Analysis of heavy-flavour particles in ALICE with the O^2 analysis framework

Vít Kučera[1,*], *Gian Michele* Innocenti[1], *Francesco* Prino[2], *Andrea* Rossi[6], *Jan Fiete* Grosse-Oetringhaus[1], *Nima* Zardoshti[1], *Luigi* Dello Stritto[3], *Chiara* Zampolli[1], *Fabio* Catalano[2,4], *Biao* Zhang[5], *Nicolo* Jacazio[1], *Mattia* Faggin[6], *Antonio* Palasciano[7], *Fabio* Colamaria[7], *Marianna* Mazzilli[1], *Deepa* Thomas[8], *Rik* Spijkers[9], *Jianhui* Zhu[10], and *Henrique* Zanoli[9]

[1]CERN
[2]INFN, Torino
[3]INFN, Salerno
[4]Politecnico Torino
[5]Central China Normal University
[6]Padua University
[7]INFN, Bari
[8]The University of Texas at Austin
[9]Utrecht University
[10]Darmstadt GSI

Abstract. Precise measurements of heavy-flavour hadrons down to very low p_T represent the core of the physics program of the upgraded ALICE experiment in Run 3 [1]. These physics probes are characterised by a very small signal-to-background ratio requiring very large statistics of minimum-bias events. In Run 3, ALICE is expected to collect up to $13\,\text{nb}^{-1}$ of lead–lead collisions, corresponding to about 1×10^{11} minimum-bias events. In order to analyse this unprecedented amount of data, which is about 100 times larger than the statistics collected in Run 1 and Run 2, the ALICE collaboration is developing a complex analysis framework that aims at maximising the processing speed and data volume reduction [2]. In this paper, the strategy of reconstruction, selection, skimming, and analysis of heavy-flavour events for Run 3 will be presented. Some preliminary results on the reconstruction of charm mesons and baryons will be shown and the prospects for future developments and optimisation discussed.

1 Introduction

The ALICE experiment is undergoing a major upgrade in view of the upcoming lead–lead (Pb–Pb), proton–lead (p–Pb) and proton–proton (pp) data taking periods in Run 3. Precise measurements of heavy-flavour (HF) hadrons down to very low p_T [3] represent the core of the physics program of ALICE in Run 3 [1]. These physics probes are characterised by a very small signal-to-background ratio requiring very large statistics. This large background makes triggering techniques very inefficient in Pb–Pb collisions, if not impossible. In Run 3, the

---

[*]e-mail: vit.kucera@cern.ch

ALICE detector plans to collect the entire lead–lead (Pb–Pb) statistics delivered by the LHC up to instantaneous luminosities of $\mathcal{L} = 6 \times 10^{27}$ cm^{-2}s^{-1}, corresponding to an interaction rate of about 50 kHz. The resulting data throughput from the detector has been estimated to be about 3.5 TB/s for Pb–Pb events, roughly two orders of magnitude larger than in Run 1 and Run 2. To minimise the cost and requirements of the computing system for data processing and storage, the ALICE Computing Model for Runs 3 and 4 is designed for a large reduction of the data volume read out from the detector as early as possible during the data-flow. The data volume reduction will be achieved by reconstructing the data in several steps synchronously with data taking. As an example, cluster finding in the Time Projection Chamber (TPC) and a first fast track reconstruction will be performed online. In the so-called asynchronous reconstruction, the final track and vertex reconstruction is performed offline using the final detector calibration in order to reach the required data quality. Only the output of the asynchronous reconstruction will be written on disk in the form of Analysis Object Data (AOD) files which serve as input for further processing with the Online–Offline (O^2) analysis framework [4].

The reconstruction of HF candidates is performed offline at the analysis stage using input AOD files, which contain the kinematic representation (*helix parameters*) of all the reconstructed tracks and the information coming from Particle Identification (PID) detectors. The production of charm and beauty hadrons will be measured in ALICE 2 by reconstructing their hadronic decay channels like $D^0 \rightarrow K^-\pi^+$, $D_s^+ \rightarrow \varphi\pi^+ \rightarrow K^+K^-\pi^+$, and $\Lambda_c^+ \rightarrow pK^-\pi^+$. This procedure relies on CPU-consuming operations like double and triple iterations over tracks in each event and reconstruction of the decay vertices. The biggest challenge associated with the reconstruction of these experimental probes in heavy-ion collisions down to very low p_T is therefore represented by the huge combinatorial background. In a high-multiplicity Pb–Pb collision (*central collisions*), indeed, the charged-particle multiplicity can reach values of about 1500–2000 per unit of rapidity at mid-rapidity, which corresponds to the multiplicity of more than 300 pp events at the same energy. To access low-p_T HF hadrons, candidates are built from tracks with p_T down to about 100 MeV/c. As a consequence, it is likely to reconstruct several thousands of two-prong and three-prong HF candidates in a central Pb–Pb collision.

In this paper, the strategy developed by ALICE to cope with these extraordinary computing challenges is presented. In Section 2, we present an overview of the O^2 computing system software framework for Run 3. In Section 3, the strategy of track pre-selection, HF reconstruction, skimming, and selection are discussed. In Section 4, some performance plots for HF analyses are presented and compared to equivalent results obtained with the software package adopted by ALICE for Run 1 and Run 2 analyses (*AliPhysics* [5]). In Section 5, the next steps in the analysis software developments are discussed.

2 Basic features of the O^2 analysis framework

The O^2 software framework is designed from the start to combine all the computing functionalities needed in a HEP experiment: detector read-out, event building, data recording, detector calibration, data reconstruction, physics simulation, and analysis. In this section, we will highlight a few key features of the O^2 *analysis* software, which includes all the utilities needed to perform the complete analysis chain, from the processing of the reconstructed events stored in AOD files to the extraction of the histograms used in the final analysis. A complete description of the various other functionalities of the O^2 framework can be found in Ref. 2.

Collections of objects (e.g. collisions, tracks, MC particles) are represented in the O^2 framework by flat Apache Arrow tables [6] in order to maximise the processing performance.

In this scheme, each object occupies one row and the columns correspond to properties of those objects. The O^2 data model supports the following most common column types:

- Static columns represent the basic type of column as they simply store values of a given type (e.g. space coordinates, momentum components, flags).

- Index columns store references to rows in other tables (e.g. indices of daughter tracks of a decay candidate, pointing to rows in the track table).

- Expression columns store results of simple calculations performed with static columns using predefined vectorised operations applied en masse on all rows of the table.

- Dynamic columns are functions operating on other types of columns in the same table and can be used to obtain values of quantities that require more complicated calculations involving values of static columns (e.g. kinematic quantities derived from momentum components).

Grouping operations can be performed by using indices, e.g. to group reconstructed objects belonging to the same collision. *Join* operations can be used to retrieve information stored in different tables belonging to the same reconstructed objects and merge them into a single joined table. This feature is for example particularly useful to create additional columns (e.g. *selection flags*) and select rows according to their values.

One of the consequences of the O^2 data model is the possibility to manipulate data in a fully declarative way, so that one can select rows using declared expressions instead of looping over rows explicitly. This feature is particularly convenient, for example, when *filtering* tracks according to various criteria or dividing tracks into *partitions*, e.g. based on their electric charge.

The O^2 analysis data model relies on a *modular* structure. Analysis tasks can subscribe to input tables, perform calculations on them and eventually produce derived tables or histograms. Task parameters are configurable from the command line or by using specific configuration files (JSON files). In the O^2 analysis framework, each component produces its output by processing the output of another component. Dependencies between tasks are dictated by their subscriptions to requested input tables. In this approach, running a chain of tasks in a full workflow simply requires providing a piped list of names of task binaries together with optional additional parameters. The topology of the workflow is determined automatically. The framework also provides the possibility to save derived tables to disk in the form of ROOT [7] trees for post-processing with external tools, e.g. for optimisation with Machine Learning techniques.

3 Heavy-flavour reconstruction, skimming, and selection

As introduced in the previous sections, the O^2 computing model paradigm relies on a large reduction of the data volume. The same guiding principle also drives the design of the O^2 HF reconstruction framework, which is meant to minimise the size of the derived datasets (also called *skims*) that contain the information about HF candidates. The reconstruction of HF candidates relies also on multiple loops over the collection of reconstructed tracks and on time-consuming algorithms of secondary vertex reconstruction and track-to-vertex propagation. As a consequence, the reconstruction of HF mesons and baryons down to very low p_T in a Pb–Pb event can take up to 50–100 seconds per central Pb–Pb collision, with a wide variability that depends on the selections applied. Achieving a good balance between data-size reduction and CPU time becomes therefore a fundamental requirement.

In this section, a detailed description of the various steps of the reconstruction and selection of a typical HF hadronic channel like $\Lambda_c^+ \rightarrow pK^-\pi^+$ is provided. This analysis, in

particular, plays a critical role in the physics program of ALICE in Run 3 since it is considered the golden probe for the study of the hadronisation mechanisms of charm in heavy-ion collisions. The need to study the production yields of this baryon, which is characterised by very small displacement with respect to the primary vertex of the collisions ($c\tau \approx 50\,\mu m$), had also a strong impact on the design of the upgraded ALICE 2 detector [1].

A detailed representation of the HF analysis workflow in O^2 is presented in the diagram in Fig. 1. Details of the various steps are provided in the following subsections.

3.1 Track pre-selection

The track tables containing the track collection of a given number of Pb–Pb collisions is accessed and filtered using declarative functions according to quality and kinematic selections. Given the need of measuring HF candidates down to very low p_T, only mild selections on the track transverse momentum ($p_T^{min} \approx 100$–$150\,\mathrm{GeV}/c$) are applied. Additional p_T-dependent selections on the distance of closest approach of the track to the primary vertex (DCA) are also considered. These selections rely on the fact that particles produced in HF decays are typically more displaced with respect to the primary vertex since they originate in weak decays. The need to measure baryons like the Λ_c^+, which present only moderate displacements with respect to the primary vertex position, limits the possibility of applying tight DCA selection on tracks (optimal choice of DCA_{min} in the range 0–20 μm, depending on the track p_T).

3.2 Track combination and proto-candidate pre-selection

Double and triple *loops* over the selected tracks belonging to each event are performed in order to reconstruct hadrons that decay in two or three charged particles (or *prongs*), respectively. At this stage tracks are combined to form *proto-candidates* by simply adding the 4-momenta of the considered tracks. Very loose selections are applied at this stage, e.g. on the invariant mass of the proto-candidates (based on hypotheses on the daughter-track masses) or on the candidate p_T. These selections provide a first significant reduction of the candidate background yield because they remove track pairs or triplets that do not present invariant masses compatible with any of the HF hadrons of interest. No significant efficiency losses for signal candidates are expected at this stage.

3.3 First secondary-vertex reconstruction and candidate building

For track pairs and triplets that pass the previous selection, the position of the vertex of decay (*secondary* or *decay vertex*) is reconstructed by minimising the distance of closest approach of the tracks to the vertex using the Newton–Raphson method. When the position of the secondary vertex is found, the tracks used in its calculations are propagated to the decay vertex using a Kalman algorithm. A complete reconstruction of the HF kinematics and of the candidate properties is then performed using the 4-vectors of the decay tracks propagated to the secondary vertex and the position of the primary and secondary vertices. A first selection at the candidate level is applied based on quantities like candidate p_T, cosine of pointing angle, and product of prong impact parameters. For the selected candidates, the indices of the decay tracks are stored as columns of a new derived table called *TrackIndexSkim*. The information of the decay channel for which the candidate was selected is also stored. The *TrackIndexSkim* tables are produced and stored permanently on disk for all the datasets. Detailed studies on Monte Carlo (MC) simulations that simulate the Run 3 detector geometry will be performed to optimise these selection criteria, with the goal of minimising the size of these derived productions while preserving large signal efficiencies.
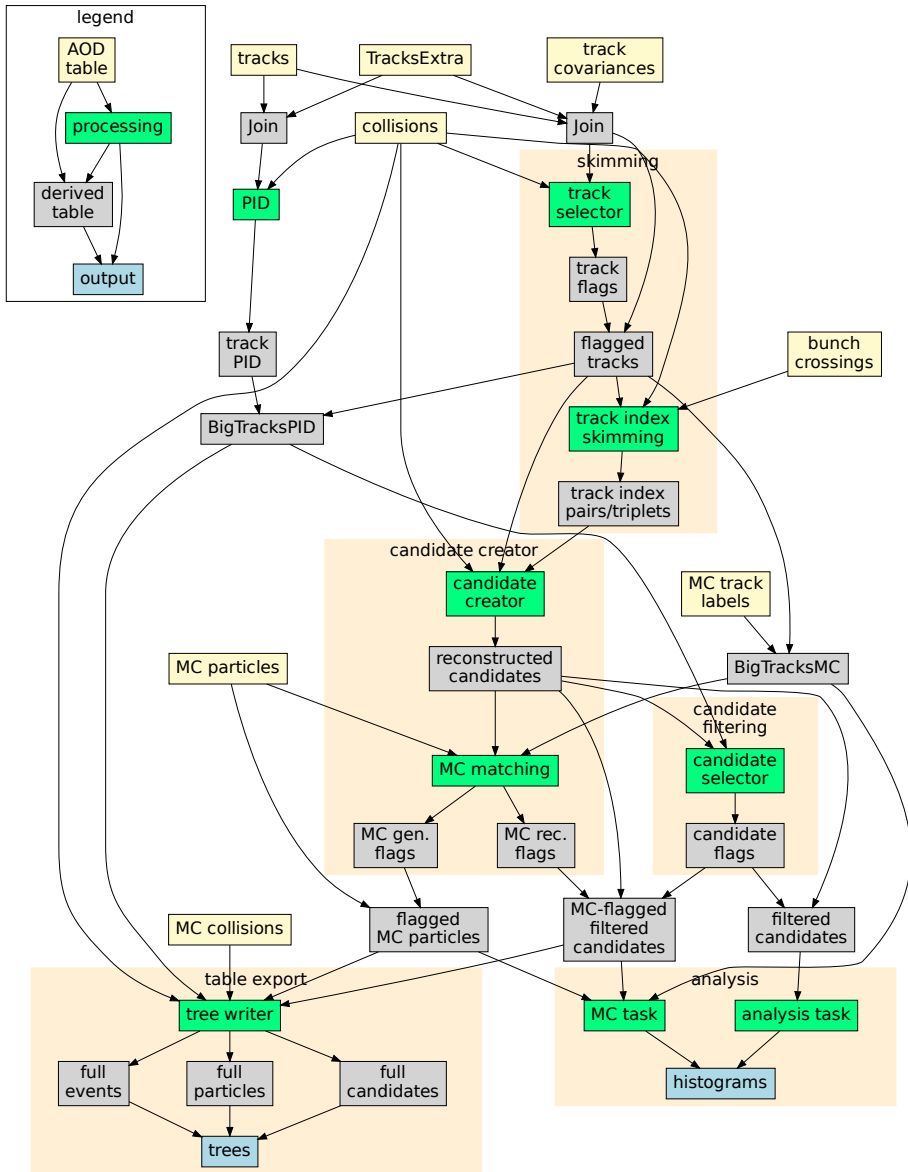
Figure 1. Diagram of the HF analysis framework in O^2.

3.4 Candidate rebuilding and Monte Carlo matching

The *TrackIndexSkim* tables are later used as a starting point of the last reconstruction step, usually called *candidate rebuilding*. At this stage, the track indices stored for each candidate are used to access the complete set of parameters of the corresponding reconstructed tracks from the AOD track tables. The secondary-vertex reconstruction is then repeated and additional candidate properties, that are needed for the signal selection and cannot be calculated

dynamically on-the-fly due to their complex calculation, are computed (e.g. uncertainties of quantities computed in the vertex reconstruction procedure). The complete list of quantities needed for the final candidate selection and analysis are stored in a derived table of reconstructed candidates, also called *CandidateSkims*. Given the large number of variables stored for each candidate, candidate skims are expected to be stored on disk only for a small fraction of the events for debugging and optimisation purposes.

For simulated events, reconstructed decay candidates are matched with their generated counterparts by checking the correspondence between the candidate prongs and the expected decay tree. The so-called MC matching procedure is performed also for generated MC particles by checking their identity and their decay tree. At this stage, derived tables with MC flags used for the estimation of the signal efficiencies and the optimisation of the signal and background selections are produced.

3.5 Final candidate selection and histogram filling

In a dedicated selector task, tailored for each decay channel, accurate analysis level selection criteria based on decay topology and PID are applied to the reconstructed candidates. The selection results are stored in an additional column of a new dedicated table that is later joined with the candidate table to filter them.

The final step of the workflow is the user analysis task where the histograms needed for the analysis, which include e.g. the distributions of the invariant mass of candidates, are saved in ROOT files. For MC events, histograms with quantities of generated MC particles and MC-matched candidates are also produced.

4 Performance and preliminary validation results

A dedicated framework was developed to validate the O^2 HF software package by comparing the results to the ones obtained with the Run 2 software (AliPhysics). Distributions of variables related to track selection, secondary vertex reconstruction, candidate selections are used to assess the performance and quality of the new analysis implementation. The analyses of $D^0 \rightarrow K^-\pi^+$ and $\Lambda_c^+ \rightarrow pK^-\pi^+$ decay channels were chosen as benchmark cases for the reconstruction of the two-prong and three-prong candidates, respectively. The reconstruction and analysis of other hadrons via different hadronic decays, including resonant channels and cascades, are also under development, profiting extensively from the O^2 framework modularity. In this section, some preliminary results of the reconstruction of $D^0 \rightarrow K^-\pi^+$ and $\Lambda_c^+ \rightarrow pK^-\pi^+$ candidates are reported. For these studies, MC simulations of proton–proton collisions generated with PYTHIA 6 [8] at $\sqrt{s} = 5.02$ TeV are considered. In order to perform the O^2 analyses, the Run 2 data were converted to the O^2 AOD format using a dedicated conversion software [5]. Similar studies performed on converted Pb–Pb real data and MC simulations are currently ongoing.

In Fig. 2, the invariant-mass distributions of $D^0 \rightarrow K^-\pi^+$ candidates reconstructed and selected with AliPhysics and O^2 are presented. The very good agreement between the results obtained with the two software packages confirms the solidity of the various steps of the O^2 HF analysis framework.

Dedicated studies were also performed to test the correctness of the MC matching procedure and to assess the accuracy of the calculation of the selection variables. As an example, the signal and background distributions of the cosine of pointing angle (left) and p_T of the first daughter track (right) for $\Lambda_c^+ \rightarrow pK^-\pi^+$ candidates are presented in Fig. 3. The same distributions were compared to the ones obtained using the Run 2 analysis software and were found to be in very good agreement.

A central aspect of the validation of the analysis framework is the study of the reconstruction and selection efficiencies for the different HF candidates. In Fig. 4, the reconstruction efficiencies of the $D^0 \to K^-\pi^+$ and $\Lambda_c^+ \to pK^-\pi^+$ decay channels as a function of the candidate p_T are presented. A detailed comparison to the efficiencies extracted using the Run 2 analysis software is currently ongoing. The possibility of performing PID selection on daughter tracks was recently included in the O^2 framework and will require an accurate validation and optimisation.

5 Prospects and next steps

Thus far, the development of the HF O^2 framework focused on the implementation of the basic features and tools needed to perform the various steps of the HF reconstruction and on
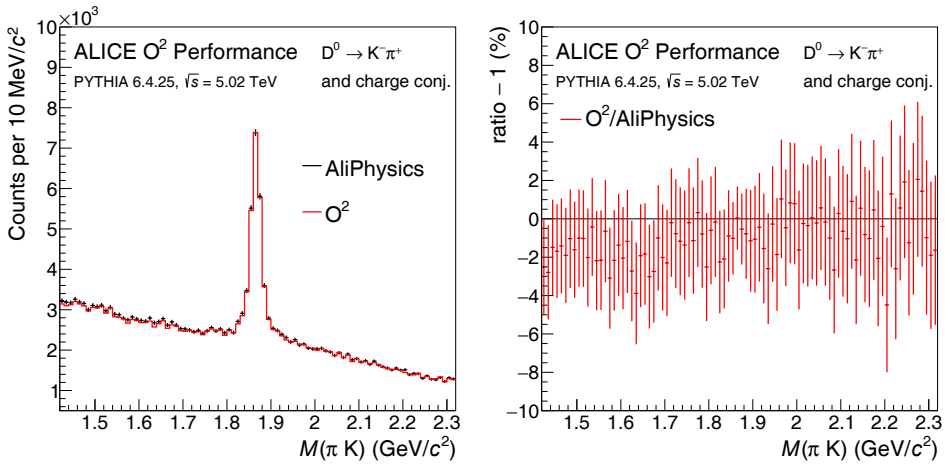

Figure 2. Comparison of the invariant-mass spectra of $D^0 \to K^-\pi^+$ candidates selected using topological criteria obtained with AliPhysics and O^2 tasks.
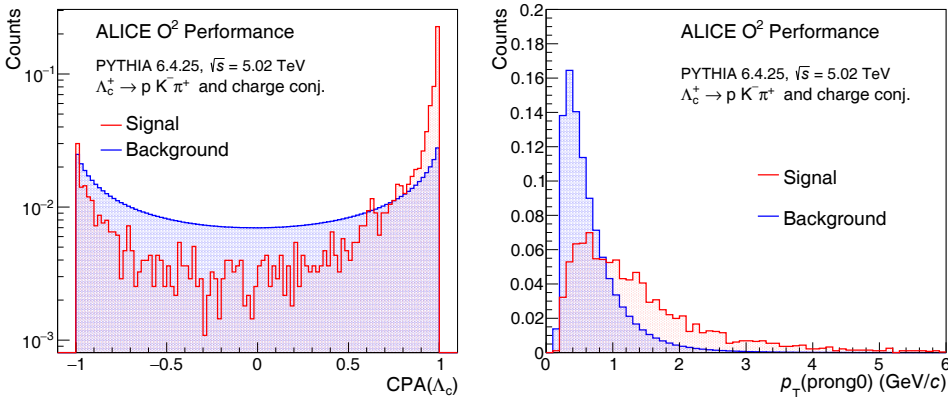

Figure 3. Signal and background distributions of cosine of the pointing angle (left) and p_T of the first prong (right) of $\Lambda_c^+ \to pK^-\pi^+$ candidates. The distributions are normalised by their integral.

the design of a modular workflow that minimise the size of the derived analysis objects. In this section, we present a brief overview of few areas of current and future developments.

5.1 Reconstruction of new and more complex decay channels

In the coming months, the framework will be expanded to include new and more complex decay channels, like fully reconstructed beauty mesons and baryons or heavier charm baryons. Heavier HF mesons and baryons are frequently measured via reconstruction of hadronic decay channels that include lighter HF hadrons as intermediate states, as in the case of $B^+ \to \overline{D^0}\pi^+$ or $\Lambda_b^0 \to \Lambda_c^+\pi^-$. Thanks to the modular design of the HF O^2 framework, the reconstruction of these channels can be easily performed using, as a starting point, the candidate skims developed for the standard two-prong or three-prong channels, like $D^0 \to K^-\pi^+$ or $\Lambda_c^+ \to pK^-\pi^+$, limited CPU and disk-space resources. By using the same approach, one can efficiently perform analyses of HF hadron correlations or HF-tagged jets.

5.2 Optimisation of the selection criteria using Run 3 simulations

Additional development of the pre-selection strategy for Run 3 is required to meet the demands imposed by the volume of gathered data. This strategy will be developed using MC simulations incorporating an accurate description of the new ALICE detector, where an enhanced tracking resolution and DCA discrimination will allow for a more effective background rejection when building and selecting HF candidates. Dedicated selection strategies will also be developed to perform online tagging of HF events to be used during the high-luminosity pp runs that are expected at the beginning of Run 3. The possibility of performing complete reconstruction and selection of HF candidates while taking data will allow to substantially increase the pp statistics that can be saved on disk.

5.3 HF vertex optimisation with Deep Neural Networks and KF particle reconstruction

By far the most computationally intensive part of the HF candidate creation is the reconstruction of two-prong and three-prong vertices and the propagation of the track parameters
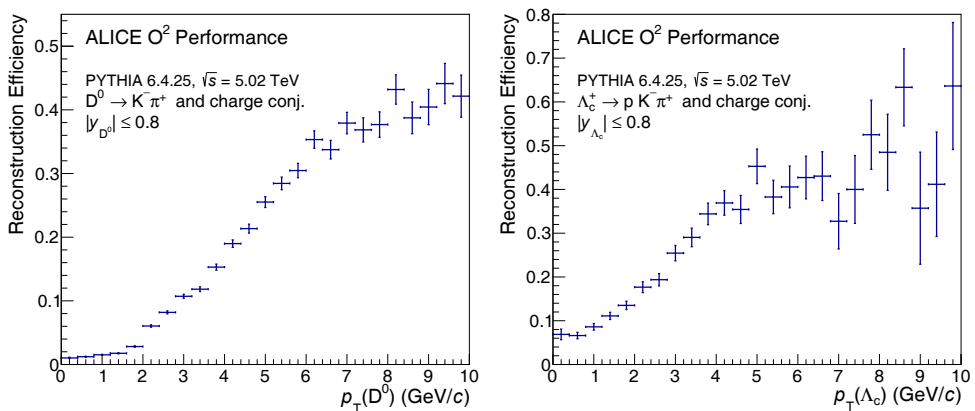

Figure 4. Reconstruction efficiencies of $D^0 \to K^-\pi^+$ (left) and $\Lambda_c^+ \to pK^-\pi^+$ (right) decays as a function of the candidate transverse momentum in the rapidity range $|y| \leq 0.8$.

to the reconstructed vertex position. Since many of the selection variables used during the pre-selection stage rely on the vertex reconstruction, it must be evaluated for a large number of track combinations. Therefore any optimisation of the vertex finder has the capacity to alleviate the toll on CPU resources. To this effect, the development of a fast vertex finder using Deep Neural Networks (DNN) was proposed. In this approach, DNN are used to map the correlations between the helix parameters of the daughter tracks and the position of the secondary vertex. The possibility of performing also the propagation of the decay-track parameters to the secondary vertex will also be explored. This algorithm could be adopted at the early stages of the HF candidate reconstruction where only a rough estimation of the vertex position is needed. If proven to be successful, this fast vertexer would provide strong benefits in terms of CPU resources since the analytic vertex reconstruction would need to be run only for a much smaller fraction of the track pairs and triplets.

The use of the KF Particle Finder package [9] for the reconstruction of HF topologies will also be explored. The KF package, which was already adopted for some HF analyses of Run 1 and Run 2 data, allows to find an optimal estimation of the parameters of short-lived particle tracks by combining already found daughter tracks of long-lived charged particles and to achieve the highest possible accuracy.

6 Conclusions

In this paper, the strategy of reconstruction, selection, skimming, and analysis of heavy-flavour hadrons for Run 3 were presented. Preliminary results on the reconstruction of charm mesons and baryons were shown, including distributions of invariant mass, selection parameters for signal and background, and reconstruction efficiencies. They were compared to and found to agree with the results obtained with the analysis package used by ALICE in Run 1 and Run 2. The current results indicate the solidity of the new framework, which is expected to provide high computing performances in the extreme environment provided by Pb–Pb collisions at the LHC. In the upcoming months, several features will be added in order to include the reconstruction of new and more complex topologies. Specific studies will be performed to optimise the workflow in order to further reduce the size of the derived datasets and the CPU processing time.

References

[1] B. Abelev et al. (ALICE Collaboration), Journal of Physics G: Nuclear and Particle Physics **41**, 087002 (2014)

[2] J. Adam et al. (ALICE Collaboration), Tech. Rep. CERN-LHCC-2015-006. ALICE-TDR-019 (2015), `https://cds.cern.ch/record/2011297`

[3] A. Andronic, F. Arleo, R. Arnaldi, A. Beraudo, E. Bruna, D. Caffarri, Z.C. del Valle, J.G. Contreras, T. Dahms, A. Dainese et al., The European Physical Journal C **76**, 107 (2016)

[4] J. Adam et al. (ALICE Collaboration), *ALICE O^2 Project*, `https://github.com/AliceO2Group/AliceO2`

[5] J. Adam et al. (ALICE Collaboration), *AliPhysics*, `https://github.com/alisw/AliPhysics`

[6] The Apache Software Foundation, *Apache Arrow*, `https://arrow.apache.org/`

[7] R. Brun, F. Rademakers, P. Canal, A. Naumann, O. Couet, L. Moneta, V. Vassilev, S. Linev, D. Piparo, G. GANIS et al., *root-project/root: v6.18/02* (2019), `https://doi.org/10.5281/zenodo.3895860`

[8] T. Sjöstrand, S. Mrenna, P. Skands, Journal of High Energy Physics **2006**, 026 (2006)

[9] P. Kisel, I. Kisel, P. Senger, I. Vassiliev, M. Zyzak, EPJ Web Conf. **173**, 04009 (2018)