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

Shyam Kumar*, Annalisa Mastroserio, Domenico Elia

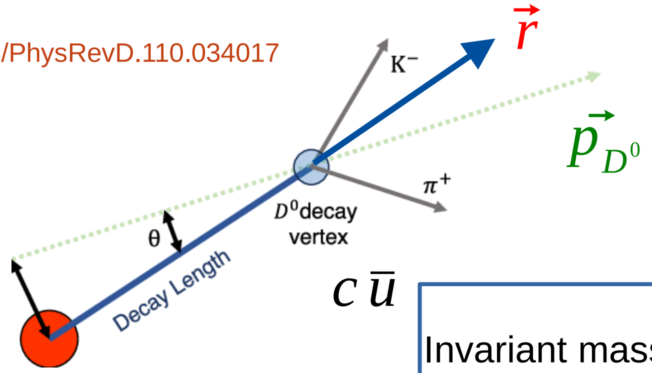
INFN Bari, Italy

hipe4ml package

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

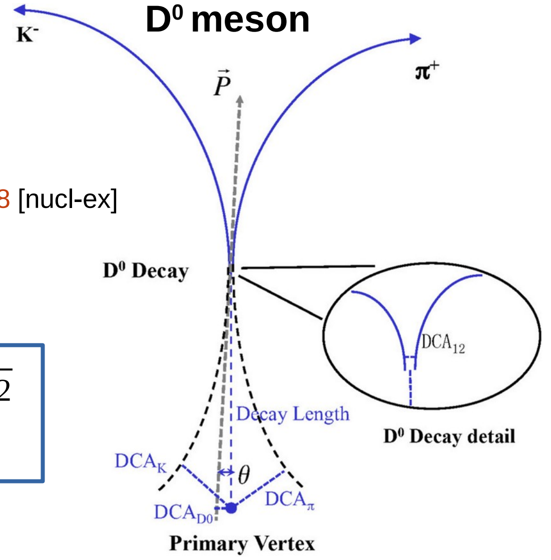
Topological Variables

10.1103/PhysRevD.110.034017



cτ = 123 μm

arXiv:1911.12168 [nucl-ex]



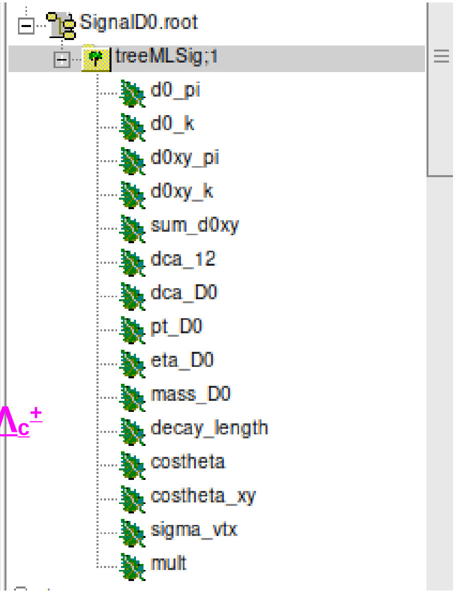
$c\bar{u}$

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

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 for D^0 meson and Truth PID for Λ_c^+ Differential in p_T and y



```
SignalD0.root
  treeMLSig;1
    d0_pi
    d0_k
    d0xy_pi
    d0xy_k
    sum_d0xy
    dca_12
    dca_D0
    pt_D0
    eta_D0
    mass_D0
    decay_length
    costheta
    costheta_xy
    sigma_vtx
    mult
```

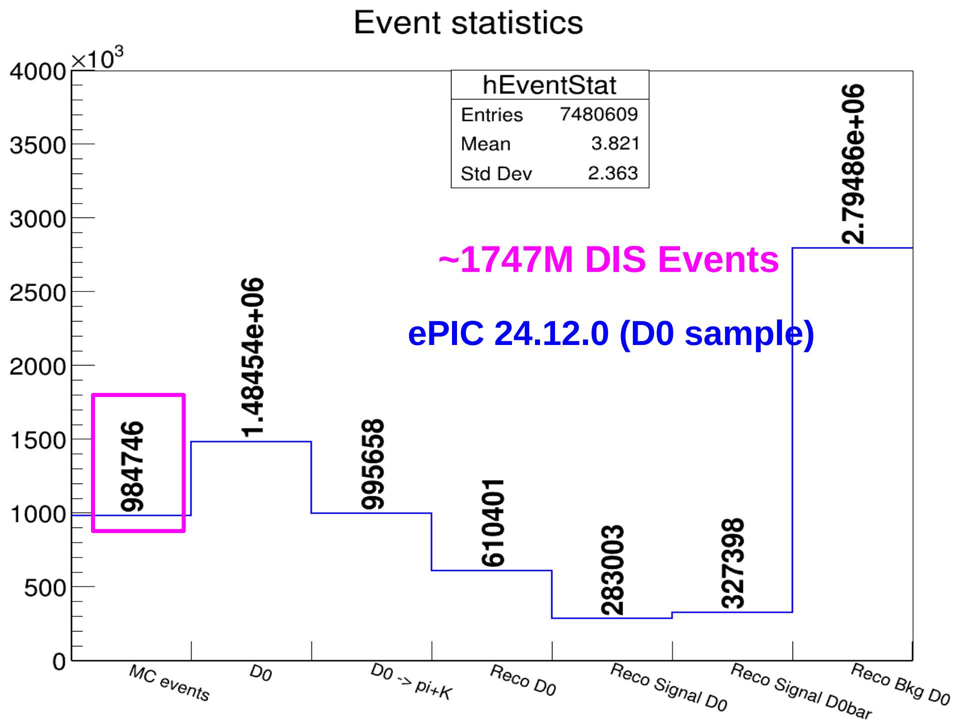
Data Sample for ML (Q^2 > 1 GeV^2)

- 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 → k-п+ known as Signal taken from 24.12.0/epic_craterlake/SIDIS/D0_ABCONV/pythia8.306-1.1/10x100/q2_1):

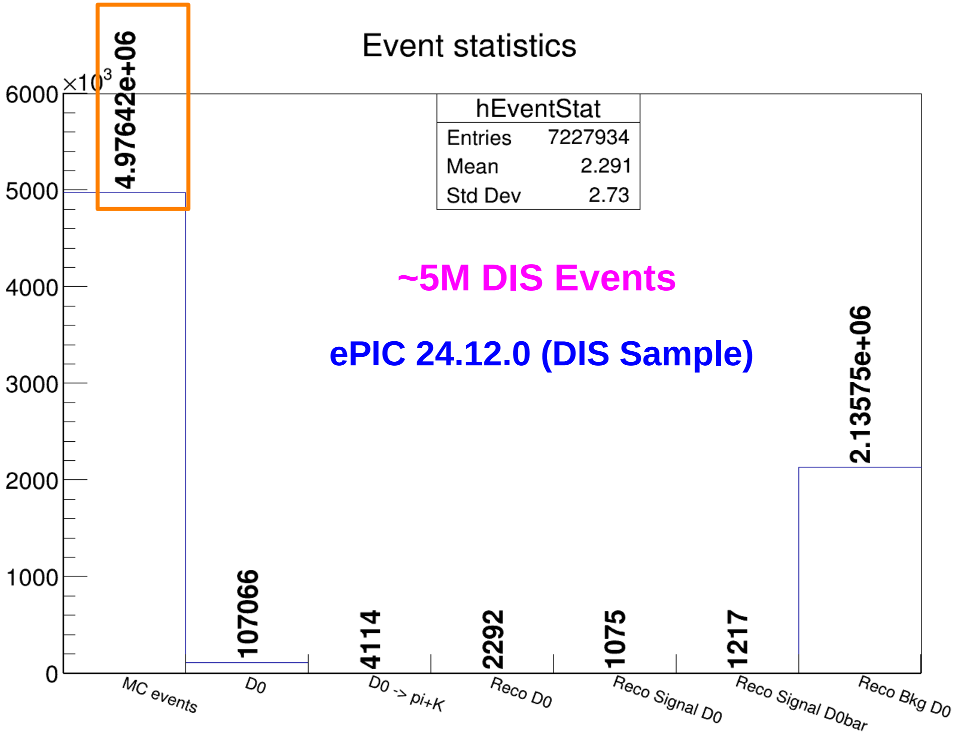
    Total files 1879 and Events = 984746

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



~1747M DIS Events

ePIC 24.12.0 (D0 sample)



~5M DIS Events

ePIC 24.12.0 (DIS Sample)

Secondary Vertex Reconstruction (D^0)

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

Secondary Vertex

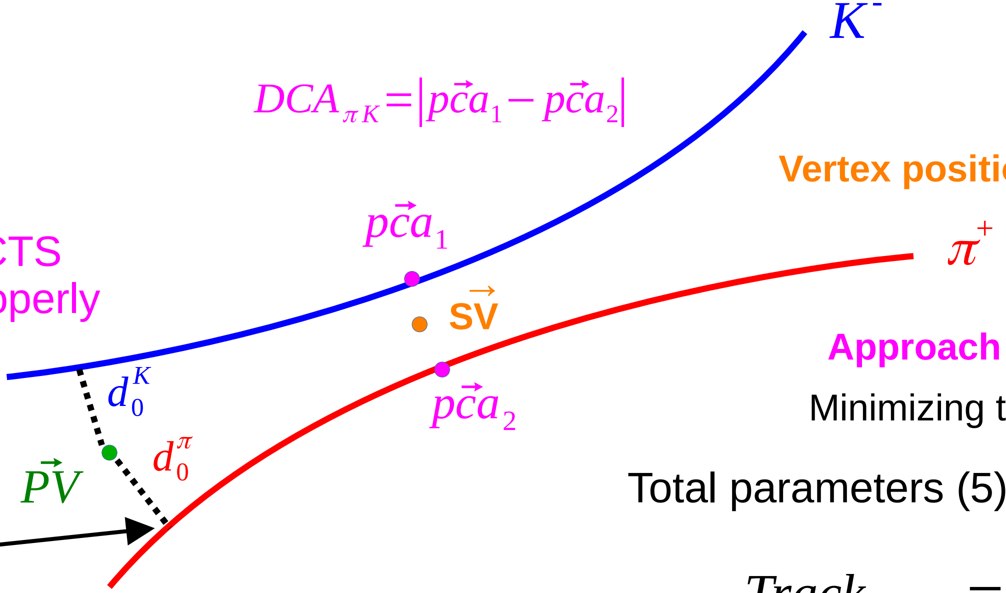
Approach 1

$$\vec{SV} = \frac{\vec{pca}_1 + \vec{pca}_2}{2}$$

$$DCA_{\pi K} = |\vec{pca}_1 - \vec{pca}_2|$$

Vertex position $\vec{SV} = (v_x, v_y, v_z)$

Secondary vertexing in ACTS considers tracking errors properly



Ignored track errors (at the moment)

$$Track_{DCA} = (\vec{r}, \vec{p}, q)$$

$$Track_{At(s)} = (\vec{r}_s, \vec{p}, q)$$ s: path length

Approach 2 (Shyam)

Minimizing the distance

Total parameters (5) = (v_x, v_y, v_z, s_1, s_2)

$$Track_{At(s1)} = (\vec{r}_{s1}, \vec{p}1, q1)$$
$$Track_{At(s2)} = (\vec{r}_{s2}, \vec{p}2, q2)$$

Minimize $d = \sqrt{(\vec{r}_{s1} - \vec{v})^2 + (\vec{r}_{s2} - \vec{v})^2}$

Comparison of four approaches:

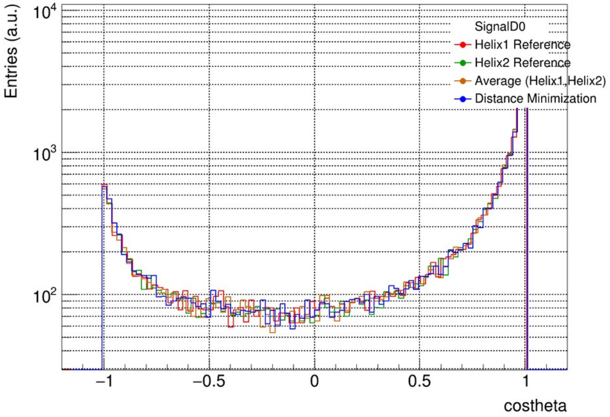
- Helix1 (using helix1 to find **pca₁** and **pca₂)**
- Distance minimization (d)
- Helix2 (using helix2 to find **pca₁** and **pca₂)**
- Using average of Helix1 and Helix2

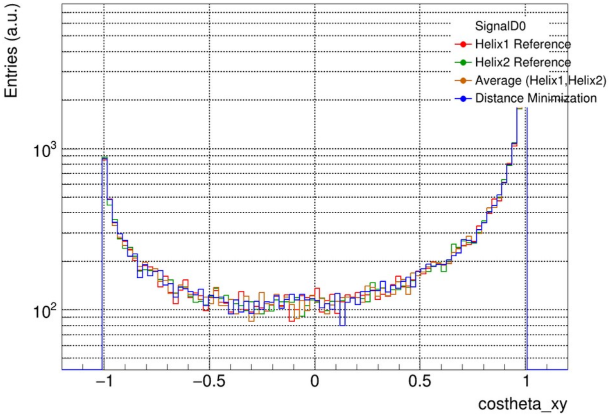
Comparison Distribution for D^0 meson (Q^2>1GeV^2)

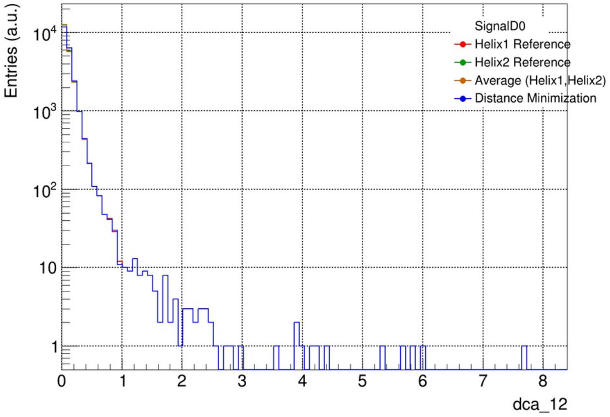
All methods are compatible

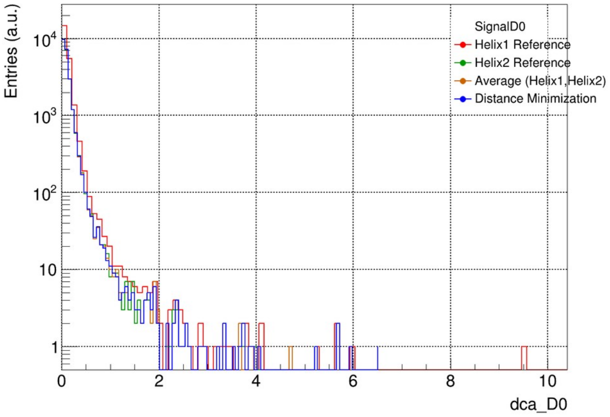
Signal D^0 meson

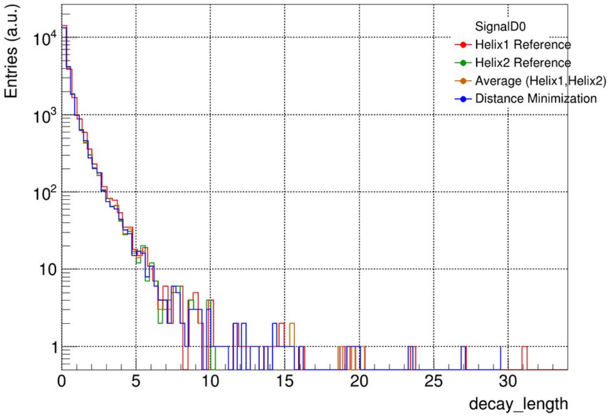
Distance minimization gives unique secondary vertex

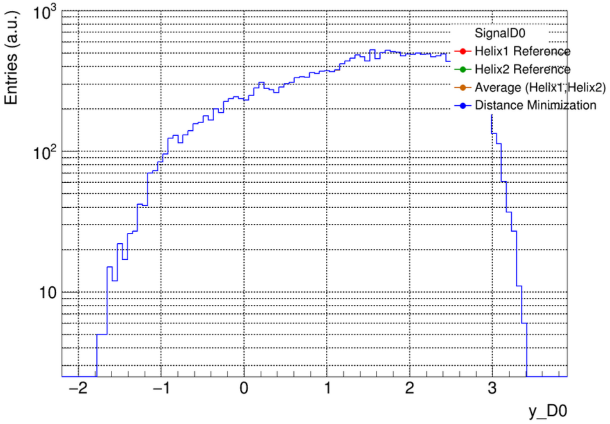








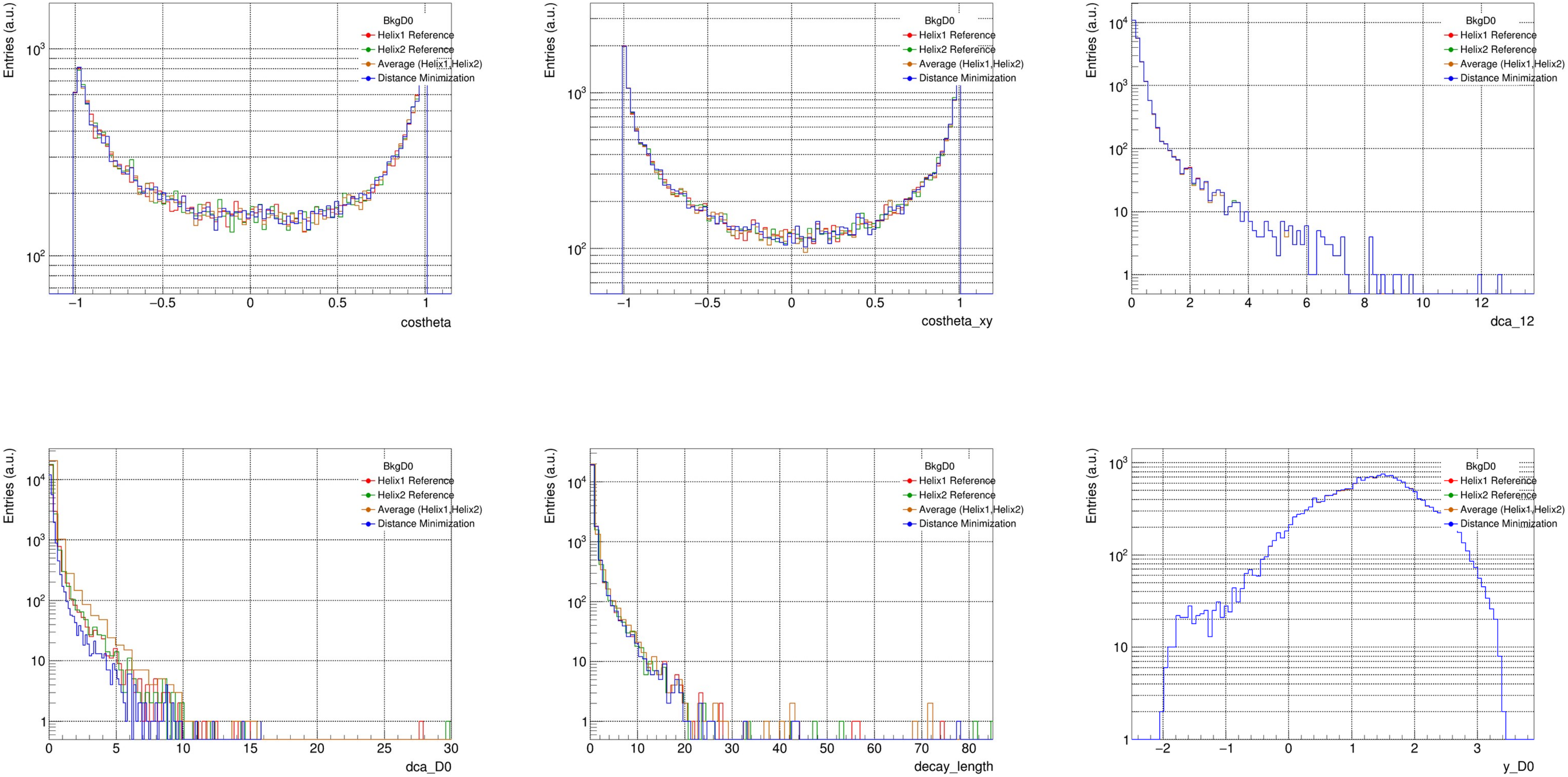




Comparison Distribution for D^0 meson (Q^2>1GeV^2)

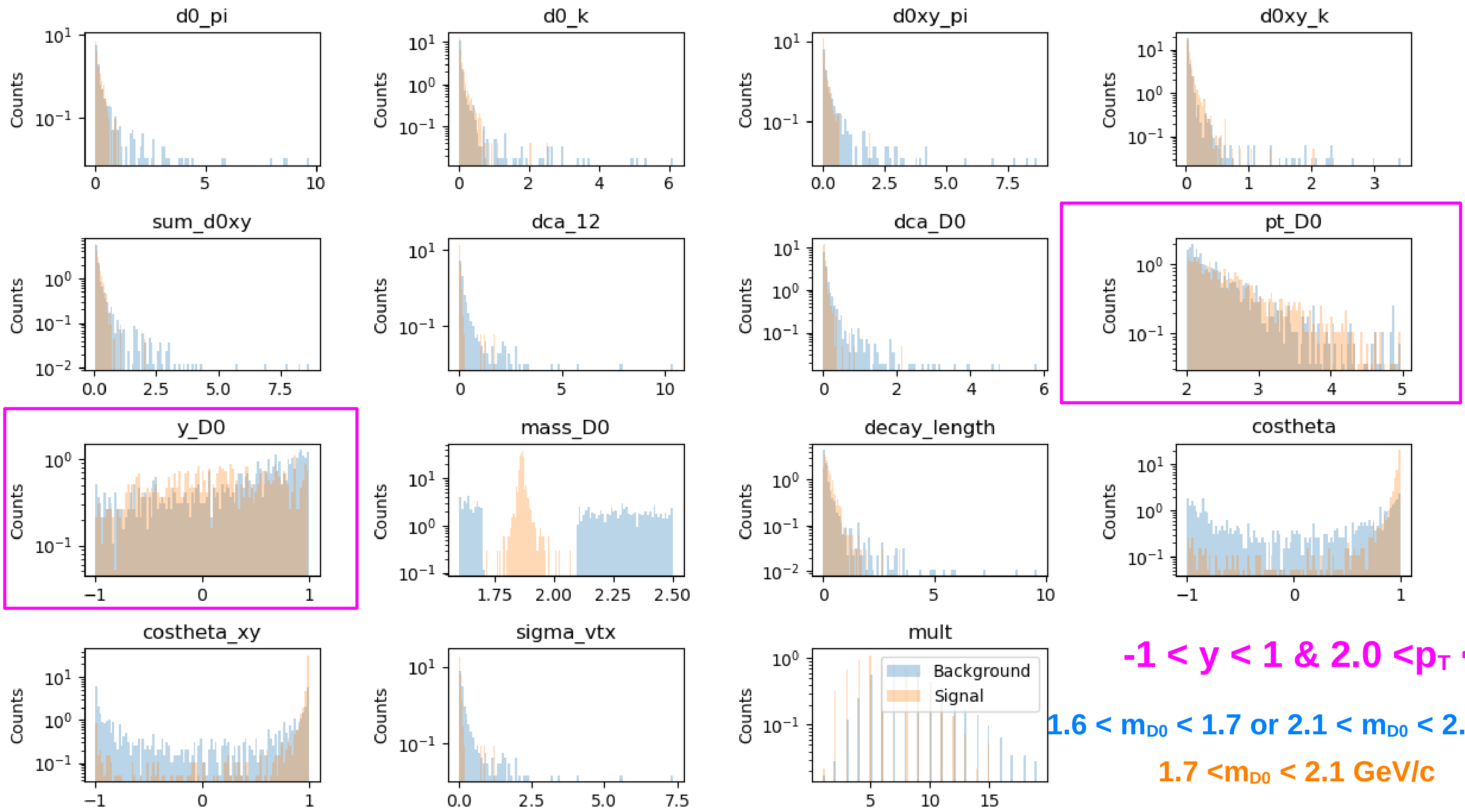
All methods are compatible

Bkg D^0 meson



Signal and Background distributions

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



$-1 < y < 1$ & $2.0 < p_T < 5.0$

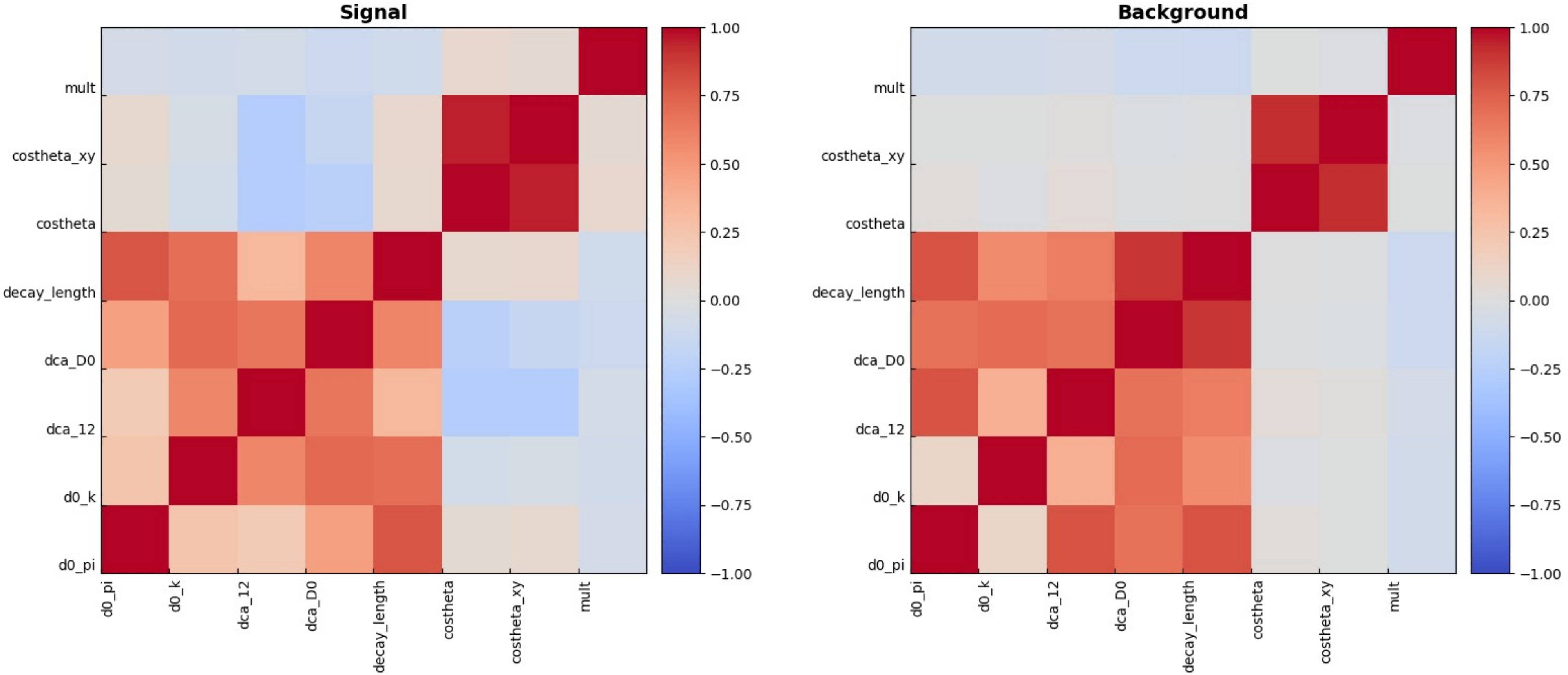
$1.6 < m_{D0} < 1.7$ or $2.1 < m_{D0} < 2.5$ GeV/c

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

Correlation Matrix

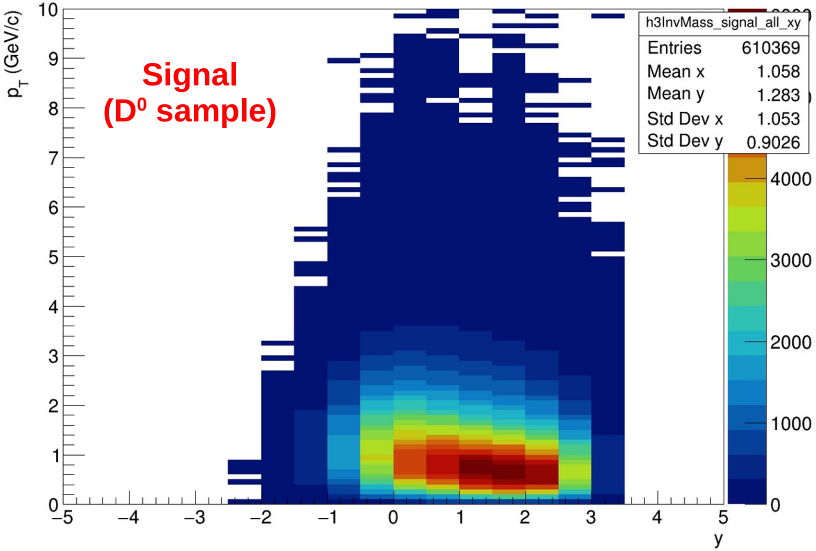
$-1.0 < y < 1.0$
$2.0 < p_T < 5.0$ GeV/c

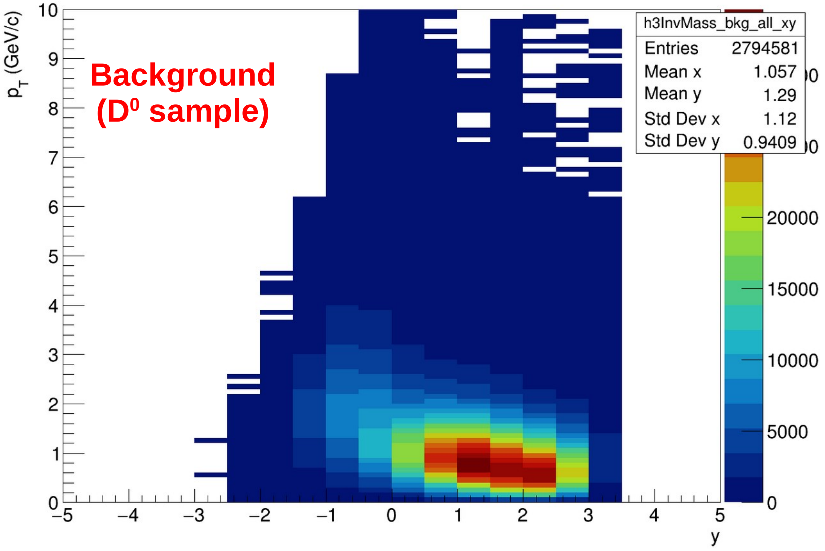
Planning to remove costheta_xy and decay_length once other cuts are available (e.g. chi2, nsigma, etc.)

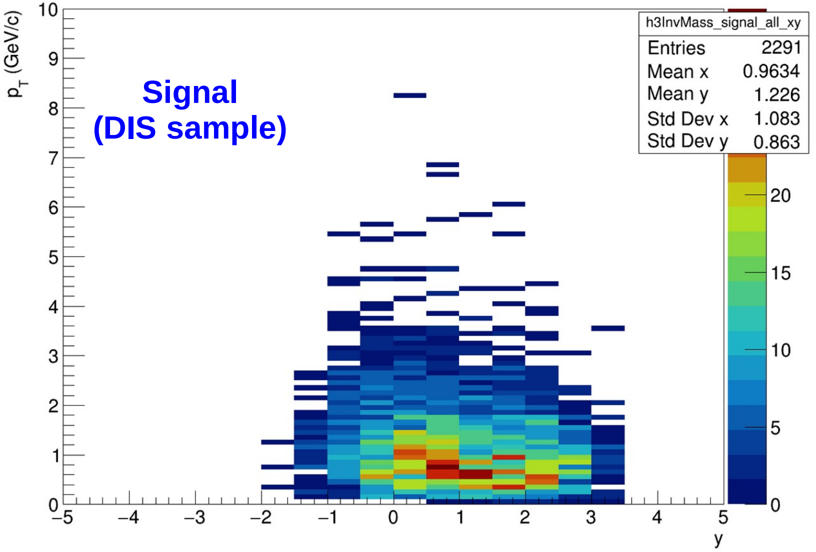


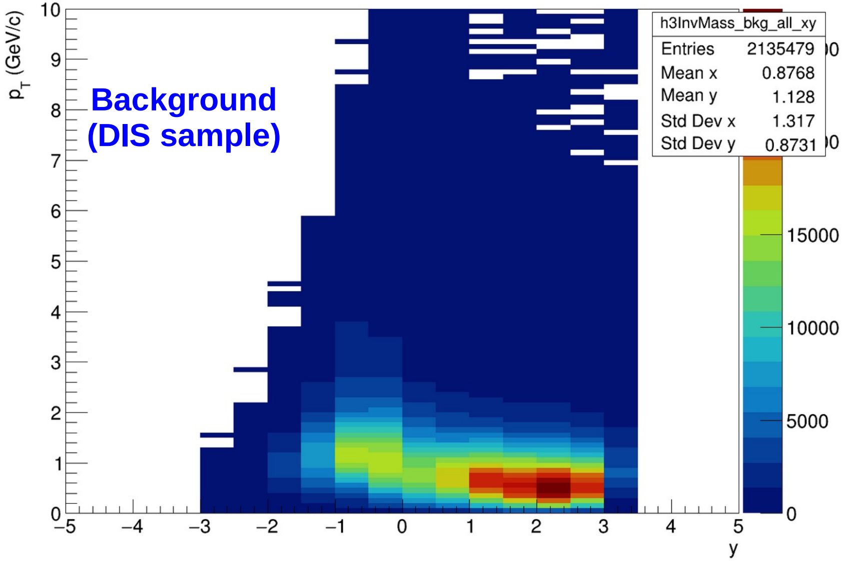
Performed a cross-check removing costheta_xy and decay_length (see backup)

Phase Space ($Q^2 > 1$ GeV^2)

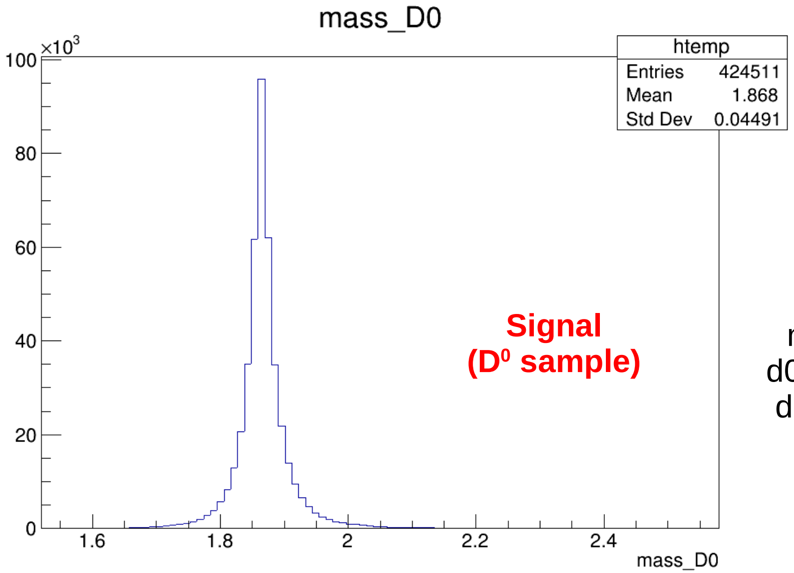






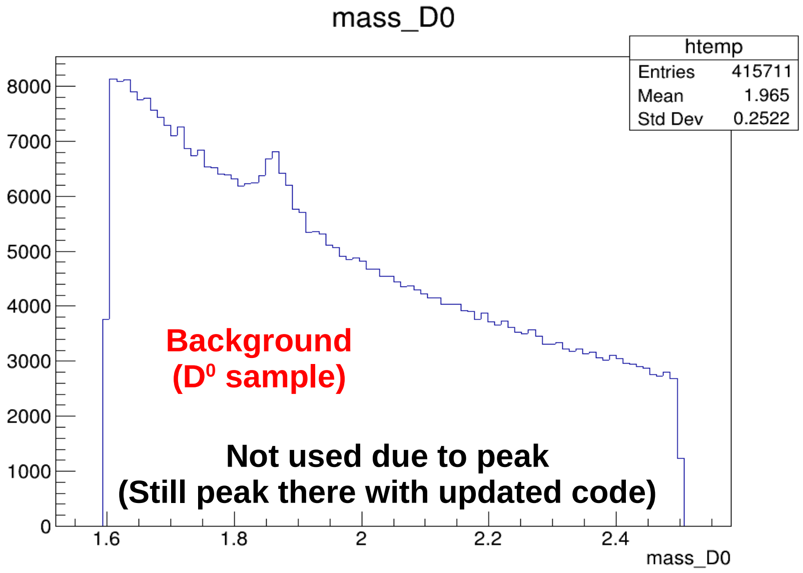


Sample After Preselection ($Q^2 > 1$ GeV^2)

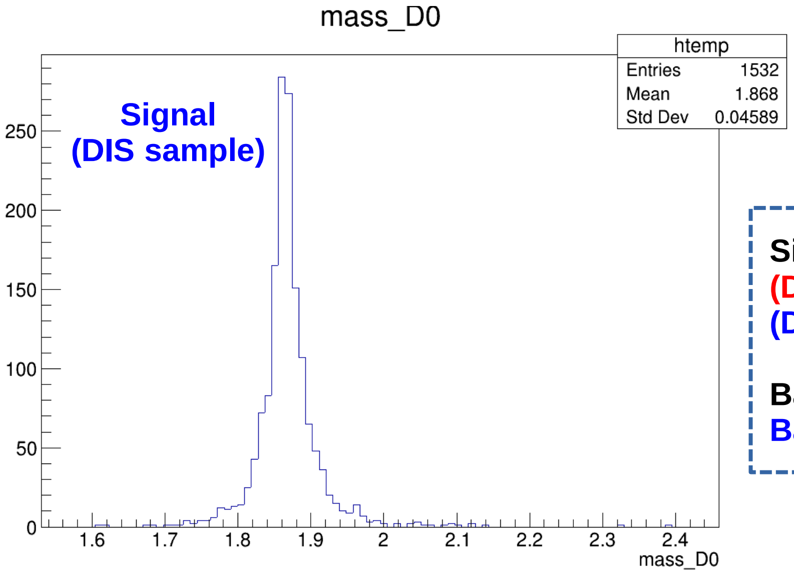


Preselection

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

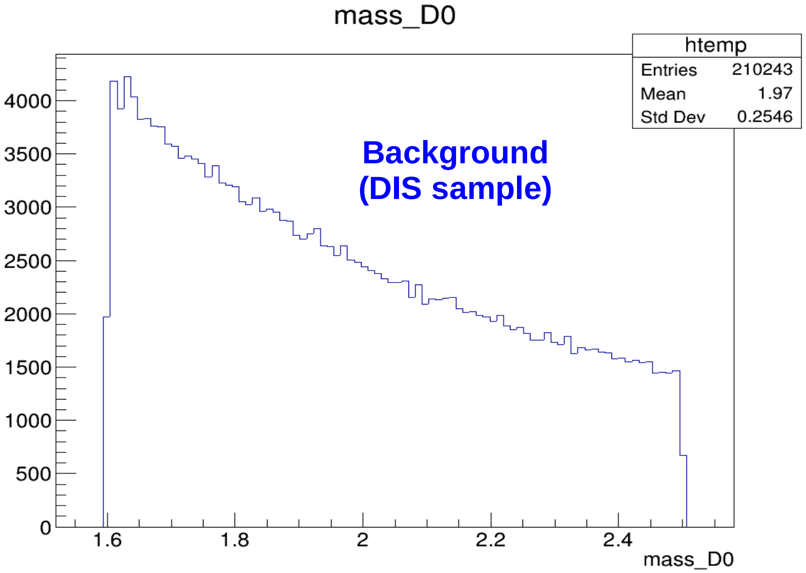


Not used due to peak (Still peak there with updated code)

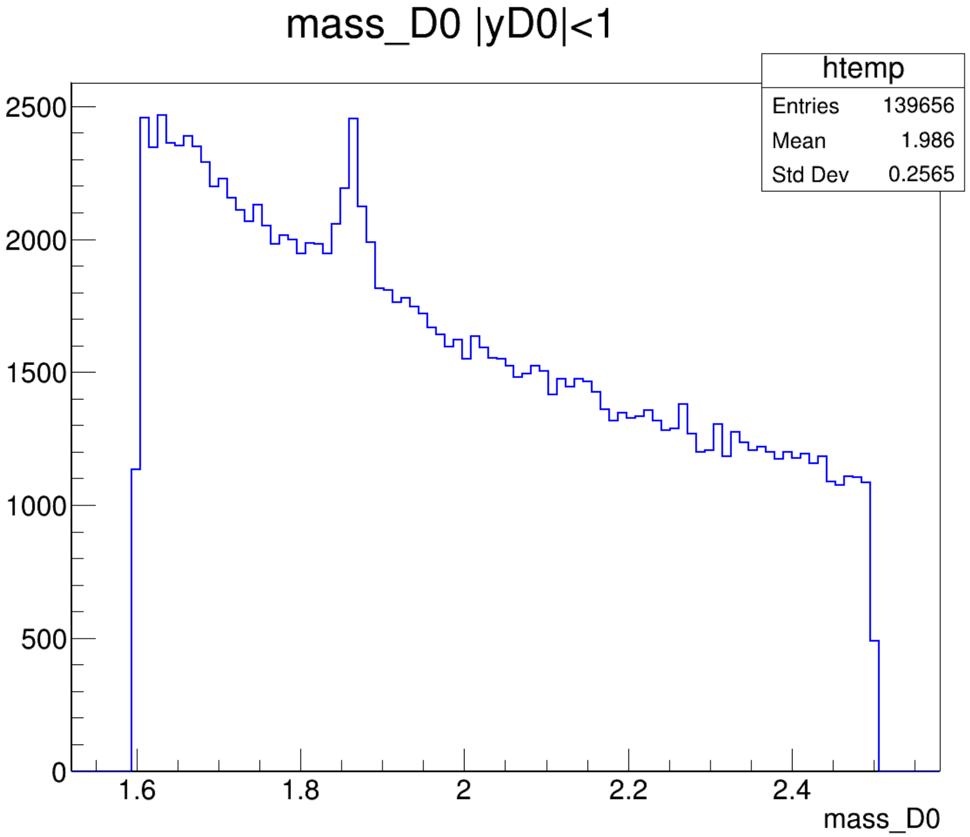


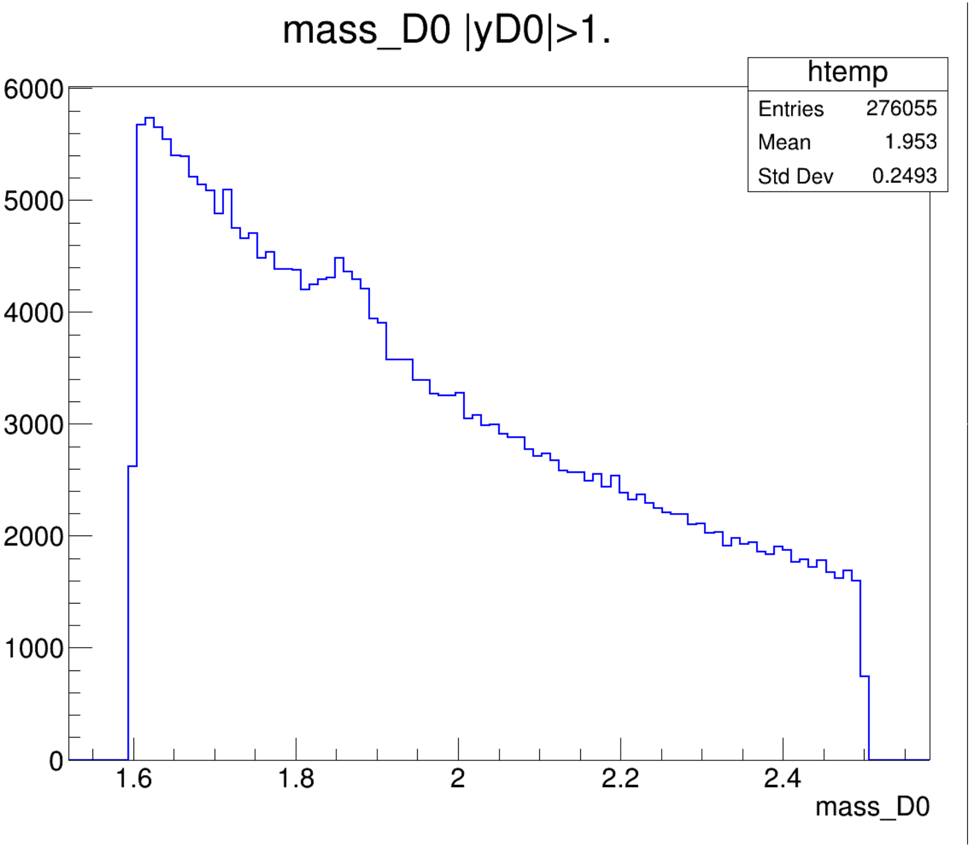
Signal for ML: Signal (D^0 sample) +Signal (DIS)

Background for ML: Background (DIS)



Background Pair in D^0 Sample





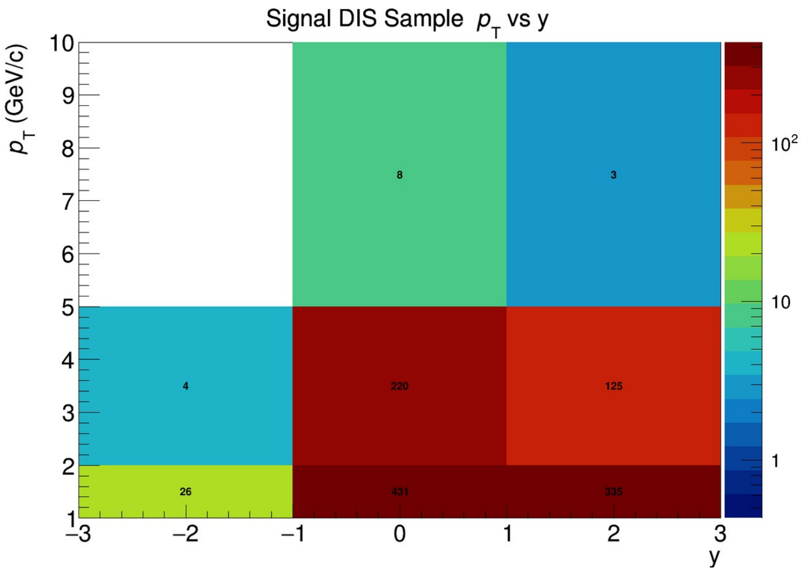
Number of Signal and Background

y(D0)	p_T(D0)	Signal	Background
-1.0 to 1.0	1.0-2.0	8211	8211
-1.0 to 1.0	2.0-10.0	993	993
1.0 to 3.0	1.0-2.0	17509	17509
1.0 to 3.0	2.0-10.0	2436	2436
-3.0 to -1.0	1.0-5.0	3228	3228

Keep the number of signal and background same for ML

There is also minor issue (std::map issue even in D^0) with associations if one reco track matches with two MC tracks, the code always considers last one, I can see print messages of two associations after changing a bit code

Method for Merging Signal and Background

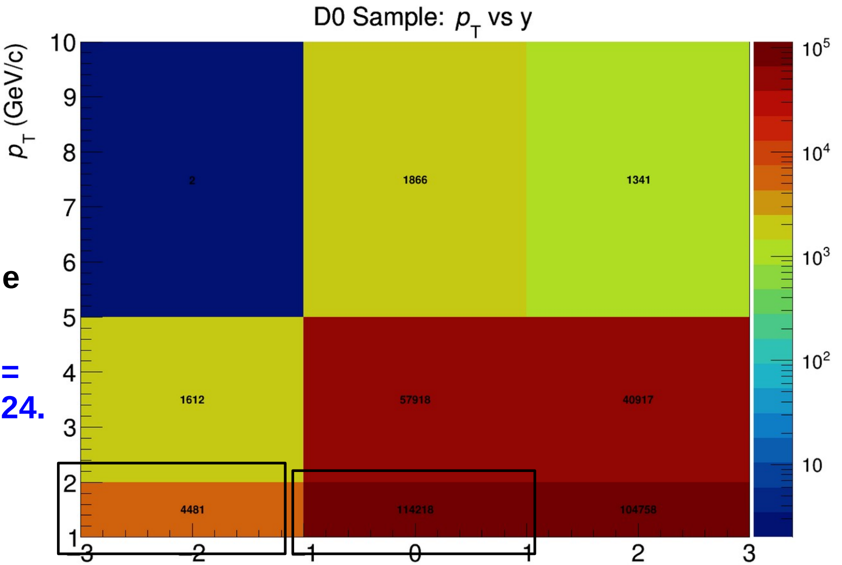

Signal DIS Sample p_T vs y

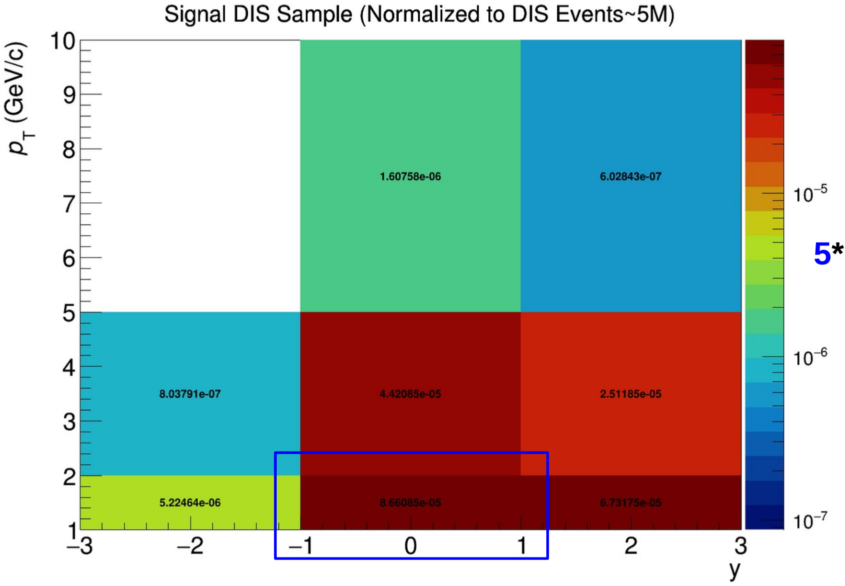
1. Reference Bin

(Max statistics)

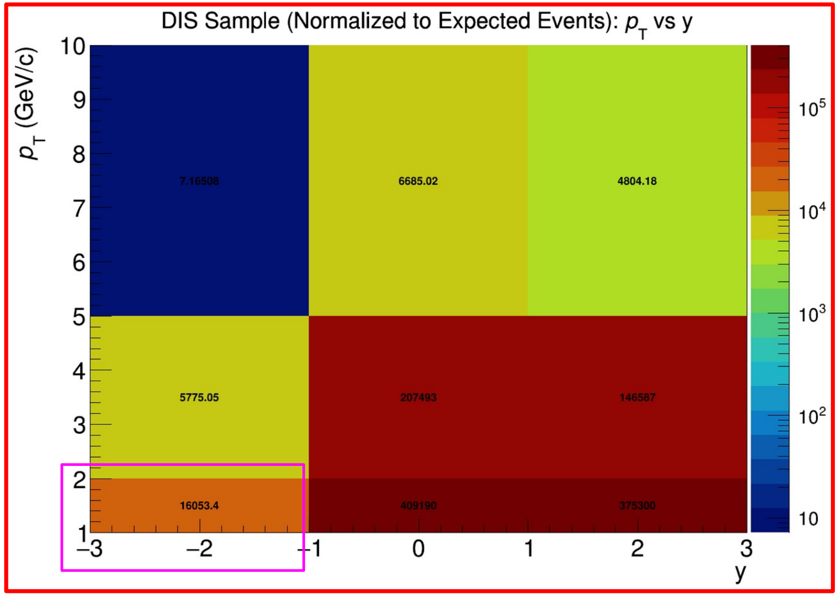
2. Fraction from D^0 sample (fraction = 4481./114218)

3. Expected D^0 (4724.6M)= ReferenceBin*fraction*4724.6M = 16053.4

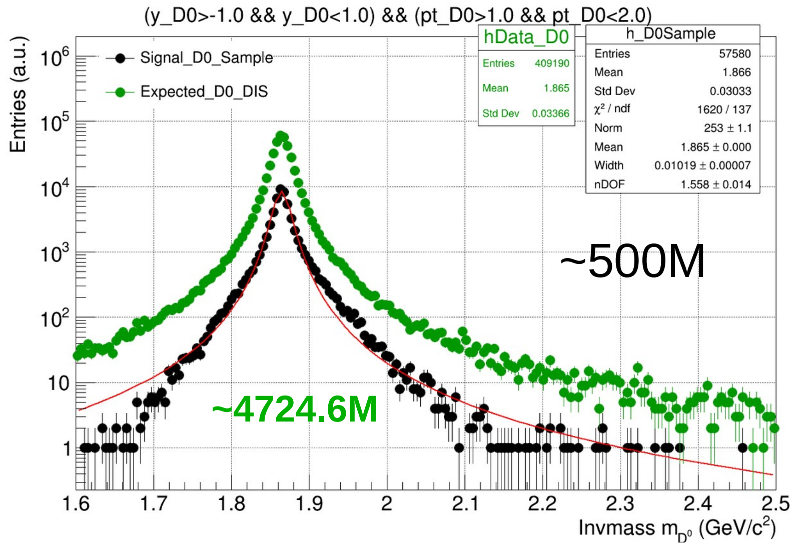

D0 Sample: p_T vs y

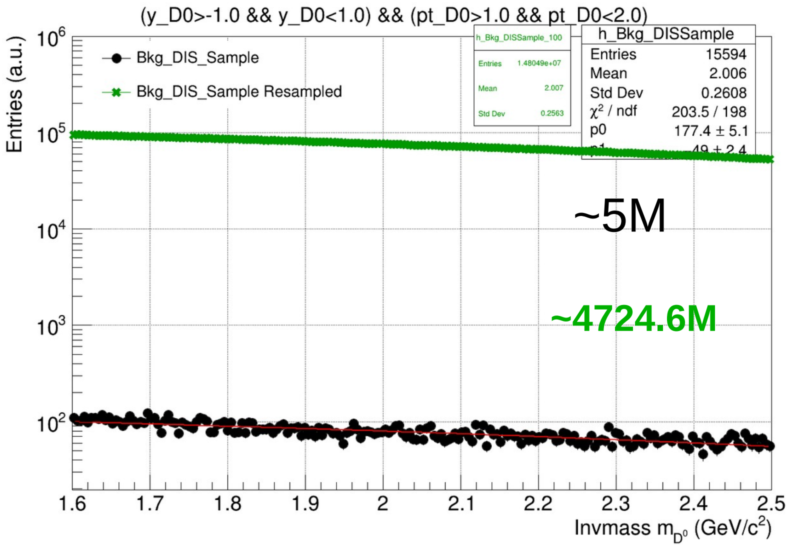

Signal DIS Sample (Normalized to DIS Events~5M)

8.66085e-5*(4481./114218)*4724.6e+6 ~ 16053.4

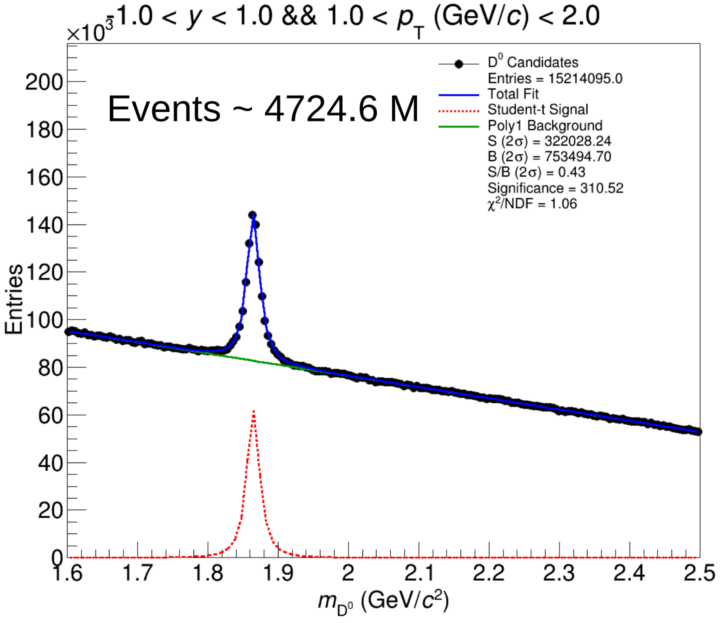

DIS Sample (Normalized to Expected Events): p_T vs y

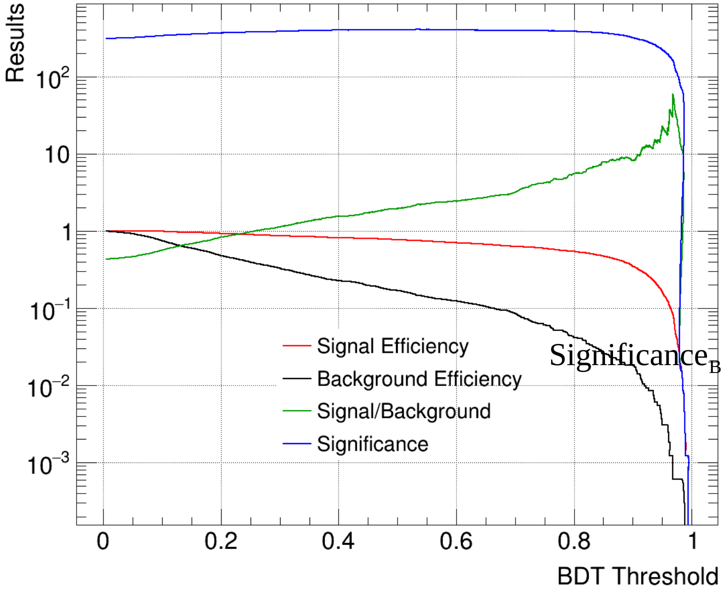
Merging Signal and Background (D0 Sample and DIS Events)





Resampling

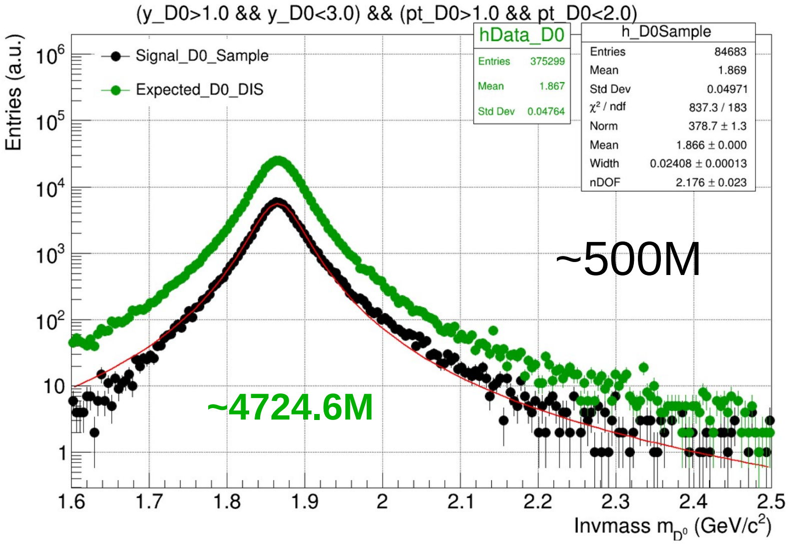


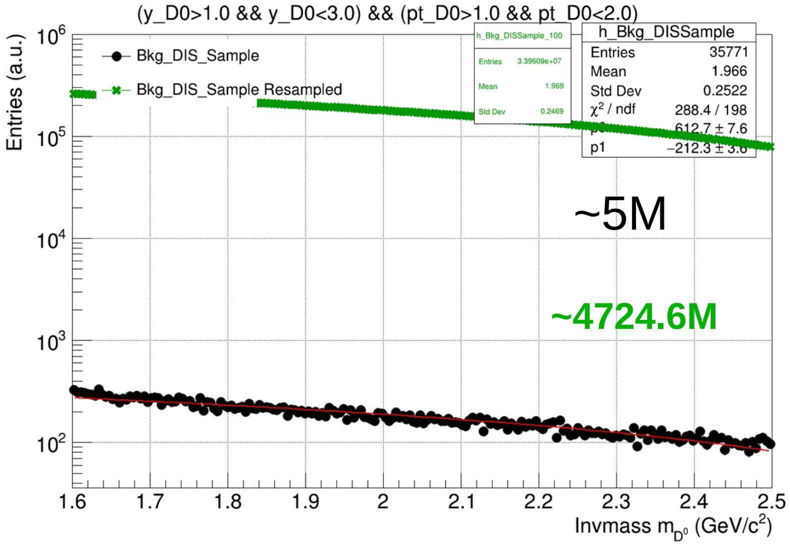


$$\left(\frac{S}{B}\right)_{\text{BDT Cut}} = \left(\frac{S}{B}\right)_{\text{No ML}} \times \frac{\epsilon_{\text{Signal}}}{\epsilon_{\text{Background}}}$$

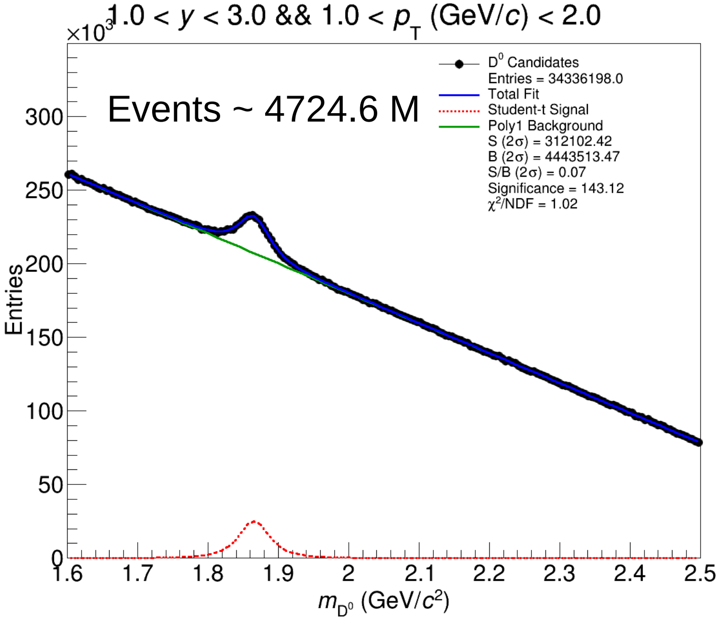
$$\text{Significance}_{\text{BDT Cut}} = \frac{S_{\text{No ML}} \times \epsilon_{\text{Signal}}}{\sqrt{S_{\text{No ML}} \times \epsilon_{\text{Signal}} + B_{\text{No ML}} \times \epsilon_{\text{Background}}}}$$

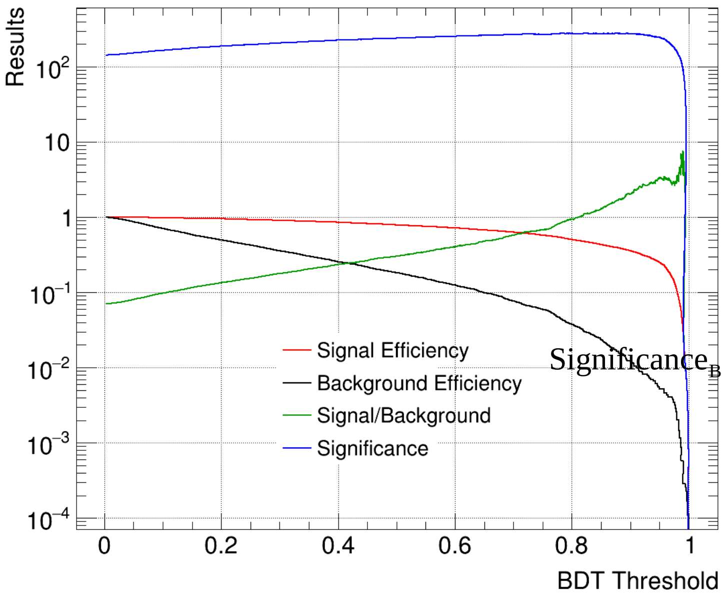
Merging Signal and Background (D^0 Sample and DIS Events)





Resampling

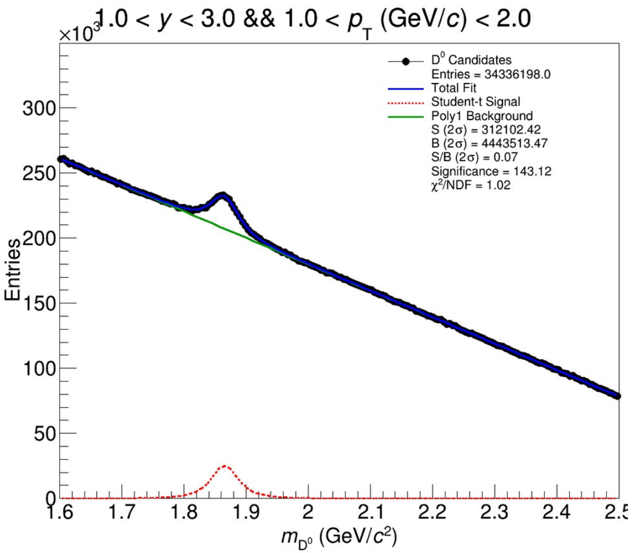


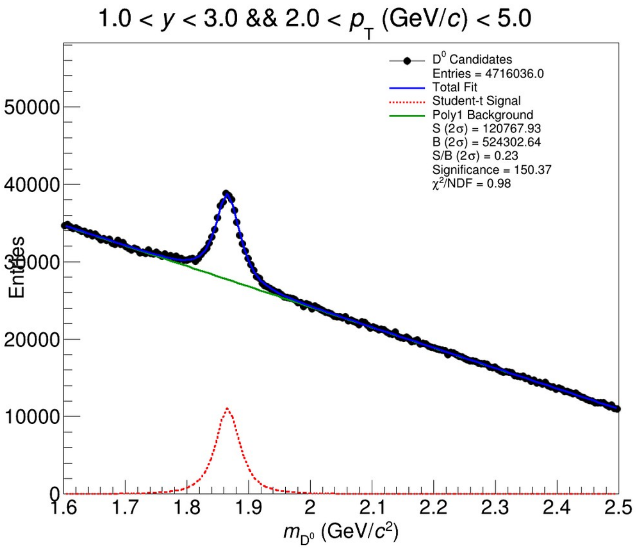


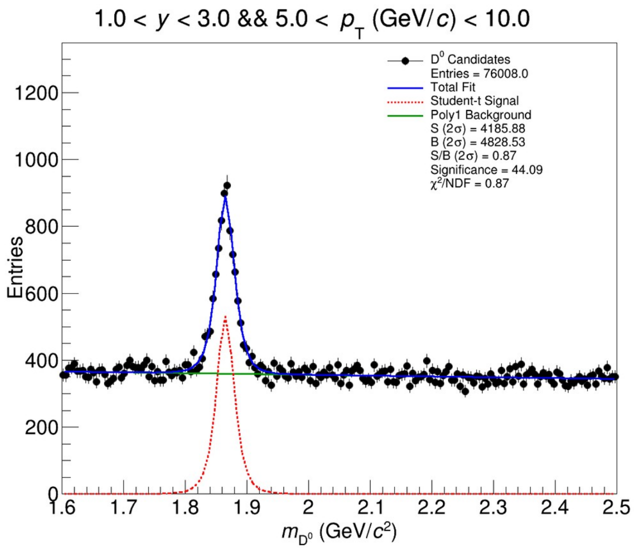
$$\left(\frac{S}{B}\right)_{\text{BDT Cut}} = \left(\frac{S}{B}\right)_{\text{No ML}} \times \frac{\epsilon_{\text{Signal}}}{\epsilon_{\text{Background}}}$$

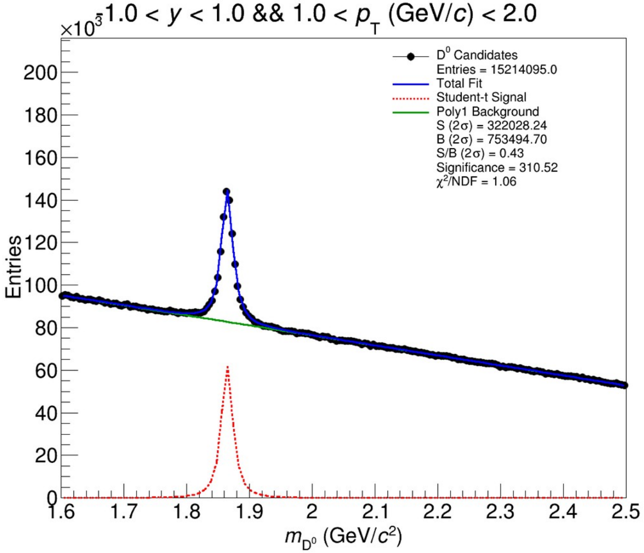
$$\text{Significance}_{\text{BDT Cut}} = \frac{S_{\text{No ML}} \times \epsilon_{\text{Signal}}}{\sqrt{S_{\text{No ML}} \times \epsilon_{\text{Signal}} + B_{\text{No ML}} \times \epsilon_{\text{Background}}}}$$

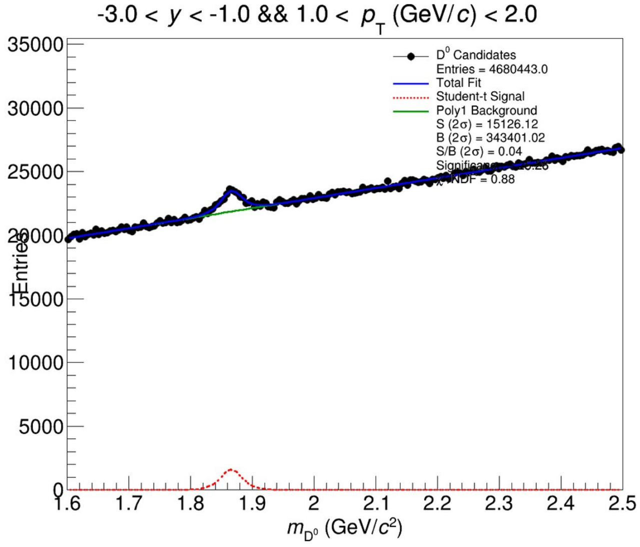
Merging Signal and Background

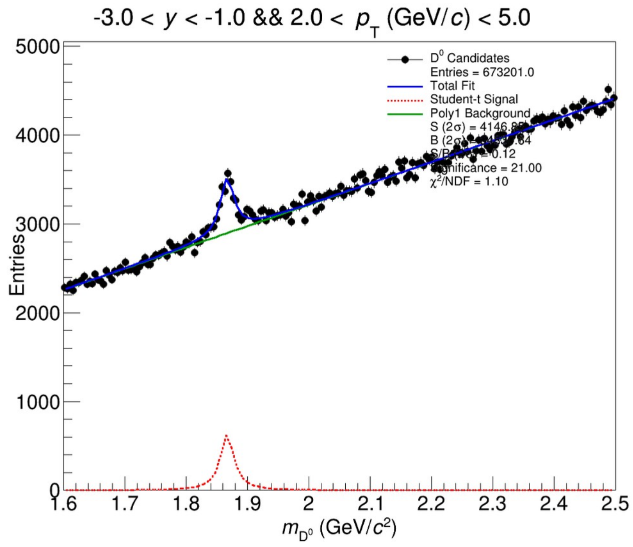












Secondary Vertex Reconstruction (Λ_c^+)

Secondary Vertex

$$\Lambda_c^+ \rightarrow p\,K^-\,\pi^+$$

$DCA_{k\pi} = |\vec{pca}_1 - \vec{pca}_2|, \quad DCA_{kp} = |\vec{pca}_1 - \vec{pca}_3|, \quad DCA_{p\pi} = |\vec{pca}_3 - \vec{pca}_2|$

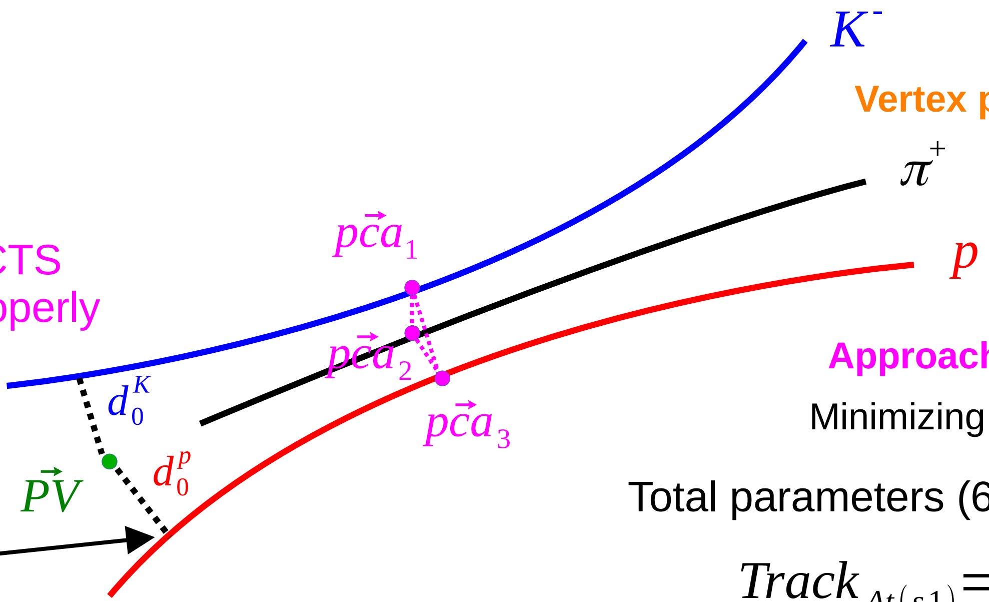
DCA_{12} = min {$DCA_{\pi K}$, DCA_{KP}, $DCA_{\pi P}$} Cut

Approach 1 (Shyam)

$$\vec{SV} = \frac{\vec{pca}_1 + \vec{pca}_2 + \vec{pca}_3}{3}$$

Secondary vertexing in ACTS considers tracking errors properly

Ignored track errors (at the moment)



$$Track_{DCA} = (\vec{r}, \vec{p}, q)$$

Vertex position $\vec{SV}$ = (v_x, v_y, v_z)

$$Track_{At(s)} = (\vec{r}_s, \vec{p}, q)$$

s: path length

Total parameters (6) = (v_x, v_y, v_z, s_1, s_2, s_3)

Approach 2 (Shyam)

Minimizing the distance

Total parameters (6) = (v_x, v_y, v_z, s_1, s_2, s_3)

$$Track_{At(s1)} = (\vec{r}_{s1}, \vec{p}1, q1)$$

$$Track_{At(s2)} = (\vec{r}_{s2}, \vec{p}2, q2)$$

$$Track_{At(s3)} = (\vec{r}_{s3}, \vec{p}3, q3)$$

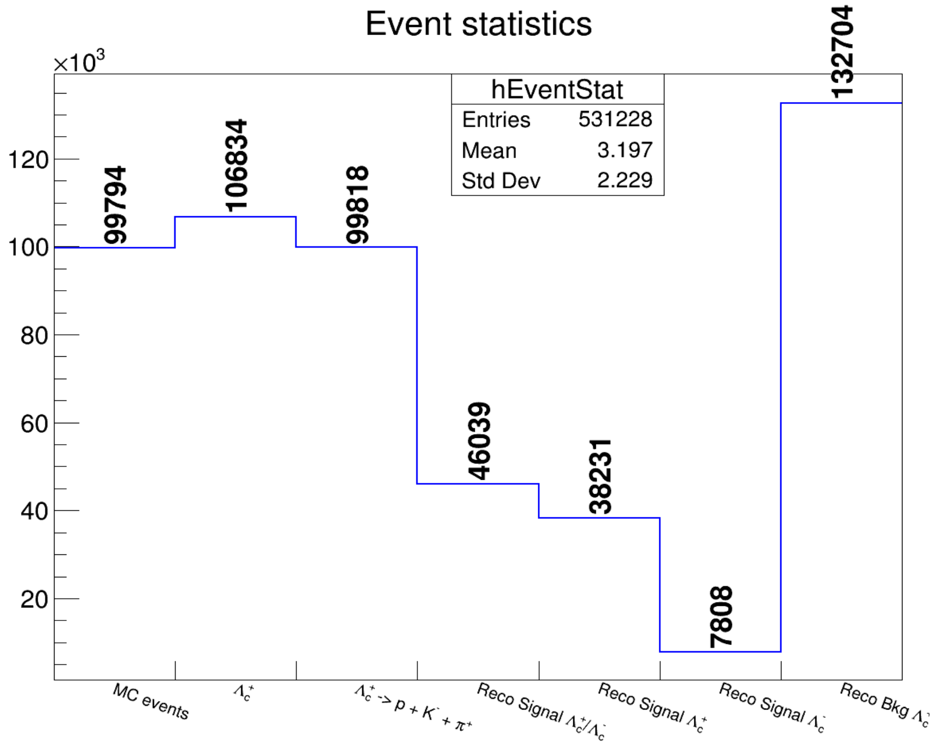
Minimize $d = \sqrt{(\vec{r}_{s1} - \vec{v})^2 + (\vec{r}_{s2} - \vec{v})^2 + (\vec{r}_{s3} - \vec{v})^2}$

Λ_c^+ Reconstruction

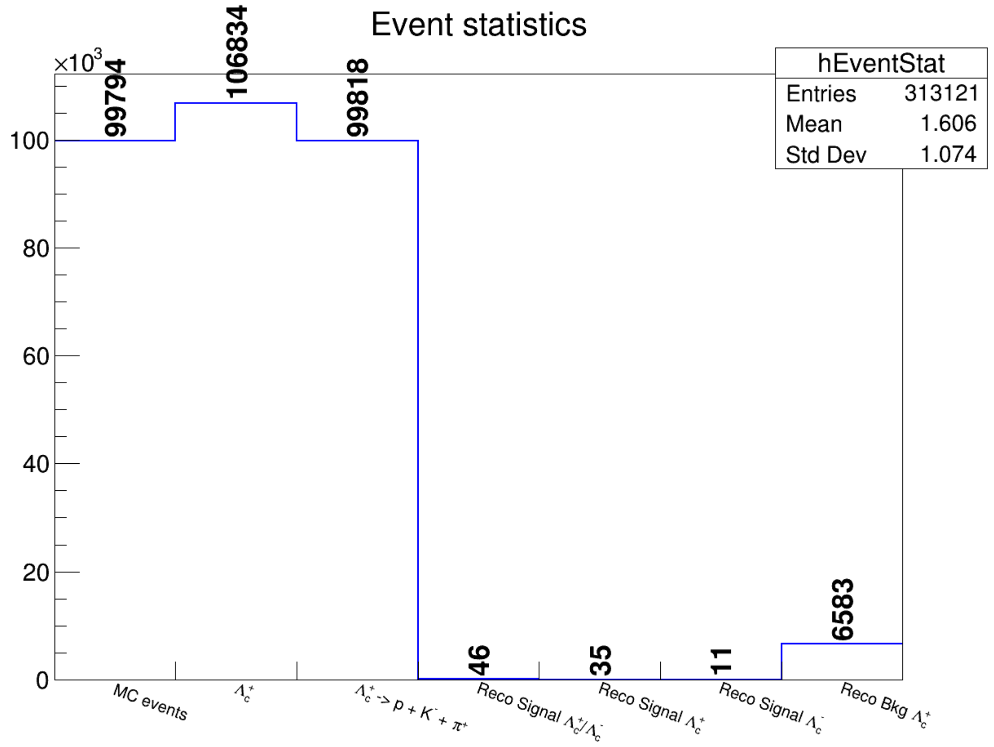
PYTHIA8 ep NC (10x100 Q^2>1) Λ_c^+ sample: by Rongrong few files for tesing

$\Lambda_c^+ \rightarrow p\, K^- \pi^+$

Truth PID

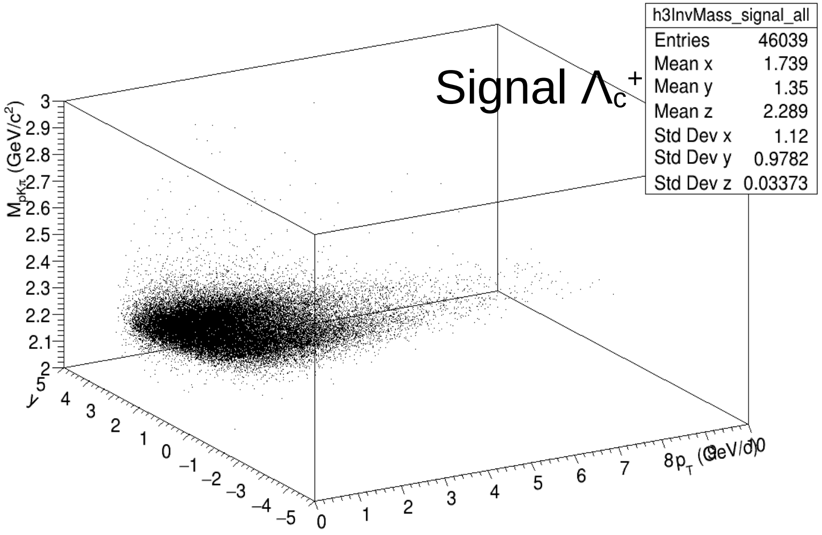


Real PID



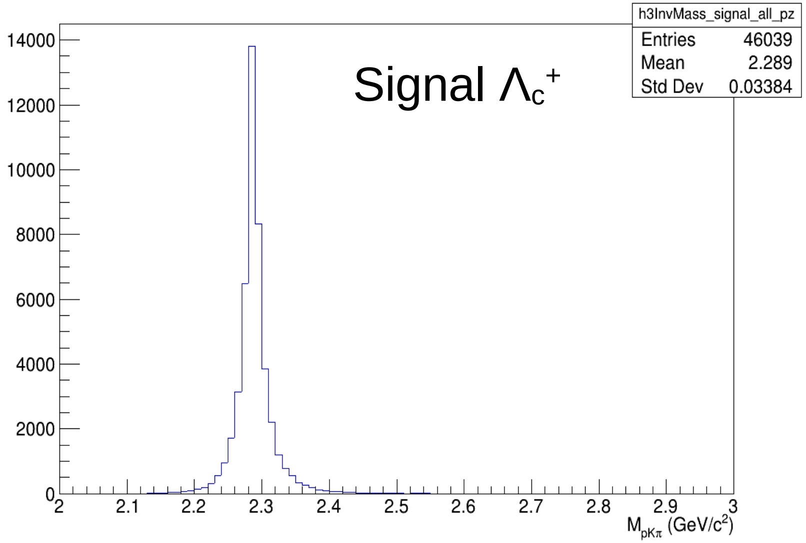
Issue coming from proton PID, it looks not properly assigned in reconstruction (loosing protons)

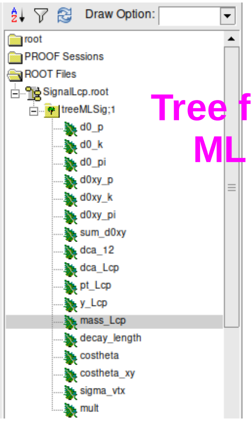
Results



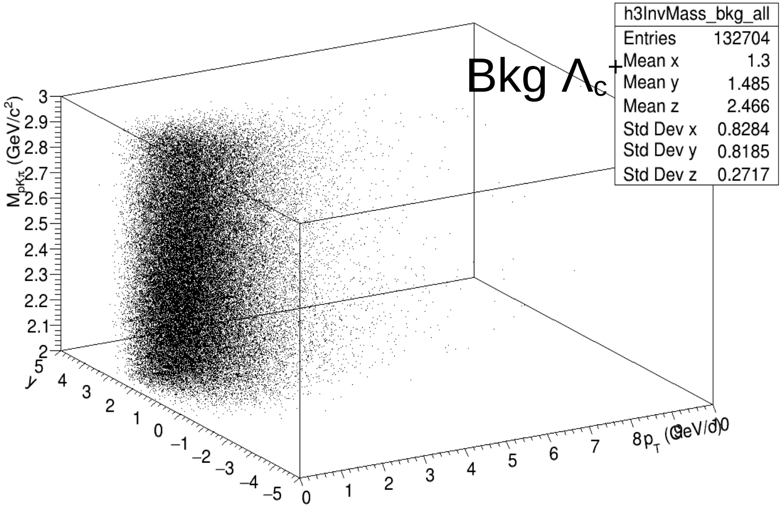
Invariant mass of unlike-sign πK pairs

Projection

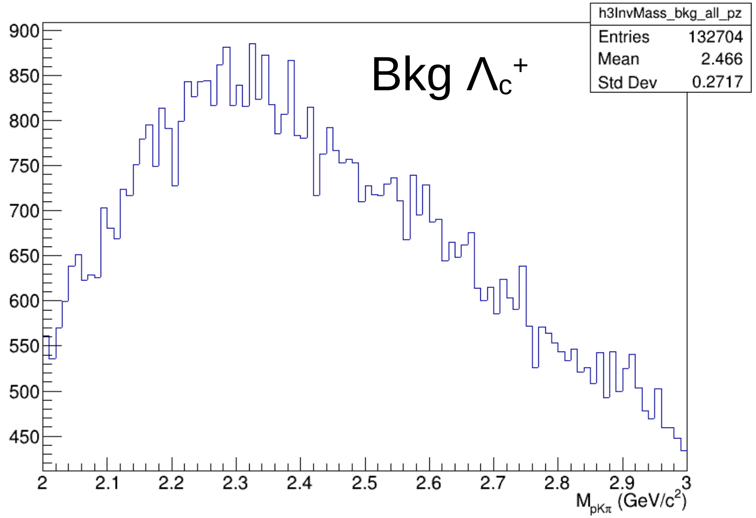




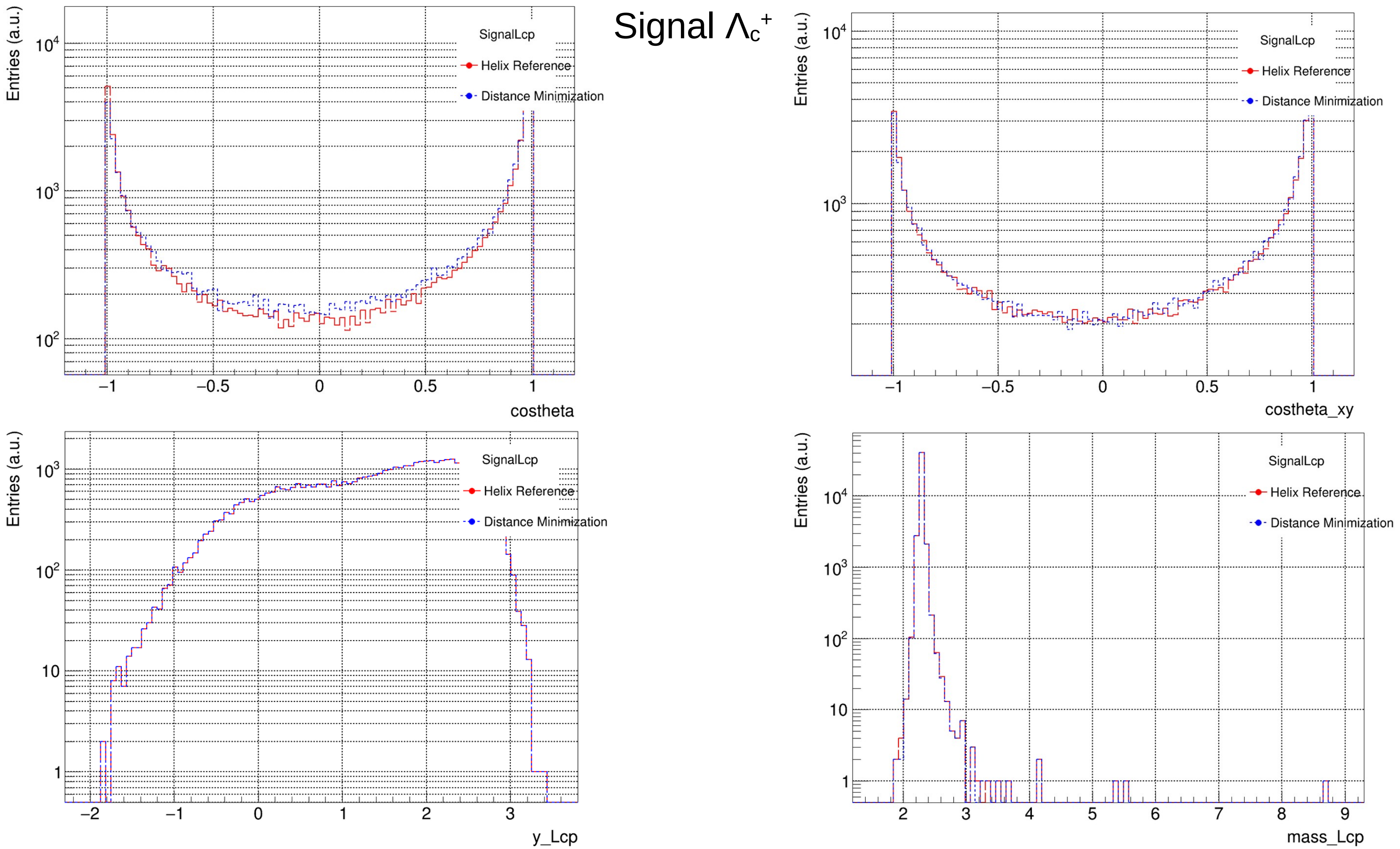
Tree for ML



Projection



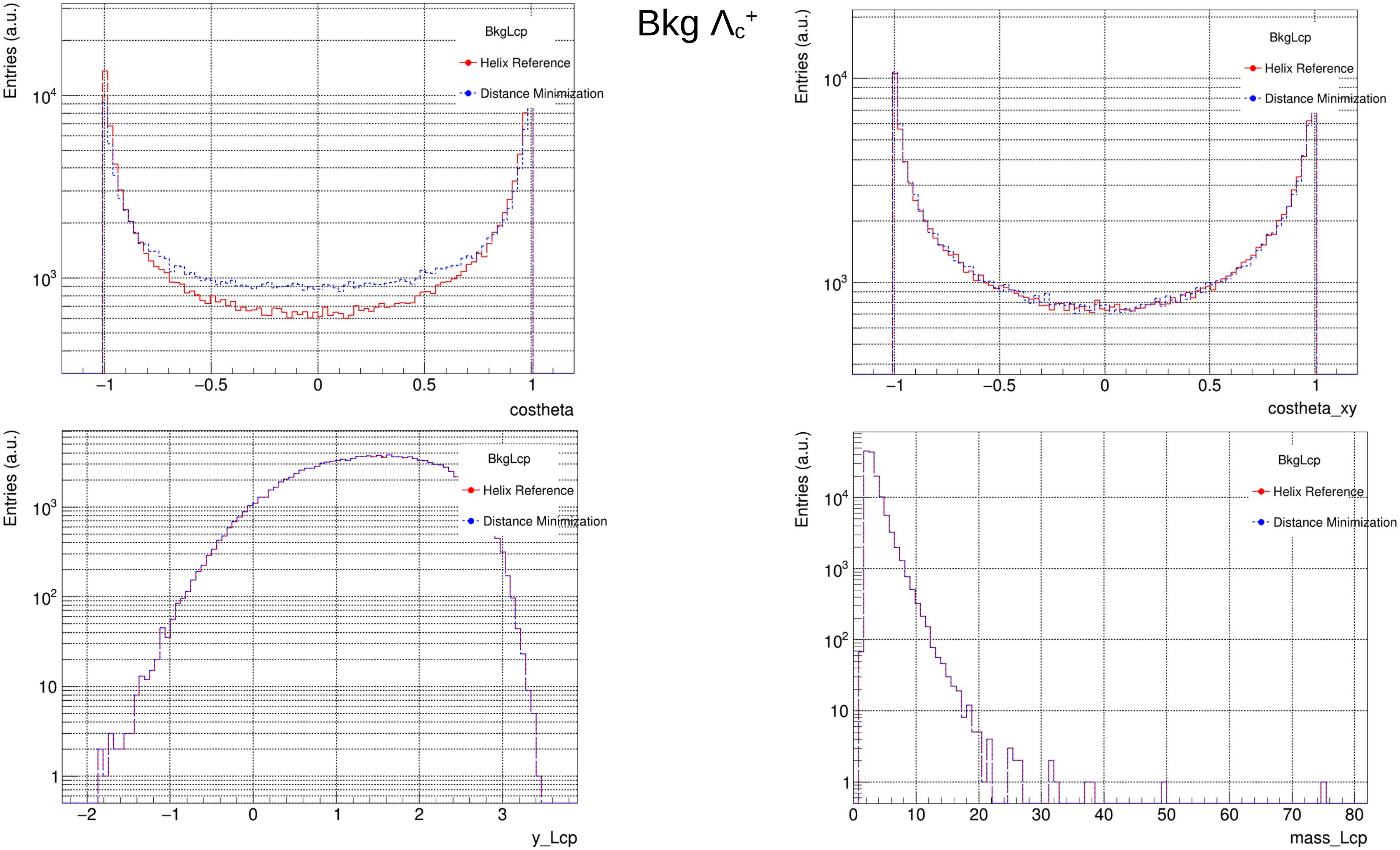
Comparison (Results)



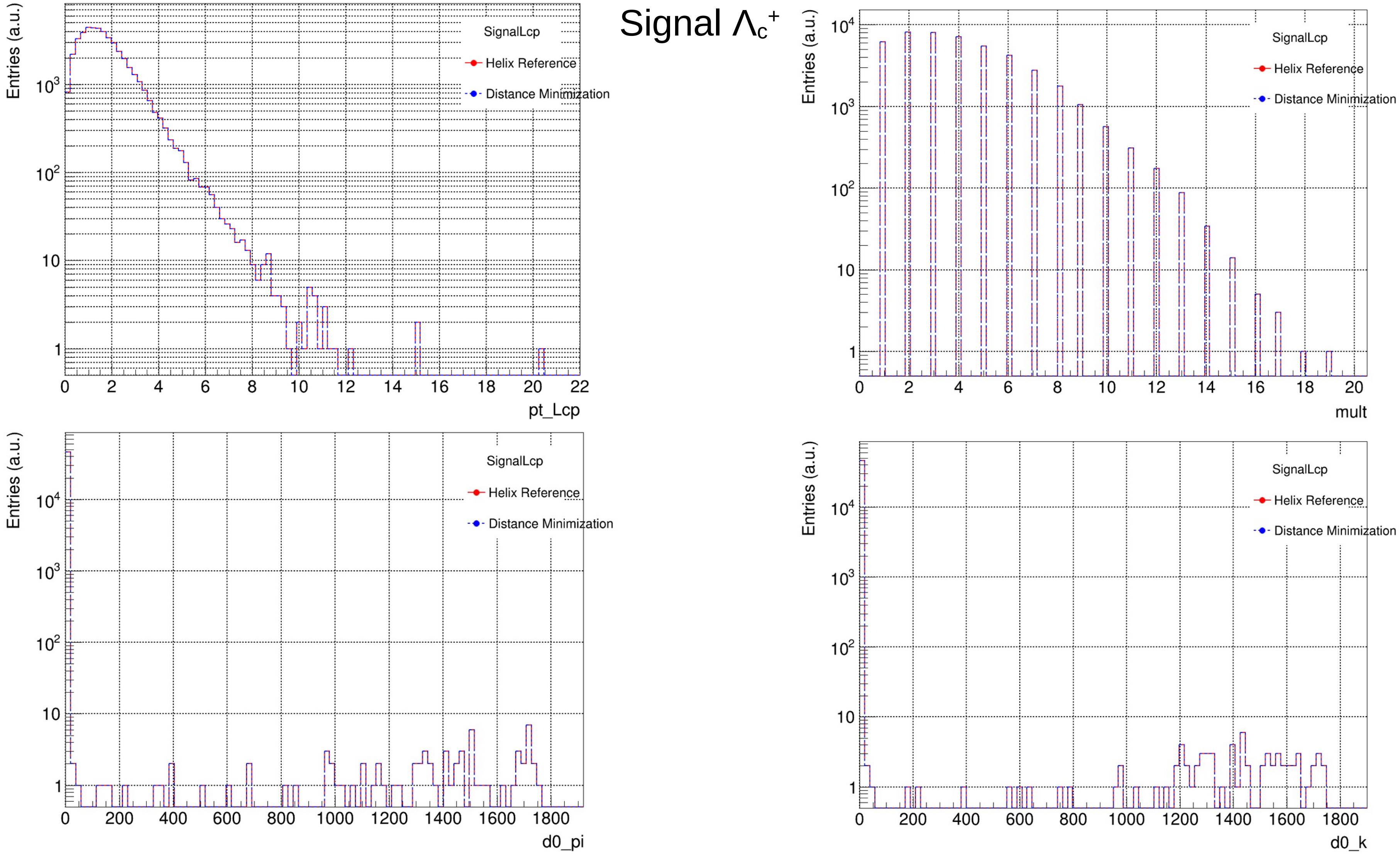
Signal Λ_c^+

Comparison (Results)

Bkg Λ_c^+



Comparison (Results)



Signal Λ_c^+

Summary and Future Plan

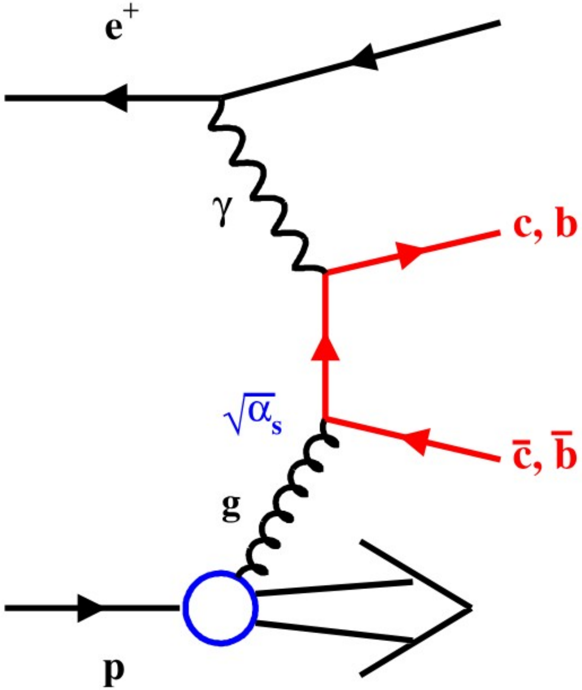
- Machine learning model studies performed for the D^0 reconstruction
- Implemented the first version of Λ_c^+ reconstruction code (will commit soon)
- Future Steps:
  - Implement secondary vertexing to improve the performances
  - Include chi2 of secondary vertex as one of the features once available
  - Extract the final results in different y and p_T bins after secondary vertexing
  - Evaluate the efficiency of D-meson and Λ_c+ baryon using preselection efficiency and BDT cut efficiency
  - Fix the reconstruction for the realistic PID of proton
  - Implement similar ML model for Λ_c^+ reconstruction (quicker)
  - Run on full stats once campaign files are available
  - Estimate Λ_c^+/D^0 ratio using machine learning
  - Implement other models e.g. Neural Network (Classifier as well as AutoEncoder)

THANK YOU !!!

Heavy-flavor Production

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

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



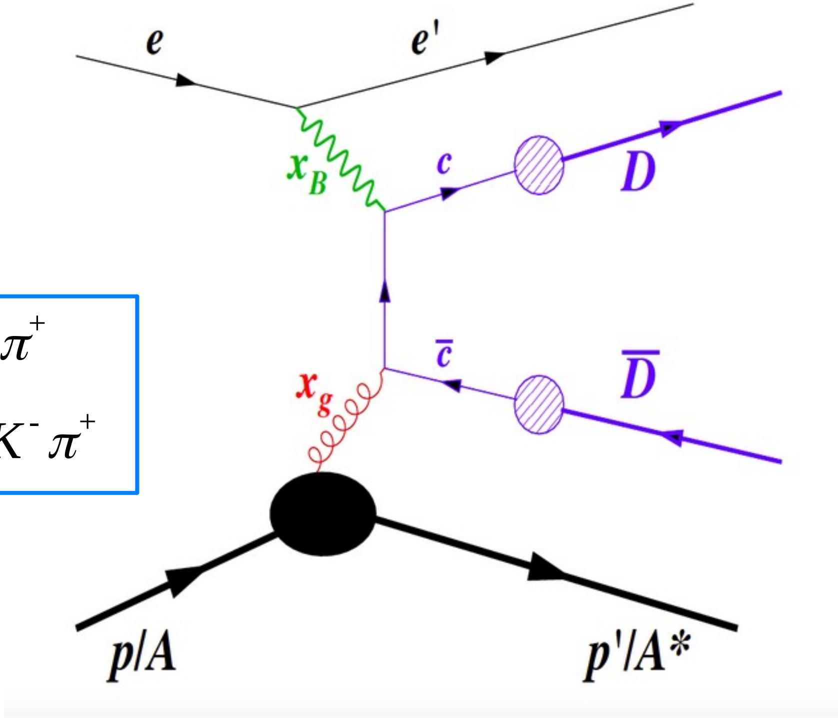
$$\gamma^* g \rightarrow c\,\bar{c}\,\text{or}\,b\,\bar{b}$$

$$c \rightarrow D^0(c\,\bar{u}) \rightarrow K^- \pi^+$$

$$c \rightarrow \Lambda_c^+(udc) \rightarrow p\,K^- \pi^+$$

m_{D0} = 1.86484 GeV/c^2

$m_{\Lambda c+}$ = 2.28646 GeV/c^2



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

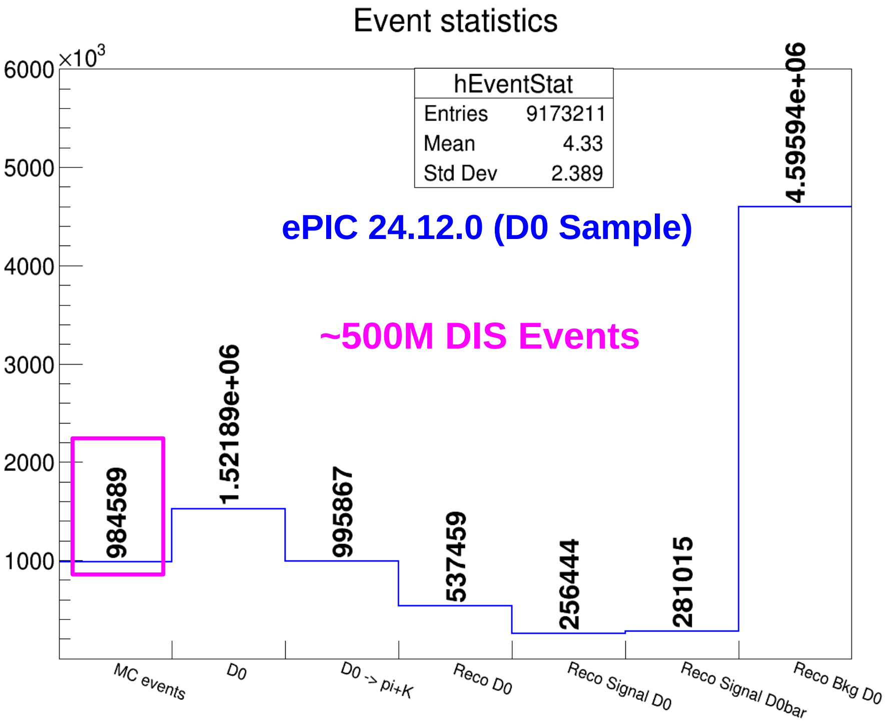
Additional NLO Mechanisms: Gluon splitting, QCD Compton Scattering

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

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

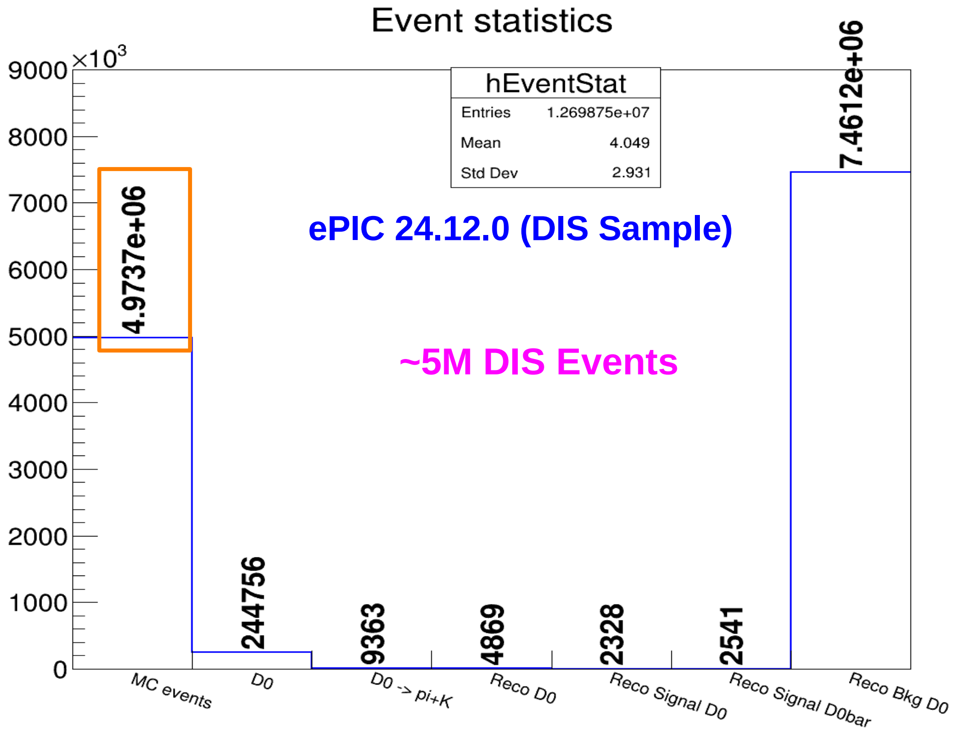
Simulation of D0 and Lc samples

- 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**



ePIC 24.12.0 (D0 Sample)

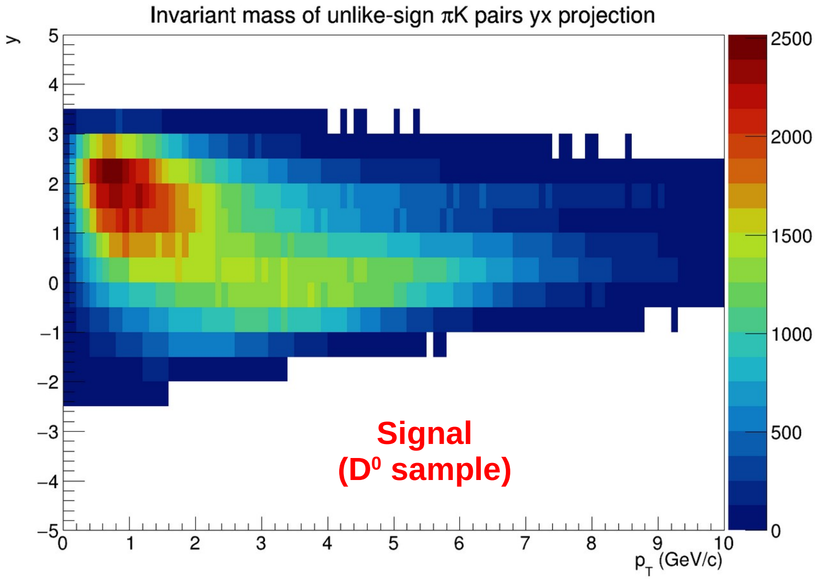
~500M DIS Events

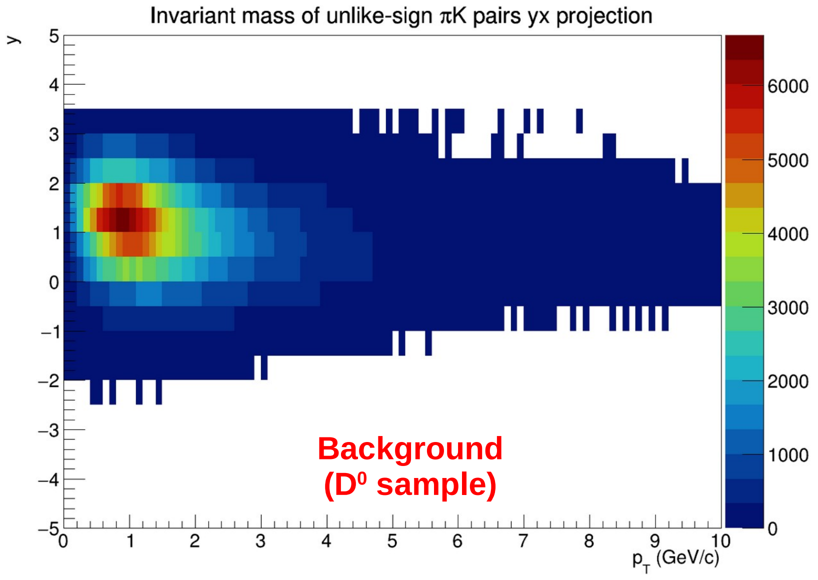


ePIC 24.12.0 (DIS Sample)

~5M DIS Events

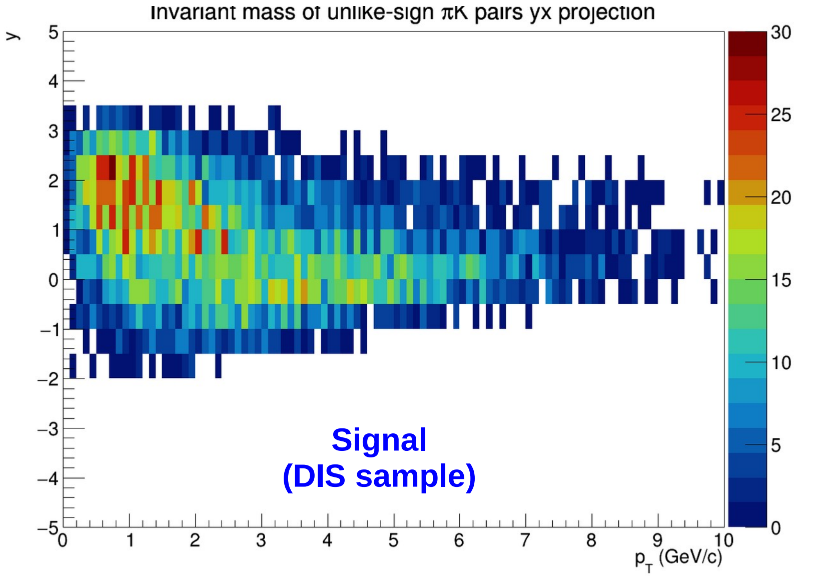
Phase Space (Q^2 = 100 GeV^2)

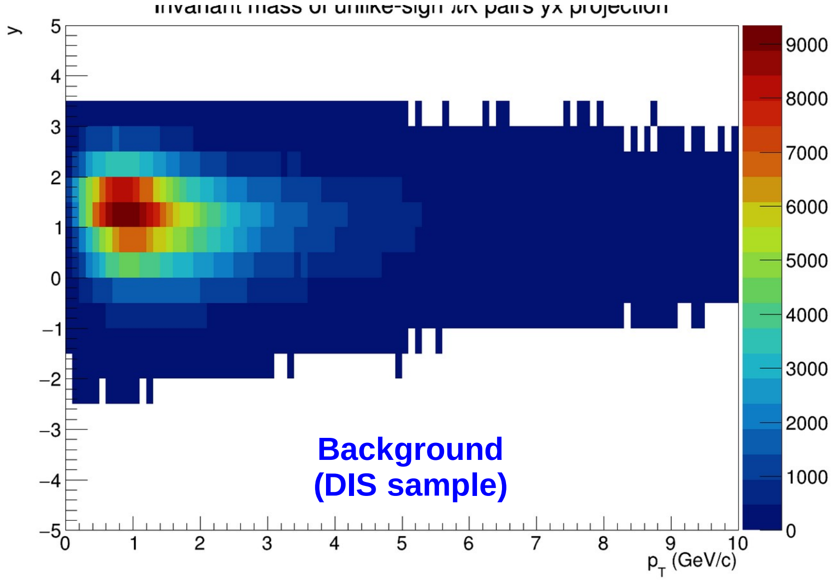




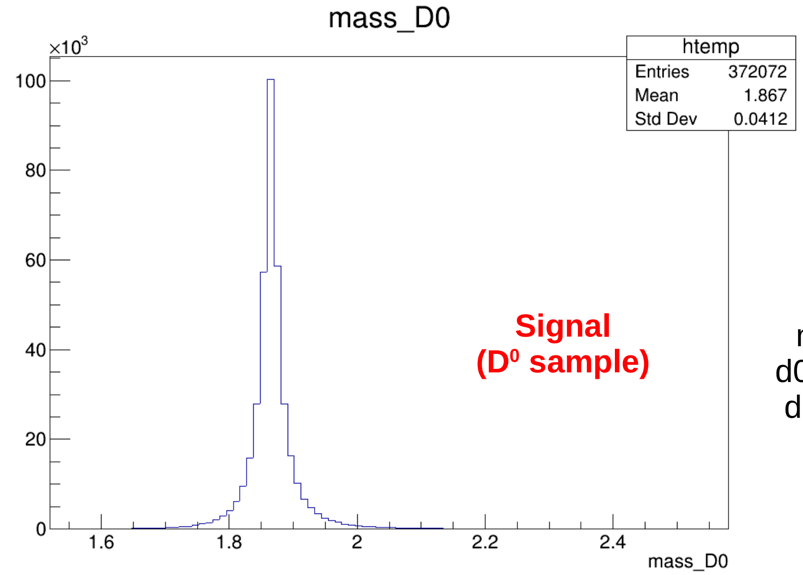
-3.0 <y < -1.0

Lowest stats





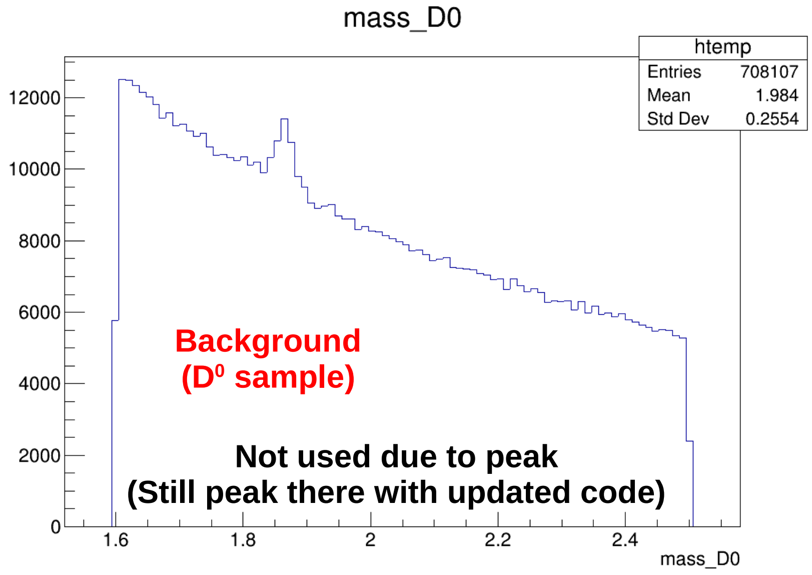
Sample After Preselection (Q^2 = 100 GeV^2)



Signal (D^0 sample)

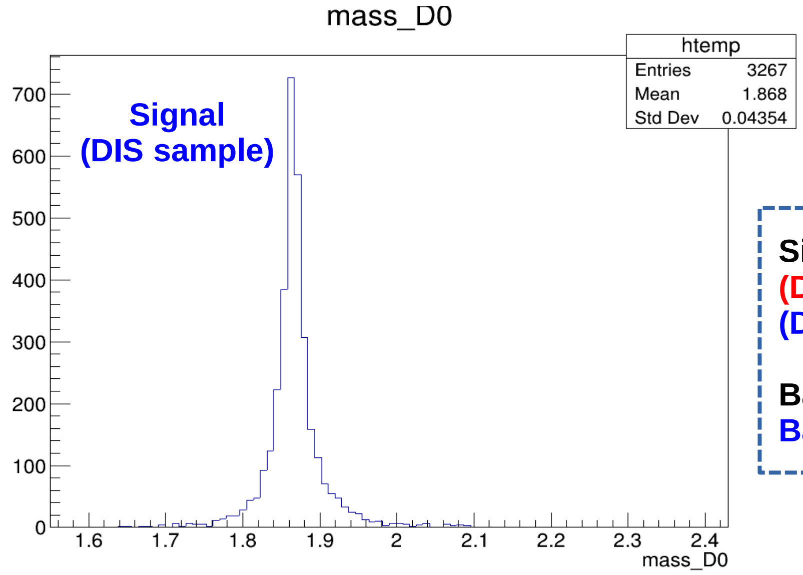
Preselection

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



Background (D^0 sample)

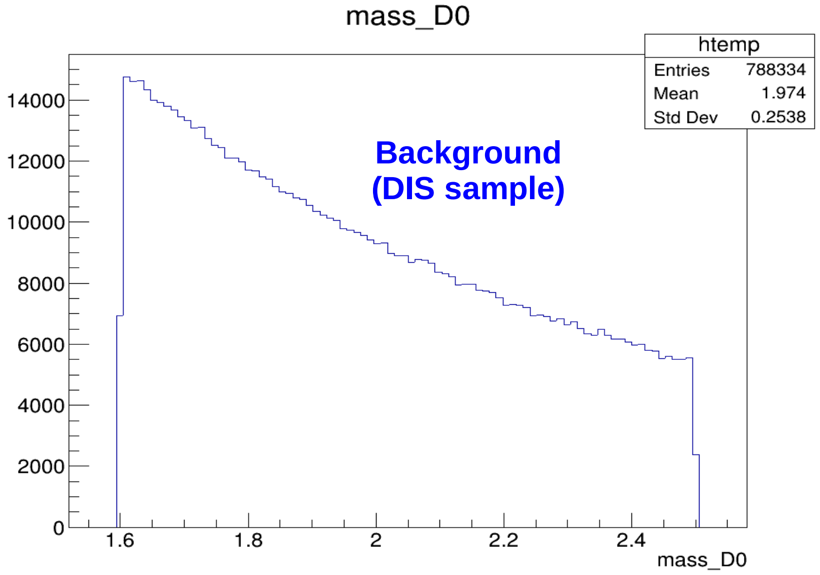
Not used due to peak
(Still peak there with updated code)



Signal (DIS sample)

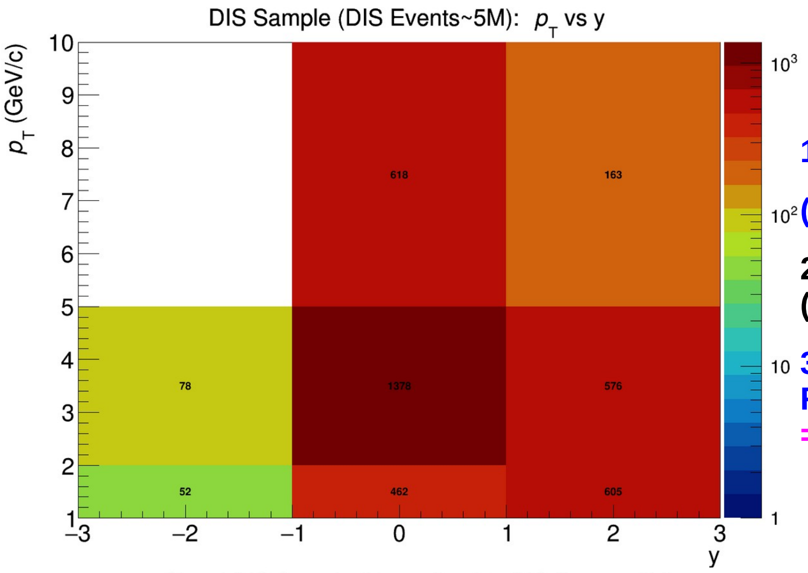
Signal for ML: Signal (D^0 sample) +Signal (DIS)

Background for ML: Background (DIS)



Background (DIS sample)

Method for Merging Signal and Background

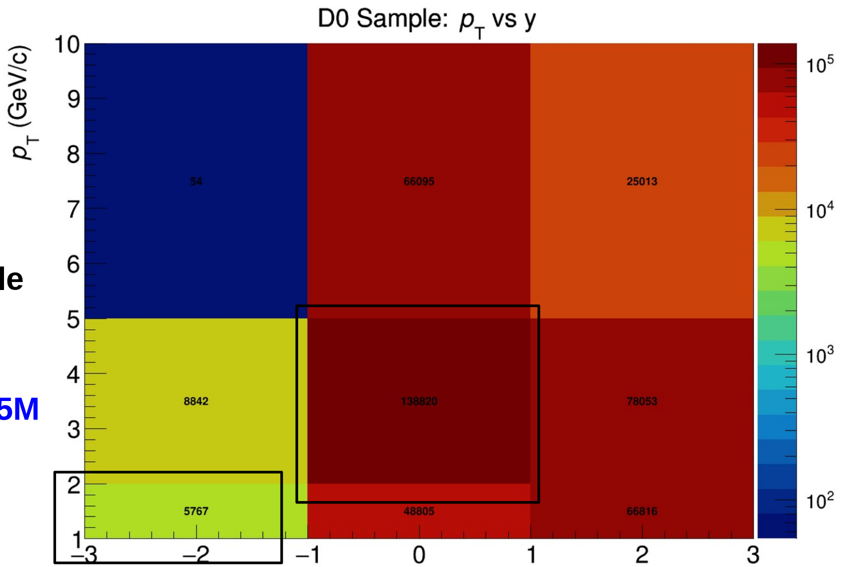

DIS Sample (DIS Events~5M): p_T vs y

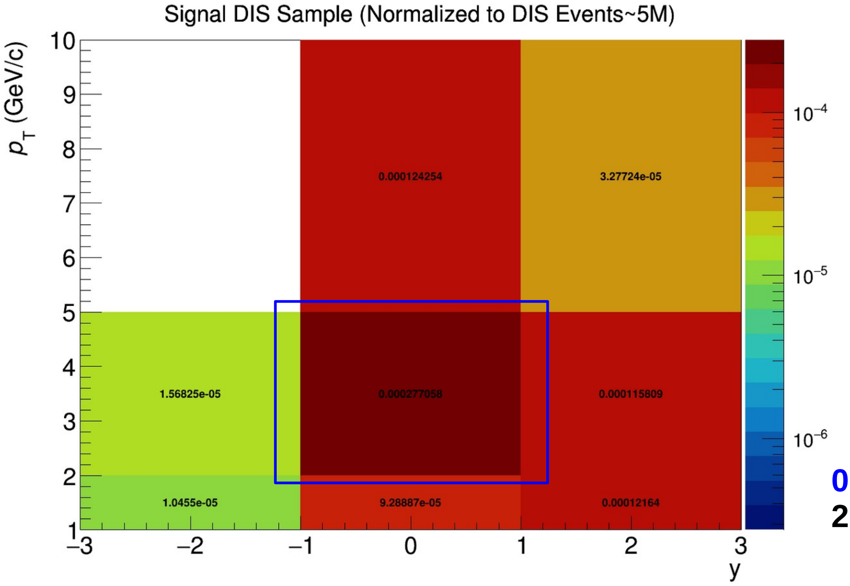
1. Reference Bin

(Max statistics)

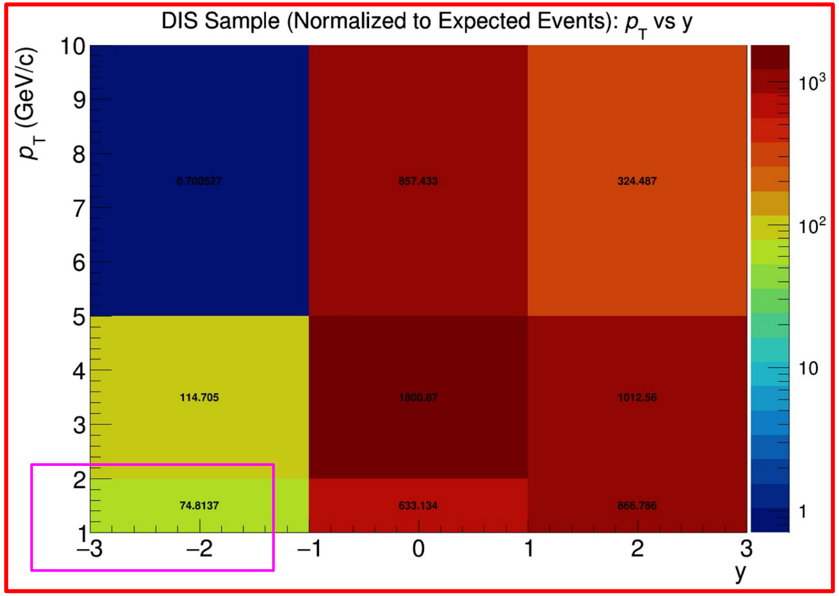
2. Fraction from D^0 sample (fraction = 5767/138820)

3. Expected D^0 (6.5 M)= ReferenceBin*fraction*6.5M = 74.81

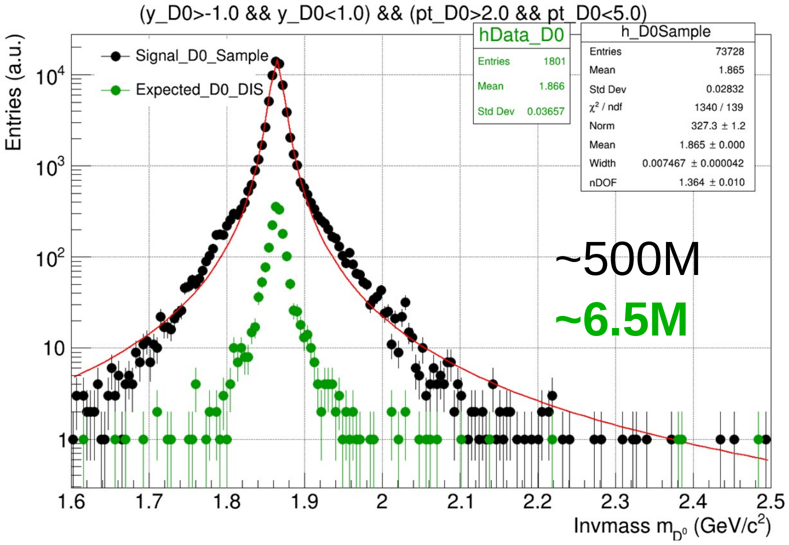

D0 Sample: p_T vs y


Signal DIS Sample (Normalized to DIS Events~5M)

0.000277*(5767./138820)*6.5e+6 ~ 74.81

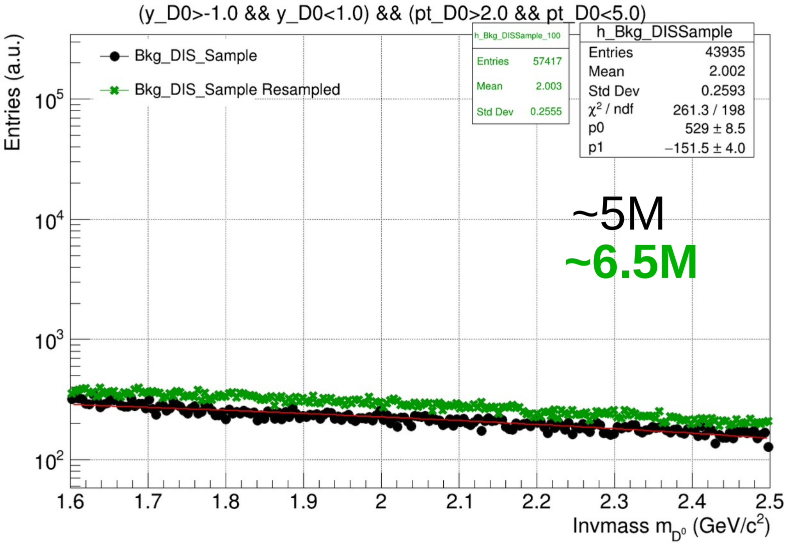

DIS Sample (Normalized to Expected Events): p_T vs y

Merging Signal and Background (D0 Sample and DIS Events)



~500M

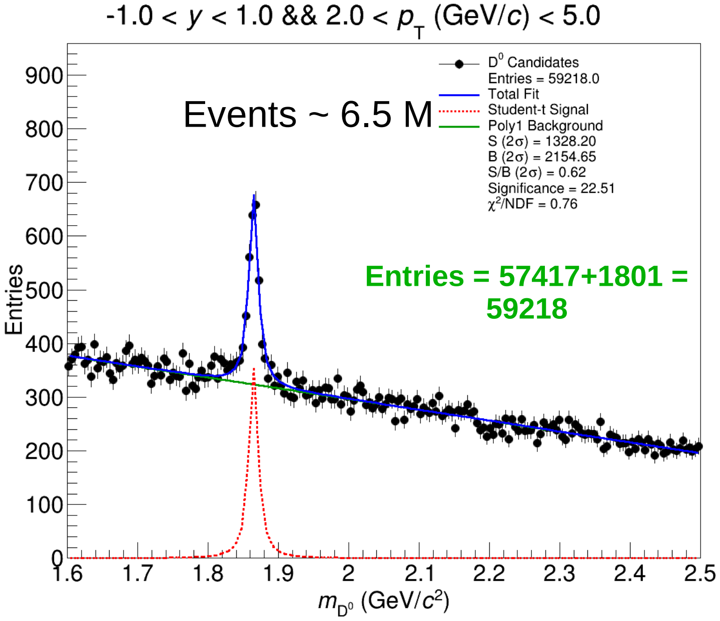
~6.5M



~5M

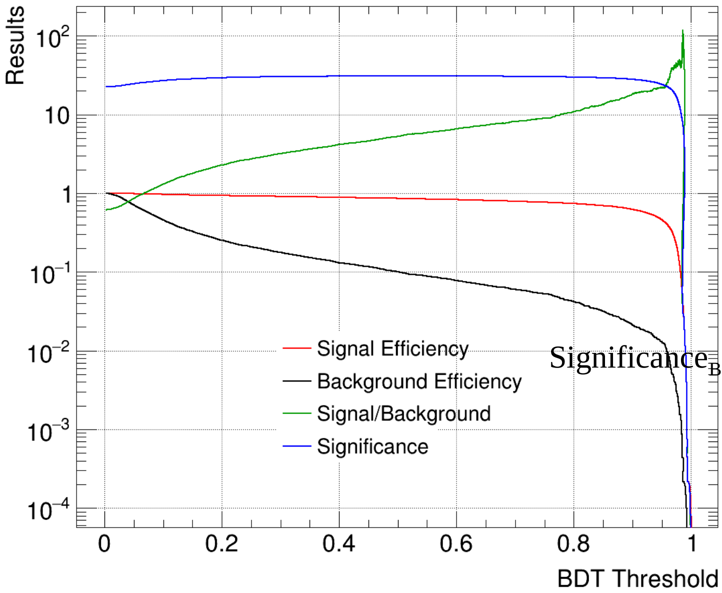
~6.5M

Resampling



Events ~ 6.5 M

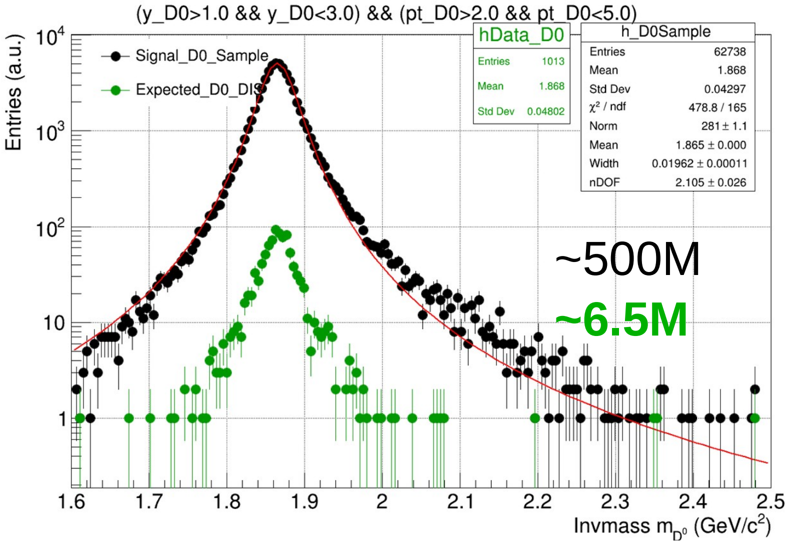
Entries = 57417+1801 = 59218



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

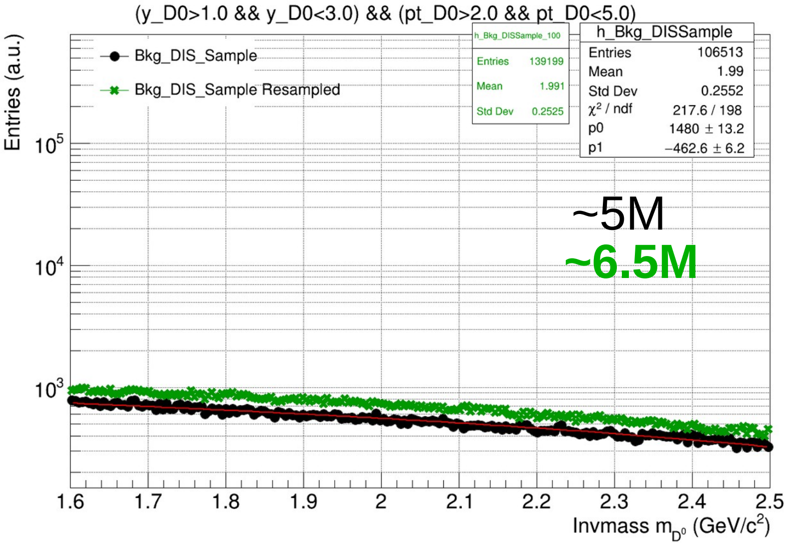
$$\text{Significance}_{\text{BDT Cut}} = \frac{S_{\text{No ML}} \times \epsilon_{\text{Signal}}}{\sqrt{S_{\text{No ML}} \times \epsilon_{\text{Signal}} + B_{\text{No ML}} \times \epsilon_{\text{Background}}}}$$

Merging Signal and Background (D0 Sample and DIS Events)



~500M

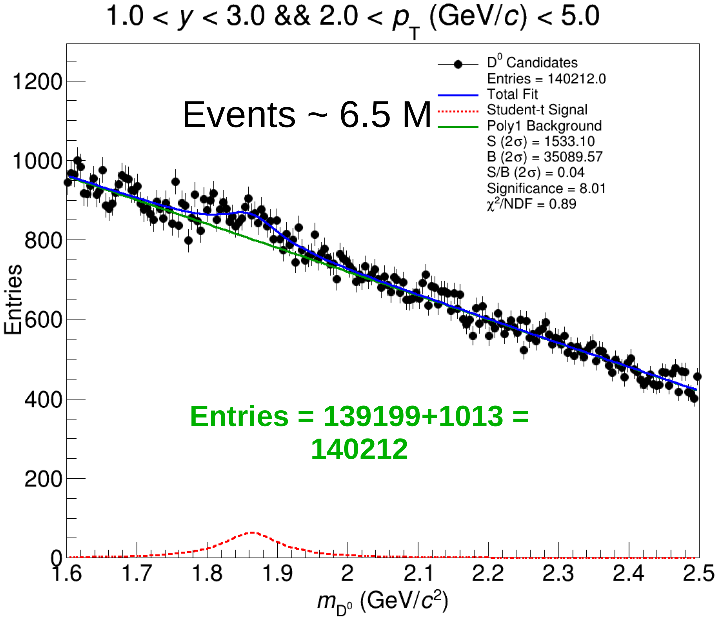
~6.5M



~5M

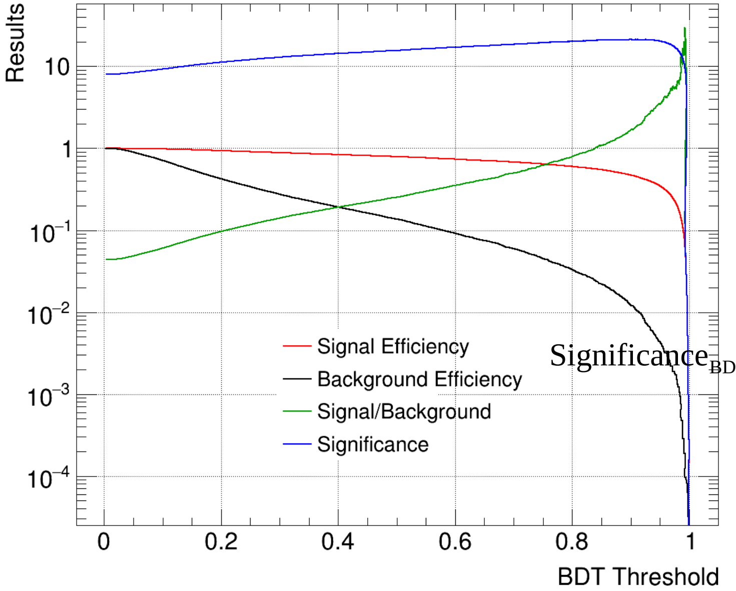
~6.5M

Resampling



Events ~ 6.5 M

Entries = 139199+1013 = 140212



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

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

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

There is also minor issue (std::map issue even in D^0) with associations if one reco track matches with two MC tracks, the code always considers last one, I can see print messages of two associations after changing a bit code

Track Parametrization (Local to Global)

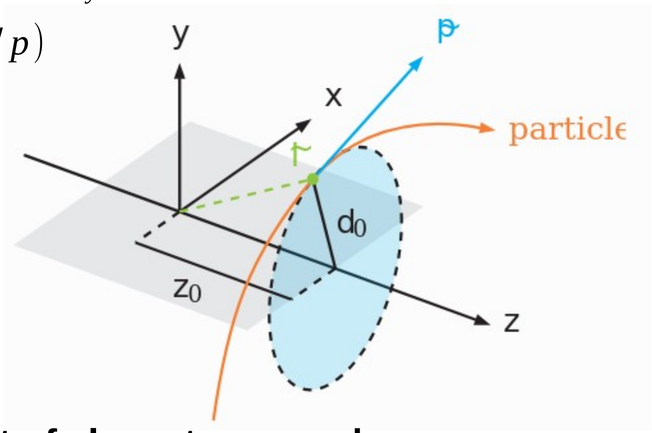
Helical Track model: $(l_0, l_1, \phi, \theta, q/p)$

$$x=-l_0\sin\phi,\quad y=l_0\cos\phi,\quad z=l_1$$

$$p_x=p\cos\phi\sin\theta,\quad p_y=p\sin\phi\sin\theta,\quad p_z=p\cos\theta$$

$$charge=sign(q/p)$$

$d_0 = l_0$
$z_0 = l_1$



At Point of closest approach (perigee surface)

$$\boxed{\begin{array}{c}(l_0, l_1, \phi, \theta, q/p)\\ \Big\downarrow \text{ Global (Lab frame)}\\ (x, y, z, p_x, p_y, p_z, q)\end{array}}$$

```
Vector3 LineSurface::localToGlobal(const GeometryContext& gctx, const Vector2& lposition, const Vector3& direction) const

{

 Vector3 unitZ0 = lineDirection(gctx);

// get the vector perpendicular to the momentum direction and the straw axis

Vector3 radiusAxisGlobal = unitZ0.cross(direction);

Vector3 locZinGlobal = transform(gctx) * Vector3(0., 0., lposition[1]);

// add loc0 * radiusAxis

return Vector3(locZinGlobal + lposition[0] * radiusAxisGlobal.normalized());

}
```

Calculation

UnitZ0: is (0,0,1) vector along the z-axis for cylinder and disks.

direction: (p Cos(phi) Sin(theta), p Sin(phi) Sin(theta), p Cos(theta))

radiusAxisGlobal = UnitZ0 Cross product direction = (-p Sin(phi) Sin(theta), p Cos(phi) Sin(theta), 0)

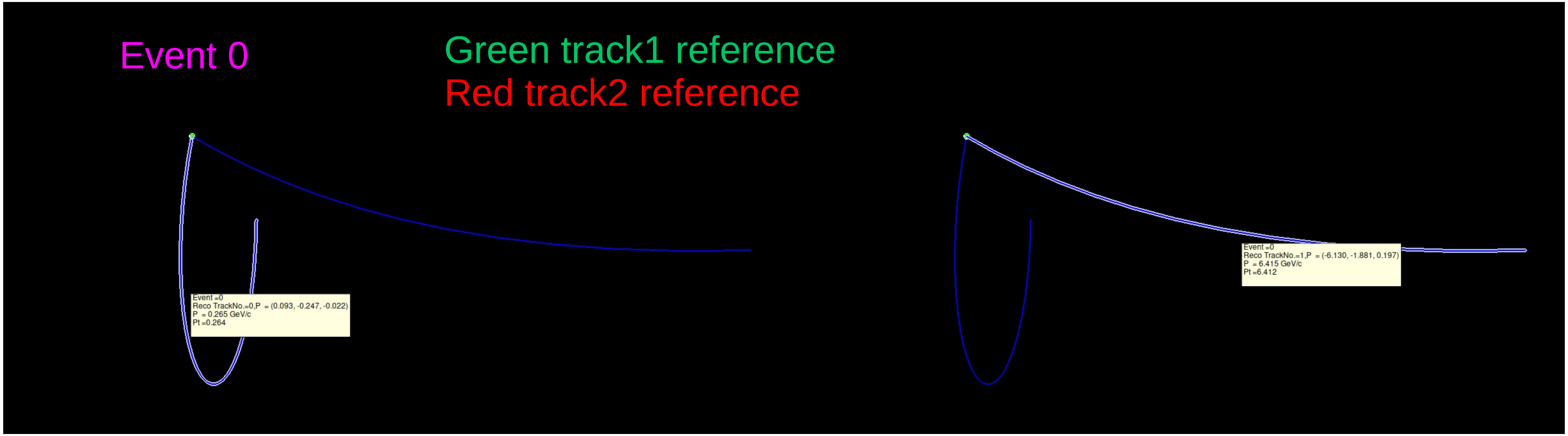
radiusAxisGlobal.Normalized = (-Sin(phi), Cos(phi), 0)locZinGlobal = (0,0,l1) (is same as global)

Global position = locZinGlobal + lposition[0] * radiusAxisGlobal.normalized() = (0,0,l1) + l0(-Sin(phi), Cos(phi), 0)Global Position = (-l0 Sin(phi), l0 Cos(phi), l1)

Returns the components, which we are using in HF analysis.

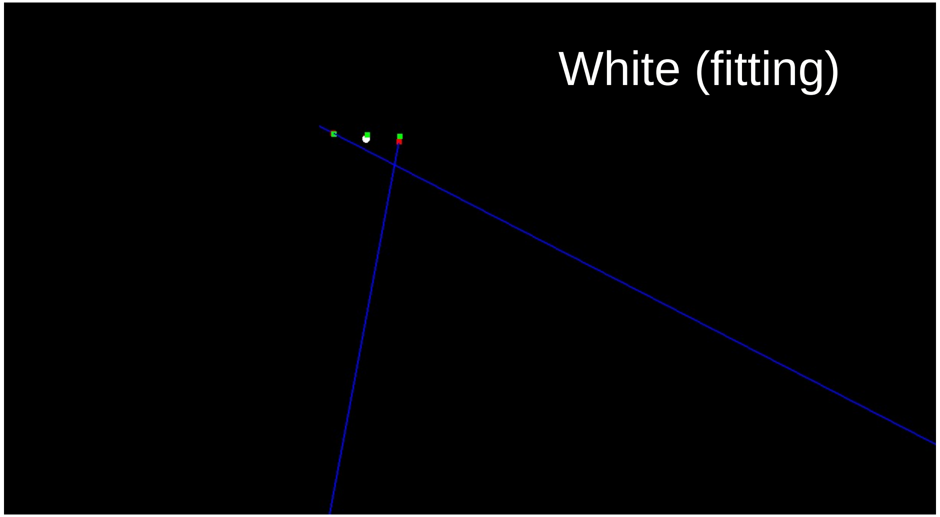
$x = -l_0 \sin\phi$, $y = l_0 \cos\phi$, $z = l_1$

Event display (First two tracks)



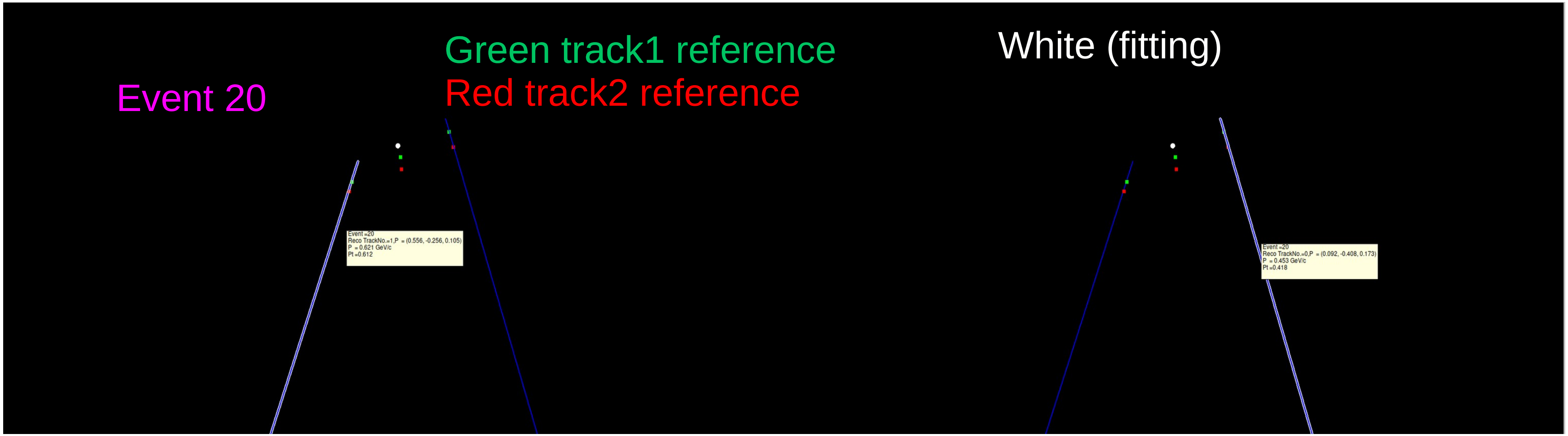
```
Event =0
Reco TrackNo.=0,P  = (0.093, -0.247, -0.022)
P  = 0.265 GeV/c
Pt =0.264
```

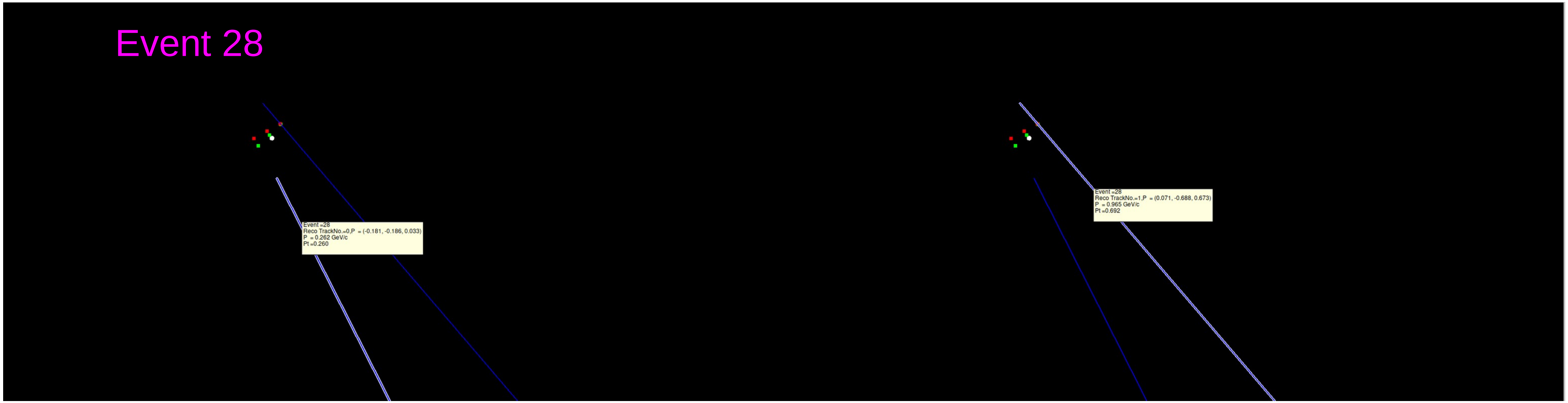
```
Event =0
Reco TrackNo.=1,P  = (-6.130, -1.881, 0.197)
P  = 6.415 GeV/c
Pt =6.412
```



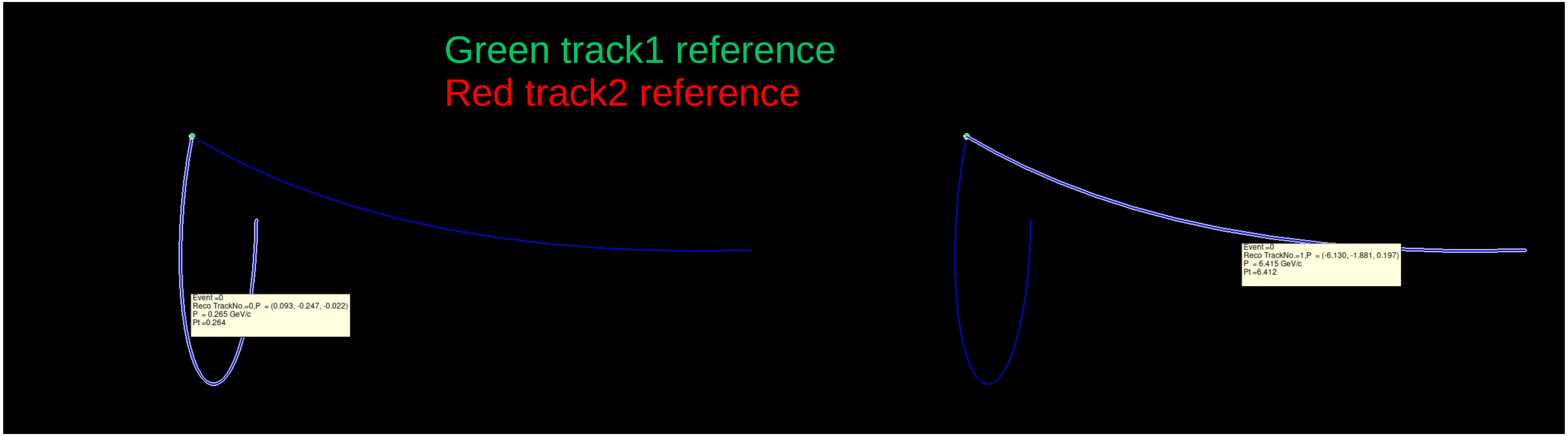
- All three methods (**track 1 reference, track2 reference, and distance minimization**) are compatible
- Minor difference is due to analytical approach
- Distance minimization returns the unique point

Event display (First two tracks)




Event 28
Event =28
Reco TrackNo.=0,P = (-0.181, -0.186, 0.033)
P = 0.262 GeV/c
Pt =0.260
Event =28
Reco TrackNo.=1,P = (0.071, -0.688, 0.673)
P = 0.965 GeV/c
Pt =0.692

Event display (First two tracks)

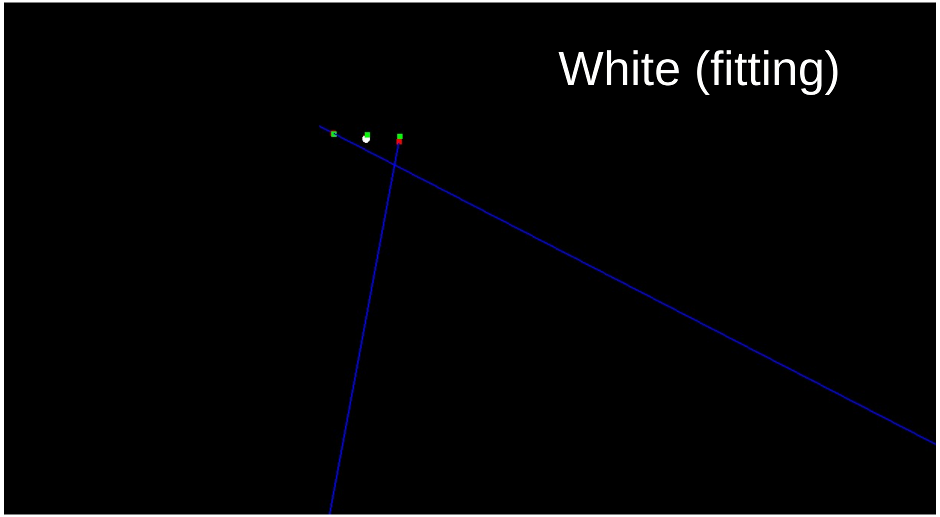


Green track1 reference

Red track2 reference

```
Event =0
Reco TrackNo.=0,P  = (0.093, -0.247, -0.022)
P  = 0.265 GeV/c
Pt =0.264
```

```
Event =0
Reco TrackNo.=1,P  = (-6.130, -1.881, 0.197)
P  = 6.415 GeV/c
Pt =6.412
```



White (fitting)

- All three methods (**track 1 reference, track2 reference, and distance minimization**) are compatible
- Minor difference is due to analytical approach
- Distance minimization returns the unique point