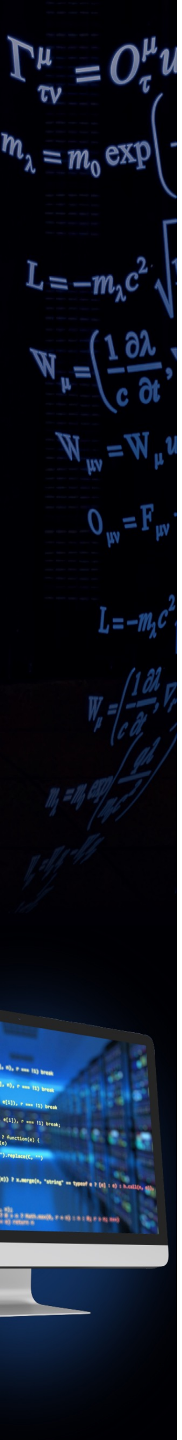


Gradient Boosted Decision Tree for Particle Identification Problem

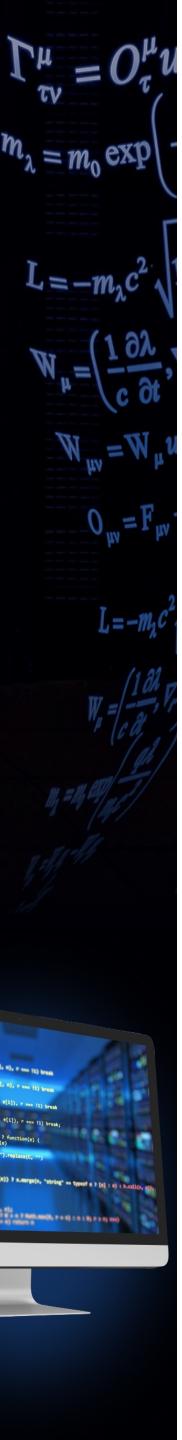
Alexander Ayriyan

SPD Physics and MC meeting N37

24 January 2023

Identification Problem of Charged Particles

- In Machine Learning terms PID can be considered as **classification** task (**Supervised learning**).
- Let
- **X** - is the input space (particle characteristics such as: **dE/dx**, **m2**, **q**, **P**, etc)
- **Y** - is the output space (particle species such as: **π**, **k**, **p**, etc.)
- Unknown mapping exists
  - $\mathbf{m : X \rightarrow Y}$,
- for values which known only on objects from the finite training set
  - $\mathbf{X^n = (x_1, y_1), \ldots, (x_n, y_n)}$,
- Goal is to find an algorithm a that classifies an arbitrary new object $x \in X$
  - $\mathbf{a : X \rightarrow Y}$.

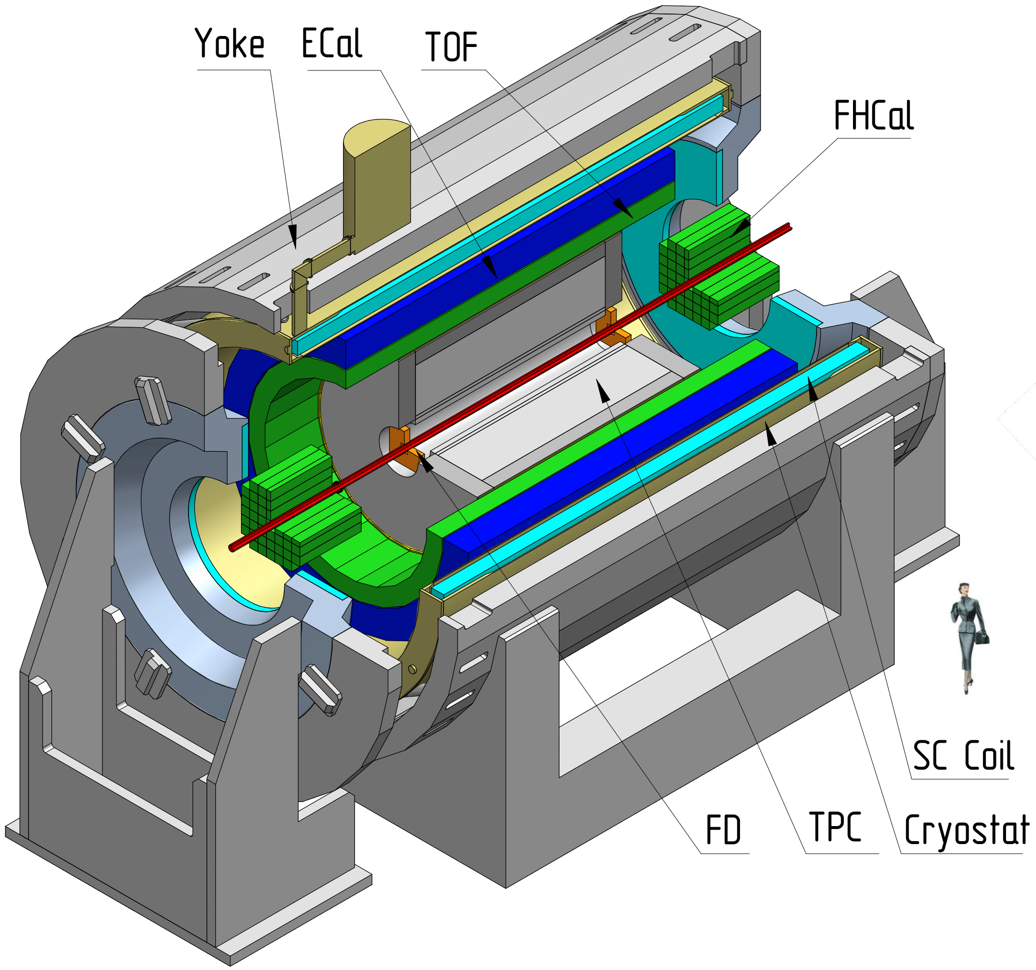


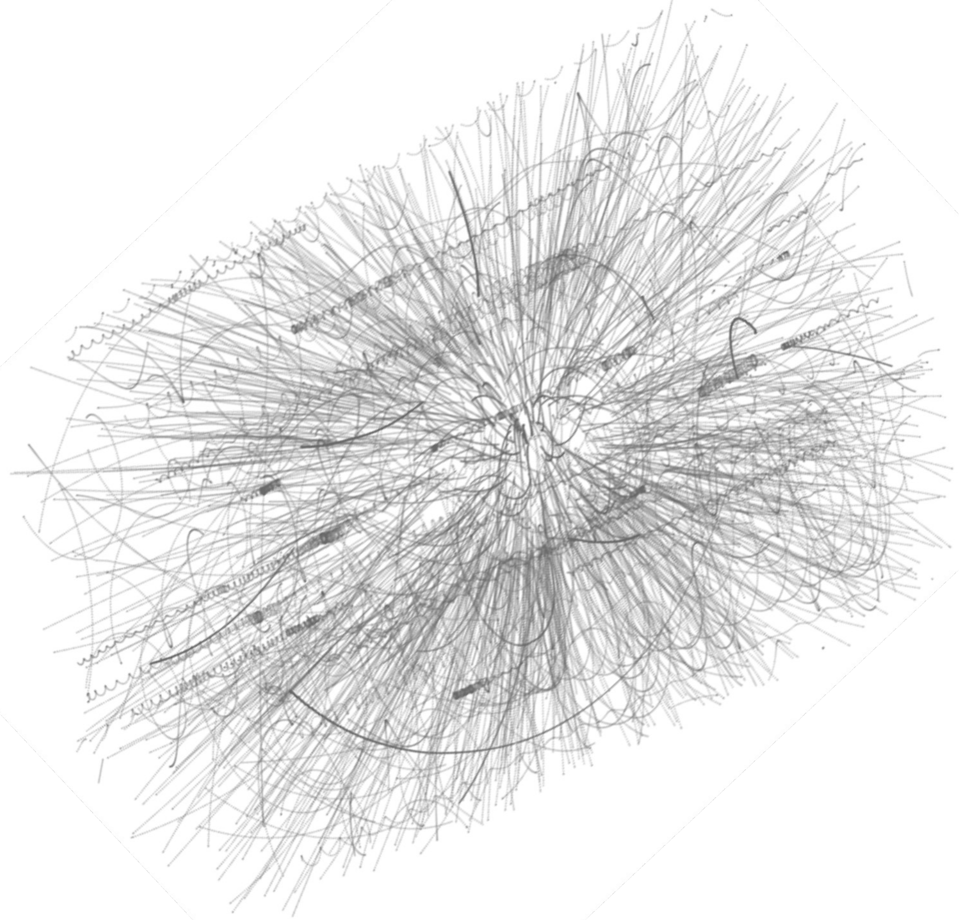
NICA

MPD

V. Papoyan, A. Ayriyan, A. Aparin, H. Grigorian, A. Korobitsin, A. Mudrokh

MPD APPARATUS AND PID

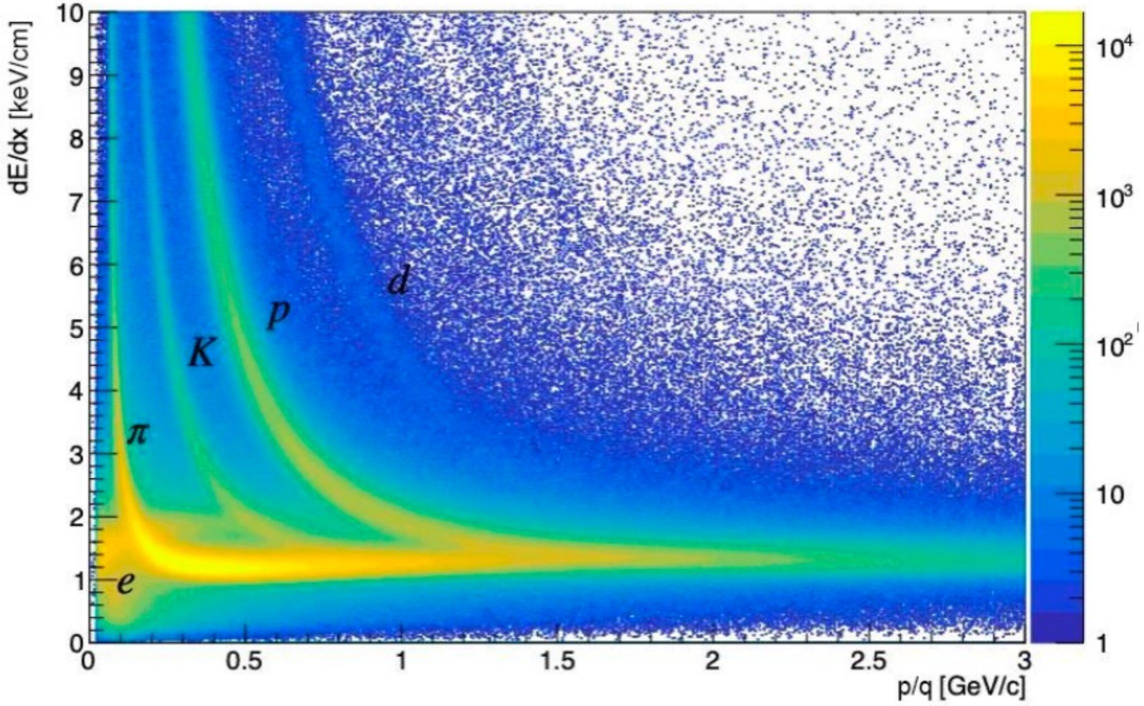


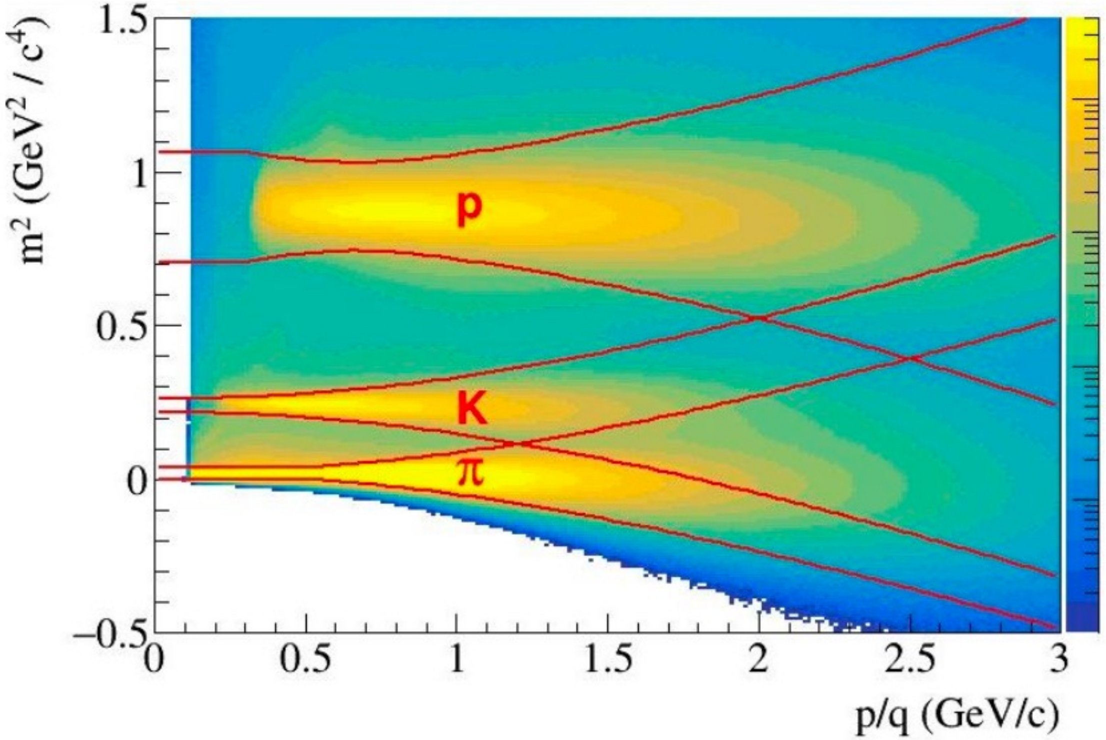


MPD particle identification (PID) based on
Time-Projection Chamber (TPC) and **Time-of-Flight** (TOF).

Particle Identification in MPD experiment

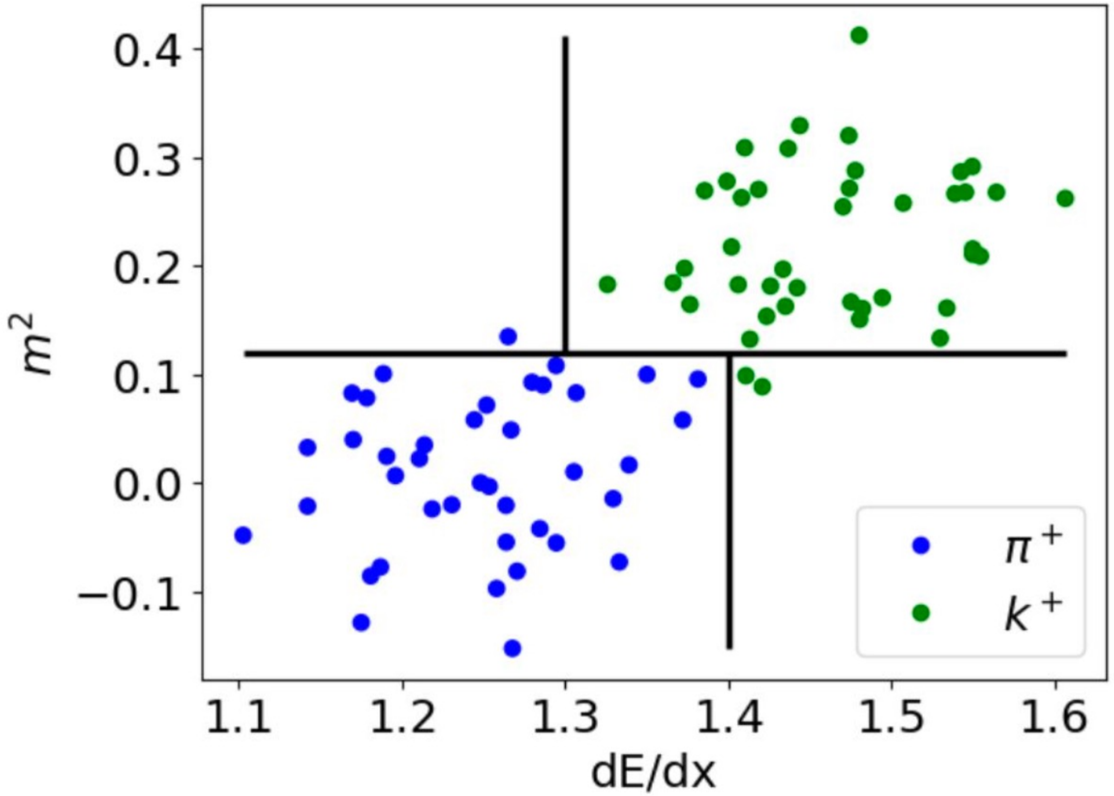
Particle identification can be achieved by using information about **momentum**, **charge**, **energy loss** (TPC) and **mass squared** (TPC + TOF).

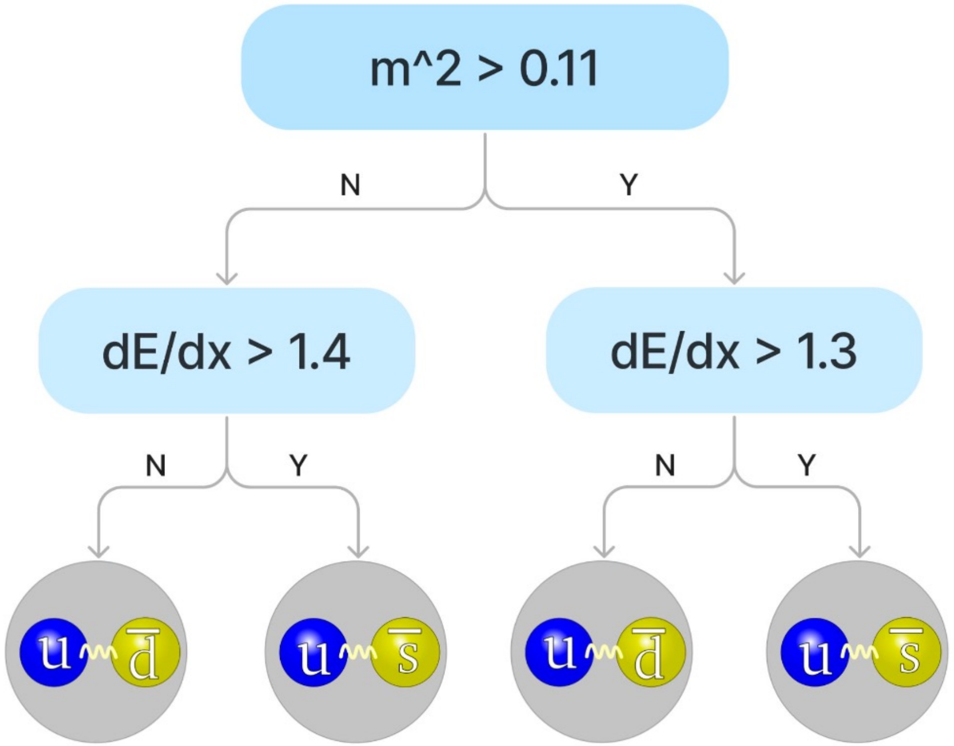




Decision Trees for PID

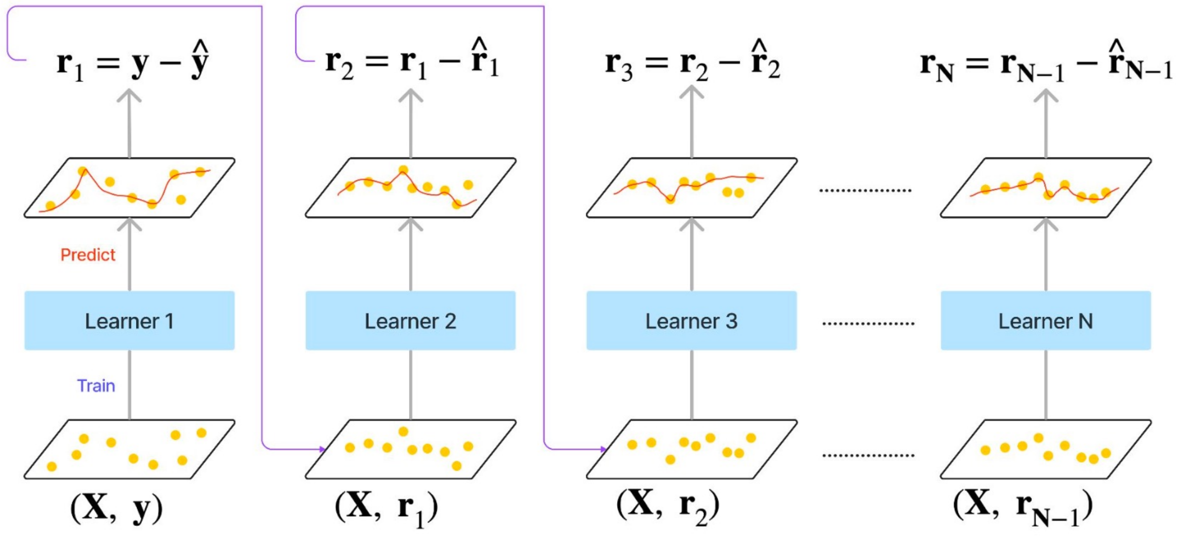
Gradient Boosted Decision Tree (GBDT) uses decision trees as weak learner. They can be considered as automated multilevel **cut-based** analysis.

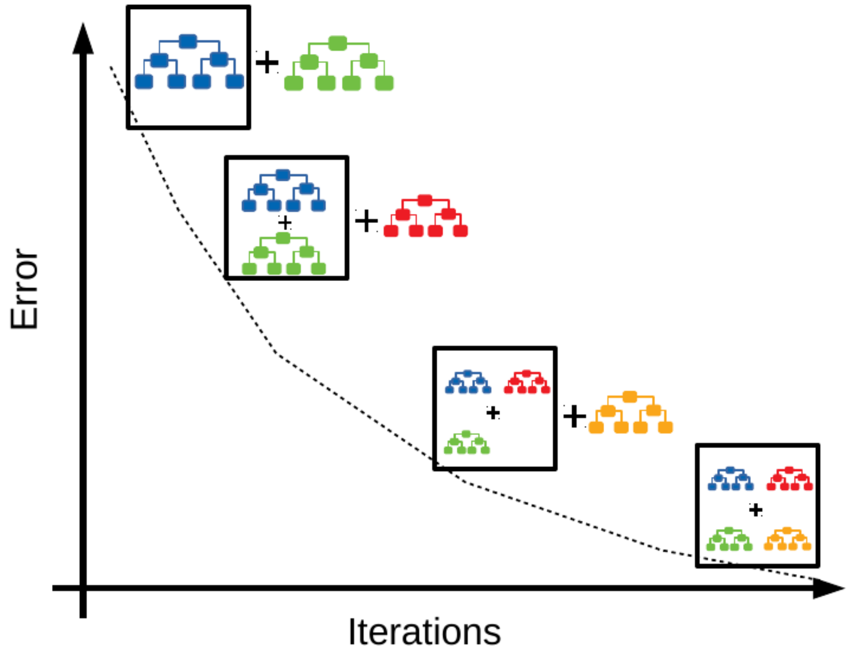




GRADIENT BOOSTING

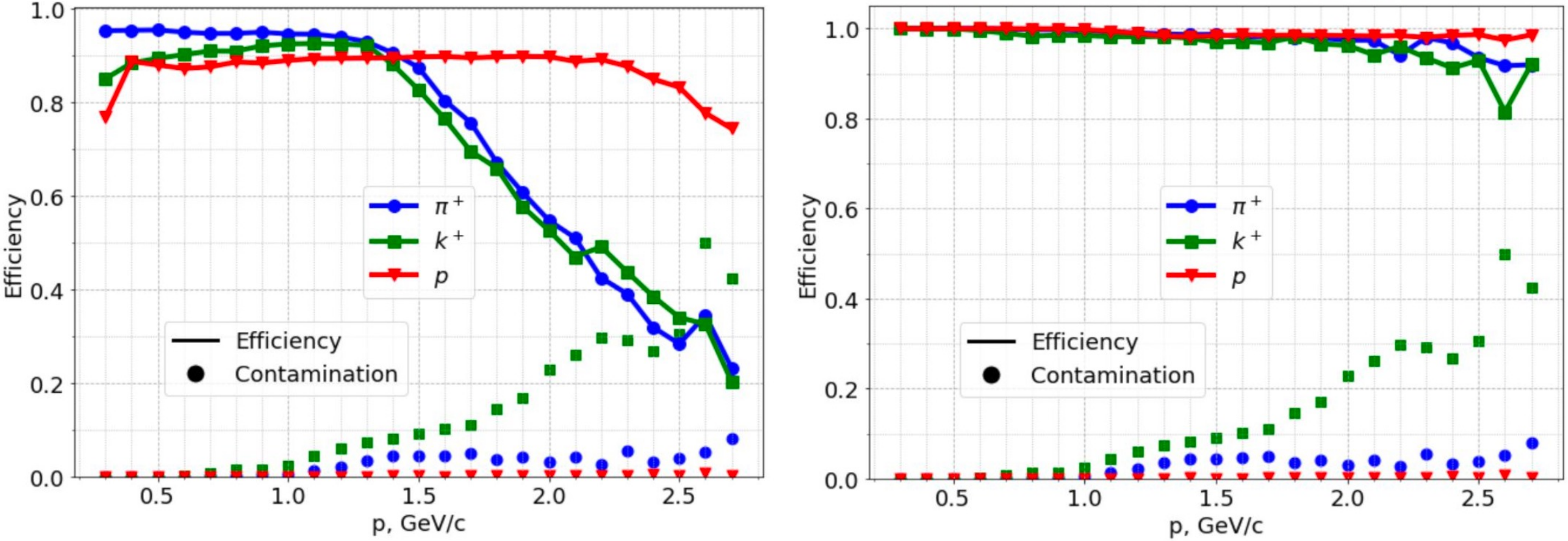
Gradient boosting is a machine learning technique which combines **weak learners** into a single strong learner in an iterative fashion.





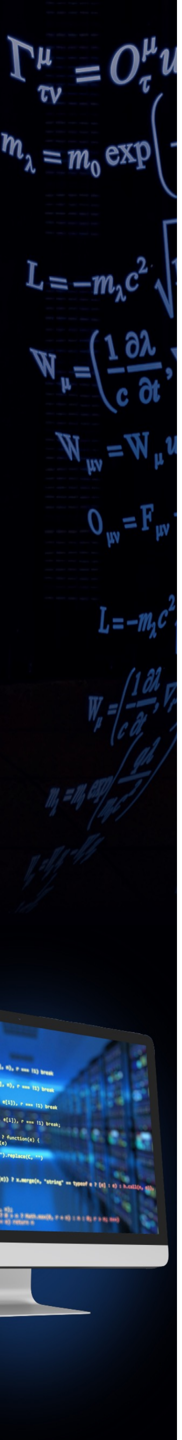
When **weak learners are decision tree**, the resulting algorithm is called **gradient-boosted decision trees**.

Baseline PID in MPD - N-sigma



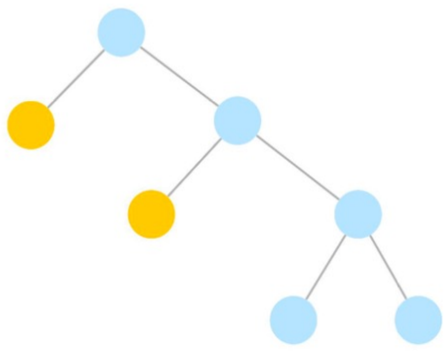
PID efficiency and contamination for all tracks (left) and only identified tracks (right) in Bi+Bi collisions at 9.2 GeV

$$E^S = \frac{N^S{}_{corr}}{N^S{}_{true}} \quad C^S = \frac{N^S{}_{incorr}}{N^S{}_{corr} + N^S{}_{incorr}}$$

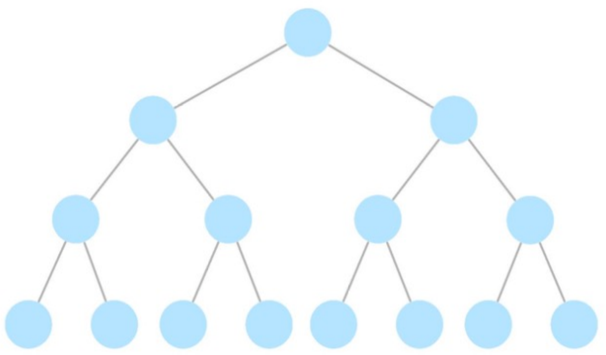


XGBoost vs LightGBM vs CatBoost vs SketchBoost

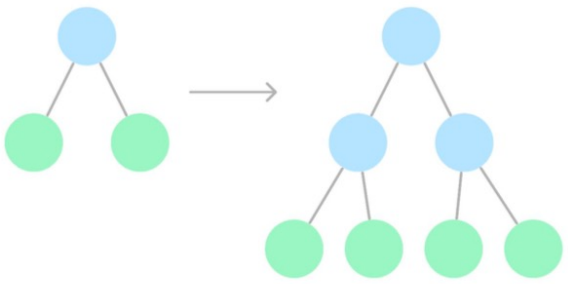
Asymmetric Tree (XGB, LGBM)



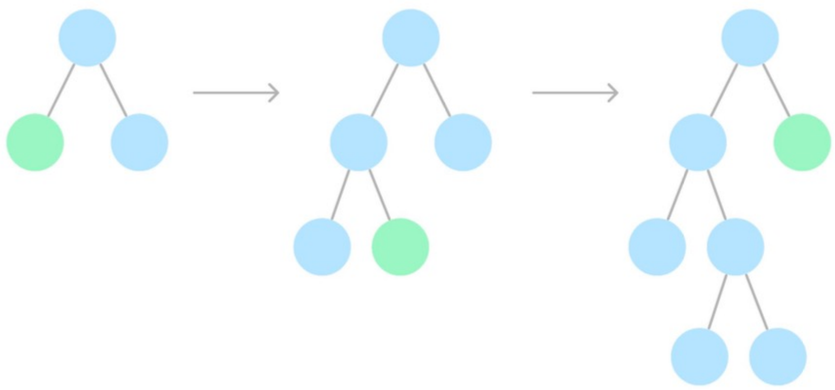
Symmetric Tree (CatBoost, SketchBoost)



Level-wise Tree Growth (XGB)



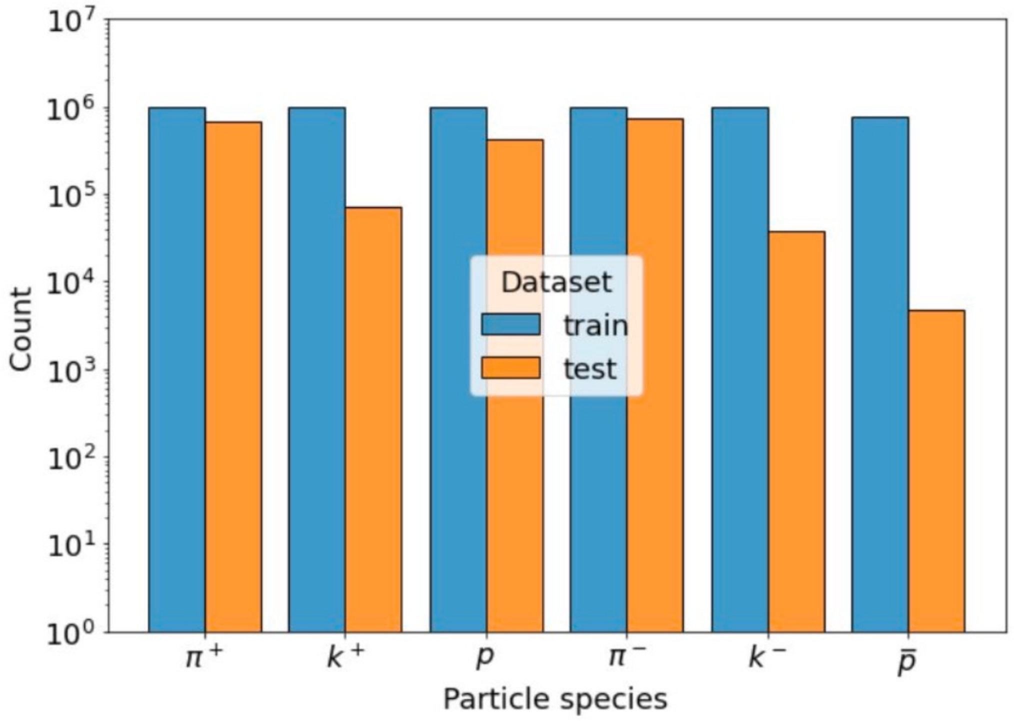
Leaf-wise Tree Growth (LGBM)



DATASET

Subsamples of the two MPD Monte-Carlo productions have been used

	prod05	prod06
Event generator	UrQMD	PHQMD
Transport	Geant 4	Geant 4
Impact parameter ranges	0-16 fm (mb)	0-12 fm
Smear Vertex XY	0.1 cm	0.1 cm
Smear Vertex Z	50 cm	50 cm
Colliding system	Bi+Bi	Bi+Bi
Energy	9.2 GeV	9.2 GeV



track selection criteria: (p < 100) & (|m^2| < 100) & (nHits > 15) & (|eta|<1.5) & (dca < 5) & (|Vz| < 100)

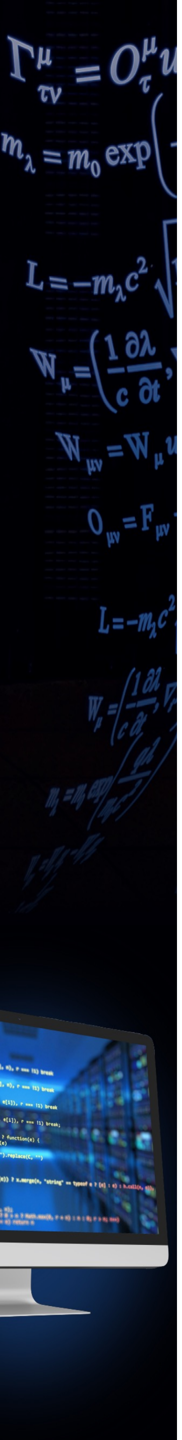
Two stages of the experiments

Some parameters for the tuning and model evaluation stages

Stage	Learning Rate	Max Number of Iterations	Early Stopping
Tuning	0.05	5 000	200
Model Evaluation	0.015	20 000	500

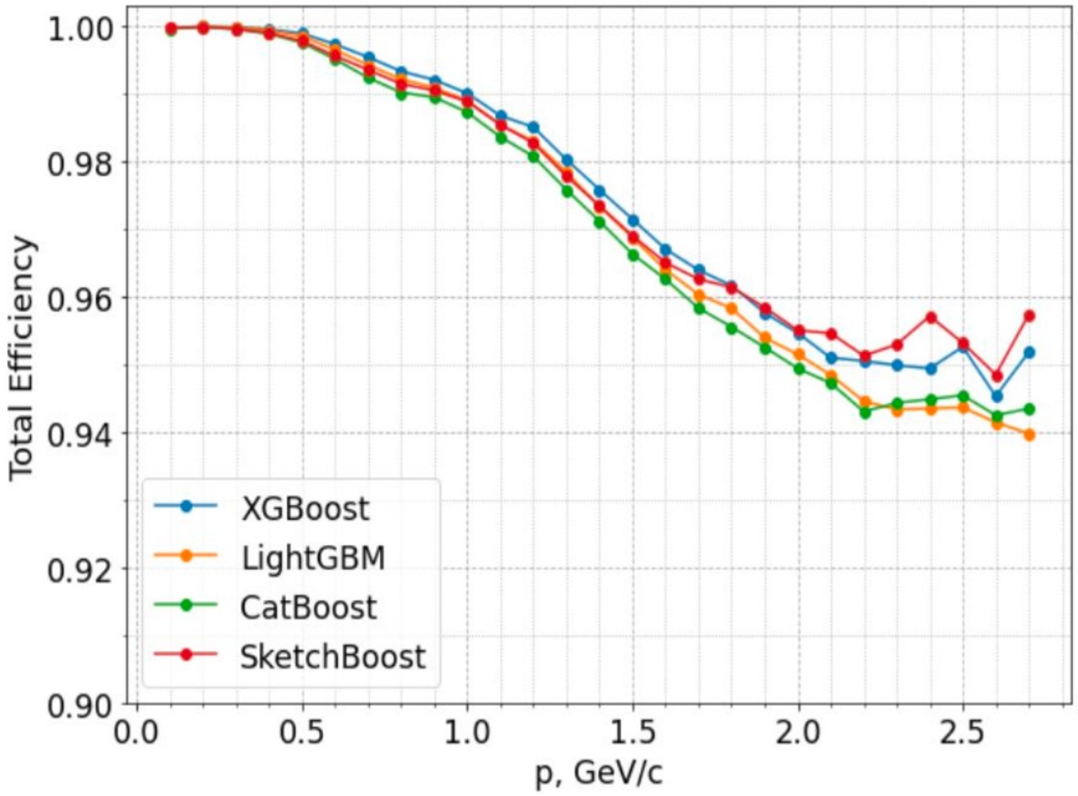
Results for hyperparameter tuning (after **30 iterations** of the TPE algorithm for each GBDT)

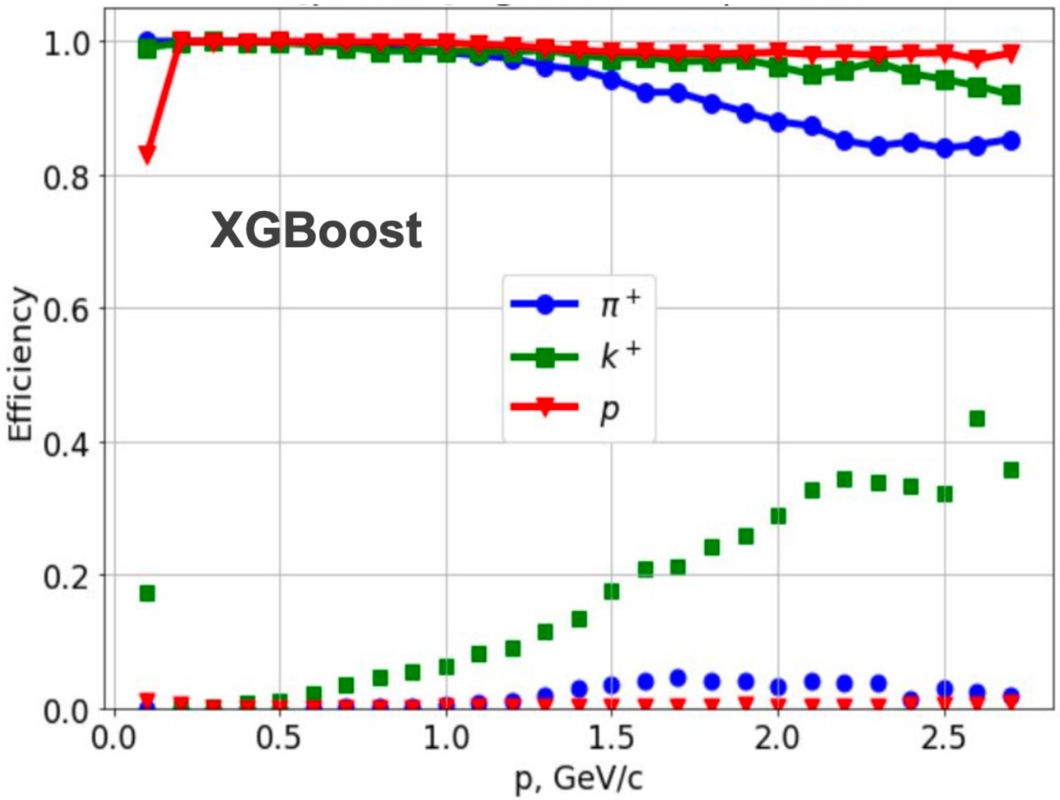
Framework	Max. Depth	L2 leaf reg.	Min. data in leaf size	Rows sampling rate
XGBoost	8	2.3	0.00234	0.942
LightGBM	12	0.1	4	0.981
CatBoost	8	3.0	5	0.99
SketchBoost	8	3.0	5	0.99



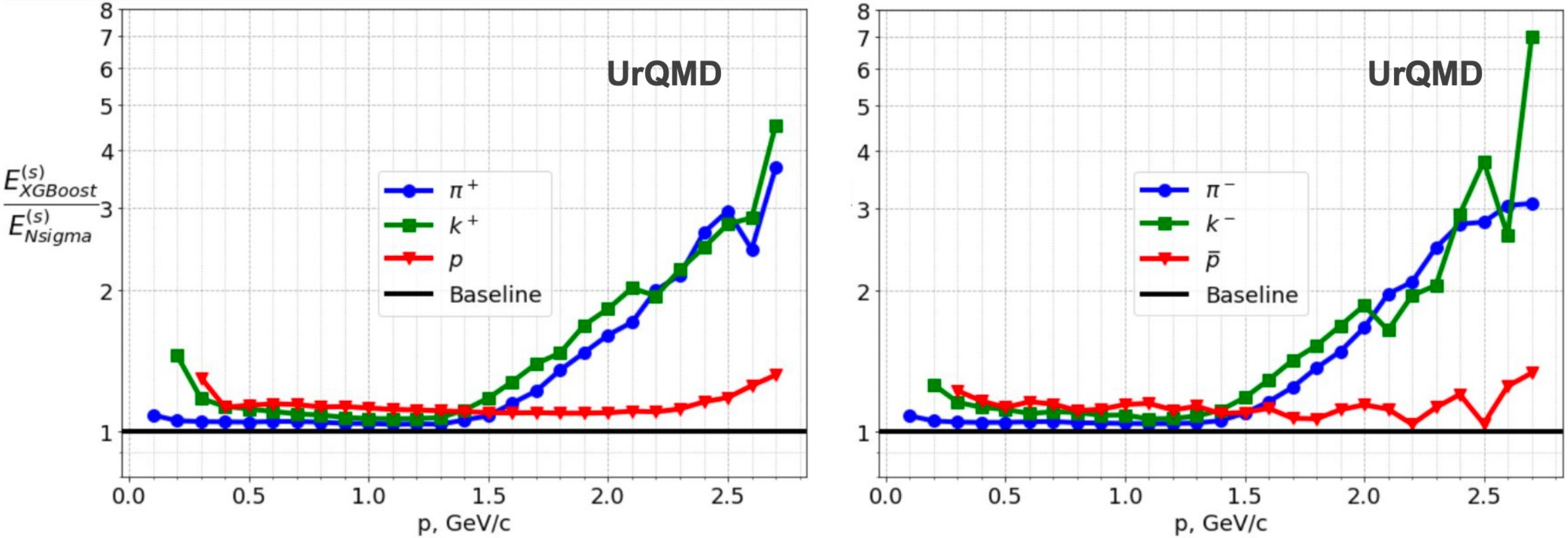
COMPARATIVE ANALYSIS OF THE ALGORITHMS

	XGBoost	LightGBM	CatBoost	SketchBoost
Total Efficiency	0.99327	0.99235	0.99138	0.99239



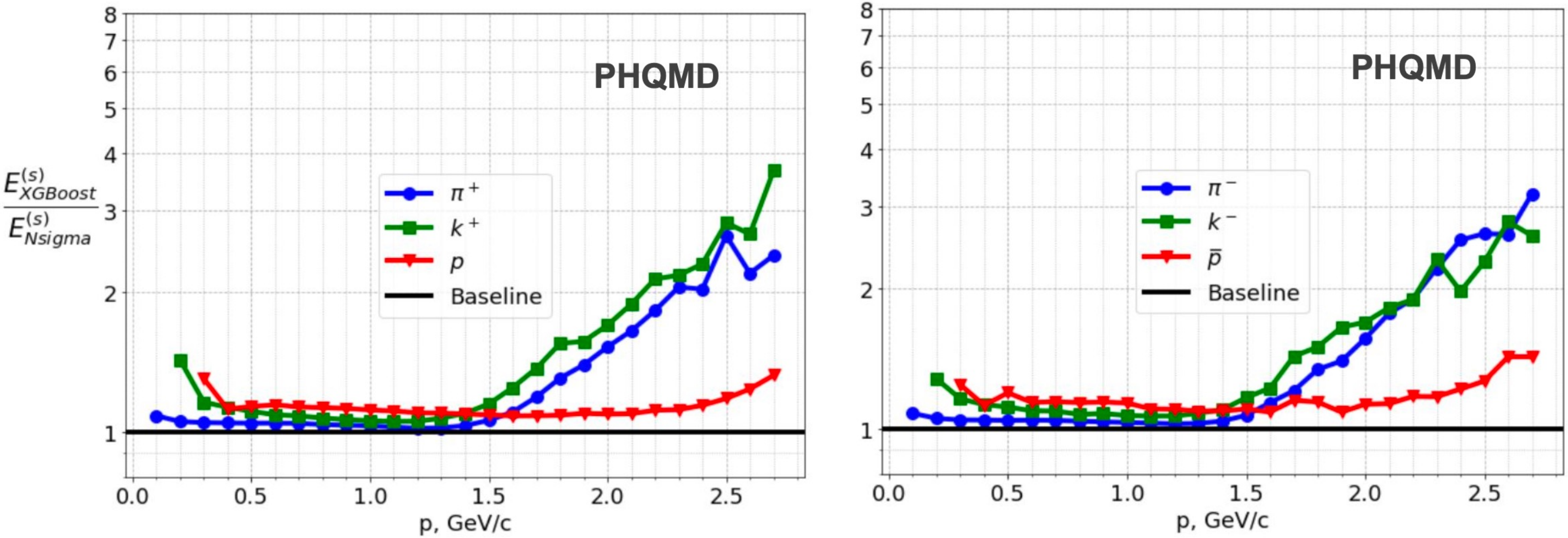


Comparison with N-sigma



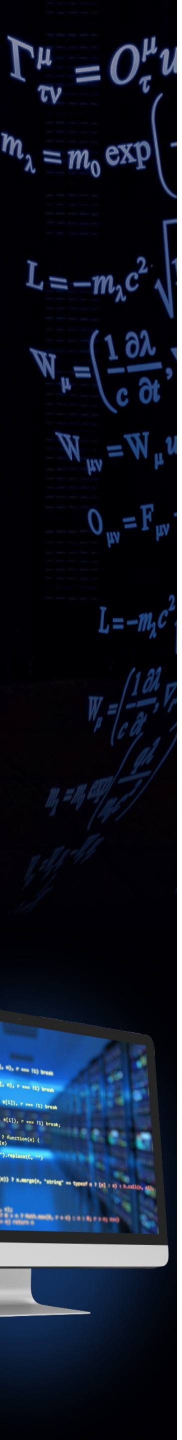
Efficiency ratio of XGBoost and n-sigma method

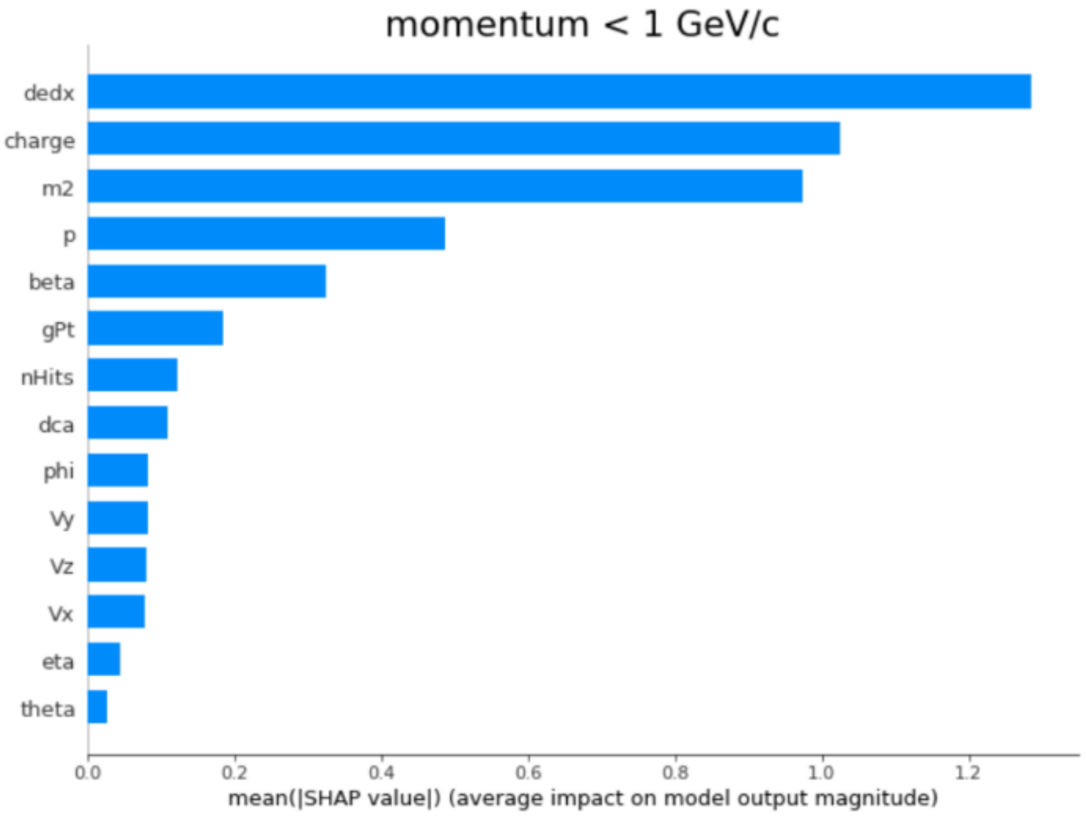
Comparison with N-sigma

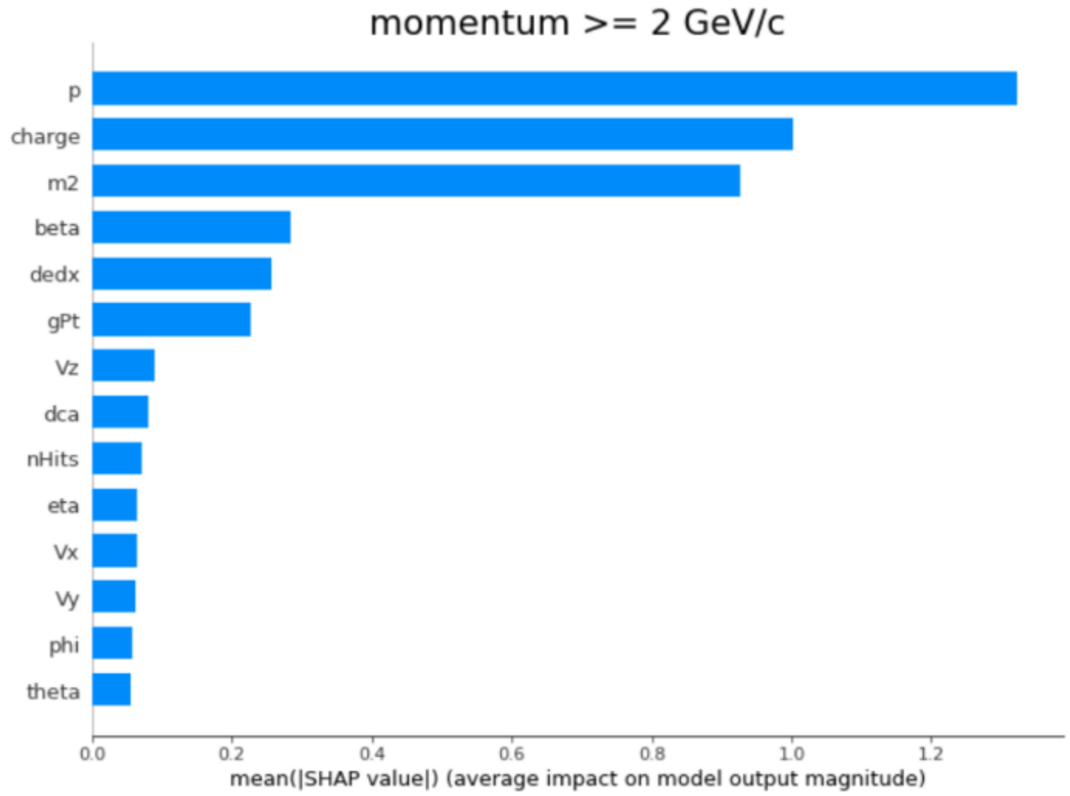


Efficiency ratio of XGBoost and n-sigma method

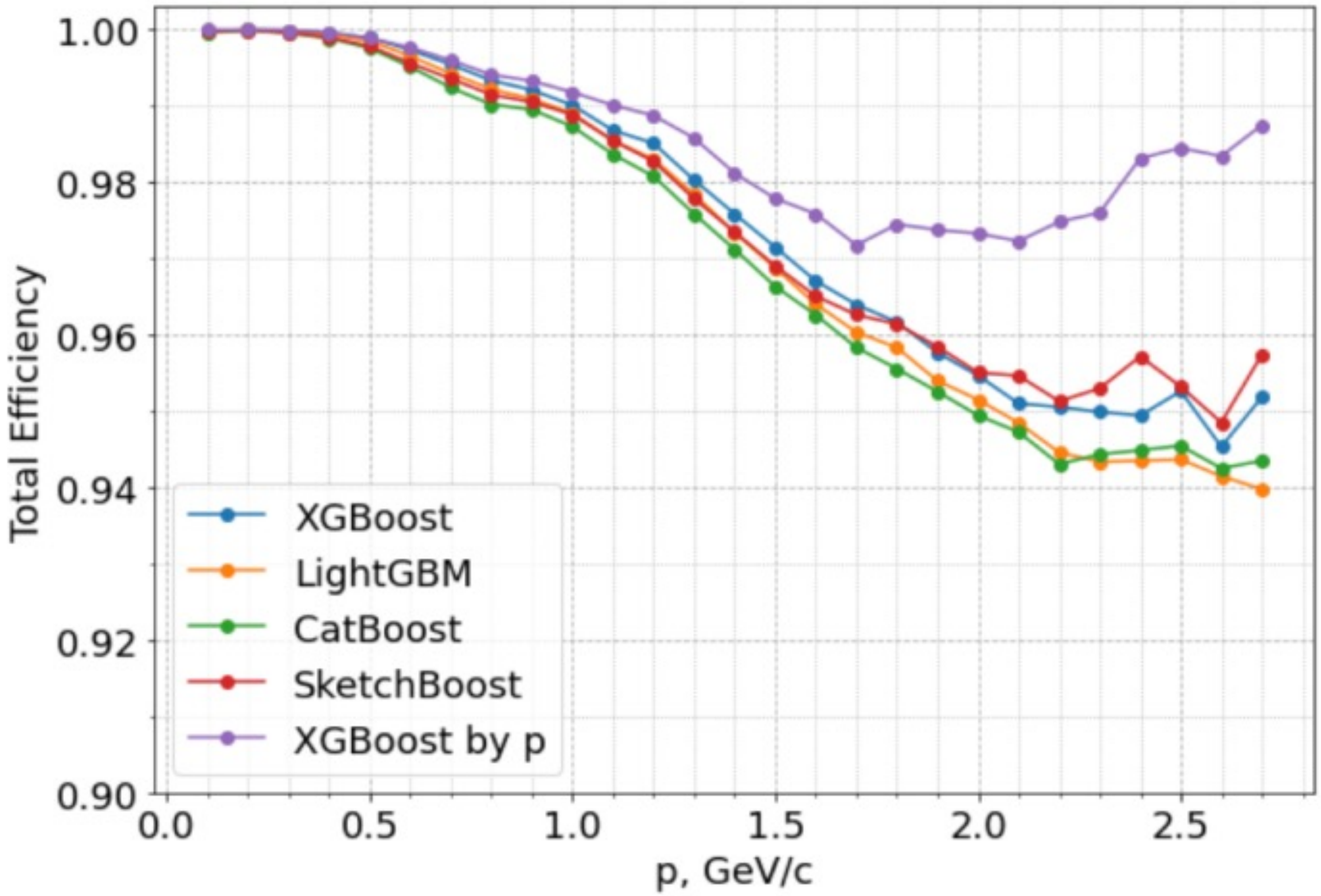
XGBoost Model Interpretation. Feature Importance







Final Efficiency of XGBoost

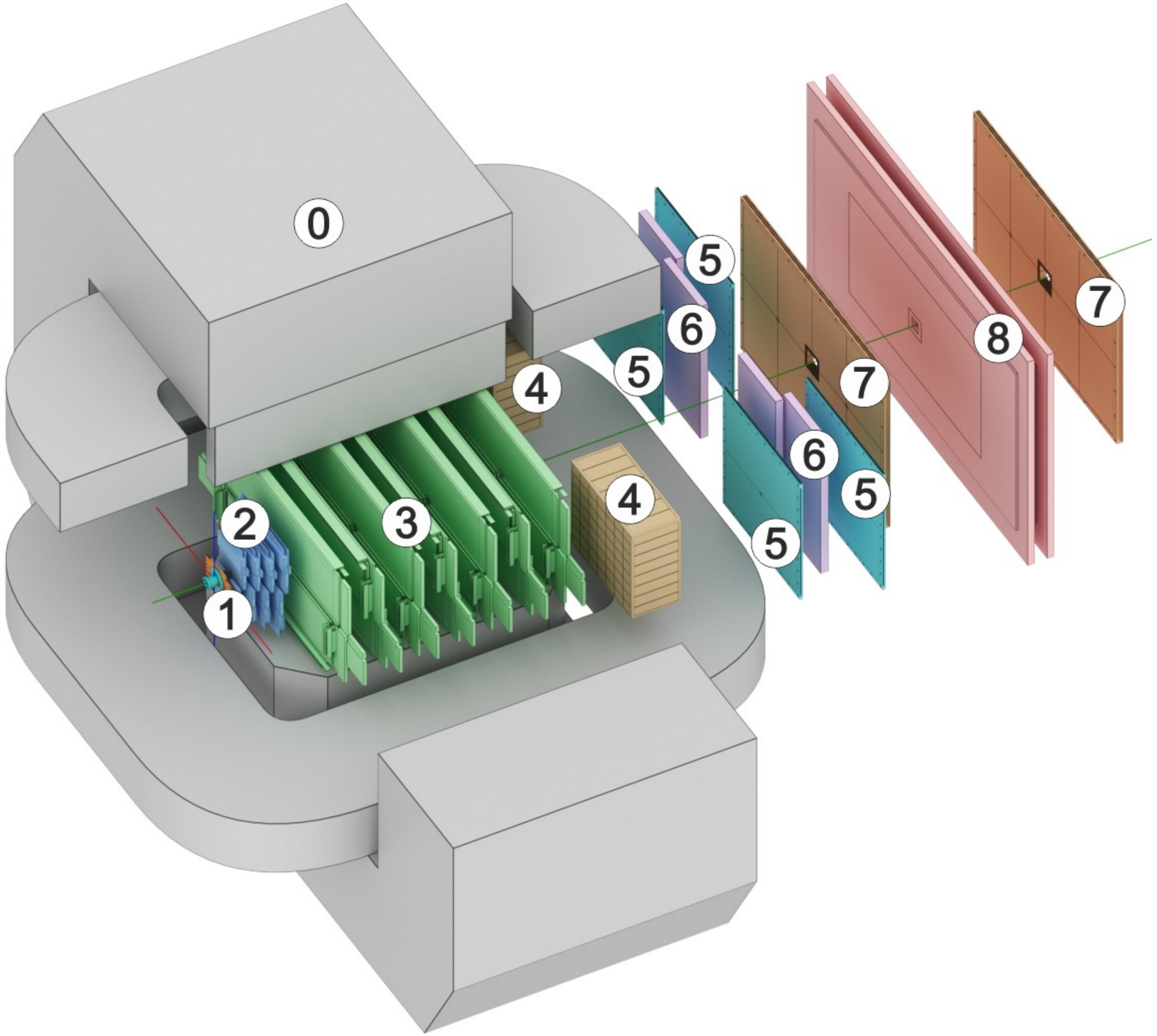


BM@N

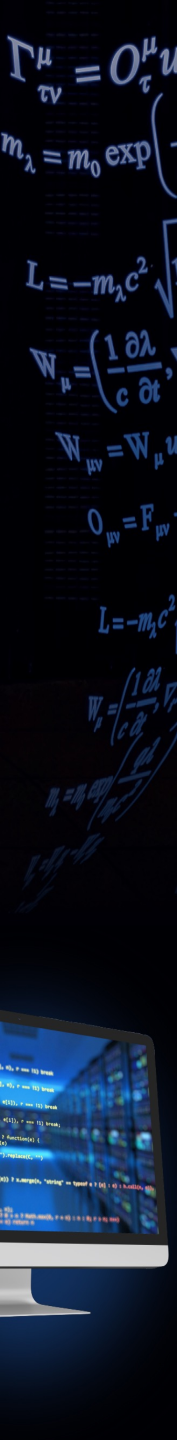
Baryonic Matter at Nuclotron

V. Papoyan, A. Ayriyan, K. Gertsenberger, H. Grigorian, S. Merts

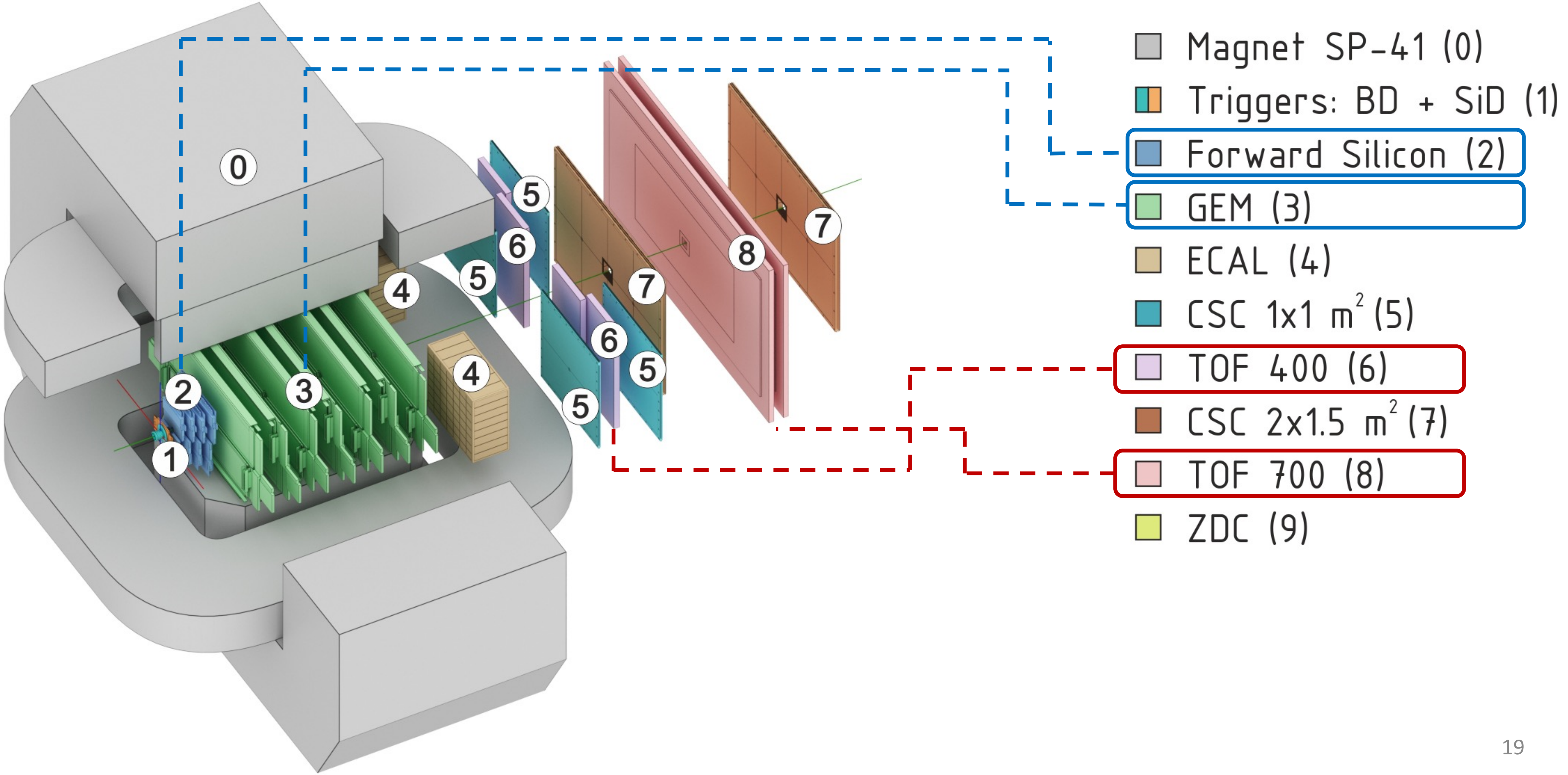
BMN DETECTOR

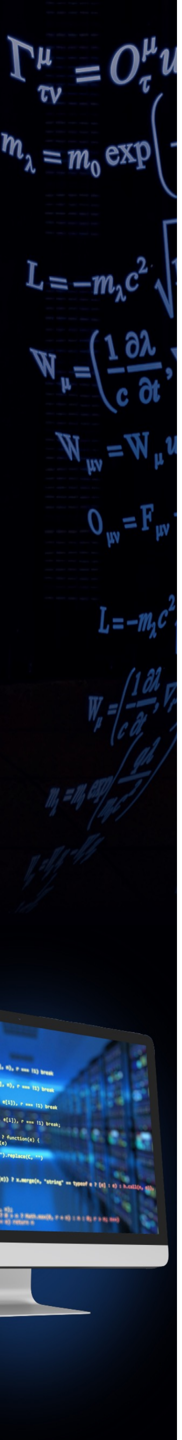


- Magnet SP-41 (0)
- Triggers: BD + SiD (1)
- Forward Silicon (2)
- GEM (3)
- ECAL (4)
- CSC 1x1 m^2 (5)
- TOF 400 (6)
- CSC 2x1.5 m^2 (7)
- TOF 700 (8)
- ZDC (9)



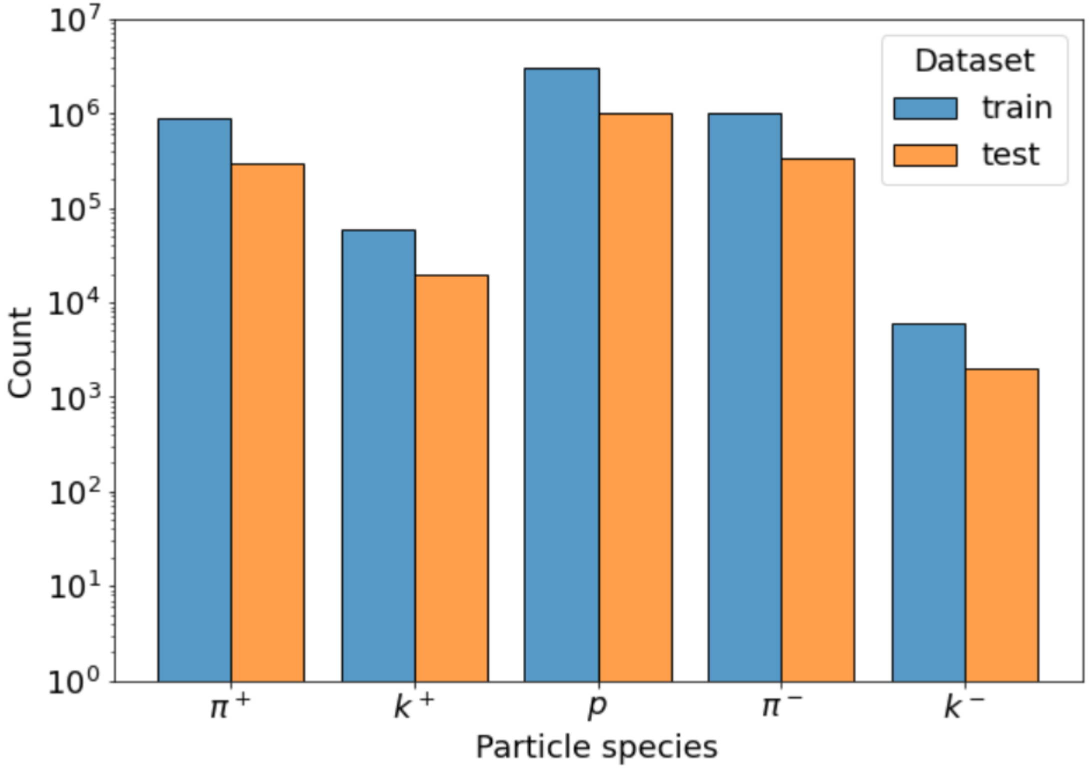
BMN Detector

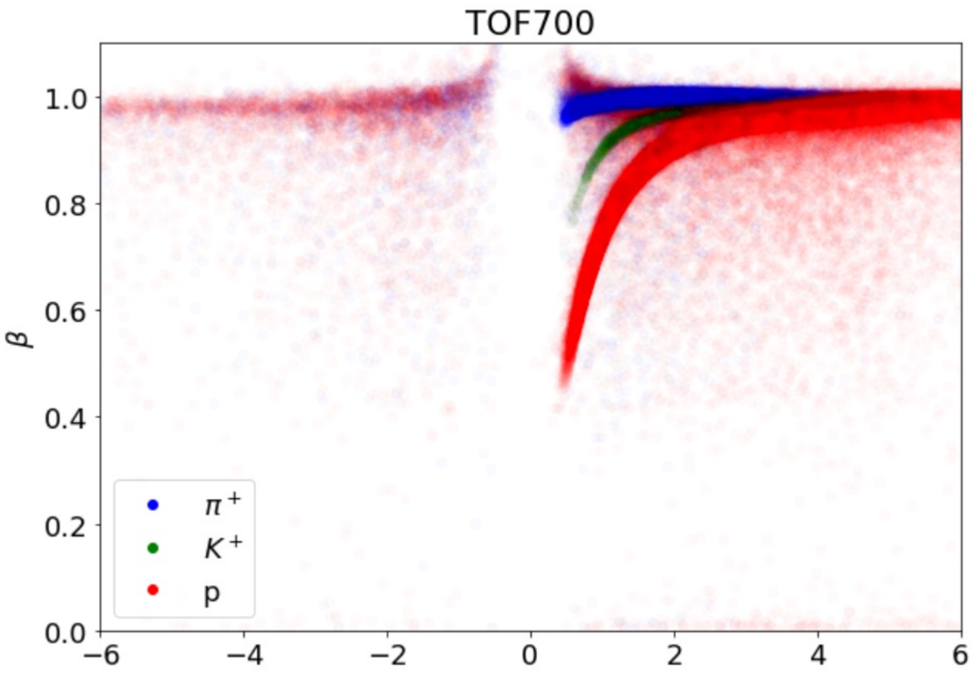




DATASET

- Number of trakcs: around 5M
  (60% protons, 40% pions, less than 1% of kaons)
- Number of traks with at least one ToF: approx. 1.4M (27%)

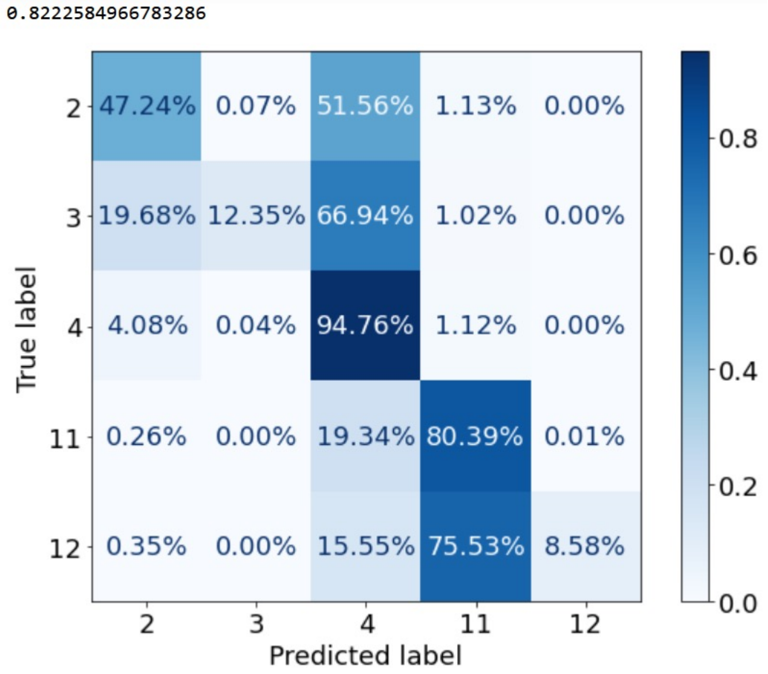


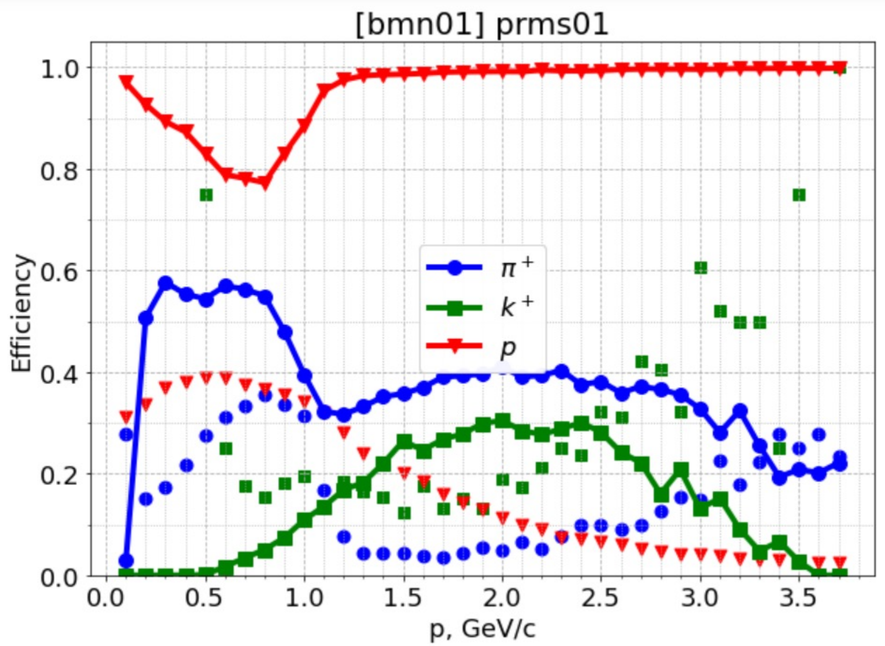


Results

- Number of trakcs: around 5M
  (60% protons, 40% pions, less than 1% of kaons)
- Number of traks with at least one ToF: approx. 1.4M (27%)

XGBoost shows identification efficiency more than 80%!





HOW?!

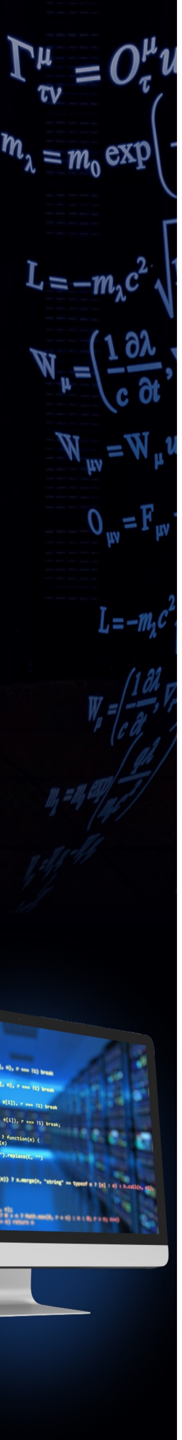
RESULTS

- Number of trakcs: around 5M
  (60% protons, 40% pions, less than 1% of kaons)
- Number of traks with at least one ToF: approx. 1.4M (27%)

XGBoost shows identification efficiency more than 80%!

HOW?!

Random efficincy: 80% minus 27% is approx 53%



RESULTS

- Number of trakcs: around 5M
  (60% protons, 40% pions, less than 1% of kaons)
- Number of traks with at least one ToF: approx. 1.4M (27%)

XGBoost shows 98.3% efficiency for traks with ToF!