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

Shyam Kumar*, Annalisa Mastroserio, Domenico Elia

INFN Bari, Italy

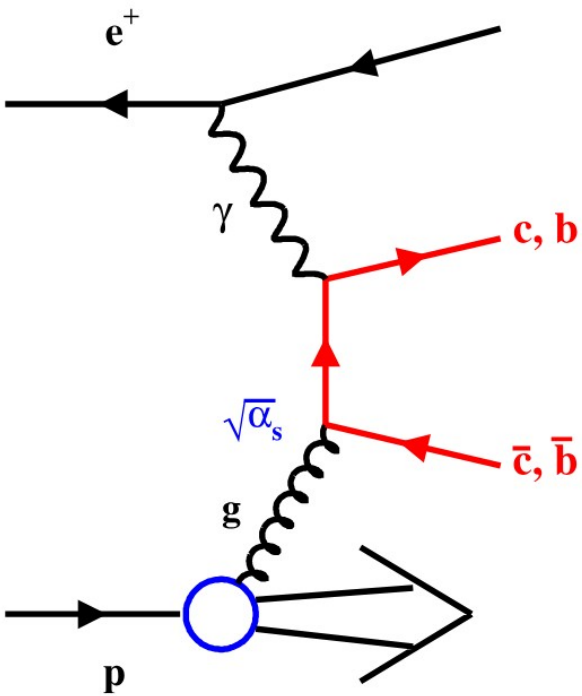
hipe4ml package

https://doi.org/10.5281/zenodo.5070131

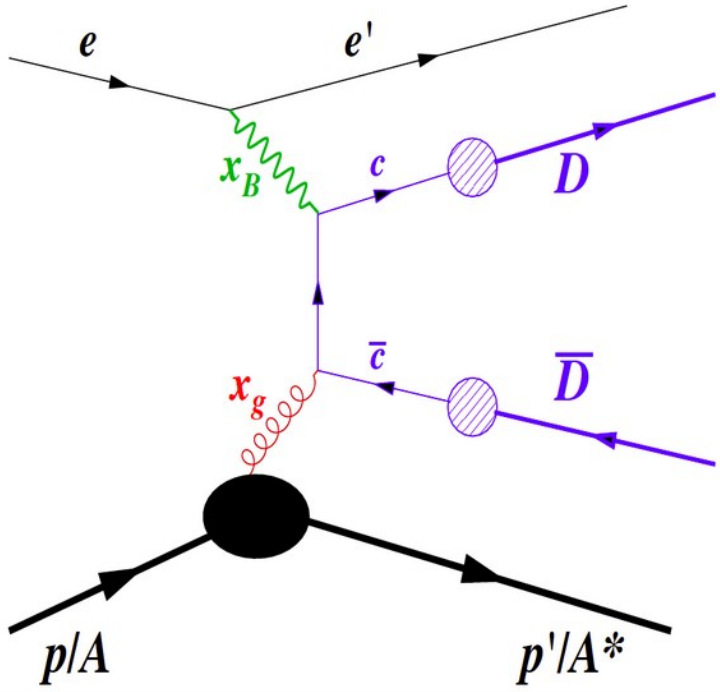
Heavy-flavor Production

Boson-Gluon Fusion (BGF) is dominant mechanism [LO]

https://doi.org/10.1016/j.ppnp.2015.06.002



$$\gamma^* g \rightarrow c\bar{c} \text{ or } b\bar{b}$$

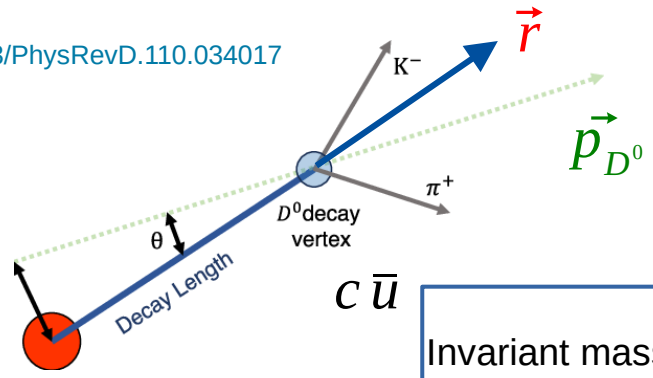


Virtual photon (γ*) from the electron interacts with a gluon from the proton, produces $c\bar{c}$ or $b\bar{b}$ pair

Additional NLO Mechanisms: Gluon splitting, QCD Compton Scattering

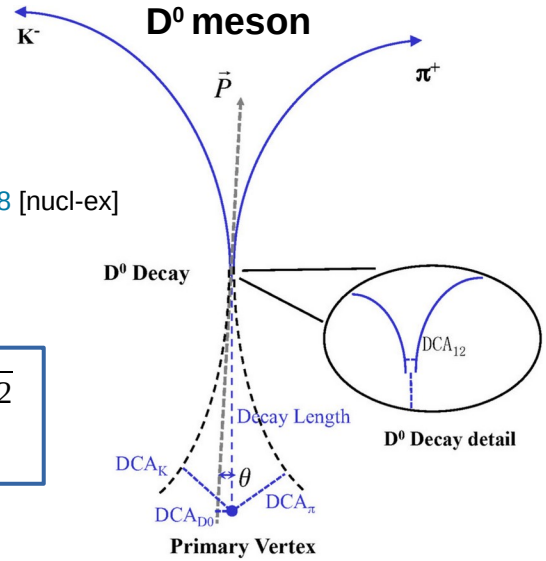
Topological Variables

10.1103/PhysRevD.110.034017



cτ = 123 μm

arXiv:1911.12168 [nucl-ex]

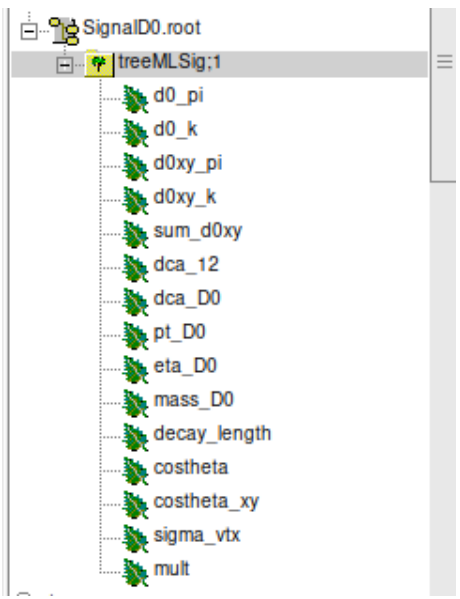


Invariant mass: $m_{D^0} = \sqrt{\left(E_{K^-} + E_{\pi^+}\right)^2 - \left(\vec{p}_{K^-} + \vec{p}_{\pi^+}\right)^2}$

Topological Variables:

- DCA_{k^-} and DCA_{π^+} with respect to the reconstructed primary vertex (d0_k, d0_pi)
- 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
- Multiplicity (mult)

Realistic PID
Differential in p_T and y

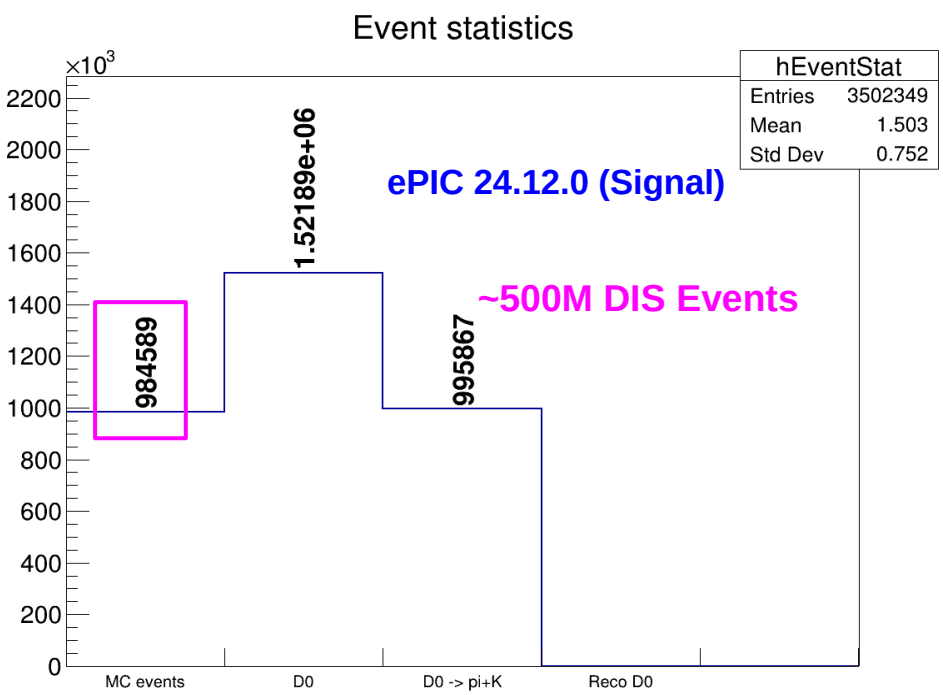


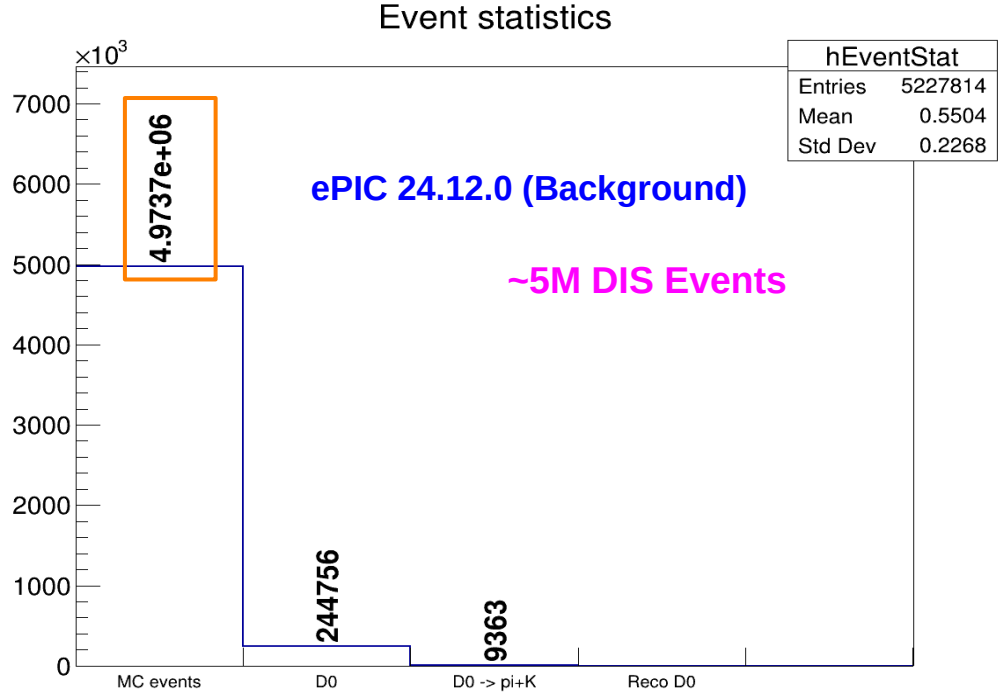
Data Sample for ML (Q^2 = 100 GeV^2)

- **ML Algorithm: BDT (Boosted Decision Tree) Binary Classifier**

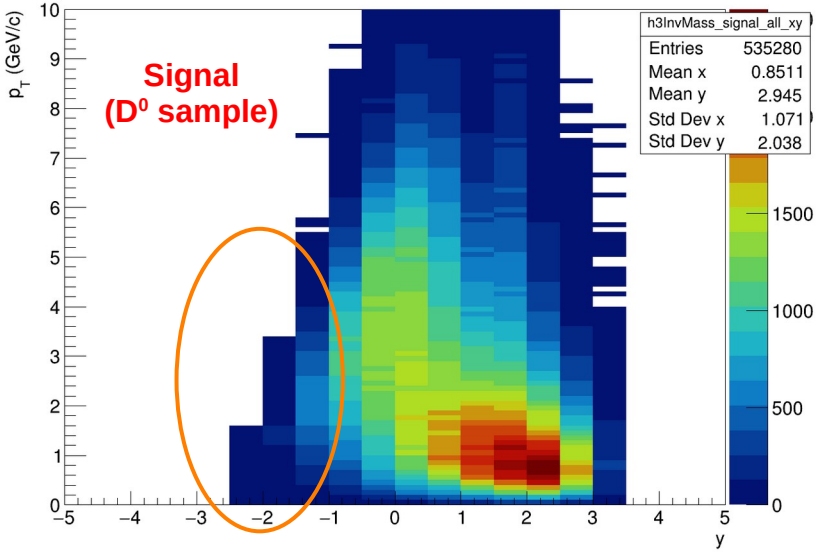
Simulation of D0 and Lc samples

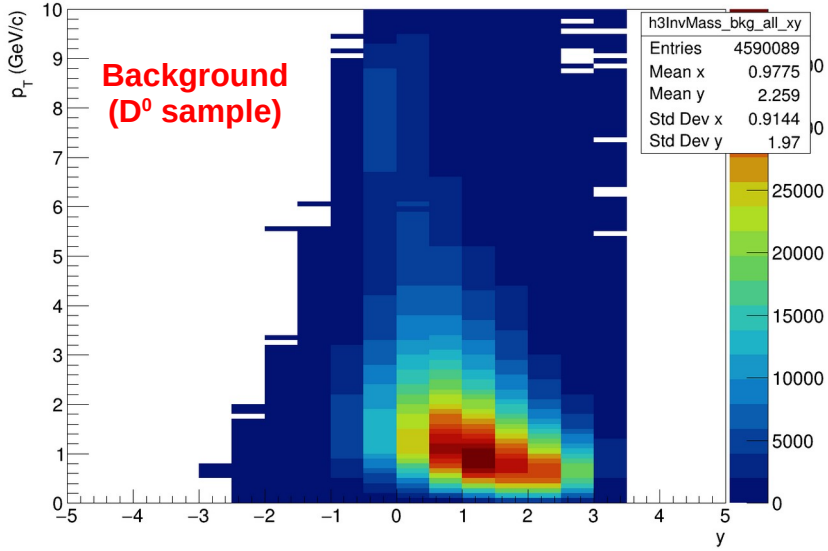
- D^0 enriched same created filtering **PYTHIA8 ep, NC, 10X100, Q^2 >100 GeV^2 events (~493M)** such that each event consist one $D^0 \rightarrow k^-\pi^+$ known as Signal taken from 24.12.0/epic_craterlake/SIDIS/D0_ABCONV/pythia8.306-1.1/10x100/q2_100): **Total files 1869 and Events = 984589**
- Background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



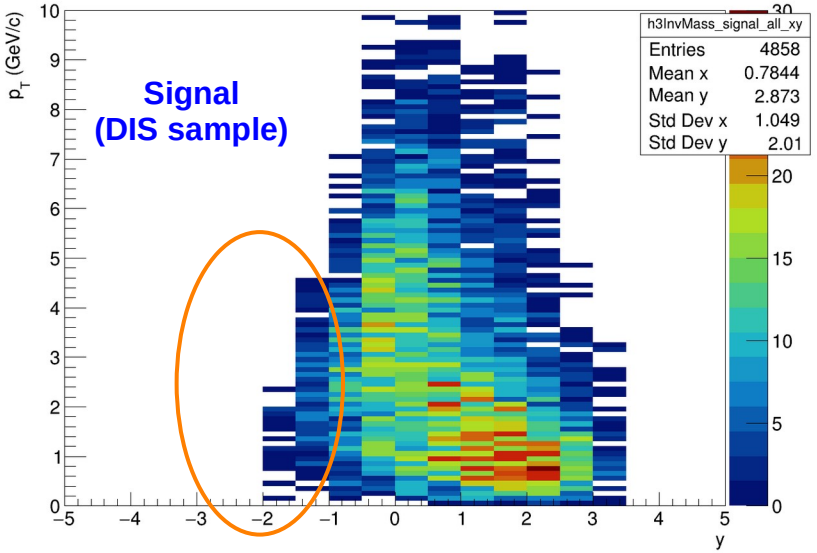


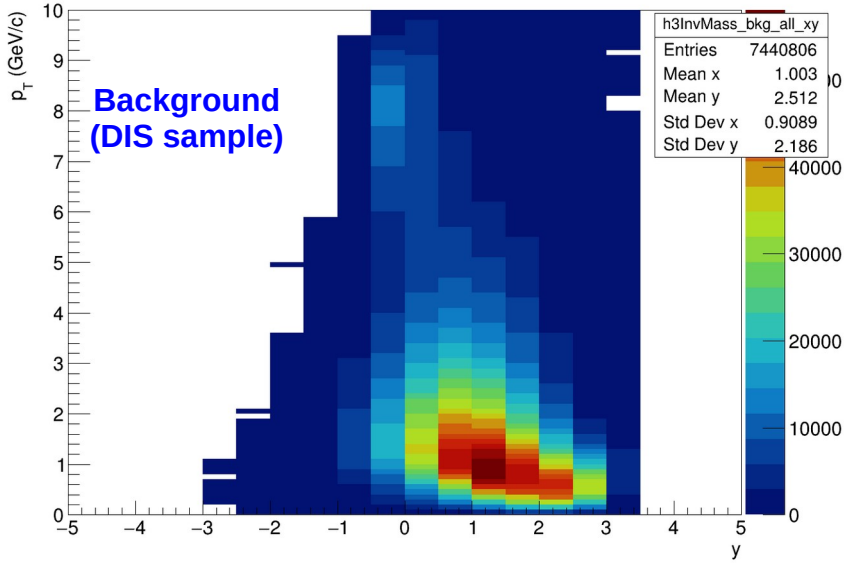
Phase Space (Q^2 = 100 GeV^2)



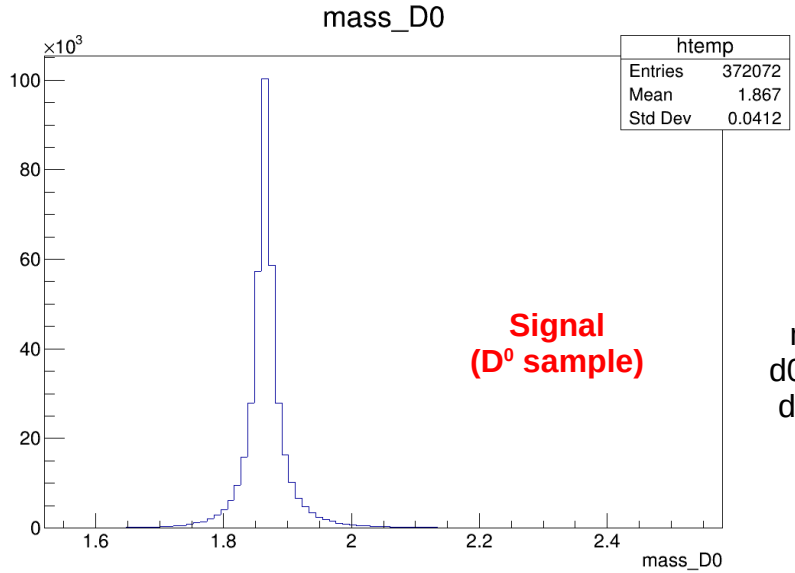


-3.0 <y < -1.0
Lowest stats



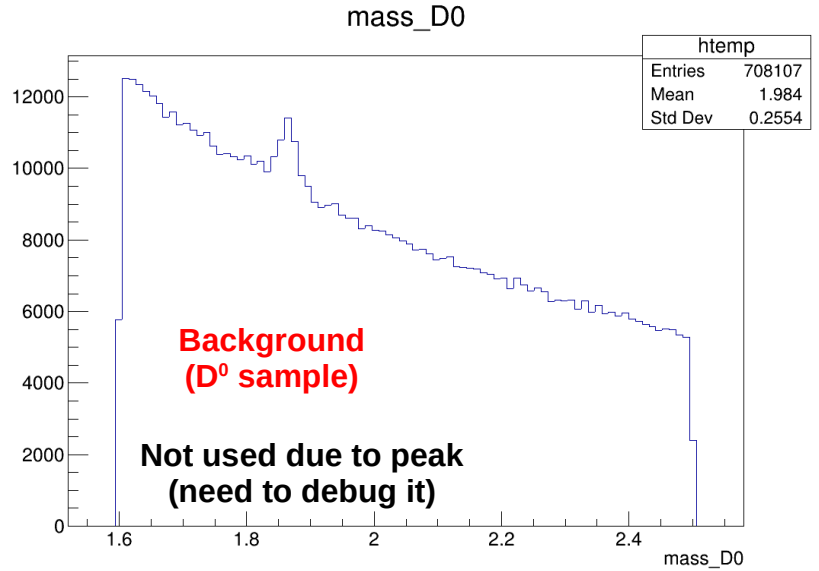


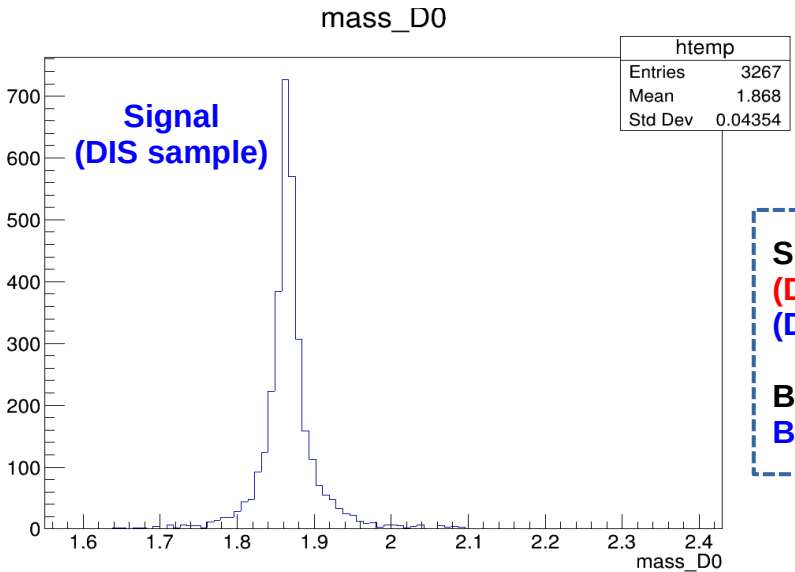
Sample After Preselection (Q^2 = 100 GeV^2)



Preselection

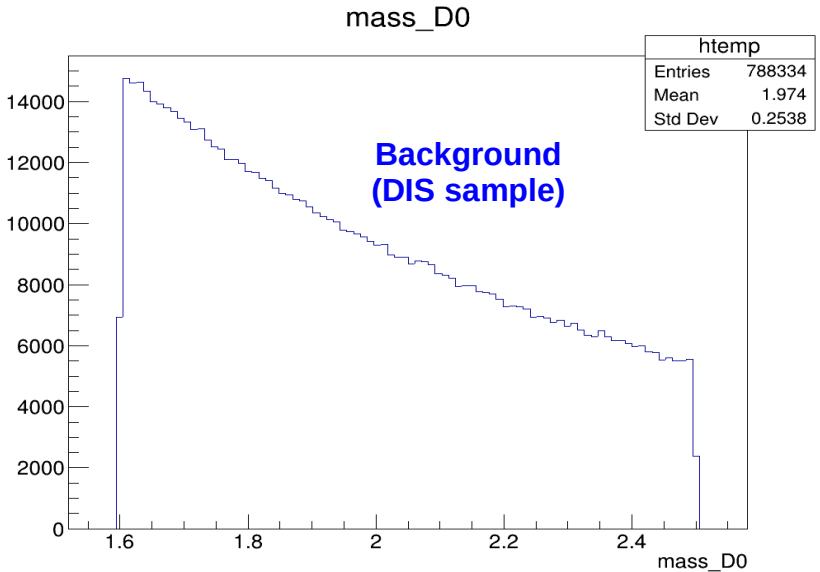
mD0 > 1.6 && mD0 < 2.5
d0xypi>0.02 && d0xypi<10.
d0xyk>0.02 && d0xyk<10.
decay length <100.





Signal for ML: Signal (D^0 sample) +Signal (DIS)

Background for ML: Background (DIS)



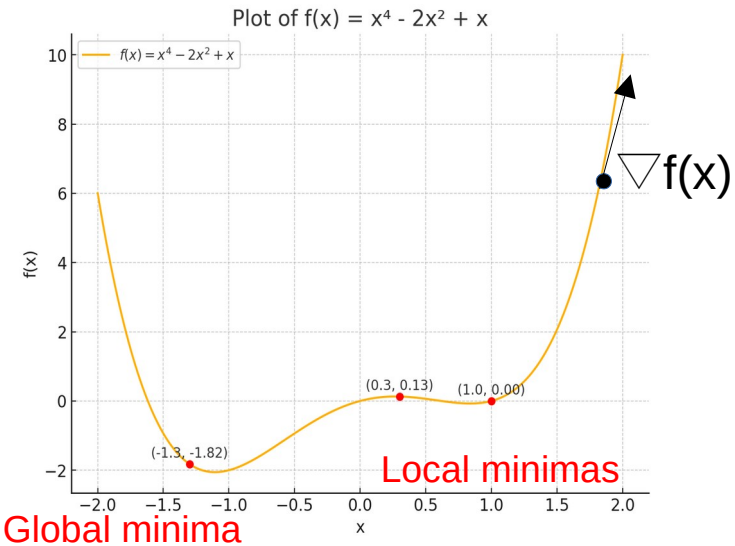
Number of Signal and Background

y(D0)	p_T(D0)	Signal	Background
-1.0 to 1.0	1.0-2.0	32624	32624
-1.0 to 1.0	2.0-5.0	22937	22937
-1.0 to 1.0	5.0-10.0	2581	2581
1.0 to 3.0	1.0-2.0	61791	61791
1.0 to 3.0	2.0-5.0	53348	53348
1.0 to 3.0	5.0-10.0	2956	2956
-3.0 to -1.0	1.0-2.0	682	682
-3.0 to -1.0	2.0-5.0	415	415

Keep the number of signal and background same for ML

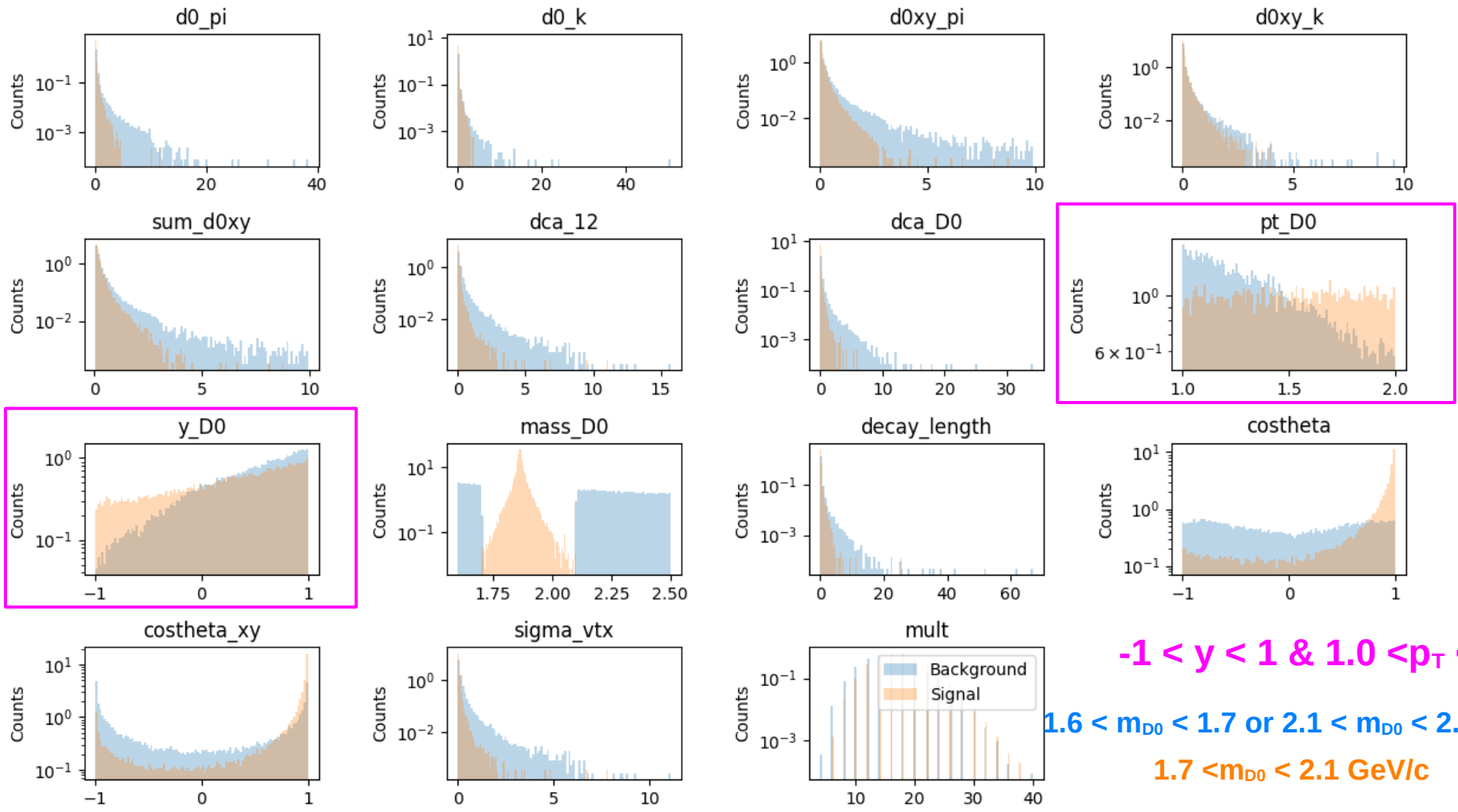
Two approaches for parameter optimization: **Optuna and Bayesian (optimization frameworks) and comparison (backup)**

Bayesian takes more computing time but catches the global minima during optimization if multiple local minima exists



Signal and Background distributions

preselection ="(mD0 > 1.6 && mD0 < 2.5) && (d0xypi>0.02 && d0xypi<10.) && (d0xyk>0.02 && d0xyk<10.) && decay length <100.";



$-1 < y < 1$ & $1.0 < p_T < 2.0$

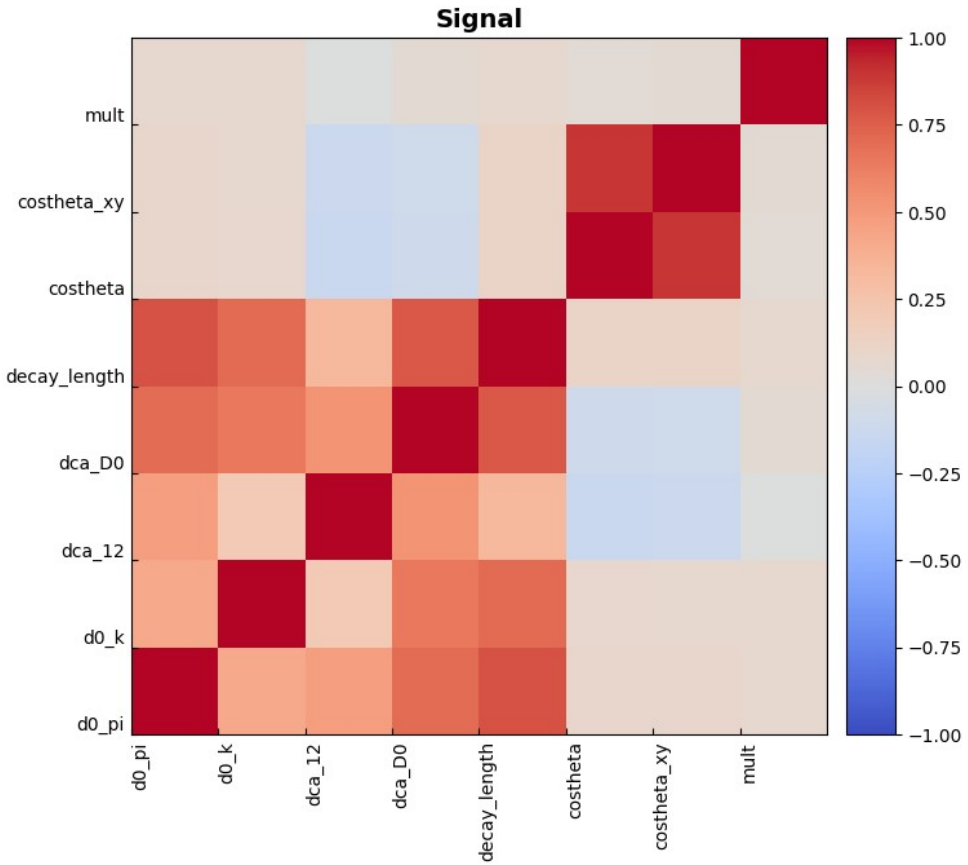
$1.6 < m_{D0} < 1.7$ or $2.1 < m_{D0} < 2.5$ GeV/c

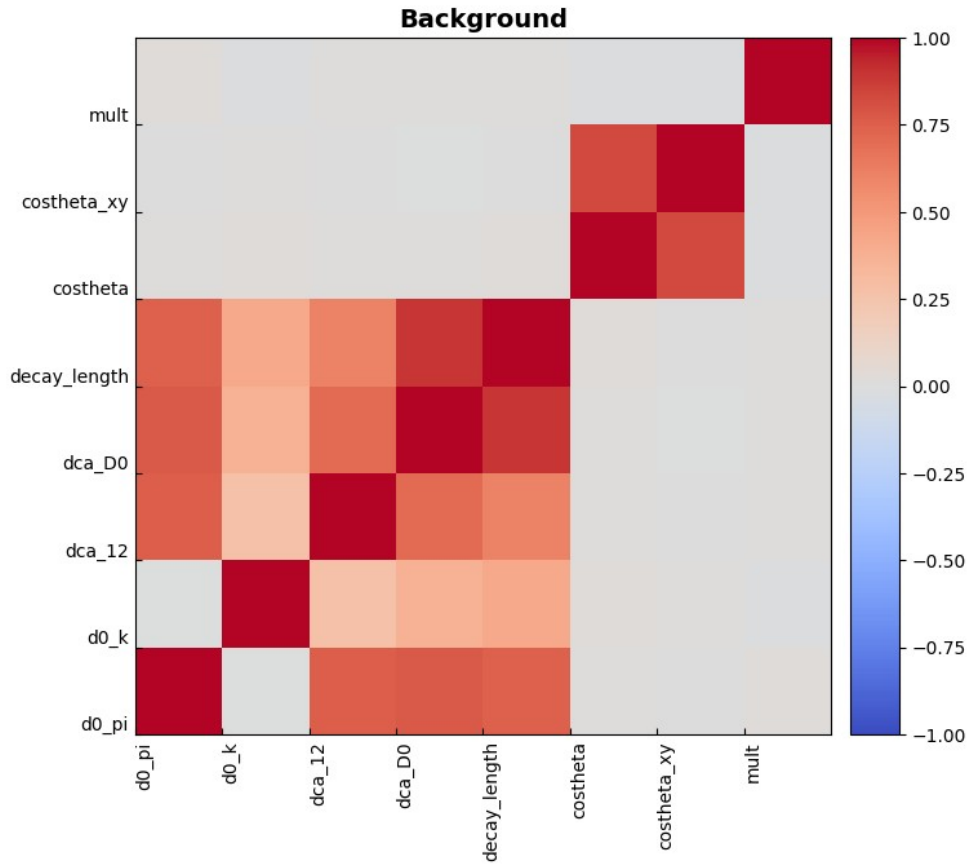
$1.7 < m_{D0} < 2.1$ GeV/c

Correlation Matrix

-1.0 < y < 1.0
1.0 < p_T < 2.0 GeV/c

Planning to remove costheta_xy and decay_length once other cuts are available (e.g. chi2, nsigma, etc.)

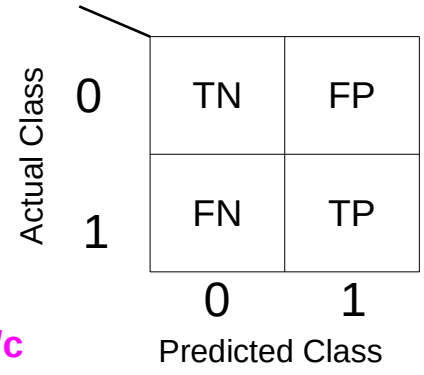




Performed a cross-check removing costheta_xy and decay_length (see backup)

Confusion Matrix

Confusion matrix



Predicted: Positive / Negative

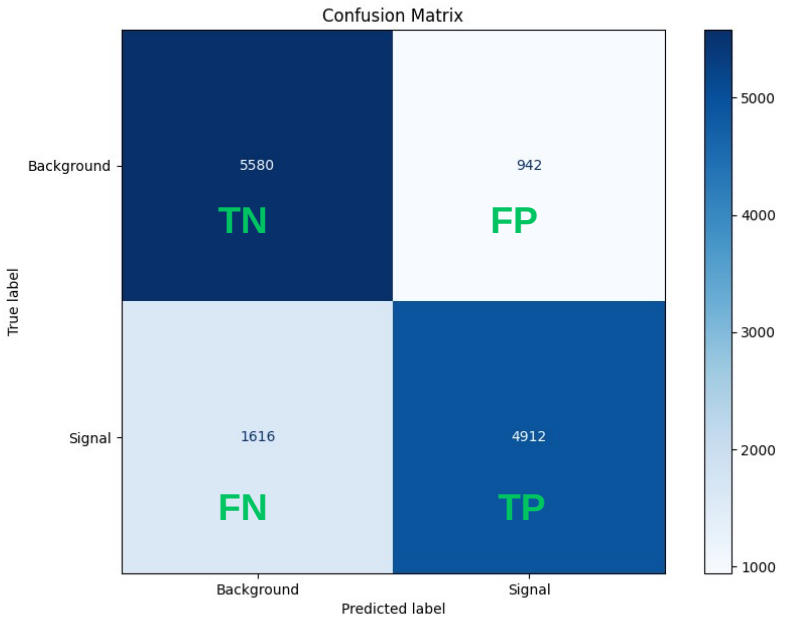
$$TPR = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

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

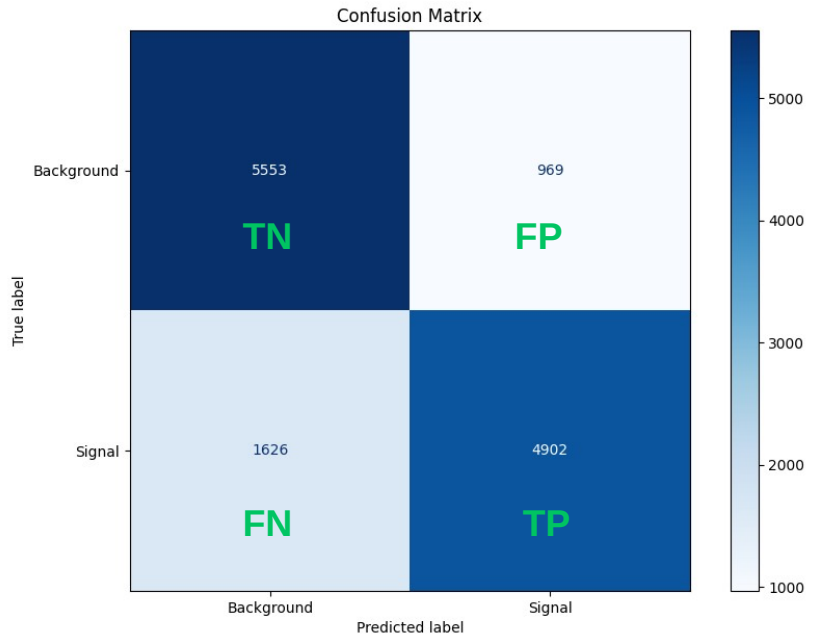
-1.0 < y < 1.0
1.0 < p_T < 2.0 GeV/c

Test sample = 0.2*65248 ~ 13050 = 5553+4902+1626+969

Test Sample (Optuna)



Test Sample (Bayesian)

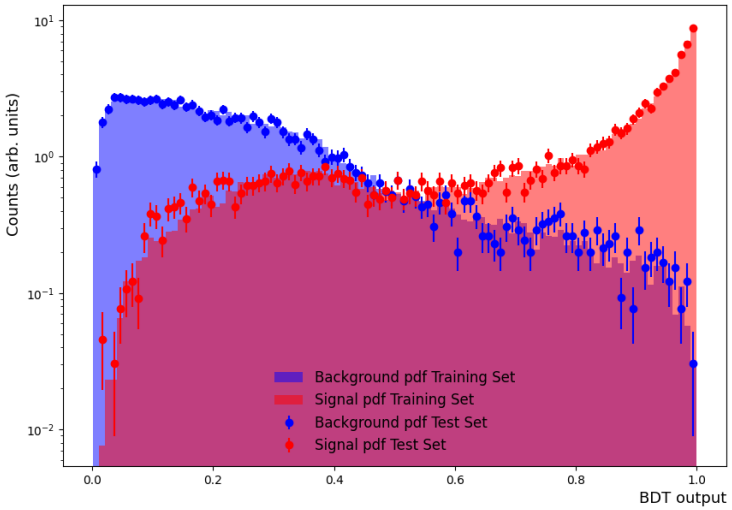


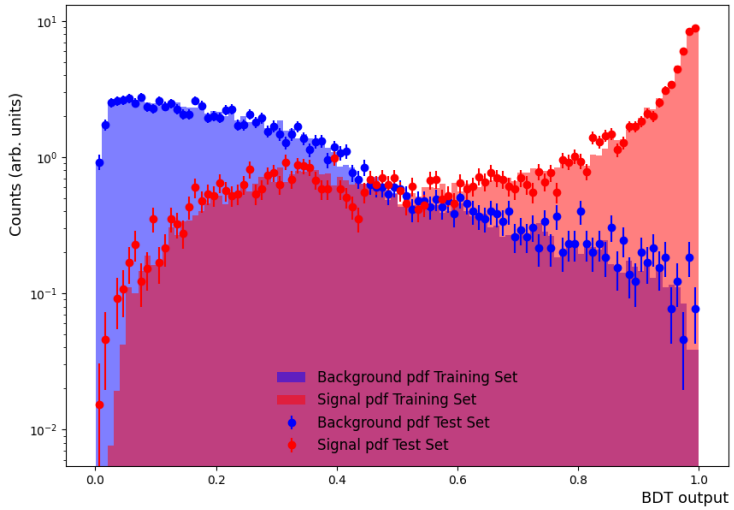
BDT Output

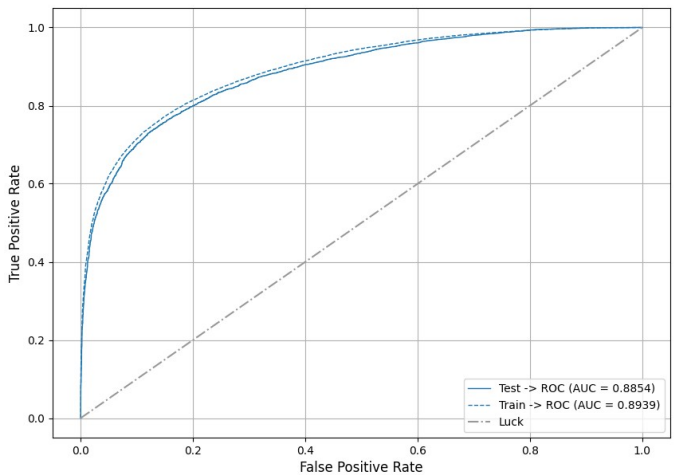
Optuna

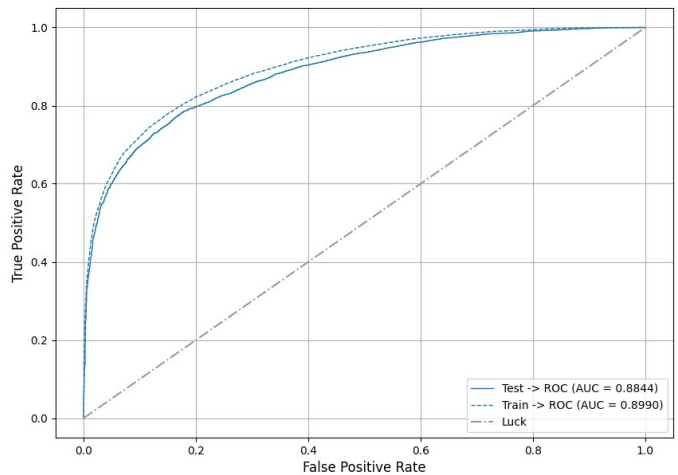
-1.0 < y < 1.0
1.0 < p_T < 2.0 GeV/c

Bayesian







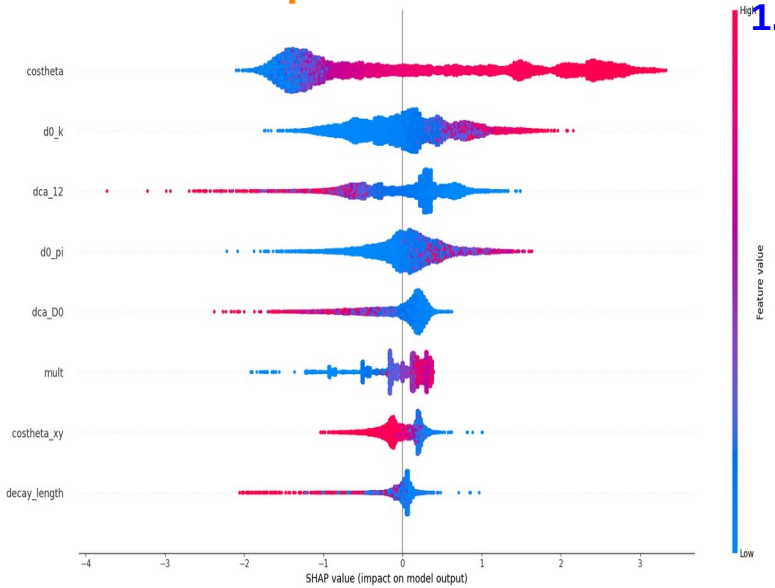


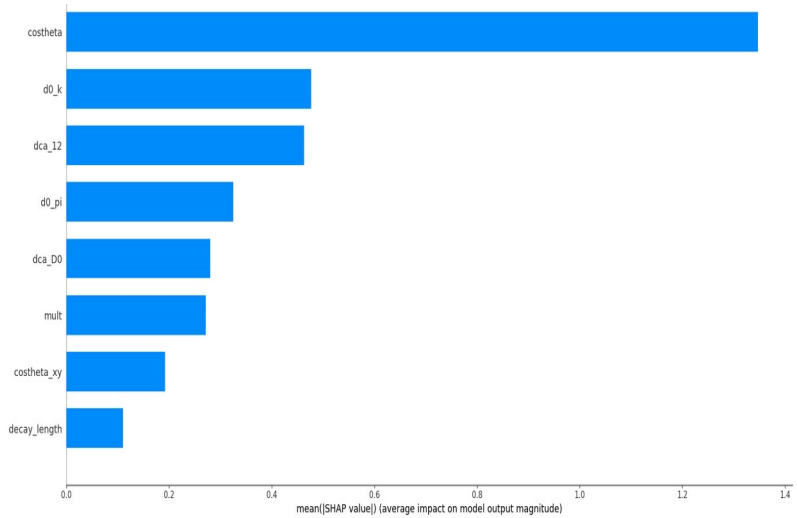
Results

Optuna

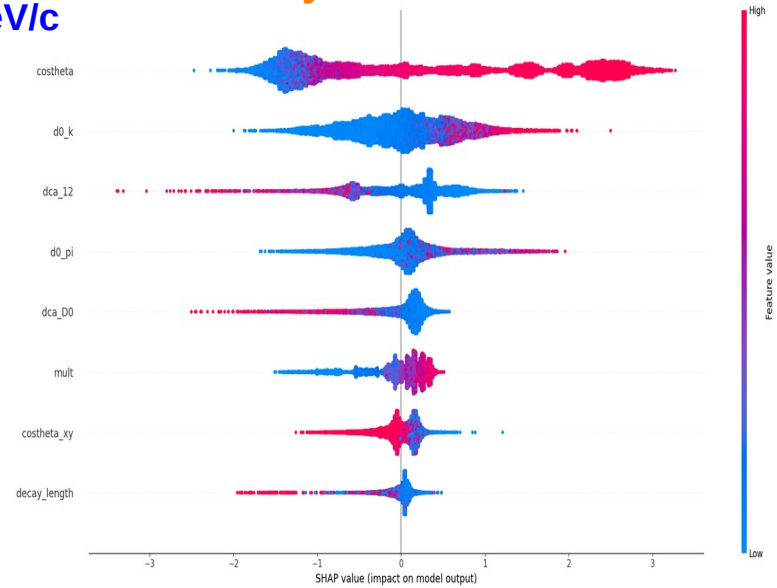
-1.0 < y < 1.0

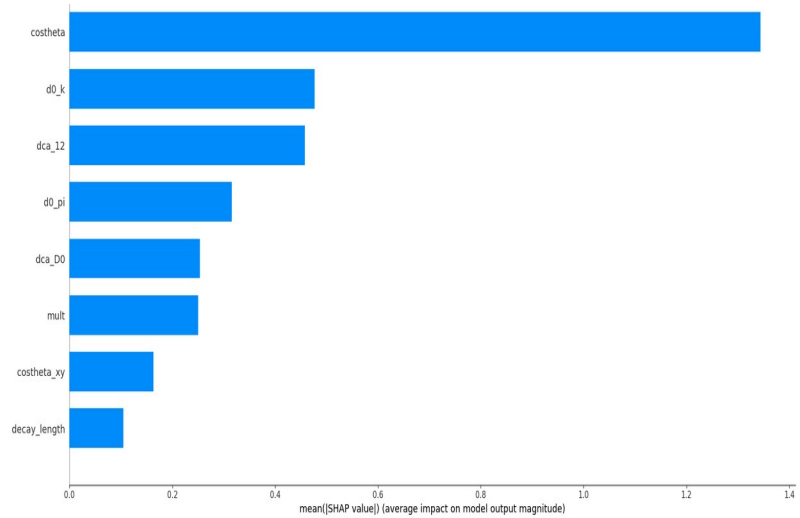
1.0 < p_T < 2.0 GeV/c





Bayesian

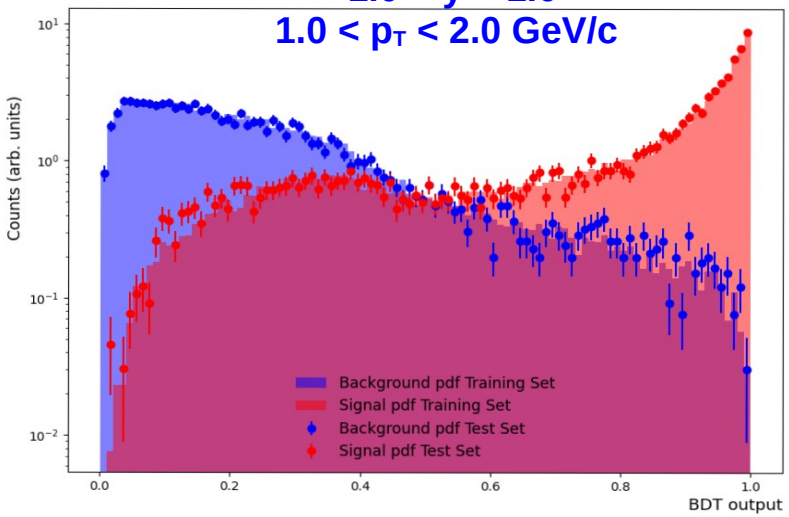


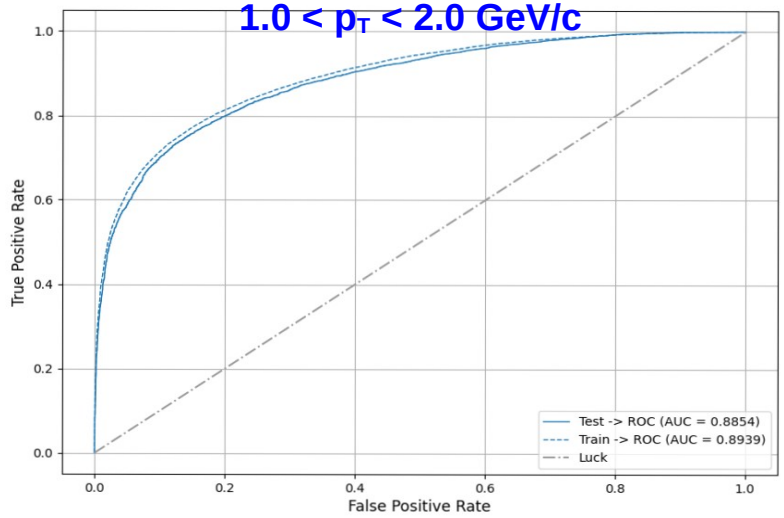


BDT Output

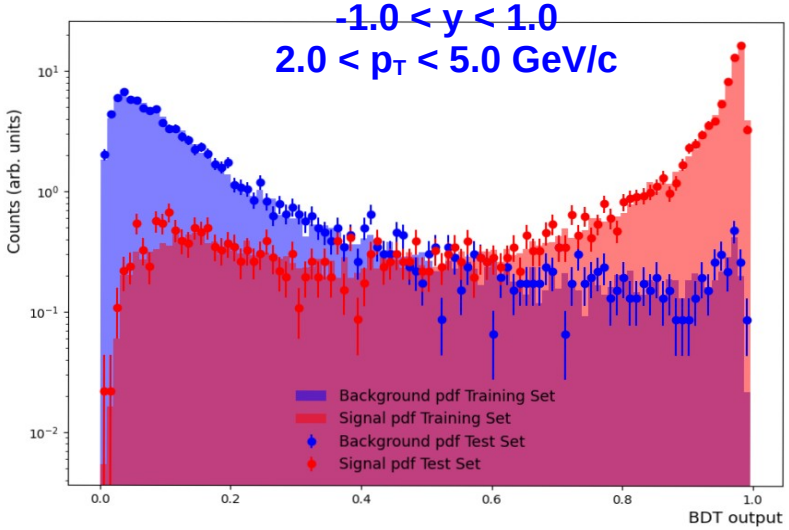
Study with Optuna

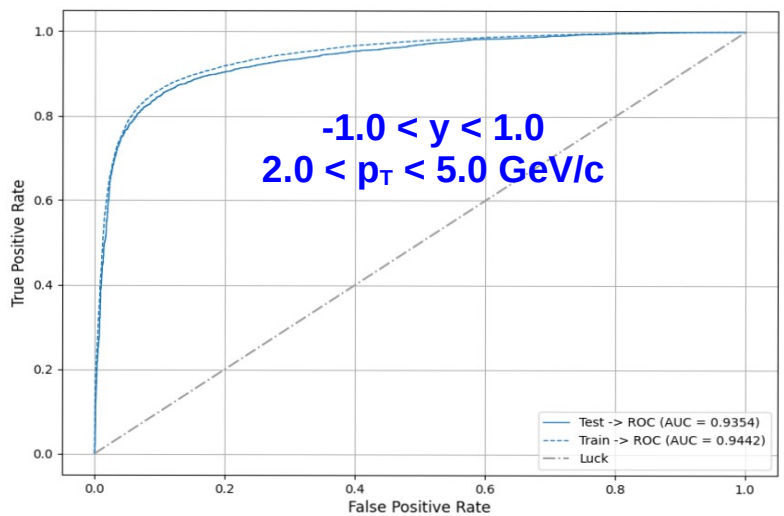
-1.0 < y < 1.0
1.0 < p_T < 2.0 GeV/c





-1.0 < y < 1.0
2.0 < p_T < 5.0 GeV/c

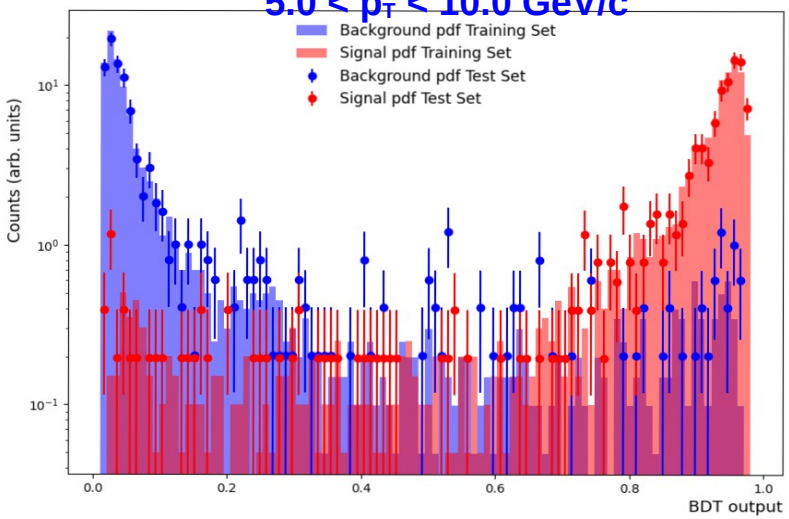


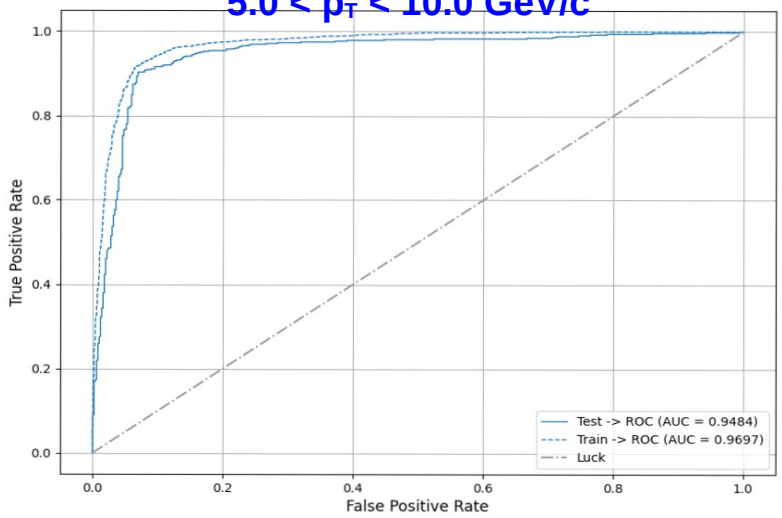


BDT Output

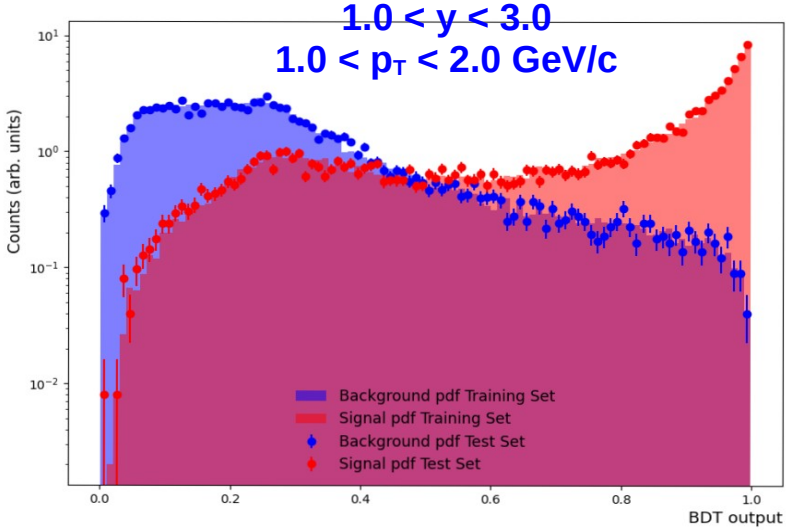
Study with Optuna

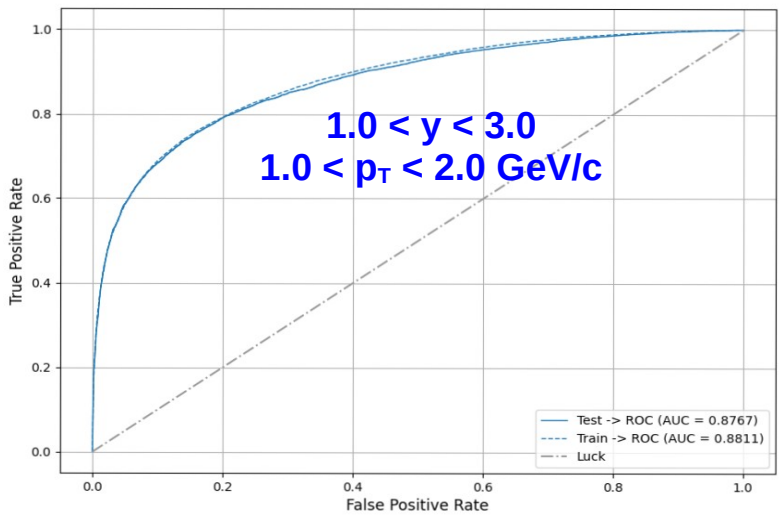
-1.0 < y < 1.0
5.0 < p_T < 10.0 GeV/c





1.0 < y < 3.0
1.0 < p_T < 2.0 GeV/c

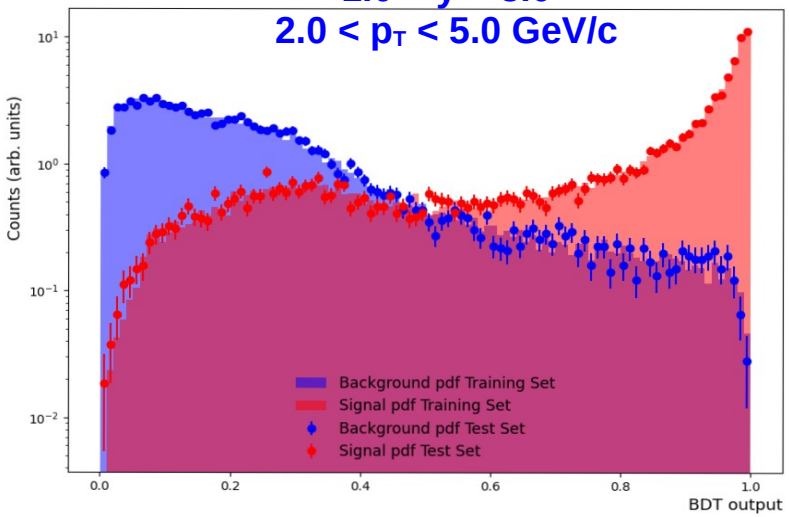


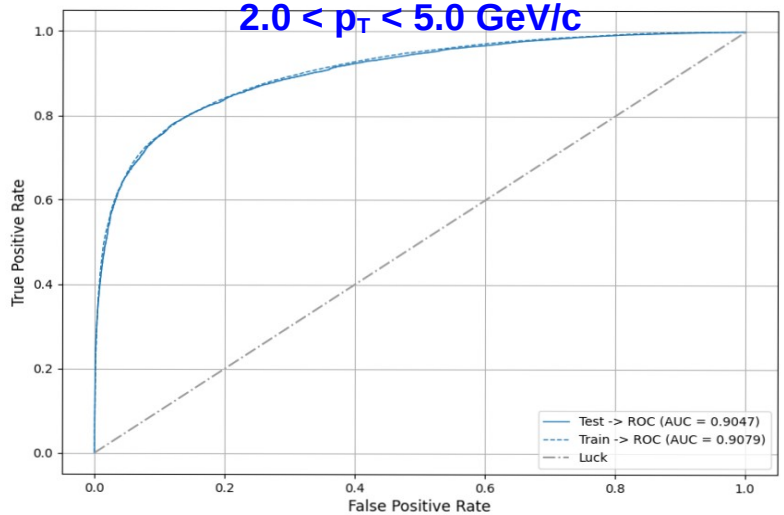


BDT Output

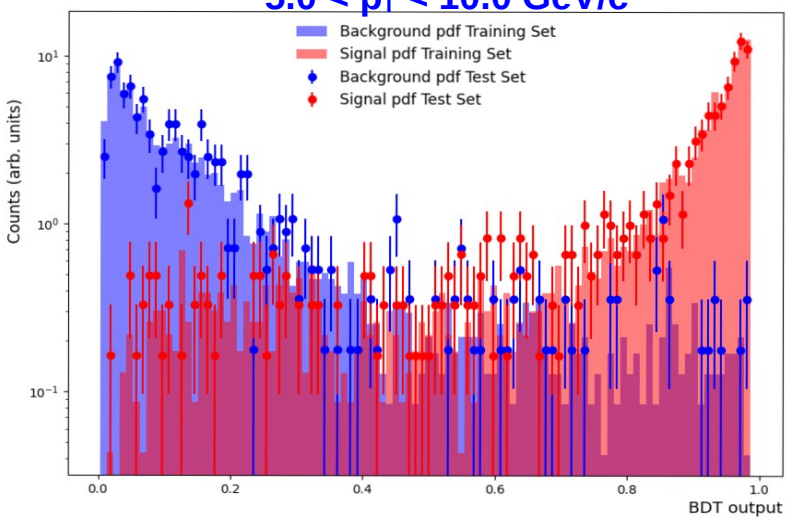
Study with Optuna

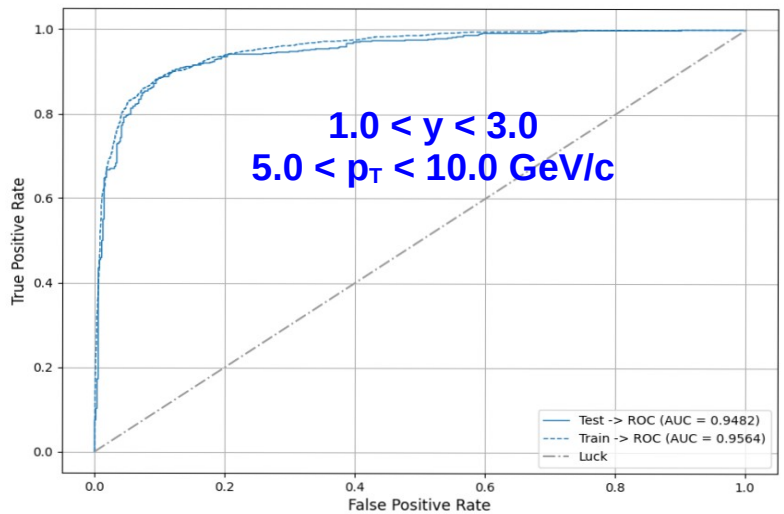
1.0 < y < 3.0
2.0 < p_T < 5.0 GeV/c





1.0 < y < 3.0
5.0 < p_T < 10.0 GeV/c

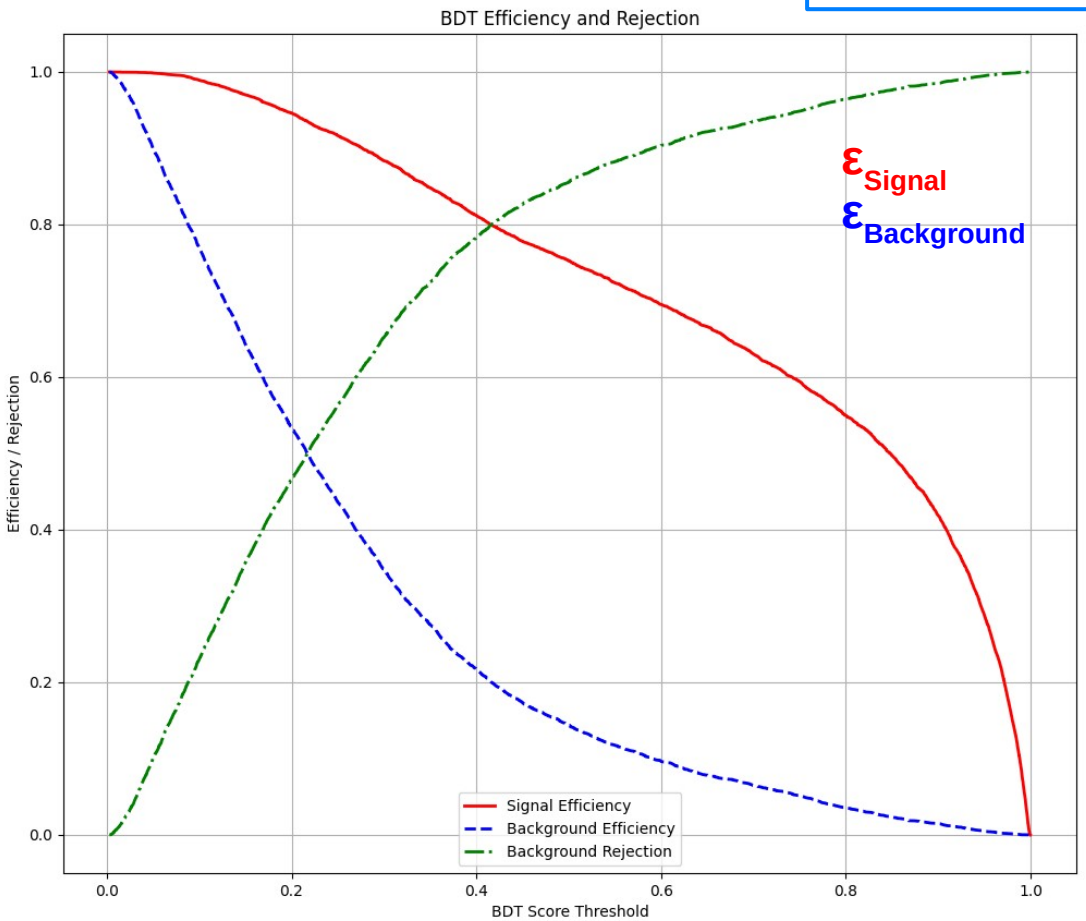




BDT Efficiencies

-1.0 < y < 1.0
1.0 < p_T < 2.0 GeV/c

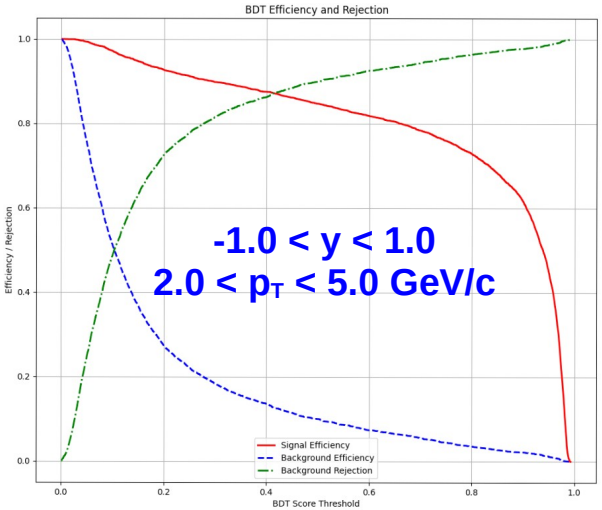
$$\text{Significance}_{\text{BDT Cut}} = \frac{S_{\text{No ML}} \times \epsilon_{\text{Signal}}}{\sqrt{S_{\text{No ML}} \times \epsilon_{\text{Signal}} + B_{\text{No ML}} \times \epsilon_{\text{Background}}}}$$

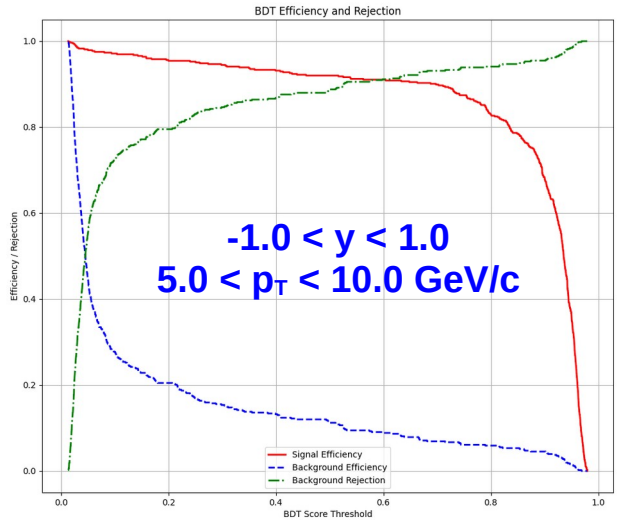


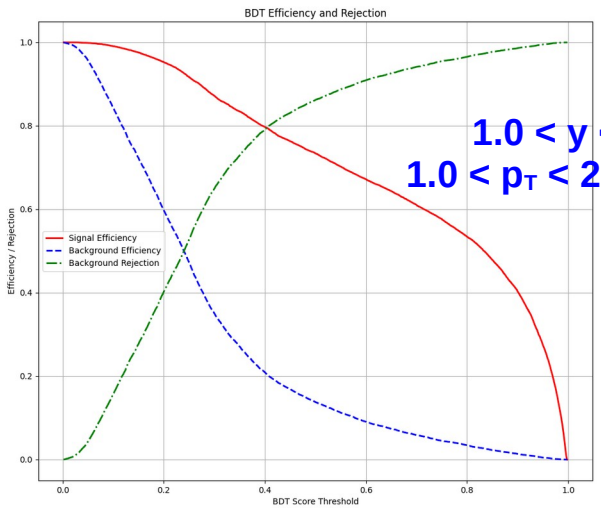
$$\left(\frac{S}{B}\right)_{\text{BDT Cut}} = \left(\frac{S}{B}\right)_{\text{No ML}} \times \frac{\epsilon_{\text{Signal}}}{\epsilon_{\text{Background}}}$$

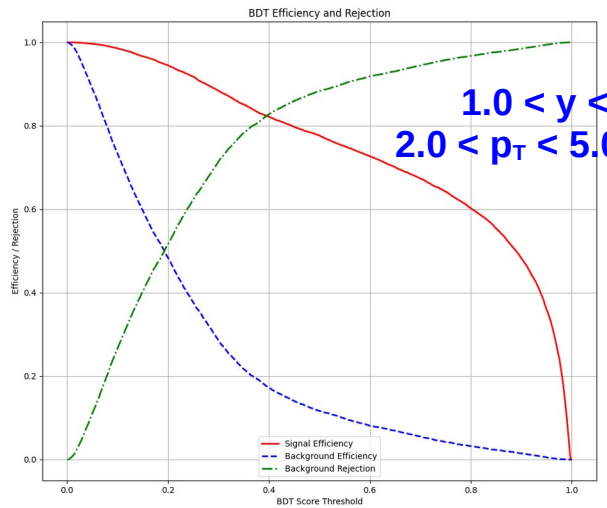
This allows to estimate the S/B and Significance at a given BDT Cut

BDT Efficiencies



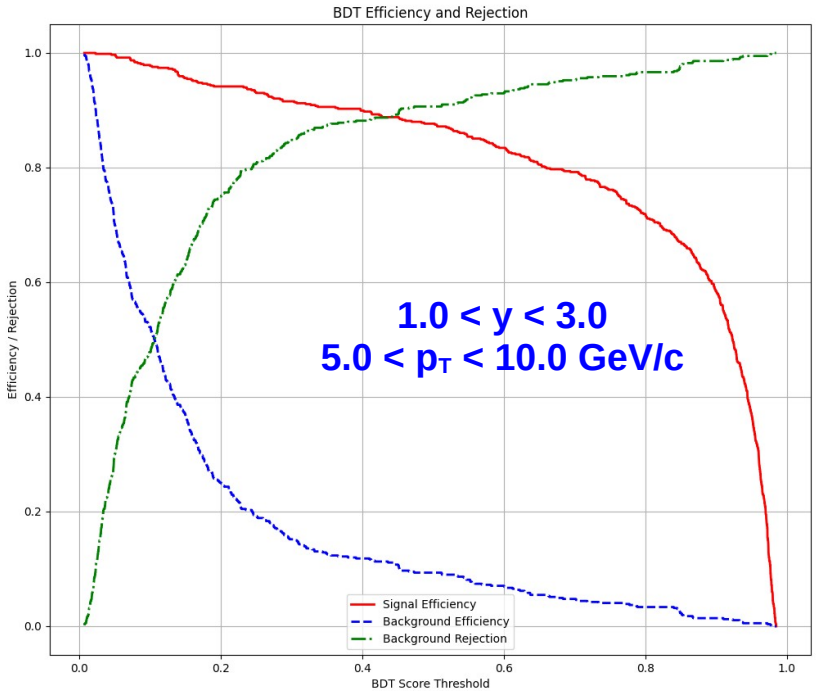


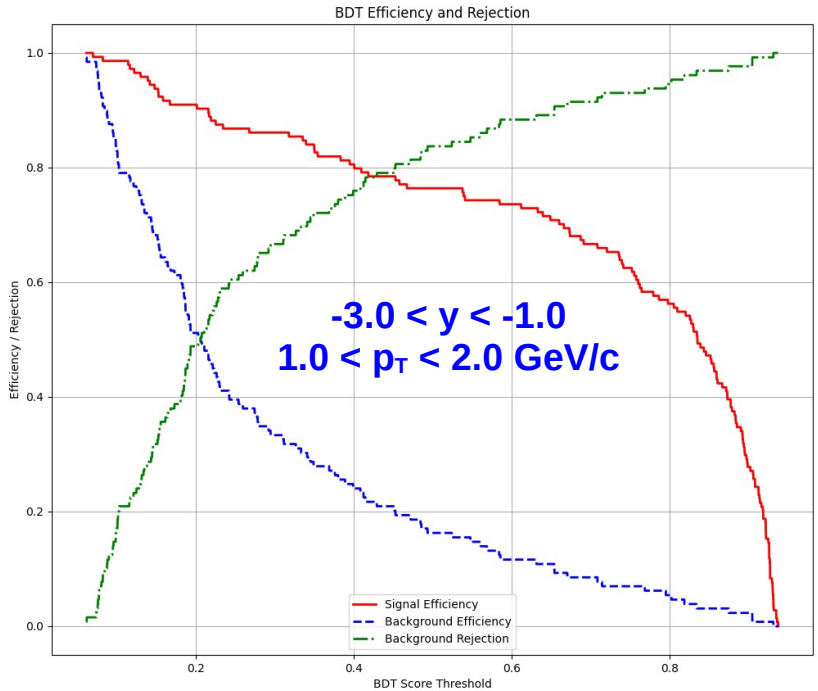




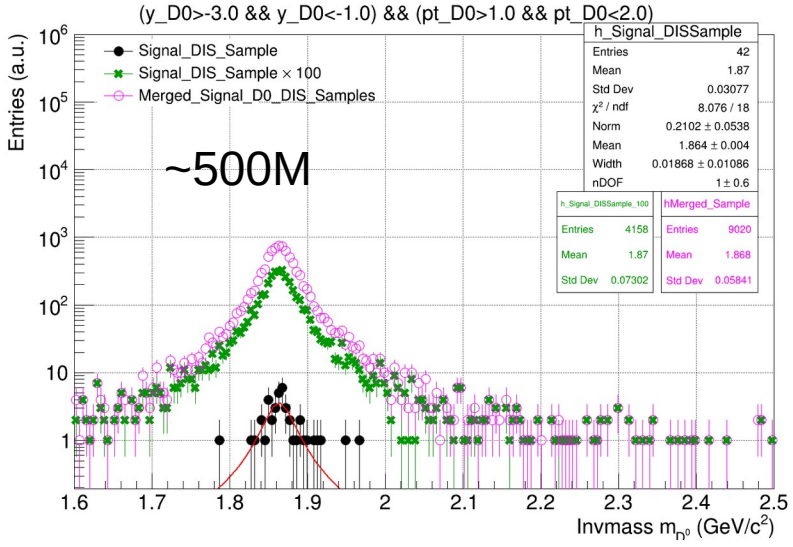
BDT Efficiencies

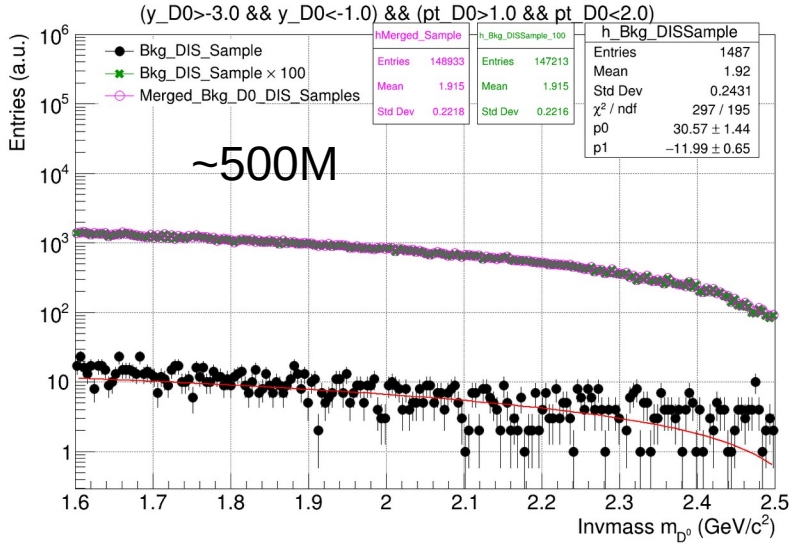
Large fluctuations (low stats)

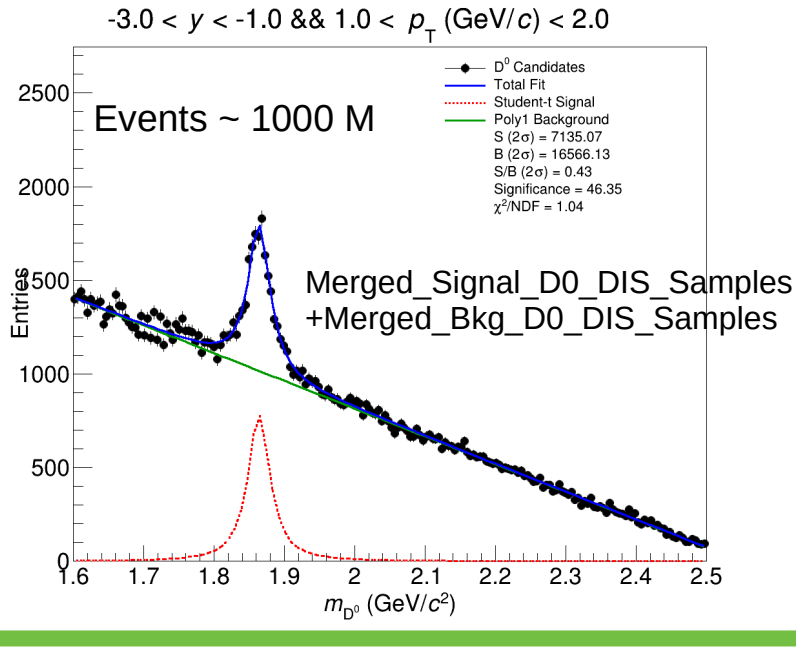




Merging Signal and Background (D0 Sample and DIS Events)







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

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

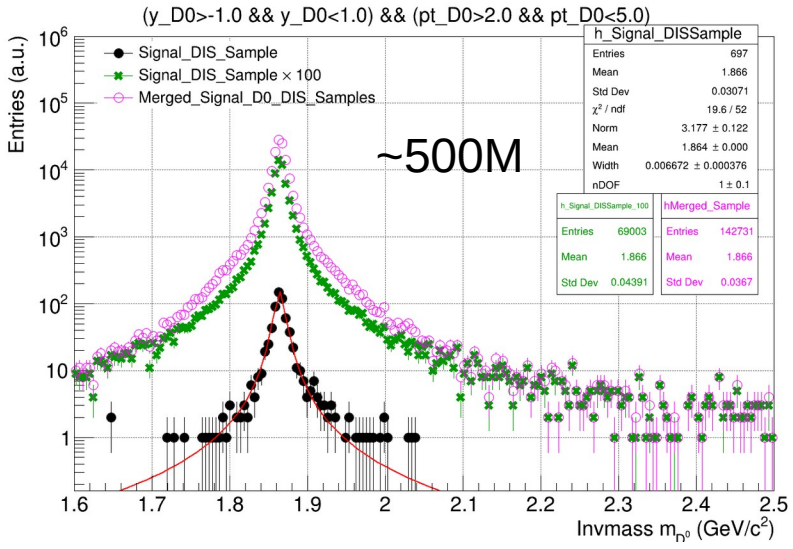
$$\left(\frac{S}{B}\right)_{\text{BDT Cut} = 0.4} \sim 0.43 * 0.8 / 0.25 = 1.37$$

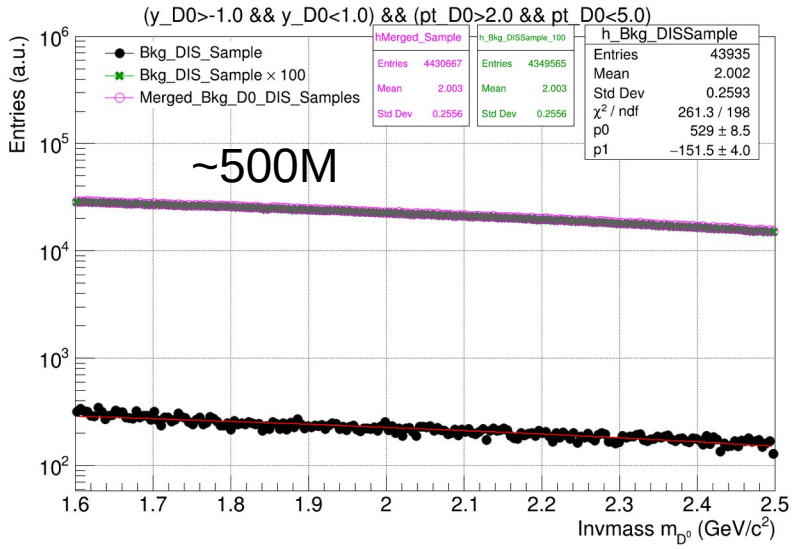
$$Significance_{\text{BDT Cut} = 0.4} \sim \frac{7135.07 * 0.8}{\sqrt{7135.07 * 0.8 + 16566.13 * 0.25}} = 57.51$$

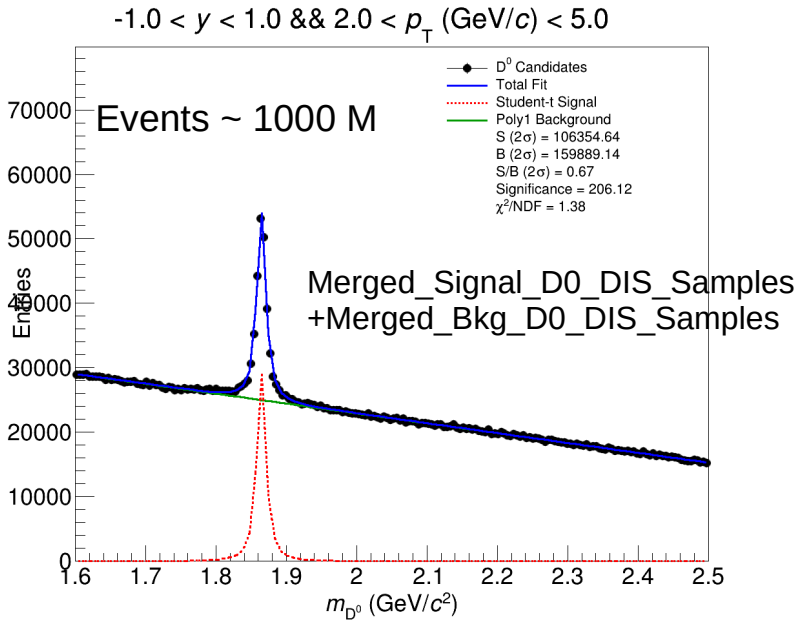
$$Expected\ Significance_{\text{BDT Cut} = 0.4} \sim 57.51 * \sqrt{(6.5/1000)} \sim 5$$

Need to scan to find Best BDT Cut

Merging Signal and Background (D0 Sample and DIS Events)







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

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

$$\left(\frac{S}{B}\right)_{\text{BDT Cut = 0.4}} \sim 0.67 * 0.85 / 0.15 = 3.80$$

$$Significance_{\text{BDT Cut = 0.4}} \sim \frac{106354.64 * 0.85}{\sqrt{106354.64 * 0.85 + 159889.14 * 0.15}} = 267.30$$

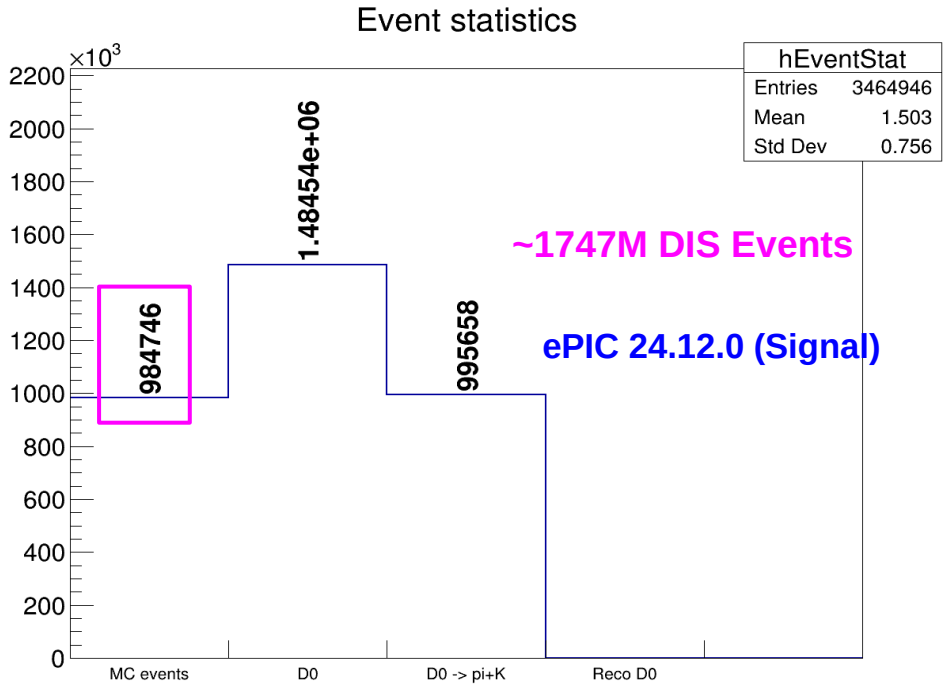
$$Expected\ Significance_{\text{BDT Cut = 0.4}} \sim 267.30 * \sqrt{(6.5/1000)} \sim 21.5$$

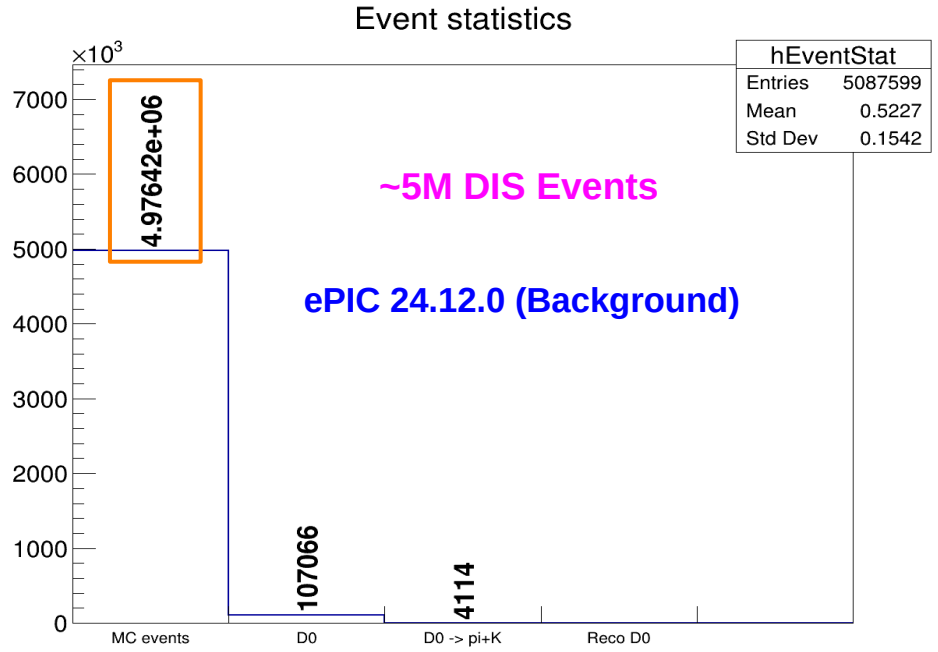
Need to scan to find Best BDT Cut

Data Sample for ML (Q^2 = 1 GeV^2)

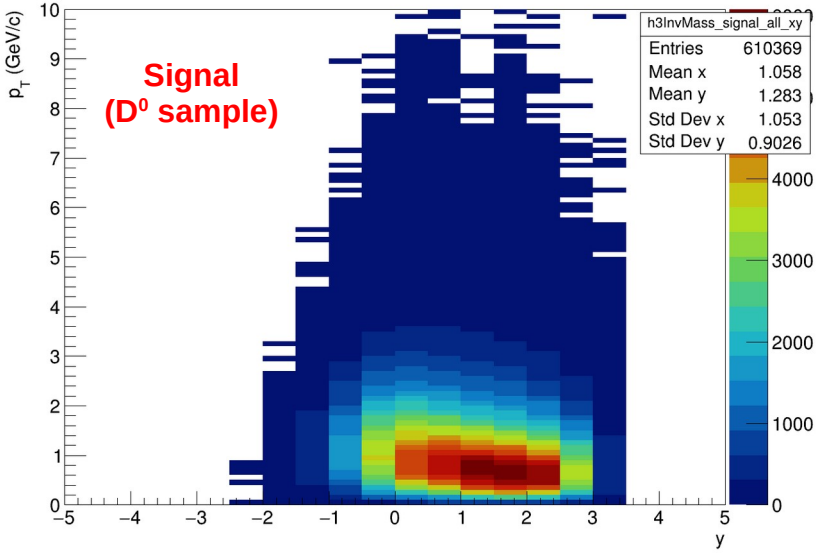
- BDT requires the features for the signal D^0 meson and background D^0 meson (fake combinations of pion,kaon)
  - D^0 enriched same created filtering **PYTHIA8 ep, NC, 10X100, Q^2 >1 GeV^2 events (~1747M)** such that each event consist one $D^0 \rightarrow$ k-π+ known as Signal taken from 24.12.0/epic_craterlake/SIDIS/D0_ABCONV/pythia8.306-1.1/10x100/q2_1):

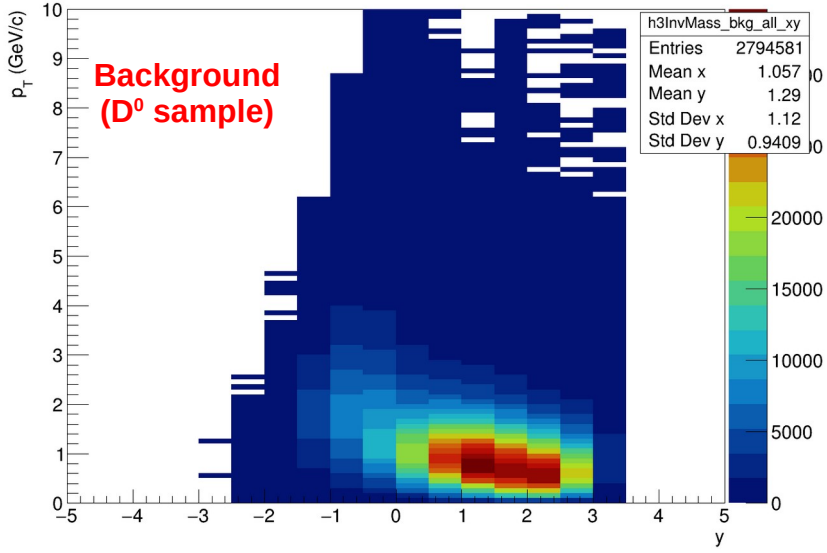
    Total files 1879 and Events = 984746
  - Background from 24.12.0/epic_craterlake/DIS/NC/10x100/minQ2=1: **Total files 5180 and Events = 4976419**

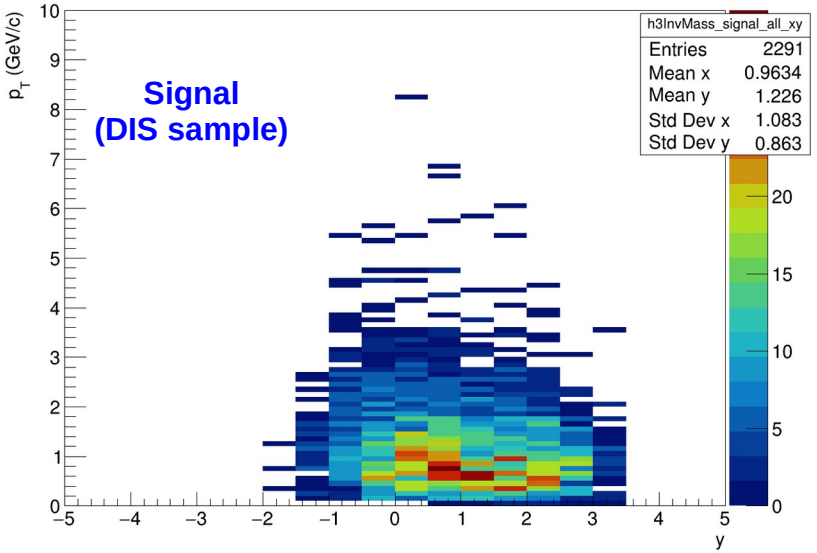


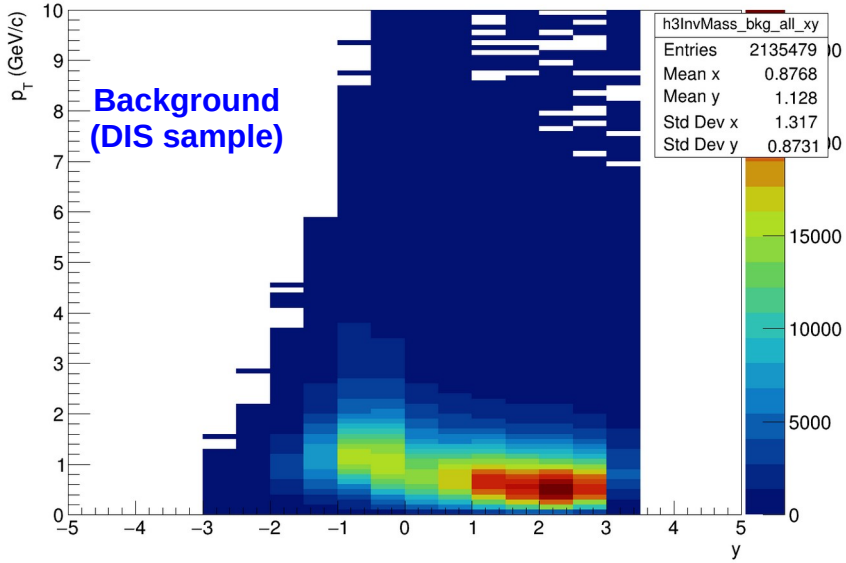


Phase Space (Q^2 = 1 GeV^2)

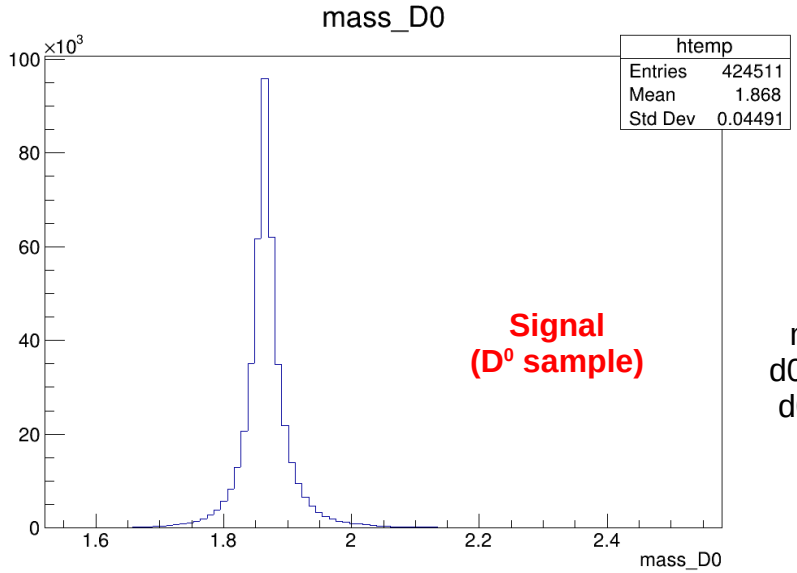






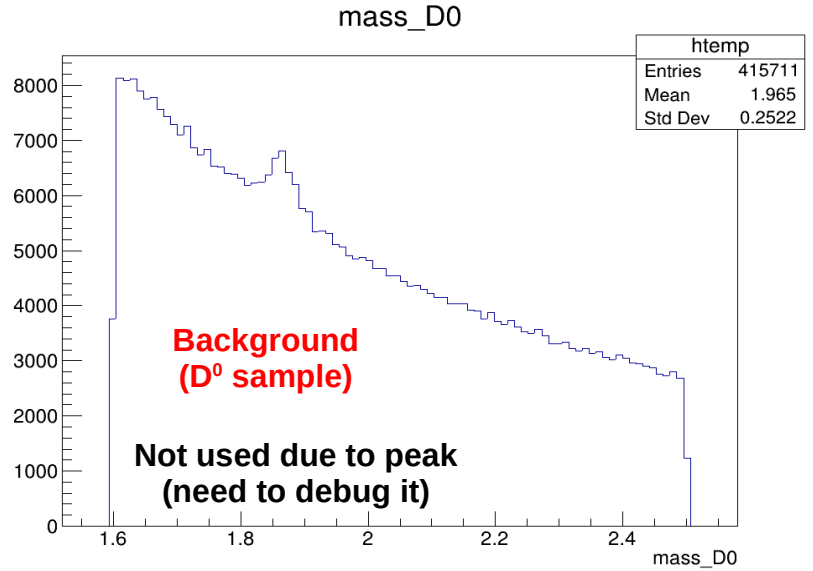


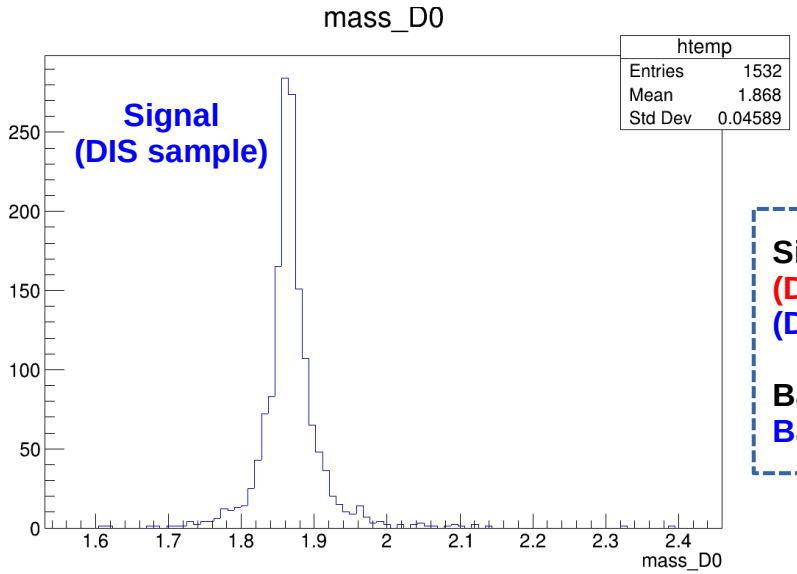
Sample After Preselection (Q^2 = 1 GeV^2)



Preselection

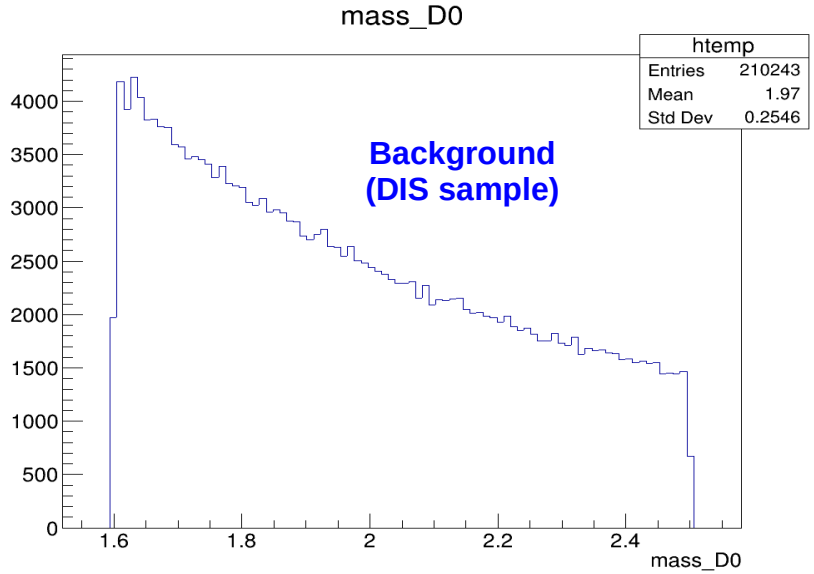
mD0 > 1.6 && mD0 < 2.5
d0xypi>0.02 && d0xypi<10.
d0xyk>0.02 && d0xyk<10.
decay length <100.





Signal for ML: Signal (D^0 sample) +Signal (DIS)

Background for ML: Background (DIS)



Number of Signal and Background

y(D0)	p_T(D0)	Signal	Background
-1.0 to 1.0	1.0-2.0	8211	8211
-1.0 to 1.0	2.0-5.0	964	964
-1.0 to 1.0	5.0-10.0	29	29
1.0 to 3.0	1.0-2.0	17509	17509
1.0 to 3.0	2.0-5.0	2398	2398
1.0 to 3.0	5.0-10.0	38	38
-3.0 to -1.0	1.0-2.0	2811	2811
-3.0 to -1.0	2.0-5.0	417	417

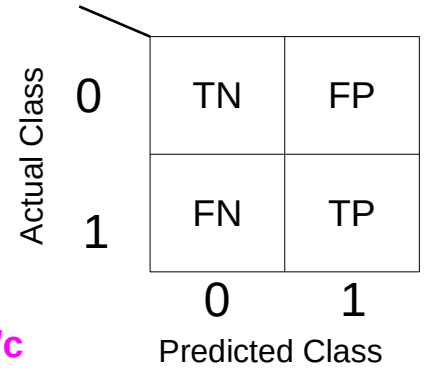
Keep the number of signal and background same for ML

Two approaches for parameter optimization: **Optuna and Bayesian**

Bayesian takes more computing time

Confusion Matrix

Confusion matrix



Predicted: Positive / Negative

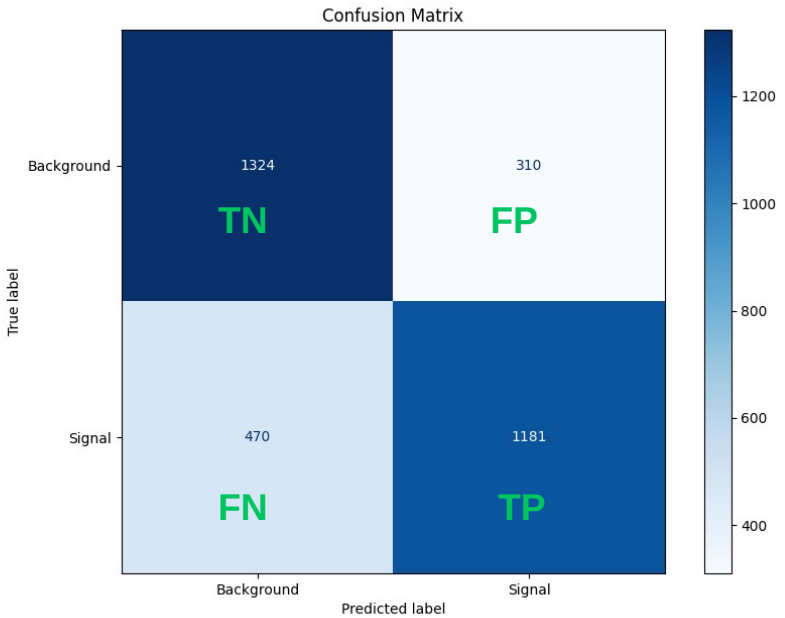
$$TPR = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

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

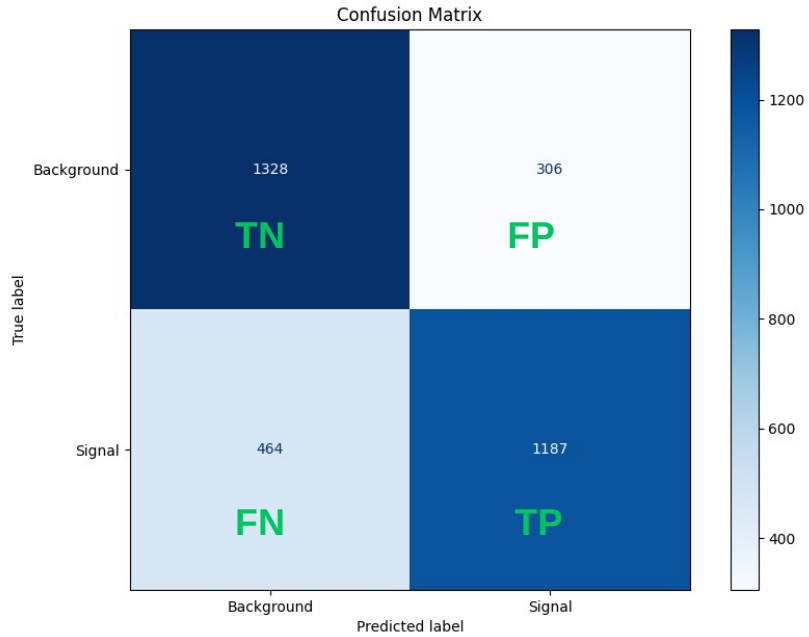
-1.0 < y < 1.0
1.0 < p_T < 2.0 GeV/c

Test sample = 0.2*16422 ~ 3285 = 1328+306+464+1187

Test Sample (Optuna)



Test Sample (Bayesian)

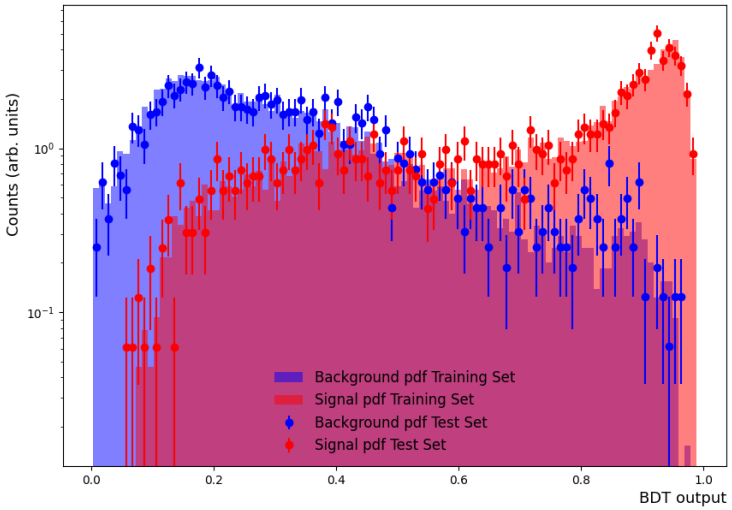


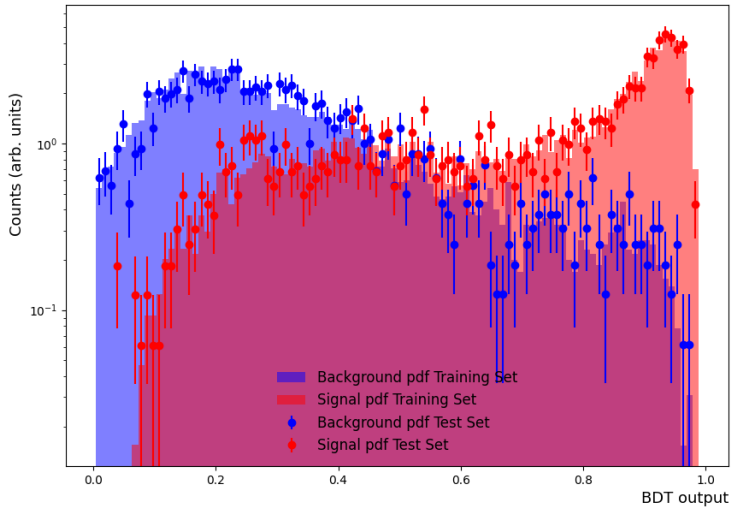
BDT Output

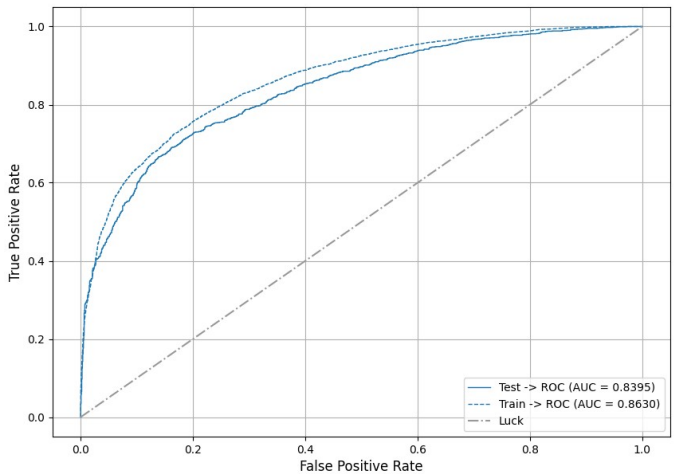
Optuna

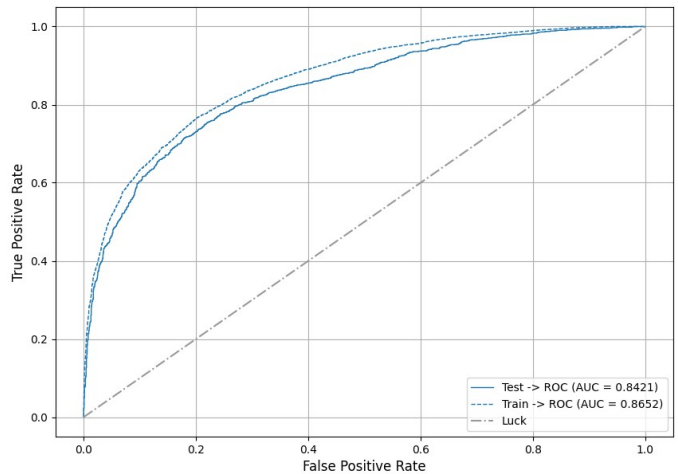
-1.0 < y < 1.0
1.0 < p_T < 2.0 GeV/c

Bayesian





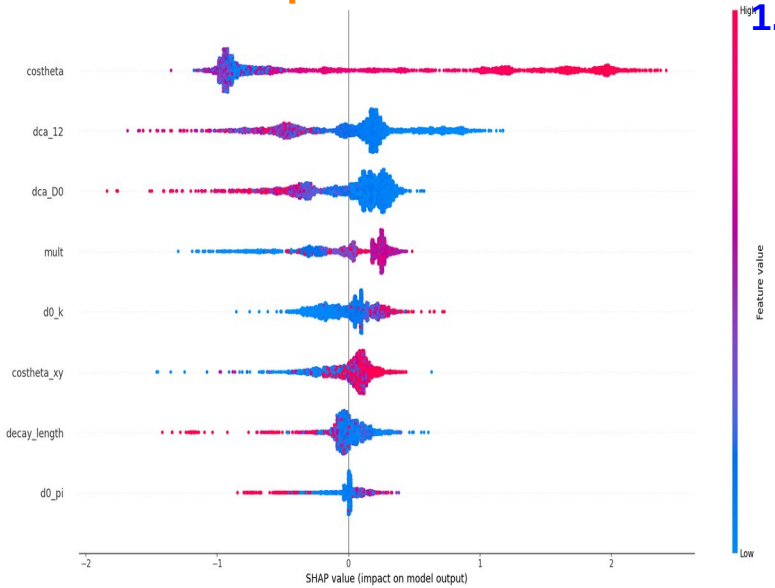


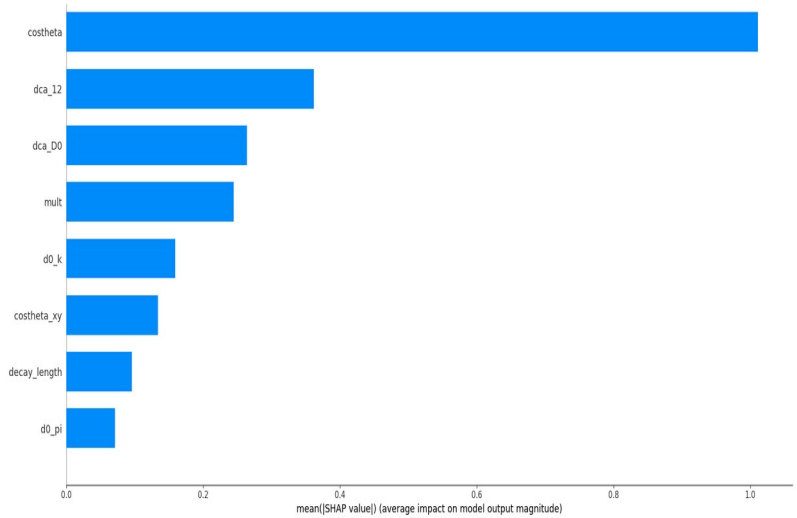


Results

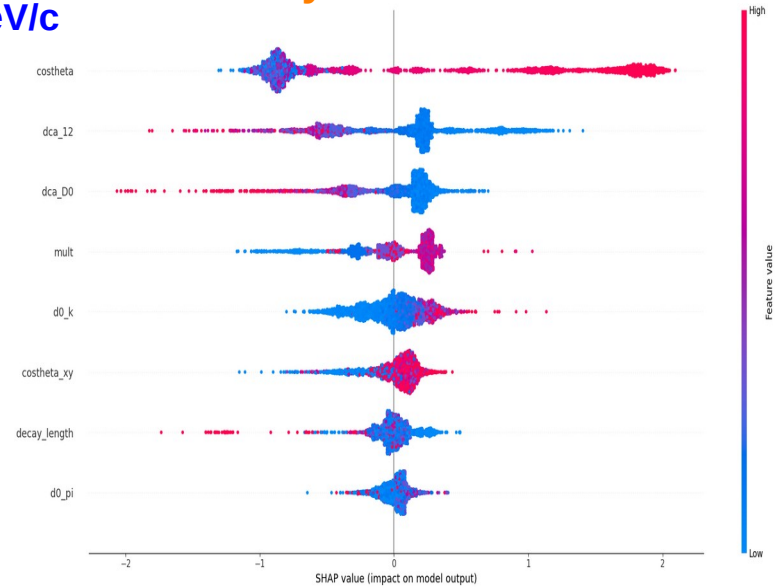
Optuna

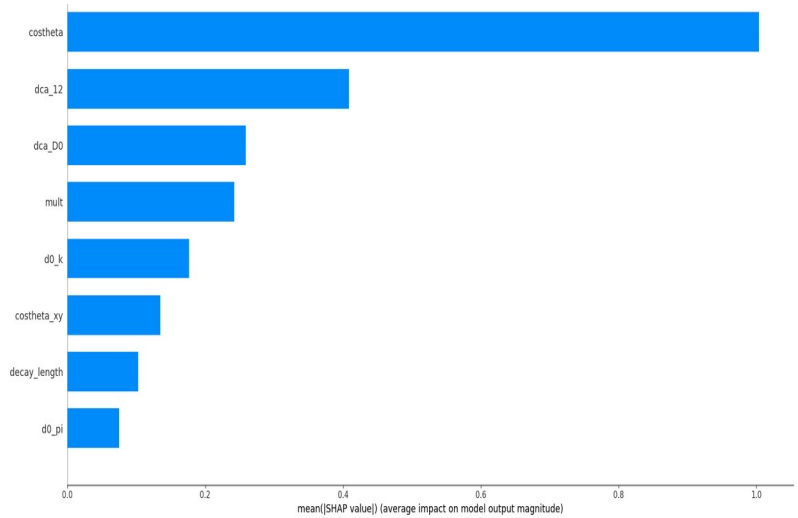
-1.0 < y < 1.0
1.0 < p_T < 2.0 GeV/c





Bayesian

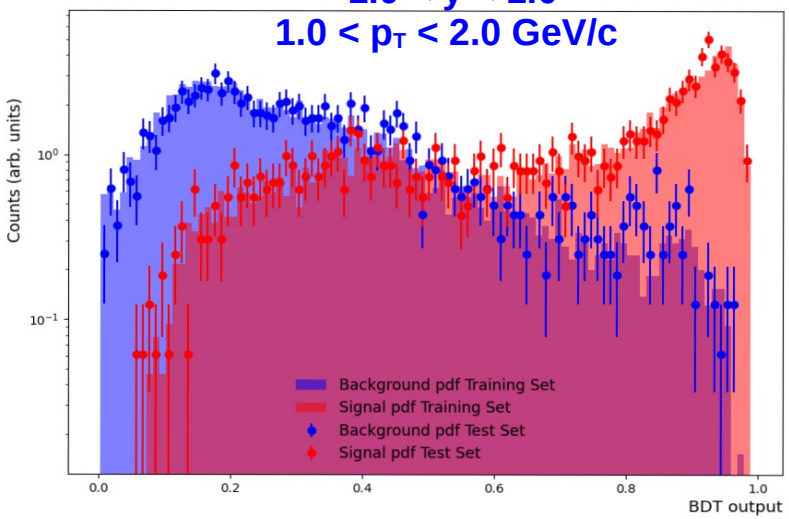


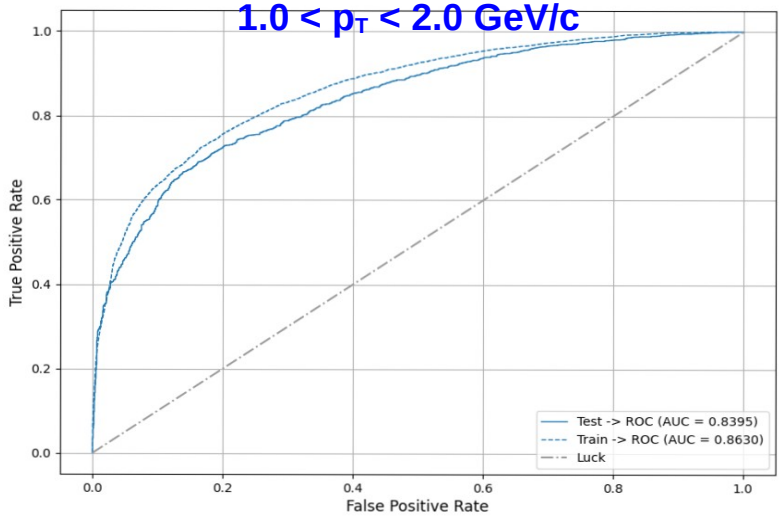


BDT Output

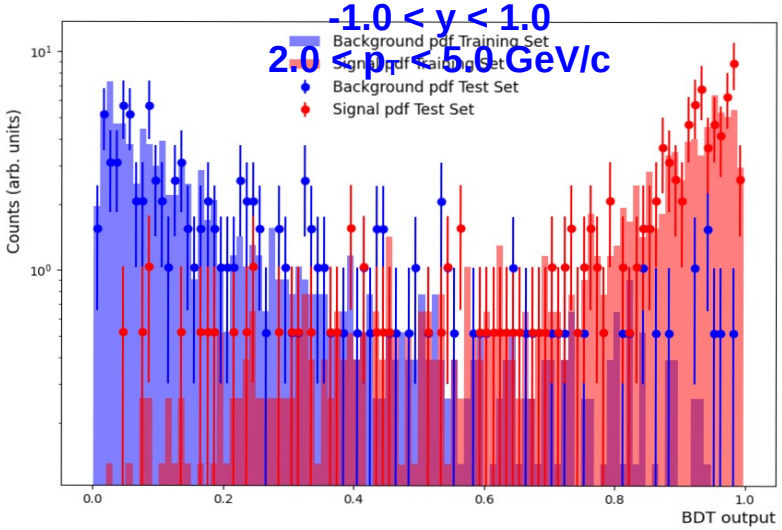
Study with Optuna

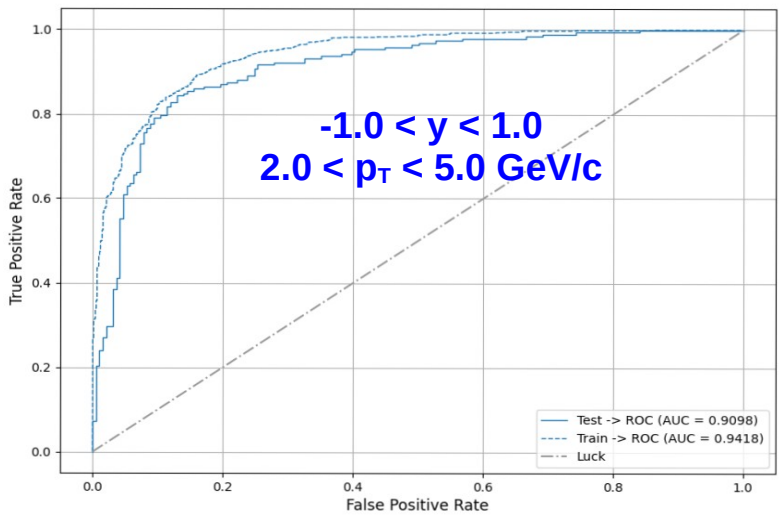
-1.0 < y < 1.0
1.0 < p_T < 2.0 GeV/c





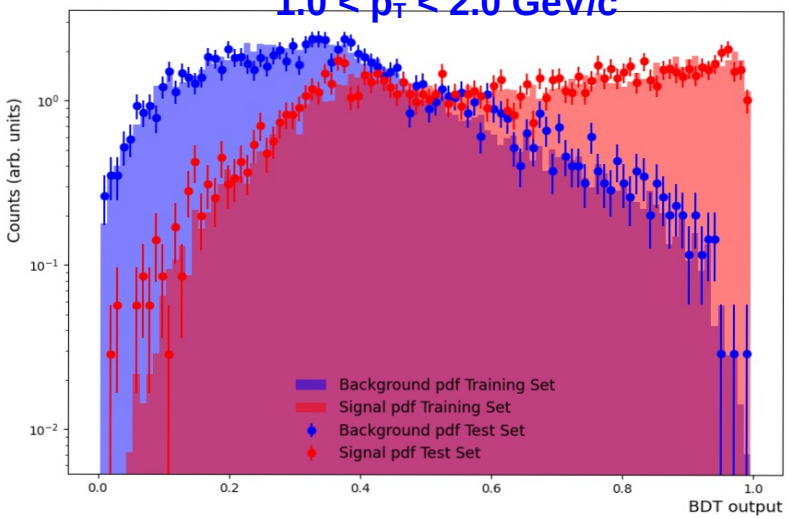
-1.0 < y < 1.0
2.0 < p_T < 5.0 GeV/c





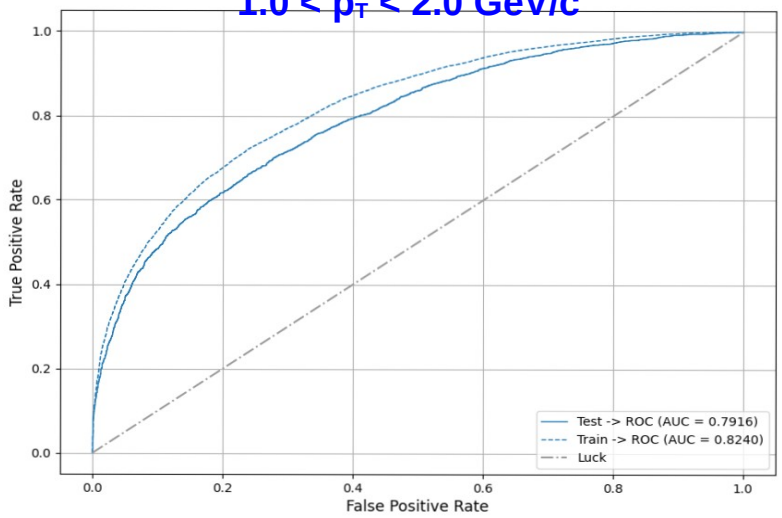
BDT Output

1.0 < y < 3.0
1.0 < p_T < 2.0 GeV/c

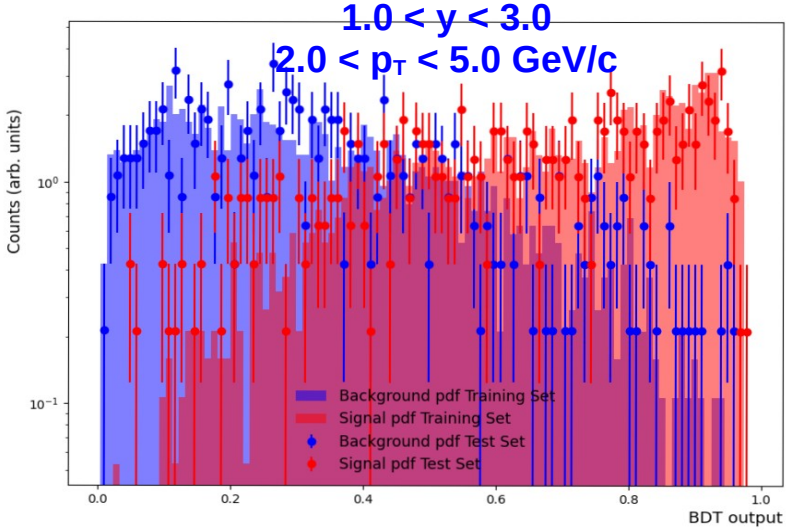


Study with Optuna

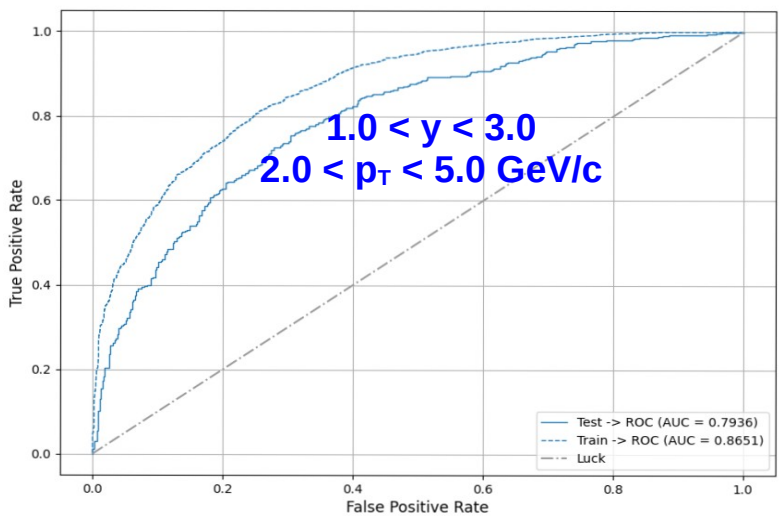
-1.0 < y < 1.0
1.0 < p_T < 2.0 GeV/c



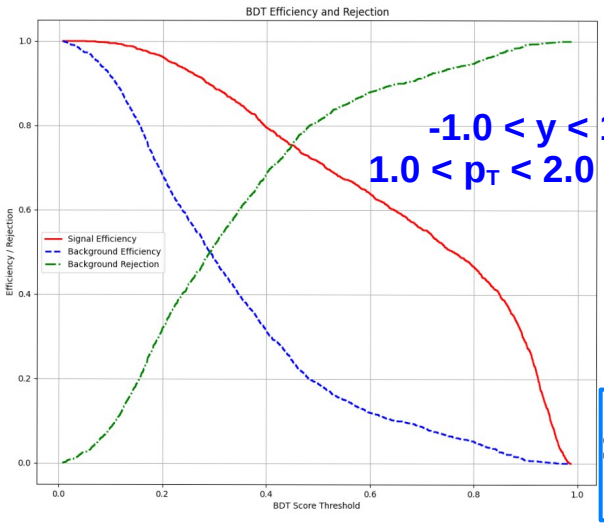
1.0 < y < 3.0
2.0 < p_T < 5.0 GeV/c



Not enough statistics

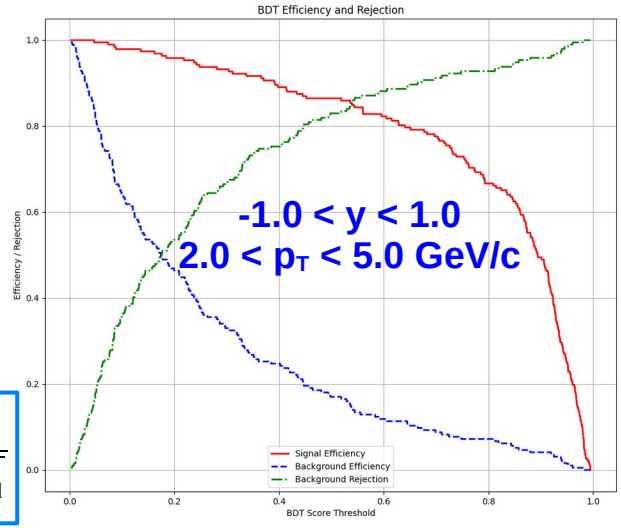


BDT Efficiencies



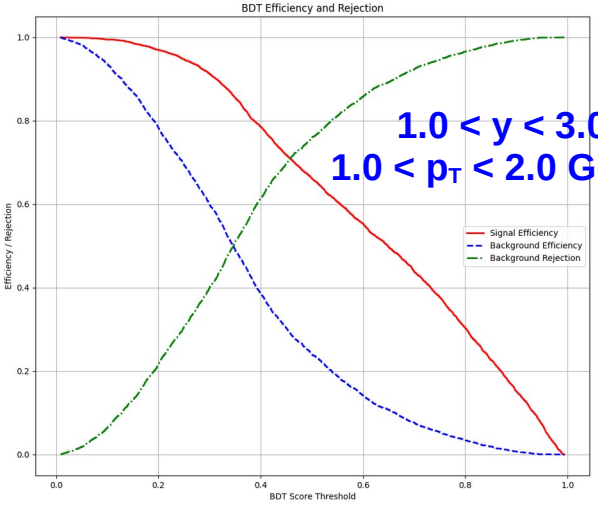
-1.0 < y < 1.0
1.0 < p_T < 2.0 GeV/c

$$\text{Signi}_{\text{BDT Cut}} = \frac{S_{\text{No ML}} \times \epsilon_{\text{Signal}}}{\sqrt{S_{\text{No ML}} \times \epsilon_{\text{Signal}} + B_{\text{No ML}} \times \epsilon_{\text{Background}}}}$$

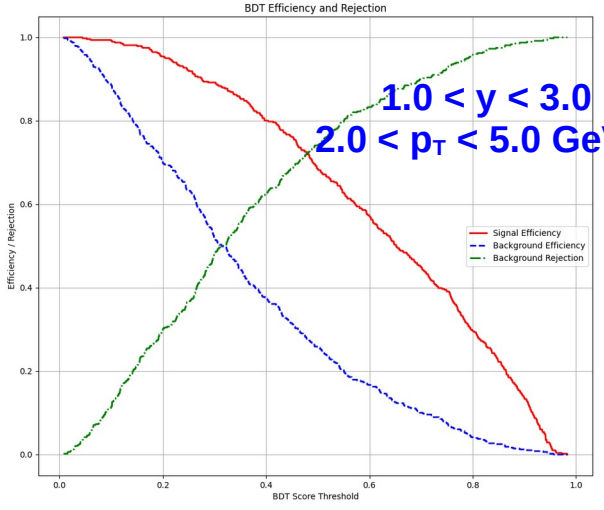


-1.0 < y < 1.0
2.0 < p_T < 5.0 GeV/c

$$\left(\frac{S}{B}\right)_{\text{BDT Cut}} = \left(\frac{S}{B}\right)_{\text{No ML}} \times \frac{\epsilon_{\text{Signal}}}{\epsilon_{\text{Background}}}$$

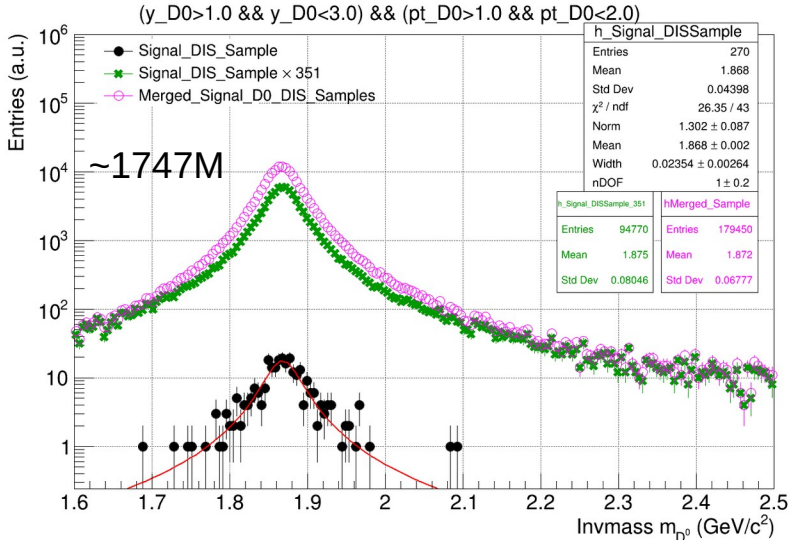


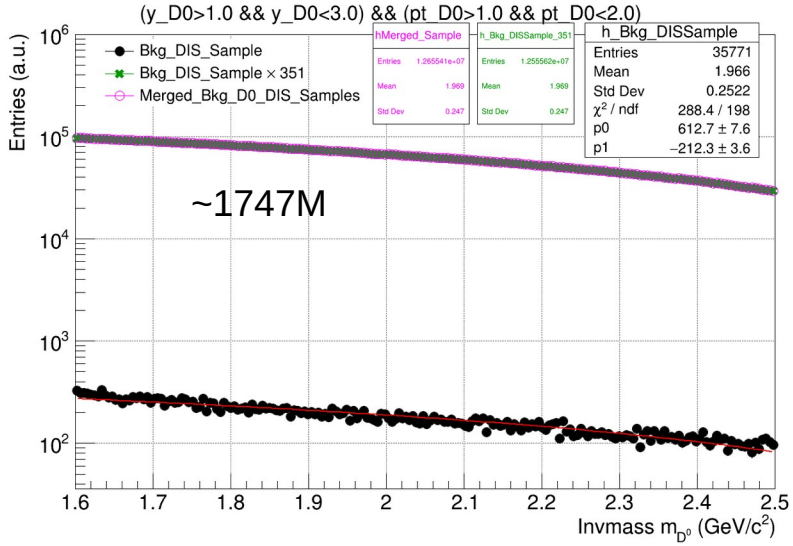
1.0 < y < 3.0
1.0 < p_T < 2.0 GeV/c

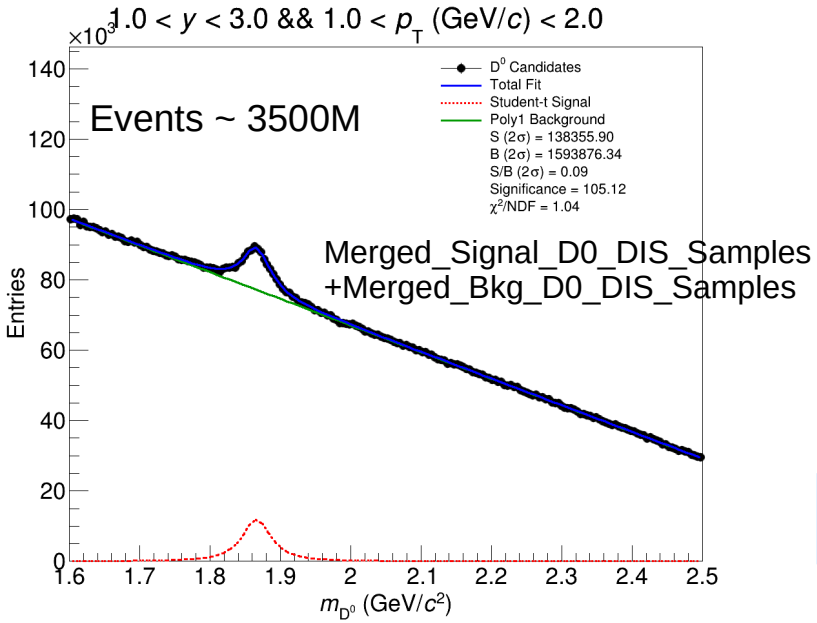


1.0 < y < 3.0
2.0 < p_T < 5.0 GeV/c

Merging Signal and Background (D0 Sample and DIS Events)







PYTHIA simulation for 10x100 ep collisions with $Q^2 > 1$ GeV^2: 990k D0 events corresponds to luminosity 3.7 fb^{-1} ~ 1747 M Events

Expected Luminosity 10 fb^{-1} corresponds to ~ 1747*10/3.7 = 4724.6 M

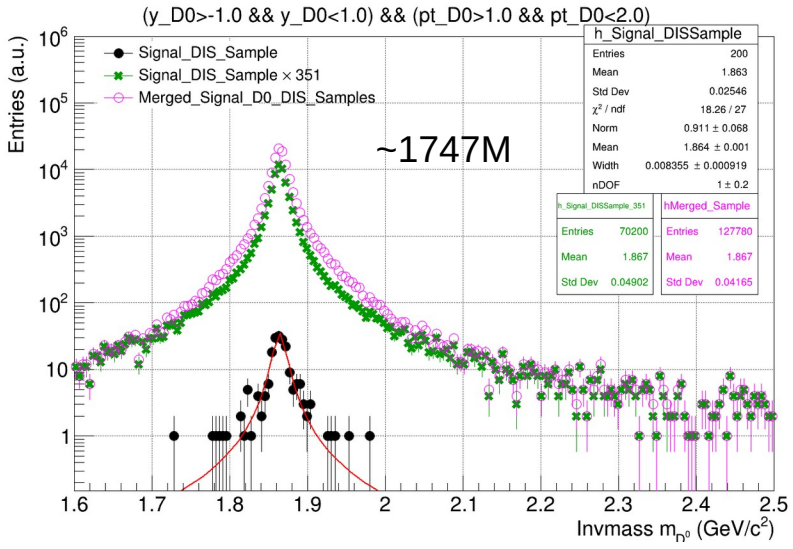
$$\left(\frac{S}{B}\right)_{\text{BDT Cut = 0.4}} \sim 0.09 * 0.8 / 0.4 = 0.18$$

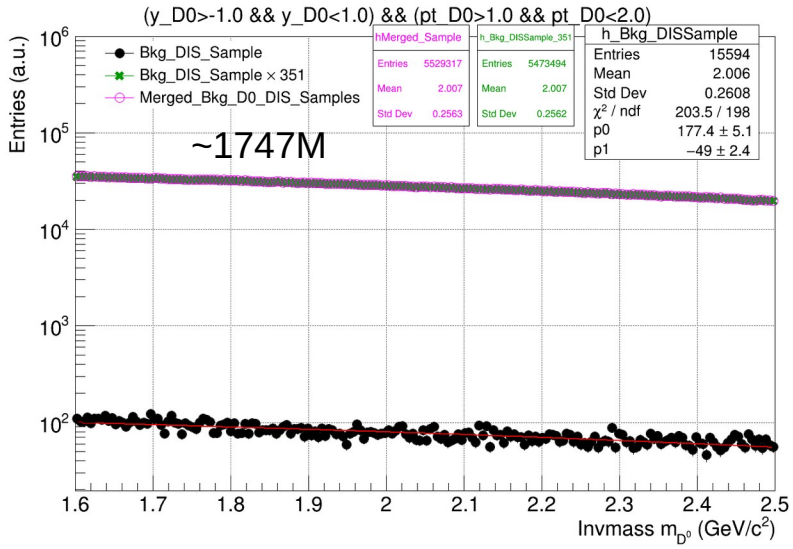
$$Significance_{\text{BDT Cut = 0.4}} \sim \frac{138355.90 * 0.8}{\sqrt{138355.90 * 0.8 + 1593876.34 * 0.2}} = 168.90$$

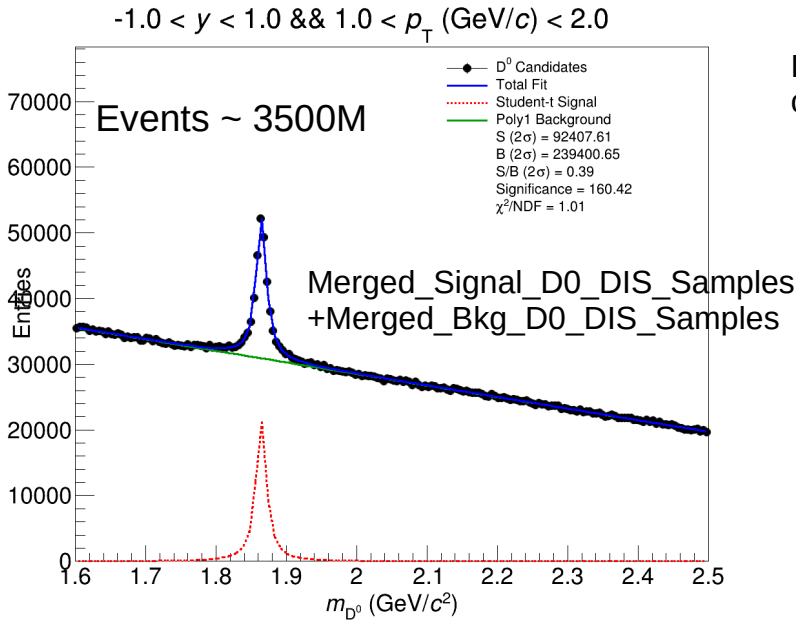
$$Expected\ Significance_{\text{BDT Cut = 0.4}} \quad 168.90 * \sqrt{(4724.6/3500)} = 196.23$$

Need to scan to find Best BDT Cut

Merging Signal and Background (D0 Sample and DIS Events)







PYTHIA simulation for 10x100 ep collisions with $Q^2 > 1$ GeV^2: 990k D0 events corresponds to luminosity 3.7 fb^{-1} ~ 1747 M DIS Events

Expected Luminosity 10 fb^{-1} corresponds to ~ 1747*10/3.7 = 4724.6 M

$$\left(\frac{S}{B}\right)_{\text{BDT Cut = 0.4}} \sim 0.39 * 0.8/0.3 = 1.04$$

$$Significance_{\text{BDT Cut = 0.4}} \sim \frac{92407.61 * 0.8}{\sqrt{92407.61 * 0.8 + 239400.65 * 0.3}} = 193.64$$

$$Expected\ Significance_{\text{BDT Cut = 0.4}} \quad 193.64 * \sqrt{(4724.6/3500)} = 225$$

Need to scan to find Best BDT Cut

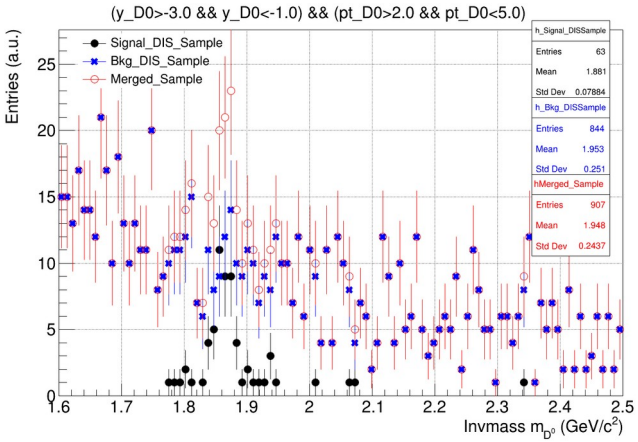
Summary and Future Plan

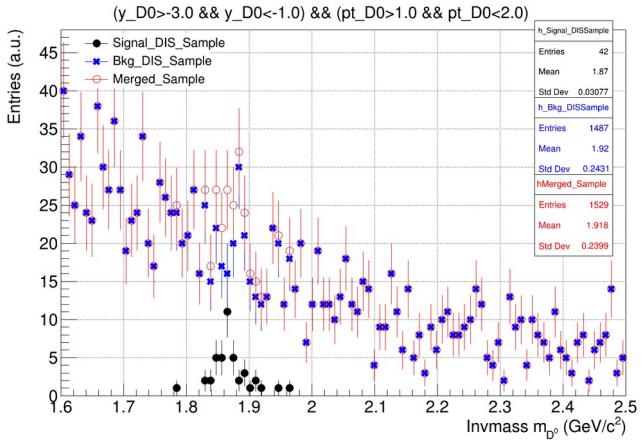
- Machine learning model studies performed for the D^0 reconstruction in differential p_T and y bins
- Implemented the Bayesian approach for optimization
- Estimation the realistic signal/background and significance is in progress
- Future Steps:
  - Scan BDT cut to get the best results
  - Implement a similar model for Λ_c^+ reconstruction
  - Estimate Λ_c^+/D^0 ratio using machine learning
  - Further include more variables e.g. chi2 of secondary vertex fit, nsigma kaon, and nsigma pion, etc.
  - Further optimize the model parameters to get highest ROC score
  - Implement other models e.g. Neural Network (Classifier as well as AutoEncoder)

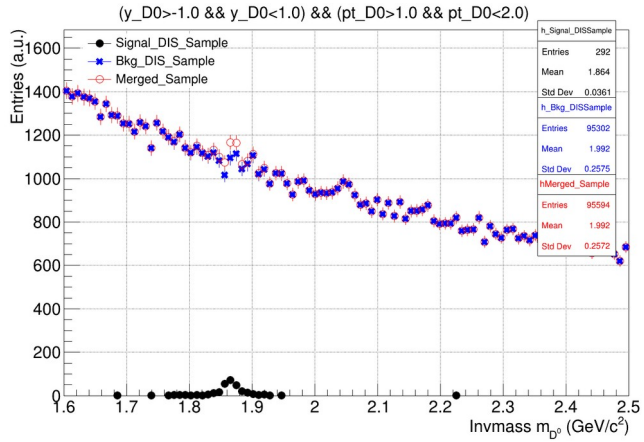
THANK YOU !!!

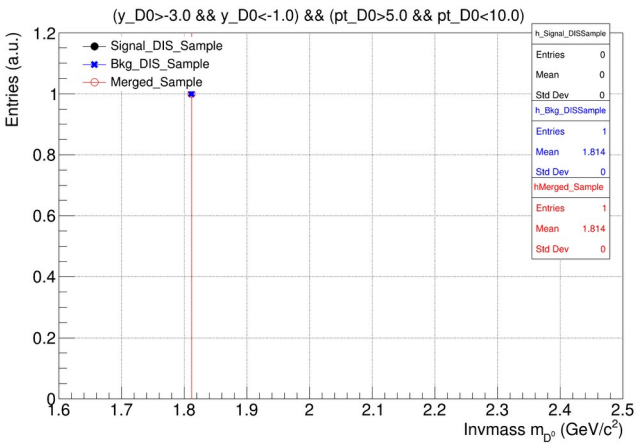
Only DIS Sample (Q2>100 GeV^2)

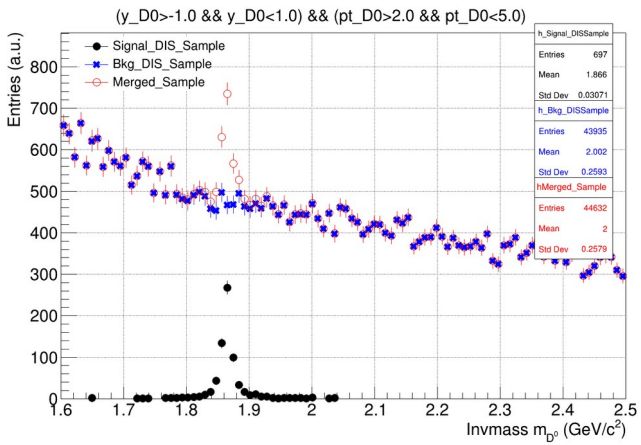
Corresponds to 5M Events merged Signal+Background

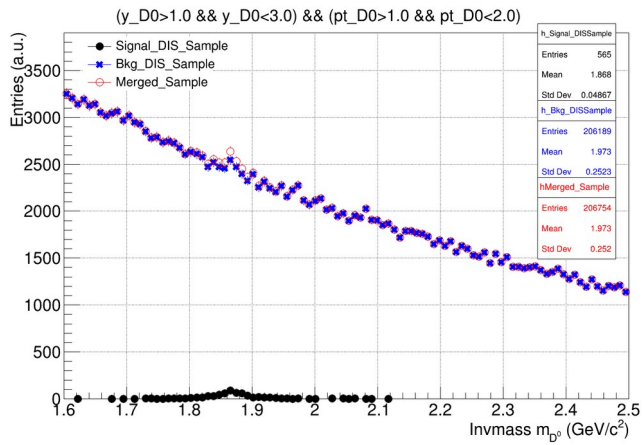






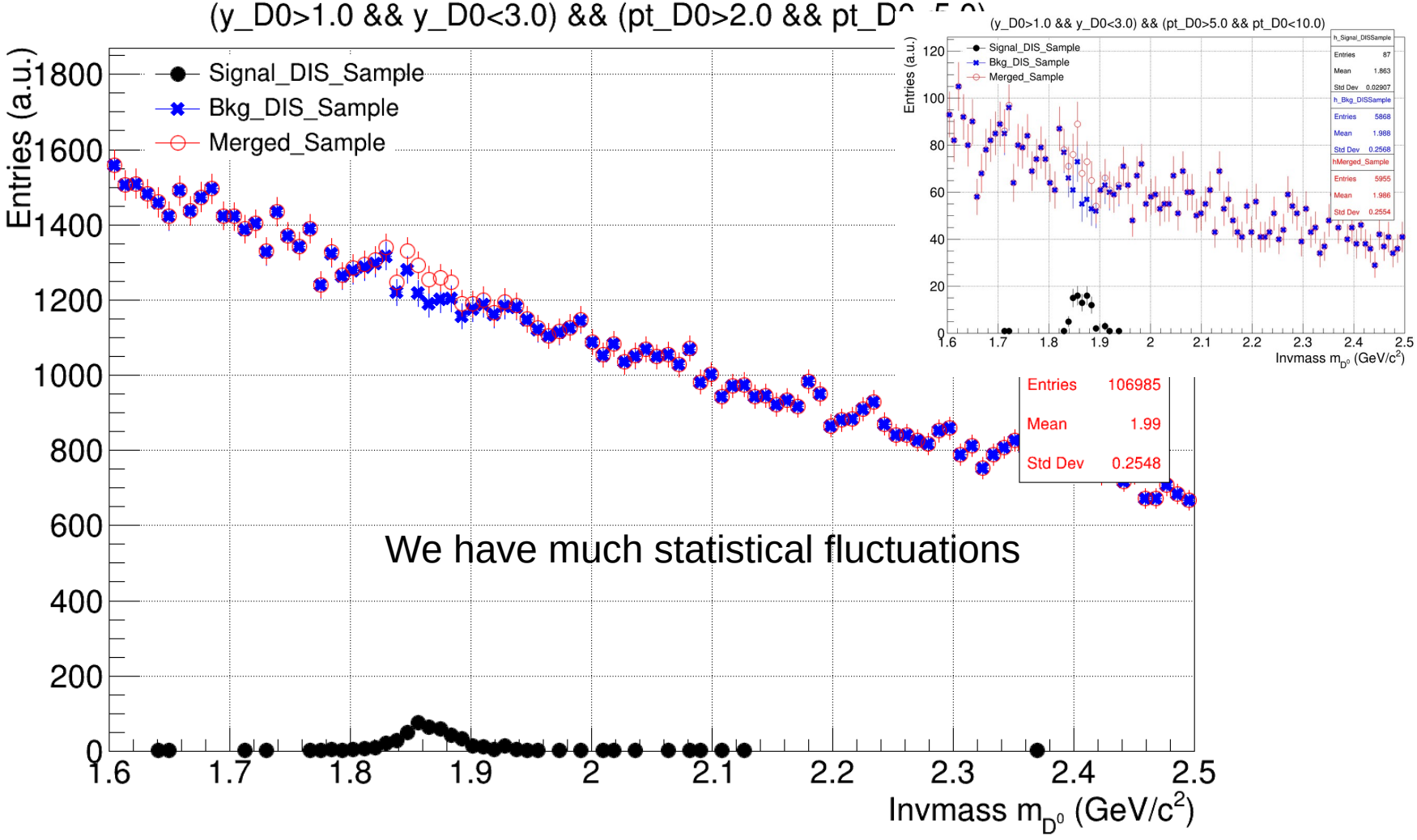






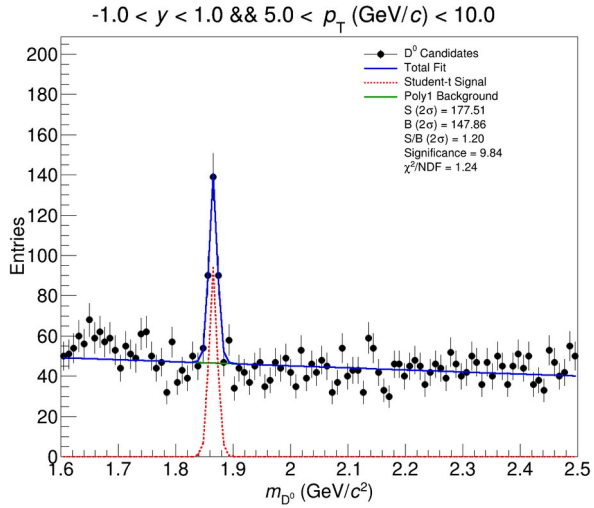
Only DIS Sample (Q2>100 GeV^2)

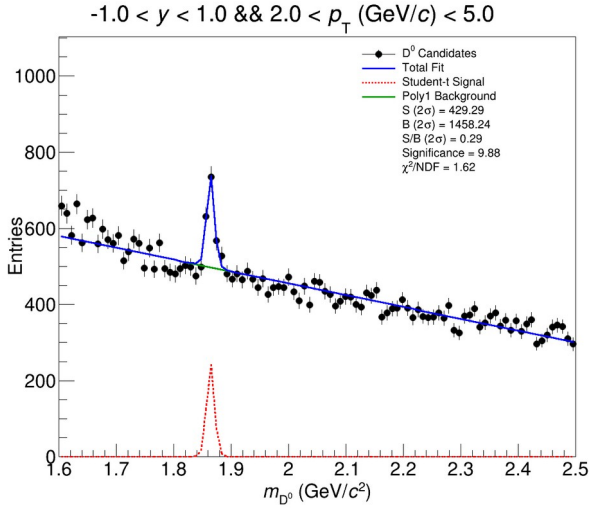
Corresponds to 5M Events merged Signal+Background

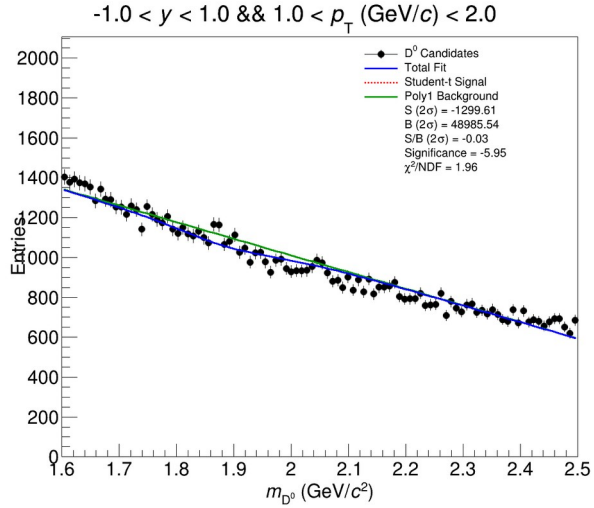


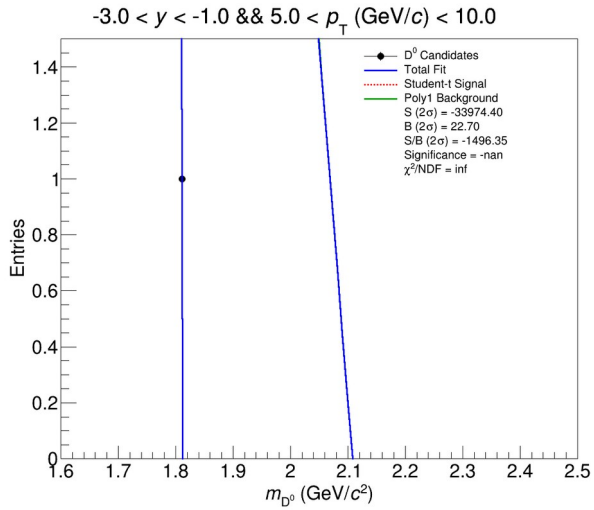
Only DIS Sample (Q2>100 GeV^2)

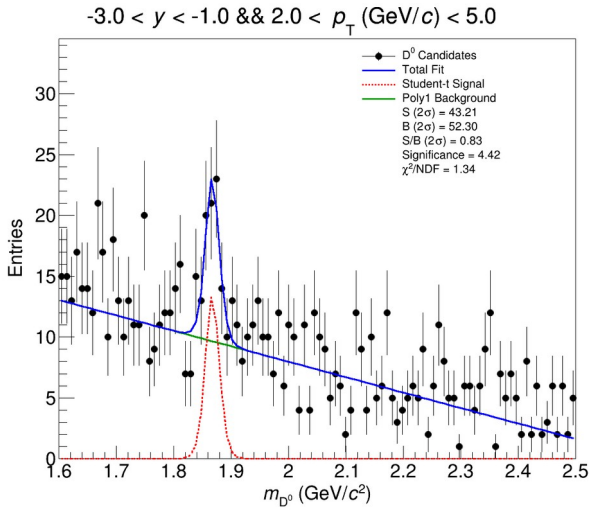
Corresponds to 5M Events

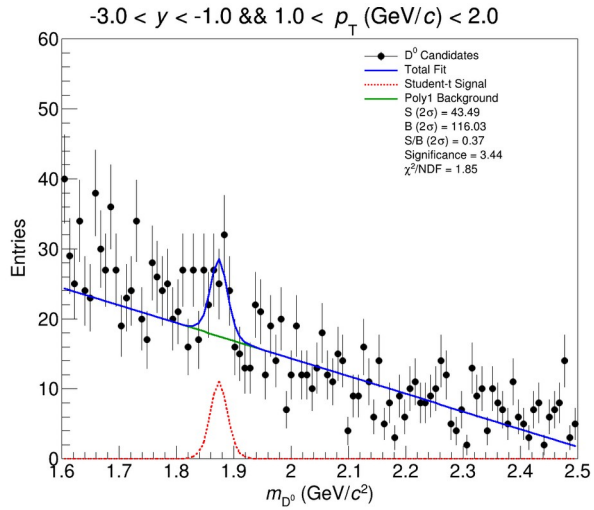






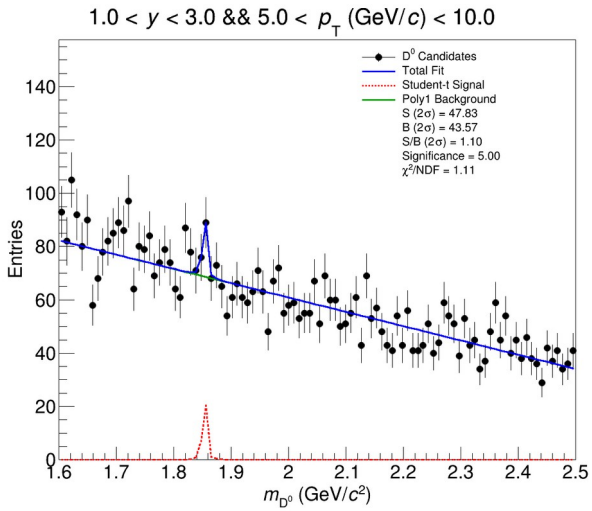


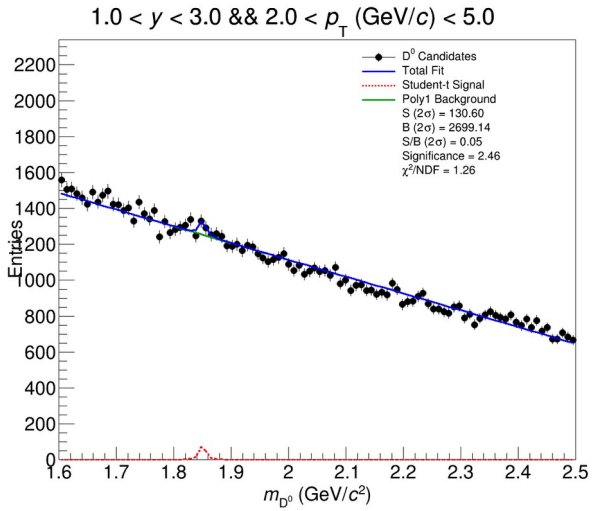


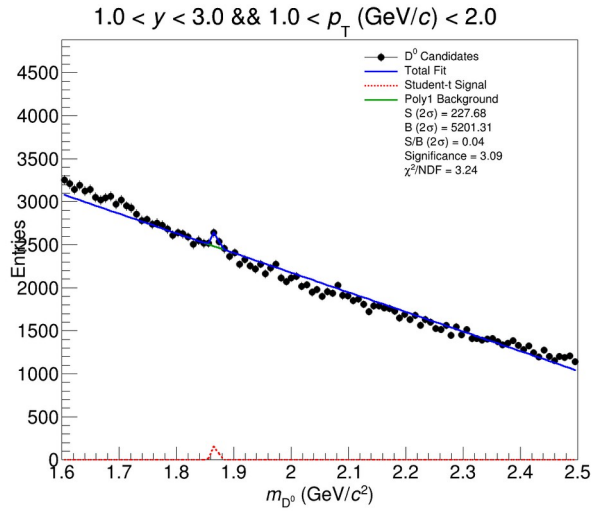


Only DIS Sample (Q2>100 GeV^2)

Corresponds to 5M Events merged Signal+Background



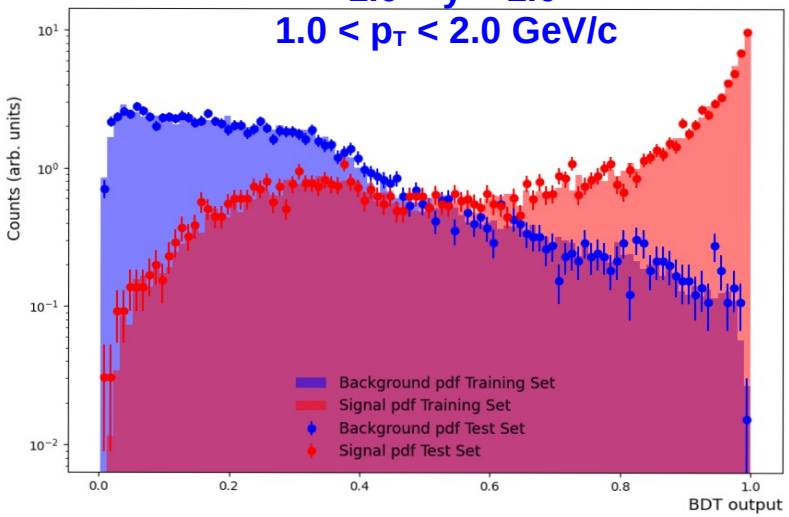




We have much statistical fluctuations

BDT Output (Q2>100)

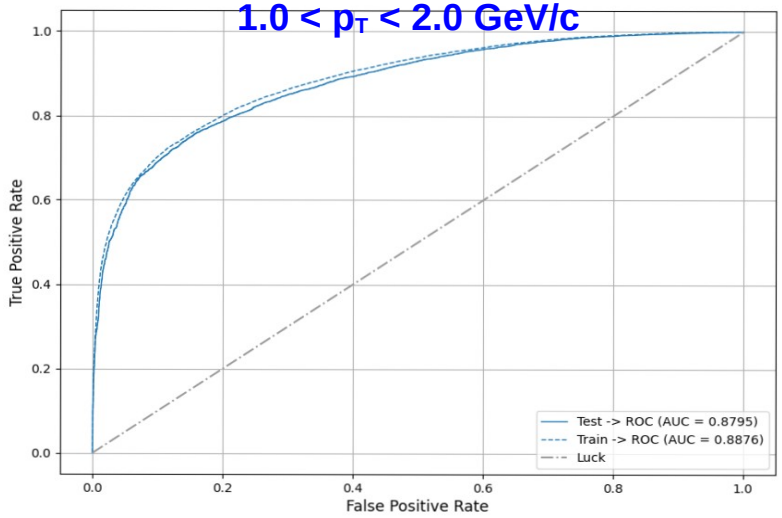
-1.0 < y < 1.0
1.0 < p_T < 2.0 GeV/c



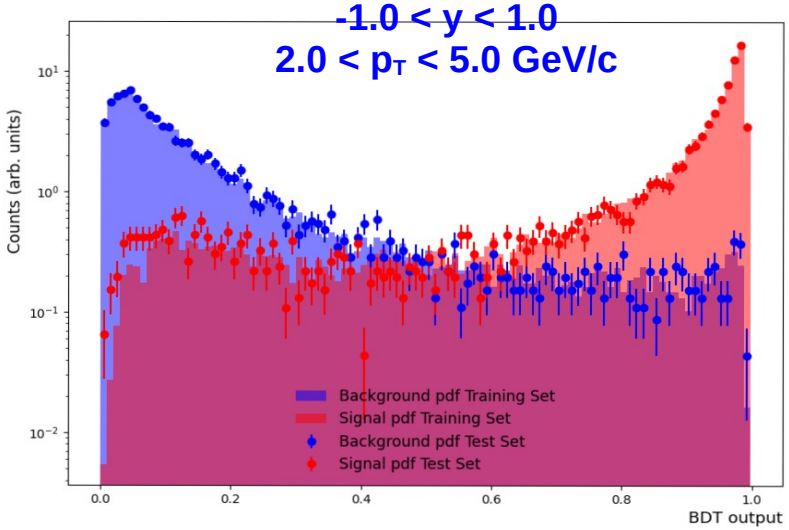
Study with Optuna

Removing
costheta_xy
decay_length

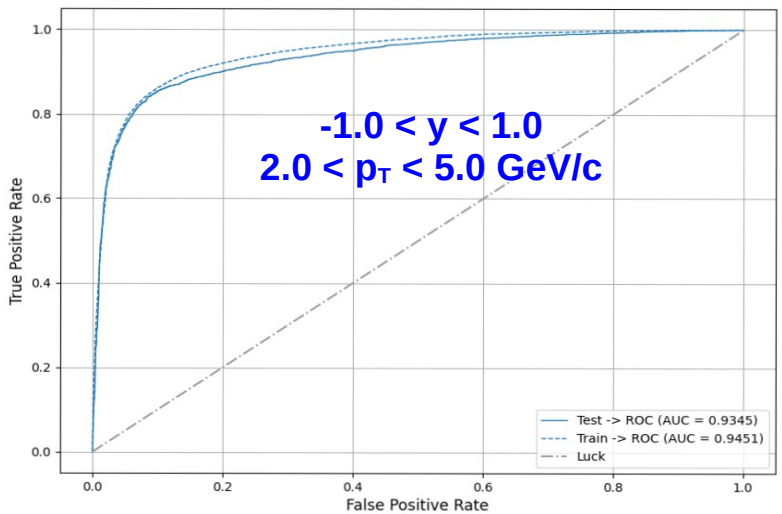
-1.0 < y < 1.0
1.0 < p_T < 2.0 GeV/c



-1.0 < y < 1.0
2.0 < p_T < 5.0 GeV/c



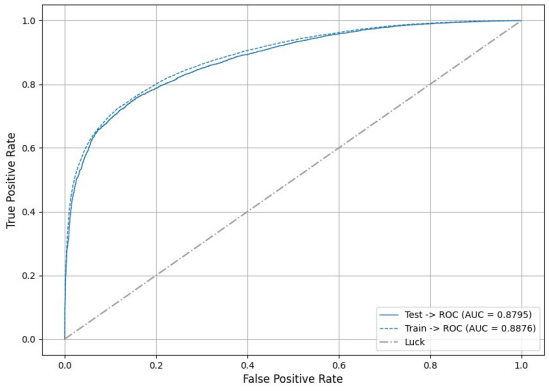
-1.0 < y < 1.0
2.0 < p_T < 5.0 GeV/c



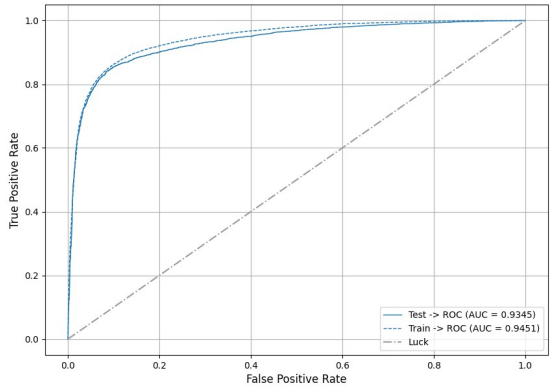
Optuna vs Bayesian (Q^2>100 and -1 < y <1)

Optuna

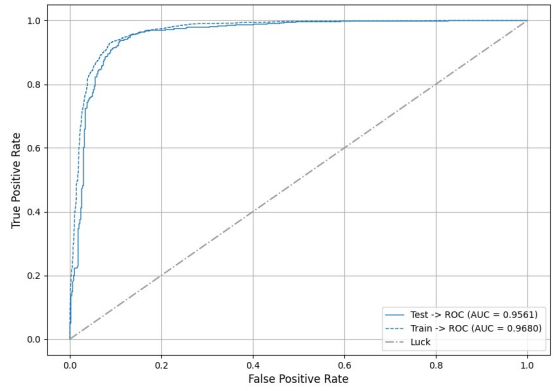
$1 < p_T < 2$ GeV/c



$2 < p_T < 5$ GeV/c

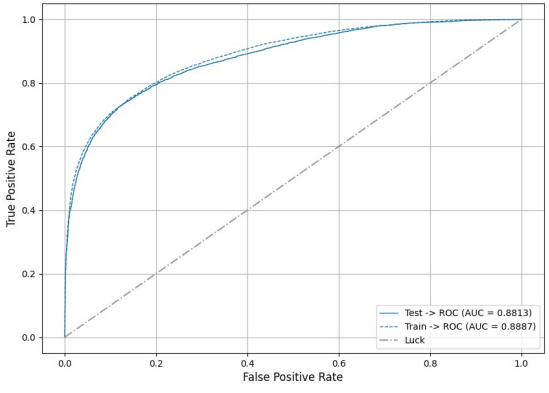


$5 < p_T < 10$ GeV/c

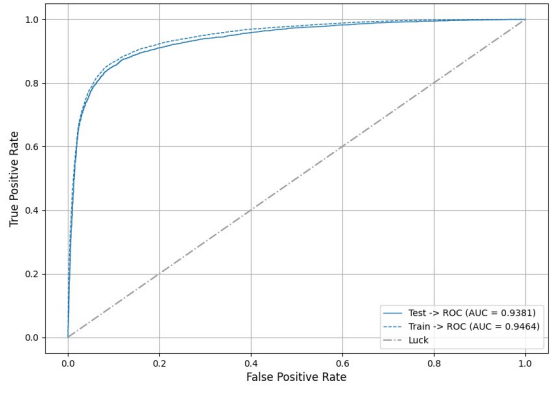


Bayesian

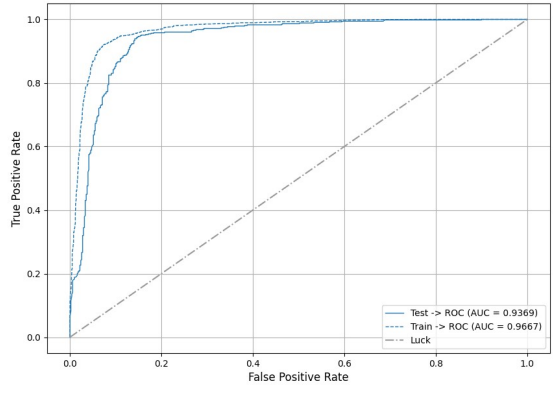
$1 < p_T < 2$ GeV/c



$2 < p_T < 5$ GeV/c



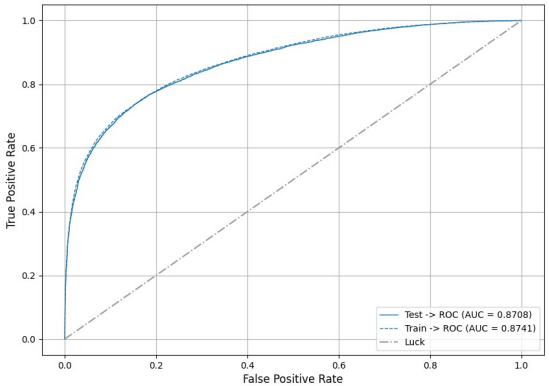
$5 < p_T < 10$ GeV/c



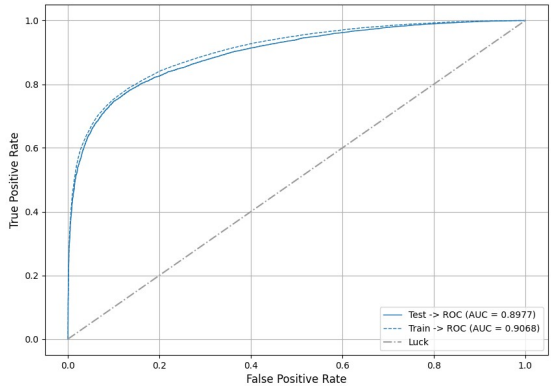
Optuna vs Bayesian (Q^2>100 and 1 < y <3)

Optuna

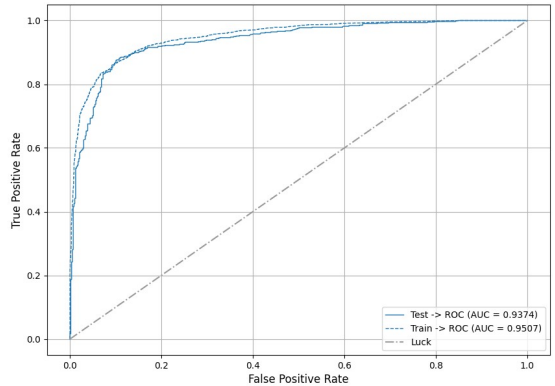
1 < p_T< 2 GeV/c



2 < p_T< 5 GeV/c

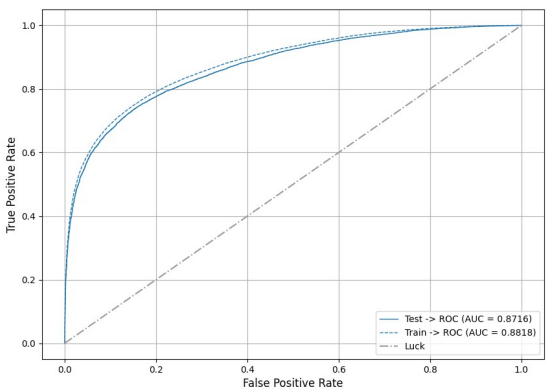


5 < p_T< 10 GeV/c

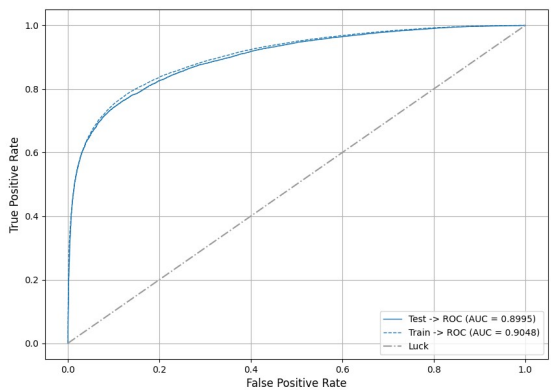


Bayesian

1 < p_T< 2 GeV/c



2 < p_T< 5 GeV/c



5 < p_T< 10 GeV/c