Identification of Charm-Quark-Initiated Jets in Fast Simulation using the ATHENA Experiment at the Electron-Ion Collider

Justine Choi, Stephanie Gilchrist, Stephen Sekula

September 3, 2021

1 Introduction

This note summarizes work done during the summer of 2021 by the SMU ATHENA group. The work is focused on the identification of charm-quark-initiated jets ("charm jets") using tracking and calorimetry information. The scattering of an electron by a proton is simulated using the charged-current (CC) deep-inelastic scattering (DIS) process at a center-of-mass energy of $\sqrt{s} = 105\,\text{GeV}$. Charm jets are initiated by various processes, notably the process $\bar{s} + W^- \rightarrow \bar{c}$. This process can be used to probe the intrinsic strange quark content ("strageness") of the proton.

A range of approaches are used to identify charm jets, including generic displaced track counting, particle-identification-based (PID-based) approaches utilizing kaons and electrons, and multivariate approaches to combine such information.

2 EIC and ATHENA Simulations

The collisions at the EIC are simulated using Pythia8 configured with the high-acceptance beam spot model. The proton's parton distribution function (PDF) is set to CT18NNLO. This results in a beam spot size consistent with the EIC Yellow Report configuration for the collider, with a transverse beam profile that is $\mathcal{O}(10)$ μm in size and a longitudinal beam profile that is $\mathcal{O}(1)$ cm in size. The "forward" direction points in the direction of the hadron beam and defines the positive z axis. The positive x axis points into the center of the EIC collider ring and the positive y direction points straight up. The common coordinate system for collisions is cylindrical and employs p_T, η (pseudorapidity), and ϕ azimuthal angle about the z axis to describe particle kinematics.

The forward direction of the ATHENA detector is designed for high-acceptance and efficiency for the hadronic components of the collisions (which tend to travel in the direction of the hadron beam), the barrel is designed to reconstruct both hadronic and leptonic components (including good missing energy reconstruction), and the backward region is designed to accept and reconstruct well the scattered beam lepton.

The ATHENA detector is modeled in one of two ways. Unless otherwise specified, the primary modeling is performed using a fast simulation framework (Delphes) configured using the `delphes_EIC` project, specifically the `delphes_card_allsilicon_3T.tcl` model. This implements

a tracking volume with acceptance, resolution, and efficiency consistent with a proposed all-silicon tracker design. The calorimeter consists of two systems, an electromagnetic calorimeter (ECal) and a hadronic calorimeter (HCal) with energy resolution consistent with the EIC Yellow Report targets. In addition, the response of dedicated PID systems (a backward mRICH, a barrel DIRC, and a forward dual RICH) for kaons and pions is simulated using efficiency maps based on the EIC User Group PID group parameterization developed for the EIC Yellow Report.

Post-Delphes work is conducted on Delphes objects using the OLeAA project (https://github.com/stephensekula/OLeAA). This results in compact ROOT files that can be used in Python- or ROOT-based analyses.

3 Event Selection

We select e-p collisions that meet the following criteria:

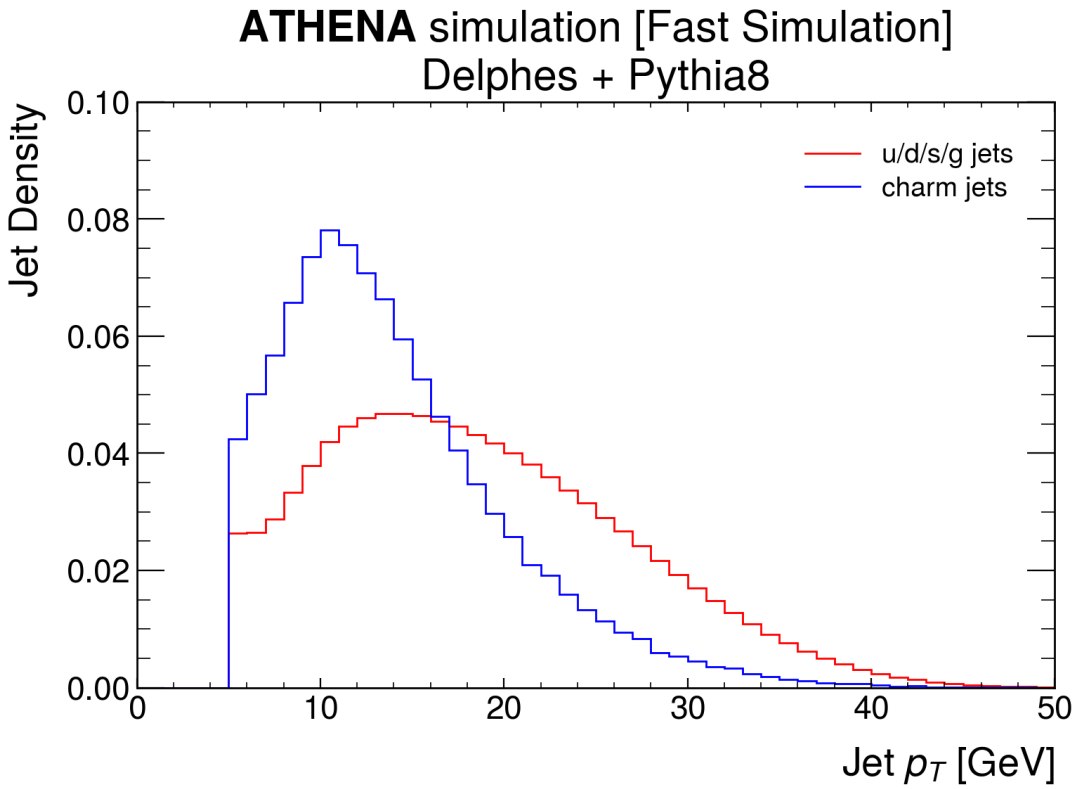
- The event missing transverse energy E_T^{miss} exceeds 10 GeV;
- The event contains at least one R=1 anti-k_T jet that has $p_T > 5\,\text{GeV}$ and $|\eta| < 3.0$.

The efficiency with which we select any CC DIS event (CC DIS events containing a true charm jet) is 59% (53%). The E_T^{miss} requirement is 86% (78%) efficient on all (charm) events. The jet requirement, relative to the E_T^{miss} requirement, is 68% (68%) efficient.

4 Charm and Light Jets

Charm and light jets are produced by different processes at the EIC and this results in general differences between the two super-classes. Light jets consist of those jets initiated by a range of quarks, including up and down, strange, and also by the gluon. Charm jets (and bottom-quark-initiated jets) are generally distinguished by the presence of decays interior to the jet structure that are displaced from the origin of the jet (the interaction point). For example, charm hadrons have lifetimes that range from $c\tau_{charm} \approx 200 - 500\,\mu\text{m}$. Such hadrons are generally further Lorentz-boosted by the collider and the subsequent jet hadronization process, resulting in decay lengths ranging from hundreds to thousands of microns.

Other particles, such as K_s^0 and Λ, readily produced in strange and lighter parton jets, have significant lifetimes and can fake the displaced decays associated with charm jets. These long-lived light hadron decays are simulated as part of the background.



Charm jets in this simulation are produced primarily by the target process, $\bar{s} + W^- \rightarrow \bar{c}$. This relies on converting sea strange quarks to charm quarks. In addition, since kinetic energy from the interaction is lost to the mass of the charm quark (which is about 4-5 times heavier than the strange quark), one expects the spectrum of the charm jets to be softer than the light jet spectrum. Light jets are primarily initiated from processes like $u + W^- \rightarrow d$ (e.g. converting a valence quark) or by interactions with the light sea; combined with the fact that such quarks have low masses and are sourced from partons that carry a significant fraction of the proton's momentum, this leads to a harder light-jet spectrum. This is observed in simulation.

5 Displaced Tracks in Jets

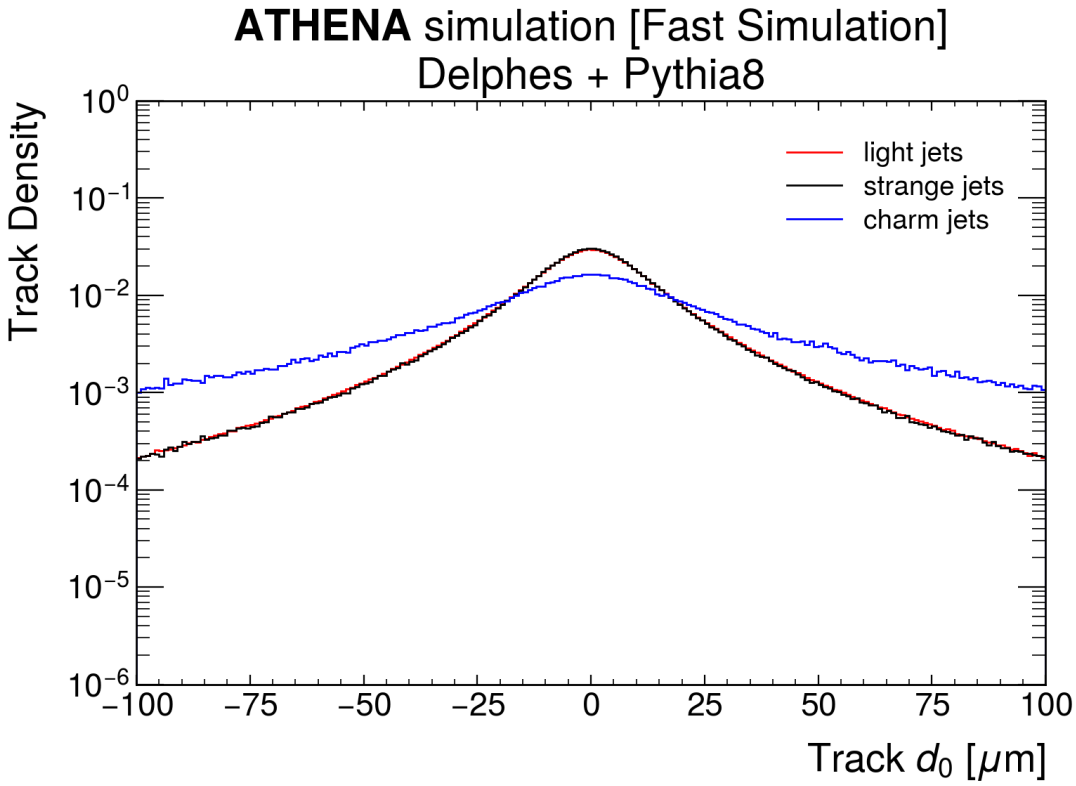
Displaced decay vertices will also necessarily lead to tracks that are displaced from the origin of the jet. A simple approach to identifying heavy flavor jets with displaced hadronic decays is to count significantly displaced tracks and place a minimum requirement on the number and quality of such tracks. The magnitude of the distance of closest approach to the interaction point, computed in either the transverse plane, the longitudinal direction, or by combining all three dimensions, contains information about the presence of such decays.

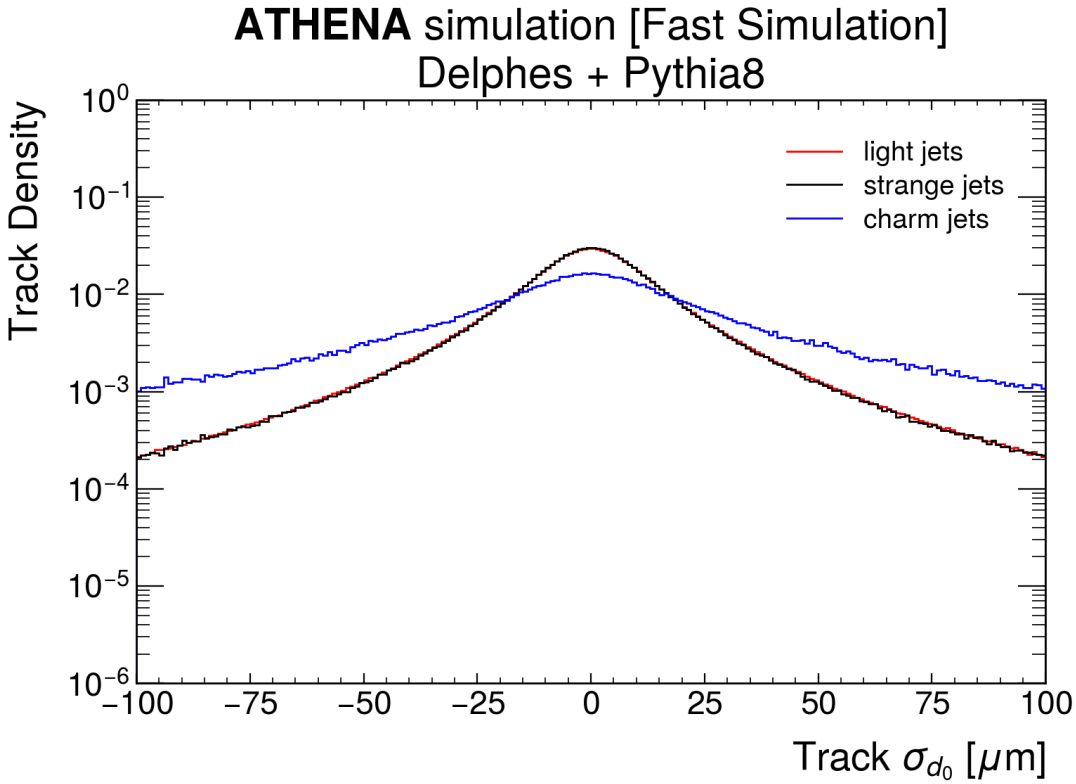
The variables used in this approach to tagging heavy flavor jets are as follows:

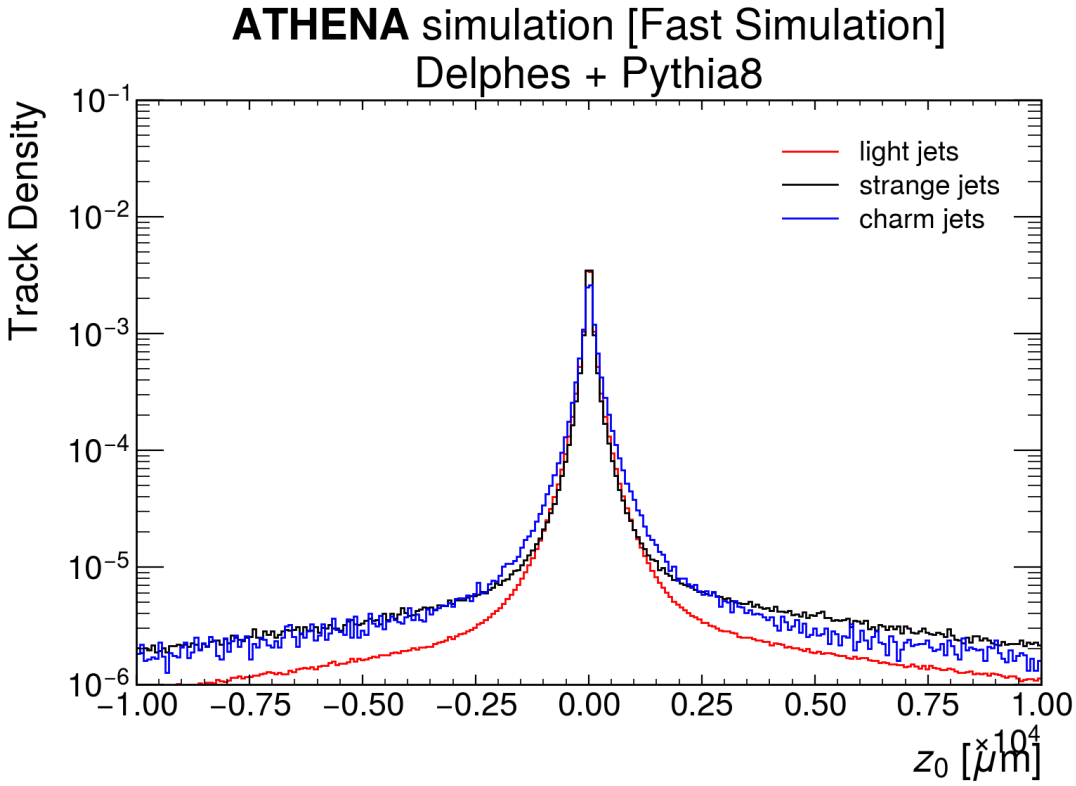
- d_0: the magnitude of the distance of closest approach in the transverse plane between a track and the IP;

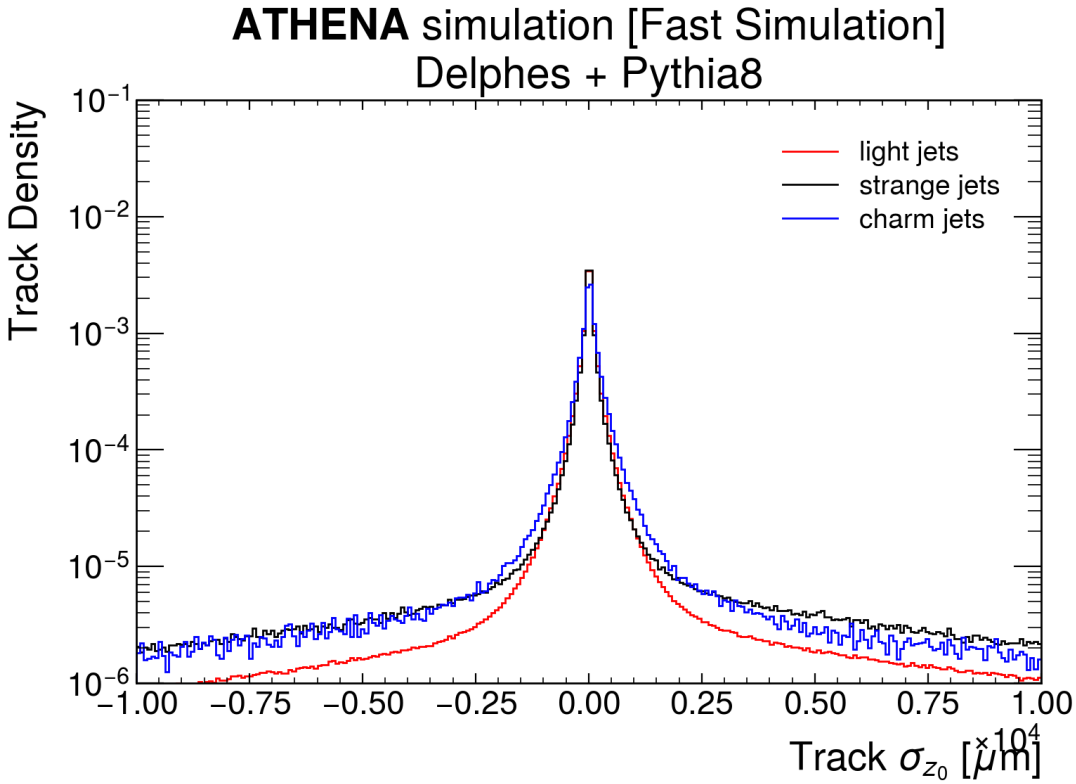
- z_0: the magnitude of the distance of closest approach in the longitudinal direction between a track and the IP;
- IP_{2D}: the significance of the track displacement in the transverse plane, defined as $IP_{2D} \equiv d_0/\sigma_{d_0}$;
- IP_{3D}: the significance of the track displacement in all three dimensions, using $r_0 = \sqrt{d_0^2 + z_0^2}$ and thus $IP_{3D} \equiv r_0/\sigma_{r_0}$;
- sIP_{3D}: the signed 3-D impact parameter significance, where the sign multiplying IP_{3D} is determined from the dot product $\vec{j} \cdot \vec{r}_0$ where $\vec{j}$ is a unit vector pointing parallel to the axis of the jet containing the track and $\vec{r}_0$ is a vector pointing from the IP to the point of closest approach on the track.

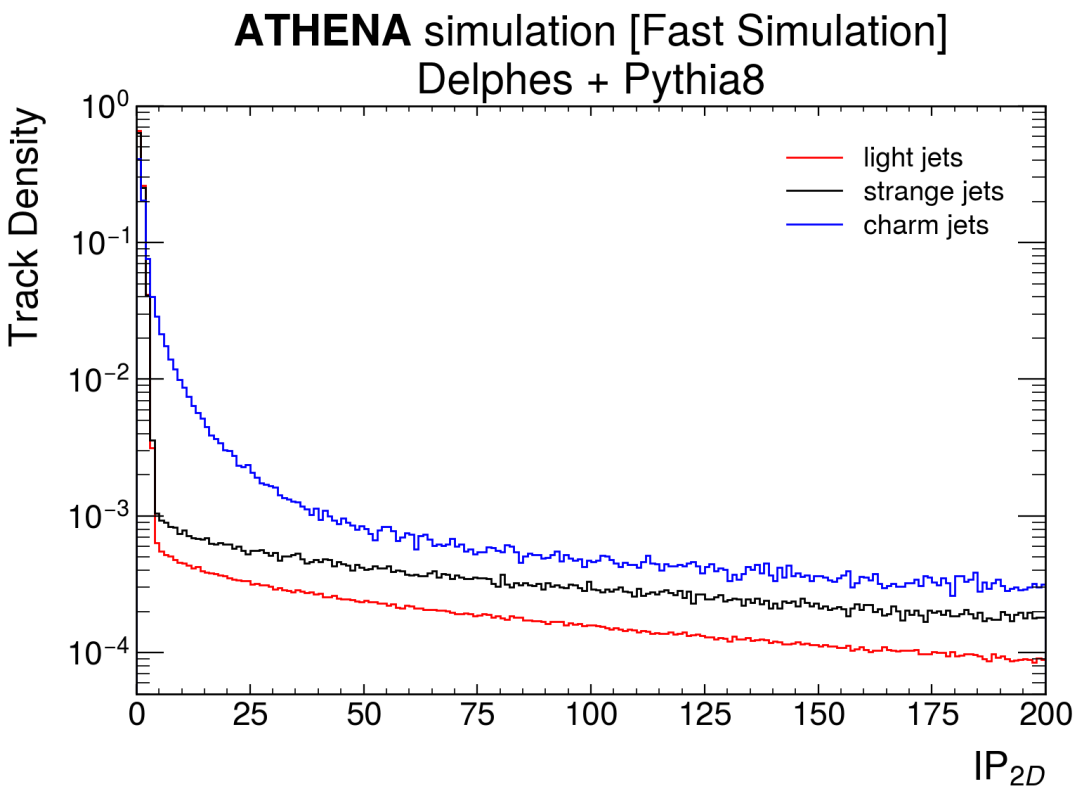
We show in the next few figures the distributions of these quantities.

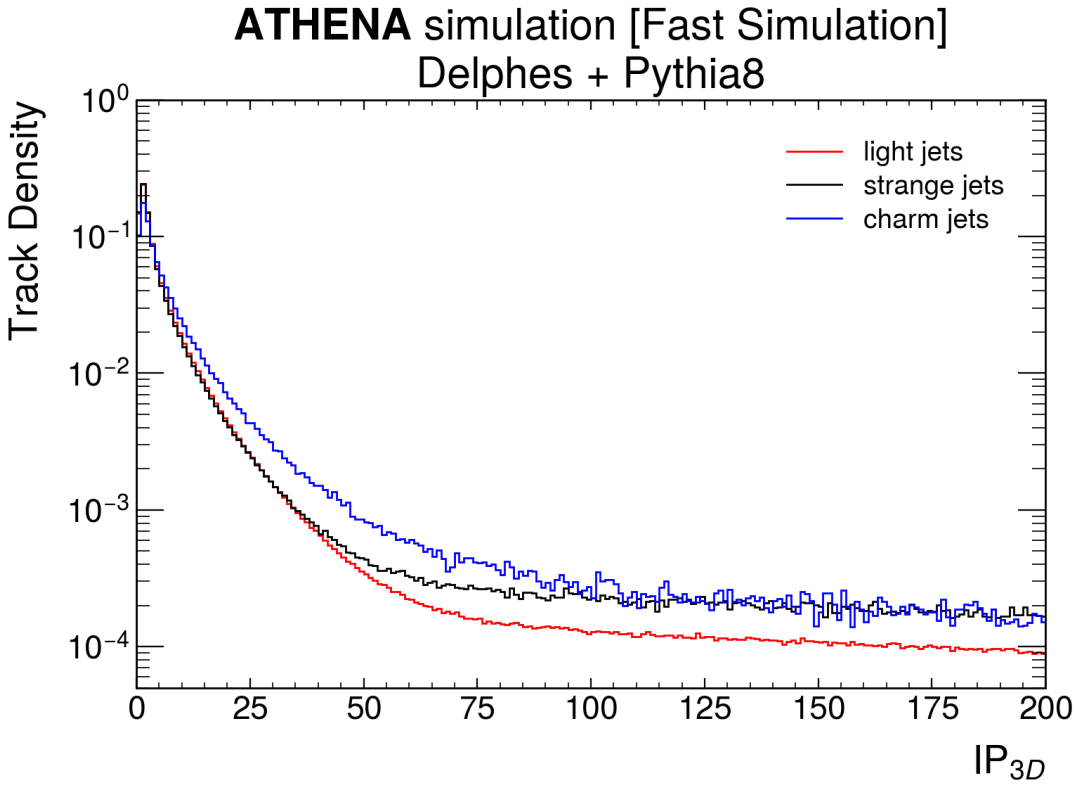


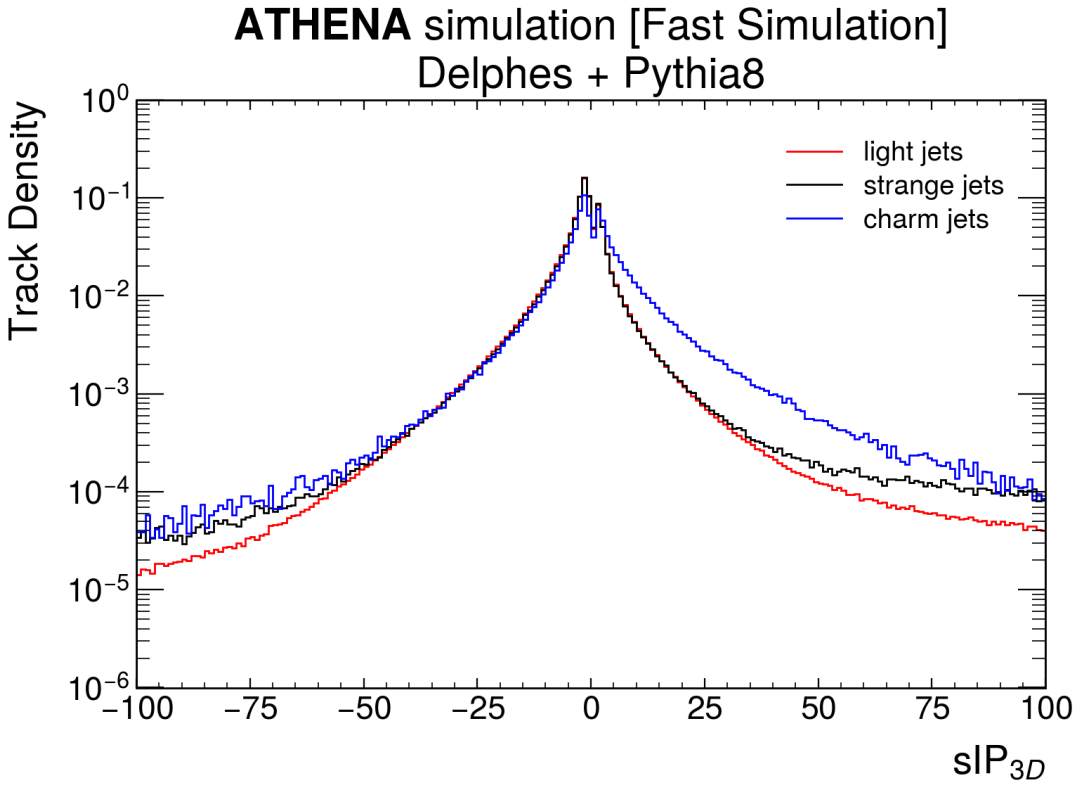

ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Track Density
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
light jets
strange jets
charm jets
−100
−75
−50
−25
0
25
50
75
100
Track σ_{d_0} [μm]


ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Track Density
light jets
strange jets
charm jets
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
×10^4
z_0 [μm]

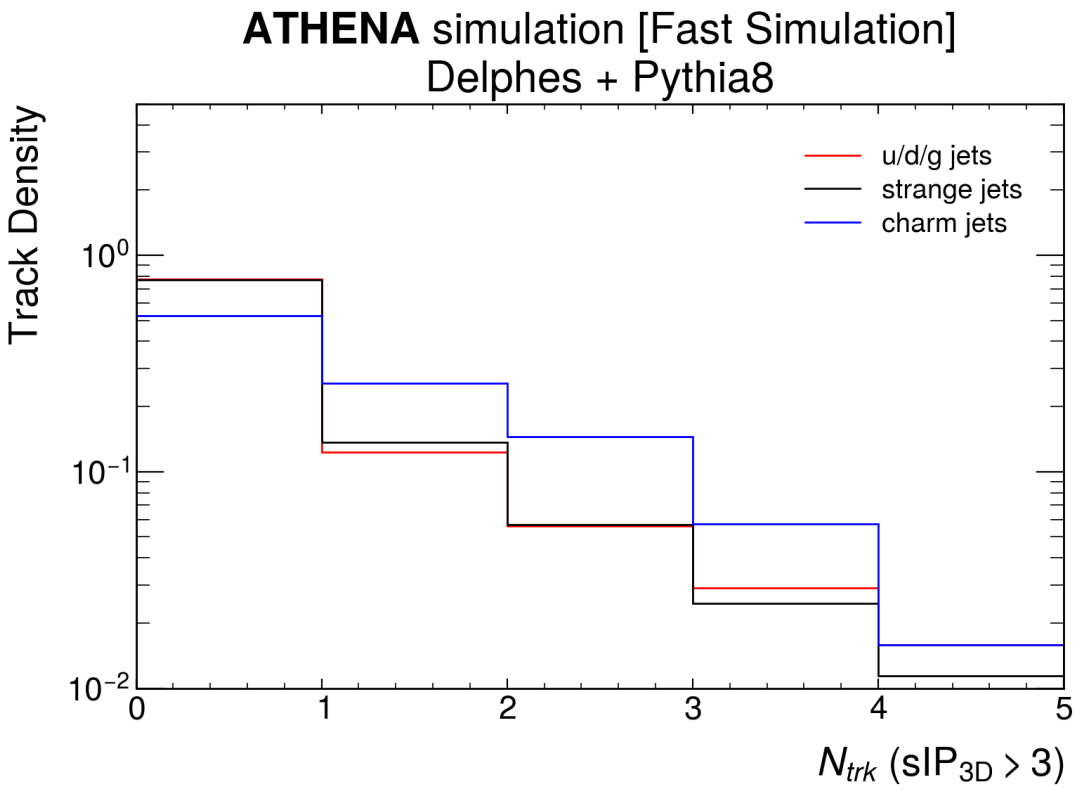

ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Track Density
light jets
strange jets
charm jets
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
$\times 10^{4}$
Track σ_{z_0} [μm]


ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Track Density
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
light jets
strange jets
charm jets
0
25
50
75
100
125
150
175
200
IP_{2D}


ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Track Density
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0
25
50
75
100
125
150
175
200
IP_{3D}
light jets
strange jets
charm jets



Given the higher mass of the charm quark, we would also expect a higher multiplicity of final-state particles in charm jets. We focus on the multiplicity of tracks with significant displacement ($sIP_{3D} > 3.0$) and compare the multiplicity (up to four such tracks) in u/d/g jets, strange jets, and charm jets. Indeed, we see that charm jets overwhelmingly tend to produce more significantly displaced tracks.



5.1 Cut-based displaced track tagging

A simple approach to displaced track tagging in jets is taken by combining the following pieces of information for the four leading (by p_T) tracks in each jet:

- The track p_T;
- The track IP_{2D}.

The counting of significantly displaced tracks is done by determining how many tracks in a jet pass the minimum criteria on each of these quantities. We optimize the selection on these hyper-parameters by scanning over combinations of criteria and identifying the one that minimizes the relative error on a final predicted charm jet yield in 100 fb^{-1} of data, σ_{charm}/n_{charm}. The results of this optimization are as follows:

- The track $p_T \geq 0.5$;
- The track $IP_{2D} \geq 3.0$;
- The number of such displaced tracks $n_{trk} \geq 3$.

As a baseline, we adopt the above requirements to define a cut-based "CharmIPXDTagger" determination. The output of this algorithm is 1.0 (true) if the above criteria are met. Based on the selection of jets using this definition of CharmIPXDTagger, we find the charm (light) jet efficiency to be 17% (0.86%). The expected yield of charm (light) jets in 100 fb^{-1} is 3.6×10^3 (7.3×10^3).

5.2 Multivariate-based displaced track tagging [BROKEN/IGNORE]

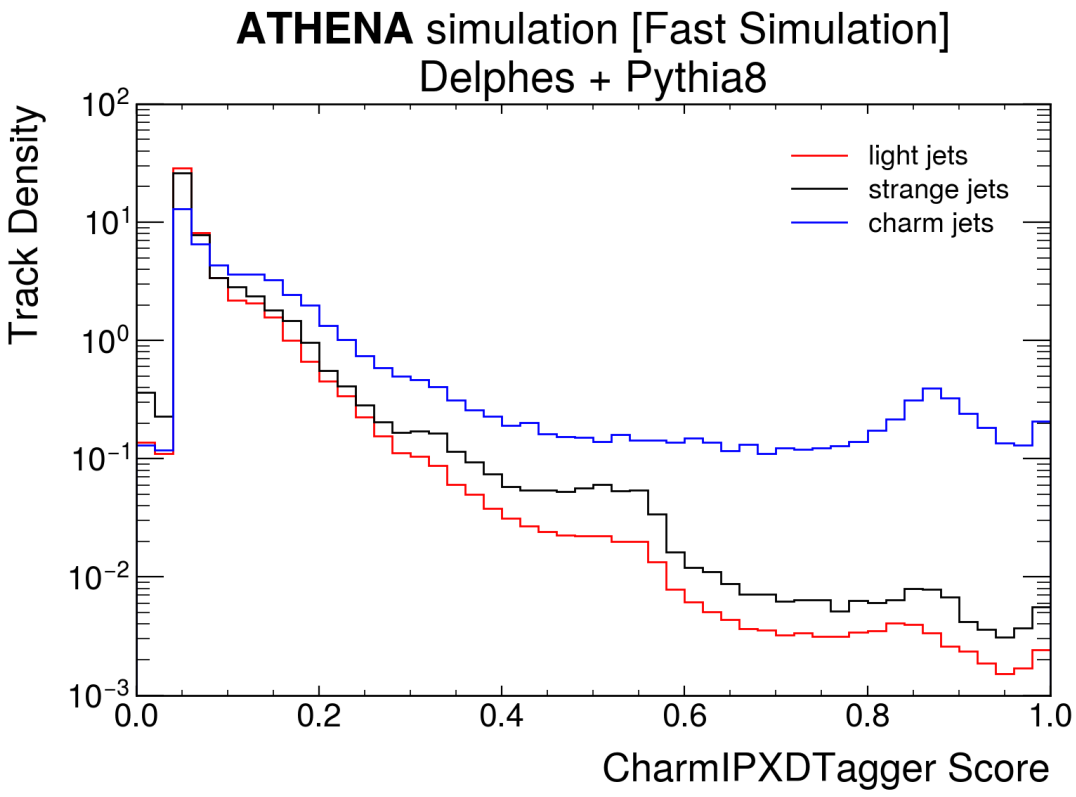
NOTE: the neural network trained for this section was flawed due to an error in the training process. It preferentially wound up selecting jets with no tracks as a result. This section should be ignored for now. The cut-based definition is used for the rest of this note.

An estimate of the level of improvement we might expect from a full multivariate tagging method (incorporating potential correlations between variables) is obtained by combining displacemnt-sensitive variables in a neural network. The network is a simple feed-forward multi-layer perceptron (MLP). The leading four tracks (based on p_T) are selected as sources of inputs for the MLP. For each track, the IP_{2D}, IP_{3D}, and sIP_{3D} variables are selected. This yields 12 total inputs to the MLP.

The MLP is architected with a single hidden layer of 36 nodes. The activation function for the hidden layer nodes is the rectified linear unit (ReLU). All inputs are transformed to have a consistent range between $[0, 1]$. A dedicated Monte Carlo sample is generated to perform the training, independent of any sample used henceforth to state performance results. A total of 10,000 (100,000) charm (light) jets are used in the training process, and 1000 epochs over the training sample are used to optimize the neural network. A single output node yields a score between 0 (light jet) and 1 (charm jet).

Bias in the training is assessed by comparing the MLP output shape for charm and light jets between the training and the testing samples used in the training process. The testing sample consistents of an independent set of jets equal in number to those used in the traininig sample. A Kolmogorov-Smirnov test of the compatibility of the training and testing MLP shapes suggests no significant evidence of over-training (e.g. the shape of the MLP in the testing sample can be explained by statistical fluctuations in the shape of the MLP in the training sample).

The output of the MLP is shown for light-flavor (u/d/g), strange, and charm jets. This MLP is henceforth referred to as the CharmIPXDTagger. A cut on the output at $>$0.58 is found to optimally select charm jets over light jets. This optimization is determined by minimizing the statistical uncertainty on the final yield of charm-tagged jets, in a target luminosity of $100\,\text{fb}^{-1}$, if we assume a perfect subtraction of all contaminating light jets by some future background subtraction method.



Based on the selection of jets using CharmIPXDTagger, we find the charm (light) jet efficiency to be 17% (0.86%). The expected yield of charm (light) jets in 100 fb^{-1} is 3.6×10^3 (7.3×10^3).

6 Kaon Tagging

Kaons are strongly associated with charm hadron decay. While not, by themselves, a unique indicator (e.g. strange jets can also produce kaons), particle identification of kaons combined with displaced track approaches at the single-track level should provide the basis for flavor discrimination of jets.

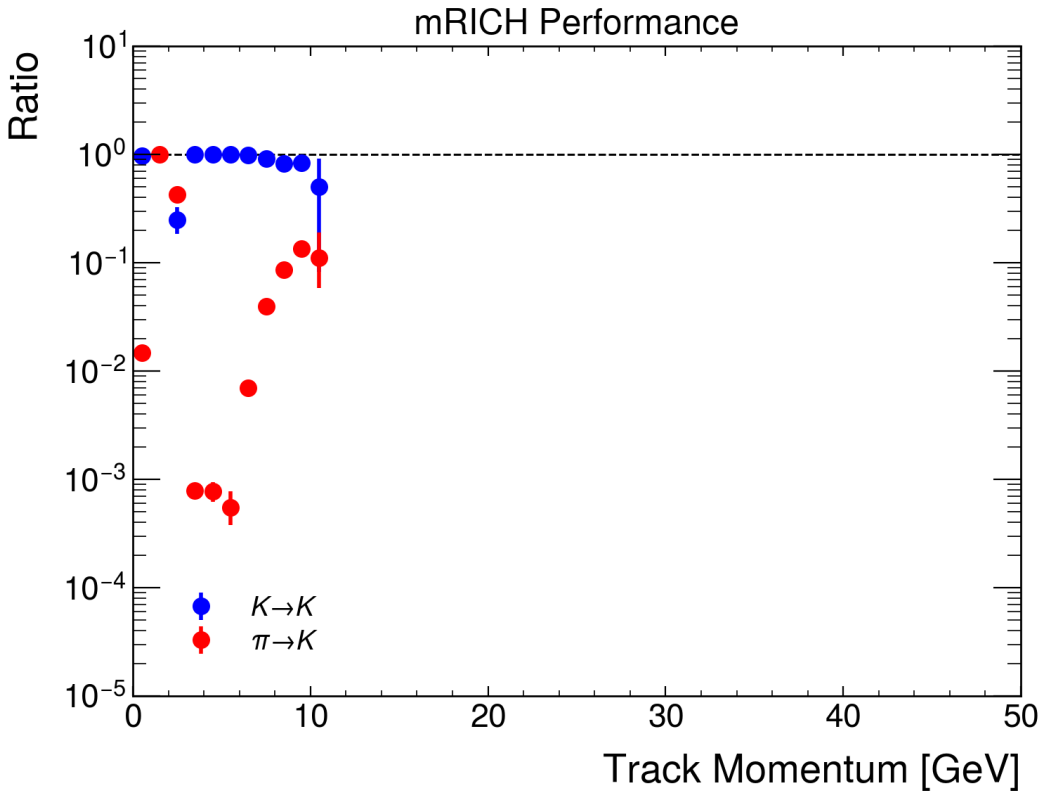
The rate at which charm hadrons are expected to produce at least one kaon is large. For example $D^{\pm}$ (D^0) mesons decay inclusively to $K^{\pm} + X$ about 25% (55%) of the time. In addition, the electric charge of the kaon should carry flavor information. For example, in the process $e^- p^+ \to \bar{c} + X$ the sign of the beam lepton (e^-) is required, by the electroweak interaction, to be the same as the final-state charm quark (assuming $\bar{s} + W^- \to \bar{c}$ dominates this production). The $\bar{c}$ quark will form either D^- or $\bar{D}^0$ mesons during hadronization. In turn, these will dominantly produce K^+ mesons in decays to kaons. Therefore, the sign of the beam lepton and the sign of the final-state kaon will be correlated, providing handles on assessing jet flavor.

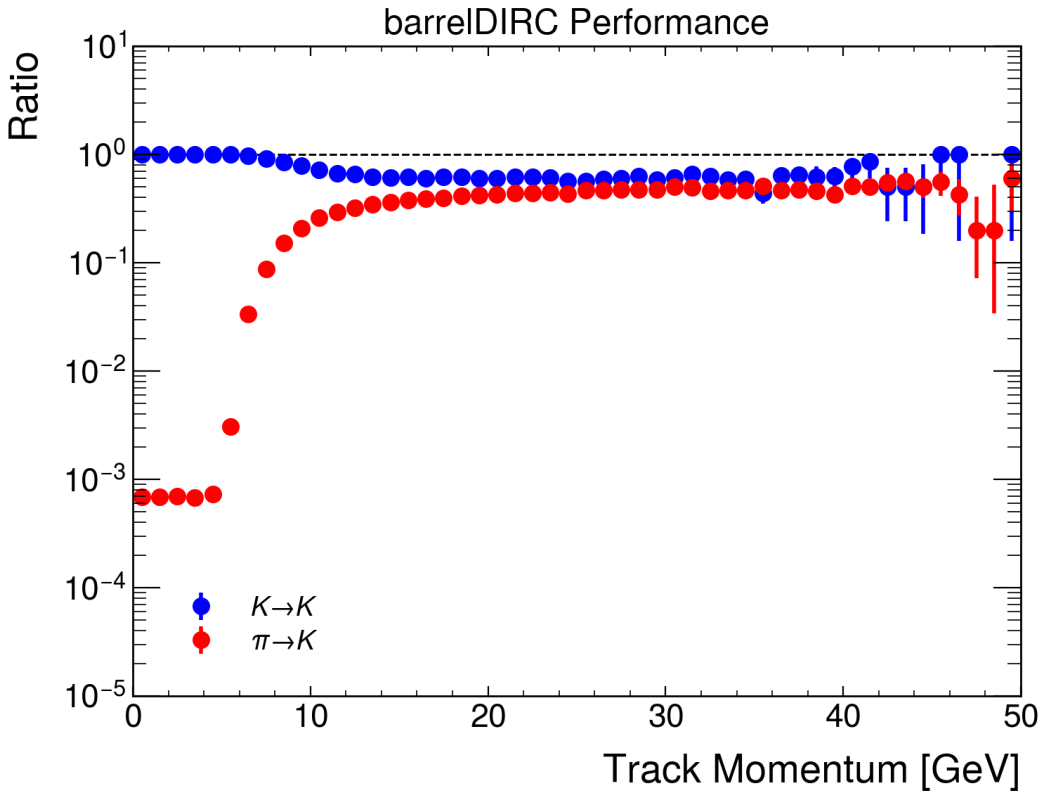
We here study kinematic, displacement, and charge information and combine that into a "K-Tagger" approach that relies solely on identified kaons to assess jet flavor.

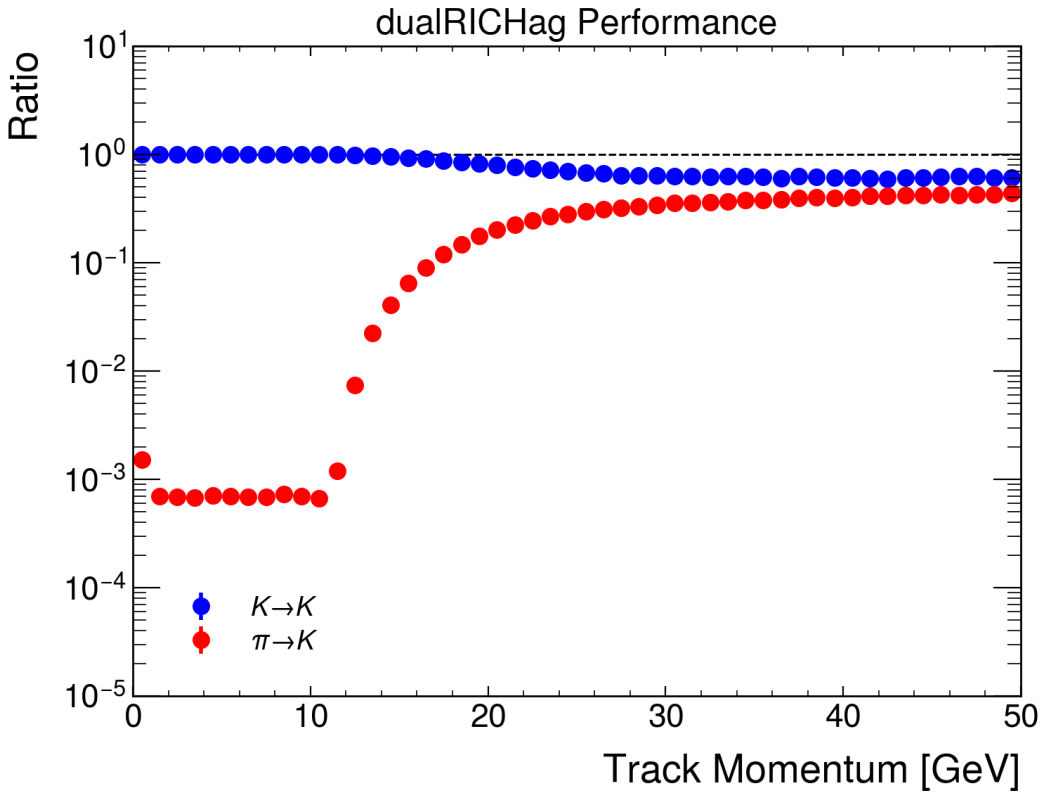
6.1 Kaon PID in ATHENA

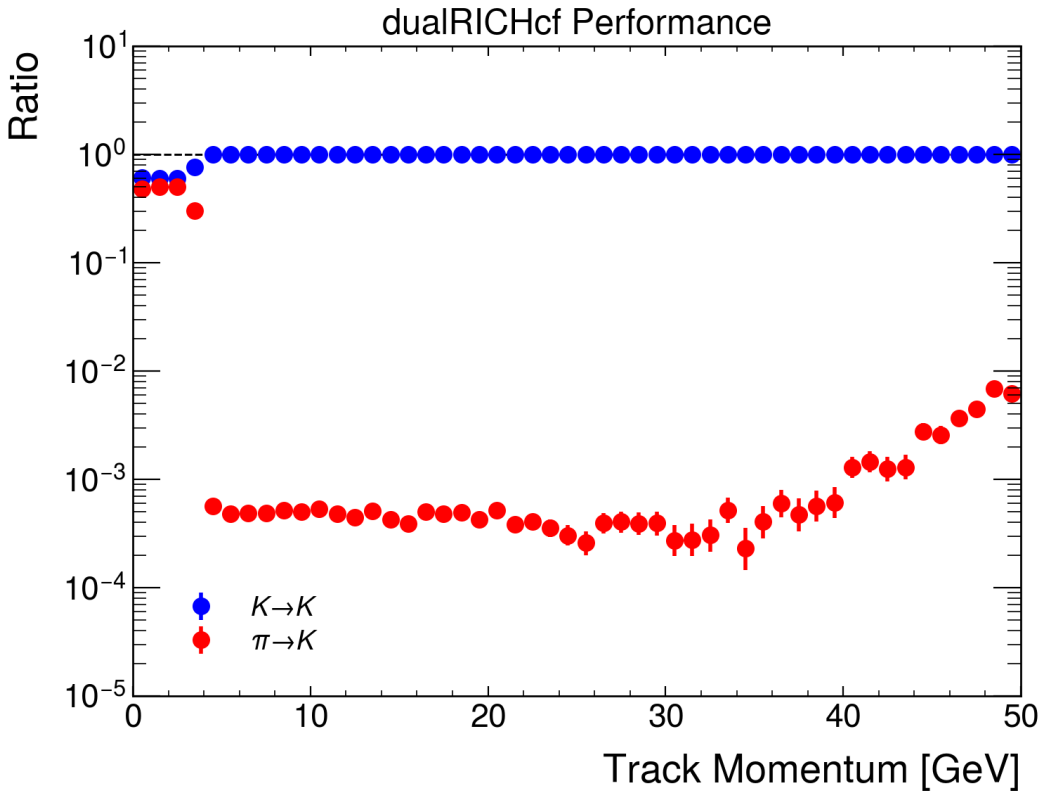
Particle identification systems are implemented in Delphes as identification maps encoding the probability that a track with truth identity H_0 will be identified as PID hypothesis H_1. For now, these systems only encode the probability of identifying charged kaons, pions, and protons. The information used to build these maps comes from the EICUG PID Group and was input for the EIC Yellow Report. Full simulation of realistic PID systems, including an mRICH in the backward direction ($-3.5 < \eta < -1.0$), a barrel DIRC ($-1.0 < \eta < 1.0$), and a forward-direction ($1.0 < \eta < 3.5$) dual RICH composed of aerogel (low-momentum PID) and C_2F_6 (high-momentum PID), is in progress in the collaboration.

The performance of the mRICH, barrel DIRC, and dRICH components in Delphes is illustrated below in a set of efficiency plots for true kaons and true pions.

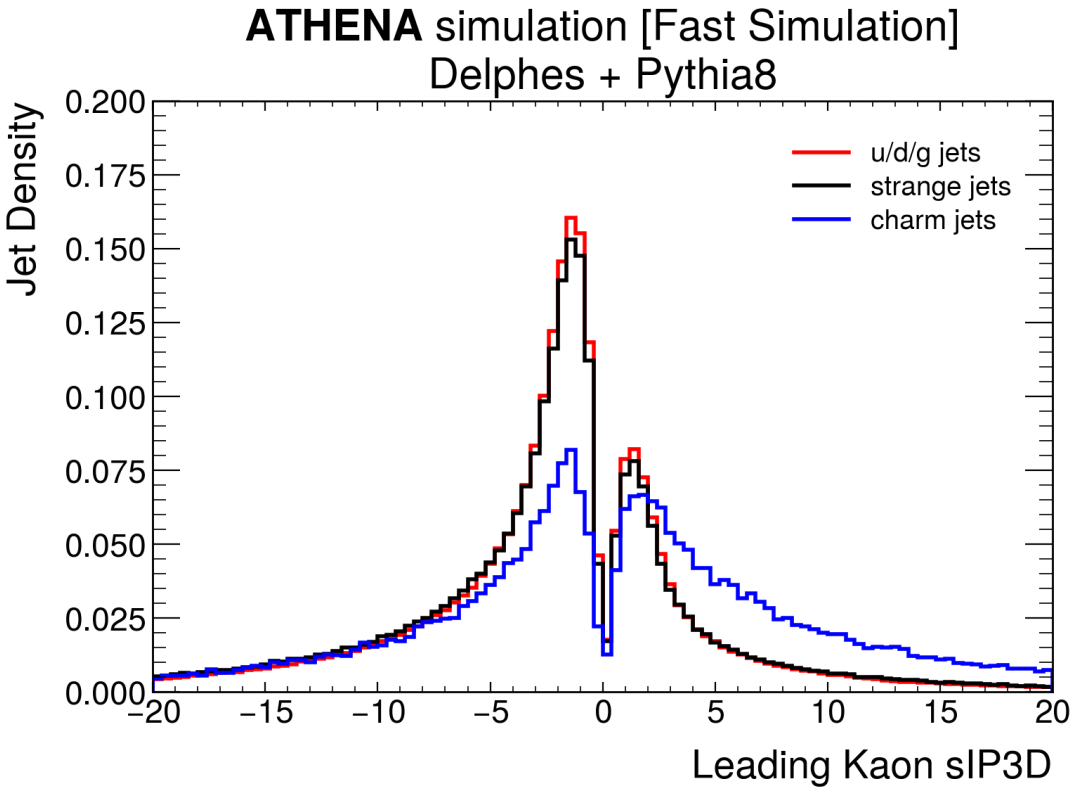


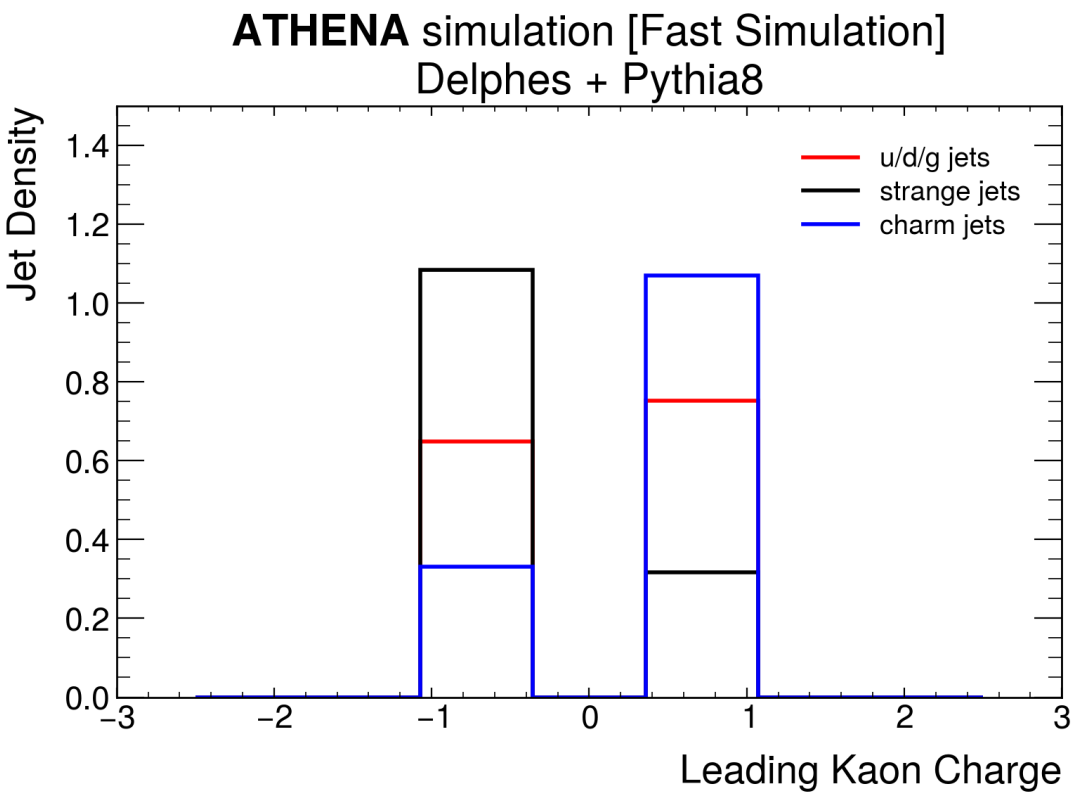

barrelDIRC Performance
Ratio
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
0
10
20
30
40
50
$K \to K$
$\pi \to K$
Track Momentum [GeV]

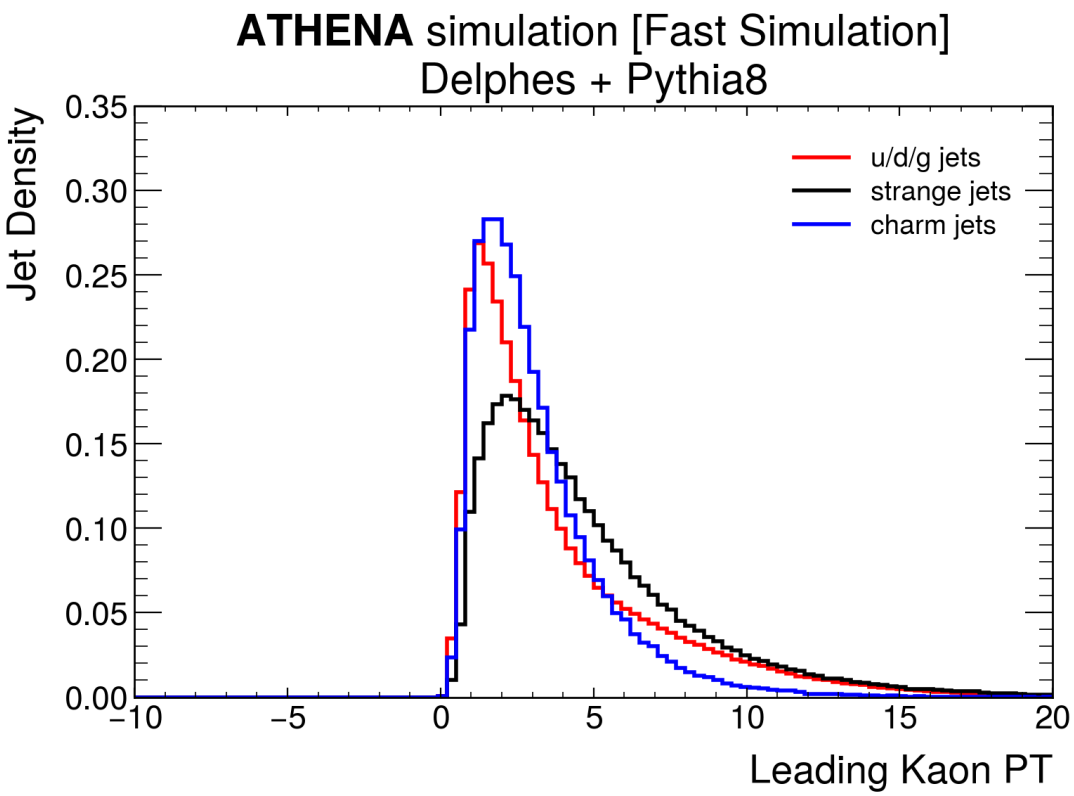

dualRICHag Performance
Ratio
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
0
10
20
30
40
50
Track Momentum [GeV]
$K \to K$
$\pi \to K$

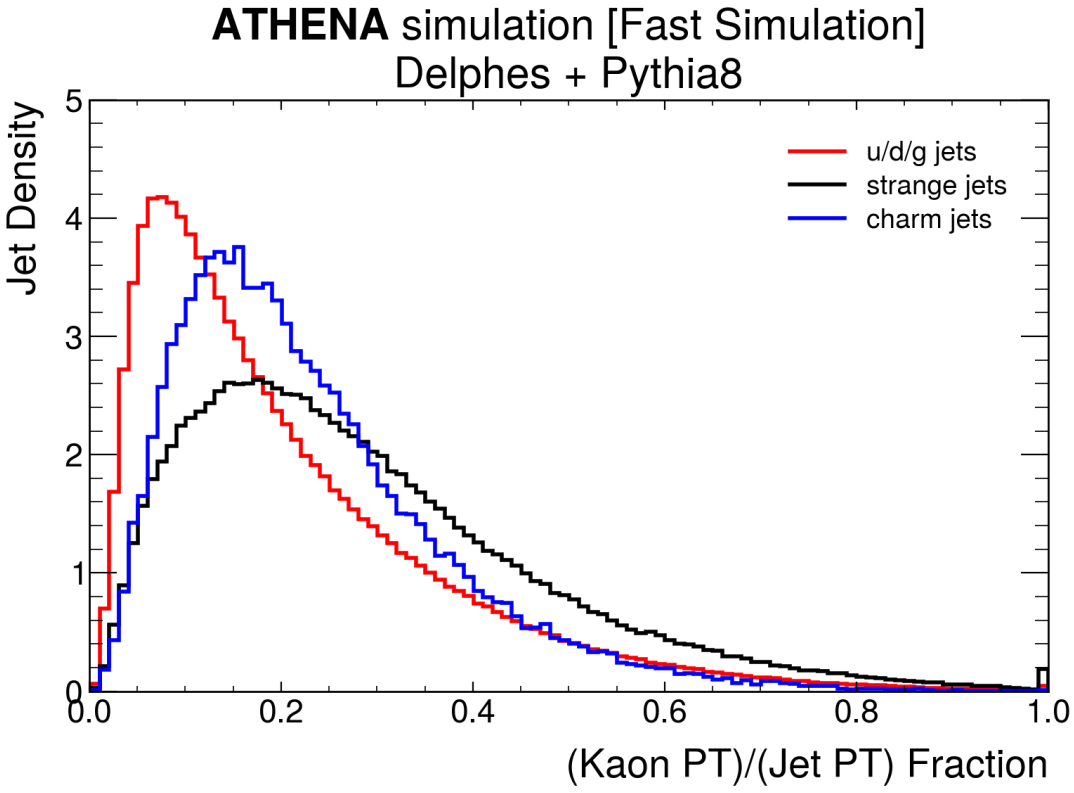

dualRICHcf Performance
Ratio
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
0
10
20
30
40
50
Track Momentum [GeV]
$K \to K$
$\pi \to K$

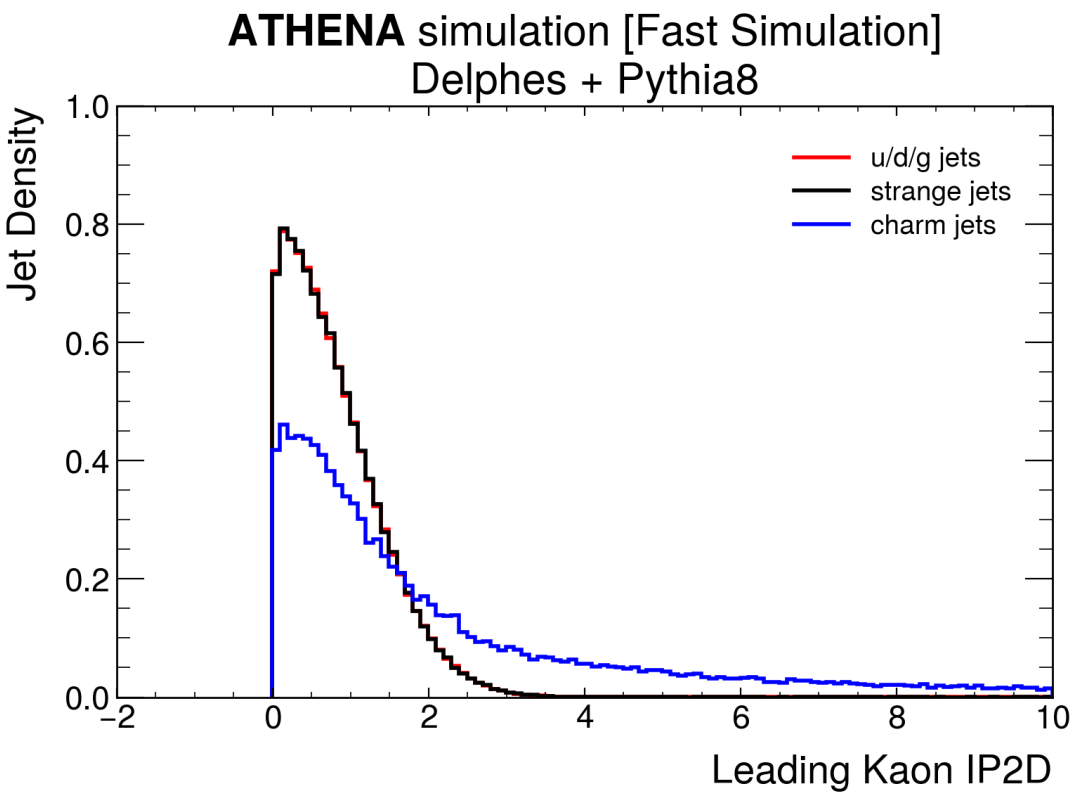
6.1.1 Kaon Features in Charm, Strange, and u/d/g Jets

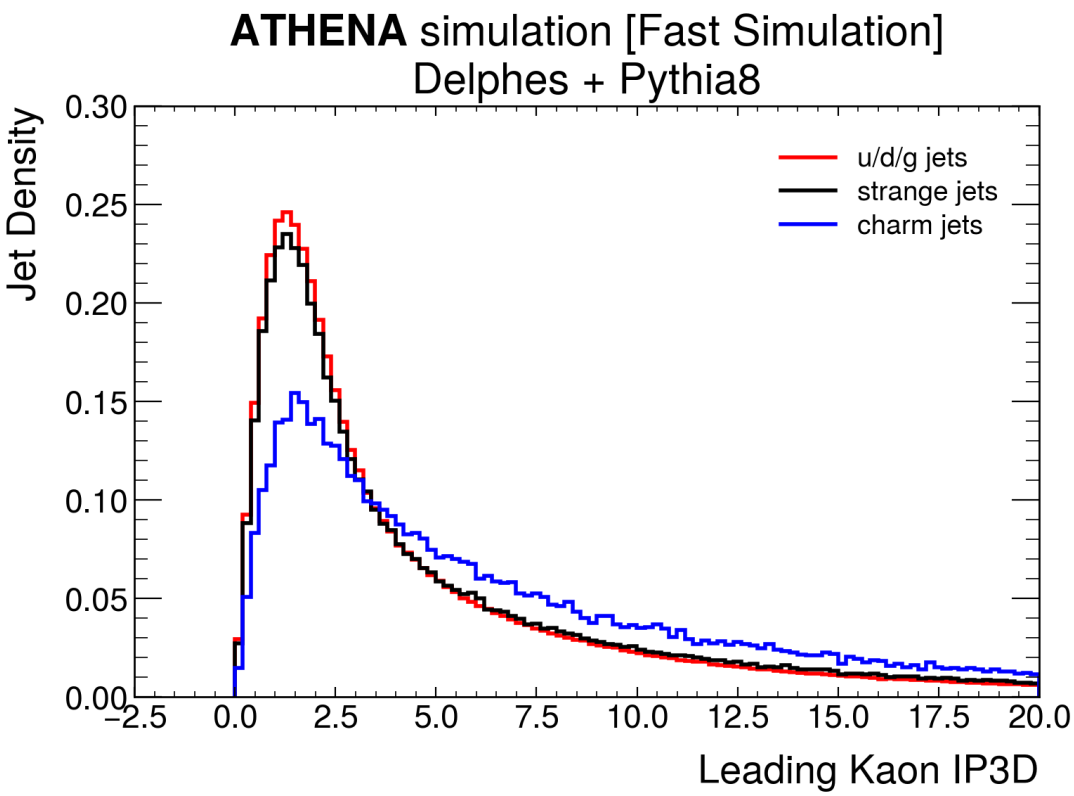


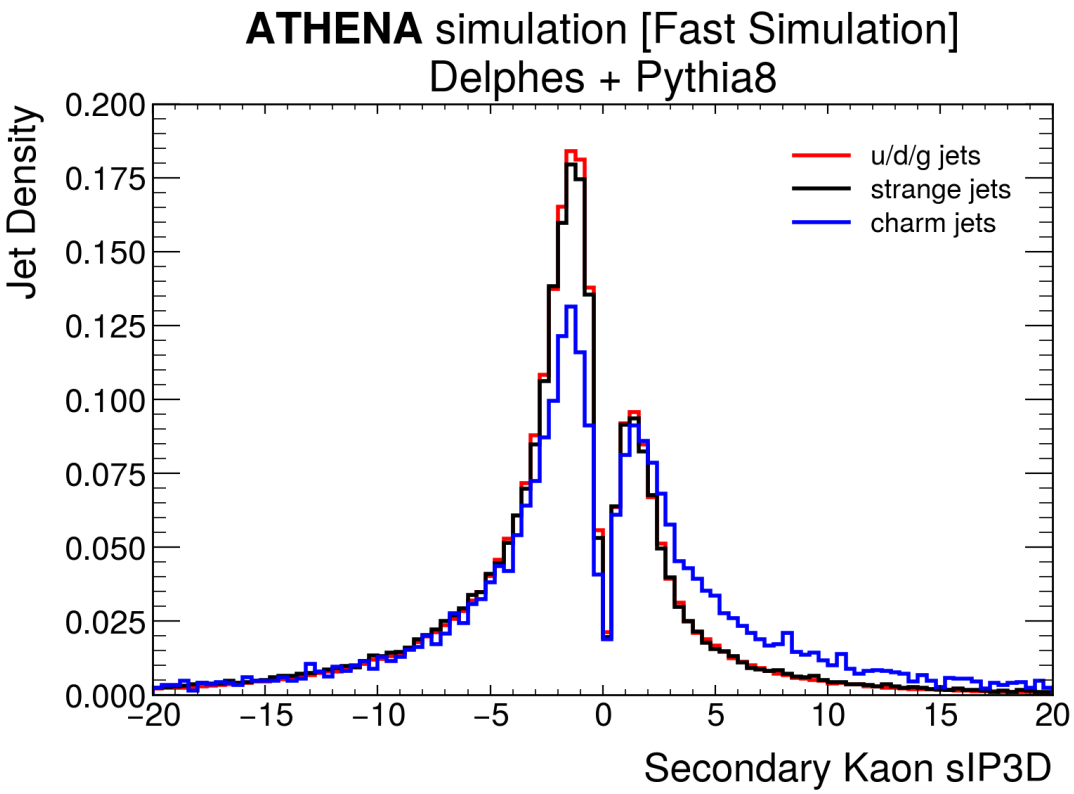

ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Jet Density
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Leading Kaon Charge
u/d/g jets
strange jets
charm jets

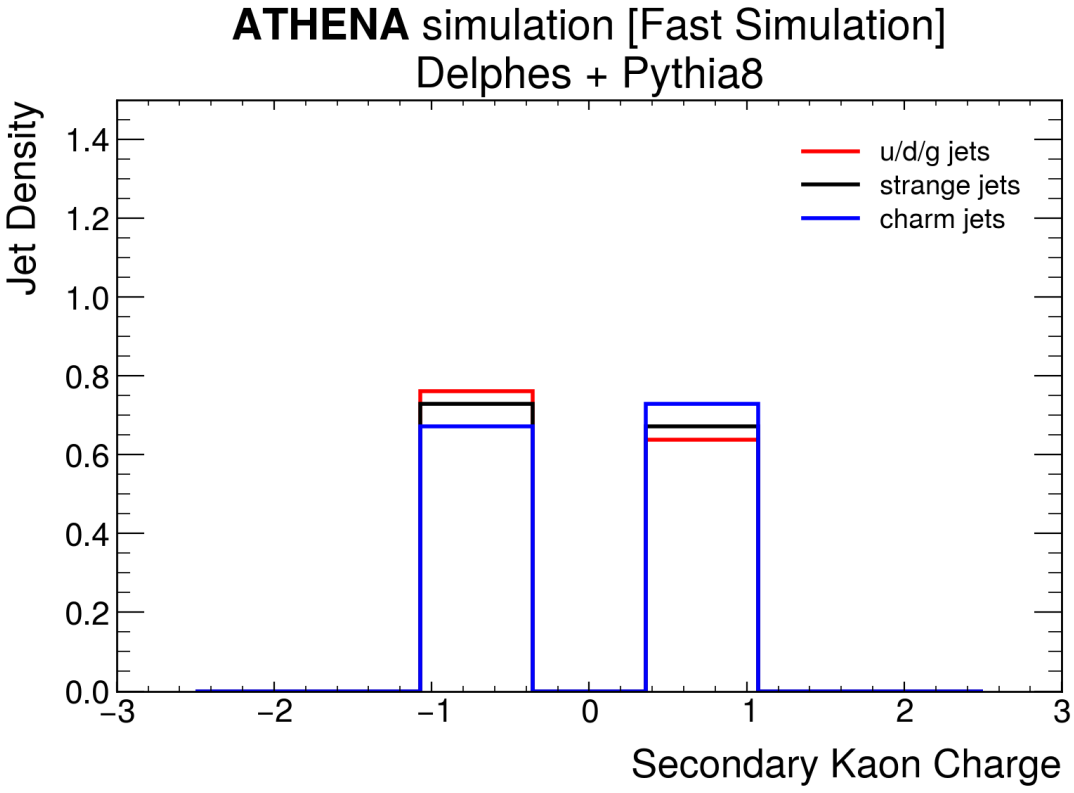

ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Jet Density
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−10
−5
0
5
10
15
20
Leading Kaon PT
u/d/g jets
strange jets
charm jets

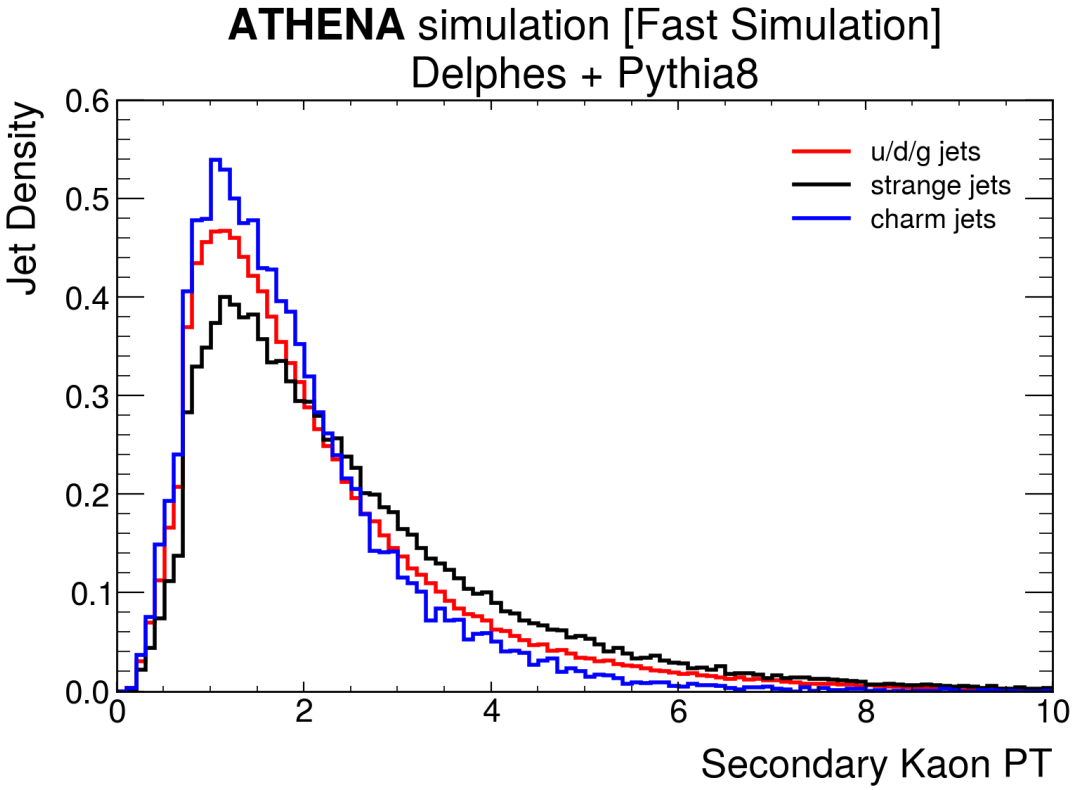

ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Jet Density
u/d/g jets
strange jets
charm jets
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
(Kaon PT)/(Jet PT) Fraction

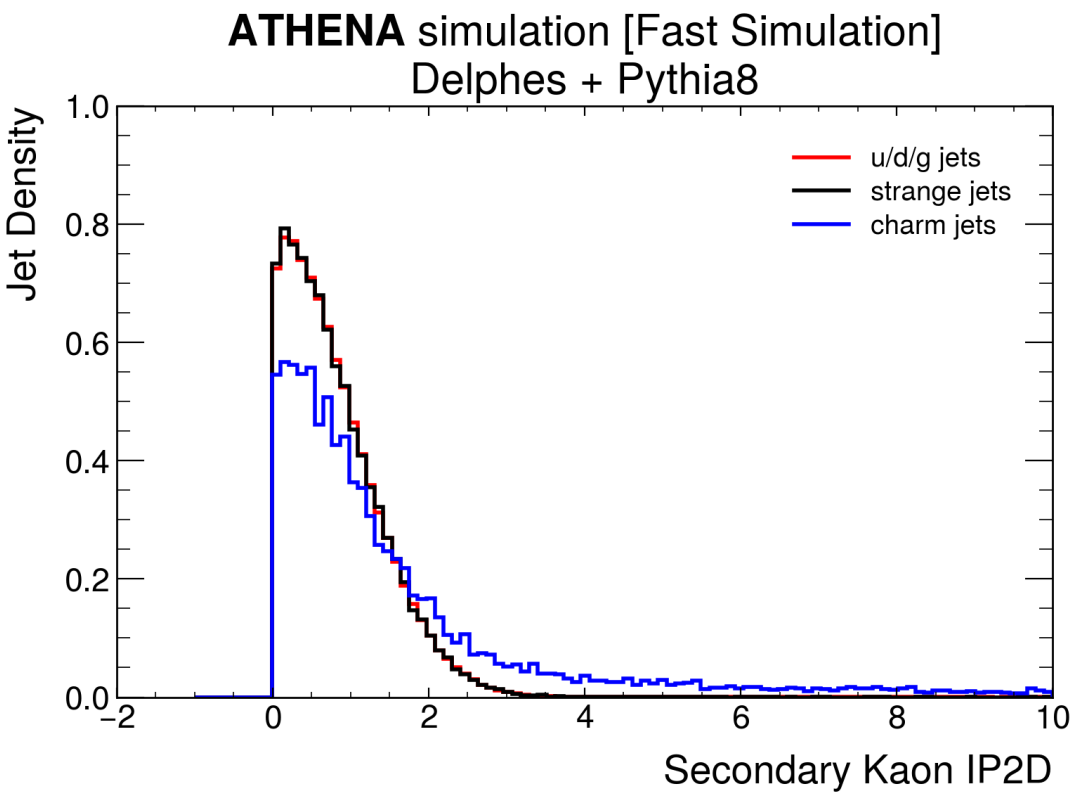

ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Jet Density
1.0
0.8
0.6
0.4
0.2
0.0
−2
0
2
4
6
8
10
Leading Kaon IP2D
u/d/g jets
strange jets
charm jets

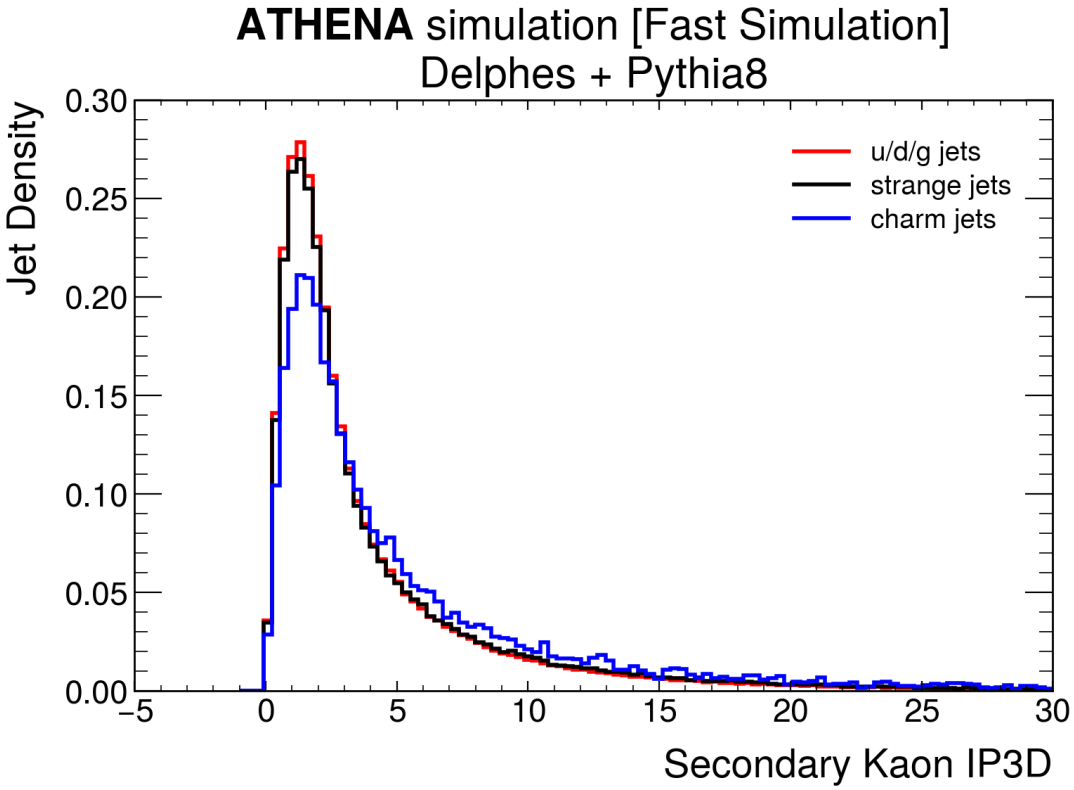

ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Jet Density
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Leading Kaon IP3D
u/d/g jets
strange jets
charm jets

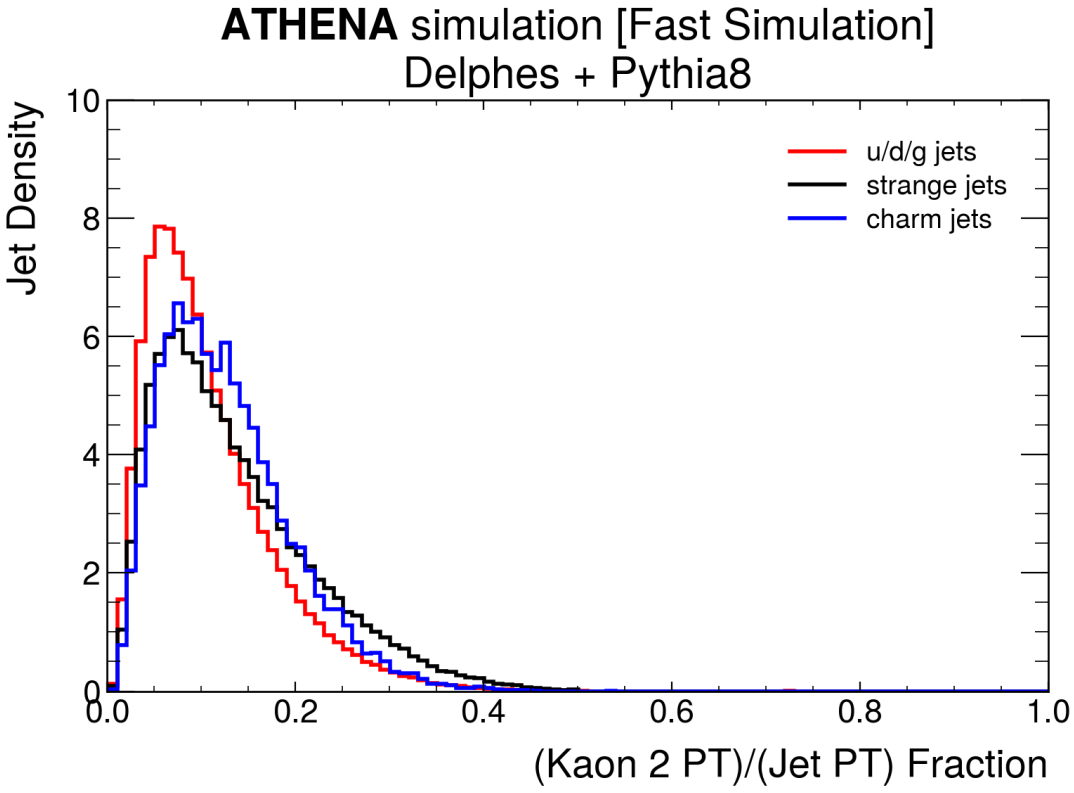

ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Jet Density
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
−20
−15
−10
−5
0
5
10
15
20
Secondary Kaon sIP3D
u/d/g jets
strange jets
charm jets


ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Jet Density
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Secondary Kaon Charge
u/d/g jets
strange jets
charm jets


ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Jet Density
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8
10
Secondary Kaon PT
u/d/g jets
strange jets
charm jets


ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Jet Density
1.0
0.8
0.6
0.4
0.2
0.0
−2
0
2
4
6
8
10
Secondary Kaon IP2D
u/d/g jets
strange jets
charm jets


ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Jet Density
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−5
0
5
10
15
20
25
30
Secondary Kaon IP3D
u/d/g jets
strange jets
charm jets



The most promising variables for selection kaons originating from charm jets are the leading kaon IP_{3D}, p_T, sIP_{3D}, and relative momentum (e.g. p_T^K/p_T^J). For other variables, especially for sub-leading kaons, the shapes of these distributions are increasingly or wholly similar among all jet types.

We note that while the IP_{2D} of the sub-leading kaon looks promising as a discriminating variable, when processed through our current optimization process our overall K-Tagger efficiency drops significantly. We recognize that our current approach - combining a series of independent one-dimensional cuts - is perhaps too naive to combine the leading and sub-leading kaon in a single approach. A more advanced approach later will be appropriate.

6.1.2 Defining the K-Tagger Using Independently Optimized Selection Criteria

We make the very naive assumption that the variables in our set are independent. We then separately optimize selection criteria on each variable, without imposing selections on the others. THe optimization goal is defined by the use of the Punzi figure-of-merit,

$$FOM_{Punzi} = \frac{\varepsilon_s}{\frac{n}{2} + \sqrt{N_b}}.$$

We recognize this is not the best approach, but use this as a baseline starting point. Here, ε_s is the efficiency of a selection on signal objects (charm jets), N_b is the of background objects (light jets) selected by the choice, and n is a hyperparameter that corresponds to the target level of separation (e.g. $n = 3$ roughly implies a goal of 3-sigma separation of signal and background).

The selection criteria ensures that the optimal location is chosen such that a minimal amount of background is chosen consistent with a goal of a certain level of background-signal separation in the post-selection sample. Note that the leading kaon electric charge, q, requires no formal optimization since a selection can clearly be made for charm jets with a postive charge in both the primary and secondary simulation data sets ($q > 0$).

Based on this approach we obtain the following constraints:

- The cut on theleading kaon $sIP_{3D} > 2.4$
- The leading kaon $p_T = [\,0.071,0.78]$ GeV.
- The leading kaon $IP_{2D} > 4.0$
- The leading kaon $IP_{3D} >$ -0.010

Combined with the requirement that $q > 0$ for the leading kaon, these selections are all imposed (boolean "and" operation) and this is referred to as the "K-Tagger".

6.1.3 K-Tagger Baseline Performance

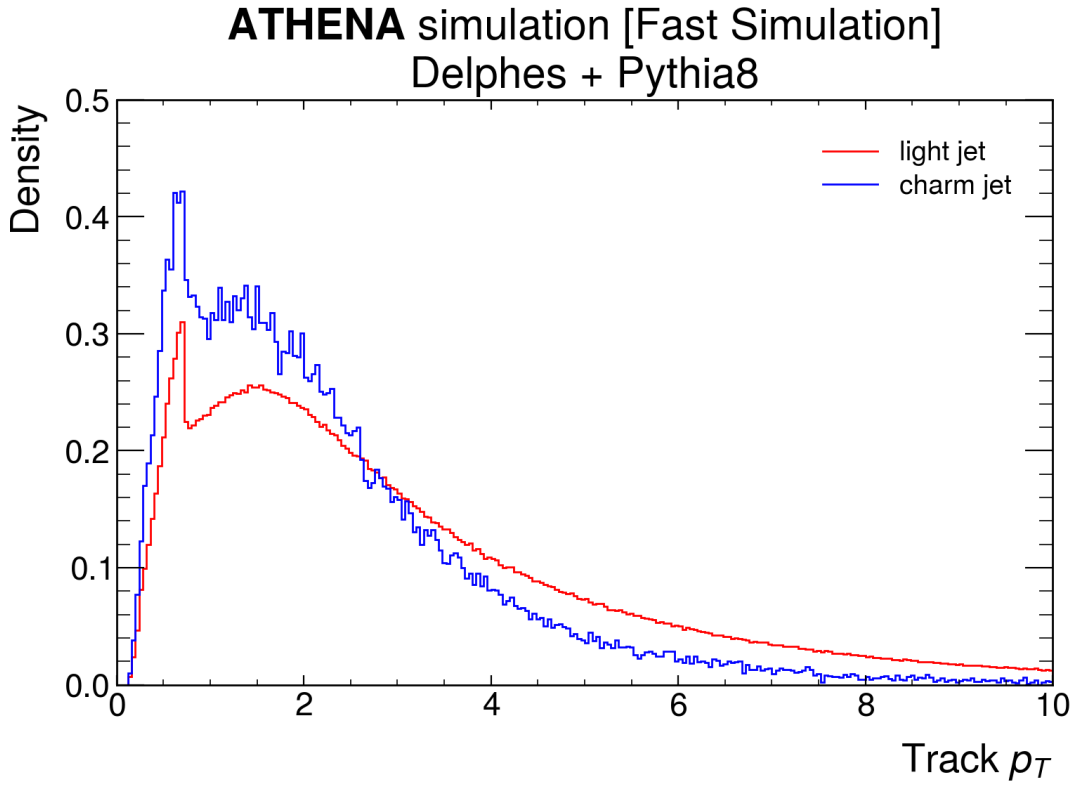
When the above cuts are made to define the K-Tagger, the following percentages of charm jets with kaons are able to be recovered from those missed by the existing CharmIPXDTagger alone. In the following, "exclusively" tagged refers to jets identified by the K-Tagger but not the CharmIPXD-tagger while "inclusively" refers to being tagged by the K-Tagger without regard to whether a jet passes or fails the CHarmIPXDTagger.

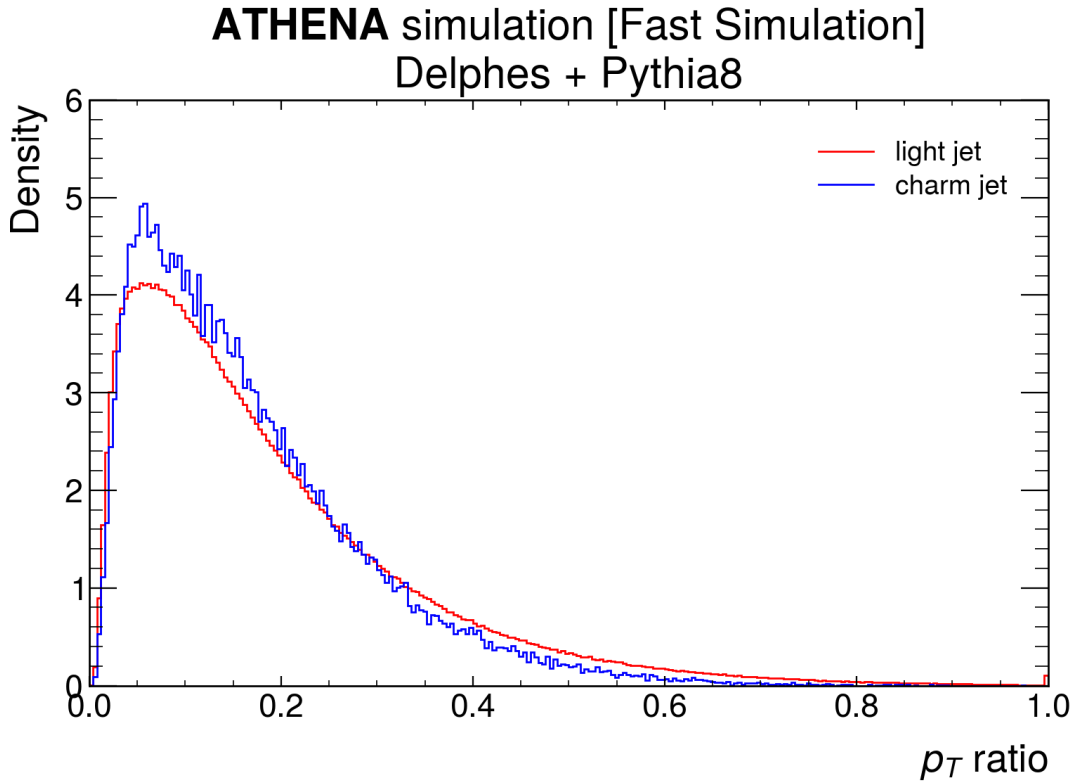
- Inclusively KTagged charm jet efficiency: 6.86%
- Inclusively KTagged strange jet efficiency: 0.0444%
- Inclusively KTagged up/down/gluon jet efficiency: 0.0187%
- Exclusively KTagged charm jet efficiency: 5.30%
- Exclusively KTagged strange jet efficiency: 0.0397%
- Exclusively KTagged up/down/gluon jet efficiency: 0.0155%

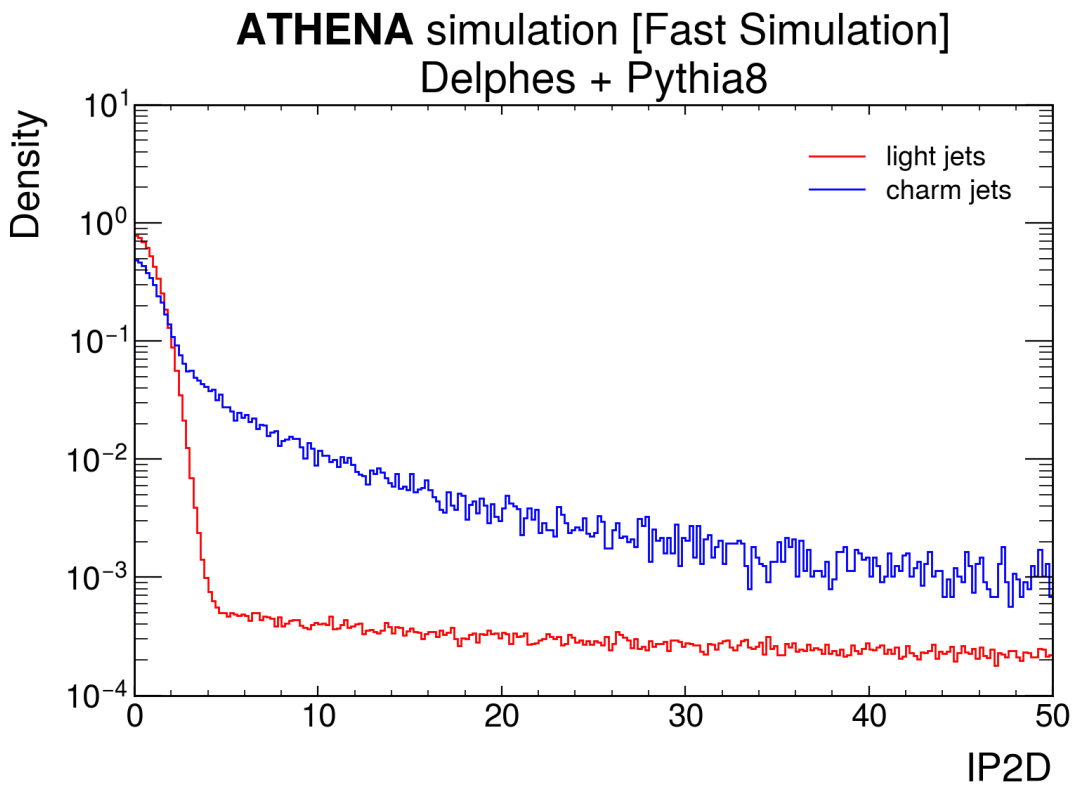
7 Electron Tagging

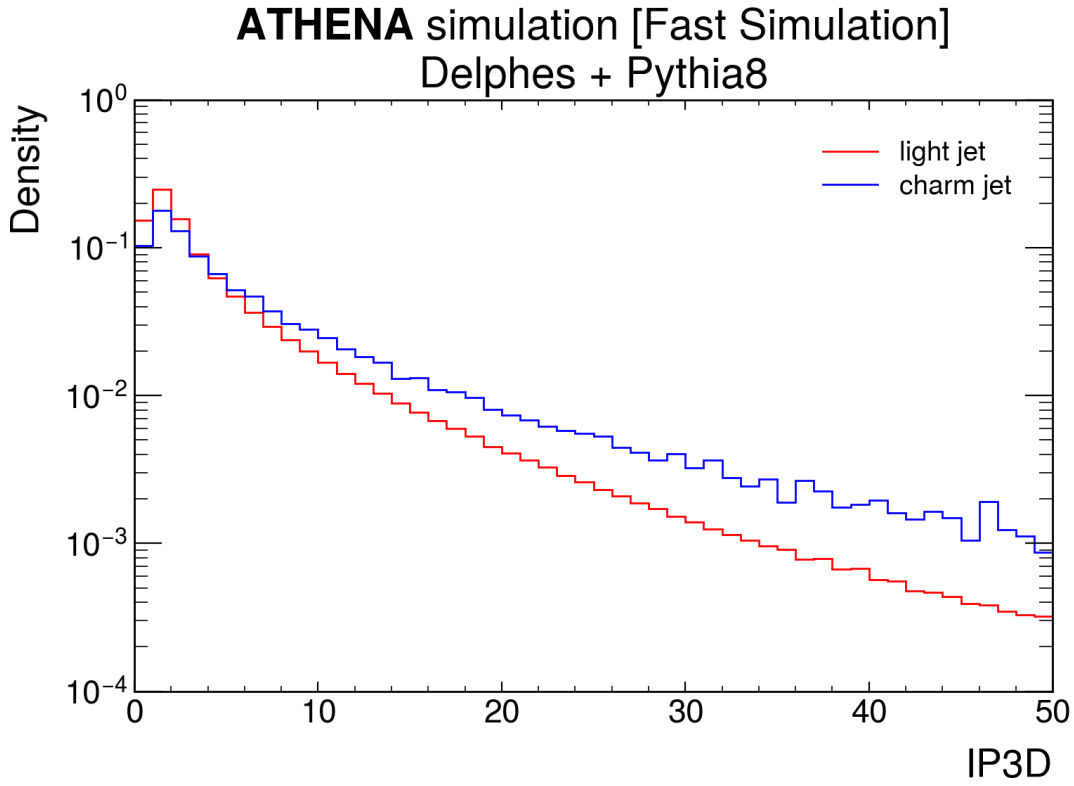
Electrons are defined as tracks matched to jets that have $E_{EM}/E_{CALO} > 0.991$. The fast simulation does not correctly model the ratio of energy deposited in the electromagnetic (EM) and hadronic (HAD) calorimeters, so we correct it using single-particle (electron or pion) simulations from a DD4HEP GEANT4-based ATHENA simulation. Knowing the true identify of a track, we allocate its calorimeter energy to either system using a Monte Carlo approach where we probabilistically set the energy-sharing fraction (based on how true electrons and pions distribute their energy) and store that along with the track. The requirement on the electromagnetic energy fraction was determined using a Punzi FOM optimization of the selection on this ratio, optimally separating real electrons and pions independent of whether or not they arose from a specific jet type.

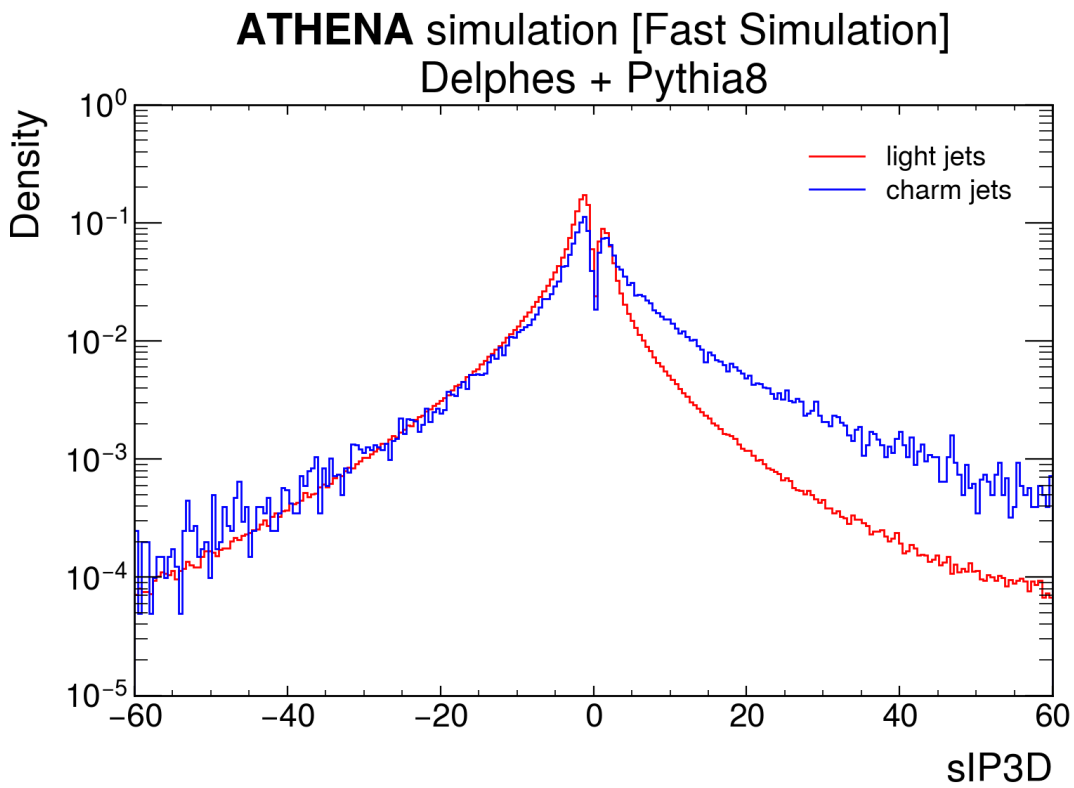
We pursue a similar strategy in the electron tagging as we did in the kaon tagging. We investigated leading electron (by p_T) and sub-leading electron variables, including the p_T, the ratio of the p_T to the parent jet p_T, and the track displacment in 2- and 3-dimensions. We find that the sub-leading electron carries not useful potential for classification of jets using these variables, and that the p_T ratio, IP_{2D}, and sIP_{3D} of the leading electron are the most promising for flavor-tagging and jet classification.

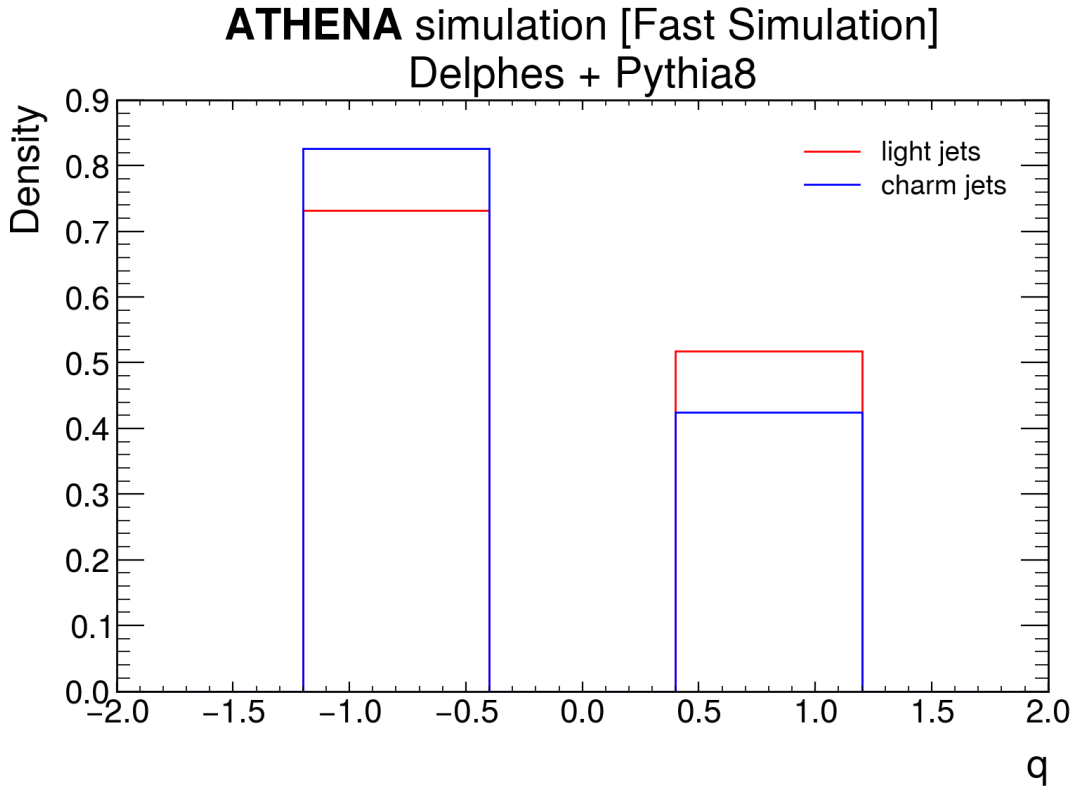



ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Density
light jet
charm jet
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
p_T ratio


ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Density
light jets
charm jets
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0
10
20
30
40
50
IP2D


ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Density
light jet
charm jet
IP3D


ATHENA simulation [Fast Simulation]
Delphes + Pythia8
Density
light jets
charm jets
sIP3D



Our optimization proceeds as with the kaons and we find the following selection criteria:

- $p_T^e / p_T^J < 0.54$
- $IP_{2D} > 3.2$
- $sIP_{3D} > 3.4$

We summarize the efficiency of these selections in the next section, where we compare the inclusive and exclusive tagging efficiencies for the algorithms.

8 Summary of Tagging Performance Estimates

We combine the displace track counting, kaon-based, and electron-based tagging approaches. We first tabulate the exclusive and inclusive efficiencies of each of these approaches.

	Inclusive Charm Efficiency	Exclusive Charm Efficiency	Inclusive Light Efficiency	Exclusive Light Efficiency
CharmIPXDTagger	17.3	14.6	0.861	0.814
KTagger	6.86	4.69	0.0187	0.0141
ETagger	6.17	3.98	0.426	0.377

Based on the current work, the total tagging efficiency naively taking the logical "OR" of all three algorithms would be 23.2%.