Measurement of non-prompt Λ_c^+ production in pp collisions at $\sqrt{s} = 13$ TeV with ALICE

Daniel Battistini

Università degli studi di Torino

IMPRS Recruiting Workshop

15 Nov 2021

Heavy Flavour production

Heavy Flavour (HF) hadrons = hadrons with beauty or charm quarks

- *c*- and *b*-quarks: produced in high-Q^2 partonic processes → perturbative approach is possible;
- In proton-proton (pp) collisions → test of perturbative Quantum Chromodynamics (pQCD) calculations and the Factorisation Theorem:

Factorisation theorem: the cross section of a hadron can be computed as the following convolution:

Parton Distribution Functions (non perturbative)

Fragmentation Function (non perturbative)

$$\sigma(pp \to H_Q + X) = \sum_{i,j=q,\bar{q},g} f(x_i, Q^2) \otimes f(x_j, Q^2) \otimes \sigma(ij \to Q\bar{Q}) \otimes D(z_Q, Q^2).$$

Partonic cross section (perturbative)

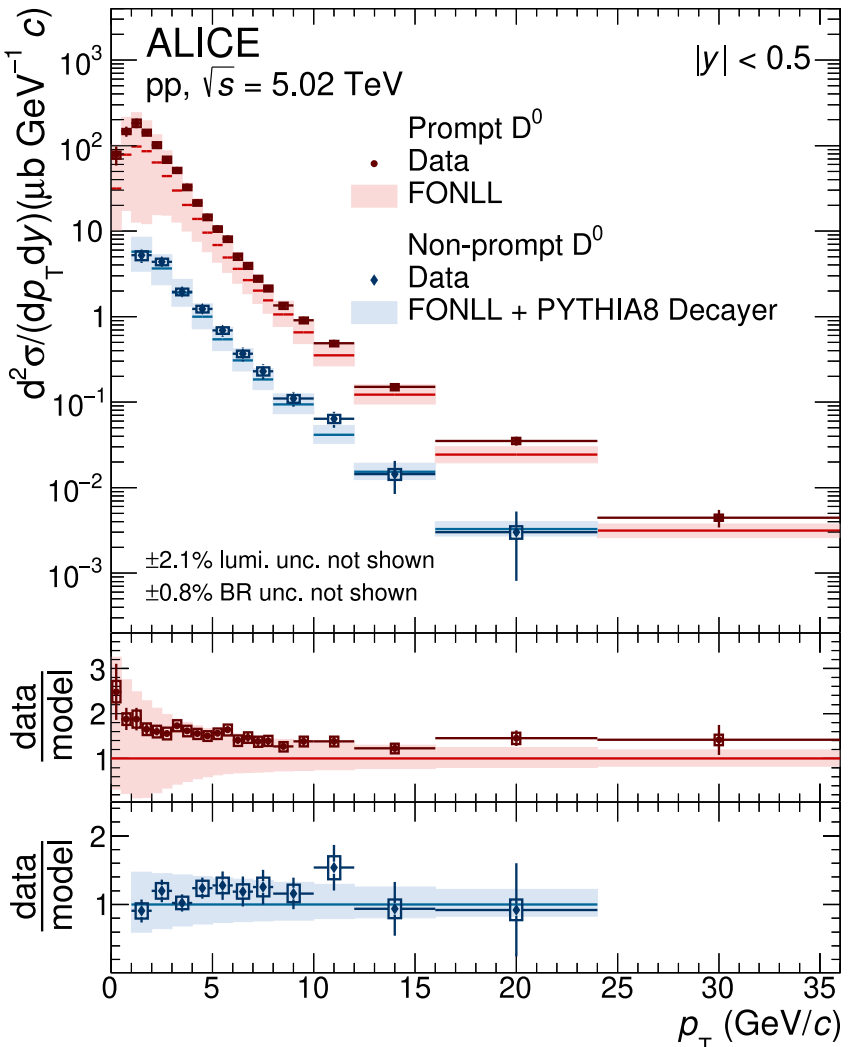
Prompt and non-prompt components

Two contributions to **charm hadron production**:

1) Prompt:
   from the hadronisation of the charm quark or decay of excited charm-hadron states.

2) Non-prompt (NP):
   from the decay of a beauty hadron.

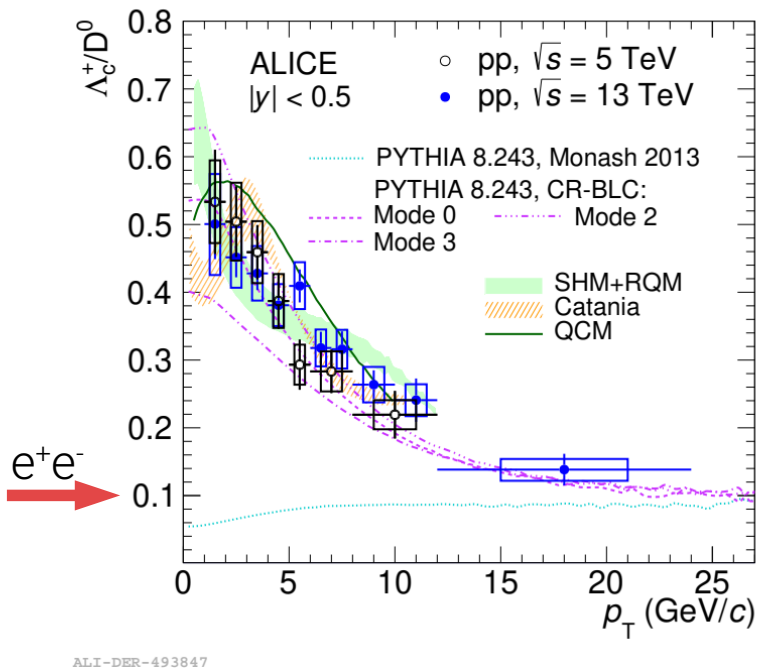
- Measured D meson cross section in good agreement with pQCD (FONLL).



Hadronisation of charm and beauty quarks

Hadronisation of charm

Prompt



Recent measurements: the hadronisation of the charm quark (→ Λ^+_c/D^0) is not independent of the collision system! Open point...

Hadronisation of beauty… ?

Exclusive H_b measurements

Difficult because of:

→ Small beauty cross section;

→ Small branching ratios;

Non-prompt (This work)

Measure the non-prompt Λ^+_c production:

→ access to the physics of the b-quark;

→ in nature: 5÷10%, depending on p_T;

→ no measurements existed;

→ decay channel: $\Lambda^+_c \rightarrow pK^0_s \rightarrow p\pi^+\pi^-$;

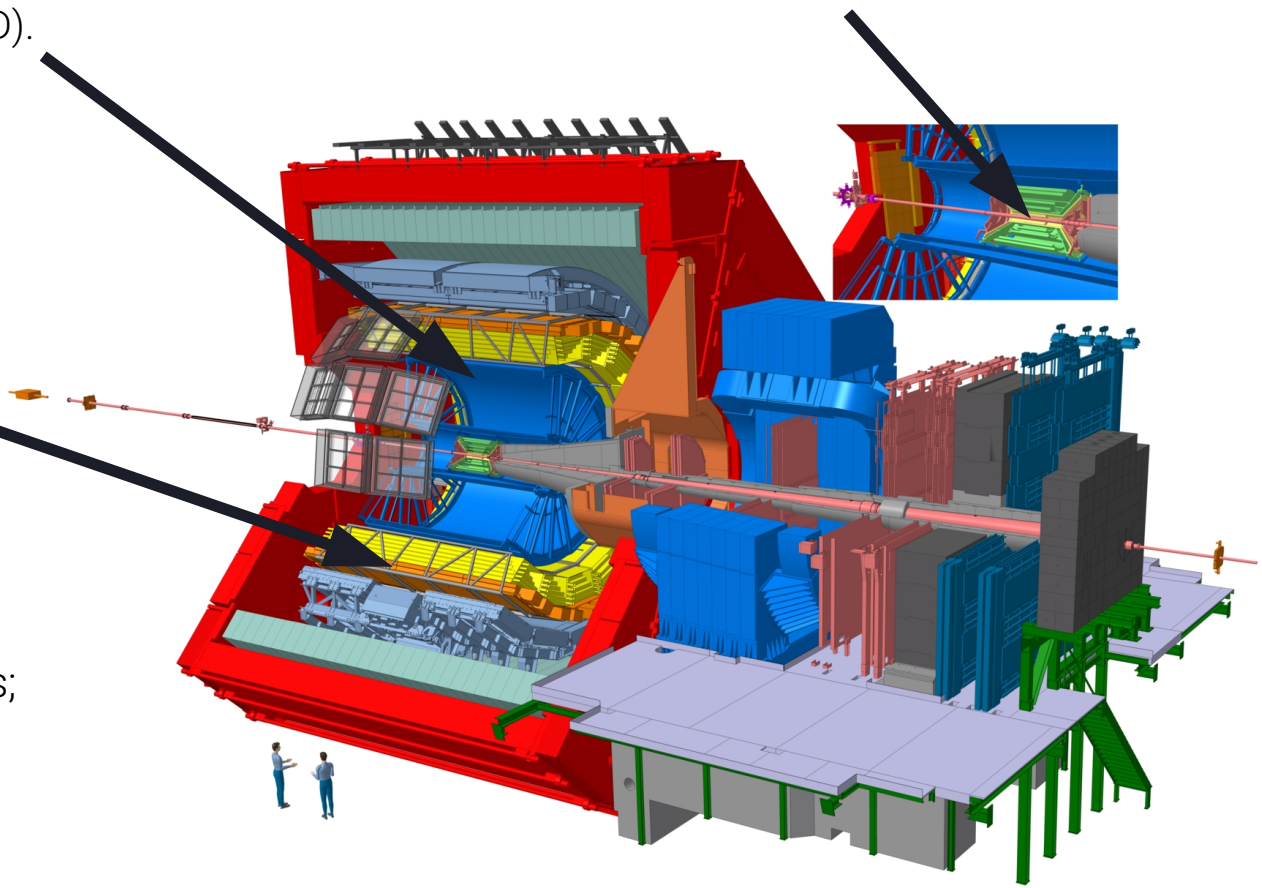
→ study non-prompt Λ^+_c/ non-prompt D^0.

The ALICE detector

TPC (Time Projection Chamber):
- track reconstruction;
- particle identification (PID).

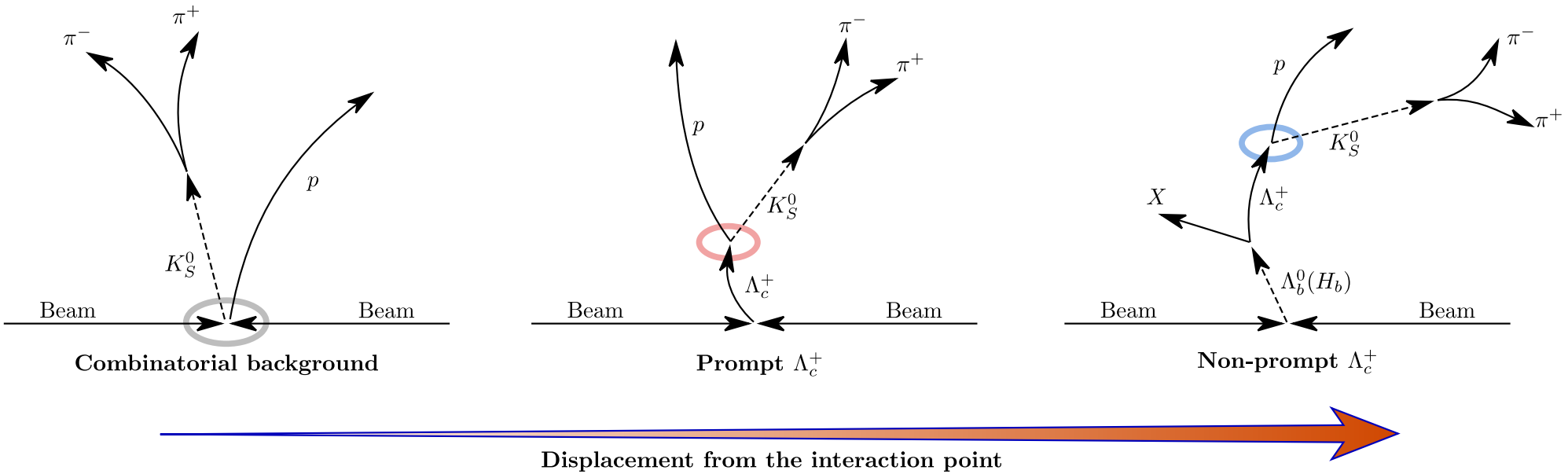
ITS (Inner Tracking System):
- track and vertex reconstruction

TOF (Time Of Flight):
- particle identification.



Data collected between 2016 and 2018:
- center-of-mass energy √s = 13 TeV;
- 1.84×10^9 Minimum Bias pp collisions;
- $L_{int} = 32\ nb^{-1}$;

Decay topologies: prompt, non-prompt and background



Background mostly from protons and K^0_s coming from the interaction point;

Mean proper decay length:

→ $c\tau(\Lambda^+_c)$ = 60 μm;

→ $c\tau(\Lambda^0_b)$ = 440 μm.

To separate the three contributions:

- exploit the decay topology;
- use Machine Learning: multi-class classification algorithm based on Boosted Decision Trees (BDTs)

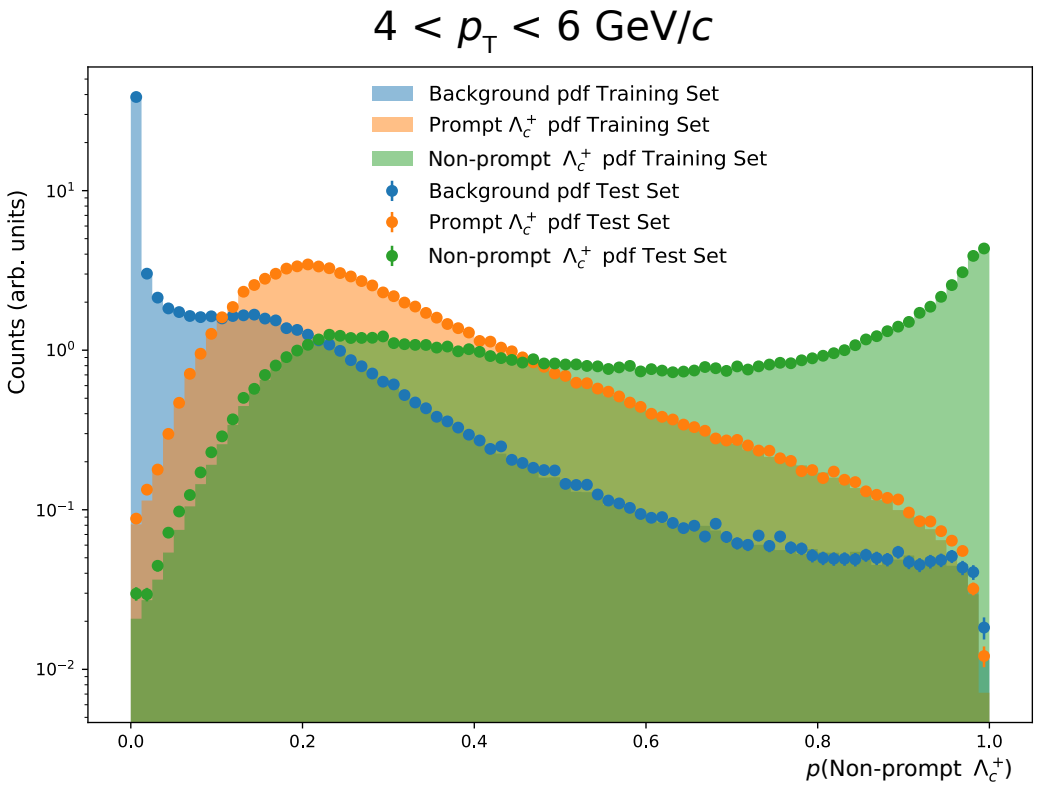
Machine learning

Datasets:
- **Background**: from the data;
- **Signal**: from Monte Carlo simulations.

Training variables:
- decay-vertex topology;
- PID of the decay products.

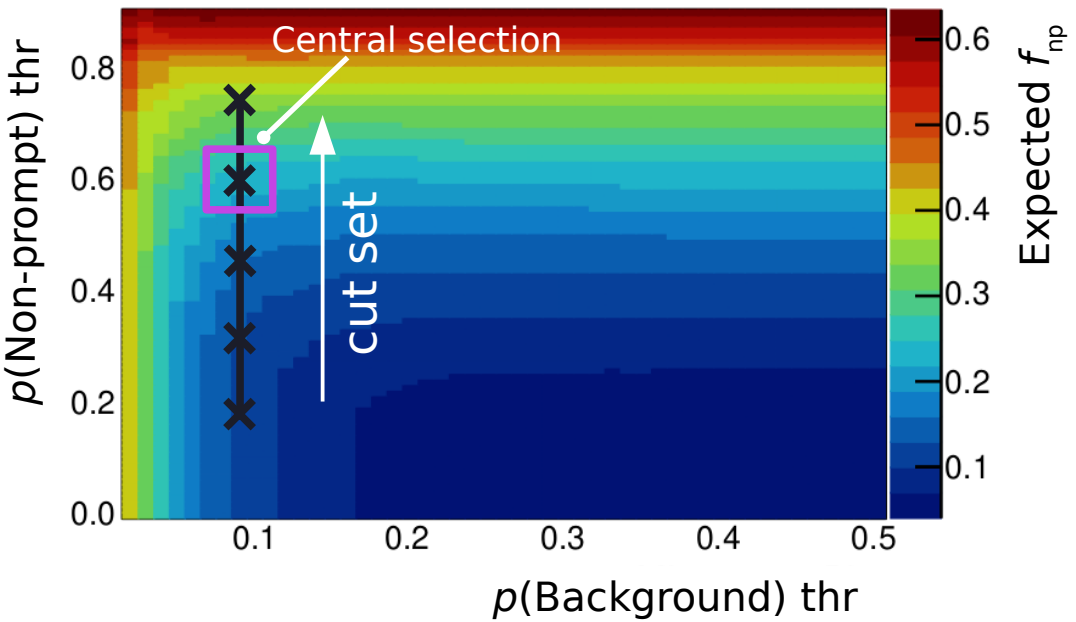
BDT

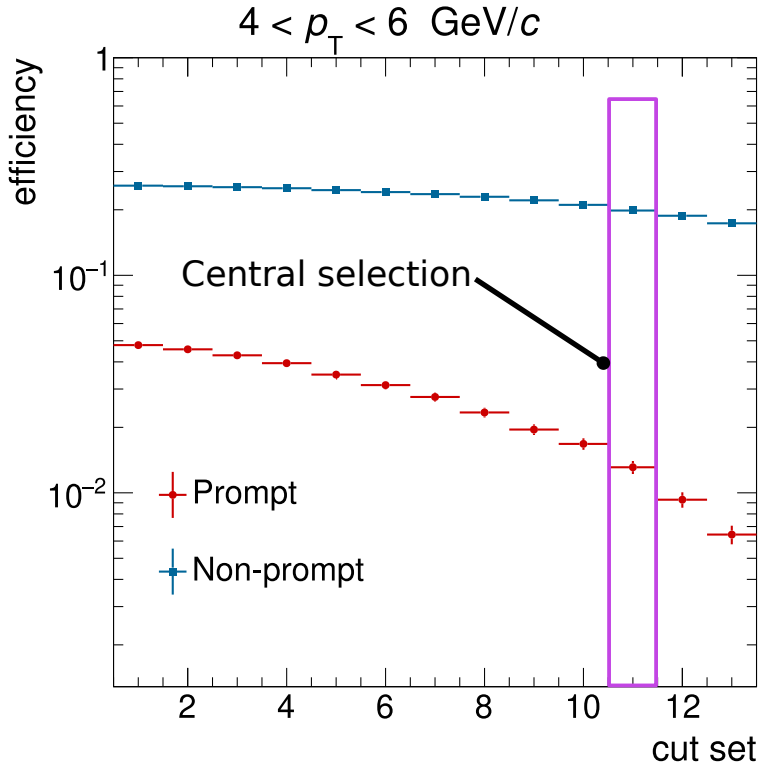
Probability (scores) of being:
- Prompt;
- Non-prompt;
- Background.



The Λ^+_c candidates are selected by requiring:
- **high** probability to be a **non-prompt** Λ^+_c;
- **low** probability to be a **background** candidate.

Estimation of the non-prompt fraction





Cut variation method:
Different cut sets → different efficiencies for prompt and non-prompt.

Compute the number of Λ^+_c candidates (raw yield Y) and efficiencies for each cut set.

Cut variation method

For each cut set *i*:

$$Y_i = \varepsilon_i^{\mathrm{p}} N^{\mathrm{p}} + \varepsilon_i^{\mathrm{np}} N^{\mathrm{np}}$$

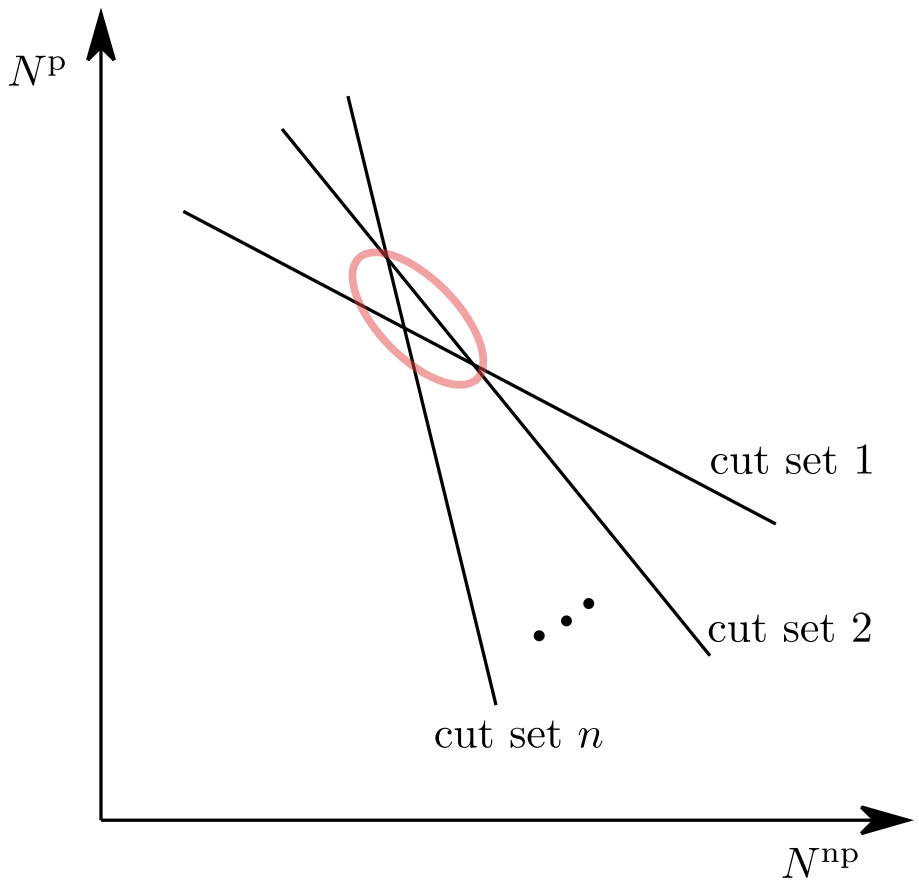
Where:

→ Y_i are the (measured) raw yields;
→ $\varepsilon_i^{\mathrm{p/np}}$ are the efficiencies;
→ $N^{\mathrm{p/np}}$ are the true yields;

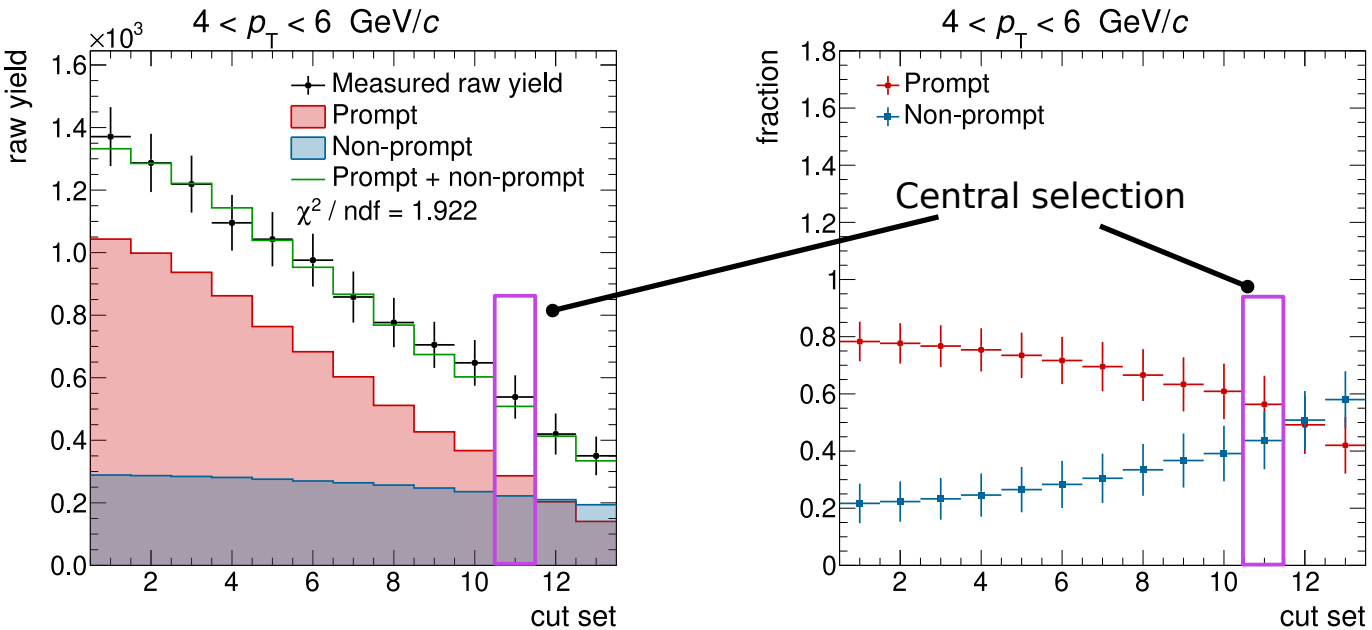
A system of equation is defined using many cut sets.

$$\begin{cases} Y_1 = \varepsilon_1^{\mathrm{p}} N^{\mathrm{p}} + \varepsilon_1^{\mathrm{np}} N^{\mathrm{np}} \\ \vdots \\ Y_n = \varepsilon_n^{\mathrm{p}} N^{\mathrm{p}} + \varepsilon_n^{\mathrm{np}} N^{\mathrm{np}} \end{cases}$$

- n equations and 2 variables (N^{p} and N^{np});
- The system is overdetermined;
- Solved minimising a χ^2-like quantity.



Prompt and non-prompt contributions



$$f_i^{\mathrm{p,np}} = \frac{\varepsilon_i^{\mathrm{p,np}} N^{\mathrm{p,np}}}{\varepsilon_i^{\mathrm{p}} N^{\mathrm{p}} + \varepsilon_i^{\mathrm{np}} N^{\mathrm{np}}}$$

The fit-like result (left) and the fraction of prompt/non-prompt (right).

- Prompt (red) and non-prompt (blue) Λ_c^+ contributions as obtained from the cut-variation method;
- Green line → fit-like function;
- For the central selection: high non-prompt fraction ~50%.

Non-prompt Λ^+_c cross section

Non-prompt Λ^+_c p_T-differential cross section.
(Not shown as it is not public yet)

Non-prompt cross section:

$$\left(\frac{d\sigma}{dp_T}\right)^{np} = \frac{f^{np}\, Y(\Lambda_c^+ + \Lambda_c^-)}{2\Delta p_T\, \varepsilon^{np}\, BR\, \mathcal{L}_{int}}$$

Models obtained from pQCD calculations.
Fragmentation fractions from LHCb.

Two possible decay tables for the H_b decay:

- PDG decay table:
  → only measured decays;
  → BR($\Lambda^0_b \rightarrow \Lambda^+_c$ + X) ~ 30%.
- PYTHIA8 decay table:
  ➡ also unobserved decays;
  → BR($\Lambda^0_b \rightarrow \Lambda^+_c$ + X) ~ 80%.

Summary and Outlook

Summary:

- First measurement of the non-prompt Λ^+_c cross section in pp was obtained;
- Described by pQCD calculations within uncertainties;
  → Better agreement with the FONLL prediction that uses the PYTHIA8 decay table, suggesting the presence of unobserved decays of type $H_b \rightarrow \Lambda^+_c + X$.

- Reference measurement for non-prompt Λ^+_c studies in PbPb collisions during the future LHC data-taking period.

Next steps:

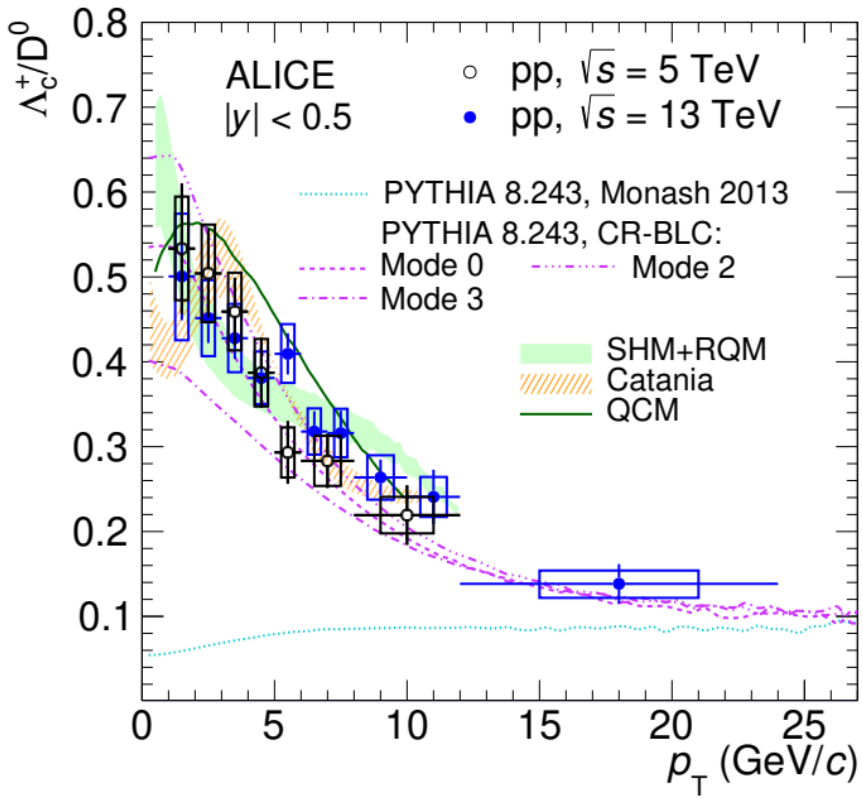
- Compare with the non-prompt $\Lambda^+_c \rightarrow pK\pi$ analysis (ongoing);
- Combine the results and publish.

Thank You!

Backup

Models for the charm and beauty production

- **PYTHIA8 (Monash):** colour string fragmentation model. Tuned on e^+e^- fragmentation fractions. Does not describe the data.
- **PYTHIA8 (CR-BLC):** Colour Reconnection Beyond Leading Colour. Colour strings fragmentation model + colour strings can reconnect → larger production of baryons. In agreement with the data.
- **Catania and Quark (re-)Combination Model (QCM):** implement a partonic coalescence mechanism, i.e., the quarks can combine with the other quarks produced in the collision. In agreement with the data.
- **SHM + RQM**: Statistical Hadronisation Model + Relativistic Quark Model: statistical model that describes the abundances of particles using a partition function. Augmented non-prompt fraction from the RQM. In agreement with the data.



The ALICE performance

Tracking and vertex reconstruction

- ITS + TPC

Resolutions:

- Imp. Par. ~70 μm at p_T = 1 GeV/c;

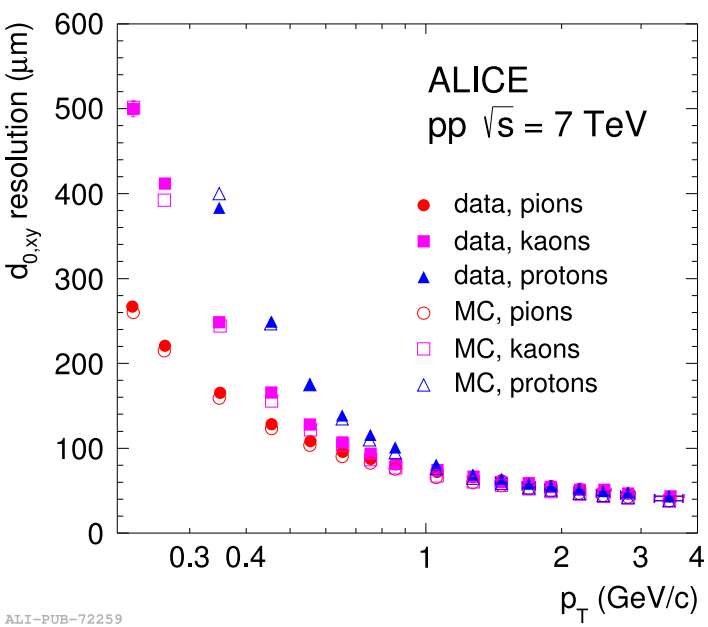
PID

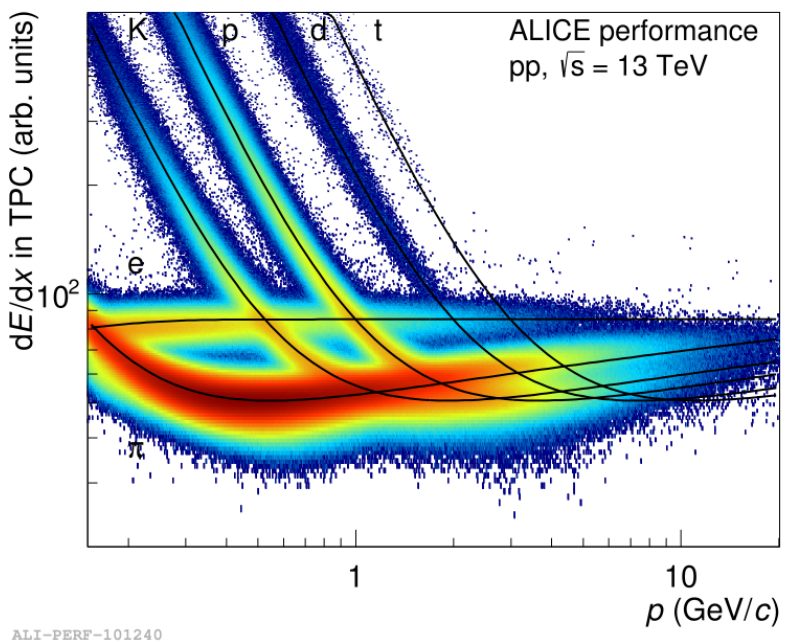
TPC (energy loss → d*E*/d*x*)
TOF (time of flight → β)

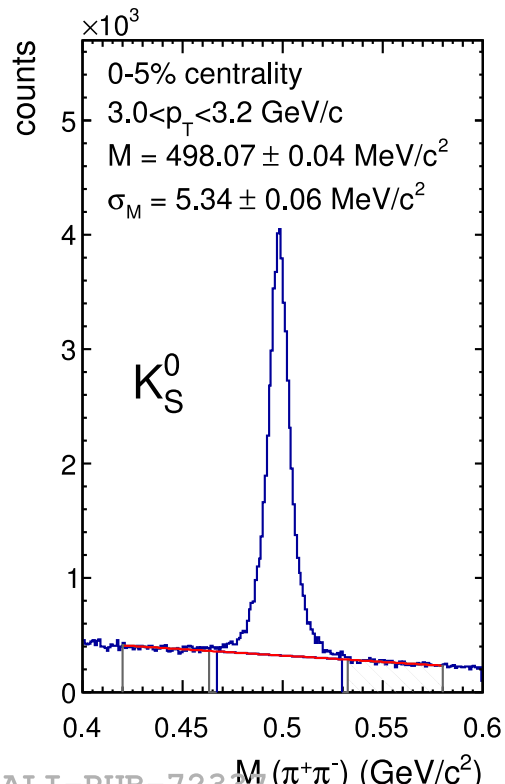
Different distribution for particles with different mass and charge.

K^0_s reconstruction

σ(Mass) ~5 MeV/c^2
High S/B ratio







Topological variables

The whole analysis relies on the displacement of the feed-down Λ^+_c baryons, therefore, topological variables of the Λ^+_c candidate, the bachelor (proton) and the V0 (K^0_S) are used.

Λ^+_c variables:

- Decay length,
- Decay length in xy plane,
- Decay length in xy plane/uncertanty,
- Cosine of pointing angle,
- Cosine of pointing angle in xy plane,
- Impact parameter,
- Impact parameter in xy plane,

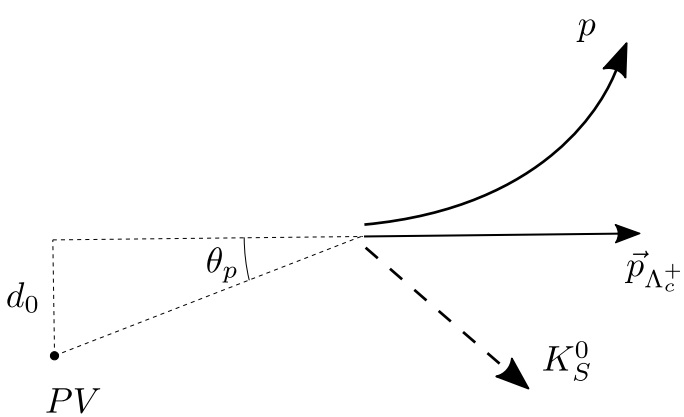
V0 variables:

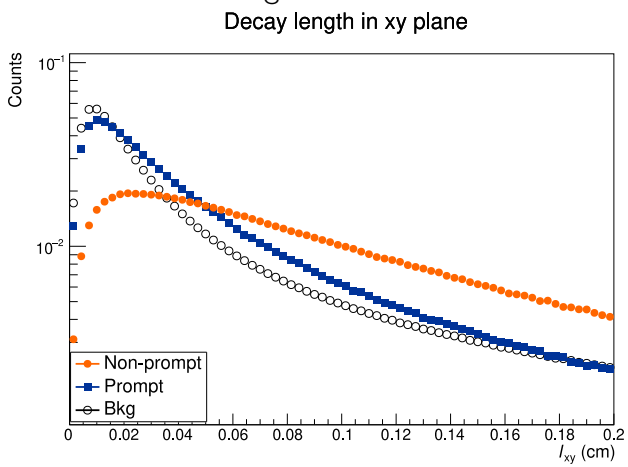
- Decay length,
- Cosine of pointing angle,
- Impact parameter.

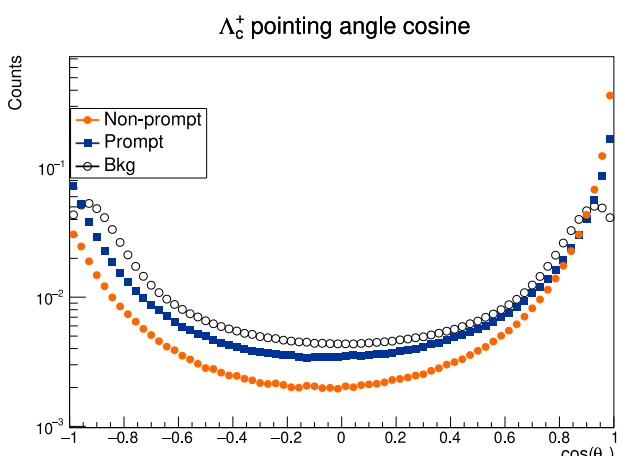
Very loose selections are applied to some of these variables in order to build the training set

Other variables:

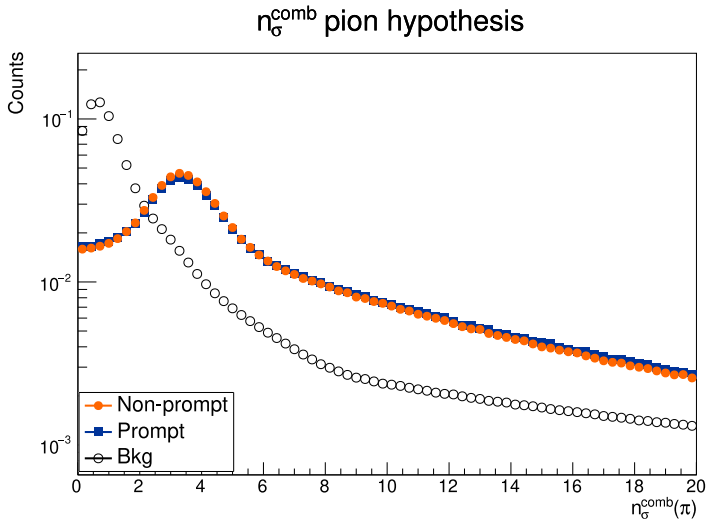
- Impact parameter of proton in xy plane with sign,
- Kalman Filter Particle topological χ^2.
- Impact parameter of prong 0
- Delta mass K^0_S,
- $c\tau(K^0_S)$.







PID



PID variables of the proton products:

n_σ combined proton, pion hypothesis;

n_σ combined proton, proton hypothesis.

The TPC detector measures the energy loss dE/dx and the TOF detector measures the time of flight. From these the n_σ variables are computed, which are defined as:

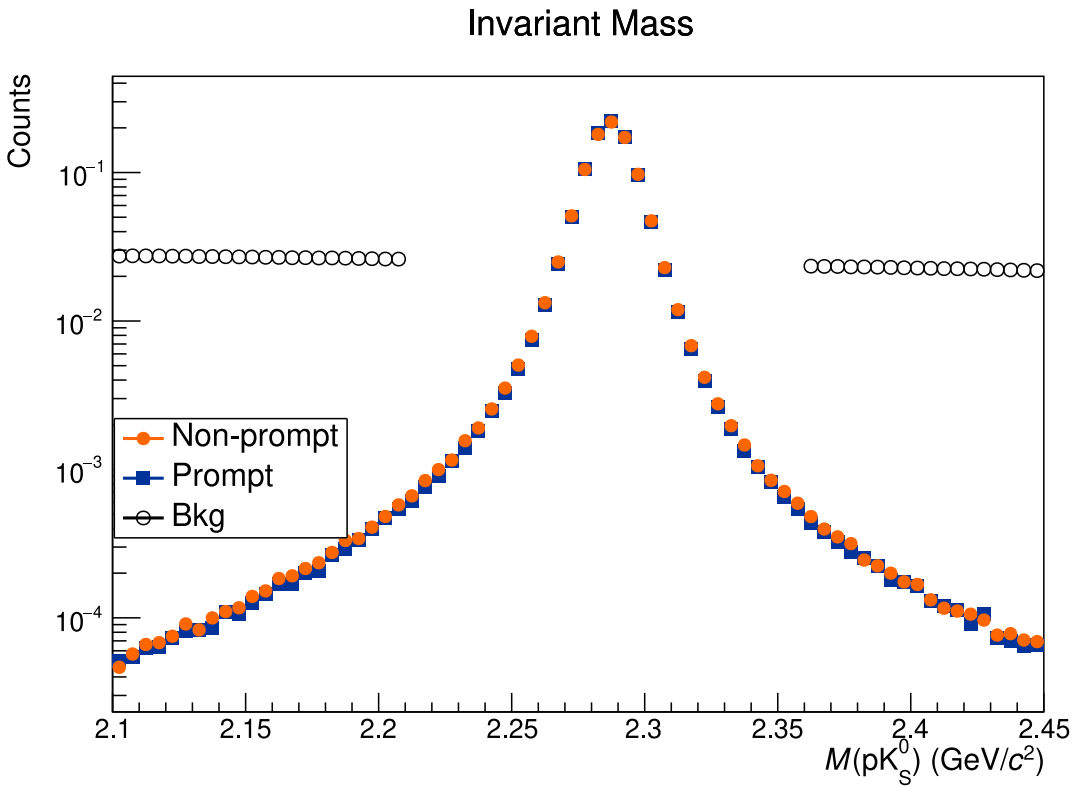
$$n_\sigma = \frac{S_{meas} - S_{exp}}{\sigma_S}$$

where S is the energy loss or the time of flight. These variables are then combined in a single one. In case of missing values, zero is used instead.

$$n_\sigma^{comb} = \frac{1}{\sqrt{2}}\sqrt{(n_\sigma^{TOF})^2 + (n_\sigma^{TPC})^2}$$

Machine Learning

For the training, a pure background sample is selected from the data, using the sidebands of the invariant mass distribution.



Machine Learning

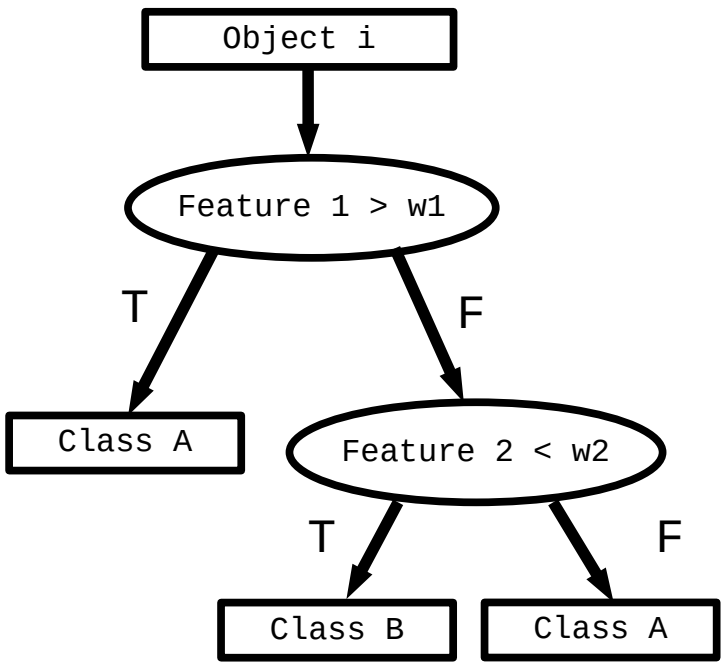
Machine Learning (ML) algorithms:

→ Exploit high-order correlation between data;
→ Better performance than rectangular selections;
→ Easily handle many variables;

Classification algorithms usually require the following step:

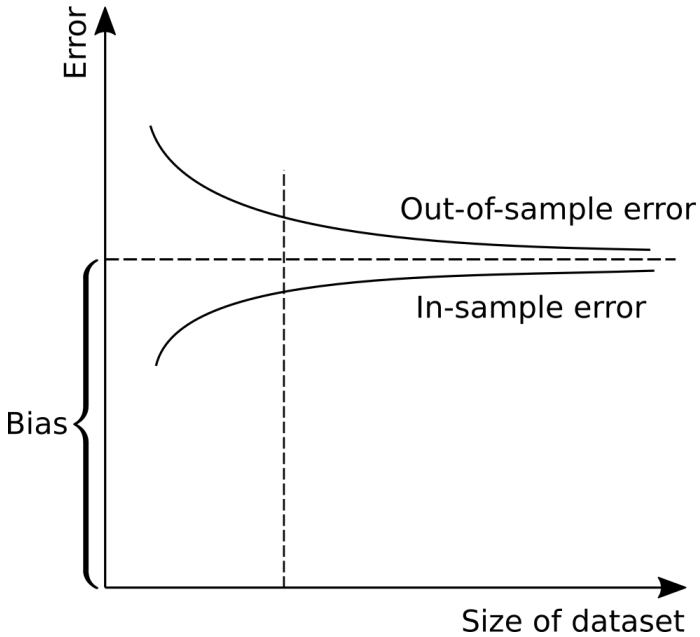
1) **Learning phase (Training):** the model learns from many examples of a labelled dataset;
2) **Hyperparameter optimisation:** find the optimal configuration of the model settings;
3) **Testing:** check the performance with an independent labelled dataset;
4) **Application:** apply the model to the unlabelled data.

The ML model used is a Boosted Decision Tree.

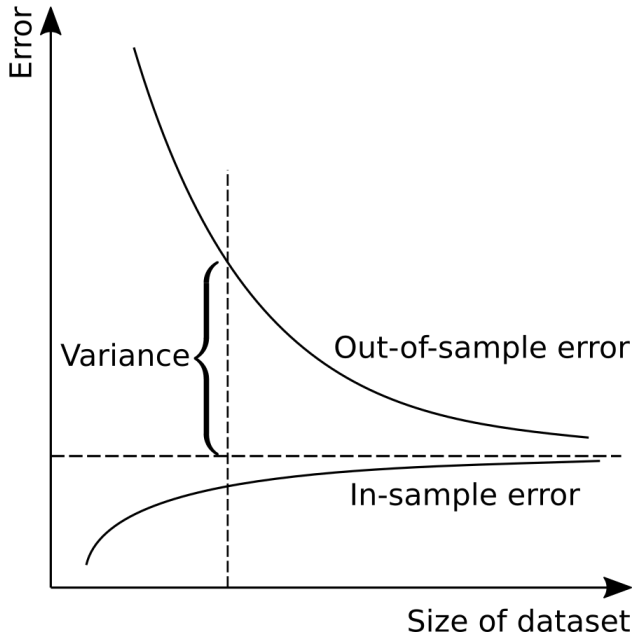


Bias-Variance trade-off

Simple model



Complex model

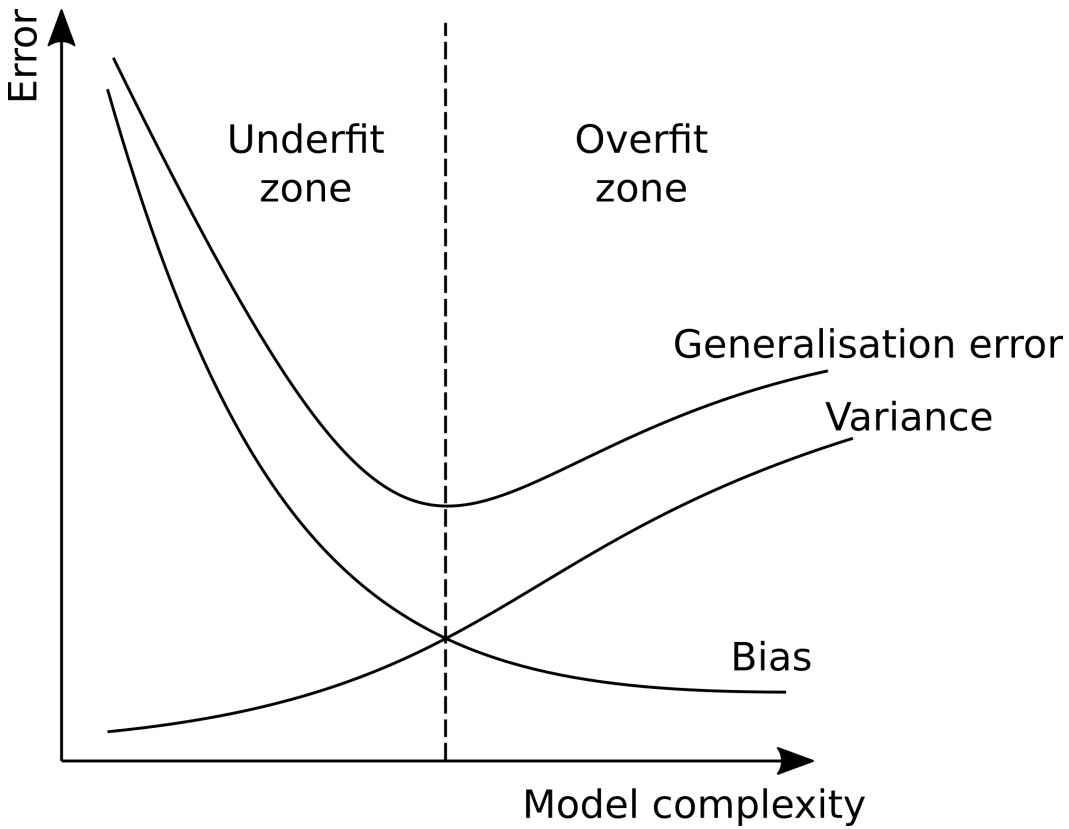


In-sample error: Mistake rate in the training set.

Out-of-sample error: Mistake rate in the test set.

The choice of the model and its hyperparameters must take into account the size of the available dataset.

Boosted Decision Trees: Bias-Variance tradeoff



The choice of the model complexity is crucial for finding a good solution to the problem.

Model complexity → hyperparameters

Overfit: the model learns the fluctuactions in the training set

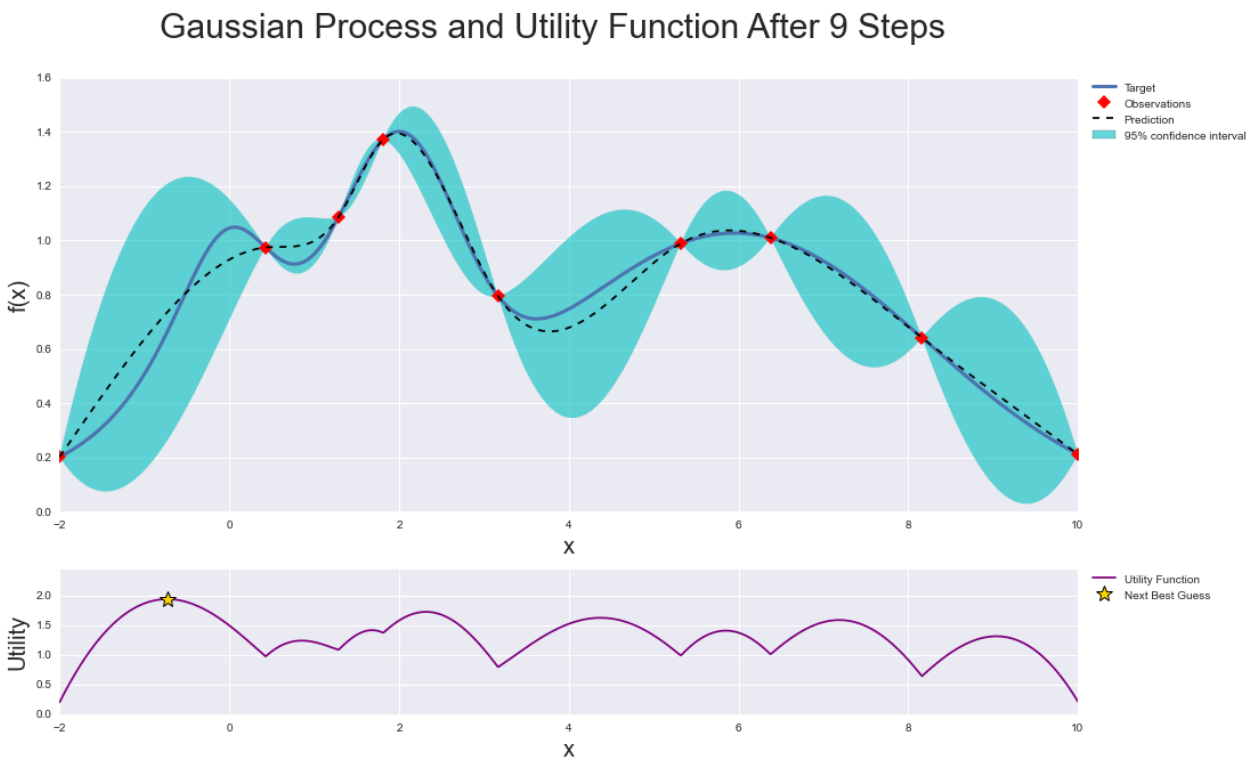
Underfit: the model is not complex enough to exploit the full information contained in the features of the data points.

Hyperparameter optimisation

Optimisation of hyperparameters:
The models depends on a number of hyper-parameters that the user chooses and that have to be optimised. For BDT, typical hyper-parameters are the number of trees, the learning rate, the depth of the trees.

→ 5-folds cross validation

→ Bayesian approach:
the configuration to test is determined according to the performance of the previous hyperparameter configurations.



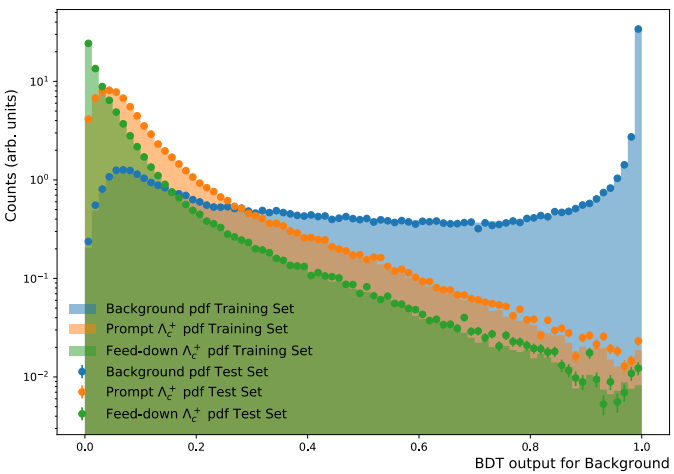
Hyperparameter optimisation

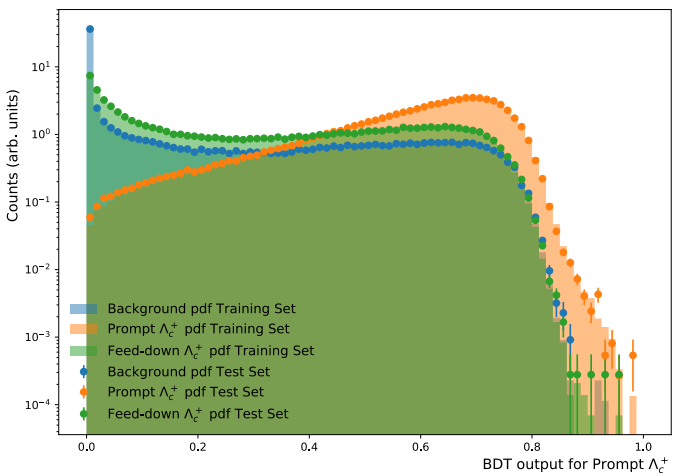
- **max_depth:** the maximum depth of the binary classifier;
- **learning_rate:** the weighting factor that is applied to the updates of the model parameters;
- **n_estimators:** the number of estimators used in the boosting of the model.
- **min_child_weight:** the sum of the Hessian of the loss function over the instances in a node. For classification, this parameter is related to the minimum purity required to stop splitting the node.
- **subsample:** the fraction of the training instances used to train the tree;
- **colsample_by_tree:** the fraction of the training variables used in the training of the tree.

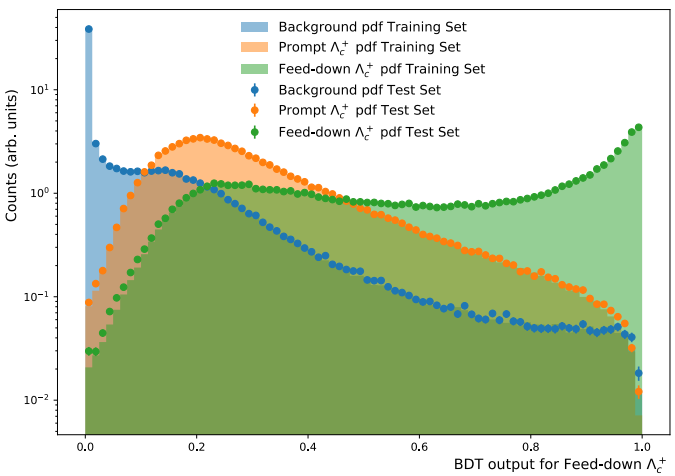
For more details see: https://xgboost.readthedocs.io/en/latest/parameter.html

BDT scores

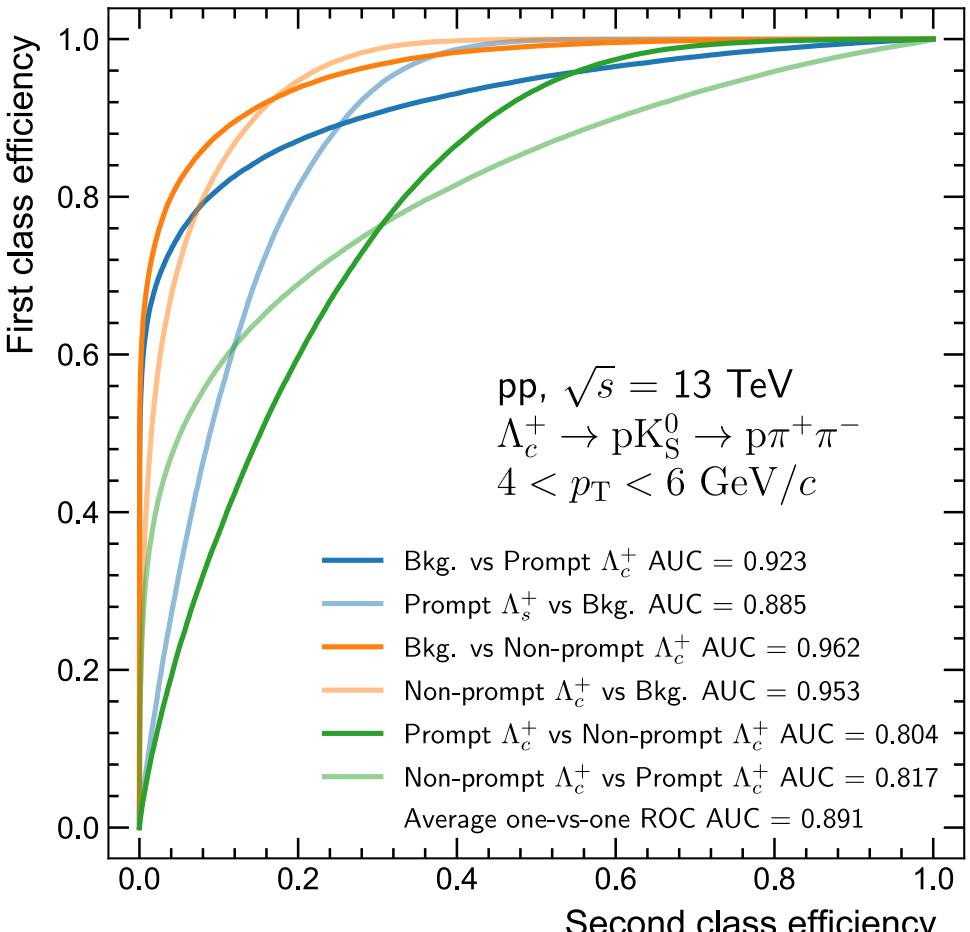
$4 < p_T < 6$ GeV/c







Performance of the model



For a given threshold t, we can define the First Class Efficiency (FCE) defined as:

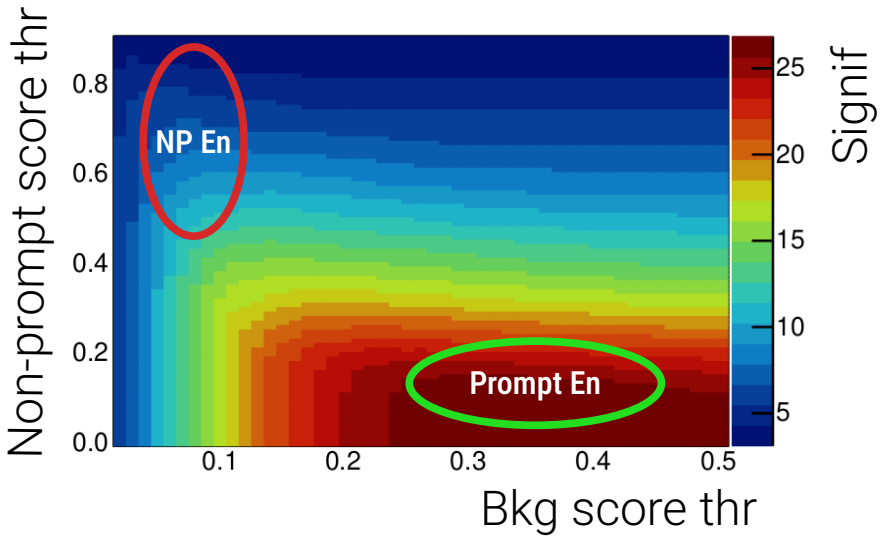
$$FCE(t) = \frac{N(p > t)}{N}$$

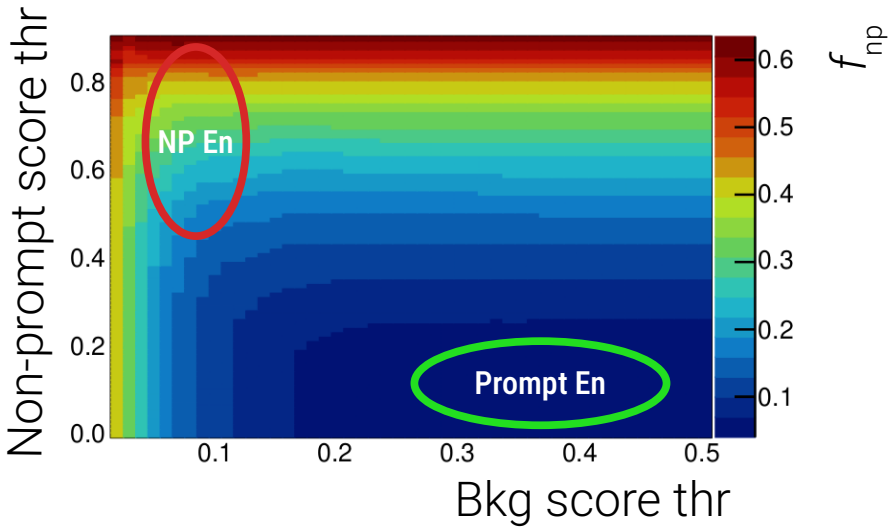
Second Class Efficiency defined analogously.

ROC (Receiver Operating Characteristic) curves correlate the FCE and SCE.

The performance of the model is quantified with the AUC (Area Under the Curve) of the ROC curve.

Optimisation of the ML selection





- (Pseudo-)signal *S* from pQCD calculations and efficiencies from simulations;
- Background yield *B* from a sub-sample of data and rescaled to full sample

Selection criteria on BDT scores optimised based on significance and f_{np}:

$$\text{Signif} = \frac{S}{\sqrt{S+B}}, \quad f_{np} = \frac{\varepsilon^{np}(d\sigma/dp_T)^{np}}{\varepsilon^{p}(d\sigma/dp_T)^{p} + \varepsilon^{np}(d\sigma/dp_T)^{np}}$$

Are selected the candidates with:

- Bkg score < Bkg_thr
- Non-prompt score > NP_thr.

The thresholds are then varied in the represented range.

2 pairs of cuts (bkg_thr, NP_thr) are chosen in order to maximise:

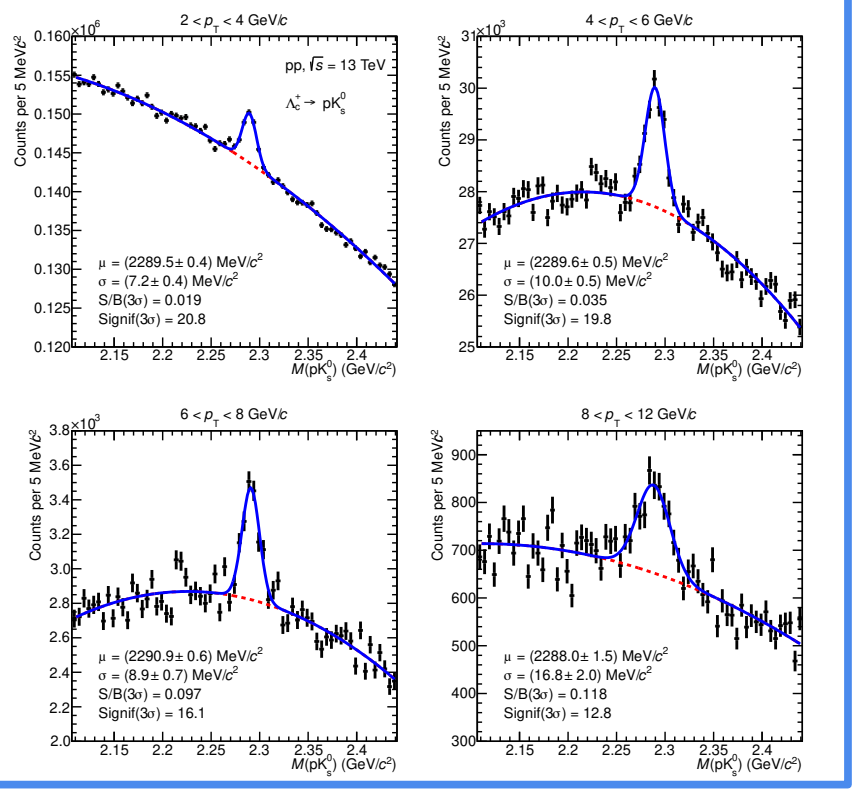
- the significance → prompt enhanced (Prompt En);
- the NP fraction → non-prompt enhanced (NP En).

The threshold of the NP enhanced configuration is chosen to maximise f_{np} while preserving a significance >7

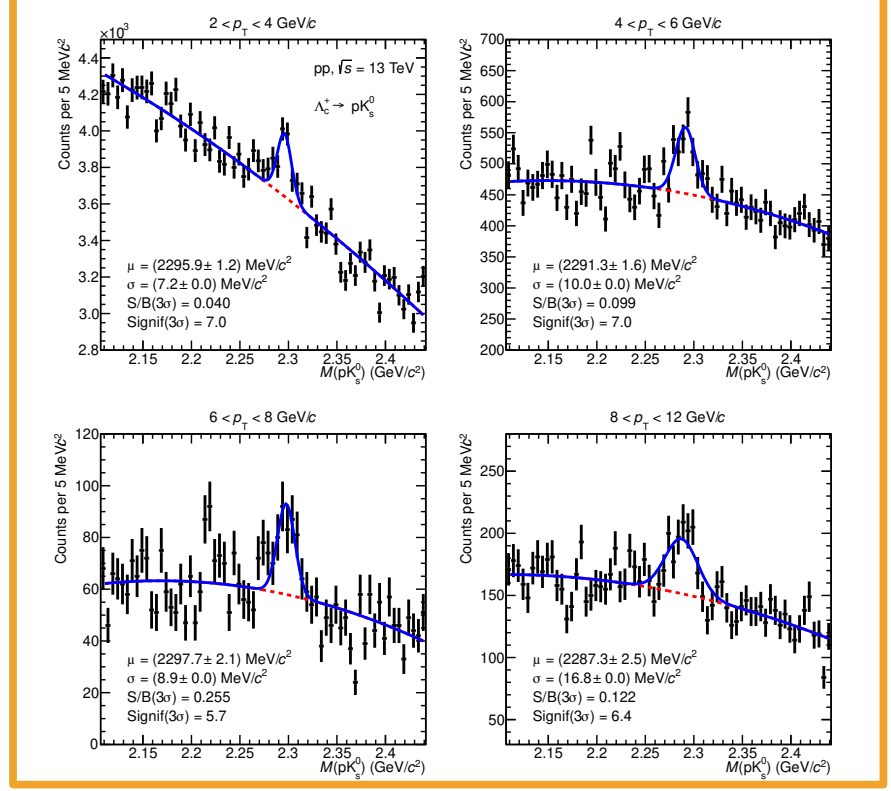
Extraction of raw yields

- Fit to the pK^0_s invariant mass distribution with the sum of a parabola for the background and a Gaussian function for the signal;

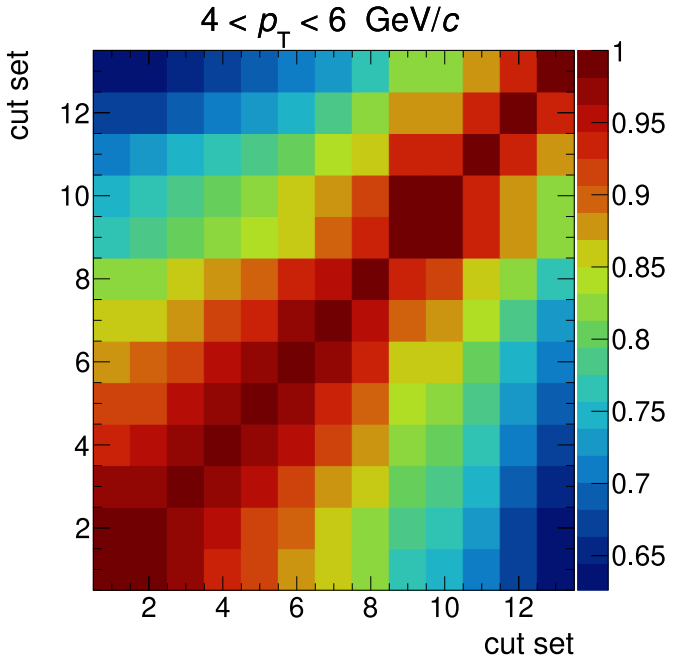
Prompt Enhanced

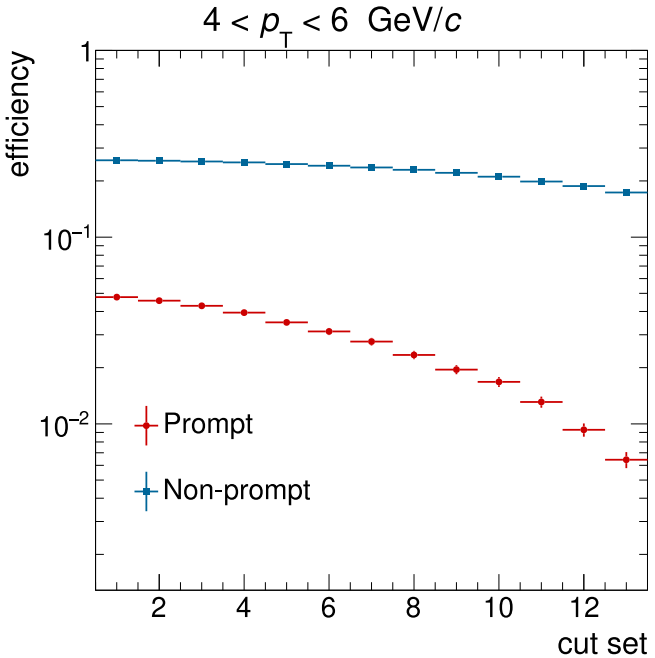


Non-prompt Enhanced



Correlation and efficiency





Correlation matrix of the yields:

- With increasing cut set number the cut on the non-prompt score is progressively more severe;
- the correlation is stronger between closer cuts;
- Assumption: full correlation between yields obtained with tighter sets of cuts;

Efficiencies:

- the efficiency of prompt candidates decreases more rapidly than the efficiency of non-prompt.
- the non-prompt enhanced selection is the number 11;

Estimation of the covariance matrix

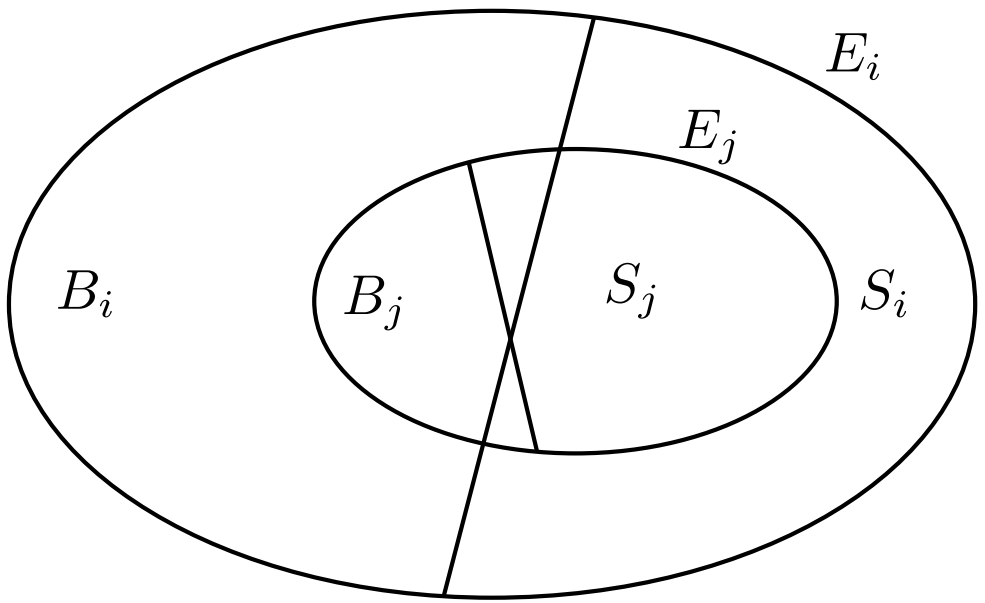
- Small S/B ratio ⇒ random fluctuations of the background might influence the RY extraction. If the background is not negligible the formula

$$\mathrm{Cov}(S_i, S_j) = \mathrm{Var}(S_j)$$

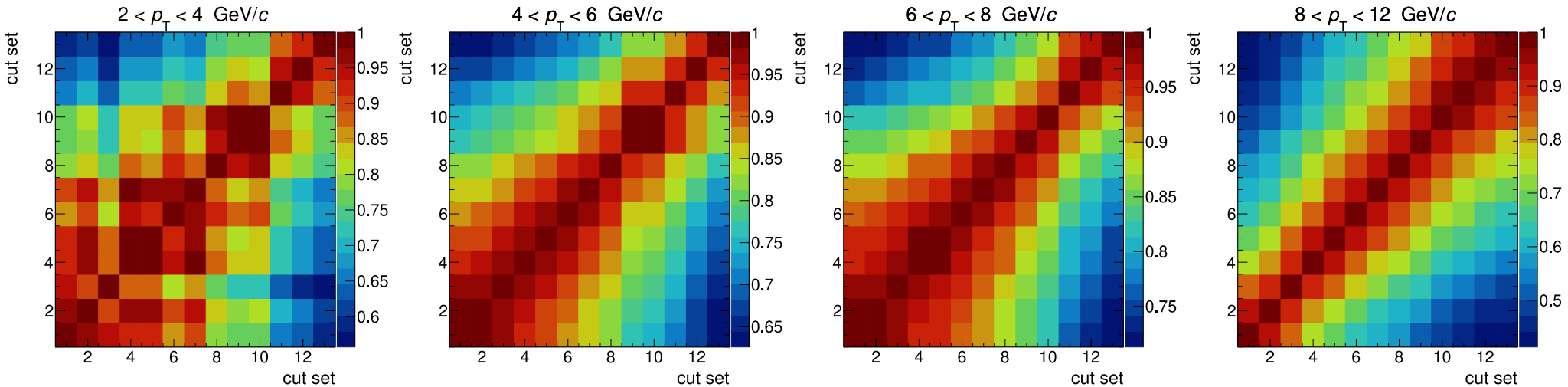
  is no longer valid. The complete analytic propagation of errors is:

$$\mathrm{Cov}(S_i, S_j) = \mathrm{Cov}(E_i, S_j) - \mathrm{Cov}(B_i, S_j)$$

- But it fails since it's not guaranteed that B_i and S_j have an empty intersection.
- Numerical estimation of the correlation matrix is performed instead via toy MC.



Estimation of the correlation matrix with toy MC

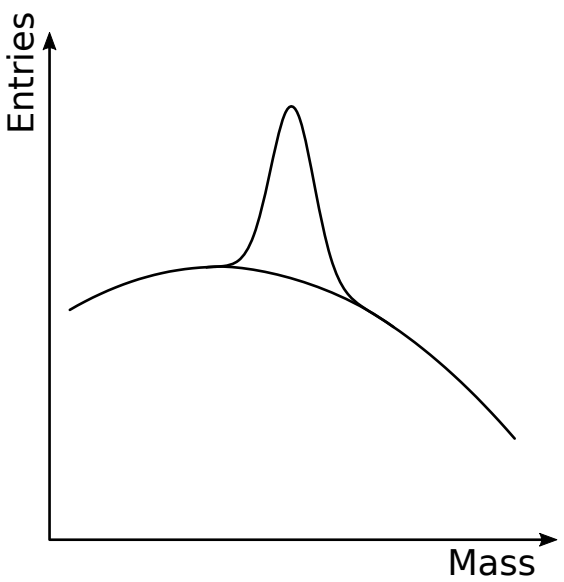


Correlation matrix of the 2-4 GeV/c p_T interval → “chequerboard pattern”.

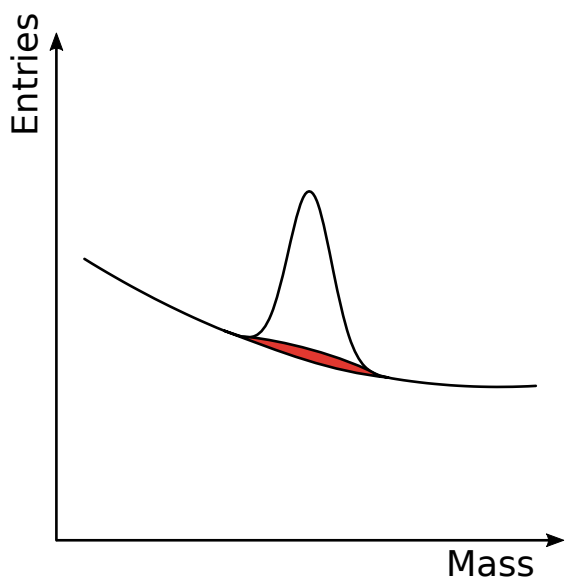
The sets are considered as fully correlated, which is true if the background is negligible. Otherwise, the extraction of the raw yields is sensitive to the background fluctuations.

Estimation of the correlation matrix with toy MC

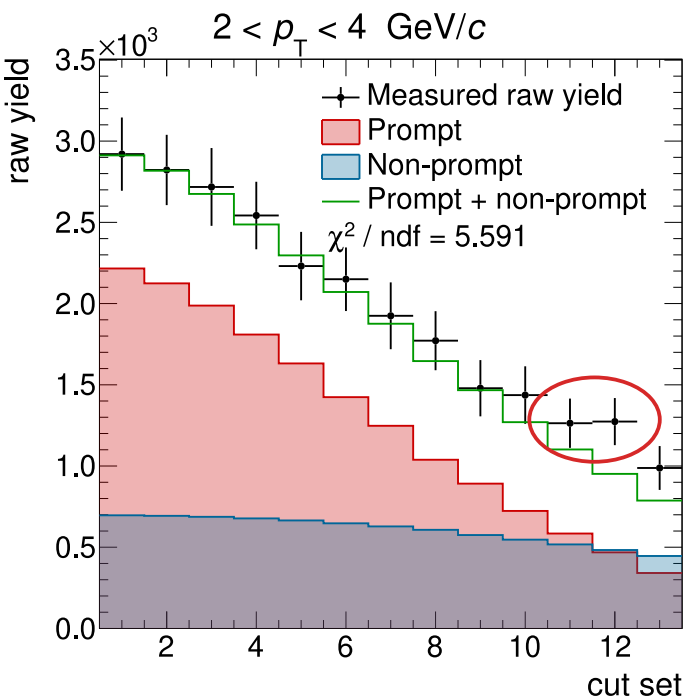
Large background ⇒ large fluctuation on the RY extraction. This might even result in a larger RY for a harder selection on the ML score (as shown).
For this reason, we should consider the RY extraction as performed on partially correlated sets.
To take into account this effect, the correlation matrix was estimated with a toy MC approach.



Looser cut



Tighter cut



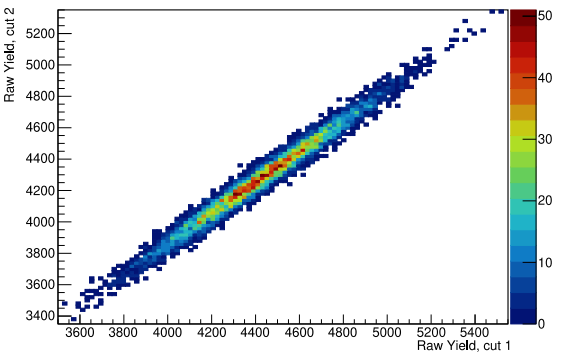
Estimation of the correlation matrix

To estimate the correlation between the RYs, many experiments are simulated with a ToyMC.

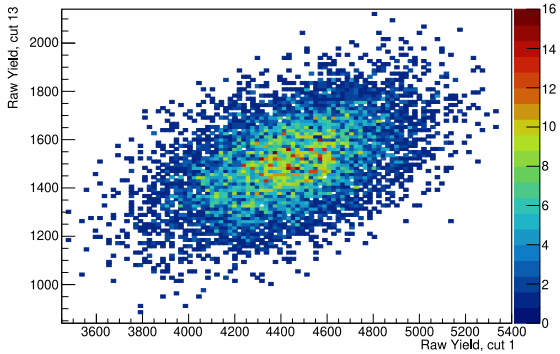
- A pair of cuts is selected, a looser and a tighter one. Then we generate a smeared version of the invariant mass spectra corresponding to these selection.
- To keep the correlation between the RYs the spectra are not smeared directly, but the tighter cut spectrum and the difference between the looser and tighter cut are smeared instead.
- The smeared spectra obtained are then are then added, so a spectrum corresponding to the looser cut is obtained.
- The looser and harder smeared spectra are fitted again and a pair of RY is extracted.

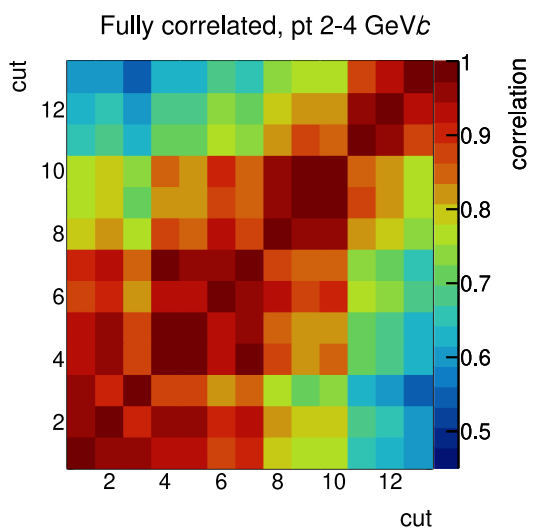
- Repeating this calculations many times leads to an accurate estimation of the correlation matrix.

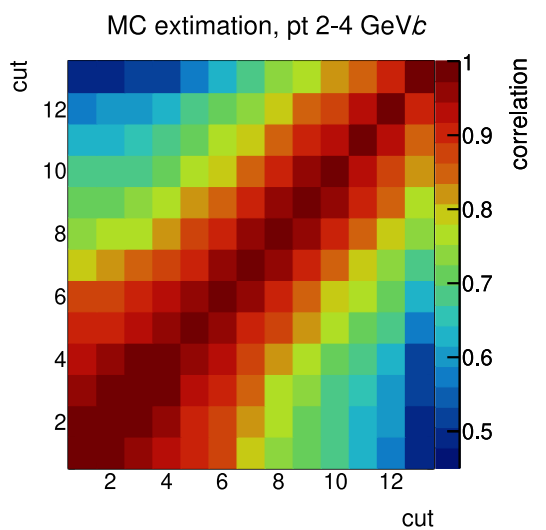
Close cuts



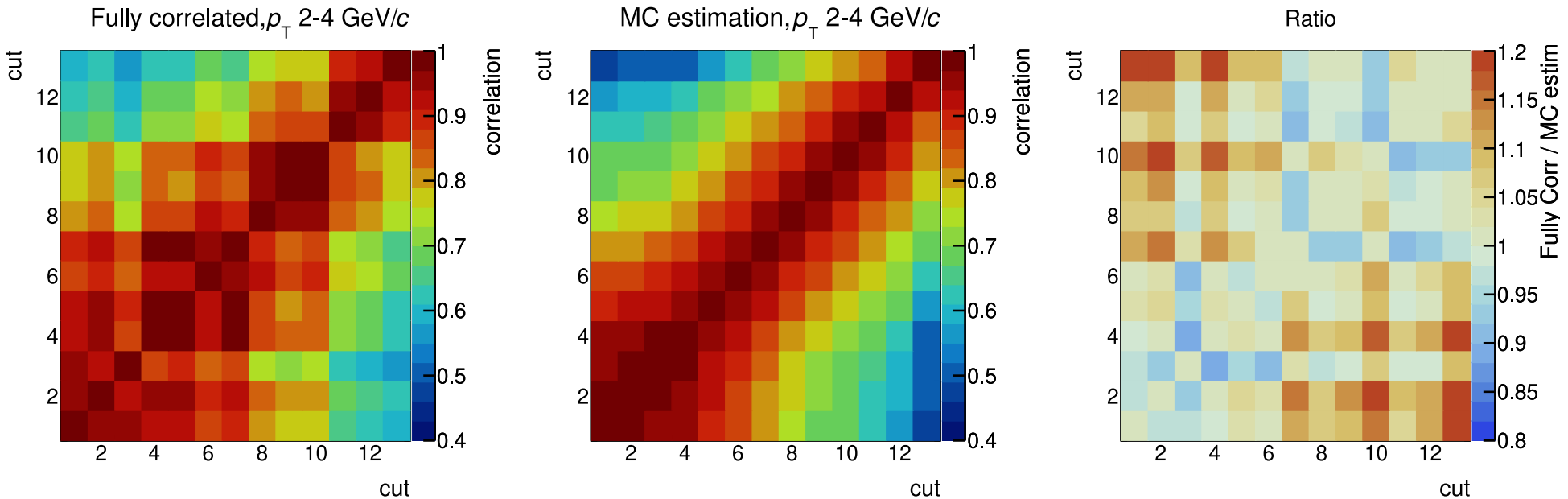
Distant cuts







2-4 GeV/*c* Correlation matrices compared



- The correlation matrix estimated with the toy MC is “smoother” than the approximated one where the sets are considered as fully correlated.
- The fully correlated one overestimates the correlation up to ~20% away from the diagonal and underestimates it up to ~10% near the diagonal.
- No significant effect on the estimation of the non-prompt fraction.
- The effect is negligible for the other transverse-momentum intervals.

The cut variation method

$$Y_i = \xi_i^p N^p + \xi_i^{np} N^{np} \quad \text{with} \quad \xi_i^{p/np} = \varepsilon_i^{p/np} \times \text{Acc}_i^{p/np}$$

- Let

$$\delta_i = \xi_i^p N^p + \xi_i^{np} N^{np} - Y_i$$

- The χ^2 is the quantity to minimise in order to estimate the $N^{p/np}$

$$\chi^2 = \delta^T C^{-1} \delta$$

- Approximation: the raw yields are considered as obtained from fully correlated sets.

- The correlation between the deltas is expressed as

$$\text{Corr}(\delta_i, \delta_j) = \frac{\sigma_j}{\sigma_i} \quad \text{with} \quad j \subset i$$

- Where the uncertainty on the deltas is computed as

$$\sigma_i^2 = \sigma_i^2(Y_i) + \sigma_i^2(\xi_i^{\text{p}})(N^{\text{p}})^2 + \sigma_i^2(\xi_i^{\text{np}})(N^{\text{np}})^2$$

Prediction for the non-prompt cross section

FONLL:
→ cross sections of heavy quarks;
→ tuned with the fragmentation fractions from e^+e^- collisions;
→ good description of the B meson cross section;
→ for non-prompt Λ^+_c a different approach is used.

For the decay $H_b \rightarrow \Lambda^+_c + X$, the most relevant beautyhadrons are B^0, B^+, B_s, Λ^+_c.
The cross section of mesons is computed as:

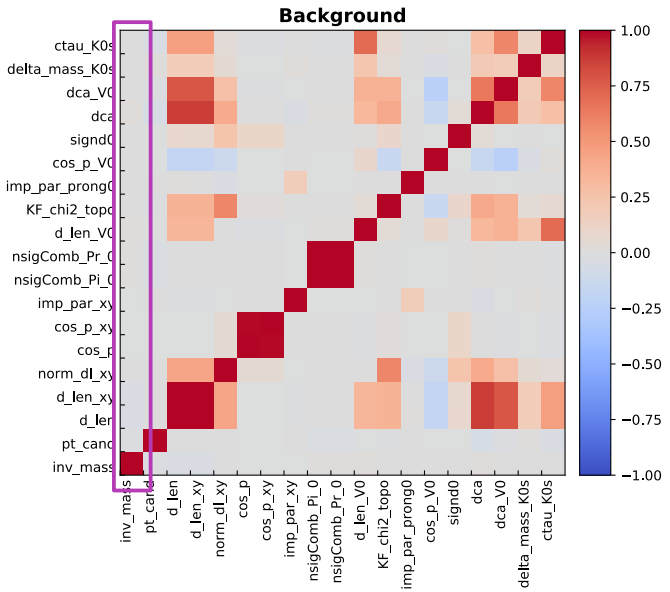
$$\frac{d\sigma(B^0, B^+, B_s^0)}{dp_T} = \left[\frac{d\sigma(b)}{dp_T}\right]_{FONLL} \times f(b \rightarrow B^0, B^+, B_s^0)_{e^+e^-}$$

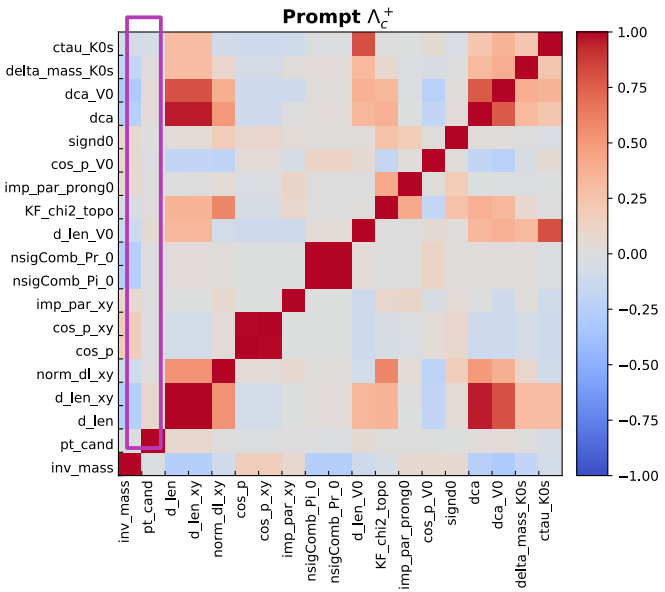
While the Λ^+_b cross section is computed as:

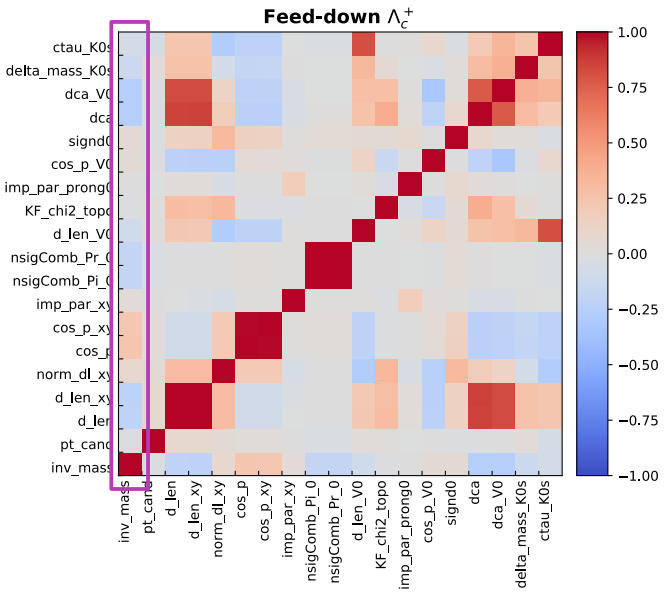
$$\frac{d\sigma(\Lambda_b^0)}{dp_T} = \left[\frac{d\sigma(b)}{dp_T}\right]_{FONLL} \times \left(\frac{f_{\Lambda_b^0}}{f_d + f_u}\right)_{LHCb} \times [f(b \rightarrow B^0) + f(b \rightarrow B^+)]_{e^+e^-}$$

Then run a MC simulation:
1) randomly choose a beauty hadron H_b in {B^0, B^+, B_s, Λ^+_c} according to their abundances;
2) assign to it a transverse momentum obtained with a sampling of the corresponding cross section;
3) let the H_b decay with PYTHIA8 using the PDG or the PYTHIA8 decay table;
4) if the decay products contain a Lc → fill a histogram with its momentum;
5) Rescale the momentum distribution to obtain the non-prompt cross section.

Mass shaping







Linear correlations between variables.

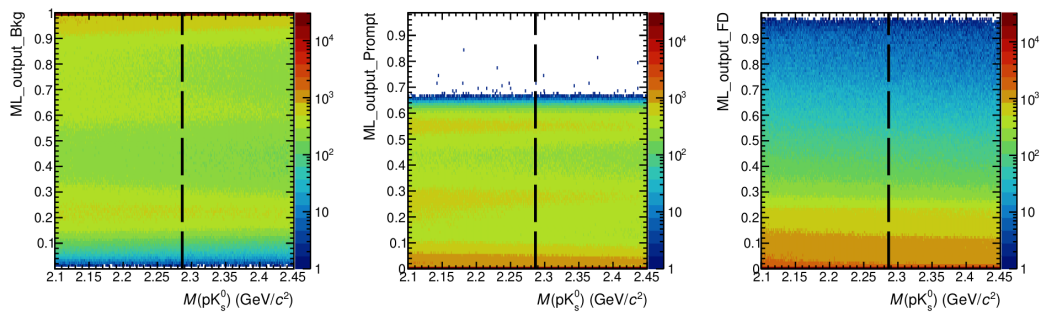
If the invariant mass of the candidates is correlated with some other variables used in the training, the model might learn these correlation and classify the ML scores might be dependent on the mass of the candidate. When this happens, a structure in the mass region is formed (mass shaping).

To check the mass shaping, a sample of pure background obtained from a general purpose MC is used to compare the shape of the spectrum before and after that the selection on ML output score is applied.

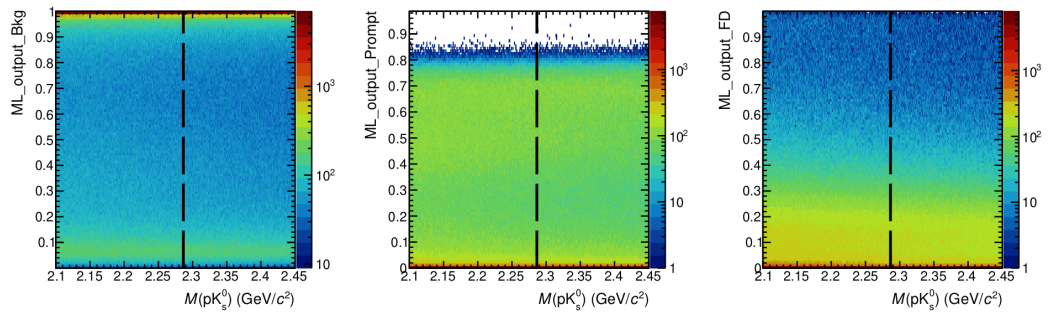
Mass shaping

- Correlations between the invariant mass of the background candidates and their ML scores.
- No structure is found in the region of the signal peak. OK.

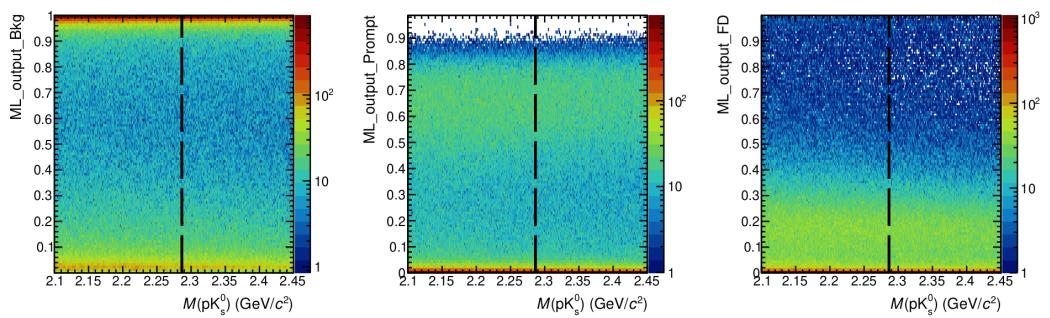
$2 < p_T < 4$ GeV/c



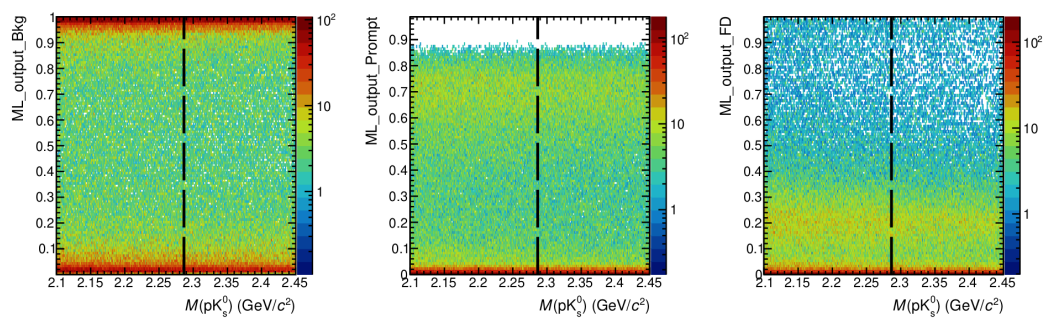
$4 < p_T < 6$ GeV/c



$6 < p_T < 8$ GeV/c



$8 < p_T < 12$ GeV/c



Systematic uncertainties

Several sources of systematic uncertainties are identified:

Signal extraction:

✔ **raw yield:** the fit to the invariant mass distribution is repeated varying the fit settings;

Sources related to the corrections:

✔ **tracking efficiency:** compare ITS-to-TPC prolongation efficiency for charged tracks in data and MC and test stability of the cross section varying the single-track selection criteria (inherited from other analysis);

✔ **ML selection efficiency:** evaluate stability of the cross section varying the selection criteria;

✔ **non-prompt fraction:** vary the sets of selections used in the χ^2-like minimisation of the system of equations;

✔ **MC p_{T} shape:** re-weight p_{T} distributions of Λ_{c}^{+} in MC to reproduce the data;

Theory-driven subtraction of the non-prompt Λ_c^+

Validation of the theory-driven subtraction of the feed-down:

Traditional way of computing the cross section of the prompt component.

The number of Λ_c^+ baryons measured is the sum of the prompt and non-prompt baryons. When calculating the production cross section of the prompt Λ_c^+ baryons, the feed-down component is subtracted with a theory-driven method based on pQCD calculations with FONLL.

From the cross section it's possible to calculate the number of Λ_c^+ baryons that we expect, then the fraction of feed-down is simply:

$$f_{\mathrm{fd}} = \frac{N^{\mathrm{fd}}_{\mathrm{raw,\ FONLL}}}{N^{\mathrm{p+fd}}_{\mathrm{raw,\ Data}}} \qquad N^{\mathrm{fd}}_{\mathrm{raw,FONLL}} = \left(\frac{\mathrm{d}\sigma}{\mathrm{d}p_{\mathrm{T}}}\right)^{\mathrm{fd}}_{\mathrm{FONLL}} \Delta p_{\mathrm{T}} \, (\mathrm{Acc}\times\varepsilon)^{\mathrm{fd}}_{\mathrm{MC}} \, BR\mathcal{L}_{\mathrm{int}}$$

A measurement of the non-prompt cross section would validate this method, the disadvantage is that the data-driven method is limited by the statistics.