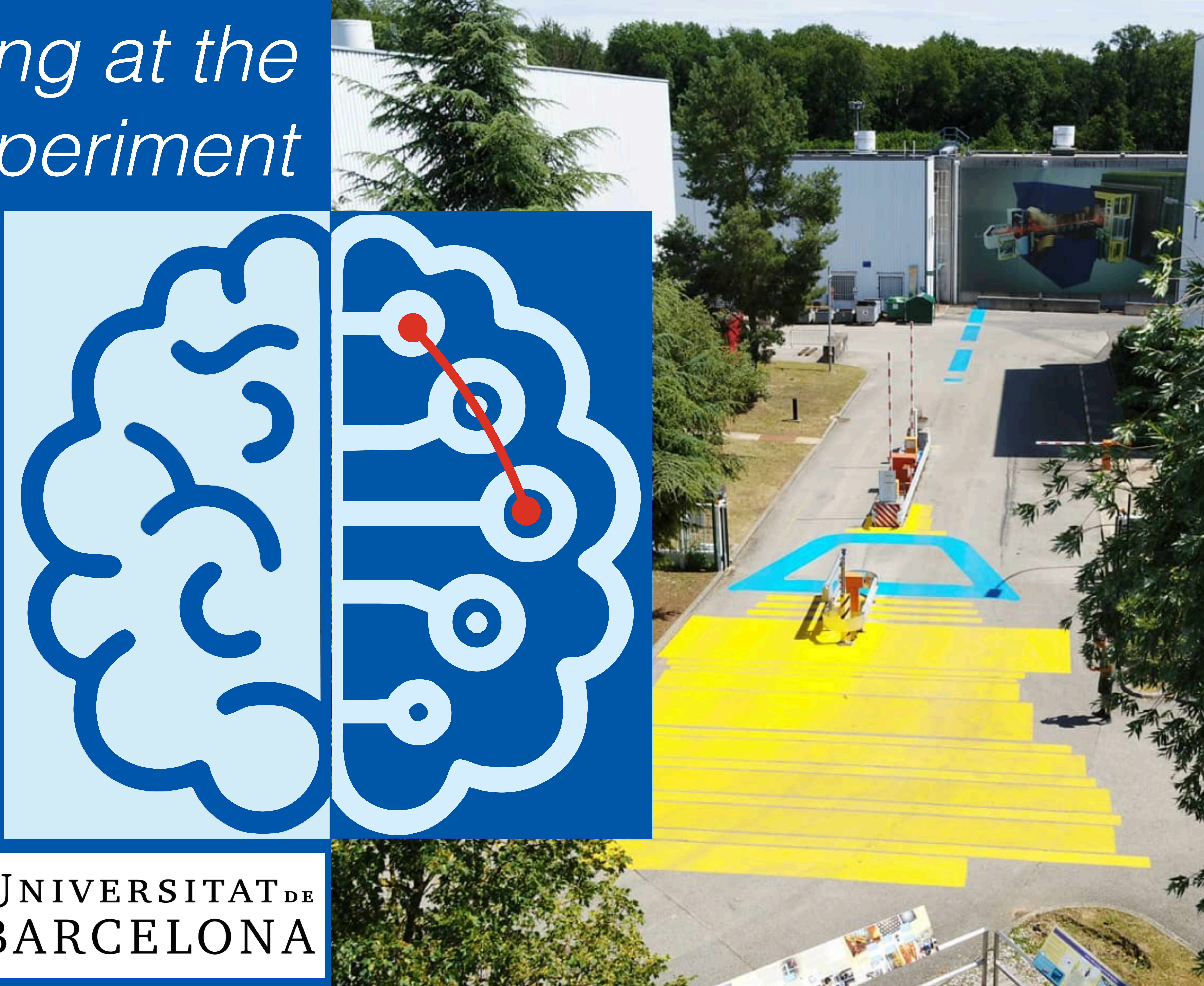
Machine learning at the LHCb experiment

Lukas Calefice[1]

[1] Universitat de Barcelona / ICCUB
lukas.calefice@cern.ch



ML Workshop at ICCUB, 24.10.2023

LHCb

erc European Research Council
Supporting top researchers from anywhere in the world

LIBERTAS PERFVNDET · OMNIA LVCE
UNIVERSITAT DE BARCELONA

Large Hadron Collider beauty experiment (LHCb)

- High-energy physics experiment at the LHC at CERN
- Precision measurements of b and c hadron decays

Large Hadron Collider beauty experiment (LHCb)

CP violation / CKM angles

Rare decays

Hadron spectroscopy

- High-energy physics experiment at the LHC at CERN
- Precision measurements of b and c hadron decays

Large Hadron Collider beauty experiment (LHCb)

CP violation / CKM angles

Rare decays

Hadron spectroscopy

Exotica searches

Electroweak

Heavy ions/ fixed target

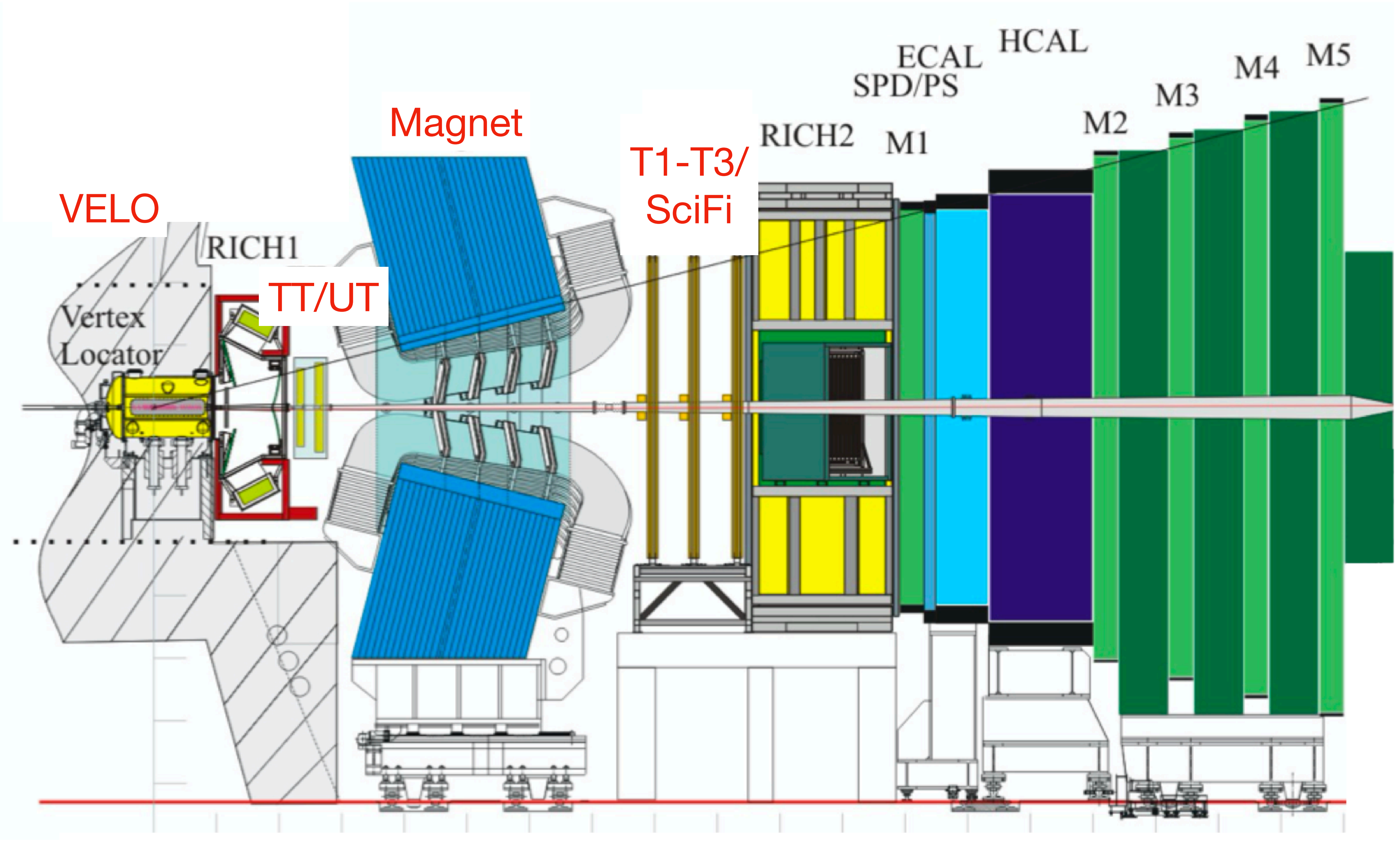
Semileptonic decays

Kaon physics

...

- High-energy physics experiment at the LHC at CERN
- Precision measurements of b and c hadron decays

Large Hadron Collider beauty experiment (LHCb)



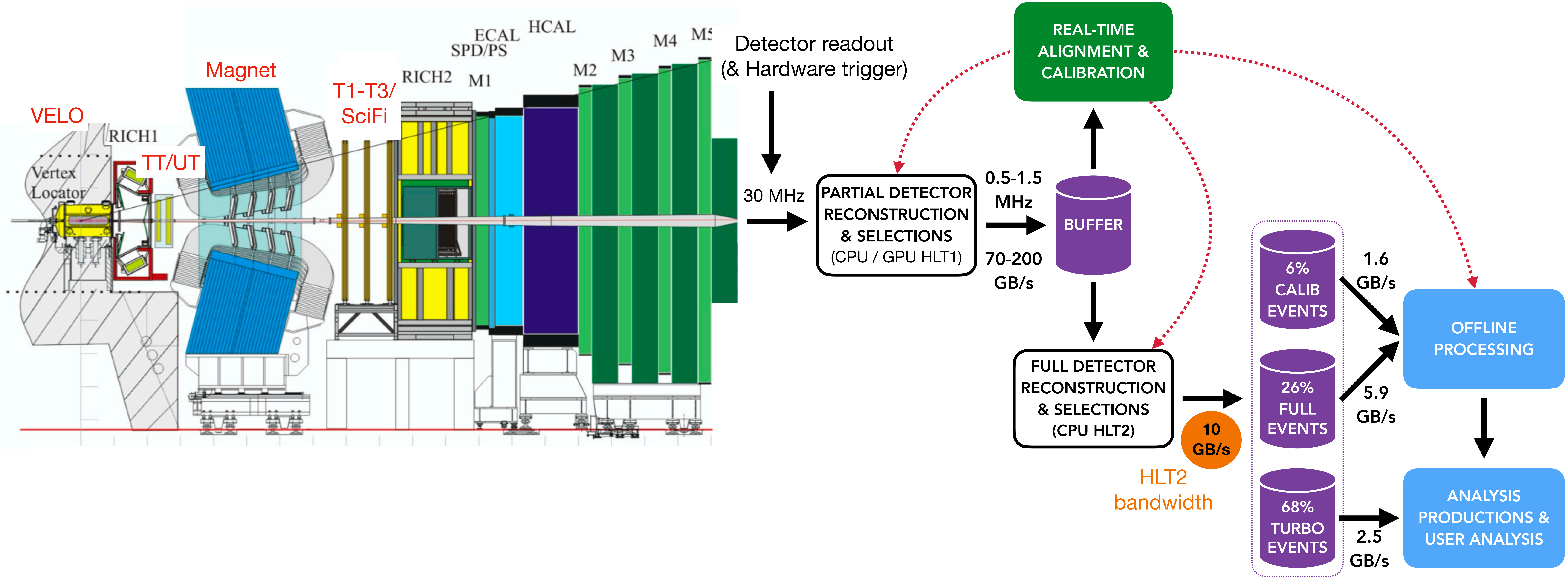
Two reconstruction tasks:

Track reconstruction

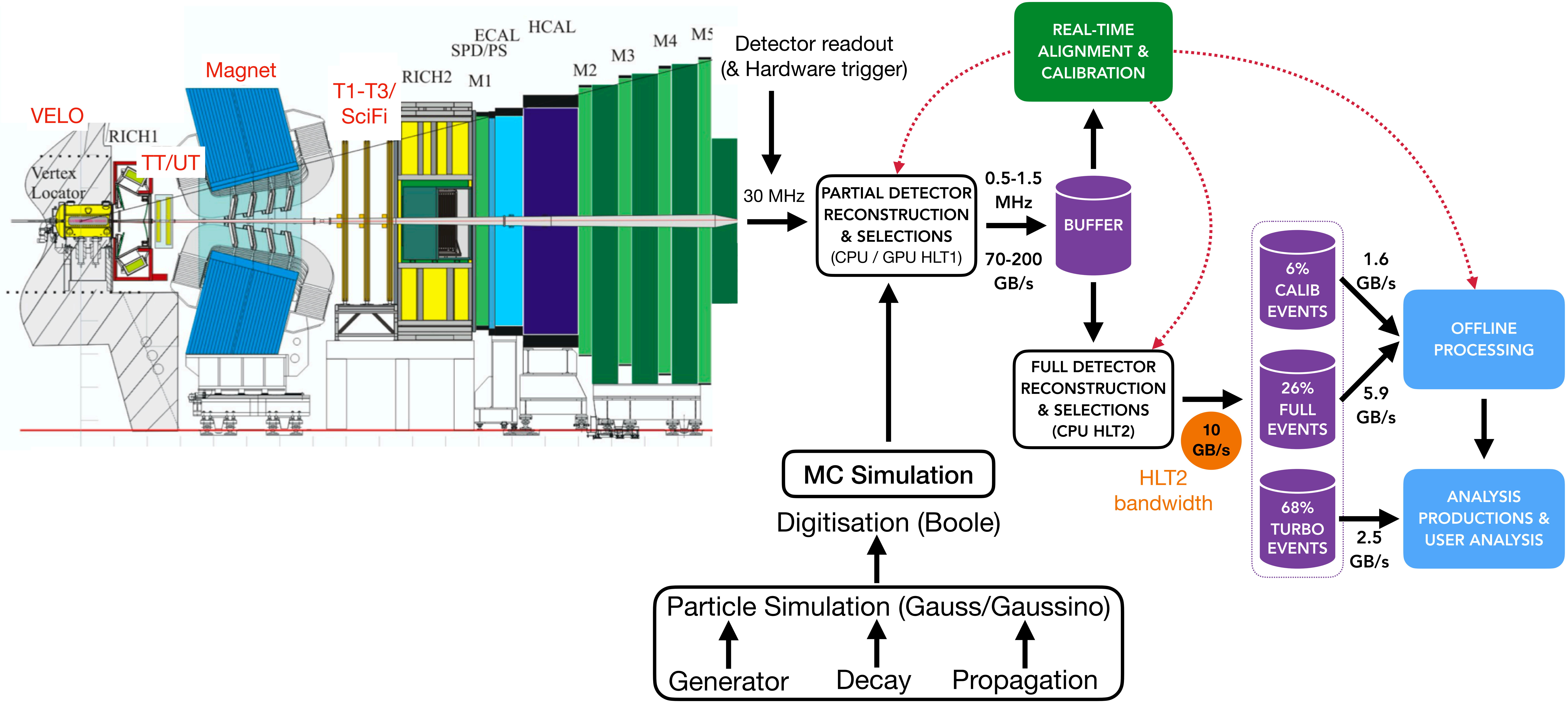
&

Particle Identification

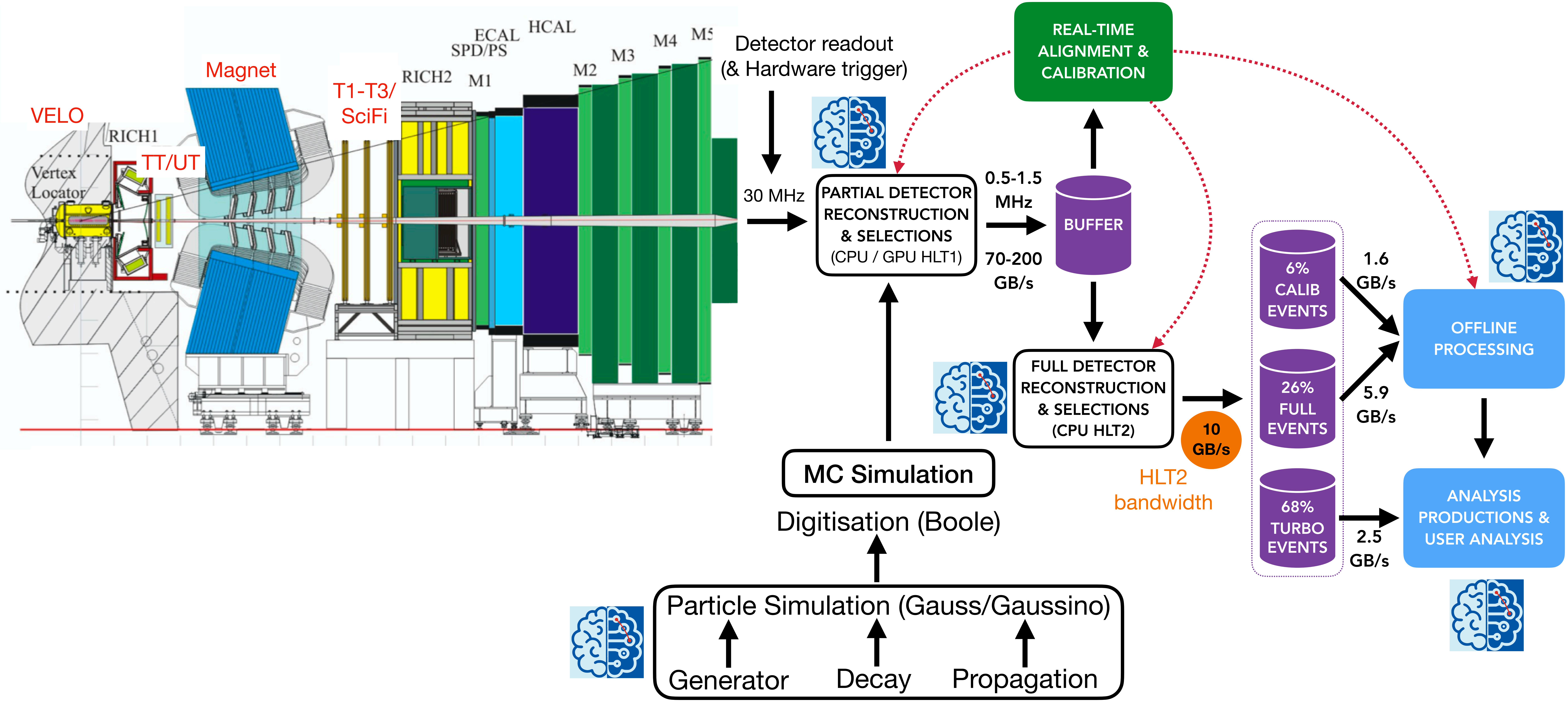
LHCb - data processing chain



LHCb - data processing chain



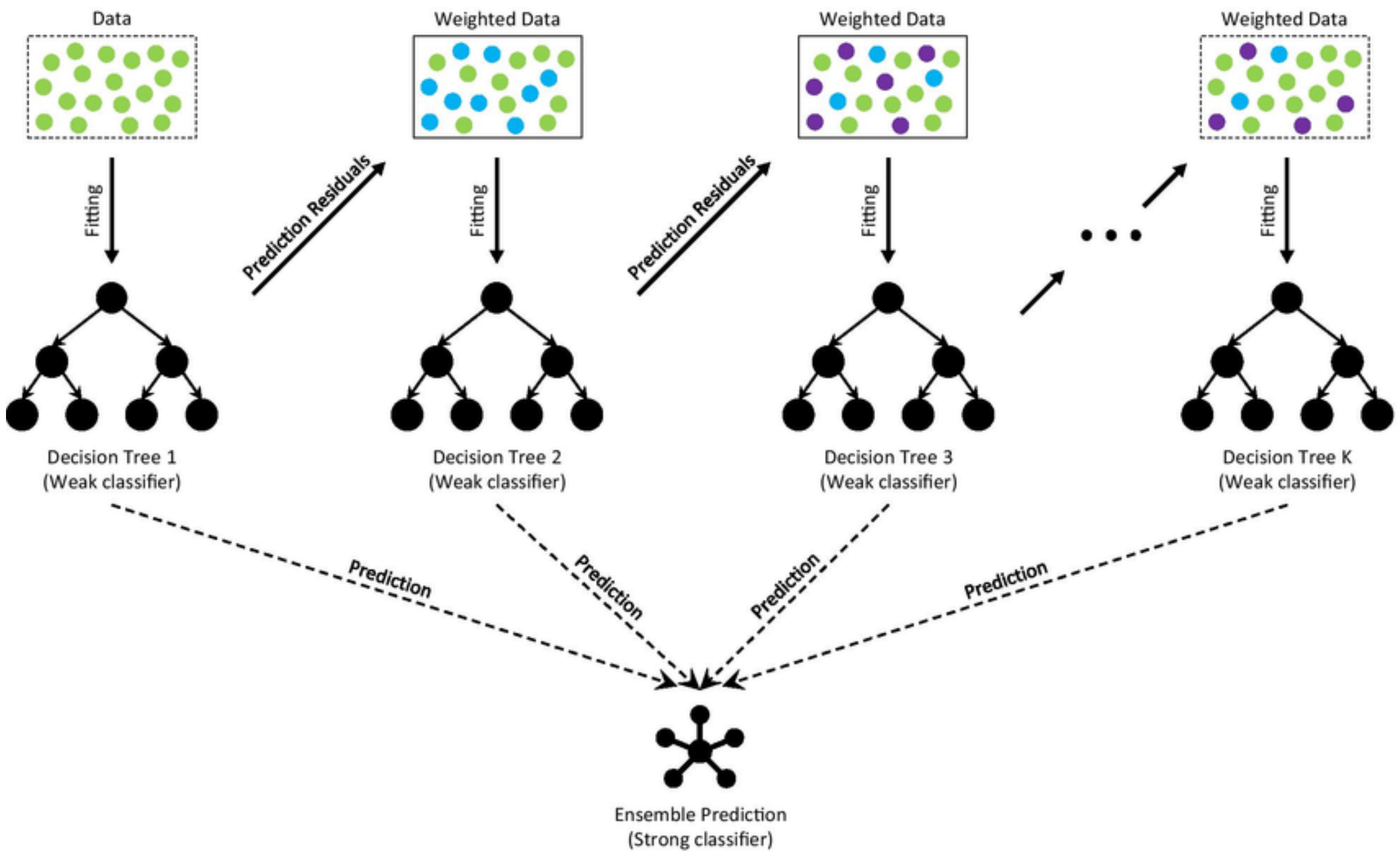
LHCb - data processing chain

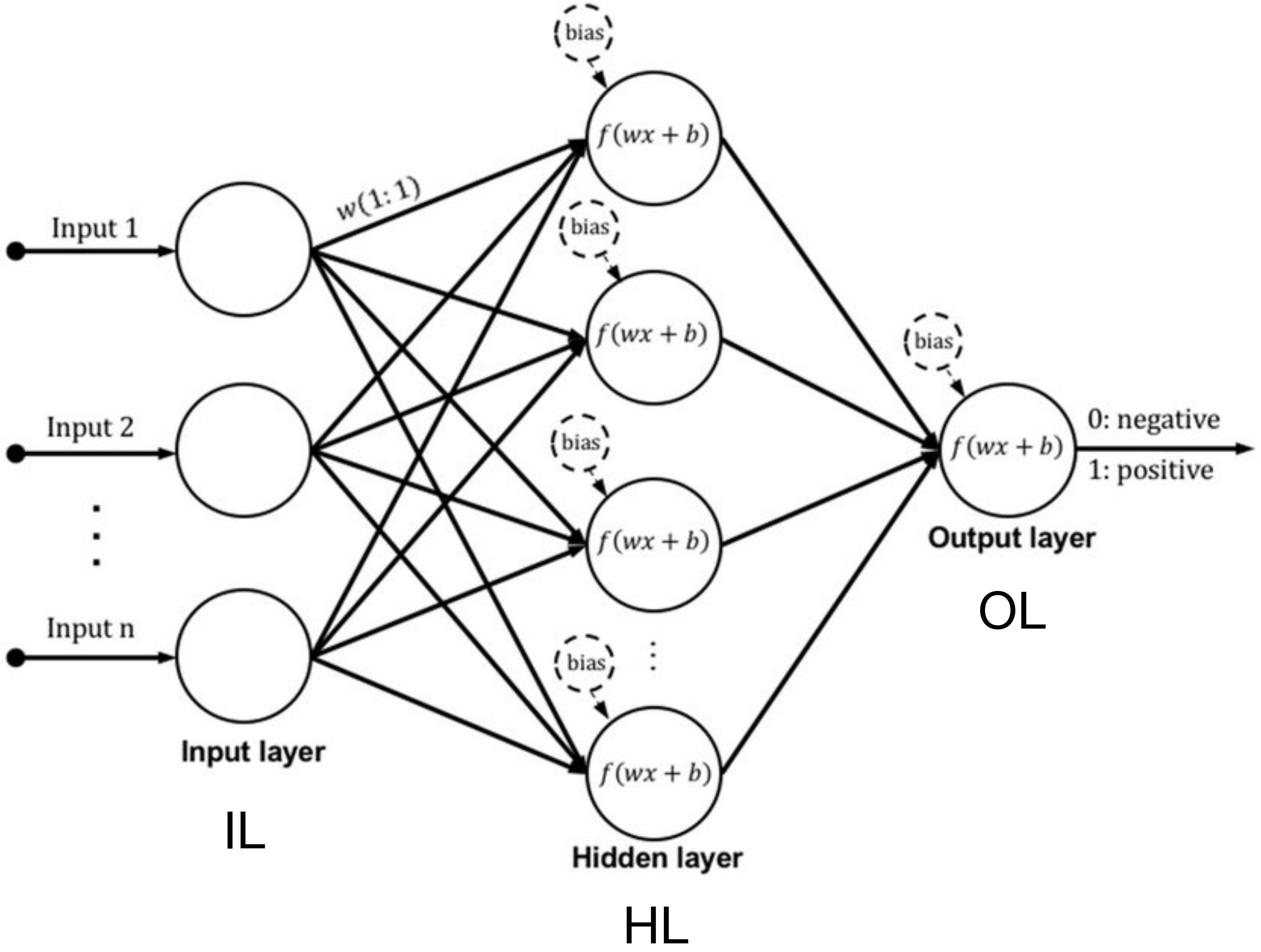


Two types of commonly used classifiers in LHCb

Gradient Boosted Decision Trees (BDT)

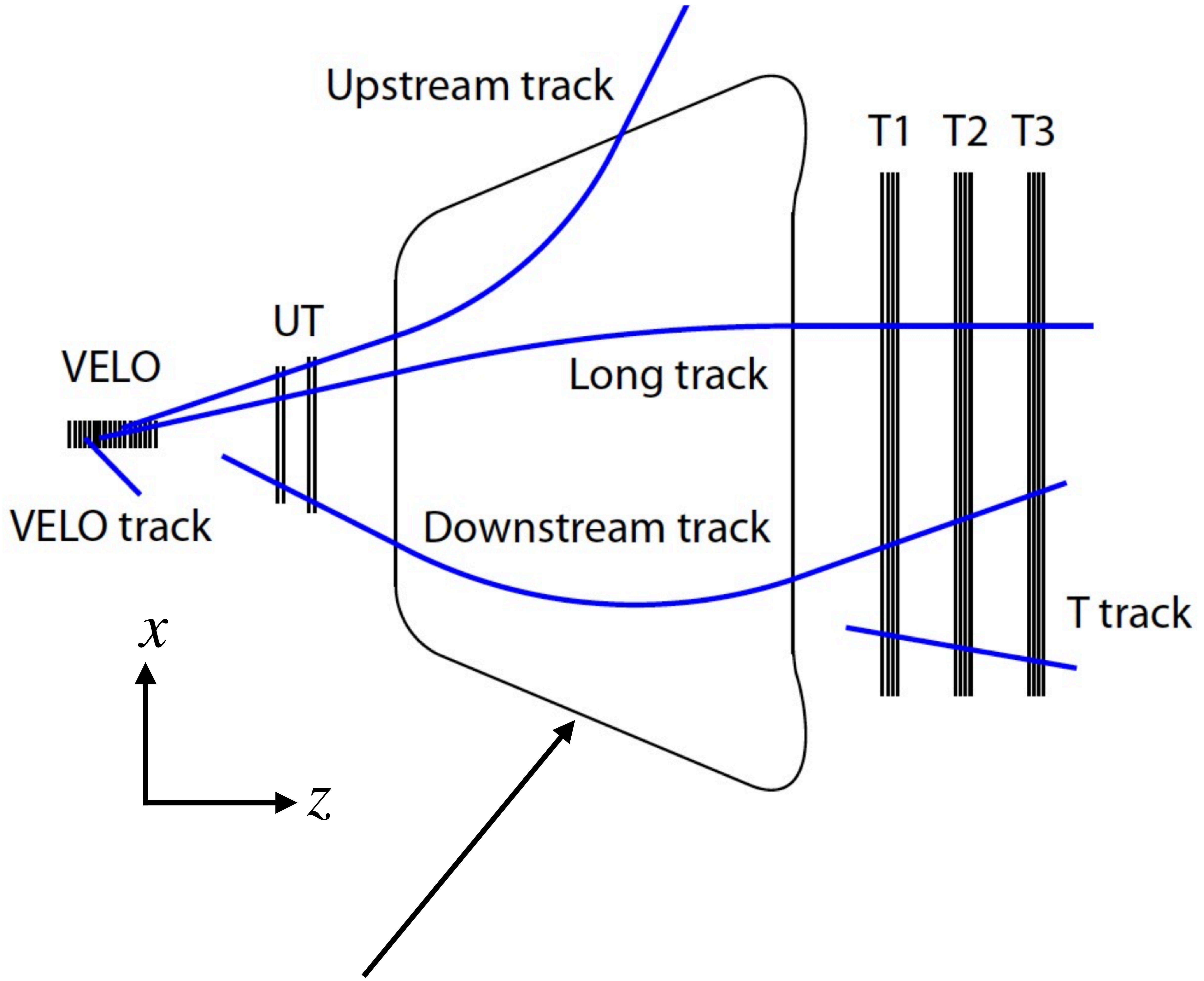
Neural Networks: Multi Layer Perceptron (MLP)





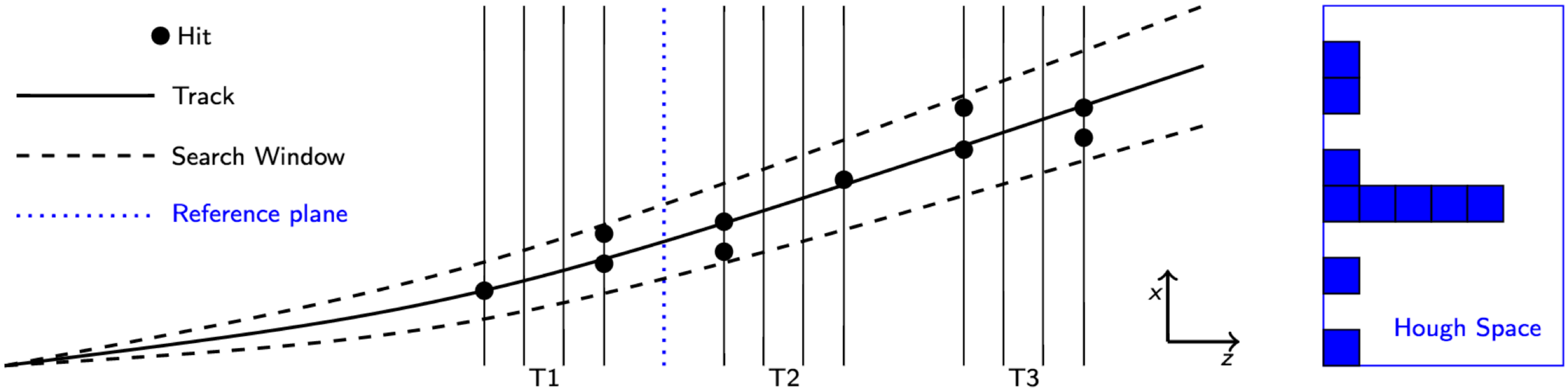
Usually we use supervised learning techniques from these software libraries: xgboost, scikit-learn, scikit-hep, TMVA

- Different track types need different tracking algorithms:
  - Forward tracking
  - Seeding & Matching
- Tracking algorithms consist of three steps:
  - Pattern recognition
  - Track fit $\rightarrow \chi^2_{\text{trk}}$
  - Removal of bad track candidates
- Performance indicators:
  - Tracking efficiency
  - Fake rate



Dipole magnet with $\vec{B} \approx B_y \vec{e}_y \rightarrow$ tracks are bent in xz-plane

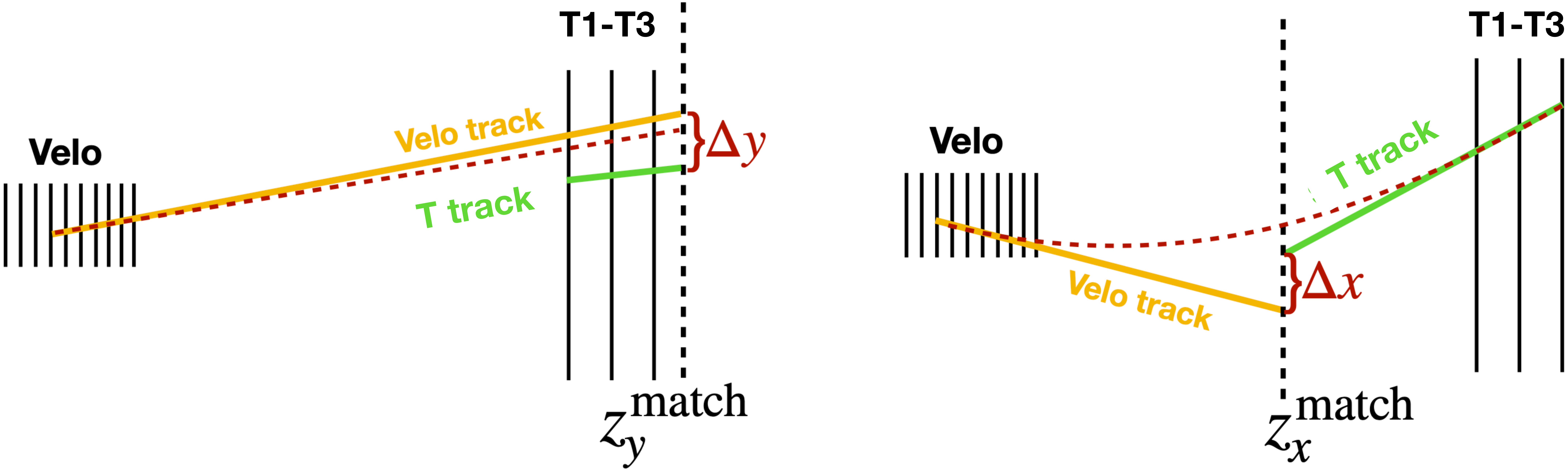
Forward tracking



- Reconstructed VELO tracks **forwarded** to the T stations looking for hits in search windows
- Clustering hits in reference plane, fitting & removing outliers, recovery loop
- Two MLPs:
  → Rejection of clusters with only 4 hits in recovery loop (2 HL, 9 IL nodes, 16,10 HL nodes)
  → Final track candidate selection before track fit is performed (3 HL, 16 IL nodes, 17,9,5 HL nodes)

Track seeding & matching

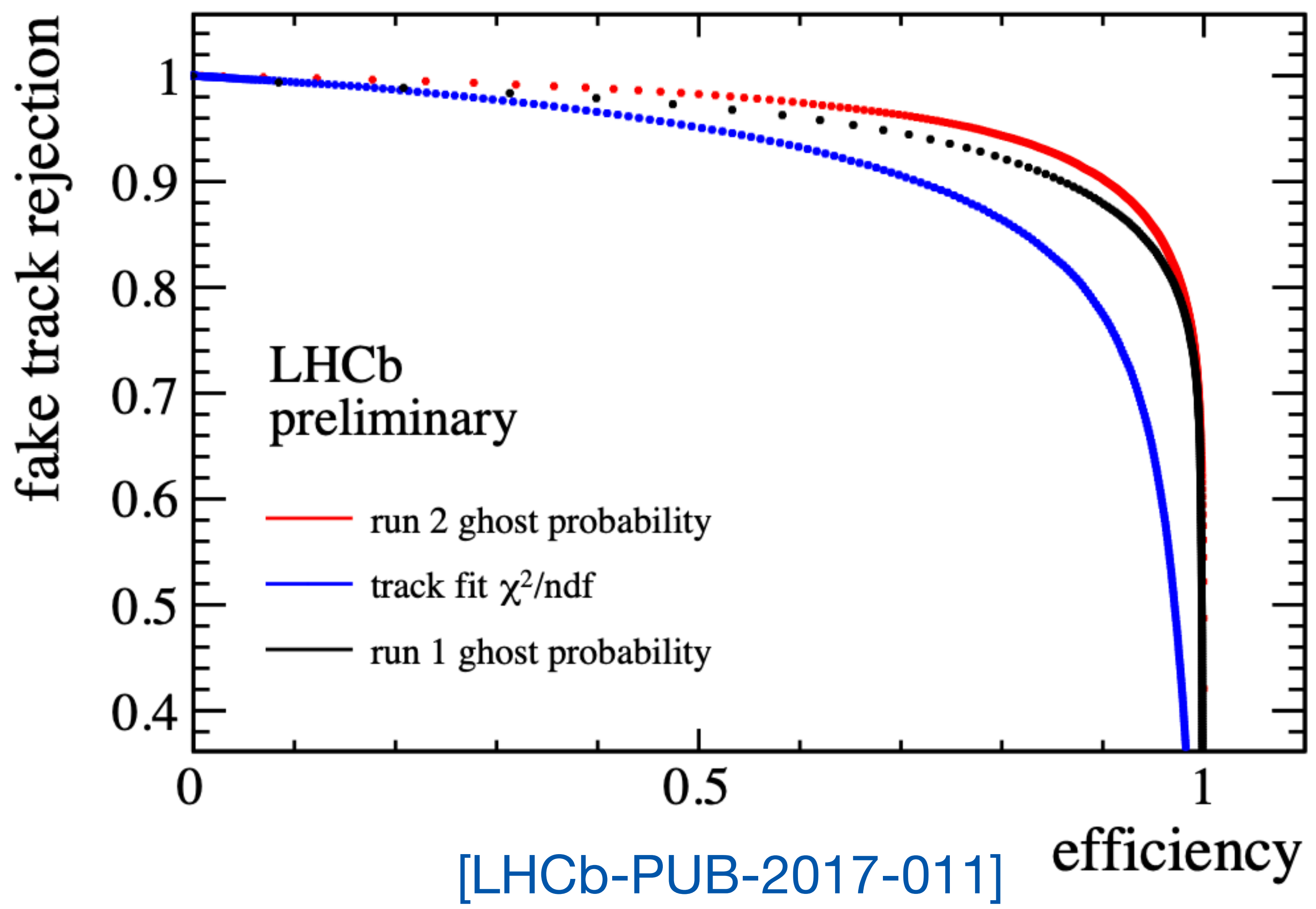
- VELO and T track segments from standalone tracking algorithms
- MLP with two hidden layers trained to match the right pairs
- Training variables: distances, track slopes, slope differences
- >80% rejection of fake pairs, retaining >97% of signal tracks



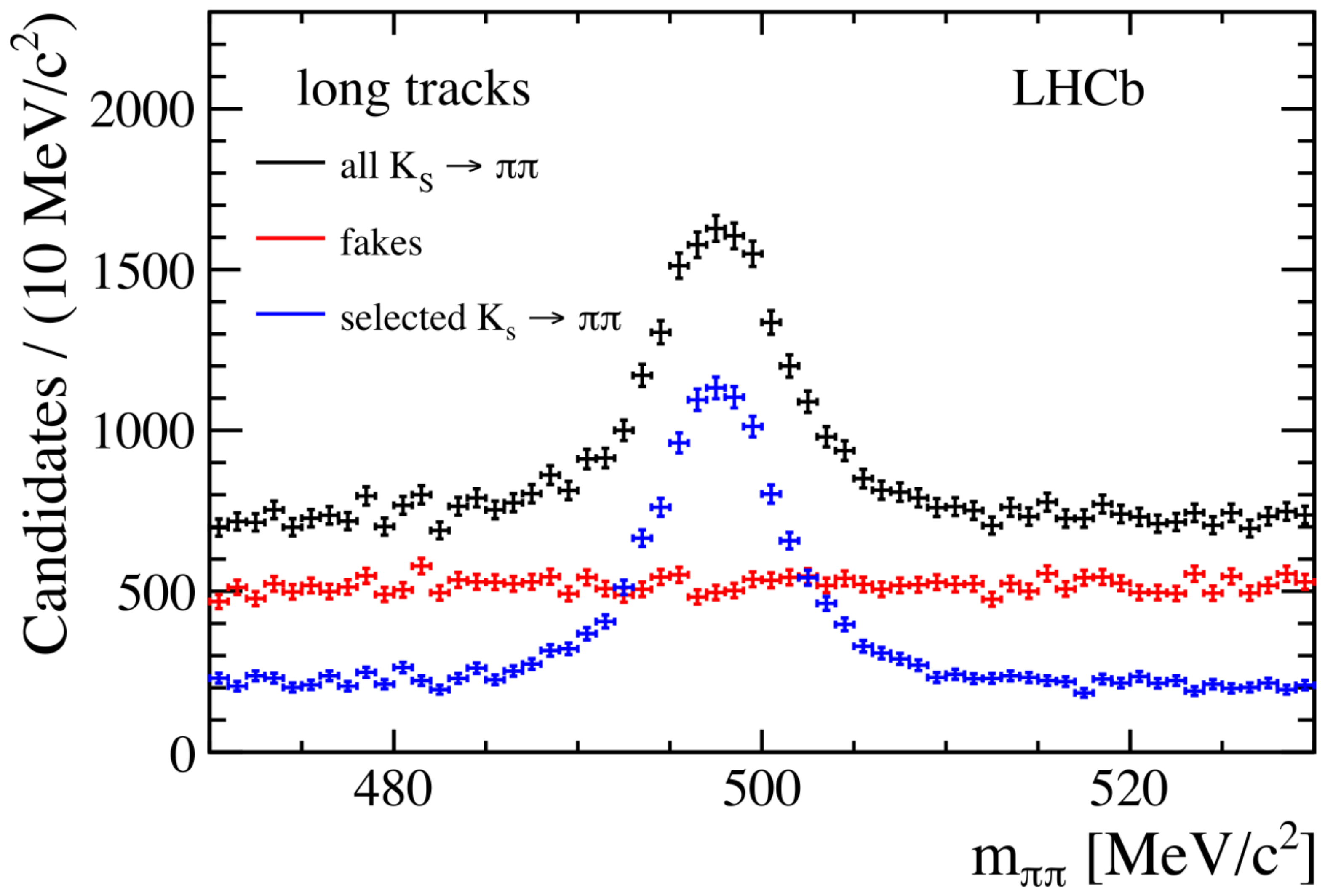
Fake track rejection → Ghost probability

- MLP, the *Ghost probability*, introduced in Run 1 in the offline reconstruction
- Adapted in Run 2 for the online reconstruction in HLT2

  → Fake reduction saves bandwidth
- One hidden layer with 26 nodes
- 21 training variables: track chi2, #Hits on track, p_T, η, detector occupancy, number of tracks competing for shared hits

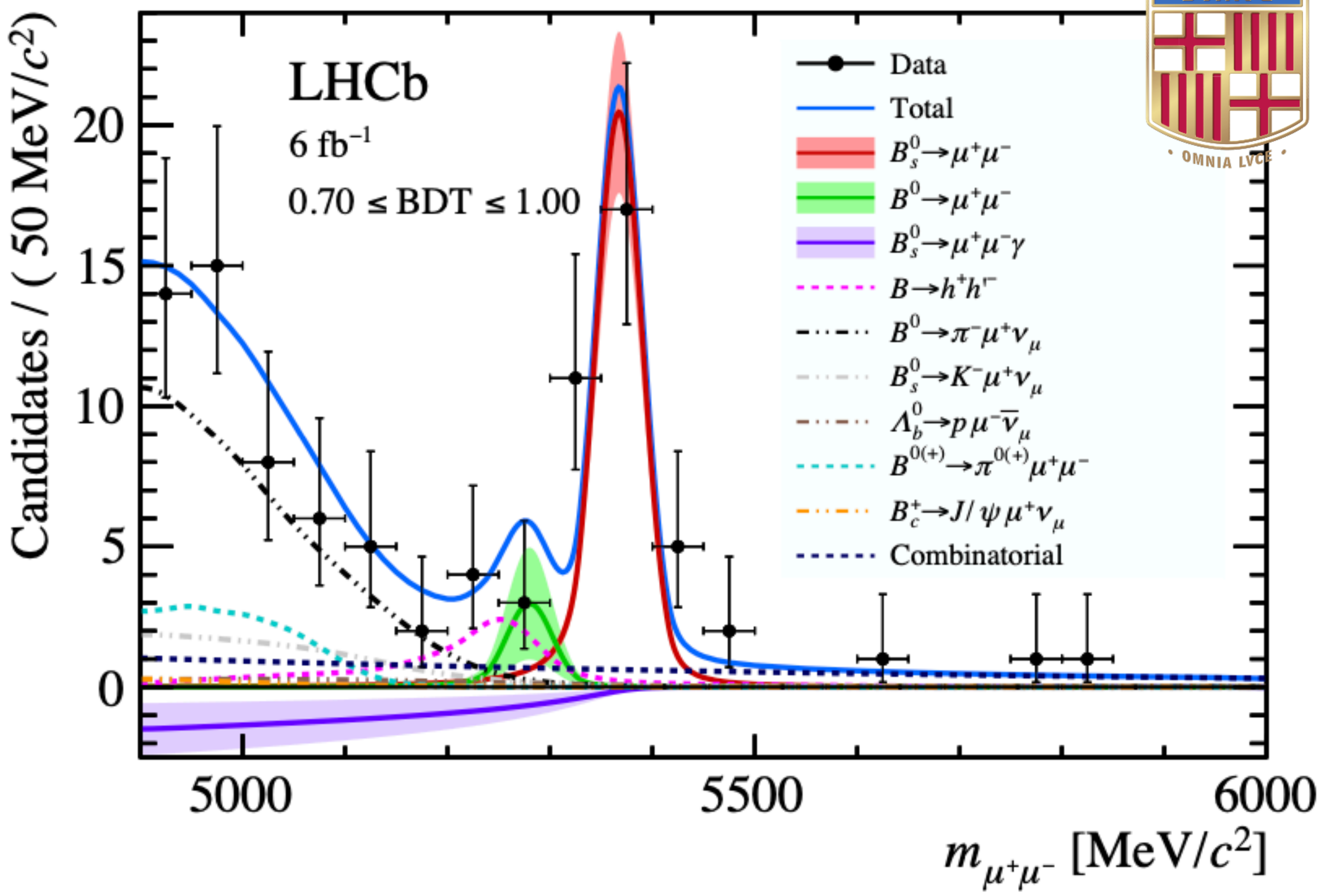


[LHCb-PUB-2017-011]



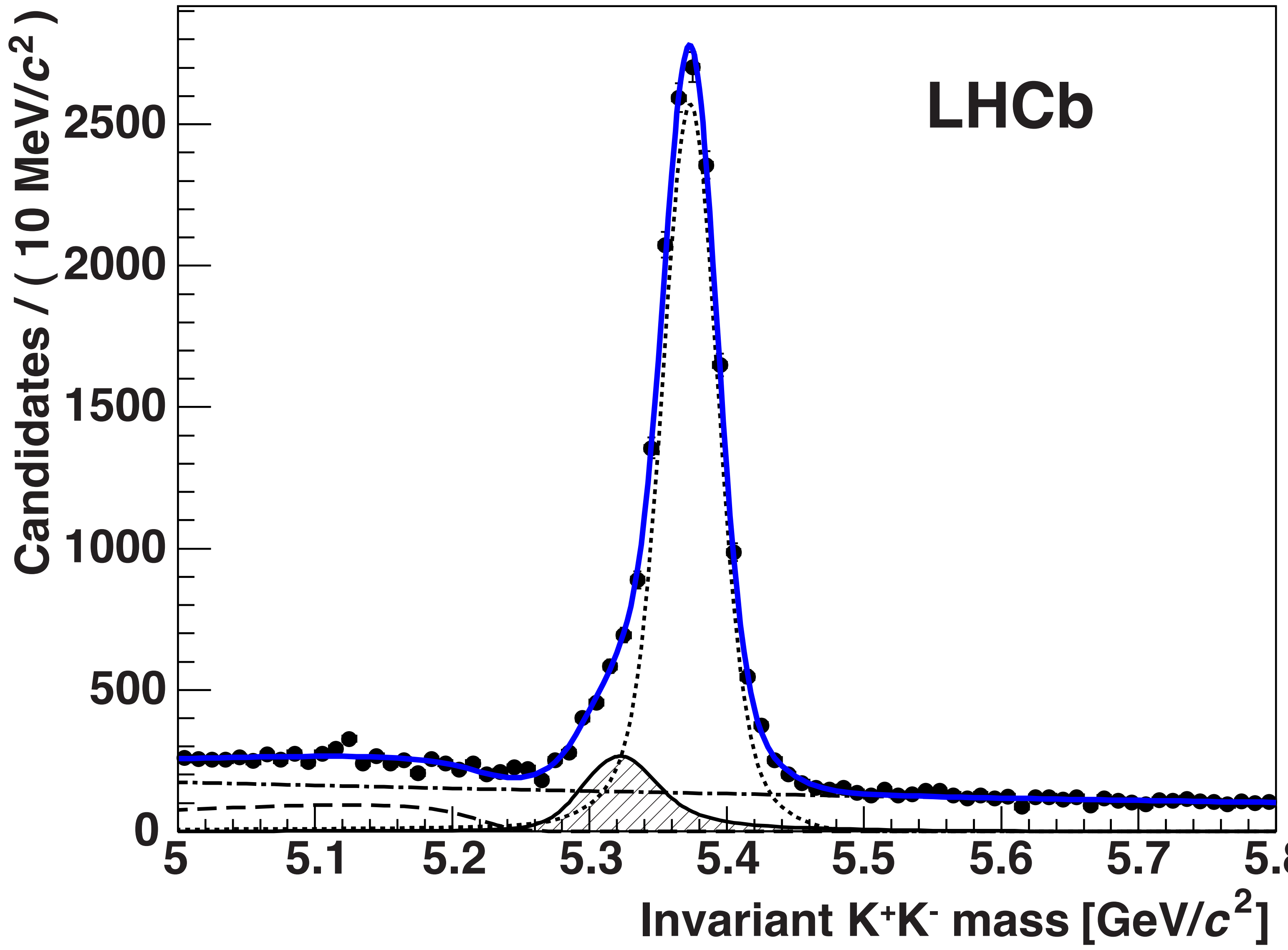
Particle Identification (PID)

$B_s^0 \to \mu^+\mu^-$



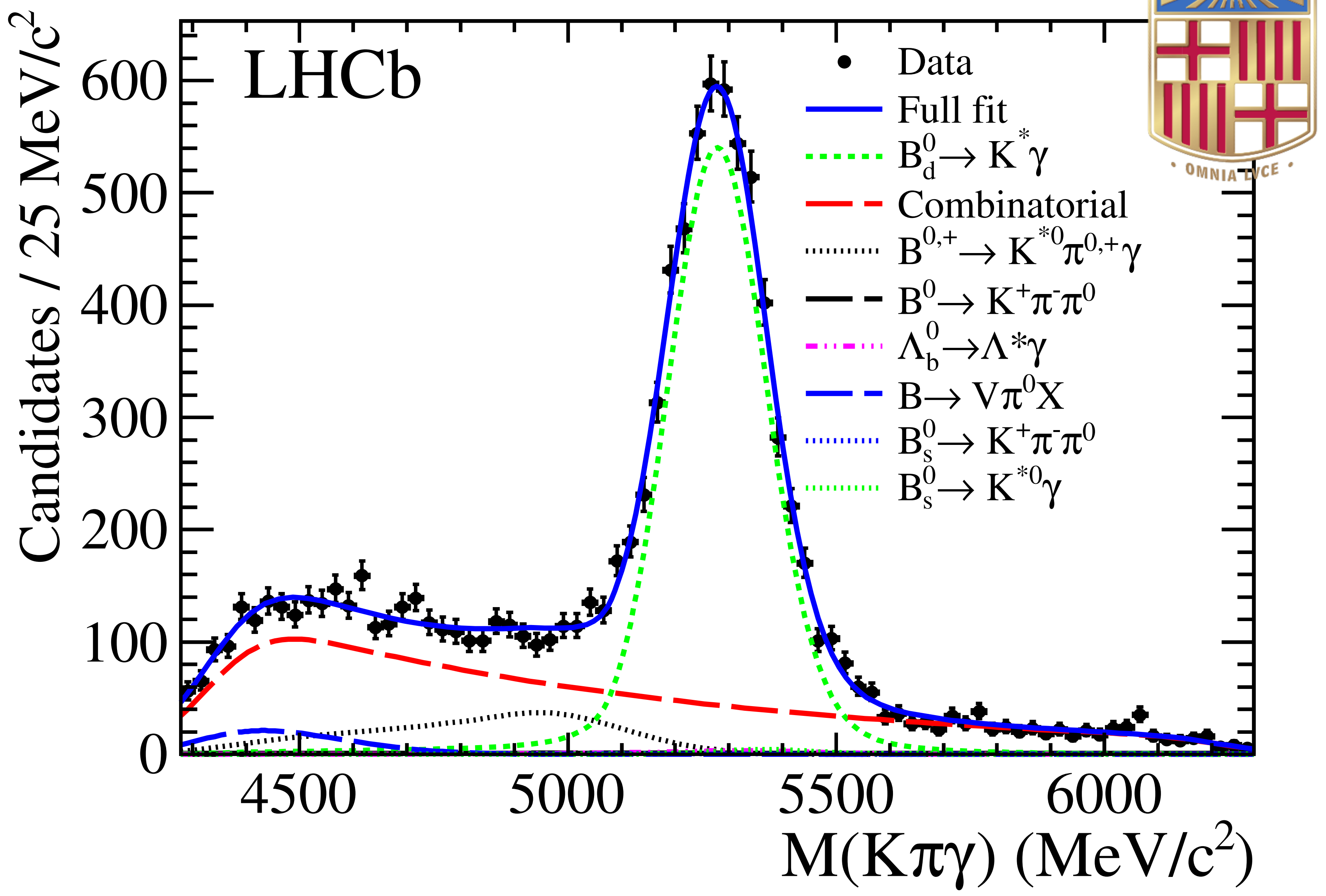
[Phys. Rev. D105 (2022) 012010]

$B_s^0 \to K^+K^-$



[JHEP 10 (2013) 183]

$B^0 \to K^*\gamma$



[Nucl. Phys. B867 (2013) 1]

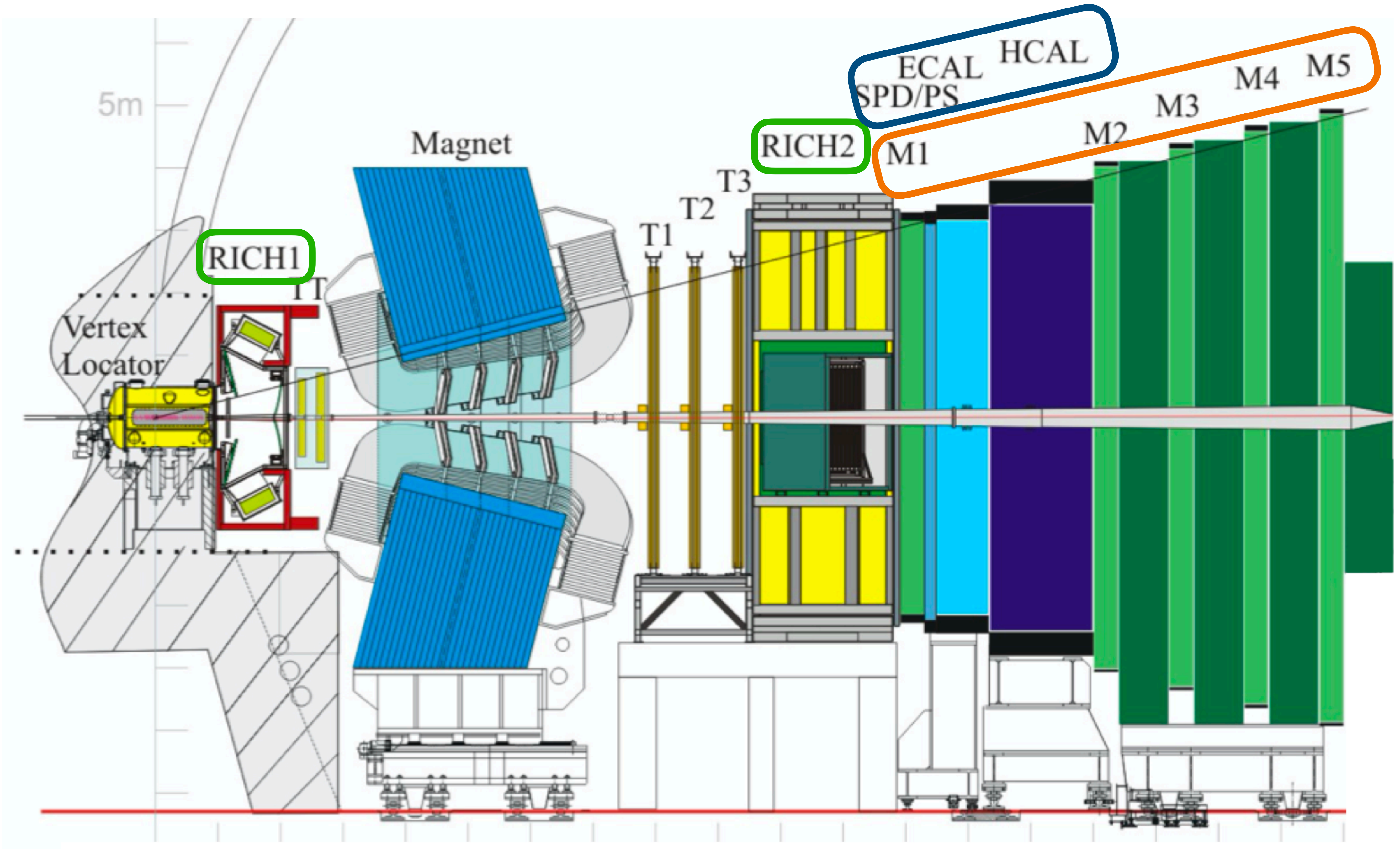
Excellent PID performance essential for the majority of LHCb analyses!

→ Background suppression for rare decay measurements such as $B_s^0 \to \mu^+\mu^-$

→ Classification between hadronic final states with same topology

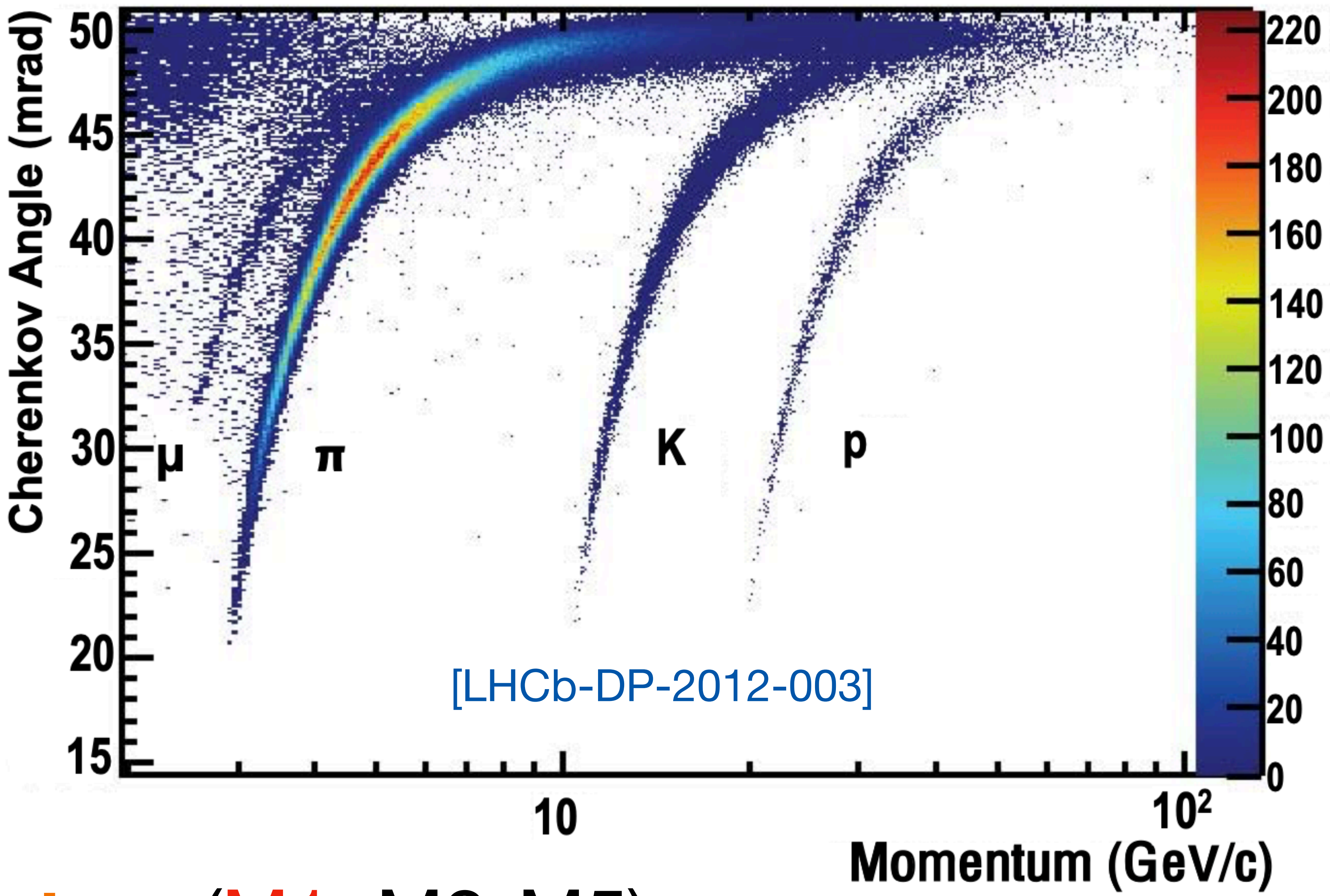
→ Bandwidth-friendly event selections in the software trigger

Particle Identification (PID)



Ring Imaging Cherenkov Detectors (RICH 1 & 2)**:**

- PID for charged hadrons (K, π, p) over large p-range



Calorimeter system (SPD/PS, ECAL, HCAL):

- $e^{\pm}$ vs. γ vs. $h^{\pm}$ discrimination, γ vs. π^0
- Measurement of energy and positions
- E_T used in hardware trigger

Muon system (M1, M2-M5):

- Muon PID for trigger and offline selections
- p_T used in hardware trigger

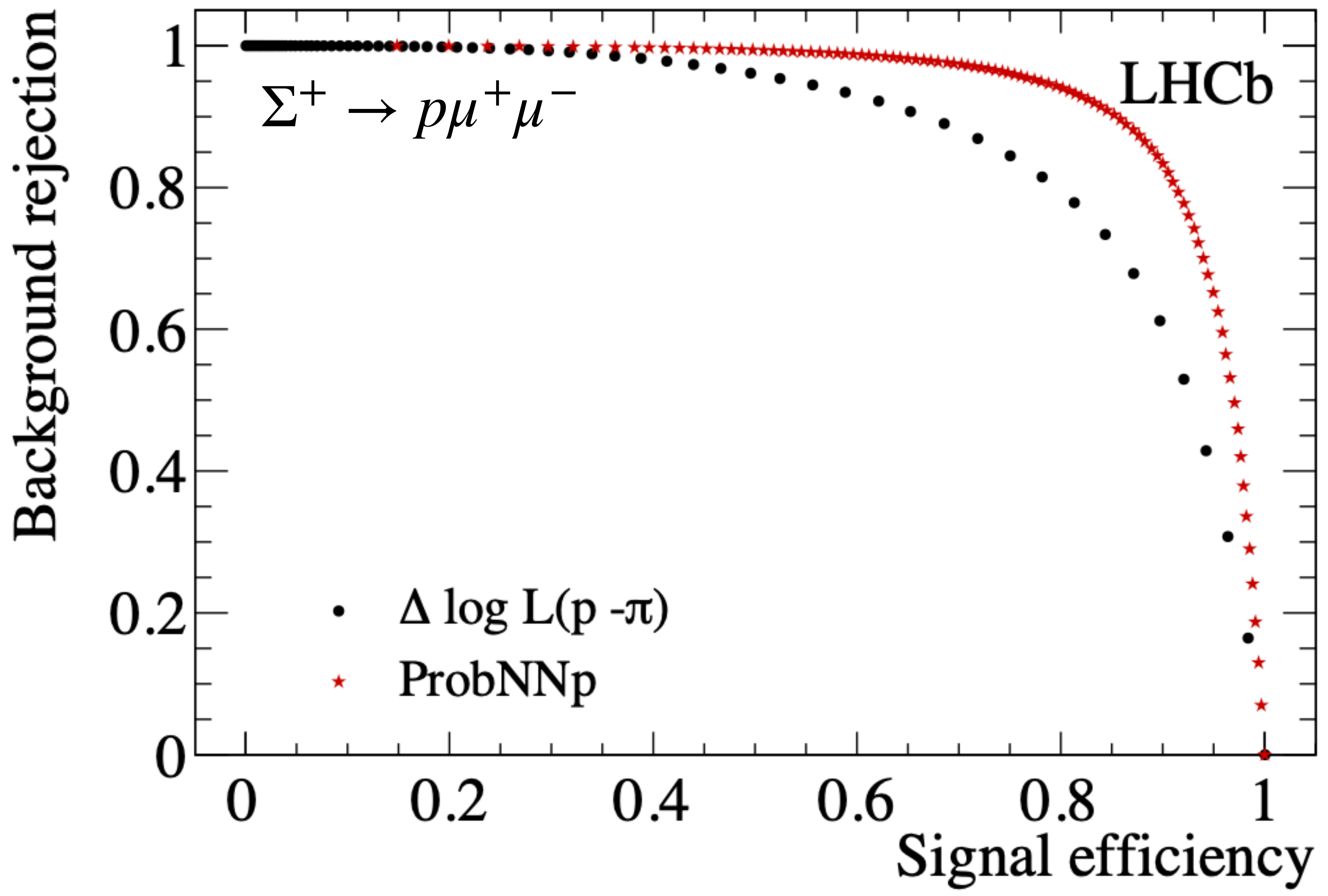
*removed for LHCb Upgrade I

ML for charged PID

https://arxiv.org/pdf/1803.00824.pdf

https://arxiv.org/pdf/1412.6352.pdf

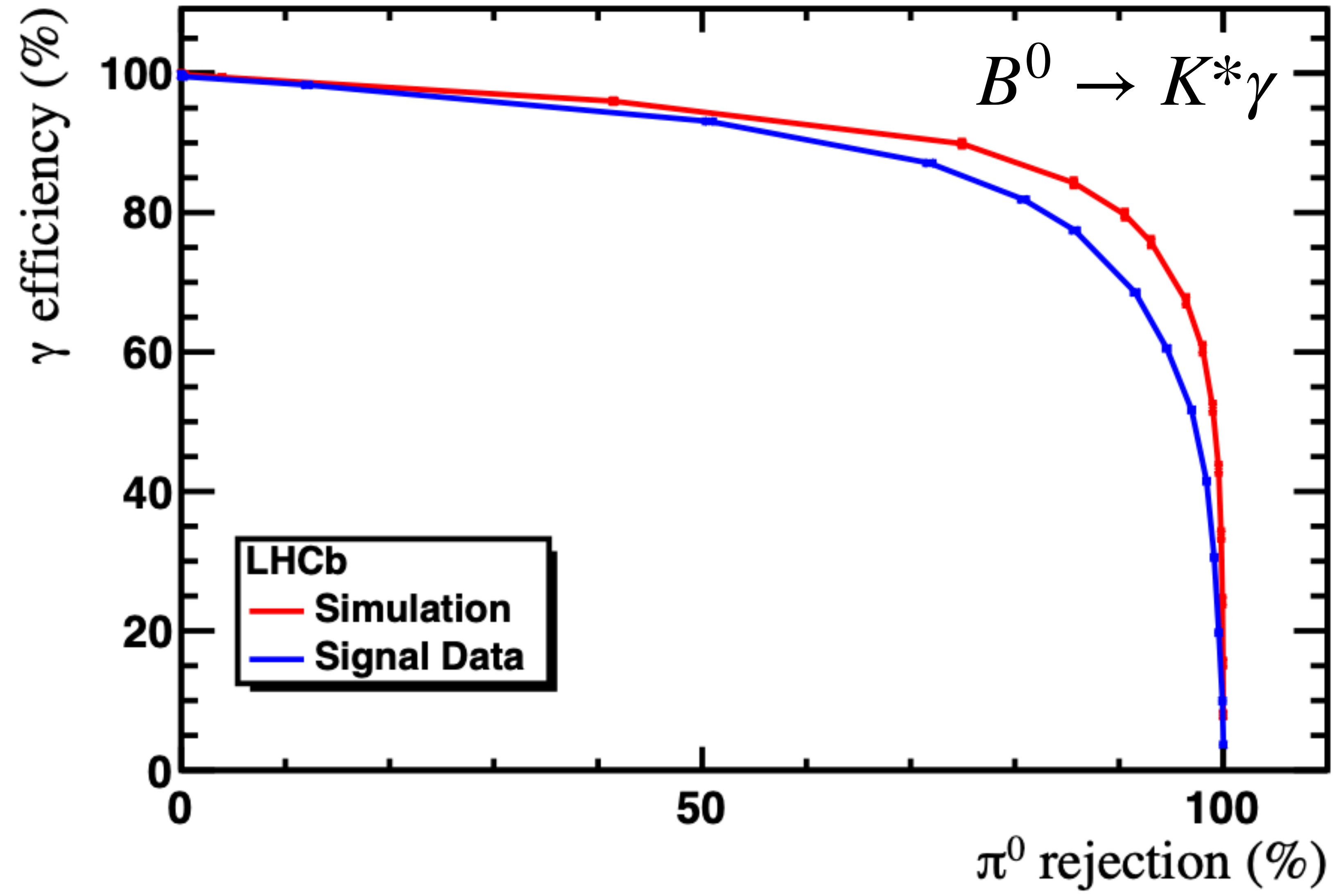
- Two main sets of variables for combined charged PID:
  - Combined Δ log-likelihood:
    - $\mathscr{L}_{X\pi} = \mathscr{L}_{X\pi}^{\text{RICH}} \cdot \mathscr{L}_{X\pi}^{\text{CALO}} \cdot \mathscr{L}_{X\pi}^{\text{MUON}}$
    - $\mathscr{L}_{X\pi}^{\text{RICH}} = \ln(\mathscr{L}_{X}^{\text{RICH}}) - \ln(\mathscr{L}_{\pi}^{\text{RICH}})$
    - $X = \{\pi, K, p, e, \mu\}$
  - Neural Network approach (ProbNNX):
    - Separate three-layer MLP for each particle
    - Different tunings for general purpose and dedicate analyses are provided
    - Training variables from tracking + PID detectors
    - Training data: simulated c and b hadron decays

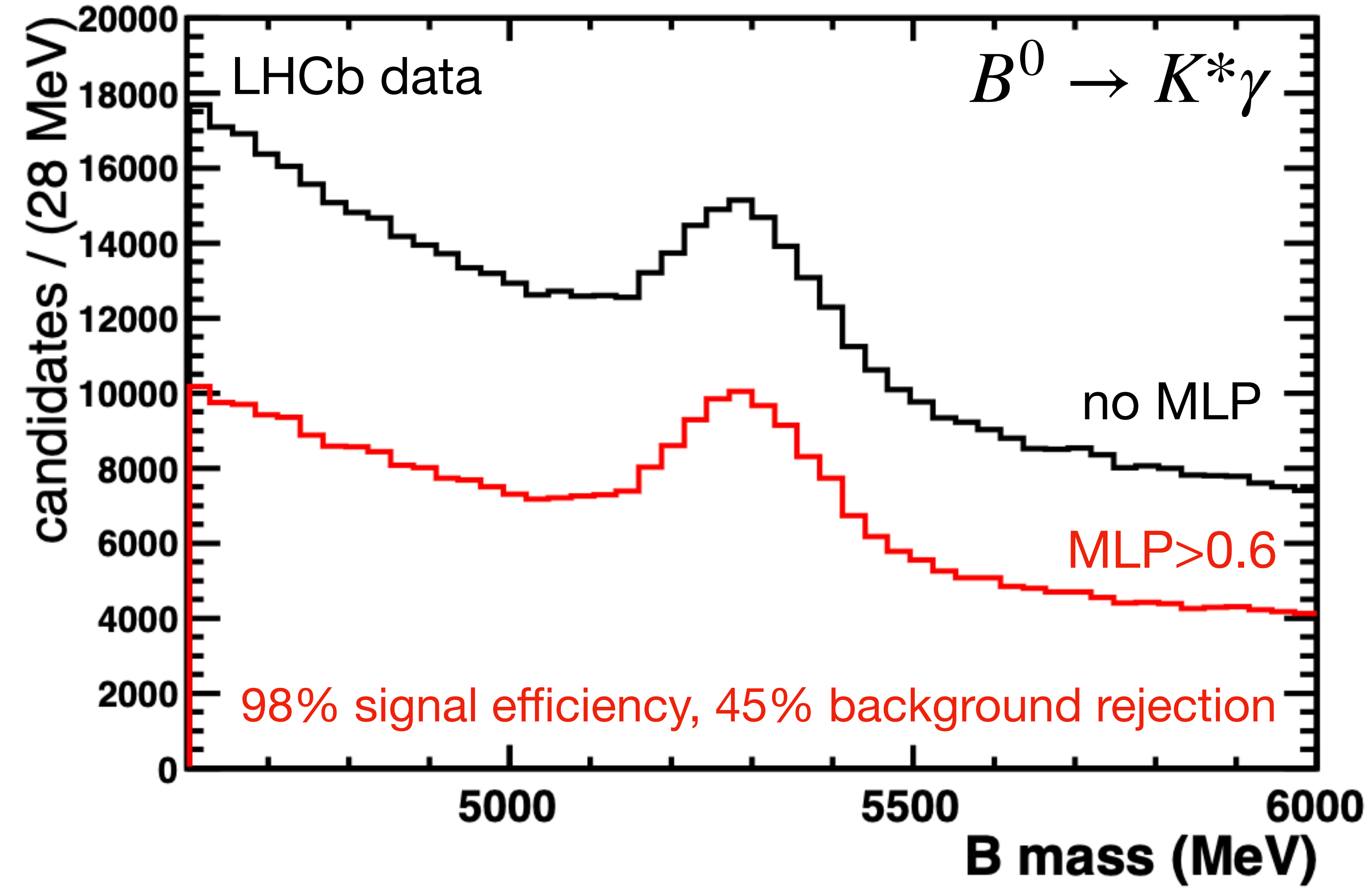


ML for neutral PID - γ/π^0 separation

[LHCb-PUB-2015-016]
[CERN-LHCb-DP-2020-001]

- Unconverted γ vs. merged $\pi^0 \to \gamma\gamma$
- Very important for radiative decays and hadronic decays including a π^0 in the final state
- Training variables based on calorimeter information
  - Shower shapes, energy deposits, hit multiplicities
- Separate four-layer MLPs for the three ECAL granularity regions
- Simulated $B^0 \to K^{*0}(\to K^+\pi^-)\gamma$ as signal, $B^0 \to K^+\pi^-\pi^0$ as background proxies

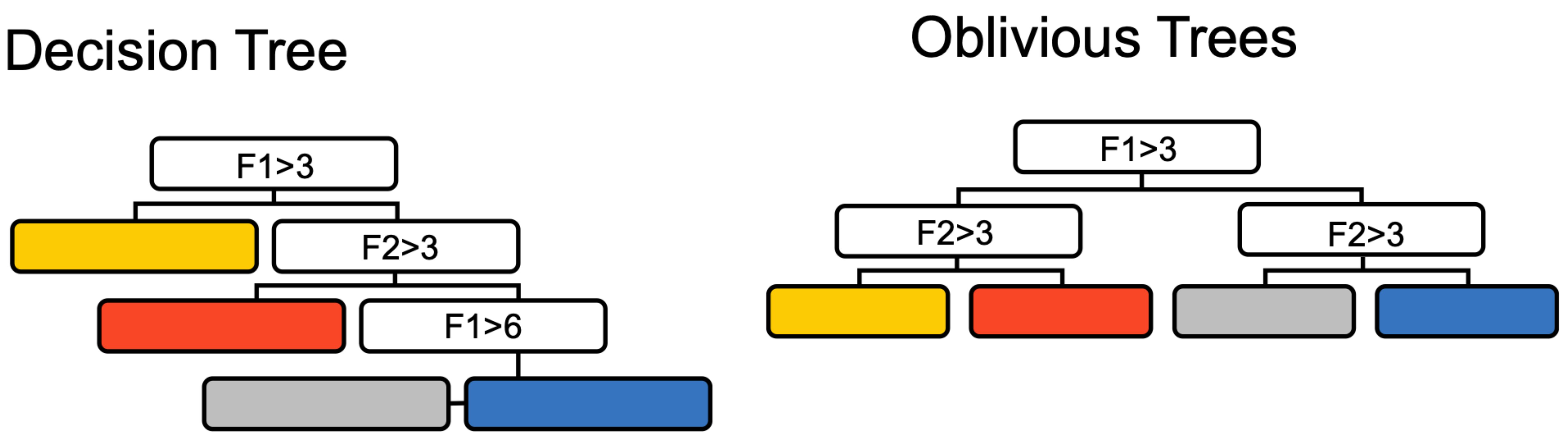




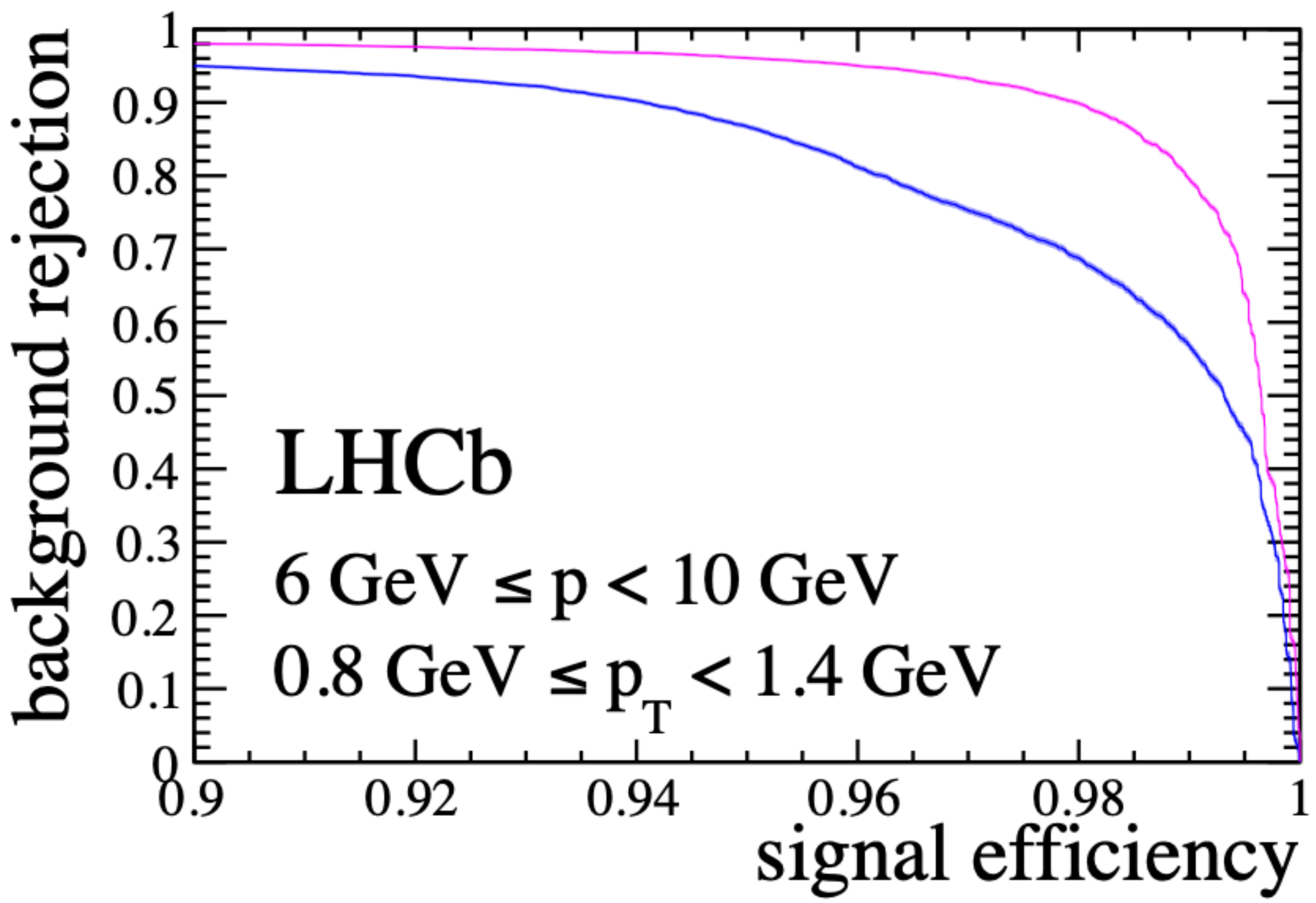
MuonID in Run 3

[CERN-LHCb-DP-2020-002]

- In HLT1: χ^2_{corr} (+isMuon)
  - spatial residuals with respect to track extrapolation
  - multiple scattering
  - correlation between hits from muon stations



- In HLT2: More timing budget
  → More complex ML applications for muonID feasible

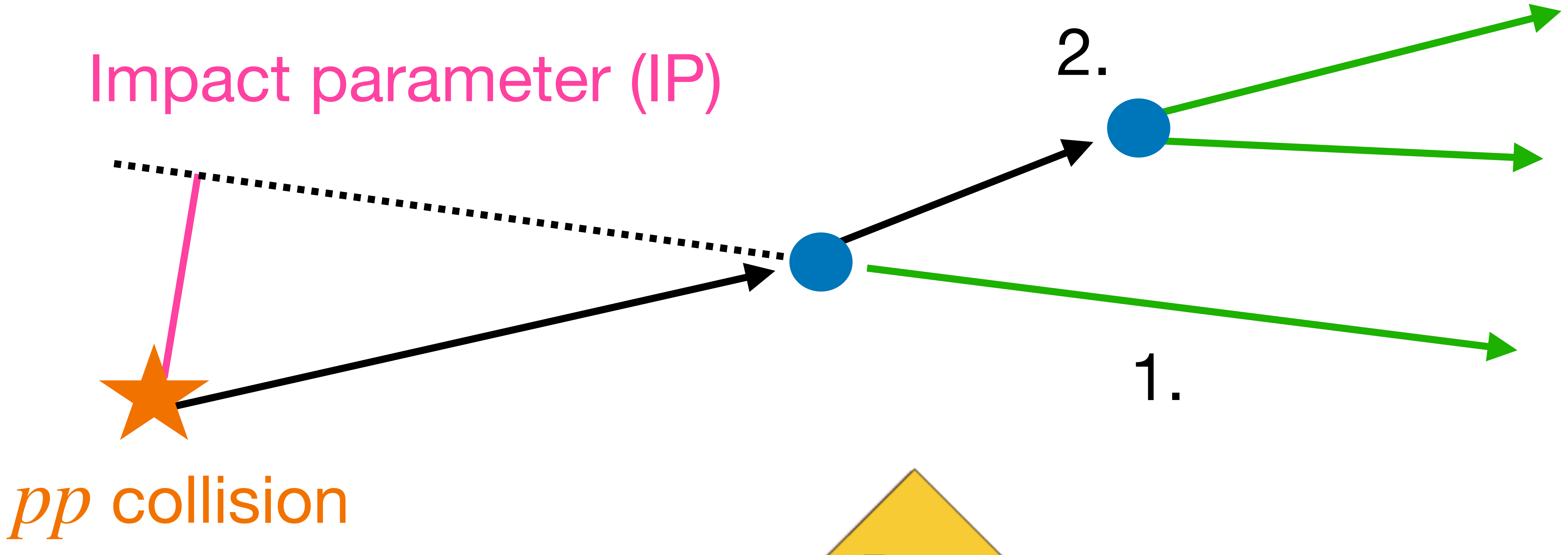


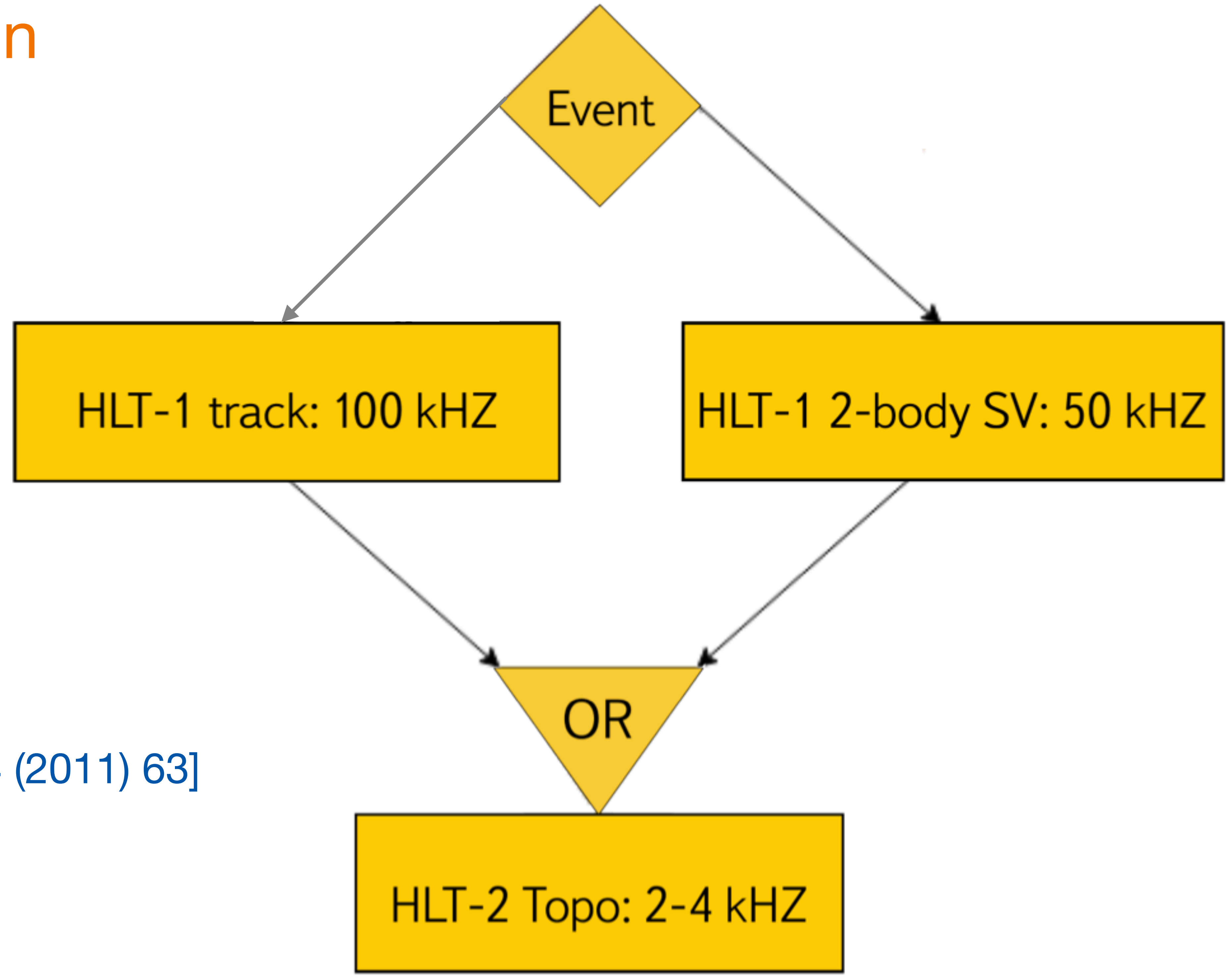
- CatBoost algorithm
  → Gradient Boosted **Oblivious Trees**
  - 23 training variables: spatial and temporal hit info + correlations

- Oblivious Trees: less expressive, but faster evaluation

Topological triggers in HLT1 and HLT2

- **Inclusive triggers** based on topology of partially reconstructed b-hadron decays
- HLT1 inclusive trigger selections:
  - 1. Single displaced high-p_{T} track
  - 2. Displaced vertex with high p_{T}
- HLT2 inclusive trigger selections:
  - multi-track displaced vertices with high p_{T}
  - Takes 40% of the total HLT2 bandwidth
- ML techniques:
  - Run 1: Bonsai Boosted Decision Trees
  - Run 2: MatrixNet by Yandex [JMLR: Workshop and Conference Proceedings 14 (2011) 63]
- Data selected by the top trigger has been used for **several hundred LHCb publications**





Topological trigger in Run 3

[arXiv:2112.00038]
[arXiv:2306.09873]

- Run 3: **Monotonic Lipschitz Neural Networks**
- Robustness against detector effects

  → constraining the gradient of the response function f by a **Lipschitz constant** λ

$$|f(x) - f(x')| \leq \lambda |x - x'|$$

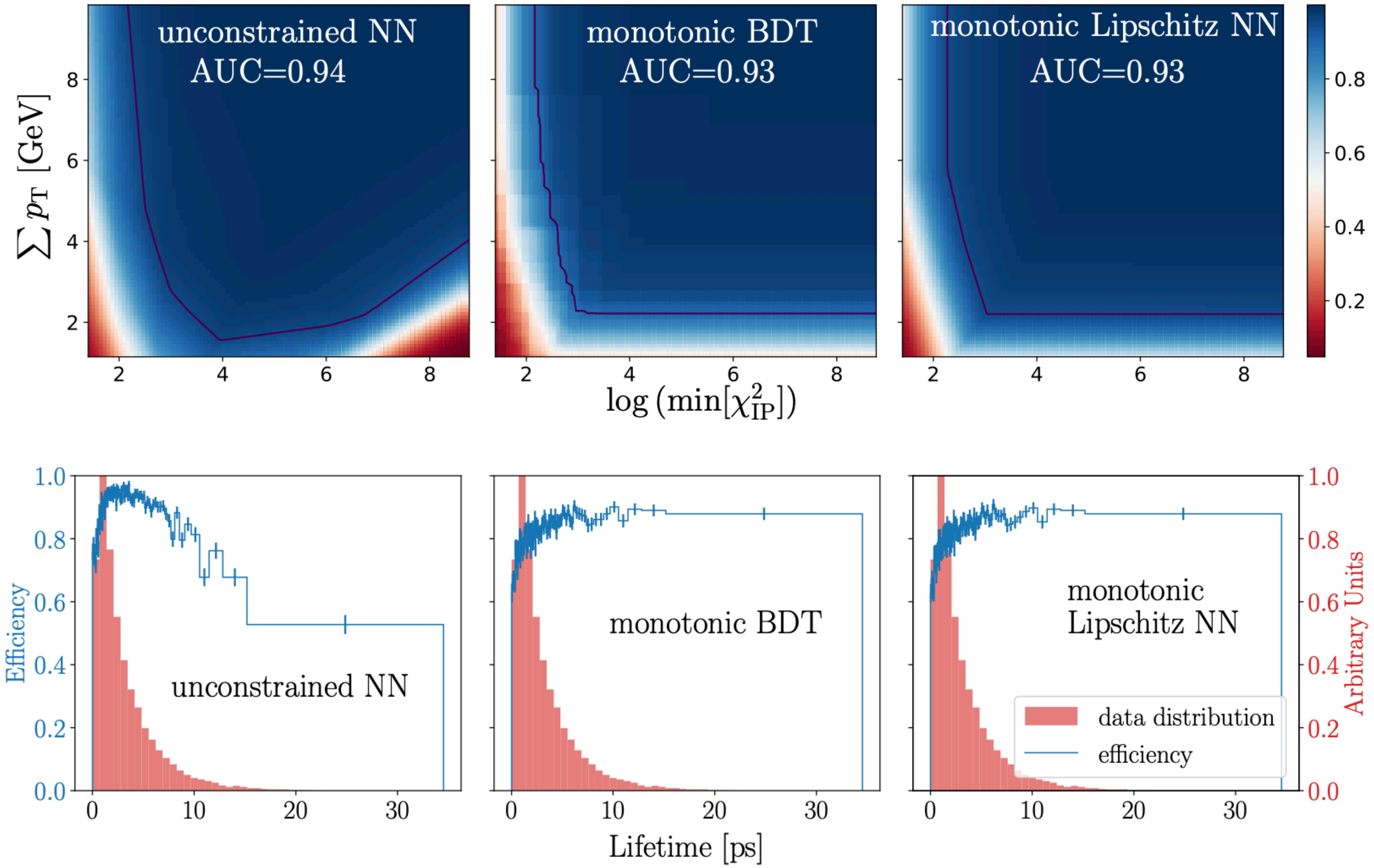
- **Monotonicity** of efficiency in variables of interest ($i \in I$)

  → Adding linear term to response function for each variable of interest

$$g(x) = f(x) + \lambda \sum_{i \in I} x_i$$

$$\frac{\partial g}{\partial x_i} = \frac{\partial g}{\partial x_i} + \lambda \geq 0$$

Simplified model of the topological trigger with two inputs



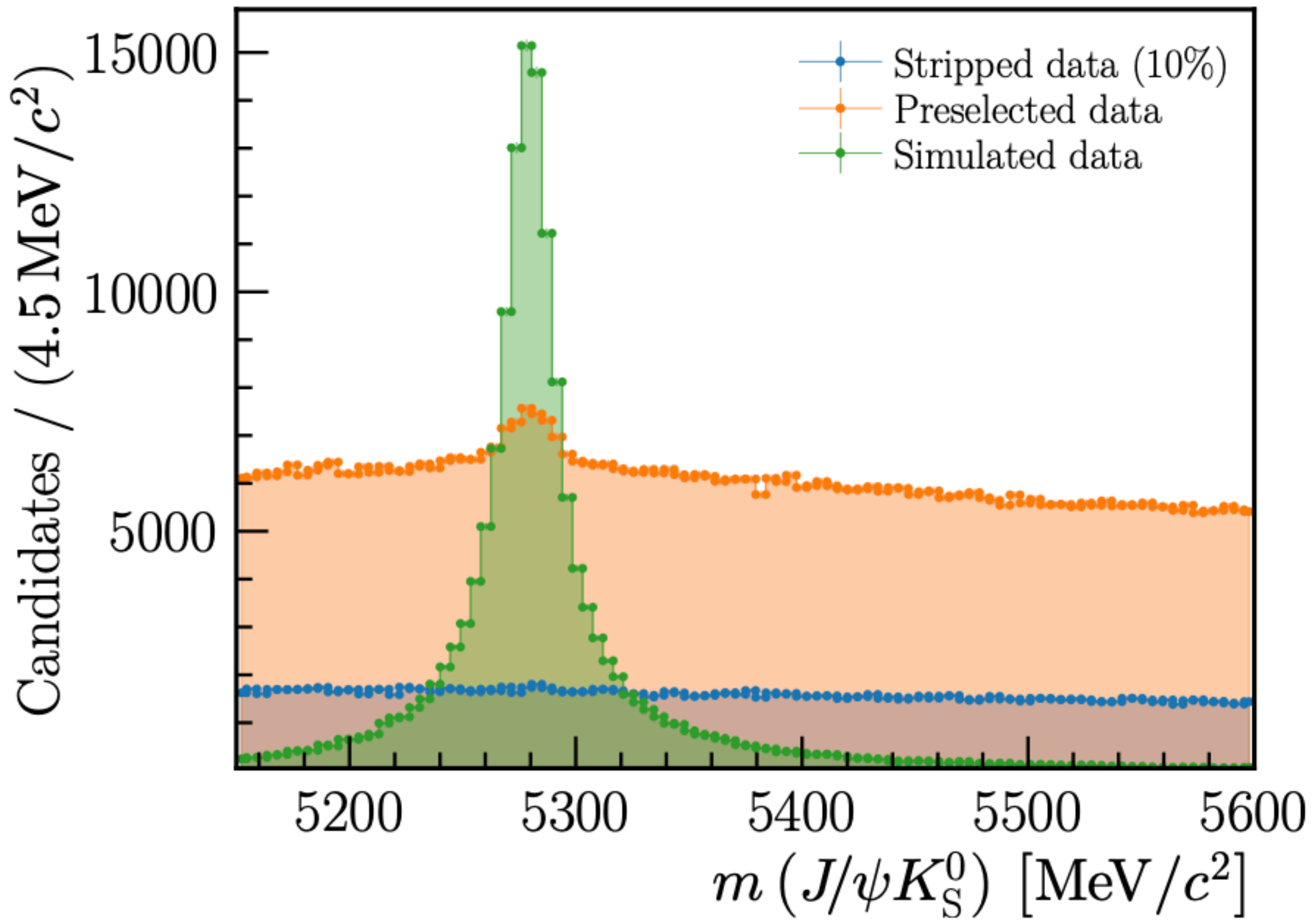
Offline data analysis

Two main ML applications relevant for the majority of LHCb data analyses among many others

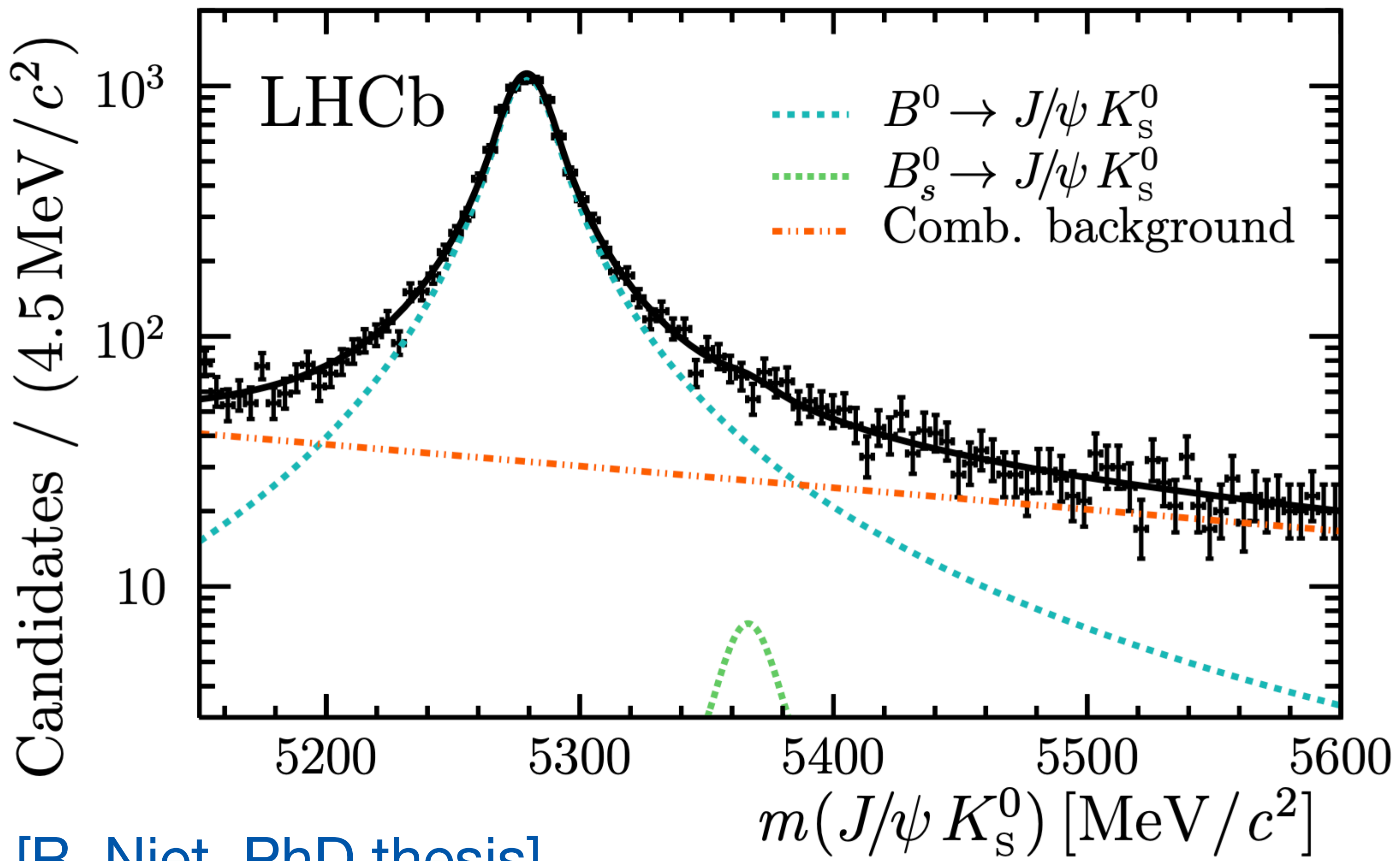
Background rejection → Reweighting of simulation

- Essential for significant mass peaks / reduce systematic uncertainties
- Training data:
  - Data from mass sidebands as background
  - simulated signal events as signal
- Most commonly used model: BDTs, e.g. from [xgboost]
- Example: $\sin(2\beta)$ from $B^0 \to J/\psi(\to e^+e^-)K^0_S$

Before BDT cut



After BDT cut



[R. Niet, PhD thesis]

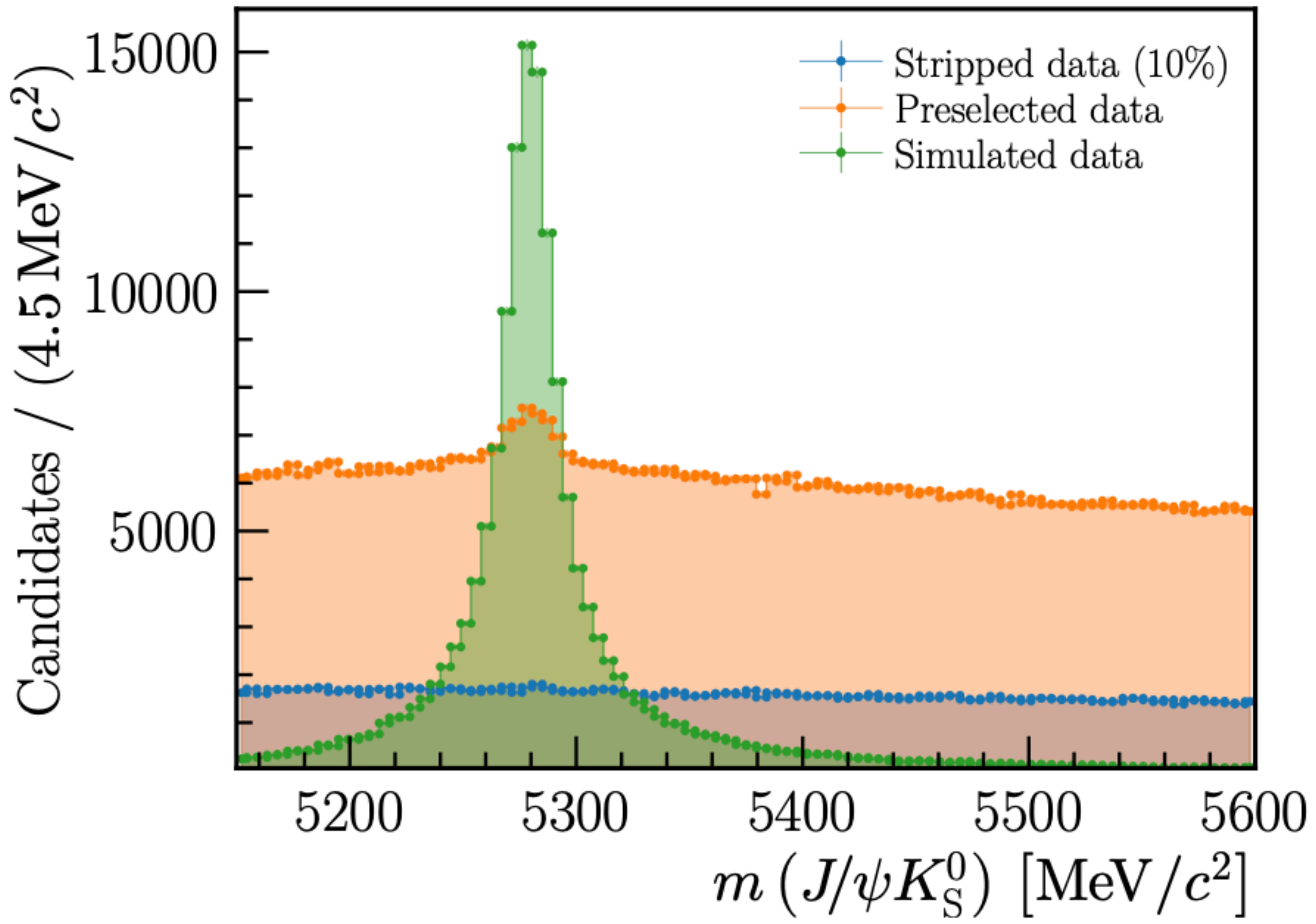
Offline data analysis

Two main ML applications relevant for the majority of LHCb data analyses among many others

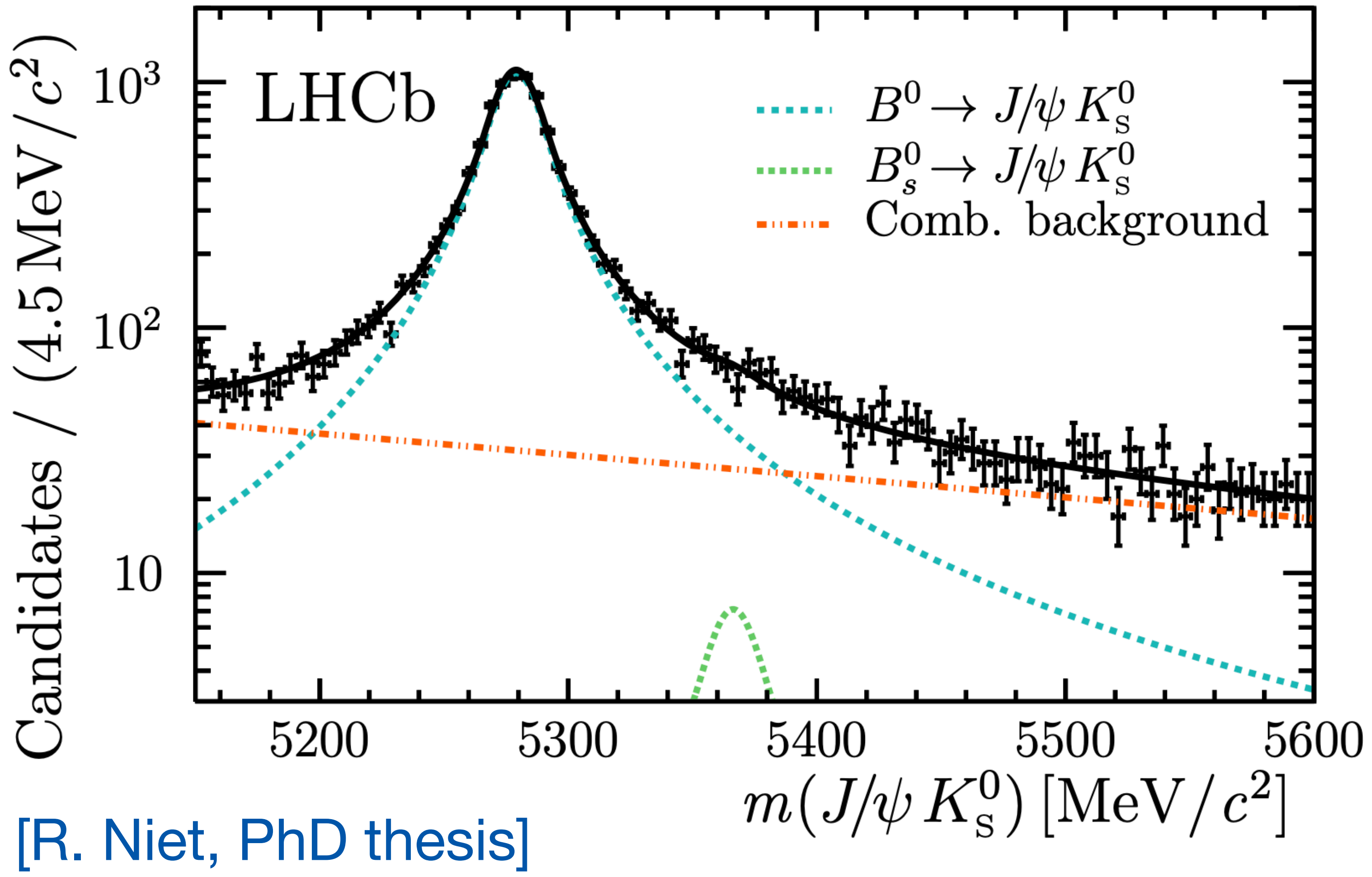
Background rejection

- Essential for significant mass peaks / reduce systematic uncertainties
- Training data:
  - Data from mass sidebands as background
  - simulated signal events as signal
- Most commonly used model: BDTs, e.g. from xgboost
- Example: $\sin(2\beta)$ from $B^0 \to J/\psi(\to e^+e^-)K^0_S$

Before BDT cut



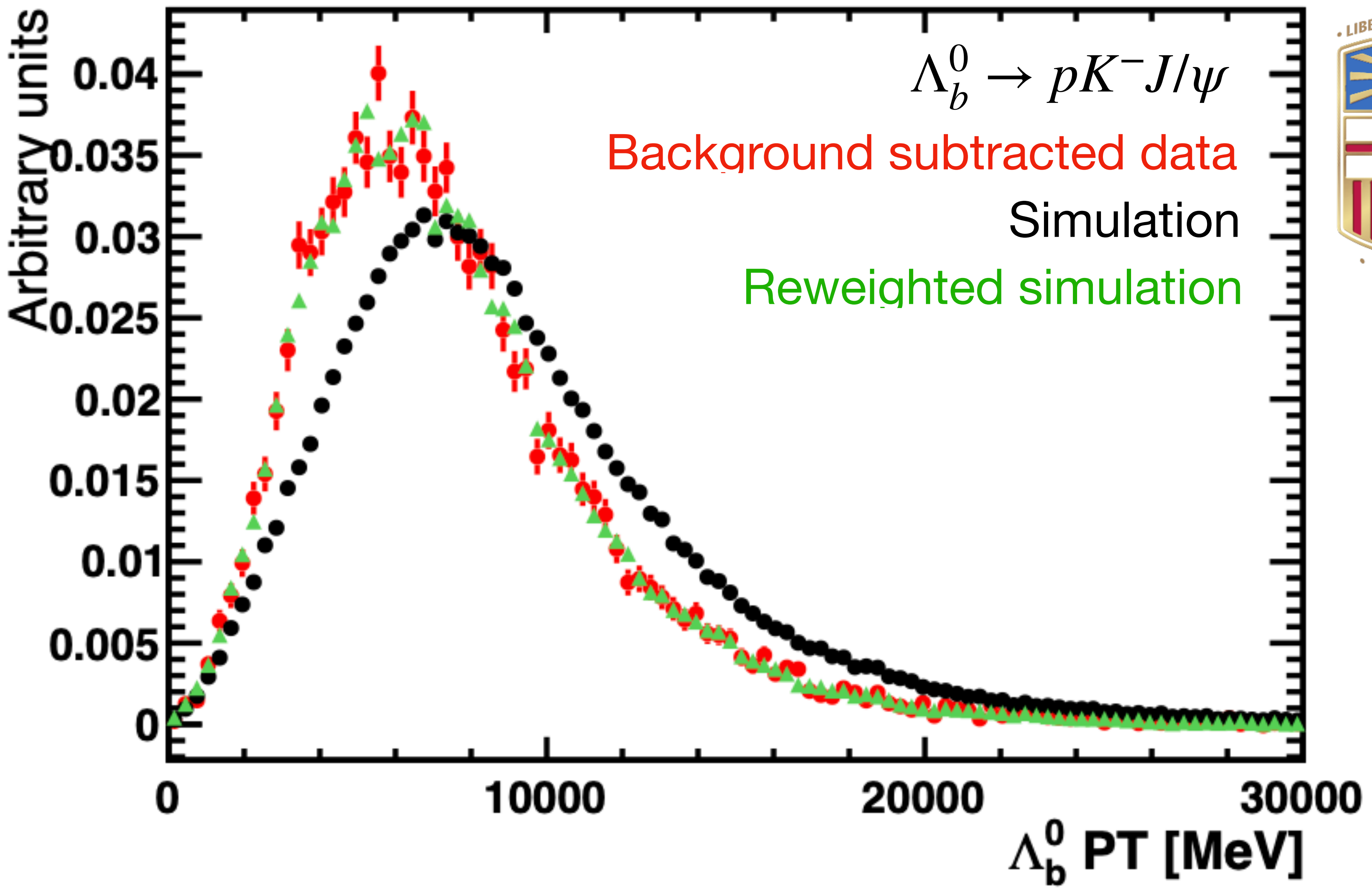
After BDT cut



[R. Niet, PhD thesis]

Reweighting of simulation

- Known and unknown mis-modellings of kinematics, PID, multiplicities, … in simulation
- Multidimensional reweighting with BDTs [arXiv:1608.05806]
- Essential for signal BDT training & efficiency calculation

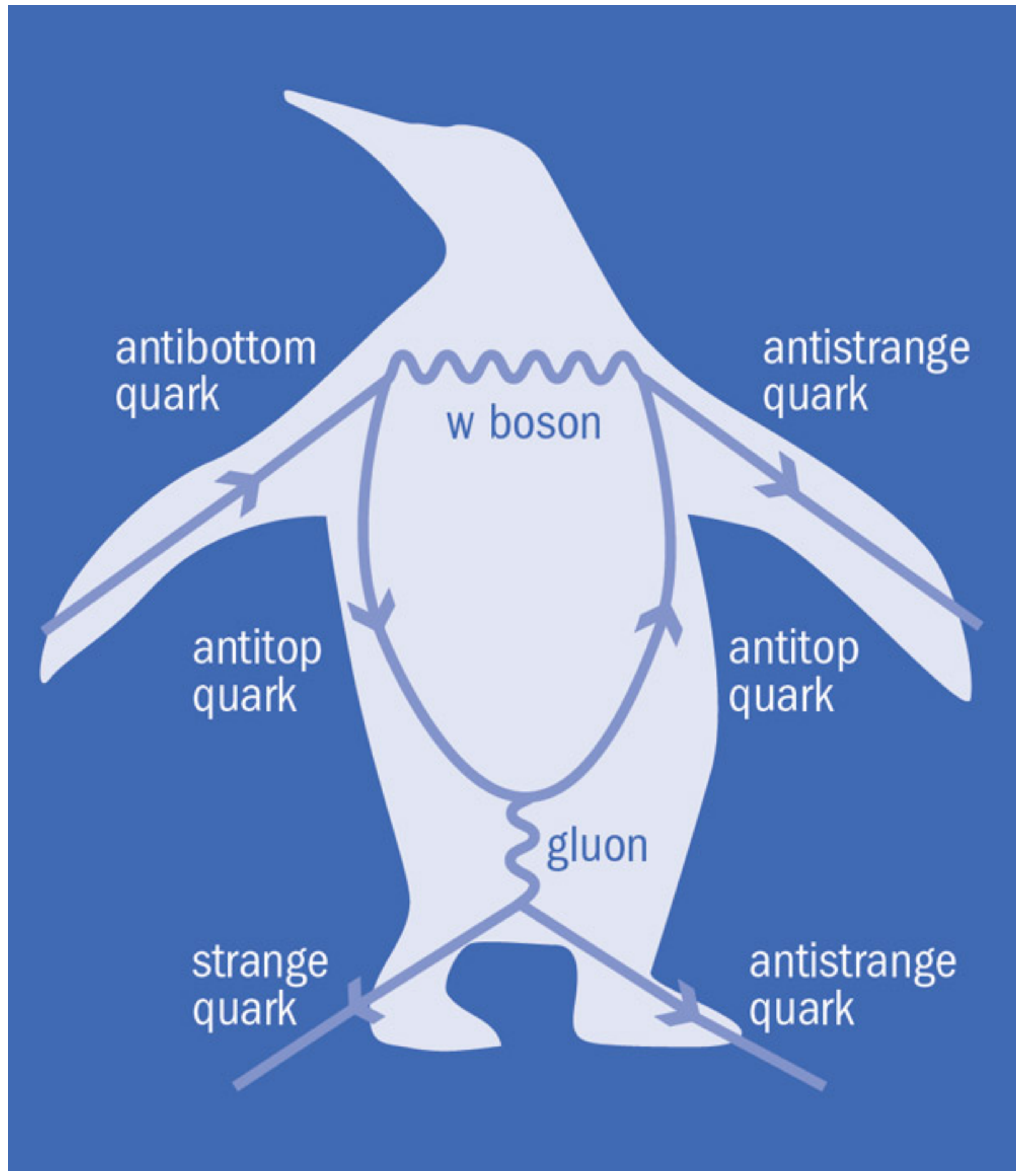


Current ML activities in UB LHCb group

- Using BDT classifiers for background rejection in various data analyses of **semileptonic or rare electroweak penguin decays**
  → **The entire UB LHCb group**

- Muon identification in Run 3
  → **Ricardo Vázquez Gómez**

- Inclusive trigger for rare radiative decays
  → Developed for the Run 2 data taking
  → Adapted for Run 3 data taking
  → **Aniol Lobo Salvia, Carla Marín Benito**

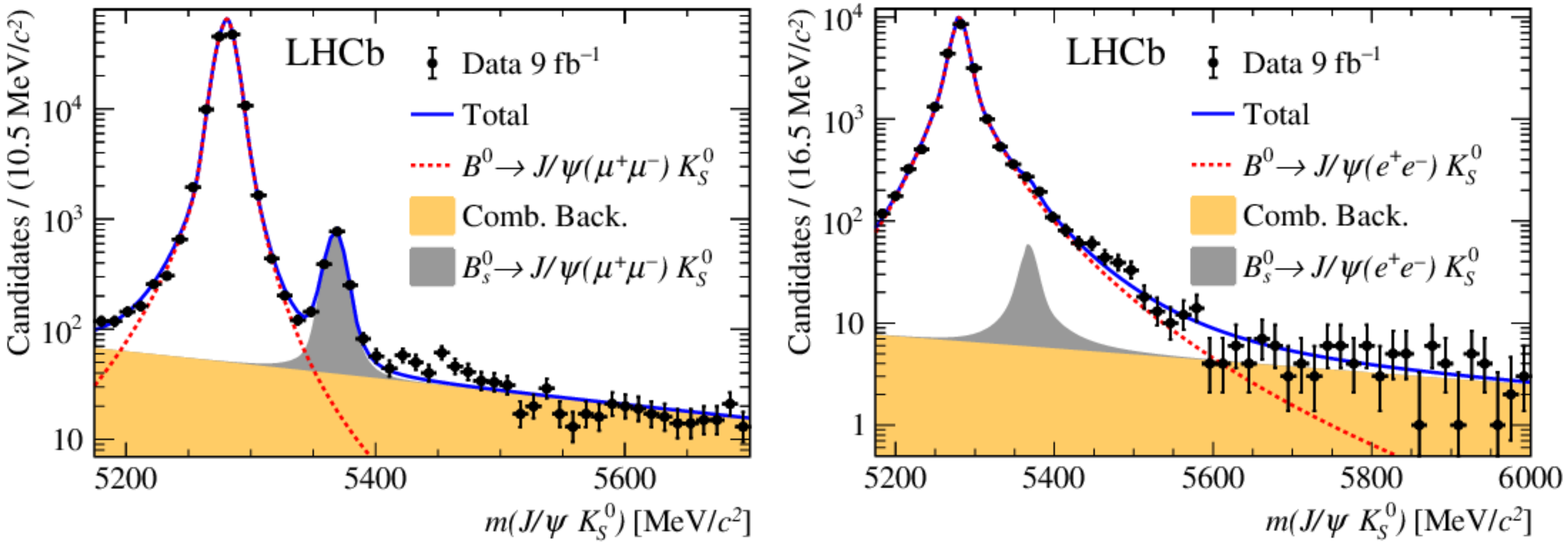


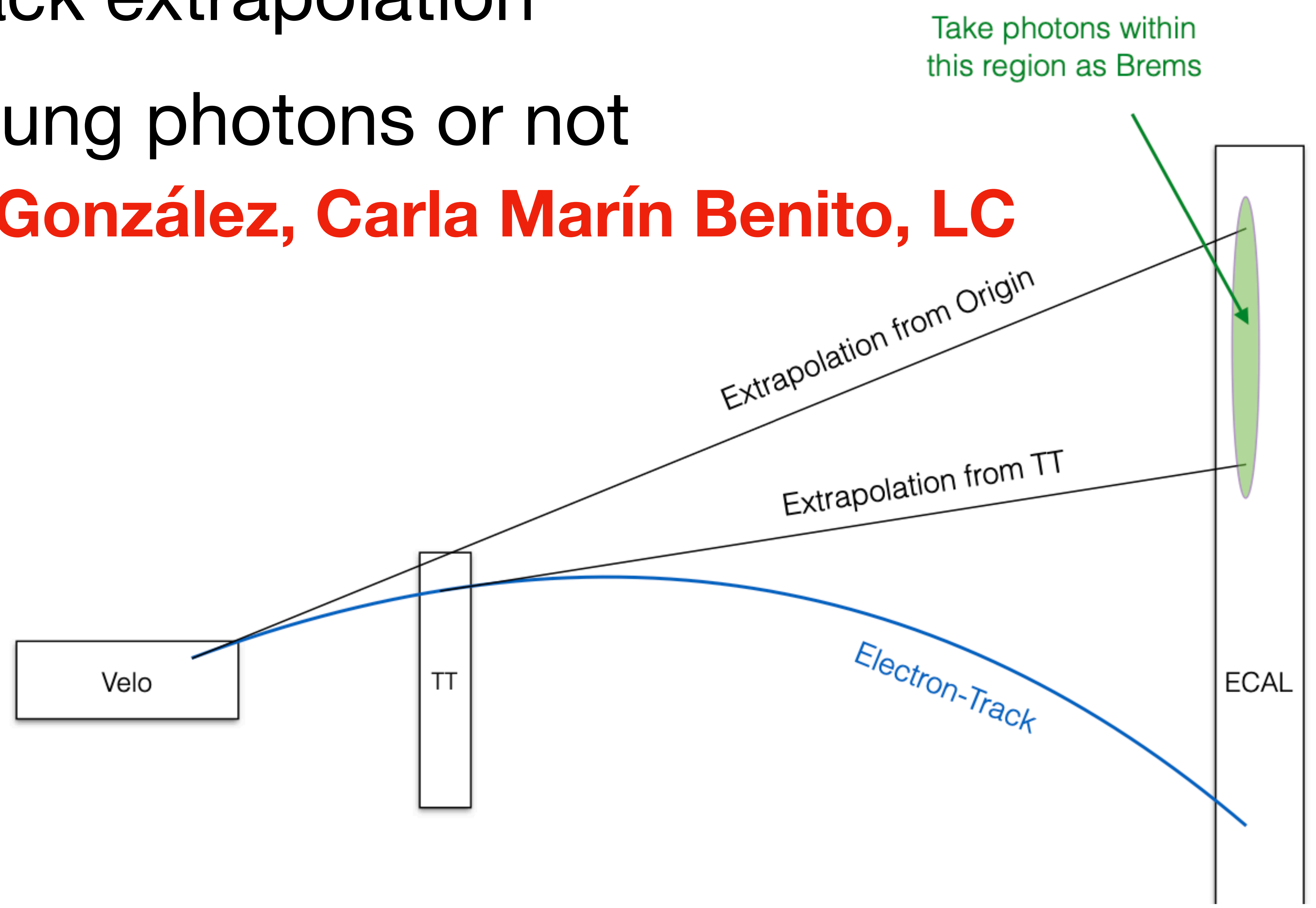
Outlook - ML for bremsstrahlung recovery

- Electron vs. muon reconstruction:
  → Electrons emit bremsstrahlung when traversing detector material
  → Momentum will be biased if emitted bremsstrahlung photons are not identified
- Bremsstrahlung recovery algorithm in the LHCb reconstruction adding photons to the electron track based on compatibility of ECAL clusters with the track extrapolation
- Idea for ML: Check if photons are the correct bremsstrahlung photons or not

→ Paloma Laguarta González, Carla Marín Benito, LC

[Phys. Rev. Lett. 128, 191802]





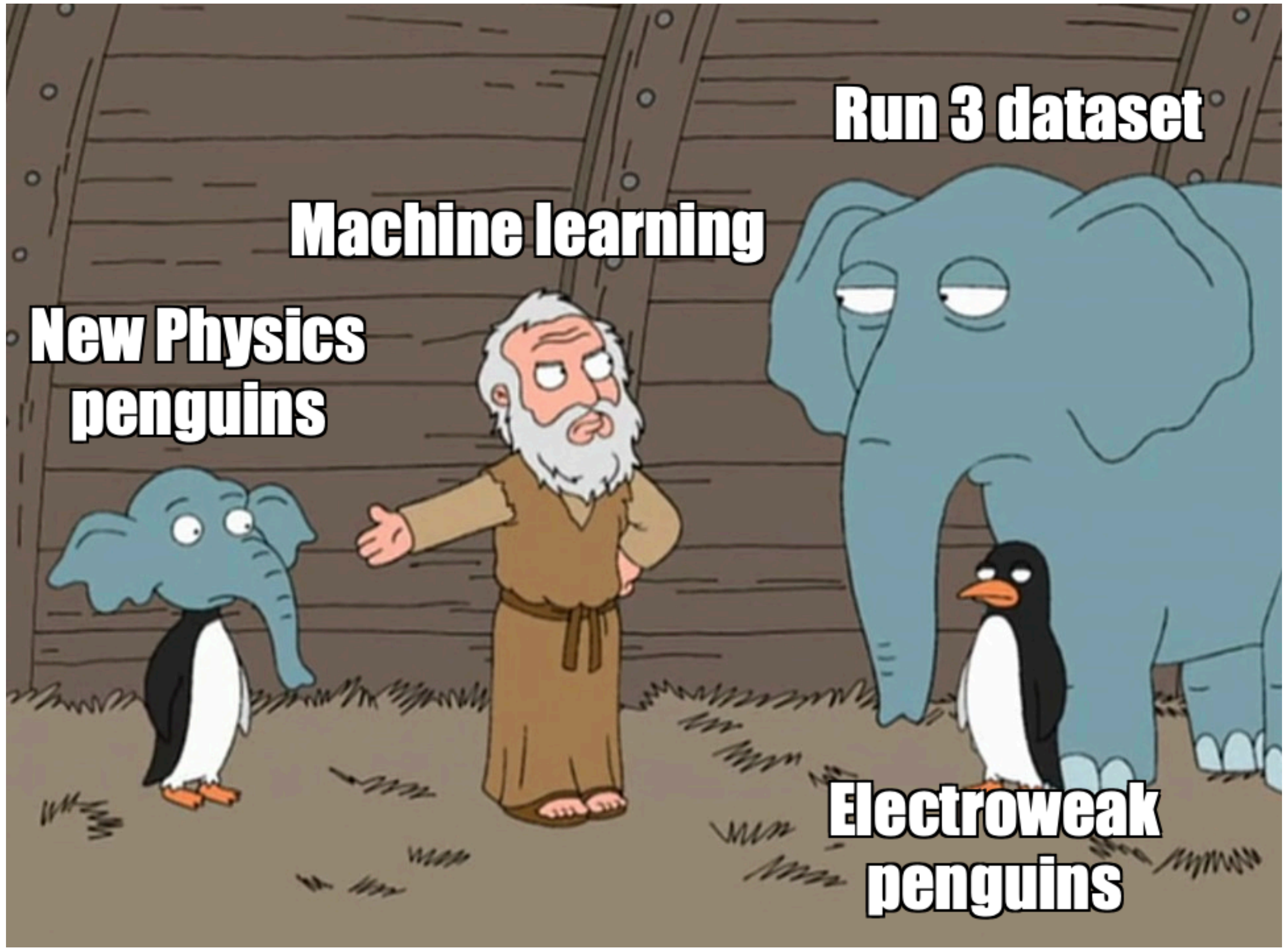
Conclusions

- LHCb has a variety of ML applications
- Insight given in this talk:
  - ML techniques in track reconstruction, particle identification
  - Inclusive trigger selections and offline data analysis
- UB LHCb group has experience in various applications
- Current and future projects in offline data analysis and bremsstrahlung recovery

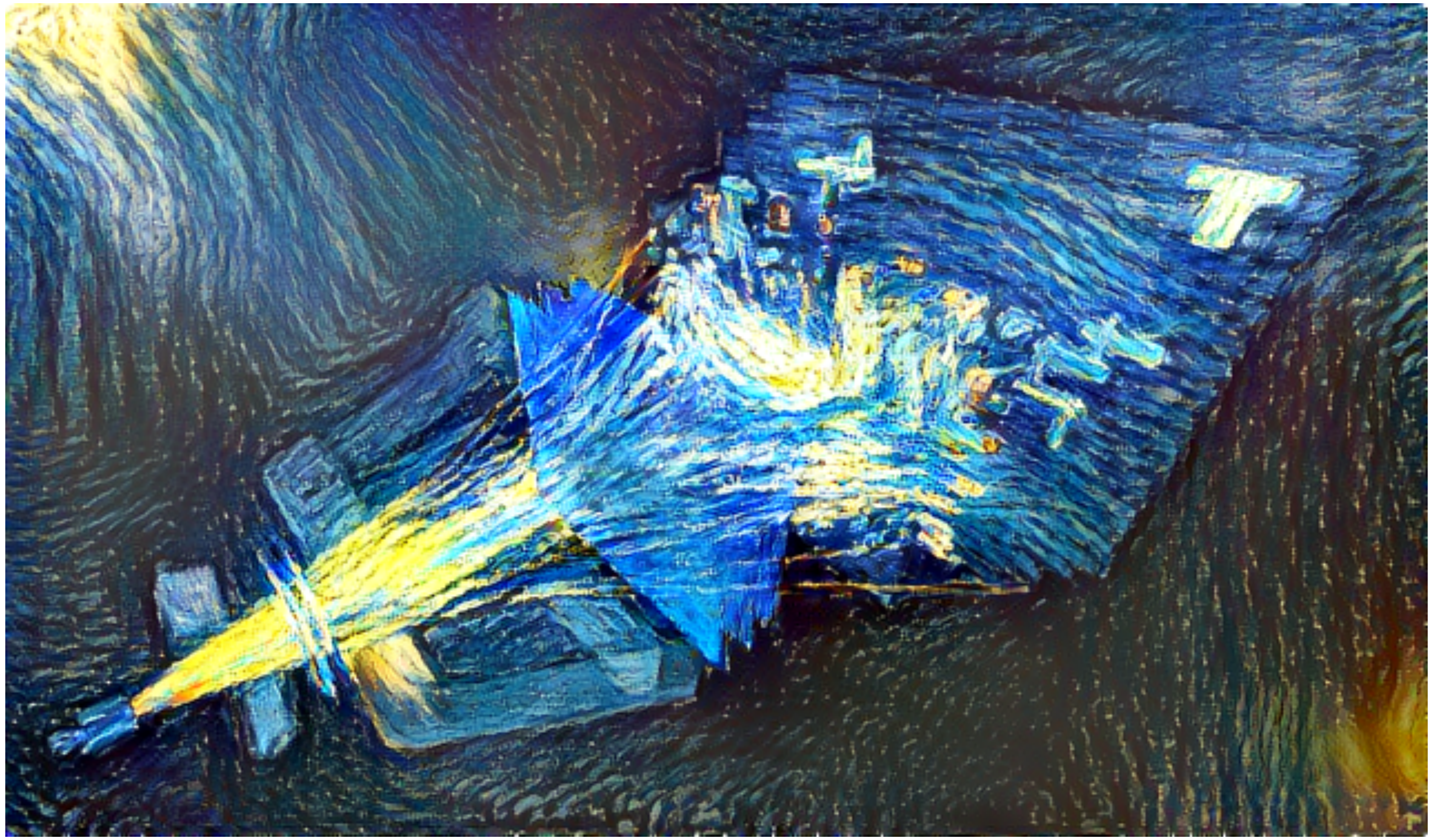
Many other applications in LHCb that I didn't cover

- data-quality monitoring
- jet reconstruction
- flavour tagging
- calorimeter reconstruction
- fast simulations
- …

Thanks for your attention!



"The wishful thinking of a LHCb physicist"



by Andrey Ustyuzhanin
https://github.com/jcjohnson/neural-style