Machine learning (ML) for D^0 reconstruction in ep collisions

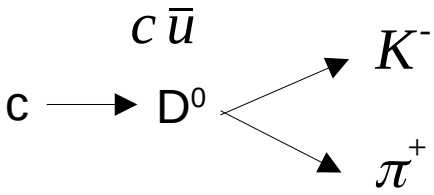
Shyam Kumar*, Annalisa Mastroserio, Domenico Elia

INFN Bari, Italy

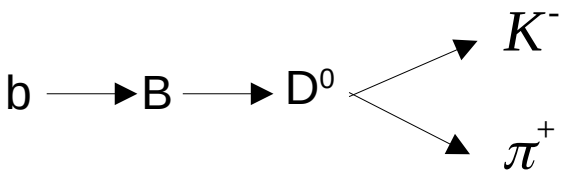
D^0 Decay

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

m_{charm} ~ 1.275 GeV/c^2 **m_{beauty} ~ 4.18 GeV/c^2**



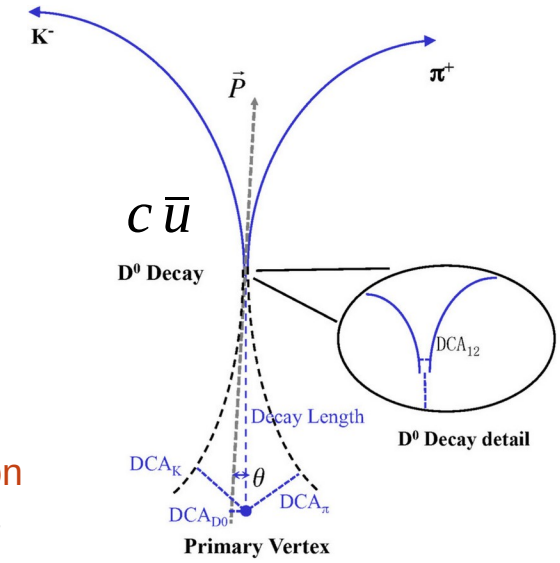
Prompt D^0-meson
(fragmentation of charm quark)



Non-prompt D^0-meson
(fragmentation of beauty quark)

Two prongs decay

m_{D0} = 1.86483 GeV/c^2



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)

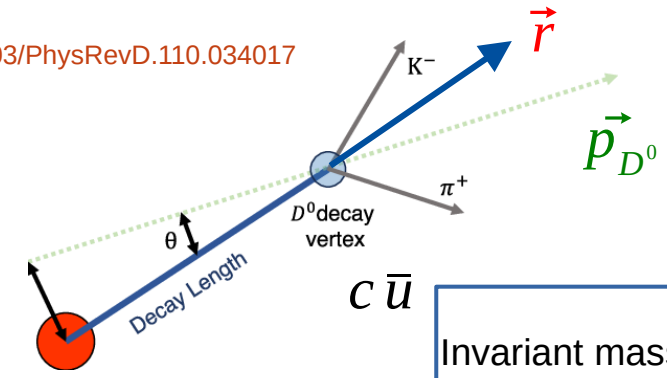
D0 reconstruction
Rongronga Ma

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

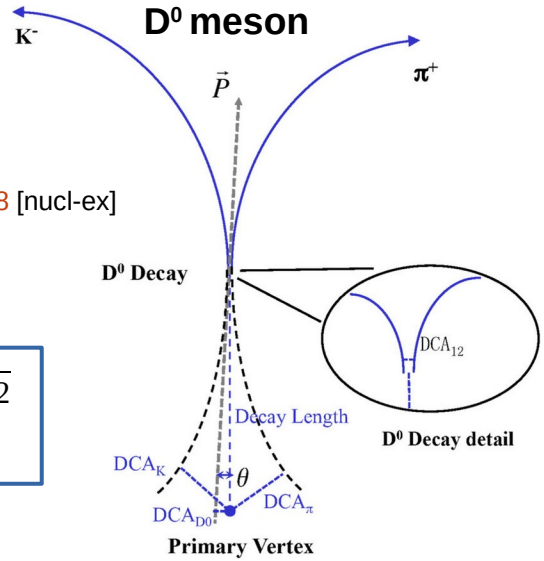
Topological Variables

10.1103/PhysRevD.110.034017



cτ = 123 μm

arXiv:1911.12168 [nucl-ex]

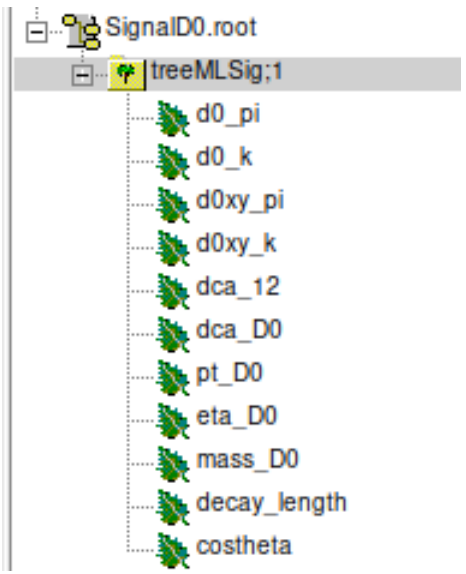


Invariant mass: $m_{D^0} = \sqrt{\left(E_{K^-} + E_{\pi^+}\right)^2 - \left(\vec{p}_{K^-} + \vec{p}_{\pi^+}\right)^2}$

Topological Variables:

- DCA_{k^-} and DCA_{π^+} with respect to the reconstructed primary vertex (d0_k, d0_pi) mm
- Decay length of D^0 meson (decaylength)
- Cosθ (angle between **r** and $\mathbf{p_{D0}}$)
- DCA_{12} distance between the daughter tracks of D^0
- DCA_{D0} impact parameter of reconstructed D^0 meson
- m_{D0} invariant mass of kaon and pion pairs
- pt_D0 reconstructed pt of the D^0 meson
- eta_D0 reconstructed η of the D^0 meson

Toplogical variables using STAR experiment classes

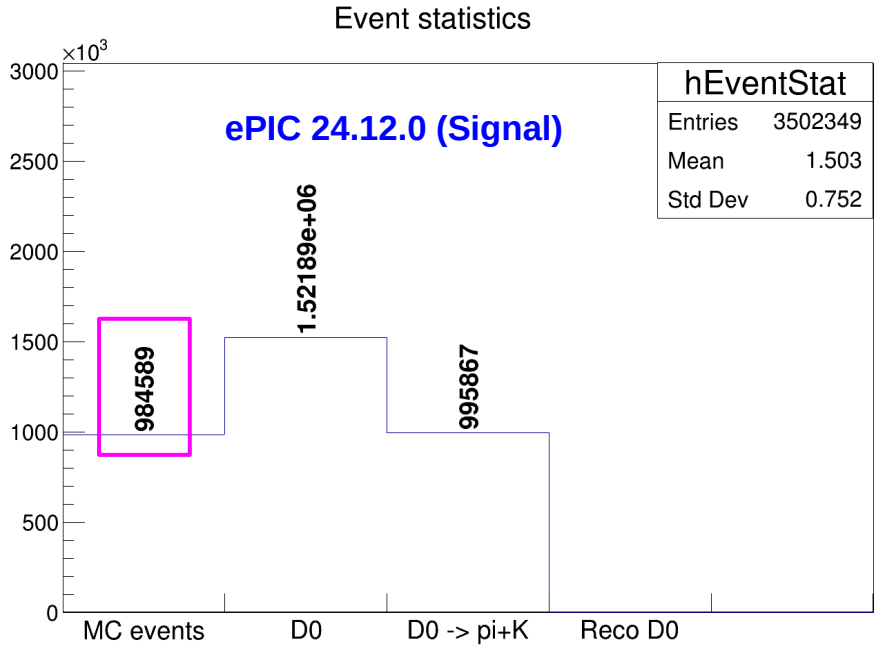


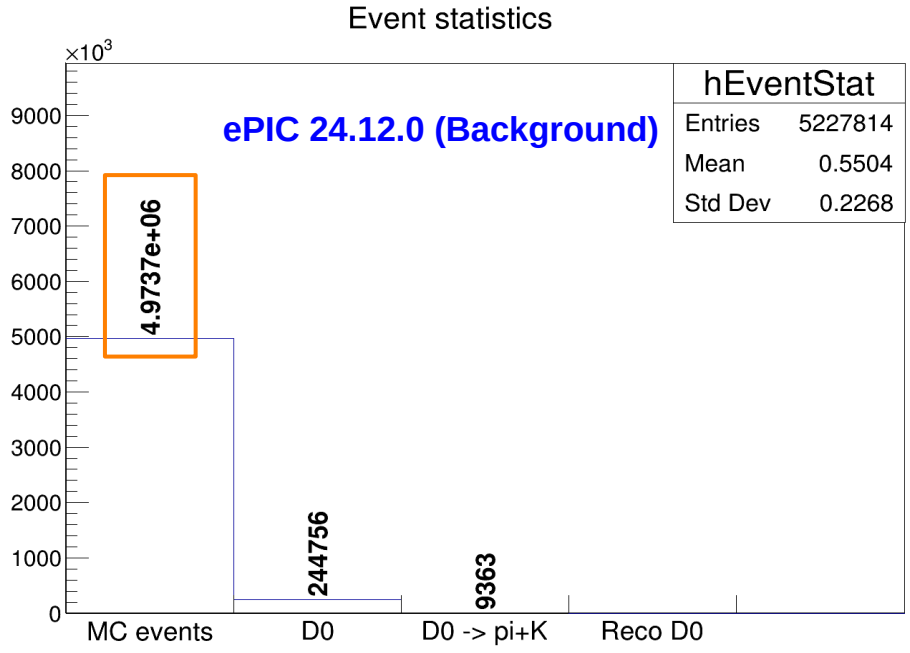
Data Sample for ML (Q^2 = 100)

- **Different algorithms: BDT (Boosted Decision Tree) XGBoost Classifier**, Convolutional Neural Network (CNN), Generative Adversarial Networks (GANs), Auto Encoder (AE)

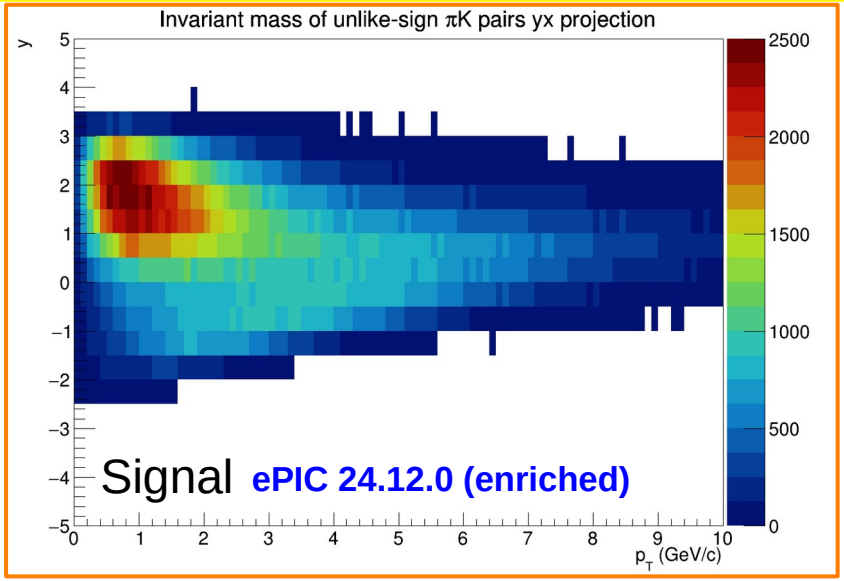
Simulation of D0 and Lc samples

  - BDT requires the features for the signal D^0 meson and background D^0 meson (fake combinations of pion,kaon)
  - D^0 enriched same created filtering **PYTHIA8 ep, NC, 10X100, Q^2 >100 events (~493M)** such that each event consist one $D^0 \rightarrow$ k-π+ known as Signal taken from 24.12.0/epic_craterlake/SIDIS/D0_ABCONV/pythia8.306-1.1/10x100/q2_100): **Total files 1869 and Events = 984589**
  - Background from 24.12.0/epic_craterlake/DIS/NC/10x100/minQ2=100: **Total files 7430 and Events = 4973695**

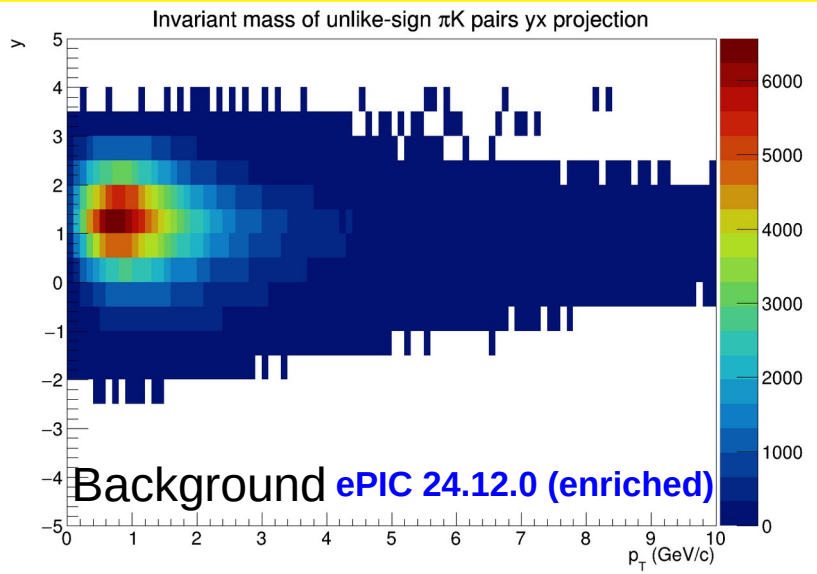


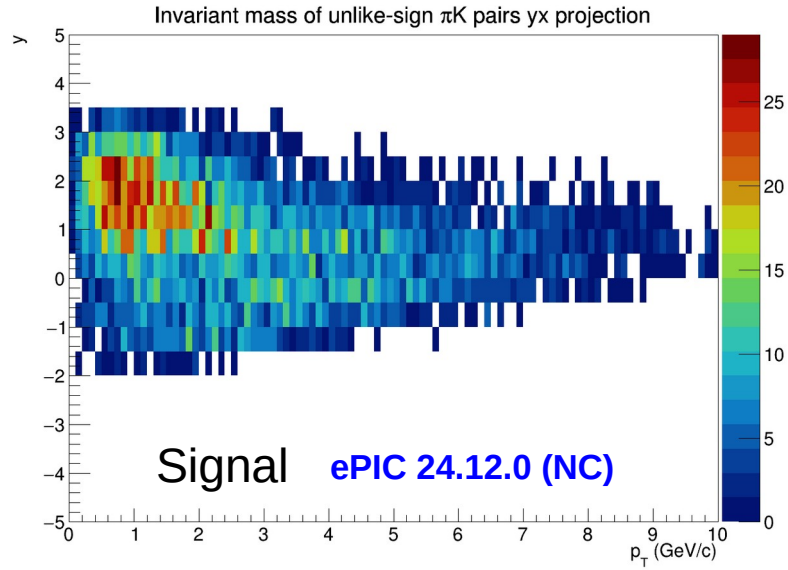


Phase Space of D^0 meson (Q^2=100)

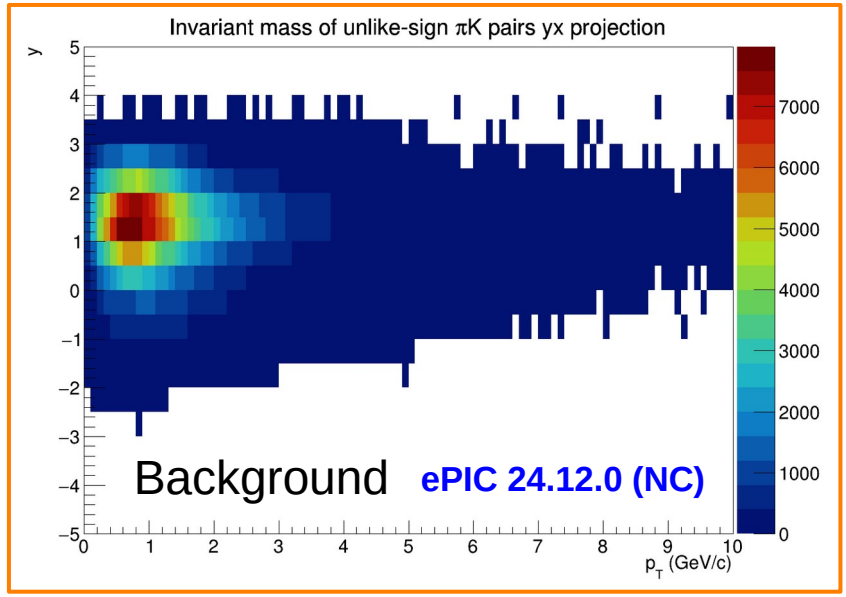


Signal





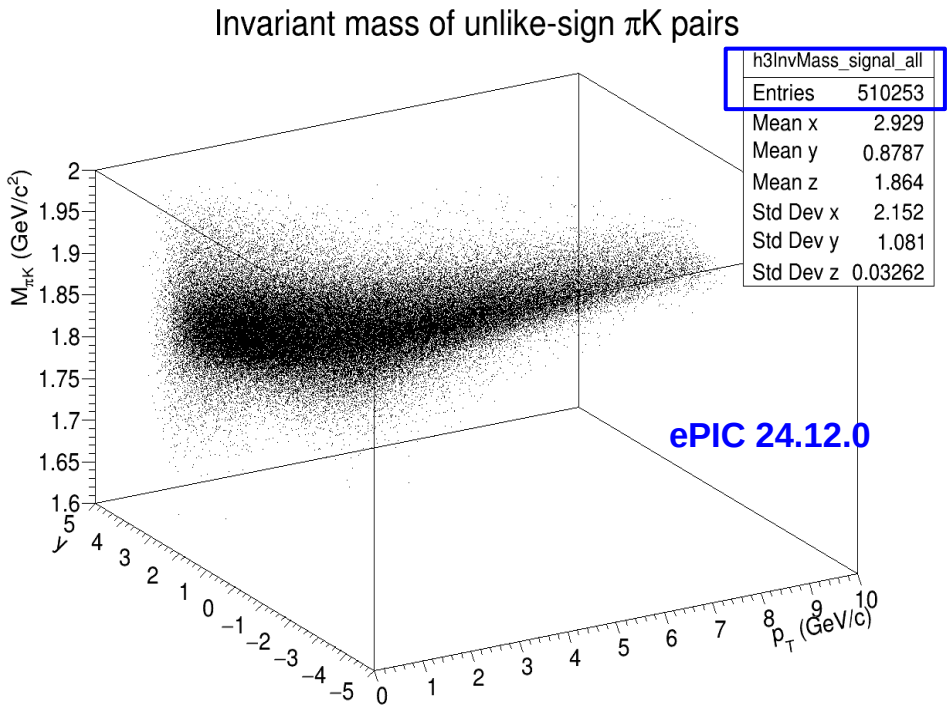
Background

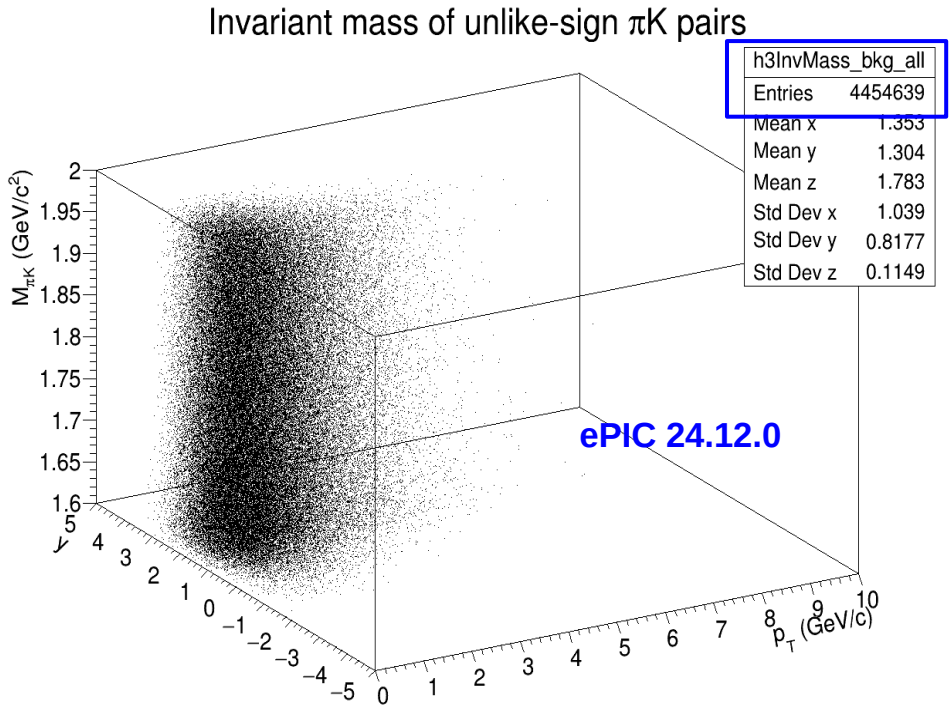


Implementation of ML Model (Q^2 = 100)

The model is developed using hipe4ml https://doi.org/10.5281/zenodo.5070131

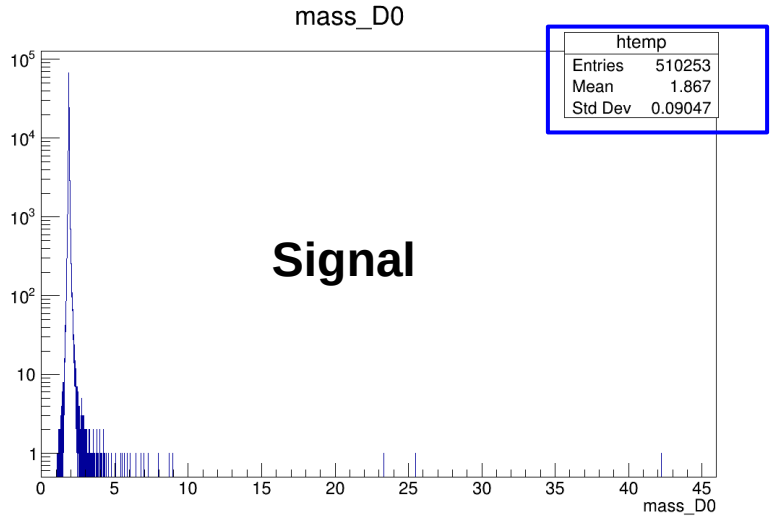
- Started with integrated p_T and η as a first implementation
- **In general we use signal from the MC and background (fake pion/kaon combinations) from the sidebands**
- Split ML data into training and test: 80% and 20% for testing (important to look if there is over-fitting/under-fitting)



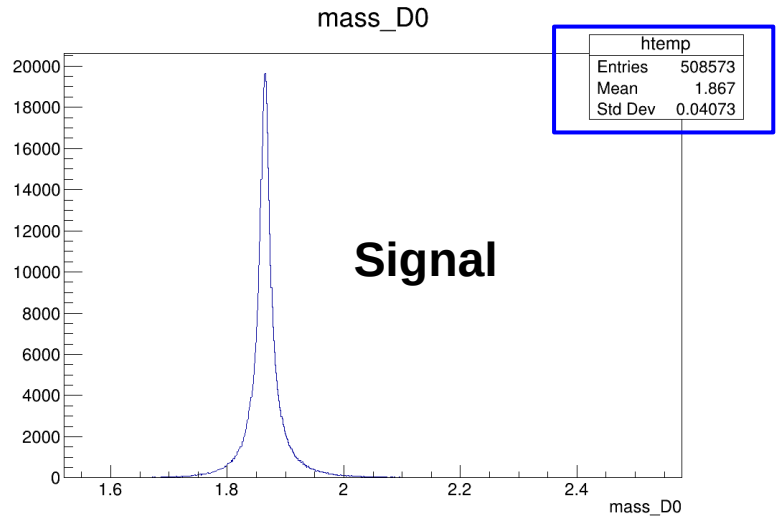


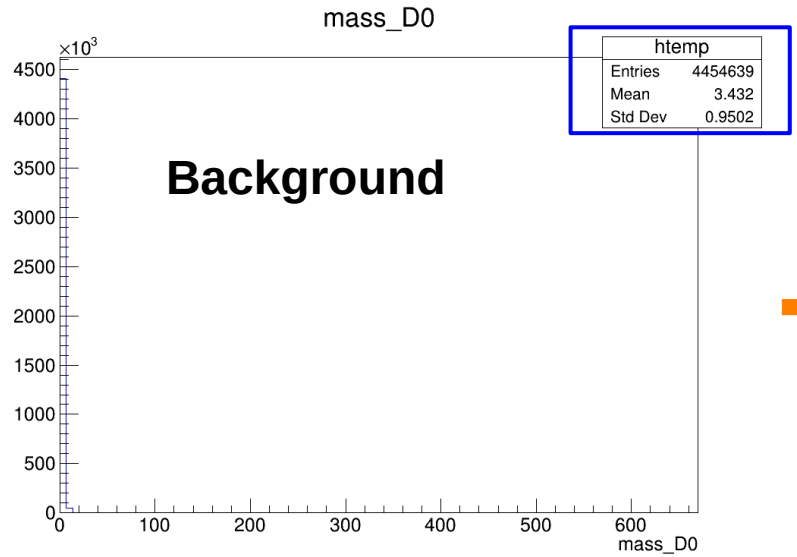
Filtering data

Filtering: (mass_D0 > 1.6 && mass_D0 < 2.5) && (d0xy_pi>0.02 && d0xy_pi<10.) && (d0xy_k>0.02 && d0xy_k<10.) && decay_length <100.

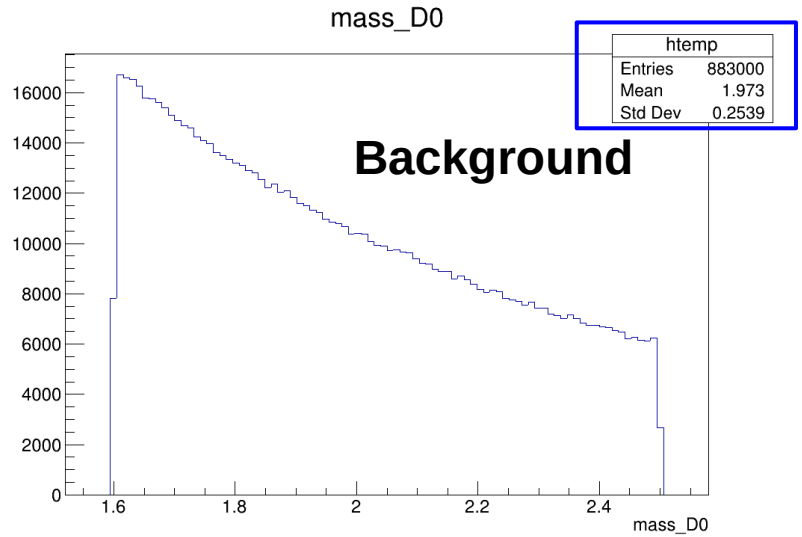


Filtering

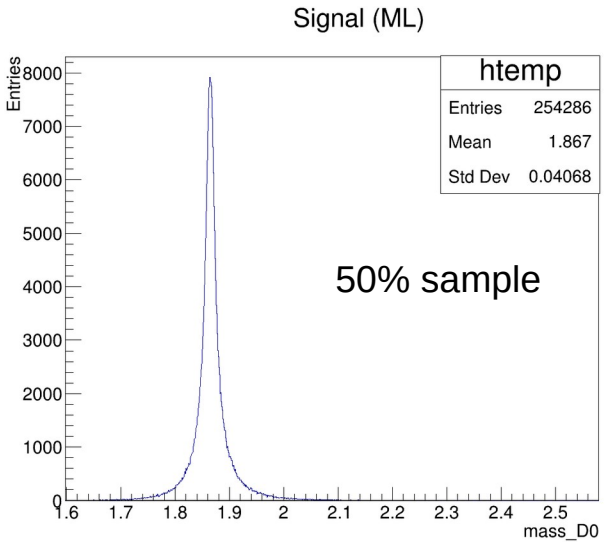


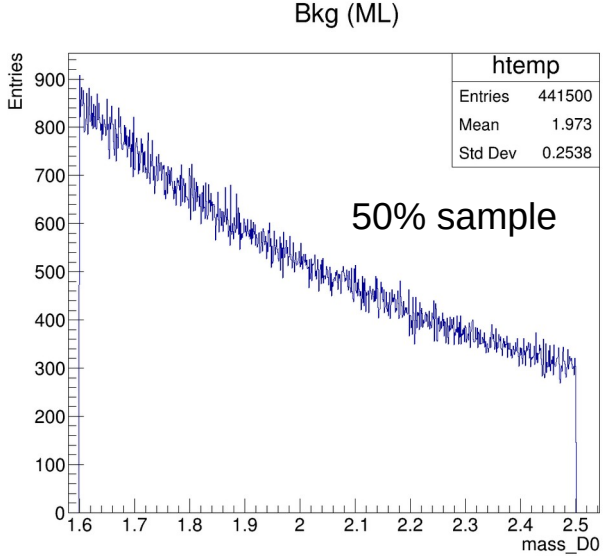


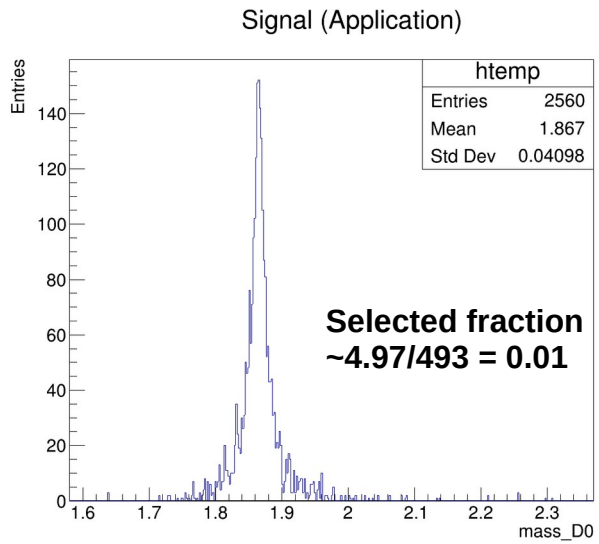
Filtering

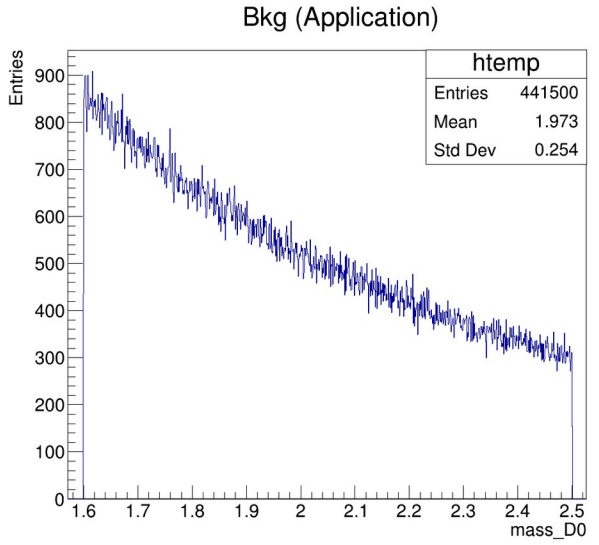


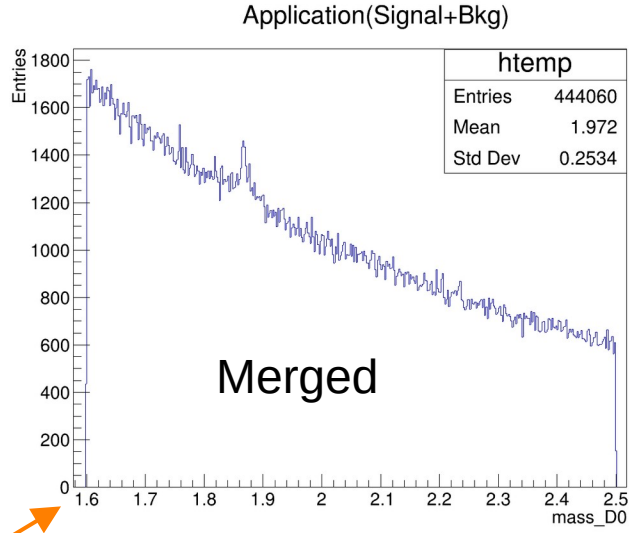
Sample Preparation











- PYTHIA8 ep, NC, 10X100, Q^2 >100: signal events (~493M) while background events (~4.97M)
- Signal and Bkg divided (50% for ML and 50% Application)
- Signal scaled by a factor 0.01 to consider unbalance in entries

Implementation Details

Integrated p_T and η

- Signal (40,000) applying mass cut of $1.8 < m_{D0} < 1.92$ GeV/c
- Background candidates (40,000) $1.6 < m_{D0} < 1.8$ or $1.92 < m_{D0} < 2.1$ GeV/c
- Removed variables (p_T, d0xy_k, d0xy_pi, η_{D0} and m_{D0})
- Total data (Signal+Background) = 80,000

Training 80% = 64,000

```
Signal candidates: 254286        Background candidates: 441500
Signal candidates for ML: 40000          Background candidates for ML: 40000
Training Data
: (        costheta      d0_k     d0_pi  ...    eta_D0   mass_D0      pt_D0
50438    0.997740  0.289914  0.131012  ...  2.048417  1.892901   2.313375
224740   0.994266  0.176400  0.131961  ...  1.474017  1.842975   1.081754
238245   0.999038  0.229977  0.359893  ...  0.476778  1.859466   5.441979
137027  -0.484314  0.043081  0.137375  ...  1.784426  1.941882   0.882545
119738   0.992224  0.595780  0.365383  ...  1.070940  1.882983  12.098245
...           ...       ...       ...  ...       ...       ...        ...
108030   0.257544  0.032144  0.056500  ...  2.198905  1.849791   0.709628
172710  -0.892203  0.178346  0.137853  ...  2.936060  1.772410   0.344631
205012  -0.827983  0.830113  0.090533  ...  2.776265  1.715847   1.052274
83094    0.787943  0.715882  0.404875  ...  0.843373  1.862297   1.591674
126186   0.999933  0.331573  0.256885  ...  0.353859  1.864653   4.925037

[64000 rows x 11 columns],        0
0       1
1       1
2       1
3       0
4       1
...    ..
63995   1
63996   0
63997   0
63998   1
63999   1

[64000 rows x 1 columns]),
```

Signal

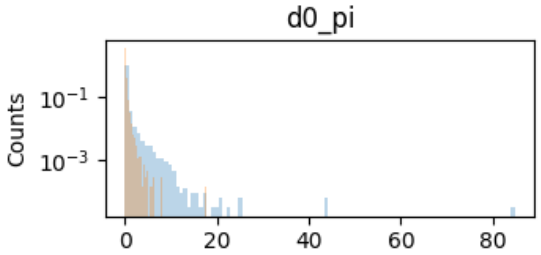
Background

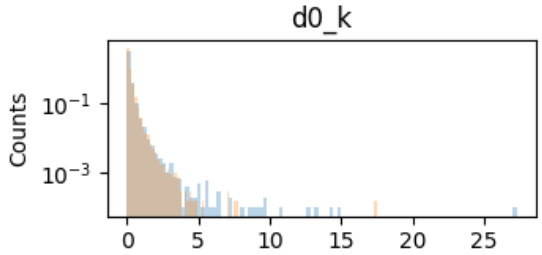
Removed variables (p_T and m_{D0})

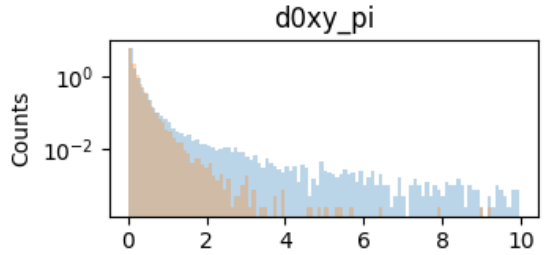
Testing 20%= 16,000

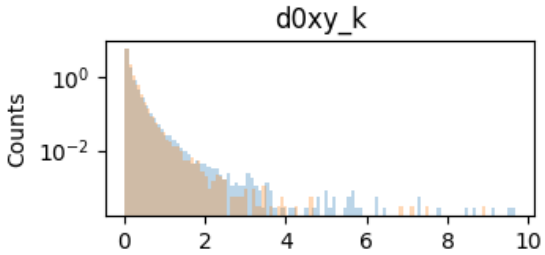
```
 Test Data
: (        costheta      d0_k     d0_pi  ...    eta_D0   mass_D0      pt_D0
82956    0.616404  0.149910  0.062005  ...  1.787789  1.925152  1.631700
136665   0.864028  0.182478  0.034499  ...  3.183362  1.966506  0.557345
313259  -0.719817  0.050165  0.537944  ...  1.498422  1.717138  1.173662
18644    0.824208  0.039054  0.117089  ... -1.289520  1.708649  3.577919
364248  -0.271628  0.037806  0.647156  ...  1.947055  1.932307  1.403166
...           ...       ...       ...  ...       ...       ...       ...
307189   0.025919  0.089753  0.036269  ... -0.432812  1.734530  1.019298
329546  -0.927165  0.138770  0.039194  ...  1.750898  1.989538  2.637261
151725  -0.869652  0.033717  0.146310  ...  1.287825  1.702242  1.485120
266828   0.914427  0.041193  0.439854  ...  1.981523  1.943916  1.401100
24231   -0.979701  0.031079  0.068232  ...  0.025866  1.872412  2.554892

[16000 rows x 11 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])
```

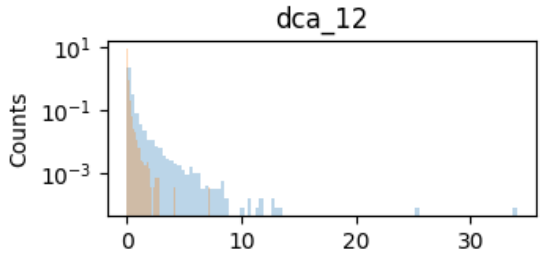
Signal and Background Distributions

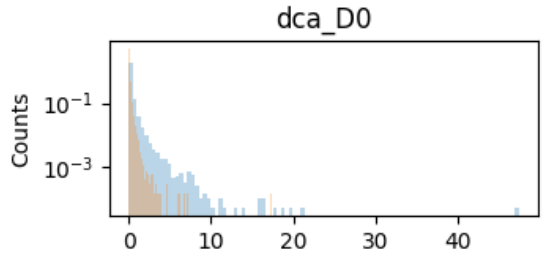


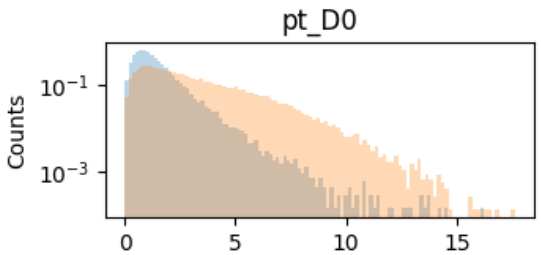


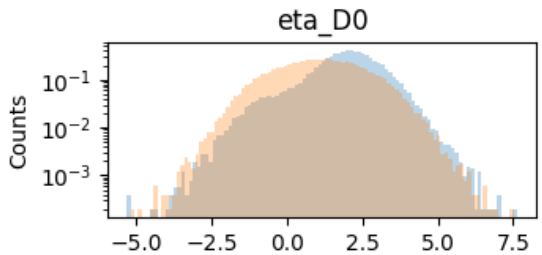


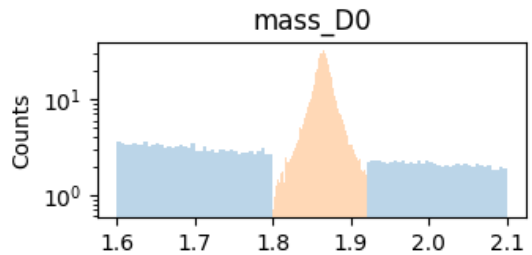


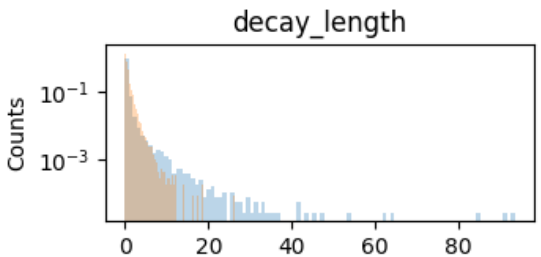


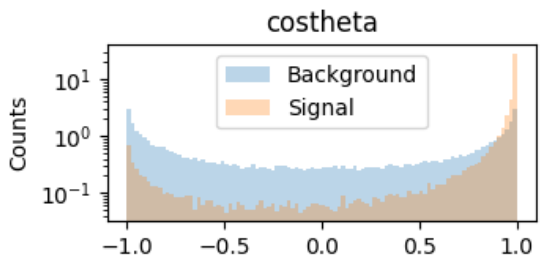










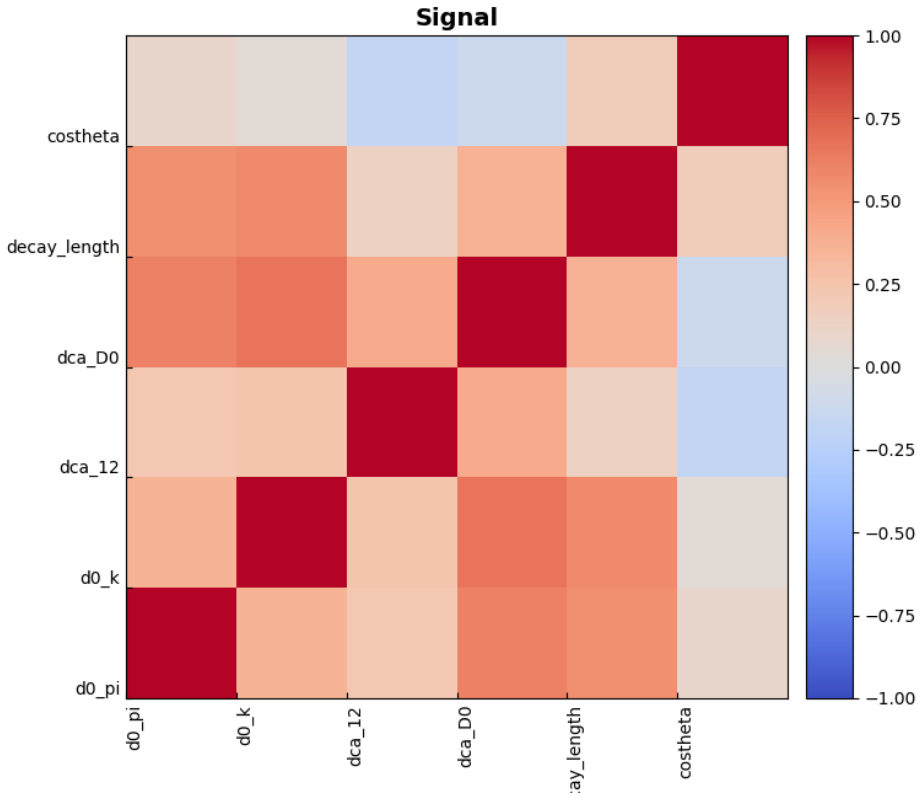


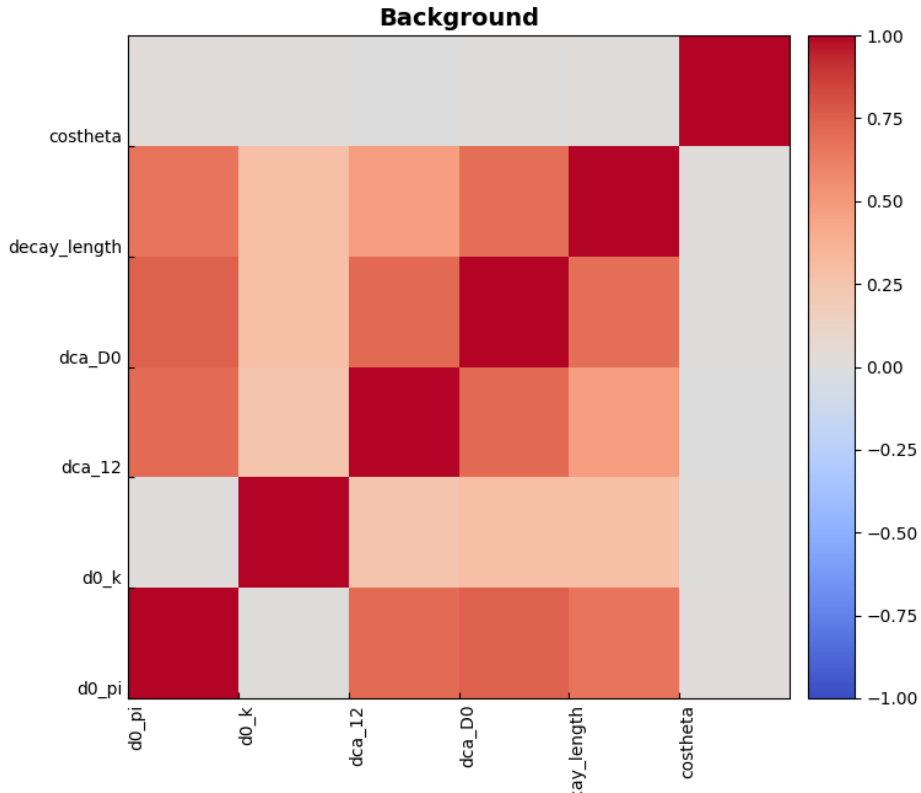
1.8 <m_{D0} < 1.92 GeV/c

1.6 < m_{D0} < 1.8 or 1.92 < m_{D0} < 2.1 GeV/c

Feature Correlations

Highly correlated variable carries similar information, one of them can be removed while training the model





Model Performances

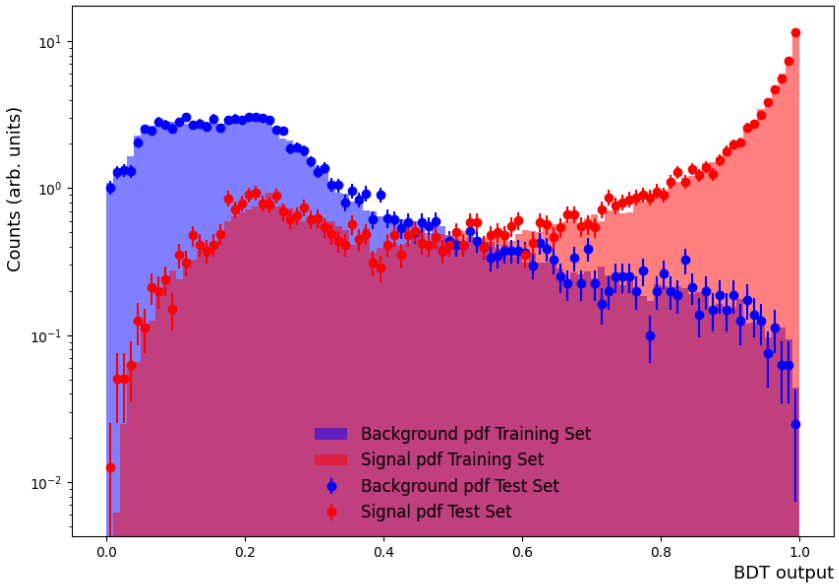
Model can be saved to .onnx format

	P	N
T	TP	TN
F	FP	FN

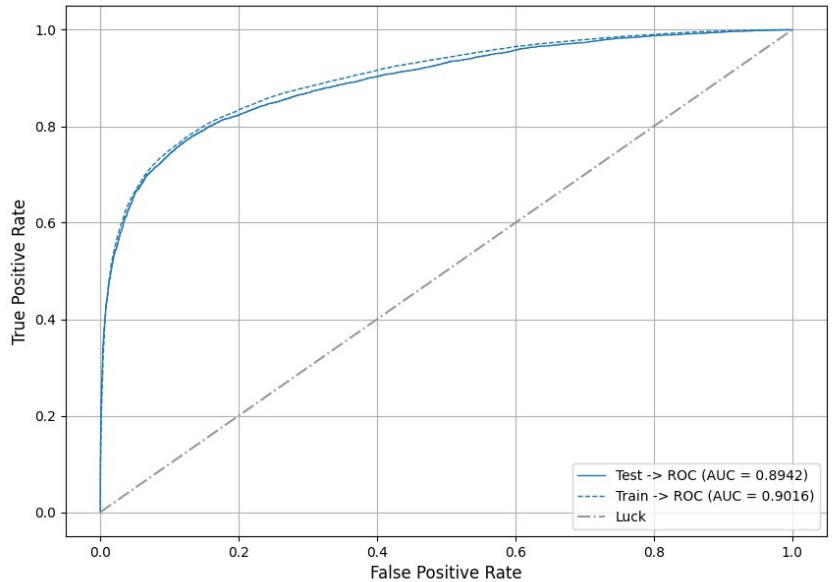
$$TPR = \frac{True\ Positive\,(TP)}{True\ Positive\,(TP) + False\ Negative\,(FN)}$$

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

AUC: Area Under Curve



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



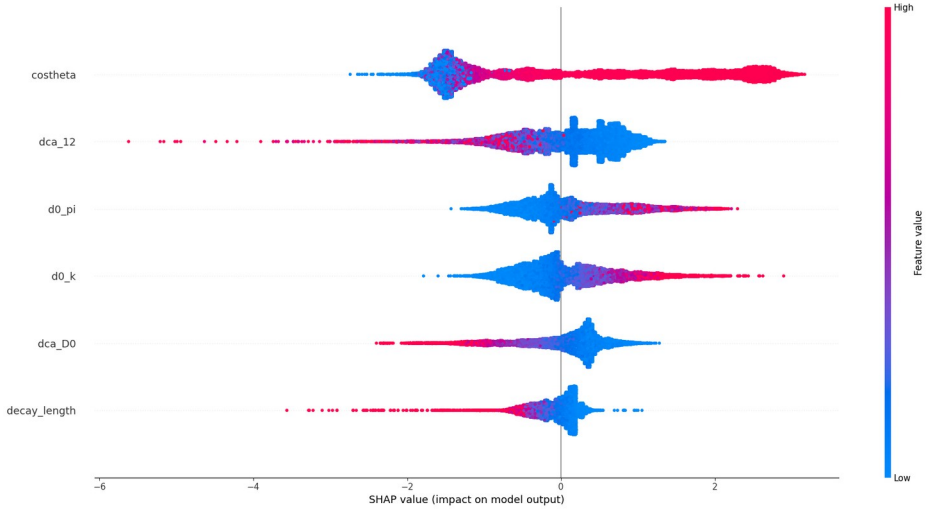
Receiver Operating Characteristic (ROC)

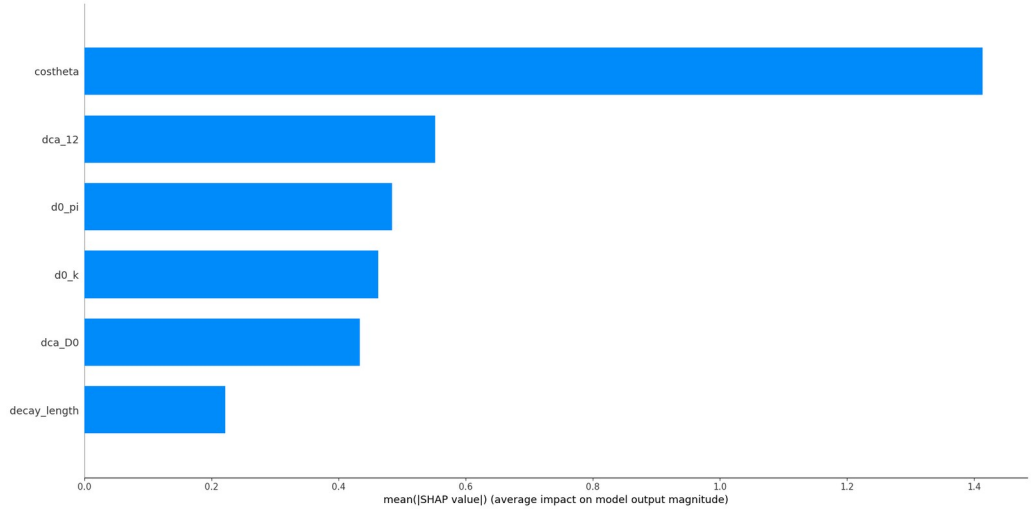
A perfect classifier would have a point at (0, 1), indicating no false positives and all true positives

Features of Importance (Training)

SHAP (SHapley Additive exPlanations)

Concept of Game theory in Mathematics





$$y_i = E[f(X)] + \sum_i SHAP_i$$

y_i: Model's prediction for i^{th} sample
E[f(X)] = Model's average prediction for entire data set
Sum of SHAP of individual feature's contribution

For the classifier y_i is transformed in probability using Softmax function

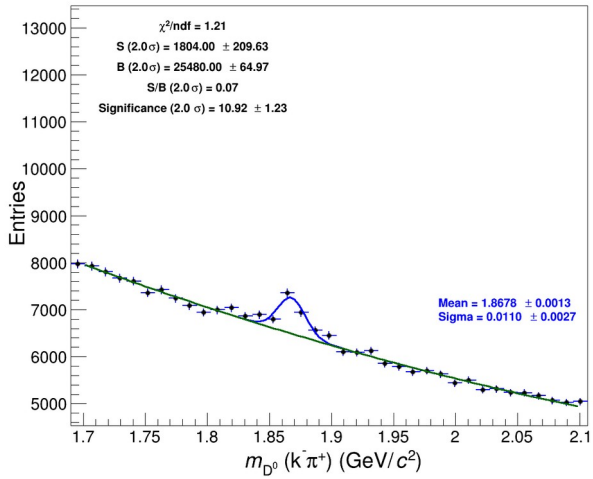
Estimation of Signal and Background

Signal: Gauss+pol2
Bkg:pol2
Sum: Gauss+pol2

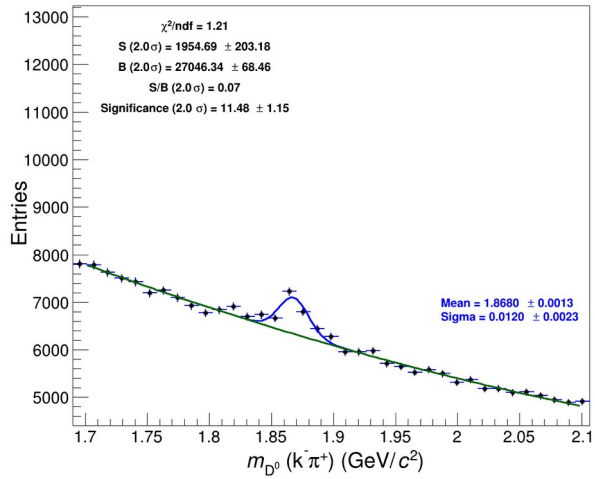
Signal weighted by 0.01 while background same

$$Signif = \frac{S}{\sqrt{S+B}}$$

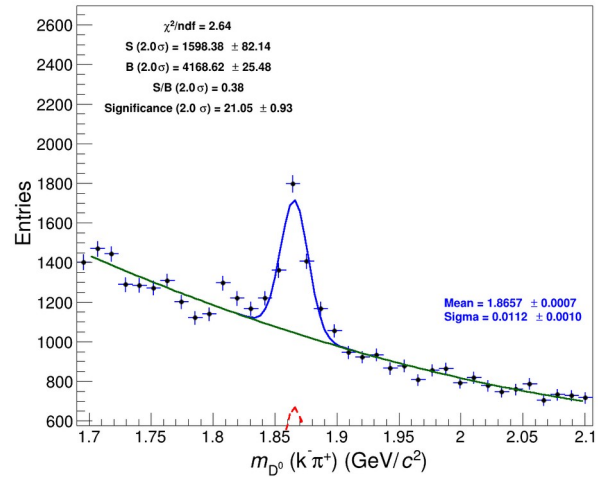
Threshold > 0.00



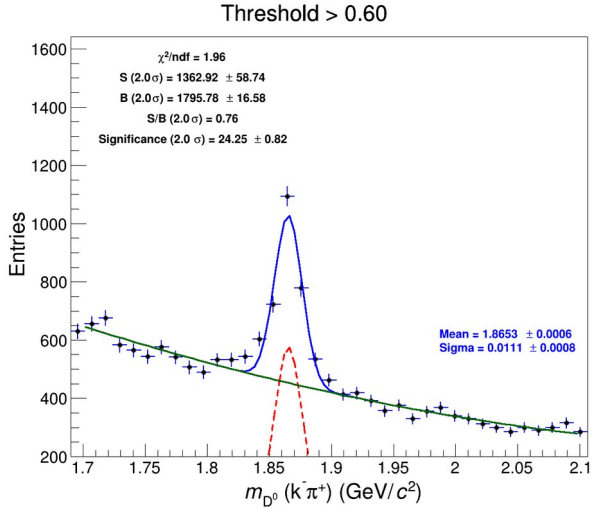
Threshold > 0.02



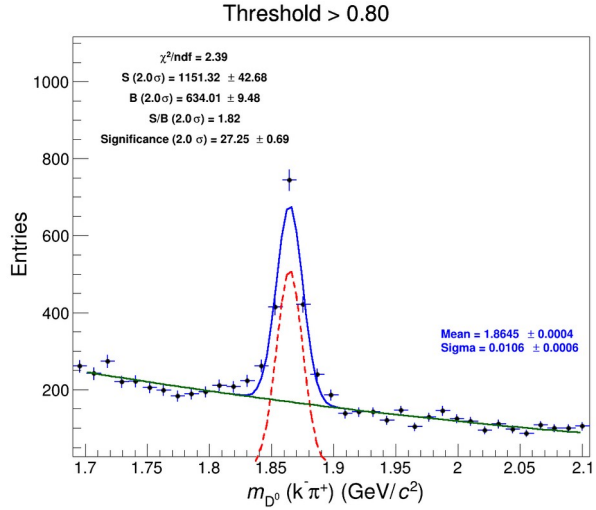
Threshold > 0.40



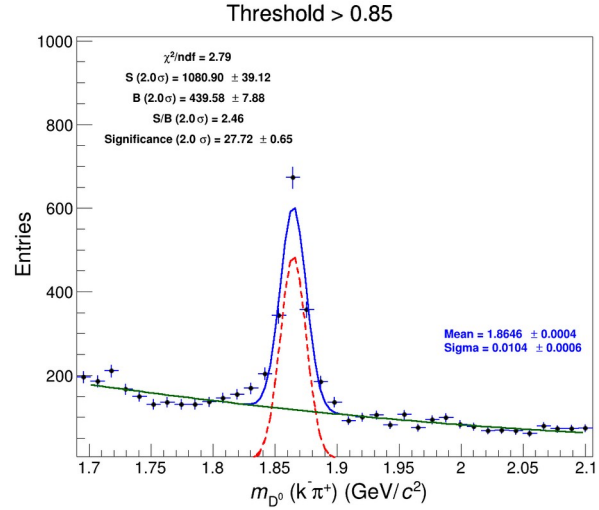
Threshold > 0.60



Threshold > 0.80



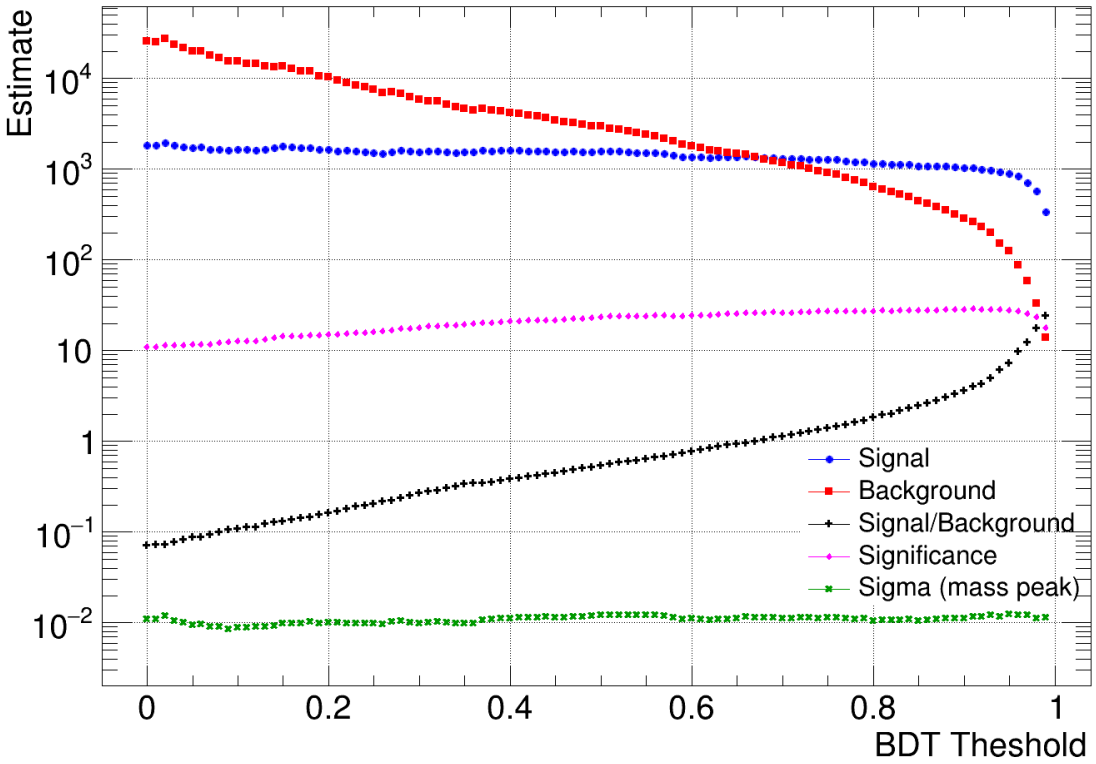
Threshold > 0.85



Estimation of Signal and Background (Q^2 = 100)

Signal scaled by 0.01 while background same

PYTHIA simulation for 10x100 ep collisions with $Q^2 > 100$: σ ~ 1.3e-6 mb, Expected luminosity of 5 fb^{-1}



$$N_{event} = L_{\text{int}} \times \sigma = 5 \times 10^{15} \times 1.3 \times 10^{-9} = 6.5\,M$$

For BDT > 0.8,D0 significance is 27 for 5M DIS events

$$\text{Expected Significance} \sim 27.0 * \sqrt{\left(\frac{6.5}{5}\right)} = 30.78$$

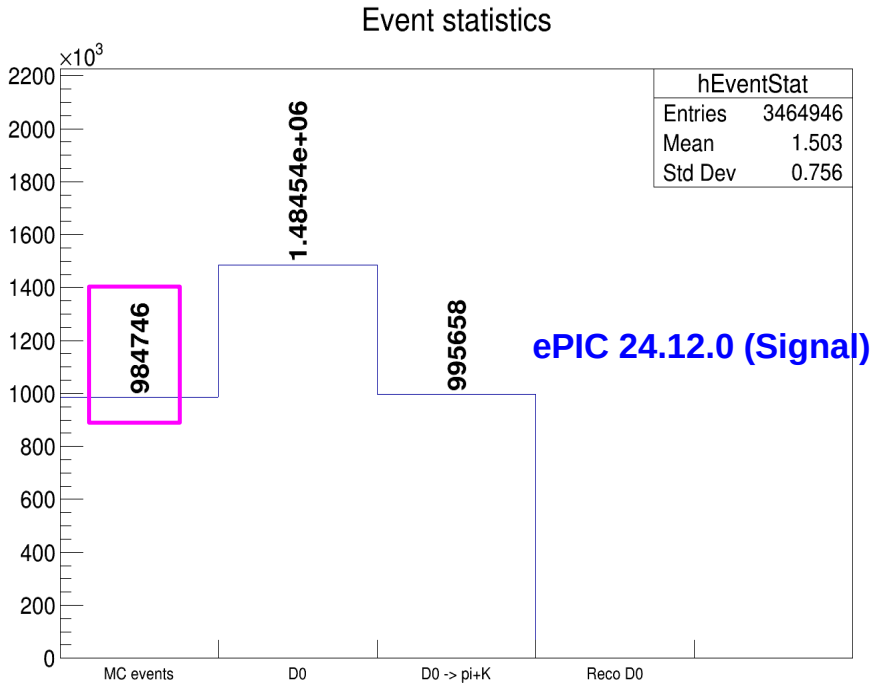
Thanks Rongronga

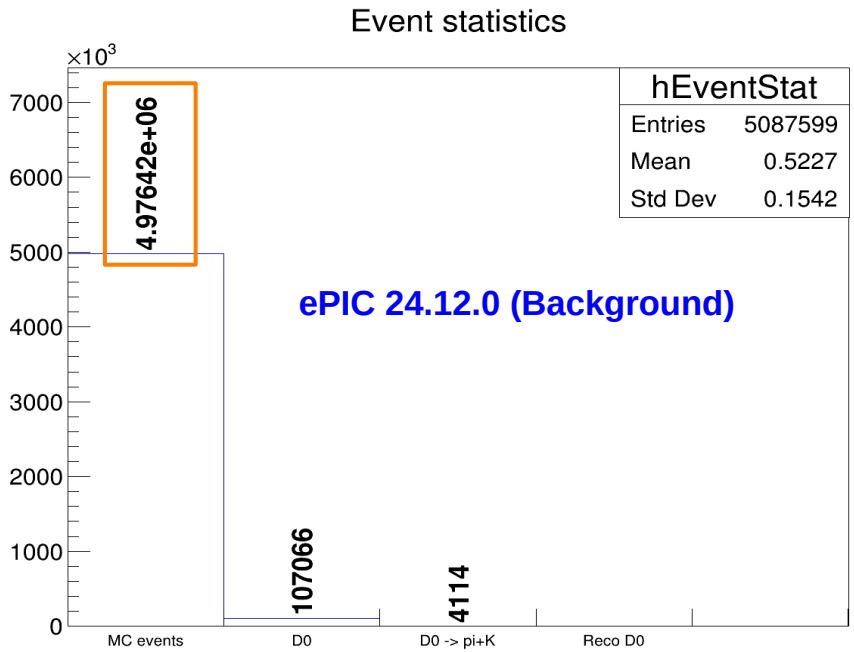
Note: BDT cut optimisation done on data (bias) but in general we need to estimate signal from MC and background from data: Need to develop a way for it

Data Sample for ML (Q^2 = 1)

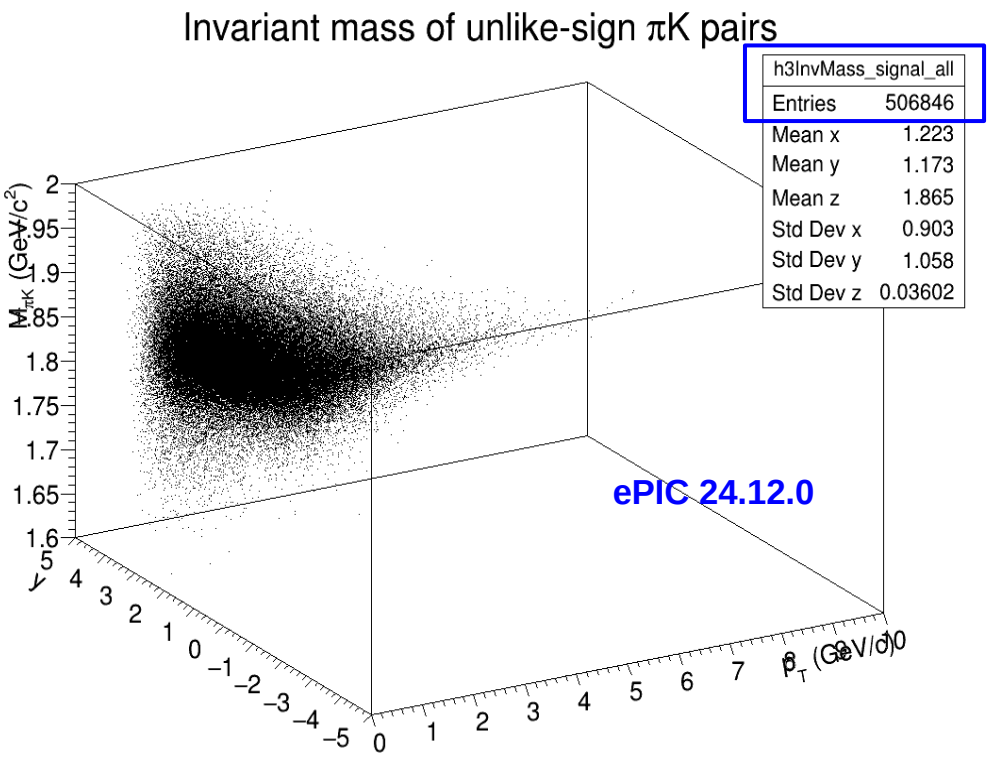
- BDT requires the features for the signal D^0 meson and background D^0 meson (fake combinations of pion,kaon)
  - D^0 enriched same created filtering **PYTHIA8 ep, NC, 10X100, Q^2 >1 events (~1747 M)** such that each event consist one $D^0 \rightarrow$ k-π+ known as Signal taken from 24.12.0/epic_craterlake/SIDIS/D0_ABCONV/pythia8.306-1.1/10x100/q2_1):

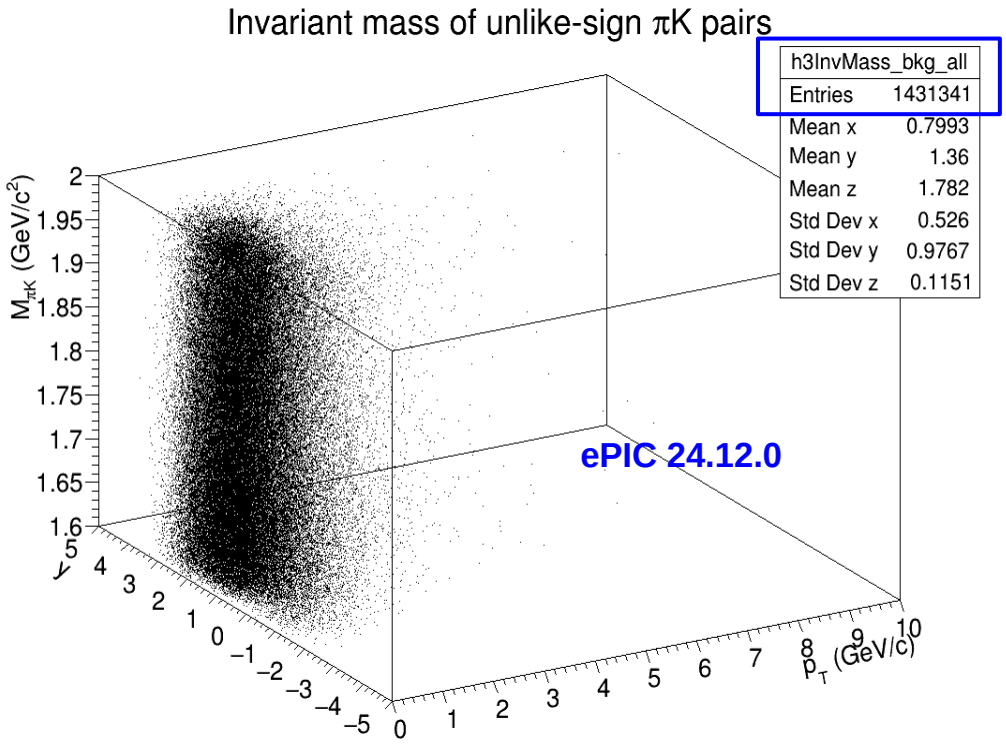
    Total files 1879 and Events = 984746
  - Background from 24.12.0/epic_craterlake/DIS/NC/10x100/minQ2=1: **Total files 5180 and Events = 4976419**





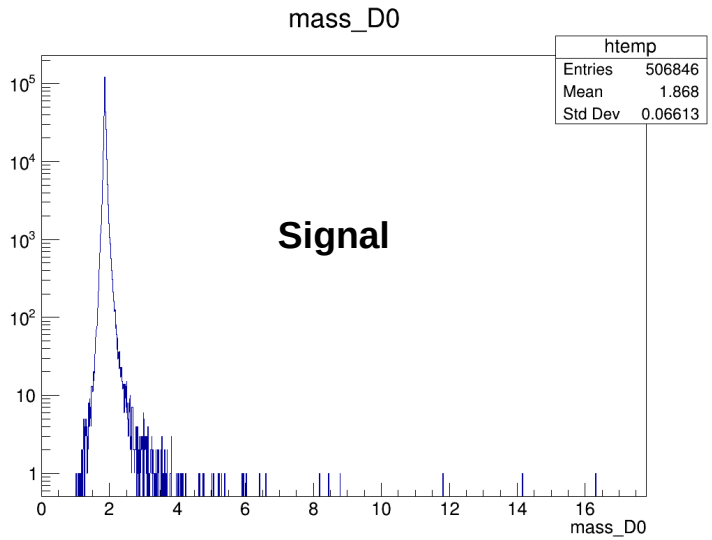
Implementation of ML Model (Q^2 = 1)





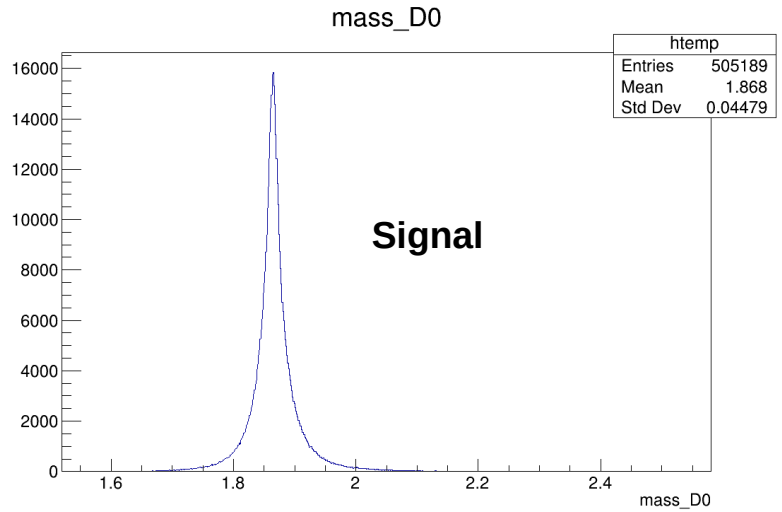
Filtering data (Q^2 = 1)

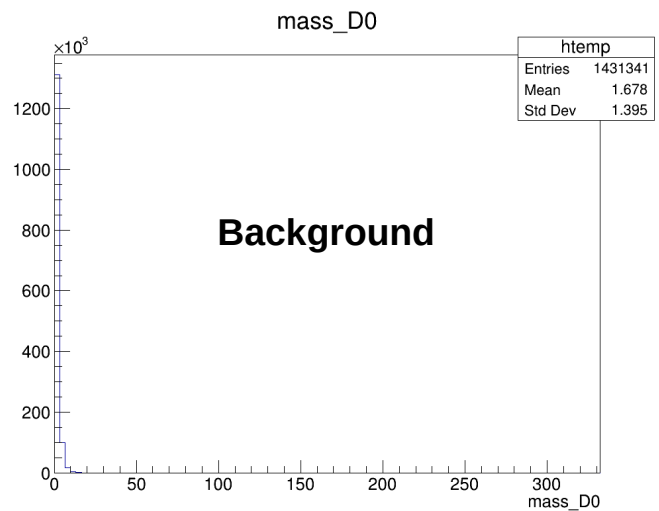
Filtering: (mass_D0 > 1.6 && mass_D0 < 2.5) && (d0xy_pi>0.02 && d0xy_pi<10.) && (d0xy_k>0.02 && d0xy_k<10.) && decay_length <100.



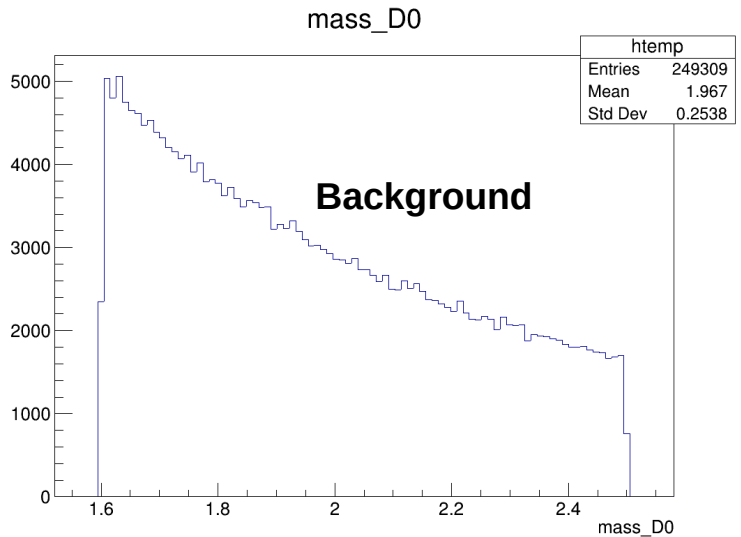
Dimension: mm

Filtering

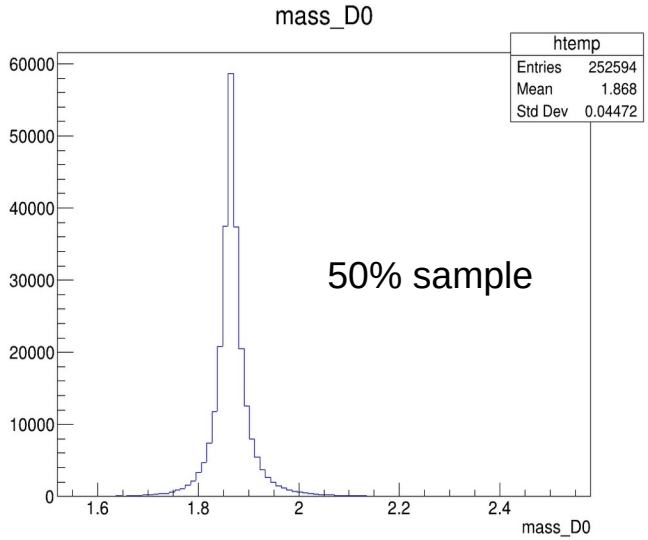


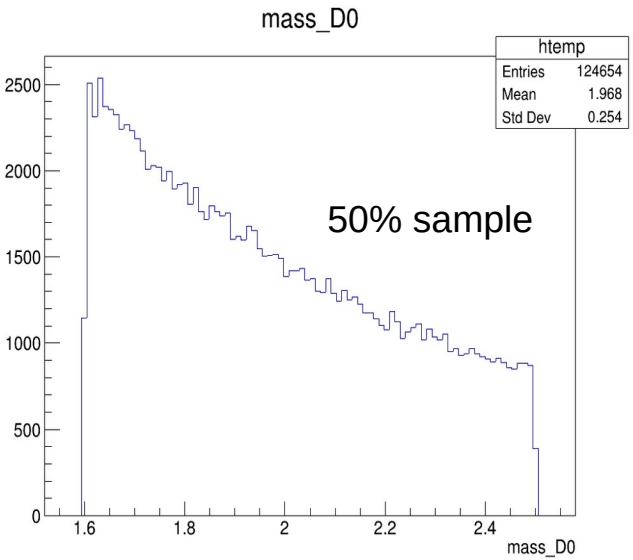


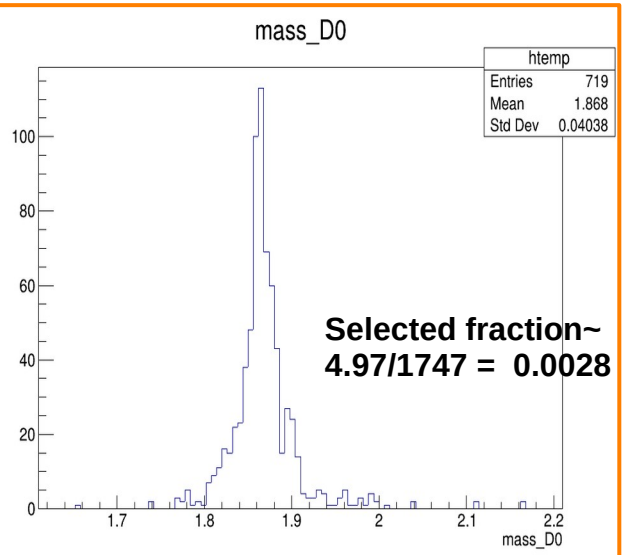
Filtering

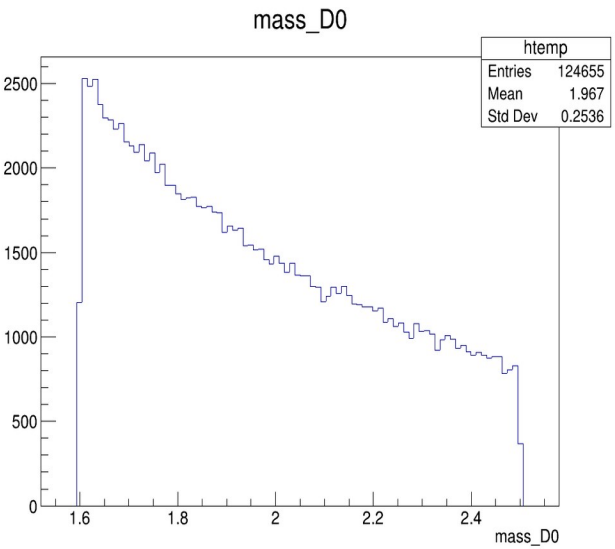


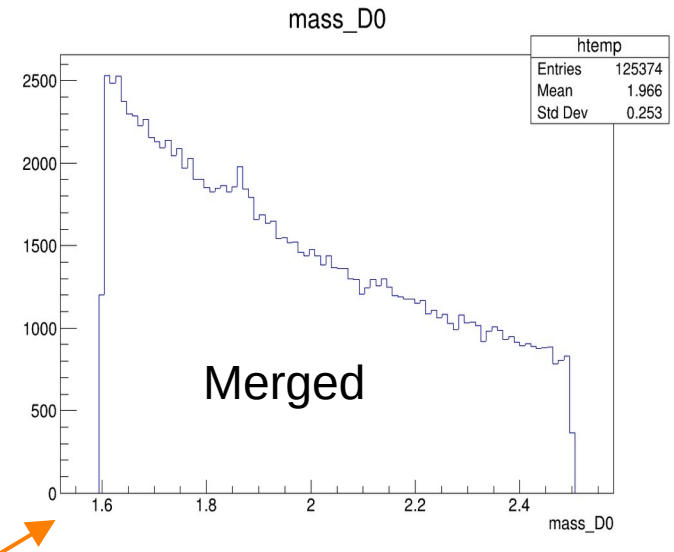
Sample Preparation (Q^2 = 1)







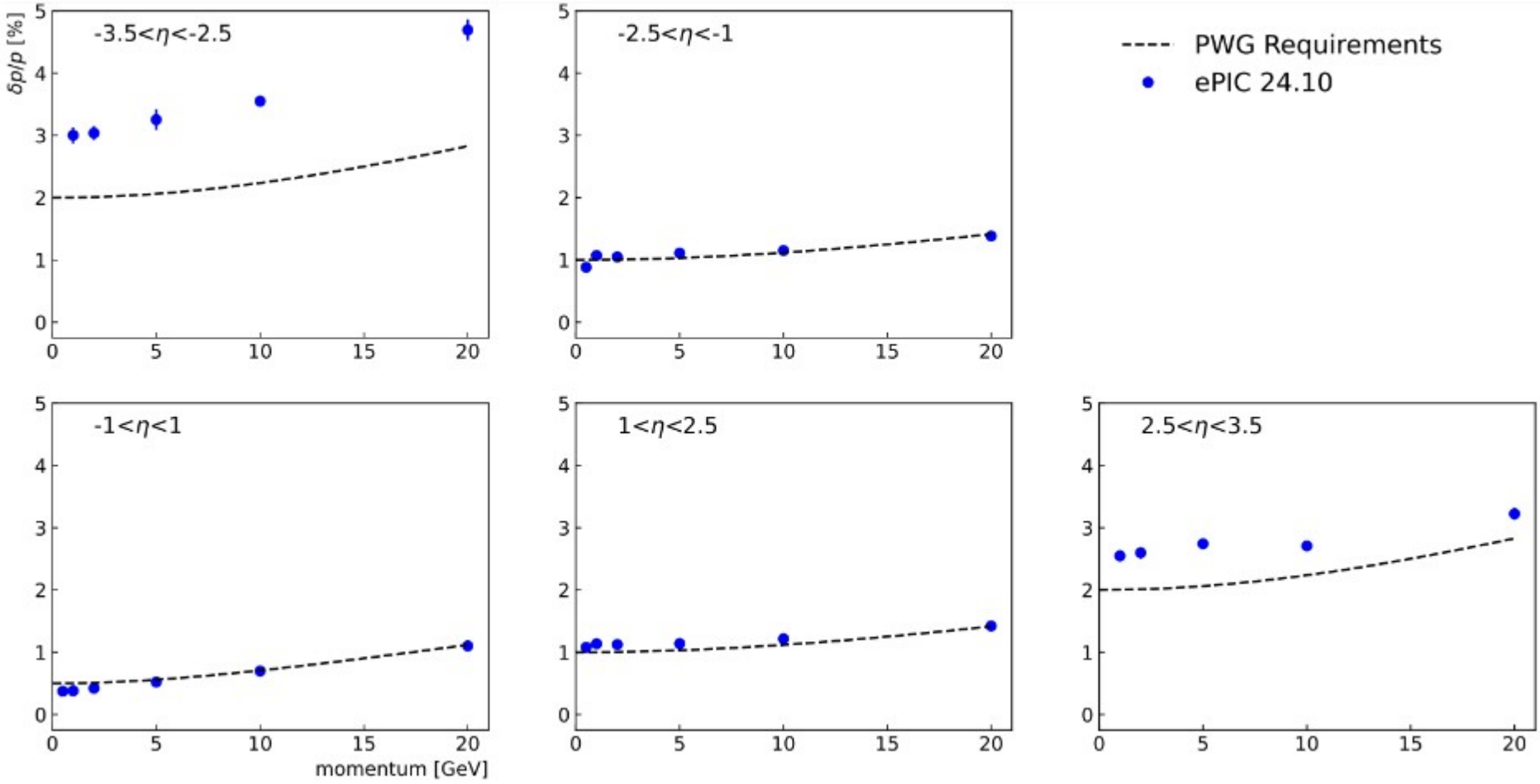




- PYTHIA8 ep, NC, 10X100, Q^2 >1: signal events (~1747M) while background events (~4.97M)
- Signal and Bkg divided (50% for ML and 50% Application)
- Signal scaled by a factor 0.01 to consider unbalance in entries

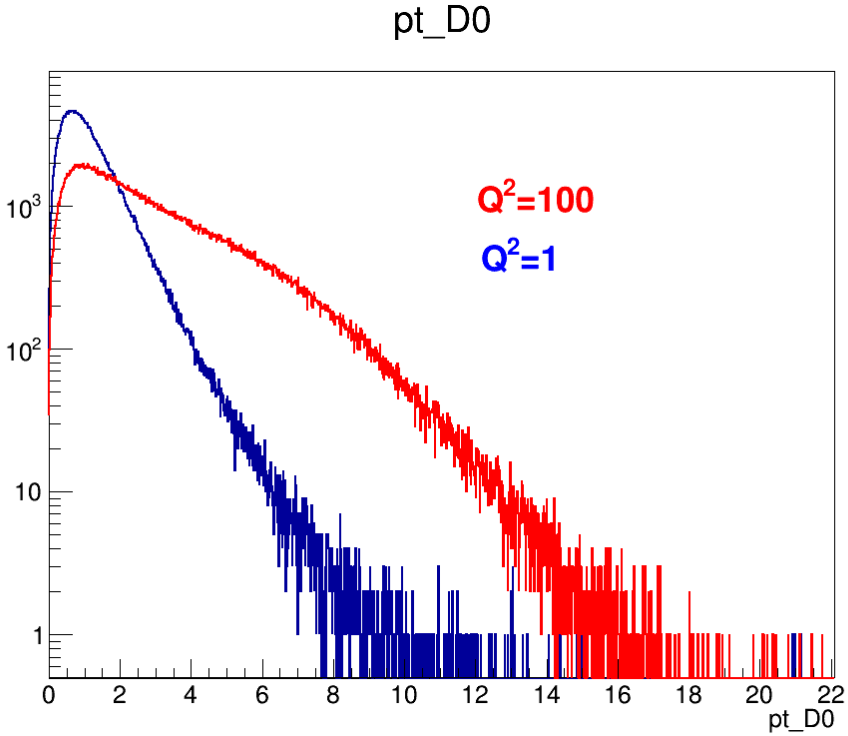
Momentum Resolutions (TDR)

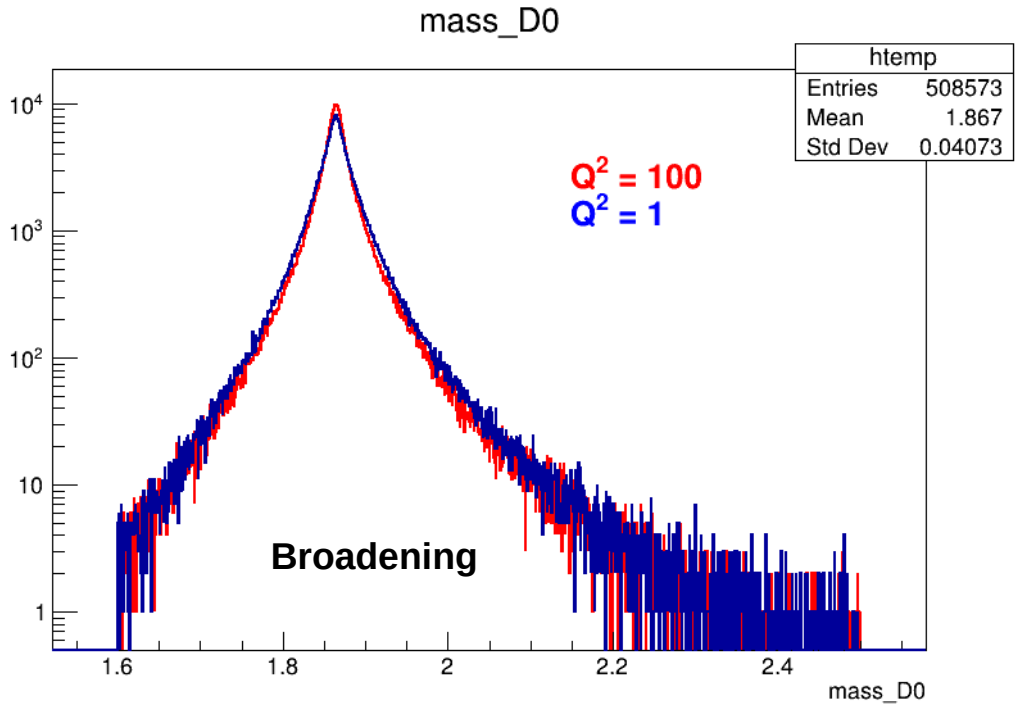
https://zenodo.org/records/14328280



Basic Distributions Comparisons

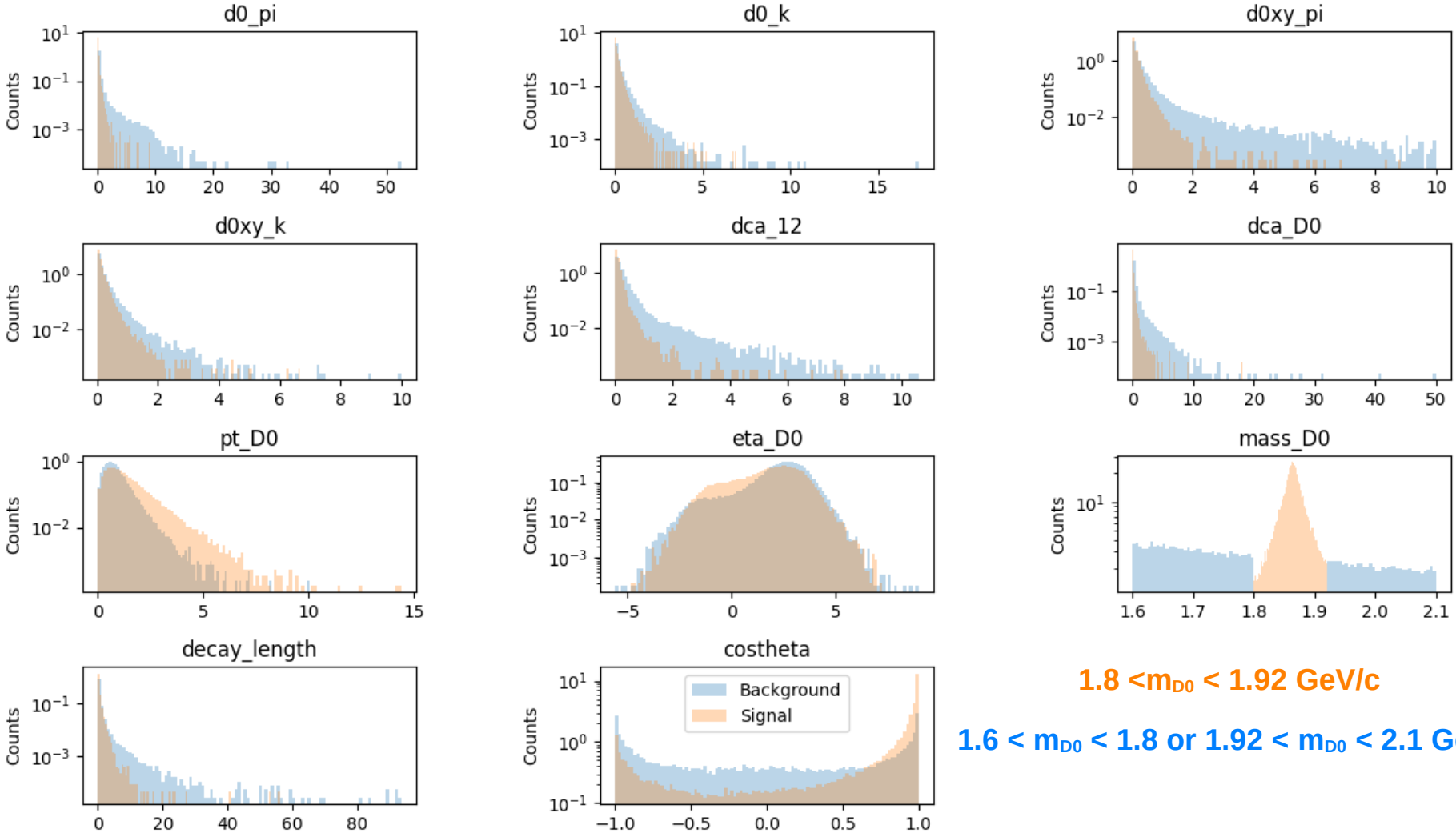
p_T and inavriant mass of signal D^0 meson: Broadening due to large momentum resolutions for η [1.0,3.5]





Signal and Background Distributions ($Q^2 = 1$)

Using similar approach as $Q^2>100$

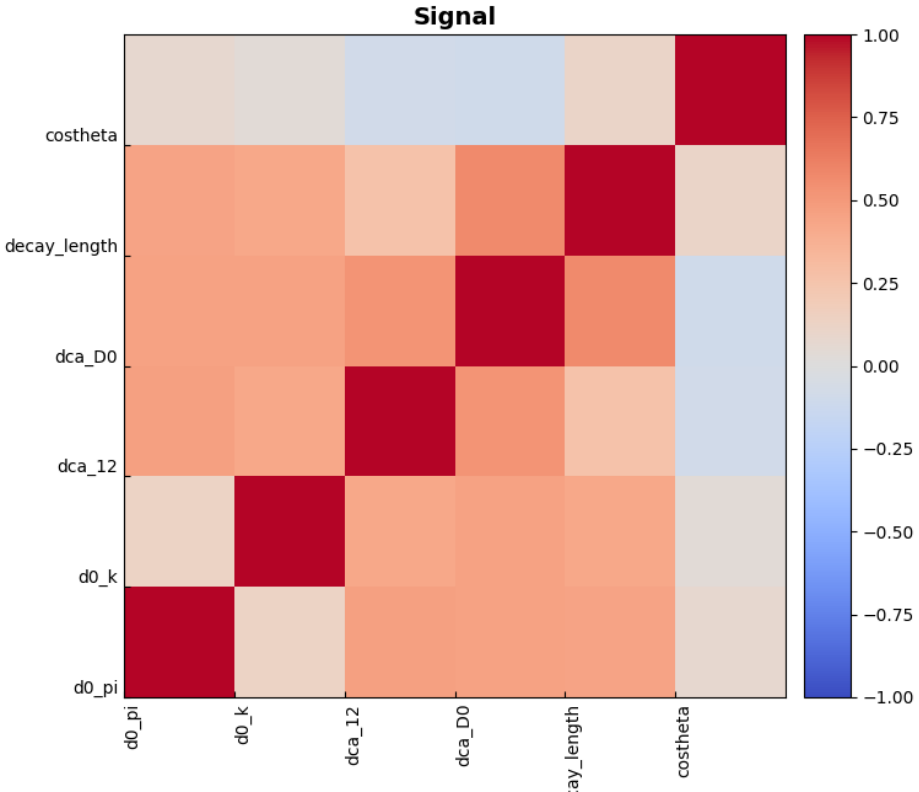


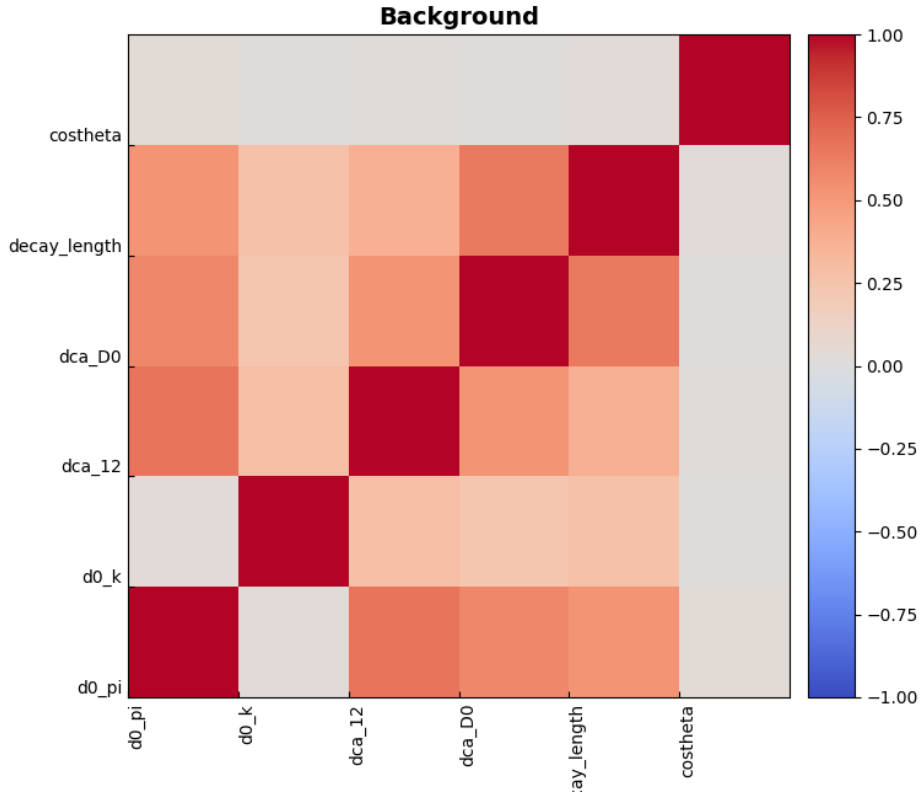
$1.8 < m_{D0} < 1.92$ GeV/c

$1.6 < m_{D0} < 1.8$ or $1.92 < m_{D0} < 2.1$ GeV/c

Feature Correlations

Highly correlated variable carries similar information, one of them can be removed while training the model





Model Performances

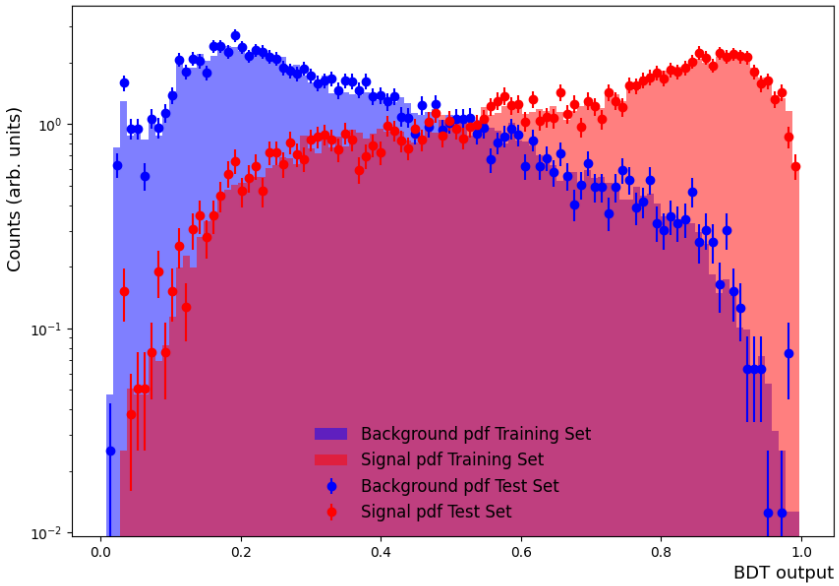
Model can be saved to .onnx format

	P	N
T	TP	TN
F	FP	FN

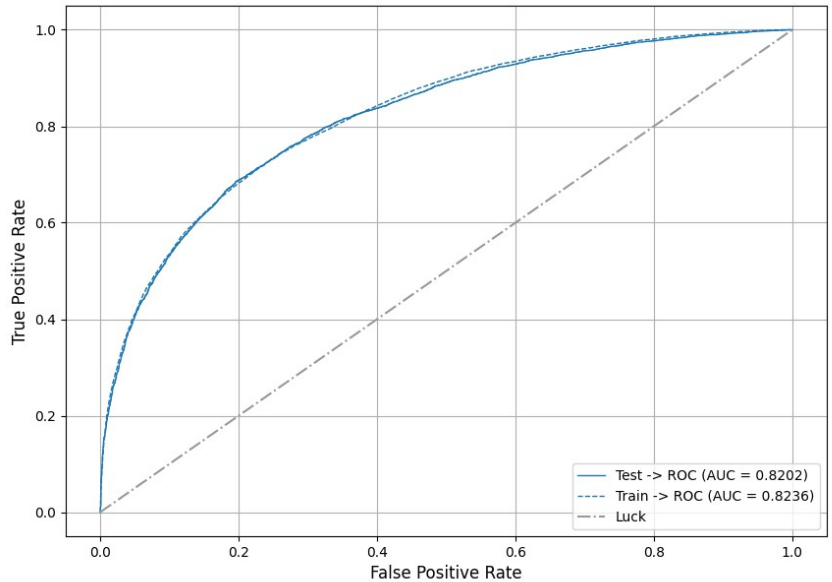
$$TPR = \frac{True\ Positive\,(TP)}{True\ Positive\,(TP) + False\ Negative\,(FN)}$$

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

AUC: Area Under Curve



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



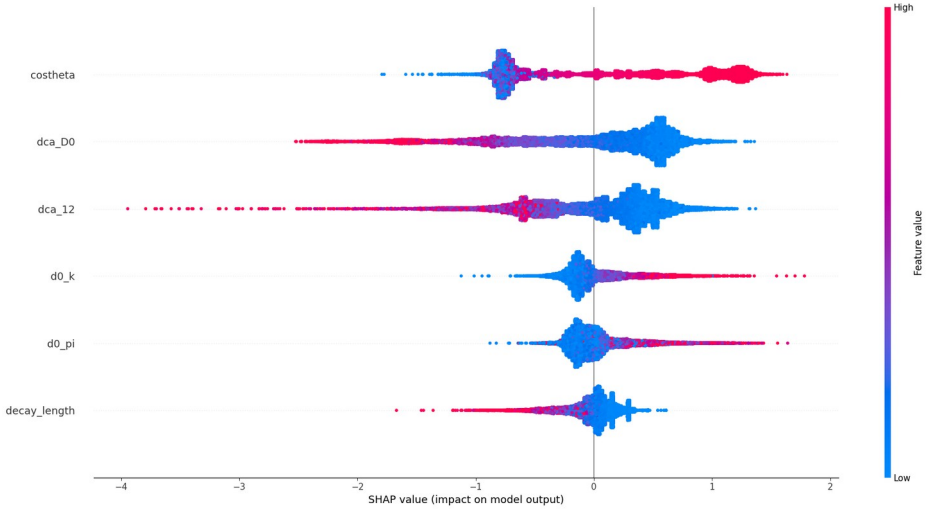
Receiver Operating Characteristic (ROC)

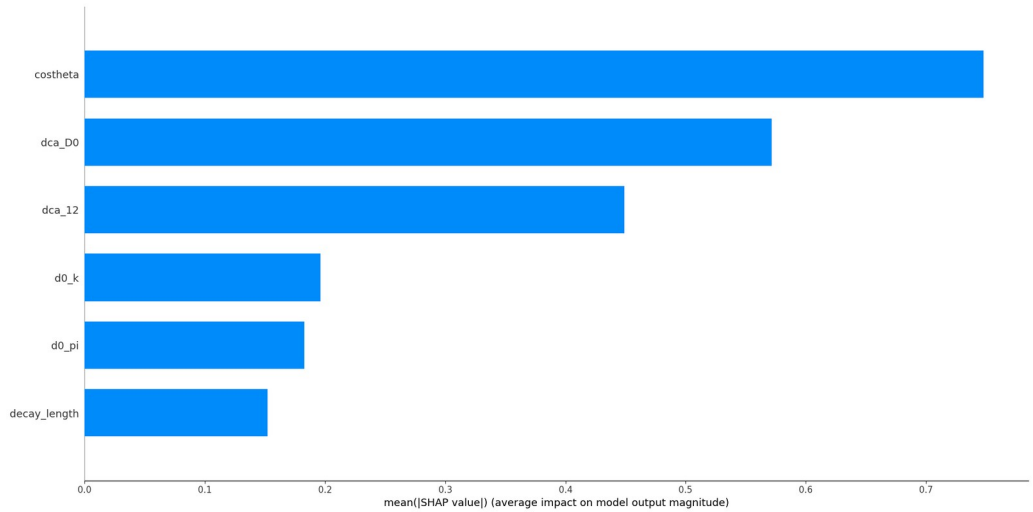
A perfect classifier would have a point at (0, 1), indicating no false positives and all true positives

Features of Importance (Training)

SHAP (SHapley Additive exPlanations)

Concept of Game theory in Mathematics





$$y_i = E[f(X)] + \sum_i SHAP_i$$

y_i: Model's prediction for i^{th} sample
E[f(X)] = Model's average prediction for entire data set
Sum of SHAP of individual feature's contribution

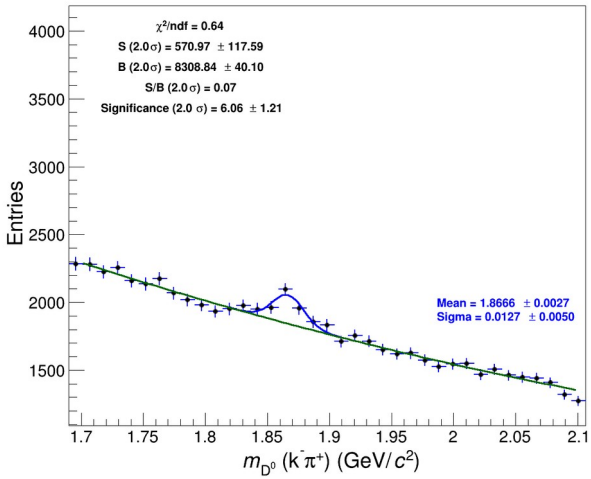
Estimation of Signal and Background

Signal: Gauss+pol2
Bkg:pol2
Sum: Gauss+pol2

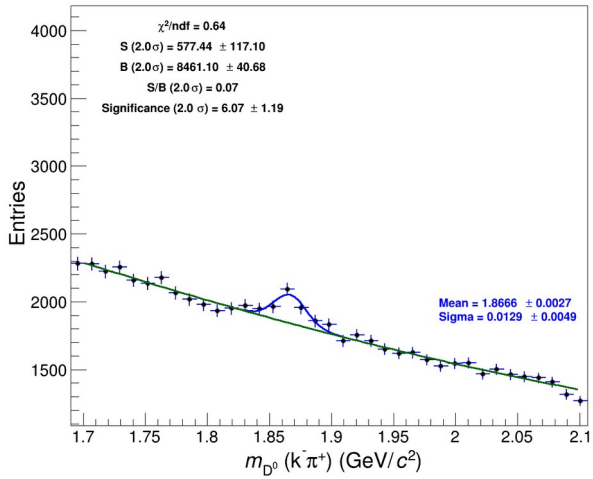
Signal weighted by 0.01 while background same

$$Signif = \frac{S}{\sqrt{S+B}}$$

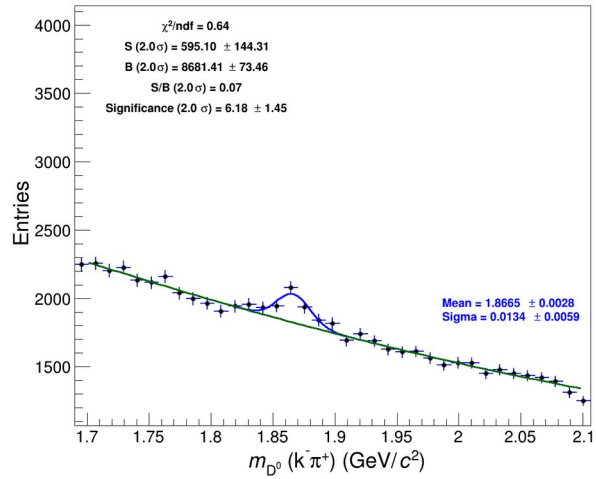
Threshold > 0.00



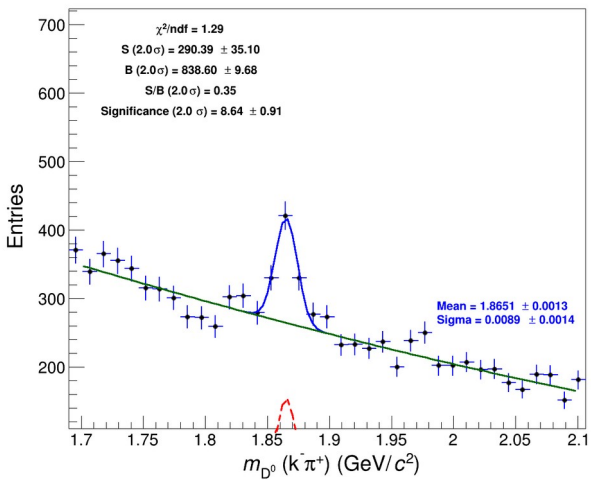
Threshold > 0.02



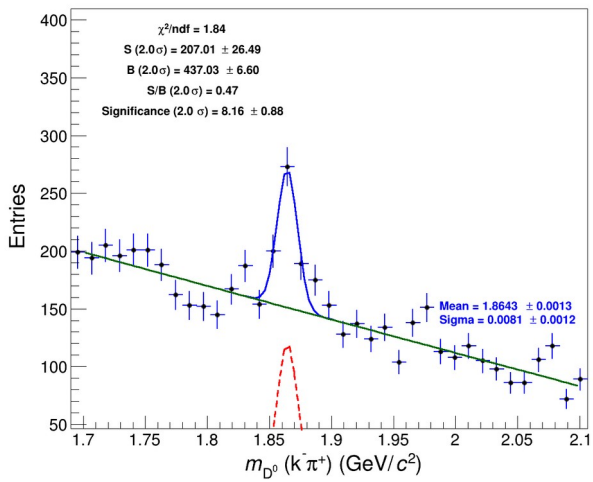
Threshold > 0.03



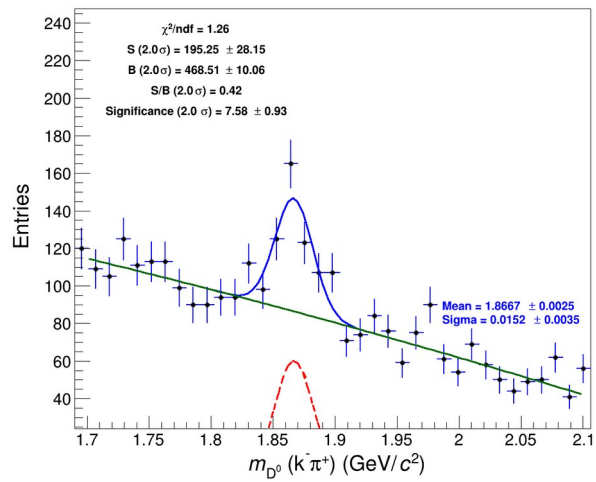
Threshold > 0.60



Threshold > 0.70

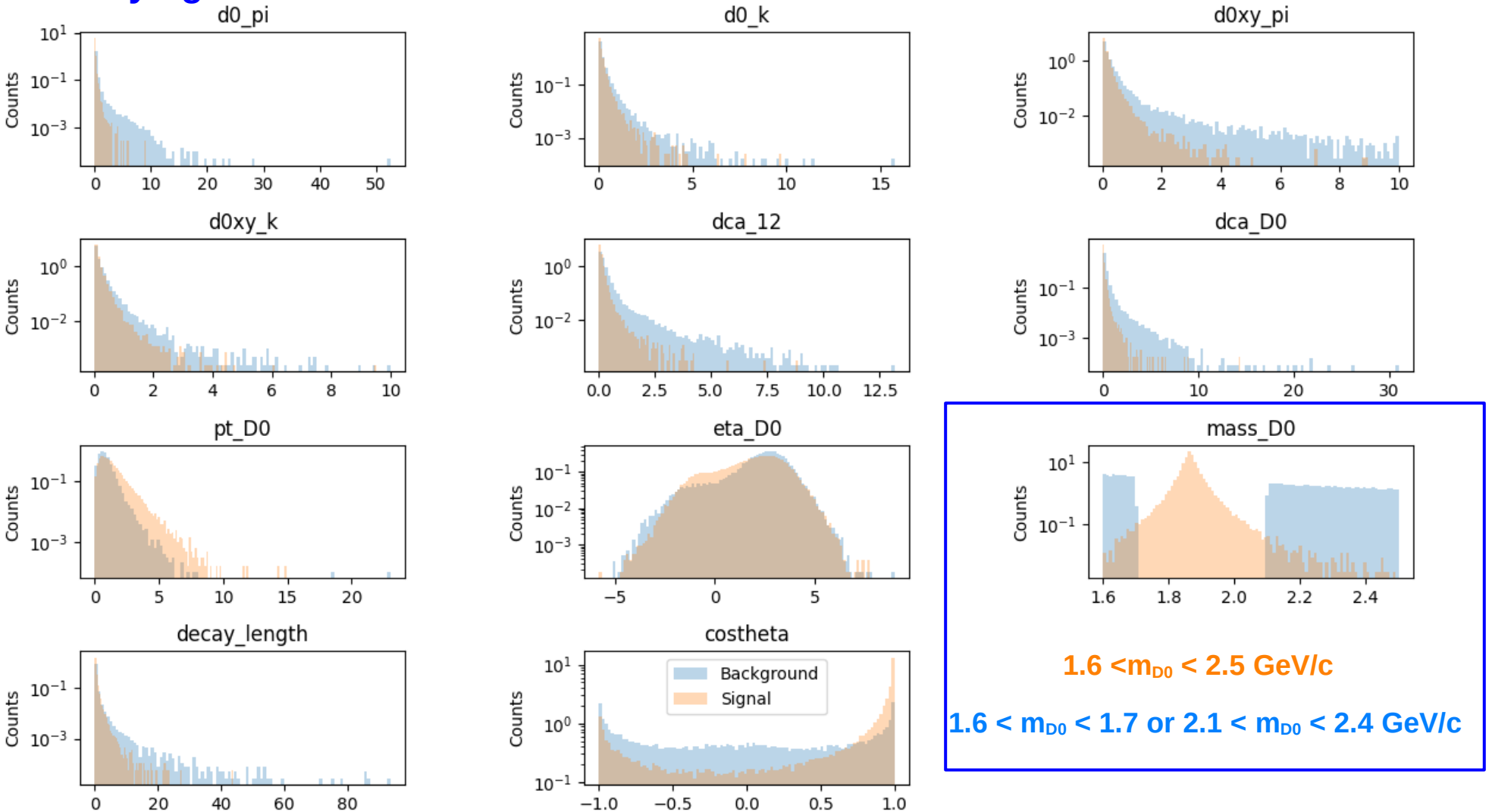


Threshold > 0.77



Signal and Background Distributions (Q^2 = 1)

Modifying mass cut

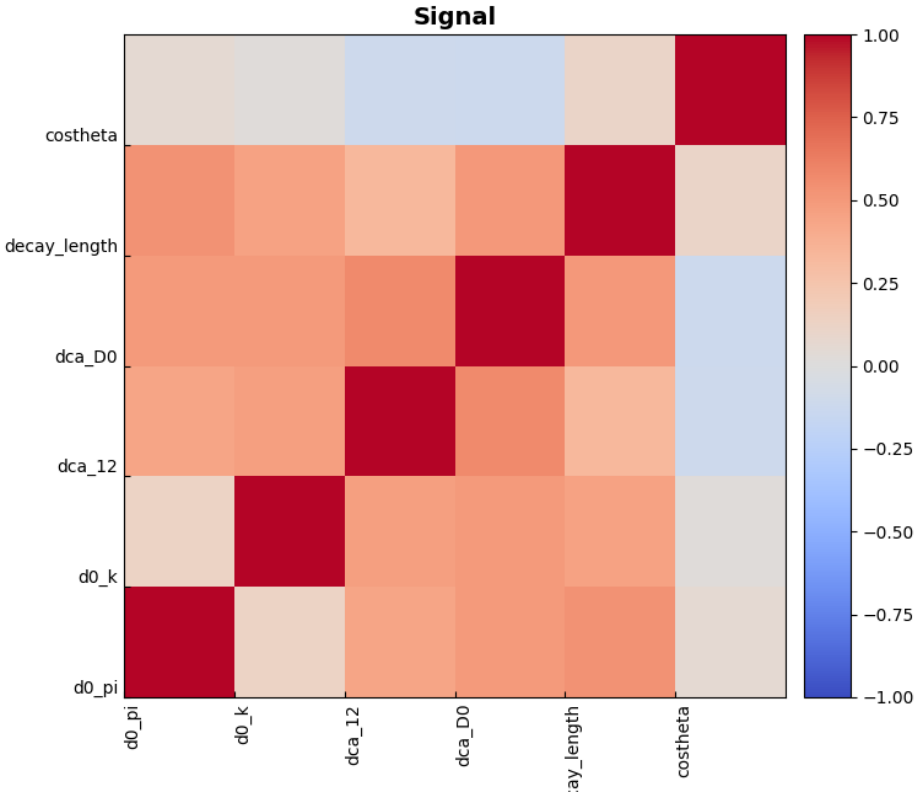


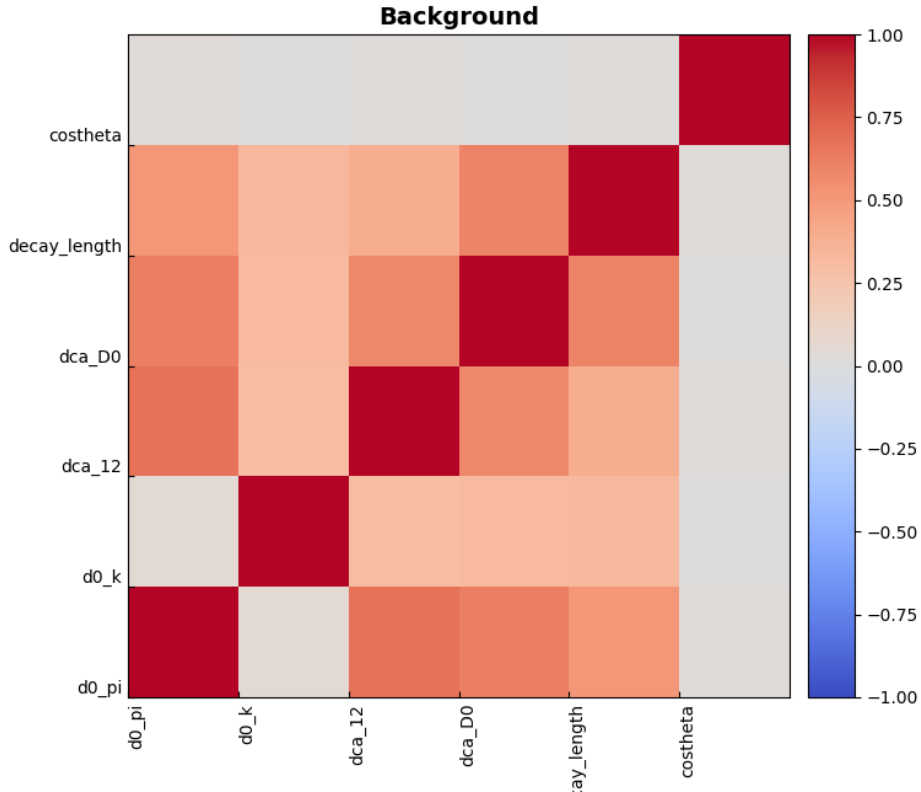
$1.6 < m_{D0} < 2.5$ GeV/c

$1.6 < m_{D0} < 1.7$ or $2.1 < m_{D0} < 2.4$ GeV/c

Feature Correlations

Highly correlated variable carries similar information, one of them can be removed while training the model





Model Performances

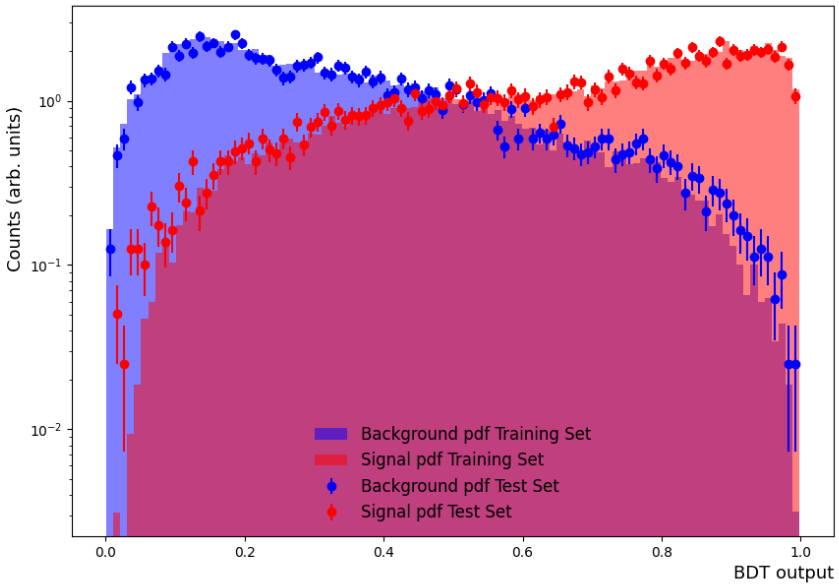
Model can be saved to .onnx format

	P	N
T	TP	TN
F	FP	FN

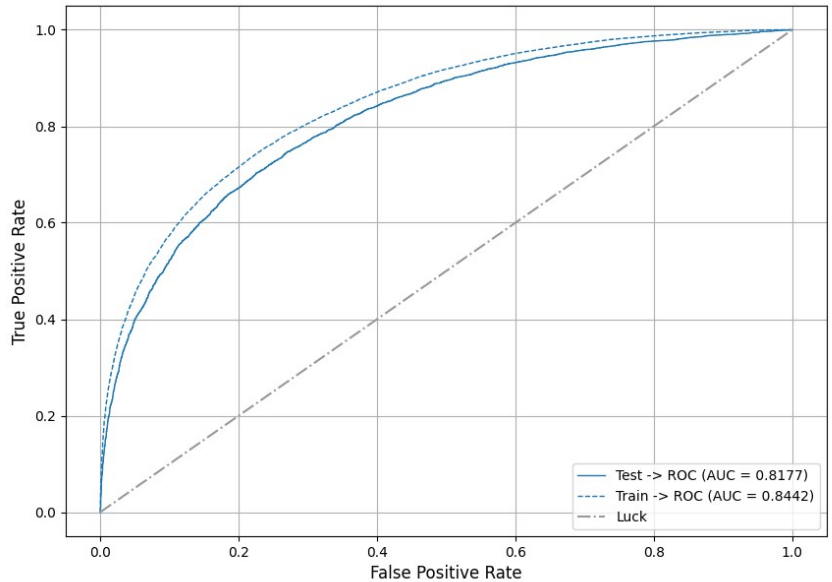
$$TPR = \frac{True\ Positive\,(TP)}{True\ Positive\,(TP) + False\ Negative\,(FN)}$$

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

AUC: Area Under Curve



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



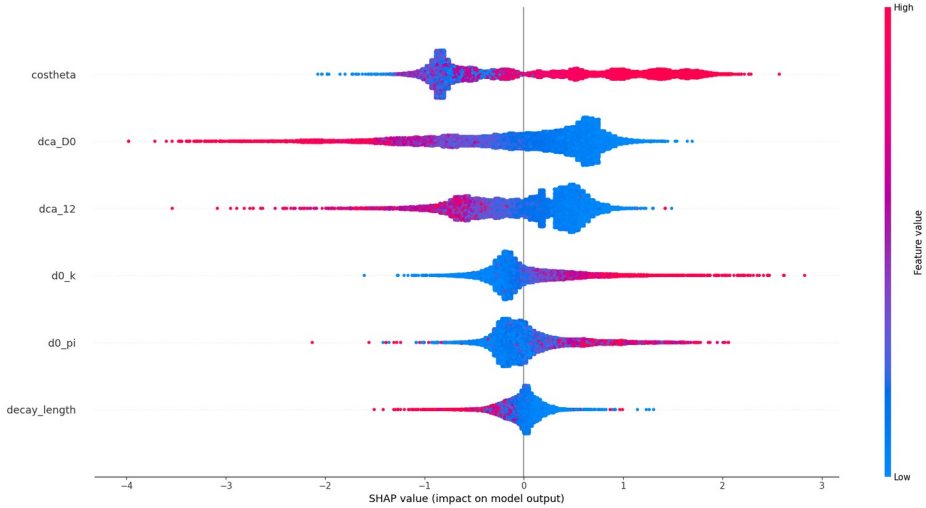
Receiver Operating Characteristic (ROC)

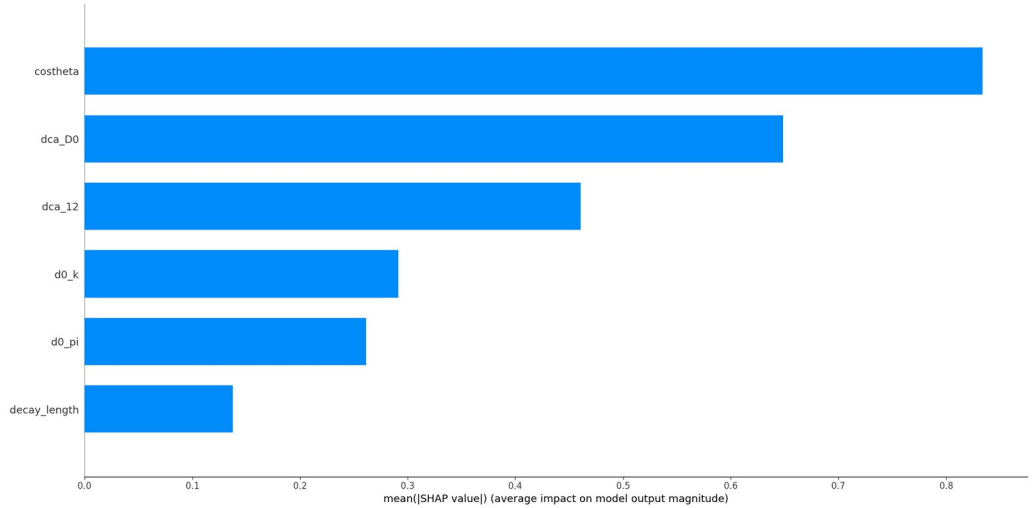
A perfect classifier would have a point at (0, 1), indicating no false positives and all true positives

Features of Importance (Training)

SHAP (SHapley Additive exPlanations)

Concept of Game theory in Mathematics





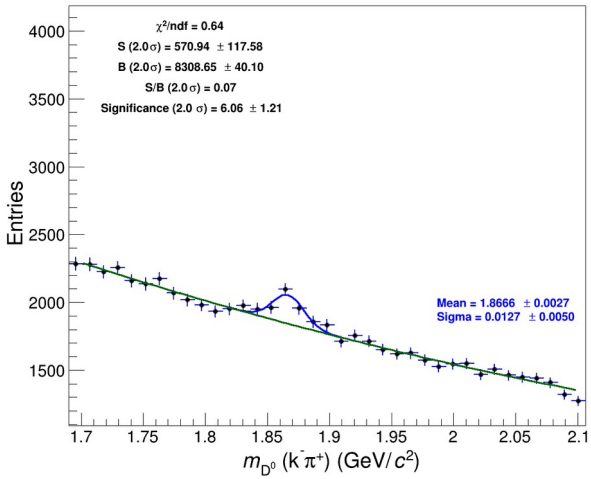
Estimation of Signal and Background

Signal: Gauss+pol2
Bkg:pol2
Sum: Gauss+pol2

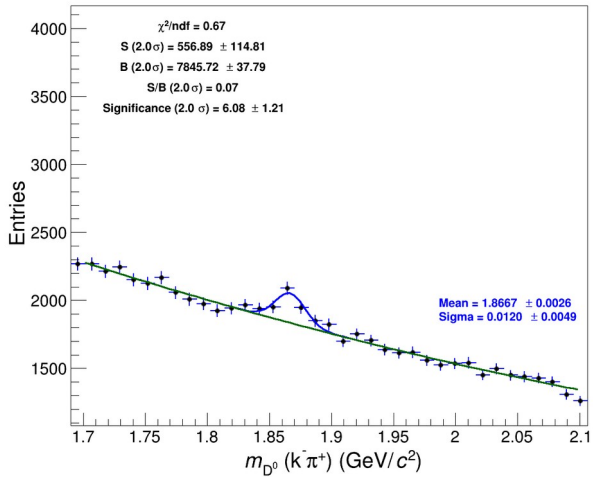
Signal weighted by 0.01 while background same

$$Signif = \frac{S}{\sqrt{S+B}}$$

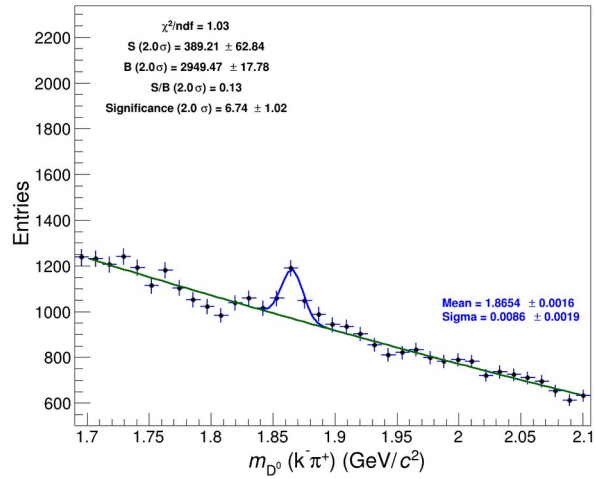
Threshold > 0.00



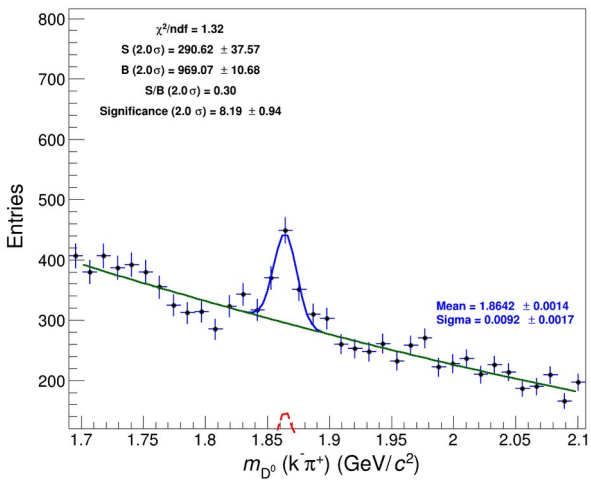
Threshold > 0.02



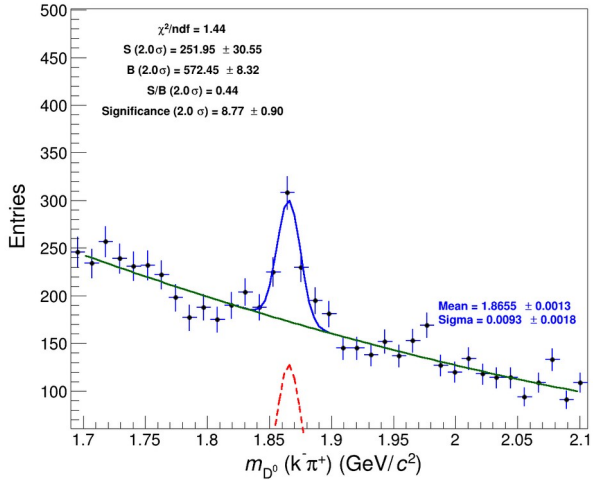
Threshold > 0.30



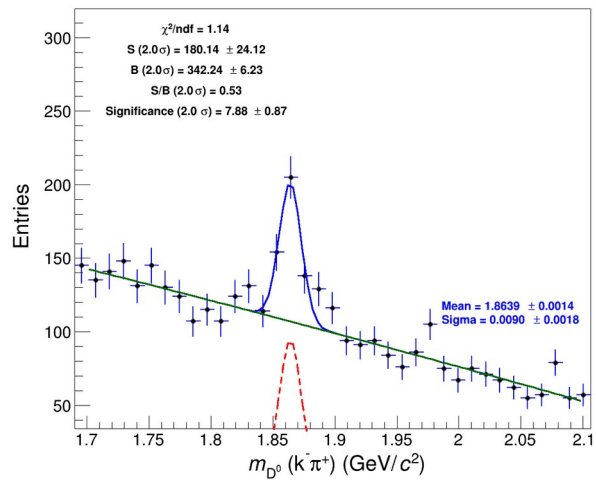
Threshold > 0.60



Threshold > 0.70

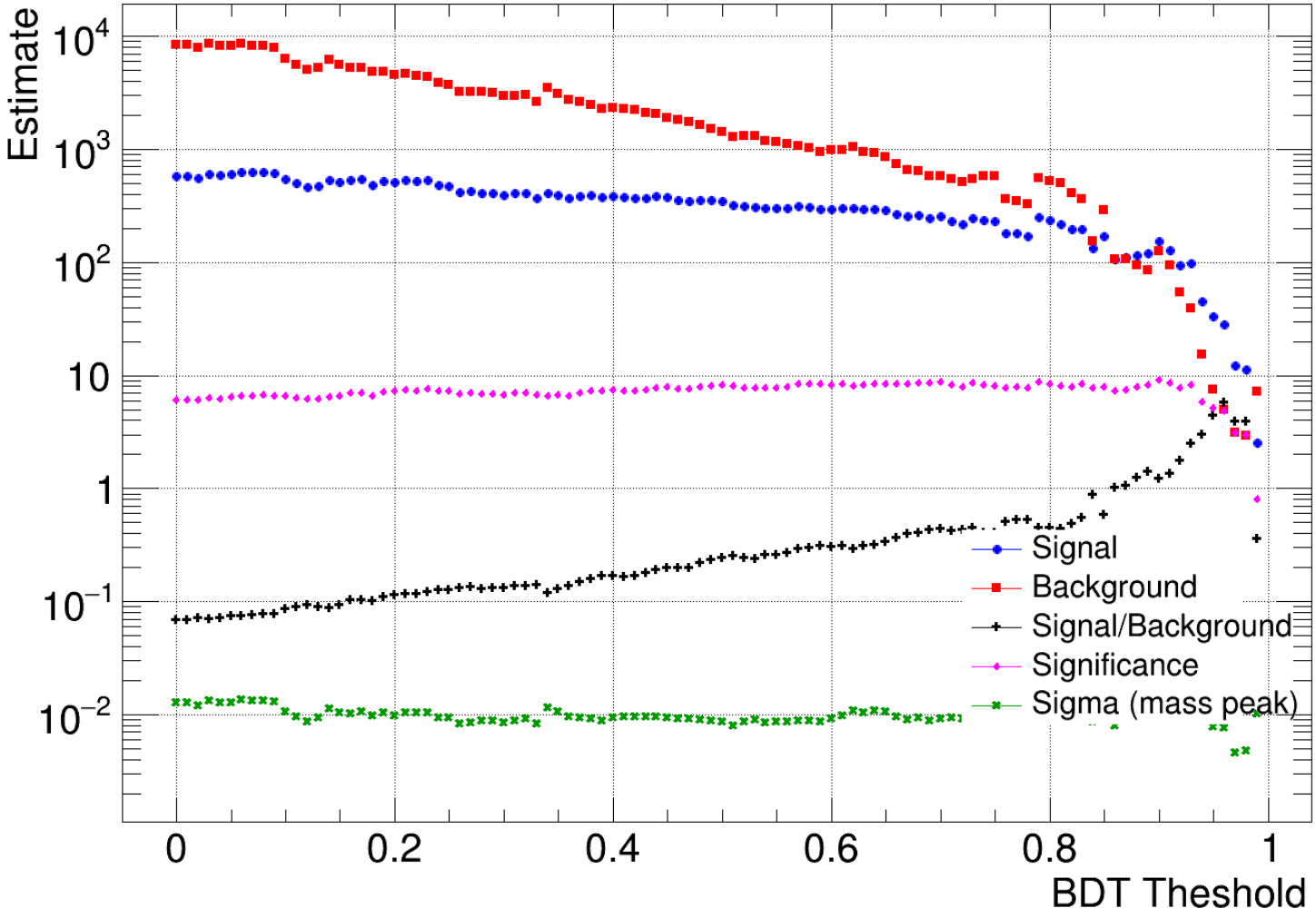


Threshold > 0.77



Estimation of Signal and Background

Signal scaled by 0.01 while background same



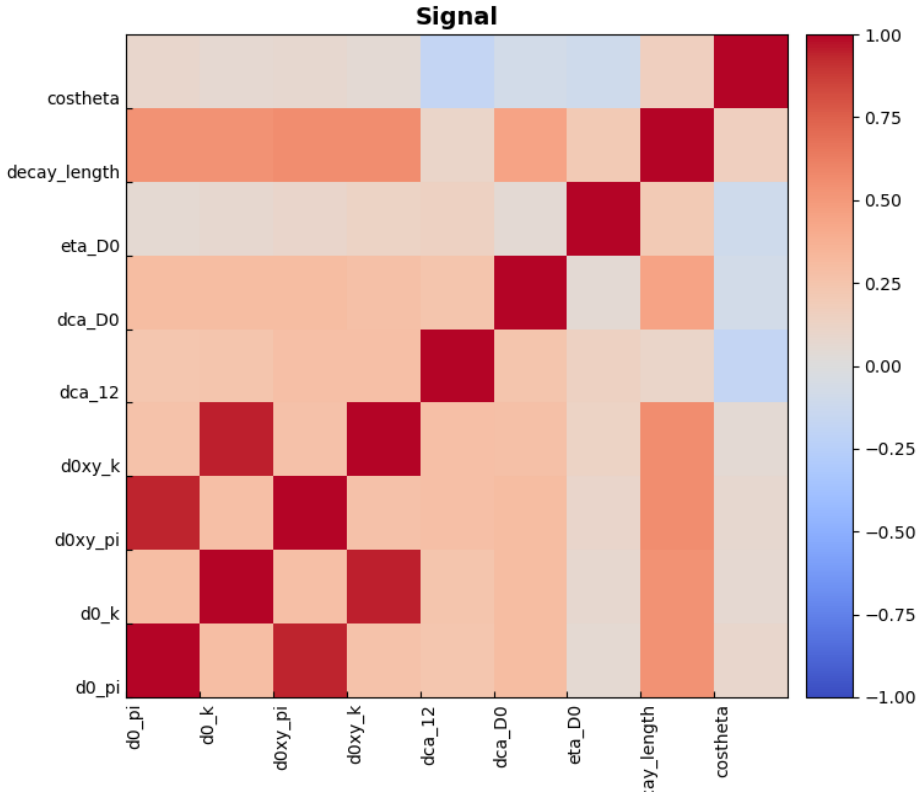
Summary and Future Plan

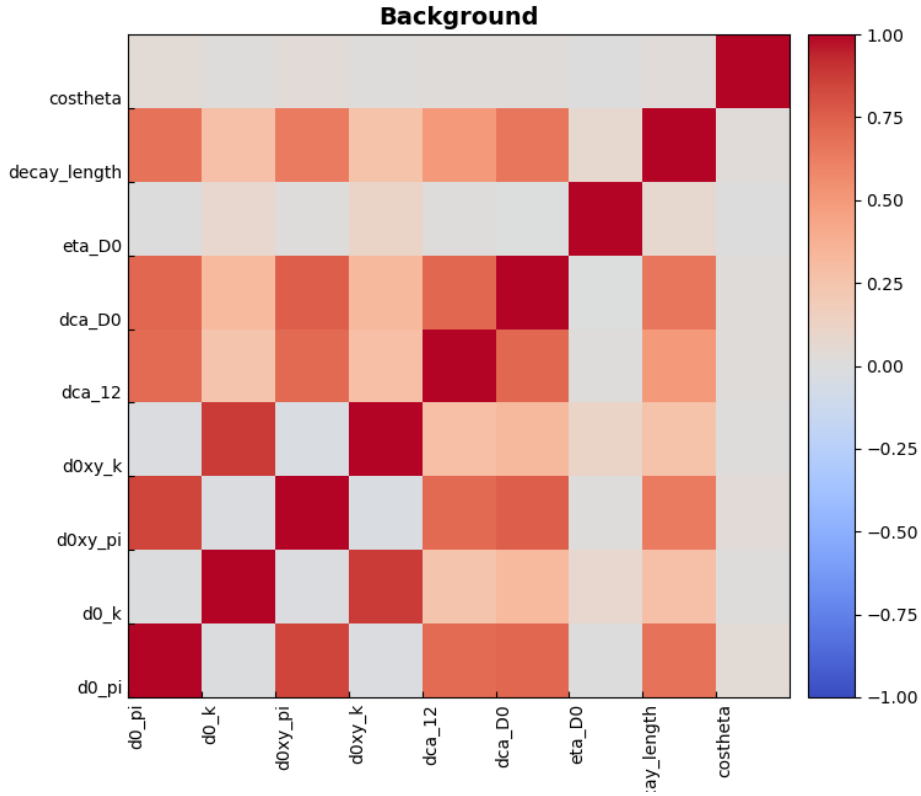
- Machine learning model studies performed for the D^0 reconstruction in ep collisions
- Future Steps:
  - Further optimize to the model to get large ROC score
  - Implement other models e.g. Neural Network (Auto Encoder)
  - Remove perfect particle identification and use particle identification from lookup tables
  - Further make it more differential in p_T and η (under testing)
  - Implement a similar model for Λ_c^+ reconstruction
  - Estimate Λ_c^+/D^0 ratio using machine learning

THANK YOU !!!

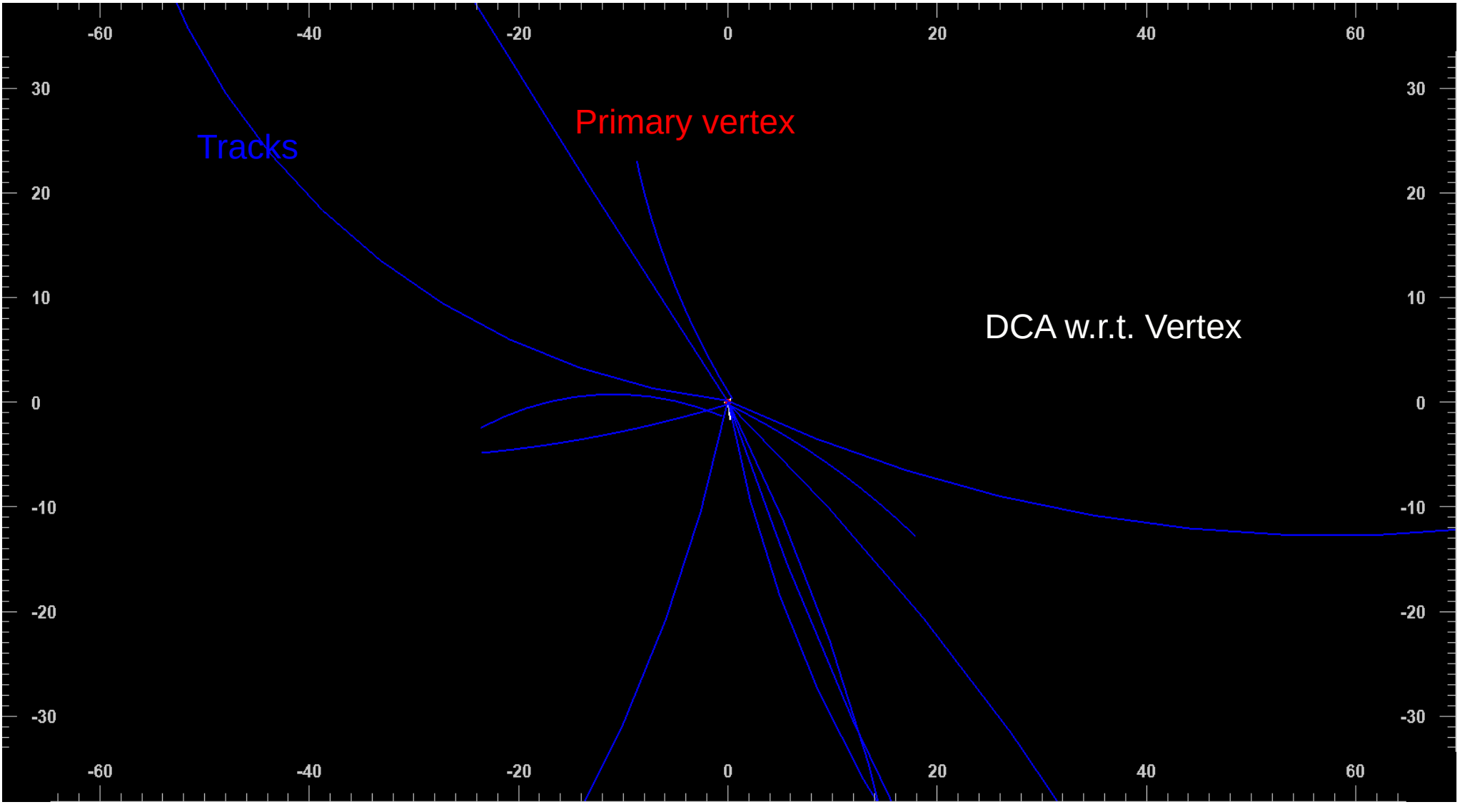
Feature Correlations

d0xy is strongly correlated to d0 so removed while training the machine learning model

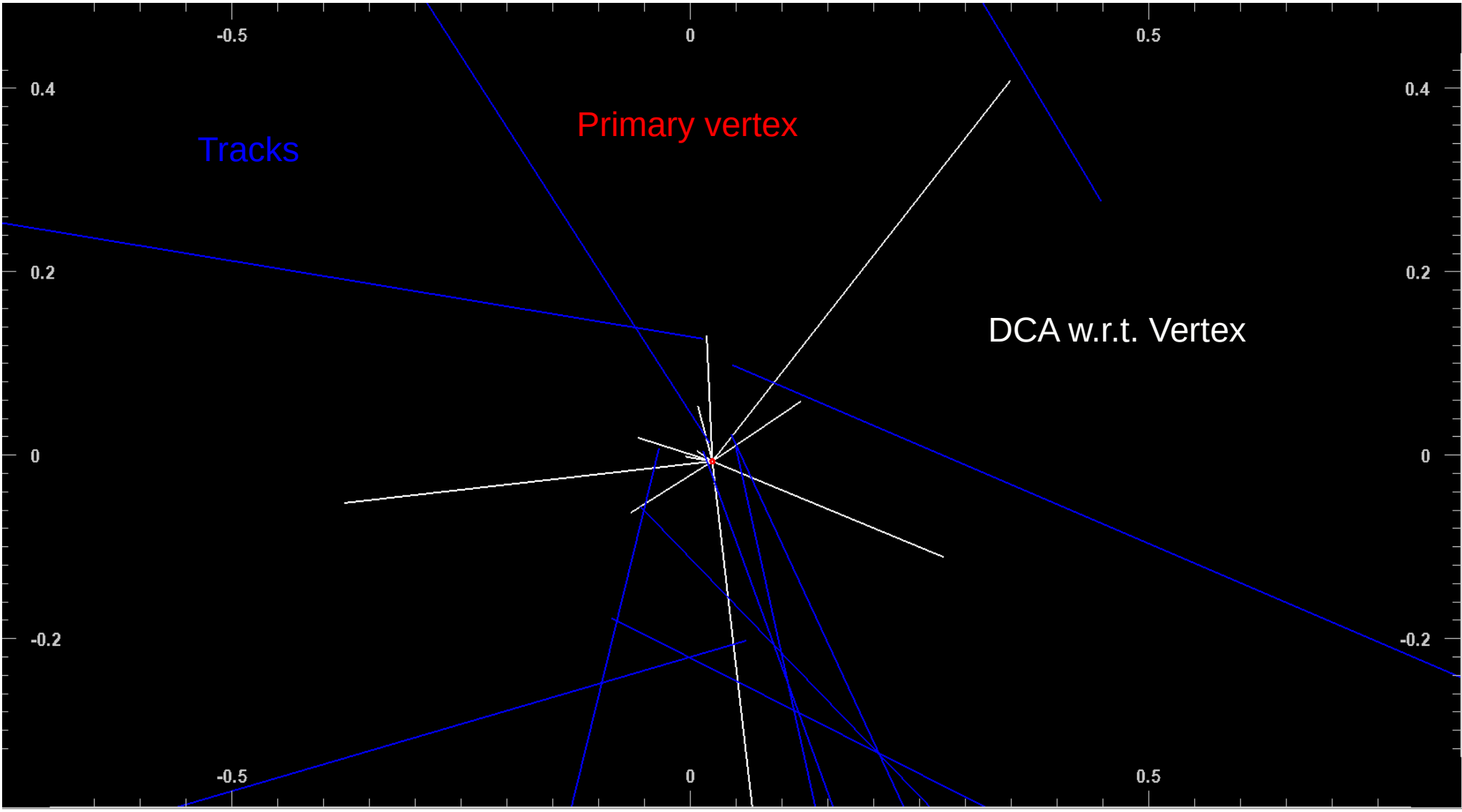




Event display Signal Reco Event (XY Plane)



Event display Signal Reco Event (XY Plane)



Simulation output

Total Events: 493857567

$D^0 \rightarrow K^- \pi^+$ 995867

Event generator	Decay channel	# of D0 events	# of DIS events sampled	Output location
PYTHIA 8 ep, NC, 10x100 $Q^2 > 100$	$D^0 \rightarrow \pi$ + K Decay daughters within |eta| < 3.5	~ 990k	493857567	root://dtn-eic.jlab.org//work/eic2/EPIC/RECO/24.12.0/epic_craterlake/SIDIS/D0_ABCONV/pythia8.306-1.1/10x100/q2_100/hiDiv
PYTHIA 8 ep, NC, 10x100 $Q^2 > 1$	$D^0 \rightarrow \pi$ + K Decay daughters within |eta| < 3.5	~ 990k	1747000357	root://dtn-eic.jlab.org//work/eic2/EPIC/RECO/24.12.0/epic_craterlake/SIDIS/D0_ABCONV/pythia8.306-1.1/10x100/q2_1/hiDiv