Машинное обучение в физике частиц

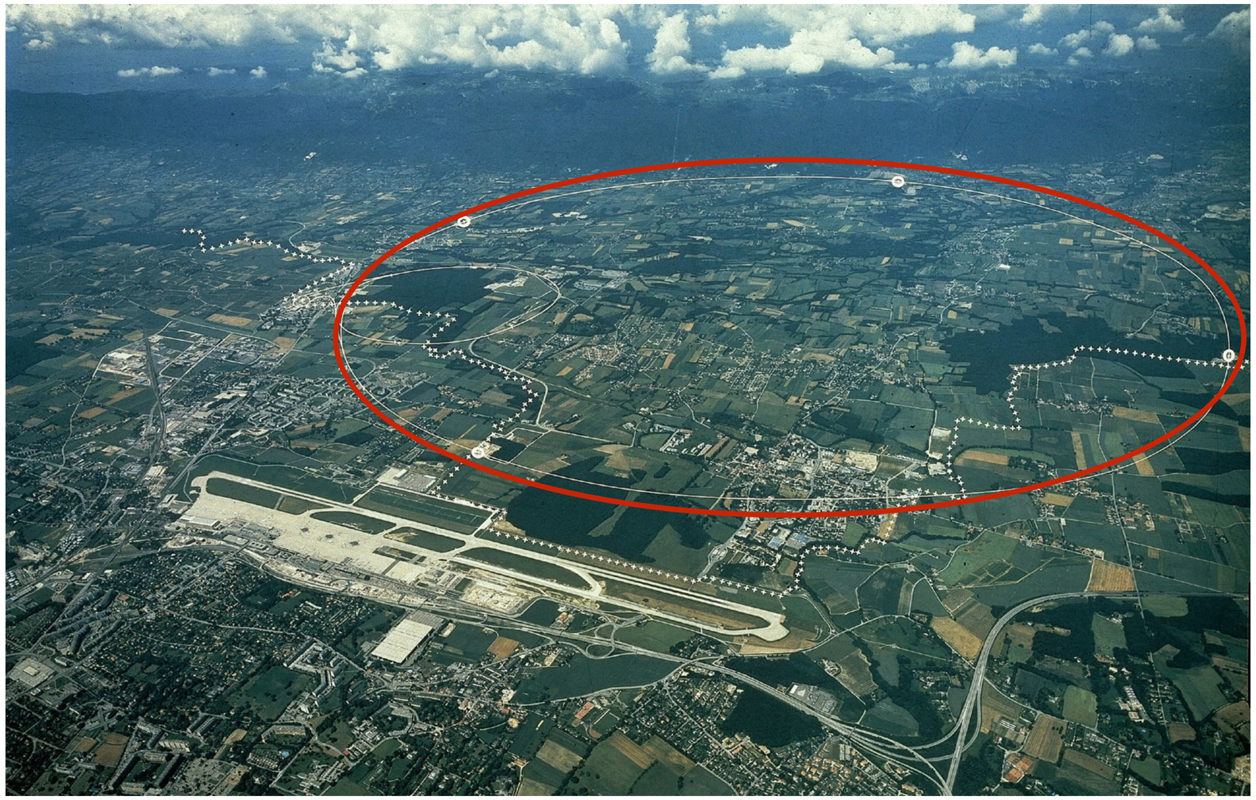
Денис Деркач
Лаборатория методов анализа больших данных (Lambda)
Национальный исследовательский университет
«Высшая школа экономики»

Race for Knowledge

With current technologies the energy depends on the linear size of collider.

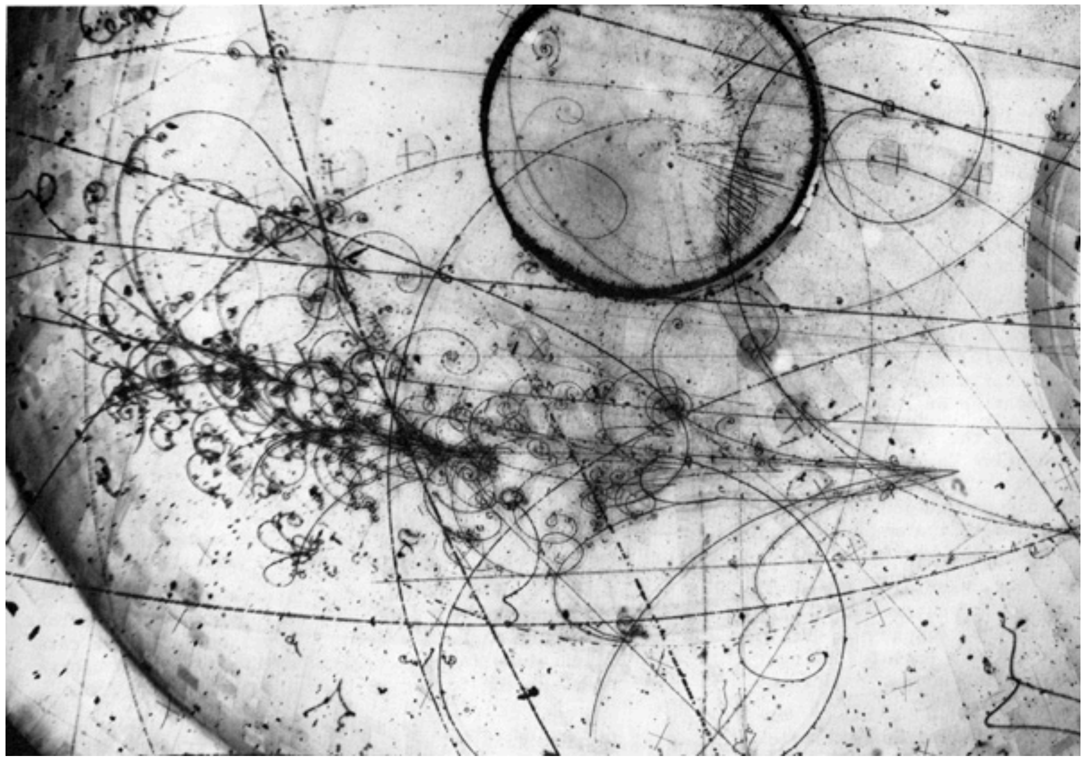
For better sensitivity we need more collisions.

Quantity	Number
Circumference	26 659 m
Dipole operating temperature	1.9 K (-271.3°C)
Number of magnets	9593
Number of main dipoles	1232
Number of main quadrupoles	392
Number of RF cavities	8 per beam
Nominal energy, protons	6.5 TeV
Nominal energy, ions	2.56 TeV/u (energy per nucleon)
Nominal energy, protons collisions	13 TeV
No. of bunches per proton beam	2808
No. of protons per bunch (at start)	1.2×10^{11}
Number of turns per second	11245
Number of collisions per second	1 billion



Four main detectors installed at Large Hadron Collider are LHCb, ALICE, CMS, ATLAS.

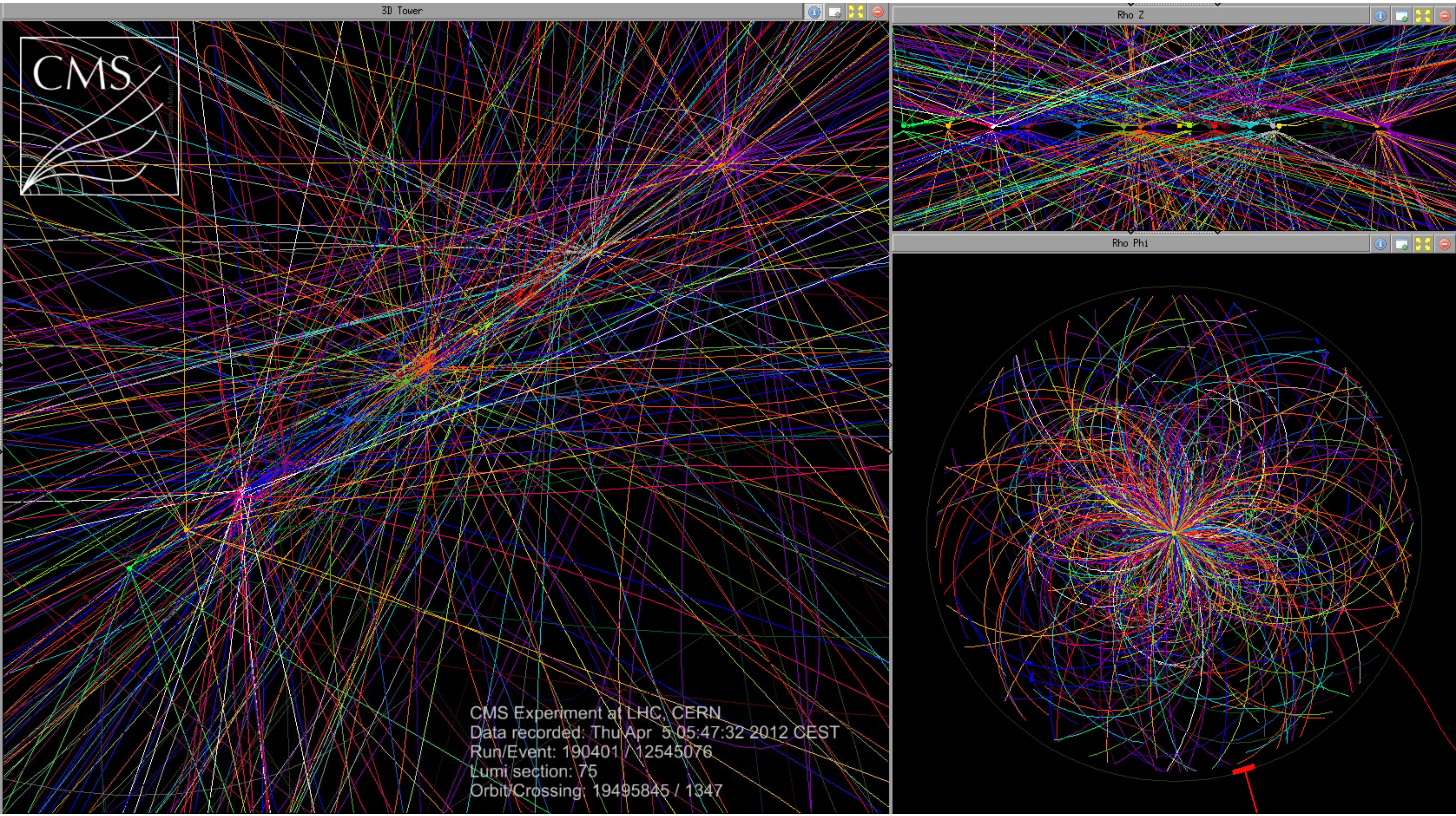
A typical discovery procedure 50 years ago



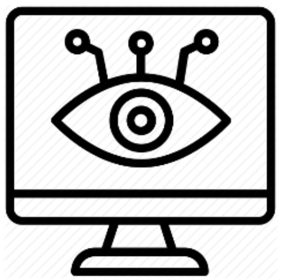
Camera was triggered by a person and than developed and analysed by another person

Fermilab and LBNL Image Database 96602983

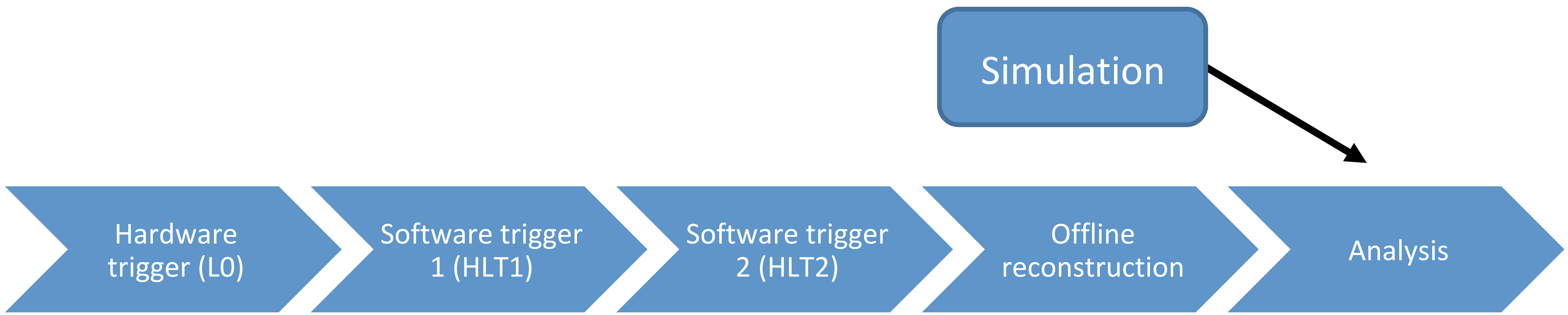
A typical CMS event in proton-proton collision



Example (LHCb) Run II data flow



Monitoring



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

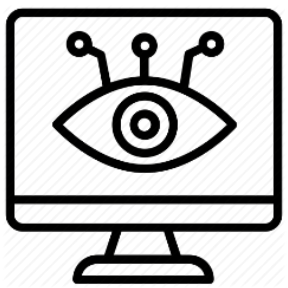
Software trigger 1 (HLT1)
- Real-time
- Selects events with muons
- Selects event s with high p_t tracks
- Selects events with high IP

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

Analysis
- Decay-specific user-specific

https://pos.sissa.it/321/226/pdf

Example (LHCb) Run II data flow



Monitoring

ML

WIP ML

ML

ML

Simulation

ML

Hardware trigger (L0) → Software trigger 1 (HLT1) → Software trigger 2 (HLT2) → Offline reconstruction → Analysis

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

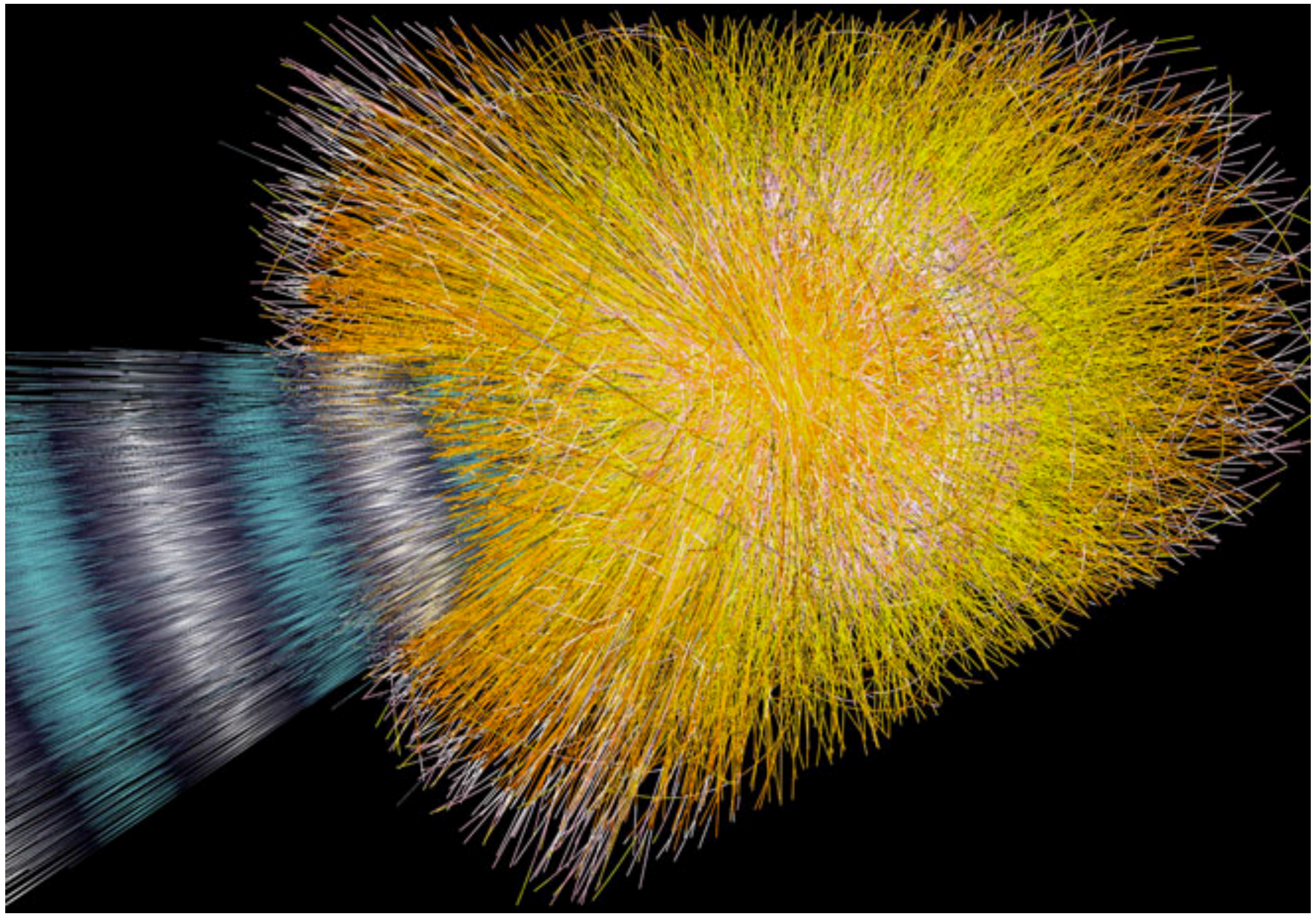
Analysis
- Decay-specific user-specific

https://pos.sissa.it/321/226/pdf

Information processing challenge

Frequency problem

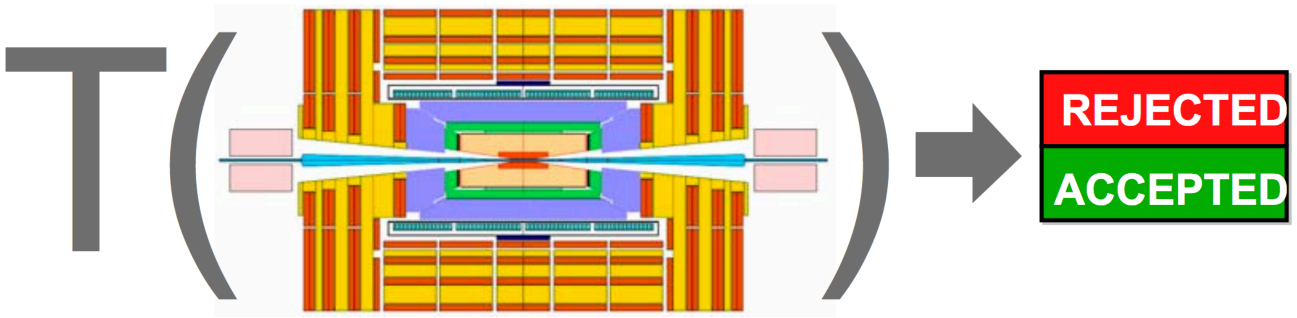
Typical ALICE event in lead-lead collision



An event is occurring 40M times per second, with a typical size of 1 MB, this makes around 40 TB/s of information.

We thus need a fast, precise and reliable to analyse the information online in search for a “good” event.

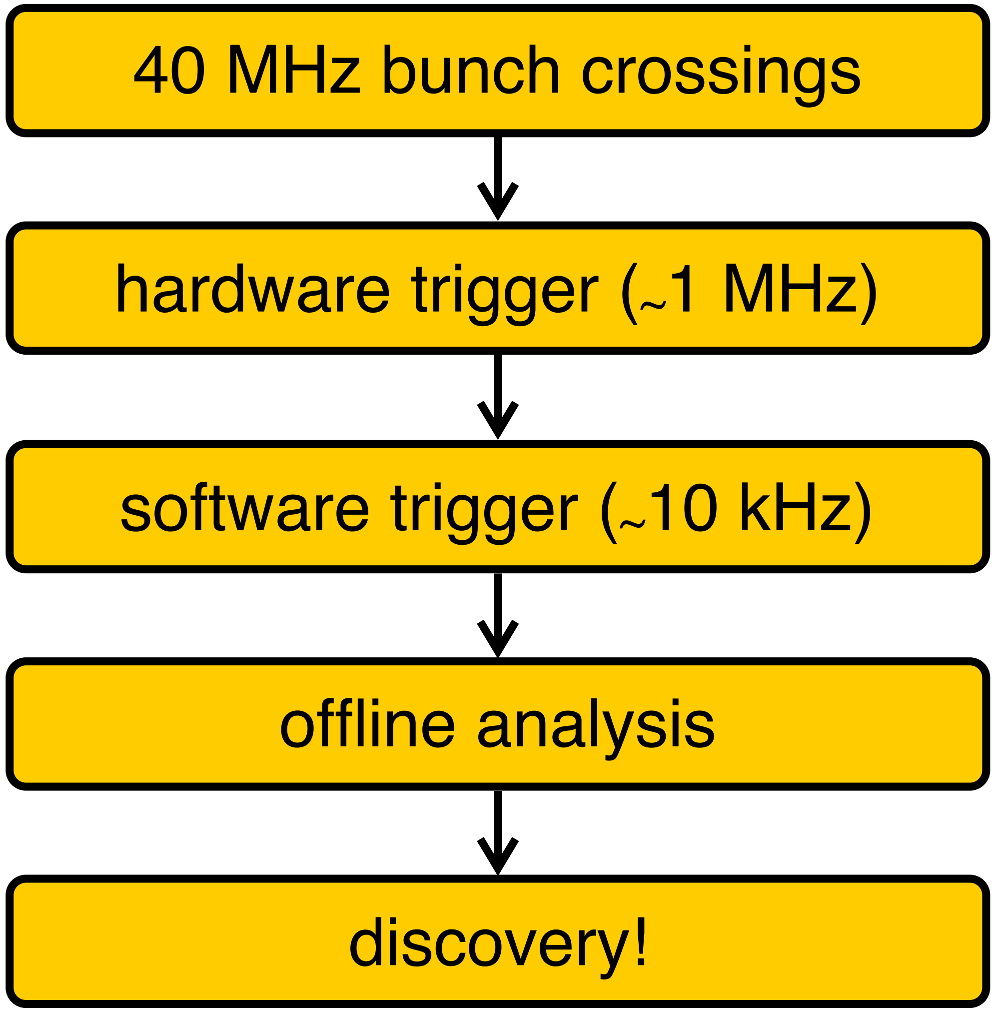
We need a trigger system.



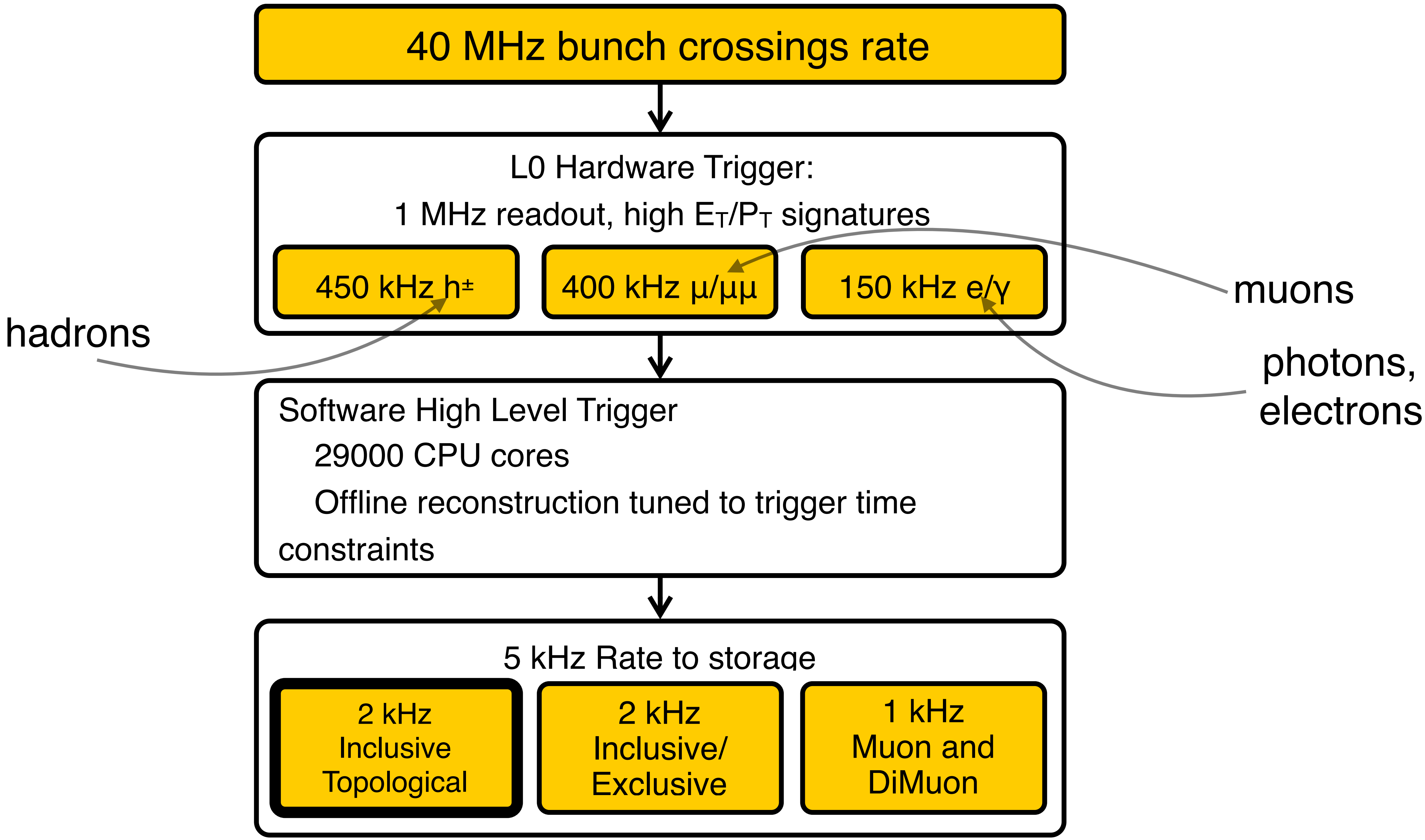
Trigger system in HEP experiments

The goal is to select interesting events (proton-proton collisions) based on detailed online analysis of measured physics information.

Trigger system often consists of two stages: hardware and software.

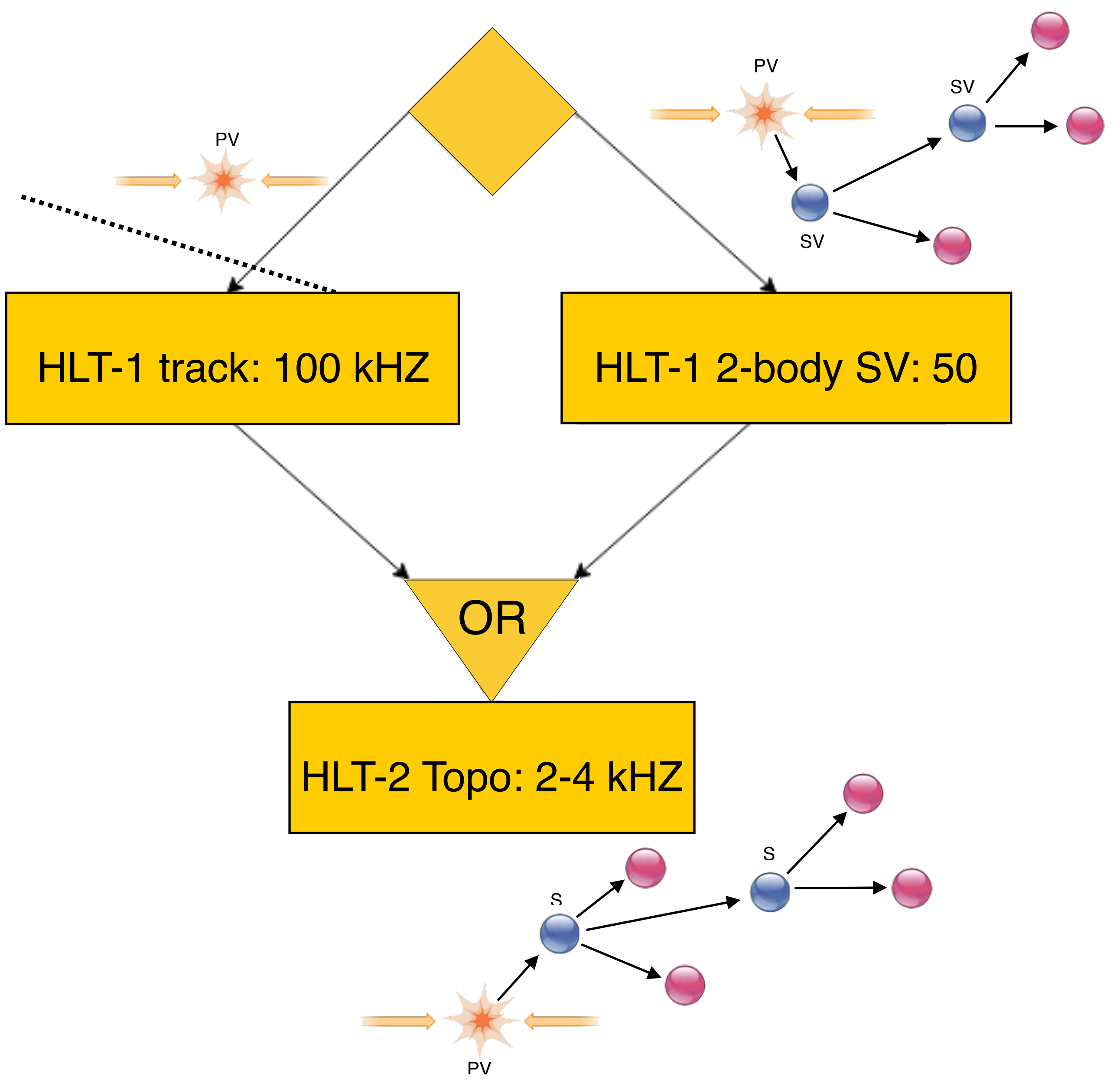


LHCb trigger



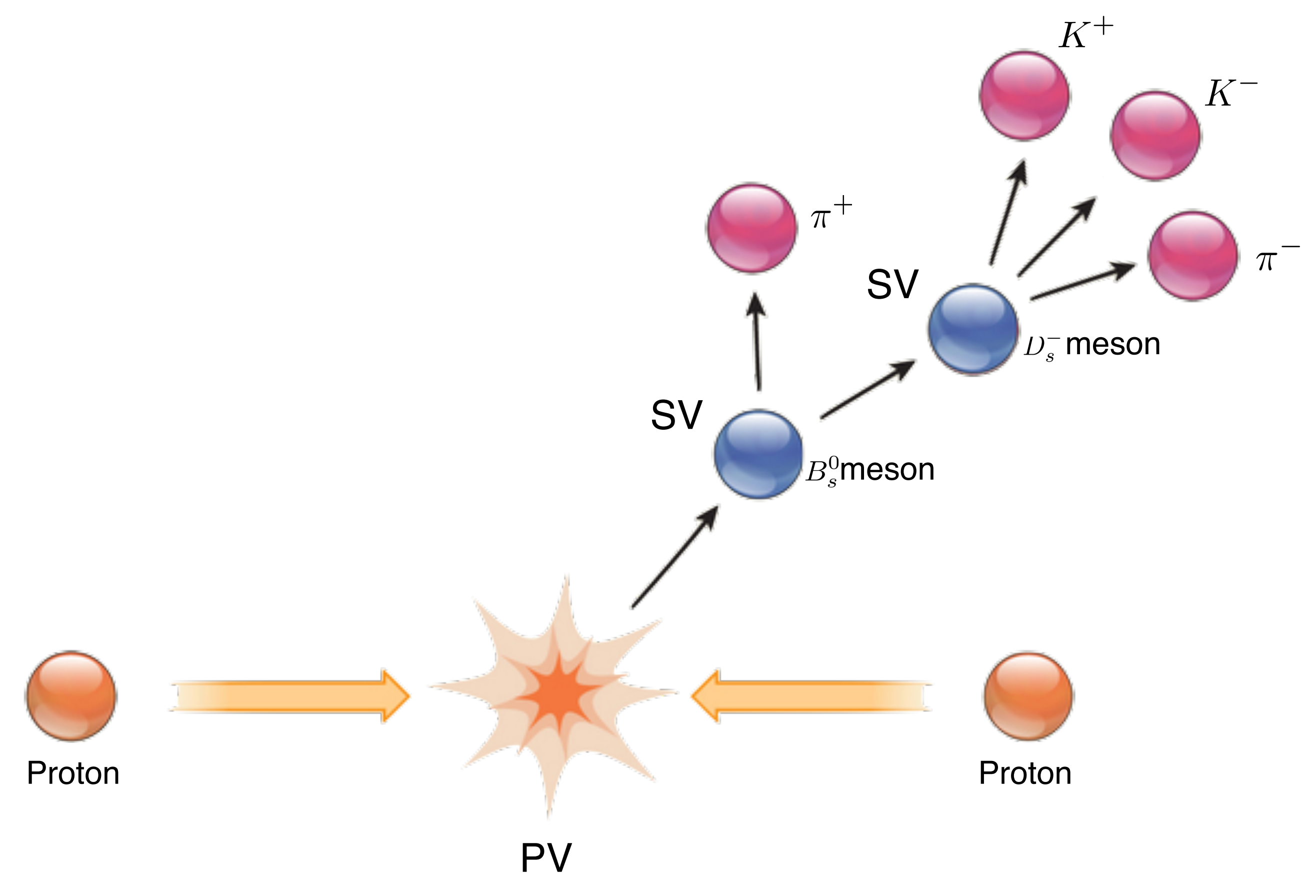
LHCb topological trigger

- 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



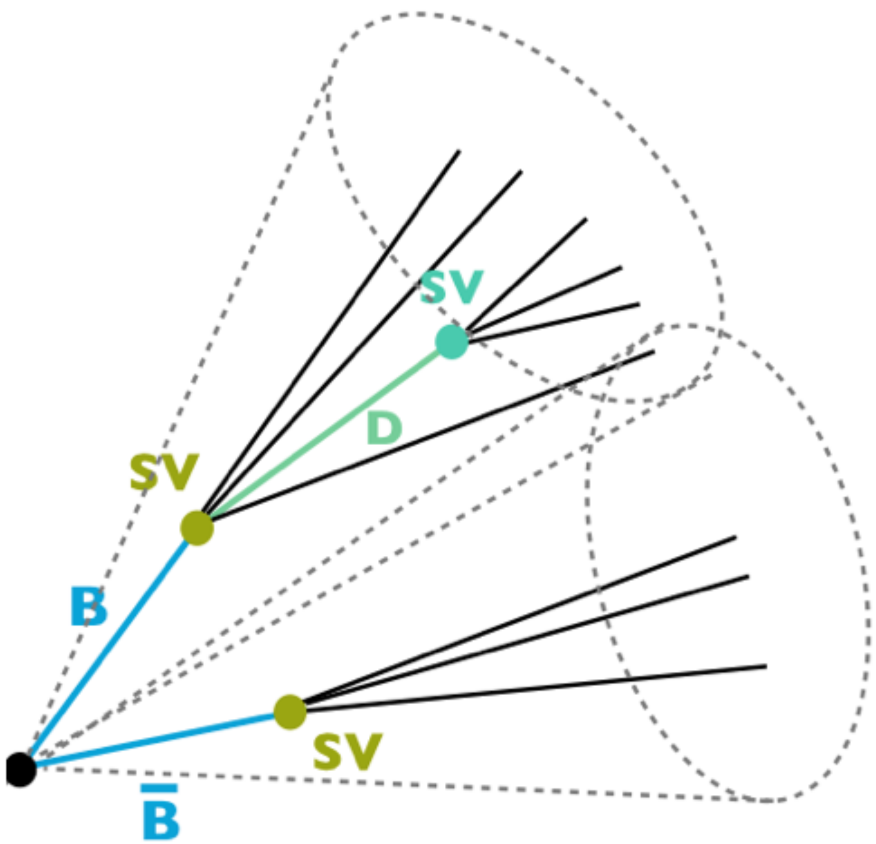
Interesting event

- Primary vertex (PV) is a collision point
- Secondary Vertex (SV) is a point where an unstable particle decayed, this particle is associated with SV
- SV is called interesting if it is associated with the decay of particle under study
- Event is interesting if it contains at least one interesting secondary vertex (SV)



LHC data

- Sample: one proton-proton collision
- Binary classification: event is interesting or not
- Event consists of:
  1. tracks (track description)
  2. secondary vertices (SV description)
- Questions:
  1. How to describe event in ML terms?
  2. How to train model on such samples?



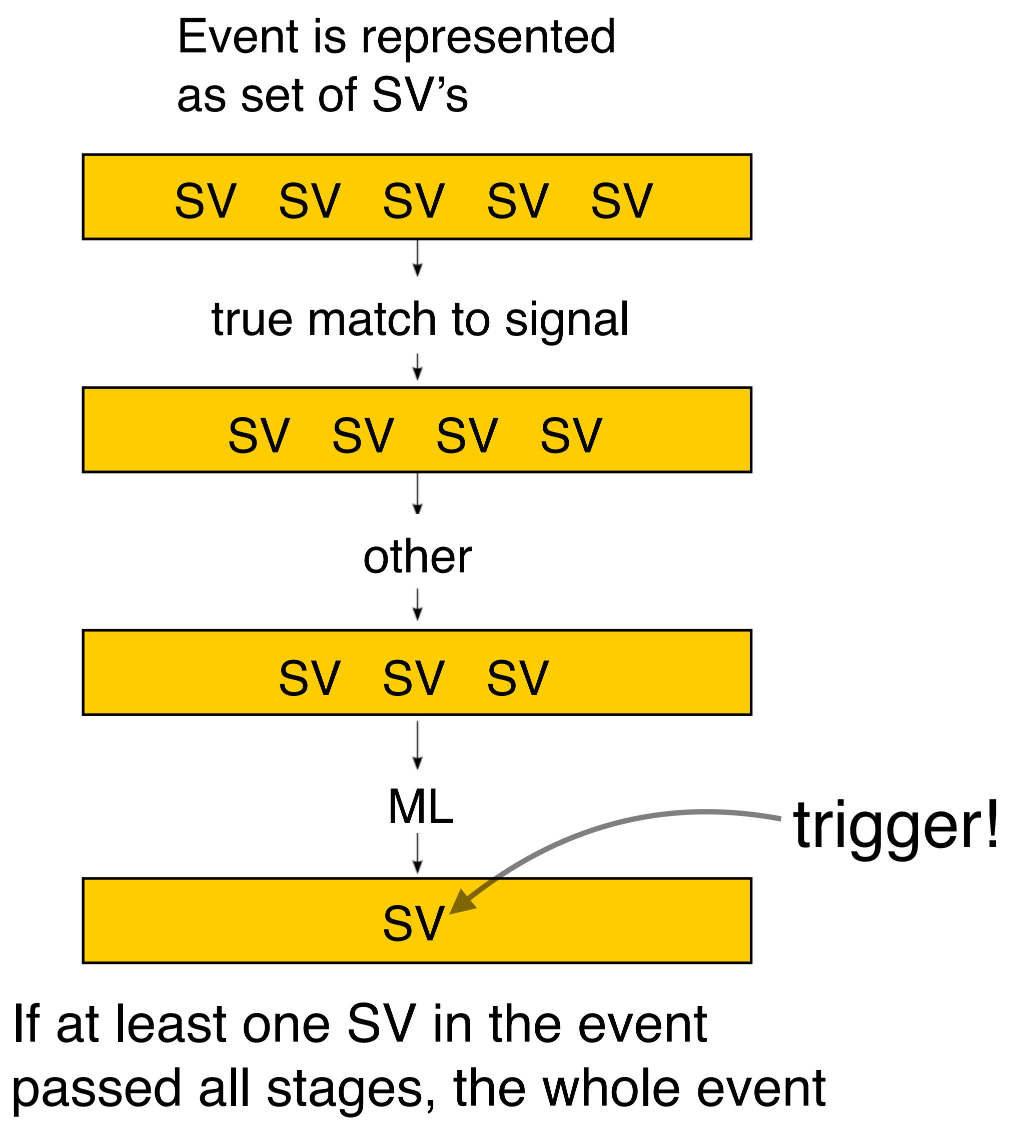
Machine learning problem

Sample is a set of SVs for all events

Features: momentum, mass, angles, impact parameter.

Task: separate "signal" signatures of B-mesons and D-mesons decays from “background".

P("signal" decay) < 10^{-4}

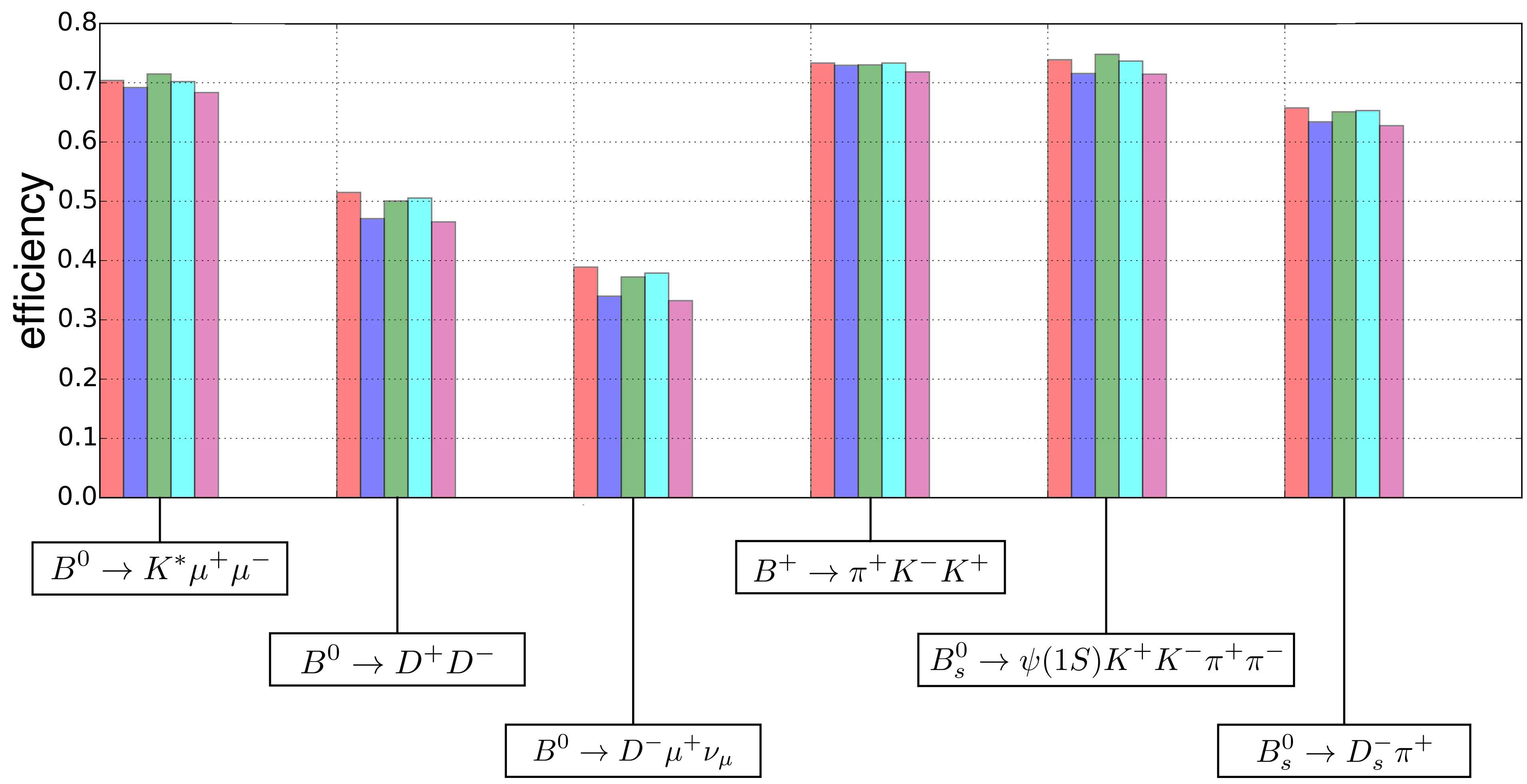


Machine learning problem

- "Signal":
  - Monte Carlo sample is simulated for various types of interesting events (different decays)
- "Background":
  - generic proton-proton collisions are simulated during a small period of time
- Imposed restriction:
  - output rate is fixed (2.5 kHz), thus, false positive rate (FPR) for events is fixed
- Goal:
  - get the highest efficiency for each type of signal events at given FPR

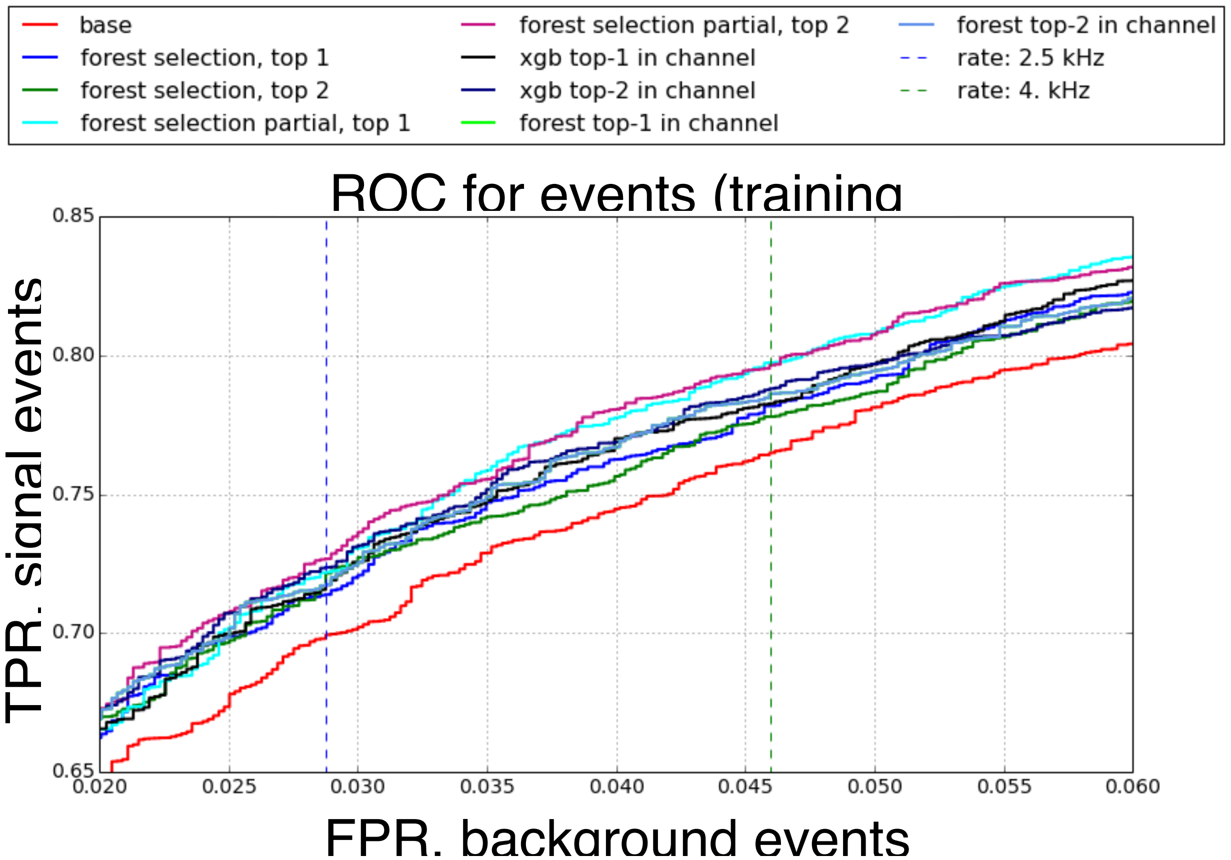
How to measure quality?

- Looking at the quality for each decay separately is not a way
- Need to have aggregative metric to measure quality



ROC curve, computed for events

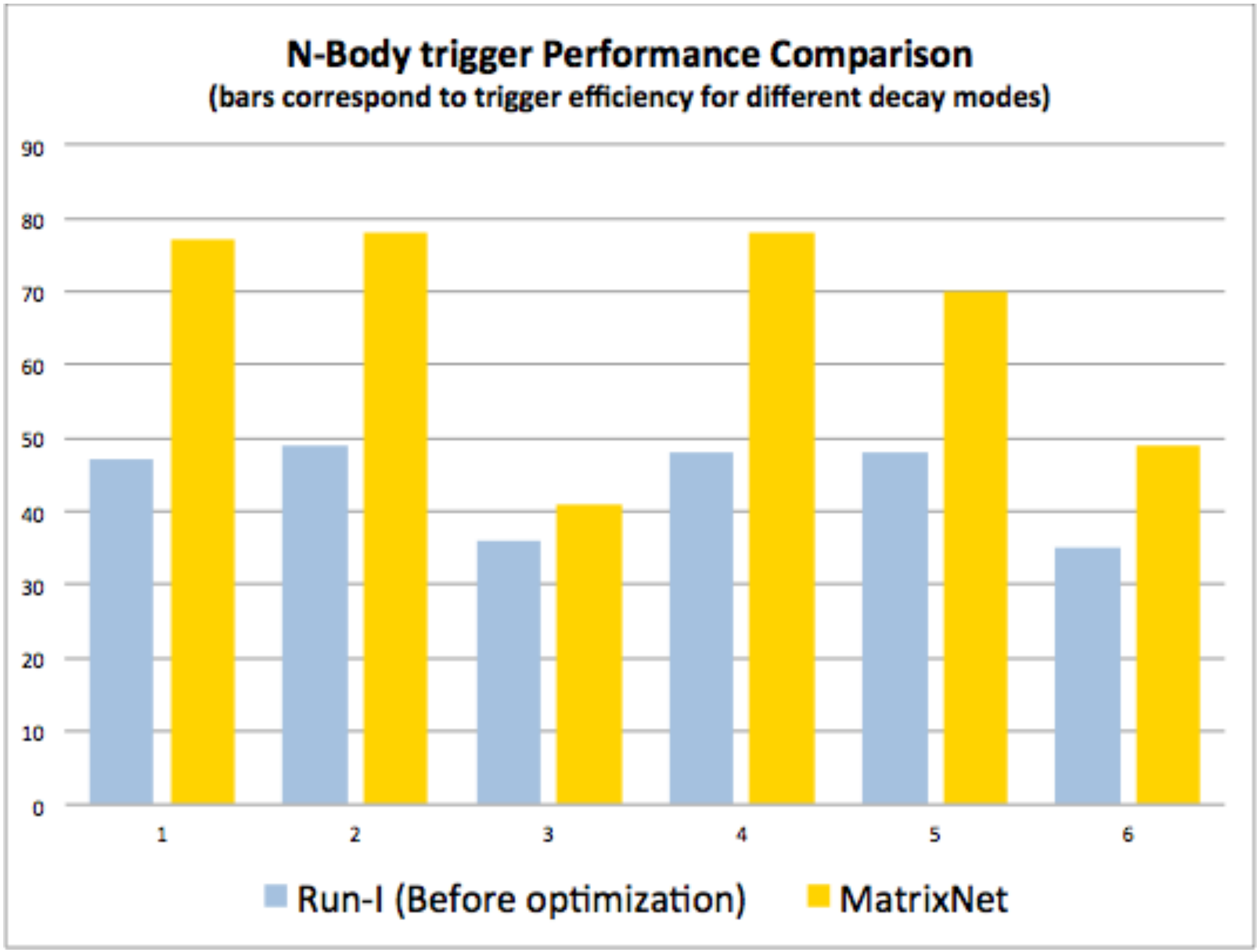
- Output rate = false positive rate (FPR) for events
- Optimise true positive rate (TPR) for fixed FPR for events
- Weight signal events in such a way that decays have the same sum of weights
- Optimise ROC curve in a region with small FPR



Topological trigger results

50% improvement implies that the same physics results would be collected during 3 years with Run I model and during 2 years with new model.

Currently, the model is run at the LHCb experiment online, collecting 60% of data.

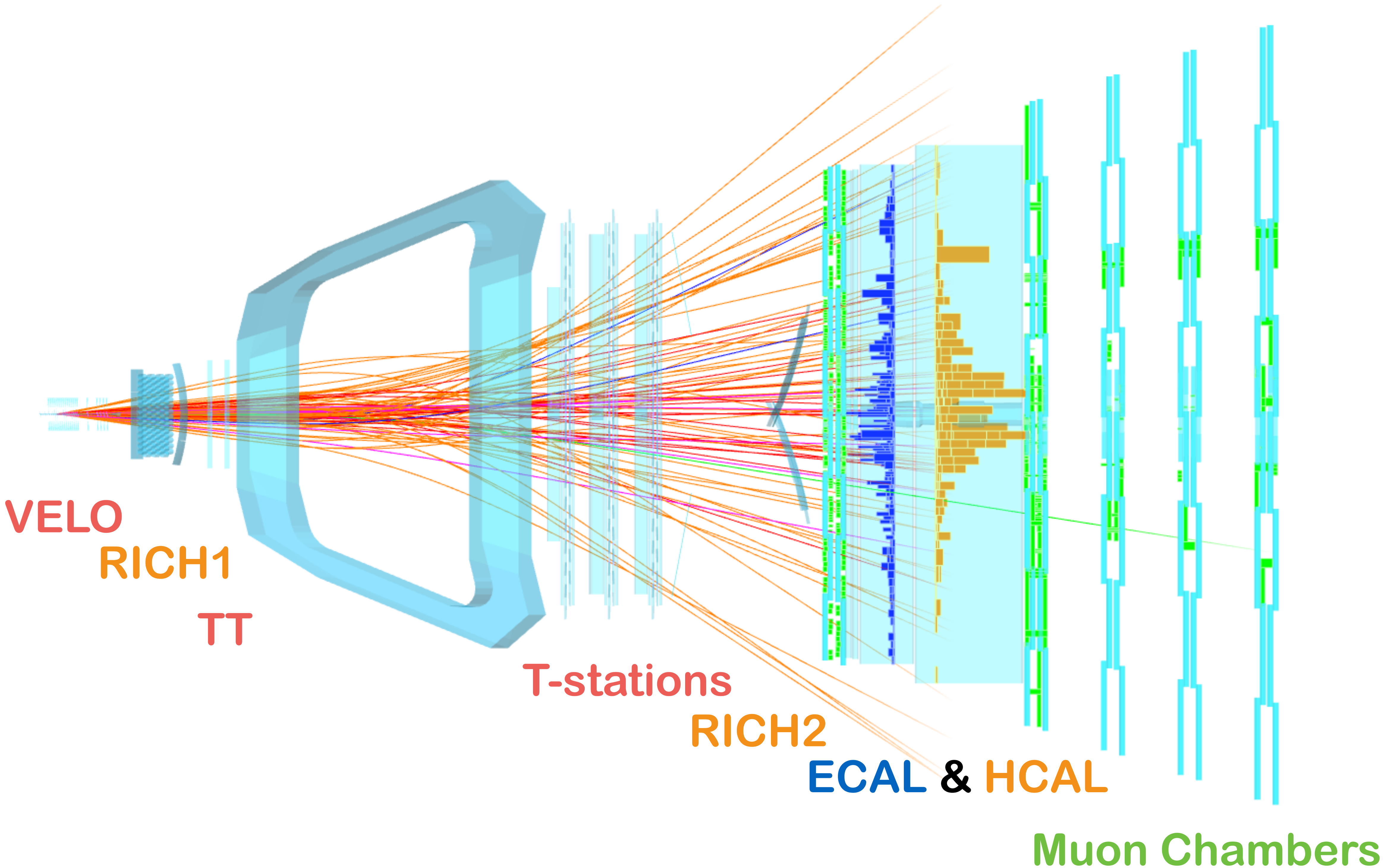


J.Phys.Conf.Ser. 664 (2015) no.8, 082025

http://iopscience.iop.org/article/10.1088/1742-6596/664/8/082025/meta

Precision analysis challenge

LHCb layout



PID at LHCb

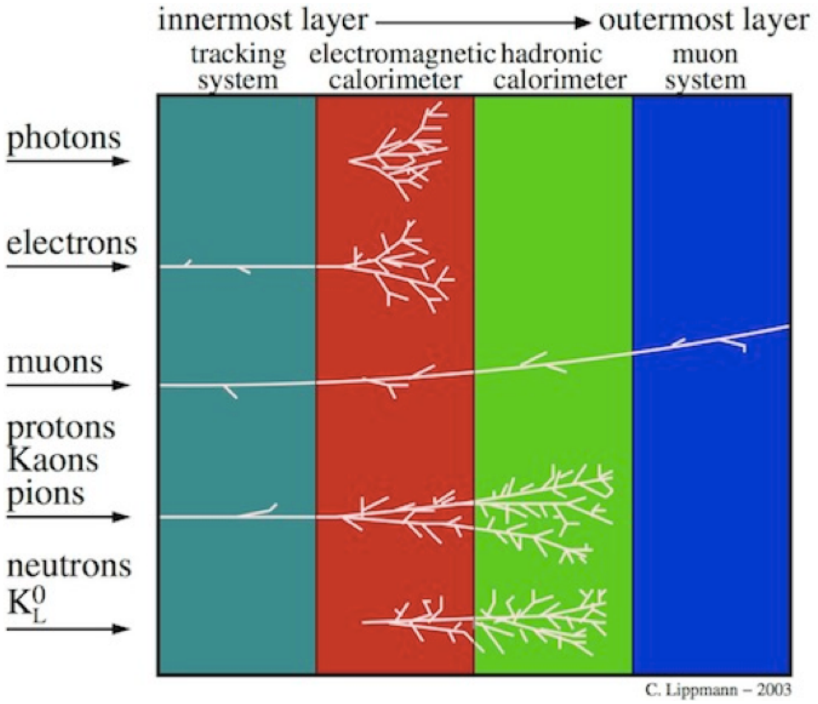
Problem: identify particle type associated with a track/energy deposited in the subdetectors

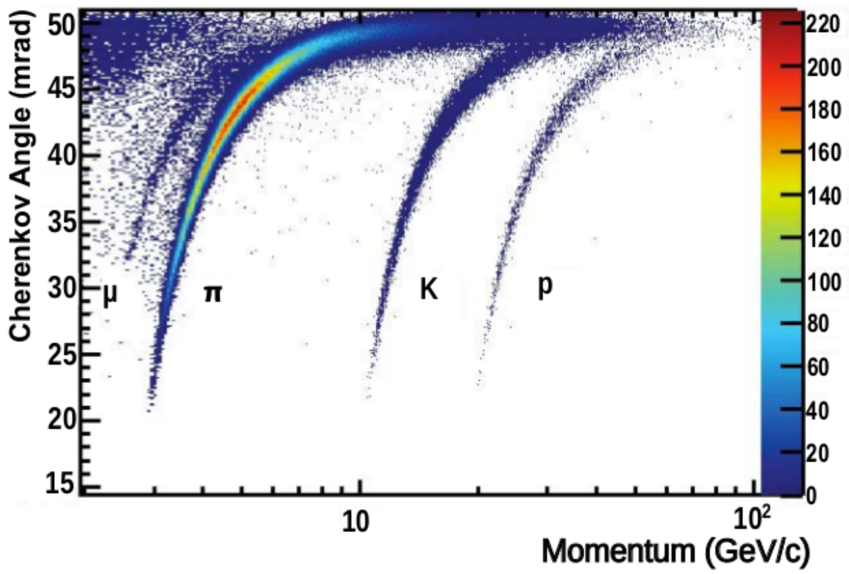
- Charged: π, e, μ, K, p
- Neutral: π^0, γ, n

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



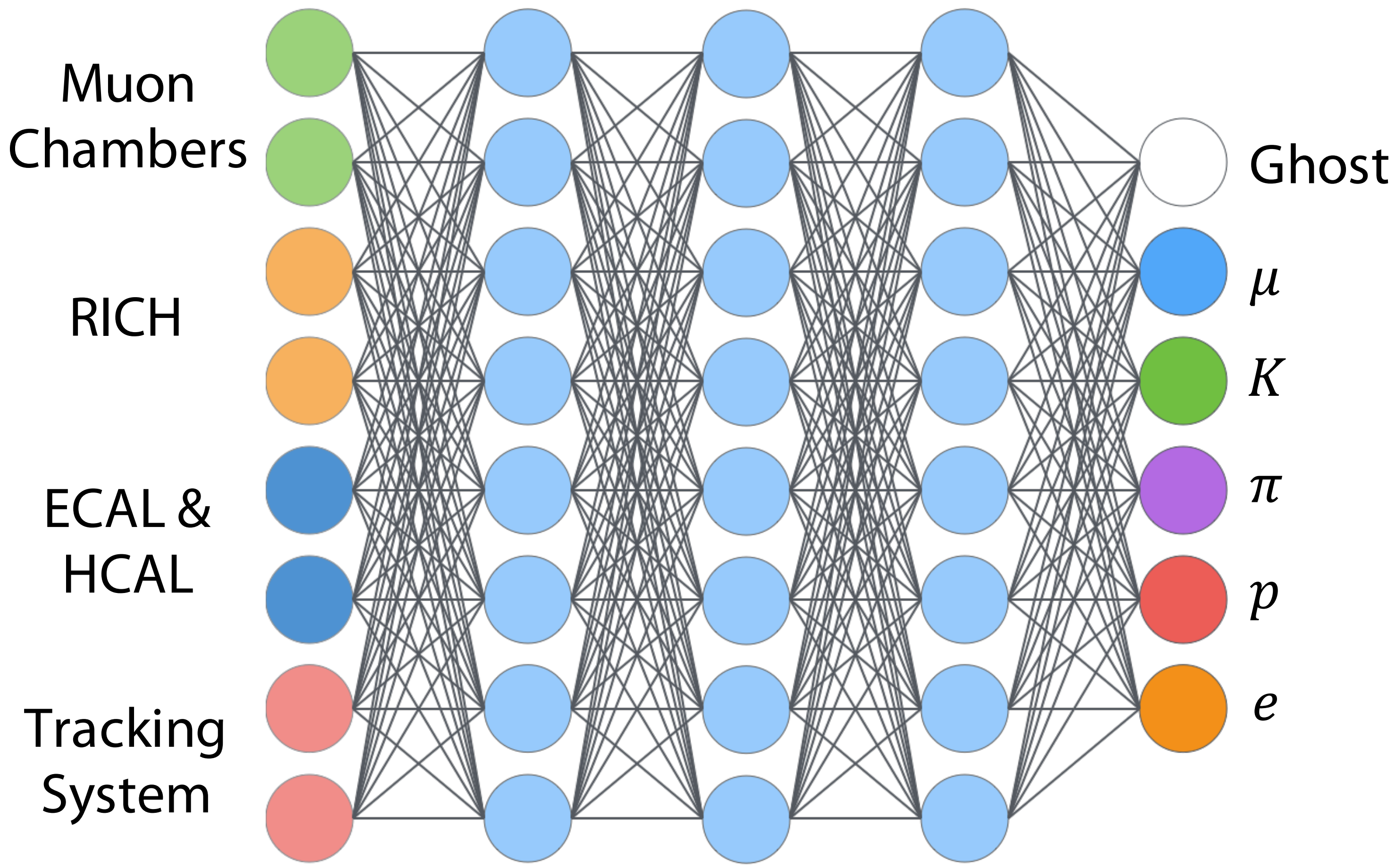


Global Particle Identification

Problem: identify particle type associated with a track.

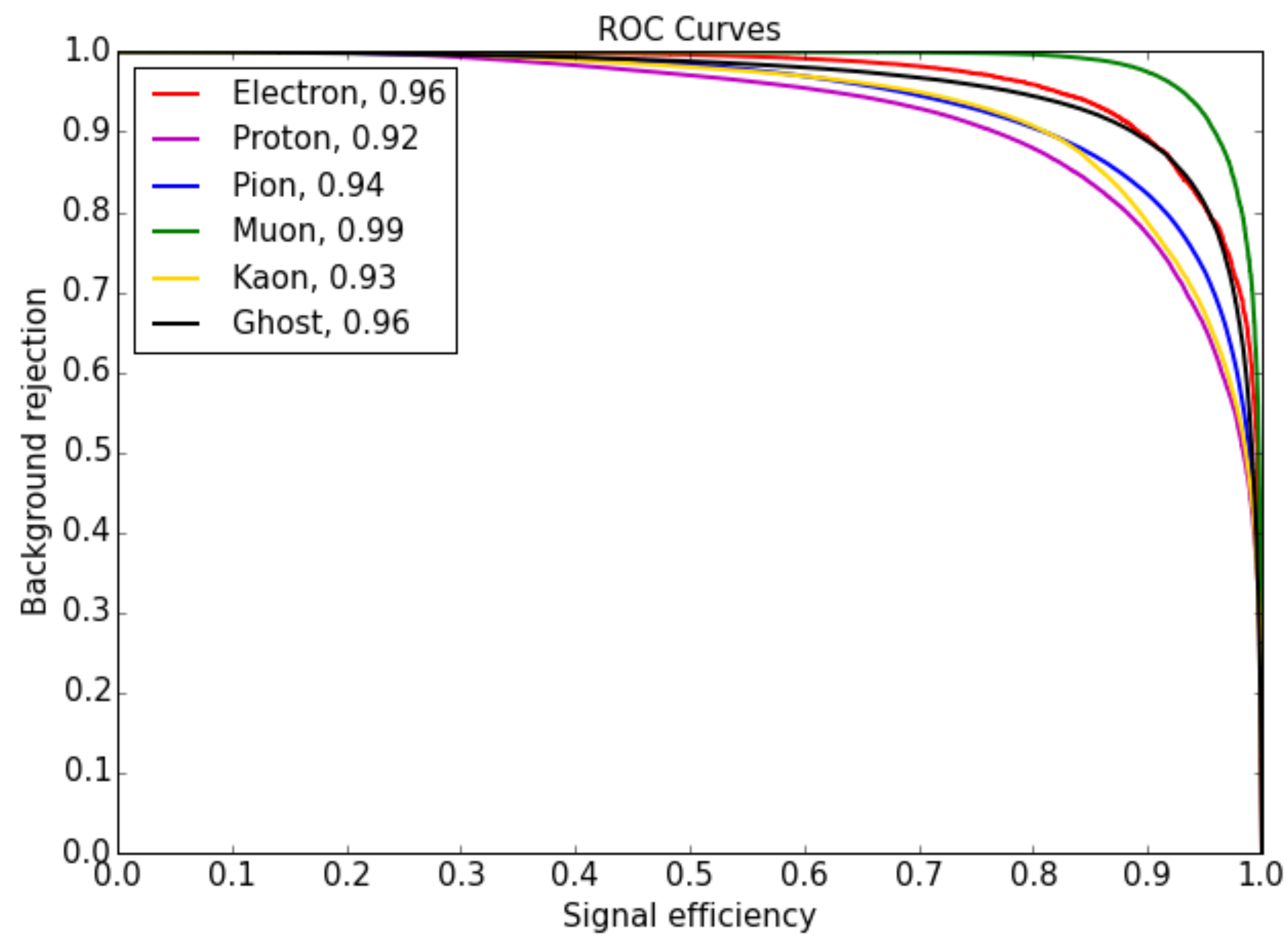
Particle types: Electron, Muon, Pion, Kaon, Proton and Ghost

Input observables: particle responses in RICH, ECAL, HCAL subdetectors, Muon Chambers and Track observables.



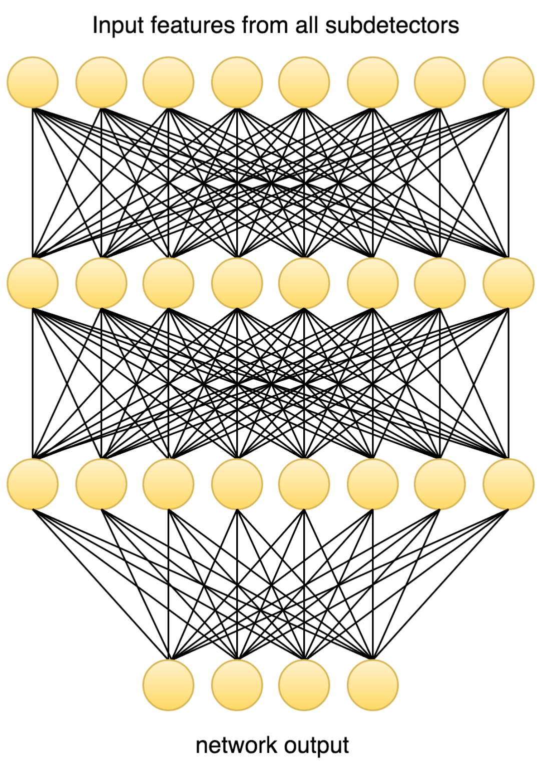
Quality Metrics

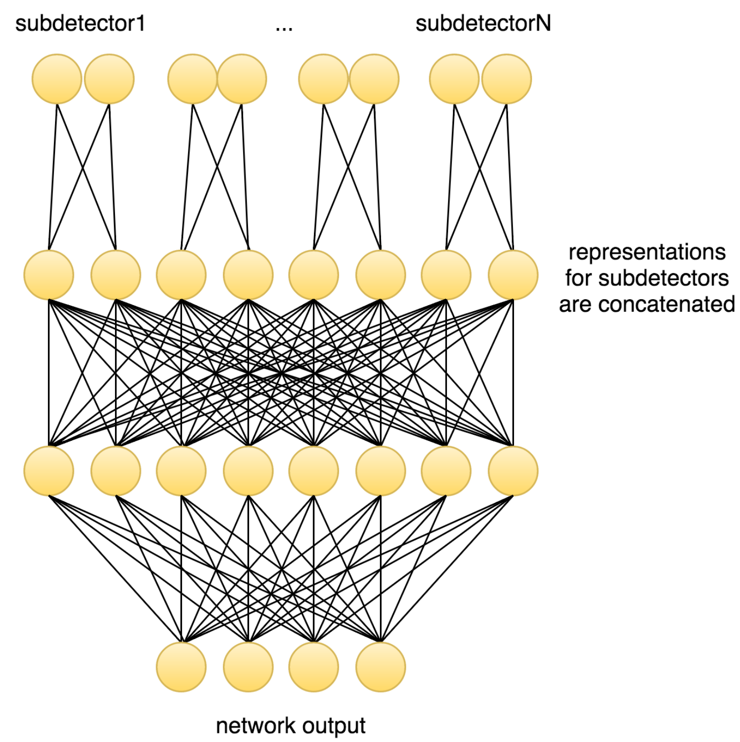
- One-vs-rest ROC curves used to measure models quality.
- Area under them (ROC AUC) are used as target metrics to select the best models.

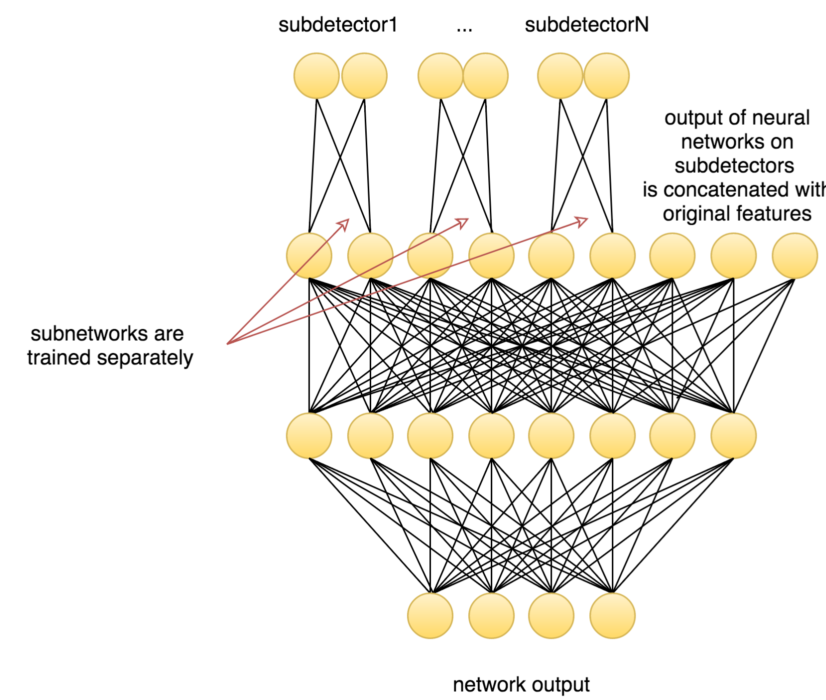


Technologies

- Several possibilities were tested, all of them were inspired by the knowledge of detector responses.



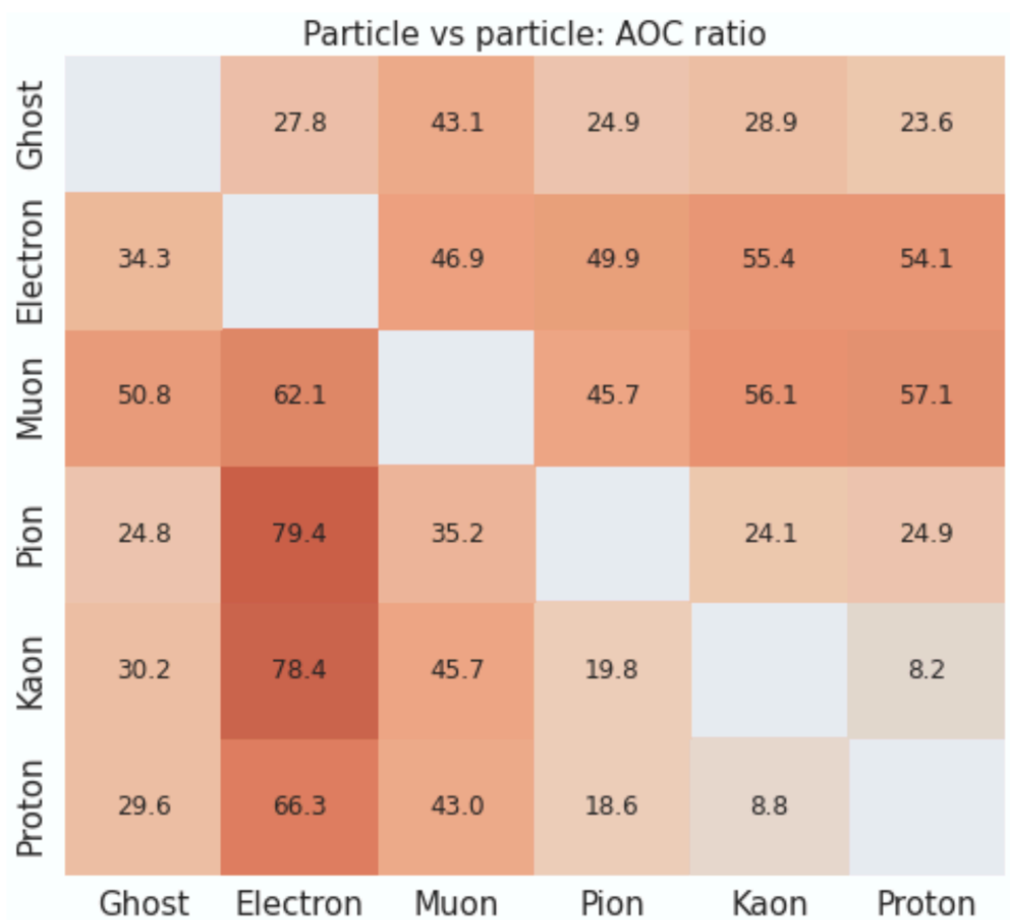


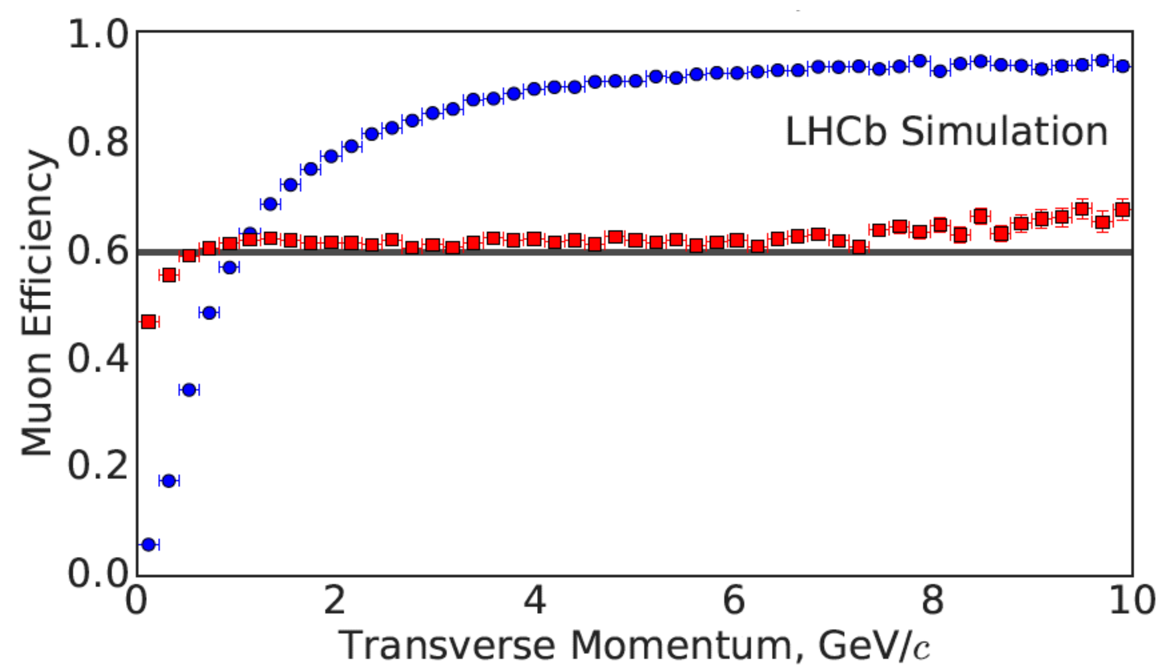


- Other approaches using Decision trees were also tested and brought competitive results.

Results

- Using the above mentioned approaches we were able to decrease the error rate by up to 80%.





- In addition to this, we were able to correct the detector acceptance function, which lead to a lower systematics.

Journal of Physics: Conference Series. 2018. Vol. 1085. No. 4. P. 1-5.

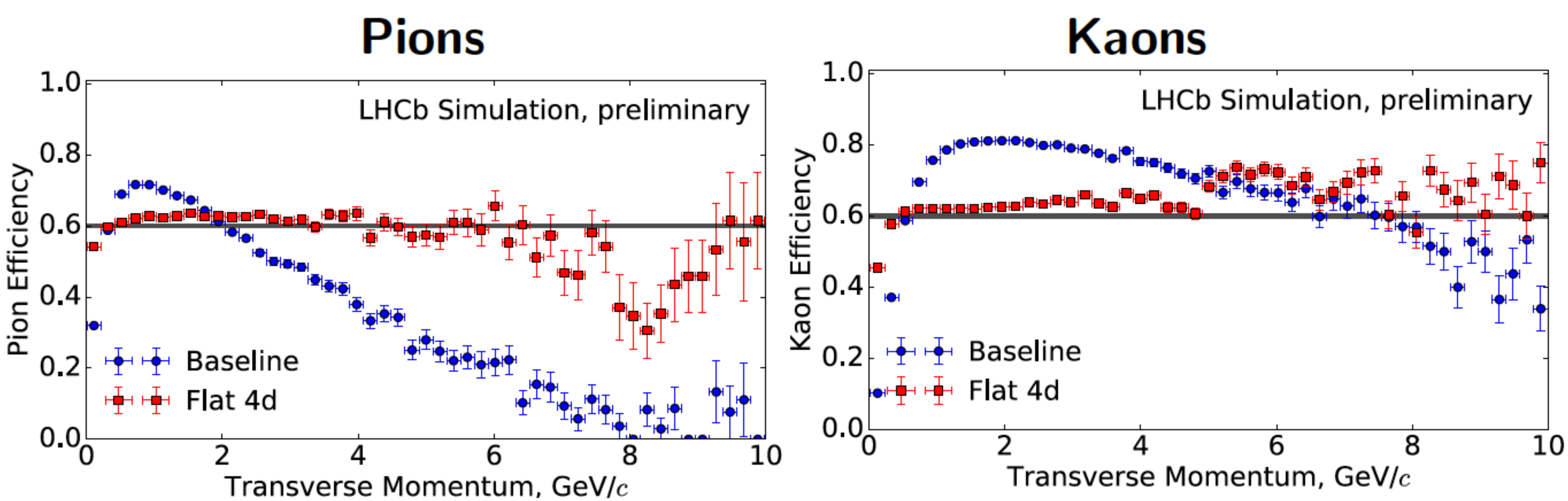
http://iopscience.iop.org/article/10.1088/1742-6596/1085/4/042038/meta

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



Flat4d, ProbNN

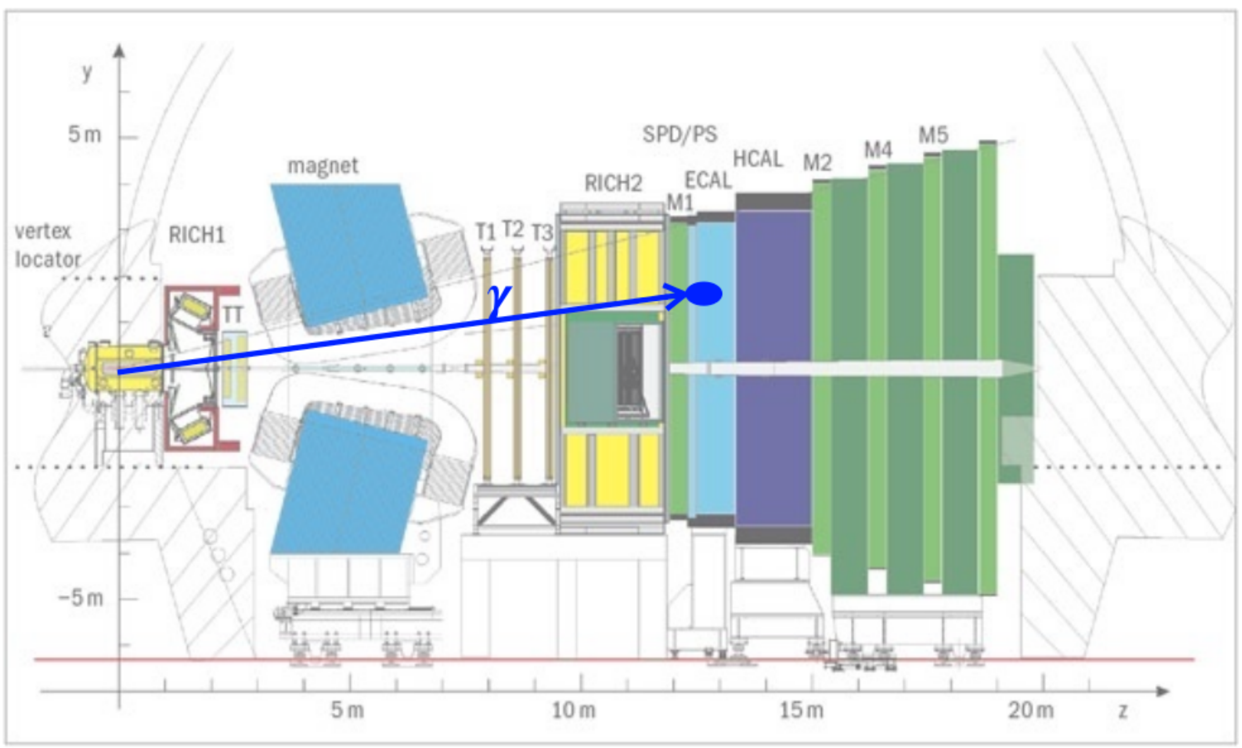
→ Better PID efficiency flatness in p, p_T, η, N_{tracks} than baseline

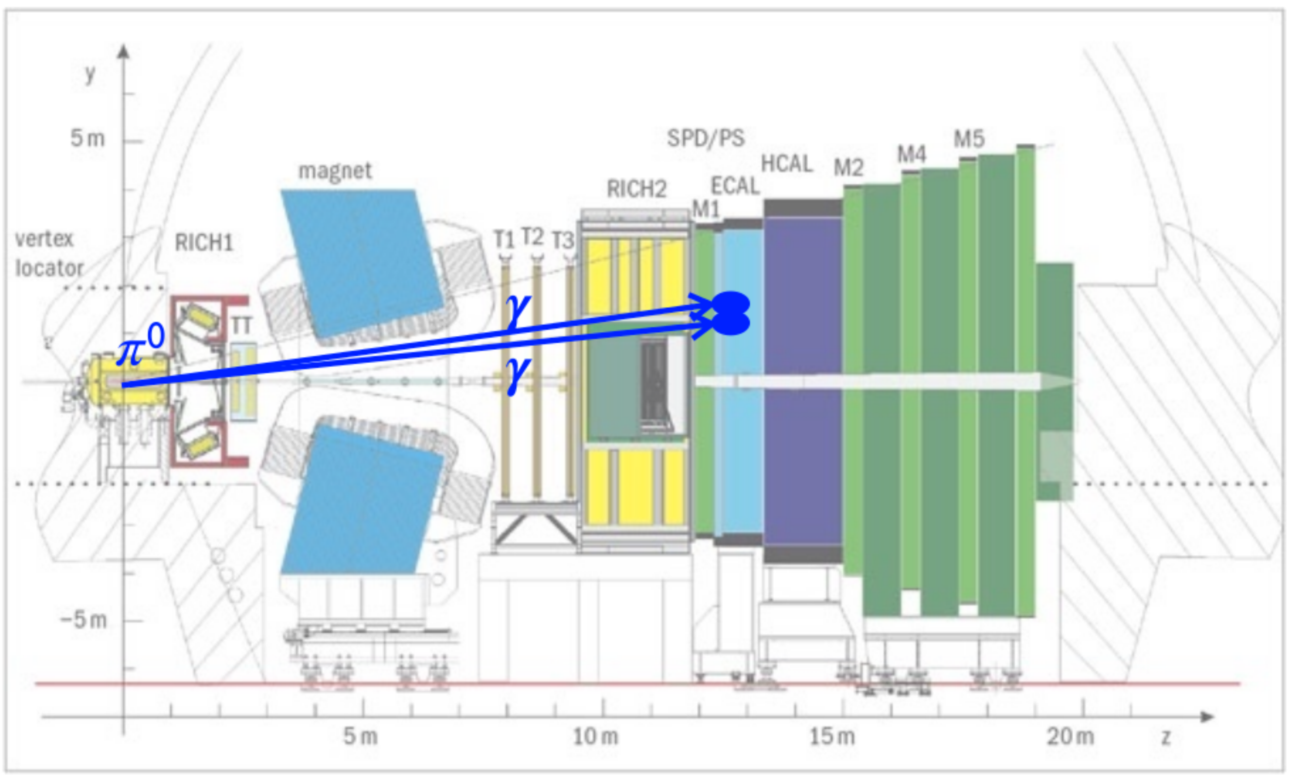
Neutral PID

π^0 copiously produced at LHCb , decay to $\gamma\gamma$

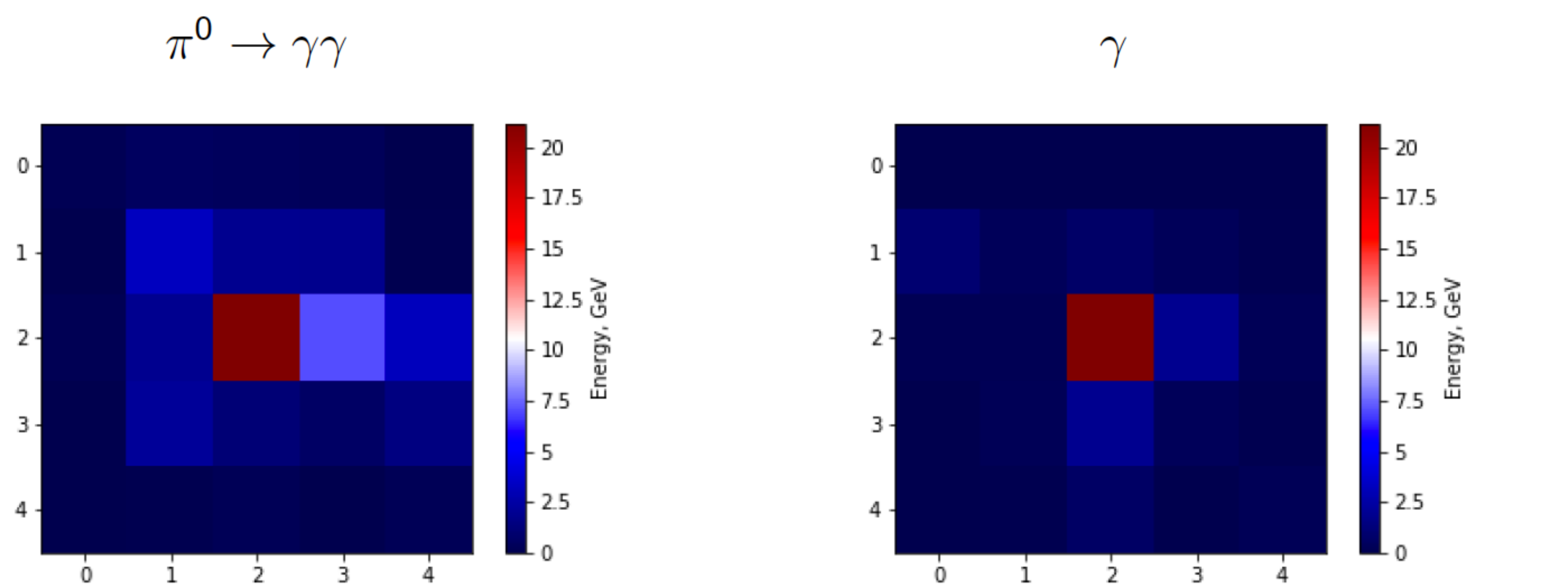
high momentum $\pi^0 \rightarrow$ merge of ECAL clusters $\rightarrow$ huge background for radiative decays

Need for a powerful tool to discriminate signal (γ) from background $\pi^0 \rightarrow \gamma\gamma$





ECAL Signatures

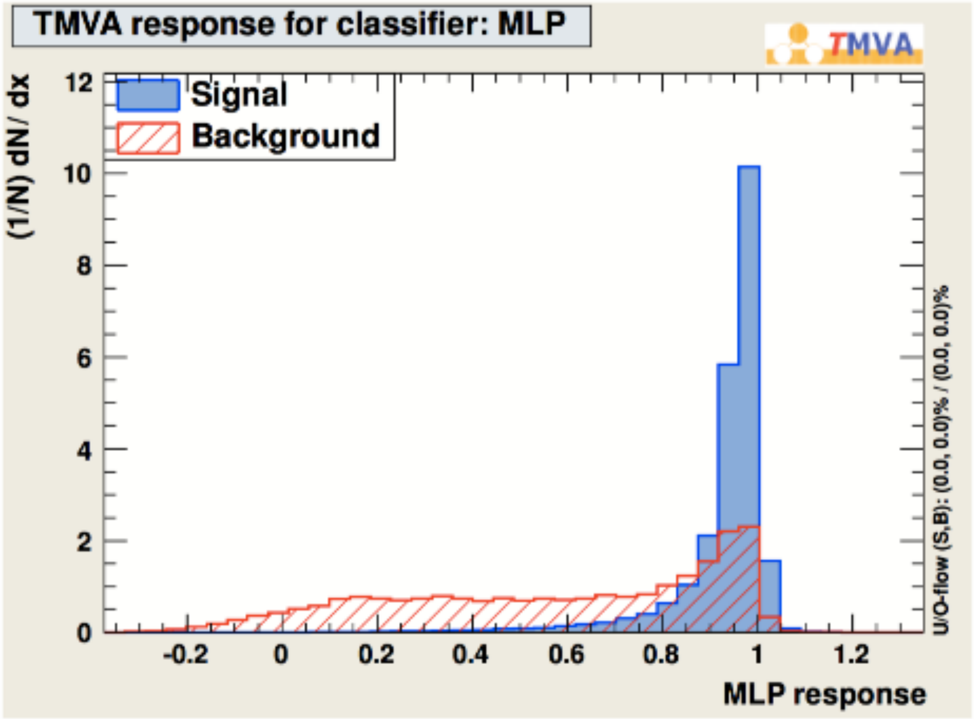


ECAL clusters (3x3 cells)

Coarse granularity → separation is not straightforward

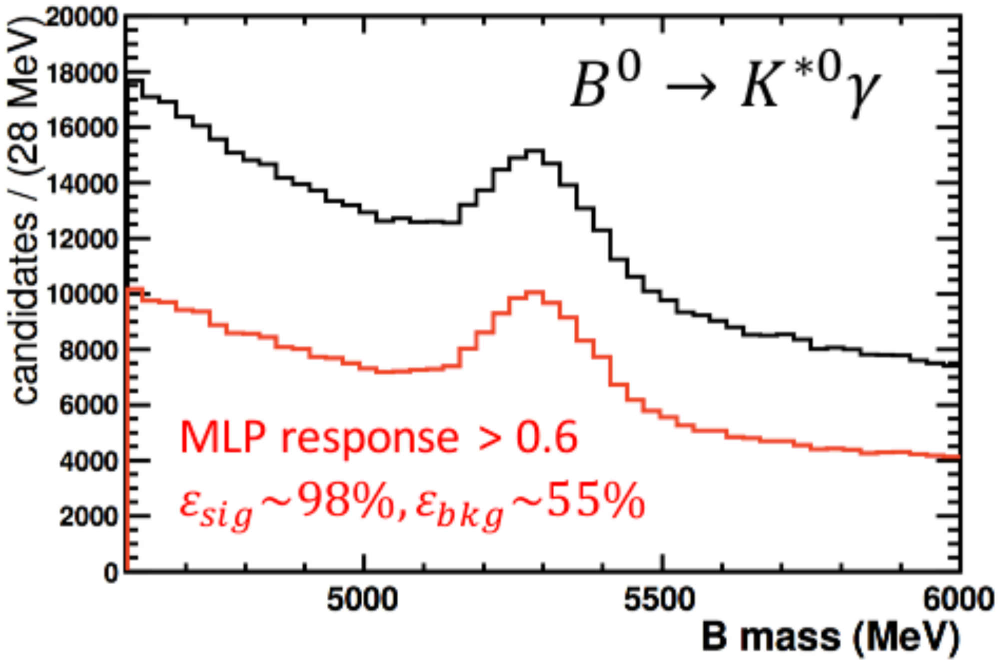
Baseline approach [LHCb-PUB-2015-016]

Neural Network with 2 hidden layers (TMVA MLP)



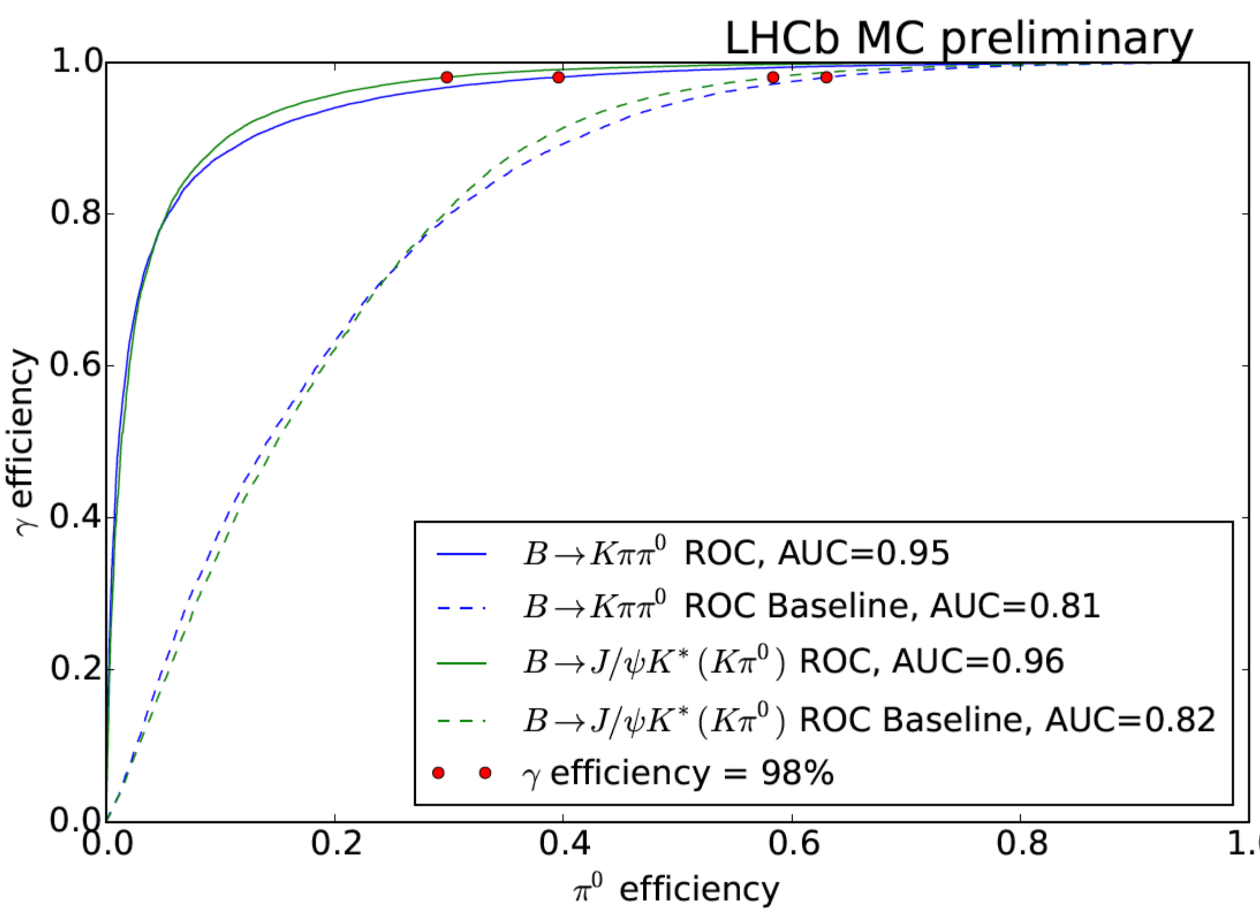
14 ECAL and Pre-Shower cluster parameters (grouped under shape and symmetry)

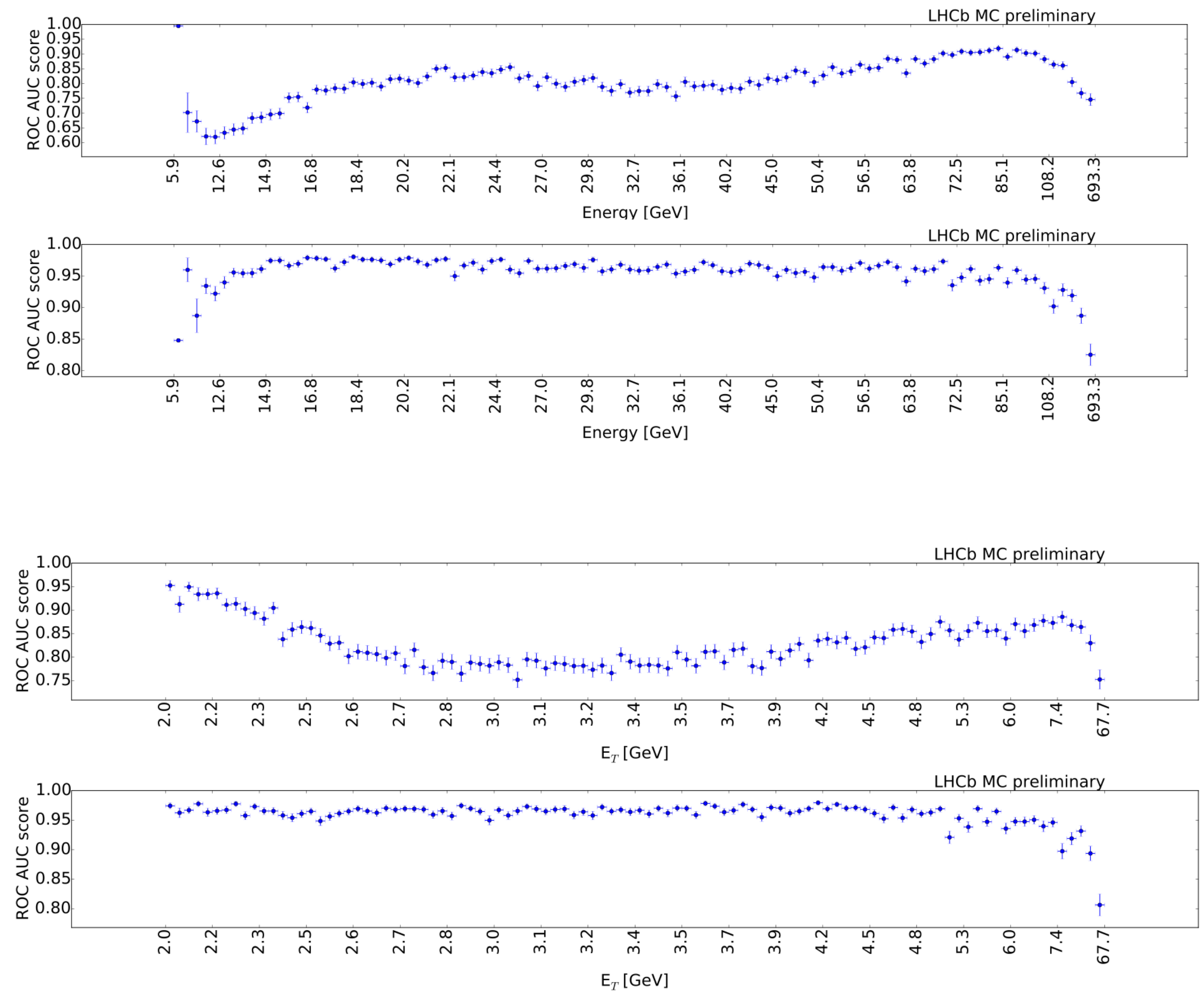
- 4 variables that account for the size & tails, semiaxes and orientation of the
- ellipse in the ECAL
- 2 variables related to the energy of the most (seed) and the second most
- energetic cells of the cluster
- 4 variables for multiplicities of hits in the PS cells matrix in front of the seed
- of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



New 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





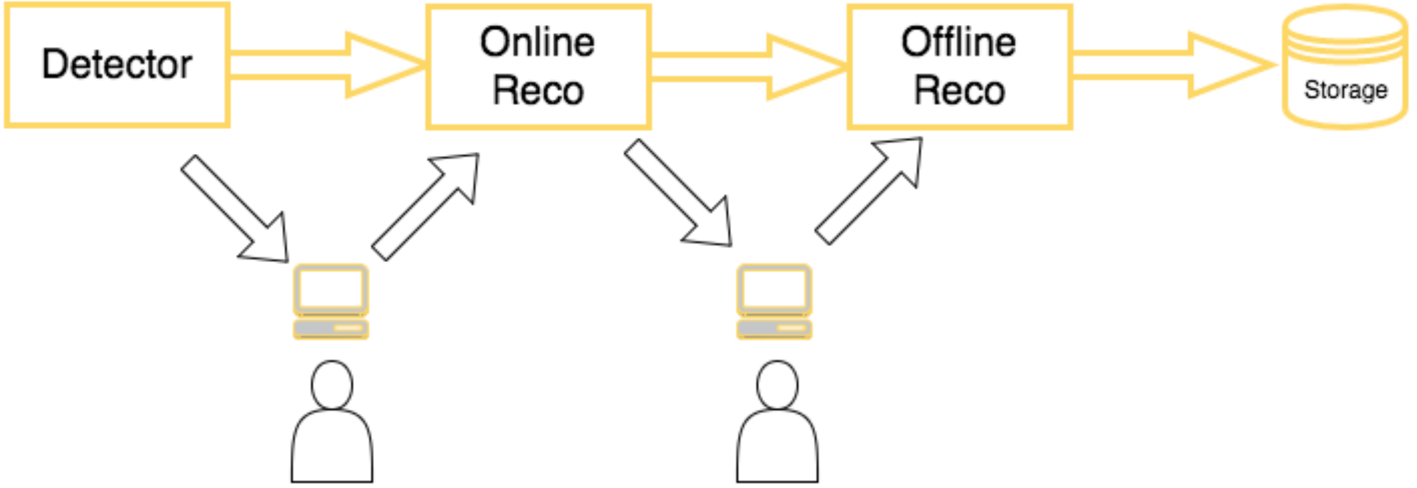
Journal of Physics: Conference Series. 2018. Vol. 1085. P. 1-5

http://iopscience.iop.org/article/10.1088/1742-6596/1085/4/042036/meta

Automation Challenge

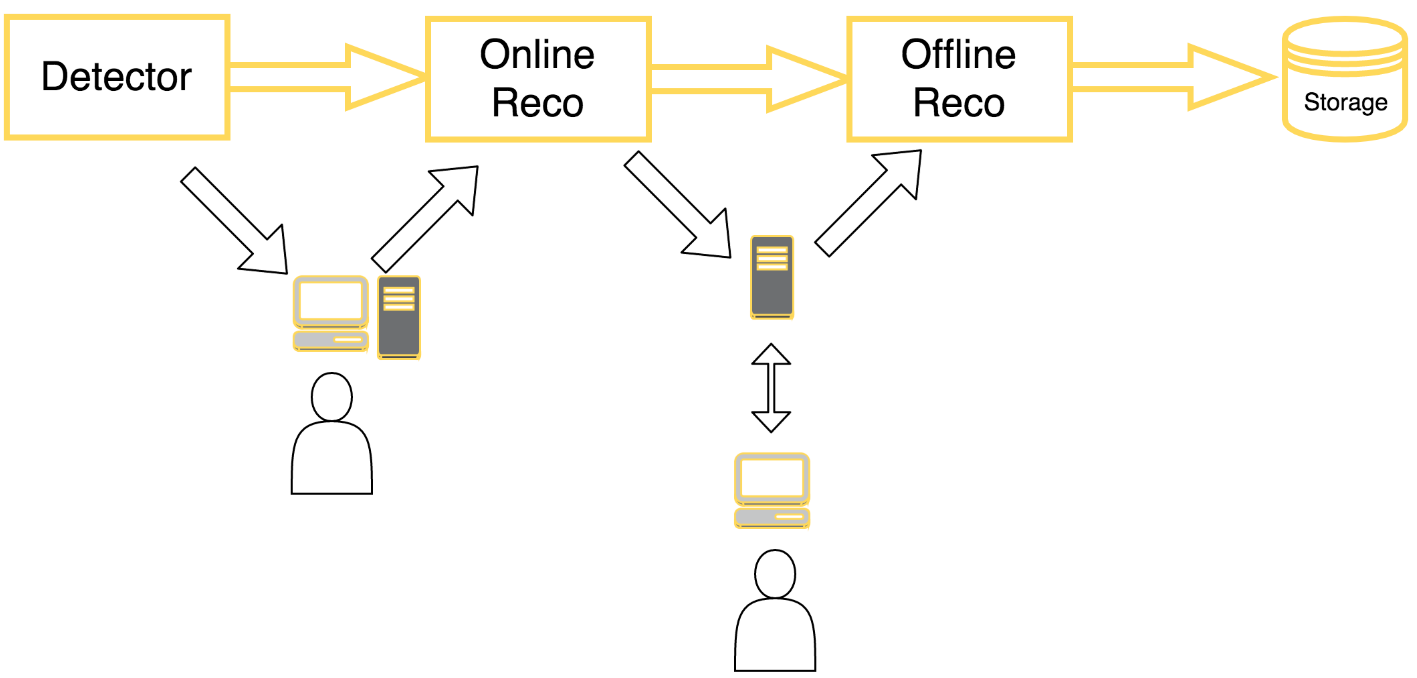
Data Quality Control

- Several people are typically on shifts controlling the flow of data from detector into the storage



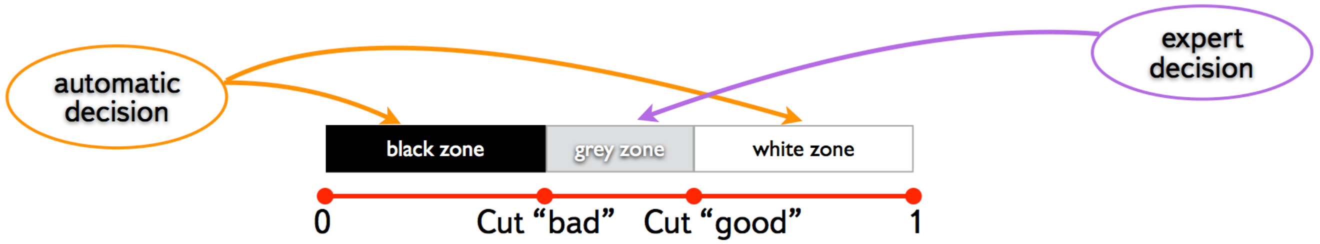
Updated Workflow

- The monitoring systems can be updated with:
  - **helper**, a recommendation system for a shifter
  - **solver**, automated decision maker
  - **both**

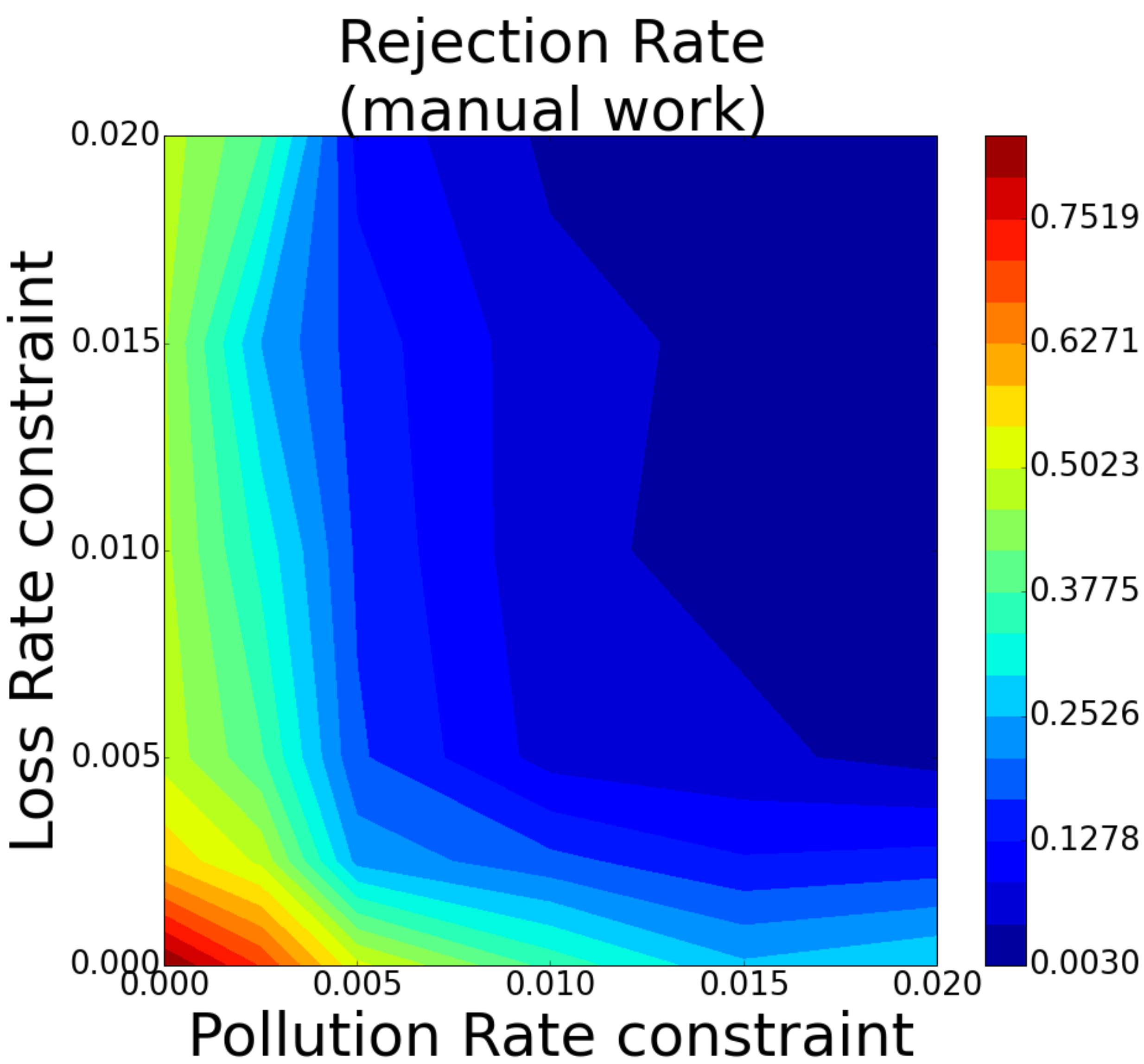


Supervised Learning

- Problem: CMS Data Certification
- Data: CMS 2010B run open data
- Aim: automated classification of LumiSections as “good” or “bad”
- Features: particle flow jets, Calorimeter Jets, Photons, Muons
- The dataset was flagged by experts (3 FTE)



More time for researchers



The aim is to minimise the Manual work with low Loss Rate ("good" classified as "bad") and Pollution Rate ("bad" classified as "good").

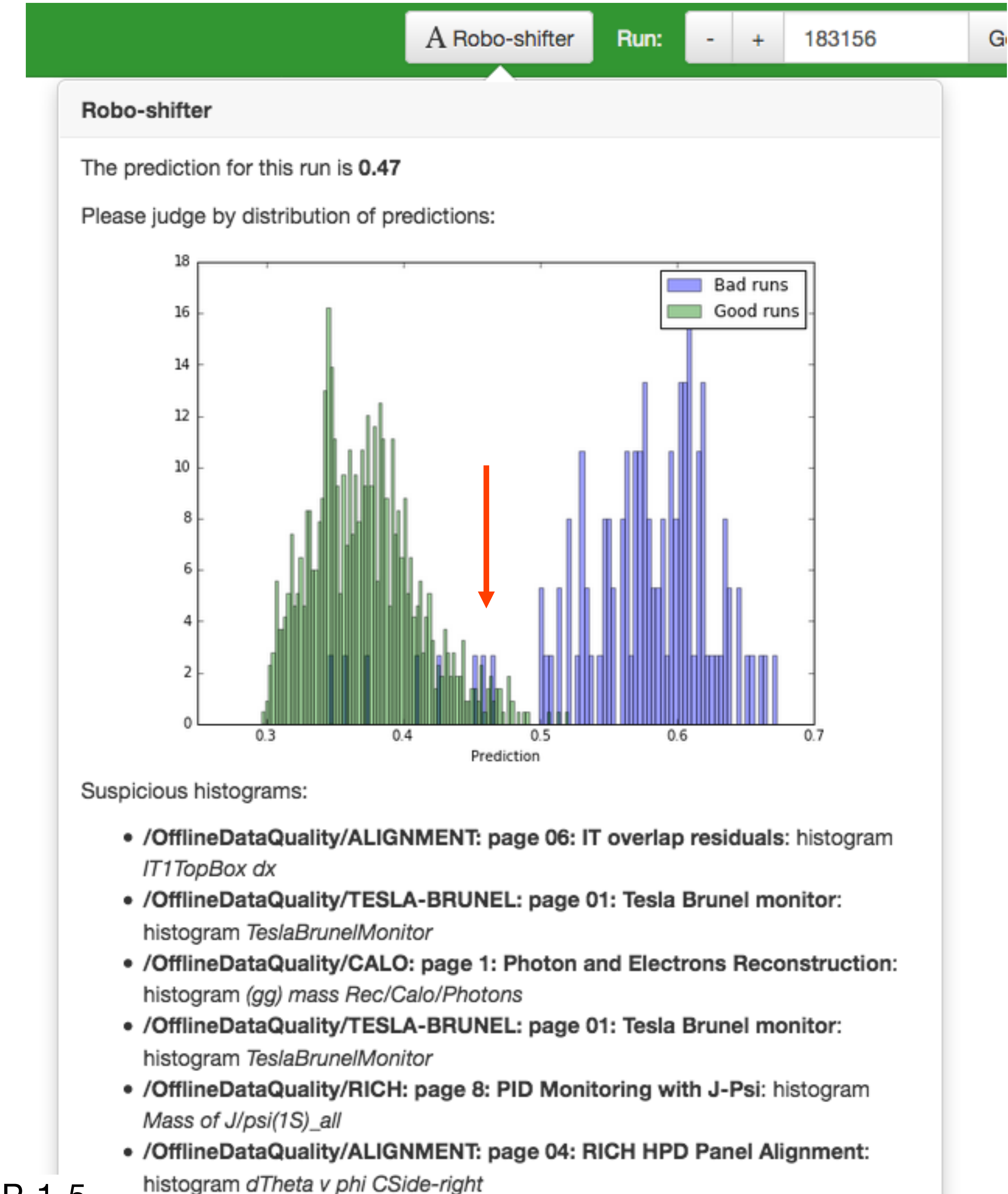
~90% saving on manual work is feasible for Pollution rate at 0.5%

J.Phys.Conf.Ser. 898 (2017) no.9, 092041

http://iopscience.iop.org/article/10.1088/1742-6596/898/9/092041/meta

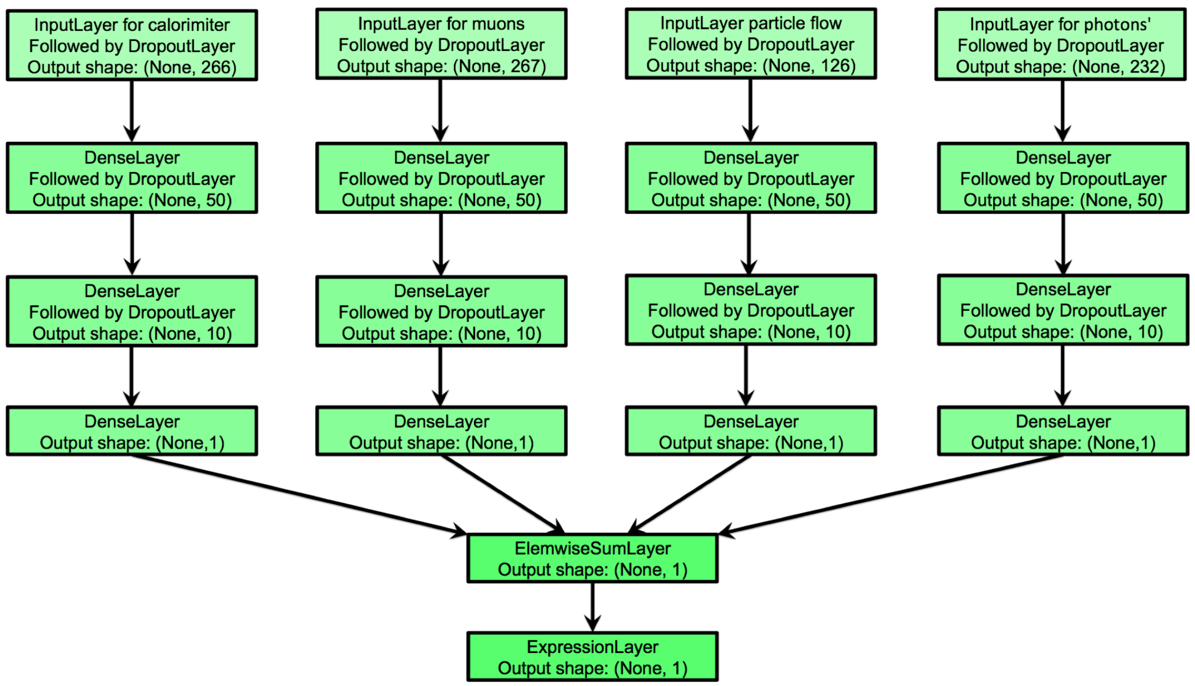
Monitoring Robo-shifter

- 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



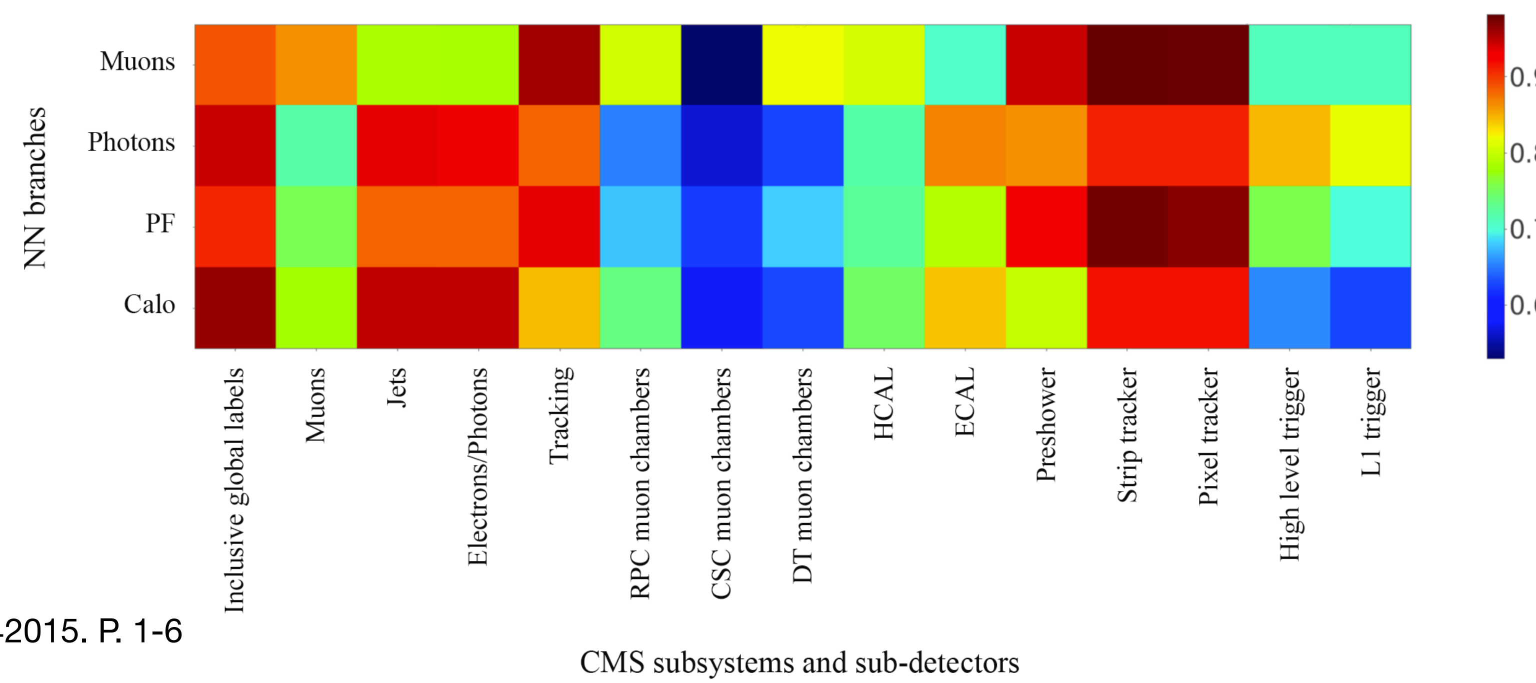
Journal of Physics: Conference Series. 2017. Vol. 898. No. 9. P. 1-5
http://iopscience.iop.org/article/10.1088/1742-6596/898/9/092027/meta

Better Localisation of Anomalies



Even more than this, we are able to identify a particular failing subsystem. The training only requires global flags.

AUC scores vs channels

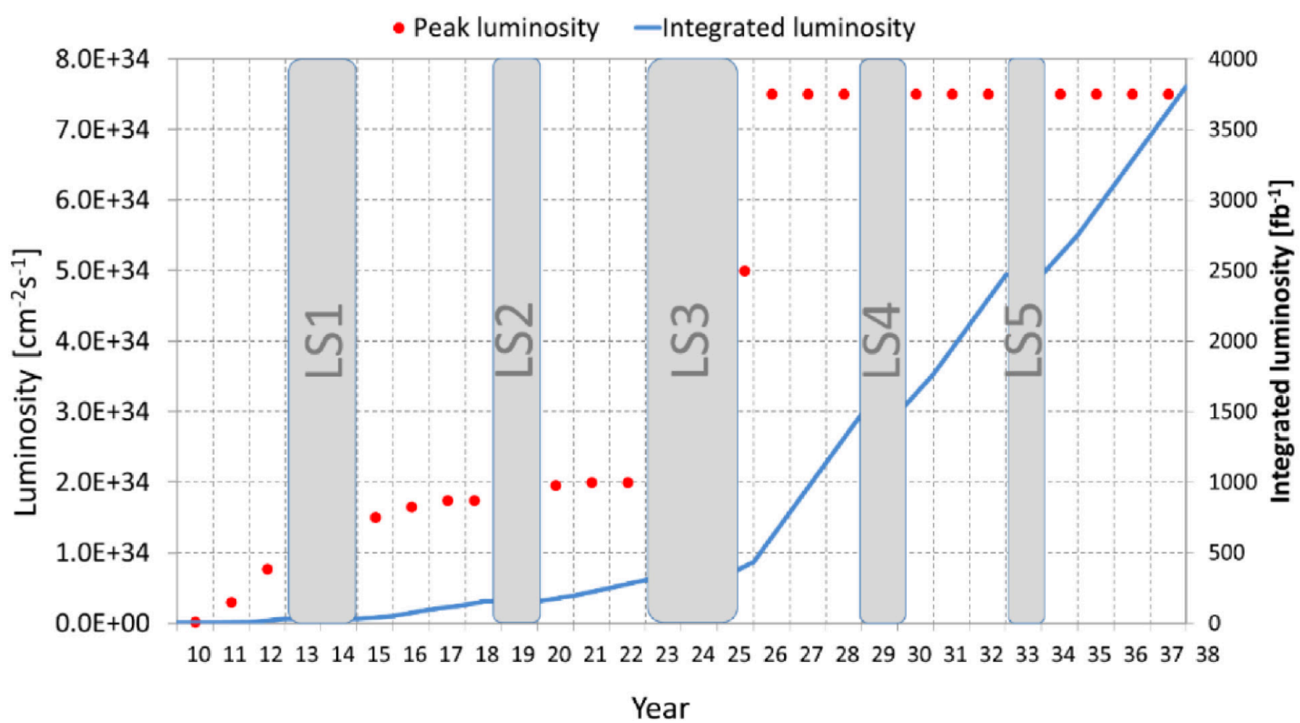


Journal of Physics: Conference Series. 2018. Vol. 1085. No. 042015. P. 1-6

http://iopscience.iop.org/article/10.1088/1742-6596/1085/4/042015/meta

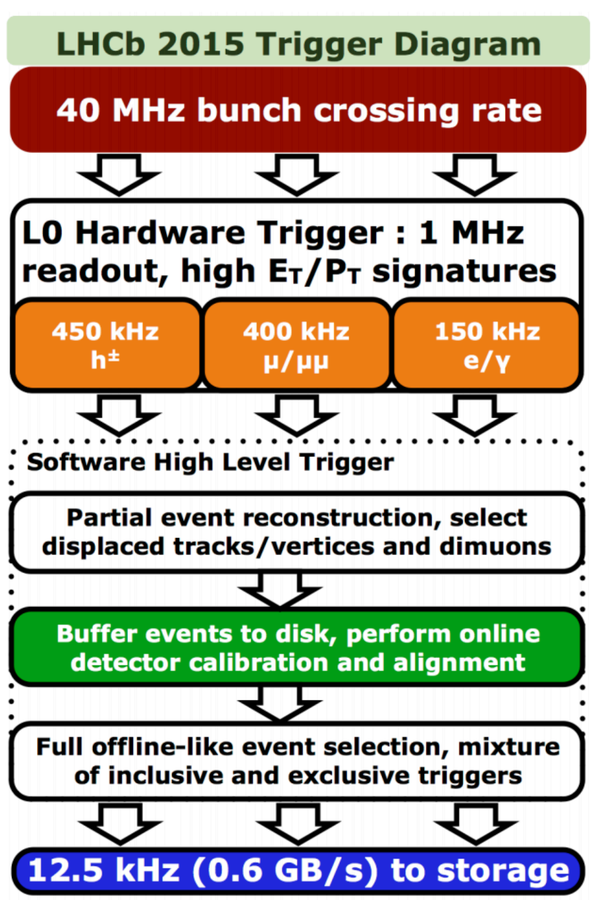
Emerging Challenges

Large Hadron Collider Upgrade



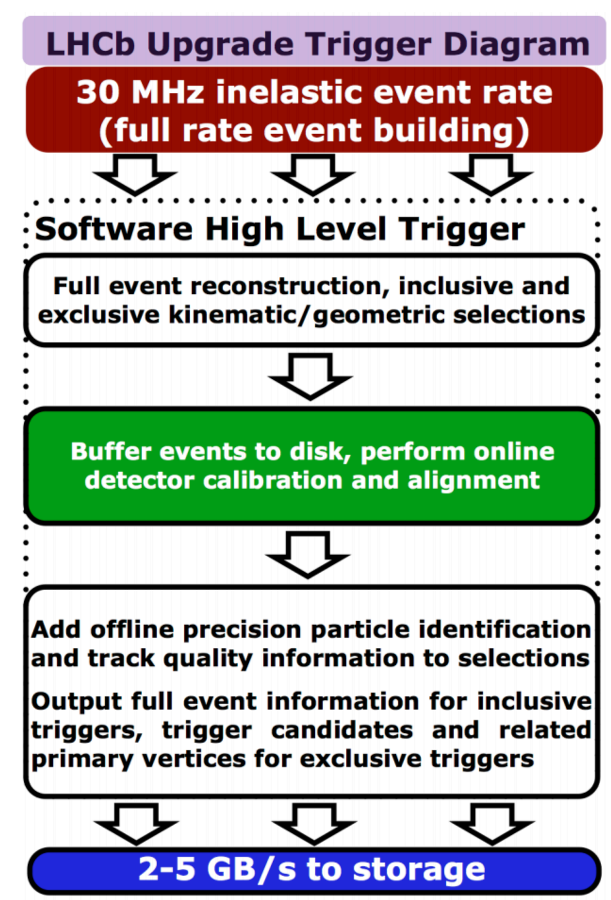
Statistics accumulated will be growing exponentially.

Thus, the challenges I mentioned before will also be harder and harder to tackle.

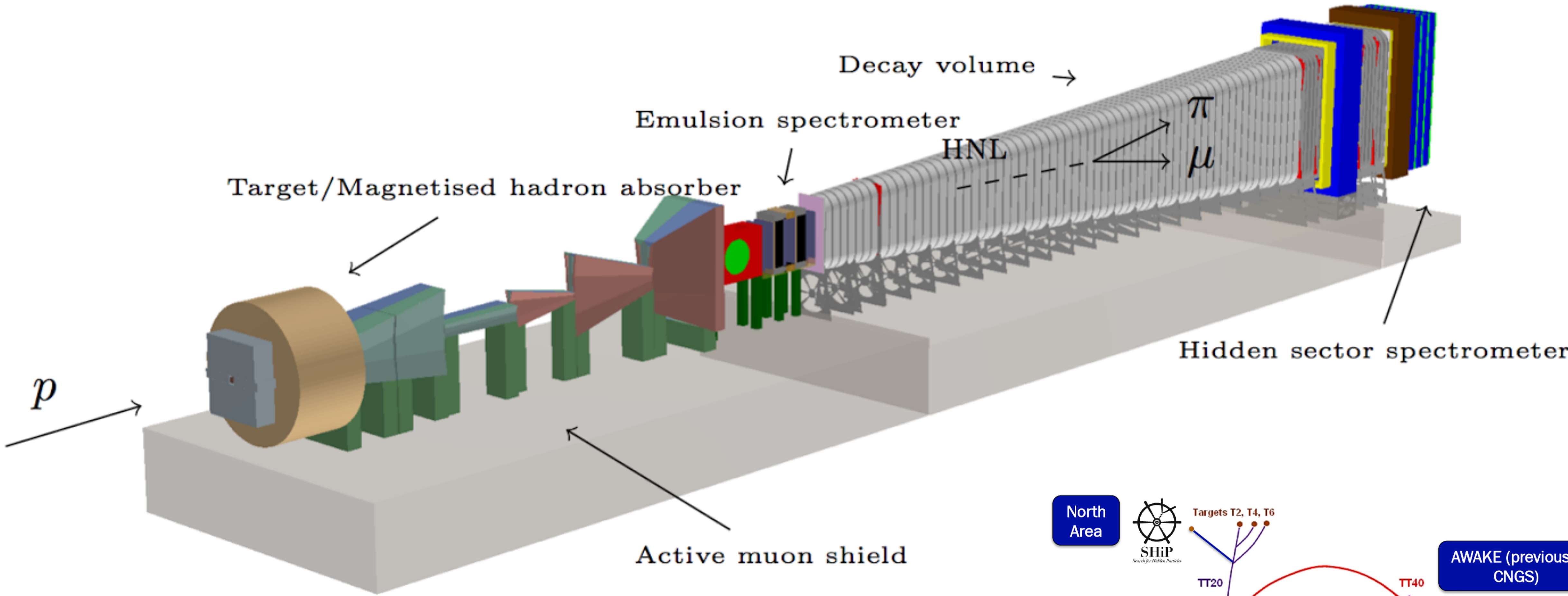


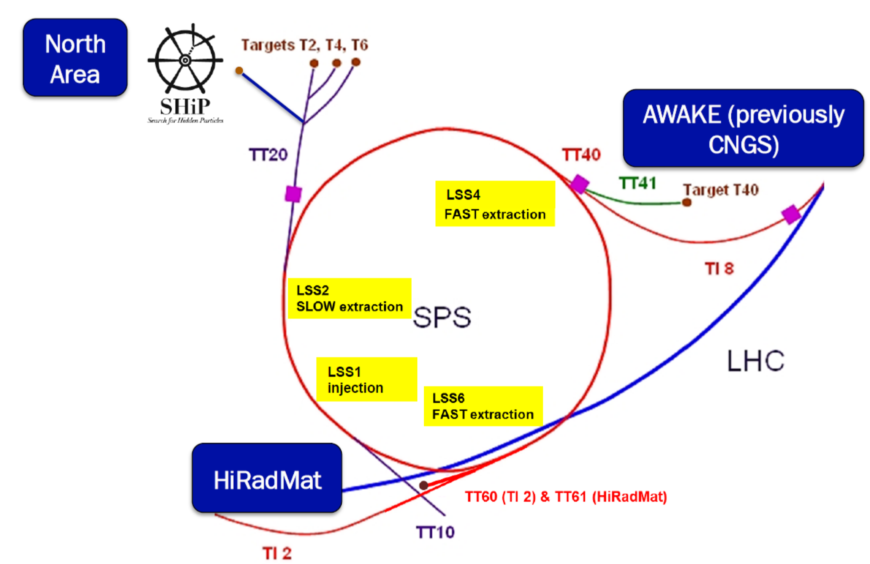
Remove hardware trigger

Increased output rate to storage



SHiP Experiment

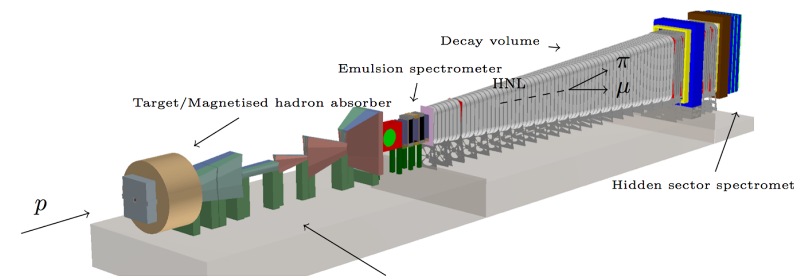


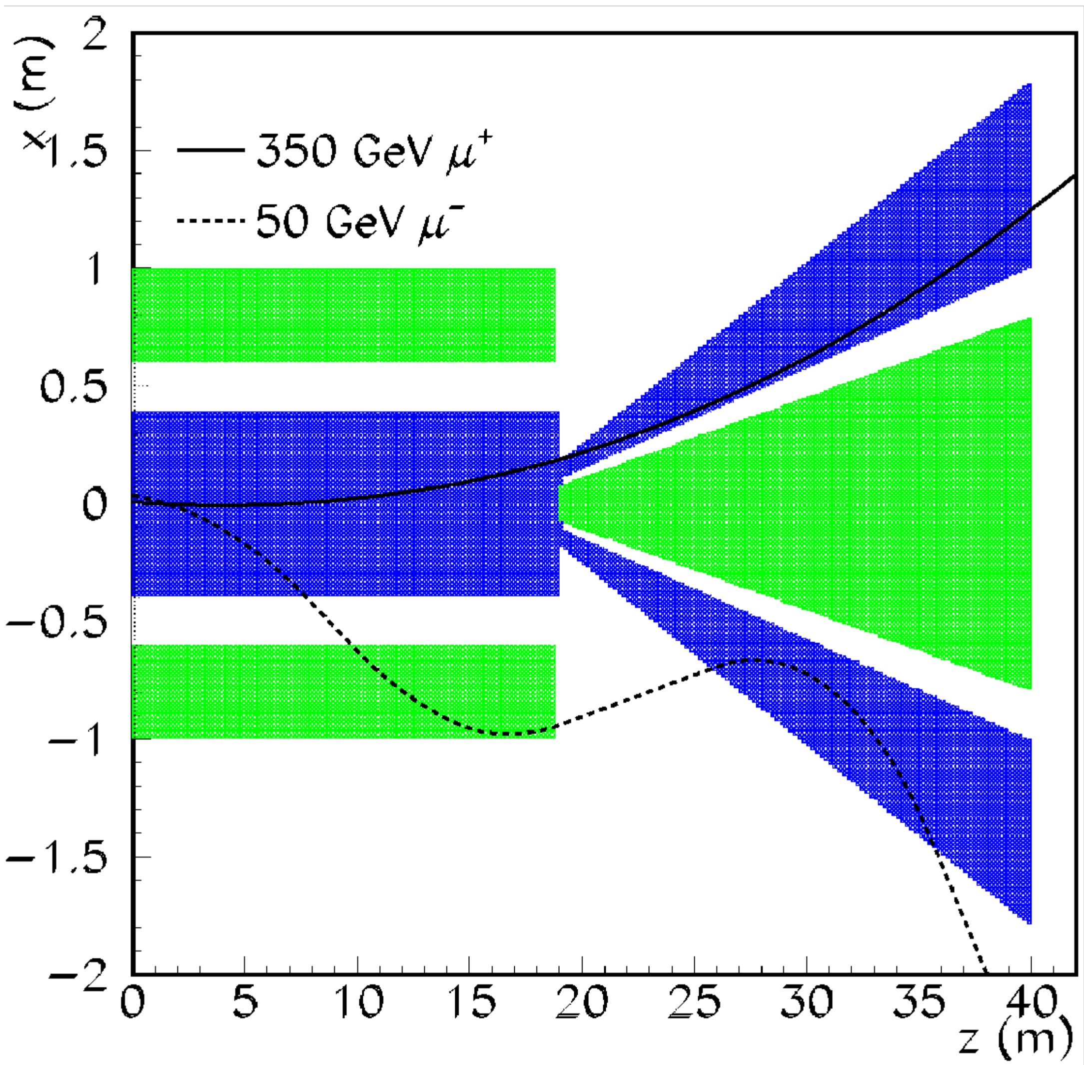


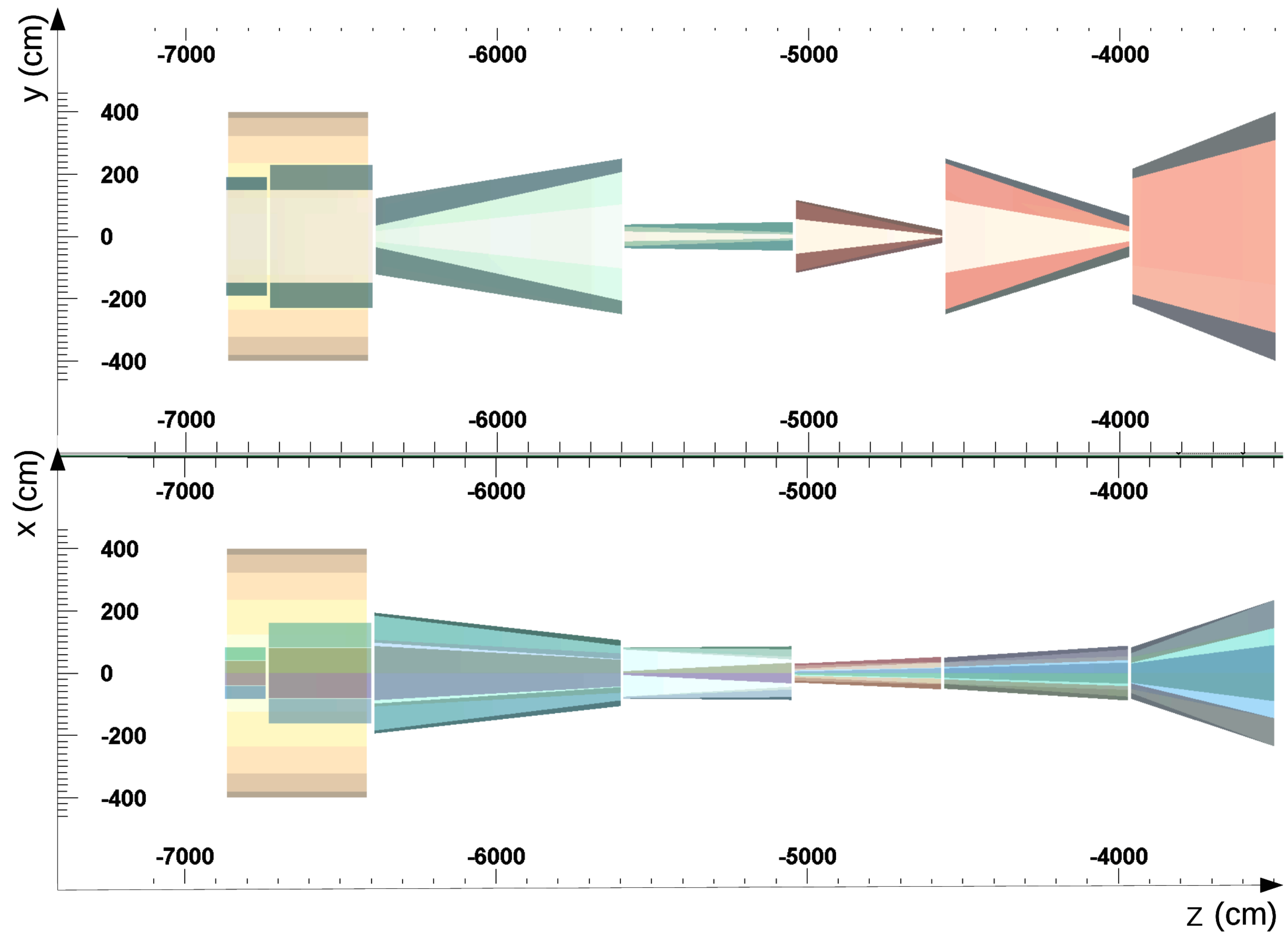
- Search for Hidden Particles
  - Post-LHC era experiment for direct search of very weakly interacting light particles

J.Phys.Conf.Ser. 934 (2017) no.1

Active Magnetic Shield



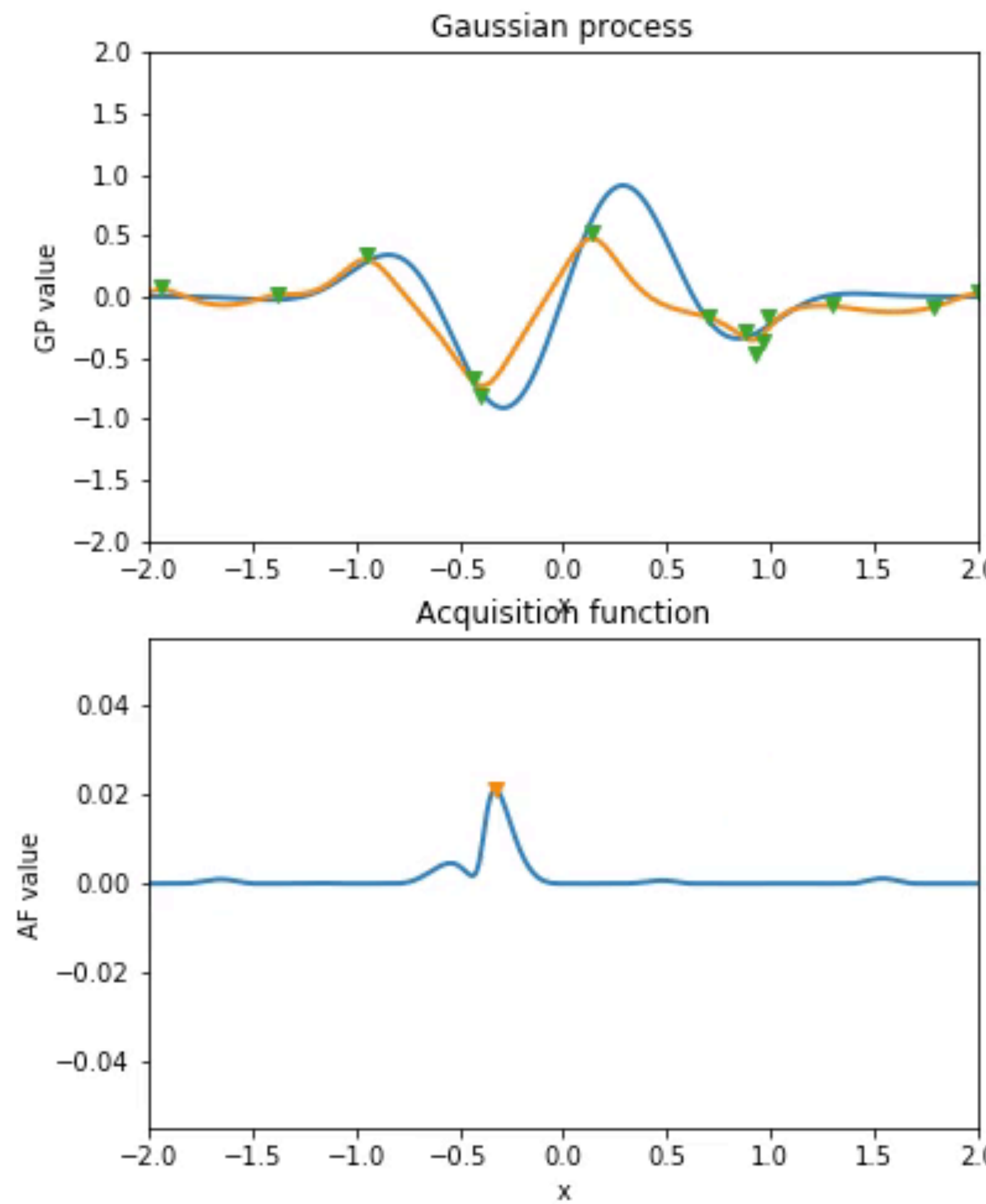




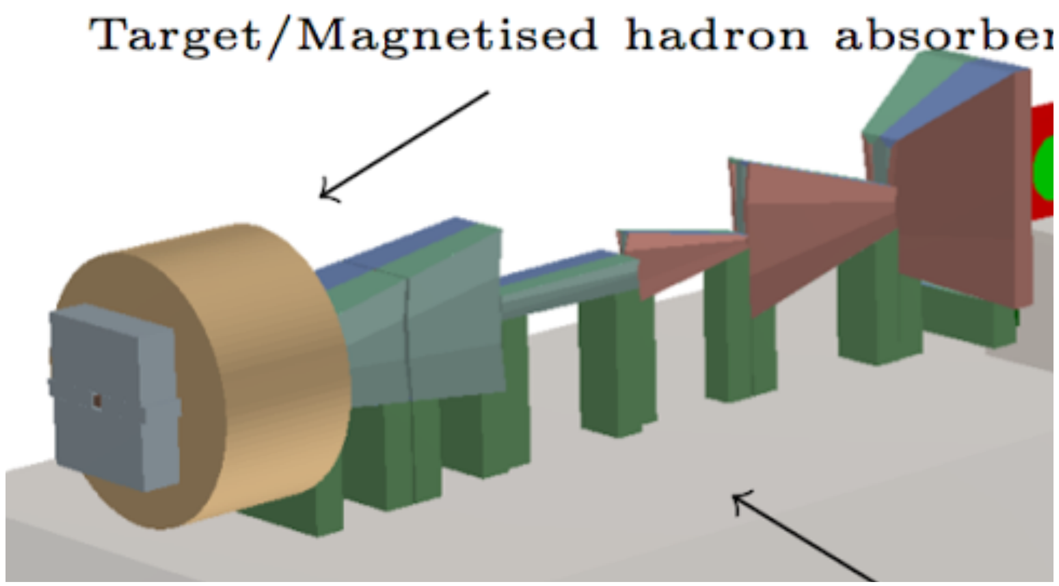
- Absorber shape optimization: background suppression at reasonable cost

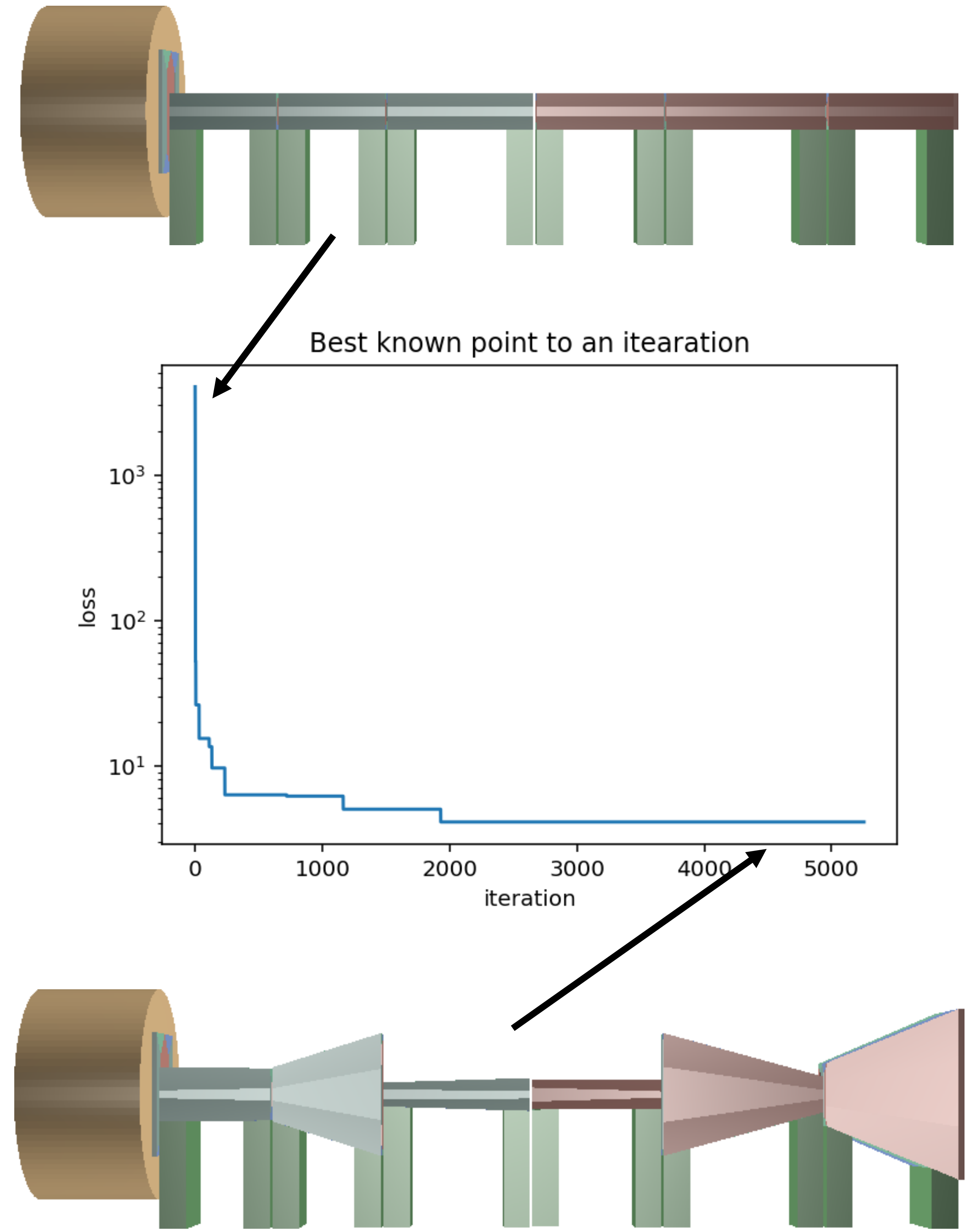
Gaussian Process Optimization

- 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

Journal of Physics: Conference Series, 2017, Vol. 934, P. 1-5

http://iopscience.iop.org/article/10.1088/1742-6596/934/1/012050/meta

Emerging Challenges: Reliable and Fast Simulation

- 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):

x **input**

K variables:
px, py, pz, ...
particle type, etc

ε **noise**

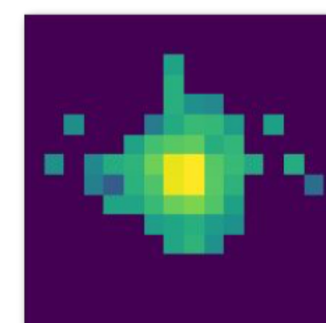
L variables:
"hidden variables"

Sampler

$$\hat{y} = G(x, \varepsilon, \theta) \sim p(y \mid x)$$

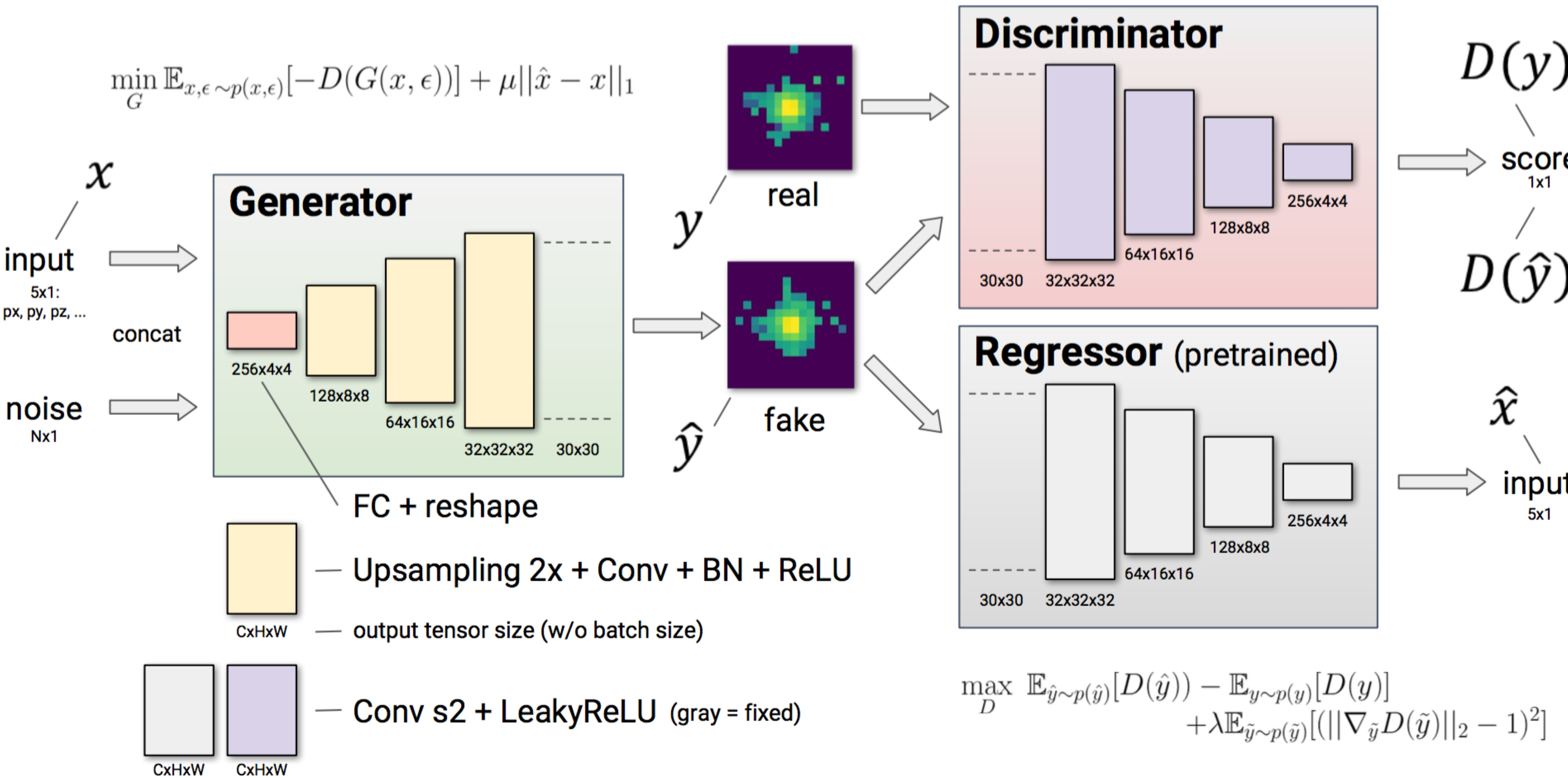
target y

HxW matrix
energy response
in cells

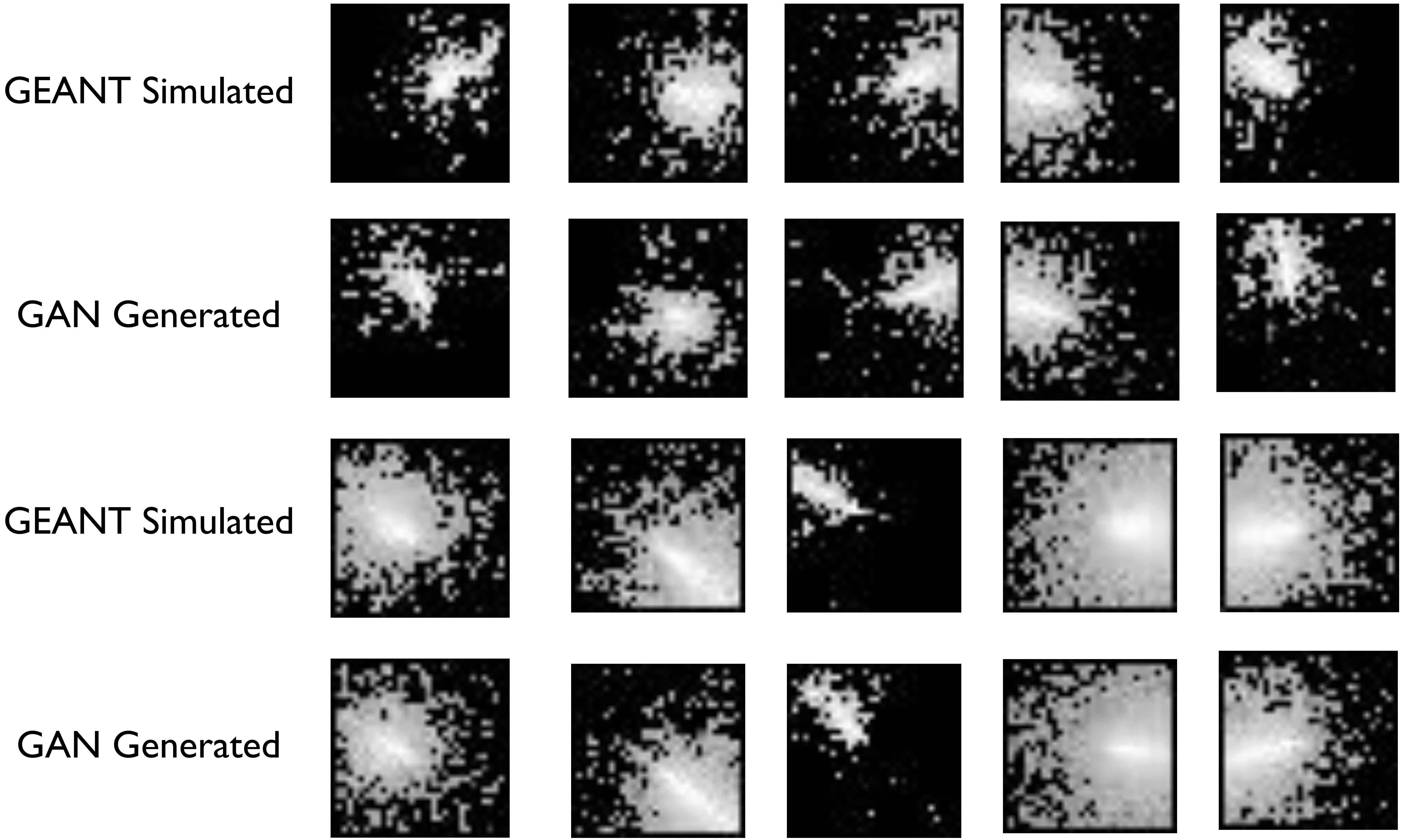


Conditional WGAN

$$\min_{G} \mathbb{E}_{x,\epsilon \sim p(x,\epsilon)}[-D(G(x,\epsilon))] + \mu||\hat{x} - x||_1$$



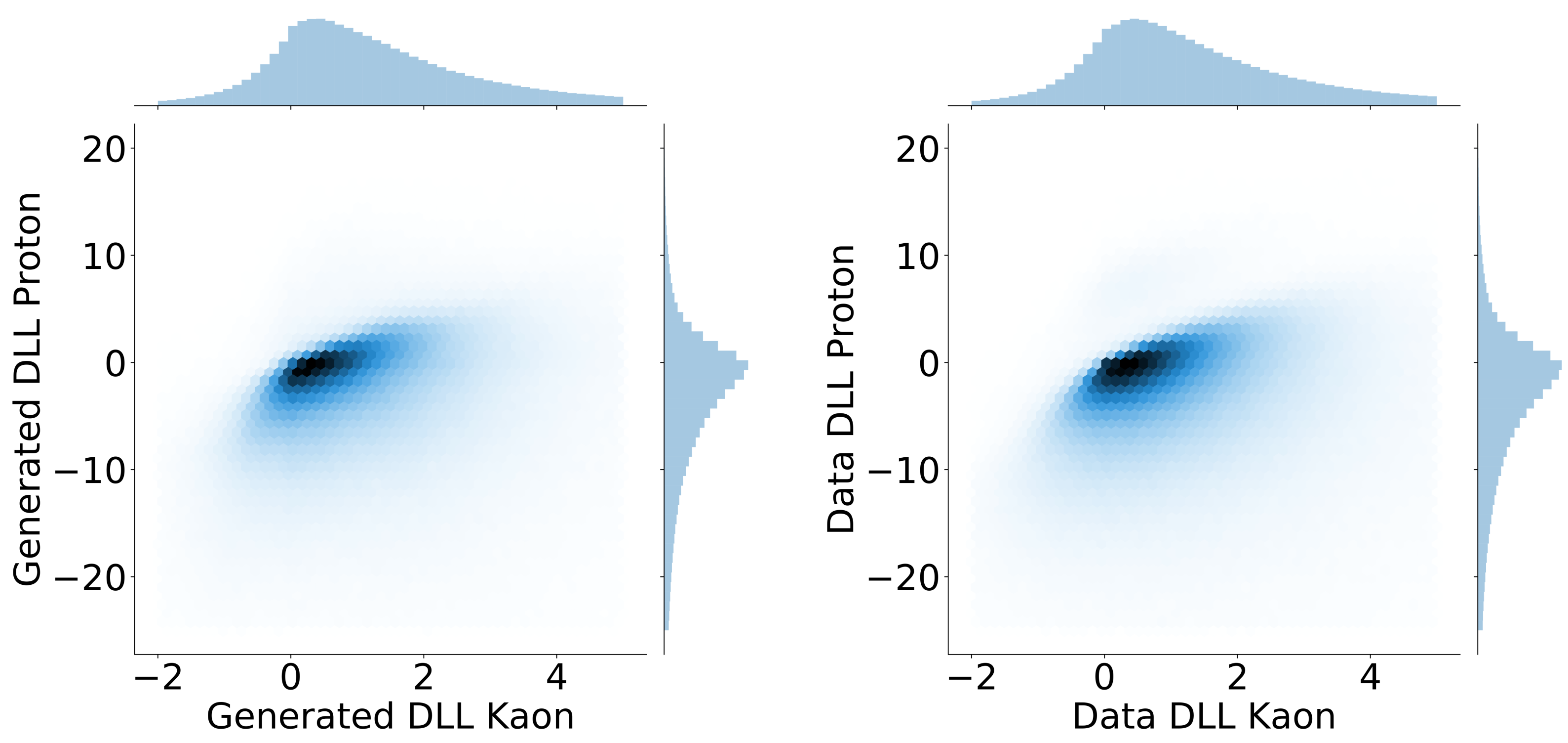
$$\max_{D} \mathbb{E}_{\hat{y}\sim p(\hat{y})}[D(\hat{y})) - \mathbb{E}_{y\sim p(y)}[D(y)] + \lambda\mathbb{E}_{\tilde{y}\sim p(\tilde{y})}[(||\nabla_{\tilde{y}} D(\tilde{y})||_2 - 1)^2]$$

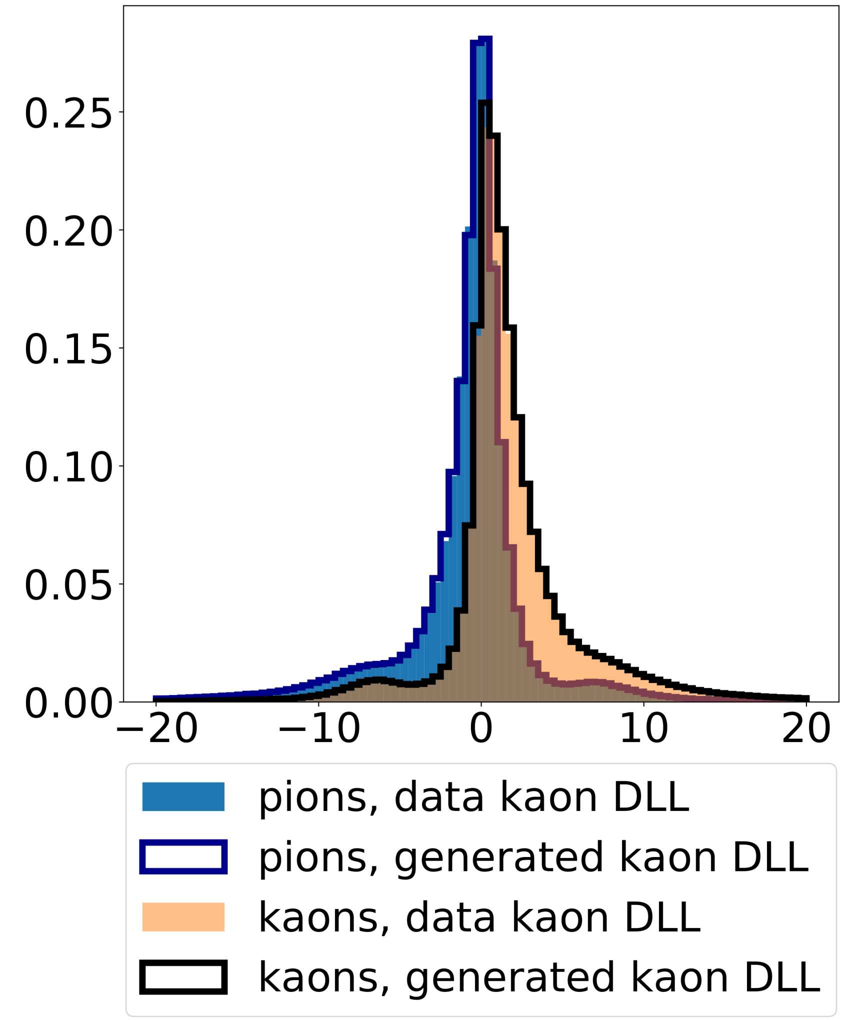


http://arxiv.org/abs/arXiv:1812.01319

Cherenkov Fast Simulation

- Plots are a pilot study on **BaBar DIRC MC**
- π vs **K** AUC difference ~0.01





https://rich2018.org/indico/event/1/contributions/89/

Conclusions

Machine learning applications in HEP are numerous.
And the amount of emerging areas is growing fast.

New challenges arise with upgrade of LHC and new experimental setups constructed around the world.

Should you have any data set with an interesting problem - let us know!

More unknown challenges ahead!