Using Machine Learning for Particle Identification in MPD

Grigorii Tolkachev[1], Alexey Aparin[2], Artem Korobitsin[2]

[1]National Research Nuclear University MEPhI, Moscow, Russia

[2]Joint Institute for Nuclear Research (JINR), Dubna, Russia

This work was supported by RSCF under grant N 22-72-10028

The 6th international conference on particle physics and astrophysics (ICCPA2022)

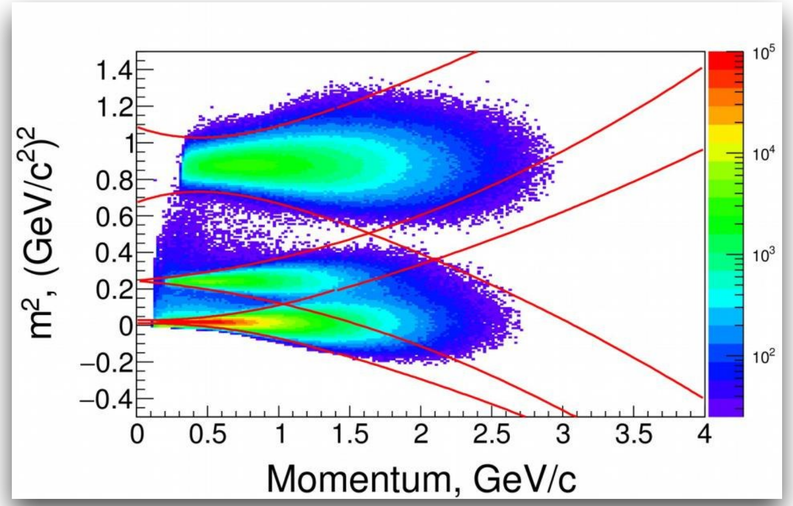
Moscow, December 2, 2022

JOINT INSTITUTE FOR NUCLEAR RESEARCH

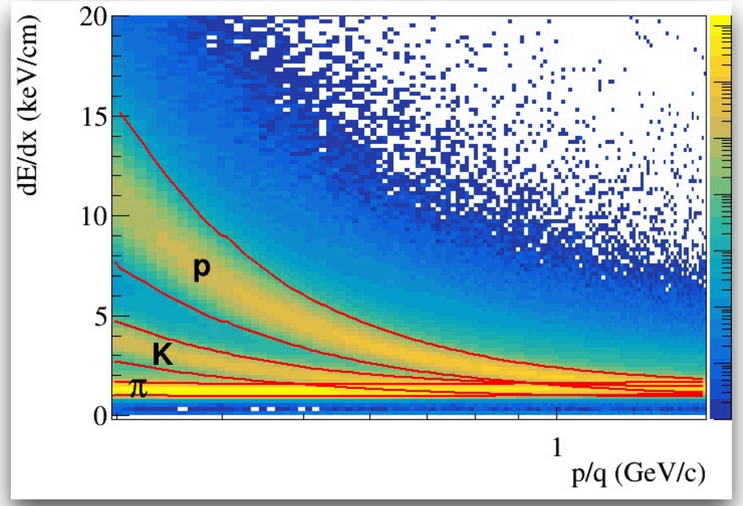
Introduction

- Particle identification is an important aspect of most particle physics experiments.
  - Identify long-lived particles that leave a trace in the detector: electrons, muons, photons, charged pions, charged kaons, etc.
  - Short-lived particles are identified by their decays into long-lived
- Various standard approaches are used for particle identification. One of them is an approach based on estimating the deviation of particles from the assumed average value in the distributions of lost energy and the mass square in different momentum range (n-Sigma) [1].
- In addition to standard approaches, machine learning methods are used for particle identification.

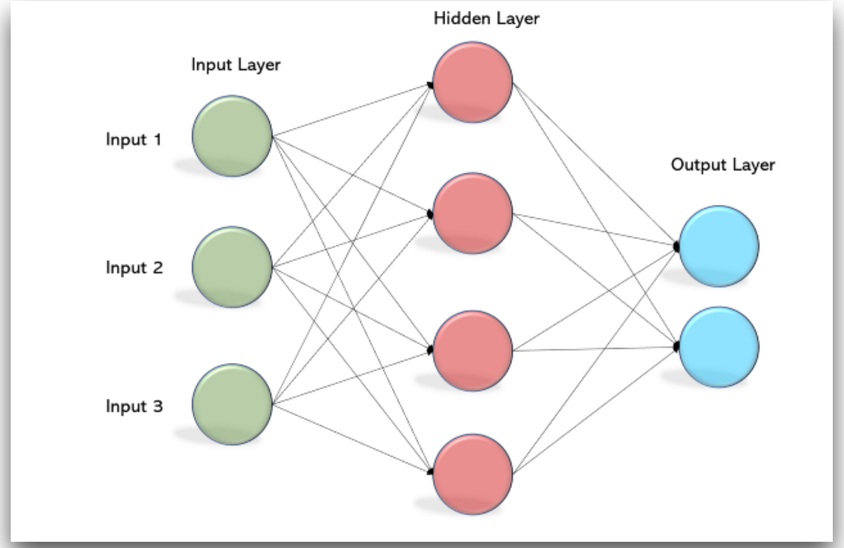
Distribution of m^2 as a function of p



Distribution of dE/dx as a function of p



Multi layer Perceptron (MLP) Models



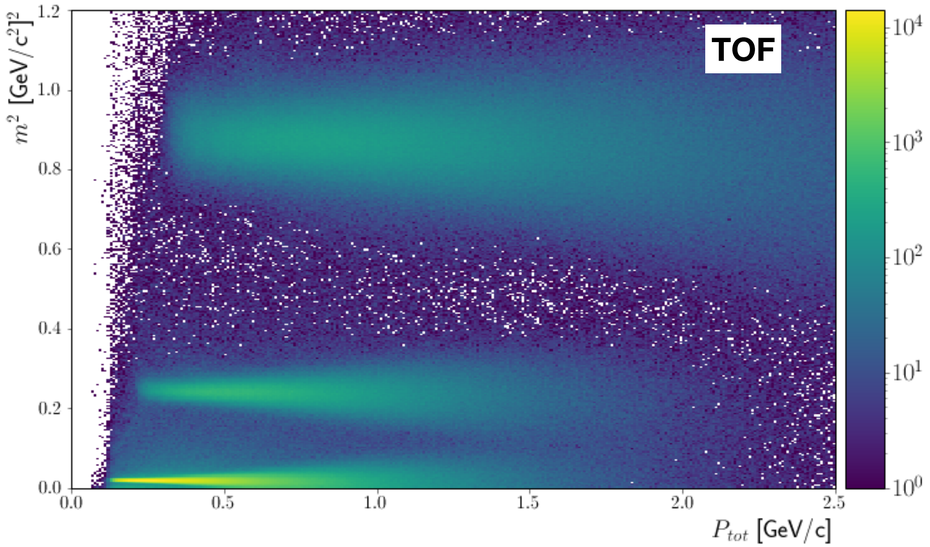
Goal: Selection of the optimal MLP model to improve the efficiency of identification of charged particles. Comparison of the efficiency of the standard and MLP approach.

[1] https://git.jinr.ru/nica/mpdroot/-/tree/dev/core/mpdPid

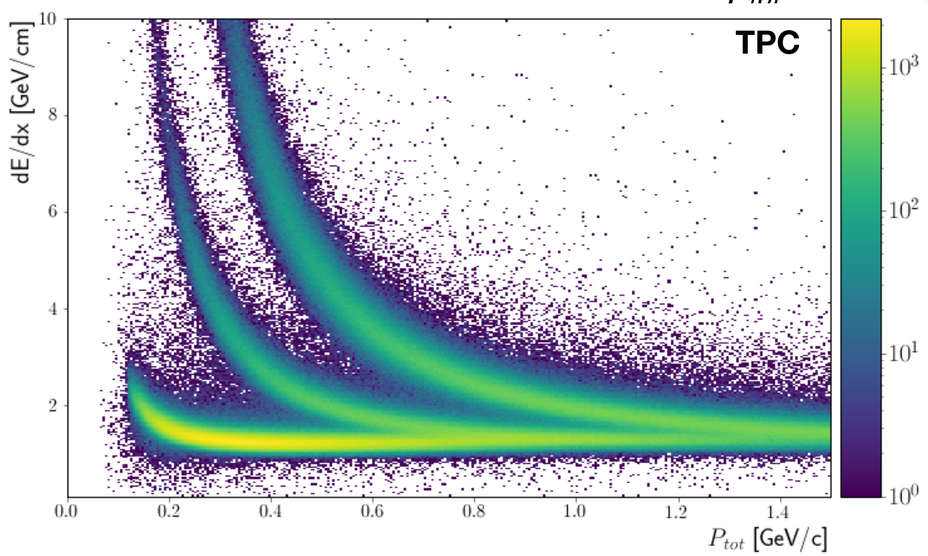
Feature selection

- Variables used: dE/dx, m^2, p_{tot}, p_T, β, ϕ, θ, η, q, $nHits$, dca, V_x, V_y, V_z.
- 6 species of particles : π^-, π^+, K^-, K^+, p, $\bar{p}$.
- 200,000 events for each class were used to train and test the models

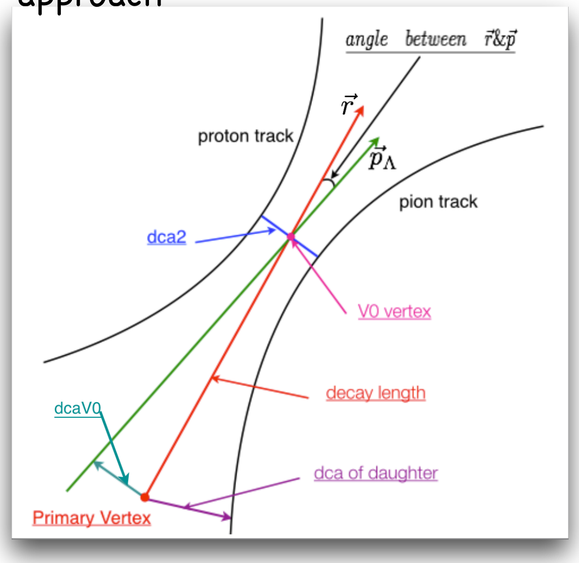
Distribution of m^2 as a function of p_{tot}



Distribution of dE/dx as a function of p_{tot}



dca - distance of the closest approach

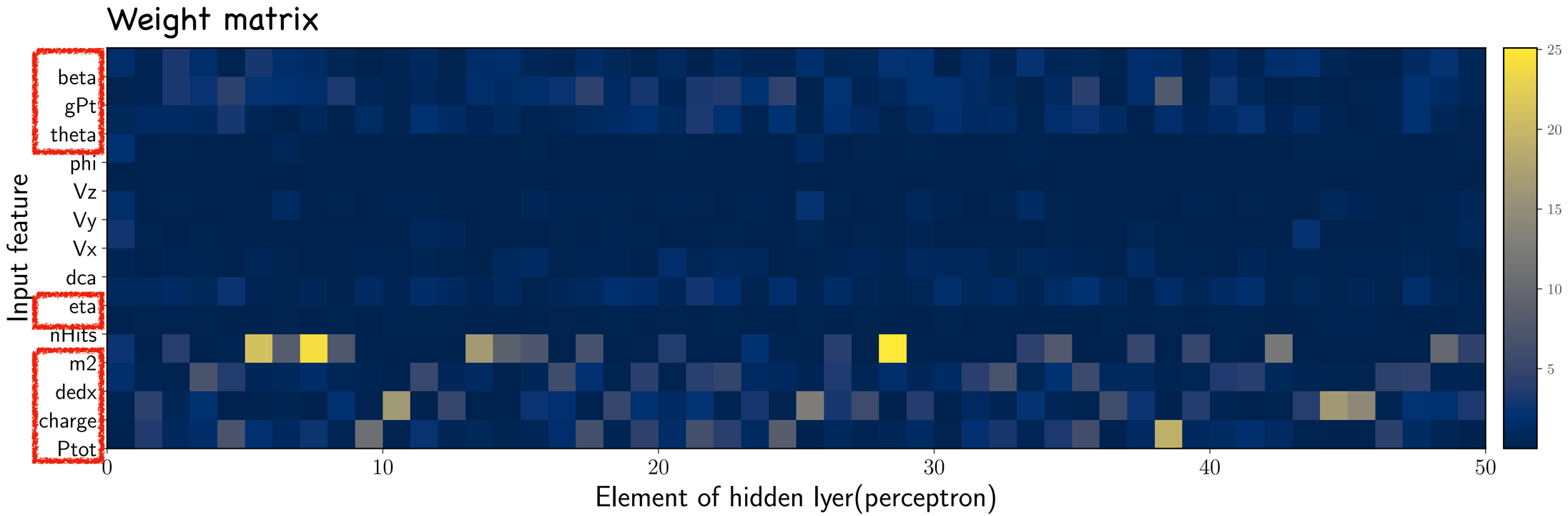


$$\beta = p/E = p/\sqrt{p^2 + m^2}$$

$$m^2 = p^2(1/\beta^2 - 1)$$

$$\frac{dE}{dx} = \frac{4\pi}{m_e c^2}\frac{nz^2}{\beta^2}\left(\frac{e^2}{4\pi\varepsilon_0}\right)^2\left[ln\left(\frac{2m_e c^2\beta^2}{I(1-\beta^2)}\right) - \beta^2\right]$$

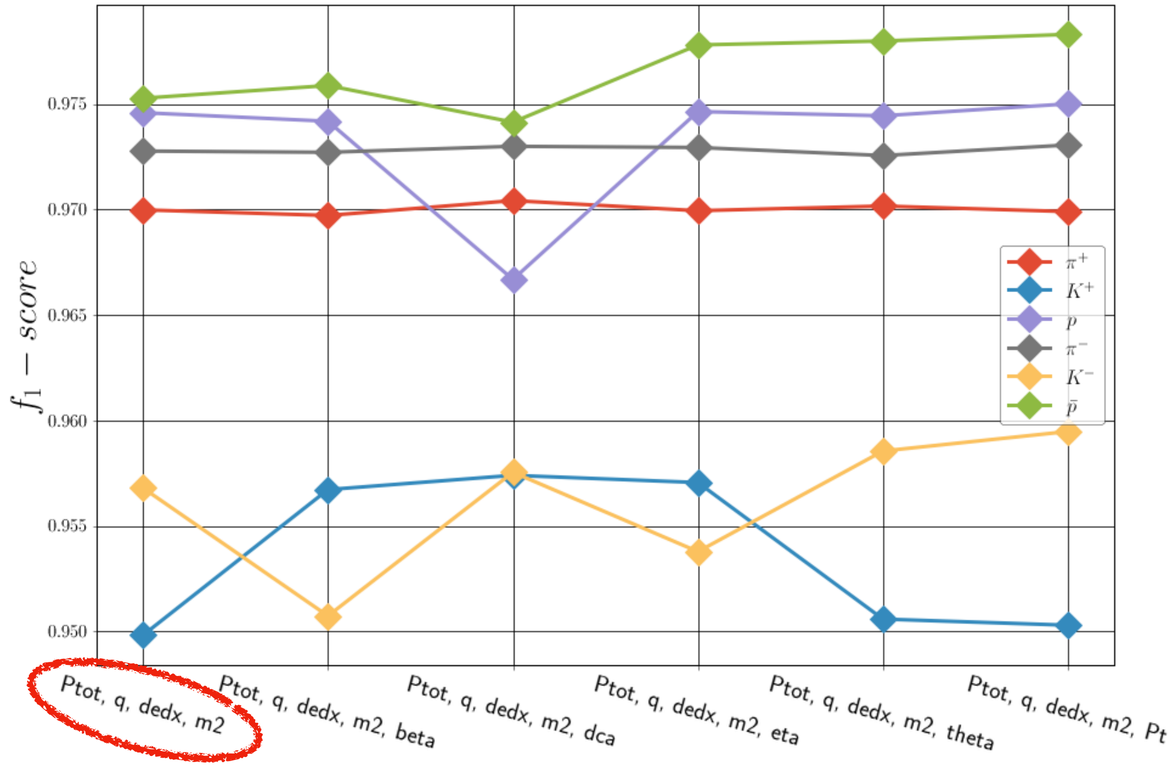
Feature selection



- dedx, m2, Ptot, charge - for almost every element of the hidden layer have a weight other than zero
- beta, eta, theta, gPt - for some elements of the hidden layer have a weight other than zero
- nHits, dca, phi , Vx, Vy, Vz - have zero weight for all elements of the hidden layer

Feature selection

Dependence of f1-score on a set of variables



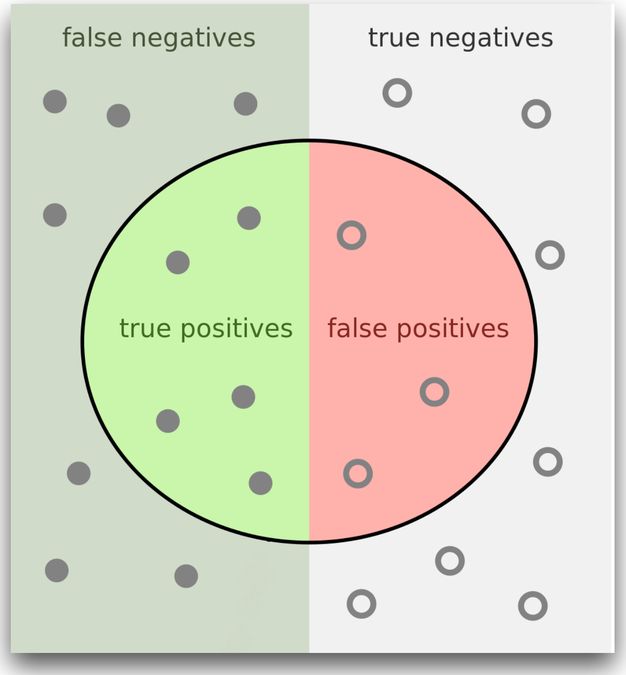
- The reason why $K^{\pm}$ have the lowest f1-score is that, for example, on the distribution m^2 they are between p and π mixed with all of them

- Some additional variables improve the f1-score for one type of particles and worsen for another type. Another additional variables do not make a significant contribution to the f1-score
- Set of parameters was used in the research: Ptot, charge, dedx, m2

$$f_1 = 2 * \frac{recall * precision}{recall + precision}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

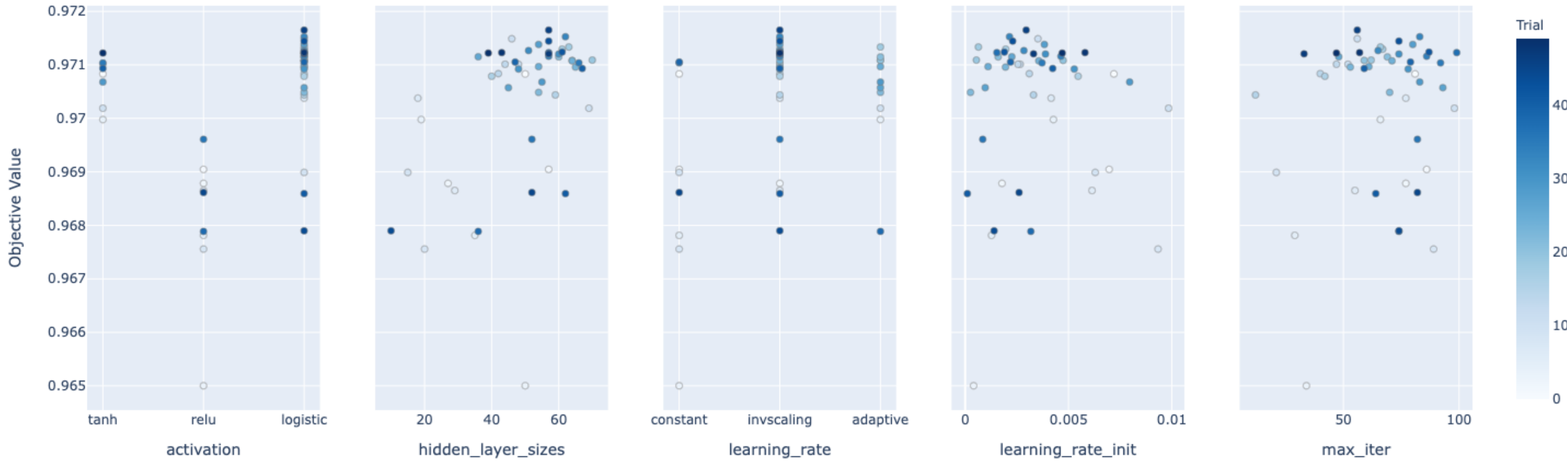


Hyperparameters selection

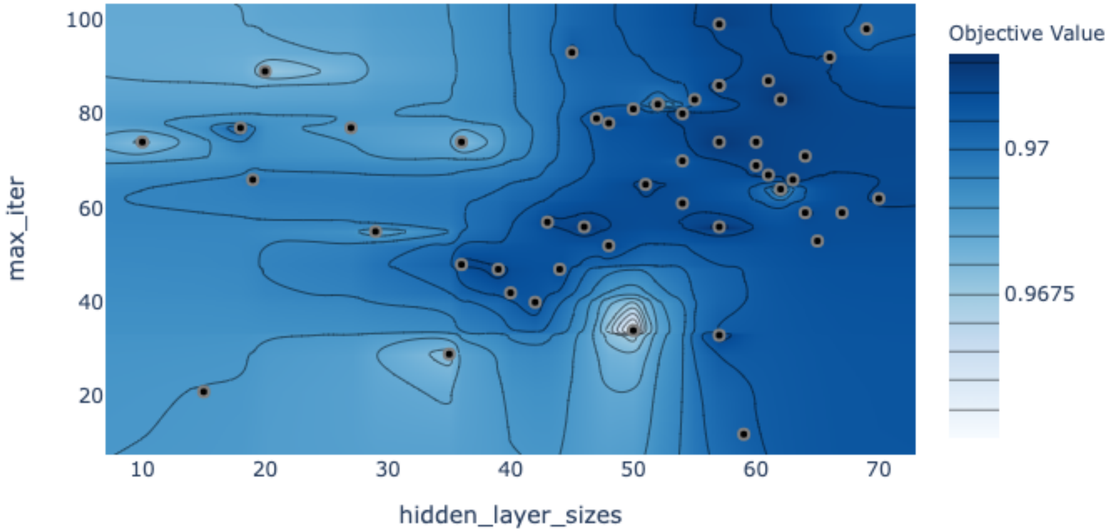
Set of hyperparameters that were used in Bayesian optimisation

hidden_layer_sizes	10 - 70
max_iter	10 - 100
learning_rate_init	0.0001 - 0.01
activation	logistic, tanh, relu
learning_rate	constant, invscaling, adaptive

Dependence of the weighted f1-score on the value of each of the hyperparameters



Map of hyperparameters hidden_layer_sizes and max_iter



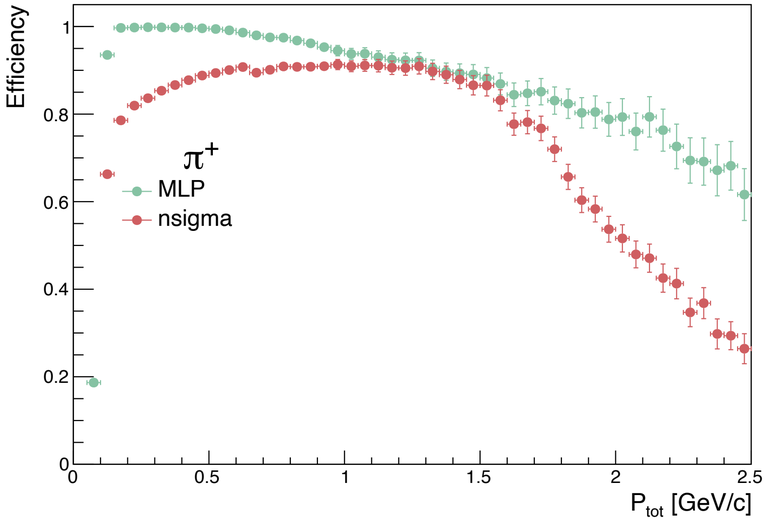
- More classifiers have f1-score > 0.97
- To simplify the model, a model with hidden_layer_sizes = 36 and max_iter = 48 was chosen. learning_rate_init = 0.006, activation = logistic, learning_rate = constant

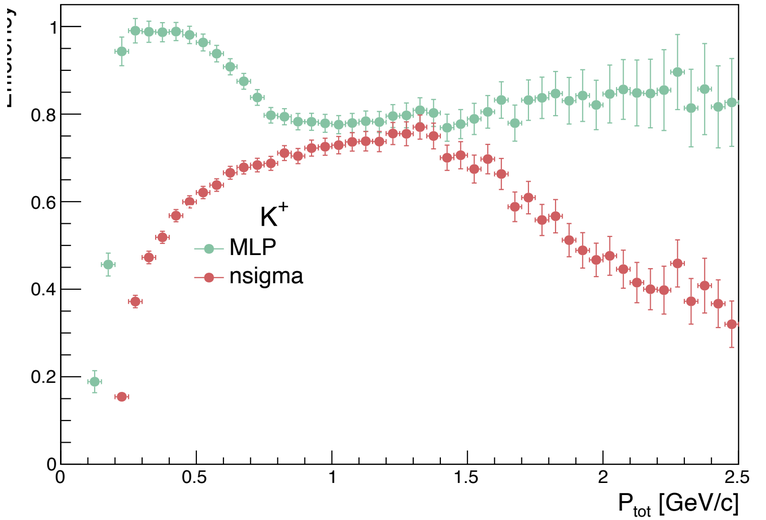
Result

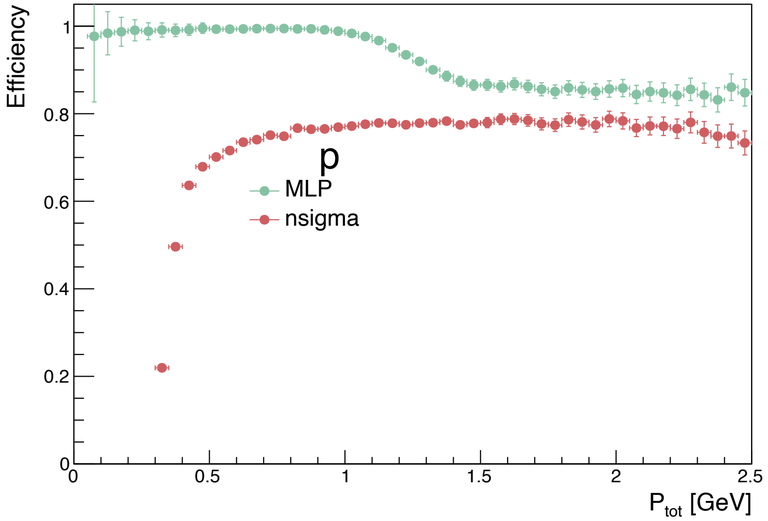
- To evaluate the approaches, a data set with a different number of particles of different species was used
- In order to estimate the quality of identification, the efficiency was used:

$$Efficiency = \frac{dN^{i}_{\text{true}}/dp}{dN^{i}_{\text{all gen.}}/dp}$$

- The efficiency of MLP model identification is compared with the efficiency of identification of the standard n-Sigma approach[1].
- For each particle species MLP approach has higher efficiency than n-Sigma approach for full range of momentum.







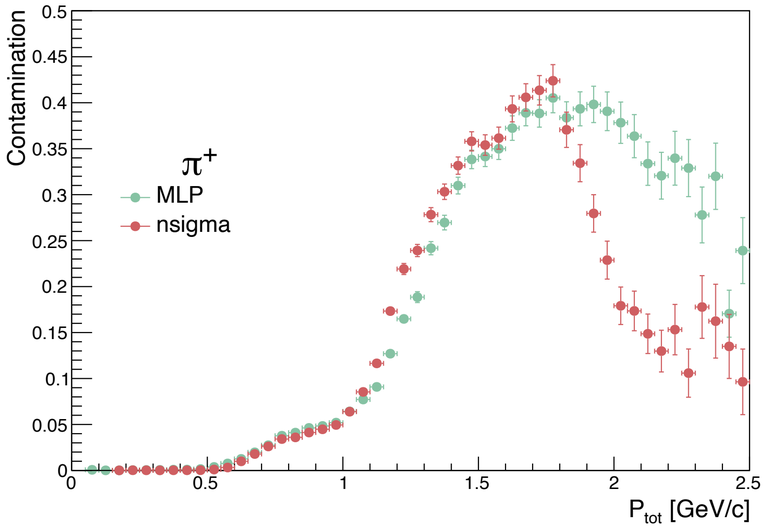
[1] https://git.jinr.ru/nica/mpdroot/-/tree/dev/core/mpdPid

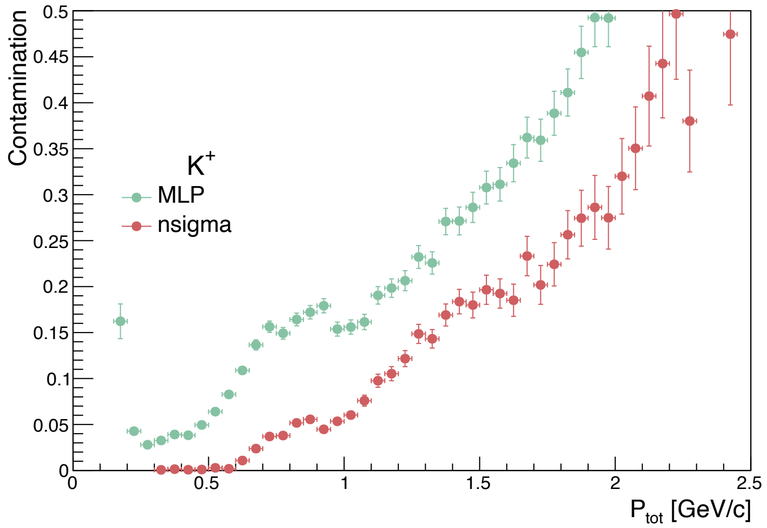
Result

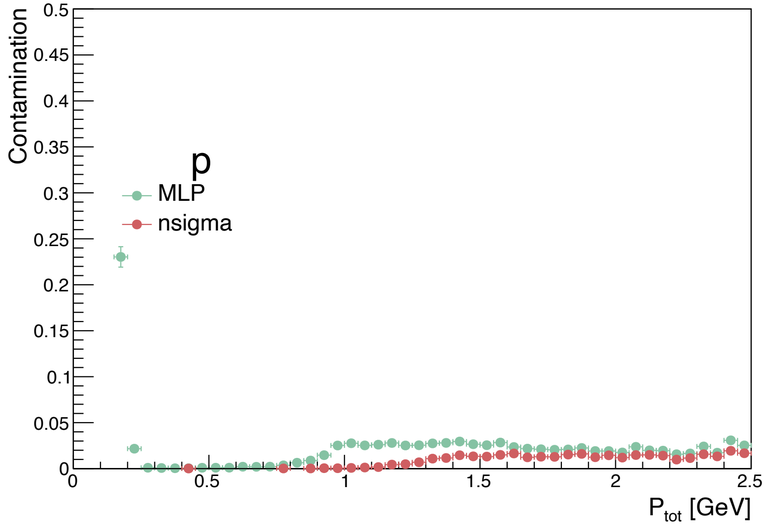
- Contamination:

$$Contamination = \frac{dN^{i}_{\text{false}}/dp}{dN^{i}_{\text{all id.}}/dp}$$

- The contamination of MLP model identification is compared with the efficiency of identification of the standard n-Sigma approach.
- For π^{+} and p particle species MLP approach has the same contamination, but for K^{+} particle species contamination is higher.



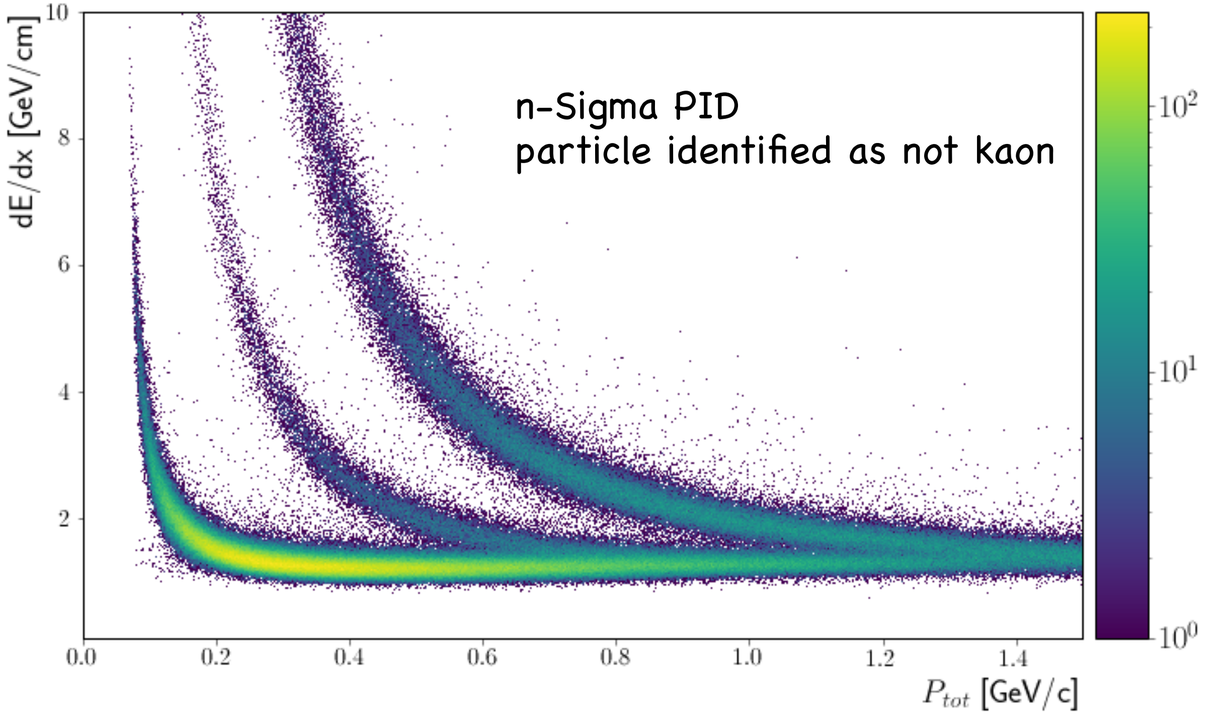


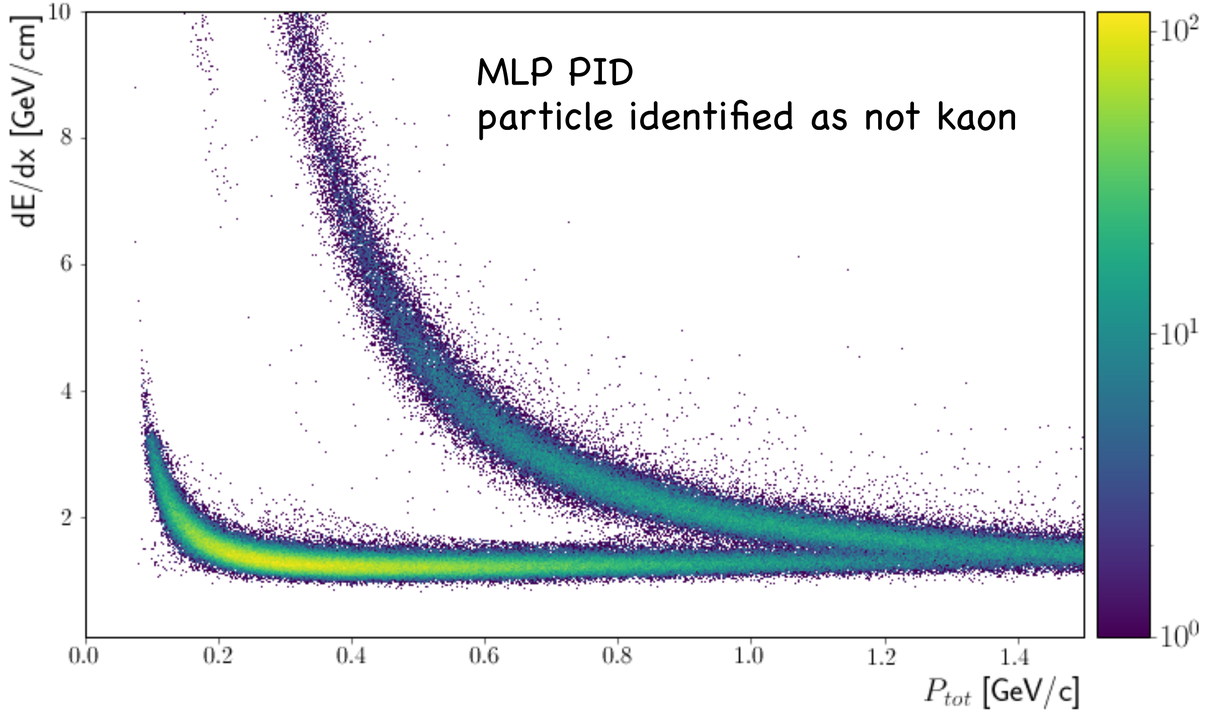


Result

Why does MLP approach have better efficiency for each particle species in all range of momentum, but it has the same or higher contamination than n-Sigma approach?

- n-Sigma approach identifies particle as particle of a i-species if $N_\sigma \leq \sqrt{N^2_{\sigma^i_{TOF}} + N^2_{\sigma^i_{TPC}}}$ (1) values are in a certain range around mean value for i-species of particle . Where $N_{\sigma^i_{TPC}} = \frac{dE/dx - \langle dE/dx \rangle^i}{\sigma^i_{TPC}}$, $N_{\sigma^i_{TOF}} = \frac{m^2 - \langle m^2 \rangle^i}{\sigma^i_{m^2}}$,
- If a particle can be compatible with more than one species, n-Sigma approach does not identify this particle.





Conclusion

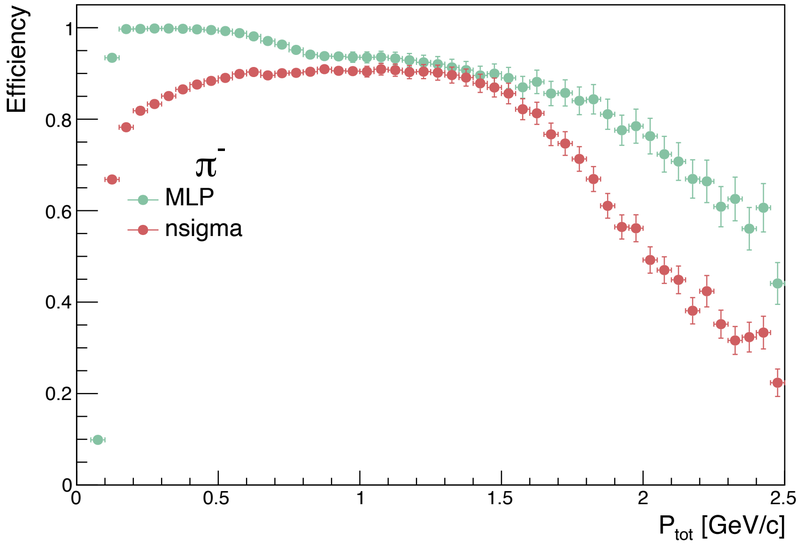
- For MLP multi-classifier the set of features that make the biggest contribution to f1-score has been chosen.
- Using Bayesian optimisation the hyperparametrs have been chosen which do not complicate MLP model and allow to get high f1-score.
- The n-Sigma approach was studied and compared with MLP approach for particle identification. It has been shown that for each particle species, the MLP approach has a higher efficiency than the n-Sigma approach for the full momentum range.
- The improvement is shown only for the certain version of MC simulation data. In the future, it is planned to conduct research for a wide set of MC data.

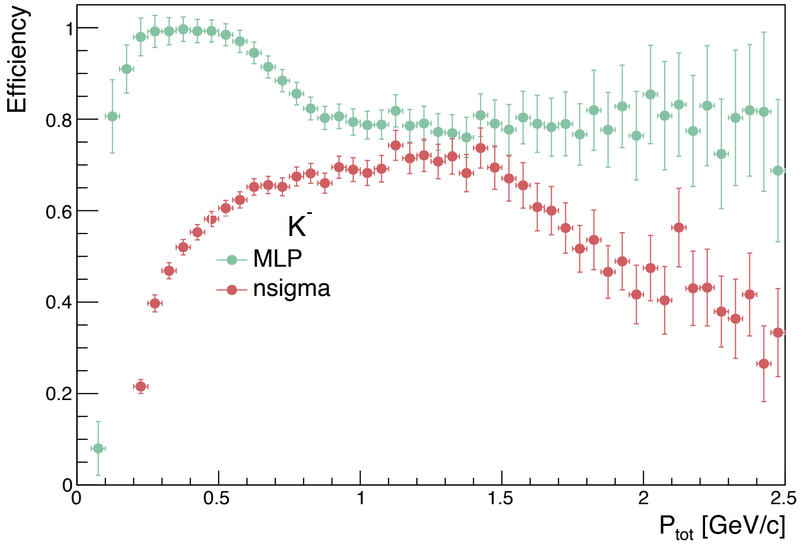
This work was supported by RSCF under grant N 22-72-10028

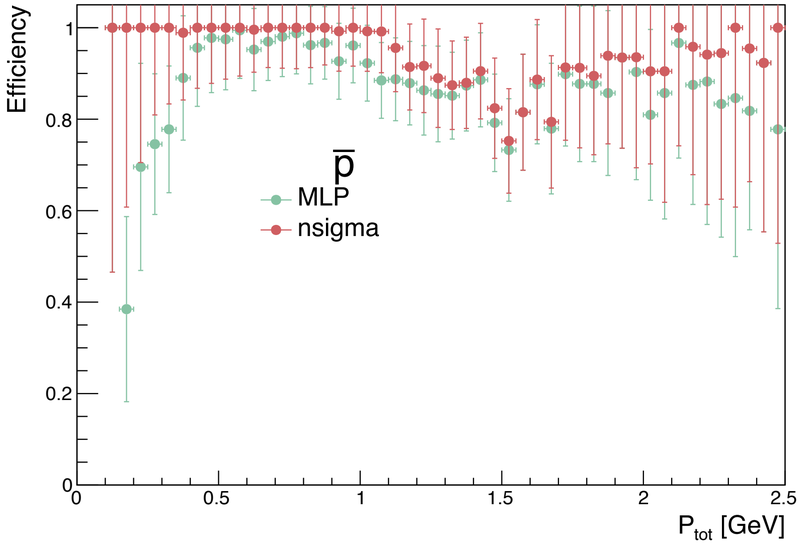
Thank You

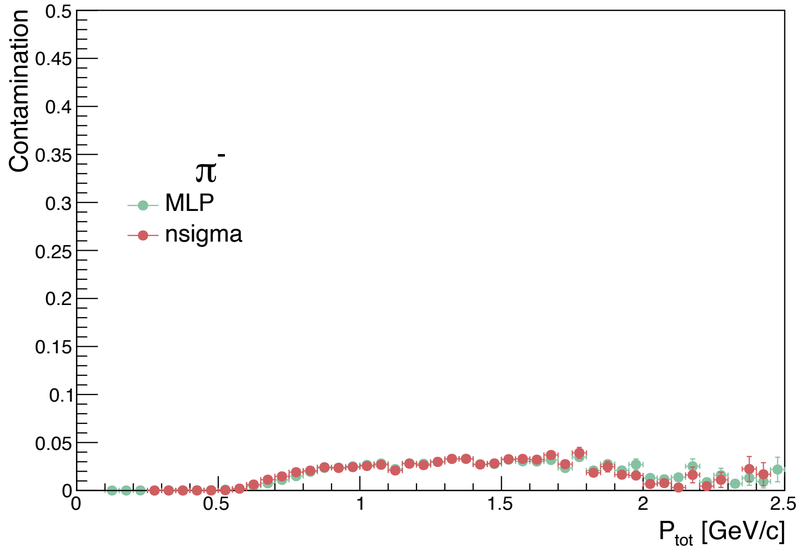
Backup Slides

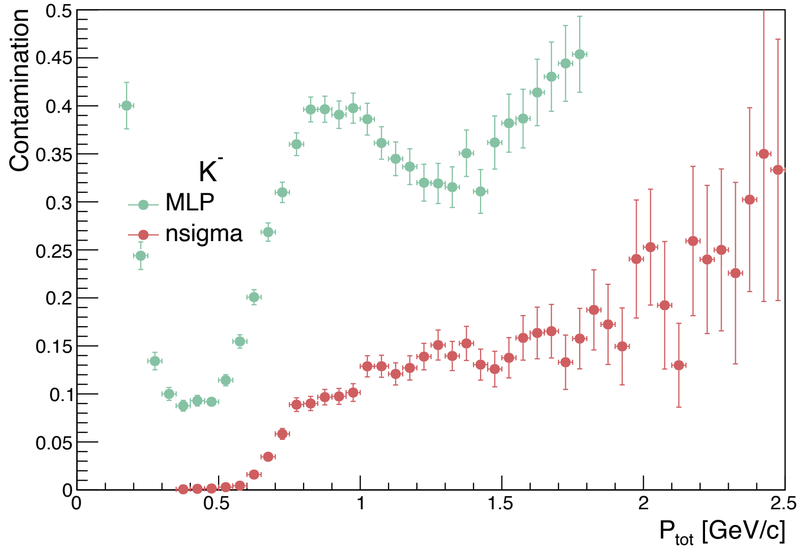
Result

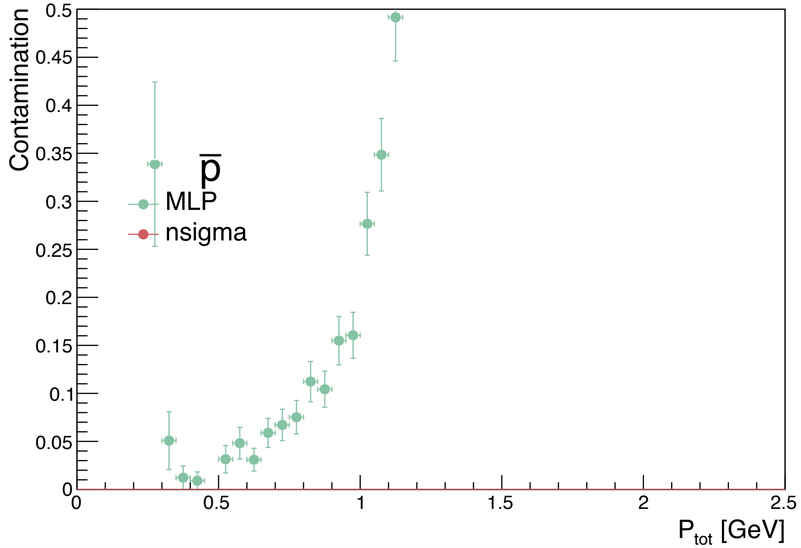

Efficiency
π^-
MLP
nsigma
P_{tot} [GeV/c]


Efficiency
K^-
MLP
nsigma
P_{tot} [GeV/c]


Efficiency
$\bar{p}$
MLP
nsigma
P_{tot} [GeV]


Contamination
π^-
MLP
nsigma
P_{tot} [GeV/c]


Contamination
K^-
MLP
nsigma
P_{tot} [GeV/c]

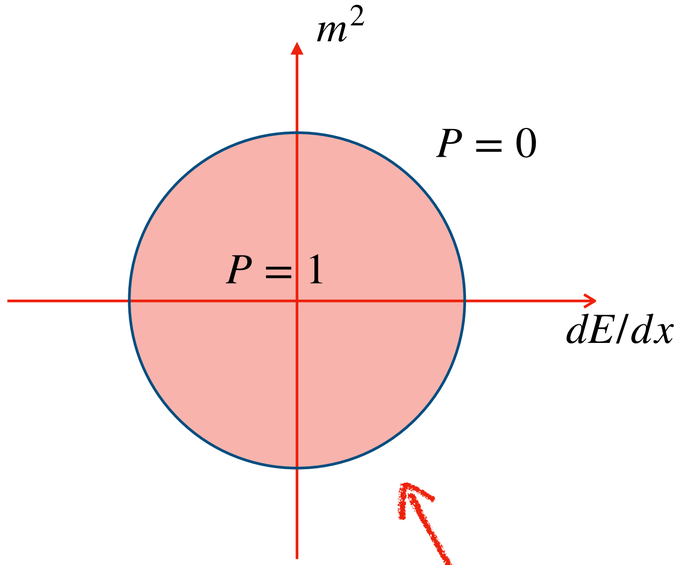

Contamination
$\bar{p}$
MLP
nsigma
P_{tot} [GeV]

Result

Why does MLP approach have better efficiency for each particle species in all range of momentum, but it has the same or higher contamination than n-Sigma approach?

- n-Sigma approach identifies particle as particle of a i-species if $N_\sigma \leq \sqrt{N^2_{\sigma^i_{TOF}} + N^2_{\sigma^i_{TPC}}}$ (1) values are in a certain range around mean value for i-species of particle . Where $N_{\sigma_{TPC}}$ and $N_{\sigma_{TOF}}$:

$$N_{\sigma^i_{TPC}} = \frac{dE/dx - \langle dE/dx \rangle^i}{\sigma^i_{TPC}}, \qquad N_{\sigma^i_{TOF}} = \frac{m^2 - \langle m^2 \rangle^i}{\sigma^i_{m^2}},$$



- If the condition (1) is met for N^i_{TPC} and N^i_{TOF}, the particle is identified as particle of i-species.
- If a particle can be compatible with more than one species, the approach does not identify this particle.
- n-Sigma has worse efficiency while having less contamination. The reason for this is that in the used dataset, the n-Sigma could not identify many particles

$$N_\sigma \leq \sqrt{N^2_{\sigma^i_{TOF}} + N^2_{\sigma^i_{TPC}}}$$

	π^+	K^+	p	π^-	K^-	$\bar{p}$
% all identified particles of given species	84.7	72.3	76.5	84.4	69.5	100