UNIVERSIDADE ESTADUAL DE CAMPINAS
Instituto de Física "Gleb Wataghin"

Rafaela Ramos Sarmento

Study of the measurement of multiply charmed baryons in Pb-Pb collisions with ALICE 3 at the LHC

Estudo da medição de bárions multi charmosos em colisões Pb-Pb com ALICE 3 no LHC

Campinas
2022

Rafaela Ramos Sarmento

Study of the measurement of multiply charmed baryons in Pb-Pb collisions with ALICE 3 at the LHC

Estudo da medição de bárions multi charmosos em colisões Pb-Pb com ALICE 3 no LHC

Dissertação apresentada ao Instituto de Física "Gleb Wataghin" da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestra em Física, na área de Física.

Thesis presented to the "Gleb Wataghin" Instiute of Physics of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Physics, in the area of Physics.

Supervisor/Orientador: Prof. Dr. David Dobrigkeit Chinellato

ESTE TRABALHO CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELA ALUNA RAFAELA RAMOS SARMENTO, E ORIENTADA PELO PROF. DR. DAVID DOBRIGKEIT CHINELLATO.

Campinas
2022

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Física Gleb Wataghin
Lucimeire de Oliveira Silva da Rocha - CRB 8/9174

Sa74s

Sarmento, Rafaela Ramos, 1997-

Study of the measurement of multiply charmed baryons in Pb-Pb collisions with ALICE 3 at the LHC / Rafaela Ramos Sarmento. – Campinas, SP : [s.n.], 2022.

Orientador: David Dobrigkeit Chinellato.

Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Física Gleb Wataghin.

1. Grande Colisor de Hádrons (França e Suiça). 2. Plasma de quarks e glúons. 3. Experimento ALICE. I. Chinellato, David Dobrigkeit, 1983-. II. Universidade Estadual de Campinas. Instituto de Física Gleb Wataghin. III. Título.

Informações Complementares

Título em outro idioma: Estudo da medição de bárions multi charmosos em colisões Pb-Pb com ALICE 3 no LHC
Palavras-chave em inglês:
Large Hadron Collider (France and Switzerland)
Quark-gluon plasma
ALICE experiment
Área de concentração: Física
Titulação: Mestra em Física
Banca examinadora:
Jun Takahashi
Ernesto Kemp
Marcelo Gameiro Munhoz
Data de defesa: 07-10-2022
Programa de Pós-Graduação: Física

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0001-9881-6953
- Currículo Lattes do autor: http://lattes.cnpq.br/2027279864774353

MEMBROS DA COMISSÃO JULGADORA DA DISSERTAÇÃO DE MESTRADO DA ALUNA RAFAELA RAMOS SARMENTO - RA 186219 APRESENTADA E APROVADA AO INSTITUTO DE FÍSICA “GLEB WATAGHIN”, DA UNIVERSIDADE ESTADUAL DE CAMPINAS, EM 07/10/2022.

COMISSÃO JULGADORA:

- Prof. Dr. Jun Takahashi – Presidente (IFGW/UNICAMP)

- Prof. Dr. Ernesto Kemp (IFGW/UNICAMP)

- Dr. Marcelo Gameiro Munhoz (Universidade de São Paulo)

OBS.: Ata da defesa com as respectivas assinaturas dos membros encontra-se no SIGA/ Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

CAMPINAS

2022

Valeu a pena? Tudo vale a pena
Se a alma não é pequena.
Quem quer passar além do Bojador
Tem que passar além da dor.
Deus ao mar o perigo e o abismo deu,
Mas nele é que espelhou o céu.
(Fernando Pessoa, Mar Português)

It's not about how hard you hit. It's about how hard you can get hit and keep moving forward. How much you can take and keep moving forward.
(Rocky Balboa, Rocky)

Per Aspera Ad Astra.

Acknowledgements

First, I would like to express my deepest and sincere gratitude to my advisor, Prof. Dr. David Dobrigkeit Chinellato for having accepted to supervise me during these two years of my master's degree. It was through his mentoring, discussions, and knowledge that the results presented in ths work could be achieved.

I also like to thank all the members of the HadrEx group for their support and discussions. I especially thank Prof. Dr. Jun Takahashi for the suggestions to improve my results and Gabriel Reis Garcia, whose contribution was fundamental to the development of my results using machine learning.

I would also like to thank my mother, Silvia Claudia Ramos, and my grandmother, Adoninha Lopes Delgado, as their support was essential to viabilize this work. I owe them much more than this thanks, I owe them my life.

I am also thankful to my family and friends who have always been there when I needed it. Life without you is meaningless. In particular, I thank my cousins Isabela Ramos de Almeida, Izadora Ramos de Almeida, Bruna Ramos de Almeida and my good friends Jully Vanessa Leandro Nascimento, Henrique Vieira de Souza, Marina Reggiani Guzzo, Matheus De Carvalho Loures and Luiza Lober de Souza Piva.

Words cannot express my gratitude to André Saugo Mazzari for all the support, affection, and especially patience he has shown me over the years.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 granted by the project 88887.517275/2020-00.

Resumo

Hádrons são partículas subatômicas compostas por quarks. Quando núcleos suficientemente pesados são acelerados e colidem entre si, é possível encontrar características fundamentais para o estudo das propriedades da matéria. Entre essas características está a criação do Plasma de Quarks e Gluons (QGP), um estado da matéria onde os quarks, partículas usualmente confinadas, tornam-se desconfinadas. O estudo da QGP pode fornecer respostas sobre o universo primordial e a restauração de uma das simetrias mais importantes da física, a simetria quiral, da qual sabe-se que a sua quebra espontânea gera as massas de uma grande parte das partículas.

O ALICE 3, versão atualizada do atual experimento ALICE, é um futuro experimento que visa responder algumas questões em aberto sobre a QGP. Um dos objetivos do ALICE 3 é o estudo de bárions multi charmosos, como o Ξ_{cc}^{++}. Previsões teóricas revelaram que esta partícula possui uma taxa relativa de produção em colisões de Pb-Pb acima de cem vezes mais do que a produção em sistemas pequenos, como colisões p-p. Se esta produção relativa de Ξ_{cc}^{++}, quando comparado a produção de outras partículas carregadas, for observada pelo experimento, ela poderá ser utilizada como uma clara evidência da formação da QGP em sistemas pesados.

O presente trabalho tem o objetivo de estudar a possibilidade da medição de bárions multi charmosos, em especial o Ξ_{cc}^{++}, em colisões Pb-Pb com o detector ALICE 3 no LHC, através de informações topológicas do decaimento das partículas reconstruídas pela técnica inovadora, o strangeness tracking. Para isto, procedimentos de seleção de candidatos foram realizados a fim de obter o maior valor possível da significância, métrica que indica a precisão estatística de uma possível medida, em dados oriundos de simulações de Monte Carlo. Em especial, o trabalho visou maximizar a significância no intervalo de momento transversal, definido por 0.0-2.0 GeV/c. Duas abordagens foram feitas: a primeira usando o procedimento usual através de cortes retangulares, do qual foi possível obter uma significância para este intervalo de aproximadamente 3σ; já na segunda abordagem utilizou-se técnicas de machine learning, onde foi obtida uma significância de aproximadamente 9σ, o que indica ser possível medir o espectro de massa invariante desta partícula até momento transversal igual a zero. Este valor final obtido para a significância demonstra o alto poder desta técnica para seleção de candidatos.

Palavras-chave: Grande Colisor de Hádrons (França e Suiça); Plasma de quarks e glúons; Experimento ALICE

Abstract

Hadrons are subatomic particles composed of quarks. When sufficiently heavy nuclei are accelerated to a speed close to that of light and collide with each other, it is possible to find fundamental characteristics for the study of the properties of matter. Among these characteristics is the creation of the Quark-Gluon Plasma (QGP), a state of matter where quarks, usually confined particles, become unconfined. The QGP study can provide answers about the early universe and the restoration of one of the most important symmetries in physics, the chiral symmetry, whose spontaneous breaking is known to generate the masses of a large part of particles.

ALICE 3, an updated version of the current ALICE experiment, is a future experiment that aims to answer some open questions about QGP. One of the goals of ALICE 3 is the study of multiply charmed baryon, such as Ξ_{cc}^{++}. Theoretical predictions revealed that this particle has a relative production rate in Pb-Pb collisions above a hundred times higher than the production rate in small systems, such as p-p collisions. If this relative production of Ξ_{cc}^{++} compared to the production of other charged particles is observed in the experiment, it can be used as clear evidence for the formation of QGP in heavy systems.

The present work aims to investigate the possibility of measuring multiply charmed baryon, especially the Ξ_{cc}^{++}, in Pb-Pb collisions with the ALICE 3 detector at the LHC by using topological information from the decay of the reconstructed particles through the innovative technique, strangeness tracking. For this, candidate selection procedures were carried out in order to obtain the highest possible value of significance, a metric that indicates the statistical accuracy of a possible measure, with Monte Carlo simulation data. In particular, this work aims to maximize the significance at the transverse momentum interval, defined by 0.0-2.0 GeV/c. Two approaches were used: the first using the usual procedure through rectangular cuts, which achieved a significance for this interval of about 3σ; and the second approach applied machine learning, where a significance of approximately 9σ obtained, which indicates that it is possible to measure the invariant mass spectrum of this particle up to zero in transversal momentum. This final significance value shows the high performance of this technique in the candidate selection procedure.

Keywords: Large Hadron Collider (France and Switzerland); Quark-gluon plasma; ALICE experiment

List of Figures

List of Tables

Contents

Chapter 1

Introduction

There are many open questions in nature, especially when it comes to the subatomic world of elementary particles. One of the major research interests in recent decades has been the Quark-Gluon Plasma (QGP), a state of matter characterized by high density and temperature in which quarks, particles confined in hadrons, become unconfined. This state occurs when heavy nuclei collide at speeds very close to that light. Numerous studies involving several research centers devoted to exploring the physics that involves this state of matter. In particular, knowledge of the QGP will open the doors to studies of the early universe, since it is believed that during the first microseconds after the Big Bang the universe was in a deconfined state. If the early Universe can be described as a state of unconfined matter, then its thermalization, expansion, and evolution would be explained in terms of how it occurs in QGP.

The Large Hadron Collider (LHC) [1] is the world's largest and highest energy particle accelerator. One of its main experiments is A Large Ion Collider Experiment (ALICE) [2], which was proposed to study the properties of QGP in heavy ion collisions. An ALICE upgrade is expected to be in operation by the beginning of the next decade. This upgrade is known as "ALICE 3", and will aim to answer open questions about QGP left by its predecessor. In particular, this experiment will have as one of its objectives the characterization of the production of multi charmed baryons and how these production rates compare to collisions in which the formation of this plasma is not expected to occur.

This work aims to explore data from Monte Carlo simulations that were generated with ALICE 3 in mind. From these data, the objective is to study the multi charm baryon Ξ_{cc}^{++}, particle that by having two charm quarks allows the verification of the coalescence process in QGP. In addition to determine candidate selection criteria to isolate the Ξ_{cc}^{++} signal, allowing for the understanding of the behavior of the signal and the background involved in the trajectory reconstruction of this particle. The information about trajectory reconstruction came from the first performance tests of an novel experimental technique “strangeness tracking”, that relies on measuring hadrons before they decay.

This thesis has been structured to provide the reader with enough knowledge to understand the results that will be presented at the end. Therefore, Chapter 2 will provide a brief introduction to the area, going through a contextualization of the physics of High-Energy Heavy-Ion Collisions that is relevant to this thesis and, especially going into the topic of the production of multi charmed particles in heavy-ion collisions. Chapter 3 aims to contextualize, focusing in an experimental interpretation, where these particles will be measured, i.e., it attempts to explain the proposals and composition of the ALICE 3 detectors.

Chapter 4 presents the properties of the baryon Ξ_{cc}^{++}, the standard procedure of candidate selection criteria, details about the characteristics and functionality of the strangeness tracking technique, the signal extraction after selection, and finally how the significance is calculated using signal and background data. Chapter 5 introduces the concepts of the machine learning technique used in this thesis, the Gradient Boosted Decision Tree (XGBoost) [3]. In addition, Section 5.2 explains how the problem was modeled to use the machine learning technique.

Chapter 6 is the chapter where the results of this work are presented. It is expected that the reader will be able to understand them with the knowledge previously provided.

The main results obtained in this thesis and an outlook for possible improvements will be presented in the “Conclusions" chapter.

Chapter 2

Introduction to High-Energy Heavy Ion Collisions

Matter as we know it is made up of two types of particles: fermions and bosons. Fermions are constituent particles and bosons are the particles that mediate interactions. Fermions are semi-integer spin particles that have the dynamics described by the Dirac equation, obey the Fermi-Dirac statistics, and comprise leptons and quarks, both with six distinct particles. Bosons are particles with an integer spin and obey the Bose-Einstein statistics.

Bosons mediate three of the four fundamental forces: electromagnetism, the strong force, and the weak force. The theory that describes all the properties and interactions of particles is known as The Standard Model of Particle Physics. Despite not being a complete theory, it explains with excellence the phenomena and experiments that involve particle physics.

Within this theory, it is possible to study phenomena such as Quark-Gluon Plasma (QGP), which is a state of matter that reaches extreme conditions such that quarks, which are confined particles, becomes asymptotically free. Studying this state of matter allows for a deeper understanding of the primordial universe since it is believed that in the first microseconds after the Big Bang, the universe was in a state of plasma of quarks and gluons.

In this chapter, a brief introduction will be made to the area of high energy particle physics, especially to the physics that involves quarks and the unconfined state of matter, the QGP.

2.1 The Standard Model of Particle Physics

The Standard Model is one of the best and most sophisticated theories developed in modernity, having been extensively tested over many years. The modern version of this theory was solidified in the 70's and it is a comprehensive theory that identifies elementary particles and specifies how they interact.

In the Standard Model, elementary particles are those that are not made up of other particles. That is, they are particles that have no internal structure. There are two groups of elementary particles that compose the ordinary matter: quarks and leptons. Quarks are particles confined in hadrons, like protons and neutrons. These particles together with gluons are the only ones that carry the color charge, the equivalent of electrical charge in Quantum Chromodynamics (QCD). Subsequent sections will be dedicated to the study of the properties of these particles.

Leptons are fermions particles of spin 1/2, that can be divided into two groups: charged and neutrinos. Both of them cannot interact via the strong force. The charged particles are the electron (e^-), muon (μ^-), and tau (τ^-). These particles define the 'flavor' of the leptons particles. Neutrinos are light neutral particles that interact only by the weak force. In nature, there one type of neutrino for each flavor: electron neutrino (ν_e), muon neutrino (ν_μ), and tau neutrino (ν_τ). Being fermions, all leptonic particles obey the Dirac Equation, which, for a free particle, is given by [4]:

$$(i\gamma_\mu \partial^\mu - m)\Psi = 0, \tag{2.1}$$

where i is the imaginary unit, γ_μ are the Pauli matrices, $\partial^\mu = (\partial^t, \partial^x, \partial^y, \partial^z)$ are the derivative operators in four dimensions, m is the fermion mass and Ψ is a vector of the fermion wave functions. Solving the equation for a particle at rest results in two states for the positive energy solution and two states for the negative energy solution, as follows:

$$\Psi_A = e^{-imt}\begin{pmatrix} 1 \\ \\ 0 \end{pmatrix} \quad , \quad \Psi_A = e^{-imt}\begin{pmatrix} 0 \\ \\ 1 \end{pmatrix} \quad \text{,positive energy solution.} \tag{2.2}$$

$$\Psi_B = e^{+imt}\begin{pmatrix} 1 \\ \\ 0 \end{pmatrix} \quad , \quad \Psi_B = e^{+imt}\begin{pmatrix} 0 \\ \\ 1 \end{pmatrix} \quad \text{,negative energy solution.} \tag{2.3}$$

Each of these states corresponds to one of two spin states of a spin-1/2 fermion. To explain the negative energy solution, Dirac proposed the 'Dirac sea', a fully filled state and, therefore, not accessible, in which there would only be electrons with negative energy and would be the description of the vacuum state in Dirac's theory. When a sufficiently energetic photon provides its energy to an electron in this sea, it excites and creates an electron with positive energy. This causes the electron to leave the vacuum state and behave as a normal electron. In the Dirac Sea, a hole would be created that would behave like a particle with the same properties as an electron, but with a positive electrical charge. This hole, in the current Quantum Field Theory, is explained as the creation of an anti-particle, named "positron" in this case. For all fermions, there exists an antiparticle, which has the same properties as its respective particle but with the opposite sign in the electrical charge.

The flavors of quarks and leptons and the forces experienced by them are shown in Tab. 2.1 . Note that all these properties can be extended to the respective antiparticles.

					strong	electromagnetic	weak
Quarks (colored)	down type	s	b	d	✓	✓	✓
	up type	u	c	t	✓	✓	✓
Leptons (color-free)	charged	e^-	μ^-	τ^-		✓	✓
	neutrinos	ν_e	ν_μ	ν_τ			✓

Table 2.1: The forces experienced by different fundamental particles. Adapted from [5].

In the boson group, there are two types of fundamental particles: gauge and scalar bosons. The first type are bosons that act as force carriers, such as the photon. The interactions of elementary particles, described by a gauge theory, are described by the exchange of gauge bosons, acting as virtual particles. For every fundamental force described by the standard model, there is at least one gauge boson: the photon for electromagnetism; gluons for the strong interaction; and W and Z bosons for the weak interaction. For gravity, the existence of a force carrier particle, the graviton, was hypothesized, however, this particle has never been measured. The second type of elementary boson is the scalar boson, which has spin equal to zero and has only one representative in the Standard Model: the Higgs boson. The interaction between the Higgs field and an initially massless particle gives the mass of the latter. The Higgs was discovered in 2013 at the Large Hadron Collider (LHC) [1] by the Compact Muon Solenoid (CMS) [6] and the A Toroidal LHC Apparatus (ATLAS) [7] experiments. This event represented a remarkable validation of the Standard Model of Particle Physics. A summary of the Standard Model, with the respective fermions and elementary bosons is shown in Fig. 2.1.

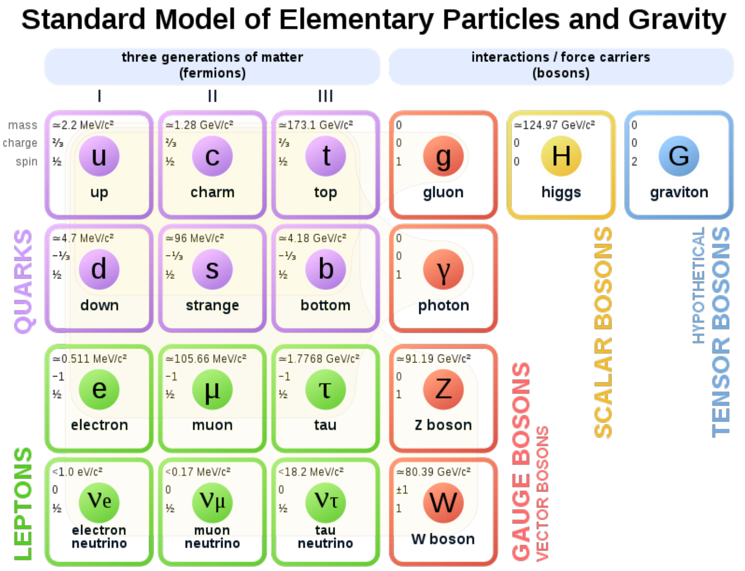

Standard Model of Elementary Particles and Gravity
three generations of matter (fermions)
interactions / force carriers (bosons)
I
II
III
mass
charge
spin
≃2.2 MeV/c² ⅔ ½ u up
≃1.28 GeV/c² ⅔ ½ c charm
≃173.1 GeV/c² ⅔ ½ t top
0 0 1 g gluon
≃124.97 GeV/c² 0 0 H higgs
0 0 2 G graviton
≃4.7 MeV/c² −⅓ ½ d down
≃96 MeV/c² −⅓ ½ s strange
≃4.18 GeV/c² −⅓ ½ b bottom
0 0 1 γ photon
≃0.511 MeV/c² −1 ½ e electron
≃105.66 MeV/c² −1 ½ μ muon
≃1.7768 GeV/c² −1 ½ τ tau
≃91.19 GeV/c² 0 1 Z Z boson
<1.0 eV/c² 0 ½ νe electron neutrino
<0.17 MeV/c² 0 ½ νμ muon neutrino
<18.2 MeV/c² 0 ½ ντ tau neutrino
≃80.39 GeV/c² ±1 1 W W boson
QUARKS
LEPTONS
GAUGE BOSONS
VECTOR BOSONS
SCALAR BOSONS
HYPOTHETICAL
TENSOR BOSONS

Figure 2.1: Elementary particles and their properties in the Standard Model of Particles Physics [8].

2.2 Quark confinement and asymptotic freedom

In 1964, the physicists Murray Gell-Mann and George Zweig separately proposed the existence of subatomic particles. They stated that the most important properties of hadrons could be better explained if, in fact, they were formed by even smaller particles. These smaller particles are now known as quarks. Quarks are elementary that have spin 1/2, and therefore are fermions. There are six types of quarks in nature, known as flavors: up (u), down (d), charm (c), strange (s), top (t), and bottom (b) (see Tab. 2.1). The first two are lighter and the most stable in nature. In addition to spin, these particles have other characteristics, such as electric charge, mass, baryonic number, weak isospin, charm, strangeness, topness, and bottomness. A summary of these properties for the six quark flavors is shown in Tab. 2.1. For every quark flavor, there is a corresponding antiquark, which differs from quarks only in the electric charge sign.

In nature, quarks are observed only confined to hadrons via the strong force, through the gluons, and have never been measured experimentally in an isolated state. Hadrons can be observed in two possible configurations: baryons (3 quarks) or mesons (2 quarks). In the

Quark Flavour	Mass [MeV/c^2]	Electric charge [e]	Spin (J)	Baryon number	Weak isospin (I_3)	Charm (C)	Strangeness (S)	Topness (T)	Bottomness (B')
up (u)	2.2	+2/3	1/2	1/3	+1/2	0	0	0	0
down (d)	4.7	-1/3	1/2	1/3	-1/2	0	0	0	0
charm (c)	1280	+2/3	1/2	1/3	0	+1	0	0	0
strange (s)	96	-1/3	1/2	1/3	0	0	-1	0	0
top (t)	173100	+2/3	1/2	1/3	0	0	0	+1	0
bottom (b)	4180	-1/3	1/2	1/3	0	0	0	0	-1

Table 2.2: Properties of quarks for each respective flavor.

first case, baryons are semi-integer spin particles, such as protons and neutrons. Mesons, in turn, are particles of integer spin, such as pions, kaons, and J/Ψ. To allow the interaction via the strong force, another property present in quarks and gluons is the color charge. For quarks, there are three different color types: red (r), green (g), and blue (b). The color charge of the gluons is a mixture of two of these three colors, totaling eight possible colors. For the antiquarks, there are three anticolors: anti-red ($\overline{r}$), anti-green ($\overline{g}$), and anti-blue ($\overline{b}$).

Differently from quarks, hadrons are white particles, that is, mesons and baryons have combinations of quarks in such a way that the resultant color is zero. For mesons, the combination is a quark (q) and an anti-quark ($\overline{q}$). For baryons, the combination of quarks must be an rgb or $\overline{r}\overline{g}\overline{b}$.

The theory which describes the interaction between quarks and gluons is Quantum Chromodynamics (QCD). In this theory, it is impossible to measure directly the colors of quarks, because color triplet states objects (quarks and gluons) are always confined into singlet states (hadrons)[9]. QCD predicts that the confinement of quarks occurs since quarks interact via gluons and these, because they also have color, interact with each other through an attractive potential. These potentials are taken to be of the form [10]:

$$V_{eff} = -\frac{4}{3}\frac{\alpha_s(r)}{r}\hbar c + \kappa r, \tag{2.4}$$

where r is the distance between colored particles, $\hbar$ is Planck constant, c is the speed of light in vacuum, $\kappa \approx$ 1GeV/fm and α_s is the strong coupling constant or QCD coupling constant. This last factor depends on the square of the scale of the momentum transfer Q and assumes the following form [9]:

$$\alpha_s(Q^2) = \frac{\alpha_0}{1 + \alpha_0 \frac{(32-2n_f)}{12\pi} \ln\left(\frac{-Q^2}{\mu^2}\right)}, \tag{2.5}$$

where α_0 is the coupling constant for momentum transfer μ and n_f is the number of flavors. A summary of the measurements of α_s at different |Q| scales is shown in Fig. 2.2. It can be seen that the value of α_s increases as Q decreases.

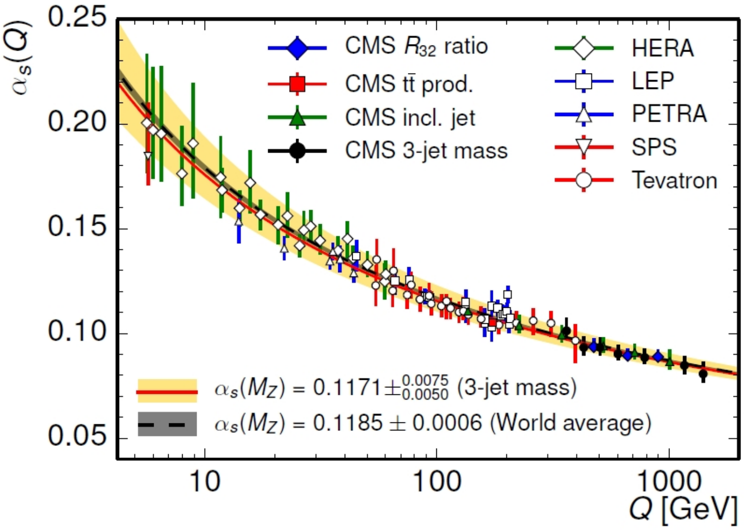


Figure 2.2: Measurements of the coupling constant α_s from different experiments for different |Q| scales. Figure from [11].

In the Eq. 2.4 there are two important limits to be studied. The first limit is when the interaction takes place at high energies or if r assumes small values in such a way as to tend to zero:

$$V_{eff}(r \longrightarrow 0) \approx -\frac{4}{3}\frac{\alpha_s(r)}{r}\hbar c. \tag{2.6}$$

In this case, the colored particles are so close that the effective potential is dominated by a behavior like Coulomb potential (1/r) and quarks become nearly free particles. In this domain, the particles are in Asymptotic Freedom and the coupling constant becomes small, as shown in Fig. 2.2. This is the regime of the perturbative QCD, or p-QCD.

The second limit is when the interaction occurs at low energies or when r assumes large values, such that the effective potential takes the following form:

$$V_{eff}(r \longrightarrow \infty) \approx \kappa r. \tag{2.7}$$

The interaction between colored particles becomes stronger as r increases, leading to the confinement of quarks and gluons. Experimentally, what is observed is that as a large amount of energy is deposited in the system in order to separate the particles, this stored energy, which behaves as κr, reaches a point that it is energetically more favorable for a new pair of particle and anti-particle to be created than to separate the initial pair. This process is qualitatively shows in Fig. 2.3. For this reason, quarks are never observed isolated.

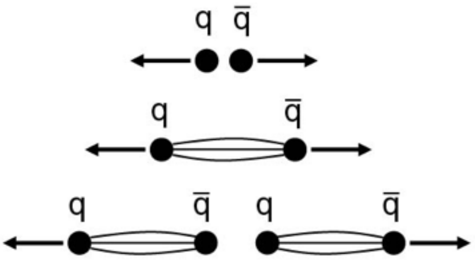


Figure 2.3: Qualitatively diagram that represents a gluon string break, creating a new pair $q\bar{q}$. Figure adapted from [5].

2.3 Quark Deconfinement & The Quark Gluon Plasma

At high energies or at small distances, the interaction between colored particles becomes small, i.e., quarks and gluons interact weakly, and the asymptotic freedom regime takes place, as described in the last section. In this condition, as the density of colored particles increases, a phase transition occurs to a state of matter known as The Quark-Gluon Plasma (QGP). In this state, quarks and gluons interact effectively by a saturated potential, shown qualitatively in Fig. 2.4. In this phase, which is characterized by a volume V, these particles behave like free particles, so it is possible for them to pass through the entire high-density volume unimpeded. The QGP was not initially predicted in the QCD theory, thus, in Fig. 2.4, the continuous red curve shows what would be the behavior of the effective potential between two quarks, which consists of the Eq. 2.4. However, studies using lattice calculations, the lattice gauge theory, indicated a new state of matter, with this new state that is observed being closer represented by the dashed blue curve.

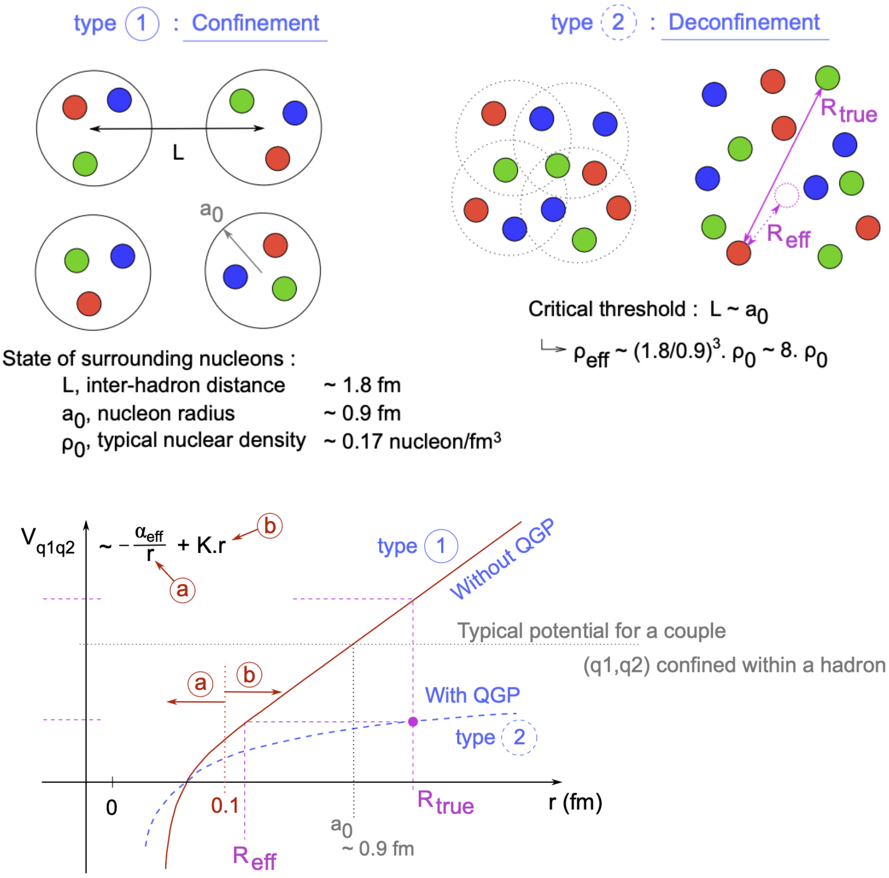

type 1 : Confinement
type 2 : Deconfinement
L
a0
Rtrue
Reff
Critical threshold : L ~ a0
↳ ρeff ~ (1.8/0.9)³. ρ0 ~ 8. ρ0
State of surrounding nucleons :
L, inter-hadron distance ~ 1.8 fm
a0, nucleon radius ~ 0.9 fm
ρ0, typical nuclear density ~ 0.17 nucleon/fm³
Vq1q2
~ - αeff/r + K.r
a
b
type 1
Without QGP
Typical potential for a couple
(q1,q2) confined within a hadron
With QGP
type 2
0
0.1
Reff
a0
~ 0.9 fm
Rtrue
r (fm)

Figure 2.4: Effective potential as function of the distance between two quarks. In the continuous line the behavior of the potential with no QGP; in the dashed blue line, with QGP. R_{eff} is the effective distance between these quarks and R_{true} is the real distance. Figure adapted from [12].

Fig. 2.5 shows an illustrative scheme of the QCD phase diagram, where the x-axis indicates the baryonic density normalized by the nuclear hadronic density $d_0 = 0.17\ nucleon/fm^3$, and the y axis the temperature in MeV. There are two extreme ways in which the phase transition from hadronic matter to quark matter can occur. The first one is commonly called 'hot QGP'. In this case, the transition occurs when the baryonic density is close to zero, that is, the number of quarks and the number of anti-quarks in the system are equal and the temperature reaches very high values. The transition, in this case, occurs vertically in the Fig. 2.5. Calculations using Monte Carlo lattice QCD simulations show that for this regime the phase transition occurs at $T_c \approx 150 - 200$ MeV, which is extremely high if compared with the temperature at the center of the sun, $1.5 \times 10^7 K = 1.3$ keV, that is, 5.0×10^5 times greater than this temperature [13]. Studies show that the hot QGP may have occurred in the first 10^{-5}s of the Universe's lifetime after the Big Bang. The reason is that if a temporal extrapolation is made to these initial seconds of the Universe the matter and the radiation was increasingly hotter and denser resulting in a “primordial fireball”, where the QGP would be formed. Thus, studies on the phase and evolution of the QGP can provide information about the early Universe and its respective evolution.

The second regime is called “Cold QGP” and it occurs when the baryonic number density increases and the temperature, relatively, becomes small. The transition for this case occurs horizontally, as in Fig. 2.5. Due to its extreme conditions, few studies are possible, and what is expected is that this regime can be found at the core of super dense stars, for example, neutron stars.

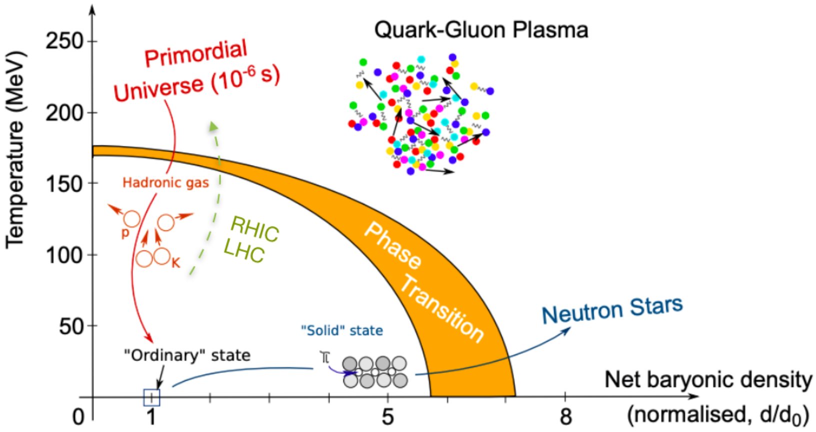


Figure 2.5: QCD phase diagram for different conditions of temperature and baryonic number density. The x axis indicates the baryonic density normalized by the nuclear hadronic density $d_0 = 0.17\ nucleon/fm^3$ and the y axis the temperature in MeV. Figure adapted from [12].

Another situation in which the occurrence of QGP can be found is in the initial stage of colliding heavy nuclei at high energies. When two heavy nuclei, such as Pb-Pb (lead-lead), for example, are accelerated such that they are both in a relativistic/ultra-relativistic regime when a head-on collision occurs, and if the center of mass is greater than 100 GeV, the resultant matter has high energy density and high temperature, but low baryonic density, so in the early stage of this collision the QGP is formed. The formation of QGP in laboratories, such as LHC, is shown in Fig. 2.6 and can be described by the following steps:

1. **Initial hard scattering:** The collision between two relativistic heavy nuclei. In the range of relativistic energies, the nuclei suffers a Lorentz contraction and the collision is characterized by a high momentum transfer. The colliding nuclei tend to pass through

each other, and the matter produced between the receding nuclei has high values of energy density and temperature, while the baryonic density is low [13].

2. **Pre-equilibrium stage and thermalization:** At this stage, the matter produced by the heavy nuclei partons (quarks and gluons) begins to thermalize, i.e., they tend to reach the thermal equilibrium through mutual interaction. In this stage, there is a high production rate of particles, in special heavy quarks, and the system expands. When the partonic matter reaches equilibrium, the QGP is formed, which is marked by a high temperature, density and pressure.

3. **Hadronization:** In this stage, there is a great expansion of the volume. The density and temperature start to decrease. It is at this moment that the first hadronic species are produced.

4. **Chemical and Kinematic Freeze-out:** In the freeze-out, the end of the expansion of the medium and the ceasing the inelastic interactions between hadrons occur. The particles decay and can then be measured in large experiments.

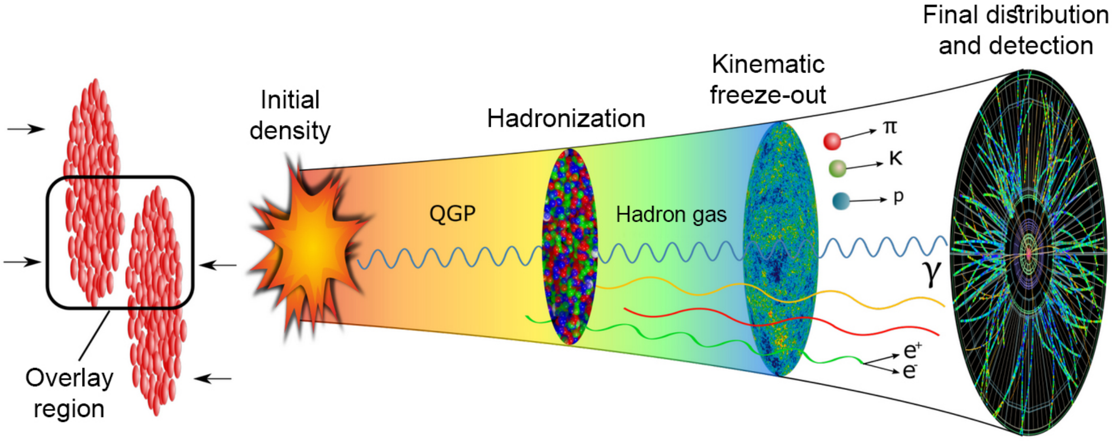


Figure 2.6: The time evolution of a high-energy heavy ion collision. Figure from [14].

It is worth mentioning that the collision of heavy ions is a very fast process, contained in a spatial length of about 10 fm, and a time scale of approximately 10 fm/c (10^{-23}) s [13]. As QGP is created in the initial stages of this process, a direct measurement of this plasma becomes unfeasible. Thus, as many particles are emitted during the process, the information from those particles can be used in order to obtain QGP signals. The following are some possible QGP signatures, that is, some measurements indicating that the plasma was created in the timeline of a heavy-ion collision:

1. **Enhanced production of strangeness from QGP**: experimental data showed that the production of strangeness is much higher than the expected abundance if they were only produced by hadronic interactions, and it saturates in a sufficiently excited QGP [15].

2. **Enhancement of thermal photons and dileptons due to emission from deconfined QCD plasma**: The deconfined medium is electrically charged and radiates photons and dileptons throughout its expansion. As photons and dileptons interact only by electromagnetic force, and this is very low in the volume that occurs the QGP, these particles can then be detected without having suffered interference from the medium. Thus, the transversal momentum distribution of these particles reflects the local properties of the QGP at their point of emission, providing information about these properties at the various stage of the collisions, including the early stages [16].

3. **Increase of an elliptic flow (v_2) of hadrons**: As the QGP is characterized by being a plasma, the particles emitted behave like a fluid, and then all of these particles must have a common transverse velocity (collective flow) in addition to their thermal motion, that is, the volume characterized by the plasma is expected to be described by hydrodynamic models. Thus, the existence of a collective expansion is important to describe the space-time evolution of a heavy-ion collision. This phenomenon can be observed due to a large azimuthal anisotropy (elliptic flow) of the emitted particles. First observations were made by the Relativistic Heavy Ion Collider (RHIC) [17].

4. **Increased production rate of multi-charmed baryons in heavy-ion collisions (A–A) if compared to proton-proton ($p - p$) collisions**. In the early stages of heavy-ion collision, heavy quarks such as the charm quark are produced. An increase in the rate of production of these particles in collisions of heavy ions with respect to collisions of small systems like $p - p$ indicates that there is a kinetic equilibrium of heavy quarks in the QGP.

About the last signature, the increased production rate of multi-charmed baryons, the next section will make a dedicated discussion about its process and measurement.

2.4 The Statistical Coalescence Model and Multi-Charmed Baryons

The Statistical Coalescence Model (SCM) or Statistical Hadronization Model expanded to quark charm ($SHMc$) [18] is a model that attempts to describe the production of multi-charmed baryons in heavy-ion collisions. With this approach, it is possible to derive formulas that allow to calculate the yield of multi charm particles and to demonstrate that in large system collisions the relative production rate of these particles is much higher than in p-p collisions. This may be intrinsically related to QGP as this phase is expected to form in heavy systems collisions such as Pb-Pb, while in p-p collisions this phase should not occur due to the fact that this system does not have a large enough volume for the formation of a QGP phase. Moreover, it shows that if more than one charmed quark-antiquark pairs are created in the collision, at the hadronization point they will coalesce into hadrons where double and triple charms can be formed. The observations of such effects would represent a great achievement for the study of the properties of deconfined matter, as the the comparison between observations and the predictions can provide a sensitive measure of the degree of equilibration of charm quarks in the medium and, beyond that, a system-size dependence.

According to the hypotheses of the SHMc model, charmed quark-antiquark pairs are created at the initial stage of a heavy-ion collision. This should occur because the time of charm equilibration in the QGP is large and exceeds the lifetime of the fireball. In addition, the rate of charm production and annihilation should be low to keep the number of heavy quark-antiquark pairs at their chemical equilibrium value at later stages.

The $SHMc$ assumes the following postulates to describe the production of charmed quarks [19] in the Quark-Gluon Plasma:

1. The charm quark (c) and anti-quark ($\overline{c}$) are created at the initial stage of A-A reaction in a heavy-ion collision;

2. Creation and annihilation of $c\overline{c}$ pairs can be neglected at later stages of the collision;

3. The formation of observed hadrons with open and hidden charms takes place near the point of chemical freeze-out in accordance with the laws of statistical physics.

From these postulates, it is possible to use the SHMc model to calculate the average multiplicities of the production of multi-charm hadronic species. For this, it is needed a partition function. In the case of the numbers charm and beauty, it is not possible to use directly the Grand Canonical ensemble, as is done for cases of electric charge, strangeness, and baryonic number, because the multiplicity of hadronic species with charm and beauty flavor is not large enough for the use of this ensemble.

Let ν_c be the number of c, $\nu_{\bar{c}}$ of $\overline{c}$, ν_b of b, and $\nu_{\bar{b}}$ of $\overline{b}$ quarks. The relevant partition function was calculated in [18] [20] and is given by:

$$Z\left(\nu_c, \nu_{\bar{c}}, \nu_b, \nu_{\bar{b}}\right) = Z_l \left[\prod_{f=c,\bar{c},b,\bar{b}} \int_{-\pi}^{\pi} \frac{\mathrm{d}\phi_f}{2\pi} \mathrm{e}^{1\nu_f \phi_f}\right] \times \exp\left[\sum_j z_j \lambda_j \mathrm{e}^{-1\nu_{cj}\phi_c - 1\nu_{\bar{c}j}\bar{\phi}_c - 1\nu_{bj}\phi_b - 1\nu_{\bar{b}j}\bar{\phi}_b}\right], \tag{2.8}$$

where Z_l is the grand-canonical partition function including all light-flavored species, λ_j are the fugacities with respect to electric, baryonic and strangeness charges, ν_{cj}, $\nu_{\bar{c}j}$, ν_{bj}, $\nu_{\bar{b}j}$ are the number of c, $\bar{c}$, b and $\bar{b}$ quarks, respectively, of the j-th hadronic species and z_j are one-particle partition functions, which are given by:

$$z_j = \frac{g_j V}{2\pi^2} m^2 T K_2 \left(\frac{m}{T}\right) \underset{m \gg T}{\simeq} g_j V \left(\frac{mT}{2\pi}\right)^{3/2} \mathrm{e}^{-m/T}, \tag{2.9}$$

where g_j is its spin degeneracy and T is the temperature. Using this partition function, it's possible to calculate, in events with fixed numbers of heavy quarks, the average multiplicities:

$$\left\langle n_j \right\rangle = z_j \lambda_j \frac{Z\left(\nu_c - \nu_{cj}, \nu_{\bar{c}} - \nu_{\bar{c}j}, \nu_b - \nu_{bj}, \nu_{\bar{b}} - \nu_{\bar{b}j}\right)}{Z\left(\nu_c, \nu_{\bar{c}}, \nu_b, \nu_{\bar{b}}\right)}. \tag{2.10}$$

To perform this calculation it is necessary to consider some numerical and physical approximations that can be found in the reference [20], where the calculations are developed in more detail. Then, the multiplicity is given by:

$$\left\langle\left\langle n_j \right\rangle\right\rangle = z_j \lambda_j \prod_{f=c,\bar{c},b,\bar{b}} \left(\frac{\left\langle \nu_f \right\rangle}{a_{f1}}\right)^{\nu_{fj}} \equiv z_j \lambda_j \prod_{f=c,\bar{c},b,\bar{b}} \eta_f^{\nu_{fj}}, \tag{2.11}$$

where a_{f1} is the sum of $z_j \lambda_j$ for hadrons with one unit of open flavour $f = c, b$. The Eq. 2.11 can be applied to estimate the average multiplicities of multiply heavy flavored hadrons in different experiments and then used to compare with the multiplicities in a $p - p$ collision. Regarding this, Fig. 2.7 shows production yields for the single parton scattering (SPS) expectation at $\sqrt{s} = 14$ TeV for mass number equal to one : A=1 (proton) [21] [22]. For the SHMc predictions on the ratio between the yields of multi-charm baryons (Ξ_{cc}^{++} (ccu), Ω_{cc}^{+} (ccs), Ω_{ccc}^{++} (ccc)) and that of the single-charm, Λ_c (udc), at $\sqrt{s_{NN}} = 5.02$ TeV for $A \neq 1$ [23] , as a function of mass number of the ions in the collision. It is possible to observe that for Ξ_{cc}^{++} there is an increase of 10^2, if compared to the value for the p-p collision, in the production yield, while there is a dramatic increase of 10^3 for the Ω_{ccc}^{++}.

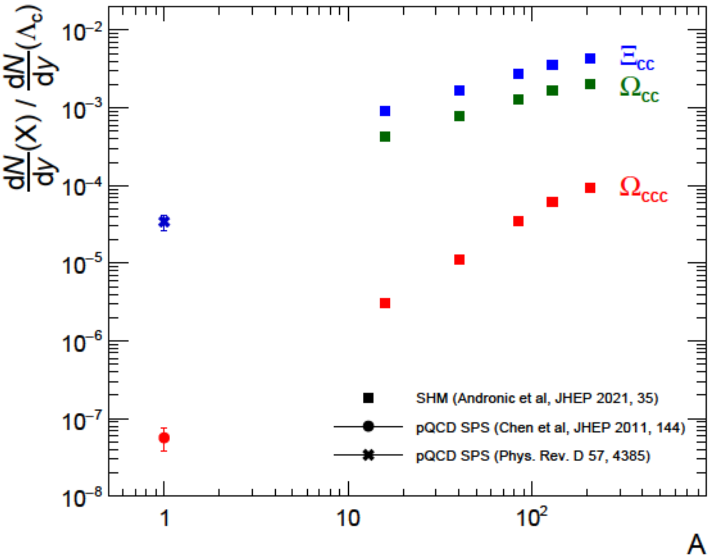


Figure 2.7: Production yields for the single parton scattering (SPS) expectation at $\sqrt{s} = 14$ TeV [21] [22] and for the SHMc predictions for the ratio between the yields of multi-charm baryons and that of the single-charm, Λ_c, at $\sqrt{s_{NN}} = 5.02$ TeV for $A \neq 1$ [23] , as a function of mass number. Figure from [24].

For the measurement of these multi-charmed baryons, it is necessary for a detector to have unprecedented high pointing and mass resolution, particle identification over a large transverse momentum range, an ultra-low material thickness, and detection layers close enough to the interaction region, with the tracking layers spaced very closely to allow the measurement of the particles before they decay. To investigate these and other properties of QGP, a new detector is being planned, called A Large Ion Collider Experiment 3 (ALICE 3) [24], which is associated with the novel experimental technique “Strangeness Tracking” that will be able to measure these production yields and provide a clear indication of QGP formation.

Chapter 3

A Large Ion Collider Experiment 3

A **L**arge **I**on **C**ollider **E**xperiment (ALICE) [2] is one of eight experiments at the Large Hadron Collider (LHC) [1]. Located in Switzerland, the LHC is the world's largest and most powerful particle accelerator ever built. It has a 27-kilometer ring of superconducting magnets, 175 meters below ground level. Inside the accelerator, there are two high-energy particle beams, using protons or nuclei of ionized atoms, that travel in opposite directions, in separate beam pipes, guided by a strong magnetic field that is maintained by superconducting electromagnets. Along the tunnel through which the particles collide, there are four main detectors: ATLAS [7], CMS [6], LHCb [25], and ALICE. These detectors collect data for various study purposes. In general, the LHC was constructed to test predictions of different theories of particle physics, including the Higgs mechanics and the unconfined state matter. In special, the ALICE experiment, which was first proposed as a central detector in 1993, in its first version had the goal to study head-on collisions between heavy nuclei at the top energy of LHC, 5.02 TeV per nucleon pair, as well as the physical properties of the strongly interacting matter at extreme energy densities, where the formation of a QGP phase is expected [2].

Despite the great progress made by the ALICE experiment, some questions are not yet possible to be answered with the current capabilities of the detector. In order to be able to tackle these problems, a novel detector is being proposed, the ALICE 3, that will own a high readout rate, unprecedented high pointing resolution, excellent tracking and particle identification over a large acceptance using advanced silicon detectors.

In this chapter, the main physics goals of ALICE 3 is presented. Notions about the basic kinematic variables used in high energy detectors, such as ALICE, and details about the new ALICE 3 detector will be presented, as well as their respective experimental aspects.

3.1 Introduction to the ALICE 3's Physics Goals

The primary science goals of ALICE 3 includes quantitative understanding of transport and hadronization of heavy flavors in the QGP medium. This study will be possible by performing, among other things, a measurement of an azimuthal correlation between charm and anti-charm mesons and the production of multi-charm baryons, like Ξ_{cc}^{++} and Ω_{ccc}^{++}. Another goal is the measurement of electromagnetic radiation from the QGP medium, which will be possible with a dilepton measurement below J/Ψ mass, down to zero p_T, to map the evolution of the collision. With the measurement of real and virtual photon emissions, it will be possible to determine the temperature of the evolution of the QGP as the flow of the particles. These measurements will allow for the study of the chiral symmetry restoration [24].

The following is a brief introduction of some of these goals.

3.1.1 QGP hadronisation and multi-charm hadrons

As mentioned in section 2.4, measuring the production yields of multi-heavy-flavor hadrons will provide essential details and unprecedented sensitivity about the production of heavy-flavor hadrons and hadron formation from a deconfined QGP. To perform measurements of multi-charm baryons, some requirements are needed, such as a very high tracking precision close to the interaction point, a large acceptance in order to study the dependence of the production of these particles on the variation of the heavy quark density with rapidity, precision particle identification over a wide transverse momentum range. The scope of this work is to study, using Monte Carlo simulation data, the possibility of measuring these particles in ALICE 3, especially the Ξ_{cc}^{++}.

3.1.2 Electromagnetic radiation

Real and virtual (dileptons) photons are emitted during all timeline of the evolution of the collision and do not interact strongly with the medium, not being affected by the high quarkonic density present in the QGP. In this way, these particles reflect the initial state that they were produced. Another important aspect is that there are different mechanisms that produce both photons and dileptons, which provides different information about the collision. These mechanisms are:

- Hard parton interactions in the early stage of the collisions. In this case, hard photons and Drell-Yan dileptons [1] are produced and with these particles it is possible to recover information about the primary/pre-equilibrium stage. In special, it is possible to recover the effective initial state parton distributions;

- Thermal photons and dileptons emission by the medium. As quarks and hadrons are interacting all the time during the collision evolution, photons and dileptons are constantly produced by the medium, with the rates increasing strongly with temperature. Thus, these particles provide information about the successive stages, from the formation of the QGP to the freeze-out process.

- Hadrons produced at any stage of the collision can decay and emit photons or dileptons. With these particles, it is possible to recover the information about the dense interacting hadronic matter and to probe in medium hadron modifications.

Hence, the study and measurement of the real and virtual photons produced during all phases of the collision will provide important information on the temperature and radial expansion velocity of QGP, that is, the radial flow velocity. Furthermore, with the invariant mass distribution of the dileptons, it will be possible to obtain a Lorentz-invariant measure of the temperature that is unaffected by the collective radial expansion of the medium [26], [27], [28].

To perform these measurements, ALICE 3 proposes a larger sampled luminosity, that is, to increase the measurement of how many collisions are occurring in the detector. Operating at high luminosity means increasing the chance to measure very rare events. Furthermore, a large acceptance and a very thin and light tracker will also be needed to have a high photon conversion [2] tracking to minimize the background from these particles.

3.1.3 Chiral symmetry restauration

It is known that most of the mass of ordinary matter is generated by the spontaneous breaking of the chiral symmetry associated with the condensate of quarks and gluons, and a negligible portion of this mass is generated by the Higgs field. Calculations using lattice QCD show that the temperature at which the phase transition from confined to deconfined

[1] Drell-Yan process is the creation of a virtual photon by the combination of quarks from the collision between two hadrons, this virtual photon will later decay into a lepton - anti lepton pair.

[2] Photon conversion is the creation of e^+e^- pair from a photon. These photons can be measured via the tracking reconstruction of these particles in the detector.

matter occurs is close to the temperature of the chiral phase transition. Because the transition temperatures are close, it is possible to use QGP phenomena to investigate the chiral symmetry restoration, since with this restoration it is expected to change the properties of hadrons as the temperature of the medium approaches the restoration point. A strong candidate for such studies is the ρ meson, with a lifetime around 4.5×10^{-24} s [29], which would have its mass reduced considerably, indicating, in this way, the restoration of chiral symmetry [30], [31].

3.2 ALICE coordinate system and relevant variables

Before discussing the ALICE 3 experiment and its detection system, it is necessary to briefly introduce the kinematic variables used to represent the results collected in this experiment.

Initially, the coordinate axis used is defined as a right-handed orthogonal cartesian system with the origin at the beam's interaction point, which can be seen schematically in Fig. 3.1. The $\hat{z}$ direction is defined to be along the direction of the incident beam of the accelerator: in ALICE, the Muon Spectrometer is at negative $\hat{z}$. The $\hat{x}$ direction is aligned with the local horizon and points to the centripetal direction of the LHC ring, thus, the $\hat{x}$ is perpendicular to the mean beam direction. The $\hat{y}$ direction is perpendicular, following the right-hand rule, to the $\hat{x}$ and $\hat{z}$ directions. The coordinate φ is the azimuthal angle with $0 < \varphi < 2\pi$ and θ is the polar angle with $0 < \theta < \pi$. The $\hat{x}$ and $\hat{y}$ axis define the reaction plan.

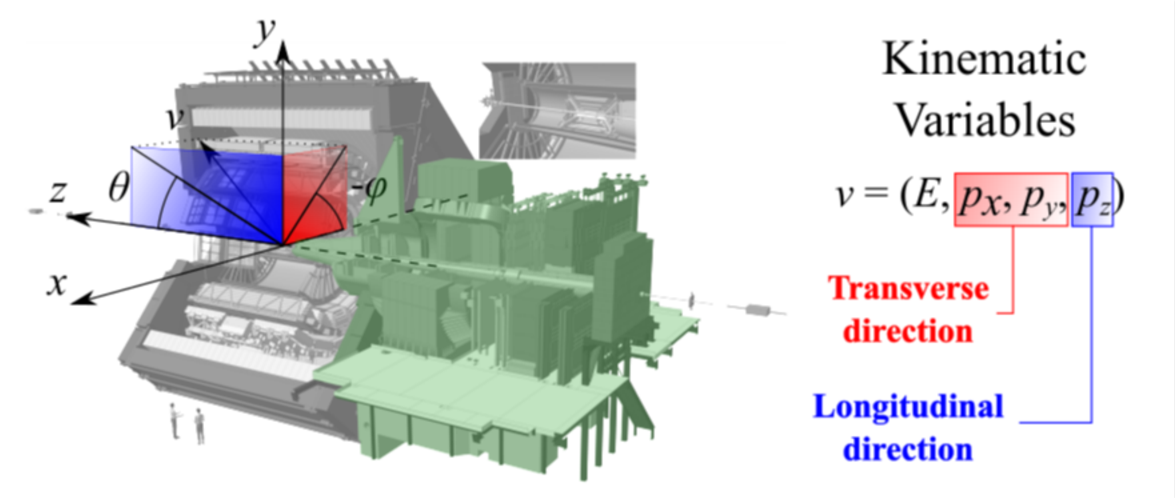


Figure 3.1: Coordinate system and relevant variables used to describe events in ALICE. The Muon Spectrometer is highlighted in green. Adapted from [32][33].

Once the coordinate system is defined, the kinematic variables of interest can be presented. These variables can then be divided with respect to the transversal and longitudinal direction of a collision. In the transverse frame, the transverse momenta of a particle is defined as follows:

$$p_T = \sqrt{p_x^2 + p_y^2}, \tag{3.1}$$

where p_x and p_y are the momentum components in the transverse momentum plane. In High Energy collisions, it is essential to measure the momentum of all particles produced in the collisions. With the moment along the beamline, only the beam particles can be observed. In contrast, the transverse momentum can be associated with any other particle produced at the vertex interaction. Another characteristic of this variable is that the transverse momentum is invariant under Lorentz transformation.

In the longitudinal direction, the main geometric variable used is the pseudorapidity η, which is important to characterize particles whose unique available information is the angle θ between its momentum and the beam axis. It is possible to obtain the pseudorapidity for any charged particle that has been detected by the experiment. The pseudorapidity η is given by:

$$\eta = \frac{1}{2}\ln\left(\frac{|\vec{p}| + p_z}{|\vec{p}| - p_z}\right) = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right], \tag{3.2}$$

where $\vec{p}$ is the momentum vector for a given particle, p_z is the particle's momentum component in the longitudinal direction $\hat{z}$, and θ is the angle between the particle's momentum and the beam axis, as indicated in Fig. 3.1.

Let $E = \sqrt{m^2 + |\vec{p}|^2}$ be the total particle energy. Another useful variable used to describe the kinematic conditions of a particle and specially to measure the relativistic velocity is the rapidity y:

$$y = \frac{1}{2}\ln\left(\frac{E + p_z}{E - p_z}\right) = arctanh\left(\frac{p_z}{E}\right) = arctanh(v_z), \tag{3.3}$$

where v_z is the longitudinal component of the velocity. The rapidity y can only be measured for identified particles, i.e, particles with known mass. In the nonrelativistic limit, the rapidity of a particle traveling in the longitudinal direction is equal to the pseudorapidity and equal to the velocity of the particle in units of the speed of light.

With the notions of transverse momentum and rapidity, it is possible to define yet another quantity, the transverse momentum distribution of charged particles:

$$\frac{d^2N}{dp_t dy} = f(p_t), \tag{3.4}$$

where N is the total particle rate production. It is commonly used the density production in the central rapidity region, dN/dy. The $d^2N/dydp_T$ spectra allows us to calculate the total production of a given particle in a certain rapidity range. The function $f(p_T)$ refers to:

$$f(p_t) = \frac{\int_{y_{\text{inf}}}^{y_{\text{sup}}} \frac{d^2N}{dp_t dy} dy}{\int_{y_{\text{inf}}}^{y_{\text{sup}}} dy} \approx \frac{1}{\Delta y} \left(\left. \frac{dN}{dp_t} \right|_{y \approx 0} \right), \tag{3.5}$$

where y_{inf} and y_{sup} are the rapidity limits of the analyzed particles and $\Delta y = y_{sup} - y_{inf}$, which is chosen to be small enough so that dN/dp_t is constant within the rapidity range considered.

3.3 Centrality

In colliding beam experiments, collisions can be initially described and classified according to their impact parameter ($\vec{b}$). This quantity is a vector that connects the center of the colliding nuclei in the transverse plane to the beam axis. The magnitude of the impact parameter is correlated with the size of the overlap region of the nuclei, and with it is possible to determine the size and shape of the resulting medium. The overlap region corresponds only to the number of nucleons participants, i.e., particles that will undergo at least one binary collision, with the spectators being particles that do not participate in the collision and, therefore, in the overlap region. A schematic collision between two nuclei showing the impact parameter and the respective spectators and participants are shown in Fig. 3.2.

The impact parameter allows the calculation of some quantities, with centrality among them. This variable is a geometric factor that defines the fraction of participant nucleons that will be included in the collision and is usually expressed as a percentage of the total nuclear interaction cross section σ:

$$c(b) = \frac{\int_0^b \frac{d\sigma}{db'} db'}{\int_0^\infty \frac{d\sigma}{db'} db'}. \tag{3.6}$$

The full centrality range is 0 to 1. Peripheral collisions have centrality closer to 1 (or 100%) while the most central events have centrality close to 0.

Since the impact parameter is a quantity that cannot be measured directly due to the order of the colliding ions (10^{-15}m), the centrality is estimated experimentally using a signal distribu-

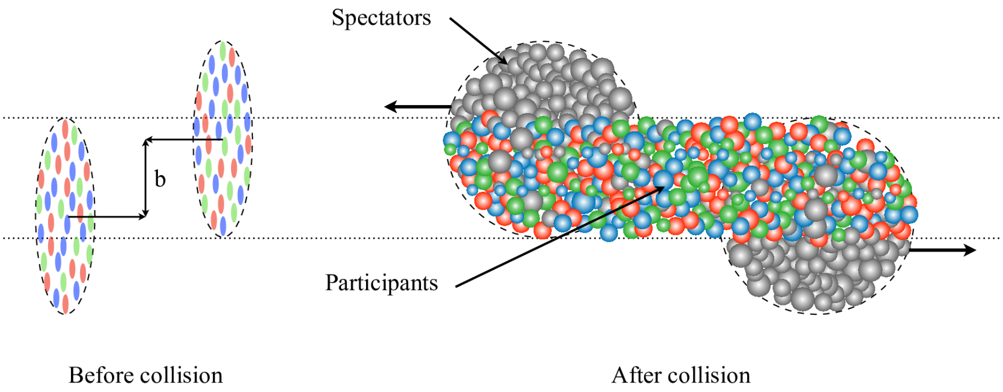


Figure 3.2: Schematic collision between two nuclei showing the impact parameter $\vec{b}$ and the respective spectators and participantes. Figure from [34].

tion of some detection system. With this, the measurements in terms of this variable are then divided into classes of centrality, which are in turn related to an impact parameter interval. For this, Monte Carlo techniques are used, i.e, an approximation is made geometrically using the Glauber model [35], which treats nuclear collisions as a superposition of binary nucleon-nucleon interactions [36]. The distribution of amplitudes in the VZERO scintillators (V0M) is shown in Fig. 3.3. This is the most common detecting system for centrality determination in the ALICE experiment, which is then fitted with the Glauber model combined with negative binomial distribution and divided into centrality classes.

When performing analysis with heavy-ion collisions, centrality is displayed graphically in centrality intervals that have a width of 10, like 0 – 10%. When referring to the 10% centrality class it is meant the interval including all events from 0% to 10%.

3.4 ALICE 3 Detectors and Systems

The ALICE experiment, which was in operation until LHC Run 2 (2015-2018), relies on a series of central detectors with $|\eta| < 0.9$, which consists of an ITS (Inner Tracking System), which is a set of layers of silicon detectors; a TPC (Time Projection Chamber), that allows the identification of particles by measuring the energy loss when they pass through the active gas; and a TRD (Transition Radiation Detector) which uses the γ-dependent threshold of

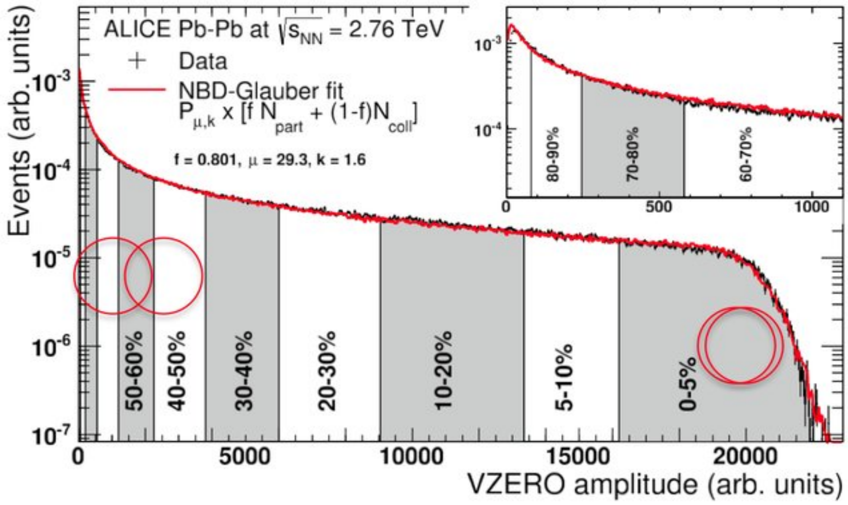


Figure 3.3: Graph with the distribution of the sum of amplitudes in the VZERO scintillators (V0M) from ALICE experiment, the fit using NBD-Glauber model and the divided centrality classes. Figure from [36].

transition radiation to distinguish between electrons and protons. With this configuration, ALICE operates at a luminosity of $\approx 5 \times 10^{30}$ cm^{-2} s^{-1}. These detectors can be employed when immersed in a 0.5 Tesla solenoid magnetic field parallel to its axis, which is responsible for particle drift, improving tracking resolution. This configuration, especially due to the presence of the TPC and ITS, makes the current detector relatively slow, despite having an excellent spatial resolution [33].

For Run 3 (2022-2025), further improvements were made to the experiment, thus designing ALICE 2. The following upgrades were made in the central detectors [37]: The diameter of the beam pipe was reduced and replaced to be closer to the center of the experiment; for ITS the first detector layer is closest to the collision point. For ALICE, the very first layer had a radius of 22.4 mm with respect to the beam pipe, while this radius was reduced to 18 mm [38] for ALICE 2. To further improve the measurement precision, this detector will also make use of the technology Monolithic Active Pixel Sensors (MAPS). With these changes, ALICE 2 will increase the track reconstruction and readout capabilities to achieve all the Pb-Pb interactions besides enhancing particle identification capabilities. ALICE 2 is expect to collect 1 nb^{-1} of Pb-Pb collisions at a peak luminosity of $\approx 5 \times 10^{27}$ cm^{-2} s^{-1}.

For Runs 5 and 6, the planned experiment is ALICE 3 (see Fig. 3.4). It is intended to be constructed in such a way to be a nearly massless barrel detector consisting of truly cylindrical layers of ultra-thin silicon sensors using MAPS technology. For this, one of the main changes between ALICE 2 and ALICE 3 is the removal of the TPC, because, as mentioned before, it is a relatively slow detector, despite generating an excellent resolution. To increase the resolution of the measurements, the magnetic field that the detectors were immersed in will be replaced by a new one of 2.0 Tesla provided by a superconducting magnet system, which in turn will produce a curvature in the particle's momentum. In other words, all ALICE 2 detectors will be removed from the cave that hosts it, and ALICE 3 will be built in the place. Fig. 3.4 shows the ALICE 3 detector concept, with emphasis on the superconducting magnet system and the tracker detector.

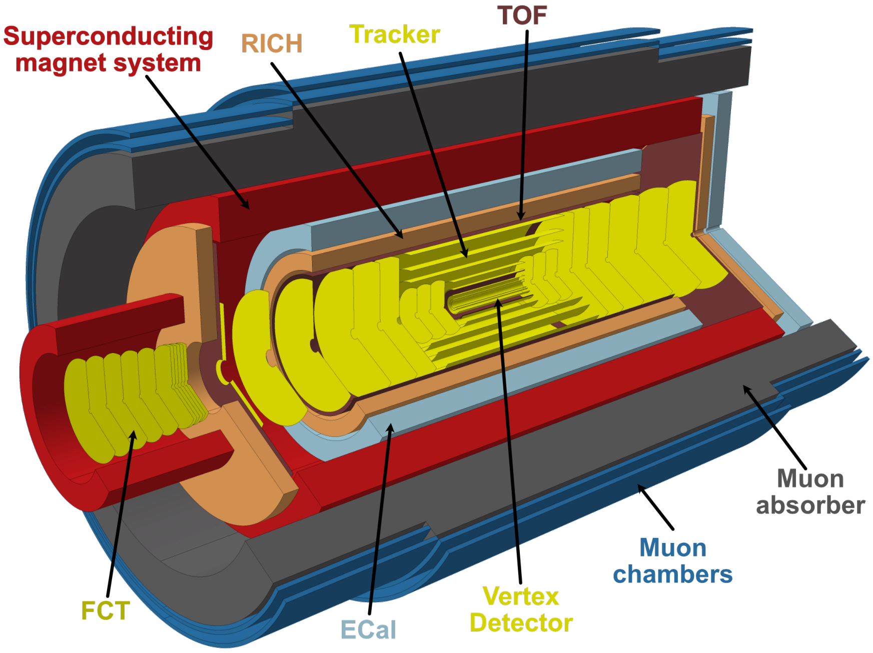


Figure 3.4: Overview of the ALICE 3 detector concept: the superconducting magnet system, a silicion tracker, the vertex tracker, a time-of-flight (TOF) detector, RICH detector, photon detector and a muon system detectors. Fig. from [24].

ALICE 3 will work with a peak luminosity of $\approx 3.3 \times 10^{27}$ cm^{-2} s^{-1} at interaction rate of $R_{max} = 93$ kHz [24]. Fig. 3.5 shows the improvement of pointing resolution [3] and effective statistics when compared to two other experiments, using the acceptance $(\Delta\eta) \times$ (Pb-Pb interaction rate (kHz)).

The following section contains a brief description of the system of ALICE 3 that is important to the tracking reconstruction.

[3] Pointing Resolution is the precision with which the position of a particle's primary vertex can be determined by the hits recorded in the detector.

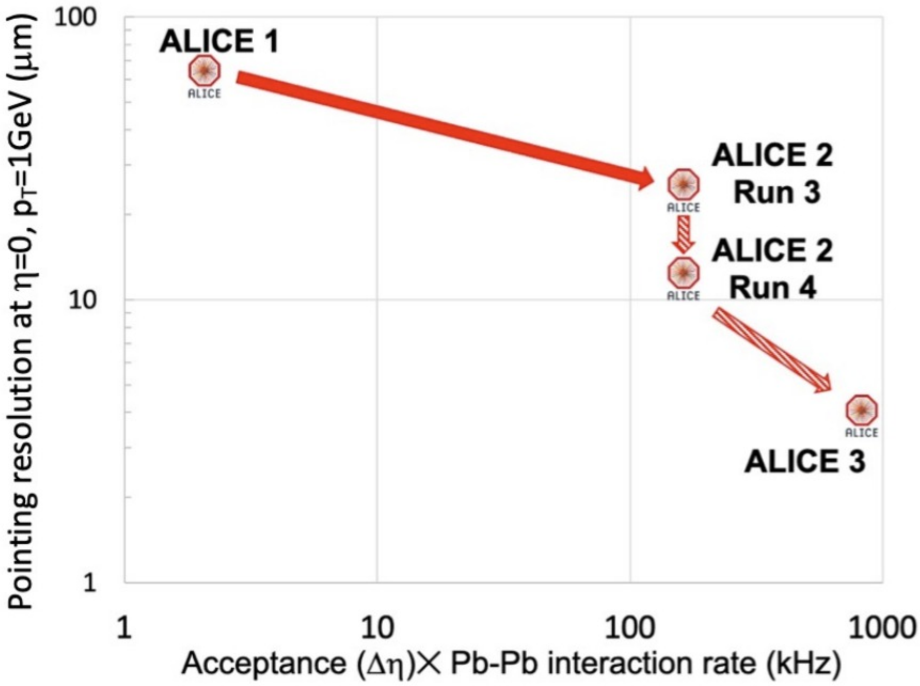


Figure 3.5: Comparation between the threes experiments: ALICE, ALICE 2 and ALICE 3 in the pointing resolution and the acceptance $(\Delta\eta) \times$ (Pb-Pb interaction rate (kHz)) . Figure from [39].

3.4.1 Detector concept

The Vertex Detector and the Outer Tracker are the most interesting detectors for particle tracking, see Fig. 3.6. These two detectors will consist of 11 barrel layers and 2×12 forward discs, which allows the possibility to cover the pseudo-rapidity interval of $|\eta| < 4$ [24]. The Vertex Detector consists of the first 3 layers and the 2×3 discs that will be retractable and installed inside a secondary vacuum. The reason for this is that in order to be as close to the interaction point as possible, the detector must be mounted so that it can be retracted when the LHC is operating in the injection stage (at least $R_{min} = 16$mm) and placed closest as possible from the interaction point when in the data collection phase. The radial distance of these first layers will be 5mm, 12mm, and 25mm from the interaction point. These layers will be responsible for measuring the first particle hits. The number of layers and their positions have been chosen to deal with fake hits. Another property of this detector that will guarantee a good reconstruction of the particle tracking is the resulting position resolution of $2.5\ \mu m$, which gives a very high intrinsic spatial resolution. For this, the technology used will be the Monolithic Active Pixel Sensors (MAPS). The MAPS consists of a pixel chip, which is made of a silicon single die with a size of 15 mm x 30 mm and which incorporates an epitaxial layer of high resistivity silicon (sensor active volume); a matrix of charge collection diodes that will perform the collection of the loads (pixels); and an electronics responsible for performing signal amplification, performing the digitization and zero-suppression.

The Outer Tracker will consist of the remaining 8 barrel layers and 9 discs on either side of the interaction point. With this, it will provide a relative momentum resolution of 1-2% over a large acceptance by measuring about 10 space points.

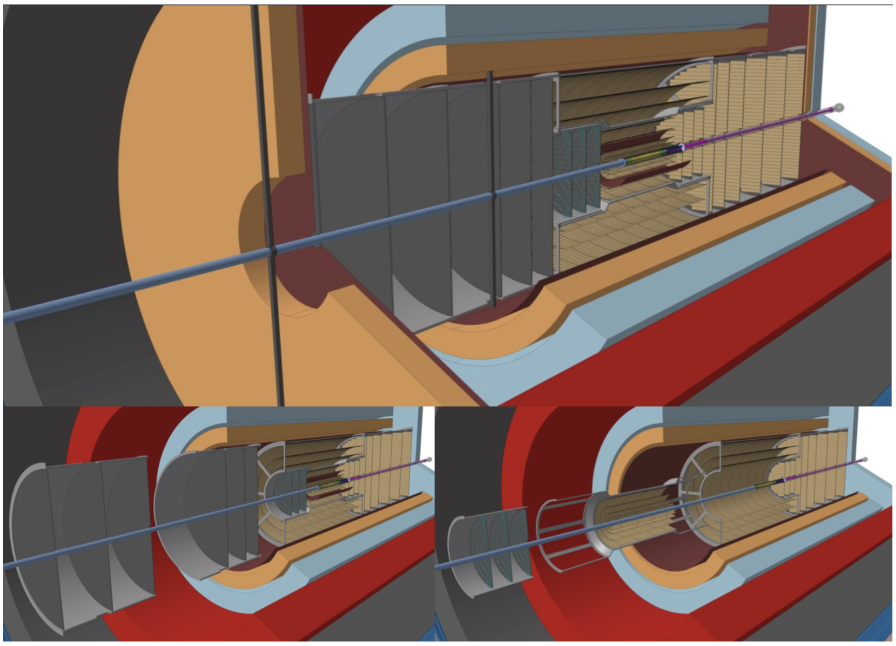


Figure 3.6: Overview of the Vertex (near to the beam pipe) and Outer Tracker detector assembly. Figure from [24].

Chapter 4

Topological Reconstruction & the Strangeness Tracking Technique

With the objective of measuring multi-charm hadrons on the ALICE 3 experiment, it is necessary to study efficient techniques of track reconstruction, which are commonly used in high energy physics for extremely rare particles. In particular, this work will focus on the candidate selection of the baryon Ξ_{cc}^{++}.

The data collected in experiments such as ALICE are massive and it is necessary to perform a complete reconstruction of the events, with an extensive data analysis performed on those particles. For this, the data analysis tool ROOT [40] was the software employed in this work, which was developed by CERN and is highly applied to variate analysis tasks in high energy physics. This framework is object-oriented with a code base written entirely in C++.

The process of reconstructing events through topological variables is referred to as Topological Reconstruction, which despite being a highly used technique in particle physics, it is not sufficient to cover extremely rare particles as the Ξ_{cc}^{++}. For this, a technique for tracking weakly decaying strange particles is being developed, called Strangeness Tracking, which will act as the second stage that follows after secondary track finding. In other words, it is a technique that makes use of topological reconstruction to increase the accuracy of the reconstruction of events.

This chapter aims to characterize the doubly charmed baryon Ξ_{cc}^{++} in Pb-Pb collisions at $\sqrt{s_{NN}} = 5.52$ TeV. It introduces the usual technique of event reconstruction, the topological reconstruction, as well as the strangeness tracking. It will also address how the standard selection of candidates is made, the extraction of signal, and the concept of the metric “statistical significance", which is used in this work to verify how successful the selection was.

4.1 Ξ_{cc}^{++}: The Doubly Charmed Baryon

The Ξ_{cc}^{++} particle is composed of two heavy charm quark and one up quark. Hence, it is a doubly charm baryon that is part of the family of weak decay qqq with the charm quantum number C = 2. This isospin doublet ($\Xi_{cc}^{++} = ccu$) have spin parity equal to $J^P = 1/2^+$ and a lifetime, measured by the LHCb collaboration, equal to $\tau\,(\Xi_{cc}^{++}) = 0.256^{+0.024}_{-0.022}$(stat) ± 0.014(syst) ps [41]. Due to the presence of two charm quarks, Ξ_{cc}^{++} is a massive baryon, with mass equal to 3621.24 ± 0.65(stat) ± 0.31 (syst) MeV/c^2 [42]. This particle have the following expected weakly decay channels, relative to the particle Ξ_c^+:

$$\Xi_{cc}^{++} \longrightarrow \Xi_c^+ + \pi^+ \quad (\,\mathrm{c}\tau \approx 77\,\mu m) \tag{4.1}$$

$$\Xi_c^+ \longrightarrow \Xi^- + 2\pi^+ \quad (\,\mathrm{c}\tau \approx 132\,\mu m) \tag{4.2}$$

and,

$$\Xi_{cc}^{++} \longrightarrow \Xi_c^+ + \pi^+ \quad (\,\mathrm{c}\tau \approx 77\,\mu m) \tag{4.3}$$

$$\Xi_c^+ \longrightarrow \pi^+ + K^- + p \quad (\,\mathrm{c}\tau \approx 132\,\mu m) \tag{4.4}$$

where $c\tau$ is the flight distance. This is the average distance, in the laboratory frame, traveled by a particle whose speed is approximately equal to the speed of light.

This work will focus on the decay defined by the Eqs. 4.1 and 4.2, for which the complete decay of the channel can be seen in the Fig. 4.1. To this channel, there are six prong decay [4] with a total Branching Ratio of 5% × 2.9% [5].

The point of interaction from which the particles are produced, immediately after the heavy nuclei collision, is called the primary vertex. All particles that come from this vertex, that is, the ones that are direct products of the collision, are called primary particles. The baryon Ξ_{cc}^{++} is one of them. Particles that are products of decays are called secondary particles.

[4] Prong decay refers to the number of charged particles in the final state of the particle decay.

[5] Branching Ratio is the probability of a particle to decay by a given process (channel) among all possible decay processes.

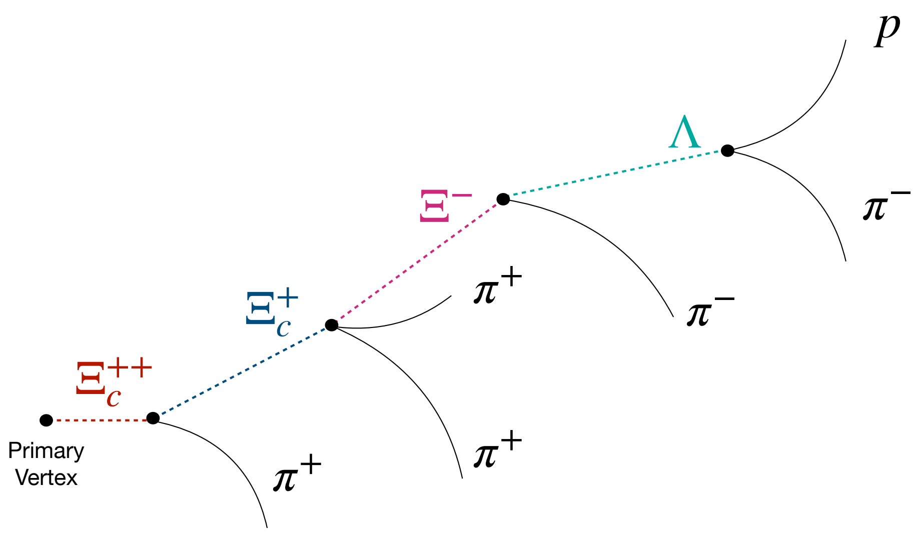


Figure 4.1: Cascade topology of Ξ_{cc}^{++} particle decay diagram for the following channel: $\Xi_{cc}^{++} \longrightarrow \pi^+\Xi^-\pi^+\pi^+$.

However, excluding all decay products to find the primary particles is experimentally infeasible. The ALICE collaboration has the convention that the primary particles are all those that originated in the high energy collision and those that are products of strong decays or decay of quarks c,b, and t, and the secondary particles are those that are products of weak decays of s quarks or that have hadronic interactions within the detector material. Because of its decay into successive secondary particles, Ξ_{cc}^{++} has a characteristic cascade topology. Both Ξ_{cc}^{++} and Ξ_c^+ have very short flight distances, which, even for ALICE 3, ends up being unfeasible for direct detection. Fig. 4.2 shows the relationship between the decay of the Ξ_{cc}^{++} with the radius layers of the ALICE 3 detectors.

The final particles that are detected leave hits in the detectors allowing for the trajectories to be obtained. With this information, it is then necessary to combine them to generate a potential cascade decay, called (cascade) candidate. Therefore, the final detected tracks (daughter particles) are used to reconstruct the decay of the particles of interest, with these named "mother particles" and being responsible for generating the former via weak interactions. The study of the descendants to reconstruct the tracks that correspond to a given decay channel is called *tracking*. As there is an immensity of particles produced in a collision, it is impossible to combine all possible particles to reconstruct a given decay. What is usually done is the application of different types of filters on the possible candidates to reject candidates that

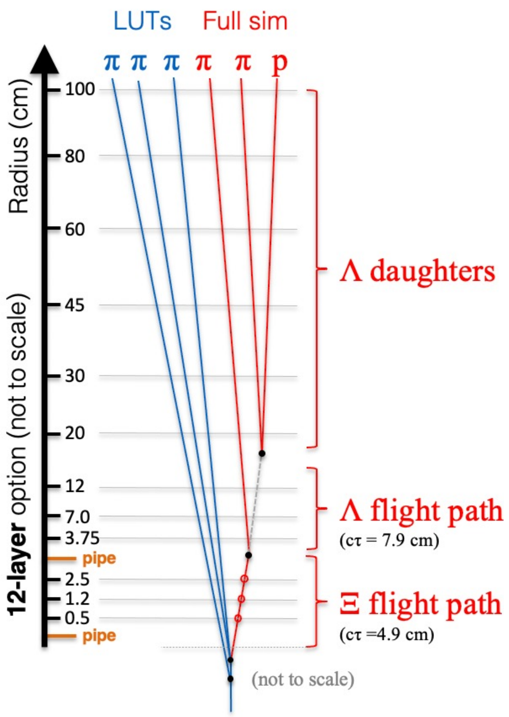


Figure 4.2: Scheme relating the decay of Ξ_{cc}^{++} to the scale of the ALICE 3 detector. Note that Ξ_{cc}^{++} and Ξ_{c}^{+} do not appear because they decay before the first layers.

are not consistent with the desired decay topology. One of these filters is the application of topological cuts, which will be discussed in the next section.

4.2 Candidate Selection Criteria

The combination of trajectories to reconstruct the decay is used to recover the mass of the original particle. In this way, the combination of trajectories to search for the cascade decay of Ξ_{cc}^{++} is done as follows: first, candidates to Λ are found performing proton-pion combinations; then these candidates for Λ are combined with one more trajectory of a pion to find a candidate for Ξ^-. These candidates for Ξ^- are combined with the trajectory of two other pions to obtain candidates for Ξ_c^+, which finally has its trajectory combined with that of a pion to obtain candidates for Ξ_{cc}^{++}.

Since the four-momentum modulus is invariant under Lorentz transformations, it is possible to consider its value in the original particle's rest frame, and hence this quantity has a modulus equal to the original particle's mass. Also, the four-momentum is conserved during the decay processes. Thus, if the modulus is computed for the decay products, and if these products indeed originated from a cascade decay, then it is possible to recover the original particle's mass. The recovered mass is called *invariant mass* and is given by:

$$M^2 = \left(\sum_i E_i\right)^2 - \left\|\sum_i \vec{p}_i\right\|^2 . \tag{4.5}$$

where the sum is over the possible decay products of the interested particle. In practice, the invariant mass calculation for the decay candidates provides a peak centered on a mean mass value. This peak, due to the high background, can be imperceptible, so it is necessary to apply cuts, that is, a candidate selection is applied to decrease the contaminations from the background without significantly affecting the signal. This selection can refer, for example, to the charges of the daughter particles, the energy deposited in the detector, the proper lifetime $c\tau$, or topological variables. It is the latter that will be discussed in this work.

If a chosen combination in the reconstruction is indeed from the decay of Ξ_{cc}^{++}, the invariant mass, calculated using the Eq. 4.5, should be close to 3621.24 MeV/c^2, with small variations arising from the analysis as well as from imprecision in the measurements of the moments.

4.2.1 Topological selection of cascade candidates

Trajectory reconstruction involves a huge number of combinations of particles and candidates since an immensity of particles is produced in Pb-Pb collisions at energies of $\sqrt{s_{NN}} = 5.52$ TeV, whether these combinations are to reconstruct the daughter particles or the original particle. Therefore, topological selections are employed to filter the tracks used in the baryonic reconstruction to reduce the number of possible combinations. This selection makes use of geometric variables that refer to properties of the topology of each particle. The determination of the intervals on these variables is used to select trajectories that have characteristics that make them more likely to have come from the decay of interest.

In Fig. 4.3, it is possible to see an example of these topological variables for the decay of Ξ^-. The same variables could be applied to the Ξ_{cc}^{++} decay. Following this, there is a brief description of each type of these variables.

- **DCA to PV**: It is the Distance of Closest Approach (DCA) from a trajectory to the Primary Vertex (PV), also know as the impact parameter of the track. It can be obtained

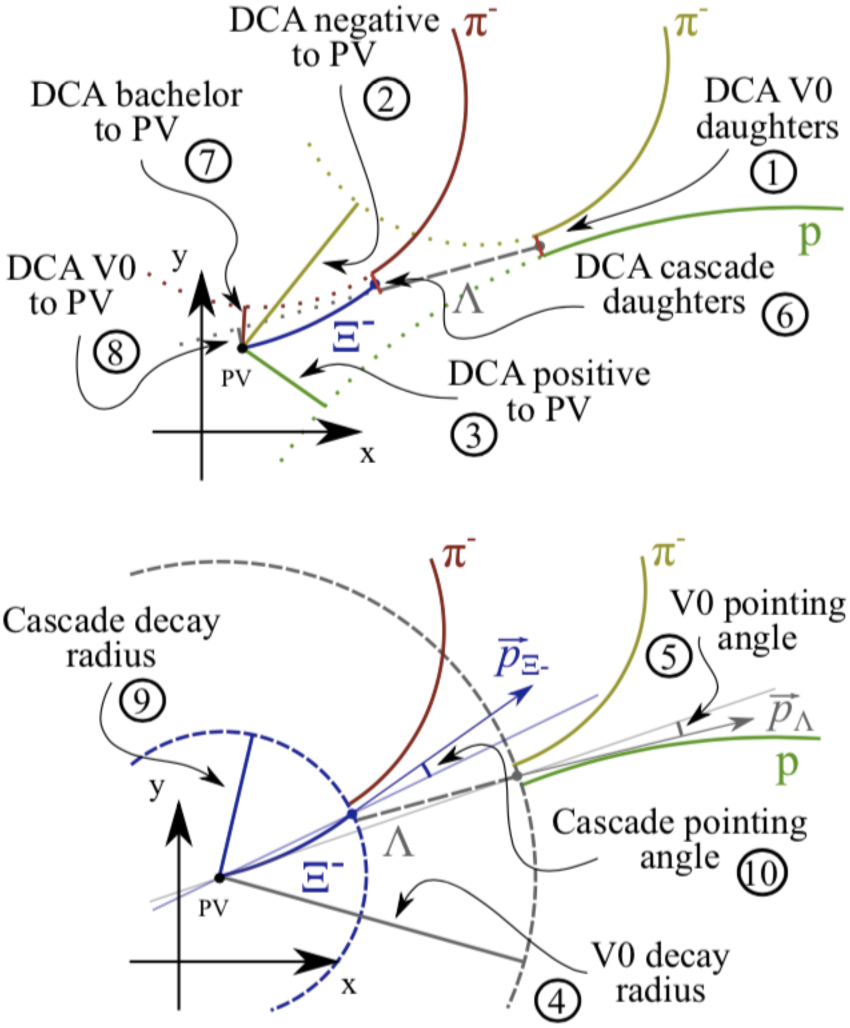


Figure 4.3: Example of topological cuts used to combining particle trajectories in candidate decays of Ξ^-. Figure from [32].

in both longitudinal and transverse directions:

$$|DCA| = \sqrt{DCA_{xy}^2 + DCA_z^2}. \tag{4.6}$$

The goal is to use a maximum value (upper bound) for primary particles and a not so small minimum value (lower bound) for secondary particles in order to be able to discriminate them.

- **DCA of the cascade daughters**: It is the Distance of Closest Approach (DCA) between the tracks of two possible daughters of the decay. It is used to ensure that the trajectories were close enough together at some point, presumably at the time of decay. Usually, an upper bound is defined to ensure that two particles come from the same vertex.

- **Decay radius**: It is the distance between the spatial position of the particle immediately before decay and the center of the detector.

As the data used in this work comes from Monte Carlo simulations, it is also possible to use topological information of the original particle itself.

With the application of these cuts into data, it is feasible to reconstruct the trajectories and select candidates for the Ξ_{cc}^{++}. Then, it is possible to graph the invariant mass and observe a peak in it. However, this is not a simple task, since it is necessary to choose a subset of variables among a great number of them and select the values of the cuts to be used with these variables. Therefore, this process demonstrates the real challenge of this work.

In addition, improving the resolution of these topological observables makes the cuts more accurate, thus eliminating more backgrounds from the candidates. To enable high-precision measurements of these variables, a new technique is being developed with the idea of finding signatures of multiply charmed baryons via their weak decays into strange baryons, called Strangeness Tracking.

4.3 The Strangeness Tracking Technique

The strangeness tracking technique is a new method that is under development to detect multiply charmed baryons using information from their decay into strange baryons. This method is based on using high-resolution detectors in the first layers, which are very close to the primary vertex, to measure the hits of weak decay hadrons before their decay and combine such information with the reconstruction of the daughters of these decays to improve significantly the pointing resolution. Strangeness tracking acts as a second stage after the topological reconstruction.

A particularity of this technique is that it has an intrinsic dependence on two factors: the proper lifetime of the particle and the detector performance. To perform the technique, it is necessary that the particles hits in the innermost detector layers before they decay. For example, in Fig. 4.4 there is a survival probability graph of Ξ^- and Ω^- baryons with a momentum equal to 1 GeV/c as a function of the distance to the primary vertex. It is possible to see that the Ξ^- has a decay length equal to $c\tau = 4.9$ cm, and the probability that this particle decays after crossing the first ALICE 3's detector layer and leaves a direct hit is around 90%, while in respect to the Ω^-, that has a decay length of $c\tau = 2.5$ cm, it is approximately 70%. Therefore, for these two cases, it is possible to apply strangeness tracking in the track reconstruction of these particles.

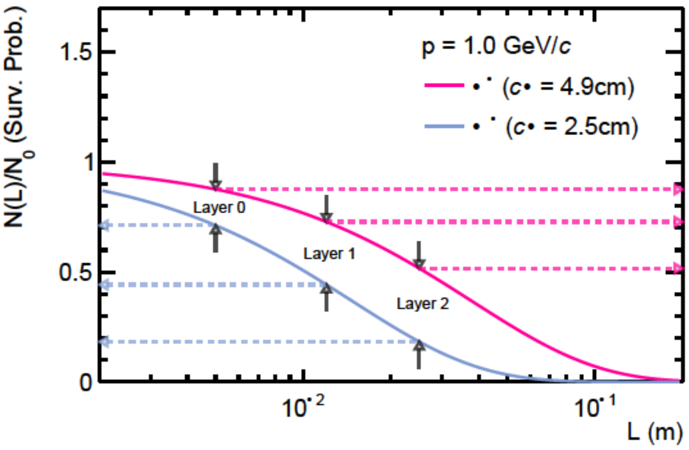
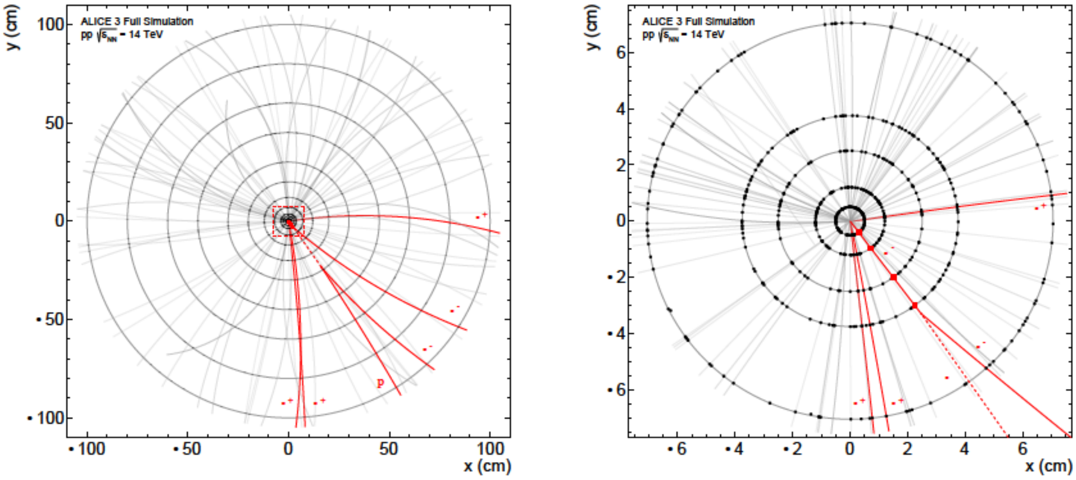


Figure 4.4: Survival probability of Ξ^- and Ω^-, with a momentum equal to 1 GeV/c, as a function of the distance to the primary vertex. The decay length of each particle and the innermost layers of ALICE 3 are indicated in the figure. Figure from [24].

The algorithm behind the technique is structured as follows: for each decay candidate, the trajectory information is calculated and used to extrapolate back to the primary vertex, as shown in Fig. 4.5 for Ξ_{cc}^{++} decay. This back propagation has a hit search window and the hits that are close enough to the propagated trajectory will be added to track parameterization to increase the precision of the decay candidate tracking reconstruction. If a hit that was added to the track parameterization turns out to be incorrect during the extrapolation, it can be deleted and the back extrapolation is updated, improving the resolution of the reconstruction. In this way, the back propagation represents a high constraining for the addition of the correct hits of the inner layers to the trajectory reconstruction [43][24].



(a)

(b)

Figure 4.5: Illustration of strangeness tracking in the ALICE 3 full detector simulation with the decay of Ξ_{cc}^{++} into $\Xi_{cc}^{++} \longrightarrow \Xi_c^+ + \pi^+$ and the decay of Ξ_c^+ into $\Xi_c^+ \longrightarrow \Xi^- + 2\pi^+$. In (a) it is possible to see the daughter decay in the last layers and in (b) is a close-up in the innermost layers showing the decay of Ξ^- and the hits that were added to its trajectory. Figure from [24].

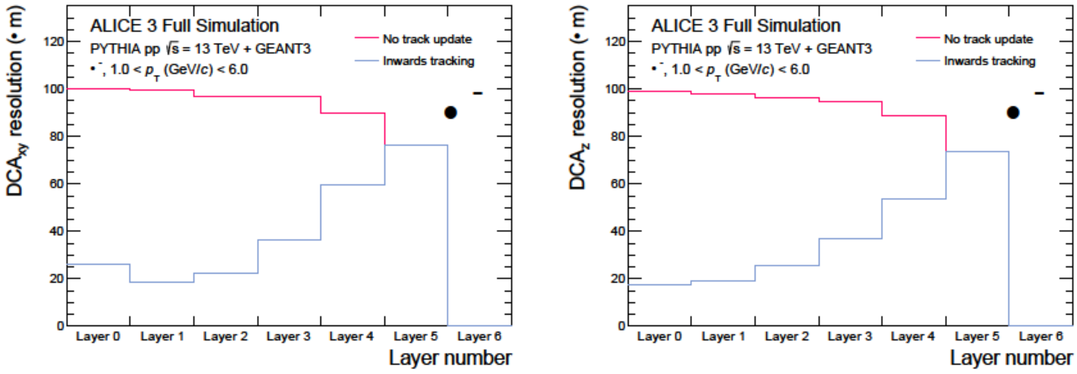
One of the concerns of the strangeness tracking technique is the association of fake hits to parameterization. The resolution in track determination must be high enough to avoid these fake hits. For this, the hit density in a given layer ρ_{hits} must obey the relation[43]:

$$\frac{1}{\rho_{hits}} >> \delta_{search}, \tag{4.7}$$

where δ_{search} is the search window. For ALICE 3 innermost layer in central Pb-Pb collisions, the expected average occupancy, using a full simulation, is approximately 10-12 hits/mm^2, which corresponds to around one hit for each 0.1 mm^2 and calculations reveal that the expected search window size of strangeness tracking δ_{search} is approximately $5 \cdot 10^{-3} mm^2$. Thus, the more hits left in the detector layers by interest candidates, the greater the pointing resolution will be.

The strangeness tracking provides a direct tracking of weakly decay hadrons and, with this, it is expected a great improvement of the pointing resolution. This improvement can be studied and verified by performing an investigation of the effect on the distribution of the

Distance of Closest Approach (DCA) of Ξ^- particle to the primary vertex. In Fig. 4.6 there is a comparison between topological reconstruction with and without strangeness tracking in (a) DCA_{xy} and (b) DCA_z resolutions in the transverse plane and longitudinal direction, respectively, for the Ξ^- decay in function of the layers of the ALICE 3 detector. In the magenta curve, there is a pure topological reconstruction, and it is possible to observe that the width of the residuals distribution deteriorates as the extrapolation distance increases (note that the distance of back propagation increases in the direction of the primary vertex). The blue curve represents the topological reconstruction with strangeness tracking, that is, including direct detection information. In this last case, it is possible to observe that the width of the residuals decreases.

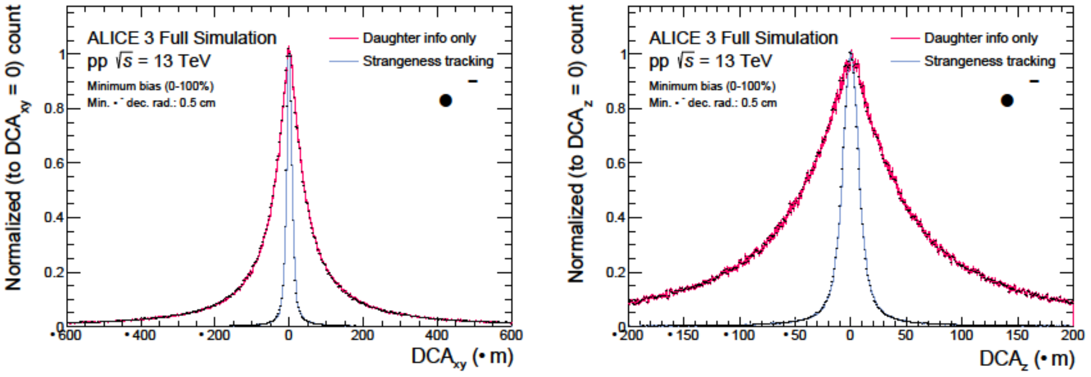


(a)

(b)

Figure 4.6: Comparison between Topological Reconstruction with (blue) and without (magenta) Strangeness Tracking in the (a) DCA_{xy} (b) DCA_z resolution in transverse plane and longitudinal direction, respectively, for the Ξ^- decay. Figure from [24].

In Fig. 4.7, there is the normalized counts for the DCA to primary vertex in (a) transverse plane and (b) longitudinal direction for Ξ^- candidates. The magenta curve shows the pure topological reconstruction, using only daughter information, while in the blue curve the strangeness tracking is included. The resulting improvement in DCA resolution from the use of strangeness tracking is approximately augmented by a factor of 4 in both transverse and longitudinal directions, which indicates an interesting potential to this method.



(a) (b)

Figure 4.7: Comparison between topological reconstruction with (blue) and without (magenta) strangeness tracking in the distributions of (a) DCA$_{xy}$ and (b) DCA$_z$ to Primary Vertex for Ξ^- . Figure from [24].

4.4 Signal Extraction

Once candidate selection has been applied, using the topological observables, it is then possible to observe a peak in the invariant mass, as shown in Fig. 4.8, where the background is quite regular and almost linear. With the invariant mass peak identified, the next step is to count the number of candidates at the peak. The analytical process for this count is done in the following manner: initially, a function constituted of a gaussian and a linear polynomial is fitted in the peak and background region of the invariant mass, respectively. After that, the gaussian parameters are used to define the peak and background regions:

- Background to the left:

$$(\bar{x} - 2L\sigma, \bar{x} - L\sigma), \tag{4.8}$$

- Background to the right:

$$(\bar{x} + L\sigma, \bar{x} + 2L\sigma), \tag{4.9}$$

- Peak region:

$$(\bar{x} - L\sigma, \bar{x} + L\sigma), \tag{4.10}$$

where σ is the standard deviation and $\bar{x}$ is the mean of the fitted gaussian, and L is a value conveniently chosen depending on the invariant mass distribution. This choice selects a region that represents the combinatorial background and should be sufficiently distant from the peak.

Once these regions are defined, the bin counting technique can be applied. Note that the two background sampling regions have together the same width as the peak region. One can subtract the count in these two regions adjacent to the peak from the number of candidates in the peak region. For each case, it is calculated the integral below the curve, which gives the counts for the background and the peak. The difference between these two counts gives the raw signal:

$$\text{Raw Signal} = N(\bar{x} - L\sigma, \bar{x} + L\sigma) - N(\bar{x} + L\sigma, \bar{x} + 2L\sigma) - N(\bar{x} - 2L\sigma, \bar{x} - L\sigma), \tag{4.11}$$

where $N(A, B)$ denotes the number of candidates inside de interval defined by $[A, B]$. This count is usually done considering the candidates in a specific transverse momentum interval (p_T) , such that it results in a set of candidate numbers indexed by to the defined p_T interval.

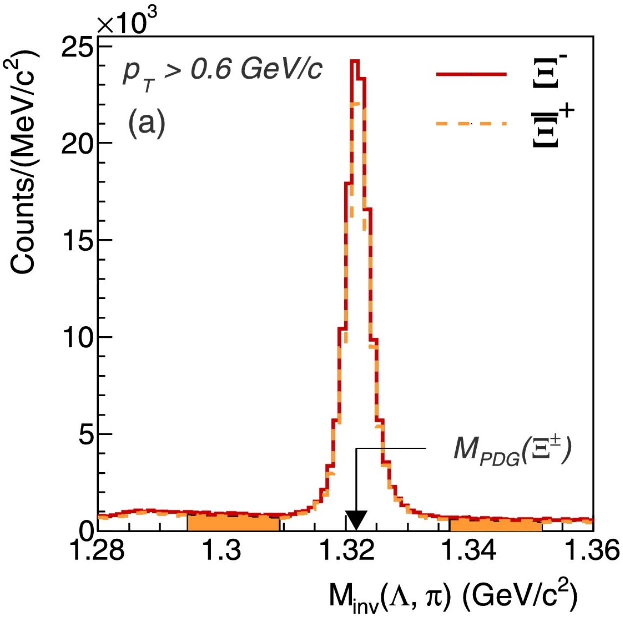


Figure 4.8: Demonstration example of the background sampling process used for signal extraction for the particle Ξ^-. The orange region is the average background estimated. Figure from [38].

4.5 Significance computation

The objective of this master's project is to study the possibility of measuring the particle Ξ_{cc}^{++} in the ALICE 3 experiment, and one way to verify this possibility is to use the significance calculation. As ALICE 3 is an experiment for the next decade only, this study is performed using Monte Carlo simulation.

In Monte Carlo simulations, the simulated candidates can be classified into two types: signal, which is the complete simulation of the final states corresponding to the reaction of interest, or background, which is all candidates that are not due to this reaction but have similar characteristics. These two types of candidates correspond to two hypotheses that must be distinguished from each other: the "signal hypothesis" H_1 and the "background hypothesis" H_0. The topological selection procedure that was described before is an hypothesis test that is applied to every single trigger[6] collected by the experiment [44].

A statistical hypothesis test is a quantity that allows quantifying the degree of confidence in a decision. When the statistical test is applied to the simulated data, a value t is calculated and compared to the test statistic data distribution following the H_1 and H_0 hypothesis, and then it is decided between H_1 and H_0. It is possible to quantify the Type-I and Type-II errors: the first refers to signal candidates that are discarded, that is, the H_1 hypothesis is declared false when it is true; and the second is about background candidates that contaminate the signal sample, in this case, H_1 is declared true when it is false. This process is shown in Fig. 4.9 where α is the Type-II error and β is the Type-I error. Notice that, in a hypothesis test, it is necessary to choose a critical t-value, or threshold, to either reject or accept the hypothesis.

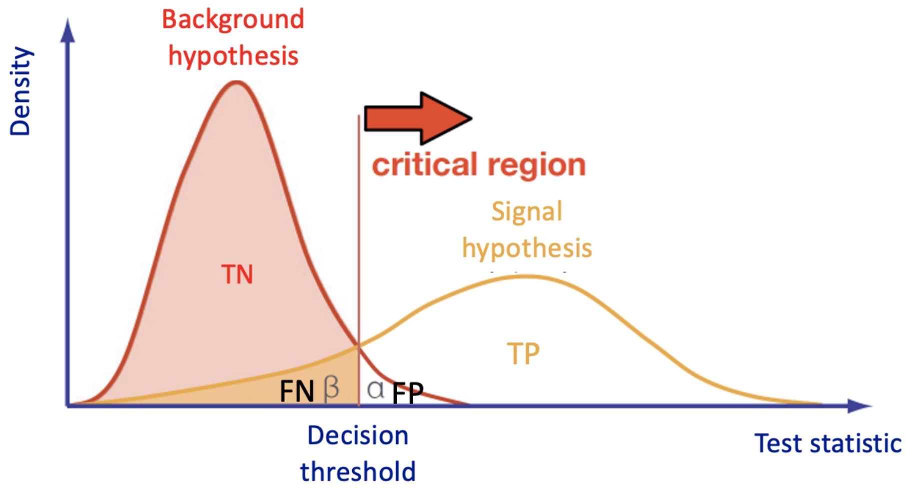


Figure 4.9: Density of signal and background test statistic data, where α is the Type-II error and β is the Type-I error. TP reefers to 'True Positive', TN to 'True Negative', FP to 'False Positive', and FN to 'False Negative'. Figure from [45].

[6] A trigger is an event that the logic of the experiment decides to retain for offline analysis. That is, is the data that is collected to pursue the analysis.

To optimize the selection procedure, it is necessary to use a score function that will define the optimum cuts applied to the data, such that the result contains the maximum possible signal in the candidate sample and the minimum as possible background contamination.

Let N be the number of candidates at the end of the selection, S the number of candidates classified as signal, obtained from the peak region using the bin counting, and B be the number of candidates classified as signal but that in reality are false-positives. With that, the best estimation of S is given by:

$$S = N - B. \tag{4.12}$$

Taking N to be a Poisson distribution, the variance of S is:

$$\sigma^2(S) = \sigma^2(N) + \sigma^2(B) = N + \sigma^2(B). \tag{4.13}$$

For large Monte Carlo statistics, the uncertainty $\sigma(B)$ is low and negligible. With this, and using Eq. 4.12, its possible to define the quantity Z:

$$Z = \frac{S}{\sigma(S)} = \frac{S}{\sqrt{N}} = \frac{S}{\sqrt{S+B}}, \tag{4.14}$$

which is given in number of standard deviations. This quantity receives the name of Significance: it is a non-dimensional number and should be as large as possible since it provides information on how well the signal can be observed in a number of standard deviations. Significance also indicates the accuracy of a measurement, that is, indicating if a particle can or not be measured.

Since in a real experiment it is not possible to know to values of S and B to be used in the calculation of Z via Eq. 4.14, the extraction of signal procedure, presented in section 4.4, is used to estimate these values.

In literature, a common standard interpretation for the values of significance are given by: [44]:

- **Z** $<$ 3σ: The H_1 hypothesis is not statistically significant, which means that there is not enough data to observe the signal.

- **3** $<$ **Z** $<$ 5σ: Characterizes an *evidence*, the signal is close to be observed.

- **Z** $>$ 5σ: The H_1 hypothesis is statistical significant and characterize a *discovery*. In other words, if there is a signal, it will be observed.

The studies dedicated to the Ξ_{cc}^{++} particle are focused on the simulation of its production and decay in the ALICE 3 and, with the data, on trying to find the best combination between the topological cuts through strangeness tracking which returns the highest possible significance value, in special a value greater than 5σ.

Chapter 5

Machine Learning Applied to Candidate Selection Criteria

The term Machine Learning (ML) was initially conceived by the pioneer in Artificial Intelligence (AI) Arthur Lee Samuel in 1959 and according to him “machine learning is a field of study that gives computers the ability to learn without being explicitly programmed" [46]. But before receiving an official name, the concept of machine learning was already being used: in 1943, the scientists Walter Pitts and Warren McCulloch published a paper titled “Logical Calculus of Ideas Immanent in Nervous Activity" [47] in which they tried, using mathematical techniques, to map the processes of thought and decision making that occur in human cognition.

Machine learning is a set of computational and mathematical techniques that uses previous information and data to improve its performance and make predictions about a given task. These computational techniques are usually applied when there is a need to process and obtain useful information from the data, especially when this process is not feasible to perform manually or when the analysis involves a large amount of data. In physics, machine learning is also applied when the goal is to discover an equation or model that governs a specific set of data, providing for example a fit on it. For these cases, it is necessary to automate tasks, thus simulating human behavior. Currently, ML techniques can be classified as supervised, unsupervised or, more recently, as reinforced learning. For this particular work, the Gradient Boosted Decision Trees (GBDT) was used as our algorithm of choice, a supervised technique, which was used to perform the selection of candidates for the Ξ_{cc}^{++}.

In this chapter, the Gradient Boosted Decision Tree algorithm will be presented, as well as the concepts of supervised and unsupervised learning, classification and regression problems. The decision tree and gradient boosting algorithm will be explained in order to construct the idea behind the GBDT. It will also describe how the pre-processing of the data used to obtain the results of this thesis was carried out and how it composes the methodology applied to the Ξ_{cc}^{++} problem.

5.1 The Gradient Boosting Decision Trees

The Gradient Boosting Decision Trees is a supervised machine learning algorithm that can be used for regression and classification problems. The main principle is to first create simple prediction models sequentially, that is, connect smaller models such that each subsequent task can cover the imprecision's of the preceding ones. In GBDT, these individual learners are decision trees. Because it is a sequential algorithm, GBDT is a slow model to build, but highly accurate. To understand and employ it for selecting candidate particles, the following sections will go through the essential details of this machine learning algorithm.

5.1.1 Supervised and Unsupervised Learning

A machine learning model is said to be learning from a dataset when it looks for patterns and relationships between these data. Learning can be divided into two major groups: supervised and unsupervised learning. The difference between these two types of learning is that the first makes use of labeled data so that the machine can perform a specific task and measure its accuracy with respect to the given labels from the test dataset (See Fig.5.1). In unsupervised learning, the data is not labeled and the algorithm must learn by itself the correlations contained in the data.

Mathematically, in supervised learning, the general rule that associates inputs with outputs correctly can be defined as follows: let $= \{(x_i, y_i)\}$ be the dataset, where x_i are the variables present in the data used as input to the predictions, called predictor variables or features, and y_i is the target output that the model will try to predict. This latter variable is often called of outcome or response variable.

The dataset will then be divided into two subsets: the training and the test datasets, where the training data will be used to build the machine learning model. This model can be represented by a mapping function f which, for a given sample of feature x_i the model will give a prediction $\hat{y}_i = f(x_i)$. To find the f mapping the algorithm will try to solve an optimization task, measuring the prediction quality for an object x_i using a loss function $(y_i, f(x_i))$ which will be minimized in order to find the best model.

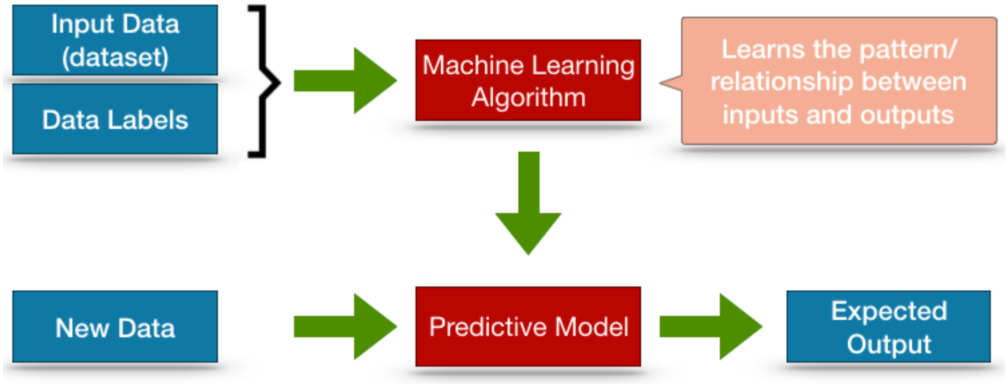

Input Data (dataset)
Data Labels
Machine Learning Algorithm
Learns the pattern/ relationship between inputs and outputs
New Data
Predictive Model
Expected Output

Figure 5.1: Schematic of supervised learning. This fluxogram represents the steps of providing the input data, training and application of the model to make predictions.

After the model is built, it is applied to the test data. With that, it is possible to make predictions for the samples in this set that, in turn, will be compared with the correct outcome to evaluate the model. Once the final model is obtained, it can be applied to make predictions to a new dataset with similar features from the training stage, for which the correct outcome is unknown.

5.1.2 Classification and Regression Problems

When subdividing supervised learning problems, two main groups of algorithms can be used to compose a ML solution: those for classification problems and others for regression tasks (see Fig. 5.2). In classification problems, the algorithm is trained to classify input data into discrete variables, that is, the algorithm will predict a class, or category, from a finite set of classes defined in the training data. For regression tasks, the algorithm is trained to predict an output from a continuous range of possible values. The algorithm in the regression process needs to identify a functional relationship between the input parameters and the output such that the output value is not discrete as in classification, but a continuous function of the input parameters, so that the response variable $\hat{y}$ will be continuous.

On the other hand, a binary classification problem, this is, when the data belong to two different classes, for example signal being equal 1 and background being equal 0, can be trans-

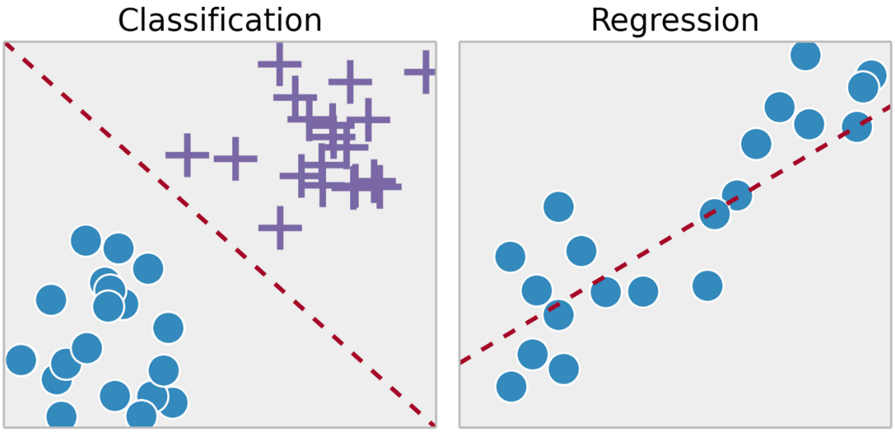


Figure 5.2: Schematization between the two types of supervised learning: classification and regression. Figure from [48]

formed into a regression problem: the regression model will provide a continuous spectrum from 0 to 1, which could be interpreted as a probability that the correct outcome is signal, and the user can then define a classification threshold that will separate what will be considered signal and what will be considered background. In this way, the regression value will be transformed into a binary category, so the output can be interpreted as a classification problem. Once the threshold is defined, it is possible to perform a visualization of the algorithm's performance by looking at the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), as in classification problems. This can be visualized in a confusion matrix, as shown in Tab. 5.1.

		Predicted Classification		
		Negative	Positive	
True Classification	Negative	TN	FP	CN
	Positive	FN	TP	CP
		RN	RP	N

Table 5.1: Structure of a confusion matrix, where it is possible to divide the data into classes like true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), depending on how the model classify the input given a threshold. CP and CN indicate the number of events in the positive or negative class and RN and RP indicate the predicted positive/negative events.

5.1.3 Decision Tree

Before introducing the GBDT model, it is necessary to understand the Decision Trees (DTs) learning technique, which are the basic components of the GBDT, as it represents the weaker learners of this algorithm [49]. Decision trees are a supervised learning method that is based on creating iterative questions. Using simple decision rules inferred from prior data, it partitions new data until it reaches a prediction. This algorithm can be used for classification and regression problems.

A decision tree is a predictive model that has the structure of a tree, that is, it starts at the root node of the tree and continues in a descending way through the nodes (see Fig. 5.3). Each node is used to denote a question about a feature of the dataset, and each branch is used to denote a decision. The first node contains the complete data sample. During training, the algorithm builds the tree recursively, in which the first nodes refer to the most important attributes as calculated from a metric such as the information gain. In this way, the deeper nodes of the tree will consequently be the ones of less importance or those that exert less influence on the resulting prediction [50]. Fig. 5.3 shows schematically an example of a decision tree for the selection of events in which the circles are called nodes and represent questions about one feature of the data sample, while the lines connecting the nodes show the outcome of different decisions such that the next node receives a certain subset of the data with the respective cut. After each cut, the sample is then divided, in the case of the figure, into two smaller sets, connected to the original sample by lines. With that, the rules are formed by the thresholds T_a, T_b, T_c and T_d applied to the variables V_a, V_b, V_c and V_d, which can be any topological or energy variable. The classification of events in the leaves of the tree is decided by a majority vote in the training data.

To perform decisions, the model needs to evaluate two fundamental criteria: 1) what are the most important variables to perform the cuts and 2) what is the optimal value of the cut. For this, the algorithm needs to quantify how well separated the classes are after each cut, that is, it is necessary to make use of an attribute that minimizes the information needed to classify the partitions in order to minimize the randomness. The algorithm is then said to measure the *impurity* of the classification.

Let S be a set of s data samples with m distinct classes labels C_i (i=1,..., m) and s_i be the number of samples of S with class equal to C_i. The common measures of impurity I are [49]:

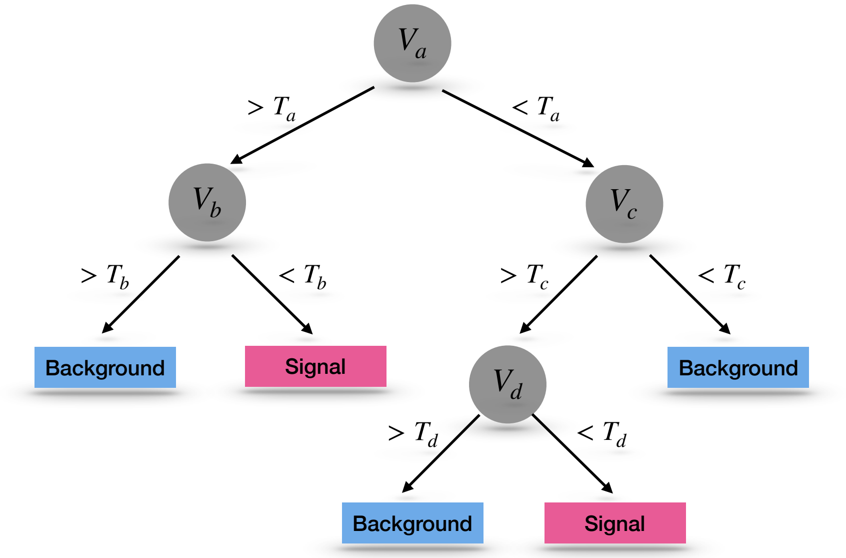


Figure 5.3: Schematic of a decision tree for selection events in signal/background. The rules are formed by the thresholds T_a, T_b, T_c and T_d applied to the variables V_a, V_b, V_c and V_d. The events are then classified as signal or background. Figure adapted from [51].

- Gini:

$$I\left(s_1, s_2, \ldots, s_m\right) = \sum_i p_i \left(1 - p_i\right), \tag{5.1}$$

- Log Loss or Shannon Entropy:

$$I\left(s_1, s_2, \ldots, s_m\right) = -\sum_{i=1}^{m} p_i \log_2\left(p_i\right), \tag{5.2}$$

where p_i is the probability that a sample belongs to class C_i and is calculated by s_i/s.

The most used impurity measure in decision tree is the entropy, actually, the Information Gain. This quantity is used to select informative features about the class of a sample, and that consequently will be employed to make the split. Hence, the feature with the highest information gain (or greatest entropy reduction) will be at the base of the tree. Shannon entropy can also be interpreted as the total information needed to classify a given sample. Thus, if an attribute A has v distinct values $a_1, a_2, a_3, ..., a_v$, it can be used to partition the set S into v subsets $S_1, S_2, S_3, ..., S_v$, such that S_j contains all samples of S with attribute A equal to a_j. If A is selected as a test attribute, i.e, best attribute for splitting, then these subsets will be distributed over the v descendant branches from the node containing the set S.

Let s_{ij} be the number of samples of class C_i in a subset S_j. The expected information based on the partitioning by A is given by [52]:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \cdots + s_{mj}}{s} I\left(s_{1j}, \ldots, s_{mj}\right), \tag{5.3}$$

where $\left(s_{1j} + \ldots + s_{mj}\right)/s$ is the weight of the j^{th} subset and corresponds to the fraction of samples from S that have the attribute A equal to a_j. The smaller the entropy value, the greater the purity of a subset of partitions. With this, it is possible to finally define the Information Gain as:

$$Gain(A) = I\left(s_1, s_2, \ldots, s_m\right) - E(A). \tag{5.4}$$

Therefore, the $Gain(A)$ can be interpreted as the expected reduction in the entropy caused by the knowledge of the value of attribute A [52].

The Information Gain is calculated for each feature, and the feature with highest gain is selected to make the split. This process is repeated considering the remaining features until some stop criteria is reached. For example, this criteria can be the maximum depth of the tree, which is the number of nodes along the longest path from the first node to a given leaf [53].

To make the split, it is possible to separate the values of the attribute into two groups. This is done when the decision tree is used for regression problems. In this case, the values are continuous and a threshold is set to define the two branches (Fig. 5.3). Inside a leaf, the value that best represents the train data that reaches this leaf is chosen as the outcome, that is, the value that minimizes some loss function.

5.1.4 Gradient Boosting

In practice, a single decision tree is not enough to be used in complex problems like candidate selection in particle physics. One of the most advanced techniques based on the decision tree is *Boosting*, in which decision trees are combined into numerous weak learners, which work sequentially and connected so that each subsequent tree attempts to minimize the errors of the previous ones. When a new tree is created, it fits into a modified version of the initial dataset, putting more weight on data that was classified incorrectly by the previous trees. This generates a collectively strong, high efficiency and accurate model.

The boosting process occurs through the following steps [3][54], as can be seen in Fig. 5.4:

1. Each weak model (tree) is created sequentially;

2. In the j^{th} step, a new tree that generates an outcome $f^{(j)}(x)$ is created together with a weight $w^{(j)}(x)$, which is relative to its accuracy.

3. The ensemble outcome in the j^{th} step is:

$$\hat{y}^{(j)}(x) = \sum_{t=1}^{j} w^{(t)}(x)f^{(t)}(x) = \hat{y}^{(j-1)}(x) + w^{(j)}(x)f^{(j)}(x). \tag{5.5}$$

4. To find $f^{(j)}(x)$ and $w^{(j)}(x)$, the algorithm tries to minimizes an objective function $Obj(x)$, which is given by:

$$\begin{aligned} Obj^{(j)} &= \sum_{i} l\left(\hat{y}^{(j)}(x_i), y_i\right) + \sum_{t=1}^{j} \Omega\left(f^{(t)}\right) \\ &= \sum_{i} l\left(\hat{y}^{(j-1)}(x_i) + w^{(j)}(x_i)f^{(j)}(x_i), y_i\right) + \sum_{t=1}^{j} \Omega\left(f^{(t)}\right), \end{aligned} \tag{5.6}$$

where the summation in the index i is over the training data samples; $l\left(\hat{y}^{(j)}(x_i), y_i\right)$ is the loss function, which is defined as a measure of the distance between the real value y_i and the value $\hat{y}^{(j)}(x_i)$ predicted by the algorithm for the i^{th} sample; and $\Omega\left(f^{(t)}\right)$ is the regularization function, which penalizes the complexity of the model $f^{(t)}$ to avoid overfits [7].

5. After each iteration, each data sample is given a weight based on its misclassification, so that data samples that are more frequently incorrectly classified will have greater weights.

Gradient boosting is a type of boosting where it uses gradient descent to minimize the objective function. Thus, in this model, the outcomes are targeted for the next model in an effort to minimize errors, that is, the weak learners are combined to fit the residuals from the previous model, improving the training. The targeted outcome for each case is based on the gradient of the error relative to the prediction. The gradient is used because the function generated in the training contains multiple variables.

[7] Overfitting is a term used to indicate when a Machine Learning model fits too well the training data in such a way that it becomes ineffective at predicting new results.

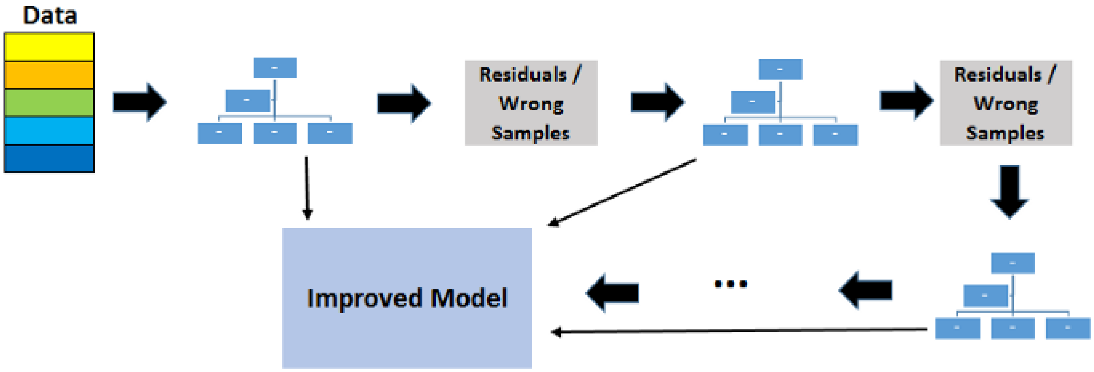


Figure 5.4: Schematic representation of working process of an GBDT: decision trees are combined into numerous weak learners, which works in sequence to allow each model to improve the error of the previous model, consequently generating a collectively strong model. Figure from [55].

For this work, the XGBoost (Extreme Gradient Boosting) [3] was used, which is a distributed Gradient Boosted Decision Tree machine learning library that is typically used in particle physics problems. This library provides an optimized solution through parallel processing, handling missing values, and regularization to avoid overfitting problems.

5.1.5 Hyperparameters used in the GBDT model

To create a GBDT model using XGBoost, it is necessary to assign the hyperparameters. This quantities are algorithm variables defined before training that are adjustable and allow to control the structure and functioning of the training process of the model. The optimal choice of these parameters can produce significant improvements in the final results, acting directly on the performance and predictivity of the model.

In XGBoost, there are many hyperparameters that can be used to construct a GBDT model. However, in this work, only five of them were used. The following is a brief description of the functionality of each of these hyperparameters [3]:

- **n_estimators**: It is the number of estimators (trees) and controls the number of boosting rounds;
- **learning_rate**: This parameter controls the step size of the gradient boosting, determining how fast or slow the model will learn. For this, a weighting factor is applied in the new corrections added by trees to the model.

- **max_depth**: This parameter sets the maximum depth of each tree that will be built. Increasing this value will make the model more complex and more susceptible to overfitting.
- **objective**: It defines the loss function that the algorithm will try to minimize. Notice that, if the objective is binary:logistic then the model will try to minimize a logistic regression for binary classification.
- **max_delta_step**: This parameter sets the maximum delta step for each tree's weight estimation, helping the update step to be more conservative. If it is set to be a positive value, the logistic regression might be improved for cases when the class is extremely imbalanced.

The hyperparameters `n_estimators` and `learning_rate` are correlated: to avoid overfitting due to a large number of trees, the `learning_rate` must be adjusted, decreasing its value, which must vary between 0 and 1.

Increasing the value of the number of estimators or the depth of the trees will make the model more complex and increases its predictive power. However, due to overffiting, increasing the value of hyperparameters is not necessarily related to increasing model performance.

5.2 The Ξ_{cc}^{++} Candidate Selection Case

The application of machine learning in this work arose from the need to maximize the significance, as discussed in Section 4.5. In this section, the modeling of the candidate selection problem for Ξ_{cc}^{++} will be explained.

Initially, note that the Ξ_{cc}^{++} observation is a complex problem, as it involves information about nine particles that participate in the decay process, with six of them being the observed daughter particles. As a consequence, many topological variables can be considered, making the standard candidate selection process very complicated. Furthermore, given that it is a very rare particle, the magnitude of the simulated/measured background is significative, with about one signal to six thousand background candidates, thus complicating the detection of its signal.

Thus, our main objective in using machine learning is to detect rare signal candidates over a large amount of background candidates. For each candidate, whether signal or background, the data from the simulation were organized as follows:

- Thirty-five topological variables about the candidate were selected. These will be the features that the model will use to make the predictions;
- The label for the candidate was chosen to be a binary output from the simulation that indicates if the prediction will be characterized as signal (1) or background (0).

With that in mind, the data will contain signal information, being it the H_1 hypothesis, and background as the alternative hypothesis (H_0). Then, the trained model will be applied to the test set to classify signal/background on an unseen dataset, which will have the same set of features with a proper distribution on the H_1 and H_0 hypotheses.

Some pre-cuts were applied to these variables: selection of centrality chosen to be $0 - 10\%$, and a range within the invariant mass centered at 3.621 GeV/c^2 with a variation of $\sigma = 0.08$ GeV/c^2, that is, only candidates with invariant mass in the range [3.541, 3.701] GeV/c^2 were selected. As for the test set, the selection of invariant mass has an sigma equal to $\sigma = 0.4$, more detail is shown in Chapter 6. The GBDT will choose the most important parameters from which the model can classify the output based on the provided features, that is, the model will be able to select, based on the topological variables, between signal and background candidates, thus optimizing the standard candidate selection method.

Despite being a classification problem, the regression method was chosen because the prediction applied to the results will return a value between 0 and 1, that in turn can be interpreted as a probability of the predicted candidate being a signal. With this, it is possible to define a threshold such that all data that are greater than or equal to this threshold will be considered signal and everything that is less will be classified as background in order to apply the significance calculation, as it was discussed in Section 4.5.

5.3 The simulation chain

The data used in this work come from Monte Carlo simulations generated by the ALICE collaboration for the ALICE 3 experiment. To use the Strangeness Tracking to study particles such as Ξ_{cc}^{++}, high statistics and an accurate description of weak decays are required. To satisfy these conditions, what is currently used is a combination of two types of simulations: full simulation and fast simulation. This combination is called hybrid simulation. The following is a brief description of these two types of simulations.

- **Full simulation**: The full simulation presents a more complete and realistic approach to event production, propagation, and interaction of particles with matter and the detector. For this case, the ALICE framework used for simulation, reconstruction, and data analysis, called O^2 [56], is used. In this framework, it is possible to develop more complex simulations to accurately simulate the ALICE 3 geometry. The events are initially generated using PYTHIA 8 [57] and the final-state particles are then passed to GEANT3 [58] which will do the entire interaction process with the matter as well as the respective detection in the active volumes. Detections are represented by hits left on the sensitive volume of the detector. In addition, Gaussian smearing is applied to the hit position to emulated the tracker's position resolution and to represent the intrinsic pixel resolution.

- **Fast simulation**: Fast simulation is used to speed up processes that do not need to be treated with high accuracy and precision. The basic principle is to apply smearing on tabulated parameters from an event generator without the need to propagate it using GEANT3. For this simulation, it is used the package *DELPHESO*^{2} which is based on DELPHES [59] fast-simulation package and the O^2. The final-state particles generated by PYTHIA 8 are passed to the DELPHES' modules where they will be propagated using the information about the curvature in the magnetic field and the decay to the desired radius. This information is stored for the next processing which will apply the detector's response using smearing on the kinematic variables. The tracking resolution and the elements of the covariant matrix for the track parameter are calculated using information like multiple scattering, detector occupancy, and energy loss. This covariant matrix is stored in multi-dimensional Look-up tables (LUTs) as a function of particle mass, pseudo-rapidity, transverse momentum and the event charged-particle density [24]. After the application of LUTs to the events, the final track list is converted to the ALICE analysis data and processed with standard analysis that is usually done for the data from the full simulation.

The hybrid simulation combines these two approaches to generate data. For the case where it is necessary high accuracy response for the detector, to be possible to use strangeness tracking, it is used the full simulation. This is the case of the weak decay hadrons K_s^0, Λ, Ξ^- and Ω^- and for the daughters of the Ξ_{cc}^{++} that comes from the decay of Ξ^-, including the pions and protons. This can be seen in the Fig. 4.2. All the other particles that do not require the complete propagation to the detector are only smeared with the fast simulation. This is the case of the pions that are daughters from Ξ_{cc}^{++} and Ξ_c^+.

Chapter 6

Results and Discussion

The methodology presented to perform candidate selection were described in Chapters 4 and 5 using the standard selection method and machine learning, respectively. This chapter will be dedicated to present the results of the application of these techniques to the Monte Carlo simulation data of Ξ_{cc}^{++} in Pb-Pb collisions at $\sqrt{s_{NN}} = 5.52$ TeV in the centrality of 0-10% in order to maximize the significance calculation. In particular, the objective is to obtain a significance greater than 5σ for the low transverse momentum defined by $0.0 < p_T < 2.0$ GeV/c, which is characterized as a delicate region since the candidate count drops significantly. In both cases, the topological variables used in this work were reconstructed using the information from the daughter particles associated with the strangeness tracking technique.

The results will be presented in two major blocks: using the standard candidate selection method and using machine learning, the latter being a computational resource that enables the intelligent combination of topological cuts. For both cases, an invariant mass spectrum is presented in which the signal extraction is performed, and thus the signal and background spectra and the respective ratio between them. Finally, the significance of the two cases is compared to determine which method has the better performance.

6.1 Candidate Selection: Standard Method

The standard candidate selection method was the first approach to study the Ξ_{cc}^{++} problem. Initially, it was necessary to study the topological variables, identifying regions where cuts could exclude a larger number of background candidates than signal candidates. Later, these variables were used to select candidates and obtain the invariant mass spectrum of the Ξ_{cc}^{++}. The invariant mass was used to perform the signal extraction described in 4.4, where the counting of background and signal candidates was performed. This can then be used to calculate significance. Each of these steps and their results is described in the following subsections.

6.1.1 Study of signal and background behavior in topological variables

Section 4.2.1 described some types of topological variables used for selection. The distribution of the different topological variables is shown in Figs. 6.1 and 6.2 for the signal (in magenta) and the background (in blue), respectively, as well as the region (dashed line) where the candidates were selected that gave the best significance value for the standard selection method. Thus, these cuts work as follows: For all Ξ_{cc}^{++} candidates, these variables are tabulated; those not within the cut are discarded, and those passing the cut are selected for the reconstruction of the invariant mass. The background candidates that pass through these cuts are called ‘false positives’ that contaminate the signal. The major limitation of this procedure lies in the fact that little is known about which regions are best suited to define the cuts and which values would be ideal for discarding primary particles, for example.

Therefore, a systematic study was conducted to test the cuts for some of these variables across multiple ranges. First, it started with default values for a given set of variables. Three possible values were examined for each variable, keeping the values of the other variables fixed. Thus, the cutoff value for the particular variable that had demonstrated greater significance, especially for low-p_T (0.0-4.0 GeV/c), was fixed. After doing that for one variable, the same process was done sequentially for all the other ones. The best values are shown in Tab. 6.1.

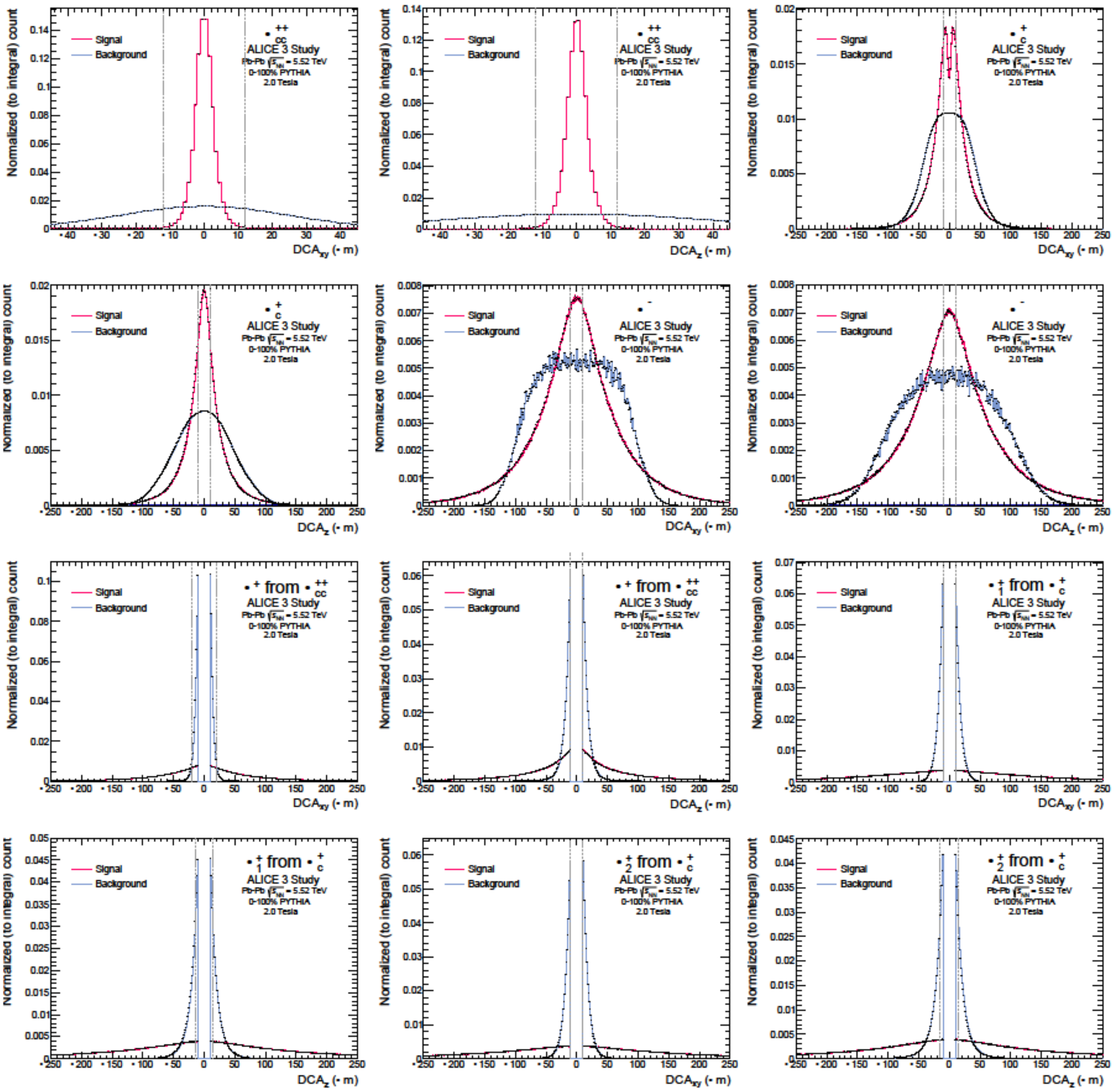


Figure 6.1: Signal (in magenta) and background (in blue) of topological variables associated to the Ξ_{cc}^{++} decay. In the dashed line it is possible to see the defined selection region.

.

Ξ_{cc}^{++} selection		Ξ_{cc}^{++} selection	
Ξ_{cc}^{++} DCA_{xy} to PV (μm)	$\|DCA_{xy}\|$ <12	(1) $\pi \leftarrow \Xi_c^+$ DCA_{xy} to PV (μm)	$\|DCA_{xy}\|$ <10
Ξ_{cc}^{++} DCA_z to PV (μm)	$\|DCA_z\|$ <12	(1) $\pi \leftarrow \Xi_c^+$ DCA_z to PV (μm)	$\|DCA_z\|$ <15
Ξ_c^+ DCA_{xy} to PV (μm)	$\|DCA_{xy}\|$ <10	(2) $\pi \leftarrow \Xi_c^+$ DCA_{xy} to PV (μ)	$\|DCA_{xy}\|$ <10
Ξ_c^+ DCA_z to PV (μm)	$\|DCA_z\|$ <10	(2) $\pi \leftarrow \Xi_c^+$ DCA_z to PV (μm)	$\|DCA_z\|$ <15
Ξ^- DCA_{xy} to PV (μm)	$\|DCA_{xy}\|$ <10	(2) Ξ_{cc}^{++} Decay Radius (cm)	>0.015
Ξ^- DCA_z to PV (μm)	$\|DCA_z\|$ <10	Ξ_c^+ Decay Radius (cm)	>0.003
$\pi \leftarrow \Xi_{cc}^{++}$ DCA_{xy} to PV (μm)	$\|DCA_{xy}\|$ <20	Ξ_{cc}^{++} DCA Between Daughters (μm)	<8
$\pi \leftarrow \Xi_{cc}^{++}$ DCA_z to PV (μm)	$\|DCA_z\|$ <10	Ξ_c^+ DCA Between Daughters (μm)	<14

Table 6.1: Topological variables and the respective intervals that define the regions used for the selection of candidates.

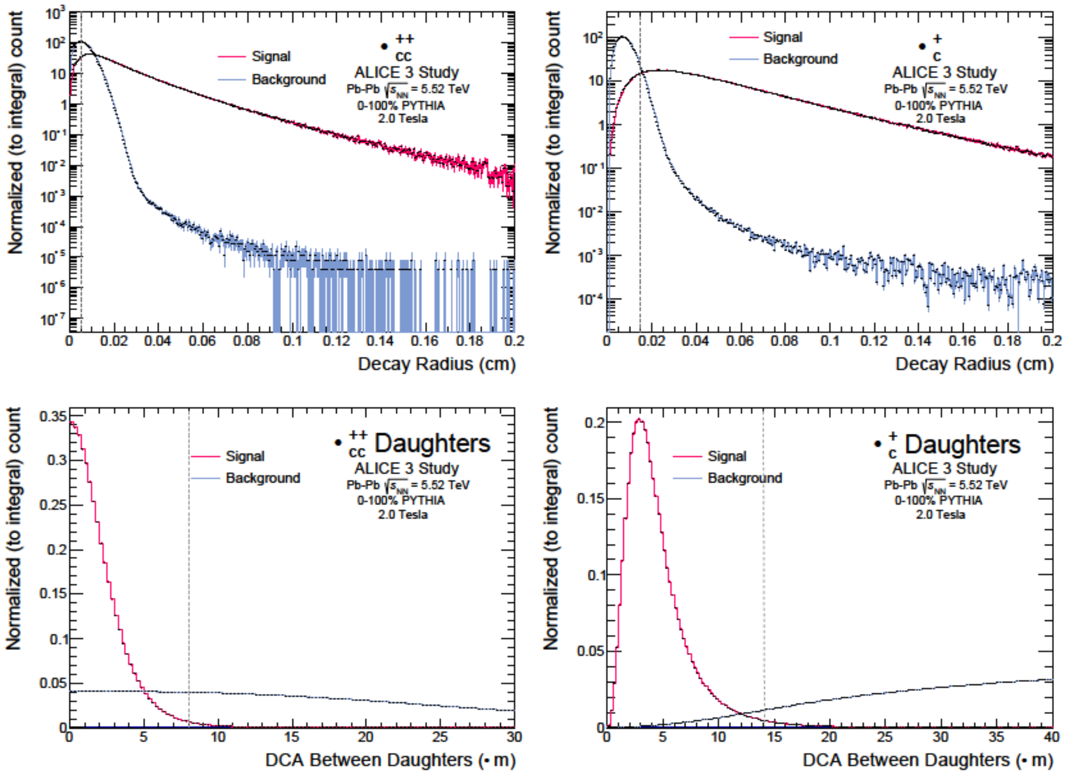


Figure 6.2: Signal (in magenta) and background (in blue) of topological variables associated to the Ξ_{cc}^{++} decay for Pb-Pb collision at $\sqrt{s_{NN}} = 5.52$ TeV. In the dashed line it is possible to see the defined selection region.

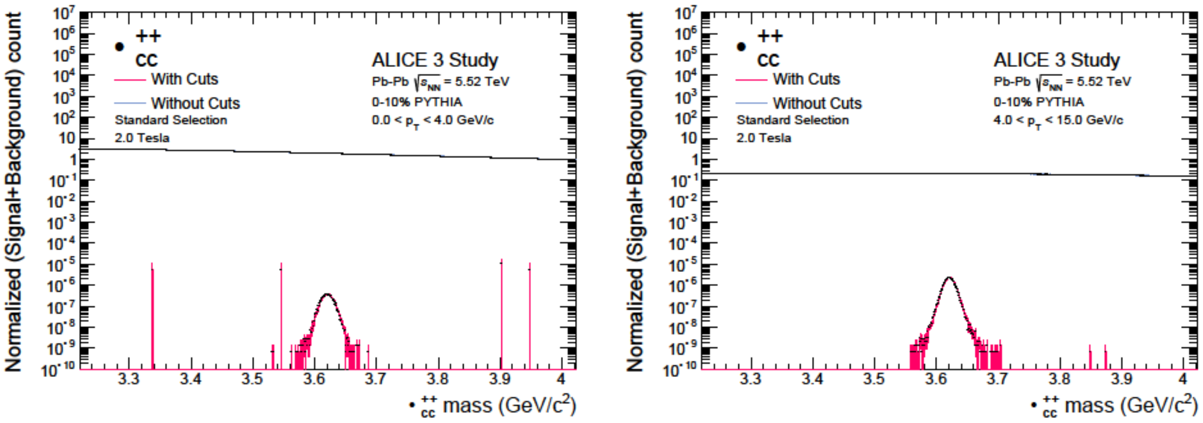
6.1.2 Invariant Mass and Ξ_{cc}^{++} Yield

The invariant mass spectrum of Ξ_{cc}^{++} particle is shown in Fig. 6.3, where the blue curve presents the spectrum without the topological selection and in magenta the same spectrum but with the cuts presented in Tab. 6.1. It is then possible to observe that once the topological cuts are applied, the peak on invariant mass turns visible. In (a) there is the invariant mass spectrum for the transverse moment region of 0.0-4.0 GeV/c and in (b) the same but for the region 4.0-15.0 GeV/c.

To create these spectra, it is necessary to apply a set of scaling factors in the invariant mass histogram that are different for signal and background. This difference occurs because the production of Ξ_{cc}^{++} is a very rare event and it is costly for the Monte Carlo simulations to produce the events in the correct proportions.

The following operations were performed for the signal:

1. Divide the count of the mass histogram by the number of event N_{events} in the centrality of $0 - 10\%$;



(a) (b)

Figure 6.3: Invariant mass spectrum of Ξ_{cc}^{++} particle for (a) low-p_T:0.0 - 4.0 GeV/c and (b) high-p_T 4.0-15.0 GeV/c. In blue is the spectrum without any topological selection and in magenta the spectrum with the selections listed in Tab.6.1.

2. Divide by the bin width, transformed into a density;
3. Multiply by per-event yield in central events (calculated by Becattini in [20]) 2/137 (factor 2, considering particle and anti-particle);
4. Divide by 400 to account for two Branching Ratios (BR) of 5% × 2.8%;
5. Divide by 5 to account for signal injection with dN / dy = 5 (increase the number of events).

As for the background, the following factors were applied:

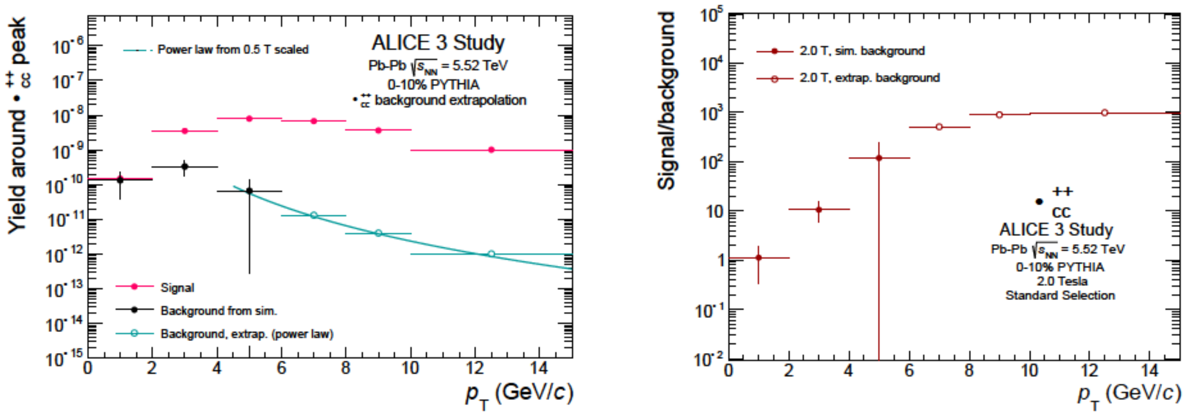
1. Divide the count of the mass histogram by the number of event N_{events} in the centrality of $0 - 10\%$;
2. Divide by bin width (same bin width as signal).

All invariant mass spectra presented in this work were constructed taking these factors into account. Once the invariant mass was calculated then the bin counting process is applied for each p_T interval. The counting of candidates for signal and background using these spectra is shown in Fig. 6.4 (a). For high-p_T bins, an extrapolation using the following power law was done:

$$f(x) = [0] \times x^{[1]}, \tag{6.1}$$

where [0] is a free renormalization parameter and [1] is a fixed parameter obtained from a fit with the same law of Eq. 6.1 to the data from the simulation with a magnetic field of 0.5 T. This procedure is necessary because at high momentum (above 6 GeV/c) the background yield is very scarce, and in some cases there is no counting. In the case of a magnetic field of 0.5 T, there is more background counting in relation to the 2.0 T. The assumed hypothesis was that the behavior of the background spectrum with a magnetic field of 2.0 T would be the same for the case where the field is equal to 0.5 T. To avoid generating gigantic and impractical amounts of background, an extrapolation that describes it sufficiently well must be done. Moreover, the power law function was chosen because, among the models that describe the behavior of this spectrum, it is the one that provides the most conservative result, i.e., the one that produces the highest count of background candidates.

The ratio of signal to background yields is shown in Fig.6.4 (b). The signal is observed to be significantly lower for low-p_T bins.



(a)

(b)

Figure 6.4: (a) Yields of signal and background around the Ξ_{cc}^{++} peak region, the counts was estimated using the bin counting procedure. (b) Ratio between signal and background yields for Ξ_{cc}^{++} particle for Pb-Pb collision at $\sqrt{s_{NN}}$ = 5.52 TeV.

6.1.3 Significance

Using the calculated yields shown in Fig. 6.4 and Eq. 4.12 it is possible to obtain the significance for each p_T interval. The results are shown in Fig. 6.5. Except for the first bin, all the other momentum intervals have significance greater than 5σ. For the first bin, which characterizes the p_T interval of 0.0-2.0 GeV/c, the significance assumes a value around 3σ.

Note that this significance value does not provide enough statistical confidence to verify that it will be possible to measure this particle in central Pb-Pb collisions down to zero p_T with ALICE 3. This measurement is important to reduce uncertainties in total yields. It allows a p_T-differential measurement with good statistical accuracy since with this measure it will not be necessary to carry out any extrapolation to zero, a process that involves a high degree of uncertainty. For this, it is necessary to explore techniques that allow to perform a particle identification that results in a significance greater than 5σ in this p_T interval.

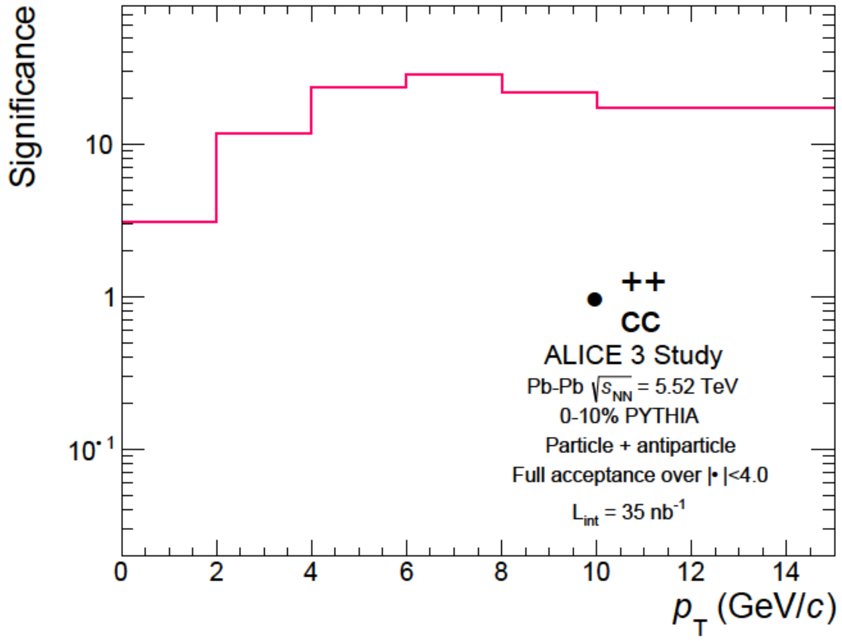


Figure 6.5: Significance obtained using the standard candidate selection for Ξ_{cc}^{++} particle for Pb-Pb collision at $\sqrt{s_{NN}}$ = 5.52 TeV for centrality 0 – 10%.

6.2 Candidate Selection: Machine Learning

To try to improve the results obtained using the rectangular candidate selection, a different approach was applied. As described in Section 5.2, due to the problem of Ξ_{cc}^{++} having many daughters, making the selection of its decay can be challenging. Machine learning techniques were then applied to optimize the selection of candidates. In particular, the technique used was Gradient Boosted Decision Tree, as described in Chapter 6.

This process involves the following steps: initially, data pre-processing is applied, in which centrality and invariant mass filtering take place; subsequently, with these data, the machine learning model is trained; the latter is applied to a test dataset to obtain an output of the model, in which there is a spectrum between 0 and 1, which can be interpreted as the probability of a candidate being a signal. This spectrum is stored and processed in an analysis chain similar to the one used to produce the results of the rectangular cuts, with the difference that this time a threshold value about the mentioned spectrum is passed as an argument. The use of different threshold values allows for the calculation of the mass distributions that contain both true positives and the remaining false positive candidates. These distributions are then used in the analysis. Note that the machine learning threshold works as an application of all cuts together, considering all topological variables passed as features of the model. Different thresholds were studied to find the best that maximizes the significance in $0.0 < p_T < 2.0$ GeV/c

6.2.1 GBDT Model training and application

To train the machine learning model, the data was separated by transverse momentum interval. Only candidates that had transverse momentum within the $0.0 < p_T < 4.0$ GeV/c interval was used for training. The values of the hyperparameters of the model that presented the highest performance to the significance value for this region are shown in Tab. 6.2. To obtain these values, the hyperparameters were systematically tested so that they had their values varied in order to maximize significance.

Note that, although the training includes data from only momentum in the region defined by $0.0 < p_T < 4.0$ GeV/c, the test data in turn include the entire transverse momentum range, i.e., the data comprises the region of 0.0-15.0 GeV/c. Furthermore, the model were trained considering an invariant mass interval of $\sigma = 0.08$ GeV/c^2 around 3.621 GeV/c^2 and the test dataset have an invariant mass window of $\sigma = 0.4$ GeV/c^2.

	n_estimators	**learning_rate**	**max_depth**	**objective**	**max_delta_step**
Low-p_T training	100	0.1	9	binary regression	10

Table 6.2: Hyperparameter values used to create the GBDT model that had the best performance for the region of $0.0 < p_T < 4.0$ GeV/c.

Thirty-five input features were passed to the trained model, but some of them have more impact and importance. Calculating the importance level of each feature to the model helps, among other things, to understand the behavior of the data as well as the relationship between the features, which in this problem are the topological observable, and the target variable. Understanding this ranking also helps to optimize the model in terms of computational consumption, since less important features can be removed for future training. The feature importance for the trained model of Tab. 6.2 are shown in Fig. 6.6.

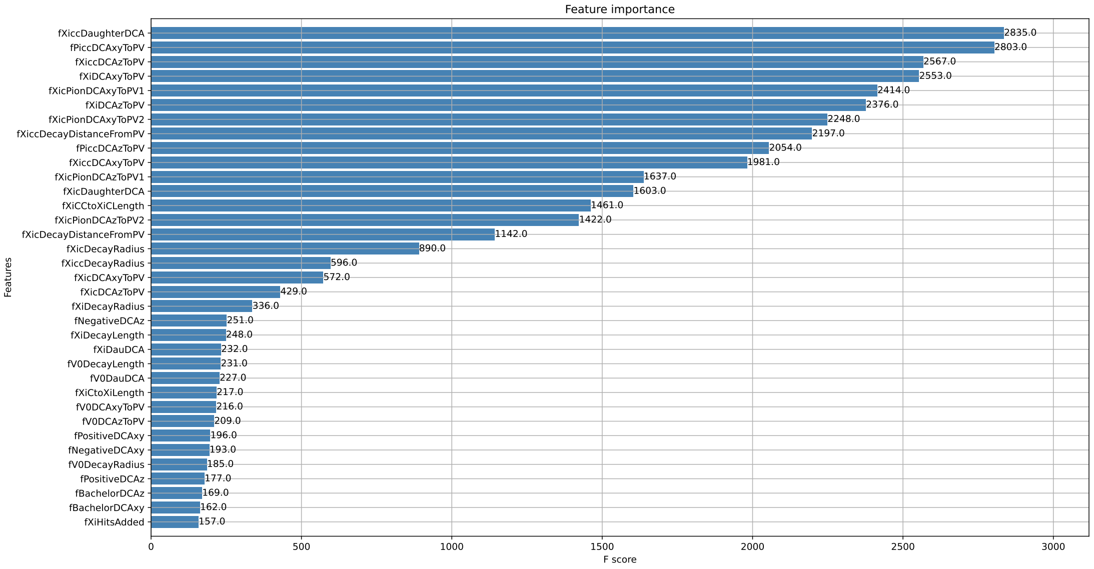


Figure 6.6: Feature Importance from the training for low-p_T using the hyperparameter of the Tab. 6.2. In the vertical axis there are the features used to train the model and in the longitudinal axis there are the score ranking.

Note that the DCA between the daughter of the directly decay of the Ξ_{cc}^{++} (XiccDaughterDCA), the DCA to the primary vertex in relation to the transverse plane of the π^{-} emitted in the decay of Ξ_{cc}^{++} (PiccDCAxyToPV) and the DCA to primary vertex in the longitudinal direction for the Ξ_{cc}^{++} (XiccDCAzToPV) represents the variables with greater influences for the learning of the model in low-p_T.

Once the model is trained it is then possible to apply it to the test dataset and visualize its response. The output of the machine learning for low-p_T applied in the train and test datasets are shown in Fig. 6.7. In magenta, the curve represents the prediction of the model to the data that was in fact signal, which presents a peak at 1. The blue curve is the prediction of the model to the data that was in fact background. The triangles are the response of the model for the train dataset, with the same colors as before for signal and background. Note that the background doesn't have a peak in 0 because a pre-cut was applied discarding all data below 0.1 to reduce the file size and allow the data to be processed.

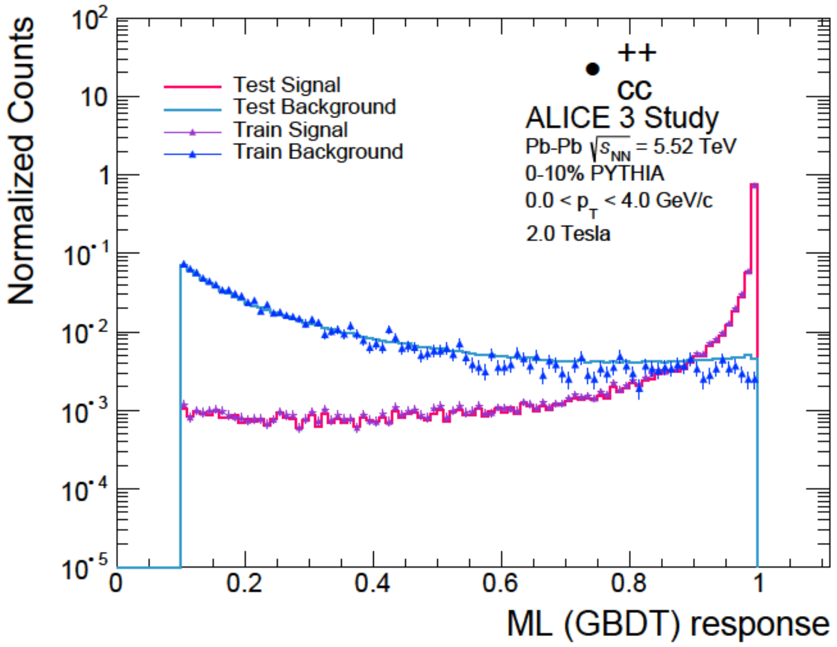


Figure 6.7: Gradient Boosted Decision Tree response for train and test dataset for low-p_T: 0.0-4.0 GeV/c.

When comparing the training and test curves in Fig. 6.7, it is noticeable that they fit each other with a proper concordance, which indicates that the model does not overtrain [44]. This comparison between model's output curves for the training and test data is usually used in high-energy physics as an indication of whether the model overfits or not.

The region of interest for analysis is located around the peak at 1. Strictly tight cuts in the threshold proved necessary, since, in general, the tighter the better the result. However, once a certain threshold is reached, this relationship no longer holds and either significance decreases or both signal and background are completely truncated. In addition, despite a high number of false positives, the region around 1 has also been shown to have a higher number of true positives, making it the region of interest.

In these machine learning outputs, one can observe that a background peak is formed in the region around 1. This can be explained by the incorrect classification of the algorithm with respect to the primary particles Ξ_c^+, since this particle has very similar tracking properties to the particle Ξ_{cc}^{++} and also decays before passing through any layer of the ALICE 3 detector.

In the current study, the threshold that showed the best performance in terms of significance was 0.9987 for low-p_T. A significance above 5σ for the region $0.0 < p_T < 2.0$ GeV/c was already obtained with a threshold above 0.9908 using the model built with the hyperparameters in Tab. 6.2.

6.2.2 Invariant Mass and Ξ_{cc}^{++} Yield:

For the threshold mentioned before, the invariant mass spectra obtained are presented in Fig. 6.8. These were then used to calculate the yield of Ξ_{cc}^{++}, which can be seen in Fig. 6.9.

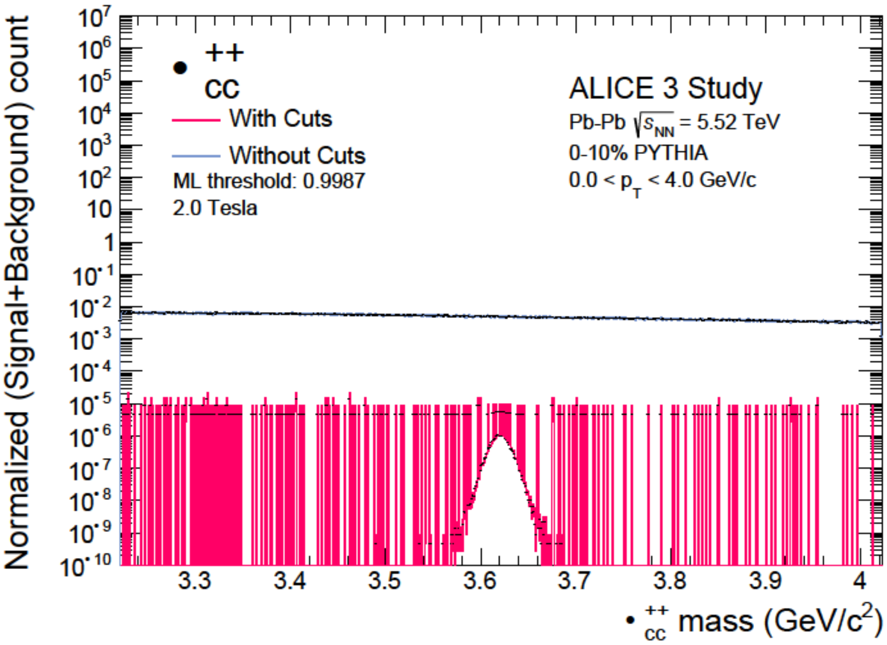
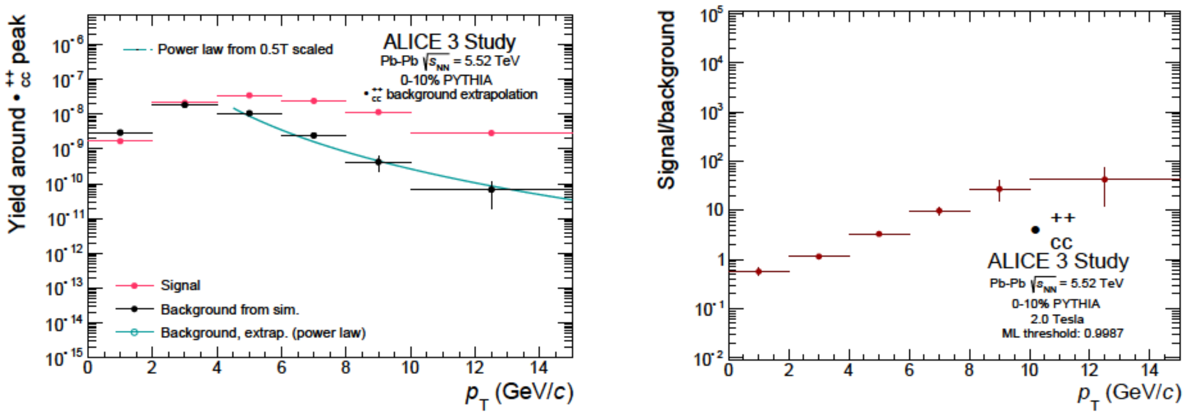


Figure 6.8: Invariant mass spectrum of Ξ_{cc}^{++} particle using training for low-p_T:0.0 - 4.0 GeV/c. In blue there is the spectrum without any topological selection and in magenta the spectrum with the machine learning cuts.

Note that for this case there was sufficient background count for the analysis and there was no need to use the extrapolation process for the background yield. It is also possible to observe that the number of candidates that were selected is greater than those of the standard procedure. This is due to all the cuts applied throughout the process that involves the use of machine learning.

It is important to mention that the amount of data used to train the model is much smaller when compared to complete statistics, especially for the background, in which only 1% of the total data was used. However, to assemble this data set, the portion that was used represent the same statistical distribution as the full sample. It was necessary to use only this amount of data due to the fact that training the machine learning model involves high consumption of memory, so, in order to perform this method, it is necessary to use a smaller background dataset, although representative, of what is used in the standard analysis.



(a)

(b)

Figure 6.9: (a) Yields of signal and background around the Ξ_{cc}^{++} peak region, the counts were estimated using the bin counting procedure and an extrapolation from a scaled power law. (b) Ratio between signal and background yields for Ξ_{cc}^{++} particle for Pb-Pb collision at $\sqrt{s_{NN}}$ = 5.52 TeV using the GBDT model trained to low-p_T:0.0-4.0 GeV/c. Note that the test dataset involves all p_T regions.

The dataset used in the test and, therefore, in the analysis presented for the calculation of significance was the complete statistics provided by the simulation of Ξ_{cc}^{++} in ALICE 3.

6.2.3 Significance

The significance calculated using the yields of the Ξ_{cc}^{++} from the selection employing the low-p_T training is shown in Fig. 6.10. Note that this low momentum training demonstrated a high significance value for low-p_T, in particular, a significance above 9σ was obtained for the 0.0-2.0 GeV/c region, which, according to the criteria presented in Chapter 4, would characterize a *discovery*. In other words, this value presents an indication that this particle can be measured in a transverse momentum down to zero.

For the p_T interval defined by 2.0-4.0 GeV/c the significance reaches the value of 38σ and for high momentum it also provided a good result, far superior to that obtained with the standard approach. Although the significance value obtained previously for high-p_T region was already relatively high, the machine learning process has managed to significantly increase this value and showed satisfactory performance even in the regions that the model was not dedicated to.

Furthermore, in this approach using the GBDT machine learning model, all transverse momentum regions had significance greater than 5σ.

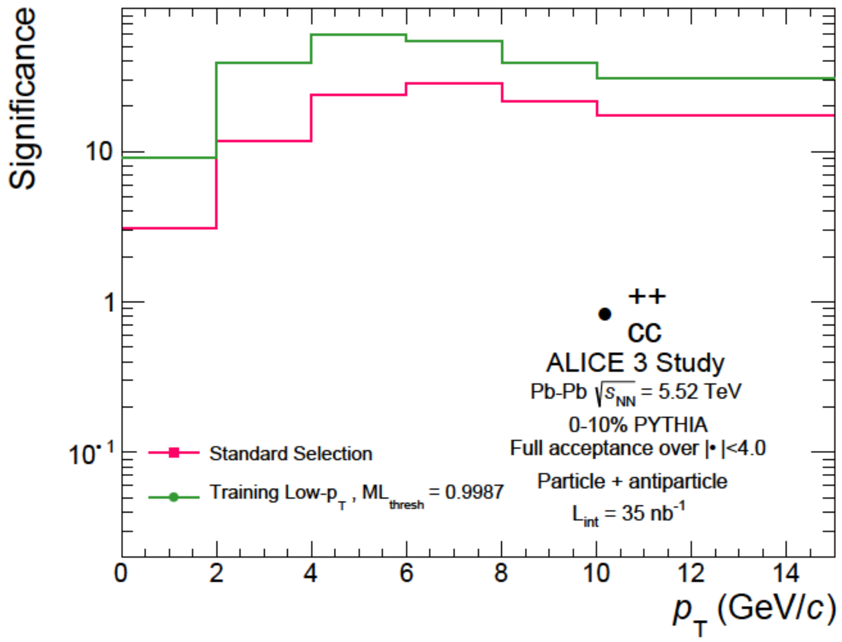


Figure 6.10: Significance obtained using the machine learning procedure to candidate selection for Ξ_{cc}^{++} particle for Pb-Pb collision at $\sqrt{s_{NN}}$ = 5.52 TeV for centrality 0 − 10%. In green is the significance using the model trained to the interval of momentum: 0.0-4.0 GeV/c, in blue the significance for model trained in 4.0-15.0 GeV/c and the magenta curve is the significance for standard selection.

These results demonstrate the high power of the application of machine learning techniques to candidate selection. It is important to comment that a machine learning model depends not only on the hyperparameters chosen but also on the dataset used for training. Therefore, this significance value, although encouraging, may suffer some fluctuations depending on the training set.

Chapter 7

Conclusion

In general, the objective of this work has been satisfactorily achieved. It was possible to explore and study the standard approach of candidate selection for the particle Ξ_{cc}^{++} as well as a new approach using machine learning techniques. The latter in turn showed a better discriminatory character between signal and background, improving the results of significance.

The results of this work demonstrated that the multi-charm baryon Ξ_{cc}^{++} can be measured in the ALICE 3 experiment with the strangeness tracking reconstruction in all transverse moments. Machine learning procedure particularly presents that the Ξ_{cc}^{++} can be measured with transverse momentum lower than 2 GeV/c.

Originally, this work was devoted to the study of the behavior of the topological variables corresponding to the decay of the particle Ξ_{cc}^{++}. This study allowed a better understanding of the behavior of the signal and background curves in order to indicate, albeit vaguely, in which regions the rectangular cuts should be made. Several combinations were tested to obtain the highest possible significance value. Since Ξ_{cc}^{++} has a complex decay with many prongs, this method, normally used in particle physics, is not practical as many combinations must be tested. Therefore, this method proved to be limited for this case.

Since the usual method of candidate selection is inefficient, it was proposed to use machine learning resources for candidate selection. It was decided to work with ensembles of decision trees, i.e., a set of models that make decisions based on thresholds for numerical features. This is similar to the process of making cuts, and the machine learning algorithm used can be seen as applying all the cuts at once. Modeling the Ξ_{cc}^{++} problem and preprocessing the data was the first difficulty with this approach, as it was necessary to investigate which of the variables provided by the simulation should be used as features and which should be used as outcomes.

Once the data were organized, it was decided to work only with the centrality of 0 – 10%, that is, with the central events, since this would reduce the problem to some extent.

Another limitation arose with respect to the range of invariant mass in which the data were selected. A larger range increased the number of candidates selected for the procedure, which significantly affected the available computing power. Therefore, the invariant mass cut for training had to be fairly restricted to include as many background candidates as possible around the peak region of the invariant mass. For this, it was necessary to apply a $\sigma = 0.08\text{GeV/c}^2$ for the training data and a $\sigma = 0.4\text{GeV/c}^2$ for the test data. These complications can be summarized in a single problem: the amount of data to be processed. A large amount of data requires high computational power, which ultimately limited the analysis of this work using this method. Consequently, the results could have shown better performance if all available data had been used.

The second point using the machine learning approach was to find a model that would provide a maximum reasonable value for significance in the interval of transverse momentum defined by $0.0 < p_T < 2.0$ GeV/c. This, in turn, was the subject of investigation until the last moment of this work. The results presented in Chapter 6 showed that the machine learning approach performed better than the usual candidate selection approach despite the circumstances. It provided a significance value around of 9σ for the interval $0.0 < p_T < 2.0$ GeV/c and 38σ for the interval $2.0 < p_T < 4.0$ GeV/c which is approximately three times greater, in both regions, than the results obtained via the standard method. A good improvement over the standard approach for high-p_T was also observed, although the data used in the training did not contain this region. However, this procedure also showed instability in terms of the significance value obtained. The reason for this is that the performance of a model depends not only on the hyperparameters used in its construction, but also on the dataset used for training. It was also observed that these results varied slightly as the amount of background data in the test set increased.

In conclusion, with this work it was possible to gain knowledge of data analysis commonly performed on large experiments in high energy physics, such as in the future experiment ALICE 3. It was also possible to explore a new approach using current machine learning methods to successfully select candidates for the double charm baryon Ξ_{cc}^{++}. The results obtained with this technique, compared to the standard approach, are a strong indication that this is the right direction for the future analysis of this particle and other multiply charmed baryons, such as Ω_{ccc}^{++}.

Bibliography

[1] The large hadron collider. URL https://home.cern/science/accelerators/large-hadron-collider.

[2] Alice collaboration publishes letter of intent. URL https://timeline.web.cern.ch/alice-collaboration-publishes-letter-intent.

[3] Chen, T. & Guestrin, C. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016). URL https://doi.org/10.1145%2F2939672.2939785.

[4] The dirac equation and the prediction of antimatter. URL http://multimidia.ufrgs.br/conteudo/frontdaciencia/dirac%20antimatter%20paper.pdf.

[5] Thomson, M. *Modern particle physics* (Cambridge University Press, New York, 2013).

[6] Bayatian, G. L. *et al.* CMS Physics: Technical Design Report Volume 1: Detector Performance and Software (2006).

[7] Collaboration, T. A. *et al.* The ATLAS experiment at the CERN large hadron collider. *Journal of Instrumentation* **3**, S08003–S08003 (2008). URL https://doi.org/10.1088/1748-0221/3/08/s08003.

[8] Standard model. URL https://en.wikipedia.org/wiki/Standard_Model.

[9] Wong, C. Y. *Introduction to high-energy heavy ion collisions* (1995).

[10] Perkins, D. H. *Introduction to high energy physics* (1982).

[11] Khachatryan, V. *et al.* Measurement of the inclusive 3-jet production differential cross section in proton–proton collisions at 7 TeV and determination of the strong coupling

constant in the TeV range. *The European Physical Journal C* **75** (2015). URL `https://doi.org/10.1140%2Fepjc%2Fs10052-015-3376-y`.

[12] Maire, A. *Production des baryons multi-étranges au LHC dans les collisions proton-proton avec l'expérience ALICE.* Ph.D. thesis, Strasbourg U. (2011).

[13] Yagi, K., Hatsuda, T. & Miake, Y. *Quark-Gluon Plasma: From Big Bang to Little Bang*. Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology (Cambridge University Press, 2005). URL `https://books.google.com.br/books?id=C2bpxwUXJngC`.

[14] Borcsik, F. S. Estudo da influência das condições iniciais em colisões nucleares ultrarrelativísticas.

[15] Rafelski, J. & Müller, B. Strangeness production in the quark-gluon plasma. *Phys. Rev. Lett.* **48**, 1066–1069 (1982). URL `https://link.aps.org/doi/10.1103/PhysRevLett.48.1066`.

[16] Paquet, J.-F. Probing the space-time evolution of heavy ion collisions with photons and dileptons. *Nuclear Physics A* **967**, 184–191 (2017). URL `https://doi.org/10.1016%2Fj.nuclphysa.2017.06.003`.

[17] Back, B. *et al.* The phobos perspective on discoveries at rhic. *Nuclear Physics A* **757**, 28–101 (2005). URL `https://www.sciencedirect.com/science/article/pii/S0375947405005282`. First Three Years of Operation of RHIC.

[18] Braun-Munzinger, P. & Stachel, J. J. (Non)thermal aspects of charmonium production and a new look at J/ψ suppression. *Phys. Lett. B* **490**, 196–202 (2000). URL `http://cds.cern.ch/record/449169`.

[19] Kostyuk, A. P. Double, triple and hidden charm production in the statistical coalescence model (2005). `nucl-th/0502005`.

[20] Becattini, F. Production of multiply heavy flavored baryons from quark gluon plasma in relativistic heavy ion collisions. *Phys. Rev. Lett.* **95**, 022301 (2005). URL `https://link.aps.org/doi/10.1103/PhysRevLett.95.022301`.

[21] Chen, Y.-Q. & Wu, S.-Z. Production of triply heavy baryons at LHC. *Journal of High Energy Physics* **2011** (2011). URL `https://doi.org/10.1007%2Fjhep08%282011%29144`.

[22] Berezhnoy, A. V., Kiselev, V. V., Likhoded, A. K. & Onishchenko, A. I. Doubly charmed baryon production in hadronic experiments. *Physical Review D* **57**, 4385–4392 (1998). URL `https://doi.org/10.1103%2Fphysrevd.57.4385`.

[23] Andronic, A. *et al.* The multiple-charm hierarchy in the statistical hadronization model. *Journal of High Energy Physics* **2021** (2021). URL `https://doi.org/10.1007%2Fjhep07%282021%29035`.

[24] ALICE, C. Letter of intent for ALICE 3: A next generation heavy-ion experiment at the LHC. Tech. Rep., CERN, Geneva (2022). URL `https://cds.cern.ch/record/2803563`.

[25] Alves, A. A., Jr. *et al.* The LHCb Detector at the LHC. *JINST* **3**, S08005 (2008).

[26] Shen, C., Heinz, U., Paquet, J.-F. & Gale, C. Thermal photons as a quark-gluon plasma thermometer reexamined. *Physical Review C* **89** (2014). URL `https://doi.org/10.1103%2Fphysrevc.89.044910`.

[27] Coquet, M., Du, X., Ollitrault, J.-Y., Schlichting, S. & Winn, M. Intermediate mass dileptons as pre-equilibrium probes in heavy ion collisions. *Physics Letters B* **821**, 136626 (2021). URL `https://doi.org/10.1016%2Fj.physletb.2021.136626`.

[28] Bratkovskaya, E. L., Linnyk, O. & Cassing, W. Electromagnetic probes of the qgp (2014). URL `https://arxiv.org/abs/1409.4190`.

[29] Tanabashi, M. *et al.* Review of particle physics. *Phys. Rev. D* **98**, 030001 (2018). URL `https://link.aps.org/doi/10.1103/PhysRevD.98.030001`.

[30] Pisarski, R. D. Where does the ρ go? chirally symmetric vector mesons in the quark-gluon plasma. *Physical Review D* **52**, R3773–R3776 (1995). URL `https://doi.org/10.1103%2Fphysrevd.52.r3773`.

[31] Hohler, P. M. & Rapp, R. Is ρ-meson melting compatible with chiral restoration? *Physics Letters B* **731**, 103–109 (2014). URL `https://doi.org/10.1016%2Fj.physletb.2014.02.021`.

[32] de Albuquerque, D. S. Multi-strange hadrons in pb–pb collisions at the lhc with alice.

[33] Chinellato, D. D. Estudo de estranheza em colisoes proton-proton no lhc.

[34] Aaij, R. *et al.* Centrality determination in heavy-ion collisions with the LHCb detector. *JINST* **17**, P05009. 38 p (2021). URL `https://cds.cern.ch/record/2789548`. All figures and tables, along with any supplementary material and additional information, are available at https://lhcbproject.web.cern.ch/Publications/p/LHCb-DP-2021-002.html (LHCb public pages), `2111.01607`.

[35] Miller, M. L., Reygers, K., Sanders, S. J. & Steinberg, P. Glauber modeling in high-energy nuclear collisions. *Annual Review of Nuclear and Particle Science* **57**, 205–243 (2007). URL `https://doi.org/10.1146/annurev.nucl.57.090506.123020`. `https://doi.org/10.1146/annurev.nucl.57.090506.123020`.

[36] Altsybeev, I. & Kovalenko, V. Classifiers for centrality determination in proton-nucleus and nucleus-nucleus collisions. *EPJ Web of Conferences* **137**, 11001 (2017). URL `https://doi.org/10.1051%2Fepjconf%2F201713711001`.

[37] Palomo, L. V. The ALICE experiment upgrades for LHC run 3 and beyond: contributions from mexican groups. *Journal of Physics: Conference Series* **912**, 012023 (2017). URL `https://doi.org/10.1088/1742-6596/912/1/012023`.

[38] Abelev, B. *et al.* Technical Design Report for the Upgrade of the ALICE Inner Tracking System. Tech. Rep. (2013). URL `https://cds.cern.ch/record/1625842`.

[39] Fabbietti, L. Heavy-ion programme for Run 5 and 6 at the LHC. 10th Edition of the Large Hadron Collider Physics Conference (2022). URL `https://cds.cern.ch/record/2811122`.

[40] Antcheva, I. *et al.* ROOT: A C++ framework for petabyte data storage, statistical analysis and visualization. *Comput. Phys. Commun.* **182**, 1384–1385 (2011).

[41] Aaij, R. *et al.* Measurement of the Lifetime of the Doubly Charmed Baryon Ξ_{cc}^{++}. *Phys. Rev. Lett.* **121**, 052002 (2018). `1806.02744`.

[42] Aaij, R. *et al.* First observation of the doubly charmed baryon decay $\Xi_{cc}^{++} \longrightarrow \Xi_{c}^{+}\pi^{+}$. *Phys. Rev. Lett.* **121**, 162002 (2018). URL `https://link.aps.org/doi/10.1103/PhysRevLett.121.162002`.

[43] and, D. D. C. Charm and multi-charm baryon measurements via strangeness tracking with the upgraded ALICE detector. *EPJ Web of Conferences* **259**, 09004 (2022). URL https://doi.org/10.1051%2Fepjconf%2F202225909004.

[44] Bini, C. Data analysis in particle physics. URL https://www.roma1.infn.it/~bini/StatEPP_new.pdf.

[45] D'Anzi, B. Signal/background discrimination for the vbf higgs four lepton decay channelwith the cms experiment using machine learning classification techniques. URL https://agenda.infn.it/event/26762/contributions/135521/attachments/80974/105943/VBF_use-case_exercise.pdf.

[46] Arthur samuel: Pioneer in machine learning. URL http://infolab.stanford.edu/pub/voy/museum/samuel.html.

[47] Mcculloch, W. & Pitts, W. A logical calculus of ideas immanent in nervous activity .

[48] Matanga, Y., Djouani, K. & Kurien, A. Analysis of user control attainment in smr-based brain computer interfaces. *IRBM* **39**, 324–333 (2018). URL https://www.sciencedirect.com/science/article/pii/S1959031818300514.

[49] Arthur samuel: Pioneer in machine learning. URL https://scikit-learn.org/stable/modules/tree.html#id2.

[50] SIMÕES, A. C. A. Mineração de dados baseada em árvores de decisão para análise do perfil de contribuintes.

[51] Adams, T. S. *et al.* Gravitational-wave detection using multivariate analysis. *Physical Review D* **88** (2013). URL https://doi.org/10.1103%2Fphysrevd.88.062006.

[52] Han, J., Kamber, M. & Pei, J. Data mining concepts and techniques, third edition (2012). URL http://www.amazon.de/Data-Mining-Concepts-Techniques-Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8&qid=1366039033&sr=1-1.

[53] Cornell, A. S., Doorsamy, W., Fuks, B., Harmsen, G. & Mason, L. Boosted decision trees in the era of new physics: a smuon analysis case study. *Journal of High Energy Physics* **2022** (2022). URL https://doi.org/10.1007%2Fjhep04%282022%29015.

[54] Woodruff, K. Introduction to boosted decision trees and hands-on tutorial¶. URL https://indico.fnal.gov/event/15356/contributions/31377/attachments/19671/24560/DecisionTrees.pdf.

[55] Abualdenien, J. & Borrmann, A. Ensemble-learning approach for the classification of levels of geometry (log) of building elements. *Advanced Engineering Informatics* **51**, 101497 (2022).

[56] Ananya *et al.* O2: A novel combined online and offline computing system for the alice experiment after 2018. *Journal of Physics: Conference Series* **513**, 012037 (2014). URL https://doi.org/10.1088/1742-6596/513/1/012037.

[57] Sjöstrand, T. *et al.* An introduction to PYTHIA 8.2. *Computer Physics Communications* **191**, 159–177 (2015). URL https://doi.org/10.1016%2Fj.cpc.2015.01.024.

[58] Brun, R., Bruyant, F., Maire, M., McPherson, A. C. & Zanarini, P. GEANT3 (1987).

[59] de Favereau, J. *et al.* DELPHES 3: a modular framework for fast simulation of a generic collider experiment. *Journal of High Energy Physics* **2014** (2014). URL https://doi.org/10.1007%2Fjhep02%282014%29057.