EPJ Web of Conferences **259**, 13008 (2022) https://doi.org/10.1051/epjconf/202225913008
SQM 2021

Machine Learning Application for Λ Hyperon Reconstruction in CBM at FAIR

Shahid Khan[1,*], *Viktor* Klochkov[1], *Olha* Lavoryk[2], *Oleksii* Lubynets[3,4], *Ali Imdad* Khan[1], *Andrea* Dubla[3], and *Ilya* Selyuzhenkov[3,5]

[1]Eberhard Karls University of Tübingen; [2]Taras Shevchenko National University of Kyiv; [3]GSI, Darmstadt; [4]Goethe University Frankfurt; [5]NRNU MEPhI, Moscow

Abstract. The Compressed Baryonic Matter experiment at FAIR will investigate the QCD phase diagram in the region of high net-baryon densities. Enhanced production of strange baryons, such as the most abundantly produced Λ hyperons, can signal transition to a new phase of the QCD matter. In this work, the CBM performance for reconstruction of the Λ hyperon via its decay to proton and π^- is presented. Decay topology reconstruction is implemented in the Particle-Finder Simple (PFSimple) package with Machine Learning algorithms providing efficient selection of the decays and high signal to background ratio.

Introduction. Theoretical calculations predict the possibility of a first order phase transition from hadron gas to a deconfined phase of strongly interacting matter and the existence of a critical point in the region of the QCD phase diagram above 450 MeV chemical potential (μ_B) and below 180 MeV temperature [1]. The experiments at SIS-18, CERN SPS, RHIC beam energy scan and fixed target programs, and future experiments at the Facility for Antiproton and Ion Research (FAIR) and NICA facilities are dedicated to explore this region of the phase diagram. The Compressed Baryonic Matter (CBM) experiment at FAIR in Darmstadt will investigate the region at 5–8 normal baryon density ($\mu_B \approx 450 - 900$ MeV) and temperatures below 120 MeV. One of the signatures of the new phase of matter is the enhanced production of strange baryons and Λ is the most abundantly produced hyperon at FAIR SIS-100 energies. In this work, the CBM performance for reconstruction of the Λ hyperon via its decay to proton and π^- using Machine Learning algorithms to achieve an efficient selection of the decays and high signal to background ratio is presented.

CBM experiment setup. The CBM experiment consists of three main components: a tracking system inside a dipole magnet, detectors for particle identification and detectors for collision geometry determination. The tracking system consists of a Micro Vertex Detector (MVD) and a Silicon Tracking System (STS). The particle identification is provided by the Ring Imaging Cherenkov detector, Muon Chamber, Transition Radiation Detector and Time-of-Flight (TOF) wall. Geometry determination is done with the help of Projectile Spectator Detector, inner part of TOF and the tracking system. For this analysis a sample of 100k minimum bias Au-Au collisions at 12A GeV/c was generated using DCM-QGSM-SMM [2] (Model A) and URQMD [3] (Model B) with collision products transported through the CBM setup using GEANT4 [4].

Machine learning. In the already existing Kalman Filter Particle Finder (KFPF) package [5] for online reconstruction and selection of short-lived particles in CBM, selection

*e-mail: shahidzafarkhan@gmail.com

criteria have been manually optimized to maximize signal to background ratio for a certain collision energy and a heavy-ion event generator. The selection criteria depend on the collision energy and centrality, decay channel and detector configuration. Machine Learning (ML) algorithms can be used to adjust these criteria automatically for different data taking scenarios. ML provides an efficient non-linear and multi-dimensional selection criteria optimization. To optimize selection criteria for Λ and other particles, ML can use many variables, associated with particle candidates, provided they are not strongly correlated with the invariant mass of the particle. In this work, we study ML performance using the same variables as used by the KFPF. Multiple ML libraries were tested for bench-marking using an automated machine learning (automl) package. The XGBoost [6] model outperformed other models in terms of numerical calculation speed and efficiency. XGBoost uses decision trees and a gradient descent optimization algorithm to minimize the loss function.

Details of the Λ reconstruction in CBM can be found in [7]. Cellular automaton and Kalman Filter Particle (KFParticle) based PFSimple package [7] are used for track finding, fitting and decay kinematics reconstruction. PFSimple interfaces the mathematics of the KFParticle package and provides a convenient interface to control the reconstruction parameters. The Λ candidates constructed from proton and π^- track pairs coming from a true Λ decay are termed as signal while all other candidates are considered as background.

The list of variables used for ML includes a squared distance between the daughter tracks and primary vertex (PV) divided by its error, the closest distance between the tracks (DCA), a squared distance between daughter tracks divided by its error, and distance between PV and secondary vertex divided by its error. A set of preselection criteria are applied to the data to remove numerical artefacts. A central Au-Au collision contains mostly combinatorial background and way less signal for Λ. Since the data has imbalanced classes, the data set used to train and test the ML algorithm is over-sampled by increasing the ratio of signal (under-represented class) to background candidates. Model A is treated as simulated signal and used for efficiency calculation while Model B is treated as background (which in real data analysis is taken from experimental data).

A sample of about 1M of Λ signal from Model A in the 5σ region around the Λ peak in the invariant mass distribution and a sample of about 3M of the background from Model B on the left and right hand sides outside of the Λ peak region and below 1.3 Gev/c^2 were used for the analysis. Using background from a different model (Model B) helps to reduce over fitting and dependence of the ML result on the signal model. These samples are divided in proportion 80% to 20% into train-test samples. To tune various parameters of the algorithm so that it fits better on the training data and performs well on the test data, Bayesian optimisation [8] package is used. It divides the train data into 5 folds and searches for optimal values of parameters using parallel computing. Area Under the Curve (AUC) parameter is used to select the best model. The trained model is then applied on the test data set and it returns a probability distribution between 0 and 1 for given input variables as shown in Figure 1 (left). Candidates which are close to 0 are most probably background and vice versa for candidates close to 1. Based on this probability distribution an approximate median significance (AMS) [9] is calculated to maximise signal to background ratio.

Results and discussion. The trained and tested XGB model, along with the AMS selection criteria, is then applied on two samples for both Models A and B containing 100k events each. Figure 1 (right) shows Λ candidates from Model B (UrQMD) before and after XGB selection is applied. For differential analysis the kinematic phase space is divided in transverse momentum (p_{T}) intervals of 0.5 GeV/c wide and laboratory rapidity (y_{lab}) intervals of 0.5 step size. Efficiency of the ML selection is calculated by dividing the number of remaining true signal (true positives in the confusion matrix) in the XGB selected Λ candidates by the number of input signal for each p_{T}-y_{lab} interval.

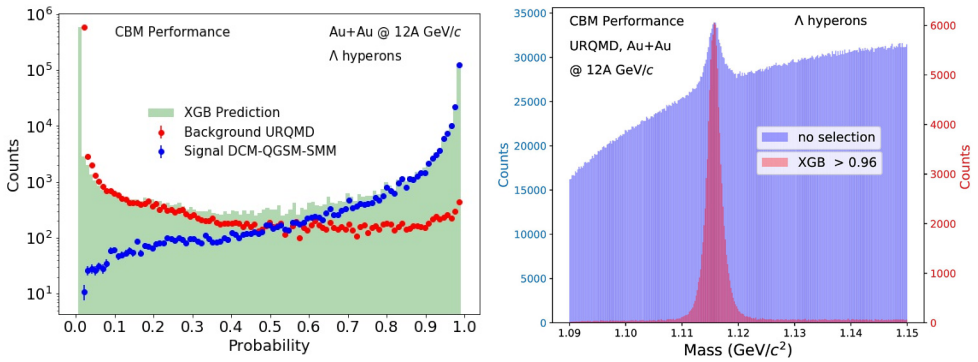

Figure 1. (left) Performance of the XGB algorithm for signal–background separation in the Λ candidates. The arrow represents the maximum AMS value at which selection was applied. (right) The invariant mass distribution of Λ candidates before (blue) and after XGB selection (red).

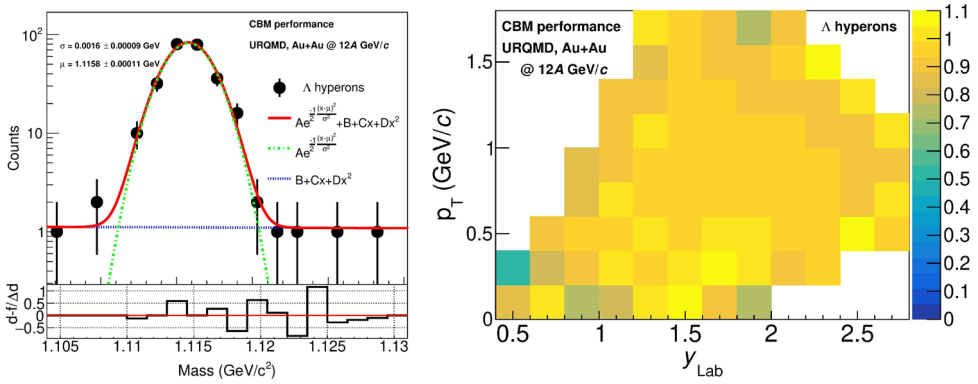

Figure 2. (left) Illustration of the procedure for Λ signal extraction after XGB selection is applied. Example is for $1.4 < p_T < 1.6$ GeV/c and $2.2 < y_{lab} < 2.4$ interval. The invariant mass yield (black circles) is fitted with the signal (Gaussian, green dotted line) and background (second order polynomial, blue line) shapes. (right) The p_T-y_{lab} distribution for the ratio of the fully corrected yield extracted after XGB selection to the simulated input.

The yield extraction is implemented as a three stage fitting procedure on Model B to each p_T-y_{lab} interval. Only the ranges which contain more than 400 candidates are considered. In stage one, the invariant mass distribution of the background, Λ candidates outside the 5σ region from the Λ peak at 1.115 GeV/c^2, is fitted with a second order polynomial ($pol2$). In the second stage, the invariant mass distribution is fitted in full range with a sum of a Gaussian ($A\exp[((x-\mu)/\sigma)^2/2]$) and $pol2$ functions. For this fit the mean (μ) and standard deviation (σ) of the Gaussian function are fixed to $\mu = 1.115$ GeV/c^2 and $\sigma = 0.0014$ GeV/c^2, while the initial values of the $pol2$ are taken from the fit at stage one. In the final stage, all parameters of the Gaussian and pol2 are released with their initial values set to the result of the fit at stage two. The number of entries under the Gaussian fit only, in the 2.5σ region around μ, are classified as Λ signal. Figure 2 (left) shows an example of the final result from the fitting procedure. The corrected Λ yield for each p_T-y_{lab} interval is extracted by dividing the signal

yield from the fitting procedure with the efficiency correction factor obtained from Model A. The corrected yield divided by the simulated yield is plotted in Figure 2 (right). Figure 3 demonstrates good agreement of the corrected yield with the simulated one as a function of p_{T} (left) and y_{lab} (right).

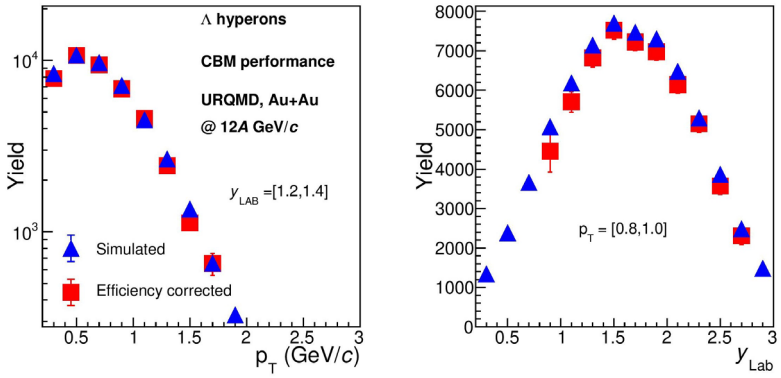

Figure 3. Yield of Λ hyperons as a function of (left) transverse momentum, p_{T}, for rapidity $1.2 < y_{\mathrm{lab}} < 1.4$ and (right) rapidity for $0.8 < p_{\mathrm{T}} < 1.0$.

Summary. The CBM performance for reconstruction of the Λ hyperon via its decay to proton and π^- in Au-Au collisions for the top FAIR SIS-100 beam momentum is presented. By combining the precise Kalman-Filter based reconstruction of the decay topology and machine learning algorithms a multi-differential analysis of the Λ yield in p_{T} and y_{lab} was performed. Reliable extraction of the Λ hyperon is demonstrated by comparison of the efficiency corrected yields after the reconstruction and ML selection with the simulated input for different heavy-ion event generators DCM-QGSM-SMM and UrQMD. In future, the method will be deployed for different collision energies and other strange and multi-strange particles and (hyper-)nuclei decays.

Acknowledgements. This work was supported by the Carlo and Karin Giersch Stiftung, HGS-HIRe Graduate School by HIC for FAIR, the Ministry of Science and Higher Education of the Russian Federation, Project "Fundamental properties of elementary particles and cosmology" No 0723-2020-0041, the Russian Foundation for Basic Research (RFBR) funding within the research project no. 18-02-40086, and the European Union's Horizon 2020 research and innovation program under grant agreement No. 871072.

References

[1] Fu et.al., Physical Review D **101**, 054032 (2020)
[2] Baznat et.al., Physics of Particles and Nuclei Letters **17**, 303 (2020)
[3] Bass et.al., Progress in Particle and Nuclear Physics **41**, 255 (1998)
[4] Allison et.al., IEEE Transactions on nuclear science **53**, 270 (2006)
[5] Zyzak, PhD thesis, 165 (2016)
[6] Chen et.al., Proceedings for "Knowledge discovery and data mining", **22**, 785 (2016)
[7] Lubynets et.al., Particles **4**, 288–295 (2021)
[8] Fernando Nogueira, https://github.com/fmfn/BayesianOptimization
[9] Adam-Bourdarios et.al., http://higgsml.lal.in2p3.fr/documentation, **9**, (2014)