Department of Physics and Astronomy

Heidelberg University

Master thesis

in Physics

submitted by

Christian Hans-Jörg Sonnabend

born in Wiesbaden

2022

Neural network regression for particle identification with the ALICE TPC detector in Run 3

This Master thesis has been carried out by

Christian Hans-Jörg Sonnabend

at the

Physikalisches Institut der Universität Heidelberg

under the supervision of

Prof. Silvia Masciocchi

"The truth is rarely pure and never simple."
∼ Oscar Wilde

Abstract

The gaseous Time Projection Chamber (TPC) of the ALICE experiment at CERN serves some of the most crucial roles in many physics analyses within the collaboration and is responsible for 92.5% of the raw data taken with the experiment. One of its major advantages is extensive particle identification over a wide range of momenta. The basic underlying physics concerns the process of ionization of gas molecules and the associated specific energy loss of traversing particles, described by the Bethe-Bloch formula.
In this thesis, a novel method for the tuning of parameters of the ALEPH parameterization of the Bethe-Bloch formula is presented based on the concept of hyperparameter optimization. A novel framework called OPTUNA and custom designed loss functions are investigated and tested against the performance on datasets with known particle identity from Run 2 of the LHC.
Besides the parameterization, further corrections to the mean as well as an estimation of the standard deviation of particle data distributions have to be made in high dimensions, which forms the main body of this thesis. Both parts are approximated with fully connected feed-forward neural networks trained on identified daughter particles from weak decays of K^0_S, Λ, $\bar{\Lambda}$ and γ conversions. An average accuracy of around 3‰ for the mean correction based on a neural network ensemble is achieved. This is compared with the results obtained from one-dimensional spline corrections in Run 2 and it is shown that the neural network introduced in this thesis can perform similarly well as the approaches from Run 2 but shows significant improvements in higher dimensions since it does not rely on a factorization approach.
The estimation of uncertainty of the distribution for each particle species is performed likewise, and an overall similar performance as the functional parameterization from Run 2 is achieved. However, due to the multidimensional mean corrections by the neural network and the limitations that a parameterized functional shape inherits, the standard deviation is captured better for all particle species by the neural network. In contrast to the Run 2 approach, this method works without additional iterations and consumes overall less time and effort for quality checks.

Zusammenfassung

Die gasbasierte Zeitprojektionskammer des ALICE Experiments am CERN spielt eine entscheidende Rolle vielen physikalischen Analysen innerhalb der Kollaboration und liefert ca. 92.5% der Rohdaten des gesamten Experiments. Einer der hauptsächlichen Vorteile des Detektors sind seine Fähigkeiten zur Teilchenbestimmung in einem breiten Impulsspektrum. Der zugrunde liegende physikalische Prozess ist die Ionisierung von Gasteilchen, die wiederum zu einem Energieverlust der traversierenden Teilchen führt, welcher durch die Bethe-Bloch Gleichung beschrieben wird.
Im Rahmen dieser Arbeit wurde eine neue Methode entwickelt, um die Parameter der ALEPH Parametrisierung der Bethe-Bloch Gleichung zu bestimmen. Sie basiert auf dem Konzept der Hyperparameter Optimierung mit dem neuen Framework OPTUNA und eigens entwickelten Loss Funktionen. Die Fitgüte wird mit der Parametrisierung aus Run 2 basierend auf Daten mit bekannter Teilchenidentität verglichen.
Darauf folgend müssen Korrekturen auf die Parametrisierung angewendet werden und die Breite der Teilchenverteilungen in höheren Dimensionen bestimmt werden. Sowohl die Korrekturen als auch die Sigma Parametrisierung werden in dieser Arbeit durch neuronale Netzwerke approximiert, die auf Tochterteilchen von V0 Zerfällen trainiert wurden. Dabei wurde für die Korrekturen eine Standardabweichung von 3‰ mithilfe von einem Netzwerk-Ensemble bestimmt. Dies wurde mit den Ergebnissen der Korrekturen von Run 2 verglichen und festgestellt, dass das neuronale Netzwerk ähnlich gute Ergebnisse liefert, jedoch einen signifikanten Vorteil, durch seine Fähigkeit höher-dimensionale Abhängigkeiten zu beschreiben, aufweist.
Die Approximation der Standardabweichung wurde ebenso mit neuronalen Netzwerken durchgeführt und erneut wurden ähnliche Ergebnisse wie in Run 2 gefunden. Durch die vorherige, multidimensionale Korrektur und die funktionale Beschränktheit der Fitfunktion von Run 2 wurden auch hier höhere Genauigkeiten für alle Teilchenspezies erzielt. Im Vergleich zu der Kalibrierung in Run 2 bedarf die hier gezeigte Methode keiner weiteren Iterationen und benötigt insgesamt weniger Zeit und Aufwand für Qualitätskontrollen.

Table of Contents

I Introduction

The Large Hadron Collider (LHC) located at CERN near Geneva is the largest particle accelerator in the world. Four major experiments, ALICE, ATLAS, CMS and LHCb, are located around the 27 kilometre (circumference) accelerator ring and measure particles emerging from collisions within their detectors.
With their groundbreaking discoveries, the experiments at the Large Hadron Collider (LHC) improve our understanding of the fundamental laws of nature and allow us to test the properties of the most precise theory of subnuclear particles and their interactions known to date, the Standard Model.

1.1 The four major experiments

While all the four major experiments at the LHC are dedicated to the studies of subnuclear matter, their constructions differ vastly and are fine-tuned to their specific physics goals. Each experiment is specialized to investigate certain aspects of subnuclear physics to high degrees of precision. In 2022 Run 3 of LHC reconvened with proton-proton beams after Long Shutdown 2 (LS2) (2018-2022). The four experiments can be seen in the schematic figure 1.1.

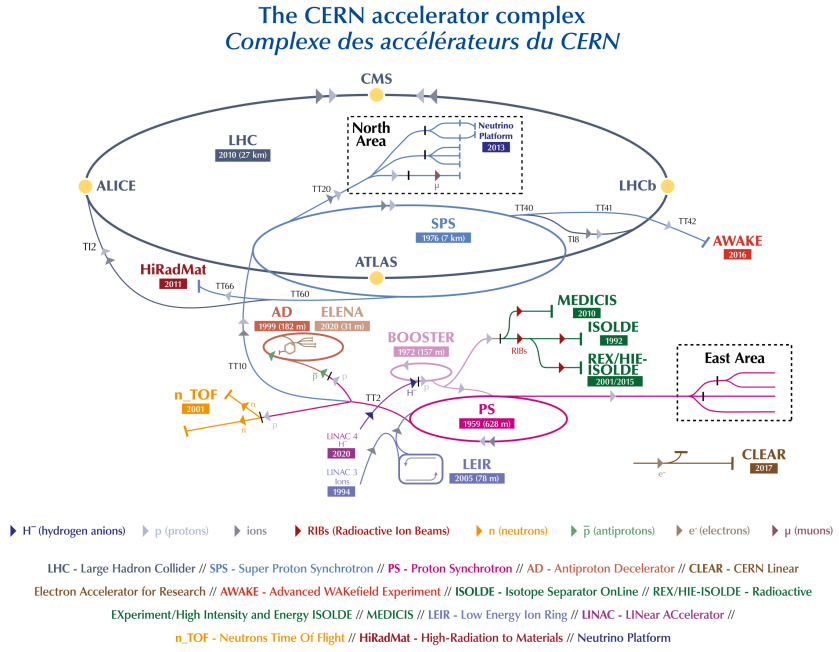


Figure 1.1: Schematic representation of the CERN accelerator complex with the four major experiments ALICE, ATLAS, CMS and LHCb.

ATLAS - **A** **T**orroidal **L**HC **A**pparatu**s**
Named after the Greek titan, ATLAS takes the Herculean tasks of searching for new particles in the Standard Model and beyond such as dark matter, investigating super symmetry and improving the precision of mass measurements of fundamental particles such as the Higgs boson. A lot of attention was focused on ATLAS in 2012 when it achieved one of its major physics goals, the discovery of the Higgs boson, which was the last missing particle in the Standard Model. ATLAS is the largest collaboration at CERN and can be found at interaction point (IP) 1 of the Large Hadron Collider.

CMS - **C**ompact **M**uon **S**olenoid
The physics goals of the CMS collaboration are similar to the ones of the ATLAS collaboration, though using a different detector setup and in particular a different magnet-system design. One of its major components is its large solenoid magnet, which can generate magnetic fields of up to 4 Tesla. In 2012 CMS codiscovered the Higgs boson and up until this day delivers some of the most precise measurements of particle masses ever made (e.g. the mass of the top quark).

LHCb - **L**arge **H**adron **C**ollider **b**eauty
The LHCb experiment focuses on the investigation of the bottom quark (called beauty quark), one of the fundamental particles of matter in the Standard Model. Its studies are dedicated to the investigation of the CP violation in the interaction of hadrons containing a bottom quark. This could provide insight to a fundamental physics question, the matter-antimatter asymmetry in the universe. Its detectors are set up in a forward configuration, in particular they are not in a cylindrical arrangement around the beam pipe such as for ATLAS or CMS.

ALICE - **A** **L**arge **I**on **C**ollider **E**xperiment
While all the above-mentioned experiments mainly focus on the interactions of proton-proton (pp) collisions, the physics programme of ALICE concentrates on the investigation of an exotic fluid, emerging from the collision of heavy-ion collisions, the Quark-Gluon Plasma (QGP). This fluid is investigated in Pb-Pb collisions and has one of the lowest viscosity over entropy ratios ever found (close to the theoretical limit), making it an almost perfect fluid. From the study of collective phenomena such as flow or the production of heavy (anti-)nuclei like ^{3}He, ^{4}He and the $^{3}_{\Lambda}$H, to individual particle interactions with the medium like jets or the study of heavy-flavour, ALICE covers a large variety of physics goals. Its extensive particle identification capabilities distinguish ALICE from the other experiments and make it unique for the study of the QGP.

II The ALICE experiment

Gaining experimental access to regimes of nuclear and sub-nuclear scale at energy densities similar to the early universe is a tremendous challenge. ALICE attempts to probe theories about the underlying physics by measuring particles emerging from collisions of protons or heavy ions inside its detectors. Each detector at the ALICE experiment has unique characteristics suited for specific tasks.

2.1 Overview

The detectors of the ALICE experiment can be divided in three main sectors, the cylindrically arranged central barrel of detectors around the beam pipe centred at the interaction point, the forward detectors located along the beam axis but shifted away from the interaction point and a small set of trigger detectors for triggering and event characterization. The forward detectors mainly deal with the investigation of muons (the muon forward spectrometer) and the remaining parts of nuclei emerging from a collision of heavy ions (mainly the ZDC (Zero Degree Calorimeter)). In contrast, the central barrel consists of the main tracking and particle identification detectors.

In radial direction, particles emerging from the collisions first traverse ITS (Inner Tracking System), TPC (Time Projection Chamber), TRD (Transition Radiation Detector) and finally TOF (Time-Of-Flight detector). These four detectors are mainly used for tracking and identification of the traversing particles. Their tracks are curved by the applied magnetic field (max. 0.5T) which allows a momentum measurement through the curvature radius. Going further outwards in radial direction, HMPID (High Momentum Particle IDentification), PHOS (PHOton Spectrometer) and EMCal (Electro-Magnetic Calorimeter) are found. A schematic representation of the detectors of ALICE in Run 3 of LHC is shown in figure 2.1. Besides a new, narrower beam pipe of LHC, many detector systems of ALICE have received major upgrades to cope with the higher interaction rates of the LHC in Run 3. In addition, continuous readout of the detectors together with substantially increased luminosities allow for strongly enhanced statistics and even the measurement of suppressed decay channels or rare processes.

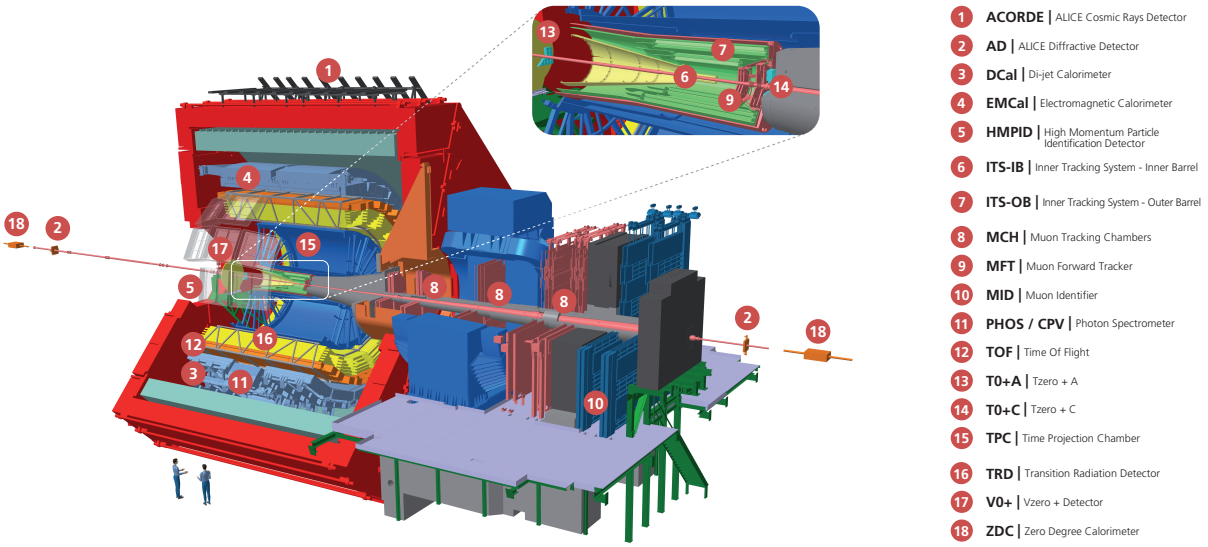


Figure 2.1: Schematic representation of the detector systems for Run 3 of the ALICE experiment, located at IP2 at the LHC.

2.2 Detectors of the central barrel

One of the most valuable characteristics of the experiment are its extensive particle identification (PID) capabilities. Many physics analyses conducted in the collaboration require a precise knowledge of particle yields in different momentum regions. Four detectors of the central barrel are crucial for this task: ITS, TPC, TRD and TOF. In combination, they can cover an approximate momentum range from ∼100 MeV/c up to ∼100 GeV/c.

2.2.1 ITS - Inner Tracking System

Located in closest proximity to the collision point, the ITS has multiple purposes for tracking and particle identification. As a silicon based detector, the primary advantage of the ITS is its excellent position resolution of tracks. The reconstruction of primary and secondary vertices greatly benefit from the innermost layers of the ITS, which consist of two layers of Silicon Pixel Detectors (SPD). In Run 2, two layers of Silicon Drift (SDD) and Silicon Strip Detectors (SSD) complemented the ITS with multi-track reconstruction capabilities. Furthermore, the analogue readout of the SDD and SSD detectors allowed particle identification via the measurement of the specific energy-loss of particles in matter ($\mathrm{d}E/\mathrm{d}x$).

Higher interaction and readout rates are foreseen for Run 3 of the LHC, hence the ITS required an upgrade from the setup used in Run 2. To reduce the material budget and improve readout rates, the ITS was fully replaced and comprises 7 layers of pixel detectors. Particle identification via the specific energy loss is thus not possible any more, but a substantially higher precision in vertex reconstruction and resolution is achieved. The precision of the impact parameter resolution increases with the transverse momentum of traversing particles and can be as good as 3 μm in the transverse plane and 4 μm in the longitudinal direction for a particle with a high transverse momentum ($\mathrm{p}_T \approx 20 GeV/c$).

All layers are built from a new 0.18 μm CMOS technology in which a matrix of charge collection diodes (pixels) is incorporated into a single, monolithic block of silicon, hence their name Monolithic Active Pixel Sensors (MAPS). This led to a reduction in the material budget

by a factor of seven for each layer (50 μm instead of 350 μm in thickness), increased the pixel density by a factor of 50 and allowed to place the innermost detection layer closer to the beam pipe axis. Moreover, higher readout rates for individual interactions of up to 400 kHz in pp and 100 kHz in Pb-Pb collisions are now supported and have been experimentally surpassed with measurements up to 4 MHz in proton-proton interactions, compared to 1 kHz in Run 2 [1].

2.2.2 TPC - Time Projection Chamber

Second in radial direction, after the ITS, comes the TPC detector. It is a gaseous detector and the main tracking device of the experiment. Cylindrically shaped, with an outer radius of 250 cm and a length of 500 cm, the ALICE TPC contains 88 m^3 of gas, making it the largest TPC ever built. A schematic view of the TPC as it is used for Run 3 can be seen in figure 2.2.

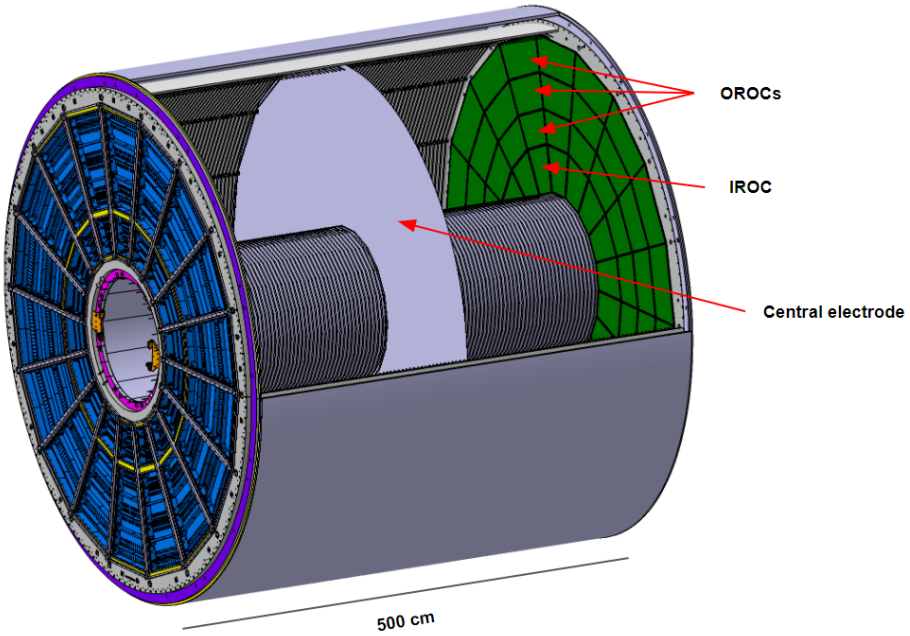


Figure 2.2: 3D representation of the ALICE TPC for Run 3 [2] with annotations.

A time projection chamber detector is mainly built from a large volume of gas or liquid in which traversing, charged particles release electrons from the detector material through the process of ionisation. These electrons drift towards the readout electronics (IROC, Inner Read-Out Chamber and OROCs, Outer Read-Out Chambers) via a constant, homogeneous electric field, generated from cathodes close to the readout at the end-caps of the TPC and a central electrode, located in the center of the TPC. Several components make the particle identification capabilities of this detector unique, such as the choice for the detector gas, the readout and front-end electronics as well as its homogeneous electric field.

Detector gas

Different gases have different properties for diffusion, quenching and drift velocities of electrons and charged ions. For the ALICE TPC, a mixture of Ne-CO_2-N_2 in the proportions 90-

10-5 is used (in order to obtain the percentages, the values have to be rescaled since they add up to 105). Experiments have shown that the gas mixture Ne-CO_2, 90-10 showed very good performance, but the addition of another quencher gas (N_2), shows beneficiary effects for the operational stability of the chamber at higher electric fields [3]. In particular, quench gases have a high cross-section for photon absorption in a broad interval of wavelengths which reduces secondary showers of photons emitted by accelerated electrons.

Readout

The drifting electrons are accelerated by the homogeneous electric field and eventually reach the readout electronics located at the end-plates (cathodes) after a maximum drift time of $\sim 100\mu s$ for electrons emitted close to the central electrode. The charge of the electrons, i.e. the number of electrons released by the process of ionisation, is proportional to the specific energy-loss per unit distance of the original particle in the detector gas. This allows for particle identification and tracking. Likewise, positively charged ions are released from the collisions of the accelerated electrons with the gas molecules.
In order to be able to measure a signal, the arriving charges must be amplified. In Run 2, the drifting electrons passed a Multi-Wire Proportional Chamber (MWPC). For a MWPC, wires are positioned in a grid-like structure, as can be seen in the illustration 2.3. The electric field around each anode wire accelerates the drifting electrons further, causing them to induce avalanches. Additional wires (gating plane) are positioned further away from the readout plane which serve the purpose of holding back positively charged ions disturbing the homogeneous electric field of the TPC (ion back-flow). However, the operation of the MWPC together with a gating grid limits the data readout rate to 1 kHz.

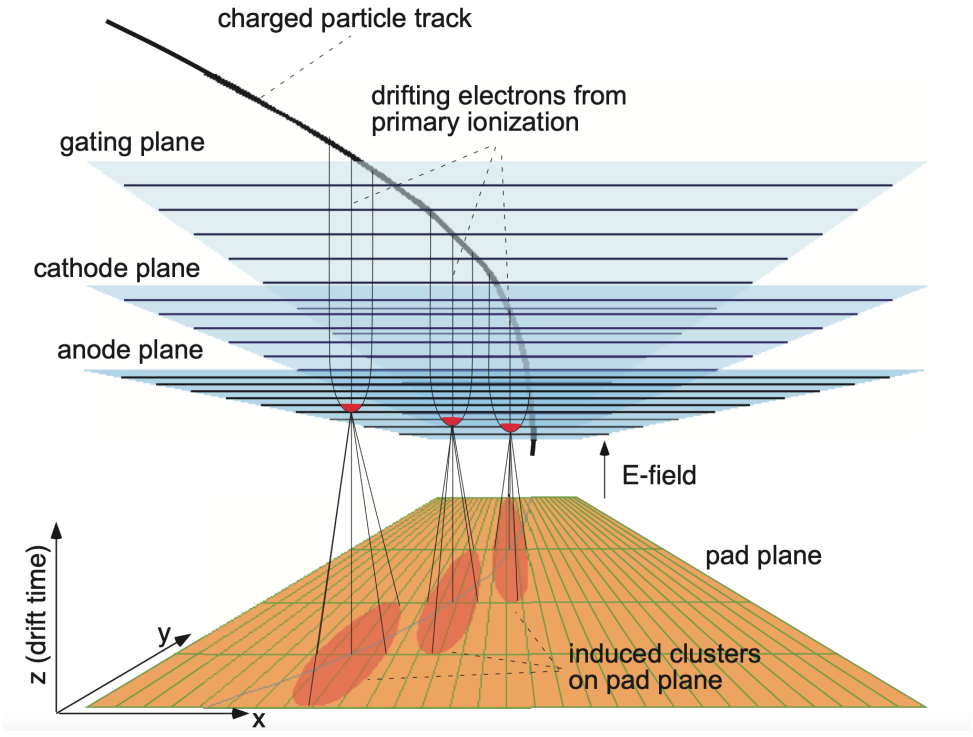


Figure 2.3: MWPC as it was used in the ALICE TPC for Run 2 [4].

For Run 3 the MWPC was replaced by Gas Electron Multipliers (GEMs) which allow continuous readout with a rate of up to 50 kHz in Pb-Pb collisions.
The GEMs for the ALICE TPC consist of a 50μm thin insulating polyimide foil, coated with copper (2-5 μm in thickness on each side). Double-conical holes with an inner diameter of around $\sim$50μm and an outer diameter of $\sim$70μm are imprinted in the material. A potential of 200-400V is applied over the two electrically conducting layers producing field strengths of $\mathcal{O}$(50 kV/cm) in each hole, sufficient for avalanche creation [3]. A simulation of the charge amplification in a GEM hole can be seen in figure 2.4.

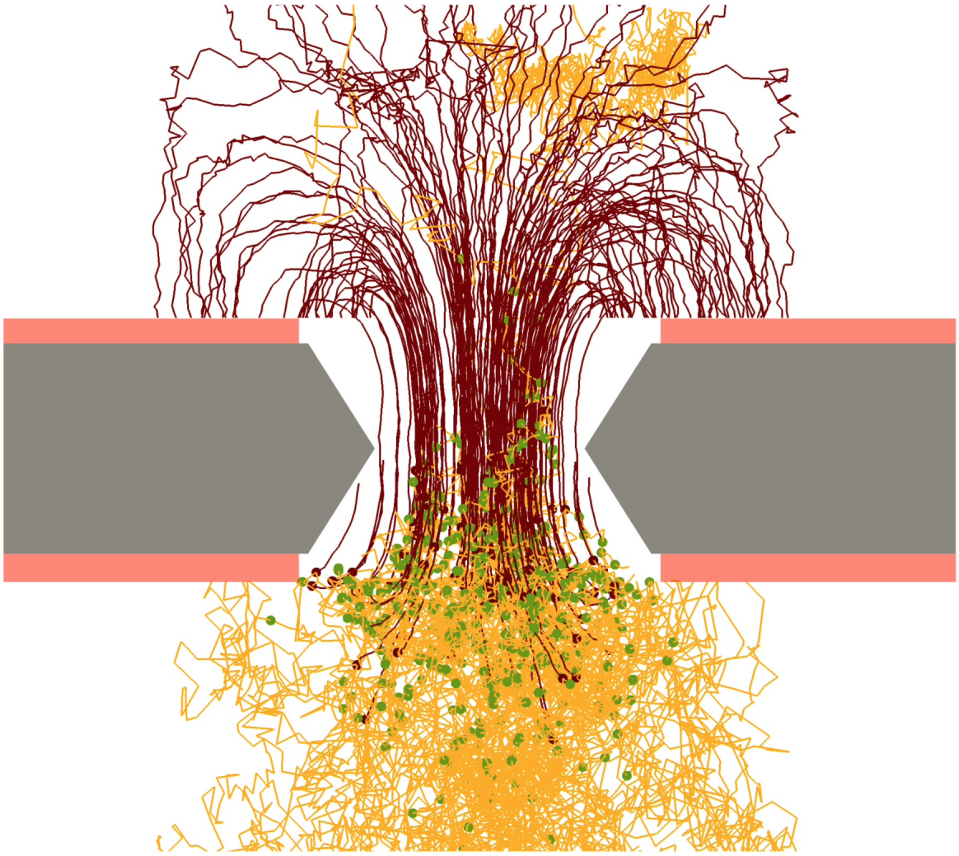


Figure 2.4: Cross-sectional view: Garfield simulation for an electron passing a GEM hole [3].

2.2.3 TRD - Transition Radiation Detector

The TRD detector is based on the emission of radiation of a particle upon crossing the boundary of two materials with different dielectric constants. Each chamber of the ALICE TRD is built from a foam/fibre radiator, followed by a drift volume filled with a Xe-CO_2 gas mixture and MWPC's which is preceded by a 3cm drift region. The main task of the TRD is the separation of electrons from heavier particles, since transition radiation is emitted for $\gamma > 1000$ which is only fulfilled for electrons. Hence, they lose proportionally more energy per unit distance than heavier particles which allows particle identification. Furthermore, the time information obtained from the track passing the detector together with the fast online reconstruction made the TRD usable as a trigger for collisions in Run 1 and 2.

2.2.4 TOF - Time-Of-Flight detector

Following in radial direction, the TOF detector shows excellent particle identification capabilities in the mid-momentum range ($0.6 - 5$ GeV/c) and additionally provides a trigger for cosmic rays and ultra-peripheral collisions. It is located at 3.7 meters in radial direction from the interaction point, covering a pseudo-rapidity range of $-0.9 < \eta < 0.9$ with 18 azimuthal sectors and a total of 152928 readout channels. This avoids high detector occupancies, even in high multiplicity events.
Particle identification is performed by measuring the time-of-flight between a collision (t_{ev}) and the detection in the sensitive volume of the detector (t_{TOF}). Based on the Multigap Resistive Plate Chambers (MRPC) technology, the TOF detector of ALICE measures traversing particles by amplifying their signal using 5 stacks of glass resistive plates, separated by gaseous layers in which the incident particle creates avalanches based on an externally applied, homogeneous electric field [5].
Measuring the velocity of an incident track, together with the momentum information inferred from the curvature radius of a particle in the magnetic field, particle identification can be performed based on the calculated mass of the track

$$m = |\vec{p}| \cdot \sqrt{\left(\frac{t}{l}\right)^2 - 1} \tag{2.1}$$

where $|\vec{p}|$ is the absolute value of the measured momentum, t is the time-of-flight ($t = t_{\text{TOF}} - t_{\text{ev}}$) and l the traversed distance.

2.3 O^2 - Online-Offline computing system

The upgrade of many detectors in ALICE for Run 3, specifically ITS and TPC, allow data-taking at higher collision rates compared to Run 2. This poses tremendous challenges for data-taking and reconstruction since computing resources and storage have to be used most efficiently. Hence, a completely new software framework (O^2) was designed and is still under commissioning at the time of writing this thesis.

The general tasks for the software framework can be categorized in two sections: Online, synchronous processing, which deals with all computing tasks that have to be executed at data-taking time of the detectors (mainly online reconstruction and data-compression) and offline, asynchronous processing, which concerns all processes decoupled from data-taking (in particular offline reconstruction, calibration and analysis).

In Pb-Pb collisions, the raw data taking rate is ≈3.5 TB/s. Cluster recognition and correlation in the TPC is used for lossless compression (Huffman coding) and allows to reduce the datarate to 1.5 TB/s for the online farm. Further reconstruction and background elimination results in a rate of 100GB/s of data being written to disk.

2.3.1 Data processing and storage

Besides the updates on the hardware, a major effort was conducted to renew the analysis software of ALICE. The key motives for designing this new framework called O^2, are computational efficiency of calculations and the lowest possible memory consumption. The computing

model was changed from assuming the Worldwide LHC Computing Grid (WLCG) as a homogeneous entity, capable of handling any type of job, to dedicated facilities that take care of specific tasks [6].

- ALICE Online-Offline Facility (O^2)
  Online reconstruction and calibration for data compression are conducted here. RAW data is reduced to Compressed Time Frames (CTFs) and finally to Analysis Object Data (AODs). Data is compressed to Sub-Time Frames where each frame contains $\sim$20ms of data meaasured with an arbitrary reference clock (the so called heartbeat trigger) and is passed from the First Layer Processors (FLPs, data compression factor $\approx$2.5) to the Event Processing Nodes (EPNs) for further reconstruction (data compression factor $\approx$8).

- Tier 0 and Tier 1 - CERN Computer Centre facility with grid site
  Provides CPU, storage and archiving resources and takes care of further reconstruction and calibration tasks as well as simulations. Receives CTF from the O^2 facility and AODs from Tier 2 to perform further compression and additions to AODs.

- Tier 2 - Regular grid site
  High bandwidth connection, running simulations. Additions to AODs via Monte Carlo simulations (MC).

- AF - Analysis facility
  Dedicated facility for analysis with High Performance Computing (HPC) infrastructure. AODs are converted to trees and histograms.

While for pp collisions synchronous and asynchronous processing of the incoming data is possible, data rates for Pb-Pb collisions are too high for the asynchronous processing such that AOD re-filtering is needed. The O^2 facility will be capable of storing and processing roughly 2/3 of the CTFs for Pb-Pb and 1/2 for pp data taking, while Tier 1 sites support the asynchronous processing of the remaining CTF data.

The central facility for storing calibration objects is the condition and calibration database (CCDB). The central aspect for storing objects on the CCDB are intervals of validity. Each object is stored with a "valid-from" and "valid-until" timestamp which specifies the range of collisions for which the object is valid. The CCDB is used from online processing down to the analysis level and provides a low-latency, scalable infrastructure for central objects of calibration and reconstruction.

2.3.2 Software design and computing model

Data processing on the analysis level is done on AODs which are produced after data has passed the reconstruction and calibration workflows. The data (tracks) in AODs are stored in ROOT histograms and is processed in the O^2 software framework in form of tables. With similarities to relational databases, tables can be joined to combine columns. These tables can be used in functions and tasks to perform physics analyses. Tasks can have dependencies on each other since they can add columns to tables. This makes O^2 highly modular and computationally efficient.
Besides local computations, O^2 further provides possibilities for decentralized computing on

the WLCG via the Hyperloop system. Tasks are submitted in the form of "wagons" which are appended to "trains" running workflows with custom configurations on specified datasets.

III Machine learning & Neural Networks

Modern machine learning (ML) techniques gain ever greater popularity in particle physics in the last decades. A prime example are boosted decision trees which have long been used to perform signal-to-background separations of measurements. However, machine learning offers many more opportunities in a variety of different tasks.

3.1 Overview of machine learning methods

Most common machine learning methods can be classified into two main categories, supervised and unsupervised learning.
Supervised learning uses a set of training data (X) to learn a specific output (Y), i.e. a supervised model adapts to map an input to an output, $f(X) = Y$. On the contrary, unsupervised models only see an input dataset (X) where a mapping to an output space is only determined by the construction of the algorithm and the dataset.

Common tasks addressed by machine learning include regression, classification, cluster finding, anomaly detection and creating synthetic data. While supervised algorithms typically learn the underlying probability distribution of a given training dataset, unsupervised methods are based on algorithmic procedures which assume a general structure in the data.

In general, many machine learning based methods focus on the approximation of a solution for a given problem by minimizing a *loss-function*. The loss-function is typically connected to the discrepancy of a model to a given dataset. This requires the mathematical definition of a "good solution" (or ideally a perfect solution), which is in many cases a non-trivial task.

Typically, a machine learning method which does not assume any form of the underlying probability distribution is called *distribution-free*. Additionally, such methods typically do not have a fixed set of parameters, making them *non-parametric*. This does not mean that the given method does not have parameters, but rather that the amount of parameters and the particular values are determined based on its performance on data.
Adjusting these parameters on a training dataset is commonly referred to as *training a model*. Many algorithmic approaches have been investigated in the literature that can minimize a loss-function while adjusting the parameters of the given model. Most well-known are two approaches called gradient descent and hyperparameter optimization.

3.2 Gradient descent and hyperparameter optimization

In physics, problems are typically expressed in mathematical equations of a closed form. In some cases, this allows the calculation of maxima or minima by using first order derivatives and finding solutions to analytically solvable problems. However, the vast majority of real-

world problems do not have analytic solutions and require iterative approaches to find approximations of solutions.

3.2.1 Gradient descent

A well-known example of finding a solution using an iterative approach is the search for the roots of real-valued functions. One such algorithm is the Newton-Raphson method, which finds better approximations of the roots at each iteration x_n given a sufficiently good starting value x_0 for a function $f(x), x \in \mathbb{R}$. At each step x_n is updated to x_{n+1} using

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{3.1}$$

To find extrema of a function $f(x)$, the Newton-Raphson method is applied to the first derivative $f'(x)$ of the function $f(x)$.

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}. \tag{3.2}$$

For functions of more than one variable, the first derivative is expressed with the vector-gradient ∇, hence the general class of such algorithms is called gradient descent methods.

Generally, gradient descent methods can be expressed as

$$x_{n+1} = x_n - \gamma \nabla F(x_n) \tag{3.3}$$

where γ is the so-called *learning rate*. A good choice for γ can significantly reduce the number of iterations needed until an approximate solution is found.

Real-world problems often require gradient calculations on a large dataset, which makes computational cost and memory consumption a significant factor of consideration. A well performing approach to reduce memory consumption is to calculate gradients only on randomly sampled subsets of data. The gradient obtained by this method still allows a good representation of the gradient of the full dataset. Common algorithms employing this method are the Stochastic Gradient Descent (SGD) or adaptive momentum estimation optimizer (Adam).

3.2.2 Hyperparameter optimization

A hyperparameter of a model can be defined as any parameter which cannot be optimized through gradient descent. In particular, this includes parameters which define the model itself, i.e. the number of parameters needed for a model to describe data well.

Classical methods for a hyperparameter tuning are exhaustive grid or random searches. The function for which the minimum needs to be found is evaluated at many different phase-space points (spanned by the model parameters) which allows the approximate determination of a minimum through interpolation of the evaluated points.

Besides exhaustive searches, algorithmic approaches have been designed to optimize the search for such parameters. One algorithmic approach which is commonly employed is Bayesian optimization which tries to find a balance between exploration (most uncertain outcomes) and exploitation (trials close to the optimum) thus mapping the phase-space on a probabilistic basis. It performs better than exhaustive searches, but more recent approaches based on pruning and evolutionary algorithms have shown even better performances for certain tasks like neural architecture optimization. Evolutionary algorithms start at random points throughout the phase-space and evaluate the performance of trial points using a fitness function. Iterations are based on previous trials where the fitness function is optimized on the explored phase-space of previous iterations. Less promising trials are stopped (pruning, early stopping) and are replaced by new trials starting at different points in the phase-space, mimicking an evolutionary behaviour, hence the name "evolutionary algorithm".

3.3 Neural Networks

Neural networks are a part of supervised learning and can perform classification and regression. Although their architectures and constructions can vary widely depending on the task at hand, some of the mathematical working principles are shared. The essential components which define a network are

- The neurons in the input layer.
  The number of neurons in the input layer is proportional to the number of features per observation in the input data. So-called "channels" like RGB (red, green, blue) in an image can add further neurons.
- The neurons in the hidden layers.
  A network with 2 or more hidden layers is typically called a deep neural network, while the architecture is highly task-dependent and allows limited freedom for design choices.
- The neurons in the output layer.
  These are typically defined by the task at hand, e.g. regression or (multi-class-) classification.
- The loss-function.
  This is the essential component to define what and how a network should learn from data. In particular, the emphasis of the task at hand is encoded in it.

3.3.1 From neurons to networks

The atomic unit of a neural network is called neuron. One neuron typically comprises an activation function $\sigma(\tilde{x})$ (where $\tilde{x}$ is an scalar value), a weight vector w and a bias value b (scalar). For a given input vector x, the output of one neuron is a scalar value corresponding to

$$\sigma(w \cdot x + b) = \sigma\left(\sum_{i=0}^{dim(w)} w_i * x_i + b \right) \tag{3.1}$$

The activation function σ can here be any function that satisfies a $\mathbb{R} \to \mathbb{R}$ mapping and is piece-wise differentiable. Commonly used functions are the hyperbolic tangent, ReLU (Rectifying Linear Unit), sigmoid and softmax function (multi-class classification) which are shown in figure 3.1. Each of them serve special purposes for different tasks.

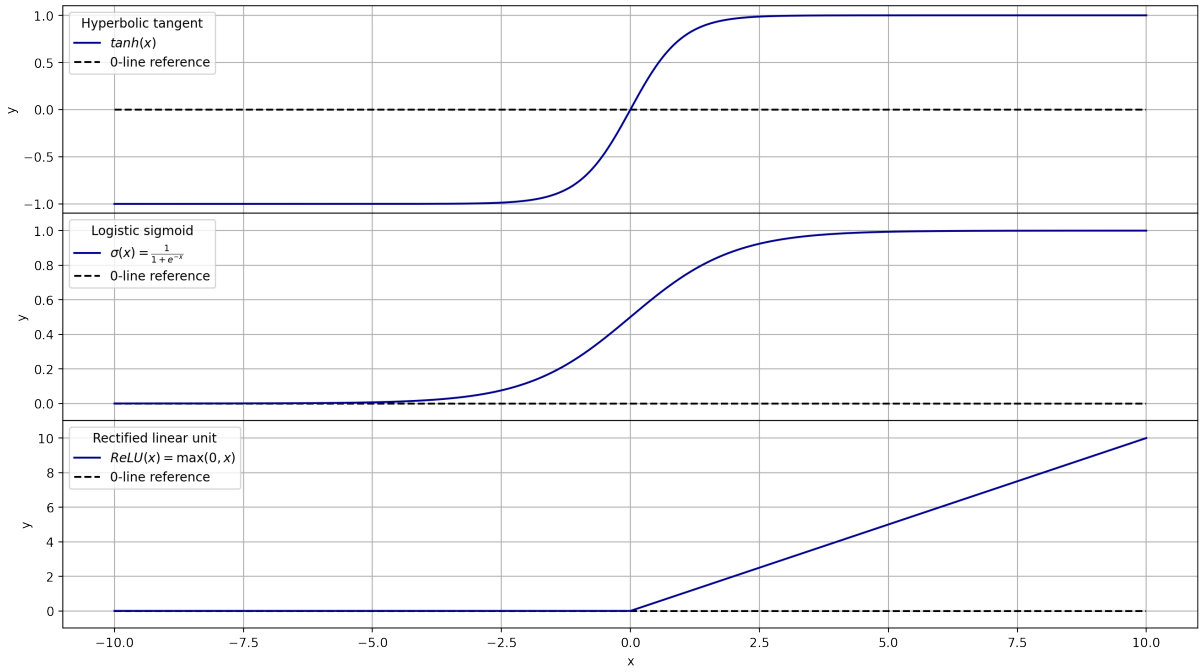


Figure 3.1: Common 1D Activation functions of neurons in a neural network.

One can also append the bias to the weight vector by appending a scalar value of 1 to the input data, making it a pure vector multiplication

$$[x_1, ..., x_n] \cdot [w_1, ..., w_n]^T + b = [x_1, ..., x_n, 1] \cdot [w_1, ..., w_n, b]^T \tag{3.2}$$

Having defined one such neuron now allows the construction of a neural network as a concatenation of neurons in layers and connecting the inputs of a layer with the outputs of the previous layer(s). Figure 3.2 shows a representation of a standard, fully connected neural network. From the input layer onwards, every neuron (circles) receives the output of all neurons from the previous layer as input (straight lines), computes a scalar value and passes it on to every other neuron of the subsequent layer.

Since every neuron is fully specified by the activation function, the weight vector and the bias, a layer of a neural network can be represented as a matrix-vector-multiplication

$$\text{layer output:} \quad \sigma\left([x_1, ..., x_n, 1] \cdot \begin{bmatrix} w_{11} & \dots & w_{1l} \\ \vdots & \ddots & \vdots \\ w_{n1} & \dots & w_{nl} \\ b_1 & \dots & b_l \end{bmatrix}\right) \tag{3.3}$$

[1]Source: https://www.ibm.com/cloud/learn/neural-networks

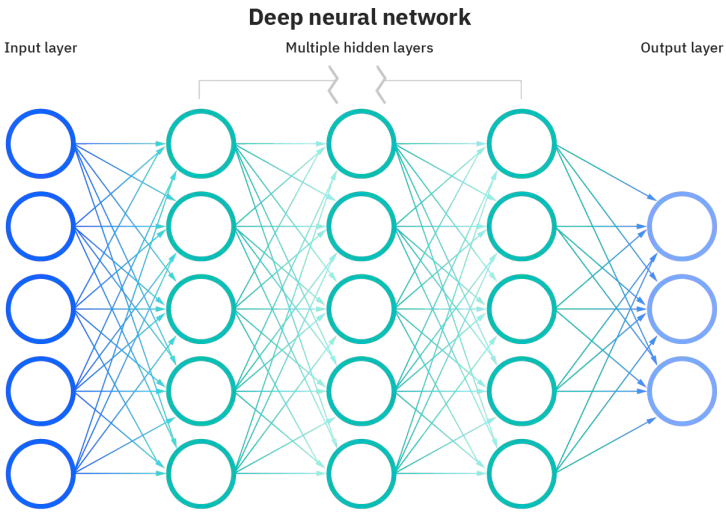


Figure 3.2: Example of a fully connected neural network built from single neurons represented in coloured circles. [1]

where the activation function σ here is meant to be applied element wise. For this reason, neural network training and execution are highly parallelizable and make them excellent candidates for calculations on Graphics Processing Units (GPUs).

For practical applications, the weights of a neural network have to be determined in a mathematical way to fulfil a meaningful task. In particular, it can be advantageous to find hidden correlations in data without imposing structure on it. Hence, the determination of weights and biases is typically purely based on a set of training data which is representative of the underlying probability distribution of the data density on which the network is being applied after training. An exception to this rule can be graph neural networks, where the architecture and values of the chosen parameters are specifically tuned to perform a certain task.

3.3.2 Backpropagation and training networks

Determining the weights of a neural network mathematically is done with an iterative approach on a set of training data. At each iteration the network performs its calculations on a part of the training dataset, the so called mini-batch and calculates a loss score based on its predictions and the known labels / values in the training and validation set. Optimally, this score represents a meaningful value for the goodness of fit of the network prediction to the mini-batch. A lower loss score implies a better fit.

It is practically impossible to initialize the network with ideal weights. Since most or all of the neurons in a neural network are interconnected, changing one weight-value can alter the output of the entire network. This implies that the "loss surface" produced by the network is highly complex, with many fluctuations in close neighbourhoods. An example of how such a loss surface can look like is shown in figure 3.3. This further demonstrates the necessity of a rigorous algorithm that can optimize the weights of a network given a training dataset in a local neighbourhood of the starting point.

Since the activation functions of the neurons are chosen to be at least piece-wise differentiable, gradients of the loss function can be computed and *back-propagated* through the network to adjust the weights. Gradient descent algorithms can then be used to find iterative steps to improve the weights, moving towards minima in the loss surface.

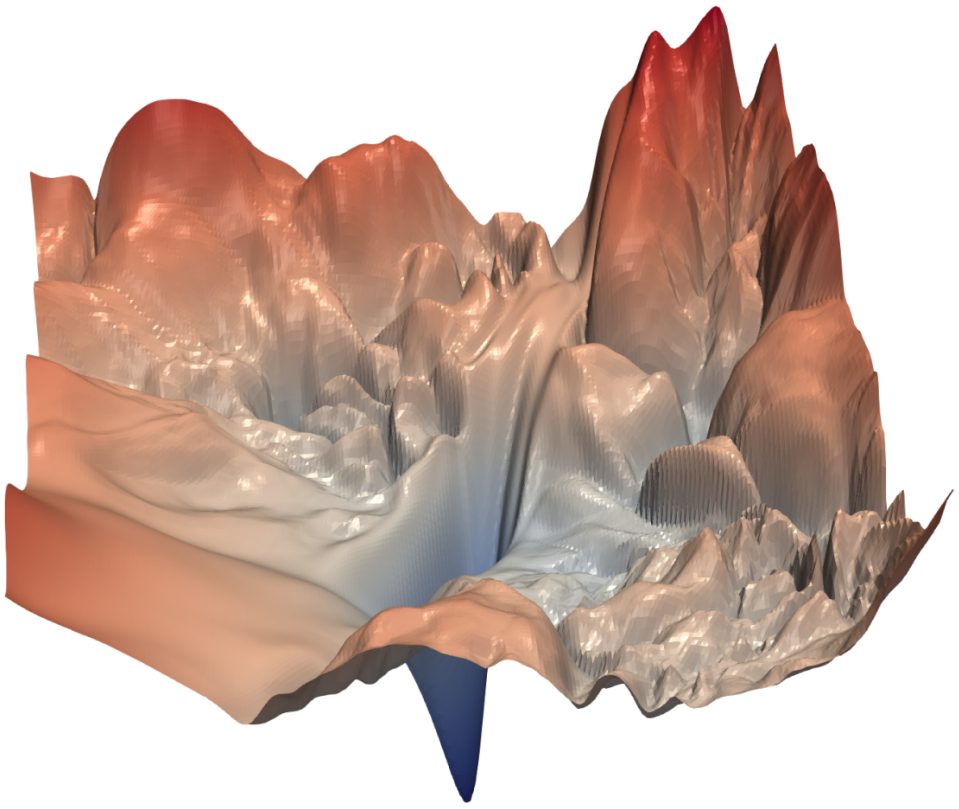


Figure 3.3: Representation of a loss surface of a neural network. The irregularities show the difficulty of finding optimal weights. [7]

If the loss surface is convex, the training is guaranteed to converge in the global minimum, while for non-convex loss surfaces, the training converges at least in a local minimum.

Computationally, back-propagation is one of the most memory consuming parts of training a network. In order to reduce the computational cost, the gradient can be calculated on a randomly chosen subset of the batch (stochastic gradient descent). Furthermore, modern libraries support automatic differentiation, such that the gradients do not have to be implemented separately for each layer of the network.

3.3.3 Architectures and design choices

Various design choices in the construction of a neural network can be made which allow for more efficient approximations of the data at hand. Starting from a single neuron, a choice for the activation function and the number of input and output connections can be made. Not all networks have to be fully connected and not all neurons in a network have to connect to just the subsequent layer.

An example of a non-subsequent connection is a so-called skip connection, shown in figure 3.4. The skip connection was popularized with the residual neural network and allows the residual block (all layers that the skip connection span over) to learn the difference between the input and the output of the layers. This can be easily seen by

$$R(x) \coloneqq \mathcal{F}(x) + x \Rightarrow \mathcal{F}(x) = R(x) - x \tag{3.4}$$

where $\mathcal{F}(x)$ is the output of the block encapsulated by the skip connection.

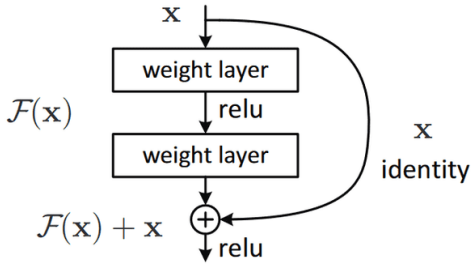


Figure 3.4: A skip connection as implemented in a residual neural network. [2]

Besides the connections of neurons among each other, the choices for the activation function of the neurons make a critical difference for the fit performance of the network. While the mathematical literature commonly uses the ReLU activation for simplicity since it is piecewise linear, it does not necessarily perform best for every task. Another common choice for regression is the tanh function, while for classification, the sigmoid and softmax functions are used.

In particular, the choice for the activation function in the last layer of the network is critical. For regression, the last layer is classically chosen to have no activation function (or for that matter, the identity function as activation). This allows the network to adjust for scaling and shifting with the weights in the last layer, and returns an output space ranging from $-\infty$ to $+\infty$. For classification, the activation in the last layer should be chosen as the sigmoid function for a two-class classification problem and the softmax function for an n-class classification problem.

Besides the activation functions, an equally important choice for training is the loss function. It encodes what the network learns, i.e. for a regression problem the network should best approximate all data points, which is typically evaluated with the mean square error (MSE) loss.

$$\text{MSE} = \sum_{i \in data} (y_i^* - \hat{y}_i)^2 \tag{3.5}$$

where y_i^* is the true value to be fitted by the network for data point i and $\hat{y}_i$ is the predicted value.

[2] Source: https://theaisummer.com/skip-connections/

For classification, the cross-entropy loss function has to be chosen in order to compare the predicted class labels $\hat{y}_i$ with the ground truth class labels y_i^*

$$\text{CE} = \sum_{i \in \text{labels}} y_i^* log(\hat{y}_i) \tag{3.6}$$

To summarize, given a task, the activation functions for the last layer and loss functions as shown in table (3.1) should be used

Problem at hand	act. function in last layer	loss function
Regression	linear	MSE
Classification (single label, two classes)	sigmoid	binary CE
Classification (single label, multiple classes)	softmax	CE
Classification (multiple label, multiple classes)	sigmoid	binary CE

Table 3.1: Activation and loss functions for neural network architectures and their application for specific problems.

IV Particle identification using the specific energy loss in the ALICE TPC

4.1 General aspects of particle identification with the TPC

Particle identification is one of the main tasks of the TPC in the ALICE experiment. Particles traversing the detector lose energy in collisions with the atoms and molecules of the detector gas, which causes the emission of electrons. A theoretical prediction for this loss of energy per unit path length in the detector gas was first proposed by Hans Bethe [8]. It depends on the electron density of the detector material n, the mean ionization energy of the atoms of the detector material I, the charge of the traversing particle z, and the velocity of the traversing particle which is encoded in the Lorentz factors $\beta = \frac{v}{c}$ and $\gamma = \frac{1}{\sqrt{1-(v^2/c^2)}}$ where $\beta\gamma$ can be calculated given the mass m and momentum p of a particle as $\beta\gamma = p/m$

$$\left\langle -\frac{dE}{dx} \right\rangle = \frac{4\pi n z^2}{m_e c^2 \beta^2} \cdot \left(\frac{e^2}{4\pi\epsilon_0}\right)^2 \cdot \left(\frac{1}{2}\ln\left(\frac{2m_e c^2}{I}\beta^2\gamma^2\right) - \beta^2 - \frac{\delta(\beta\gamma)}{2}\right). \tag{4.1}$$

Here the electron density of the detector material n can be given as an expression of the atomic number Z, the relative atomic mass A, the material density ρ and the molar mass constant M_u as

$$n = \frac{N_A \cdot Z \cdot \rho}{A \cdot M_u}. \tag{4.2}$$

This allows the calculation of the energy loss in different materials for an approximate kinematic range of $0.1 \lesssim \beta\gamma \lesssim 1000$. First approximations for the mean excitation potential of the detector material were first performed by Felix Bloch [9] and states $I \approx (10\text{eV}) \cdot Z$. Hence, formula (4.1) is called Bethe-Bloch formula.

As highly energetic particles traverse the detector medium, their electric field is subject to Lorentz contraction in the direction of flight, but extends further in the transverse direction. This causes electromagnetic interactions with more distant molecules and a contribution to the Bethe-Bloch formula corresponding to a factor $\beta^2\gamma^2$ in the logarithmic term. This term is referred to as the relativistic rise. However, the larger field extension in the transverse direction polarizes the surrounding molecules, which partly shields the electric field and partly limits the relativistic rise. This so-called "density effect" is described by the term $\delta(\beta)$, which adds even stronger contributions in dense media (e.g. solids).

Formula 4.1 is constructed for heavier particles (e.g. pions, protons) scattering on comparably light shell electrons. Due to their lighter mass, the kinematics of the ionization process changes significantly for electrons. Additionally, quantum effects have to be taken in consideration which, together with the energy loss from Bremsstrahlung, makes formula 4.1 not

applicable for electrons in this form.
Since particles with high momenta can transfer a lot of energy to shell electrons upon scattering, it is possible that electrons are liberated (δ-electrons) from the shells, forming secondary tracks which are no longer attributed to the original track. In order to take this effect into account, an energetic cut-off E_{max} has to be introduced to the Bethe-Bloch formula

$$\left\langle -\frac{dE}{dx}\right\rangle = \frac{4\pi n z^2}{m_e c^2 \beta^2}\cdot\left(\frac{e^2}{4\pi\epsilon_0}\right)^2\cdot\left(\ln\left(\frac{\sqrt{2m_e c^2 E_{\max}}\beta\gamma}{I}\right)-\beta^2-\frac{\delta(\beta\gamma)}{2}\right). \tag{4.3}$$

E_{max} is strongly dependent on the detector gas/material and effectively limits the maximum considered recoil energy of the two colliding particles. With the energetic limit, the Bethe-Bloch formula is also valid for electrons.

Figure 4.1 shows a measurement of the $\langle d\mathrm{E}/d\mathrm{x}\rangle$ with the ALICE TPC. At low $\beta\gamma$, the specific energy loss falls steeply with $1/\beta^\alpha$ where $\alpha \approx 1.6-2$ until the minimum ionizing region is reached at $\beta\gamma \approx 3.6$. The energetic cut-off together with the density effect lead ultimately to the complete cancellation of the relativistic rise and cause the function to reach a limit at very high $\beta\gamma$, the Fermi plateau.

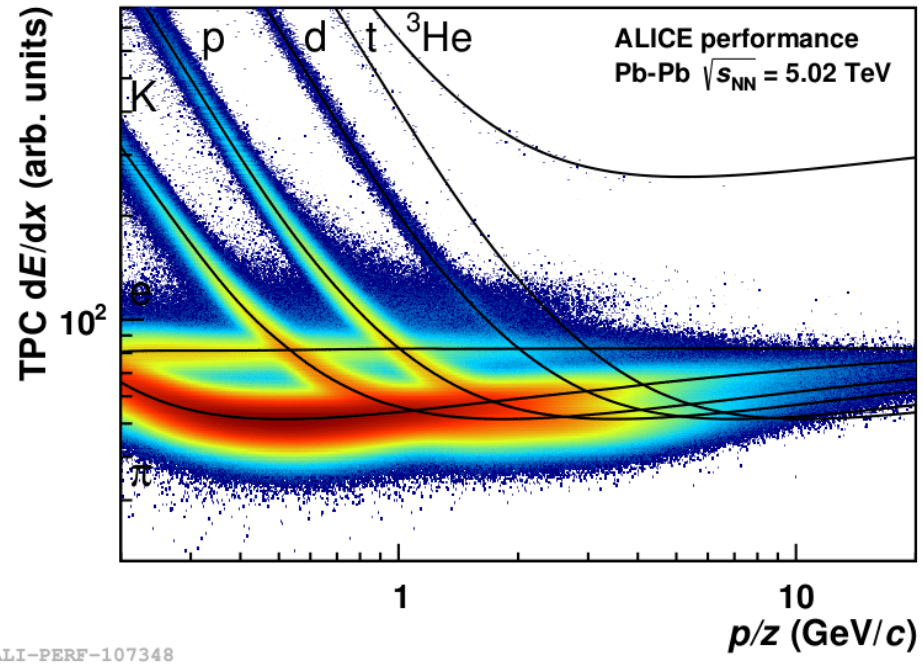


Figure 4.1: $\langle d\mathrm{E}/d\mathrm{x}\rangle$ performance plot of the ALICE TPC at $\sqrt{s_{NN}} = 5.02$ TeV [10]. Several particle species are visible together with their parameterized Bethe-Bloch curve.

In order to obtain such a dE/dx distribution, it is necessary to assign one specific value of dE/dx for each track. For this, the truncated mean estimator is used.

4.1.1 The truncated mean estimator

The energy loss of a particle in the TPC is read out at the pad rows, where 159 (152) for Run 2 (Run 3) charge clusters are collected. The deposited charge per cluster ΔQ is proportional to

the energy loss per unit distance $\frac{\Delta E}{\Delta x}$ of the original particle.
The measured values over the pad row follow a Landau distribution. This distribution has non-finite first (mean) and second (variance) order moments, which would lead to large fluctuations if an average value is calculated. To overcome this problem, the αN-lowest values ($\alpha \in]0, 1]$) are used to assign a dE/dx value for any given track. This estimator, the *truncated mean* $\langle S \rangle_\alpha$, has shown good performance for the mean calculation while keeping fluctuations small. For N measurements, the truncated mean $\langle S \rangle_\alpha$ is defined as

$$\langle S \rangle_\alpha \coloneqq \frac{1}{\alpha N} \sum_{i=1}^{\lceil \alpha N \rceil} \left(\frac{\Delta E}{\Delta x} \right)_i . \tag{4.4}$$

Using the truncated mean results in a reliable dE/dx estimation, where the final distribution of dE/dx values follows approximately a Gaussian distribution.

Based on the separation power of the particle species in different momentum regions, the parameter α of the truncated mean is chosen. For this, the dE/dx-distance between the minimum ionising region and the Fermi-plateau divided by their average resolution is considered. Typical values for α are found to be between 0.5 and 0.7 [11].

4.1.2 Parameterization of the Bethe-Bloch function

Due to the use of the truncated mean estimator and in order to better fit the obtained particle distributions, a parametrization of the Bethe-Bloch function is used, where seven parameters allow adjustments in different kinematic regions. A commonly used functional shape is the ALEPH parametrization which can be written as a function of $\beta\gamma$, the charge z and the parameters $a_1, ..., a_5, f_z$ (the charge factor) and f_{MIP} (the (dE/dx)-value of the minimum ionising region)

$$\langle dE/dx \rangle_{ALEPH} = a_1 * (a_2 - \log(a_3 + (\beta\gamma)^{-a_5})/\beta^{a_4} - 1) * z^{f_z} * f_{MIP} \tag{4.5}$$

Optimal parameters for this parametrization are not known a priori, but have to be determined from a fit to data. The regions of phase-space in which the parameters are found can be approximated by calculating the appropriate pre-factors in the Bethe-Bloch function. For each track, two properties are accessible as physical observables, the dE/dx measured in the TPC and approximated by the truncated mean as well as the momentum which is typically measured at the inner wall of the TPC. Particle species assignments can then be made by testing different mass hypotheses for each measured track and comparing with the parametrization.

In order to assess the statistical significance of the particle species assignment, the uncertainty of the data distribution has to be estimated. Expected values for the width of each particle distribution are determined by statistical fluctuations in the measurements, as well as residual uncertainties of readout and detector effects. An estimate of the intrinsic uncertainty of the energy loss is given at ≈5-7% of the expected dE/dx value [3].

4.1.3 Relevant observables for particle identification

While the one-dimensional functional form of the Bethe-Bloch parametrization already performs well for particle identification, a dependence on several other parameters exists. In particular, detector effects have to be taken into consideration.

The inclination angle of the track, which is measured with the local polar angle $\tan(\lambda)$. A higher inclination angle leads to a larger charge deposition per pad row in the readout and thus to an enhanced dE/dx signal. The $\tan(\lambda)$ is also connected to the more commonly used pseudo-rapidity of the track with $\eta = -\ln(\tan(\frac{\pi}{4} - \frac{\lambda}{2}))$. Their dependence is depicted in 4.2.

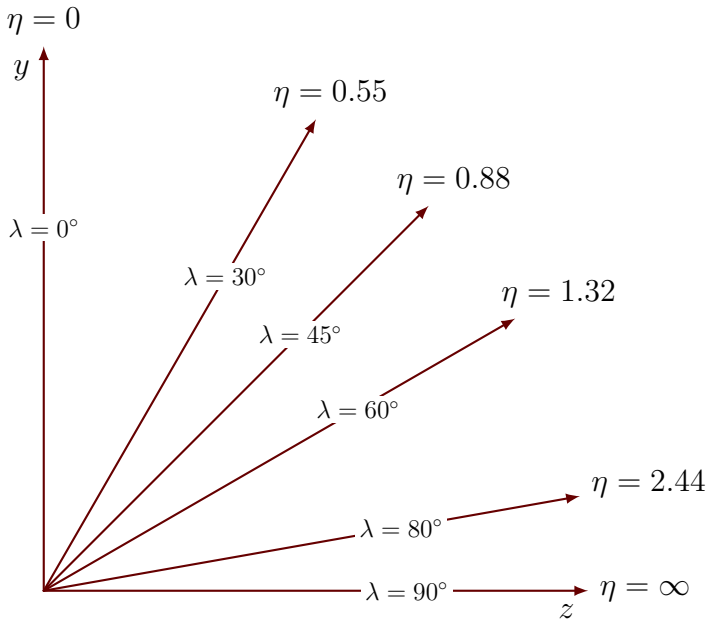


Figure 4.2: Pseudo-rapidity and $\tan(\lambda)$ for different values of λ.

Besides track inclination, the occupancy (also called multiplicity) of the TPC can have a significant influence on the dE/dx signal due to an increased probability of overlapping tracks, which influence their mutual charge measurement (pileup). Furthermore, the quality of the dE/dx signal increases with the number of clusters on the pad-rows on which a charge is collected. Higher number of clusters decrease the statistical uncertainty and fluctuations in the truncated mean calculation of the dE/dx measurement.

Further corrections have to be applied for low-momentum particles. In particular, particles with momenta below 0.1 GeV/c show strong deviations from the expected Bethe-Bloch curves. Here, the energy loss has a significant impact on the momentum of the particle. For kaons and protons at such momenta, the dE/dx rises steeply ($1/\beta^2$) as momentum decreases. Additionally, due to a curvature radius which decrease with momentum, the effective track length of a particle per pad row increases and significantly impacts the measured dE/dx signal. Low momentum corrections have to be applied below 1 GeV/c and are typically performed down to 0.15 GeV/c, limited by the available statistics.

Since the momentum of a track is significantly modified as it passes the ITS and inner field

cage of the TPC, the momentum measured at the inner wall of the TPC is used for further analysis. Residual tension between the Bethe-Bloch parametrization and data can predominantly be explained by detector effects, as mentioned above. This results in the necessity for a multidimensional correction in order to gain reliable particle identification information from the TPC.

4.2 Clean sample selection

In order to provide multidimensional corrections for the detector effects, it is necessary to extract data samples for which the particle identity is known a priori since these samples can be selected cleanly and hence the residual tension between the measured and predicted values for the specific energy loss is purely induced by detector and environment effects. Clean selections can be conducted for several particle species as they emerge from decays of other unstable particles (so called V0 particles) or can be identified by detectors in regions of kinematic separation.

4.2.1 V0 selection

Unstable particles emerging from high-energetic collisions can decay into stable particles. The initial particles are called mother particles, while the decay products are called daughter particles. For the investigated decays, the mother particle is uncharged and is hence not seen in the detector, while each daughter particle carries charge and can thus be detected. The V-shaped decay topology of the daughter tracks in combination with the zero charge of the mother particle gives the decay its typical name, V0 decay.

The Λ (uds quark combination) baryon or K_S^0 ($\frac{d\bar{s}-s\bar{d}}{\sqrt{2}}$ quark combination) meson are examples of such unstable particles. They decay into

$$\Lambda \to \pi^- p \quad , \quad \bar{\Lambda} \to \pi^+ \bar{p} \quad , \quad K_S^0 \to \pi^+ \pi^-$$

However, a similar conversion called pair production can also occur for highly energetic photons

$$\gamma + Z \to e\bar{e} + Z$$

where the product is an electron-positron pair and Z is the nucleus of a molecule which participates for energy and momentum conservation purposes.

Since the daughter particles emerge from charge-neutral mother particles which are not detectable with the detectors of the inner barrel (no electromagnetic interactions with the detector material), their electric charges are always opposite to each other. Hence, their tracks curve in opposite directions in the magnetic field, emerging from the same (secondary) vertex, making them identifiable by their topological separation.

Several physical variables obtained by measurements and the reconstruction procedure can then identify clean samples of V0 particles and their corresponding daughter tracks. Commonly, kinematic variables obtained from the reconstruction are used, such as a selection criterion on the reconstructed invariant mass of the mother particle. For this, the reconstructed

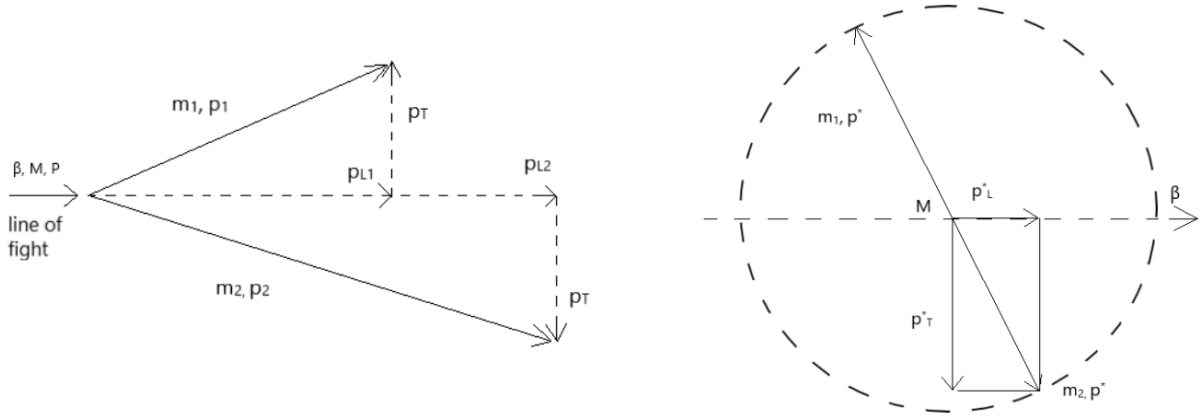
mass of the mother particle is compared to the expected mass and accepted if it is found within a (user-defined) range of validity, otherwise rejected.

Another common selection criterion for clean V0 selection is based on the *Armenteros-Podolanski* variables, which employs a selection on the kinematic variables α and q_T defined as

$$q_T \coloneqq p_T \tag{4.1}$$

$$\alpha \coloneqq \frac{p_L^+ - p_L^-}{p_L^+ + p_L^-} \tag{4.2}$$

where p_L^+ (p_L^-) is the longitudinal and p_T^+ (p_T^-) the transverse momentum of the positively (negatively) charged daughter particle in the direction of flight of the mother particle, see figure 4.3. In the reference frame of the mother particle it holds true that $p_T^+ = p_T^- \coloneqq p_T$.



(a) LAB (b) CM

Figure 4.3: Representation of the Armenteros-Podolanski variables α and q_T in the laboratory ((a) LAB) and the center-of-mass ((b) CM) frame [12].

Using the four momentum conservation and the ultra-relativistic approximation $\beta \to 1$, an equation for the kinematic occurrence of the V0 particles can be derived [12]

$$\frac{(\alpha - \alpha_0)^2}{r_\alpha^2} + \frac{q_T^2}{p^{*2}} = 1 \tag{4.3}$$

where

$$\text{I.} \quad \alpha_0 = \frac{m_1^2 - m_2^2}{M^2} \tag{4.4}$$

$$\text{II.} \quad r_\alpha = \frac{2p^*}{M} \tag{4.5}$$

$$\text{III.} \quad p^{*2} = \frac{1}{4M^2}(M^4 + m_1^4 + m_2^4 - 2M^2(m_1^2 + m_2^2) - 2m_1^2 m_2^2) \tag{4.6}$$

This forms an ellipse in the (α, q_T)-space. If the daughter particles have the same mass (e.g. for K_S^0), the equations simplify since $\alpha_0 = 0$. Accordingly, the ellipse of the K_S^0 particle will be centred around $(\alpha, q_T) = (0, 0)$ while for the decay of the Λ and $\bar{\Lambda}$ particles, the centre of the ellipse will be offset with

$(\alpha, q_T) = (\alpha_0(M = m_\Lambda, m_1 = m_p, m_2 = m_\pi), 0)$ for Λ

and

$(\alpha, q_T) = (\alpha_0(M = m_\Lambda, m_1 = m_\pi, m_2 = m_p), 0)$ for $\bar{\Lambda}$.

The expected ellipses can be clearly recognized in data (see figure 4.4) and a selection can be applied accordingly.

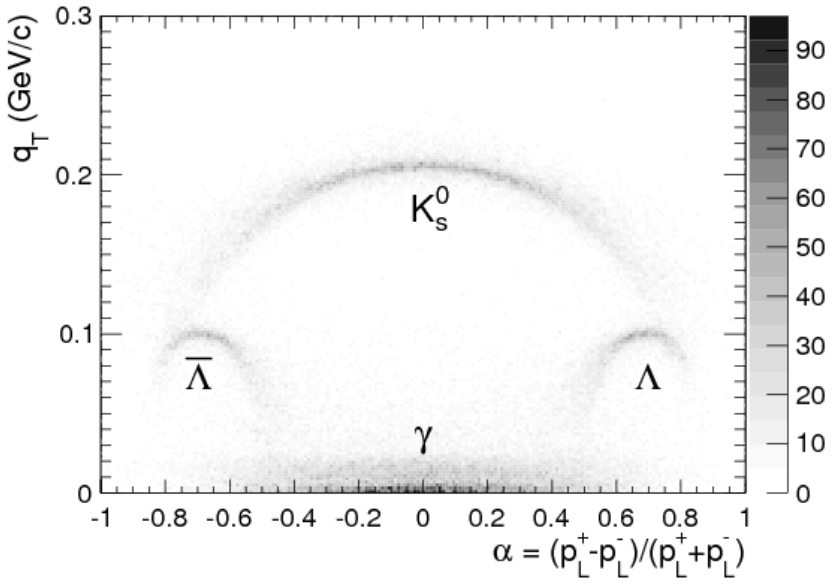


Figure 4.4: Plot of the Armenteros-Podolanski variables with reconstructed mother particles [13].

Since a decay occurs from highly energetic mother particles, the daughter particles will have the highest proportion of their momentum in the direction of flight of the V0 particle from which they originate. Reconstructing the angle θ_{PA} between the momentum vector of the mother particle and the line connecting the primary vertex with the V0 decay-vertex can bring significant improvement for the selection of clean samples. A selection on the $\cos(\theta_{\text{PA}})$ is typically performed, where the index "PA" stands for the abbreviation of the name of the angle θ between the direction of flight of the mother particle and its reconstructed momentum, the pointing angle. An illustration of a V0 decay, together with $\cos\theta_{\text{PA}}$ is shown in figure 4.5.

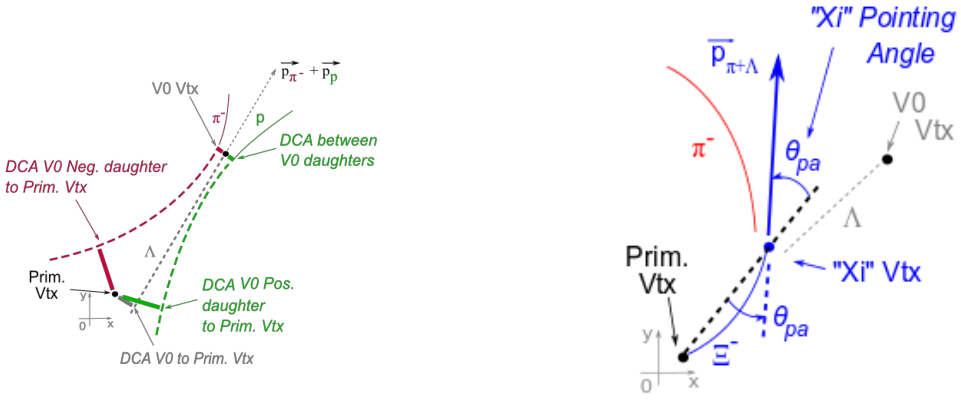


Figure 4.5: Topological V0 decay and relevant kinematic variables for particle identification

4.2.2 Selection by detection

Besides the clean samples gathered from the decay of V0 particles, different detectors can kinematically separate particle species. This depends on their separation power and the quality of their calibration. Typically, TPC and TOF selections are used to extract clean samples, where their separation power can be seen in figure 4.6.

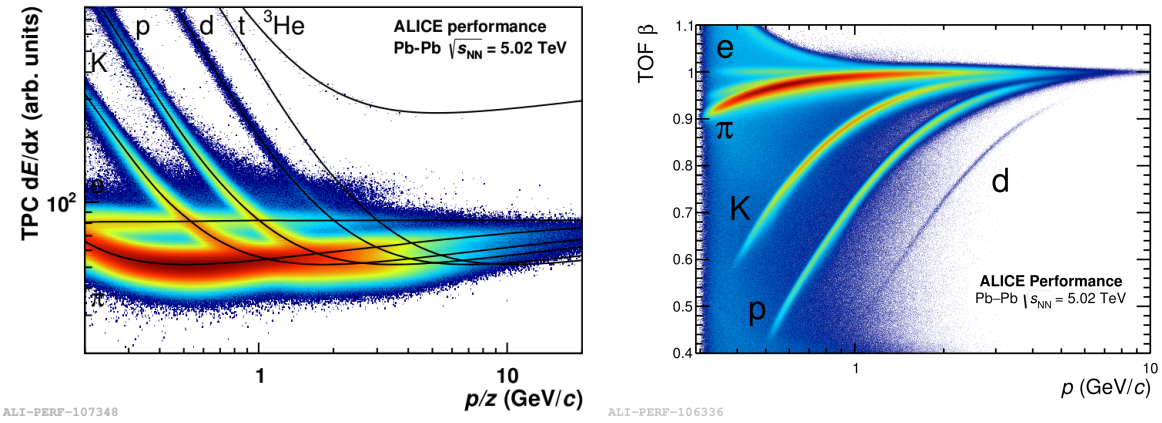


Figure 4.6: Separation power of the TPC (left) and TOF (right) detector

Given a theoretical prediction for the expected kinematic region in which particles are detected, a fit to data can be performed and the particle species is determined through an Nσ selection.

For the TPC, such a selection is performed on the measured dE/dx signal with the Bethe-Bloch as the theoretical prediction for each species. Using the parametrization of the Bethe-Bloch with its corrections as $\mathrm{dE/dx}_{corr}$ and an estimation of the spread of data as σ_{exp}, an Nσ cut can be applied as

$$N = \frac{\mathrm{dE/dx_{meas}} - \mathrm{dE/dx_{corr}}}{\sigma_{exp}}. \tag{4.7}$$

Typically, the TOF-only selection is used up to $p = 1.1$ GeV/c for light particles like muons, pions and electrons and up to even higher momenta for heavier particles like kaons, protons, and deuterons.
TPC selections can typically be used for pions up to a momentum of 0.5 GeV/c, kaons up to 0.3 GeV/c and protons up to 0.6 GeV/c, at higher momenta a TPC + TOF selection has to applied. Exact selection criteria can vary for individual datasets.

4.3 Mean correction and uncertainty estimation in Run 2

Once clean samples have been selected, a correction to the initial parametrization of the Bethe-Bloch function can be performed. During Run 1 and 2 of LHC, corrections were performed on a per-dimension basis by fitting piece-wise polynomial functions to identified clean samples. In particular, this includes corrections for a dependence on $\tan(\lambda)$ and the occupancy of the TPC, as well as low momentum corrections. The corrections were applied on a factorization approach, where it is assumed that if corrections are small, they can be applied to the Bethe-Bloch parametrization as

$$\left\langle \frac{dE}{dx} \right\rangle_{corr} = \left\langle \frac{dE}{dx} \right\rangle_{\text{param}} \cdot f(\tan(\lambda)) \cdot g(\text{MULT}_{\text{TPC}}) \cdot h(p_{\text{low}}) \tag{4.1}$$

where f, g and h are one-dimensional spline functions. Several iterations had to be performed in order to find a correction which calibrated the data on the order of few permille along the momentum axis. It can further obscure cross variable correlations by correcting projections in each dimension separately as a projection of all data points.

The uncertainty of data was estimated using a physically motivated fit function, for which the parameters were also adjusted using clean samples. In contrast to the mean corrections, the sigma estimation was in fact a multidimensional polynomial constructed to also capture cross-variable correlations. Vice versa, its predictions and potential to captured correlations was limited by the rigidity of the polynomial approach.

V Analysis and results on particle identification with neural networks

5.1 Outline of the research conducted in this thesis

Particle identification with the ALICE TPC is one of the most powerful tools in many analyses performed within the ALICE collaboration. Several steps have to be conducted in sequence in order to gain reliable PID information from the measured dE/dx signal.
At first, the parametrization which is used to describe the mean dE/dx signal in the (p, dE/dx)-space without corrections for detector and environment effects has to be calibrated. A set of initial parameters has to be determined from a fit to data and represents a parameterized theoretical prediction of the mean dE/dx signal based purely on the measured momentum and charge of the particles.
Two approaches were chosen in this thesis, depending on the availability of data. If clean samples are not available, a hyperparameter optimization (HPO) framework was constructed which does not require initial particle identities. Given a region of phase-space in which a sufficiently good set of parameters can be found, the HPO framework assigns a particle species based on proximity to a Bethe-Bloch curve and calculates local density maxima or mean values (both modes are available). Based on the found values, the points are approximated with least square minimization. The parameters improve over the iterations of the framework. If clean samples can be selected, a gradient descent method was implemented to approximate the best possible parameters. Gaussian profiles are fitted in bins of $\beta\gamma$ to data and the loss-function is designed based on least square fitting to the mean values of the fitted Gaussian distributions.
After this, detector effects and further deviations have to be calibrated by applying a correction of the mean dE/dx in a high-dimensional space spanned by observables, which can impact particle identification (see chapter IV). Additionally, the width of the particle distributions has to be estimated in high dimensions. Using the truncated mean, the final distribution of data around the mean dE/dx value follows in good approximation a Gaussian distribution in dE/dx and as such the uncertainty can be estimated by the standard deviation.
In this thesis, both the mean correction and estimation of the standard deviation are performed with neural networks with high precision which are fitted to a cleaned dataset of identified particles obtained from V0 selections.

Since the O^2 software framework was still in the commissioning phase during the writing of this thesis, the approaches were tested on data collected in Run 2 of LHC. In particular, the clean sample selection is demonstrated on the LHC18b dataset, which was converted into a Run 3 compatible data format and could thus be processed by the O^2 framework. LHC18b follows an internal nomenclature of the ALICE experiment and stands for a dataset which was recorded in April 2018. The data was recorded at a centre of mass energy of $\sqrt{s} = 13$ TeV for proton-proton collisions with ≈ 200 a million recorded minimum bias events (no trigger, i.e. selection, was applied for the data taking).

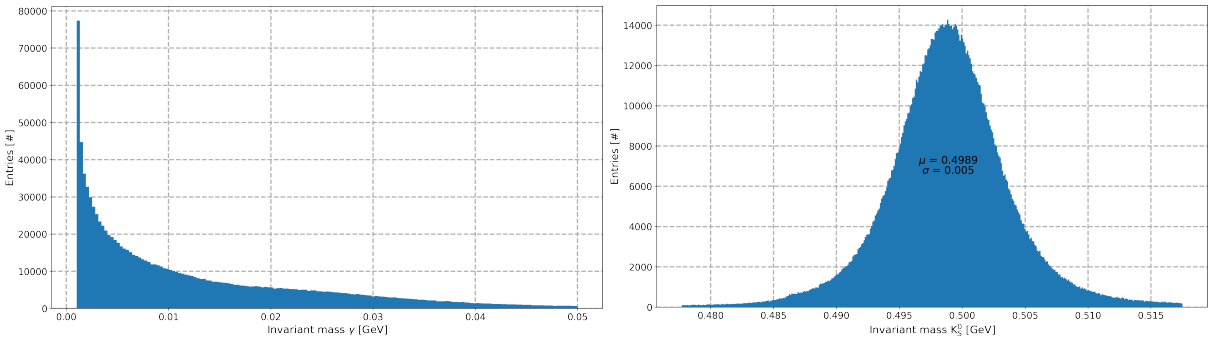
However, in order to provide a direct comparison to the fit performance in Run 2, the clean samples produced during Run 2 for the LHC18b dataset (so called filtered trees) were used to uncover shortcomings and benefits of the chosen approaches. The initial parametrization was then determined on cleaned data and Run 3 data without selection criteria, while the final neural network corrections were tested on the clean samples and compared with the results from the Run 2 calibration.

5.2 Clean sample selection

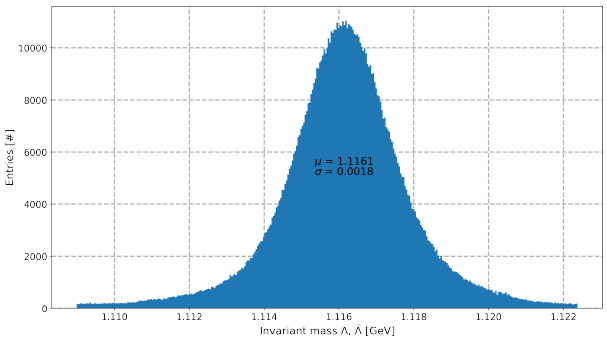
As shown in the previous section, reconstructed V0 particles can be identified clearly by the topology of their decay and selected by passing selection criteria on kinematic variables. The major criteria applied to data are demonstrated in this section on the LHC18b converted data and were performed in the O^2 software framework.

Selection criteria on the invariant mass and pointing angle

Based on the four-momenta of the daughter particles, the four-momentum of the mother particle can be reconstructed and hence the invariant mass is calculated based on the mass assumption and momenta of the daughter particles. The reconstructed masses can be collected in histograms, where clear peaks at the expected masses of the mother particles (Λ, K^0_S) are visible, see figure 5.1, (a)-(c).



(a) (b)



(c)

Figure 5.1: Reconstructed invariant mass distributions for γ (top left), K^0_S (top right) and Λ / $\bar{\Lambda}$ (bottom center).

A tight selection on the reconstructed invariant mass of the mother particle can be applied without biasing the momentum distribution of the daughter particles. Typical selection criteria are chosen between $2-3\sigma$ based on the width of the distributions.

On the contrary, a tight selection on the $\cos(\theta_{PA})$ (pointing angle) can bias the resulting momentum distribution of the daughter particles by shifting it towards higher momenta, but can also bring a significant improvement for the purity of the clean selection. The criterion

is typically chosen to be at $\cos(\theta_{PA}) \geq 0.999$ but still on the plateau of the distribution. The distribution of the $\cos(\theta_{PA})$ for all decays mentioned above is shown in figure 5.2.

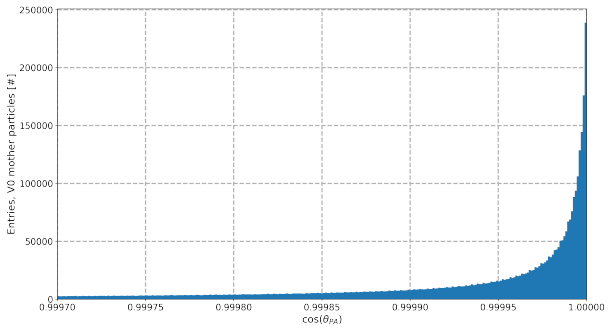


Figure 5.2: Distribution of the $\cos(\theta_{PA})$ combined for all decays.

Armenteros-Podolanski

The longitudinal and transverse momentum of the daughter particles from a Λ or K^0_S decay result in ellipses in the (α, q_T)-space for the location of the mother particle. However, the ellipses contain cross-over regions, such as the ones of Λ and $\bar{\Lambda}$ overlapping with K^0_S at $q_T \approx 0.11$ GeV/c. Thus, the kinematic overlap regions between the particle species are excluded to reduce misidentification errors. Similarly, γ-particles are found close to $\alpha \approx 0$ and thus show an overlap with Λ and $\bar{\Lambda}$ ellipses, which are then excluded accordingly. Figure 5.3 shows the clean selection based on the applied Armenteros-Podolanski selections (selections on invariant mass and $\cos(\theta_{PA})$ are already applied).

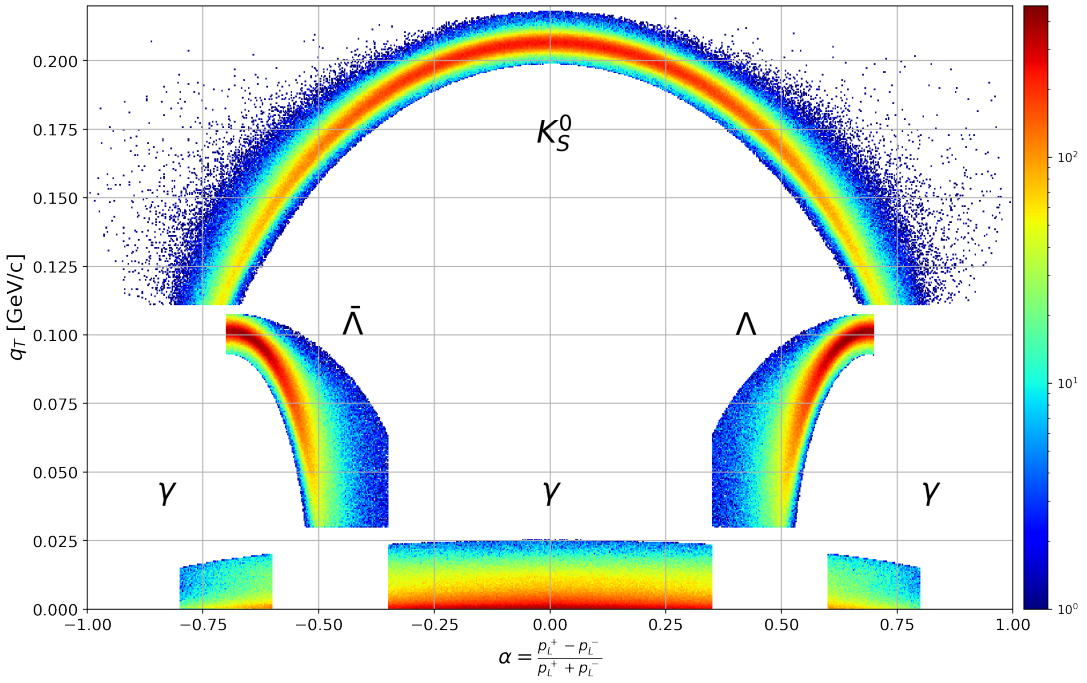
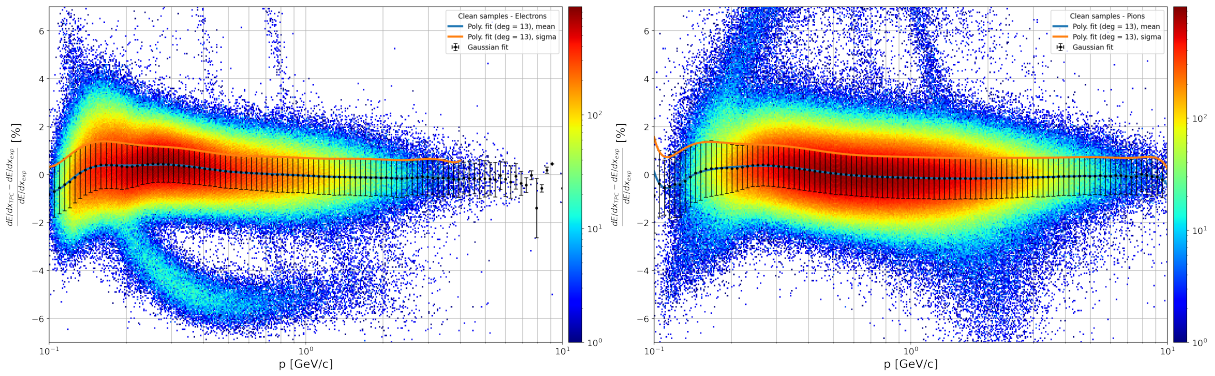


Figure 5.3: Pure V0 selection based on the Armenteros-Podolanski selections for the LHC18b converted dataset together with applied selection criteria on the invariant mass and $\cos(\theta_{PA})$ (this work).

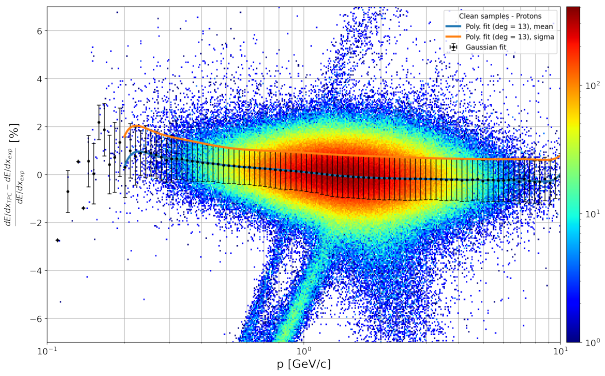
5.2.1 Purity of the obtained data

Residual contamination of the pure samples can be quantified as a comparison between the contamination and clean samples by using a sufficiently good initial calibration and an Nσ selection based on an estimation of Gaussian fitting in the (p, dE/dx) space.
The mean and standard deviation of each particle distribution on bins in momentum and the resulting points were fitted with a polynomial of 13th order to capture all features and fluctuations. This allows the estimation of impurity in regions where there is clear, kinematic separation in the measured dE/dx distribution of the TPC. For pions in the minimum ionizing region and protons at low momenta, the contamination was found in both datasets to be approximately 1%. This shows that a similar purity is achieved, however the implementation of selection criteria on kinematic variables and their specific values is a topic of investigation at the time of writing this thesis.
Potential contamination can be explained by suboptimal or missing kinematic selections and by misidentification of certain decays (e.g. electron-positron pairs can be misidentified for the decay of a Lambda baryon). The dE/dx distributions as a ratio to the Bethe-Bloch parametrization from Run 2 together with the fitted polynomials to determine the purity are shown in figures 5.4 (a-c).



(a)

(b)



(c)

Figure 5.4: Distributions of the TPC dE/dx as a ratio of the Bethe-Bloch function against the measured momentum for electrons (a), pions (b) and protons (c) together with Gaussian fits and polynomials for assessing the purity.

5.3 Initial calibration of the Bethe-Bloch parameterization using hyperparameter optimization

The parameters for the Bethe-Bloch function have to be determined from a fit to data. However, certain preconditions determine the optimal method for a fit. One possible approach is to perform a simple least-squares fitting of the parameters based on a cleaned dataset. However, especially at the start of Run 3 of LHC, not all detectors are calibrated with sufficient precision (e.g. TOF calibration not sufficient, ITS alignment missing) which requires an approach that does not rely on any initial PID information. Even if a selection with these detectors can be made, the assigned PID information might not be reliable enough for a fitting procedure.

The initial parametrization will receive additional corrections for detector effects and hence exhaustive searches (e.g. grid-based approaches) are not taken into consideration since they are computationally expensive. Furthermore, the parameters can at best be estimated to a certain region of phase-space, which requires an algorithmic approach that investigates the phase-space.
Hence, a novel hyperparameter optimization framework called Optuna [14] is employed. It allows efficient parsing of the phase-space using pruning methods for unpromising trials evaluated on various internal metrics.

The construction of the score function which defines the goodness of a parameter set is a critical step and is based on the search for density maxima or the estimation of mean values of data in a local neighbourhood. Particle identities are assigned dynamically at each iteration and by proximity to the nearest Bethe-Bloch curve in dE/dx. The underlying assumption is that in the vicinity of a good parameter set, the Bethe-Bloch curves overlap with the local maxima / mean values of data density for each assigned particle species. An example of local maximum searches (grey points) based on the initial parametrization from Run 2 and the dynamic assignment of particle identities for the LHC18b dataset can be seen in figure 5.5.

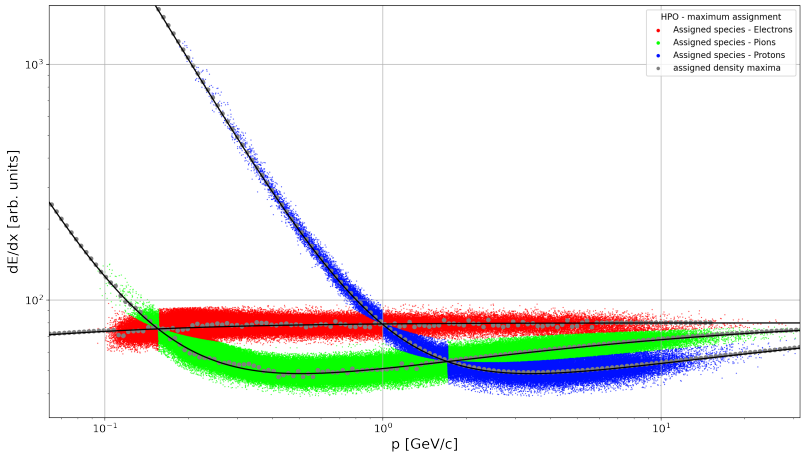


Figure 5.5: Locally assigned particle species and density maxima based on binning in $\beta\gamma$ and dE/dx.

The score function for the fitting procedure is purely based on the assigned maxima / mean values and gets evaluated at every iteration. The final score is calculated with the sum of square residuals from these points to the predicted Bethe-Bloch curves for all bins

$$\text{Score}_{\text{optim}} = \sum_{i \in \text{species}} \sum_{j \in \text{bins}(\beta\gamma)} \frac{1}{\text{N}_{\text{bins}}} \cdot \left(\frac{\max(\rho_{ij}) - BB(\beta\gamma = j)^2}{BB(\beta\gamma = j)} \right)^2 . \tag{5.1}$$

A fit can succeed if the data density is high enough and a sufficiently wide range of values in $\beta\gamma$ is covered. In particular, the amount of data is crucial for the maximum-density approach, since otherwise the found values for the maxima are subject to large fluctuations based on the distribution of single tracks.
If no tracks or too few tracks are found in a bin of $\beta\gamma$, then the density maximum is set to overlap with the Bethe-Bloch value at this point. Like this, the score is not artificially inflated by fluctuations in data or regions of low data density.
In order to reduce fluctuations further, a region of validity around the Bethe-Bloch curves is defined based on a local sigma estimation, in which the maximum or mean value is calculated.

The performance is first demonstrated on the cleaned filtered tree of the LHC18b dataset and compared to the parametrization found in Run 2. Based on the availability of clean samples, electrons, pions and protons are used for the fit procedure. The ratios $((dE/dx)_{meas.}$-$(dE/dx)_{exp.})/(dE/dx)_{exp}$ for each particle species are shown in figure 5.6. The momentum was divided into bins and mean and sigma values were calculated in each bin. Since final corrections are not applied here, a fit on the precision of few percent is sufficient.
Overall, the performance of the HPO algorithm without particle assigned is similar to the parameters found with the method in Run 2. Qualitatively, it is notable that for pions, the HPO algorithm performs better than the method used in Run 2 at the minimum ionization region, while for protons the binned mean values show a better fit for the Run 2 parametrization. For electrons, the results from the HPO algorithm are preferable at higher momenta ($p \geq 0.5$ GeV/c). Overall, both mean and maximum estimation perform similarly well and result in parametrizations which capture the underlying data distributions on the order of few percent.

Comparing the results of the loss function, the mean estimation achieves an overall lower score of 0.001866 compared to the maximum estimation with a score of 0.03186. This indicates higher fluctuations of the assigned maxima over the bins in $\beta\gamma$. From this, it can be concluded that the maximum estimation needs stronger regulation for the minimum number of points per bin in order to minimize statistical fluctuations of the density maxima. It further implies that a higher amount of data is needed in order to obtain stable and thus reliable results.

In order to investigate the performance with known particle identities, the fit was reproduced with an initially fixed particle identity for each track based on the identity provided in the clean samples. The results are shown in figure 5.7. Here it can be clearly seen that the mean estimation performs better than the maximum estimation for protons, while the performance over pions and electrons is similar for both methods. This can indicate an excessively granular binning was chosen for the maximum estimation to work.

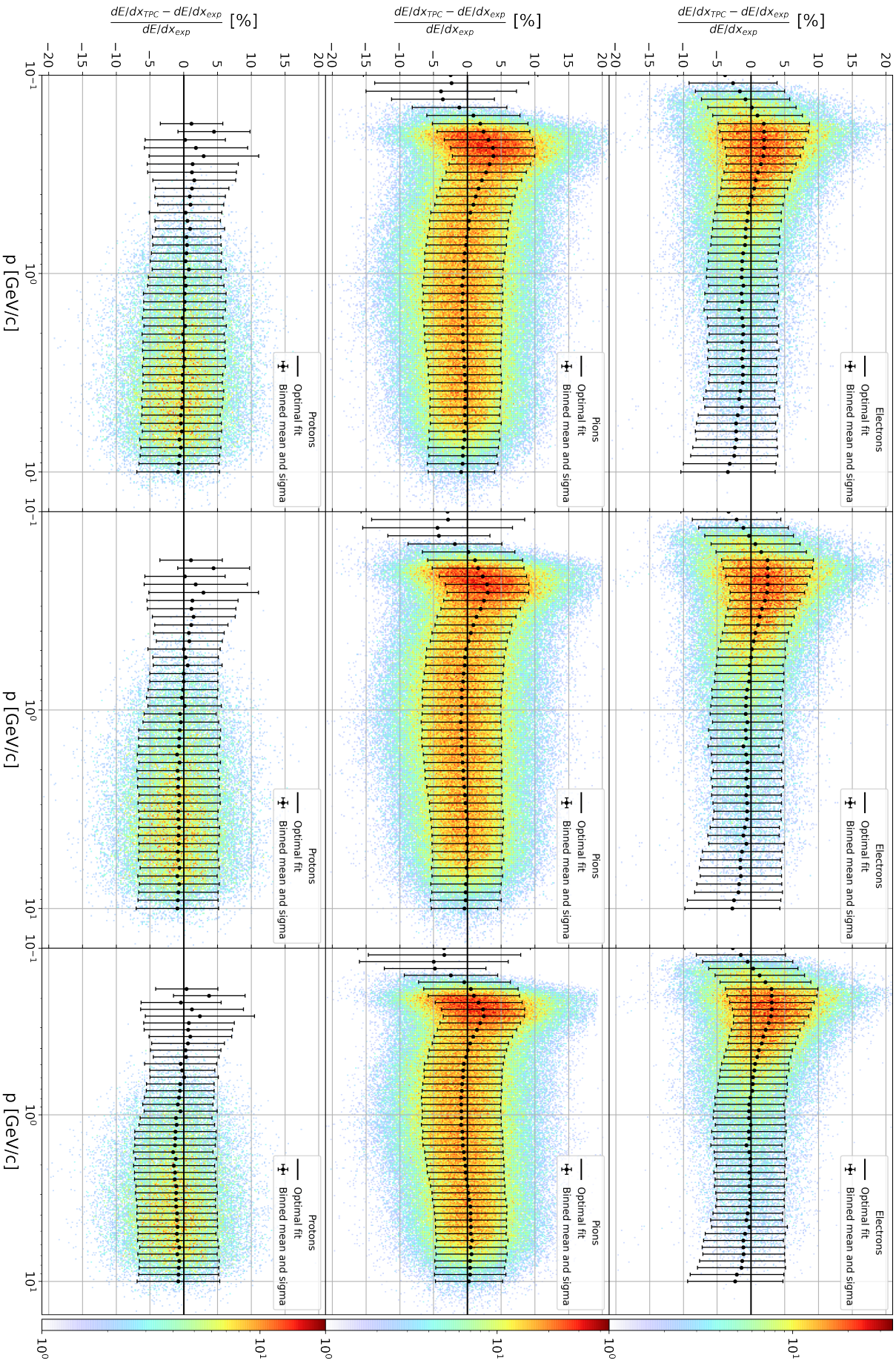


Figure 5.6: HPO performance for unfixed particle identities compared to the Run 2 parametrization for identified electrons, pions and protons.

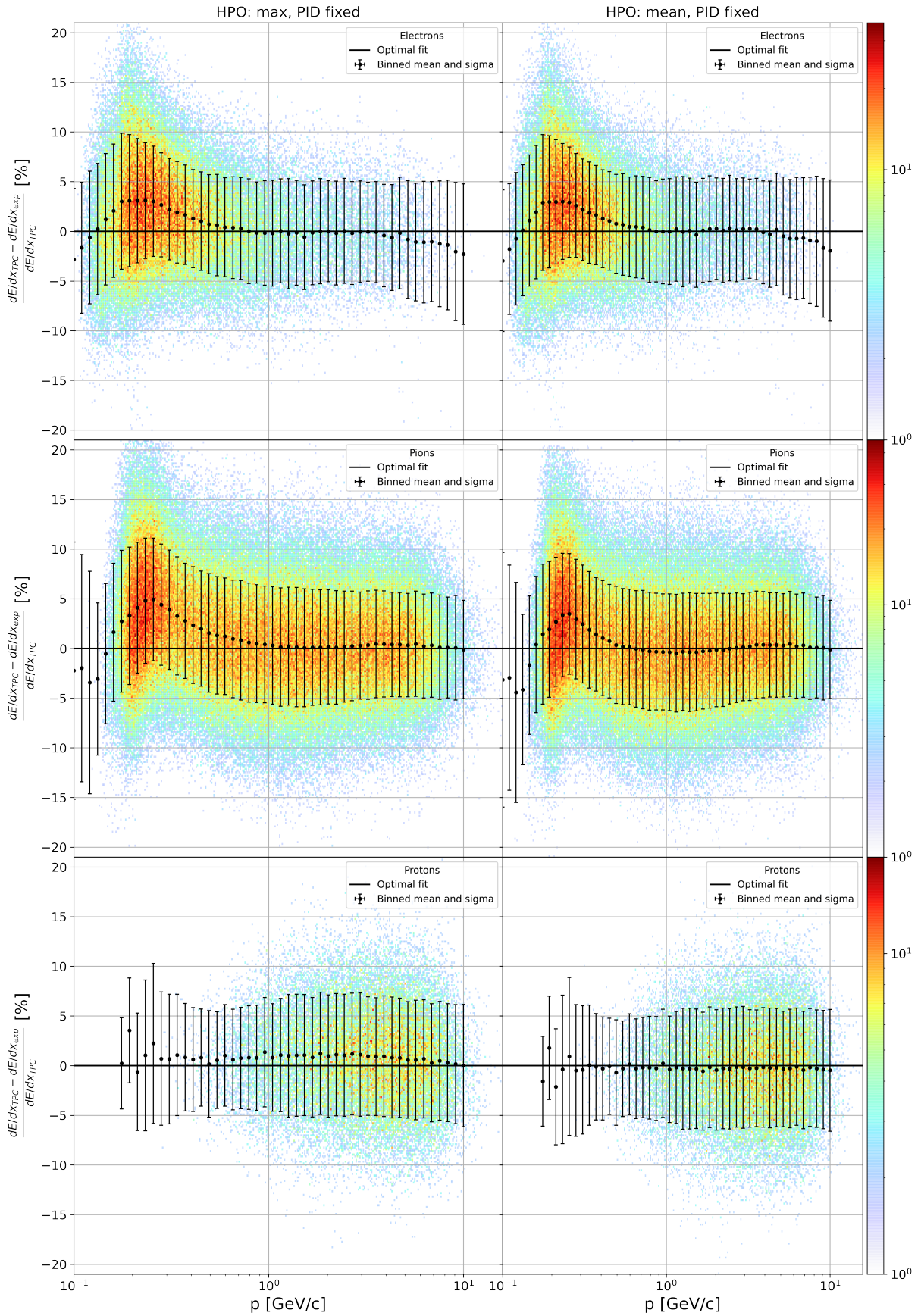


Figure 5.7: Comparison of the HPO performance for fixed particle identities with the Run 2 parameterization for identified electrons, pions and protons.

In comparison with the previously presented results from Run 2 (see figure 5.6) it can be noted that the HPO algorithm performs overall better and in particular the mean estimation algorithm performs better than 1% of deviation to the Gaussian fits in regions where the particle identity is kinematically separated from other particle species ($p(\pi) \geq 0.5$ GeV/c, protons in the full kinematic region, $p(e) \geq 0.5$ GeV/c).
By the output of the score functions, it can be noted again that the approach of calculating mean values performs more stably due to an overall lower loss score of 0.0191 compared to 0.0523 for the maximum estimation approach. Further investigations have to be conducted in order to determine an optimal working point and an optimal amount of data for both versions of the algorithm.

Besides potentially suboptimally tuned parameters, it can be noted that the hyperparameter optimization framework can perform at least similarly well if not better than the framework used in Run 2 and shows promising results for clean samples.
Finally, the resulting curves were compared to the parametrization from Run 2 to investigate the overall difference in different kinematic regions. The result is shown in figure 5.8 together with the approximated, data-covered regions for each particle species.

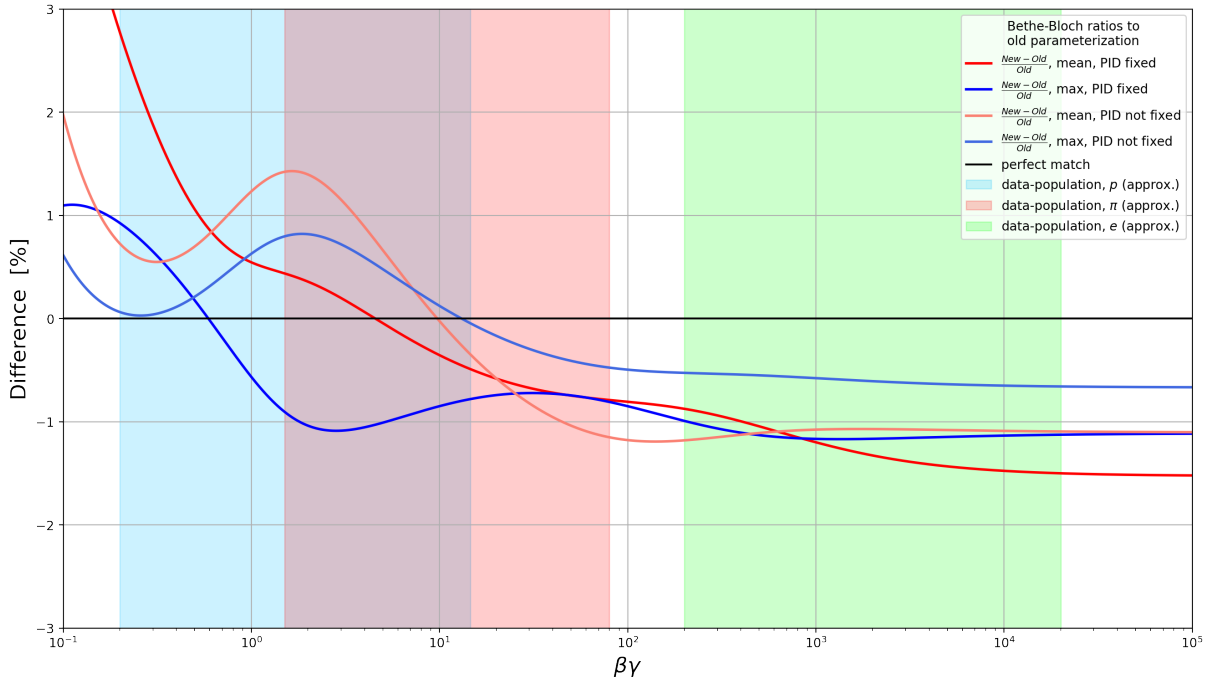


Figure 5.8: Difference in [%] between the Bethe-Bloch parametrization obtained in Run 2 and the parametrizations obtained from the HPO algorithm with different settings.

It can be noted that all parametrizations show a tendency towards higher values for dE/dx at the same values for $\beta\gamma$ for the low-$\beta\gamma$ region where protons dominate the fit and a tendency towards lower values than the parametrization from Run 2 for high values of $\beta\gamma$, close to the Fermi plateau where electrons dominate the data distribution.

5.3.1 Improvement on clean data with gradient descent

In order to improve the fit, gradient descent can be used for fixed particle identities on clean samples or with a particle assignment based on a sufficiently good initial parametrization from the hyperparameter optimization. Gaussian profiles are fitted to each of the assigned distributions in slices of momentum and hence, mean values can be estimated. With a sufficient number of bins, gradients can be estimated for all parameters and the fit can be conducted based on update steps to each of the parameters entering the parametrization. The results of the gradient descent are compared to the results obtained with the hyperparameter optimization in figure 5.9.

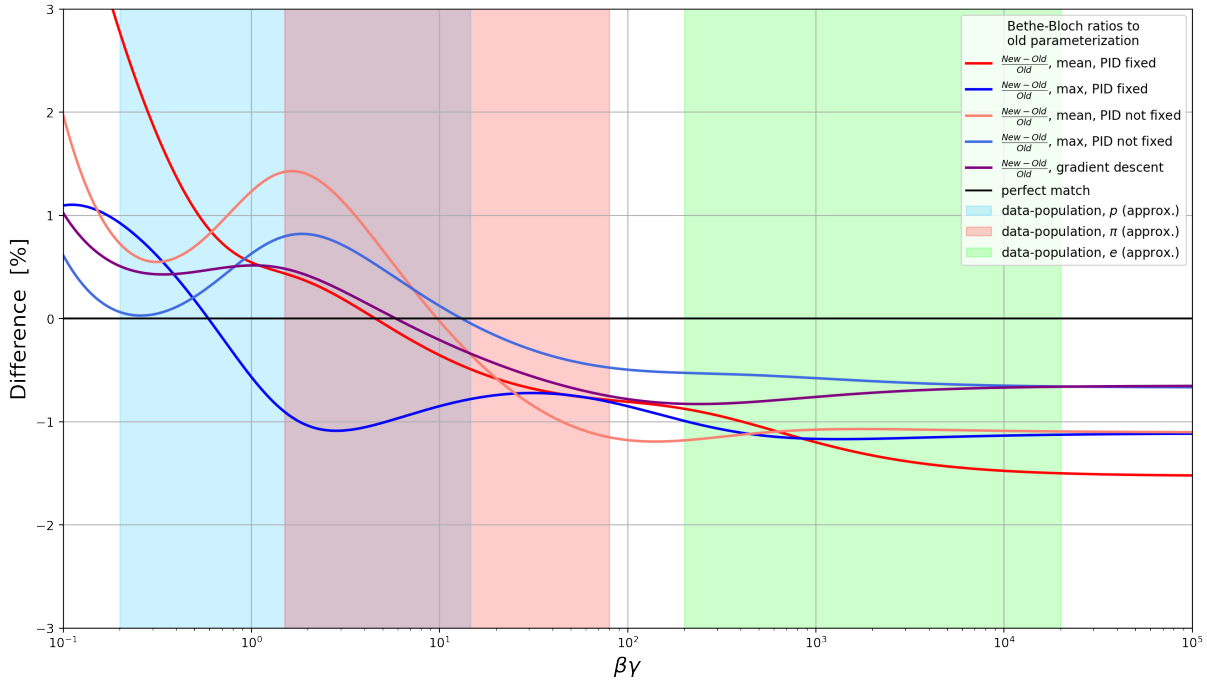


Figure 5.9: Gradient descent method together with the HPO results as a ratio to the Run 2 parametrization.

As can be seen, the gradient descent shows great similarities with the results found by the hyperparameter optimization. On the Fermi plateau, the parametrization shows agreement with the results from the HPO for the mean estimation and fixed particle identities. At low momenta however, the parametrization shows higher agreement with the results from Run 2 than with the mean estimation of the HPO. Since data density is rather sparse in this region, it can hardly be determined which curve fits the data better. Overall, all methods show good agreement within few per-cent, with the gradient descent being a further tool for improvement which can be applied after sufficiently good initial parameters were found by the hyperparameter optimization framework.

5.3.2 Performance on full data (LHC22f)

The possibility for applying the algorithm on contaminated data samples opens the opportunity to investigate data collected in Run 3 of LHC. Tracks can be propagated from the primary vertex to the TPC, giving reliable momentum and dE/dx information. Hence, the HPO algorithm was applied to perform the initial calibrations for LHC22f (and further datasets of the

LHC22 periods, the performance is demonstrated here on LHC22f). This dataset was taken in July 2022 with a record energy of LHC of $\sqrt{s} = 13.6$ TeV in proton-proton collisions. The result for data taken from a small fraction of the dataset are shown in figure 5.10.

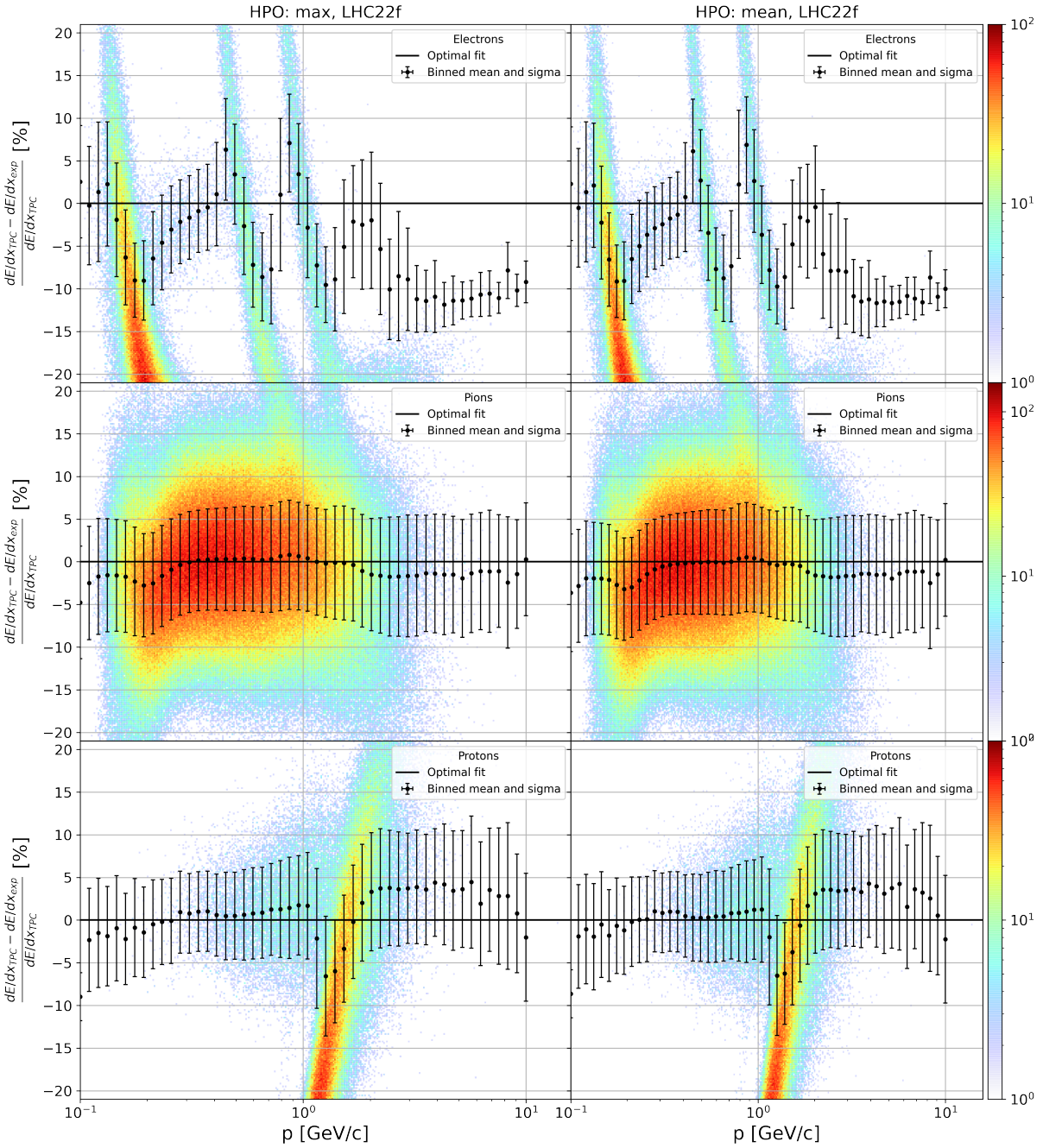


Figure 5.10: Performance of the mean and maximum density estimation on uncleaned data of the LHC22f dataset.

A noticeable difference between the data and the fitted curve is apparent for the electrons. Although data density is limited, the fit was performed before also on different datasets from Run 3 which all showed a similar behaviour. This could indicate that the wrong region of phase-space is used for the Bethe-Bloch parameters, or the curve itself cannot capture the features for low momentum electrons. A feature that can be observed in the data is that the

gradient for low momentum electrons is much stronger than what was observed in the Run 2 data. Especially close to the cross-over with pions, the electrons show disagreement with the fitted curves. Ultimately, this feature will be captured by the neural network corrections (see next section), but it remains a topic of investigation to test different regions of the phase space spanned by the Bethe-Bloch parameters.

Pions and protons are showing agreement in kinematic regions where they can be cleanly selected by their TPC signal. However, only with clean samples can the cross-over regions between different particle species be investigated and potential differences between the fit and the data be visualized.

It can be noted that both the mean and maximum searches perform similarly well. For electrons, the maximum search performs slightly better and shifts the curve towards lower values, thus centering more on the distribution. In contrast, the mean parametrization performs slightly better for low momentum protons. For pions, both parametrizations perform almost identically. This illustrates, that the minimum ionizing region (dominated by pions) is well captured by the curve, while the relativistic rise where particle species overlap (mainly pions, kaons and protons) still shows residual tension.

5.4 Mean correction with neural networks

After clean samples have been selected, a mean correction using neural networks is applied in this thesis. This represents a possible replacement for the per-dimension corrections applied by the spline fit used in Run 2.

The choice of the network architecture is based on the training performance (MSE score) and fit performance on final datasets. Since the networks will be applied in many analyses within the ALICE collaboration on (multi-)CPU powered machines, a trade-off between accuracy, computation time and memory consumption has to be made. For the purposes of the mean correction, a large fully connected network with 10 hidden layers and 12 neurons per layer is trained initially. This network learns an initial correction for the mean, but is computationally too expensive for a large scale inferencing. Hence, a second, smaller network (3 layers, 8 neurons per layer) is trained on the output of the larger model which can then be applied in the O^2 software framework at runtime.

The network is trained to learn the ratio $dE/dx_{meas.}/dE/dx_{exp.}$ where the expected signal is given by the Bethe-Bloch function with parameters determined from the initial fit. As an input, the network obtains properties of the tracks measured in the TPC.
These are

1. momentum at the inner wall of the TPC: p in GeV/c
2. tangent of the local track inclination angle: $\tan(\lambda)$
3. sign of the charge of the particle divided by the transverse momentum: $\text{sign}(q)/p_T$
4. mass (hypothesis) of the particle: m in GeV
5. multiplicity of the TPC normalized to 11000: $\text{MULT}_{\text{TPC}}/11000$
6. number of clusters of a track measured in the TPC normalized to the maximum number of clusters of 159 rows (Run 2) / 152 rows (Run 3) in the TPC:
   $\sqrt{159/\text{NCL}_{\text{TPC}}}$ (Run 2) ; $\sqrt{152/\text{NCL}_{\text{TPC}}}$ (Run 3)

The normalization for each variable is chosen to keep the values close to the range of [0,1]. The choice of the parameters is inspired by the most significant variables influencing the mean corrections and sigma estimations determined in Run 2. In particular, the dependence on momentum and the local track inclination showed significant dependencies throughout Run 2.

5.4.1 Mean correction values from training on cleaned data

Using the MSE as the loss function during training, the network learns the mean values of $dE/dx_{meas.}/dE/dx_{exp.}$ in the six-dimensional space spanned by the input variables and thus a correction to the Bethe-Bloch parametrization can be applied as

$$dE/dx_{corr.} = dE/dx_{exp.} \cdot \text{net}_{\text{mean}}. \tag{5.1}$$

In order to compare the neural network corrections directly to the corrections applied in Run 2, the same initial parameters as in Run 2 are being used for the training of the network. The

corrections are applied for both cases. The resulting distributions are binned in momentum and the mean is taken in each bin. The results can be found in the figures (5.11).

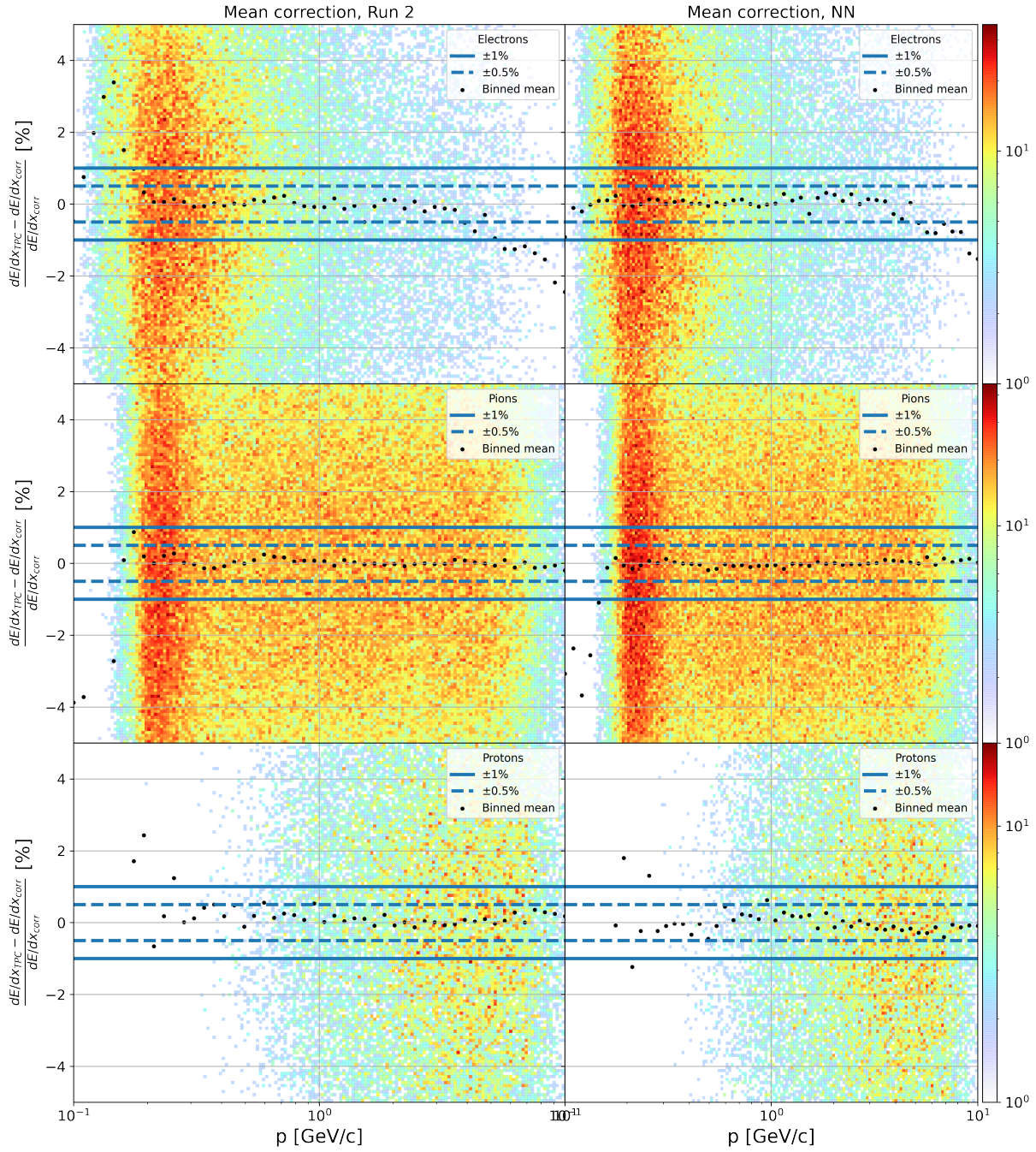


Figure 5.11: Mean corrections applied to the clean samples for the Run 2 and neural network approach.

As can be seen on the right-hand plots, the network correction results in fitted mean values better than 3‰ and performs overall similarly well as the spline fit from Run 2. Larger tension is observable for electrons at low momenta for the corrections from Run 2 since low-momentum corrections are not applied for $p \leq 0.15$ GeV/c. However, this region is well captured by the neural network and shows deviations of typically ≤ 3‰ from the expected values of the mean at 0. Slight tension can be observed for both methods for protons at lower

momenta, which could however also be due to limited statistics. The overall success of the small architecture of the neural network shows that the correction factors do not fluctuate strongly over the phase-space.

5.4.2 Comparison of η-map corrections

In order to compare the corrections on a per-dimension basis, the corrections in the ($\tan\lambda$, 1/(dE/dx$_{\text{exp}}$))-space are plotted (so called η-maps). This showed one of the most dominant dependencies throughout Run 2 and indicates whether similar corrections were found by the neural network as compared to the spline fit. The results are shown in figures 5.12 and 5.13.

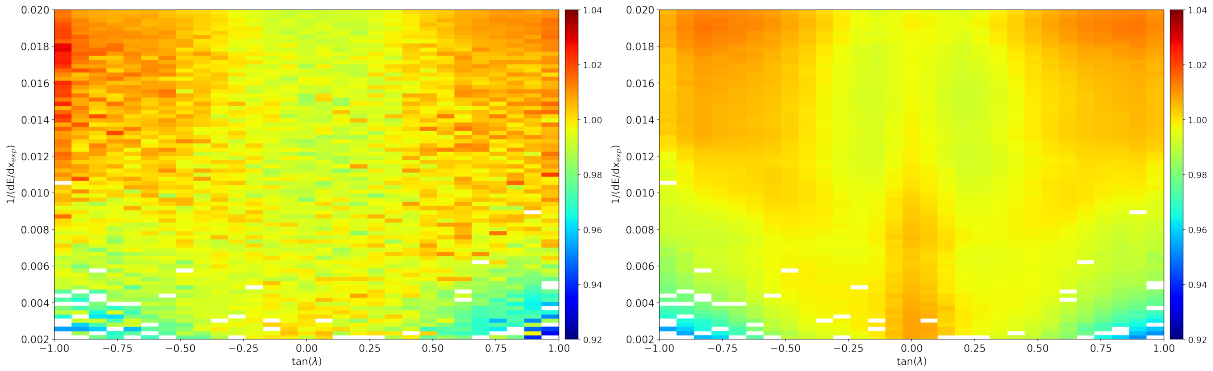


Figure 5.12: Mean corrections for protons in the ($\tan(\lambda)$,1/(dE/dx$_{exp}$)) space for the neural network (left) and the spline fit (right).

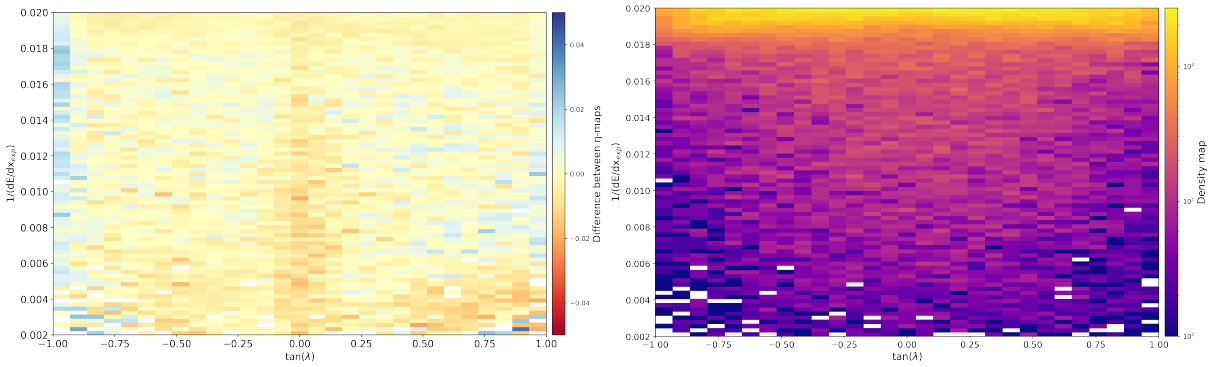


Figure 5.13: Difference between the η-maps (left) and the density of data points (right).

The comparison between the η-maps shows a smoother overall behaviour for the corrections from the spline approach. However, this can indicate that high-dimensional features or correlations are not captured by the two-dimensional approach, but can be seen by the fluctuations in the η-map for the network corrections. Overall, the corrections show similarities for both methods, which is also shown by taking the difference between the two η-maps (see figure 5.13, left). It is noticeable that around $tan(\lambda) \approx 0$ a structure is visible in the spline correction, but hardly noticeable in the neural network fit. This could indicate potential problems for small values of the input in the neural network fit.
However, further, more extensive testing would need to be conducted to investigate the behaviour in all corners of the phase space, which reaches outside the scope of this thesis.

5.4.3 Fluctuations of the mean correction values

In order to investigate the fluctuations of the correction values to the mean Bethe-Bloch parametrization, an ensemble of 15 neural networks was trained on the same dataset with identical settings. The only difference between the networks is the random sampling for training / validation data and the initialization of the weights and biases of each neuron which are randomly sampled from a standard normal Gaussian distribution. Based on the results, the standard deviation of the correction values to the mean parametrization can be determined and is shown for each species in figure 5.14.

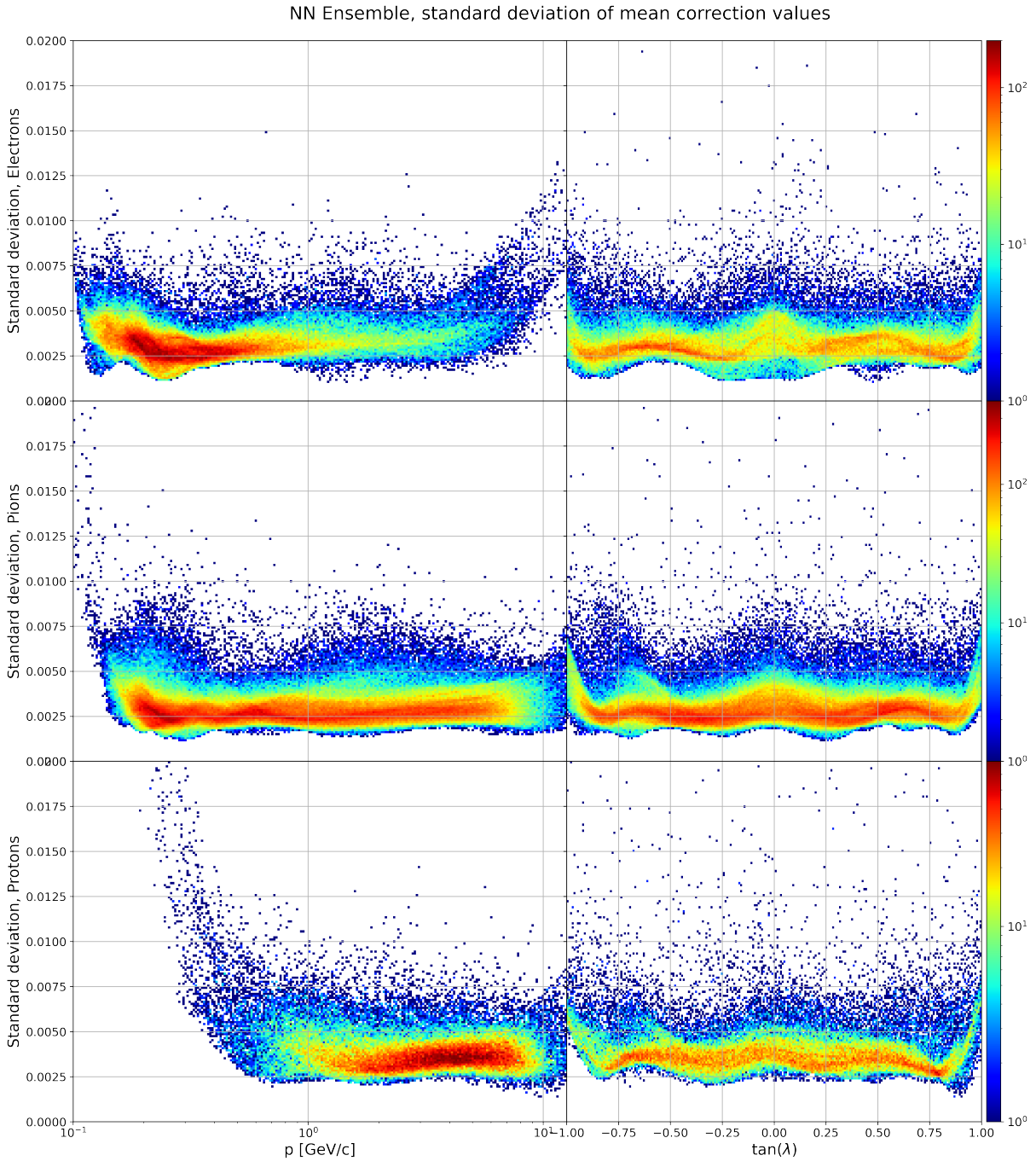


Figure 5.14: Deviation of the mean correction factors produced by a neural network ensemble for momentum and $\tan(\lambda)$.

From these plots it can be seen that low data-densities clearly impact the stability of the network fits (e.g. electrons at $p \geq 6$ GeV/c, protons at $p \leq 0.7$ GeV/c). It can further be observed that a large fraction of the correction values are typically found within 5‰ around the mean of the ensemble. This shows that the network fits will typically produce corrections with an accuracy of ≤ 5‰. The clearly visible outliers in these distributions (e.g. standard deviation $\geq 1\%$) were investigated based on a random selection and were found in regions of low data-density or at the corners of the phase space (e.g. at low and high momenta for each particle species). This further demonstrates the necessity for an equal distribution of data across the phase space of interest which can be achieved by e.g. down-sampling in p_T using the Levi-Tsallis distribution of particle yields since the overall statistics will cover the phase space well in Run 3.

Increasing tension can be observed for pions and protons in the region of $\tan(\lambda) \sim 0$. This could indicate that larger fluctuations in the dE/dx signal are apparent since $\tan(\lambda) \sim 0$ implies $\lambda \approx 0$ and hence this also correlates to tracks which get absorbed by the central electrode. Hence, missing statistics in this phase-space region could be an explanation of this phenomenon, but further investigations have to be conducted before a final verdict can be given.

The mean fluctuations could possibly be further suppressed by training the networks with more epochs. In the case demonstrated in this thesis, a learning rate scheduling together with decreasing batch-sizes and 200 training epochs were used to produce the neural networks with $\approx 6.2 \times 10^5$ data-points in total.

5.5 Sigma estimation with neural networks

The estimation of the uncertainty of the data is an essential part of particle identification, since it defines the significance with which a selection is made and is required to do Nσ selections on the final distributions. In case of the TPC PID, the data distribution for each particle species can be well approximated by a Gaussian distribution in dE/dx for every slice in momentum. The mean value and standard deviation change as a function of the observables that span the phase-space.
In order to estimate the standard deviation of the data, a reliable mean estimation has to be done in the first place. After this, the distribution can be centred accordingly (the mean of the distribution is at / close to 0 in the full phase-space) and the standard deviation can be fitted as shown in the following.

The Gaussian distribution $\mathcal{N}(x, \mu = 0, \sigma), x \in \mathbb{R}$ is considered. This corresponds to the centred point-density distribution with $x =$ dE/dx at every point in the considered phase-space (the mean correction has been applied) where the standard deviation should now be estimated. Taking the absolute values in x (i.e. $x =$ dE/dx) and normalizing the distribution yields

$$\mathcal{N}^*(x) \coloneqq 2\mathcal{N}(x, \mu = 0, \sigma)\theta(x) \tag{5.1}$$

where $\theta(x) = I_{\mathbb{R}_0^+}(x)$ is the indicator function and σ is the standard deviation of the original Gaussian distribution. The $\mathcal{N}^*(x)$-function in comparison to the original Gaussian distribution is shown in figure 5.15. The mean μ^* of this function can be calculated as

$$\begin{aligned}
\mu^* &= \int_{-\infty}^{\infty} x \cdot \mathcal{N}^*(x)dx \\
&= \int_{-\infty}^{\infty} x \cdot 2\mathcal{N}(x, \mu = 0, \sigma)\theta(x)dx \\
&= 2\int_{0}^{\infty} x \cdot \mathcal{N}(x, \mu = 0, \sigma)dx \\
&= \sqrt{\frac{2}{\pi}}\frac{1}{\sigma}\int_{0}^{\infty} x \cdot \exp\left(-\frac{x^2}{2\sigma^2}\right)dx \qquad \Big| \quad \nu \coloneqq \frac{x^2}{2\sigma^2} \quad ; \quad \frac{d\nu}{dx} = \frac{x}{\sigma^2} \\
&= \sqrt{\frac{2}{\pi}}\sigma\int_{0}^{\infty} e^{-\nu}d\nu \\
&= \sqrt{\frac{2}{\pi}}\sigma
\end{aligned} \tag{5.2}$$

Hence the neural network training can be performed on the centred distribution of absolute values of dE/dx. By using the MSE loss and fitting $\mathcal{N}^*(x)$, the standard deviation of the original dataset can be learned by a neural network where $\sigma(x) = \sqrt{\frac{\pi}{2}} \cdot \text{NN}(\mathcal{N}^*(x))$. The $\mathcal{N}^*$-distribution can be constructed from the centred dE/dx distribution by taking the absolute values, $\mathcal{N}^*(|\frac{\text{dE/dx}_{\text{meas.}} - \text{dE/dx}_{\text{exp.}} \cdot \text{net}_{\text{mean}}}{\text{dE/dx}_{\text{exp.}} \cdot \text{net}_{\text{mean}}}|)$. Figure 5.15 illustrates the location of the mean of the $\mathcal{N}^*$-distribution together with the original Gaussian distribution ($\sigma = 1$ was chosen for illustration purposes).

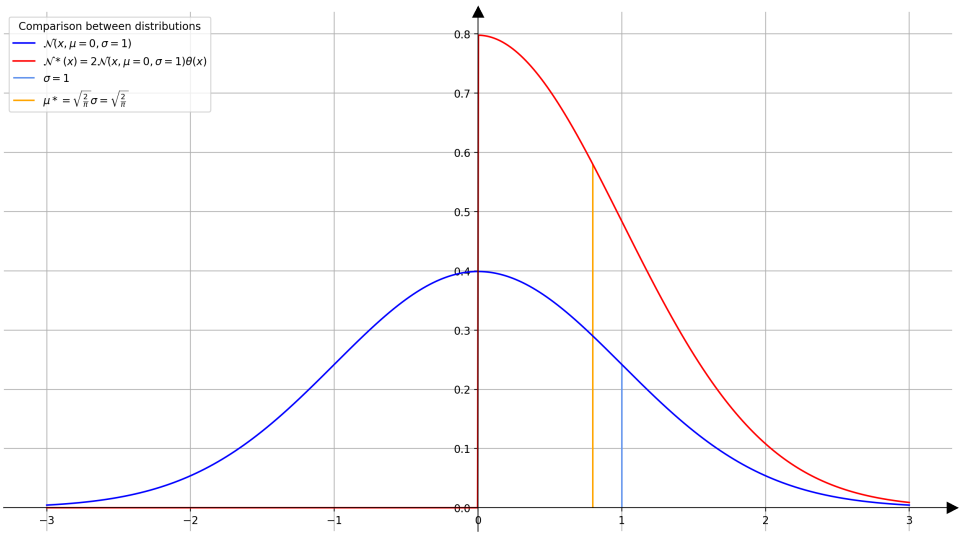


Figure 5.15: Comparison between original Gaussian distribution and the $\mathcal{N}^*$ distribution.

In order for the mean and sigma to be on similar scales and in order not to skew the neural network values by learning values with larger discrepancies, the final network is trained to learn the mean correction in the first output dimension and the mean correction + sigma estimation in the second output dimension. Like this, both dimensions that need to be learned are close to 1.
In figure 5.16, a representation of a fully trained network together with the connecting weights (line thickness) can be seen. Neurons are represented as circles, edges between neurons are represented with lines.

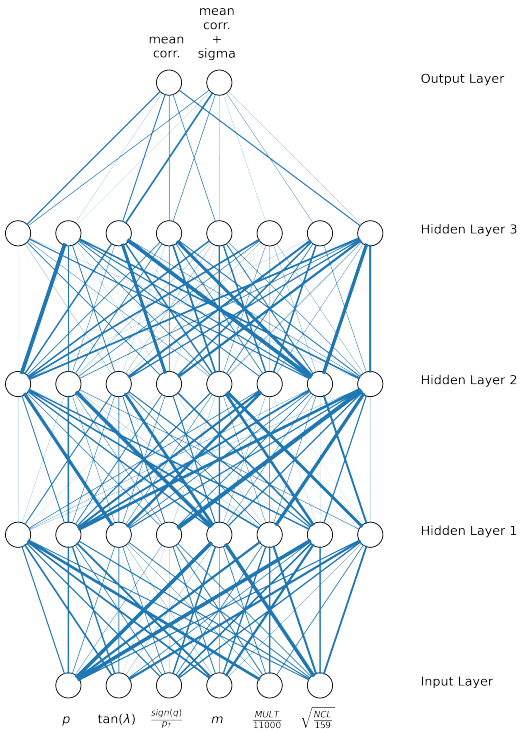


Figure 5.16: Neural network (final architecture) with the connecting weights represented by the line-thickness between adjacent layers.

5.5.1 Mean correction with sigma estimation applied to data

Using the mean correction from the neural network from the previous section and estimating mean and sigma values in slices of momentum results in figure 5.17.

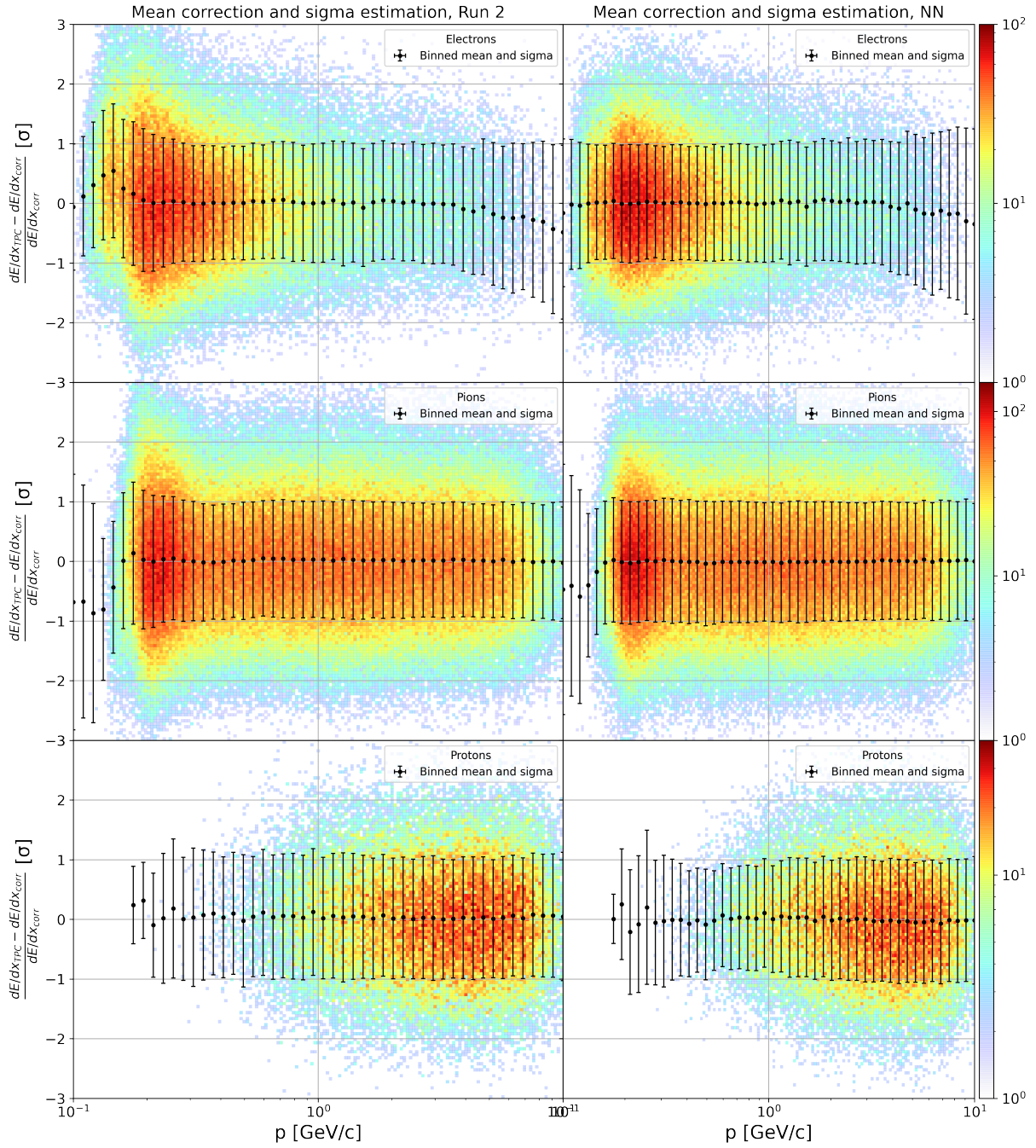


Figure 5.17: Nσ distributions for identified electrons, pions and protons with the method from Run 2 and the neural network application.

As can be seen from the plots, the Nσ distributions are captured well by both methods. However, the network shows an overall smoother behaviour for the sigma estimation and shows fewer fluctuations in the mean (e.g. electrons at high momenta, pions at low momenta) in two dimensions. Since the Nσ selections for the TPC will be performed in the (p, dE/dx) space by

the analysers, this plot represents the final correction quality and ultimately determines which tracks pass the selection.

To show the capability of fitting a high dimensional space and that there are cross-variable correlations, the Nσ plots were performed against $\tan(\lambda)$. The result is found in figure 5.18.

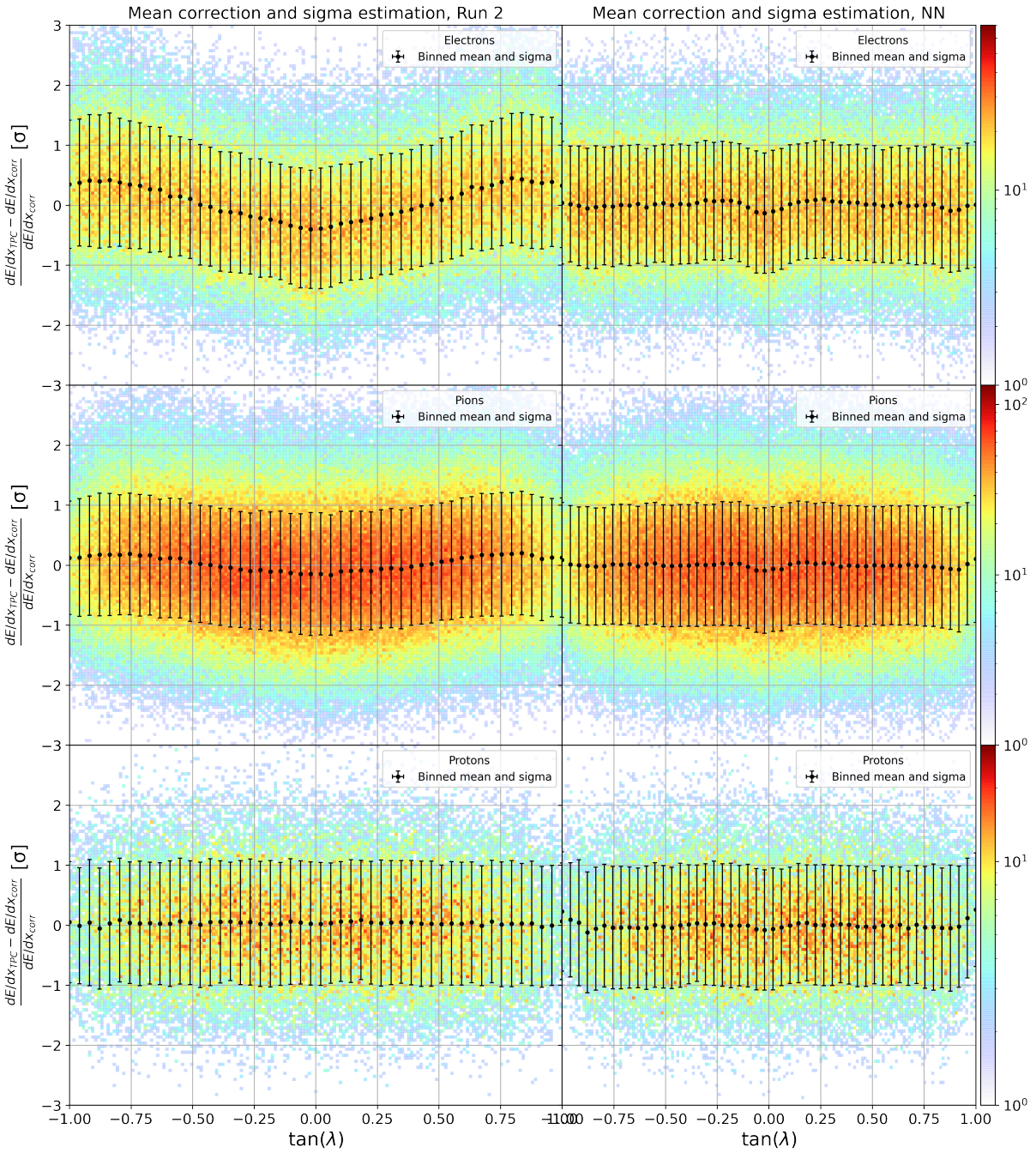


Figure 5.18: Nσ distributions for identified electrons, pions and protons with the method from Run 2 and the neural network application.

This clearly demonstrates that, besides the application of an η-map correction, correlations between the variables in higher dimensions exists, which are not captured by the approach of a spline fit in each observable individually. Tension between the corrected data and the calcu-

lated mean values is clearly apparent in almost all regions of $\tan(\lambda)$ for the approach of Run 2 and is particularly strong for electrons. While slight tension is also visible for the corrections of the neural network at $\tan(\lambda)$ close to 0, the deviations are far less pronounced. It remains a topic of investigation which dependencies cause the fluctuation at $\tan(\lambda) \approx 0$. A physical reason for such a behaviour could be tracks which get absorbed by the central electrode, since $\tan(\lambda) \approx 0$ corresponds to tracks passing the TPC in parallel to the central electrode. Upon investigation of the location of points with $\tan(\lambda) \approx 0$ in other dimensions, no clear dependence or outliers in any dimension could be found. This could indicate that crucial observables for understanding and balancing this behaviour are not yet included in the input to the neural network.

Similar to the mean correction, the predicted sigma values by the neural network have been measured with the ensemble for each track. The result is shown in figure 5.19.
As can be seen in the figure, the fluctuations for the sigma estimation are of similar scale and even slightly better than the fluctuations for the mean correction values. Enhanced fluctuations can be observed for higher momenta with pions and protons, where data density gets sparser and potential contamination from other particle species interferes. An overall higher spread is observed at large absolute values of $\tan(\lambda)$ (close to 1 and -1) is found for pions.

The apparent double structure at $|\tan(\lambda)| > 0.5$ (most prominently seen for pions) was investigated and it was found that the points with an overall lower standard deviation of the ensemble have (on average) a higher mean correction factor. These points are typically located close to the minimum ionizing region where the data point density is high and mean correction factors greater than 1 were typically found by the network. The secondary ridge with higher values of the standard deviation of the ensemble was found in regions of lower data density with momenta over $p \geq 2$ GeV/c for pions and mean correction values close to and below 1. This further demonstrates the necessity for sufficient data in all regions of the phase space. In particular, in the high momentum range, highly energetic cosmic particles could provide great value by being a reliable source and covering the full phase space uniformly. However, it remains to be seen whether enough statistics can be gathered from cosmic tracks.

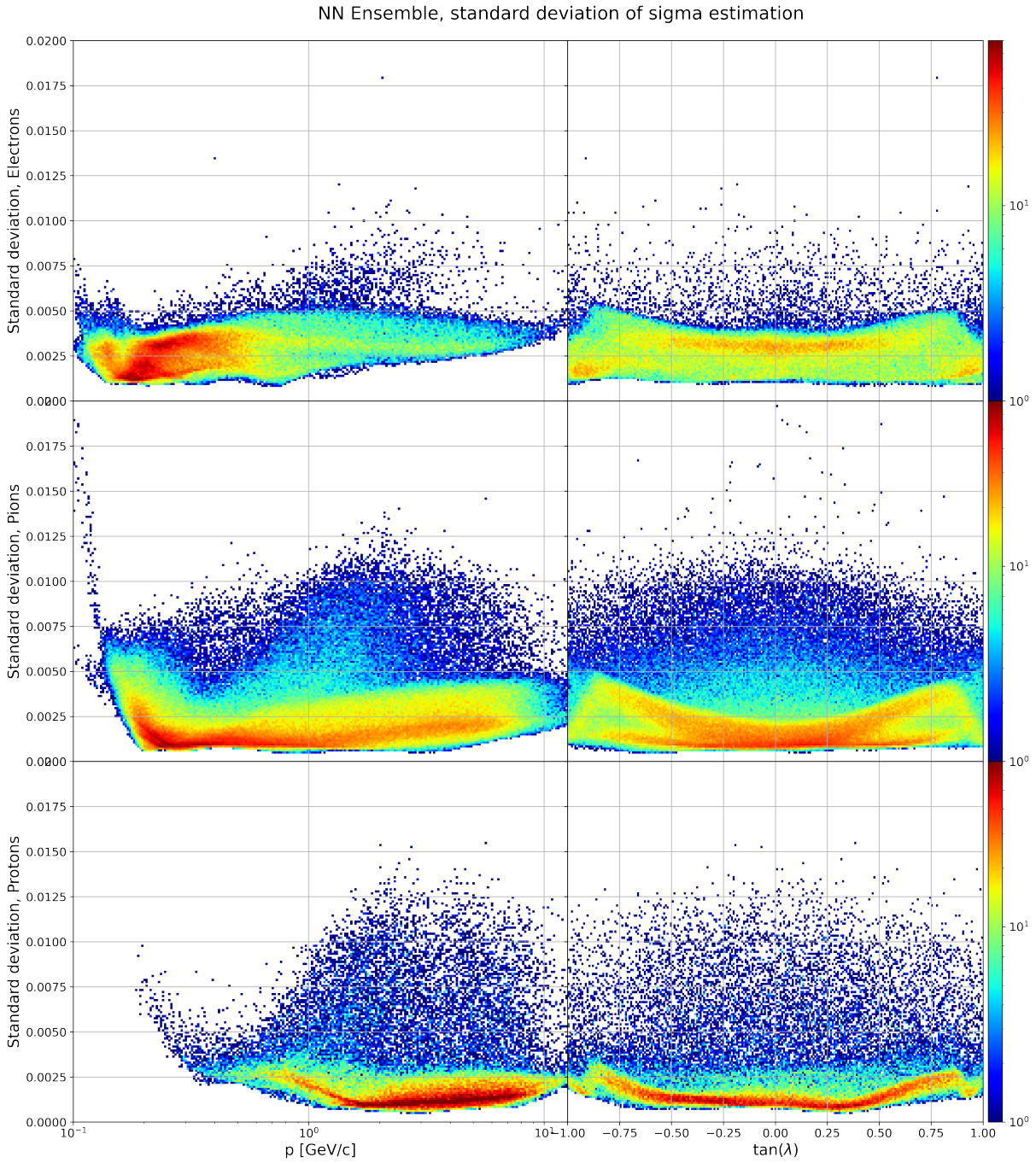


Figure 5.19: Deviation of the sigma estimation produced by a neural network ensemble for momentum and $\tan(\lambda)$.

5.6 Technical aspects on training and inferencing

The neural networks used in this thesis will have the main purpose of being applied at analysis time in the O^2 software framework. Since it will be used in many analyses, the main focus is put on efficient calculations, while it must be ensured that the network is not overtraining on the training data.

Firstly, the loss scores over the training epochs is investigated. If the network overtrains on the training data, it will be immediately visible on the validation score. The validation score will increase again after a certain number of epochs, while the training score decreases. This would correspond to a good fit of the network to its training data while missing the datapoints in the validation set. For the training, at each epoch 10% of all data is kept in aside as validation data while 90% is used for training the network. A more typical split would be 80% in the training data and 20% in the validation set, however since physics-wise all data in which samples are clearly identifiable can be used to train the network, this split was chosen such that the a higher amount of data is used for training. The obtained loss scores of the networks are shown in figure 5.20 as a function of the number of training epochs.

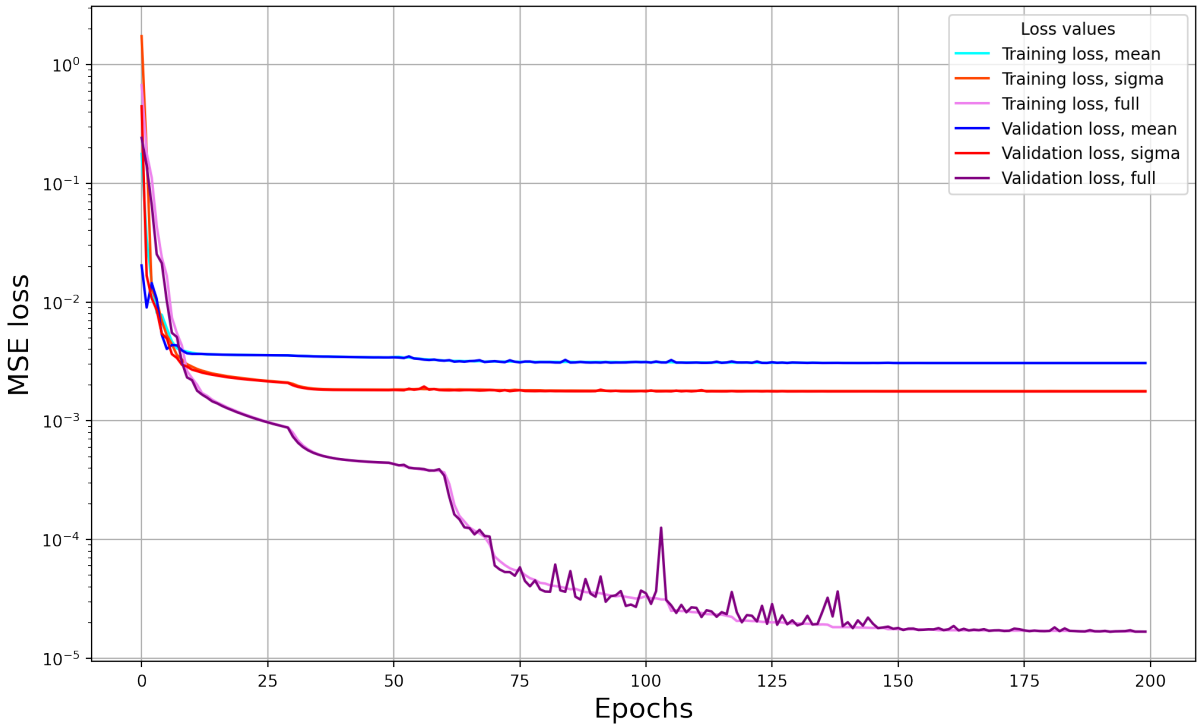


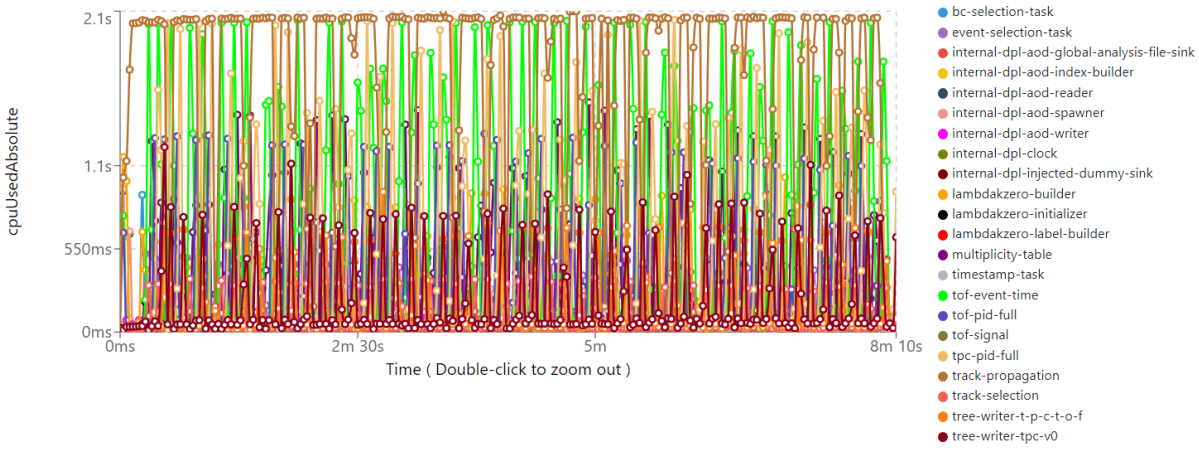
Figure 5.20: Loss scores of the networks trained for the mean and sigma estimation, as well as the full network for the application in O^2.

Firstly, it can be seen that no over-training has occurred. The low loss score of the full network can be explained since it does not retrain on the data itself, but rather only on the values produced by the previous two networks for the mean correction and sigma estimation. This demonstrates that the high-dimensional features can already be learned well by a comparably small network, which makes inferencing more efficient and allows the application at runtime. It is further noticeable that the loss for the estimation of the standard deviation tends to systematically lower values. An explanation for this behaviour can be given as data is denser for

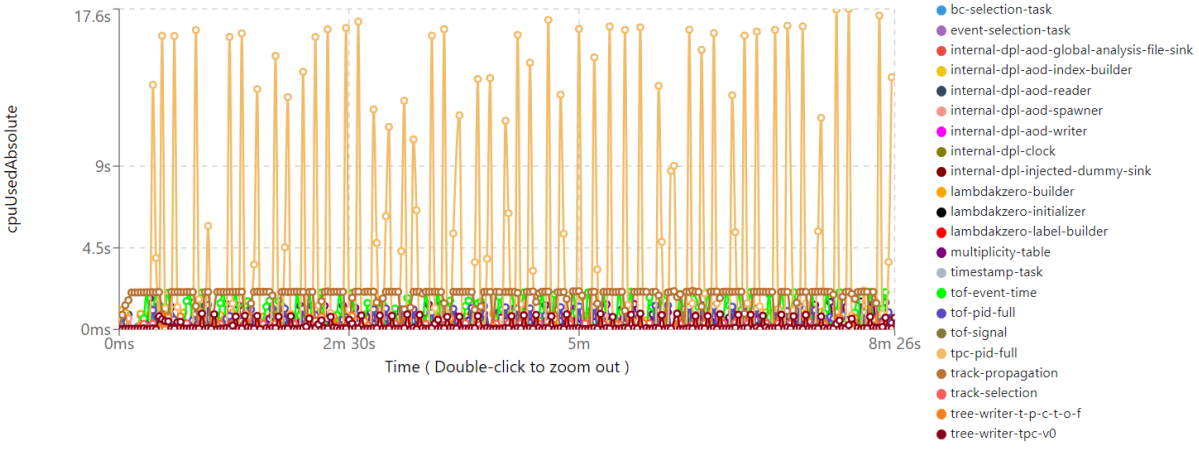
the sigma estimation due to the use of the absolute values in the $\mathcal{N}^*$-function and hence a better fit is achieved.

Fluctuations in the loss score are minimized over the epochs by using a learning rate scheduling which decreases the learning rate after a defined amount of epochs (here 5 epochs) if the loss score has not improved significantly. This adds beneficial effects on reaching the local minimum in the loss surface.

The application in the O^2 software framework shows no significant increase in computation time (wall time) on multi-CPU machines, but it was observed, that CPU time increases drastically on calling the Run()-function of the ONNX framework with spikes of up to 18 seconds per call. Typical functions of the tasks are found at around 2.2 seconds (CPU time). The difference shown in 5.21 (a) and (b). One network evaluation (i.e. one track and one mass hypothesis) takes around 52 ns (wall-time) and is hence comparable with the application of a multidimensional or convoluted C++ function.



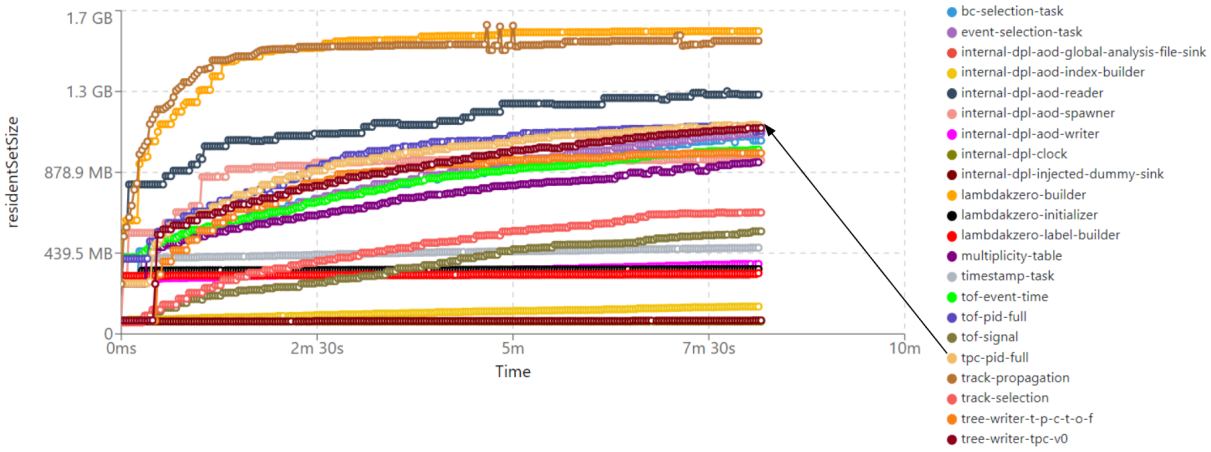
(a)



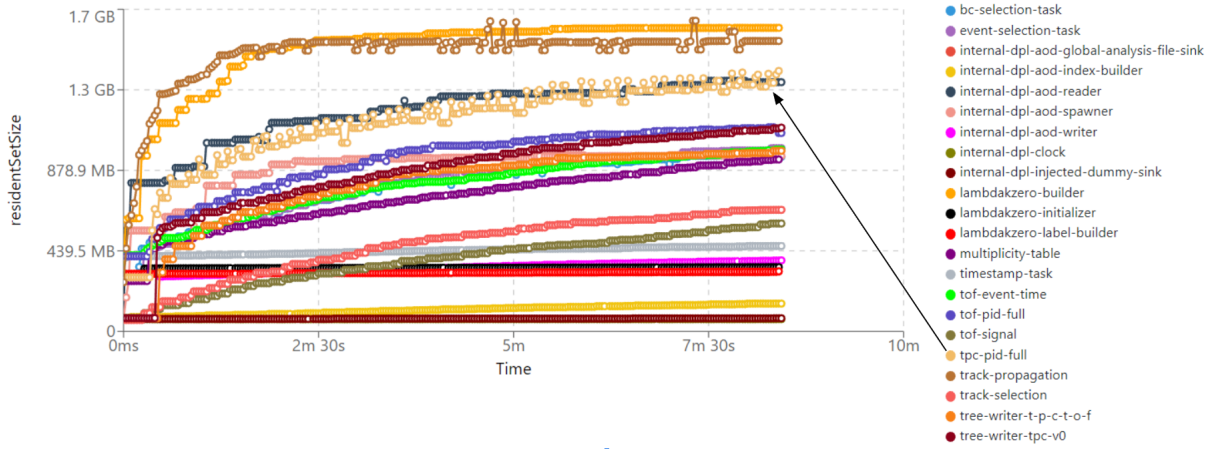
(b)

Figure 5.21: CPU time used for the tasks in O^2 without (a) and with the neural network application (b).

Various optimizations are under investigation at the time of writing this thesis to reduce this time and make calculations more efficient. Further tests have to be conducted on the hyperloop analysis system on which analyses will be run by many analysers and the impact on single- and multi-CPU machines.
Further considerations had to be taken for the optimization of the memory consumption. Each call of the ONNX Run()-function is computationally expensive, hence tracks are stored in an array and the Run()-function is only called once. Storing all tracks of a collision in one array and applying the network for each particle hypothesis posed a problem to the memory consumption of the task. Hence the network was applied on a per-species basis (total of eight calls of the Run()-function, instead of one call). This did not measurably increase wall time, but reduced the memory consumption to a sufficient level. Applying the neural network shows ≈250 MB higher memory-consumption (resident set size) of the tpc-pid-full task (here the network is applied) while leaving all other tasks untouched. This corresponds to a 23% higher memory consumption of $\approx$ 1.35 GB (accumulated over runtime) compared to $\approx$ 1.1 GB without the network. The memory consumption of all tasks is shown in figure 5.22 (a) and (b).



(a)



(b)

Figure 5.22: Memory consumption for the tasks in O^2 without (a) and with the neural network application (b). The tpc-pid-full task, where the network is applied, is marked with an arrow

Extrapolation capabilities of the neural network pose an interesting question about fluctuations and the behaviour in regions outside the training data of the neural network. This topic is of particular interest for the clean selection of light nuclei for which training data is typically to sparse. Unfortunately it cannot be guaranteed that the neural network performs well for regions of extrapolation, since this is also an open question in the machine learning community. Considering their masses, deuteron, triton and helium nuclei are on the order of magnitude protons, hence it is very much plausible that an extrapolation is feasible, however this will ultimately be determined once sufficient statistics are available during the Run 3 data taking.

VI Conclusion & Outlook

In conclusion, new, powerful tools for particle identification with the TPC detector were presented in this thesis. Starting with the initial parameterization, a novel method for the determination of initial parameters based on hyperparameter optimization was demonstrated which does not rely on initially assigned particle identities but can perform similarly well, if not better, than the approach used during Run 2 with identified tracks for each species.
Based on neural network regression, the new framework is capable of finding corrections to a parametrization of the mean energy loss per unit distance given by the theoretical prediction of the Bethe-Bloch formula. It is further capable of estimating the uncertainty of a Gaussian shaped particle density distribution and in contrast to the spline fit used in Run 2 of LHC, the neural network can perform both tasks in high dimensions.
In connection with the work performed in this thesis, neural network applications have been integrated into the new ALICE analysis software (O^2) and are available for users at runtime.

Overall, the methods are mainly limited by the availability of data in different corners of the phase-space, which will be no obstacle in Run 3 of LHC due to detector upgrades allowing for ≥ 50 times higher raw data taking capabilities. A particular advantage of the presented method is that it does not rely on Monte-Carlo driven data for training, but it rather represents a fully data-driven approach with full functional flexibility for the a priori unknown correction factors and sigma values at all corners of the covered phase-space. Moreover, it does not require multiple iterations to find sufficiently good corrections to the initial parametrization. The correction factors in every corner reach stable values with a precision of $\leq 3‰$ once data-density is sufficiently high and enough epochs are used to train the network. Since statistics were overall limited in the tests performed in this thesis, it remains to be seen how much data needs to be gathered to make a firmly reliable neural network fit and how often a fit has to be performed.
Further research has to be conducted in order to optimize the amount and distribution of data over the phase-space for training, to find optimal working points of the neural network and the hyperparameter optimization and to investigate the extrapolation behaviour of the neural network to other particle species such as e.g. deuteron, triton or Helium-3.
Especially at high momentum (≥ 8 GeV/c), cosmic tracks are foreseen to cover a wide region of the underlying phase-space while being a uniformly distributed. Additionally, kaon samples can be selected kinematically in low momentum with the TOF detector and weak decays of omega baryons ($\Omega \rightarrow K\Lambda$).

Finally, it has to be said that ultimately the performance can only be investigated by performing a specific physics analysis (e.g. a D^0 meson analysis in the golden channel $D^0 \rightarrow K^-\pi^+$, where particle identification is key due to the large combinatorial background at low momentum). This would show misidentifications and inefficiencies of the neural network fit and could guide the direction for further developments.
The full potential of this neural network regression is yet to be uncovered, as several criteria for clean selections were not fulfilled for the high statistics in Run 3 data at the time of writing

this thesis. However, the research conducted in this thesis on Run 2 data shows very promising results and extends the achievements of deep learning in ALICE.

Bibliography

[1] ALICE Collaboration. Technical Design Report for the Upgrade of the ALICE Inner Tracking System. Technical report, Nov 2013. CERN-LHCC-2013-024, ALICE-TDR-017.

[2] J. Adolfsson et al. The upgrade of the ALICE TPC with GEMs and continuous readout. *JINST*, 16:P03022. 87 p, Dec 2020. doi:10.1088/1748-0221/16/03/P03022.

[3] ALICE Collaboration. Upgrade of the ALICE Time Projection Chamber. Technical report, Oct 2013. CERN-LHCC-2013-020, ALICE-TDR-016.

[4] Jens Wiechula. Commissioning and Calibration of the ALICE-TPC, 2008. URL https://www.uni-frankfurt.de/46491136/Generic_46491136.pdf. PhD thesis, Institut für Kernphysik, Goethe Universität.

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

[6] P Buncic, M Krzewicki, and P Vande Vyvre. Technical Design Report for the Upgrade of the Online-Offline Computing System. Technical report, 2015. URL https://cds.cern.ch/record/2011297.

[7] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. 2017. doi:10.48550/ARXIV.1712.09913.

[8] Hans Bethe. Zur Theorie des Durchgangs schneller Korpuskularstrahlen durch Materie. 1930. doi:10.1002/andp.19303970303. Annalen der Physik.

[9] Bloch, Felix. Zur Bremsung rasch bewegter Teilchen beim Durchgang durch Materie. *Annalen der Physik. Band 408, Nr. 3, 1933.* doi:10.1002/andp.19334080303.

[10] Schicker, Rainer. Overview of ALICE results in pp, pA and AA collisions. *EPJ Web of Conferences*, 12 2016. doi:10.1051/epjconf/201713801021.

[11] PDG. 33. Passage of Particles Through Matter, 2019. URL https://pdg.lbl.gov/2019/reviews/rpp2018-rev-passage-particles-matter.pdf#section.33.6.

[12] Pablo Baladron Rodriguez, et. al. Calibration of the momentum scale of a particle physics detector using the armenteros-podolanski plot. 16, Jun 2021. doi:10.1088/1748-0221/16/06/p06036.

[13] Aamodt K., et al. Strange particle production in proton-proton collisions at $\sqrt{s}$ = 0.9 TeV with ALICE at the LHC. *Eur. Phys. J. C*, 71:1594. 34 p, 2011. doi:10.1140/epjc/s10052-011-1594-5.

[14] Takuya A., et. al. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019. URL https://arxiv.org/abs/1907.10902.

Acknowledgement

No work is done alone and this thesis was no exception. It has been the greatest honor and pleasure to conduct the research on this project and applications on the new software framework of the ALICE collaboration in person at CERN in Geneva in the first half of 2022. This has happened solely because of the trust, effort and time my supervisor Silvia Masciocchi has put in me and our work since I started in the group as a bachelor student. Thank you, Silvia!

As a master student one sometimes feels lost in all the details that such exciting physics as the LHC has to offer. Even one detector, like the TPC, is enough to spend a lifetime on it and still discover new aspects. With this I would like to thank Kai Schweda and Jens Wiechula for the guidance on the underlying physics and the discovery of problems as well as their input on possible solutions for them. Likewise I also want to thank the TPC team I worked with over this year (Jeremy, Annalena and Tiantian) who were always there to help and support.

I want to especially thank also a new friend and brilliant physicist I got to know during my time at CERN, Nicolo' Jacazio, who believed in the idea of neural networks for this project from the start onwards. Without his help and guidance in the new software framework, this work would not have been possible in this time.

Most importantly I want to thank my family and loved ones for their constant support and help even in difficult times.

Finally, I want to thank you, the reader, for investing your precious time to read this thesis. I hope you enjoyed it!

Declaration of authorship

Erklärung:

Ich versichere, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den 27.11.2022 . . . . . .