Funded by
the European Union
NextGenerationEU

Ministero
dell'Università
e della Ricerca

Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA

Future
Artificial
Intelligence
Research

Charmed-hadron reconstruction in ALICE using AI/ML

M. T. Camerlingo on behalf of ALICE Collaboration

FUTURE AI RESEARCH

INFN

With support by the FAIR (Future Artificial Intelligence Research) Spoke 6 Project, funded by the NextGenerationEU program (Italy)

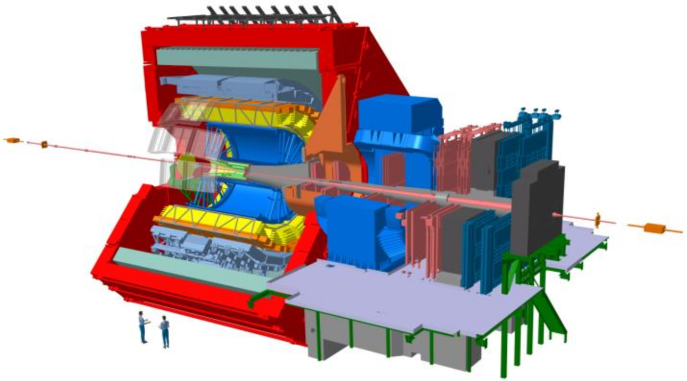
ALICE

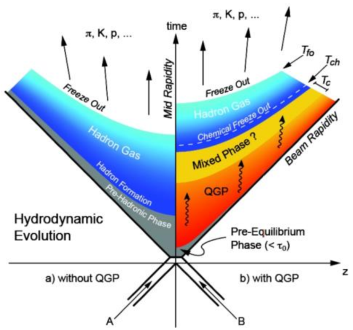
Contents

- **The ALICE experiment**
- **Selection of results from charm hadronization studies**, combination of standard and ML-based analyses of LHC Run 2 ALICE data
- **Charm-hadron reconstruction**
  - Example of a ML-based analysis
- **New ML perspectives for Run 3**
- **NRRP FAIR project on AI/ML thematic**
  - ALICE's Use Case (UC.2)
- **Conclusions**

The ALICE experiment

The ALICE experiment is dedicated to studies of the **q**uark-**g**luon **p**lasma (**QGP**): "deconfined state of matter created under extreme energy densities"





p-p collision-like time-space evolution

A-A collision-like time-space evolution

QGP reproducible in ultra-relativistic **heavy-ion (A-A) collisions** at the CERN LHC;

pp collisions: reference for A-A and tests of pQCD calculations;

p-A collisions to assess cold nuclear matter (CNM) effects in initial and final states;

- ➢ Upgrades for LHC Run 3,
- ➢ Upgrades for LHC Run 4 (ITS3 and FoCal) and Run 5 (ALICE 3)

In this talk: focusing on hadronization studies

Hadron Production in pp collisions

According to QCD factorisation approach :

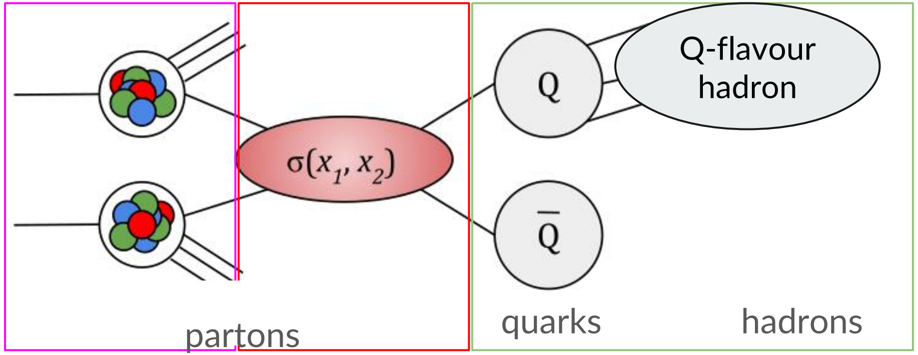
$$\frac{d\sigma^{Hadr}}{dp_{\mathrm{T}}}(p_{\mathrm{T}};\ \mu_F;\ \mu_R) =$$

Parton Distribution Functions
(x_1, x_2, μ_F)

pQCD hard scattering cross section
$\sigma^{Q\bar{Q}}(x_1, x_2, \mu_F, \mu_R)$

Fragmentation function
$(z = p_{Hadr}/p_{Quark}, \mu_F)$

protons



$x_{1\,(2)}$ = momentum fraction of parton 1 (2);
μ_F = factorisation scale;
μ_R= renormalisation scale;
z = momentum ratio of the produced hadron and quark Q

D-meson results in LHC Run 2 pp collisions

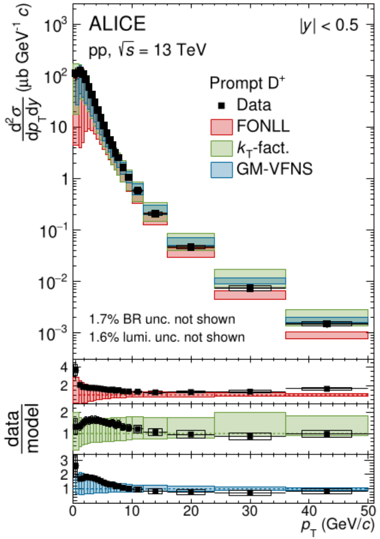
According to QCD factorisation approach :

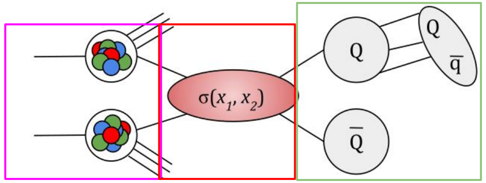
$$\frac{d\sigma^{Hadr}}{dp_{\mathrm{T}}}(p_{\mathrm{T}};\ \mu_F;\ \mu_R) =$$

Parton Distribution Functions (x_1, x_2, μ_F)

pQCD hard scattering cross section $\sigma^{Q\bar{Q}}(x_1, x_2, \mu_F, \mu_R)$

Fragmentation function $(z = p_{\mathrm{Hadr}}/p_{\mathrm{Quark}}, \mu_F)$

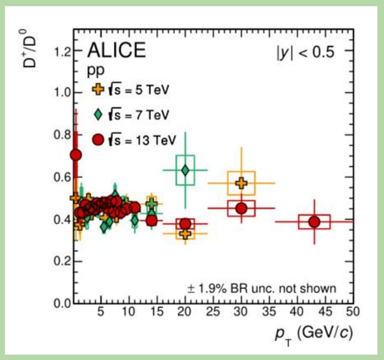




1. The D-meson production cross sections **are described within uncertainties** by **pQCD calculations** assuming **universal fragmentation functions (FF) evaluated from e⁺e⁻, ep** collision measurements.

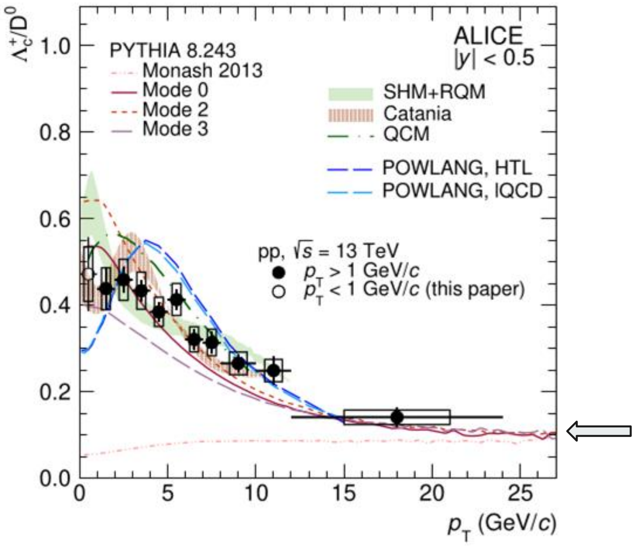
2. The measured cross section ratios of different D-meson species are **sensitive to FF** and help to investigate **hadronization mechanisms.**

FF are independent on the collision energy.



Baryon results in LHC Run 2 pp collisions

Looking at ALICE Λ_c^+/D^0 measurements (down to p_T ~ 0):

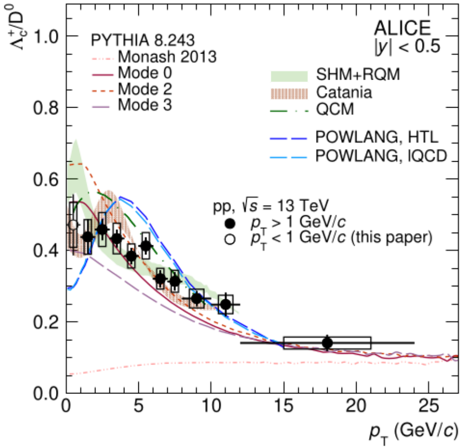


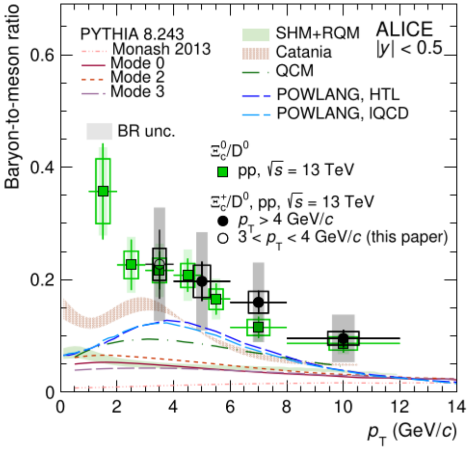
- ➢ Pythia 8 Monash model with standard Lund fragmentation underestimates pp results.
- ➢ These results support the **scenario of charm-quark hadronization in pp collisions via other mechanisms than those in e^+e^- collisions (in-vacuum fragmentation)**
- ➢ Models implementing an enhanced baryon production with different mechanisms (as color reconnection beyond leading color approximation, statistical hadronization + augmented set of baryons as predicted by relativistic quark model, quark (re)combination) can describe Λ_c^+/D^0 results at pp collisions.

Baryon results in LHC Run 2 pp collisions

Looking at a different baryon species (including charm and strange flavours), as $\boxed{\Xi_c^0}$ and $\boxed{\Xi_c^+}$, can the previous models describe Ξ_c/D^0 ?

- ➢ Model with standard Lund fragmentation (Pythia 8 Monash) underestimates pp results
- ➢ These results support the **scenario of charm-quark hadronization in pp collisions at the LHC via mechanisms other than those in e$^+$e$^-$ collisions.**
- ➢ Models implementing an enhanced baryon production with different mechanisms can describe Λ_c^+/D^0 results at pp collisions.
- ➢ These models can not describe Ξ_c/D^0.

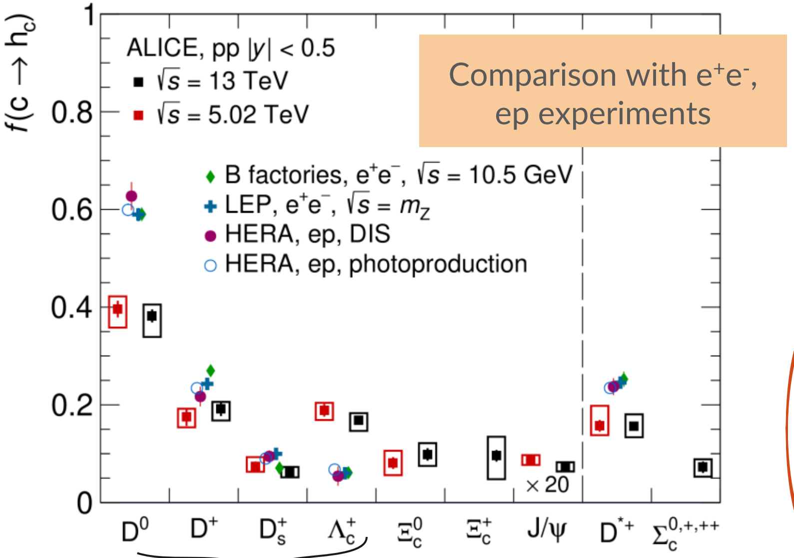




Heavy-flavour baryon production not fully understood, important to perform precise measurements down to p_T ~ 0

Charm-quark fragmentation fractions

ALICE measured the p_T-integrated cross section of all the ground-state charm mesons and baryons → corresponding production fraction f(c→ h_c), i.e. frequency of a c-quark producing a given charmed-hadron h_c



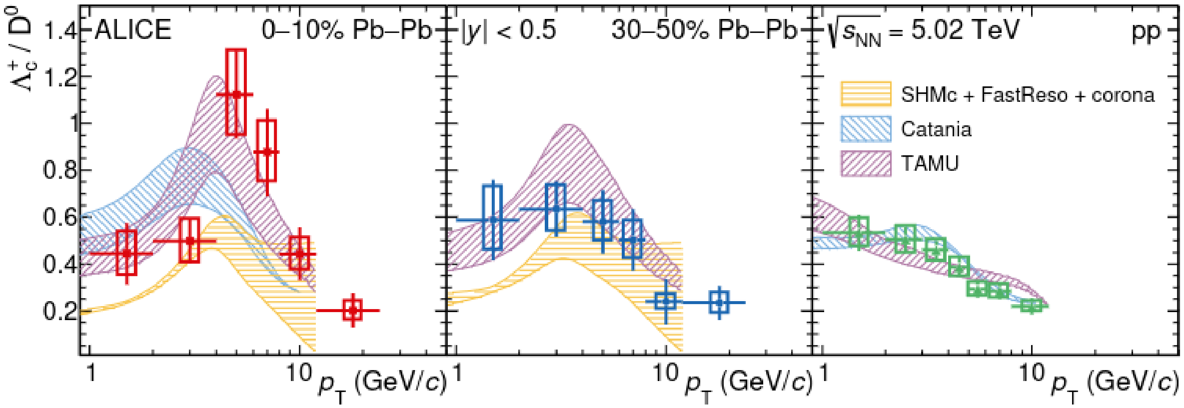
Significant enhanced Λ_c^+ baryon production in pp collisions w.r.t. e^+e^- collisions.

$f(c \to h_c)$	pp, $\sqrt{s} = 5.02$ TeV (%)	pp, $\sqrt{s} = 13$ TeV (%)
D^0	39.6 ± 1.7 (stat.) $^{+2.6}_{-3.8}$ (syst.)	38.2 ± 1.3 (stat.) $^{+2.3}_{-4.3}$ (syst.)
D^+	17.5 ± 1.8 (stat.) $^{+1.7}_{-2.1}$ (syst.)	19.1 ± 1.4 (stat.) $^{+1.5}_{-2.3}$ (syst.)
D_s^+	7.4 ± 1.0 (stat.) $^{+1.9}_{-1.1}$ (syst.)	6.1 ± 0.5 (stat.) $^{+1.2}_{-0.9}$ (syst.)
Λ_c^+	18.9 ± 1.3 (stat.) $^{+1.5}_{-2.0}$ (syst.)	16.8 ± 0.8 (stat.) $^{+1.5}_{-2.1}$ (syst.)
Ξ_c^0	8.1 ± 1.2 (stat.) $^{+2.5}_{-2.5}$ (syst.)	9.9 ± 1.3 (stat.) $^{+2.3}_{-2.4}$ (syst.)
Ξ_c^+	Assumed to be the same as Ξ_c^0	9.6 ± 1.2 (stat.) $^{+3.9}_{-4.8}$ (syst.)
J/ψ	0.44 ± 0.03 (stat.) $^{+0.04}_{-0.06}$ (syst.)	0.37 ± 0.02 (stat.) $^{+0.04}_{-0.05}$ (syst.)
D^{*+}	15.7 ± 1.2 (stat.) $^{+4.1}_{-1.9}$ (syst.)	15.6 ± 0.7 (stat.) $^{+2.5}_{-2.2}$ (syst.)
$\Sigma_c^{0,+,++}$	—	7.2 ± 1.2 (stat.) $^{+1.6}_{-1.9}$ (syst.)

For Ξ_c^0 and Ξ_c^+ baryons, ALICE could not measure the low-p_T range during Run 2. **Ongoing Run 3 ALICE campaign to acquire data for more precise and new measurements down to p_T ~ 0.**

Λ_c^+/D^0 ratio in Pb-Pb

The p $_T$-differential Λ_c^+/D^0 ratios increase from pp to central Pb–Pb collisions for $4 < p_T < 8$ GeV/c with a significance of 3.7σ, while the pp and Pb-Pb p_T-integrated ratios are compatible within 1σ.



The measurements are described by **theoretical calculations that include both coalescence and fragmentation processes when describing the hadronization of heavy flavours in the QGP.**

➢ Also for Pb-Pb collisions, **neither** the measurement precision can point out the closest theoretical model **nor** a single model can describe the data in the full p_T range and baryon species.

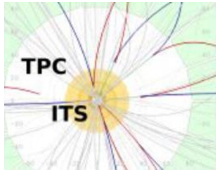
Messages to convey up to now

- Previous slides are just a short review of measurements for hadronization in charm sector;
- The ALICE Heavy Flavour (HF) includes many further studies:
    - QGP characterisation ,
    - QCD studies,
    - Studies of small system collectivity

  including measurements in the beauty sector.
- Recent review of ALICE studies and results:
    - The ALICE experiment: a journey through QCD
- Current hadronization models do not entirely describe the observations, **more precise** observation could help to discriminate among the available models. **Machine-learning-based analyses** (plus higher stat, and detector upgrades) can “boost” this search.

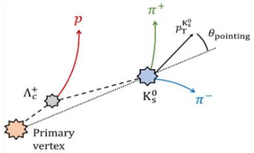
Charm-hadron reconstruction (in a nutshell)

1. Track selections



- ITS-TPC matched tracks
  → selection of primaries (ALICE-PUBLIC-2017-005)

2. Secondary-vertex reconstruction

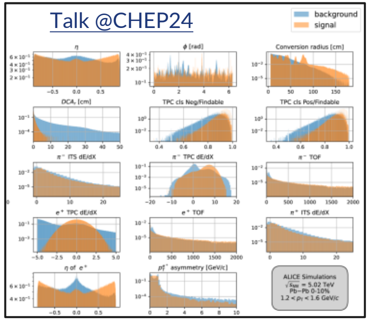


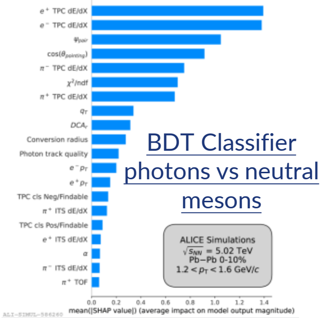
- Impact parameter resolution to primary vertex ~75 µm @ p_T = 1 GeV/c
- Evaluation of topological features
  → intrinsic displacement (i.e. large decay length $c\tau$)

3. Particle identification (PID) and topological selections

- **Separate signal from background candidates**
- Extract signal yields using an invariant-mass analysis

RUN 2 ML - TPC response calibration
NN for energy-loss (dE/dx) calibration

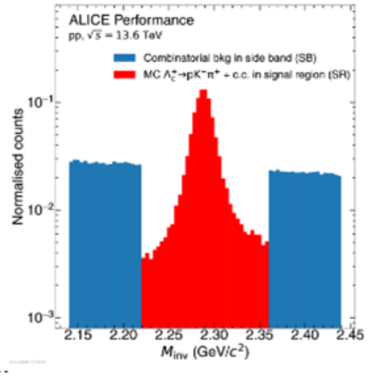




RUN 2 ML - BDT classifier
exploiting/enhancing the discrimination from the **signal decay-vertex topology** and **PID** based on a not-ML Bayesian approach (CERN-EP-2016-023).

Signal-vs-Background classification using BDT

Binary classifier workflow



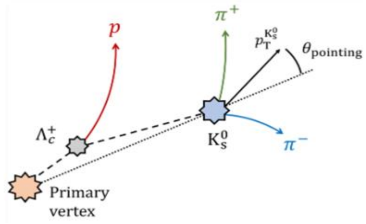
1. Define the Side Bands (SB) and Signal Region (SR) of invariant mass
   - CLASS 0: background candidates, from ALICE data in SB
   - CLASS 1: Signal candidates obtained via ALICE Monte Carlo simulations

Data-preparation
- Search variables independent on M_{inv}
- CLASS 1 vs CLASS 0 to find the most discriminant features
- Mixing the data of the two classes to create the labelled input dataset

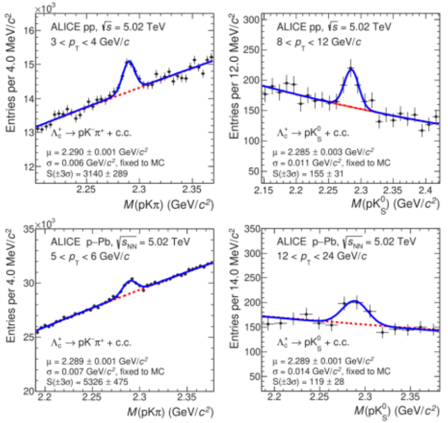
2. BDT training and Test
   - Monitor figures of merits
   - Calculate score threshold @MaxSignificance

3. Cut on the predicted BDT score of the data in SR

Λ^+_c secondary vertex not reconstructed → pK^0_s propagation to primary vertex not precise enough



In p-Pb collisions analysis, ML-based analysis was adopted, reaching compatible M_{inv} peak widths wrt standard analysis in pp collisions PhysRevC.104.054905.



Already-adopted more complex models: for D-meson, a BDT multiclassifier to also discriminate the third class of non-prompt candidates CERN-EP-2021-034

ML in the ALICE charmed-hadron reconstruction for Run 3

- **Signal-vs-Background classification**
- **TPC response calibration**

NEW PID in RUN 3 (CHEP2023, *JINST* 19 C07013)
NN to combine tracking and PID info from different detectors;
PID in ITS2 using BDT regression.

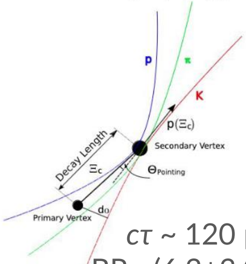
worth to mention for HF sector:

NEW Heavy flavour hadron trigger (for B-flavour)
BDT to perform offline trigger selection on displaced decay-vertex topologies

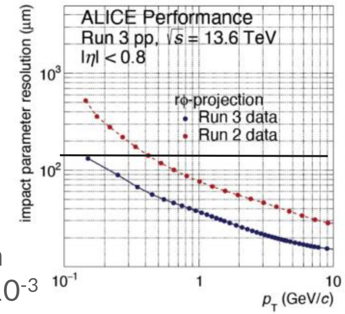
Larger statistics, ALICE detector upgrades → access to other charmed-baryon decay channels like

$$\Xi_c^+ \rightarrow pK^-\pi^+ + c.c.$$

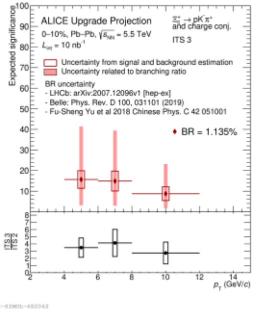
decay topology



$c\tau$ ~ 120 μm
BR ~(6.2±3.0)10^{-3}



Pb-Pb perspectives



Huge combinatorial bkg and small BR: ML-based analysis improves significance and S/B w.r.t. rectangular cut approach, allowing to perform the measurement.

FAIR project

FAIR foundation is an extended partnership based on the Hub & Spoke model.

- Its research topic is AI in different humanistic and scientific fields.
- INFN joined Spoke 6 of FAIR project to develop AI/ML infrastructure solutions (back-up) and to propose HEP Use Cases (UCs) synergic to JLab and CERN LHC experiments:

UC1: “AI-supported algorithms in Streaming ReadOut for HEP data acquisition systems” F. Rossi’s talk

UC2:”Machine Learning-based reconstruction of (multi-)charm baryons in ALICE”

FAIR-ALICE UC2

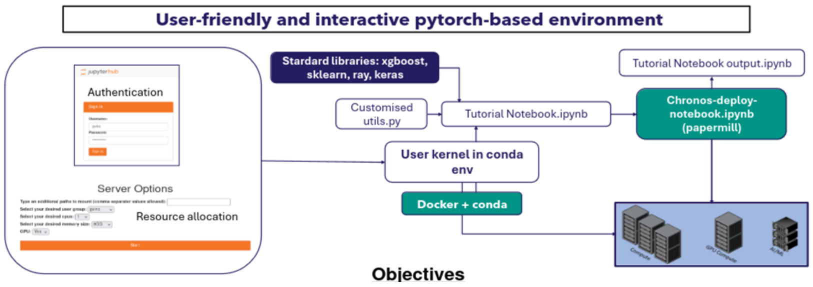
Objectives:

- Technological: to provide a UC for testing the infrastructure in upgrade
    - Wide grid-searches in the hyperparameter optimisation, and cross-validation
    - Comparing BDT with more complex ML models (common to both objectives)
- HEP Physics: to contribute of $\Xi^{+}_{c} \rightarrow pK^{-}\pi^{+}$ + c.c. and $\Xi^{+}_{c} \rightarrow pK^{-}\pi^{+}/\Lambda^{+}_{c} \rightarrow pK^{-}\pi^{+}$ measurements using Run 3 data

UC2 version/notebook 1: BDT binary classifier

UC2 version/notebook 2: anomaly detection using autoencoder (AE)

Actual open-source AI/ML environment (INFN, ReCas Bari)



Even with the upgrades (kubernetes, INFN cloud …), the environment will be very **user-friendly**.

Inside INFN Cloud project, a PaaS dedicated to AI/ML is under development.

Anomaly detection using autoencoder

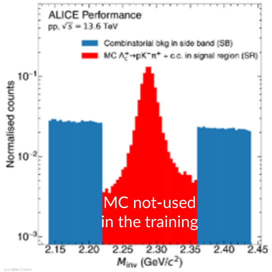
ALICE Interest

Anomaly detection workflow: “signal event is the anomaly”

used a preparatory dataset

Motivations:

- Comparative study with BDT classifier;
- exploring data-driven method



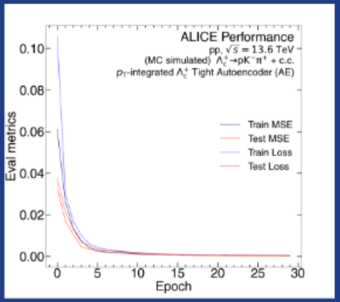
1. Define the Side Bands (SB) and Signal Region (SR) of invariant mass

Data-preparation (only **background candidates, from ALICE data in SB**):

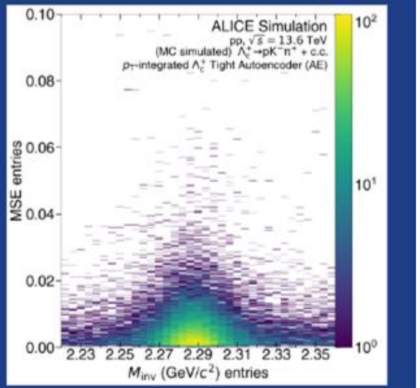
- Search independent variables on M_{inv}
- Preprocessing transformation

2. AE training and test: to represent background from the variables that are independent on the invariant mass

Λ_c^+ - TIGHT (AE learning curves of ex-class 0)



Λ_c^+ - TIGHT (AE prediction of ex-class 1 (MC))



3. Calculate the MSE between the reconstructed the original features for
   - Test data in SB (i.e. comb bkg) - to define a sata-driven MSE threshold
   - if MC available, MC signal in SR to cross-validate the AE and to calculate the MSE @MaxSignificance

4. MSE Cut application on data in SR

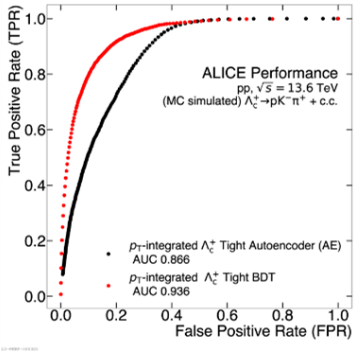
Model comparison w BDT

used a preparatory dataset

Model performance:

Test ROC curves

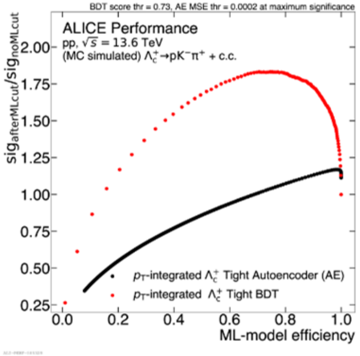
- BDT shows a larger Area Under Curve (AUC) than AE



Test significance (sig)

$$\mathrm{sig}_{\mathrm{afterMLcut}} = \frac{\mathrm{Signal}}{\sqrt{\mathrm{Signal+Background}}}$$

after a cut on the min BDT class 1 score (or AE MSE)



Computing performance:

used a preparatory dataset

As expected, BDT has faster times but heavy resident memory size (RES) usage than AE

BDT	Wall time (s)	RES (GiB)	Class 0 train/val stat	Class 1 train/val stat	# of trainings
Hyp opt	(52 ± 1)*10	56	4*class 1 stat	32k/8k	100
Cross-val	(9 ± 1)*10	5.5	200k/-		5 folds (x 20 rounds)
Training	8.0 ± 0.5	3.9	207k/52k	32k/8k	1
Prediction	0.060 ± 0.005	3.9	-/52k	-/8k	-

AE	Wall time (s)	RES (GiB)	Class 0 train/val stat	# of training
Hyp opt	(115 ± 6)*10	57.5	150k/10k	100
Training	40 ± 2	1.6	207k/30k	1
Prediction	1.5±0.5	1.6	-22k	-
MC Prediction	2.0±0.5	1.6	-/40k	-

Computing performance during interactive execution of the Use Case Jupyter Notebooks in which 16 CPUs and 0.1 partitioned shared GPU were allocated (standard privileges)

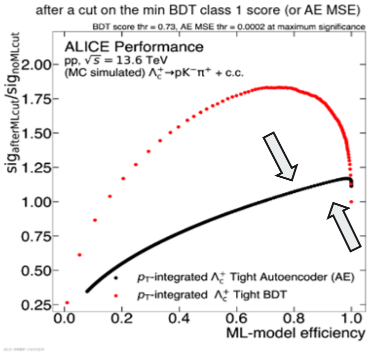
- Model architectures in back-up
- BDT also shows a larger plateau @max significance
- Significance after the cut on BDT score improves more but it corresponds to a lower efficiency
- AE has a better efficiency after the MSE cut

Autoencoder performance by including signal in its training

Study in a very preliminary phase/under evaluation

Unsupervised AE training vs signal contamination

Repeating the training&test



- to find a correlation between the curve increase and the contamination fraction;
- to investigate a possible procedure to estimate sys unc.

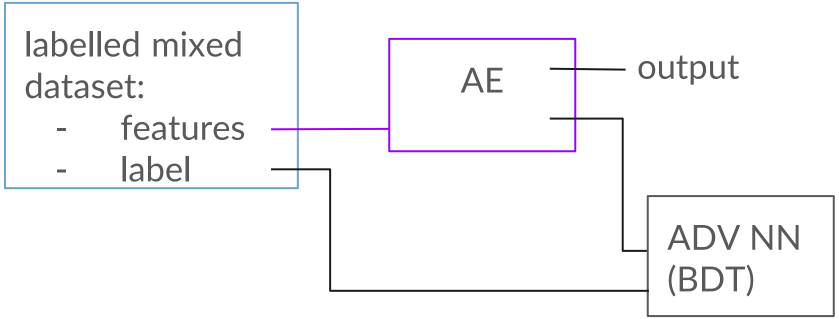
Comparison with a semi-supervised model (ref):

1st training: only AE → $Loss_{AE}$

2nd training: only ADV NN → $Loss_{adv}$

3rd training: AE again → $Loss_{TOT}$ = $Loss_{AE}$ - α$Loss_{adv}$

Adversial autoencoder



(other than vs BDT)

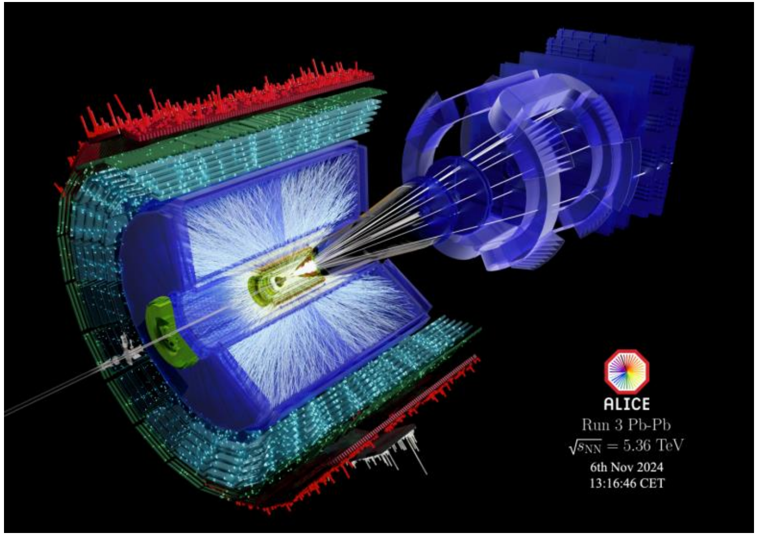
(Not-exhaustive) List of ALICE AI/ML activities

- **Signal-vs-Background classification**
  already mentioned **hype4ml** package for BDT
- **Jet pT reconstruction**
  correction for the background from the underlying events using a shallow NN
- **Heavy flavour hadron trigger**
  BDT to perform offline trigger selection on displaced decay-vertex topologies
- **TPC response calibration**
  NN for energy-loss (dE/dx) calibration
- **PID**
  NN to combine tracking and PID info from different detectors
  PID in ITS2 using BDT regression
- **Fast simulation for ZDC**
  using GANs and VAEs
- **ML for quality control/assurance (LLM)**
  alert AI-system in data-taking
- **flavour jet-tagging**

Conclusions

Run 2 results showed an enhancement of heavy-flavour baryon-to-meson ratios w.r.t. e^+e^-, that the actual theoretical models can not reproduce completely → Further investigations and a better precision are required to shed further light on heavy-quark hadronisation.

Detector upgrades for **Run 3 and 4,** the increase of statistics, and a **more extensive use of ML techniques** lead to **improved results and access to new particles and observables.**



Thanks for the attention

Back-up

Model architectures in slide 17

BDT

Hyperparameter	Value	Hyperparameter	Value
input features	12	gradient boosting library	XGBoost
tree_method	hist	device	cuda
objective	binary logistic	eval_metric	auc (rmse, error)
colsample_bytree	0.952	subsample	0.941
max_depth	7	min_child_weight	3
η	0.168	num_round	100 (default in fit)
α	1 (default)	γ	0 (default)

AE

Layer (type)	Output Shape	Param #	Activation
In Encoded (InputLayer)	[(None, 13)]	0	regularizers.L2(lr=0.0687)
Encoded (Dense)	(None, 239)	3346	ReLU
Latent (Dense)	(None, 20)	4800	ReLU
Decoded (Dense)	(None, 239)	5019	ReLU
Out decoded (Dense)	(None,13)	3120	sigmoid
Hyperparameters	**Optimizer**	**Batch**	**epochs**
	Adam	1996	30

INFN DataCloud-like deployments

INFN is responsible to install a complete stand-alone server GPU NVIDIA in ReCas datacenter, bought by FAIR NRRP.

The proposed configuration is inspired by “INFN DataCloud”-like deployments, concerning:

- **Identity and Access Management** → INDIGO IAM
- **Workload offloading for interactive analyses** → Jupyter and INFN Cloud PaaS (with Dask library integrated for parallel computing)
- **Environment preparation for data science/machine learning** → CernVM Filesystem (CVMFS) operated via a SQUID hierarchy, and Docker/Conda combo (for small projects)
- **Federation Data Infrastructure and management** → open-source Rucio project and the File Transfer component is based on the open-source FTS software (for data orchestration both by CERN), Cloud storage S3 based to the WebDAV server, mainly based on the open-source StoRM WebDAV software (for the storage edge service), file-level caching service under evaluation
- **Cloudify any resource** → offloading k8s workflows