Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea in Fisica

Machine learning techniques for heavy-flavour baryon production measurements at the LHC

Relatore:
Prof. Andrea Alici

Presentata da:
Marco Cruciani

Anno Accademico 2020/2021

Abstract

Il quark-gluon plasma (QGP) è uno stato della materia previsto dalla cromodinamica quantistica. L'esperimento ALICE a LHC ha tra i suoi obbiettivi principali lo studio della materia fortemente interagente e le proprietà del QGP attraverso collisioni di ioni pesanti ultra-relativistici. Per un'esaustiva comprensione di tali proprietà, le stesse misure effettuate su sistemi collidenti più piccoli (collisioni protone-protone e protone-ione) sono necessarie come riferimento. Le recenti analisi dei dati raccolti ad ALICE hanno mostrato che la nostra comprensione dei meccanismi di adronizzazione di quark pesanti non è completa, perché i dati ottenuti in collisioni pp e p-Pb non sono riproducibili utilizzando modelli basati sui risultati ottenuti con collisioni e^+e^- ed ep. Per questo motivo, nuovi modelli teorici e fenomenologici, in grado di riprodurre le misure sperimentali, sono stati proposti. Gli errori associati a queste nuove misure sperimentali al momento non permettono di verificare in maniera chiara la veridicità dei diversi modelli proposti. Nei prossimi anni sarà quindi fondamentale aumentare la precisione di tali misure sperimentali; d'altra parte, stimare il numero delle diverse specie di particelle prodotte in una collisione può essere estremamente complicato. In questa tesi, il numero di barioni Λ_c^+ prodotti in un campione di dati è stato ottenuto utilizzando delle tecniche di machine learning, in grado di apprendere pattern e imparare a distinguere candidate di segnale da quelle di fondo. Si sono inoltre confrontate tre diverse implementazioni di un algoritmo di Boosted Decision Trees (BDT) e si è utilizzata quella più performante per ricostruire il barione Λ_c^+ in collisioni pp raccolte dall'esperimento ALICE.

Contents

Introduction

The ALICE (A Large Ion Collider Experiment) experiment at the LHC (Large Hadron Collider) is dedicated to studying heavy-ion ultra-relativistic collisions, measuring their properties and comparing them to those of proton-proton and proton-ion collisions. Its main purpose is to study QCD matter and the medium known as Quark-Gluon Plasma (QGP). One way to probe the QGP and learn its properties is to observe heavy quarks that are produced in the early stages of the collision and propagate through the medium, interacting with it. Recent analyses of data taken from the ALICE experiment in pp and p-Pb collisions have measured the baryon over meson ratio Λ_c^+/D^0 and have shown that models tuned to e^+e^- and ep collisions considerably underestimate this ratio. This means that the fragmentation models previously used are incomplete, and new theories have been advanced to try to reproduce these new measurements. The current measurements are not precise enough to let us understand which theory best fits the data, so we will only be able to understand the mechanisms at play once more precise measurements are carried out.

Measuring the mentioned Λ_c^+/D^0 ratio is not easy. Λ_c^+ particles are hard to detect, they decay very fast, so we have to reconstruct them from their decay. The best way to approach this issue is to use machine learning algorithms, allowing to consider multiple event properties simultaneously. We trained and compared three different multivariate methods based on Boosted Decision Trees (BDT), to learn to distinguish signal from background candidates by using simulated signal over real background data from the ALICE experiment. We studied in particular which method provides the best performance in handling missing data in Decision Trees. Once we determined it, we were able to use it to reconstruct the number of Λ_c^+ baryons in our sample data of ALICE pp collisions.

1 Heavy-flavour hadronization in pp and heavy-ion collisions

1.1 The Standard Model

According to our current understanding of physics, there are four fundamental forces in nature: electromagnetic, weak, strong and gravitational. The Standard Model is the theory that describes the first three forces. It states that all matter is made out of three kinds of particles: quarks, leptons and mediators, shown in fig. 1.1 [1].

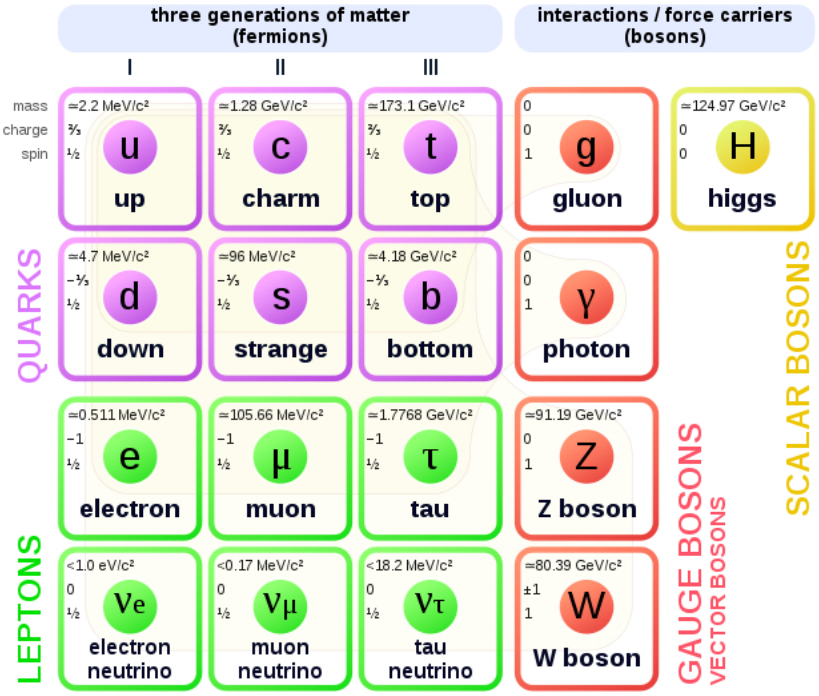

Figure 1.1: All particles of the Standard Model with their mass, charge and spin.

There are six different *flavors* of quarks with different electrical charge (Q), strangeness (S), charm (C), bottomness (B) and topness (T). Similarly, there are six different leptons classified by electrical charge (Q), electron number (L_e), muon number (L_μ) and tau number (L_τ). Both quarks and leptons are fermions with spin 1/2 which means they obey the Fermi-Dirac statistics.

Every interaction has a mediator: photons γ for the electromagnetic force, $W^\pm$ and Z^0 particles for the weak force and eight kinds of gluons g for the strong force. These particles are bosons with spin 1, and they obey Bose-Einstein statistics.

The Standard Model also includes the Higgs Boson, a scalar boson with spin 0 which is responsible for the mass of the weak force mediators.

The electromagnetic force acts between particles with electrical charge, the strong force affects particles with color charge i.e. quarks and gluons, while the weak force acts on all particles of the Standard Model.

All particles described have a corresponding anti-particle which has the same mass but opposite charges. Some neutral particles, such as the photon or Z^0 are their own anti-particle

This theory has had a great experimental success, however it's not a complete theory of all interactions: it doesn't include the theory of gravitation described by general relativity, whose effects are infinitesimal on a quantum scale. Also, there are some open questions, one of the main ones is the origin of the model's numerous parameters that can only be obtained experimentally and are not derived from within the theory itself, such as the particles' masses.

1.1.1 QCD: Quantum Chromodynamics

Quantum Chromodynamics (QCD) is the theory that describes the strong interaction that acts between quarks and gluons [2]. It is a non-abelian gauge theory with symmetry group SU(3). The strong interaction acts between particles with color, which is the QCD analogous of the electric charge of QED. Experimental evidence leads us to conclude that there are three different colors (and anti-colors) which are usually referred to as red, green, blue: r, g, b.

The interaction between quarks is mediated by gluons, which are massless particles with spin 1. Gluons themselves also carry color and anti-color: this means that, unlike QED photons, they can interact with each other directly.

Free quarks have never been observed: they always bind together to form mesons (quark anti-quark pairs) or baryons (three quarks) with total color charge equal to zero. If one were to try and separate a quark from an anti-quark, the system would have enough energy to create a new quark anti-quark pair, thus making it impossible to get an isolated quark. This phenomenon is known as *color confinement*.

Another important property of QCD is the *asymptotic freedom*. At low energy the intensity of the strong force is extremely high, leading to the tightly bound states of hadrons. On the other hand, quarks at high energy are weakly interacting: this is very important because in these conditions we can make perturbative approximations.

1.1.2 QGP: Quark-Gluon Plasma

Since quarks are confined inside a hadron, a useful description is given by the *bag model* [3]. In the bag model, quarks are massless particle in a bag of finite dimension and are infinitely massive outside the bag. In this model confinement is a result of the balance of the inward bag pressure and the stress given by the kinetic energy of the quarks. The gluons exchanged by the quarks are also confined in the bag. The total color charge of the bag has to be colorless.

This heuristic model gives us an intuitive understanding of why one might expect new phases of quark matter: if the pressure of the quarks inside is increased it will eventually be greater than the bag pressure, leading to a state where quarks and gluons are no longer bound, which is referred to as *Quark Gluon Plasma* (QGP). To increase the pressure we may increase the temperature or density. The approximate phase diagram of QCD matter is shown in fig. 1.2.

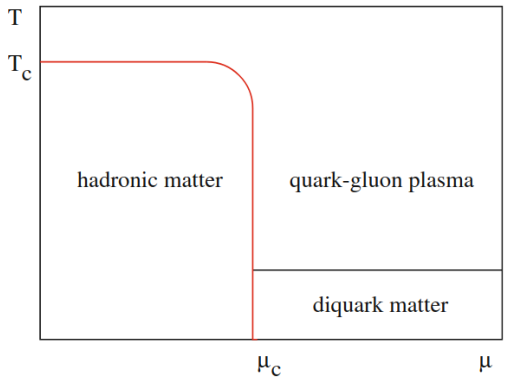

Figure 1.2: Phase diagram of QCD matter. Diquark matter is the formation of colored bosonic quark pairs, analogous to Cooper pairs of superconductors.

With a deep analysis using finite-temperature lattice QCD we can estimate the transition temperature T_C at $T_C \simeq 150 - 200\ MeV$. A very simple model can give remarkably close results [4].

Suppose that for an ideal gass of massless pions the pressure is given by the Stefan-Boltzmann formula:

$$P_\pi = 3\frac{\pi^2}{90}T^4$$

where the factor 3 takes into account the three charge states of the pion. For an ideal quark-gluon plasma this turns into:

$$P_{qg} = \left\{2 \times 8 + \frac{7}{8}(3 \times 2 \times 2 \times 2)\right\}\frac{\pi^2}{90}T^4 - B = 37\frac{\pi^2}{90}T^4 - B$$

where the first term in the curly brackets accounts for the degrees of freedom of the gluons, the second one the degrees of freedom of the quarks and B accounts for the non-zero pressure at $T = 0$ because of Fermi-Dirac statistics. By comparing these equations we get the temperature T_C at which the transition occurs:

$$T_C = \left(\frac{45}{17\pi^2}\right)^{1/4} B^{1/4}$$

and given that $B^{1/4} \simeq 200\ MeV$ from hadron spectroscopy, we get

$$T_C \simeq 150\ MeV$$

The energy densities are given by

$$\epsilon_\pi = \frac{\pi^2}{10}T^4 \qquad \epsilon_{qg} = 32\frac{\pi^2}{30}T^4 + B$$

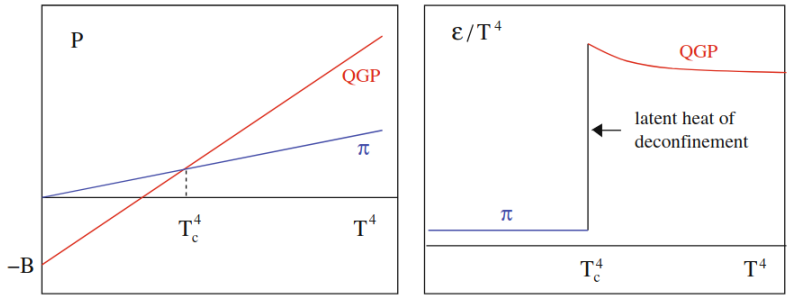

Figure 1.3: Pressure and energy density in a two-phase ideal gas model.

so this transition is first order. The energy density has a sudden increase given by the *latent heat of deconfinement* as is shown in fig. 1.3.

1.2 Heavy-flavour hadronization in pp collisions

Given the momentum scale Q^2 of the pp collision process, we can separate the high energy perturbative QCD production of a leading parton from the subsequent conversion in the hadronic state in low energy non-perturbative QCD. This technique is known as the factorization theorem [5].

The overall process is

$$p + p \to h + X$$

the hadron h is given by the decay of a parton c coming from an $a+b$ scattering of the protons' partons $a + b \to c + d$.

This theorem allows us to express the invariant cross-section of hadron production at mid-rapidity in pp collisions as

$$\frac{d\sigma_{pp}^{h}}{dyd^2p_T} = K\sum_{abcd}\int dx_a dx_b f_a(x_a,Q^2) f_b(x_b,Q^2)\frac{d\sigma}{d\hat{t}}(ab\to cd)\frac{D_{h/c}^{0}}{\pi z_c}$$

where the terms are:

- the *Parton Distribution Function* (PDF) $f_i(x_i, Q^2)$ of the species inside the proton
- the elementary perturbative QCD cross-section of leading c particle production from $a+b$ partonic scattering $\frac{d\sigma}{d\hat{t}}(ab\to cd)$
- the *Fragmentation Fucntion* (FF) $D_{h/c}^{0}$ which is a dimensionless object that gives us the probability that the parton c will hadronize by spraying soft gluons into the final hadron h carrying a fraction the fragmenting parton's momentum.

1.2.1 Parton Distribution Functions

Early deep inelastic scatterings of electrons against hadrons suggested that hadrons are not point-like particles and are instead composed of partons i.e. quarks and gluons. The typical well known quarks that compose an hadron, such as *udd* for the proton, are known as *valence quarks* whereas all the other partons are known as *sea partons*, which includes both gluons and sea quarks produced as virtual quark anti-quark pairs. In order to better understand PDFs let's consider the example of the proton.

We'll use $q^v(x)$ to indicate a valence quark probability density and $q^s(x)$ to indicate a sea quark probability density and $g(x)$ for the gluon probability density, where x is the fraction of the total momentum carried by q or g. We know that the valence quarks of a proton are uud, which gives us the condition:

$$\int_0^1 dx u^v(x) = 2, \qquad \int_0^1 dx d^v(x) = 1$$

The sea quarks are always produced in $q\overline{q}$ pairs, so they give a zero contribution to the baryon number

$$\int_0^1 dx[u^s(x) - \overline{u}^s(x)] = 0, \qquad \int_0^1 dx[d^s(x) - \overline{d}^s(x)] = 0$$

and the same is valid for s^s, c^s, b^s and t^s.

The total momentum carried by all the partons must add up to the proton momentum, which means that:

$$\int_0^1 dx x[u^v(x) + d^v(x) + \sum_q (q^s + \overline{q}^s)] = 1$$

Heavy quarks are included but are only active at scales $Q > m_q$. It's interesting to note that the gluon term by itself carries about half of the total momentum. Fig. 1.4 shows the proton distribution functions at $Q^2 = 10\ GeV^2$.

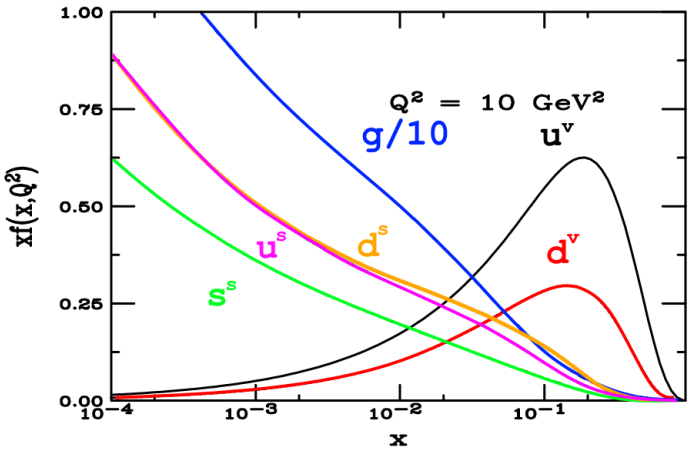

Figure 1.4: The CTEQ6M parametrization of the proton PDF at $Q^2 = 10\ GeV^2$. $xf(x)$ is the parton momentum distribution. The gluon distribution is multiplied by 0.1 for better visualization.

1.2.2 Fragmentation Functions

In order to better understand fragmentation functions, we will consider the simple example of an electron positron annihilation to produce a quark anti-quark pair [6]

$$e^- e^+ \rightarrow q\overline{q}$$

If the center-of-mass energy of the e^+e^- collision is Q, the electron beam energy is $E_{beam} = Q/2$. The produced quark has energy E_q equal to the beam energy. So if the final hadron has energy E_h, it will carry a fraction z of the quark energy given by

$$z = \frac{E_h}{E_q} = \frac{2E_h}{Q}$$

The differential cross-section for inclusive hadron production as a function of z is:

$$\frac{d\sigma(e^+e^- \to hX)}{dz} = \sum_q \sigma(e^+e^- \to q\overline{q})[D_q^h(z) + D_{\overline{q}}^h(z)]$$

which is basically an application of the factorization theorem without the PDF since electrons are fundamental particles.

The fragmentation function $D_q^h(z)$ represents the probability that the final-state hadron h carries a fraction z of the initial quark momentum. The sum of the energies of all produced hadrons has to add up to give the energy of the parent quark, so we have the condition:

$$\sum_h \int_0^1 dz z D_q^h(z) = 1$$

(the same is required for $\overline{q}$). The multiplicity of h is given by:

$$\sum_q \int_{z_{min}}^1 dz[D_q^h(z) + D_{\overline{q}}^h(z)] = n_h$$

where z_{min} is the threshold energy of producing a hadron of mass m_h, $z_{min} = 2m_h/Q$.

The fragmentation functions can have different parametrizations. In any case the parameters are obtained experimentally by fitting the wide range of data available for e^+e^- collisions. These functions are thought to be universal i.e. once calculated for e^+e^- collisions they should also be appliable to other cases such as ep, pp and $p\overline{p}$.

1.3 Heavy-flavour hadronization in A-A collisions

The first observations of hadron production in heavy-ion collisions showed that hadronization occurs differently compared to the vacuum fragmentation. Some models which tried to explain this difference make use of a mechanism called *recombination* or *coalescence*. In fragmentation, the initial momentum is distributed among fragments, whereas in recombination two or three comoving partons produce a hadron with transverse momentum given by the sum of their momenta, as shown in fig. 1.5.

Calculating the effects of recombination in heavy-ion collisions is quite complicated because we cannot write a simple wavefunction of the partons in the QGP.

The probability of finding two or three partons close in the phase-space decreases as momentum increases, so coalescence becomes less important at higher p_T where fragmentation is the dominant mechanism. Also, recombination effects should be more important in central collisions, while fragmentation should play a bigger role in peripheral collisions.

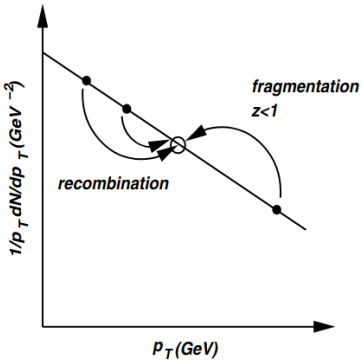

Figure 1.5: The competing mechanisms of recombination and fragmentation can lead to the same p_T final state hadron.

1.4 Experimental results

1.4.1 Charmed hadrons in pp and p-Pb collisions at $\sqrt{s} = 5.02\ TeV$

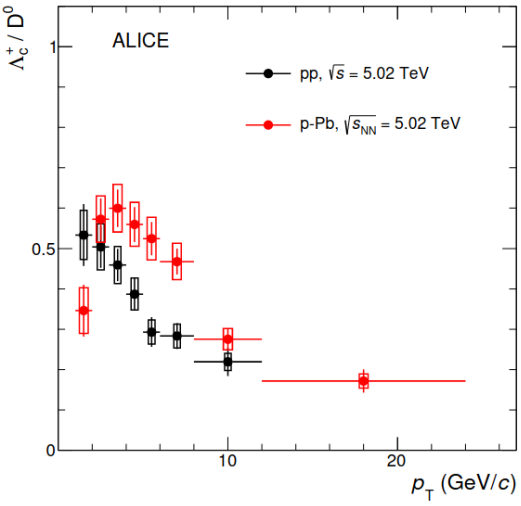

Figure 1.6: Λ_c^+/D^0 ratio as a function of p_T in pp and p-Pb collisions at $\sqrt{s} = 5.02\ TeV$.

Fig. 1.6 shows the Λ_c^+/D^0 ratio as a function of p_T in pp and p-Pb collisions at $\sqrt{s} = 5.02\ TeV$ measured with the ALICE detector [7]. Both the ratios show a decreasing trend for $p_T > 2\ GeV/c$. The ratios have some differences but are qualitatively consistent with each other. The values of the p_T-integrated Λ_c^+/D^0 ratios are ~ 0.51 and ~ 0.43 for pp and p-Pb respectively and these values are consistent with each other within the experimental uncertainties. If we compare

these ratios with those calculated from e^+e^- or e^-p collisions we see that the ratios are enhanced by a factor of about $2-5$, indicating that the hadronization mechanisms must be different.

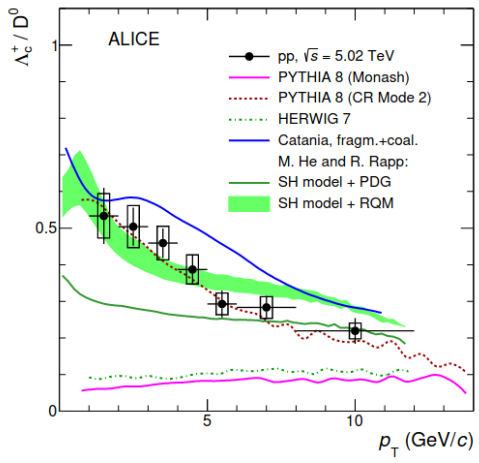

Figure 1.7: Λ_c^+/D^0 ratio as a function of p_T in pp and collisions at $\sqrt{s} = 5.02\ TeV$ compared to different models.

Fig. 1.7 shows the Λ_c^+/D^0 as a function of p_T in pp collisions at $\sqrt{s} = 5.02\ TeV$ compared to several model predictions.

- PYTHIA 8 (Monash) and HERWIG 7 are both Monte Carlo (MC) generators implementing fragmentation processes tuned on charm production measured from e^+e^- collisions. They predict a ratio of about 0.1 with almost no p_T dependence. These models underestimate the data at low p_T by a factor of $5-10$ while at high p_T by a factor of 2. This may support the hypothesis that fragmentation mechanisms dominate at high p_T.

- PYTHIA 8 (CR) is an MC generator that implements *color reconnection* [8]. This model is based on the string model for hadronization but includes additional 'junctions' which fragment into baryons, thus increasing the baryon production.

- Catania model assumes that QGP is formed in pp collisions and hadronization occurs via both coalescence and fragmentation.

- Statistical Hadronization models (SH) calculate branching fractions of charm quarks based on thermal densities, therefore depending on the state mass and the spin-degeneracy factor. The SH + PDG model is based on the currently measured particles of the Particle Data Group (PDG) while the SH + RQM

model uses additional excited baryon states which have not been measured but are assumed to exist in the Relativistic Quark Model (RQM).

1.4.2 Charmed hadrons in pp collisions at $\sqrt{s} = 13\ TeV$

ALICE was the first experiment to measure $\Sigma_c^{0,++}(2455)$ charmed hadrons in hadronic collisions [9]. The $\Sigma_c^{0,+,++}$ baryon triplet with isospin $I = 1$ is the partner of the Λ_c^+ baryon ($I = 0$). It decays with a branching ratio of $\sim 100\%$ to Λ_c^+, since it's the only strong decay allowed. The three isospin states are assumed to be equally produced, so the published data reports $3/2 \times \Sigma_c^{0,++}$ to indicate the assumed overall number of produced $\Sigma_c^{0,+,++}$.

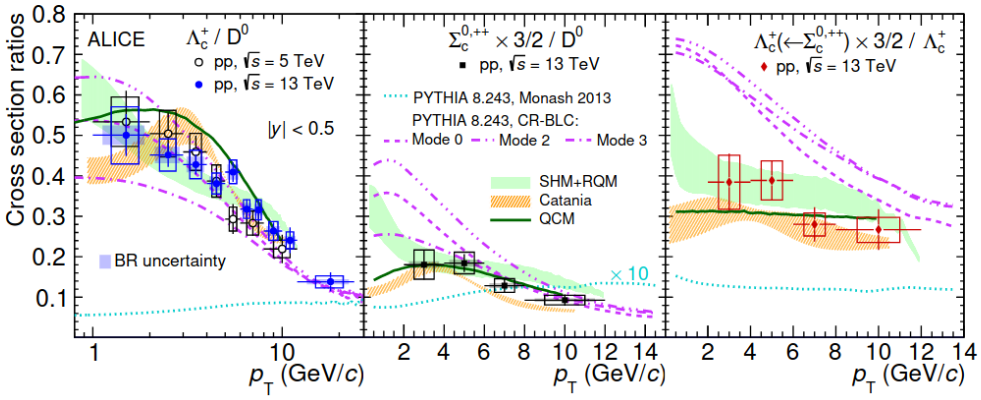

Figure 1.8: charmed hadron cross-section ratios: Λ_c^+/D^0 (left), $\Sigma_c^{0,+,++}/D^0$ (middle) and $\Lambda_c^+ \leftarrow \Sigma_c^{0,+,++}/\Lambda_c^+$. The PYTHIA Monash 2013 curve is scaled by 10 in the middle panel.

Fig. 1.8 (left) shows the Λ_c^+/D^0 ratio as a function of p_T in pp collisions compared with different model expectations. The values measured at $\sqrt{s} = 13\ TeV$ are compatible within uncertainties with those measured at $\sqrt{s} = 5.02\ TeV$. Most of the models used for comparison were discussed above. Quark (re-)combination mechanism (QCM) is a model that employs recombination mechanisms in which charm quarks form hadrons by combining with equal-velocity light quarks.

Fig. 1.8 (middle) shows the $\Sigma_c^{0,+,++}/D^0$ ratio, which is close to 0.2 for low p_T and close to 0.1 for high p_T, however uncertainties are such that we cannot make definitive conclusions about p_T dependence. This data shows that PYTHIA 8 Monash severely underestimates the ratio, while the other models give a generally good agreement.

Fig. 1.8 (right) show the fraction of Λ_c^+ coming from $\Sigma_c^{0,+,++}$. The p_T-integrated value is ~ 0.39, which is significantly different from measurement of the same ratio

in e^+e^- collisions ($\sim$ 0.17) and from PYTHIA 8 Monash simulations ($\sim$ 0.13). This larger feed-down from $\Sigma_c^{0,+,++}$ partially explains the increase in the measured Λ_c^+/D^0 ratio.

The results showed, in both $\sqrt{s} = 5.02\ TeV$ and $\sqrt{s} = 7\ TeV$, indicate that a pure fragmentation model based on the FFs measured from e^+e^- collisions is insufficient. There are different models all giving a reasonably close description of the data points by hypothesising different mechanisms. The data we have so far does not allow us to give more credit to one of these models. Further higher-precision measurements may shed more light about which of these mechanisms best describe the experimental data.

2 The ALICE experiment

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is the world's largest and most powerful particle accelerator, which was first started up in 2008. It lies in a tunnel with a circumference of 27 kilometers near Geneva [10]. Unlike fixed-target particle accelerators, where a beam collides with a stationary target, in the LHC two high-energy particle beams travel at relativistic speed before colliding, allowing to reach much higher energies of up to 6.5 TeV per beam [11]. To avoid other collisions, the beam pipes in which the particles travel are kept at ultra-high vacuum with pressures below 10^{-13} atmospheres. The particles are guided by a strong magnetic field which is produced by coils made of a special electric superconducting cable. To maintain the superconducting state of the cables they need to be kept at the extremely low temperatures of -271.3 °C with a distribution system of liquid helium. The beams inside the LHC collide in four different locations which correspond to the position of the main four particle detectors: ATLAS, CMS, ALICE and LHCb, as shown in fig. 2.1.

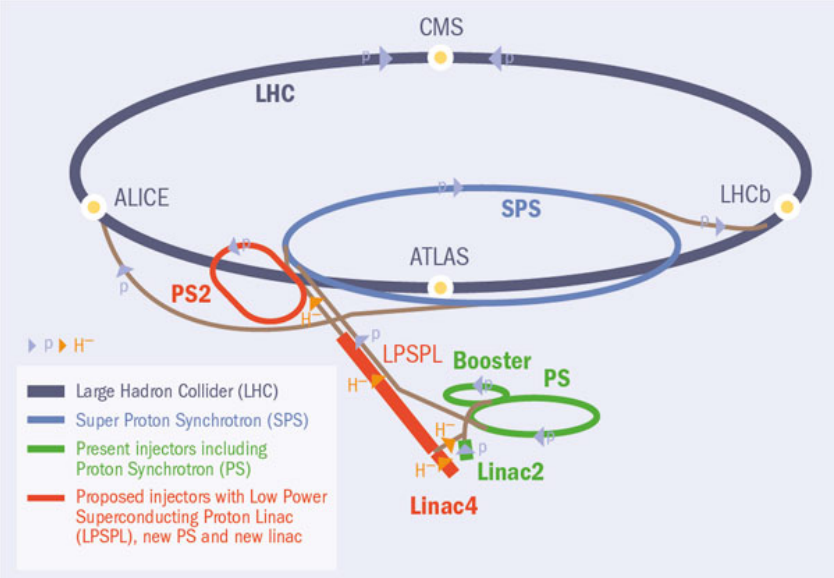

Figure 2.1: Essential scheme of the LHC layout [12].

2.1.1 LHC magnets

Particles accelerated in the LHC are usually protons or lead ions. Before reaching the main ring, the particles have to be sped up in a series of different kinds of

linear and circular accelerators. In order to keep the particles together and send them along these complex paths, several kinds of powerful electromagnets are used. These coils have to be in a superconducting state to be able to withstand the extreme currents needed to reach magnetic fields as high as 8.3 tesla [13].

Dipole magnets are one of the most complex parts of the LHC. There are 1232 main dipoles, each one is 15 meters long and weighs about 35 tonnes. These superconducting magnets are an important part of the LHC's design: to reach the same energies with 'normal' magnets, the ring would have to be 120 kilometers long. A cross-section of an LHC dipole magnet is shown in fig. 2.2.

Particles should be very close together to increase the chances of collisions. Quadrupole magnets are the ones responsible for keeping them in a tight beam: they are magnets with four magnetic poles symmetrically arranged around the beam pipe.

When the particle beams are about to enter the detectors a further set of three quadrupole magnets, called inner triplets, are used to tighten the beam even further making it as narrow as 16 micrometers across.

To dispose of the particles, they are deflected along a straight line to the beam dump, where they collide with a block of concrete and graphite.

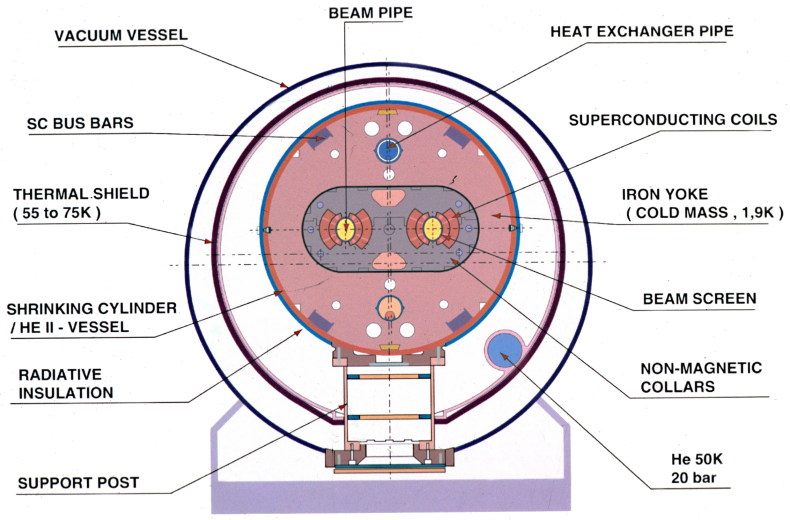

Figure 2.2: Cross-section of an LHC superconducting dipole magnet [14].

2.2 The ALICE detector

ALICE stands for *A Large Ion Collider Experiment*. It's a detector dedicated to studying strongly interacting matter by colliding heavy lead ions. These collisions

generate temperatures high enough to 'melt' protons and neutrons and create quark-gluon plasma. The purpose of the experiment is to study key issues of QCD such as the properties of QGP, understanding color confinement and chiral-symmetry restoration [15].

The overall dimensions of the detector are $16 \times 16 \times 26\ m^3$, and it weighs approximately 10000 tonnes. It consists of a central barrel part (shown in fig. 2.3) dedicated to measuring hadrons, electrons and photons and a forward muon spectrometer. The central part covers polar angles from 45° to 135° and is inside a large solenoid magnet. From the inside out, the barrel contains:

- Inner Tracking System (ITS)
- Time-Projection Chamber (TPC)
- Transition Radiation Detector (TRD)
- Time-of-Flight (TOF)
- High Momentum Particle Identification Detector (HMPID)
- two calorimeters, PHOS and EMCal

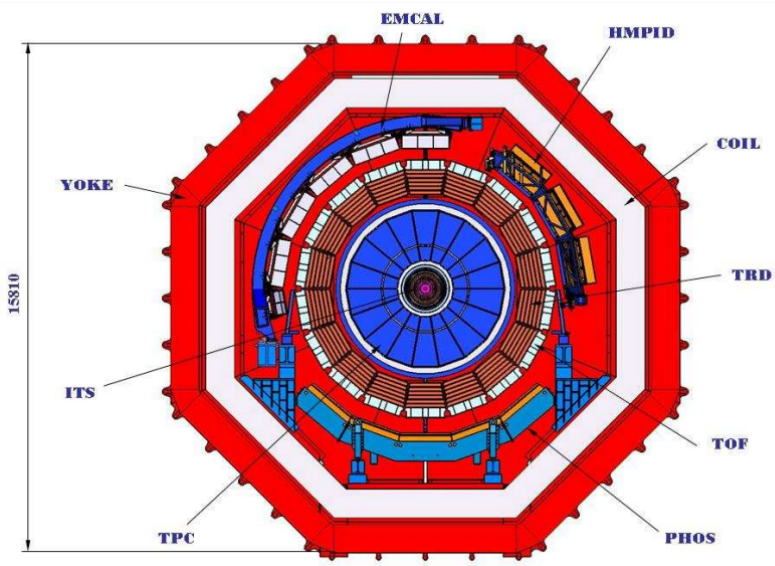

Figure 2.3: Cross-section of the central barrel of the ALICE apparatus.

All detectors except HMPID, PHOS and EMCal cover the full azimuth.

The muon arm consists of an arrangement of absorbers, a large dipole magnet and fourteen planes of tracking and triggering chambers. Furthermore, a number

of small and specialized detector systems are used for triggering or measuring global event characteristics.

The event time is measured by the T0 detector with very good precision (< 25 ps). The V0 detector is used as a minimum bias trigger and for rejection of beam-gas background.

The Alice Cosmic Ray Detector (ACORDE) is an array of large scintillators that trigger on cosmic rays for calibration and alignment purposes and also cosmic ray physics.

The Forward Multiplicity Detector (FMD) provides multiplicity information over a wide pseudo-rapidity range: it counts particles in rings of silicon strips detectors placed in three locations along the beam pipe. The Photon Multiplicity Detector (PMD) measures the multiplicity and spatial distribution of photons event-by-event.

The Zero Degree Calorimeter is a set of two compact calorimeters on either side of the machine tunnel at 116 m from the interaction point. Each ZDC set is made of two detectors, the ZN for spectator neutrons and ZP for spectator protons. They are used to measure and trigger on the impact parameter of the collision. Another pair of small electromagnetic calorimeters (ZEM) are installed on one side to improve centrality selection.

2.2.1 Innter Tracking System

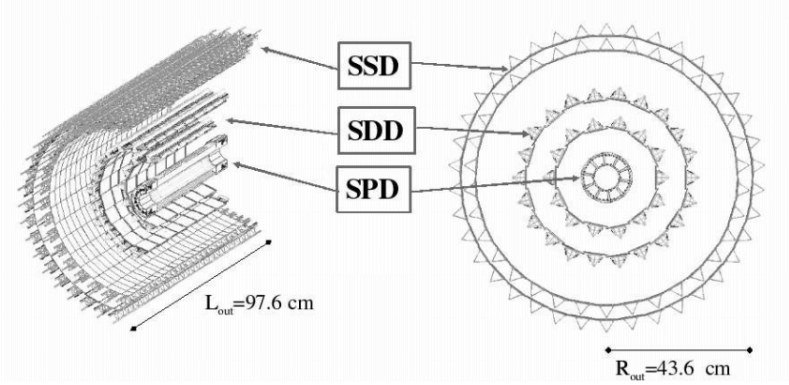

Figure 2.4: Layout of the Inner Tracking System.

The main tasks of the Inner Tracking System (ITS) are [15]:

- localizing the primary vertex with a resolution better than 100 micrometers;

- reconstructing the secondary vertices from the decays of hyperons and D/B mesons;
- tracking and identifying particles with momentum below 200 MeV/c;
- improving the momentum and angle resolution for particles reconstructed by the TPC;
- reconstructing particles traversing dead regions of the TPC.

The ITS is coaxial with the beam pipe and consists of six cylindrical layers of silicon detectors with different radii varying from 4 to 43 cm, as shown in fig 2.4.

The innermost layers use Silicon Pixel Detectors (SPD) and Silicon Drift Detectors to deal with the predicted high density of particles produced in heavy-ion collisions (as many as 50 per cm^2). The two outer layers are expected to read less than one particle per cm^2 and are equipped with double-sided Silicon micro-Strip Detectors (SSD)

2.2.2 Time-Projection Chamber

The Time-Projection Chamber is the main tracking detector of the central barrel. Its purpose is to provide charged particle momentum measurement with good two-track separation, particle identification and vertex determination. The covered p_T range goes from about 0.1 GeV/c to 100 GeV/c [15].

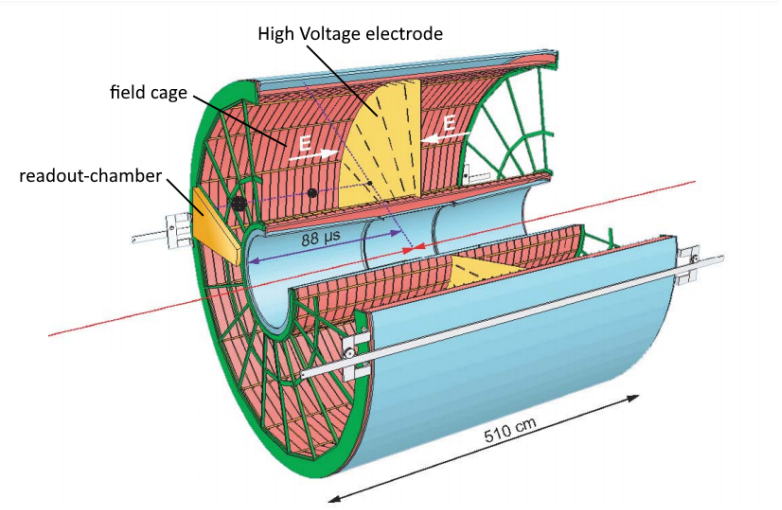

Figure 2.5: Layout of the Time Projection Chamber. Adtapted from [16].

The TPC has a cylindrical shape, it extends on the beam direction by about 500 cm, while the inner and outer radii are 85 cm and 250 cm, respectively. The detector itself is made of a large field cage, filled with a gas mixture of $Ne/CO_2/N_2$ (90/10/5). At the center of the cage is a high-voltage electrode and two opposite axial potential dividers to create a highly uniform electrostatic field. The field cage is operated at high voltage gradients of about 400 V/cm with about 100 kV at the central electrode. The layout of the TPC is shown in fig. 2.5.

When a charged particle traverses the gas, it will leave behind a long trace of ionized gas. The trace will have different properties based on the charge and momentum of the particle. This ionization traces move to reach either of the two end plates on the sides where they induce signals on read-out detectors equipped with Multi-Wire Proportional Chambers (MWPCs). The hit positions at the endcaps together with accurate measurements of the arrival time allow for a 3D reconstruction of the complete trajectory of all charged particles traversing the TPC [17].

2.2.3 Transition Radiation Detector

The Transition Radiation Detector (TRD) allows the identification of electrons and positrons with momenta above 1 GeV/c (electrons with lower momentum can be identified by the TPC). Combining its data with ITS and TPC it's possible to study the production of light and heavy vector-meson resonances in both pp and Pb-Pb collisions [15].

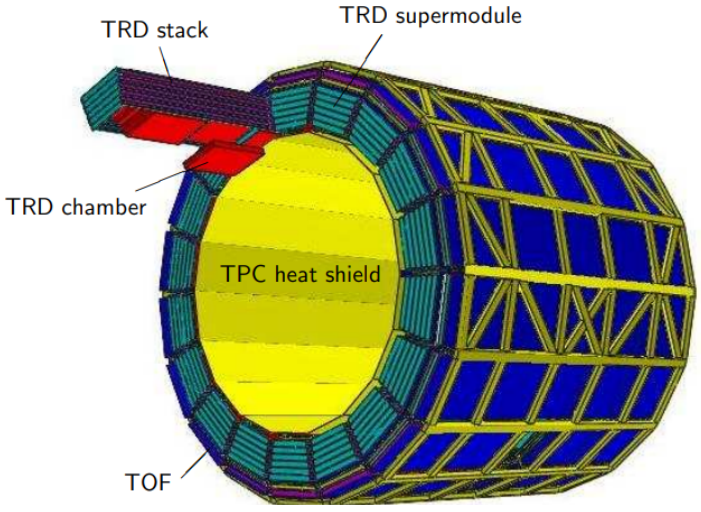

Figure 2.6: Schematic drawing of the TRD layout. On the outside the TRD is surrounded by the TOF (dark blue).

The TRD consists of six layers of Xe/CO_2-filled time expansions wire chambers after a composite foam and fibre radiator. Ionizing radiation produces electrons in the counting gas. In addition to this, particles with very high speed ($\gamma \sim 1000$) will produce transition radiation in the X-ray energy range. These X-ray photons are converted to electrons by the high-Z counting gas. All electrons drift towards the anode wires and induce signals on the readout pads.

The detector consists of 18 super modules containing 30 modules each, for a total of 540 individual read-out modules. The overall length of the super module is 7.8 m, and it weighs about 1650 kg. A drawing of its layout is shown in fig. 2.6.

2.2.4 Time-of-flight

The Time-Of-Flight (TOF) detector is a large area array for particle identification. It covers a pseudo-rapidity range of $|\eta| \lesssim 0.9$ and provide 3σ π/K and K/p separation for momenta up to 2.5 GeV/c and 4 GeV/c, respectively. The TOF coupled with the ITS and TPC can identify large samples of pions, kaons and protons [15].

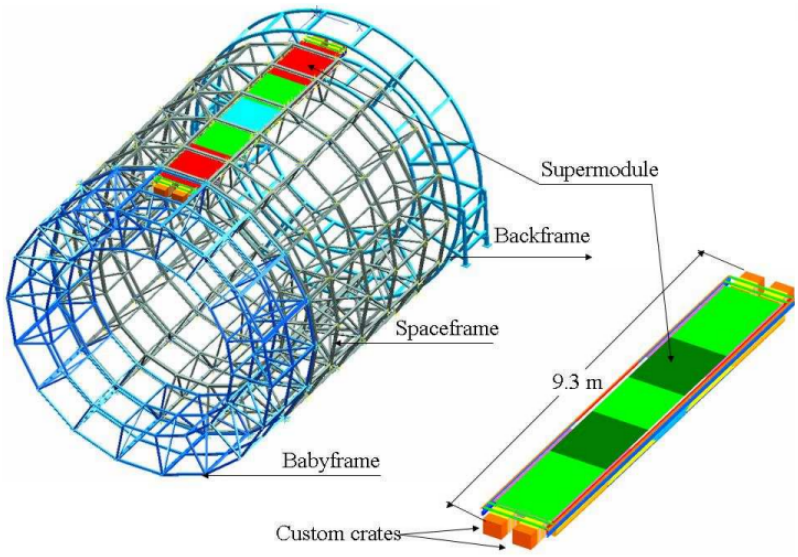

Figure 2.7: Schematic drawing of one TOF supermodule, consisting of 5 modules. 18 supermodules are placed around the frame to cover the full range of ϕ.

The detector covers a cylindrical surface of polar acceptance $|\theta - 90°| < 45°$. It's divided in 18 sectors around ϕ and 5 modules along the z direction, as shown in fig. 2.7. The whole device is inside a cylindrical shell with inner radius of 370 cm and outer radius of 399 cm.

A very large area needs to be covered by this detector, so the best choice is a gaseous detector. The basic unit of the TOF is the Multi-gap Resistive-Plate Chamber (MRPC). Every module of the TOF consists of a group of MRPC strips in a box that seals the gas volume. This technology is able to maintain a high and uniform electric field over the whole volume, so that any charged particle going through the medium will start a gas avalanche process and generate the observed signals on the electrodes.

2.2.5 High-Momentum Particle Identification Detector

The High-Momentum Particle Identification Detector (HMPID) is dedicated to inclusive measurements of identified hadrons at $p_T > 1\ GeV/c$. Its purpose is to enhance the particle identification capabilities of ALICE beyond the intervals that ITS, TPC and TOF can reach. In addition to this it's also able to identify light nuclei and anti-nuclei at high p_T in the central rapidity range. The detector consists of seven modules covering about 11 m^2 mounted on a cradle fixed at two o'clock position [15].

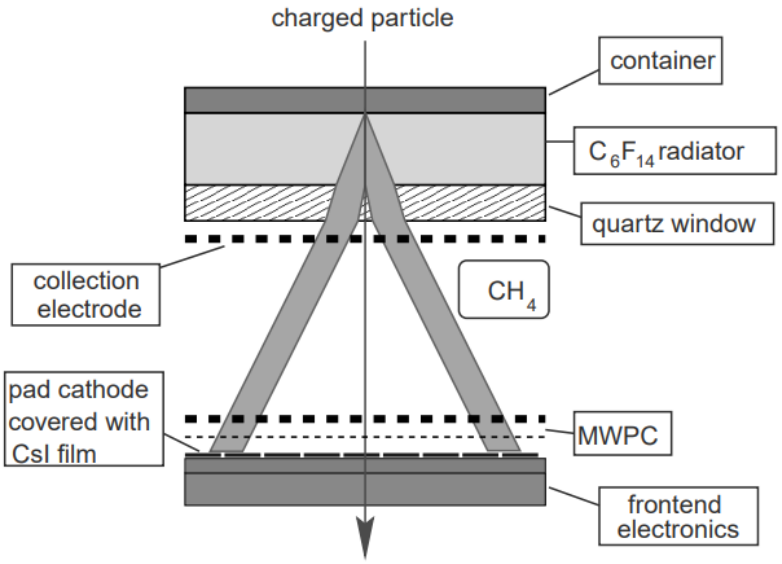

Figure 2.8: Schematic drawing of the HMPID working mechanism.

The HMPID is a Ring Imaging Cherenkov (RICH) detector. It exploits the Cherenkov radiation emitted by charged particles when going through a medium faster than the speed of light in that medium. The working principle of the detector is shown in fig 2.8. The Cherenkov cone refracts out of a layer of C_6F_{14} and expands in a volume of methane (CH_4) until it reaches the MWPC photon detector.

2.2.6 Calorimeters

The Photon Spectrometer (PHOS) is a high-resolution electromagnetic calorimeter. It allows for testing the thermal and dynamical properties of the initial phase of the collision by measuring low p_T direct photons and also studies jet quenching by measuring high p_T π^0 and γ-jet correlations [15].

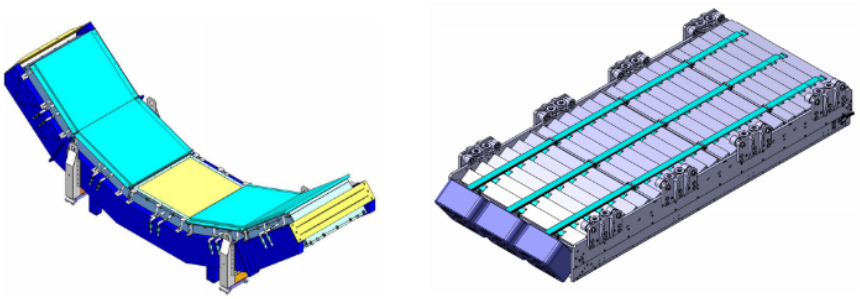

Figure 2.9: Left: 5 PHOS modules configuration. Right: an EMCAL super module.

The detector is structured in a single arm consisting of a highly segmented electromagnetic calorimeter (PHOS) and a Charged-Particle Veto (CPV) detector. The detector is divided into five independent PHOS+CPV modules, as shown in fig 2.9 (left). It's positioned at the bottom of the ALICE setup at 460 cm from the interaction point. It covers a range of pseudo-rapidity of $|\eta| < 0.12$ and 100° in azimuthal angle.

The EMCAL is a large Pb-scintillator sampling calorimeter with cylindrical geometry. It was constructed with the aim to explore the details of jet quenching over the large kinematic range accessible in the collisions at the LHC.

It's located at a radius of about 450 cm from the beam line. It covers a pseudo-rapidity range of $|\eta| < 0.7$ and $\Delta\phi = 107°$. The detector consists of 10 'full-size' super modules, shown in fig 2.9 (right), and 2 'one-third size' super modules. Its design was limited by the physical constraint of available space and maximum supported weight.

2.2.7 Muon spectrometer

The muon detector allows to study the spectrum of heavy-quark vector-mesons resonances through the $\mu^+\mu^-$ decay channel. Measuring all of the quarkonia species with the same apparatus allows a direct comparison of the production rate as a function of different parameters like p_T or centrality [15].

The muon spectrometer's design was a compromise between acceptance and detector cost. It detects muons in the polar range $171° - 178°$ which corresponds to a pseudo-rapidity range of $-4.0 \leq \eta \leq -2.5$. The layout of the detector is shown in fig 2.10.

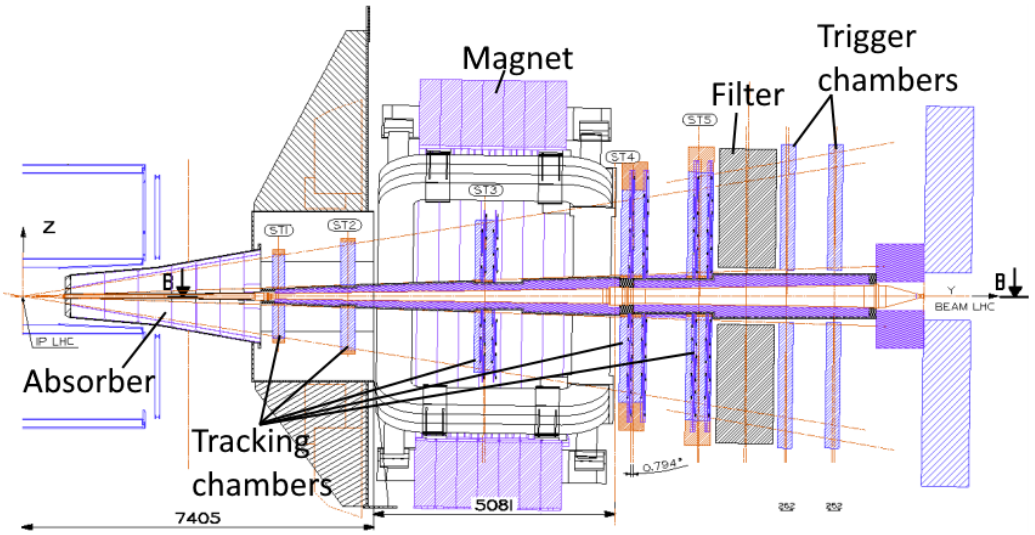

Figure 2.10: Longitudinal section of the muon spectrometer.

The spectrometer consists of the following components: a passive front absorber; a high-granularity tracking system; a large dipole magnet; a passive muon-filter wall followed by four planes of trigger chambers; an inner beam shield.

The front absorber is about 4.13 meters long and has conical geometry. It blocks hadrons and photons produced in the interaction vertices. It's mainly made of carbon and concrete to limit small-angle scattering and energy loss by traversing muons. The spectrometer is also shielded throughout its length by an absorber tube around the beam pipe. This tube (beam shield) is made of tungsten, lead and stainless steel. The muon filter is an additional protection needed for the trigger chambers, it's an iron wall about 1.2 meters in thickness.

The tracking chambers were designed to achieve a spatial resolution of about 100 μm and to operate at the hit density of about 5×10^{-2} which is expected in central Pb-Pb collisions. The tracking system covers a total area of about 100 m^2.

The trigger system consists of four planes of Resistive Plate Chambers (RPC) arranged in two stations, placed behind the muon filter. The total active area is about 140 cm^2. It has a spatial resolution better than 1 cm.

3 Λ_c^+ reconstruction with Boosted Decision Trees

3.1 Introduction

The study of charmed baryon production is a fundamental tool to verify the theoretical predictions of QCD and the properties of QGP.

As we discussed in section 1.4, measurements using the data collected by the ALICE experiment showed that charm hadron production in pp and p-Pb collisions is not well understood. Models using the fragmentation functions obtained from e^+e^- and ep collisions severely underestimate the baryon-over-meson ratios measured at the LHC energies, and new theories and mechanisms have been put forward to explain the measurements.

The study of the relative production of heavy-flavour hadron allows investigating the hadronization processes. In this respect, the Λ_c^+ baryon plays a fundamental role being the most abundantly produced baryon; thus, measuring the Λ_c^+ production cross-section relative to D^0 mesons provides insight into the hadronization mechanisms of charm quarks into baryons. In this work, we will focus on the reconstruction of Λ_c^+ produced through the $\Lambda_c^+ \to pK_S^0$ decay channel.

Λ_c^+ is a baryon with quark content udc and $I(J^P) = 0(1/2^+)$. Its mass and mean life are $(2286.46 \pm 0.14)\ MeV/c^2$ and $(2.024 \pm 0.031) \times 10^{-13}\ s$, respectively. The decay channel we consider $\Lambda_c^+ \to pK_S^0$, schematically represented in 3.1, has a branching ratio of $(1.59 \pm 0.08)\%$ [18].

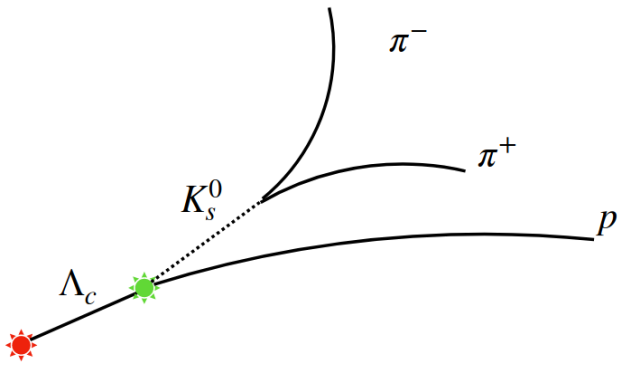

Figure 3.1: $\Lambda_c^+ \to pK_S^0$ decay graphical representation.

Trying to estimate the number of Λ_c^+ is not easy. The main issue is the low signal to background ratio. Another issue is that this baryon has a very short mean life, and it decays after $\sim 60\ \mu m$ while the ITS of ALICE has a spatial resolution of about 100 μm, which means that the particles coming from Λ_c^+ decay are basically seen as coming from the primary vertex.

A solution that allows us to get the most out of our large amount of data and tell signal and background apart is to employ machine learning techniques. We will use multivariate analysis techniques provided by the TMVA package.

3.2 TMVA

The Toolkit for Multivariate Analysis (TMVA) [19] provides a ROOT-integrated environment to process, evaluate and apply multivariate classification. These techniques use training events for which the desired output is known and determine the mapping function that describes a decision boundary (for classification) or an approximation of a functional behavior (for regression). Training and testing is performed with user data. A preanalysis calculates linear correlation coefficients of the input variables. TMVA provides performance data that allows to compare between different MVA methods using the same training and test data.

This work will compare three different techniques, all based on Boosted Decision Trees (BDT).

3.3 Boosted Decision Trees

A Decision Tree is a very simple classifier, that can be easily understood by looking at the example in fig. 3.2. Basically, a Decision Tree asks a series of binary questions about the input variables $\mathbf{x}$. This is equivalent to partitioning the input-variable space in several regions [20].

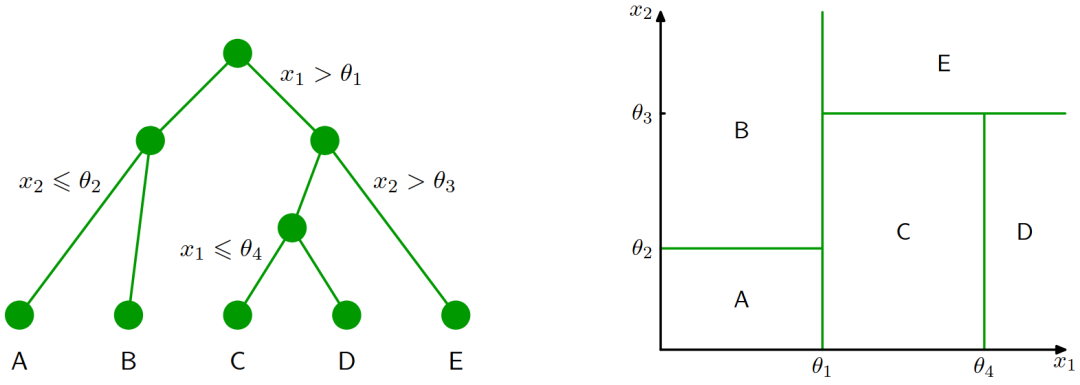

Figure 3.2: Scheme of a simple decision tree (left) and corresponding two-dimensional space partitioning (right).

For our purpose, the decision tree has to decide whether an event is signal or background, so the split choice at every node is determined to give the best separation between the two classes. Every leaf node at the bottom of the tree will

be labeled as signal or background based on the majority of events that end up in that node. A completed tree will work as a function $y(\mathbf{x})$ that outputs a number $[-1, 1]$ for every input-variable set $\mathbf{x}$, with 1 meaning the event is certain to be signal and -1 meaning that the event is certain to be background. Boosting is a powerful technique that combines multiple 'base' classifiers and can significantly improve the performance. Boosting techniques can even give good results even with *weak learners*, i.e. base classifiers that are barely better than random. The most widely used form of boosting is *AdaBoost*, which is the default choice for TMVA BDT, and it's the one used in this work.

The base classifiers are trained in sequence using a weighted form of the data that depends on the previous classifier's performance: points that are mislabeled by the current classifier will be given a greater weight when used to train the following one. Once all base classifiers $y_m(\mathbf{x})$ have been trained they're combined to give the final classifier which looks like:

$$Y_M(\mathbf{x}) = \sum_m \alpha_m y_m(\mathbf{x}), \qquad Y_M(\mathbf{x}) \in [-1, 1]$$

where the coefficients α_m are calculated to give greater weight to the accurate classifiers.

3.4 Data and input variables

The training sample for signal-like candidates was taken from pp collisions simulated with the PYTHIA8 [21] event generator. The presence of at least one Λ_c^+ decaying via the hadronic decay channel under consideration in each simulated event was required in order to maximize the number of candidates. The generated particles are then transported through the ALICE apparatus by using the GEANT3 package [22] via a simulation that reproduces the detector layout and the data-taking conditions.

The training sample for background data was taken from ALICE measurements of the LHC Run2 data taking (2016-2017-2018). This same data will be used for application after the algorithm is trained. To ensure we only select background data for the training phase, we use data points that give a reconstructed invariant mass which is outside an interval of 3σ around the known mass of Λ_c^+.

Machine learning (ML) is defined as the study of computer algorithms who can learn to mimic or to find patterns in the training data automatically through experience. In general, a ML algorithm takes input arrays of variables, and applies functions to them to sort them into the known categories. The choice of such input variables is of great importance because they have to allow the algorithm to learn how to separate the classes of events, in our case signal and background, in the most efficient way.

Using fig. 3.1 as a reference, we refer to the potential proton as 'bachelor' particle and to the potential K^0_S as V^0 particle. These are the variables we selected as input:

- **massK0S**: this is the invariant mass of the V^0 particle. It's calculated by finding two particles coming from the same vertex with opposite charge and reconstructing the mass of the particle that generated them, with energy and momentum conservation, assuming the particles' masses to be $m_\pi \simeq 138$ MeV/c^2. The expected value for K^0_S is ~ 497 MeV/c^2.
- **tImpParBach**: impact parameter of the bachelor particle, defined as the minimum distance from the bachelor track and the primary vertex, on the plane that is normal to the track.
- **tImpParV0**: impact parameter of the V^0 particle
- **ctK0S**: $c\tau$ of the V^0. The expected value for K^0_S is ~ 2.68 cm
- **cosPaK0S**: cosine of the pointing angle, i.e. the angle between the direction of V^0 and the line connecting the primary vertex to the secondary. It's expected to be close to 1.
- **CosThetaStar**: cosine of the angle between the direction of the bachelor particle in the frame of reference where Λ_c^+ is at rest and the direction of Λ_c^+ in the frame of reference of the laboratory. The decay should be isotropic, so a uniform distribution is expected for signal, whereas background should have values closer to 1 or -1
- **nSigmaTOFpr**: this is the probability that the bachelor particle is an actual proton according to the TOF detector. It's calculated by comparing the time a proton would take to reach the TOF detector to the time taken by the bachelor particle.
- **nSigmaTOFpi**: this is the probability that the bachelor particle is actually a pion according to the TOF detector.
- **nSigmaTOFka**: this is the probability that the bachelor particle is actually a kaon according to the TOF detector
- **nSigmaTPCpr, nSigmaTPCpi, nSigmaTPCka**: these are the probabilities that the bachelor particle is a proton, a pion or a kaon according to the TPC detector. This is calculated by comparing the dE/dx energy loss of the bachelor particle to the expected value for protons pions or kaons.

Figs. $3.3 - 3.8$ show the variables distribution for signal and background in all p_T ranges.

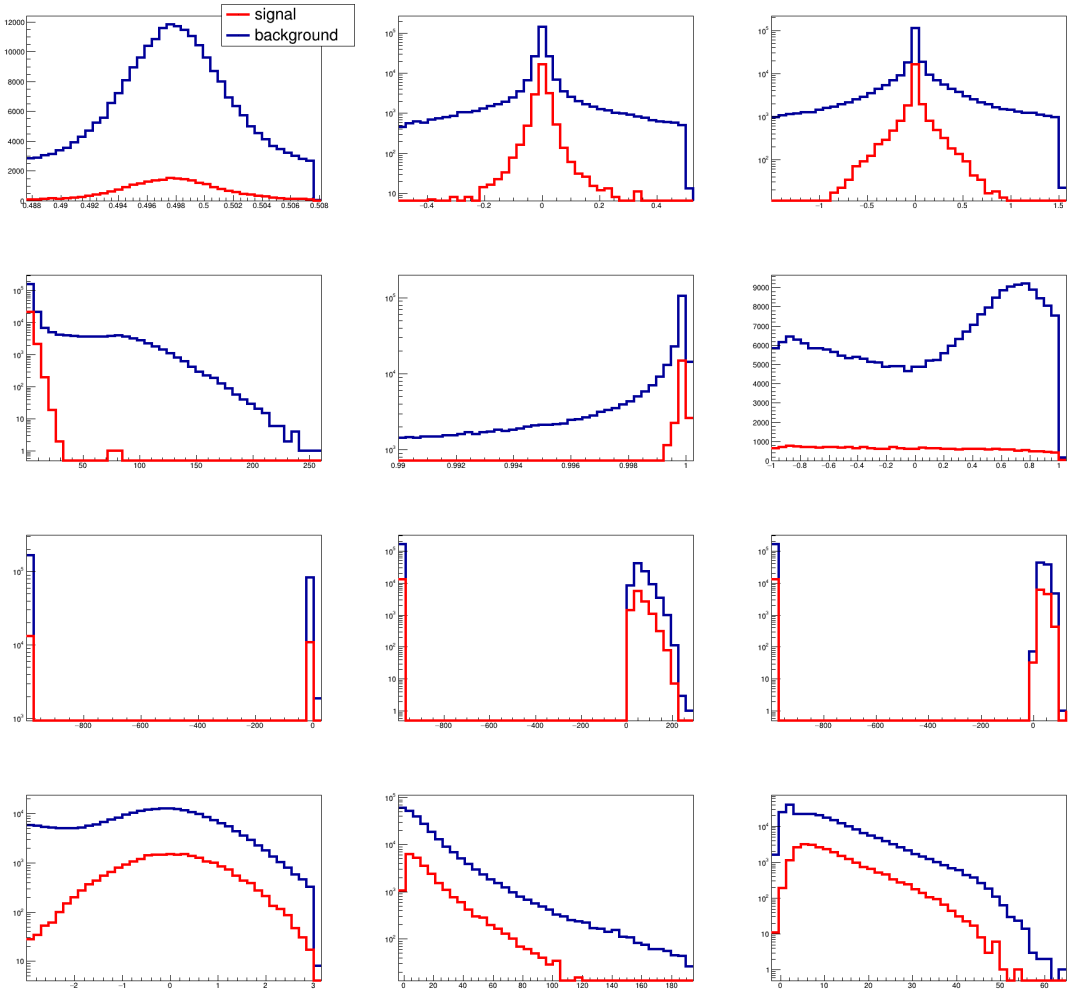

Figure 3.3: Input variables distribution in p_T range $[0, 1]$. From left to right, top to bottom: massK0S, tImpParBach, tImpParV0, CtK0s, cosPAK0S, CosThetaStar, nSigmaTOFpr, nSigmaTOFpi, nSigmaTOFka, nSigmaTPCpr, nSigmaTPCpi, nSigmaTPCka. Signal is red and background is blue.

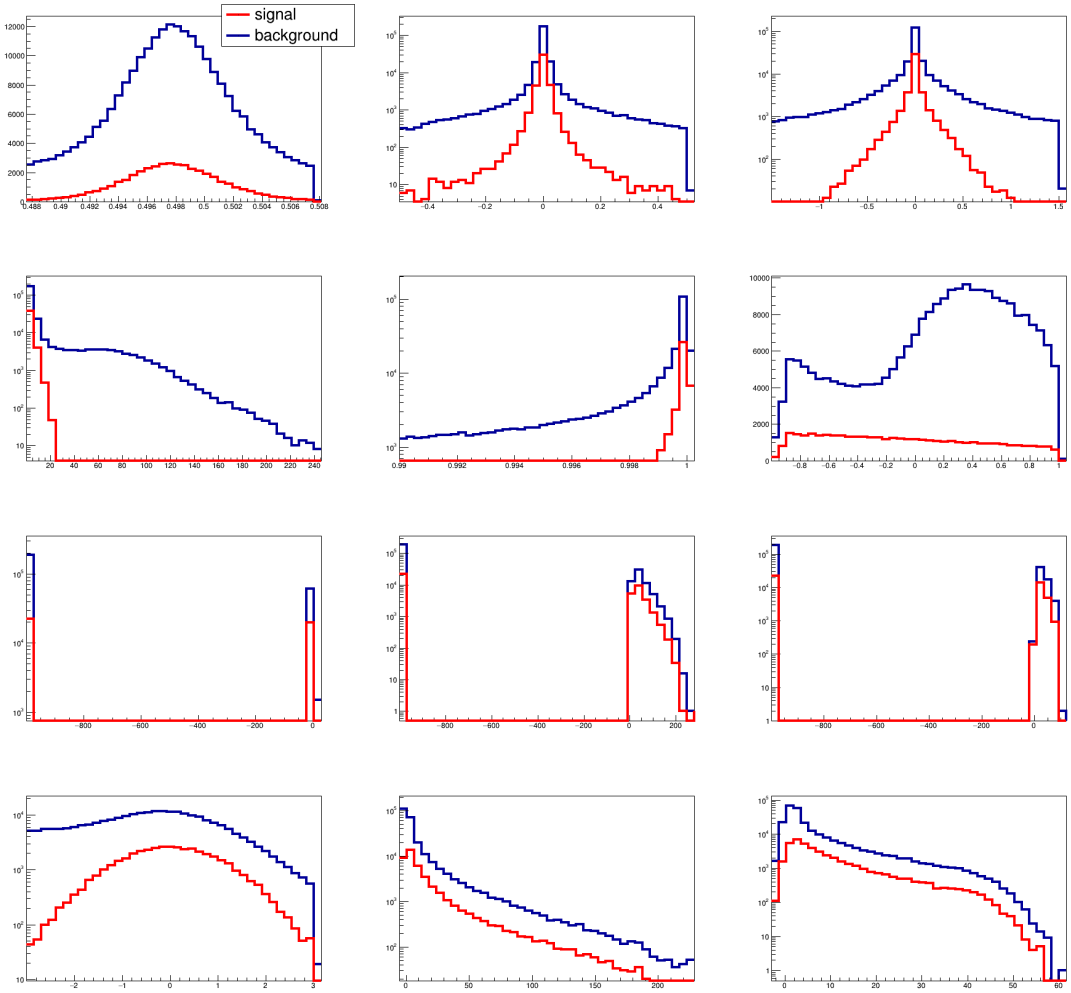

Figure 3.4: Input variables distribution in p_T range $[1, 2]$. From left to right, top to bottom: massK0S, tImpParBach, tImpParV0, CtK0s, cosPAK0S, CosThetaStar, nSigmaTOFpr, nSigmaTOFpi, nSigmaTOFka, nSigmaTPCpr, nSigmaTPCpi, nSigmaTPCka. Signal is red and background is blue.

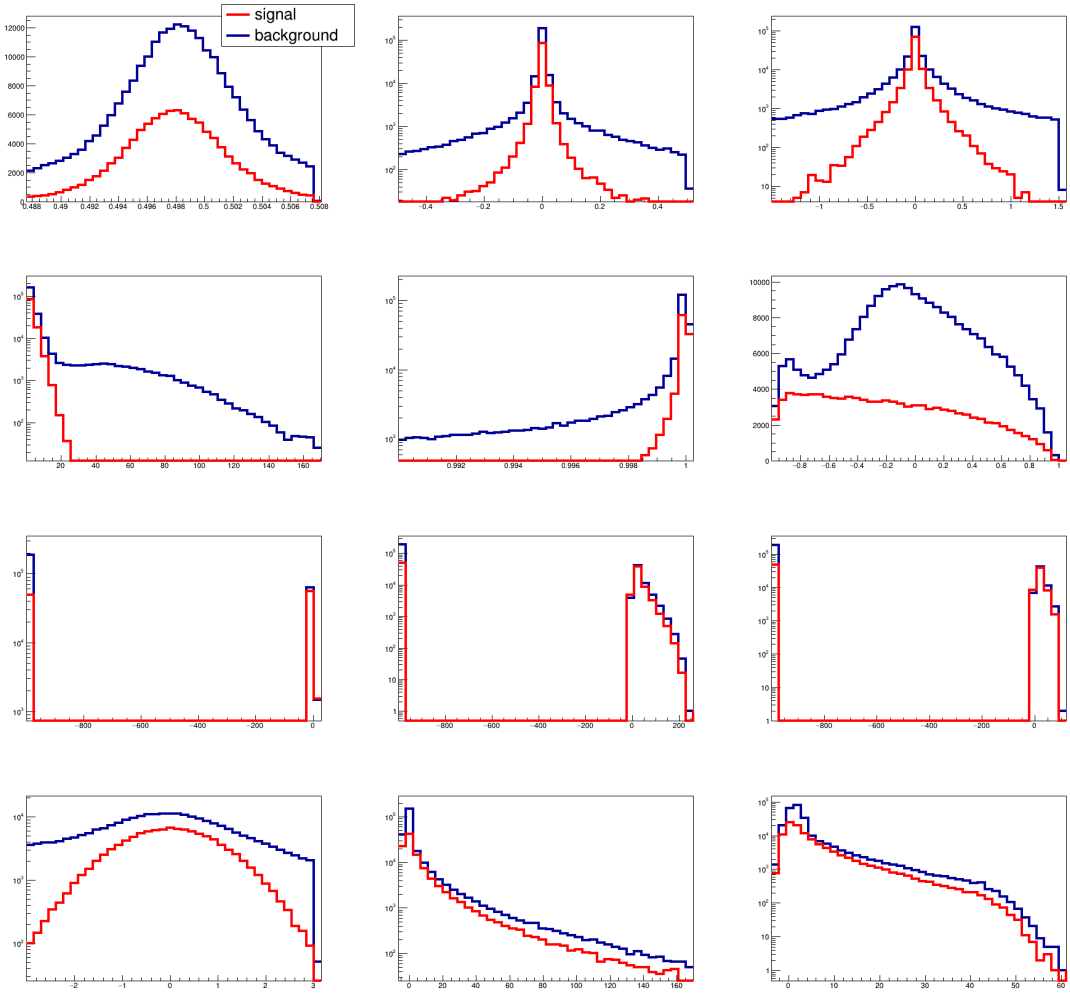

Figure 3.5: Input variables distribution in p_T range $[2, 4]$. From left to right, top to bottom: massK0S, tImpParBach, tImpParV0, CtK0s, cosPAK0S, CosThetaStar, nSigmaTOFpr, nSigmaTOFpi, nSigmaTOFka, nSigmaTPCpr, nSigmaTPCpi, nSigmaTPCka. Signal is red and background is blue.

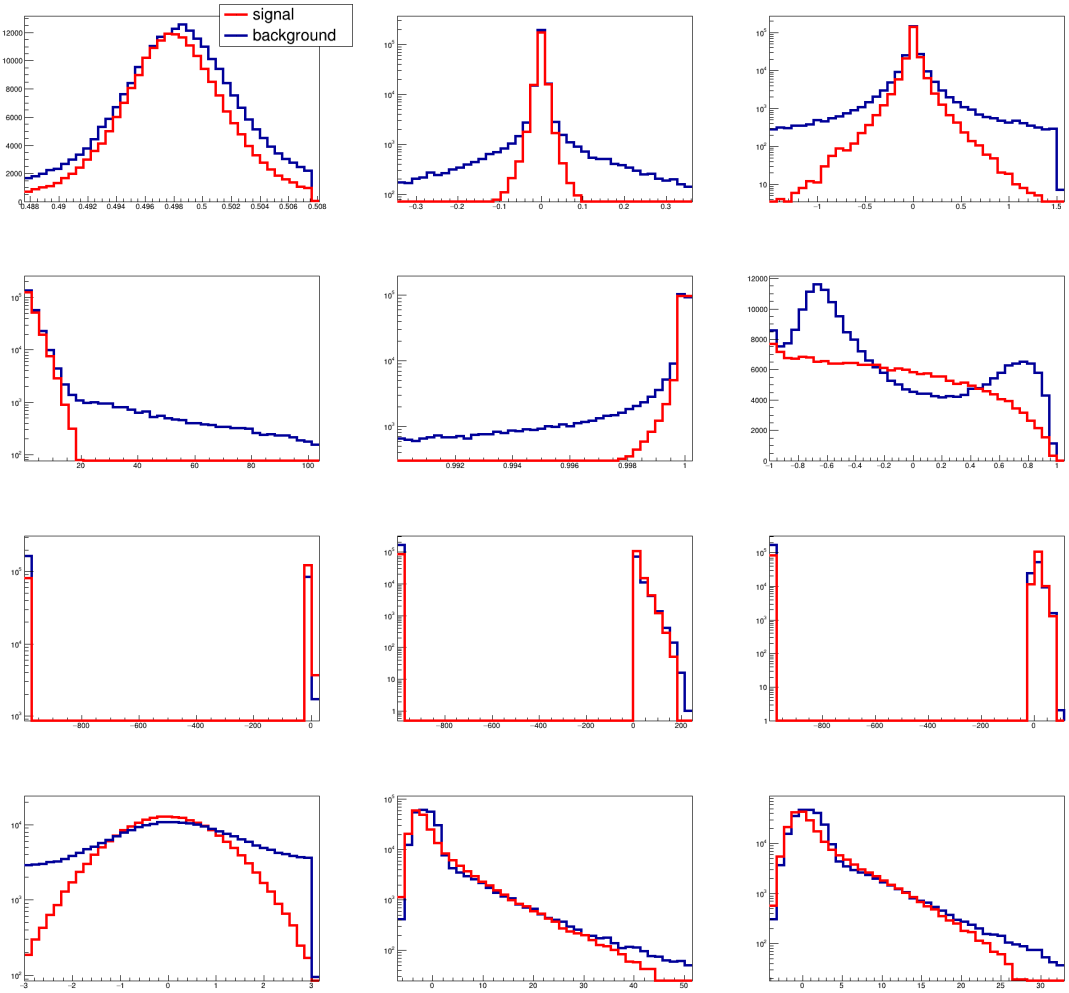

Figure 3.6: Input variables distribution in p_T range $[4, 6]$. From left to right, top to bottom: massK0S, tImpParBach, tImpParV0, CtK0s, cosPAK0S, CosThetaStar, nSigmaTOFpr, nSigmaTOFpi, nSigmaTOFka, nSigmaTPCpr, nSigmaTPCpi, nSigmaTPCka. Signal is red and background is blue.

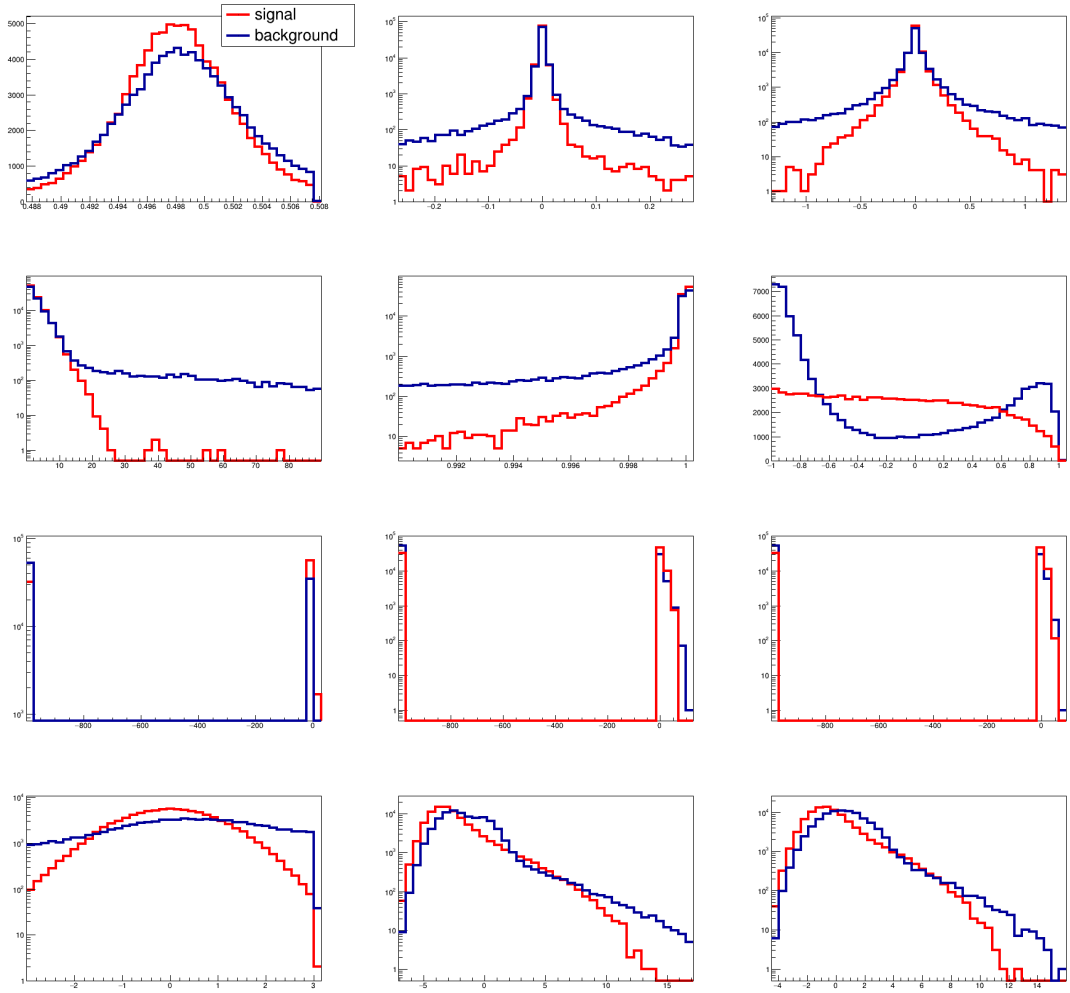

Figure 3.7: Input variables distribution in p_T range $[6, 8]$. From left to right, top to bottom: massK0S, tImpParBach, tImpParV0, CtK0s, cosPAK0S, CosThetaStar, nSigmaTOFpr, nSigmaTOFpi, nSigmaTOFka, nSigmaTPCpr, nSigmaTPCpi, nSigmaTPCka. Signal is red and background is blue.

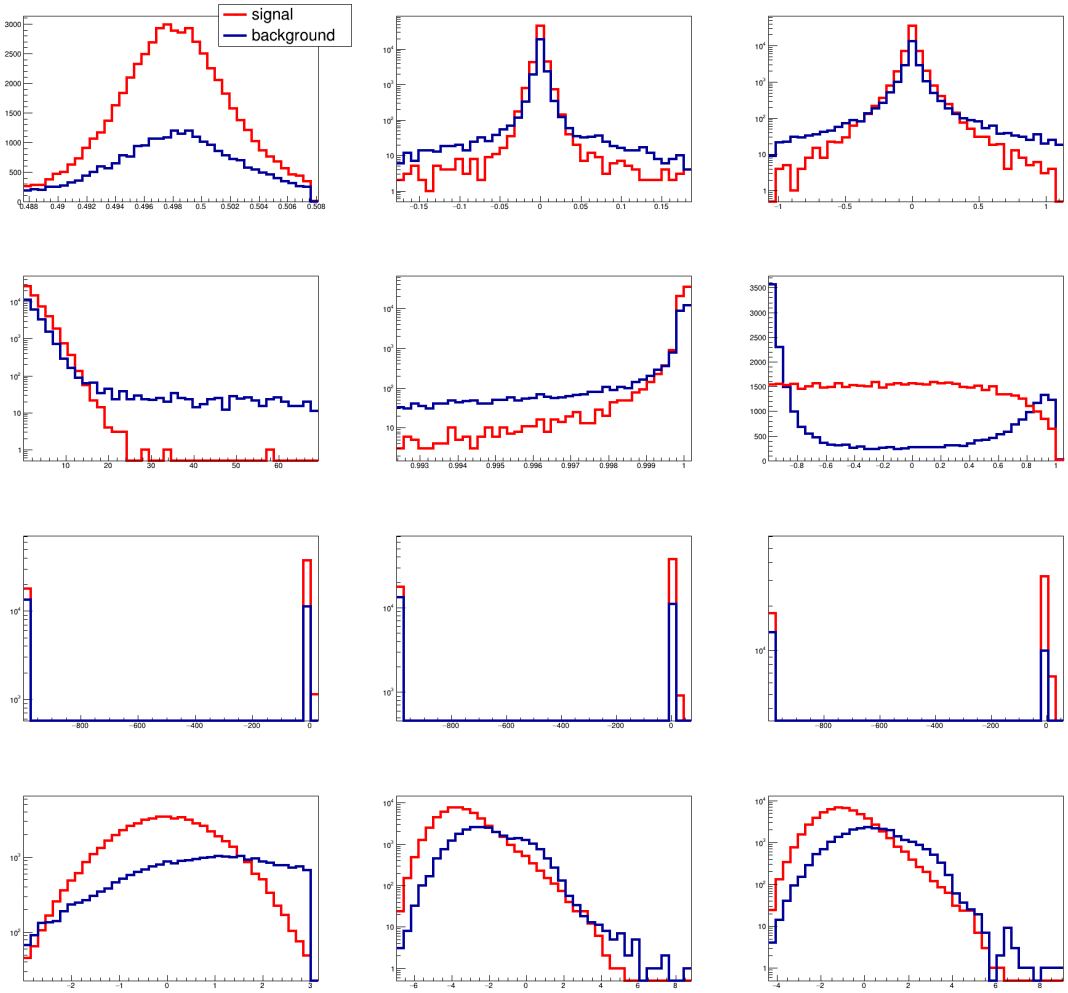

Figure 3.8: Input variables distribution in p_T range $[8, 12]$. From left to right, top to bottom: massK0S, tImpParBach, tImpParV0, CtK0s, cosPAK0S, CosThetaStar, nSigmaTOFpr, nSigmaTOFpi, nSigmaTOFka, nSigmaTPCpr, nSigmaTPCpi, nSigmaTPCka. Signal is red and background is blue.

3.5 Method description and settings

The ALICE TOF detector isn't always able to identify the proton. The particle is first identified by the TPC, but it may never be detected in the TOF because of several reasons, such as interactions with the TRD, algorithm inefficiencies, holes in the TOF acceptance due to hardware problems or MRPC inefficiency. All of these effects give us an overall matching efficiency of about 70%, which also has a p_T dependence, since low p_T particles detected by the TPC may be deflected so much they don't even reach the TOF detector.

When the data is missing, the variables are given the default value of -999. We thought this may negatively impact the performance of the BDT algorithm, since this value is the same for both signal and background. So we decided to use two more approaches, in addition to the simple BDT, that may account for this issue.

The first alternative approach is using the *Category* method available in the TMVA library. This method lets the user separate the training data into disjoint subpopulations with different properties. An independent training is performed in each of these regions using possibly different MVA methods. In our case we used two subpopulations, one where TOF data is available and one where it's missing, and trained both of them with the BDT method.

The second alternative approach is combining the TOF and TPC data in a single variable nSigmapr defined as:

$$nSigmapr = \begin{cases} \sqrt{nSigmaTPCpr^2 + nSigmaTOFpr^2} & nSigmaTOF \neq -999 \\ nSigmaTPCpr & nSigmaTOF = -999 \end{cases}$$

which is used instead of the nSigmaTPCpr and nSigmaTOFpr variables. This way we are able to give a well-defined value to every variable for each candidate.

So overall we trained and tested three different methods: the simple BDT algorithm applied with the input variables described in the previous section (we will refer to this as `BDT`); the BDT algorithm applied to the two different subpopulations, with those same variables (we will refer to this as `BdtCat`); the BDT algorithm applied to the new set of variables that uses nSigmapr (we will refer to this as `BdtSqrt`).

In all cases, the data was split and the BDTs were trained in 6 different intervals of p_T (GeV/c): $[0,1]$, $[1,2]$, $[2,4]$, $[4,6]$, $[6,8]$ and $[8,12]$. The following BDT settings are also the same for all cases:

Option	Value	Description
NTrees	850	Number of trees in the forest
MaxDepth	3	Max depth of the decision tree allowed
MinNodeSize	2.5 %	Minimum percentage of training events required in a leaf node
BoostType	AdaBoost	Boosting type for the trees in the forest
AdaBoostBeta	0.5	Learning rate for AdaBoost algorithm
BaggedSampleFraction	0.6	Relative size of the bagged event sample to original size of the data sample
SeparationType	GiniIndex	Separation criterion for node splitting
nCuts	20	number of points in variable range used to find optimal cut in splitting node

3.6 Linear correlation

Before training any method, the software produces linear correlation matrices for the variables in case of signal and background, shown in figs. 3.9 – 3.14. Even though it has been demonstrated that BDT performances are not affected by correlation between the input variables, it's good practice to try to avoid it since highly correlated variables might unnecessarily increase training processing time. The high correlation between the three TOF PID variables, shown in the figures, is expected and shouldn't be a problem.

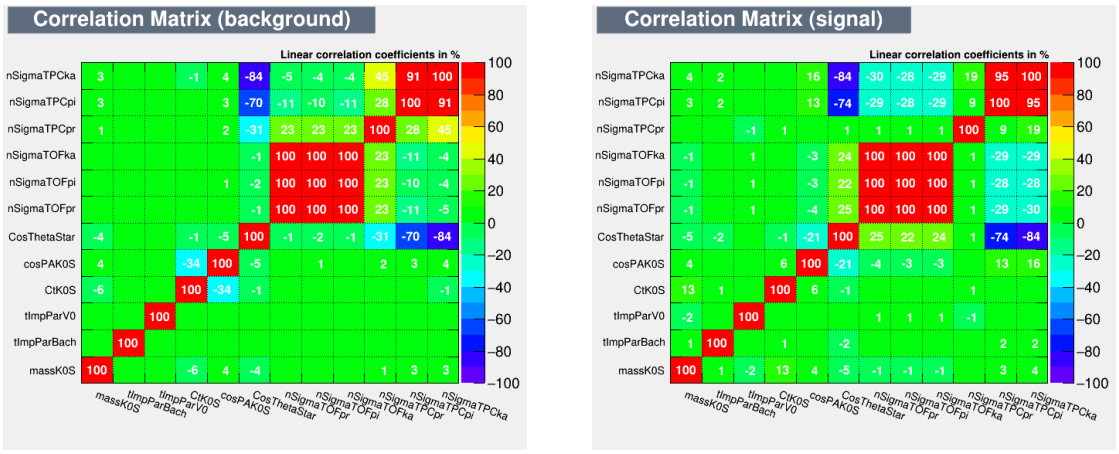

Figure 3.9: Input variable linear correlation coefficients for background (left) and signal (right) in the p_T range $[0, 1]$ GeV/c.

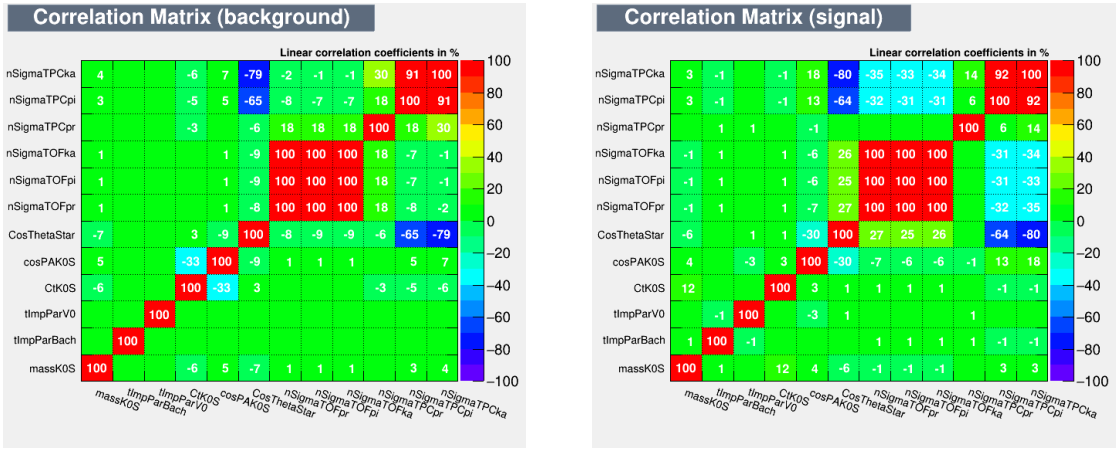

Figure 3.10: Input variable linear correlation coefficients for background (left) and signal (right) in the p_T range $[1, 2]$ GeV/c.

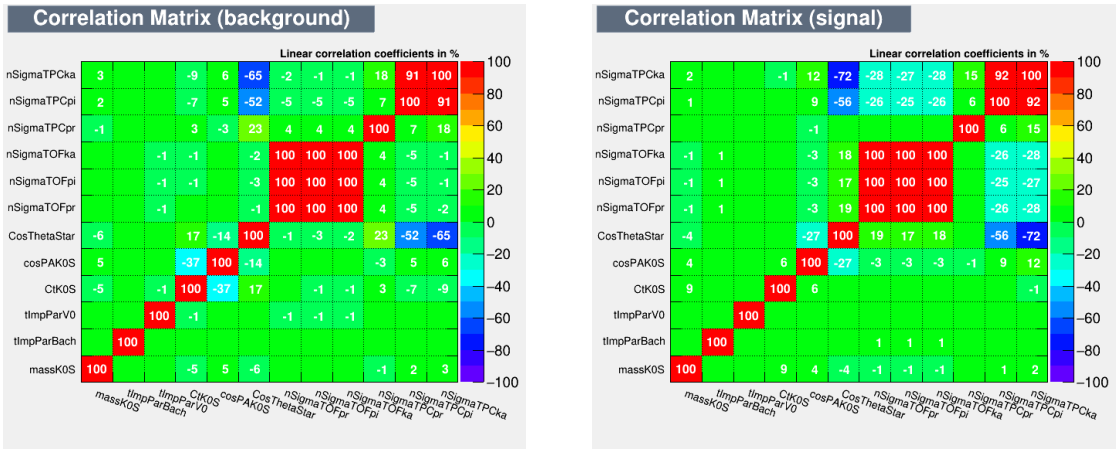

Figure 3.11: Input variable linear correlation coefficients for background (left) and signal (right) in the p_T range $[2, 4]$ GeV/c.

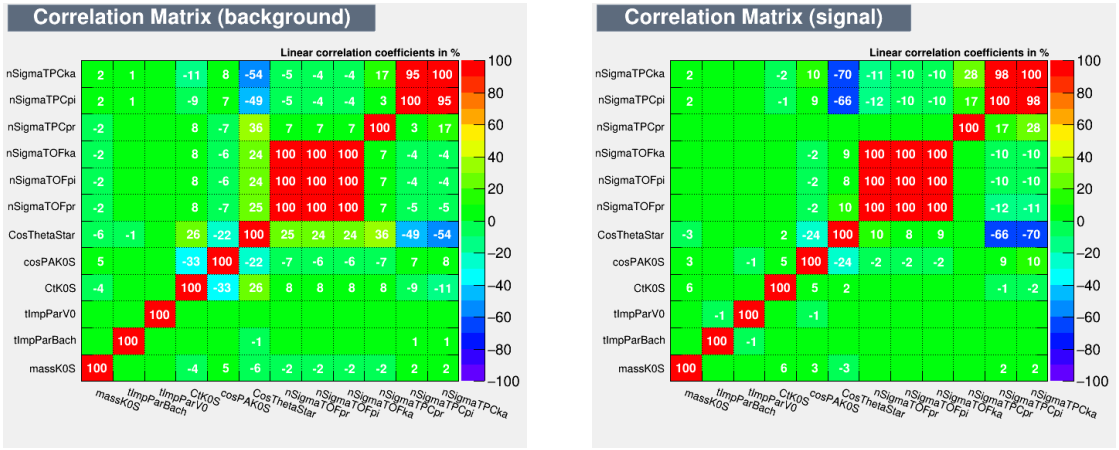

Figure 3.12: Input variable linear correlation coefficients for background (left) and signal (right) in the p_T range $[4, 6]$ GeV/c.

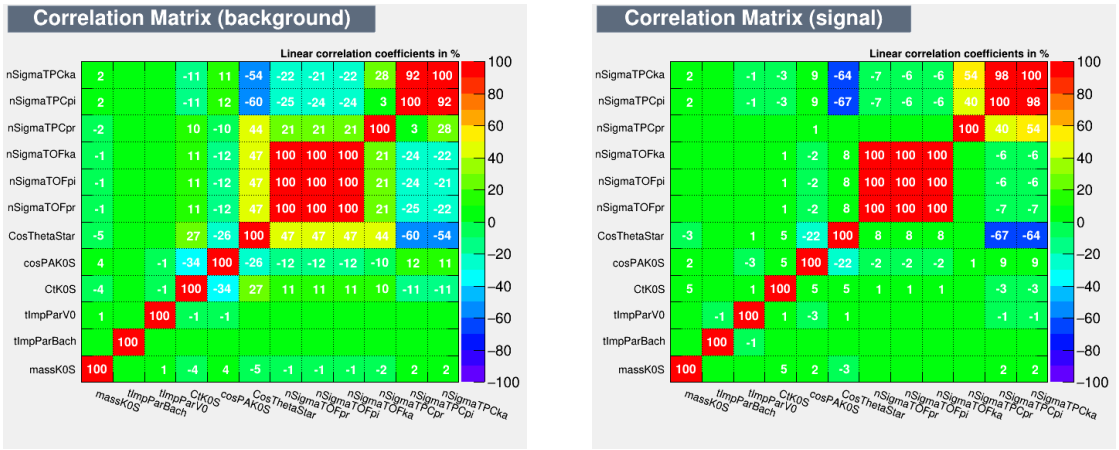

Figure 3.13: Input variable linear correlation coefficients for background (left) and signal (right) in the p_T range $[6, 8]$ GeV/c.

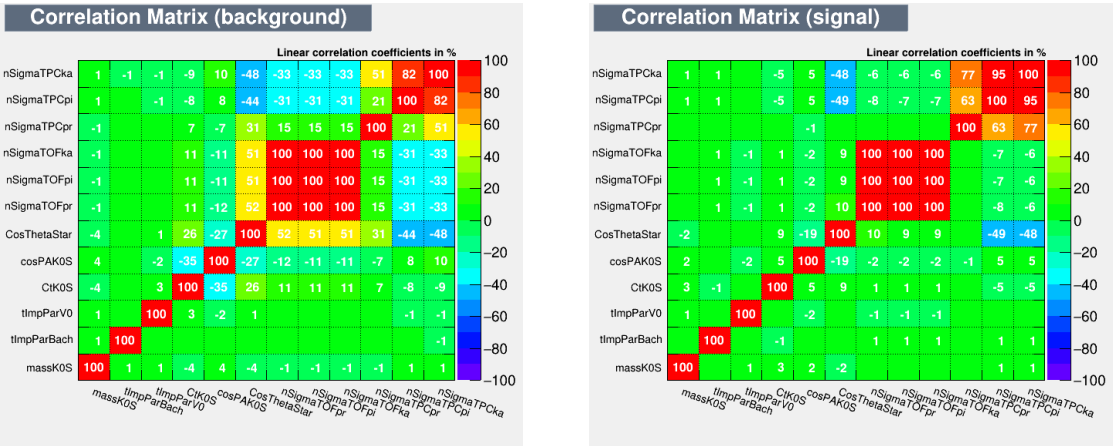

Figure 3.14: Input variable linear correlation coefficients for background (left) and signal (right) in the p_T range $[8, 12]$ GeV/c.

3.7 BDT response and testing for overtraining

Before running the classification algorithm, TMVA splits the input sample in two subsamples (training and testing samples) by randomly sampling them with an equal amount of candidates; only the first is used in the training phase, and the algorithm will apply the weights calculated during training to the second one. Figs. 3.15 − 3.20 show the response distribution for training and testing data. Comparing the distributions is a way to check that the algorithm was not *over-trained*. Overtraining happens when the trained classifier starts to pickup patterns in the training data which are not significant and just statistical fluctuations. This makes the classifier perform very well on the training data, but makes the general performance worse. In all histograms, the training and testing response are very close, so we can say that no overtraining has occurred.

These histograms also let us make early considerations about which of the three methods is best. We want to separate signal and background data by applying a 'cut' on the x-axis. A good cut will remove most of the background data, but also keep a good portion of the signal.

In some p_T ranges such as $[0, 1]$ GeV/c and $[1, 2]$ GeV/c it's easy to understand that `BdtCat` performs worse than the other two, since it's hard to make a clear cut that separates signal and background. `BDT` and `BdtSqrt` are hard to compare visually, and seem to have a similar performance.

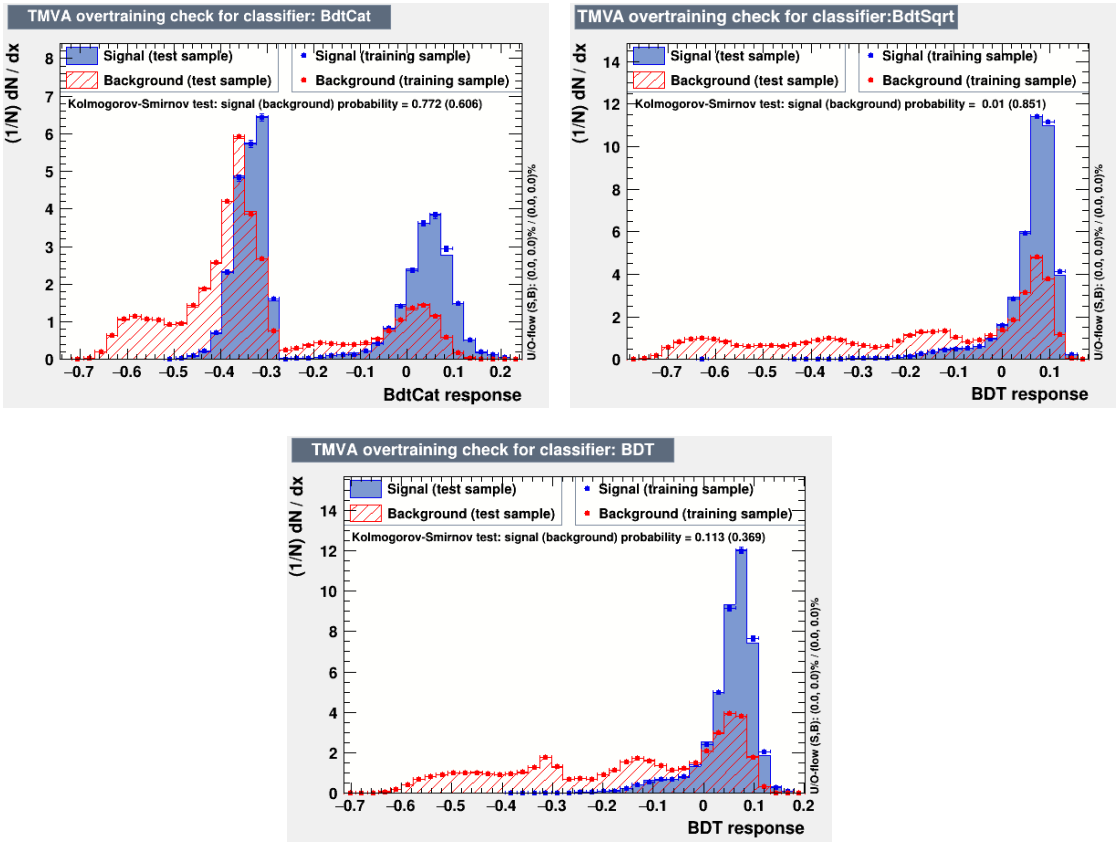

Figure 3.15: BDT response for test and training data in the p_T range $[0, 1]$ GeV/c for `BdtCat` (top left), `BdtSqrt` (top right) and `BDT` (bottom).

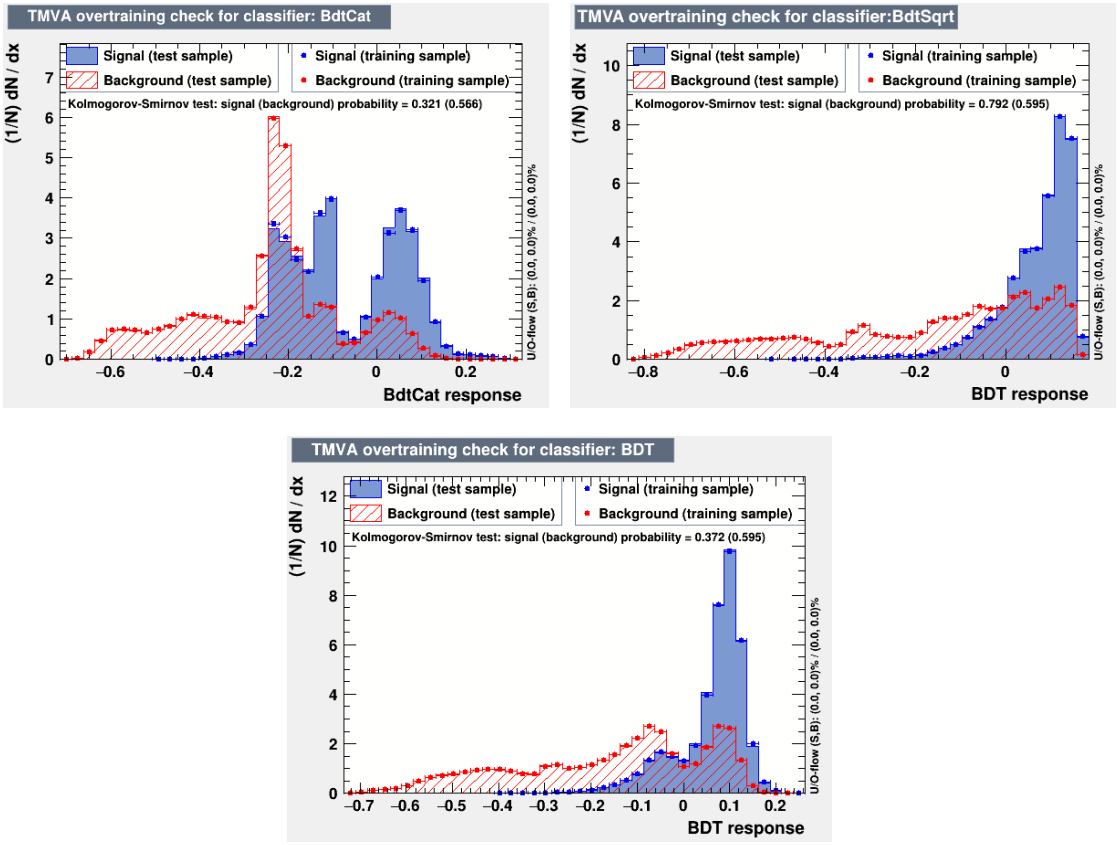

Figure 3.16: BDT response for test and training data in the p_T range $[1, 2]$ GeV/c for `BdtCat` (top left), `BdtSqrt` (top right) and `BDT` (bottom).

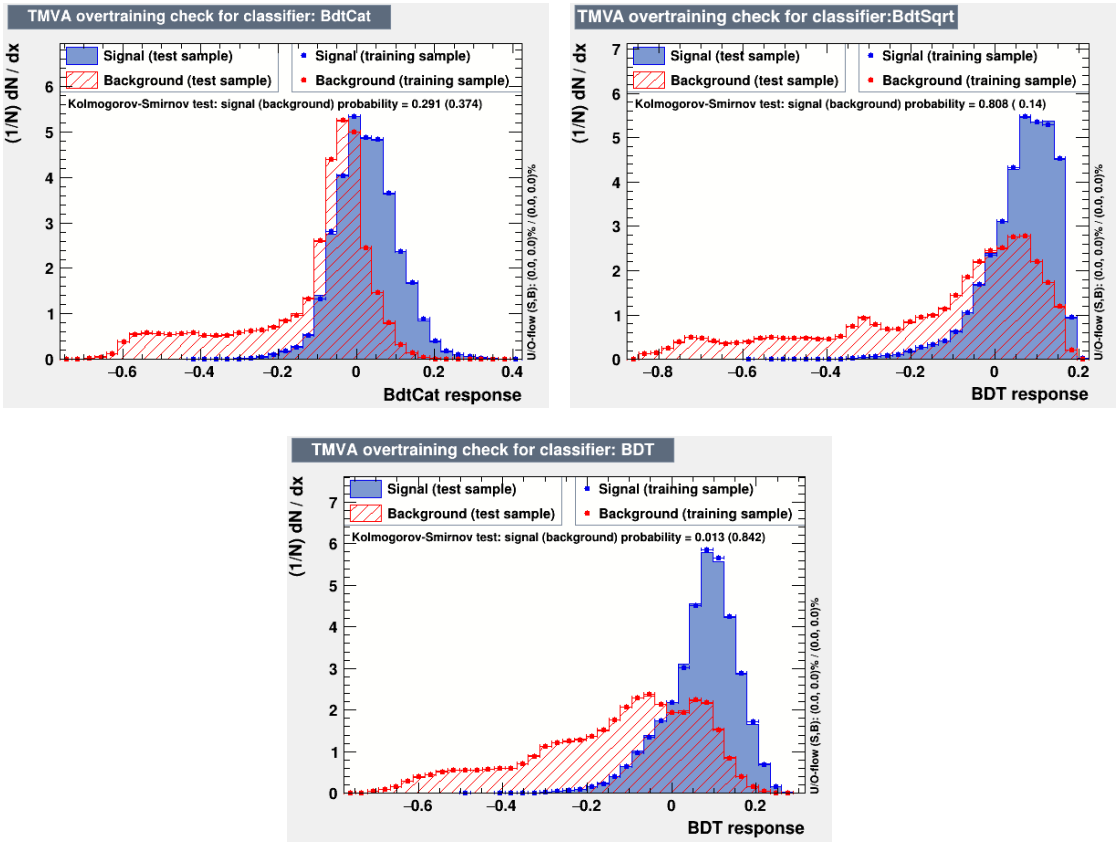

Figure 3.17: BDT response for test and training data in the p_T range $[2, 4]$ GeV/c for `BdtCat` (top left), `BdtSqrt` (top right) and `BDT` (bottom).

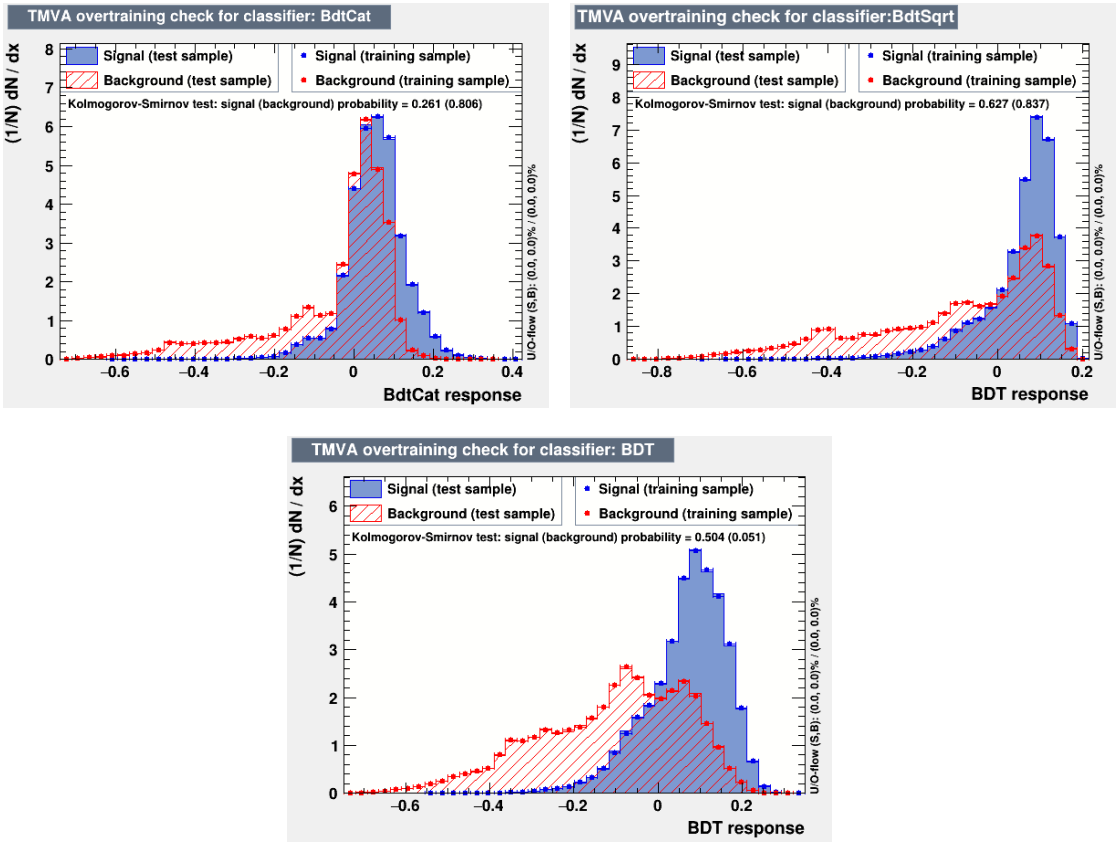

Figure 3.18: BDT response for test and training data in the p_T range $[4, 6]$ GeV/c for `BdtCat` (top left), `BdtSqrt` (top right) and `BDT` (bottom).

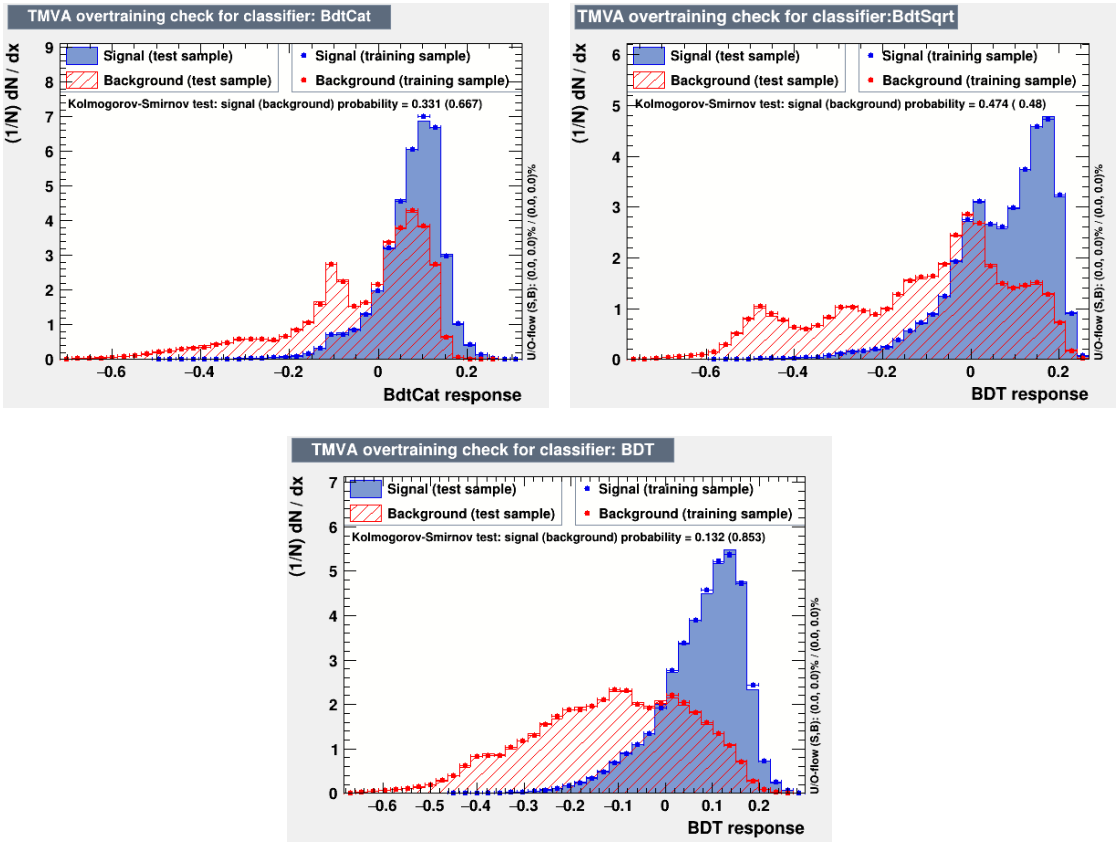

Figure 3.19: BDT response for test and training data in the p_T range $[6, 8]$ GeV/c for `BdtCat` (top left), `BdtSqrt` (top right) and `BDT` (bottom).

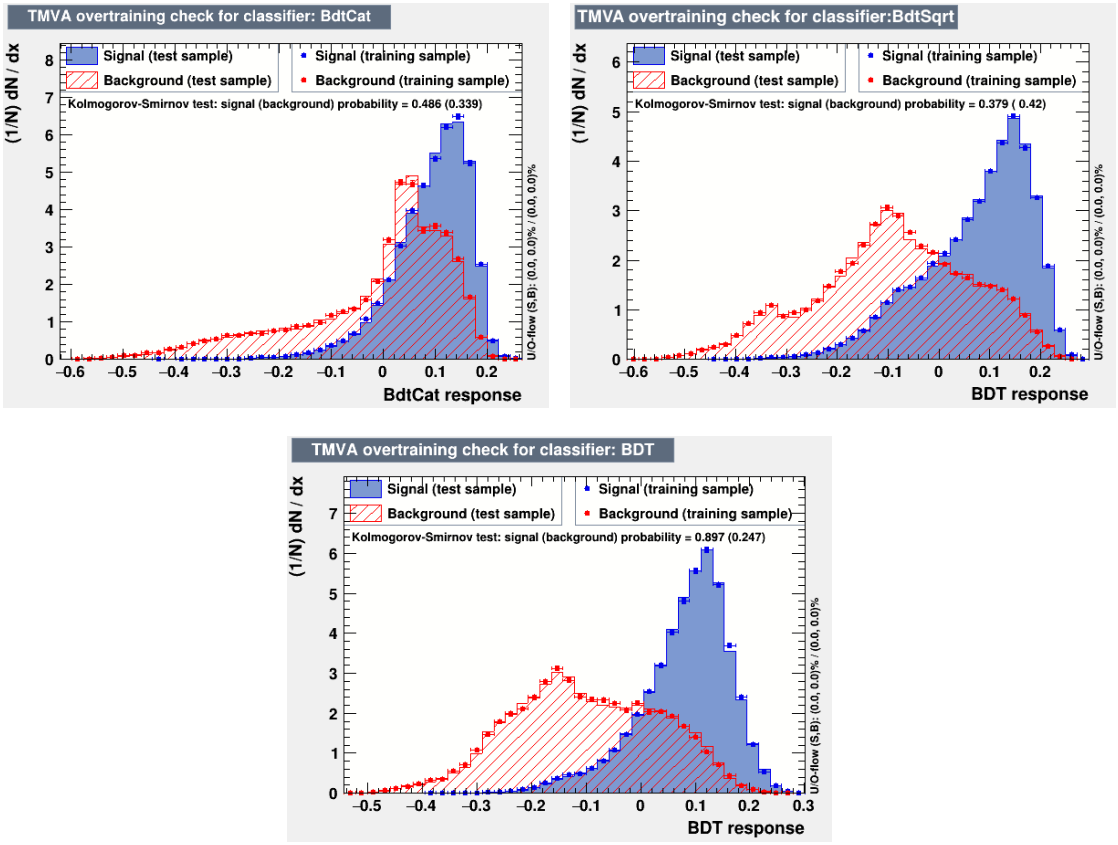

Figure 3.20: BDT response for test and training data in the p_T range $[8, 12]$ GeV/c for `BdtCat` (top left), `BdtSqrt` (top right) and `BDT` (bottom).

3.8 ROC curves

A *receiver operating characteristic* curve, or ROC curve, plots background rejection against signal efficiency. The ROC curve for an ideal classifier would be rectangular, with a background rejection of 100% and signal efficiency of 100% and an integral of 1. The closer a method is to this ideal curve, the better it is at discriminating signal and background. Table 3.1 shows the values of ROC integrals and figs. $3.21 - 3.26$ compare the ROC curves of the three methods in every p_T range.

ROC Integral	p_T range $[0,1]$	$[1,2]$	$[2,4]$	$[4,6]$	$[6,8]$	$[8,12]$
`BDT`	0.831	0.841	0.841	0.816	0.845	0.868
`BdtSqrt`	0.804	0.819	0.783	0.743	0.781	0.808
`BdtCat`	0.744	0.802	0.795	0.714	0.746	0.737

Table 3.1: ROC integrals for every method in all p_T ranges

Like we expected by looking at the BDT output distributions, `BdtCat` has the worst performance. We expected that using the category method would improve the performance, however our results suggest that this scenario may not be appropriate for the use of this technique. A possible explanation is that there are no real physical differences between the two categories, as we would have, for example, by looking at two angular or rapidity regions of the experiment where the types of detectors and the resolutions are different. Here with the category we are training separately the two methods with fewer candidates with respect to the other two approaches, and this could somehow explain the worsening in the performance. `BDT` turned out to be the best in spite of the issue with the TOF variables. `BdtSqrt` performance is comparable to `BDT` but slightly worse. This indicates that the usage of the BDT out-of-the-box is able to deal reasonably well with missing data values.

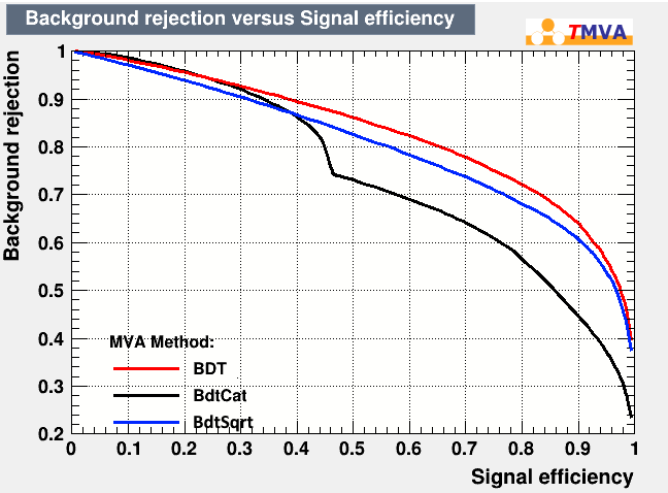

Figure 3.21: ROC curves for the p_T range $[0, 1]$ GeV/c.

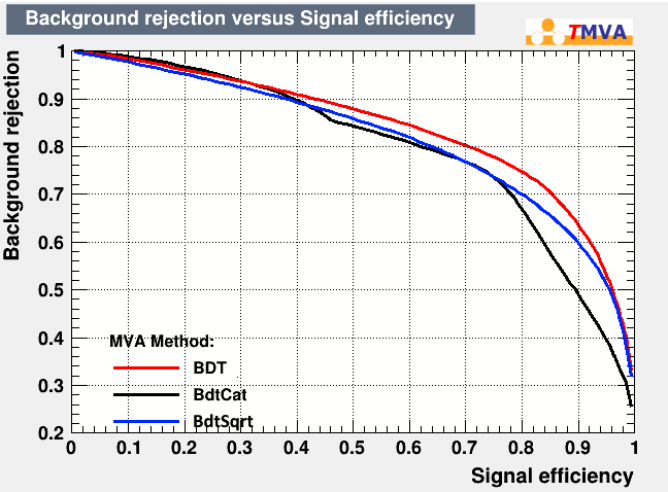

Figure 3.22: ROC curves for the p_T range $[1, 2]$ GeV/c.

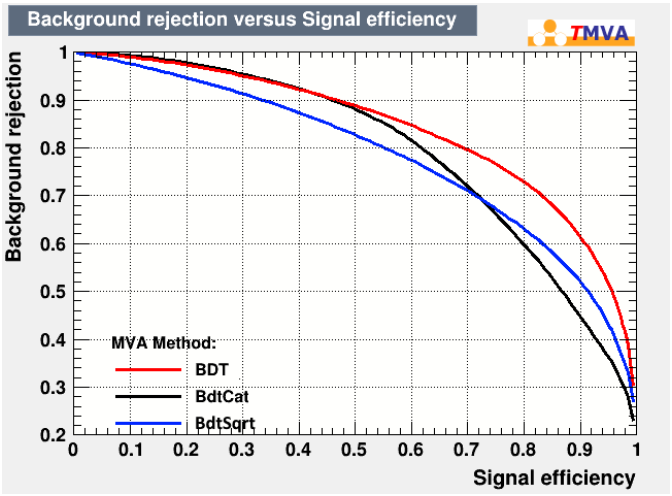

Figure 3.23: ROC curves for the p_T range $[2, 4]$ GeV/c.

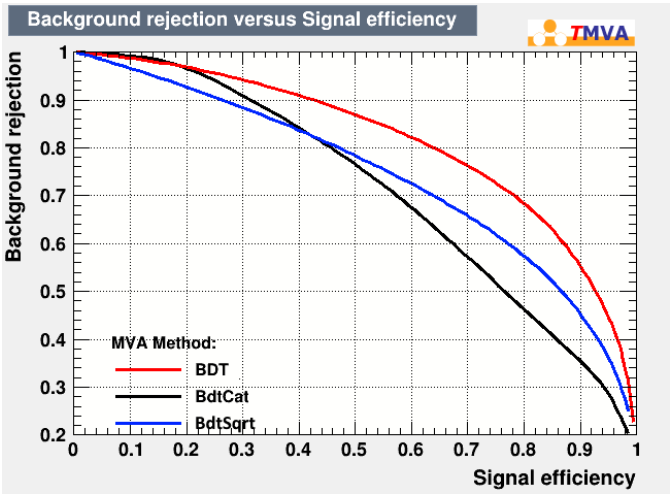

Figure 3.24: ROC curves for the p_T range $[4, 6]$ GeV/c.

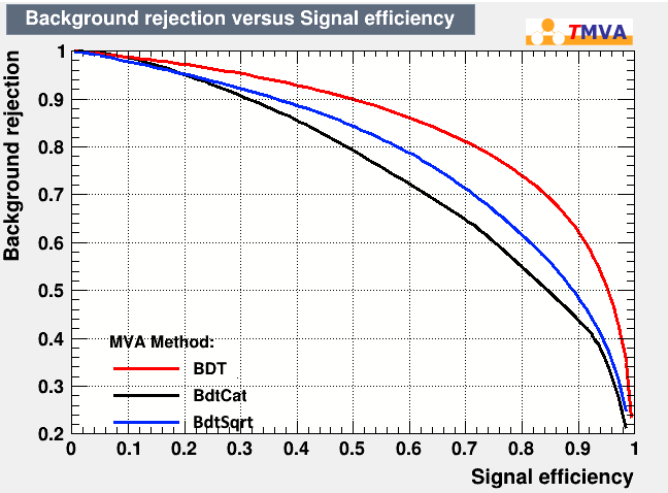

Figure 3.25: ROC curves for the p_T range $[6, 8]$ GeV/c.

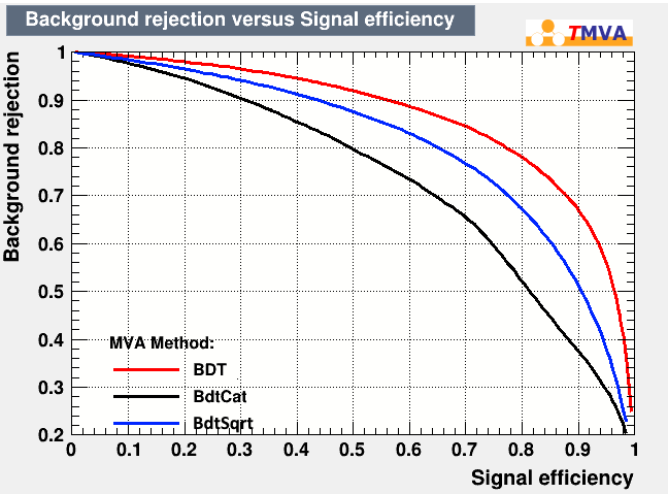

Figure 3.26: ROC curves for the p_T range $[8, 12]$ GeV/c.

3.9 BDT variable ranking

TMVA provides a ranking of the input variables, based on how important they were in determining the splitting of the nodes. The following tables show variable ranking for `BDT` method in all p_T ranges.

rank	variable	importance
1	CtK0S	2.031e-01
2	nSigmaTPCka	1.239e-01
3	CosThetaStar	1.217e-01
4	nSigmaTPCpr	1.148e-01
5	cosPAK0S	1.089e-01
6	nSigmaTPCpi	1.002e-01
7	tImpParBach	8.215e-02
8	massK0S	8.168e-02
9	tImpParV0	6.353e-02

rank	variable	importance
1	CtK0S	2.004e-01
2	nSigmaTPCpr	1.289e-01
3	cosPAK0S	1.181e-01
4	nSigmaTPCka	1.159e-01
5	CosThetaStar	1.140e-01
6	nSigmaTPCpi	8.778e-02
7	massK0S	8.676e-02
8	tImpParV0	7.630e-02
9	tImpParBach	7.178e-02

Table 3.2: BDT variable ranking in p_T range $[0, 1]$ (left) and $[1, 2]$ (right).

rank	variable	importance
1	CtK0S	1.991e-01
2	cosPAK0S	1.365e-01
3	nSigmaTPCka	1.191e-01
4	nSigmaTPCpr	1.151e-01
5	CosThetaStar	1.103e-01
6	nSigmaTPCpi	1.023e-01
7	tImpParBach	8.267e-02
8	massK0S	7.563e-02
9	tImpParV0	5.918e-02

rank	variable	importance
1	cosPAK0S	1.530e-01
2	nSigmaTPCpr	1.421e-01
3	CtK0S	1.289e-01
4	nSigmaTPCpi	1.259e-01
5	nSigmaTPCka	1.217e-01
6	CosThetaStar	1.164e-01
7	tImpParBach	9.532e-02
8	massK0S	6.324e-02
9	tImpParV0	5.354e-02

Table 3.3: BDT variable ranking in p_T range $[2, 4]$ (left) and $[4, 6]$ (right).

rank	variable	importance
1	nSigmaTPCpr	1.836e-01
2	nSigmaTPCpi	1.524e-01
3	cosPAK0S	1.389e-01
4	CosThetaStar	1.382e-01
5	massK0S	9.497e-02
6	nSigmaTPCka	8.994e-02
7	CtK0S	7.845e-02
8	tImpParV0	6.865e-02
9	tImpParBach	5.496e-02

rank	variable	importance
1	nSigmaTPCpi	1.612e-01
2	nSigmaTPCpr	1.595e-01
3	CosThetaStar	1.486e-01
4	massK0S	1.206e-01
5	nSigmaTPCka	1.151e-01
6	cosPAK0S	1.002e-01
7	tImpParV0	6.983e-02
8	CtK0S	6.443e-02
9	tImpParBach	6.057e-02

Table 3.4: BDT variable ranking in p_T range $[6, 8]$ (left) and $[8, 12]$ (right).

3.10 BDT Application

Once trained, the algorithm can be applied to a sample of data where the identity of the candidates is unknown; here we want to verify that our approach based on a multivariate analysis can be effectively used in the reconstruction of the Λ_c^+ baryon in real data with a good statistical significance. Based on previous considerations, we select the method we referred to as `BDT`, since it's the one showing the best overall performances, and we apply it to the data collected by the ALICE experiment in pp collisions at $\sqrt{s} = 13\ TeV$ during the LHC Run2 data taking (2016-2017-2018). To reconstruct the number of particles, we must get rid of a significant part of the background, by applying a cut and only keeping events with a BDT response above the cut. We used the cuts suggested by TMVA during the training phase that maximize the significance, which are the following:

	p_T range					
	$[0, 1]$	$[1, 2]$	$[2, 4]$	$[4, 6]$	$[6, 8]$	$[8, 12]$
applied cut	-0.08	-0.07	-0.06	-0.08	-0.06	-0.03

Now we consider the invariant mass histograms we get after applying these cuts, and we fit this data with a gaussian around the expected value of m_{Λ_c} and a second degree polynomial (except for the $[0, 1]$ range where we used a third degree polynomial) to model the background. Figs. $3.27 - 3.32$ show the histograms and the fit results. Table 3.5 shows the calculated signal and significance for every p_T range. Signal and background were calculated from the integral of the gaussian and background fit functions in a range of 3σ around the mean value. Significance was calculated as $S/\sqrt{S+B}$.

So we have shown that this approach looks feasible, and it does indeed allow us to calculate the number of Λ_c^+ particles produced in the decay channel we

p_T range	Signal(3σ)	significance(3σ)
$[0, 1]$	4748 ± 1193	3.9 ± 1.0
$[1, 2]$	13517 ± 1509	8.6 ± 1.0
$[2, 4]$	18591 ± 1264	13.8 ± 0.9
$[4, 6]$	7413 ± 523	13.1 ± 0.9
$[6, 8]$	2034 ± 183	11.0 ± 1.0
$[8, 12]$	770 ± 91	7.9 ± 0.9

Table 3.5: Signal and significance for every p_T range in an interval of 3σ around the mean value.

considered, with good statistical significance. Taking into account the algorithm's efficiency and any other selection efficiency for cuts that have been applied to the data before the BDT analysis, we can then extract the Λ_c^+ corrected yield and eventually the p_T-differential production cross-section.

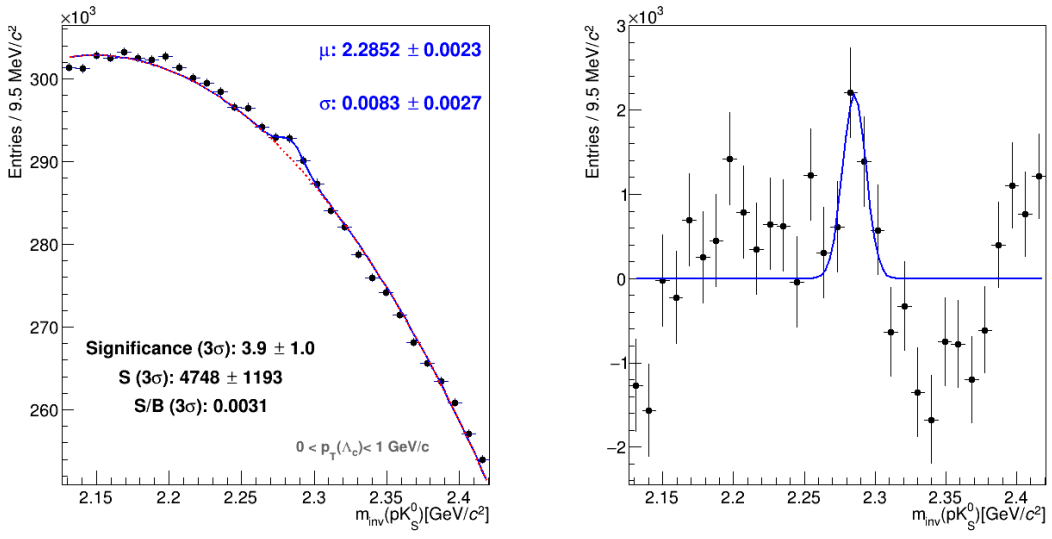

Figure 3.27: Invariant mass histogram for the $[0, 1]$ p_T range, fitted with gaussian and a third degree polynomial background. Background was subtracted in the right histogram.

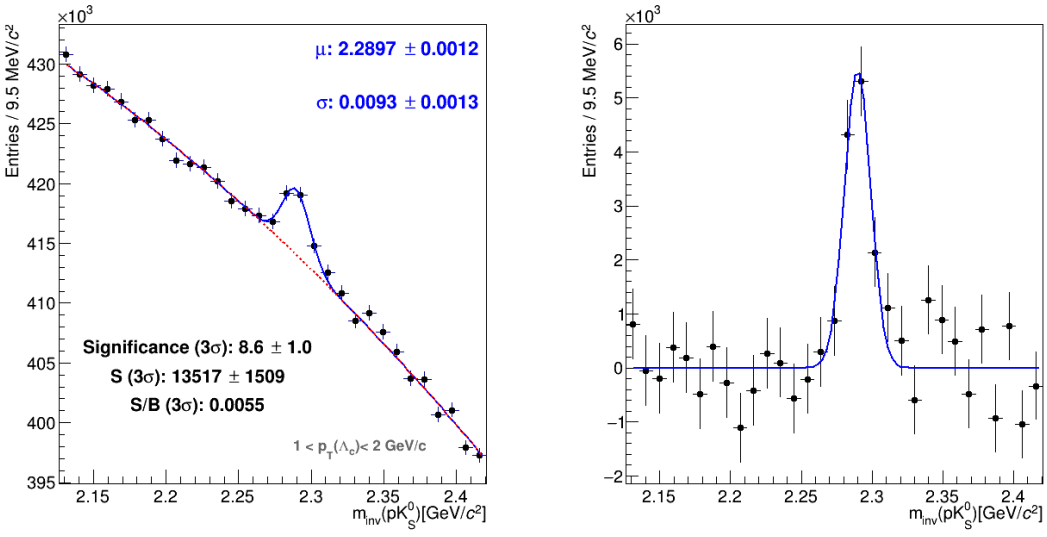

Figure 3.28: Invariant mass histogram for the $[1, 2]$ p_T range, fitted with gaussian and a second degree polynomial background. Background was subtracted in the right histogram.

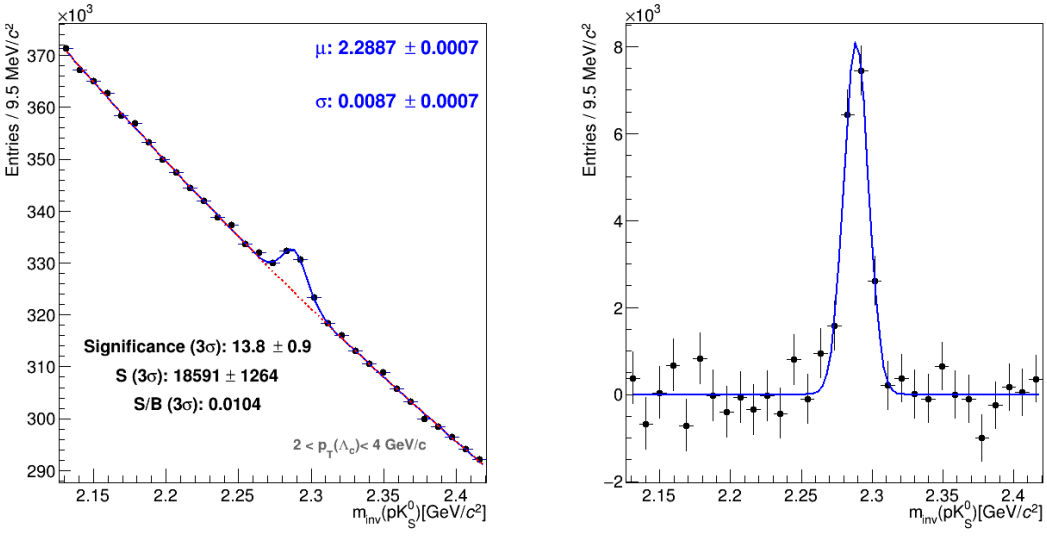

Figure 3.29: Invariant mass histogram for the $[2, 4]$ p_T range, fitted with gaussian and a second degree polynomial background. Background was subtracted in the right histogram.

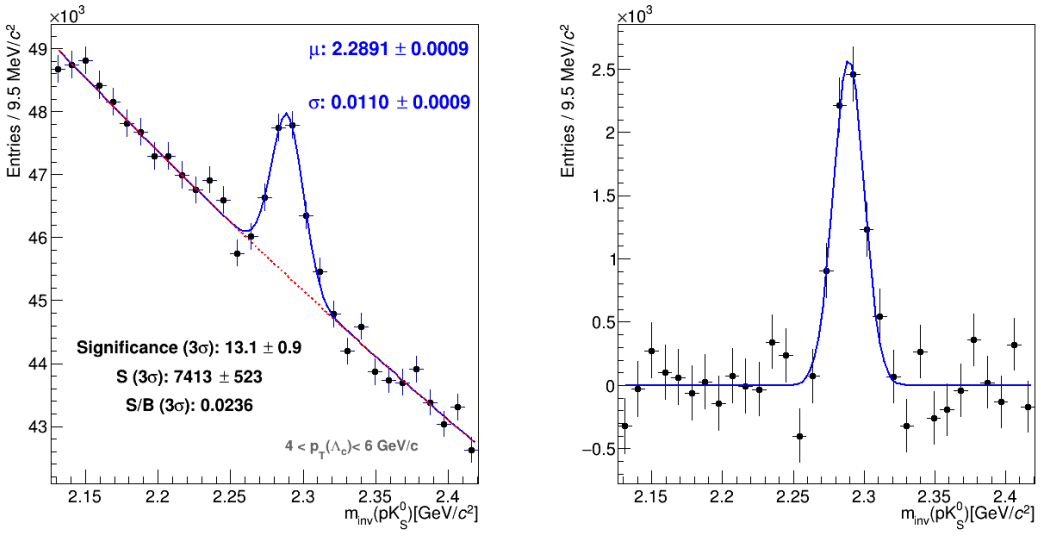

Figure 3.30: Invariant mass histogram for the $[4, 6]$ p_T range, fitted with gaussian and a second degree polynomial background. Background was subtracted in the right histogram.

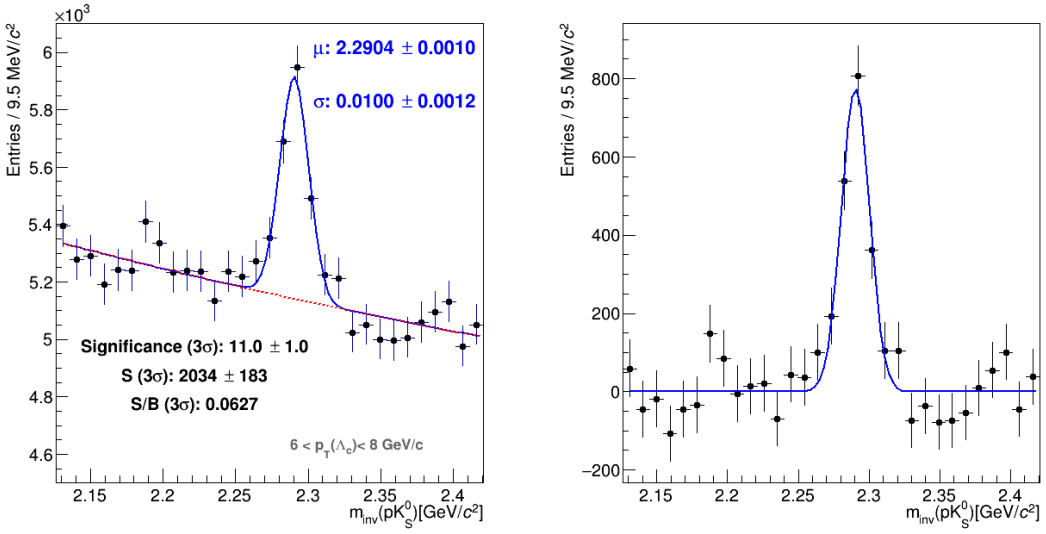

Figure 3.31: Invariant mass histogram for the $[6, 8]$ p_T range, fitted with gaussian and a second degree polynomial background. Background was subtracted in the right histogram.

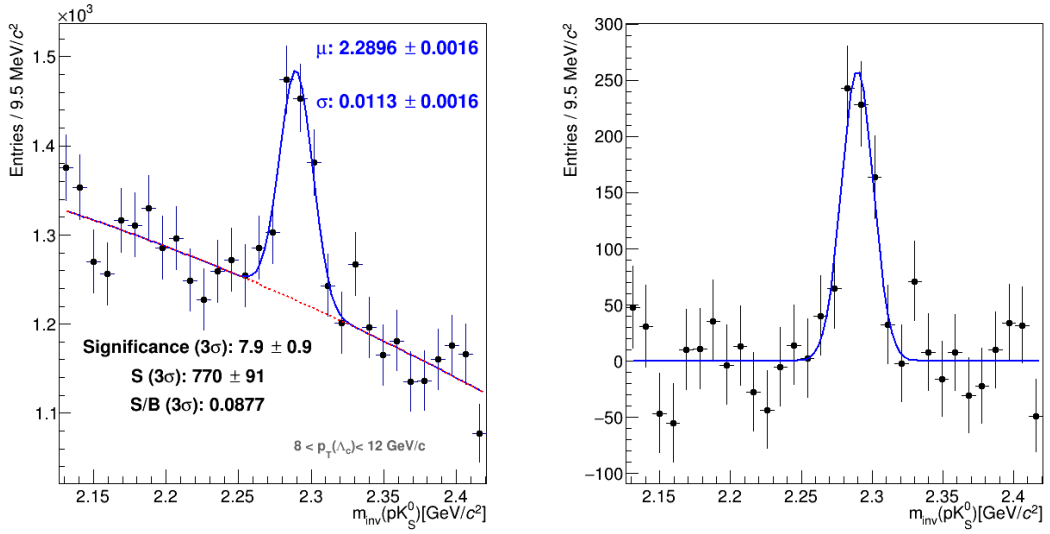

Figure 3.32: Invariant mass histogram for the $[8, 12]$ p_T range, fitted with gaussian and a second degree polynomial background. Background was subtracted in the right histogram.

Conclusions

QCD is a key part of the Standard Model, and understanding QGP and its properties is one of QCD's main aims. Studying QCD matter at extreme energy densities and temperatures is the purpose of the ALICE detector at LHC. Recent studies analyzing pp and p-Pb collisions at ALICE have shown that the conclusions made about heavy-flavour hadron production in e^+e^- collisions are not universal. New models have tried to explain the measured enhancement in the Λ_c^+/D^0 ratio by hypothesizing different mechanisms. The increase in baryon production has been explained by considering the onset of additional hadronization mechanisms at the LHC energies and with multi-parton interactions, by taking into account an increased set of excited and still unobserved baryon states beyond those listed by the Particle Data Group or assuming the creation of a small-size QGP also in small colliding systems. More than one theory is able to reproduce the measured values, and the experimental uncertainties don't allow to draw firm conclusions about the models; new measurements with higher precision are then mandatory.

In order to measure the Λ_c^+/D^0 we have to estimate the number of Λ_c^+ produced in the collision; this measurement however is challenging especially due to the short Λ_c^+ lifetime. The best approach is to make use of multivariate analysis algorithms, exploiting the most of the available information through machine learning techniques, and to train them to recognize patterns and learn to distinguish between signal and background events. In this work, we used the TMVA library and tested three different approaches, all based on the Boosted Decision Trees (BDT) method. The simple out-of-the-box BDT turned out to be the best choice. The use of this categorization method allowed us to remove a significant part of the background from the data and to be able to reconstruct the number of Λ_c^+ baryons produced in our data sample in the decay channel $\Lambda_c^+ \to pK_S^0$.

References

[1] W. N. Cottingham and D. A. Greenwood. *An introduction to the Standard Model of Particle Physics*. Cambridge University Press, 2007.

[2] Sylvie Braibant, Giorgio Giacomelli and Maurizio Spurio. *Particles and Fundamental Interactions: An Introduction to Particle Physics*. Springer, 2012.

[3] Cheuk-Yin Wong. *Introduction to High-Energy Heavy-Ion Collisions*. World Scientific, 1994.

[4] Sourav Sarkar, Helmut Satz and Bikash Sinha (Eds.) *The Physics of the Quark-Gluon Plasma: Introductory Lectures*. Springer, 2010.

[5] John C. Collins, Davison E. Soper and George Sterman. *Factorization of Hard Processes in QCD*. 2004. arXiv: 0409313 [hep-ph].

[6] Ramona Vogt. *Ultrarelativistic Heavy-Ion collisions*. Elsevier Science, 2007.

[7] ALICE Collaboration. Λ_c^+ *production in pp and in p-Pb collisions at* $\sqrt{s_{\mathrm{NN}}} = 5.02$ *TeV*. 2021. arXiv: 2011.06079 [nucl-ex].

[8] Jesper R. Christiansen and Peter Z. Skands. "String formation beyond leading colour". In: *Journal of High Energy Physics* 2015.8 (Aug. 2015). ISSN: 1029-8479. DOI: 10.1007/jhep08(2015)003. URL: http://dx.doi.org/10.1007/JHEP08(2015)003.

[9] ALICE Collaboration. *Measurement of prompt* D^0, Λ_c^+, *and* $\Sigma_c^{0,++}$*(2455) production in pp collisions at* $\sqrt{s} = 13$ *TeV*. 2021. arXiv: 2106.08278 [hep-ex].

[10] CERN Collaboration. *The Large Hadron Collider*. URL: https://home.cern/science/accelerators/large-hadron-collider.

[11] CERN Collaboration. *Accelerators*. URL: https://home.cern/science/accelerators.

[12] Lyn Evans and Lucie Linssen. "The Super-LHC is on the starting blocks". In: *CERN Courier* (July 8, 2008). URL: https://cerncourier.com/a/the-super-lhc-is-on-the-starting-blocks/.

[13] CERN Collaboration. *Pulling together: Superconducting electromagnets*. URL: https://home.cern/science/engineering/pulling-together-superconducting-electromagnets.

[14] CERN Collaboration. URL: https://home.cern/resources/image/physics/infographics-gallery.

[15] The ALICE Collaboration et al. "The ALICE experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08002–S08002. DOI: 10.1088/1748-0221/3/08/s08002. URL: https://doi.org/10.1088/1748-0221/3/08/s08002.

[16] C. Fabjan and J. Schukraft. *The story of ALICE: Building the dedicated heavy ion detector at LHC*. 2011. arXiv: 1101.1257 [physics.ins-det].

[17] CERN Collaboration. *Alice TPC*. URL: https://alice-collaboration.web.cern.ch/menu_proj_items/tpc.

[18] P.A. Zyla et al. (Particle Data Group), Prog. Theor. Exp. Phys. 2020, 083C01 (2020).

[19] A. Hoecker et al. *TMVA - Toolkit for Multivariate Data Analysis*. 2007. arXiv: 0703039.

[20] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[21] Torbjörn Sjöstrand et al. "An introduction to PYTHIA 8.2". In: *Computer Physics Communications* 191 (June 2015). ISSN: 0010-4655. DOI: 10.1016/j.cpc.2015.01.024. URL: http://dx.doi.org/10.1016/j.cpc.2015.01.024.

[22] René Brun et al. *GEANT: Detector Description and Simulation Tool; Oct 1994*. CERN Program Library. Long Writeup W5013. Geneva: CERN, 1993. DOI: 10.17181/CERN.MUHF.DMJ1. URL: http://cds.cern.ch/record/1082634.