

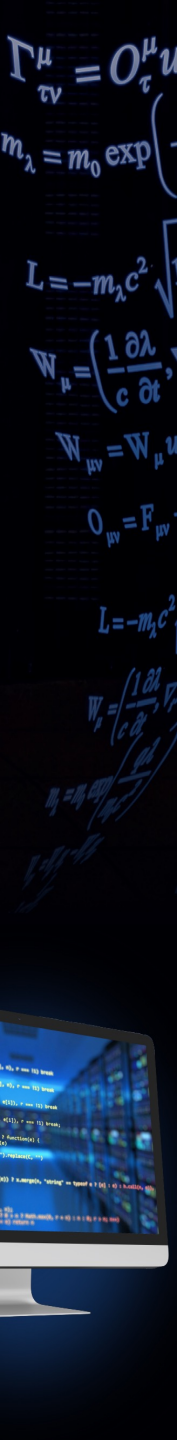
ML METHODS FOR PID IN HIC EXPERIMENTS

ALEXANDER AYRIYAN

JINR SCHOOL OF INFORMATION TECHNOLOGIES

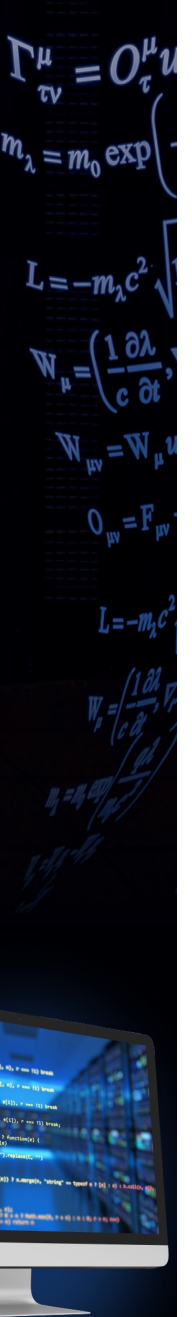
16 – 20 OCTOBER 2023

MESHCHERYAKOV LABORATORY OF INFORMATION TECHNOLOGIES, DUBNA



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
 - **$m : X \rightarrow Y$** ,
- for values which known only on objects from the finite training set
 - **$X^n = (x_1, y_1), \dots, (x_n, y_n)$** ,
- Goal is to find an algorithm **a** that classifies an arbitrary new object $x \in X$
 - **$a : X \rightarrow Y$** .

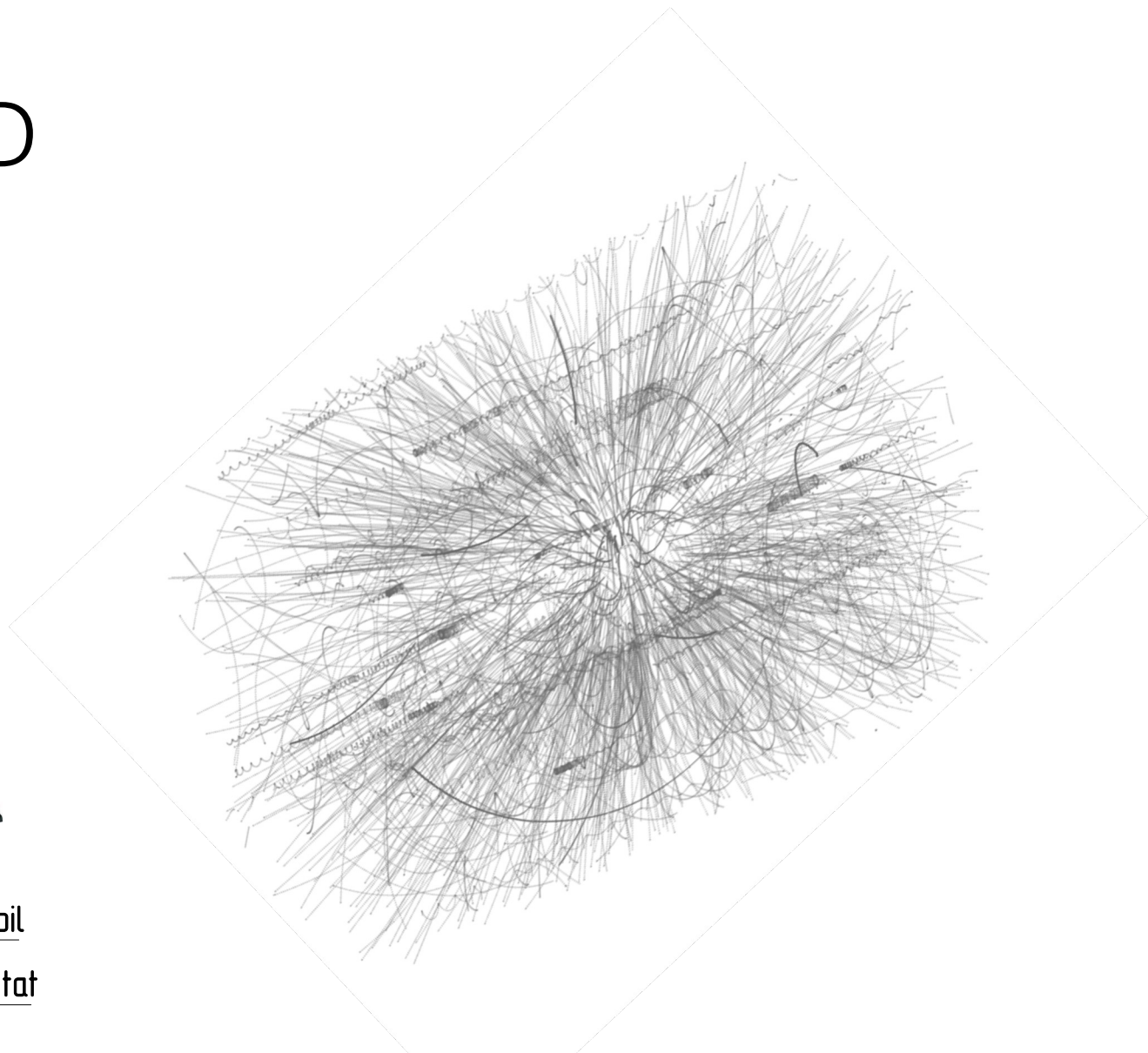
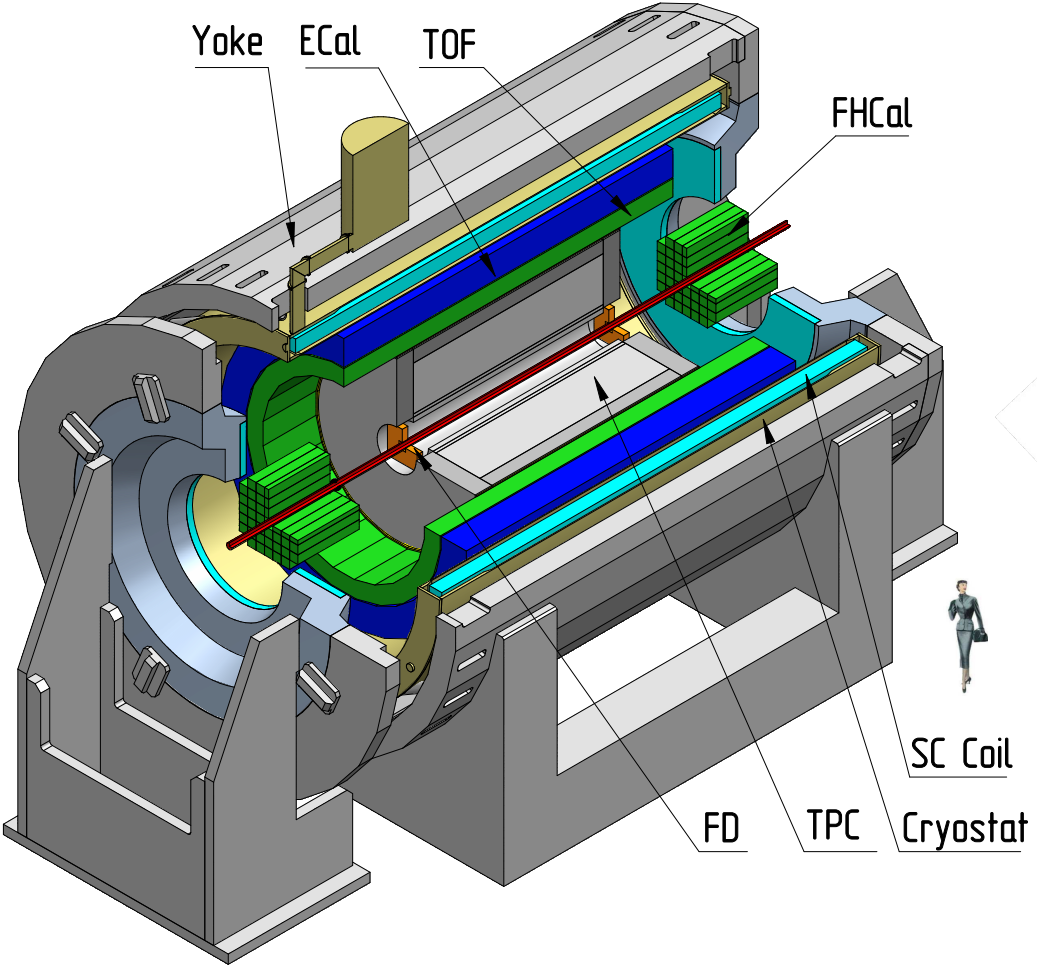


MPPD



$\Gamma_{\tau\nu}^{\mu} = O_{\tau}^{\mu}$
 $m_{\lambda} = m_0 \exp$
 $L = -m_{\lambda} c^2$
 $W_{\mu} = \left(\frac{1}{c} \frac{\partial \lambda}{\partial t} \right)$
 $W_{\nu} = W_{\mu}$
 $O_{\nu} = F_{\nu}$
 $L = -m_{\lambda} c^2$
 $W_{\mu} = \left(\frac{1}{c} \frac{\partial \lambda}{\partial t} \right)$
 $\mu = \frac{1}{c} \frac{\partial \lambda}{\partial t}$

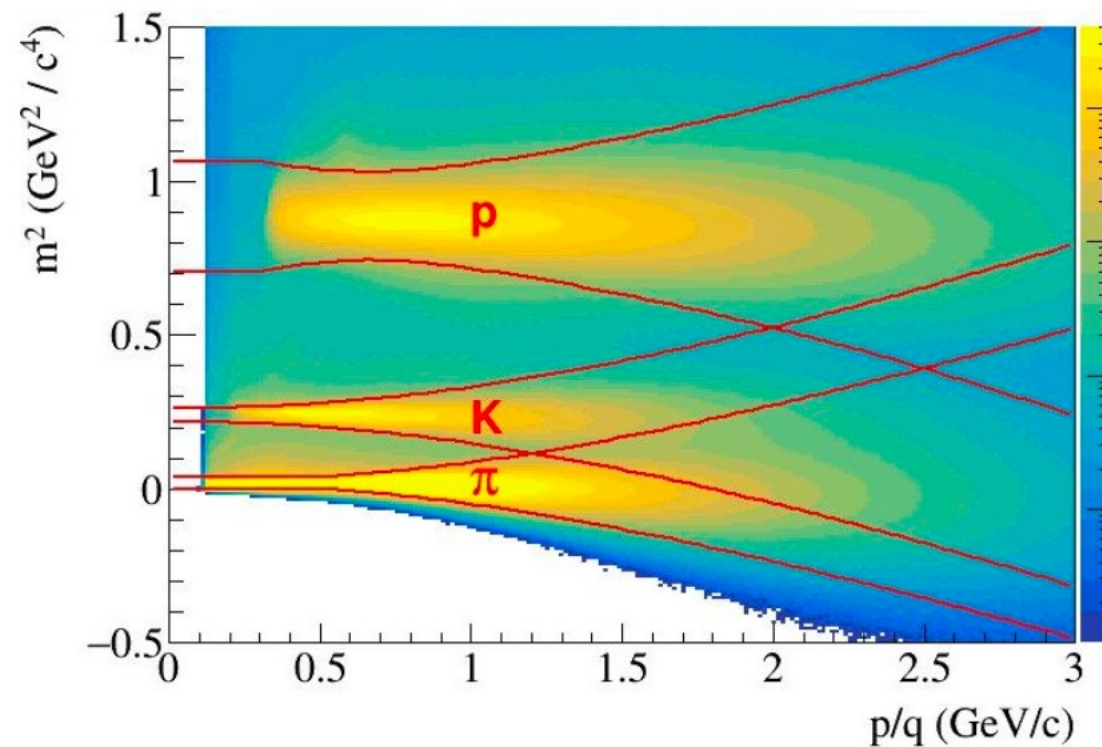
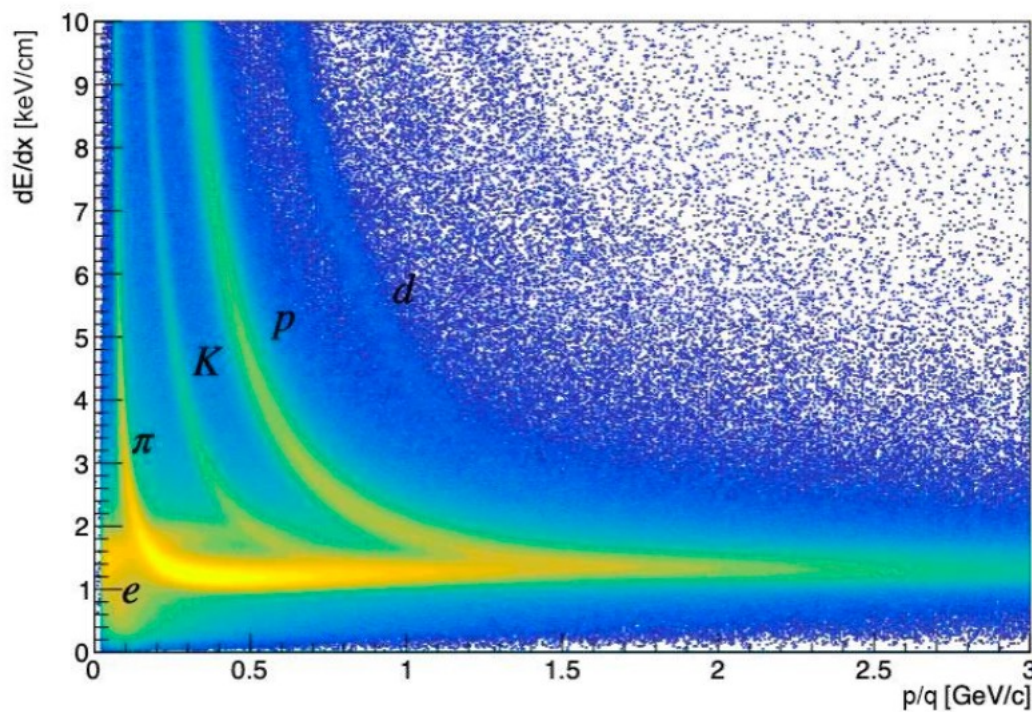
MPD APPARATUS AND PID



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

PARTICLE IDENTIFICATION IN MPD EXPERIMENT

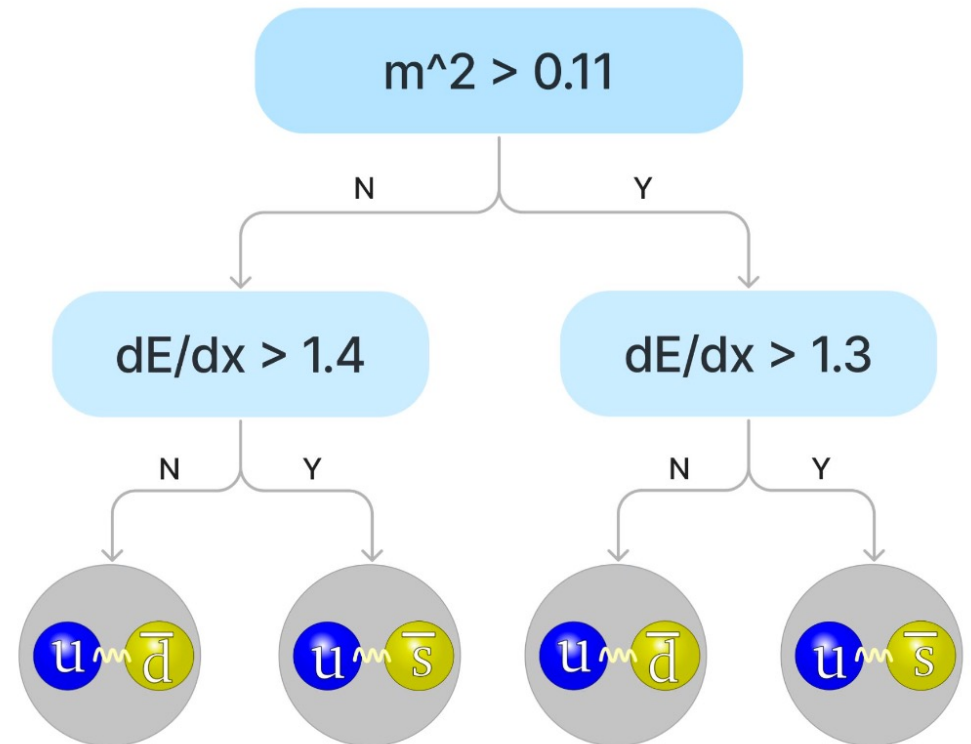
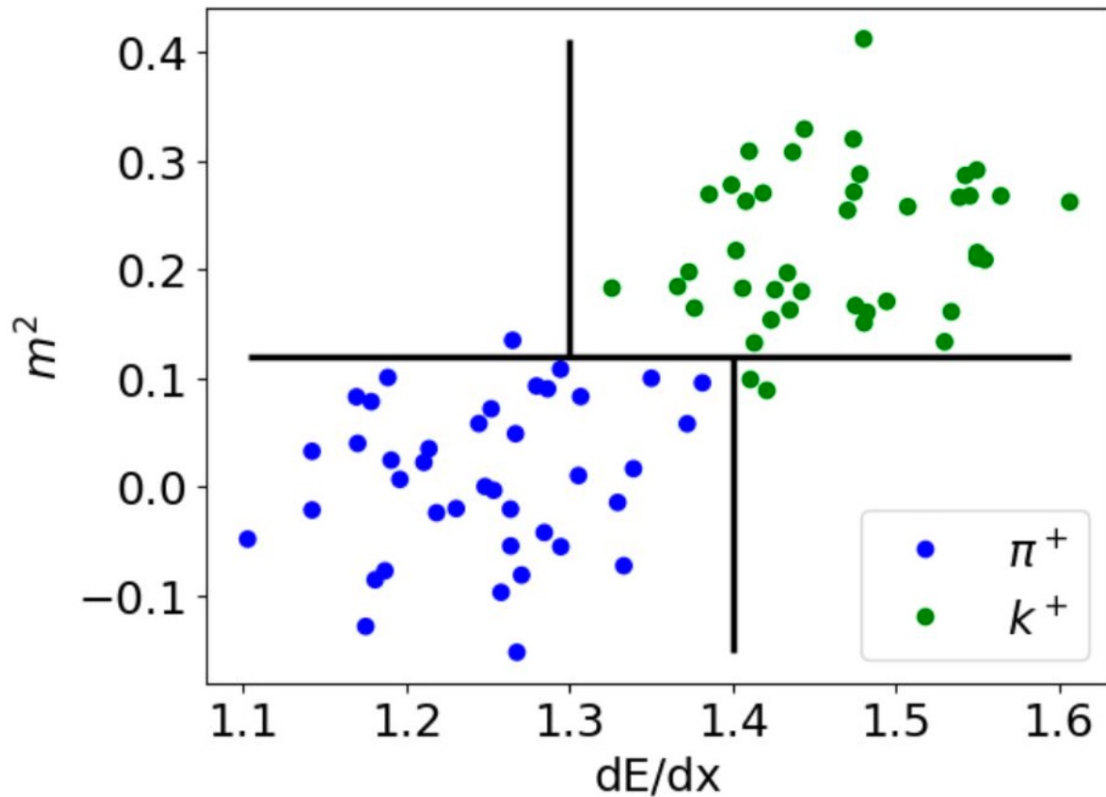
Particle identification can be achieved by using information about **momentum, charge, energy loss (TPC)** and **mass squared (TPC + TOF)**.



$\Gamma_{\tau\nu}^{\mu} = O_{\tau}^{\mu}$
 $m_{\lambda} = m_0 \exp$
 $L = -m_{\lambda} c^2$
 $W_{\mu} = \left(\frac{1}{c} \frac{\partial \lambda}{\partial t} \right)$
 $W_{\nu} = W_{\mu}$
 $O_{\nu} = F_{\nu}$
 $L = -m_{\lambda} c^2$
 $W_{\mu} = \left(\frac{1}{c} \frac{\partial \lambda}{\partial t} \right)$

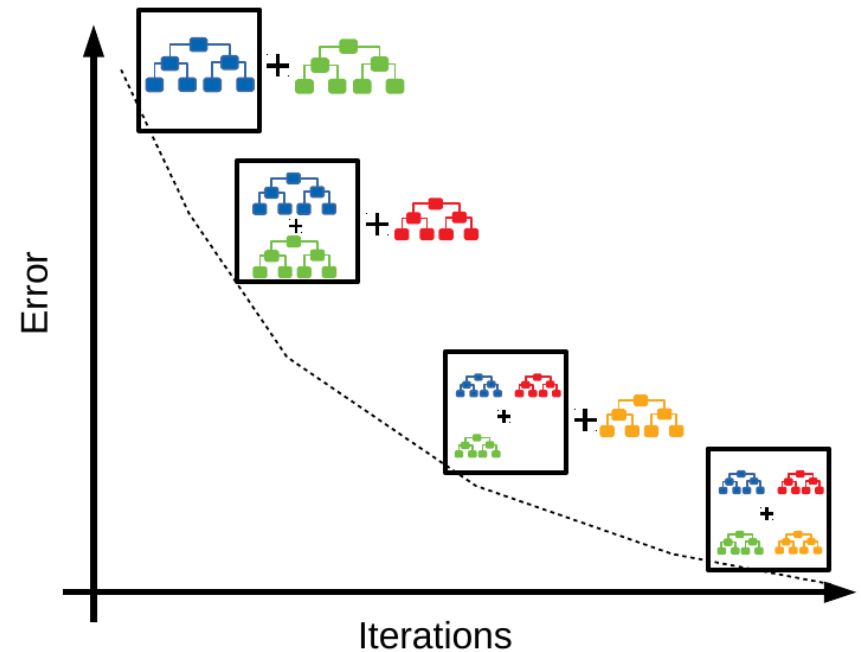
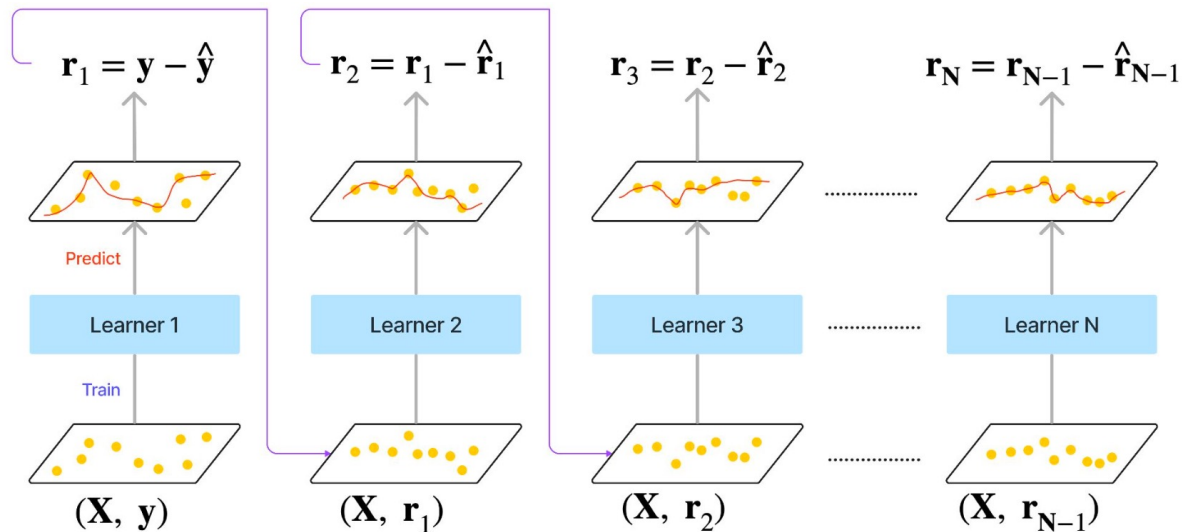
DECISION TREES FOR PID

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



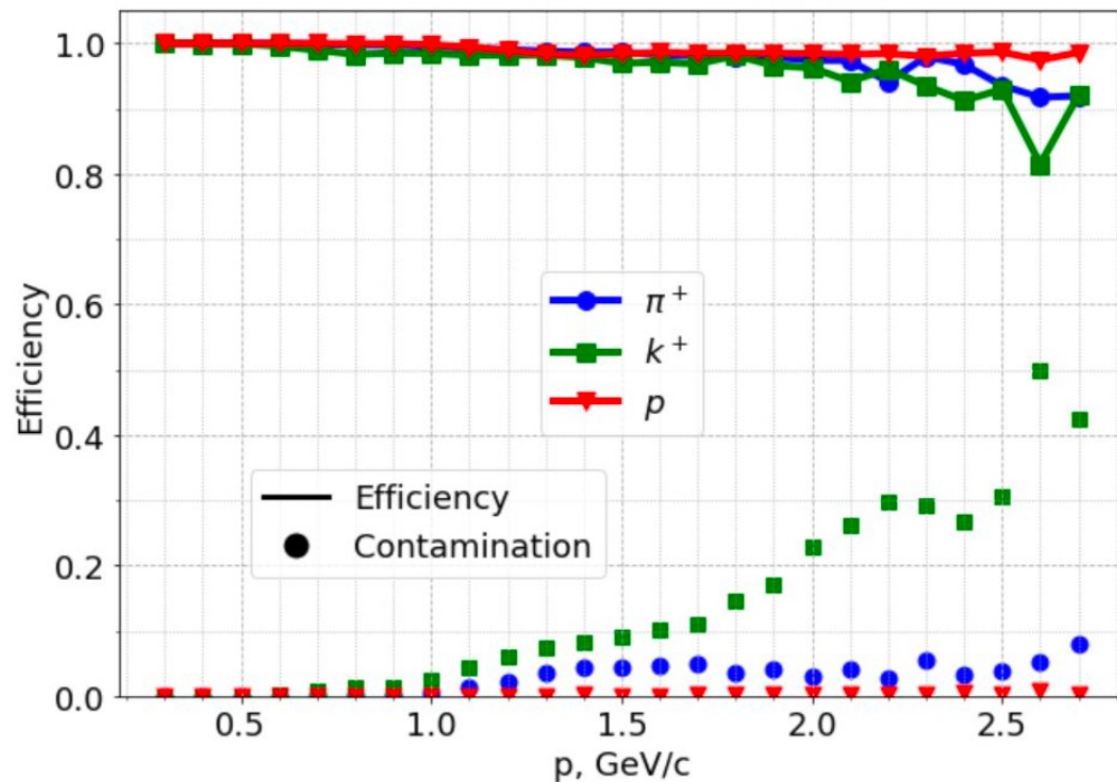
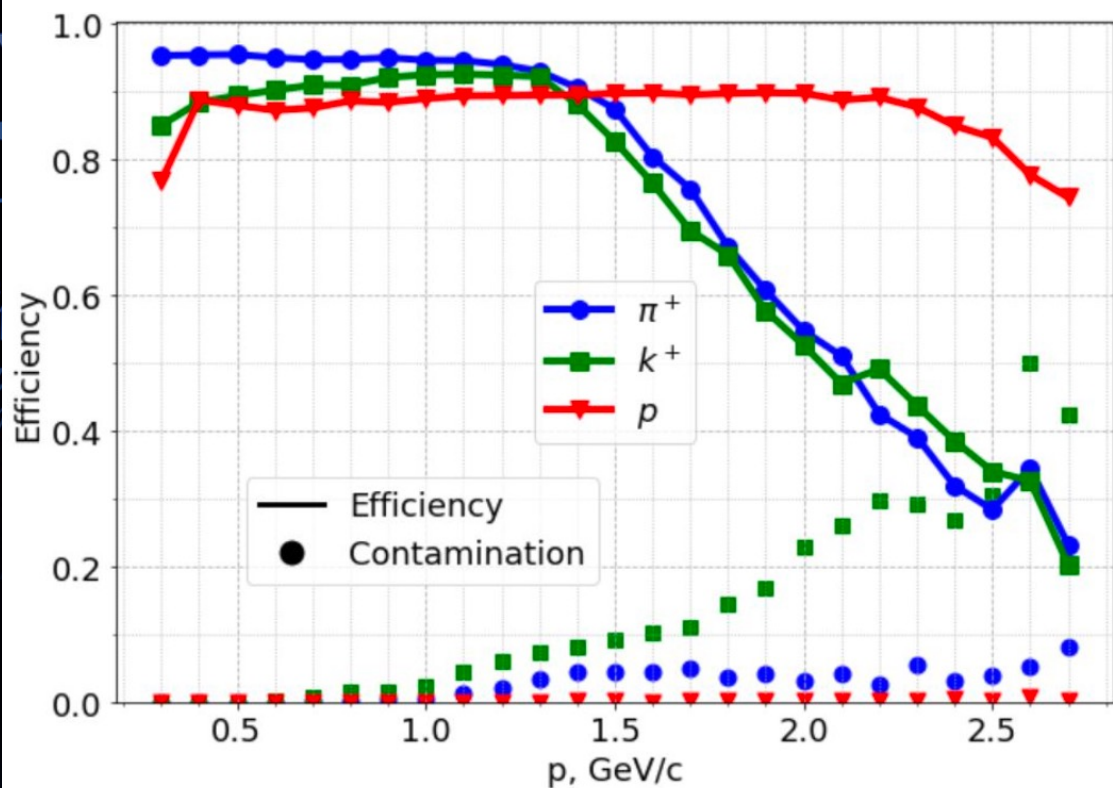
GRADIENT BOOSTING

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



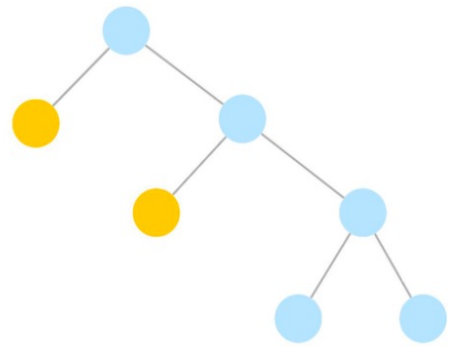
PID efficiency and contamination for all tracks (left) and only identified tracks (right) in Bi+Bi collisions at 9.2 GeV

$$E^s = \frac{N^s_{corr}}{N^s_{true}} \quad C^s = \frac{N^s_{incorr}}{N^s_{corr} + N^s_{incorr}}$$

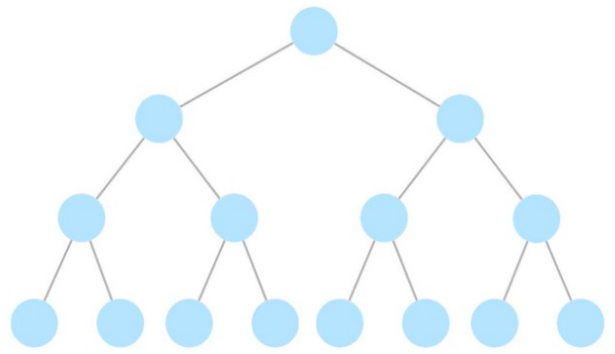
$\Gamma_{\tau\nu}^{\mu} = O_{\tau}^{\mu}$
 $m_{\lambda} = m_0 \exp$
 $L = -m_{\lambda} c^2$
 $W_{\mu} = \left(\frac{1}{c} \frac{\partial \lambda}{\partial t}\right)$
 $W_{\nu} = W_{\mu}$
 $O_{\nu} = F_{\nu}$
 $L = -m_{\lambda} c^2$
 $W_{\mu} = \left(\frac{1}{c} \frac{\partial \lambda}{\partial t}\right)$
 $\mu = \frac{1}{c} \frac{\partial \lambda}{\partial t}$

XGBOOST VS LIGHTGBM VS CATBOOST VS SKETCHBOOST

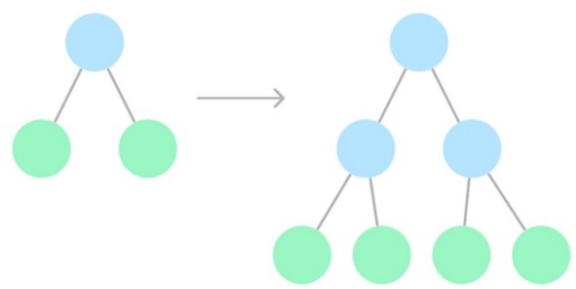
Asymmetric Tree (XGB, LGBM)



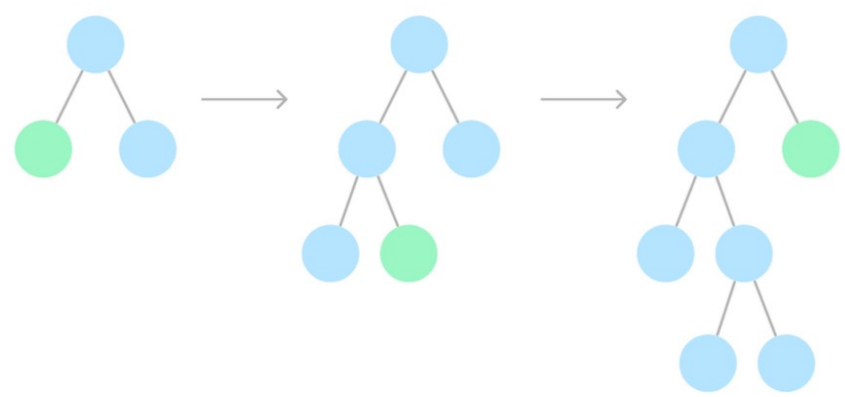
Symmetric Tree (CatBoost, SketchBoost)



Level-wise Tree Growth (XGB)



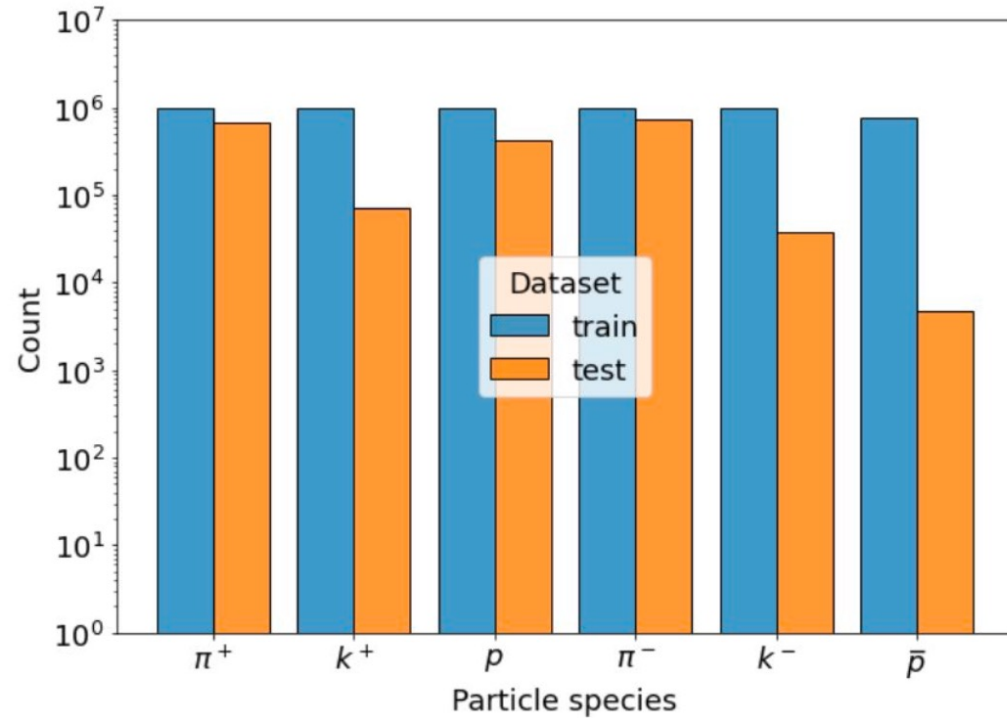
Leaf-wise Tree Growth (LGBM)



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

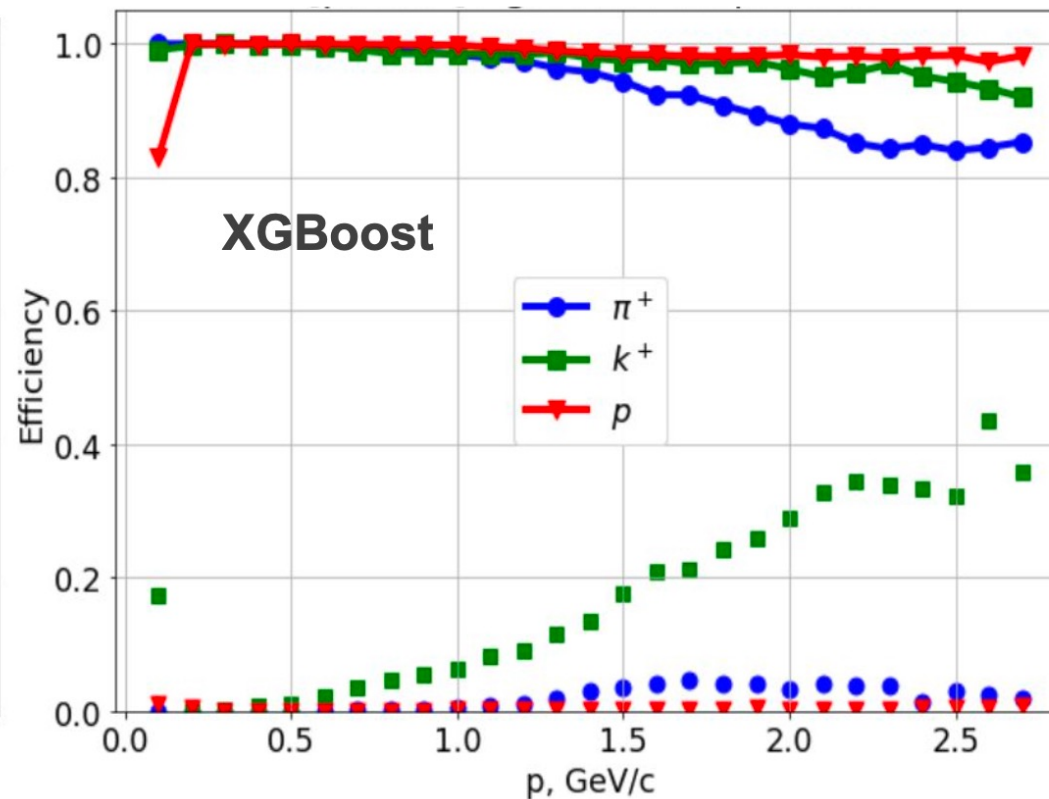
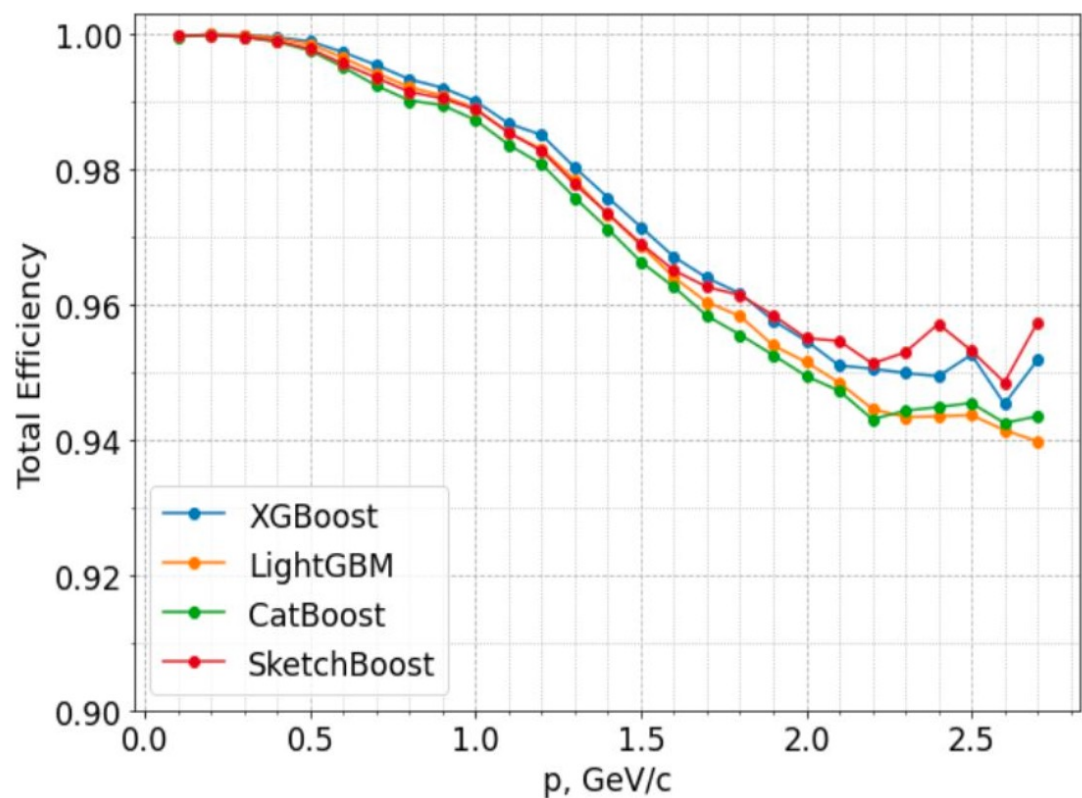
Stage	Learning Rate	Max Number of Iterations	Early Stopping
Tuning	0.05	5 000	200
Model Evaluation	0.015	20 000	500

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

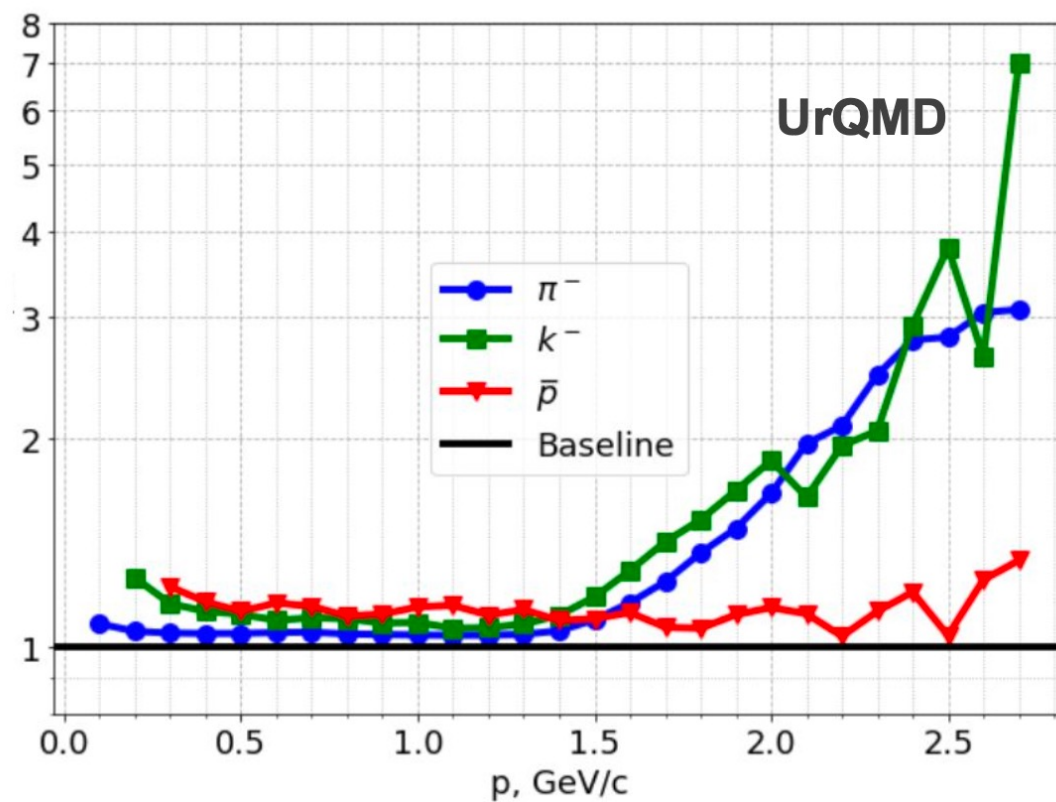
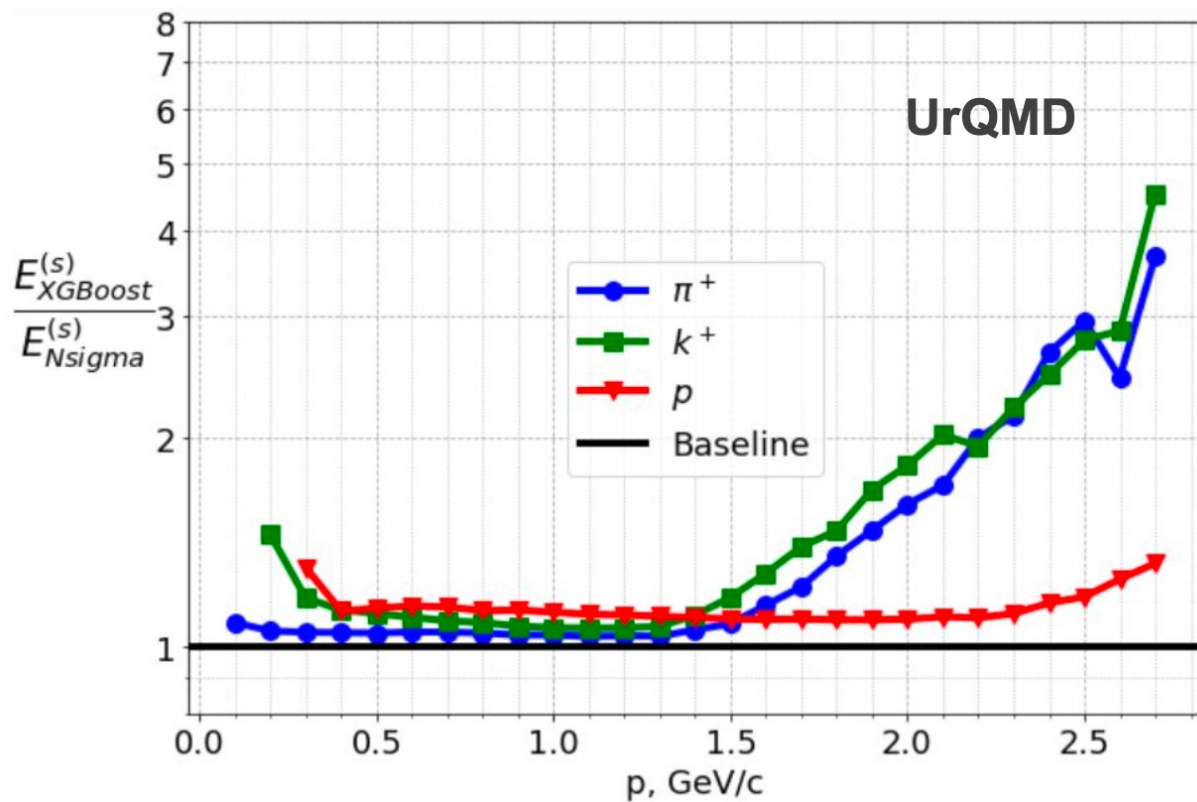
Framework	Max. Depth	L2 leaf reg.	Min. data in leaf size	Rows sampling rate
XGBoost	8	2.3	0.00234	0.942
LightGBM	12	0.1	4	0.981
CatBoost	8	3.0	5	0.99
SketchBoost	8	3.0	5	0.99

COMPARATIVE ANALYSIS OF THE ALGORITHMS

	XGBoost	LightGBM	CatBoost	SketchBoost
Total Efficiency	0.99327	0.99235	0.99138	0.99239

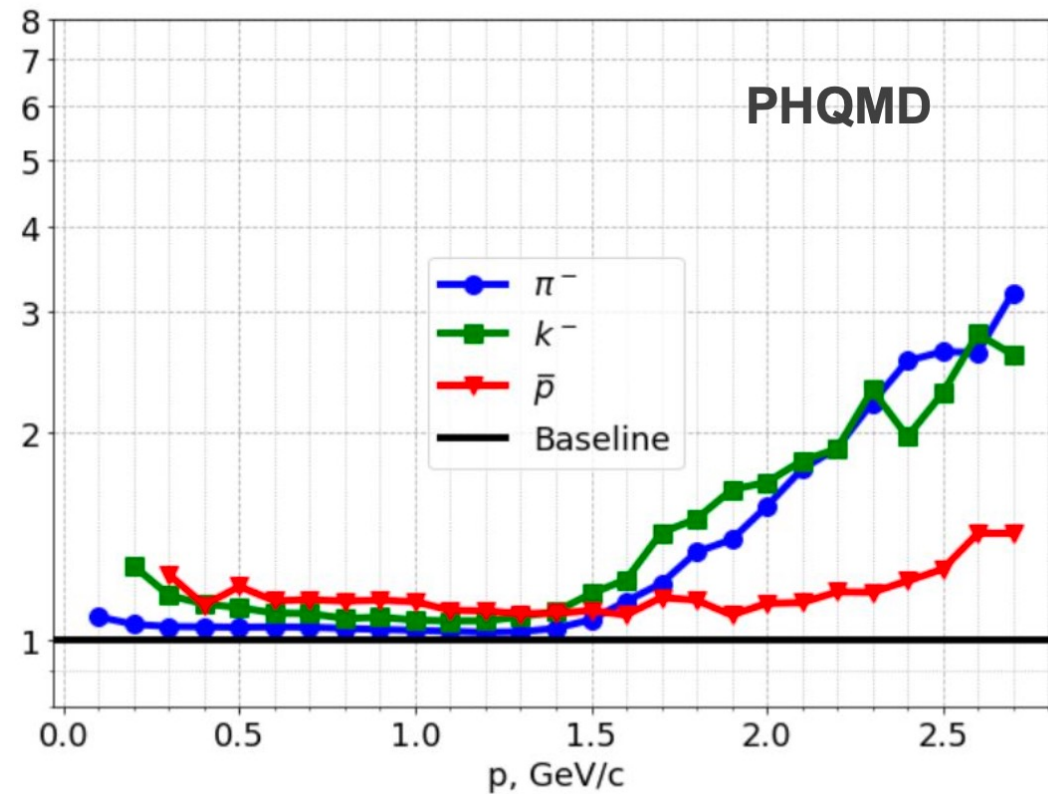
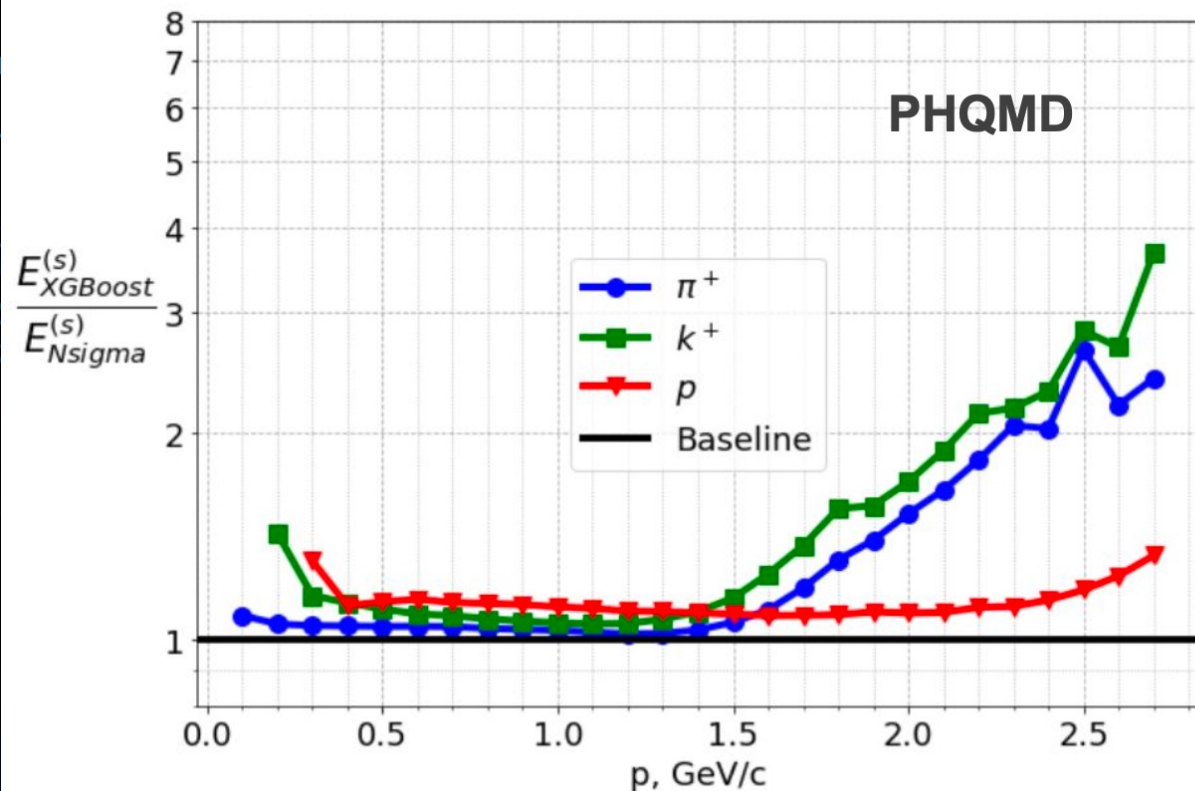


COMPARISON WITH N-SIGMA



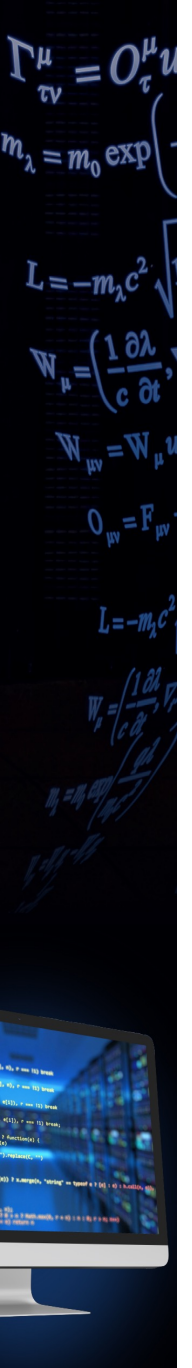
Efficiency ratio of XGBoost and n-sigma method

COMPARISON WITH N-SIGMA

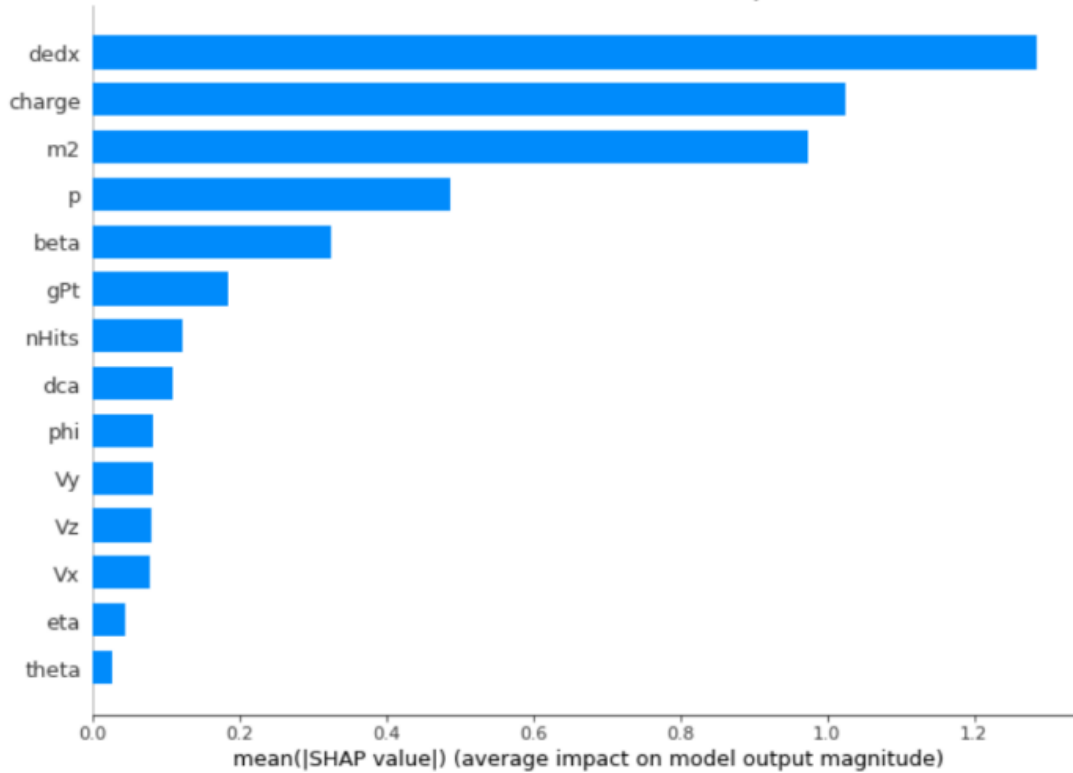


Efficiency ratio of XGBoost and n-sigma method

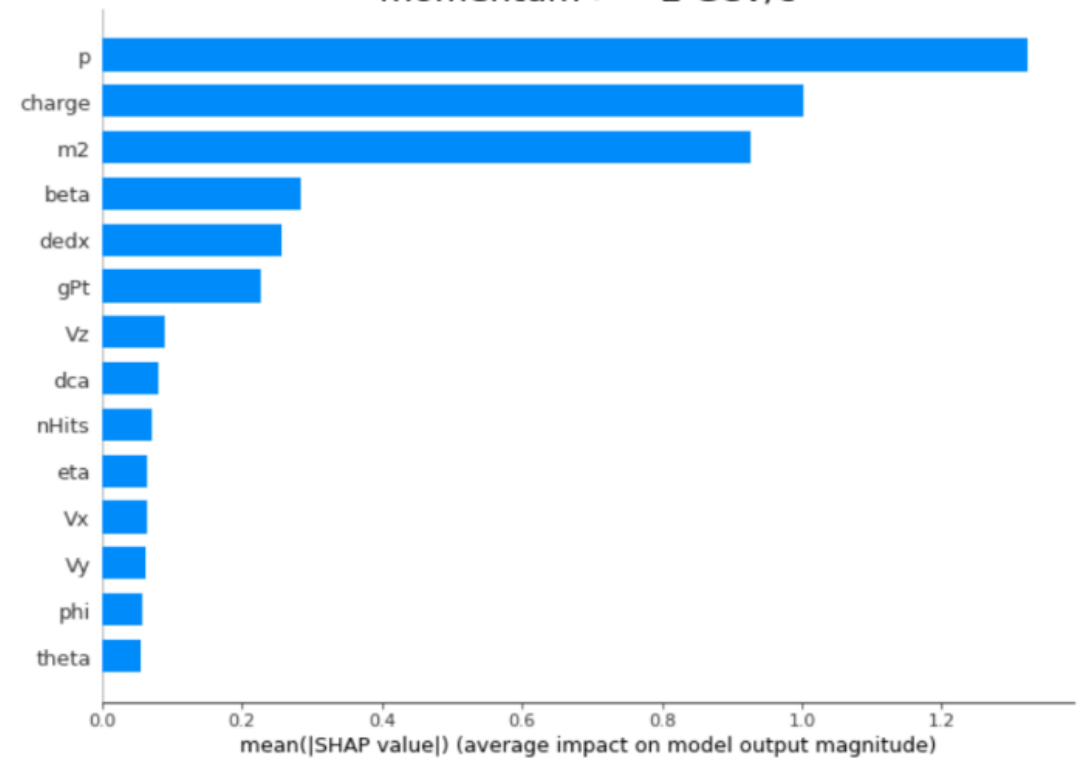
XGBOOST MODEL INTERPRETATION. FEATURE IMPORTANCE

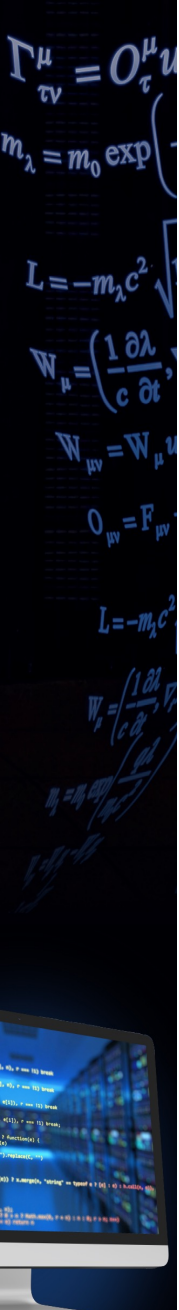


momentum < 1 GeV/c

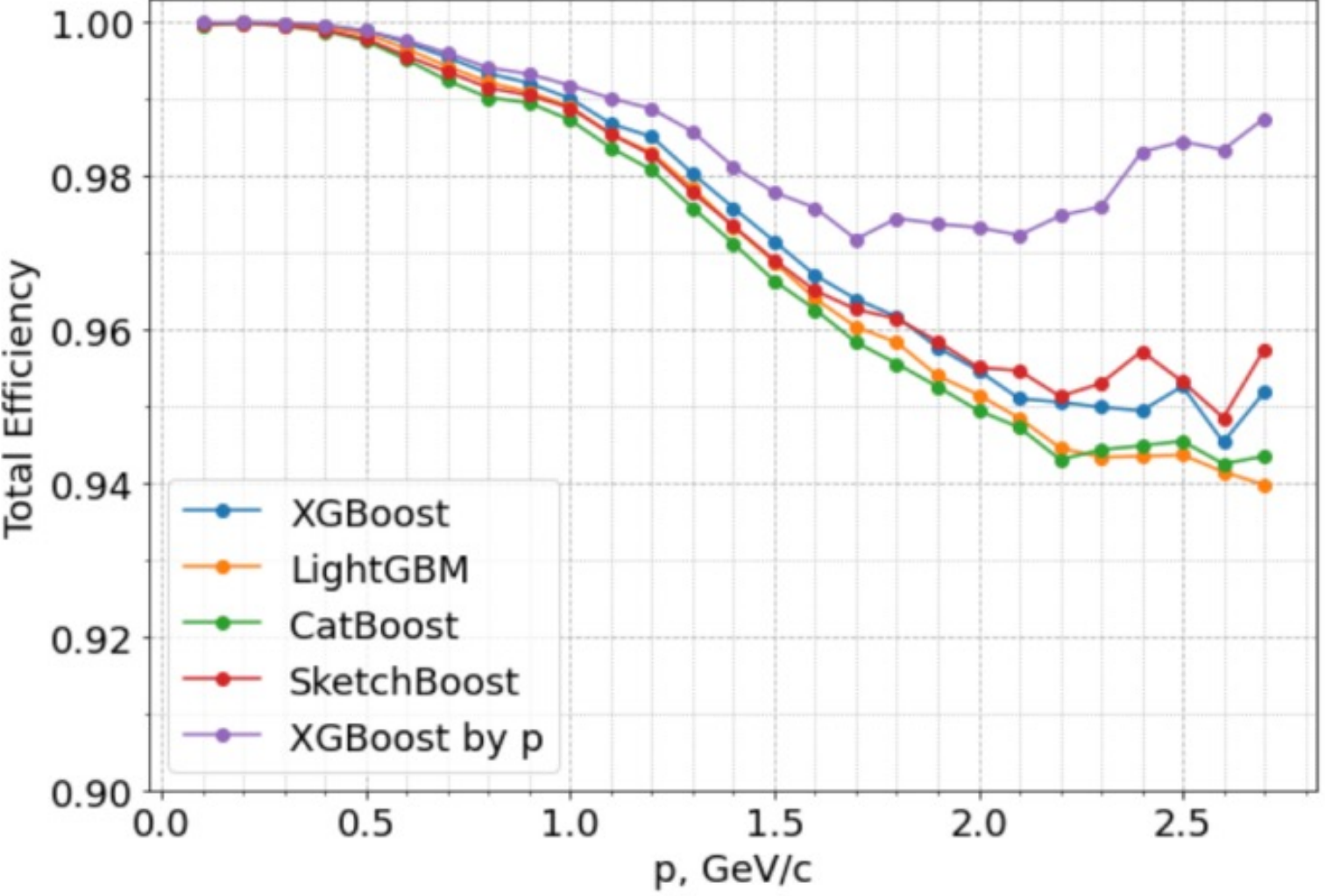


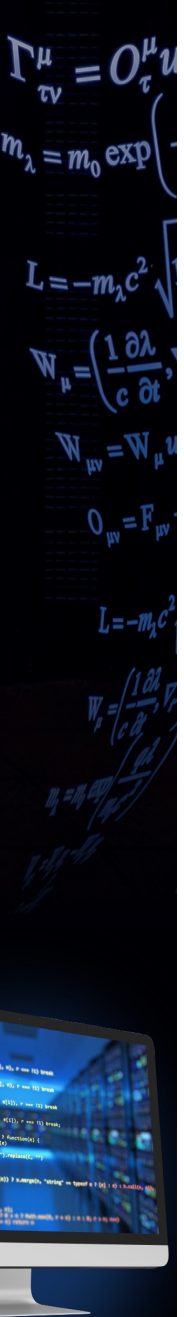
momentum >= 2 GeV/c



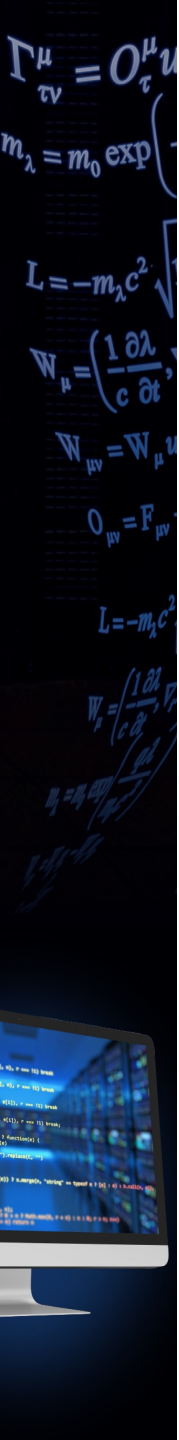


FINAL EFFICIENCY OF XGBOOST

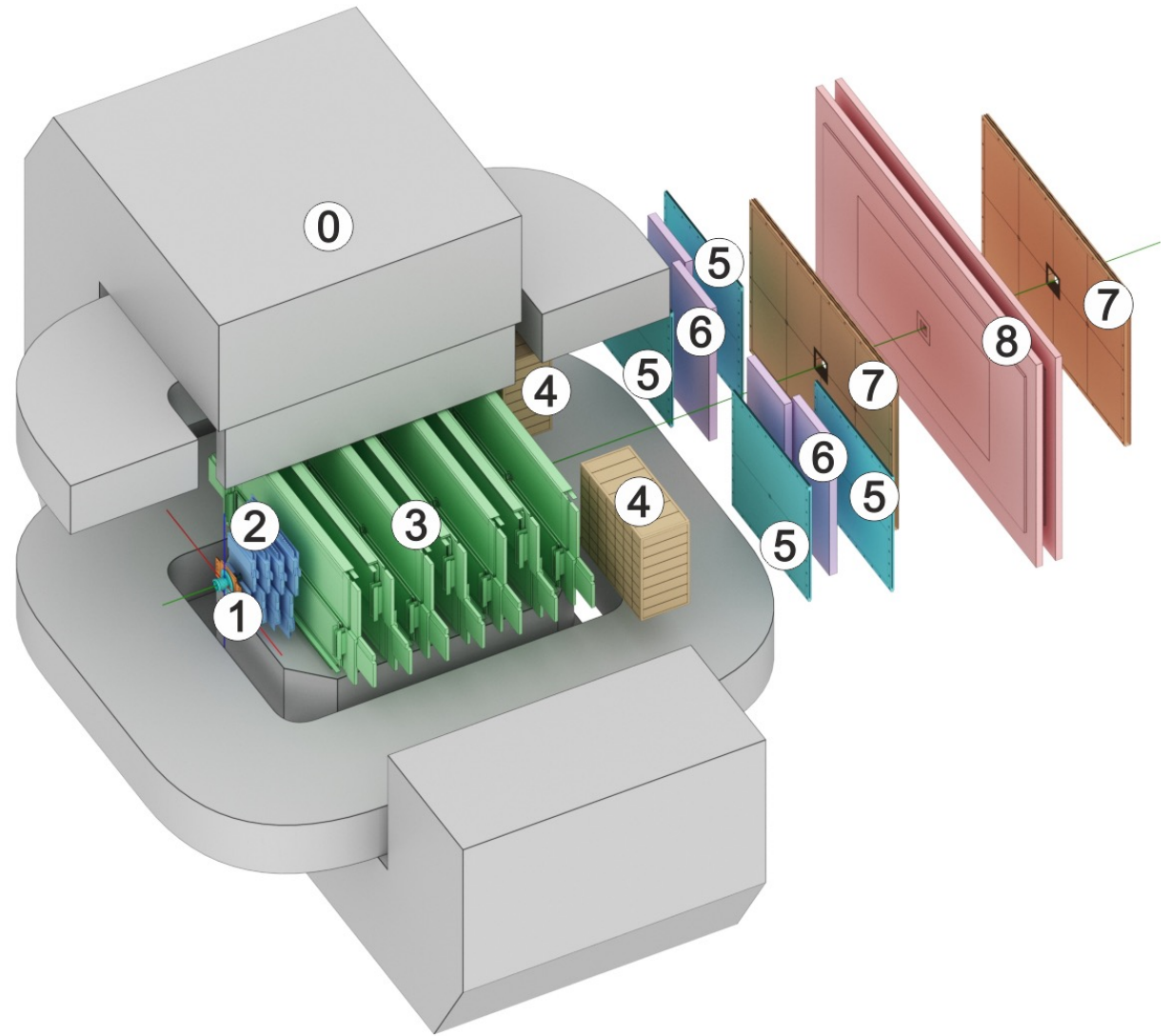




Baryonic Matter at Nuclotron



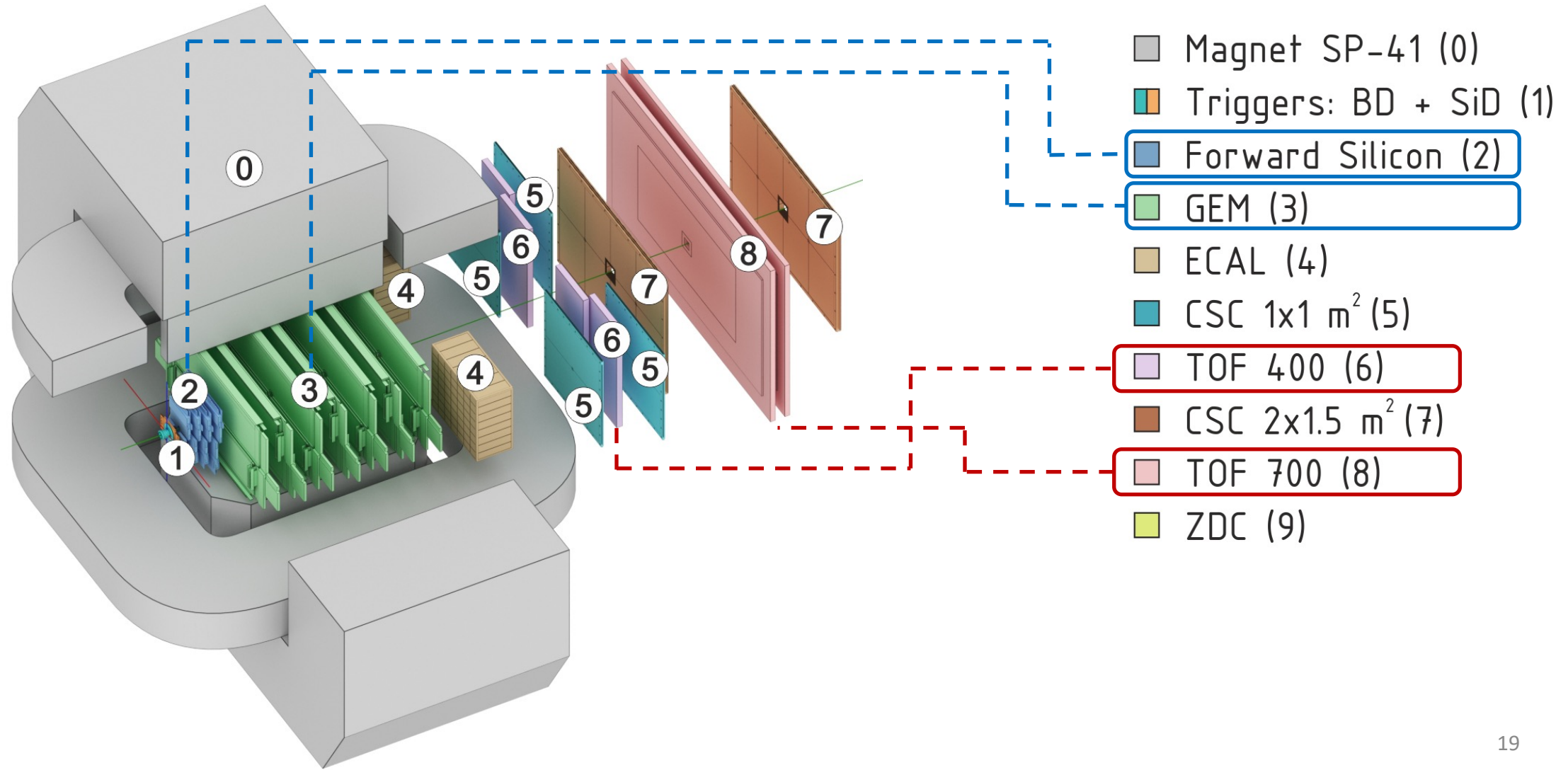
BMN DETECTOR

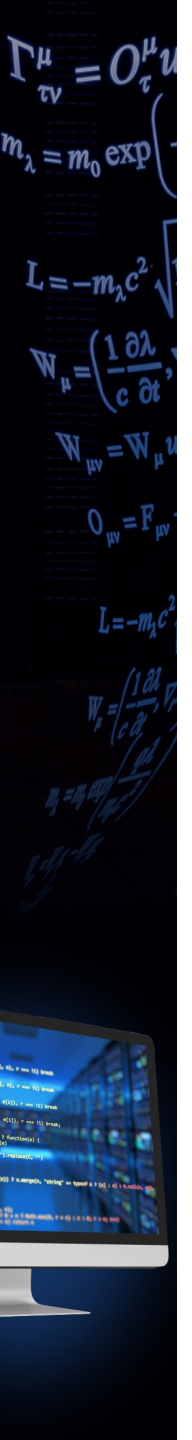


- Magnet SP-41 (0)
- Triggers: BD + SiD (1)
- Forward Silicon (2)
- GEM (3)
- ECAL (4)
- CSC 1x1 m² (5)
- TOF 400 (6)
- CSC 2x1.5 m² (7)
- TOF 700 (8)
- ZDC (9)

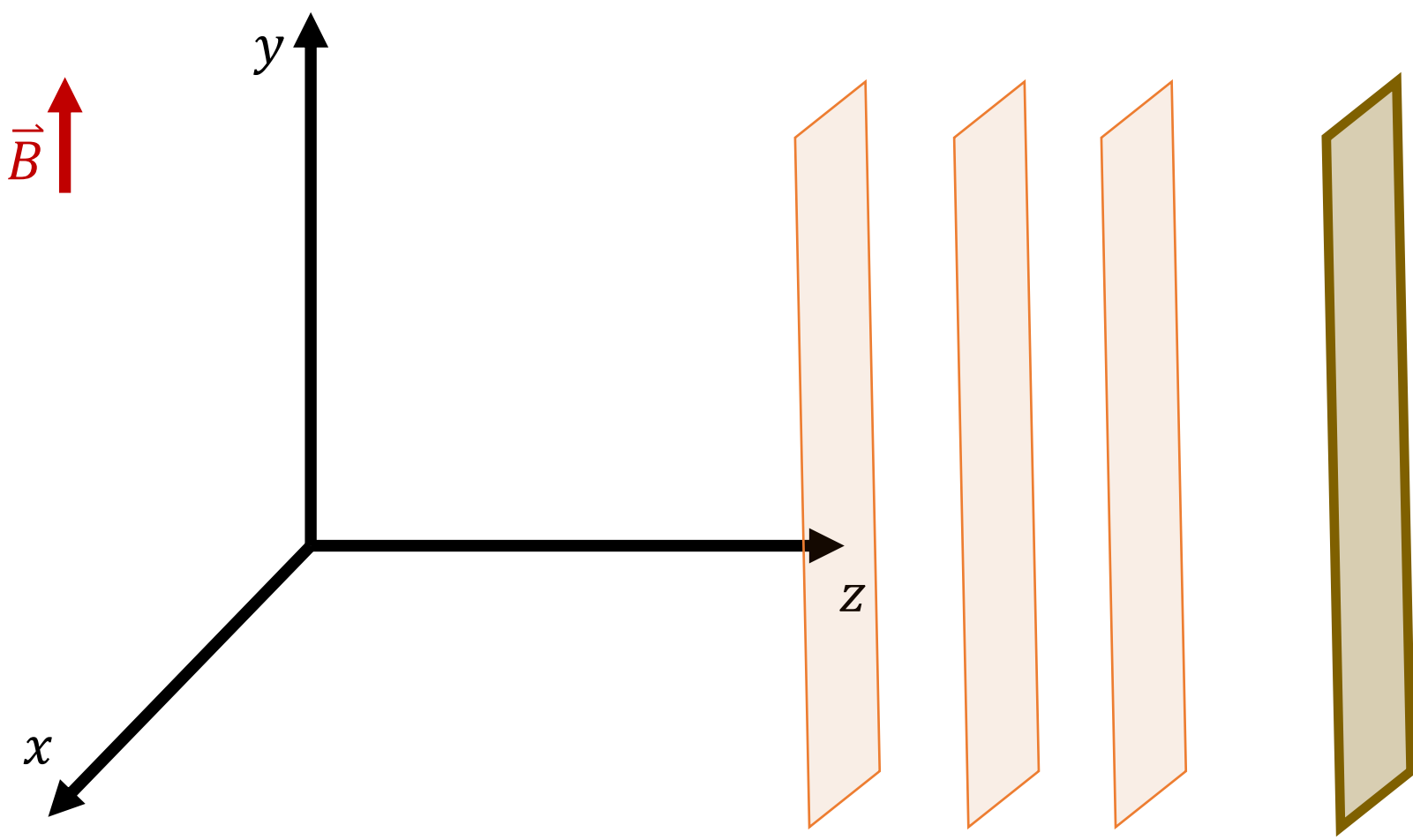
$\Gamma_{\tau\nu}^{\mu} = O_{\tau}^{\mu}$
 $m_{\lambda} = m_0 \exp$
 $L = -m_{\lambda} c^2$
 $W_{\mu} = \left(\frac{1}{c} \frac{\partial \lambda}{\partial t} \right)$
 $W_{\nu} = W_{\mu}$
 $O_{\nu} = F_{\nu}$
 $L = -m_{\lambda} c^2$
 $W_{\mu} = \left(\frac{1}{c} \frac{\partial \lambda}{\partial t} \right)$

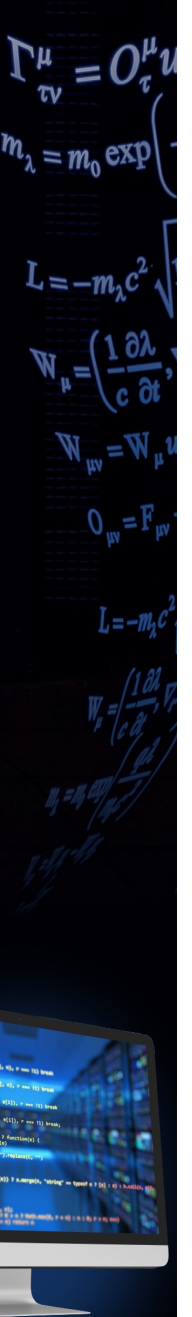
BMN DETECTOR



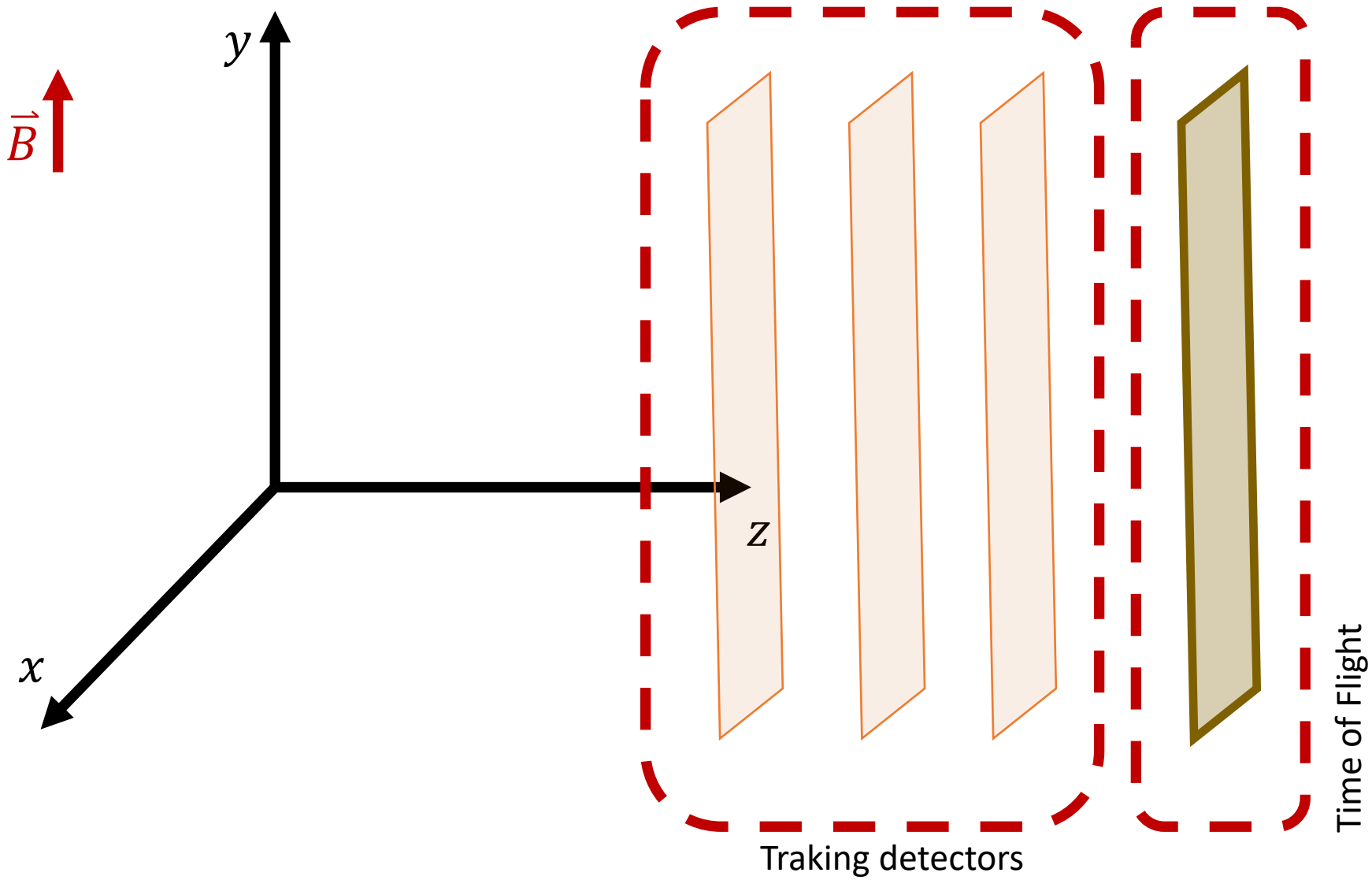


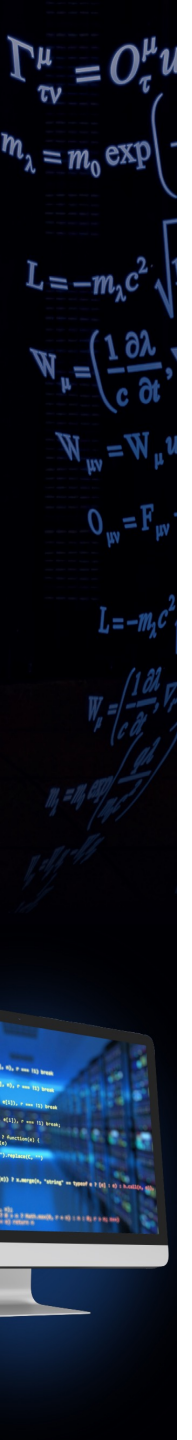
PID IN BMN



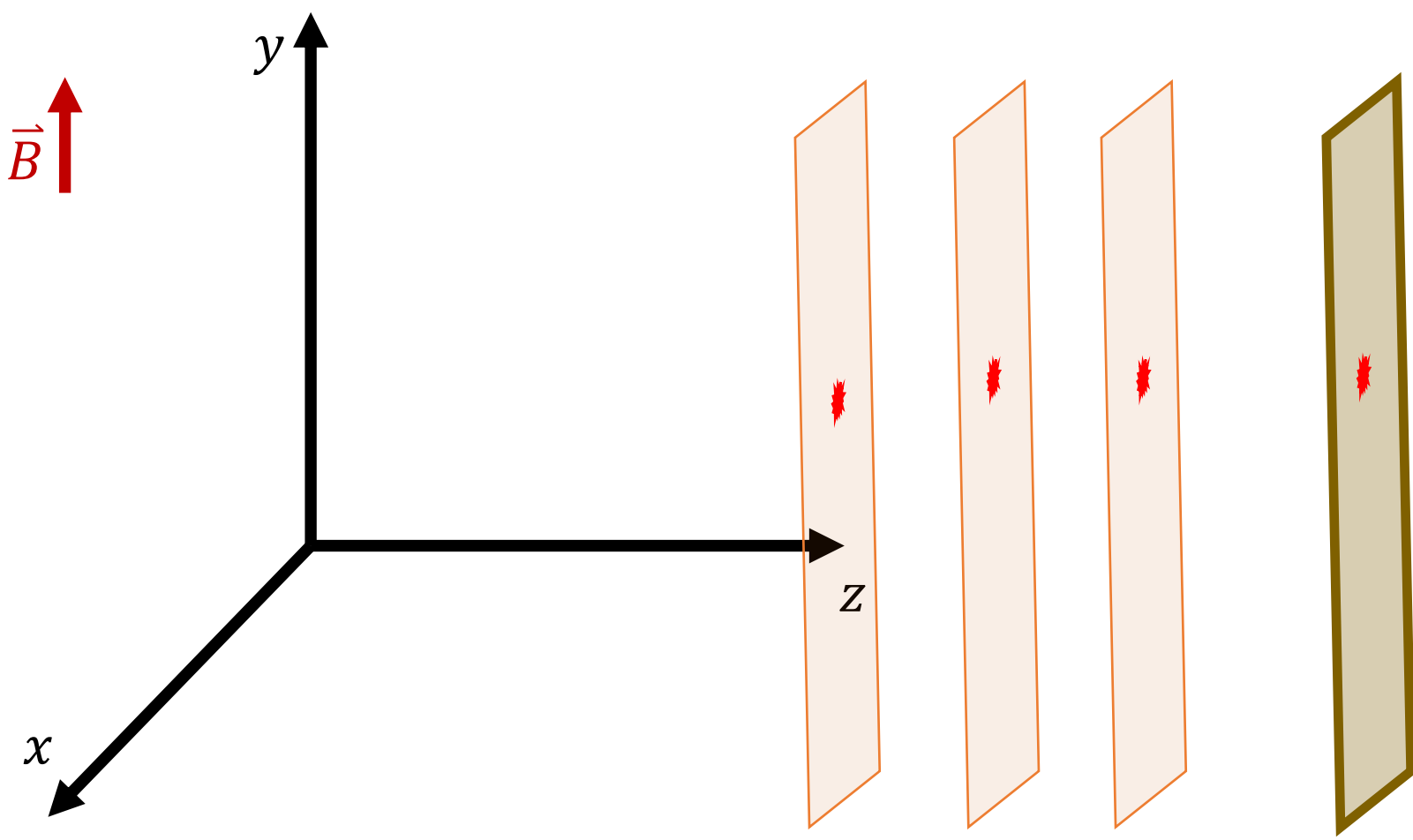


PID IN BMN



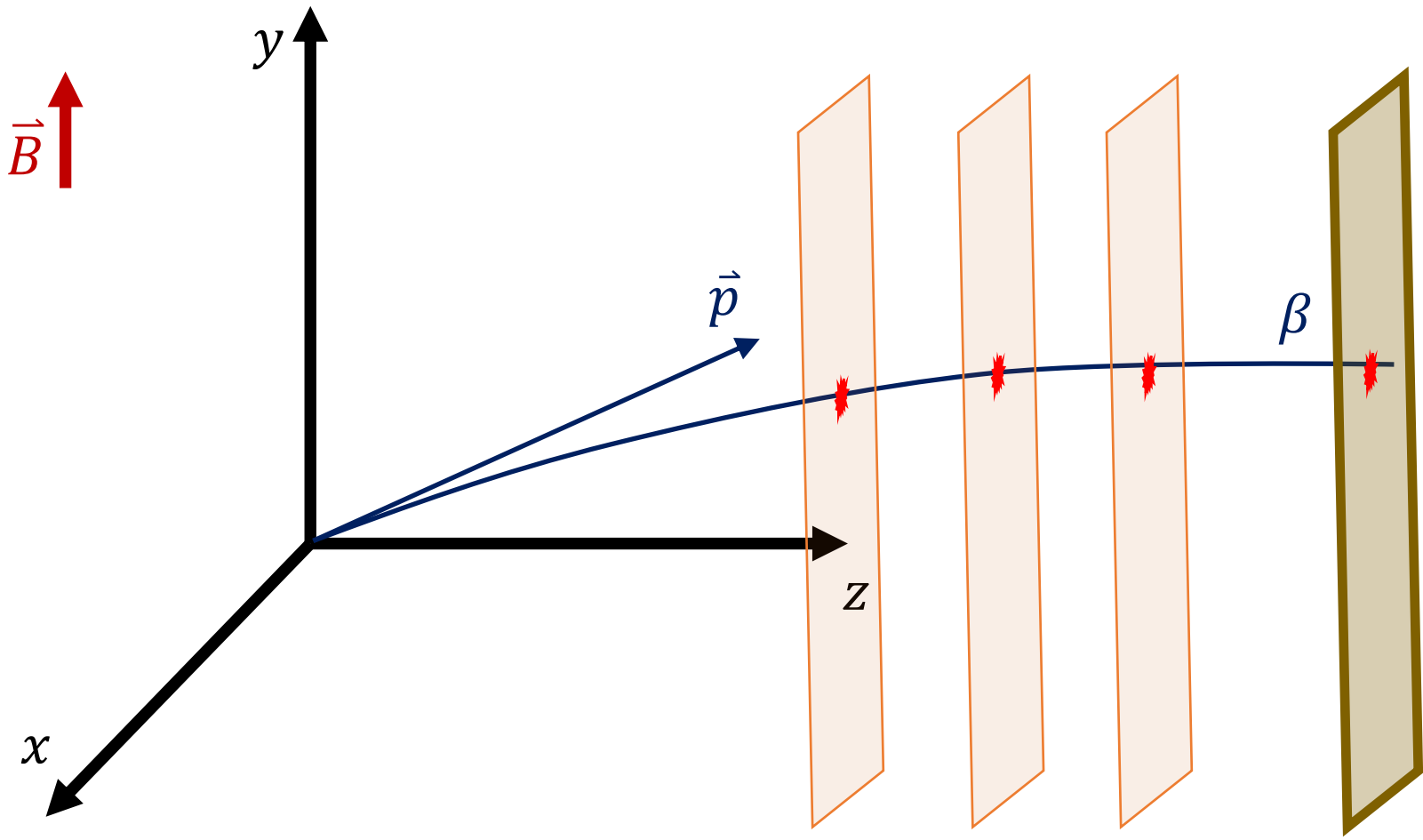


PID IN BMN



$\Gamma_{\tau\nu}^{\mu} = O_{\tau}^{\mu}$
 $m_{\lambda} = m_0 \exp$
 $L = -m_{\lambda} c^2$
 $W_{\mu} = \left(\frac{1}{c} \frac{\partial \lambda}{\partial t} \right)$
 $W_{\nu} = W_{\mu}$
 $O_{\nu} = F_{\nu}$
 $L = -m_{\lambda} c^2$
 $W_{\mu} = \left(\frac{1}{c} \frac{\partial \lambda}{\partial x} \right)$

PID IN BMN



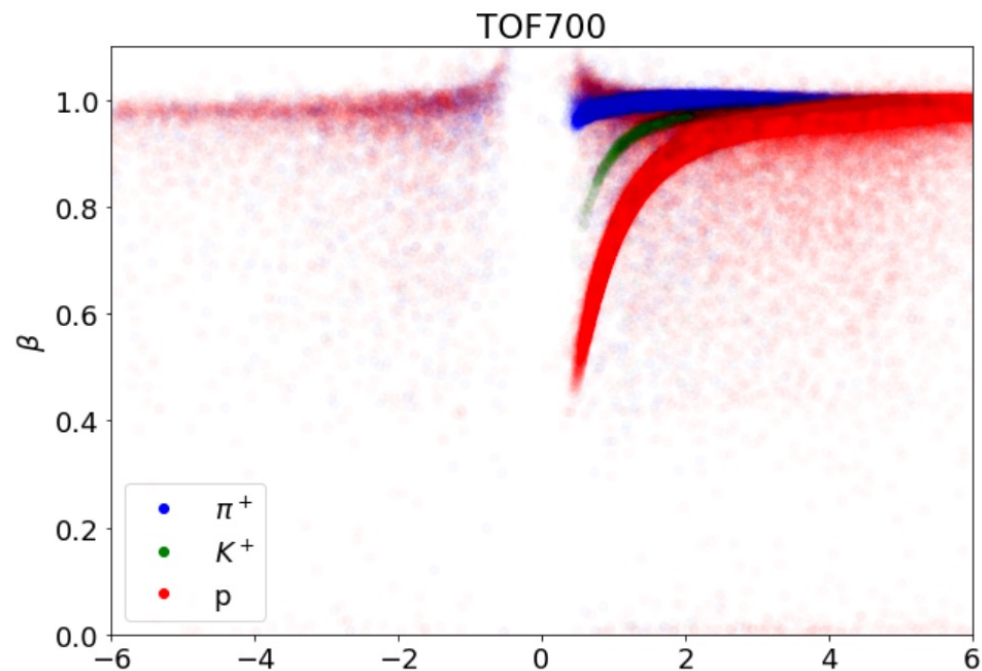
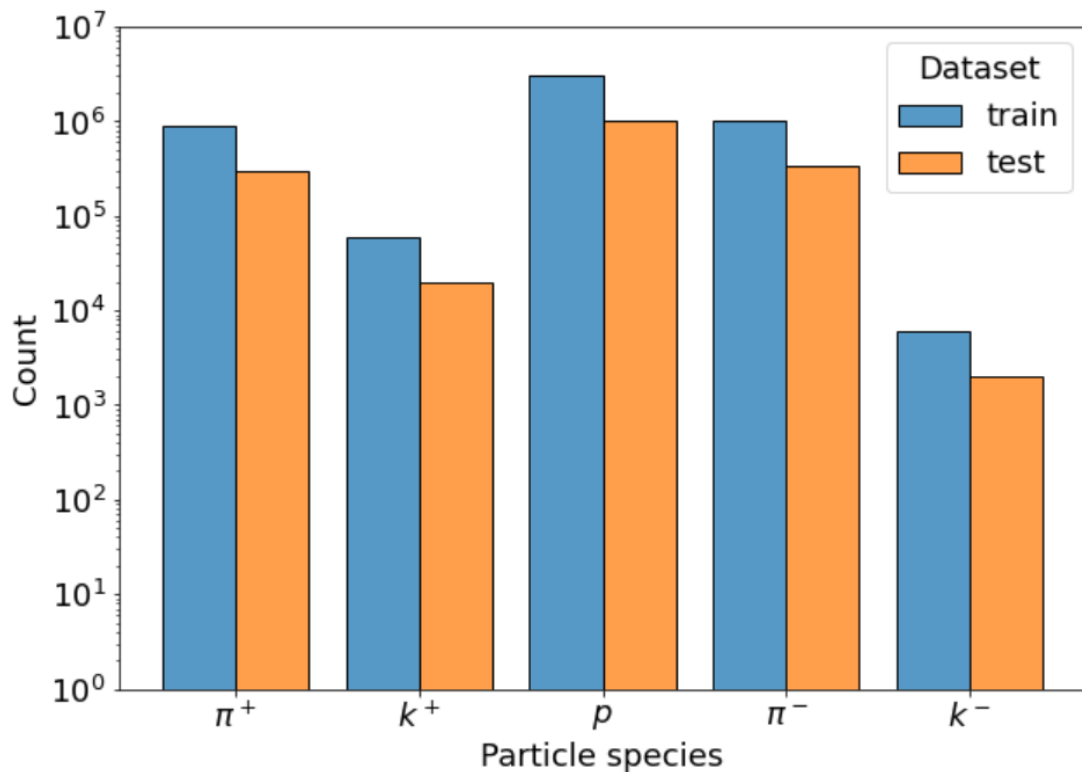
$$\beta = \frac{p}{\sqrt{p^2 + m^2}}$$

$$m^2 = \frac{p^2}{\beta^2} - p^2$$

Название частицы	Символ		Заряд, ед. e	Масса покоя, ед. m _e	Масса, МэВ
	частицы	анти-частицы			
Пионы	π ⁰	π ⁻	0	264,1	135
	π ⁺		1	273,1	140
Каоны	K ⁰	\tilde{K}^0	0	974,0	498
	K ⁺	K ⁻	1	966,2	494
Протон	p	\bar{p}	1	1836,2	938
Нейтрон	n	\bar{n}	0	1838,7	940

DATASET

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

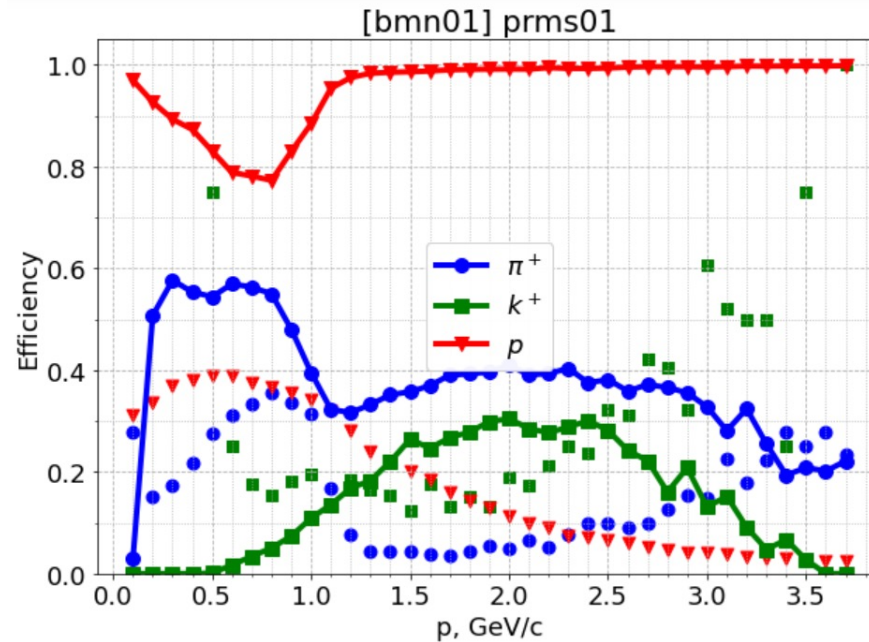
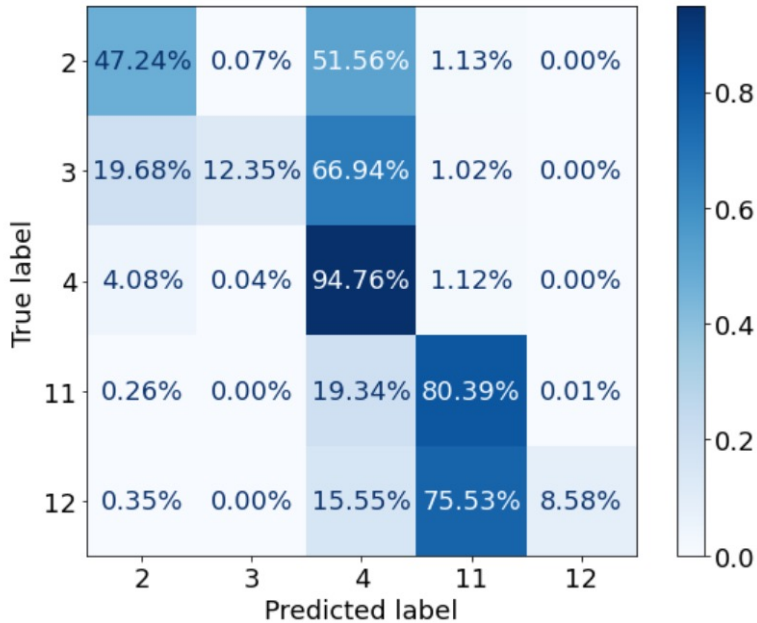


RESULTS

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

XGBoost shows identification efficiency more than 80%!

0.8222584966783286



HOW?!

RESULTS

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

XGBoost shows identification efficiency more than 80%!

HOW?!



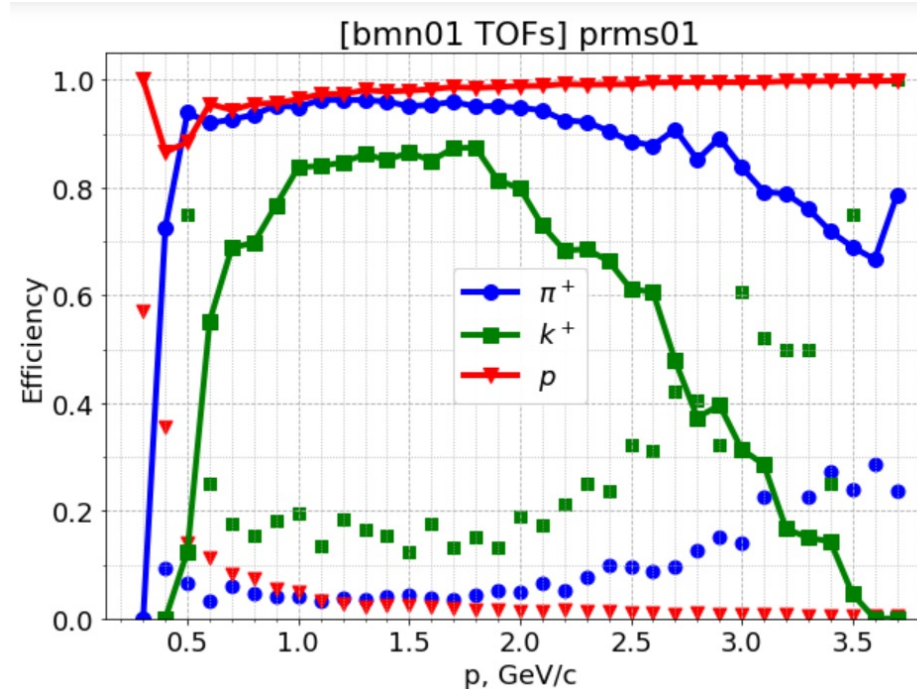
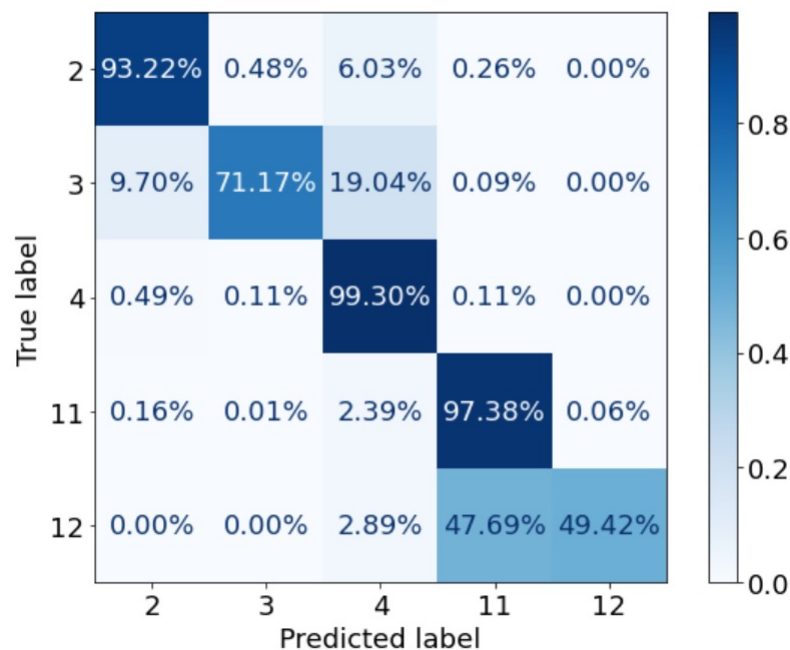
Random efficiency: 80% minus 27% is approx 53%

RESULTS

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

XGBoost shows 98.3% efficiency for tracks with ToF!

0.9828742299942589



СПАСИБО ЗА ВНИМАНИЕ!

Тема: Исследование и разработка методов и подходов применения методов машинного обучения в задачах теоретической физики (ТФ) и физики высоких энергий (ФВЭ)

Руководители: к.ф.-м.н. Айриян А.С., к.ф.-м.н. Григорян О.А.

Аннотация

В исследованиях в рамках ТФ и ФВЭ возникают проблемы, которые плохо формализуемые, либо их формальная математическая постановка требует привлечения сложного (а возможно еще не существующего) математического аппарата для их решения. В таких случаях может быть полезно применение методов машинного обучения. Планируется исследовать возможность использования машинного обучения при решении прямых и обратных задач для нелинейных уравнений, описывающих исследуемые физические процессы. Такой подход должен быть общим, т.е. независимым от физической сути решаемой проблемы, и эффективным, чтобы получить реальное применение на практике. Основной проблемой в данном направлении исследований является постановка задачи с точки зрения машинного обучения, а также формирование выборки, обучение на которой позволит решать поставленную задачу.

Что приобретет студент: практику решения актуальных научных задач, повышение квалификации в области машинного обучения, дипломную работу, обладающую научной новизной и актуальностью.

Возможные темы дипломных работ

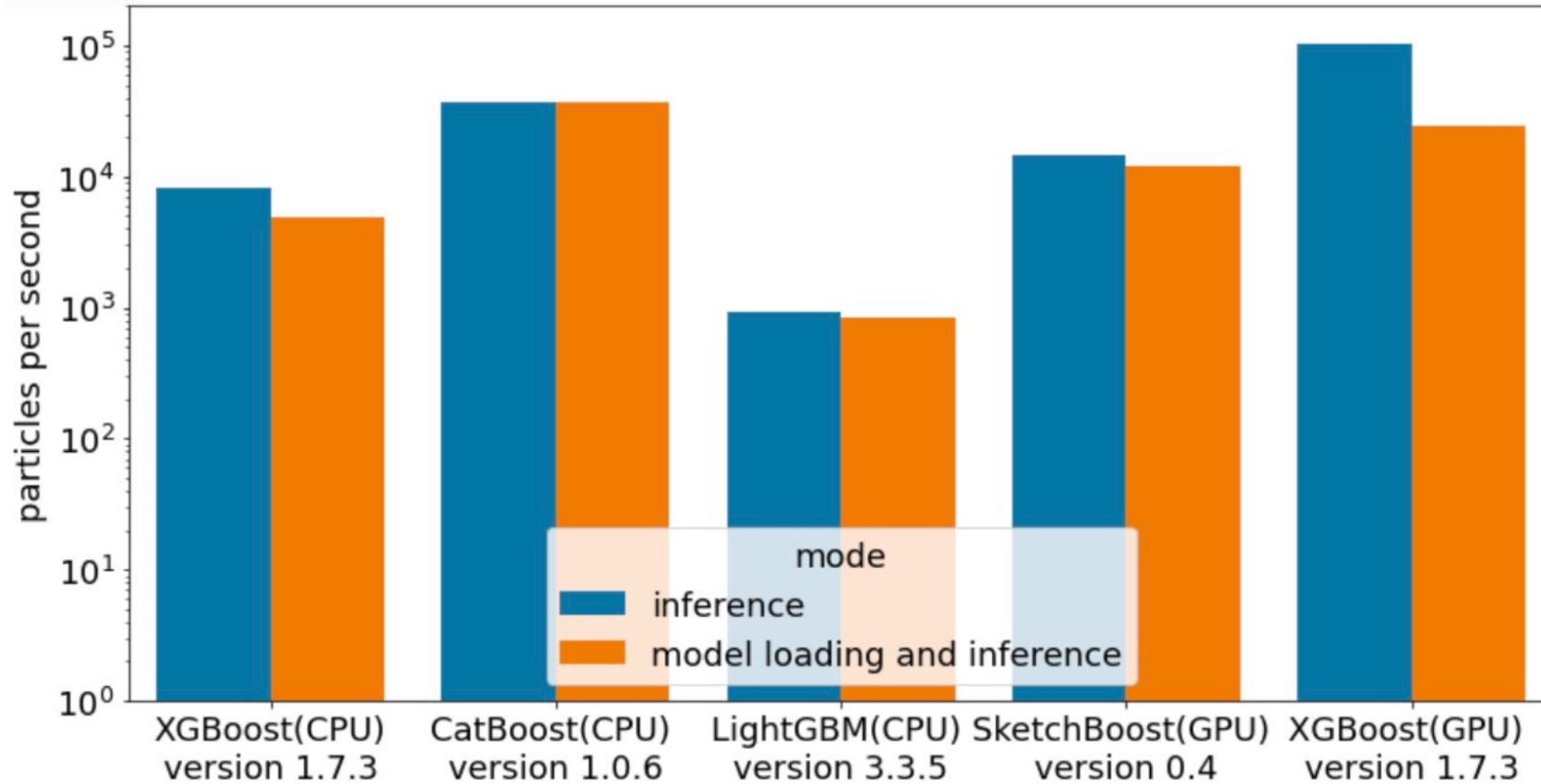
1. Решение обратной задачи Толмана-Оппегеймера-Волкова с применением глубоких нейронных сетей;
2. Нейросетевой подход к прямому и обратному вейвлет-преобразованию;
3. Деревья решений для распознавания элементарных частиц по данным детекторов физики высоких энергий.

Общие требования к студентам

- Знание математических основ дифференциальных уравнений.
- Знание основ машинного обучения.
- Элементарное владение Python, желательно элементарное владение библиотеками NumPy, TensorFlow, Keras, Pandas, Pytorch.



COMPARATIVE ANALYSIS OF THE ALGORITHMS. TIMING



GPU: Nvidia Tesla V100-SXM2 NVLink 32GB HBM2

CPU: Intel Xeon Gold 6148 CPU @ 2.40 GHz 20 Cores / 40 Threads