

Comparison of Jet-Flavor Tagging in full and fast Simulation at CLD

Sara Aumiller^{1*}, Dolores Garcia^{2*} and Michele Selvaggi^{2*}

¹Technical University of Munich (TUM), Munich, Germany.

²CERN, Geneva, Switzerland.

*Corresponding author(s). E-mail(s): sara.aumiller@cern.ch;
dolores.garcia@cern.ch; michele.selvaggi@cern.ch;

1 Introduction

At future Higgs factories like FCC-ee, we are particularly interested in precision measurements of the Higgs bosons such as the determination of the branching ratios to fermions and their couplings. Jet-flavor tagging is indispensable to perform these measurements.

A jet can be viewed as an ensemble of particles originating from the same source. The source can either be a quark, gluon, or even a hadronically decaying τ lepton, which hadronizes and creates a narrow cone of traceable particles in the detector. Determining the particle type of the jet origin is referred to as jet-flavor tagging.

We introduce jet-flavor tagging at CLD, a proposed general-purpose detector at FCC-ee. This is the first study of tagging on full simulation at FCC-ee. Tagging on fast simulation using DELPHES [1] has already been performed on the IDEA detector [2] which can derive particle identification (PID) using a cluster counting method in contrast to CLD. For a meaningful comparison between fast and full simulation, we have modified the extensively tested DELPHES card normally used for IDEA in order to include a silicon track detector with the geometry and resolutions as in the CLD full simulation. This setup will be referred to as "CLD fast simulation" in the following. We compare the performances of tagging in fast and full simulation at CLD and discuss the bottlenecks of the tagger, as well as potential solutions to enhance its performance.

2 Jet tagging observables in fast and full simulation

2.1 Track reconstruction in fast and full simulation

As tracking is responsible for reconstructing charged particles, it is of great importance for jet flavor tagging. Therefore, we summarize the track reconstruction methods used in fast and full simulation at FCC-ee.

DELPHES [1] performs a fast simulation of the tracking performance for detectors at FCC-ee [2]. It describes two tracking system geometries: cylinders coaxial to the beam direction and planar disks orthogonal to the beam direction. Strip, wire and pixel measurement geometries are included in the framework. Each layer of the tracking system geometry describes either a measurement or accounts as passive material contributing to multiple scattering. As tracking is performed in a magnetic field, charged particles move on a helix trajectory. Therefore, the track can be described with five parameters $\vec{\alpha}$. To reconstruct the track parameters $\vec{\alpha}$ from measured coordinates \vec{d}^* , a χ^2 minimization with respect to the track parameters $\vec{\alpha}$ is performed:

$$\chi^2 = (\vec{d} - \vec{d}^*)^t S^{-1} (\vec{d}^* - \vec{d}) \quad (1)$$

where \vec{d} are the predicted coordinates derived from $\vec{\alpha}$ and the geometry of each measurement layer. S is the covariance matrix of all measurements including detector resolutions and multiple scattering [2].

Full simulation at CLD uses conformal tracking [3]. This method combines two strategies: conformal mapping and cellular automaton-based track finding. Conformal mapping transforms point coordinates in Euclidean space into the conformal space. This is useful as circles passing through the origin in Euclidean space turn into straight lines in conformal space. Therefore, the problem of helix fitting is simplified to finding straight lines in this chosen space. But in real physics processes, there are deviations from a perfect path e.g. through multiple scattering. As displaced particles do not pass through the origin, their straight line can also only be mathematically approximated. Therefore, cellular automaton is used for pattern recognition in conformal space consisting of two steps: building and extending cellular track candidates.

Full simulation relies on very accurate track reconstruction as the particle flow algorithm pandora [4] is used for reconstructing whole physics events.

2.2 Comparison of jet observables in fast and full simulation for $H \rightarrow b\bar{b}$

To compare the jet description in fast and full simulation, we use $e^+e^- \rightarrow ZH$ events at 240 GeV with an invisible Z decay into two neutrinos $Z \rightarrow \nu\bar{\nu}$. We choose to investigate the dominant Higgs decay channel $H \rightarrow b\bar{b}$.

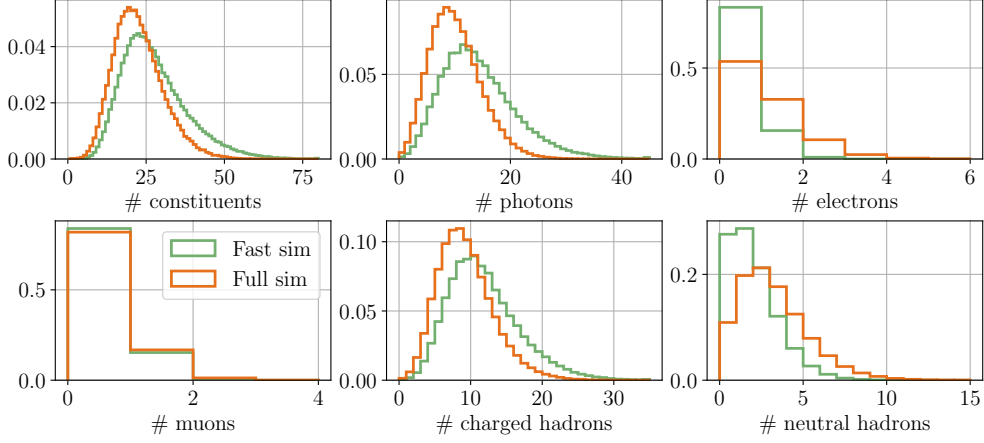


Fig. 1: Distributions of jet-constituent multiplicities in fast (green) and full (orange) simulation using 250k $H \rightarrow b\bar{b}$ jets.

Figure 1 shows the multiplicity of all particles and specific particle types in jets in fast and full simulation. Full simulation shows an overall lower multiplicity of particles in jets. We want to highlight that there are more neutral but less charged hadrons in full simulation. We are particularly interested in comparing the observables of these jet constituents as they define the characteristics of a jet. We can describe them with kinematic observables, their identification and, if charged, with track characteristics. Most of the compared jet observables in our study are in good agreement. As an example, we show the longitudinal impact parameter z_0 for the leading order charged particles in Figure 2a. The distributions including the tails match. Other track displacement observables for $H \rightarrow u\bar{u}$, $H \rightarrow c\bar{c}$ and $H \rightarrow g\bar{g}$ comparing fast and full simulation at CLD are shown in the Appendix, Figure 12.

We note that the description of tracks in full simulation is still missing the shift of the primary vertex to the true interaction point. We manually apply the shift to all parameters of the helix but the relative angles and the covariance matrix as done in fast simulation.

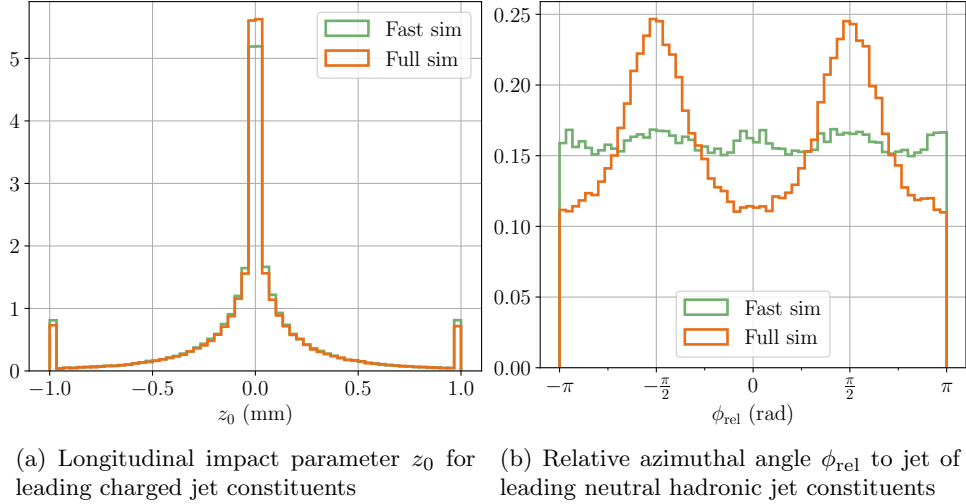


Fig. 2: Distributions of jet constituent observables in full (orange) and fast (green) simulation using 250k $H \rightarrow b\bar{b}$ jets.

Two major problems arise in full simulation compared to fast simulation: fake neutrals, see Section 2.2.1 and lost tracks, see Section 2.2.2.

2.2.1 Fake neutrals in full simulation

The main difference observed through analyzing the distributions of jet-constituent observables is the creation of fake neutral hadrons in full simulation. It becomes apparent when examining the relative azimuthal angle ϕ between the jet and its components. The distribution of ϕ_{rel} is expected to be flat, as we observe in fast simulation and for charged jet constituents in full simulation. The distribution of ϕ_{rel} for neutral particles in full simulation is not flat but shows smeared out peaks around $\pm \frac{\pi}{2}$, see Figure 2b. It can be mathematically derived that if the jet and its constituents have similar angles in θ and ϕ , then ϕ_{rel} approaches $\pm \frac{\pi}{2}$. High energetic charged particles leave large showers in the calorimeter. Such a large shower might get wrongly split into two, one high energetic shower associated to the track and one less energetic shower. The less energetic shower without track will be falsely reconstructed as a neutral particle. The problem is sketched in Figure 3. A fake neutral will have similar angles as the original charged particle. As high energetic charged particles often dominate the jet kinematics, the jet and the fake neutral have similar angles too. In this case $\phi_{\text{rel}} \rightarrow \pm \frac{\pi}{2}$ which we observe in the data. Other analysis seem to face the same problem of fake neutral creation in pandora [5].

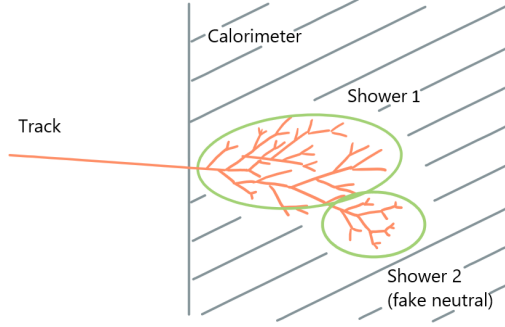


Fig. 3: Sketch of fake neutral reconstruction. A high energetic charged particles leaves a track and a large shower in the calorimeter. The shower is split in two of which one has no track associated and is therefore reconstructed as a fake neutral.

2.2.2 Lost tracks in full simulation

Particle Flow Objects (PFOs) used in reconstruction algorithms like pandora [4] match tracking information and clustering information from calorimetry to form reconstructed particles. Neutral particles only contain calorimetric shower information as they do not leave tracks while charged particles are build from tracking and shower (cluster) information. A possible bottleneck for the performance of the network in full simulation are lost charged particles. As tracks are more relevant for jet-flavor tagging due to the unique track displacements for different flavors, we might loose important information. Therefore, we investigate the performance of reconstructing charged PFOs in full simulation.

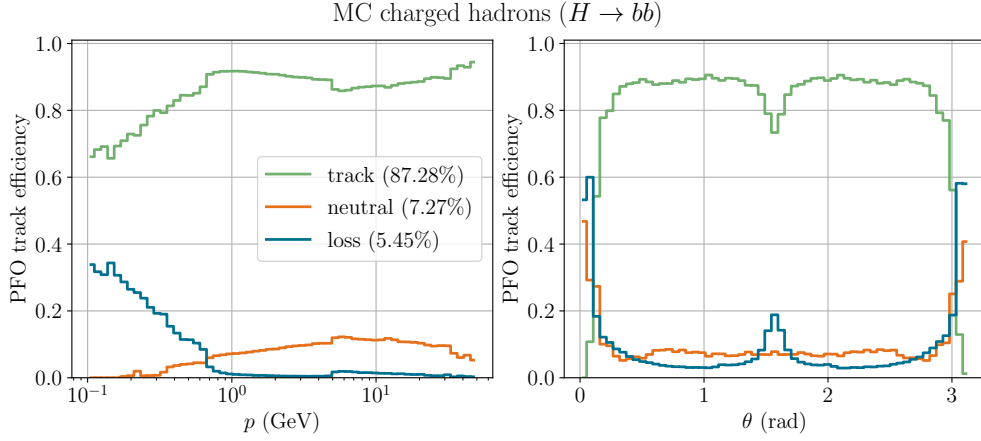


Fig. 4: MC charged hadrons from full simulation $H \rightarrow b\bar{b}$ events shown as a fraction of which are reconstructed with track (green), reconstructed as neutral (orange) or not reconstructed at all (blue) as a function of MC momentum p (left) and MC polar angle θ (right).

Considering all charged MC particles that leave at least four tracker hits, we check whether there is a reconstructed particle associated that uses at least 30 % of the MC tracker hits or MC calorimeter cluster hits. If there is no such reconstructed particle, we mark it as lost. Otherwise we check if an associated track got reconstructed using at least 30 % of the MC tracker hits. If so, we mark the particle as reconstructed with track, otherwise it is marked as neutral. The results of that study is exemplary shown in Figure 4 for charged hadrons of the $H \rightarrow b\bar{b}$ channel. We see the fraction of MC charged hadrons which got reconstructed with track information (green), reconstructed without track information (orange) and not reconstructed (blue) as a function of MC momentum p (left) and MC polar angle θ (right). The efficiency of reconstructing tracks rises with increasing momentum until 1 GeV while less tracks get lost. There is a clear drop in the performance at $p = 5$ GeV which originates from the pandora reconstruction. Above 5 GeV the algorithm requires a track to point towards a cluster to be used to reconstruct a charged particle as it assumes that the energy of the particle is high enough to reach the calorimeter. The additional constraint causes MC charged particles to get lost or wrongly reconstructed as neutrals because unmatched tracks get lost. Furthermore, we see a loss of charged particles perpendicular to the beam line. These are particles that move on spiraling paths in the magnetic field never reaching the calorimeter. These are probably marked as lost as they leave many hits in the tracker which are not all used for reconstruction and therefore do not fulfill the selection requirement of 30 % used tracker hits.

2.3 Comparison of jet observables for all Higgs channels in full simulation

To check whether other Higgs decay channels than $H \rightarrow b\bar{b}$ in full simulation behave as expected, we compare all observables of $H \rightarrow u\bar{u}$, $H \rightarrow d\bar{d}$, $H \rightarrow c\bar{c}$, $H \rightarrow s\bar{s}$, $H \rightarrow gg$, $H \rightarrow \tau\bar{\tau}$ to $H \rightarrow b\bar{b}$. All channels demonstrate a consistent agreement with each other. Exemplary, we show the significance of the signed IP in 2D in Figure 5. The observed asymmetry from the left to right in the case of b and c quarks originates from secondary vertices of short lived mesons and is the main discriminant for b and c jet compared to other flavors.

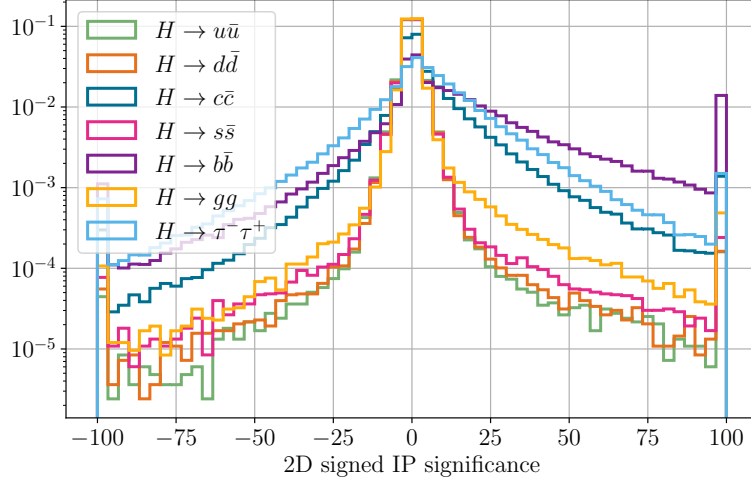


Fig. 5: Distribution of track displacement parameter for different Higgs decay channels using 250k jets in CLD full simulation.

3 Jet-flavor tagging at CLD in fast and full simulation using deep learning

Due to the complex nature of hadronisation, deep learning has become the standard approach for jet-flavor tagging. We are using the state-of-the-art neural network Particle Transformer (ParT) [6] for jet-flavor tagging at CLD. This classification problem has seven labels representing seven different decay channels of the Higgs boson: $H \rightarrow u\bar{u}$, $H \rightarrow d\bar{d}$, $H \rightarrow c\bar{c}$, $H \rightarrow s\bar{s}$, $H \rightarrow b\bar{b}$, $H \rightarrow gg$ and $H \rightarrow \tau\bar{\tau}$.

3.1 Input parameter to the network

Jet-flavor tagging at FCC-ee is already explored at IDEA in fast simulation [2]. Therefore, we begin by training the ParT with the same input variables used at IDEA describing different properties of the jet constituents such as kinematic variables, track properties for charged particles and PID. If particles are neutral and therefore do not

have tracks, the track variables are set to a dummy value lying outside the distribution. We must point out that compared to IDEA, CLD has no time-of-flight (ToF) and no cluster counting (dN/dx) information. An overview of the parameters used for the CLD fast and full simulation training can be seen in Table 1.

| Variable type | Description | Number of parameters |
|---------------------|--|----------------------|
| Kinematics | $\log E_{\text{rel}}, \theta_{\text{rel}}, \phi_{\text{rel}}$ | 3 |
| Identification | reco PID, charge, PID flags | 7 |
| Track displacements | d_0, z_0 , covariance matrix c_{ij} SIP in 2D and 3D (& significance), jet-track distance $d_{3\text{D}}$ (& significance) | 23 |

Table 1: Overview of the 33 input parameters used for training ParT. For a thorough description see [2].

3.2 Results and comparison

We train both, fast and full simulation, on the input parameters described in Table 1 performing jet-based classification. The dataset contains 14 million jets (2 million jets per Higgs decay channel) of which we use 95 % for training the network and 5 % for evaluation. The results of the tagging performance and comparison between fast and full simulation can be seen in Figure 6. The receiver operating characteristic (ROC) curves show the jet misidentification probability (false positive rate) in logarithmic scale against the jet tagging efficiency (true positive rate). The closer the curves to the right lower corner, the better the results. Fast simulation outperforms full simulation as expected in all cases. Picking c -tagging as an example, the efficiency of c vs. ud drops from over 80 % to under 70 % at a misidentification probability of 10^{-2} .

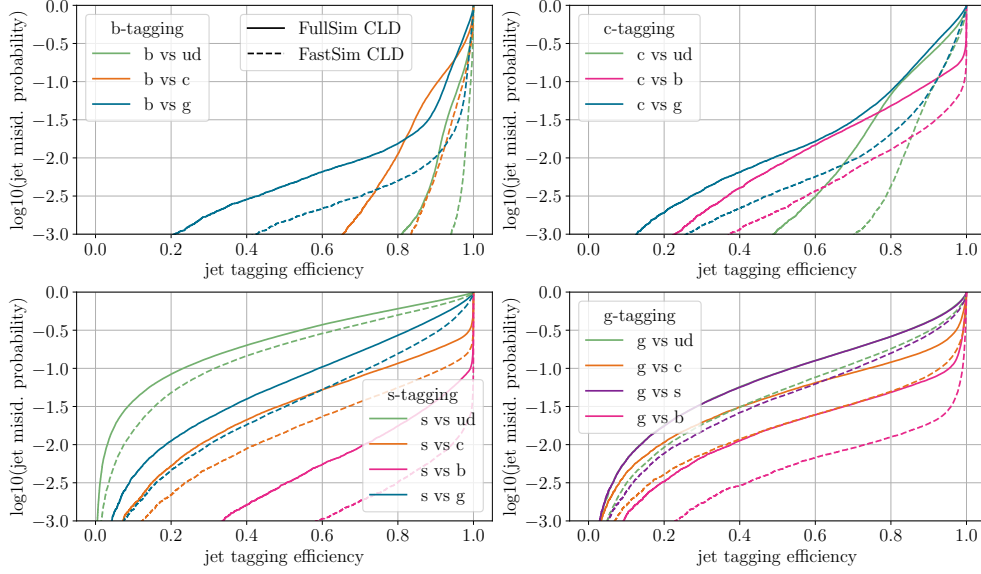


Fig. 6: ROC curves for b -, c -, s - and g -tagging (from upper left to lower right). The jet misidentification probability for other particles vs. the tagged particle is shown as a function of efficiency. The solid (dashed) line shows full (fast) simulation CLD results at 240 GeV. The closer the curves to the lower-right corner the better the results.

3.3 Results with complete track information in full simulation

To investigate whether the missing tracks (via loss or neutral PFO reconstruction) described in Section 2.2.2 have an effect on the jet-flavor tagging results, we consider an alternative approach to retrieve information on the jet content than considering PFOs. We retrieve charged particles by considering all reconstructed tracks and match them via their polar angle to the two jets in the events. Neutral particles are retrieved by considering neutral PFOs in the jets while checking their MC PID to be neutral to not double count charged particles. With this method we also avoid the problem of fake neutral particles described in Section 2.2.1.

The results (dashed lines) and comparison to the standard full simulation CLD tagging performance (solid lines) are shown in Figure 7. The ROC curves suggest that using complete track information leads to a significant improvement in the tagging performance. Taking b -tagging as an example, the tagging efficiency for b vs. ud quarks for a 10^{-2} misidentification probability rises from 90 % to 95 %.

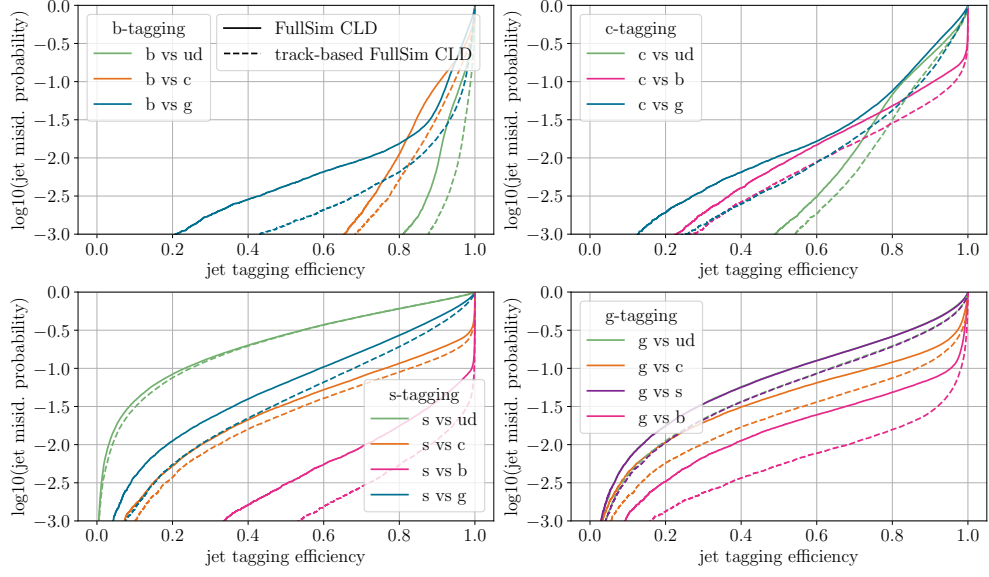


Fig. 7: ROC curves for b -, c -, s - and g -tagging (from upper left to lower right). The jet misidentification probability for other particles vs. the tagged particle is shown as a function of efficiency. The solid line shows full simulation CLD results at 240 GeV using PFOs. The dashed line corrects the reconstruction by considering all reconstructed tracks instead of charged PFOs and tests the PID of neutral PFOs to avoid double counting.

In the Appendix, we show the ROC curves for fast simulation vs. this track-based full simulation approach in Figure 13. In Figure 8, we show an alternative illustration of the tagging performance comparing the track-based full simulation performance to fast simulation: the non-binary discriminates as $\log \frac{p_i}{1-p_i}$ with p_i being the probabilities for a class $i \in ud, s, c, b, g, \tau$. ud refers to the average probability of u and d . The more separated the distributions of the classes to the distribution of the class of interest, the better the result. Some more uncommon discriminants for s -, u - and τ -tagging can be seen in the Appendix, Figure 14 as well as some uncommon ROC curves showing u -, d - and τ -tagging in the Appendix, Figure 15

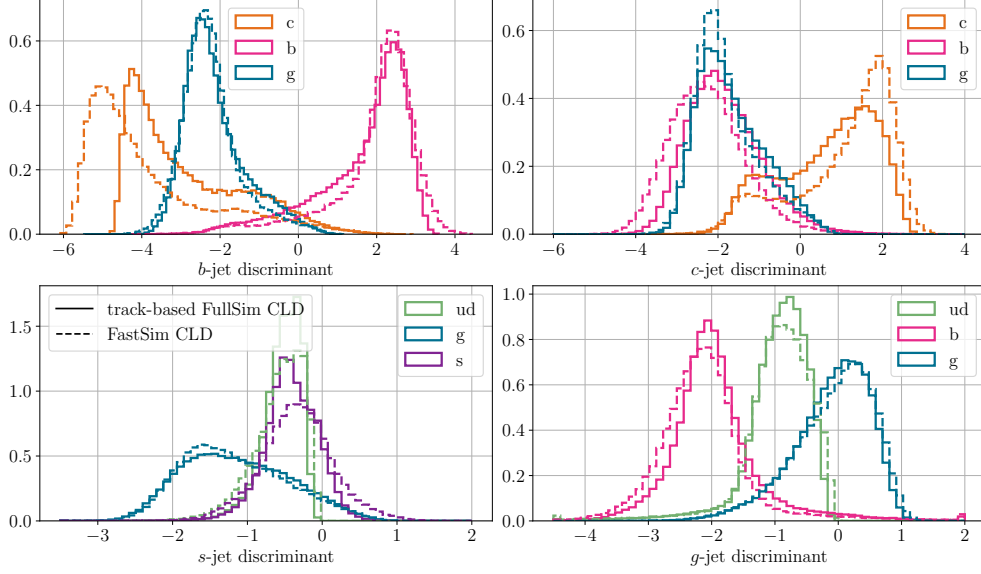


Fig. 8: Tagging performance of track-based full simulation CLD (solid line) vs. fast simulation CLD (dashed line) showing the discriminates as $\log \frac{p_i}{1-p_i}$.

3.4 Comparison of CLD vs. IDEA jet-flavor tagging in fast simulation

The IDEA detector [7] is an other proposed detector at FCC-ee. Jet-flavor tagging at IDEA in fast simulation was already explored using ParticleNet [8]. Unlike CLD, IDEA has the possibility to perform particle identification which is especially important for s -tagging. We perform an equivalent training to CLD on IDEA with the only difference that we provide time of flight (ToF) and dN/dx as an additional input to the network. While the performance on b - and g -tagging are comparable, the CLD performance worsens on s -tagging and on c vs. ud as expected without PID information. The non-binary discriminates as $\log \frac{p_i}{1-p_i}$ are shown in the Appendix, Figure 16, where we see that the s -tagging discriminates are more entangled at CLD than at IDEA leading to worse separation between ud - and s -jets. s -jets are non distinguishable from first generation jets without PID information of the hadronized baryons and mesons carrying the s quarks information e.g. as kaons or lambdas.

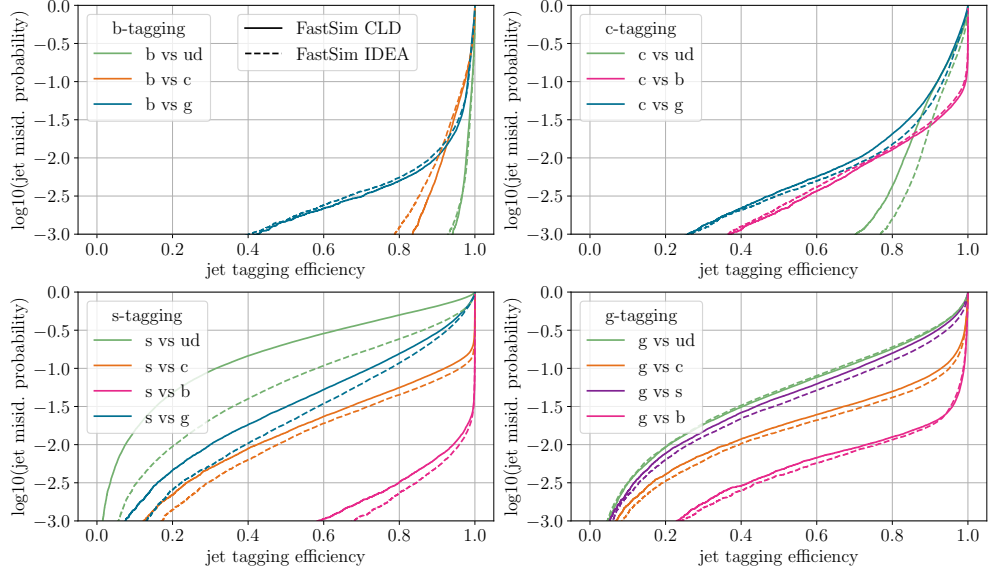


Fig. 9: ROC curves for b -, c -, s - and g -tagging (from upper left to lower right). The jet misidentification probability for other particles vs. the tagged particle is shown as a function of efficiency. The solid (dashed) line shows fast simulation CLD (IDEA) results at 240 GeV. CLD fast simulation refers to the IDEA detector concept with a silicon tracker.

3.5 Results with secondary vertex information

Until now, we have only considered basic track information for tagging. Although CLD does not have the capabilities to perform PID like IDEA, we can provide vertex information retrieved from tracking algorithms. Therefore, we retrain the tagger on full simulation CLD data but with added vertex position information (x, y, z) and invariant mass of the vertices. We consider V^0 s and secondary vertices for this purpose. The distribution of the invariant mass for V^0 and tranverse radius for secondary vertices can be seen in the Appendix, Figure 17. The V^0 invariant mass distribution shows a clear peak at the K^0 mass of 497.6 MeV. The radius distribution of secondary vertices shows more displaced tracks for b -jets characterized through the longer tail similarly to Figure 5. The first two tracker layers leave noise in the distribution.

The performance does not improve using vertex information, see Figure 10. The performance fluctuates around the standard full simulation CLD performance instead. This indicates that the network can learn vertexing on its own through track displacements and track kinematics so that the added vertex information does not add valuable information to the network.

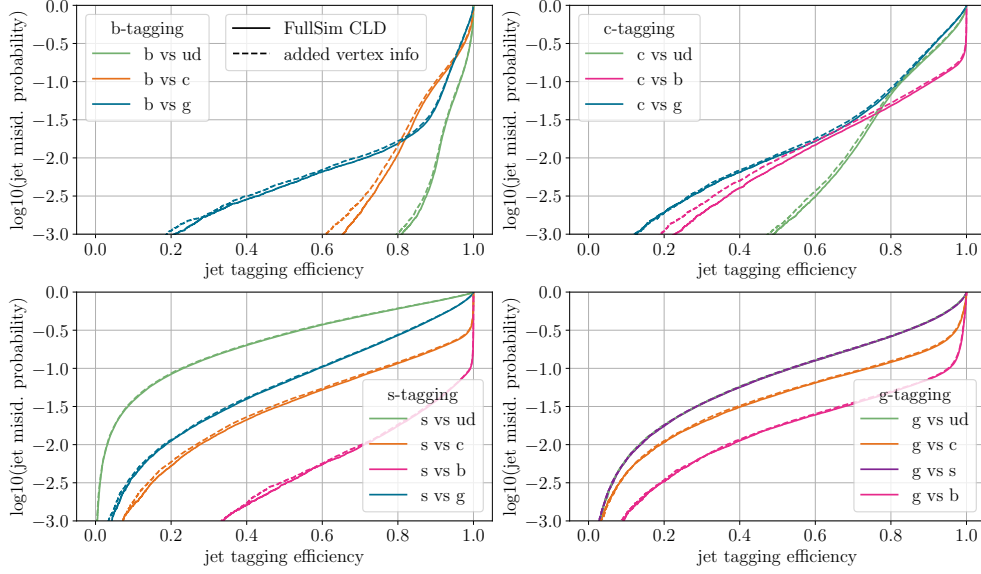


Fig. 10: Tagging performance of CLD full simulation at 250 GeV without (solid line) and with (dashed line) vertex information (invariant mass and coordinates of V^0 s and secondary vertices).

3.6 Influence of calorimetric energy resolution on jet-tagging

The CLD-like fast simulation that we are using for comparison to CLD full simulation is a modified IDEA DELPHES card. The IDEA DELPHES card has been extensively used and debugged in the past. To simulate a CLD-like detector, we exchange the tracking system to a silicon track detector with the same geometry and resolutions as in the CLD full simulation. This setup might raise the question of how much influence the energy resolution of the unchanged calorimetric system has on the tagging results. Therefore, we train an other modified version of the IDEA detector with worse HCAL resolution simulating the CMS detector at LHC using the exact same training setup. The results can be seen in Figure 11. The difference of the tagging performance due to the worse HCAL resolution seems negligible, only a small deviation on s vs. ud is visible.

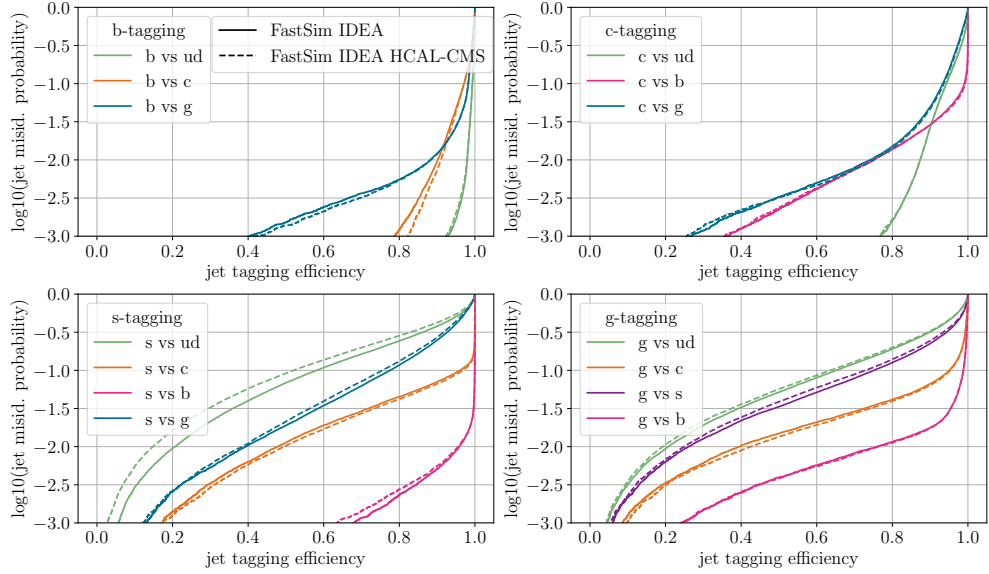


Fig. 11: Tagging performance of IDEA fast simulation at 240 GeV with dual readout calorimeter (solid line) and CMS HCAL resolution (dashed line).

4 Summary and Discussion

We present the first study of jet-flavor tagging on full simulation at CLD at 240 GeV. When comparing the results to a CLD-like detector in fast simulation, the first results in full simulation do not yet achieve the same level of performance. Full simulation jet-flavor tagging improves if tracks are used instead of charged PFOs as mistakes in the reconstruction chain such as lost tracks or fake neutral particles are avoided. This encourages future work on the CLD reconstruction for more precise physics analysis.

We have also compared the tagging performance between CLD and IDEA in fast simulation, showing the influence of PID information on s -tagging. As adding secondary vertex information to CLD full simulation does not improve the tagging performance, we conclude that the neural network learns a basic level of vertexing on its own via the other provided track observables. Using a modified IDEA detector in fast simulation, we have shown that the calorimetric energy resolution has a negligible effect on jet flavor tagging.

Although CLD has no dedicated detector for Time-of-Flight (ToF) or dN/dx measurements, dE/dx information could be retrieved from the silicon trackers. In the past, these measurements have not been considered sufficiently precise for PID but might provide valuable information for tagging with machine learning methods. Leveraging dE/dx might further push the limits of jet-flavor tagging at CLD. The CLD-community might envisage a RICH detector in the future which could provide similar PID capabilities as dN/dx at IDEA. Although s -tagging will improve with this setup, the overall

tagging performance might suffer due to more detector material influencing the crucial particle flow performance.

Acknowledgements

This work has been partly supported by the Future Circular Collider Innovation Study (FCCIS) project, that has received funding from the European Union’s Horizon 2020 research and innovation programme under grant No 951754.

References

- [1] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, M. Selvaggi, Delphes 3: a modular framework for fast simulation of a generic collider experiment. *Journal of High Energy Physics* **2014**(2) (2014). [https://doi.org/10.1007/jhep02\(2014\)057](https://doi.org/10.1007/jhep02(2014)057). URL [http://dx.doi.org/10.1007/JHEP02\(2014\)057](http://dx.doi.org/10.1007/JHEP02(2014)057)
- [2] F. Bedeschi, L. Gouskos, M. Selvaggi, Jet flavour tagging for future colliders with fast simulation. *The European Physical Journal C* **82**(7) (2022). <https://doi.org/10.1140/epjc/s10052-022-10609-1>. URL <http://dx.doi.org/10.1140/epjc/s10052-022-10609-1>
- [3] E. Brondolin, E. Leogrande, D. Hynds, F. Gaede, M. Petrič, A. Sailer, R. Simoniello, Conformal tracking for all-silicon trackers at future electron–positron colliders. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **956**, 163304 (2020). <https://doi.org/10.1016/j.nima.2019.163304>. URL <http://dx.doi.org/10.1016/j.nima.2019.163304>
- [4] J.S. Marshall, M.A. Thomson. The pandora particle flow algorithm (2013). URL <https://arxiv.org/abs/1308.4537>
- [5] D. Jeans, K. Yumino. Ild benchmark: a study of $e^-e^+ \rightarrow \tau^-\tau^+$ at 500 gev (2020). URL <https://arxiv.org/abs/1912.08403>
- [6] H. Qu, C. Li, S. Qian. Particle transformer for jet tagging (2024). URL <https://arxiv.org/abs/2202.03772>
- [7] F. Bedeschi, A detector concept proposal for a circular e+e- collider. *PoS ICHEP2020*, 819 (2021). <https://doi.org/10.22323/1.390.0819>
- [8] H. Qu, L. Gouskos, Jet tagging via particle clouds. *Physical Review D* **101**(5) (2020). <https://doi.org/10.1103/physrevd.101.056019>. URL <http://dx.doi.org/10.1103/PhysRevD.101.056019>

Appendix

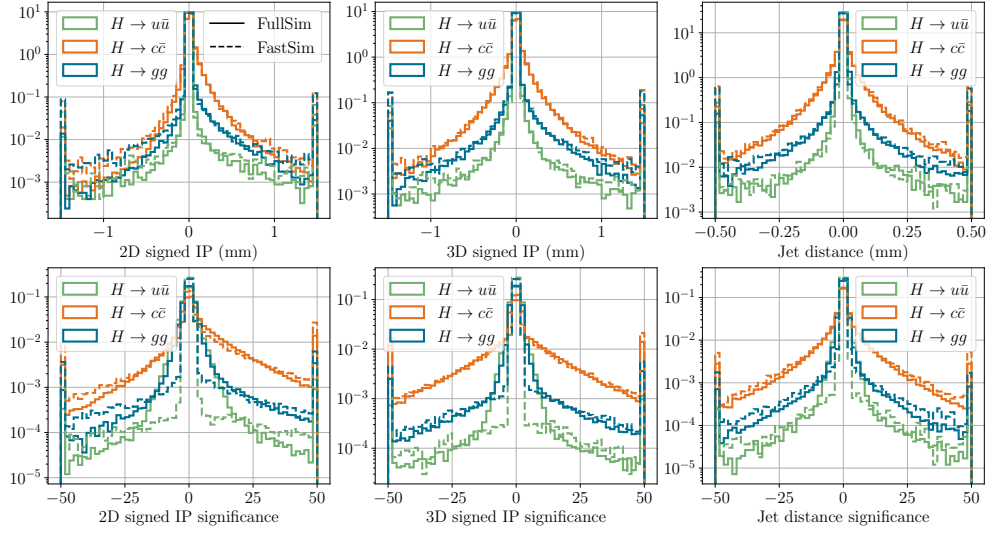


Fig. 12: 2D and 3D signed impact parameters, jet distance and their significances from leading tracks of $H \rightarrow u\bar{u}$ (green), $H \rightarrow c\bar{c}$ (orange) and $H \rightarrow g\bar{g}$ (blue) in full (fast) simulation shown as a solid (dashed) line.

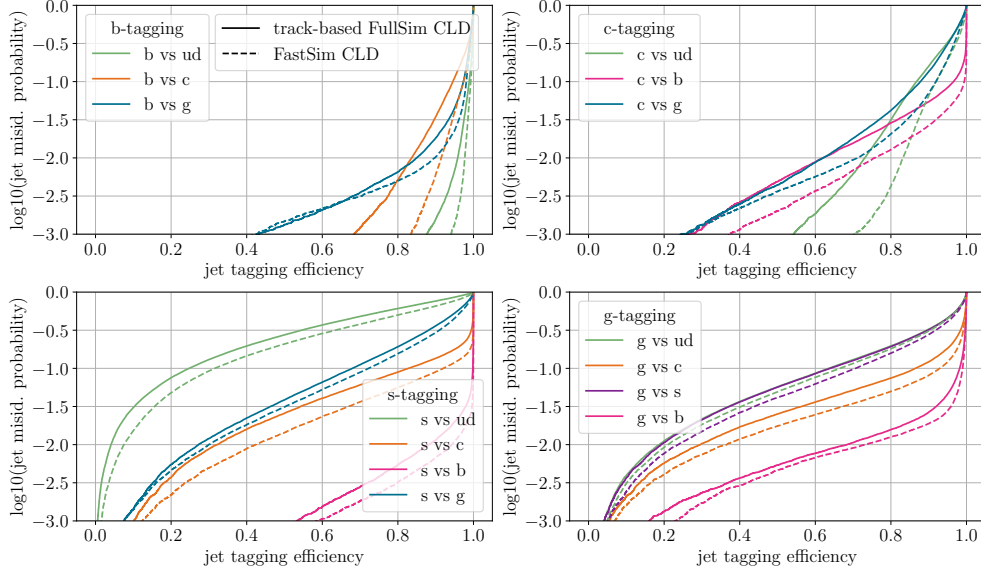


Fig. 13: ROC curves for b -, c -, s - and g -tagging (from upper left to lower right). The jet misidentification probability for other particles vs. the tagged particle is shown as a function of efficiency. The dashed line shows fast simulation CLD results at 240 GeV where fast simulation CLD refers to the modified IDEA detector with a silicon tracker. The solid line shows track-based full simulation CLD results which correct the reconstruction by considering all reconstructed tracks instead of charged PFOs and tests the PID of neutral PFOs to avoid double counting.

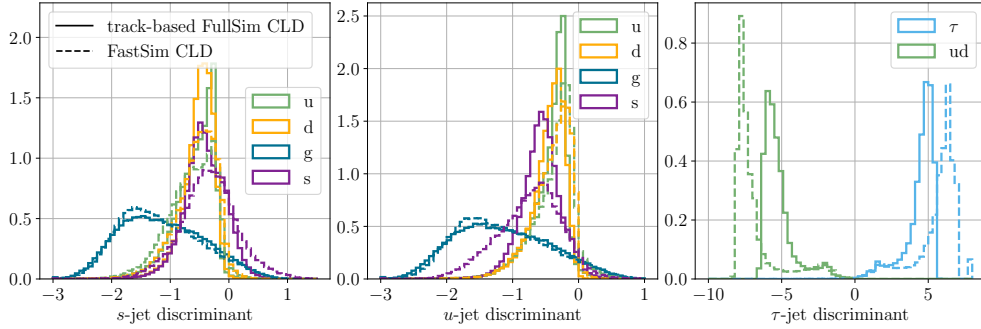


Fig. 14: Tagging performance of track-based full simulation CLD (solid line) vs. fast simulation CLD (dashed line) showing the discriminates as $\log \frac{p_i}{1-p_i}$.

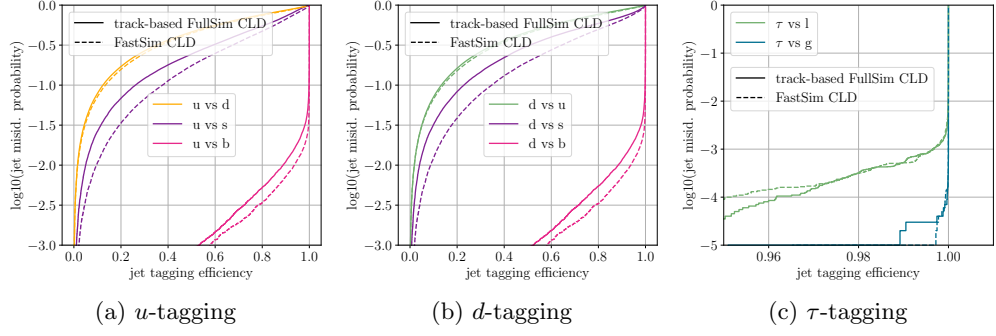


Fig. 15: ROC curves for different classes (u , d , τ from left to right) for track-based full simulation CLD (solid line) and fast simulation CLD (dashed line). The jet misidentification probability for other particles vs. the tagged particle is shown as a function of efficiency.

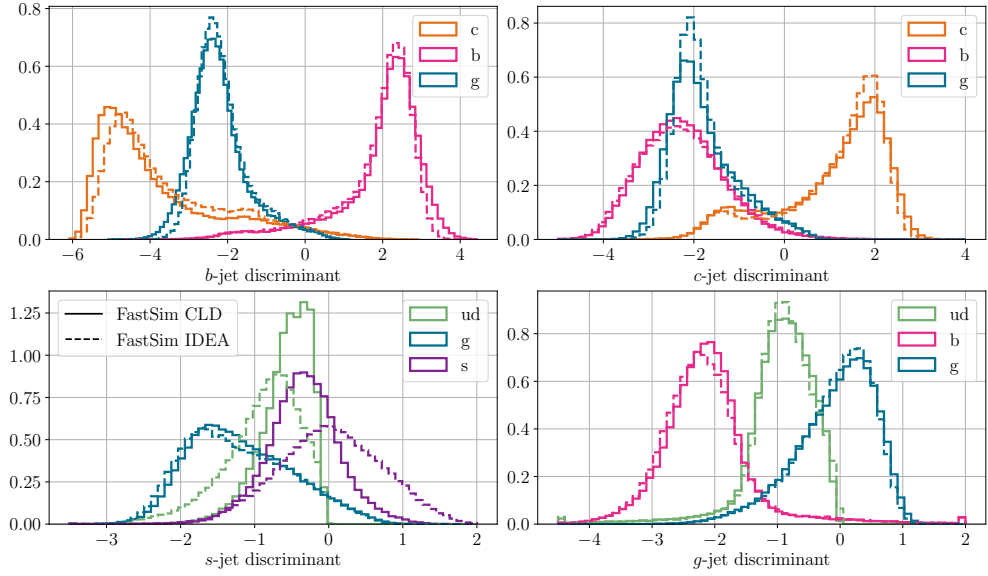


Fig. 16: Tagging performance of fast simulation CLD (solid line) vs. fast simulation IDEA (dashed line) showing the discriminates as $\log \frac{p_i}{1-p_i}$.

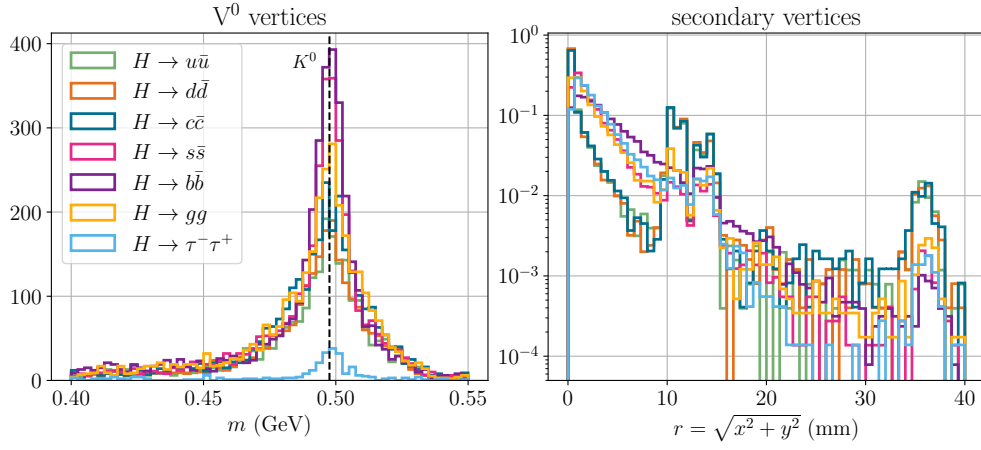


Fig. 17: Invariant-mass spectrum of V^0 (left) and transverse radius of secondary vertices (right) for different Higgs decay channels (250k jets per channel). PDG masses of relevant particles are indicated with a black dashed line.