



TECHNISCHE
UNIVERSITÄT
WIEN



DIPLOMARBEIT

A Multivariate Approach to Dilepton Analyses in the Upgraded ALICE Detector at CERN-LHC

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Technische Physik

eingereicht von

Sebastian Templ BSc.

Matrikelnummer 01125396

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

Betreuung

Betreuer: Prof. Dr. Eberhard Widmann

Mitwirkung: Dr. Michael Weber

Wien, 28.03.2018

(Unterschrift Verfasser)

(Unterschrift Betreuer)

CERN-THESIS-2018-034
25/04/2018



Abstract

ALICE, the dedicated heavy-ion experiment at CERN–LHC, will undergo a major upgrade in 2019/20. This work aims to assess the feasibility of conventional and multivariate analysis techniques for low-mass dielectron measurements in Pb-Pb collisions in a scenario involving the upgraded ALICE detector with a low magnetic field ($B = 0.2$ T). These electron-positron pairs are promising probes for the hot and dense medium, which is created in collisions of ultra-relativistic heavy nuclei, as they traverse the medium without significant final-state modifications. Due to their small signal-to-background ratio, high-purity dielectron samples are required. They can be provided by conventional analysis methods, which are based on sequential cuts, however at the price of low signal efficiency. This work shows that existing methods can be improved by employing multivariate approaches to reject different background sources of the dielectron invariant mass spectrum. The major background components are dielectrons from photon conversion and combinatorial pairs. By implementing deep neural networks, the signal-to-background ratio can be improved by up to 60 % over existing results in the case of pure conversion rejection and up to 30 % in the case of additional suppression of all combinatorial background components. In both cases, the gain in significance is about 15 % compared to conventional approaches. Additionally, different strategies for rejecting heavy flavor pairs (*i. e.*, dielectrons originating from $c\bar{c}$ or $b\bar{b}$) are studied and some of their major challenges identified. In general, it is concluded that multivariate techniques are a powerful and promising approach to dielectron analyses since they significantly improve the results over conventional methods in terms of signal-to-background ratio and significance. Moreover, these techniques remove complexity from existing implementations as they allow to (1) base the analyses on individual tracks (instead of track pairs), essentially without sacrificing analysis performance, (2) render some of the existing and involved analysis methods obsolete and, to some degree, (3) obviate the need for manual input feature engineering.

Zusammenfassung

ALICE, eines der vier großen Experimente am CERN-LHC, wird in den Jahren 2019/20 einem umfassenden Upgrade unterzogen werden. Die vorliegende Arbeit dient der Abschätzung der Realisierbarkeit von konventionellen und multivariaten Analysetechniken für die Messung von Di-Elektronen mit geringer Masse in Pb-Pb-Kollisionen in einem Szenario, das den verbesserten ALICE-Detektor mit einem geringen Magnetfeld ($B = 0.2\text{ T}$) involviert. In Kollisionen von ultrarelativistischen, schweren Kernen wird ein heißes und dichtes Medium erzeugt. Elektron-Positron-Paare stellen vielversprechende Sonden für dieses Medium dar, da sie es ohne signifikante Änderungen ihres finalen Zustandes durchqueren. Aufgrund ihres geringen Signal-Untergrund-Verhältnisses werden Proben mit hoher Reinheit benötigt. Diese können durch konventionelle Analysemethoden, welche auf sequentiellen Schnitten basieren, auf Kosten von geringer Signaleffizienz bereitgestellt werden. Die vorliegende Arbeit zeigt, dass existierende Ansätze durch den Einsatz multivariater Methoden zur Unterdrückung verschiedener Untergrund-Quellen im Di-Elektronen-Massenspektrum verbessert werden können. Die Hauptkomponente des Untergrund sind Di-Elektronen von Photon-Konversionen und kombinatorische Di-Elektronenpaare. Der Einsatz von “tiefen” neuronalen Netzwerken (*deep neural networks*) ermöglicht eine Verbesserung des Signal-Untergrund-Verhältnisses von bis zu 60 % im Vergleich zu existierenden Ergebnissen im Falle von multivariater Unterdrückung des Konversionsuntergrundes und bis zu 30 % im Falle von zusätzlicher Unterdrückung aller kombinatorischer Komponenten. In beiden Fällen erhöht sich die Signifikanz um etwa 15 % im Vergleich zu herkömmlichen Ansätzen. Außerdem werden verschiedene Strategien zur Unterdrückung von *heavy flavor*-Paaren (*i. e.*, von $c\bar{c}$ oder $b\bar{b}$ herrührende Di-Elektronen) untersucht und die damit verbundenen Herausforderungen identifiziert. Da multivariate Methoden im Vergleich zu konventionellen Methoden die Ergebnisse hinsichtlich des Signal-Untergrund-Verhältnisses und der Signifikanz bedeutend verbessern, wird allgemein die Schlussfolgerung gezogen, dass diese Methoden einen leistungsfähigen und vielversprechenden Ansatz für die Analyse von Di-Elektronen darstellen. Weiters verringern multivariate Methoden die Komplexität existierender Implementierungen, indem sie ermöglichen, (1) die Analysen auf Basis einzelner Teilchenspuren (*i. e.*, *tracks*) durchzuführen, anstelle von Paaren von *tracks*, (2) manche eingesetzten und aufwendigen Analysemethoden zu ersetzen, ohne die Analyseergebnisse dabei nennenswert zu verschlechtern, und (3) das manuelle Konstruieren von Inputvariablen bis zu einem gewissen Grad hinfällig zu machen.

Contents

Abstract	iii
List of Figures	ix
List of Tables	xi
Acknowledgements	xiii
1 Motivation	1
2 Theoretical background	3
2.1 Properties and states of hadronic matter	3
2.1.1 Principles of quantum chromodynamics	4
2.1.2 QCD phase transition and quark-gluon plasma	6
2.1.3 QGP probes	11
2.2 The dielectron spectrum	12
2.3 The ALICE experiment	15
2.3.1 Current detector setup	15
2.3.2 Detector upgrade for Run 3 and 4	16
2.3.3 ALICE physics program	17
2.4 Multivariate analysis and Deep Learning	18
2.4.1 General motivation for a multivariate approach	18
2.4.2 General aspects of multivariate classification	19
2.4.3 Evaluation of classifier performance	20
2.4.4 Neural networks and Deep Learning	22
3 Multivariate conversion rejection	26
3.1 Used dataset	26
3.2 Used frameworks	27
3.3 Data preparation	27
3.4 Analyses	28
3.4.1 Bivariate <i>prefilter</i> cuts for conversion pair rejection (reference analysis)	28
3.4.2 Multivariate <i>prefilter</i> cut for conversion pair rejection	33
3.4.3 Multivariate conversion rejection on the single-track level	36
3.4.4 Multivariate rejection of combinatorial conversion pairs	36
3.5 Results	37
3.5.1 General classifier performances	38
3.5.2 Classifier performances at the optimized working points	39

3.6	Discussion	44
4	Multivariate heavy flavor rejection	48
4.1	Used datasets	48
4.2	Used frameworks	49
4.3	Data preparation	49
4.4	Analyses	49
4.5	Results	52
4.5.1	Non-combinatorial vs. heavy flavor pairs	53
4.5.2	Combinatorial vs. heavy flavor pairs	54
4.5.3	Non-conversion vs. heavy flavor tracks	55
4.6	Discussion	55
5	Summary and outlook	60
	Appendices	64
A	Track cuts	66
A.1	Default track cuts	66
A.2	Loose track cuts	67
A.3	Minimal track cuts	67
B	Input variables	68
C	<i>Prefiltering</i> algorithms	70
D	Conversion rejection studies assuming perfect PID (supplement)	72
E	Results for multivariate conversion rejection (supplement)	74
F	Parameters of heavy flavor rejection classifiers	76
	Bibliography	78

List of Figures

2.1	QCD phase diagram	7
2.2	Order parameters at the QCD transition	9
2.3	Finite temperature phase structure for three quark flavors	9
2.4	Expected contributions to the dielectrom mass spectrum	13
2.5	Feynman diagram for the Drell-Yan process.	14
2.6	Feynman diagrams of heavy quark production	14
2.7	Annihilation of two pions to a pair of leptons	15
2.8	The ALICE detector (schematic)	17
2.9	Examples of ROC curves	21
2.10	Graphical representation of a neural network	23
3.1	PID efficiencies as a function of p_T	28
3.2	Dielectron spectrum for Pb-Pb collisions at $\sqrt{s_{NN}} = 5.5$ TeV	29
3.3	Nomenclature for pairs from hadron decays or photon conversions	29
3.4	Misreconstruction of conversion pair opening angle	30
3.5	Distribution of φ_V for different contributions	31
3.6	Non-combinatorial (non-)conversion pairs in the φ_V - m_{ee} plane	31
3.7	Dielectron spectrum after the bivariate <i>prefilter</i> cut	32
3.8	Mass-dependent cut efficiencies of the bivariate <i>prefilter</i> cut	33
3.9	Learning curve of the neural net rejecting real-pair conversions	34
3.10	MVA cut optimization (non-combinatorial conversion pair rejection)	35
3.11	MVA cut optimization (single-track conversion rejection)	37
3.12	MVA cut optimization (combinatorial conversion rejection)	38
3.13	ROC curves of the different conversion rejection approaches	39
3.14	ROC curves of the different conversion rejection approaches (zoomed).	40
3.15	Significance vs. signal efficiency (conversion rejection)	40
3.16	Significance vs. signal efficiency (conversion rejection, zoomed)	41
3.17	Cut efficiencies for different dielectron sources and classifiers (1/2)	42
3.18	Cut efficiencies for different dielectron sources and classifiers (2/2)	43
3.19	Mass-dependent significances (conversion rejection analyses)	44
3.20	Mass-dependent $S_{\text{class}}/B_{\text{class}}$ (conversion rejection analyses)	45
3.21	Mass-dependent significances (combinatorial and conversion pair rej.)	45
3.22	Mass-dependent $S_{\text{exp}}/B_{\text{exp}}$ (combinatorial and conversion pair rej.)	46
4.1	Dielectron invariant mass spectrum for pp collisions at $\sqrt{s} = 13$ TeV	50
4.2	Visualization of prompt vs. delayed decays w.r.t. DCA	51
4.3	ROC AUCs of the different HF-rejecting classifiers	52
4.4	MVA cut efficiencies for the “RP::HF” classifiers	54
4.5	MVA cut efficiencies for the “CombBG::HF” classifiers	54

4.6	MVA cut efficiencies for the “NonConvs::HF” classifiers	55
4.7	m_{ee} vs. pair- p_T for two dielectron sources in the GP MC production .	56
4.8	MVA cut efficiencies vs. m_{ee} for two dielectron sources and different classifiers for the GP MC production as functions of m_{ee}	56
4.9	MVA cut efficiencies vs. pair- p_T for two dielectron sources and different classifiers for the GP MC production as functions of pair- p_T . .	57
D.1	ROC curves in the case of perfect PID (conversion rejection)	73
E.1	Dielectron spectrum & MVA cut efficiencies (multivariate <i>prefilter</i> cuts)	74
E.2	Dielectron spectrum & MVA cut efficiencies (single-track conv. rej.) .	75
E.3	Dielectron spectrum & MVA cut efficiencies (combinatorial conv. rej.)	75

List of Tables

3.1	ROC AUCs of all MVA approaches (conversion rejection)	39
A.1	Default track cuts (conversion rejection)	66
A.2	Loose track cuts (conversion rejection)	67
A.3	Minimal track cuts (conversion rejection)	67
B.1	Relevant single-track variables used for the analyses	68
B.2	Relevant pair variables used for the analyses	69
D.1	ROC AUCs of all MVA approaches (conversion rejection, perfect PID)	73
F.1	Parameter specifications for the HF-rejecting neural networks	77

Acknowledgements

This work would not have been possible without the support of others, some of whom deserve special mention.

First and foremost, I would like to thank my supervisor Eberhard Widmann, who not only supported me in the course of this work, but also over and over again during my last years as a physics student. As director of the Stefan Meyer Institute (SMI), he is also jointly responsible for the particularly pleasant working atmosphere at the institute, which I have always enjoyed being a part of.

Most of all, I would like to extend my sincere thanks to my co-supervisor Michael Weber, on whose excellent support I could always rely throughout my time in his group at SMI. His guidance, his vision and his versatile expertise were of utmost importance for the success of this work. Michael provided me with numerous opportunities to deepen my knowledge, improve my skills, present my work at various occasions and get to know parts of the communities in high-energy physics and machine learning. I am thankful for his support, his patience and the trust he has put in me from day one. But most of all, I would like to thank him for having created an inclusive environment that fostered my enthusiasm and made me enjoy working and learning.

I would also like to thank my colleagues Aaron Capon and Sebastian Lehner for their part in this work. They always had an open ear for my questions and were always keen to make suggestions for improvement. Apart from that – and this applies equally to all other office colleagues – they were exceptionally pleasant contemporaries who made everyday life much more enjoyable.

Last but not least, I would like to express my deep gratitude to my family and my friends for their continued support and for the foundation they are giving me on which to build my life.

This work was supported by the Austrian Academy of Sciences and by the Nationalstiftung für Forschung, Technologie und Entwicklung.

Chapter 1

Motivation

Modern physics connects the microscopic world of fundamental particles with the macroscopic world, *i. e.*, the universe at its grandest scale throughout its evolution since the Big Bang about 13.8 billion years ago [1]. Some of the fundamental questions of physics, too, appear on both the micro- and the macroscopic level: What are the ultimate building blocks of matter? What are the fundamental forces between them? And, given those basic constituents of matter and their interactions, which states of matter are there? How do the transitions between these states happen?

As a humble step in the quest of answering these questions, this work focuses on the study of states and transitions of *hadronic* matter, *i. e.*, matter that interacts via the strong force (one of four fundamental forces in the known universe). Its experimental foundation is the ALICE detector at CERN-LHC, Switzerland, an experiment that is optimized for the study of strongly interacting matter in collisions of ultra-relativistic heavy nuclei. In order to probe the hot and dense medium that is created in heavy-ion collisions with ALICE detector, this work uses electron-positron pairs (*dielectrons*) to extract information about the collision and its different stages. Dielectrons have the great advantage that they do not interact via the strong force and, therefore, do not show significant in-medium modifications of their final states. The goal is to eventually reconstruct the dielectron invariant mass spectrum, from which information about the entire collision history as well as fundamental quantities about the states and transitions of hadronic matter can be extracted. ALICE will undergo a major upgrade in 2019/20 in order to prepare the detector for higher LHC luminosities. After this upgrade, a dedicated run is planned to access low-mass dilepton physics. This work studies dielectrons in a scenario that involves the upgraded ALICE detector.

In dielectron analyses, one usually has to deal with various sources of background that are larger than the signal by a few orders of magnitude ($S/B \approx 10^{-3}$). The rejection of these background components requires the use of sophisticated analysis techniques. In this work, various multivariate analysis (MVA) approaches are employed. Their feasibility and performances are studied with respect to conventional analyses, which are based on sequential cuts that are applied to a relatively small set of input variables. Specifically, deep neural networks are used as they are a powerful tool to learn complex and non-linear relations between *all* input variables in order to differentiate between signal and background.

One of the major background components of the dielectron invariant mass spectrum are e^+e^- pairs from photon conversions. This work will investigate different

multivariate ways to reject this component and compare them to conventional analyses. Furthermore, the performance of the trained neural networks to reject all combinatorial pairs in addition to conversion pairs will be studied. Especially in the higher-mass region, pairs from heavy-flavor decays (*i. e.*, $c\bar{c} \rightarrow Xe^+e^-$ or $b\bar{b} \rightarrow Xe^+e^-$, with c/\bar{c} ...charm/anticharm quarks and b/\bar{b} ...bottom/antibottom quarks) constitute a large contribution to the dielectron spectrum. Different multivariate ways to characterize and suppress heavy flavor will be also be explored in addition to the studies of multivariate conversion rejection. All in all, this work aims to investigate the potential of multivariate techniques for dielectron analyses in ALICE and to give an estimate of where some of the strengths and weaknesses of these methods lie.

This work is structured as follows: First, Chap. 2 introduces the theoretical aspects of the properties and states of hadronic matter (Sec. 2.1) and the dielectron invariant mass spectrum (Sec. 2.2). Subsequently, in Sec. 2.3, the ALICE experiment in its current as well as in its upgraded state will be described. A general discussion of multivariate analysis approaches with a particular focus on Deep Learning will be given in Sec. 2.4. After this introductory chapter, the two main studies of this work will be detailed, *i. e.*, multivariate conversion rejection (see Chap. 3) and multivariate heavy flavor rejection (see Chap. 4). Finally, Chap. 5 will summarize the results of this work and give an outlook for future dielectron studies.

Chapter 2

Theoretical background

2.1 Properties and states of hadronic matter

Most of the phenomena that happen in our day-to-day lives can be described by two fundamental forces of nature: gravity and electromagnetism. They have far-reaching fields of application, ranging from the atomic to the galactic scale. However, two more fundamental interactions are needed¹ in order to explain subnuclear physics: the weak and the strong interactions. While the weak interaction most prominently works at the core of nuclear processes like β decay, the strong interaction explains the dynamics and interplay of the constituents (quarks and gluons) of nuclear matter particles like protons and neutrons. One can estimate the strengths of these fundamental interactions by comparing the timescales of different particle decays. Roughly speaking, the lifetime of a quantum state is inversely proportional to the strength of the interaction responsible for its decay [2]. Comparing these lifetimes τ , one finds that the strong interaction ($\tau_{\text{strong}} \approx 10^{-23}$ s) is by far the strongest force, followed by the electromagnetic ($\tau_{\text{em}} \approx 10^{-16}$ s) and then the weak ($\tau_{\text{weak}} \approx 10^{-8} - 10^{-6}$ s) force.

For the purpose of this work, the most interesting of the above forces is the strong interaction. The theory which describes the processes and dynamics of strongly interacting particles is called *Quantum Chromodynamics* and will be briefly introduced in Sec. 2.1.1. Under normal conditions, strongly interacting particles form composite states of quarks that are held together by the strong interaction, like protons or pions². These compound particles are called *hadrons*. When in a hot and/or dense state, however, hadrons “melt” and form a new state in which the quarks and gluons are no longer *confined* into their compound hadrons. Strongly interacting matter in this state, as well as the transition from “cold” hadronic matter to *deconfined* hot and/or dense matter, will be described in Sec. 2.1.2. Finally, Sec. 2.1.3 will give an overview of different ways to probe the properties of nuclear matter under such extreme conditions.

¹One can think of simple reasons for this observation. For example, electromagnetism is unable to explain the stability of nuclei with many positively charged protons that repel each other electromagnetically; and gravity is just too weak to account for the processes happening on the (sub)atomic scale.

²The *valence quark* composition of a proton is two *up* and one *down* quark. Hadrons with three valence quarks are called *baryons*. In contrast, pions consist of a light quark and a light antiquark (e. g., $\pi^0 = u\bar{u}$ or $d\bar{d}$, $\pi^+ = u\bar{d}$, $\pi^- = d\bar{u}$). Particles in such a quark-antiquark arrangement are called *mesons*.

2.1.1 Principles of quantum chromodynamics

Quantum Chromodynamics (QCD) is the gauge field theory of the strong interaction, which describes the interactions of quarks and gluons. It does so in a way similar to Quantum Electrodynamics (QED), the gauge field theory of the electromagnetic interaction. Both theories describe spin- $\frac{1}{2}$ fermions (quarks or electrons) as matter fields that interact via the exchange of massless vector gauge bosons (gluons or photons). There is however a fundamental difference between QCD and QED. Quarks and gluons carry an additional quantum number (the *color* charge), which is associated with the non-Abelian gauge group $SU(3)$. In contrast, the electric charge is described by the Abelian $U(1)$ group [3]. Consequently, there are three different color charge states of quarks (*red*, *green* and *blue*, and their respective anticolors) and eight gluons states, which transform according to the adjoint representation of $SU(3)$. In QED, instead, there is only one charged fermion and one uncharged photon.

The Lagrangian density \mathcal{L} of QCD [3], yielding the dynamics of quarks and gluons at their fundamental level, is given in the following.³

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}^a F_{\mu\nu}^a - \bar{\psi}_f^\alpha \left[i\delta_{\alpha\beta}\gamma^\mu\partial_\mu - \frac{g}{2}(\lambda_a)_{\alpha\beta}\gamma^\mu A_\mu^a \right] \psi_f^\beta + m_f \bar{\psi}_f^\alpha \psi_{\alpha,f}, \quad (2.1.1)$$

with

$$F_{\mu\nu}^a = (\partial_\mu A_\nu^a - \partial_\nu A_\mu^a - gf_{bc}^a A_\mu^b A_\nu^c), \quad (2.1.2)$$

where A_μ^a is the gluon vector field with color a ($a = 1, 2, \dots, 8$), ψ_f^α is the quark spinor field with flavor⁴ f ($f = u, d, c, s, t, b$) and color α ($\alpha = 1, 2, 3$), g is the strong coupling constant and f_{bc}^a are the structure functions corresponding to the color gauge group. The generators of these structure functions are the Gell-Mann matrices λ_a [4] with

$$[\lambda_a, \lambda_b] = if_{ab}^c \lambda_c. \quad (2.1.3)$$

The crucial difference between QCD and QED can be readily seen in the above equations: Direct interactions between gluons, which are due to the intrinsic charge of the gluon field, manifest themselves in the last term of Eq. 2.1.2. For $f_{bc}^a = 0$, the Lagrangian density of QCD (Eq. 2.1.1) changes into the one of QED, where there are no interactions between the quanta of the corresponding gauge field (*i. e.*, the photons).

The last term in Eq. 2.1.1 is the only term in the Lagrangian responsible for quark masses. Without it, the QCD Lagrangian is symmetric under the so-called *chiral transformations*,

$$\Lambda_V : \psi \rightarrow e^{-i\alpha_{i,V} \cdot \frac{\lambda_i}{2}} \psi \quad (\text{vector transformation}), \quad (2.1.4)$$

$$\Lambda_A : \psi \rightarrow e^{-i\alpha_{i,A} \cdot \frac{\lambda_i}{2}} \gamma_5 \psi \quad (\text{axial transformation}), \quad (2.1.5)$$

where $\gamma_5 = i\gamma^0\gamma^1\gamma^2\gamma^3$ (with γ^μ being the Dirac matrices). Quarks have small yet finite masses, so the QCD Lagrangian is still symmetric under vector transformations, but no longer under axial transformations. In other words, chiral symmetry

³If not stated otherwise, indices that appear multiple times per term and are not otherwise defined are to be summed over their respective ranges (*Einstein's summation convention*).

⁴The standard model of particle physics includes quarks in six different flavors, u (“up”), d (“down”), c (“charm”), s (“strange”), t (“top” or “truth”) and b (“bottom” or “beauty”), and their corresponding antiquarks

is *explicitly* broken. However, since the masses of u and d quarks are significantly smaller than the scale parameter of QCD, $\Lambda_{\text{QCD}} \approx 200 \text{ MeV}$, chiral symmetry can be seen as approximately realized in nature. Vector transformations Λ_V correspond to continuous rotations in isospin⁵ space and thus, by virtue of Noether's theorem [6], the strong interaction conserves isospin. Axial transformations Λ_A rotate particles with opposite chirality, like the ρ and a_1 mesons, into each other. If the QCD Lagrangian were symmetric under Λ_A transformations, these two chiral partner particles would have to have the same masses (or very similar masses, in case of an approximate symmetry). However, experimentally it is shown that $m_\rho = 760 \text{ MeV}/c^2$ and $m_{a_1} = 1230 \text{ MeV}/c^2$ [7]. This large mass difference cannot be explained by the explicit chiral symmetry breaking due to the small u and d masses alone. Instead, chiral symmetry is also *spontaneously* broken, which is a consequence of the chiral symmetry of the Lagrangian not corresponding to the symmetry of its ground state. At sufficiently high temperatures, however, chiral symmetry is expected to be restored, resulting in chiral partners having the same mass⁶.

Furthermore, the non-vanishing values of the quark masses are also responsible for the phase transition of nuclear matter being a rapid *cross-over* transition instead of a first- or second-order transition (in a thermodynamical sense). This will be further discussed in Sec. 2.1.2. For now, let us briefly consider the QCD Lagrangian in the case of vanishing quark masses, *i. e.*, $m_f = 0$ for all flavors f . In this case, Eq. 2.1.1 determines all of the dynamics of quarks and gluons in terms of a single dimensionless coupling constant, g . Rather than predicting observables in terms of absolute values of physical units, the theory merely predicts ratios of physical quantities and cannot provide us with a dimensional scale for QCD physics [3]. However, once the scale is fixed, the predicted masses of all known mesons and baryons meet the experimentally found values [8, 9].

QCD not only has to describe hadron masses, but also hadronic scattering processes. A well-established framework for the mathematical treatment of particle scattering processes is perturbation theory. In order to describe the applicability of these techniques for QCD processes, a comparison to QED will be done again.

In QED, the (electromagnetic) interaction is communicated by means of virtual photons. These photons exist for very short times and are therefore allowed to have non-zero masses. They interact with charged fermions but not with themselves, since they have no intrinsic charge. The strength of the electromagnetic interaction is encoded in the coupling constant $g_e = \sqrt{4\pi\alpha}$, with $\alpha = e^2/(\hbar c \cdot 4\pi\epsilon_0)$ being the *fine-structure constant*. According to QED, α increases logarithmically with increasing momentum transfers q . For large $|q^2|/m$ (with m being the mass of the charged particle),

$$\alpha(|q^2|) \approx \left(\frac{1}{\alpha(0)} - \frac{1}{3\pi} \ln \frac{|q^2|}{m^2} \right)^{-1}. \quad (2.1.6)$$

The decrease of the electromagnetic coupling strength with growing distance

⁵Isospin is the quantum number related to the strong interaction and was first introduced by W. Heisenberg in 1932 in order to explain symmetries regarding the then newly discovered neutron [5]. The mathematical formalism for isospin is closely related to that of angular momentum in QED, although it does not have units of angular momentum and thus is not a type of spin.

⁶Again, in case of an approximate chiral symmetry, the masses cannot be restored to the exact same value, but merely become very similar. If however the temperature at which chiral symmetry restoration coincides with the deconfinement temperature, the notion of a hadron, and thus its mass, becomes meaningless. For more details on this, see Sec. 2.1.2.

(*i. e.*, smaller momentum transfer) can be attributed to the creation of virtual e^+e^- pairs out of the vacuum via quantum fluctuations. In the presence of electric charges (*e. g.*, electrons), these virtual pairs act as a dipole and effectively screen the charges. Only by probing an electron at short distances (*i. e.*, via large momentum transfers), one is able to “plunge” into the polarizing cloud of virtual e^+e^- pairs and measure an increasing effective charge of the electron.

In contrast to photons in QED, gluons are (color-)charged and therefore interact with themselves (in addition to their interaction with quarks), as discussed earlier. In other words, gluons can not produce only virtual quark-antiquark pairs but also pairs of gluons. The strong coupling constant α_s depends on the momentum transfer q according to

$$\alpha_s(|q^2|) = \frac{12\pi}{(11N_c - 2N_f) \ln(|q^2|/\Lambda_{\text{QCD}}^2)}, \quad (2.1.7)$$

where N_c is the number of color charge states, N_f the number of flavors and $\Lambda_{\text{QCD}}^2 \approx 200 \text{ MeV}$ the scale parameter of QCD. In the standard model of particle physics, $N_c = 3$ and $N_f = 6$ (thus, $11N_c - 2N_f > 0$), which results in the magnitude of the strong interaction between two quarks (or, equivalently, the strength of the color charge) to increase with growing distance between them. Therefore, virtual gluon loops, that are created from the QCD vacuum by quantum fluctuations, provide an anti-screening effect which is dominant over the color screening effect from quark-antiquark pairs.

The fact that the strong coupling approaches zero for $|q^2| \rightarrow \infty$ is known as *asymptotic freedom*. It is due to this very phenomenon that hard scattering processes at large momentum transfer ($|q^2| \gtrsim 1 \text{ GeV}/c$) can be successfully described via perturbative methods in QCD [10]. On the other hand, a direct quantitative application of QCD to hadronic scattering processes with low momentum transfer (*i. e.*, non-perturbative regime) is infeasible [3], as non-negligible long-range correlations and multi-particle interactions at the core of QCD result in higher-order terms that cannot be assumed to be small. So far, only one non-perturbative regularization scheme for solving this relativistic quantum field theory is available, *i. e.*, the lattice formulation of QCD [11]. The enabling aspect of Lattice QCD is its capability to provide thermodynamic observables that can be assessed numerically by computer simulations [12].

2.1.2 QCD phase transition and quark-gluon plasma

Quantum Chromodynamics has proved itself as a well-established theory of the strong interaction. Still, there are many fundamental aspects yet to be fully explained.⁷ In the quest of studying nuclear matter and its phase transitions many dedicated experiments have been carried out or are currently ongoing (*e. g.*, ALICE⁸ at the Large Hadron Collider (LHC) at CERN, Switzerland; STAR⁹, PHENIX¹⁰,

⁷Examples of open questions in the field are the exact nature of (a) quark confinement, (b) the parton-hadron transition, (c) hadronic matter at high temperatures, (d) chiral symmetry breaking and (e) the origin of light-quark masses [13].

⁸ALICE – “A Large Ion Collider Experiment”, <http://aliceinfo.cern.ch/Public/Welcome.html>.

⁹STAR – “Solenoidal Tracker at RHIC”, <https://www.star.bnl.gov/>.

¹⁰PHENIX – “Pioneering High Energy Nuclear Interaction Experiment”, <http://www.phenix.bnl.gov/>.

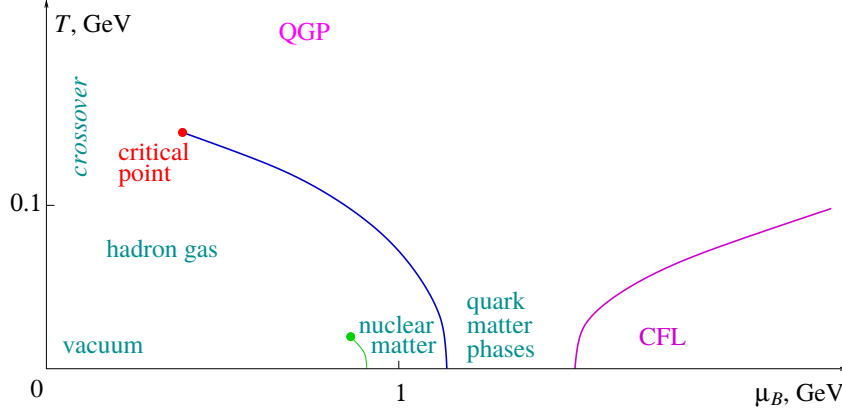


FIGURE 2.1: A semiquantitative sketch of the QCD phase diagram [17].

BRAHMS¹¹ or PHOBOS¹² at the Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory (BNL), New York) and some are planned (*e. g.*, the CBM¹³ experiment at the Facility for Antiproton and Ion Research (FAIR), Germany).

The different phases of strongly interacting matter and their boundaries are usually depicted in the so-called *QCD phase diagram*, which schematically represents thermodynamic states as functions of temperature T and net baryon density or chemical potential μ_B . An illustration of the present-knowledge QCD phase diagram is shown in Fig. 2.1. Ordinary nuclear matter exists at almost zero temperature and $\mu_B \approx 940$ MeV. For larger values of μ_B , corresponding to a pressure of about 1 MeV/fm³, hadronic matter is a degenerate gas of neutrons¹⁴; and for even larger μ_B , at pressures above 1 GeV/fm³, a degenerate gas of quarks¹⁵. Above that, strongly interacting matter forms a so-called color-flavor locked (CFL) phase¹⁶. Exploring the phase diagram at lower μ_B , but finite temperatures $T \lesssim 160$ MeV, one finds quarks and gluons being confined into hadrons that form an interacting gas state. At higher temperatures, asymptotically free quarks and gluons form a quark-gluon plasma (QGP) [14, 15]. In this state, the quark and gluon degrees of freedom are liberated; quarks and gluons are no longer *confined* inside atomic nuclei or other hadrons. Conditions suitable for the formation of a QGP were present in the early Universe [16], but these are now shielded from astronomical observations by its subsequent evolution. However, it is possible to recreate similar conditions in high-energy particle collisions, such as ultra-relativistic heavy-ion collisions.

In order to reach the QGP state, (cold) nuclear matter has to undergo a QCD phase transition. Despite its name, the QCD phase transition is expected to consist of actually two transitions: the *deconfinement* and the *chiral phase transition*,

¹¹BRAHMS – “Broad Range Hadron Magnetic Spectrometer”, <http://www4.rcf.bnl.gov/brahms/WWW/>.

¹²PHOBOS, <https://www.bnl.gov/phobos/>.

¹³CBM – “Compressed Baryonic Matter”, <http://www.fair-center.eu/for-users/experiments/cbm-and-hades/cbm.html>.

¹⁴This phase is expected to be present in the core of neutron stars.

¹⁵This phase is considered to be a color superconductor, *i. e.*, a degenerate Fermi gas of quarks with a condensate of Cooper pairs near the Fermi surface.

¹⁶This phase is superfluid, electromagnetically insulating and chiral symmetry breaking.

which, arguing from first principles, need not happen at the same temperature or density. The former is caused by breaking of the Z_3 symmetry [18], which is an exact symmetry in the limit of pure-gauge QCD (*i. e.*, $m_f \rightarrow \infty$ in Eq. 2.1.1). For temperatures T below the critical transition temperature $T_C^{(\text{deconf})}$, the Z_3 symmetry is present and quarks and gluons are confined inside color-neutral objects. Above $T_C^{(\text{deconf})}$, the matter descriptions via hadronic degrees of freedom ceases to apply; quarks and gluons become deconfined. The corresponding order parameter is the expectation value of the Polyakov loop $L(T)$ [19–21],

$$L(T) \approx \lim_{r \rightarrow \infty} \exp[-V(r)/T], \quad (2.1.8)$$

where $V(r)$ is the potential¹⁷ between a quark and an antiquark with distance r . In contrast, the chiral phase transition is connected with the generation of an effective quark mass which is due to the existence of a quark-antiquark vacuum condensate at low temperatures $T < T_C^{(\text{chiral})}$ [13]. The order parameter for this transition is the expectation value for the density of the quark-antiquark condensate in vacuum, $\langle \bar{\psi}\psi \rangle(T)$. It is possible that the system is in a state with $\langle \bar{\psi}\psi \rangle \neq 0$, even if $m_f = 0$ for all flavors f in Eq. 2.1.1. If so, an effective, non-vanishing quark mass is spontaneously generated by virtue of the quarks being “dressed” by mainly massless gluons. When the temperature rises above $T > T_C^{(\text{chiral})}$, the vacuum condensate $\langle \bar{\psi}\psi \rangle$ decreases and the quark masses drop to their bare values while the chiral symmetry is restored during the transition (*c. f.*, Fig. 2.2).

As indicated above, the respective phase transitions are real first-order transitions only in the limiting cases of vanishing or infinite quark masses. This is illustrated in Fig. 2.3 for the case of three quark flavors of masses m_u , m_d and m_s . In the case of pure-gauge QCD with no quarks present (upper right corner of the diagram in Fig. 2.3), the Polyakov loop is a genuine order parameter, *i. e.*, it has a discontinuity at the temperature of the deconfinement phase transition. This discontinuity remains for a little if quarks with large, but finite masses are introduced. It however disappears when the quark masses fall below a certain value m_q^c (see Fig. 2.3). This finite region of discontinuous first-order behavior is bounded by a line of second-order transitions. For still smaller quark masses, the Polyakov loop does not show a singular behavior at the phase transition anymore. However, it keeps changing very fast over a small temperature interval at the transition temperature, such that in practice the transition from one state into another is quite clearly recognizable (*c. f.*, Fig. 2.2). This is commonly referred to as a *rapid cross-over*. In the case of zero or small quark masses (lower left corner in Fig. 2.3), the situation is similar. There is a first-order chiral phase transition in some domain of small m_q . The discontinuity in the order parameter $\langle \bar{\psi}\psi \rangle$, after crossing the confining line of second-order transition¹⁸, is again replaced by a rapid cross-over behavior at the transition temperature. It is also confirmed by Lattice QCD calculations that the

¹⁷In pure-gauge QCD, $V(r) \approx \sigma r$, with σ being the string tension, and hence $V(r \rightarrow \infty) \rightarrow \infty$, resulting in $L = 0$. In a deconfined medium, due to color screening among the gluons leading to melting of the string, $V(r)$ is finite at large r and thus L does not vanish [22]. Therefore, the onset of deconfinement happens at the temperature T_C at which L becomes finite.

¹⁸The line of second-order phase transition continues from the point m_s^{tri} towards $m_u = m_d = 0$, $m_s = \infty$. We do not want to further elaborate on this observation since we are primarily interested in the location of the “physical point” in Fig. 2.3. For a brief discussion of the two-flavor second-order point see, *e. g.*, [3, chapter 6.4].

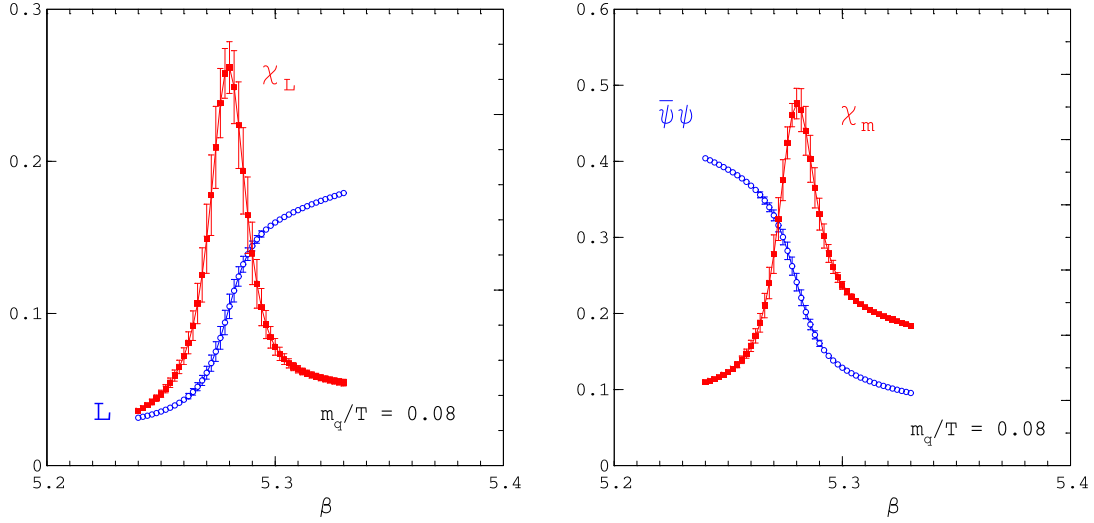


FIGURE 2.2: Order parameters for deconfinement and chiral symmetry restoration in 2-flavor QCD. Left: Expectation value of the Polyakov loop, $\langle L \rangle$, which is the order parameter for deconfinement in pure-gauge QCD ($m_q \rightarrow \infty$). Right: Expectation value of the $q\bar{q}$ vacuum condensate density, $\langle \bar{\psi}\psi \rangle$, which is the order parameter for chiral symmetry breaking in the chiral limit ($m_q \rightarrow 0$). Additionally, the corresponding susceptibilities χ are shown as functions of the coupling $\beta = 6/g^2$ [23].

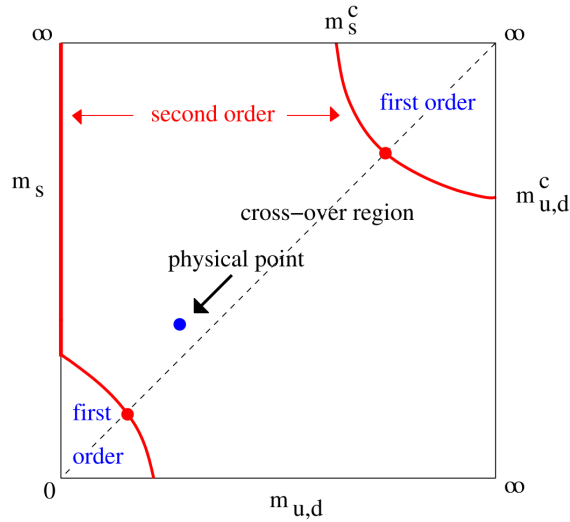


FIGURE 2.3: Finite temperature phase structure for three quark flavors u , d and s with respective masses m_u , m_d and m_s . The domains of first-order transitions are separated from the cross-over region by a line of second-order phase transitions at critical quark masses m_q^c (with $q = u, d, s$). Figure adapted from [3].

phase transition (at zero net baryon density) is not a sharp phase transition, but a smooth cross-over of the two underlying transitions. The situation actually realized in nature is schematically marked by the “physical point” in Fig. 2.3. It is fairly certain today that the point for two light u and d quarks and a heavier s quark lies in the cross-over region [3].

Up to this point, we have not discussed at which temperatures the deconfinement and the chiral phase transitions occur. As mentioned before, they need not coincide. In order to pin down their respective transition temperatures, one first has to find basic indications of the two phenomena happening. In the case of deconfinement, such an indicator is the abrupt change of the number of degrees of freedom when going from bound to unbound color states (*i. e.*, from color-neutral hadrons to colored quarks). Indicative of chiral symmetry restoration (*i. e.*, the transition from a state of massive to one of massless quarks), is a sudden drop of the finite effective quark mass to zero. Recent Lattice results for the temperature behavior of the entropy density¹⁹ $s(T, m_q)$ for the physical case of two light and one heavy quark flavors suggest that the two cross-over phase transitions indeed occur at the same temperature – the chiral critical temperature $T_\chi = 160 \pm 10$ MeV, obtained from the chiral condensate $\langle \bar{\psi}\psi \rangle$ in the limit $m_{u,d} \rightarrow 0$ [24–27]. This observation is compatible with the picture that the constituent quark mass is a polarization cloud excited by the quark in the surrounding medium [3]. This cloud melts away with increasing temperature T . At $T = T_C$ it has dissolved entirely, resulting in pointlike quarks and gluons. The consequence is a single transition²⁰ with color confinement and chiral symmetry restoration happening at the same time.

For many years, the theoretical expectations were that the QGP is a weakly-interacting gas of quarks and gluons at high temperature, with the constituents traveling long distances²¹ between interactions. This picture has been challenged by the results of recent heavy-ion experiments which suggest a behavior of the hot and dense nuclear matter that is fundamentally different²² from the expected one up to all yet-probed energies: The created matter acts as a plasma of strongly-coupled particles, which have very short mean free paths, and shows a high degree of collectivity and flows. This has led to the QGP to be commonly seen as an almost-perfect inviscid liquid [13]. Particularly, this image is based on measurements of inclusive transverse momentum spectra (*radial flow*) and azimuthal anisotropy of particle production (*elliptic flow*) for identified hadrons at soft transverse momenta (up to about 3 GeV/ c). It is further supported by a comparison to model calculations based on viscous relativistic hydrodynamics.

¹⁹The entropy density is a measure for the degree of deconfinement [3].

²⁰This is however not necessarily valid for situations with low T and large baryochemical potential μ_B since the melting of the effective quark mass is a consequence of the hot gluonic environment. In such circumstances, an intermediate plasma state of massive quarks, which separates hadronic matter from QGP, seems quite imaginable [28].

²¹relative to the size of a proton

²²It should be noted that hot and dense matter might still behave as expected at much more extreme conditions. But in the energy regimes that are accessible with current technology, hadronic matter behaves very differently from expectations and the above statement is true.

2.1.3 QGP probes

Strongly interacting matter at sufficiently high temperatures and/or densities forms a new state of deconfined quarks and gluons, as discussed in the previous section. For the sake of illustration, this section assumes that this new QGP-like state has been created (*e. g.*, in a small volume, following the collision of two heavy nuclei) and briefly introduces the different ways in which its properties can be probed and studied as a function of temperature and density. For a more in-depth discussion of these methods, please refer to standard literature treatments of this topic, *e. g.*, [3, 22, 29].

Hadron radiation

The newly created state of matter has a temperature that is assumed to be much higher than the one of the environment [22]. Therefore, the produced medium radiates. One of the major radiation components is *hadron radiation*, *i. e.*, the emission of hadrons which are composed of light quarks (up, down and strange quarks). These have to be created at the transition surface (at a temperature $T = T_C \approx 160$ MeV) between the deconfined medium inside the “fireball” and the physical vacuum. Since hadronization always takes place at the same temperature T_C , the physics of the transition surface does not depend on the processes happening in the enclosed volume of deconfined matter. Therefore, by studying radiated hadrons, one can only learn about the physics of hadronization, but not about the hot and dense medium. This also means that the relative hadron abundances correspond to those of an ideal resonance gas at the freeze-out temperature for a given baryochemical potential [30–34].

However, the situation described above is a little too simple and there are still ways to learn something about the pre-hadronic collision stages from hadron radiation, *i. e.*, *radial* and *elliptic* flow. Radial flow occurs when the created medium can expand freely and thus gives rise to a global hydrodynamic flow [35–37] that boosts the produced hadrons depending on the initial energy density. Elliptic flow happens if the initial conditions are spherically asymmetric²³. In this case, pressure gradients will lead to a directed flow that, again, depends on the initial conditions following the collision [22].

Quarkonia dissociation

Another probe to study the QGP is the dissociation of *quarkonia*, *i. e.*, bound states of heavy quarks and their antiquarks²⁴. In heavy-ion collisions, quarkonia are created in the very early stages of the collision, even before the hot QGP medium is formed. Quarkonia are expected to survive in a QGP through some temperatures $T > T_C$ [38] due to their binding energies being much larger than the typical hadronic scale ($\Lambda \approx 0.2$ GeV) [22]. As a result, these states have much smaller radii²⁵ which are specific²⁶ to their excitation state. Since the deconfinement transition is related to

²³Spherical symmetry is clearly broken, *e. g.*, in case of peripheral heavy-ion collisions.

²⁴Examples for such states consisting of charm and bottom (anti)quarks are the J/Ψ and the Υ .

²⁵Typical hadrons made of light quarks (u,d,s) have radii of about $1 \text{ fm} \approx 1/(200 \text{ MeV})$ (hadronic scale), while quarkonia have ones of about $0.1\text{--}0.2 \text{ fm}$ [22].

²⁶Larger radii correspond to more highly excited quarkonia states, as these are less tightly bound.

the offset of color screening, the amount of quarkonia dissociation is determined by the relation of binding to screening radius [22]. Different quarkonia states have different “melting temperatures” in a QGP. Thus, spectral analysis of in-medium quarkonia dissociation provides a QGP thermometer [39].

Like quarkonia, *jets* (see the following paragraph) are also formed in the early collision stages; together with quarkonia, they constitute so-called *hard probes*. Since the production of hard probes requires high energies and momenta, perturbative QCD techniques can be applied and tested in pp and pA collisions [22].

Energy loss of hard jets

Another way to probe the QGP is to study the energy loss of energetic partons (quarks or gluons) when they traverse the hot medium. Like in the case of quarkonia, these *hard probes* are created together with the medium in the collision of heavy nuclei. The amount of energy loss of such partons gives information about the density of the created medium [40–42]. This density rapidly increases during the deconfinement transition, and so does the energy loss of a passing color charge. Therefore, *hard jets* (*i. e.*, partons showering – or hadronizing – in collimated hadron jets) can be used to study the QGP in the hot as well as in the transitioning state.

Electromagnetic radiation

Another type of radiation from the hot and dense medium is *electromagnetic radiation*, which consists of photons and dileptons (e^+e^- and $\mu^+\mu^-$ pairs) [43, 44]. These particles are produced by interactions of quarks and/or gluons or by quark-antiquark annihilation. Since they do not interact strongly, they can carry information about the spacetime point of their creation out of the hot QGP without significant in-medium modifications. Photons and leptons are produced during all stages of the fireball evolution, so they provide a promising probe of the hot QGP [22]. However, the main difficulty in dilepton analyses is that they are created not only in the hot medium, but also elsewhere. The major background components to the dilepton spectrum in detectors like ALICE²⁷ are dileptons from hadronic decays and ones from photon conversions in the detector material [13]. The following section will further detail the analysis of one specific class of electromagnetic QGP probes that is important for the analyses carried out in this work, *i. e.*, low-mass dielectrons.

2.2 The dielectron spectrum

As described in Sec. 2.1.3, dileptons (*i. e.*, e^+e^- and $\mu^+\mu^-$ pairs) provide a unique probe to the properties of the hot and dense matter that is created in the collision of heavy nuclei at sufficiently high energies, since they do not interact via the strong force and thus do not experience significant in-medium modifications. Dileptons are emitted throughout the collision history and are able to carry information about the conditions at their point and time of creation out of the medium.

For the purpose of this work, the rest of this chapter will focus on dielectrons (*i. e.*, e^+e^- pairs) and will describe a generic approach to a (low-mass) dielectron

²⁷The ALICE experiment will be described in Sec. 2.3.

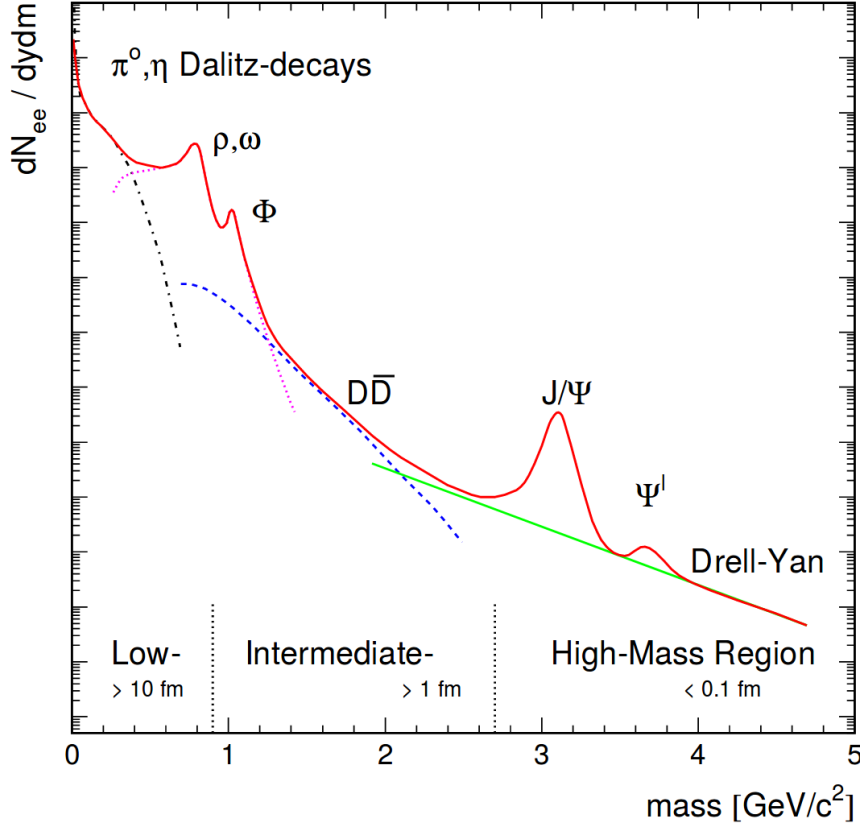


FIGURE 2.4: *Expected contributions to the dielectron spectrum as a function of invariant mass in ultra-relativistic heavy-ion collisions [45].*

analysis in terms of their invariant mass spectrum. Such a spectrum is schematically shown in Fig. 2.4 and the following discussion is with reference to it.

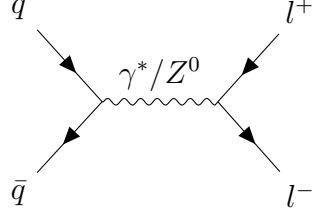
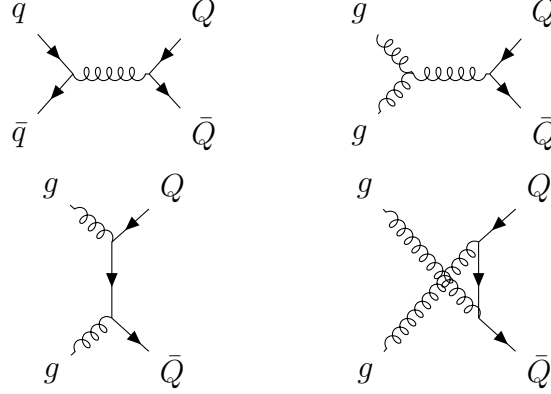
Generally speaking, there is an approximate time ordering in the invariant mass of electron pairs, which means that pairs with larger masses are generally produced earlier in the collision.

In the earliest collision stages, dielectrons are predominantly produced via *Drell-Yan* annihilation [46]. In this process, a quark of one nucleus annihilates with a sea antiquark of the other nucleus, thus creating a virtual photon (or a Z^0 boson) that decays into a lepton pair. The Feynman diagram²⁸ for the Drell-Yan process is shown in Fig. 2.5. This production mechanism is responsible for a major contribution to the invariant mass spectrum in the high-mass regime, where $m_{ll} > 3 \text{ GeV}/c^2$, and scales with the mass number A of the nucleus according to $A^{4/3}$, with respect to the nucleon-nucleon case [48]. In this mass region, a large dilepton contribution is also expected from $q\bar{q}$ annihilation which, however, follows a thermal distribution in lower-mass regions.

For $1 < m_{ll} < 3 \text{ GeV}/c^2$, *i. e.*, the intermediate mass region, a dominating dilepton contribution stems from semi-leptonic decays of *open charm* or *bottom* mesons²⁹. Quarks and antiquarks can produce $Q\bar{Q}$ states (where $Q\bar{Q} = c\bar{c}$ or $b\bar{b}$) via processes

²⁸The Feynman diagrams in this work were created using the **TikZ-Feynman** LaTeX package [47].

²⁹Mesons consisting of one charm (bottom) and one light quark are commonly referred to as *open charm* (*bottom*).


 FIGURE 2.5: *Feynman diagram for the Drell-Yan process.*

 FIGURE 2.6: *Feynman diagrams of leading-order processes of heavy quark production.*

like $q\bar{q} \rightarrow g^* \rightarrow Q\bar{Q}$ and $gg \rightarrow g^* \rightarrow Q\bar{Q}$ (see Fig. 2.6). By means of fragmentation, such a produced heavy quark pair $Q\bar{Q}$ can generate, *e. g.*, a D^+D^- meson pair³⁰, which inherits most of the initial back-to-back correlation of the $Q\bar{Q}$ due to the large mass of the heavy quark. The charmed (or bottom) mesons can then weakly decay (*e. g.*, $D^+ \rightarrow \bar{K}^0 l^+ \nu_l$). A dilepton pair is created if both mesons decay semi-leptonically, leading to a continuum of lepton pairs that constitutes the largest contribution to the intermediate mass region.

The latest collision stages, when the medium has expanded and its temperature has fallen below the critical value, can be probed towards lower masses, $m_{ll} < 1 \text{ GeV}/c^2$. In this regime, the main production of dileptons comes from annihilation of pions and kaons and scattering between other hadrons. Particularly, the light vector mesons ρ^0 , ω and ϕ can decay directly into pairs of leptons, which makes this channel especially relevant for dilepton analyses at low masses (see Fig. 2.7). Due to the direct coupling, the invariant mass of the produced lepton pair is directly connected to the mass of the vector meson at the time of its decay. Compared to the lifetime of the hadron gas ($\tau \approx 10 \text{ fm}/c$), the ρ^0 has a very short lifetime ($\tau \approx 1.3 \text{ fm}/c$) [49]. Therefore, the majority of the ρ^0 mesons will decay in the hot and dense medium and imprint information about any medium effect onto the created lepton pair. Such medium effects are expected to result in the shifting of the ρ^0 and the a_1 meson pole masses towards the same value [50, 51] and/or the broadening (“melting”) of the ρ^0 mass distribution [52], which are both scenarios that are commonly considered to be indicative of chiral symmetry restoration.

However, it is not only the *light* vector mesons that are able to provide informa-

³⁰ $D^+ = c\bar{d}$, $D^- = \bar{c}d$

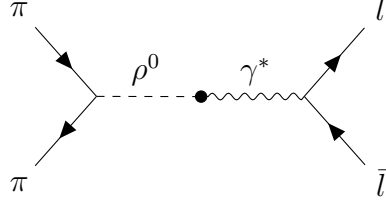


FIGURE 2.7: *Feynman diagram of two-body annihilation of pions into the light vector meson ρ^0 , which eventually decays into a lepton pair.*

tion about in-medium modifications, but also heavier ones, like the J/Ψ or the Υ , which have a longer lifetime and therefore mostly decay after the thermal freeze-out. These mesons, *i. a.*, hold information about heavy quark modifications³¹.

In the very late collision stages, after the thermal freeze-out has taken place, resonance and Dalitz decays of light mesons (*e. g.*, π^0 , η , η') are the dominant sources of lepton pairs.

2.3 The ALICE experiment

ALICE (*A Large Ion Collider Experiment*) [53, 54] is one of the major experiments at the CERN Large Hadron Collider (LHC) located near Geneva, Switzerland. Currently³², the ALICE collaboration encompasses more than 1800 researchers from 176 institutes in 41 countries.

In the center of the ALICE detector, counter-circulating beams of ultra-relativistic particles are brought together close enough that some of the particles collide. Their resulting decay products are measured by 17 subdetector systems and the data is subsequently stored for later analysis. Although ALICE has been optimized to study QCD matter created at the collision point of high-energy lead nuclei, it also takes data from pp and pA collisions. Not only are these measurements interesting in themselves, but they also serve the purpose of being a reference for Pb–Pb collisions. A schematic view of the detector is shown in Fig. 2.8.

In the following, an overview of the ALICE detector in its current state, as well as an outlook on the future detector upgrade, will be given, focusing on the two subsystems that are most relevant to this work, *i. e.*, the Inner Tracking System and the Time Projection Chamber. Afterwards, the ALICE scientific program for the next running period after the long LHC shutdown in 2019/20 will be briefly outlined.

2.3.1 Current detector setup

The ALICE detector measures $16 \times 16 \times 26 \text{ m}^3$ and weighs about 10 000 tons. Its 17 subdetectors are categorized into the *central-barrel detectors*, which are dedicated to the study of, *e. g.*, hadronic signals and dielectrons, and the *forward MUON spectrometer*, which is devoted to the study of, *e. g.*, quarkonia behavior in hot and dense

³¹However, the crucial information is encoded in their signal magnitude rather than the spectral shape.

³²as of October 2017

matter. The central-barrel detectors cover $\pm 45^\circ$ of the polar angle θ (corresponding to a pseudorapidity $|\eta| \equiv |-\ln[\tan(\theta/2)]| < 0.9$) and the full azimuthal angle φ ; they are housed in a large solenoid magnet which typically operates with a magnetic field of $B = 0.5$ T. Among them are the *Inner Tracking System* (ITS) [55], *Time Projection Chamber* (TPC) [56], *Transition Radiation Detector* (TRD) [57], *Time-of-Flight* detector (TOF) [58], *Photon Spectrometer* (PHOS) [59], *Electromagnetic Calorimeter* (EMCal) [60] and *High Momentum Particle Identification Detector* (HMPID) [61]. The forward detectors are mainly divided into V0 [62], T0 [62] and ZDC [63].

A major challenge for ALICE is the identification of particle tracks due to the extremely high track density in heavy-ion collisions. Three-dimensional hit information with many points on each track has to be used. The main tracking detectors are the ITS and the TPC, which are cylindrically aligned around the collision point in the LHC beam pipe. Despite its drawbacks in terms of read-out frequency and data volume, the latter was chosen as a main tracking device due to the need for a large number of points on each track. The TPC can reliably cope with up to 8000 charged particles per unit of rapidity at midrapidity over such a large volume³³. Its dimensions are determined by the maximally acceptable hit density (leading to a minimum possible radius of $r_{\text{in}} \approx 90$ cm) and the minimum length required for the specific energy loss resolution of a track of better than 10% (leading to an outer radius of $r_{\text{out}} \approx 250$ cm). For larger hit densities (*i. e.*, smaller radii), the tracking is handled by the ITS, which fills up the volume between the LHC beam pipe ($r_{\text{Pipe}} = 3$ cm) and the TPC. In its present state, the ITS consists of six layers: two layers of Silicon Pixel Detectors (SPD), two of Silicon Drift Detectors (SDD) and two of Silicon Strip Detectors (SSD). These detectors are capable of providing the required high granularity and spatial precision. In addition to tracking, ITS and TPC provide charged particle identification (PID) by also measuring the specific energy loss of each track.

2.3.2 Detector upgrade for Run 3 and 4

ALICE will undergo a major upgrade during the long LHC shutdown in 2019/20 (*LS2*). The main cornerstones for this upgrade are improved tracking efficiencies for tracks with low transverse momentum p_T and an increased readout rate (up to 50 kHz for Pb–Pb collisions) in order to address the physics goals for the LHC Run 3 and 4. The upgrade requires a new ITS and significant improvements of the TPC as well as an adaption of all other subdetectors to the higher readout rate. The online and offline systems, too, have to be upgraded to enable successful operation of the detector at increased data taking rates. The rest of this section will focus on the upgrade of the main tracking devices in ALICE, *i. e.*, the ITS and the TPC.

The ITS in its current state has a maximum readout rate³⁴ of about 1 kHz (with 100% deadtime), irrespective of detector occupancy [55]. It is inaccessible for maintenance work and repairs during the yearly LHC shutdowns, which can be problematic and potentially lead to lower-quality data during LHC run periods. These restrictions, together with the need of an improved low- p_T tracking efficiency, result in an alternative configuration of the upgraded ITS with more layers at dif-

³³The TPC has a drift volume of about 90 m³.

³⁴The given values correspond to the capacities of the SDD.

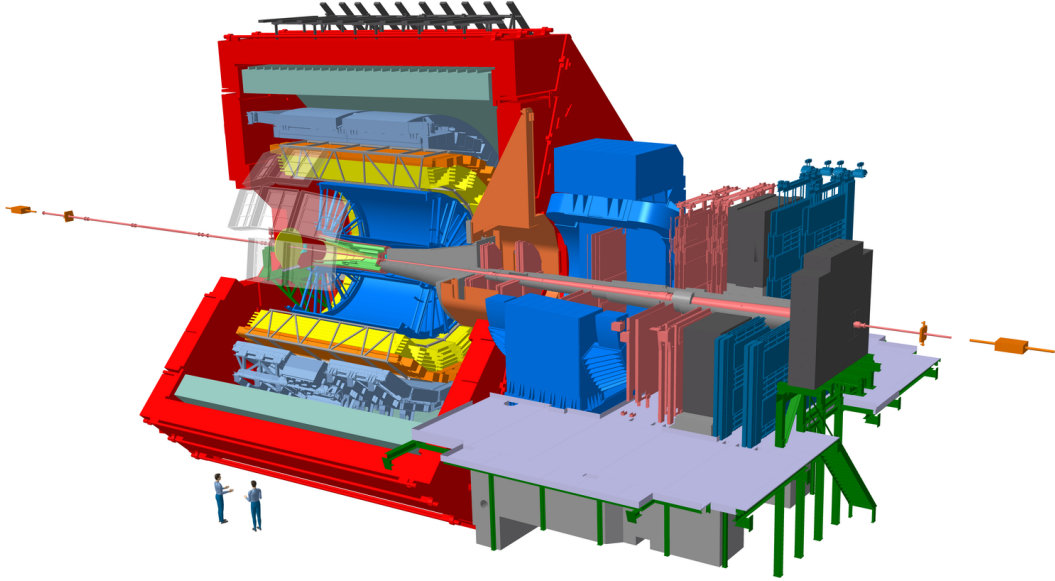


FIGURE 2.8: *A schematic drawing of the ALICE detector for LHC Run 3/4.*

ferent radii. More specifically, the upgrade entails the first ITS layer to be closer at the beamline (from the present radial distance of 39 mm to 22 mm). Further, a considerable reduction of material budget will significantly improve the tracking performance and momentum resolution. A radiation length of at most $0.3\%X_0$ per layer should be feasible. The new ITS will include seven concentric and cylindrical layers of silicon pixel detectors (each having an intrinsic resolution of $4\text{ }\mu\text{m}$), covering radii between 22 mm and 430 mm with respect to the beamline. In terms of readout time, the new ITS is required to achieve 50 kHz for Pb–Pb and several hundred kHz for pp collisions. After the upgrade, it is further expected to improve the track position resolution at the primary vertex by a factor of three or more [64].

The TPC consists of a cylindrical drift field cage volume that is filled with a Ne-CO₂ (with a ratio of 90:10, for Run 1) or Ar-CO₂ (with a ratio of 88:12, for Run 2) gas mixture and has two readout planes at the end caps [56]. At the present time, the detector is read out via Multiwire Proportional Chambers (MWPC) at a rate of 3.5 kHz. This rate is limited mainly by the closing frequency of the gating grid. Currently, it is not possible to operate the TPC in an ungated mode since this would lead to charge pile-ups that distort the electron drift paths. The planned upgrade involves a replacement of the MWPC with GEM (*Gas Electron Multiplier*) detectors which feature exceptional rate capabilities and intrinsic ion blocking without the need for additional gating. Together with the intended replacement of the front-end electronics, this will make a continuous readout of the TPC possible for the LHC Run 3 and 4 [65].

2.3.3 ALICE physics program

The ALICE experiment is designed to study hadronic matter at extremely high temperatures and/or densities. This includes the experimental demonstration of the QCD phase transition, *i. e.*, the deconfinement and chiral phase transitions,

as well as the verification of Lattice QCD predictions regarding the fundamental symmetries of the theory. Ultimately, ALICE aims at precisely determining the properties of the QGP state, like critical temperature, degrees of freedom, speed of sound, transport coefficients and the equation of state [13].

In the quest of studying matter at extreme conditions, ALICE focuses, *i. a.*, on the study of rare probes, their coupling with the hot medium and hadronization [13]. In order to do so, all the different ways to probe the QGP, which were described in Sec. 2.1.3, are utilized. Regarding the production of heavy-flavors, ALICE mainly concentrates on massive charm and bottom quarks, not only to use them as hard probes, but also to measure their azimuthal-flow anisotropy, since this is strongly related to the partonic equation of state. Regarding quarkonia, the dissociation of low-momentum quarkonia is studied, with the expectation of obtaining information about the color deconfinement and the medium temperature. As another probe, the in-medium parton energy loss is characterized to assess the multi-particle aspects of QCD and probe the density of the deconfined medium. This also determines the dependency of the in-medium energy loss on the color charge, mass, energy and medium density. Further, ALICE investigates heavy nuclear states and bound states with multi-strange baryons as well as perform systematic studies of the production of light nuclei and anti-nuclei. Finally, ALICE measures low-mass dileptons and thermal photons in order to estimate the medium's initial temperature and its equation of state, as well as studying the chiral nature of the QCD phase transition. However, with the current detector setup, precise measurements of these observables are not possible to a satisfactory degree. Together with an increased luminosity of the LHC, the upgraded ALICE detector will be able to fulfill the experimental requirements for such measurements [13]. Dileptons – whose analysis is a major subject of this work – will be among the main observables in the LHC Run 3/4 period.

2.4 Multivariate analysis and Deep Learning

This work is mainly concerned with the study of dielectron spectra (see Sec. 2.2) using data from the ALICE detector (see Sec. 2.3). Such analyses require the subtraction of background contributions that are larger than the actual signal by a few orders of magnitude. For this purpose, sophisticated analysis techniques are of crucial importance. In our analyses, we want to employ an approach that can be considered quite novel in the field of particle physics, *i. e.*, a multivariate one in the form of Deep Learning. Our motivation for doing so will be outlined in Sec. 2.4.1. Subsequently, some generic aspects (see Sec. 2.4.2) as well as specific techniques (see Sec. 2.4.3) of multivariate classification will be discussed. Finally, a short introduction to neural networks and Deep Learning will be given in Sec. 2.4.4.

2.4.1 General motivation for a multivariate approach

Conventional analysis techniques for classification of data into different categories (*e. g.*, “signal” and “background”) usually apply specific requirements in a sequential manner to observables³⁵ of an event. Most commonly, these requirements are

³⁵In this work, we use the terms “observables”, “variables” and “features” synonymously in order to refer to the input variables of a given dataset.

rectangular cut selections on individual observables. In contrast, multivariate analysis (MVA) techniques use the statistical distribution of the events in the entire high-dimensional observable space to classify events.

In general, multivariate approaches can be regarded as mappings from the m -dimensional observable space to a real number $\mathbb{R}^m \rightarrow \mathbb{R} \mid y = y(\mathbf{x} = \{x_1, x_2, \dots, x_m\})$. Hypersurfaces in the original observable space are represented by constant values $y(x) = c$ of this MVA variable and can potentially be very complex. Classification of events is usually done by assigning all events with $y(x) > c$ to the signal class while rejecting all others as background events. Since the hypersurface $y(x) = c$ separates signal from background events in the original observable space, it is commonly referred to as the *decision boundary*. Optimizing the decision boundary (*i. e.*, finding the best value for c) is a problem-specific task which heavily depends on the objective of the analysis.

The motivation for employing MVA techniques instead of classical cuts is the better performance of the former in terms of higher efficiency at the same misclassification rate [66]. By repeatedly performing cuts only in lower-dimensional subspaces of the observable space, one cannot fully exploit possible dependencies between the input variables. Further, a signal event might look like a background event in only one variable, which might lead to misclassification using the traditional cut-based analysis approach. Multivariate techniques, however, might still be able to correctly classify this event since they can compensate for this background-like variable by simultaneously considering all other variables which potentially look very signal-like [66].

There are, however, several potential drawbacks when using MVA methods in high-energy particle physics. A major issue is the realization that the training data (usually in form of dedicated Monte Carlo simulations) does not necessarily represent the data to which the trained algorithms are eventually applied (*i. e.*, the *real data* recorded in the detectors during beamtime) up to the required degree of accuracy. While solutions for some use cases exist (*e. g.*, *domain-adversarial* training of neural networks [67] or *reweighting* in the case of a covariate shift between simulated and real data [68]), this remains an open issue that requires further investigation. Another argument often used against the utilization of Machine Learning techniques is the “black-box” nature of algorithms like neural networks, *i. e.*, the claim that the criteria on which the trained model bases its predictions are fully inaccessible to the analyst due to the high degree of internal model complexity. While currently this is true (especially when the model has learned non-linear correlations between input variables), there exist several approaches to “open” the black box of such algorithms (see, *e. g.*, [69] or [70] for fairly recent ones) and research in the field of *interpretable Machine Learning* is actively ongoing.

2.4.2 General aspects of multivariate classification

The basic working principle of a multivariate classification task is to separate observations of different categories from each other by finding an appropriate decision boundary in the space of input variables. In the case of two categories (*e. g.*, “signal” and “background”), the problem is a *binary classification* problem. Otherwise, it is called a *multilabel classification*.

For a given input feature matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, which consists of n samples having

m features each, the goal is to find a function $f : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^n \mid \mathbf{X} \mapsto \mathbf{y} = f(\mathbf{X})$ which, for all $i \in \{1, 2, \dots, n\}$, best maps the input variables $\mathbf{x}_i = \{x_1, x_2, \dots, x_m\}_i$ to the respective output variable y_i . In order to find this function, labeled data is used to build a predictive model that aims to minimize the “distance” between predicted and true class labels by defining a loss function. In *supervised learning*, this task is called *training* of the model. If trained properly, the constructed model is then able to predict the classes for data without known labels.

There are several pitfalls when training a model on the available data. One of them is called *overfitting* and occurs when the model tries to extract useful information not only from the underlying distribution of the data, but also from noise which the training data (a subsample of the entire data) inevitably contains [71]. Since these statistical fluctuations are specific to the training sample, the predictive performance of the model decreases on data which it has not seen during training.³⁶ Generally speaking, while more complex and flexible models are more prone to overfitting, simpler ones might fail to fully exploit the relationships between the input features in order to find the optimal decision boundary. This issue is known as the *bias-variance trade-off* [66]. In order to be able to estimate the amount of overfitting, it is essential to separate the available labeled data into a *training*, a *validation* and a *test* sample beforehand. The model is then built using the training data, while overfitting is controlled by minimizing the performance difference between the training and the validation sample. Since the training error is not a good measure of classifier performance, the final evaluation of the model performance has to be done on yet another sample, *i. e.*, the test sample, from which information has neither directly nor indirectly been able to leak into the model building process. Choosing the “right” fractions of training, validation and test data is not a straight-forward task. There is the conflict of simultaneously wanting to (1) learn the model from the available data as accurately as possible and (2) estimating its predictive power as accurately as possible. Both objectives would improve with more data in the respective samples, so one has to find a compromise.³⁷

2.4.3 Evaluation of classifier performance

Usually, the classification performance of a model on a given dataset is assessed by comparing its predicted class labels with the true ones. To this end, one usually computes the so-called *confusion matrix*, *i. e.*, a table containing the number of correctly predicted signal events (*true positives*, TP), correctly predicted background events (*true negatives*, TN), misclassified signal events (*false positives*, FP) and misclassified background events (*false negatives*, FN). From the confusion matrix, all common metrics can be calculated, *e. g.*, accuracy = $(\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$, precision = $\text{TP} / (\text{TP} + \text{FP})$, recall = $\text{TP} / (\text{TP} + \text{FN})$, $F_1 = 2 \cdot \text{precision} \cdot \text{recall} / (\text{precision} + \text{recall})$, ...

Simple metrics like accuracy alone do not constitute good performance measures.³⁸ Therefore, it is necessary to report multiple metrics at once or choose ones

³⁶In other words, overfitting does not introduce systematical errors to the model predictions, but still should be avoided at all cost in order to maximize the generalization power of the trained algorithm.

³⁷If there is little labeled data available, a training-validation-test split might be too wasteful and one prefers to use resampling techniques like k -fold cross-validation or bootstrapping [71].

³⁸This is particularly apparent in case of *imbalanced data*, *i. e.*, data having one or more classes

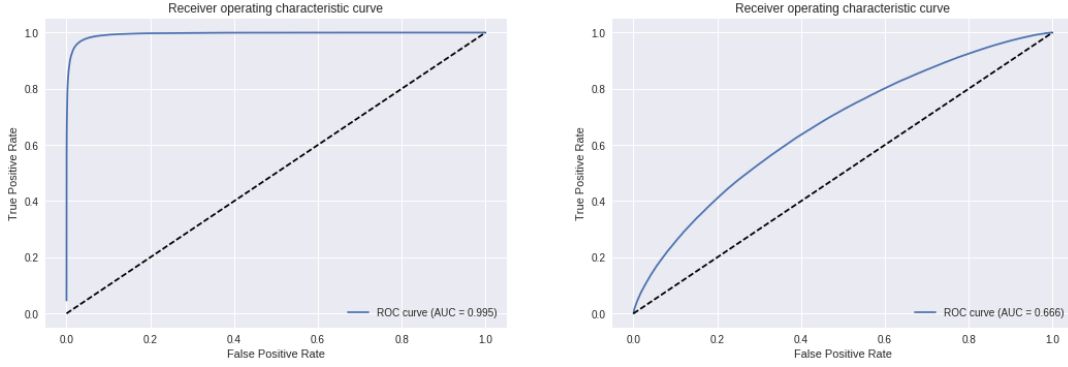


FIGURE 2.9: *Examples of ROC curves. In this particular case, a trained model was evaluated on the training data (left) and on hold-out test data (right). The significantly differing ROC AUCs (corresponding to the classification performances of the model) indicate that overfitting has occurred.*

that combine information of several metrics.

A metric which is commonly used in physics is the area under the *receiver-operating characteristic* curve (in short: “ROC AUC”). The ROC corresponds to signal efficiency vs. background acceptance – or in terms of the entries of the confusion matrix: true positive rate $\text{TPR} = \text{TP}/(\text{TP} + \text{FN})$ vs. false positive rate $\text{FPR} = \text{FP}/(\text{FP} + \text{TN})$. The ROC curve is constructed by varying the cut value on the MVA output variable and calculating signal and background efficiencies for each operating point. As a last step, the area under the resulting curve is calculated and reported as the ROC AUC. A schematic ROC curve is shown in Fig. 2.9. The ROC AUC can be interpreted in a statistical way: If you choose a pair of observations at random, one from the positive and one from the negative class, the ROC AUC is the probability of obtaining a larger classification score for the positive observation³⁹ [71]. ROC AUC as a metric is particularly useful for estimating and comparing classifier capabilities, but less useful if one has some knowledge about the optimal operating point, *i. e.*, the chosen cut value on the MVA output. In other words, the more one knows about the operating point, the less relevant the ROC AUC is [71].

When it comes to choosing an optimal working point, there is no general rule to follow. The decision is primarily driven by the specifics and requirements of the performed analyses, such as acceptable background levels or minimal signal efficiency. Since the MVA variable distributions for signal and background events overlap in most real-world applications, one has to find the best balance between false positives and false negatives [66]. It should also be mentioned that the ROC curve is not influenced by the class prior probabilities (*i. e.*, the “imbalancedness”

which constitute a relative majority among all samples: A classifier might, for example, learn to completely ignore a minority class in order to maximize the number of correctly classified samples. In this case, the model achieves high accuracy while not having learned anything about the particular class at all.

³⁹One can visualize this interpretation most easily by considering the limiting cases of $\text{AUC} = 1$ (perfect separation of the signal and background distributions in the MVA output) and $\text{AUC} = 0.5$ (identical signal and background distributions resulting in a model with the same predictive power as a randomly guessing classifier).

of the data), but the operating point is. Our approach of optimizing the operating point automatically takes these class prior probabilities into account. It will be described in Sec. 3.4.2.

2.4.4 Neural networks and Deep Learning

While there are numerous Machine Learning algorithms – some of them having proven to be useful and reliable in many domains of application, both theoretically and in practice (*e. g.*, Support Vector Machines [72–74]) – a large focus on methods that are based on ensembles of decision trees as well as on variants of neural networks seems to have been developed in recent years. For example, in high-energy physics, *boosted decision trees* (BDTs) and neural networks are widely and successfully used for data analysis. In this work, we employ large neural networks that consist of many layers of nodes (commonly called *deep neural networks*) for supervised classification tasks.

In order to briefly outline⁴⁰ the working principle of basic neural networks, it seems adequate to start with the definition of a linear classifier $y(\mathbf{x})$,

$$y(\mathbf{x}) = w_0 + \sum_{k=1}^m x_k \cdot w_k = w_0 + \mathbf{x}^T \cdot \mathbf{w}, \quad (2.4.1)$$

where $\mathbf{x} = \{x_1, x_2, \dots, x_m\}$ is the vector of basis functions x_k , $\mathbf{w} = \{w_1, w_2, \dots, w_m\}$ the weight vector, w_0 the *bias* term (*i. e.*, the offset from the origin in phase space) and m the number of features (*i. e.*, the dimensionality of the problem). Hyperplanes $y(\mathbf{x}) = c = \text{constant}$ correspond to linear decision boundaries in the feature space.

While this model might perform sufficiently well on data with classes that are linearly separable, it will not do so in case of data that requires non-linear decision boundaries. However, it is easy to extend this model to the non-linear case by replacing the linear basis functions, x_k , with non-linear ones, $h_k(\mathbf{x})$,

$$y(\mathbf{x}) = w_0 + \sum_{l=1}^L w_l \cdot h_l(\mathbf{x}). \quad (2.4.2)$$

Now, the summation is not limited to $l \in \{1, 2, \dots, m\}$ (with m being the number of dimensions of the feature space) any longer and can increase up to an arbitrary number L .

Until lately, among the most popular choices for h_l were the sigmoid function, $h(t) = 1/(1 + e^{-t})$, and the hyperbolic tangent, $h(t) = \tanh(t)$. In recent times, they seem to have been mostly replaced by functions like *ReLU* [75] (“rectified linear unit”), $h(t) = \max(0, t)$, or its variants⁴¹ for the hidden⁴² layer(s) and the *softmax* function [79], $h_j(t) = e^{t_j} / \left(\sum_{k=1}^K e^{t_k} \right)$ for $j = 1, \dots, K$ (K ...number of classes; $K = 2$ in case of binary classification) for the output layer. In any case,

⁴⁰Here, we closely follow the steps described in [66].

⁴¹ReLU variants include *Leaky ReLU* [76], *ELU* [77] and *SELU* [78].

⁴²Typically, neural networks consist of an input layer (where each feature of the training data corresponds to an input node), one or many hidden layers and an output layer (which encodes the network predictions by means of output node activations, whose number equals the number of target classes, *i. e.*, two output nodes in case of binary classification). For an illustration, see Fig. 2.10.

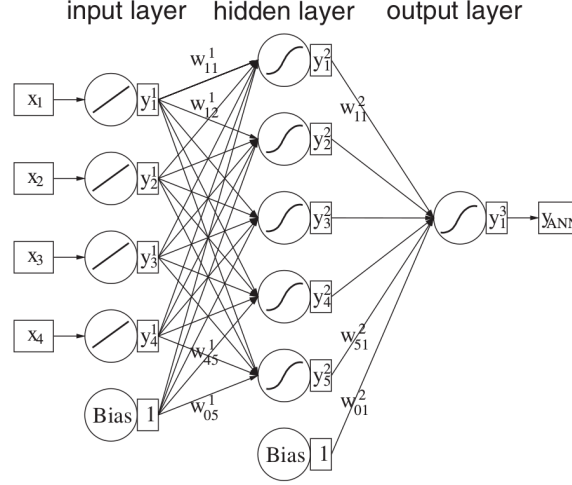


FIGURE 2.10: Graphical representation of the neural network described in Eq. 2.4.3 with $m = 4$ and $L = 5$ [66].

the output of the h_l depends non-linearly on the input t for a limited range while being (almost) constant elsewhere. This property makes the combination of the non-linear basis functions h_l in Eq. 2.4.2 well-suited for piece-wise approximation of the desired decision function and one can easily imagine that, in principle, any decision boundary can be approximated by this model [80–82]. As the input t for the basis functions $h(t)$, suitable linear combinations $t = \mathbf{w}^T \mathbf{x}$ of the observables \mathbf{x} are used. Particularly, for the l th basis function, $h_l(\mathbf{x}) = h\left((\mathbf{w}_l)^T(\mathbf{x})\right)$. Using just this one type of so-called *activation function*, Eq. 2.4.2 becomes

$$y(\mathbf{x}) = w_0^{(2)} + \sum_{l=1}^L \left[w_l^{(2)} \cdot h \left(w_{0l}^{(1)} + \sum_{k=1}^m w_{kl}^{(1)} \cdot x_k \right) \right], \quad (2.4.3)$$

where a constant bias with weight $w_{0l}^{(1)}$ was added. This bias allows the linear combination of the input variables to the node to be shifted to any working point of the activation function.

The non-linear classifier in Eq. 2.4.3 can be interpreted as a neural network having an input layer, one hidden layer and an output layer: The activation functions h_l correspond to the hidden nodes of the network which receive inputs via the connections to the nodes in the input layer, *i. e.*, the input data. The strength of a connection is given by its weight $w_{kl}^{(1)}$ in the linear combination with the input features x_k . The output node is fed with a linear combination of the outputs of the hidden nodes. The output node itself is just another activation function which maps its input to values between zero and one. This kind of neural network is usually trained in a way that the output node response peaks close to one in case of signal events and close to zero in case of background events. The neural network in Eq. 2.4.3 is visualized in Fig. 2.10. Conceptually, it is straightforward to enlarge the network architecture by adding more hidden layers or hidden nodes.

The network type described above and depicted in Fig. 2.10 is known as *feed-forward network* (FFN) or *multi-layer perceptron* (MLP) [66].

Having fixed the network architecture, one is still left with the problem of finding suitable weights such that the model is able to separate signal from background

events for the specific task. In order to find well-performing weights, one first has to define a loss function to quantify the performance of the model on training data. A suitable and common loss function for classification problems is the so-called cross-entropy.

$$L(\mathbf{w}) = \sum_i [\gamma^{(i)} \ln(\gamma(\mathbf{x}^{(i)}, \mathbf{w})) + (1 - \gamma^{(i)}) \ln(1 - \gamma(\mathbf{x}^{(i)}, \mathbf{w}))], \quad (2.4.4)$$

where $\gamma^{(i)}$ is the coded class membership⁴³ of training sample i and $\gamma(\mathbf{x}^{(i)}, \mathbf{w})$ is the network output for the i th event with weights \mathbf{w} . The sum in Eq. 2.4.4 runs over all samples in the training data. The most common approach to finding the loss function minimum by adjusting the network's weights is called *backpropagation* and described in the following.

As a preparatory step for backpropagation, one typically initializes the network weights with small and random values. Then, using events of the training sample, the gradient of the loss function is calculated as a function of the weights. The basic idea of backpropagation is to adjust the weights in a stepwise manner towards smaller losses. Thanks to the particular nature of the network (*i. e.*, a series of identical activations that are repeatedly combined in a linear way), the calculation of the gradient is possible by simply using the chain rule of differentiation. Its evaluation at a given position in *weight space* using training events is straightforward for the same reasons. The calculated gradients are fed back through the network by adjusting the weights by an amount $\mathbf{w} \rightarrow \mathbf{w} - \eta \cdot \nabla_{\mathbf{w}} L(\mathbf{w})$. The *learning rate* η specifies how fast the network moves into the indicated direction of smaller loss with each iteration.

Updating of the weights can be invoked either after the entire training sample has been fed through the network (*batch learning*) or after a small number of training samples (*mini-batches*) has been propagated through. It is also possible to update the weights after each forward pass of an event (*online learning*), but this approach is less commonly used since, due to the sequential nature of the weight optimization in this case, the network training can hardly be parallelized and usually takes unnecessarily long.

When extending the architecture of the neural network in Eq. 2.4.3 to a few hidden layers or beyond, it is common to refer to such a model as *deep neural network*. The mathematics of a network or of backpropagation does not change in principle when entering the realm of *Deep Learning*. However, networks with deep architectures seem to inhibit a remarkable property that is worth mentioning: Given enough flexibility, they seem to be able to learn abstract, high-level features from low-level input data, thus potentially obviating the need for manual feature engineering (see [83, 84] and references therein). Moreover, for the high-energy physics use case in particular, it has been shown that deep neural networks are capable of extracting information from low-level data which is not fully captured by the set of high-level features constructed by physicists [85]. This capability of deep networks to learn abstract and latent features on their own, as well as recent advances on the hardware, software and algorithmic side (see references in [86, sec. 5]), have arguably contributed to the revolutionary successes of Deep Learning applications in fields like computer vision, speech recognition, natural language processing, Reinforcement Learning, machine translation, bioinformatics and drug design, to name but a

⁴³ $\gamma^{(i)} = 1$: signal; $\gamma^{(i)} = 0$: background

few. A summary of much of the relevant work in the area of Deep Learning can be found in [86].

Especially in the field of high-energy physics there are enormous amounts of data available. The greatest part of it is *real* data that has been taken during experimental runs. However, sophisticated frameworks (*e. g.*, [87–89]) basically allow one to generate an arbitrarily large amount of labeled data in form of Monte Carlo simulations. These simulations are capable of describing the real data to a high degree of accuracy. Therefore, a multivariate analysis approach in form of Machine Learning or Deep Learning for high-energy physics seems promising and worth pursuing.

For more details on the exact implementations of the Machine Learning algorithms used for this work, see Sec. 3.4.2–3.4.4.

Chapter 3

Multivariate conversion rejection

The dielectron invariant mass spectrum has two main sources of background, *i. e.*, combinatorial e^+e^- pairs and pairs that contain tracks from photon conversions. This chapter sets out to study different ways to reject conversion pairs in a scenario involving the upgraded ALICE detector. In doing so, one of the central questions is whether multivariate approaches are able to increase the analysis performance with respect to conventional methods, which are based on sequential cuts. In Sec. 3.4.1, a conventional approach to conversion rejection in ALICE will be described, while the multivariate approaches will be detailed in the following sections (Sec. 3.4.2–3.4.4). All of these studies involve realistic particle identification (PID). Finally, the results of all analyses will be presented and compared in Sec. 3.5 and discussed in Sec. 3.6. Additionally, these sections will investigate the potential of the different analyses to reject not only pairs with conversion tracks, but *all* background contributions to the dielectron spectrum (*i. e.*, conversion pairs and all combinatorial pairs).

The analysis code that has been developed in the course of this work is publicly accessible on GitHub (https://github.com/tempse/ALICE_ITSupgrade).

3.1 Used dataset

The data used for this particular study is a Monte-Carlo (MC) simulation¹ of the upgraded ALICE detector (see Sec. 2.3.2) containing tracks of 1 million central Pb–Pb collisions at a nucleon-nucleon center-of-mass energy $\sqrt{s_{\text{NN}}} = 5.5$ TeV in a magnetic field of $B = 0.2$ T. This scenario represents the conditions that are expected during the dedicated ALICE low- B -field run in the LHC Run 3 period. At the event generator level, HIJING (*Heavy Ion Jet INteraction Generator* [90]) is used. Full tracking and detailed transportation of simulated particles are carried out for the LHC beam pipe and for the Inner Tracking System (ITS). For the remaining parts of the ALICE detector, a *Fast Simulation Tool*² is utilized, which does not perform detailed track propagation but instead makes use of different detector response parametrizations in order to speed up the MC production phase.

¹Internally, this data is referred to as “LHC16j6a”.

²This tool has been developed by R. Shahoyan and J. Stiller.

3.2 Used frameworks

The analyses carried out in this work require several different frameworks to be used and combined. In general, for all tasks *not* related to Machine Learning, the data analysis frameworks ROOT v6.12/04 [91] and AliRoot [92–94] (an extended version of ROOT v5.34/30 which is used within the ALICE collaboration) are employed. For all Machine Learning tasks (*i. e.*, defining, training and applying Machine Learning classifiers), Keras 2.0.8 [95] is used, with Tensorflow 1.4.0 [96] as its backend. The necessary code is written in Python 3.4.5 and heavily relies on NumPy³ 1.14.0 and Pandas⁴ 0.19.2. The conversion between ROOT trees and Pandas data structures is handled by root-numpy⁵ 4.4.0.dev0. The resulting plots are created using either ROOT or seaborn⁶ 0.8.1 on top of matplotlib⁷ 2.1.1.

3.3 Data preparation

For our analyses, unless stated otherwise, we use the track cuts given in Tab. A.1 in the appendix and select electrons/positrons only, which are identified by their MC information⁸. For realistic particle identification (PID), PID efficiencies from pp collisions at $\sqrt{s} = 13\text{ TeV}$ which were recorded by the ALICE detector with low magnetic field in 2016 are implemented⁹ (see Fig. 3.1). However, in order to additionally explore the limiting case of perfect PID (*i. e.*, efficiencies of 100 % at no contamination for all values of p_T), all analyses described in this chapter are also performed for this idealistic case and their results presented in App. D.

As one of the main preparatory steps, all electron/positron tracks undergo a pairing procedure, *i. e.*, all pair-wise combinations of tracks within a given event are formed. All pairs with oppositely charged¹⁰ tracks (the so-called *unlike-sign (ULS) pairs*) are subsequently stored for later analysis.¹¹ During pairing, all relevant quantities of a particle pair are calculated, such as the opening angle θ between the two momentum vectors \vec{p}_1 and \vec{p}_2 , the pair invariant mass $m_{ee} = \sqrt{2(\vec{p}_1 \cdot \vec{p}_2)(1 - \cos\theta)}$ or the angle φ_V between the magnetic field direction and the orientation of the opening angle (see Sec. 3.4.1 for a more detailed discussion of the variable φ_V). A list of all used pair variables is given in App. B (see Tab. B.2). Furthermore, the PID efficiency of a pair is assumed to be the product of the PID efficiencies of the individual tracks, *i. e.*, $\text{Eff}_{\text{pair}} = \text{Eff}_{\text{track 1}} \times \text{Eff}_{\text{track 2}}$.

³<http://www.numpy.org/>

⁴<https://pandas.pydata.org/index.html>

⁵http://scikit-hep.org/root_numpy/

⁶<http://seaborn.pydata.org/>

⁷[https://matplotlib.org/\protect\char"007D\relax](https://matplotlib.org/\protect\char)

⁸In the simulation data, each track stores information about its actual particle type, which is used for identifying electrons and positrons for the purposes of this work. In contrast, data measured by the detector (*i. e.*, “real” data) does not provide this kind of information and observables like particle type have to be assessed by other means.

⁹These PID efficiencies are considered in the analyses by associating each track with a weight that corresponds to the PID efficiency at the given p_T value. Then, whenever a track is counted or a histogram bin is filled with track variables, the counters are not incremented by one, but by the track’s PID efficiency value (*i. e.*, a real number between zero and one).

¹⁰The charges are obtained via the MC information of the tracks.

¹¹*Like-sign (LS)* pairs are discarded as combinatorial background before the pairing.

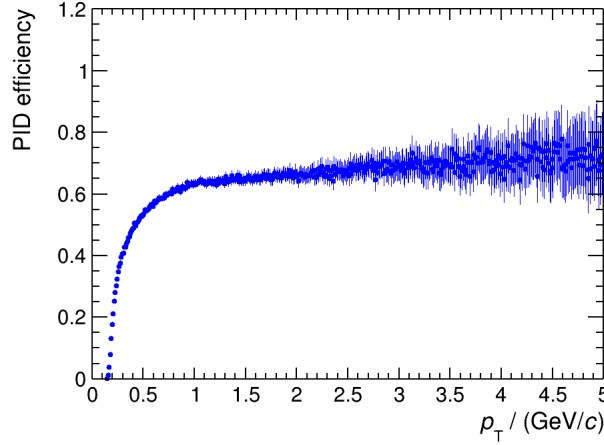


FIGURE 3.1: Particle identification (PID) efficiencies for electrons and positrons from pp collisions at $\sqrt{s} = 13$ TeV recorded by the ALICE detector with low magnetic field ($B = 0.2$ T) in 2016 as a function of transverse momentum p_T .

3.4 Analyses

The dielectron mass spectrum that is obtained after event-wise track pairing is shown in Fig. 3.2. The two main contributions to the overall yield are of combinatorial nature – they either contain at least one track per pair that comes from a photon conversion or no conversion tracks at all. Other contributions are non-combinatorial (*i. e.*, real) pairs from hadronic decays, from correlated heavy-flavor decays¹² and from photon conversions. The different ways¹³ to obtain e^+e^- pairs are graphically explained in Fig. 3.3. Some of the major difficulties of the analysis of low-mass dielectrons can be seen by merely looking at the dielectron spectrum: Over the entire mass range, it is dominated by a combinatorial background that is several orders of magnitude larger than the non-combinatorial contributions. Further, both the signal distribution and the conversion pair background have their largest contribution in the first mass bin, making the task of differentiating between them a difficult one.

The following sections will describe different approaches to reject conversion pairs, *i. e.*, pairs containing either one or two tracks from photon conversions. Eventually, their respective results are presented and compared in Sec. 3.5.

3.4.1 Bivariate *prefilter* cuts for conversion pair rejection (reference analysis)

At its point of creation, an e^+e^- pair from a photon conversion does not have an intrinsic opening angle since it has no invariant mass¹⁴. Its tracks are bent into the azimuthal direction only by the presence of the magnetic field that points along the z axis. This results in an opening angle θ between the two electrons that is expected

¹²These are decays of $c\bar{c}$ or $b\bar{b}$ quarks to e^+e^- pairs (see also Sec. 2.2).

¹³In principle, one is also interested in thermal radiation, which does not involve hadron decays. The simulation data used in this work, however, does not include such processes. For this reason, these processes are ignored in Fig. 3.3.

¹⁴This neglects the recoil momentum of the nucleus that is involved in the conversion process.

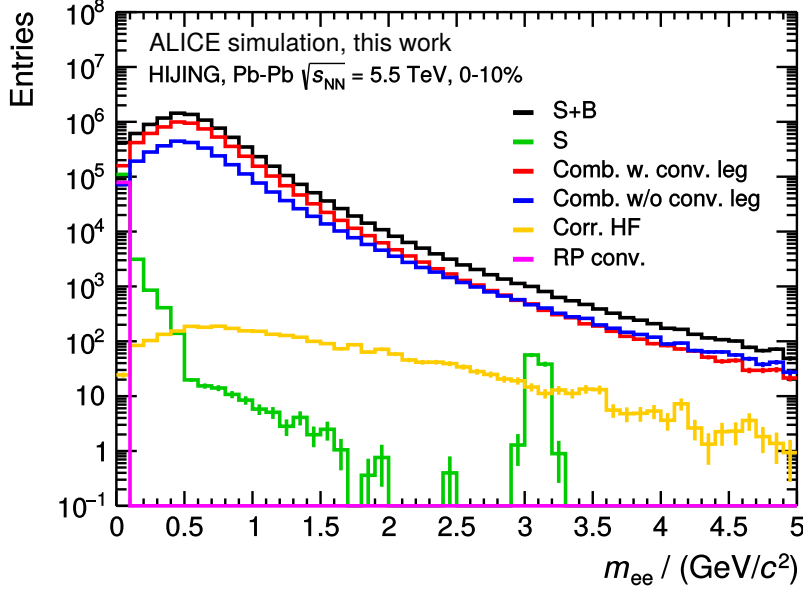


FIGURE 3.2: Dielectron invariant mass spectrum obtained for simulated Pb-Pb collisions at $\sqrt{s_{NN}} = 5.5$ TeV (with a centrality of 0–10 %) after track pairing. At the basis of the simulations, the event generator HIJING has been used. The overall yield (black) mainly consists of combinatorial pairs that either do (red) or do not (blue) contain conversion tracks, but also of non-combinatorial pairs with (magenta) and without (green) conversion tracks. Furthermore, the contribution from correlated heavy flavor pairs (yellow) is shown.

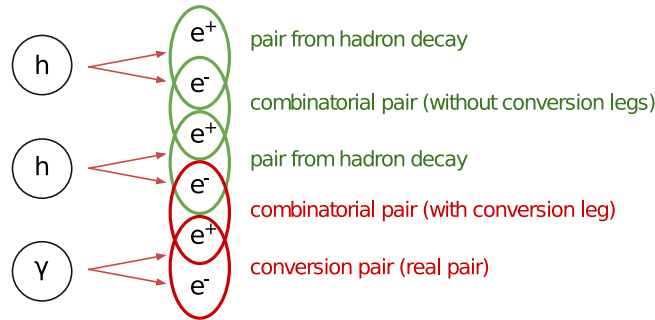


FIGURE 3.3: Nomenclature for all different pairs from hadron decays and photon conversions. The color coding corresponds to the signal (green) and background (red) definitions in the case of conversion rejection.

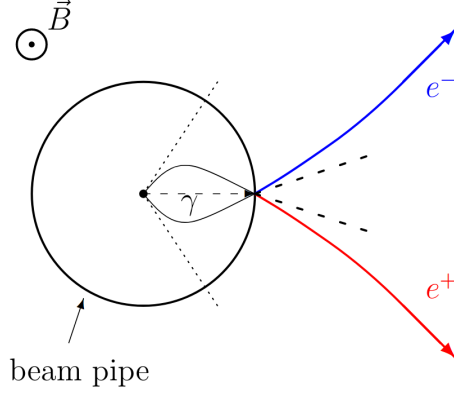


FIGURE 3.4: *Misreconstruction of the opening angle of a conversion pair. Since the e^+e^- pair is created off-vertex (contrary to the assumption of the tracking algorithm), the reconstructed opening angle (dotted lines) differs from the actual one (dashed lines) [97].*

to have the orientation

$$\hat{w}_c = \hat{u} \times \hat{z}, \quad (3.4.1)$$

where $\hat{u} = (\vec{p}_+ + \vec{p}_-)/|\vec{p}_+ + \vec{p}_-|$ is the unit vector in direction of pair momentum and \hat{z} the unit vector along the magnetic field. However, the tracking algorithm assumes all tracks to have their origin at the primary vertex location. Conversion pairs are usually created off-vertex and, consequently, they are assigned an apparent opening angle (and thus a non-zero mass) that is dependent on the distance between the primary vertex and the point of the conversion pair creation (see Fig. 3.4). The angle between \hat{w}_c and the orientation of the actual opening angle \hat{w} ,

$$\hat{w} = \hat{u} \times \vec{v}, \quad (3.4.2)$$

where $\hat{v} = (\vec{p}_+ \times \vec{p}_-)/(|\vec{p}_+| |\vec{p}_-|)$ is the unit vector perpendicular to the e^+e^- plane, is called φ_V [98],

$$\varphi_V = \arccos(\hat{w} \cdot \hat{w}_c). \quad (3.4.3)$$

For conversion pairs, $\varphi_V \approx \pi$, if consistent charge ordering of tracks within each pair is performed¹⁵. In contrast, both pairs from hadronic decays and combinatorial pairs do not show a preferred orientation of their opening angle (see Fig. 3.5). Using this information in combination with mass, one can identify conversion pair candidates at small values of mass m_{ee} and φ_V (*c.f.*, Fig. 3.6).

In this reference analysis, four different sets of sequential cuts on φ_V and m_{ee} are used in order to tag conversion pairs (see Definition 3.1).

¹⁵Without this charge ordering, $\varphi_V \approx 0$ is an equally viable solution for conversion pairs. It should also be noted that the charge ordering can be equally done in a way that conversion pairs peak at $\varphi_V \approx 0$ instead of $\varphi_V \approx \pi$.

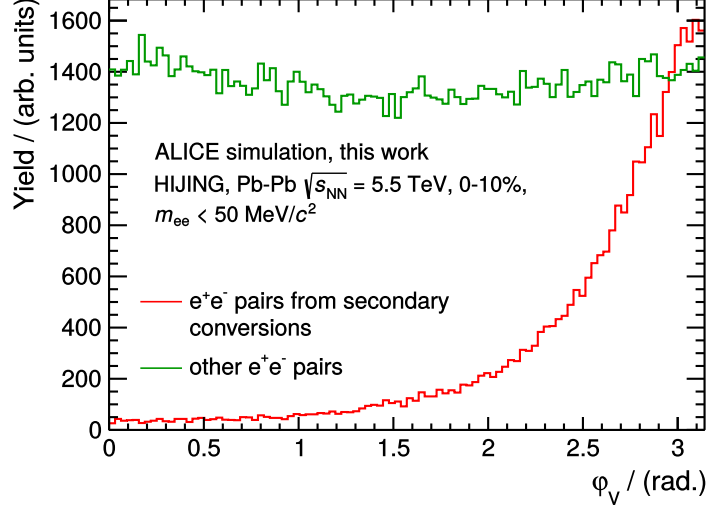


FIGURE 3.5: *Distribution of φ_V for e^+e^- pairs from secondary conversions (red) and for all other pairs (green). Only pairs with $m_{ee} < 50 \text{ MeV}/c^2$ are considered.*

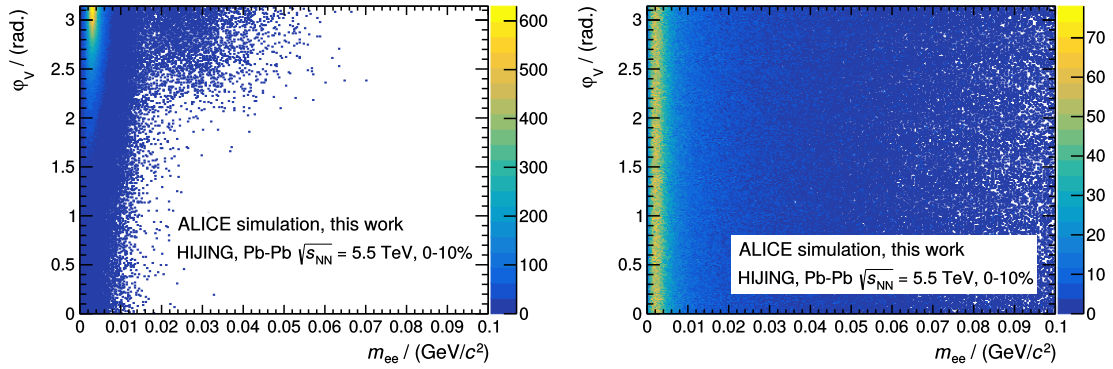


FIGURE 3.6: *Non-combinatorial conversion (left) and non-conversion (right) pairs in the φ_V - m_{ee} plane (identified by their MC information).*

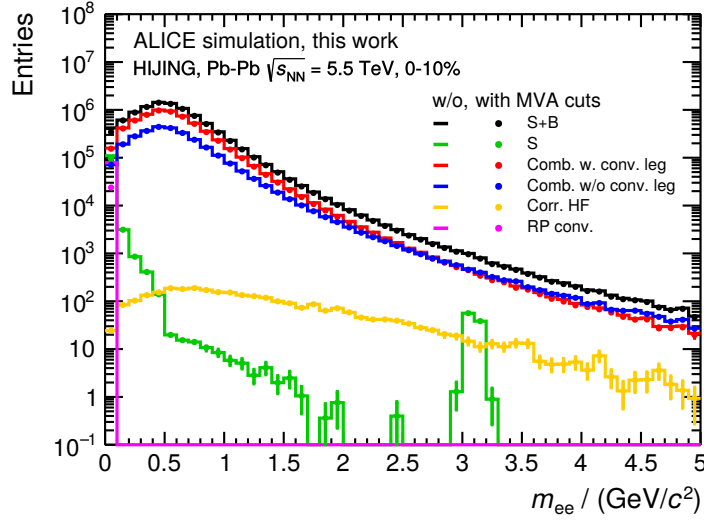


FIGURE 3.7: Dielectron invariant mass spectrum after the bivariate prefilter cut. The continuous lines correspond to the yield before the cut, the dotted distributions to the yield after the cut (using the “Bivar. Pref. Cuts B” setting of Definition 3.1).

Definition 3.1: Cut settings for bivariate *prefilter* analyses

- “Bivar. Pref. Cuts A”: $\varphi_V > \pi/2 \quad \wedge \quad m_{ee} < 50 \text{ MeV}/c^2$
- “Bivar. Pref. Cuts B”: $\varphi_V > 2 \quad \wedge \quad m_{ee} < 40 \text{ MeV}/c^2$
- “Bivar. Pref. Cuts C”: $\varphi_V > 2.4 \quad \wedge \quad m_{ee} < 10 \text{ MeV}/c^2$
- “Bivar. Pref. Cuts D”: $\varphi_V > 2.9 \quad \wedge \quad m_{ee} < 3.5 \text{ MeV}/c^2$

Pairs fulfilling one of these conditions are treated as conversion pairs in the respective analyses. Due to the nature of the pairing algorithm, the tracks of such identified pairs individually appear in other pairs as well. In fact, a large portion of the (combinatorial) pairs over the entire mass range has common tracks with conversion pairs, for which $m_{ee} \gtrsim 0$ and $\varphi_V \lesssim \pi$. In a two-step procedure called *prefiltering*, both types of conversion pairs are rejected as background by (1) identifying conversion pairs via rectangular cuts in the φ_V – m_{ee} plane and (2) subsequently finding all pairs of the same event that share tracks with them (see Algorithm 1 in the appendix). The effect of this bivariate *prefilter cut* on the invariant mass spectrum is shown in Fig. 3.7, where the continuous histogram lines correspond to the yield before the cut and the dotted distributions to the yield after the *prefilter* cut has been applied¹⁶. The efficiencies of the cut – *i. e.*, the yield in a given bin after the cut has been applied divided by the yield in the same bin before the cut – for each of the individual spectrum components are plotted in Fig. 3.8.

¹⁶The shown plots are obtained by using the *Bivar. Pref. Cuts B* setting (see Definition 3.1).

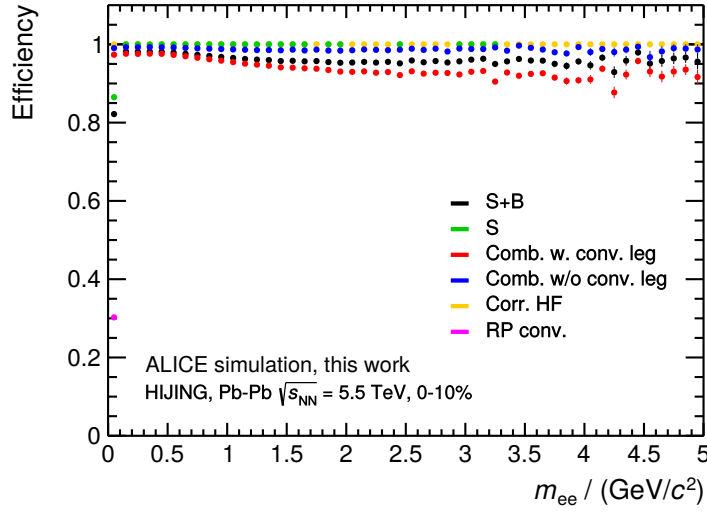


FIGURE 3.8: *Mass-dependent efficiencies of the bivariate prefilter cut for the different contributions of the dielectron invariant mass spectrum.*

3.4.2 Multivariate *prefilter* cut for conversion pair rejection

This analysis is very similar to the reference analysis which uses bivariate *prefilter* cuts (see Sec. 3.4.1). However, instead of manually defining rectangular cuts in the φ_V - m_{ee} plane during the first *prefiltering* step (*c.f.*, Algorithm 2 in the appendix), the task of identifying non-combinatorial conversion pairs is given to a Machine Learning algorithm¹⁷. In this way, the decision boundary that separates conversion pairs from non-conversion pairs is optimized in the entire space of input variables (instead of just in the lower-dimensional space of φ_V and m_{ee} like before). The second part of the *prefiltering* procedure is then performed in an analogous way.

As the chosen Machine Learning algorithm, a neural network is implemented that consists of two hidden layers with 100 nodes each. The initial network weights are randomly drawn from a truncated normal distribution which is centered at zero and has a standard deviation $\sigma = \sqrt{2/(n_j + n_{j+1})}$, where n_j is the number of input units and n_{j+1} the number of output units for the respective layer. This commonly used initialization scheme is known as *Glorot* or *Xavier normal initialization* [99]. The bias nodes of each layer are initialized with zeros. Throughout the network, *Rectified Linear Units (ReLU)* [75] are used as activation functions; except for the last layer, which implements the sigmoid activation function. During training, the mini-batch size is set to 128 samples and *batch normalization*¹⁸ [100] is performed for all layers consistently. The loss function is chosen to be the binary cross-entropy (see Eq. 2.4.4). As an optimizer, *Adam* [101] is employed, which is a first-order,

¹⁷The chosen Machine Learning algorithm is a neural network. For general notions of neural networks and Deep Learning as well as explanations of classifier evaluation, see Sec. 2.4.3 and Sec. 2.4.4.

¹⁸Batch normalization applies a transformation at each layer that maintains the mean activation close to zero and the activation standard deviation close to one. This technique not only improves the model performance in many cases, but also has a regularizing effect and thus helps preventing overfitting.

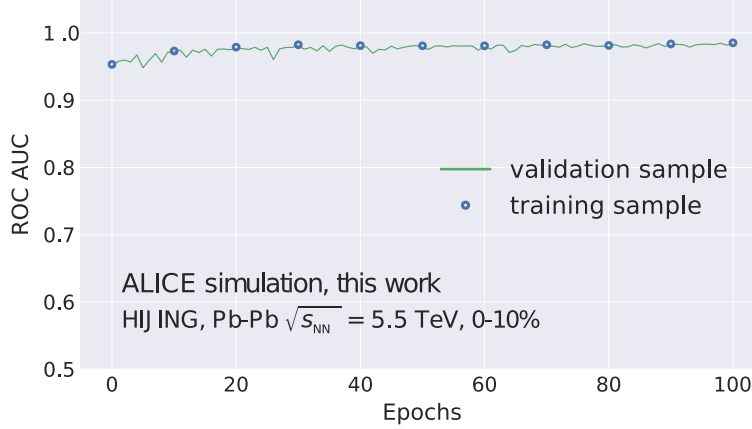


FIGURE 3.9: *Learning curve of the neural network that rejects non-combinatorial conversion pairs at low masses. In order to keep overfitting at a minimum, one has to make sure that the training ROC AUCs (blue) and the validation ROC AUCs (green) do not start to differ with growing number of epochs.*

gradient-based optimization algorithm for stochastic loss functions that is based on adaptive estimates of lower-order moments.

The training of the neural network is performed¹⁹ on about one million pairs²⁰ with $m_{ee} < 50 \text{ MeV}/c^2$. The restriction of the training samples to the low-mass range is necessary since, otherwise, the classifier tends to learn to report high classification scores by just treating the entire first mass bin as background, thus failing its actual classification task at low masses. This happens because of the very different signal–background compositions of the first mass bin(s) with respect to the higher-mass bins. A classifier that has mistakenly learned to focus too much on specific variable ranges due to different abundancies of signal and background samples along this variable is considered *biased*. (In the above case, such a classifier would be *mass-biased*²¹.) For the validation and training datasets, further 260 000 and 648 000 samples are used, respectively. As a default preprocessing step, the distribution of values of each feature is transformed such that it has zero mean and unit variance (*standard scaling*).

During training, overfitting is monitored by comparing the differences between training and validation ROC AUCs²². In order to maximize the generalization potential of the classifier, this difference has to be kept at a minimum (see Fig. 3.9). The performance of the trained classifier, that is ultimately reported in Sec. 3.5, is evaluated on the held-out test sample.

After successful classifier training, one has to decide on a working point, *i. e.*, on a cut value on the MVA output variable. The chosen approach of MVA cut optimization is to maximize the significance of the signal–background separation

¹⁹The number of training epochs is 100. In the context of the training of a neural network, an “epoch” refers to one forward pass and one backward pass of all training samples.

²⁰The fraction of positive- to negative-class samples in this dataset is about 80 % to 20 %.

²¹Generally, it is difficult to ascertain if a classifier is mass-biased, since many (kinematic) variables correlate with mass and can introduce such a bias. Several approaches to remedy this shortcoming have been proposed; see [102] for a recent example.

²²See Sec. 2.4.3 for a discussion of ROC AUC.

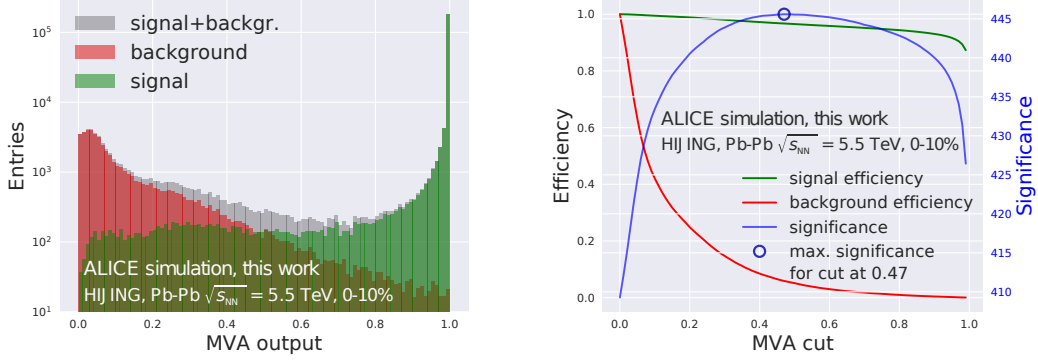


FIGURE 3.10: *Optimization of the MVA cut value (non-combinatorial conversion pair rejection). Left: MVA output of the classifier (gray), superimposed with its contributions from positive- (green) and negative-class (red) events. Right: Signal (green) and background (red) efficiencies as well as significances (blue) for each value of MVA output. The circle indicates the maximum significance.*

in the MVA output variable. This procedure is visualized in Fig. 3.10. In the MVA output distribution, the underlying signal and background distributions are identified using Monte-Carlo information. Subsequently, for each value of MVA output, signal and background efficiencies are calculated (see Fig. 3.10). These are used to compute the significance $S/\sqrt{S+B}$ (blue) for each MVA cut value. Eventually, the optimized working point is defined as the one corresponding to the MVA output value that maximizes the MVA cut significance. It is indicated by the small circle in Fig. 3.10.

Having optimized the working point, the trained classifier is applied to the rest of the data (463 million e^+e^- pairs). In doing so, one has to distinguish between pairs with $m_{ee} < 50 \text{ MeV}/c^2$ and ones with $m_{ee} \geq 50 \text{ MeV}/c^2$. In the former case, the trained classifier is allowed to cast a prediction and all pairs producing an MVA output smaller than the optimized MVA cut value are classified as background (*i. e.*, as non-combinatorial conversion pairs). In the latter case, the classifier predictions cannot be trusted, since the presented data differs from the training data, and all pairs with $m_{ee} \geq 50 \text{ MeV}/c^2$ are defined as belonging to the signal class²³.

Having identified non-combinatorial conversion pairs in the lowest-mass region in a multivariate way, the first part of the *prefiltering* procedure is concluded. What remains is the subsequent rejection of pairs which share tracks with those identified pairs. This is done in an entirely analogous way as in case of the sequential-cut analysis in Sec. 3.4.1.

By comparing the different dielectron spectrum contributions, one obtains mass spectra before and after the *prefilter* cut as well as efficiency plots similar to the ones presented before. For reasons of clarity, they have been moved to App. E and simply their mass-dependent efficiencies are presented in Sec. 3.5 (see Fig. 3.17 and Fig. 3.18), combined with the efficiencies of the other analyses.

All of the above steps in the multivariate *prefilter* analysis are repeated for

²³This classification is physically motivated as there are no non-combinatorial conversion pairs above this mass threshold.

datasets containing tracks with wider acceptance cuts. The corresponding analyses are referred to as “MVA 3 with loose tracks” (see Tab. A.2 in the appendix for the specific set of track cuts) and “MVA 3 with all tracks” (see Tab. A.3). Wider cuts allow for tracks with, *e. g.*, larger distances to the primary vertex (“distance of closest approach”, *DCA*) to be available for the pairing. This, in turn, is hoped to further improve the identification rate of non-combinatorial conversion pairs, since less strict track cuts enhance the probability of recovering partner tracks which have not been reconstructed with the “default” track cuts (see Tab. A.1). After the identification of conversion pairs using looser cuts, only tracks with the “default” cuts are processed again. The results of these analyses are presented together with the others in Sec. 3.5.

3.4.3 Multivariate conversion rejection on the single-track level

In this analysis, the quite involved and computationally expensive *prefiltering* logic (see Algorithm 1 and Algorithm 2) is abandoned. Instead, a conceptually much simpler approach is pursued: Rather than rejecting conversions on a pair level, a classifier is trained to reject conversion tracks²⁴ on the basis of individual, unpaired tracks. A list of all used track variables is given in App. B (see Tab. B.1).

For this purpose, a neural network with six hidden layers, each consisting of 100 nodes, is trained. The network weights are initialized with the *Glorot/Xavier normal initializer*, the bias nodes with zeros. Again, *ReLU*s serve as activation functions for the hidden layers, while the sigmoid function is used in the output layer. The mini-batch size is 128 and batch normalization is performed at each layer during training. Binary cross-entropy is used as a loss function and *Adam* as the optimizer algorithm.

This network is trained²⁵ on 5.32 million individual tracks²⁶, whose input variables have undergone *standard scaling*. The validation and test sample sizes are 591 000 and 656 000, respectively. Overfitting checks and working point optimization are performed in exactly the same way as before (see Sec. 3.4.2). The corresponding figures are shown in Fig. 3.11.

The trained classifier is applied to 37 million tracks. Subsequently, these tracks are combined to pairs. The MVA outputs of the individual tracks are consistently stored as sets of variables for each pair. Eventually, a pair which contains at least one track with an MVA output smaller than the optimized MVA cut value is rejected as conversion background. See Sec. 3.5 (Fig. 3.17 and Fig. 3.18) for the resulting MVA cut efficiencies and App. E for more detailed results regarding this specific analysis.

3.4.4 Multivariate rejection of combinatorial conversion pairs

Yet another approach of conversion rejection is the training of a classifier to identify and discard combinatorial pairs that contain at least one conversion track.

²⁴The class definitions therefore are: “signal” \leftrightarrow all tracks that do not originate from a photon, “background” \leftrightarrow all tracks that do originate from a photon.

²⁵The number of training epochs is 300.

²⁶Of these tracks, 56 % are non-conversion tracks and 44 % are daughter tracks of photons.

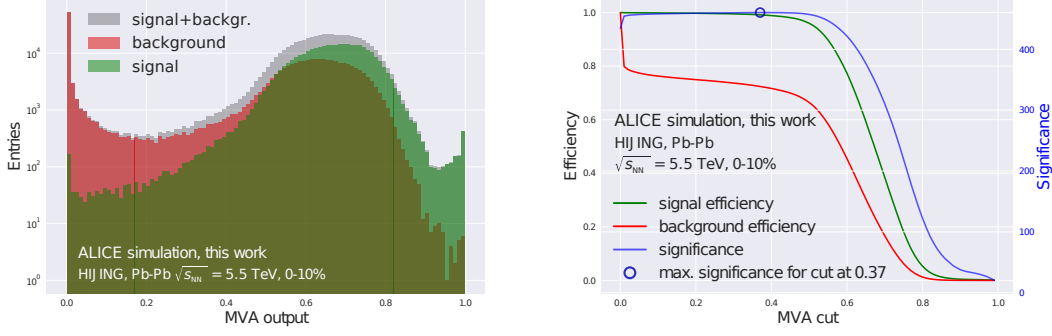


FIGURE 3.11: *Optimization of the MVA cut value (single-track conversion rejection). Left: MVA output of the classifier (gray), superimposed with its contributions from positive- (green) and negative-class (red) events. Right: Signal (green) and background (red) efficiencies as well as significances (blue) for each value of MVA output. The circle indicates the maximum significance.*

To this end, again, a neural network is used. Its architecture, activation functions and initialization scheme are the same as the network described in Sec. 3.4.3, but as an additional regularization technique, *dropout*²⁷ [103] at a rate of 0.25 at each network layer has been found to be helpful.

The neural network is trained²⁸ on 7.96 million e^+e^- pairs²⁹ with $m_{ee} \geq 50$ MeV/ c^2 . The exclusion of low-mass pairs for the classifier training is necessary for similar reasons as discussed in Sec. 3.4.2: Since the composition of the first mass bin differs significantly from the other mass bins (especially with regard to conversions), a classifier trained on the entire mass range tends to be *mass-biased*, *i. e.*, it focuses too much on mass-related quantities (*i. e.*, kinematic features like m_{ee} , θ , ...) and thereby fails to capture crucial details of the classification task. The validation and test data samples are created from additional 885 000 and 983 000 e^+e^- pairs, respectively. Again, overfitting checks and working point optimization are done analogously to the analyses before (see Sec. 3.4.2). Figure 3.12 shows the corresponding figures.

The trained and optimized classifier is subsequently applied to the remaining 463 million e^+e^- pairs. However, pairs with $m_{ee} < 50$ MeV/ c^2 are outside of the training scope and their class predictions therefore considered untrustworthy. They are consistently excluded from all results plots. In Sec. 3.5, the corresponding MVA cut efficiencies are presented, while further, more detailed results for the current analysis are given in App. E.

3.5 Results

In this section, the results of the analyses described in Sec. 3.4 are presented. First, the general performances of the different analysis approaches are compared in

²⁷In a nutshell, dropout refers to the practice of randomly dropping a given fraction of network nodes (along with their connections) during training, which prevents the network nodes from co-adapting too strongly. At test time, dropout schemes are no longer applied, but the entire network is used again.

²⁸The number of training epochs is 300.

²⁹The percentage of positive- and negative-class events is 32 % and 68 %, respectively.

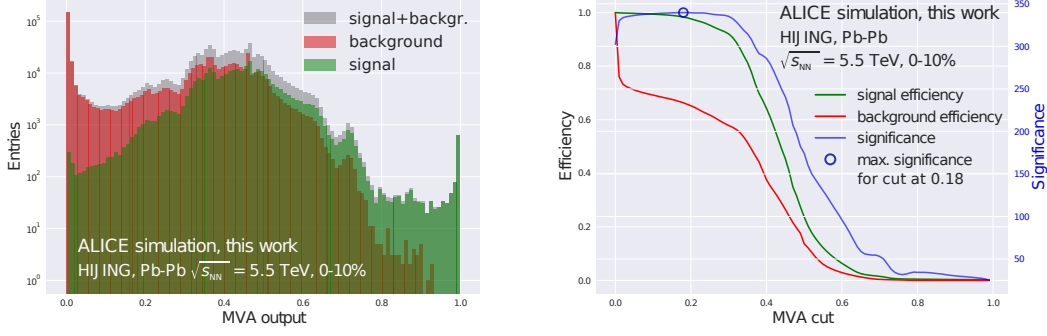


FIGURE 3.12: *Optimization of the MVA cut value (combinatorial conversion rejection). Left: MVA output of the classifier (gray), superimposed with its contributions from positive- (green) and negative-class (red) events. Right: Signal (green) and background (red) efficiencies as well as significances (blue) for each value of MVA output. The circle indicates the maximum significance.*

Sec. 3.5.1. Subsequently, Sec. 3.5.2 shows the effectiveness of the different conversion rejection methods at their respective optimized working points.

3.5.1 General classifier performances

As discussed in Sec. 2.4.3, a suitable metric for the general comparison of classifier performances is the area under the *receiver-operating characteristic* curve (“ROC AUC”). The possible values for ROC AUC range between $AUC = 1$ (perfect signal–background discrimination) and $AUC = 0.5$ (classification performance equal to a “random-guessing classifier”).

In Fig. 3.13, the ROC curves of all the multivariate approaches that have been discussed in Sec. 3.4 are shown as continuous lines. They have been given pseudonyms in order to keep the plot legends legible (see Definition 3.2).

Definition 3.2: Pseudonyms for multivariate analysis approaches

- *MVA 1* (green) \leftrightarrow rejection of combinatorial conversion pairs (see Sec. 3.4.4)
- *MVA 2* (red) \leftrightarrow rejection of conversion on the single-track level (see Sec. 3.4.3)
- *MVA 3* (blue) \leftrightarrow rejection of conversion pairs via a multivariate *prefilter* cut (see Sec. 3.4.2).

For each of these ROC curves, a small circle indicates the optimized working point (belonging to the MVA output value that maximizes the significance of the signal–background separation). The yellow curve represents a combination³⁰ of the methods *MVA 1* (green) and *MVA 3* (blue). The bivariate *prefilter* cut based reference

³⁰To be exact, this combined classifier rejects a pair as conversion background as soon as either the *MVA 1* or the *MVA 3* classifier identifies at least one conversion track in the pair at a given MVA cut value.

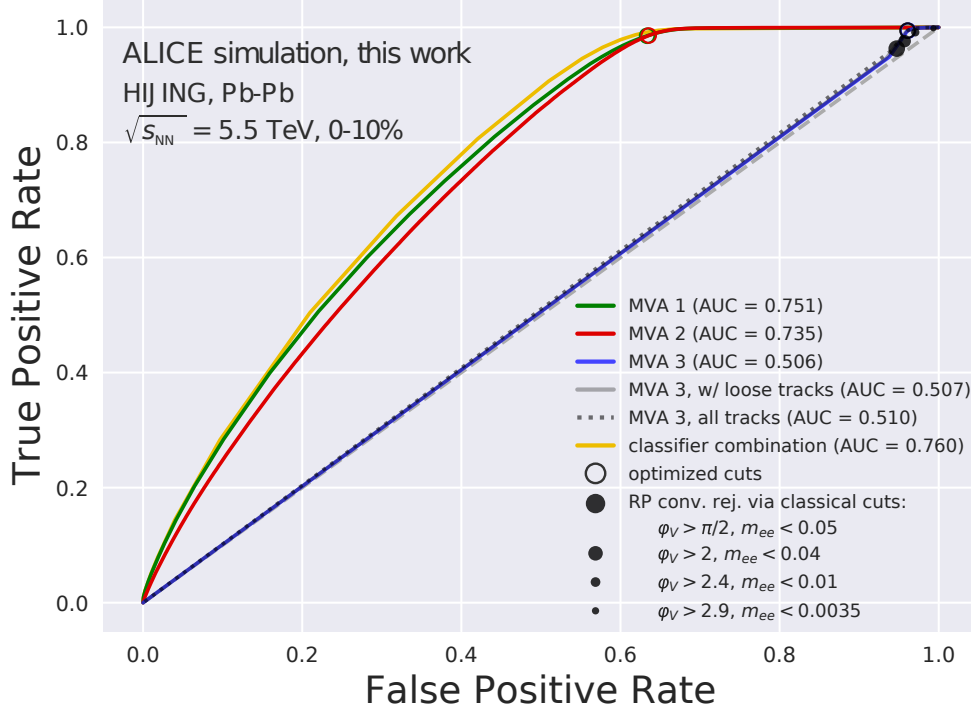


FIGURE 3.13: ROC curves of the different conversion rejection analysis approaches. For an explanation of the pseudonyms in the legend, see Definition 3.2.

analyses with the four different cut settings (see Definition 3.1) are represented by the black dots near the top right of Fig. 3.13. A zoomed version of this ROC region is additionally provided in Fig. 3.14. The resulting values for ROC AUCs are listed in Tab. 3.1.

For each classifier, as well as for each of the reference analysis working points, the significance of the conversion background rejection is plotted as a function of signal efficiency in Fig. 3.15. Again, a zoomed version is supplied in Fig. 3.16.

3.5.2 Classifier performances at the optimized working points

While the general performances of the different conversion rejection methods constitute informative and vital quantities (*c.f.*, Sec. 3.5.1), it is particularly interesting

MVA approach	ROC AUC in %
MVA 1	75.1
MVA 2	73.5
MVA 3	50.6
MVA 3 (with loose tracks)	50.7
MVA 3 (all tracks)	51.0
MVA 1+MVA 3	76.0

TABLE 3.1: ROC AUCs of the different MVA approaches (conversion rejection).

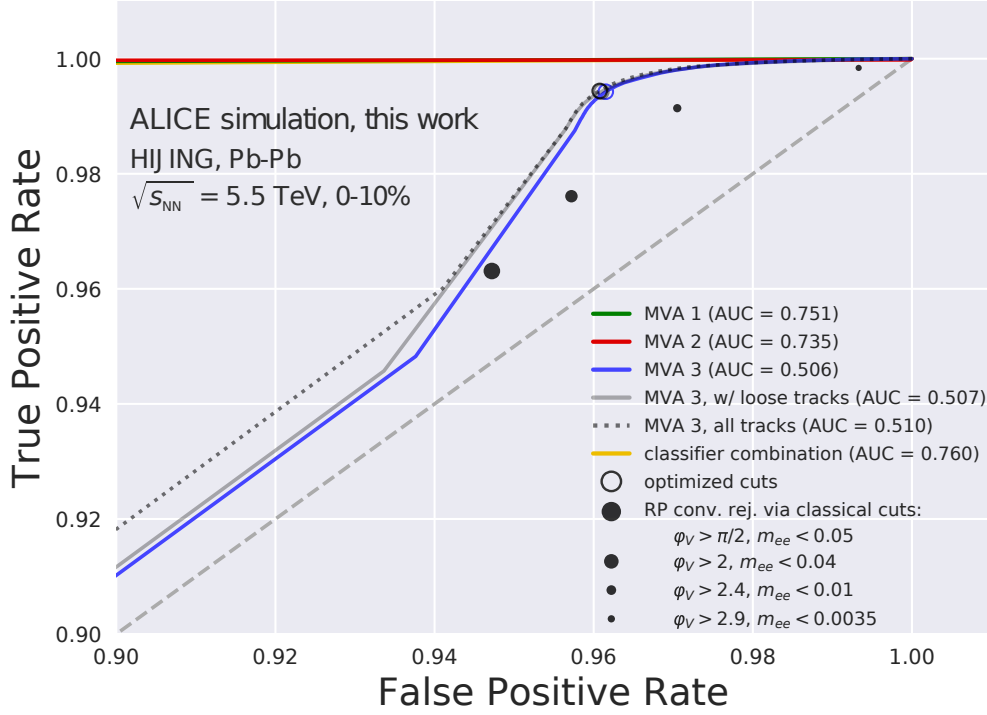


FIGURE 3.14: *ROC curves of the different conversion rejection analysis approaches (zoomed). For an explanation of the pseudonyms in the legend, see Definition 3.2.*

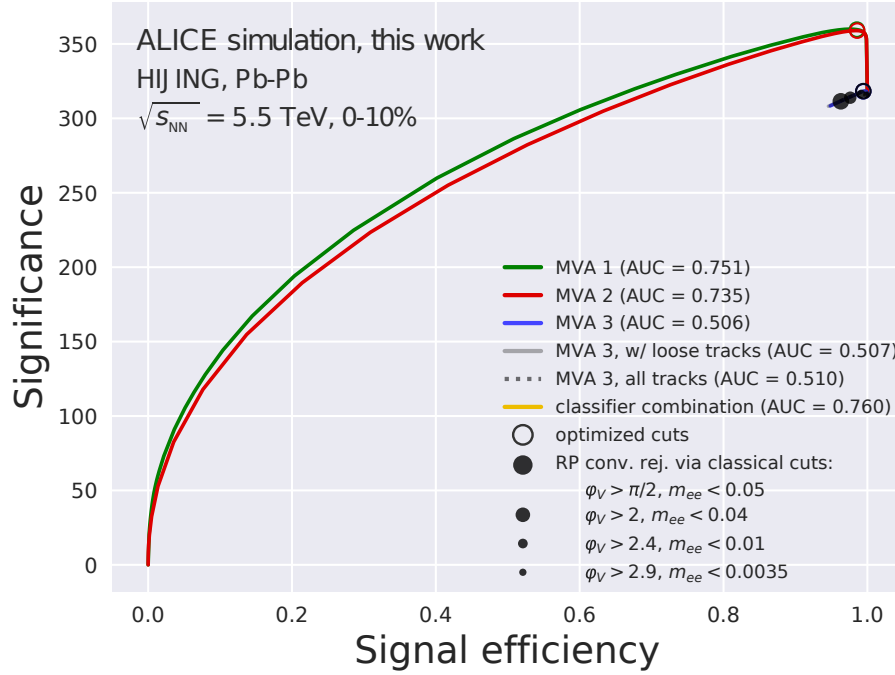


FIGURE 3.15: *Significance vs. signal efficiency for the different conversion rejection analysis approaches. For an explanation of the pseudonyms in the legend, see Definition 3.2.*

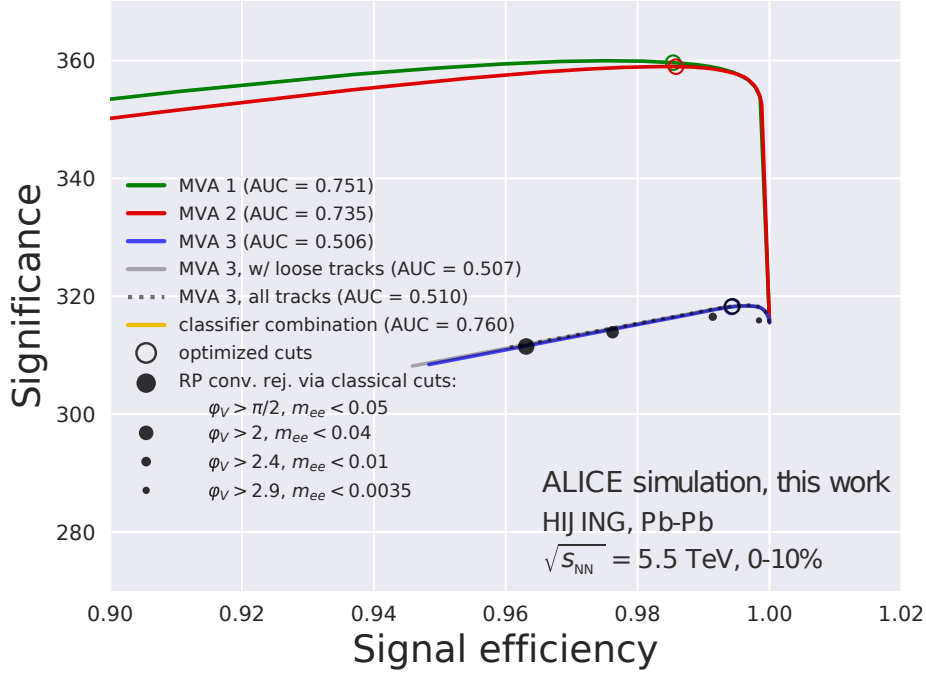


FIGURE 3.16: *Significance vs. signal efficiency for the different conversion rejection analysis approaches (zoomed). For an explanation of the pseudonyms in the legend, see Definition 3.2.*

to know which effects the respective classifiers have on the different contributions of the dielectron invariant mass spectrum *at their chosen working points*. In order to visualize the behavior of the optimized classifiers, their respective cut efficiencies are plotted as a function of invariant mass for each contribution (see Fig. 3.17 and Fig. 3.18). In addition to the different multivariate cut efficiencies, those of the bivariate *prefilter* cuts are plotted as black squares.

Furthermore, for all analysis approaches, the significance³¹ $S/\sqrt{S+B}$ (with $S \leftrightarrow S_{\text{class}}$ and $B \leftrightarrow B_{\text{class}}$) of the respectively applied cuts and the resulting signal-to-background ratios ($S_{\text{class}}/B_{\text{class}}$) are calculated and plotted for each bin of invariant mass (see Fig. 3.19 and Fig. 3.20). These figures also contain ratio plots, which show the respective values for significance or $S_{\text{class}}/B_{\text{class}}$ normalized to the values in the original dataset, where no cuts have been applied.

The studies carried out in this entire chapter focus on rejection of pairs with either one or two conversion tracks, which is reflected in the implicit signal and background definitions in the significances and $S_{\text{class}}/B_{\text{class}}$ values in Fig. 3.19 and Fig. 3.20: Pairs that contain at least one conversion track are considered background, while all other pairs are treated as signal. Ultimately, from a physics point of

³¹Since in the definition of significance, the signal S appears in both the numerator and the denominator, the errors cannot be assumed to be uncorrelated in general. In this work, therefore, a simple bootstrapping approach for estimating the magnitude of the significance error has been chosen: 50 bootstrap samples (*i. e.*, samples with the same size of the original dataset which have been randomly drawn with replacement) are created and their individual significance values are calculated. The standard deviation of the resulting distribution is then assigned to the error of the original significance.

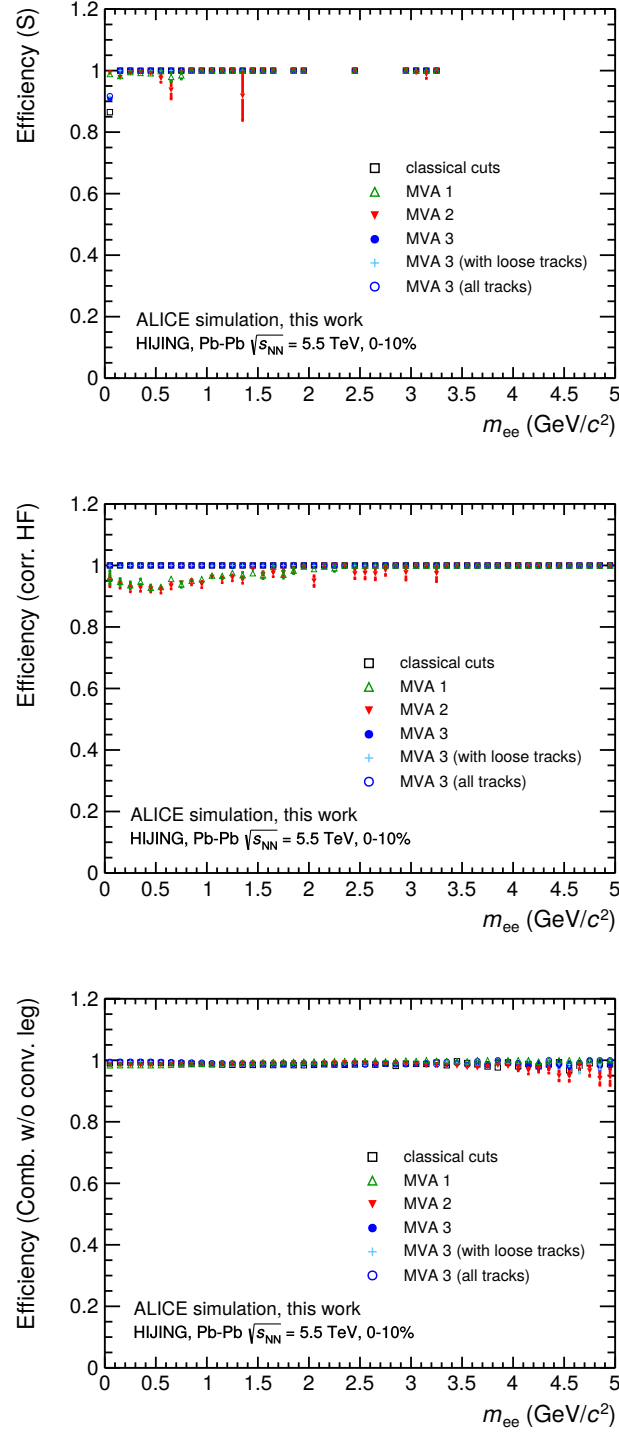
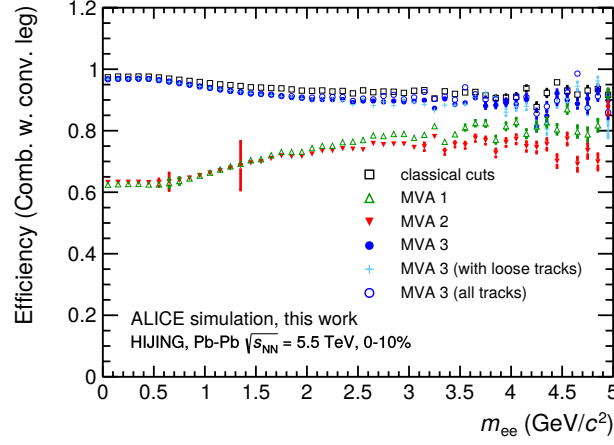
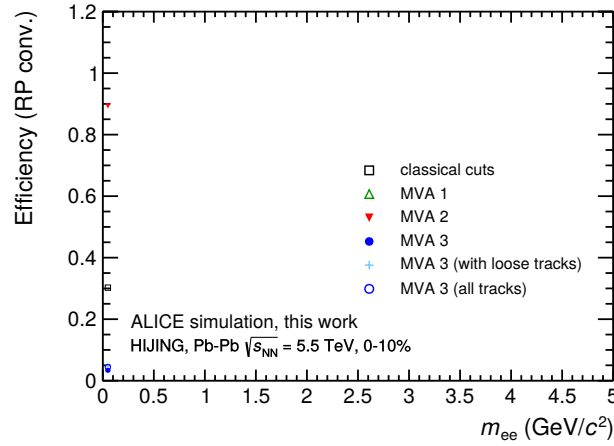


FIGURE 3.17: *Cut efficiencies for different dielectron sources for different classifiers at their optimized working points (1/2). Top: non-combinatorial non-conversion pairs. Center: correlated heavy-flavor pairs. Bottom: combinatorial pairs without conversion tracks. For an explanation of the pseudonyms in the legend, see Definition 3.2.*



(A)



(B)

FIGURE 3.18: *Cut efficiencies for different dielectron sources for different classifiers at their optimized working points (2/2). Top: combinatorial pairs with one or two conversion tracks. Bottom: non-combinatorial conversion pairs. For an explanation of the pseudonyms in the legend, see Definition 3.2.*

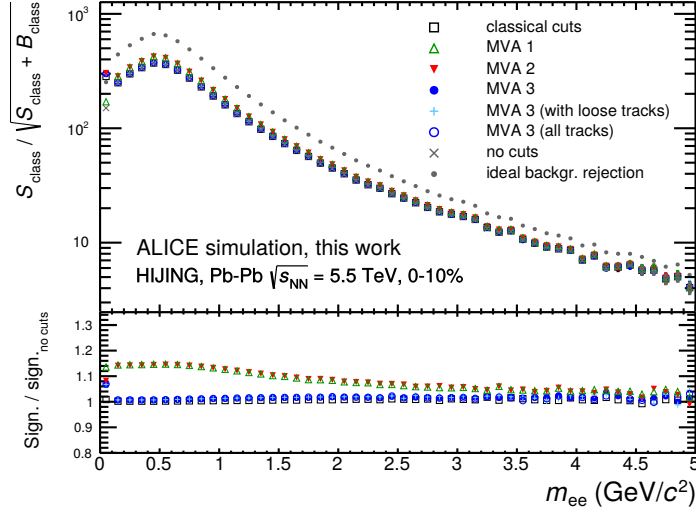


FIGURE 3.19: Mass-dependent significances $S_{\text{class}}/\sqrt{S_{\text{class}} + B_{\text{class}}}$ (with S_{class} and B_{class} corresponding to the signal and background definitions for pure conversion rejection) of the different conversion rejection analyses. Also shown are the limiting cases of no applied cuts and of ideal (i. e., complete) conversion background rejection. The ratio plot on the bottom shows the respective significances normalized to the case of no applied cuts. For an explanation of the pseudonyms in the legend, see Definition 3.2.

view, one wants to reject conversions and *all* combinatorial pairs (not only the combinatorial pairs with conversion tracks). Applying the same classifiers, as they were trained before, to the data, but changing the signal–background definition accordingly³² ($S_{\text{class}} \rightarrow S_{\text{exp}}$, $B_{\text{class}} \rightarrow B_{\text{exp}}$), results in the distributions for significance $S_{\text{exp}}/\sqrt{S_{\text{exp}} + B_{\text{exp}}}$ and $S_{\text{exp}}/B_{\text{exp}}$ that are shown in Fig. 3.21 and Fig. 3.22. It should be noted that these distributions do not necessarily represent the optimal performances for this signal–background definition, since the classifiers have not been trained for this particular classification task. (Nevertheless, one can visualize and study the behavior of these classifier with respect to S_{exp} and B_{exp} .)

3.6 Discussion

In this work, different multivariate analysis (MVA) approaches for the rejection of pairs that contain conversion tracks were studied and compared to “conventional” analysis, which are based on sequential (bivariate) cuts. The results were presented in Sec. 3.5.

It has been shown that all of the multivariate analyses outperform the conventional approach consistently: Not only are the general classifier performances superior to the bivariate *prefilter* cuts by means of ROC AUC (see, *e. g.*, Fig. 3.13), but all multivariate classifiers are also able to significantly improve the rejection of con-

³²That is, $B_{\text{exp}} \leftrightarrow$ conversion pairs and all combinatorial pairs; $S_{\text{exp}} \leftrightarrow$ non-combinatorial pairs without conversion tracks and correlated heavy-flavor pairs.

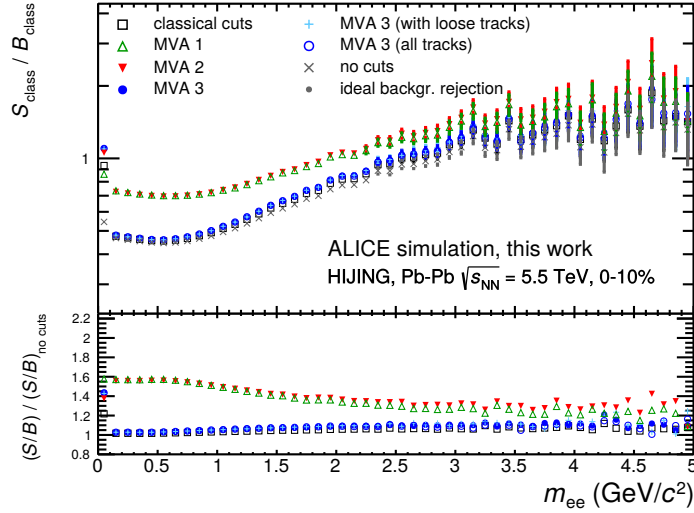


FIGURE 3.20: Mass-dependent $S_{\text{class}}/B_{\text{class}}$ (with S_{class} and B_{class} corresponding to the signal and background definitions for pure conversion rejection) of the different conversion rejection analyses. Also shown are the limiting cases of no applied cuts and of ideal (i. e., complete) conversion background rejection. The ratio plot on the bottom shows the respective signal-to-background ratios normalized to the case of no applied cuts. For an explanation of the pseudonyms in the legend, see Definition 3.2.

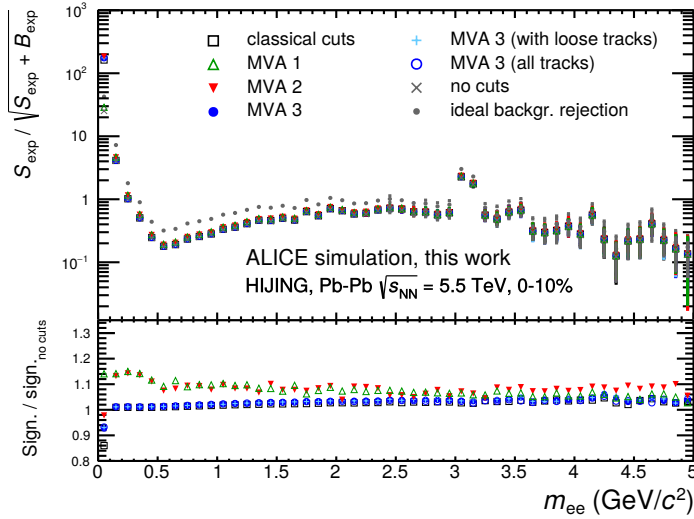


FIGURE 3.21: Mass-dependent significances $S_{\text{exp}}/\sqrt{S_{\text{exp}} + B_{\text{exp}}}$ (with S_{exp} and B_{exp} corresponding to the signal and background definitions for the rejection of conversions and all combinatorial sources). Also shown are the limiting cases of no applied cuts and of ideal (i. e., complete) conversion background rejection. The ratio plot on the bottom shows the respective significances normalized to the case of no applied cuts. For an explanation of the pseudonyms in the legend, see Definition 3.2.

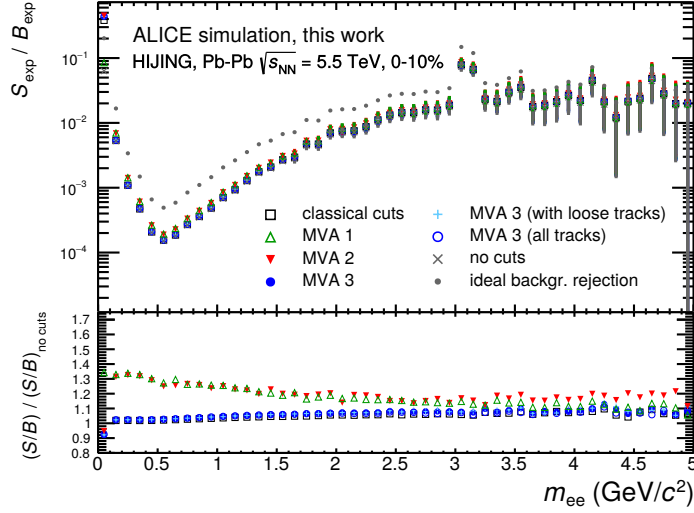


FIGURE 3.22: Mass-dependent $S_{\text{exp}}/B_{\text{exp}}$ (with S_{exp} and B_{exp} corresponding to the signal and background definitions for the rejection of conversions and all combinatorial sources). Also shown are the limiting cases of no applied cuts and of ideal (i. e., complete) conversion background rejection. The ratio plot on the bottom shows the respective signal-to-background ratios normalized to the case of no applied cuts. For an explanation of the pseudonyms in the legend, see Definition 3.2.

version pairs over the entire mass range when applied at their respective optimized working points (see Fig. 3.19 and Fig. 3.20). In terms of signal-over-background ($S_{\text{class}}/B_{\text{class}}$), two MVA approaches yield an improvement of up to 60 % over the conventional analyses, i. e., the classifier trained to reject combinatorial conversion pairs (*MVA 1*) and the classifier trained to rejection conversions on a single-track level (*MVA 2*). The gain in the corresponding significances is of the order of up to 15 %. The superiority of those methods still holds when one reformulates the definitions of signal and background ($S_{\text{class}} \rightarrow S_{\text{exp}}$, $B_{\text{class}} \rightarrow B_{\text{exp}}$) such that, in addition to conversion pairs, *all* combinatorial pairs (even those which do *not* contain conversion tracks) are considered background. In this case the same two MVA methods as before achieve an improvement of up to 30 % regarding $S_{\text{exp}}/B_{\text{exp}}$ and up to about 15 % regarding the corresponding significances. One should also note that the values of significance gain for both signal–background definitions are similar only for the lowest mass bins, where there is no additional background from other combinatorics.

One classifier – the multivariate *prefilter* approach (*MVA 3*) – shows similar performance as the conventional, bivariate *prefilter* method. However, both are barely able to improve the results over the case where no background rejection whatsoever is applied. This is a result of the implemented particle identification (PID) efficiencies (see Fig. 3.1), which suppress low- p_T pairs and thus most of the non-combinatorial conversion pairs. Simply put, after PID efficiency application, there are almost no conversion pairs left that could be rejected via bi- or multivariate cuts. Therefore, opening the track cuts for the *prefilter* does not significantly improve the situation,

as can be seen by the almost identical performances of the *MVA 3* analysis variants with looser track cuts (see Fig. 3.14). Another reason why the use of looser track cuts does not dramatically improve the classifier performances is indicated by studies which assume perfect PID (see App. D). In these analyses, too, the loose track cuts do not lead to a major performance boost, from which it can be concluded that one of the main reasons for this observation is the fact that the “default” track cuts (see Tab. A.1) are quite loose already.

Another interesting result is the realization that the two approaches *MVA 1* and *MVA 2* perform almost equally well, both in terms of ROC AUCs and observables like S/B at their optimized working points. Normally, one could expect that the Machine Learning algorithms have more information available when being trained on e^+e^- pairs, compared to the case of individual tracks where there are no pair variables. Although the pair-approach is indeed a little better in terms of ROC AUC, the optimized working points have almost the same signal and background efficiencies and, therefore, basically produce the same results when applied to the data. This result, although maybe surprising, is encouraging as the classification of individual tracks arguably has some advantages over methods based on e^+e^- pairs. Most importantly, using individual tracks without information about pair variables, one is less prone to obtain classifiers that are mass-biased (*i. e.*, ones that have learned to overly focus on specific mass regions in order to maximize their classification score). As another side benefit, compared to working with pairs and thus with large combinatorics, the handling and processing of individual tracks seems much easier, at least on a conceptual level.

In conclusion, it has been shown that multivariate approaches are able to improve the analysis of low-mass dielectrons in ALICE, in some cases even significantly. Furthermore, not only do these approaches yield better analysis results, but they also take away a lot of complexity from the entire analysis process. Concretely, it has been demonstrated that the training of a classifier on individual tracks yields results that are basically equivalent to multivariate pair-level analyses and much better than any methods involving *prefilter* cuts. Future analyses might therefore be able to abandon the numerically and conceptually quite involved *prefilter* approaches in favor of simpler and better-performing multivariate classifiers that are based on individual tracks.

Chapter 4

Multivariate heavy flavor rejection

A major contribution to the dielectron invariant mass spectrum (see Sec. 2.2) stems from e^+e^- pairs that are decay products of heavy flavor (HF) quarks, *i. e.*, the processes $c\bar{c} \rightarrow Xe^+e^-$ and $b\bar{b} \rightarrow Xe^+e^-$. They can be categorized into *correlated* and *combinatorial* HF pairs. The former are dielectrons originating from the very same HF quark pair, while the latter are pairs whose individual particles originate from different (“independent”) c or b quarks. HF pairs are an important component of the dielectron spectrum, especially for intermediate masses ($1 < m_{ee} < 3 \text{ GeV}/c^2$) and above. However, the cross-section and invariant mass shape of HF decays in heavy-ion collisions are not well understood. Therefore, a viable analysis strategy is to treat HF pairs as another physical background source and try to reject them as much as possible.

This chapter probes different ways to characterize and suppress HF pairs via multivariate analysis (MVA) techniques. The aim of these studies is to estimate their feasibility for future analyses of the HF content of the dielectron invariant mass spectrum. The analysis methods employed here are mostly the same as in the case of multivariate conversion rejection (see Chap. 3), but involve different specifications of the used classifiers and different datasets with other signal/background definitions, which will be detailed in the following sections.

4.1 Used datasets

For this study, an ALICE Monte Carlo (MC) production of more than 66 million pp collisions at $\sqrt{s} = 13 \text{ TeV}$ in a reduced magnetic field ($B = 0.2 \text{ T}$) is used¹. Due to the small amount of HF tracks in this so-called *general-purpose* (GP) production, the training of multivariate classifiers for HF rejection is difficult. In order to compensate for this, a *heavy flavor enhanced* MC production² with about 33 million events is used in addition³. In contrast to the previous studies in this work (*c. f.*, Chap. 3), the simulated geometry in both MC productions represents the ALICE detector in its current state, *i. e.*, before the planned upgrade for Run 3/4 (see Sec. 2.3). At the event generator level, these simulations employ PYTHIA [104]. The track cuts used for both MC productions are listed in Tab. A.1 in the appendix.

¹This data is internally labeled as “LHC17d1”.

²In the heavy flavor enhanced simulation scenario, only those events which originate from heavy flavor quarks are kept and stored.

³This data is internally labeled as “LHC17d12”.

4.2 Used frameworks

The frameworks utilized for these studies equal the ones that have been used for the conversion rejection studies. See Sec. 3.2 for more details.

4.3 Data preparation

Having investigated the rejection of conversion tracks/pairs in Chap. 3 and intending to focus on the rejection of HF pairs only, conversion tracks/pairs are ignored altogether and removed from both datasets for the training of multivariate classifiers. As another preparatory step, from the HF enhanced MC production, only those e^+/e^- tracks that originate from either one $c\bar{c}$ or one $b\bar{b}$ pair are selected⁴.

From the accepted tracks in each dataset, 20 % of all tracks are split off and reserved for the training of the multivariate classifiers. Subsequently, all data samples separately undergo the same pairing procedure as described in Sec. 3.3. Finally, in preparation to the classifier training, the training data samples are combined and their entries (*i. e.*, tracks or pairs, depending on the analysis scenario) are shuffled. Eventually, this procedure yields training data samples without conversion electrons, but with an enhanced number of HF tracks/pairs relative to the HF contents in the GP production.

4.4 Analyses

After track pairing, the dielectron invariant mass spectrum at the basis of the following analyses is obtained and shown in Fig. 4.1.

For the purpose of HF pair rejection, several multivariate classifiers are developed for different training scenarios, *i. e.*, for the setting of signal and background according to Definition 4.1 (the first two cases involve the training on e^+e^- pairs, whereas for the last one, the MVA aspects are performed on individual tracks before the pairing).

Definition 4.1: Training scenarios for multivariate HF rejection

1. “RP::HF”:
 $signal \leftrightarrow non\text{-}combinatorial\ pairs\ (without\ conversion\ tracks\ and\ HF),$
 $background \leftrightarrow heavy\ flavor\ pairs$
2. “CombBG::HF”:
 $signal \leftrightarrow combinatorial\ pairs\ (without\ conversion\ tracks\ and\ HF),$
 $background \leftrightarrow heavy\ flavor\ pairs$
3. “NonConvs::HF”:
 $signal \leftrightarrow tracks\ from\ the\ GP\ data\ (without\ conversion\ tracks\ and\ HF),$
 $background \leftrightarrow heavy\ flavor\ tracks$

⁴To a small extent, the HF enhanced production also contains events which show *two* pairs of $c\bar{c}$ or $b\bar{b}$ at the lowest level of their particle decay history.

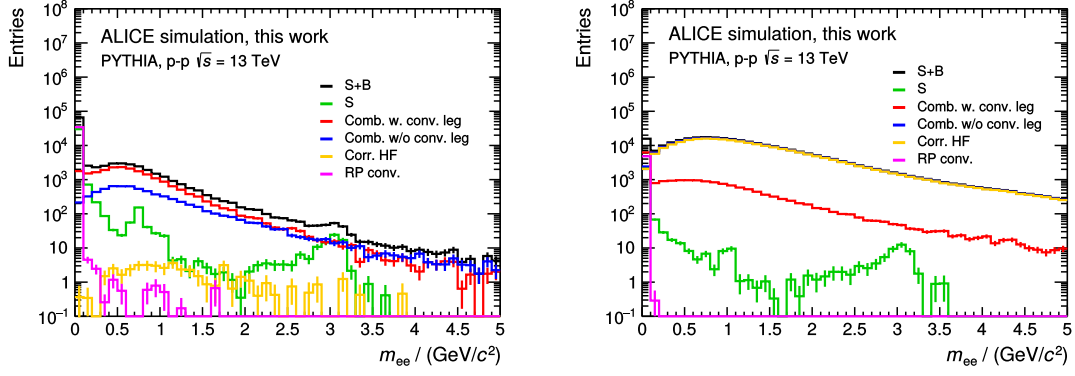


FIGURE 4.1: Dielectron invariant mass spectrum for simulated pp collisions at a center-of-mass energy $\sqrt{s} = 13$ TeV for the general-purpose Monte Carlo production (left) and the heavy flavor enhanced production (right). At the basis of the simulations, the event generator PYTHIA has been used. The overall yield (black) mainly consists of combinatorial pairs that either do (red) or do not (blue) contain conversion tracks, but also of non-combinatorial pairs with (magenta) and without (green) conversion tracks. Furthermore, the contribution from correlated heavy flavor decays (yellow) is shown.

For each of these scenarios, the following two aspects of the underlying data are investigated in particular:

- **Kinematic features:** Classifiers which are trained on data that contains kinematic features⁵ are usually prone to obtain a *mass bias* (*c.f.*, Sec. 3.4.2), *i.e.*, they might focus too preferentially on specific mass regions and thus fail their actual classification task on the entire mass domain. In order to investigate this behavior, classifiers are separately trained on data that contains kinematic features and on data that does not contain these features for each of the above analysis scenarios.
- **DCA features:** An important track variable is the so-called “distance of closest approach” (DCA), *i.e.*, the shortest distance between a reconstructed track and the primary vertex. Electrons and positrons originating from HF hadrons (*i.e.*, delayed decays) are expected to have a larger DCA on average, compared to tracks from light-flavor hadrons or from J/Ψ (*i.e.*, prompt decays). The reason for this is that delayed decays involve mesons (like D^0/\bar{D}^0) that produce electrons and positrons via the weak interaction and thus further away from the primary vertex. A schematic depiction of these decay scenarios is shown in Fig. 4.2. Especially in the case of multivariate HF rejection, the performance of a classifier is expected to heavily depend on the nature and quality of the provided DCA input features. A question of interest is whether classifiers are able to extract all the information required for the HF discrimination task from “low-level” features like the DCA vector components of each track⁶, or whether one can gain in classifier performance by manually

⁵See Tab. B.1 and Tab. B.2 in the appendix for a listing of the kinematic features in the used datasets.

⁶That is, x , y and z components of the DCA vector of each track in a pair.

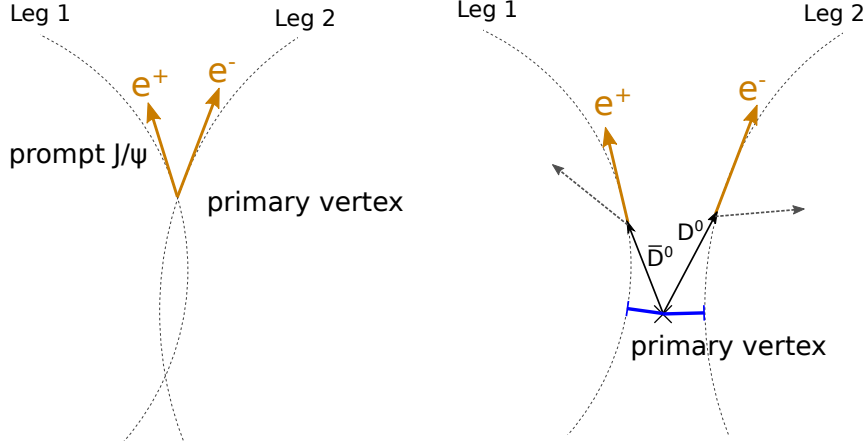


FIGURE 4.2: *Visualization of prompt and delayed decays. Prompt decays (left) from light-flavor hadrons or J/Ψ have vanishing values of DCA, whereas delayed decays (right) from heavy-flavor hadrons usually result in larger DCA values.*

constructing DCA features. In order to approach this question, three use cases for the above analysis scenarios 1. and 2. (see Definition 4.1) are investigated: classifier training on data with (1) DCA vector components as DCA features only, (2) manually engineered DCA features only and (3) both DCA vector components and engineered DCA features.

Considering all combinations of analysis scenarios, kinematic and DCA features cases, the number of performed analysis adds up⁷ to 14. Additionally, a classifier is trained on data containing DCA vector components (but no manually engineered DCA features) and all kinematic variables except the invariant mass. Altogether, this amounts to 15 different classifiers that are trained for the purpose of multivariate HF rejection.

In each of these 15 cases, the input variables are subjected to *standard scaling* and the used classifiers are neural networks similar to those implemented in Chap. 3. For detailed specifications of the individual network implementations, see Tab. F.1 in the appendix. The performances of all trained classifiers are presented in the following Sec. 4.5.

The optimization of the respective classifiers' working points is performed analogously to the methods described in Sec. 3.4.2: In the MVA output distribution that is obtained using the training data, the underlying signal and background distributions are identified via MC information. Subsequently, a scan over the MVA output variable is performed and the significance of the signal-background separation at each MVA cut value is computed. Finally, the optimized working point is defined as belonging to the MVA cut that maximizes this significance.

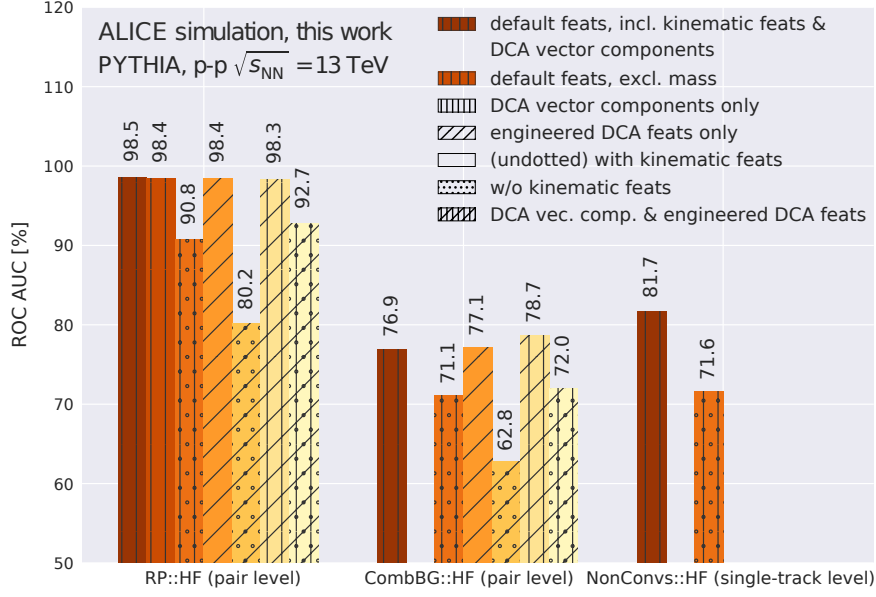


FIGURE 4.3: *ROC AUC comparison of the different HF-rejecting classifiers. See the text for a detailed description.*

4.5 Results

The performances of the 15 HF-rejecting classifiers are presented in terms of ROC AUC in Fig. 4.3. They are grouped according to the respective analysis scenarios, where the two pair-level classifiers are called “RP::HF” (corresponding to the MVA training of real-pair non-conversions vs. HF pairs) and “CombBG::HF” (corresponding to combinatorial non-conversion pairs vs. HF pairs). The third group (“NonConvs::HF”) belongs to the classifiers that have been trained on individual tracks to differentiate between non-conversion tracks and HF tracks. The different variations of the training procedure regarding kinematic and DCA features are encoded in the face patterns of the ROC AUC bars in Fig. 4.3. The baseline performances are given by the leftmost bars of each group, which correspond to the “default” features, *i. e.*, all features described in the respective tables in App. B. These features include the individual DCA vector components, but no manually engineered DCA features, which is indicated by the vertical lines on the bars. In contrast, diagonal lines correspond to scenarios with custom-engineered⁸ DCA features, but no DCA vector components. In those cases where both the DCA vector components and the DCA features are available during MVA training, vertical and diagonal lines are superimposed. Bars with dotted face patterns belong to classifiers that were trained on data without kinematic features, whereas undotted bars indi-

⁷Two different pair analyses \times two different scenarios regarding kinematic features \times three different scenarios regarding DCA features + single-track analyses with and without kinematic features = 14 classifiers.

⁸These custom DCA features include the angle between the two tracks’ DCA vectors, the difference between those vectors and its absolute value as well as the component-wise reciprocal values of the difference vector.

cate a training scenario with kinematic features. As Fig. 4.3 shows, no DCA feature sets variations were performed for the single-track classifiers.

There are some general patterns that can be seen across all analyses. For example, the best overall classifier performances are achieved for those scenarios that involve kinematic feature during training, almost regardless of the nature of the used DCA features. This also applies to the classifier that has been trained on “default” features, but has not seen the pair invariant mass as an training input variable.⁹ Only in the case of no kinematic features, there is a strong performance dependence on the used set of DCA features. When providing both the DCA vector components and the manually-engineered DCA features during training, the classifiers gain a performance boost relative to the respective cases where only DCA vector components are used. In some cases, however, this boost is relatively small, which means that most of the information that is necessary for the signal-background discrimination task is retained by the neural networks already in the case of “DCA vector components only”. Another notable observation is the fact that the training on individual tracks yields performances that are similar to (or even better than) the ones for pair-level training in the case of combinatorial background vs. HF pairs.

In the remainder of this section, two classifiers of each group will be discussed in further detail. For reasons of comparability among training scenarios, the chosen classifiers are the ones corresponding to training data with “default” features and the ones trained on data without kinematic features¹⁰.

4.5.1 Non-combinatorial vs. heavy flavor pairs

In analogy to the multivariate conversion rejection studies (*c.f.*, Chap. 3), mass-dependent MVA cut efficiency plots are created for all classifiers in this chapter. For each bin of invariant mass, these plots show the yield of a given contribution after a particular MVA cut has been applied divided by the total yield of the given contribution in the same bin. Figure 4.4 shows the MVA cut efficiencies for the classifier that has been trained to separate non-combinatorial, non-conversion pairs from heavy flavor pairs (“RP::HF”) for data with “default” features (*i.e.*, all features that are described in Tab. B.2 in the appendix) and for data without kinematic features.

The classifier corresponding to Fig. 4.4 (left) clearly indicates a significant mass bias since it rejects pairs with $m_{ee} \gtrsim 0.4 \text{ GeV}/c^2$ entirely. For this reason, despite its good overall performance (ROC AUC = 0.985), this classifier is strongly disfavored for the task of HF rejection.

By excluding kinematic features from the classifier training, the classifier generally performs worse (ROC AUC = 0.908) but seems to be less mass-biased, which is apparent by the non-vanishing values of MVA cut efficiencies over the entire mass range in Fig. 4.4 (right). However, there is still a large difference between the classifier’s HF rejection capability in the first mass bin and in all other bins, where the difference between HF and non-HF efficiencies is rather small.

⁹There are several kinematic variables that strongly correlate with invariant mass (*e.g.*, the opening angle θ), which can explain the observed independence of the mass as input feature.

¹⁰These scenarios are the only ones in the case of single-track training (*c.f.*, Fig. 4.3).

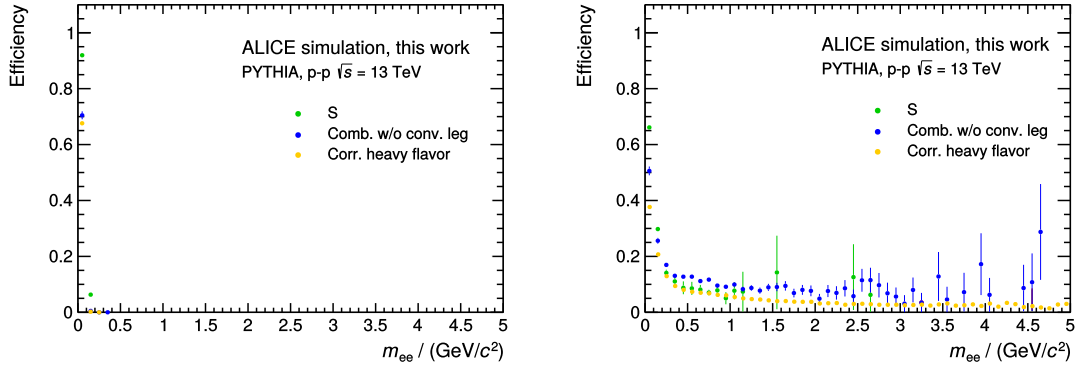


FIGURE 4.4: *MVA cut efficiencies for the different non-conversion dielectron sources for the “RP::HF” classifier trained on data with “default” features (left) and on data without kinematic features (right).*

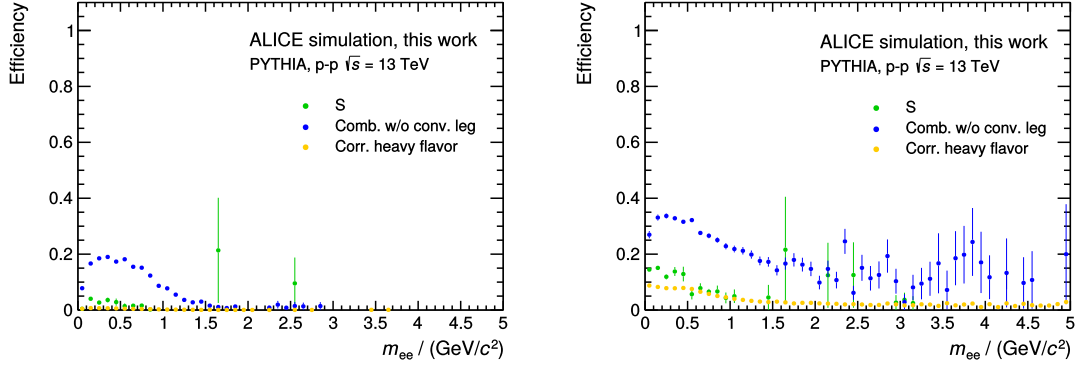


FIGURE 4.5: *MVA cut efficiencies for the different non-conversion dielectron sources for the “CombBG::HF” classifier trained on data with “default” features (left) and on data without kinematic features (right).*

4.5.2 Combinatorial vs. heavy flavor pairs

Analogously to Sec. 4.5.1, MVA cut efficiency plots are generated for the classification task of combinatorial, non-conversion pairs vs. HF pairs (“CombBG::HF”). The classifier which has been trained with “default” features (see Tab. B.2) corresponds to the efficiencies shown in Fig. 4.5 (left); the one trained without kinematic features corresponds to Fig. 4.5 (right). Compared to the efficiencies in Sec. 4.5.1, the classifiers achieve worse general performance (*i. e.*, ROC AUCs of 76.9 % and 71.1 %, respectively), but much better signal–background separation over a wider range of invariant mass. However, while HF pairs are convincingly suppressed with respect to combinatorial non-conversions, the separation between HF and non-combinatorial, non-conversion pairs is quite poor.

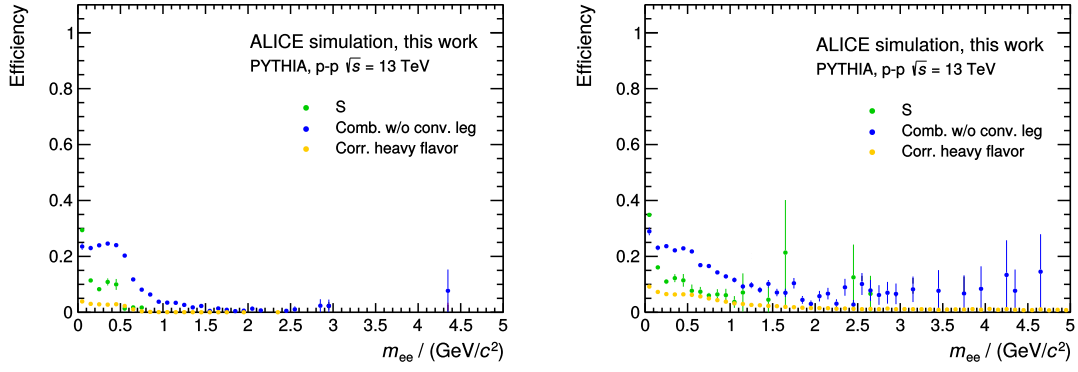


FIGURE 4.6: *MVA cut efficiencies for the different non-conversion dielectron sources for the “NonConvs::HF” classifier trained on data with “default” features (left) and on data without kinematic features (right).*

4.5.3 Non-conversion vs. heavy flavor tracks

In the case of HF rejection on the single-track level (“NonConvs::HF”), classifiers are trained on individual tracks and subsequently, when creating pairs from these tracks, the MVA outputs of the respective classifiers are consistently stored as pair variables. At classification time, an e^+e^- pair is rejected as background if at least one of its tracks has been labeled as an HF track by a given MVA algorithm. The resulting efficiency plots are given in Fig. 4.6 (left) for the classifier trained on data with “default” features (see Tab. B.1) and in Fig. 4.6 (right) for the data without kinematic features.

Apparently, the efficiency distributions in Fig. 4.6 show a peculiar behavior. The classifiers seem to treat combinatorial and non-combinatorial pairs¹¹ differently. At first glance, this behavior is unexpected for classifiers that have only been trained on individual tracks as they should not be able to know a difference between those two kinds of pairs. However, plotting these two contributions as functions of m_{ee} and transverse pair momentum $p_{T,\text{pair}} = [(p_{x,1} + p_{x,2})^2 + (p_{y,1} + p_{y,2})^2]^{1/2}$ before and after MVA cut application, for both cases of “default” features and no kinematic features, helps to estimate the underlying reason for this unexpected classifier behavior. See Fig. 4.7–4.9 for the corresponding results and Sec. 4.6 for their discussion.

4.6 Discussion

One of the major observations in the studies to suppress the HF content in a multivariate way is the realization that the training of classifiers is not straight forward. This is due to the small amounts of HF pairs relative to other dielectron sources in the invariant mass spectrum. Generally speaking, when training Machine Learning algorithms on extremely imbalanced data (*i. e.*, in this case, data with a vanishing number of background samples compared to signal samples), they usually learn to

¹¹Needless to say, these pairs do not contain conversion tracks and, in contrast to the conversion rejection studies in Chap. 3, it is not necessary to differentiate between combinatorial pairs with and without conversion tracks.

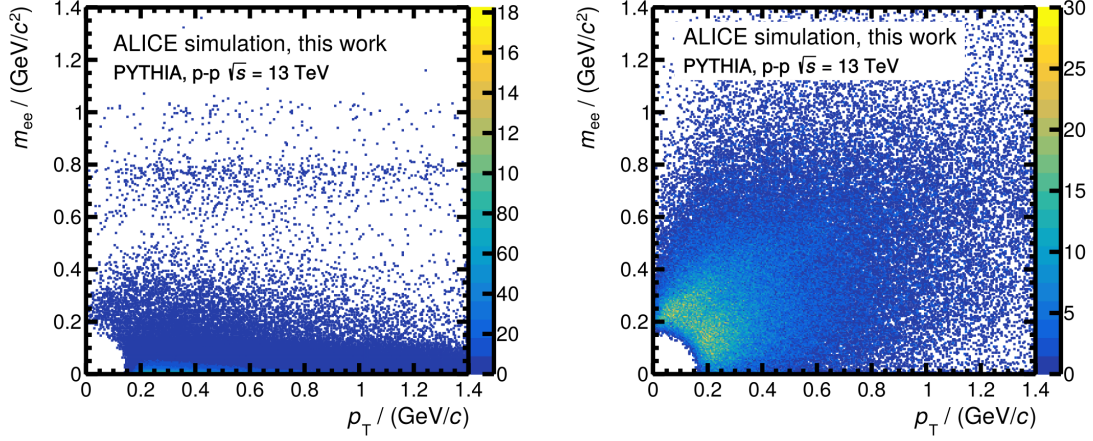


FIGURE 4.7: *Invariant mass vs. transverse pair momentum of real (left) and combinatorial non-conversion pairs (right) for pp collisions at $\sqrt{s} = 13$ TeV in the general-purpose MC production.*

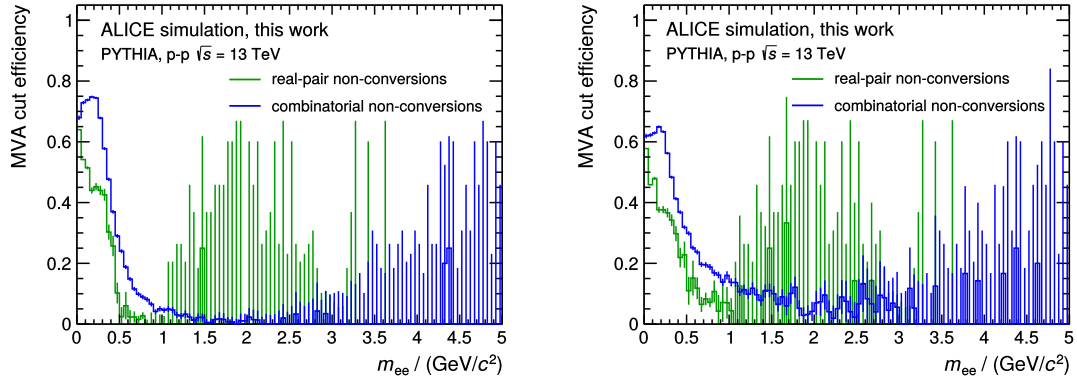


FIGURE 4.8: *MVA cut efficiencies for real (green) and combinatorial (blue) non-conversion pairs in the general-purpose MC production as functions of invariant mass, obtained from classifiers that have been trained on data with “default” features (left) and on data without kinematic features (right).*

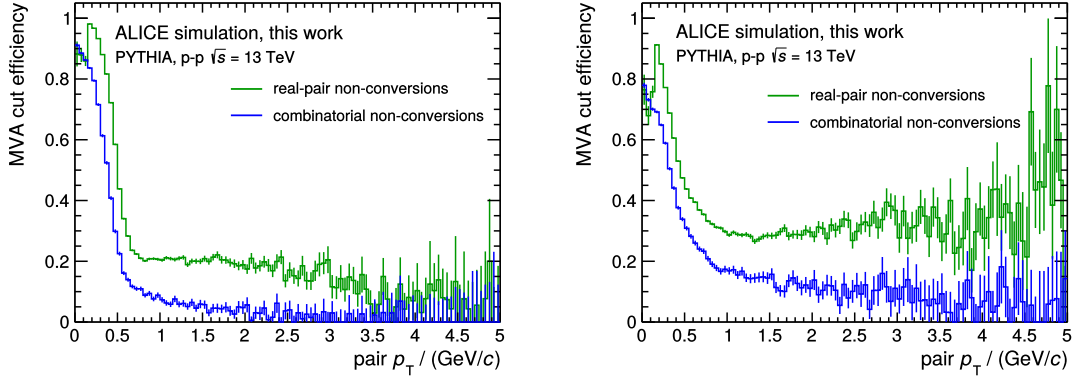


FIGURE 4.9: *MVA cut efficiencies for real (green) and combinatorial (blue) non-conversion pairs in the general-purpose MC production as functions of pair- p_T , obtained from classifiers that have been trained on data with “default” features (left) and on data without kinematic features (right).*

ignore the minority class altogether in order to maximize their classification scores. One way of compensating for the small amount of HF tracks in the data is to additionally utilize HF-enhanced datasets. As it has been shown in Sec. 4.4, this makes it possible for the classifiers to actually learn the signal–background discrimination task.

The performances of all trained classifiers were reported in terms of their ROC AUCs in Fig. 4.3. As a general trend, one can see clear performance differences between the classifiers trained to separate non-combinatorial pairs¹² from HF pairs (“RP::HF”) and the other two training approaches. The overall superiority of the “RP::HF” classifiers in terms of ROC AUC can be explained by the observation that these classifiers mainly concentrate on the first bins of invariant mass, where the yield of signal samples is very large and the one of background samples very small (*c.f.*, Fig. 4.1). This statement is supported by the MVA cut efficiencies in Fig. 4.4, which show that the background is convincingly rejected in the first mass bins only. The situation seems to be improved over a wider range of invariant mass in the case of the classifiers that discriminate combinatorial pairs against HF pairs (see Fig. 4.5): Despite these classifiers reporting worse overall performance (which arguably is due to the fact that their background rejection capabilities in the first mass bins is worse compared to the “RP::HF” classifiers), they show better signal–background separation potential over the entire mass range. Similar statements can be made for the classifiers trained to reject HF on the single-track level (“Non-Convs::HF”), whose cut efficiencies are shown in Fig. 4.6. Generally speaking, based on these considerations, it can be seen that ROC AUC alone is not necessarily a good indicator of overall “usefulness” of a classifier in the case of multivariate HF rejection.

For each of those three classification tasks, there are some general trends among the different training scenarios, too. On the one hand, it appears that the exact nature of the DCA features does not produce large performance differences as long

¹²Again, all pairs that contain conversion tracks are ignored during training, thus obviating the need to distinguish between pairs with conversions and ones without conversions.

as the training involves kinematic features. Put simply, this can be explained by the classifiers focusing entirely on these kinematic features, while basically ignoring many of the other features during training. On the other hand, once kinematic features are excluded from the training, the differences between the respective DCA feature implementations become apparent. The use of “low-level” information like the DCA vector components always results in better performance compared to the use case of manually-engineered DCA features. However, when making both DCA vector components and custom DCA features available during training, the classifiers only minimally improve over their performance in the case of only DCA vector components. From this, it can be concluded that most of the necessary information is already contained in the “low-level” features like the DCA vector components. The fact that deep neural networks are able to extract useful information from rudimentary input features, thus obviating the need for manual feature engineering (at least to some degree), is well-known and has been studied for the high-energy physics case, *e. g.*, in [85].

Once more, the problem of mass bias should be discussed here. As the studies in this chapter have shown, it is particularly difficult not to obtain classifiers that are mass-biased. In the case of training on e^+e^- pairs, where different mass bins show strongly varying yields of different dielectron sources, the fact that classifiers easily obtain such a bias does not seem too surprising. However, particular care should be also taken regarding the classifiers that have been trained on individual tracks, which are expected to be much less prone to such biases. As Fig. 4.6 shows, these classifiers seem to make a difference between combinatorial and non-combinatorial pairs, which is manifested in the different MVA cut efficiencies of these dielectron sources. Looking at these MVA cut efficiencies as functions of invariant mass and $p_{T,\text{pair}}$, one can see that even the classifier cuts are applied differently in those two variables. Figure 4.7 shows that real and combinatorial pairs populate different regions in the subspace of m_{ee} and $p_{T,\text{pair}}$. This difference should be unknown to classifiers that are unbiased regarding kinematic variables like these. However, as Fig. 4.8 and Fig. 4.9 demonstrate, the single-track classifier indeed shows a different cut behavior in these variables for the discussed dielectron sources. This observation is true, regardless of the inclusion (or exclusion) of kinematic features during training. After all, despite all efforts, even the single-track classifiers seem to learn mass-dependent decision criteria. This undesired property has to be addressed in future analyses.

Another aspect of the analyses that needs improvement is the optimization of the working point of a classifier. A working point that has been optimized on training data with an artificially enhanced amount of HF tracks is not necessarily optimal in the case of realistic HF content. In principle, this can be compensated for by scaling the background distributions in the classifiers’ MVA outputs to the HF content of the GP production before the working point optimization. However, in this work, the factor between the HF content in the GP and in the HF enhanced production is about 5×10^{-5} . When applying such extreme scaling to the respective MVA outputs and computing the optimized working points, the obtained classifiers always turn out to be ones that do not apply cuts at all. This is due to the vanishing amount of HF pairs in this case, which yields maximum classification scores when simply labeling all samples as belonging to the signal class. It should be noted, however, that this scaling of MVA output distributions of a given class does not affect the

ROC AUCs of the classifiers and that the validity of their reported performances therefore still holds.

Due to the artificially enhanced HF contents in the training data and the fact that the scaling approach during working point optimization does not yield meaningful classifiers, it is difficult to compare the results of this work to existing analyses. Plausible cut efficiencies for the rejection of HF as compared to the ALICE upgrade Letter of Intent [13] are 2 % for HF and 10 % for signal. Comparing the results in this section (see, *e. g.*, Fig. 4.6), one might be tempted to reason that some of them are already on the same scale regarding cut efficiencies. However, direct comparisons to existing studies are difficult at this stage and one has to think of a better approach to MVA training in further investigations.

On a more positive note, the HF studies in this work suggest that one could save time by omitting manual DCA feature engineering, as deep neural networks are able to extract almost all the necessary information already from low-level features like the DCA vector components. Furthermore, classifier training on individual tracks, again, is not inferior to variants of pair-level training, which is also one of the findings in the multivariate conversions rejection study (see Chap. 3). Since single-track MVA is still less prone to obtaining mass-biased classifiers and might be considered to be conceptually simpler, this is an encouraging result as one could pursue analysis approaches based on individual tracks in the future without sacrificing classification performance.

Chapter 5

Summary and outlook

The phase diagram of hadronic matter, *i. e.*, matter that is made of composite states of quarks and gluons, is yet to be fully understood and experimentally probed. Despite knowing the dynamics and interactions of strongly interacting particles at a fundamental level by means of QCD (*i. e.*, *Quantum Chromodynamics*, the gauge field theory of the strong interaction), it has been proven difficult to make precise theoretical predictions on the exact nature of nuclear matter states at high temperatures and/or densities, as well as the corresponding phase transitions from “cold” hadronic matter to “deconfined” matter. At high temperatures and/or densities, a new state of matter is predicted to exist – the so-called Quark-Gluon Plasma (QGP). In the laboratory, it is expected that this state can be created in collisions of ultra-relativistic heavy-ion nuclei. The ALICE experiment at CERN–LHC is designed to study hadronic matter under extreme conditions via the collision of ultra-relativistic lead ions. Different ways to probe the hot and dense medium that is created in such collisions exist. One of these probes is electromagnetic radiation in form of e^+e^- and $\mu^+\mu^-$ pairs, which are able to provide information about the conditions inside the medium as these particles do not interact strongly and, thus, do not experience significant in-medium modifications of their final states. Furthermore, these probes are created during all stages of a collision and can therefore give insight into the entire evolution of the hot “fireball”. This work is particularly concerned with the study of low-mass dielectrons (*i. e.*, electron-positron pairs) by means of the reconstruction of their invariant mass spectrum. The major background sources in this spectrum are dielectrons from photon conversions and combinatorial pairs, and one has to deal with signal-to-background ratios as small as 10^{-3} . To some extent, the dielectron spectrum also contains e^+e^- pairs from heavy flavor (HF) decays. Since these decays are not well understood in heavy-ion collisions in terms of their cross-sections and invariant mass shape, a viable analysis approach is to treat e^+e^- from HF decays as another physical background source. In order to suppress all these background components in the dielectron spectrum, sophisticated analysis techniques are required.

This work studied the potential of multivariate analysis (MVA) methods to reject different background components of the dielectron invariant mass spectrum. It consisted of two major parts, *i. e.*, the multivariate rejection of conversion pairs (see Chap. 3) and of pairs from HF decays (see Chap. 4).

For the conversion rejection studies, several multivariate classifiers in form of neural networks were trained for different analysis scenarios (*e. g.*, training on pairs or on

individual tracks to reject tracks from photon conversions). Their performances were compared to a conventional analysis approach that is based on a bivariate *prefilter*¹ cut. It has been shown that MVA methods can improve the signal-to-background ratio by up to 60 % (with a significance gain of about 15 %) in the case of pure conversion rejection and by up to 30 % (with a significance gain of about 15 %) in the case of rejection of conversions as well as all combinatorial pairs. Furthermore, it turned out that the classifier trained on individual tracks are *not* inferior to ones trained on pairs in terms of MVA cut efficiencies and observables like S/B . This can be considered an encouraging result since classifiers, which have been trained on individual tracks, are less prone to obtaining mass biases. Moreover, one might consider the processing and handling of single tracks to be conceptually simpler, compared to pairs of tracks.

The classifier training for HF-pair rejection did not turn out to be straight forward. The major problem in this case is the vanishing amount of HF tracks compared to other dielectron sources in the data. In order to compensate for this, an additional data sample with enhanced HF content was used to make classifier training possible. Ultimately, 15 different MVA approaches for HF-pair rejection were carried out. They can be grouped into three categories, according to the signal and background definitions in their respective classification tasks: (1) “RP::HF” \leftrightarrow non-combinatorial, non-conversion pairs vs. HF pairs, (2) “CombBG::HF” \leftrightarrow combinatorial, non-conversion pairs vs. HF pairs and (3) “NonConv::HF” \leftrightarrow non-conversion *tracks* vs. HF *tracks*. It has been observed that the best-performing classifiers (*i. e.*, the “RP::HF” ones) obtain significant mass biases, which makes them impractical for the purposes of HF rejection over a wide range of invariant masses. Considering the other two classes of trained models, one again observes that single-track training is in no way inferior to pair-level training. This is an encouraging result for the same reasons as mentioned before. However, investigating the single-track classifiers in more detail, even these ones turn out to be mass biased to some degree, as they perform cuts differently for combinatorial and non-combinatorial pairs in the invariant mass–transverse pair momentum subspace (in which the populations of these pairs are indeed different).

In general, the HF rejection classifiers indicated a strong dependence of their performances on kinematic input variables: If these variables are present during training, the classifiers tend to almost exclusively focus on them. If they are not present, other variables become important, *e. g.*, features involving DCA (“distance of closest approach”, the minimum distance between a track and the primary vertex). For these DCA features, it has been observed that, using sufficiently deep neural networks, most of the information needed for the signal–background discrimination task can be obtained by merely using DCA vector components as DCA features, thus obviating the need for manual DCA feature engineering.

In general, this work has shown that multivariate HF rejection works in principle. Particularly, some approaches turned out to be less prone to mass biases than others, low-level features like DCA vector components seemed to suffice as input variables and the classifier training on individual tracks turned out to be a reasonable choice

¹In the bivariate *prefiltering* method, non-combinatorial conversion pairs are identified via rectangular cuts in a two-dimensional variable subspace at low masses. Subsequently, this information is used to also reject combinatorial pairs at higher masses which contain at least one of those conversion-tagged tracks. See Sec. 3.4.1 for further details.

of training strategy. However, several issues have to be dealt with and further investigated in future analyses. First, even the single-track classifiers seem to obtain some form of undesired mass bias. Second, the chosen working points in this work are not necessarily optimal for the use case with a realistic amount of HF tracks in the data. The usual approach of scaling the signal and background distributions in the MVA output before working point optimization does however not work, which is again due to the vanishing amount of HF in realistic data. A better solution (probably in form of a different training strategy) has to be found in future studies.

In conclusion, MVA techniques have proven themselves to be a promising approach to low-mass dielectron analyses in ALICE as they are able to improve existing results considerably. Furthermore, the use of such techniques seems to take away a lot of complexity from the conventional analyses tasks. For example, deep neural networks allow to transition from pair-level analysis to analysis on individual tracks without significant performance loss. Moreover, due to their enhanced performances, multivariate approaches render some of the more complicated and numerically involved background rejection techniques (*e. g.*, *prefiltering*) obsolete, thus further simplifying dielectron analyses.

Appendices

Appendix A

Track cuts

In this section, the different sets of track cuts which are used throughout this work are presented.

A.1 Default track cuts

AliESDtrackCuts function	Value	Comment
accept kink daughters	false	reject tracks with a kink
require sigma to vertex	false	no sigma cut to vertex
DCA to vertex 2D	true	cut on the quadratic sum of distance of closest approach (DCA) in xy and z direction
max DCA to vertex Z	3 cm	maximum DCA to the primary vertex in longitudinal direction
max DCA to vertex XY	1 cm	maximum DCA in transverse direction
require ITS refit	true	require ITS refit
cluster requirement ITS	ESD track cuts: kSPD, kFirst	require at least one hit in the SPD
min N clusters ITS	3	minimum number of clusters in the ITS
require TPC refit	true	require TPC refit
min N clusters TPC	70	minimum number of cluster in the TPC
max χ^2 per cluster TPC	3.5	maximum χ^2 per TPC cluster in the first iteration

TABLE A.1: *Default track cuts for the conversion rejection studies in this work.*

A.2 Loose track cuts

AliESDtrackCuts function	Value	Comment
accept kink daughters	false	reject tracks with a kink
require sigma to vertex	false	no sigma cut to vertex
DCA to vertex 2D	true	cut on the quadratic sum of DCA in xy and z direction
max DCA to vertex Z	6 cm	maximum distance of closest approach (DCA) to the primary vertex in longitudinal direction
max DCA to vertex XY	2 cm	maximum DCA in transverse direction
require ITS refit	true	require ITS refit
min N clusters ITS	3	minimum number of clusters in the ITS
require TPC refit	true	require TPC refit
min N clusters TPC	50	minimum number of cluster in the TPC
max χ^2 per cluster TPC	3.5	maximum χ^2 per TPC cluster in the first iteration

TABLE A.2: *Loose track cuts for the conversion rejection studies in this work.*

A.3 Minimal track cuts

AliESDtrackCuts function	Value	Comment
require ITS refit	true	require ITS refit
cluster requirement ITS	ESD track cuts: kSPD, kFirst	require at least one hit in the SPD
min N clusters TPC	4	minimum number of cluster in the TPC

TABLE A.3: *Minimal track cuts for the conversion rejection studies in this work.*

Appendix B

Input variables

This section lists the variables that are used in the analyses carried out in this work. These variables are categorized into single-track variables (see Tab. B.1) and pair variables (see Tab. B.2).

Name	Description	kin. feature?
event	event ID	
η	angle η	✓
ϕ	angle ϕ	✓
p_T	transverse momentum	✓
DCA_a	distance of closest approach (DCA) in direction a (with $a = x, y, z$)	
p_a	momentum in direction a (with $a = x, y, z$)	✓
nITS	number of clusters in the Inner Tracking System (ITS)	
nITSshared	number of shared ITS clusters	
nTPC	number of cluster in the Time Projection Chamber (TPC)	
χ^2_{ITS}	quality parameter for the fit of the reconstructed track in the ITS	
χ^2_{TPC}	quality parameter for the fit of the reconstructed track in the TPC	

TABLE B.1: *Relevant single-track variables used for the analyses. Kinematic variables are indicated by a checkmark symbol in the last column.*

Name	Description	kin. feature?
EventID1, EventID2	event ID of the individual tracks in a pair	
IsRP	1 if pair is a <i>real</i> (<i>i. e.</i> , non-combinatorial) pair, 0 otherwise	
IsConv	1 if pair contains at least one track from a photon conversion, 0 otherwise	
IsCorrHF	1 if pair comes from the same heavy-flavor quark pair ($c\bar{c}$ or $b\bar{b}$), 0 otherwise	
IsCombHF	1 if pair comes from a charm or bottom quark and is combinatorial, 0 otherwise	
opang	opening angle θ of the pair's tracks	✓
diffz	$\text{diffz} = \arccos([\vec{p}_1 / \vec{p}_1 - \vec{p}_2/ \vec{p}_2] \cdot \hat{z})$	✓
sumz	$\text{sumz} = \arccos([\vec{p}_1 + \vec{p}_2 \cdot \hat{z})$	✓
mass	invariant mass of the pair	✓
phiv	value of φ_V of the pair	✓
$p_{a,i}$	momentum of track i in direction a (with $i = 1, 2$, $a = x, y, z$)	✓
PIDeff1, PIDeff2	particle identification efficiencies of the individual tracks in a pair	
$\text{DCA}_{a,i}$	distance of closest approach (DCA) of track i in direction a (with $i = 1, 2$, $a = x, y, z$)	
nITS1, nITS2	number of clusters in the Inner Tracking System (ITS) of the individual tracks in a pair	
nITSshared1, nITSshared2	number of shared ITS cluster of the individual tracks in a pair	
nTPC1, nTPC2	number of cluster in the Time Projection Chamber (TPC) of the individual tracks in a pair	
$\chi^2_{\text{ITS},i}$	quality parameter for the fit of the reconstructed track i in the ITS (with $i = 1, 2$)	
$\chi^2_{\text{TPC},i}$	quality parameter for the fit of the reconstructed track i in the TPC (with $i = 1, 2$)	
ϕ_i	angle ϕ for track i (with $i = 1, 2$)	✓
η_i	angle η for track i (with $i = 1, 2$)	✓
$p_{T,i}$	transverse momentum p_T for track i (with $i = 1, 2$)	✓

TABLE B.2: *Relevant pair variables used for the analyses. Kinematic variables are indicated by a checkmark symbol in the last column.*

Appendix C

Prefiltering algorithms

The following algorithms outline the basic working principles of the *prefiltering* methods used in the multivariate conversion rejection studies (see Sec. 3). Algorithm 1 describes the identification of conversion pairs via bivariate *prefiltering*, whereas Algorithm 2 the conversion pair tagging via the multivariate *prefilter*.

Algorithm 1 Identification of conversion pairs via bivariate *prefiltering*.

```
1: for all events do
2:   for all pairs in the current event do
3:     if pair meets  $\varphi_V$  and  $m_{ee}$  conditions then
4:       mark pair as conversion pair
5:     else
6:       mark pair as non-conversion pair
7:     end if
8:   end for
9:   for all pairs in the current event do
10:    rand  $\leftarrow$  random number between 0 and 1
11:    pairPIDeff  $\leftarrow$  get_pair_PID_efficiency(current pair)
12:    if pair shares a track with one of the previously marked conversion pairs
      and rand < pairPIDeff then
13:      mark pair as conversion pair
14:    end if
15:  end for
16: end for
```

Algorithm 2 Identification of conversion pairs via multivariate *prefiltering*.

```

1: for all events do
2:   for all pairs in the current event do
3:     if classifier identifies pair as conversion then
4:       mark pair as conversion pair
5:     else
6:       mark pair as non-conversion pair
7:     end if
8:   end for
9:   for all pairs in the current event do
10:    rand  $\leftarrow$  random number between 0 and 1
11:    pairPIDeff  $\leftarrow$  get_pair_PID_efficiency(current pair)
12:    if pair shares a track with one of the previously marked conversion pairs
      and rand < pairPIDeff then
13:      mark pair as conversion pair
14:    end if
15:  end for
16: end for

```

Appendix D

Conversion rejection studies assuming perfect PID (supplement)

All the different approaches for conversion rejection that are described in Sec. 3.4 use p_T -dependent particle identification (PID) efficiencies from pp collisions at $\sqrt{s} = 13$ TeV which were recorded by the ALICE detector with a low magnetic field ($B = 0.2$ T) in 2016 (see Fig. 3.1 on page 28). These PID efficiencies suppress e^+e^- pairs predominantly at small $p_T \lesssim 500$ MeV/c and thus a large amount of non-combinatorial conversion pairs. This, in turn, leads to a bad performance of some approaches, especially of the *prefilter* cut based methods (see Sec. 3.4.1 and Sec. 3.4.2).

One might wonder whether the comparison of the different conversion rejection approaches (see Sec. 3.5 and Sec. 3.6) still holds in the case of better PID capabilities. Therefore, this section discusses the most optimistic situation and explores the performances of the different methods in the limiting case of perfect PID (*i. e.*, an efficiency of 100 % and a contamination of 0 % for all values of p_T).

By analogy to the ROC curves in Fig. 3.13 (page 39), the corresponding ROC curves for the case of perfect PID are plotted in Fig. D.1. Although some of the analyses perform very differently compared to the case of realistic PID, the basic statements of Sec. 3.6 still hold:

- All multivariate approaches outperform the analyses that are based on conventional (*i. e.*, sequential) cuts.
- The classifier arrangement according to their respective performances (*i. e.*, ROC AUCs) is the same as before (*c. f.*, Tab. D.1).
- The optimized working points of *MVA 1* and *MVA 2* show similar values for true and false positive rates, meaning that they perform comparably well in terms of MVA cut efficiencies, significance and signal-over-background.
- Opening the track cuts for the *prefilter*, does not add much in terms of classifier performance. This can be explained by the fact that the “default” cuts (see Tab. A.1) are quite loose already.

However, some of the qualitative differences to the studies with realistic PID efficiencies are:

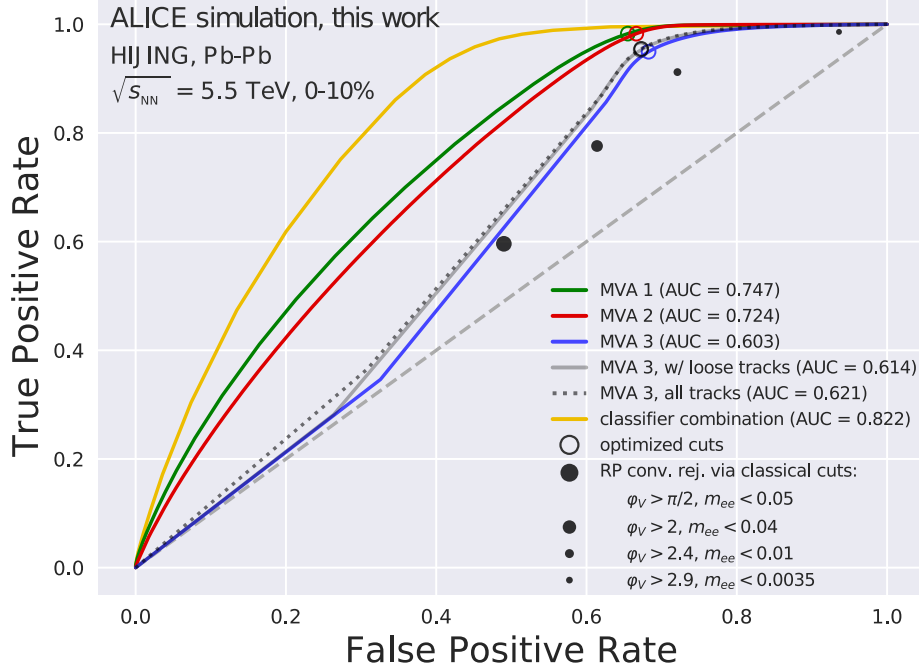


FIGURE D.1: *ROC curves of the different conversion rejection approaches for the limiting case of perfect PID. For an explanation of the pseudonyms in the legend, see Definition 3.2 on page 38.*

MVA approach	ROC AUC in %
MVA 1	74.7
MVA 2	72.4
MVA 3	60.3
MVA 3 (with loose tracks)	61.4
MVA 3 (all tracks)	62.1
MVA 1 + MVA 3	82.2

TABLE D.1: *ROC AUCs of the different MVA approaches in the case of perfect PID (conversion rejection).*

- The optimized working points of the MVA methods are much closer together in the ROC plane. With realistic PID, the distance between the *MVA 1*/*MVA 2* curves and any of the curves belonging to *prefilter* cut based analyses is much larger.
- By combining the classifiers of *MVA 1* and *MVA 3* one can reach a much higher performance of the combined classifier than before (*c. f.*, $\text{ROC AUC}_{(\text{realistic PID})} = 0.76$ vs. $\text{ROC AUC}_{(\text{perfect PID})} = 0.82$).

To summarize, using perfect PID significantly boosts the performances of the *prefilter*-based methods with respect to a scenario with realistic PID. However, the basic statements and the results of this work continue to be true. Multivariate approaches consistently show performances that are superior to any of the conventional conversion rejection methods.

Appendix E

Results for multivariate conversion rejection (supplement)

In Sec. 3.4, the different multivariate analysis (MVA) approaches are described and their results presented in Sec. 3.5 in terms of combined MVA cut efficiency plots. However, further and more detailed results plots are omitted there for reasons of clarity. This section complements the main part of this work by displaying the missing plots.

Figure E.1 shows the effects of the multivariate *prefilter* cuts on the dielectron invariant mass spectrum (*c.f.*, Sec. 3.4.2) as well as the corresponding cut efficiencies.

Equivalent plots that correspond to the multivariate conversion rejection on the single-track level (*c.f.*, Sec. 3.4.3) are presented in Fig. E.2.

Finally, Fig. E.3 shows the mass spectrum before and after the applied MVA cuts as well as the cut efficiencies for the multivariate approach of combinatorial conversion pair rejection (see Sec. 3.4.4).

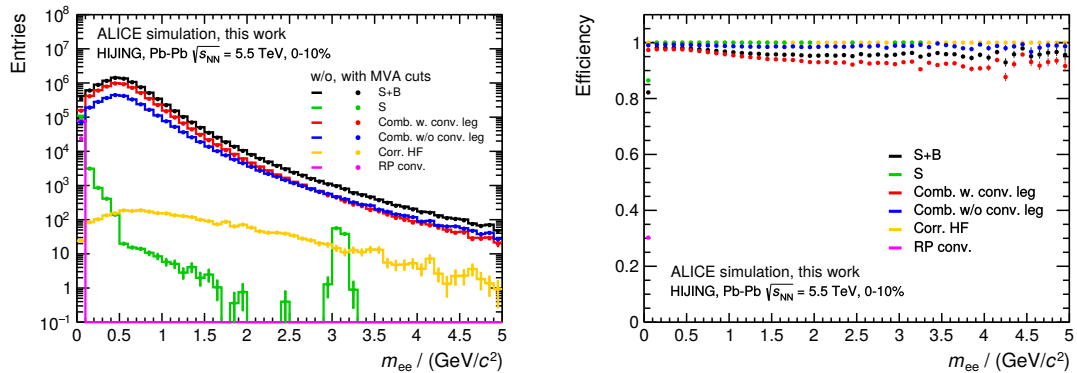


FIGURE E.1: *Left: Dielectron invariant mass spectrum before and after applied MVA cuts. Right: corresponding MVA cut efficiencies (conversion rejection via multivariate prefilter cut).*

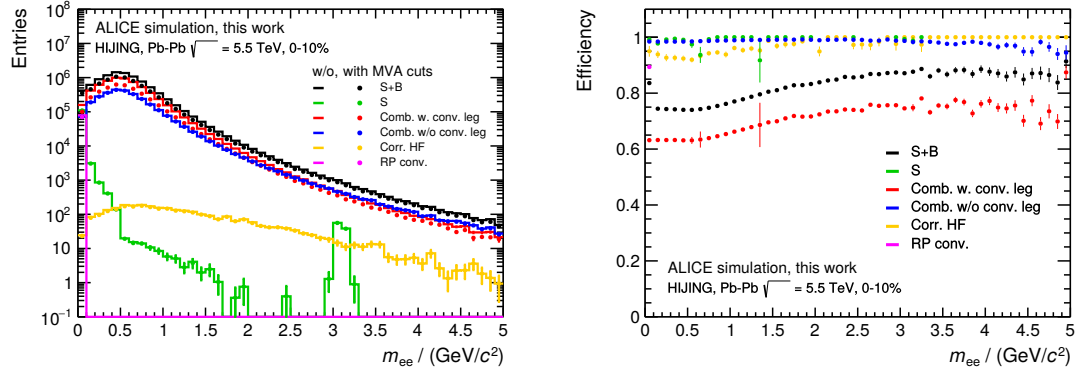


FIGURE E.2: *Left: Dielectron invariant mass spectrum before and after applied MVA cuts. Right: corresponding MVA cut efficiencies (multivariate conversion rejection on the single-track level).*

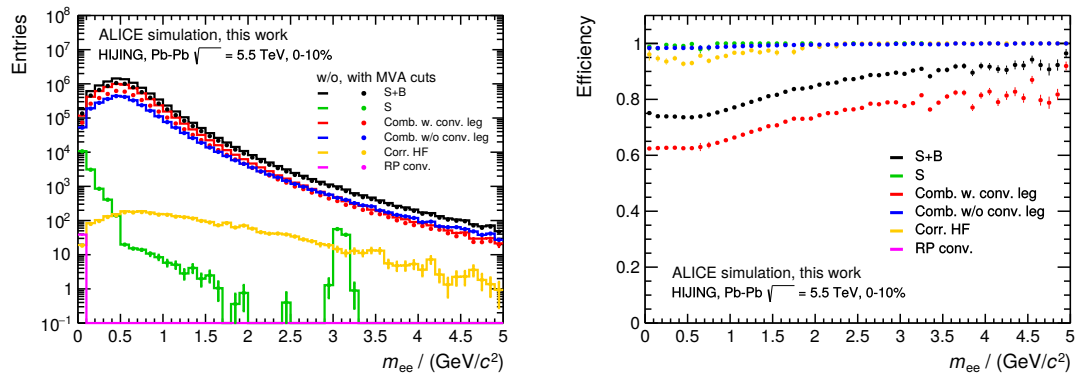


FIGURE E.3: *Left: Dielectron invariant mass spectrum before and after applied MVA cuts. Right: corresponding MVA cut efficiencies (B) (multivariate rejection of combinatorial conversion pairs).*

Appendix F

Parameters of heavy flavor rejection classifiers

For the purpose of multivariate heavy flavor (HF) rejection (see Chap. 4), several neural networks have been trained. All these classifiers consist of densely connected nodes. Except from the output layer, where the sigmoid function serves as activation, all these nodes employ Rectified Linear Units (ReLU) [75] as their activation functions. The kernel nodes are initialized via the Glorot/Xavier normal initialization scheme [99]. All networks define the binary cross-entropy (Eq. 2.4.4) as their loss functions and use Adam [101] as an optimizer. Table F.1 lists further specifications which differ between classifiers. Here, the first column describes the classification task and additionally gives the values for the respective ROC AUC scores. Together with Fig. 4.3 on page 52, this should help to unambiguously identify the respective classifiers.

Classification task (ROC AUC [%])	# layers	# nodes per layer	batch size	regularization
RP::HF (98.5)	4	50	512	Dropout (50%)
RP::HF, without mass as input variable (98.4)	4	50	512	Dropout (50%)
RP::HF (90.8)	4	50	512	Dropout (25%), batch normalization
RP::HF (98.4)	5	100	512	Dropout (20%)
RP::HF (80.2)	8	100	512	Dropout (20%)
RP::HF (98.3)	4	50	512	Dropout (50%)
RP::HF (92.7)	8	100	512	Dropout (20%)
CombBG::HF (76.9)	6	100	512	Dropout (25%), batch normalization
CombBG::HF (71.1)	6	100	1024	Dropout (25%), batch normalization
CombBG::HF (77.1)	6	100	512	Dropout (40%)
CombBG::HF (62.8)	7	100	512	Dropout (20%)
CombBG::HF (78.7)	6	100	512	Dropout (40%)
CombBG::HF (72.0)	4	100	512	Dropout (25%)
NonConvs::HF (81.7)	6	100	512	batch normalization
NonConvs::HF (71.6)	6	100	512	Dropout (5%)

TABLE F.1: *Parameter specifications for the HF-rejecting neural networks*

Bibliography

- [1] P. A. R. Ade et al. Planck 2015 results. XIII. Cosmological parameters. *Astronomy & Astrophysics*, 594:A13, 2016.
- [2] L. Álvarez Gaumé and M. Á. Vázquez-Mozo. An invitation to quantum field theory. *Lecture Notes in Physics*, 839, 2012.
- [3] H. Satz. *Extreme states of matter in strong interaction physics: An Introduction*, volume 841. Springer Science & Business Media, 2012.
- [4] M. Gell-Mann. Symmetries of baryons and mesons. *Physical Review*, 125(3):1067, 1962.
- [5] W. Heisenberg. Über den Bau der Atomkerne. *Zeits. f. Physik*, 77(1–2):1–11, 1932.
- [6] E. Noether. Invariante Variationsprobleme. *Nachr. d. König. Gesellsch. d. Wiss. zu Göttingen, Math-phys. Klasse (1918) 235–257*, page 57, 1918.
- [7] W. M. al Yao et al. Review of particle physics. *Journal of Physics G: Nuclear and Particle Physics*, 33(1):1, 2006.
- [8] C. T. H. Davies et al. High-precision lattice QCD confronts experiment. *Physical Review Letters*, 92(2):022001, 2004.
- [9] S. Dürr et al. Ab initio determination of light hadron masses. *Science*, 322(5905):1224–1227, 2008.
- [10] G. Altarelli. Experimental tests of perturbative QCD. *Annual Review of Nuclear and Particle Science*, 39(1):357–406, 1989.
- [11] K. G. Wilson. Confinement of quarks. *Physical Review D*, 10:2445–2459, Oct 1974.
- [12] M. Creutz. Monte Carlo study of quantized SU(2) gauge theory. *Physical Review D*, 21(8):2308, 1980.
- [13] B. Abelev et al. Upgrade of the ALICE Experiment: Letter of Intent. Technical Report CERN-LHCC-2012-012. LHCC-I-022. ALICE-UG-002, CERN, Aug 2012.
- [14] E.V. Shuryak. Quark-gluon plasma and hadronic production of leptons, photons and psions. *Physics Letters B*, 78(1):150 – 153, 1978. ISSN 0370-2693.

- [15] N. Cabibbo and G. Parisi. Exponential hadronic spectrum and quark liberation. *Physics Letters B*, 59(1):67–69, 1975.
- [16] P. Braun-Munzinger and J. Wambach. The Phase Diagram of Strongly-Interacting Matter. *Review of Modern Physics*, 81:1031–1050, 2009.
- [17] M. A. Stephanov. QCD phase diagram: an overview. *arXiv preprint hep-lat/0701002*, 2006.
- [18] B. Svetitsky and L. G. Yaffe. Critical behavior at finite-temperature confinement transitions. *Nuclear Physics B*, 210(4):423–447, 1982.
- [19] L. D. McLerran and B. Svetitsky. A Monte Carlo study of SU(2) Yang-Mills theory at finite temperature. *Physics Letters B*, 98(3):195–198, 1981.
- [20] L. D. McLerran and B. Svetitsky. Quark liberation at high temperature: a Monte Carlo study of SU(2) gauge theory. *Physical Review D*, 24(2):450, 1981.
- [21] J. Kuti, J. Polonyi, and K. Szlachányi. Monte Carlo study of SU(2) gauge theory at finite temperature. *Physics Letters B*, 98(3):199–204, 1981.
- [22] S. Sarkar, H. Satz, and B. Sinha. *The physics of the quark-gluon plasma: introductory lectures*, volume 785. Springer, 2009.
- [23] F. Karsch. Lattice QCD at high temperature and density. In *Lectures on quark matter*, pages 209–249. Springer, 2002.
- [24] A. Bazavov et al. Equation of state and QCD transition at finite temperature. *Physical Review D*, 80:014504, Jul 2009.
- [25] C. A. Dominguez and M. Loewe. Deconfinement and chiral-symmetry restoration at finite temperature. *Physics Letters B*, 233(1-2):201–204, 1989.
- [26] A. Barducci and others. Heuristic argument for coincidence or almost coincidence of deconfinement and chirality restoration in finite temperature QCD. *Physics Letters B*, 244(2):311–315, 1990.
- [27] A. Ayala et al. QCD phase diagram from finite energy sum rules. *Physical Review D*, 84(5):056004, 2011.
- [28] P. Castorina, K. Redlich, and H. Satz. The phase diagram of hadronic matter. *The European Physical Journal C-Particles and Fields*, 59(1):67–73, 2009.
- [29] W. Florkowski. *Phenomenology of ultra-relativistic heavy-ion collisions*. World Scientific Publishing Co Inc, 2010.
- [30] R. Hagedorn. Statistical thermodynamics of strong interactions at high energies. *Nuovo Cimento, Suppl.*, 3(CERN-TH-520):147–186, 1965.
- [31] J. Cleymans and H. Satz. Thermal hadron production in high energy heavy ion collisions. *Zeitschrift für Physik C Particles and Fields*, 57(1):135–147, 1993.

- [32] K. Redlich, J. Cleymans, H. Satz, and E. Suhonen. Hadronisation of quark-gluon plasma. *Nuclear Physics A*, 566:391–394, 1994.
- [33] P. Braun-Munzinger, J. Stachel, J. P. Wessels, and N. Xu. Thermal equilibration and expansion in nucleus-nucleus collisions at the AGS. *Physics Letters B*, 344(1):43–48, 1995.
- [34] F. Becattini. A thermodynamical approach to hadron production in e^+e^- collisions. *Zeitschrift für Physik C Particles and Fields*, 69(3):485–492, 1995.
- [35] L. D. Landau. On the multiparticle production in high-energy collisions. *Izvestiya Akademii Nauk SSR, Seriya Fizicheskaya*, 17:51–64, 1953.
- [36] J.-P. Blaizot and J.-Y. Ollitrault. Hydrodynamics of quark-gluon plasmas. In *Quark-Gluon Plasma*, pages 393–470. World Scientific, 1990.
- [37] P. F. Kolb, J. Sollfrank, and U. Heinz. Anisotropic transverse flow and the quark-hadron phase transition. *Physical Review C*, 62(5):054909, 2000.
- [38] S. Datta, F. Karsch, P. Petreczky, and I. Wetzorke. Behavior of charmonium systems after deconfinement. *Physical Review D*, 69(9):094507, 2004.
- [39] T. Matsui and H. Satz. J/ψ suppression by quark-gluon plasma formation. *Physics Letters B*, 178(4):416–422, 1986.
- [40] J. D. Bjorken. Energy loss of energetic partons in quark-gluon plasma: possible extinction of high p_T jets in hadron-hadron collisions. *Fermilab-Pub-82/59-THY*, 1982.
- [41] R. Baier, Y. L. Dokshitzer, S. Peigne, and D. Schiff. Induced gluon radiation in a QCD medium. *Physics Letters B*, 345(3):277–286, 1995.
- [42] B. G. Zakharov. Fully quantum treatment of the Landau-Pomeranchuk-Migdal effect in QED and QCD. *JETP letters*, 63(12):952–957, 1996.
- [43] E.V. Shuryak. Quantum chromodynamics and the theory of superdense matter. *Physics Reports*, 61(2):71–158, 1980.
- [44] K. Kajantie and H. I. Miettinen. Temperature measurement of quark-gluon plasma formed in high energy nucleus-nucleus collisions. *Zeitschrift für Physik C Particles and Fields*, 9(4):341–345, 1981.
- [45] R. Rapp and J. Wambach. Chiral symmetry restoration and dileptons in relativistic heavy-ion collisions. In *Advances in Nuclear Physics*, pages 1–205. Springer, 2002.
- [46] S. D. Drell and T.-M. Yan. Massive lepton-pair production in hadron-hadron collisions at high energies. *Physical Review Letters*, 25(5):316, 1970.
- [47] J. P. Ellis. TikZ-Feynman: Feynman diagrams with tikz. *Computer Physics Communications*, 210:103–123, 2017.
- [48] C.-Y. Wong. *Introduction to high-energy heavy-ion collisions*. World scientific, 1994.

- [49] R. D. Pisarski. Phenomenology of the chiral phase transition. *Physics Letters B*, 110(2):155–158, 1982.
- [50] G.-Q. Li, C. M. Ko, and G. E. Brown. Enhancement of low-mass dileptons in heavy ion collisions. *Physical Review Letters*, 75(22):4007, 1995.
- [51] G. E. Brown and M. Rho. Scaling effective Lagrangians in a dense medium. *Physical review letters*, 66(21):2720, 1991.
- [52] G. Chanfray, R. Rapp, and J. Wambach. Medium modifications of the rho meson at CERN super proton synchrotron energies (200 gev/nucleon). *Physical review letters*, 76(3):368, 1996.
- [53] The ALICE Collaboration et al. The ALICE experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):S08002, 2008.
- [54] The ALICE Collaboration et al. Performance of the ALICE experiment at the CERN LHC. *International Journal of Modern Physics A*, 2014.
- [55] G. Dellacasa et al. ALICE technical design report of the inner tracking system (ITS). Technical report, 1999.
- [56] The ALICE Collaboration et al. Technical design report of the time projection chamber. *CERN/LHCC*, 1(2000):375, 2000.
- [57] The ALICE Collaboration et al. Technical design report of the transition-radiation detector. 2001.
- [58] ALICE Time-Of-Flight system (TOF): Technical Design Report. 2000.
- [59] G. Dellacasa et al. ALICE technical design report of the photon spectrometer (PHOS). 1999.
- [60] P. Cortese et al. ALICE Electromagnetic Calorimeter Technical Design Report. (CERN-LHCC-2008-014. ALICE-TDR-14), Aug 2008.
- [61] The ALICE Collaboration et al. Technical design report of the high momentum particle identification detector. *CERN/LHCC*, 98:19, 1998.
- [62] The ALICE Collaboration et al. ALICE technical design report on forward detectors: FMD, T0 and V0. *CERN-LHCC-2004-025*, 2004.
- [63] G. Dellacasa et al. ALICE technical design report of the zero degree calorimeter (ZDC). *CERN-LHCC-99-5*, 1999.
- [64] The ALICE Collaboration et al. Technical design report for the upgrade of the ALICE inner tracking system. (CERN-LHCC-2013-024. ALICE-TDR-017), Nov 2013.
- [65] The ALICE Collaboration et al. Upgrade of the ALICE time projection chamber. (CERN-LHCC-2013-020. ALICE-TDR-016), Oct 2013.
- [66] O. Behnke, K. Kröninger, G. Schott, and T. Schörner-Sadenius. *Data analysis in high energy physics: a practical guide to statistical methods*. John Wiley & Sons, 2013.

- [67] Y. Ganin et al. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.
- [68] A. Rogozhnikov. Reweighting with boosted decision trees. In *Journal of Physics: Conference Series*, volume 762, page 012036. IOP Publishing, 2016.
- [69] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4768–4777, 2017.
- [70] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.
- [71] I. Narsky and F. C. Porter. *Statistical analysis techniques in particle physics*. Wiley-VCH, 2014.
- [72] V. Vapnik. *The nature of statistical learning theory*. Springer, 1995.
- [73] V. Vapnik. *Statistical learning theory*. Wiley, 1998.
- [74] V. Vapnik. The support vector method of function estimation. In *Nonlinear Modeling*, pages 55–85. Springer, 1998.
- [75] V. Nair and G. E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.
- [76] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of Machine Learning Research*, volume 30, 2013.
- [77] D. A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). *CoRR*, abs/1511.07289, 2015.
- [78] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks. *CoRR*, abs/1706.02515, 2017.
- [79] J. S. Bridle. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In *Advances in Neural Information Processing Systems 2*, pages 211–217. 1990.
- [80] A. N. Kolmogorov. The representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables. *Doklady Akademii Nauk SSSR*, 108(2):179–182, 1956.
- [81] R. Hecht-Nielsen et al. Theory of the backpropagation neural network. *Neural Networks*, 1(Supplement-1):445–448, 1988.
- [82] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

- [83] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.
- [84] G. Zhong, L.-N. Wang, X. Ling, and J. Dong. An overview on data representation learning: From traditional feature learning to recent deep learning. *The Journal of Finance and Data Science*, 2(4):265 – 278, 2016.
- [85] P. Baldi, P. Sadowski, and D. Whiteson. Searching for Exotic Particles in High-Energy Physics with Deep Learning. *Nature Communications*, 5:4308, 2014.
- [86] J. Schmidhuber. Deep learning in neural networks: An overview. *Computing Research Repository*, abs/1404.7828, 2014.
- [87] S. Agostinelli et al. Geant4 – a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250 – 303, 2003.
- [88] J. Allison et al. Geant4 developments and applications. *IEEE Transactions on Nuclear Science*, 53(1):270–278, 2006.
- [89] J. Allison et al. Recent developments in Geant4. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 835(Supplement C):186 – 225, 2016.
- [90] M. Gyulassy and X.-N. Wang. HIJING 1.0: A Monte Carlo program for parton and particle production in high energy hadronic and nuclear collisions. *Computer Physics Communications*, 83(2-3):307–331, 1994.
- [91] R. Brun and F. Rademakers. ROOT — an object oriented data analysis framework. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 389(1-2): 81–86, 1997.
- [92] R. Brun et al. Computing in ALICE. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 502(2-3):339–346, 2003.
- [93] F. Carminati and A. Morsch. Simulation in ALICE. *arXiv preprint physics/0306092*, 2003.
- [94] A. Morsch. ALICE simulation framework. Technical report, 1999.
- [95] F. Chollet et al. Keras. <https://github.com/keras-team/keras>, 2015.
- [96] M. Abadi et al. TensorFlow: A system for large-scale machine learning. In *Operating Systems Design and Implementation*, volume 16, pages 265–283, 2016.
- [97] T. Dahms. *Dilepton spectra in p+p and Au+Au collisions at RHIC*. PhD thesis, SUNY, Stony Brook, 2008.

- [98] A. Adare et al. Detailed measurement of the e^+e^- pair continuum in $p + p$ and Au+Au collisions at $\sqrt{s_{NN}} = 200$ GeV and implications for direct photon production. *Physical Reviews C*, 81:034911, 2010.
- [99] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feed-forward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.
- [100] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456, 2015.
- [101] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [102] G. Louppe, M. Kagan, and K. Cranmer. Learning to Pivot with Adversarial Networks. *arXiv preprint arXiv:1611.01046*, 2016.
- [103] N. Srivastava et al. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- [104] T. Sjöstrand et al. An introduction to PYTHIA 8.2. *Computer physics communications*, 191:159–177, 2015.