PID4SMOG: A Neural-Network-defined Gaussian Mixture Model for particle identification applied to the LHCb fixed-target programme

Saverio Mariani
CERN

ML INFN, 06/03/22

About me…

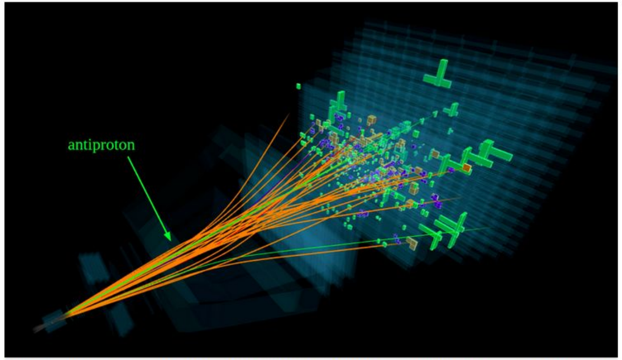
saverio.mariani@cern.ch
PhD at Florence INFN (fixed-target physics at LHCb)
Now senior research Fellow at CERN

My research activities:
LHCb fixed-target beam-gas physics, especially for cosmic rays interest (antiproton production in *p*He collisions)

LHCb reveals secret of antimatter creation in cosmic collisions

The finding may help determine whether or not any antimatter seen by experiments in space originates from dark matter

7 APRIL, 2022



A proton-proton collision event recorded by the LHCb detector, showing the track followed by an antiproton formed in the collision (Image: CERN)

But, more importantly:

I love seeing new places and knowing new people

I do cook/bake a lot, especially if shared with friends/family

Never tired of my dog and my nephews (and their chaos)

About this project

- **Why?** PID calibration for LHCb fixed-target beam-gas data suffers from the low statistic
  - PID efficiencies are **one of the dominant uncertainties** in analyses
- **What? Learn from fixed-target most abundant sample** how the PID depends on the event features and robustly **extrapolate for a lower-statistic one**

JINST 17 (2022) Published by IOP Publishing for Sissa Medialab

Received: *November 4, 2021*
Accepted: *January 21, 2022*
Published: *February 9, 2022*

A Neural-Network-defined Gaussian Mixture Model for particle identification applied to the LHCb fixed-target programme

G. Graziani,[a] L. Anderlini,[a] S. Mariani,[a,b,c,*] E. Franzoso,[d,e] L.L. Pappalardo[d,e] and P. di Nezza[f]

- **How?** Model the training PID classifiers through a **maximum-likelihood fit** with the composition of **multinormal functions** initialized with **neural networks** fed with the feature values
- **Who?**

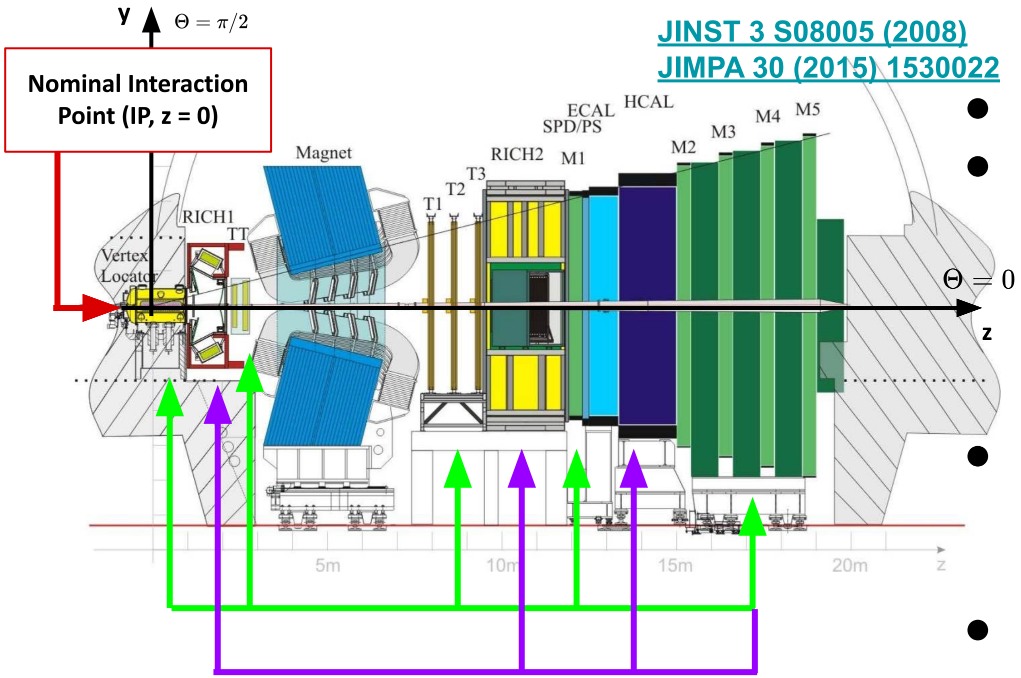
1

Introduction and motivation

The LHCb detector

- Designed for **heavy flavour physics**, the instrumented region covers $\Theta \in [10, 250]$ mrad

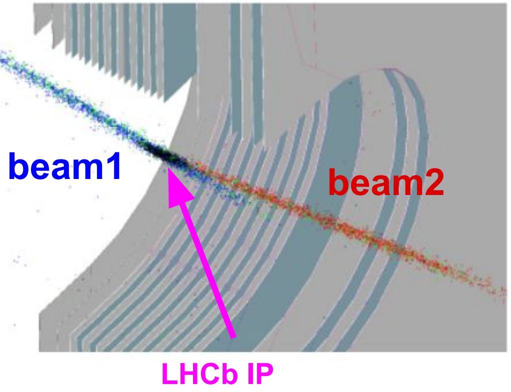
JINST 3 S08005 (2008)
JIMPA 30 (2015) 1530022

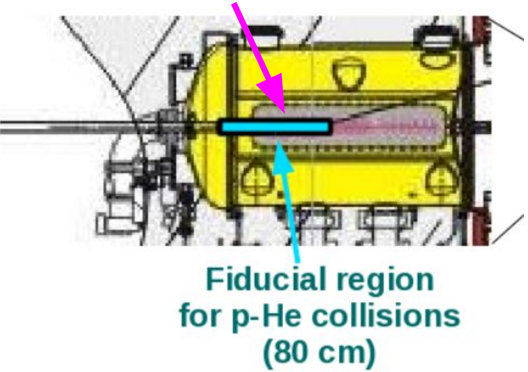


- **Complementary** wrt other LHC experiments
- Tracking system: **VErtex LOcator** + tracking stations upstream and downstream of a magnet
  - 0.5-1% p resolution for p < 300 GeV/c
  - 10-80 µm IP resolution
- Particle identification (PID): Two **Cherenkov detectors (RICH)** + calorimetric and muon systems
- **Flexible and versatile trigger**

The LHCb detector in fixed-target mode (I)

JINST 9, (2014) P12005





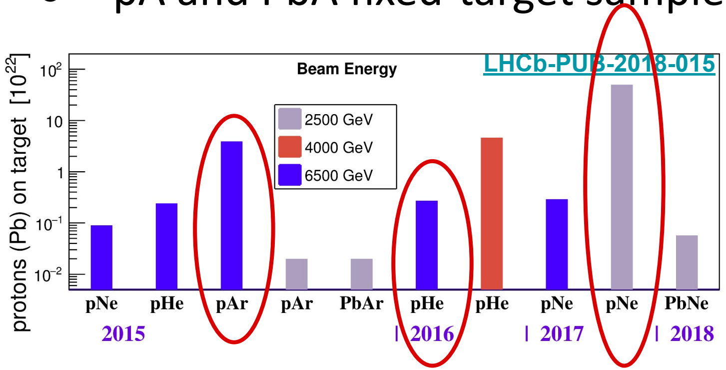
- Since 2011, LHCb is equipped with the **System for Measuring Overlap with Gas (SMOG)**
  - Used to complement the LHC luminosity measurement by reconstructing the **LHC beams transverse profiles** via proton collisions with the **small quantity of injected gas (10^{-7} mbar)**

- In proximity of the LHCb IP, **the proton-nucleus interaction can be fully reconstructed**!

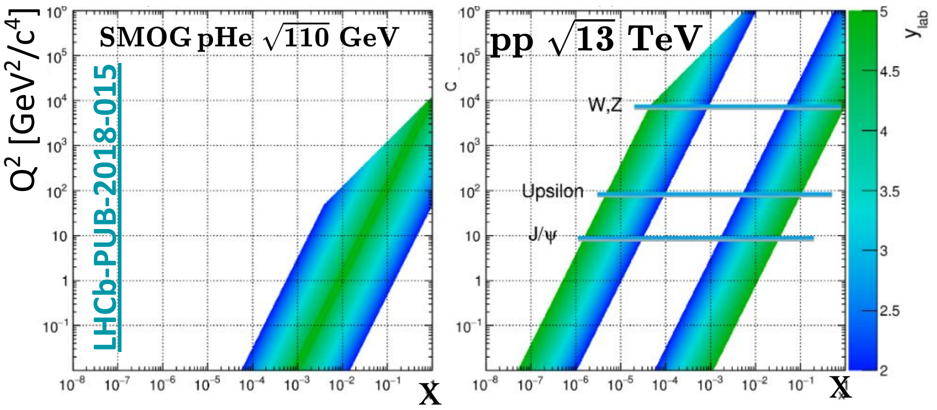
- Forward detector + gas target = **highest-energy fixed-target ever!**

The LHCb detector in fixed-target mode (II)

- pA and PbA fixed-target samples **collected during special runs in 2015-2018**



e.g. 6.5 TeV LHC protons on at-rest He correspond to a nucleon-nucleon centre-of-mass energy $\sqrt{s_{\mathrm{NN}}}$ =110 GeV

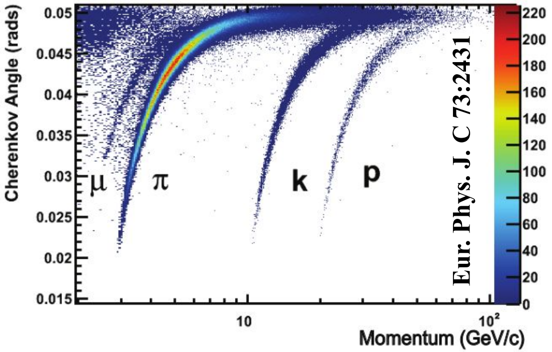


- **Intermediate energy** to SpS and LHC scales
- Many collision systems (Z dependence)
- **Access to the moderate Q^2 and large target Bjorken-x** (the nucleon momentum fraction carried by the colliding parton) region

→ **Unique experimental inputs**

Particle identification at LHCb

- How to distinguish **pions, kaons and (anti)protons** produced in each collision?

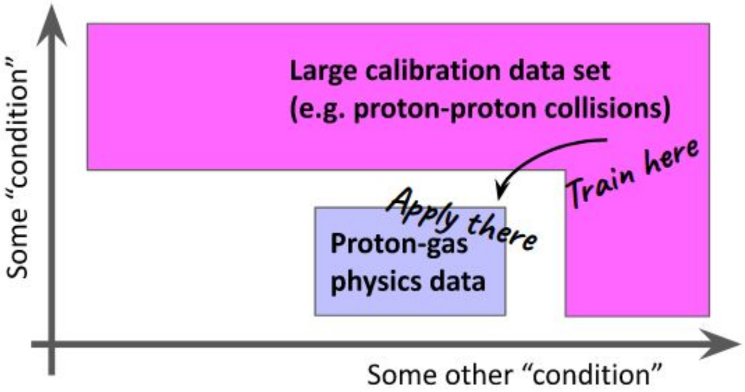


$$cos\theta_c = \frac{1}{n\beta} = \frac{1}{n}\sqrt{1+\left(\frac{mc^2}{pc}\right)^2}.$$

Reconstruction of the Cherenkov angle

$$\mathrm{DLL}_{h1,h2} = log\left(\frac{\text{h1 likelihood}}{\text{h2 likelihood}}\right)$$

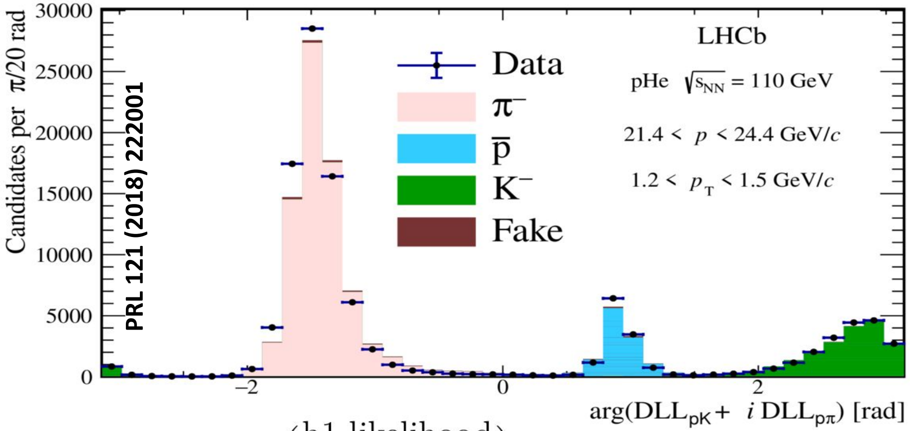
Fit to Cherenkov rings to define likelihood functions for each particle hypothesis



- **Problem:** the simulation cannot be fully trusted, hence PID is **calibrated on decays** selected with no PID info and then applied to the signal of interest
- **How robust is the extrapolation**, provided that **the calibration and application phase-spaces differ**?

Fixed-target particle identification at LHCb

- Calibration channels can be reconstructed and selected with high statistics in *pp* data, but **statistics is not sufficient in some of the fixed-target** collected samples
- PID calibration from *pp* cannot be efficiently applied to fixed-target because of the **poor phase-space coverage** (different occupancy, momentum, *z* distribution…)



$$\mathrm{DLL}_{h1,h2} = log\left(\frac{\text{h1 likelihood}}{\text{h2 likelihood}}\right)$$

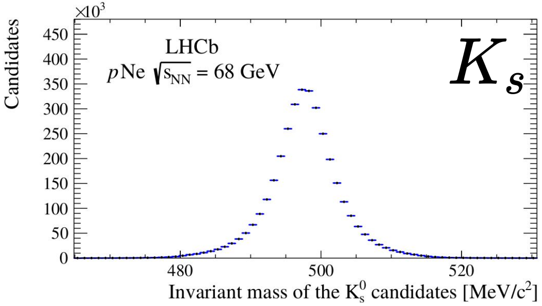
- **Example**: $\sigma(p\mathrm{He} \to \bar{\mathrm{p}}\mathrm{X}, \sqrt{s_{\mathrm{NN}}} = 110\,\mathrm{GeV})$
- Prompt antiprotons are counted with a **template fit to PID variables**
- PID fit quality not satisfactory and PID found as one of the **dominant contributions to the systematic uncertainty**

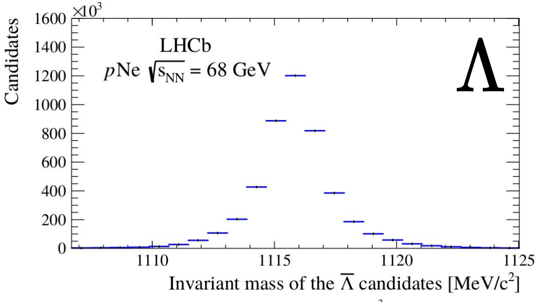
2

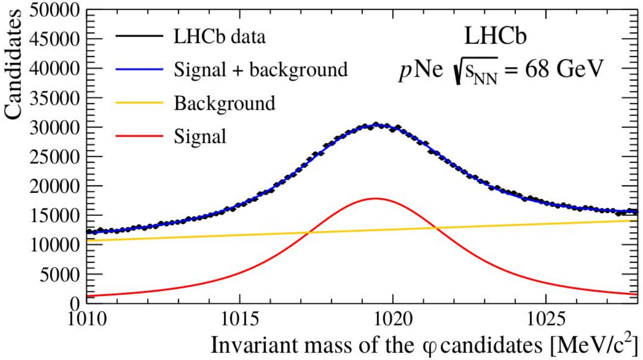
ML model

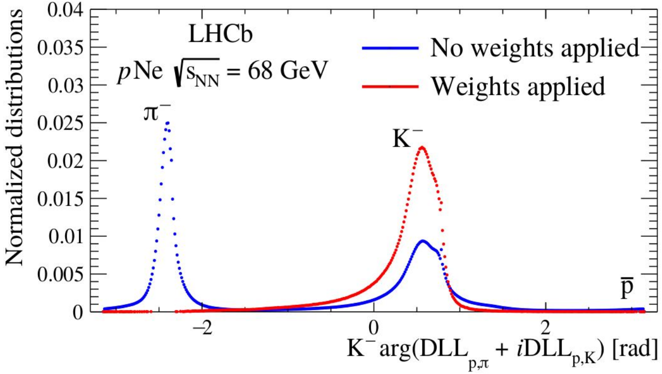
Calibration channels

- The $\Lambda \to p\pi$ ($\bar{\Lambda} \to \bar{p}\pi$), $K_s \to \pi\pi$ and $\phi(1020) \to KK$ decays are reconstructed and selected (with no PID cuts) in the **SMOG largest-statistics sample (*p*Ne)**
- Large purity achieved for Λ and Ks thanks to their description in the Armenteros plot









- *sPlot* performed on *φ* by fitting the invariant mass with a Voigtian + first-order polynomial
- Weights validated as being the pion contamination suppressed

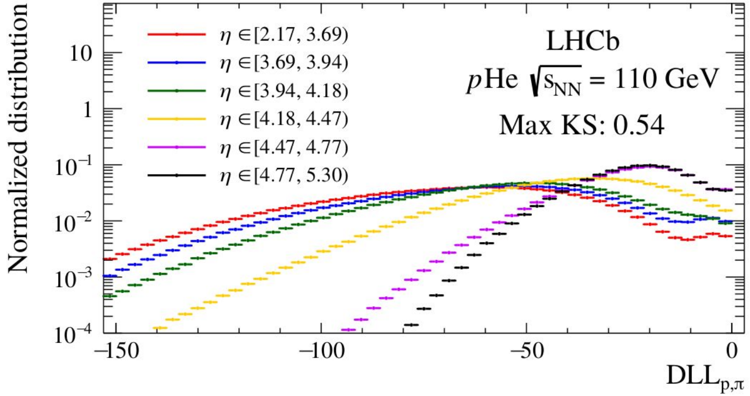
Gaussian Mixture Model

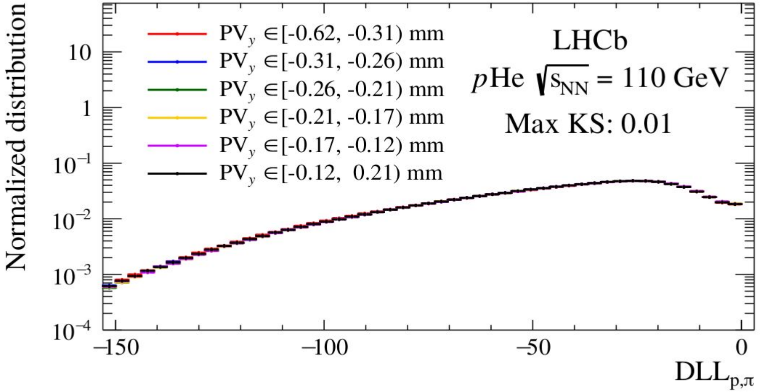
- For each decay channel, the 2D DLL distribution $\underline{x}_p$ (*e.g* $\mathrm{DLL}_{p,\pi}$ and $\mathrm{DLL}_{p,K}$) is modelled with a sum of N_g multinormal distributions:

$$\underline{x}_p \sim \sum_{j=1}^{N_{g,p}} \alpha_{j,p}(\underline{\theta}) \frac{\exp(-\frac{1}{2}(\underline{x}_p - \underline{\mu}_{j,p}(\underline{\theta}))^T \, \Sigma_{j,p}^{-1}(\underline{\theta}) \, (\underline{x}_p - \underline{\mu}_{j,p}(\underline{\theta})))}{2\pi\sqrt{\det(\Sigma_{j,p}(\underline{\theta}))}} \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

- All multinormal parameters are a function of the **features $\boldsymbol{\theta}$**, representing the **physical quantities affecting the RICH response**
- To properly take into account correlations and to enhance the template tails statistical significance, each parameter is the output of a **Neural Network (NN) fed with $\boldsymbol{\theta}$**
- Number of multinormal N_g and the NN structure (depth, nodes..) defined by the user

Feature variables choice





- Which features do affect most the RICH response?
- Ordered according to the **max. Kolmogorov-Smirnov (KS) distance** between all pairs of DLL histograms plotted in bins of each variable

Variable	Max KS	Variable	Max KS	Variable	Max KS
p_z	0.64	p	0.64	η	0.54
p_T	0.51	$yz\ slope$	0.38	$track\ ndf$	0.34
$xz\ slope$	0.34	$nTracks$	0.34	$nRich2Hits$	0.33
$nSpdHits$	0.32	$nRich1Hits$	0.28	$track\ \chi^2/\mathrm{ndf}$	0.26

- Relevant features reflect **particle kinematics, detector occupancy and reconstruction quality**
- **Geometry added** to consider the difference between training (detached) and validation (prompt) particles

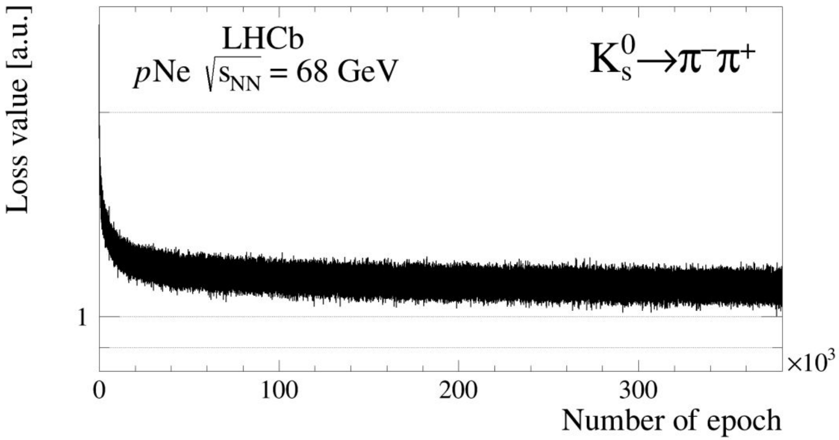
Preprocessing and training

- To ease the convergence, DLL variable are rescaled to [0, 1] (MinMaxScaler algorithm), features are converted into Gaussians (QuantileTransformer algorithm)
- For each calibration decay, training on n_p *p*Ne events by minimizing a **loss** defined as **the opposite of the maximum likelihood**:

[1], [2]

$$\mathcal{L} = -\sum_{i=1}^{n_p} w_i \log \left[\sum_{j=1}^{N_{g,p}} \alpha_{j,p}(\underline{\theta}_i) \mathcal{G}(x_i, \mu_{j,p}(\underline{\theta}_i), \sigma_{j,p}(\underline{\theta}_i)) \right]$$

(being w_i the *sPlot* weights for the φ line)

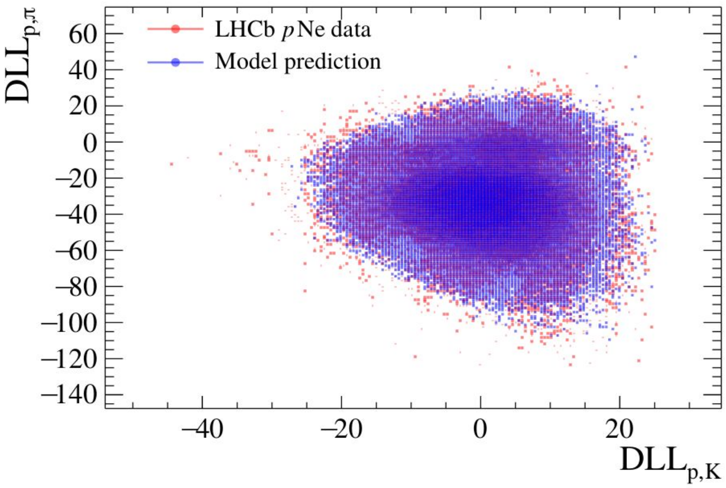


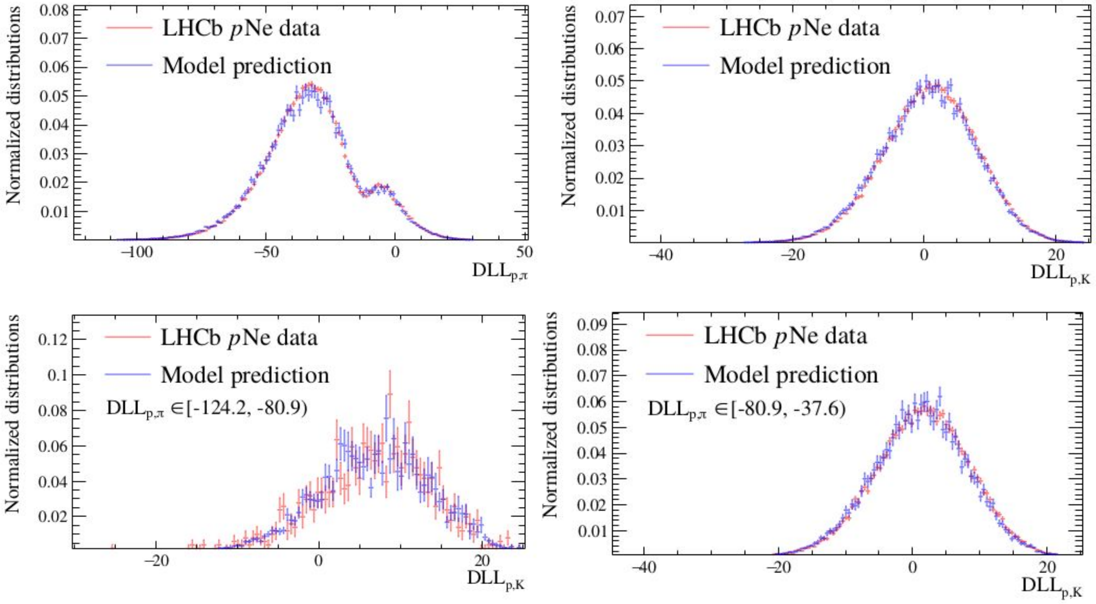
- NN weights are adjusted as a function of θ to maximize the likelihood wrt training data
- **The $x_p(\theta)$ relation is learned!**
- Steep decreasing, followed by a gentle one and an oscillation around the minimum

Validation

- To verify that the **trained model** has correctly learned to reproduce **the data**, these are compared in bins of all possible feature pairs (below p є [12.0, 15.5) MeV/c, η є [4.1, 4.4))

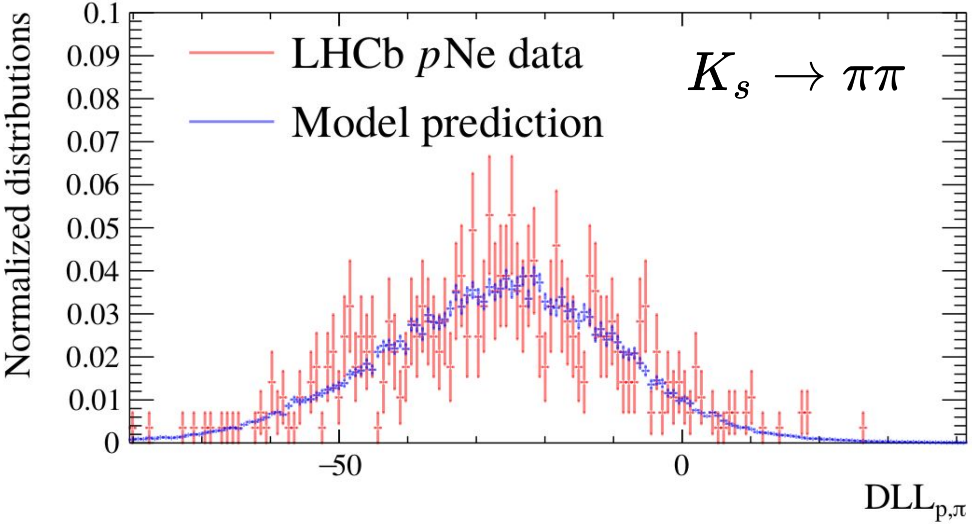
$K_s \to \pi\pi$

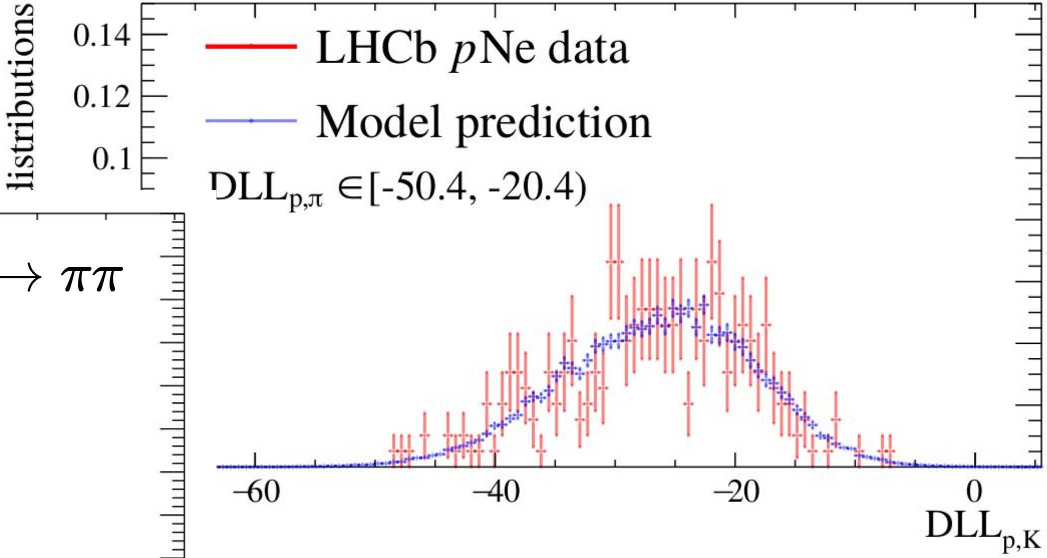




Validation (II)

- Also, based on the available information, the model is able to draw **a smooth template in low statistics phase-space regions**!



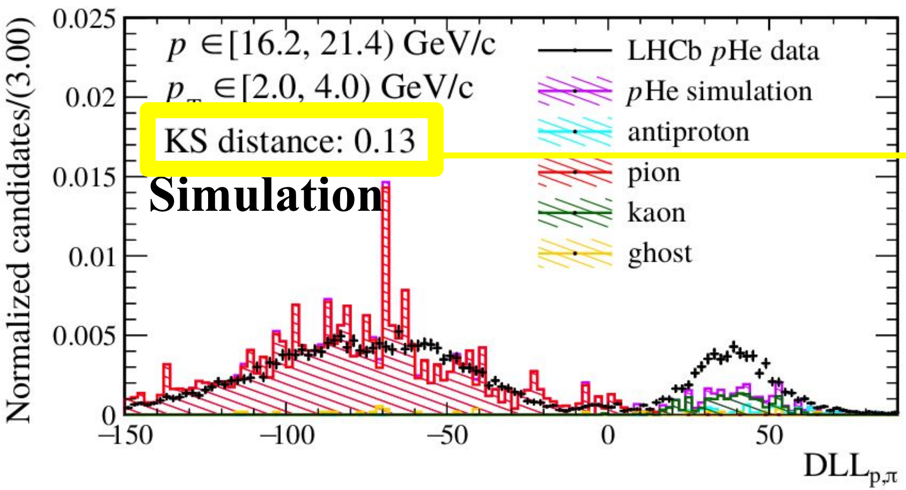


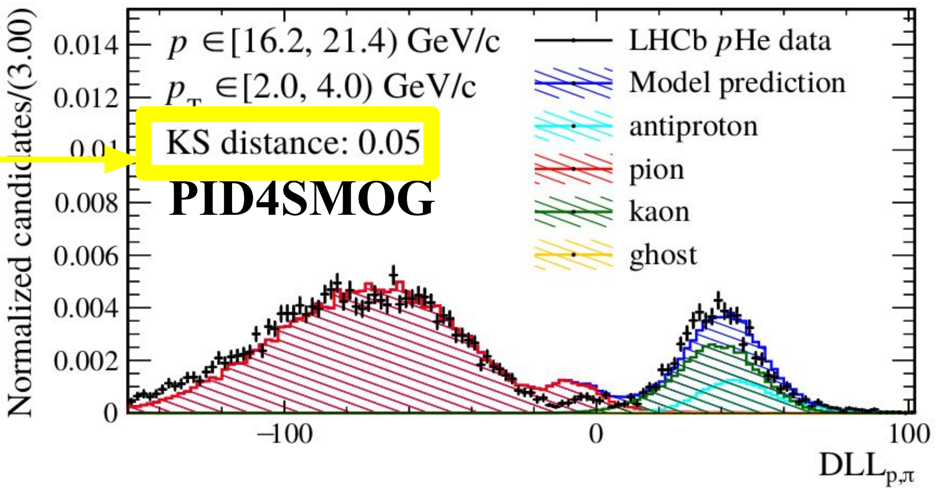
3

Use cases and prospects

Generalization to *p*He and *p*Ar data samples

- Using the trained models, templates are produced for prompt antiproton candidates in the 2016 *p*He and 2015 *p*Ar data, **according to their feature distributions** (different wrt pNe!)
- **Fit procedure** followed in the antiproton measurement repeated with the composition of simulated and predicted templates and compared

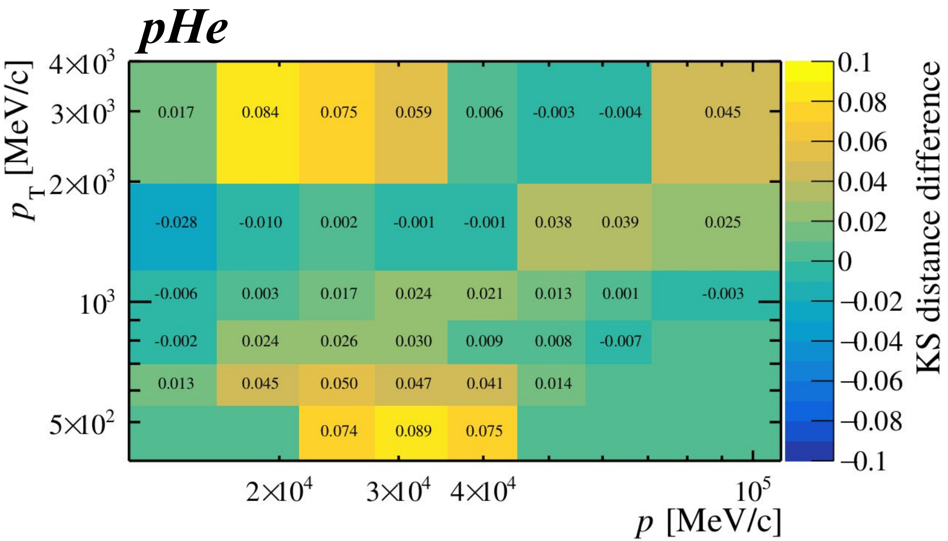


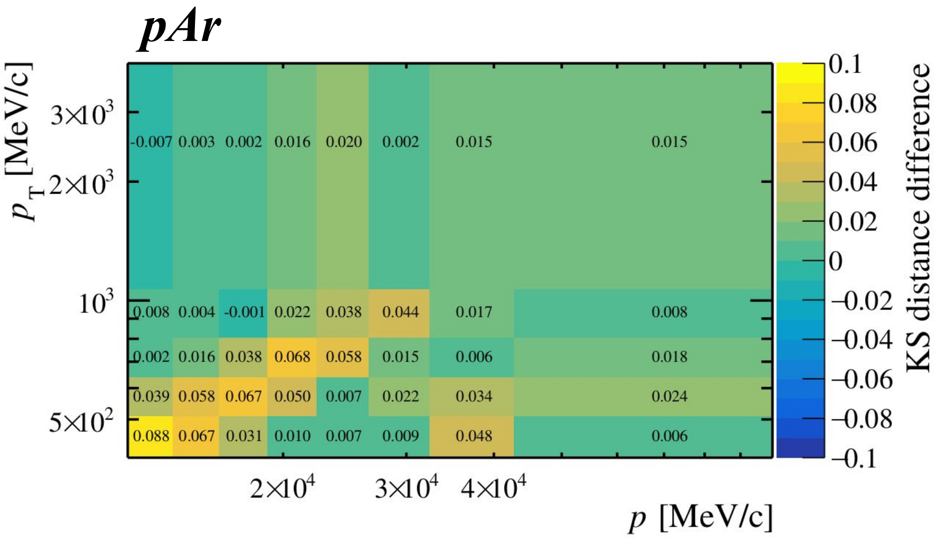


- **Improvement in the data description** evident and measured in the KS distance!

Generalization to *p*He and *p*Ar data samples (II)

- Procedure iterated in kinematic bins and **KS distance between data and simulated or predicted templates** composition measured



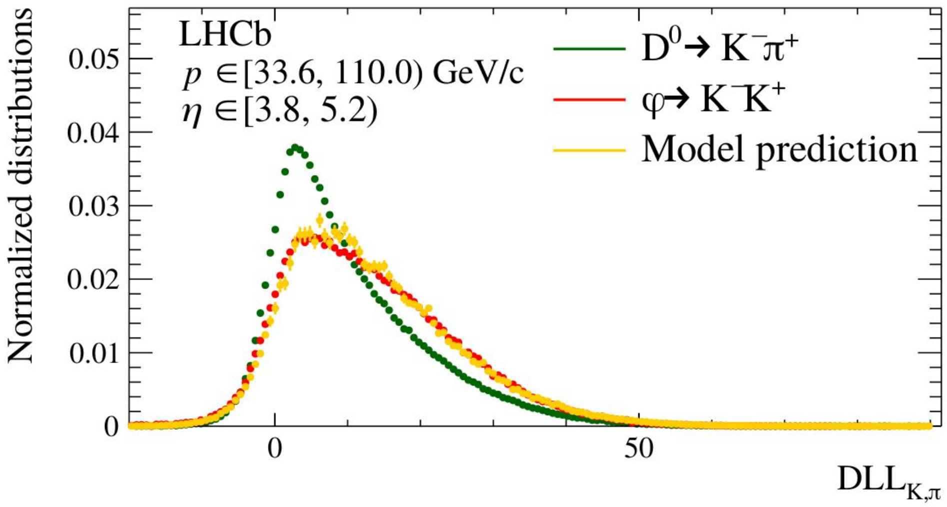


- Difference between KS with simulation and prediction **mostly show positive values**
- **Our model offers an equal or better data description than the detailed simulation**

Other use cases

- **PID eff calculated via PID4SMOG** already in two other analyses:
  - Detached-to-prompt antiproton ratio in *p*He (LHCb-PAPER-2022-006)
  - Quarkonia and D^0 production in 2017 *p*Ne (LHCb-PAPER-2022-015)

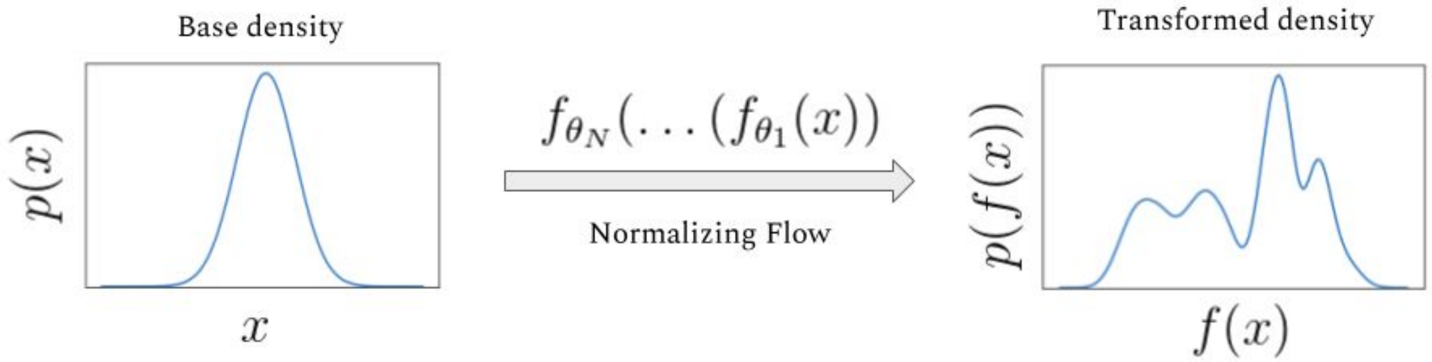
- Also, **for *pp*,** where statistics is sufficient, PID4SMOG can be used to compare different calibration channels:



  - A possible **correlation between kaon tracks from φ decays** inducing a bias in PID studies is investigated by training a model on **2017 *pp* D^0** data and predicting DLLs for **kaons from 2017 pp φ**
  - The match of the prediction (not taking into account the possible correlation) with **φ data** excludes the effect

Prospects

- **Main limitation** of the model atm is that it only supports a bidimensional target (which was motivated, being the goal the π-K-p separation)



- Plan is to move to **density estimation via normalizing flows,** efficiently supported in TF2, to overcome the dimensionality limitation
- **PID5SMOG is on the horizon**, but, unfortunately, people power is very limited atm

4

Conclusions

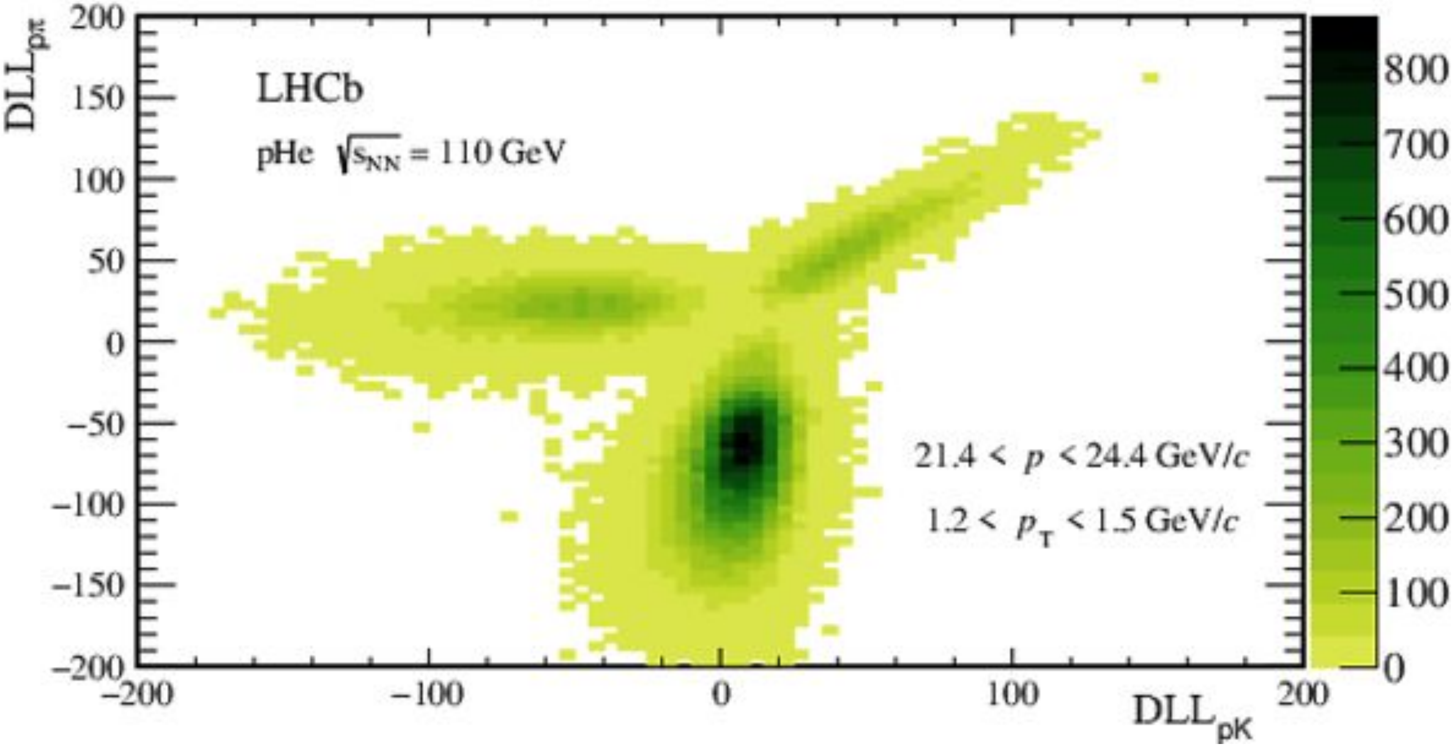
Conclusions

- **PID4SMOG: Data-driven machine-learning-based** approach to the PID conceived to perform robust extrapolations (for SMOG, this **mitigates one of the dominant uncertainties**)
  - **Calibration channels** reconstructed and selected in *p*Ne data for pions, kaons, protons
  - Training data modelled as a **Gaussian Mixture Model** with all parameters determined by **Neural Networks** fed with a set of relevant experimental features
  - **Significant improvement in the description** of *p*He and *p*Ar samples wrt simulation
  - Some use-cases for SMOG and *pp* data presented and **prospects to overcome limitations are clear**

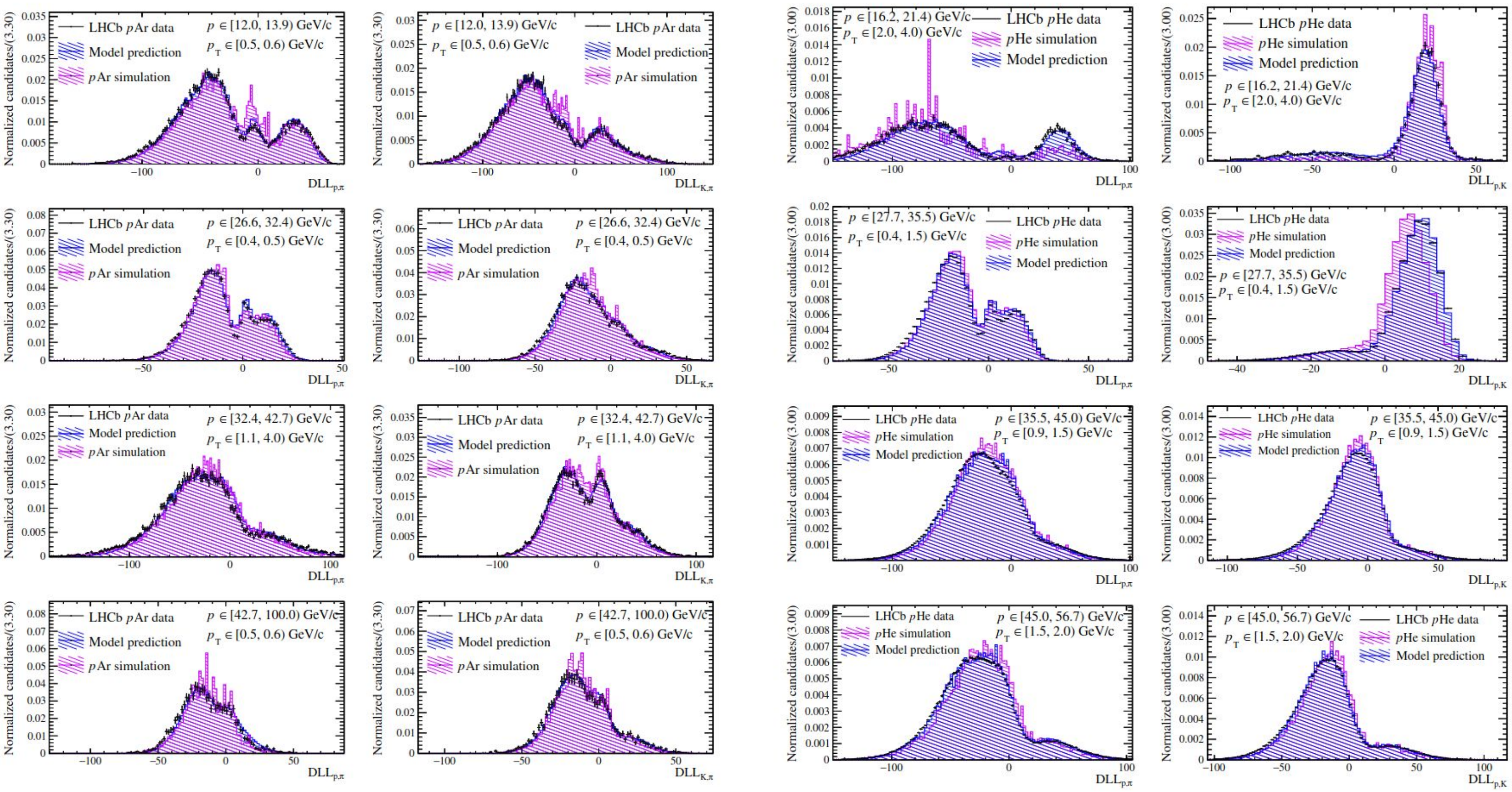
Thanks for your attention!

Follow up? saverio.mariani@cern.ch

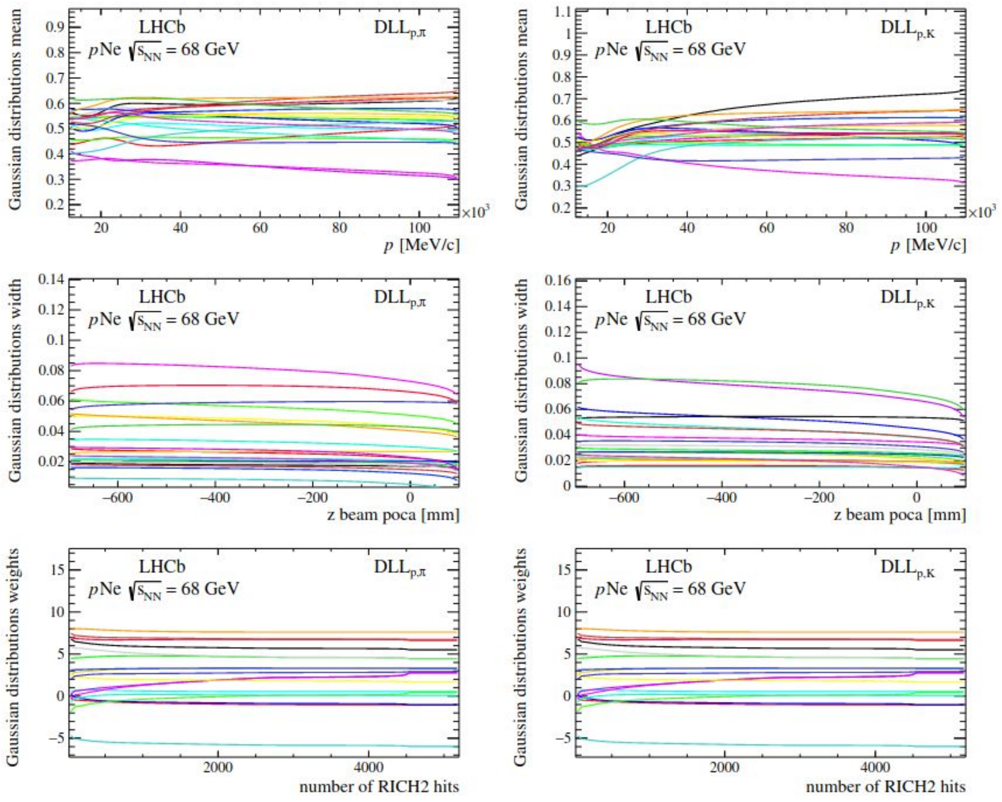
DLL distribution for fixed-target data



Model application in kinematic bins



Overtraining?



- **Overtraining** not a worrying issue in this application, since goal is to learn a relation
- Possibly, multinormal parameters could be rapidly adapted to training data in phase-space corners

- Smooth parameters evolution as a function of features indicates this is not the case