Universiteit Utrecht

Faculty of Science

Reconstruction of D^+ mesons in proton-proton collisions at $\sqrt{s} = 5.02$ TeV using boosted decision tree techniques with ALICE at the LHC

Bachelor Thesis

C.H.B. Koenders

Physics & Astronomy

Supervisors:

Dr. André Mischke (†) Project Supervisor
Institute for Subatomic Physics
Dr. Alessandro Grelli Project Supervisor
Institute for Subatomic Physics

Luuk Vermunt Msc Daily Supervisor
Institute for Subatomic Physics

January 2019

Abstract

The use of machine-learning techniques in particle physics is relatively new and unexplored when compared to more conventional methods that use rectangular cuts. This thesis has its focus on the reconstruction of $\mathrm{D}^{\pm}$ mesons in transverse momentum range $2 \leq p_{\mathrm{T}} \leq 24$ GeV/c, that are produced in proton-proton collisions at centre-of-mass energy $\sqrt{s} = 5.02$ TeV with ALICE. The goal of this thesis is to try to reduce the current statistical uncertainties on the extraction of the $\mathrm{D}^{\pm}$ raw yield by leveraging boosted decision trees for signal selection. This is especially important because, besides being an excellent test for (perturbative) quantum chromodynamics, this type of measurement constitutes the baseline for the corresponding measurement in heavy-ion collisions. Currently the uncertainties on the baseline measurement are a limiting factor when comparing to models, to shed light on the in-medium energy loss of charm quarks in the quark-gluon plasma. A precise estimation of the baseline helps with extracting key plasma properties and therefore it contributes to paint a picture of the early Universe, only a few microseconds after the Big Bang, since it is believed that at that point all the matter in the Universe was in the quark-gluon plasma phase. The results presented in this thesis show a promising increase in significances, which directly correlates to an improved statistical uncertainty, compared to those obtained from the topological selection cuts used for the standard ALICE analysis, along with some interesting deviations in efficiency-corrected yields. The stability of the results obtained from the use of boosted decision trees is shown to be high, especially at mid-to-high p_{T}.

Contents

1 Introduction

For centuries scientists have tried to uncover and understand the intrinsic nature of the world and Universe we live in. We have now arrived at a time where we can start to unveil the fundamental properties of matter itself, including how it behaves at extremely high energy, pressure and temperature. This could teach us about the structure of the early Universe, the origins of matter and the fundamental forces of nature.

Analysing collision data recorded by A Large Ion Collider Experiment (ALICE) [1] at the Large Hadron Collider (LHC) [2] at CERN in Geneva gives us a way to probe this behaviour of matter at high energy, pressure and temperature. A new state of matter, the quark-gluon plasma, is expected to be created, and current leading theories can be subjected to empirical testing.

This thesis will analyse proton-proton collision data with the help from Boosted Decision Trees (BDT's) [3] for signal selection, and compare it to the standard rectangular selection method that is used by the ALICE collaboration. The focus of this analysis will be on the production of D^+ mesons, which is of fundamental importance to the verification of (perturbative) quantum chromodynamics. In addition, proton-proton collisions are needed as a baseline measurement for probing the quark-gluon plasma which is believed to be created during heavy-ion collisions at ALICE. For some of the current heavy-ion analyses, the dominant uncertainty arises from this baseline measurement. It is desired to have the statistical and systematic uncertainties as small as possible for these analyses, so the baseline measurements needs to be improved. In this thesis, it is shown that the use of BDT's can serve this need and bring the field of heavy-ion physics a step closer to a deeper understanding of the properties of the quark-gluon plasma.

Before the analysis itself, the needed theory and background will be presented, along with details concerning the experimental setup at ALICE, the Monte-Carlo simulations, and the used data from ALICE. Following that, the latest ALICE D-meson analyses results will be shown, before moving on to the analysis performed in this thesis using BDT's and discussing the obtained results.

2 Introduction to heavy-ion physics

2.1 The Standard Model

Any field of physics relies on theoretical models to describe the fundamentals of that field. Particle physics is no exception and uses the so-called 'Standard Model' [4]. This model divides all fundamental particles into two distinct classes: fermions and bosons. Here, fermions are what matter is made of, whereas bosons are tied to the *interactions* that happen at the fundamental particle level. It should be mentioned that this is only true for *gauge* bosons, the reason of it will be covered later. All (currently known) fundamental particles are shown in Figure 1. Fermions are displayed on the left side, and bosons on the right. Keep in mind that any particle also has a corresponding anti-particle, although some particles, such as the photon, are their own anti-particle. When a particle is its own anti-particle it is called truly neutral.

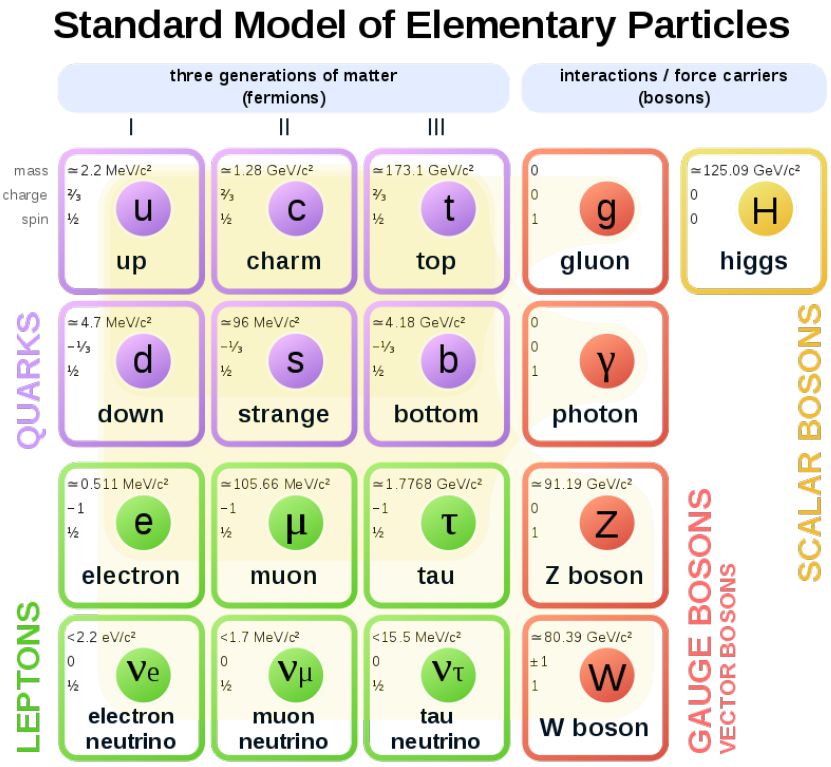

Figure 1: Elementary particles of the Standard Model [5].

2.1.1 Fermions and their generations

The fundamental fermions are again divided in two: quarks and leptons. The most important difference between these two are that quarks need to be in a bound state, meaning that they don't exist as single and free particles. Leptons, on the other hand, *can* and do exist as single, free particles.

Next to this division into quarks and leptons, there is the important classification of generations. Each of the generations has two quarks and two leptons. Of the two quarks there is a positively and a negatively charged one. The two leptons consist of a negatively charged particle and its corresponding neutrino, which is neutral. The three generations are ordered from lowest-mass to highest mass, which also means that the higher the generation the more rare and unstable the corresponding particles are. Particles that belong to the first generation are the only particles that have been observed outside of particle accelerators, with the exception of neutrinos [6] and muons [7]. Why there are three generations, and why there exist similarities between them, are questions that are still outstanding.

Quarks can combine into baryons or mesons, the difference being that the first consist of three quarks and the latter of a quark anti-quark pair. The most well-known baryons are the proton and neutron, the particles which form the nucleus of an atom. Combined with the electron, a lepton, atoms can be formed. The different type of quarks are referred to as flavours, these flavours are: up, down, strange, charm, bottom (also known as beauty), and top. Even though quarks almost exclusively combine into states that consist of two or three quarks, it is not impossible for other combinations to exist. In the last years both pentaquark [8] and tetraquark [9] particles have been identified during the analysis of LHCb data.

2.1.2 Gauge bosons

As mentioned before, gauge bosons play an important role in the interaction between particles. They are thought of as force carriers, making it possible for particles to decay into and interact with one another. Current physical knowledge tells us that there are four fundamental forces: electromagnetic, gravitational, strong, and weak forces. Of these, only the gravitational force is not described by the Standard Model.

Interactions between fundamental particles is done through the emitting and absorbing of the carriers of the force at hand. Some decay processes also involve emitting a force carrier that consequently decays.

Electromagnetic interaction has the photon as its force carrier. This interaction has an infinite range and is described by quantum electrodynamics (QED). All, and only, charged particles can emit and absorb photons as part of their interaction with other particles.

The weak force has Z and W bosons as force carriers and has the shortest effective range of 10^{-18} m. A very special property of these bosons is that they can change the flavour of a quark, as there is no other way for this to happen. Two well-known processes where this flavour change takes place are the fusion that takes place in stars and radioactive beta decay.

The strong force is tied to a property called colour charge, which applies only to quarks and their force carrier. The force that comes with this colour charge is carried by the gluon and has an effective range of 10^{-15} m. Colour charge and its behaviour is described by quantum chromodynamics (QCD).

2.1.3 Higgs boson

All the way on the right in Figure 1 we find the, famously postulated and recently found [10,11], Higgs boson, which is classified as a scalar boson. It deviates in the fact that it is the only fundamental particle with spin 0. Why it is called a scalar boson and how it is tied to the mass of other particles is beyond the scope of this thesis. However one interesting consequence of the existence of the Higgs boson is the possibility that at high energies the electromagnetic and weak force fuse into a so-called electroweak force carried by massless W and Z bosons. This fusing is part of the grand unification theory, where at high enough energies there is only one fundamental force left.

2.1.4 Spin

Each fundamental particle has a property called spin. It should be made clear that spin is mostly a mathematical structure, which *can* be measured, and does not so much as directly describe what is physically going on. However it does help to classify particles and their behaviour. Spin is one of the so-called quantum numbers, which help to define the quantum properties of a particle in a very compact way. Now, spin can either be an integer (0, 1, 2, ...) or a half-integer (1/2, 3/2, 5/2 ...).

Spin is also important for the distinction between fermions and bosons. Fermions are characterised by half-integer spin and bosons by integer spin. This means that when combining three quarks into a baryon, that baryon will also have half-integer spin and therefore be a fermion. On the other hand, when combining just two quarks to get a meson, one will find the resulting particle to be a boson since the spin will always be an integer when adding two half-integer spins. This is the reason why it should be specified what kind of fermion or boson is being considered, because combined particles are a whole different breed than the fundamental ones that were discussed here.

2.2 Quantum Chromodynamics

As mentioned before, QCD describes the way quarks and gluons interact via the strong interaction. Each quark has a certain colour charge: red, green or blue and their respective anti-colours. When combining quarks into baryons, mesons, or any other multi-quark particle, the resulting colour has to be the neutral colour white. For baryons this means each colour is represented once, since red, blue and green together make white. When combining a colour and its anti-colour, the resulting colour is neutral as well, which makes it possible to have mesons.

Gluons can also carry colour charge, and in that way facilitate a quark changing colour charge by carrying away colour- and anti-colour charge to and from other quarks. Next to facilitating interactions and colour charge changes for quarks, gluons also interact with each other by exchanging gluons themselves. One of the effects of this interaction is that gluons attract each other, resulting in a rubber-band-like term in the potential. This rubber-band-like potential has an interesting consequence called confinement: single quarks cannot be isolated when in a bound state. Imagine a quark anti-quark pair held together by the strong force, which manifests itself in the constant exchange of gluons between the pair. When separating these quarks, the attractive force increases as if one stretches a rubber band. When the potential energy of the system exceeds that of two quark masses, the connection 'snaps' creating a new quark and anti-quark that bind with the original quark anti-quark pair, forming two quark anti-quark pairs. [12]

2.3 Quark-Gluon Plasma

At extremely high temperature and/or energy density, regular hadronic matter breaks down into an almost perfect fluid [13] consisting of unbound quarks and gluons, interacting with each other through the strong force. This hot dense QCD matter is called the Quark-Gluon Plasma (QGP). A schematic view of the temperature-density space including the QGP conditions is shown in Figure 2. It also indicates what conditions the LHC can create, which has a clear overlap with those of the QGP. It is believed that the matter in the early Universe, a few microseconds after the Big Bang, consisted of a QGP. This is why experimentation on the QGP is of special interest to fundamental physics; it could help one understand how and why the Universe formed the way it did.

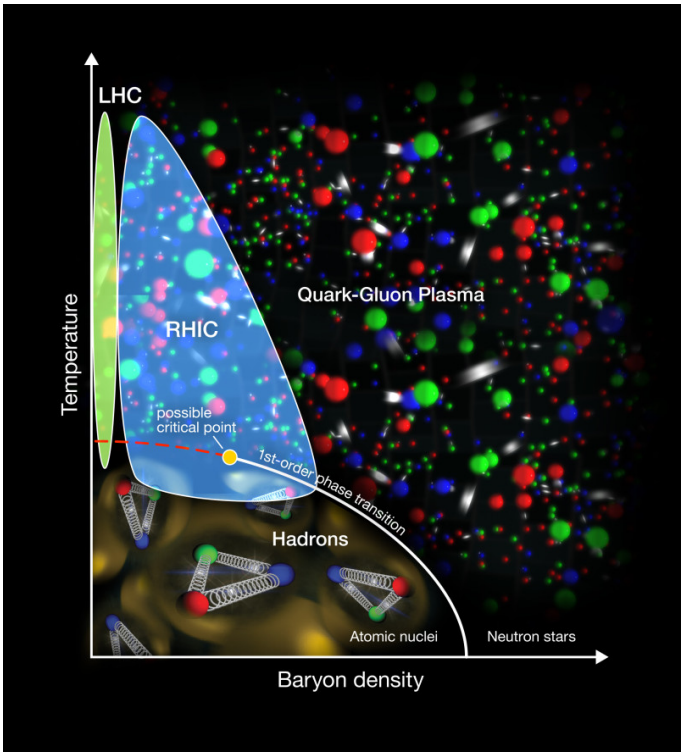

Figure 2: Schematic phase diagram of nuclear matter showing the phase transition into the quark-gluon plasma [14].

2.4 Probing the Quark-Gluon Plasma

The lifetime of the QCP at ALICE is around 10 fm/c or about $3.336 \cdot 10^{-11}$ picoseconds[15]. Since this is too little time to make direct observations, the only way to measure its properties is by probing the outcome. As the formation time of a quark is inversely proportional to its mass, heavy quarks are formed early in the collision, with charm quarks formed on a time scale of about 0.6 $\frac{\text{fm}}{c}$. The plasma is expected to form later, at about 1 fm/c. So when measuring charmed particles that are creating during a QGP-inducing collision, one has a very high certainty that those particles consist of charmed quarks formed before the plasma. Quarks and gluons that traverse the plasma are prone to thermalisation, which is caused by the loss of energy to gluon radiation and elastic scattering. For heavy quarks, gluon radiation is greatly suppressed by the dead-cone effect, due to their large mass[16]. Therefore the heavy quarks are expected to loose less energy than lighter quarks (and gluons) while traversing the plasma, and have a thermalisation time longer than the lifetime of the QGP[17]. So the heavy quarks that are created before the QGP formation can escape the plasma and hadronise forming the heavy-flavour particles that can be observed in the ALICE detector. The aforementioned properties make heavy quarks, and therefore heavy-flavour particles, ideal to probe the QGP.

To quantify the effects of the QGP on heavy quarks in heavy-ion collisions, a baseline measurement is done by analysing collisions that do not produce the conditions needed for the plasma to form, but do produce the heavy quarks of interest. This is where the analysis of D^+ and D^- mesons created during proton-proton collisions, which lack the energy density required for the QGP, contributes to the currently ongoing QGP research.

3 A Large Ion Collider Experiment at CERN

Situated near Geneva is the European Laboratory for Particle Physics, or CERN. Historically this was the European Council of Nuclear Research, from which (when translated to French) the acronym CERN originates. At CERN all efforts are focused on research concerning what the Universe is made of and how it works. For this they built the Large Hadron Collider (LHC), the worlds largest and most powerful particle accelerator (at the time of writing). It accelerates particles in a circular ring, with circumference of 27 km, to velocities close to the speed of light. Powerful and superconducting magnets are used to keep the particles in the circular trajectory, and to 'squeeze' the particles together before collision. There are multiple detectors and experiments at CERN to measure and understand what happens during these collisions. This thesis will use data from 'A Large Ion Collider Experiment' (ALICE). The corresponding detector is discussed in the next few sections. Probing the QGP is the main purpose of the ALICE experiment. It should then be no surprise that the experiment is optimised for heavy-ion collisions and for the highest possible LHC energy proton-proton collisions.

3.1 The ALICE detector

There are many sub-detectors within ALICE's main detector, many of which are not directly relevant to this thesis. Therefore only the detectors which are used to reconstruct heavy-flavour candidates will be discussed shortly in the next sub-sections. For more in-depth information about ALICE's detector and its sub-detectors see the publication on ALICE at CERN [1].

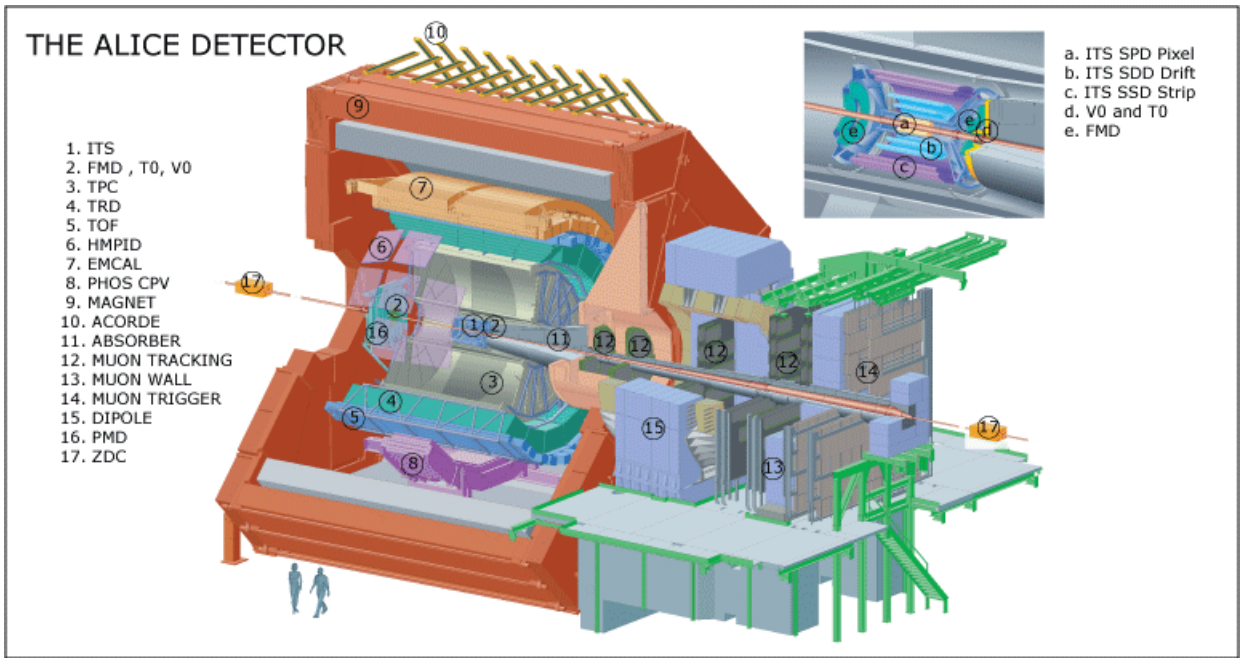

Figure 3: The ALICE detector and its sub-detectors [18].

3.1.1 Inner Tracking System (ITS)

The ITS is located directly around the beam pipe where the collisions take place. It consists of six cylindrical layers. The first two layers, at a radius of 3.9-7.6 cm, are the Silicon Pixel Detectors (SPD), followed by two Silicon Drift Detectors (SDD) at 15.0-23.9 cm, and then finally two double-sided Silicon micro-Strip Detectors (SSD) at 38.0-43.0 cm. Figure 3 shows a three-dimensional view of the ITS in the top-right corner.

One of the main uses for the ITS is the localisation of the primary vertex of the collision, and the decay vertices of short living resonances (like D-mesons or charmed baryons). Furthermore, the ITS is well-suited to track and identify particles with momentum between 80 MeV/c and 200 MeV/c, and it contributes to improve the momentum and angle resolution of the time projection chamber.

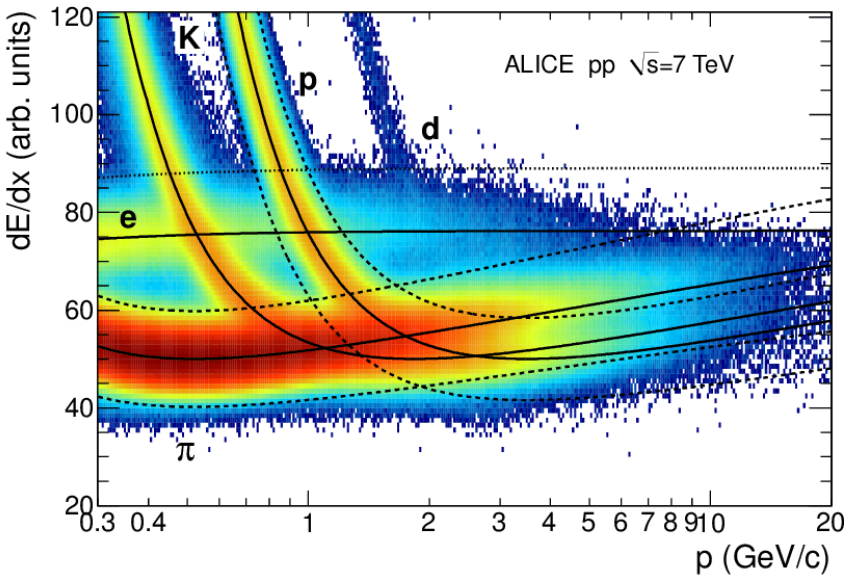

Figure 4: Specific energy loss in the TPC as a function of momentum with superimposed Bethe-Bloch lines for various particle species [19].

3.1.2 Time Projection Chamber (TPC)

Surrounding the ITS is the TPC, labelled '3' in Figure 3. This cylindrical detector, with an inner radius of about 85 cm and an outer radius of 250 cm, tracks all charged particles that move through its detection gas comprised of Ne/CO2/N2 (90/10/5). It is the main tracking device of ALICE, giving up to 160 points per track. The curvature of reconstructed tracks due to the active magnetic field is used to calculate the momentum of a tracked particle. In addition the TPC assures particle identification via specific energy loss, as the energy lost per distance $\frac{dE}{dx}$ is an excellent identifier for different particle species. This energy loss arises from ionizing the gas a charged particle is traversing, and is described by the Bethe Bloch formula. Figure 4 shows the specific energy loss in the TPC as a function of momentum, along with superimposed Bethe-Bloch lines for various particle species. The TPC can measure the specific energy loss of a particle with a resolution between 5.5% and 6.8%.

3.1.3 Time Of Flight (TOF) detector

The TOF detector starts at a radius of 370 cm and ends at 399 cm. It consists of 90 modules, each with 24 Multi-gap Resistive-plate Chamber (MRPC) strips. These MRPC strips each have a time resolution of 40 ps or better.

Measuring the time from collision to a particles arrival at the TOF detector is essential for the final particle identification procedure for pions and kaons with momentum below 2.5 GeV/c and protons with momentum below 4 GeV/c. This identification is done by using the time-of-flight of a particle to calculate its velocity, which, when paired with the momentum of the particle, gives great identification power. Figure 5 shows the particle velocity, in $\beta = \frac{v}{c}$, distribution as a function of momentum, along with labels for the different particle species that can be identified.

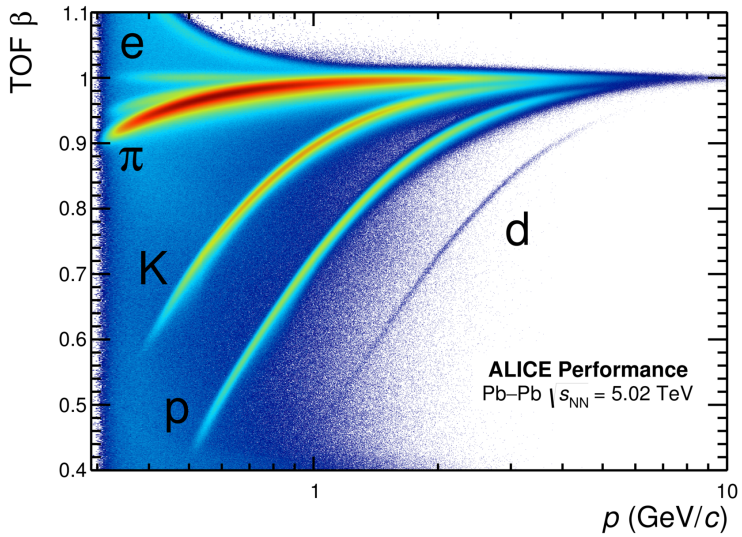

Figure 5: Particle velocity (β) distribution as a function of momentum, retrieved from Pb-Pb collisions at centre-of-mass energy $\sqrt{s} = 5.02$ TeV[20].

4 Introduction to decision trees and boosting

4.1 Decision trees

A decision tree is a structure used to make binary decisions depending on certain input variables. When working with particle physics data this decision is whether a certain input should be classified as either signal or background. In Figure 6 a simple two-layered tree is shown. This tree is merely an example with some real variables, it does not make any useful selections. The exact physical meaning of these variables will be discussed in a later section. When a data point is passed into it, it first checks the top condition: whether the variable 'pT', which stands for the momentum of the candidate particle in the transverse direction from the beam, is below the value of 5.2. If so, the data point is passed to the left, if not it is passed to the right. On the left and right the same type of check is done, but with 'invM' (invariant mass of the candidate particle) and 'cos_p' (cosine of pointing angle) as the variables being checked, respectively. In the end, each data point will end up in the locations labelled 'A', 'B', 'C', or 'D' called 'leaves', where the data point is classified as either signal or background. For instance, if after testing this tree on a dataset with known signal and background data points leaf 'B' is filled with 72% signal, it will be a signal-type leaf. So when a data point ends up in this leaf 'B' it will be classified as signal.

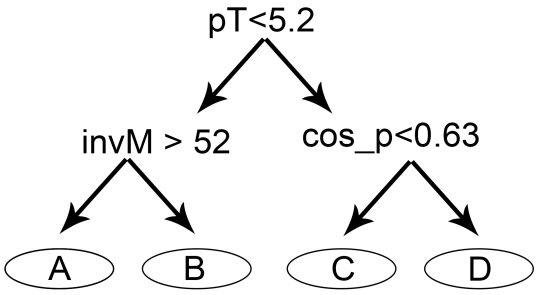

Figure 6: Simple example of a decision tree.

4.2 Growing a decision tree

Before one can grow a working decision tree, a training dataset is needed. The training set should be representative for the actual data to which the tree will be applied. In the case of this thesis the training data is obtained from Monte-Carlo simulations. Growing the tree starts by picking the first selection criterium. The simplest way to do this is in a 'greedy' manner: pick the variable, and condition on that variable, that splits the dataset the best. A good split is characterised by both resulting partitions of the data having a good purity. After finding the best criterium, the same process is recursively applied to the partitions. This algorithm stops when for instance the amount of maximum amount layers has been reached, the amount of data in a leaf is too low or a partition can not be improved by splitting it, or any other criterium that can be defined by the user.

4.3 Boosting

A single decision tree is very sensitive to changes in a dataset, leaving a lot desired to its performance. An improvement to decision trees called 'boosting' can be very beneficial. In general boosting uses a multitude of smaller trees (typically around 3-5 layers) to make decisions on the classification of a data point. Given a certain input, the boosted decision tree (BDT) will output a real number between -1 and 1. The idea is that an output close to -1 describes a background-like data point and an output close to 1 a signal-like data point. Where to cut this output for the decision on a data point being background or signal (the working point) is done afterwards by the user, or by the program when supplied by the criteria to be optimised.

The boosting technique used in this thesis is called 'gradient boosting'. The boosting procedure starts with building a small decision tree on the training data. After this the quality of the tree is assessed with a loss-function. The goal is to gradually reduce this loss function by sequentially adding the weighted

output of differently trained trees. Every time a differently trained tree is added, it should increase the performance on data points that were classified wrongly, without altering it on correctly classified data points. At each step gradient descent is leveraged to both weight the data on which the new decision tree is grown, and weight the new decision tree itself.

4.4 Testing

After training a BDT, one needs to test it before application. For this, part of a Monte-Carlo dataset is used, which has not been used for training. The BDT is applied to this testing Monte-Carlo dataset, and then compared to the results when applying the BDT on the training set. An important comparison is the so-called 'response' from the BDT. Here, all output values from the BDT are plotted in a histogram for both signal and background type data. If the response from the BDT differs too much, 'overtraining' could be in effect. Overtraining is when a machine learning tool has learned to give the right answer specifically to the training dataset, and not to the general problem at hand.

5 Implementing boosted decision trees

The implementation of BDTs in this thesis is done using ROOT [21] v6.14/08 and its Toolkit for Multivariate Variable Analysis (TMVA) [22]. All scripts are written in C++, all running on the Ubuntu Linux distribution version 18.04.1 LTS.

5.1 Preprocessing

Before BDTs are used, the raw track info from the ALICE detector needs to be preprocessed. Using standard ALICE filtering, an AOD friendfile is created, from which good D^+ candidates are selected that undergo some very loose topological cuts (so signal efficiency is basically 100%). All candidates are then subjected to the topological selection cuts of the standard analysis and given a flag that tells whether it is selected or not. The final result is a TTree that contains only candidates without any information about the relevant collision events, which can be used as input in TMVA.

The D^+-meson reconstruction is done from the track and PID data by selecting particle triplets that match the decay products of the decay channel $D^+ \rightarrow \pi^+\pi^+K^-$ (and its anti-particle equivalent). A schematic view of this reconstruction from a particle triplet is shown in Figure 7. The D^+ decay vertex (blue dot) and momentum vector is reconstructed from the track and kinematics information of the decay products. The dashed line in Figure 7 shows the vector from the primary vertex (red dot, where the collision took place) to the D^+ decay vertex. The angle between this dashed line and the D^+ momentum vector is shown as θ, which is known as the 'pointing angle'. Taking the cosine of this angle tells you how much the two vectors point in the same direction, with a value of 1 being a perfect overlap, and a value of -1 resulting from them pointing in exactly opposite directions. This makes it very useful for selection in both earlier and later stages.

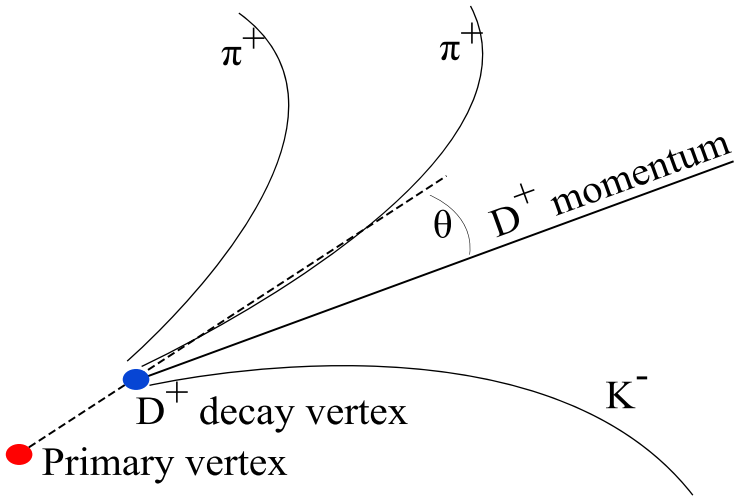

Figure 7: Schematic view of a D^+-meson reconstruction from a particle triplet.

5.2 Training

5.2.1 Training and testing data

The data used for training and testing consists of two types: background and signal. All training data is partitioned into the p_T-bins [2,3], [3,4], [4,5], [5,6], [6,7], [7,8], [6,8], [8,10], [10,12], [12,16], and [16,24] GeV/c. Background data is divided into 75% for training and 25% for testing. The signal dataset is divided into 90% for training and 10% for testing, since there is less signal statistics available. The testing set is used to evaluate over-training of the BDT and to retrieve signal and background efficiencies. Each p_T bin is trained separately, resulting in a BDT used only for application on data within that p_T bin.

Signal data is taken from the Monte-Carlo simulation dataset (LHC18a4a_cent), for which heavy-flavour enriched events with forced $D^{+(-)} \rightarrow K^{-(+)}\pi^{+(-)}\pi^{+(-)}$ decay channels were generated with the Pythia 6.4.25 event generator [23] (using the Perugia 11 tune [24]). Background data, on the other hand, is taken

from the ALICE dataset sidebands at three sigma from the mean[1], for each p_{T}-bin separately. This is done to minimise any bias in the analysis, because the simulation does not match the data completely, especially in the low p_{T} regions. In Figure 8 the variables for a D^+ candidate are shown, for both data and MC, within p_{T} range from 3 to 4 GeV/c. Each plot has a bottom panel that shows the data/MC ratio, highlighting the extent of this mismatch.

5.2.2 Training variables and parameters

All variables used for training are shown in Table 1, along with their definitions. Even though redundant, the transverse momentum of the candidate particle has been left in as training variable. For future reference, it is better practice to leave it out. Decay length (d_len) is the distance from the point of collision (primary vertex) to where the candidate particle decays (secondary vertex). The pointing angle is the angle between the momentum of the candidate particle, and the path from primary to secondary vertex, also known as the decay path, and is depicted as θ in Figure 7. The impact parameter is defined as the transverse distance between the two colliding proton nuclei. Training is not done on the invariant mass, since it would result in signal being strongly selected on the invariant mass creating a bigger invariant mass peak which causes the fitting procedure to overestimate yield.

pt_cand	Transverse momentum of candidate
d_len	Decay length of candidate
d_len_xy	Decay length of candidate in the plane perpendicular to the beam direction (xy-plane)
norm_dl_xy	d_len_xy divided by its error
cos_p	Cosine of the pointing angle
cos_p_xy	Cosine of the pointing angle in the xy-plane
imp_par_xy	Impact parameter in the xy-plane
pt_prong0	Transverse momentum of the first pion from the D^+ decay
pt_prong1	Transverse momentum of the Kaon from the D^+ decay
pt_prong2	Transverse momentum of the second pion from the D^+ decay
sig_vert	Dispersion of the primary vertex

Table 1: All variables used for training and their definitions.

Each BDT is trained using gradient boosting for its boosting algorithm with a total number of 500 trees. Each tree has a maximum depth of 2, and a minimum node size of 2.5%. Shrinkage, which controls how much the learning rate is reduced at each step, is set to 0.10. Lastly, nCuts, the number of cuts done on a variable to find a value with good separation, is set to 50. All training options and their meaning can be found in TMVA's user guide[22].

[1]Here, both sigma and the mean are taken from a fit to the Monte-Carlo pure signal invariant mass spectrum.

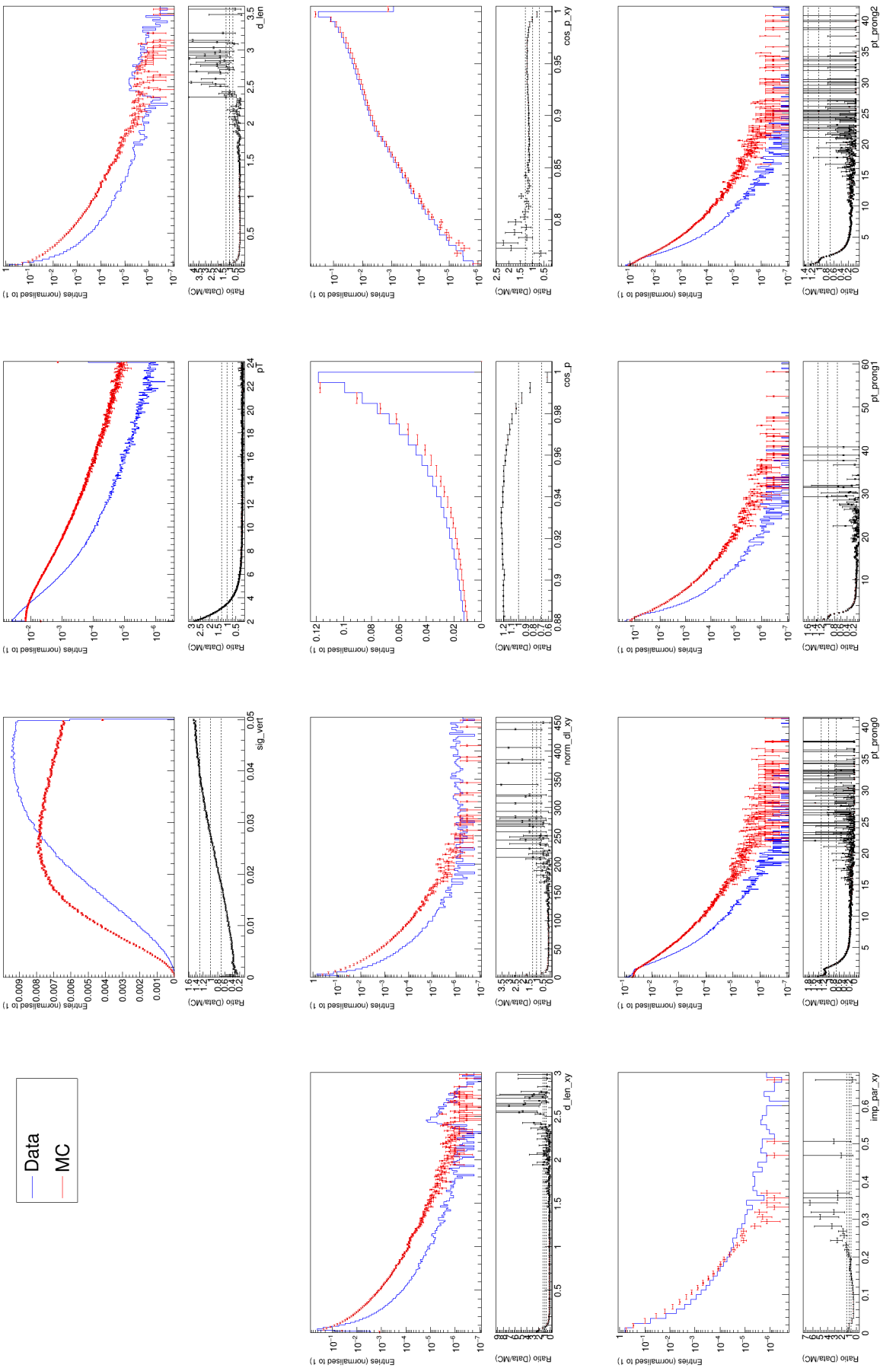

Figure 8: Spectra for all training variables for a D^+ candidate, for both data and MC. Histograms are normalised by integration and plotted together for each variable, each with a bottom panel showing the data/MC ratio after normalisation.

5.3 Validation and application

Before applying a trained BDT, it must be validated. In this thesis the validation that takes place directly after training is done by checking for over-training. As mentioned before, this is done by applying the BDT to both the training and testing dataset, and comparing their responses. An example of such a comparison can be seen in Figure 9, with all such comparison plots available in Figures 21 and 22 in the Appendix. The plot shows a clear agreement between both responses, indicating that there is no over-training. The final validation takes place after obtaining results, where the BDT is checked for its stability with respect to the chosen cut value on its output.

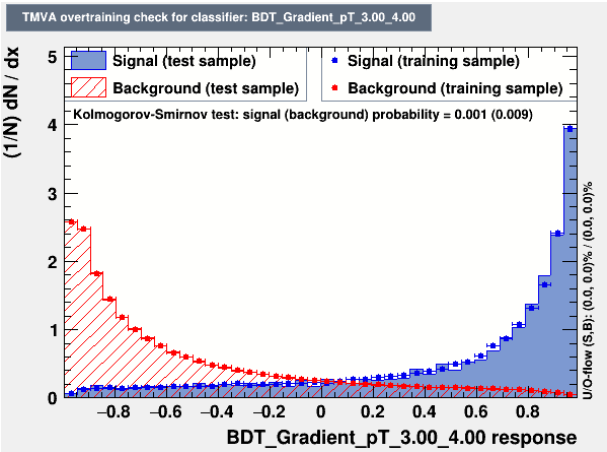

Figure 9: An example of the over-training check by comparing the training and testing sample responses of the BDT for $3 \leq p_{\mathrm{T}} \leq 4$ GeV/c.

Data for application is taken from ALICE proton-proton collisions at centre-of-mass energy $\sqrt{s} = 5.02$ TeV in period LHC17p using the CENT with SDD reconstruction. Note that only one of the two reconstructions[2] is used, so there is more statistics available than used here. During application, each D-meson candidate has its variables used as input for the BDT, generating a response which is saved along with the values of all variables. This response is also added to a histogram to eventually display the total response of the BDT. When classifying candidates, a cut value on the BDT response is chosen. When the response for a candidate is higher than that cut value, it is classified as signal and added to a histogram. When this is done for all candidates in a p_{T} bin, the histogram is saved and the yield extraction procedure can start.

[2]The FAST reconstruction is not considered.

6 D^+-meson signal extraction

After classifying the data, the next challenge is to extract the amount of signal and background in the sample. For this a model needs to be fit to the invariant mass spectrum of the sample, after which we can integrate the fitted function. After extracting the amount of signal in the sample, a correction needs to be performed to retrieve the amount of signal in the pre-processed dataset. This is done by correcting for the efficiency with which signal is selected. This entire procedure is described in the next subsections.

6.1 Fitting procedure and raw yield extraction

The combined background and signal fitting function is defined in Equation 1 as a Gaussian superposed with a second-order polynomial, where the Gaussian represents the signal and the polynomial represents the background:

$$f(x) = \frac{A}{\sigma \cdot \sqrt{2\pi}} \cdot \mathrm{e}^{\frac{(x-\mu)^2}{2\cdot\sigma^2}} + a \cdot x^2 + b \cdot x + c. \tag{1}$$

First the second order polynomial is fit on the sidebands three sigma from the Gaussian mean[3], where both the sigma and the mean values are extracted from a Gaussian fit on the appropriate Monte-Carlo data. This is done to get good starting points for the parameters $a, b,$ and c. Starting points for σ and μ are also extracted from the fit to the Monte-Carlo data. For the parameter A, the amplitude is estimated and divided by the normalisation factor to get a good starting point. Then the combined function is fitted to the sample. An example of a completed fit is shown in Figure 10, where the combined function is plotted in red. The polynomial, or background, part of the fitted function is plotted as a dashed black line. The fit itself clearly follows the data, with a mean close to the mass of the D^+-meson at 1869.5 $\pm$ 0.4 MeV/c^2 [25].

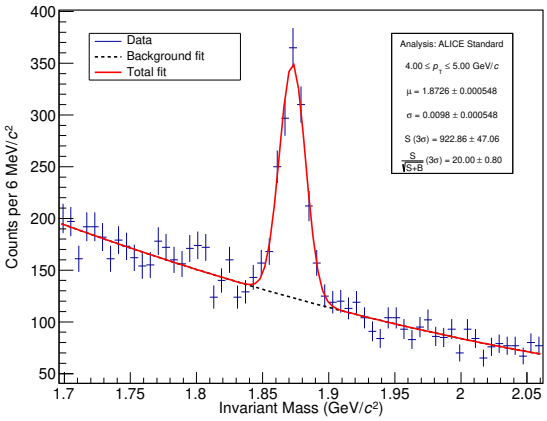

Figure 10: Fit to the invariant mass spectrum from the ALICE standard selection in p_{T} range from 4 to 5 GeV/c.

With the function fitted to the data, the extraction of raw yield can start. The implementation of this is taken directly from AliPhyscis, the standard software used by the ALICE heavy-flavour analysis group, functions[4]. All signal and background yields are extracted from within three sigma around the mean from the fitted Gaussian function.

For the raw background yield, the fitted background function is integrated and divided by the binwidth of the histogram. The error is estimated by the average squared error of the bins in the three sigma sidebands, normalised to the yield count. The raw signal yield is extracted by integrating the Gaussian

[3] this is done by excluding points that fall within three sigma of the mean.

[4] Functions are named 'Signal', 'Background' and 'Significance'. They are available at `https://github.com/alisw/AliPhysics/blob/master/PWGHF/vertexingHF/AliHFInvMassFitter.cxx`.

signal function. The error is estimated by normalising the error of fit parameter A with respect to the yield count.

The significance is computed as $\frac{S}{\sqrt{S+B}}$, where S and B are, respectively, the extracted raw signal and background yields. The error is estimated by propagating the errors of the raw signal and background yields. This significance is directly related to the statistical uncertainty on the measurement, where a higher significance indicates a more precise measurement. The signal-to-background ratio is computed as $\frac{S}{B}$ and errors are again estimated by propagation.

6.2 Efficiency correction

The efficiency is obtained by classifying the candidates from the testing data, and counting how many signal and background entries are classified as signal or background. The formula used here is shown in Equation 2, which can be applied to get both the background and signal efficiencies. The correction is then done by dividing the raw yield by the corresponding efficiency, after which the errors are correctly propagated. This gives the final efficiency-corrected yield.

$$\text{Efficiency} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}. \tag{2}$$

6.3 Choosing a BDT response cut

When classification is done using a BDT, a cut value on the output needs to be chosen. In this thesis the choice is made to pick the value at which the significance is maximized. To do so, a ROOT function is defined that cuts on the BDT output, goes through the fitting and yield extraction procedures and returns the significance. This function is plotted in Figure 11 for $2 \leq p_{\mathrm{T}} \leq 3$ GeV/c (left plot) and $4 \leq p_{\mathrm{T}} \leq 5$ GeV/c (right plot). Finding the maximum of this function is then done by ROOTs build-in functionality, of which the result is shown as a vertical blue line.

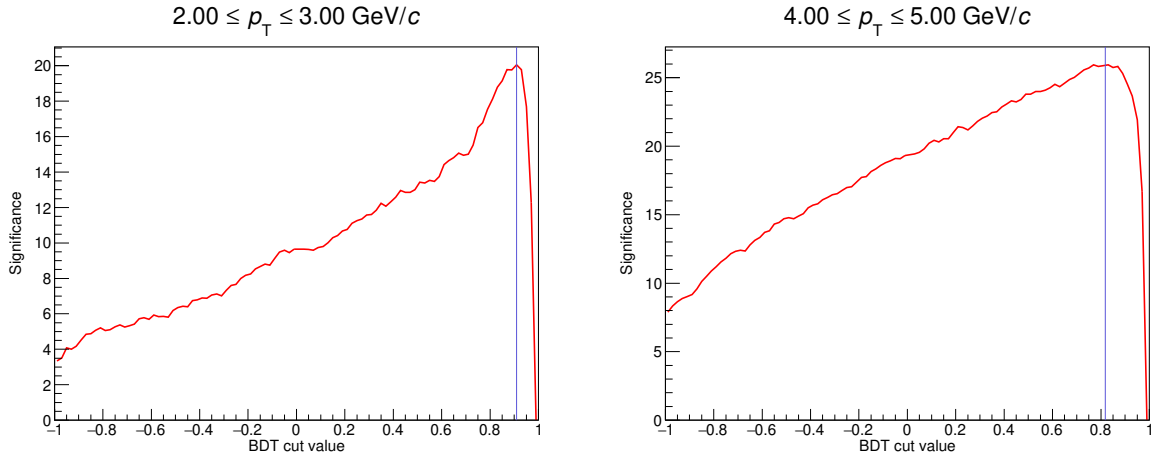

Figure 11: Significance obtained at different boosted decision tree cut values, shown in two p_{T}-bins. The maximum-significance cut is shown as a vertical blue line in both plots.

7 ALICE measurement of D^+ meson production

During the development of the BDT analysis in this thesis, the ALICE collaboration has been writing a paper based on the standard analysis, which was concluded in May of 2018. At the moment of writing this thesis, only the Preliminary Physics Summary is available. They have measured the production of D^0, D^+, D^{*+}, and D_s^+ mesons in proton-proton collisions at $\sqrt{s} = 5.02$ TeV. When compared to the previous measurement at $\sqrt{s} = 7$ TeV, this measurement should allow for a more accurate determination of the nuclear modification factor R_{AA}, which is a measure for the ratio between the yields obtained in lead-lead and proton-proton collisions. The data used in this thesis is a subset of the data used by this analysis from ALICE, containing about half the statistics. Since only the D^+-meson production is of interest in this thesis, only those results will be discussed here.

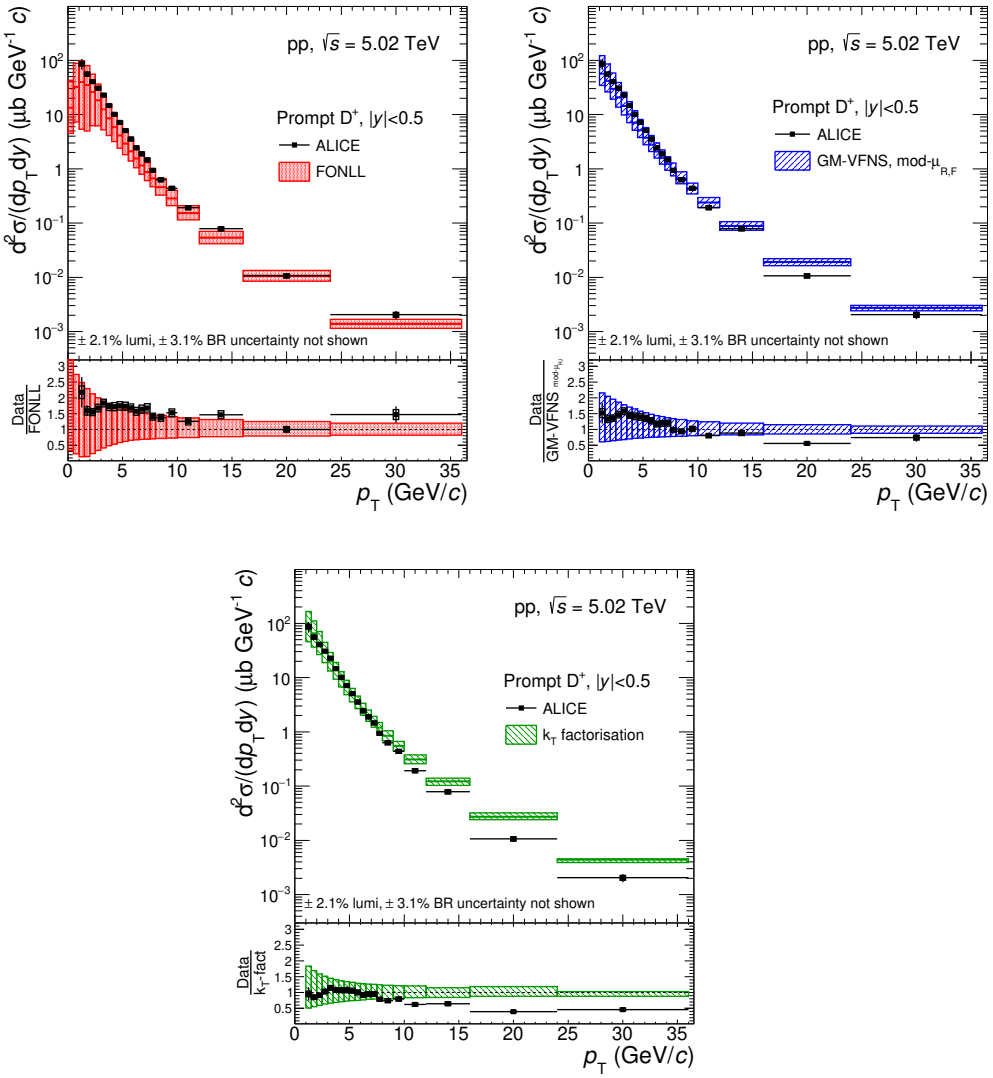

Figure 12: p_T-differential production cross sections for prompt D^+ mesons compared to pQCD calculations. The ratios of the data to the theoretical predictions are shown in the lower part of each panel. [26]

The measured prompt D^+-meson p_T-differential cross section is compared with the results of several perturbative QCD (pQCD) calculations in Figure 12. For the FONLL [27,28] and GM-VFNS [29,30] calculations the measured cross sections fall within uncertainties, lying systematically on the upper edge of the uncertainty band of the FONLL predictions. GM-VFNS tends to underestimate at low to moderate p_T, overestimating at higher p_T. The k_T factorisation model [31] starts overestimating at p_T above 7 GeV/c, describing the measurements at lower p_T values.

8 Results and comparison to standard ALICE selection

8.1 Comparing the results

To properly compare the BDT analysis to the standard ALICE analysis, the only step that should differ is that of signal selection. As mentioned before, during the pre-processing each D^+ meson candidate is given a flag that indicates whether it passes the standard ALICE cuts or not. A simple C++ ROOT script takes the data, cuts it on the p_T values depending on which p_T bin is requested, and then cuts on the before mentioned flag generating a signal and a background histogram which are saved in a .root file. The signal histogram then undergoes the same fitting and signal yield extraction procedure as the BDT generated histogram would. Finally the efficiency correction is applied, where the efficiency is calculated from the same Monte-Carlo dataset that is used for the BDT efficiency calculation. A comparison between the signal histogram of the standard ALICE and the boosted decision tree selection is shown in Figure 13. Both show a clear signal-like Gaussian peak, with the ALICE standard selection showing a little lower Gaussian peak and a bigger amount of entries in the sidebands, which is especially prominent on the left side.

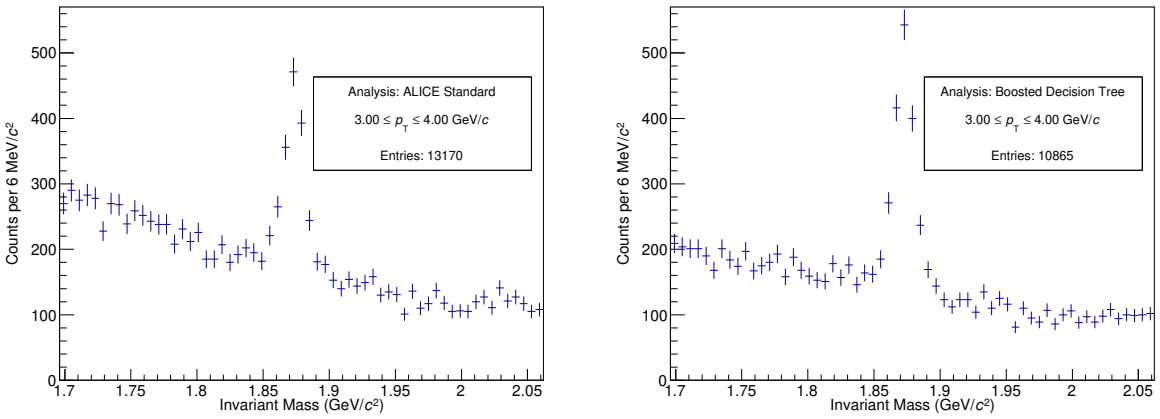

Figure 13: Invariant mass spectra of D^+ candidates with $3 \leq p_\mathrm{T} \leq 4$ GeV/c after ALICE standard (left) and boosted decision tree (right) signal selection.

Moving to the p_T-differential prompt production cross section of D^+ mesons would give a result that is directly comparable to most theoretical and other analyses at the same centre-of-mass energy. However, in this thesis the focus is a comparison between BDT-obtained results and that of the standard ALICE analysis. The equation for this cross section is:

$$\frac{\mathrm{d}^2\sigma^\mathrm{D}}{dp_\mathrm{T}\mathrm{d}y} = \frac{1}{c_{\Delta\mathrm{y}}(p_\mathrm{T})\Delta p_\mathrm{T}} \cdot \frac{1}{BR} \cdot \frac{\frac{1}{2} f_\mathrm{prompt}(p_\mathrm{T}) \cdot N^{\mathrm{D}+\overline{\mathrm{D}},\mathrm{raw}}(p_\mathrm{T})|_{|\mathrm{y}|<\mathrm{y_{fid}}(p_\mathrm{T})}}{(\mathrm{Acc}\times\varepsilon)_\mathrm{prompt}(p_\mathrm{T})} \cdot \frac{1}{L_\mathrm{int}}. \tag{3}$$

Here the raw signal yield ($N^{\mathrm{D}+\overline{\mathrm{D}},\mathrm{raw}}$) is multiplied by the fraction of signal created by the main collision (f_prompt) and halved, since both particle and anti-particles are measured. It is then divided by the fraction of signal measured by the detector and the signal efficiency, the branching ratio of the decay (BR), the width of the p_T-bin and the correction factor for rapidity correction ($c_{\Delta\mathrm{y}}$), and the integrated luminosity (L_int) which is a measure for the number of collisions. All these factors are the same for both the BDT and ALICE standard selection, except for f_prompt, which is assumed to not be too different. For the standard analysis values of f_prompt are normally of the order 0.95. Feed-down D^+ production is about 50 times less probable to occur, and the extracted signal of both analyses is of the same order, which all together support the assumption. When moving to the inclusive cross section, there should be no differences.

8.2 Raw yield

All obtained raw signal and background yields are shown in, respectively, the upper and lower plot of Figure 14. For all but the lowest p_{T}-bin, the boosted decision tree selection results in a higher raw signal yield than that of the ALICE standard selection. The raw background yield results show lower values for the ALICE standard selection when $p_{\mathrm{T}} \geq 5$ GeV/c. Plots of the fits are shown for $4 \leq p_{\mathrm{T}} \leq 5$ GeV/c and $12 \leq p_{\mathrm{T}} \leq 16$ GeV/c in Figure 15. A complete collection of all fits relevant to this thesis is available in the appendix in Figures 23 and 24.

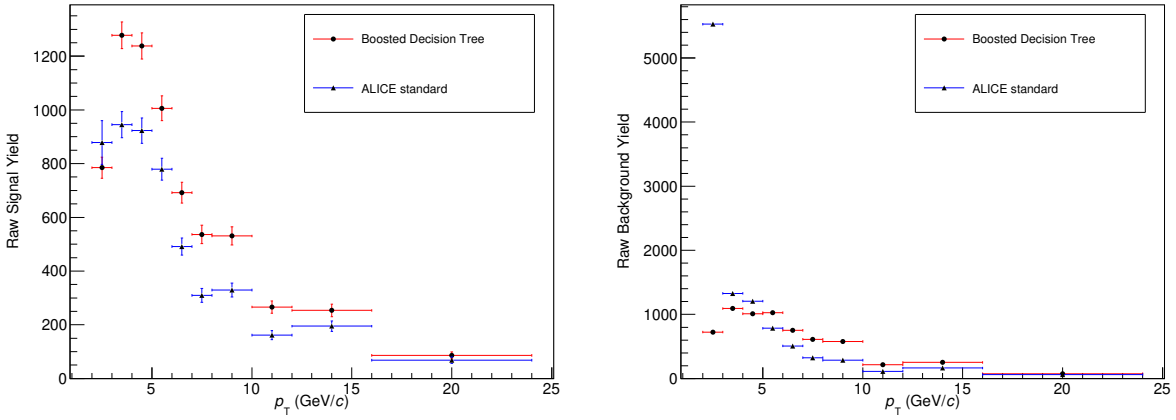

Figure 14: Raw signal yield (left) and raw background yield (right) obtained by the boosted decision tree and ALICE standard analysis.

The significance of the results is shown in the upper plot of Figure 16, with the signal-to-background ratio shown in the lower plot. The significance obtained with the BDT selection follows roughly the same shape as that from the ALICE standard selection, and has central values higher in all p_{T} bins. There is an overlap of the uncertainty bands in the last two bins, with all other bins having a clear gap between between them. Signal-to-background ratio hovers around 1 for the BDT selection, ranging from just below 0.9 to just above 1.2. The ALICE standard selection shows a slowing increase of this ratio, starting at just below 0.2 and stabilising for $p_{\mathrm{T}} \geq 10$ GeV/c around a ratio of 1.1.

The mean (μ) and sigma (σ) of the Gaussian fit are displayed in Figure 17 in the upper and lower plot, respecively. Values for μ and σ obtained from fitting pure Monte-Carlo signal in the appropriate p_{T} bin are also displayed. All BDT and ALICE standard obtained values for μ fall within their uncertainties, both mostly deviating from the Monte-Carlo values. Almost all σ values follow the Monte-Carlo values within uncertainty bands. The exception here is the ALICE standard obtained value with $10 \leq p_{\mathrm{T}} \leq 12$ GeV/c, which falls below the value of both the BDT and Monte-Carlo values. However, since the gap between the uncertainty bands is not very large compared to the uncertainty itself, and the amount of signal in this bin is relatively low, this could be attributed to statistical fluctuation. Furthermore, this lower value with relation to the Monte-Carlo is always observed in D-meson analyses in ALICE.

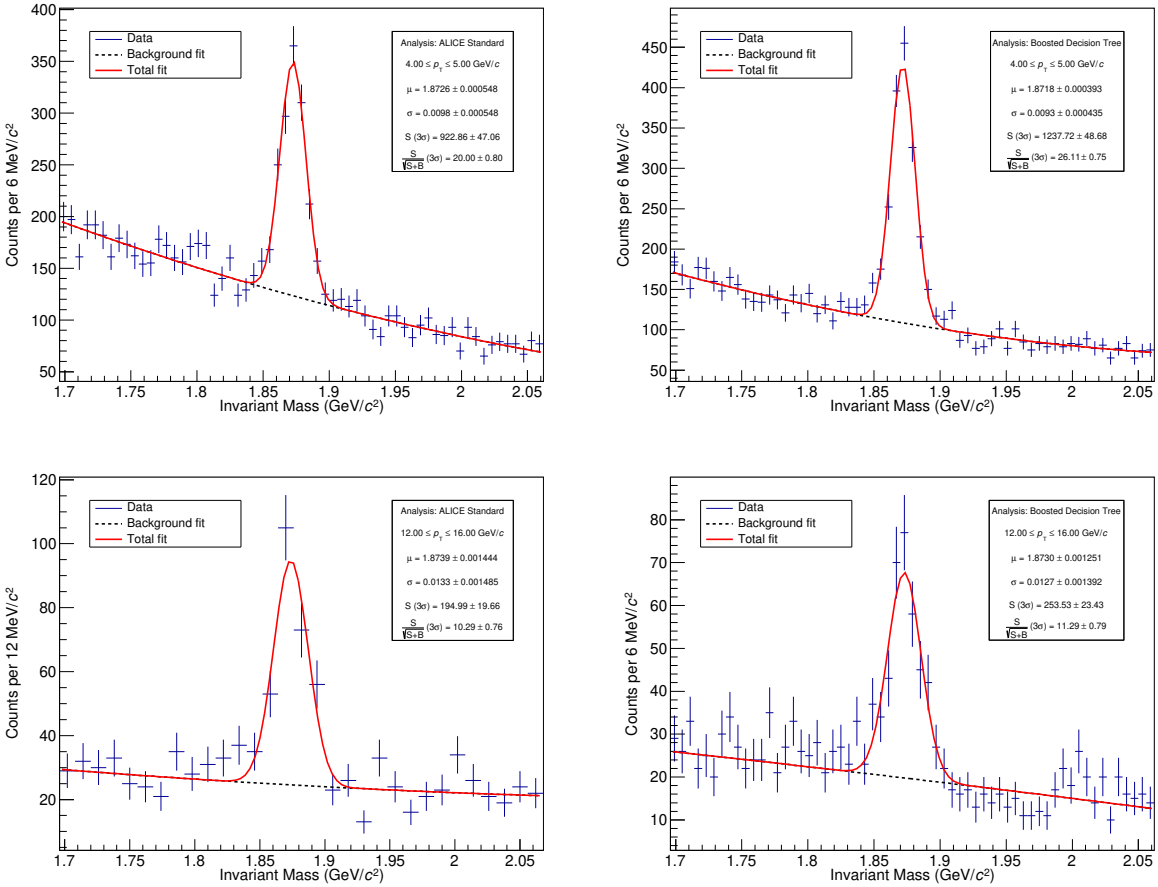

Figure 15: Fits on the invariant mass spectra from the ALICE standard (left side) and boosted decision tree (right side) selections for $4 \leq p_{\mathrm{T}} \leq 5$ (upper plots) and $12 \leq p_{\mathrm{T}} \leq 16$ (lower plots) GeV/c.

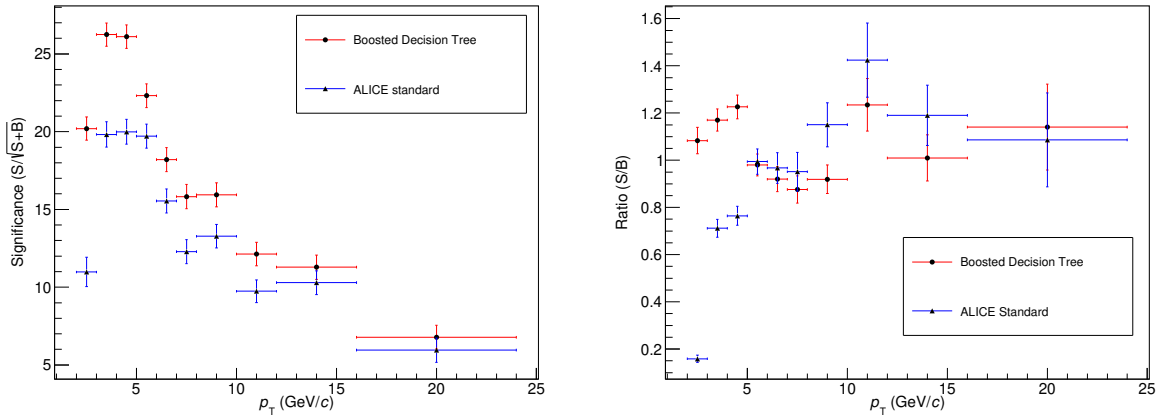

Figure 16: The significance (left) and signal-to-background ratio (right) obtained by fitting the boosted decision tree and ALICE standard selection invariant mass spectra.

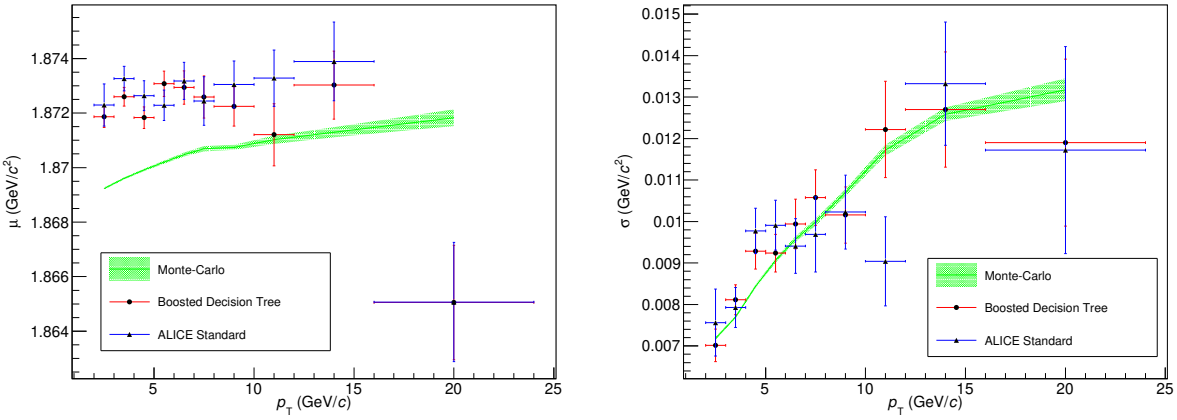

Figure 17: The Gaussian mean (left) and sigma (right) obtained by fitting the boosted decision tree and ALICE standard selection invariant mass spectra.

8.3 Efficiency-corrected yield

All results from correcting the raw signal yields for efficiency are shown in Figure 18. The efficiencies used for this correction are presented in Figure 19. The BDT and ALICE standard efficiency-corrected yields follow a similar shape, with the BDT analysis consistently producing lower central values. Most bins have values with clearly, or close to, overlapping uncertainty bands. The only p_{T}-bin that shows a significant difference between the obtained value is that of 5 to 6 GeV/c. In Figure 20 the efficiency-corrected yield in this bin is shown as a function of BDT cut value, including a line that marks the optimal significance cut value that is used in this thesis. This plot shows a consistent and stable output for cut values from -0.75 up to 0.85, solidifying the choice of the BDT cut at 0.66. A collection of this type of plot for all p_{T}-bins is available in the Appendix in Figure 25. The relatively big jump in BDT signal efficiency for $5 \leq p_{\mathrm{T}} \leq 6$ GeV/c in Figure 19 could be a hint at the root cause of the much lower efficiency-corrected yield produced by the BDT in that bin, since the corresponding raw yield in Figure 14 does not show a relatively big fall compared to the ALICE standard fall and surrounding data points.

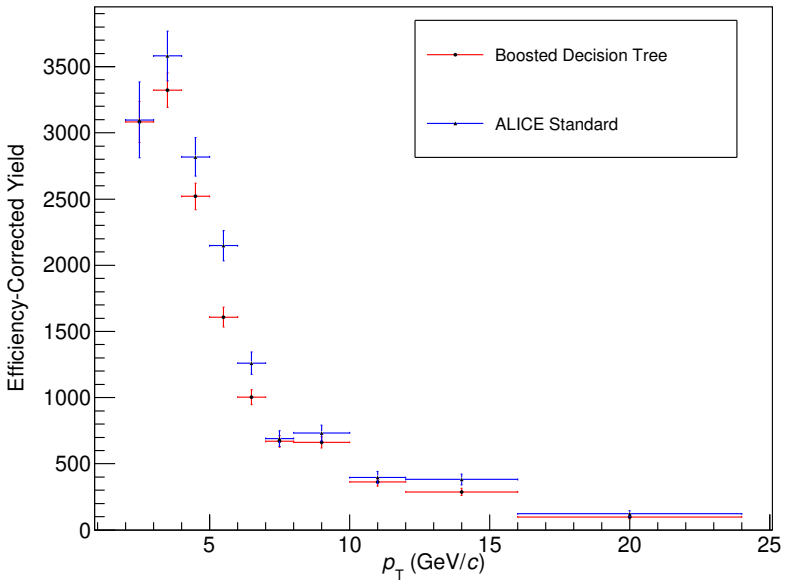

Figure 18: Efficiency-corrected yield obtained by the boosted decision tree and ALICE standard analysis.

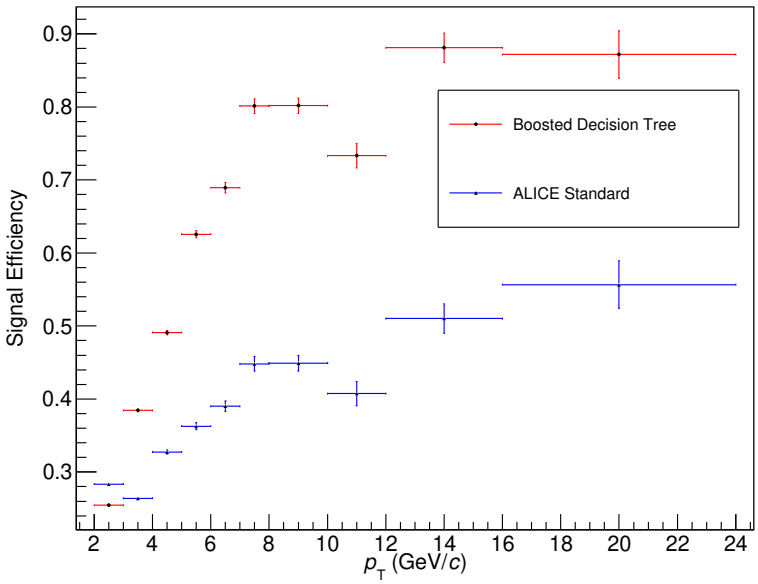

Figure 19: Signal efficiency for the boosted decision tree and ALICE standard selection.

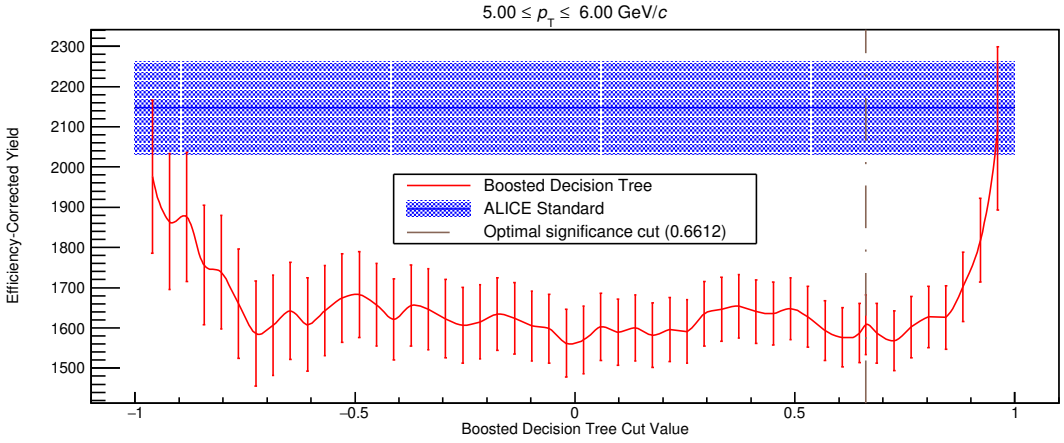

Figure 20: Efficiency-corrected yield obtained at different cut values on the boosted decision tree output.

9 Conclusion, Discussion, and Outlook

This thesis has shown that a boosted decision tree is able to make useful selections on ALICE D^+ meson candidates for raw yield extraction purposes, resulting in higher significance values than obtained with the ALICE standard selections and high signal efficiency for mid to high p_T. The efficiency-corrected yields have shown to be relatively stable when varying the cut value that is applied to the BDT output. A tension with the ALICE standard analysis is observed for $4 \leq p_T \leq 7$ GeV/c, where efficiency-corrected yields are lower. This may raise questions about the reliability of the BDT-produced invariant mass spectra. Sadly, any further investigation lies outside of the scope of this thesis. A possibly interesting use-case for the BDT-based selection, which was briefly considered for this thesis, would be to make a distinction between prompt and feed-down D^+ meson production. This would give a way to indirectly measure beauty quark energy loss in the QGP, which is very valuable since a direct measurement is extremely hard for ALICE. Furthermore, it would give yet another probe of the QGP that is expected to have even less energy loss when traversing the plasma than charm quarks.

An estimation of the systematic uncertainties is needed to properly quantify the extent of the observed deviations in efficiency-corrected signal yield, and better assess the overall quality of the BDT-produced selections. Calculating the D^+ meson p_T-differential prompt production cross-section should be done for a more in-depth comparison of the results, giving the option to include theoretical prediction models and results obtained by other researchers. There is also room for optimisations to be done on the training variables, as some variables can be excluded without losing functionality. One of the biggest limitations when using a BDT lie in the used Monte-Carlo dataset. This thesis has shown the extent to which variables from ALICE and Monte-Carlo data disagree. BDTs, and most other basic machine-learning techniques, are particularly sensitive to training data that does not properly match real-world data. It could therefore be beneficial to apply some cuts to the Monte Carlo data before training to minimise the impact of some variables that have too large of a mismatch. New and improved Monte-Carlo simulations are being worked on, leveraging machine-learning for the simulation itself. This is a great outlook for both the application of BDTs and the ALICE standard selection procedure, especially in low p_T regions.

All in all, the use of a boosted decision trees instead of the standard rectangular cuts has shown to be promising for obtaining an even more precise p_t-differential cross section measurement due to the higher significances on the extracted raw yields, validating the current ongoing development of BDT-based analysis for ALICE and giving a great prospect at supplementing the tools that particle physicists work with today.

A Appendix

A.1 Plots

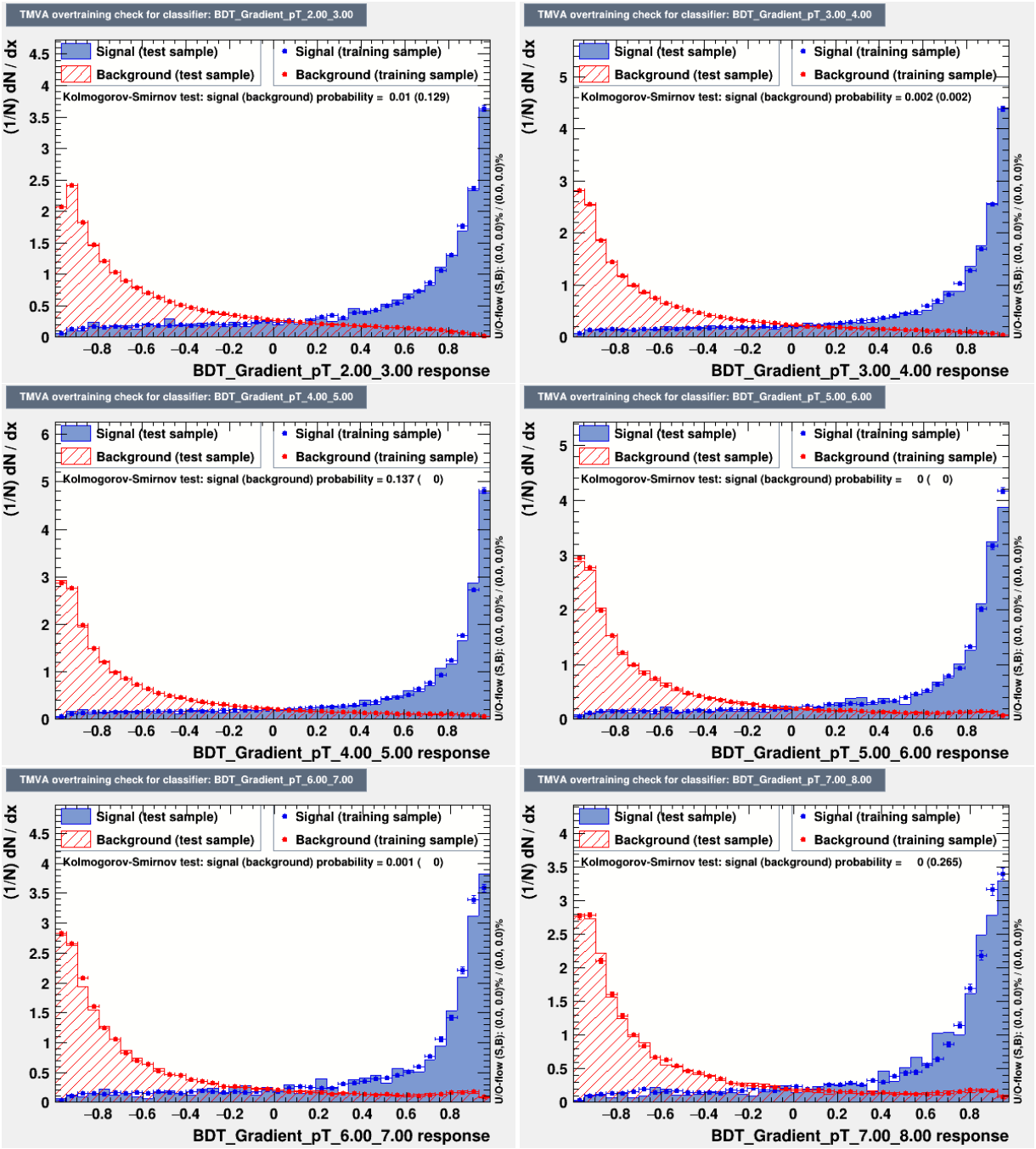

Figure 21: Over-training check in the first six p_{T}-bins.

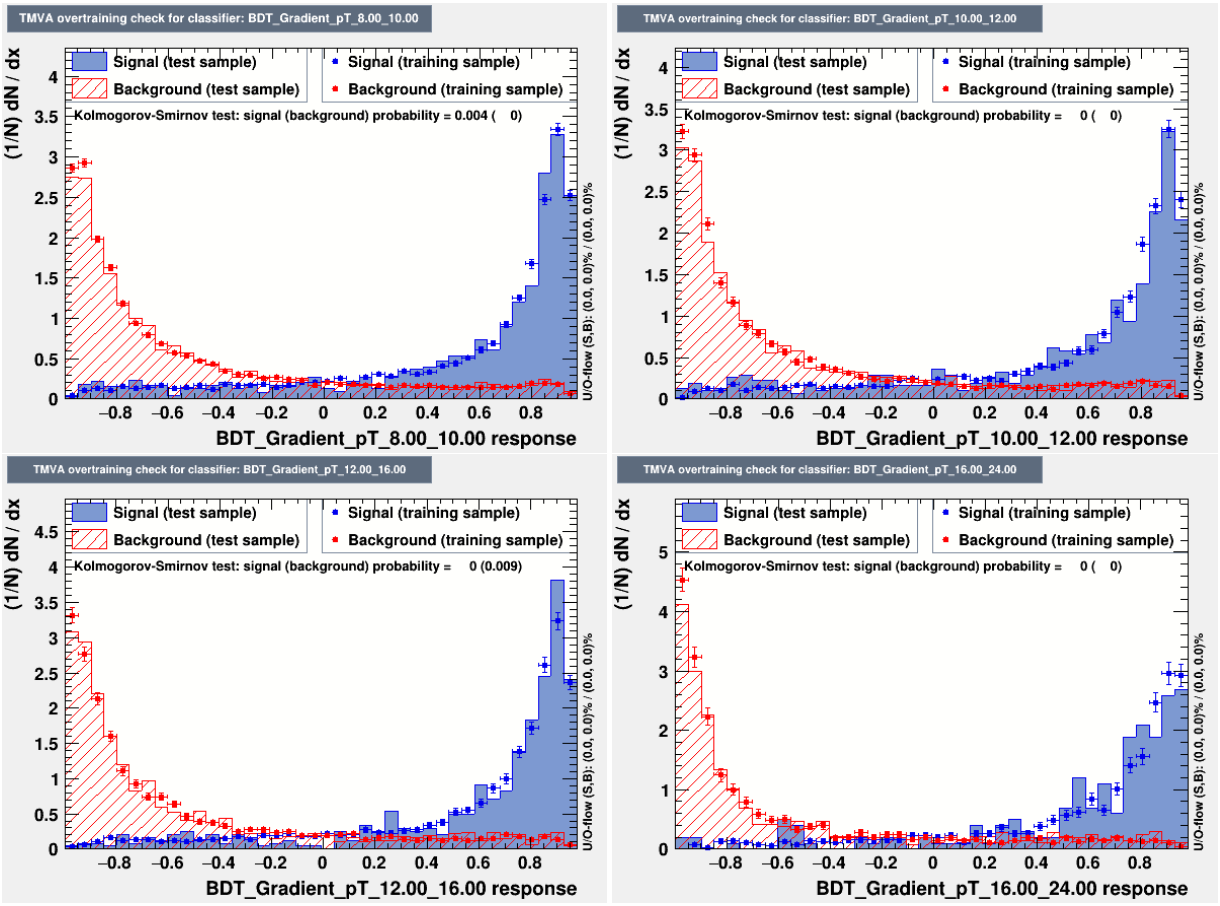

Figure 22: Over-training check in the last four p_{T}-bins.

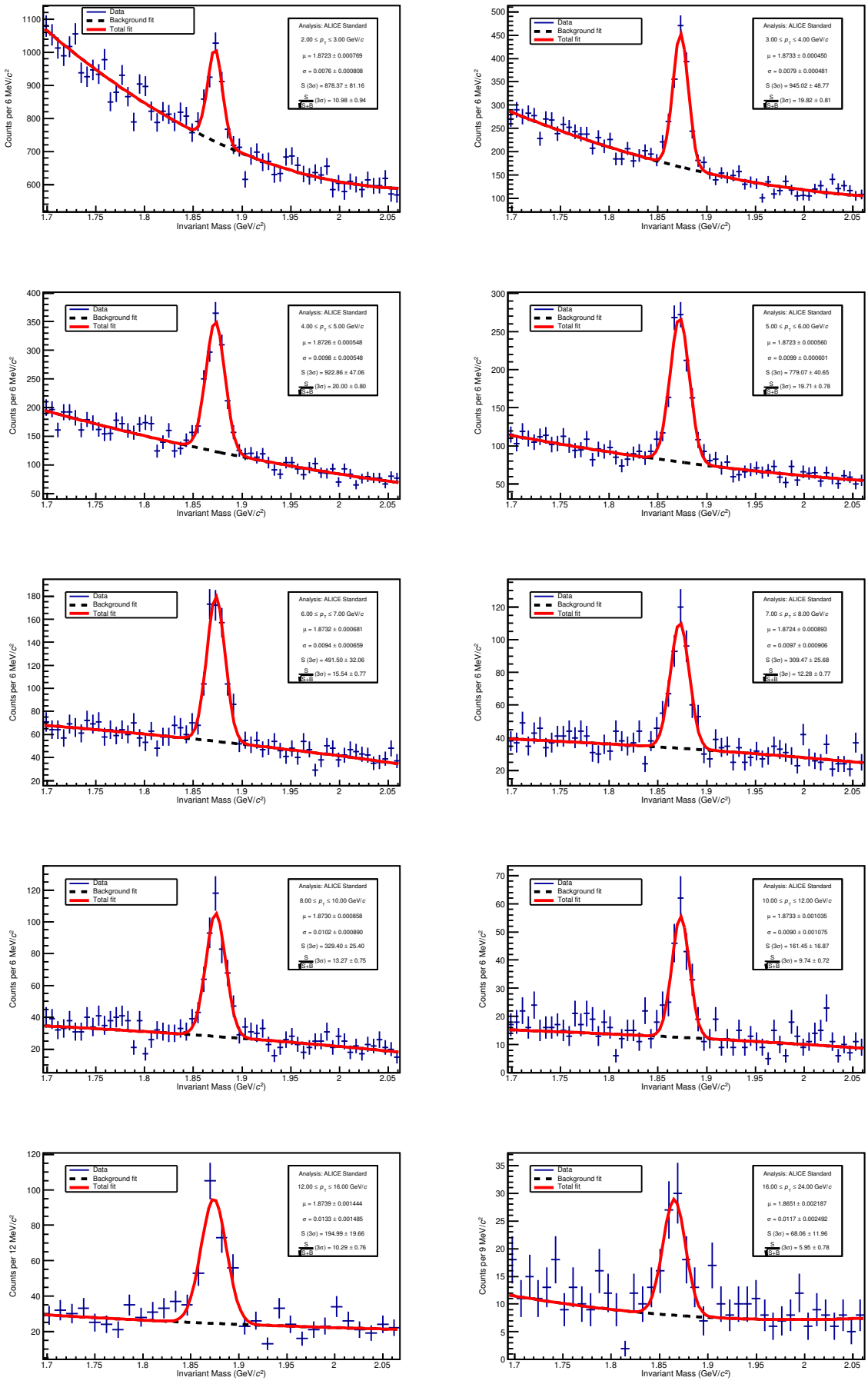

Figure 23: All fits from the ALICE standard analysis.

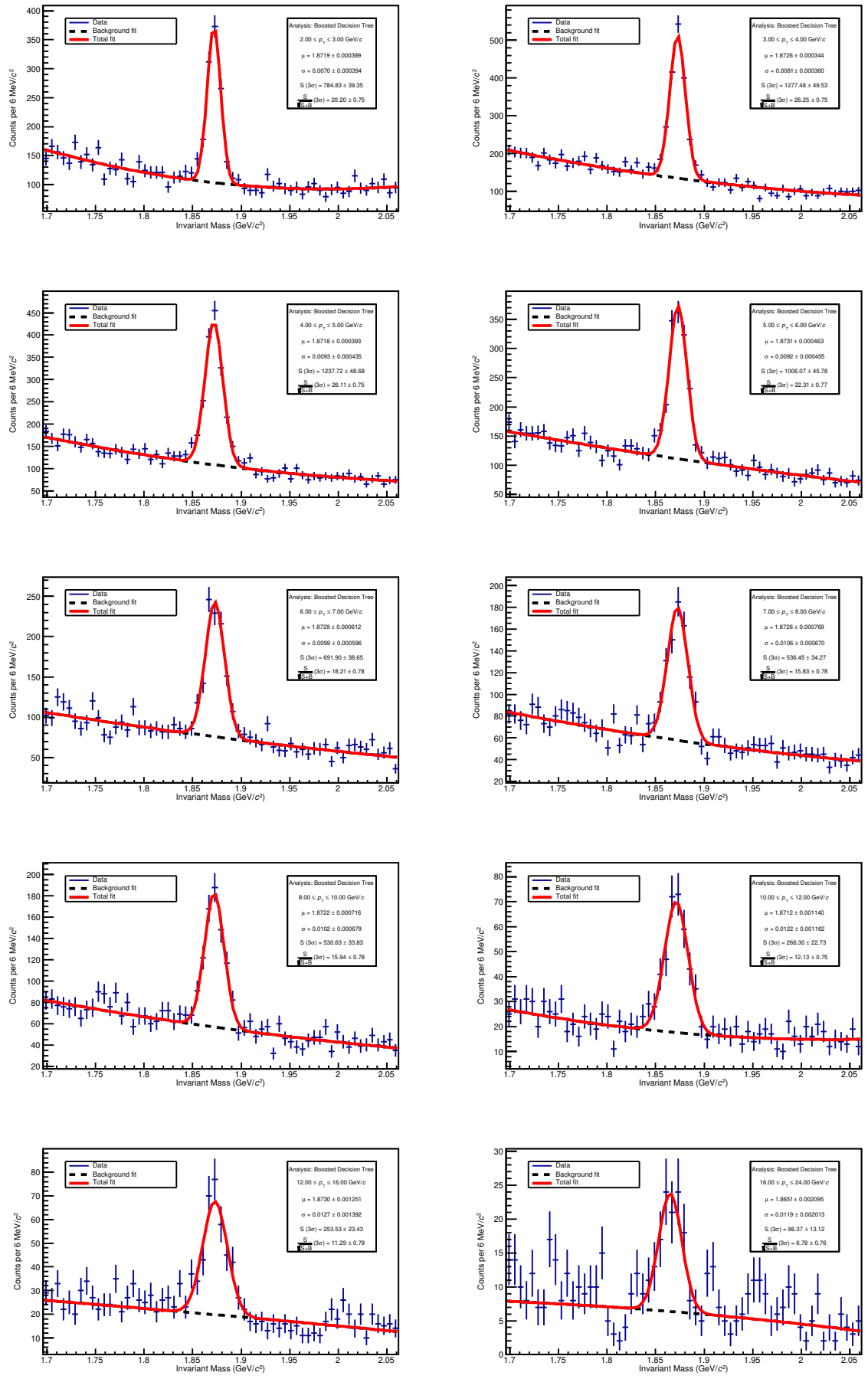

Figure 24: All fits from the boosted decision tree analysis.

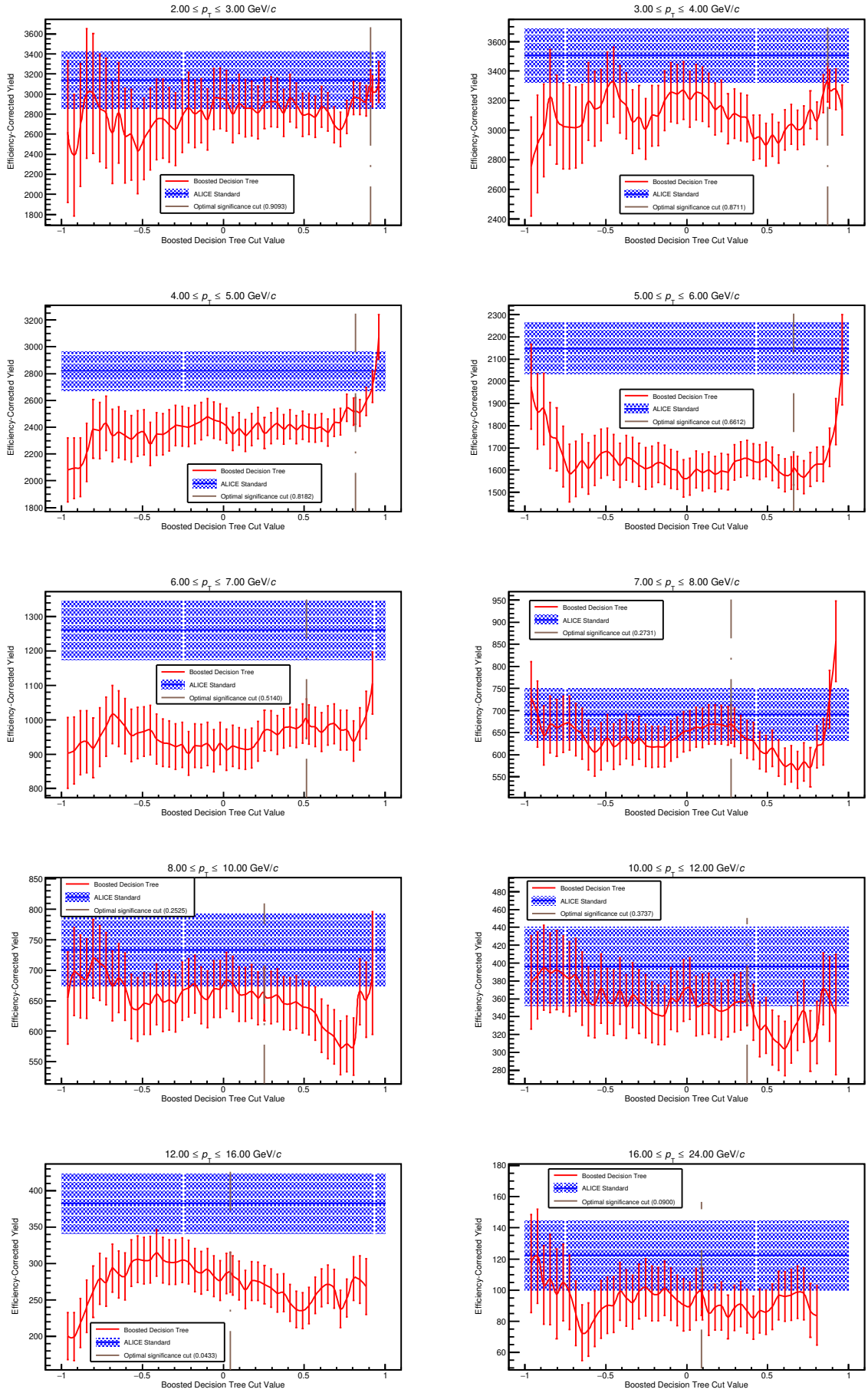

Figure 25: All efficiency-corrected yields as function of BDT cut plots.

A.2 Files and scripts

The used ALICE data and Monte-Carlo files are available at `https://drive.google.com/drive/folders/1VPpj18n8NAJO2bO2EXNhTW2c92biR5YO?usp=sharing`.

All C++ ROOT scripts are available at `https://github.com/CHBKoenders/BONZ_ALICE_BDT`.

References

[1] ALICE Collaboration, K. Aamodt, et al. The ALICE experiment at the CERN LHC. *JINST*, 3: S08002, 2008. doi: 10.1088/1748-0221/3/08/S08002.

[2] Lyndon Evans and Philip Bryant. LHC Machine. *JINST*, 3:S08001, 2008. doi: 10.1088/1748-0221/3/08/S08001.

[3] Harris Drucker and Corinna Cortes. Boosting decision trees. In *Proceedings of the 8th International Conference on Neural Information Processing Systems*, NIPS'95, pages 479–485, Cambridge, MA, USA, 1995. MIT Press. URL `http://dl.acm.org/citation.cfm?id=2998828.2998896`.

[4] Guido ALTARELLI. The Standard model of particle physics. 2005.

[5] Wikimedia Commons and MissMJ. Standard Model of Elementary Particles, 2008. URL `https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg`.

[6] IceCube Collaboration, M. G. Aartsen, et al. Observation of high-energy astrophysical neutrinos in three years of icecube data. *Phys. Rev. Lett.*, 113:101101, Sep 2014. doi: 10.1103/PhysRevLett.113.101101. URL `https://link.aps.org/doi/10.1103/PhysRevLett.113.101101`.

[7] G. Bellini et al. Cosmic-muon flux and annual modulation in borexino at 3800 m water-equivalent depth. *Journal of Cosmology and Astroparticle Physics*, 2012(05):015, 2012. URL `http://stacks.iop.org/1475-7516/2012/i=05/a=015`.

[8] LHCb Collaboration, R. Aaij, et al. Observation of $J/\psi p$ Resonances Consistent with Pentaquark States in $\Lambda_b^0 \to J/\psi K^- p$ Decays. *Phys. Rev. Lett.*, 115:072001, 2015. doi: 10.1103/PhysRevLett.115.072001.

[9] LHCb Collaboration, R. Aaij, et al. Observation of $J/\psi\phi$ structures consistent with exotic states from amplitude analysis of $B^+ \to J/\psi\phi K^+$ decays. *Phys. Rev. Lett.*, 118(2):022003, 2017. doi: 10.1103/PhysRevLett.118.022003.

[10] CMS Collaboration, S. Chatrchyan, et al. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett.*, B716:30–61, 2012. doi: 10.1016/j.physletb.2012.08.021.

[11] ATLAS Collaboration, G. Aad, et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett.*, B716:1–29, 2012. doi: 10.1016/j.physletb.2012.08.020.

[12] J. Bartke. *Introduction to Relativistic Heavy Ion Physics*. World Scientific Publishing Co. Pte. Ltd., 2009.

[13] U.W. Heinz. 'RHIC serves the perfect fluid': Hydrodynamic flow of the QGP. In *Proceedings, Workshop on Extreme QCD, 2005*, pages 3–12, 2005.

[14] Brookhaven National Laboratory. Tracking the Transition of Early-Universe Quark Soup to Matter-as-we-know-it, 2014. URL `https://www.bnl.gov/newsroom/news.php?a=24473`.

[15] Panagiota F. and the ALICE Collaboration. Overview of results from alice at the cern lhc. *Journal of Physics: Conference Series*, 455(1):012004, 2013. URL `http://stacks.iop.org/1742-6596/455/i=1/a=012004`.

[16] Y.L. Dokshitzer and D. E. Kharzeev. Heavy quark colorimetry of QCD matter. *Phys. Lett.*, B519: 199–206, 2001. doi: 10.1016/S0370-2693(01)01130-3.

[17] Shanshan Cao, Guang-You Qin, and S.A. Bass. Heavy quark energy loss and thermalization in hot and dense nuclear matter. *Journal of Physics: Conference Series*, 420(1):012022, 2013. URL `http://stacks.iop.org/1742-6596/420/i=1/a=012022`.

[18] ALICE Collaboration. The Alice Experiment, 2008. URL `http://aliceinfo.cern.ch/Public/en/Chapter2/Chap2Experiment-en.html`.

[19] ALICE Collaboration, Kjeld Aamodt, et al. Rapidity and transverse momentum dependence of inclusive J/Ψ production in pp collisions at $\sqrt{s}$=7 TeV. *Physics Letters B*, 704:442–455, 05 2011. doi: 10.1016/j.physletb.2011.09.054.

[20] Nicolò Jacazio and the ALICE Collaboration. PID performance of the ALICE-TOF detector in Run 2. *PoS*, LHCP2018:232, 2018. doi: 10.22323/1.321.0232.

[21] R. Brun and F. Rademakers. ROOT: An object oriented data analysis framework. *Nucl. Instrum. Meth.*, A389:81–86, 1997. doi: 10.1016/S0168-9002(97)00048-X.

[22] A. Hoecker et al. TMVA - Toolkit for Multivariate Data Analysis. *[arXiv] physics/0703039*, March 2007.

[23] T. Sjostrand, S. Mrenna, and P.Z. Skands. PYTHIA 6.4 Physics and Manual. *JHEP*, 05:026, 2006. doi: 10.1088/1126-6708/2006/05/026.

[24] P.Z. Skands. Tuning Monte Carlo Generators: The Perugia Tunes. *Phys. Rev.*, D82:074018, 2010. doi: 10.1103/PhysRevD.82.074018.

[25] Particle Data Group Collaboration, M. Tanabashi, et al. Review of particle physics. *Phys. Rev. D*, 98:030001, Aug 2018. doi: 10.1103/PhysRevD.98.030001. URL `https://link.aps.org/doi/10.1103/PhysRevD.98.030001`.

[26] ALICE Collaboration. Preliminary Physics Summary: Measurement of D^0, D^+, D^{*+} and D_s^+ production in pp collisions at $\sqrt{s} = 5.02$ TeV with ALICE. May 2018. URL `https://cds.cern.ch/record/2317187`.

[27] M. Cacciari, M. Greco, and P. Nason. The P(T) spectrum in heavy flavor hadroproduction. *JHEP*, 05:007, 1998. doi: 10.1088/1126-6708/1998/05/007.

[28] M. Cacciari et al. Theoretical predictions for charm and bottom production at the LHC. *JHEP*, 10: 137, 2012. doi: 10.1007/JHEP10(2012)137.

[29] G. Kramer and H. Spiesberger. Study of heavy meson production in p–Pb collisions at $\sqrt{S}$= 5.02 TeV in the general-mass variable-flavour-number scheme. *Nucl. Phys.*, B925:415–430, 2017. doi: 10.1016/j.nuclphysb.2017.10.016.

[30] M. Benzke et al. Prompt neutrinos from atmospheric charm in the general-mass variable-flavor-number scheme. *JHEP*, 12:021, 2017. doi: 10.1007/JHEP12(2017)021.

[31] R. Maciuła and A. Szczurek. Production of Λ_c baryons at the LHC within the k_T-factorization approach and independent parton fragmentation picture. *Phys. Rev.*, D98(1):014016, 2018. doi: 10.1103/PhysRevD.98.014016.