Charm Hadron Reconstruction with ML in the ePIC Experiment

Shyam Kumar*, Annalisa Mastroserio, Domenico Elia

INFN Bari, Italy

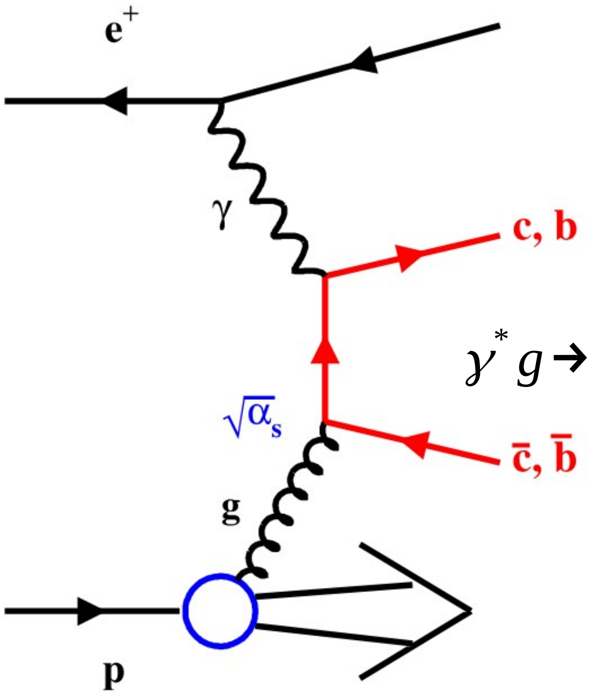
hipe4ml package

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

Heavy-flavor Production in the ePIC Experiment

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

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/c²

$m_{\Lambda c+}$ = 2.28646 GeV/c²

$c \longrightarrow D^0$ — Prompt D^0-meson (fragmentation of charm quark)

$b \longrightarrow B \longrightarrow D^0$ — Non-prompt D^0-meson (coming from beauty-hadron decays)

Signal refers to prompt and non-prompt charm hadrons

Boson-Gluon Fusion (BGF) is dominant mechanism [LO] at high energy ep collisions

ep (10X100) collisions with Q^2>1 GeV²

Machine Learning (ML) techniques useful to extract signal:

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

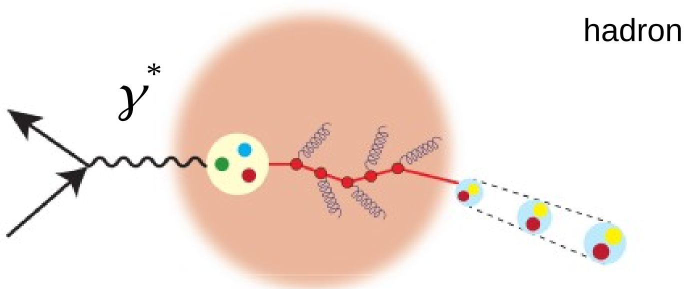
Multi class classifier: Machine learning model to separate prompt, non-prompt D^0 meson, and background

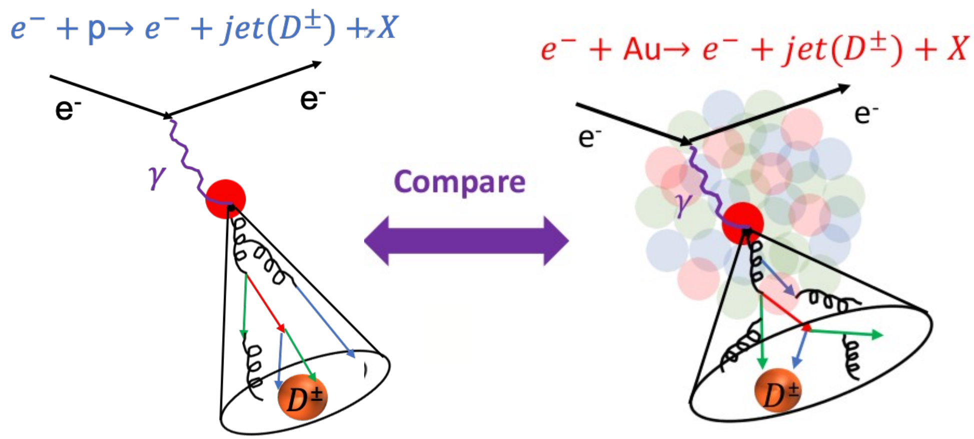
Comparison between ep and eA collisions

- Study hadronization mechanisms
- D mesons and Λ_c^+ baryon R_{eA} measurements to investigate energy loss and transport of charm in the medium

Nuclear modification factor: $R_{eA} = \frac{\sigma_{eA}}{A \times \sigma_{ep}}$

Parton energy loss
(Collisional+Radiative)





ep, eAu collisions (10X100 GeV) DIS collisions with $Q^2 > 1\text{GeV}^2$

- Heavy-flavour hadrochemistry and collectivity:
  - ➔ **Hadronization modification in presence of cold-nuclear matter**

Hadronization: heavy-flavour baryon over meson ratios

Factorization Approach:

$z = p_{HQ}/p_Q$

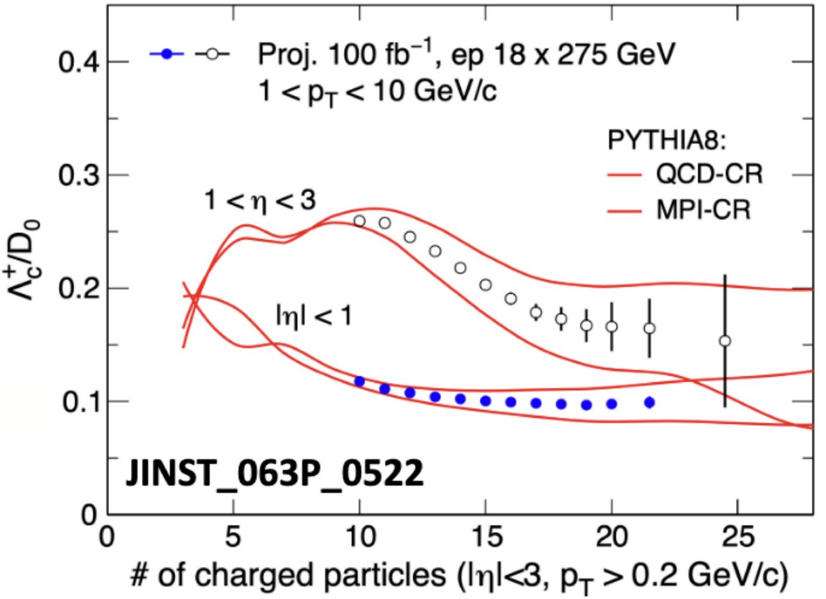
$$\frac{d\sigma^{NN\to H_Q X}}{dp_T^{H_Q}}(\sqrt{s_{NN}}, M_Q, \mu_F^2, \mu_R^2) = \sum_{i,j=q,\bar{q},g} \boxed{f_i(x_1,\mu_F^2)\otimes f_j(x_2,\mu_F^2)} \otimes \boxed{d\hat{\sigma}^{ij\to Q\bar{Q}}(\alpha_s(\mu_R^2),\mu_F^2, M_Q, x_1, x_2, s_{NN})} \otimes \boxed{D_Q^{H_Q}(z,\mu_F^2)}$$

Parton distribution functions — Hard Scattering Cross-section — Fragmentation function

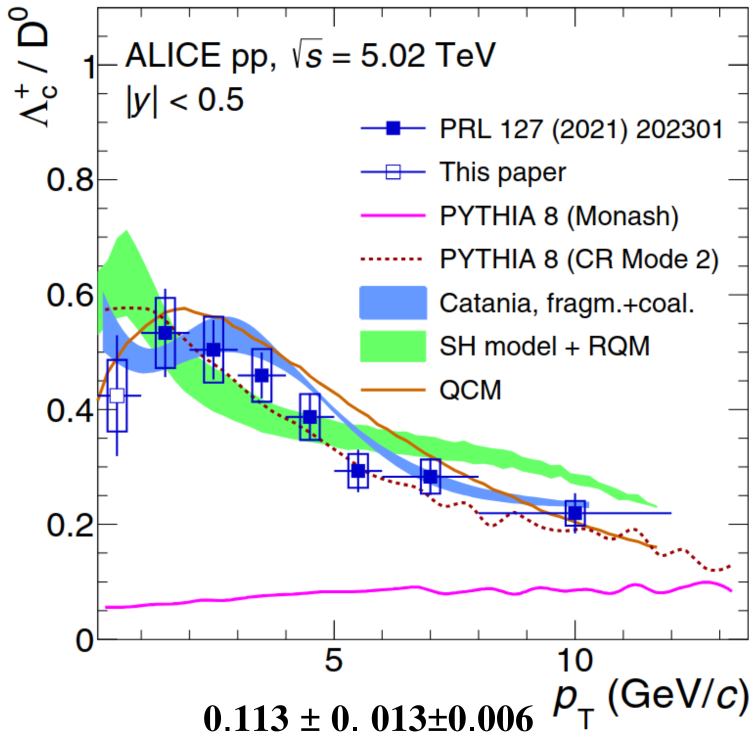
- Charm/beauty baryon-to-meson ratios: higher than measurements at e^+e^-, ep collisions for p_T < 10 GeV/c

  ➔ Predictions that include baryon enhancement mechanisms describe data

Projection based on fast simulation



https://doi.org/10.1103/PhysRevC.107.064901



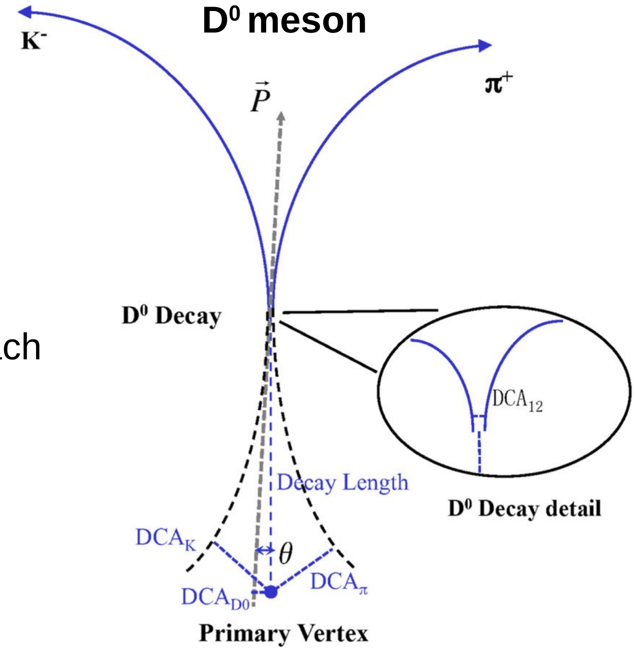
0.113 ± 0. 013±0.006
(LEP average, EPJC 75, 19 (2015))

Main objective is to measure Λ_c^+/D^0 ratio in ep, and their possible modification in eAu (10X100 GeV) DIS collisions with $Q^2 > 1GeV^2$

Topological Variables for Reconstruction of D^0 meson

arXiv:1911.12168 [nucl-ex]

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



DCA: Distance of closest approach

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 **dl** 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
- y_D0 reconstructed y of the D^0 meson
- Multiplicity (mult)

Decay length (dl), Primary Vertex (PV), Secondary Vertex (SV)

$$\vec{dl} = \vec{SV} - \vec{PV}$$

$$\cos\theta = \frac{\vec{dl}\,.\,\vec{p}}{|\vec{dl}||\vec{p}|}$$

$$DCA_{D0} = dl \sin\theta$$

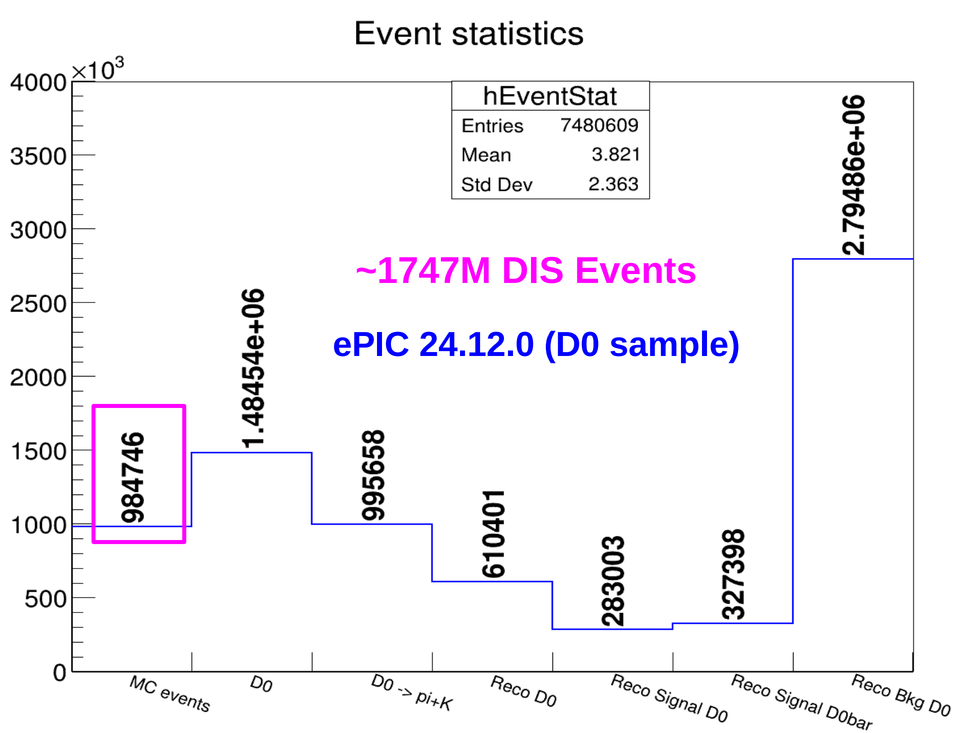
Realistic PID (PID lookup tables) for D^0 meson and Truth PID (MonteCarlo) for Λ_c^+ Baryon

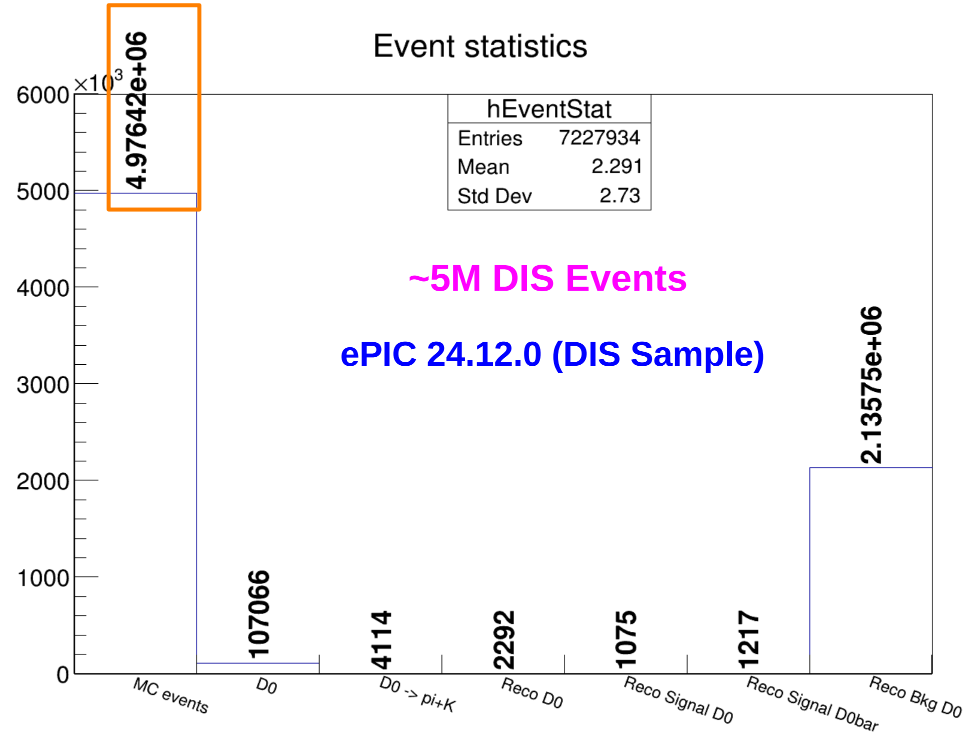
Data Sample (Q^2 > 1 GeV^2)

- Boosted Decision Tree (BDT) requires the features for the signal D^0 meson and background D^0 meson (fake combinations of pion,kaon)
  - D^0 sample created filtering **PYTHIA8 ep, NC, 10X100, Q^2 >1 GeV^2 events (~1747M)** such that each event consists 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

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





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

HF Tasks

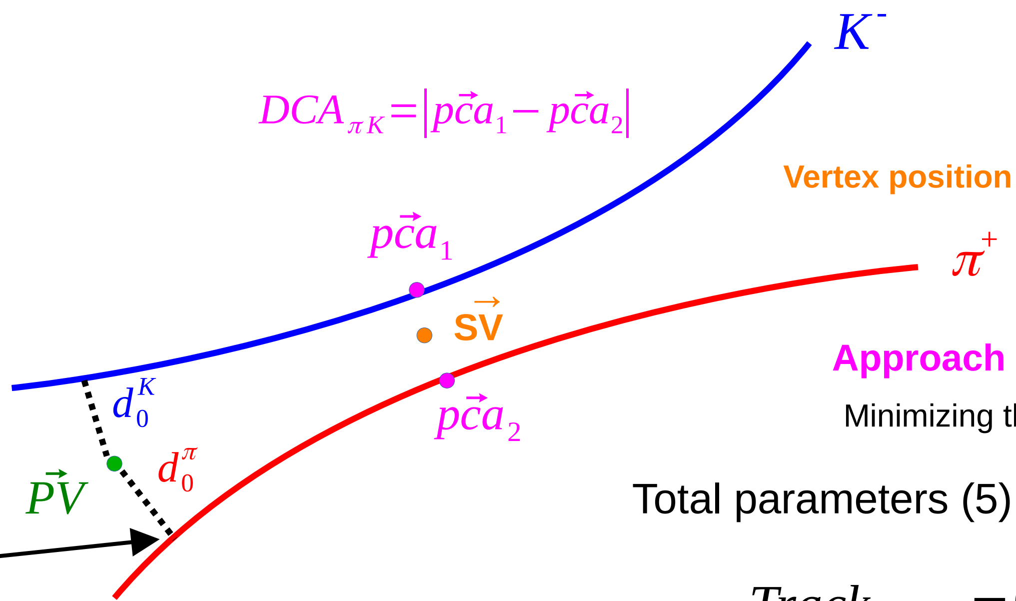
My Slides from Jets and Heavy-flavor meeting

Secondary Vertex

Approach 1 (Analytical)

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

Ignored track errors (at the moment)



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

$$Track_{DCA} = (\vec{r}_0, \vec{p}_0, q)$$

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

Approach 2 (Shyam)

Minimizing the distance

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

$$Track_{At(s1)} = (\vec{r}_{s1}, \vec{p}_1, q_1)$$
$$Track_{At(s2)} = (\vec{r}_{s2}, \vec{p}_2, q_2)$$

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

Comparison of four approaches:

- Helix1 (using helix1 to find **pca_1** and **pca_2)**
- Distance minimization (d)
- Helix2 (using helix2 to find **pca_1** and **pca_2)**
- Using average of Helix1 and Helix2

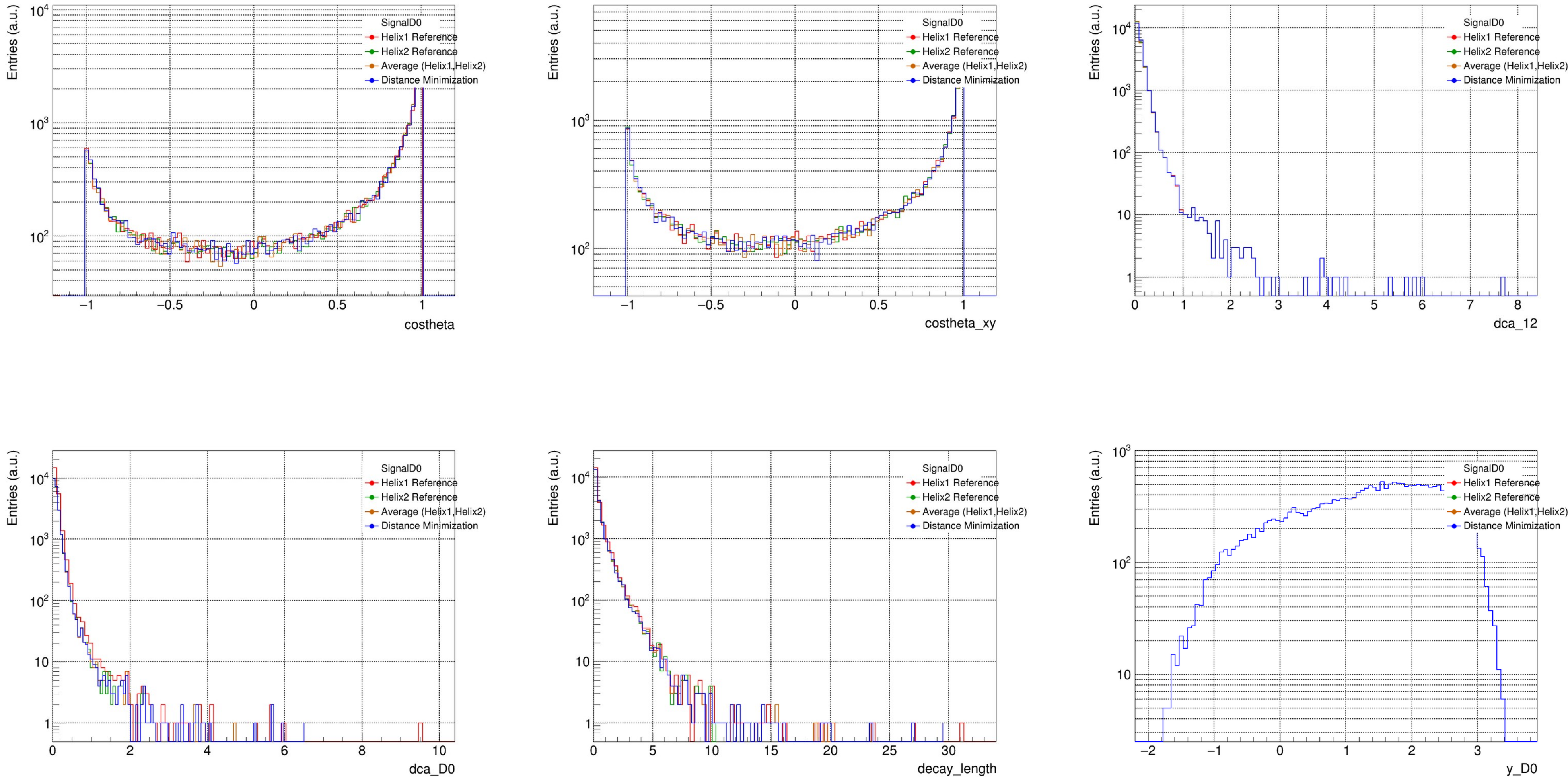
Final Approach: Secondary vertexing in ACTS using AdaptiveMultiVertexFinder (considers track error properly)

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

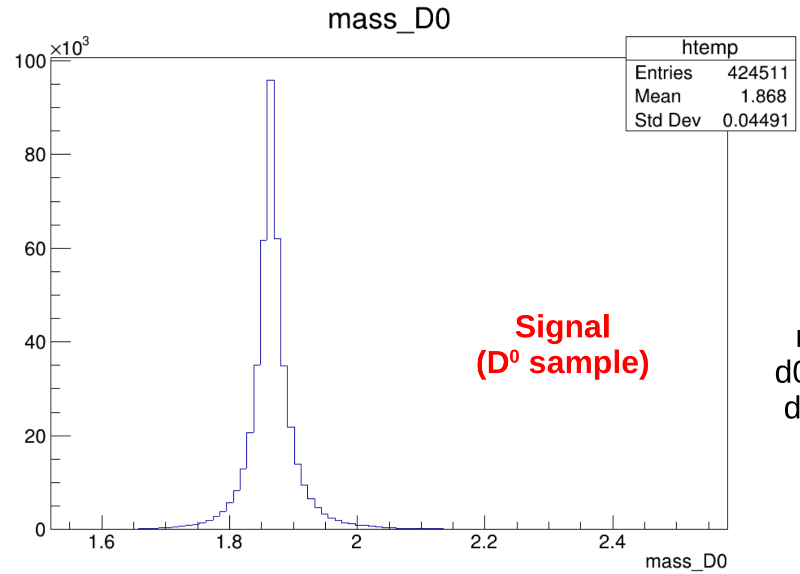
All methods are compatible

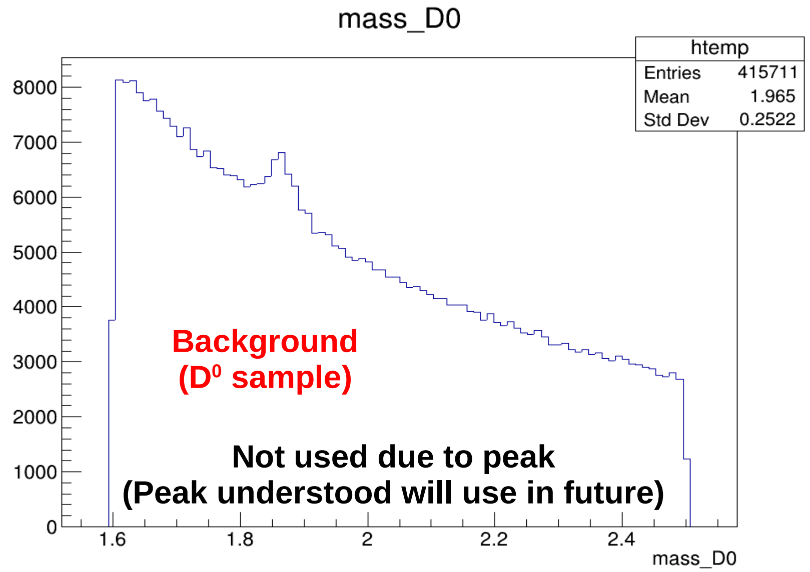
Signal D^0 meson

Distance minimization gives unique secondary vertex



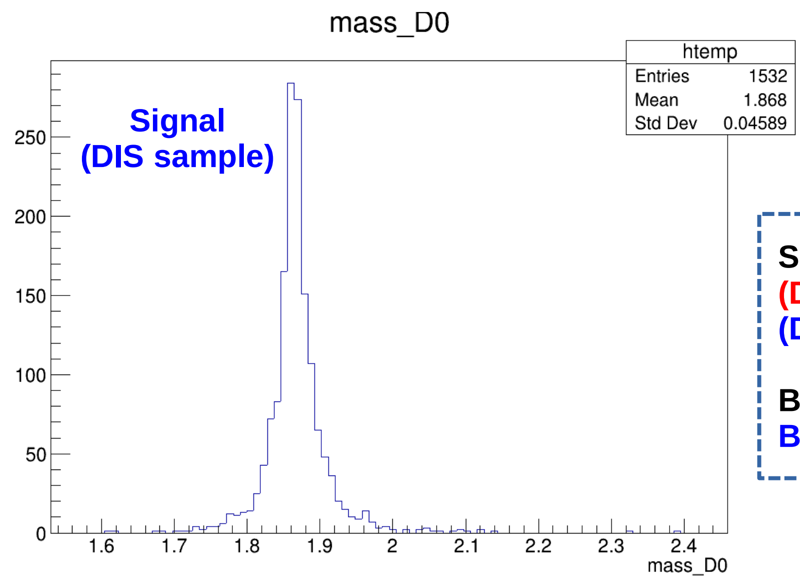
Sample After Preselection (Q^2 > 1 GeV^2)





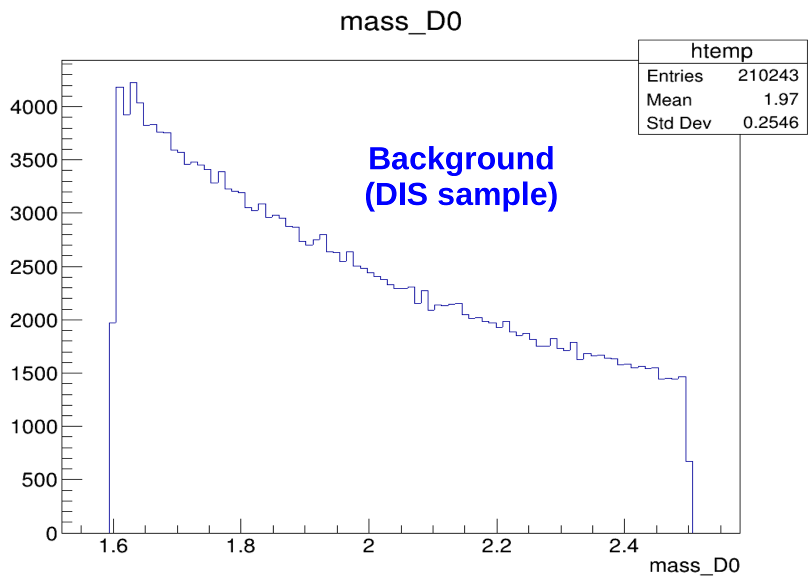
Preselection

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



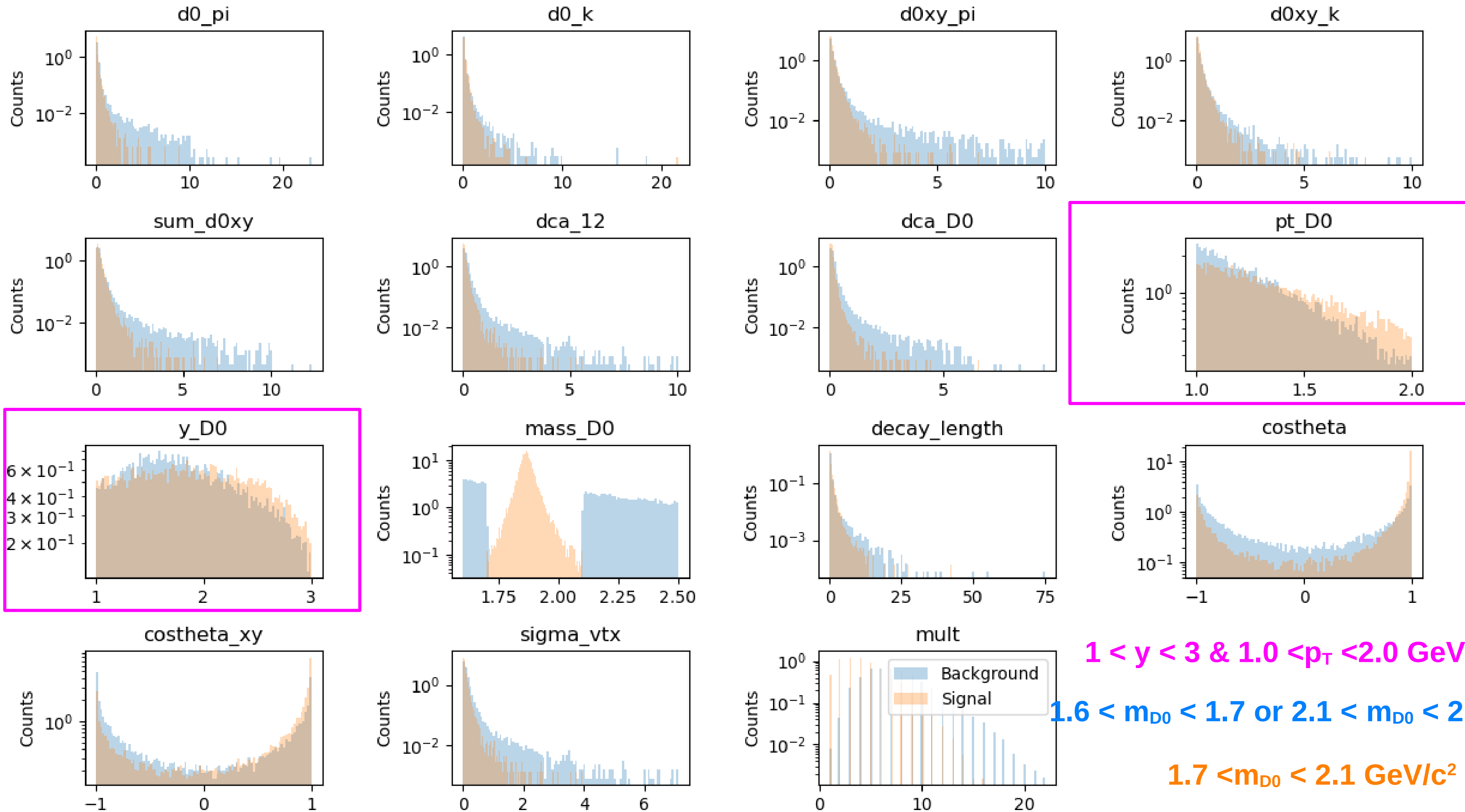
Signal for ML: Signal (D^0 sample) +Signal (DIS)

Background for ML: Background (DIS)



Signal and Background distributions

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



$1 < y < 3$ & $1.0 < p_T < 2.0$ GeV/c

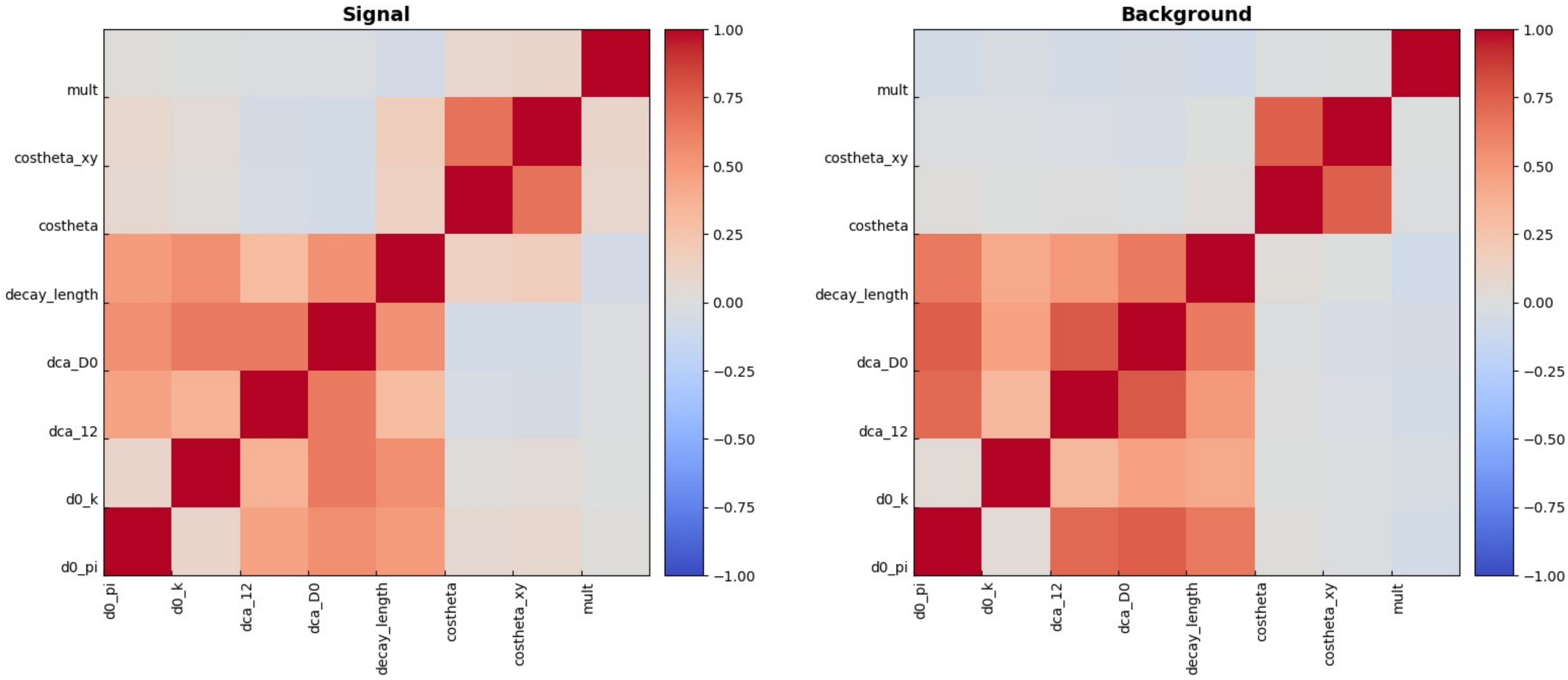
$1.6 < m_{D0} < 1.7$ or $2.1 < m_{D0} < 2.5$ GeV/c^2

$1.7 < m_{D0} < 2.1$ GeV/c^2

Correlation Matrix

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

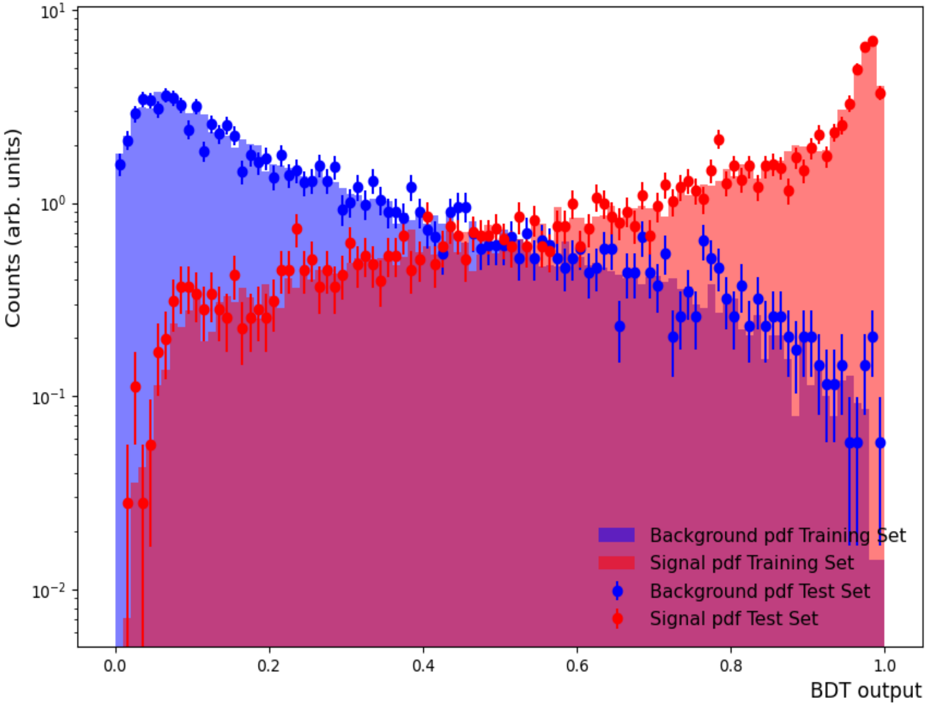
Planning to remove costheta_xy and decay_length once other cuts are available (e.g. chi2, nsigma, etc.)



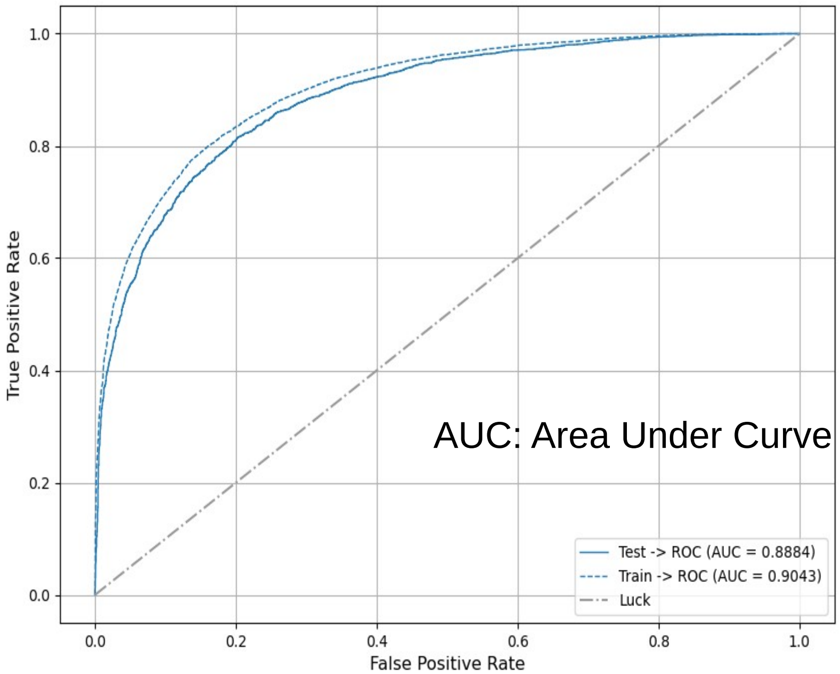
1.0 < y < 3.0
1.0 < p_T < 2.0 GeV/c

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

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



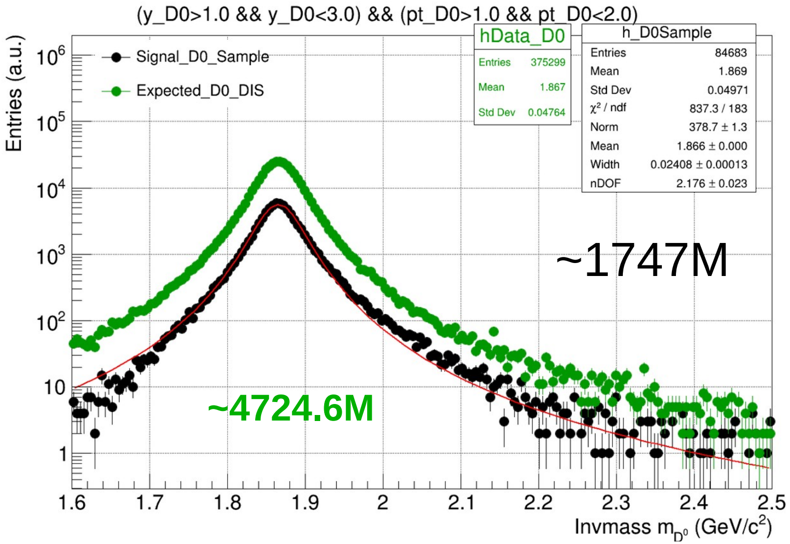
How Boosted Decision Tree (BDT) classifier separates signal from background

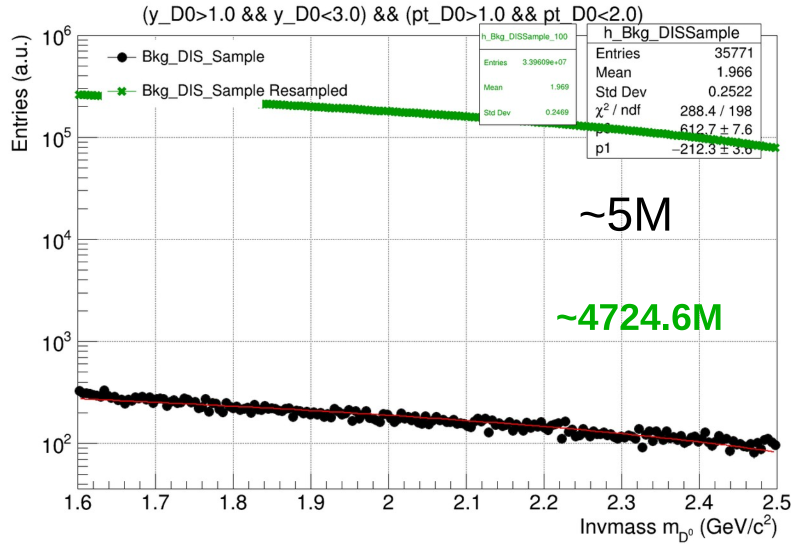


Receiver Operating Characteristic (ROC)

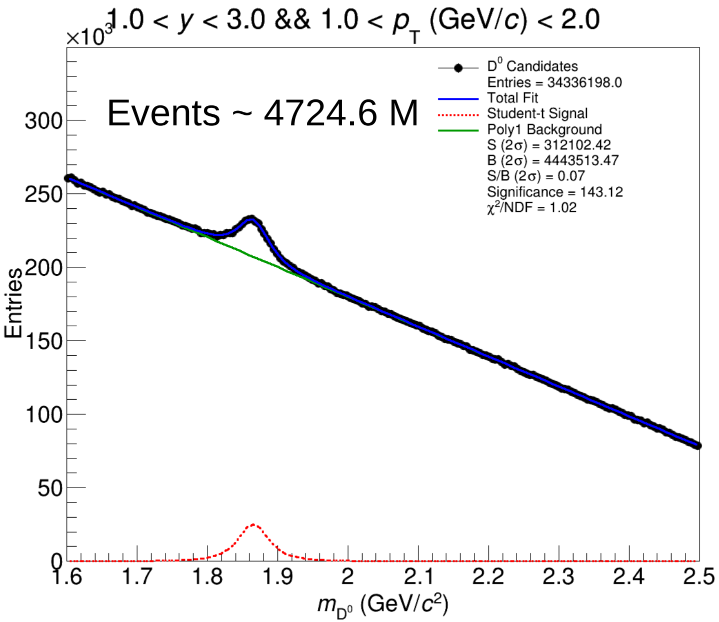
A perfect classifier would have a point at (0, 1), indicating no false positives and all true positives

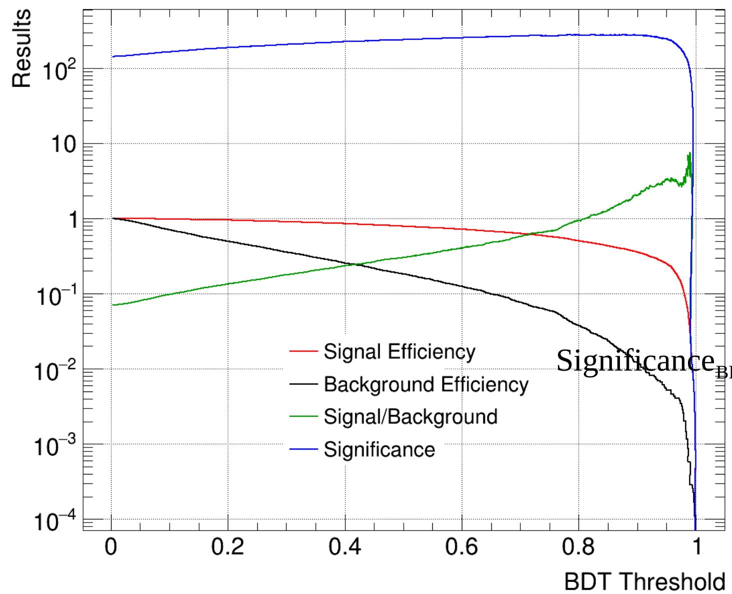
Signal and Background (D^0 Sample and DIS Events)





Resampling



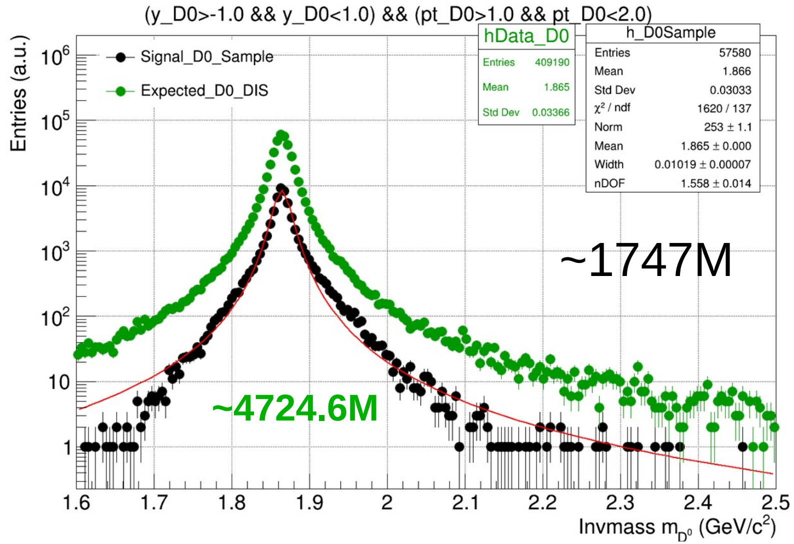


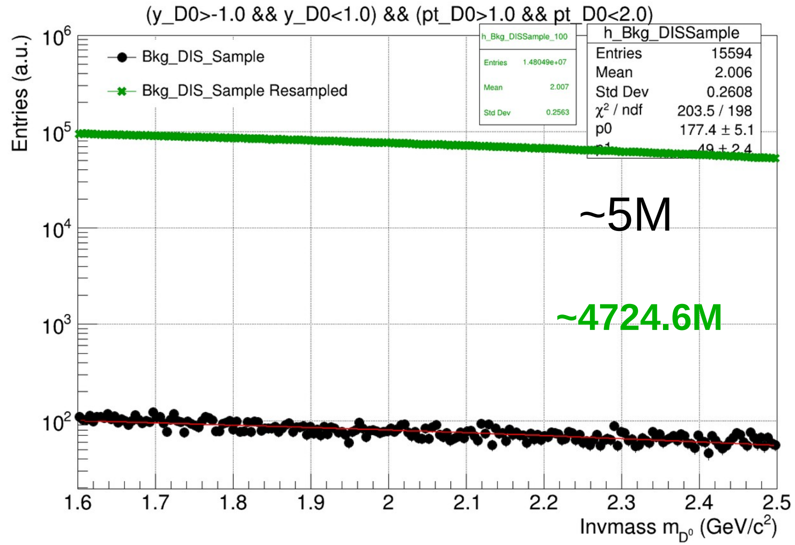
$$L = 10\, fb^{-1} \sim 4724.6\, M$$

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

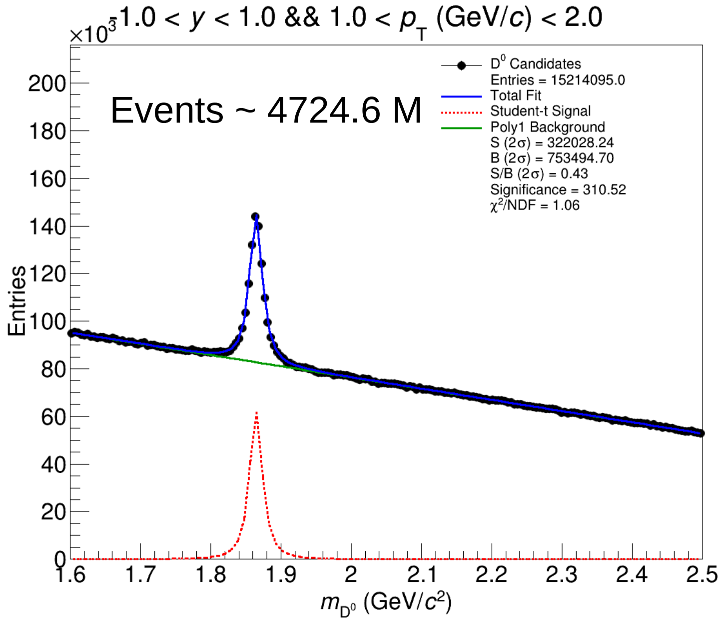
$$\text{Significance}_{\text{BDT Cut}} = \frac{S_{\text{No ML}} \times \epsilon_{\text{Signal}}}{\sqrt{S_{\text{No ML}} \times \epsilon_{\text{Signal}} + B_{\text{No ML}} \times \epsilon_{\text{Background}}}}$$

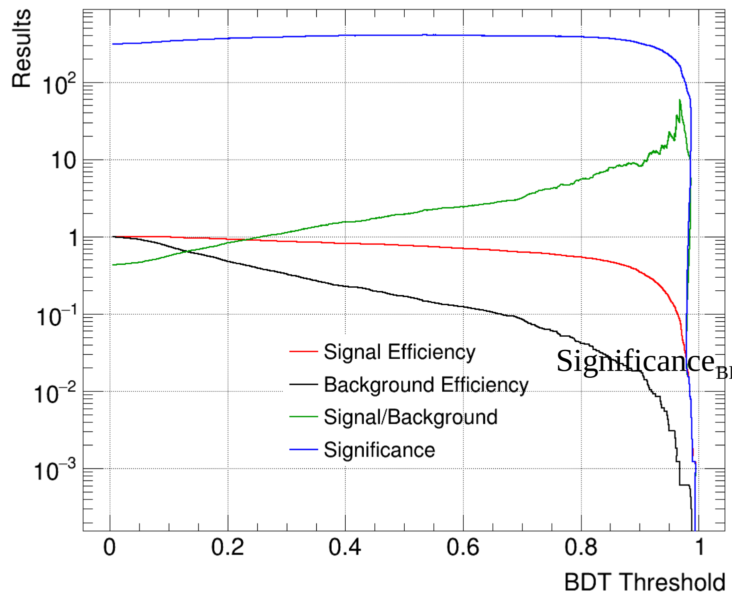
Signal and Background (D0 Sample and DIS Events)





Resampling





$$L = 10\, fb^{-1} \sim 4724.6\, M$$

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

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

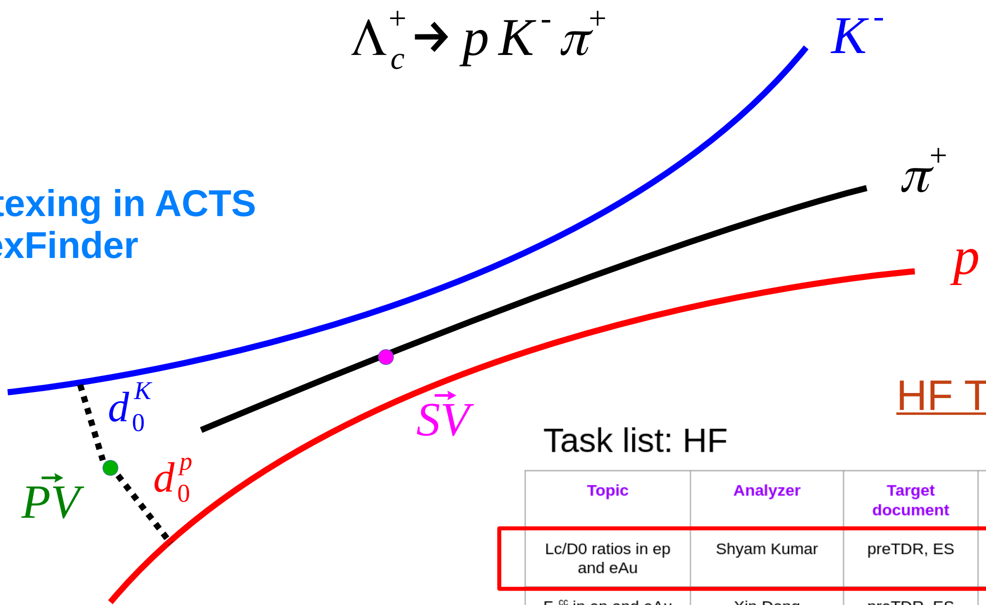
Λ_c^+ Baryon Reconstruction

Analytical Approach for secondary vertex

Primary Vertex (PV)

Secondary Vertex (SV)

Final Approach: Secondary vertexing in ACTS using AdaptiveMultiVertexFinder



My Slides from Jets and Heavy-flavor meeting

HF Tasks

Task list: HF

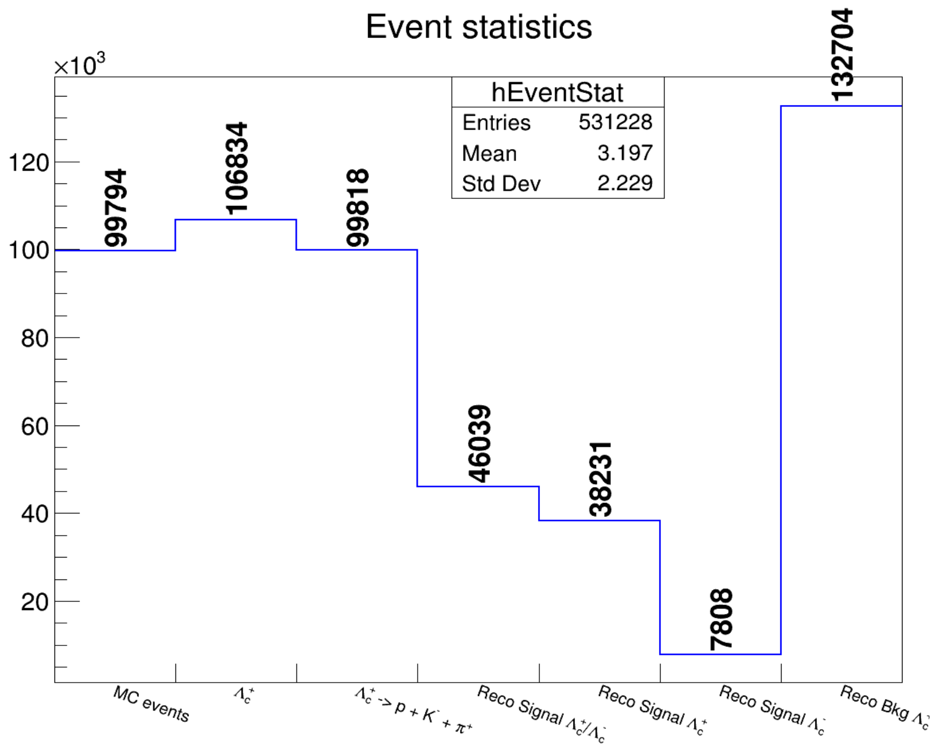
ES: early science

Topic	Analyzer	Target document	Simulations needed	Simulation status
Lc/D0 ratios in ep and eAu	Shyam Kumar	preTDR, ES	ep, eAu@10x100, DIS, D0, Lc, Q2 > 1	
F_2^{cc} in ep and eAu	Xin Dong	preTDR, ES	ep@10x100, 10x250, 5x41, DIS, D0, Q2 > 1 eAu@10x100, 5x41, DIS, D0, Q2 > 1	
D0 R_{eAu} vs. pT, x, z	Rongrong Ma Connie Yang Deepa Thomas	preTDR, ES	ep, eAu@10x100, DIS, D0, Q2 > 1	
D0 mass peak, pointing resolution, Primary vs. secondary track DCA	Rongrong Ma	preTDR	ep, eAu@10x100, DIS, D0, Q2 > 1	

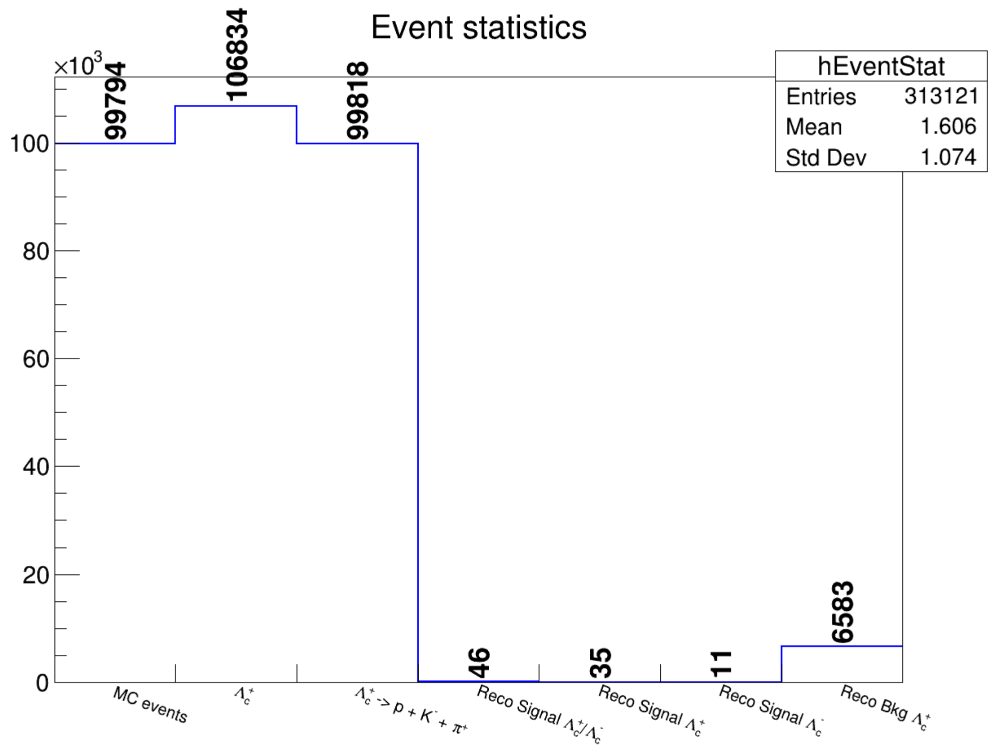
PYTHIA8 ep NC (10x100 Q^2>1) Λ_c^+ sample: by Rongrong

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

Truth PID



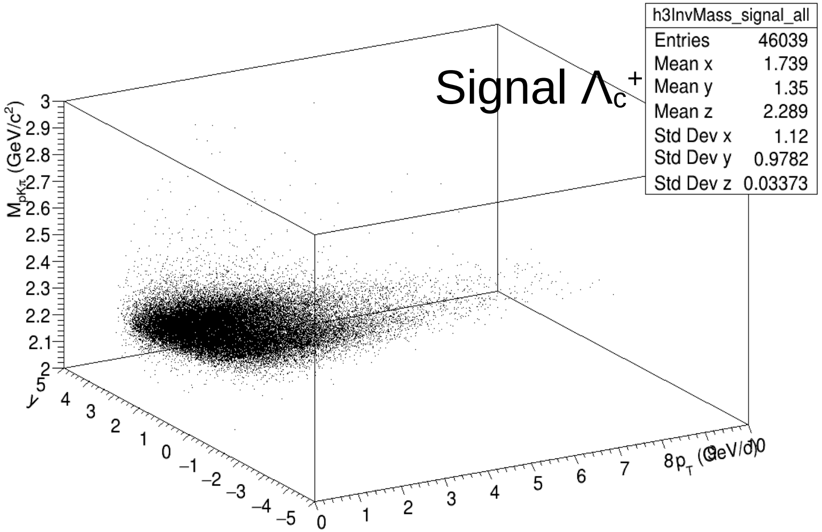
Real PID



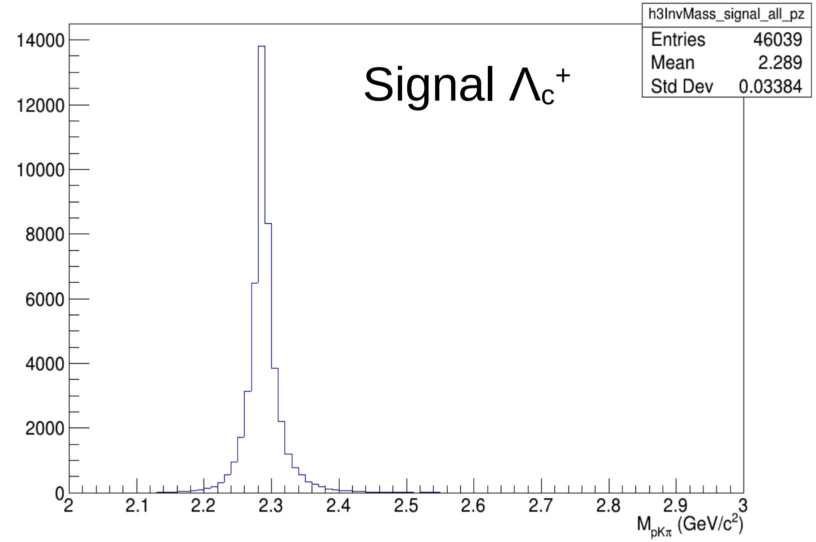
Issue coming from proton PID, it looks not properly assigned in reconstruction (loosing protons)

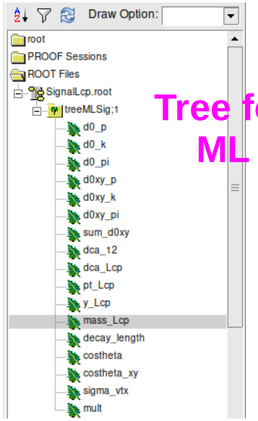
Results

Truth PID



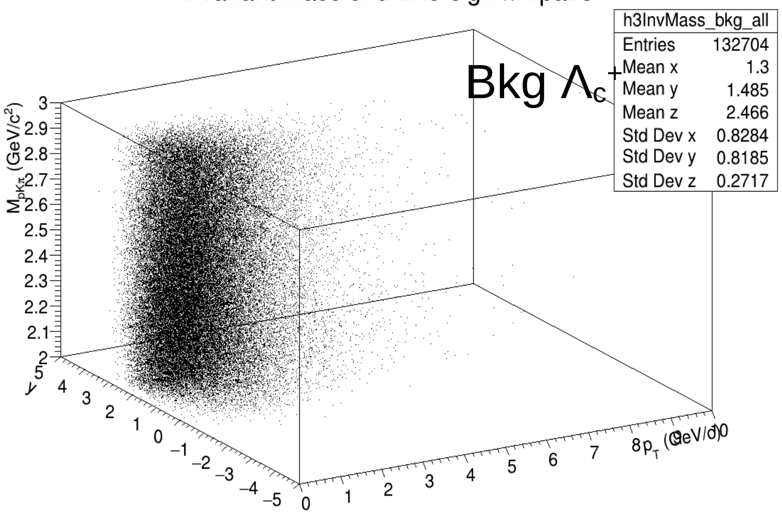
Projection



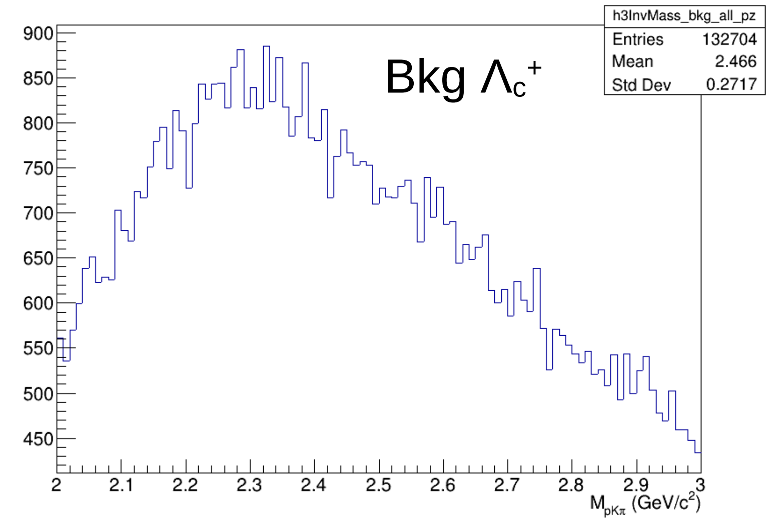


Tree for ML

Invariant mass of unlike-sign πK pairs



Projection



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 (**AdaptiveMultiVertexFinder**) from ACTS 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
  - Fix the reconstruction for the realistic PID of proton
  - Implement similar ML model for Λ_c^+ reconstruction (quicker)
  - Run on full statistics 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 !!!

Track Parametrization (Local to Global)

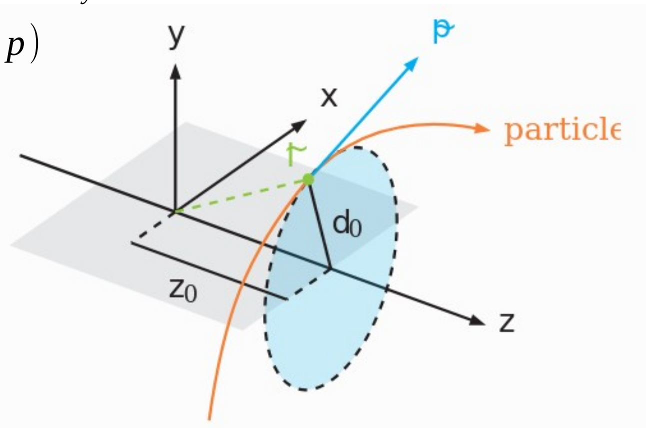
Helical Track model: $(l_0, l_1, \phi, \theta, q/p)$

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

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

$charge = sign(q/p)$

$d_0 = l_0$
$z_0 = l_1$



At Point of closest approach (perigee surface)

$(l_0, l_1, \phi, \theta, q/p)$

Global (Lab frame)

$(x, y, z, p_x, p_y, p_z, q)$

```
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.

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

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

The peak observed in the background of the D^0 sample was caused by mis-associations, but we resolved it a few days ago. We also plan to use the background D^0 sample for machine-learning studies in the future

New strategy

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

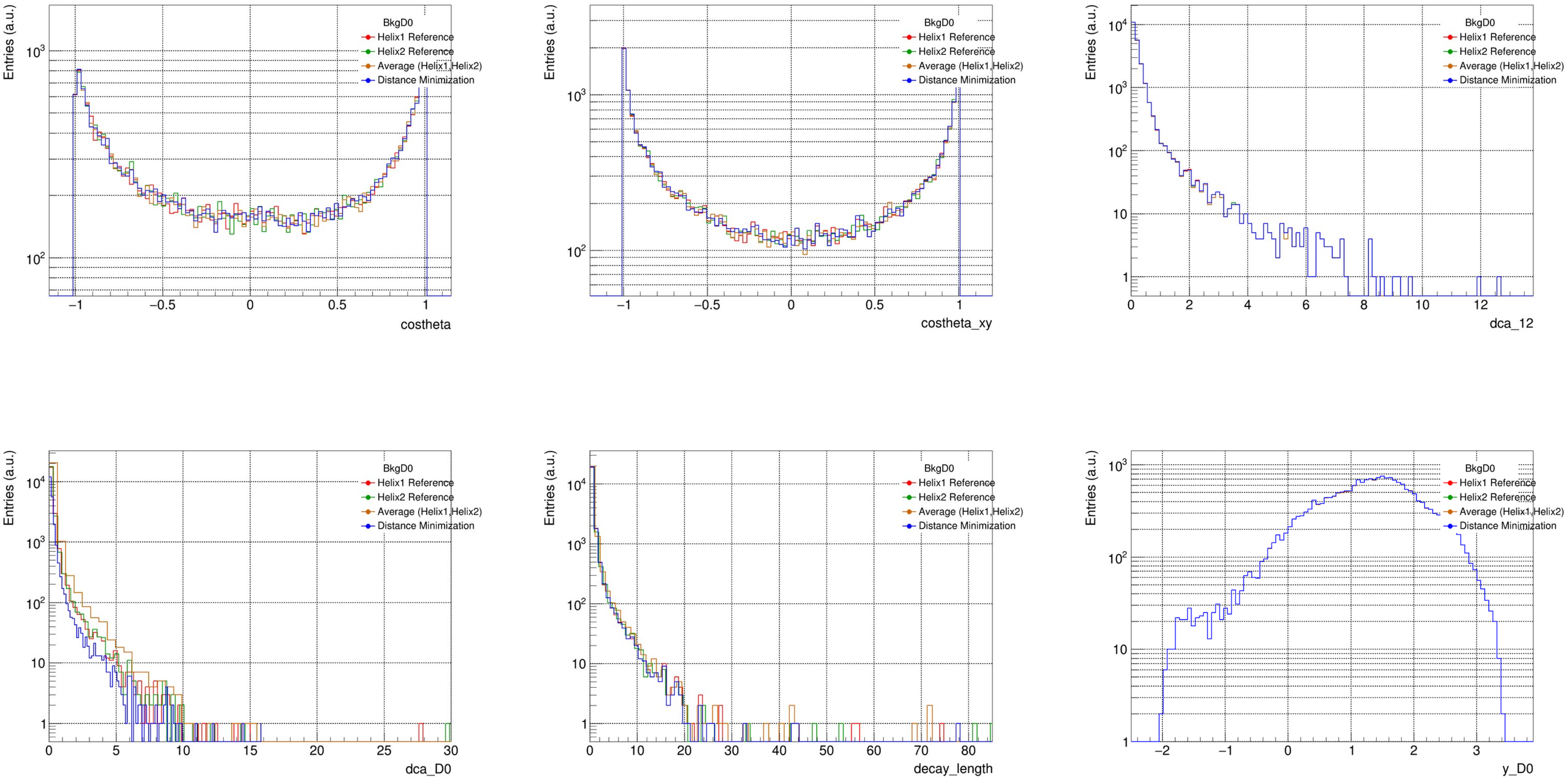
Background for ML: Background (D^0 sample) + Background (DIS)

I will produce the results soon with this strategy

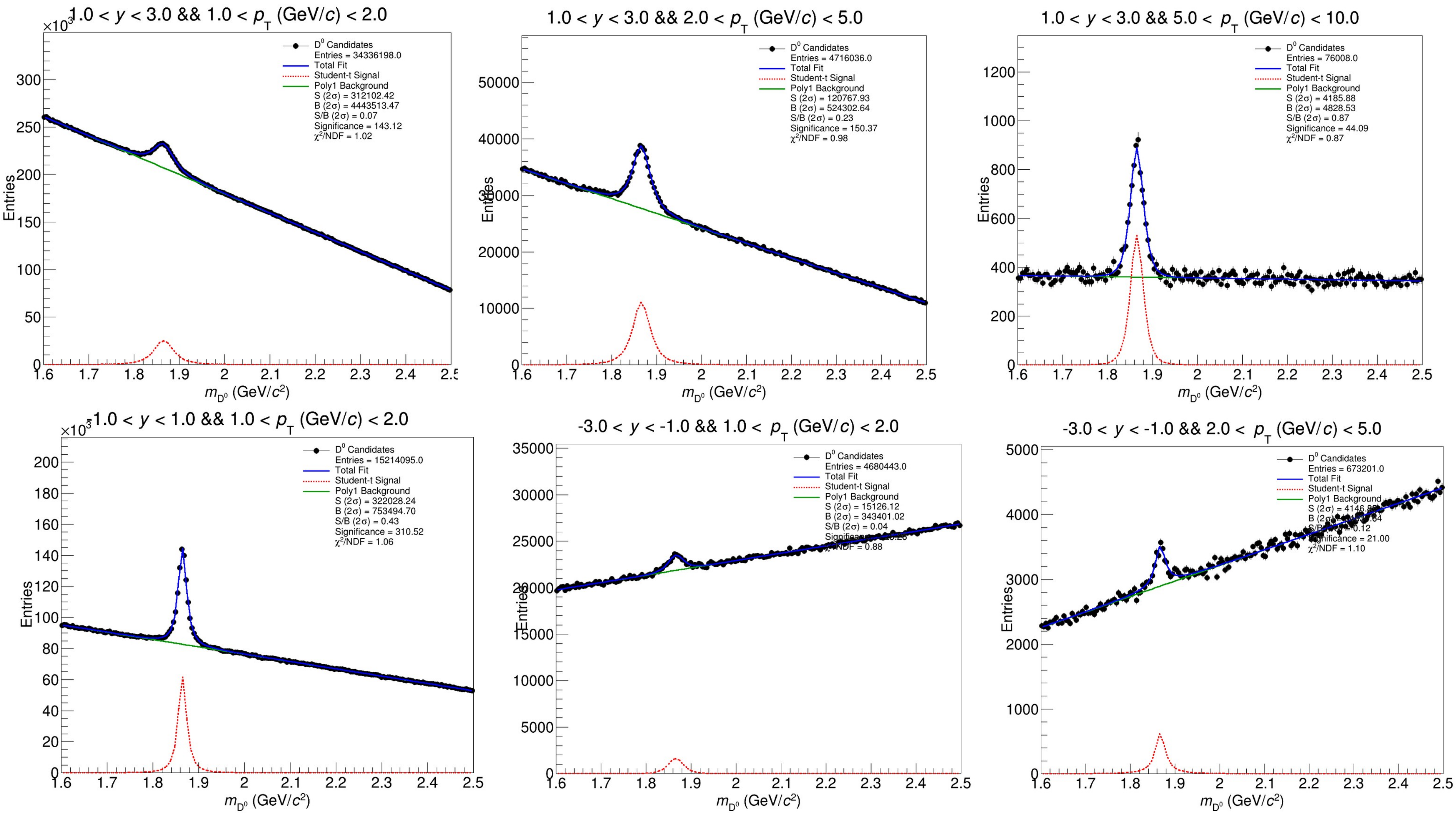
Comparison Distribution for D^0 meson (Q^2>1GeV^2)

All methods are compatible

Bkg D^0 meson



Merging Signal and Background



Secondary Vertex Reconstruction (Λ_c^+)

Secondary Vertex

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

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

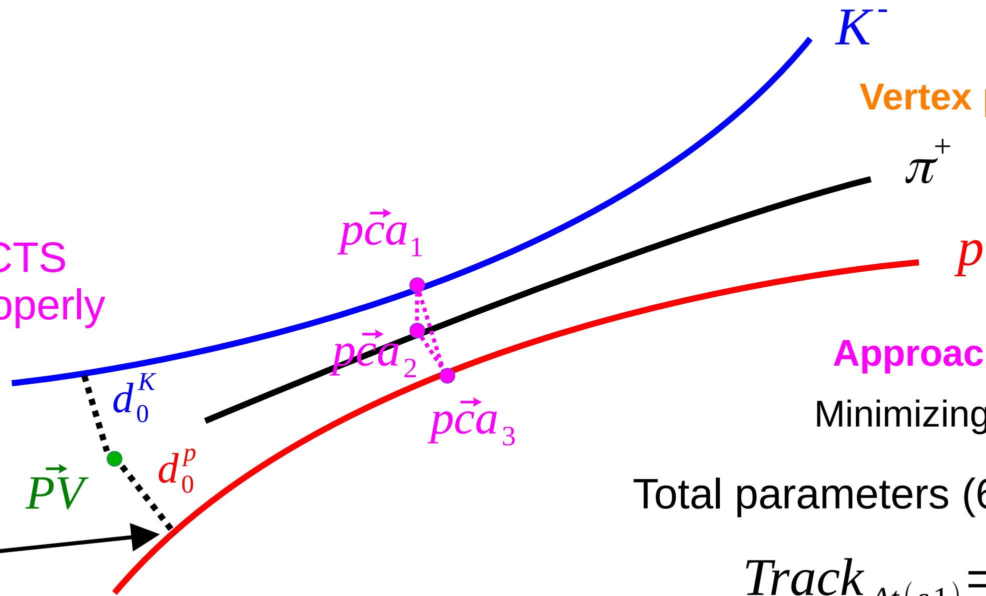
DCA_{12} = min {$DCA_{\pi K}$, DCA_{KP}, $DCA_{\pi P}$} Cut

Approach 1 (Shyam)

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

Secondary vertexing in ACTS considers tracking errors properly

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



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

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

Approach 2 (Shyam)

Minimizing the distance

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

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

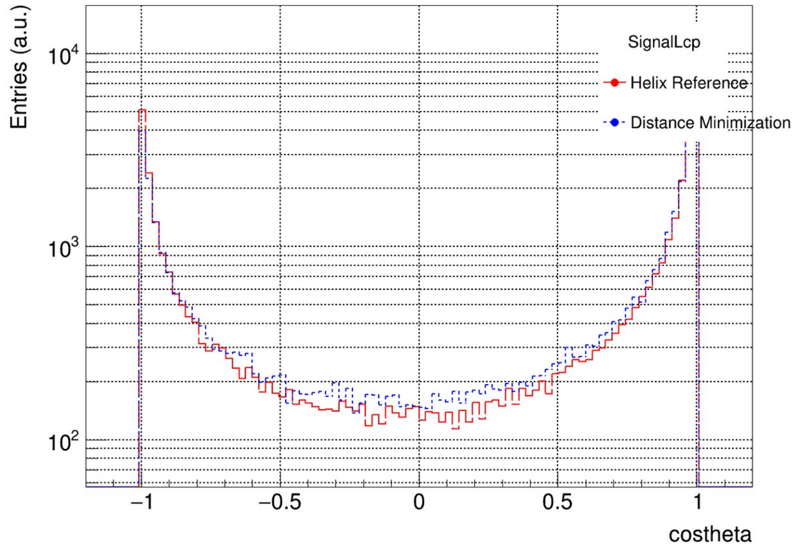
$$Track_{At(s2)} = (\vec{r}_{s2}, \vec{p}2, q2)$$

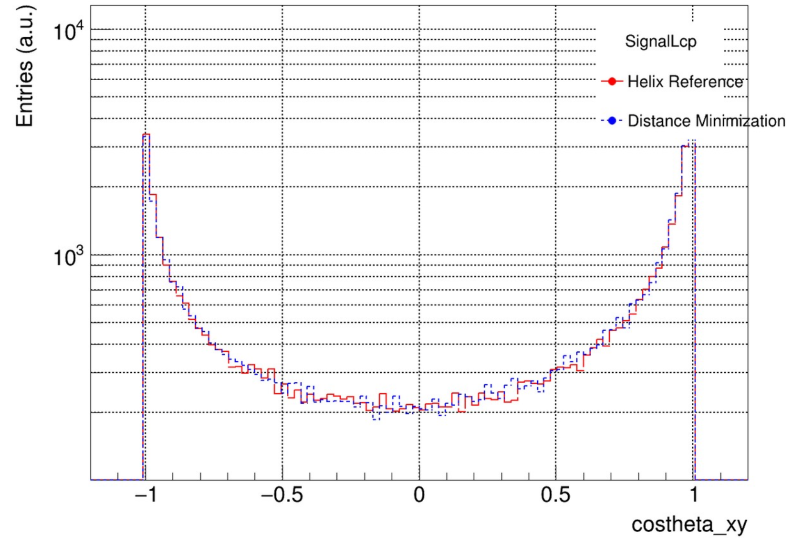
$$Track_{At(s3)} = (\vec{r}_{s3}, \vec{p}3, q3)$$

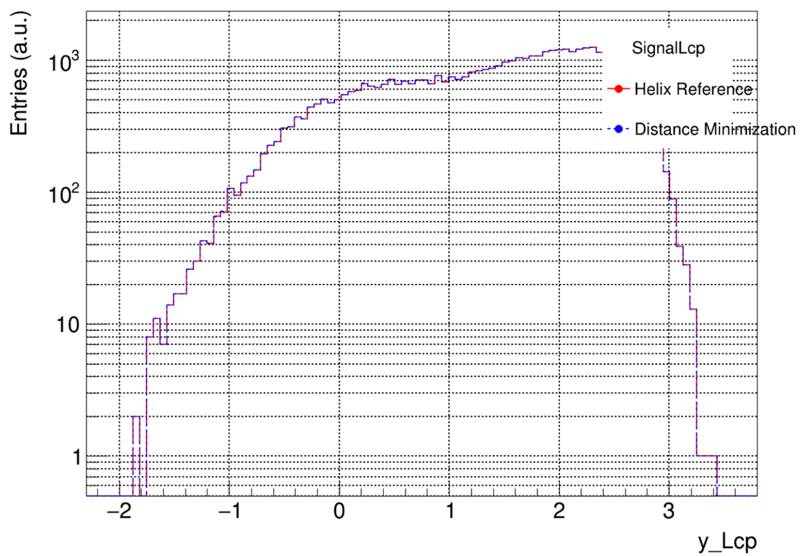
Minimize $$d = \sqrt{(\vec{r}_{s1} - \vec{v})^2 + (\vec{r}_{s2} - \vec{v})^2 + (\vec{r}_{s3} - \vec{v})^2}$$

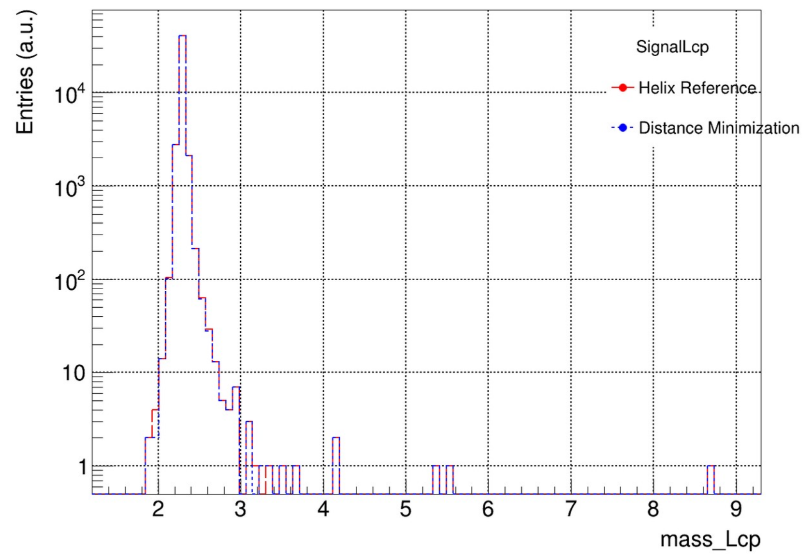
Comparison (Results)

Signal Λ_c^+



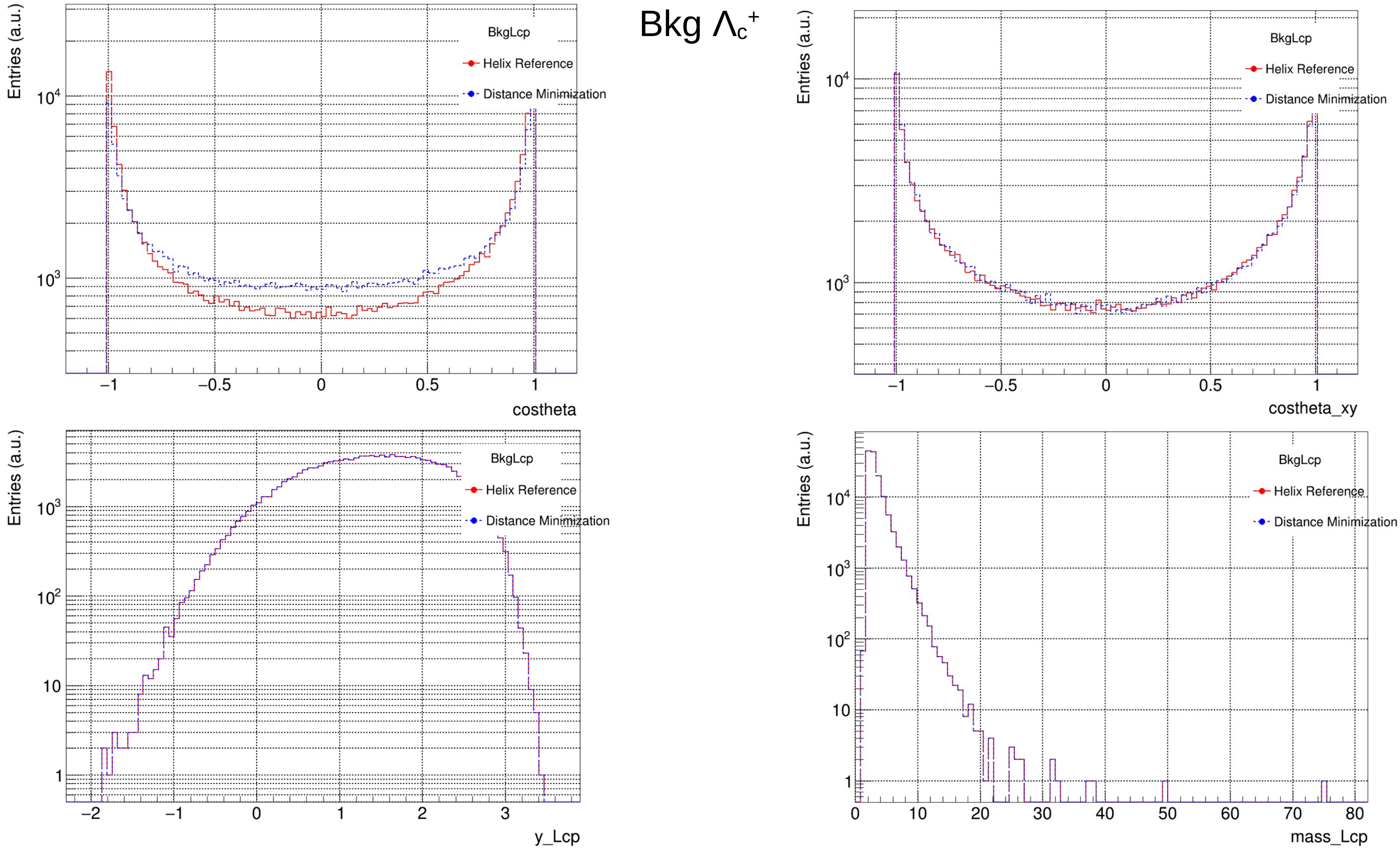






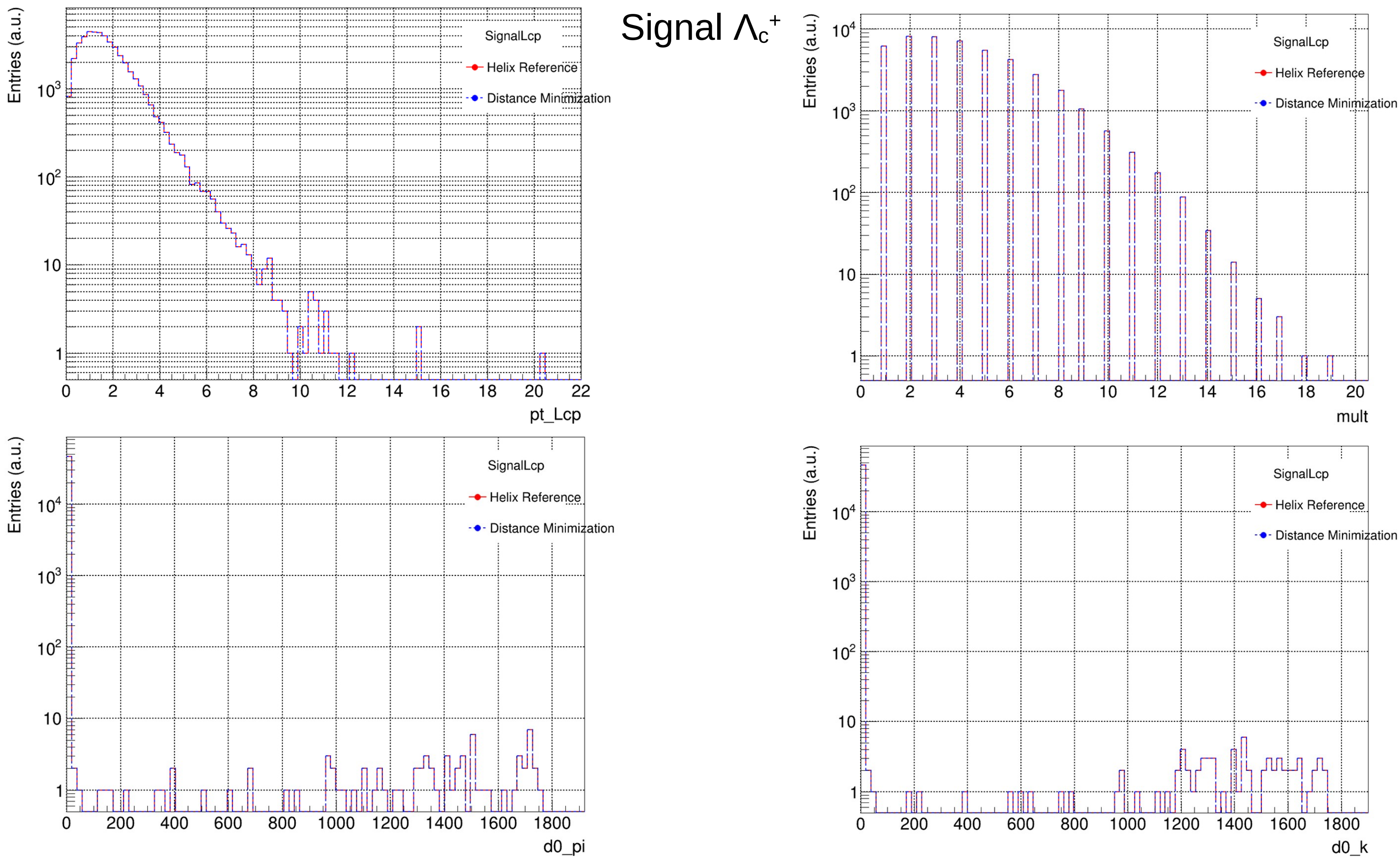
Comparison (Results)

Bkg Λ_c^+

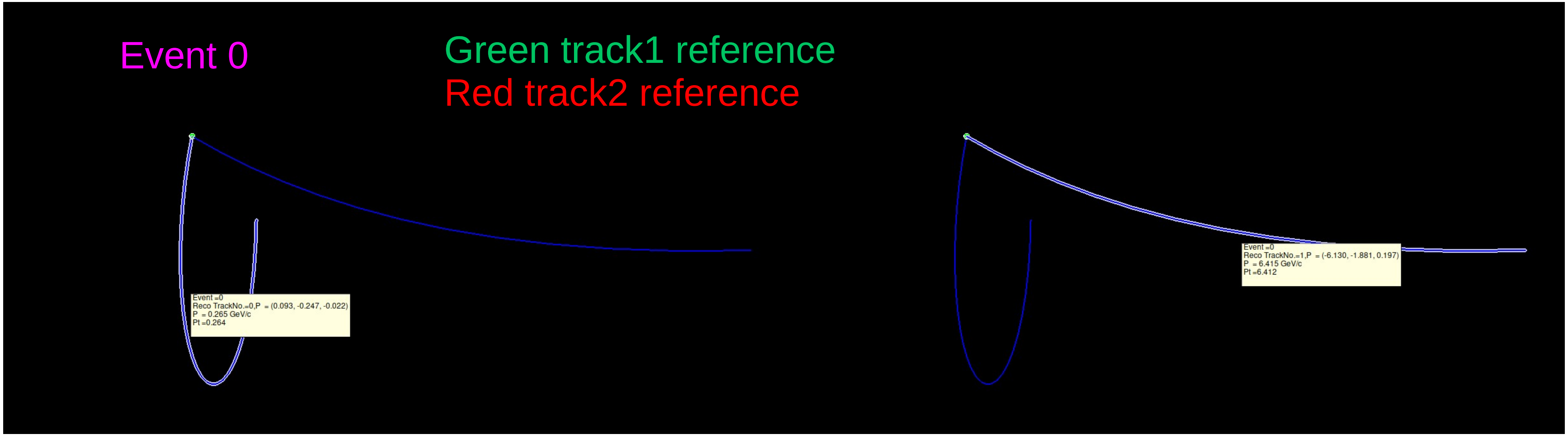


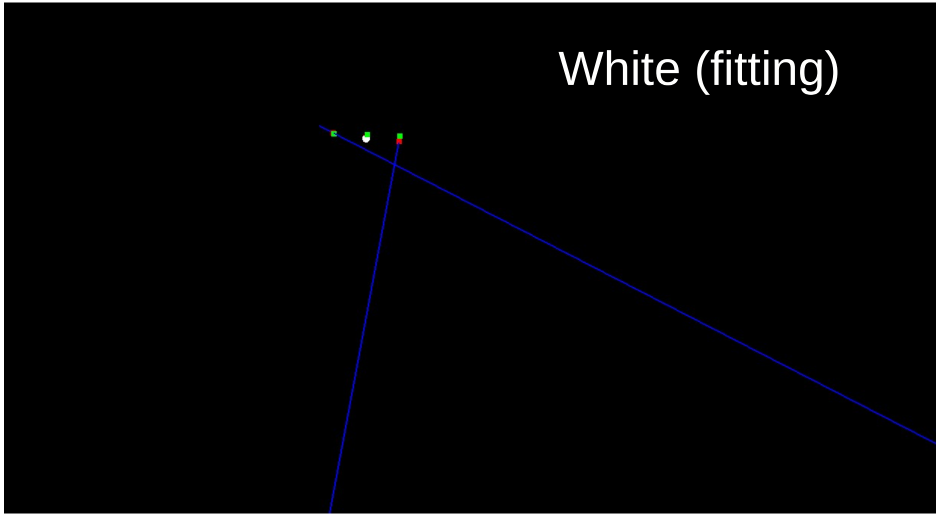
Comparison (Results)

Signal Λ_c^+



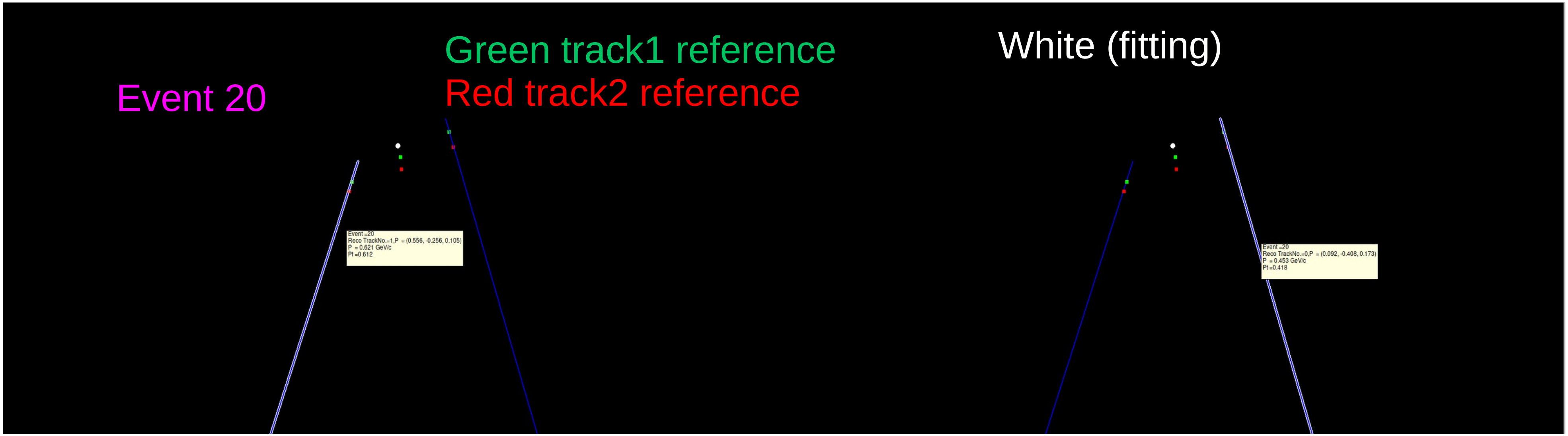
Event display (First two tracks)



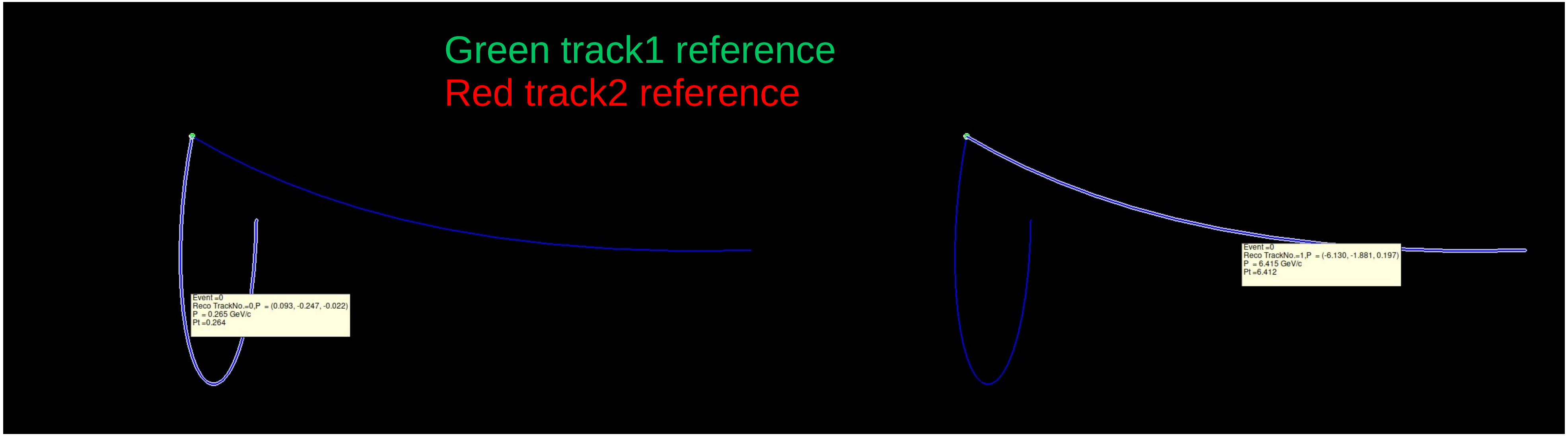


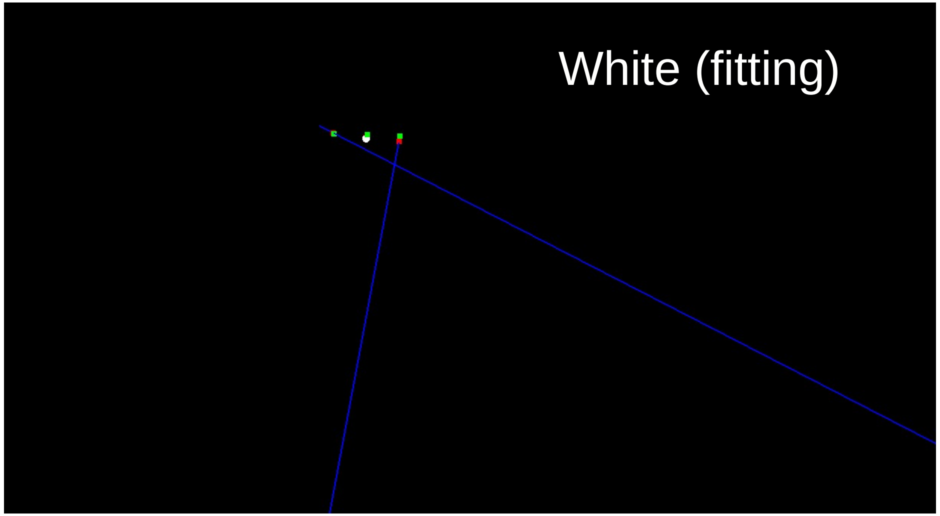
- All three methods (**track 1 reference, track2 reference, and distance minimization**) are compatible
- Minor difference is due to analytical approach
- Distance minimization returns the unique point

Event display (First two tracks)



Event display (First two tracks)





- 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

More hints on Hadronization: Fragmentation Fractions

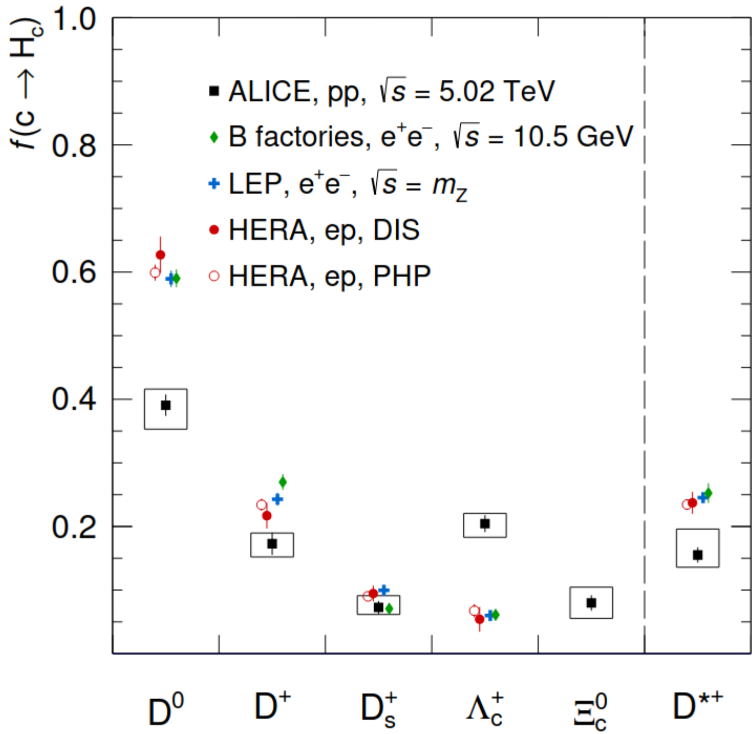
Violation of universality of fragmentation fractions (FF) already in pp collisions

➔ Cannot rely on e^+e^- FF to get charm cross section

$z = p_{HQ}/p_Q$

$$\frac{d\sigma^{NN\rightarrow H_Q X}}{dp_T^{H_Q}}(\sqrt{s_{NN}}, M_Q, \mu_F^2, \mu_R^2) = \sum_{i,j=q,\bar{q},g} \boxed{f_i(x_1,\mu_F^2) \otimes f_j(x_2,\mu_F^2)} \otimes \boxed{d\hat{\sigma}^{ij\rightarrow Q\bar{Q}}(\alpha_s(\mu_R^2), \mu_F^2, M_Q, x_1, x_2, s_{NN})} \otimes \boxed{D_Q^{H_Q}(z, \mu_F^2)}$$

Parton distribution functions **Hard Scattering Cross-section** **Fragmentation function**



Measure fragmentation fractions at ePIC with different nuclei systems for more understanding!!