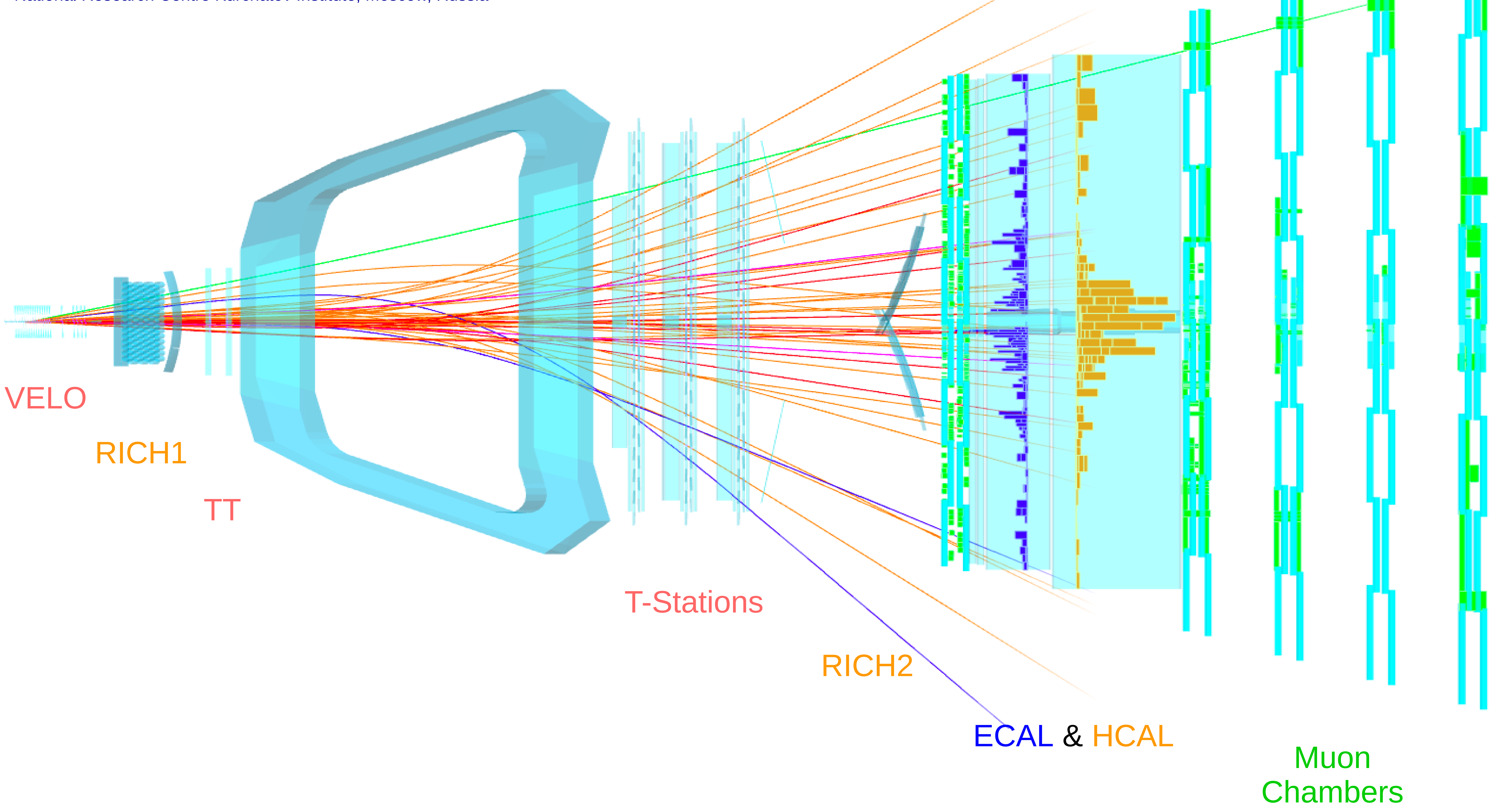
Machine Learning Based Global Particle Identification at LHCb

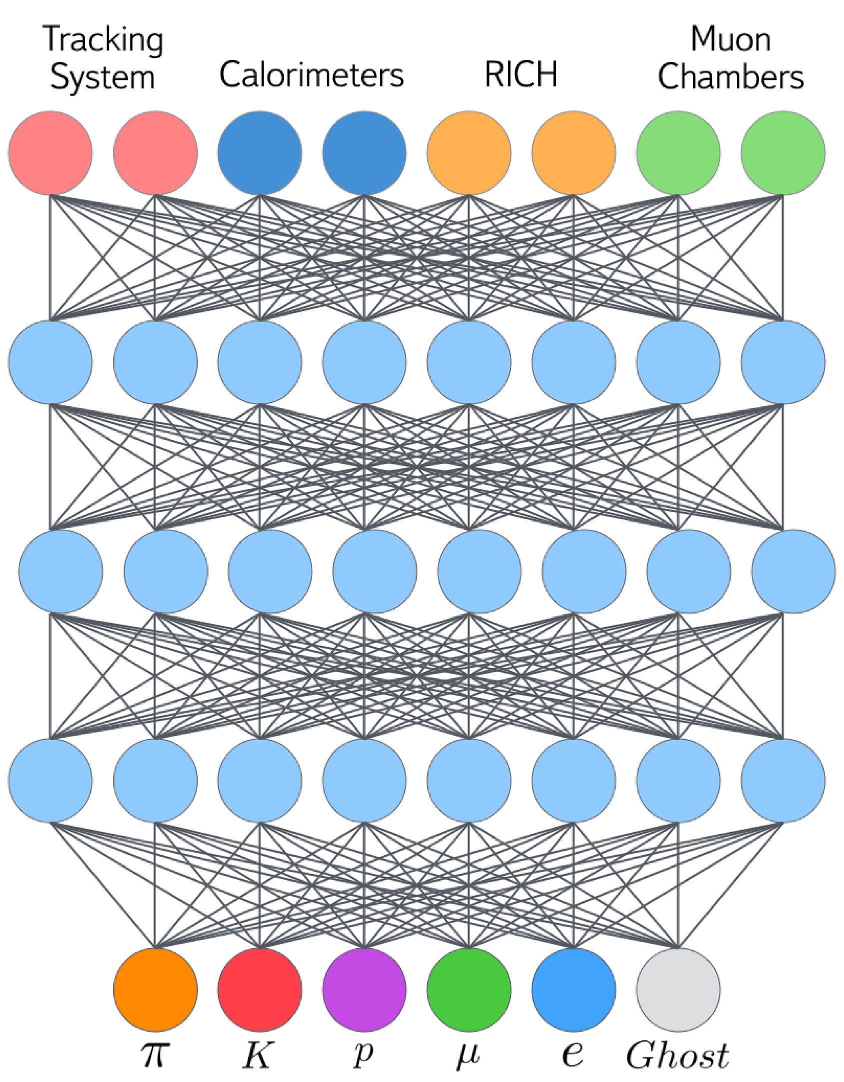
Denis Derkach[1,2], Mikhail Hushchyn[1,3], Tatiana Likhomanenko[1,2,4], Alex Rogozhnikov[1,2], Nikita Kazeev[1,2], Victoria Chekalina[1], Radoslav Neychev[1,3], Stanislav Kirillov[1] , Fedor Ratnikov[1,2] on behalf of the LHCb collaboration

18th International Workshop on Advanced Computing and Analysis Techniques in Physics Research, Seattle, August 21-25, 2017

[1] Yandex School of Data Analysis, Moscow, Russia
[2] Higher School of Economics, Moscow, Russia
[3] Moscow Institute of Physics and Technology, Moscow, Russia
[4] National Research Centre Kurchatov Institute, Moscow, Russia



Best-Efficiency Models



The problem is to identify the charged particle type which a given track is associated. There are five particle types: electron, muon, pion, kaon, proton plus the "not classifiable" ghost for a total of 6 hypotheses.
This problem can be considered as a multiclass classification problem. For PID infromation from Ring-Imaging Cherenkov Detector (RICH), Electromagnetic Calorimeter, Hadronic Calorimeter and Muon Chambers sub-detectors and track reconstruction is used.

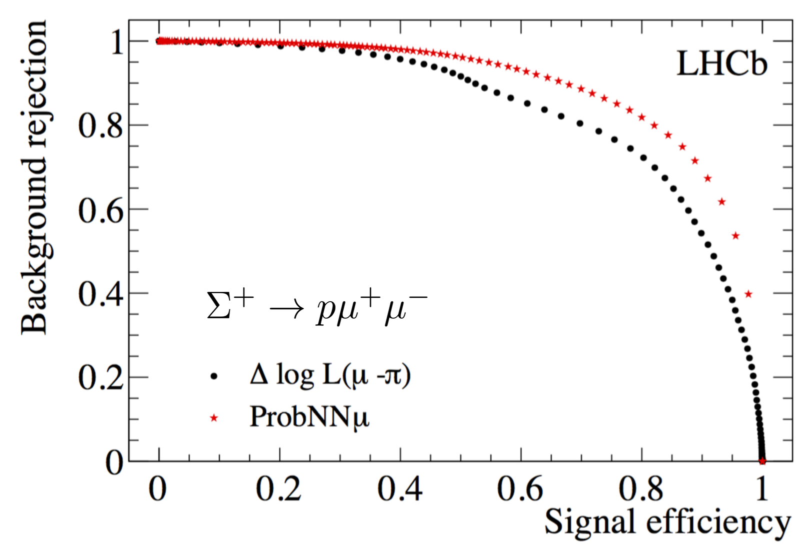
Current particle ID approaches:

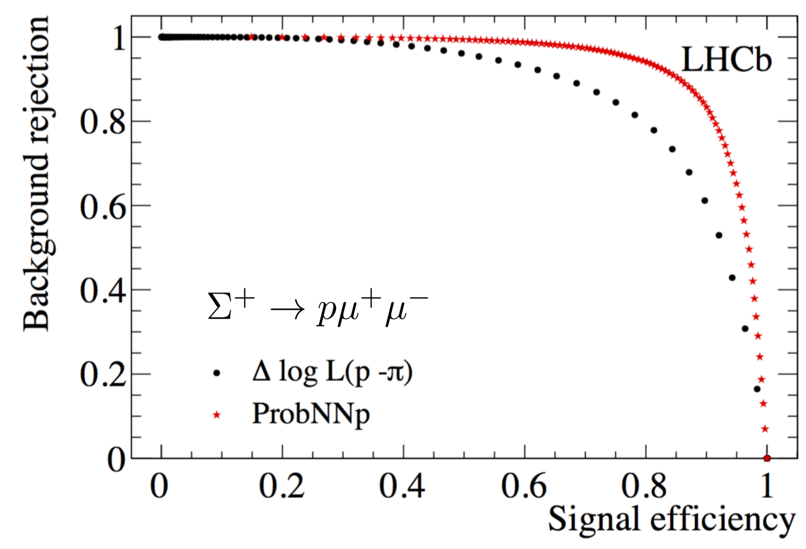
- **Δ log L**: Estimate likelihood of a particle type based on a subdetector response. Likelihoodod the subdetectors are combined into global likelihood of the particle type.
- **ProbNN** (baseline): The subdetector responses are combined using one hidden layer neural network (TMVA MLP) in one-particle-versus-rest mode.

The PID was improved using gradient boosting algorithms (**XGBoost**, **CatBoost**) and deep neural network (**deep NN**) in multiclassification mode.

$(1\text{-AUC})/(1\text{-AUC}_{baseline})$ LHCb Simulation, preliminary

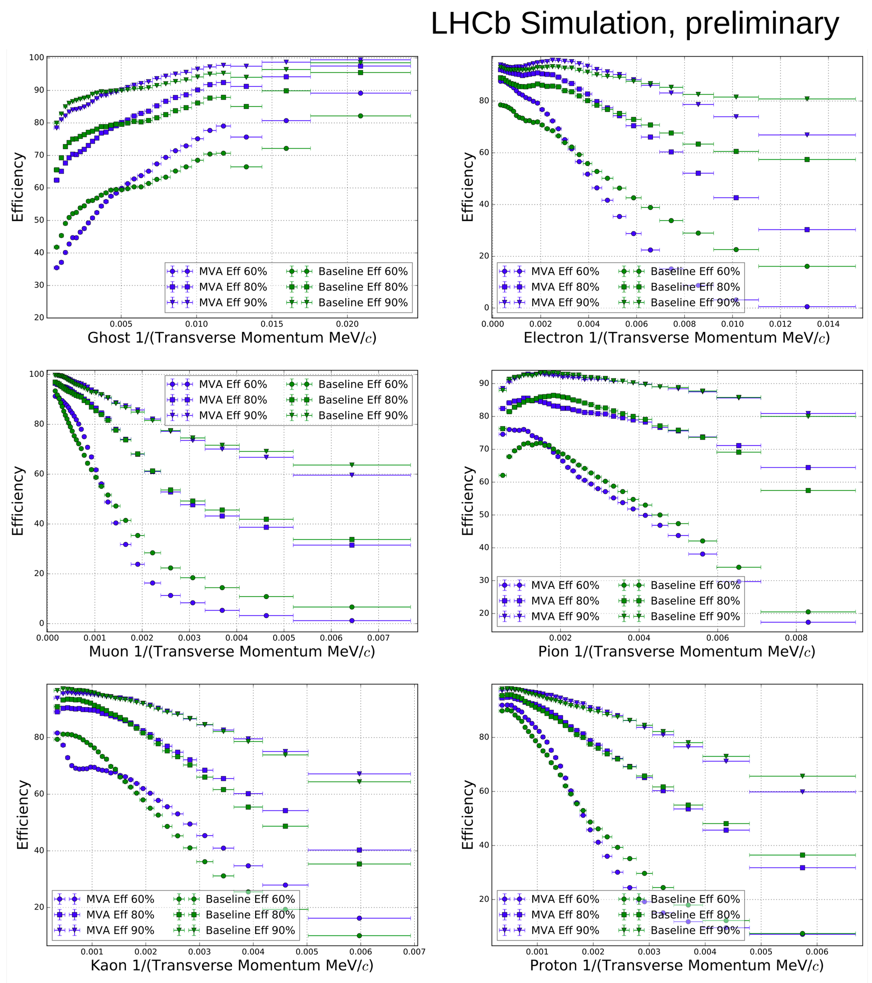
	Ghost	Electron	Muon	Pion	Kaon	Proton
baseline	1	1	1	1	1	1
deep NN	-29 %	-41 %	-52 %	-37 %	-20 %	-17 %
XGBoost	-24 %	-37 %	-50 %	-34 %	-18 %	-15 %
CatBoost	-30 %	-43 %	-54 %	-37 %	-20 %	-18 %





Best-Flatness Models

LHCb Simulation, preliminary



Non-flat deep NN model (blue)

The PID information strongly depends on the kinematic variables. This relationship leads to strong dependency between PID efficiency and kinematic variables as it is shown in the left figure .In practice it could be helpful to have flat PID efficiencies along chosen control variables to reduce systematics effects.
To provide uniformity along some observables models were trained using the modified loss funstion:

$$\mathcal{L} = \mathcal{L}_{ExpLoss} + \alpha \mathcal{L}_{FL}$$

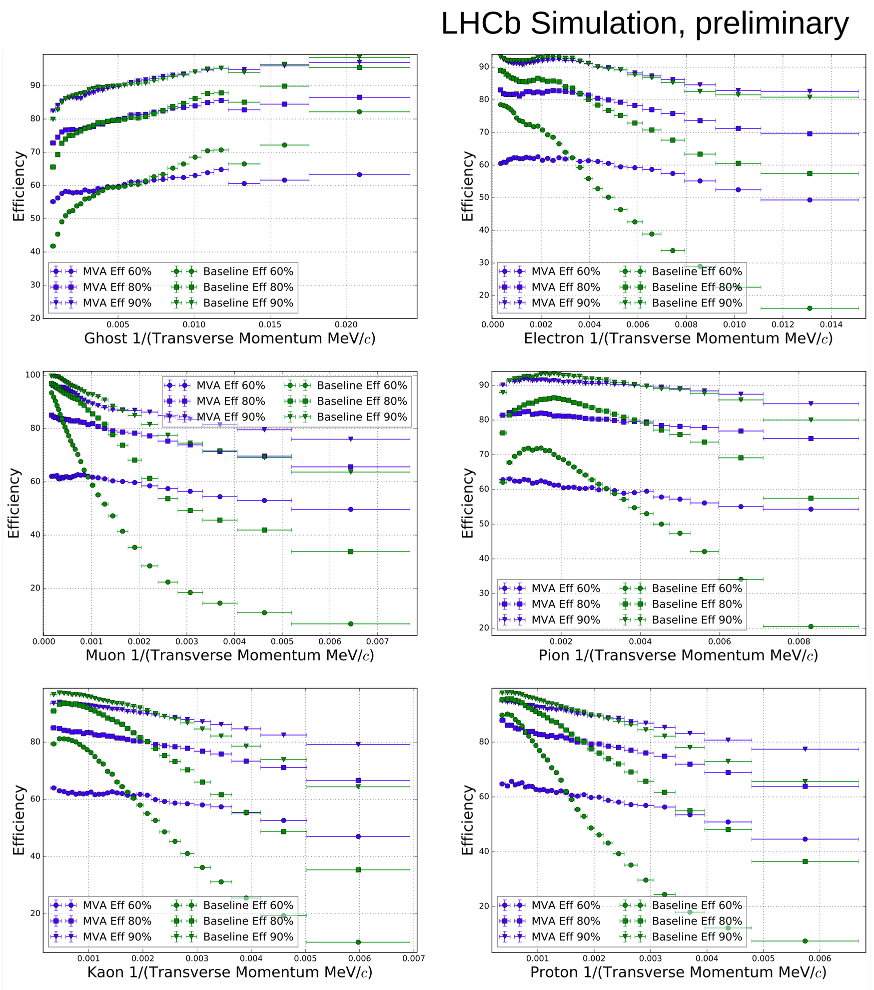
where the fisrt term is the classification loss function and the second term is the the uniformity loss written in differential form of *Cramer-von Mises* measure. The right figure demonstrates the flatness improvement.

$(1\text{-AUC})/(1\text{-AUC}_{baseline})$ LHCb Simulation, preliminary

	Ghost	Electron	Muon	Pion	Kaon	Proton
baseline	1	1	1	1	1	1
p + p_T flatness	-23 %	-20 %	-27 %	-26 %	+2 %	+5 %
2d(p, p_T) flatness	-21 %	-9 %	-13 %	-23 %	+12 %	+23 %
p + p_T + η + nTracks flatness	-22 %	-13 %	-26 %	-24 %	+2 %	+6%
4d(p, p_T, η, nTracks) flatness	-21 %	-4 %	-13 %	-20 %	+10 %	+25 %

There is a trade off between the PID models efficiency and flatness. It is possible to archine any efficiency flaness of a model along any variable. However, the efficiency decreases with increasing its flatness.

LHCb Simulation, preliminary



Flat 4d model (blue)

SCHOOL OF DATA ANALYSIS

NATIONAL RESEARCH UNIVERSITY

MIPT MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY

NRC "KURCHATOV INSTITUTE"