite-sign combinations of events with different N are plotted. The panel on the left features the $3-10$ track background created

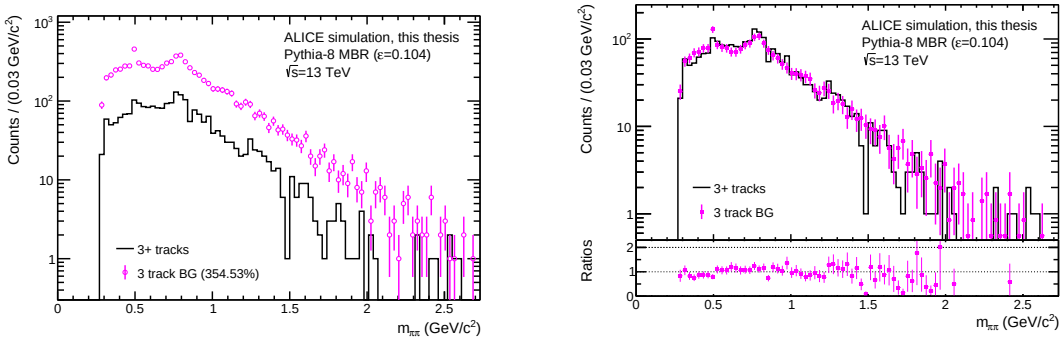


(A) (B)

FIGURE 3.13: *Comparison of the* 3+ *background approximation for different numbers detected tracks: Left: The* $3-10$ *track background created by stacking* N*-track combinatorial* $\pi^+\pi^-$ *invariant mass spectra for* $N = \{3, 4, \ldots, 10\}$ *and the individual* N *track contributions. Right: A direct shape comparison between the* $3-10$ *background approximation (black) and the individual* N *track background templates (in color) for various* N*.*

by combining the individual N-track combinatorial $\pi^+\pi^-$ invariant mass spectra for $N = \{3, 4, \ldots, 10\}$ and the individual N track contributions. A large portion of the $3-10$ BG consists of contributions from 3 and 4 track events. This dominance of low N track contributions arises due to the strict event and track filter (Tab. 3.1) which is applied to the data. This pre-filter exponentially suppresses events with a high number of tracks. The right panel features a direct shape comparison of the $3-10$ background approximation (black) with the individual N track background templates (in color) for various N yields a constant agreement across all N. As the choice of N does not affect the resulting shape, $N = 3$ is used to approximate the 3+ background shape.

In Fig. 3.14 a comparison of the 3+ background approximation with the 3 track estimation is plotted. The right panel shows a direct shape comparison between the 3 track background approximation (pink) and the 3+ background (black) yielding



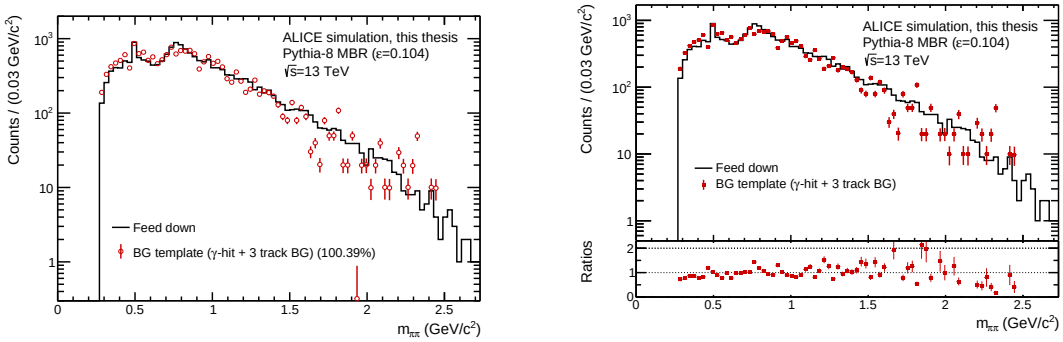
(A) (B)

FIGURE 3.14: *Comparison of the* 3+ *background approximation with the* 3 *track estimation. Left: The* 3 *track estimation (pink) exceeds the* 3+ *background (black) by a factor of* ~ 3.5. *Right: A direct shape comparison between the* 3 *track background approximation (pink) and the* 3+ *background (black) yields suitable results.*

suitable results. The left panel illustrates the 3 track estimation (pink) exceeding the 3+ background (black) by a factor of ~ 3.5. Therefore, the relative scaling needed to obtain a true estimate of the 3+ background has to be considered. The scaling is composed of two parts. First, the relationship between the number of possible unique opposite-sign pair combinations N_{pair} with respect to the number of detected charged tracks N is considered. In the case of $N = 3$ detected charged pions, the possible combinations is 2. This number decreases as other charged particles are detected as well (*e. g.* $K^{\pm}$). The exact factor can be obtained by dividing the number of 3 track events processed by the total number pairs created. In this study a value of $N_{pair}/N_{evts} = 1.85$ is obtained. Second, a compensation term which includes contributions from likelihood/efficiency considerations (similar to the γ-hit background) has to be taken into account. This includes *e. g.* the ratio between the likelihood of detecting two tracks in a multiple charged tracks event and the likelihood of detecting three tracks in the same event. And, *e. g.* the efficiency difference between an event with three detected tracks and an event with two detected tracks which pass the applied prefilter cut in Tab. 3.1 . These considerations, however, exceed the scope of this thesis and have to be the subject of further studies.

3.3.3 Results

As mentioned in the previous sections (Sec. 3.3.1 and Sec. 3.3.2), careful efficiency-corrections have to be carried out in order to estimate the feed-down components correctly. Here, we use the available MC information to rescale the histograms in Fig. 3.14a and Fig. 3.12a accordingly. A final approximation of the feed-down is made by using a combination of the 3-track and γ-hit background template, amounting to 12% and 88% of the final approximation, respectively (according to the relative ratios, see Fig. 3.5). The relative numbers ($12-88\%$: 3-track $-$ γ-hit) have to be the subject of further discussions as *e. g.* the γ-hit approximation also includes events with 3+ tracks, *i. e.* many decay channels with more than two charged tracks also frequently have final state gammas. The result is plotted in Fig. 3.15. Despite



(A) (B)

FIGURE 3.15: *Left: Feed down (black) approximation with a combination of the γ-hit and 3 track template (red). A direct shape comparison between the feed-down (black) and the combined approximation (red).*

the need for some further adjustments, the feed-down shape approximation obtained by a combination of the γ-hit and 3 track template provides reasonable results on MC data. This result seems to be a promising application to real data.

Chapter 4

Multivariate feed-down rejection

In this chapter background suppression techniques are discussed, which provide a complementary approach to background estimation studies (see Chap. 3). Instead of subtracting a background model from the data the goal is to find characteristics in the data which help to identify and thereby reduce background contamination. Conventional methods for BG rejection usually apply sequential rectangular cuts to various (individual) observables followed by a statistical analysis on the selected sample. An example for the use of of sequential rectangular cuts is the prefilter which is applied to the raw data summarized in Tab. 3.1. To obtain a maximally pure data sample with high signal efficiency a multivariate analysis approach is attempted. In contrast to single-variate methods a multivariate analysis (MVA) treats the data in its full high-dimensional feature[17] space in order to make predictions of its signal or background nature. This signal/background prediction process is commonly referred to as a *classification task*. In the following sections a motivation as well as an introduction to MVA techniques are discussed.

4.1 Motivation for using MVA for BG rejection

To suppress the background component in the data one typically tries to find observables where a scalar cut value c_{S-B} can be introduced to obtain a signal and background sample (see Sec. 3.3.1) A cut can be regarded as a simple *if*-statement which is chosen so that on one "side" (*e. g.* $> c_{S-B}$) the data behaves more signal-like and on the other side more background-like ($< c_{S-B}$). Traditionally, these individual decisions are made using the distribution of a *single* observable motivated by physical considerations. However, this scheme does not easily scale to higher dimensions as correlations between the variables come into play. Therefore, sequential one-dimensional cuts lack the potential to fully utilize the complex and high dimensional feature dependencies [35].

MVA by definition employs multiple variables simultaneously. Classification tasks can be regarded as mapping d input variables $\boldsymbol{x} = \{x_1, x_2, \ldots, x_d\}$ onto the real numbers such as $\mathbb{R}^d \rightarrow \mathbb{R}$ via a function $y = y(\boldsymbol{x})$. The input variables $\boldsymbol{x}$ range from kinematic variables, *e. g.* a particle's energy and momentum, to global event variables like the total number of clusters produced in various detector systems. The purpose of the function y is to combine the input information $\boldsymbol{x}$ in such

[17] The words *feature*, *variable* and *observable* are used interchangeably.

a way that the discrimination of signal and background is possible. As the output space is one-dimensional[18] signal and background events have to be separated again via a one-dimensional cut. Therefore, output values with $y(\boldsymbol{x}) > c_{S-B}$ are regarded as signal, while events with $y(\boldsymbol{x}) \leq c_{S-B}$ are considered background (or the other way around). Since the function y maps d-dimensional inputs onto the real numbers a constant value in the output space $y = const$ corresponds to a potentially highly intricate hypersurface in the input space [43]. This hypersurface with $y(\boldsymbol{x}) = c_{S-B}$ is what is called the *decision boundary* in the input space between signal and background events. Machine learning refers to the automated task of distinguishing signal from background. This entails finding the optimal mapping function $y(\boldsymbol{x})$ and hence the best possible decision boundary to separate the two classes. Since individual cuts (to some degree) ignore the possible high dimensional correlations between input features, the use of MVA methods such as machine learning is motivated by a performance increase in terms of higher efficiency for the same misclassification rate [43].

4.1.1 General aspects of MVA

In the following section general aspects of MVA will be discussed. The content of this section is a summary of concepts presented in the following sources [35, 43, 44], where additional, more in-depth information is provided.

The *Neyman-Pearson* lemma states that a classification algorithm which makes decisions on the likelihood ratio

$$y(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathrm{S})}{p(\boldsymbol{x}|\mathrm{B})} \tag{4.1.1}$$

provides the highest signal efficiency for a given background efficiency [43]. The exact probability functions $p(\boldsymbol{x}|\theta)$ for signal ($\theta = \mathrm{S}$) and background ($\theta = \mathrm{B}$) tend to be unknown, *i. e.* $p(\boldsymbol{x}|\theta)$ is not explicit formulated mathematically as an equation which can be evaluated. For low dimensional data[19] histograms or kernel-based density estimates can be used to assess the unknown source probability density function from simulated samples [35]. In order for these methods to provide reliable results the sample space has to be represented to some reasonable extend. If a one dimensional data set needs N samples to describe the underlying PDF, d-dimensional data require in the order of $O(N^d)$ samples. Therefore, PDF estimation techniques need massive amounts of data in high dimensions and, thus, fail simply due to limiting computational resources[20], regardless of the speed of the sample generator [35].

Alternatively, a variable y can be constructed depending on the d-dimensional input $\boldsymbol{x} = \{x_1, x_2, \ldots, x_d\}$ like $y = y(\boldsymbol{x})$ which is used as a multivariate classifier. For n samples a feature matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is constructed from the individual samples. The goal is to find the best mapping $Y : \mathbb{R}^{n \times d} \to \mathbb{R}^n \mid \boldsymbol{X} \mapsto \boldsymbol{y} = Y(\boldsymbol{X})$ between the collective inputs $\boldsymbol{X}$ to their corresponding desired target labels $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$. The quality of the function Y at point $\boldsymbol{x}$ is measured by a *loss function* written as $L(\boldsymbol{y}, Y(\boldsymbol{x}))$. It can be interpreted as the distance between the true class labels

[18] Which is the case for a binary, *i. e.* signal & background, classification task. For a n-label classification problem with $n > 2$ the output space is in general n-dimensional.

[19] *I. e.* in the order of $d < 5$ dimensions.

[20] This problem is referred to as the *curse of dimensionality*.

$\boldsymbol{y}$ and the predicted labels $Y(\boldsymbol{x})$ [44]. A leaning algorithm is tasked to minimize the loss L, and, by doing so, finding the optimal mapping function Y. Ideally, the algorithm finds the best function for all possible sets of ($\boldsymbol{x}$, $\boldsymbol{y}$). However, due to the curse of (the large) dimensionality and the infinite number of possible functions to choose from this becomes an impossible task. In supervised learning, instead a set of labeled data $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ for $i = (1, 2, \ldots, n)$ is sampled from the set of all possible values of ($\boldsymbol{x}$, $\boldsymbol{y}$) via the probability $p(\boldsymbol{x}, \boldsymbol{y})$ [35]. In order to find the function Y one chooses an algorithm, *e. g.* neural networks or decision tree based methods such as random forests. These algorithms all distinctly restrict the function space to families of highly adjustable functions $Y_\phi(\boldsymbol{X})$ with finite sets of tunable parameters ϕ [35]. This constitutes the hierarchy of multivariate analysis, which is regarded as an umbrella term for all analysis methods exceeding one dimension. Machine learning (ML) is considered a sub-part of MVA, which is in turn a comprehensive term for all algorithms capable of autonomously "learning" specific traits about a data set. Choosing an algorithm, also called a model, restricts the space of functions with which these data set traits can be found.

The target labels $\boldsymbol{y}$ are chosen to be 0 for background and 1 signal[21]. In order to adapt the parameters ϕ the loss function is minimized reducing the distance between the predicted and the true labels. This process is called *training the model* which aims to be effective across a range of inputs, not only on known data (seen during training) but also on unseen one. This goal is referred to as generalization and is sometimes rather tricky to achieve. An important goal in attaining generalization is to find a balance between *overfitting* and *underfitting*. Typically, overfitting is more prominent in more complex, flexible models and underfitting is common in very simple learners. As the model aims to extract as much information from the training data as possible, data specific artifacts such as random noise contributes to the model. The knowledge obtained from the random noise leads to a performance loss if tested on data not used during training [44]. This describes the problem of overfitting. Underfitting happens when the model complexity is too low, making it impossible for the model to learn important characteristics present in the data. Instead of overfitting vs. underfitting this problem is also often referred to as the *bias-variance* trade-off [43]. Due to the high potential complexity of commonly used models such as neural nets, the problem of overfitting is the prominent one. Therefore, to maximize the generalization power of a classifier, overtraining has to be reduced to a minimum. In a first step to do so the available data is split into three groups: the training set, a validation set, and a test set. While the model is fitted to the training set, overfitting is monitored on the validation sample. which is done via a performance comparison. In the case of overfitting the performance of the training sample drifts away from the validation set performance. Once a model yields satisfying and unbiased results on training *and* validation samples, its final generalization power is determined on a third sample: the test sample. As the training happens over more than one epoch[22] information can leak from the validation sample into the training sample. In this case the test sample provides a

[21] Consequently, MC data is necessary in the case of fully-supervised learning as it provides information on the signal or background nature of an event.

[22] An epoch is a single step in training a neural network; a neural network can be trained on every training sample more than one time. Each time all training samples have passed, the performance is evaluated on the validation sample. We say that one epoch is finished.

truly unbiased classifier performance report.

Since both training and testing require a statistically well-balanced data set, both samples want to maximize the number of data points in them. This poses a problem due the limited amount of available data. The general goal is to split the data in such a way that during training a representative and diverse set is available while keeping a sufficient amount of data to adequately test the model on a balanced representation of the underlying PDF. In this thesis a train-validation-test split of $60\% - 20\% - 20\%$ has been chosen.

4.1.2 Assessing classifier performance

A crucial task when training a model is its performance evaluation. Simply put, one checks how many times the classifier makes correct and incorrect predictions by comparing them to the true class labels from the MC truth. Usually, it is more interesting to find a specific class (*i. e.* commonly signal) while reducing contamination from the other one (*i. e.* background). In this case, a confusion matrix is useful. It contains information on the number of true positives (TP) *i. e.* correctly predicted signal events, true negatives (TN), *i. e.* correctly predicted background, and misclassified signal and background events: false positives (FP, or type I error) and false negatives (FN, or type II error), respectively. The confusion matrix is the basis of multiple performance measures (also called *metrics*). *E. g.* one popular metric is called accuracy which is defined as the ratio of all correctly classified events over the number of total events: (TP + TN)/(TP + FP + FN + TN). However, accuracy runs into problems in the case of sample imbalance where one class is overly present shadowing the performance of the small class. Other performance metrics include *e. g.* precision, recall, and the $f1$-score. All these metrics have their own advantages and disadvantages. Therefore, in order to best reflect the prediction performance of the classifier either multiple metrics, or a powerful metric like the ROC[23] curve should be reported. Due to the widespread use of the ROC curve in HEP analysis [44] and its special properties (reported in the next paragraph) it will be employed to assess model performances in this thesis.

The ROC curve plots the false positive rate (FPR = FP/(TN + FP)) against the true positive rate (TPR = TP/(TP + FN)) corresponding to background acceptance versus signal efficiency, respectively. It is constructed via the MVA output $Y = Y(\boldsymbol{X})$ of the classifier. As one slides across the range of outputs Y, the FPR and TPR are computed for each cut value along the MVA output: from the lowest (most background-like region) to the highest (signal-like) region, *i. e.* from $0 \to 1$ (in this thesis). The MVA output of a classifier is plotted in Fig. 4.1 (in gray). In green the contribution from positive-class (signal) events and in red negative-class events (background) is shown. In the case of a perfect classifier the signal and background distributions no longer overlap making them totally separable. The nature of the ROC curve restricts itself between zero and one on both axes. The best possible model has a working point in the top left corner at (0,1) with no false negatives and 100% signal efficiency. A truly random classifier would lie on the 45° line, regardless of sample imbalance. Consequently, the area under the ROC curve (ROC AUC) can be used as a *scalar* metric to report the model performance. That means the ROC curve can be condensed into a single classifier performance measure which it useful

[23] ROC stands for *receiver operating characteristic*.

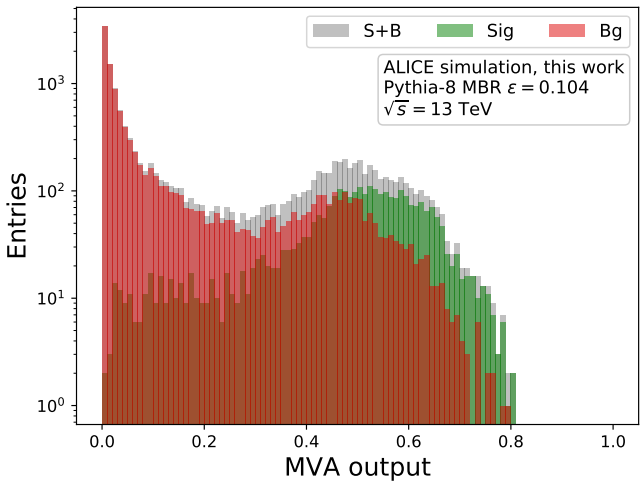
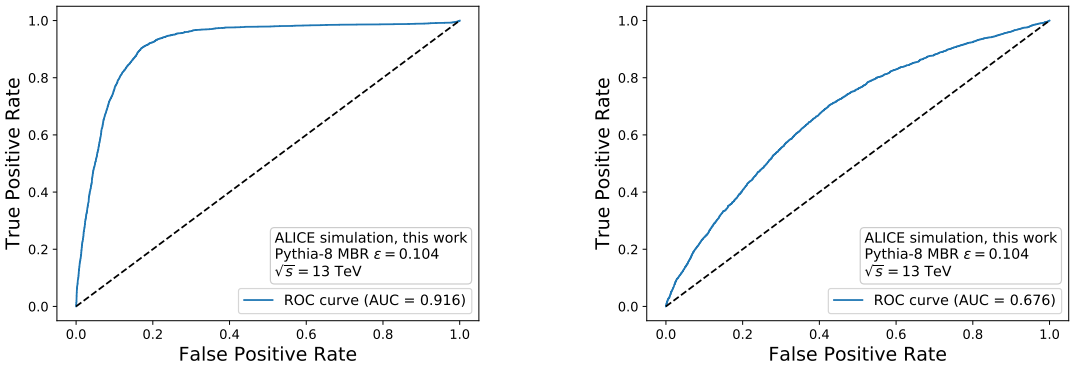


FIGURE 4.1: *MVA output of a classifier shown in gray together with contributions from positive- (green) and negative-class (red) events.*

to directly compare the quality of trained classifiers. In Fig. 4.2 two examples of differently performing classifiers and their associated ROC curves are shown. The classifier in the left panel shows a higher ROC AUC than the one on the right indicating that the left model can separate signal from background more clearly. We conclude that the model in the left panel outperforms the one in the right panel. The statistical interpretation of the ROC AUC can be formulated as follows. It is



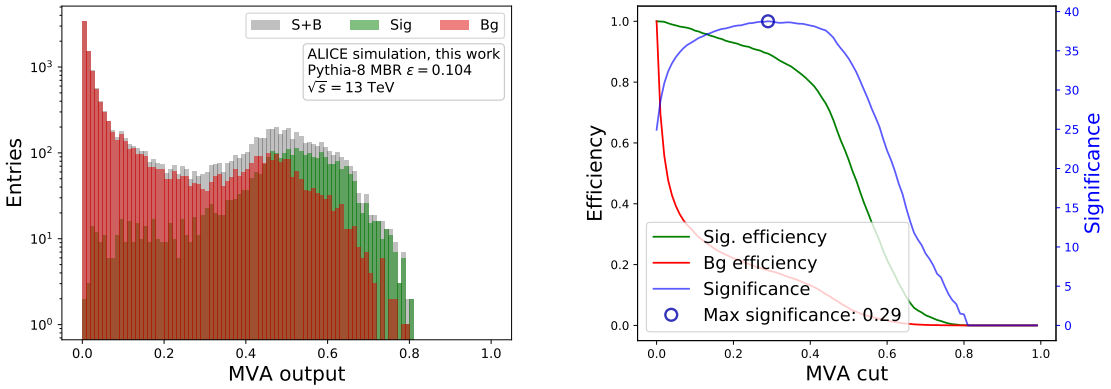
(A) (B)

FIGURE 4.2: *Examples of different performing classifiers, which are presented via their associated ROC curves: The classifier in the left panel shows a higher ROC AUC than the one on the right, indicating that the left model can separate signal from background more clearly. We conclude that the model in the left panel outperforms the one in the right panel.*

the expectation value that a randomly drawn signal sample is ranked higher than a randomly drawn background sample [44]. Despite ROC AUC being a powerful metric its application is limited to comparing classifier performances. Flach et *al.* [45] conclude that ROC AUC is a coherent metric when including non-optimal operat-

ing points, *i. e.* the optimal cut value on the MVA output Y. Choosing an optimal working point is a separate problem altogether which depends on the individual classification problem. It dependents heavily on finding a suitable balance between type I and type II errors [43]. Since the aim of this thesis is to find a classifier with a high degree of generalization power, the working point is a secondary task and defined after a model is selected. Therefore, we can use the powerful metric ROC AUC without restrains in the search for a proper classifier. Afterwards the optimal working point is determined by maximizing the signal significance (as a function of MVA output) which is often approximated by $\mathrm{TP}/\sqrt{\mathrm{TP}+\mathrm{FN}}$ for a given cut on the MVA output. It describes the ratio of the signal strength over the uncertainty of the total number of events $\sqrt{N}$ assumed to be signal (*i. e.* all events on the left hand side of an MVA cut) with $N = \mathrm{TP} + \mathrm{FN}$ where a Poisson statistics is assumed [43].

Fig. 4.3 illustrates the MVA cut optimization via significance calculations (in the right panel). As signal efficiency (green) and background efficiency (red) decrease at different rates, the significance (blue) peaks somewhere in between.



(A) (B)

FIGURE 4.3: *Significance calculations to find the optimal MVA cut value. Left: MVA output, see Fig. 4.1 Right: Optimization of the MVA cut value via significance calculations. As signal efficiency (green) and background deficiencies (red) decrease at different rates, the significance (blue) peaks somewhere in between. The circle indicates the maximum significance and the MVA cut value.*

4.1.3 Neural networks and deep learning

Machine learning has become a popular tool in high energy physics using different algorithms such as boosted decision trees and neural networks suited for the large amount and intricacy of HEP data. With the rise of deep learning the immense data surge can be handled more adeptly as higher-dimensional, more complex problems became more feasible [35]. Following an introduction into deep neural networks and deep learning, this work focuses on the application of such classifier structures to a binary classification task in order to discriminate signal from background events. In this section the theory of neural networks and deep learning is explored, summarizing the core concepts outlined in the following sources [35, 43]. A holistic summary of (nearly) all important work published in the area of neural networks and deep

learning can be found in [46].

In neural networks the structure of the network pre-defines the space of functions y_ϕ[24]. The structure is comprised of a series of transformations mapping the input $\boldsymbol{x}$, of dimensionality d, onto a so-called *hidden* state $\boldsymbol{h}^{(i)}$, with dimensionality m. This hidden state $\boldsymbol{h}^{(i)}$ is often called the *embedding* and it is the i'th transformation of the input $\boldsymbol{x}$. The subscript i denotes the i'th hidden layer. In principle there can be an arbitrary number M of hidden states as long as the final transformation maps these hidden states onto the function output y. The transformation $\boldsymbol{h}^{(i)} \to \boldsymbol{h}^{(i+1)}$ from the i'th layer to the $i+1$'layer can be expressed by the following relationship

$$\boldsymbol{h}^{(i+1)} = \Phi^{(i)}(W^{(i)}\boldsymbol{h}^{(i)} + \boldsymbol{b}^{(i)}) = A(\boldsymbol{h}^{(i)}) \tag{4.1.2}$$

An M layer neural network can therefore be written as

$$y_\phi(\boldsymbol{x}) = A_M(A_{M-1}(\dots A_1(\boldsymbol{x}))) \tag{4.1.3}$$

The function $\Phi^{(i)}$ in Eq. 4.1.2 is called the *activation function* which can differ from layer to layer. Popular choices include functions of error-like behavior, *e. g.* the sigmoid $\Phi(t) = 1/(1+e^{-t})$ or the hyperbolic tangens. With the rise of deep learning more activation functions are used including the ReLU[25] [48] $\Phi(t) = max(0,t)$ and the softmax [49] function $\Phi(t)_j = e^{t_j}/\sum_{k=1}^{K} e^{t_k}$ for $j \in \{1,\dots,K\}$ where K are the number of classes with $K = 2$ for binary classification. $W^{(i)}$ denotes the so called *weight matrix* of the i'th layer with $W^{(i)} \in \mathbb{R}^{m\times n}$. It transforms the i'th embedding $\boldsymbol{h}_i \in \mathbb{R}^n$ to m-dimensional space. The m-dimensional vector $\boldsymbol{b}^{(i)}$ is called the bias term causing a translation. The first embedding is the input vector $\boldsymbol{h}_0 \equiv \boldsymbol{x}$ and the final embedding is the one-dimensional output variable y.

In the simplest case the input $\boldsymbol{x} \in \mathbb{R}^d$ (d features) gets multiplied by a weight matrix $W^{(1)} \in \mathbb{R}^{1\times d}$ resulting in a weighted sum. If Φ_1 represents the identity function $\Phi(t) = t$ then the network can be mathematically expressed as

$$y(\boldsymbol{x}) = b + \sum_{l=1}^{d} W_{1,l}x_l \tag{4.1.4}$$

This is the simplest form of a *single layer perceptron* (with an identity activation function). It describes a linear classifier, as Eq. 4.1.4 contains no non-linear operations. *I. e.* hyperplanes with $y(\boldsymbol{x}) = const$ correspond to linear decision boundaries in the d-dimensional feature space of $\boldsymbol{x}$. Typically, linear models oversimplify the problem at hand, *i. e.* HEP data where multidimensional correlation between the input variables exist.

A non-linear model is attained by adding a hidden layer $\boldsymbol{h}^{(1)} \in \mathbb{R}^m$ accompanied by a transition matrix $W^{(1)} \in \mathbb{R}^{m\times d}$ and a bias term $\boldsymbol{b}^{(1)}$. Together they apply an affine transform to the input vector $\boldsymbol{x}$ transforming it into m-dimensional space. The non-linearity is introduced by a non-linear activation function $\Phi^{(1)}$ (*e. g.* a sigmoid) which transforms the space by point-wise application of the function. These operations can be formulated by

$$y(\boldsymbol{x}) = \Phi^{(2)}\left(\boldsymbol{b}^{(2)} + \sum_{l=1}^{L}\left[W^2_{1,l}\cdot\Phi^{(1)}\left(b^{(1)}_l + \sum_{k=1}^{m} W^{(1)}_{kl}\cdot x_k\right)\right]\right) \tag{4.1.5}$$

[24]To demonstrate the workings of a neural net we focus on a single data point in order to increase readability by decreasing the number of necessary indices. *I. e.* $Y_\phi(\boldsymbol{X}) \to y_\phi(\boldsymbol{x}) = y_\phi$

[25]ReLU stands for rectified linear unit. It constitutes the identify function truncated at 0. Its adoption in deep learning solved (to some degree) the vanishing gradient problem; see [47]

The network described by Eq. 4.1.5 can be regarded as having an input layer, a fully connected[26] hidden layer and a fully connected output layer. This general structure of fully connected layers is referred to as a *feed forward* network (FFN) or *multi-layer perceptron* (MLP). A schematic example is shown in Fig. 4.4.

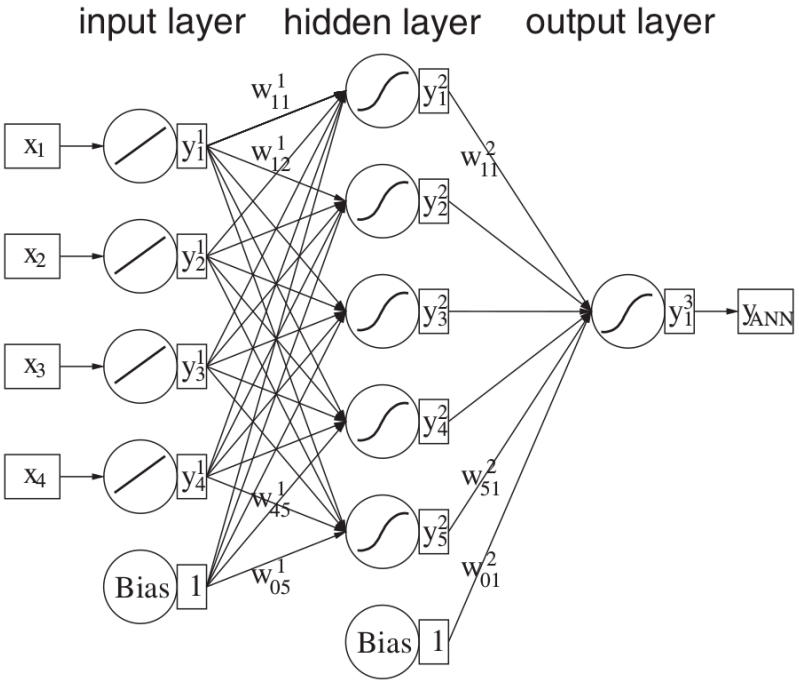


FIGURE 4.4: *Schematic drawing [43] of a feed forward neural network described by the Eq. 4.1.5. This network takes a $m = 4$ dimensional input and transforms it into a $L = 5$ dimensional embedding before returning a one-dimensional output y.*

The composition of the network, *i. e.* the dimensionality of each $W^{(i)}$ as well as the choice of the activation functions $\Phi^{(i)}$ is referred to as the network architecture, which is defined by the analyst. For a set architecture the goal of classifier training is to find the right values for the affine transforms ($W^{(i)}$ & $\boldsymbol{b}^{(i)}$) which optimize signal and background discrimination. For this task a loss function with respect to the model parameters $L(f_\Phi(\boldsymbol{x}, y))$ is defined, quantifying the performance of the classifier on the training data. Due to the nature of the fully connected layer structure the different weights are highly correlated. Consequently, as the network grows in size, the loss function becomes more complicated resulting in the existence of many saddle points and local minima. This poses a problem for standard minimization techniques which are not adjusted to such a complex task. Usually, the optimum weights, resulting in the best classification (for an MLP), are found by a method called *backpropagation*. Backpropagation is an efficient method to compute the gradient of the loss function - $\nabla_\Phi L$. Training is then done in the following way:

Before the training starts the architecture is fixed and the weights of the network are initialized, usually with small and random values. Then the data is fed into the network and propagated to the output following the randomized affine transformations. This result is compared with the expected output (the target labels) via the

[26]Fully connected refers to the fact that each hidden node is connected to all inputs x_i with $i \in \{1, \ldots, d\}$ (or the previous layer nodes) via the weights w_i.

loss function. Then the gradient of the loss function is calculated as a function of the individual weights (*i. e.* of each individual layer) using the chain rule of differentiation. To minimize the loss, backpropagation is used to adjust the weights towards smaller losses like $W \to W - \eta \cdot \nabla_W L(W)$. The parameter η is called the learning rate and defines the step size towards the direction of smaller losses (indicated by $\nabla_W L(W)$). The leaning rate η is part of the training parameters which have to be tuned as well[27] to ensure optimal generalization power of a classifier. Together with the network architecture they constitute the *hyperparameters* of the network.

Initially, multilayer perceptrons were distinguished from deep neural networks by the number of hidden layers, with deep nets having more than one layer and MLPs being shallow networks with only one hidden layer. As Hornik *et al.* have shown in [50] using a shallow network poses no restriction as any function can be approximated by introducing a single hidden layer. However, an effective one-hidden-layer-network may require a large number of nodes in the hidden layer resulting in a highly non-linear decision boundary. These intricate, high-dimensional models often fail to find some underlying truth about the data. Deep networks (*i. e.* more hidden layers) in contrast, are faced with the problem of the so-called vanishing gradient [51, 52]. During backpropagation the difference between $y_\Phi(\boldsymbol{x})$ and the desired output y_{true} is propagated from the output backwards through the various embeddings. For each embedding the gradient of the loss function with respect to its weight matrix $\nabla_{W^{(i)}} L(W)$ is calculated. As the gradient is calculated via the chain rule of differentiation for every layer a derivation close to zero (or zero) in one layer will force the whole product towards zero. Therefore, the gradient rapidly approaches zero as the network grows in depth and the parameter adjustments eventually die out at some point in the network. The problem of vanishing gradients makes it very difficult to improve the performance of deep architectures. With the recent introduction of multiple strategies including new activation functions (*e. g.* ReLU), larger training samples, and regularization techniques like dropout [53] this problem has been largely mastered. One advantage of using deep networks is their ability to learn abstract, high-level features from low-level input data, provided that enough flexibility is given. This opens the potential to obviate the need for manual and often time-consuming feature engineering[28] (see [54, 55] for more information). Moreover, the layers of deep architectures can be interpreted as constituting a hierarchical representation of the data. These properties have lead to a wide success in the application of deep learning: especially the methods of computer vision and natural language processing have become almost entirely dependent on deep learning. These fields use specific architectures tailored to their needs like convolutional nets [56] (popular in image recognition) and recurrent/recursive nets [57].

Especially recurrent networks are relevant to this analysis as they posses the ability to process variable length sequences of a common dimensionality. *I. e.* an input of fixed dimension d which can occur several times N (with N not predictable), *e. g.* N particle tracks in an event. This issue arises *e. g.* if a variable amount of detected pions (*i. e.* 2,4,6) should be considered. In simple feed forward networks variable-length input can in principle be processes by cropping or zero-padding the

[27] The tuning of the training parameters such as the learning rate η happens outside of the network training.

[28] Feature engineering is the process of creating observables which make a signal and background discrimination more easy. This often requires specific domain knowledge.

input to a fixed length/dimension d. However, these solutions either neglect useful information or introduce a placeholder value in the network with no physical meaning. An optimal solution would entail a self adjusting network which dynamically adopts to the required input size. Illustrated in the following expression is an example of such a network mapping n individual inputs $\{\boldsymbol{h}^{(i)}\} i \in \{1, \dots n\}$ onto a single output $\boldsymbol{h}$

$$\boldsymbol{h} = \Phi(W^{(1)}\boldsymbol{h}^{(1)} + W^{(2)}\boldsymbol{h}^{(2)} + \cdots + W^{(n)}\boldsymbol{h}^{(n)}) \tag{4.1.6}$$

This architecture is called recursive or recurrent as input can be fed recursively into the network which is then condensed into a single (arbitrary) length representation $\boldsymbol{h}$. Due to the possibly endless depth of recursive networks such rudimentary recursive units in Eq. 4.1.6 can encounter vanishing and exploding gradient problems. By selectively applying the activation function and transformations $W^{(i)}$ these problems can be mitigated. This procedure is called *gating* and results in a more complex recurrent unit, with for example long-short-term-memory (LSTM [58]) and gated recurrent units (GRU [59]). These units use shared weights that can be considered as creating a sort of "memory" of recent states.

Usually, deep neural networks effectively have tunable hyperparameters in the order of 10^5 up to 10^7 depending on their width and depth. Consequently, it is (nearly) impossible to explain the prediction just from inspecting the final weights and biases. Algorithms like neural networks and deep learning are commonly referred to as black box models. However, recent advances have been made to partially entangle the workings of black box algorithms and research towards interpretable machine learning is actively conducted (see *e. g.* [60, 61]). An additional, algorithm independent problem is described as a *covariate shift* between training data and *real*[29] data. As MC simulations provide only an approximation to real data the distribution describing the training samples varies (at least slightly) from the actual data. This leads to a performance decrease when the classifier is used to make predictions on real data. Since covariate shift is a common problem in machine learning, domain specific approaches such as *re-weighting* [62] or domain-adversarial training of neural nets [63] exist. For an overview of deep learning and its applications outside of physics see [64, 65].

In the next section machine learning methods using neural networks with deep architectures are deployed to discriminate signal and background events in order to reduce feed-down contamination in the data.

4.1.4 Used frameworks & data sets

The data preprocessing and filtering has been done as described in Sec. 3.1. In addition, a gamma-filter has been added discarding events with calorimeter clusters exceeding a distance of $d_{C-T} > 0.51$ rad (see Sec. 3.3.1 for details) which suggests the detection of a photon.

All machine learning tasks are performed with Python 3.5.2 and the Keras 2.1.3 [66] framework using the Tensorflow 1.4.1 [67] backend. Data handling is performed using the Python packages NumPy 1.14.0 and Pandas 0.22.0. The conversion between ROOT trees and the Python data structures is done with the package uproot 2.6.14 [68]. Plotting in Python is done via the matplotlib 2.1.2 library.

[29] *I. e.* data collected in the experiment unlike the simulated training data which is an approximation of these events.

The data items considered for this study include event-, and track-level features. While event-level information describes overall event behavior such as total energy depositions in various sub-detector systems, track-level information contains details about single track characteristics. Baldi *et al.* [69] conclude that classifiers using raw *low-level* information from detectors combined with *high-level* features (*i. e.* constructed from low-level information) outperform models trained solely on either only low-, or high-level data. Technically, all features obtained within the ALICE framework are to some degree composite features obtained from simple digital signals from the detector. Here, high-level features refers to observables which are constructed from features obtained via the ALICE software framework.

4.1.5 Data preparation

In order to construct sensible features the background nature of high mass feed-down events is considered. Feed down is characterized by its missing undetected energy and momentum which is expected to yield a dissimilar event topology compared to fully reconstructed signal events. To describe the event topology two variables are generated: First, the distance in the $\phi - \eta$ space as $d_{\phi-\eta} = \sqrt{(\phi_1 - \phi_2)^2 + (\eta_1 - \eta_2)^2}$ is calculated where 1 and 2 refers to the first and second particle. Second the enclosed angle φ_{1-2} between the two track three-momenta $\vec{p}_1$ and $\vec{p}_2$ is constructed (φ_{1-2} yields a similar quantity as $d_{\phi-\eta}$). Further information available in the ALICE software framework include the particles four-momentum, the length of the track, and the distance of closest approach (DCA) to the main vertex. The DCA is obtained as the minimum distance from any point of the prolongated track fit[30] to the expected decay vertex. The invariant mass itself is not used as a training variable as it relies heavily on theoretical assumptions. The network should not focus too much on the mass observable as its distribution may not represent the real data accurately. The goal is to extract information about the signal and background nature of events via their individual topology introduced by different "production" mechanisms: *i. e.* feed-down mass/energy loss vs. total event reconstruction. This underlying topology difference should, to some degree, be model-independent[31], thus hopefully producing a robust model which can be applied to general missing mass/energy cases.

4.2 Multivariate feed-down rejection

In this section the multivariate approach to reject the feed-down background component in the two pion invariant mass spectrum in $X \to \pi^+ + pi^-$ decays of the centrally produced system X is described. Specifically, a summary of the algorithms used, as well as the optimization procedures implemented is presented. Eventually, the results produced are discussed in Sec. 4.3.

The machine learning algorithm in use is a neural network. Training a neural net is done by updating the weights after a single event has been fed through

[30]The tracks are obtained via performing a Kalman filter [70]. Simply speaking a curve is fitted through clusters in the detector produced by the track.

[31]That is the model of the simulator generating the diffractive mass which depends on theoretical assumptions.

the network, this is called *online learning*. As the amount of data grows, this approach becomes less and less tractable due to the hardly parallelizable weight-update scheme. In practice, one feeds a certain number of training samples into the network and then updates the weights once. This procedure is referred to as *mini-batch learning*, in case only a subsample of the data is fed into the network at a time, or *batch-learning* in case the entire training sample is propagated trough the network before the weights are updated. In the following paragraph the specific choices of different hyperparameters are briefly explained. These choices are mostly motivated by being the state of the art method in training a neural network. Additionally, the availability of the hyperparameter setting in Keras is an important criterion as it facilitates their application.

Before the training starts the weights are initialized via the Glorot [71] normal initialization scheme. The initial weights are drawn randomly following a truncated normal distribution with zero mean and a standard deviation of $\sigma = \sqrt{2/(n+m)}$. n and m refer to the dimensionality of the weight matrix $W \in \mathbb{R}^{m \times n}$ (the layer takes n - dimensional input and produces m - dimensional embeddings). The bias nodes $\boldsymbol{b}^{(i)}$ are initialized with zeros. The activation functions are chosen to be ReLUs within network layers and a sigmoid in the final layer returning a value between zero and one. In order to prevent the network from overfitting, regularization methods are employed. They include *dropout* [72] layers and *batch normalization* [73] which are installed for every hidden layer. Dropout among them is the simplest yet quite effective technique of randomly ignoring a certain fraction of nodes in a layer, consequently preventing the network from focusing on one specific connection. Extreme dropout fractions include 0 and 1, which describe no dropout regularization and total dropout which renders the network untrainable. Somewhere in between the two extreme dropout fractions sits an optimal dropout fraction $f_{drop} \in (0, 1)$ for a given architecture. Therefore, this f_{drop} constitutes to the hyperparameters of the network. Contrary to dropout, batch normalization alters the output of a network layer. It attempts to maintain the activation close to a mean of zero and its standard deviation close to one. This has two major benefits: for one it prevents the activations jointly to fall towards zero (with near zero variance) which increases the model's performance and, moreover, introduces regularizing effects (*i. e.* by preventing internal covariate shift, for more information see [73]). Due to the relatively small training sample consisting of about 10^5 data points the mini-batch is chosen to be of size 32, which allows for reasonable training times in the order of $2-4$ minutes depending on the specific architecture. To prevent features from highly varying in magnitudes, units, and range the data is *standard scaled* before it is entered into the network. Standard scaling transforms each feature distribution centering its distribution around a zero mean with a variance of one. The model loss is determined via the binary *cross-entropy* (CE) - a quantity originating from information theory which is commonly implemented in classification tasks. In order to optimize the loss function a popular gradient-based optimization algorithm called *Adam* [74] is used.

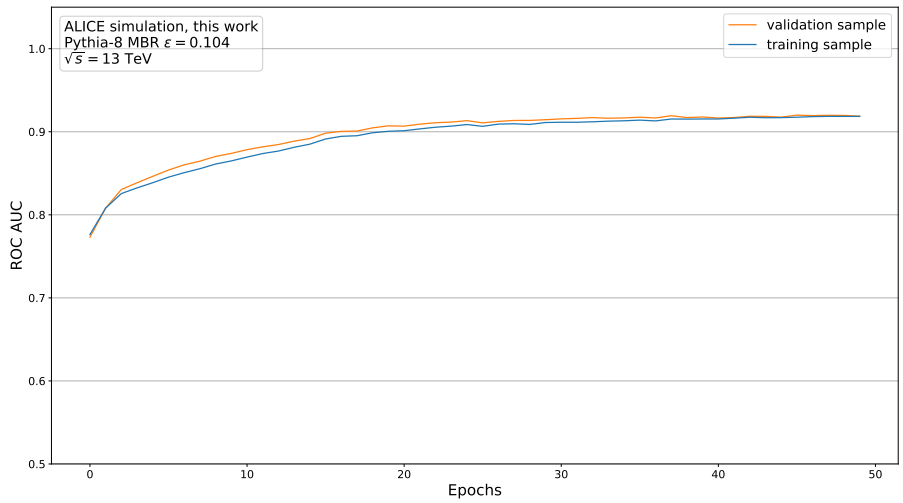
4.2.1 Evolution of the classifier architecture

The evolution, *i. e.*the search via trial and error of the optimal classifier architecture, is done manually in the following way: At first an architecture is selected. Hereby, two major components can be altered: first, the input layer type, and sec-

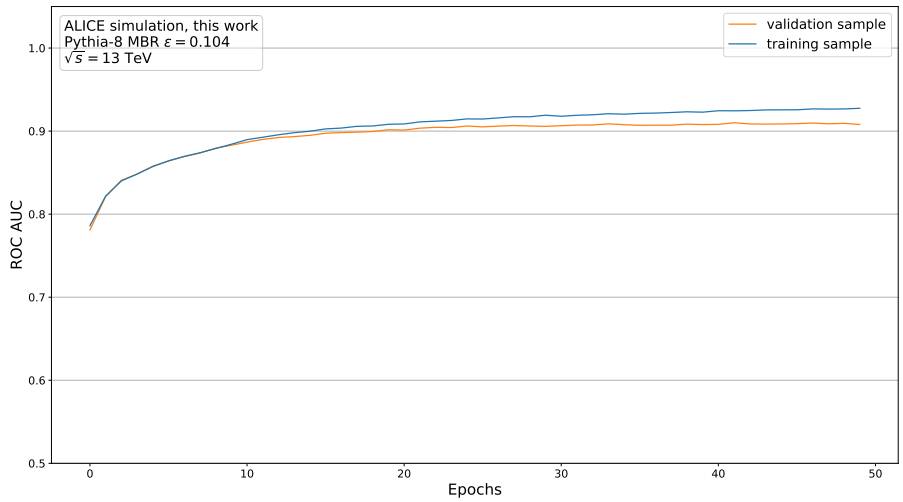
ond, the number of hidden layers as well as the dimensionality of each embedding (except the first,*i. e.* the input, and last one, *i. e.* the output). To describe the first component more precisely the data at hand has to be considered, which consists of event- and track-level information. This implies that the data cannot be simply entered into a "flat" neural network (*i. e.* with a standard d - dimensional input) as the number of tracks may not be constant, resulting in a variable input dimension. However, this thesis considers only the measured two pion spectrum. In this case a flat, fixed length input vector can be constructed by stacking the feature vectors of both recorded tracks on top of the event-level features. This results, depending on the exact features used (see Tab. C.1 in the appendix), in an approximately 50-dimensional vector with about 10 event observables and 20 for each track (the exact number of features used is discussed below). However, stacking the particles introduces an ordering, *i. e.* a top and a bottom particle, which may introduce biases. Additionally, stacking the inputs increases the total dimensionality drastically. The most obvious way of handling multiple similar inputs[32] is via recurrent units. Especially long-short-term-memory units are relevant as they have the ability to store important information in a hidden state causing it to entangle details about multiple n-dimensional inputs into a single n-dimensional output. Hence, the first choice of architecture regards the use of a *flat* network versus a *recursive* one. The output of the recursive layer is then concatenated with the event input (in the same way the flat stacked NN concatenates the three input vectors). Thereafter, a (deep) neural net is attached, whose width and depth, *i. e.* the dimensionality and the number of layers, respectively, is subject to the second variation. The performance is tested on a fixed set of features (later referred to as "BLF$_{\phi\eta}$ Bayes", see below). Finally, the performance of different feature combinations is studied. During training, overtraining is monitored by comparing the training and validation performance via their CE-losses and their respective ROC AUC's. If at least one pair of these measures starts to drift apart from one another (illustrated in Fig. 4.5b) the model is discarded and further more restrictive regularization is applied: *i. e.* the dropout rate $f_{drop} = min(f_{drop} \mathrel{+}= 0.1, 0.8)$. If a classifier is still overfitting with a dropout rate of 80% the architecture is not reported and not considered any further. After successful training (*i. e.* no overfitting) over 50 epochs[33] the performance of a model is evaluated on the unseen test sample. The number of epochs is restricted to 50 as overtraining occurs quite frequently if training happens over $\gtrsim 50$ epochs (can *e. g.* be improved by a larger sample size). The optimal performance for a certain architecture is reported in Tab. 4.1. The results can be summarized as follows: In general, the *recurrent* architecture slightly outperforms the *flat* one. Flat architectures have a considerably higher input dimensionality which innately makes them more difficult to train as they require way more fine-tuning to prevent overfitting. While architectures implementing recurrent units produce reasonable results across multiple network designs, the *flat* setup is struggling to yield stable generalizable results due to overfitting during training, apparent in Fig. 4.5b. This overfitting trend can be seen in Tab. 4.1. Models with recurrent units tend to have a lower variance around their mean value compared to classifiers trained on the flat architecture

[32]Similar in the sense that each input has the same features, *i. e.* the same dimensionality as well as identical variables. In the context of this thesis this applies to the detected pions which share observables.

[33]Describes one forward pass and one backward pass of the entire training sample.



(A)



(B)

FIGURE 4.5: *Model performance comparison between training and validation set during training to highlight overtraining: In (A) no overtraining occurs on the classifier structure* 2×50*, (*$f_{drop} = 0.7$*) implementing a* recurrent *cell where in (B) the* flat *approach leads to overfitting.*

which experience standard deviations as high as 1.21. This means these models are quite unstable in their predictions making them unreliable predictors. In general, overfitting in flat architectures requires more restrictive dropout rates of $\geq 80\%$. Additionally, deeper and wider > 70 networks are also more prone to overfitting as the number of complexity increases (applies to flat and recurrent setup). Underfitting also happens for architectures with a width < 30 where the performance between the training and validation set constantly drifts apart. Therefore, flat and very deep structures are avoided in favor of more stable and generalizing results. These architecture tests are performed on a feature set which excludes the specific use of the p_T variable. The reported performances can be improved by adding p_T dependent variables (see below). However, as stated above, p_T is strongly dependent on the choice of the underlying model (mass biased) which should generally

Architecture	ROC-AUC [%]	ROC-AUC DL-test [%]
2×30, *recurrent* ($f_{drop} = 0.7$)	90.4 ± 0.98	85.7 ± 0.62
2×40, *recurrent* ($f_{drop} = 0.7$)	91.2 ± 0.83	86.1 ± 0.68
2×50, *recurrent* ($f_{drop} = 0.7$)	91.2 ± 0.55	86.8 ± 0.33
2×60, *recurrent* ($f_{drop} = 0.7$)	**91.3**$\pm$0.47	**87.7**$\pm$0.12
2×70, *recurrent* ($f_{drop} = 0.7$)	91.0 ± 0.71	85.8 ± 0.24
3×40, *recurrent* ($f_{drop} = 0.7$)	90.2 ± 0.81	85.5 ± 0.41
3×60, *recurrent* ($f_{drop} = 0.7$)	89.9 ± 0.78	85.7 ± 0.46
2×50, *flat* ($f_{drop} = 0.8$) (70 epochs)	89.9 ± 1.21	85.8 ± 1.01
2×60, *flat* ($f_{drop} = 0.8$)	90.0 ± 1.09	85.8 ± 0.95

TABLE 4.1: *Performance comparison of various neural net architectures measured via their ROC-AUC. The reported performance is calculated as the mean of three runs (with as little overfitting as possible) with the corresponding standard deviation. The notation is as follows: $a \times b$ represents the dimensions of the network where a refers to the number of hidden layers and b to the number of nodes in each layer. The subsequent text describes how the track level data is handled;* flat *suggests a stacking of the feature vectors whereas* recurrent *means that the track information is processed in a recurrent unit, i. e. a LSTM cell.*

be avoided. Nevertheless, many track features experience a p_T dependence to some degree, which cannot be eliminated.

In order to test the full generalization power of the network the classifier is tested on another simulated sample. This sample does not follow the CEP simulation scheme of MBR [37] but is modeled according to the theory of Donnachie-Landshoff [75] (DL). Like the MBR simulation scheme, the DL data are generated based on a Pomeron approach. However, the DL parametrisation uses intrinsically different PDF shapes in order to simulate the kinematics. The results are reported also in Tab. 4.1 under ROC-AUC DL-test[34]. The mass distribution is slightly shifted towards lower masses (using the default setting in both the MBR and the DL simulation) compared to the MBR hypothesis, plotted in Fig. 4.6. The performance on this data set provides important insights as a potential disagreement in the invariant mass distribution cannot be ruled out when transitioning to real data. Despite the invariant mass difference, the remaining one-dimensional feature distributions, which were used to train the model overlap almost entirely. However, high dimensional correlations between theses variables are likely to be different for the two CEP simulation approaches. Hence, a performance drop is assumed by testing the model on the DL data, which is also what we obtain, illustrated in the performance Tab. 4.1. This drop is quite moderate which implies a high level of consensus in the

[34]DL stands for the Donnachie-Landshoff parametrisation.

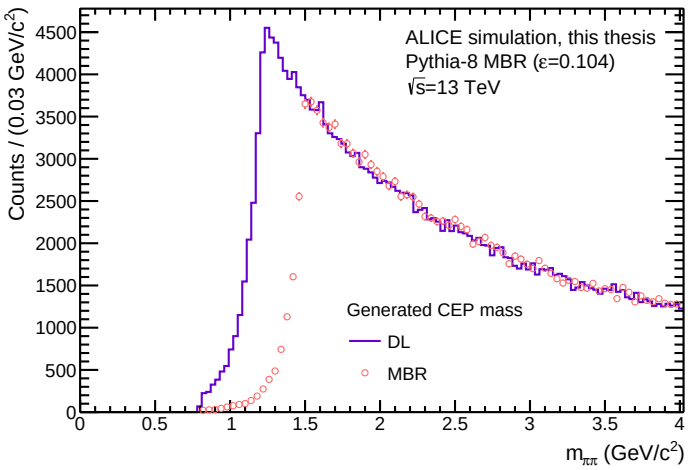


FIGURE 4.6: *Invariant mass comparison of MBR and DL simulated events. The DL parametrisation assumes lower invariant masses compared to the MBR model of the high mass continuum background (default values).*

modeling of the simulation[35] (which is also the case).

In order to study the effect of certain features on the classifier we fix the architecture to 2×60, *recurrent* ($f_{drop} = 0.7$). It should be noted that regarding overfitting, the recurrent architectures reported in Tab. 4.1 are quite stable and perform rather similarly. In order to choose a generalizing architecture the lowest standard deviation is considered which is a measure for a reliable and stable configuration. Again, the reported performance metrics are the ROC-AUC on the test set, and the DL sample (average of three runs with the corresponding standard deviation). As a baseline a set of features listed in 4.2.1 are used.

[35]The re-weighting scheme has also been applied to combat a covariate shift expected in differently simulated data. However due to the similarities the procedure does not increase the performance on the Donnachie-Landshoff data.

Baseline features (BLF) used in every training

1. Event features:
   - *Number of tracklets*
   - *Number of singles*
   - *Number of tracks (in total)*
   - *Number of residuals*
   - *Total FMD, V0 & AD multiplicity*
   - *Number of V0s*

2. Track features:
   - η
   - ϕ
   - *Number of clusters in ITS, TPC, TRD*
   - *Number of shared clusters in TPC*
   - *PID TPC signal*
   - *Golden* χ^2
   - *Track length*
   - *ITS* χ^2
   - *TPC* χ^2
   - DCA_{xy} *&* DCA_z

A more detailed description of features used in this analysis is shown in Tab. C.1 in appendix A. Event level features contain observables like the number of tracklets,[36] and the total accumulated signal in the veto detector systems FMD, V0, and AD[37]. Particle features include kinematic observables such as a tracks η and ϕ direction, as well as detector specific signals, *e. g.* the number of clusters track produced in the TPC. The results are shown in Tab. 4.2. The standard feature configuration on its own does not perform exceptionally well. Most of these features have rather similar signal and background distributions, thus making it hard to disentangle signal and background events. As additional information gets introduced the classifier performance increases quite noticeably. The first feature added is the distance of both tracks in $\phi - \eta$ which reveals information about the event topology, plotted in Fig. 4.7. It shows a distinct pattern for signal and background which the classifier is able to utilize. Generally, a trend to better performances can be observed as additional kinematic information is introduced via p_T correlated features. This performance increase is immediately noticeable by simply adding the variable p_T to the baseline features (BLF + p_T). A similar performance can be obtained by introducing $d_{\eta\phi}$ as well as six features describing the particle properties via TPC number of sigmas (nσ) and the Baysian PID probability of pions, kaons, and protons, respectively. However, the PID features themselves are not independent of the transverse momentum and carry some information either directly or indirectly about p_T. Therefore, in order to achieve a well performing model, some information of the particles momentum has to be added. This arises from the particular way the

[36]Tracklets are segments formed with hits on two layers of SPD which is part of the ITS detector system.

[37]The total signal is the sum of all signal lying below the trigger threshold.

Features	ROC-AUC [%]	ROC-AUC DL-test [%]
BLF	78.5 ± 0.12[38]	76.1 ± 0.05
BLF + $d_{\phi-\eta}$ (BLF$_{\eta\phi}$)	85.4 ± 0.68	83.0 ± 0.18
BLF$_{\eta\phi}$ + Bayes PID probabilities	91.3 ± 0.47	87.7 ± 0.12
BLF$_{\eta\phi}$ + Bayes + TPC nσ	94.9 ± 0.35	89.8 ± 0.12
BLF + p_T	94.3 ± 0.26	89.2 ± 0.46
BLF$_{\eta\phi}$ + p_T + φ_{1-2}	95.1 ± 0.19	91.0 ± 0.05
All features	95.3 ± 0.35	90.7 ± 0.09

TABLE 4.2: *Performance comparison of networks trained with various feature compositions using the* 2×60, recurrent *architecture with adjustable dropout rate (to avoid overfitting). The ROC-AUC score reported is the mean of three trials with the corresponding standard deviation (reported to two significant figures as it would sometimes appear to vanish).*

feed-down background is produced. The missing mass and energy from undetected particles results in a shift to lower momenta (see Fig. 4.8). In the following section on the results the reduction power of neural networks is discussed depending on the features used listed in Tab. 4.2. Specifically, the effects on the invariant mass distribution are discussed.

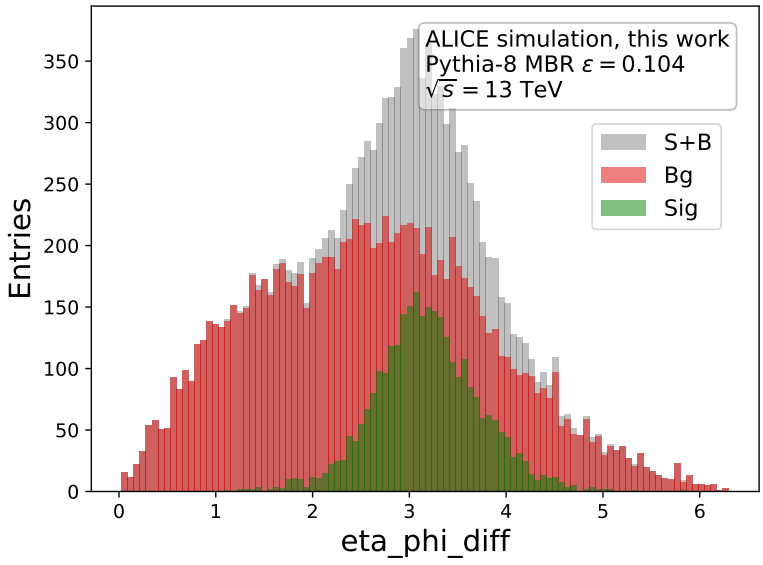


FIGURE 4.7: *Signal and background distributions of the distance in* $\phi - \eta$ *space of the two measured tracks.*

[38]Slight overtraining was unavoidable.

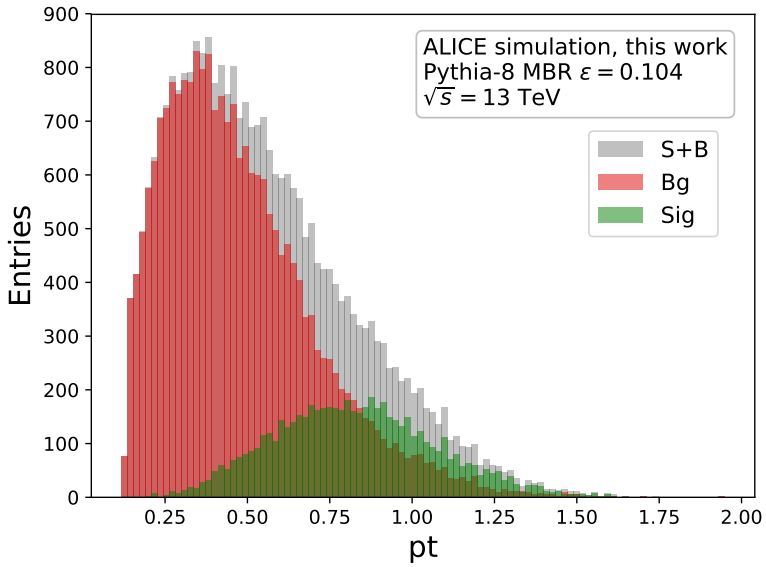


FIGURE 4.8: *Track transverse momentum distribution of signal and background events. Due to the nature of feed-down events energy and momentum is lost leading to a shifted momentum distribution to lower values.*

4.3 Results & Discussion

In this section, the results of the classifiers reported in Sec. 4.2.1 are presented. Before discussing classifier performances, the invariant mass spectrum of signal and background before applying an MVA cut plotted in Fig. 4.9 is again examined. Here

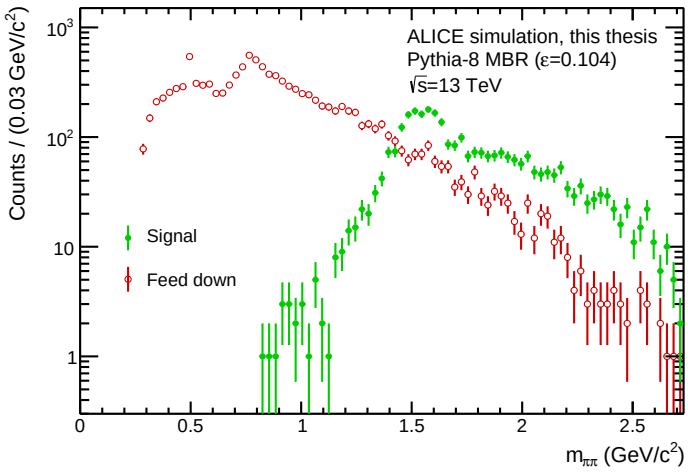


FIGURE 4.9: *Invariant mass spectrum of signal and feed-down data used to train the network.*

the effect of feed-down is apparent as it results in a shift to lower invariant masses of the detected particles. It is tempting to introduce a mass cut at roughly $\lesssim 1$ GeV/c^2 eliminating a large portion of the feed-down. However, as mentioned in Sec. 3.1, the simulated data sample includes only high mass continuum events. These states con-

stitute particles with higher masses of the CEP event spectrum. Not contained in the data sample are final states from resonantly produced X particles which occupy the mass spectrum around and below the 1 GeV/c^2 range. Therefore, a classifier performance is not only measured via its ROC-AUC but most importantly by the level of background reduction across the whole invariant mass spectrum while keeping the mass dependent signal efficiency as close to one as possible.

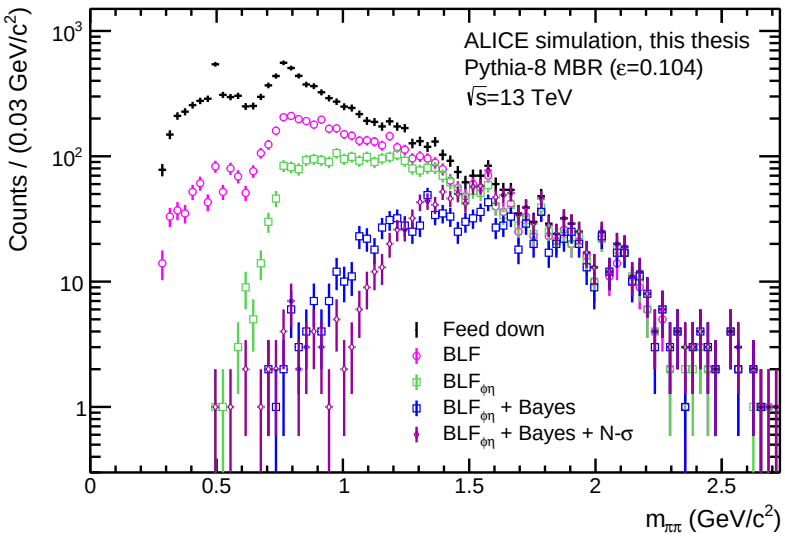
In order to make predictions with a classifier, the MVA output is used to determine the maximum signal significance (using the MC truth) for a given cut on the output. Data points to the left of the cut are regarded background events, data points on the right as signal events (see Sec. 4.1.2 for more details). The resulting invariant mass distribution and background reduction of feed-down events is plotted in Fig. 4.10 and Fig. 4.11 for various classifier predictions. The models themselves differ in the set of features used to train the classier. Models reported in Fig. 4.11 explicitly use the observable p_T during training, whereas the plots in Fig. 4.10 do not. The baseline classifier (BLF) is reported for comparison in both figures. In addition, the performance of each classifier is reported via the signal purity and signal efficiency in Tab. 4.3. The purity P is defined as the fraction of signal

Features	Signal purity [%]	Signal efficiency [%]
No classifier	21.0	100.0
BLF	33.9	88.1
$\text{BLF}_{\eta\phi}$	45.7	82.8
$\text{BLF}_{\eta\phi}$ + Bayes PID probabilities	66.1	70.6
$\text{BLF}_{\eta\phi}$ + Bayes + TPC nσ	67.8	93.3
BLF + p_T	64.8	87.4
$\text{BLF}_{\eta\phi}$ + p_T + φ_{1-2}	69.4	92.1
All features	69.9	93.1
Mass cut at 1.39 GeV/c^2 [39]	69.9	93.8

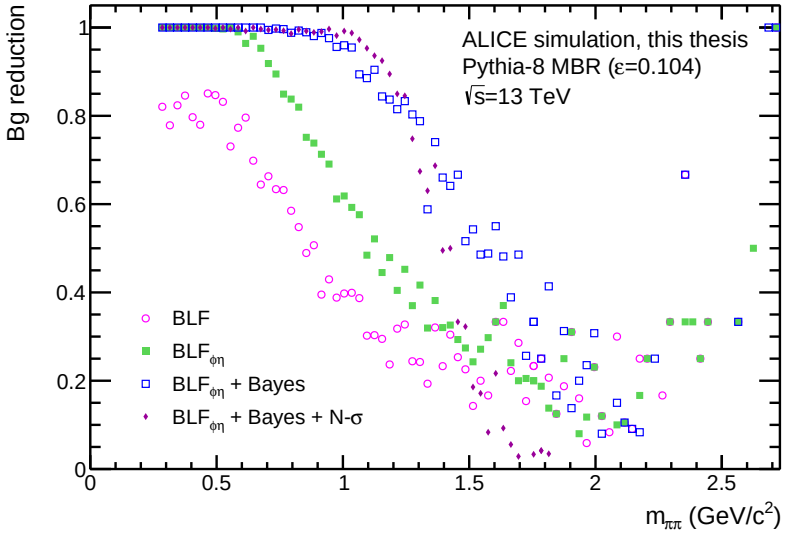
TABLE 4.3: *Performance comparison of networks trained with various feature compositions using the* 2×60, recurrent *architecture. The reported scores are signal purity P defined as the fraction of signal events in the event sample: $P = S/(S + B)$ and the signal efficiency defined as the fraction of signal events that survive the classifier cut.*

events in the event sample: $P = S/(S + B)$ and the signal efficiency defined as the fraction of signal events that survive the classifier cut. The optimal classier maximizes both, signal efficiency and purity which is achieved by completely reducing the background while leaving signal events unchanged. The true performance of a classifier is then assessed via a combination of these measures. The mass dependent background reduction reveals a clear pattern as more features get added to the

[39] Mass cut value is obtained via maximum signal significance for a cut along the invariant mass spectrum. See Fig. 4.12.

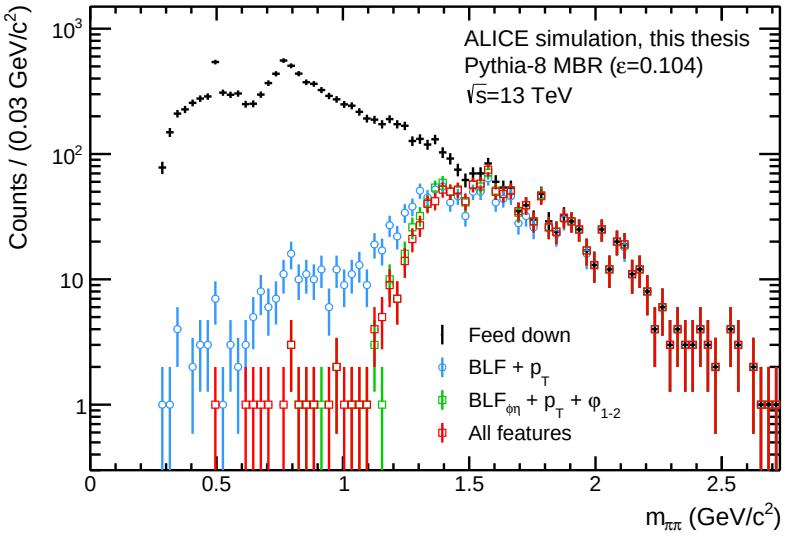


(A)

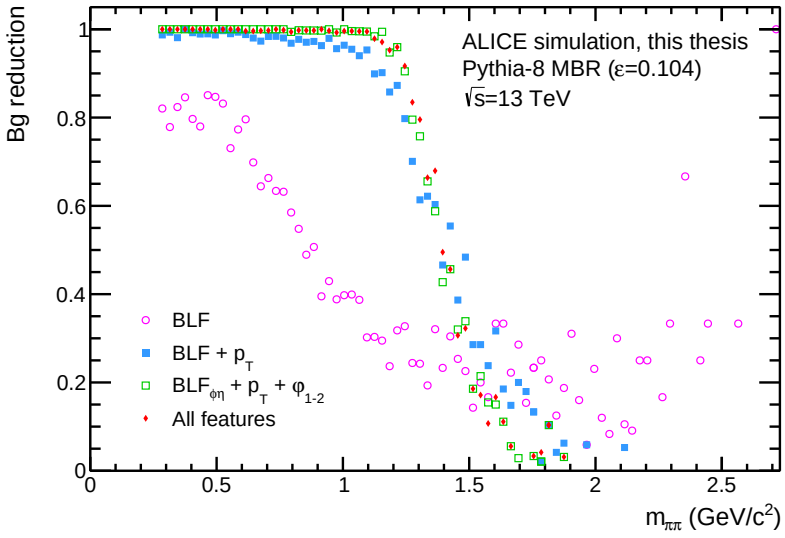


(B)

FIGURE 4.10: *Invariant mass distribution (A) and background reduction (B) in feed-down events for various MVA models: The choice of features used to train the classifier depicted do* not *explicitly contain the p_T observable.*



(A)



(B)

FIGURE 4.11: *Invariant mass distribution (A) and background reduction (B) in feed-down events for various MVA models: In contrast to Fig. 4.10 the training features for the displayed classifier do explicitly contain the p_T observable. Additionally the baseline classifier (BLF) is reported for comparison.*

baseline features (BLF): *i. e.* the classifier gradually becomes a more defined mass cut between roughly $1.2 - 1.5$ GeV/c^2. This is also evident in the signal efficiency plotted in Fig. 4.13 which increases as the background reduction approaches zero. The introduction of the variable p_T, either directly or indirectly via *e. g.* nσ the classifier acts as a mass cut. This effect is most prominent for a classifier using *all features* plotted in Fig. 4.14. The classifier makes an error function like signal and background separation in the invariant mass spectrum at around $1.3 - 1.4$ GeV/c^2. Based on this assumption a mass cut for comparison is introduced. The optimal cut value is obtained via the maximum significance calculated along the invariant mass, illustrated in Fig. 4.12. This results in an optimal mass cut of 1.39 GeV/c^2. This optimal cut value lies precisely in the transition region where background reduction transitions from $1 \rightarrow 0$ for classifiers using p_T variables. The cut also results in a very similar purity and signal efficiency compared to the *all features* classifier (see Tab. 4.3). The network seems capable of extracting invariant mass informa-

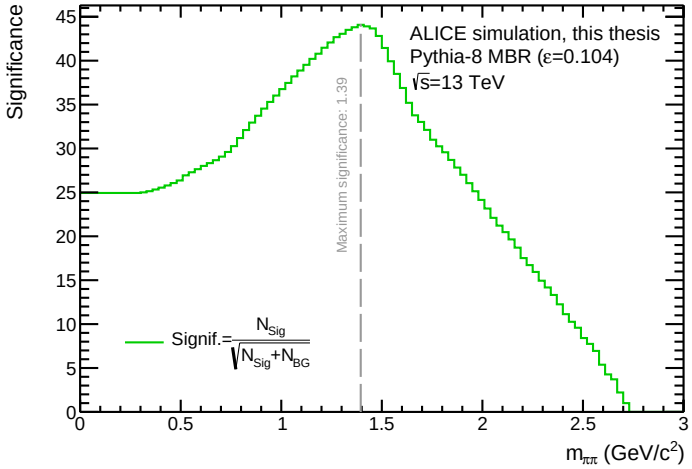
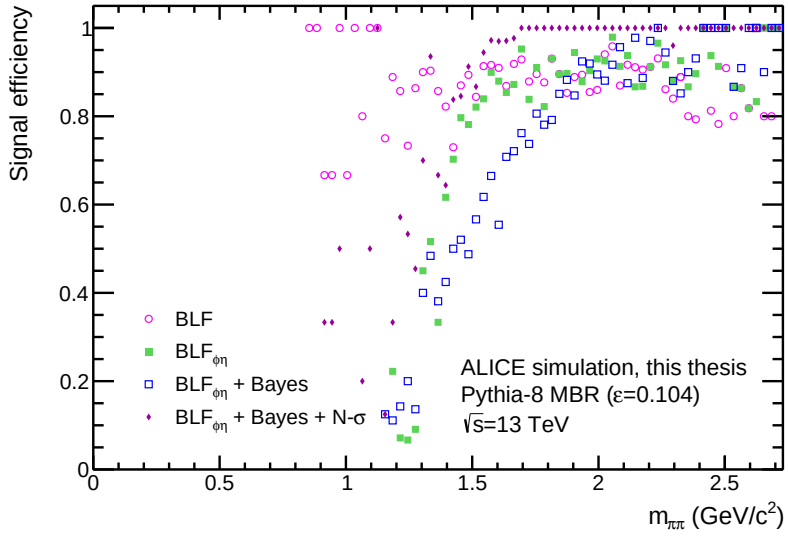


FIGURE 4.12: *Signal significance for a cut along the invariant mass variable.*

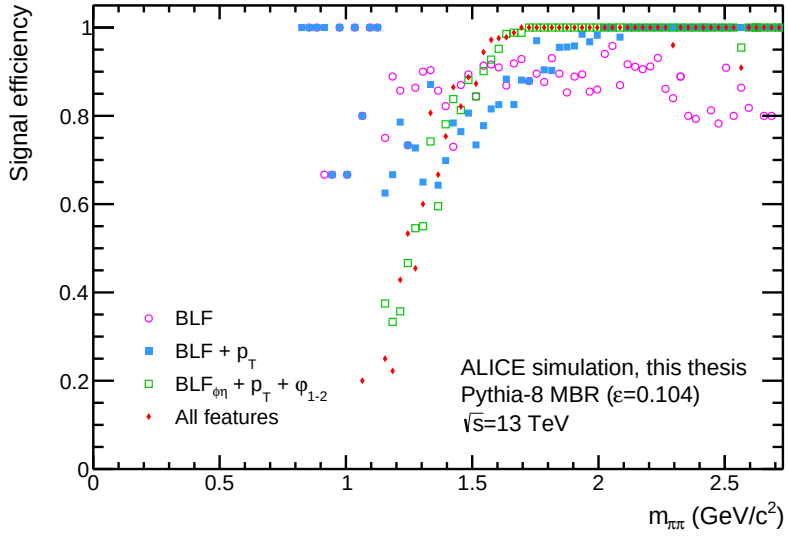
tion about the measured tracks and subsequently using it in the decision process. Furthermore, some traces of the combined invariant mass appears to be present in several features, as classifiers with no direct momentum information behave very similarly to those with access to the full kinematics (*i. e.* via p_T). This becomes evident if we compare the mass dependent signal efficiency and background reduction of the classifiers (BLF + p_T) and (BLF$_{\phi\eta}$ + Bayes + nσ) in Fig. 4.15 which show similar behavior.

Despite relatively high background suppression rates reported in Tab. 4.2 the use-case of models trained in this thesis are constrained to high mass continuum CEP events (where MVA can be replaced by a one-dimensional cut on the mass). In order to avoid a strict mass cut on real data a multivariate analysis trained on Pythia-8 simulated data using currently available parametrisations (*e. g.*MBR/DL) can only be practical if rudimentary features (*e. g.* BLF) are used to train the model.

To increase background reduction in the regions > 1.3 GeV/c^2 a classifier may only be trained in this specific mass region. However, training in this specific mass region would require at least 10^2 more events to be simulated as the background



(A)



(B)

FIGURE 4.13: *Signal efficiency over invariant mass for different classifiers without (A) and with (B) the full kinematic variables (p_T) among the training features. Additionally, the baseline classifier (BLF) is reported for comparison.*

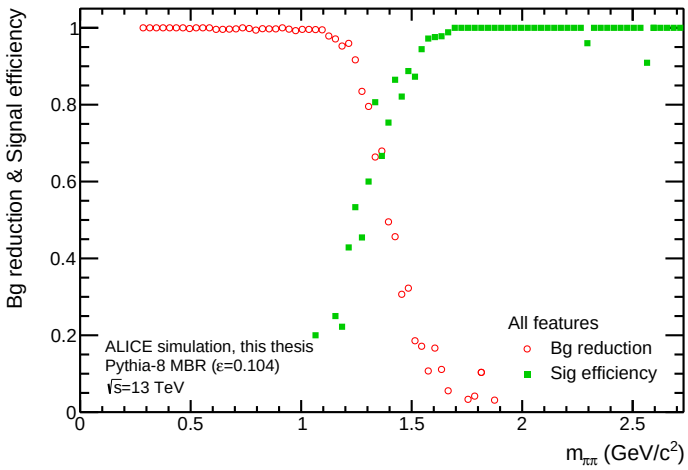
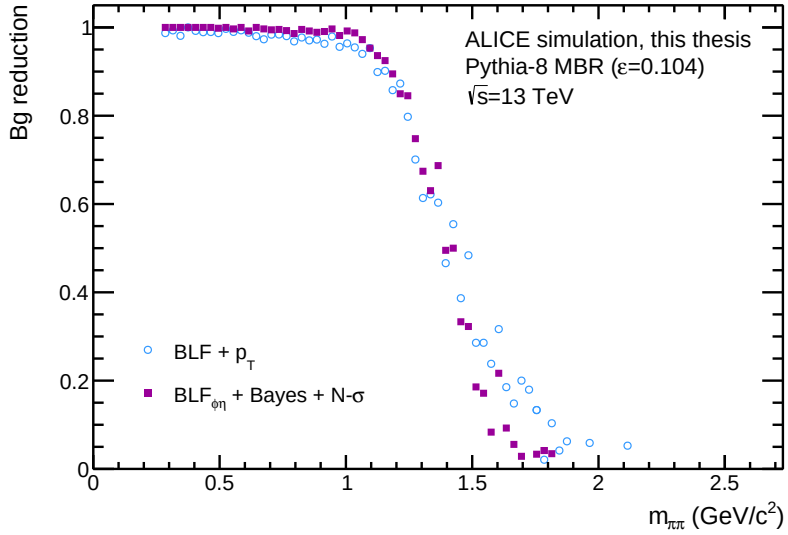


FIGURE 4.14: *Signal efficiency and background reduction as a function of invariant mass of a classifier trained on* all features*: This results in error function like behavior of both the signal efficiency and background reduction at around* $1.3 - 1.4$ *GeV/*c^2 *indicating that the neural net focuses heavily on mass dependent features such as* p_T*.*

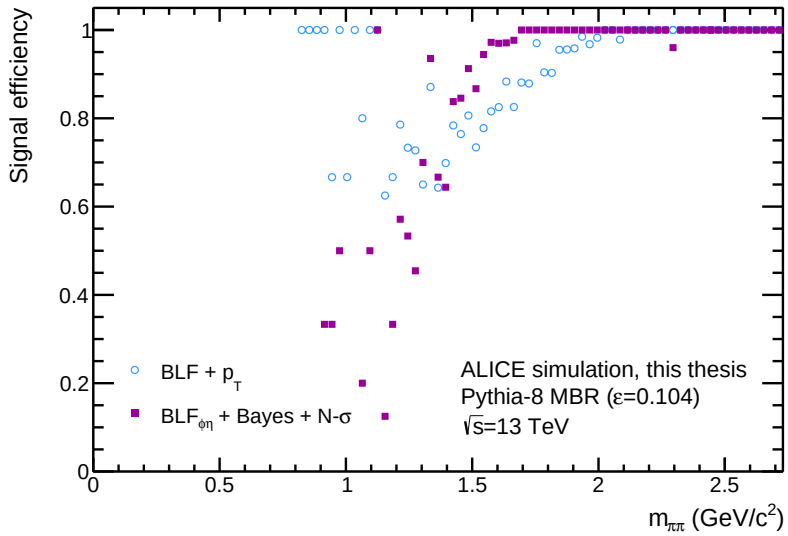
component drops roughly exponentially above $\sim$ 1 GeV/c^2. That implies an increase to roughly 10^8 total simulated CEP events.

An alternative strategy is described in a paper by Metodiev et al. [76]. The method described is called *classification without labels* (CWoLa) and has the advantage of being easily implementable in the case of real data. Instead of providing a pure signal and background sample the classifier is trained to separate statistical mixtures of classes. Metodiev *et al.* show that even without information on individual labels and class proportions the optimal classifier can still be found as in the case of fully-supervised learning (*i. e.* all label information is available). The training is performed on two data sets where one predominantly contains background samples and the other contains more signal instances. As described in Chap. 3 a fairly representative background sample can be created by using tracks with a γ-hit in the EMCal or more than two measured tracks in the TPC. Here a sample of 3 and 4 measured tracks is used. This negative-class sample is labeled *generated background*. The positive-class sample consists of signal and background events, *i. e.* events with two tracks, passing the filter in Tab. 3.1, and lacking EMCal hits. It is labeled "S-B mixture" (contains real signal and background samples) and is considered the "signal" class when training the classifier.

The classifier is tasked to distinguish generated background samples from a mixture of signal & background events (S-B mixture). The feature distributions of the generated background sample and the background in the mixed sample should share similar characteristics. Therefore, in order to distinguish the two data sets the classifier has to focus on traits present in signal events. It is expected that the MVA output of signal events within the "S-B mixture" sample gets shifted more towards 1 compared to background events in the "S-B mixture" as they provide a more dis-



(A)

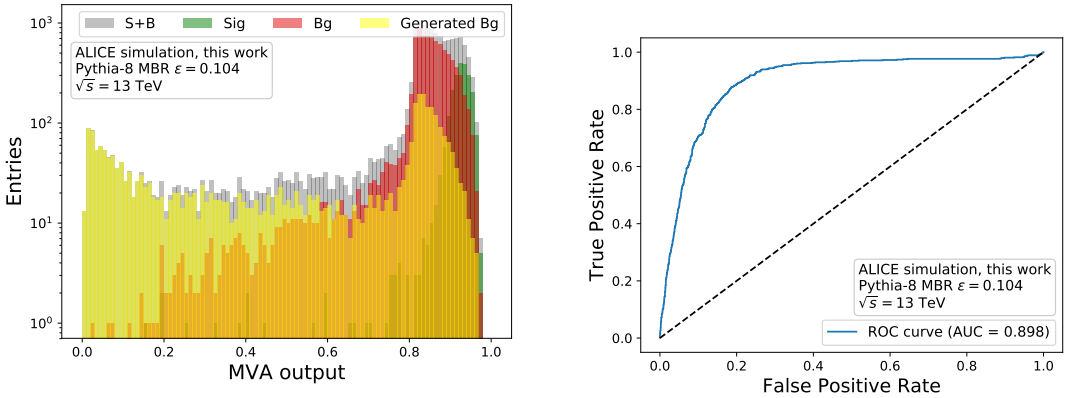


(B)

FIGURE 4.15: *Model comparison of (BLF + p_T) and (BLF$_{\phi\eta}$ + Bayes + nσ) regarding invariant mass dependent signal efficiency and background reduction. Both models behave similarly, despite having different additional (on top of the BLF) features. Therefore, the model (BLF$_{\phi\eta}$ + Bayes + nσ) is able to extract information on the transverse momentum from the additional features it gets.*

tinct difference from "generated background" events. In addition, the model should be less prone to kinematic variables as background samples contained in the "S-B mixture" set alter the distribution of said variables.

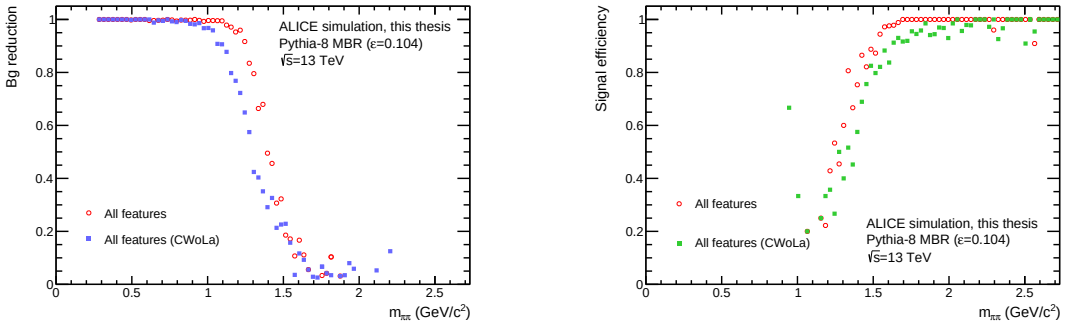
The MVA output and ROC curve tested on (real) signal (green) and background (red) events contained in the "S-B mixed" sample can be seen in Fig. 4.16. The classifier is able to distinguish signal events (green) contained in the mixed sample more clearly from the generated background events (yellow) than real background (in red) contained in the mixed set. However, as can be seen in Fig. 4.17, this training scenario results in a very similar classifier (as predicted by [76]), which again obtains a mass bias. p_T dependent variables provide the classifier with an immediate implication to originate from the "S-B mixed" sample. To eliminate p_T dependencies a CWoLa network is trained using only the set of baseline features (BLF). A comparison of the mass dependent background reduction and signal efficiency between a CWoLa network and a standard network trained with BLF is plotted in Fig. 4.18. Both purity and signal efficiency are reported in Tab. 4.4



(A) (B)

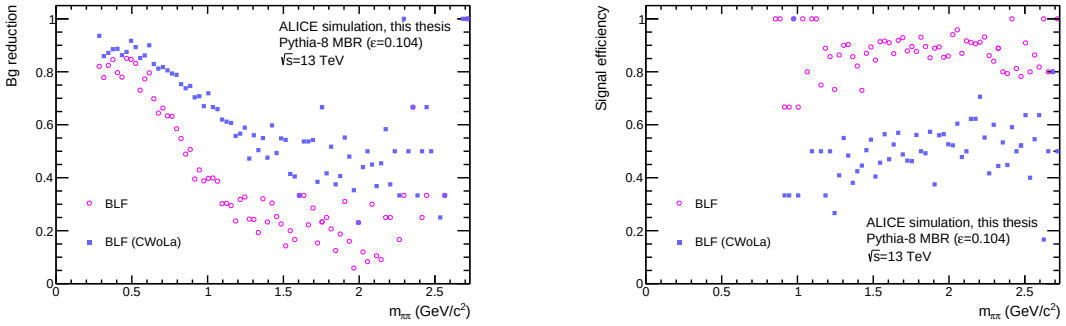
FIGURE 4.16: *MVA output and ROC curve for a classifier trained to distinguish generated background samples (yellow) from a mixture of positive-class (green) and negative-class (red) instances. The generated background sample is obtained via γ - hit and 3+ track methods (with 3 and 4 tracks in the TPC) described in chapter 3. The classifier can distinguish signal events in the mixed samples more clearly from the generated background samples than the remaining background sample. The final ROC curve is then evaluated with the ground truth information on only signal and background events (not including generated background samples).*

Despite yielding a classifier which tends to be mass independent the potential of this method is again limited by the missing low mass resonant component, if the training is performed on MC simulated data. However, as the negative-class sample is composed of (γ-hit & 3+ track) feed-down events and the positive-class sample of a mixture of signal and background events the method is transferable to real data. The network could be trained following the steps outlined above. Ideally, the true signal should emerge from the background data in a similar way as it is depicted in Fig. 4.16. Despite some difficulties regarding the optimal working point when training on real data, classification without labels provides a potential implementation



(A) (B)

FIGURE 4.17: *Background reduction (A) and signal efficiency (B) comparison of classifiers trained on* all features *using two different methods. The classifier plotted in red is trained using the standard fully-supervised learning approach where all label information is available. The classifier represented by the blue dots uses a different strategy called* classification without labels. *Here the negative-class training set consists of generated background events (γ hit, and* 3+ *background - see Chap. 3), where the positive-class sample is a mixture of signal and background samples without any other background indications. Both training scenarios result in very similar classifiers.*



(A) (B)

FIGURE 4.18: *Same scenario as in Fig 4.17, with the difference that the classifiers are trained only on the set of baseline features. Both training scenarios result in very similar classifiers. However, the resulting CWoLa model seems to be less depending on the invariant mass as other models before.*

of MVA methods in the analysis of CEP events, which makes it relevant for further investigations.

Ultimately, standard supervised machine learning applications (using simulated training data) involving central exclusive production require a more complete CEP simulation to be sought out to truly deploy the potential of multivariate analyses. A package currently in development is called *ExDiff* [77], which simulates centrally produced low mass resonances $f_0(1500)$, $f_0(1710)$, $f_2(1950)$, and $f_2(2220)$ at the de-

Features	Signal purity [%]	Signal efficiency [%]
BLF	33.9	88.1
BLF (CWoLa)	33.5	50.1
All features	69.9	93.1
All features (CWoLa)	60.8	86.2

TABLE 4.4: *Performance comparison of networks trained with various feature compositions using the CWoLa training scheme. The reported scores provide an extension to Tab. 4.3*

sired center of mass energy of $\sqrt{s} = 13$ TeV. The recently published version 2.0 is also connected to Pythia 8.2 for resonance decays and hadronization making it easy to include and interesting for further studies.

Chapter 5

Summary and outlook

The framework of the strong force is quantum chromodynamics (QCD). QCD is experimentally well established at high energies, where theoretical assumptions can be made using perturbative methods. Interactions in this energy regime are characterized in terms of basic quark and gluon exchanges. At lower energies, however, accurate descriptions become increasingly difficult to outright impossible - even when knowledge of higher energy dynamics are considered - as complex (high-order) interactions become the dominating processes. Diffraction physics at LHC energies lies in-between these two energy scales, describing strong interactions outside the QCD framework via *Regge theory*. To resolve the issue of the rising total cross section at high energies a *Pomeron* ansatz is used to describe the mediation of the strong force. Central exclusive production (CEP) events represent a particular interesting diffractive process. CEP processes are events where the two interacting protons stay intact but exchange sufficient energy and momentum to create a particle X. According to Regge theory, theses states (at LHC energies) are produced by a fusion of two Pomerons, which are emitted by the interacting protons. The production of X via double Pomeron fusion is a colorless mechanism which results in a clear experimental signature with large voids of particles between the outgoing protons and the centrally produced system in the pseudorapidity variable η (referred to as a rapidity gap).

The ALICE experiment consists of a central barrel and a forward muon spectrometer. Additional smaller detectors for global event characterization and triggering are located at small angles outside of the central barrel. Such a geometry allows the investigation of central exclusive production (CEP). A CEP trigger is constructed by requiring hits in the central barrel and no activity outside of it, creating a rapidity gap filter. Measuring the decay products of X at ALICE allows for a detailed study of the Pomeron.

In this thesis the charged dipion invariant mass spectrum of the decay $X \to \pi^+\pi^-$ has been studied for possible background sources. This was done by employing *Pythia-8* simulations of these processes, which indicated a drastic reduction of non-diffractive events (background) by enforcing the rapidity gap condition (described above). Furthermore, these studies revealed that the remaining background is largely composed of sources from high mass states which decay into $X \to \pi^+\pi^- + N$, *i. e.* two charged pions, which end up in the detector, and N additional unobserved particles. This background is referred to as feed-down (FD). To successfully extract information about the composition of the X particle from the invariant mass

spectrum, the feed-down background source has to be reduced.

This thesis focused on (1) understanding, and describing the feed-down background (see Chap. 3) and (2) studied the potential of multivariate analysis (MVA) methods to reject feed-down contributions in the invariant mass spectrum (see Chap. 4).

The goal of the feed-down description study (in Chap. 3) was to find feasible methods which are able to replicate the background shape. *I. e.* a representative template of the background shape which can be subtracted from the whole data yielding the excessive data as signal. A well-known method for approximating the combinatorial background is the so-called *like-sign* estimation. It is constructed as follows: particle pairs of identical charge cannot solely constitute the central system X (due to charge conservation, X has vacuum quantum numbers) and, consequently, their mass spectrum is expected to coincide with the continuum background in the data. However, often feed-down events (83 %) are accompanied solely by neutral particles. *I. e.* the majority of background decay channels do not feature additional charged pions and can, thus, not be modeled using the like-sign approximation. Furthermore, the invariant mass spectrum of feed-down events show emerging structures not describable in the continuum assumption. Therefore, the like-sign method was substituted by two other FD estimation methods: the γ-hit, and the 3+ background approximations.

First, roughly 95% of all feed-down events are accompanied by gammas. However, of those gammas entering the calorimeters (EMCal) at ALICE only a small fraction get measured. This is due to low detection efficiency ($\sim 10^{-3} - 10^{-4}$) of the EMCal at the expected gamma energies. If photons actually do produce a signal in the EMCal they were used as veto information discarding roughly 10% of feed-down events. Additionally, as detected gammas identify a background event, such events were used to approximate the background distribution. In contrast to the like-sign approximation, where a like-sign pair estimation is constructed, the so-called γ-hit background template is constructed with two opposite-sign pions if an additional photon was measured in the EMCal. This results in a very good approximation of the feed-down shape (as 95% of the feed-down come with at least one gamma). However, due to the low detection efficiency at the expected gamma energies this method provides a relatively small sample size.

Second, to combat the small statistics obtained in the γ-hit method a track based estimation was used. A feed-down composition study revealed that about 12% of all feed-down events have more than two tracks. The detection efficiency of a charged track entering the detector acceptance - within a certain momentum range - equals nearly one. Therefore, charged pions entering the detector acceptance are a few orders of magnitude more likely to be detected than photons in the EMCal, guaranteeing an increase in statistics. The background was constructed in the following way: A data set of events with more than two detected tracks was created (containing at least a $\pi^+\pi^-$ pair). Within one event $\pi^+\pi^-$ pairs are formed and their invariant mass is calculated. This procedure, called the 3+ track estimation, has the advantage over the like-sign method in that it retains the structures observed in the BG spectrum.

Both the γ-hit and the 3+ background estimations represent their respective feed-down contributions quite well. Despite the need for further adjustments, *i. e.* careful efficiency corrections in order to estimate the true extend of each background

component, the combined background template provided reasonable results on simulated data. Therefore, the γ-hit and 3+ background approximation seems to be a promising application to real data.

In the feed-down suppression study, multivariate classification with neural networks has been implemented. A comparison of different architectures and feature selections has been conducted. For training event- and track-level data were used. When training the classifiers overfitting was prominent for deep and wide architectures. Consequently, rather shallow and narrow architectures were used which provided consistent results. In addition, a recursive unit (LSTM) has shown improved results when dealing with multiple track inputs compared to a simple flattened network input (*i. e.* the track inputs have been stacked on top of each other), which continuously showed signs of overtraining. Despite a relatively decent purity increase of 30%, while keeping the signal efficiency at $> 93\%$, a closer look at the invariant mass dependent background reduction (& signal efficiency) shows that the classifier is seemingly adopting a simple mass cut. Subsequently, the classifier was compared to a simple one dimensional mass cut (obtained by maximizing the signal efficiency over the invariant mass) which yielded a striking similarity. The trained classifiers seem to introduce a cut at the same position which was obtained via the maximum signal significance, thus, resulting in approximately the same purity and signal efficiency as a regular 1-dimensional cut. In general, all classifiers (*i. e.* classifiers differing in the set training features) adopted a mass/p_T bias to some degree. This can be attributed to the defining characteristic of feed-down events which is missing energy/momentum. However, this mass bias is not desirable as the simulated data only describe the high mass continuum production of X. Other CEP events from resonant decays (not described in the simulations used in this thesis) feature lower mass states, whose contribution to the dipion invariant mass spectrum in $X \to \pi^+\pi^-$ decays would be truncated by the networks obtained mass bias.

In an effort to prevent classifiers from obtaining a mass/p_T bias, a method called *classification without labels* (CWoLa) was implemented. Instead of providing a pure signal and background sample, the classifier was trained to separate statistical mixtures of classes. Therefore, the model should be less prone to kinematic variables as the additional background samples in the mixture data set alter the distribution of said variables. The negative-class data set was constructed via γ-hit and 3+ background methods which resulted in a pure background sample. The positive-class data set was composed of a mixture of signal and background events, *i. e.* events with two tracks, passing the track filters without γ signals in the EMCal. As expected, the MVA output of signal events in the positive-class sample got shifted more towards 1 compared to background events in the same sample as they provide a more distinct difference from the negative-class event sample consisting of generated backward events (γ-hit, 3+ BG). This training scenario resulted in a very similar classifier, which is again strongly biased towards kinematic (*i. e.* p_T dependent) variables.

In conclusion, the use-case of models trained in this thesis are constrained to high mass continuum CEP events. However, as negative-class samples are composed of generated background events (γ-hit & 3+ track) and positive-class samples of a mixture of signal and background events, CWoLa is easily transferable to a

real data. Despite the difficulty of finding an optimal working point when training on real data, classification without labels provides a potential implementation of MVA methods in the analysis of CEP events, which makes it relevant for further investigations.

Ultimately, standard supervised machine learning applications (using simulated training data), involving CEP events, require a more complete CEP simulation to truly deploy the potential of multivariate analyses. An interesting package currently in development is called *ExDiff*, which simulates resonantly produced X particles in the low mass region. This simulation package provides a reference point for future studies.

Furthermore, besides being useful in MVA application such as classification without labels, the background approximations γ-hit-, and 3+ track-background can be used to estimate the high mass continuum background (feed-down) present in real data.

Appendices

Appendix A

γ-hit algorithm

The following algorithm is used to discriminate a charged pion from a potential gamma in the EMCal.

Algorithm 1 Identification of γ cluster in the EMCal.

```
for all clusters do
    min d_{C-T} ← 999 rad
    for all tracks do
        if current track not propagatable to EMCal surface then
            continue to next track
        else
            current d_{C-T} ← get_phi_eta_dist(current track, current cluster)
        end if
        if current d_{C-T} < min d_{C-T} then
            min d_{C-T} ← current d_{C-T}
        end if
    end for
    if min d_{C-T} > 0.51 rad then
        mark current cluster γ-hit
    end if
end for
```

Appendix B

Decay table

This section lists an extended version of the decay table Tab. 3.1 featuring decay channels until a cumulative occurrence of 80%. The color scheme is as follows: red colored decay channels describe 2π final states accompanied by additional final state gammas. Blue colored rows represent decay channels with more than two charged particles in the final state.

Decay	Occurrence[%]	Cumulative [%]
$X \to \pi^+ \rho^-$, $\rho^- \to \pi^0 \pi^-$, $\pi^0 \to \gamma\gamma$	21.82	21.82
$X \to \pi^+\pi^-$	19.66	41.48
$X \to \pi^0 \rho^0$, $\pi^0 \to \gamma\gamma$, $\rho^0 \to \pi^+\pi^-$	7.75	49.23
$X \to \pi^0 \pi^- \rho^+$, $\pi^0 \to \gamma\gamma$, $\rho^+ \to \pi^0 \pi^+$, $\pi^0 \to \gamma\gamma$	5.37	54.60
$X \to \pi^0 \pi^+ \pi^-$, $\pi^0 \to \gamma\gamma$	4.20	58.80

X π^0 $\gamma\gamma$ ω π^0 $\gamma\gamma$ π^+ π^-	3.64	62.44
X ρ^+ π^0 $\gamma\gamma$ π^+ ρ^- π^0 $\gamma\gamma$ π^-	3.52	65.96
X π^0 $\gamma\gamma$ ρ^+ π^0 $\gamma\gamma$ π^+ ρ^- π^0 $\gamma\gamma$ π^-	1.23	67.19
X π^0 $\gamma\gamma$ π^0 $\gamma\gamma$ ρ^0 $\pi^+\pi^-$	1.20	68.40
X π^0 $\gamma\gamma$ π^0 $\gamma\gamma$ π^- ρ^+ π^0 $\gamma\gamma$ π^+	1.08	69.48

X π^0 $\gamma\gamma$ π^0 $\gamma\gamma$ π^+ π^-	0.93	70.40
X ρ^0 $\pi^+\pi^-$ π^+ π^-	0.90	71.30
X ρ^0 $\pi^+\pi^-$ η $\gamma\gamma$	0.86	72.16
X ρ^0 $\pi^+\pi^-$ η π^0 $\gamma\gamma$ π^0 $\gamma\gamma$ π^0 $\gamma\gamma$	0.80	72.96
X π^+ ρ^- π^0 $\gamma\gamma$ π^- η $\gamma\gamma$	0.77	73.73
X π^+ π^- η $\gamma\gamma$	0.74	74.48

X π^0 $\gamma\gamma$ η π^0 $\gamma\gamma$ π^+ π^-	0.68	75.15
X π^- ρ^+ π^0 $\gamma\gamma$ π^+ η π^0 $\gamma\gamma$ π^0 $\gamma\gamma$ π^0 $\gamma\gamma$	0.52	75.68
X π^+ π^- π^- ρ^+ π^0 $\gamma\gamma$ π^+	0.49	76.17
X π^+ π^- ω π^0 $\gamma\gamma$ π^+ π^-	0.46	76.64
X π^+ ρ^- π^0 $\gamma\gamma$ π^- N $\overline{N}$	0.46	77.10

X π^0 $\gamma\gamma$ π^0 $\gamma\gamma$ ω π^0 $\gamma\gamma$ π^+ π^-	0.46	77.56
X K^0 K^0_L $\overline{K^0}$ K^0_S	0.43	77.99
X η $\gamma\gamma$ ω π^0 $\gamma\gamma$ π^+ π^-	0.43	78.43
X π^0 $\gamma\gamma$ π^+ π^- η $\gamma\gamma$	0.46	78.89
X ρ^0 $\pi^+\pi^-$ π^+ ρ^- π^0 $\gamma\gamma$ π^-	0.40	79.29

X π^+ π^- η π^0 $\gamma\gamma$ π^0 $\gamma\gamma$ π^0 $\gamma\gamma$	0.37	79.66
X π^0 $\gamma\gamma$ ρ^0 $\pi^+\pi^-$ η $\gamma\gamma$	0.34	80.00

TABLE B.1: *Extended decay table sorted by highest relative occurrence including decay channels until a cumulative occurrence of* 80%. *The color scheme is as follows: red colored decay channels describe* 2π *final states accompanied by additional final state gammas. Blue colored rows represent decay channels with more than two charged particles in the final state.*

Appendix C

Input variables

Here a short description of the training variables used in MVA is provided.

Feature name	Description	Signature
Event features		
nTrklets	Total number of tracklets	BLF
nSingles	Number of cluster on SPD layer 1 or 2 not associated with a tracklet on other SPD layers	BLF
nTrks	Total number of tracks	BLF
nResidual	Number of tracklets not associated with a track	BLF
nV0s	Number of V0s	BLF
FMDmult	Total multiplicity measured in FMD	BLF
V0mult	Total multiplicity measured in V0	BLF
ADmult	Total multiplicity measured in AD	BLF
Track features		
η	Pseudorapidity of particle	BLF
ϕ	Azimuthal angle of particle	BLF
nClusITS	Number of clusters in ITS	BLF
nClusTPC	Number of clusters in TPC	BLF
nClusTRD	Number of clusters in TRD	BLF
nSharedClusITS	Number of shared clusters in TPC	BLF
TPCsig	TPC signal	BLF
trkLen	Length of measured track	BLF
χ^2_{ITS}	Quality measure of the fit of the reconstructed track in the ITS	BLF
χ^2_{TPC}	Quality measure of the fit of the reconstructed track in the TPC	BLF
χ^2_{golden}	χ^2 between the TPC track constrained to the primary vertex and the global track	BLF
DCA_{xy}	Closest approach of the track to the primary vertex in the xy-plane	BLF
DCA_z	Closest approach of the track to the primary vertex in the z-direction	BLF
$d_{\phi-\eta}$	Distance in the 2D $\phi-\eta$ space between the two tracks	$\mathrm{BLF}_{\phi-\eta}$
Bayes	Bayesian PID probabilities to be of particle type a (with $a = \pi$, K, P)	Bayes
TPC $n\sigma$	Number of σ away from the mean expected TPC signal for the particle a (with $a = \pi$, K, P)	TPC $n\sigma$
p_T	Transverse momentum	p_T
φ_{1-2}	Angle between the two tracks	φ_{1-2}

TABLE C.1: *Variables used in the multivariate analysis. Additionally, the feature group is added.*

Bibliography

[1] O. J. P. Éboli, E. M. Gregores, and F. Halzen. Are two gluons the QCD Pomeron? *Nuclear Physics B Proceedings Supplements*, 71:349–357, 1999.

[2] F. Halzen and A. D. Martin. *Quarks and leptons: An introductory course in modern particle physics*. John Wiley & Sons, 1984.

[3] F. Mandl and G. Shaw. *Quantum field theory*. John Wiley & Sons, 2010.

[4] I. J. R. Aitchison and A. J. G. Hey. *Gauge Theories in Particle Physics: Volume I: From Relativistic Quantum Mechanics to QED, Third Edition*. Gauge Theories in Particle Physics: A Practical Introduction. Taylor & Francis, 2003.

[5] R. K. Ellis, W. J. Stirling, and B. R. Webber. *QCD and collider physics*. Cambridge monographs on particle physics, nuclear physics, and cosmology; 8. Cambridge University Press, 1996.

[6] S. Donnachie, G. Dosch, P. Landshoff, and O. Nachtmann. *Pomeron Physics and QCD*. Cambridge University Press, 2002.

[7] V. Barone and E. Predazzi. *High-Energy Particle Diffraction*. Springer Berlin Heidelberg, 2002.

[8] T. Regge. Introduction to complex orbital momenta. *Il Nuovo Cimento*, 14(5): 951–976, 1959.

[9] T. Regge. Bound states, shadow states and mandelstam representation. *Il Nuovo Cimento*, 18(5):947–956, 1960.

[10] P. D. B. Collins. *An Introduction to Regge Theory and High Energy Physics*. Cambridge Monographs on Mathematical Physics. Cambridge University Press, 1977.

[11] J. R. Forshaw and D. A. Ross. *Quantum Chromodynamics and the Pomeron*. Cambridge University Press, 1997.

[12] V. A. Petrov. Nonlinear Regge Trajectories in Theory and Practice. *American Institute of Physics Conference Proceedings*, 1105:266–269, 2009.

[13] T. Csörgö et al. Elastic Scattering and Total Cross-Section in $p + p$ reactions measured by the LHC Experiment TOTEM at $\sqrt{s} = 7$ TeV. *Progress of Theoretical Physics Supplements*, 193:180–183, 2012.

[14] M. Albrow. Central exclusive production issue: Introduction. *International Journal of Modern Physics A*, 29(28), 2014.

[15] O. Nachtmann. Considerations concerning diffraction scattering in quantum chromodynamics. *Annals of Physics*, 209(2):436–478, 1991.

[16] V. A. Khoze, A. D. Martin, and M. G. Ryskin. Prospects for new physics observations in diffractive processes at the LHC and Tevatron. *European Physics Journal*, C23:311–327, 2002.

[17] V. A. Khoze, A. D. Martin, and M. G. Ryskin. Double diffractive processes in high resolution missing mass experiments at the Tevatron. *European Physics Journal*, C19:477–483, 2001.

[18] K. Nakamura et al. Review of particle physics. *Journal of Physics*, G37:075021, 2010.

[19] The ALICE Collaboration et al. The ALICE experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):S08002, 2008.

[20] The ALICE Collaboration et al. Performance of the ALICE experiment at the CERN LHC. *International Journal of Modern Physics A*, A29:1430044, 2014.

[21] G. Dellacasa et al. ALICE technical design report of the inner tracking system (ITS). Technical report, 1999.

[22] The ALICE Collaboration et al. Technical design report of the time projection chamber. *CERN/LHCC*, 1(2000):375, 2000.

[23] ALICE Time-Of-Flight system (TOF): Technical Design Report. 2000.

[24] The ALICE Collaboration et al. Technical design report of the high momentum particle identification detector. *CERN/LHCC*, 98:19, 1998.

[25] The ALICE Collaboration et al. Technical design report of the transition-radiation detector. 2001.

[26] P. Cortese et al. ALICE Electromagnetic Calorimeter Technical Design Report. (CERN-LHCC-2008-014. ALICE-TDR-14), 2008.

[27] J. Allen and other. ALICE DCal: An Addendum to the EMCal Technical Design Report Di-Jet and Hadron-Jet correlation measurements in ALICE. Technical Report CERN-LHCC-2010-011. ALICE-TDR-14-add-1, 2010.

[28] G. Dellacasa et al. ALICE technical design report of the photon spectrometer (PHOS). 1999.

[29] S. Evdokimov. Diffraction Physics with ALICE at the LHC. In *Proceedings ot the 30th International Workshop on High Energy Physics*, pages 83–90, 2015.

[30] Alice Collaboration. Performance of the ALICE experiment at the CERN LHC. *International Journal of Modern Physics A*, 29:1430044, 2014.

[31] The ALICE Collaboration et al. ALICE technical design report on forward detectors: FMD, T0 and V0. *CERN-LHCC-2004-025*, 2004.

[32] A. Villatoro Tello. AD, the ALICE diffractive detector. *American Instsitute of Physics Conference Proceedings*, 1819(1):040020, 2017.

[33] R. Schicker. Central Diffraction in ALICE. *American Institute of Physics Conference Proceedings*, 1350:107–110, 2011.

[34] D. G. d'Enterria. Forward Physics at the LHC. In *Proceedings of the 15th International Workshop on Deep-inelastic scattering and related subjects (DIS 2007)*, volume 1 and 2, pages 1141–1152, 2007.

[35] D. Guest, K. Cranmer, and D. Whiteson. Deep Learning and its Application to LHC Physics. 2018.

[36] T. Sjöstrand et al. An introduction to PYTHIA 8.2. *Computer physics communications*, 191:159–177, 2015.

[37] Robert Ciesielski and Konstantin Goulianos. MBR Monte Carlo Simulation in PYTHIA8. *Proceedings of Science*, ICHEP2012, 2013.

[38] P. Lebiedowicz, O. Nachtmann, and A. Szczurek. Central exclusive diffractive production of the $\pi^+\pi^-$ continuum, scalar, and tensor resonances in pp and $p\overline{p}$ scattering within the tensor pomeron approach. *Physical Review D*, 93:054015, 2016.

[39] R. Brun et al. *GEANT 3: user's guide Geant 3.10, Geant 3.11; rev. version.* CERN, Geneva, 1987.

[40] R. Brun et al. Computing in ALICE. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 502(2-3):339–346, 2003.

[41] R. Brun and F. Rademakers. ROOT — an object oriented data analysis framework. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 389(1-2):81–86, 1997.

[42] A. Alan et al. *Proceedings Of The 29th International Conference On High Energy Physics: Ichep '98 (In 2 Volumes)*. World Scientific Publishing Company, 1999.

[43] O. Behnke et al. *Data analysis in high energy physics: a practical guide to statistical methods*. Wiley-VCH, 2013.

[44] I. Narsky and F. C. Porter. *Statistical analysis techniques in particle physics*. Wiley-VCH, 2014.

[45] P. Flach, J. Hernandez-Orallo, and C. Ferri. A coherent interpretation of auc as a measure of aggregated classification performance. pages 657–664, 2011.

[46] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[47] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 315–323, 2011.

[48] V. Nair and G. E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[49] J. S. Bridle. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In *Advances in Neural Information Processing Systems 2*, pages 211–217. 1990.

[50] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[51] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 06(02):107–116, 1998.

[52] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2): 157–166, 1994.

[53] E. H. Geoffrey et al. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

[54] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[55] G. Zhong, L.-N. Wang, X. Ling, and J. Dong. An overview on data representation learning: From traditional feature learning to recent deep learning. *The Journal of Finance and Data Science*, 2(4):265 – 278, 2016.

[56] Y. Lecun et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[57] C. Goller and A. Kuchler. Learning task-dependent distributed representations by backpropagation through structure. In *IEEE International Conference on Neural Networks*, volume 1, pages 347–352 vol.1, 1996.

[58] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[59] K. Cho et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014.

[60] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[61] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4768–4777, 2017.

[62] A. Rogozhnikov. Reweighting with boosted decision trees. In *Journal of Physics: Conference Series*, volume 762, page 012036. IOP Publishing, 2016.

[63] Y. Ganin et al. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[64] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[65] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[66] F. Chollet et al. Keras. `https://github.com/keras-team/keras`, 2015.

[67] M. Abadi et al. TensorFlow: A system for large-scale machine learning. In *Operating Systems Design and Implementation*, volume 16, pages 265–283, 2016.

[68] J. Pivarski et al. Scikit-hep/uproot: 2.9.6, 2018.

[69] P. Baldi, P. Sadowski, and D. Whiteson. Searching for Exotic Particles in High-Energy Physics with Deep Learning. *Nature Communications*, 5:4308, 2014.

[70] Y. Belikov, K. Safarík, and B. Batyunya. Kalman Filtering Application for Track Recognition and Reconstruction in ALICE Tracking System. 1997.

[71] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 249–256, 2010.

[72] N. Srivastava et al. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[73] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of Machine Learning Research*, volume 37, pages 448–456, 2015.

[74] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[75] A. Donnachie and P. V. Landshoff. Elastic scattering and diffraction dissociation. *Nuclear Physics B*, 244(2):322–336, 1984.

[76] E. M. Metodiev, B. Nachman, and J. Thaler. Classification without labels: Learning from mixed samples in high energy physics. *Journal of High Energy Physics*, 10:174, 2017.

[77] R. A. Ryutin. ExDiff Monte Carlo generator for Exclusive Diffraction. Version 2.0. Physics and manual. 2018.