Data selection tools based on machine learning for hyperon form factor studies with the Belle II experiment

Academic Year 2022-2023

Federico Bonaldo

Supervisor: Prof. Simonetta Marcello

Co-Supervisors:
Prof. Karin Schönning
Dr. Bianca Scavino
Dr. Martina Laurenza

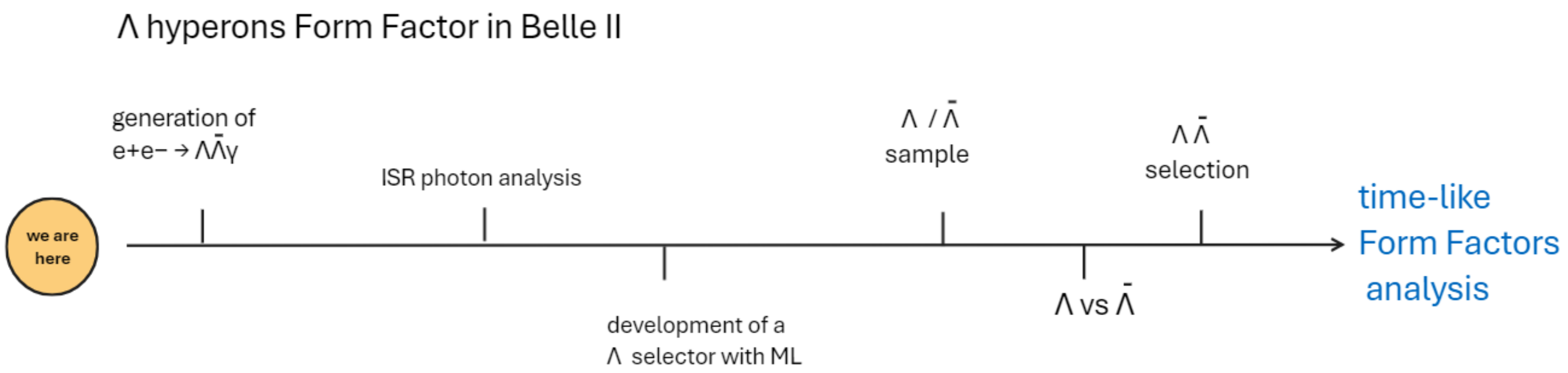
Examiner: Prof. Elena Botta

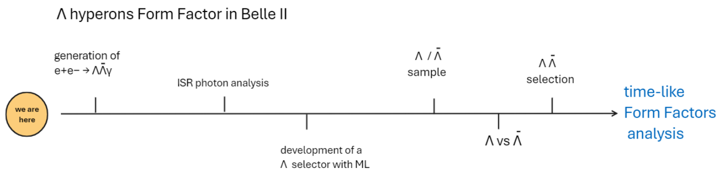
Thesis goals

Development of an event selection toolkit for the Λ hyperons form factor studies in the Belle II experiment:

- Study the reconstruction performance of the Initial State Radiation (ISR photon) in the Belle II detector

- Development of Λ hyperon selector for Belle II data



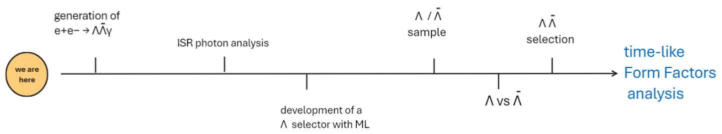


Λ Hyperon

Hyperon: nucleon with one or more **u** or **d** quark replaced by a **s** quark

- Baryon composed by uds quark
- $M_\Lambda = 1.115\ GeV/c^2$
- Neutral
- S = -1
- $\tau_\Lambda = 261\ ps$

- Main Λ decay modes:
  - $\Lambda \to p\,\pi^-$ (64%)
  - $\Lambda \to$ n π^0 (36%)

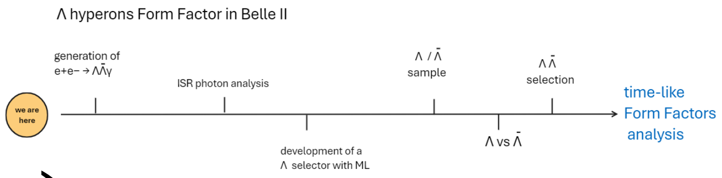


Λ Hyperon

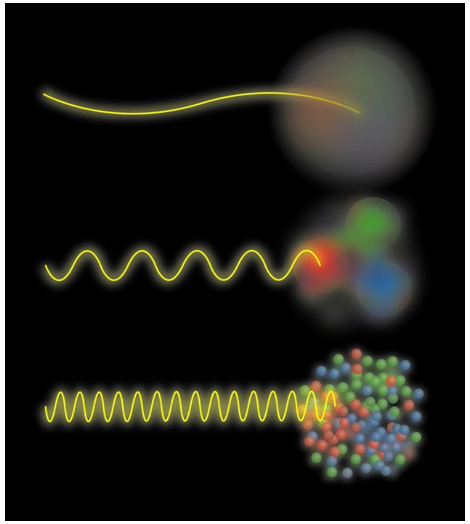
Hyperon: nucleon with one or more **u** or **d** quark replaced by a **s** quark

- Baryon composed by uds quark
- $M_\Lambda = 1.115\ GeV/c^2$
- Neutral
- S = -1
- $\tau_\Lambda = 261\ ps$

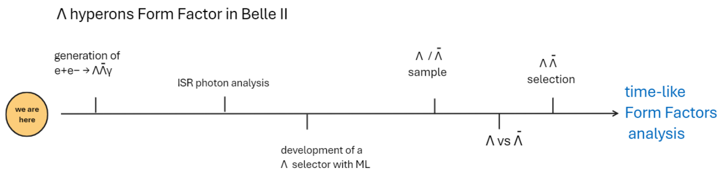
- Main Λ decay modes:
  - $\Lambda \to p\,\pi^-$ (64%)
  - $\Lambda \to \mathrm{n}\,\pi^0$ (36%)



Electromagnetic Form Factors (EMFFs)



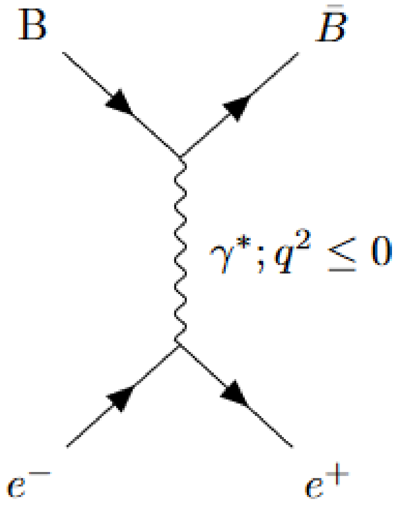
- A tool for studying the structure of the hadrons
- EMFFs quantify the deviation from the point-like particle
- It is defined as a function of the momentum squared transferred to the baryon via a virtual photon q^2



Hyperon Form Factor

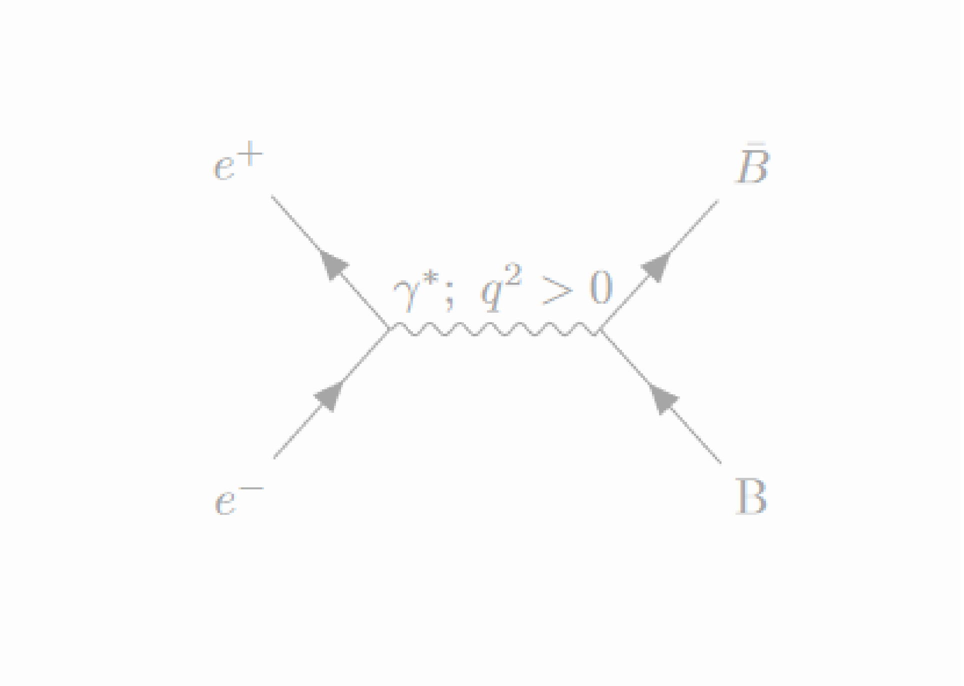
Space-like EMFF $q^2 \leq 0$

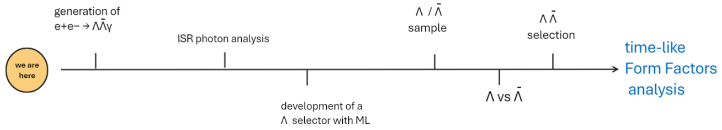
$$e^- B \rightarrow e^- B$$



Time-like EMFF $q^2 > 0$

$$e^+ e^- \rightarrow B\bar{B}$$

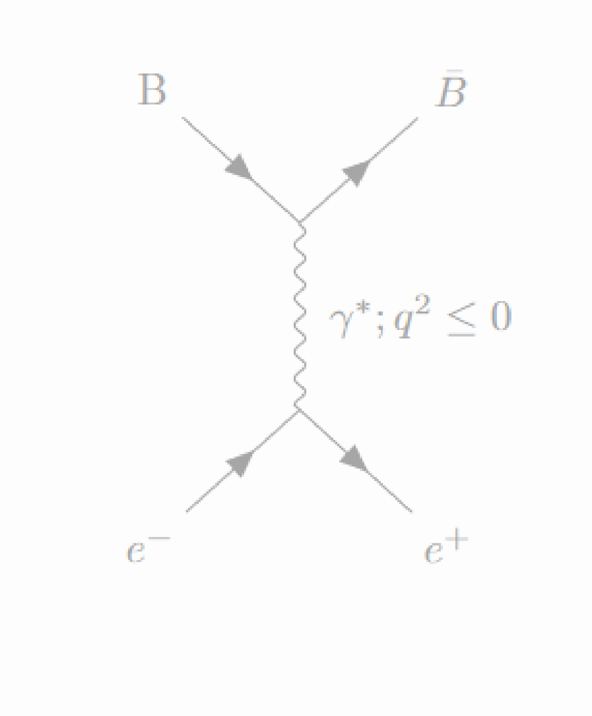




Hyperon Form Factor

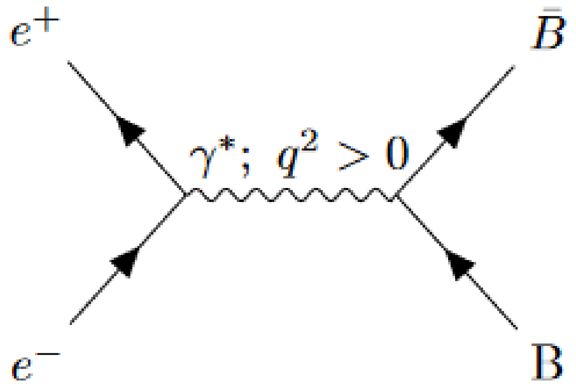
Space-like EMFF $q^2 \leq 0$

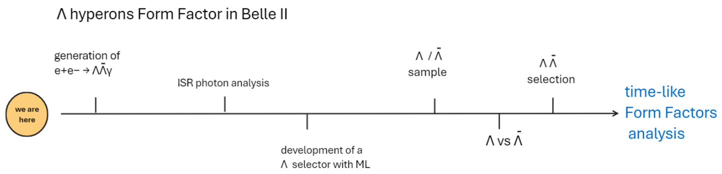
$$e^- B \to e^- B$$



Time-like EMFF $q^2 > 0$

$$e^+ e^- \to B\bar{B}$$

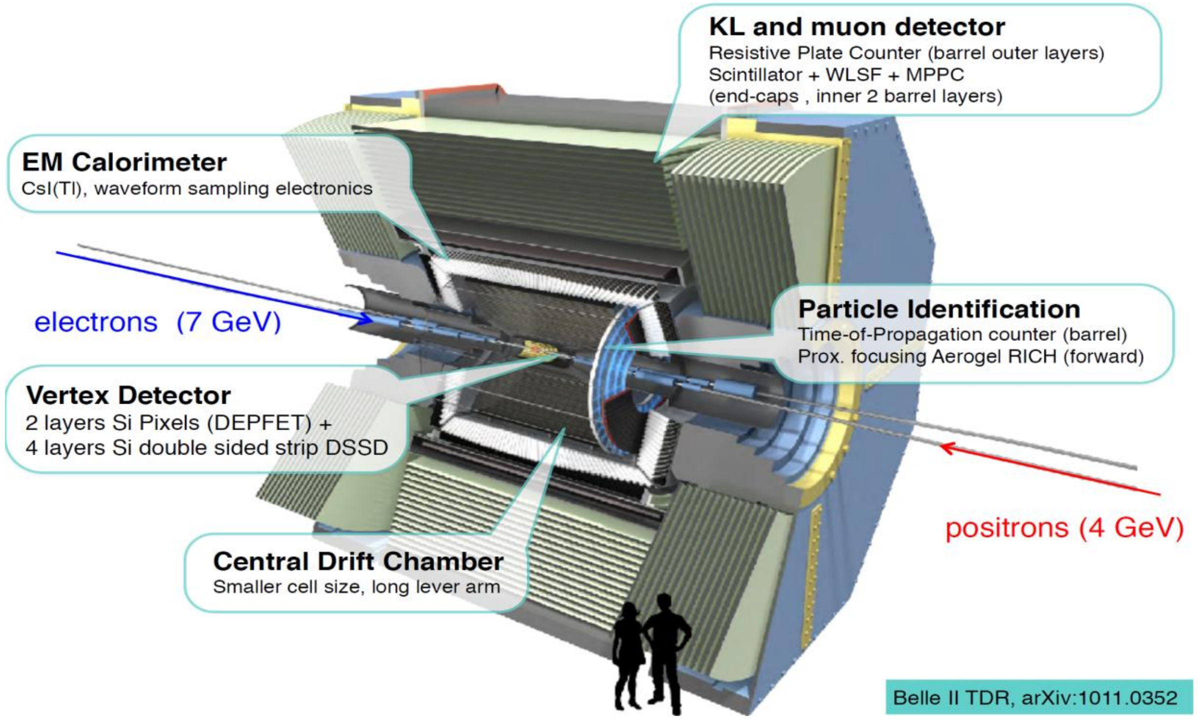




Belle II experiment

- At SuperKEKB e^+e^- collider, in Tsukuba, Japan

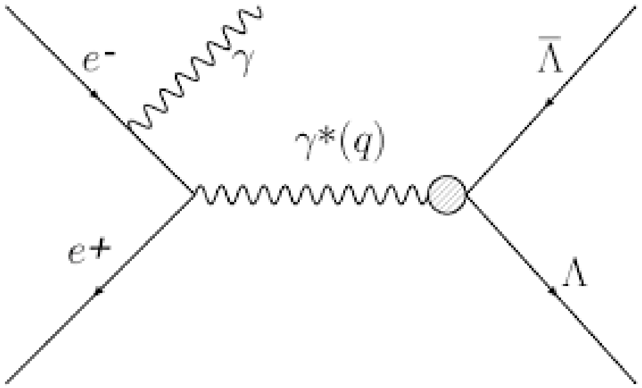
- Set to work at the center-of-mass energy of 10.58 GeV to study the B-meson physics.



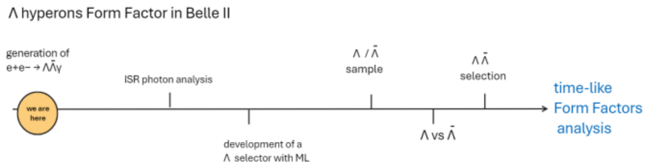
Hyperons in the Belle II experiment

In Belle II the $\Lambda\bar{\Lambda}$ pairs can be generated by:

- $e^+e^- \to e^+e^-\gamma_{ISR} \to \Lambda\bar{\Lambda}\gamma_{ISR}$

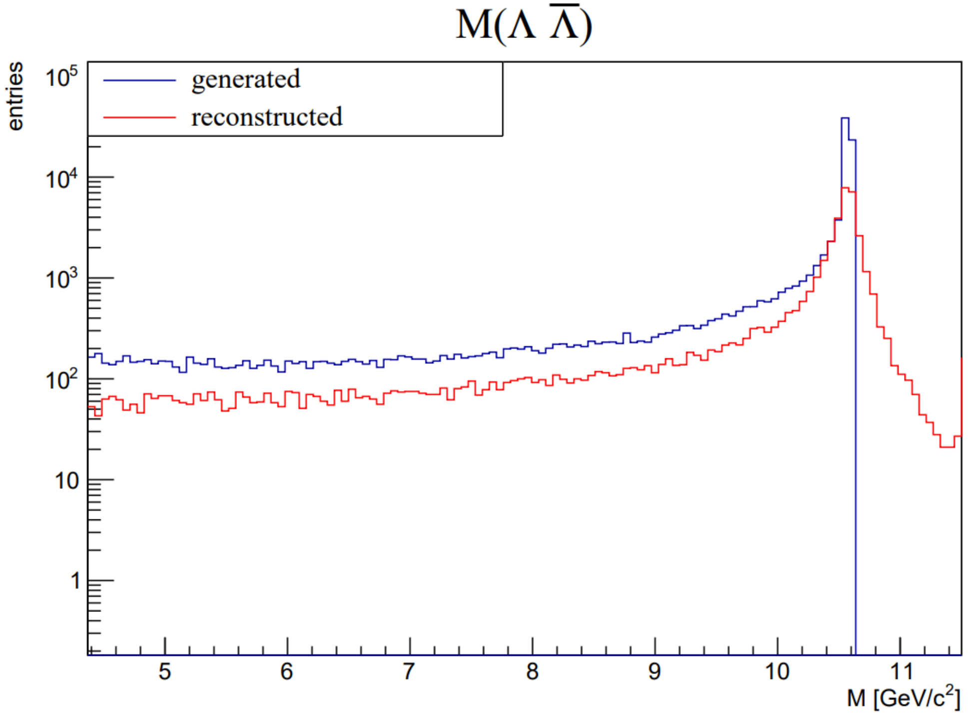


- Initial State Radiation process
  - The e^- or e^+ beam irradiates one photon reducing the effective center-of-mass energy of the annihilation

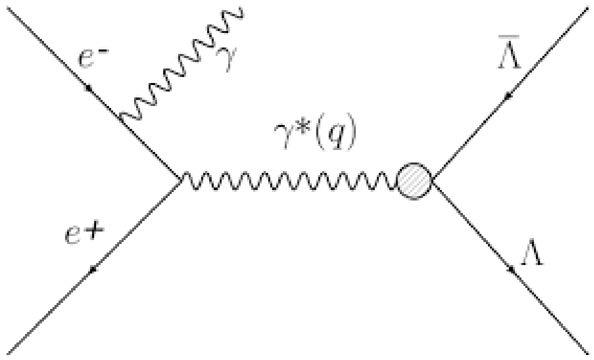


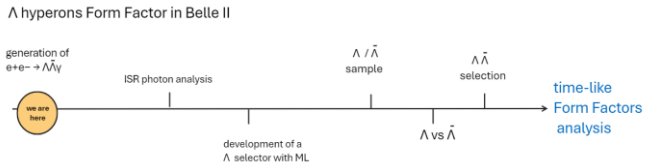
Monte Carlo generation

Generation of: $e^+e^- \rightarrow e^+e^-\gamma_{ISR} \rightarrow \Lambda\bar{\Lambda}\gamma_{ISR}$



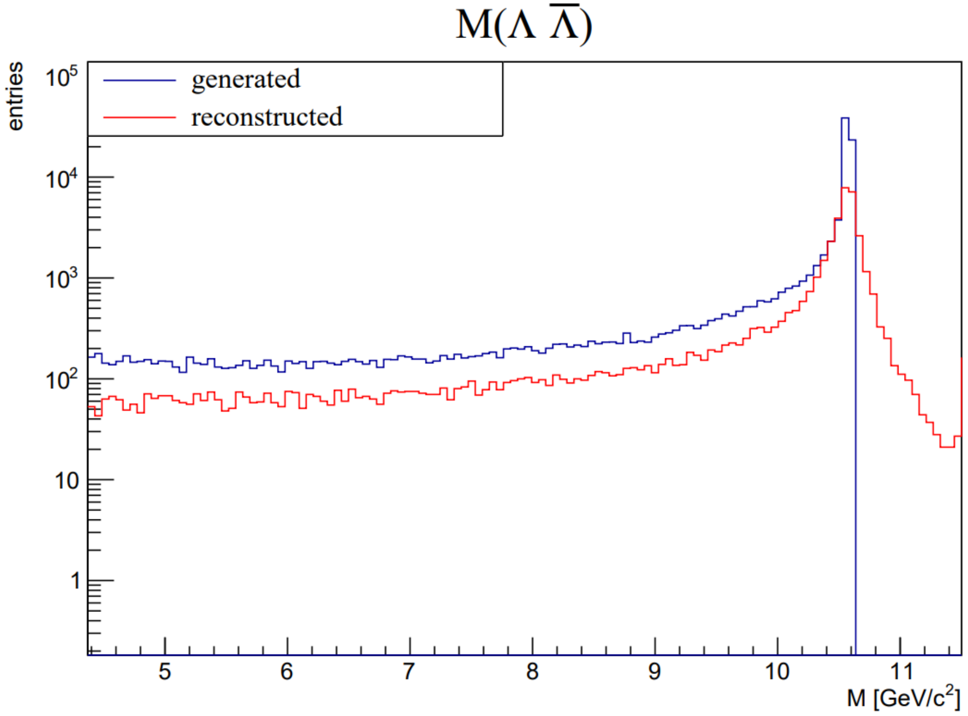
Which particles do we need to reconstruct in the final state in order to identify the full event?



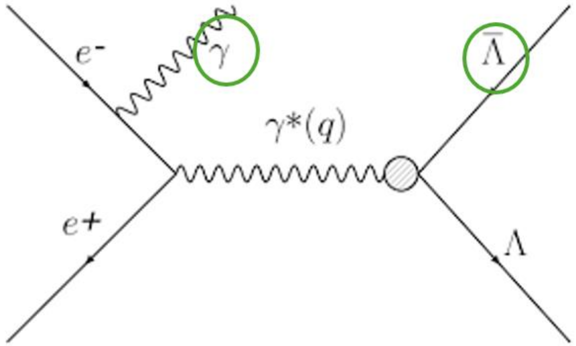


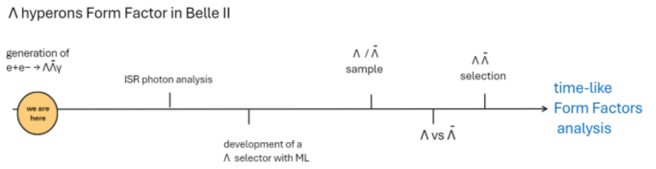
Monte Carlo generation

Generation of: $e^+e^- \to e^+e^-\gamma_{ISR} \to \Lambda\bar{\Lambda}\gamma_{ISR}$



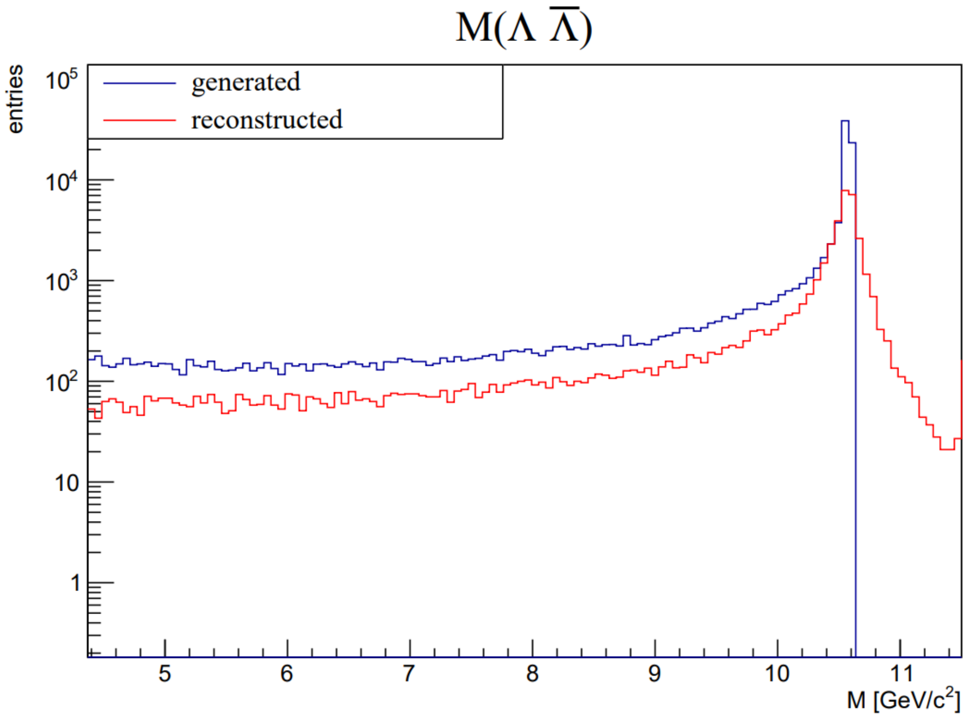
Tagging the ISR photon approach:



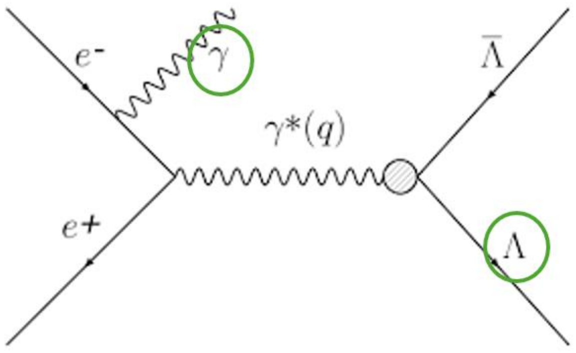


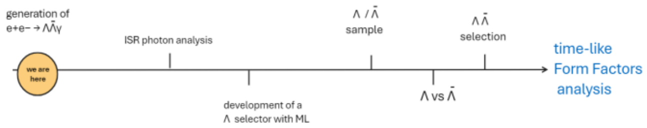
Monte Carlo generation

Generation of: $e^+e^- \to e^+e^-\gamma_{ISR} \to \Lambda\bar{\Lambda}\gamma_{ISR}$



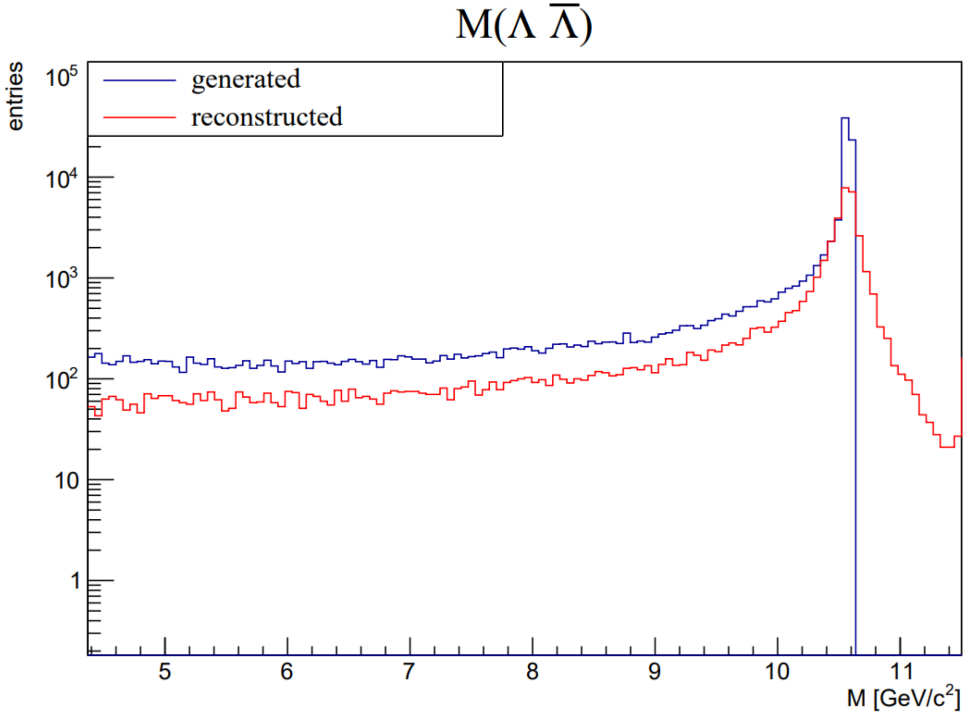
Tagging the ISR photon approach:



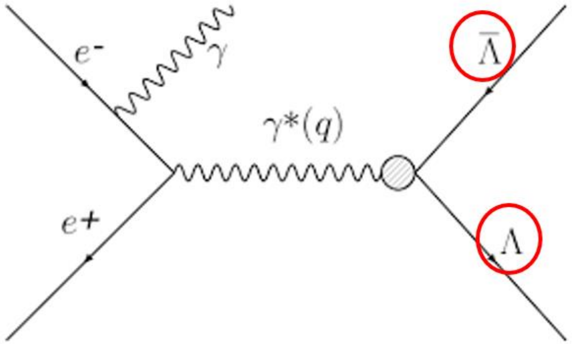


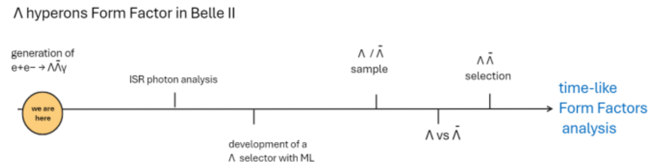
Monte Carlo generation

Generation of: $e^+e^- \rightarrow e^+e^-\gamma_{ISR} \rightarrow \Lambda\bar{\Lambda}\,\gamma_{ISR}$



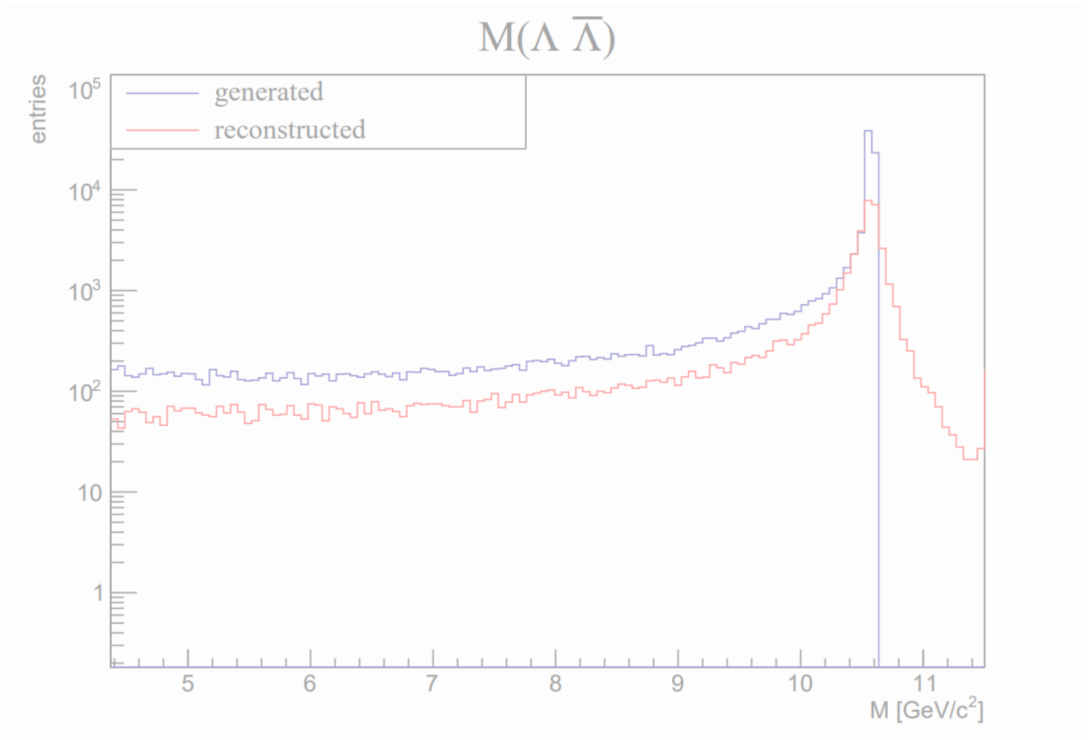
Without tagging the ISR photon:



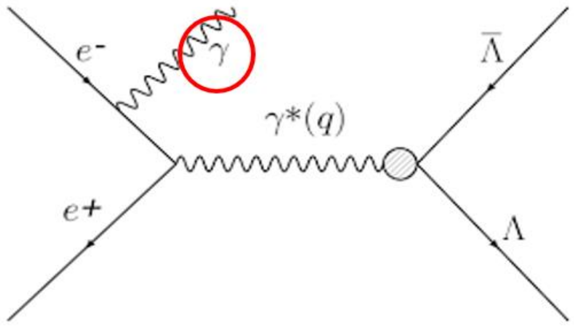


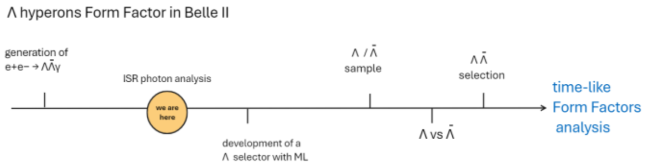
Monte Carlo generation

Generation of: $e^+e^- \to e^+e^-\gamma_{ISR} \to \Lambda\bar{\Lambda}\,\gamma_{ISR}$



How efficiently can we reconstruct the γ_{ISR}?

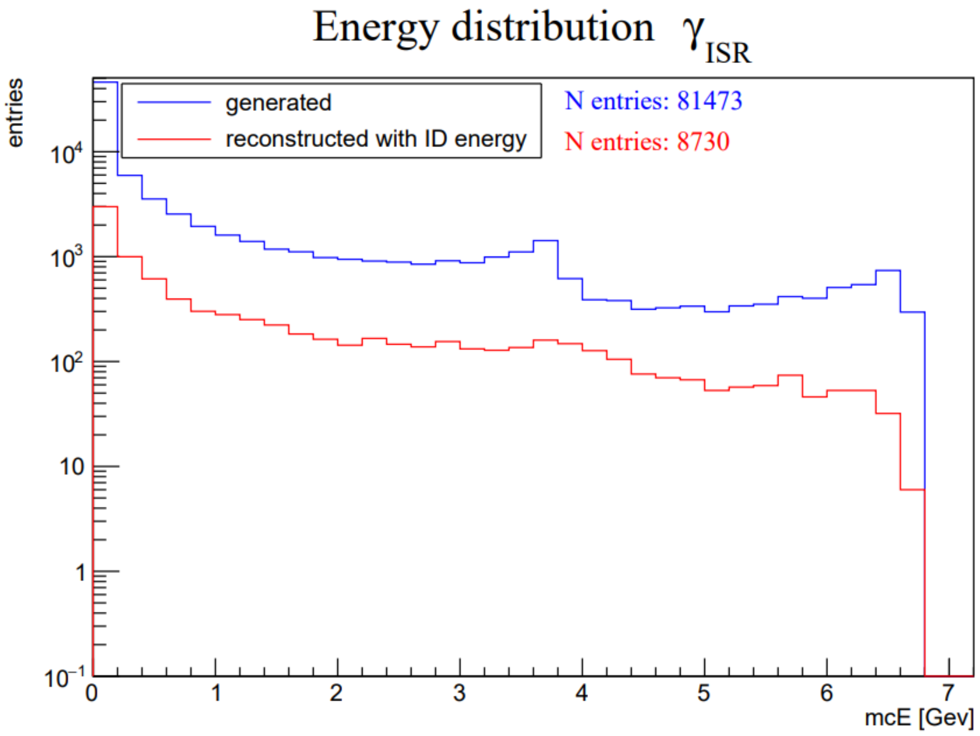




Energy distribution of the ISR photon

Channel: $e^+e^- \to e^+e^-\gamma_{ISR} \to \Lambda\bar{\Lambda}\,\gamma_{ISR}$

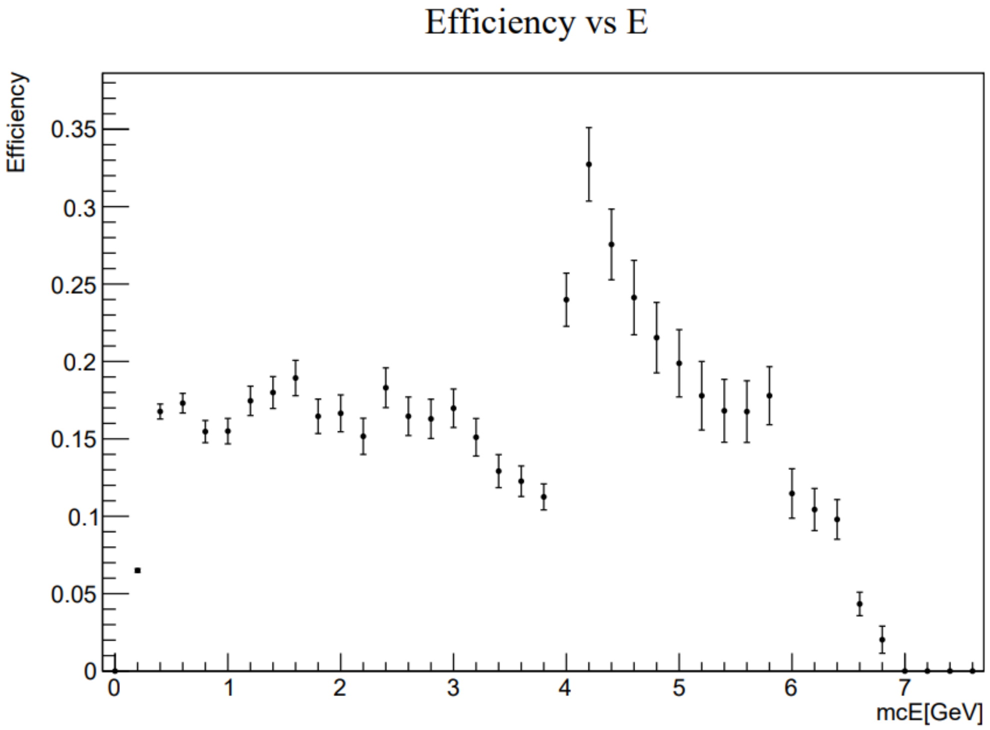
The energy distribution of the γ_{ISR} defines the energy of the recoiling $\Lambda\bar{\Lambda}$ system



- The performance of the reconstruction algorithm may be evaluated thanks to the MC truth.

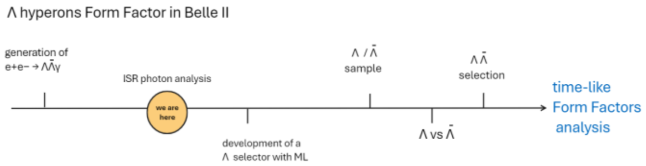
ISR efficiency dependence on energy

Efficiency defined as the ratio of the reconstructed events histogram and the generated event histogram



$$\varepsilon^i = \frac{N^i_{rec}(E_{gen})}{N^i_{gen}(E_{gen})}$$

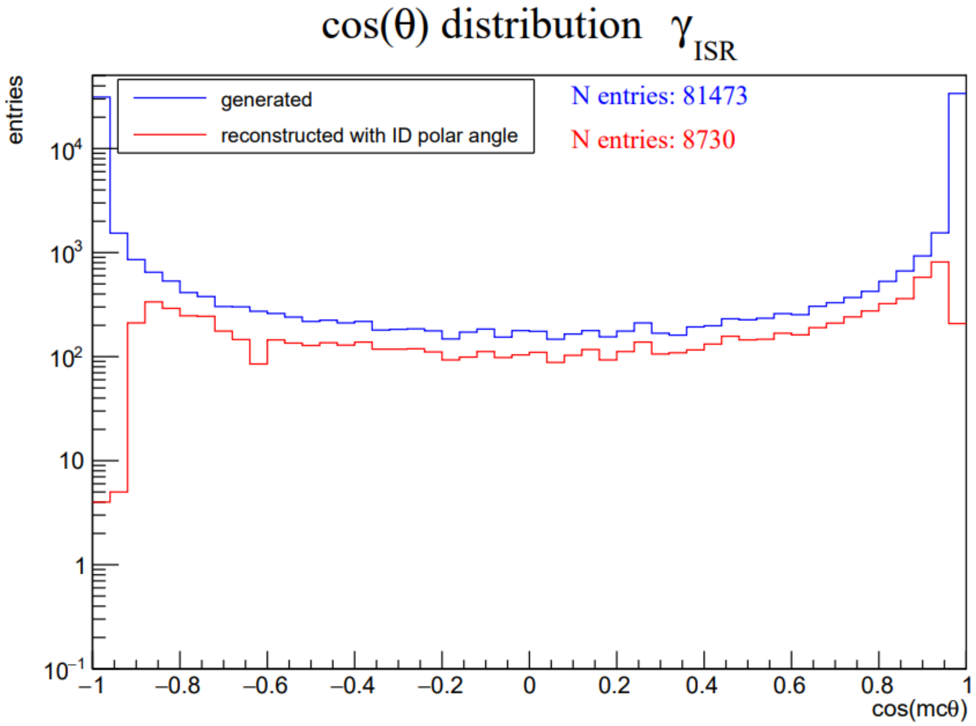
- Efficiency level is below 30% over almost all the energy range
- We can learn at which energies of the ISR photons the detector is missing more events



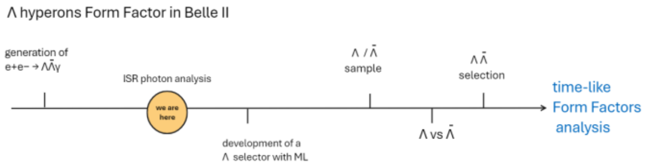
Angular distribution of the ISR photon

Channel: $e^+e^- \rightarrow e^+e^-\gamma_{ISR} \rightarrow \Lambda\bar{\Lambda}\,\gamma_{ISR}$

How well the detector is able to reconstruct the ISR photon along the polar angle

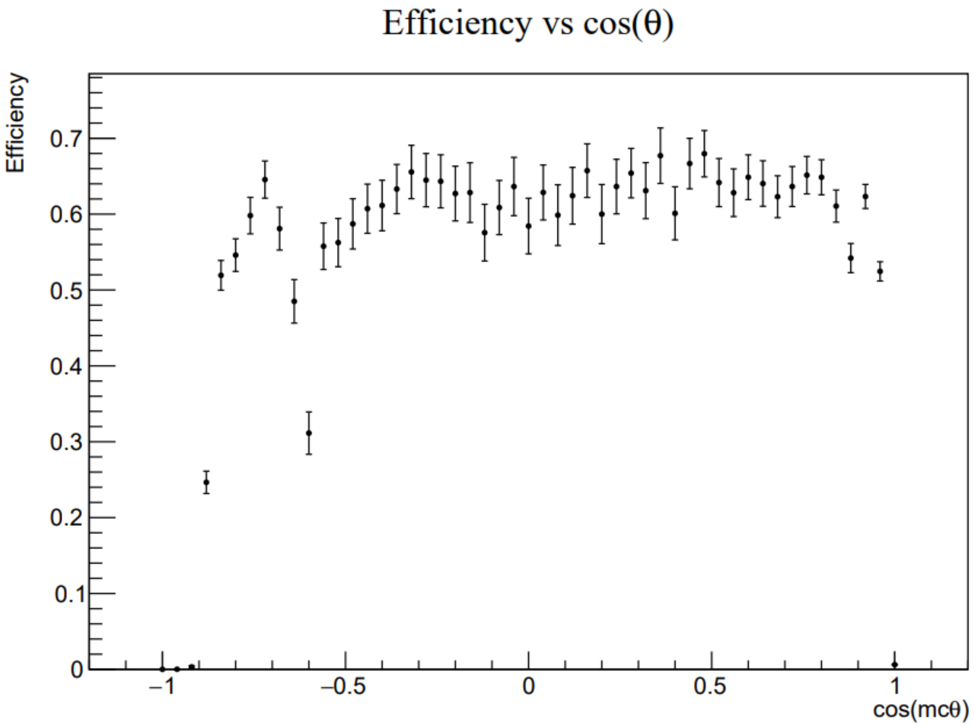


- $\cos\theta$ defined as the cosine of the angle between the momentum direction of the γ_{ISR} and the e^- beam
- Most of the γ_{ISR} are emitted at extreme angles

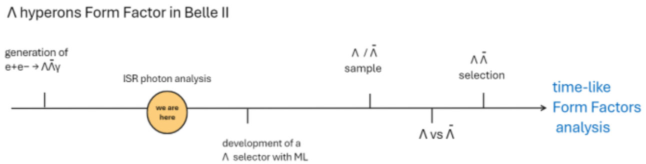


ISR efficiency dependence on angular distribution

Efficiency defined as the ratio of the reconstructed events histogram and the generated event histogram

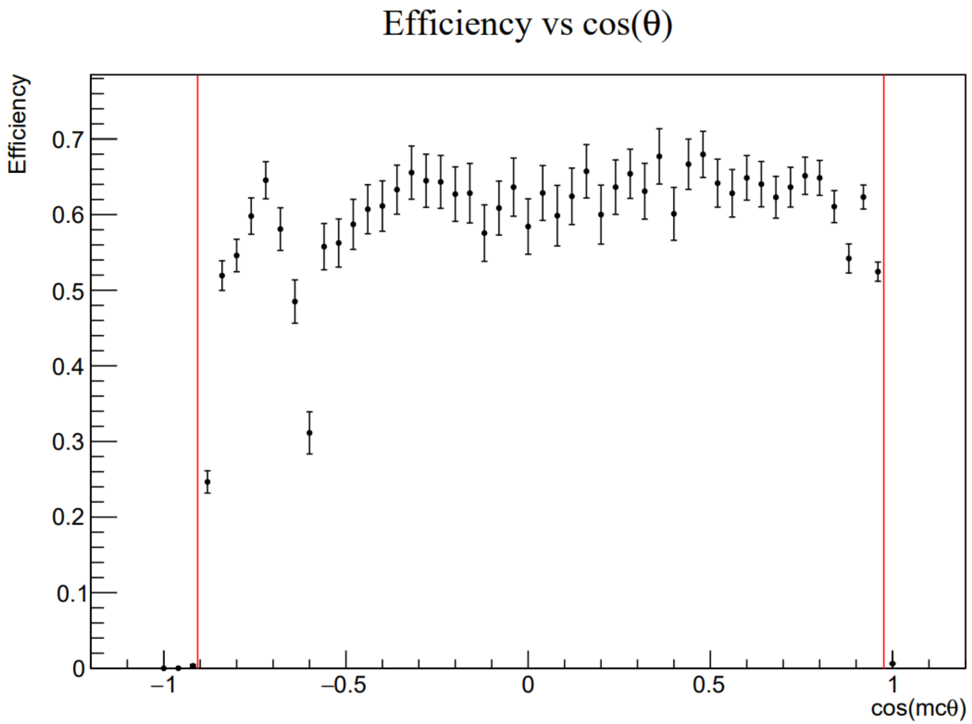


Efficiency	
Full angular range 0° < θ < 180°	Detector angular acceptance 12.4° < θ < 155.1°
10%	55%

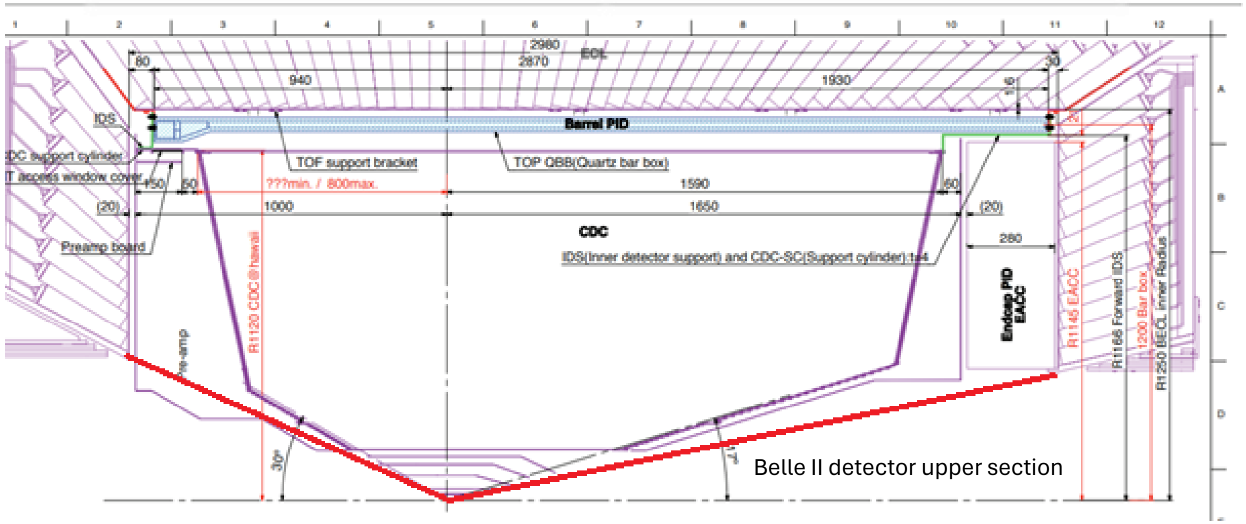


Efficiency dependence on angular distribution

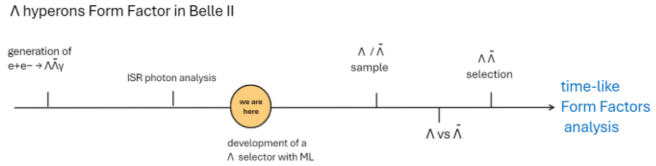
Efficiency defined as the ratio of the reconstructed events histogram and the generated event histogram



Efficiency	
Full angular range $0° < \theta < 180°$	Detector angular acceptance $12.4° < \theta < 155.1°$
10%	55%



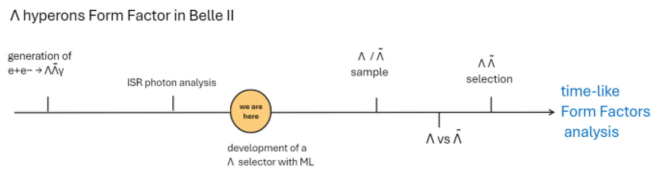
Belle II detector upper section



Selecting Λ events from MC samples

Development of Λ ($\bar{\Lambda}$) selector for Belle II data samples

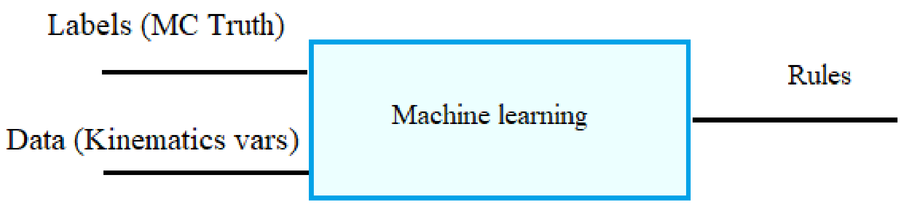
- Machine Learning (ML) based selection algorithm
- Task: select the Λ ($\bar{\Lambda}$) hyperons distinguishing them from the background

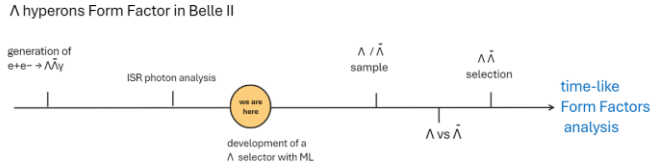


Machine Learning approach

ML uses the statistics to enable machines to recognize patterns by learning and through experience on a set of provided data.

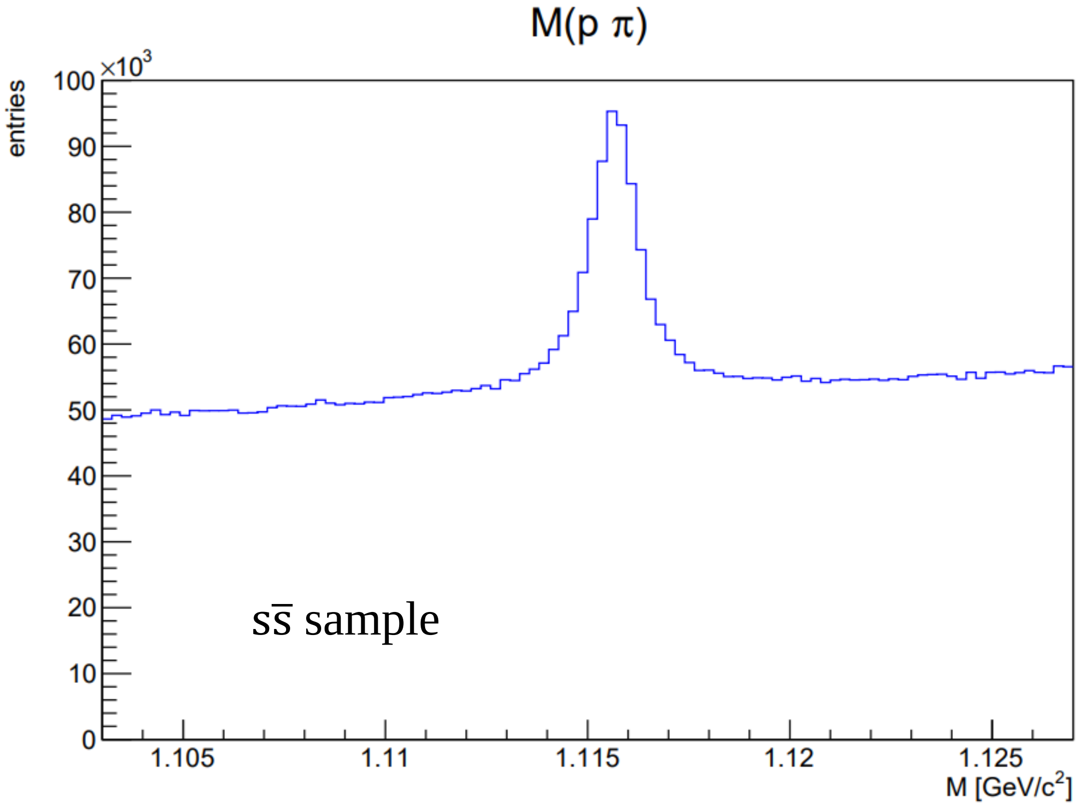
- Supervised ML tool uses data and labels to discover the rules behind a problem.
  - ✓ Appropriate for a classification problem

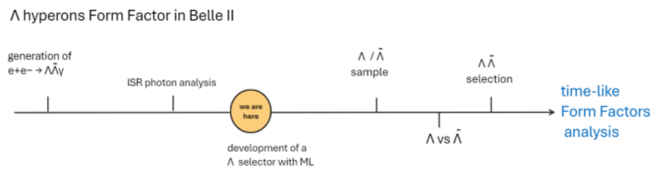




ML analysis: data collection

- $s\bar{s}$ hadronic MC sample with L = 100 fb^{-1}
- Training sample: L = 90 fb^{-1}
- Test sample: L = 10 fb^{-1}





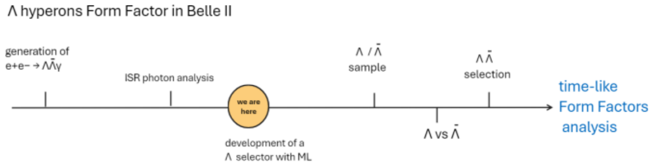
ML analysis: features extraction

List of kinematics variables used in my analysis:

- Proton ID
- Pion ID
- Angle between two Λ daughters: Φ
- Pion momentum: p_{π}
- Proton momentum: p_{p}
- Angle between vertex vector and reconstructed Λ momentum: ξ
- Flight distance significance: FDS

$$\text{proton ID} = \frac{L_p}{\sum_i L_i}$$

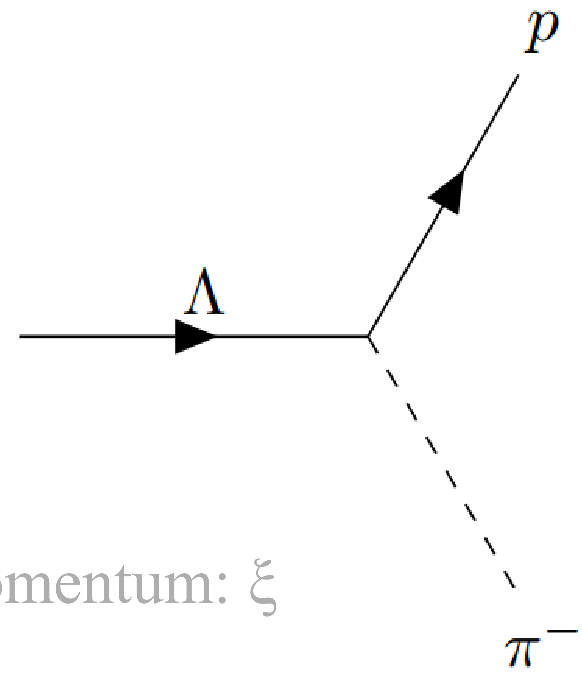
$$\text{pion ID} = \frac{L_{\pi}}{\sum_i L_i}$$

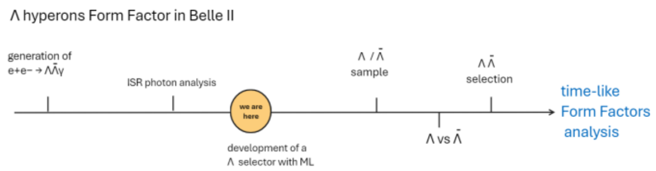


ML analysis: features extraction

List of kinematics variables used in my analysis:

- Proton ID
- Pion ID
- Angle between two Λ daughters: Φ
- Pion momentum: p_{π}
- Proton momentum: p_{p}
- Angle between vertex vector and reconstructed Λ momentum: ξ
- Flight distance significance: FDS

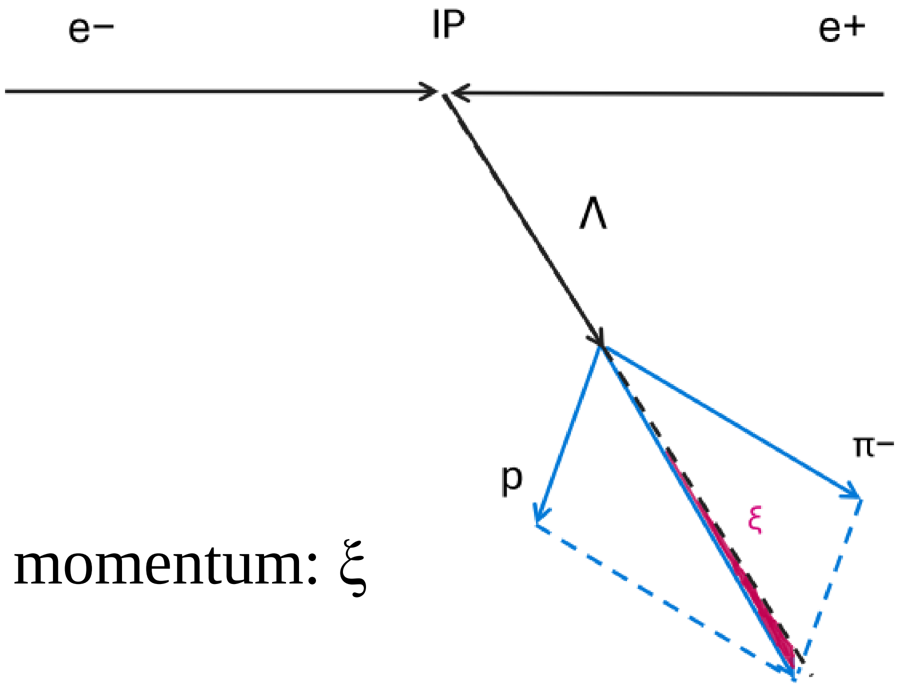


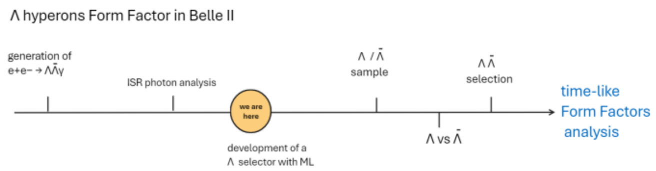


ML analysis: features extraction

List of kinematics variables used in my analysis:

- Proton ID
- Pion ID
- Angle between two Λ daughters: Φ
- Pion momentum: p_{π}
- Proton momentum: p_{p}
- Angle between vertex vector and reconstructed Λ momentum: ξ
- Flight distance significance: FDS



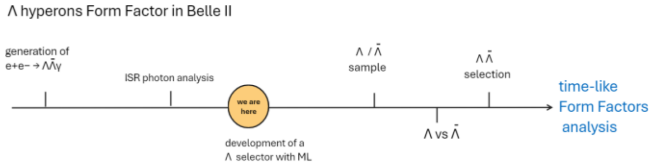


ML analysis: features extraction

List of kinematics variables used in my analysis:

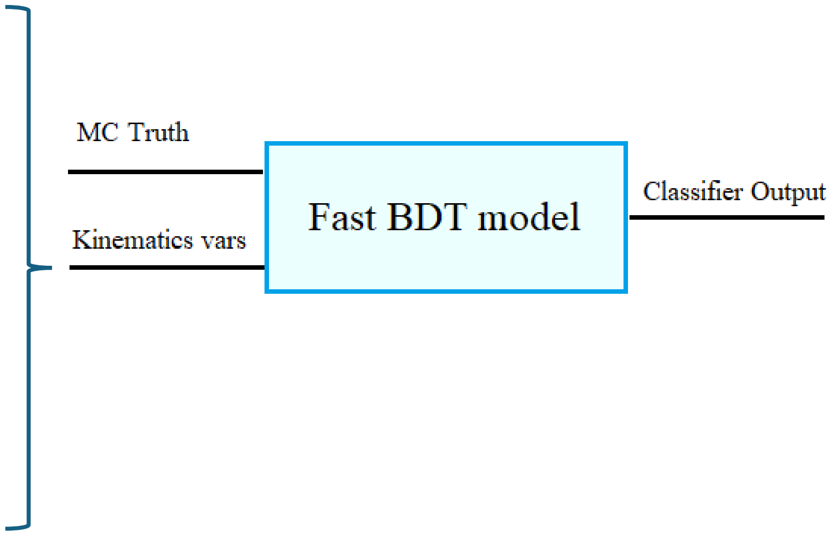
- Proton ID
- Pion ID
- Angle between two Λ daughters: Φ
- Pion momentum: p_π
- Proton momentum: p_p
- Angle between vertex vector and momentum: ξ
- Flight distance significance: FDS

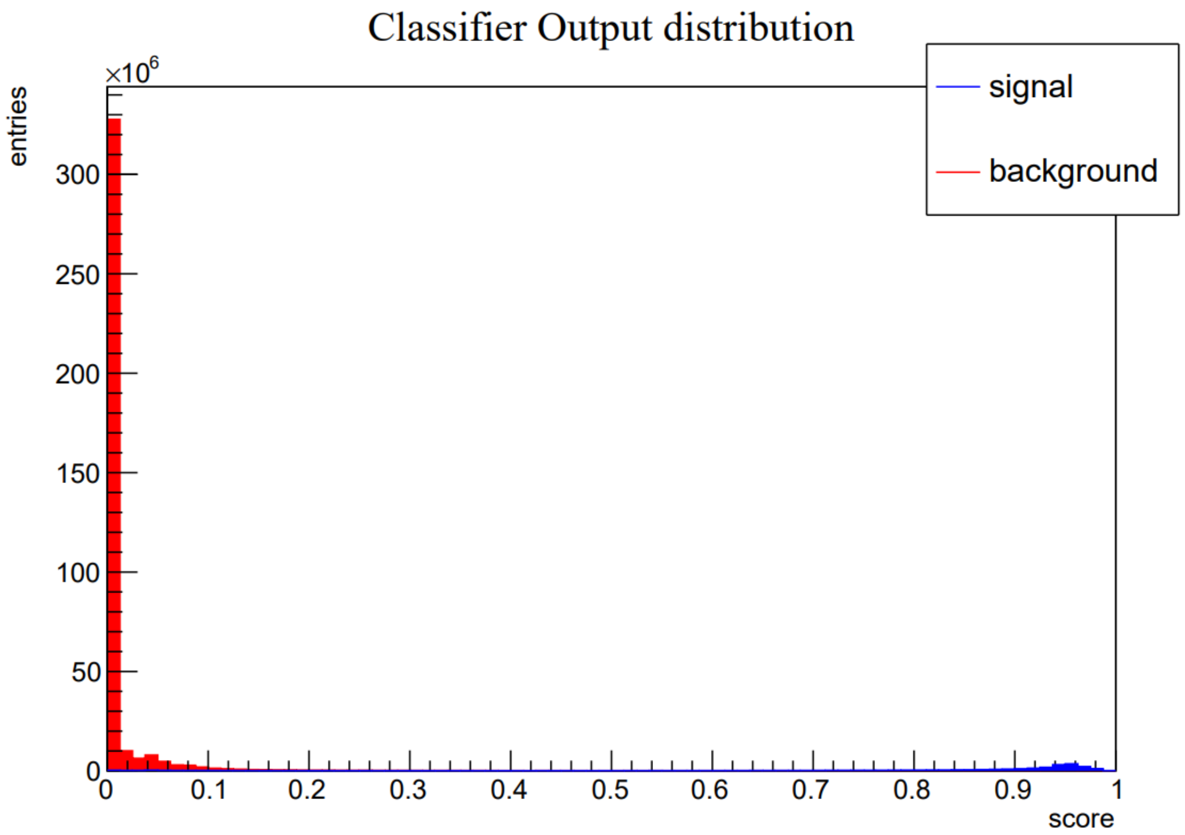
$$FDS = \frac{Flight\ distance}{\sigma_{Flight\ distance}}$$

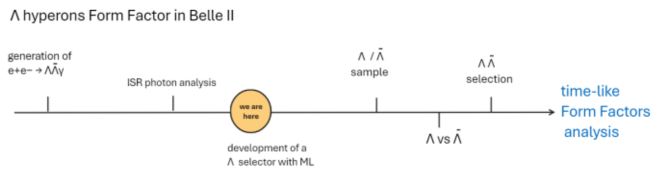


ML analysis: training & test

- Proton ID
- Pion ID
- Φ
- p_{π}
- p_{p}
- ξ
- FDS





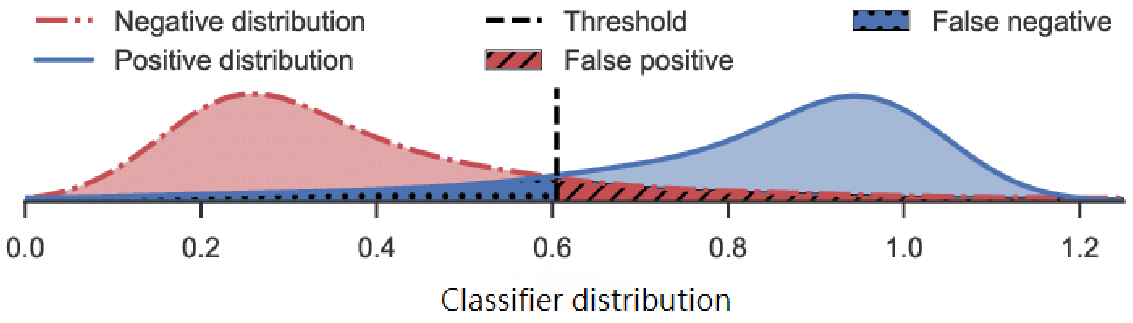


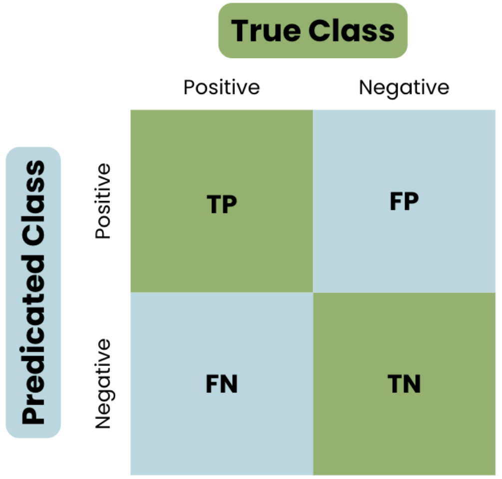
ML analysis: evaluation

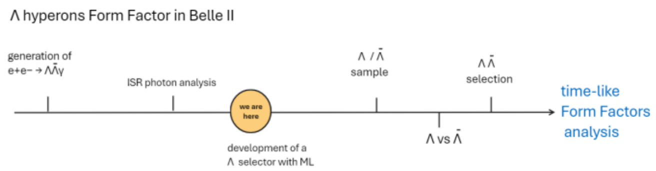
$$signal\ efficiency = \frac{TP}{TP + FN}$$

$$background\ rejection = \frac{TN}{TN + FP}$$

$$purity = \frac{TP}{TP + FP}$$

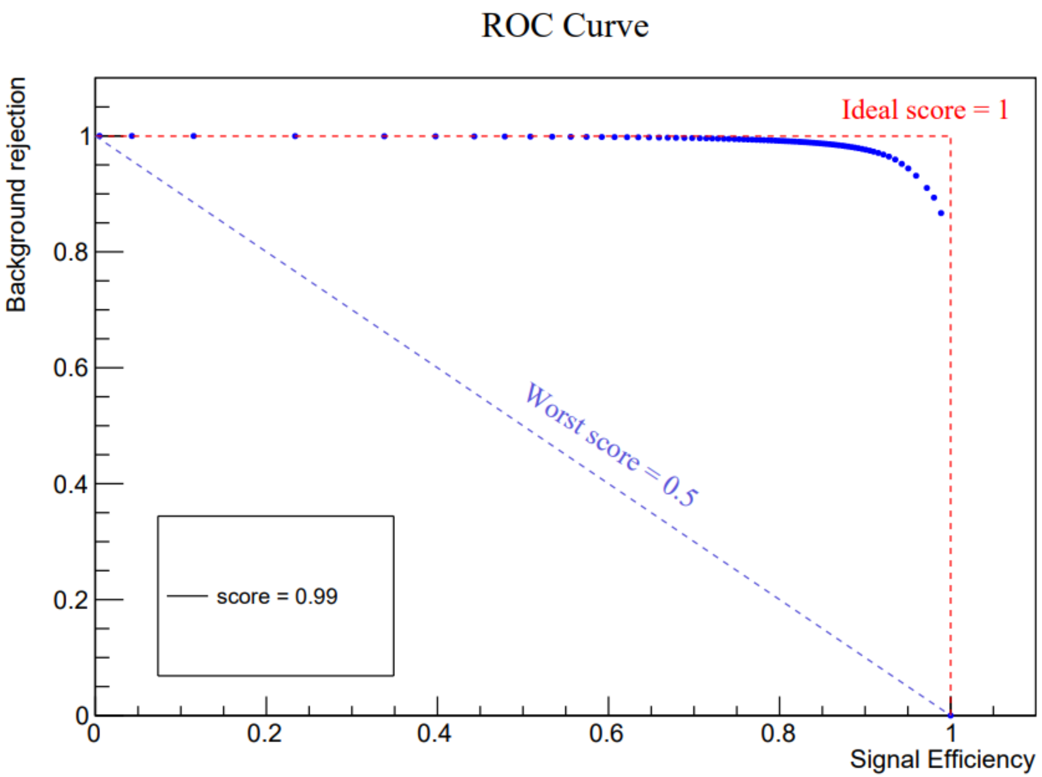


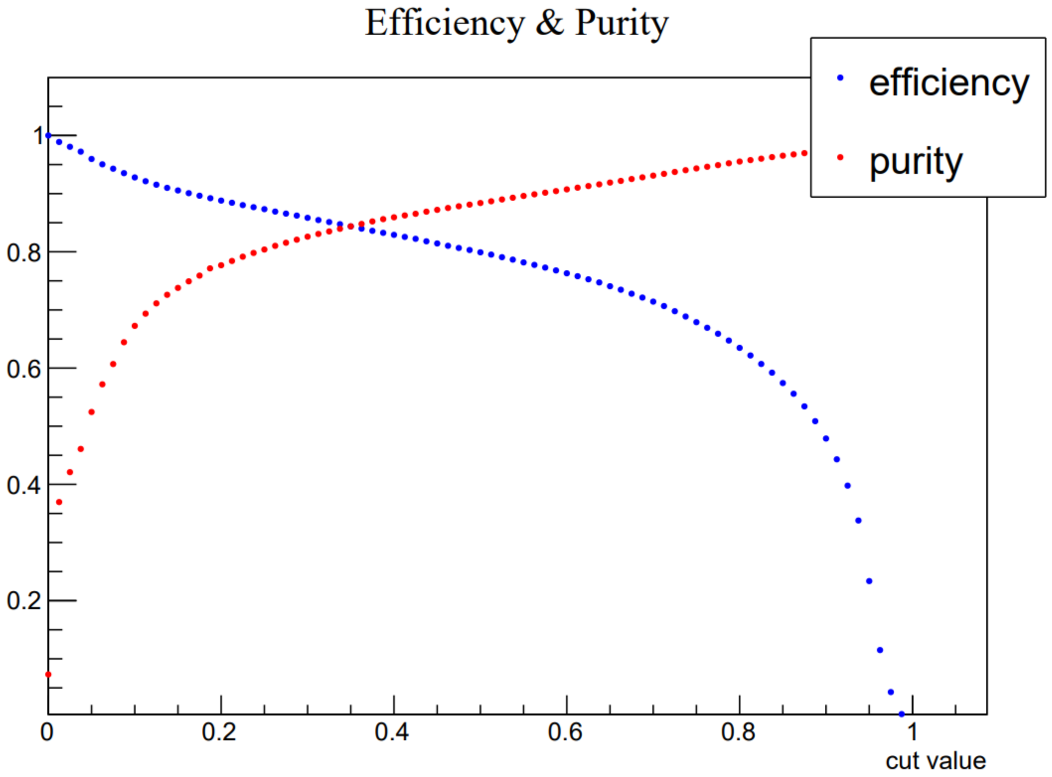


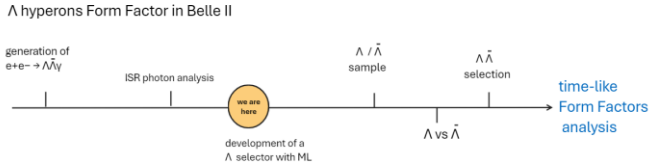


ML analysis: evaluation

Model: Fast Boosted Decision Trees (Fast BDT) with 200 trees



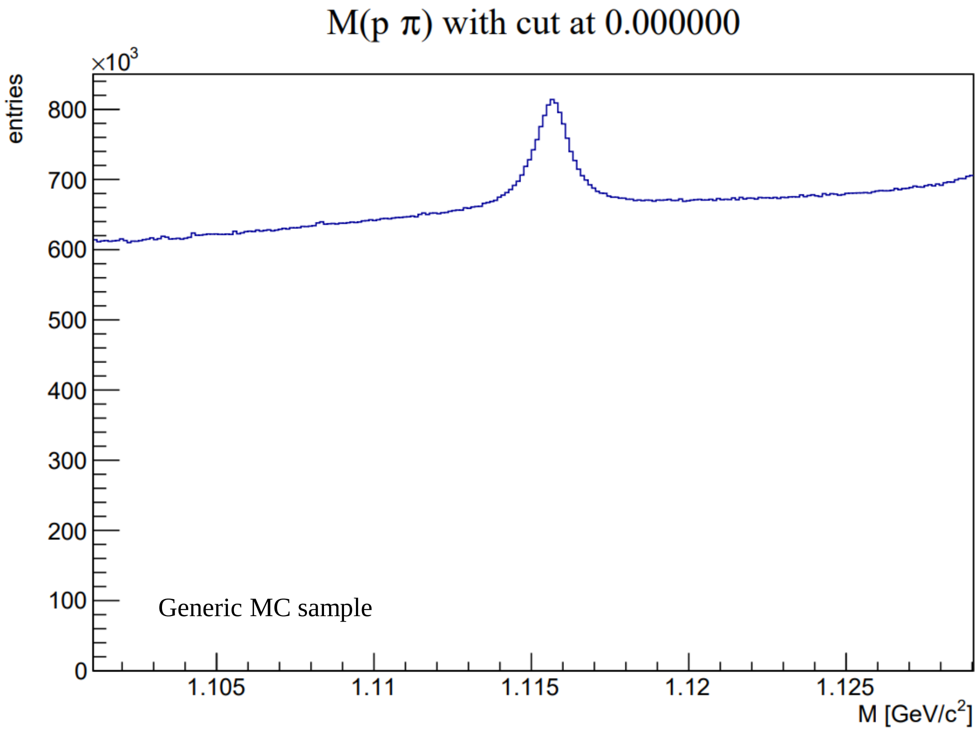




Which cut value is the best?

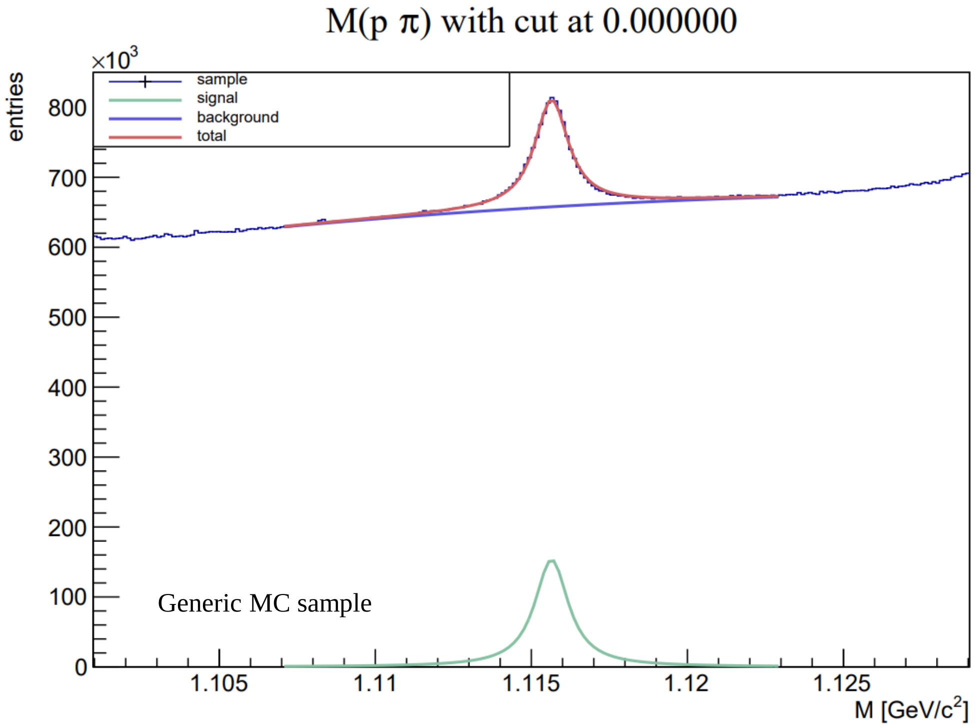
The best cut value choice is the one that maximizes a **Figure of Merit (*FoM*)**

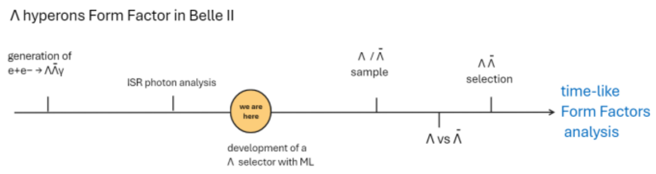
$$\boldsymbol{FoM} = \frac{\boldsymbol{N_{signal}}}{\sqrt{\boldsymbol{N_{signal}} + \boldsymbol{N_{background}}}}$$



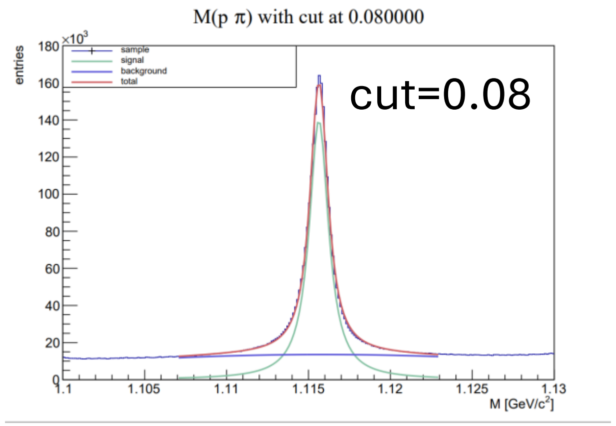
FoM evaluation

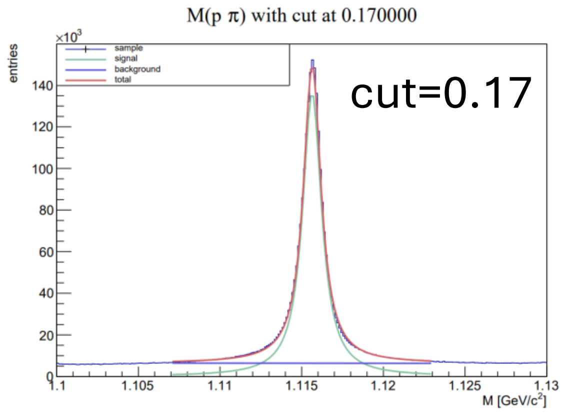
- Fit on the M(pπ)
- Signal is the area under the peak
- Background is the area further the peak

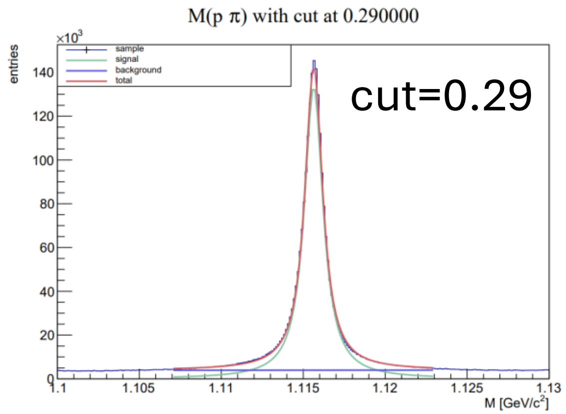


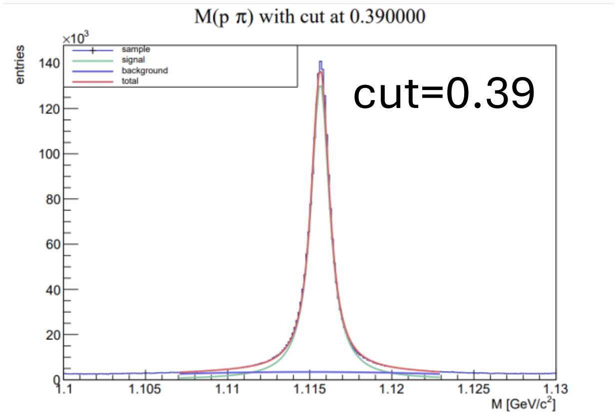


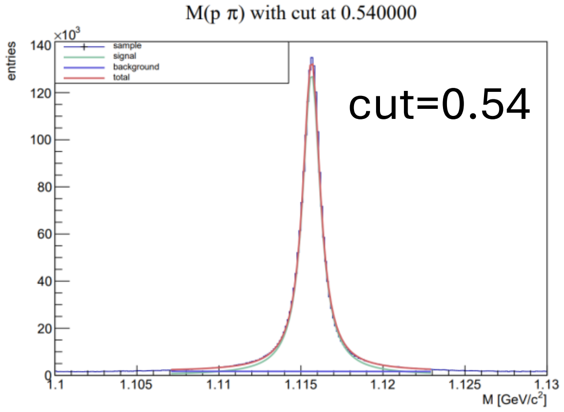
FoM evaluation

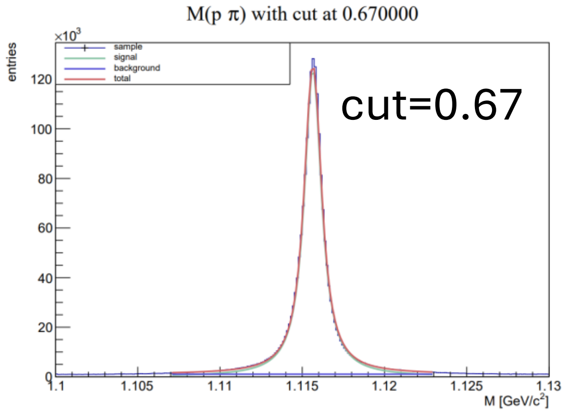


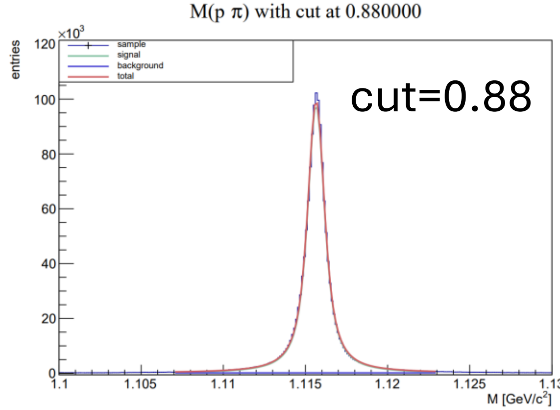


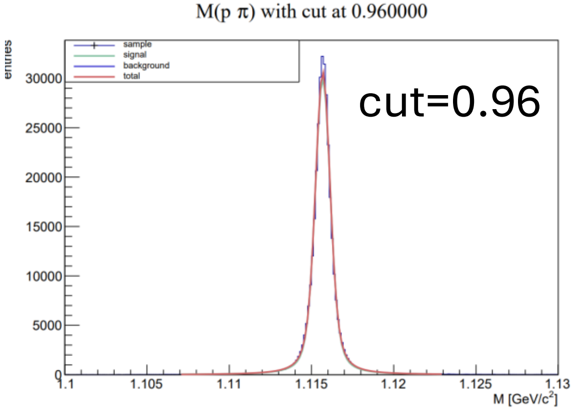


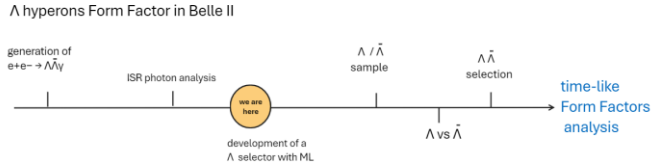








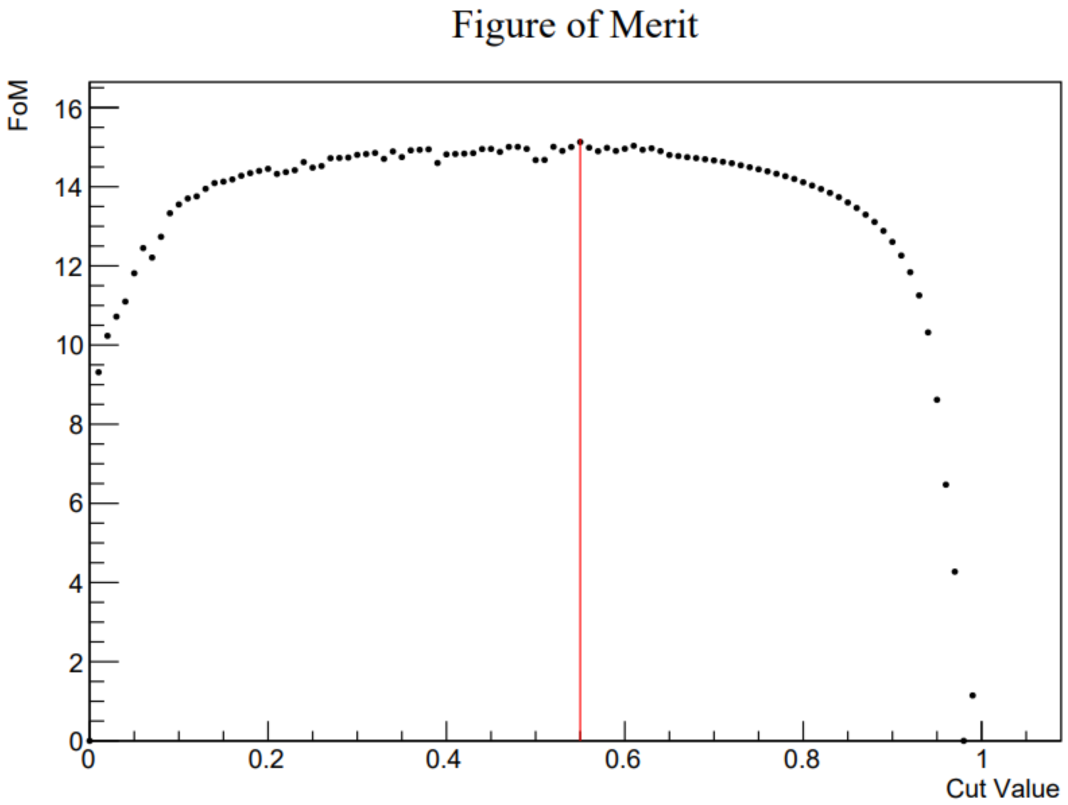


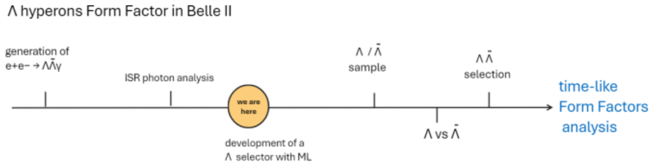


Cut value optimization

- Max value of FoM is at 0.55

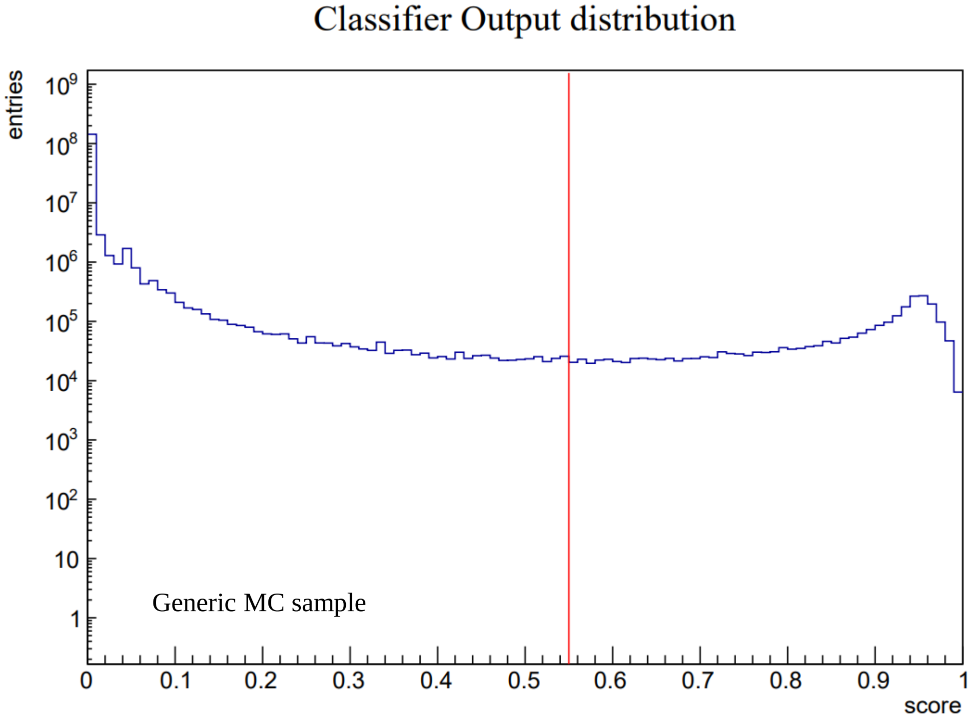
- The prediction over the threshold is classified as signal by the ML algorithm.

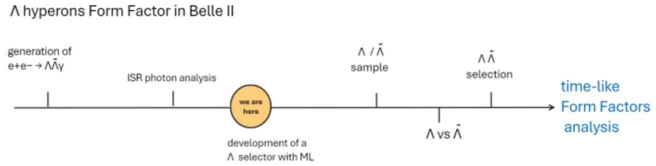




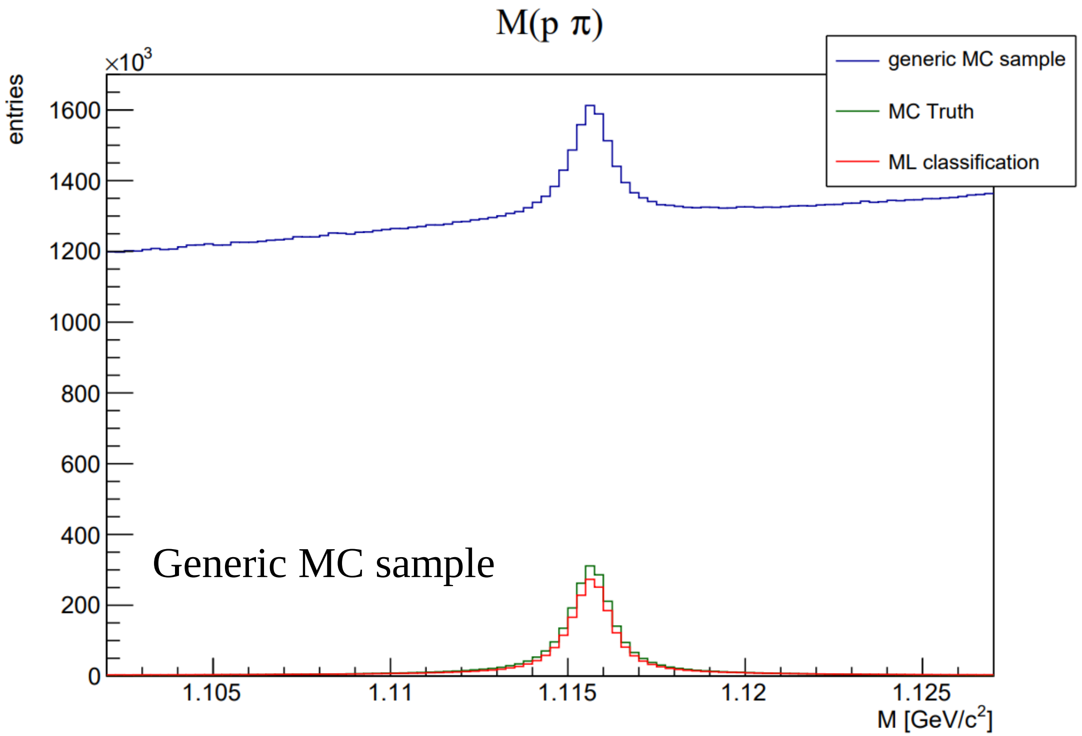
Final step: application of the model

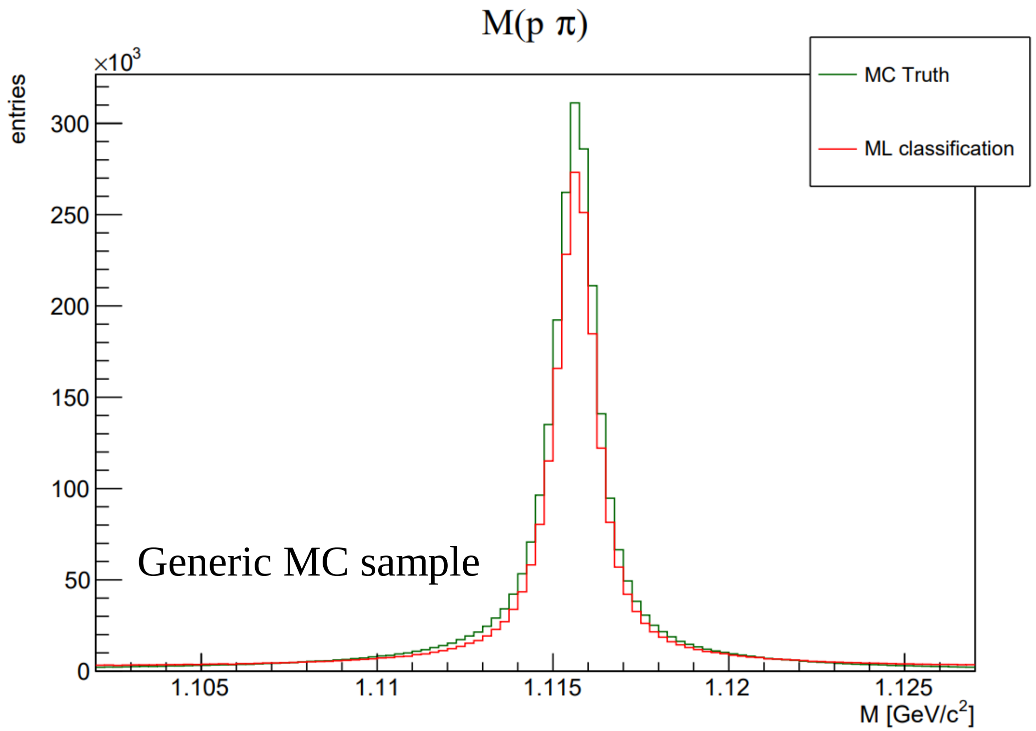
- Dataset: independent generic hadronic MC sample with $\mathrm{L} = 60fb^{-1}$
- Application of 0.55 on the classifier output distribution
- The datapoints with prediction over the threshold are classified as signal





Results and performance of the Λ hyperon selector



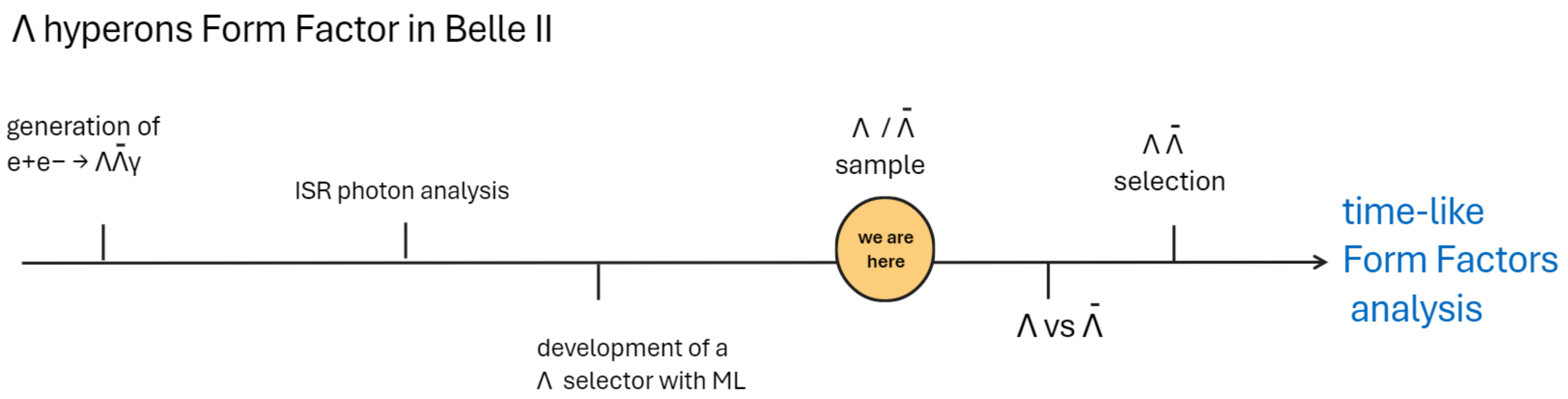


Final performance: signal efficiency = 77%

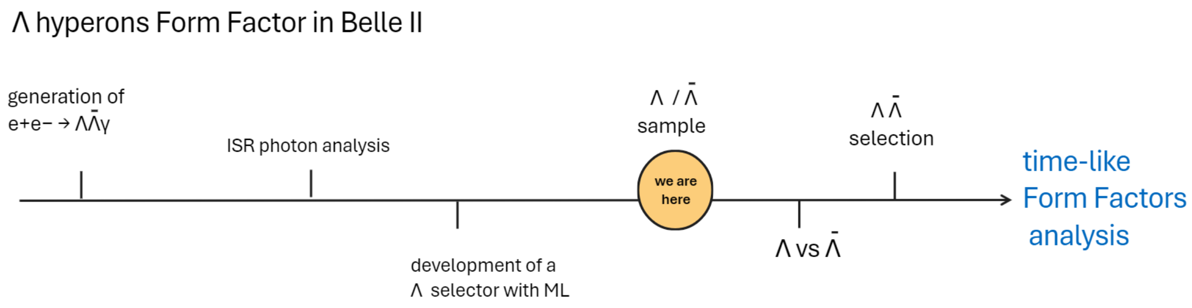
purity = 88%

A similar work has been conducted using a “classical approach” with:
signal efficiency = 65%
purity = 86%

Conclusions and future steps



References



- Karin Schönning, "Production and decay of polarized hyperon-antihyperon pairs", Chinese Physics C (2023).
- Bianca Scavino, "Development of Λ baryons reconstruction and its application to the search for a stable hexaquark at Belle II", PhD thesis, University of Mainz.
- Viktor Thoren, "Hadron physics in a polarized world:exploring electromagnetic interaction with spin Observables", PhD thesis, Uppsala University.
- Elisabetta Perotti, "Electromagnetic and spin properties of hyperons", PhD thesis, Uppsala University.
- J.Pettersson, "From Strange to Charm: meson production in electron positron collision", PhD thesis, Uppsala University.
- Chinese physics C (August 2014), "An exclusive event generator for e+e- scan experiments".
- E. Kou et al. "The Belle II Physics Book", In Progress of Theoretical and Experimental Physics (Dec 2019).
- T. Abe et al. "Belle II Technical design report", 2010.

Backup slides

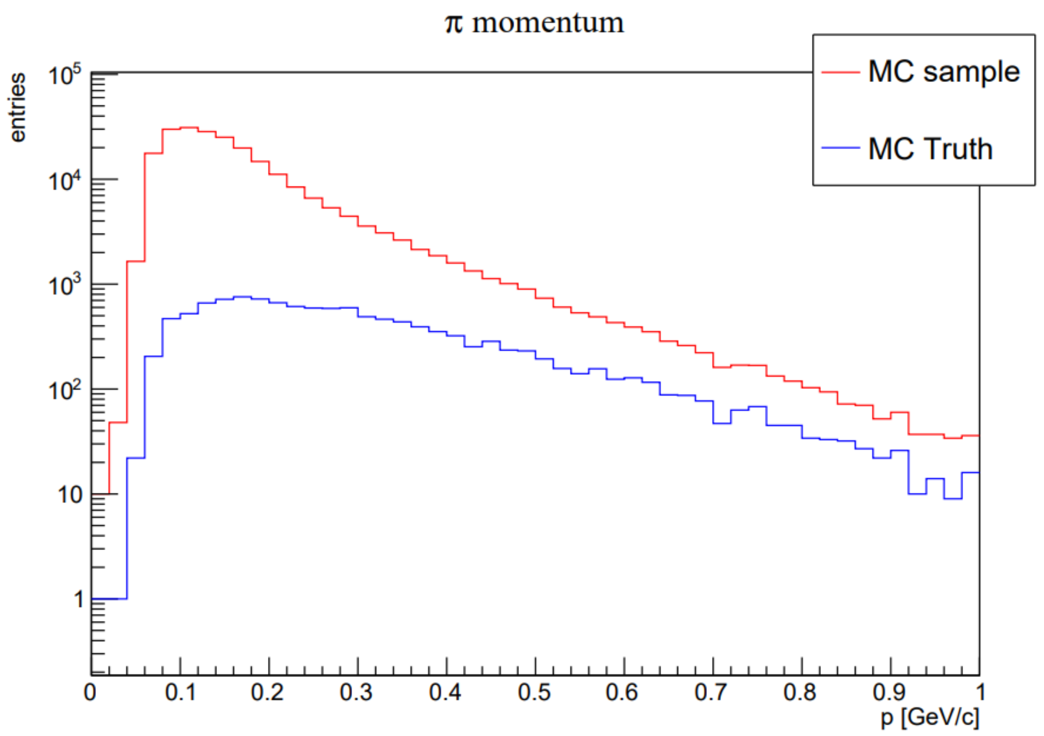
MVA analysis: training & test

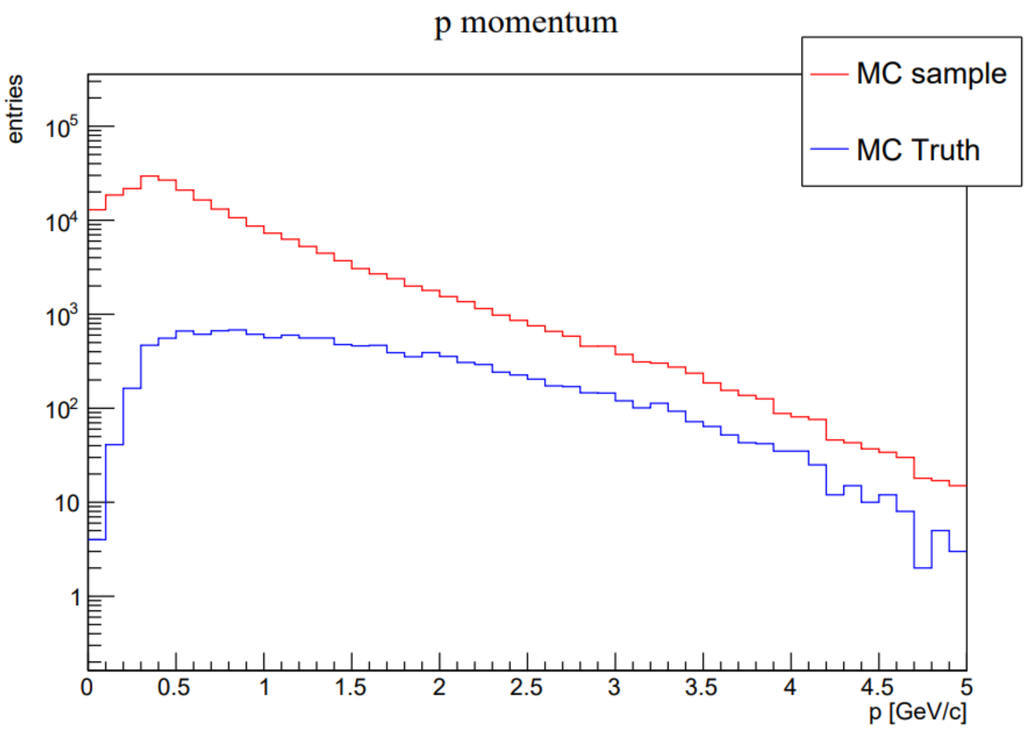
Training and test evaluated on a $s\bar{s}$ hadronic sample with L=90fb^{-1} and L=10 fb^{-1}

- Variable ranking is a tool available on the Belle II framework (basf2)

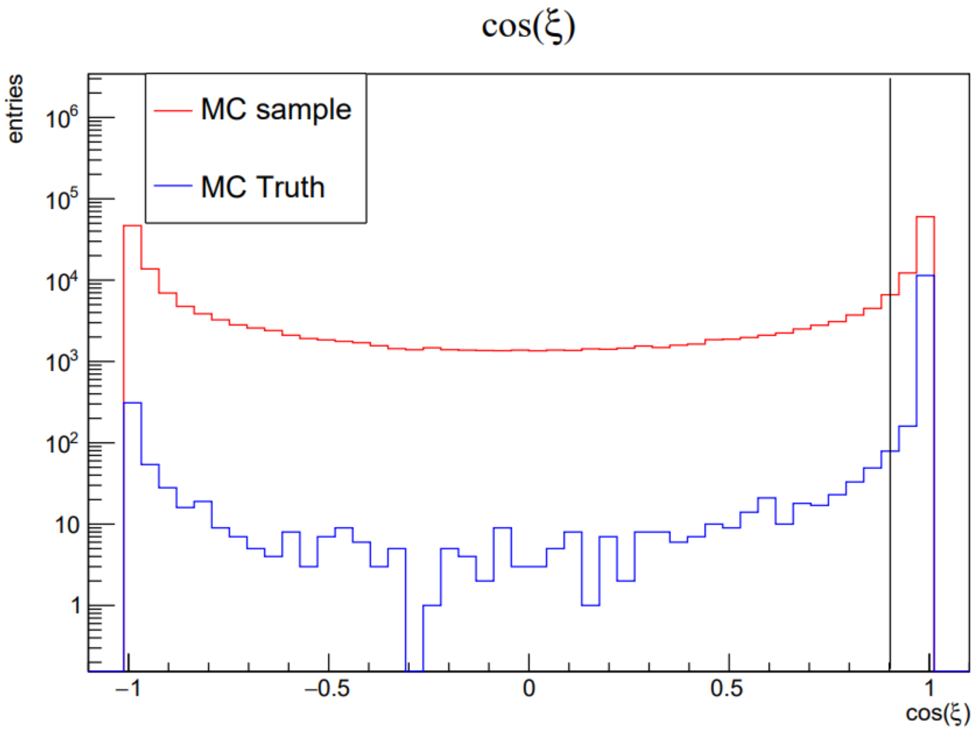
Variable ranking	importance
Proton ID	100
ξ	21
FDS	14
Pion ID	8
p_{π}	7
$p_{\pi,x}$	6
$p_{\pi,y}$	4
p_{p}	3
Φ	0

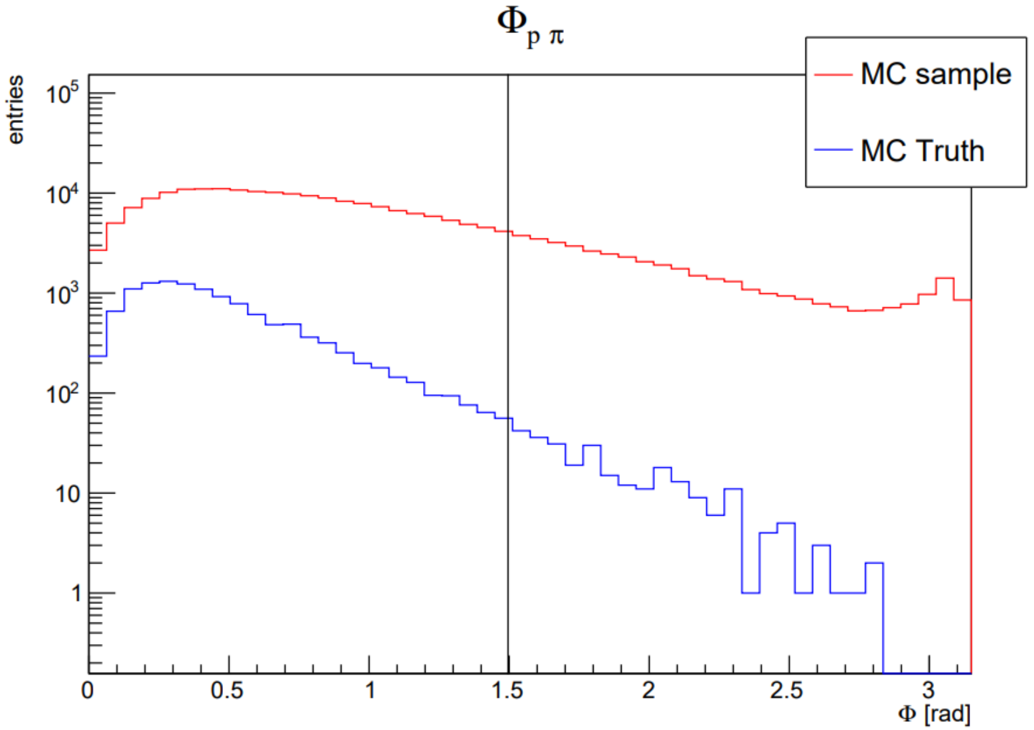
Kinematics variable distributions





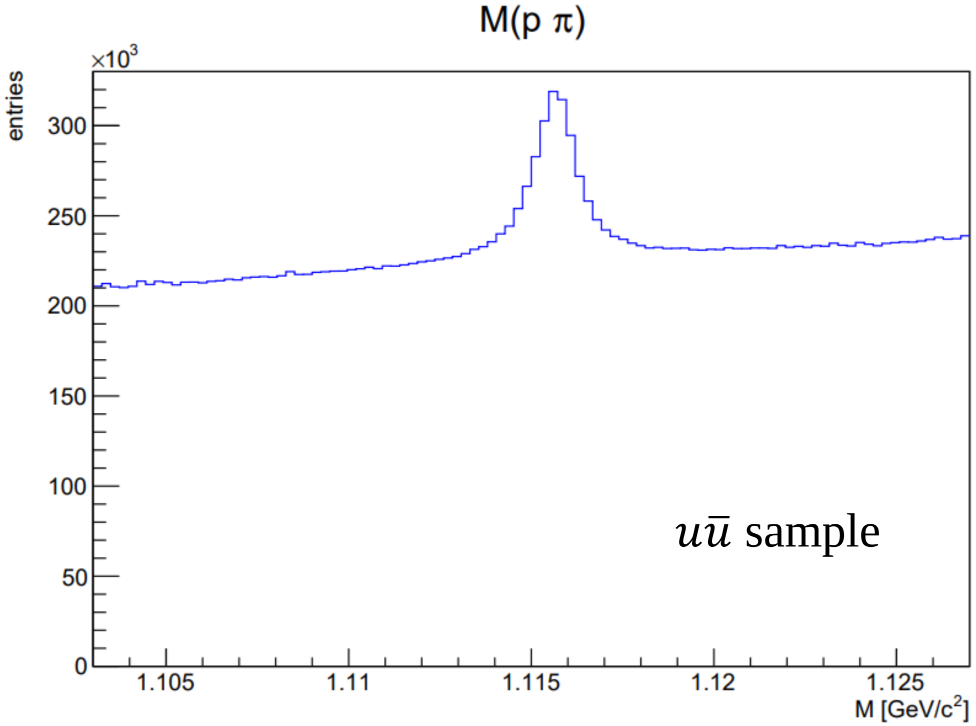
Kinematics variable distributions





Generation of the Monte Carlo samples

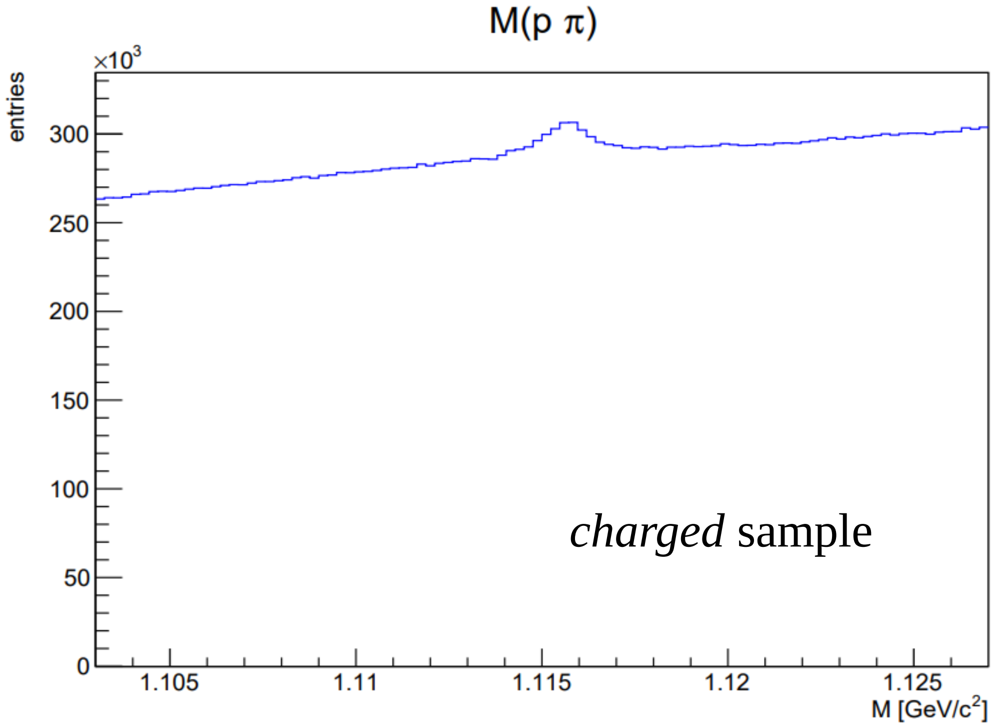
Generic hadronic Monte Carlo samples:



- Continuum: $q\bar{q}$ *with q=u,d,s,c*
- Charged: B^+B^- production
- Mixed: $B^0\bar{B}^0$ production

Generation of the Monte Carlo samples

Generic hadronic Monte Carlo samples:

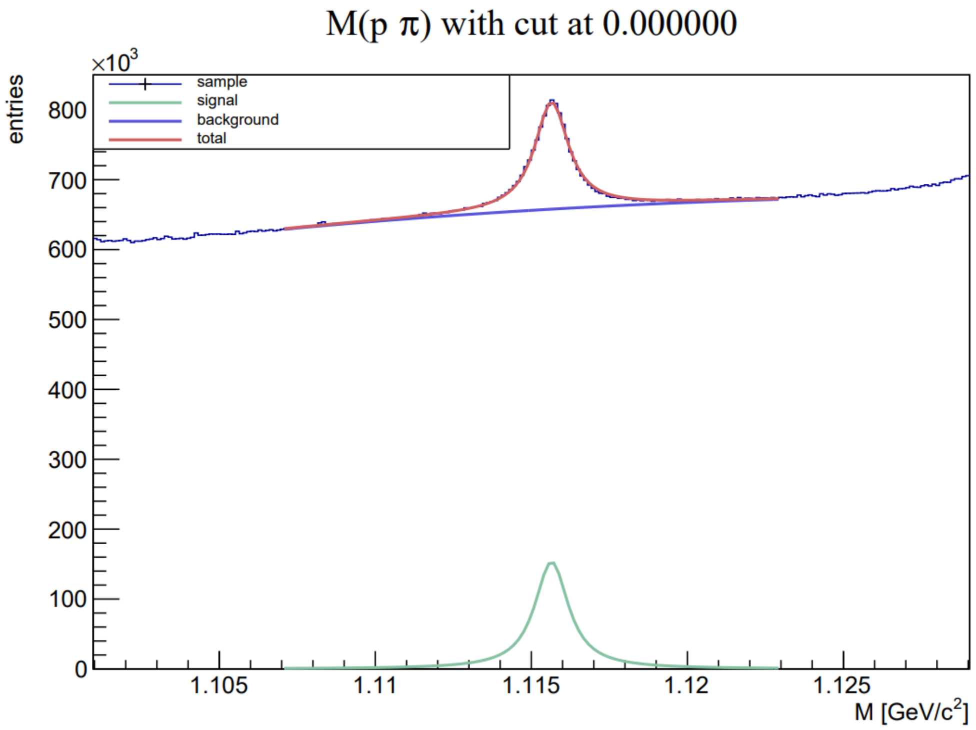


- Continuum: $q\bar{q}$ *with q=u,d,s,c*
- Charged: B^+B^- production
- Mixed: $B^0\bar{B}^0$ production

Cut value choice

- $N_{signal} = \int_{1.107GeV/c^2}^{1.123GeV/c^2} \left(Voigt(x) + pol2(x)\right)dx - \int_{1.107GeV/c^2}^{1.123GeV/c^2} pol2(x)dx$

- $N_{background} = \int_{1.107GeV/c^2}^{1.123GeV/c^2} pol2(x)dx$

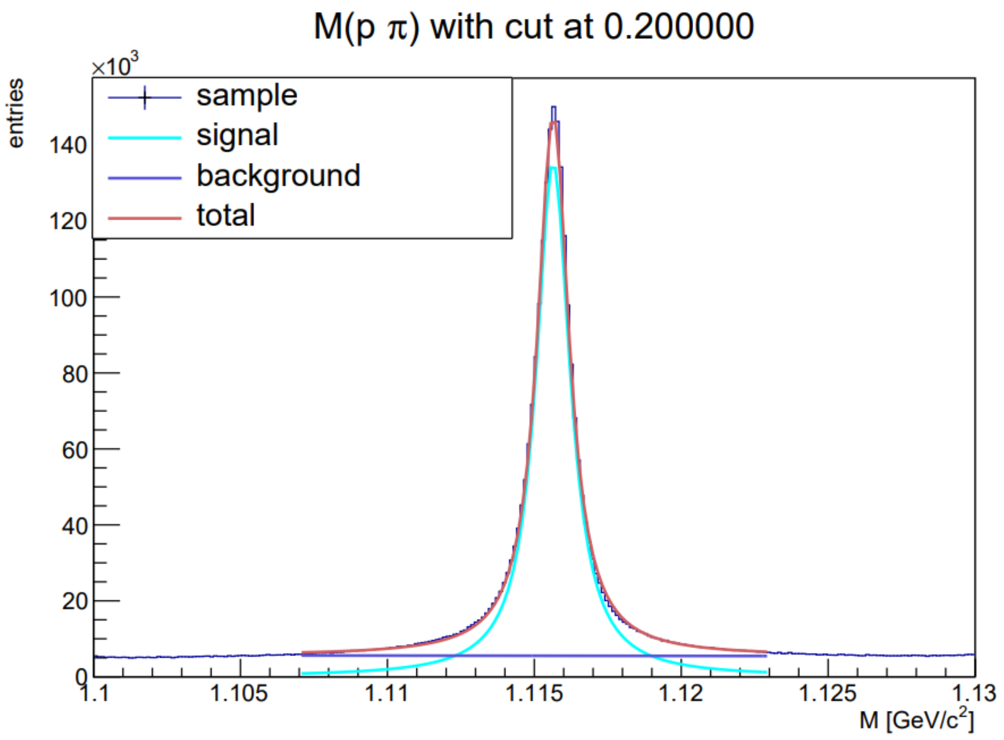


Cut value choice

Voigt function defined as a convolution of two different functions:

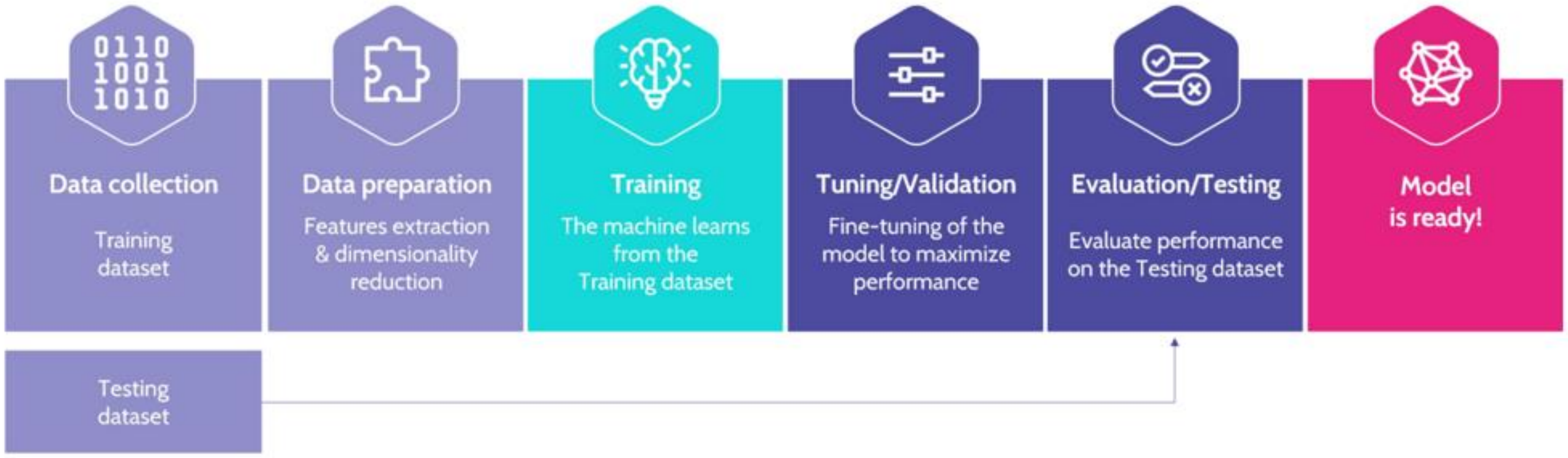
$$gauss(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2}{2\sigma^2}}$$

$$lorentz(x) = \frac{1}{\pi} \frac{\frac{\Gamma}{2}}{x^2 + \frac{\Gamma^2}{4}}$$



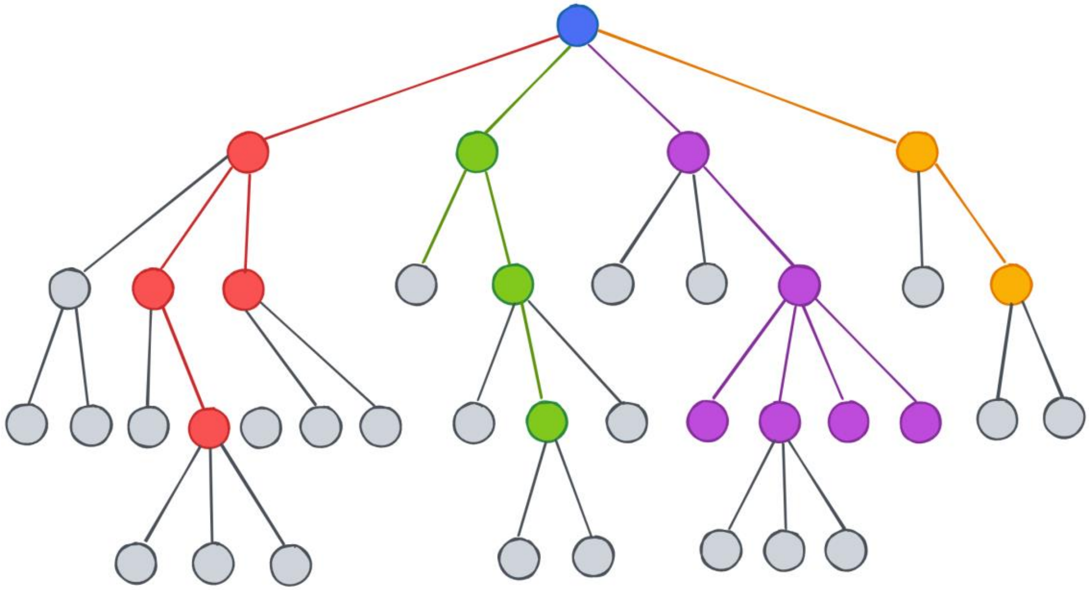
Machine learning algorithm in Λ selection

- The Dataset is a collection of datapoints
- The datapoint is a multidimentional array composed by kinematic variables related to the Λ hyperon and its daughters
- The label of each datapoint is the MC truth: signal (Λ) or background (no Λ).



Decision tree (DT)

- A DT reaches its decision by performing a sequence of test
- Each internal node in the tree corresponds to a test of one of the input features.
- A datapoint is classified by starting by the root node of the tree, testing the feature specified by this node, and then moving down the tree for other different test corresponding into other nodes.
- Every node corresponds a selection criterion, splitting the features in two or more other sub-node
- Each node represents a fraction of signal, the nodes with a higher fraction of signal are called leaf and no other splits are applied



Boosted Decision trees (BDT)

- A single decision tree has low performance
- The performance can be improved by combining a set of decision tree forming a *decision forest*
- This procedure is called *boosting.*
- Each individual tree is built sequentially iterating over the previous one
- Each tree in the sequence is fitted giving more importance to observations in the dataset that were poorly handled by the previous models in the sequence
- Each new model focuses its effords on the most difficult observations obtained in the last iteration