Machine learning at LHCb

Nikita Kazeev on behalf of the LHCb collaboration

NRU Higher School of Economics

Yandex School of Data Analysis

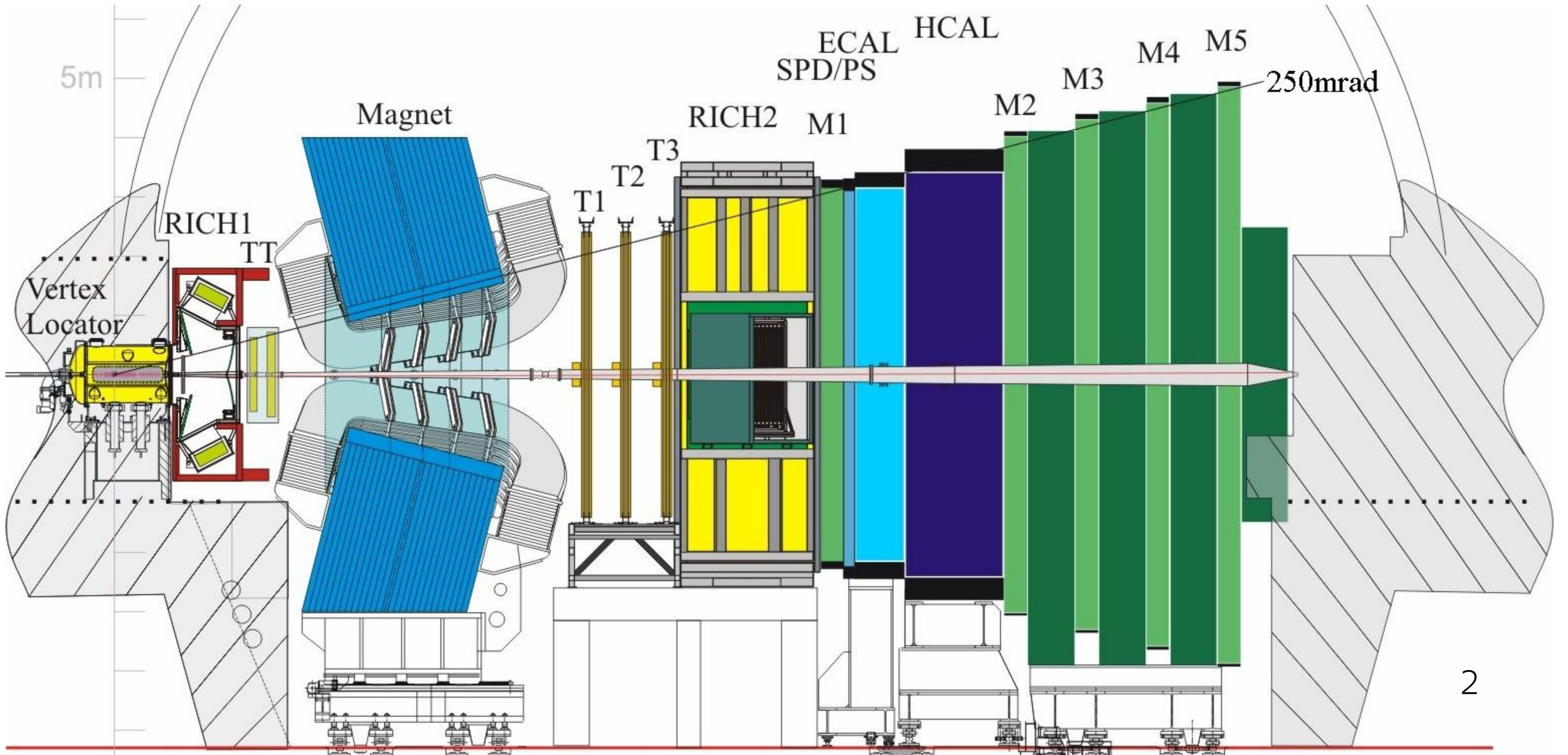
Sapienza University of Rome

ICPPA 2018, 22-26 October, Moscow

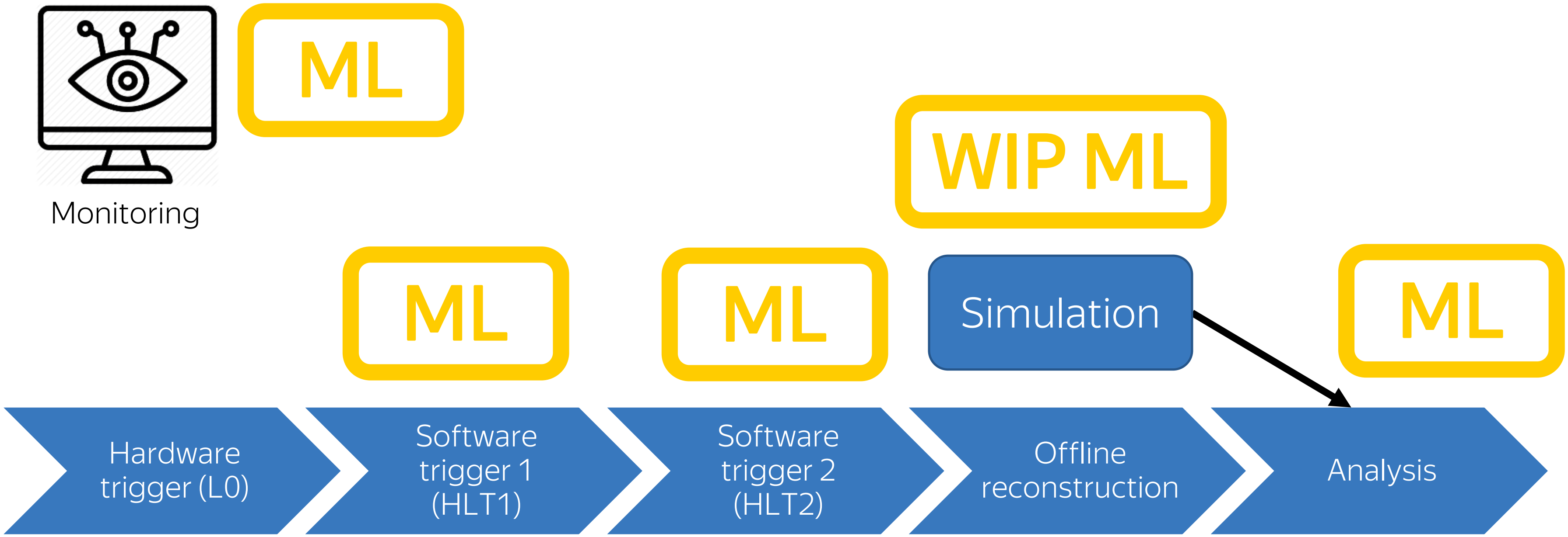
The LHCb detector

[LHCb detector performance]

- LHCb is a single-arm forward spectrometer
- The main goal of the detector is to search for indirect evidence of new physics in CP violation and rare decays of beauty and charm hadrons



LHCb Run II data flow



Hardware trigger (L0)
- Real-time
- Selects events with hits in muon chamber
- Selects events with significant amount of transverse energy in hadronic calorimeter

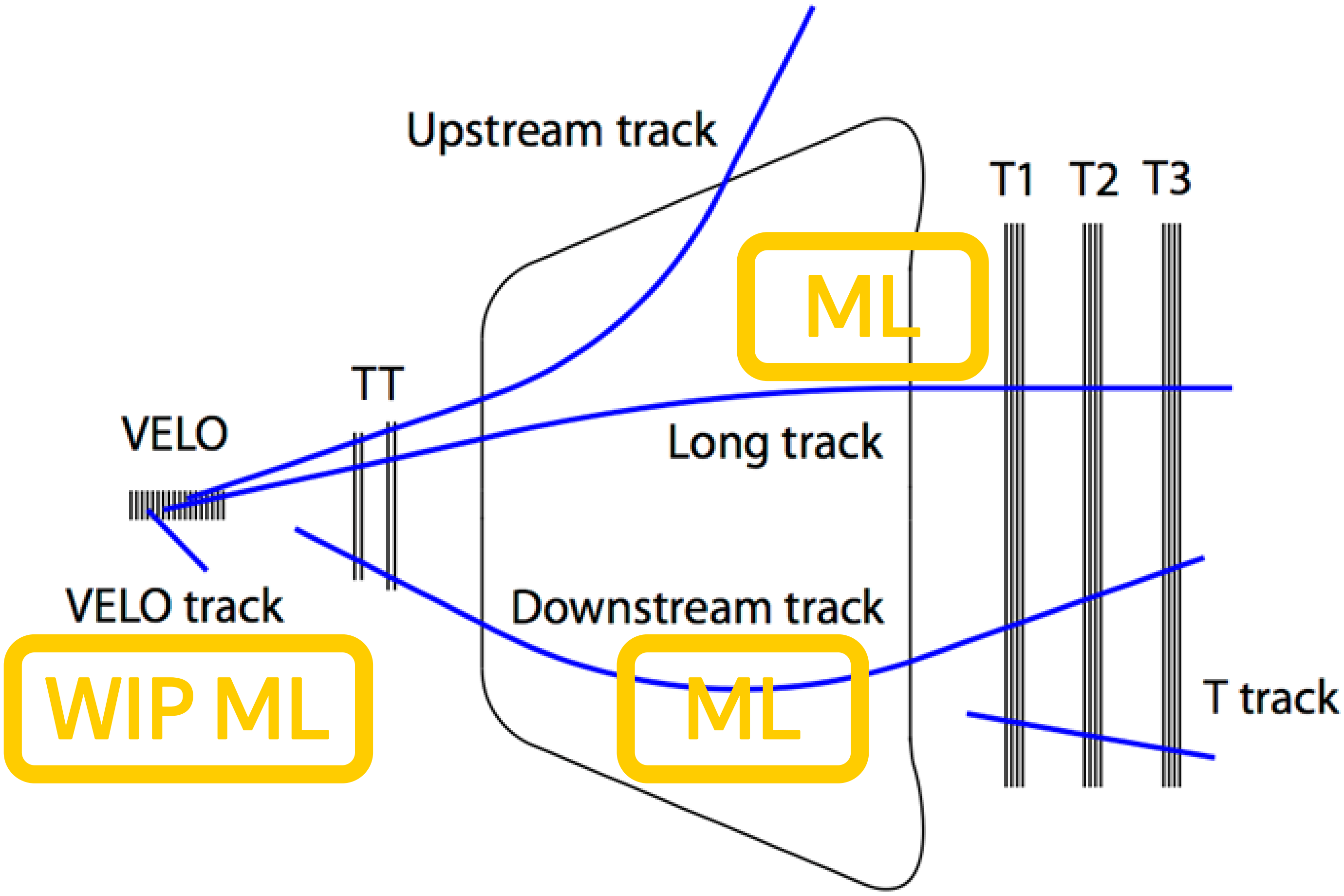
Software trigger 1 (HLT1)
- Real-time
- Selects events with muons
- Selects event s with high p_t tracks
- Selects events with high IP

Software trigger 2 (HLT2)
- Complete event reconstruction
- Decay-specific selection

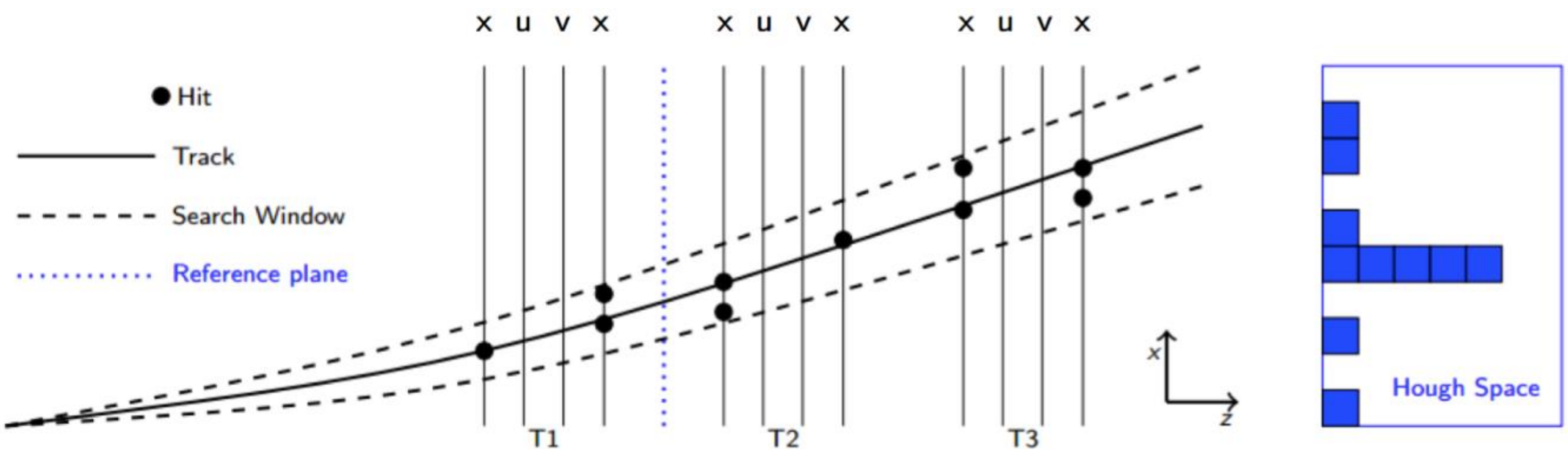
Analysis
- Decay-specific user-specific

[Real-time physics, alignment, and reconstruction in the LHCb trigger]

Track Pattern Recognition



Long Track Reconstruction



Starting from seeds in the VELO, tracks are searched in T stations:

- Search window in T stations defined by VELO track
- Project x-hit hits into reference plane – Hough transformation
- Fit **4-layer-x-cluster** and remove outliers
- Add and fit track with stereo hits

Deep Neural Networks:

1. Rejection of bad 4-layer-x-clusters in recovery loop
2. Candidates selection after stereo fit (HLT1 and HLT2)

[Tracking and Vertex reconstruction at LHCb for Run II] [Machine learning and parallelism in the reconstruction of LHCb and its upgrade]

Downstream Track Reconstruction

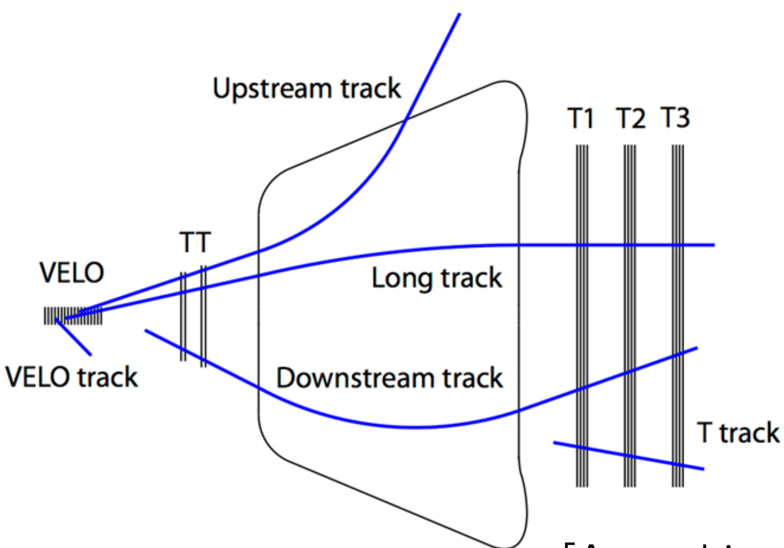
T-stations tracks → Find matching TT hits → Select best track candidate

Find matching TT hits

- Uses Bonsai Boosted Decision Trees to reject ~30% of fake T-seeds

Select best track candidate

- Uses a Multilayer Perceptron to gain 3-5% in fake tracks rejection and real tracks efficiency

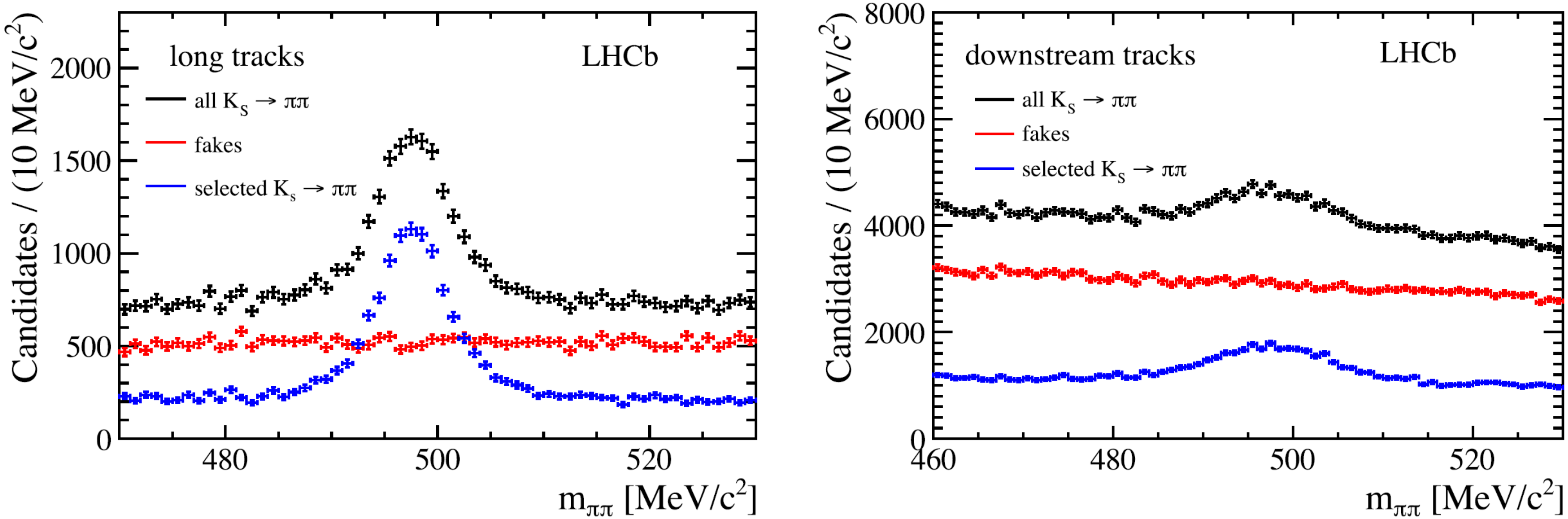


[A tracking algorithm for the reconstruction of the daughters of long-lived particles in LHCb]

Fake Track Rejection

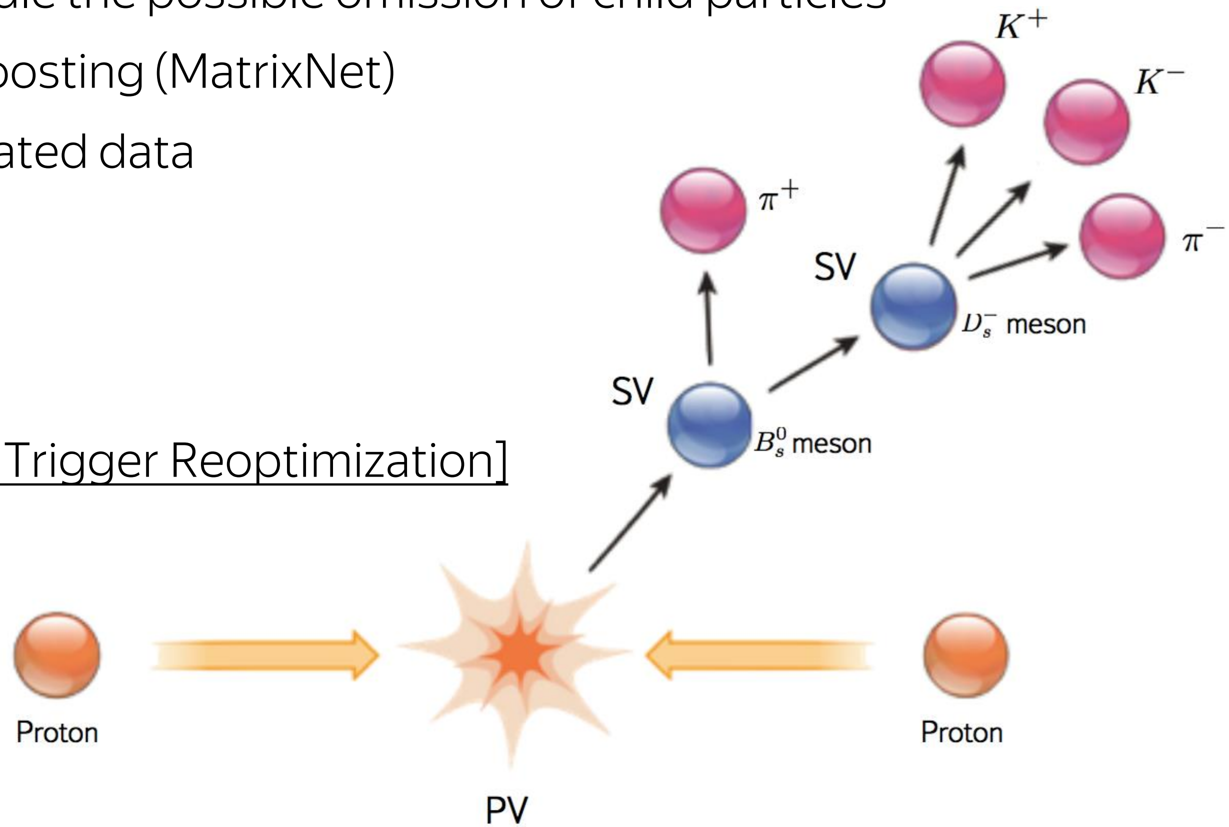
- Multilayer Perceptron reduces the fake track rate from 22% to 14%
- Multilayer Perceptron takes 0.5-2% of run time of forward algorithm, but the whole reconstruction sequence is faster due to less fakes

[Fast neural-net based fake track rejection in the LHCb reconstruction]



Topological trigger (HLT2)

- Selects of any B (and D) decay with at least 2 charged daughters
- Designed to handle the possible omission of child particles
- Uses Gradient Boosting (MatrixNet)
- Trained on simulated data

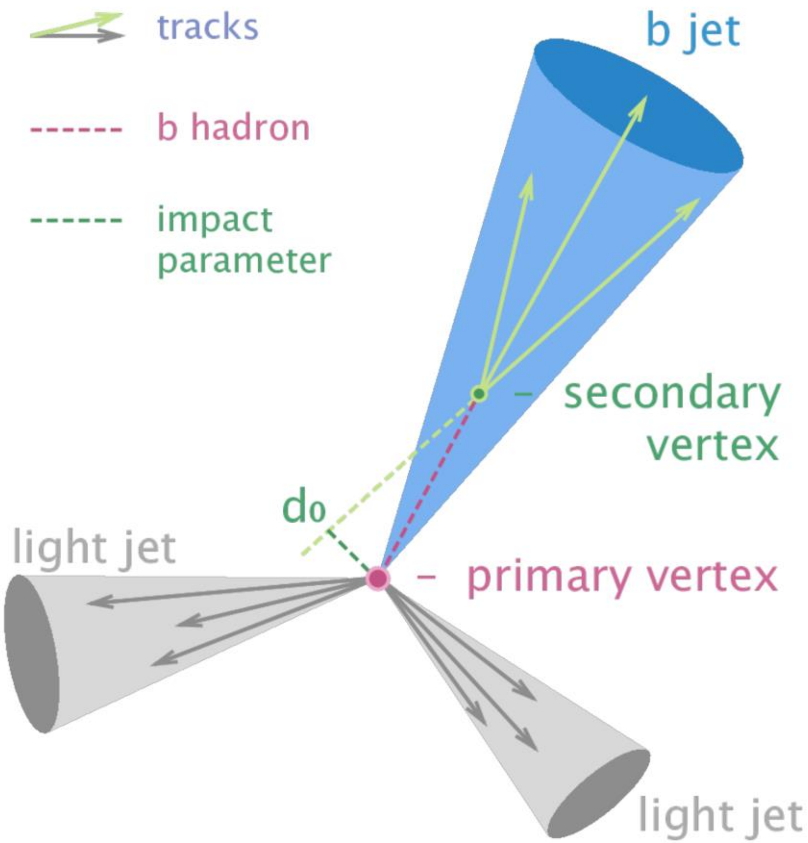


[LHCb Topological Trigger Reoptimization]

Jet tagging

- Identifies b, c and light jets
- Trained on simulated data
- Uses kinematic observables of SVs as inputs
- Uses Boosted Decision Trees
- The efficiency for identifying a b(c) jet is ~65%(25%)
- Probability for misidentifying a light-parton jet of 0.3% for jets

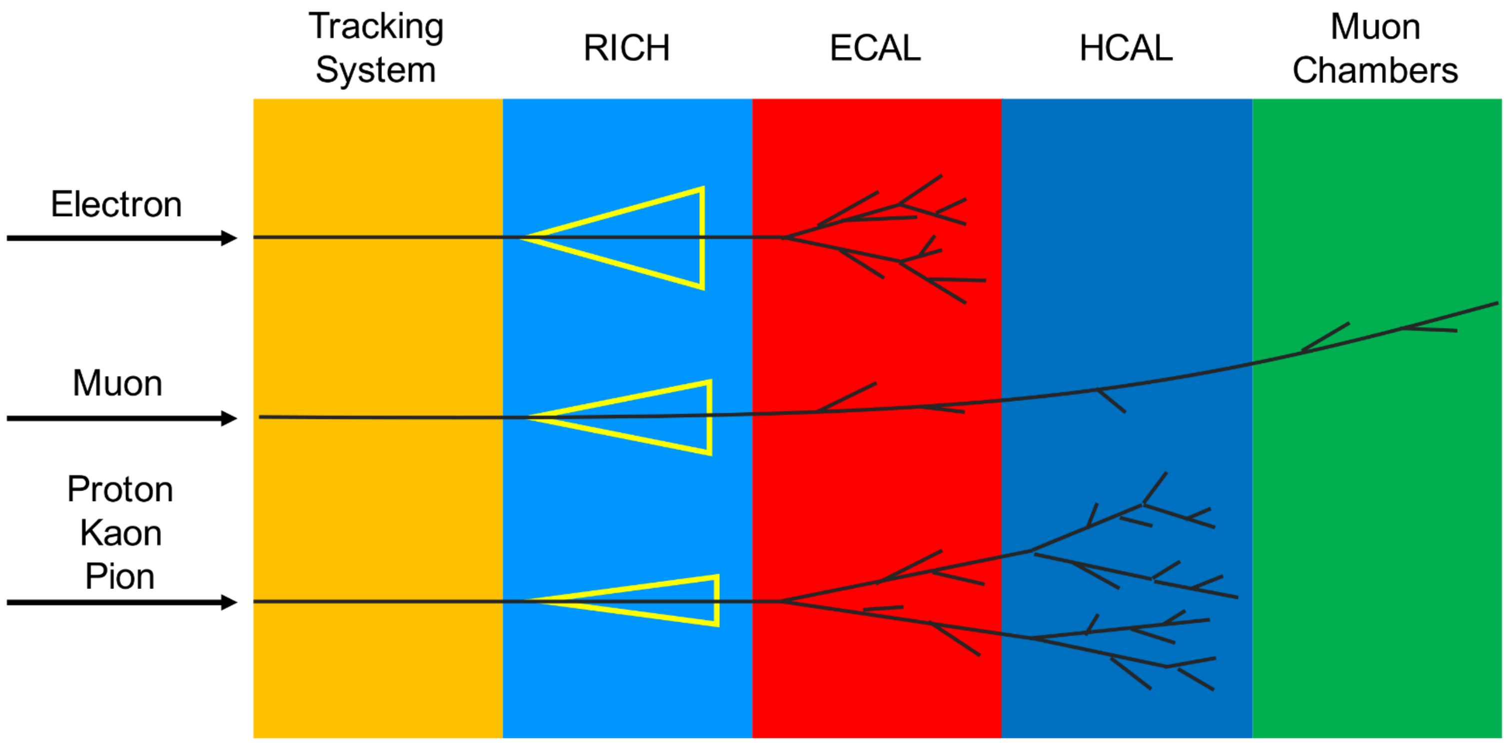
For Run I, $p_T > 20$ GeV, $2.2 < \eta < 4.2$



[Identification of beauty and charm quark jets at LHCb]

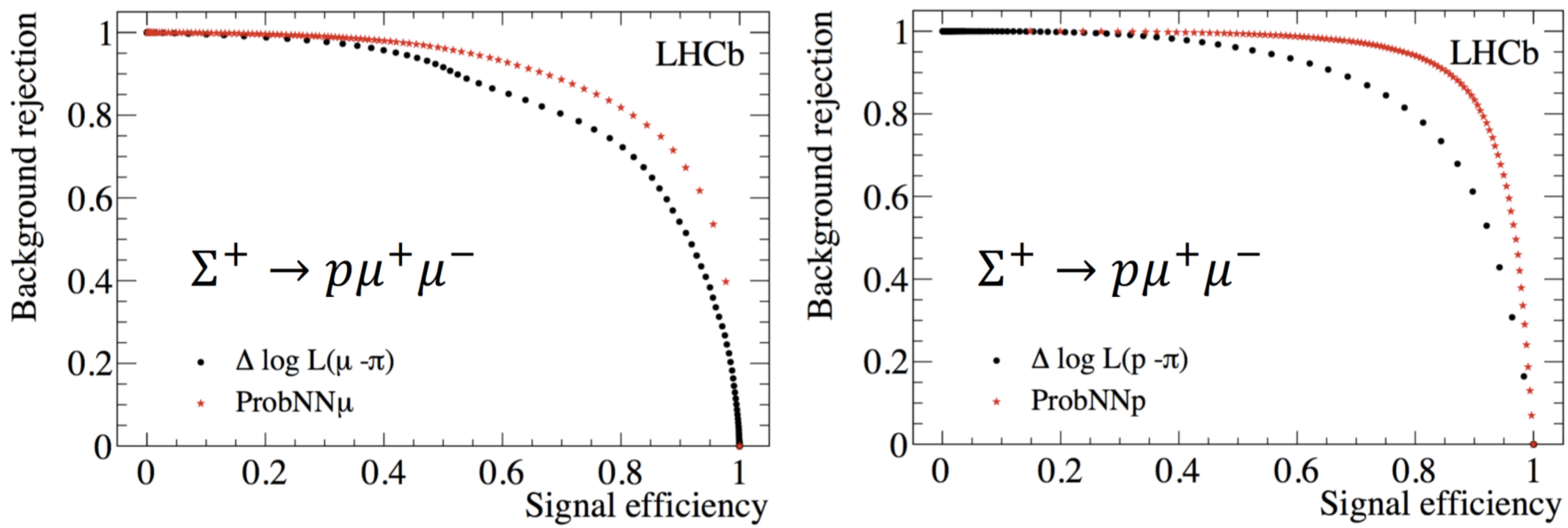
Charged particle identification

Objective: combine information from all subdetectors into a single decision on particle type



Charged particle identification

- Deployed: ProbNN, neural network with one hidden layer
- Each particle type has its own binary neural network trained in one-particle-vs-rest mode



Plots: using data sidebands for backgrounds and Monte Carlo simulation for the signal

9

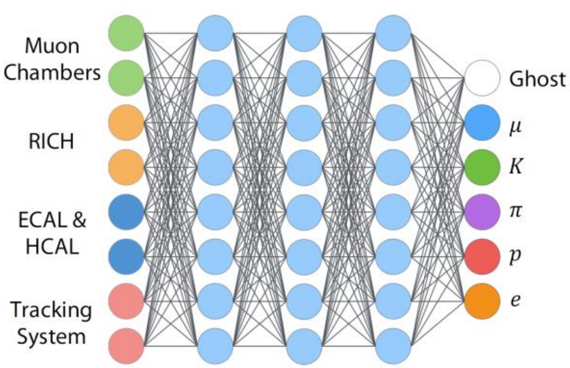
[LHCb detector performance]

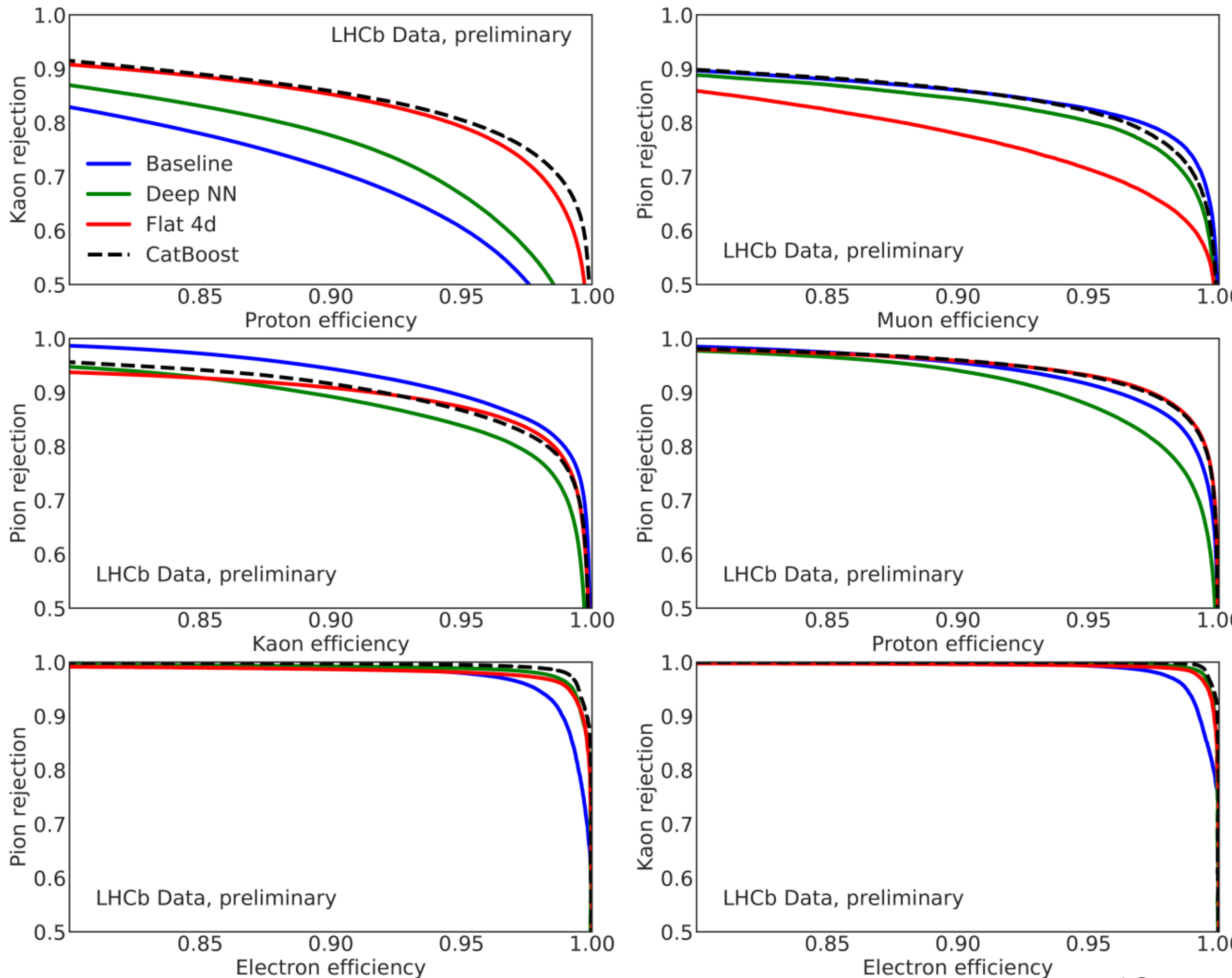
Charged particle identification

Preliminary

[Machine Learning based global particle identification algorithms at the LHCb experiment]

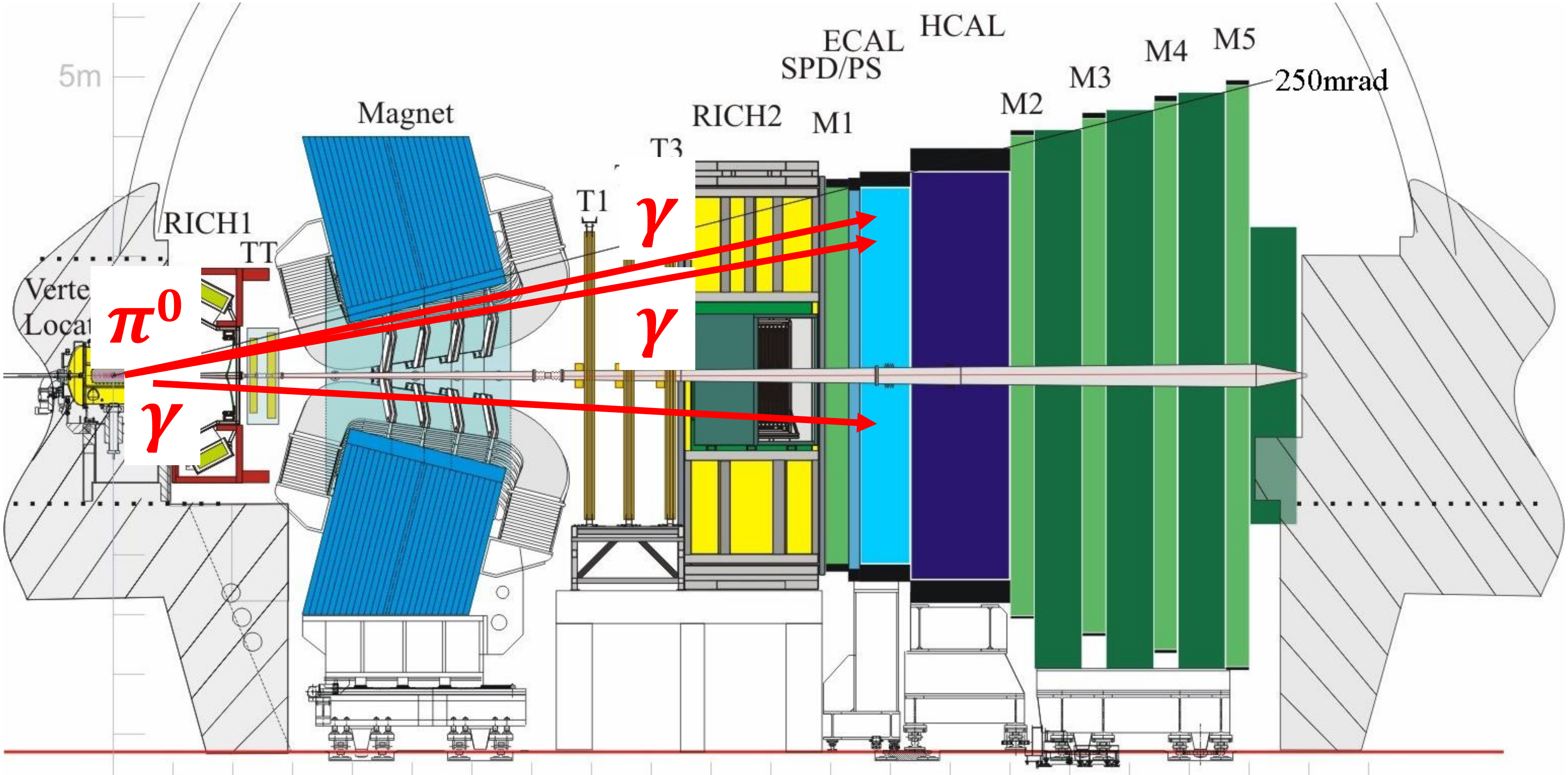
We work on improving Global PID with state-of-the-art algorithms





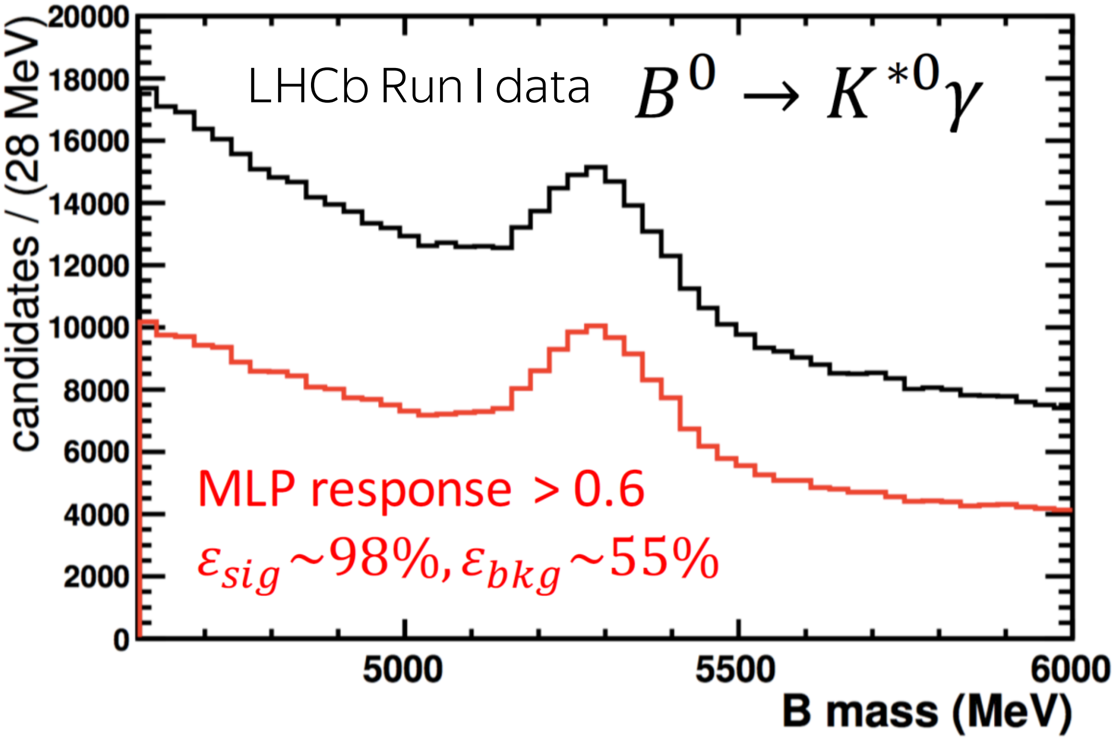
$\pi^0 - \gamma$ separation

- Signal: single photon γ
- Background: photons from $\pi^0 \rightarrow \gamma\gamma$ decay



$\pi^0 - \gamma$ separation: deployed

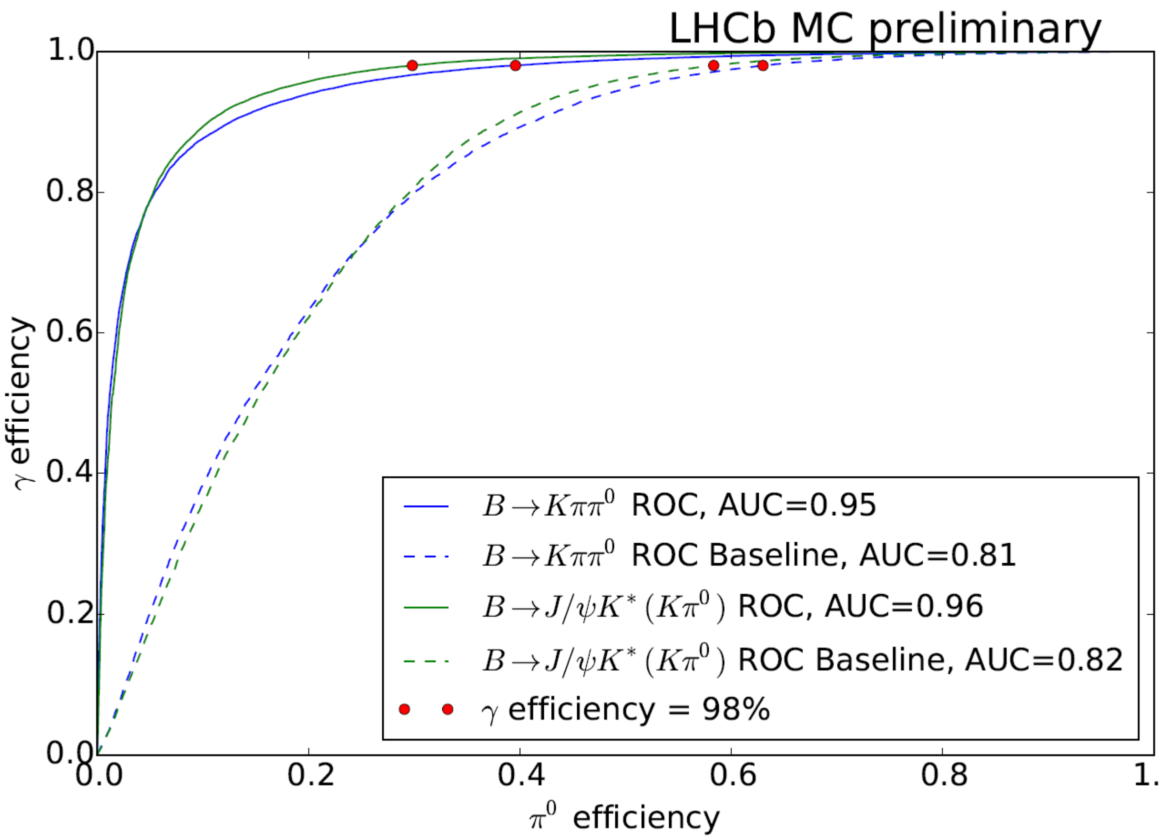
- Clusters shape and symmetry are described by set of features
- 2-layer MLP is trained to separate signal and background clusters



[A tool for γ/π^0 separation at high energies]

$\pi^0 - \gamma$ separation
Work in progress

- Features: responses in 5x5 cell clusters for ECAL and pre-shower detectors
- State-of-the-art gradient boosting algorithms
- Trained on MC

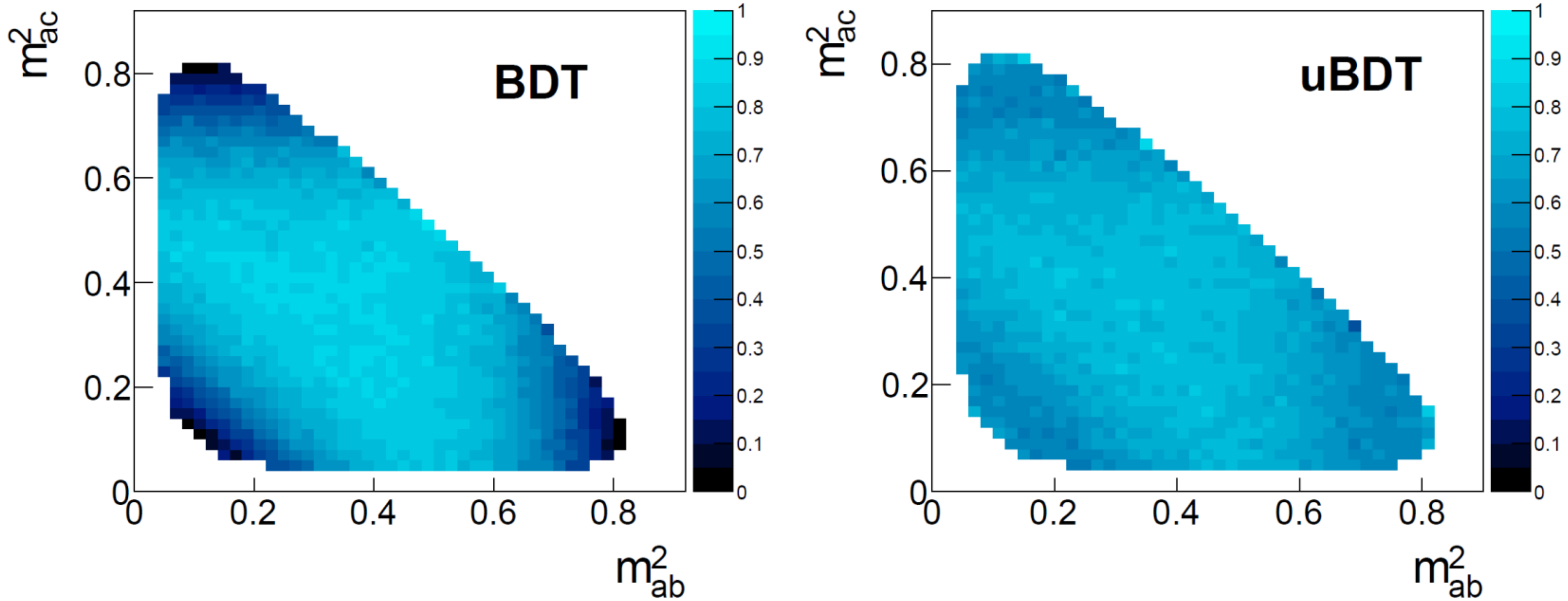


[Boosting Neutral Particles Identification by Boosting Trees: LHCb case]

uBoost

[uBoost: a boosting method for producing uniform selection efficiencies from multivariate classifiers]

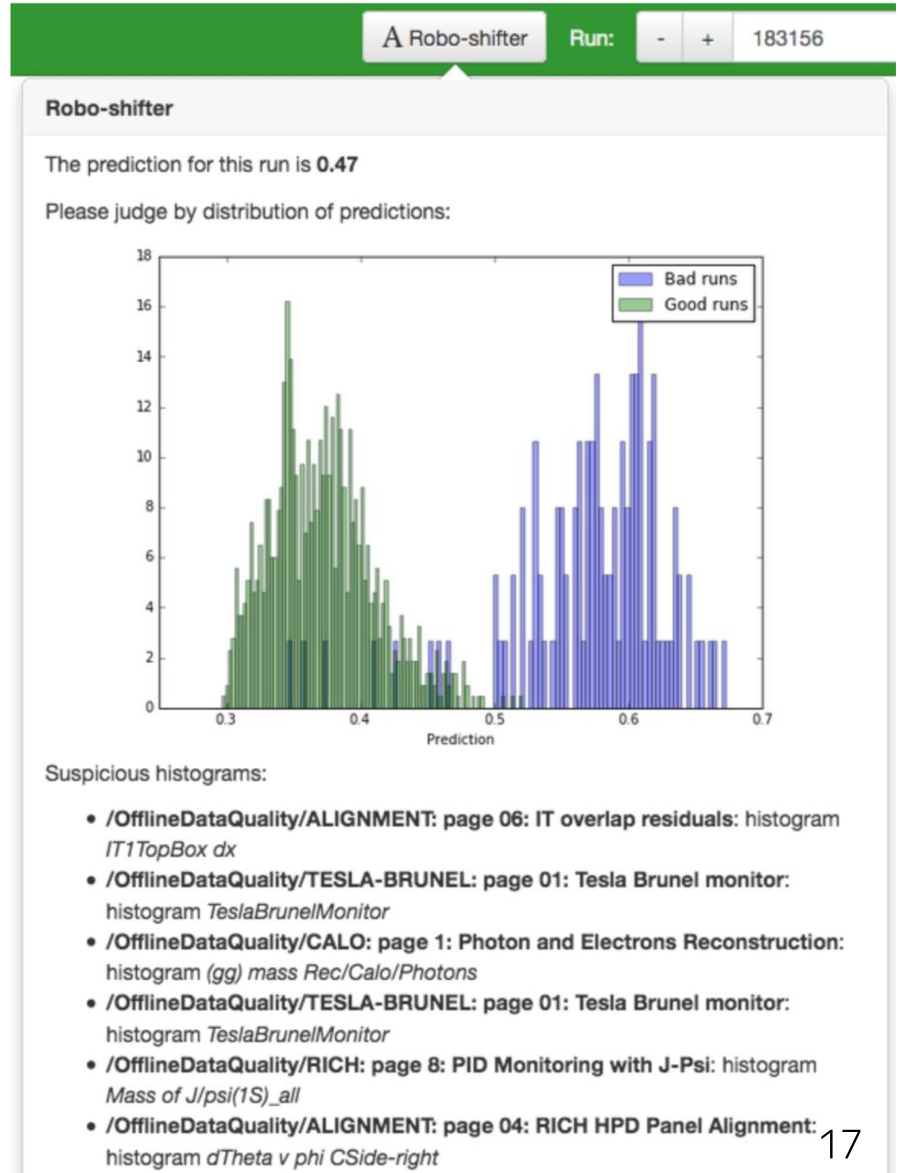
- Classification algorithm, boosting over decision trees
- Uniform selection efficiency with the respect to mass
- Used in some LHCb analyses



Selection efficiency distributions for 70% overall signal efficiency using (left) AdaBoost and (right) uBoost on toy data

Data Quality: RoboShifter

- Aids shifter in data quality monitoring
- Predicts probability of give run being bad
- Provides list of features that contributed the most to the decision
- AdaBoost with trees of depth 1

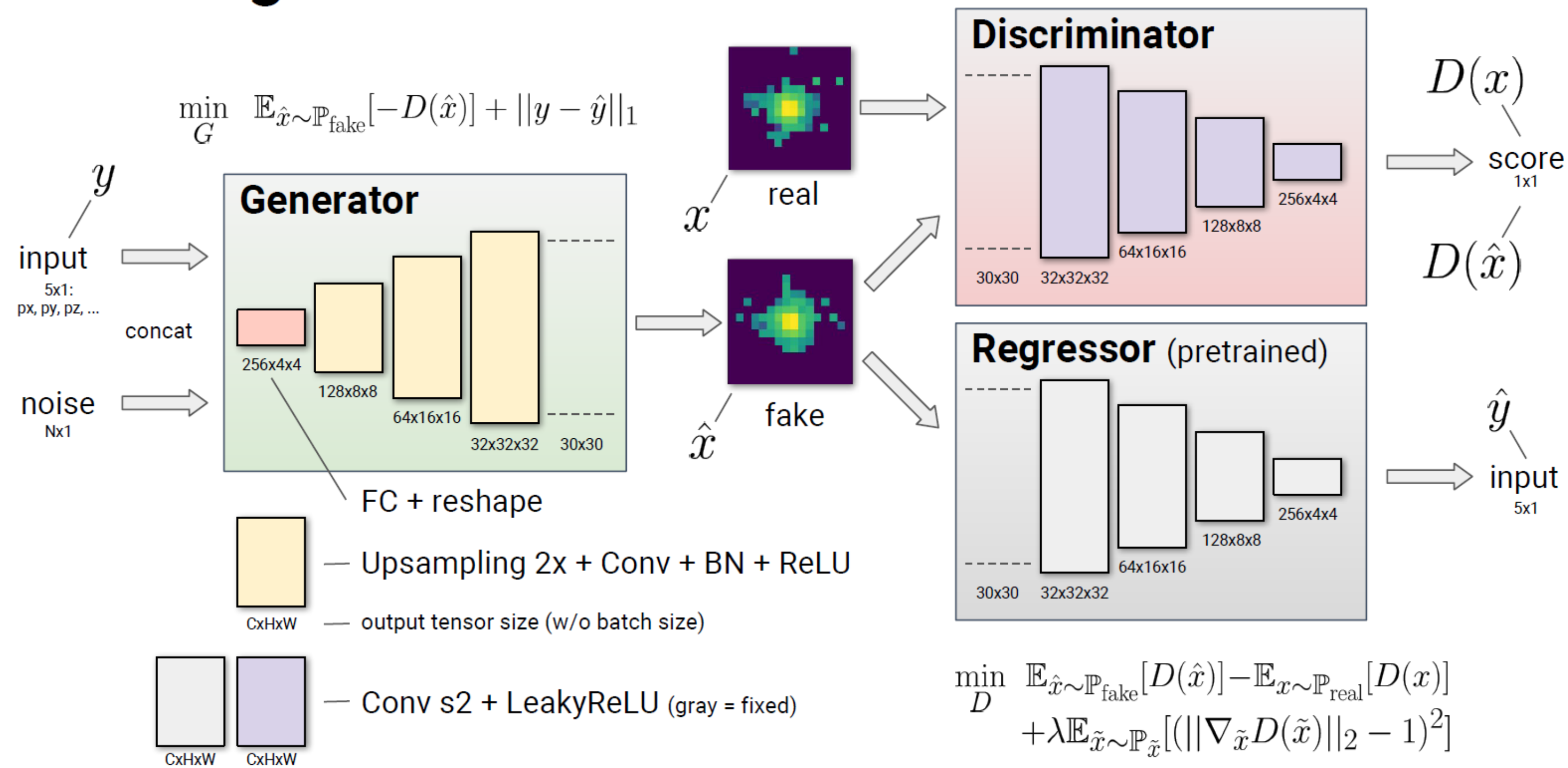


[LHCb data quality monitoring]

Calorimeter Fast Simulation

Work in progress

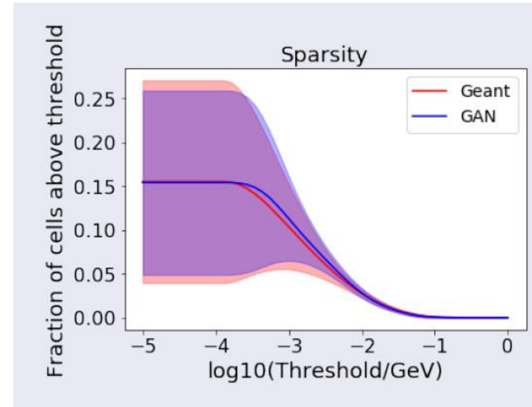
Training scheme

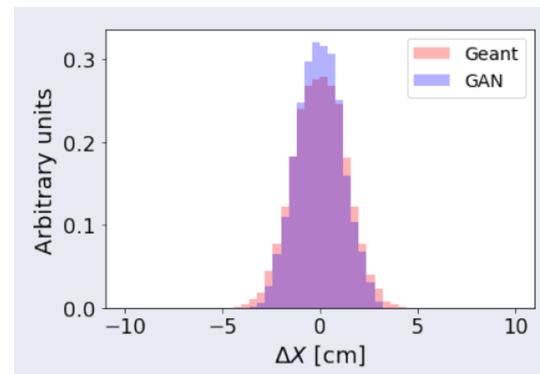


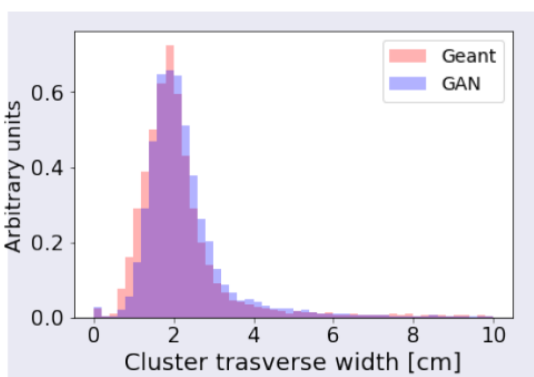
[Generative Models for Fast Calorimeter Simulation: LHCb Case]

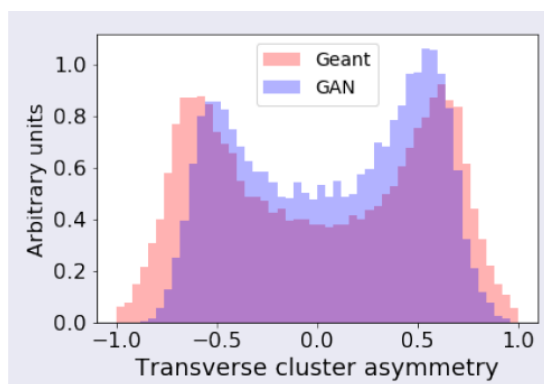
Calorimeter Fast Simulation
Work in progress

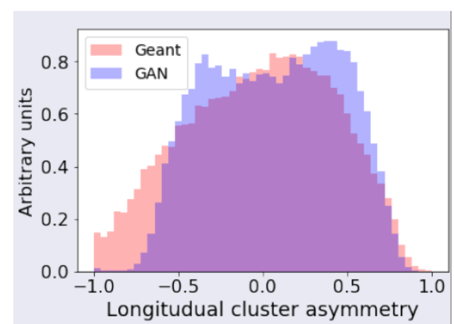
- Outputs raw calorimeter response in 30x30 squares
- Plots: pilot using stand-alone LHCb-like calorimeter, GEANT4 simulation











RICH Fast Simulation
Work in progress

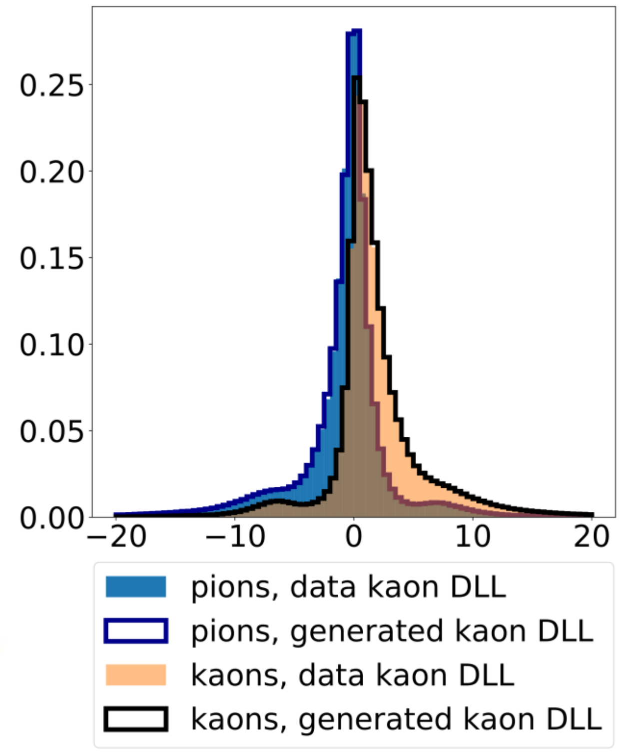
- Trained on real data (calibration samples)
- Directly samples P(PID DLL | kinematics, particle type) bypassing simulation and reconstruction
- Plain Cramer GAN using fully-connected deep NNs
- Pro: simple data structure (just 8D) allows for high fidelity
- Con: parametrization limited to variables included in training

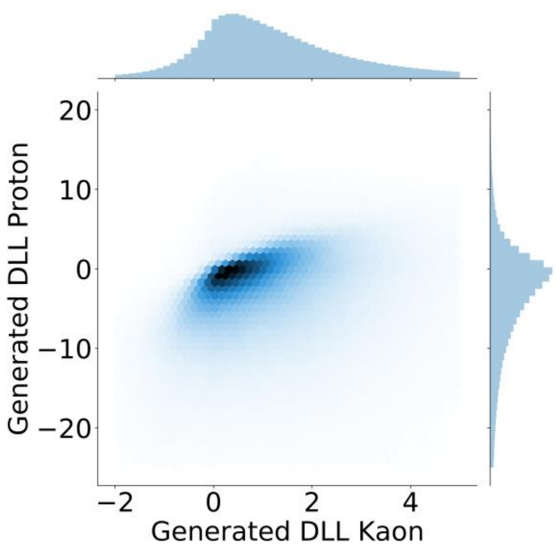
[Cherenkov Detectors Fast Simulations Using Neural Networks]

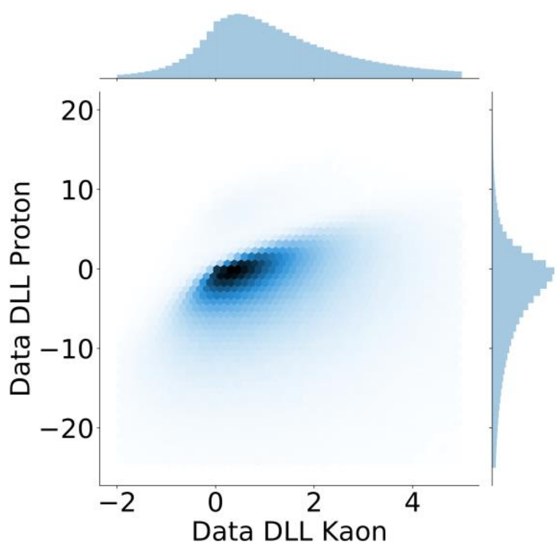
RICH Fast Simulation

Work in progress

- Plots are a pilot study on **BaBar DIRC MC**
- π vs K AUC difference ~0.01
- No public plots for LHCb at the moment, sorry







Summary

- ML is used almost at every stage of LHCb data processing
  15 minutes is not nearly enough to present everything

- ~70% of all data retained are classified by machine learning

- Greatly improved performance while satisfying the robustness requirements of a system that makes irreversible decisions

- *“As an example, achieving the same sensitivity as a recent LHCb search for the dark-matter analogue of the photon without the use of machine learning would have required 10 years of data collection instead of 1” **

- Looking forward to exciting new developments: GANs, LSTMs and other fun things

[LHCb Topological Trigger Reoptimization]

[Search for dark photons in 13 TeV pp collisions]

* [Machine learning at the energy and intensity frontiers of particle physics]

Thank you!

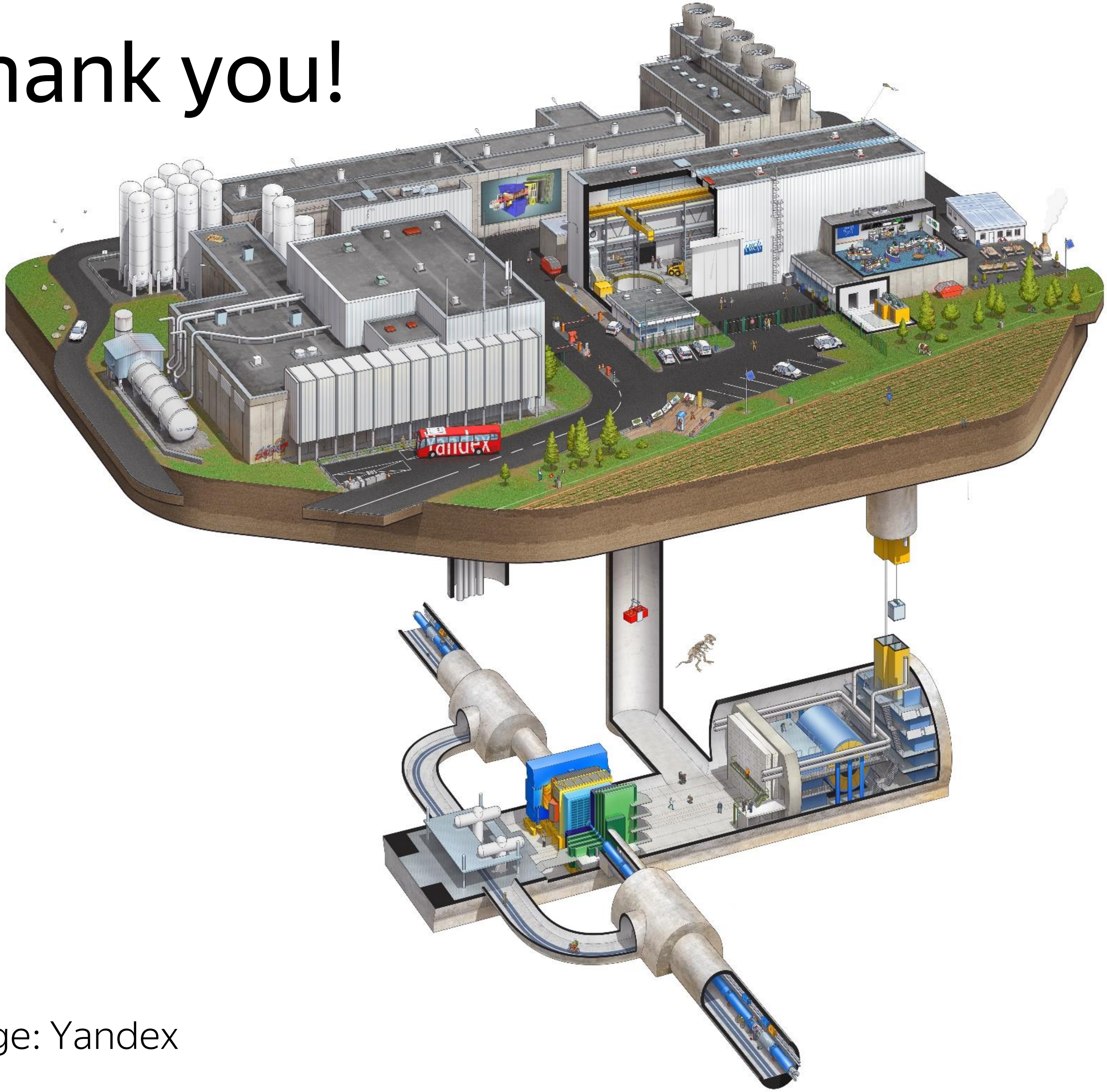
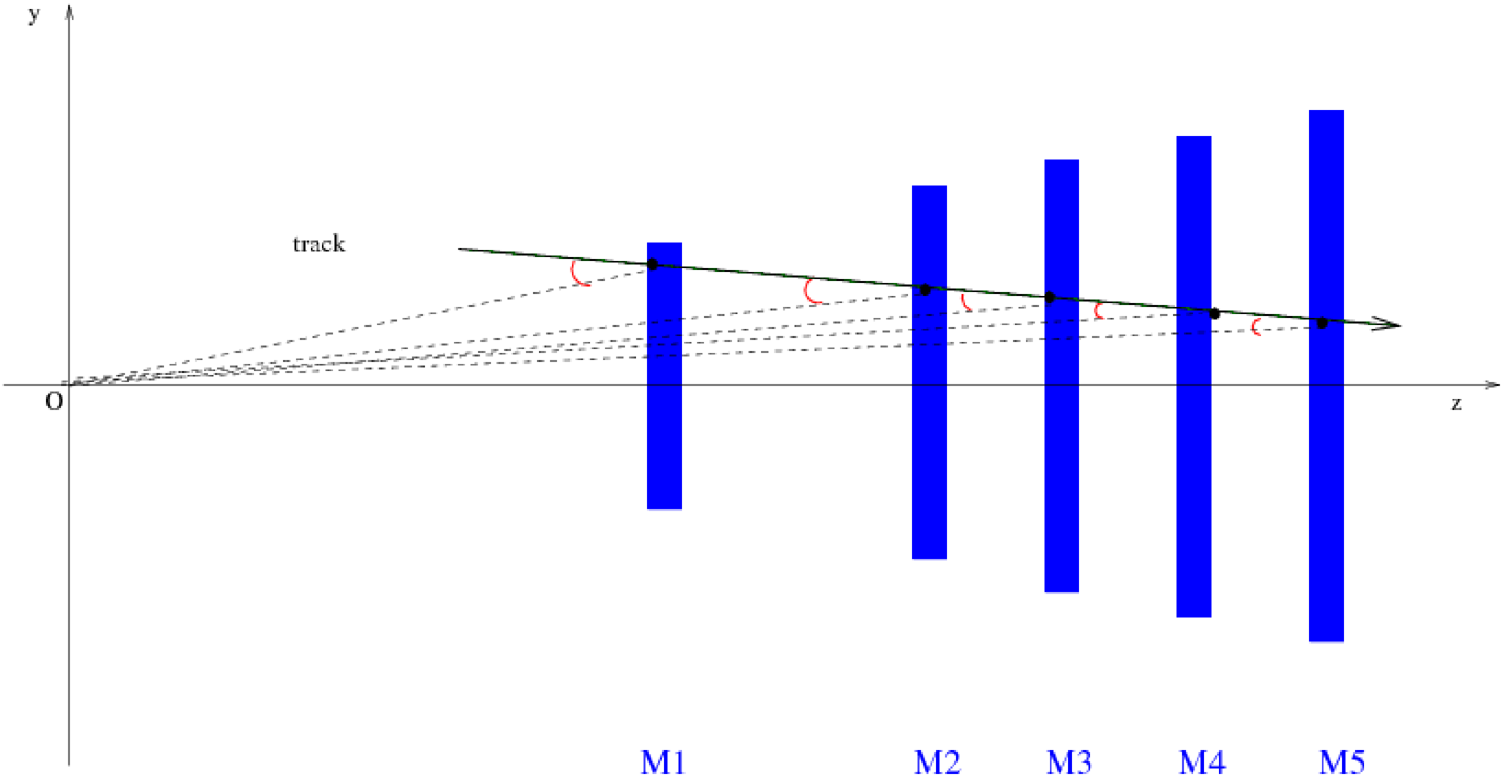


Image: Yandex

Backup

Muon identification

- Muons are distinguished as the particles penetrating through the whole detector and reaching the muon chambers
- Muon ID in nutshell: checking whether there are muon chambers hits associated with the track
- Muon ID HLT1: IsMuon, uses multiple scattering theory to define a cone around the track checks whether there are hits in it. ~98% efficient in Run II, will fall to ~90% after Upgrade

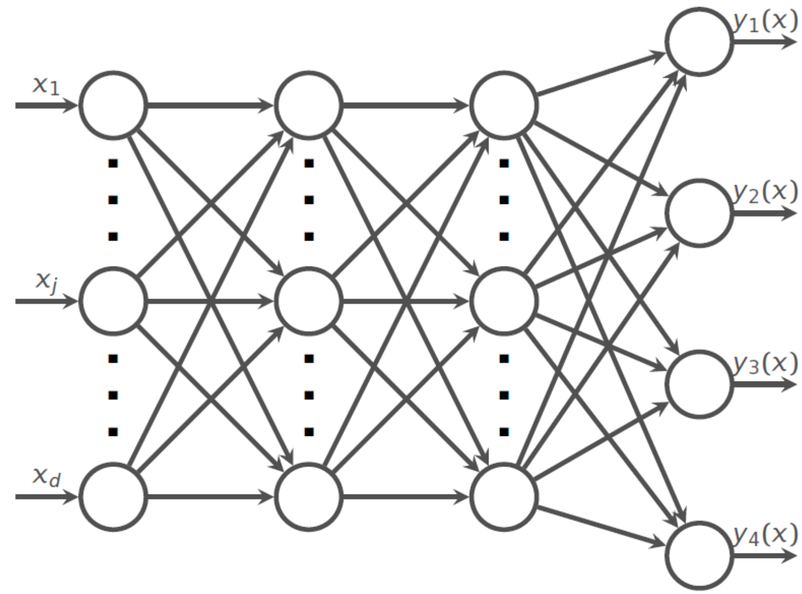


ML for Muon Identification
Work in progress

- We develop MuonID based on gradient boosting
- To be run after IsMuon
- Training on real data: calibration samples
- Features: hits residuals, timing, technical information
- No public plots (yet)

NN for Upgrade VELO
Work in progress

- Inputs **x** are seed & target layer (r, ϕ) coodinates
- *One* seed (r, ϕ) pair and *several* target (r, ϕ) pairs
- Outputs are target index (class) probabilities
- Network topology & dimension of **y** not *a priory* ovious
- Trained on labeled data from *full* simulation
- Train one classifier for each pair of layers to be connected



Efficient data preparation an book-keeping required.

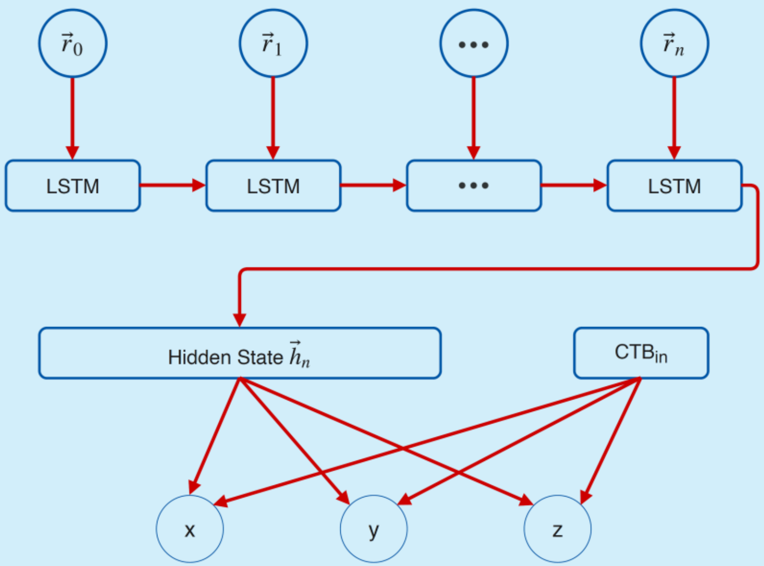
[Novel Approaches to Track & Vertex Reconstruction in the Upgraded LHCb VELO]

LSTM for Upgrade VELO

Work in progress

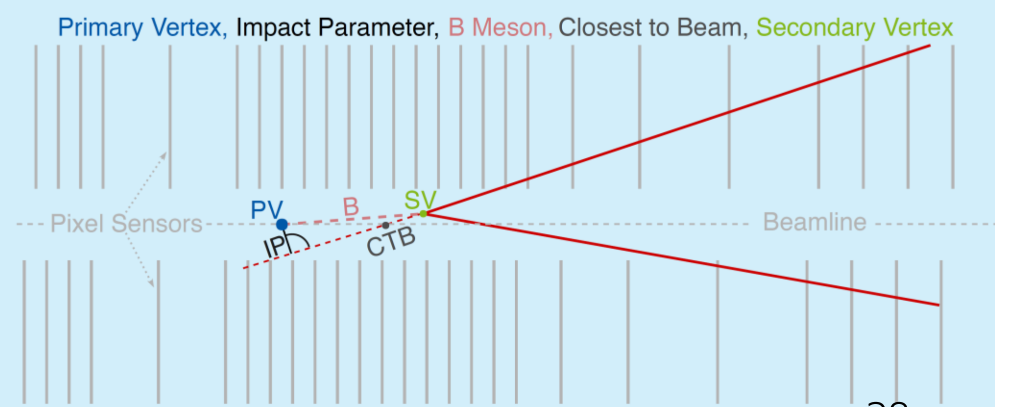
Track Reconstruction in the Vertex Locator

1. Reconstruct tracks via track forwarding from the outer to the inner region
2. Simplified Kalman Filter to account for multiple scattering and predict a track's closest to beam (CTB) position. Problem: missing momentum information
3. Idea: use a special Neural Network architecture to handle variable number of hits in a track



Model architecture to predict CTB position

[New approaches for track reconstruction in LHCb's Vertex Locator]



Outline of the VELO including an exemplary B decay