Machine learning for D^0 reconstruction in ep collisions

Shyam Kumar*, Annalisa Mastroserio, Domenico Elia
INFN Bari, Italy

D0 Reconstruction

Rongrong Ma

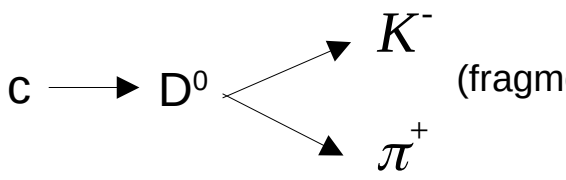
D^0 meson

Heavy quarks (charm and beauty) are produced through hard parton scatterings in the initial stage of the collisions

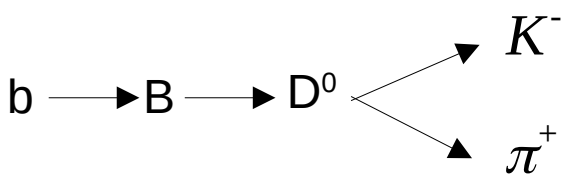
m_{D0} = 1.86483 GeV/c^2

m_{charm} ~ 1.275 GeV/c^2

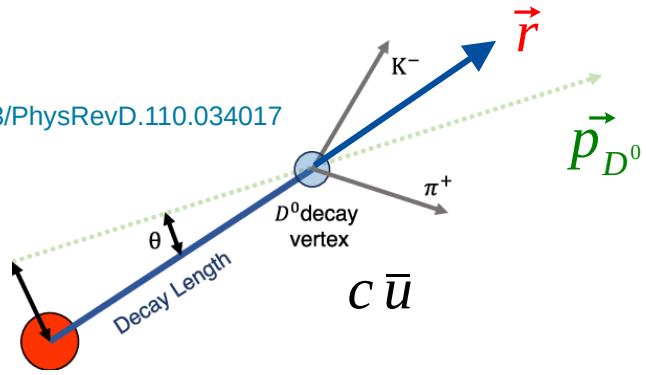
m_{beauty} ~ 4.18 GeV/c^2



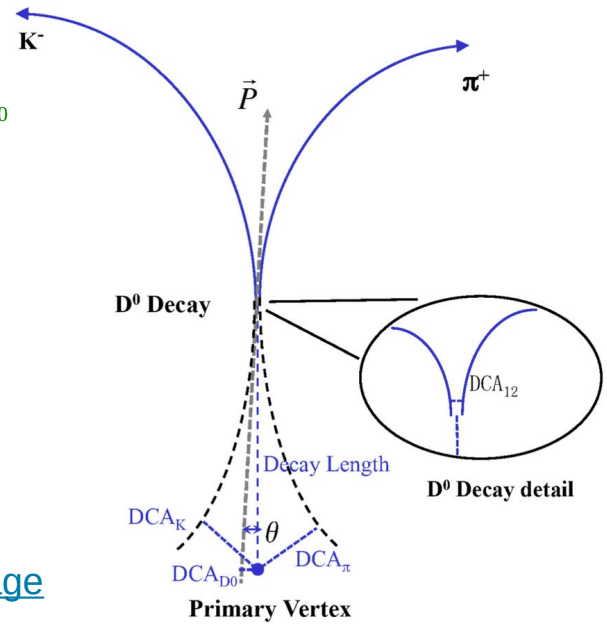
Prompt D-meson
(fragmentation of charm quark)



Non-prompt D-meson
(fragmentation of beauty quark)



Two prongs decay



Signal means simply D0 meson (prompt or non-prompt)

Reconstruction of D^0 meson using combinations of pion and kaon:

- Pion and kaon from true D0 meson (signal)
- Pion and kaon from others (combinatorial background)

hipe4ml package

Thanks Rongrong

Binary classifier: Machine learning model to separate signal D^0 meson from background

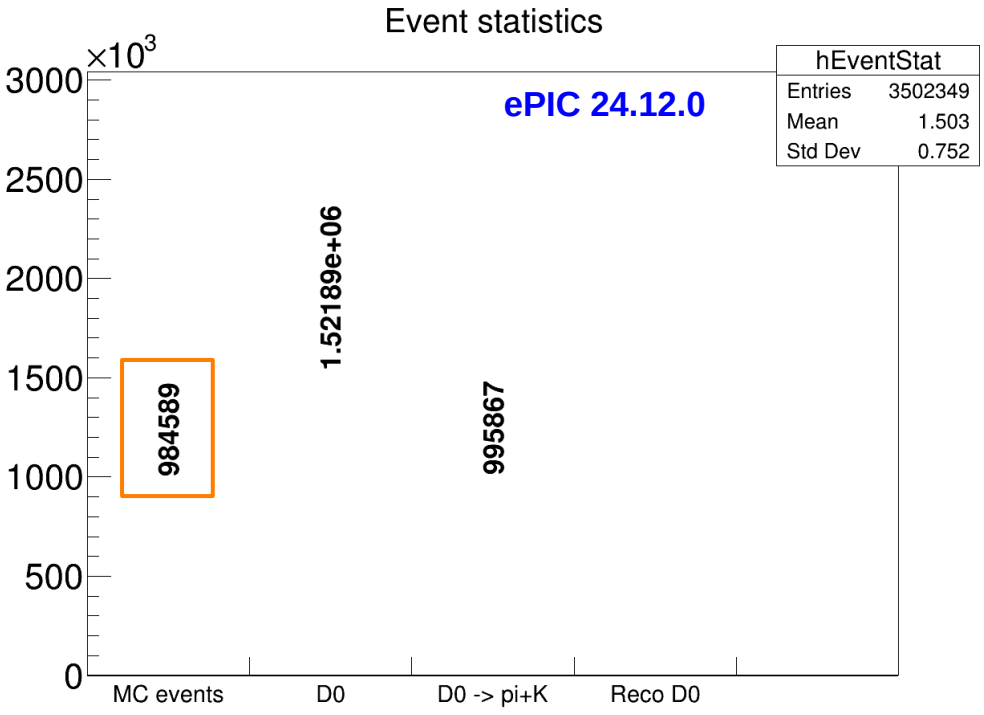
Multi class classifier: Machine learning model to separate prompt, non-prompt D^0 meson, and background

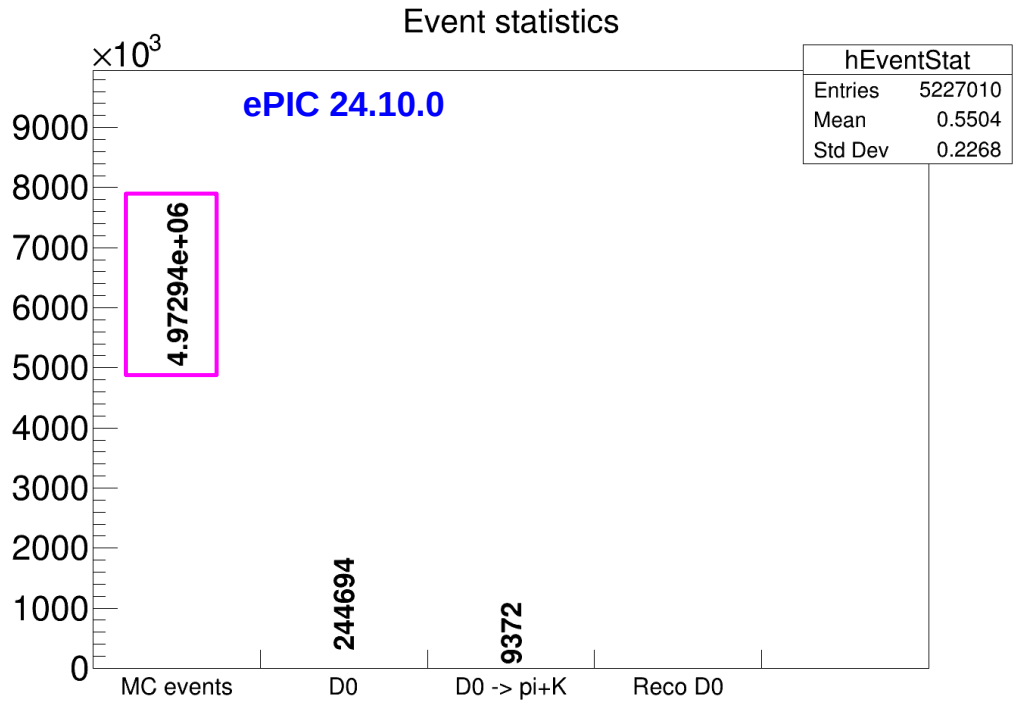
ML Model and Data Sample

- BDT (Boosted Decision Tree) XGBoost Classifier

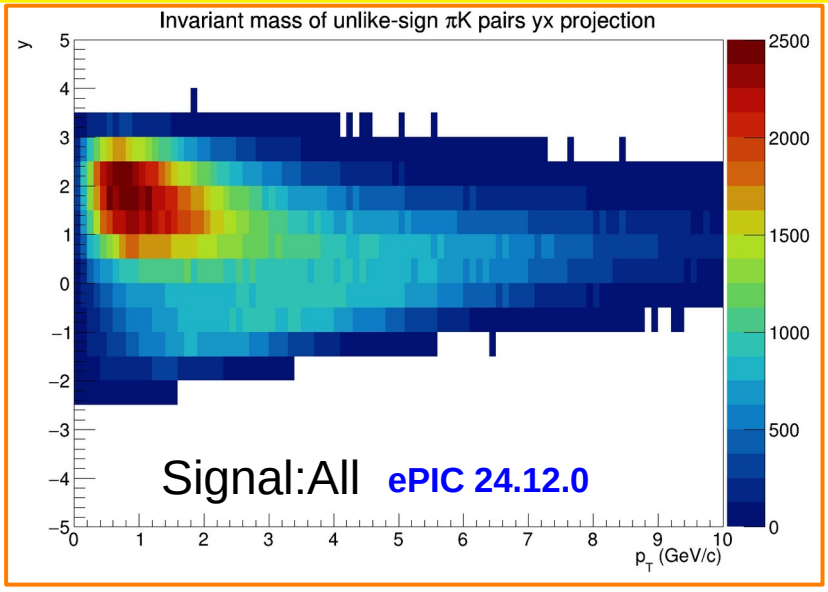
Simulation of D0 and Lc samples

  - Training sample signal and background
  - Signal from the campaign (EPIC/RECO/24.12.0/epic_craterlake/SIDIS/D0_ABCONV/pythia8.306-1.1/10x100/q2_100): **Total files 1869 and Events 984589**
  - Background from EPIC/RECO/24.10.0/epic_craterlake/DIS/NC/10x100/minQ2=100/pythia8NCDIS_10x100_minQ2=100:**Total files 7823 and Events 4.97 M**



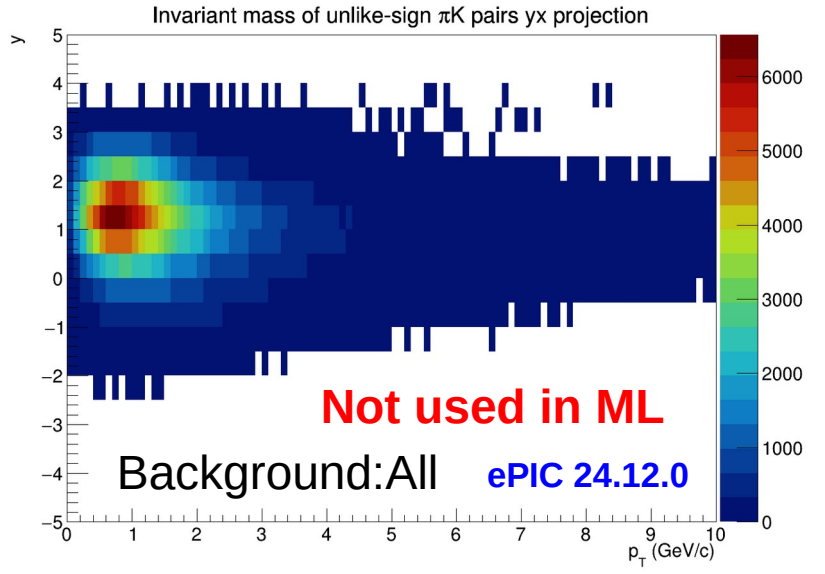


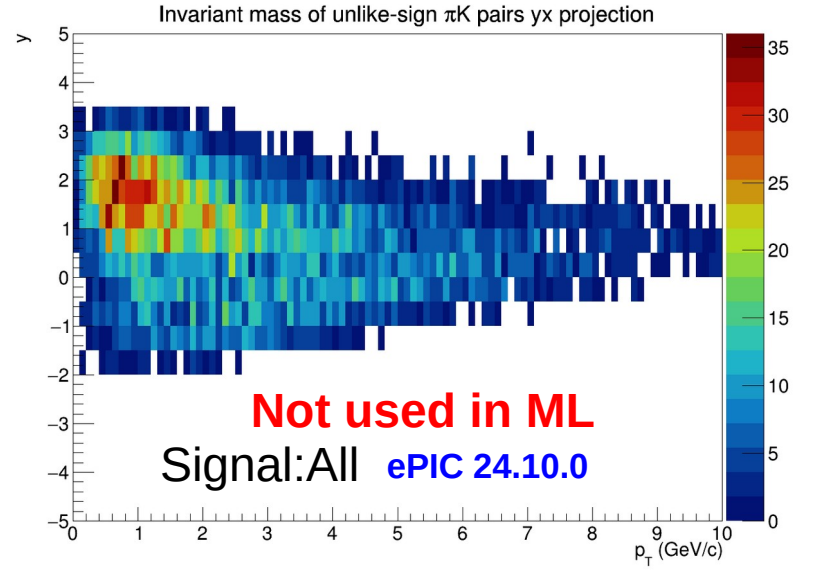
Signal and Background



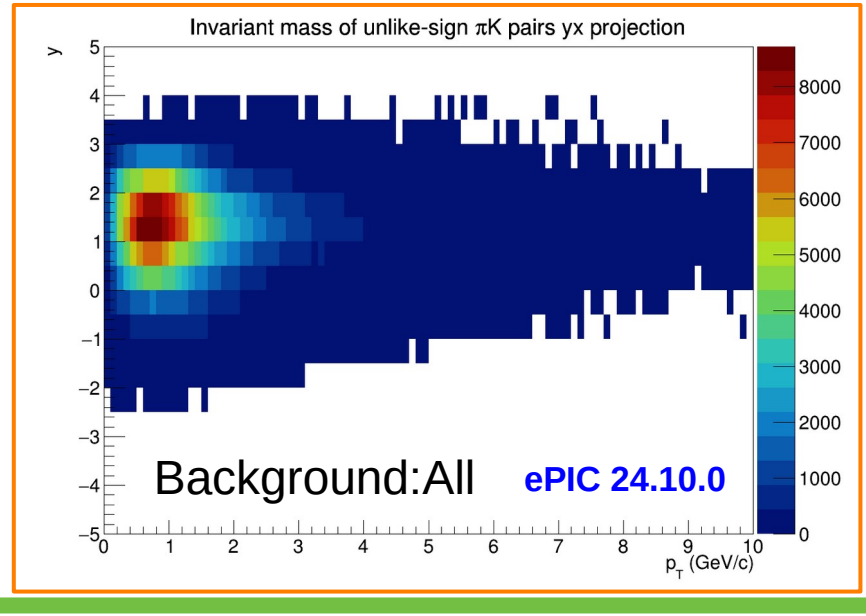
Signal

Features ML:
dca12, dcaD0,
decaylength, costheta





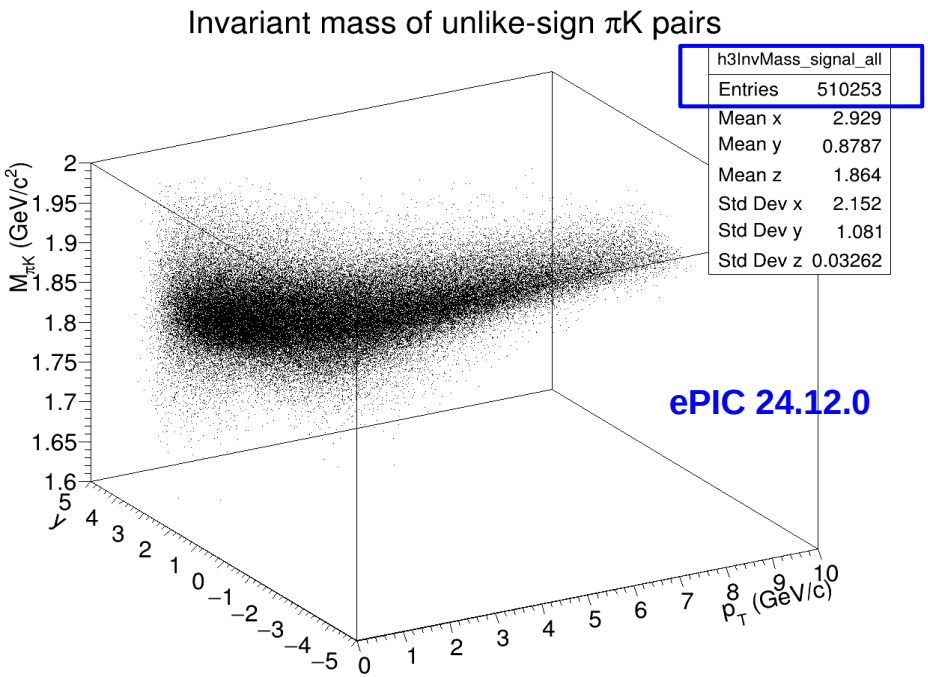
Background

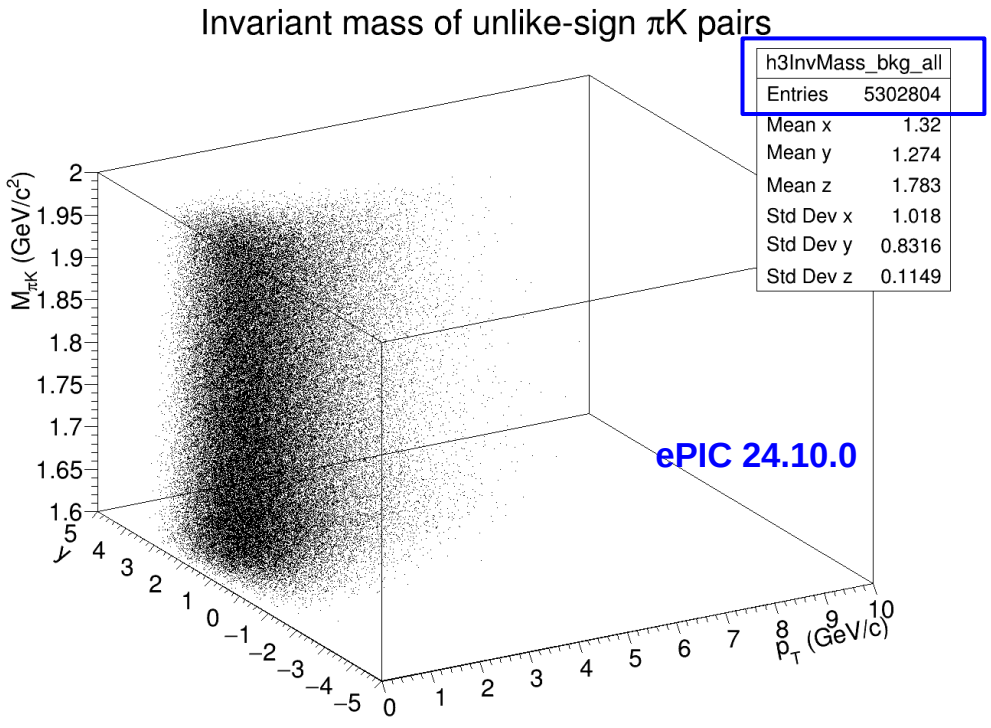


Implementation of ML Model

The model is developed using hipe4ml software

- Started with integrated p_T and η as a first implementation
- Store the features of the true D0 meson (Signal) and background
- **In general we use signal from the MC and background from the sidebands of collected data**
- Split data into training and test: 80% and 20% for testing (important to look if there is over-fitting/under-fitting)





Implementation of ML Model Details

Integrated p_T and η

- Signal (20,000) selected after applying mass cut of 1.7 <m_{D0} < 2.1 GeV/c
- Background candidates (3 times of Signal = 60,000) after applying 1.0 < m_{D0} < 1.70 or 2.1 < m_{D0} < 2.8 GeV/c
- Removed variables (p_T, η, and m_{D0})
- Total data (Signal+Background) = 80,000

Training 80% = 64,000

```
Signal candidates: 510253        Background candidates: 5302804
Signal candidates for ML: 20000          Background candidates for ML: 60000
Training Data
: (         costheta    dca_12    dca_D0  decay_length    eta_D0   mass_D0     pt_D0
410450    0.995605  0.097374  0.090773      0.969308  2.025764  1.837988  1.600895
2979584 -0.900979  0.058452  0.209460      0.482781  2.816687  2.165168  1.579016
204688    0.998793  0.059634  0.090288      1.838040  2.233483  1.839649  1.512166
4435375   0.978284  0.189250  0.190188      0.917589  3.240764  1.512329  0.527592
1507210   0.331427  0.127600  0.111846      0.118546  0.927330  1.093507  2.539548
...            ...       ...       ...           ...       ...       ...       ...
110490    0.788185  0.053176  0.191500      0.311159  2.311648  1.848967  0.665862
2810700   0.369116  0.024220  0.131331      0.141309  0.368601  1.169959  0.843409
4033685 -0.720506  0.261037  0.196921      0.283973  2.850389  1.401140  1.541461
418881    0.998541  0.066653  0.020650      0.382414  0.932611  1.860036  2.738884
412263    0.894591  0.210925  0.164350      0.367767  1.231883  1.889590  1.294997

[64000 rows x 7 columns],        0
0       1
1       0
2       1
3       0
4       0
...    ..
63995   1
63996   0
63997   0
63998   1
63999   1
```

Signal

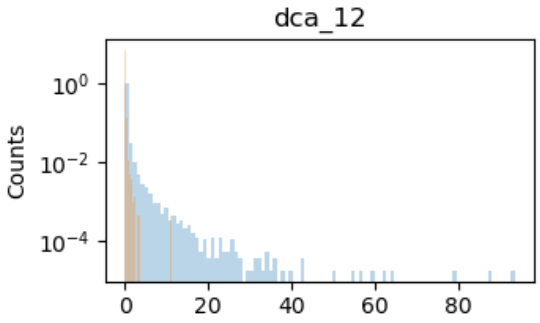
Background

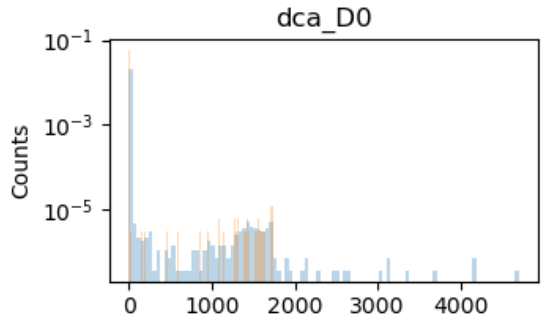
Removed variables (p_T, η, and m_{D0})

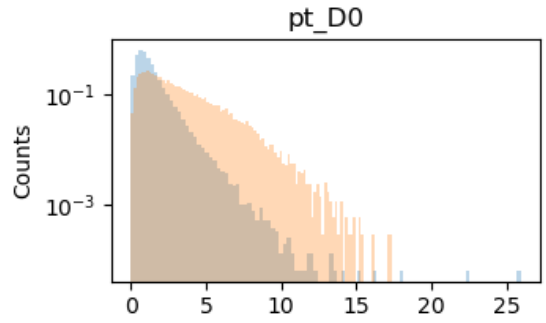
Testing 20%= 16,000

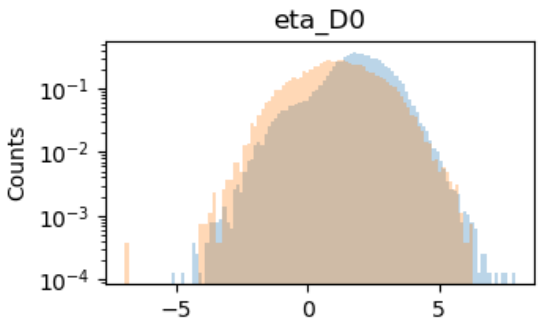
```
 Test Data
: (         costheta    dca_12    dca_D0  decay_length    eta_D0   mass_D0     pt_D0
3351989   0.991298  0.124306  0.039428      0.299528  0.799423  1.069003  1.974903
1025495 -0.394583  0.037872  0.077913      0.084793  1.295227  1.493813  0.789670
4168518   0.464340  0.520904  0.316954      0.357875  2.126908  1.137351  0.659704
935579    0.996984  0.008807  0.195899      2.524239  1.937303  1.000394  1.885553
2186766   0.007942  0.181243  0.124535      0.124539  2.550948  1.205659  0.293283
...            ...       ...       ...           ...       ...       ...       ...
1102953 -0.879209  0.213283  0.040562      0.085137 -0.422315  1.378439  1.181337
4007541   0.530057  0.030857  0.127466      0.150320  0.772105  1.100929  0.360912
2062653   0.909685  0.035019  0.084668      0.203873  0.250959  1.333800  1.179283
2677121   0.917194  0.001137  0.022295      0.055954  0.149519  2.339304  1.122834
392776    0.993571  0.011772  0.055925      0.493987  1.484205  1.819660  2.440594

[16000 rows x 7 columns],        0
0       0
1       0
2       0
3       0
4       0
...    ..
15995   0
15996   0
15997   0
15998   0
15999   1

[16000 rows x 1 columns])
```

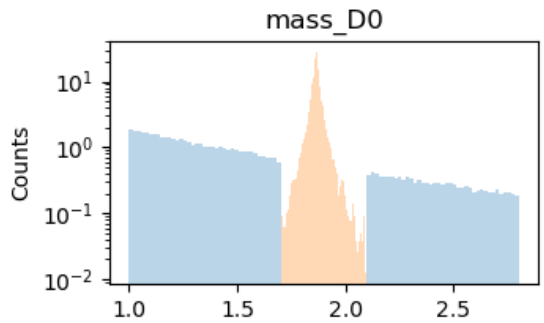
Signal and Background Distributions

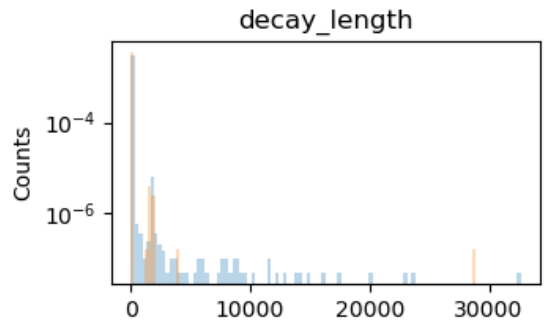


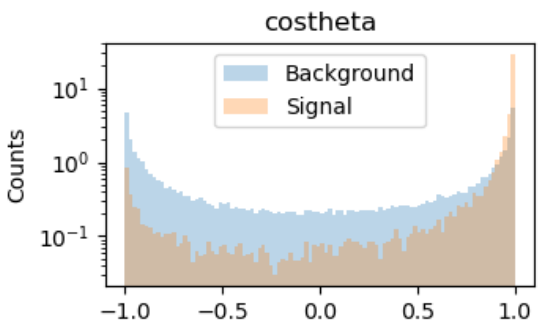








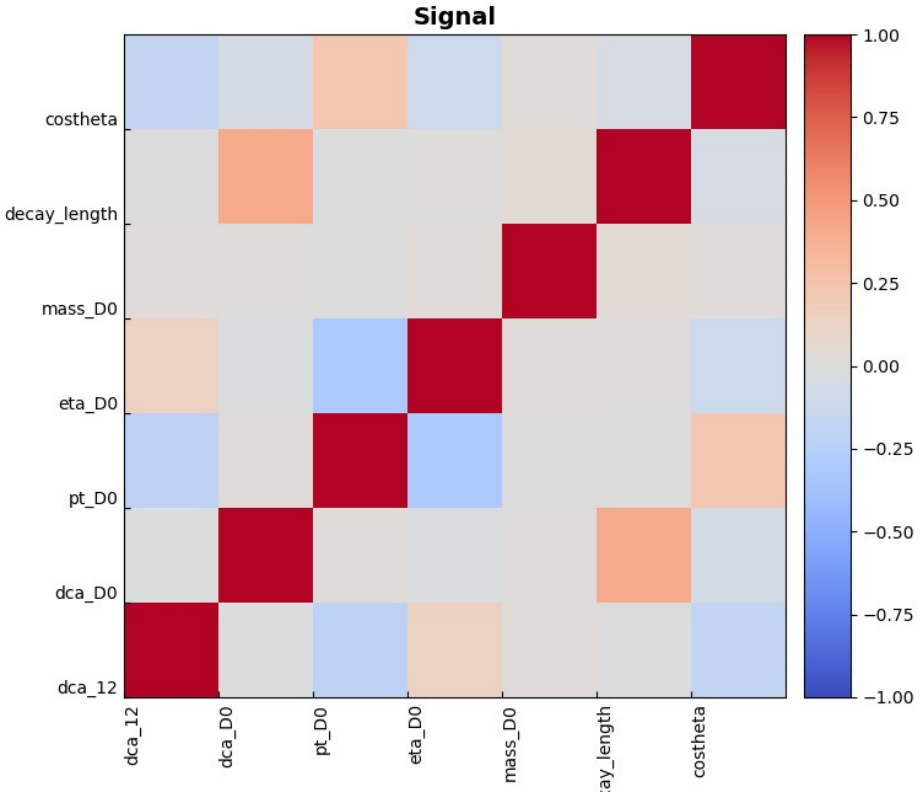


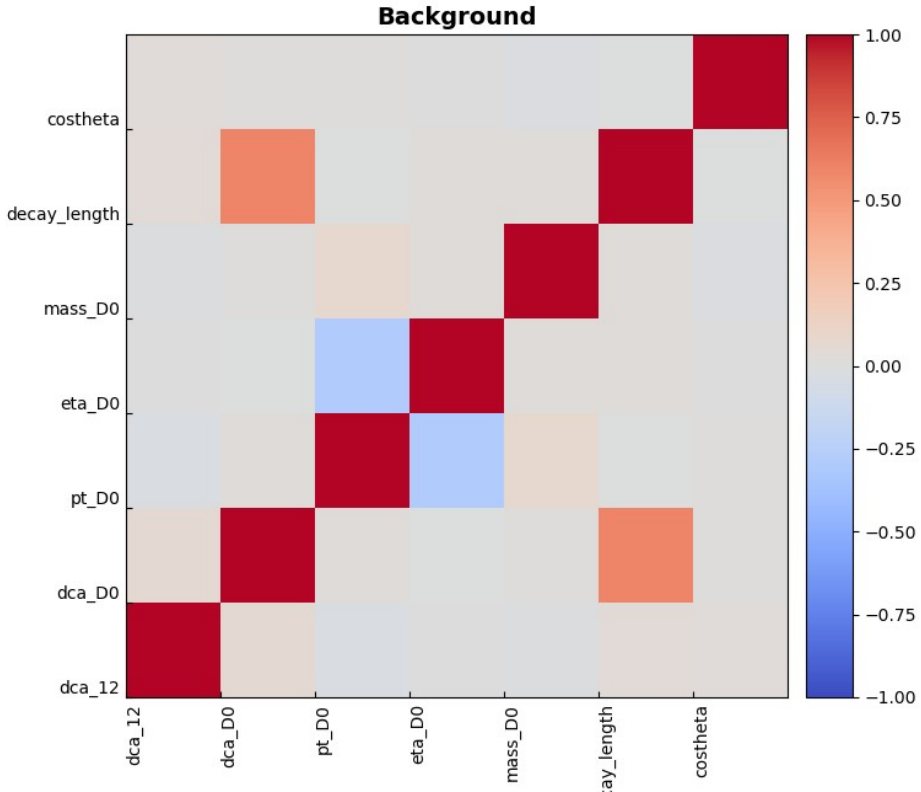


1.7 <m_{D0} < 2.1 GeV/c

1.0 < m_{D0} < 1.70 or 2.1 < m_{D0} < 2.8 GeV/c

Signal and Background Correlations





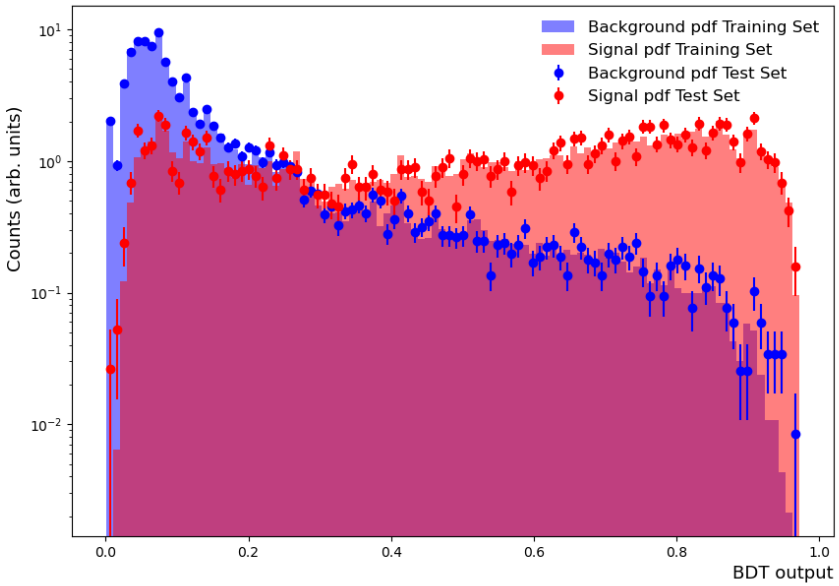
Model Performances

Model can be saved to .onnx format

	P	N
T	TP	TN
F	FP	FN

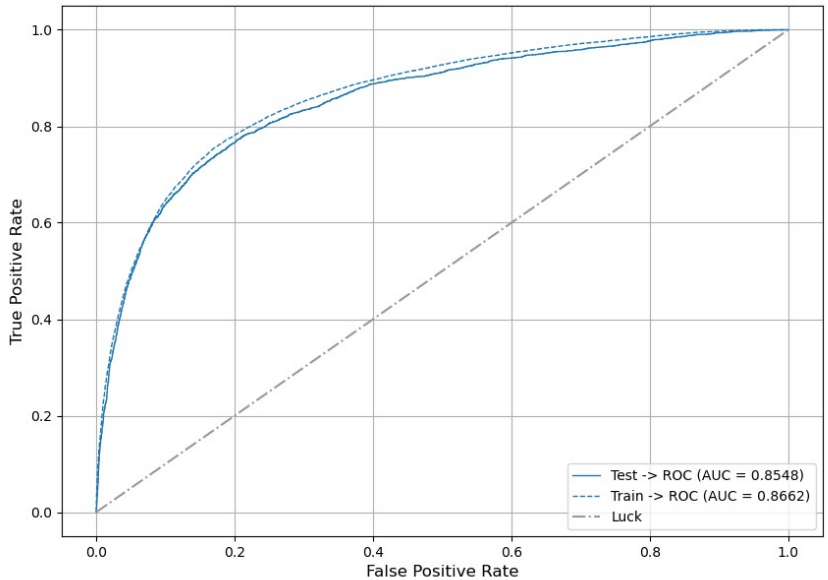
$$TPR = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

$$FPR = \frac{False\ Positive(FP)}{False\ Positive(FP) + True\ Negative(TN)}$$



How Boosted Decision Tree (BDT) classifier separates signal from background

AUC: Area Under Curve



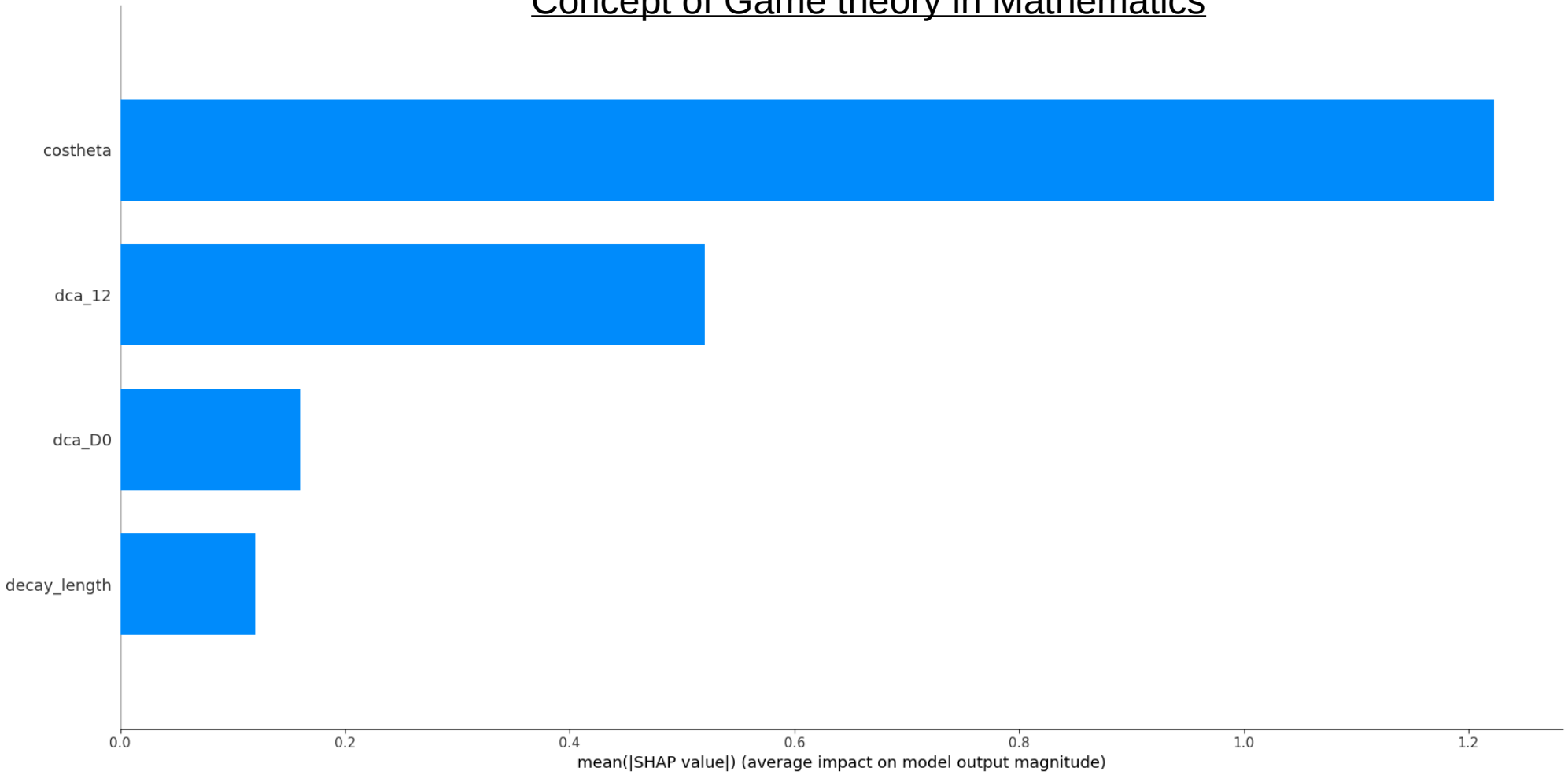
Receiver Operating Characteristic (ROC)

A perfect classifier would have a point at (0, 1), indicating no false positives and all true positives

Features of Importance (Training)

SHAP (SHapley Additive exPlanations)

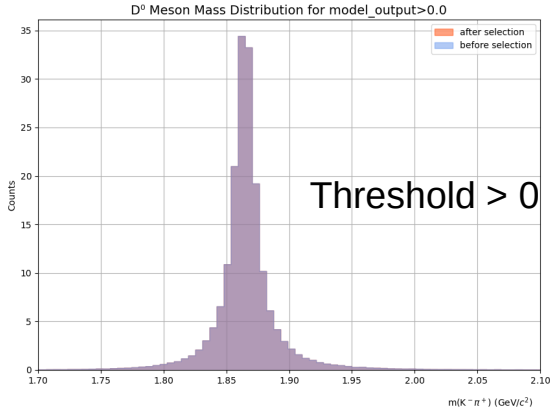
Concept of Game theory in Mathematics

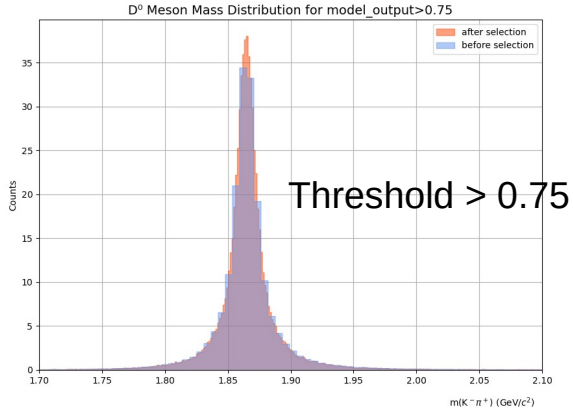


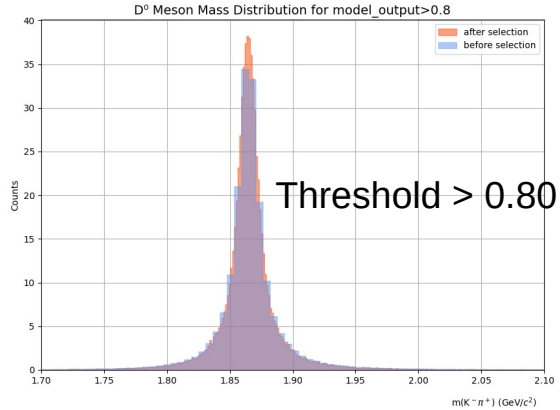
Application of model

I applied it again on signal with topological cut for **ePIC 24.12.0** (Before selection)

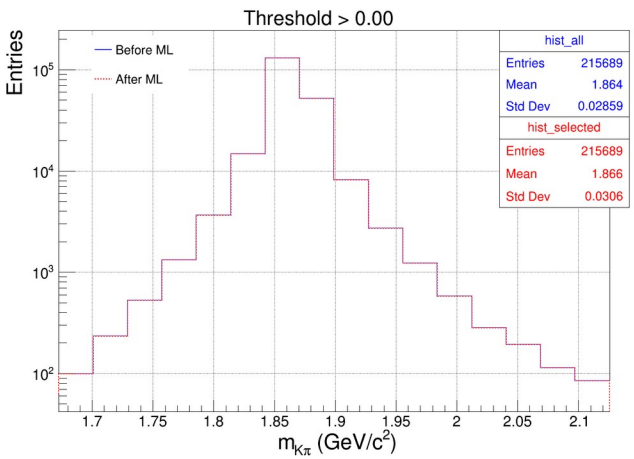
Normalized Plots **After selection:** Applying a BDT threshold cut

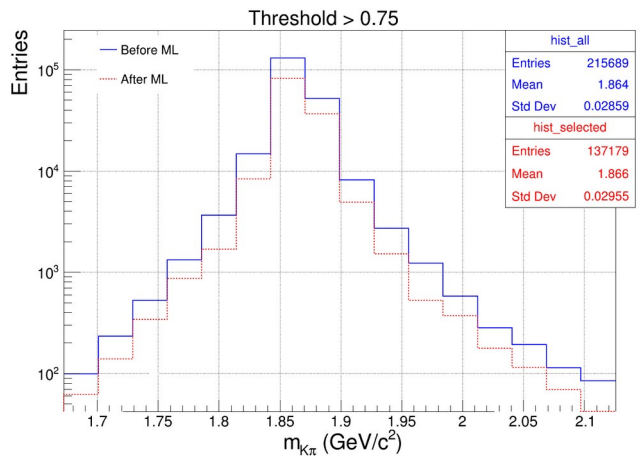


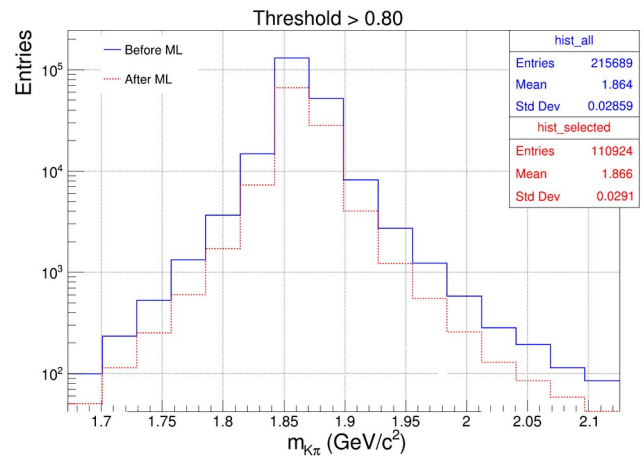




Absolute Entries



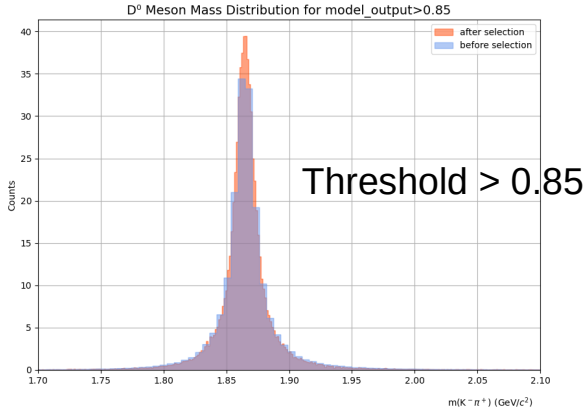




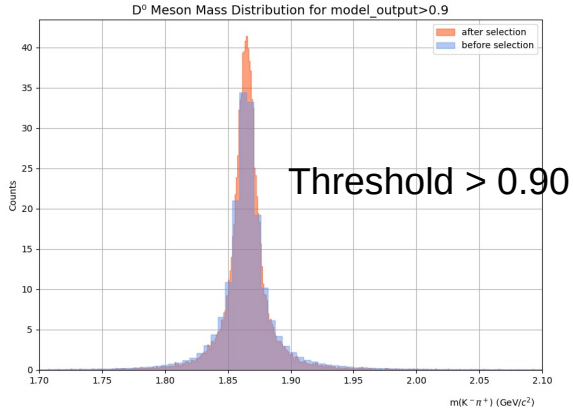
Application of model

I applied it again on signal with topological cut for **ePIC 24.12.0** (Before selection)

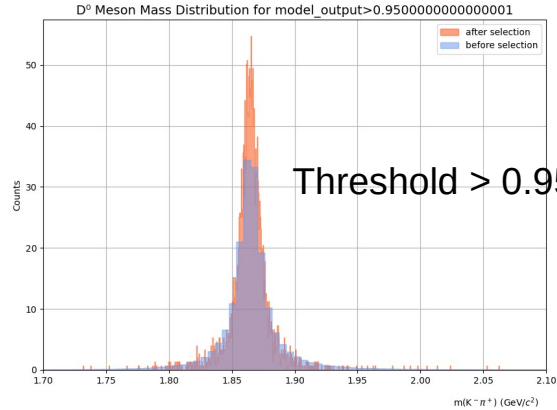
Normalized Plots **After selection:** Applying a BDT threshold cut



Threshold > 0.85

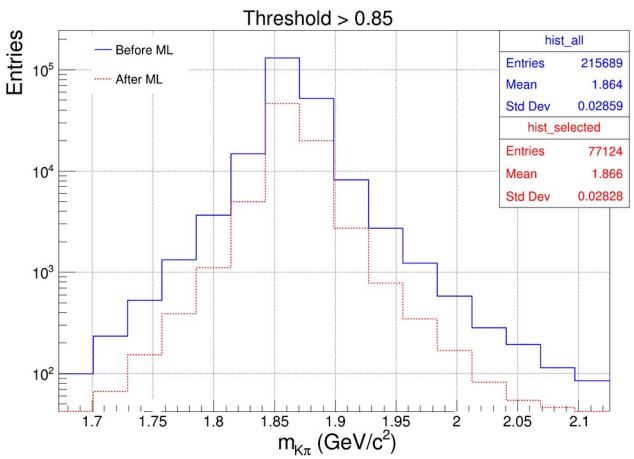


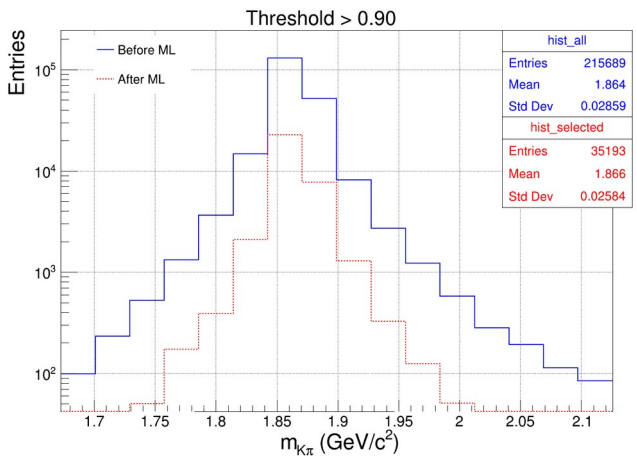
Threshold > 0.90

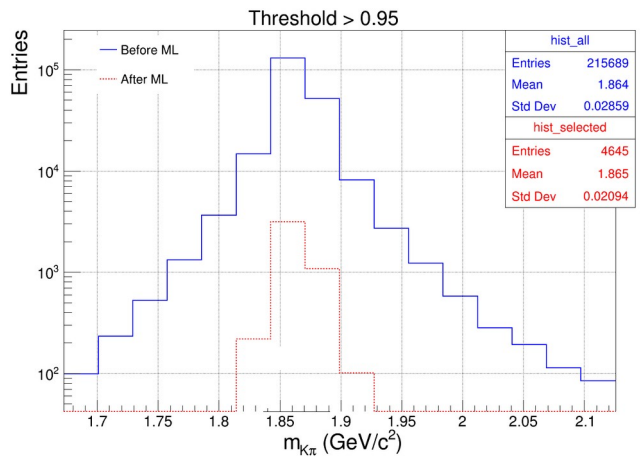


Threshold > 0.95

Absolute Entries







Summary and Future Plan

- First version of Machine learning model implemented for the D^0 reconstruction in ep collisions
- Next Steps:
  - ➔ Add more feature e.g. single track impact parameters, etc.
  - ➔ Further apply it to the collisions with backgrounds
  - ➔ Remove perfect particle identification using only topological variables (data)
  - ➔ Further make it more differential in p_T and η (under testing)
  - ➔ Similar model will implement of Λ_c^+ reconstruction
  - ➔ Once we have prompt, non-prompt tagging then I will implement the multi class classifier

THANK YOU !!!