Finding Charm at the LHCb

Optimisation of $D^0 \to K\pi$ selections at HLT1

Robert Hartley

10622219

School of Physics and Astronomy

University of Manchester

MPhys Report

May 13, 2024

This project was performed in collaboration with *Elisabeth Peak*
With special thanks to Conor Fitzpatrick and Timothy David Evans.

Abstract

This report builds on previous work to further analyse the $D^0 \to K^{\pm}\pi^{\mp}$ decay to find a more optimised set of selections. A study of the background composition found that the Monte Carlo truth-matching system wrongly identifies 5.7±0.2% of $D^0 \to K\pi$ decays as background. An algorithm was then implemented to optimise the signal significance of selections recursively. The optimal combination of selections resulted in a signal significance of 18.6, a signal purity of 59.3±1.6% and a retention rate of 29.9±1.1 kHz on a 2024 Monte Carlo minimum bias dataset. These were then applied to real data taken in April 2024, resulting in a retention rate of 13.7±0.4 kHz and an estimated signal significance of 11.9.

1 Introduction

The LHC is the world's largest and most powerful particle collider, comprised of four detectors along a 26.7 km circumference beam pipe. One detector, LHCb, focuses on studying rare beauty and charm quark decays in proton-proton collisions. The LHCb experiment generates more data than can be stored offline for further analysis and therefore utilizes a high-level trigger (HLT1) responsible for determining which events to save. This report focuses on optimising the selections that are applied during the HLT1 stage to save the most statistically significant sample of data. This builds on the work completed last semester [1] [2] where $K^0 \to \pi^\pm\pi^\mp$ and $D^0 \to K^\pm\pi^\mp$ decays were studied on data taken from run 3 in 2023. A combination of selections was found for both decays that resulted in a retention rate of less than 1 MHz, the maximum rate of data LHCb can process. However, these results were not optimised to be statistically significant and did not cut harshly enough on the background decays. Consequently, and in anticipation of the changes in conditions and subdetectors expected in 2024 data taking, the decay $D^0 \to K^\pm\pi^\mp$ is more thoroughly analysed in this report.

1.1 Bunch crossing rate

The layout of the LHC is shown in Fig. 1 where the LHCb experiment is located at interaction point eight (IP8). Both beams of protons are injected into the beam pipe in bunches. The minimum time between bunches that the LHC can handle is 25 ns. This corresponds to a maximum rate of bunch crossings of 40 MHz. The number of colliding bunches, n_c can be calculated through Eq. 1,

$$\text{Bunch crossing rate} = f n_c, \tag{1}$$

where f is the LHC orbit frequency, 11245 Hz [4], and the maximum bunch crossing rate is 40 MHz. This results in a maximum number of colliding bunches of 3500, however, in practice it is not possible to reach this maximum for multiple reasons. The LHC beam, at its peak, contains a stored energy of 350 MJ. This destructive power prompts the need for an external beam dump responsible for extracting, diluting and reducing power before absorbing the beam into a dedicated system [5] [6], located at IP6 in Fig. 1. The beam dump system powers up magnets that divert the proton bunches out of the LHC ring. A gap of 3 ms is required in the beam for the magnets to power up to their working value, without this gap, the magnets will pull the bunches into the side of the beam pipe, causing mass damage. This gap in the beam causes a start and end to the beam, an exaggerated version is drawn in Fig. 1, reducing n_c to 2808, corresponding to 31.5 MHz using Eq. 1. As the beams are calibrated to collide at ATLAS and

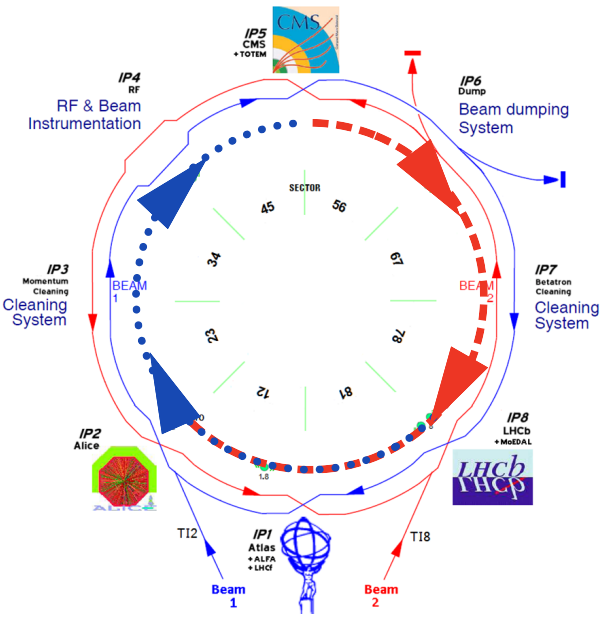


Figure 1: Showing the LHC beam pipe layout. Modified from [3]

CMS, at LHCb, some bunch crossings only have one of the two beams filled. This results in no collision and reduces the overall bunch crossing rate. Fig. 1 shows that in 2023, the maximum number of bunches at the LHCb was 2191 [7]. In this report, the event rate will therefore be assumed to be 25 MHz.

1.2 Issues with the VELO

One of the main detectors at LHCb, the vertex locator (VELO) tracks charged particles and determines the location of the primary and secondary vertices [1]. This information's precision is crucial for event selection and affects other calculated variables. The VELO is retractable, and during beam injection and stabilisation sits 3 cm away to avoid damage, before moving to 5.11 mm away from the beam axis. However, during the start of the 2023 run, there was a vacuum incident at LHCb which affected the function of the VELO. There was a loss of control of the protection system, which led to a pressure imbalance between the beam and VELO vacuums, deforming the RF foil that protects the VELO. The decision was taken that the VELO would not be at a distance closer than 16mm during data taking until it was safe to repair and replace [8]. This means that for the 2023 data, the VELO was set at a distance between 16 mm and 24 mm, whereas in 2024, it is moved to its normal position of 5.11 mm. This further motivates carrying on studying the $D^0 \to K\pi$, as conditions have changed.

2 Data sets

The project aimed to expand on the work completed last semester by focusing on the 2024 data-taking. The pile-up, which is the average number of proton-proton collisions per event, was expected to increase in 2024, as well as changes to detectors, such as the VELO. Data sets were created to replicate the conditions expected and enable deeper insight into the composition of each event. The four datasets used in this project are;

Monte Carlo Signal: Ensures at least one $D^0 \to K\pi$ is created per event. Contains approximately 100,000 events.

Monte Carlo Minimum Bias: Replicates the expected conditions and number of $D^0 \to K\pi$ decays per event. Contains approximately 580,000 events.

Real 2023: Real data taken in 2023 from run 3. Contains approximately 65,000 events.

Real 2024: Real data taken in April 2024. Contains approximately 1,900,000 events

The Monte Carlo Signal data set is used when the selections are not very optimised, and so a When the selections are relaxed, the background becomes predominant, reducing the significance of the mass peak. To mitigate this effect, the Monte Carlo Signal data set is used to generate a higher proportion of signal events, resulting in an enhanced mass peak. This allows changes to initial selections to be easily observed. The Monte Carlo minimum bias data set is used during optimisation to identify the most effective selections on real data. The two real data sets are then used to verify the quality of the selections applied.

2.1 Monte Carlo generation and truth-matching

The Monte Carlo simulations used in this project are created in four stages [9].

Generation: Monte Carlo particles are generated according to Standard Model physics and probabilistic cross-sections.

Digitisation: This generated data is processed through a simulated detector to replicate the readout of an actual event. Accounting for real-world uncertainties, such as detector efficiency.

Reconstruction: In each subdetector, hits are combined to create partial tracks called stubs. Some particles will travel through multiple subdetectors and therefore, multiple stubs can be combined to create long tracks.

Monte Carlo truth-matching: The reconstructed tracks are then compared with Monte Carlo particles to pair up each track with the particle that created it.

The Monte Carlo reconstruction stage is the same as the real reconstruction stage, with the addition of truth matching. The hits that the particles create as they travel through the detector are reconstructed into full tracks, and if 70% of the hits on one track are determined to be from the same Monte Carlo particle, it is assigned an MCTruthID. The MCTruthID is vital for this project as it allows the identity of individual tracks to be known, and pairings of tracks to be classified into signal, background and ghosts. Pairings that consist of a $\pi^{\pm}$ and a $K^{\mp}$ both with a $D^0/\bar{D^0}$ mother particle are categorized as signal decays. Ghosts refer to pairings where at least one of the tracks fails the Monte Carlo truth matching, while the remaining decays are considered background. This classification allows the purity, Eq. 2, and signal significance, Eq. 7, to be accurately determined. It is important to note that the truth matching system is not 100% accurate [10] [11]. There are many ways that the truth-matching system can fail.

Confidence level: As discussed above, the truth matching requires 70% of the hits on a track to be from the same Monte Carlo particle to assign a match. If a track fails this, it is classified as a ghost track. A pairing is then classified as a ghost pair if at least one of the two daughter tracks fails the Monte Carlo truth-matching.

Mismatched stubs: Long-lived tracks will hit multiple subdetectors within LHCb. The hits left in each subdetector are combined into track stubs. These stubs are then matched together to recreate the full-length track. However, in regions with a high particle flux, matching stubs together becomes challenging, leading to an increase in the number of ghost tracks, or instances where tracks are assigned an incorrect MCTruthID.

Material interaction: As particles propagate through the material, they can interact in multiple ways, such as Bremmstralung radiation. Sudden changes in momentum or spontaneously emitted electrons could confuse the truth-matching algorithm resulting in misidentified particles.

Inflight decays: Although the truth matching system is designed to deal with inflight decays, it can still have trouble matching rapidly decay particles to the right tracks.

3 Selection metrics

3.1 Signal purity

The signal purity, ϵ_p, is defined as the ratio between decays classified as signal and the total number of decays. It is calculated through Eq. 2,

$$\text{Purity} = \frac{N_s}{N_s + N_b} = \epsilon_p, \tag{2}$$

where N_s is the number of signal decays saved and N_b is the number of non-signal (background and ghost) decays saved. The uncertainty on a purity calculation can be determined using the error propagation formula,

$$\sigma(\epsilon_p)^2 = \left(\frac{d\epsilon_p}{dN_s}\right)^2 \sigma^2(N_s) + \left(\frac{d\epsilon_p}{dN_b}\right)^2 \sigma(N_b)^2. \tag{3}$$

The error propagation formula assumes that both variables are independent of each other, and therefore the total number of decays, N, is split into $N_s + N_b$ to remove the dependence between N_s and N. The uncertainty on the number of signal decays, $\sigma(N_s)$, and non-signal decays, $\sigma(N_b)$, can be approximated to $\sqrt{N_s}, \sqrt{N_b}$ using Gaussian statistics, which then results in,

$$\sigma(\epsilon_p) = \sqrt{\frac{\epsilon_p(1-\epsilon_p)}{N_s + N_b}}, \tag{4}$$

for the error on purity.

3.2 Retention rate

Another essential metric for evaluating selections is the retention rate, which is the number of events saved by the HLT1 per second. As outlined in [1], LHCb's maximum processing rate stands at 1 MHz. However, as multiple decay channels are run concurrently, the retention range for these elections is required to be within the tens of kHz. Retention is calculated as the ratio of the number of accepted events, N_a, to the total number of events in the original sample, N_e. Following the same rationale as previously stated, this can be expressed as $N_e = N_a + N_{na}$, where N_{na} represents the count of events not accepted, thus resulting in,

$$\text{Retention Rate} = fn_c \left(\frac{N_a}{N_a + N_{na}}\right) = \epsilon_r. \tag{5}$$

The uncertainty on the retention can then be calculated in the same manner, resulting in Eq. 6,

$$\sigma(\epsilon_r) = fn_c \sqrt{\frac{\epsilon_r(1-\epsilon_r)}{N_a + N_b}}. \tag{6}$$

3.3 Signal significance

To identify the optimal combination of selections, an equilibrium must be achieved between signal purity and signal efficiency. To optimize both parameters, signal significance (S) is used as a crucial metric. It is defined as,

$$S = \frac{N_s}{\sqrt{N_s + N_b}}. \tag{7}$$

The signal significance parameterises how many standard deviations a signal peak is from the background. The higher the signal significance, the more statistically prominent the invariant mass peak is. At the LHC, for a discovery to be accepted statistically, a signal significance above 5σ is required. Although the $D^0 \to K\pi$ is a well-known and studied decay, the aim is to reach this significance at the first level of the trigger.

4 Particle decay parameters

This section provides an overview of the main selections implemented in the project. A theoretical and conceptual explanation of each cut is given before experimental evidence of the effectiveness is shown later in Sect. 6.

4.1 Impact parameter chi-squared

The impact parameter is the distance between a track and the primary vertex as it crosses the primary vertex's x-y plane. It is discussed in [1] [2] and used throughout both projects. It is a useful way of determining if a track originated from a certain point. Last semester, the impact parameter of the mother particle was used as an additional selection to ensure the mother particle originated close to a primary vertex. This selection can be upgraded to the impact parameter chi-squared, χ^2_{IP}, which accounts for the relative uncertainties and is proportional to $(\mathrm{IP}/\sigma_{\mathrm{IP}})^2$. The mother particle χ^2_{IP} is more computationally expensive than the track χ^2_{IP}. This is because the χ^2_{IP} of a composite particle has to account for each track's uncertainty, which extends to the uncertainty of the decay vertex as well as the composite particle's momentum. Therefore, at the level of the HLT1, the impact parameter chi-squared is approximated.
Approximating the composite χ^2_{IP} requires the uncertainty on the inverse of the track momentum. This can be found by using the Monte Carlo MCTruth properties discussed above. The difference, δp, in reconstructed track momentum, p_{TR}, and Monte Carlo particle momentum, p_{MC} can be found through,

$$\frac{\delta p}{p} = \frac{p_{TR} - p_{MC}}{p_{TR}}. \tag{8}$$

The uncertainty on a track with momentum p_{TR} can be better characterised by grouping the data into bins and then calculating the root mean square error (RMSE) on each bin. This is calculated as,

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i}^{N}\left(\frac{\delta p}{p}\right)_i^2}, \tag{9}$$

where N is the number of values in each bin. The error on inverse track momentum, $\sigma(\frac{1}{p})$, is then related to the track momentum uncertainty, $\sigma(p)$, by

$$\sigma\left(\frac{1}{p}\right) = \frac{\sigma(p)}{p^2}, \tag{10}$$

through normal error propagation.
A small composite χ^2_{IP} suggests that the decaying particle was produced near the primary vertex. D^0 particles are mainly produced at the primary vertex or through rapidly decaying resonances,

such as $D^*(2007)^0$, which only travel ≈ 10 fm [12]. Therefore, the composite χ^2_{IP} can be implemented to remove a high percentage of combinatoric decays that are less likely to look like they originate from the primary vertex.

4.2 Radial distance travelled

The distance that a decaying particle travels in the azimuthal plane, ΔR, is calculated by

$$\Delta R = \sqrt{\Delta x^2 + \Delta y^2}, \tag{11}$$

where Δx, Δy is the distance between the primary vertex and composite particles decay vertex in the x, y direction. Combinatoric background decays are expected to be more densely populated where many tracks are intersecting. Following the inverse square law, combinatorial decays are expected to have a shorter radial distance travelled. The signal decays on the other hand have a lifetime and so on average travel further before decay. The combination of a ΔR and composite χ^2_{IP} cut ensures that the mother particles are produced near the primary vertex, and then travel a substantial distance before decaying. This combination will help drastically reduce the amount of combinatoric background being selected.

4.3 Decay vertex position reduced chi-squared

The decay vertex position is determined by performing a fit of the point where the daughter tracks are the closest together in terms of their uncertainties. The vertex position chi-squared, χ^2_v, is then calculated by seeing how consistent the tracks are with the fitted position. The decay vertex position reduced chi-squared, $\tilde{\chi}^2_v$, is then calculated by dividing by the degrees of freedom available, N_{DOF},

$$\tilde{\chi}^2_v = \frac{\chi^2_v}{N_{\mathrm{DOF}}}. \tag{12}$$

During this project, this cut has been set at $\tilde{\chi}^2_v < 10$, which is the standard for LHCb. However, with the anticipated rise in pile-up for the 2024 data collection, harsher cuts will be necessary. These stricter measures are likely to eliminate more signal decays. Therefore, implementing selections that specifically target ill-defined signal decays will mitigate the impact of these harsher cuts. Signal decays with a higher uncertainty are inherently less useful for further analysis, and removing them increases the overall quality of the sample.

4.4 Pseudorapdity

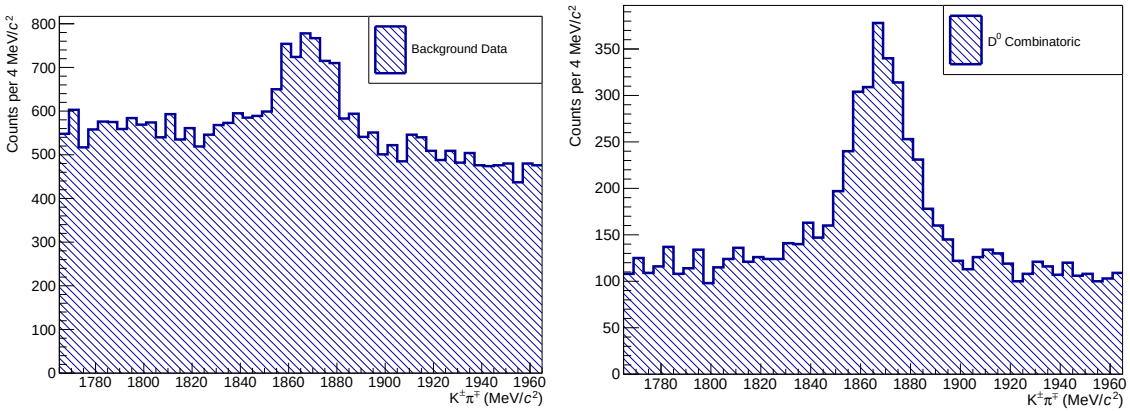
The pseudorapidity, η is related to θ, the angle between particle momentum and the positive z axis by,

$$\eta = -\ln(\tan\theta/2). \tag{13}$$

This parameter reveals more spatial information about the track trajectory, allowing another cut to be implemented that can remove decays that travel at a certain solid angle. For example, if a detector is known to have reduced spatial coverage, the pseudorapidity cut can remove tracks that miss the detector, increasing the overall quality of the sample.

Initial Selections	Retention Rate	Signal Purity	Signal Significance
p_T >500MeV/c			
p >4000MeV/c			
Track χ^2_{IP} >3	8700±15 kHz	1.13±0.01%	10.47±0.11
DIRA>0.995			
D^0 IP <0.15mm			

Table 1: The final set of selections from last semester, [1], applied to Monte Carlo minimum bias 2024



(a) Shows a peak in the background sample. **(b)** Shows a peak in the D^0 combinatoric sample.

Figure 2: Invariant mass plots from the 2024 Monte Carlo signal data with Tab. 1 selections.

5 Looking further into $D^0 \to K\pi$ selections

The research conducted in the previous semester concluded with a combination of selections that achieved a retention rate of 750±15 kHz, which was under the target of 1 MHz. As discussed in last semester's report [1], the 1 MHz limit was set as this is the maximum amount of data LHCb can store offline for further analysis, however, multiple decay channel selections will be running at the same time so the retention rate target has been changed to the tens of kHz. Furthermore, when applied to the Monte Carlo 2024 minimum bias data, the starting selections were two orders of magnitude too large, as shown in Tab. 1. This motivates the deeper analysis into the $D^0 \to K\pi$ selections.

In this section, the report aims to discover more about the form of the background to more successfully remove it from selections. The 2024 signal Monte Carlo data set will be used to study the composition of the background. This will identify where improvements to existing cuts, or in the introduction of selections could be made.

5.1 Analysis of the background sample

The starting background sample, shown in Fig. 2a, is the invariant mass distribution of all the background track pairings that survive the initial selections, Tab. 1, as defined in Sect. 2.1. Fig. 2a has also had an invariant mass selection added, which ensures that only decays with an invariant mass inside 1865 ± 100 MeV/c^2 are saved. This is done to drastically reduce the

amount of decays saved, whilst also allowing enough space on either side of the peak to model the background in further analysis. This is kept for the remainder of the project. Of the 28218 track pairings that make up the background sample, 2.8% are from actual decays, where both tracks have the same mother particle MCTruthID and are thus assumed to be part of an actual particle decay. The other 97.2% are combinatoric, where two random tracks have been paired together and satisfy the selections in place. The actual decays were mainly from $\rho^0 \to \pi\pi$, which could have been removed by changing the mass hypothesis of the daughter tracks and implementing a strict invariant mass cut at the ρ^0 mass (770 MeV). However, as the actual decays only contributed to 2.8% of the background, this was deemed inefficient.

5.1.1 False mass peak

During this investigation, a false mass peak was found in the background sample as shown in Fig. 2a. Having a peak in the background sample limits the purity that can be achieved, the significance of the signal mass peak and thereby, reduces the quality of the selections. It was found that roughly 3% of the background decays were peaking at the D^0 invariant mass.
To analyse what was causing this issue, the combinatoric background section was broken down further into;

D^0 **Combinatoric:** Decays where only one track has a $D^0(\bar{D}^0)$ mother particle.

Non D^0 **Combinatoric:** Decays where no tracks have a $D^0(\bar{D}^0)$ mother particle.

It was found that the combinatoric D^0 decays almost fully account for the invariant mass peak, whilst only containing 27.2% of the total background, as shown in Fig. 2b. As these are peaking at the D^0 invariant mass, they must be signal decays that have been misidentified as background. The difference between D^0 and non-D^0 combinatoric decay categories was then studied further. The majority of the combinatoric decays should be similar to Fig. 3, where an actual decay has occurred but a random track has been paired with one of the daughter tracks. Trivally, D^0 combinatoric decays therefore involve a D^0 decay and a random track, and non-D^0 decays involve a random decay and a random track. There should therefore be no difference in the distributions of the random track MCTruthID in these scenarios. Fig. 4 shows the MCTruthID of the random track involved in D^0 and non-D^0 combinatoric decays. It shows that when the D^0 is involved, the chance of the random track being a pion or kaon decreases by 23.7% and 4.5% respectively, whilst the possibility of it being an electron increases by nearly the same amount, 28.5%. It is then concluded that the Monte Carlo truth-matching system is mistaking the daughter tracks of D^0 decays for electrons, causing signal decays to be labelled as background. If all of the peaking decays in the background sample are signal that have been misidentified, then 5.78% $\pm$0.15% of signal decays are misidentified by the Monte Carlo truth matching system. Although this is a significant error, it does not invalidate the

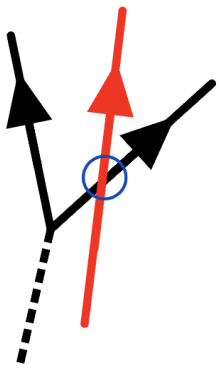


Figure 3: Example of a combinatoric decay. A daughter track from a real decay, in black, is paired up with a random track, in red.

results and conclusions made in this report. It introduces a systematic uncertainty that causes the reported signal purities and efficiencies to be lower than if the Monte Carlo truth matching system was perfect.

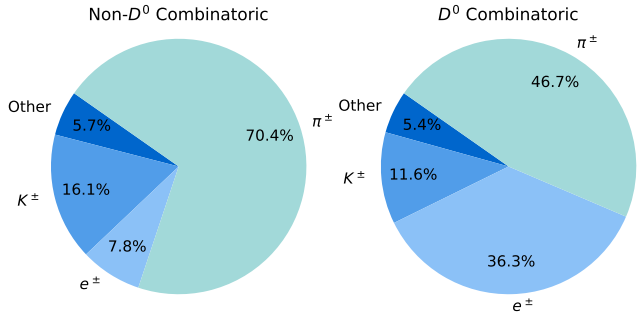


Figure 4: Distribution of MCTruthID for the random track, (Fig. 3), involved in combinatoric decays

5.1.2 Monte Carlo charge misidentification

The discovery of a significant error in the Monte Carlo truth matching system spurred a deeper investigation into how the reconstructed tracks were being wrongly assigned MCTruthIDs. With the information available at the HLT1 reconstruction level, the only definite way to check if a track has been assigned the wrong MCTruthID is to check if the charge of the reconstructed track is the same as the Monte Carlo particle's charge. This checks for the most extreme cases of particle misidentification. Around 3% of the tracks in the Monte Carlo 2024 signal data were classified as MC misidentified, with 86% being electrons. These tracks can be studied further to see what kind of interaction is causing this error. It is important to note that this error is different from the one discussed above. A misidentified decay that causes the background mass peak in Fig. 2a, is not the same as a track that has been assigned a Monte Carlo particle of opposite charge. Fig. 5 shows the difference between reconstructed track momentum and the assigned Monte Carlo particle momentum, in cases where they have undergone Monte Carlo charge error. Normally, this would be very tightly distributed around 0 MeV, however, in the case of the charge error, it shows that 93% of reconstructed tracks have more momentum than the assigned Monte Carlo particle. The average reconstructed momentum of tracks with the charge error is 28 GeV, whereas the assigned Monte Carlo particles on average have 2 GeV. The charge of a track is determined by the direction it bends when a magnetic field is applied. The higher the momentum of the track, the less it will curve, eventually appearing as a

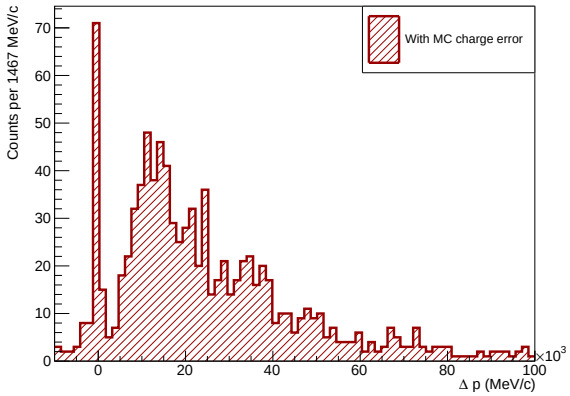


Figure 5: Plot comparing reconstructed track to Monte Carlo particle momentum for charge misidentified tracks

straight line in the detector. This makes their charge difficult to resolve. It is, therefore, more likely that these charge-misidentified particles have high momentum tracks, however, an average momentum of 28 GeV is not high enough to be the dominant effect causing this. The simplified reconstruction at HTL1 will also play a part in causing these tracks to be truth-matched to the wrong Monte Carlo particle.

5.1.3 Cause of the charge misidentification

As discussed in Section. 2.1, there are four main ways the Monte Carlo truth matching fails; multiple hits confidence level, mismatched stubs, material interactions and inflight decays. These Monte Carlo charge error tracks can be analysed to identify the cause of the error.
The angular distribution of tracks is shown in Fig. 6. If the charge error was due to mismatched stubs, then the Fig. 6b distribution would be expected to be heavily clustered around the high concentration areas in Fig. 6a. However, as the Monte Carlo charge error distribution is uniform, this is improbable that this is the primary source of the error. Therefore, it can be inferred that the charge errors are not caused by mismatched stubs, and it is assumed that this holds true for the background mass peak found earlier in Fig. 2a. Consequently, it is deduced that the error associated with the false mass peak stems from the daughter pions and kaons decaying in flight or interacting with the material, resulting in the production of a collinear electron. This electron is then misidentified as the reconstructed track by the truth-matching system.

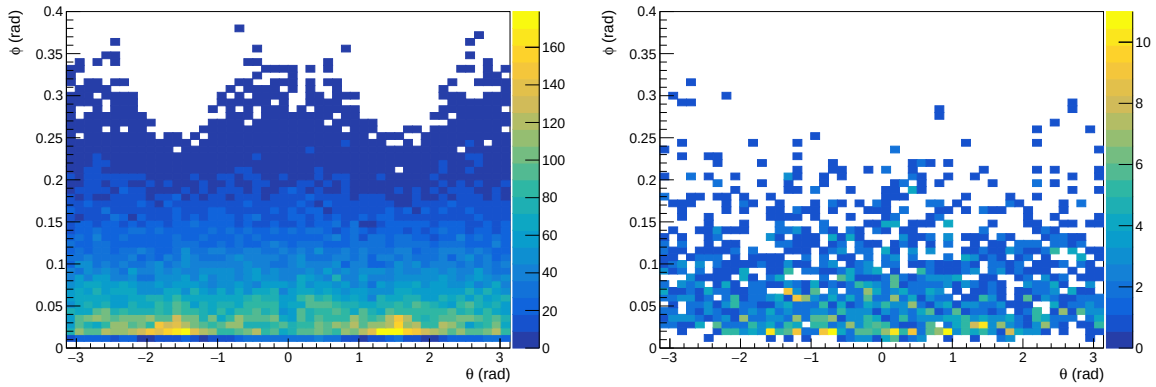
6 Tightening $D^0 \to K\pi$ selections

Although a deeper understanding of the background distribution has been gained the retention rate on the Monte Carlo minimum bias dataset is still too high. Attention is shifted to introducing new cuts before tightening and optimising the most effective ones.

6.1 Mother particle impact parameter chi-squared

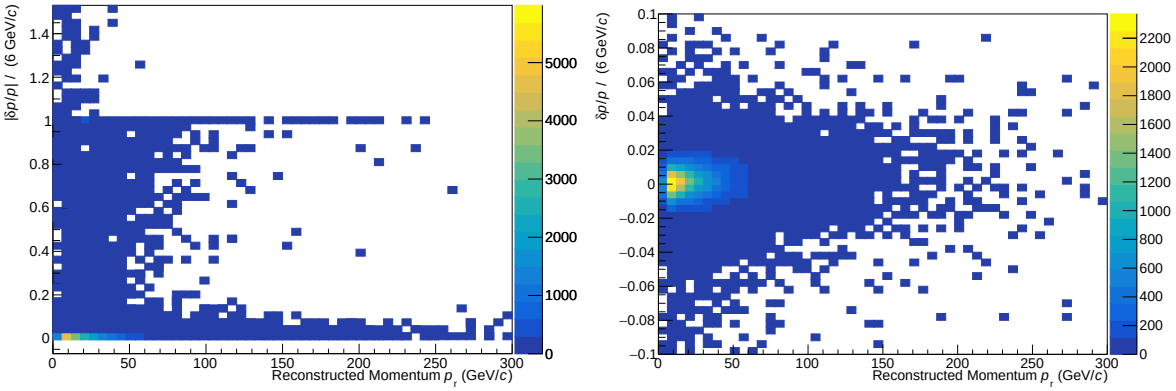
As explained in Sect. 4.1, the composite impact parameter chi-squared, or D^0 χ^2_{IP}, needs to be approximated. First, $\delta p/p$ is calculated through Eq. 8. The $|\frac{\delta p}{p}|$ distribution is then plotted against p_{TR} in Fig. 7a, which shows a much wider spread than expected. There is a significant amount of tracks that have $|\delta p/p| \approx 1$, which indicates that high-momentum reconstructed tracks are being assigned to low-momentum Monte Carlo particles. This seems to be the same effect as discussed in Sect. 5.1.2. It is assumed that the wider spread is due to Monte Carlo-related errors and therefore the region of $|\delta p/p| < 0.1$ was focused on for the estimation of the track momentum uncertainty, as shown in Fig. 7b. This is done as the composite χ^2_{IP} calculation should only include detector uncertainties. To finish the track momentum uncertainty parameterisation, Fig. 7b was sorted into bins of equal size $N = 1000$ and the root mean squared error was calculated using Eq. 9. This is approximately the standard deviation of the track momentum, as the mean in the $(-0.1 < \delta p/p < 0.1)$ region is approximately 0. The RMSE of each bin was fitted against the average momentum of the bin, resulting in Fig. 8a and the fit equation,

$$\sigma(p) = 1.05 \times 10^{-4} \times p_{TR}[\mathrm{GeV}/c] + 8.22 \times 10^{-3}. \tag{14}$$



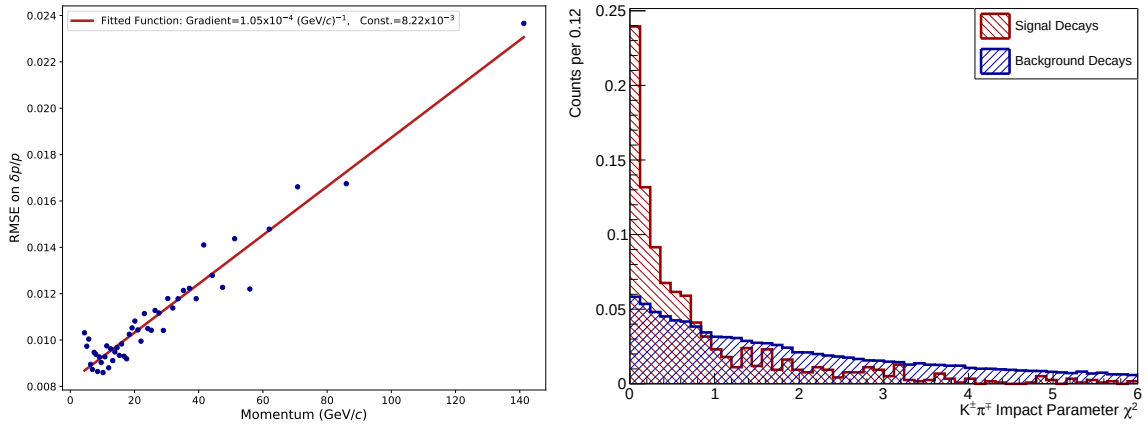
(a) Distribution of all tracks **(b)** Distribution of charge misidentified tracks

Figure 6: Graphs that show the angular distribution of tracks throughout the detector.



(a) The absolute difference, $|\delta p/p|$, of all decays **(b)** $\delta p/p$, in the region $(-0.1 < \delta p/p < 0.1)$

Figure 7: Plots of the differences between reconstructed and Monte Carlo momentum against reconstructed momentum.



(a) Fit of RMSE of $\delta p/p$ against track p_{TR}. **(b)** Normalized signal vs background D^0 χ^2_{IP} Plot

Figure 8: Shows the result of the RMSE of $\delta p/p$ bins against p to parameterise $\sigma(p)$. This is then used to create the composite particle χ^2_{IP} distributions shown in Fig. 8b.

The fitted equation allows the error, $\sigma(p)$, on a particle track of momentum p_{TR} to be found. The uncertainty on the inverse momentum can then be calculated through Eq. 10 and input into the covariance matrix to estimate the composite χ^2_{IP}. The approximation of the D^0 χ^2_{IP} results in the normalized distributions of signal and background decays, plotted in Fig. 8b. This shows that, as expected, the D^0 mesons have a much smaller χ^2_{IP} than the combinatoric background. This is because they are predominantly prompt particles produced at the primary vertex, justifying the implementation of a maximum impact parameter χ^2 cut.

6.2 Radial distance

Using Eq. 11, the normalized signal and background distributions have been plotted in Fig. 9a. This confirms that the ΔR distribution between background and signal decays is very different, allowing a very efficient minimum distance travelled cut.

6.3 Decay vertex position reduced chi-squared

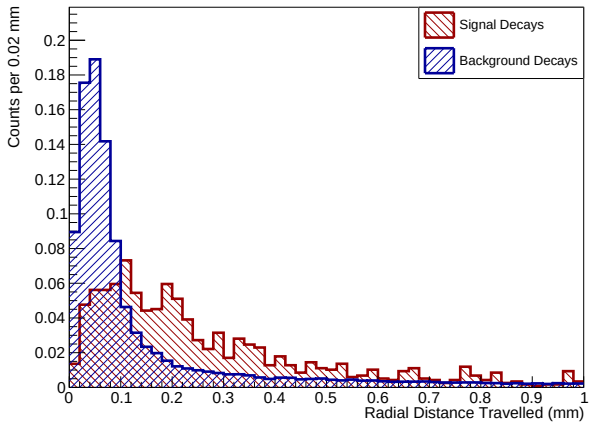
Fig. 9b shows the distribution of σ_V for both signal and background decays and indicates that this cut can be stricter. Although this cut will not be as effective as others in terms of pure signal significance, its aim is to remove decays with higher uncertainty, increasing the overall quality of the sample saved.

6.4 Pseudorapidity cuts

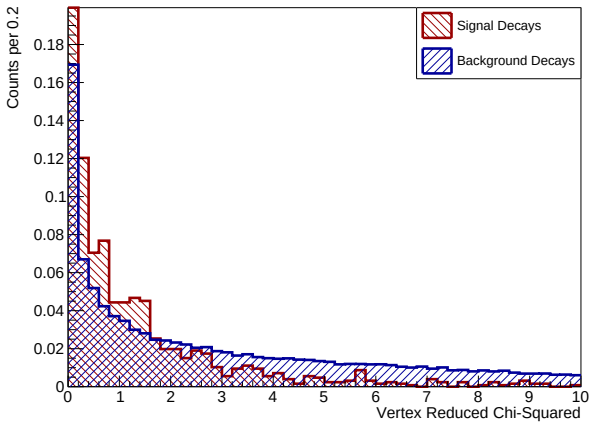
The most important detector for particle identification in charm physics is the RICH-1 detector which spans 25-300 mrad. If a track misses this range it reduces the certainty of a track's particle identity, decreasing the usefulness of the decay. Using Eq. 13, an angle of 25 mrad will correspond to a pseudorapidity of 4.38, any η above this will likely miss the RICH-1 detector. If a track misses the RICH-1 detector, its identity will be harder to calculate, decreasing its usefulness in later analysis. As shown in Fig. 9c, this cut removes more background than signal, and, as explained above, the signal removed is of a lower quality than what is kept.

7 Optimisation of selections

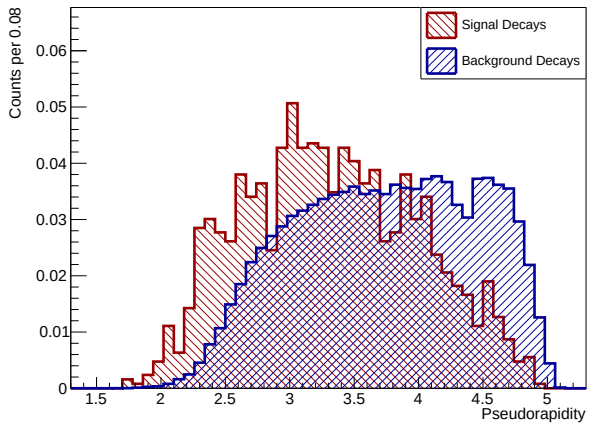
An algorithm was created to run thousands of combinations and find the optimal selections with the maximum signal significance. The algorithm was written to iterate through a range of selections and return the retention rate and signal significance for each one. To confidently find a global maximum for signal significance, over 15,000 combinations of selections had to be run, which meant the process of applying a selection had to be made quicker. When a selection was made at the start of the project, every pair that made it through the initial cuts was paired up to create composite particles. This would take up to six minutes. To increase the speed of the selection application, every pairing of tracks, that passed a relaxed initial set of cuts, was cached with all the relevant information. This allowed each selection to skip the composite particle section, reducing the run time to under two seconds per selection. To reduce the amount of selections that had to be run, only the most effective cuts were varied. These were the track impact parameter chi-squared, track transverse momentum, direction angle (DIRA),



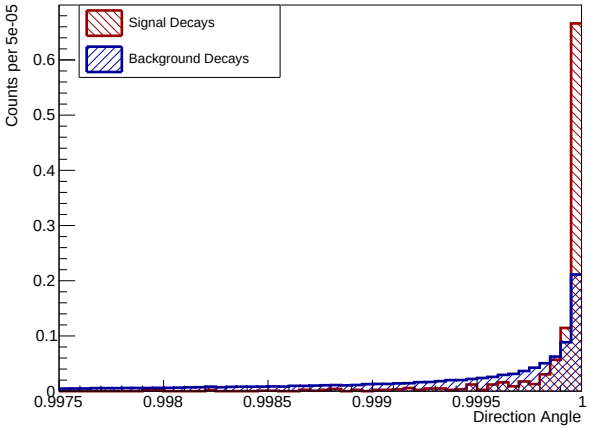
(a) Normalised radial distance distributions



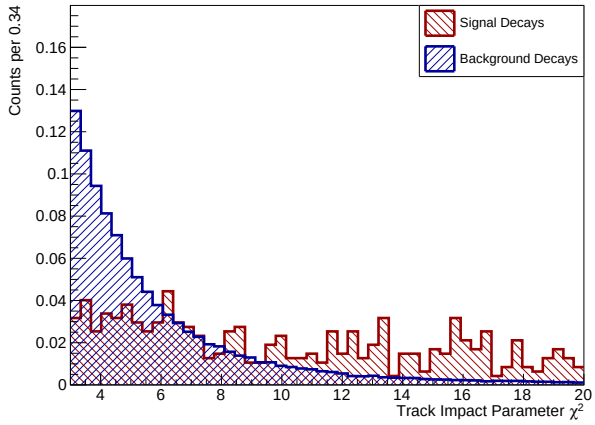
(b) Normalised decay vertex χ^2 distribution



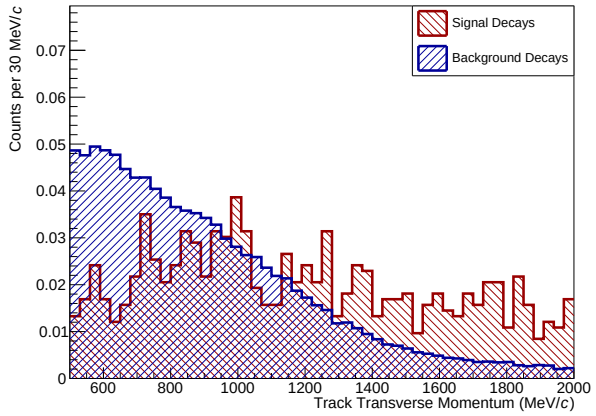
(c) Normalized pseudorapidity distribution



(d) Normalised DIRA distributions

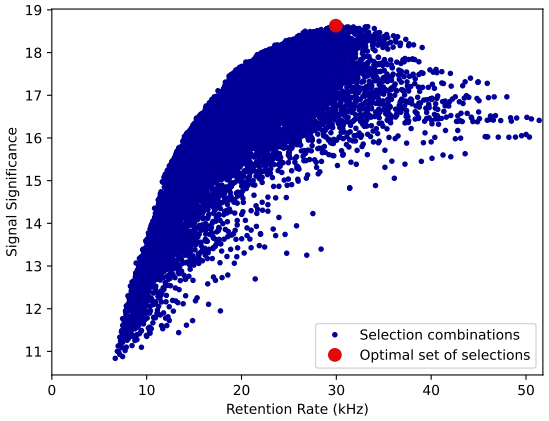


(e) Normalised Track χ^2_{IP} Distributions

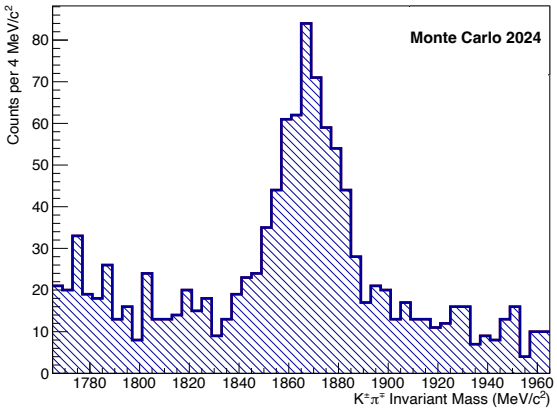


(f) Normalised track p_T distribution

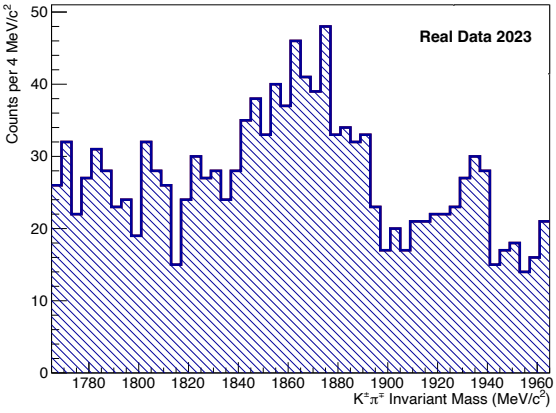
Figure 9: Graphs showing the difference in the signal and background distribution for multiple parameters.



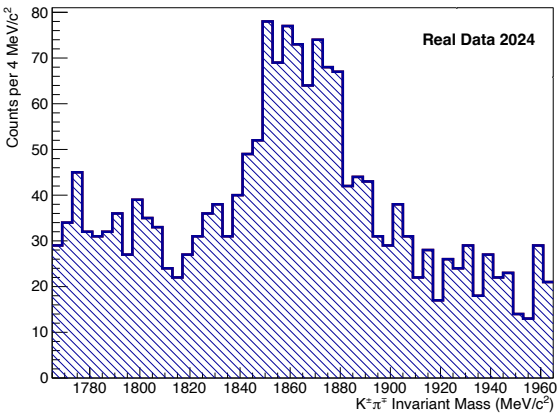
(a) Plot of signal significance against retention rate, resulting in Tab. 2



(b) Invariant mass distribution of the 2024 Monte Carlo minimum bias data.



(c) Invariant mass distribution of the 2023 real data



(d) Invariant mass distribution of the 2024 real data

Figure 10: Graphs showing the results of the optimisation algorithm. The optimal combination of selections, shown in Tab. 2, is applied to three data sets, resulting in a clear peak.

radial distance travelled, mother particle impact parameter chi-squared and the decay vertex error. Although Fig. 9c indicates that the pseudorapidity cut can be further optimised, other cuts have been proven more effective and are therefore varied instead. Using Fig. 9, and Fig. 8b each parameter's optimal value was roughly estimated before a sensible range was set around this. If the combination that resulted in the highest signal significance had any values at the edge of the range set, the range was widened, such that the old value sat centrally instead. This was repeated until a clear global maximum was found as shown in Fig. 10a, with the optimal selections in Tab. 2. These final selections were applied to the Monte Carlo minimum bias and resulted in a retention rate of 29.9±1.1 kHz, with a purity of 59.3±1.6% and a signal significance of 18.6. As shown in Fig. 10b, this results in a prominent peak that is statistically significant. At the LHC, a signal significance of 5 σ is needed for a discovery. Although the $D^0 \to K\pi$ decay is well-known, the selections have been able to get a very high significance at just the HLT1 stage, which is very promising for future rare decays.

Cut	Range of Value
Track χ^2_{IP}	> 10
Track p_T	$> 500\text{MeV}/c$
DIRA	> 0.9995
Radial Distance	$> 0.1\text{mm}$
D^0 χ^2_{IP}	< 3.25
Decay Vertex $\tilde{\chi}^2_v$	< 7

Table 2: The optimal combination of selections that were varied in the algorithm

Dataset	Retention Rate (kHz)	Purity	Signal Significance
Monte Carlo 2024	29.9 ± 1.1	ϵ_p =59.3±1.6% $\epsilon_p^e = 40.1 \pm 2.9\%$	$S = 18.6$ $S^e = 12.8$
Real 2023	350.1 ± 11.5	$13.2 \pm 2.7\%$	4.9
Real 2024	13.7 ± 0.4	$30.8 \pm 2.5\%$	11.9

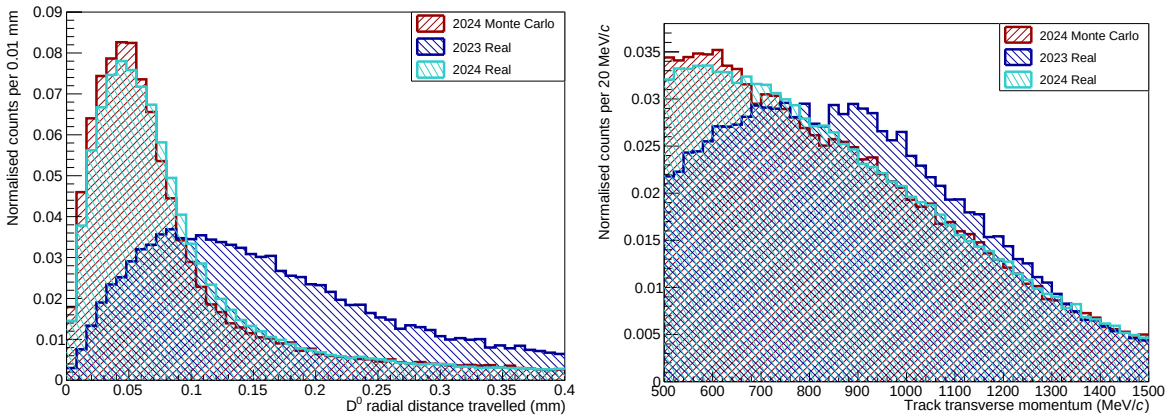
Table 3: Retention rates of the final selections, Tab. 2, on three different data sets. ϵ_p^e and S^e are the estimated signal purity and signal significance, using the method discussed in Sect. 7.2.

7.1 Retention rates

The final optimal selections found in Tab. 2 were applied to the 2023 real, 2024 real and the 2024 minimum bias Monte Carlo data sets. The resulting invariant mass distributions are shown in Fig. 10. A clear peak is shown in all three data sets, with the most prominent being Fig. 10b, followed by Fig. 10d and Fig. 10c. These three data sets had different retention rates when the same selections were run. Tab. 3 shows that the 2023 data's retention was more than 10x higher than the Monte Carlo, whereas the 2024 data was a factor of 2 less.

The factor of ten difference between real 2023 and Monte Carlo 2024 can be explained by the fact that the VELO detector was open, as explained in Sect. 1.2. When the VELO is open, a track will need more p_T or to travel further to be detected. This can be seen in Fig. 11a and Fig. 11b. Fig. 11a shows the radial distance distributions for the different data sets. From here it is clear that by applying a $\Delta R > 0.1\text{mm}$ cut, proportionally more data is being cut in the 2024 MC and real samples compared to the 2023 real sample. The difference in distribution shape between the 2024 and 2023 data sets implies that the radial distance resolution is much better in 2024, which aligns with the VELO being open in 2023. The track transverse momentum, shown in Fig. 11b has a similar effect. When the VELO is retracted, it will catch a smaller amount of low p_T tracks, resulting in the lower resolution shown in Fig. 11b. Therefore, a $p_T > 500\text{MeV}/c$ cut will remove proportionally less decays in the 2023 sample compared to the 2024 samples.

The 2024 Monte Carlo and real distributions are very similar over all of the selection parameters, as expected. However, this means that the factor of 2 difference seen in Tab. 3 has to come from somewhere else. The Monte Carlo data sets simulate 5.3 inelastic proton-proton collisions per event, as this was the goal for 2024 data taking. At the time of writing, LHCb has only been able to run stably at an average of 2-3 collisions per event. This would fully account for the difference in retention rates seen in Tab. 3, as in each event number of $D^0 \to K\pi$ decays produced has approximately decreased by a factor of two.



(a) Distributions of D^0 radial distance travelled. **(b)** Distributions of track transverse momentum.

Figure 11: The distributions of the radial distance travelled and track transverse momentum are compared across three data sets, 2024 Monte Carlo, 2024 Real and 2023 Real.

7.2 Estimated signal significances

To finalize the results on real data, the purity and signal significance is estimated using a crude method. In Fig. 10b, 10c, 10d, there are 50 bins over 200 MeV/c^2. It is assumed that all decays outside of the 1815-1915 MeV/c^2 range are background. The number of decays outside the 1815-1915 region divided by the number of bins calculates the average background decays per bin. This can be used to calculate the number of signal decays and therefore purity, as

$$\epsilon_p^e = \frac{N_i - N_o \frac{B_i}{B_o}}{N_i + N_o}, \tag{15}$$

where N_i and B_i are the number of decays and bins inside 1815-1915 MeV/c^2 and N_o, B_o are the number of decays and bins outside the region. As $B_o = B_i = 25$, Eq. 15 simplifies easily. The estimated purity uncertainty can be calculated through error propagation Eq. 3, to result in

$$\sigma(\epsilon_p^e) = 2\sqrt{\frac{N_i^2 N_o + N_o^2 N_i}{(N_i + N_o)^4}}. \tag{16}$$

The signal significance can also be estimated through

$$S^e = \frac{N_i - N_o}{\sqrt{N_i + N_o}}. \tag{17}$$

These results are shown in Tab. 3. It is important to note that this method is not the most accurate, assuming that all decays outside of the 1815-1915 region are background is not a good estimate, as it is known that the signal decays can span across 1865 ± 100MeV/c^2. This means that the estimated purity and signal significance will be lower than calculated by Monte Carlo properties. This can be seen in Tab. 3, where ϵ_p^e is 1.5x smaller than ϵ_p, and a similar relation for S and S^e. To increase the accuracy of this estimation, the background distribution could be better modelled. As seen in Figs. 10b, 10c, 10d, the distribution is higher on the left side of the peak. In the current estimation, the background is assumed to be uniformly distribution, however, if a linear fit is performed and integrated, the average number of background decays per bin can be better calculated.

Single Track Selections	**Value**
Track p_T	> 500 MeV/c
Track p	> 4000 MeV/c
Track χ^2_{IP}	> 10
Track η	< 4.38
Composite Particle Selections	**Value**
Track charges	Opposite
Distance of nearest approach	< 0.2 mm
Decay vertex $\tilde{\chi}^2_v$	< 7
Composite mass cut	1865 ± 100 MeV/c^2
DIRA	> 0.9995
ΔR	> 0.1 mm
D^0 χ^2_{IP}	< 0.325

Table 4: A list of all the cuts in order of application. Cuts not mentioned in this report have not been changed since [1] [2].

The estimated signal significance of the 2024 Monte Carlo and real data are very similar. This has confirmed what is expected, despite the difference in retention rate and pile-up, the composition of each event should still be similar and therefore, the selections should have a similar effect. The improvement that the VELO caused in the quality of decays is further shown in the increase in purity and significance between 2023 and 2024 real data.

8 Conclusion

This report has thoroughly investigated the composition of background decays from the initial selections in Tab. 1. Through this investigation, the source of an unexpected peak has been identified as the Monte Carlo truth matching system misidentifying kaons and pions for electrons, classifying $5.78\pm0.15\%$ of signal decays wrongly as background. The cause of this error has been investigated further, narrowing it down to be due to material interactions or inflight decays. This knowledge will help when corrections are being implemented to the truth-matching system. The report has also shown the effectiveness of new cuts such as the radial distance and impact parameter-chi squared and has optimised such selections to achieve an 18.6 signal significance at a retention rate of 29.9 ± 1.1 kHz on the Monte Carlo minimum bias data set. These same selections also resulted in an 11.9 signal significance at a retention rate of 13.7 ± 0.4 kHz on some of the first available 2024 real data. This is expected to improve as the average number of proton-proton collisions per event increases. While this report has focused on optimising specific selections, there are many more that have been used that have not been changed since last semester's report. For completeness, Tab. 4 shows all the selections implemented in order. The next step to continue this project would be to implement the optimal selections into the actual HTL1 trigger. During this project, the selections have been applied to a saved data set, that has already undergone track reconstruction and a loose trigger stage. To make sure these selections can be used on live data, the computational complexity has to be considered. The HLT1 only has 25 ns to determine if the event is to be saved before the next event occurs. This

could slightly shift the weighting of selections away from computationally expensive calculations, such as the D^0 χ^2_{IP} and towards quicker selection such as p_T. The methodology used throughout this project can also be applied to other, rarer decays such as multibody $D^0 \to \pi^{\pm}\pi^{\mp}\pi^0$, which could prove more illuminating in the search for charge-parity violation [13].

References

[1] R. Hartley, "Finding Charm at the LHCb - An Analysis of Trigger Selections on $K_s \to \pi\pi$ and $D^0 \to K\pi$ Decays," *MPhys Report - University of Manchester*, 2024.

[2] E. Peak, "Finding Charm at the LHC - Analysing LHCb Trigger Selections," *MPhys Report - University of Manchester*, 2024.

[3] J. Wenninger, "Machine Protection and Operation for LHC," *CERN Yellow Reports*, p. Vol 2 (2016): Proceedings of the 2014 Joint International Accelerator School: Beam Loss and Accelerator Protection, 2016.

[4] M. Schaumann, D. Gamba, H. Garcia Morales, R. Corsini, M. Guinchard, L. Scislo, and J. Wenninger, "The effect of ground motion on the LHC and HL-LHC beam orbit," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 1055, p. 168495, 2023.

[5] LHC Machine Outreach, "The LHC Beam Dumps." Accessed at `https://lhc-machine-outreach.web.cern.ch/components/beam-dump.htm` on 26th April 2024.

[6] O. Brüning et al, *LHC Design Report*. CERN Yellow Reports: Monographs, Geneva: CERN, 2004. Chapter 17, pg. 441 .

[7] LHC Program Coordination, "Filling Scheme Viewer." Accessed at `https://lpc.web.cern.ch/cgi-bin/fillTable.py` on 30th April 2024.

[8] E. Minucci, "154th LHCC meeting Open Session." Accessed at `https://cds.cern.ch/record/2862130/files/LHCCJune2023_LHCb_EMinucci.pdf` on 30th April 2024.

[9] T. Sjöstrand, "Monte Carlo event generation for LHC," 1992.

[10] M. Needham, "Classification of Ghost Tracks," tech. rep., CERN, Geneva, 2007.

[11] S. Tolk, J. Albrecht, F. Dettori, and A. Pellegrino, "Data driven trigger efficiency determination at LHCb," tech. rep., CERN, Geneva, 2014.

[12] R.L. Workman *et al.* (Particle Data Group), "Review of Particle Physics," *Progress of Theoretical and Experimental Physics*, vol. 2020, p. 083C01, 08 2020.

[13] M. Gaspero, V. Crede, P. Eugenio, and A. Ostrovidov, "Study of the $D^0 \to \pi^+\pi^-\pi^0$ decay at BABAR," in *AIP Conference Proceedings*, AIP, 2010.