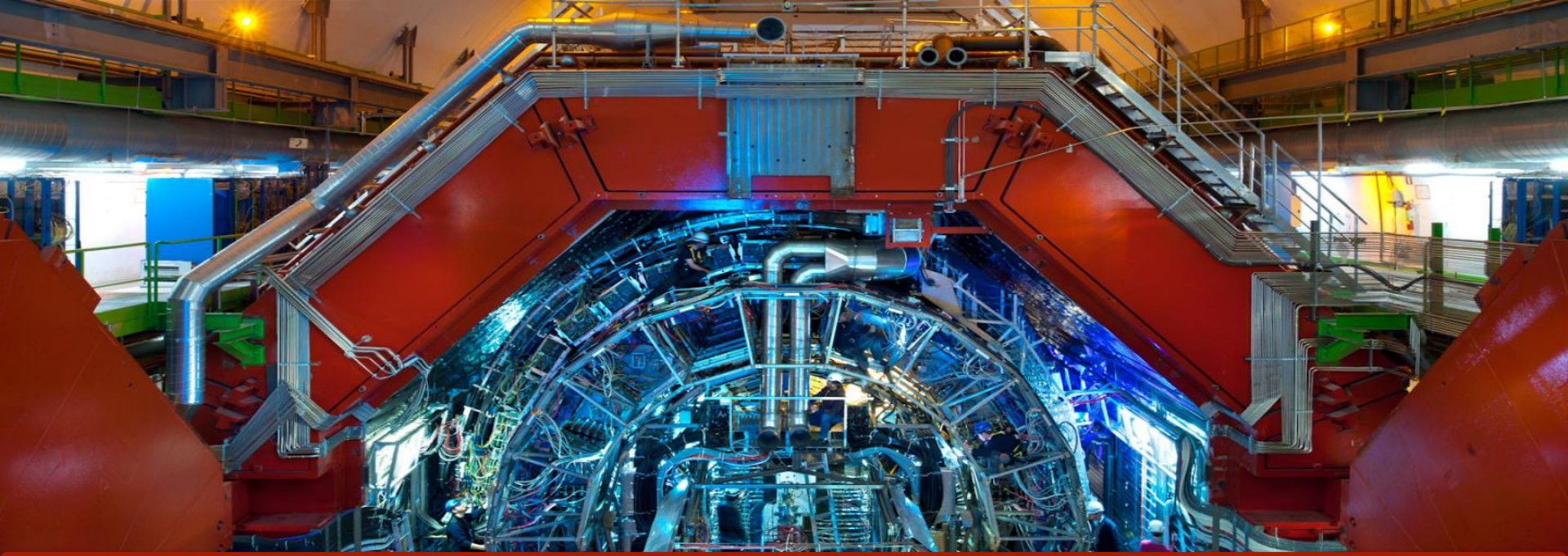


Machine learning in ALICE: status and perspectives

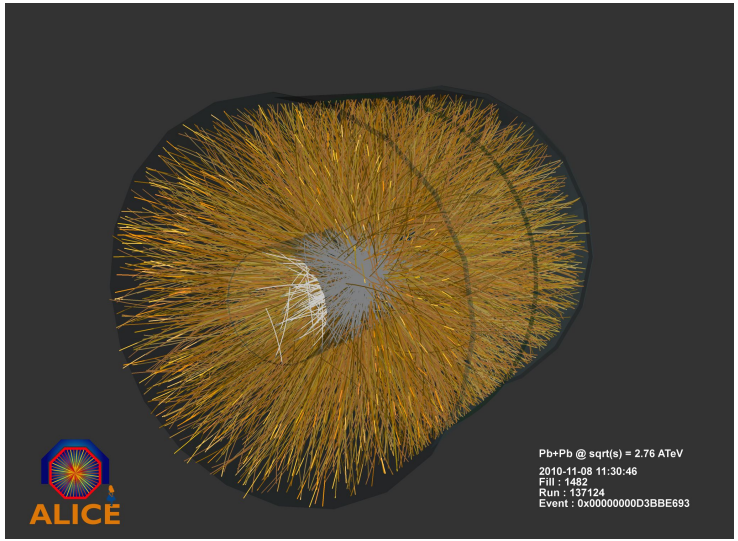
Francesco Mazzaschi [1, 2]

Workshop di CCR, 24/05/21-28/05/21

Università degli Studi di Torino [1], INFN Torino [2]

Machine learning in ALICE

- Machine learning in ALICE
  - fundamental tool to cope with the huge amount of data produced at LHC



- ALICE activities involving ML
  - physics analyses
  - detector calibrations
  - data quality control (QC)
  - Monte Carlo simulations

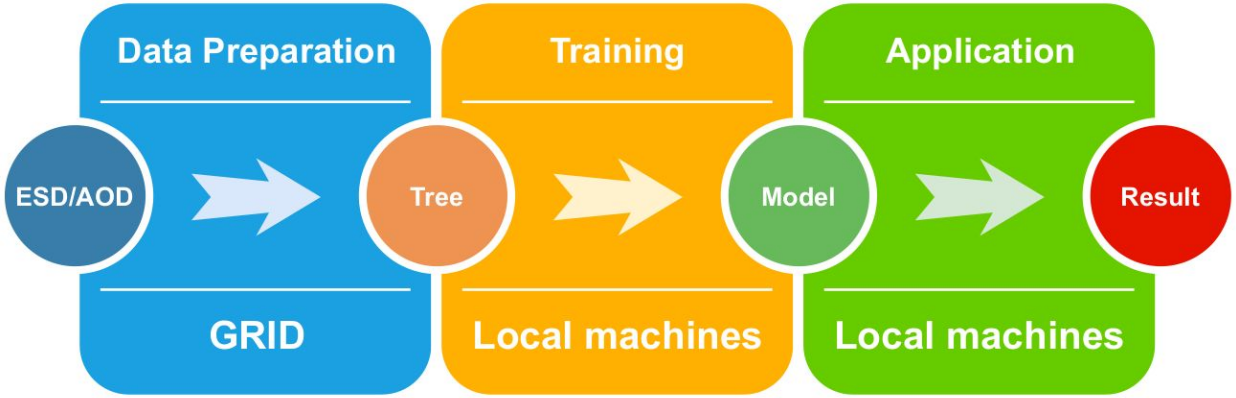
Run 2

Run 3

- INFN group of ALICE strongly involved in many of these activities, focusing mostly on light and heavy flavour analyses

Physics analyses in Run 2

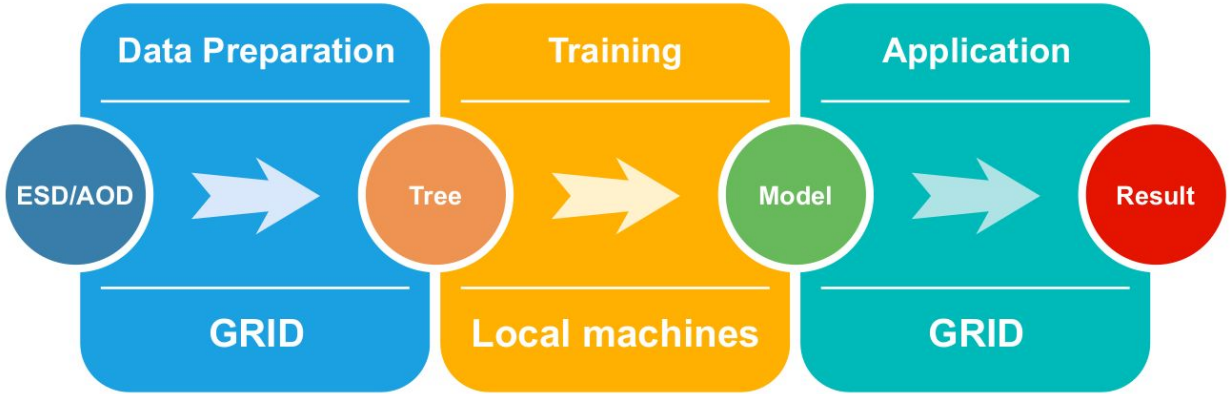
Analysis workflow



- Training and testing on ROOT TTrees of particle candidates satisfying loose selections
  - signal candidates from Monte Carlo simulations
  - background candidates both from Monte Carlo and data (like-sign, sidebands)

- Application on local machines:
  - feasible if the TTree size is lower than few hundreds of GBs (pp and pPb collisions)
  - personal laptop for TTree size up to ~3-4 GBs, otherwise dedicated machine (e.g. Turin ML workstation: 256 GBs RAM, 128 threads)

Analysis workflow

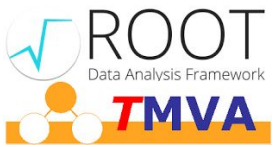


- Training and testing on ROOT TTrees of particle candidates satisfying loose selections
  - signal candidates from Monte Carlo simulations
  - background candidates both from Monte Carlo and data (like-sign, sidebands)

- Application on grid:
  - used when the size of the TTree exceeds the TB
  - if training is done with a python package, require an interface with the ALICE environment (treelite, onnx-runtime)

Boosted Decision Trees

- Widely used algorithms in HEP analyses
  - simple but powerful algorithms, less sensitive to hyperparameter tuning wrt neural networks
  - few comparisons done between BDTs an DNNs, no significant differences for the physics analyses
  - two different approaches used: TMVA (ROOT) and python



- Naturally integrated in the ALICE software and on the grid
- Limited selection of models
- Usually BDT based on AdaBoost algorithm

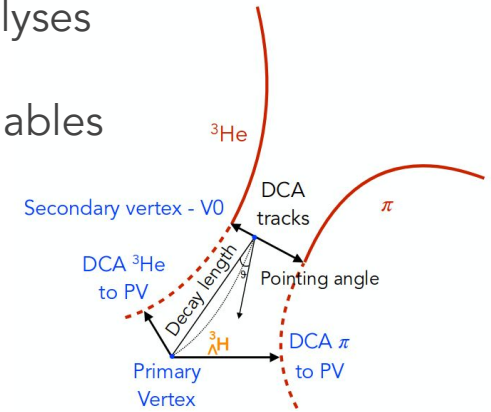
- Widely used outside HEP
  - more models and analysis tools available
- Need interface with ALICE ecosystem
  - treelite, onnx-runtime

Light and heavy flavour analyses with ML

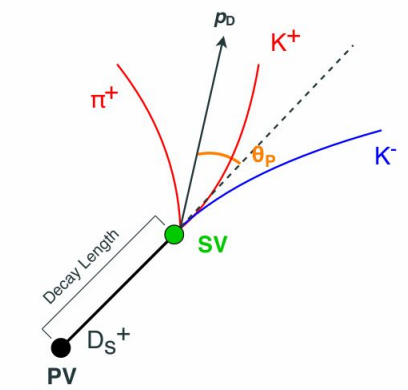
- From the ML point, both heavy and light flavour analyses are very similar
  - BDTs trained employing high-level physical variables (decay length, PID, …)

- ML employed for reconstructing rare signals
  - low S/B ratio, especially in the very high multiplicity environment of Pb-Pb collisions
  - small systems (p-p, p-Pb): no specific triggers -> not negligible statistical uncertainties

- Next slides: examples of physics analyses in LF and HF sectors
  - not-comprehensive overview (e.g. applications to jet analyses not covered), focused on activities with large participation of INFN members



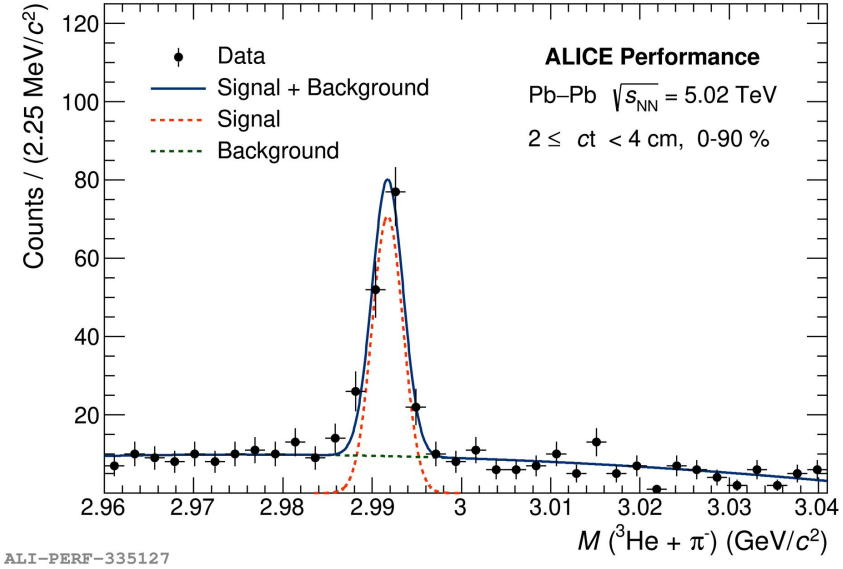
Topology of the hypertriton decay



Topology of the D_s meson decay

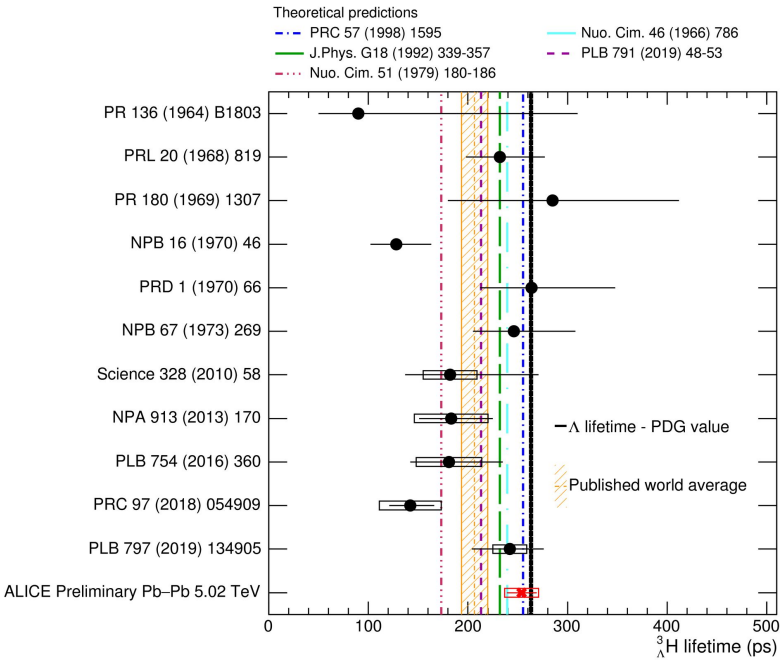
ML Analyses - Hypertriton

- Lightest hypernucleus: p, n, Λ
- Hypertriton lifetime and binding energy are powerful tools for investigating Λ-Nucleon interactions
- Reconstructed via its two-body decay channel
  - ${}^{3}_{\Lambda}\mathrm{H} \rightarrow {}^{3}\mathrm{He} + \pi$
- Pb-Pb @ √s = 5.02 TeV
- Selection using XGBoost BDT
  - signal extraction with high significance in 9 *c*t bins

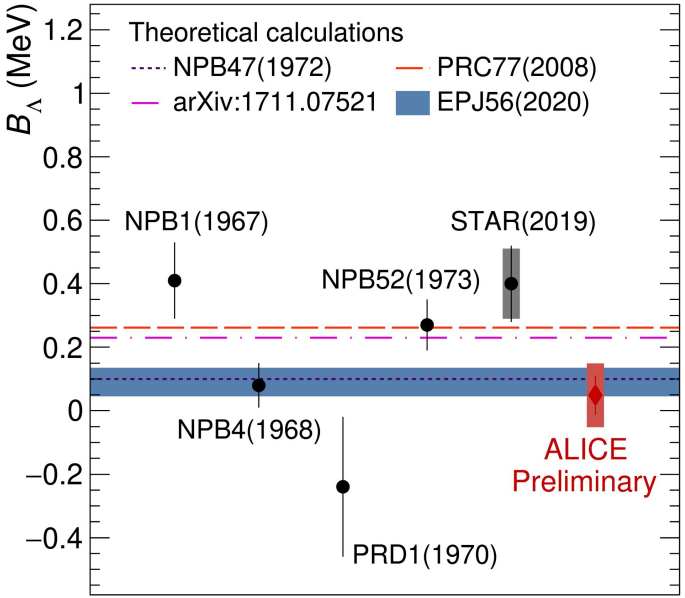


ML Analyses - Hypertriton

- Selection using XGBoost BDT
  - Most precise measurements of the lifetime and binding energy



ALI-DER-358857



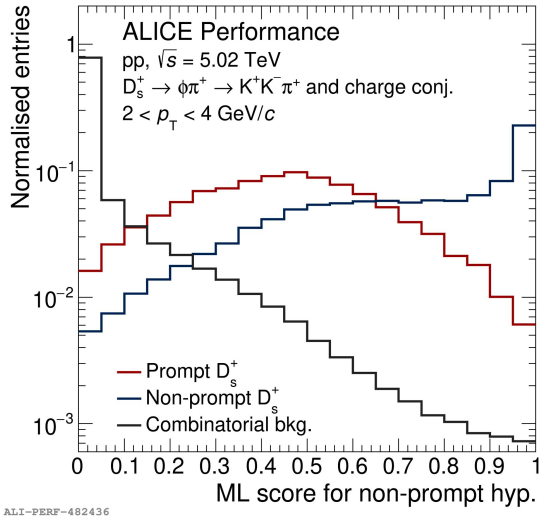
ALI-PREL-486370

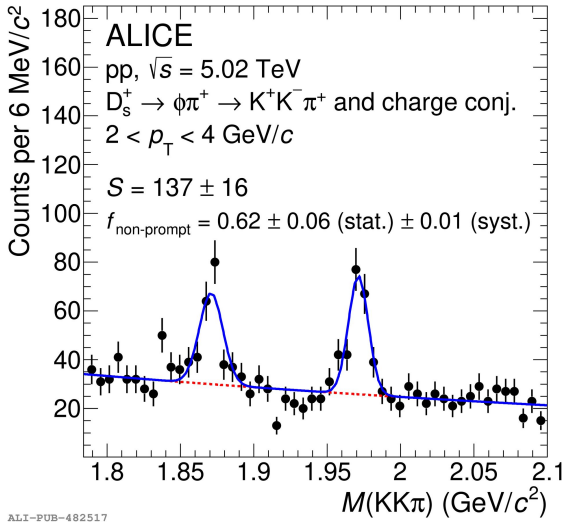
ML Analyses - D mesons

- Measurements of prompt and non-prompt D^0, D^+, D^+_s mesons
  - prompt: from charm-quark hadronisation or excited charm hadron decay
  - non-prompt: from beauty hadron decays

- p-p @ √s = 5.02 TeV arXiv:2102.13601

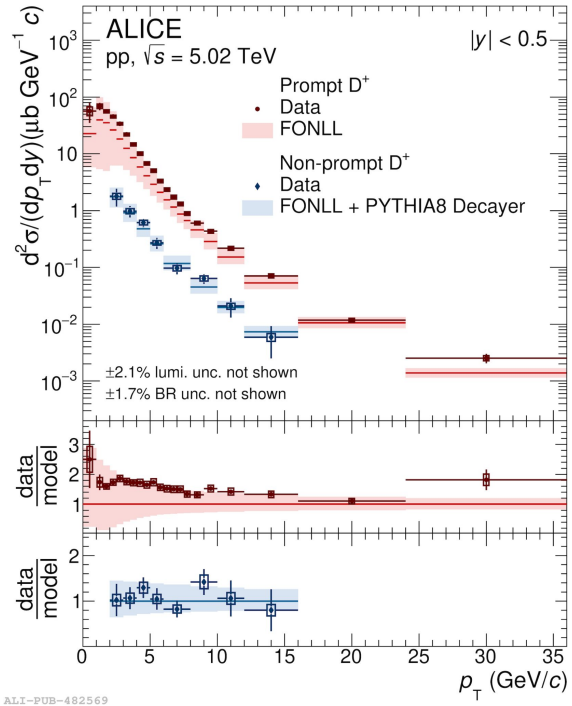
- Multi-classification BDT (XGBoost) to separate prompt, non-prompt and bkg (D^0 with 2-step BDTs, details in backup)



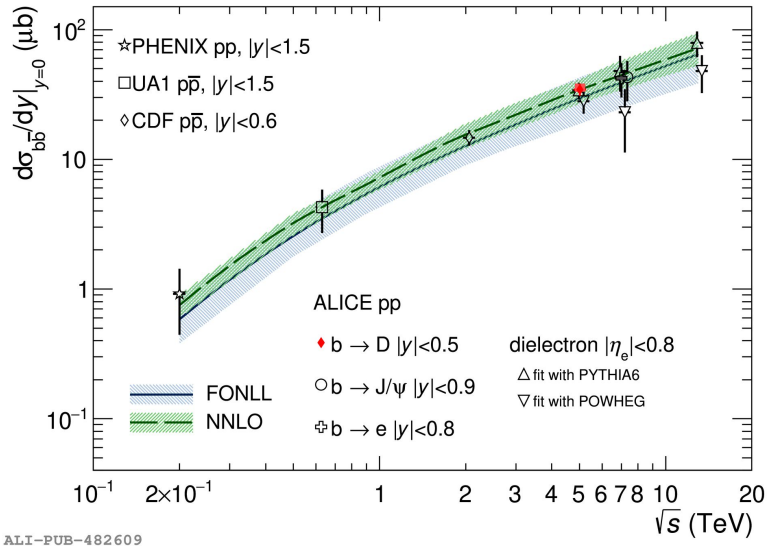


ML Analyses - D mesons

- Non-prompt fraction determined by sampling the raw yields for different BDT outputs
- Most precise charm and beauty production measurements at midrapidity at the LHC



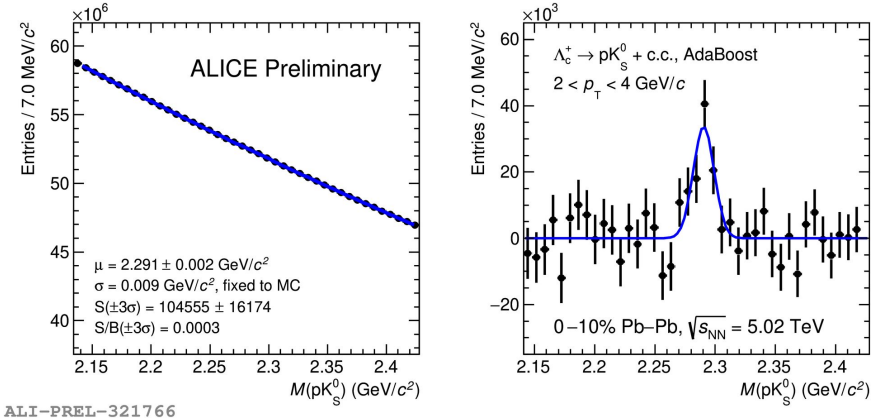
arXiv:2102.13601



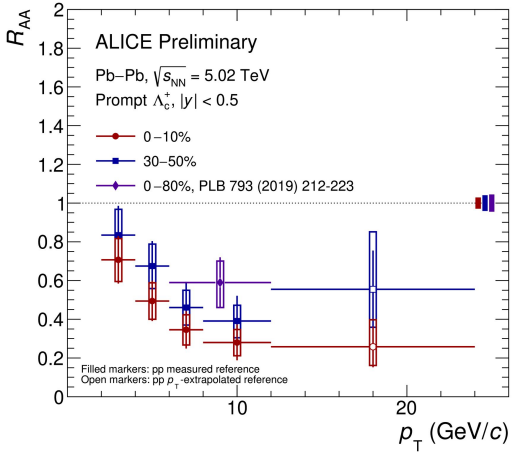
ML Analyses - Λ_c

- Measurements of Λ_c production
  - Reconstruction of the Λ_c -> pK_s^0 decay channel
- Pb-Pb @ $\sqrt{s}$ = 5.02 TeV
- Λ_c selection using TMVA BDT (based on AdaBoost)
- Possible to measure Λ_c R_{AA} in a wide p_T range

$$R_{\mathrm{AA}}\ (p_{\mathrm{T}}) = \frac{1}{\langle N_{coll}^{\mathrm{AA}} \rangle} \frac{\mathrm{d}N_{\mathrm{AA}}/\mathrm{d}p_{\mathrm{T}}}{\mathrm{d}N_{\mathrm{pp}}/\mathrm{d}p_{\mathrm{T}}}$$



ALI-PREL-321766



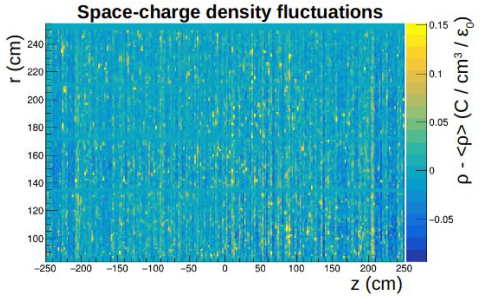
Towards LHC Run 3

TPC space-charge distortion corrections with CNN

- ALICE Time Projection Chamber (TPC):
  - identification of charged particles via reconstruction of the trajectory in the magnetic field and the specific energy loss
- Upgrade to allow for continuous readout up to 50kHz
  - significant amount of positive ions (space charge) from the readout chamber cause large space-charge distortions
    - average corrections used, but fluctuations of several mm are expected (TPC resolution: ~200 μm)

Space-charge density fluctuations

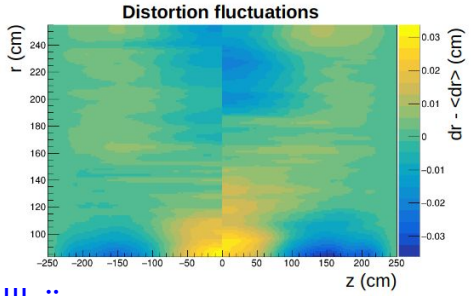
- TPC pad signals integrated over 1 ms intervals (IDCs) as an estimator



Train a ML algorithm for fast data-driven prediction

Space-charge distortion fluctuations

- Solve Poisson equation of the space-charge density to obtain the electrical field deviation
- Solve Langevin equations to obtain the deflection of the ionization electrons along their drift path



See vCHEP2021 talk of Ernst Hellbär

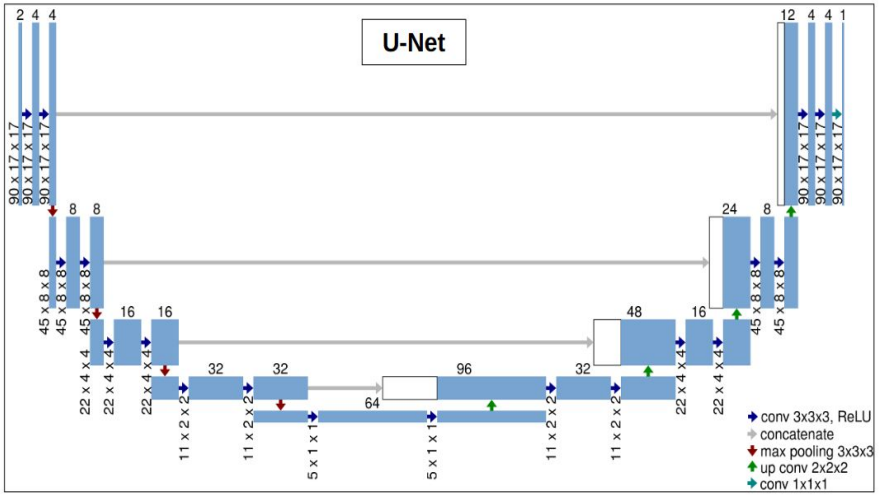
TPC space-charge distortion corrections with CNN

- Convolutional Neural Network (CNN) originally developed for biomedical image used
  - U-Net

- Implementation in python
  - Keras, Tensorflow
  - ROOT and RootInteractive for evaluation of trained models in multi-dimensional TPC phase space

- Input variables
  - space-charge density fluctuations
  - mean space-charge density

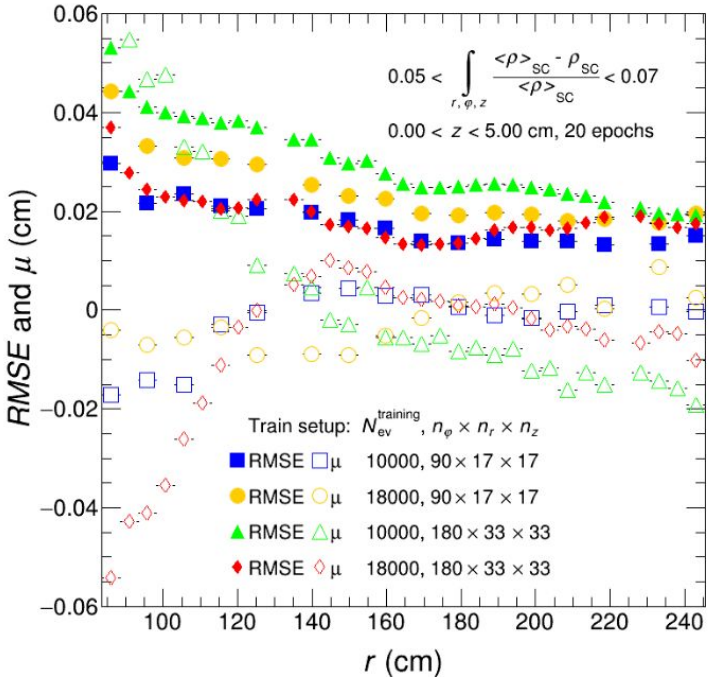
- Output
  - distorsion fluctuations in radial direction



See vCHEP2021 talk of Ernst Hellbär

TPC space-charge distortion corrections with CNN

- Evaluation
  - mean (μ) and RMSE of the difference between predicted and true distortion fluctuations
- Variation of
  - grid size: 90 x 17 x 17, 180 x 33 , x 33
  - number of training samples: 5k, 10k, 18k
- Promising preliminary results:
  - for 10k training events the fluctuations are always lower than the intrinsic TPC resolution
  - systematic dependence of the predicted results on the TPC radius -> investigations on-going
  - aim to extend the correction also to the **φ** and z coordinates



See vCHEP2021 talk of Ernst Hellbär

What's next

- In last years, the use of Machine Learning has spread to many areas of the ALICE experiment: this is a non-reversible process

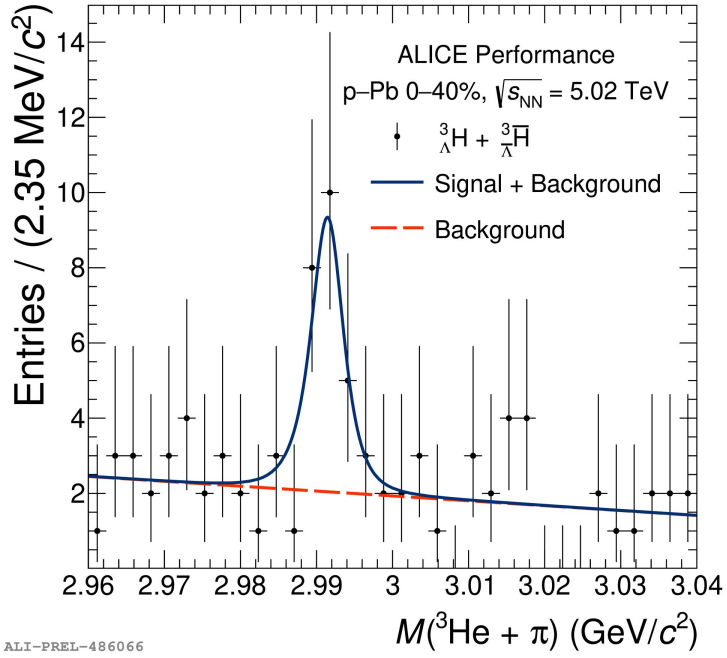
- Run 3 is coming, we have many projects based on ML ongoing
  - online data quality monitoring/data anomaly detection
    - ➢ GANs, BDTs
  - centrality determination in proton-nucleus and nucleus-nucleus collisions
    - ➢ BDTs
  - n dimensional reweighting of the Monte Carlo samples
    - ➢ DNNs
  - centralized framework in the ALICE O^2 software for ML application on Grid
    - ➢ onnx-runtime, treelite
  - training of ML models on large datasets (≳ 1 TB)
    - ➢ new ALICE Analysis Facility (see talk of M.Concas on Friday)
    - ➢ HPC resources available in our local sites
    - ➢ onnx-runtime

Thanks for your attention!

Backup

ML Analyses - Hypertriton

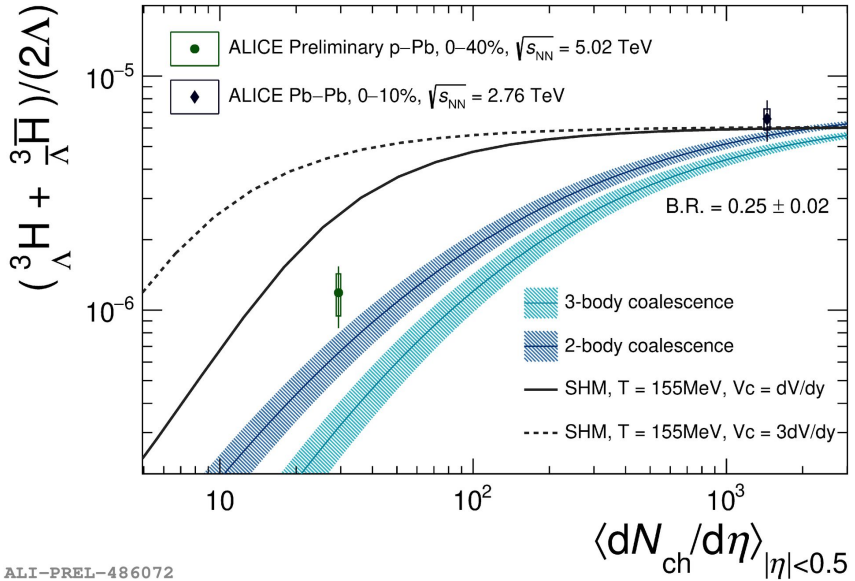
- ${}^{3}_{\Lambda}\mathrm{H}$ production extremely sensitive to the nucleosynthesis model at low charged-particle multiplicity density
- Reconstructed via its two-body decay channel:
  - ${}^{3}_{\Lambda}\mathrm{H} \rightarrow {}^{3}\mathrm{He} + \pi$
- p-Pb @ $\sqrt{s}$ = 5.02 TeV
- Selection using XGBoost BDT



First measurement of hypertriton production in p-A collisions: significance of 4.6**σ**

ML Analyses - Hypertriton

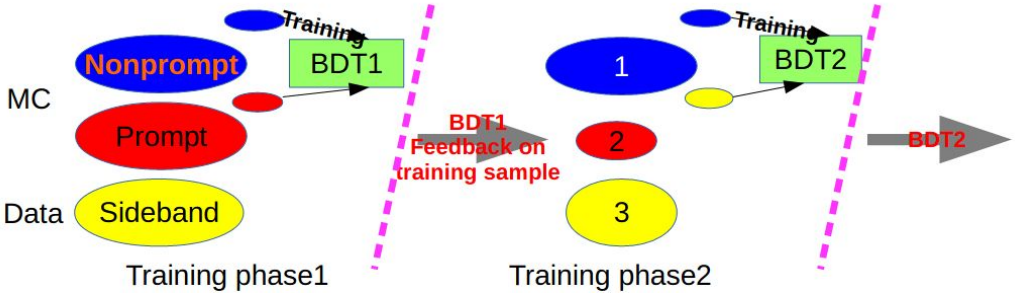
- ${}^{3}_{\Lambda}\mathrm{H}$ production extremely sensitive to the nucleosynthesis model at low charged-particle multiplicity density
- Reconstructed via its two-body decay channel:
  - ${}^{3}_{\Lambda}\mathrm{H} \rightarrow {}^{3}\mathrm{He} + \pi$
- p-Pb @ √s = 5.02 TeV
- Selection using XGBoost BDT

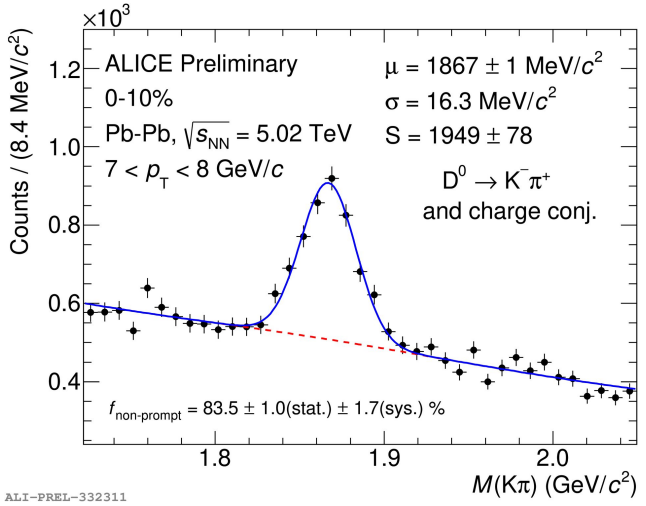


Comparison with available theoretical models: ${}^{3}_{\Lambda}\mathrm{H}$ /**Λ** compatible within 1.2**σ** with 2-body coalescence

ML Analyses - D mesons

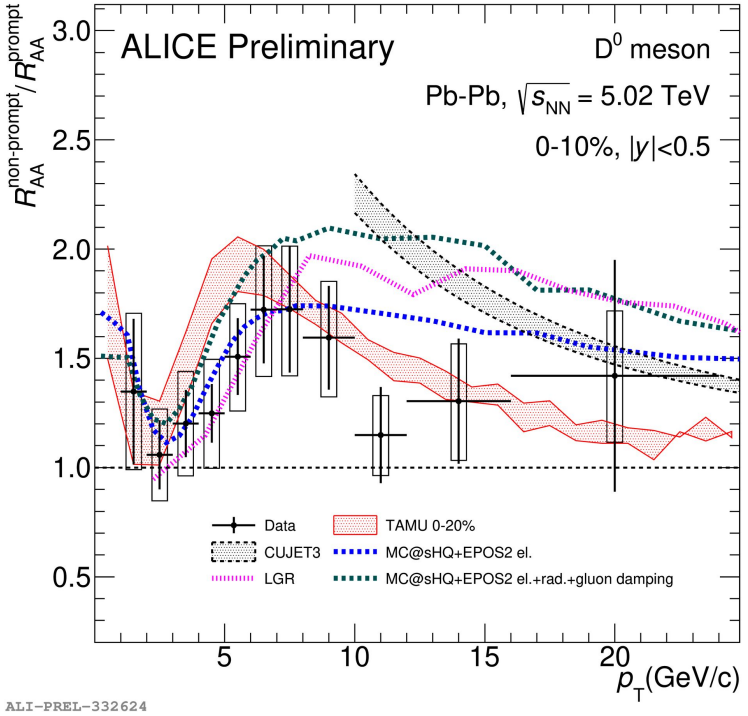
- Measurements of non-prompt D^0 meson
  - prompt: from charm-quark hadronisation or excited charm hadron decay
  - non-prompt: from beauty hadron decays
- Pb-Pb @ $\sqrt{s}$ = 5.02 TeV
- 2-step binary classification:
  - 2 BDTs (XGBoost) to separate prompt, non-prompt and bkg





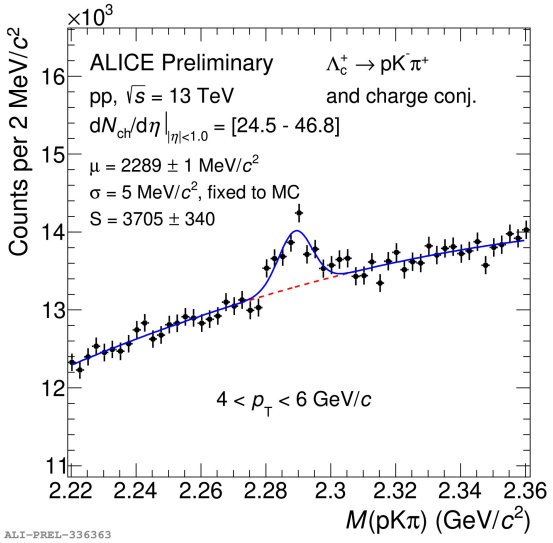
ML Analyses - D mesons

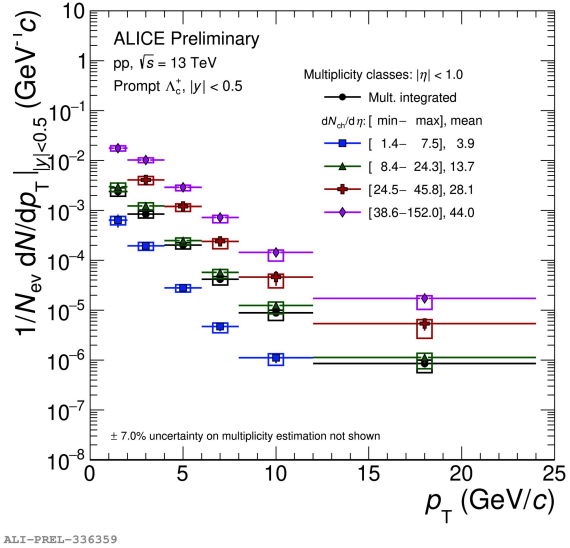
- Measurements of non-prompt D^0 meson
  - prompt: from charm-quark hadronisation or excited charm hadron decay
  - non-prompt: from beauty hadron decays
- Pb-Pb @ $\sqrt{s}$ = 5.02 TeV
- Probe beauty R_{AA} in central Pb-Pb collisions down to very low p_T
- Ratio of R_{AA} of non-prompt and prompt D^0
  - close agreement with most of the available theoretical models



ML Analyses - Λ_c

- Measurements of Λ_c production
  - Reconstruction of the Λ_c -> $pK^-\pi^+$ decay channel
- p-p @ $\sqrt{s}$ = 13 TeV
- Λ_c selection using TMVA BDT (based on AdaBoost)





- Precise production measurement of Λ_c vs average charged-particle multiplicity