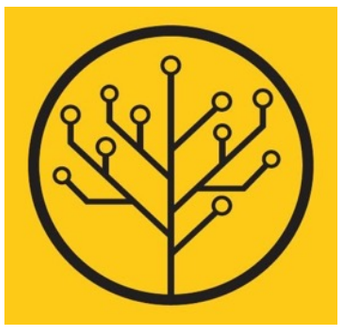
Laboratory of Methods for Big Data Analysis “LAMBDA”

National Research University Higher School of Economics,
Faculty of Computing Science

MPD@NICA Physics WG Meeting

May 25, 2018



Mission

- Solve tough natural science challenges with aid of advanced data analysis and machine learning
- Develop new data analysis methods
- Bridge the gap between CS and HEP communities

Group in HEP

- Member of LHCb Collaboration
  - trigger
  - charged and neutral particles ID
  - monitoring
  - anomalies detection
  - computing resources
- Member of SHIP Collaboration
  - detector optimisation
  - computing resources
- Active participation in CRAYFIS
- Cooperating with CMS Collaboration
  - data certification
- Particular projects with Atlas, Opera

Group PhD

- Fedor Ratnikov - physicists
  - ARGUS - HERA-B - CDF - CMS - LHCb - SHiP
  - triggers, data handling, calo reco, offline reco, simulation
  - heavy flavour, tau-physics, searches: rare decays, exotica, SUSY
- Andrey Ustyuzhanin (head of the lab) - computer scientist
  - LHCb - Opera - SHiP - CRAYFIS
  - applying different aspects of ML to different problems in science
- Denis Derkach - physicist
  - Alice - NA61 - BaBar - LHCb
  - heavy flavour, SM combination
- Andrey Sapronov - physicist (JINR PhD)
  - ATLAS
  - numerical computation of higher order corrections

PID at LHCb

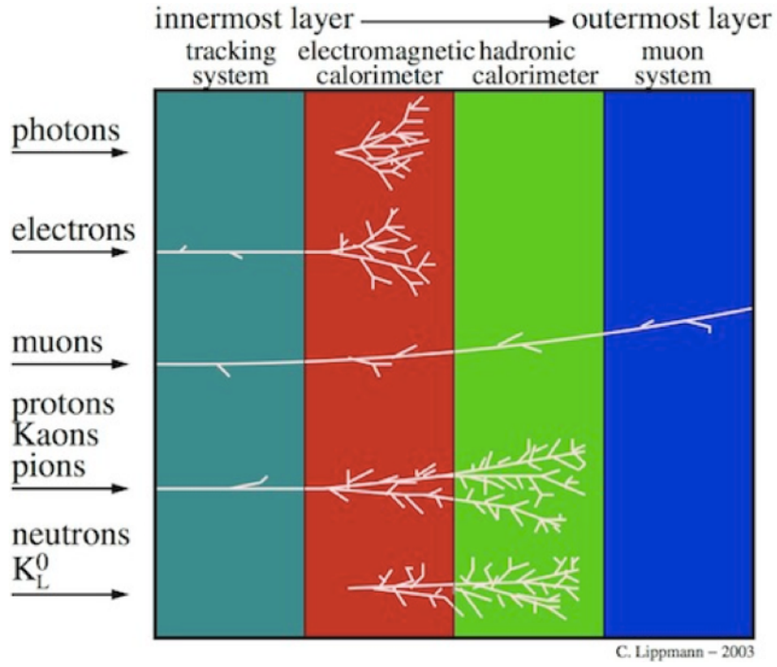
Problem: identify particle type associated with a track/energy deposited in the subdetectors

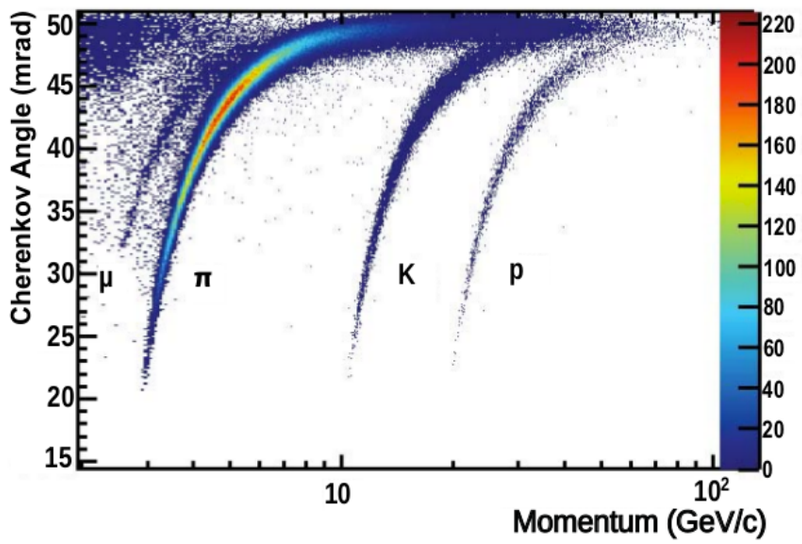
- Charged: π, e, μ, K, p
- Neutral: π^0, γ, n

Better PID performance → better bkg rejection → more precise results.

PID also used for trigger (in particular for upgrade): less background → less resources (less bandwidth)

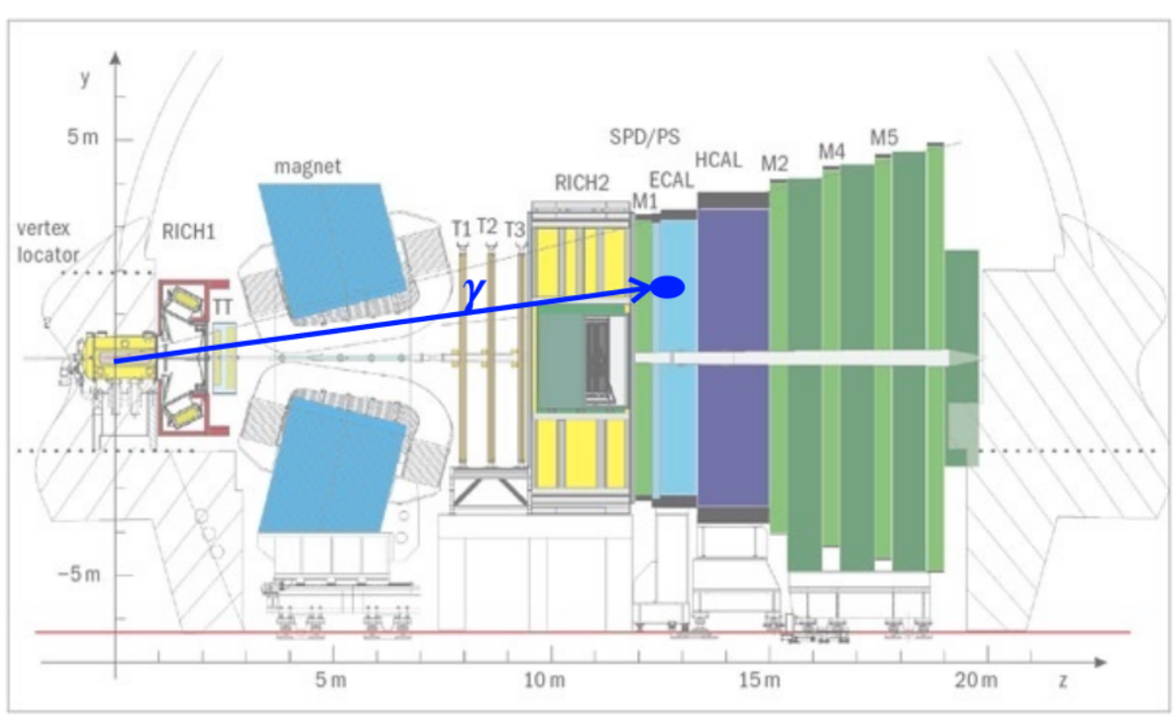
High-level info from subdetectors + track quality info → multi-class classification in machine learning

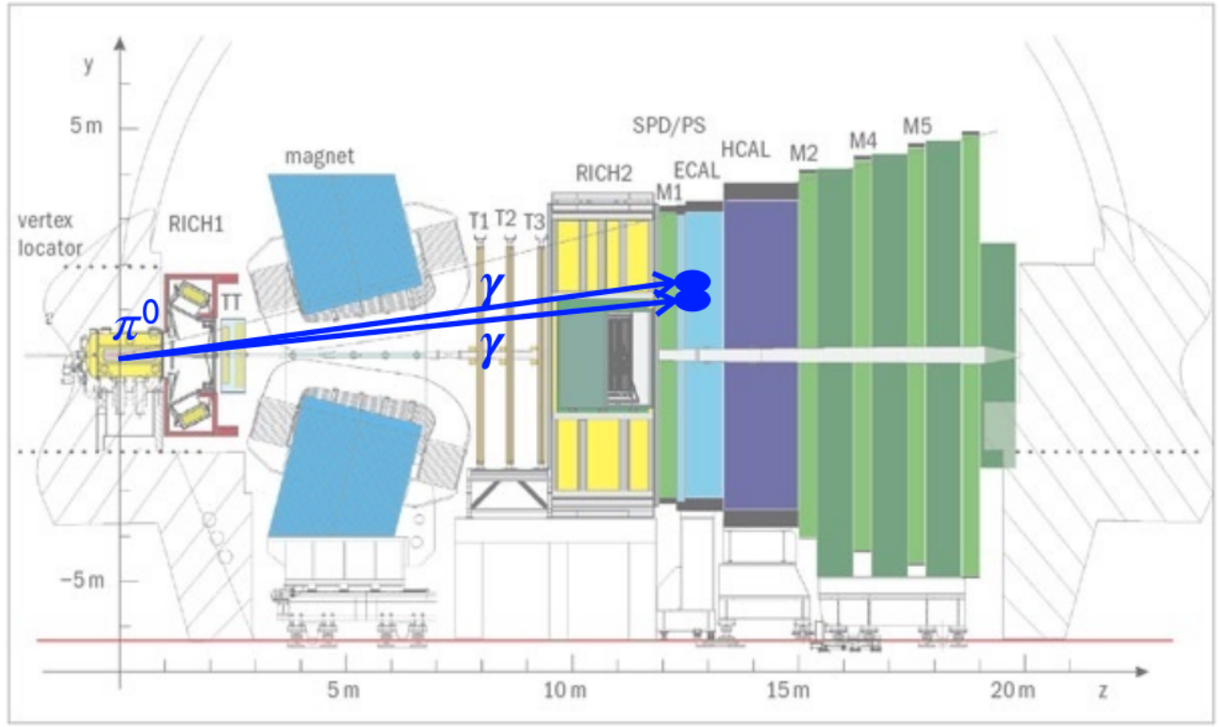




Neutral PID

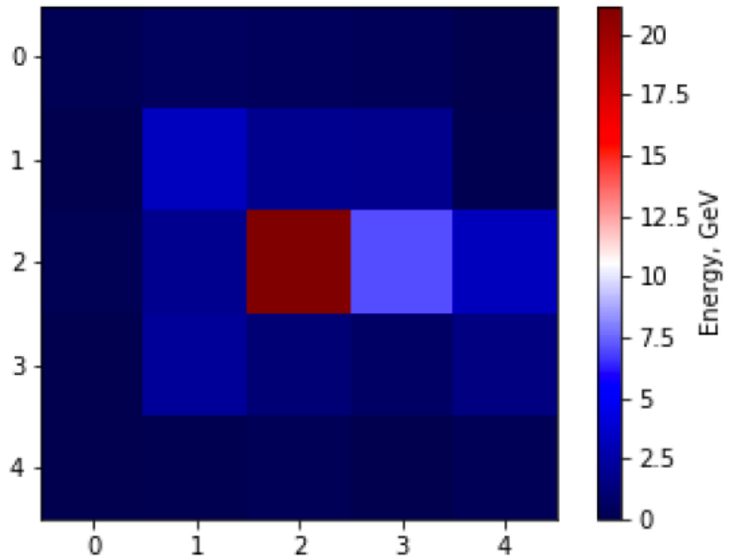
π^0 copiously produced at LHCb , decay to $\gamma\gamma$
high momentum $\pi^0 \rightarrow$ merge of ECAL clusters $\rightarrow$ huge background for radiative decays
Need for a powerful tool to discriminate signal (γ) from background $\pi^0 \rightarrow \gamma\gamma$



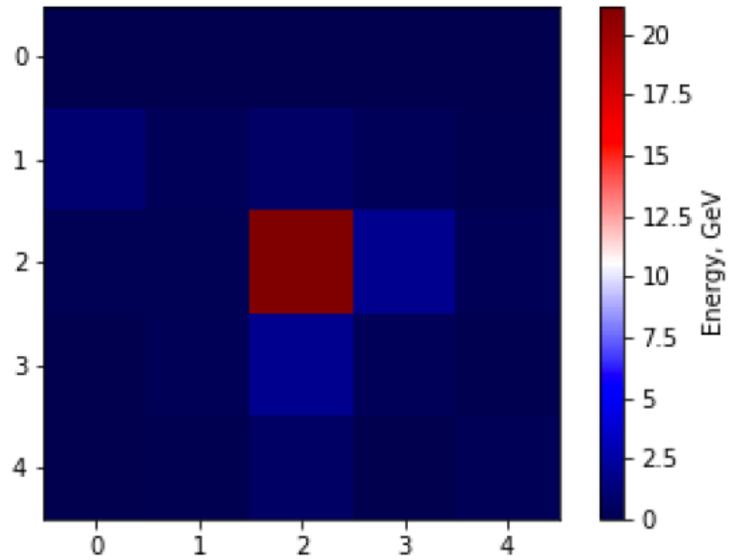


ECAL Signatures

$\pi^0 \to \gamma\gamma$



γ



ECAL clusters (3x3 cells)

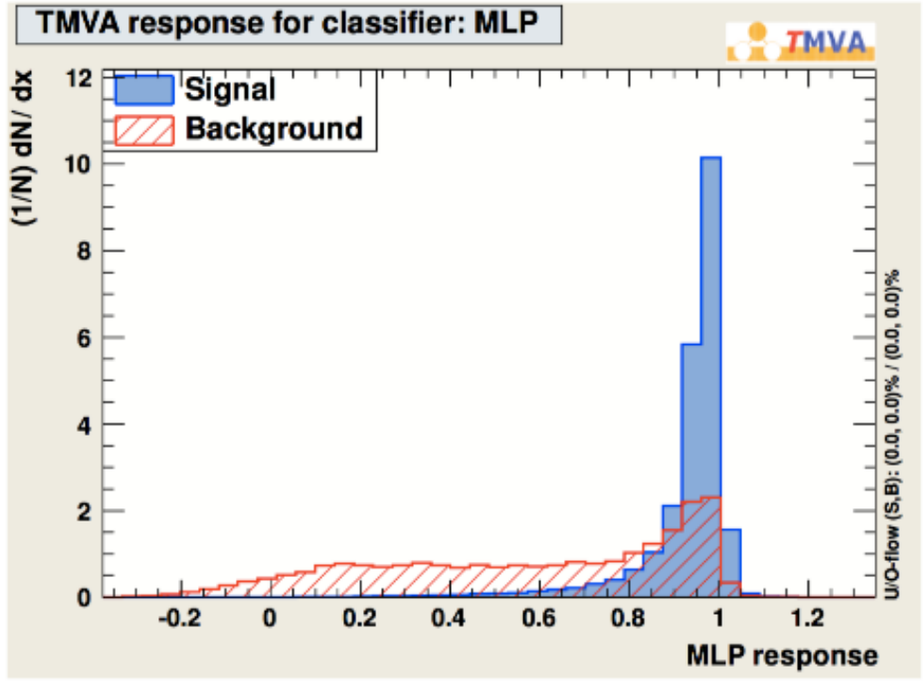
Coarse granularity → separation is not straightforward

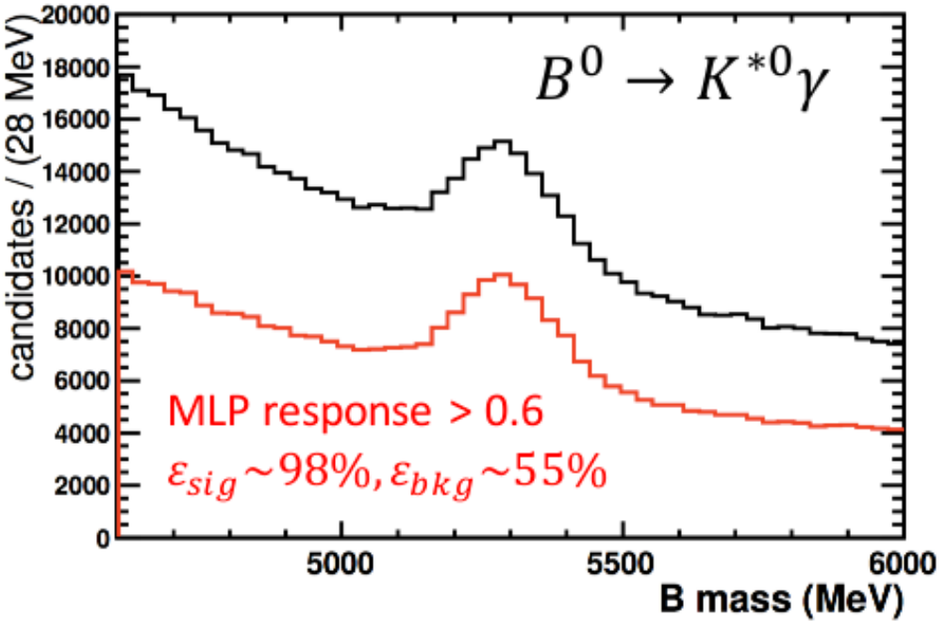
Baseline approach [LHCb-PUB-2015-016]

Neural Network with 2 hidden layers (TMVA MLP)

14 ECAL and Pre-Shower cluster parameters (grouped under shape and symmetry)

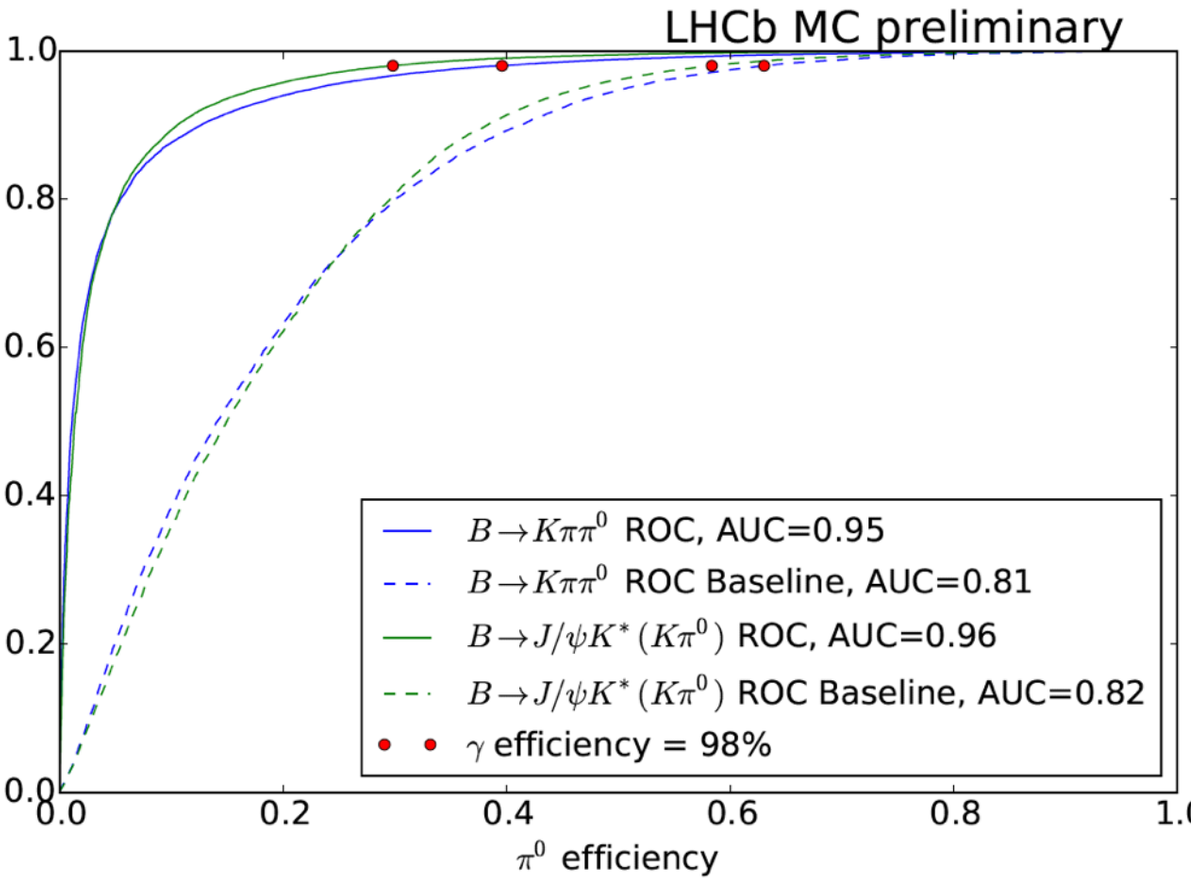
- 4 variables that account for the size & tails, semiaxes and orientation of the
- ellipse in the ECAL
- 2 variables related to the energy of the most (seed) and the second most
- energetic cells of the cluster
- 4 variables for multiplicities of hits in the PS cells matrix in front of the seed
- of the electromagnetic cluster
- 4 shape and asymmetry variables in the 3x3 PS cells

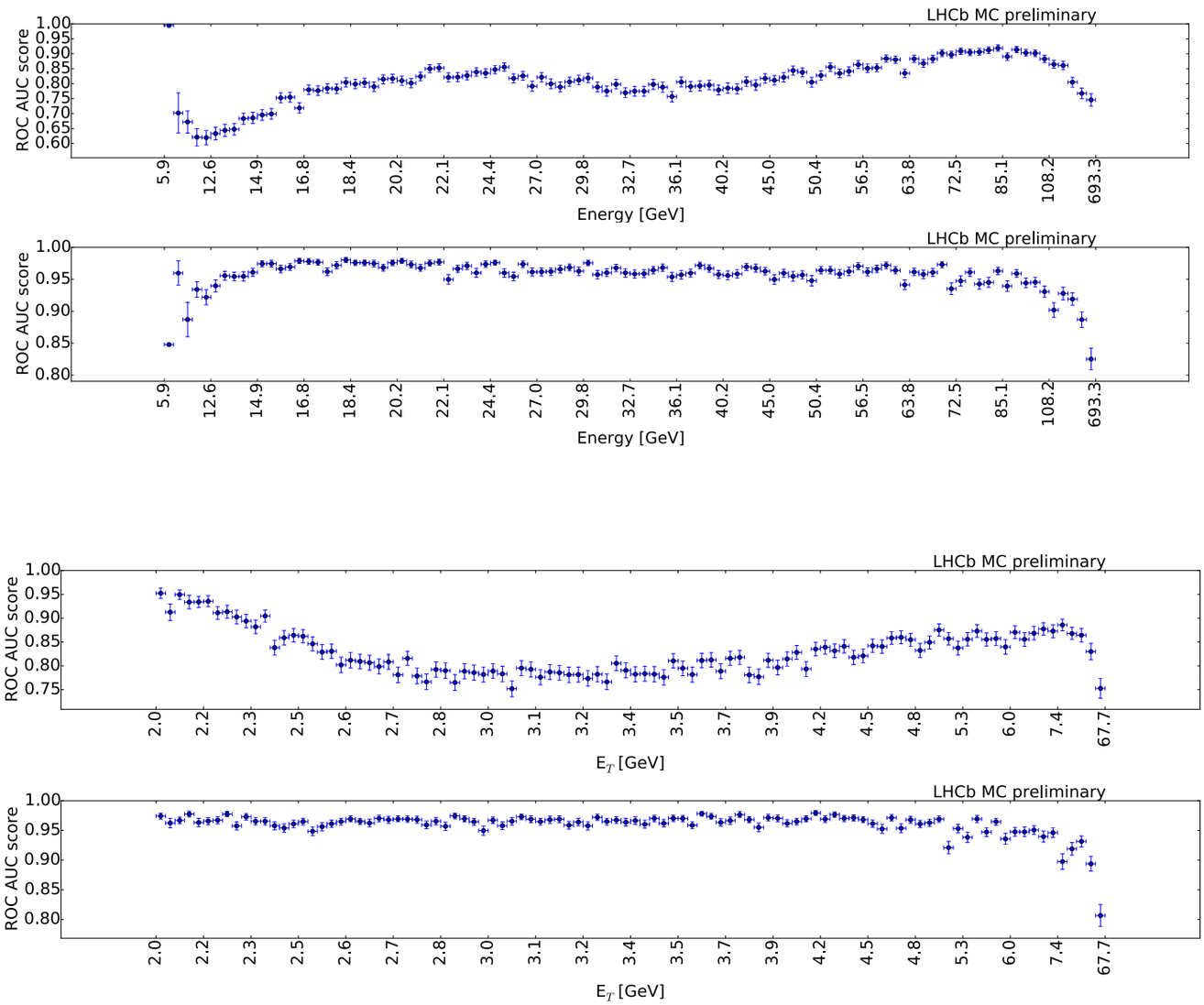




Our approach

New method: XGBoost classifier which is a Gradient Boosting over Decision Trees classifier.
Inputs
are raw energy values in 5 5 ECAL and PS cells around the cell seed. There are no any additional input features





Charged PID

Problem: identify particle type associated with a track/energy deposited in the subdetectors

- Charged: π, e, μ, K, p also we have ghosts

Standard MVA used for PID LHCb

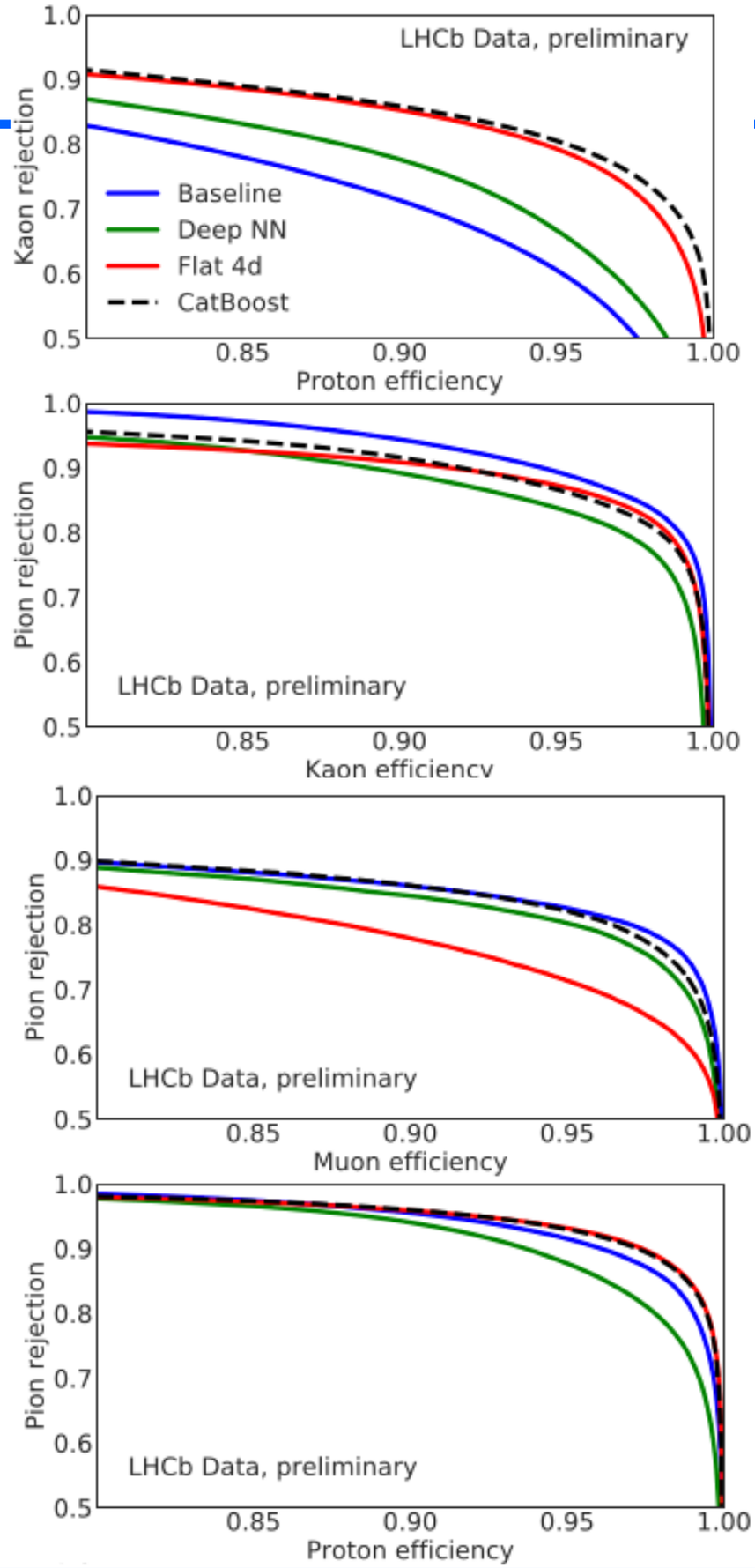
Artificial neural networks with 6 binary classification models
(One-versus-rest approach: separate one type from the others)
1 hidden layer, TMVA MLP [arXiv:0703039]

Gradient boosting:

- CatBoost [arXiv:1706.09516]

Artificial neural networks (NN)

- Deep neural networks, miulti-class approach



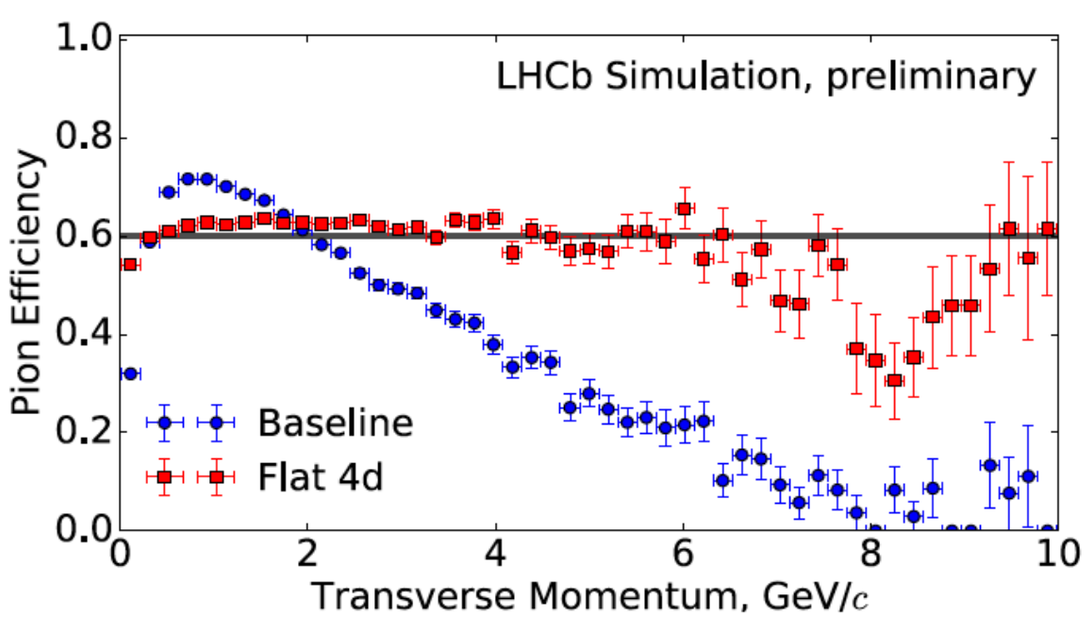
Flat efficiency approach

- PID performace depends on **particle kinematics** (p, p_T, η) and $\mathbf{N_{tracks}}$
- Flat PID efficiencies:
  - ★ Good discrimination for different analyses
  - ★ Unbiased background discrimination
  - ★ Reduced systematic uncertainties

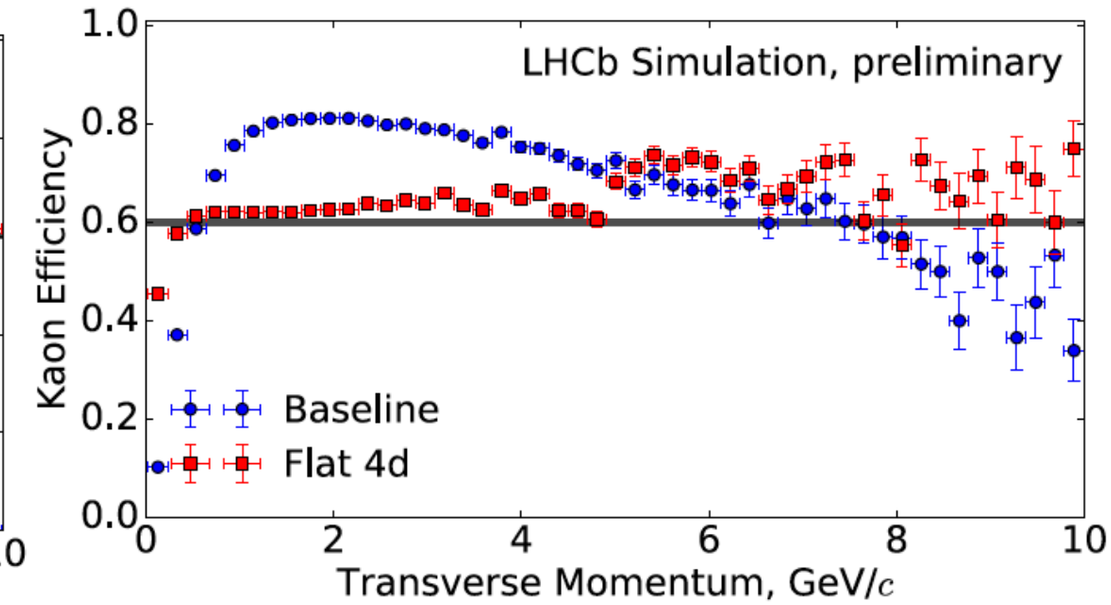
Introduce flatness term in loss function: $\mathcal{L} = \mathcal{L}_{AdaLoss} + \alpha \mathcal{L}_{Flat}$

- **Flat4d:** $\mathcal{L}_{Flat_{4d}} = \mathcal{L}_{Flat\_P} + \mathcal{L}_{Flat\_PT} + \mathcal{L}_{Flat\_nTracks} + \mathcal{L}_{Flat\_\eta}$

Pions



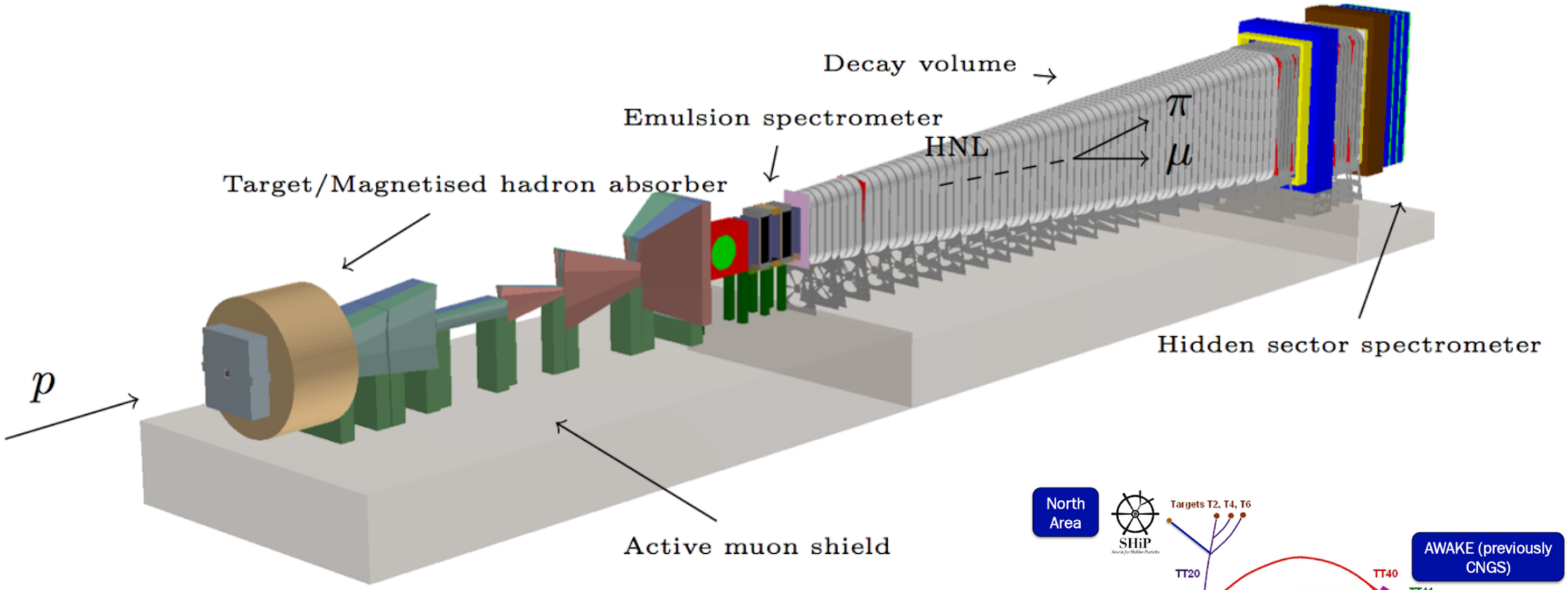
Kaons

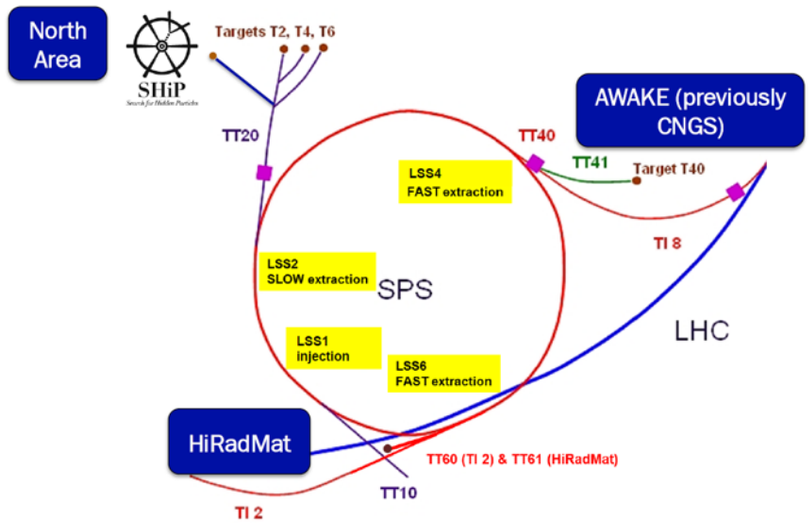


Flat4d, ProbNN

$\rightarrow$ Better PID efficiency flatness in p, p_T, η, N_{tracks} than baseline

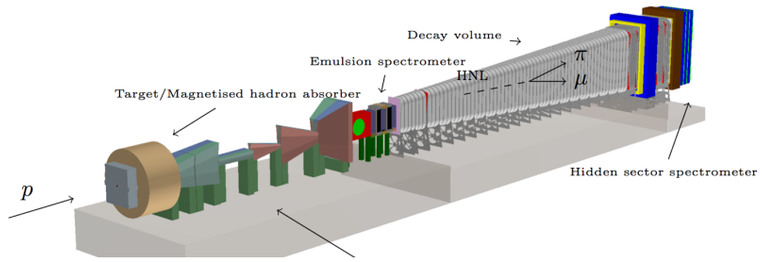
SHiP Experiment

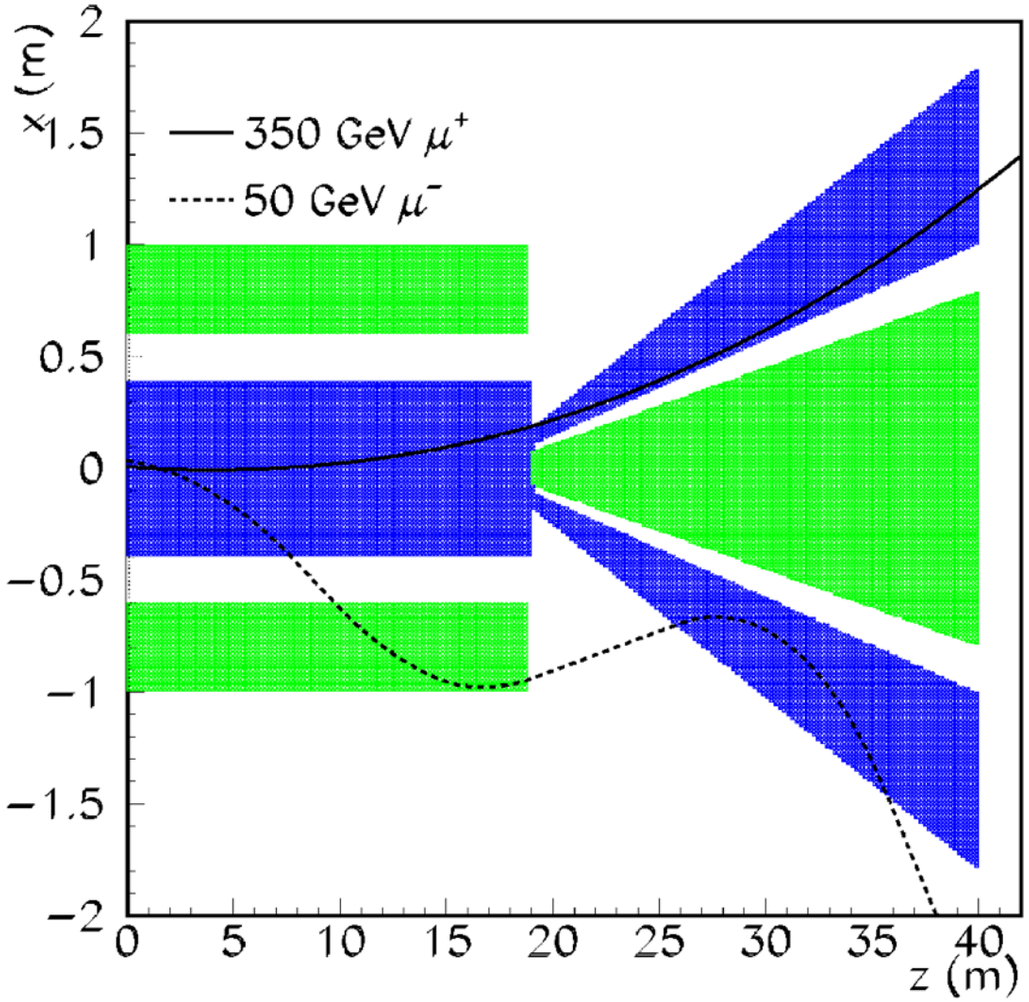


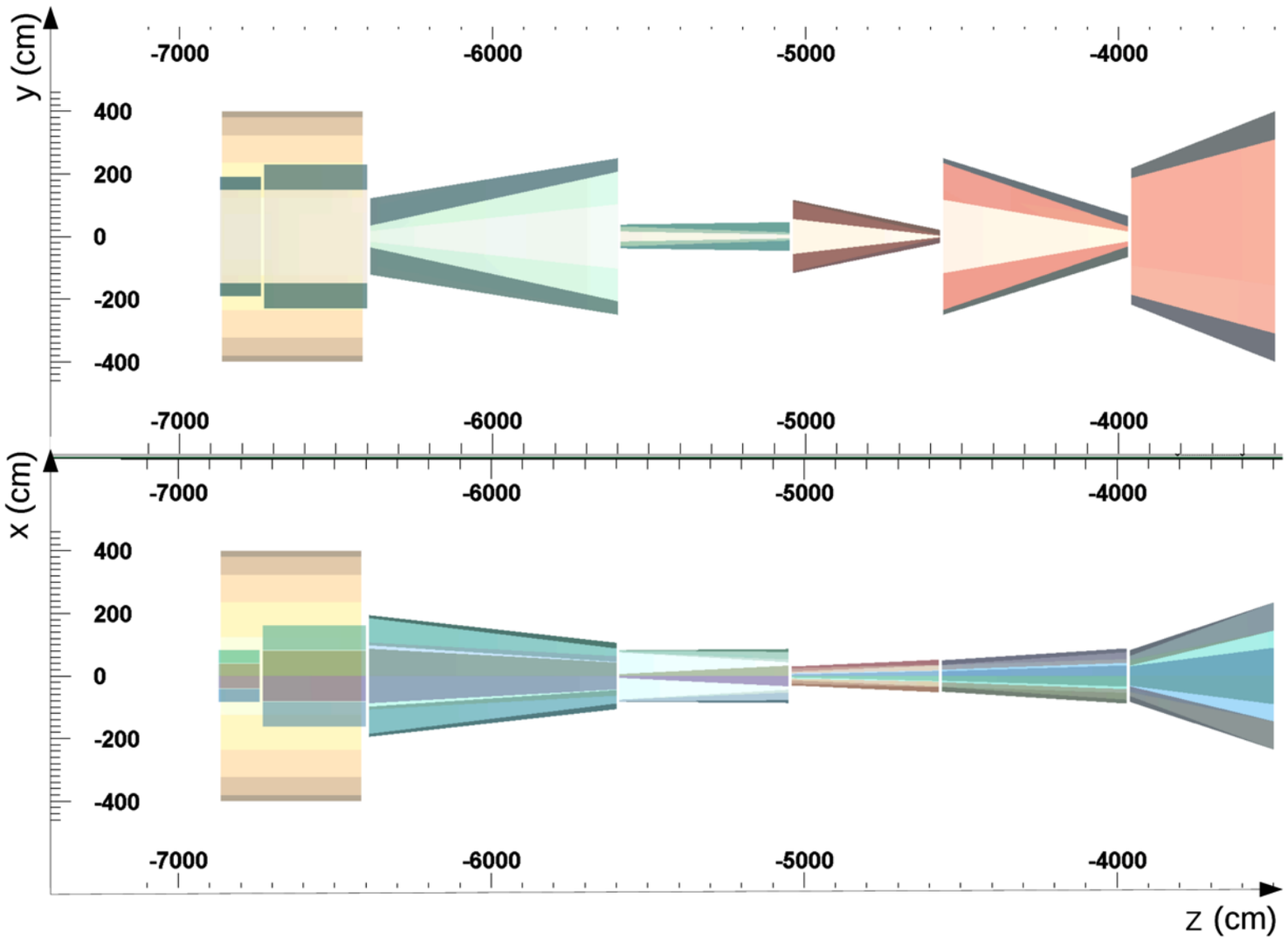


- Search for Hidden Particles
  - Post-LHC era experiment for direct search of very weakly interacting light particles

Active Magnetic Shield



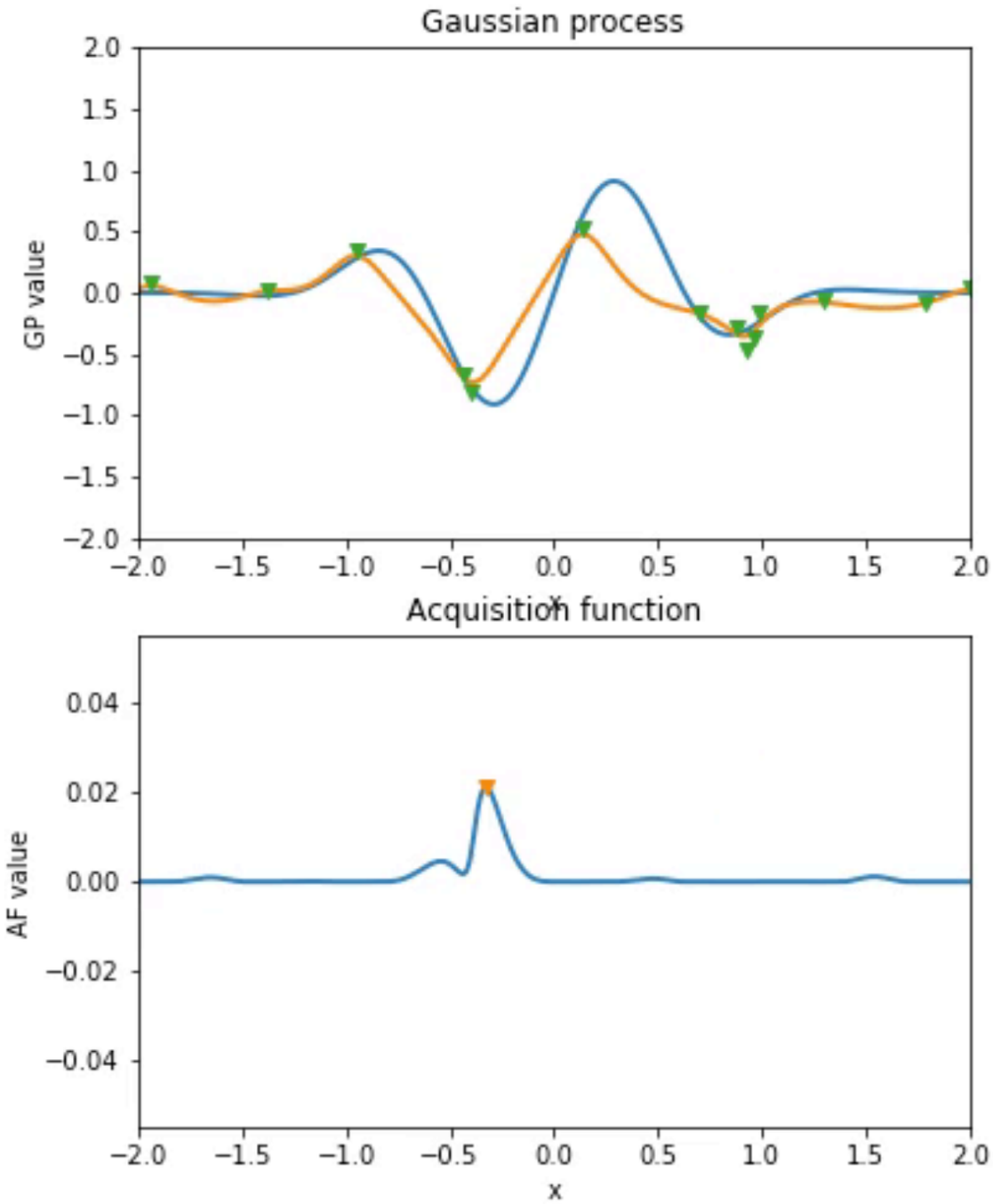




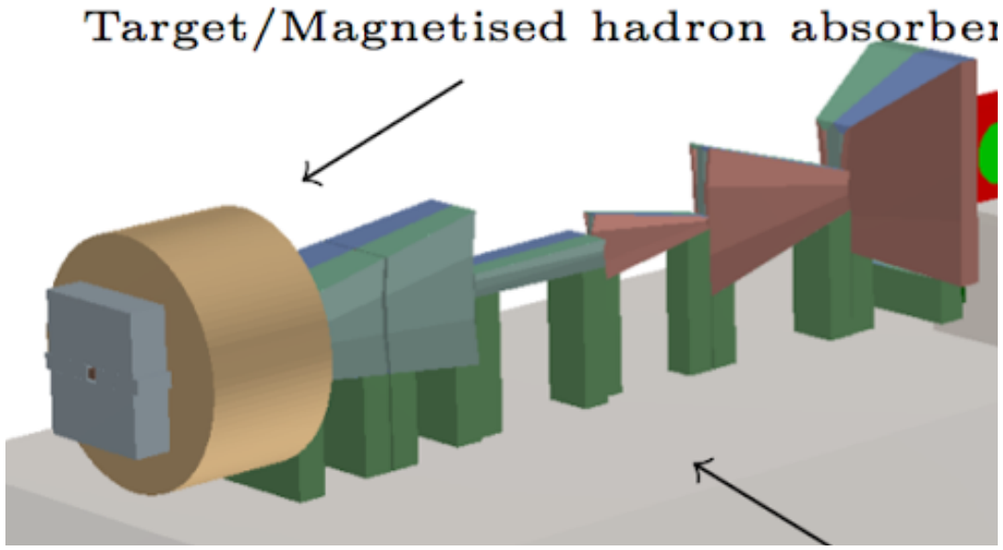
- Absorber shape optimization: background suppression at reasonable cost

Gaussian Process Optimization

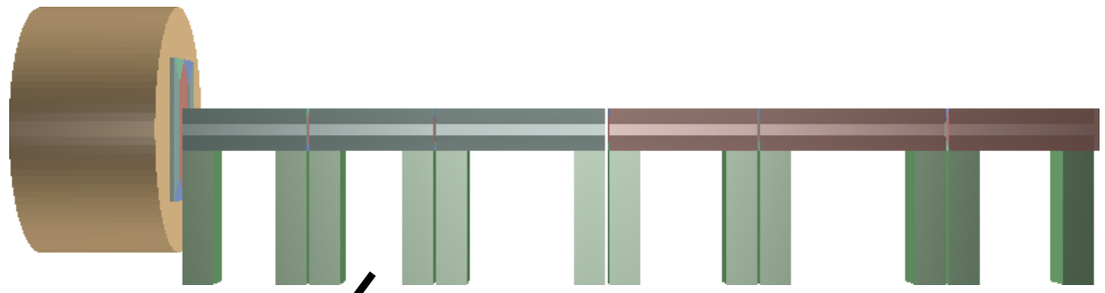
- Loss function includes both background level and cost
- 50+ configuration parameters
  - estimation in every point takes significant time
    - full GEANT simulation of 10+M muons passing through iron
  - loss function is very irregular in the multidimensional parameter space
- Use Gaussian Processes

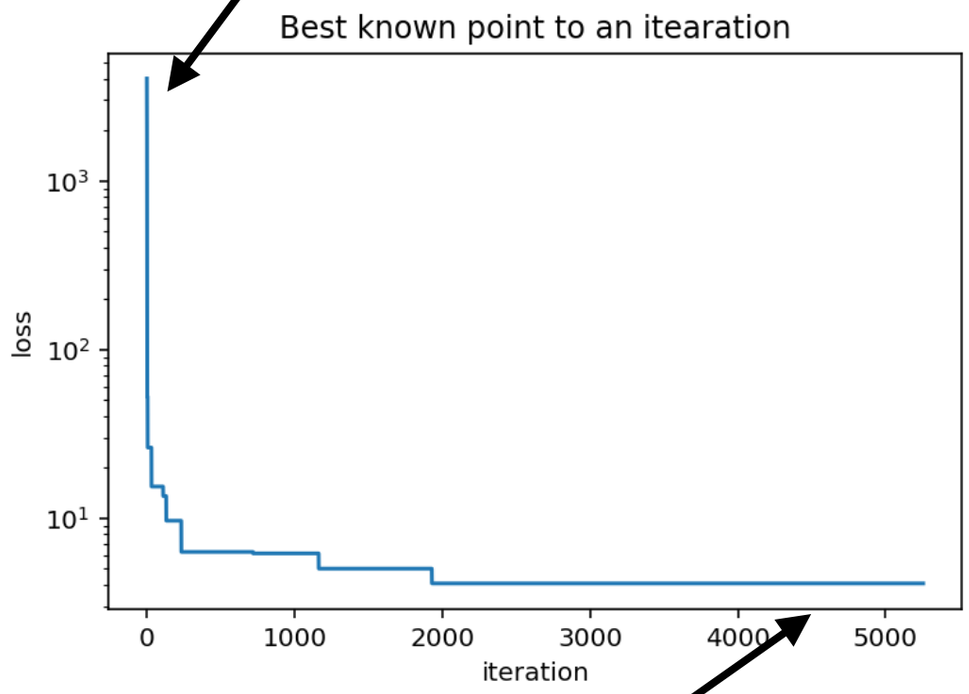


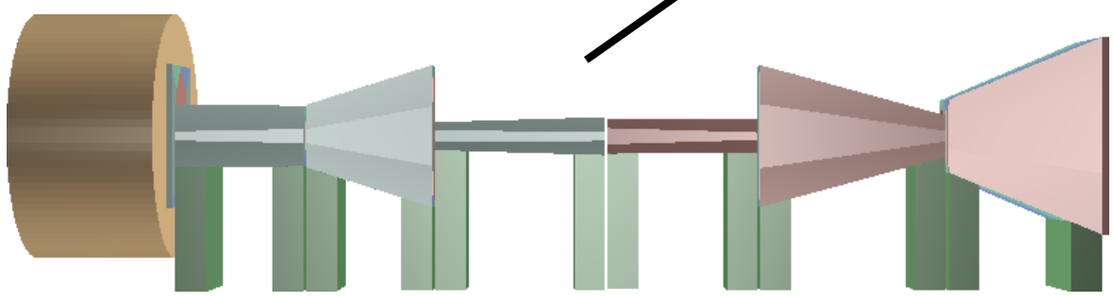
Shield Optimization



- The same background suppression
- Twice lighter
  - save $$



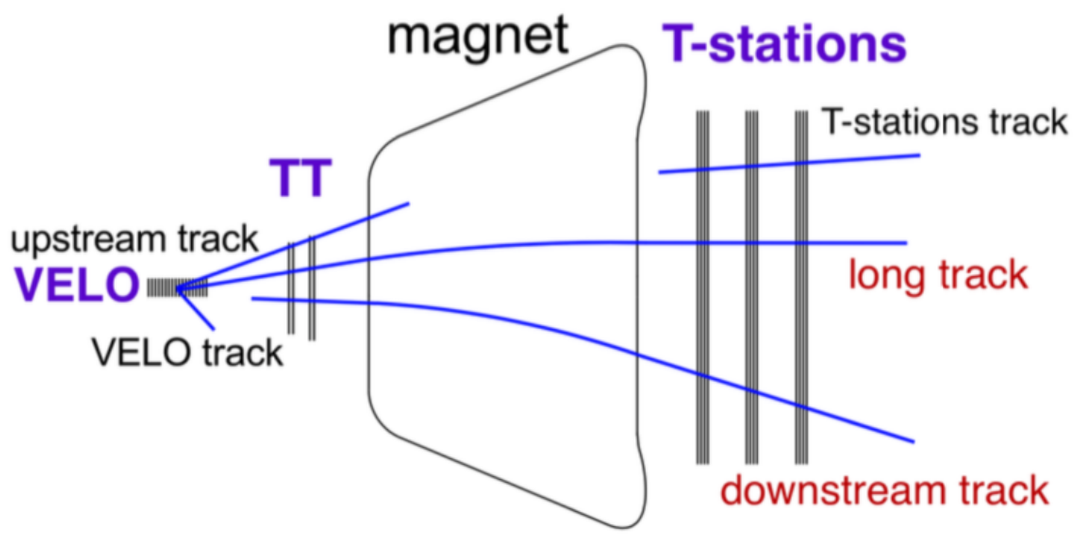




Advanced optimization methods rule in multidimensional space

Track reconstruction at LHCb

- The Tracking System consist of:
  - VErtex LOcator (VELO)
  - Two stations downstream magnet (TT)
  - Three stations upstream magnet (T-stations)
- Two phases in the LHCb tracking:
  - finding (pattern recognition)
  - fitting (Kalman-filtering)



- Main track types for physics analyses:
  - Long tracks: hits in VELO, T stations (and eventually TT).
    Used in majority of analyses (B/D decays)
  - Downstream tracks: hits in TT and T stations.
    Tracks from daughters of long lived particles (Λ, K_s^0)
- The same reconstruction in the trigger and offline.
- Need to fit in the tight timing budget (35 ms for HLT1 and 650 ms for HLT2)

Track reconstruction at LHCb

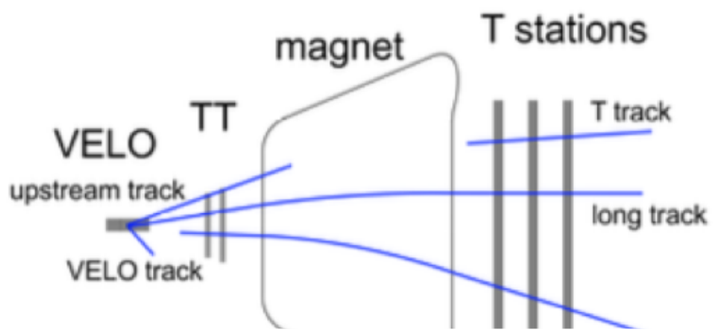
HLT1 sequence (35 ms)

Offline VELO tracking with simplified Kalman Filter

Primary Vertex finding

VELO → TT:
initial momentum estimate
p_T>200 MeV

TT → T-stations



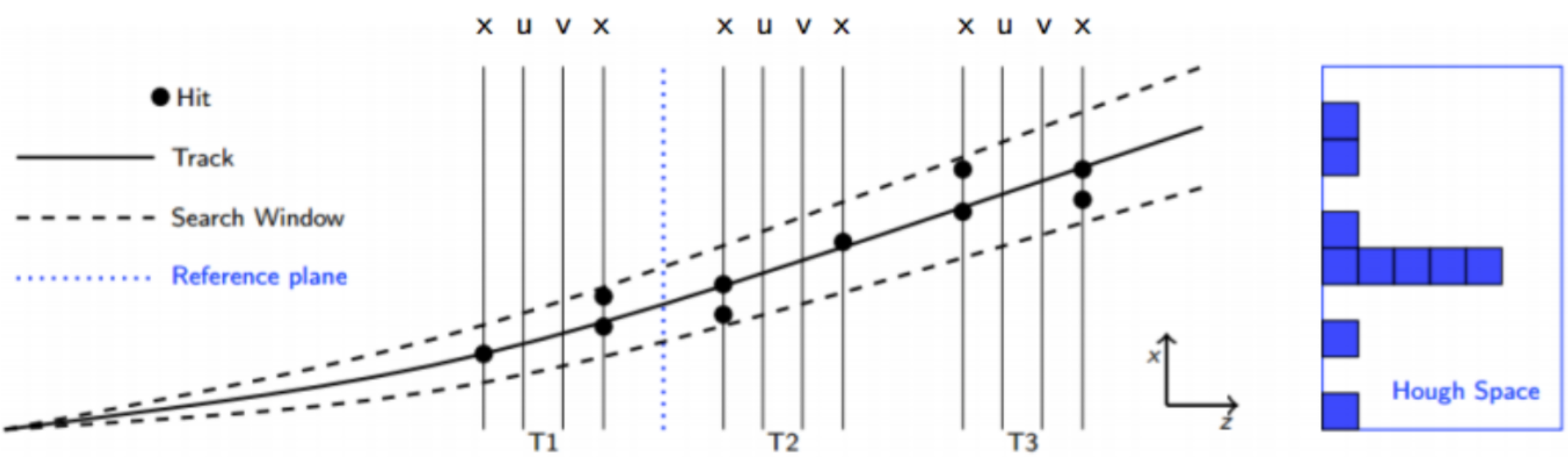
HLT2 sequence (650 ms)

VELO → T-stations
no p_T threshold

T-station standalone tracking

Downstream tracking

Matching tracking



Track reconstruction at LHCb

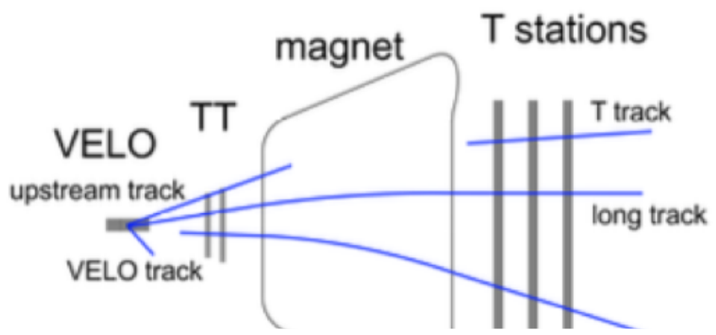
HLT1 sequence (35 ms)

Offline VELO tracking with simplified Kalman Filter

Primary Vertex finding

VELO → TT:
initial momentum estimate
p_T>200 MeV

TT → T-stations



HLT2 sequence (650 ms)

VELO → T-stations
no p_T threshold

T-station standalone tracking

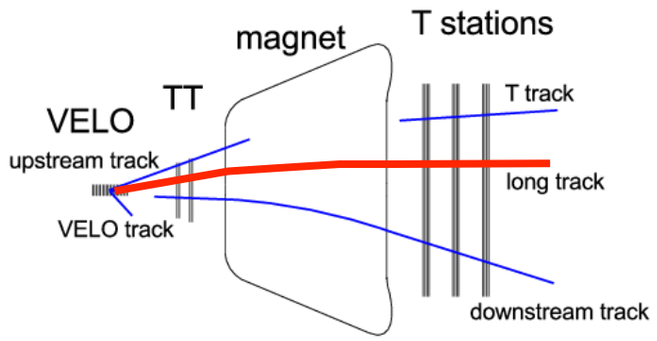
Downstream tracking

Matching tracking

Machine Learning in the LHCb Tracking

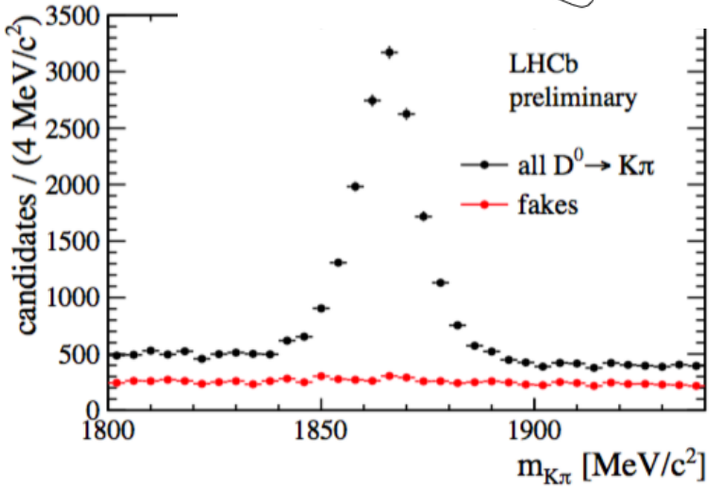
Tracking algorithms are also good places for using Machine Learning.

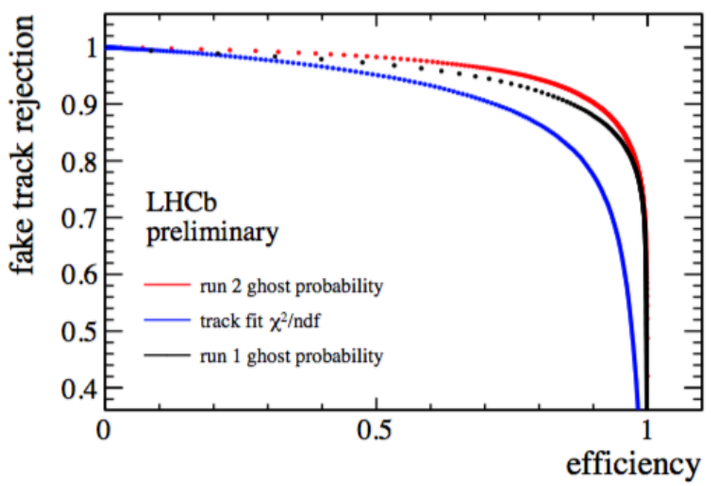
- Neural Network (most often Multi Layer Perceptron (MLP))
- bonsai Boosted Decision Trees (bBDT)
- ...



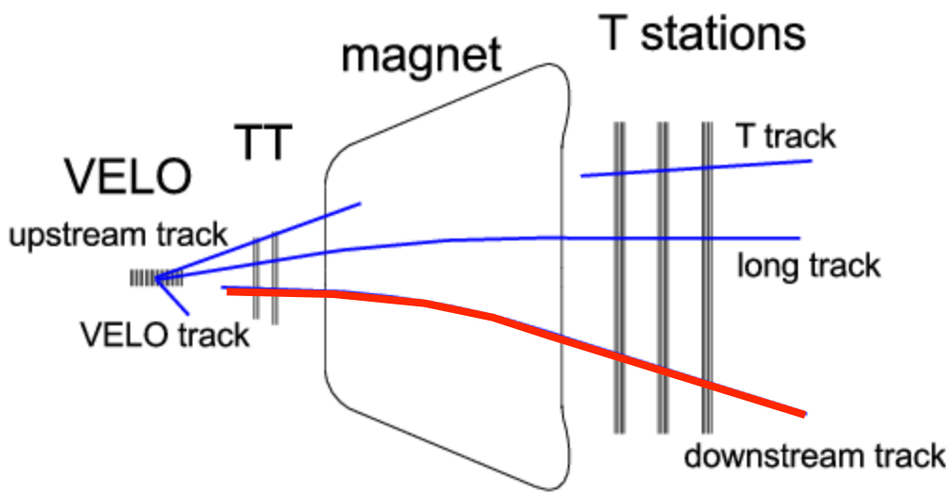
Example of using Machine Learning: Fake track rejection

- After Kalman Filtering still significant amount of fake tracks
- Further ∼30% fake rate reduction with MLP.
- First tuning for 2015
  - speed up by factor $O(90)\%$,
  - less than 0.5 ms per event,
  - used in HLT2.
- Further improvements for 2016:
  - reduces HLT2 combinatorics by 40% with negligible efficiency loss,
  - used in HLT1.





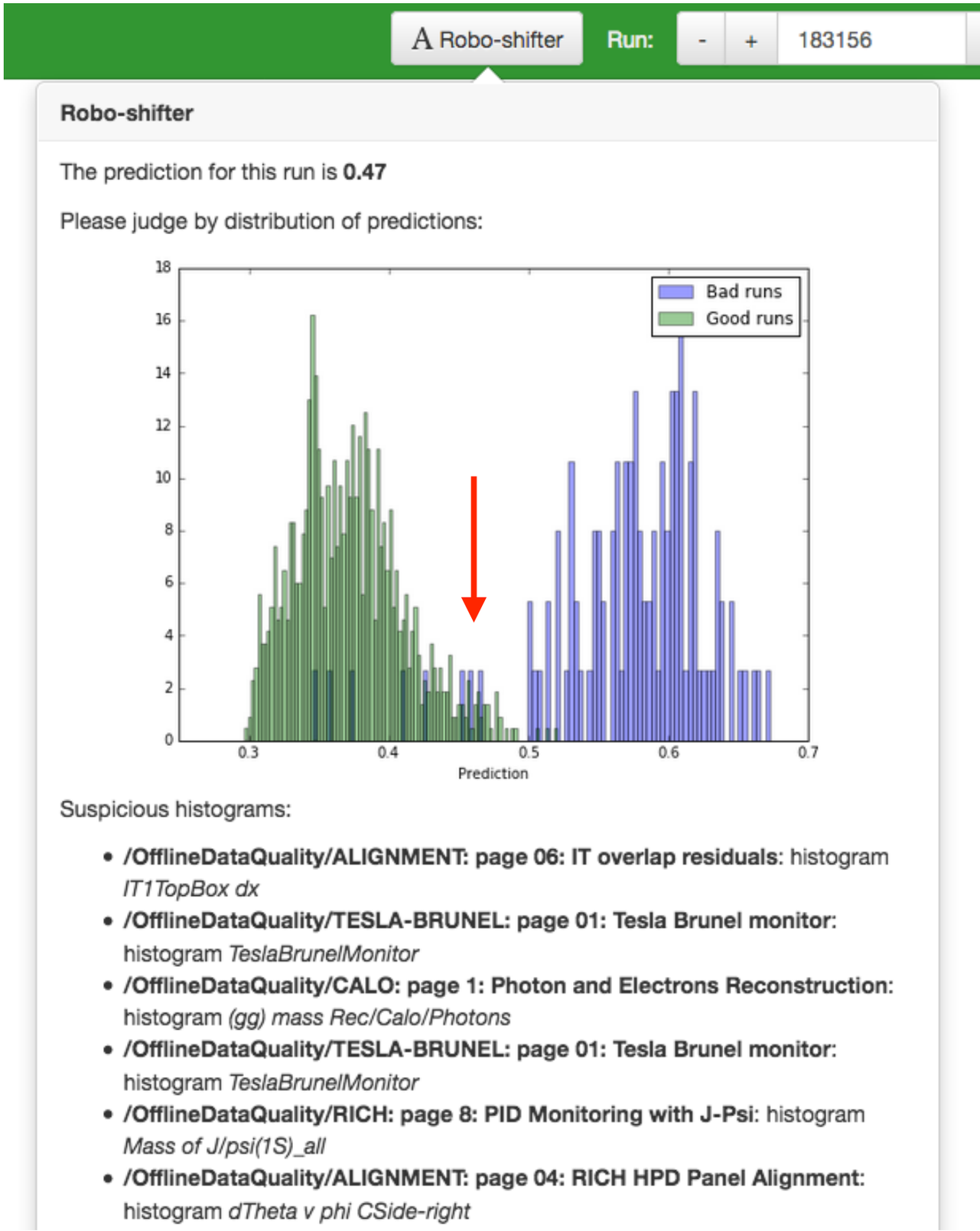
Downstream Tracking



- Downstream tracking contains two Multivariate classifiers
  - reject as much fake T-seeds as possible: avoiding unnecessary reconstruction
    - BBDT
  - final accepting tracks: further reducing fake tracks
    - MLP
- Improved both fake tracks reduction and signal efficiency gain by 3-5%
  - implemented for 2017 operation

Monitoring Robo-shifter

- Robo-shifter is machine-learning based system designed to assists the DQ shifter
- Given run data it can predict probability of run being good or bad
- Hint for potential problem sources is extracted from decision trees
- Commissioned for LHCb Data Quality Monitoring

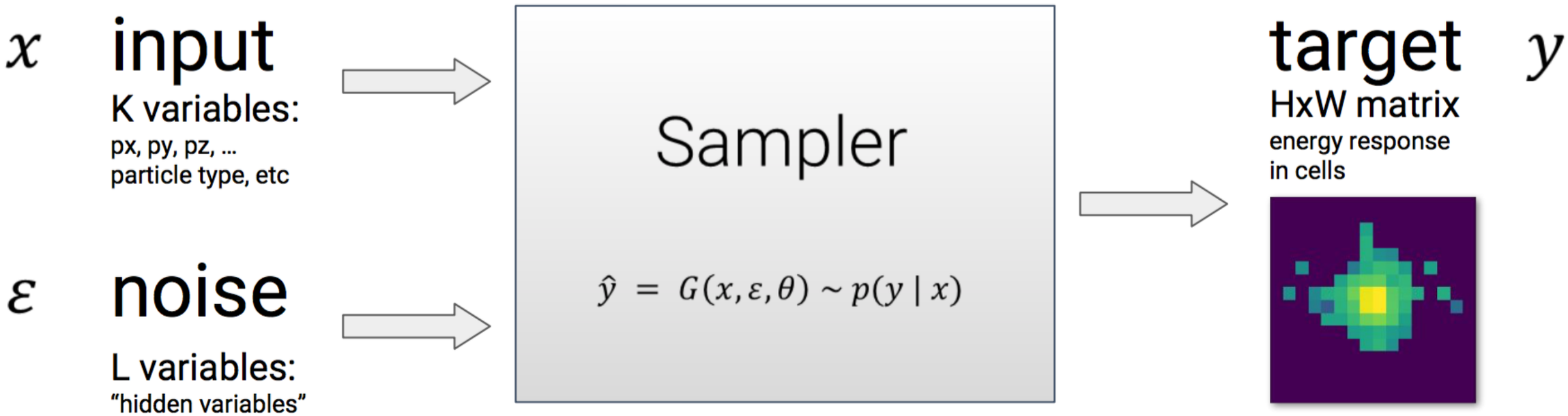


Generative Models

- Computationally heavy tasks
  - e.g. simulating shower development in the calorimeter
- May be substituted by generative models trained on the original task
  - save orders of magnitude in computing performance
  - challenge is to keep physics performance high

Problem

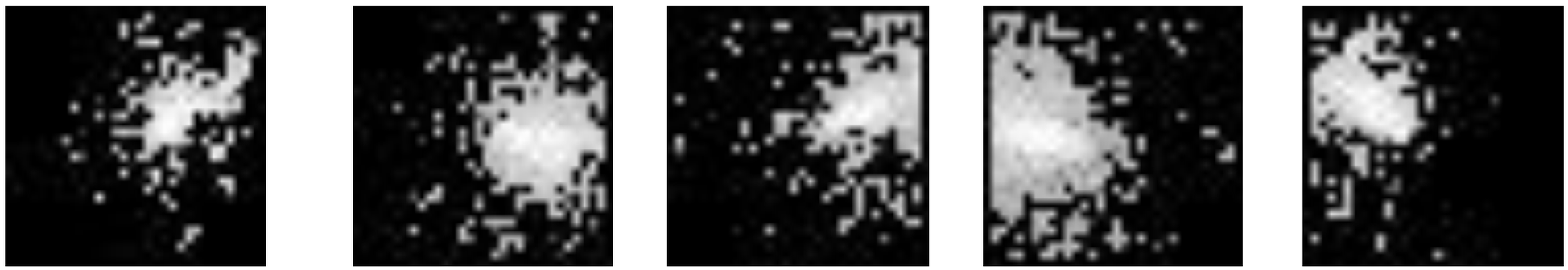
- We want to speed up calorimeter simulation (calorimeter showers) while keeping reasonable simulation accuracy (correctly reproducing simulation behavior)
  - consider LHCb ECAL as a practical goal
- Our ML problem formulation (hidden variables model):



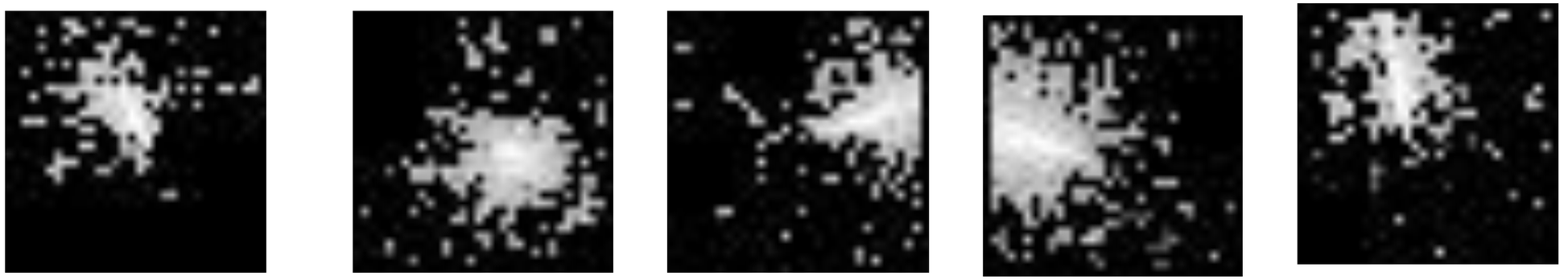
Conditional WGAN

Very Preliminary Results

GEANT Simulated

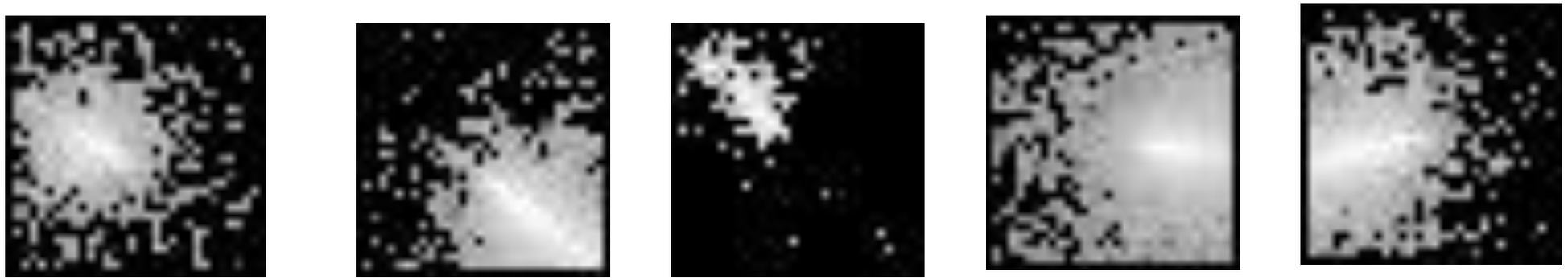


GAN Generated



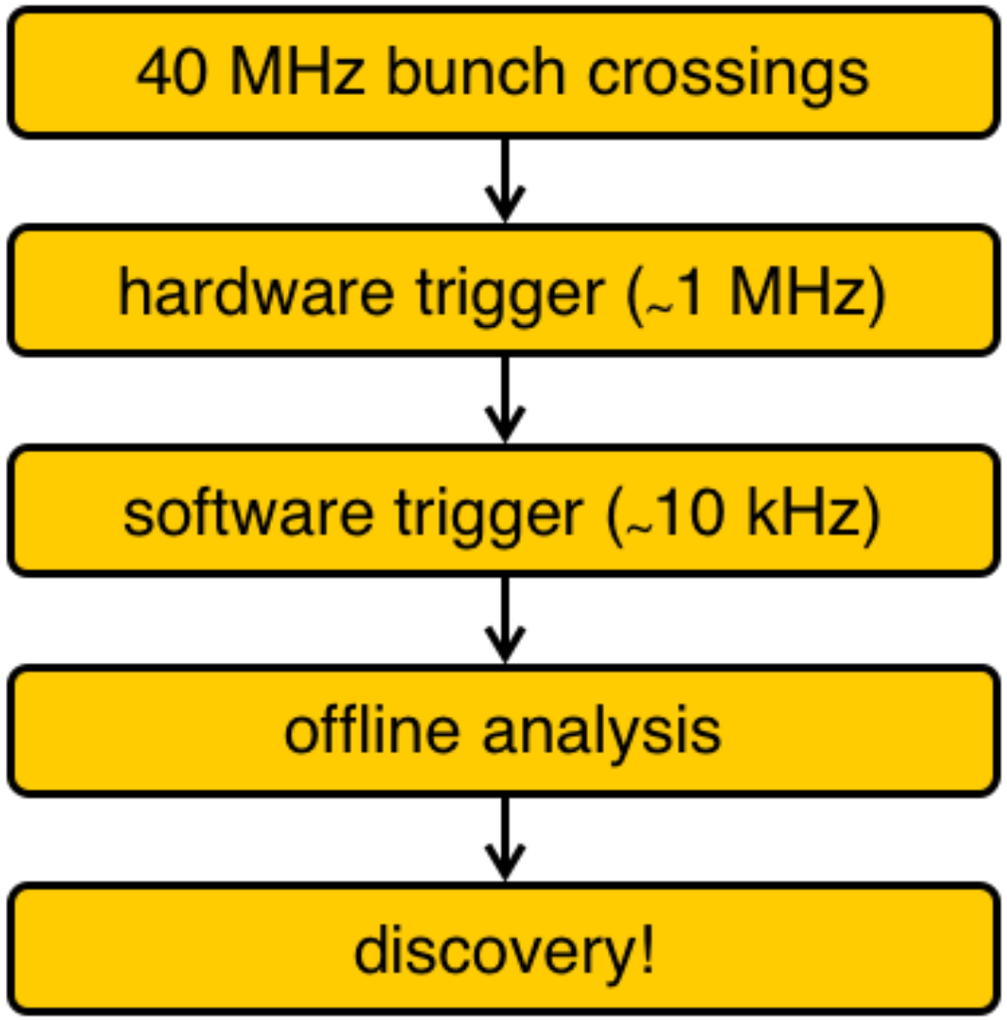
GEANT Simulated

GAN Generated

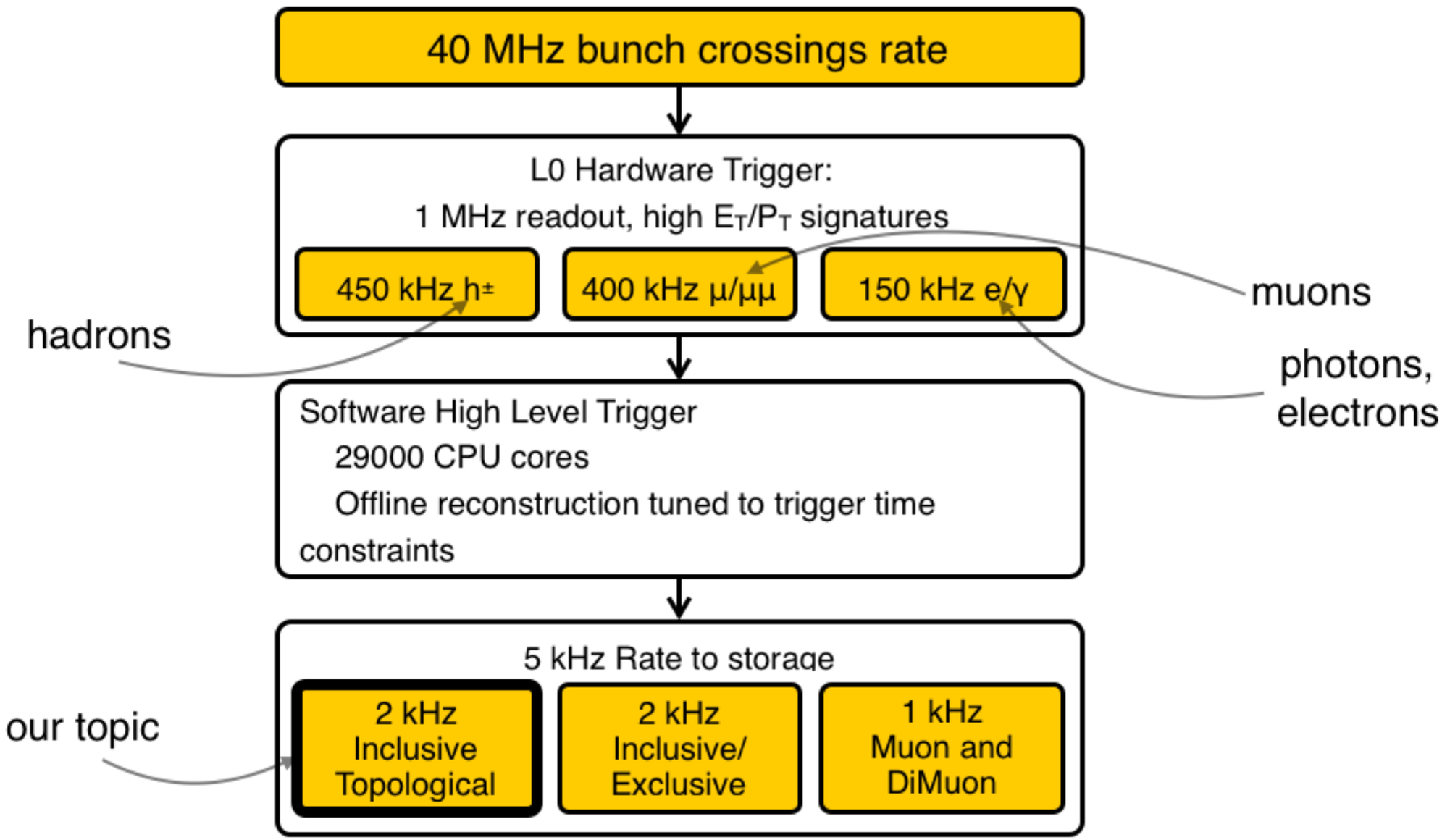


Triggers at LHCb

Need to collect many **different** interesting events
For this, we need to cut very hard in order to collect enriched data sample



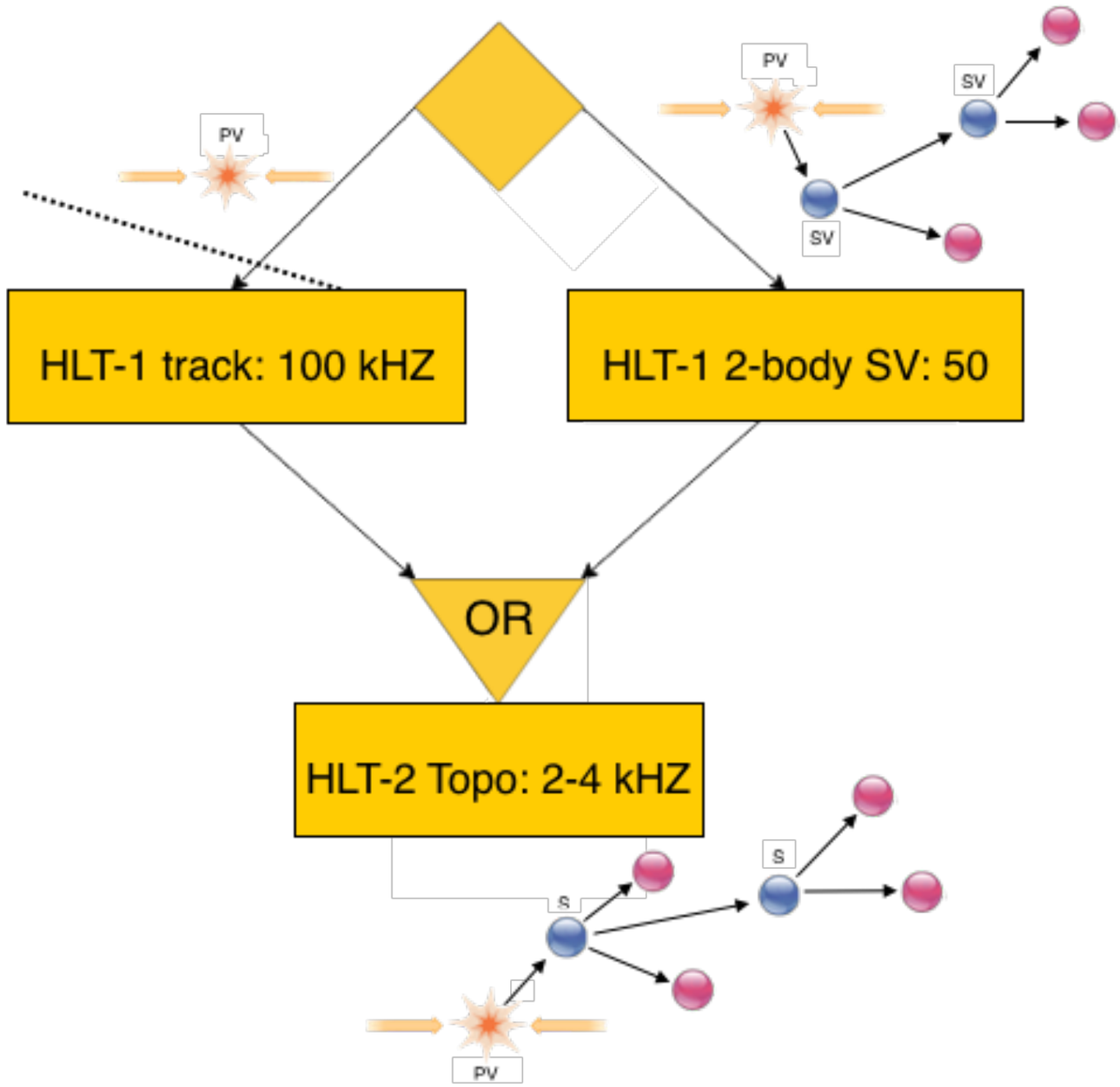
Triggers at LHCb



Muons, hadrons, photons are particle species

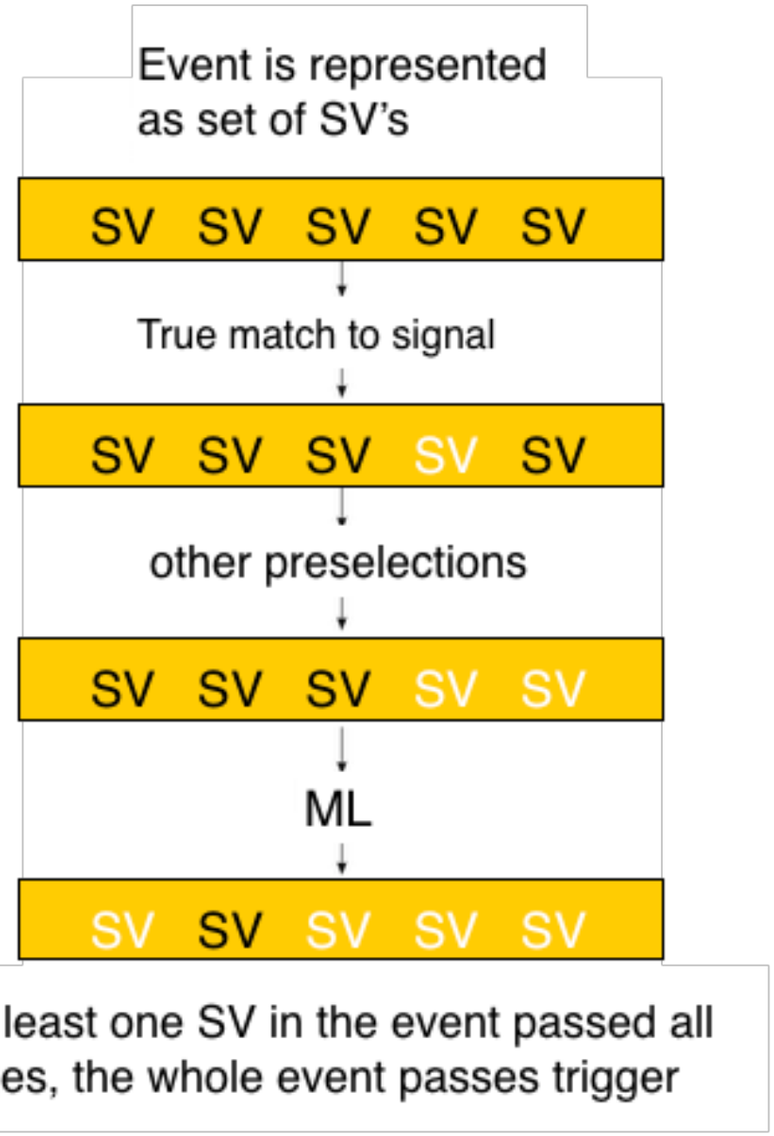
LHCb topo-trigger

- HLT-1 track is looking for either one super high PT or high displaced track
- HLT-1 2-body SV classifier is looking for two tracks making a vertex
- HLT-2 improved topological classifier uses full reconstructed event to look for 2, 3, 4 and more tracks making a vertex



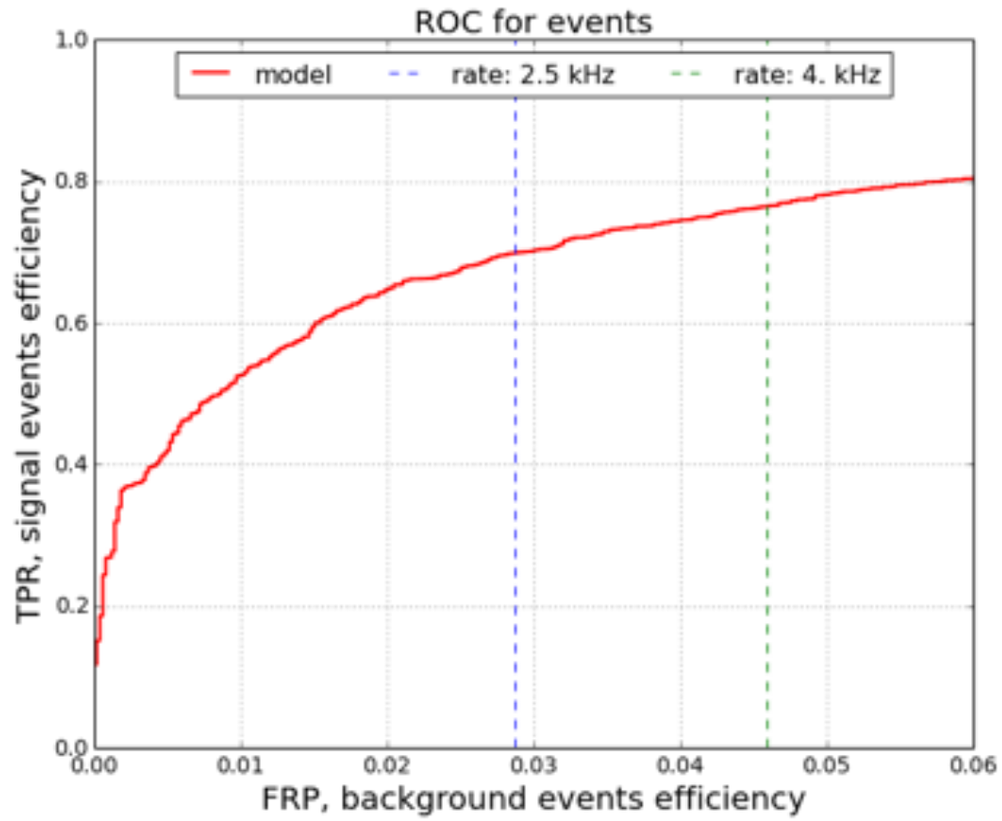
Data

- Monte Carlo samples (used as signal) are simulated with B decays of various topologies
- Generic proton-proton collisions are used as background sample (also includes some signal)
- Most events have many secondary vertices
- Goal is to improve efficiency for each type of signal events with fixed efficiency for background

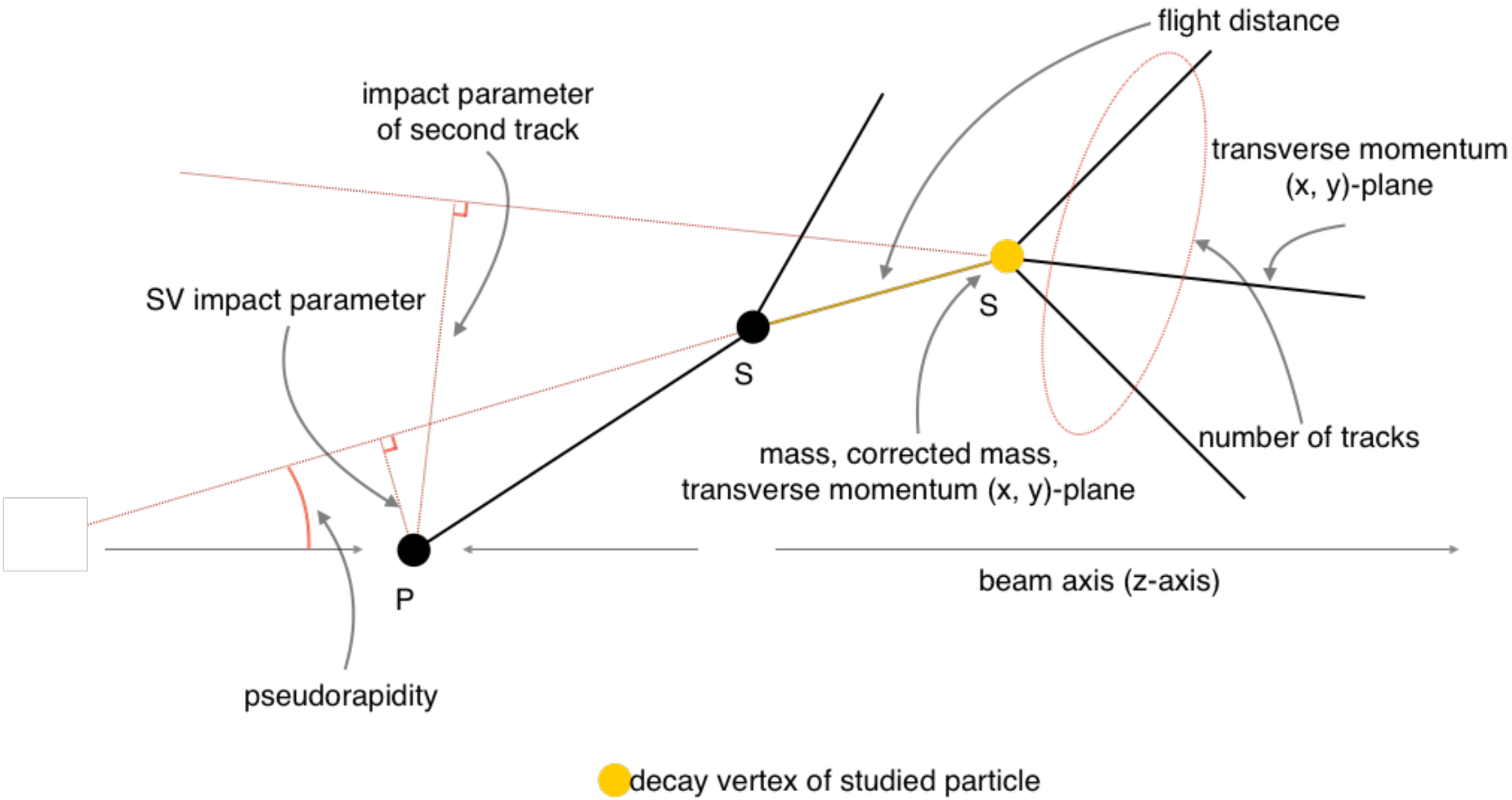


ROC curve, computed for events

- Output rate = false positive rate (FPR) for events (since background = generic event)
- Optimise true positive rate (TPR) for fixed FPR for events
- Weight signal events in such way that channels have the same amount of events.
- Optimise ROC curve in a small FPR region

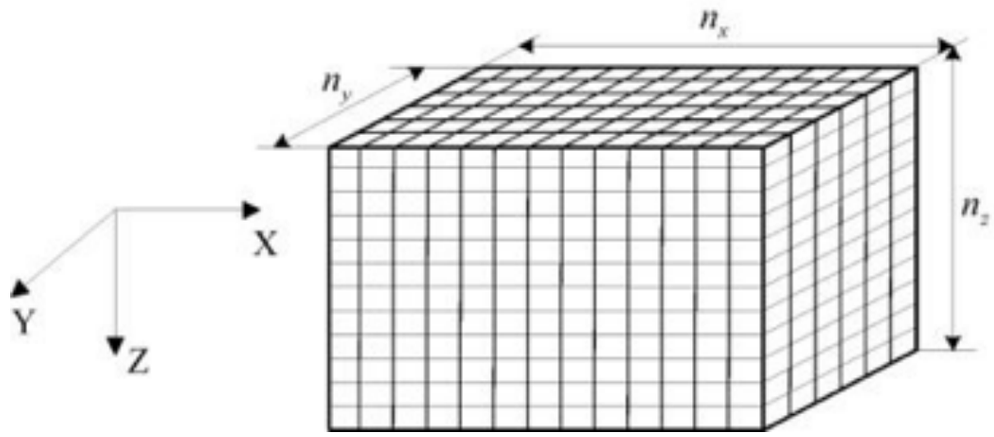


Available decision variables

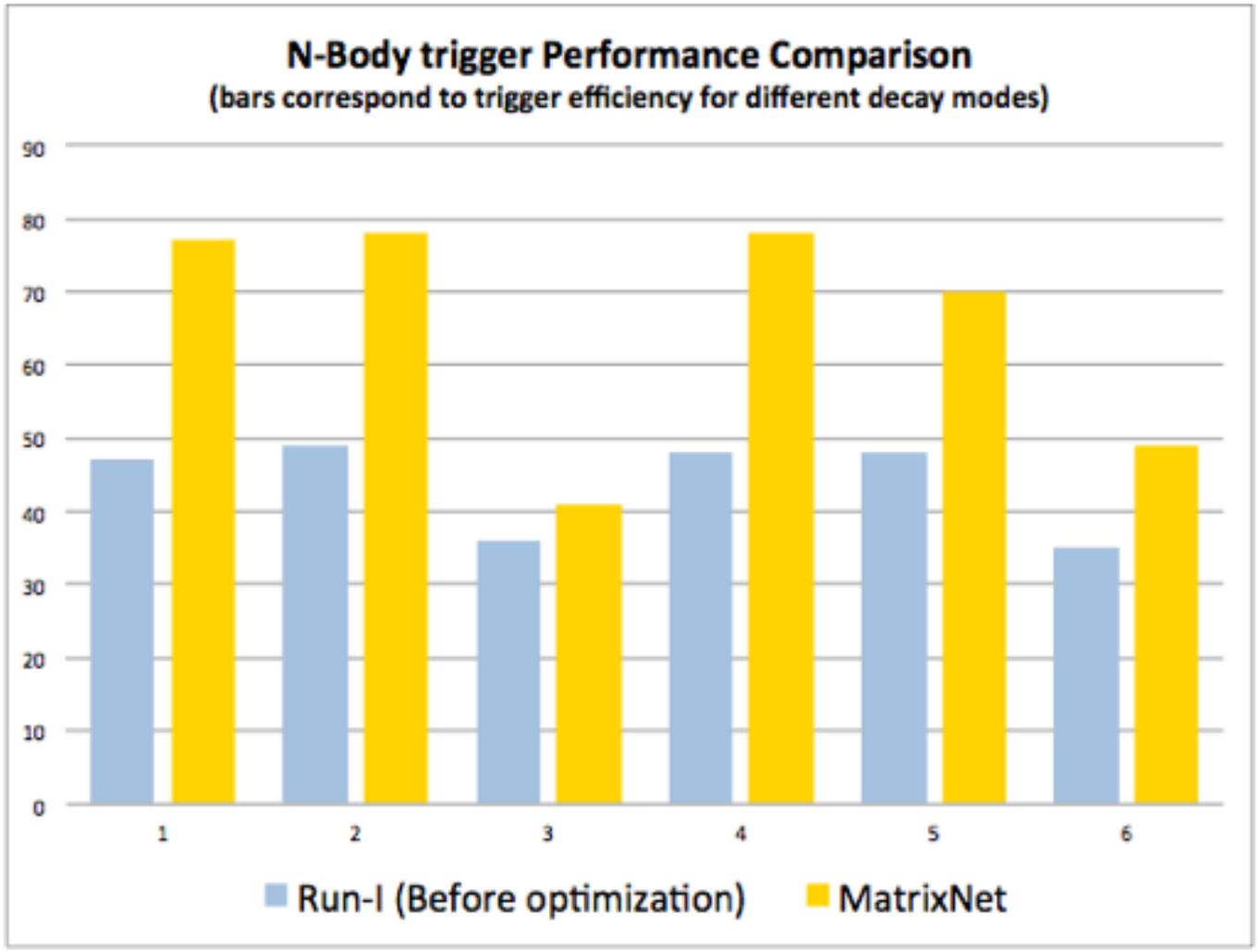


Online part using Bonsai BDT

- Features hashing using bins before training
- Converting decision trees to n-dimensional table (lookup table)
- Table size is limited in RAM (1Gb), thus count of bins for each features should be small (5 bins for each of 12 features)
- Discretisation reduces the quality



Topological trigger results



$$B^0 \to K^*[K^+\pi^-]\mu^+\mu^-$$
$$B^+ \to \pi^+K^-K^+$$
$$B_s^0 \to D_s^-[K^+K^-\pi^-]\mu^+\nu_\mu$$
$$B_s^0 \to \psi(1S)[\mu^+\mu^-]K^+K^-\pi^+\pi^-$$
$$B_s^0 \to D_s^-[K^+K^-\pi^-]\pi^+$$
$$B^0 \to D^+[K^-\pi^+\pi^+]D^-[K^+\pi^-\pi^-]$$

https://github.com/yandexdataschool/LHCb-topo-trigger

Summary

- Our group has quite a lot of successful stories to demonstrate power of modern CS approaches to the High Energy Physics
- MPD@NICA is a very interesting project
  - attractive physics program
  - active development phase
  - nice and convenient location
  - many possibilities for applying our expertise
- We would like to discuss a possible task(s) to get involved into the detector R&D in the most efficient way