National Research University Higher School of Economics

as a manuscript

Nikita Kazeev

MACHINE LEARNING FOR PARTICLE IDENTIFICATION IN THE LHCB DETECTOR

PhD Dissertation Summary

for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Moscow – 2020

The PhD dissertation was prepared at National Research University Higher School of Economics and Sapienza Unviersity of Rome

Academic Supervisors:

- Andrey Ustyuzhanin, Candidate of Physical and Mathematical Sciences, Head of Laboratory of Methods for Big Data Analysis, Faculty of Computer Science, National Research University Higher School of Economics
- Dr. Barbara Sciascia, Researcher at the Istituto Nazionale di Fisica Nucleare (INFN) Laboratori Nazionali di Frascati
- Prof. Davide Pinci, Professor at the Laboratory of Nuclear and Subnuclear Physics, Sapienza Unviersity of Rome; Researcher at Istituto Nazionale di Fisica Nucleare (INFN) Rome section

1 Dissertation topic

1.1 Relevance – the LHCb experiment and its role in studying the Universe

The theory of strong and electroweak interactions, the so-called Standard Model (SM) of particle physics, has achieved outstanding successes. Yet there are unexplained phenomena, such as dark matter and the mass of the neutrinos; and gaps in theory, e.g. the lack of explanation for the predominance of matter over antimatter (CP violation). This incompleteness suggests that the SM is only an effective theory at the energies explored so far and that a more complete theory should exist.

When looking for more complete theories, the physics beyond the Standard Model, one of the best places to start is where existing theory says an event is not likely to happen: any deviations will be large compared to what we expect. For example, on the one hand, the branching fractions BR($B_{d,s} \to \mu^+\mu^-$) are tiny in the SM and can be predicted with high accuracy. On the other hand, a large class of theories that extend the Standard Model, like supersymmetry, allows significant modifications to these branching fractions and therefore an observation of any significant deviation from the SM prediction would indicate a discovery of new effects.

These decays have been extensively studied, most recently at the LHC experiments: LHCb [6, 1], CMS [16] and ATLAS [17]. There is also the combined analysis of the two results from the joint efforts of LHCb and CMS collaborations [23]. Thus far the decay $B_s \to \mu^+\mu^-$ has been observed. For branching fraction of $B_d \to \mu^+\mu^-$ only upper limits have been set.

Finding and studying rare decays presents many practical challenges. The rejection of background while preserving signal is of crucial importance. It is fundamental to develop selection criteria and muon identification to enable the discovery of $B_d \to \mu^+\mu^-$, and to place even more stringent limits on supersymmetry and other new physics models.

Another way to look for new physics is the study of fundamental properties within the SM. One of the more appealing currently is the lepton universality which requires equality of couplings between the gauge bosons and the three families of leptons. Hints of lepton non-universal effects have been reported in $B^+ \to K^+e^+e^-$ and $B^+ \to K^+\mu^+\mu^-$ [4, 3] decays. But there is no definitive observation of a deviation yet. A large class of models that extend the SM contains additional interactions involving enhanced couplings to the third-generation that would violate the lepton universality principle [2]. Semileptonic decays of b hadrons to third-generation leptons provide a sensitive probe for

such effects. In particular, the presence of additional charged Higgs bosons, which are often required in these models, can have a significant effect on the rate of the semitauonic decays of b-quark hadron $b \to c\tau^+\nu_\tau$ [12].

The LHCb experiment is one of the world's premier b-physics experiments. It is designed to exploit the high $pp \to c\bar{c}$ and $pp \to b\bar{b}$ cross-sections at the LHC in order to perform precision measurements of CP violation and rare decays.

The majority of physics analyses using data from the LHCb detector [9] rely on particle identification variables to separate charged tracks of different species: pions, kaons, protons, electrons, and muons. LHCb has three subdetectors that are used for particle identification: Ring Imaging Cherenkov (RICH) detectors, calorimeter (CALO), and muon identification system (MUON). Particle identification variables are used extensively to increase signal purity, reduce the processing time required to reconstruct high-multiplicity signal decays, and allow selections to reduce biases on quantities of physical interest, such as decay time.

Simulated data are needed at every stage of data processing for algorithm development and performance evaluation. To minimise the statistical uncertainty, the number of simulated events must be greater or equal to the number of events in real data. Once the LHCb experiment restarts in 2021, the luminosity will be increased by a factor of five. There is a pressing need for correspondingly faster simulation to be able to produce the required amount of the simulated data under the available computing budget.

Data-driven methods for measuring the performance of particle identification are needed, as the usage of simulated data has several serious shortcomings. First, particle identification variables are poorly reproduced by the simulation. Second, for some purposes, it is prohibitively expensive in terms of computing resources to accumulate a sufficiently large sample of simulated events. The LHCb collaboration developed data-driven methods based on so-called calibration samples. Calibration samples consist of charged tracks of different particle species that have been selected without the use of particle identification information about the RICH, CALO or MUON systems response to those tracks. In addition to being used in physics analyses, these samples can be used to monitor temporal variations in performance and to study future improvements in reconstruction algorithms [26]. Calibration samples are contaminated by background, i. e. they contain not only the particles of the desired type. In order to use those samples for algorithms development and evaluation, the contamination must be taken into account.

The role of the particle identification in achieving the physics goals of the upgraded experiment will remain critical. Furthermore, the calorimeter and muon system will continue to be essential in triggering because of their input

to the low-level trigger, which has the most stringent timing requirements [25]. All particle identification systems will contribute to the decision of the software trigger.

The higher collision energy available at the LHC makes that the underlying event of collisions from Run 2 is more complicated than those from Run 1. This effect has a severe impact on particle identification reducing the probability of correct particle identification. In order to overcome this complication, new identification techniques are being developed, also to be applied in the future upgrade scenario, where the number of interaction vertices will increase by a factor 5. These new techniques need to be executed already at the trigger level in order not to compromise the online/offline alignment achieved for Run 2. The muon identification is executed upfront on the trigger chain. Therefore, what is needed at the end is a single fast, high-performance algorithm.

1.2 The Study Objectives

The top-level goal of the dissertation is to improve the quality of event reconstruction in the LHCb experiment – the accuracy of the match between the real physical process that occurred in the detector and our results of reconstructing it from the detector readout. Reconstruction is handled by a sophisticated system of data processing algorithms. My contribution is concentrated around the following objectives:

Muon ID The objective is to develop an algorithm to distinguish muons from the other charged particles using only information from the muon subdetector. It must have better quality than state-of-the-art approaches. The algorithm must be fast enough to be used for online data processing. The performance of the algorithm on real data (calibration samples) needs to be experimentally evaluated.

Machine learning on background-contaminated data The objective is to develop a universal and theoretically-proven way to train machine learning algorithms on data contaminated with background. In machine learning terms, this is a particular model of label noise in a binary classification problem. Each label might be flipped from the true value. The flipping is random and occurs independently for each example. For each example, we know the probability that its label has been flipped, the probabilities for different examples are not necessarily equal.

Global PID The objective is to develop an algorithm that combines all available information about a particle candidate in the detector into a single rule

about its type and achieves better quality than the competing approaches.

Fast simulation The objective is to develop a method for fully-parametric simulation of Cherenkov detectors, which must be at least an order of magnitude faster than ab initio simulation while providing good precision. It also needs to be applicable for different detector hardware model and data taking conditions with minimal manual adaptations.

2 Key results and conclusions

Muon identification. I developed a novel method based on oblivious decision trees. When evaluated on data after preliminary filtration and setting the decision threshold to maintain 90% signal efficiency (true positive rate), it reduces background passthrough (false positive rate) from 14% to 10%, compared to the competing physics-based χ^2_{CORR} approach in the low-momentum region. The algorithm is implemented in the LHCb software stack.

Machine learning on background-contaminated data. My contribution is a novel and formally proven way to use any machine learning methods on such data. It is evaluated on two datasets from high energy physics and exhibits stable training and greater or equal performance compared to alternatives.

Global particle identification. I developed a novel method based on gradient boosting decision trees. It allowed reducing error, as measured by 1-vs-all $1 - \text{AUC}$ (area *over* the ROC curve) score, by 18%-54% compared to the baseline solution (a shallow fully-connected neural network) on simulated data. The method is implemented in the LHCb software stack.

Fast simulation. I developed a method based on generative adversarial network (GAN) for fully-parametric simulation. It is evaluated on two datasets corresponding to two different Cherenkov detectors: simulated BaBar DIRC data and real LHCb RICH data. The same method works in both of the cases with minimal tuning and provides good precision. The method is $1 - 2$ orders of magnitude faster than ab initio simulation even when running on a single CPU without batching. Aside from speed and quality, the novel characteristics of the method, compared to the previous approaches to fully-parametric simulation, are the following. The method is easily tunable in case of detector or data taking condition modifications, in contrast to explicit expert-written parametrisations. The final point of novelty is the usage of GANs to precisely model a relatively low-dimensional distribution, in contrast to the vast majority of published GAN

applications being about approximate modelling of high-dimensional objects, such as images.

3 Publications and approbation of research

3.1 First-tier publications

1. Denis Derkach, Nikita Kazeev, Fedor Ratnikov, Andrey Ustyuzhanin, and Alexandra Volokhova. "Cherenkov detectors fast simulation using neural networks". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 952 (2020), p. 161804. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2019.01.031, I am the main author, Scopus Q1

2. M. Borisyak and N. Kazeev. "Machine Learning on data with sPlot background subtraction". In: *Journal of Instrumentation* 14.08 (Aug. 2019), P08020–P08020. DOI: 10.1088/1748-0221/14/08/p08020, I am the main author, Scopus Q1

3.2 Second-tier publications

1. A Maevskiy, D Derkach, N Kazeev, A Ustyuzhanin, M Artemev, and L Anderlini and. "Fast Data-Driven Simulation of Cherenkov Detectors Using Generative Adversarial Networks". In: *Journal of Physics: Conference Series* 1525 (Apr. 2020), p. 012097. DOI: 10.1088/1742-6596/1525/1/012097, Scopus Q3

2. M Borisyak and N Kazeev. "Machine Learning on sWeighted data". In: *Journal of Physics: Conference Series* 1525 (Apr. 2020), p. 012088. DOI: 10.1088/1742-6596/1525/1/012088, I am the main author, Scopus Q3

3. Denis Derkach, Mikhail Hushchyn, Tatiana Likhomanenko, Alex Rogozhnikov, Nikita Kazeev, Victoria Chekalina, Radoslav Neychev, Stanislav Kirillov, and Fedor Ratnikov and. "Machine-Learning-based global particle-identification algorithms at the LHCb experiment". In: *Journal of Physics: Conference Series* 1085 (Sept. 2018), p. 042038. DOI: 10.1088/1742-6596/1085/4/042038, Scopus Q3

3.3 Conferences

1. Data Driven Simulation of Cherenkov Detectors using Generative Adversarial Network, Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada, 8 Dec 2019 – 14 Dec 2019

2. Training machine learning algorithms on background-contaminated data, Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada, 8 Dec 2019 – 14 Dec 2019

3. Machine Learning on sWeighted data, 19-th International Workshop on Advanced Computing and Analysis Techniques in Physics Research, 10 – 15 March 2019, Saas-Fee, Switzerland

4. Machine Learning for Muon Identification at LHCb, 9-th International Workshop on Advanced Computing and Analysis Techniques in Physics Research, 10 – 15 March 2019, Saas-Fee, Switzerland

5. Machine learning at LHCb, IV international conference on particle physics and astrophysics, 22 – 26 October 2018, Moscow, Russia

6. Cherenkov Detectors Fast Simulations Using Neural Networks, 10-th International Workshop on Ring Imaging Cherenkov Detectors (RICH 2018), Moscow, Russia from July 29 – August 4, 2018

4 Contents

The thesis consists of 9 chapters. Chapter 1 is the introduction equivalent to section 1 in this resume. Chapter 2 contains the descriptions of the machine learning methods used in the dissertation. Chapter 3 gives an overview of published applications of machine learning methods in high-energy physics. Chapter 4 describes the hardware and data processing of the LHCb experiment. Chapters 5 to 8 contain my contributions. Chapter 9 is the conclusion that contains the information from section 2 of the resume.

4.1 Introductory materials

Chapter 2 begins with a historical outlook of the AI field. Next, it formally defines different tasks addressed by machine learning. After that, it contains the descriptions of various metrics used to assess the quality of machine learning models. The next section discusses the limitations of machine learning – the so-called no free lunch theorem. It states that the performance of all machine learning algorithms is the same when averaged over all possible datasets. Machine learning works in practice in the cases where the algorithm incorporates, even if implicitly, some assumption about the data, usually some form of smoothness. The next section formally defines fully-connected neural networks and discusses their practical implementation. Next, there is an introduction to

gradient boosting decision trees. Afterwards, we introduce generative adversarial networks (GANs).

Chapter 3 begins with an overview of the main aspects of using machine learning in high-energy physics. The use of machine learning is enabled by the availability of large amounts of training data, usually, but not always simulated. Its results are validated using several described techniques. Next in the chapter are the descriptions of two related machine learning algorithms developed specifically for high-energy physics use cases. They provide a way to train a model who's output is independent of the defined nuisance parameters. The next section looks at three established areas of machine learning applications: event selection, event reconstruction and monitoring.

Chapter 4 begins with the description of the Large Hadron Collider. Then it continues to the LHCb detector hardware with a focus on the particle identification subsystems. The last section describes the LHCb data processing: real-time online trigger, more precise offline selection, and the ongoing implementation of the first level of the online trigger in GPU.

4.2 My contributions

4.2.1 Muon Identification – Chapter 5

Muons are present in the final states of many decays sensitive to new physics that are studied by the LHCb experiment [11, 6, 5]. Efficient and precise muon identification is of paramount importance for the LHCb.

The schematic view of the LHCb Muon subdetector is presented in figure 1. It consists of 5 sensitive planes (M1-M5) placed perpendicular to the beam axis. The muon detector is placed downstream of the rest of the detector, except for M1, which is in front of the calorimeters. Stations M2 to M5 are placed downstream the calorimeters and are interleaved with 80-cm thick iron absorbers.

The physical principle behind muon identification is the high penetration power of muons. If a charged particle leaves hits in the muon chambers, it is highly likely a muon. A muon identification algorithm is given two inputs:

- The reconstructed particle track's extrapolation into the muon chambers
- The coordinates and technical information about the hits in the muon chambers

There are two main sources of misidentification. The first is combinatorial – when unrelated hits align by chance with the track. The second is decays in flight for pions and kaons. Formally, this is a binary classification problem, where we strive to achieve the best quality on the given datasets.

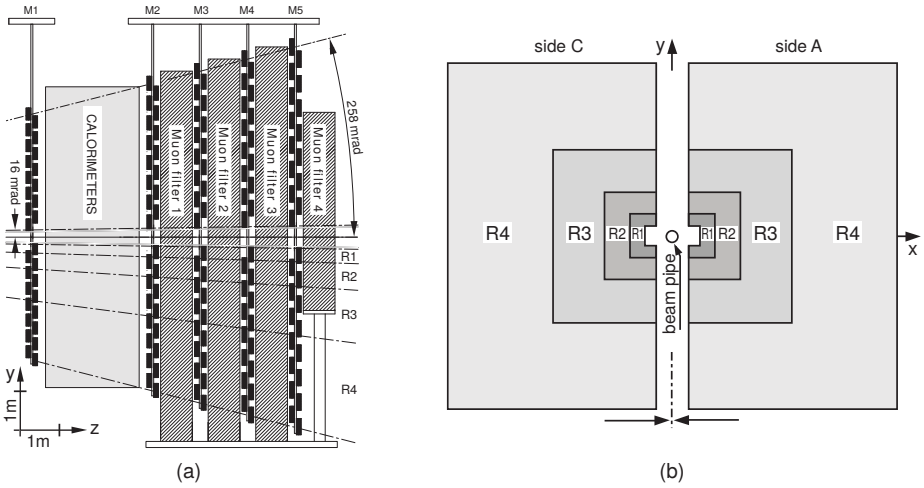

Figure 1: Side (a) and front (b) views of the LHCb Muon system. Reproduced from [8]

.

The baseline algorithm is based on a simple probabilistic model. The decision variable is defined as follows:

$$\chi^2_{\rm CORR} = \delta\mathbf{x}^T V_x^{-1} \delta\mathbf{x} + \delta\mathbf{y}^T V_y^{-1} \delta\mathbf{y}, \tag{1}$$

where $\delta\mathbf{x}$ and $\delta\mathbf{y}$ are the vectors with distances, in the x and y directions, between the track extrapolation points and the closest hit position in the M2-M5 stations. V_x and V_y are the covariance matrices, defined as a sum of two parts: $V^{\rm RES}$ and $V^{\rm MS}$. $V^{\rm RES}$ is a diagonal matrix, computed separately for x and y direction, that takes into account the spatial resolution of the muon stations:

$$V^{\rm RES}_{jj} = \frac{d_j^2}{12} \tag{2}$$

$$V^{\rm RES}_{jk} = 0 \text{ for } j \neq k, \tag{3}$$

where d_j is the x or y pad size corresponding to the muon hit. $V^{\rm MS}$ accounts for the uncertainty introduced by multiple scattering (MS):

$$V^{\rm MS}_{jk} = \sum_{z_i < z_j, z_k} (z_j - z_i)(z_k - z_i)\sigma^2_{{\rm MS},i}. \tag{4}$$

$\sigma^2_{{\rm MS},i}$ is estimated as [28, 10]:

$$\sigma^2_{{\rm MS},i} = \frac{13.6{\rm MeV}}{\beta cp} q \sqrt{\frac{z_i}{X_0}} [1 + 0.038 \ln(z_i/X_0)] \approx \frac{13.6{\rm MeV}}{\beta cp} q \sqrt{\frac{z_i}{X_0}}, \tag{5}$$

where p is the particle's momentum, βc the velocity, and q the charge; z_i/X_0 is the thickness of the scattering medium in units of radiation length. The

position along the $z-$axis and the thickness of the considered scattering centres are presented in the table 1.

MS contribution	z position (m)	z_i/X_0
ECAL+SPD+PS	12.8	28
HCAL	14.3	53
M23 filter	15.8	47.5
M34 filter	17.1	47.5
M45 filter	18.3	47.5

Table 1: Position and thickness in units of radiation length for the scattering media contributing to the multiple scattering experienced by particles traversing the muon system. Reproduced from [10].

The probability of a muon to reach a given muon station depends on its momentum [24]. For a low-momentum particle, hits in the farthest stations are likely to be combinatorial background. Therefore, for low-momentum particles ($p < 6$ GeV/c), only M2 and M3 stations are used, while for particles with $p \geq 6$ GeV/c all stations are used. The 6 GeV/c threshold coincides with one of the used in `IsMuon`, around 90% of muons with this momentum reach M5.

Next in the chapter is the analysis of several machine learning approaches culminating in the algorithm that is proposed as a key contribution of the dissertation. In consists of CatBoost trained on muon, pion and proton calibration samples, using the background subtraction method proposed in chapter 6 of the thesis. The following features are used:

- Information about the closest hits in each station: space residuals, hit time, hit delta time, whether the hit is crossed
- χ^2_{CORR}
- Track momentum and transverse momentum

Since muon identification is used for many different analyses in different regions of the phase space, it is hard to put a single figure of merit. To compare quality of different algorithms, we used a proxy method. We compared the ROC curves in different momentum bins for true positive rate greater than 90%. The method proposed in the dissertation achieves the best quality among the evaluated algorithms, as presented in figures 2 and 3.

The final section of the chapter describes an outreach activity – the International Data Analysis Olympiad, where this problem was offered for the online part of the competition.

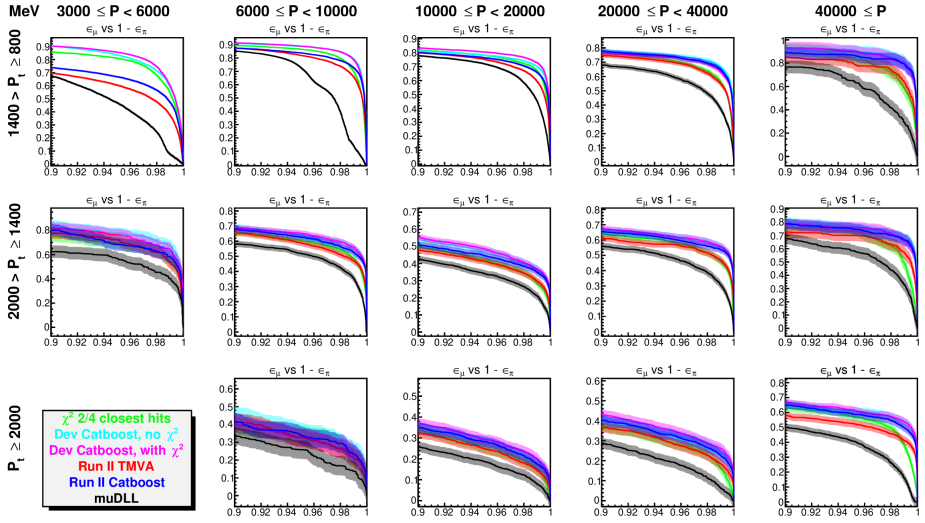

Figure 2: Muon efficiency versus pion rejection after applying `IsMuon` selection using 2016 calibration data that is weighted in momentum and nPVs. "χ^2 2/4 closest hits" is described in the algorithm based on physics principles "Run II Catboost" and "Run II TMVA" are the early machine learning solutions. "Dev Catboost" are the best machine learning solutions. "Dev Catboost, with χ^2" uses χ^2_{CORR} as a feature, "Dev Catboost, no χ^2" does not.

4.3 Machine Learning on Data With sPlot Background Subtraction – Chapter 6

Experimental data obtained in high energy physics experiments usually consists of contributions from different processes. A large part of a typical data analysis consists of selecting the target decay from all collected data. A common technique used as a part of this process is sideband subtraction. It requires a signal-enriched and a signal-poor phase-space regions be identified (usually by an invariant mass fit). A commonly used method is sPlot [29]. It assigns weights (sWeights) to events, some of them negative. This does not present a problem if the next analysis steps use simple single-dimensional tools, like histograms, but is an obstacle for some multivariate machine learning methods. In this chapter, we prose a mathematically rigorous way of training machine learning algorithms on such data.

In machine learning terms, we are dealing with a particular model of label noise coupled with prior knowledge. For each example, we know the probability that its label has been flipped. The flipping probabilities are not constant but are sampled from some distribution independently of the features' distribution. As an example, consider the case where you have a dataset of with stars

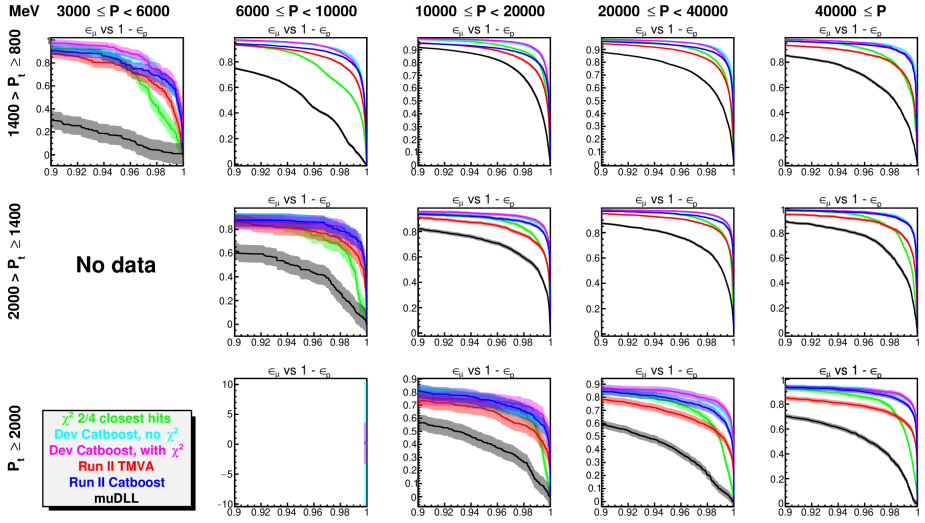

Figure 3: Muon efficiency versus proton rejection after applying IsMuon selection using 2016 calibration data that is weighted in momentum and nPVs. "χ^2 2/4 closest hits" is described in the algorithm based on physics principles "Run II Catboost" and "Run II TMVA" are the early machine learning solutions. "Dev Catboost" are the best machine learning solutions. "Dev Catboost, with χ^2" uses χ^2_{CORR} as a feature, "Dev Catboost, no χ^2" does not. 2

labelled from telescope observations. The observations were taken in different atmospheric conditions. For each observation the atmospheric conditions are known; the probability of misclassifying a star is a known function of atmospheric conditions.

The sPlot method works as follows. Take a dataset populated by events from N_s sources. Assume that distribution of some variables is known for each source, call these variables discriminative. Usually, the discriminative variable is the reconstructed invariant mass, and the probability densities are estimated by a maximum likelihood fit. sPlot reconstructs the distributions of the rest of the variables (call them control), provided they are independent of the discriminative variables for each event source.

Let $p_k(m)$ be probability density function of the discriminative variable m of the k-th species, N_k be the number of events expected on the average for the k-th species, N be the total number of events, m_e be the value of m for the e-th event. Compute the covariance matrix of the signal and background probability density functions $\mathbf{V}$:

$$\left(\mathbf{V}^{-1}\right)_{nj} = \sum_{e=1}^{N} \frac{p_n(m_e)p_j(m_e)}{\left(\sum_{k=1}^{N_s} N_k p_k(m_e)\right)^2} \tag{6}$$

The sWeight for the e-th event corresponding to the n-the species is obtained using the following transformation:

$$\text{sWeight}_n(m_e) = \frac{\sum_{j=1}^{N_s} \mathbf{V}_{nj} p_j(m_e)}{\sum_{k=1}^{N_s} N_k p_k(m_e)} \tag{7}$$

The problem with using the sWeights directly as example weights is that they remove the lower bound of an algorithm's loss function – the more an algorithm errors on an example with negative weight, the less is the value of the loss.

My solution for the problem is to replace the classification problem with example negative weights with a regression problem, where the model is asked to predict said weights. Let m be the variable that was used to compute the sWeights, x is the rest of variables. Let $p_{\text{signal}}(x)$ be the probability density function of signal, $p_{\text{mix}}(x)$ be the probability function of the signal and background mixture from which the data were sampled. It is possible to weight the dataset, so distribution of x will be that of the signal component with positive weights equal to class probabilities $W(x) = \frac{p_{\text{signal}}(x)}{p_{\text{mix}}(x)}$. The data weighted with the sWeights achieve the same distribution. Therefore the signal probability for an example with features x will be the average of the sWeights over examples with features x:

$$E_m\left[w(m) \mid x\right] = \frac{p_{\text{signal}}(x)}{p_{\text{mix}}(x)}. \tag{8}$$

Since the optimal output lies in $[0, 1]$, one can easily avoid a priori incorrect solutions by applying the sigmoid function to the model output. The resulting loss function is the following:

$$L = \sum_i \left(w_i - \frac{e^{f_\theta(x_i)}}{1 + e^{f_\theta(x_i)}} \right)^2, \tag{9}$$

where w_i is the sWeight and $f_\theta(x_i)$ is the model output. Therefore the loss is called Constrained MSE. We have implemented this loss for the CatBoost machine learning library [30] and the source code is available on GitHub.[1]

Next in the section is an alternative method developed by a coauthor. Instead of using the sPlot method, he directly applies the maximum likelihood principle. The loss function is the following:

$$L(\theta) = -\sum_i \log\left[f_\theta(x_i) p_{\text{signal}}(m_i) + (1 - f_\theta(x_i)) p_{\text{background}}(m_i) \right], \tag{10}$$

where $f_\theta(x_i)$ is the output of the model, $p_{\text{signal}}(m_i)$ and $p_{\text{background}}(m_i)$ are the probability densities of the signal and background m distributions.

[1] https://github.com/kazeevn/catboost/tree/constrained_regression

Finally, the proposed methods are evaluated on two datasets: UCI Higgs and muon identification. In our experiments, CatBoost training converges for both our methods and naive training on the sWeights; due to a large number of regularisations embedded in the algorithm, it is able to avoid the problems with unbounded loss. The quality is similar in all cases, with a slight advantage of Constrained MSE on the muon identification. For a fully-connected neural network, naive training on the sWeights diverges, our methods converge and provide similar quality.

4.4 Global Charged Particle Identification – Chapter 7

The chapter begins with addressing the difficulty of formalising the quality of a particle identification algorithm. The physical goal of the algorithm is to match the reconstructed particle types with the real particle types in the detector. The problem is that the value of any quality metric depends on the dataset over which it is measured. Different plausible alternatives are discussed. The LHCb collaboration uses a simulated dataset which includes a representative sample of processes under study.

Global PID is the last step of particle identification; its goal is to combine the responses of all the subdetectors into a single decision about the particle class. The earliest and most straightforward approach is to compute the log-likelihoods separately for each subdetector. This approach is far from ideal, as it disregards potentially informative correlations between different features and relies upon the log-likelihoods to be well-defined and computed using the same priors.

The next logical step is to put all the available information into a machine learning algorithm which is described in the following sections. We investigate two state-of-the-art approaches: a deep neural network and gradient boosting decision trees. The performance is evaluated on simulated and real data. Both of the approaches outperform the baseline (a shallow neural network) and show similar performance, with gradient boosting having a slight lead. The results of an evaluation performed on simulated data presented in figure 4. Since two very different methods perform similarly, we have likely achieved the limit of performance for this stage, and further improvement must be thought either in better subdetector-specific preprocessing (e. g. muon identification) or training dataset composition.

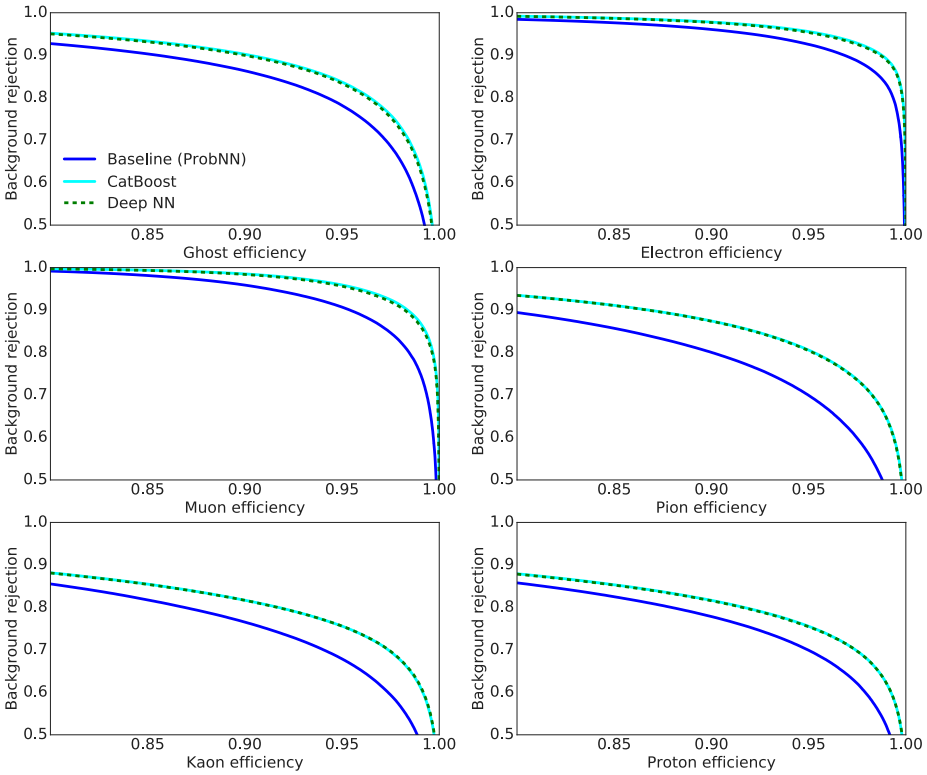

Figure 4: Performance of the models in terms of one-vs-rest ROC curves. Reproduced from [22].

5 Fast Simulation of the Cherenkov Detector – Chapter 8

The chapter begins with a description of the role of simulated data in high-energy physics – it is used at every stage of experiments, from detector design to reconstruction algorithms evaluation and final analysis.

Simulation in LHCb consists of three stages. First, the p-p collision and the processes resulting from it are simulated. Second, the interaction between the resulting particles and the detector hardware is simulated. Third, the simulated readout is reconstructed with the same algorithms that are used for reconstructing the collected experimental data.

Simulation is computationally expensive. Given the budget constraints, the LHCb collaboration won't be able to produce the required amount of simulated data after the experiment is restarted with 5x luminosity in 2021. Therefore, faster simulation is needed. The core idea behind fast simulations is that the detector is the same for different events and simulating it from scratch for each

event is a waste of effort. A possible way to take advantage of this is the so-called ReDecay method, where the background part is shared between the different events. The issue with this method is that the resulting events are not independent. This complicates the analysis. An alternative is parametrisation, where a heuristic parametrisation replaces some part of ab initio computations. If this includes the reconstruction step, such simulation is called fully-parametrised.

My contribution is a method for fully-parametrised simulation of Cherenkov detectors with a generative adversarial network. It takes as its input the track true particle type, kinematic parameters and a proxy for detector occupancy. The model outputs the likelihoods for particle types. The model is evaluated for two use Cherenkov detectors: BaBar DIRC detector and LHCb RICH.

DIRC training and evaluation are done on a dataset with two particles simulated with FastDirc software [21]. The model inputs are the signal particle type and the energies, pseudorapidities and the transverse coordinates of both particles in the event — 7 variables in total. Its outputs are the delta log-likelihoods with pion hypothesis for the electron, kaon, muon, proton and "beyond threshold" particle type hypotheses — 5 variables. We used the Cramer (energy) GAN [13] for the sake of unbiased gradient estimates. The generator and discriminator are fully connected neural networks with 10 layers, each containing 128 neurons with ReLu activation. We trained a separate model for each signal particle type. Each model was trained using 1 million generated events. We transformed each observable distribution into a Gaussian using quantile transformation before passing them to the neural network. The speed improvement with respect to the full simulation in GEANT 4 [7] is $8 \cdot 10^4$ times on a single CPU core. The speed is also improved with respect to the FastDIRC generation, where a factor up to 80 can be achieved. The batch generation on GPU produces up to 1 million track predictions per second.

The LHCb collaboration is developing an application for parametrised simulation of the detector response called Lamarr. Our RICH simulation method has been implemented as a part of the application. Lamarr architecture is described in the corresponding section. The model for RICH simulation is the same as for DIRC simulation; it is presented in figure 5. The model input parameters are momentum, pseudorapidity and the number of reconstructed tracks in the event. The model outputs are `RichDLL*` – the delta log-likelihoods between a particle type hypothesis and the pion hypothesis. The model is trained on the 2016 calibration samples. On a single CPU core, the proposed model is at least 2 orders of magnitude faster than the full simulation with Geant4. We test the quality on the same calibration channel that was used for training, and on a different one. GAN shows a good approximation of the real data distributions in the same channel. The results of one of the tests are presented in figure 6.

The main source of systematic uncertainty in Lamarr will likely be the selection of parametrisation and the training data.

References

[1] R. Aaij et al. "Measurement of the $B_s^0 \to \mu^+\mu^-$ Branching Fraction and Effective Lifetime and Search for $B^0 \to \mu^+\mu^-$ Decays". In: *Phys. Rev. Lett.* 118 (19 May 2017), p. 191801. DOI: 10.1103/PhysRevLett.118.191801. URL: https://link.aps.org/doi/10.1103/PhysRevLett.118.191801.

[2] R. Aaij et al. "Measurement of the Ratio of Branching Fractions $\mathcal{B}(\overline{B}^0 \to D^{*+}\tau^-\overline{\nu}_\tau)/\mathcal{B}(\overline{B}^0 \to D^{*+}\mu^-\overline{\nu}_\mu)$". In: *Phys. Rev. Lett.* 115 (11 Sept. 2015), p. 111803. DOI: 10.1103/PhysRevLett.115.111803. URL: https://link.aps.org/doi/10.1103/PhysRevLett.115.111803.

[3] R. Aaij et al. "Search for Lepton-Universality Violation in $B^+ \to K^+\ell^+\ell^-$ Decays". In: *Phys. Rev. Lett.* 122 (19 May 2019), p. 191801. DOI: 10.1103/PhysRevLett.122.191801. URL: https://link.aps.org/doi/10.1103/PhysRevLett.122.191801.

[4] R. Aaij et al. "Test of Lepton Universality Using $B^+ \to K^+\ell^+\ell^-$ Decays". In: *Phys. Rev. Lett.* 113 (15 Oct. 2014), p. 151601. DOI: 10.1103/PhysRevLett.113.151601. URL: https://link.aps.org/doi/10.1103/PhysRevLett.113.151601.

[5] R Aaij et al. "Differential branching fraction and angular analysis of the decay B 0→ K* 0 μ+ μ-". In: *Physical review letters* 108.18 (2012), p. 181806.

[6] R Aaij et al. "Measurement of the ${B_s}^0 \to \mu^+\mu^-$ branching fraction and search for $B^0 \to \mu^+\mu^-$ decays at the LHCb experiment". In: *Physical review letters* 111.10 (2013), p. 101805.

[7] John Allison et al. "Geant4 developments and applications". In: *IEEE Transactions on nuclear science* 53.1 (2006), pp. 270–278.

[8] A Augusto Alves Jr et al. "Performance of the LHCb muon system". In: *Journal of Instrumentation* 8.02 (2013), P02022.

[9] A. A. Alves Jr. et al. "The LHCb detector at the LHC". In: *JINST* 3 (2008), S08005. DOI: 10.1088/1748-0221/3/08/S08005.

[10] Lucio Anderlini et al. *New muon identification operators*. Tech. rep. LHCb-INT-2019-020. CERN-LHCb-INT-2019-020. Geneva: CERN, Aug. 2019. URL: https://cds.cern.ch/record/2687369.

[11] F Archilli et al. "Performance of the muon identification at LHCb". In: *Journal of Instrumentation* 8.10 (2013), P10020.

[12] "Beauty quarks test lepton universality". In: *CERN Courier* 58.3 (Mar. 2018). URL: https://cds.cern.ch/record/2315229.

[13] Marc G Bellemare et al. "The Cramer distance as a solution to biased Wasserstein gradients". In: *arXiv preprint arXiv:1705.10743* (2017).

[14] M. Borisyak and N. Kazeev. "Machine Learning on data with sPlot background subtraction". In: *Journal of Instrumentation* 14.08 (Aug. 2019), P08020–P08020. DOI: 10.1088/1748-0221/14/08/p08020.

[15] M Borisyak and N Kazeev. "Machine Learning on sWeighted data". In: *Journal of Physics: Conference Series* 1525 (Apr. 2020), p. 012088. DOI: 10.1088/1742-6596/1525/1/012088.

[16] Serguei Chatrchyan et al. "Measurement of the $B_s^0 \to \mu^+\mu^-$ branching fraction and search for $B^0 \to \mu^+\mu^-$ with the CMS experiment". In: *Physical review letters* 111.10 (2013), p. 101804.

[17] ATLAS collaboration et al. "Study of the rare decays of B_s^0 and B^0 mesons into muon pairs using data collected during 2015 and 2016 with the ATLAS detector". In: *Journal of High Energy Physics* 2019.4 (2019), p. 98.

[18] Denis Derkach et al. "Cherenkov detectors fast simulation using neural networks". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 952 (2020), p. 161804. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2019.01.031.

[19] Denis Derkach et al. "Data Driven Simulation of Cherenkov Detectors using Generative Adversarial Network". In: *Machine Learning and the Physical Sciences Workshop at the 33rd Conference on Neural Information Processing Systems (NeurIPS)* (Dec. 2019). URL: https://ml4physicalsciences.github.io/files/NeurIPS_ML4PS_2019_40.pdf.

[20] Denis Derkach et al. "Machine-Learning-based global particle-identification algorithms at the LHCb experiment". In: *Journal of Physics: Conference Series* 1085 (Sept. 2018), p. 042038. DOI: 10.1088/1742-6596/1085/4/042038.

[21] John Hardin and Mike Williams. "FastDIRC: a fast Monte Carlo and reconstruction algorithm for DIRC detectors". In: *JINST* 11.10 (2016), P10007. DOI: 10.1088/1748-0221/11/10/P10007. arXiv: 1608.01180 [physics.data-an].

[22] Mikhail Hushchyn and Denis Derkach. *Plots for yPID*. URL: https://indico.cern.ch/event/668115/.

[23] Vardan Khachatryan et al. "Observation of the rare $B_s^0 \to \mu^+\mu^-$ decay from the combined analysis of CMS and LHCb data". In: *Nature* 522 (2015), pp. 68–72. DOI: 10.1038/nature14474. arXiv: 1411.4413 [hep-ex].

[24] G Lanfranchi et al. *The muon identification procedure of the LHCb experiment for the first data*. Tech. rep. 2009.

[25] *LHCb Trigger and Online Upgrade Technical Design Report*. Tech. rep. CERN-LHCC-2014-016. LHCB-TDR-016. May 2014. URL: https://cds.cern.ch/record/1701361.

[26] Oliver Lupton. "Studies of $D^0 \to K^0_S h^+ h'^-$ decays at the LHCb experiment". Presented 14 Sep 2016. July 2016. URL: https://cds.cern.ch/record/2230910.

[27] A Maevskiy et al. "Fast Data-Driven Simulation of Cherenkov Detectors Using Generative Adversarial Networks". In: *Journal of Physics: Conference Series* 1525 (Apr. 2020), p. 012097. DOI: 10.1088/1742-6596/1525/1/012097.

[28] C. Patrignani et al. "Review of Particle Physics". In: *Chin. Phys.* C40.10 (2016), p. 100001. DOI: 10.1088/1674-1137/40/10/100001.

[29] Muriel Pivk and Francois R Le Diberder. "sPlot: A statistical tool to unfold data distributions". In: *NIMA* 555.1-2 (2005), pp. 356–369.

[30] Liudmila Prokhorenkova et al. In: *Advances in Neural Information Processing Systems*. 2018, pp. 6638–6648.

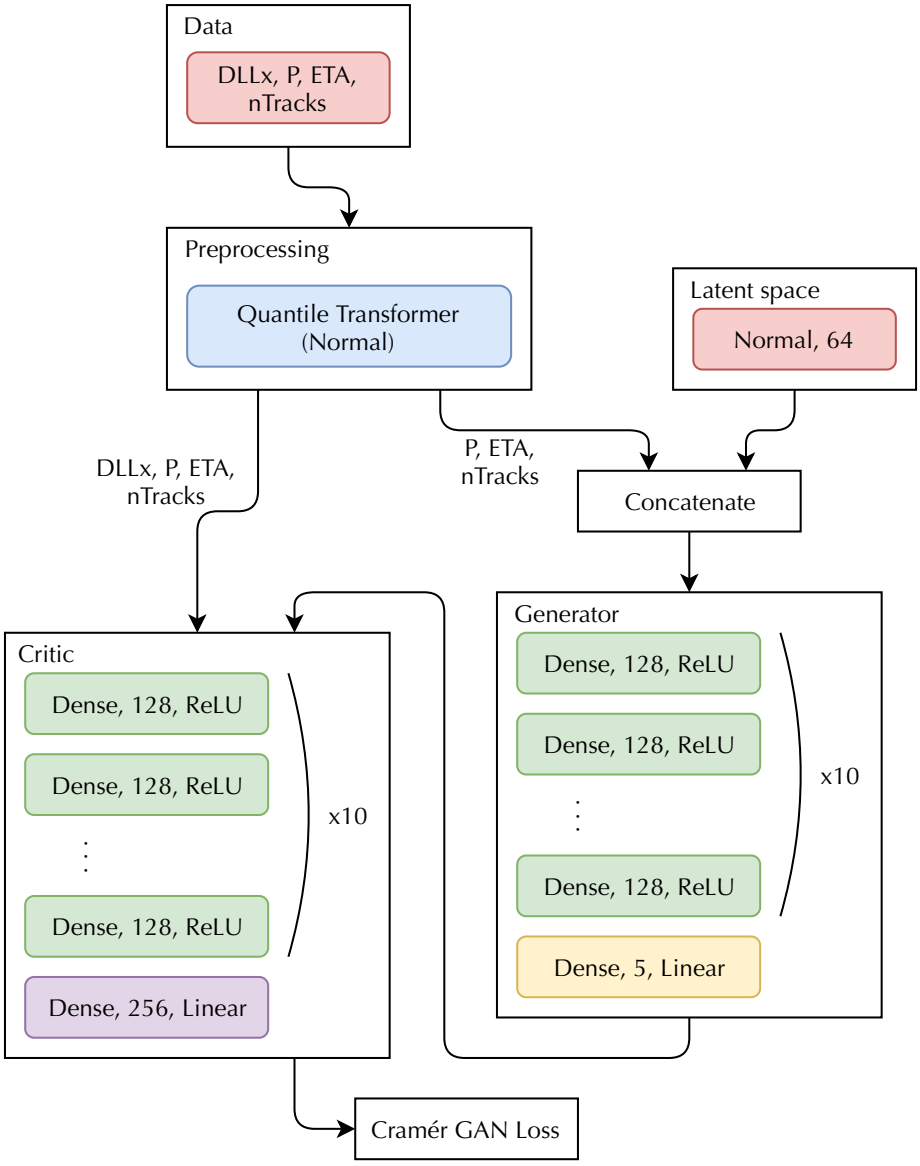

Figure 5: Architecture of the RICH GAN. First presented at [19].

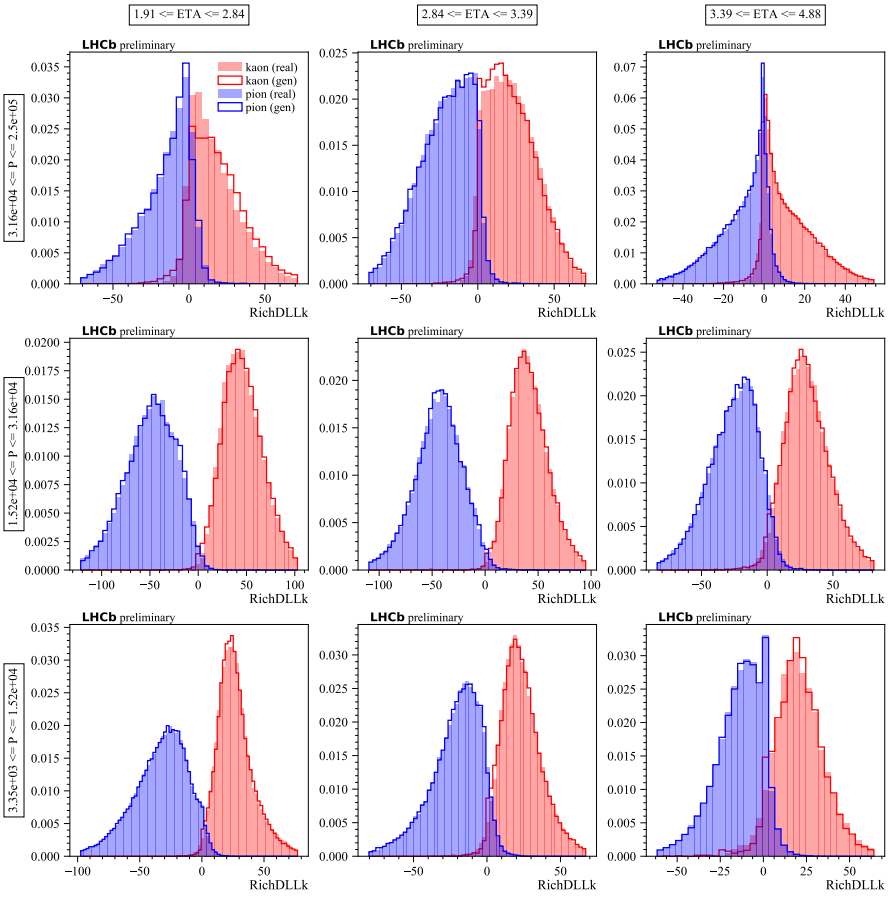

Figure 6: Weighted real data and generated distributions of `RichDLLk` for kaon and pion track candidates in bins of pseudorapidity (ETA) and momentum (P, MeV) over the full phase-space. First published in [27].