LHCb ML Challenges Highlights

2020 January 23, CERN openlab technical meeting

Andrey Ustyuzhanin on behalf of LHCb

NRU HSE

YSDA

ICL

LHCb experiment intro

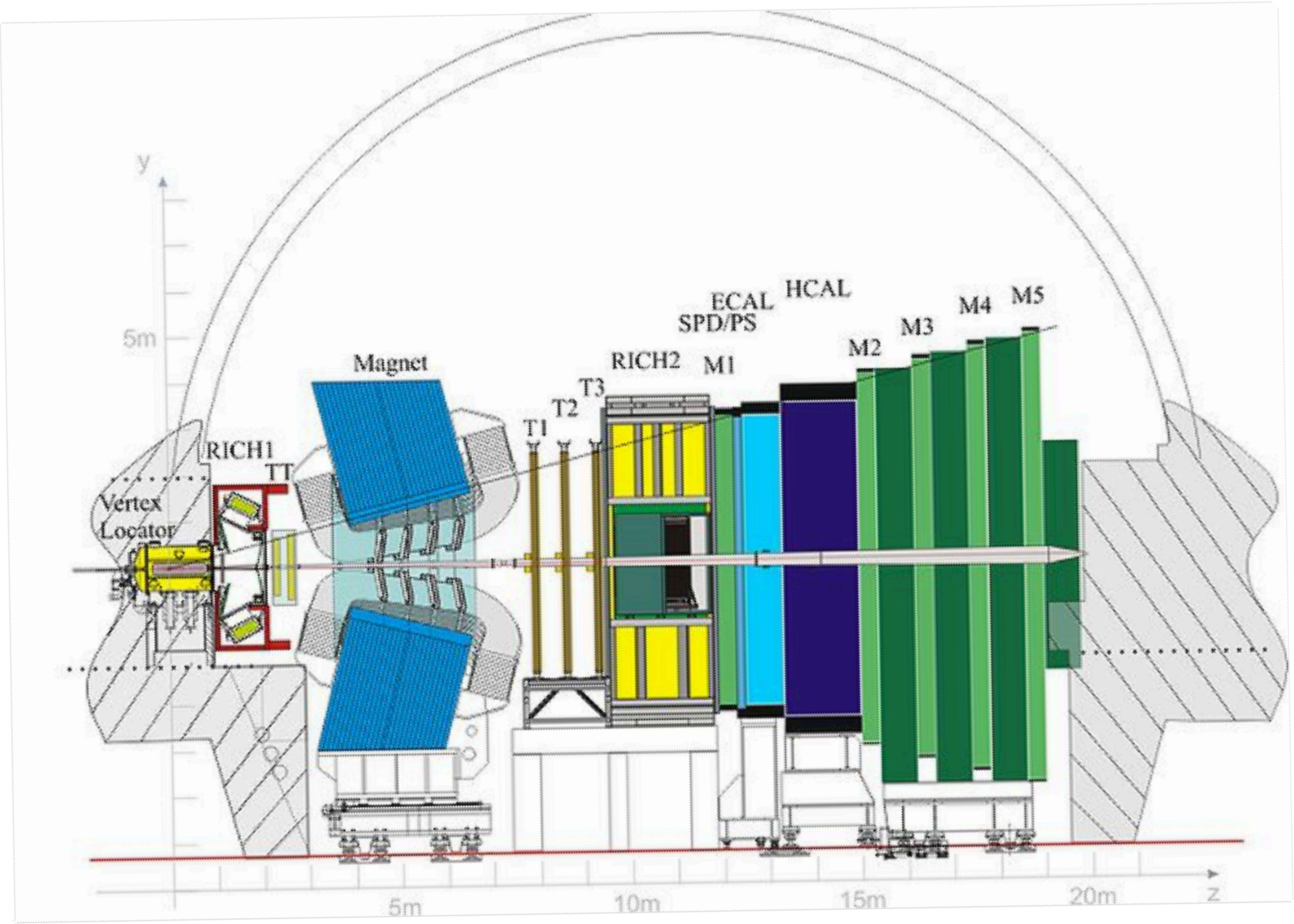
Physics channels

- flavour physics,
- Electro-Weak,
- high PT,
- Lepton-Flavour Violation.

2019-21 Upgrade challenges:

- order of magnitude higher signal yield,
- Increased pile-up,
- upgrade detector hardware,
- change in the data analysis paradigm

(real-time analysis, RTA).

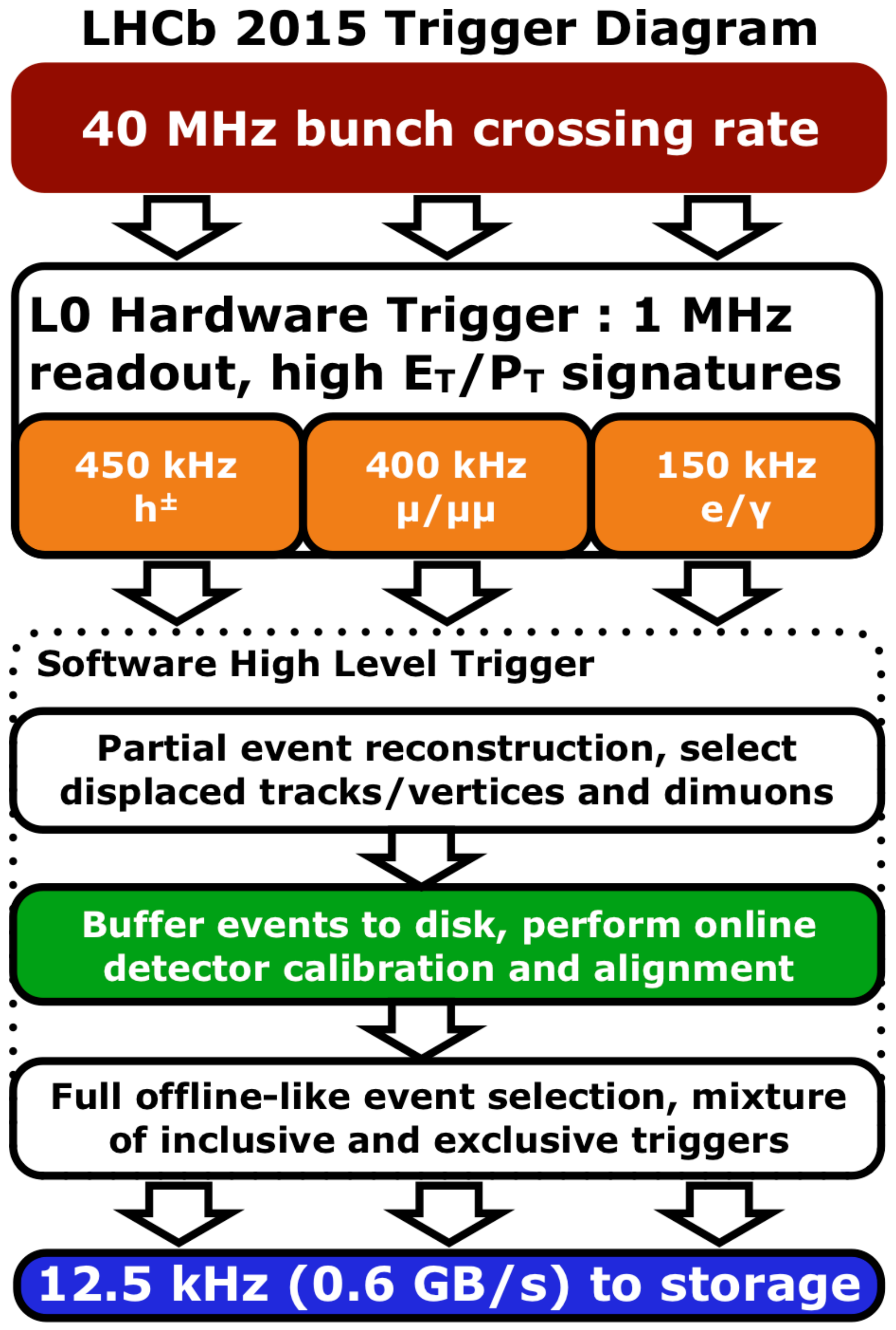


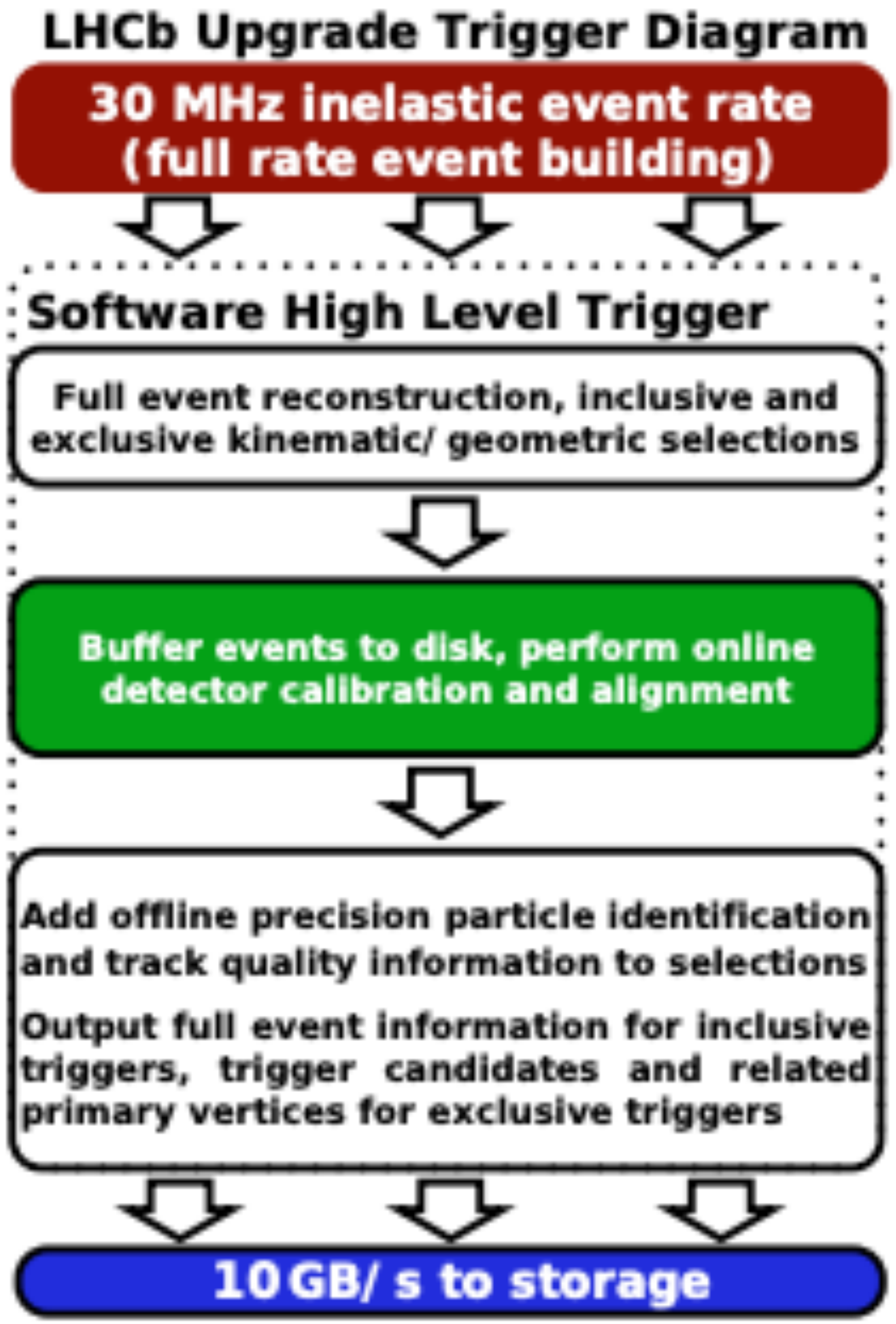
Real Time Analysis Project

RTA develops and maintains the real-time processing of LHCb's data for Run 3 and beyond.

Project work packages:

- Data structures
- Event Reconstruction
- Event Selection
- Align & Calibration
- Data QA
- Hardware Accelerators

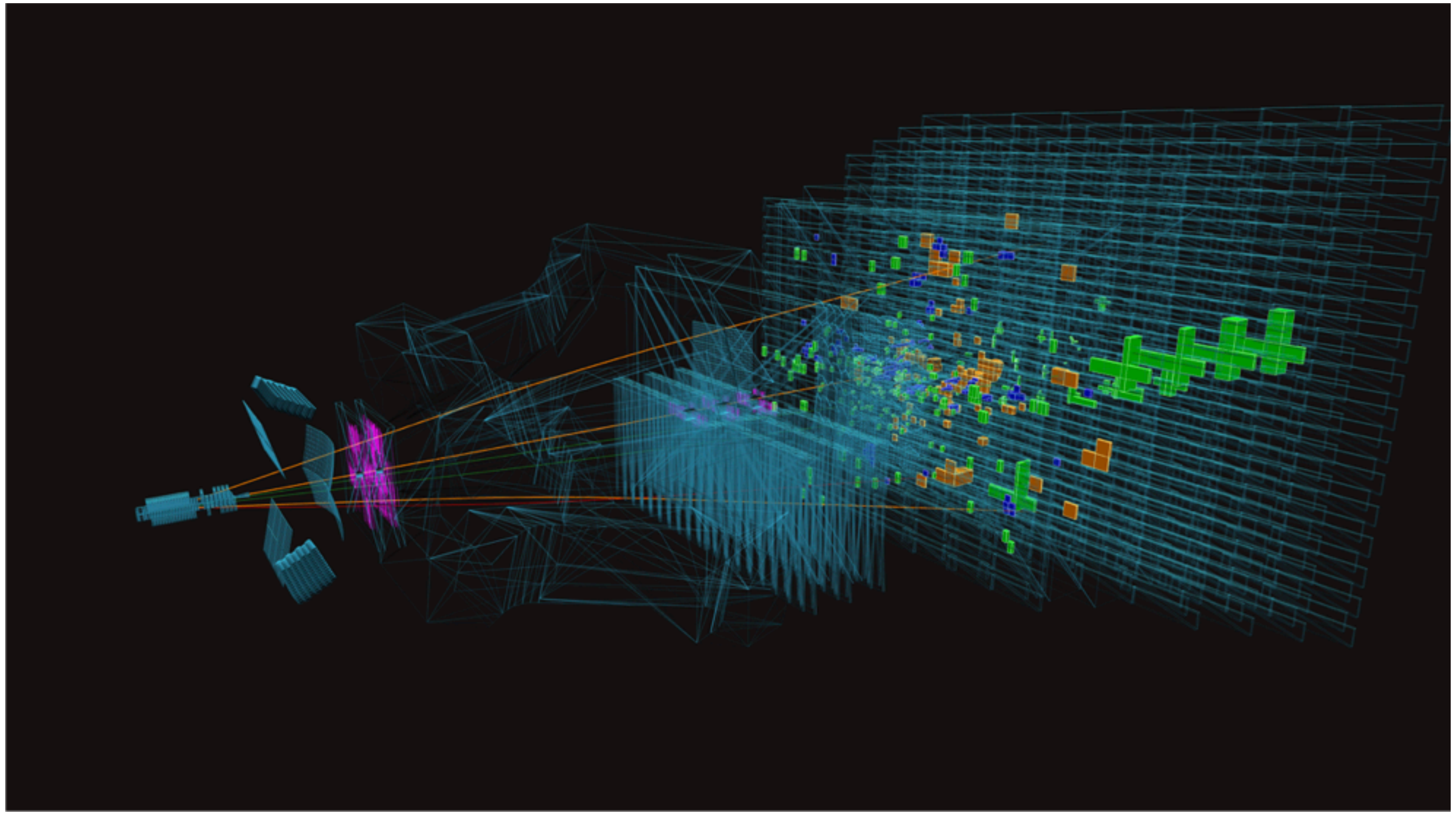




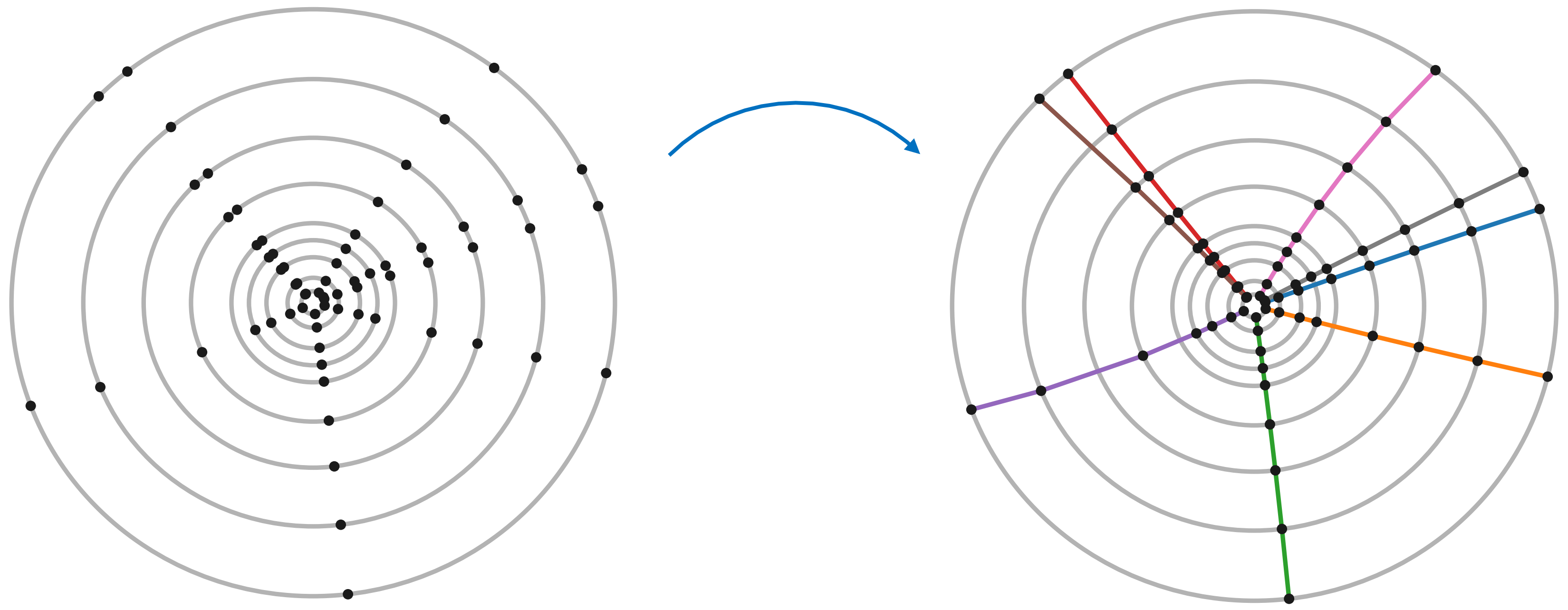
Event Reconstruction

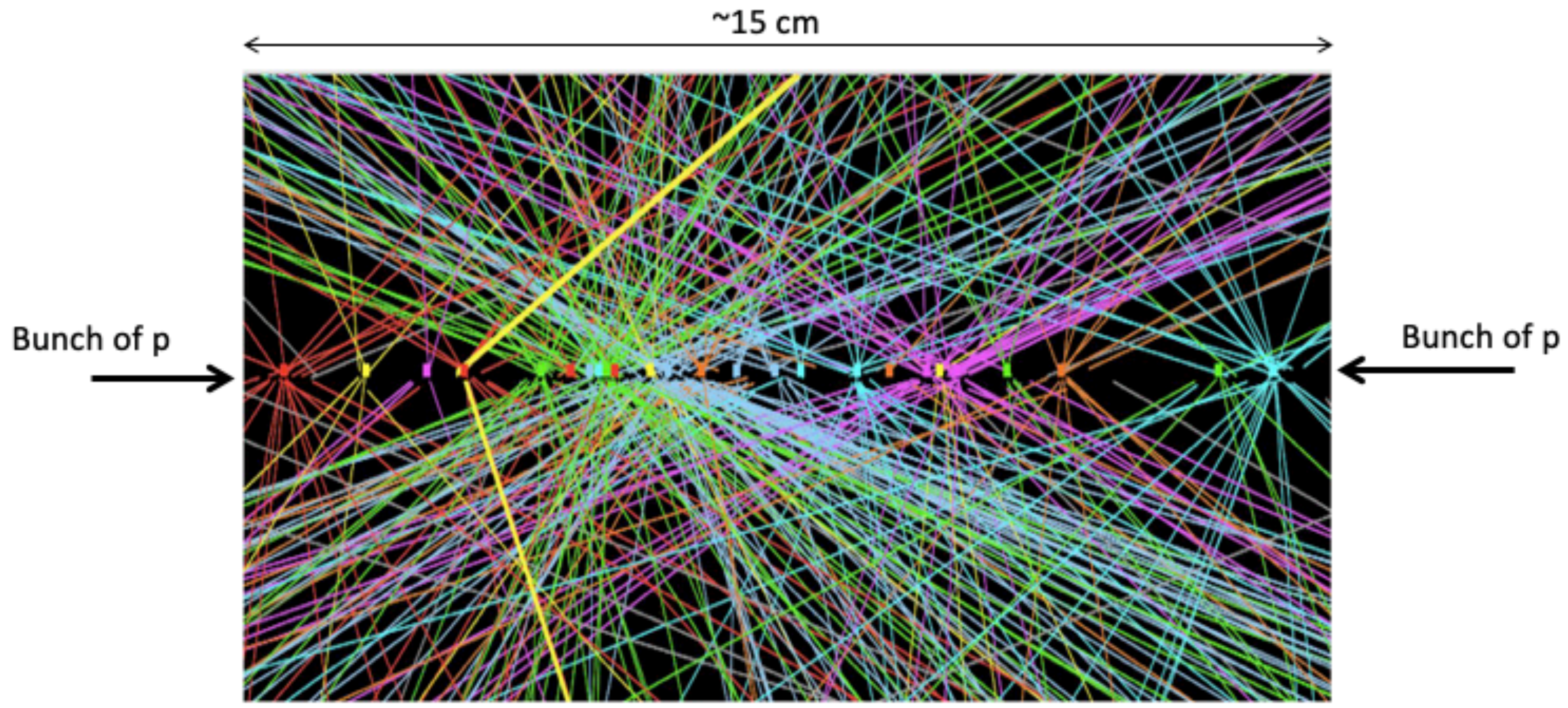
Reduce dimensionality of raw event by analyzing and combining information from subdetectors :

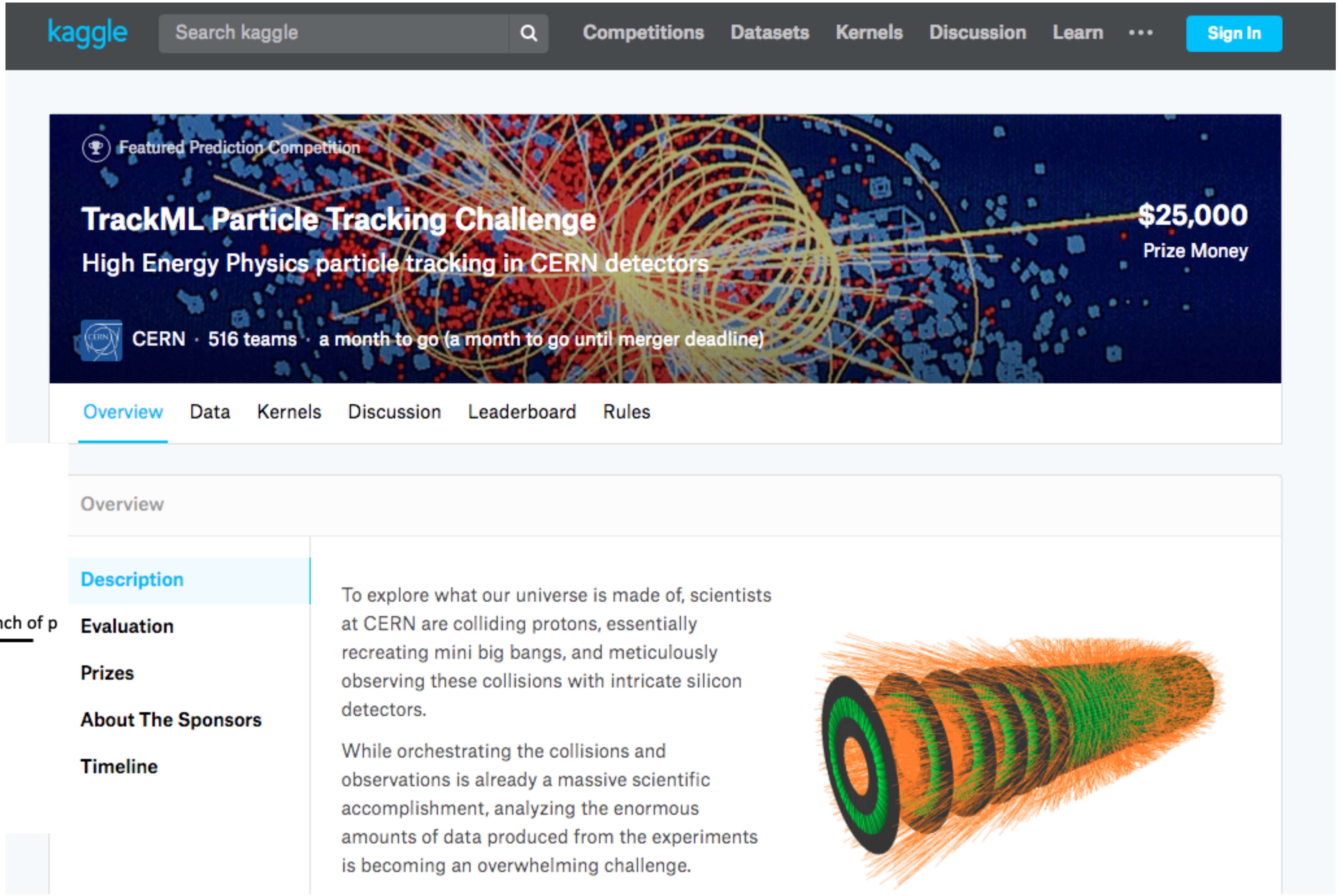
VELO
Tracker
RIng CHerenkov
Calorimeter
Muon Chambers



Tracking Machine Learning (ML) challenge





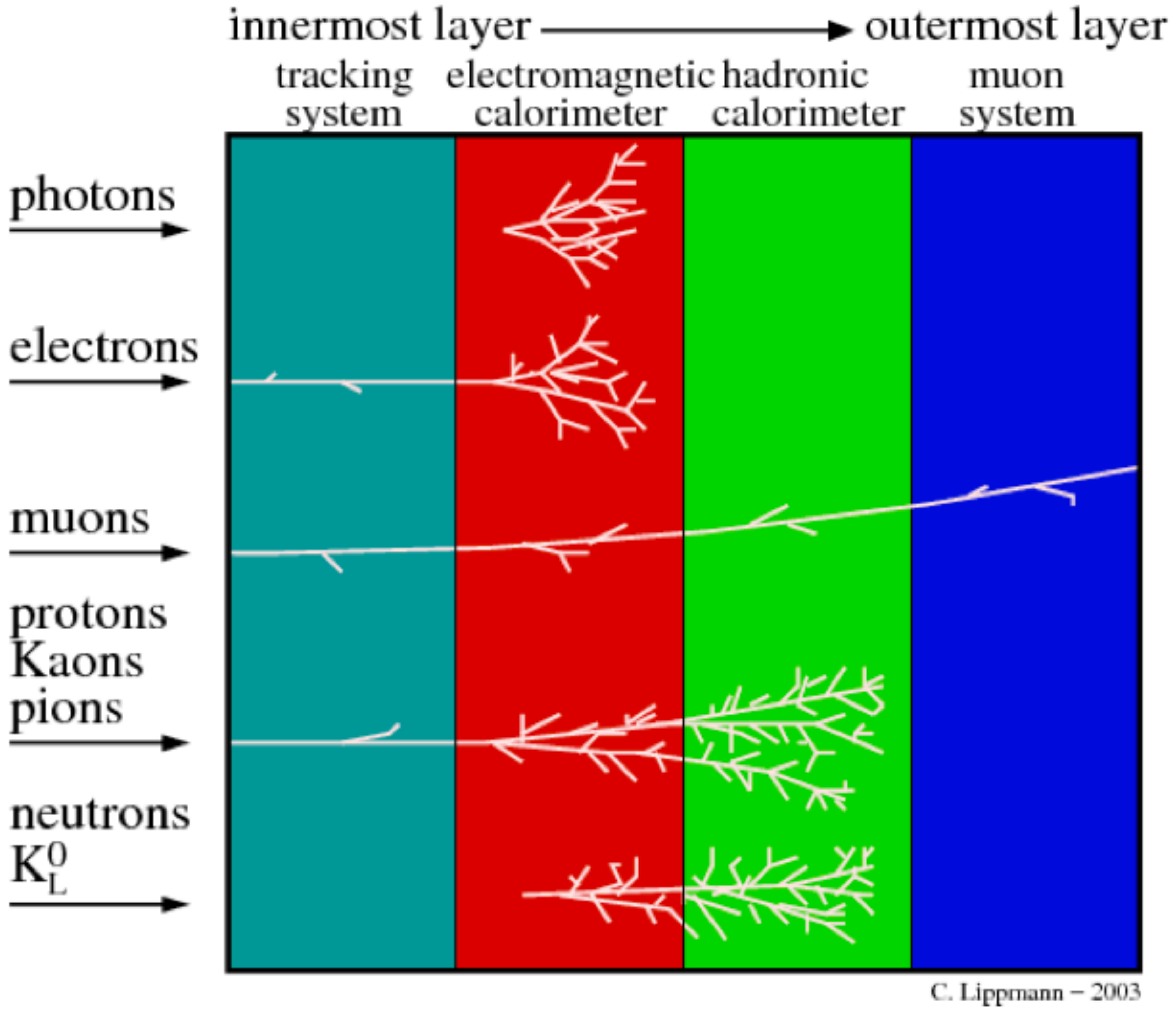


https://indico.cern.ch/event/813759/

Particle Identification (PID)

Combine information from sub detectors for identifying type of a track or particle

- Ring Cherenkov (RICH)
- Electromagnetic Calorimeter
- Hadron Calorimeter
- Muon Chambers



RICH PID using convolutional neural networks

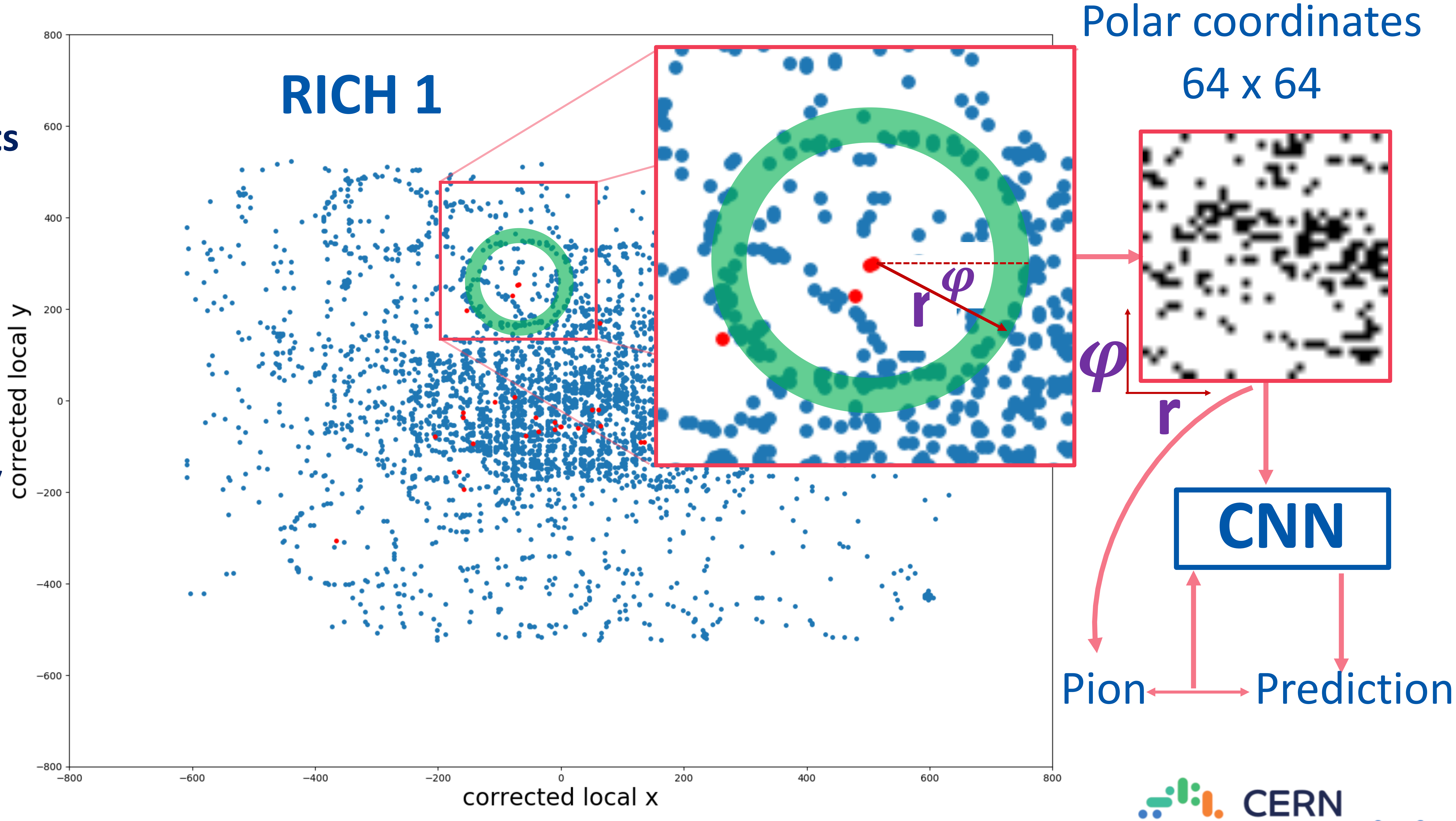
Michele Blago, Daniel Campora, Chris Jones

IN PROGRESS

MC data from LHCb reconstruction:**31K events in current dataset**.

Region around track centres **(depending on momentum range)** translated into polar coordinate **64 x 64** binary pixel images**.**

Labelled images used to train CNN.



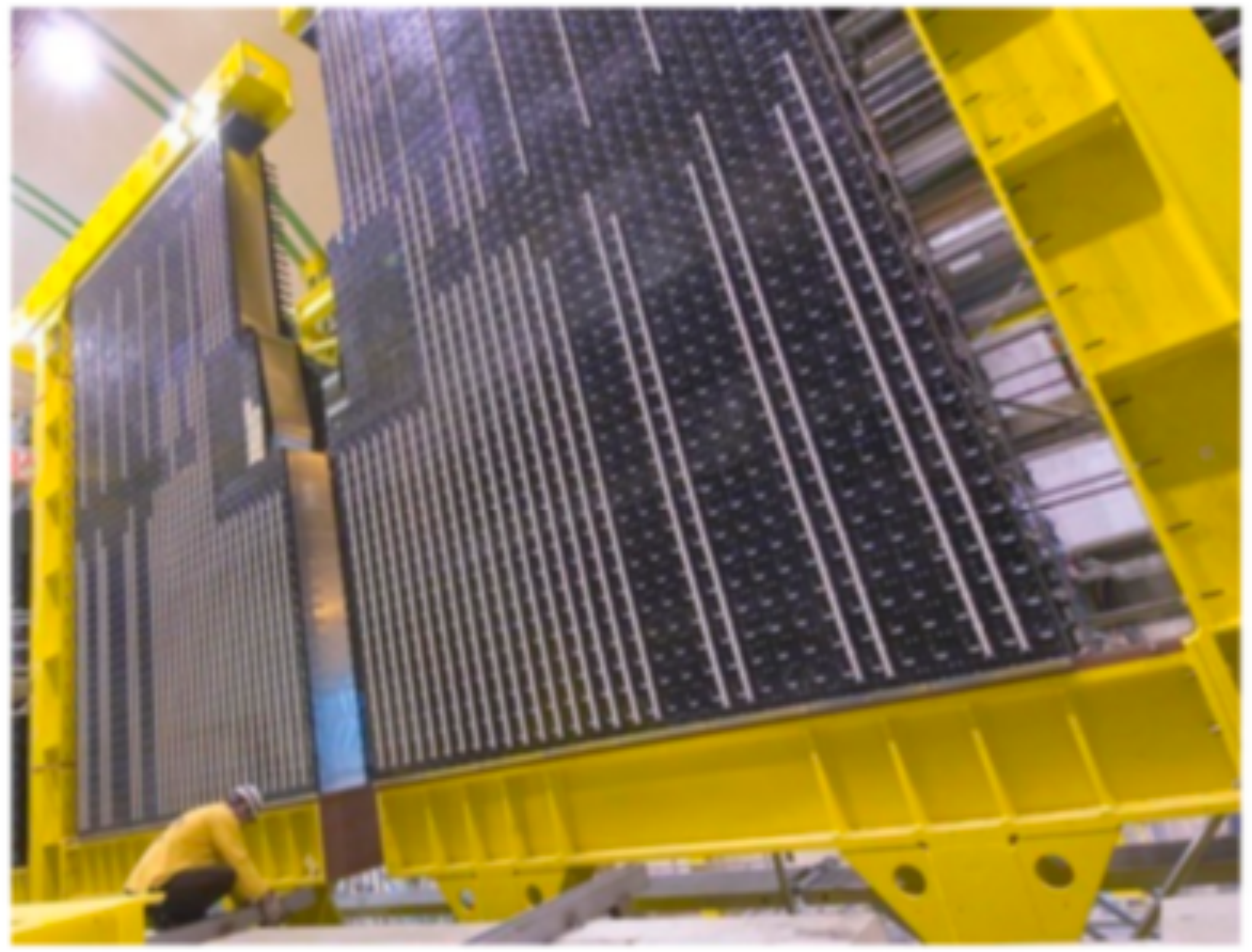
CERN openlab

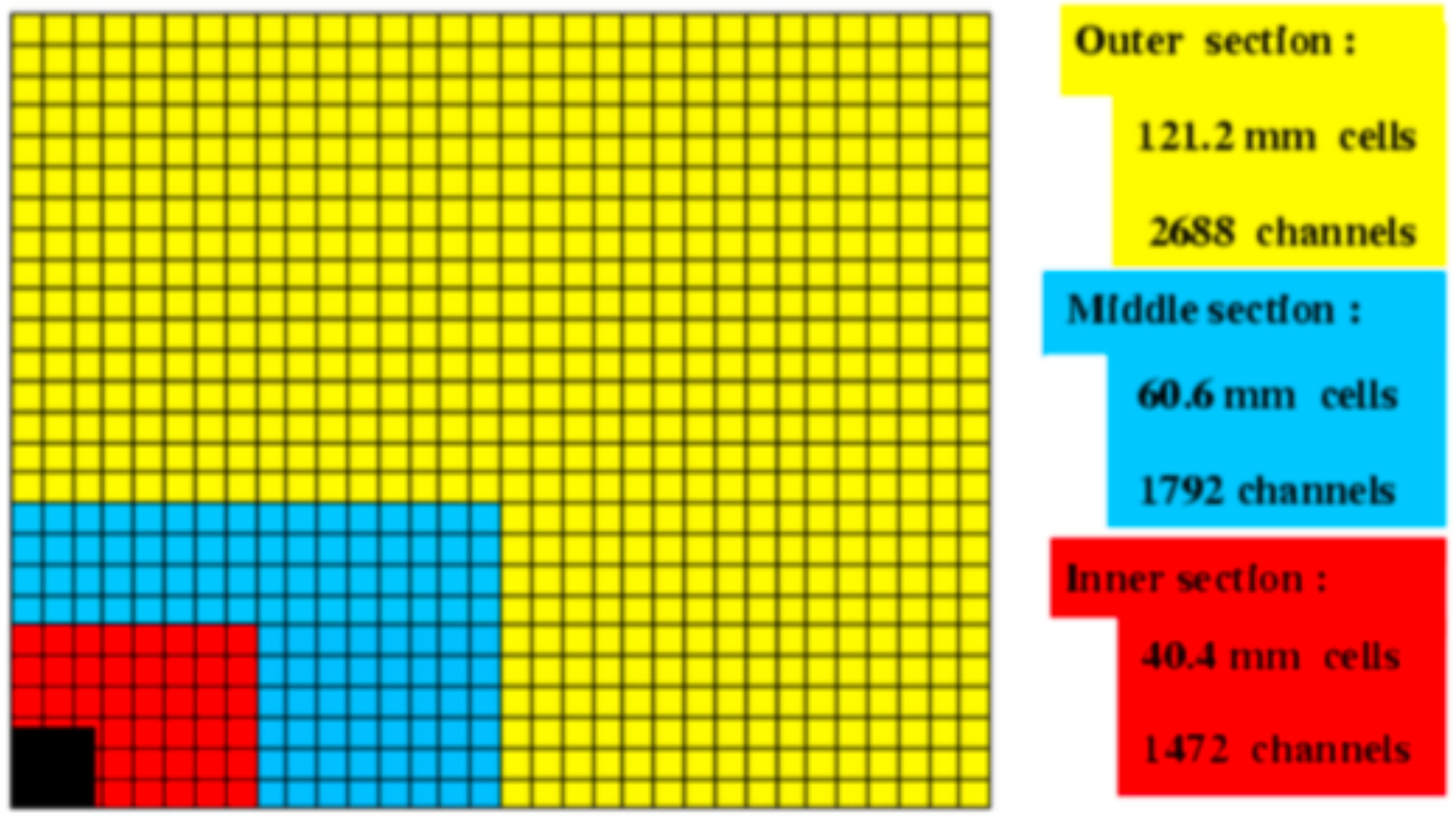
Deep Learning on LHCb Calorimetry

Blaise Delaney, Joao Coelho

IN PROGRESS

- Purpose of LHCb calorimeter system: trigger on e, γ, hadrons + measure energy and position from particle showers
- Broadly speaking, can think of such tasks as clustering, regression and classification
- Develop algorithms that can deal with realistic calorimeter geometry
- Use Graph-Neural Network based approach





CERN openlab

Event Selection challenges

LHCb will have O(1000) individual selections (filters) in HLT2 in Run 3, and many of these will be ML-based

Reproducibility of the model training

Interpretability of trained models:

- How can we ensure they're inclusive enough to select things we haven't thought about but selective enough to fit in the rate constraints?
- How big is the overlap?

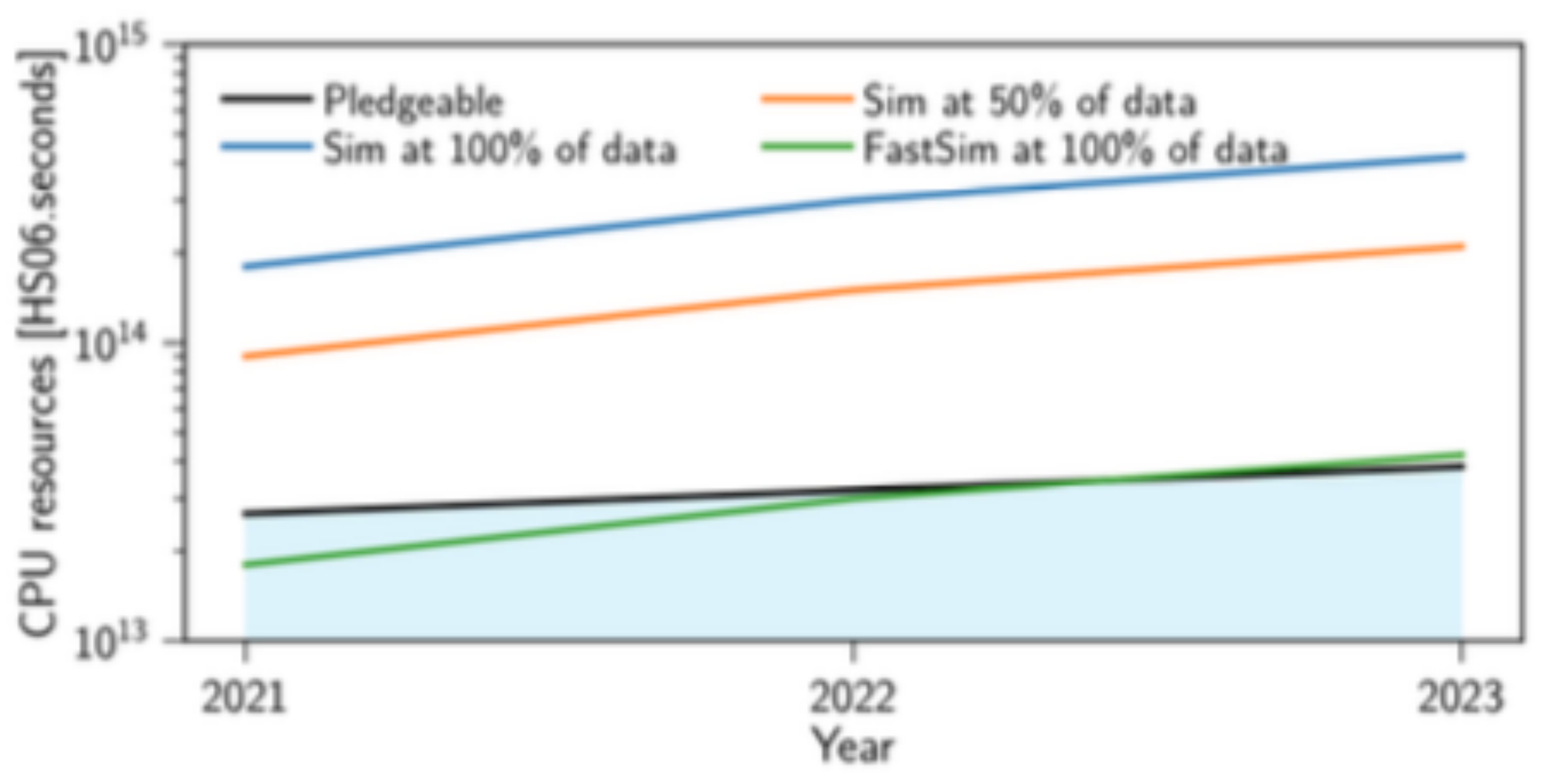
ML frameworks: support and transition from research to production

Generic RTA challenges

Inference speed vs accuracy of ML models for CPU & GPU
Model conversion from CPU to GPU
Pipeline for porting trained ML models to C++ stack
ML model uncertainty estimation and interpretability
Data acquisition quality certification and anomaly detection
New way of triggering on holistic event information https://arxiv.org/abs/1808.00711

Fast Simulation

Number of events to be simulated scales with the luminosity, and that the simulation time scales with pile-up, the CPU requirements will scale accordingly.



Fast simulation:

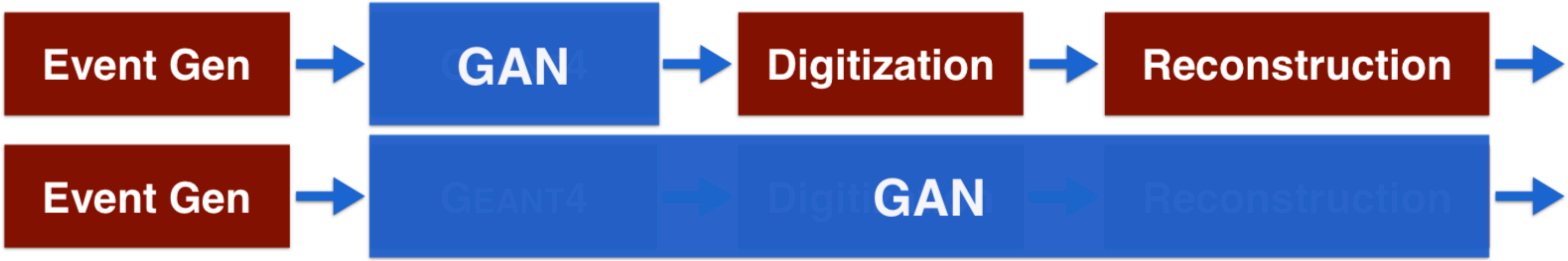
- ReDecay: only the signal part is simulated, while the same underlying event(s) are re-used several times
- RICHless: the Cherenkov photons and their computationally expensive propagation in the RICH detectors are not simulated
- TrackerOnly: only the tracking detectors are simulated
- ParticleGun: only signal or a small number of particles are simulated
- Shower library, https://bit.ly/2TPM0MG
- Generative models: use NN-based simulation

LHCb PID Simulation

- The LHCb PID response makes use of information from several subdetectors, namely the RICH detectors, the calorimeters and the muon detectors
- Simulation of the subdetectors devoted to PID is non-trivial – computing the detector response requires modelling of particle kinematics, detector occupancy and experimental conditions (alignments, temperature etc.)
- Simulation of the detector response using Geant is the most time-consuming stage of the full LHCb MC – time taken scales linearly with particle multiplicity

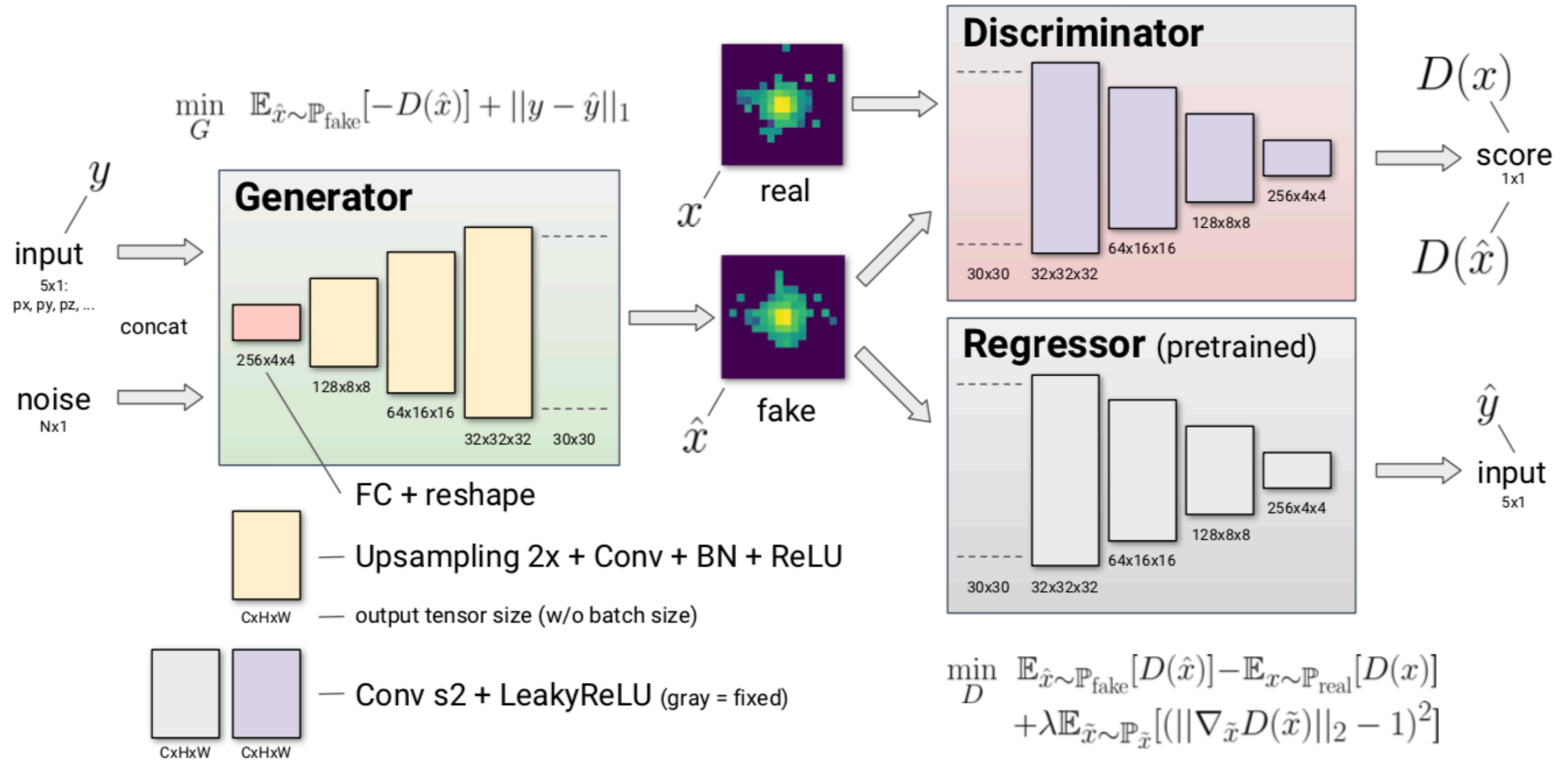
Typical simulation workflow

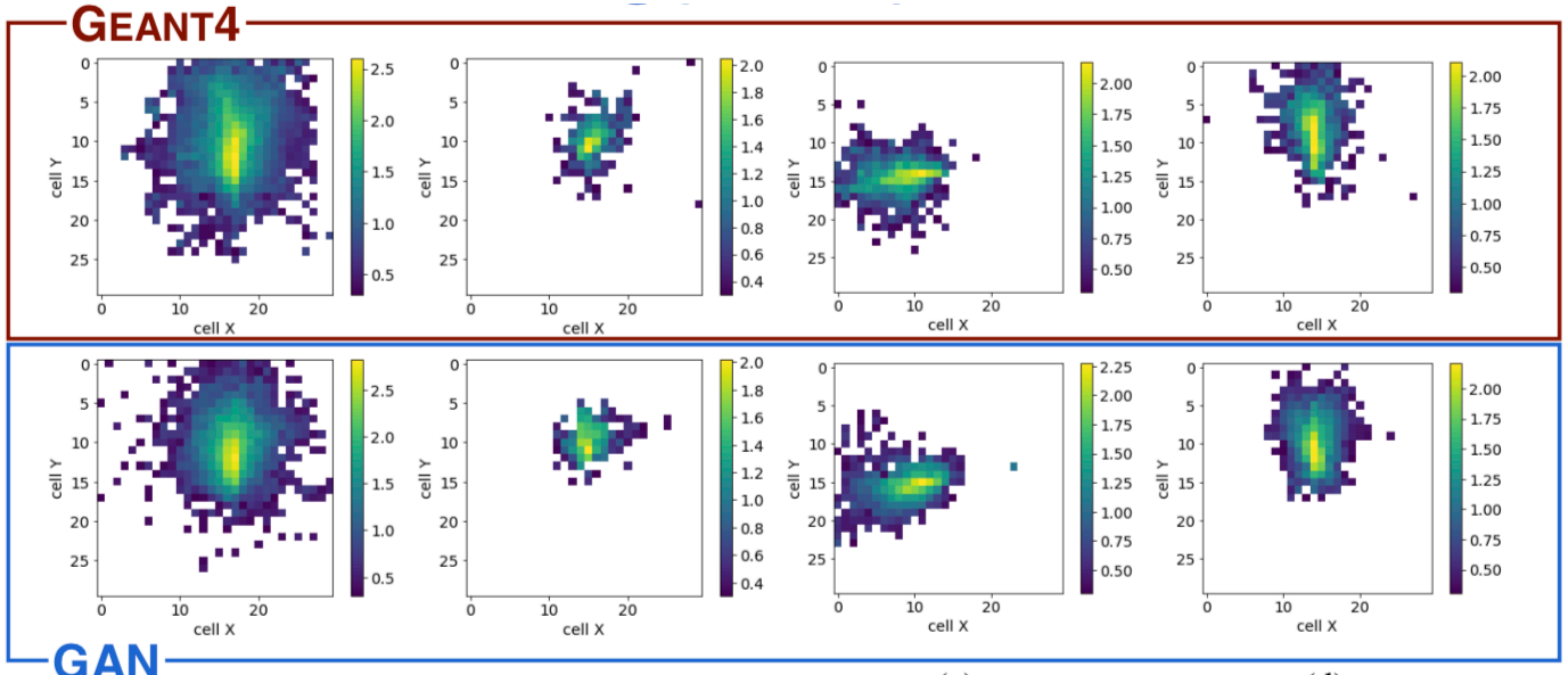
- One may imagine any part of this chain to be replaced by GAN
- Here we demonstrate two approaches:



Fast Calorimetry Simulation

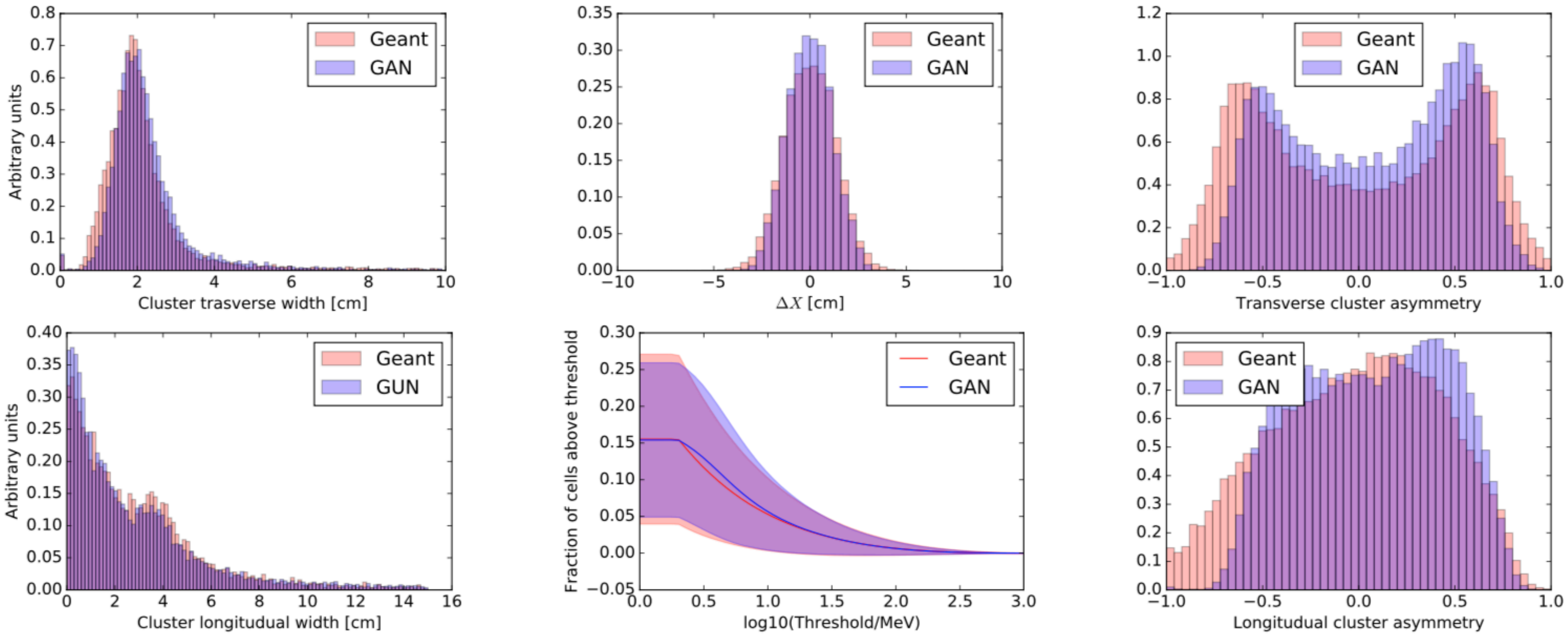
- LHCb-like calorimeter 30x30
- 5 conditional parameters per particle (3D momentum, 2D coordinate)
- Electrons from particle gun shot at 1x1 cm square at the center of the calorimeter face
- Approach: use GANs





(a) $E_0 = 63.7$ GeV (b) $E_0 = 6.5$ GeV (c) $E_0 = 15.6$ GeV (d) $E_0 = 15.9$ GeV

Quality assessment and open questions



- Visual similarity of raw features does not guarantee the similarity of higher-level characteristics
- How can we make sure tails of distribution are reproduced carefully enough?
- How can we estimate statistic and systematic uncertainty of such a model?

Very fast RICH simulation

Bypass all accurate simulation steps from Cherenkov light generation up to the high-level likelihood parameters (DLLs)
Learn the distribution of DLLs for given track parameters and sample from it,
P(DLLs | <track params>)

Derkach et al, NIMA 2019 (01) 031

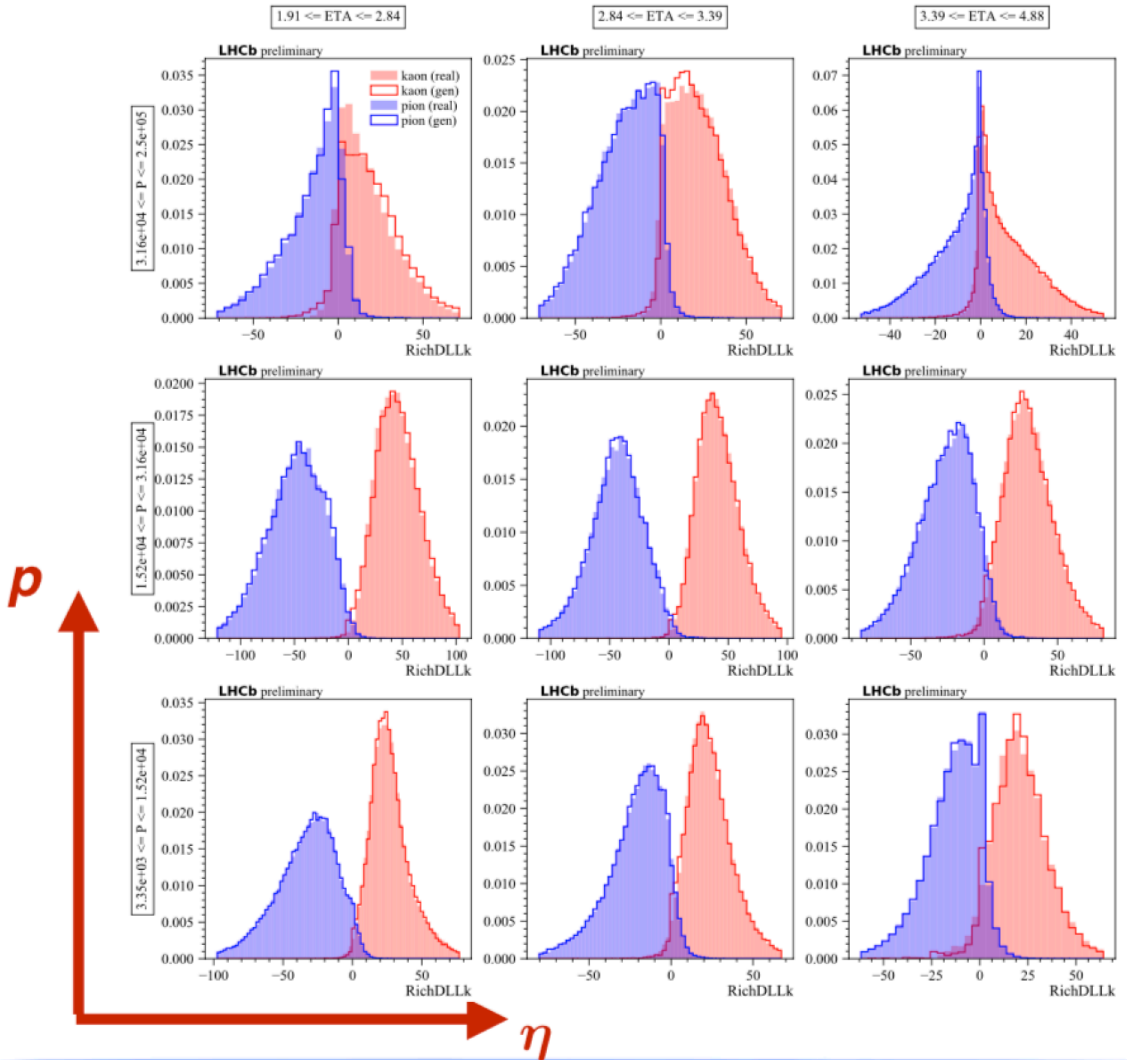
Number of input features:

- track momentum, pseudorapidity (+2)
- total number of tracks in that event (+1)

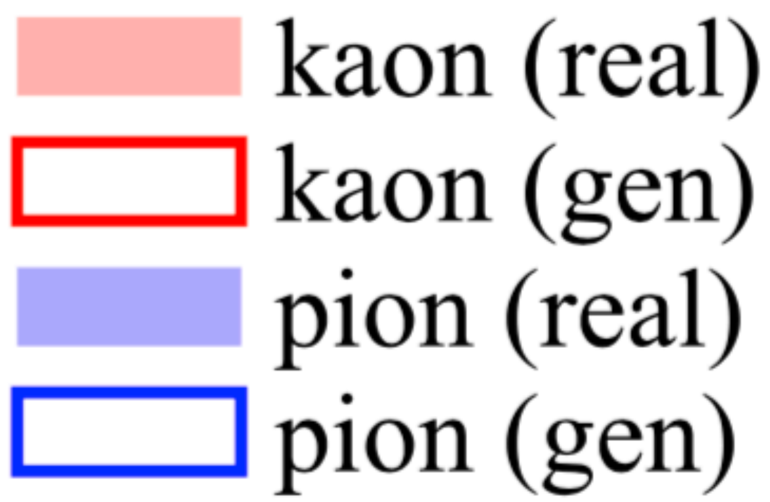
Number of output features: 5 DLLs
Training on real data (calibration channels) using sPlot technique1 to extract signal distributions

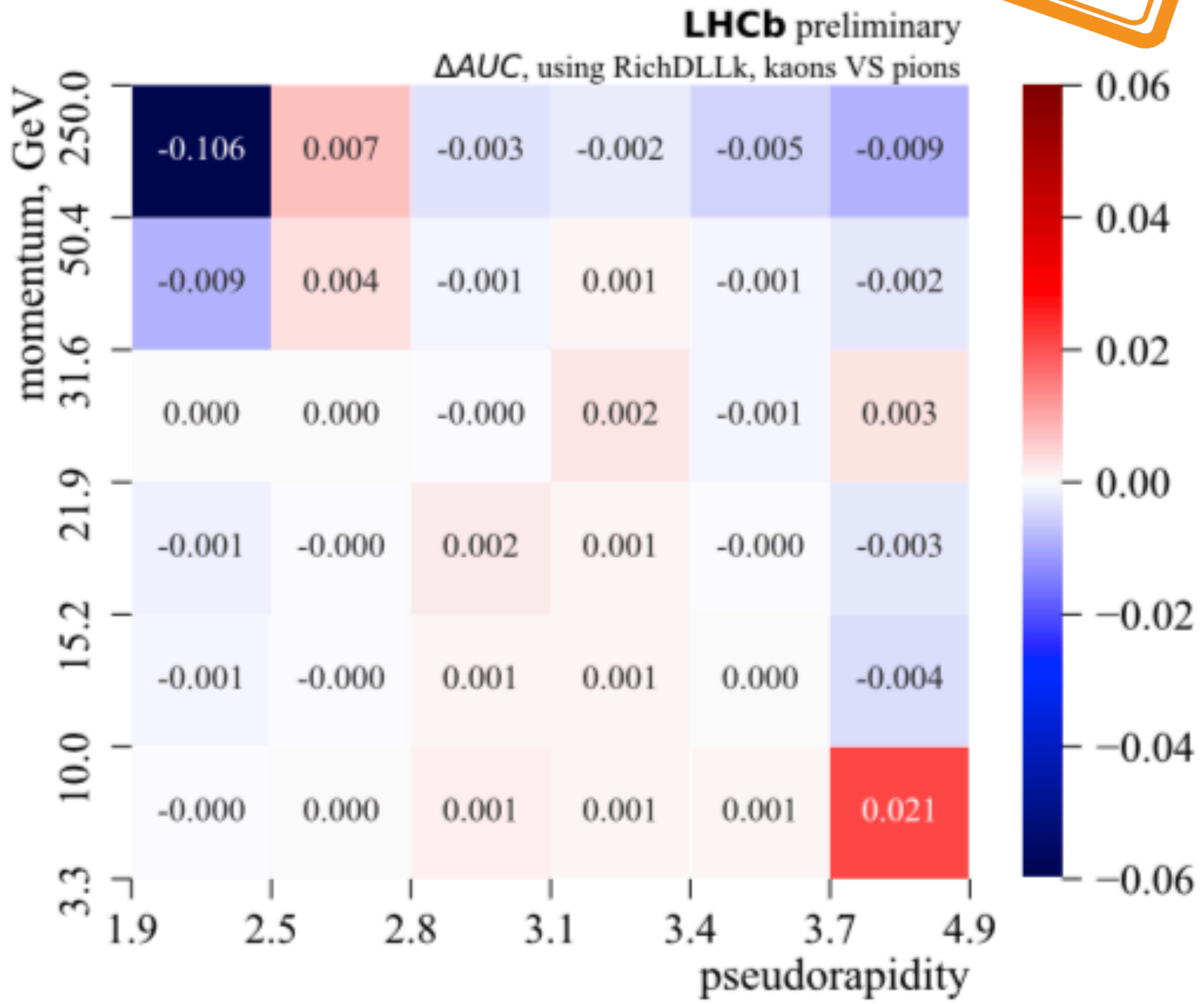
loss function is weighted
some of the weights are negative

Comparison



RichDLLk
(π vs K)

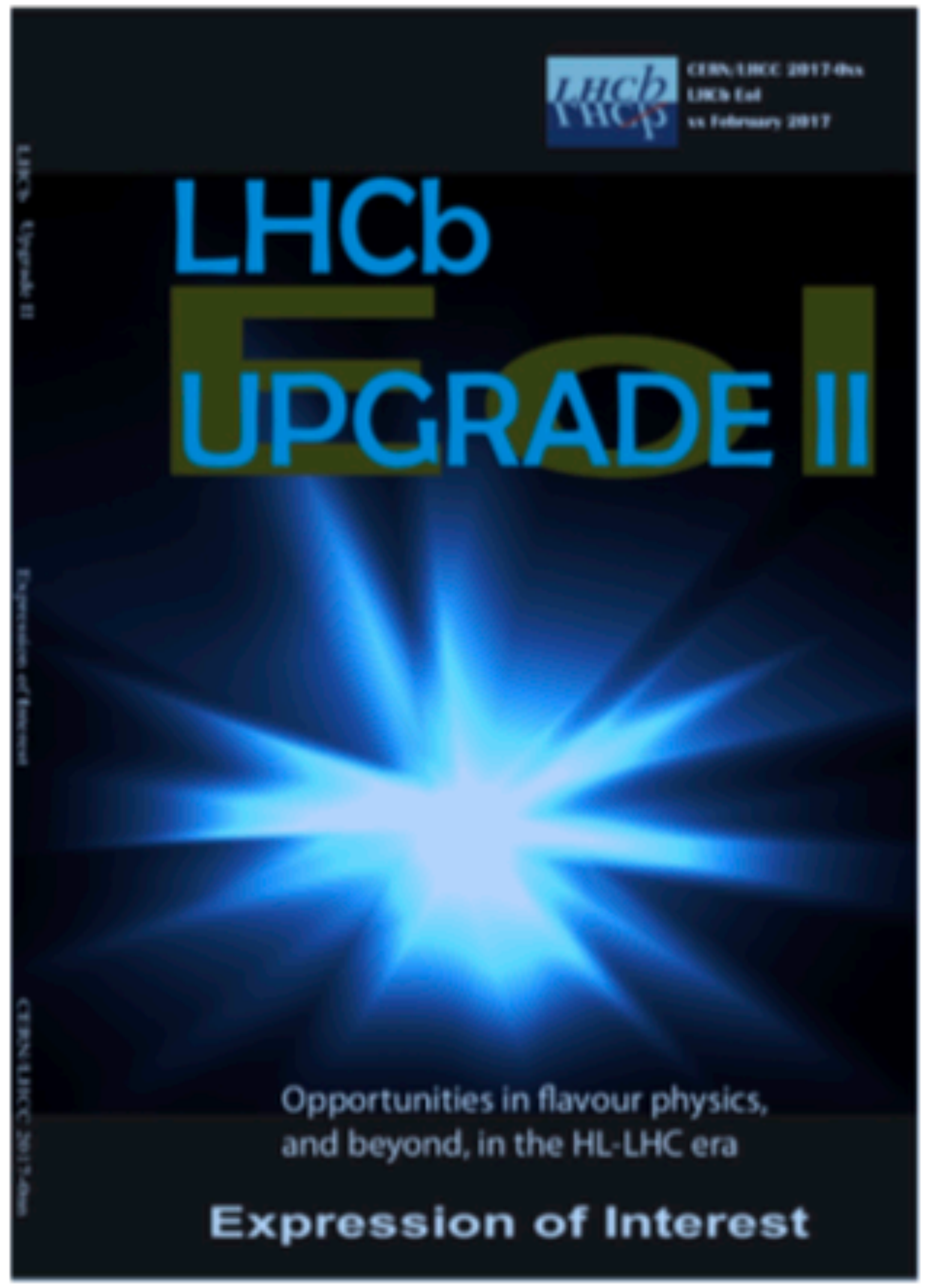


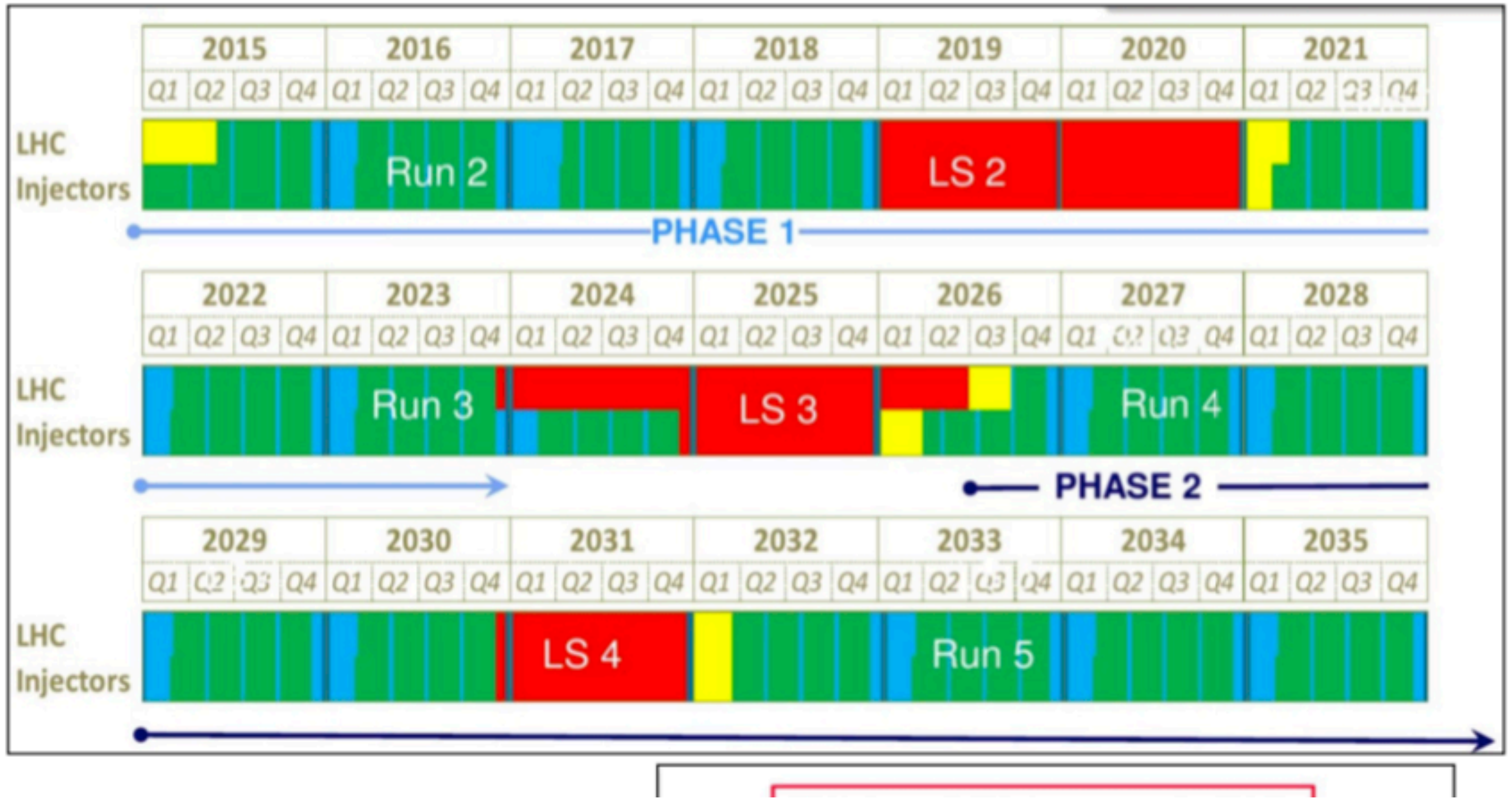


3x3 bin plot over full P-ETA rang

How to evaluate? **test in a physics analysis environment.**

Design optimisation

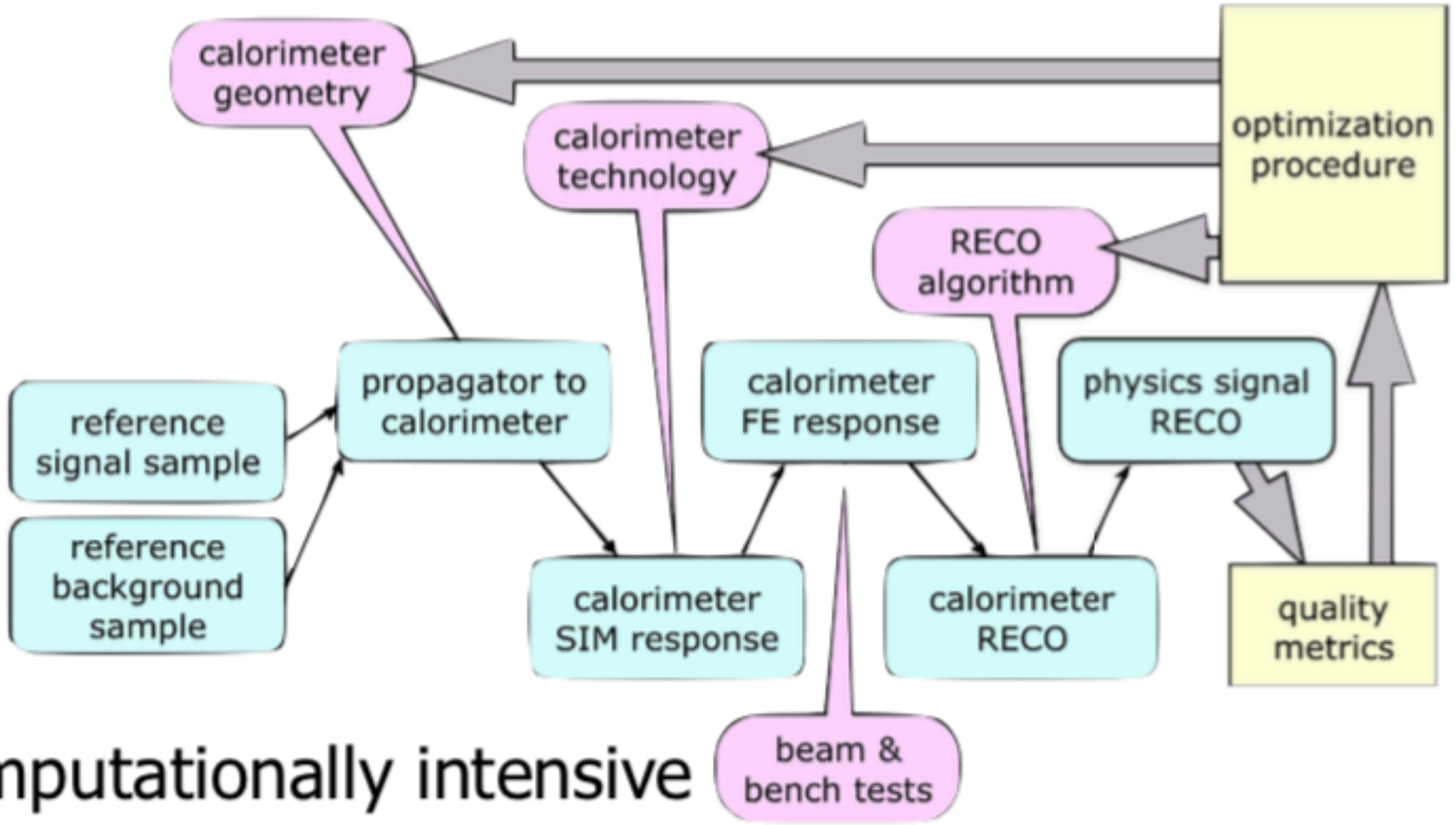




LHCb Upgrade II targets Run 5&6: 1.5e34 $cm^{-2}c^{-2}$ instantaneous luminosity

Requires extensive R&D studies for U2 LHCb ECAL including module technology, model configuration, readout properties, timing property, installation geometry

Optimization Cycle



Bottlenecks:

- calorimeter simulation is computationally intensive
  - shower development
  - photons transport
- direct beam and bench tests hard to directly include into simulation stack
- RECO algorithm needs tuning for the particular module technology/ geometry/configuration
- multi-parametric optimization may be expensive

https://bit.ly/2NMe4Nv

ML in the Optimization Cycle

- Machine Learning provides a set of tools and methods which allow effective fit of multi-dimensional data to non-parametric (generic) functions
  - allows quick turn over for the optimization cycle, when parameters are changed
  - eliminates manual work for re-tuning simulation and reconstruction
- ML model may be suboptimal comparing to “the best” solution
  - however it catches main features, that is usually good enough to estimate physics performance and give feedback to ongoing detector R&D

Optimisation Challenges

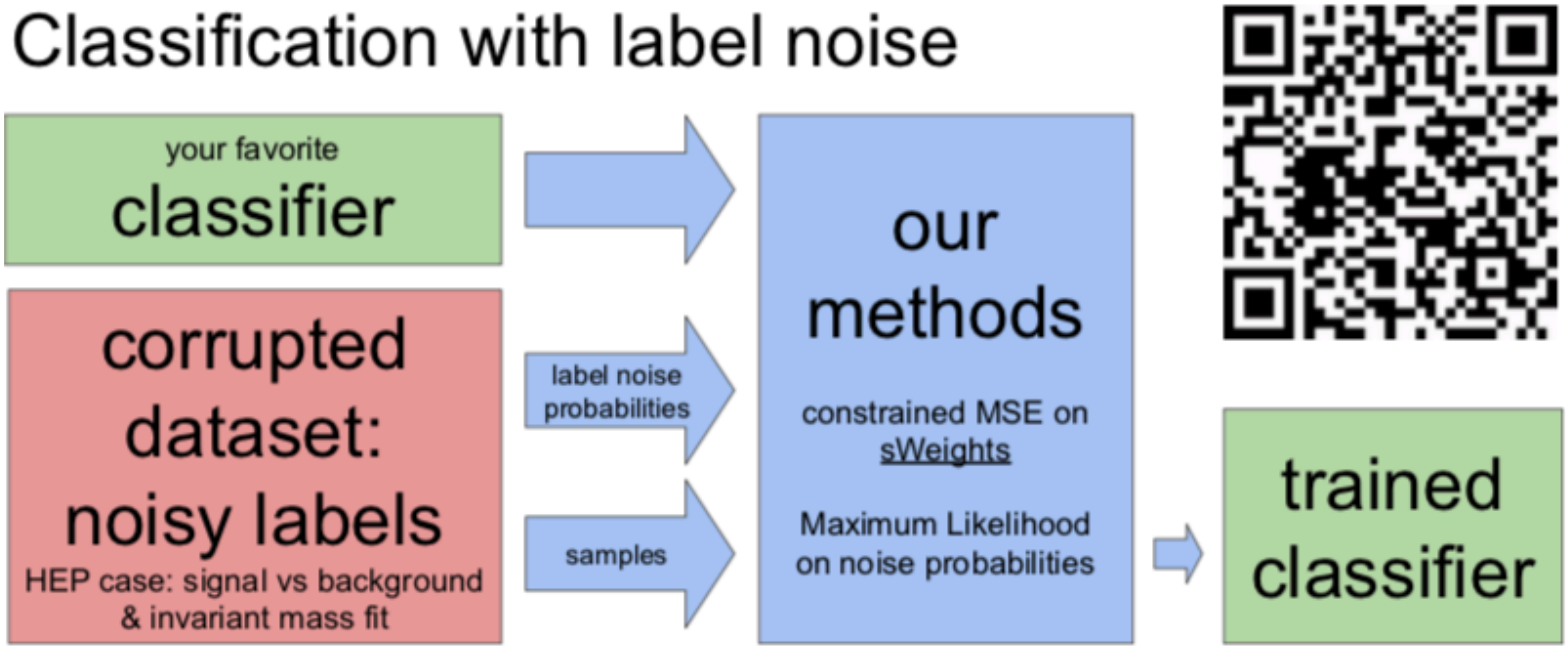
IN PROGRESS

- Many parameters to optimize simultaneously
  - E.g. granularity distribution in LHCb U2 ECAL
- Trade off between physics performance and costs
  - not obvious measure of success
  - non-differentiable optimization loss function
- Relatively long single iteration
- ML provides special methods developed for such use cases (e.g. Bayesian optimization)

Other

- For offline analysis: batch scheduling system support for TensorFlow, GPUs, multi-core training and inference
- Unsupervised algorithms e.g. Data Quality and for the new physics search (https://arxiv.org/abs/1811.10276)
- Efficient sampling algorithms
- Training with noisy labels (next slide)

ML on background-contaminated data



https://arxiv.org/abs/physics/0402083, sWeights intro

https://ml4physicalsciences.github.io/files/NeurIPS_ML4PS_2019_122.pdf

Conclusion

LHCb has Ambitious Physics goals for Run3-6
Long road aided with technical/infrastructure development
There is plenty of space for ML to shine, but it requires tailoring of generic methods to LHCb specifics