

МІНІСТЕРСТВО ОСВІТИ ТА НАУКИ УКРАЇНИ  
Київський національний університет імені Тараса Шевченка  
Фізичний факультет  
Кафедра ядерної фізики та високих енергій

На правах рукопису

CBM performance for  $K^0_S$  meson measurement using  
Machine Learning

Галузь знань: 10 Природничі науки  
Спеціальність: 104 Фізика та астрономія  
Освітня програма: Фізика високих енергій

Кваліфікаційна робота  
магістра  
студентки 2 року навчання  
Лаворик Ольги Сергіївни

Науковий керівник:  
кандидат фіз. – мат. наук,  
доцент  
Безшийко О.А.  
Співкерівник:  
Dr. Ilya Selyuzhenkov (GSI,  
Germany)

Робота заслухана на засіданні кафедри ядерної фізики та високих енергій та  
рекомендована до захисту на ЕК, протокол No 15 від «19» 05 2022 р.

Зав. кафедри



Ігор КАДЕНКО

Київ-2022

ВИТЯГ  
з протоколу No \_\_\_\_\_  
засідання Екзаменаційної комісії

Визнати, що студентка Лаворик О. С. виконала та захистила  
кваліфікаційну роботу магістра з оцінкою \_\_\_\_\_.

Голова ЕК \_\_\_\_\_

«\_\_\_\_\_» \_\_\_\_\_ 2022 р.

## Анотація

**Лаворик О.С.** "Вимірювання  $K_s^0$  мезонів за допомогою машинного навчання в експерименті CBM "

*Кваліфікаційна робота магістра, спеціальністю 104 Фізика та астрономія, освітня програма «Фізика високих енергій». — Київський національний університет імені Тараса Шевченка, фізичний факультет, кафедра ядерної фізики. — Київ — 2022*

**Науковий керівник:** кандидат фіз. – мат. наук, доцент Безшийко О.А (КНУ ім. Т.Шевченка, Київ, Україна)

**Співкерівник:** Dr. Ilya Selyuzhenkov (GSI, Germany)

Продуктивність CBM для мультидиференційного вимірювання врожайності буде повідомлено про підтвердження дивного адрону  $K_s^0$ . Описано реконструкцію  $K_s^0$ , вилучення сигналу за допомогою алгоритму машинного навчання та процедуру обчислення виходів.

**Ключові слова:** ефективність, машинне навчання, CBM

## Abstract

**Lavoryk O.S.** "CBM performance for  $K_s^0$  meson measurement using Machine Learning"

*Qualifying work of the master on a speciality 104 — physics, specialization "high energy physics". — Taras Shevchenko National University of Kyiv, Faculty of Physics, Department of Nuclear Physics. — Kyiv, 2020.*

**Research supervisor:** Dr. Oleg Bezshyyko, TSKNU

**Co-supervisor:** Dr. Ilya Selyuzhenkov (GSI, Germany)

CBM performance for the multi-differential yield measurements of strange  $K_s^0$  hadron will be reported. Reconstruction  $K_s^0$ , signal extraction via machine learning algorithm and yield computation procedure is described.

**Key words:** yield, efficiency, machine learning , CBM

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>CBM experiment physics motivation</b>	<b>2</b>
2.1	Investigation of the QCD phase diagram	2
2.2	Dynamics of heavy-ion collision	2
2.3	Strangeness production as an evidence of the deconfined state	3
2.4	Motivation of $K_S^0$ yield measurement	4
2.5	FAIR - Facility for Antiproton and Ion Research	4
2.6	CBM experiment at FAIR	5
2.7	CBM detector setup	6
2.8	(Multi-)strange analysis workflow	8
<b>3</b>	<b>Short-lived particles reconstruction in CBM experiment</b>	<b>9</b>
3.1	PFSimple reconstruction	9
3.2	Selection variables	10
3.3	Quality assurance cuts	12
3.4	Data simulation	13
<b>4</b>	<b>Machine learning framework for analysis of particle decays</b>	<b>14</b>
4.1	Machine learning methology	14
4.2	Machine learning advantages	14
4.3	Hipe4ml	14
4.4	Input data for the ML algorithm	15
4.5	Correlation studies	16
4.6	Training variables	18
4.7	ML framework configuration with TOML	19
4.8	ROC curve and BDT threshold optimization	20
4.9	Confusion matrix	21
4.10	Variables importance	22
4.11	pT-rapidity distribution	22
4.12	XGBoost model	23
4.13	Model performance analysis	25
<b>5</b>	<b>Yields extraction</b>	<b>29</b>
5.1	Yield extraction procedure	29
<b>6</b>	<b>Summary</b>	<b>33</b>
<b>7</b>	<b>Acknowledgment</b>	<b>34</b>

# 1 Introduction

The Compressed Baryonic Matter (CBM) experiment at FAIR will investigate the QCD phase diagram at high net-baryon density ( $\mu_B > 400$  MeV) in the energy range of  $\sqrt{s_{NN}} = 2.7\text{-}4.9$  GeV. Precise determination of dense baryonic matter properties requires multi-differential measurements of strange hadron yields, both for most copiously produced kaons and  $\Lambda$  as well as for rare (multi-)strange hyperons and their anti-particles.

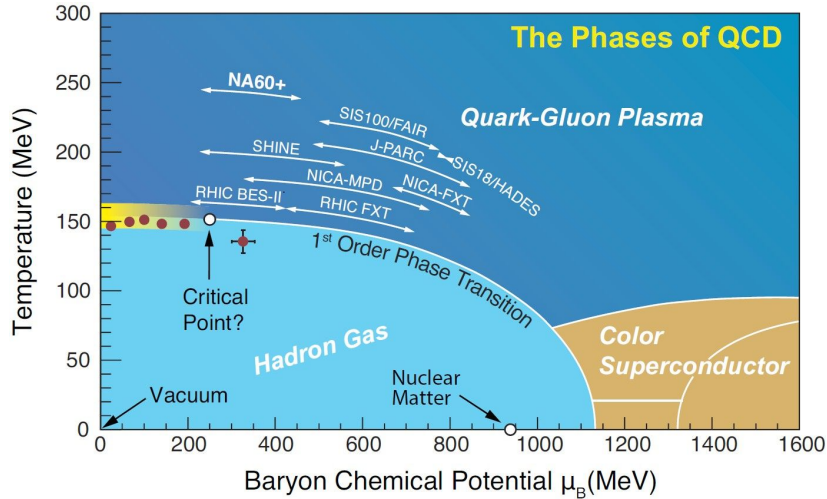
In this presentation, the CBM performance for the multi-differential yield measurements of strange  $K_s^0$  hadron will be reported. The strange hadrons are reconstructed via their weak decay topology using the Kalman Filter algorithm. Machine Learning techniques, such as XGBoost, are used for non-linear multi-parameter selection of weak decay topology, resulting in high signal purity and efficient rejection of the combinatorial background. Yield extraction and extrapolation to unmeasured phase space is implemented as a multi-step fitting procedure, differentially in centrality, transverse momentum, and rapidity at different collision energies. Variation of the analysis parameters allows estimating systematic uncertainties. A novel approach to study feed-down contribution to the primary strange hadrons using Machine Learning algorithms will also be discussed.

## 2 CBM experiment physics motivation

### 2.1 Investigation of the QCD phase diagram

The strong interaction between quarks and gluons is described by Quantum Chromodynamics (QCD). Diagram of strongly-interacting matter is shown in Fig. 2.1. At low temperature and moderate baryon chemical potential ( $\mu_B$ ) quarks and gluons are bounded within hadrons and cannot be observed in the free state. This phenomenon is called confinement. At high temperatures and baryon chemical potential quarks and gluons can be in unconfined state called Quark Gluon Plasma (QGP). The lattice QCD describes transition between these two phases at zero and low  $\mu_B$  and temperature about 150 MeV. There are some challenges how to extend this theory for higher  $\mu_B$  [1]. The most interesting phenomena is predicted at higher  $\mu_B$ . Going along the phase-space trajectory towards higher  $\mu_B$  values one expects to observe first-order phase transition; for extreme values of  $\mu_B$  color superconductivity state is predicted.

There are some ways to investigate the properties of the strongly interacting matter. Relativistic heavy ion collisions proved to be a useful tool for probing above processes in the laboratory.



**Figure 2.1:** Sketch of the phase diagram for strongly-interacting matter [2]

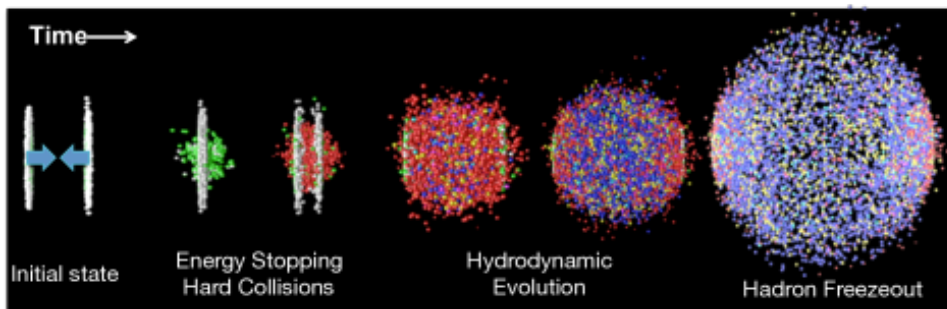
### 2.2 Dynamics of heavy-ion collision

The heavy-ion collision evolves through 4 stages:

- Initial state: two bunches approach each other at relativistic velocities

- Energy stopping hard collisions: QGP formation. Heavy ions approach to the distances of strong interaction. The system evolves while not reaching the thermal equilibrium. At this stage, the first order phase transition takes place.
- Hydrodynamic evolution: systems reaches the thermal equilibrium and expands.
- Hadron freez-out: system cools down and reaction products fly away to hadrons after the reaching the critical temperature

The evolution of the system happens within few femtoseconds; thus, its stages are indistinguishable time-wise for the observer: reaction products from all stages of the ion collision are being registered at once.



**Figure 2.2:** The stages of the heavy-ion collision [3]

### 2.3 Strangeness production as an evidence of the deconfined state

Strangeness enhancement is one of the most important probes of new deconfined state proposed in [4]. There are two main reasons for that: high temperatures at higher densities in heavy-ion collision and additional enhancement at large baryon densities.

The QCD Lagrangian has an approximate symmetry; in the limit of vanishing quark masses ( $m_q \rightarrow 0$ , where  $m_q$  are the quark masses entering the Lagrangian, i.e. the so-called “bare” or “current” masses), it reveals chiral symmetry.

At zero and low temperature the chiral symmetry is broken. With the rise of the temperature chiral symmetry is tend to be restored.

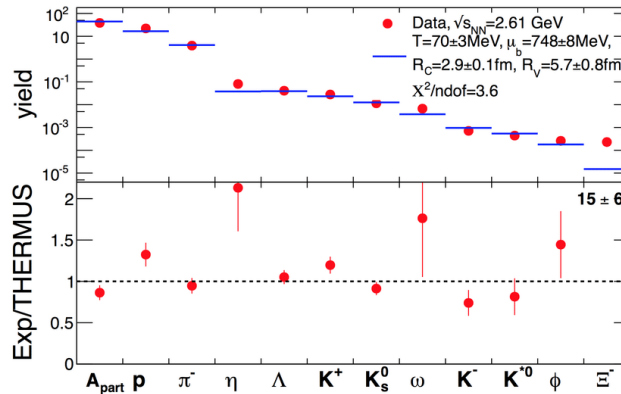
At low temperatures production of the strange quark is suppressed because of the large dynamical mass value. After chiral symmetry restoration the effective mass of the s quarks decreases, which leads to the strange hyperon production.

The second reason for the strangeness enhancement at energies 13-150 GeV per nuclei in laboratory frame rises when there is large baryon stopping. At these energies heavy-ion collision have large baryon density which corresponds to high baryonic potential. Therefore, if the hadronic matter is deconfined during the collision, the production of u and d quarks will be suppressed by Pauli blocking. Then at these energies we expect a global increase of strangeness production.

Strangeness enhancement is an important evidence of the deconfined state. Therefore strange hyperon precise reconstruction is important part of the CBM physics analysis [5].

## 2.4 Motivation of $K_S^0$ yield measurement

This thesis is focused on the  $K_S^0$  mesons. In the Lagrangian one can see that the neutral  $K^0$  consists of d and  $\bar{s}$  quarks and is a superposition of two weak eigenstates:  $K_S^0$  and  $K_L^0$  with different lifetimes.  $K_S^0$  has a mean lifetime  $0.8954 \times 10^{-10}$  s which corresponds to  $c\tau = 2.6844$  cm.  $K_S^0$  has invariant mass 0.4976 GeV. The most probable decay is  $K_S^0 \rightarrow \pi^+\pi^-$  with branching ratio  $69.20 \pm 0.05\%$  [6].



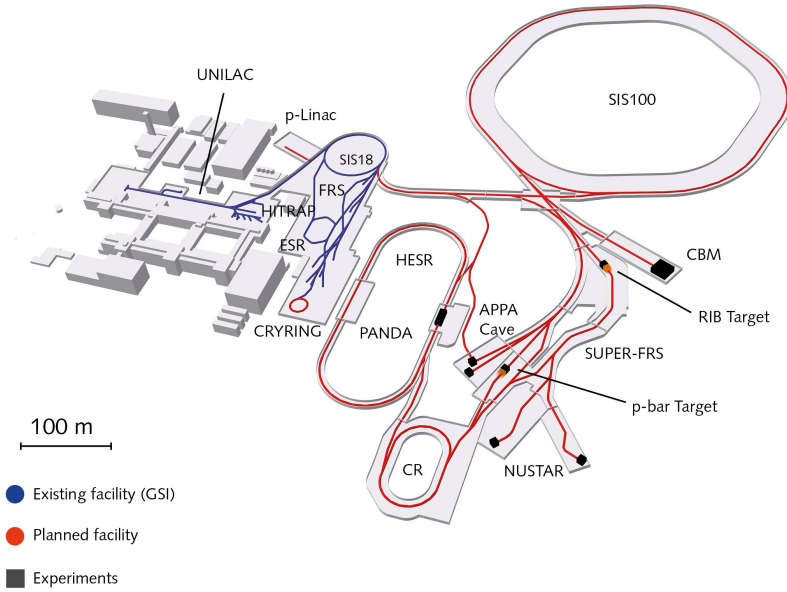
**Figure 2.3:** Multistrange hyperons production at SIS-100 energies [7]

## 2.5 FAIR - Facility for Antiproton and Ion Research

The Facility for Antiproton and Ion Research (FAIR) is a future accelerator complex at GSI, Darmstadt, which is designed to provide high-intensity heavy ion



beams with the SIS-100 accelerator ring with magnetic rigidity of 100 Tm. The beam kinetic energy range is 2–12A GeV for gold ions and 5–11 and 14–29 GeV for protons. The schematic plan of the FAIR accelerator complex is shown in Fig 2.4.



**Figure 2.4:** The schematic plan of the FAIR accelerator complex [8]

## 2.6 CBM experiment at FAIR

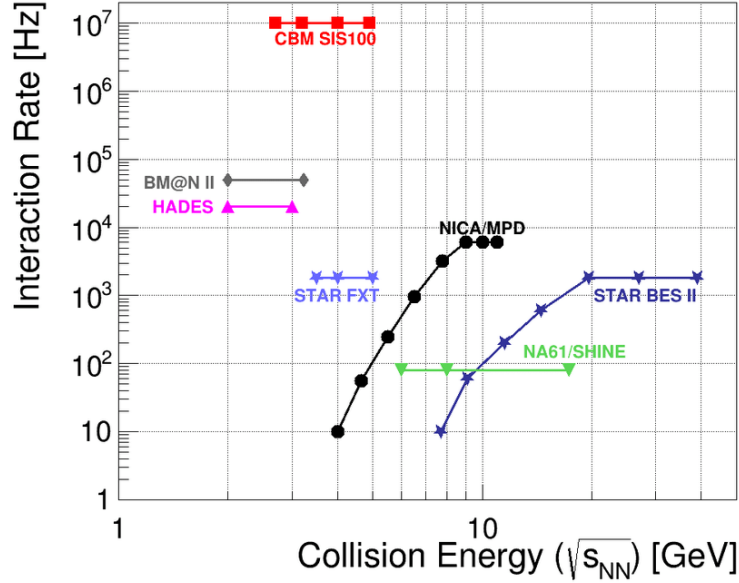
Compressed Baryonic Matter (CBM) is a future experiment at FAIR. CBM physics program includes study QCD matter in extreme conditions (high net-baryon densities, moderate temperatures), equation of state of nuclear matter at densities similar to the densities in the core of neutron stars.

The major observables to be studied during the experiment operation:

- particles containing strange or charm quarks: (multi-)strange hyperons ( $K$ ,  $\Lambda$ ,  $\Sigma$ ,  $\Xi$ ,  $\Omega$ ),  $J/\psi$
- light mass vector mesons decaying via dilepton channel
- the excitation functions of yields, spectra, and collective flow of these particles;
- the in-medium mass alteration of low-mass vector mesons;

- event-by-event fluctuations

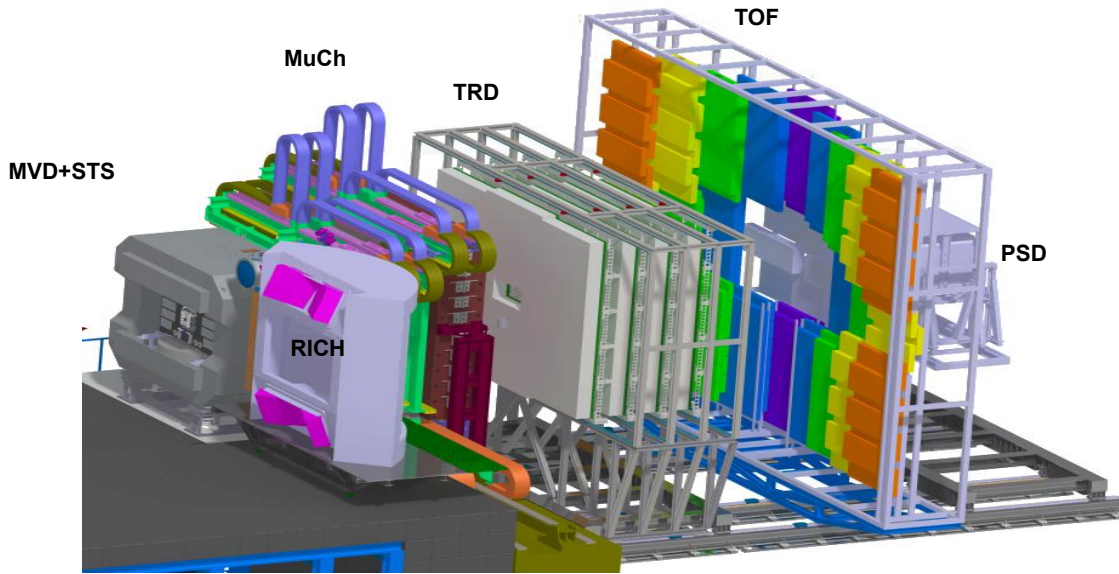
CBM will operate with event rate up to the  $10^7$  Au+Au collisions per second which will provide us with ample number of events containing the above phenomena. This is a uniquely high rate in comparison to other experiments.



**Figure 2.5:** CBM event rate with respect to the collision energy in comparison to other experiments [9]

## 2.7 CBM detector setup

CBM experiment is a single-arm forward spectrometer. There are several detectors composing the CBM experimental setup: they provide information necessary for the event building, tracking and vertex reconstruction and particle identification. The tracking system in the volume of the superconductive 1 Tm dipole magnet provides charged particle tracking and momentum measurements. It consists of 2 detectors: a Micro Vertex Detector (MVD) and a Silicon Tracking System (STS). To identify particles the PID setup is used along with tracking setup. The PID system has 4 parts: Ring Imaging Cherenkov (RICH), Muon Chamber (MuCh), Transition Radiation Detector (TRD) and Time-of-Flight wall (TOF). Projectile Spectator Detector (PSD) is used for collision geometry characterization.



**Figure 2.6:** CBM detector subsystems [10]

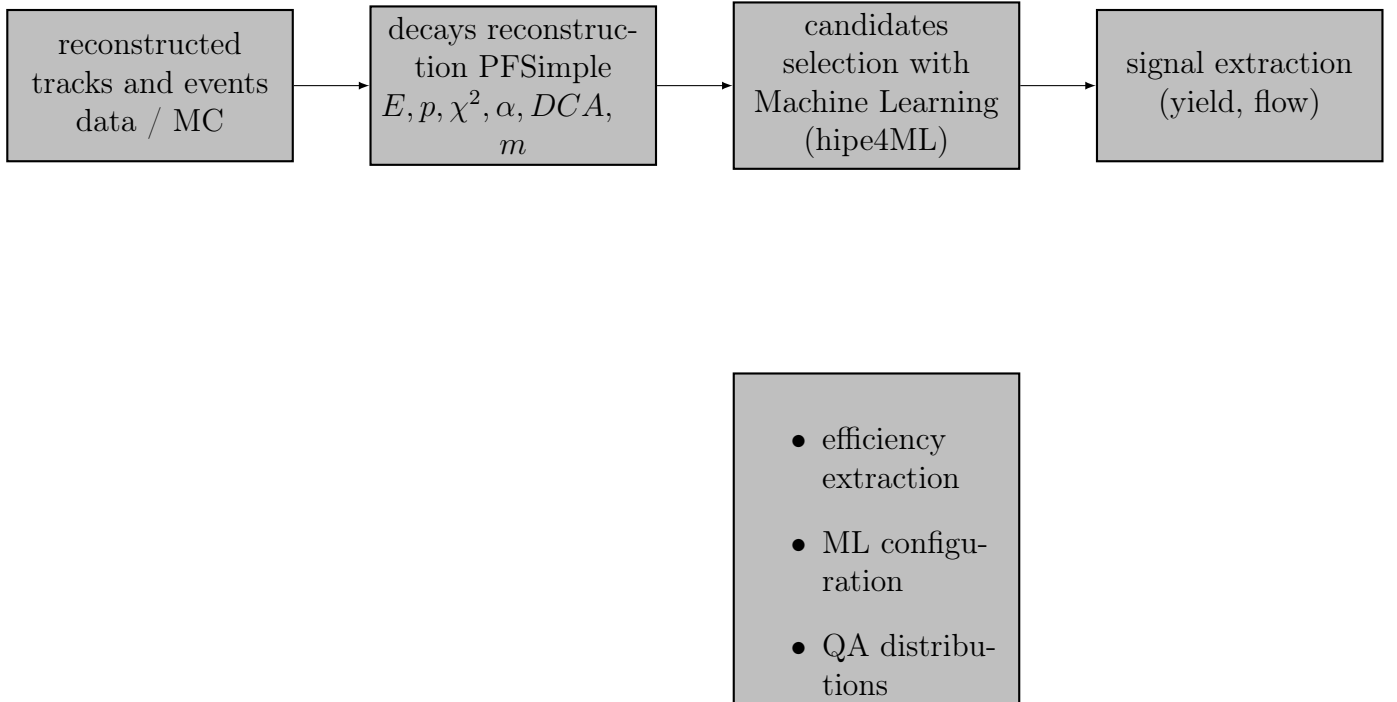
- a superconducting dipole magnet
- MVD is a system for track position resolution of few microns in the target region. It consists of four layers of silicon monolithic active pixel sensors.
- STS is a system for particle trajectory and momentum determination. It is based on double-sided silicon micro-strip sensors.
- MuCh is a system for muon identification consisting of a set of gaseous micro-pattern chambers sandwiched between hadron absorber plates made of graphite and iron.
- RICH is a system dedicated to the electron/pion discrimination. It is a detector comprising a CO<sub>2</sub> radiator and a UV photon detector realized with multianode photomultipliers for electron identification.
- TRD is a system for pion suppression, particle tracking, and identification using specific energy loss.
- TOF is a system for hadron identification. It is based on Multi-Gap Resistive Plate Chambers (MRPC) with low-resistivity glass.
- a Projectile Spectator Detector (PSD) is a system for centrality and reaction plane angle measurement. It is a hadron calorimeter with 44 modules.

TOF and STS are the most important detector components for short-lived particles reconstruction, because identification of charged particles is performed using Time-of-Flight technique, and STS provides information about particle's momentum.

## 2.8 (Multi-)strange analysis workflow

(Multi-)strange hyperon analysis is one of the most important analyses on the CBM experiment.

The first step includes track and event reconstruction from Monte Carlo simulated data of the particles interaction with the tracking system. Then one can reconstruct decays using and computing variables which describe decay's kinematic. After that one can select candidates with the use of machine learning techniques. This process comes with configuring of the machine learning model. One needs to estimate a quality assurance(to check if the model is not overfitted and able to generalize predictions) of the selection, as well as the efficiency. After the proper selection advanced analysis such as yields and flows computation takes place. Fig. 2.7 demonstrates the (multi-)strange analysis is performed in the CBM experiment.



**Figure 2.7:** The scheme of the CBM (Multi-)strange analysis workflow

## 3 Short-lived particles reconstruction in CBM experiment

### 3.1 PFSimple reconstruction

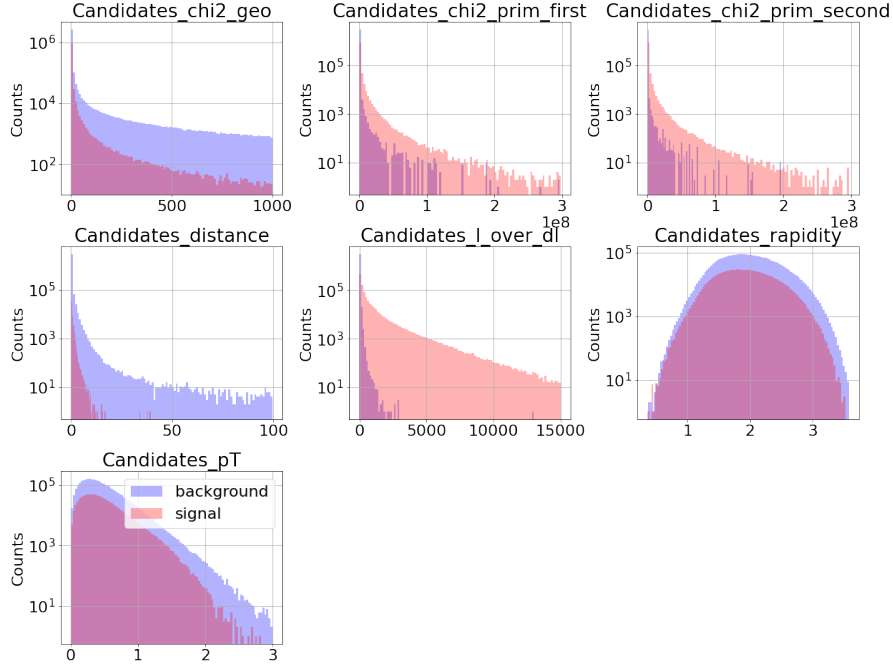
Short-lived particles can't be reconstructed directly. However, they can be reconstructed indirectly from their stable long-lived decay products. Stable particles trajectories are reconstructed in the tracking system and following parameters are measured: coordinates and momentum projections ( $x, y, z, p_x, p_y, p_z$ ).

For example, for  $K_S^0$  reconstruction, positive and negative charged pions are reconstructed in MVD and STS and identified using TOF. Then, every pion-pion pair is considered as  $K_S^0$  candidate. For this a point of the closest approach between daughter particles trajectories is found by numerical methods and parameters of the particles are extrapolated to this point. Then the obtained momentum and energy of daughters are summed up [11].

To perform this reconstruction efficient and fast tool KFParticle Finder was developed based on KF Particle package [12, 11].

The Particle Finder Simple package is simplified version of the KFPartice Finder package based on its mathematical apparatus. It is developed for the complete reconstruction of short-lived particles with their momentum, energy, mass, lifetime, decay length, rapidity, etc [13]. It takes input information about daughter particles(including track parameters, track charge, covariance matrix of track parameters etc) and returns as an output kinematics information and topological variables, matching with MC-true information for mother and daughter particles.

Fig. 3.1 shows the distributions of  $K_S^0$  baryon selection variables that were obtained from PFSimple reconstruction.



**Figure 3.1:**  $K_S^0$  baryon selection variables that were obtained from PFSimple reconstruction

To provide highest accuracy for further analysis we need to be sure that some daughter pair belongs to the signal, not background. For instance, in  $K_S^0 \rightarrow \pi^+\pi^-$  case signal is denoted as  $\pi^+\pi^-$  pair that originates from  $K_S^0$  decay. Other  $\pi^+\pi^-$  pairs are denoted as background.

KFParticle Finder is fast and more powerful online tool, while PFSimple has the same functional with more flexibility for physics analysis.

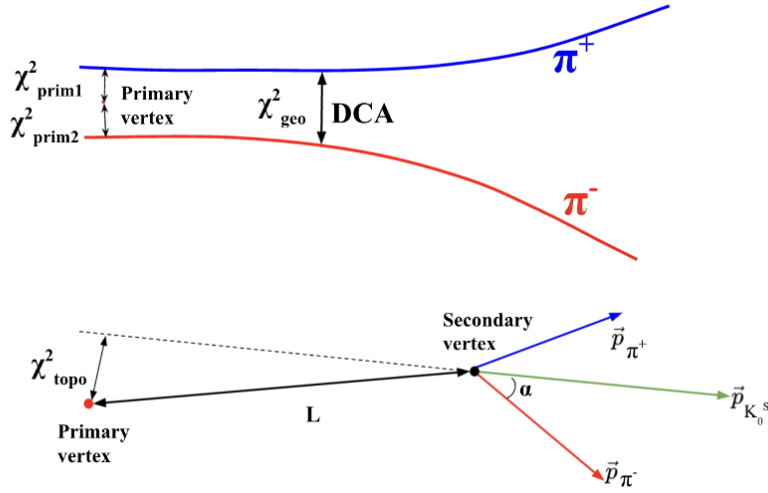
PFSimple provides manual optimization of selection criteria to distinguish signal from background. Unlike the current implementation, the selection criteria based on Machine Learning (ML) algorithms can be adjusted multi-dimensionally by the algorithm for different input data sets.

### 3.2 Selection variables

PFSimple returns selection variables. For  $K_S^0 \rightarrow \pi^+\pi^-$  decay these variables are:

- $\chi_{prim}^2$  - squared distance  $\Delta r$  between the daughter track and the primary vertex (PV) divided by its error covariance matrix

- DCA - distance of closest approach between  $\pi^+$  and  $\pi^-$  tracks
- $\chi_{geo}^2$  - squared distance  $\Delta r$  between daughter tracks divided by its error covariance matrix
- cosinepos - cosine of the angle between  $\vec{P}_{K_S^0}$  and  $\vec{P}_{\pi^+}$
- cosineneg - cosine of the angle between  $\vec{P}_{K_S^0}$  and  $\vec{P}_{\pi^-}$
- $L/\Delta L$  - distance between PV and secondary vertex, the point of lambda decay, divided over its error
- $\chi_{topo}^2$  - squared distance  $\Delta r$  between candidate trajectory and the PV divided by its error covariance matrix
- cosine topological - cosine of the angle between PV and point of  $K_S^0$  origin
- issignal - target binary variable which defines type of event (signal or background)



**Figure 3.2:**  $K_S^0$  selection variables. (a): variables associated with the decay tracks of  $K_S^0$  candidates. (b): variables associated with angles between  $\vec{P}_{K_S^0}$ ,  $\vec{P}_{\pi^+}$  and  $\vec{P}_{\pi^-}$

These variables need to be optimized, so we need an algorithm to find the minimum of the cost function in this multidimensional space.

### 3.3 Quality assurance cuts

Prior to the application of machine learning a data set needs to be cleaned. Cleaning refers to removing nan values, infinite and numerical artifacts. Nan( numeric data type used to represent any value that is undefined or unrepresentable) and infinite entries are dropped from the data and the numerical artifacts which don't have physical meaning and come from the mistakes of reconstruction are eliminated by applying the following selection criteria on the data:

- $\chi_{prim}^2, \chi_{geo}^2, \chi_{topo}^2 > 0$ , since  $\chi^2 > 0$
- $\chi_{prim}^2 < 3 \times 10^8, \chi_{geo}^2 < 10^3, \chi_{topo}^2 < 3 \times 10^5$  to reduce amount of data
- invariant mass  $> 0.28$  GeV; conservation of mass imposes that the mass of the  $K_S^0$  should be greater than the sum of the rest masses of  $\pi^+(0.139$  GeV) and  $\pi^-(0.139$  GeV)
- invariant mass  $< 1$  GeV to reduce amount of data
- $-5 < L < 80$  cm - Decay kinematics require that the distance between secondary vertex and PV should be equal or greater than 0, however, some candidates have negative L values. To reconstruct the track of a charged particle, the particle has to register hits in three stations of the tracking system, but the last two stations of the tracking setup are above 80 cm. Therefore, any candidate decaying above 80 cm will not be reconstructed
- $-25 < L/\Delta L < 15000$
- $-1 < z < 80$  cm for the same reason
- $DCA > 0$  cm; distance between two distinct points is always positive, therefore, proton and pion tracks for which distance of closest approach is negative will be discarded
- $DCA < 100$  cm; the largest station of the tracking setup i.e. the 8th station of STS has a surface area smaller than  $100 \text{ cm}^2$ , therefore, two tracks which have DCA greater than 100 are discarded [14].
- $|x|, |y| < 50$  cm restriction comes from the size of STS



- $p_z > 0$ ,  $p_T > 0$ ,  $p > 0$  because detector has a fixed target geometry
- $p < 20$  GeV,  $p_T < 3$  GeV to reduce amount of data
- $1 < \eta < 6.5$  because STS covers polar angle from  $2.5^\circ$  to  $25^\circ$ , where  $\eta = -\ln \tan(\frac{\theta}{2})$ ,  $\theta$  is a polar angle and magnetic field has an impact on pseudorapidity

The results of cuts application could be found here [15].

### 3.4 Data simulation

Data simulation for CBM experiment physics analysis is based on two Monte Carlo simulation packages for heavy-ion collision simulations: Ultra relativistic Quantum Molecular Dynamics(UrQMD) and Dubna Cascade Model and Statistical Multifragmentation Model(DCM-QGSM-SMM) [16, 17, 18, 19]. Both treat the production of new particles via formation and fragmentation of specific colored objects, strings. The differences between the models arise on different stages of a string formation and fragmentation [20].

UrQMD/DCM-QGSM-SMM simulated dataset containing  $\approx 3.6$  M events with Au+Au collisions at  $p_{beam} = 12.4$  GeV/ $c$  is analyzed in this study.

Simulation of interaction of these products with different parts of the detector was performed using GEANT4 [21, 22]. From hits tracks are reconstructed and put to the special data format AnalysisTree [23] or the ROOT format.

All pion tracks (MC PID is used) are combined into  $K_S^0$  candidates. Pion pair coming from a  $K_S^0$  decay is termed as signal (MC=1). Pion pair not originating from a  $K_S^0$  decay is considered as background (MC=0).

## 4 Machine learning framework for analysis of particle decays

### 4.1 Machine learning methodology

Machine learning (ML) is a complex of computer algorithms that learn to perform a specific task by learning from data. ML algorithms make predictions based on previously analysed information. It is widely-used in a wide range of applications: beam adjustment in accelerators, track finding and fitting, and data analysis in physics performance studies

A ML model is a mathematical model that has found patterns in the data and can now make predictions based on that. The part of making a model in which an ML algorithm learns from the data is referred to as training while the making predictions part is called as testing.

There are three ways to a train model: supervised, unsupervised and reinforcement learning.

Supervised learning is a way to teach model using previously labeled training dataset. Then during the testing stage the trained model is applied to unlabeled dataset.

This supervised machine learning approach can be also used to classify particle candidates.

### 4.2 Machine learning advantages

Machine learning approach has some advantages in comparison with manual selection. Existing KFPPF particle based selection criteria optimization maximizes 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 provides efficient multidimensional, automatized optimization of selection criteria.

### 4.3 Hipe4ml

Hipe4ml is selection optimization tool developed in ALICE Collaboration, minimal heavy ion physics environment for machine learning [\[24\]](#). It provides simple

and efficient functional for physics analysis, such as:

- Data selection
- Quality assurance of input data
- Finds Search for the best parameters for model to be trained(hyperparameters tuning)
- Model training and testing
- Model application to the user's specified data

Despite the hipec4ml basic analysis toolkit, some extended options are in need:

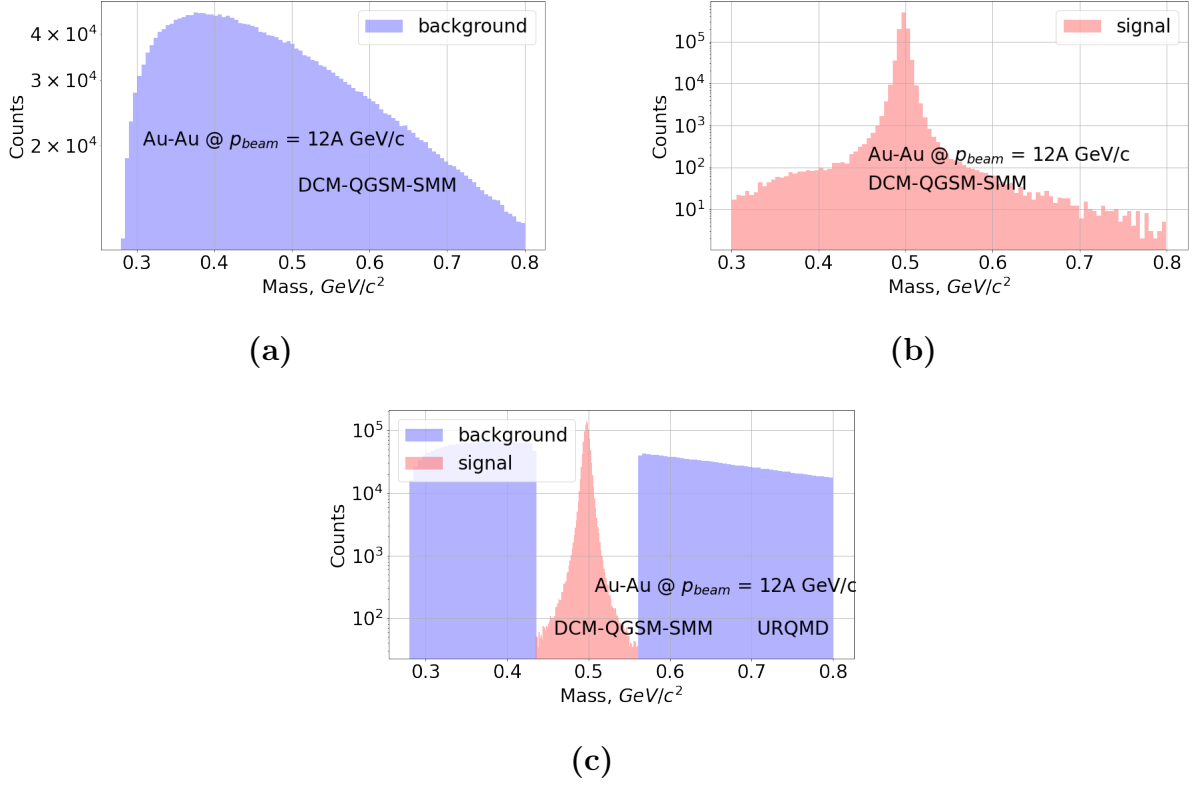
- Plot (non-)linear correlations
- Check results after selection
  - confusion matrix
  - possibility to visualize the selection
  - $p_T$ -rapidity distributions
  - variables distributions before and after ML cut (signal and background)
- Save model as C++ library

Integration with already existing hipec4ml is a solution of this problem [\[25\]](#).

#### 4.4 Input data for the ML algorithm

Machine learning algorithm requires some selection of the input data except the quality cuts.

The dataset consists of background samples generated from URQMD model and signal samples generated from DCM-QGSM-SMM model. We take signal samples within  $5\sigma$  range from  $K_S^0$  invariant mass peak(0.4976 GeV). We cut from background samples region within  $5\sigma$  range from  $K_S^0$  invariant mass peak. Fig. [4.1](#) shows how the dataset was created.



**Figure 4.1:** Data selection for train and test dataset

The ranges of signal and background invariant masses respectively:

- $0.4349 \text{ GeV} < m_{inv} < 0.5614 \text{ GeV}$
- $0.28 \text{ GeV} < m_{inv} < 0.4349 \text{ GeV}$  and  $0.5614 \text{ GeV} < m_{inv} < 1 \text{ GeV}$

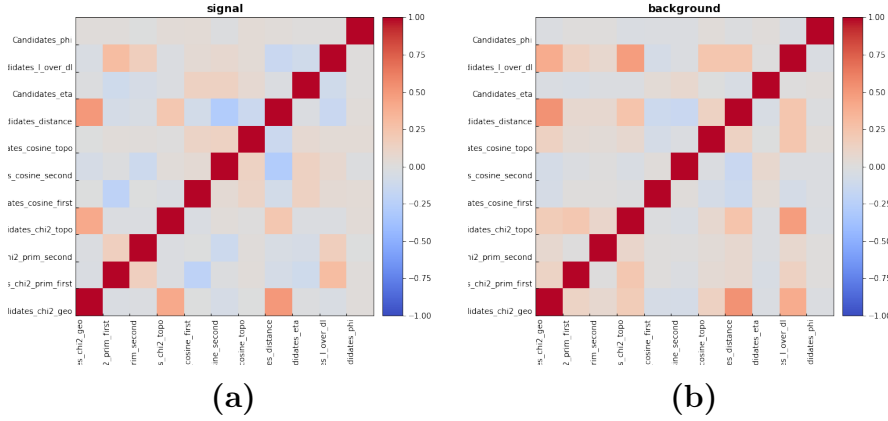
That was done to eliminate the bias of the algorithm, because it's not known if the area under the peak has some signal or not.

## 4.5 Correlation studies

Correlation study is an important aspect of training variables selection. If two variables have strong correlation, one of them should be excluded from the analysis to make the learning algorithm faster and decrease bias. If one of the variables has strong correlation with invariant mass it should be eliminated, as well, to exclude model's bias.

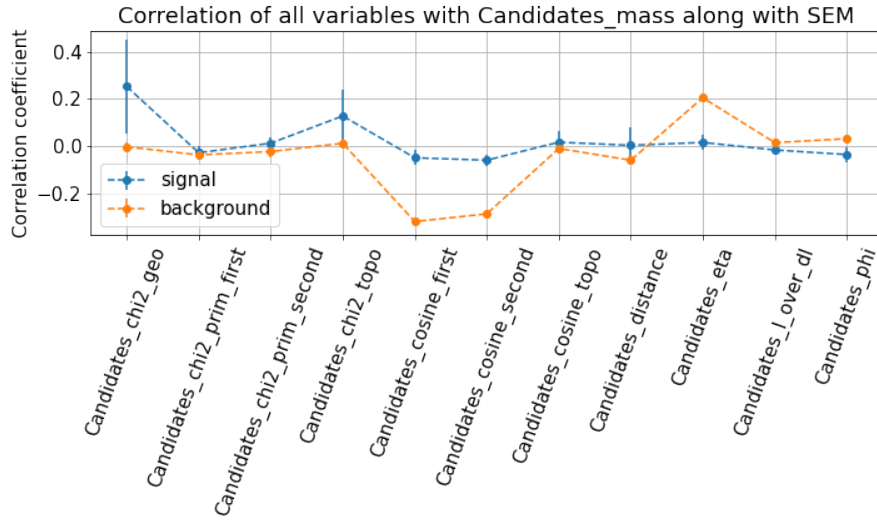
Correlation between invariant mass and all the variables is checked to make sure that certain variable could be included to the analysis. Firstly, Pearson correlation coefficient  $\rho_{X,Y} = \text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$  is calculated for all variables with invariant mass and the matrix is shown in Fig [4.2](#)

$\text{Cov}(X,Y) = E[(X-E[X])(Y-E[Y])]$ , where  $E[X]$  is expected value known as the mean of  $X$ ,  $\sigma_X$  is standard deviation,  $\sigma_X = \sqrt{E[X^2] - E[X]^2}$ .



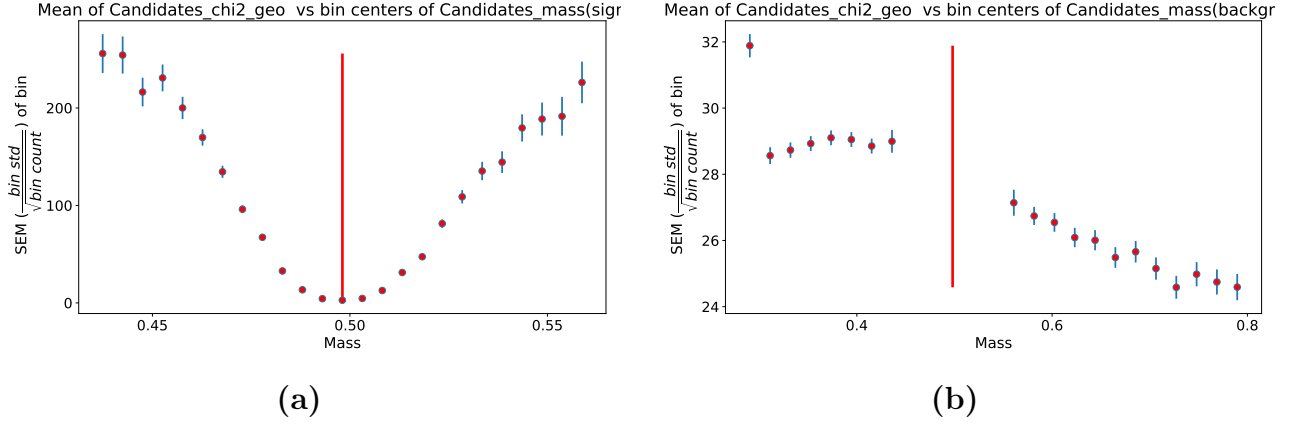
**Figure 4.2:** Pearson coefficient correlation matrix after applying quality cuts. (a): signal. (b): background

Standard error of the mean (SEM),  $\frac{\sigma}{\sqrt{n}}$ . Correlations of all variables with invariant mass along with SEM are shown in Fig. 4.3.



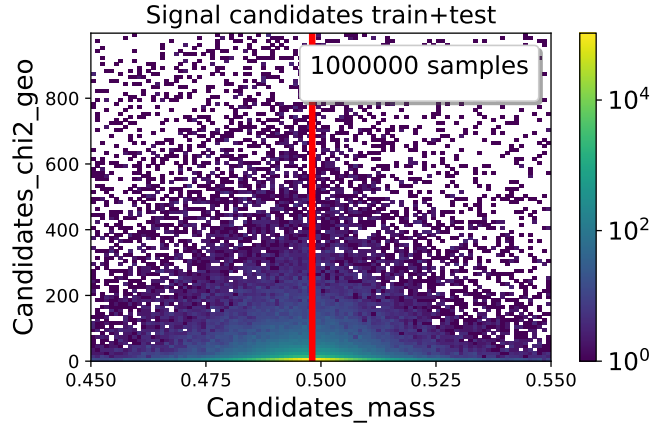
**Figure 4.3:** Correlations of all the variables with invariant mass

Since Pearson correlation coefficient only shows linear correlation between two variables, therefore, a different approach is also followed. The variable to be checked for correlation is taken and its data is divided into 25 bins. Similarly, the distribution of invariant mass variable is also divided into 25 bins. Fig. 4.4 shows the correlations for  $\chi_{geo}^2$  variable.



**Figure 4.4:** Mean of each bin with the SEM of  $\chi^2_{geo}$  bin versus mass. Bin center of invariant mass plotted versus the bin center of invariant mass. The error bars represent the SEM of each bin and the red line shows the mean of the  $K_S^0$  peak according to PDG value. (a): signal. (b): background.

One can build 2D correlation plot between variables and invariant mass to detect suspicious structures.

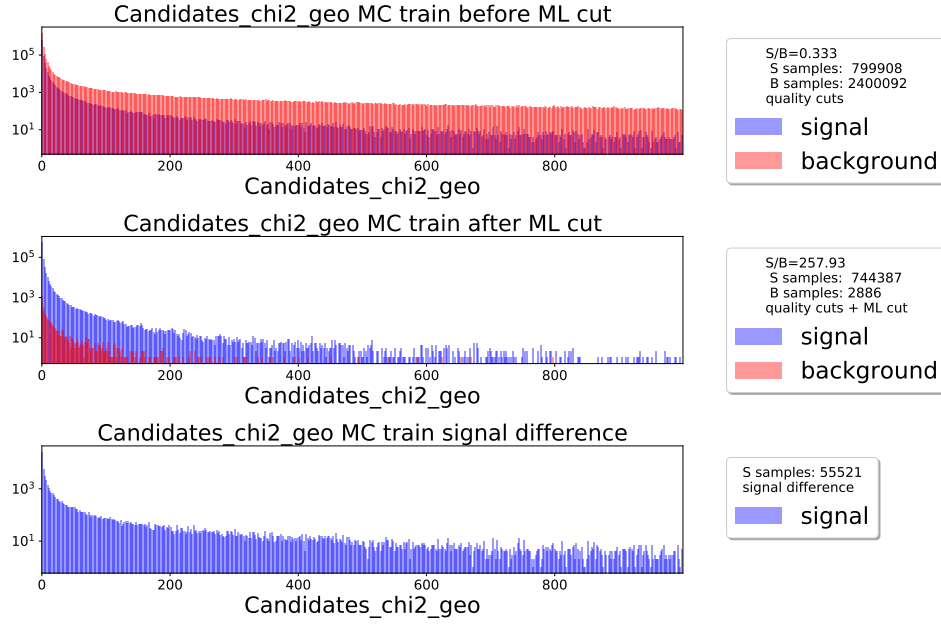


**Figure 4.5:** Example of correlation between variable and invariant mass plot

## 4.6 Training variables

Correlation studies showed that the following variables are the most suitable for the model training:  $\chi^2_{prim}$ ,  $\chi^2_{geo}$ , DCA,  $L/\Delta L$ . These variables do not strongly correlate neither with invariant mass nor with each other.

The one can easily check the distribution of each variable before and after XGBoost selection(Fig. 4.6).



**Figure 4.6:** Distribution of the  $\chi_{geo}^2$  variable before and after XGBoost selection

## 4.7 ML framework configuration with TOML

At the beginning of the work user specifies input data (signal and background path, candidate mass etc) to configuration file in TOML format. TOML format [26] is a minimal configuration file format that's easy to read due to obvious semantics. It was designed to map unambiguously to a hash table and is easy to parse into data structures in a wide variety of languages.

```
[signal]
path = "/home/olha/CBM/dataset10k_tree/dcm_1m_prim_signal.root"
tree = "PlainTree"

[background]
path = "/home/olha/CBM/dataset10k_tree/urqmd_100k_cleaned.root"
tree = "PlainTree"

[log_scale]
variables = [
  'chi2geo',
  'chi2primneg',
  'chi2primpos',
  'chi2topo',
  'distance'
]

[non_log_scale]
variables = ['cosineqneg', 'cosineqpos', 'cosinetopo', 'mass', 'pT',
  'rapidity', 'phi', 'eta', 'x', 'y', 'z', 'px', 'py', 'pz',
  'l', 'ldl']

[peak_range]
peak_mass = 1.115683
sgn_left_edge = 1.108
sgn_right_edge = 1.1227
bgr_left_edge = 1.07
bgr_right_edge = 1.3

[number_of_events]
number_of_signal_events = 10000
number_of_background_events = 50000
```

**Figure 4.7:** Implemented for CBM: User can specify parameters via configuration files

## 4.8 ROC curve and BDT threshold optimization

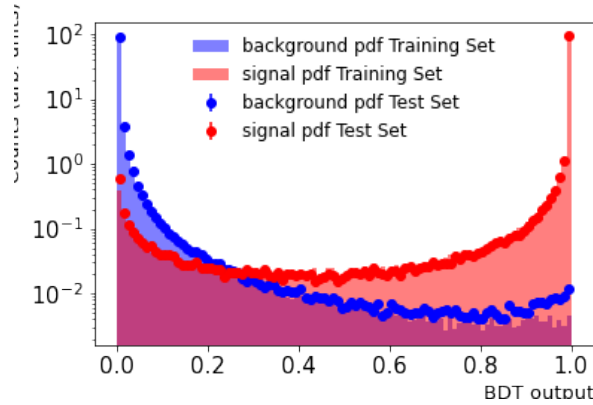
Receiver operating characteristic(ROC) curve is the plot that illustrates binary classifier performance as its discrimination threshold is varied.

Any classifier isn't able to separate signal samples from background. Some of them anyway will be classified as background. After classification all the samples are compared with their Monte Carlo label. If the sample's MC label is signal(MC=1) and it was classified as signal it is denoted as true positive. If the sample's MC label is background(MC=0) and it was classified as signal it is denoted as false positive.

ROC curves typically feature true positive rate on the Y axis, and false positive rate on the X axis. This means that the top left corner of the plot is the "ideal" point - a false positive rate of zero, and a true positive rate of one. This is not very realistic, but it does mean that a larger area under the curve (AUC) is usually better.

ROC curves are typically used in binary classification to study the output of a classifier [27].

There are two ways to choose optimal BDT cut. The first approach requires to check BDT outputs distribution Fig.4.8 and choose the optimal value.



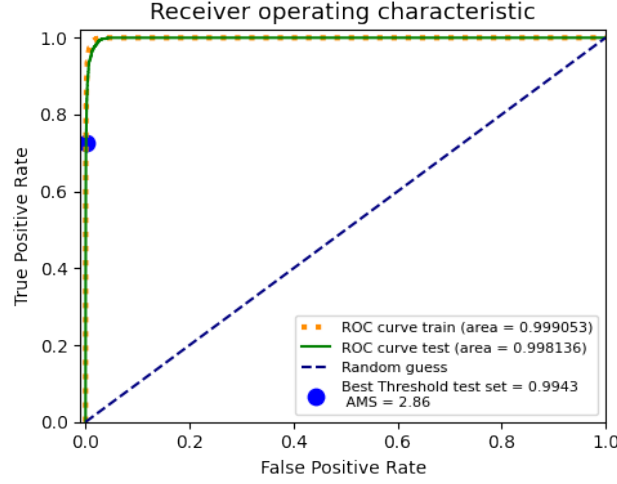
**Figure 4.8:** BDT outputs distribution

The second one requires to choose an optimal threshold with respect to the metric optimization. In this work it was decided to use the Approximate Median Significance(AMS) [28] which was used for The Higgs Machine Learning Challenge.



$$\text{AMS} = \sqrt{2(tpr + fpr) \ln \left(1 + \frac{fpr}{tpr}\right)} - tpr$$

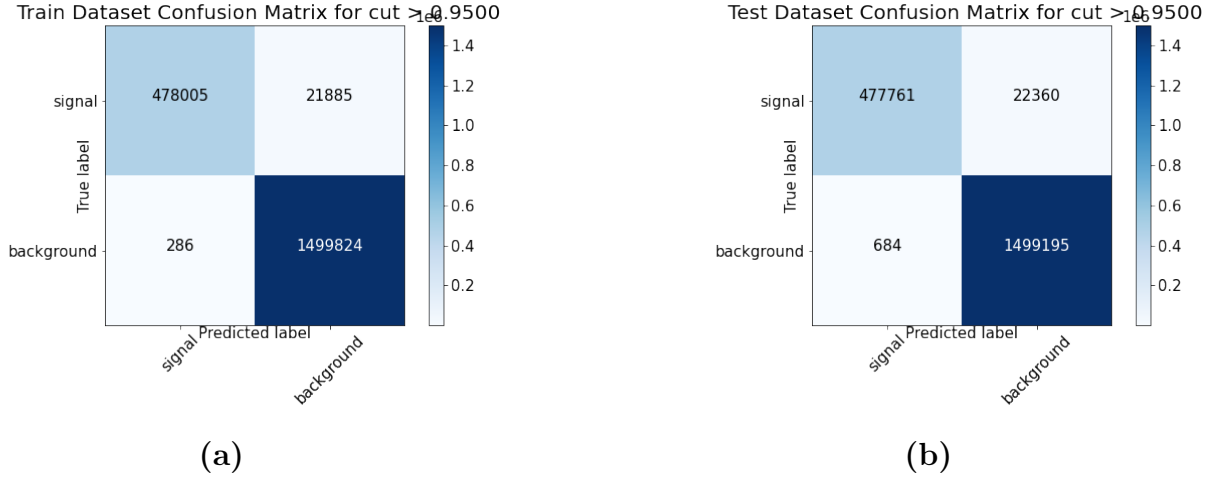
Fig. 4.9 demonstrates roc curve with the optimal threshold with respect to AMS maximum.



**Figure 4.9:** Receiver operating characteristic

## 4.9 Confusion matrix

Confusion matrix is the matrix that describes binary classifier performance. It shows how many real(Monte Carlo=1) signal samples were classified as signal and how many as background. Main diagonal consists of samples with predicted label the same as Monte Carlo label. Side diagonal consists of samples with MC label doesn't coincide with prediction. Fig. 4.10 shows the confusion matrix for BDT cut  $>0.95$  for training and testing datasets.

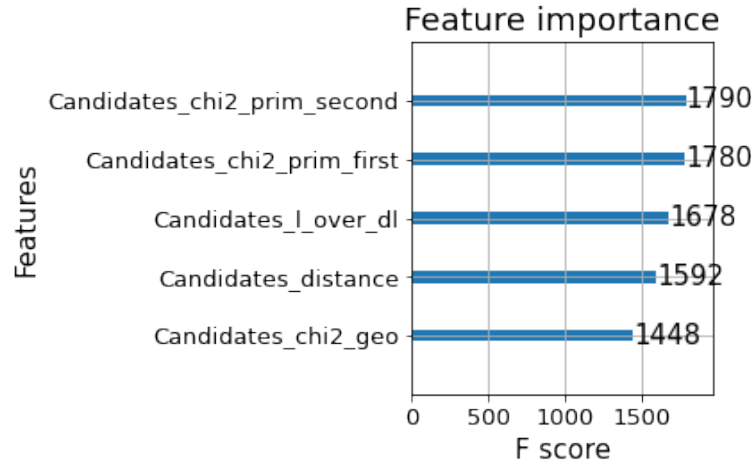


**Figure 4.10:** Confusion matrix of training(a) and testing(b) dataset

#### 4.10 Variables importance

XGBoost provides features importance rank that allows to estimate how useful or valuable each feature was in the construction of the boosted decision trees within the model [29].

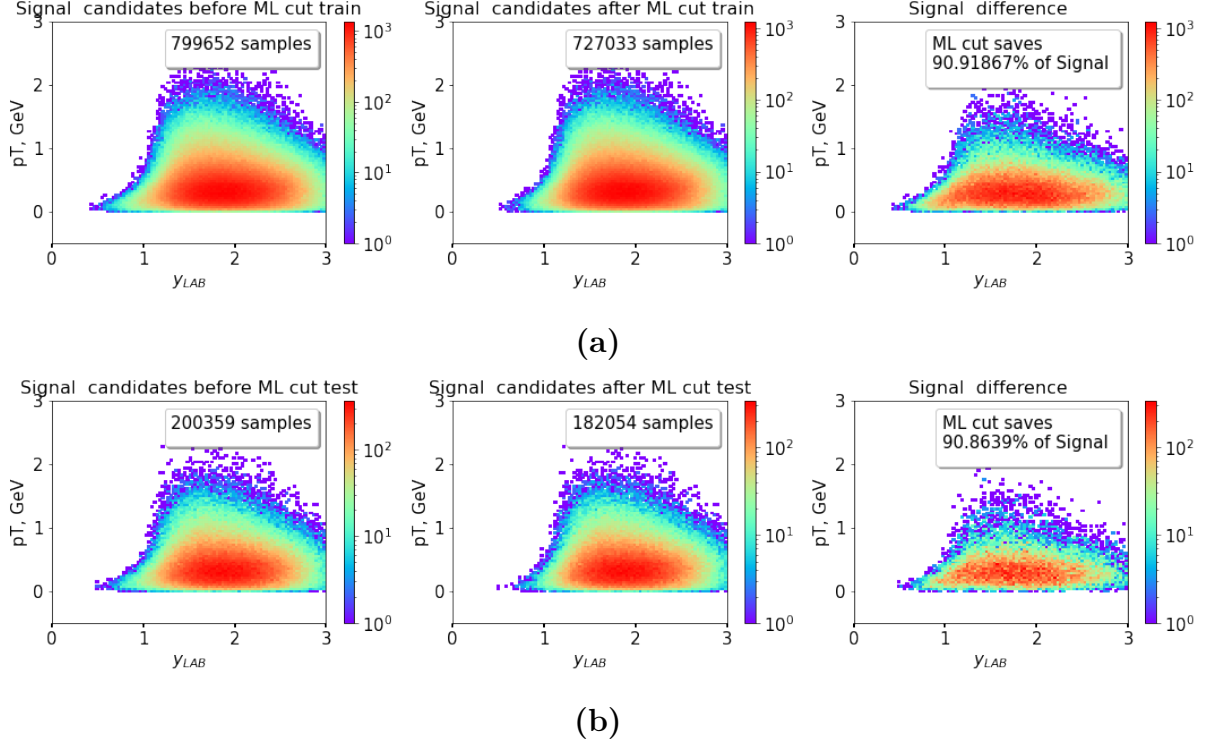
Fig. 4.11 shows the feature importance which was done during the training of the machine learning algorithm.



**Figure 4.11:** XGBoost variables importance

#### 4.11 pT-rapidity distribution

Input for selection efficiency determination is implemented. Efficiency is determined as  $\epsilon = \frac{N_{rec}}{N_{all}}$ , where  $N_{rec}$  is number of reconstructed samples,  $N_{all}$  - number of reconstructable (the samples reconstructed while no selection is applied) samples.



**Figure 4.12:**  $p_T$ -rapidity differential selection and reconstruction efficiency training (a) and testing (b) dataset

The one can see that BDT cut saves approximately 90.8% of the signal samples. These plots could show the potential bias of the model. If the tails of the  $p_T$ -rapidity distribution are cut non-uniformly, then the model is biased. In Fig. 4.12 once can observe no bias.

## 4.12 XGBoost model

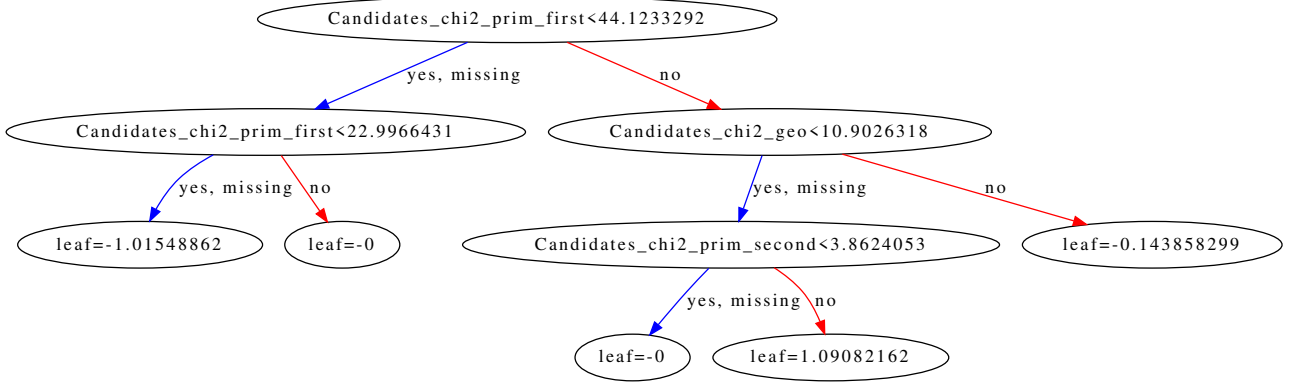
This framework is based on the XGBoost machine learning algorithm. XGBoost is a decision tree ensemble based on gradient boosting designed to be highly scale-able. XGBoost builds an additive expansion of the objective function by minimizing a loss function. XGBoost provides efficient and fast performance, so it was decided to use it for the current study.

This algorithm's advantages include:

- Boosting combines weak learners (error rate  $<50\%$ ) to make a strong learner (error rate  $<25\%$ )
- Decision trees (weak learners) are combined together to make a Gradient Boost algorithm

- In each step a new tree is used to improve the previous prediction

The example of the decision tree for  $K_S^0$  candidates selection is shown in the Fig. 4.13 using XGBoost optimized selection criteria. This selection criteria was optimized for small dataset with 1000 signal candidates and 3000 background candidates, so the structure of the tree is simplified in comparison with bigger datasets.



**Figure 4.13:** The example of the decision tree using XGBoost optimized selection criteria. This optimization was performed on 1000 signal and 3000 background candidates.

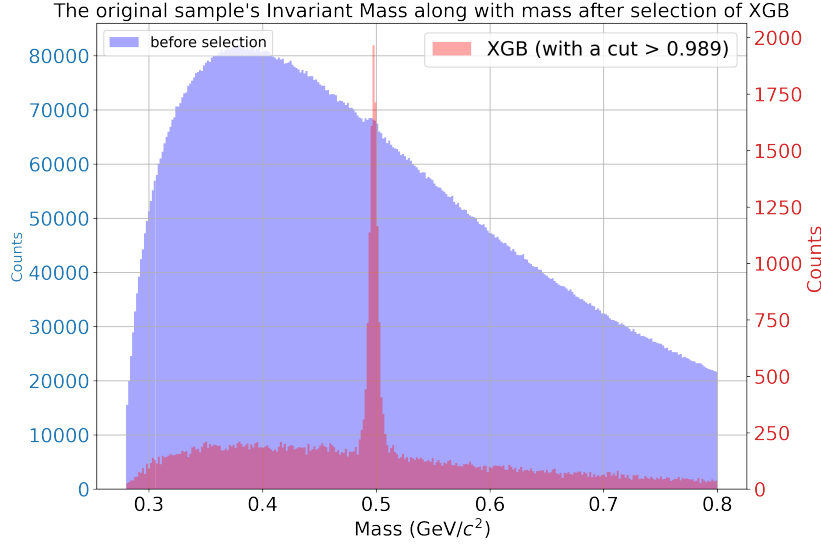
The optimal XGBoost parameters were optimized via Bayesian Optimization [30] as following:

- max\_depth(optimized value = 8): maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit.
- gamma (optimized value = 0.975): minimum loss reduction required to make a further partition on a leaf node of the tree.
- alpha(optimized value = 16) L1 regularization term on weights.
- learning\_rate(optimized value = 0.732): step size shrinkage used in update to prevents overfitting.

Model had 2 training epochs. Dataset includes 1000000 signal samples(MC=1) and 3000000 background samples(MC=0).

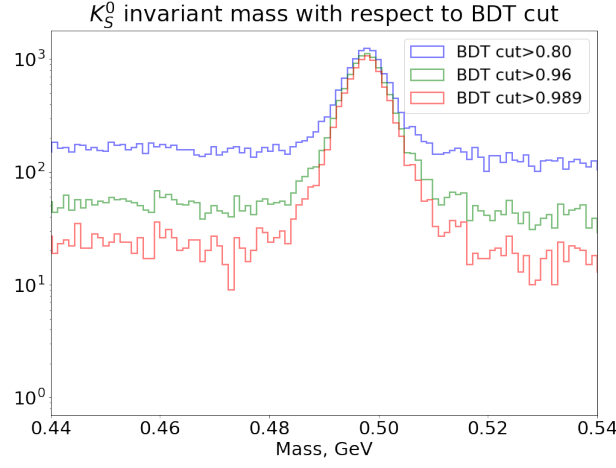
### 4.13 Model performance analysis

One can precisely optimize selection criteria with using XGBoost algorithm. Fig. 4.14 shows the  $K_S^0$  candidates invariant mass distribution before and after XGB optimized selection criteria application. This dataset consists of 10000 events with reconstructable 16113794  $K_S^0$  candidates. With the BDT cut  $>0.989$  we have ratio reconstructable  $K_S^0$  / reconstructed  $K_S^0 = 0.915$ .



**Figure 4.14:**  $K_S^0$  candidates invariant mass distribution before and after XGB optimized selection criteria application

The number of the reconstructed candidates depends on the BDT cut. Tighter cut rejects more background while rejection more signal candidates. The impact of the BDT cut is shown in the Fig. 4.15.



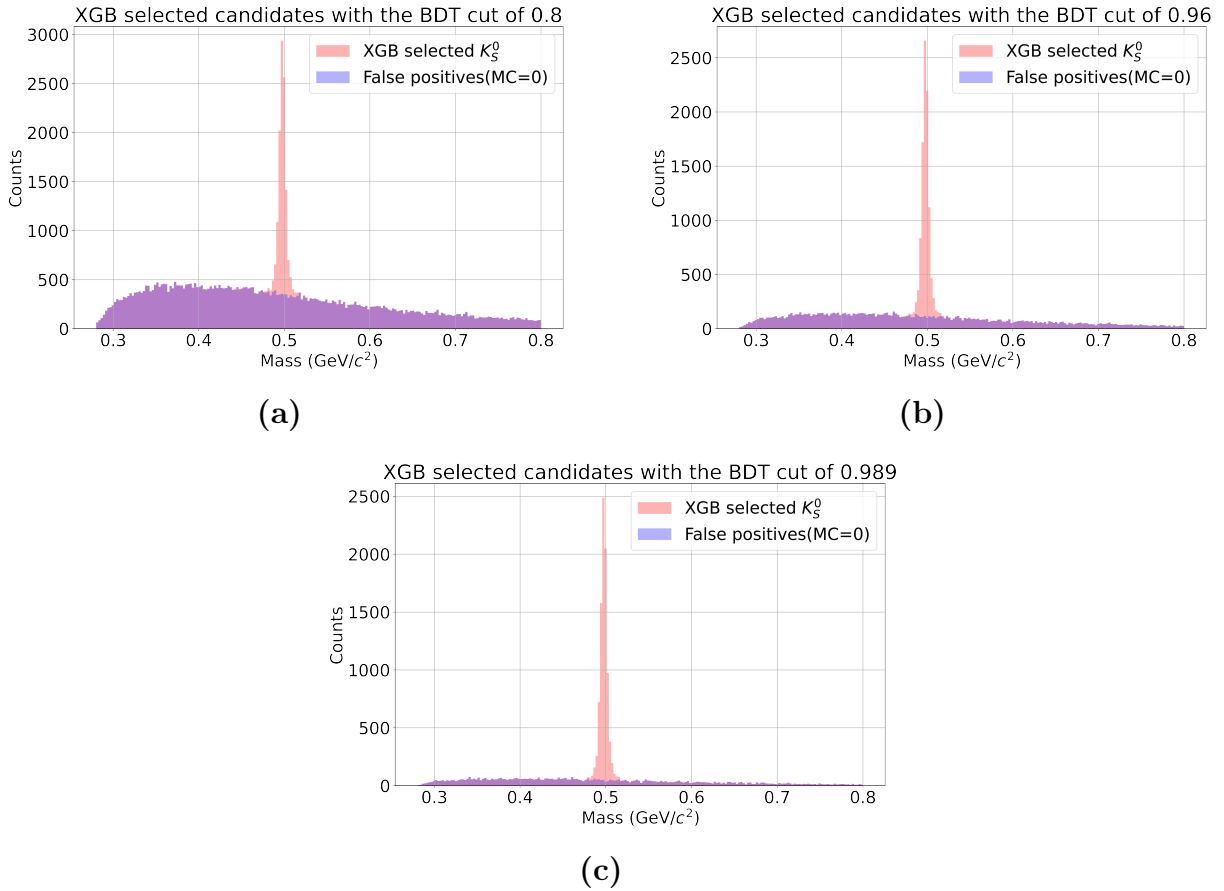
**Figure 4.15:**  $K_S^0$  candidates invariant mass distribution with respect to BDT cuts

Fig. 4.16 shows how the application of the BDT cut transforms background shape and make it linear.

Tab. 1 shows how many signal candidates are reconstructed and how many background is rejected with respect to the BDT cut.

BDT cut	reconstructable/reconstructed	background rejection
0.8	94.95%	99.66 %
0.96	90.65%	99.89 %
0.989	85.15%	99.95%

**Table 1:** Ratio between reconstructable/reconstructed  $K_S^0$  sigal candidates; background rejection - ratio between candidates which were classified as back-ground and false true(MC=0) candidates. Both values are with respect to the BDT cut.

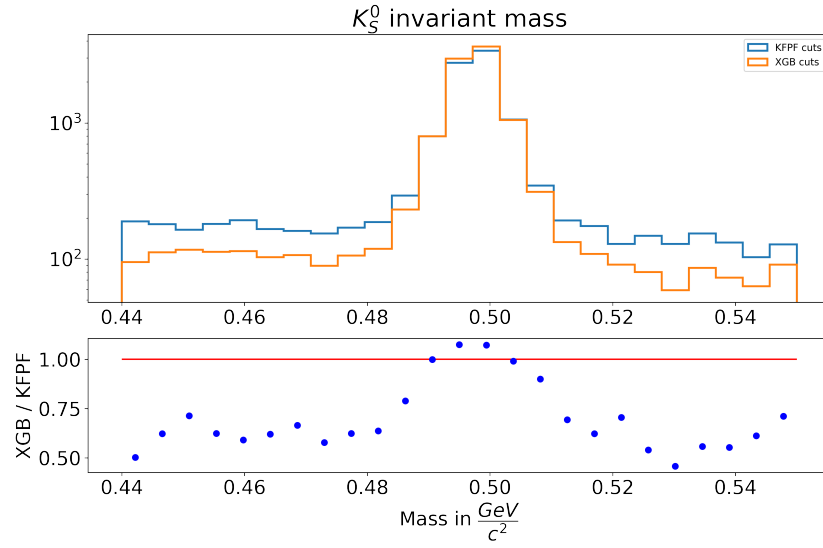


**Figure 4.16:**  $K_S^0$  candidates invariants mass distribution with respect to BDT cuts. The red peak shows XGBoost selected  $K_S^0$ , while blue shape shows false positive samples which are classified as signal while being background

Machine learning algorithms select signal more efficiently in comparison with manual tools(Fig. 4.17). Existing KFPPF package selection criteria optimization is based on the maximization signal to the background ratio. The default KFPPF-optimized selection criteria:

- $L\Delta L > 5$

- $DCA < 1 \text{ cm}$
- $\chi^2_{geo} < 3$
- $\chi^2_{prim} > 18.4$



**Figure 4.17:** Comparison between KFPP selection criteria and XGB



## 5 Yields extraction

### 5.1 Yield extraction procedure

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 (shown in Fig. 5.1).

The yield extraction is implemented as a three stage fitting procedure on URQMD to each  $p_T$ - $y_{LAB}$  interval [31].

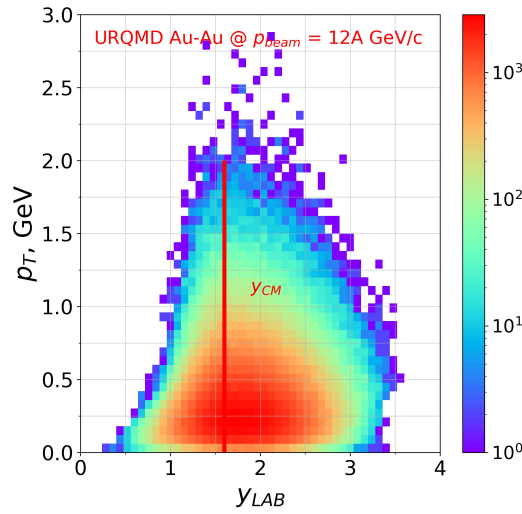


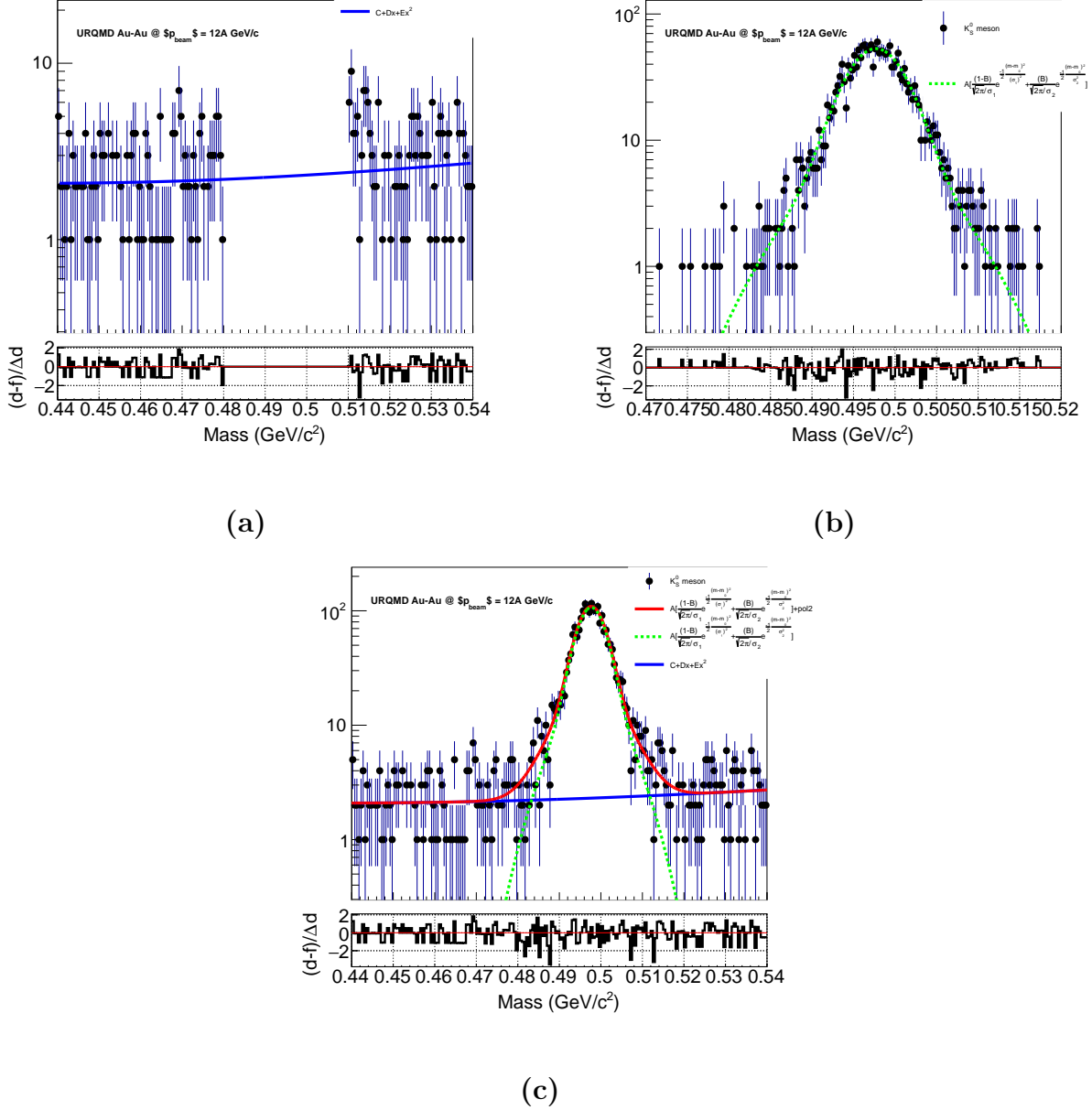
Figure 5.1

In stage one, the invariant mass distribution of the background,  $K_S^0$  candidates outside the  $5\sigma$  region from the  $K_S^0$  peak at 0.498 GeV/c<sup>2</sup>, 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 Double Gaussian ( $A/(1-B) \exp[(x - \mu)/\sigma_1]^2/2] + B \exp[(x - \mu)/\sigma_2]^2/2]$ ) and pol2 functions. For this fit the mean ( $\mu$ ) and standard deviations ( $\sigma_1$ ,  $\sigma_2$ ) of the Gaussian function are fixed to  $\mu = 0.498$  GeV/c<sup>2</sup> and  $\sigma_1 = 0.004$  GeV/c<sup>2</sup>,  $\sigma_2 = 0.007$  GeV/c<sup>2</sup> 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. Fig. 5.2 shows three stage fitting procedure.

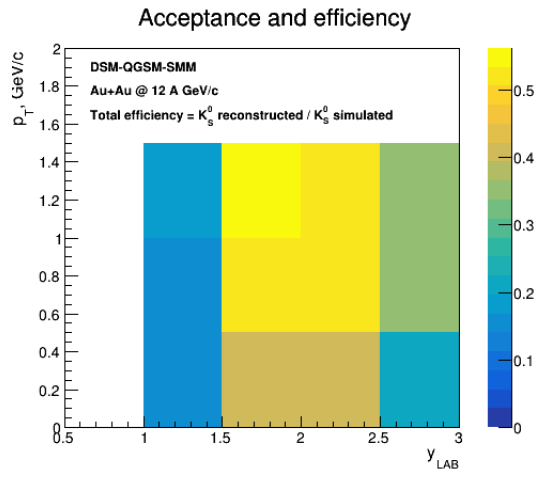
The corrected  $K_S^0$  yield for each  $p_T$ - $y_{lab}$  interval is extracted by dividing the signal yield from the fitting procedure with the efficiency correction factor ob-

tained from DSM-QGSM-SMM Model. In Fig. 5.3 reconstructed  $K_S^0$  acceptance and efficiency, corrected  $K_S^0$  yield, and corrected  $K_S^0$  yield and true  $K_S^0$  yield are shown.

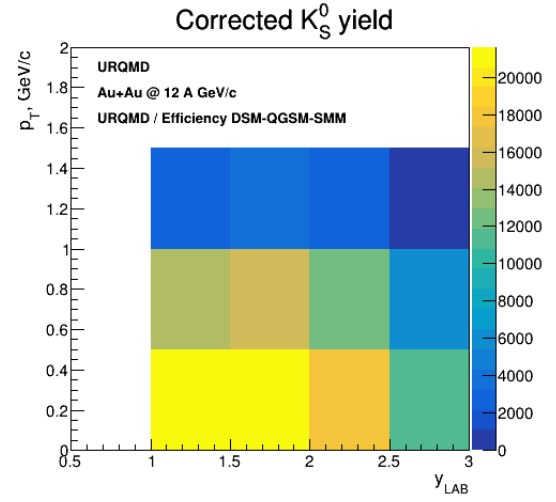
In Fig. 5.4 efficiency and acceptance corrected  $K_S^0$  yields on the  $y_{LAB}$  projection are shown. They are in the agreement with the simulated ones.



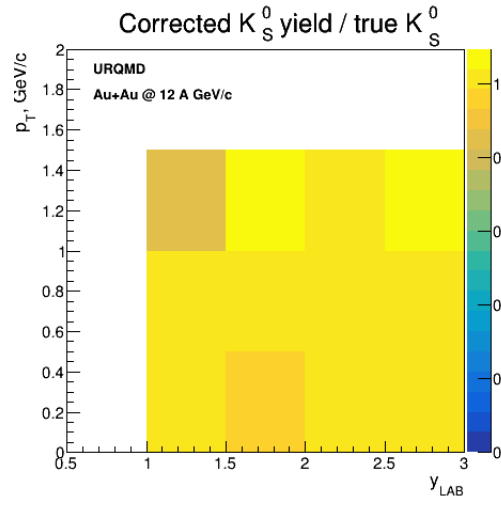
**Figure 5.2:** Three stage fitting procedure implementation: (a): Exclude signal region and fit background with pol2, (b) Use background fit parameters as initial values for next iteration, where signal (double Gaussian) fit function has fixed parameters, (c) Use fit parameters as initial values for unconstrained fit to the whole inv. mass range



(a)

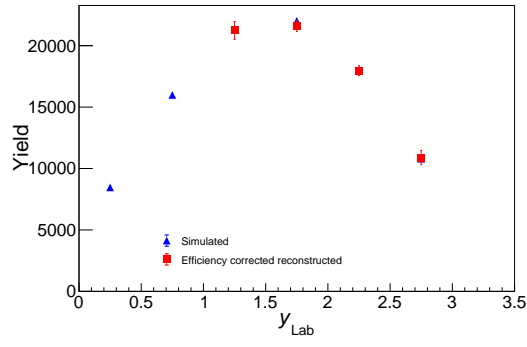


(b)

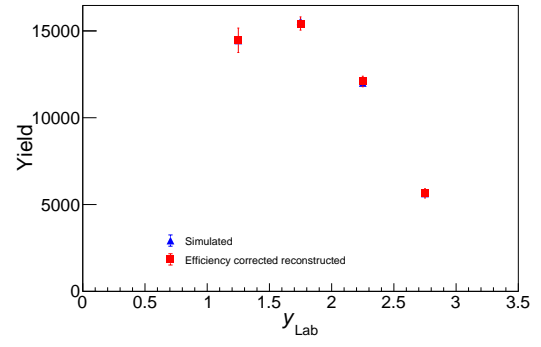


(c)

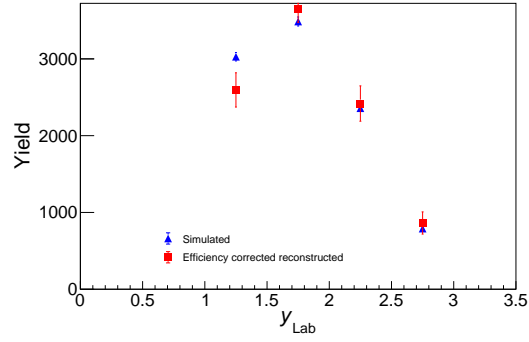
**Figure 5.3:** (a) reconstructed  $K_S^0$  acceptance and efficiency, (b) corrected  $K_S^0$  yield, (c) corrected  $K_S^0$  yield and true  $K_S^0$  yield



(a)



(b)



(c)

**Figure 5.4:** Efficiency and acceptance corrected  $K_S^0$  yields as a function of  $y_{LAB}$

## 6 Summary

Strangeness enhancement is one of the most important signatures of the deconfined state called quark gluon plasma. That's why (multi-)strange hadron reconstruction is crucial for the CBM experiment.

The CBM performance for the  $K_S^0$  mesons reconstruction via its decay to  $\pi^+$  and  $\pi^-$  is presented.

Machine learning XGBoost is powerful and robust tool for signal selection. The core of machine learning framework for physics analysis environment for (multi-)strange, hypernuclei, and other decays was written. Integration with already existing hipe4ML package developed for ALICE was implemented. It provides easy and user-friendly tool for physics analysis: input data quality assurance, variables selection, model training and testing, model's performance checking. The framework was used in this analysis for the signal selection.

$K_S^0$  yield extraction procedure was performed for two different heavy ion collision generators DSM-QGSM-SMM and URQMD. The yields of the both generators are withing the agreement.

The result of this analysis were presented on the 38<sup>th</sup> CBM collaboration meeting [32], the proceeding to the The 19<sup>th</sup> International Conference of Strangeness in Quark Matter (SQM 2021) was published [31] which includes the author of the thesis as the co-author. Author will present the current results in the FAIRness 2022 - the workshop for the yound scientists [33].

## 7 Acknowledgment

First, I would like to express my gratitude to the brave Ukrainian defenders. I was able to continue this work in the safety because of them. They protect our lives, our future, our tomorrow.

Thank my GSI supervisor Dr. Ilya Seliuzhenkov for the great experience and fruitful collaboration. To Dr. Andrea Dubla for the comments, help, and guidance.

To the great guys Dr. Viktor Klochkov, Shahid Khan and Oleksii Lubynets. They introduced me to a new area of physics, encouraged me, and helped me. They were the the best company to work with.

To my good friends Serhii Hamotskyi and Mariia Tkachenko for always being there for me and never letting me down.

To my brilliant friend Anton Rudakovskiy who encouraged me to stay grit and bold even in the darkest times.

To Dr. Tetyana Galatyuk for her help organizing my talk in the FAIRNESS workshop.

To colleagues from the STS group Dr. Maksym Teklishyn for his comments on my work and Dr. Anton Lymanets who kindly agreed to review my thesis.

To my supervisor from Nuclear Physics Department Oleg Bezshyyko and Larysa Golinka-Bezshyyko for their enthusiasm and encouragement.

## References

- [1] Jana N. Guenther. *Overview of the QCD phase diagram – Recent progress from the lattice*. 2020. DOI: [10.48550/ARXIV.2010.15503](https://doi.org/10.48550/ARXIV.2010.15503). URL: <https://arxiv.org/abs/2010.15503>.
- [2] [https://indico.cern.ch/event/985460/contributions/4264615/attachments/2211234/3742919/adf\\_cpod.pdf](https://indico.cern.ch/event/985460/contributions/4264615/attachments/2211234/3742919/adf_cpod.pdf).
- [3] Tapan K. Nayak. “Heavy Ions: Results from the Large Hadron Collider”. In: *Pramana* 79 (2012). Ed. by Rohin Godbole and Naba K. Mondal, pp. 719–735. DOI: [10.1007/s12043-012-0373-7](https://doi.org/10.1007/s12043-012-0373-7). arXiv: [1201.4264](https://arxiv.org/abs/1201.4264) [nucl-ex].
- [4] P. Koch, Berndt Muller, and Johann Rafelski. “Strangeness in Relativistic Heavy Ion Collisions”. In: *Phys. Rept.* 142 (1986), pp. 167–262. DOI: [10.1016/0370-1573\(86\)90096-7](https://doi.org/10.1016/0370-1573(86)90096-7).
- [5] Spyridon Margetis, Karel Safarik, and Orlando Villalobos Baillie. “Strangeness Production in Heavy-Ion Collisions”. In: *Annual Review of Nuclear and Particle Science* 50.1 (2000), pp. 299–342. DOI: [10.1146/annurev.nucl.50.1.299](https://doi.org/10.1146/annurev.nucl.50.1.299). eprint: <https://doi.org/10.1146/annurev.nucl.50.1.299>. URL: <https://doi.org/10.1146/annurev.nucl.50.1.299>.
- [6] P.A. Zyla et al. “Review of Particle Physics”. In: *PTEP* 2020.8 (2020). and 2021 update, p. 083C01. DOI: [10.1093/ptep/ptaa104](https://doi.org/10.1093/ptep/ptaa104).
- [7] Subhasis Chattopadhyay. “Physics of strongly interacting matter at high net-baryon density”. In: *Eur. Phys. J. ST* 230.3 (2021), pp. 689–696. DOI: [10.1140/epjs/s11734-021-00024-0](https://doi.org/10.1140/epjs/s11734-021-00024-0).
- [8] Bengt Friman et al., eds. *The CBM physics book: Compressed baryonic matter in laboratory experiments*. Vol. 814. 2011. DOI: [10.1007/978-3-642-13293-3](https://doi.org/10.1007/978-3-642-13293-3).
- [9] P Senger. “Probing dense QCD matter in the laboratory—The CBM experiment at FAIR”. In: *Physica Scripta* 95.7 (May 2020), p. 074003. DOI: [10.1088/1402-4896/ab8c14](https://doi.org/10.1088/1402-4896/ab8c14). URL: <https://doi.org/10.1088/1402-4896/ab8c14>.
- [10] T. Ablyazimov et al. “Challenges in QCD matter physics –The scientific programme of the Compressed Baryonic Matter experiment at FAIR”. In: *Eur. Phys. J. A* 53.3 (2017), p. 60. DOI: [10.1140/epja/i2017-12248-y](https://doi.org/10.1140/epja/i2017-12248-y). arXiv: [1607.01487](https://arxiv.org/abs/1607.01487) [nucl-ex].
- [11] Maksym Zyzak. “Online selection of short-lived particles on many-core computer architectures in the CBM experiment at FAIR”. PhD thesis. Frankfurt U., 2016.
- [12] Sergey Gorbunov. “On-line reconstruction algorithms for the CBM and ALICE experiments”. PhD thesis. 2013, pp. 104, 9.
- [13] [https://git.cbm.gsi.de/pwg-c2f/analysis/pf\\_simple](https://git.cbm.gsi.de/pwg-c2f/analysis/pf_simple).
- [14] Johann Heuser et al., eds. *[GSI Report 2013-4] Technical Design Report for the CBM Silicon Tracking System (STS)*. Darmstadt: GSI, 2013, 167 p. URL: <https://repository.gsi.de/record/54798>.

- [15] <https://github.com/julnow/JupyterNotebooks/blob/kaon/img/histograms/histogramsBeforeAndAfter.pdf>.
- [16] M. Baznat et al. “Monte-Carlo Generator of Heavy Ion Collisions DCM-SMM”. In: *Physics of Particles and Nuclei Letters* 17.3 (May 2020), pp. 303–324. ISSN: 1531-8567. DOI: [10.1134/s1547477120030024](https://doi.org/10.1134/s1547477120030024). URL: <http://dx.doi.org/10.1134/S1547477120030024>.
- [17] W. Trautmann. “Multifragmentation in relativistic heavy ion reactions”. In: *International Summer School on Correlations and Clustering Phenomena in Subatomic Physics*. Aug. 1996, pp. 115–135. arXiv: [nuc1-ex/9611002](https://arxiv.org/abs/nuc1-ex/9611002).
- [18] S Bass. “Microscopic models for ultrarelativistic heavy ion collisions”. In: *Progress in Particle and Nuclear Physics* 41 (1998), pp. 255–369. ISSN: 0146-6410. DOI: [10.1016/s0146-6410\(98\)00058-1](https://doi.org/10.1016/s0146-6410(98)00058-1). URL: [http://dx.doi.org/10.1016/S0146-6410\(98\)00058-1](http://dx.doi.org/10.1016/S0146-6410(98)00058-1).
- [19] M Bleicher et al. “Relativistic hadron-hadron collisions in the ultra-relativistic quantum molecular dynamics model”. In: *Journal of Physics G: Nuclear and Particle Physics* 25.9 (Sept. 1999), pp. 1859–1896. ISSN: 1361-6471. DOI: [10.1088/0954-3899/25/9/308](https://doi.org/10.1088/0954-3899/25/9/308). URL: <http://dx.doi.org/10.1088/0954-3899/25/9/308>.
- [20] M. Baznat et al. “Monte-Carlo Generator of Heavy Ion Collisions DCM-SMM”. In: *Physics of Particles and Nuclei Letters* 17.3 (May 2020), pp. 303–324. DOI: [10.1134/s1547477120030024](https://doi.org/10.1134/s1547477120030024). URL: <https://doi.org/10.1134/s1547477120030024>.
- [21] S. Agostinelli et al. “Geant4—a simulation toolkit”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303. ISSN: 0168-9002. DOI: [https://doi.org/10.1016/S0168-9002\(03\)01368-8](https://doi.org/10.1016/S0168-9002(03)01368-8). URL: <https://www.sciencedirect.com/science/article/pii/S0168900203013688>.
- [22] J. Allison et al. “Geant4 developments and applications”. In: *IEEE Transactions on Nuclear Science* 53.1 (2006), pp. 270–278. DOI: [10.1109/TNS.2006.869826](https://doi.org/10.1109/TNS.2006.869826).
- [23] <https://github.com/HeavyIonAnalysis/AnalysisTree>.
- [24] <https://github.com/hipe4ml/hipe4ml>.
- [25] <https://github.com/CBM-ML/CandidatesClassifier>.
- [26] <https://toml.io/en/>.
- [27] [https://scikit-learn.org/stable/auto\\_examples/model\\_selection/plot\\_roc.html](https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html).
- [28] C. Adam-Bourdarios et al. “The Higgs Machine Learning Challenge”. In: *J. Phys. Conf. Ser.* 664.7 (2015), p. 072015. DOI: [10.1088/1742-6596/664/7/072015](https://doi.org/10.1088/1742-6596/664/7/072015).
- [29] <https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/>.



- [30] Peter I. Frazier. *A Tutorial on Bayesian Optimization*. 2018. DOI: [10.48550/ARXIV.1807.02811](https://doi.org/10.48550/ARXIV.1807.02811). URL: <https://arxiv.org/abs/1807.02811>.
- [31] Shahid Khan et al. “Machine Learning Application for  $\Lambda$  Hyperon Reconstruction in CBM at FAIR”. In: *EPJ Web of Conferences* 259 (2022). Ed. by G. David et al., p. 13008. DOI: [10.1051/epjconf/202225913008](https://doi.org/10.1051/epjconf/202225913008).
- [32] <https://indico.gsi.de/event/13089/contributions/56279/attachments/37179/49724/CBM%20collab%20meet%20olha%20lavoryk%2028%20september.pdf>.
- [33] <https://indico.gsi.de/event/13960/>.