sPH-HF-2023-001
v0.1
January 13, 2023

$D^0 \to K^-\pi^+$ modelling and selection in min-bias Au+Au simulations at sPHENIX

Cameron Dean

Massachusetts Institute of Technology

Abstract

The reconstruction and selection of $D^0 \to K^-\pi^+$ in simulated Au+Au collisions with the sPHENIX detector is presented. Approximately 22 million minimum bias events were generated at $\sqrt{s_{NN}} = 200\,\text{GeV}$ using the HIJING event generator, corresponding to roughly 20 minutes of data taking at a collision rate of 15 kHz.

Contents

1 Introduction

The reconstruction of open-charm hadrons is a key requirement of the sPHENIX experiment which will commence data-taking in Spring 2023. As with any new detector, the initial data-taking period will be devoted to commissioning the experiment. This involves ensuring the correct operation of all subsystems, the tracking, calorimeter calibration and calculation of physics objects. Preparations for the latter requirement is the focus of this note. A detailed commissioning timeline has been agreed by the collaboration which is publicised within the Beam Use Proposal [1]. The timeline is given in Table 1 for completeness.

Weeks	Details
2.0	low rate, 6-28 bunches
2.0	low rate, 111 bunches, MBD L1 timing
1.0	low rate, crossing angle checks
1.0	low rate, calorimeter timing
4.0	medium rate, TPC timing, optimization
2.0	full rate, system test, DAQ throughput
12.0	**Total**

Table 1: Timeline for sPHENIX commissioning period in 2023, the first year of operation. Copied from the sPHENIX 2022 beam use proposal [1].

A challenging aspect of the commissioning period will be the reconstruction of $D^0 \to K^-\pi^+$[1] as it is both reasonably rare and the decay products often have a transverse momentum (p_T) less than 1 GeV. This channel is also key to realising the b-physics program of sPHENIX as the separation of prompt and non-prompt open-charm decays is used to tag probable b-hadrons [2, 3, 4]. The number of $D^0 \to K^-\pi^+$, $N_{D^0\to K^-\pi^+}$, produced is proportional to the integrated luminosity, $\mathcal{L}$, and is roughly given by

$$N_{D^0\to K^-\pi^+} \approx \mathcal{L}\sigma_{c\bar{c}} 2 N_{\text{coll}} f_{D^0} \text{BF}(D^0 \to K^-\pi^+) \varepsilon^2_{\text{acc}} \varepsilon^2_{\text{track}} \varepsilon_{\text{sel}} \tag{1}$$

where $\sigma_{c\bar{c}}$ is the production cross section for charm-quarks (about $1/60^{\text{th}}$ of the total inelastic cross-section [5, 6]), N_{coll} is the number of binary collisions, f_{D^0} is the D^0 fragmentation fraction (about 40% [7]), BF($D^0 \to K^-\pi^+$) is the $D^0 \to K^-\pi^+$ branching fraction (about 4% [8]), ε_{acc} is the geometrical acceptance efficiency (about 23%, see Section 3.2), $\varepsilon_{\text{track}}$ is the tracking efficiency (about 80%, see Section 3.3) and ε_{sel} is the selection efficiency. The factor of two comes from producing 2 charm quarks which are equally likely to go through a $D^0 \to K^-\pi^+$ decay and the squares on the acceptance and tracking efficiencies comes about from both efficiencies applying equally to each daughter track. This means the the expected yield of $D^0 \to K^-\pi^+$ is given by

$$N_{D^0\to K^-\pi^+} \approx \mathcal{L} N_{\text{coll}} \varepsilon_{\text{sel}} \times 10^{-5} \tag{2}$$

[1] Charge conjugation is implied throughout this note unless otherwise stated.

The aim of this note is to understand the topological and kinematic variable distributions of D^0 decays in Au+Au collisions at sPHENIX and hence optimise ε_{sel}. This note is intended to act as a baseline reference for the initial selection of $D^0 \to K^-\pi^+$ in the commissioning period. Several aspects of the experiment data pipeline will be improved during the commissioning period as the collaboration sees and understands the initial data, most importantly the tracking. It is expected that the momentum resolution will undergo improvements in the initial period and this resolution has a direct impact on the D^0 resolution. Due to the large data volume and need for a fast turnaround on physics object construction, we will pre-select tracks that share signatures with a heavy flavor decay (as examples, tracks with a large p_T or large distance-of-closest approach to the primary vertex) and then save these track seeds and clusters to a smaller data file to re-run the track fitting with improved alignment parameters. It is important then to understand what these variables look like and select tracks with loose enough cuts to catch tracks that may not have the best resolution yet but the cuts need to be tight enough to reject a large portion of the background. The reconstruction of D^0 then serves as an important indicator that the sPHENIX collaboration is ready to produce physics results.

The note after this introduction is arranged as follows; the general simulation setup of the generators and detector, along with the generated statistics is detailed in Section 2, the tools developed within sPHENIX to understand the decay topology are described in Section 3, the base selections applied to the simulation data set is detailed in Section 4, the models used to describe the invariant mass distributions are described in Section 5 and the results are presented in Section 6.

To avoid biases, the signal was modelled using $D^0 \to K^-\pi^+$ decays simulated in p+p collisions using PYTHIA8 while the background was modelled using HIJING events outside of the invariant mass search window. Comparisons of the signal-to-background ratio are made between candidates selected using direct cuts and machine-learning methods.

2 Simulation

This study used two different data sets. One data set consisted of p+p collisions at $\sqrt{s} = 200\,\text{GeV}$ using PYTHIA8 [9] as an event generator. PYTHIA8 was tuned to approximate the minimum bias collision environment at RHIC [10] but, after the initial generation, each event was required to have a $c\bar{c}$ pair produced to enrich the $D^0 \to K^-\pi^+$ statistics. The RHIC collision rate for p+p collisions is almost 10 MHz and some of the sPHENIX detectors integrate their hits over a longer period that this and hence, there is an out-of-time pileup effect. This effect was implemented as part of our simulation. A total to 50 million events were generated in this fashion and they were used to model the $D^0 \to K^-\pi^+$ signal shape and variable distributions. This was intended to avoid biases in the final selection and fitting by not using the same $D^0 \to K^-\pi^+$ candidates for both selection and testing.

The second data set consisted of minimum bias Au+Au collisions at $\sqrt{s_{NN}} = 200\,\text{GeV}$ using HIJING [11]. Unlike the p+p sample, there was no further requirements on the generator to enrich the $D^0 \to K^-\pi^+$ purity. This sample was used for both background modelling and for the final test of the selection. The background sample was taken from a region outside the invariant mass range to ensure there is no bias in the final selection.

Almost 22 million events were generated which corresponds to a little over 20 minutes of data taking at the peak RHIC Au+Au collision rate of 15 kHz.

The detector was simulated using the GEANT4 package [12]. The simulation consists of the beam pipe and all detectors, except the event plane detector (sEPD) which is not used for this study. In the interaction region, the beam pipe consists of a beryllium section with an inner radius of 2 cm and a thickness of 762 µm. The volume inside the beam pipe is simulated as a vacuum. The monolithic active pixel sensor vertex detector (MVTX) is simulated after the beam pipe and extends from 2.4 cm to 5.3 cm. It consists of a three layer vertexing detector with a timing resolution of approximately 5 µs and hence sees the out of time pile-up effect. This effect is accounted for by duplicating the hits with a time delay in GEANT4. The next detector radially is the intermediate tracker (INTT) which consists of a two-layer silicon strip detector with a timing resolution capable of resolving the 10 MHz collision rate. The INTT sits between 7.2 cm and 10.3 cm, radially. A compact time projection chamber (TPC) sits beyond the INTT up to a radius of 80 cm and provides the momentum measurement for tracks in sPHENIX. The TPC integrates over a period of 34 µs and also sees the out-of-time pileup effect. The final tracking detector is the TPC outer tracker (TPOT) which consists of 8 GEM detectors below the TPC in the vertical direction. This acts as a final measurement point in the tracking to account for the distortions and drift in the TPC tracks. As well as the tracking detectors, there are two hadronic calorimeters and an electromagnetic calorimeter in the simulation. As they are not used to identify the D^0 candidates, they are not described in this note. Between the two hadronic calorimeters, there is a 1.5 T superconducting solenoidal magnet which is also simulated. The tracking is performed using A Common Tracking Software (ACTS) [13, 14].

3 Toolkit

3.1 KFParticle

The candidate reconstruction was performed by KFPARTICLE, originally designed by the CBM collaboration [15] and adapted to be used within the Fun4All framework [16]. The internal logic of the package has been largely unchanged since its previous description in the sPHENIX framework [17] except for small bug fixes or additional output variables for users. The main update since the previous note is the addition of a "decay descriptor" to simplify the user interface.

The decay descriptor is a string that users write to specify the decay topology and is parsed by the top interface class before any event processing. If the descriptor can not be understood by the parser, a warning will be raised and the module will not be added to the node tree although Fun4All will still run. The decay parser checks each particle the user specifies against the particle database found in ROOT's TDatabasePDG to ensure it exists. If the particle does not exist, this will be written to the user. Similarly, if the parser cannot interpret the charge of a track, the user will also be notified of the offending track. All other instances of the parser not understanding the string will be written as a standard warning. An example decay descriptor is as follows

```
[B+ -> {D0bar -> K^- pi^+} pi^+]cc
```

This is interpreted as a B^+ hadron decaying to an intermediate D^0 and an isolated charged pion. The mother is always written to the left of a `->` and intermediate decays are always contained within {} braces. The charge of a track is written as either +, - or 0 directly after a caret (`^`). If you wish to also search for the charge conjugate decay, the entire descriptor must be contained with square braces and appended with `cc` or `CC` for charge-conjugate. From this, KFParticle knows how many intermediate decays there are, how many tracks are in each intermediate state and final state as well as the charges of all tracks. It also knows if it needs to bring the intermediate states back to a common vertex or associate them to any other final state tracks.

3.2 DecayFinder

DecayFinder is a new package that runs over the truth record of an event. It searches for decays specified by the user and accepts input ranges for the final state particles pseudorapidity, η, and p_T. It effectively measures the geometrical acceptance of a decay. DecayFinder also uses a decay descriptor with the same logic as KFParticle, so users can just repeat the string they wrote previously. DecayFinder begins by looking for the HepMC2 record [18] on the node tree, if this doesn't exist then it will fall back to using the Geant4 truth record. The tool loops over the truth container to find the required mother and, when this is found, the decay products will be analysed. There is an internal list of resonances that will be further analysed if seen in the mothers decay chain. For example, a $\phi(1020)$ resonance often decays to two kaons. If a user is looking for two kaons in the final state, then the $\phi(1020)$ will be studied further. This resonance would not automatically be seen by sPHENIX without reconstructing the dikaon pair first. If the $\phi(1020)$ is specified in the decay descriptor, it will be removed from the internal resonance list and treated as an integral part of the decay chain.

It is possible to limit the decay volume of some event generators such as Pythia8. If this occurs, the HepMC2 record will not contain the full information of the event. In this case, DecayFinder can detect that the decay volume was limited and switch to searching the Geant4 truth record when this boundary is discovered[2]. When DecayFinder detects a final state track ($e^{\pm}$, $\mu^{\pm}$, $\pi^{\pm}$, $K^{\pm}$, p or $\bar{p}$), it calculates the track's η and p_T and compares this to the user specified values[3]. If all tracks pass the requirement, DecayFinder tags the decay as reconstructable. The barcodes, embedding IDs and PDG IDs of all particles in a reconstructable decay can be written to the node tree for further analysis. When DecayFinder finishes, it can write a report of how many decays were generated, how many had at least one track fail the p_T requirement or had at least one track fail the η requirement or had at least one track fail both the p_T and η requirements and finally how many decays fell within the required acceptance. DecayFinder is also capable of triggering on events that have all particles within the sPHENIX acceptance.

[2]This boundary appears as a null pointer when calling a particle's end vertex in HepMC2

[3]The default values are $p_T \geqslant 0.2\,\text{GeV}$ and $|\eta| \leqslant 1.1$

3.3 HFTrackEfficiency

HFTrackEfficiency uses the output of DecayFinder to match final state tracks to the decay products that DecayFinder claims are all in the required acceptance region. When each trackable particle is found in the truth record, the reconstructed track map is iterated over and the 3-momentum of the truth and reconstructed objects is compared. If all three values match within a specified limit[4] then HFTrackEfficiency counts that particle as reconstructed. The user can specify that they would like to have the results of the search output to an nTuple. This nTuple contains the truth momenta, PID and η for each trackable particle as well as the reconstructed momenta of the track, if it exists. There is also a boolean flag to say whether that track was or was not reconstructed. The mothers true momenta, PID and η are also written to the file along with the reconstructed mass as seen by ACTS, assuming all tracks were reconstructed. This files gives a direct look at the tracking efficiency for heavy flavor particles.

As with DecayFinder, HFTrackEfficiency is also capable of triggering when all final state particles are reconstructed. As well as this triggering, it can also write the reconstructed tracks back to the node tree in a subset of the track map. This is only done when all tracks are reconstructed for a decay and so allows users to build a decay without any selections which is useful for studying the decays kinematic distributions.

4 Selection

The D^0 candidates were selected using kinematic and topological variable cuts. The cuts are applied to the daughter tracks, secondary vertex and reconstructed mother candidates and requires knowledge of the primary vertex (PV). The tracks are selected based on their transverse momentum (p_{T}), track χ^2 per number of degrees of freedom and minimum distance of closest approach with respect to each reconstructed primary vertex (IP). The secondary vertex (SV) is selected by making pairs of tracks and measuring their distance of closest approach with respect to each other (DCA) and the χ^2 per number of degrees of freedom of the reconstructed SV. The mother candidates are selected by requiring they lie within an invariant mass range ($m_{K^-\pi^+}$), they have a minimum IP χ^2 with respect to its selected PV, a minimum cosine of the angle between the flight direction and mother momentum vector (DIRA), and p_{T}. Further variables were also studied such as the quality of the separation of the primary and secondary vertices (the flight distance χ^2) and the quality of the track IP. These variables were found to have little separation power and improving their calculations will be a task to undertake before the initial sPHENIX data taking. A visual description of the variables is given in Figure 1.

To ensure the final selection was unbiased, signal and background samples were taken from events that were not used in the final simulation. The signal events were taken from a simulated sample of minimum bias p+p collisions using Pythia8 which were filtered at the generator level to ensure each event contained a $c\bar{c}$ pair. DecayFinder was run on each event to select $D^0 \to K^-\pi^+$ decays and required the minimum track p_{T} was greater than 0.16 GeV with a track pseudorapidity (η) between $-1.6 \leqslant \eta \leqslant 1.6$. DecayFinder was required to

[4]The default matching limit is 5%

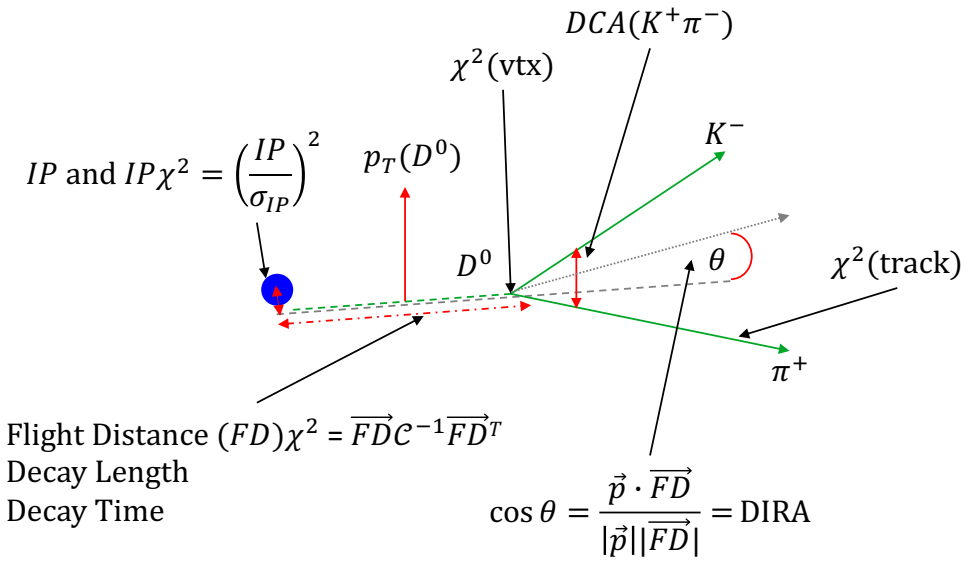

Figure 1: Visual description of the kinematic and topological variables used to select the D^0 candidates.

trigger on this selection. For all events which passed this check, HFTrackEfficiency was then run to match the true kaon and pion tracks to a reconstructed track within 5% for p_x, p_y and p_z and write these tracks to a new track map. KFParticle was then configured to run on this new track map to ensure a pure $D^0 \to K^-\pi^+$ sample. The only selection requirement in KFParticle was that the $K^-\pi^+$ pair falls within the range $1.70 \leqslant m_{K^-\pi^+} \leqslant 2.00$.

The background sample was taken from a simulated sample of minimum bias Au+Au collisions using HIJING. 10 thousand events were used with no selection other than the $K^-\pi^+$ pair falls within the range $2.00 \leqslant m_{K^-\pi^+} \leqslant 2.10$.

ROOT's multi-variate analysis toolkit (TMVA) [19] can read in data sets where users specify variables to use for multivariate analyses. Useful features of this kit is it can automatically overlay the signal and background distributions of each variable, measure the variable correlations and show the agreement between training and testing samples. The signal and background samples were fed into TMVA where the event number that is stamped by Fun4All is used to separate training and testing samples. The training samples used exclusively odd numbered events while the testing samples used even numbered events. The training and testing samples' invariant mass distributions are shown in Figure 2. 277005 candidates were used for the signal sample (138720 for training) and 884752 were used for the background sample (449199 for training). The variable comparisons are shown in Figure 3 while the variable correlations are shown in Figure 4. TMVA was run only using cuts-based methods, the associated receiver operating characteristic (ROC) plot is shown in Figure 5. Using TMVA and assuming that there are 10000 background candidates to every signal candidate, the package predictions an optimal signal efficiency of approximately 2%.

With this information, the baseline cuts to be applied to the minimum bias Au+Au sample are given in Table 2. Before these cuts, there were 300755 D^0 signal candidates from the p+p simulation and 1782139 background candidates from the Au+Au. After applying these cuts, there were 39808 D^0 signal candidates from the p+p simulation and 978 back-

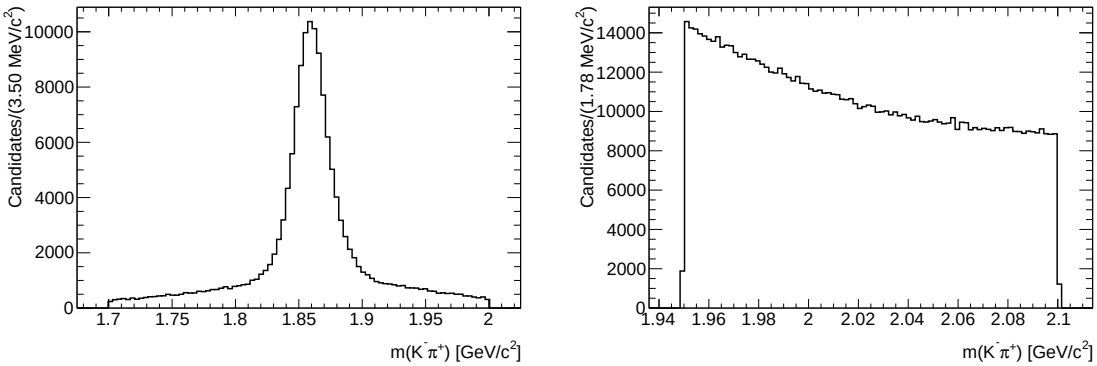

Figure 2: Invariant mass distributions of the signal and background samples used to make the selection decisions. The signal sample is shown on the left and the background sample is shown on the right.

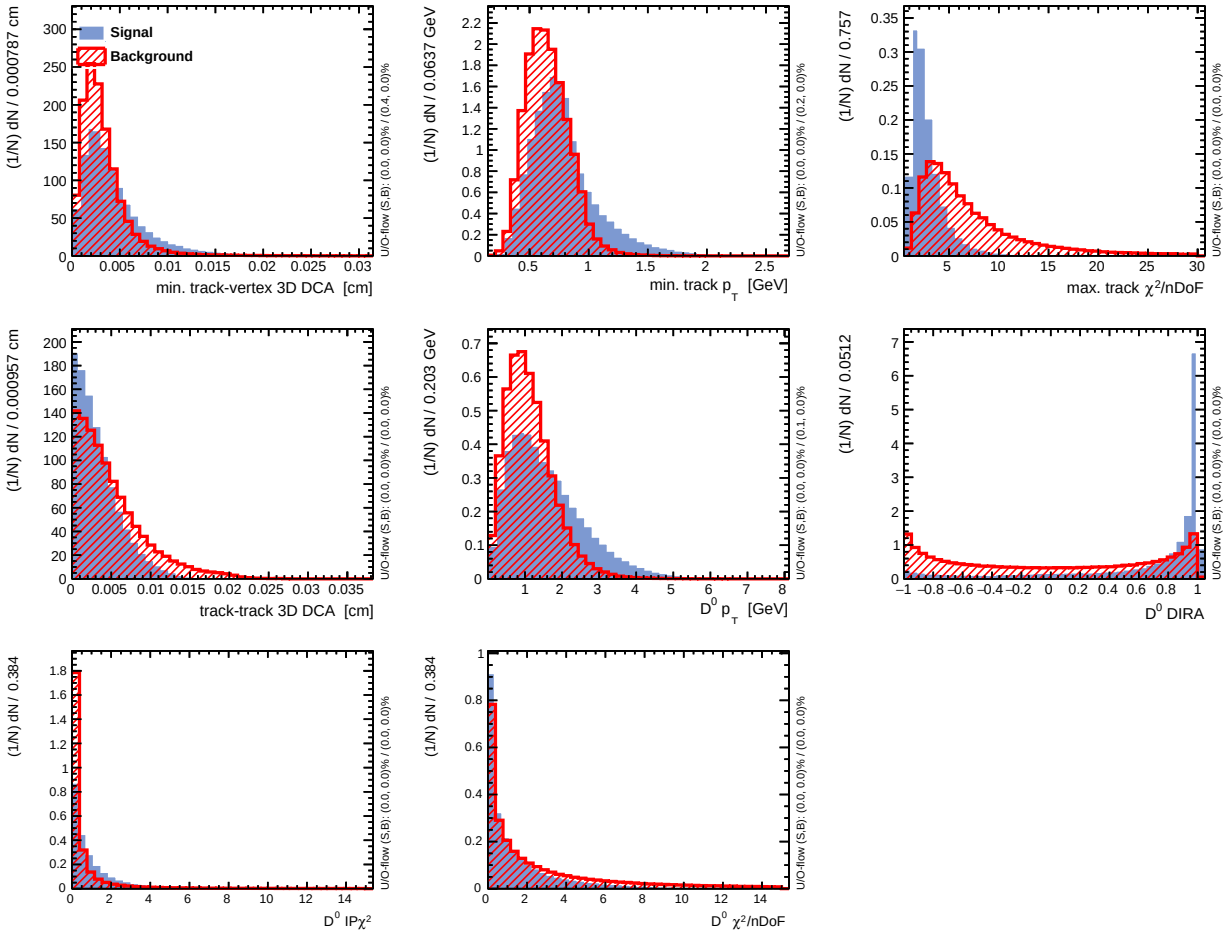

Figure 3: Input variable distributions of the signal (blue, whole) and background (red, dashed) samples used to make the selection decisions.Variables from top to bottom, left to right: minimum track IP, minimum track p_{T}, maximum track χ^2 per number of degrees of freedom, track-track DCA, D^0 p_{T}, D^0 DIRA, D^0 IPχ^2, D^0 χ^2 per number of degrees of freedom.

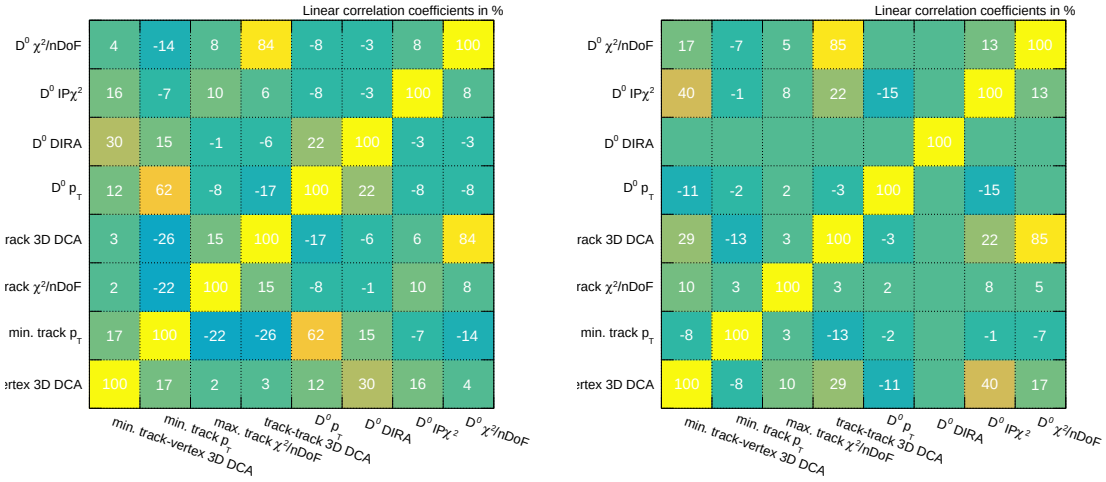

Figure 4: Input variable correlation coefficients of the signal (left) and background (right) samples used to make the selection decisions.

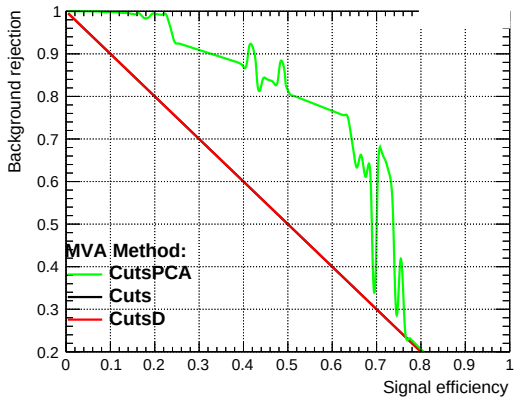

Figure 5: Receiver operating characteristic plot from the pre-selected variable distributions using cut methods.

ground candidates from the Au+Au. This gives a 13.24% efficiency on the signal and 0.05% efficiency on the background selection. It should be noted that the invariant mass range in Table 2 is slightly larger than what is used in the final fitting ($1.70 \leqslant m_{K^-\pi^+} \leqslant 2.00$). This is to obtain sufficient background samples on either side of the D^0 mean mass to allow for a machine-learning study.

Variable	Cut
min. track IP [cm]	0.0025
min. track p_{T} [GeV]	0.7
max. track χ^2/nDoF	5
max. track-track DCA [cm]	0.008
D^0 IPχ^2	3
D^0 p_{T} [GeV]	1.5
D^0 χ^2/nDoF	5
D^0 DIRA	0.90
$m_{K^-\pi^+}$ [GeV]	1.65 $\rightarrow$ 2.10

Table 2: Baseline cuts used to select $D^0 \rightarrow K^-\pi^+$ candidates.

5 Models

The modelling of the $K^-\pi^+$ invariant mass distribution is performed using the RooFit fitting package [20] provide with ROOT.

5.1 Signal

As there is no PID at sPHENIX, it is possible to reconstruct a D^0 but invert the mass hypothesis for both tracks. The two hypotheses will return different values for the mass. However, if the mass window is sufficiently large, it's increasingly likely to select the incorrect combination. Due to this effect, the invariant mass of the $K^-\pi^+$ pair will appear as the composition of two Gaussian functions, one with a very large width. The PDF describing the signal shape is

$$f(x;\mu,\sigma_{\mathrm{cor-ID}},\sigma_{\mathrm{mis-ID}},k_{\mathrm{cor-ID}}) = N \cdot \left[k_{\mathrm{cor-ID}} \exp\left(-\frac{(x\ -\ \mu)^2}{2\sigma^2_{\mathrm{cor-ID}}}\right) + (1-k_{\mathrm{cor-ID}}) \exp\left(-\frac{(x\ -\ \mu)^2}{2\sigma^2_{\mathrm{mis-ID}}}\right)\right] \tag{3}$$

where μ is the mean value, $\sigma_{\mathrm{cor-ID}}$ is the width of the Gaussian describing the candidates with the correct mass hypothesis, $\sigma_{\mathrm{mis-ID}}$ is the width of the Gaussian describing the candidates with the incorrect mass hypothesis and $k_{\mathrm{cor-ID}}$ is the fraction of candidates belonging to the Gaussian with the correct mass hypothesis. When fitting the signal extracted from

the p+p baseline, all parameters are left floating. However, when fitting the final distribution $k_{\mathrm{cor-ID}}$ is fixed to the value from the p+p simulation and $\sigma_{\mathrm{mis-ID}}$ is scaled from $\sigma_{\mathrm{cor-ID}}$ by the ratio of widths as measured in the p+p simulation to account for the difference in momentum resolution between p+p and Au+Au events

$$\sigma^{\mathrm{Au+Au}}_{\mathrm{mis-ID}} = \sigma^{\mathrm{Au+Au}}_{\mathrm{cor-ID}} \frac{\sigma^{p+p}_{\mathrm{mis-ID}}}{\sigma^{p+p}_{\mathrm{cor-ID}}} \tag{4}$$

The results of the fit to the signal extracted from p+p simulated events are given in Table 3 and Figure 6 respectively.

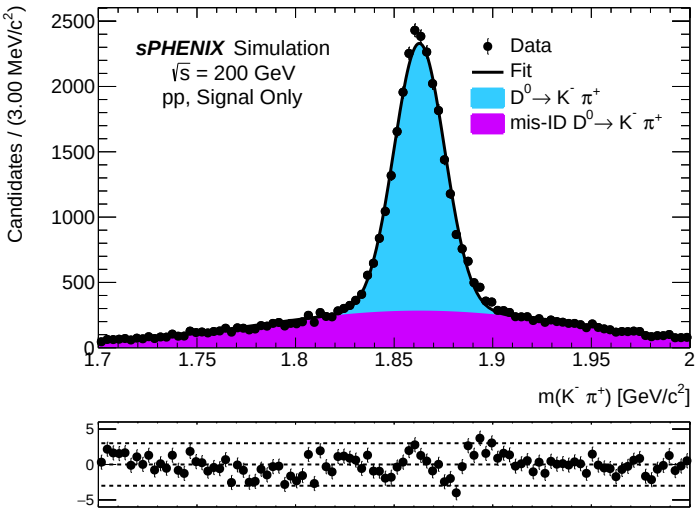

Figure 6: Default fit to the $K^-\pi^+$ invariant mass distribution using simulated p+p events with truth matching.

Parameter	Value
μ[MeV]	1862.78 ± 0.11
$\sigma_{\mathrm{cor-ID}}$[MeV]	12.83 ± 0.12
$k_{\mathrm{cor-ID}}$ [%]	54.62 ± 0.43
$\sigma_{\mathrm{mis-ID}}$[MeV]	82.89 ± 1.00

Table 3: Default fit parameters to the $D^0 \to K^-\pi^+$ p+p signal sample.

5.2 Background

The background is modeled using an exponential function as it was felt this would have a better handle on the combinatorial background from the Au+Au simulation than a linear function would. The background PDF is

$$f(x;k) = N \cdot \exp(\lambda x) \tag{5}$$

where λ is the decay constant.

5.2.1 Additional background sources

At this stage, there are no additional background sources modeled. However, there are several sources that could contribute:

- $D^0 \to K^-\pi^+\pi^0$ where the diphoton daughters are not used in the reconstruction. The branching ratio of $D^0 \to K^-\pi^+\pi^0$ to $D^0 \to K^-\pi^+$ is approximately 3.5 and would lie to the lower mass region.

- $D^+ \to K^-\pi^+\pi^+$ where one of the pions are missed. The fragmentation fraction times branching ratio of $D^+ \to K^-\pi^+\pi^+$ to $D^0 \to K^-\pi^+$ is approximately 1.0 and would lie to the lower mass region. The decay $D^+ \to K^-K^+\pi^+$, where a kaon is missing is unlikely to contribute to the background in this study as either the mass loss from the kaon would be too great or the mis-ID of one kaon track as a pion as well as a missing track would likely make the channel fall out of the invariant mass window.

- $D_s^+ \to K^-K^+\pi^+$ where one of the kaons are missed. The fragmentation fraction times branching ratio of $D_s^+ \to K^-K^+\pi^+$ to $D^0 \to K^-\pi^+$ is approximately 0.5 and would lie close to the D^0 peak.

- $\Lambda_c^+ \to pK^+\pi^+$ is unlikely to contribute either due to the mis-ID of one track or the missing proton, which carries a significant portion of the energy.

The contribution of these background sources to this channel could be studied using dedicated simulations of these decays where the $D^0 \to K^-\pi^+$ selection is applied and the $K^-\pi^+$ mass hpothesis is applied to the selection. The invariant mass distributions would not follow a predictable shape and so the Kernel method could be applied to model the shapes [21].

6 Results

The baseline selection in Table 2 was applied to the Au+Au simulation without using DecayFinder and HFTrackEfficiency as a trigger to allow for a realistic background contamination. DecayFinder and HFTrackEfficiency were both run alongside KFParticle to calculate the acceptance, tracking, and selection efficiencies. The resulting $m_{K^-\pi^+}$ spectrum is shown in Figure 7 where there is no apparent D^0 mass peak. The baseline cuts were chosen to be very loose so it was expected that they would need to be tightened at a later stage. The loose cuts were chosen to allow for a machine learning study which is detailed in Appendix B.

DecayFinder and HFTrackEfficiency were run requiring that a maximum track $\eta \leqslant 1.6$ and minimum track $p_{\mathrm{T}} \geqslant 0.16\,\mathrm{GeV}$. HFTrackEfficiency also required that the true value of p_x, p_y and p_z match to the reconstructed values within 5%. The p_{T} and η values are right at the periphery of the sPHENIX tracking abilities and only significantly displaced tracks can meet these requirements or else they become loopers or fall outside of

the tracking acceptance. By using these values, the acceptance and tracking efficiency can be underestimated by counting decays that could never be reconstructed. Thus, an improvement in calculating the acceptance and tracking efficiency would be to add a recalculation of the p_{T} and η thresholds based on the position of the secondary vertex. This is outside of the scope of this study but a second calculation of the tracking efficiency is performed with a maximum track $\eta \leqslant 0.5$ and minimum track $p_{\mathrm{T}} \geqslant 1.00\,\mathrm{GeV}$. Further, the track matching requirement can reject lower p_{T} tracks where the momentum calculation is less accurate.

Using the loose requirements to accept a decay, the geometric acceptance is calculated to be 34.5% and the tracking efficiency is calculated to be 7.8%. The geometric acceptance includes $D^0 \to K^-\pi^+$ decays that have an associated photon. The energy of these photons is not known and so it is unknown if these decays could be meaningfully reconstructed if the energy loss is too great. If all decays with an associated photon are assumed to fall outside of the invariant mass window, then the acceptance is reduced to 24.3%. Table 4 details the number of generated decays and how they do or do not fall inside the loose sPHENIX acceptance. If the tracking requirement is tightened to the values in the previous paragraph, then the efficiency increases by a factor of 3 to 20.4%. Table 5 also details the number of decays with an associated photon within the loose cut and how many had all their tracks reconstructed via the tracking algorithms. A comparison of the tracking efficiency as a 2D plot of minimum and maximum track p_{T} and the efficiency as a function of the true D^0 p_{T} is given in Figure 8.

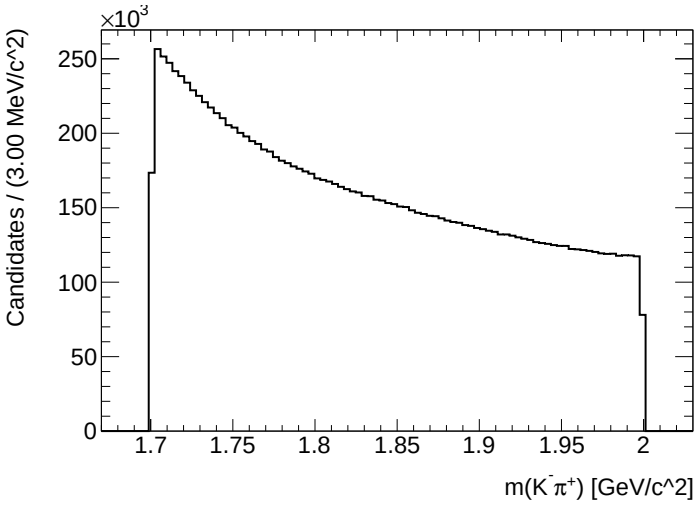

Figure 7: Distribution of the $K^-\pi^+$ invariant mass pairs using simulated Au+Au events with the baseline selection.

A large volume of the D^0 daughter tracks appear to sit at the edge or beyond the sPHENIX fiducial region where one track has a either large value of $|\eta|$ or a low p_{T} (or both). This can be seen in Figure 9 where there is a large concentration of events with a p_{T} less that 0.2 GeV and an $|\eta \geqslant 1.1$.

A tighter selection was applied to the data set to extract the D^0 peak. The cuts were chosen by studying the kinematic distributions of signal and background samples after the baseline selection was applied. To avoid biases, the p+p sample was used for the signal

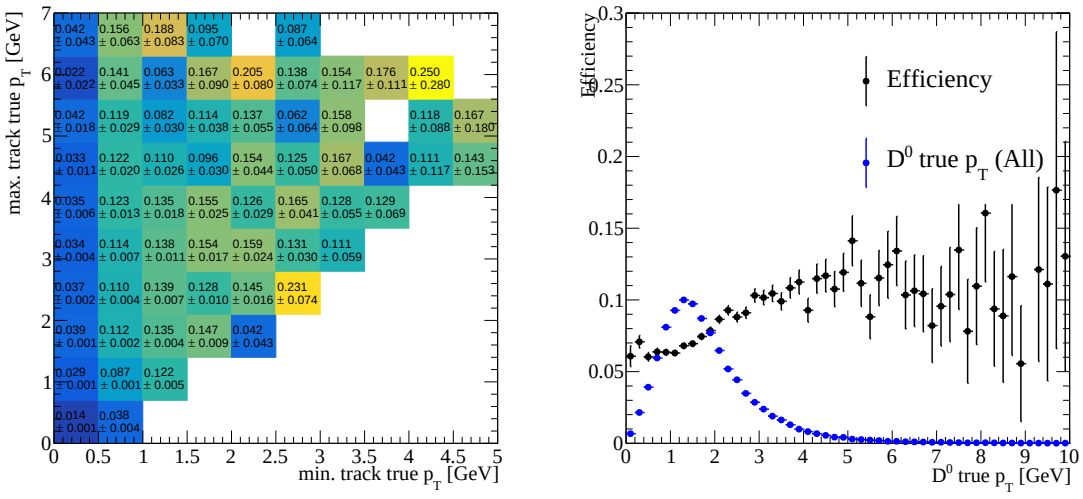

Figure 8: The tracking efficiency as measured in the Au+Au simulation. A 2D plot of the tracking efficiency as a function of the minimum track p_{T} (x-axis) and maximum track p_{T} (y-axis) is shown on the left while the tracking efficiency as a function of the true D^0 p_{T} is shown on the right..

Requirement	Count
Generated decays	636070
Generated decays that fail p_T	16766
Generated decays that fail η	300839
Generated decays that fail p_T and η	98869
Reconstructable decays	219596
Reconstructable decays with an associated γ	65048
Reconstructable decays with an associated π^0	0
Reconstructable decays with an associated γ and π^0	0
Reconstructable decays with no associated γ no π^0	154548
Efficiency [%]	34.5

Table 4: Comparison of the generated number of $D^0 \to K^-\pi^+$ decays in the Au+Au simulation to the number of decays that fall in the sPHENIX acceptance of $|\eta| \leqslant 1.6$ and $p_{\mathrm{T}} \geqslant 0.16\,\mathrm{MeV}$.

Requirement	Count
Decays in acceptance	219596
Decays fully reconstructed	17138
Efficiency [%]	7.8

Table 5: Comparison of the number of $D^0 \to K^-\pi^+$ decays fully in the sPHENIX acceptance in the Au+Au simulation to the number of decays that have all track reconstructed.

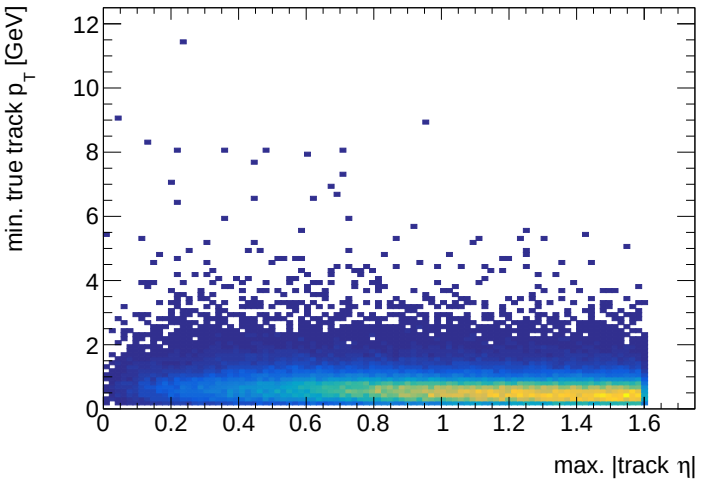

Figure 9: 2D distribution of the maximum absolute η value of a track and the minimum p_{T} of a track. Both of these values come from the truth information and do not necessarily come from the same track. Only one track needs to fall outside the fiducial range to be unable to reconstruct a decay.

and the upper and lower mass sidebands from the Au+Au sample that sit outside of the fit range were used for the background sample. These samples and the corresponding variable distributions can be found in Appendix B while the new cuts are listed in Table 6.

Two fits were performed using the models detailed in Section 5 at two different values of the D^0 p_{T}: 2 and 4 GeV. The latter cut is determined to be a very tight cut capable of rejecting a significant fraction of the background and could be achieved at an early stage of the commissioning period while the tracking calibrations are being studied. The former cut represents a physics goal of sPHENIX, to reconstruct D^0 to a p_{T} of 2 GeV or less. The results of the fit for D^0 $p_T \geqslant 4$ GeV are given in Table 7 and Figure 10 where the yield was measured to be 129 ± 20. This is a statistical significance of $\sigma = 6.45$. The results of the fit for D^0 $p_T \geqslant 2$ GeV are given in Table 8 and Figure 11 where the yield was measured to be 536 ± 79. This is a statistical significance of $\sigma = 6.78$.

Variable	Cut
min. track IP [cm]	0.008
max. track-track DCA [cm]	0.005
D^0 p_{T} [GeV]	2 or 4
D^0 DIRA	0.98
$m_{K^-\pi^+}$ [GeV]	1.70 $\rightarrow$ 2.00

Table 6: Enhanced cuts used to select $D^0 \rightarrow K^-\pi^+$ candidates.

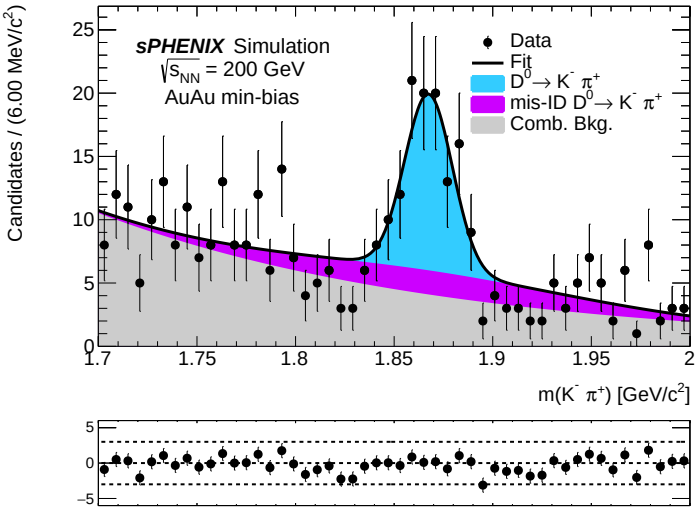

Figure 10: Fit to the $K^-\pi^+$ invariant mass distribution using simulated Au+Au events with the tighter selection and a D^0 $p_T \geqslant 4\,\mathrm{GeV}$.

Parameter	Value
$\mu[\,\mathrm{MeV}]$	1867.45 ± 2.1
$\sigma_{\mathrm{cor-ID}}[\,\mathrm{MeV}]$	12.1 ± 1.7
$\lambda\ [\,\mathrm{MeV}^{-1}]$	-5.6 ± 1.0
N_{cand}	383
f_{D^0} [%]	33.7 ± 5.3

Table 7: Fit results with the increased selection requirements and a D^0 $p_T \geqslant 4\,\mathrm{GeV}$.

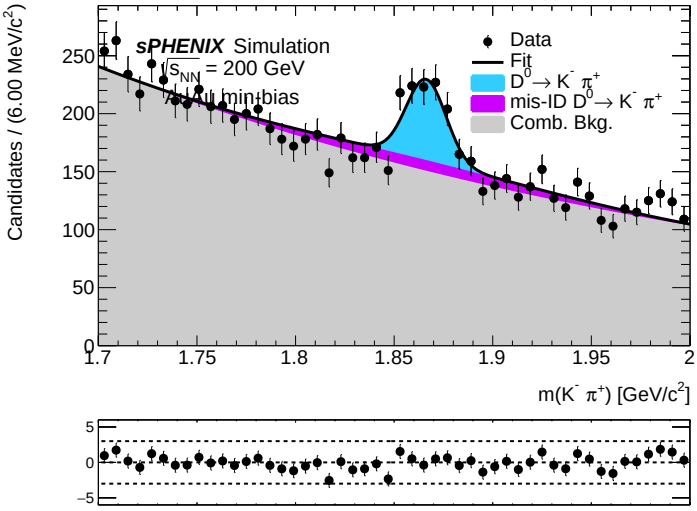

Figure 11: Fit to the $K^-\pi^+$ invariant mass distribution using simulated Au+Au events with the tighter selection and a D^0 $p_T \geqslant 2\,\mathrm{GeV}$.

Parameter	Value
μ[MeV]	1866.23 ± 1.6
$\sigma_{\mathrm{cor-ID}}$[MeV]	10.1 ± 1.5
λ [MeV^{-1}]	-2.80 $\pm$ 0.14
N_{cand}	8664
f_{D^0} [%]	6.2 ± 0.9

Table 8: Fit results with the increased selection requirements and a D^0 $p_T \geqslant 2\,\mathrm{GeV}$.

7 Conclusion

This note details a study performed using simulated Au+Au events with the sPHENIX detector, and tracking and reconstruction algorithms that will be deployed during the commissioning period which is due to commence in Spring 2023. Approximately 22 million minimum-bias events were simulated, corresponding to just over 20 minutes of data taking at the full RHIC collision rate of 15 kHz. This study is aimed at defining base cuts to be deployed during the commissioning period where the collision rate is lower and so this sample corresponds to a longer integrated time than 20 minutes.

By applying the cuts detailed in Tables 2 and 6 and the models described in Section 5, $D^0 \rightarrow K^-\pi^+$ decays were extracted with a significance larger than 5σ for a mother p_{T} $\geqslant 2\,\mathrm{GeV}$ It is currently recommended that these cuts be applied to the initial sPHENIX data sample to obtain a sufficient quantity and quality of D^0 candidates to validate the tracking and heavy flavor programs for full physics analyses.

There are several areas where this study could be improved and some of these are detailed in the appendices in this note such as the application of machine learning, improved background and signal modelling, and the use of statistical techniques to unfold data for reaching a lower p_{T} region that 2 GeV. There are other areas where this study could be improved that are not discussed in detail of this note. One are is the improvement of the acceptance and tracking efficeincy calculations. It was mentioned in Section 6 that the efficiency calculations do not account for the position of the secondary vertex and how this can alter the track η and p_{T} requirements. A module that can calculate the position dependent requirements would improve the understanding of these efficiencies. Further, the calculated values of the DCA variables appears to be smaller than would be expected for a particle that travels around 1 mm before decaying. A study will be performed that compares the DCA values calculated by KFParticle with a typical straight line extrapolation of a track to the primary vertex to see if there are any discrepancies.

Beyond the additional studies detailed in the appendices, this work could be expanded easily to study the $K^+K^-\pi^+$ spectrum to obtain $D^+_{(s)}$ candidates with little alterations to the selection beyond the invariant mass window. This study could also be used to serve as a baseline for a Λ_c^+ study (which has a shorter lifetime than the D^0) and a K_s^0 study (which has a longer lifetime than the D^0 but a lower p_{T} spectrum).

References

[1] sPHENIX Collaboration. sPHENIX Beam Use Proposal, May 2022.

[2] G. Aad et al. Measurement of the nuclear modification factor for muons from charm and bottom hadrons in Pb+Pb collisions at 5.02 TeV with the ATLAS detector. *Physics Letters B*, 829:137077, 2022.

[3] ALICE Collaboration. Measurement of beauty production via non-prompt D^0 mesons in Pb-Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV, 2022.

[4] X. Chen. Prompt and non-prompt D0-meson production in Au+Au collisions, Ph.D. Thesis, STAR, 2019.

[5] P. Jackson et al. Measurement of the total cross section from elastic scattering in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. *Physics Letters. Section B: Nuclear, Elementary Particle and High-Energy Physics*, 761:158–178, 2016.

[6] R. Aaij et al. Prompt charm production in pp collisions at $\sqrt{s} = 7$TeV. *Nuclear Physics B*, 871(1):1–20, 2013.

[7] S. Acharya et al. Charm-quark fragmentation fractions and production cross section at midrapidity in pp collisions at the LHC. *Physical Review D*, 105(1), jan 2022.

[8] Particle Data Group. Review of Particle Physics. *Progress of Theoretical and Experimental Physics*, 2020(8), 08 2020. 083C01.

[9] T. Sjöstrand, S. Mrenna, and P. Skands. A Brief Introduction to PYTHIA 8.1. *Computer Physics Communications*, 178(11):852–867, 2008.

[10] S. Lim, W Park. PYTHIA8 tune in pp 200 GeV. Presentation at Heavy Flavor Topical Group Meeting, Dec. 14, 2020, https://indico.bnl.gov/event/10309/contributions/44139/attachments/31909/50542/sPHENIX_HF_shlim_20201215.pdf.

[11] M. Gyulassy and X.-N. Wang. HIJING 1.0: A Monte Carlo program for parton and particle production in high energy hadronic and nuclear collisions. *Computer Physics Communications*, 83(2):307–331, 1994.

[12] S. Agostinelli et al. Geant4—a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250–303, 2003.

[13] X. Ai et al. A Common Tracking Software Project, 2021.

[14] J. D. Osborn, A. D. Frawley, J. Huang, S. Lee, H. Pereira Da Costa, M. Peters, C. Pinkenburg, C. Roland, and H. Yu. Implementation of ACTS into sPHENIX Track Reconstruction. *Computing and Software for Big Science*, 5(1), Oct 2021.

[15] A. Gorbunov and I. Kisel. Reconstruction of decayed particles based on the Kalman filter. *CBM-SOFT-note-2007-003*, May 2007.

[16] C. Dean. Event Reconstruction and MDC1. Presentation at sPHENIX 10^{th} Collaboration Meeting, Jan. 22, 2021, https://indico.bnl.gov/event/10568/contributions/45092/attachments/32421/51581/HeavyFlavor_EventRecoAndMDC_C_Dean_20210122.pdf.

[17] S. T. Araya, C. Dean, J. Huang, H. Okawa, , and Z. Shi. First MDC1 Results from Heavy Flavor Topical Group, April 2021.

[18] A. Buckley, P. Ilten, D. Konstantinov, L. Lönnblad, J. Monk, W. Pokorski, T. Przedzinski, and A. Verbytskyi. The HepMC3 event record library for Monte Carlo event generators. *Computer Physics Communications*, 260:107310, March 2021.

[19] A. Hoecker et al. TMVA - Toolkit for Multivariate Data Analysis, 2007.

[20] W Verkerke and D. P. Kirkby. The RooFit toolkit for data modeling. *eConf*, C0303241:MOLT007, 2003. arXiv:physics/0306116.

[21] K. Cranmer. Kernel Estimation in High-Energy Physics. *Computer Physics Communications*, 136(3):198–207, 2001.

[22] J. Gaiser. Charmonium Spectroscopy From Radiative Decays of the J/ψ and ψ'. *Ph.D. Thesis, SLAC*, 1982.

[23] B. Knuteson and H. Miettinen. Mass Analysis and Parameter Estimation with PDE. *D0 notes*, 9 1997.

[24] I. S. Abramson. On Bandwidth Variation in Kernel Estimates - A Square Root Law. *The Annals of Statistics*, pages 1217–1223, 1982.

[25] M. Pivk and F.R. Le Diberder. sPlot: A Statistical Tool to Unfold Data Distributions. *Nuclear Instruments and Methods in Physics Research A*, 555(1):356 – 369, 2005.

Appendices

A Alternative Fit Models

In Figure 6 it can bee seen that the default model does not perfectly describe the data as it overshoots the mean while the base seems to be shifted in the opposite direction (it over fits before the mean and under fits after the mean). This implies that simply adjusting the fit parameters will not help. Alternative fit models are proposed that can better describe the data. The fits are described only to emphasise the default fit shortcomings. The double Gaussian model is used in the final fit as the D^0 shape is known to be well described by a single Gaussian. The reason for the asymmetric peak shape is assumed to be due to an earlier version of the tracking that was run on the p+p simulation compared to the Au+Au simulation. This earlier version of the tracking had a simpler version of the TPC clustering which was improved for the Au+Au simulation and was known to affect the momentum resolution of the tracks.

A.1 Bifurcated Gaussian

The first model used is a bifurcated Gaussian which has different widths on either side of the mean value. This shape is described by the PDF

$$f(x;\mu,\sigma_L,\sigma_R) = N \cdot \begin{cases} \exp\left(-\frac{(x\ -\ \mu)^2}{2\sigma_L^2}\right), & \text{for } x < \mu \\ \exp(-\frac{(x\ -\ \mu)^2}{2\sigma_R^2}), & \text{for } x \geqslant \mu \end{cases} \tag{6}$$

where μ describes the mean of the Gaussian and $\sigma_{[L/R]}$ describe the Gaussian widths on the left and right hand side of the mean value.

The model is fit to a subset of the p+p signal sample. This subset has no cuts applied and uses 38420 signal candidates in the fit. Using a smaller number of events avoids the issue of over-fitting the distribution but having no selection means the shape will differ from the final distribution. The fit values and plot are given in Table 9 and Figure 12 respectively. It can be seen that the alternative fit model better describes the data, however there is still a prominent bias.

Parameter	Value
μ[MeV]	1858.93 ± 0.28
$\sigma_{\text{cor-ID,L}}$[MeV]	12.02 ± 0.22
$\sigma_{\text{cor-ID,R}}$[MeV]	14.98 ± 0.25
$k_{\text{cor-ID}}$ [%]	57.73 ± 0.44
$\sigma_{\text{mis-ID}}$[MeV]	89.21 ± 1.31

Table 9: Bifurcated Gaussian fit parameters to the $D^0 \to K^-\pi^+$ p+p signal sample.

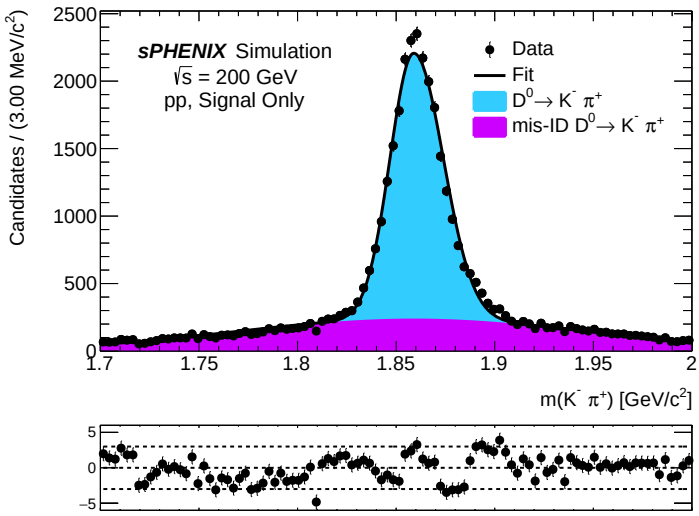

Figure 12: Bifurcated Gaussian fit to the $K^{-}\pi^{+}$ invariant mass distribution using simulated p+p events with truth matching.

A.2 Double Crystal Ball

A second alternative model is proposed to describe the data. This model is a double-sided Crystal Ball function [22] and consists of a Gaussian core with a polynomial tail. This tail is capable of describing radiative losses. The double-sided Crystal Ball has this tail on either side of the Gaussian core unlike the single-sided which only has the polynomial on the low-mass mass side. The shape is described by the PDF

$$f(x;\alpha_{\mathrm{L}},n_{\mathrm{L}},\alpha_{\mathrm{H}},n_{\mathrm{H}},\mu,\sigma) = N\cdot\begin{cases}\exp(-\frac{(x-\mu)^2}{2\sigma^2}), & \text{for } -\alpha_{\mathrm{L}} < \frac{x-\mu}{\sigma} < -\alpha_{\mathrm{H}}\\ A_{\mathrm{L}}\cdot(B_{\mathrm{L}} - \frac{x-\mu}{\sigma})^{-n_{\mathrm{L}}}, & \text{for } -\alpha_{\mathrm{L}} \geqslant \frac{x-\mu}{\sigma}\\ A_{\mathrm{H}}\cdot(B_{\mathrm{H}} - \frac{x-\mu}{\sigma})^{-n_{\mathrm{H}}}, & \text{for } \frac{x-\mu}{\sigma} \geqslant -\alpha_{\mathrm{H}}\end{cases} \tag{7}$$

where

$$A_{[\mathrm{L/H}]} = \left(\frac{n_{[\mathrm{L/H}]}}{\left|\alpha_{[\mathrm{L/H}]}\right|}\right)^{n_{[\mathrm{L/H}]}}\cdot\exp\left(-\frac{\left|\alpha_{[\mathrm{L/H}]}\right|^2}{2}\right) \tag{8}$$

$$B_{[\mathrm{L/H}]} = \frac{n_{[\mathrm{L/H}]}}{\left|\alpha_{[\mathrm{L/H}]}\right|} - \left|\alpha_{[\mathrm{L/H}]}\right|. \tag{9}$$

The mean and width of the central Gaussian distribution are given by μ and σ respectively while L and H refer to the parameters defining the low and high mass regions of the distribution. The parameter $\alpha_{[\mathrm{L/H}]}$ describes the boundary between the Gaussian and power law components while $n_{[\mathrm{L/H}]}$ describes the order of the power law (it is not necessarily an integer value).

The model is fit to the same subset of the p+p signal sample. The fit values and plot are given in Table 10 and Figure 13 respectively. It can be seen that the double Crystal Ball describes the data well.

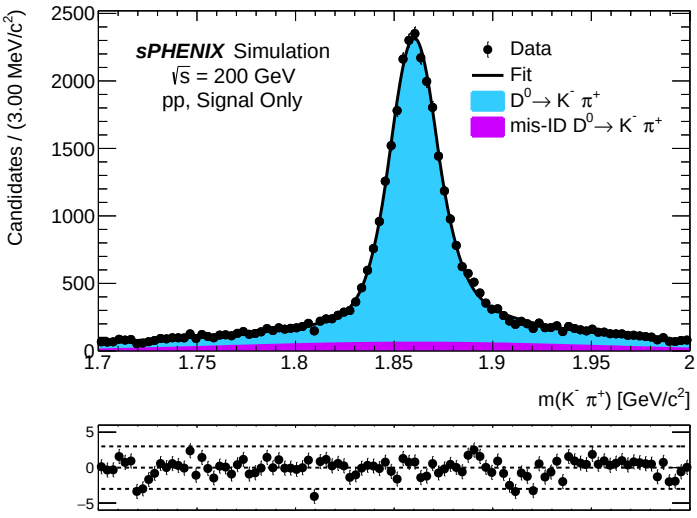

Figure 13: Double Crystal Ball fit to the $K^-\pi^+$ invariant mass distribution using simulated p+p events with truth matching.

Parameter	Value
μ[MeV]	1860.15 ± 0.12
$\sigma_{\mathrm{cor-ID}}$[MeV]	12.67 ± 0.15
α_{L}	1.07 ± 0.03
n_{L}	0.89 ± 0.20
α_{H}	0.93 ± 0.03
n_{H}	1.31 ± 0.30
$k_{\mathrm{DCB,\ Gauss.\ core}}$ [%]	41.53 ± 2.86
$k_{\mathrm{cor-ID}}$ [%]	87.29 ± 6.26
$\sigma_{\mathrm{mis-ID}}$[MeV]	97.7 ± 19.4

Table 10: Double Crystal Ball fit parameters to the $D^0 \to K^-\pi^+$ p+p signal sample.

A.3 Kernel density estimated backgrounds

No partially reconstructed backgrounds are modelled in the default fit. However, if one wishes to model these non-parametric distributions an option would be to use a kernel density estimation (KDE). As the invariant mass distribution of the events that pass the selection requirements tend to become non-parametric due to the selection requirements and the invariant mass substitutions imposed on the sample, then the shape of the PDFs are extracted using the kernel density estimation method with an adaptive bandwidth [21]. The use of kernel density estimation to describe non-parametric distributions in invariant masses was proposed by the D0 collaboration for analysis of the Higgs boson [23] but has been developed for use in other collaborations.

In a kernel estimation the events of a distribution are substituted for a kernel function so that the distribution, $\hat{f}_0(x)$, can be described as

$$\hat{f}_0(x) = \frac{1}{nh} \sum_{\mathrm{i}=1}^{n} K\left(\frac{x - t_i}{h}\right) \tag{10}$$

where t_i is the value of event i and h is the bandwidth or smoothing parameter. It has been suggested [23] that a suitable kernel function for use in describing the invariant mass distribution would be a Gaussian as it is positive definite and infinitely differentiable.

To avoid issues of function overspill at boundaries, underestimations of the distribution within regions with low event density, and overestimation within regions of high density then the bandwidth is allowed to alter on an event-by-event basis which is known as an *adaptive kernel estimation* where the per-event bandwidth is given by [24]

$$h_i = \frac{h}{\sqrt{\hat{f}_0(x)}}. \tag{11}$$

The PDFs determined by the kernel method were obtained by using DecayFinder and HFTrackEfficiency to select $D_s^+ \to K^-K^+\pi^+$ from the p+p to $c\bar{c}$ simulation. The new track map of these decays was then passed to KFParticle where the default selection given in Table 6 was applied using the $D^0 \to K^-\pi^+$ decay descriptor. The resulting shape is given in Figure 14. The expected contamination of this decay with respect to $D^0 \to K^-\pi^+$ would be estimated from the ratio of the hadronisation fractions, branching fractions and the number of each decay that pass the selection divided by the total number of each decay generated,

$$k_{\mathrm{chann}}^{\mathrm{mid}} = \frac{f_{\mathrm{mid}}}{f_{\mathrm{chann}}} \frac{\mathcal{B}_{\mathrm{mid}}}{\mathcal{B}_{\mathrm{chann}}} \frac{\omega_{\mathrm{mid}}}{\varepsilon_{\mathrm{chann}}} \tag{12}$$

where f_{P} is the hadronization fraction of the mother particle, $\mathcal{B}_{\mathrm{P}}$ is the branching fraction of the decay, ω_{mid} is the misidentification efficiency and $\varepsilon_{\mathrm{chann}}$ is the signal efficiency. As both the signal and background samples are drawn from the same minimum bias p+p simulation where all $c\bar{c}$ events are saved, then the contamination fraction becomes

$$k_{\mathrm{chann}}^{\mathrm{mid}} = \frac{N_{\mathrm{sel}}(D_s^+ \to K^-K^+\pi^+)}{N_{\mathrm{sel}}(D^0 \to K^-\pi^+)} \tag{13}$$

where $N_{\rm sel}(X)$ is the number of decays of X that pass the selection.

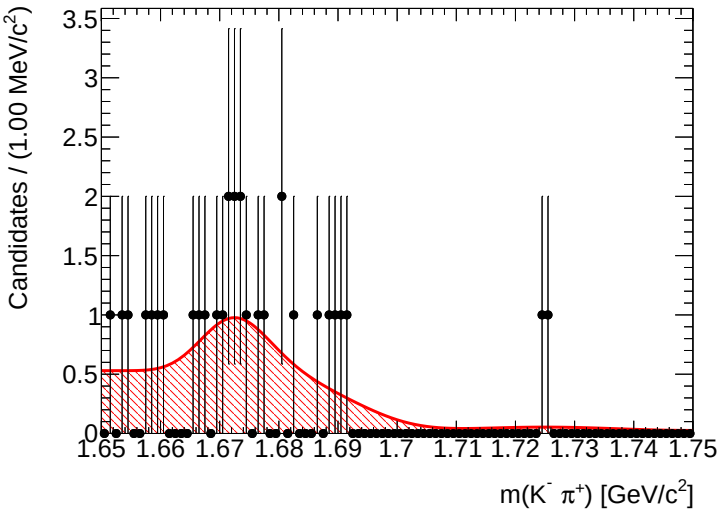

Figure 14: Kernel estimated distribution of $D_s^+ \to K^-K^+\pi^+$ in the $m_{K^-\pi^+}$ spectrum using the tighter $D^0 \to K^-\pi^+$ selection.

B Machine Learning Selection

When selecting the baseline cuts, it was known that they would be so loose that the background would overwhelm the signal. The reason behind this cut was to allow for a machine-learning study to see if the low-p_{T} reach of D^0 could be improved. To this end, signal and background samples were passed to ROOT's TMVA machine learning package [19] to train various algorithms and see if this goal could be achieved. The signal sample was comprised of the p+p sample with the baseline cuts applied while the background sample was taken from the upper and lower mass side bands of the Au+Au simulation with the baseline selection applied. The background sample was taken from the ranges $1.65 \leqslant m_{K^-\pi^+} < 1.70$ and $2.00 < m_{K^-\pi^+} \leqslant 2.10$. The samples were split into even and odd samples by their event numbers. The odd samples were used for training while the even samples were used for testing. All of this was to avoid biases in the final selection by not re-using events. The number of events used in each sample is given in Table 11 and the invariant mass distributions are given in Figure 15.

Sample	Number of candidates
Signal, training	20082
Signal, testing	19726
Background, training	199228
Background, testing	200662

Table 11: Number of candidates used for training and testing the machine learning algorithms.

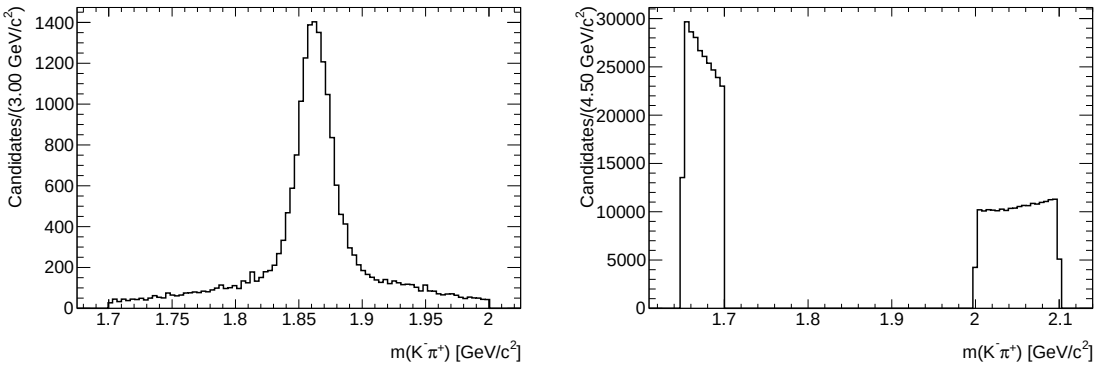

Figure 15: Invariant mass distributions of the signal and background samples used to train the machine learning algorithms. The preselection given in Table 2 has been applied to both samples. Each sample consists of events with an odd event number. The signal sample is shown on the left and the background sample is shown on the right.

Thirteen variables were chosen to train the algorithms: the minimum and maximum DCA of the selected tracks with respect to their primary vertex in both two and three dimensions, the minimum and maximum track p_{T}, the minimum and maximum track χ^2 per number of

degrees of freedom, the DCA of the tracks with respect to each other, the D^0 p_{T}, the D^0 DIRA, the D^0 DCA χ^2 with respect to the primary vertex and the D^0 χ^2 per number of degrees of freedom. The comparison of the signal and background distributions for these variables can be seen in Figure 16 and the correlations between variables an be seen in Figure 17. These variables were used to train six algorithms: two multi-layer perceptrons (MLP), ROOTs own neural net, a boosted decision tree (BDT), a BDT with gradient boost and a BDT with de-correlation and adaptive boost. The receiver operating characteristic (ROC) curve and response of each algorithm with training and testing samples overlaid can be seen in Figures 18 and 19 respectively.

The ROC curve legend lists the algorithms in order of best performance. Thus, the old MLP is predicted to give the best performance. The algorithm was applied to the Au+Au data within the range $1.70 \leqslant m_{K^-\pi^+} \leqslant 2.00$ and the response of each candidate to each ML algorithm was calculated. The fit to the invariant mass was then performed using the baseline cut and two values of the MLP. For an MLP response $\geqslant 0.99$, the yield was measured to be 536 ± 74 while for an MLP response $\geqslant 0.999$, the yield was measured to be 339 ± 48. The fit to both distributions is given in Figure 20 and the fit results with an algorithm response $\geqslant 0.999$ are given in Table 12.

Parameter	Value
$\mu[\,\mathrm{MeV}]$	1865.55 ± 1.5
$\sigma_{\mathrm{cor-ID}}[\,\mathrm{MeV}]$	10.4 ± 1.4
$\lambda\ [\,\mathrm{MeV}^{-1}]$	-6.60 ± 0.29
N_{cand}	3139
f_{D^0} [%]	10.8 ± 1.5

Table 12: Fit results for the ML study with a multi-layer perceptron with an algorithm response $\geqslant 0.999$.

While this study was brief, it demonstrates that machine learning algorithms can be used to extract the $D^0 \to K^-\pi^+$ decay at sPHENIX. This could be applied to the entire sPHENIX data set but it is not felt to be suited to the commissioning period where we wish to pre-select a small subset of tracks that we can quickly reprocess at a later date. The baseline selection resulted in 15 million D^0 candidates from a sample of 21 million minimum bias events and thus would result in an unnecessarily large data sample to reprocess.

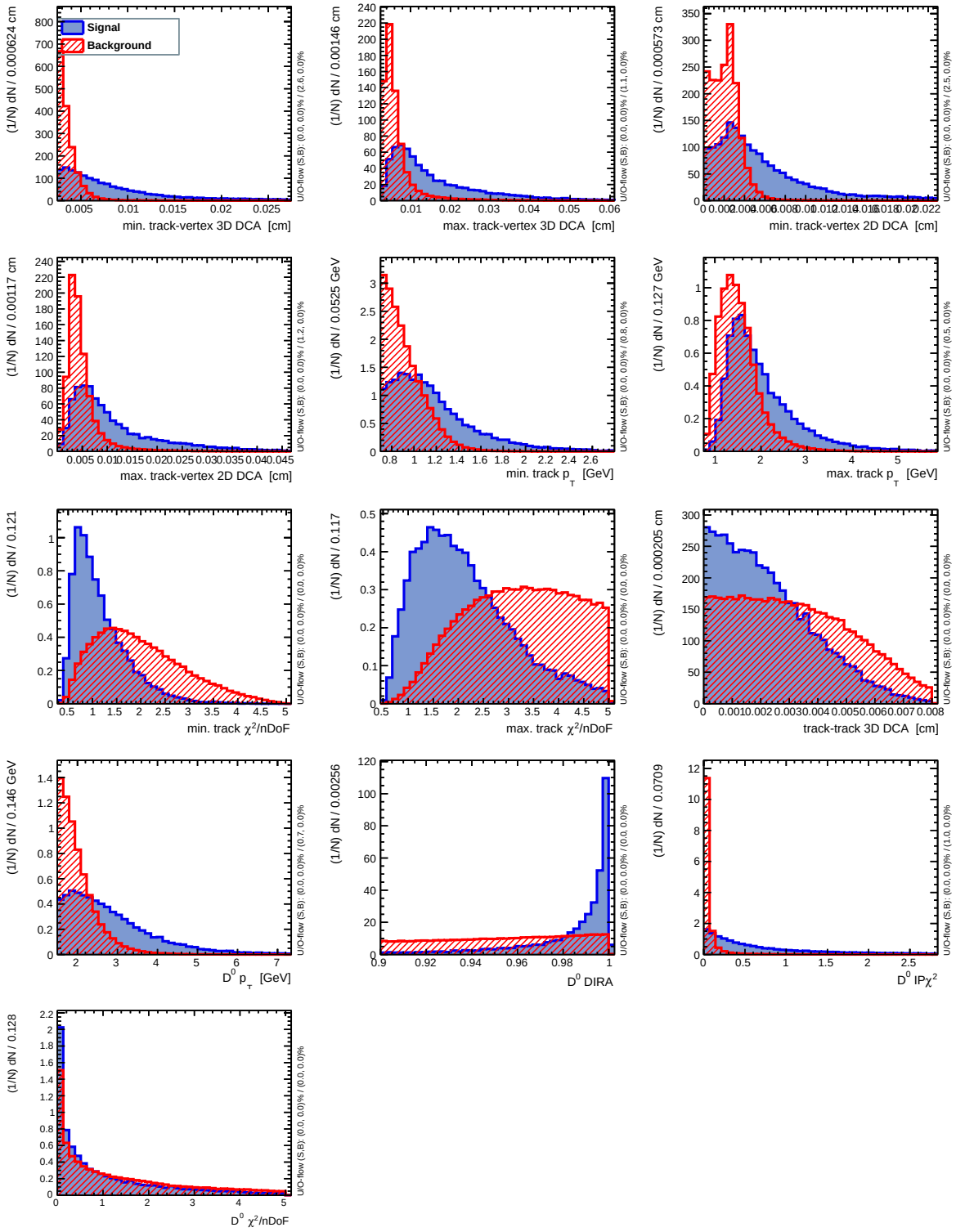

Figure 16: Input variable distributions of the signal (blue) and background (red) samples used to train the machine learning algorithms. Variables from top to bottom, left to right: minimum track IP, min. track p_{T}, max. track χ^2 per no. of degrees of freedom, track-track DCA, D^0 p_{T}, D^0 DIRA, D^0 IPχ^2, D^0 χ^2 per no. of degrees of freedom.

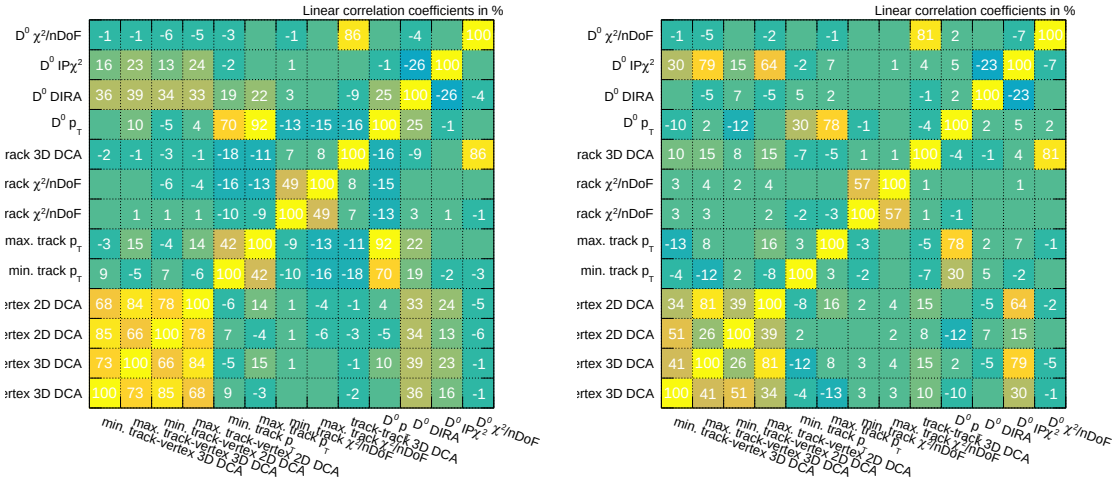

Figure 17: Input variable correlation coefficients of the signal (left) and background right) samples used to train the machine learning algorithms.

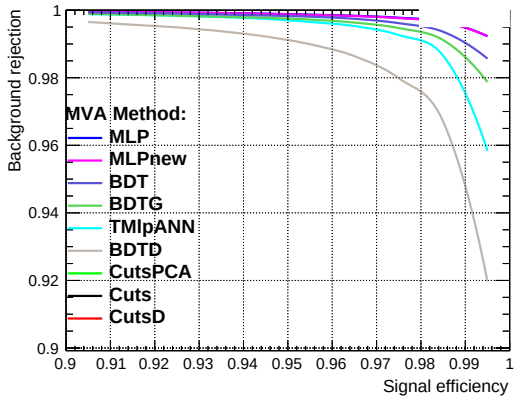

Figure 18: Receiver operating characteristic plot from the machine learning study.

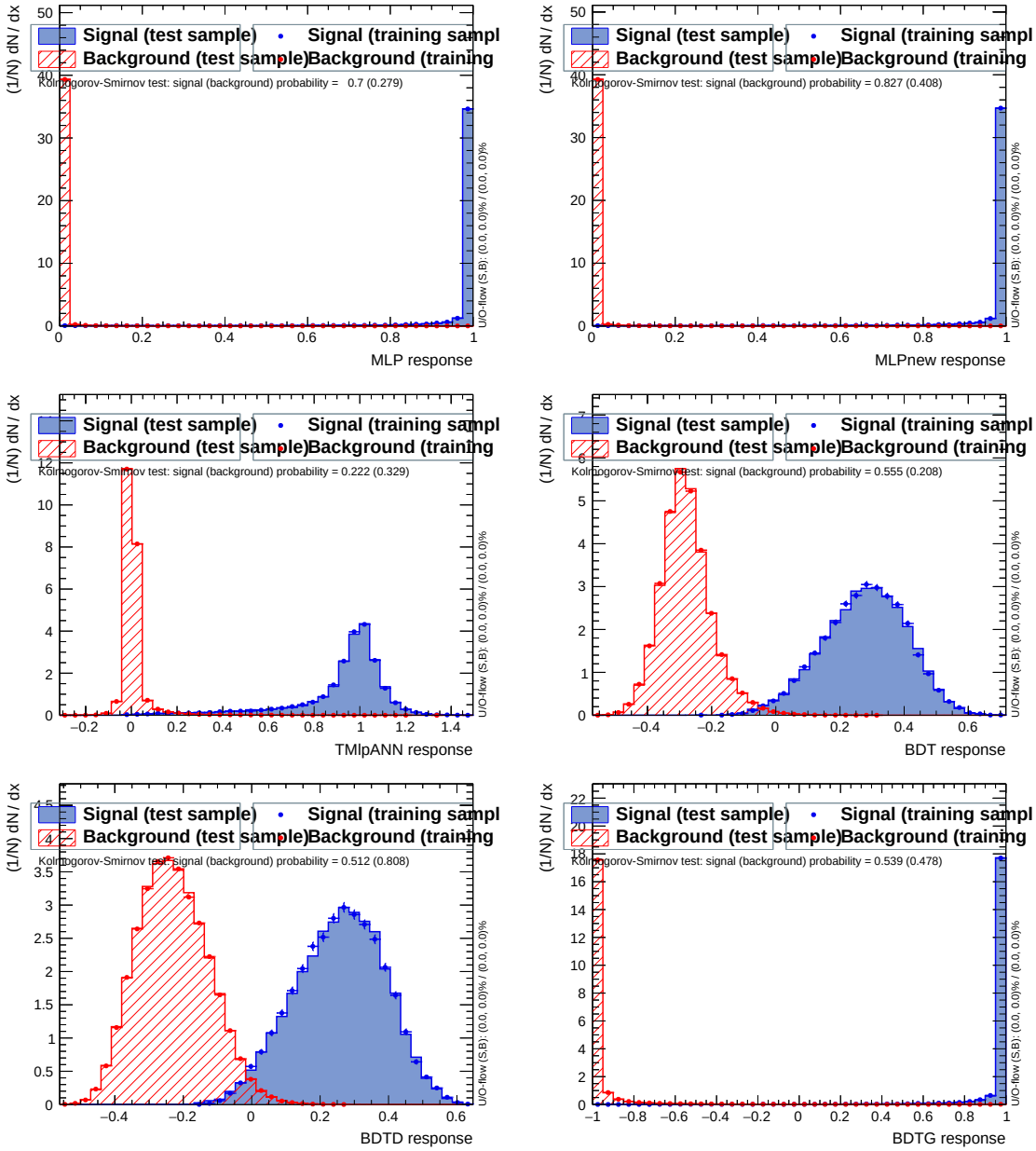

Figure 19: Classifier output distributions for the machine learning algorithms that were trained. From top to bottom, left to right: the multi-layer perception, a second multi-layer perceptron, ROOT's own neural net, a boosted decision tree, a boosted decision tree with decorrelation and adaptive boost, and a boosted decision tree with gradient boosting.

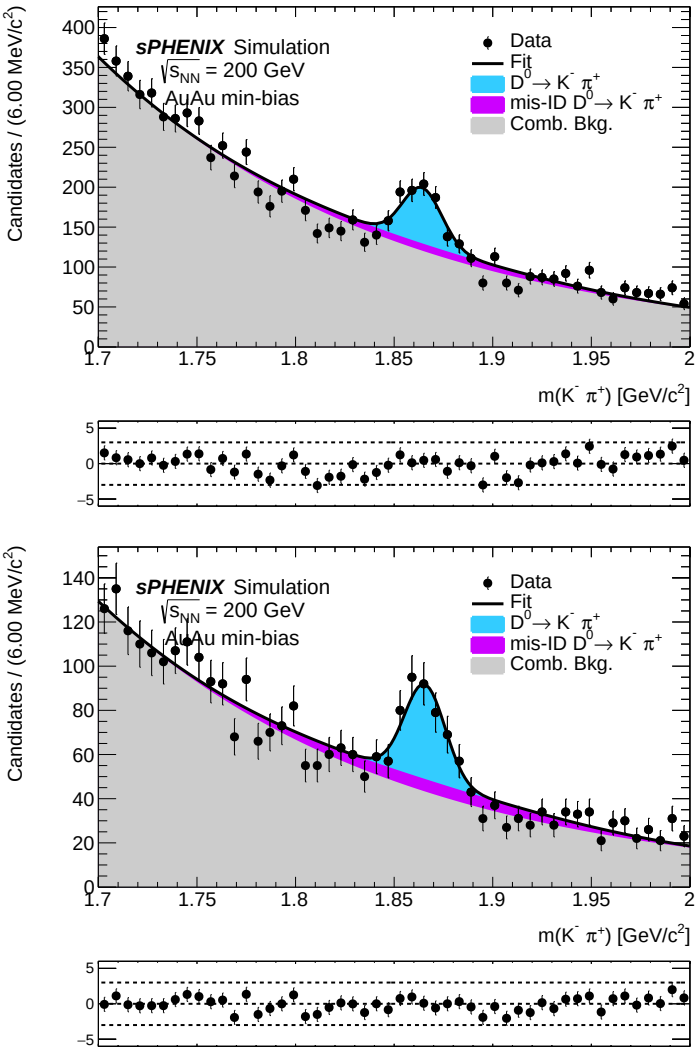

Figure 20: Fit to the $K^{-}\pi^{+}$ invariant mass distribution using simulated Au+Au events with the baseline selection. Machine learning algorithms were trained to reject more background and this fit was performed using a multi-layer perceptron with an algorithm response $\geqslant 0.99$ (top) and $\geqslant 0.999$ (bottom).

C sWeighting

It is often the case that data distributions are composed of a composite of sources such a D^0 signal and combinatorial background. Often we want to look at the distribution from a single source and removing the other sources can be difficult. One method for doing this is to cut on variables that have a large separation between sources but this can lead to contamination and biases in the final distribution due to overlaps. Another method, known as *sWeighting* [25], involves maximising a likelihood function for a discriminating variable (such as a candidates's mass) where each candidate's contribution to a specific class is determined by calculating the likelihood with and without that event then assigning an event probability where the probabilities of an event belonging to that class are required to sum to one over all classes in the model. These weights can then be applied to a control variable (such as a particle's p_{T}) assuming there is no correlation between the two variables.

To investigate whether *sWeights* can be used to improve the low p_{T} reach of the sample, the fit was redone using the baseline selection with an ML response $\geqslant$ 0.999 then the weights were calculated for each candidate to be a signal or background event. For simplicity, the second Gaussian that models the mis-ID was removed. The weight for each candidate to come from a $D^0 \to K^-\pi^+$ decay was then added to the histogram of the mother's p_{T} distribution and compared to the unweighted sample. Of 3149 candidates in the total fit, the yield of correctly reconstructed D^0 was measured to be 209 $\pm$ 26. The fit to the invariant mass distribution is given in Figure 21 and the comparison of the weighted and unweighted p_{T} distributions is given in Figure 22. From the weighted sample, it appears that we can reach a p_{T} as low as 1.5 GeV which is the baseline cut.

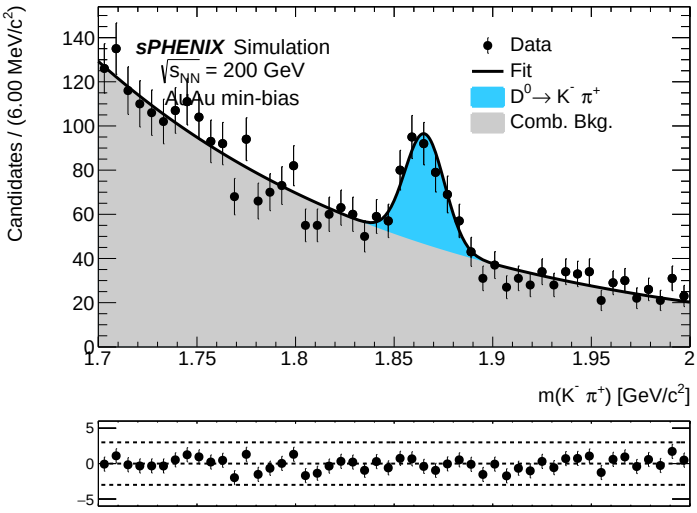

Figure 21: Fit to the $K^-\pi^+$ invariant mass distribution using simulated Au+Au events with the baseline selection and an MLP response $\geqslant$ 0.999. The mis-ID'd D^0 model was removed in this fit and the fitter was required to *sWeight* each candidate.

RooFit has an internal class that is capable of calculating the *sWeights* so it is a reasonably simple addition to add these weights to an nTuple. It should be noted that there are some requirements:

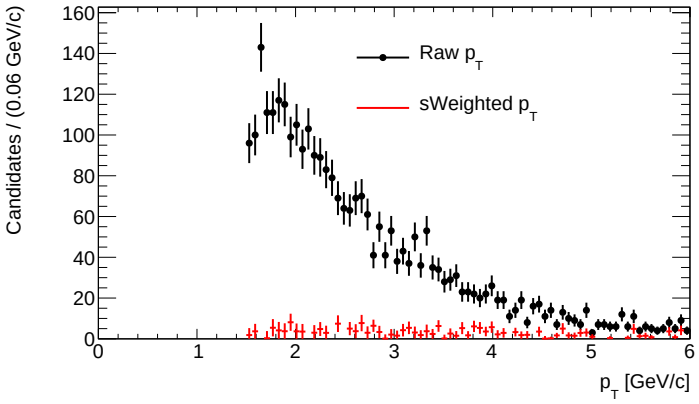

Figure 22: p_{T} distribution of the D^0 candidates with *sWeighting* in red and without *sWeighting* in black.

1. You must have a fit value for each component of your fit. This means you can't have a fit fraction for the signal and then subtract that from 1 to get the background fraction.

2. You must fit for the yield of each component of the fit, not the fraction. This is because the log-likelihood minimisation requires a value that is not less than 1.

An example code follows,

```
stringstream cutStream;
cutStream << "1.7 <= D0_mass && D0_mass <= 2.0";
TCut masscut = cutStream.str().c_str();
TFile* dataFile = new TFile("myInputFile.root");
TTree* dataTree = (TTree*)dataFile->Get("DecayTree");

string datasWeight =  inputFile.substr(0, inputFile.size()-5) + "_sWeighted.root";
TFile* sWeightedDataFile  = new TFile(datasWeight.c_str(), "RECREATE");
TTree* dataSWTree = dataTree->CopyTree(masscut);
TTree* sWeightedDataTree = dataSWTree->CloneTree(-1);

RooRealVar mass(branch.c_str(), "mass", minMass, maxMass);
RooDataSet dataSet(branch.c_str(), "data", mass, Import(*sWeightedDataTree));

/*
 * Signal Model
 */
RooRealVar  mean("mean", "mean", 1.865, 1.835, 1.875);
RooRealVar  sigma("sigma", "sigma", 0.006, 1e-3, 0.030);
RooGaussian D0("D0", "D0", mass, mean, sigma);

RooRealVar fSig("fSig", "fSig", 0.1*dataSet.numEntries(), 0, 2*dataSet.numEntries());
RooRealVar fBkg("fBkg", "fBkg", 0.9*dataSet.numEntries(), 0, 2*dataSet.numEntries());
```

```
/*
 * Background Model
 */
RooRealVar expConst("expConst", "expConst", -10, -1e2, 0.);
RooExponential background("background", "background", mass, expConst);

/*
 * Fitting to the data
 */
RooArgList fitModelList(D0, background), fitFracList(fSig, fBkg);

RooAddPdf model("model", "model", fitModelList, fitFracList);
model.fitTo(dataSet);

RooStats::SPlot* sData = new RooStats::SPlot("sData", "An sPlot",  dataSet,
                                             &model, RooArgList(fSig, fBkg));
double sig_sw;
TBranch* b_sig_sw = sWeightedDataTree->Branch("sWeight", &sig_sw, "sWeight/D");

std::cout << "Check sWeights:" << std::endl;
std::cout << "Yield of signal is "  << fSig.getVal()
          << ".  From sWeights it is " << sData->GetYieldFromSWeight("fSig")
          << std::endl;

for (int i = 0; i < dataSet.numEntries(); ++i)
{
  if (i < 5)
  {
    std::cout << "Signal Weight = " << sData->GetSWeight(i,"fSig")
              << ", Background Weight = " << sData->GetSWeight(i,"fBkg")
              << ", Total Weight = " << sData->GetSumOfEventSWeight(i)
              << std::endl;
  }

  const RooArgSet* row = dataSet.get(i);
  sig_sw = (double) row->getRealValue("fSig_sw");
  b_sig_sw->Fill();
}

sWeightedDataFile->Write();
sWeightedDataFile->Close();
```