LHCb-INT-20XX-YYY
September 22, 2023

ML-based charged isolation tool for Run 3 and a comparative assessment

Ł. Hartman[1,2], A. Mathad[3], J. Garcia Pardinas[2]

[1] *EPFL, Lausanne, Switzerland*
[2] *CERN, Geneva, Switzerland*
[3] *University of Zurich, Zurich, Switzerland*

Abstract

During Runs 1 and 2, many LHCb analyses have relied on multivariate isolation techniques to inspect extra particles in a collision event that accompany a signal decay, to suppress/control backgrounds and to reconstruct excited states. There are currently no such tools for Run 3, where the isolation problem becomes more complicated due to the increased particle multiplicity per event. The isolation of interesting particles in hadron collisions at LHCb becomes more challenging as the instantaneous luminosity increases in Run III. A gradient boosted decision tree is used to select associated particles. A systematic study of the input variables leads to a reduced correlation between the parameters. The input variables have been chosen to be topological, and independent of the decay kinematics. They are a combination of cone, track and vertex isolation variables. This model presents better background rejection and signal efficiencies than cone and impact parameter based track isolation tools. Similar performance is achieved in Run III conditions compared to Run II. This report is part of a summer student project.

Contents

Acknowledgements

I would like to thank my supervisors Dr. Abhijit Mathad and Dr. Julian Garcia Pardinas for their help and support throughout this summer student project. Their unbounded enthusiasm and passionate explanations have made these weeks a fascinating learning experience. I would like to thank CERN for the organisation of the program, allowing me to fulfill my dream of participating in scientific research at CERN. Finally, I would like to thank the LHCb collaboration for providing the data and software used in this project.

1 Introduction

In hadron collisions, a pletora of particles are created. With around five collisions per bunch crossing at the LHCb experiment, the number of tracks reconstructed reaches the hundreds [1]. The associations of tracks to form signal candidates is therefore a challenge. The aim is thus to isolate those tracks from the rest of the event. Such isolation is crucial to many semi-leptonic analyses, and is often used in other LHCb studies. Isolation consists of selecting tracks that are well apart from the other tracks in the event. Isolation is used to reduce combinatorial background, that is, when tracks that do not belong together are combined as a signal candidate [2]. Partially reconstructed decays can also mimic the signal. Isolation tools can be used to reduce this background by rejecting candidates with compatible extra tracks [3]. In contrary, excited state with very short lifetimes require the addition of extra tracks produced in the de-excitation to the signal candidate (e.g., a decay to ground state). These extra tracks need to be persisted in order to fully reconstruct the decay [4]. In that case, a tool usually used to select isolated tracks can be inverted to find groups of tracks that belong together. Generally, the isolation is used to enrich the data sample in the decay of interest. To precisely study the background, the isolation can also be used to select a control sample, where the signal is suppressed.

With the increase in luminosity, the disk storage limitations become more challenging [5]. Physics analyses require the pre-processing of the full events to select the interesting tracks. A more efficient selection is therefore required to match the storage limitations. Simply tightening the selection criteria risks loosing important information for later studies. There is therefore a clear need to develop a general isolation tool with good performances when applying conservative requirements.

The isolation can be done using multiple techniques. The simplest one is the cone isolation, which is based on the opening angle between the candidate and the extra track. The track isolation requires the tracks to come from the same primary vertex or to be completely detached, using the common origin of all the signal tracks. The complexity of the challenge has led to the development of multiple multivariate analysis (MVA) techniques. Such tools have been designed and are used for Run I and Run II. The aim of this project is to compare the performances of the cone and track isolation techniques, and to develop the first isolation MVA tool for Run III. This new tool is based on the boosted decision tree (BDT).

In this report, the different techniques are presented in further details in Section 2. The Monte Carlo simulations and selections are described in Section 3. The variables used in the previous Run II MVA are studied, and the performances of the cone isolation, track isolation, previous MVA and the new MVA are compared in Section 4. In Section 5, for Run III, the set of variables is extended. The distributions and correlations are discussed. The performances of cone isolation, track isolation and the new MVA are compared. The efficiencies are also studied for two further exclusive decays.

2 Techniques

2.1 Cone isolation

The cone isolation is a simple isolation criterion, based on the opening angle between the candidate and the extra track. The cosine of the opening angle is defined as

$$\cos\theta = \frac{\vec{p}_{\text{cand}} \cdot \vec{p}_{\text{track}}}{|\vec{p}_{\text{cand}}| \cdot |\vec{p}_{\text{track}}|}, \tag{1}$$

where $\vec{p}_{\text{cand}}$ is the momentum of the candidate obtained by summing over the momenta of the particles coming from its decay, and $\vec{p}_{\text{track}}$ is the momentum of the extra track. This technique makes use of the fact that particles coming from the same decay are boosted in the same direction, due to the large momentum of the parent particle produced at LHCb. For signal tracks (i.e., originating the same heavy hadron decay as the candidate), the opening angle should be small. On the other hand, background tracks are mostly obtained by combining two random independent tracks in the event. The opening angle between these tracks should therefore be larger. The forward acceptance of the LHCb detector nonetheless favours relatively small opening angles, attenuating the discriminating power of this method.

2.2 Track isolation

The track isolation technique uses the fact that tracks coming from the same mother particle originate in the same PV. This is translated on the requirement that the best PV of the extra track is the same as the best PV of the candidate. The best PV is defined as the PV with the smallest impact parameter (IP) significance. The impact parameter is defined as the distance between the PV and the point of the closest approach of the track. To include tracks that start further away from the PV, another criterion is added. If the IP significance for the extra track with respect to any PV is larger than a given threshold, the track is also considered as a signal track. All in all, this technique selects all the tracks coming from the same PV as the candidate, and all the tracks that do not come from any PV. The discrimination power of this technique is limited by the high number of tracks coming from any given PV, and the high PV multiplicity in an event. This increases the likelihood of mis-associating tracks to the PV. Therefore, the background rejection efficiency is quite poor, reaching at best 80% background rejection, assuming on average 5 PVs per events.

2.3 Boosted decision tree

Because of the relatively poor performance offered by single-variable techniques discussed previously, a multivariate analysis (MVA) is used to improve the signal to background discrimination. Such a MVA for track isolation has been previously developed for Run I [6]. The MVA is based on a boosted decision tree, a type of MVAs based on decision trees [7]. It was trained on a cocktail of exclusive decays $B \to D^{**}\mu\nu$, where the D^{**} decays to a D^* meson and a pion. The mesons D^* and the muon constitute the candidate; the extra pion is the extra track used for training. The background consists of every other track in the event. This MVA uses a set of variables that are based on the topological properties of the extra track, in order to have a model that generalizes well to other decays. The variables are presented in Section 4.1.

For the new model, the BDT library used is *XGBoost* [8]. To improve the performance of the model, further variables have been added. The strategy to choose the variables is in three parts. Firstly, variables related to track, cone, and vertex isolation techniques are compiled. The distributions for the signal and background tracks are then compared to select the variables with the largest discriminating power. A choice between the correlated variables is then made.

The hyperparameters are the number of trees, the maximal depth and the learning rate. These values are found by minimizing the difference in predictions for the train and test samples to avoid over-training, using the Kolmogorov-Smirnov test statistic between the two samples [9]. The optimizing library used is *optuna* [10].

3 Monte Carlo sample preparation

Two Monte-Carlo (MC) simulated data sample are used. The event is propagated through a simulation of the LHCb detector, including the signal digitalization. The reconstruction is done using the same software as the data. Firstly, a typical semi-leptonic exclusive decay is simulated. The decay is

$$\Lambda_b \to \Lambda_c^* \mu\nu, \quad \Lambda_c^* \to \Lambda_c \pi^+\pi^-, \quad \Lambda_c \to pK\pi. \tag{2}$$

The tracks originating in the Λ_c decay and the muon form a signal candidate, and the two pions are the extra signal particles that are used to train the model. The truth matching from the reconstructed tracks to the true MC particles is done by requiring the exact match for the full decay chain. No requirement is applied on the extra tracks.

The second sample is an inclusive sample, which contains all the $b\bar{b}$ decays. The signal candidates are two body combinations h^+h^- missing one extra track $h^{\pm}_{\text{extra}}$, where h is any long-lived charged particle. In the inclusive sample, the signal is defined as any triplet of particles coming from the same long-lived b hadron. The lifetime threshold is 10^{-13} s. Decays passing these requirements, but with an intermediate long-lived charmed hadron are not considered signal. Such decays would not present the same topological features.

Run II The exclusive line selected is `Strippingb2LcMuXB2DMuForTauMuLine`. Loose track and vertex quality selections are applied to the inclusive sample, along with a selection on the distance of the closest approach (DOCA) of the two body candidate, and very loose selection on the h^+h^- combination in the inclusive sample. The transverse momentum of the any extra track is required to be larger than 0.8 GeV and the sum larger than 2 GeV for the second extra track, with a total larger than 2.5 GeV. The candidate must be displaced with respect to the PV with significance larger than 50. The DOCA is required to be less than 0.1 mm. The best PV DIRA must be larger than 0.999.

Run III The exclusive line has very loose selection criteria on the $\Lambda_c\mu$ combination. The candidate DIRA must be larger than 0.7. The transverse momentum of the Λ_c larger than 0.8 GeV, its total transverse momentum must be larger than 2.5 GeV, and DOCA significance less than 40. The extra signal candidate must have a momentum larger than 1 GeV, and transverse momentum larger than 0.1 GeV. The track goodness of fit must be less than 6, the ghost probability less than 0.5, and the track minimal impact parameter significance larger than 2. There is a very loose selection on the h^+h^- combination on the inclusive sample, requiring the candidate momentum to be larger than 6 GeV, the transverse momentum to be larger than 1 GeV, the track goodness of fit less than 3, the ghost probability less than 0.3, and the track minimal impact parameter significance larger than 9. The two body candidate is required to have a maximal DOCA less than 0.1.

File size reduction For Run II, only events with partially reconstructed signal are kept, with a reconstructed final state of 4 particles. To reduce the size of the files used for training and evaluation, only the signal tracks and 10 background tracks per event are kept. Only one signal candidate is picked per event and only events containing at least one signal track are used for the background tracks.

4 Run II

4.1 Nominal variable distributions

The set of variables used in the previous BDT algorithm is used as the starting point. The variables are listed below. For three of the five variables, their significance has been selected over the variable itself; internal LHCb studies have shown that the significance is more discriminating than the variable itself. Large peaks or long tails have been avoided by rescaling or taking the logarithm (base 10). The variables are:

- The cosine of the opening angle between the candidate momentum and the momentum of the extra track, transformed to avoid the peak at 1 following the mapping $x \to 1-(1-x)^{0.2}$, written as $\text{Tranformed}\cos\theta$.
- The logarithm of the transverse momentum of the extra track $\log(p_T[GeV])$.
- The logarithm of the minimum IP significance of the extra track for all primary vertices $\log \min\chi^2_{\text{IP wrt PV}}$.
- The logarithm of the IP significance of the extra track for the SV $\log\chi^2_{\text{IP wrt SV}}$.
- The logarithm of the flight distance significance $\log\chi^2_{\text{FD wrt PV}}$.
- The logarithm of the difference in flight distance significance with and without the extra track $\log\Delta\chi^2_{\text{FD wrt PV}}$.

The cosine of the opening angle is shown on Figure 1a, the transverse momentum on Figure 1b, the IP significance with respect to the PV on Figure 1c and with respect to the SV. Very good discrimination between signal and background is visible for all variables. The flight distance significance is shown on Figure 2a and the difference in flight distance significance with and without the extra track on Figure 2b. These variables show much lower discriminating power.

The correlation plot of the variables is shown on Figure 3a for the signal tracks, and on Figure 3b for the background tracks. The correlation between most variables are low, except for the flight distance significance and the difference in flight distance significance, as expected. Since these variables also show the lowest discrimination power, the latter has been removed when the extended set of variables.

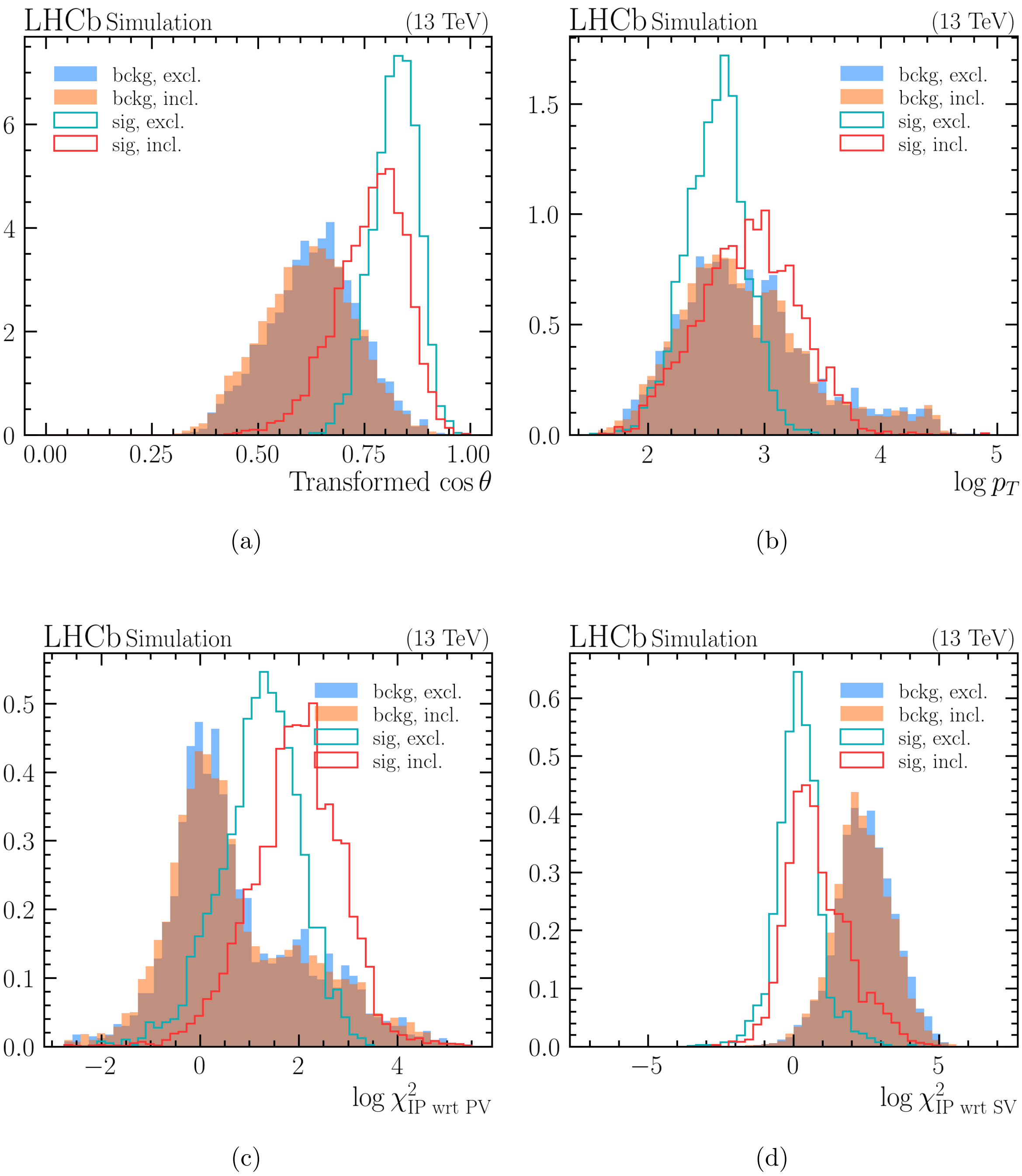
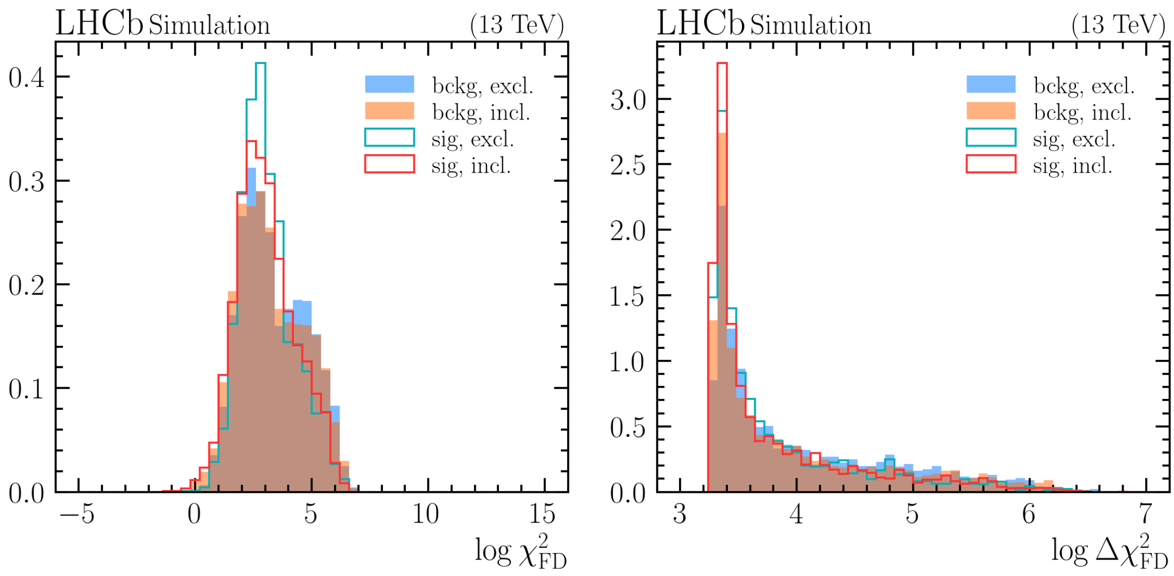


Figure 1: Signal and background distributions for (a) the cosine of the opening angle, as defined in the text, (b) the transverse momentum p_T of the extra track, (c) the IP significance with respect to the PV, and (d) with respect to the SV. The exclusive decay mode and the inclusive sample are overlaid; the brown part represents the overlap between the backgrounds of both samples.

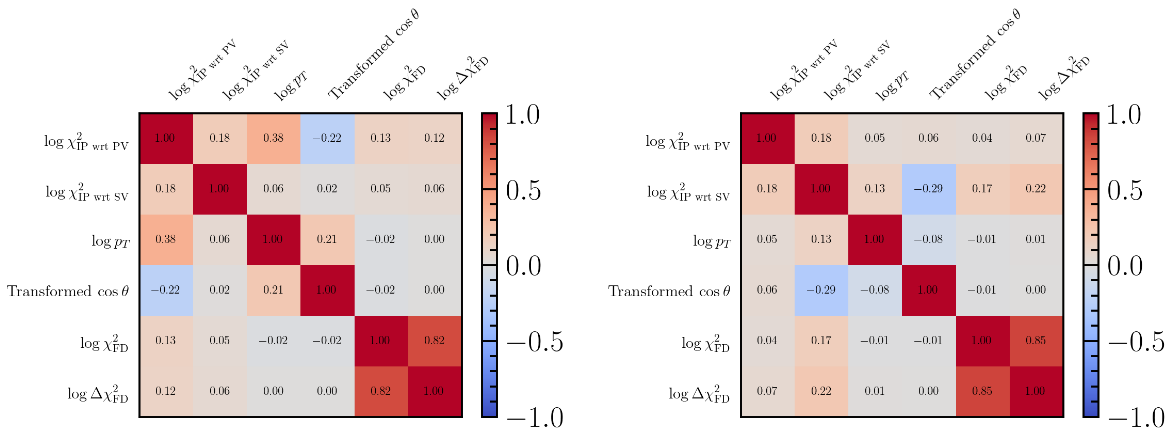
4.2 Extra Run II variables

Aiming at improving the performance of the new BDT compared to the previous one, further variables are used to train the model. A full list of variables has been compiled



(a) (b)

Figure 2: Signal and background distributions for (a) the flight distance significance and (b) difference in flight distance significance PV, and (d) with respect to the SV. The exclusive decay mode and the inclusive sample are overlaid; the brown part represents the overlap between the backgrounds of both samples.

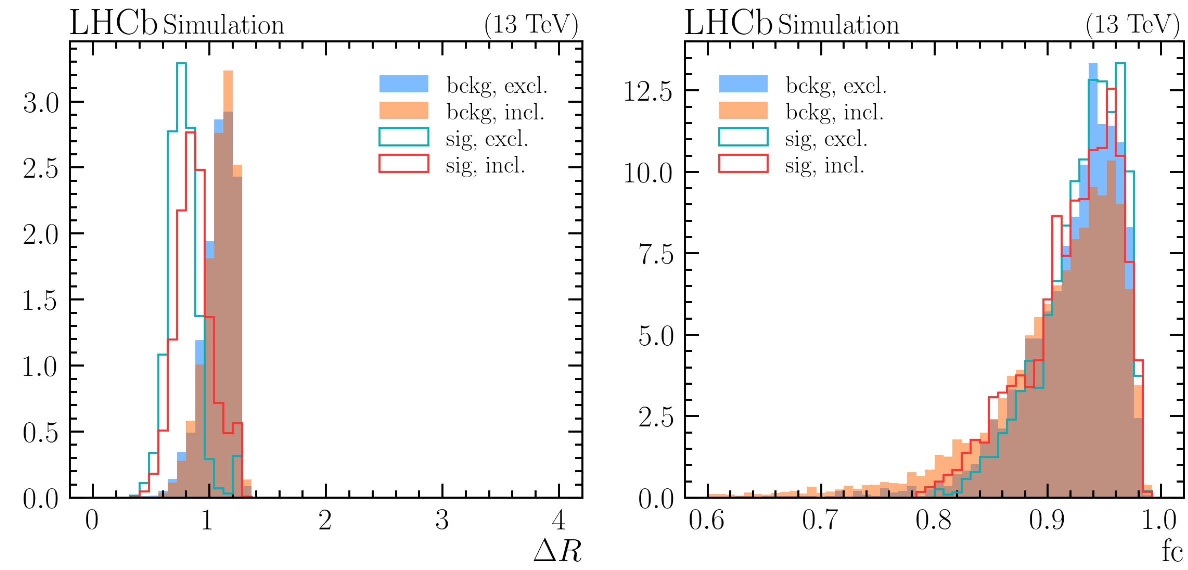


(a) (b)

Figure 3: Correlation plot of the variables for (a) the signal tracks and (b) the background tracks.

from different sources for the Run III model. For Run II, only two extra variables have been added, because the other variables have not been computed during the Run II event reconstruction. The two variables are:

- The angular distance between the extra track and the candidate $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}^{0.2}$. This value is taken to the power 0.2 to avoid the



(a) (b)

Figure 4: Signal and background distributions for (a) the angular distance ΔR between the candidate and the extra track, and (b) the composite variable fc. The exclusive decay mode and the inclusive sample are overlaid; the brown part represents the overlap between the backgrounds of both samples. A DIRA cut has been applied in the pruning.

peak at 0.

- A composite variable called flight cosine (fc) [11]. It is based on the angle between the direction of flight of the candidate and the total momentum of the candidate and the extra track. It also contains the total transverse momentum of the candidate and the extra track. The formula is

$$\mathrm{fc} = \frac{|\vec{P}_\mu + \vec{P}_\mathrm{B}| \cdot \alpha^{h+\mathrm{B},\ \mathrm{PV}}}{|\vec{P}_\mu + \vec{P}_\mathrm{B}| \cdot \alpha^{h+\mathrm{B},\ \mathrm{PV}} + P_{T,h} + P_{T,\mathrm{B}}}. \tag{3}$$

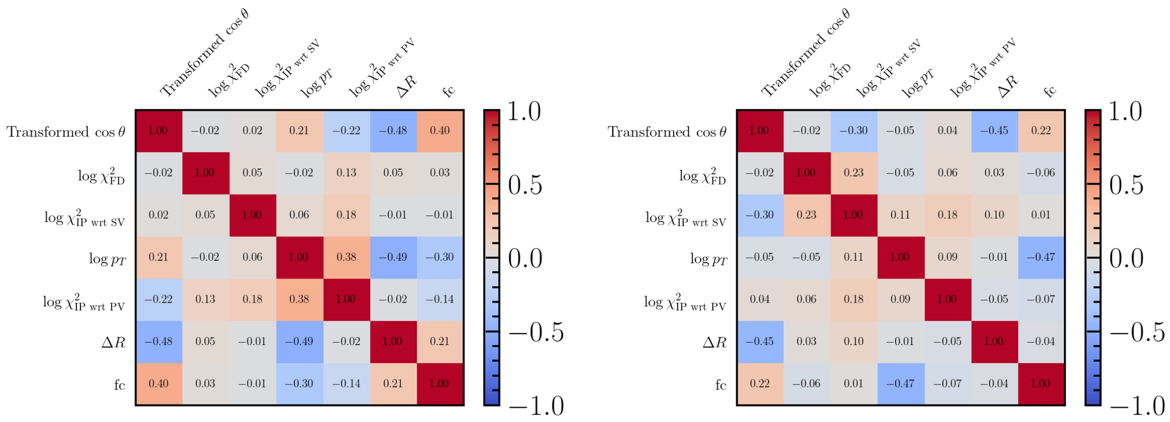
  This variable tends to 0 for signal tracks (the extra track is aligned with the candidate and the transverse momentum is large), but goes to 1 for background tracks.

The distributions of these variables are shown on Figure 4a and Figure 4b, respectively. The new variable fc shows poor discrimination power, caused by the DIRA cut applied during the pruning.

The difference in flight distance significance has been removed from the set of training variables. With the two extra parameters, the correlation plot is shown for the signal tracks on Figure 5a and for the background tracks on Figure 5b.

4.3 MVA performance

The performance of the different techniques are evaluated using the receiver operating characteristic (ROC) curve. Such figures are obtained by drawing the background rejection as a function of the signal efficiency. Graphically, an ideal classifier will look be represented



(a) (b)

Figure 5: Correlation plot of the extended set of variables for (a) the signal tracks and (b) the background tracks.

by two straight lines in the (x, y) plane, connecting the point $(0, 1)$ to $(1, 1)$ and then to $(1, 0)$. A random classifier would be the diagonal from $(0, 1)$ to $(1, 0)$. To summaries this behaviour, the area under the curve (AUC) is computed. A more performant classifier usually has a larger AUC, but this figure of merit alone is not sufficient. Indeed, the interest mostly lays in the high signal efficiency ($\leq 90\%$). The model is trained separately on the exclusive decay and inclusive sample. They are then both separately evaluated on both decays (leading therefore to four combinations). The ROC curve for the model trained on and evaluated on the exclusive decay (inclusive sample) is shown on Figure 6a (Figure 6b). For the cross-training, the model trained on the inclusive sample (exclusive decay) and evaluated on the exclusive decay (inclusive sample) is presented on Figure 6c (Figure 6d). In every case, the signal efficiency obtained using the track isolation technique is very high, but at the cost of very poor background rejection. The cone isolation shows higher background rejection, but only at lower signal efficiency. The previous BDT has similar performance compared to the cone. The new XGBoost based BDT is much more powerful. It shows larger signal efficiency and background rejection.

When considering for example 95% signal efficiency in the exclusive decay, the background rejection is at 25% using the track isolation, 80% using the cone isolation or the previous BDT, and 95% (90%) using the new BDT trained on the exclusive decay (inclusive sample). On the inclusive sample, the cone isolation technique shows worse background rejection than the previous BDT, but performs better at high signal efficiency. The new BDT trained on the inclusive sample outperforms all other technique. The performance from inclusive training on the exclusive decay is worse than when trained on itself, as expected, but remains better than the other techniques. On the contrary, the results obtained by training on the exclusive decay and extrapolating on the inclusive sample is more nuanced. The background rejection is tighter using the previous BDT, but the performance at high signal efficiency is better using the new BDT.

Since the full signal decay contains two extra tracks, there is a risk that the MVA is performant at recognizing the tracks independently but rarely predicts both tracks as being signal like. To measure this potential effect, the signal efficiency is defined

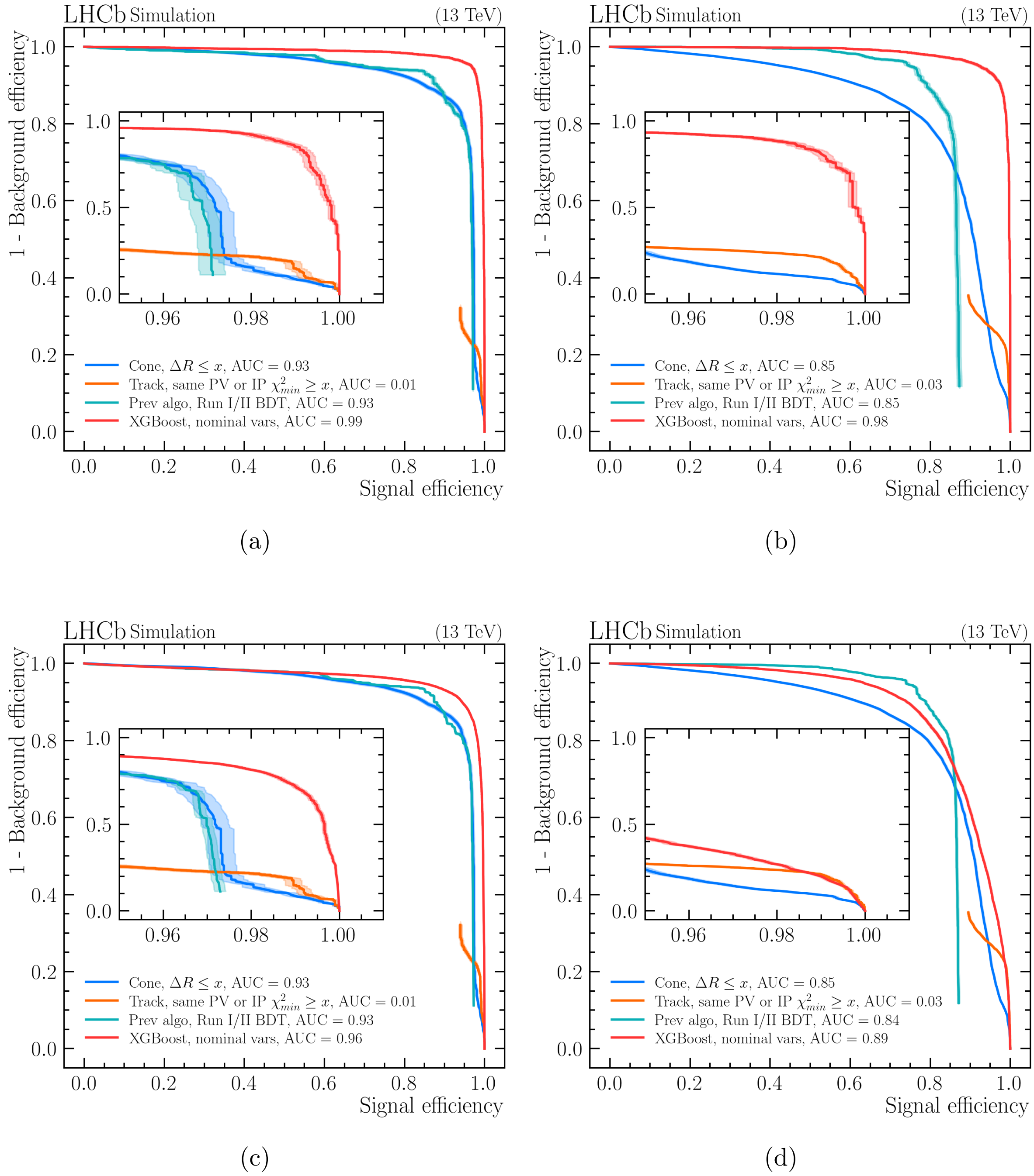
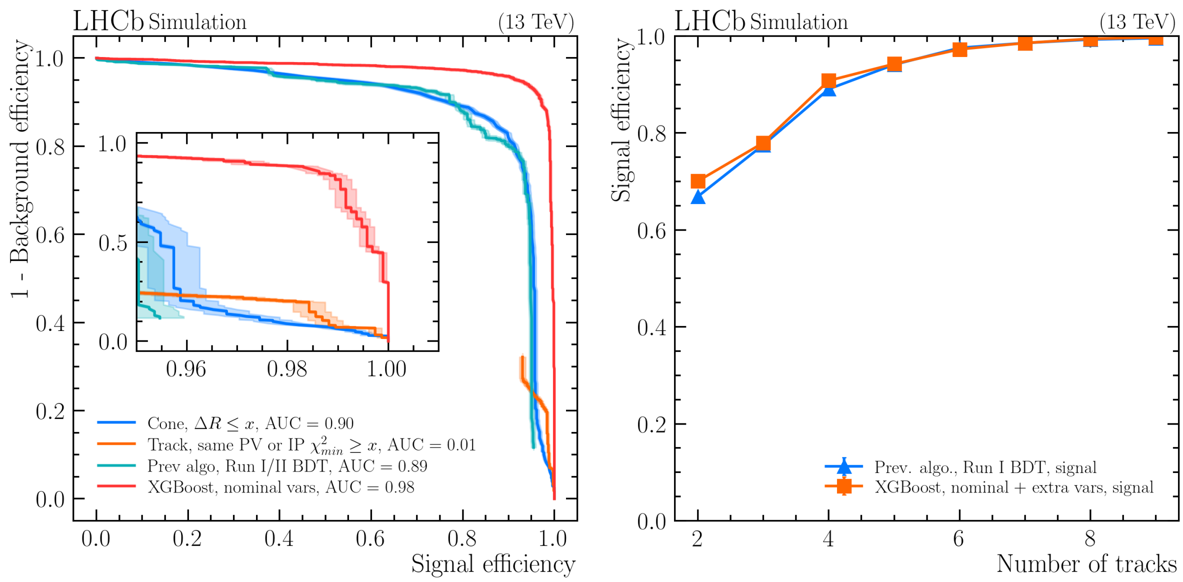


Figure 6: ROC curves for Run II (a) trained and evaluated on the exclusive decay, (b) trained and evaluated on the inclusive sample, (c) trained on the inclusive sample and evaluated on the exclusive decay, and (d) trained on the exclusive decay and evaluated on the inclusive sample. The XGBoost model is trained using the nominal variables.



(a) (b)

Figure 7: (a) ROC curve for at the candidate level for Run II. The XGBoost model is trained using the nominal variables. (b) The signal efficiency when selecting the n tracks with the most signal like MVA prediction, for both the previous and the new BDT.

differently than before. It is computed by counting how often both tracks pass the criteria BDT predicition $\geq x$, divided by the total number of candidates. Again, the new BDT outperforms any other technique. When analysing the data, there is often interest in reducing the storage requirements; this is often done by only keeping the tracks of interest. The idea is therefore to keep only the n tracks with the most signal like BDT prediction. The signal efficiency as a function of n is shown on Figure 7b, using the output of both the previous and the new BDT. This selection reaches 90% signal efficiency when selecting the 4 best tracks out of the around 100 tracks in any given event, with only two background tracks per full signal candidate[1]. The performance is very similar for both BDTs.

[1]This study is done on a different reduced-size file. For each signal candidate, 350 background tracks are kept (or every track in the event if there are less than 350 background tracks).

5 Run III

5.1 Variable distributions and correlations

Similar to Run II, the distributions of the variables are studied for the signal and background tracks. The exclusive decay and inclusive sample are compared. The normalized distributions are shown on Figure 8a to Figure 10c. Most variables show strong signal to background discrimination. For the fc variable, there is a clear bump at low values for the signal, which is not visible in Run II (Figure 4b). The new added variables are defined as:

- PV distance: the distance between the computed PV position with and without the extra track. It is expected to be positive (more forward) for signal tracks, and centred around 0 for background tracks.
- SV distance: the distance between the computed SV position with and without the extra track. It is expected to be centred around 0 for signal tracks, but shifted backwards (negative) for background tracks. This is caused by the association of a track coming directly from the PV.
- DIRA: the cosine of the angle between the momentum of the extra track and the vector between the PV and the SV. It is expected to be peaked at 1 for signal, and broader for background.
- Difference of end vertex significance $\Delta\chi^2_{\text{end vtx}}$: The difference in fit significance of the candidate end vertex, with and without the extra track. It is expected to be larger (worse) for background tracks, and centred around 0 for signal tracks.
- Distance of the closest approach χ^2_{DOCA}: The fit significance of the distance of the closest approach between the candidate and the extra track. It is expected to be large for background tracks, and smaller for signal tracks.

There are significant differences between the signal distributions of the exclusive decay and inclusive sample in variables such as ΔR, $\log p_T$, PV distance, $\Delta\chi^2_{\text{end vtx}}$, $\log\chi^2_{\text{IP wrt SV}}$ and $\log\chi^2_{\text{DOCA}}$.

The correlation plot is shown on Figure 11 for the signal tracks, and on Figure 12 for the background tracks. The correlation between the different variables is mostly well controlled. There is significant correlation for the DIRA and fc, which is expected since fc is computed using the DIRA. There is also strong correlation between the two angular variables, the cosine of the opening angle and ΔR. In general, the correlation is slightly lower for background tracks compared to the signal tracks, except for the PV and SV distances.

5.2 MVA performance

As in Run II, the performances of the different techniques are evaluated using the ROC curve. The ROC curve for the model trained on and evaluated on the exclusive decay (inclusive sample) is shown on Figure 13a (Figure 13b). For the cross-training, the model trained on the inclusive sample (exclusive decay) and evaluated on the exclusive decay (inclusive sample) is presented on Figure 13c (Figure 13d). Compared to Run II in

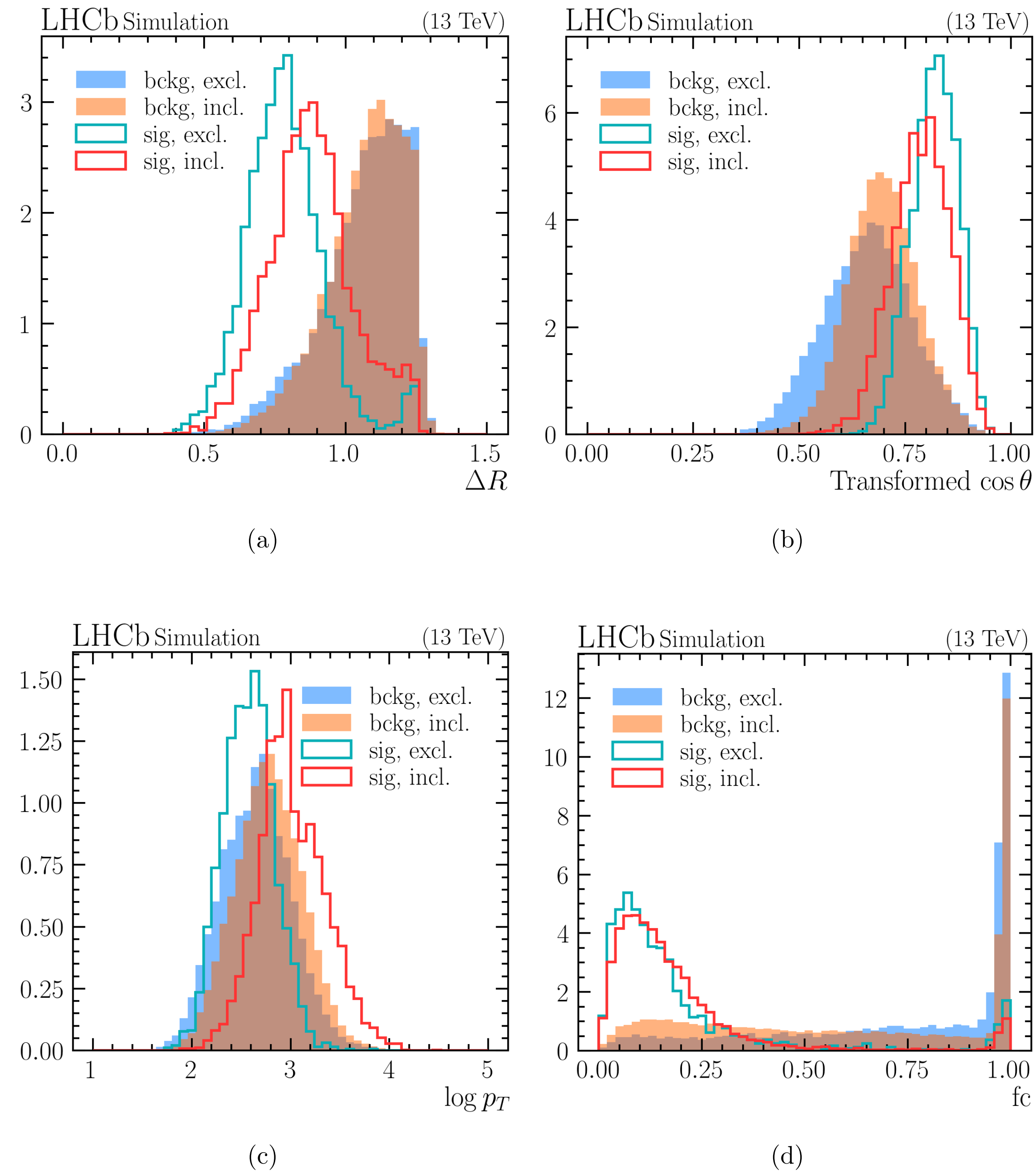


Figure 8: Signal and background distributions for (a) the cone ΔR, (b) the transformed cosine of the opening angle $\cos\theta$, (c) the transverse momentum $\log p_T$, and (d) the variable fc. The exclusive decay mode and the inclusive sample are overlaid; the brown part represents the overlap between the backgrounds of both samples.

Section 4.3, the cone isolation technique shows lower background rejection. The increase in background tracks is interpreted as the cause of it. The simple track isolation technique shows very poor background rejection. The new BDT with the nominal set of variables is more performant than the non-MVA tools, but shows about 10% less background rejection than the new BDT with the extended set of variables on the exclusive decay. For the

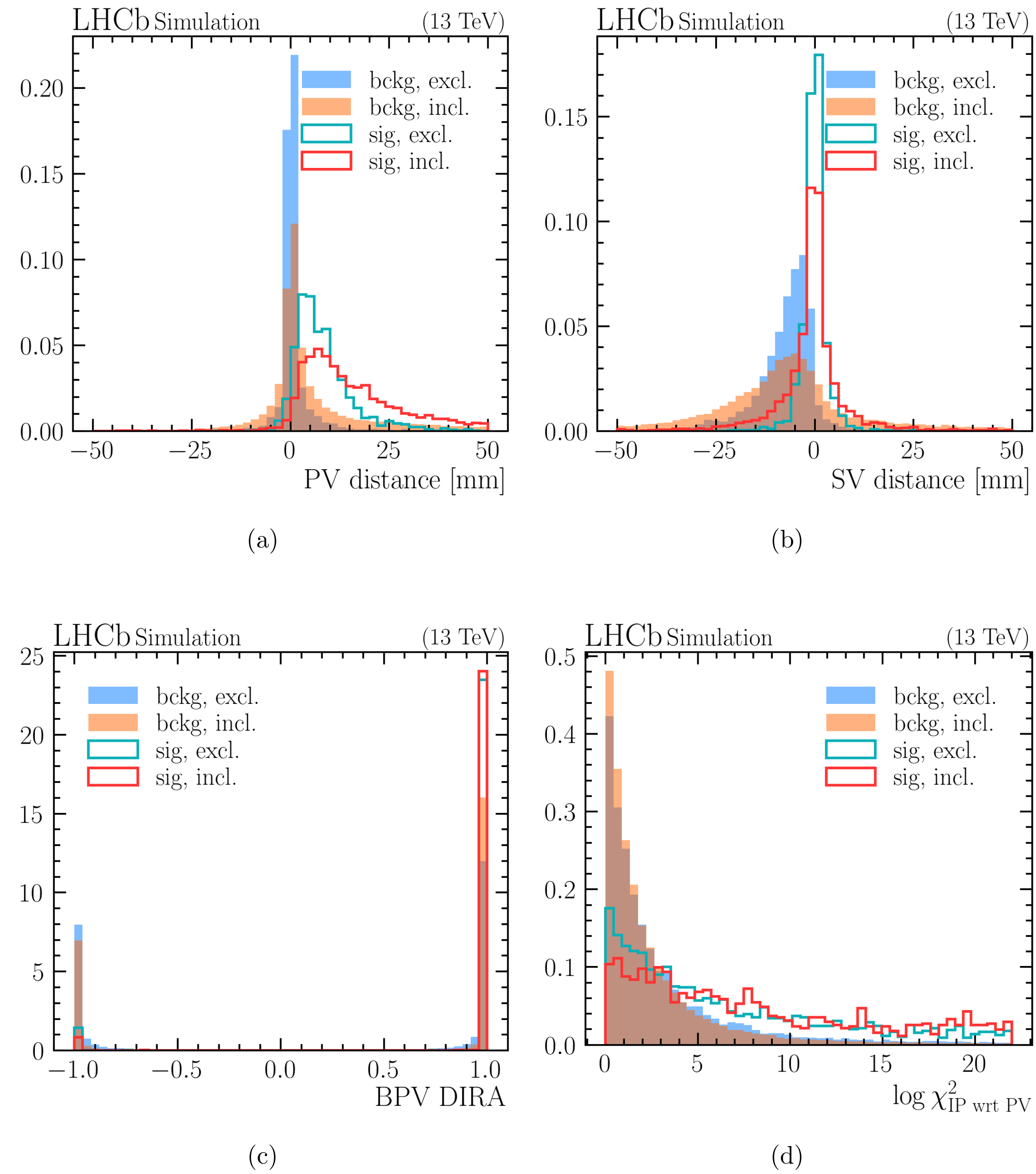
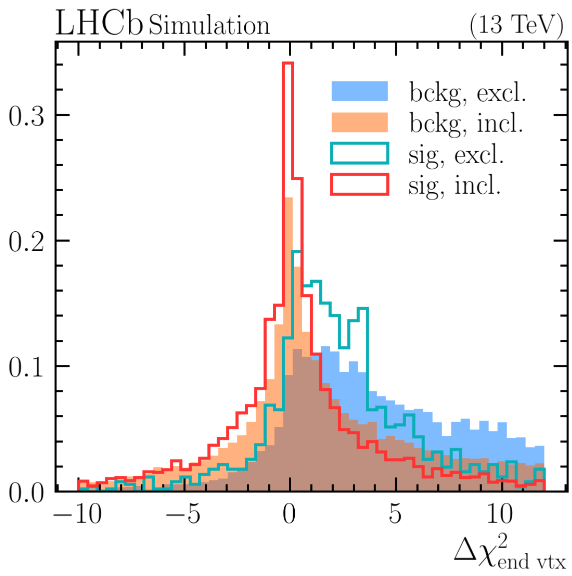
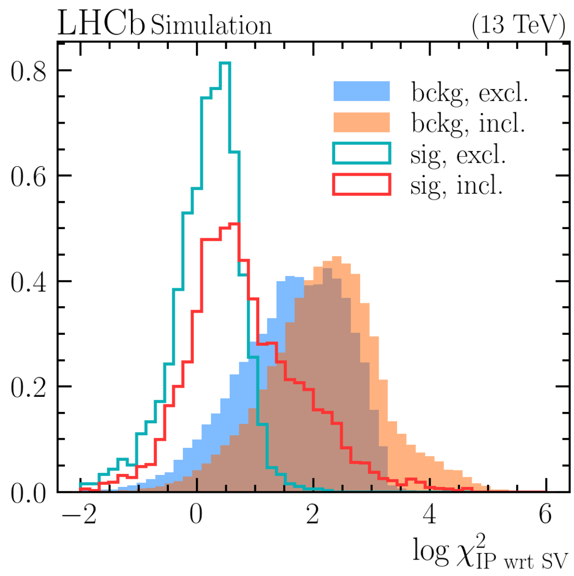


Figure 9: Signal and background distributions for (a) the PV distance, (b) the SV distance, (c) the BPV DIRA, and (d) the significance of the IP with respect to the PV $\log \chi^2_{\mathrm{IP\ wrt\ PV}}$. The exclusive decay mode and the inclusive sample are overlaid; the brown part represents the overlap between the backgrounds of both samples.

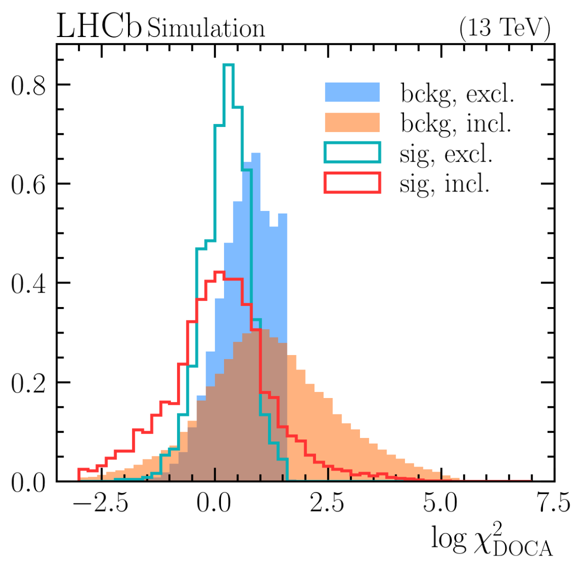
inclusive sample, the cone isolation technique has lower signal efficiency than on the exclusive decay. The new BDT is more performant for both sets of variables, with a smaller difference between the two sets. The new BDT trained using the extended set remains the most performant technique. With the extended set of variables, the evaluation of the inclusively trained model on the exclusive decay is only slightly worse performing



(a)



(b)



(c)

Figure 10: Signal and background distributions for (a) difference in candidate end vertex significance $\Delta\chi^2_{\text{end vtx}}$, (b) the significance of the IP with respect to the SV $\log \chi^2_{\text{IP wrt SV}}$, and (c) the DOCA significance $\log \chi^2_{\text{DOCA}}$. The exclusive decay mode and the inclusive sample are overlaid; the brown part represents the overlap between the backgrounds of both samples.

than the exclusively trained model. The reduced set drops performance to about the level of the cone isolation. Lastly, the generalization from the exclusive decay onto the inclusive sample follows the same trend. Interestingly, the nominal set of variables shows a slightly better background rejection at high signal efficiencies. The ROC curve for the reconstruction of the full signal candidate is shown on Figure 14a. The performance of

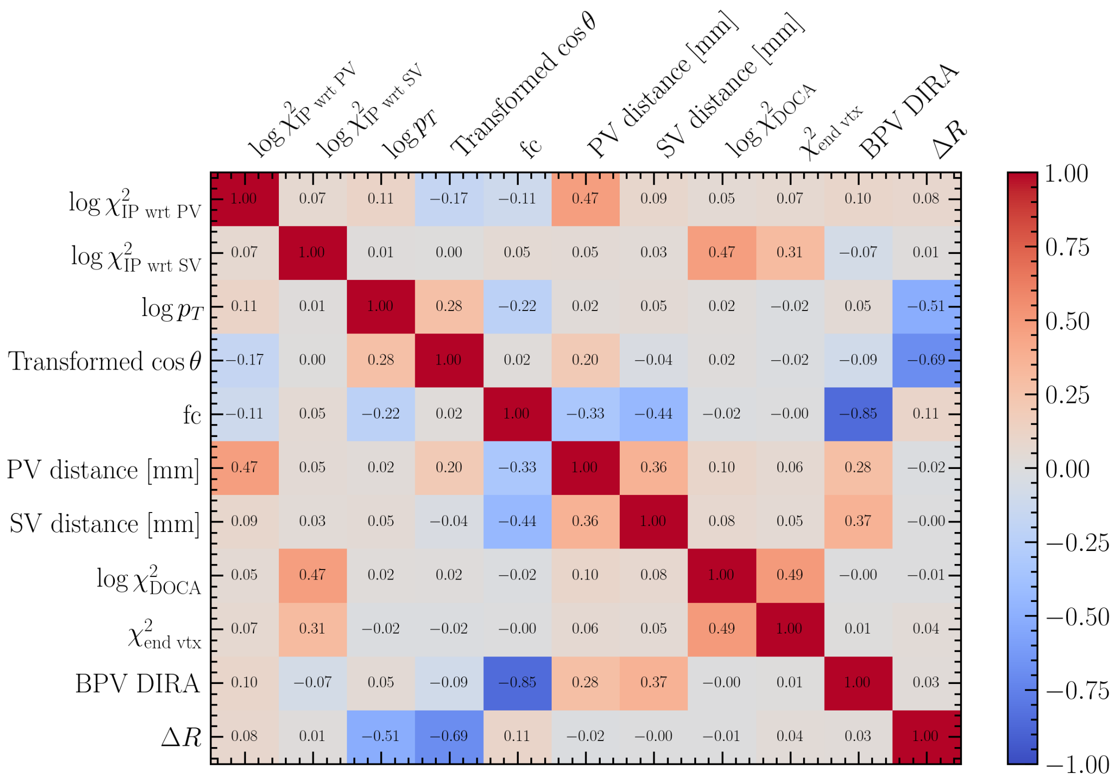


Figure 11: Correlation plot of the variables for the signal tracks.

the new BDT with the extended set of variables is slightly lower, but remains very high. The cone isolation and simple track isolation techniques both have a stronger reduction in signal efficiencies. Finally, the signal efficiency when selecting the n tracks with the most signal like prediction is shown on Figure 14b. Compared to Run II (Figure 7b), the signal efficiency is much lower. To reach 90% efficiency, about 70 tracks are needed in Run III, whereas only 4 where needed in Run II.

5.3 Other exclusive decays

In order to study how well the model generalizes to other exclusive decays, two further decay chains are considered. The first one is

$$B_u \to D^*(2010)\tau^+, \quad D^*(2010) \to D^0\pi^-, \quad D^0 \to K^-\pi^+, \quad \tau^+ \to \mu^+\bar{\nu}\nu, \tag{4}$$

whilst the second one is

$$\Lambda_b^0 \to \Lambda_c^+\mu^-\bar{\nu}\nu, \quad \Lambda_c^+ \to p^+K^-\pi^+. \tag{5}$$

The extra track in the first decay chain is the π^-, which is very soft. The candidate consists of the D^0 meson (reconstructed from the Kaon and the pion) and the muon. In the second decay chain, the two extra signal tracks are the $(K^-\pi^+)$ pair, with kinematics close to that of the $\Lambda_c^* \to \Lambda_c\pi^+\pi^-$ decay. The candidate is made of the anti-proton and the muon. The distributions of the variables are shown in the Appendixes. A new BDT model is trained on the combination of the three exclusive decays. The performance of

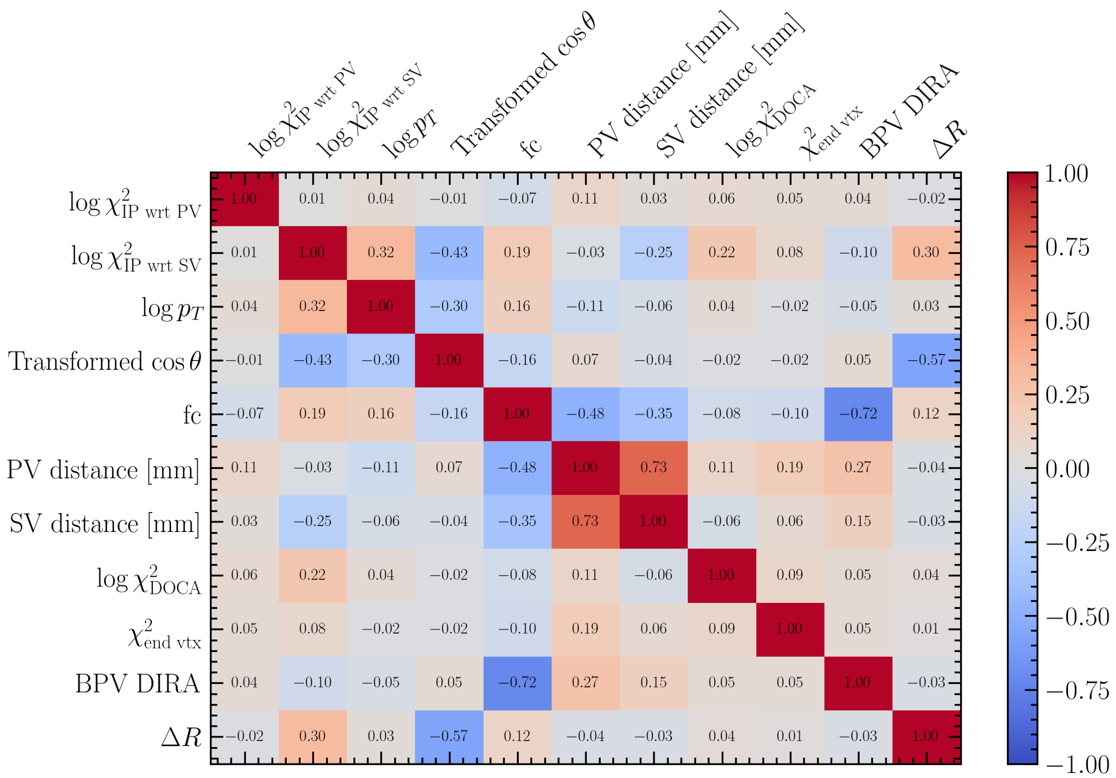


Figure 12: Correlation plot of the variables for the background tracks.

each sample is then evaluated separately, as shown on Figure 15a. The two Λ_b decays have very similar performances. On the other hand, the B_u decay performs more poorly. It comes from the differences in the variables distributions, caused by the different decay kinematics. The model is also trained on the inclusive sample only, before being evaluated separately on the exclusive decays. The ROC curves are shown on Figure 15b. The same behaviour for the three decays as in the previous figure is observed. The AUC of the $\Lambda_c^* \to \Lambda_c \pi^+ \pi^-$ decays is only marginally lower than the one obtained when training on itself. The results indicate that the B_u decay is significantly different from the inclusive sample and the two other exclusive decays. The main difference is the softness of the extra track, as can be seen in Appendix A. Therefore, the B_u decay is combined with the inclusive sample to train the model. The performance is shown on Figure 15c. The performances of the $\Lambda_c^* \to \Lambda_c \pi^+ \pi^-$ and $\Lambda_b \to \Lambda_c \mu^- \bar{\nu} \nu$ decays are very similar to the one obtained when training on the combination of the three exclusive decays or on the inclusive sample. The $B_u \to D^*(2010)\tau^+$ decay now shows similar performance, vastly improving the previous result.

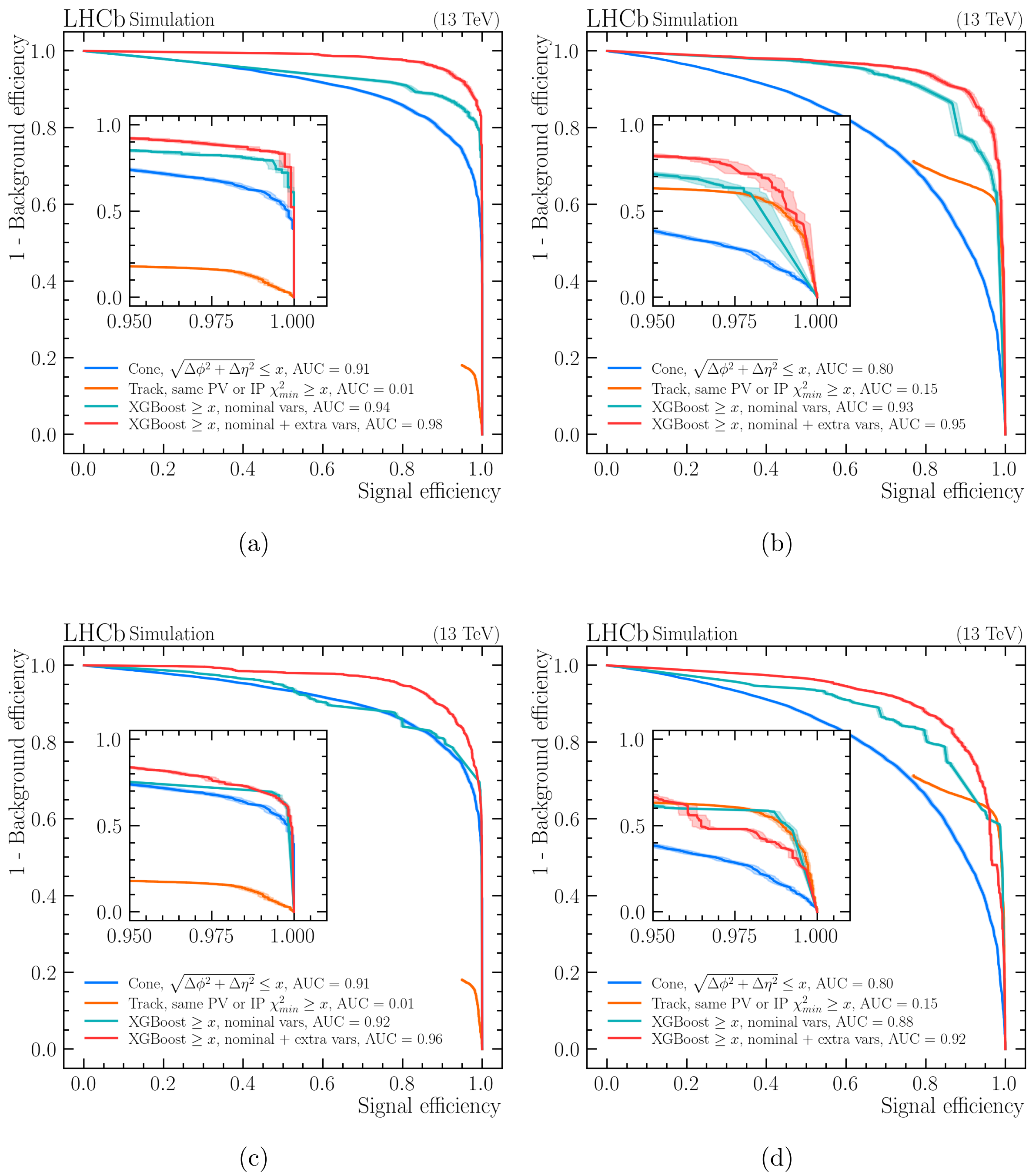
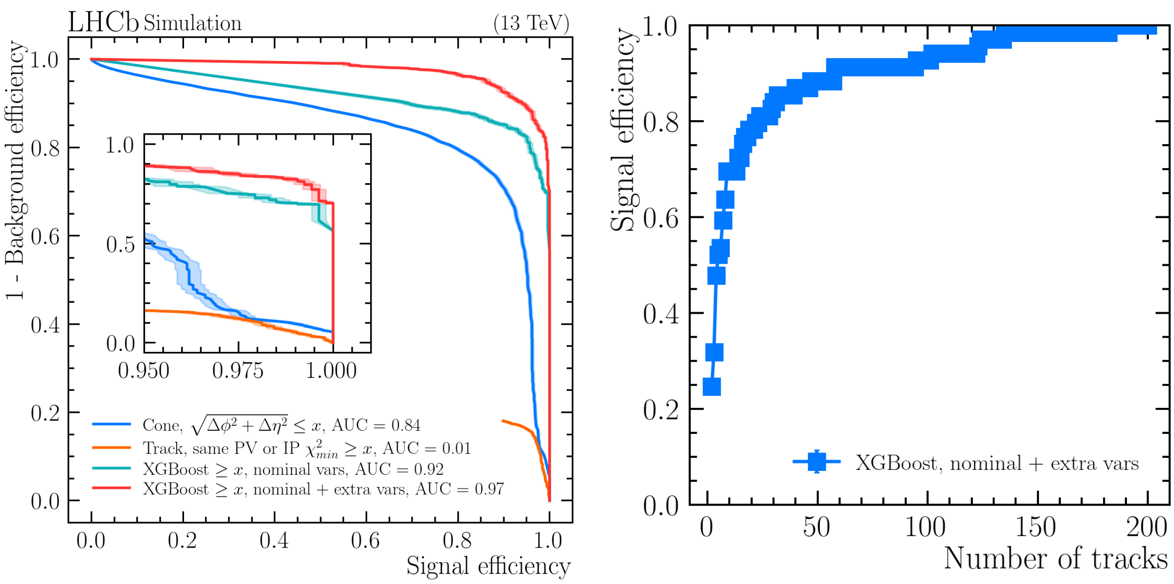
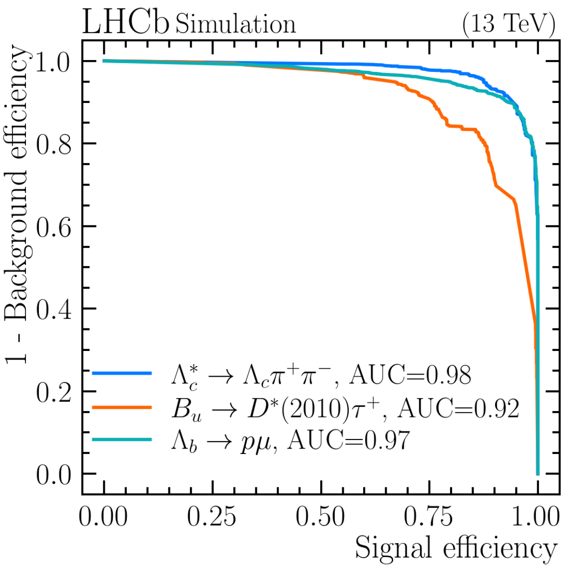


Figure 13: ROC curves for Run III (a) trained and evaluated on the exclusive decay, (b) trained and evaluated on the inclusive sample, (c) trained on the inclusive sample and evaluated on the exclusive decay, and (d) trained on the exclusive decay and evaluated on the inclusive sample. The XGBoost model is trained using the nominal and the extended set of variables.

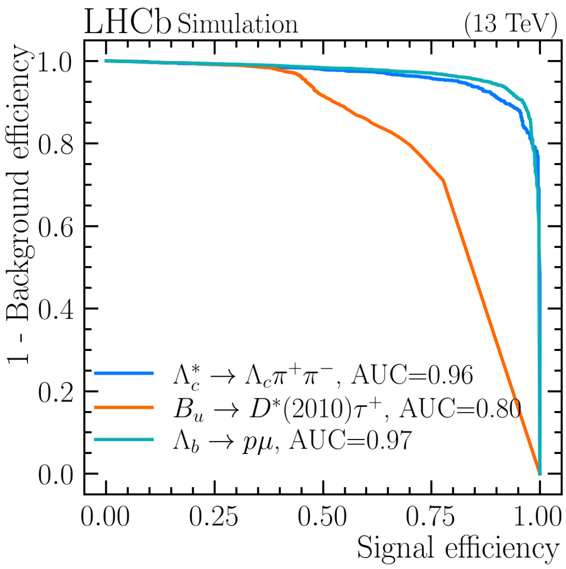


(a) (b)

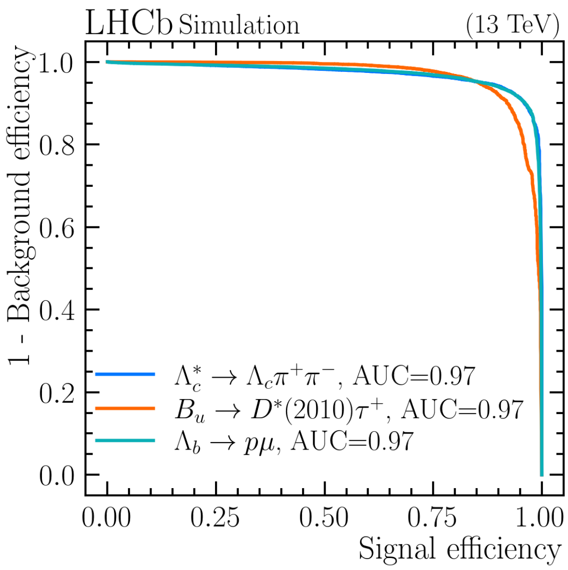
Figure 14: (a) ROC curve for at the candidate level for Run III. The XGBoost model is trained using the nominal variables. (b) The signal efficiency when selecting the n tracks with the most signal like MVA prediction, using the new BDT.



(a)



(b)



(c)

Figure 15: ROC curves for the different exclusive decays (a) trained on the combination of the three exclusive decays and evaluated on the exclusive decays, (b) trained on the inclusive sample and evaluated on the exclusive decay, and (c) trained on the combination of the inclusive sample and the $B_u \to D^*(2010)\tau^+$ decay, and evaluated on the three exclusive decays.

6 Conclusion

In this report, the first systematic study of the performance of different isolation tools in nominal Run III conditions is presented, and compared to Run II conditions. The background rejection and signal efficiency of the cone isolation technique are similar to those obtained using the previous BDT, whilst being more powerful than the track isolation. The new BDT outperforms all three techniques, even when using only the same variables as the previous model. These results are obtained both on a semi-leptonic exclusive decay and on an inclusive $b\bar{b}$ sample. Furthermore, the model trained on the inclusive sample provides improved background rejection and signal efficiency compared to the other techniques. The efficiencies of selecting the full candidate within an event are improved using the new BDT. The removal of a variable and the addition of two new ones is found to reduce the correlations.

Comparing simulations of Run II to those of Run III, the background rejection of the cone isolation technique drops significantly, because of the increase in background tracks. The new BDT using the extended set of variables yields similar performances than in Run II using the nominal set of variables, whilst keeping them mostly uncorrelated. The incorporation of the extra variables leads to a tighter background rejection. The new BDT is significantly more performant. The signal efficiency when selecting the tracks with the most signal like predictions is much lower in Run III than in Run II, because of the larger background. Nonetheless, this approach remains a powerful tool to reduce the disk storage required, as the current strategy is to save every track in the event.

The inclusive model evaluated on different exclusive decays show good efficiencies, outperforming the other techniques. Soft extra tracks are found to diminish the discriminating power of the new BDT. The combination of an exclusive decay with a soft extra track and the inclusive sample is found to improve the performance of the model on that decay, whilst keeping the performance on the other decays. The new BDT is therefore found to be a powerful tool for the isolation of signal tracks in the LHCb experiment.

To fully validate this approach, the background rejection and signal efficiency will be studied on more exclusive decays, with varied kinematics. The accuracy of the simulations of the different variables will also be evaluated. If the need arises, two separate models for different decay kinematics could be trained. A specific model can be trained on a specific decay to obtain the best performance.

Appendices

A Other exclusive decays variable distributions

The distributions of the variables used in the new BDT model for the exclusive decay $B_u \to D^*(2010)\tau^+$ are shown in the figures Figure 16a to Figure 18c. The correlations of the signal and background tracks are shown on Figure 19 and Figure 20 respectively. The extra track is much softer and with larger opening angle than in the inclusive sample. For the exclusive decay $\Lambda_b \to p\mu$, the distributions of the variables used in the new BDT model are shown in the Figure 21a to Figure 23c. The correlations of the signal and background tracks are shown on Figure 24 and Figure 25 respectively. The distributions almost exactly match that of the inclusive decay. It explains the excellent performance of the inclusive model on this exclusive decay.

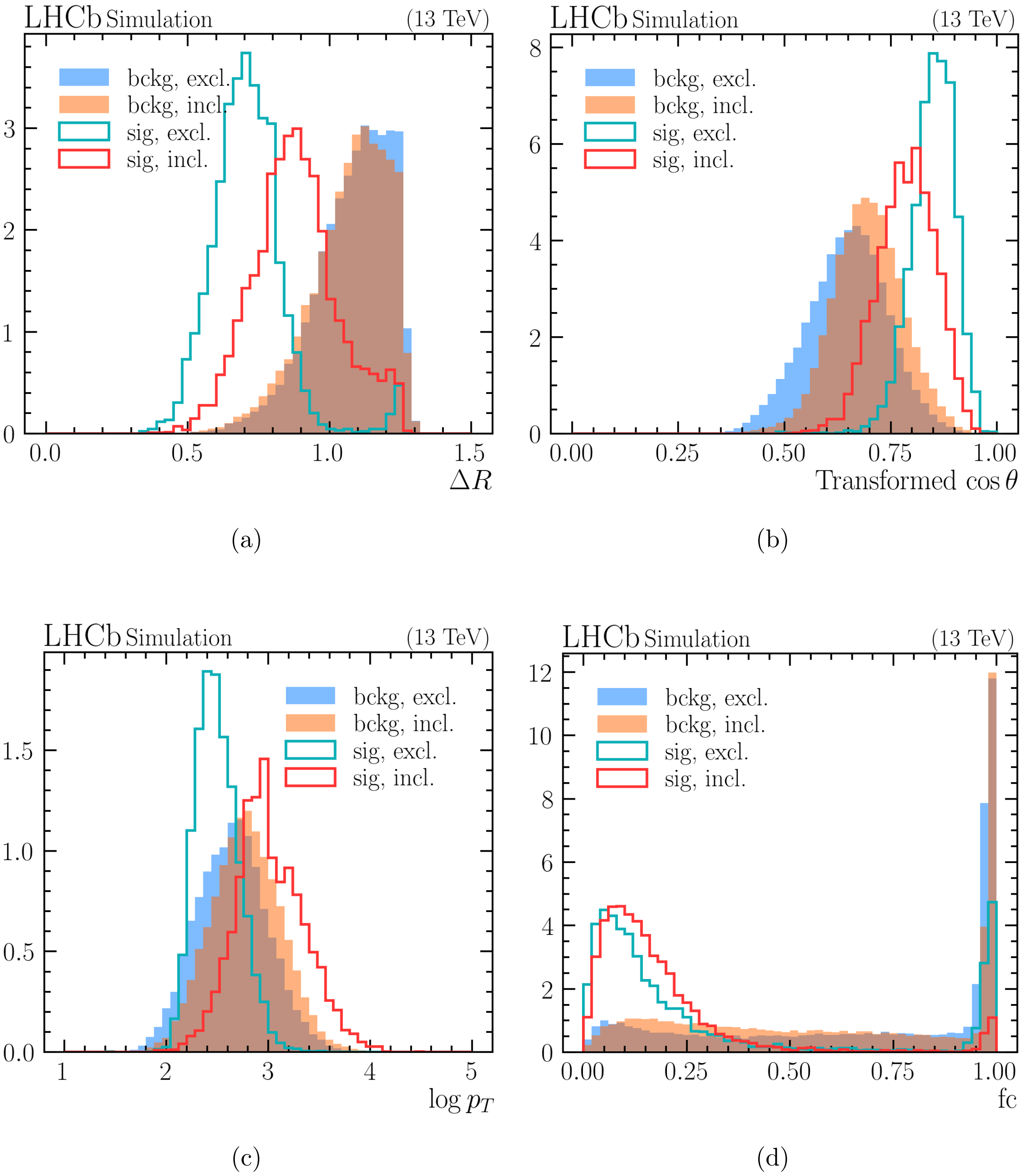


Figure 16: Signal and background distributions for (a) the cone ΔR, (b) the transformed cosine of the opening angle $\cos\theta$, (c) the transverse momentum $\log p_T$, and (d) the variable fc. The exlcusive decay mode and the inclusive sample are overlaid; the brown part represents the overlap between the backgrounds of both samples. The distributions are for the $B_u \to D^*(2010)\tau^+$ decay.

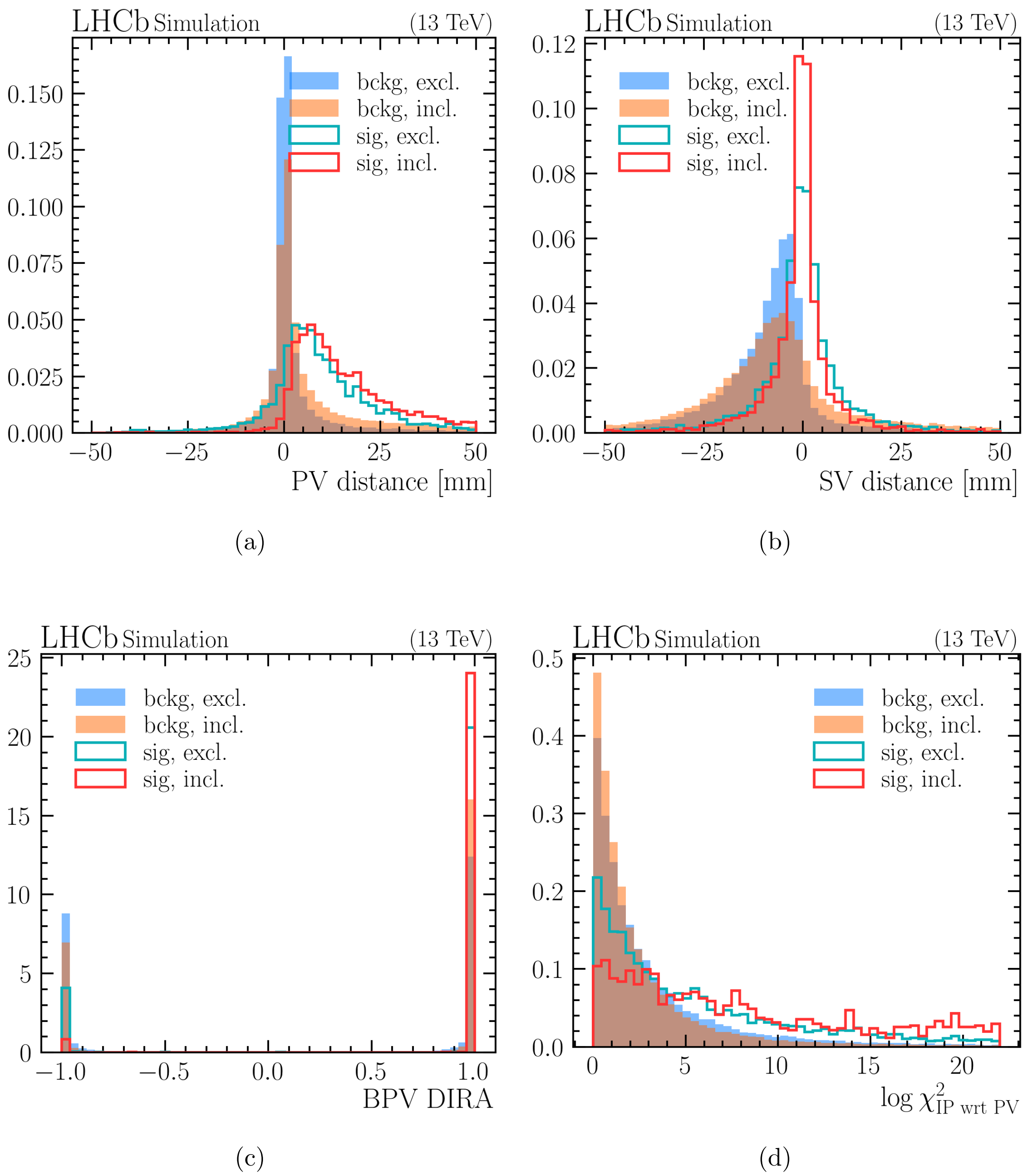
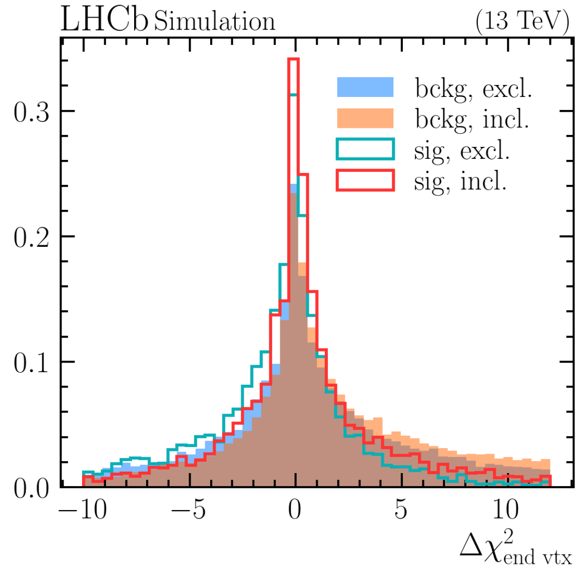
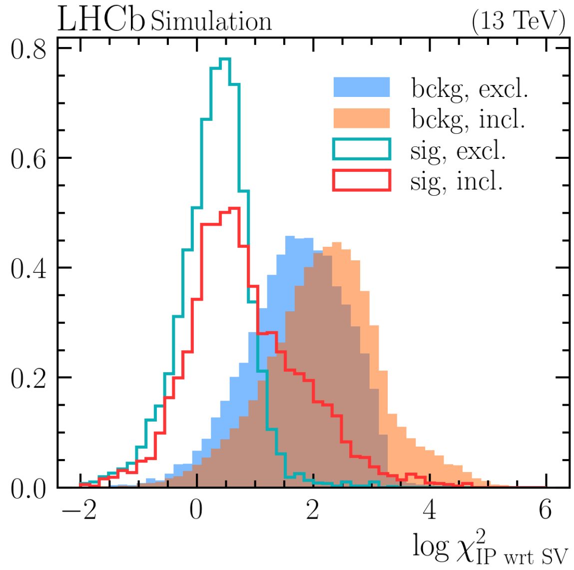


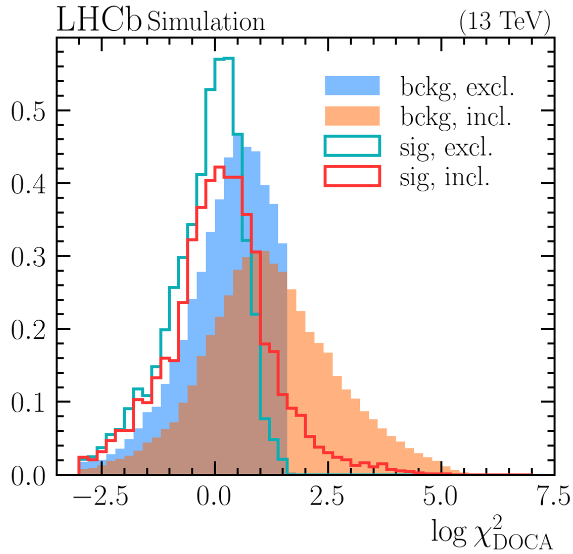
Figure 17: Signal and background distributions for (a) the PV distance, (b) the SV distance, (c) the BPV DIRA, and (d) the significance of the IP with respect to the PV $\log \chi^2_{\mathrm{IP\ wrt\ PV}}$. The exlcusive decay mode and the inclusive sample are overlaid; the brown part represents the overlap between the backgrounds of both samples. The distributions are for the $B_u \to D^*(2010)\tau^+$ decay.



(a)



(b)



(c)

Figure 18: Signal and background distributions for (a) difference in candidate end vertex significance $\Delta\chi^2_{\text{end vtx}}$, (b) the significance of the IP with respect to the SV $\log\chi^2_{\text{IP wrt SV}}$, and (c) the DOCA significance $\log\chi^2_{\text{DOCA}}$. The exlcusive decay mode and the inclusive sample are overlaid; the brown part represents the overlap between the backgrounds of both samples. The distributions are for the $B_u \to D^*(2010)\tau^+$ decay.

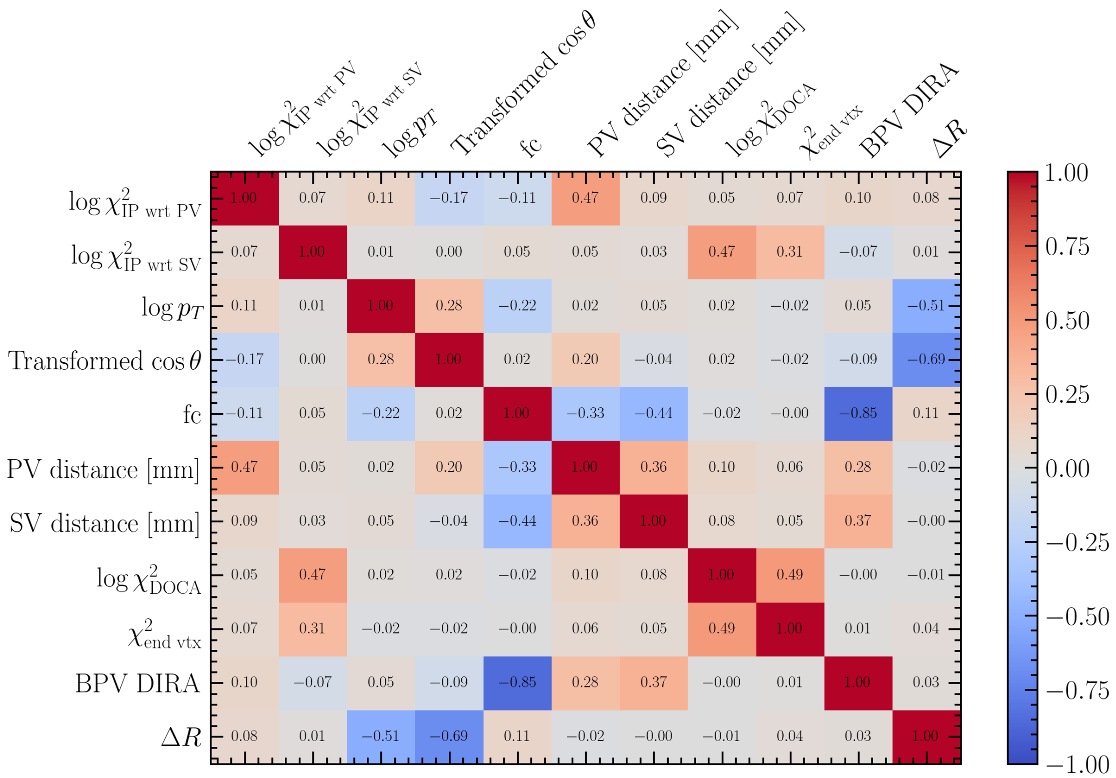


Figure 19: Correlation plot of the variables for the signal tracks for the $B_u \to D^*(2010)\tau^+$ decay.

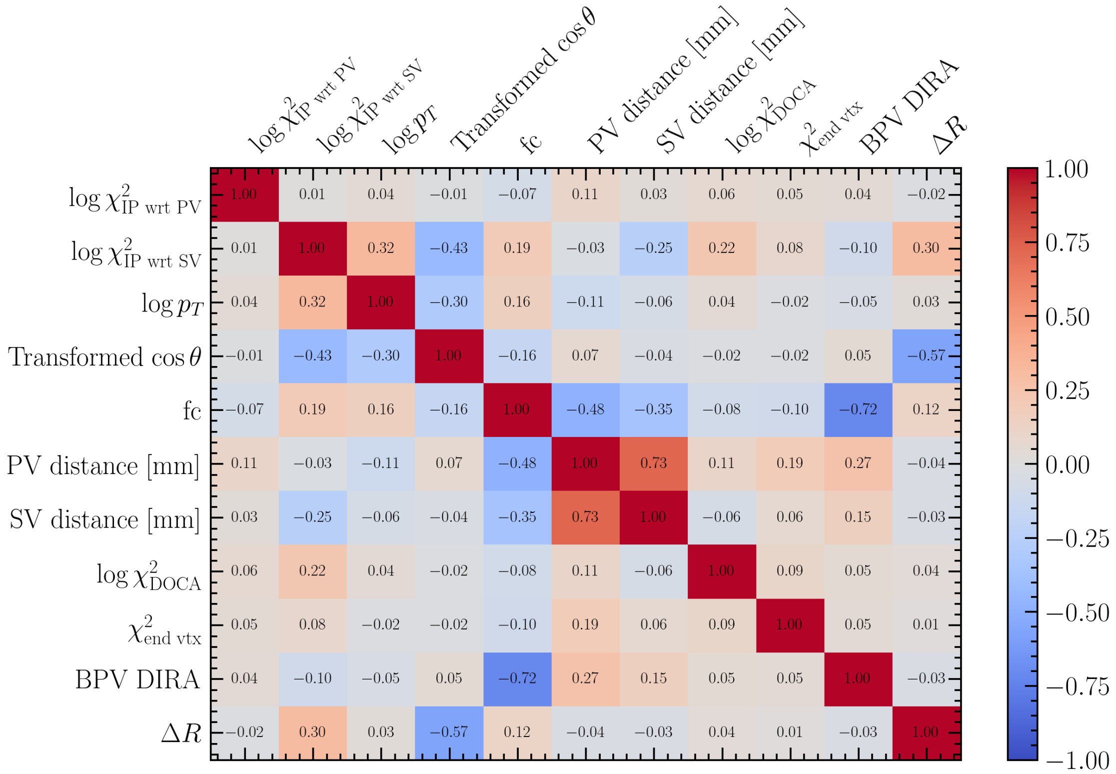


Figure 20: Correlation plot of the variables for the background tracks for the $B_u \to D^*(2010)\tau^+$ decay.

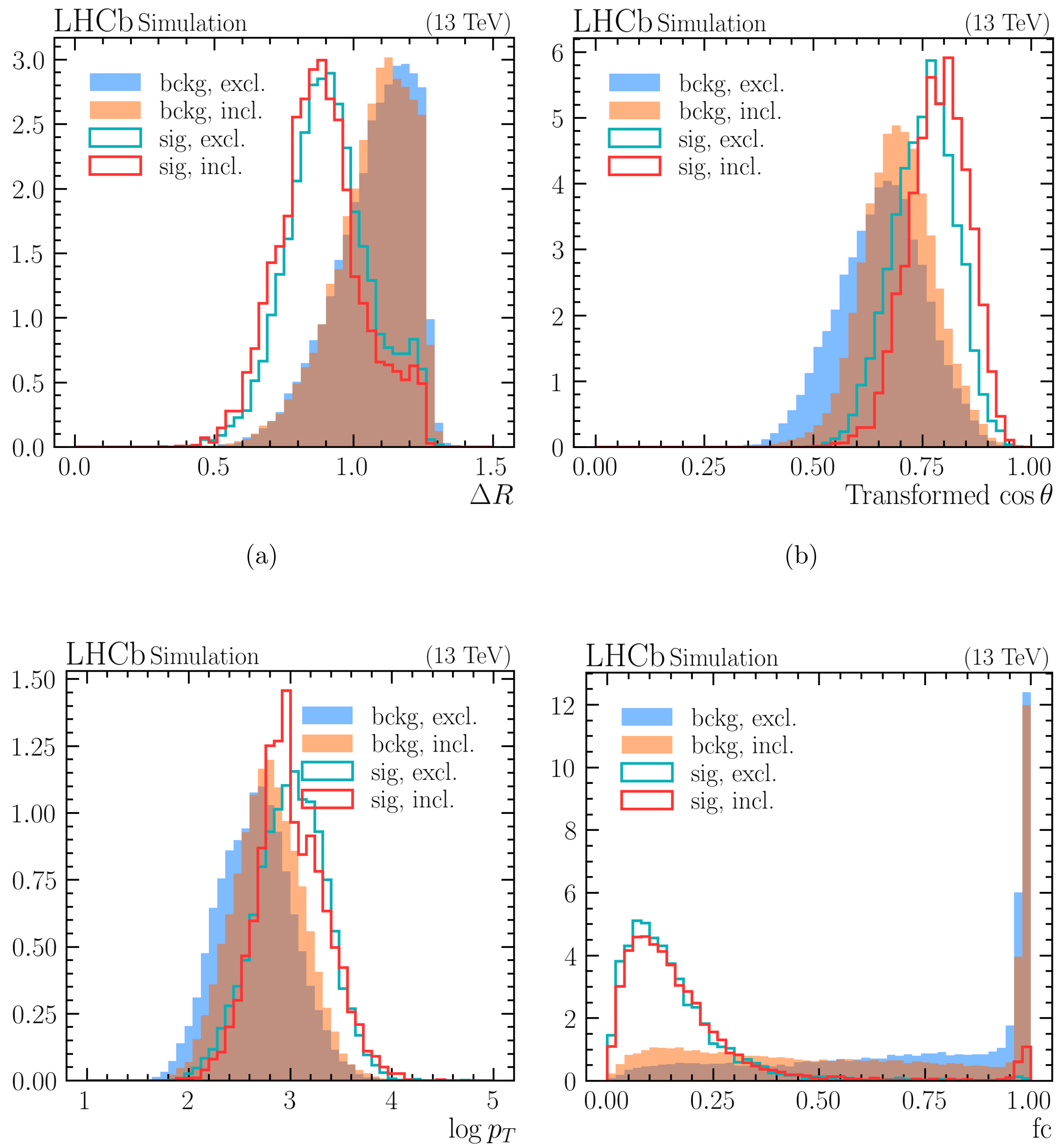


Figure 21: Signal and background distributions for (a) the cone ΔR, (b) the transformed cosine of the opening angle $\cos\theta$, (c) the transverse momentum $\log p_T$, and (d) the variable fc. The exlcusive decay mode and the inclusive sample are overlaid; the brown part represents the overlap between the backgrounds of both samples. The distributions are shown for the $\Lambda_b \to p\mu$ decay.

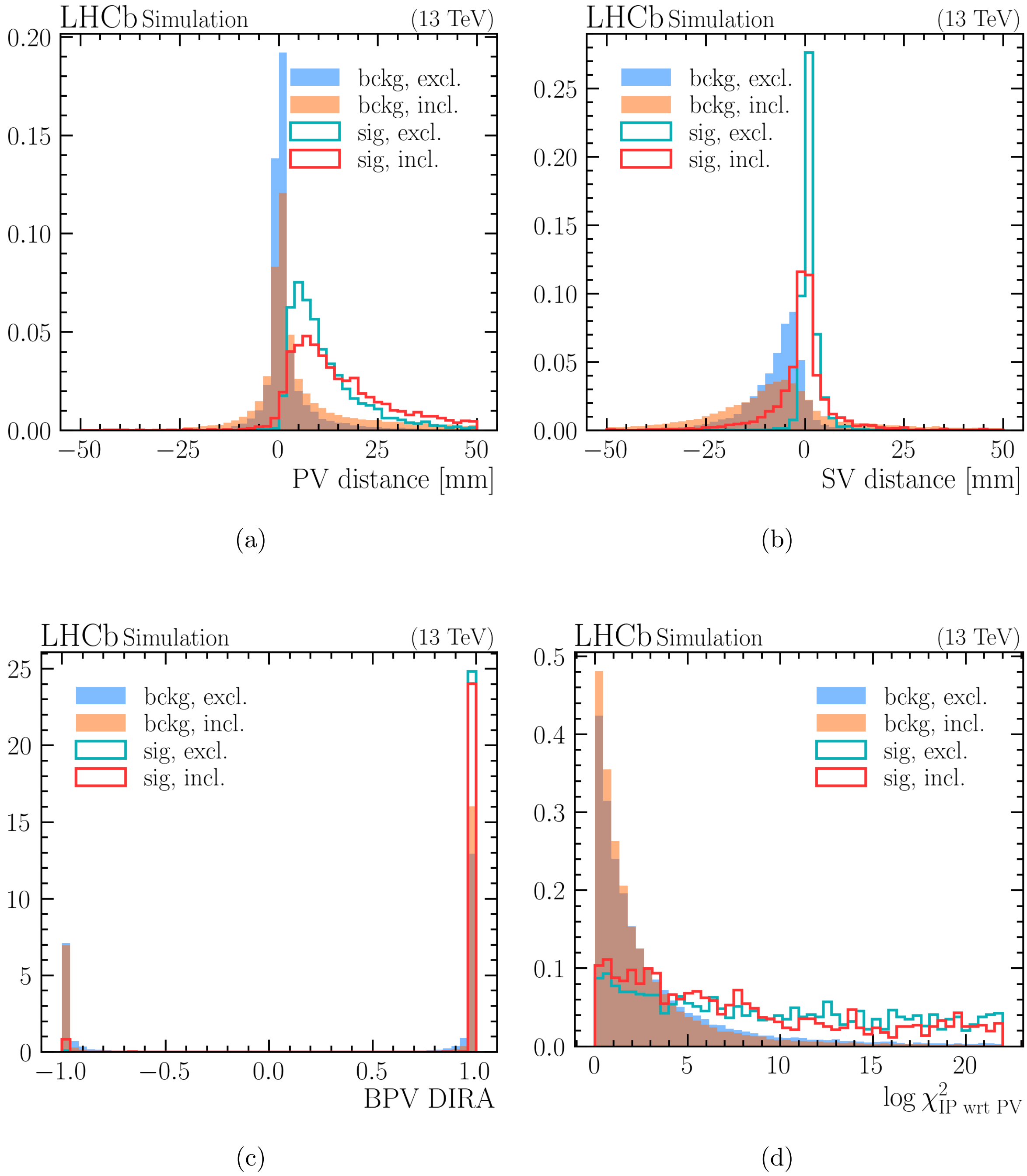
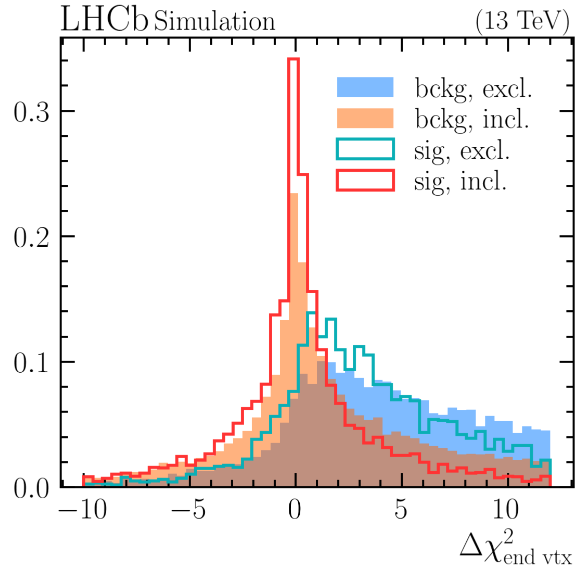
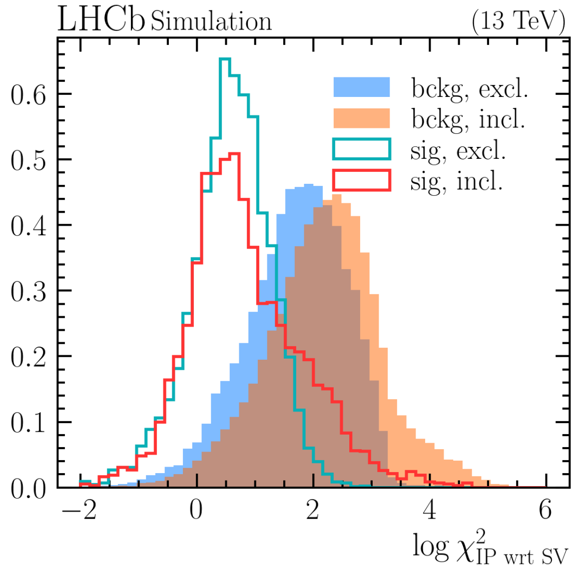


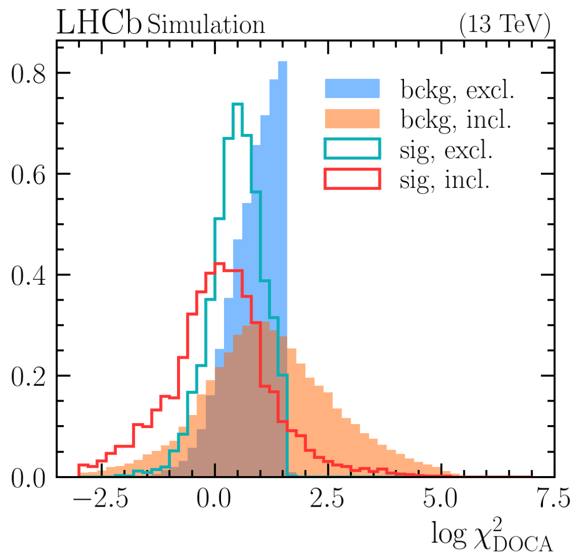
Figure 22: Signal and background distributions for (a) the PV distance, (b) the SV distance, (c) the BPV DIRA, and (d) the significance of the IP with respect to the PV $\log \chi^2_{\mathrm{IP\ wrt\ PV}}$. The exlcusive decay mode and the inclusive sample are overlaid; the brown part represents the overlap between the backgrounds of both samples. The distributions are shown for the $\Lambda_b \to p\mu$ decay.



(a)



(b)



(c)

Figure 23: Signal and background distributions for (a) difference in candidate end vertex significance $\Delta\chi^2_{\text{end vtx}}$, (b) the significance of the IP with respect to the SV $\log\chi^2_{\text{IP wrt SV}}$, and (c) the DOCA significance $\log\chi^2_{\text{DOCA}}$. The exlcusive decay mode and the inclusive sample are overlaid; the brown part represents the overlap between the backgrounds of both samples. The distributions are shown for the $\Lambda_b \to p\mu$ decay.

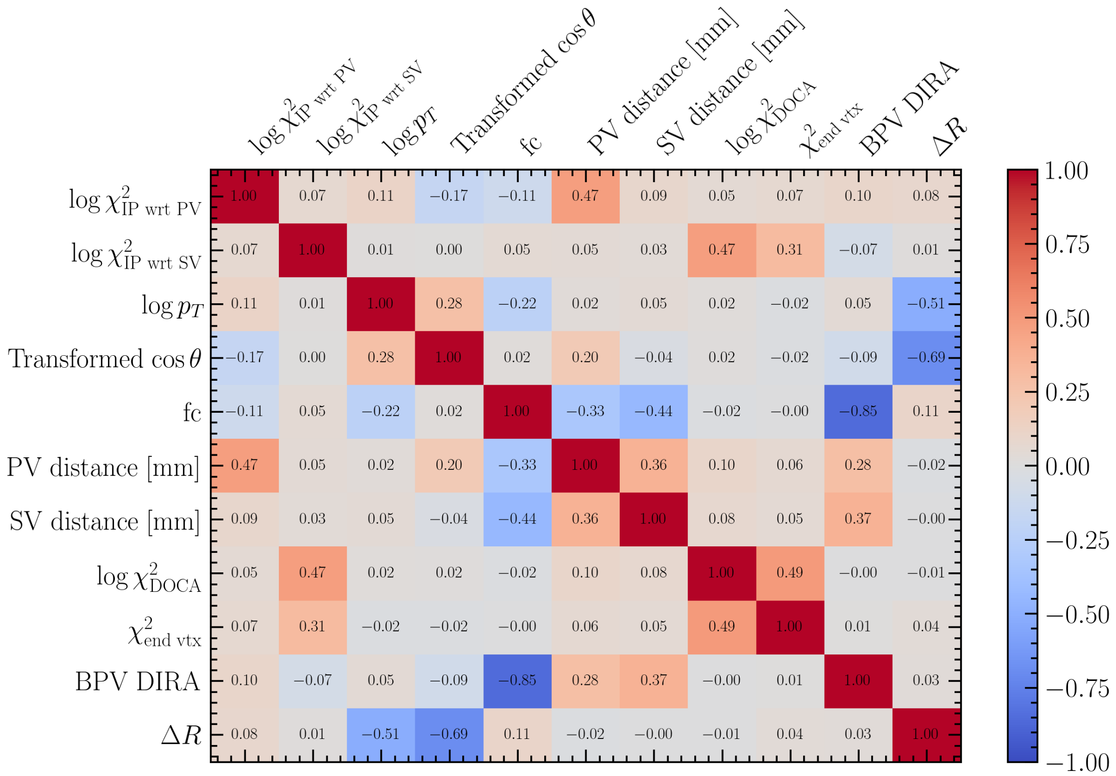


Figure 24: Correlation plot of the variables for the signal tracks for the $\Lambda_b \to p\mu$ decay.

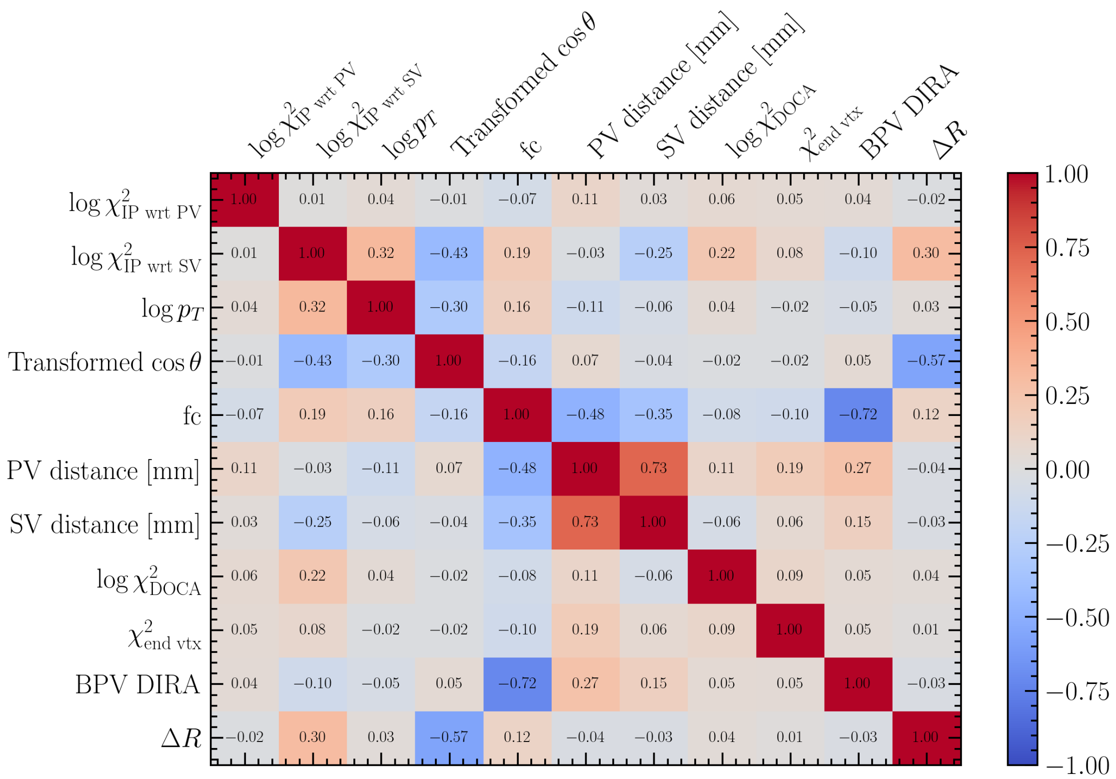


Figure 25: Correlation plot of the variables for the background tracks for the $\Lambda_b \to p\mu$ decay.

References

[1] LHCb collaboration, R. Aaij *et al.*, *The LHCb upgrade I*, arXiv:2305.10515.

[2] LHCb collaboration, R. Aaij *et al.*, *Measurement of the $B_s^0 \to \mu^+\mu^-$ branching fraction and effective lifetime and search for $B^0 \to \mu^+\mu^-$ decays*, Phys. Rev. Lett. **118** (2017) 191801, arXiv:1703.05747.

[3] LHCb collaboration, R. Aaij *et al.*, *Measurement of the shape of the $\Lambda_b^0 \to \Lambda_c^+ \mu^- \overline{\nu}_\mu$ differential decay rate*, Phys. Rev. **D96** (2017) 112005, arXiv:1709.01920.

[4] LHCb collaboration, R. Aaij *et al.*, *Measurement of the ratio of branching fractions $\mathcal{R}(D^*)$ and $\mathcal{R}(D^0)$* , arXiv:2302.02886, submitted to Phys. Rev. Lett.

[5] R. Aaij *et al.*, *A comprehensive real-time analysis model at the LHCb experiment*, JINST **14** (2019) P04006, arXiv:1903.01360, 15 pages, 4 figures, 1 table. All figures and tables are available at https://cern.ch/lhcbproject/Publications/LHCbProjectPublic/LHCb-DP-2019-002.html.

[6] LHCb collaboration, R. Aaij *et al.*, *Measurement of the ratio of branching fractions $\mathcal{B}(\overline{B}^0 \to D^{*+}\tau^-\overline{\nu}_\tau)/\mathcal{B}(\overline{B}^0 \to D^{*+}\mu^-\overline{\nu}_\mu)$*, Phys. Rev. Lett. **115** (2015) 111803, Publisher's Note ibid. **115** (2015) 159901, arXiv:1506.08614.

[7] Y. Coadou, *Boosted decision trees*, doi: $10.1142/9789811234033_0002$ arXiv:2206.09645.

[8] T. Chen and C. Guestrin, *Xgboost: A scalable tree boosting system*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), 785–794, ACM, 2016.

[9] R. H. C. Lopes, in *Kolmogorov-Smirnov Test*, M. Lovric, ed., pp. 718–720, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[10] T. Akiba *et al.*, *Optuna: A next-generation hyperparameter optimization framework*, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[11] L. Gavardi, *Search for lepton flavour violation in τ decays at the LHCb experiment*, 2013. Presented 28 Nov 2013.