Eur. Phys. J. C (2025) 85:165
https://doi.org/10.1140/epjc/s10052-025-13785-y

THE EUROPEAN
PHYSICAL JOURNAL C

Regular Article - Experimental Physics

Tagging more quark jet flavours at FCC-ee at 91 GeV with a transformer-based neural network

Freya Blekman[1,2,4], Florencia Canelli[3], Alexandre De Moor[1], Kunal Gautam[1,3,a], Armin Ilg[3], Anna Macchiolo[3], Eduardo Ploerer[1,3,b]

[1] Inter-university Institute for High Energies, Vrije Universiteit Brussel, 1050 Brussels, Belgium
[2] Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany
[3] Universität Zürich, Winterthurerstr. 190, 8057 Zurich, Switzerland
[4] Universität Hamburg, Luruper Chaussee 149, 22761 Hamburg, Germany

Received: 15 July 2024 / Accepted: 2 January 2025

Abstract Jet flavour tagging is crucial in experimental high-energy physics. A tagging algorithm, `DeepJetTransformer`, is presented, which exploits a transformer-based neural network that is substantially faster to train than state-of-the-art graph neural networks. The `DeepJetTransformer` algorithm uses information from particle flow-style objects and secondary vertex reconstruction for b- and c-jet identification, supplemented by additional information that is not always included in tagging algorithms at the LHC, such as reconstructed K_S^0 and Λ^0 and $K^\pm/\pi^\pm$ discrimination. The model is trained as a multiclassifier to identify all quark flavours separately and performs excellently in identifying b- and c-jets. An s-tagging efficiency of 40% can be achieved with a 10% ud-jet background efficiency. The performance improvement achieved by including K_S^0 and Λ^0 reconstruction and $K^\pm/\pi^\pm$ discrimination is presented. The algorithm is applied on exclusive $Z \rightarrow q\bar{q}$ samples to examine the physics potential and is shown to isolate $Z \rightarrow s\bar{s}$ events. Assuming all non-$Z \rightarrow q\bar{q}$ backgrounds can be efficiently rejected, a 5σ discovery significance for $Z \rightarrow s\bar{s}$ can be achieved with an integrated luminosity of 60 nb^{-1} of e^+e^- collisions at $\sqrt{s} = 91.2$ GeV, corresponding to less than a second of the FCC-ee run plan at the Z boson resonance.

1 Introduction

The Standard Model (SM) of particle physics [1–4] is one of the most successful scientific theories describing the fundamental particles and their interactions. The last piece of this model, the Higgs boson, was discovered [5,6] at the Large Hadron Collider (LHC) [7] in 2012, and the precise study of its properties will remain mostly superficial at the LHC due to high irreducible backgrounds from other SM processes while isolating Higgs boson events.

One of the main motivations for proposed future lepton colliders [8–11] is the precise measurement of SM parameters, like precision studies of the hadronic decay of the Z boson and greatly improved sensitivity to the couplings of the Higgs boson to the bottom (b) and charm (c) quarks and gluons (g) [12–14]. Achieving these objectives requires an efficient reconstruction and identification of the hadronic decays of these particles. The feasibility of studying the decay of the Higgs boson to the strange (s), up (u), and down (d) quarks depends on the collider and detector performance and is currently under investigation in the field. It is well established that efficient and accurate jet flavour identification is essential to exploit the maximal physics potential of future collider experiments [15–19].

The state-of-the art is shortly reviewed in the rest of Sect. 1. Section 2 summarises the FCC-ee collider, the IDEA detector concept, and the used simulated samples and provides minimal event selection requirements. Section 3 briefly describes the algorithms used to reconstruct displaced decay vertices and their performance. Section 4 introduces the attention mechanism and Transformer models and outlines the description of the input features and the network architecture used for tagging. Finally, the obtained results and the performance of the flavour-tagging algorithm in Z boson signatures are presented in Sects. 5 and 6, respectively.

[a] e-mail: kunal.gautam@cern.ch (corresponding author)
[b] e-mail: eduardo.ploerer@cern.ch

Published online: 10 February 2025

1.1 Review of jet flavour tagging

Jets originating from the b and c quarks contain hadrons with significant lifetimes that travel distances of the order of millimeters from the interaction point before decaying into lighter hadrons. The heavy flavour tagging algorithms used at the Large Electron-Positron collider (LEP) [20,21] and the Tevatron [22,23] experiments exploited variables derived from the displaced charged tracks originating from these decayed b or c hadrons to distinguish the heavy flavoured jets from s, u, d quark and gluon jets. These charged tracks are commonly clustered to reconstruct the original decay vertices of the b and c hadrons, also called secondary vertices (SVs). The properties of these SVs, like their mass and displacement, are some of the most important inputs used to identify b- and c-jets.

The understanding and performance of jet flavour tagging at the LHC has steadily been improving and heavily relies on machine learning (ML) [24,25], which also inspires flavour tagging algorithms for the FCC-ee [26,27].

ML approaches are uniquely suited to classify jet flavours, where training samples are abundant in the form of Monte Carlo (MC) simulation. Still, the underlying dynamics of jet formation and hadronisation are not always well understood. With the advent of ML techniques, including Neural Networks (NNs) and Boosted Decision Trees (BDTs), approaches relying on single physics-motivated variables for jet flavour discrimination were significantly outperformed [25,28–30]. Since then, a multitude of architectures and jet representations have found success in discriminating jet flavours, including Dense Neural Networks (DNNs) [31], Recurrent Neural Networks (RNNs) [32], Convolutional Neural Networks (CNNs) [33,34], and Graph Neural Networks (GNNs) [26,35,36].

Among the most successful of these are Graph-based architectures such as `ParticleNet` [26] that represent jets as sets of nodes (jet constituents) and edges (some pairwise defined feature, often the difference in a given variable of jet constituents). In particular, networks combining a self-attention mechanism [37] to exploit the relative importance of constructed features, dubbed Transformer Networks, have achieved state-of-the-art performance in the task of jet flavour tagging [30,38–40]. Particle Transformer (ParT) [40] combines a graph representation of jets with an attention mechanism. In this work, a pure Transformer architecture, `DeepJetTransformer`, similar to the ParT (plain) variant introduced in Ref. [40], is presented for the task of jet flavour identification at future lepton colliders, using the FCC-ee with the IDEA detector concept as a benchmark [8,41]. `DeepJetTransformer` is relatively lightweight and requires much less computational time compared to graph-based architectures [40,42,43], yet achieves comparable tagging performance.

1.2 The Z boson at the FCC-ee

After the discovery of the Z boson at the Super Proton Synchrotron (SPS) at CERN in 1983 [44,45], this neutral vector boson was extensively studied at the LEP collider and the SLAC Linear Collider. The existence of the Z boson confirmed the electroweak mixing [46,47] and the measurement of its width constrained the number of neutrino generations to three [48–52].

The proposed FCC-ee program provides a unique opportunity to push the Z boson measurements to their ultimate limit. The four-year-long FCC-ee run at and around the Z resonance will produce an unprecedented 6×10^{12} total decays. The integrated luminosity expected at the Z resonance at FCC-ee is 125 ab^{-1}, about 10^6 times that of LEP. The statistical errors on the mass and width of the Z boson can be reduced from 1.2 MeV and 2 MeV to 5 KeV and 8 KeV [8], respectively. Lower center-of-mass energy spread due to beam energy calibration will benefit in reducing the systematic uncertainty of these quantities. Measuring the forward-backward and polarisation asymmetries is a powerful method to estimate the effective weak mixing angle, $\sin^2\theta_{\mathrm{W}}^{\mathrm{eff}}$, for which the statistical uncertainty is expected to reduce to about 10^{-6}, corresponding a more than thirty-fold improvement [8].

Studying the hadronic decay channels of the Z boson is a very important aspect of the FCC-ee physics program. The couplings and decay widths of the Z boson have only been measured to the heavier quarks, b and c. The only study of the s quark decay of the Z boson available in the literature is preliminary [53]. For the lighter quarks, s, u, and d, these properties are typically only listed collectively for up-type and down-type quarks [54]. Similarly, the axial and vector couplings have also been collectively measured for up-type and down-type quarks [54].

Future colliders with a dedicated Z boson run, like FCC-ee, will improve the precision of all these measurements and make the s quark, and potentially the u, and d quarks, accessible. Individual measurements of the quark vector and axial couplings should be possible via their forward-backward asymmetries, corresponding partial decay widths of the Z boson, and the precise knowledge of A_e, the asymmetry parameter of the e^-e^+ pair. The experimental systematic uncertainties corresponding to these measurements are also expected to drastically improve due to better detector designs [8].

1.3 Strange jet tagging

The discrimination of s-jets is widely regarded as one of the most challenging types of jet discrimination. Thus, it has received considerably less attention than its heavy-flavour counterparts, or indeed gluon discrimination. At the core of the problem is the fact that unlike in the discrimination of quarks vs gluons, which relies heavily on properties following from their differing colour factors $C_F = 4/3$ vs $C_A = 3$, or heavy flavour tagging, which relies on displaced vertices of b/c hadrons, strange quarks are treated identically to down quarks by QCD and Electroweak theory in the massless limit prior to their decay. Discriminating strange and down jets is particularly challenging due to the same fractional charge of the initiating quarks. In practice, however, strange hadrons carry a larger fraction of the total scalar momentum of strange jets, compared to hadrons consisting of up and down (ud) quarks. The total scalar momentum is obtained by summing over the scalar momentum of all jet constituents. This idea was also explored in the context of hadron colliders [55]. Strange jets tend to have a higher kaon multiplicity and a lower number of pions than u- and d-jets. Therefore distinguishing $K^{\pm}$ and $\pi^{\pm}$ and reconstructing K_S^0 is crucial for strange jet identification [55–57].

SLD [58] tagged $Z \rightarrow s\bar{s}$ events by looking for the absence of reconstructed b and c hadrons and the presence of $K^{\pm}$ or K_S^0 [59]. Particle identification (PID) was performed at SLD, as at DELPHI [60], with a RICH detector. At most other detectors, energy loss (dE/dx) was used for PID [61,62], with the addition of timing at ALEPH [63]. The detector concepts at the FCC-ee foresee the use of techniques like energy loss (dE/dx) [64], ionisation cluster counting (dN/dx) [65], time-of-flight [66], and compact-Ring Imaging CHerenkov (RICH) detectors.

Tagging strange jets at future colliders has been explored as a probe to perform precision measurements in the Higgs sector [18,67], and the impact of using dN/dx and time-of-flight on strange tagging performance for jets originating from Higgs boson decay was studied using a graph neural network [42]. In this work, DeepJetTransformer is used to isolate $Z \rightarrow s\bar{s}$ events from the exclusive hadronic decays of the Z boson in the FCC-ee environment. The excess momentum carried by strange hadrons is exploited, firstly by including V^0 variables and secondly through $K^{\pm}/\pi^{\pm}$ discrimination. The cleaner environment at lepton colliders and the powerful PID capabilities of the proposed detectors facilitate making strange jet tagging feasible.

2 Experimental environment

2.1 FCC-ee

The Future Circular Collider (FCC) integrated project [68,69] aims to build e^+e^-, pp, and ep colliders in a 90.7 km circular tunnel in the Geneva region. FCC-ee [8] is a proposed e^+e^- collider and the first stage of the FCC integrated project. It is currently planned to run at four different center-of-mass energy modes, starting from around 91.2 GeV at the Z-pole to 365 GeV, over the $t\bar{t}$ threshold. The unprecedented luminosities at the FCC-ee uniquely facilitate tests of the SM and, at the same time, present novel challenges in reducing systematic errors. The circular collider design provides the opportunity for four interaction points, each of which can host a different detector design. Such detector concepts [41,70,71] are currently being studied, of which the IDEA detector concept [41] has been used in this study.

2.2 IDEA detector concept

A fast simulation of the IDEA detector concept [72] has been implemented in Delphes [73] and used for the simulation of the samples used in this work. A spherical coordinate system is used with its origin at the center of the detector system and the positive z axis in the direction of travel of the incoming electron. The polar angle, θ, is defined as the angle between the radial line and the positive z axis and the azimuthal angle, ϕ, is defined as the angle of rotation of the radial line around the positive z axis.

The innermost part of the IDEA detector is the monolithic active pixel sensor (MAPS) based vertex detector, which consists of three inner layers with a space point resolution of 3 µm, and two outer barrel and three disk layers on each side with a space point resolution of 7 µm. The innermost layer is positioned at a radius of 1.7 cm. The vertex detector is enclosed by the drift chamber incorporating 112 layers of 100 µm resolution. The multiple scattering of particles is minimal thanks to the main gas component being Helium. Two layers of silicon sensors surround the drift chamber to provide a very precise space point measurement. A single-hit resolution of 7 µm (90 µm) along ϕ (z) is assumed. These sit inside a solenoid magnet with a 2 T magnetic field. It is followed by a dual-readout calorimeter that is sensitive to independent signals from the scintillation and the Cerenkov light production. This results in a good energy resolution for both electromagnetic and hadronic showers. The calorimeter is enveloped by the muon system consisting of layers of chambers embedded in the magnet return yoke. The detector geometry has been modified since generating the event samples, and further optimisation is in progress.

2.3 Event samples and jet reconstruction

The simulated event samples used for training and evaluation consist of the process $e^+e^- \rightarrow Z \rightarrow q\bar{q}$, where $q \equiv b, c, (u, d, s)$, at the center-of-mass energy ($\sqrt{s}$) of 91.2 GeV. Pythia8.303 [74] is used for event generation, parton showering, and hadronisation. Delphes [73] is used for event reconstruction assuming the IDEA detector concept [41,72]. A tracking efficiency of 99.7% is assumed for electrons, muons and charged hadrons with 3-momentum magnitude $|p| > 0.5$ GeV that lie within acceptance. This efficiency is reduced to 65% (4%) for $0.5 > |p| > 0.3$ GeV ($|p| < 0.3$ GeV). Fake tracks are not considered.

Jet clustering is performed on the particle flow-style objects reconstructed by `Delphes` with `FastJet-3.3.4` [75] using the exclusive e^+e^- $k_{\rm T}$ algorithm [76]. Other jet clustering algorithms like the anti-$k_{\rm T}$ algorithm [77] and the generalised e^+e^- $k_{\rm T}$, also referred to as the inclusive e^+e^- $k_{\rm T}$, algorithm [75] were also considered. The exclusive e^+e^- $k_{\rm T}$ algorithm, which creates irregularly shaped jets and, in this study, requires exactly two jets, is very robust against gluon emissions and gluon splitting. Since the exclusive e^+e^- $k_{\rm T}$ algorithm clustered jets include all reconstructed final particles, they were observed to satisfy the requirements of this study by most accurately reproducing the Z boson reconstructed invariant mass signature. No additional selections were applied to the samples for training and evaluation of the jet flavour tagger.

In this study, the jets are assigned an MC flavour as the flavour of the quarks to which the Z boson decays. Besides simplicity, this has the added benefit that other studies for future facilities use the same definition.

A separate set of event samples was generated with the process $e^+e^- \rightarrow Z(\rightarrow \nu\nu)H \rightarrow q\bar{q}$, where $q \equiv b, c, (u, d, s)$, at $\sqrt{s}$ of 240 GeV. The same reconstruction and jet clustering were applied as for the samples at the Z resonance. Training and evaluation of `DeepJetTransformer` were performed with these samples for comparison with other taggers.

3 Vertex reconstruction

Vertex reconstruction is essential to find the primary interaction vertex and the secondary decay vertices of the long-lived b, c, and s hadrons. It helps improve the b- and c-tagging performance and aids in s-tagging. Charged tracks can be fitted to reconstruct the primary and the displaced secondary vertices. These displaced vertices can either be the decay vertices of b and c hadrons (SVs) or those of the long-lived hadrons containing s quarks, like K^0_S or Λ^0, commonly referred to as V^0s, which are particles that decay into a pair of oppositely charged tracks. All displaced vertices except V^0s are referred to as SVs. The properties of SVs and V^0s, such as their masses, displacements, and charged track multiplicities, can be used to identify the decaying hadrons and, in effect, the jet flavour. The SVs can even be used to reconstruct the entire hadronic decay chain. Similarly, reconstructing and identifying the V^0 vertices can be used to identify s-jets, as K^0_S and Λ^0 are the particles carrying most of the momentum of some s-jets [18]. Distinguishing V^0s from SVs also helps to reduce the misidentification of some b- and c-jets as s-jets.

The vertex reconstruction in this study has been performed using an implementation of the vertexing module of the `LCFIPlus` framework [78,79]. It has been implemented in `FCCAnalyses` [80], the FCC software framework, using a χ^2-based vertex fitter [81]. The constraints and parameters have been kept the same as in Ref. [78]. The algorithm first identifies the tracks forming V^0s. Unlike in standard vertex reconstruction algorithms, the V^0s are not discarded but stored and assigned a particle ID based on the set of constraints that they pass, summarised in Table 1. The tracks originating from the primary vertex or V^0 candidates are not considered while reconstructing SVs.

The properties of the SVs and V^0s, along with more variables, are used as input to train the neural network tagger described in Sect. 4.

3.1 V^0 vertex reconstruction

Two processes are considered: $K^0_S \rightarrow \pi^+\pi^-$ and $\Lambda^0 \rightarrow p\pi^-$. The invariant mass of the reconstructed K^0_Ss can be seen in Fig. 1a, demonstrating a good reconstruction of V^0s and their properties. The mass of the tracks used to calculate the invariant mass of the V^0 is decided based on the set of constraints the V^0 passes with a certain permutation of the two tracks. In contrast, all tracks are assumed to be pions in the invariant mass calculation for the SVs.

Figure 1b displays the V^0 multiplicity in jets from $Z \rightarrow q\bar{q}$ events. No reconstructed V^0s are found for most of the jets. But, a higher fraction of heavy- and strange-flavoured jets contain reconstructed V^0s than u- and d-jets, which justifies the importance of V^0 rejection before attempting to reconstruct SVs. It is also evident that more s-jets have one or more reconstructed V^0s than u- and d-jets, making V^0s an important discriminator of s-jets against lighter quark jets.

3.2 Secondary vertex reconstruction

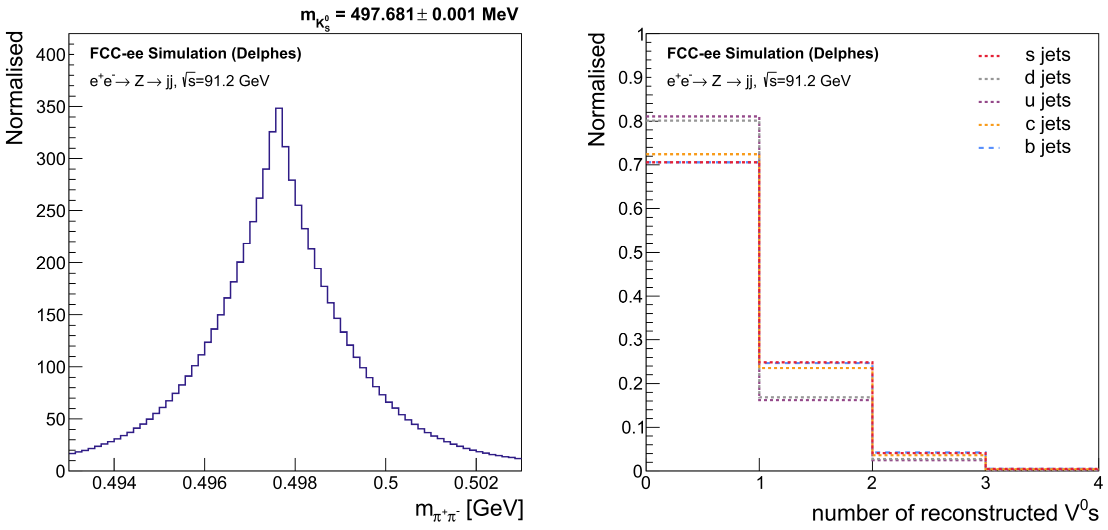
Due to the near-diagonal CKM matrix, the cascading decay chain of heavier quarks is expected to be $b \rightarrow c \rightarrow s \rightarrow (u, d)$. Hence, the SV multiplicity tends to be higher in b-jets compared to c-, s-, u-, and d-jets, as shown in Fig. 2.

4 DeepJetTransformer

Since the introduction of `ParticleNet` [26], the concept of a Particle Cloud has become the prevailing representation of jet structure. A Particle Cloud considers the jet as an unordered set of jet constituents of varying length. Elements of differing nature, such as charged, neutral particles, or SVs associated with the jet, are considered to create the most complete and accurate representation. This representation concept was used to build the presented model, the key element of which, the unordered set of particles, requires the construction of a model invariant under the permutation of the jet

Table 1 Summary of the default V^0 selection criteria [78]. M is the invariant mass, and p is the momentum of the V^0 candidate. r is the distance of the V^0 candidate from the primary vertex. The collinearity of the V^0 candidate is defined as $\hat{p} \cdot \hat{r}$. The set of 'tight' constraints has been used to identify V^0s in this study, while the set of 'loose' constraints has been used to remove the V^0 background while reconstructing SVs

	K_S^0		Λ^0	
	tight	loose	tight	loose
M [GeV]	[0.493, 0.503]	[0.488, 0.508]	[1.111, 1.121]	[1.106, 1.126]
r [mm]	> 0.5	> 0.3	> 0.5	> 0.3
$\hat{p} \cdot \hat{r}$	> 0.999	> 0.999	> 0.99995	> 0.999



(a) Reconstructed K_S^0

(b) V^0 multiplicities

Fig. 1 Performance of V^0 reconstruction. **a** Invariant mass distribution of reconstructed K_S^0 vertices. The quoted mass is the mean and the error on the mean of the distribution. **b** The reconstructed V^0 multiplicity in jets from $e^+e^- \rightarrow Z \rightarrow q\bar{q}$ events at $\sqrt{s} = 91.2$ GeV, where $q \equiv u, d, s, c, b$. The distributions for b- and s-jets overlap almost perfectly

constituents,[1] a benefit also used by other transformer-based taggers [40]. Moreover, Transformers possess the essential property of full connectivity between jet constituents via the attention mechanism [37]. This enables the model to capture subtle correlations among jet constituents, enhancing the high-level features used for jet discrimination.

A structure based on Transformer blocks was thus chosen for this study. Previous research has indicated that Transformer models offer enhanced performance and increased efficiency, particularly compared to graph models [40,43]. The subsequent sections will elaborate on the inputs to the neural network and the fundamental characteristics of Transformer models and provide a detailed description of the specific model, `DeepJetTransformer`, which has been developed for this study.

4.1 Input features

The properties of each jet and its constituents represent different categories of input features available for model training. All input features are built using information reconstructed with `Delphes` detector simulation unless stated otherwise. The jet kinematics are represented by variables defined using its 4−momentum, as detailed in Table 2. Many future collider detector concepts are designed to be used with a particle flow algorithm [83,84]. Therefore, jet constituents are subdivided into five sets according to the typical particle flow candidate categories: charged hadrons, neutral hadrons, electrons and positrons ($e^\pm$), photons (γ), and muons ($\mu^\pm$). Kinematic variables are defined for each jet constituent using its 4-momentum, as listed in Table 3. For each jet up to 25 charged jet constituents and 25 neutral jet constituents are considered. This is enforced by truncating the input feature array of a given jet if the number of charged/neutral jet constituents is more than 25. Conversely, if the number of charged/neutral

[1] Permutation invariance is in opposition to most Transformer models established around the principle of causality [37,82].

Fig. 2 SV multiplicity in jets from $e^+e^- \rightarrow Z \rightarrow q\bar{q}$ events. The term “light jets” here collectively refers to u-, d-, and s-jets

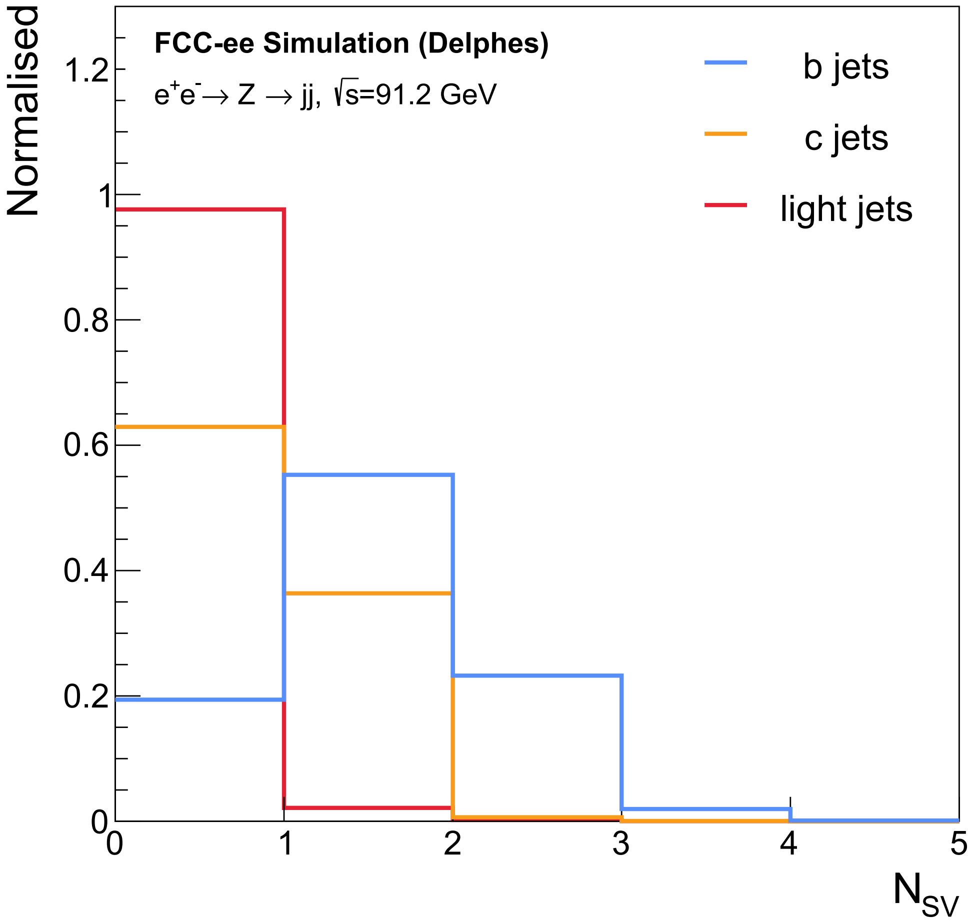


Table 2 Description of global features associated with each jet

Input feature	Description
$\lvert p \rvert,\ E,\ m$	3-Momentum magnitude, energy, and invariant mass of the jet
θ, ϕ	Polar and azimuthal angle of the jet axis
$\mathrm{N}_{\mathrm{charged}}$	Charged particle (track) multiplicity in the jet
$\mathrm{N}_{\mathrm{neutral}}$	Neutral particle multiplicity in the jet
$\lambda^{\kappa}_{\beta} = \Sigma_{i \in \mathrm{jet}} z^{\kappa}_{i} R^{\beta}_{i}$	Jet angularity [85] as sum of normalized jet constituent energy (z_i) and angular distance to jet axis (R_i) for ($\kappa = 0, \beta = 0$), ($\kappa = 1, \beta = 0.5$), ($\kappa = 1, \beta = 1$), ($\kappa = 1, \beta = 2$), ($\kappa = 0, \beta = 2$)
`isU/D/S/C/B`	MC flavour assigned to the jet

jet constituents is less than 25, then the input feature array is zero-padded.

Charged tracks are first fitted to find the V^0s and the remaining tracks are used to reconstruct SVs. Feature variables are defined separately for both classes of reconstructed vertices (V^0s and SVs) and are listed in Table 4. Up to 4 V^0s and 4 SVs are considered per jet. The V^0 and SV input feature arrays are likewise truncated/zero-padded. The distinguishing power of some of these variables is discussed below.

The jet 3-momentum magnitude distribution of b- and c-jets tends to be more spread out than that of s-, u-, and d-jets, as seen in Fig. 3a. This is due to the longer decay chain in c-jets than s-, u-, and d-jets, and even longer decay chains in b-jets, where more momentum can be lost through neutrinos than in s-, u-, and d-jets.

An important distinguishing variable for b-jet identification is the transverse impact parameter (D_0), which is higher for heavier flavour jets as the decaying b hadrons have a significantly longer lifetime than c or s, u, d hadrons (except for V^0s). The differentiating effect between flavours caused by this can be seen more clearly in the transverse impact parameter significance, defined as $S(D_0) = D_0/\sigma_{D_0}$, where σ_{D_0} is the uncertainty in the measurement of the transverse impact parameter. It is depicted in Fig. 3b.

Table 3 Description of features associated with each jet constituent. The sets of variables are divided into charged particles (tracks) and neutral particles

Input feature	Description
$D_0 (z_0)$	Signed transverse (longitudinal) impact parameter
$D_0/\sigma_{D_0} (z_0/\sigma_{z_0})$	Signed transverse (longitudinal) impact parameter significance
$\theta_{\text{rel}} (\phi_{\text{rel}})$	Polar (azimuthal) angle of track with respect to the jet axis
R	Angular distance of track and jet axis
C	Half-curvature of the track
$m_{\text{ch.}}, q$	Track invariant mass and charge
$\frac{\lvert p\rvert_{\text{ch.}}}{\lvert p\rvert_{\text{jet}}}, \ln(\lvert p\rvert_{\text{ch.}}), \ln\left(\frac{\lvert p\rvert_{\text{ch.}}}{\lvert p\rvert_{\text{jet}}}\right)$	(Normalised) magnitude of track momentum and logarithms
$\frac{E_{\text{ch.}}}{E_{\text{jet}}}, \ln(E_{\text{ch.}}), \ln\left(\frac{E_{\text{ch.}}}{E_{\text{jet}}}\right)$	(Normalised) track energy and logarithms
`isKaon`	If the particle is identified as a $K^{\pm}$
`isMuon`	If the particle is identified as a $\mu^{\pm}$
`isElectron`	If the particle is identified as an $e^{\pm}$
$\theta_{\text{rel}} (\phi_{\text{rel}})$	Polar (azimuthal) angle of particle with respect to the jet axis
R	Angular distance of neutral particle and jet axis
$\frac{\lvert p\rvert_{\text{neut.}}}{\lvert p\rvert_{\text{jet}}}, \ln(\lvert p\rvert_{\text{neut.}}), \ln\left(\frac{\lvert p\rvert_{\text{neut.}}}{\lvert p\rvert_{\text{jet}}}\right)$	(Normalised) magnitude of particle momentum and logarithms
$\frac{E_{\text{neut.}}}{E_{\text{jet}}}, \ln(E_{\text{neut.}}), \ln\left(\frac{E_{\text{neut.}}}{E_{\text{jet}}}\right)$	(Normalised) neutral particle energy and logarithms
`isPhoton`	If the particle is identified as a Photon

Table 4 Description of features associated with each reconstructed secondary vertex. Similar features, with the addition of PDG ID [54], are also defined for V^0s while comparing the performance of the tagger trained with and without V^0s

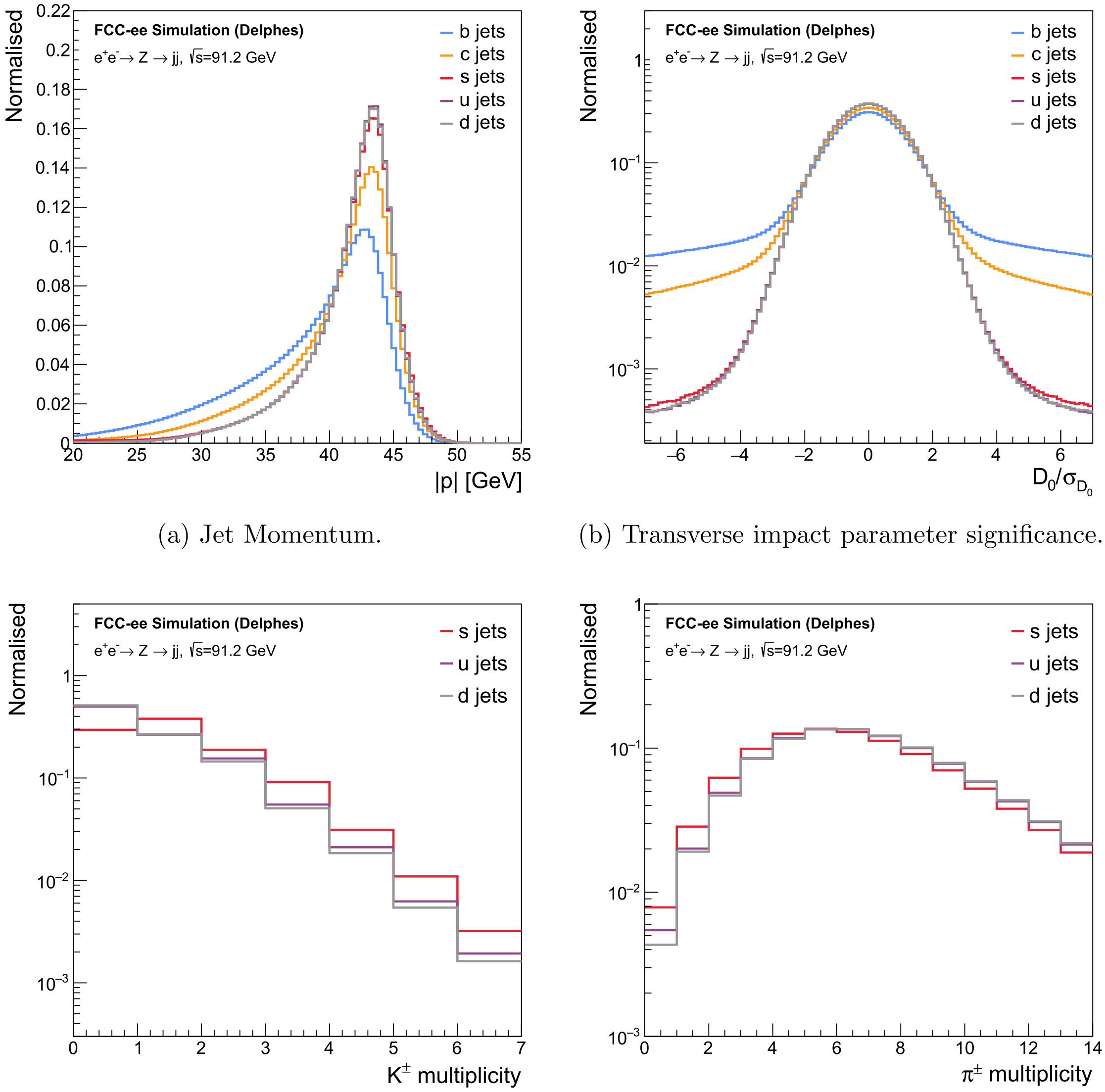
Input feature	Description
$\lvert p\rvert,\ m$	3-Momentum magnitude and invariant mass of the SV
N_{tracks}	Track multiplicity of the SV
$\chi^2, \text{N}_{\text{DoF}}$	χ^2 and number of degrees of freedom of the SV
$\theta_{\text{rel}}, \phi_{\text{rel}}$	Polar and azimuthal angle of the SV with respect to the jet axis
$\hat{\text{p}}.\hat{\text{r}}$	Collinearity of SV with respect to PV
$\text{d}_{\text{3D}}, \text{d}_{xy}$	3D and transverse distance of the SV from the PV

As mentioned in Sect. 3.2, b-jets tend to have a higher SV multiplicity than c-, s-, u-, and d-jets. It is a dominant property in identifying b-jets and, to some extent, c-jets.

The most challenging background for s-tagging is ud-jets. Two powerful distinguishing variables tend to be the multiplicities of charged and neutral Kaons and Pions, exploiting the conservation of strangeness during hadronisation in strange jets. These can be seen in Figs. 3c and 3d. To distinguish between $K^{\pm}$ and $\pi^{\pm}$, PID techniques like energy loss (dE/dx) [64], ionisation cluster counting (dN/dx) [65], time-of-flight [66], etc. are traditionally used. The $K^{\pm}/\pi^{\pm}$ classification is generically emulated, instead of relying on any particular PID technique, with several scenarios of different efficiency to correctly identify $K^{\pm}$, the baseline scenario being 90% efficiency and a 10% efficiency of misidentifying $\pi^{\pm}$ as $K^{\pm}$. The $K^{\pm}$ identification efficiency and the $\pi^{\pm}$ misidentification efficiency are chosen to be constant over the entire momentum range for all the scenarios. The other scenarios considered to study the impact of PID on flavour tagging are summarised in Sect. 5.2. The baseline PID scenario was deliberately conservative with respect to the state-of-the-art $K^{\pm}$ identification, which is expected to provide better than 3σ $K^{\pm}/\pi^{\pm}$ separation using cluster counting at FCC-ee [86], potentially supplemented by time-of-flight [8,87]). This study instead follows PID studies at Belle, which found the average efficiency and fake rate for charged particles between 0.5 and 4 GeV/c to be $(87.99 \pm 0.12)\%$ and $(8.53 \pm 0.10)\%$, respectively [88]. The reconstructed V^0s, as shown in Fig. 1, further improve PID by identifying the neutral strange hadrons, K^0_S and Λ^0. These variables, as described in Tables 2, 3 and 4, are fed into a neural network, the architecture of which is described below.

4.2 Transformer models

Inspired by the success of attention mechanism in Natural Language Processing (NLP) [37,82] or Computer Vision (CV) [89] tasks, this model adopts Transformer blocks as its



(a) Jet Momentum.

(b) Transverse impact parameter significance.

(c) Kaon multiplicities.

(d) Pion multiplicities.

Fig. 3 Distinguishing features in the clustered jets of $e^+e^- \to Z \to q\bar{q}$ events at $\sqrt{s} = 91.2$ GeV, separated by flavour. **b** Shows a property of the jet constituents, while **a**, **c**, and **d** show properties of the clustered jets. The IDEA detector concept was used for reconstruction

primary architectural component. Transformers belong to a class of neural networks that leverage the scaled dot-product attention (SDPA) mechanism [37]. The attention mechanism enables the model to selectively focus on specific segments of the input sequence while processing each constituent element. In contrast to earlier architectures, such as recurrent models that utilise fixed-size windows or recurrent connections, the attention mechanism dynamically assigns weights to individual elements within the jet based on their relevance, capturing intricate dependencies across the entirety of the jet structure. This adaptive and global weighting scheme empowers the Transformer to effectively model contextual information, a crucial element for understanding and generating coherent high-level features.

4.2.1 Scaled dot-product attention and heavy flavour transformer block

The SDPA mechanism uses three inputs: a query matrix Q, a key matrix K, and a value matrix V. In general, the query matrix represents the items for which the attention weights are computed, while the key and value matrices represent all items in the sequence. In this study, the items can be understood to be jet constituents. After being fed into linear layers, the query tensor Q of dimension (B, N, d_k), the key tensor K of dimension (B, L, d_k), and the value tensor V of dimension (B, L, d_k') are fed into the scaled dot-product attention as:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{4.1}$$

The attention mechanism in this study is employed in a specific configuration where the input query, key, and value tensors are identical ($Q = K = V$), and derived from jet constituent features. The tensors Q, K, and V are each passed through linear layers, facilitating the transformation and projection of the input tensors to the attention space. The SDPA is then computed on these transformed tensors as $\text{Attention}(QW^Q, KW^K, VW^V)$, where W^Q, W^K, W^V, represent distinct linear transformations. This particular case is commonly referred to as self-attention [37].

SDPA is extended to enhance the discriminating power of the model by allowing it to attend to multiple subspaces of attention in parallel. This extension, referred to as Multi-Head Attention (MHA), facilitates the capture of diverse and complementary high-level features from the jet constituent input by projecting the Query, Key, and Value matrices independently for each of the h attention heads. Each attention head performs an SDPA operation, yielding distinct representations. These head representations are then concatenated and passed through a linear layer to integrate the information across heads. The MHA layer can mathematically be represented by the following equations:

$$\text{MHA}(Q, K, V) = \text{Concat}(h_1, ..., h_n)W^O, \tag{4.2}$$

$$h_i = \text{Attention}(QW^{Q,i}, KW^{K,i}, VW^{V,i}). \tag{4.3}$$

The presented approach, employing the Particle Cloud representation [26], intentionally refrains from employing positional encoding. This decision stems from the absence of a hierarchical structure or positional ordering among the components of the jets, in contrast to sequences such as sentences or image patches. Consequently, the MHA module operates without incorporating positional encoding and instead only leverages permutation invariant mechanisms to capture and process the interrelationships between particles in the jet, yielding meaningful results. The permutation invariance of `DeepJetTransformer` is established by the properties of permutation equivariance and invariance of function composition [90]. The permutation equivariance of each function of the transformer blocks ensures that the network produces a representation of the jet constituents respecting the Particle Cloud properties. It is made sure that the network's flavour predictions remain invariant under the permutation of jet constituents by applying a permutation invariant attention pooling followed by linear layers for classification. By analogy with graph structures, the attention mechanism can be interpreted similarly to the ones used in fully connected graph networks, with the attention scores playing a role similar to the edge features by capturing relationships within the jet structure.

After establishing the fundamental components of the utilised model's architecture, the foundational block forming the backbone of the model can be defined. This essential building block, referred to as the Heavy Flavour Transformer block (HFT), is structured in the following manner:

- The jet constituent inputs are fed into a basic Multilayer Perceptron (MLP) layer followed by a ReLU activation function.
- The product of the MLP layer is then fed in an MHA layer before using a residual connection and layer normalisation.
- In addition to the MHA layer, a fully connected feed-forward layer is also added, identical to the original Transformer implementation [37] followed by a last residual connection and layer normalisation.

Unlike other Transformer models applied to jet (sub)structures [30,40], a *cls* token is not employed to embed the information of the jet structures into relevant features for classification. Instead, an attention pooling is introduced, behaving similarly to a Max or Average pooling layer with an attention mechanism and learnable parameters. The attention pooling operates by employing an MLP projection layer, which enables local feature extraction. Subsequently, a softmax activation function is applied to calculate attention weights, allowing the layer to emphasise relevant elements in the sequence. The attention weights are then used to aggregate the sequence information by performing a weighted sum. To enhance the layer's performance, batch normalisation is applied, the ReLU [91] activation function is used to introduce non-linearity, and dropout regularisation is incorporated to prevent overfitting. The attention pooling layer can effectively capture essential information from the sequence and produce a condensed representation by incorporating these components that can be utilised for jet flavour classification. In the context of jet flavour tagging, Transformer models can be interpreted as fully connected graph networks using the jet's constituents as the nodes, and the SDPA as a mechanism connecting all the node information for enhancing the feature engineering of the model.

4.2.2 *DeepJetTransformer architecture*

With all the components of `DeepJetTransformer` defined, the global structure of the model can be described. Figure 4 illustrates the detailed structure of `DeepJetTransformer`, which is as follows:

- The features of distinct jet constituents first undergo embedding via a series of three MLPs with output fea-

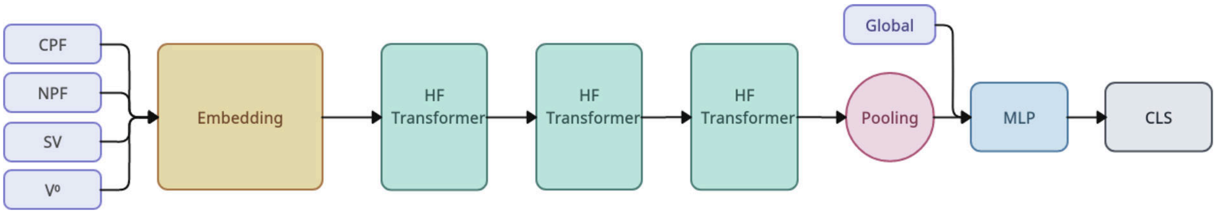


Fig. 4 Schematic structure of `DeepJetTranformer` model

ture dimensions of (64, 128, 128), employing ReLU activation, residual connections, and batch normalisation. Dropout regularisation with a rate of 0.1 is applied following each batch normalisation operation.

- The resulting feature tensors are then concatenated to form a single tensor containing all the comprehensive information of the jet constituents.
- This global tensor is subsequently passed through three HFT blocks, each possessing a feature dimension of 128. Each block contains eight attention heads and incorporates a dropout rate of 0.1.
- The representation of the jet structure, obtained through the HFT blocks, is further condensed via attention pooling. The resulting tensor is concatenated with jet-level features, yielding a vector containing 135 relevant features for heavy flavour classification. Among these, 128 features originate from attention pooling, while the remaining seven variables represent the jet-level attributes.
- The jet representation is subsequently fed to three MLPs with output feature dimensions of (135, 135, 135), mirroring the structure of the input embedding MLPs.
- A single MLP followed by a SoftMax function is applied finally for classification.

In summary, the three main differences in the architectures of `DeepJetTransformer` and `ParT (plain)` [40], which both implement a pure Transformer architecture, are the use of attention pooling instead of the typical *cls* token, additional linear layers prior to the MHA, and the inclusion of the of jet-level variables (listed in Table 2) in addition to jet constituent variables.

4.2.3 Training methodology

`PyTorch (v1.10.1)` [92] was employed as the deep learning library in this study for the neural network model construction and the training process. The optimiser utilised was the Lookahead optimiser [93], with hyperparameters $k = 6$ and $\alpha = 0.5$ and a RAdam [94] as the base optimiser with a learning rate of 5e−3 and decay rates $(\beta_1, \ \beta_2)$ set to $(0.95, \ 0.999)$. The training was conducted over 70 epochs with a batch size of 4000, accompanied by a per-epoch linear learning rate decay starting after 70% of the training, gradually decreasing to 5e−5 by the final epoch. A cross-entropy loss function was used for optimisation. The training dataset comprised of 1 million jets, divided into an 80/20% train-validation split. Finally, the model was evaluated on a separate dataset of 1 million jets for performance assessment. Documentation for the sample preparation and training methodology, along with the relevant code, is publicly available here: DeepJetFCC.[2]

5 Classifier performance

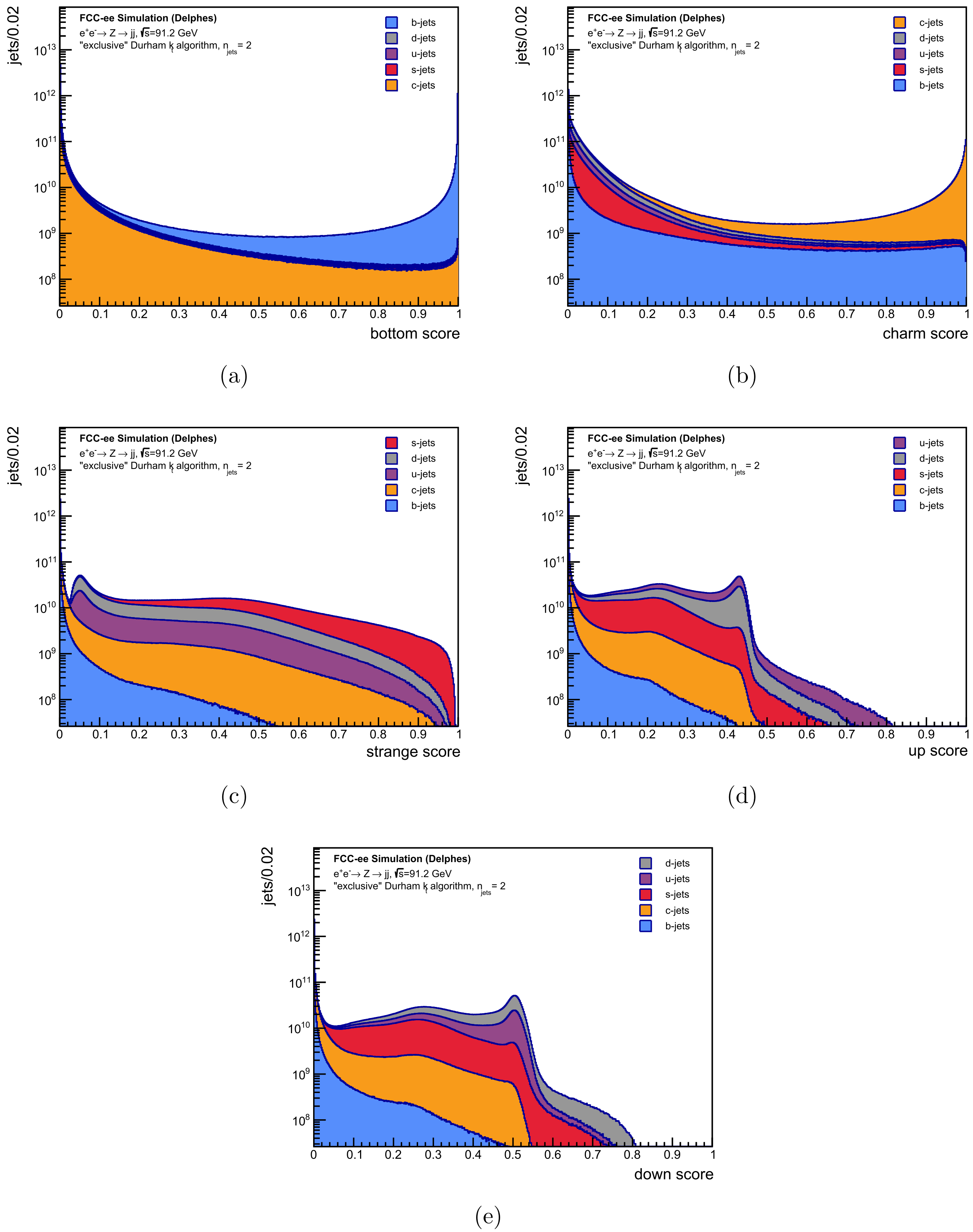
To evaluate the performance of `DeepJetTransformer`, clustered jets from $Z \rightarrow q\bar{q}$ events at $\sqrt{s} = 91.2$ GeV and $Z(\rightarrow \nu\nu)H(\rightarrow q\bar{q})$ events at $\sqrt{s} = 240$ GeV were considered. The tagger was trained separately for each process. The emphasis was placed on the Z resonance for these studies, with the classification of $H \rightarrow q\bar{q}$ events serving primarily as a comparison to the classification performance of other jet flavour taggers for future colliders, like `ParticleNetIDEA` [42,95]. A binary classifier was constructed for each jet flavour $q \equiv u, d, s, c, b, (g)$ with a signal flavour (i) and a background flavour (j):

$$S_{ij} = \frac{S_i}{S_i + S_j}, \tag{5.1}$$

where S_i are the outputs of the classifier normalised using the SoftMax function, as described in Sect. 4.2.2. These normalised outputs, which are constrained to lie between 0 and 1 and sum to 1 across all jet flavours, are referred to as "softmaxed classifier outputs" for brevity.

The five softmaxed classifier outputs of `DeepJetTransformer` are shown in Fig. 5. ROC curves were computed for each S_{ij} combination and are depicted in Fig. 6 for the Z resonance and the ZH training. Predictably, the strongest discrimination is between b-jets and s-, u-, d- jets and is roughly equivalent for all three background jets. The dominant background is from c-jets, originating from the similarity of b- and c-jets with a single reconstructed SV. Discriminating c-jets from u-, d- and s-jets exhibits similar performances, with

[2] https://github.com/Edler1/DeepJetFCC/tree/master/docs.



(a)

(b)

(c)

(d)

(e)

Fig. 5 The softmaxed classifier outputs (S_i) of the five output nodes of `DeepJetTransformer` trained with clustered jets of $e^+e^- \to Z \to q\bar{q}$ events at $\sqrt{s} = 91.2$ GeV. The contributions of different MC flavours have been displayed

relatively worse discrimination of the s-jet background. Figure 6b shows that as the efficiency increases from the right to the left side of the plot, s-, u- and d-jets are discriminated worse than b-jets in the high-efficiency regime for c-jets until a turnover point at $\epsilon_{sig}^{c} \approx 80\%$, after which distinguishing s-, u- and d-jets becomes considerably easier than b-jets. Such a turnover can also be found in `ParticleNetIDEA` [42]. The sub-leading background comes from s-jets, clustered at low to mid charm scores, as also evident in Fig. 5b, primarily as no SVs can be reconstructed for a significant number of c-jets, leaving few variables to distinguish c- and s-jets.

When s-jets are taken to be the signal, as shown in Fig. 6c, c- and ud-jets present the most challenging backgrounds, with c-jets being easier to discriminate against at all signal purities. The c-jet background comes from jets where a charm hadron decays to a strange hadron, and only the V^0 can be reconstructed, or a strange hadron carries excess momentum. Some discrimination against the dominant ud-jets background can be achieved at higher cuts on the strange score, owing to the $K^{\pm}/\pi^{\pm}$ separation and V^0 reconstruction. Finally, Fig. 5d and e show almost overlapping distributions of classifier scores for u- and d-jets. Figure 6d validates that classification is most challenging for u- and d-jets. When u-jets are taken to be the signal, it can be seen that `DeepJetTransformer` learns to discriminate u- vs d-jets with a $\epsilon_{sig}^{u} \approx 15\%$ at a $\epsilon_{bkg} = 10\%$, which is better than a random classifier, although not considerably. The discrimination is likely related to a mapping to the initiating parton's charge, such as the jet charge [96,97], the effect of which is diluted by the presence of antiquarks.

While considering the performance for $H(\rightarrow q\bar{q})$ jets, depicted as dashed lines in Fig. 6, no clear trend can be observed. Slight degradation in performance can be observed in the case of b tagging, compared to $Z \rightarrow q\bar{q}$ jets, particularly when c-jets are taken to be the background. The discrimination of c-jets vs s-, u-, and d-jets is found to perform relatively the best with respect to the $Z \rightarrow q\bar{q}$ jets when considering the percent-improvement in the ROC Area Under the Curve metric.

Figure 6e shows that the best quark-gluon discrimination can be achieved against the b quarks. This performance can be attributed to several discriminating variables, like jet-constituent multiplicity, constituent momentum distribution, etc., but is dominated by the presence or absence of reconstructed SVs. It is the most challenging to discriminate the s, u, and d quarks from gluons due to their similar jet composition.

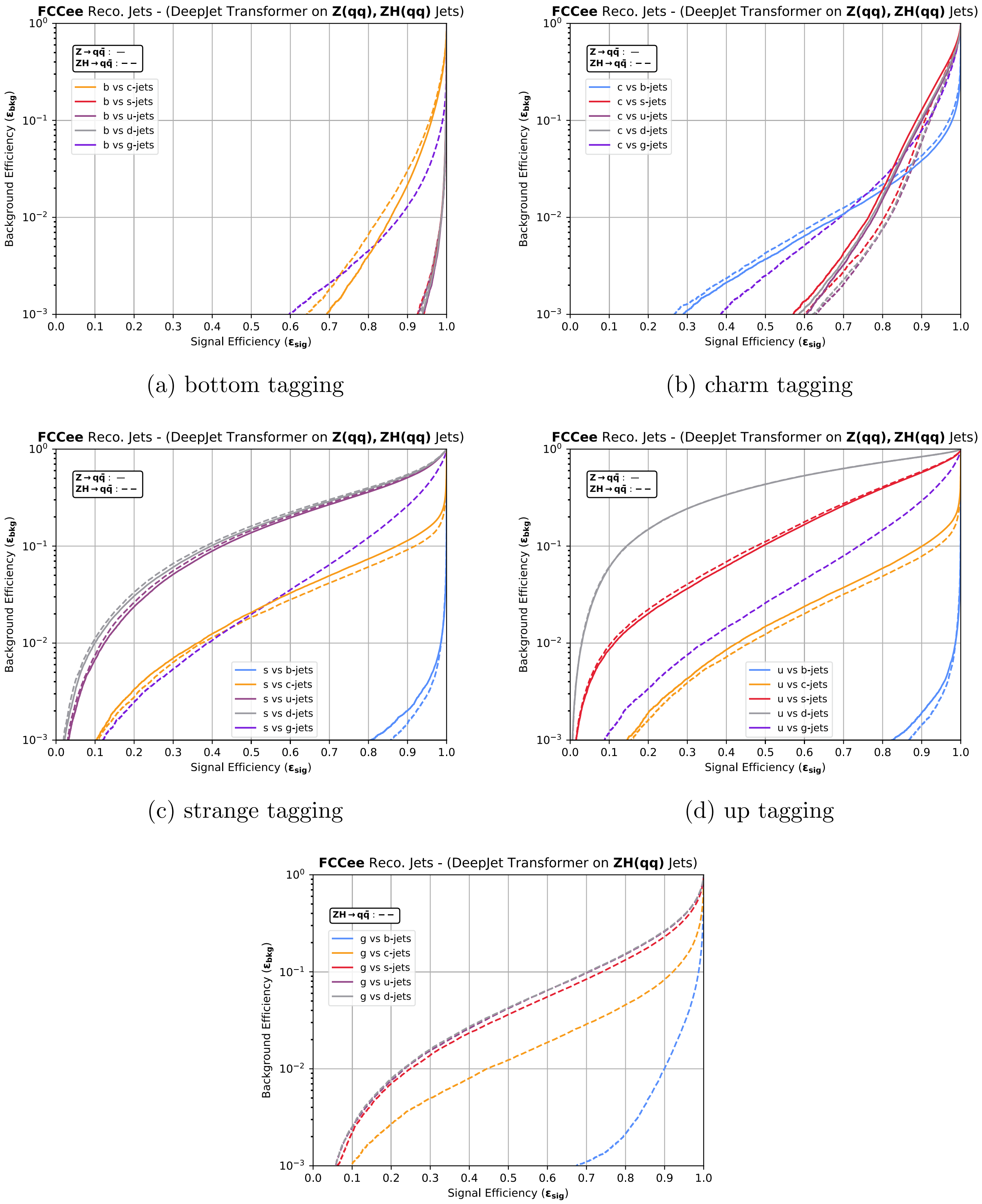
The tagging efficiency of `DeepJetTransformer` was evaluated for three cases: b vs c tagging, c vs s tagging, s vs ud tagging. Figure 7 shows the efficiency of `DeepJetTransformer` over the entire jet momentum range and the jet-axis polar angle (θ) range for all three cases for two working points. The efficiency for b vs c tagging and c vs s tagging is mostly uniform, showing a good performance for all jet momenta. Similarly, the performance is largely uniform over the θ range for all three cases, degrading at the extremes due to jet constituents being lost by fiducial cuts.

However, the s vs ud tagging efficiency displays a peculiar distribution over the momentum range of interest, as shown in Fig. 7e. This was found to be dependent on the two most distinguishing features for identifying s-jets: $K^{\pm}/\pi^{\pm}$ discrimination and V^0 reconstruction. The *low-momentum* ($24 < |p| < 35$ GeV) strange jets, on average, have lower $K^{\pm}$ multiplicities, which leads to a reduced tagging efficiency. The *very-low-momentum* ($|p| < 24$ GeV) strange jets have a significantly low total charged-particle multiplicity, making V^0 reconstruction crucial. The majority of such jets have a single reconstructed V^0, helping identify the s-jets. On the other hand, the *low-momentum* strange jets tend to have multiple V^0s, splitting the already low jet momentum among these V^0s and other hadrons. This is expected to make the strange jet identification more ambiguous. Hence, the s-tagging efficiency rises at very low momenta.

A similar but exaggerated trend in the distribution is seen for the looser working point of 10% mistag rate for jets with momentum values below 25 GeV. The efficiency is observed to be stable in momentum above this value. As stated above, some of this increase in s-tagging efficiency can be attributed to the presence of a reconstructed V^0 in jets with low particle multiplicities. Another important aspect to note is that only a small fraction of jets ($< 1\%$) with such very low momenta are present in Z boson decays. This means that these low-momentum jets will not have a large contribution to the training of the neural network or the working point determination, which will both be dominated by the bulk of the momentum distribution. The fact that the 10% u, d-jet background efficiency also increases to 40% for momenta less than 25 GeV implies that this part of the jet momentum phase space is likely not optimally examined by the neural network. A potential method to improve would be to use training weights flattened over the jet momentum and train on much larger samples with this part of the momentum distribution sufficiently populated. But since these jets contribute to a very small fraction of the total Z boson decays, the improvement in analyses requiring strange tagging would likely not be significant unless the physics case is specific.

5.1 Qualitative comparison with other taggers

A fair quantitative comparison with other taggers developed for future colliders is not feasible due to differing event samples and input features. However, the jet tagging performance trends are very similar to those of `ParticleNetIDEA` [42,95]. The strange tagging efficiency of `ParticleNetIDEA` against the u-, d-jets sur-



(a) bottom tagging

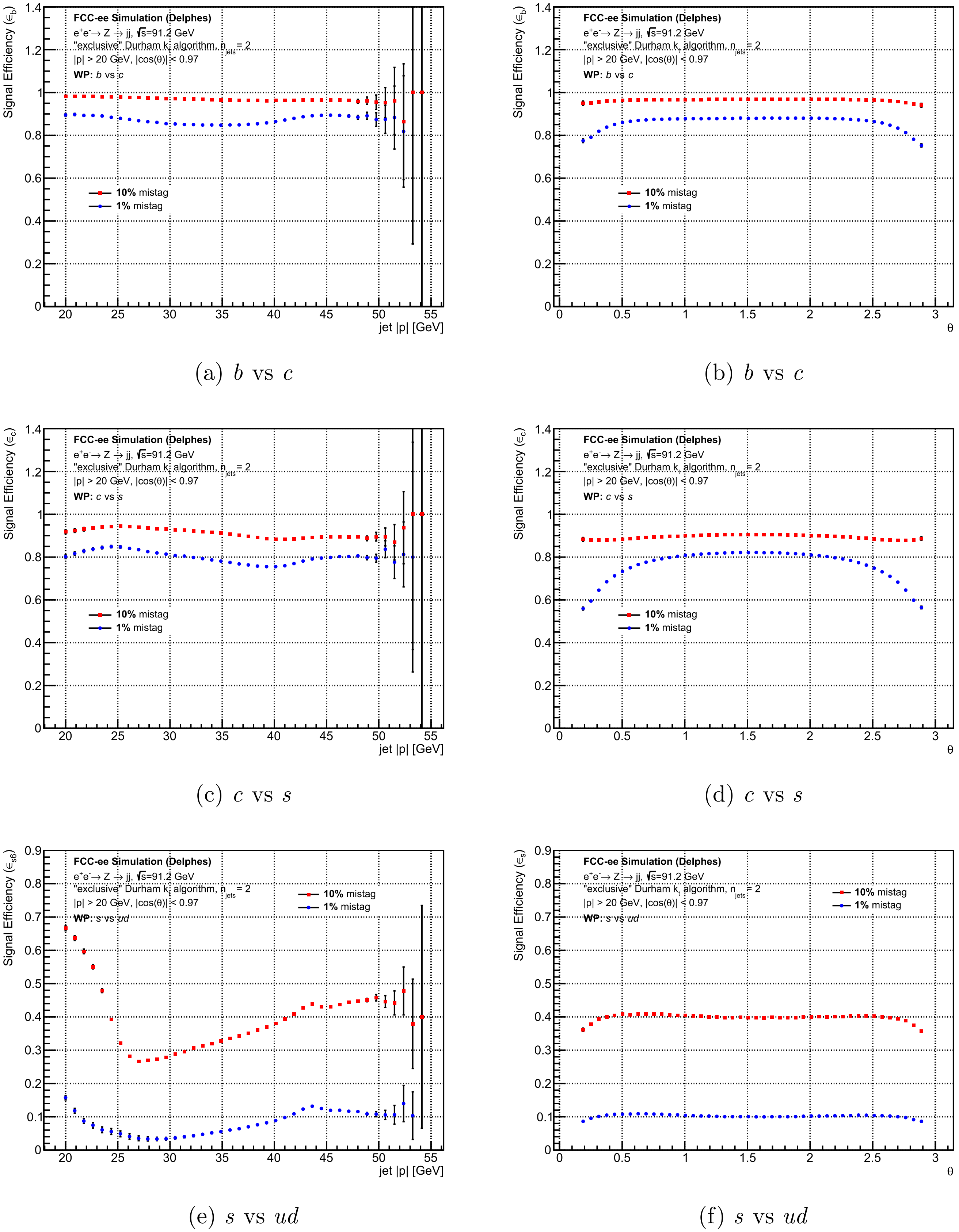
(b) charm tagging

(c) strange tagging

(d) up tagging

(e) gluon tagging

Fig. 6 ROC curves for each S_{ij} combination, as defined in Eq. 5.1, where i is the signal parton flavour and j is the background flavour. The solid lines correspond to the classification of jets at the Z resonance at $\sqrt{s} = 91.2$ GeV, while the dashed lines correspond to the classification of jets from $Z(\to \nu\nu)H(\to q\bar{q})$ events at $\sqrt{s} = 240$ GeV. The tagger was trained separately for each process. No quark-gluon discrimination results are presented for jets from $Z \to q\bar{q}$ events as the Z boson does not decay into gluons



(a) b vs c

(b) b vs c

(c) c vs s

(d) c vs s

(e) s vs ud

(f) s vs ud

Fig. 7 The jet flavour tagging efficiency over the range of jet momentum and the jet axis polar angle for jets of $e^+e^- \rightarrow Z \rightarrow q\bar{q}$ events at $\sqrt{s} = 91.2$ GeV. Three cases at 1% and 10% background efficiencies are shown: b vs c tagging, c vs s tagging, s vs ud tagging

passes that of DeepJetTransformer, owing to PID techniques like cluster counting and time-of-flight used by ParticleNetIDEA and the conservative PID estimates of DeepJetTransformer. A more detailed training dataset including such PID variables is expected to improve the tagging efficiencies of DeepJetTransformer.

DeepJetTransformer outperforms ParticleNetIDEA in bottom-gluon discrimination, especially for efficiencies lower than 90%. DeepJetTransformer also has a better discrimination of b-jet background for all other signal quark jet flavours. This efficient discrimination can be attributed to the inclusion of SVs.

With about 10^6 parameters and efficient transformer blocks as the workhorse, training DeepJetTransformer converges within 2 h after approximately 50 epochs on an NVIDIA Tesla V100s GPU. The computational complexity, measured in FLOPs, is approximately 19.7 MFLOPs. Comparatively, DeepJetTransformer requires fewer FLOPs than competing architectures [26,40], making it an excellent choice to efficiently test the impact of the constantly evolving detector design on flavour tagging.

5.2 Dependence on the quality of particle identification

Several $K^{\pm}$ classification scenarios were defined by fixing the efficiency of misidentification to $\pi^{\pm}$ and varying the $K^{\pm}$ identification efficiency. In addition, the limiting cases of Kaon identification with 0% and 100% efficiencies were considered. These are referred to henceforth as the no $K^{\pm}$ID and the perfect $K^{\pm}$ID scenarios. The considered efficiencies and the misidentification rates are shown in Table 5.

The no $K^{\pm}$ID scenario is used as the reference in this section to assess the impact of adding PID variables as input features for jet flavour tagging. The largest performance gain with the addition of $K^{\pm}$ID information is predictably in the classification of s vs ud jets, shown in Fig. 8. Relative to the reference no $K^{\pm}$ID scenario, with a ϵ_{sig} of 31.6% at a ϵ_{bkg} of 10%, strange tagging efficiency improvements of 11.4%, 25.9%, and 32.9% are evident as the $K^{\pm}$ID efficiency is increased to 60%, 90%, and 95%, respectively. The perfect $K^{\pm}$ID scenario shows the most sizeable performance gain in ϵ_{sig} of 82.9%. This large performance improvement over the 95% $K^{\pm}$ID efficiency with the efficiency of misidentification to $\pi^{\pm}$ of 10% scenario suggests that minimising this misidentification is crucial to tagging strange jets, given their high $\pi^{\pm}$ multiplicity [55].

The performance gain for other forms of classification was marginal, with the exception of c vs ud and u vs d discrimination. For c vs ud, a performance gain of 1.8% from a ϵ_{sig} of 89.3% to 90.9% at a ϵ_{bkg} of 10% is observed while comparing the no $K^{\pm}$ID and the perfect $K^{\pm}$ID scenarios. In the case of u vs d, a 12.5% performance gain from a ϵ_{sig} of 13.6% to 15.3% at a ϵ_{bkg} of 10% is observed.

These results confirm the importance and necessity of particle identification techniques, especially for strange quark studies, as was also noted by some previous studies [15,18,42].

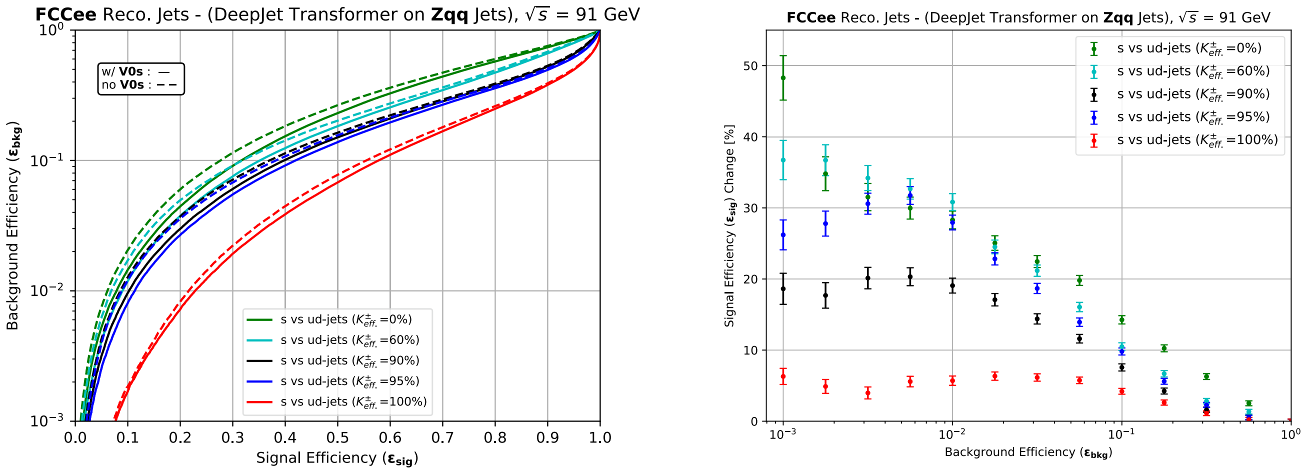
5.3 Dependence on the presence of neutral kaons

As noted earlier, an excess of V^0s, reconstructed K_S^0 and Λ^0, carrying the bulk of the jet momenta is also a distinguishing feature of strange jets and these are expected to be more significant in the scarcity of charged Kaons. The inclusion of V^0 variables, as Fig. 8 shows, results in an improvement of signal efficiency ranging from 14.3% in case of no $K^{\pm}$ID to 4.2% in the case of perfect $K^{\pm}$ID at a background efficiency of 10% for s vs ud discrimination. The percent improvement in signal efficiency for each of the $K^{\pm}$ID scenarios listed in Table 5 is depicted separately in Fig. 8b. This trend proves the importance of V^0s to identify strange jets with low $K^{\pm}$ multiplicities or substandard $K^{\pm}/\pi^{\pm}$ discrimination. The performance gain in other forms of classification was again marginal.

5.4 Importance of variable classes and individual variables

Aiming to estimate the relative importance of a given variable class (e.g. SV variables), the classifier performance was evaluated using the Permutation Feature Importance [98,99] method.

In particular, the variable class under investigation was shuffled amongst all other jets, keeping the rest of the variables unchanged. Specifically, the values for the variable class under investigation were randomly permuted across all jets in the dataset, while the remaining variables for all jets were left unchanged. This disrupts the relationship between the permuted variable and the jet classification, allowing for an estimate of how much the performance of the classifier depends on the given variable. The resulting performance change was considered for discriminating between b- vs c-, c- vs s-, and s- vs ud- jets, compared to the baseline where no variable classes were permuted. Charged jet constituent variables, listed in Table 3, were found to be the most impactful variable class for all types of discrimination at a background efficiency of $\epsilon_{bkg} = 10\%$, as depicted in Table 6. This is presumably due to charged particles being the majority of the reconstructed particles in the jets. SV variables, listed in Table 4, primarily benefited c vs s discrimination, with s vs ud tagging particularly insensitive. Of the remaining three variable classes, V^0 variables and neutral jet constituent variables were found to almost exclusively impact the performance of s vs ud discrimination, with little impact on both b vs c and c vs s discrimination, justifying the inclusion of V^0s for identifying s-jets through conservation of strangeness. Jet-level variables were found to be the least



(a) ROC curves

(b) Percent change in signal efficiency

Fig. 8 The dependence of strange jet tagging performance on the inclusion of V^0s and charged Kaon identification scenarios. **a** ROC curves for s vs ud tagging at the Z resonance at $\sqrt{s} = 91.2$ GeV. Solid lines represent results with the inclusion of V^0s, while dashed lines show the results without them. **b** Percent change in signal efficiency (ϵ_{sig}) with the inclusion of V^0s for s vs ud tagging for each of the $K^{\pm}$ID scenarios listed in Table 5. The axes are swapped with respect to **a** to present the percent change in signal efficiency (ϵ_{sig}) as a function of 12 fixed background efficiencies (ϵ_{bkg})

Table 5 Considered scenarios for $K^{\pm}$ and $\pi^{\pm}$ particle identification performance

$K^{\pm}$ ID efficiency	0%	20%	40%	60%	80%	90%	95%	100%
$\pi^{\pm}$ misID efficiency	0%	10%	10%	10%	10%	10%	10%	0%

significant, marginally impacting s vs ud discrimination, and having virtually no impact on heavy flavour discrimination. Moving to the high purity regime at a background efficiency of $\epsilon_{bkg} = 0.1\%$, primarily the same trends were observed, with the impact of any variable type being amplified. SV variables, in particular, became hugely important to heavy flavour tagging, reaching almost equal in impact to the charged jet constituent variables, proving that the presence and properties of SVs are definitive indicators for identifying heavy flavour jets.

The above studies were repeated to estimate the relative importance of individual variables (e.g. m^{SV}), where rather than shuffling an entire variable class amongst jets, one individual variable was shuffled amongst itself. The 64 variables can be loosely split into the following categories:

- Kinematic ($|p|$, E, $|p|/|p|_{\mathrm{jet}}$, θ, $\Delta\theta$,...)
- PID ($isPhoton$, $K^{\pm}$ID,...)
- Track (D_0, z_0,...)

It was found that, at a background efficiency of 10%, kinematic variables of charged particle constituents, including $\frac{E_{\mathrm{ch.}}}{E_{\mathrm{jet}}}$ and $\frac{|p|_{\mathrm{ch.}}}{|p|_{\mathrm{jet}}}$, were generally impactful, particularly for c vs s discrimination. Track variables, such as D_0/σ_{D_0} and z_0, were the most impactful, though less for b vs c than other types of discrimination, possibly due to their redundant information after the inclusion of SVs. PID variables had little impact on b vs c and c vs s discrimination, but $K^{\pm}$ID and photon ID were the most important for s vs ud discrimination, as was observed earlier. The high purity regime at a background efficiency of 0.1% resulted in similar trends, though with PID variables, including $K^{\pm}$ID and photon ID, decreasing in importance and being somewhat replaced by kinematic ones. It should be stated that the baseline $K^{\pm}$ID scenario, as mentioned in Sect. 5, is deliberately pessimistic, which could account for its decrease in importance. Track variables remained the most impactful. The secondary vertex mass m^{SV} became the most impactful variable in b vs c discrimination at high purity by a sizeable margin, as SV kinematics store essential information about the decaying hadrons. The results of this study are summarised in Table 7 below.

5.5 Dependence on the flavour definition

Defining the flavour of a reconstructed jet is a complex task. Several definitions have been used in past and current experi-

Table 6 Performance decrease in signal efficiency (ϵ_{sig}) after permutation of variable classes defined in Sect. 4.1 for fixed background efficiencies (ϵ_{bkg}) of 10% and 0.1%

Variable class	Jet-level (%)	Charged (%)	Neutral (%)	SV (%)	V^0 (%)
$\epsilon_{bkg} = 10\%$					
b vs c	2.4	62.4	2.2	13.9	0.1
c vs s	1.2	65.7	2.9	29.6	0.2
s vs ud	7.6	59.4	21.8	5.0	16.4
$\epsilon_{bkg} = 0.1\%$					
b vs c	6.6	97.0	8.0	89.9	0.6
c vs s	9.3	96.1	11.0	77.9	0.2
s vs ud	35.9	91.0	57.3	7.4	43.8

Table 7 Performance decrease in signal efficiency (ϵ_{sig}) after permutation of individual variables defined in Sect. 4.1 for fixed background efficiencies (ϵ_{bkg}) of 10% and 0.1%. A set of seven variables, chosen among the most impactful, is presented here

Variable	$\ln(E_{\mathrm{ch.}})$ (%)	`isPhoton` (%)	$K^{\pm}$ID (%)	m^{SV} (%)	$\lvert p\rvert^{\mathrm{V}^0}$ (%)	z_0 (%)	D_0/σ_{D_0} (%)
$\epsilon_{bkg} = 10\%$							
b vs c	3.5	0.3	0.2	3.0	0.1	7.8	11.6
c vs s	23.8	0.7	0.5	0.3	0.2	20.9	39.1
s vs ud	12.8	16.6	38.8	0.0	9.2	23.3	26.7
$\epsilon_{bkg} = 0.1\%$							
b vs c	13.8	1.3	0.9	67.2	0.8	34.1	45.0
c vs s	57.6	0.9	4.8	7.0	0.3	56.2	79.5
s vs ud	35.0	28.0	59.0	0.4	34.7	60.5	80.1

ments to assign the flavour of MC-generated jets. The flavour definition can impact the classifier performance for Z boson decay events because this definition can lead to jets being assigned a different flavour than the original quark.

In the Z boson definition used throughout this work as introduced in Sect. 2.3, the flavour of a jet is defined as the flavour of the quark to which the Z boson decays. The hadronisation and fragmentation of the quark are ignored in this definition. One flavour definition that accounts for fragmentation and hadronisation effects is the Ghost Matching algorithm used at CMS [100], which defines the flavour of a jet by finding the hadrons or partons from the MC history of the jet, clustered with the same jet clustering algorithm as the reconstructed jet.

The largest performance differences of the algorithm after changing flavour definitions can be observed in the discrimination of s-jets vs ud-jets, where the Ghost Matching definition leads to a 11.8% higher tagging efficiency than the Z boson definition at a fixed background efficiency of 10%. Such significant changes in performance make it essential to account for the used flavour definitions while comparing different flavour tagging algorithms.

6 Example of performance: the Z boson at the FCC-ee

The Z boson decays relatively uniformly to the five quark flavours, and none of the decay channels to $q\bar{q}$ pairs are suppressed. Thus, tagging a particular jet flavour entails discrimination against every other flavour. Especially, isolating $Z \to s\bar{s}$ events from the exclusive decays of the Z boson provides a challenging case to tag the s-jets by eliminating both the heavy jets and u-, d-jets. The dominant discriminating variable against the heavy jets is the reconstructed SVs, while it is the presence of a high-momentum strange hadron against u- and d-jets. This makes isolating $Z \to s\bar{s}$ events from the exclusive hadronic decays of the Z boson an ideal metric to assess the performance of `DeepJetTransformer` in the FCC-ee environment and allows for a unique opportunity to access a hitherto scarcely studied channel. Further backgrounds are not considered but are expected to be well under one per cent of the total expected yield around the Z boson resonance.

6.1 Event and jet selection

The $e^+e^- \to Z \to q\bar{q}$ event samples with $q \equiv b, c, (u, d, s)$ described in Sect. 2.3 are used. These are the same samples used to evaluate the performance of `DeepJetTransformer`.

Events are selected if exactly two jets could be reconstructed with their final constituents. Jets with low momentum or jet axes outside the fiducial region of the detector are excluded. An event is selected if both of its jets have a 3-momentum magnitude ($|p|$) greater than 20 GeV and the polar angle (θ) of their jet axes within 14 and 176 degrees. Events are required to have jets of the same MC flavour, defined as the flavour of the quarks to which the Z boson decays.

6.2 Performance and working points

All jets from $Z \to q\bar{q}$ events are independently evaluated using DeepJetTransformer. Discriminants are defined to sequentially remove the heavy flavour background (b- and c-jets) and the light flavour background (u- and d-jets). The s-jets are first tagged to be discriminated from b- and c-jets by defining the discriminant as in Eq. 5.1 with s-jets as signal and b- and c-jets as background. For the jets tagged by introducing a cut on this discriminant, another discriminant is defined to distinguish s-jets from u- and d-jets through the same method. The signal efficiencies after each subsequent cut, corresponding to four working points with increasing purity, are reported in Table 8.

The Z boson resonance is reconstructed from the 4-momentum of the two jets. The reconstructed invariant dijet mass distribution, separated by the MC flavour of the resulting hadronic jets, is shown in Fig. 9a. The hadrons in b-jets tend to have longer decay chains, which causes more momentum to be lost via neutrinos, resulting in a wider invariant mass distribution for $Z \to b\bar{b}$. Similarly, the $Z \to c\bar{c}$ reconstructed invariant mass distribution also shows a tail, but for the lighter flavour jets, s, u, and d, a clear Gaussian peak can be seen at the Z resonance.

These jets are first tagged to remove the background of b- and c-jets by defining the discriminant, as described above. If both jets from a Z boson decay event pass this tagging requirement, they are used to reconstruct the invariant mass. The distribution of this invariant mass is displayed in Fig. 9b, with the contributions of the MC flavours of the jets indicated. The jets from the events passing the anti-b/c tag requirement are subsequently tagged with the s vs ud quark tagger to remove the background of u- and d-jets. Figures 9c and d show the distribution of the reconstructed invariant mass of the Z boson after this additional tag. Both jets of every event are required to pass the tagging requirements in each stage of the selection.

The reconstructed tagged Z resonance in Fig. 9 shows that the $Z \to s\bar{s}$ sample is extremely pure after requiring the two consecutive tags on each jet from Z boson decay events. Table 8 lists events corresponding to an integrated luminosity of 125 ab^{-1}. The discovery significance, Z, in σ, is defined [101] as,

$$Z = \sqrt{2\left[(N_{sig} + N_{bkg})\log\left(1 + \frac{N_{sig}}{N_{bkg}}\right) - N_{sig}\right]}. \quad (6.1)$$

N_{sig} and N_{bkg} refer to the number of signal and background events, respectively. Signal is defined as $Z \to s\bar{s}$ events while the background is composed of $Z \to q\bar{q}$ (all quarks but s quarks) events. It is apparent that all four working points are significantly above the canonical discovery significance of 5σ. It is important to realise that machine backgrounds and irreducible backgrounds from other standard model processes are not considered in this study, and are at the per cent level. However, the remarkable sensitivity warrants investigation of how limited the integrated luminosity can be to observe $Z \to s\bar{s}$ in the considered scenario.

Figure 10 shows the discovery significance of the process $Z \to s\bar{s}$, under the background-free scenario, as a function of integrated luminosity. The corresponding values of N_{sig} and N_{bkg} at each working point can be referred to from Table 8. It can be seen that a 5σ significance can be achieved with minuscule luminosities compared to the FCC-ee run plan, even at the tightest working point. For WP3, corresponding to Fig. 9d, a 5σ significance can be reached with a luminosity of 60 nb^{-1}, equivalent to *less than a second* of the FCC-ee run at the Z resonance.

Data-to-simulation scale factors for b-jets can be measured with a precision of approximately ±2.5% for jets with $30 < p_{\mathrm{T}} < 50$ GeV at the LHC experiments. Tagging algorithms at the future colliders are expected to achieve smaller uncertainties.

These findings will open up avenues at FCC-ee for measurements that require ultra-pure $Z \to q\bar{q}$ samples, at least for the three heaviest flavours to which the Z boson decays. Some examples are vector and axial couplings of the Z to up- and down-type quarks and possibly even individual quark flavours and asymmetry parameters of the Z boson in the hadronic decay channels. LEP and SLD performed comprehensive measurements of the forward-backwards charge asymmetry for $e^+e^- \to b\bar{b}$ [54], similar precise measurements for the charm and the strange quark, and possibly the u, d quarks, will become feasible at the FCC-ee.

7 Summary

The transformer-based model presented in this work can be trained considerably more quickly than the state-of-the-art graph neural network-based taggers [40,43]. The discrimination power of this framework called DeepJetTransformer is presented for FCC-ee, allowing the classification of all jet flavours in e^+e^- collisions at the Z resonance.

Table 8 Presented are the efficiencies to select s quark jets and the mistag rate for other flavours at four different working points. Also listed are the expected yields calculated for an integrated luminosity of 125 ab^{-1}. Signal is defined as $Z \rightarrow s\bar{s}$ events while the background is composed of $Z \rightarrow q\bar{q}$ (all quarks but s quarks) events. The number of observed events is significantly above the canonical discovery significance of five standard deviations for all selections

	Mistag rate [%]	Efficiency [%]	N_{sig}	N_{bkg}
WP1				
s vs bc	10	98.93 ± 0.03	7.35×10^{11}	1.35×10^{12}
s vs ud	10	40.03 ± 0.04	1.45×10^{11}	3.25×10^{10}
WP2				
s vs bc	1	54.18 ± 0.04	2.38×10^{11}	2.06×10^{11}
s vs ud	10	39.28 ± 0.06	5.10×10^{10}	5.57×10^{9}
WP3				
s vs bc	1	54.18 ± 0.04	2.38×10^{11}	2.06×10^{11}
s vs ud	1	10.05 ± 0.11	1.12×10^{10}	4.77×10^{8}
WP4				
s vs bc	0.1	17.96 ± 0.06	3.23×10^{10}	6.98×10^{9}
s vs ud	0.1	1.98 ± 0.33	3.56×10^{8}	3.38×10^{6}

A tagging efficiency for b-jets of about 99%(86%) can be achieved against s-, u-, and d-jets (c-jets) at a background efficiency of 0.1%, pointing to an excellent b-jet discrimination, dominantly owing to the secondary vertex reconstruction coming from the expected excellent detector resolution. A c-jet tagging efficiency of about 90%(70%) can be achieved when discriminating from b-jets, at a background efficiency of 10%(1%).

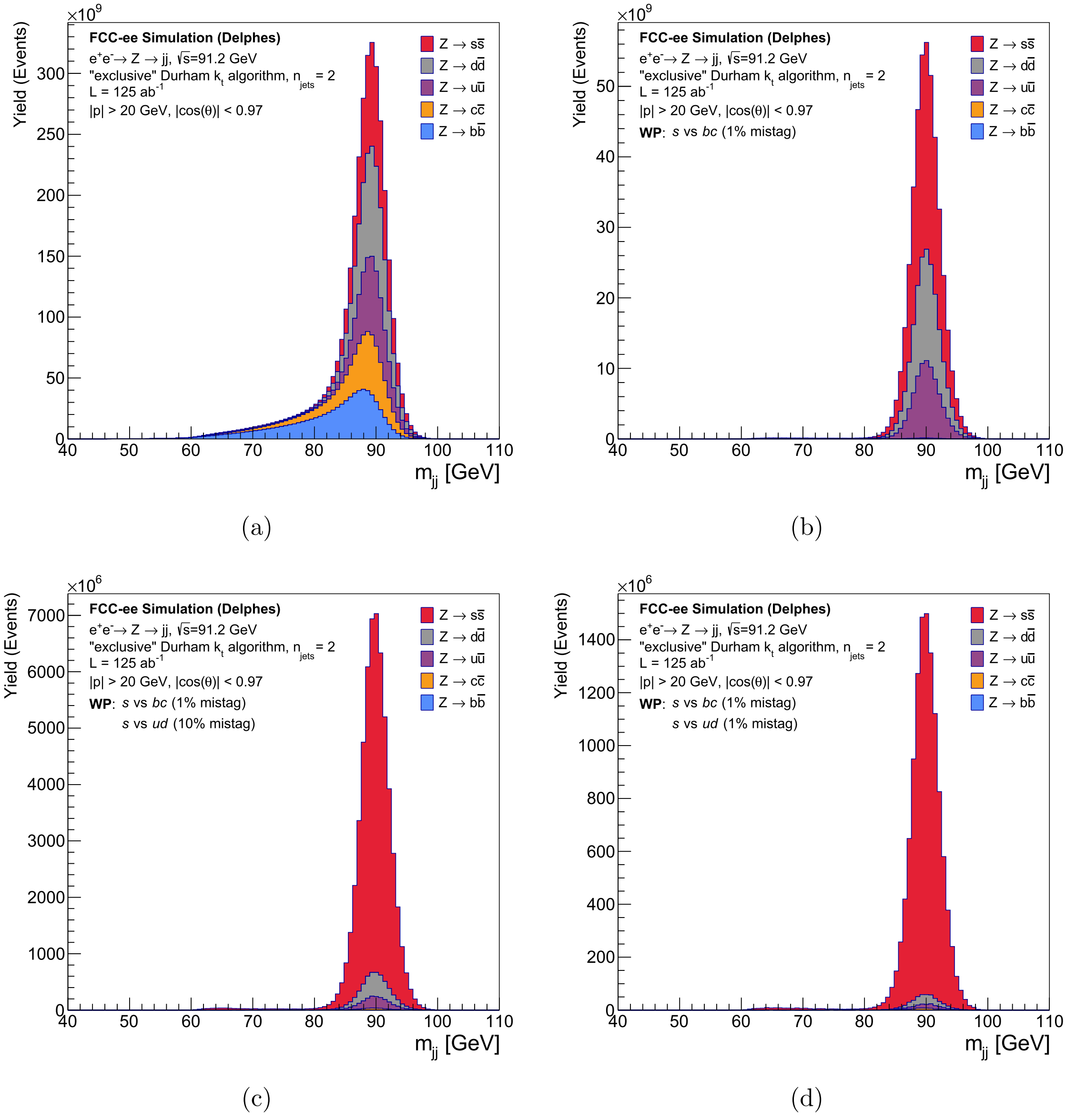
Excellent discrimination can be achieved for s-quark tagging against the b- and c-quark jet background. Against the most challenging background of ud-jets, a 40% efficiency for s quark jets can be achieved at a background efficiency of 10%. This performance is partially attributed to the inclusion of V^{0}s. Further significant performance enhancement in strange tagging is seen when $K^{\pm}/\pi^{\pm}$ discrimination is included. Minor discrimination can even be achieved between u- and d-jets. The $Z \rightarrow s\bar{s}$ process can be efficiently isolated from other hadronic decays of the Z boson, and an extremely pure $Z \rightarrow s\bar{s}$ sample can be obtained.

8 Outlook

The current input feature set is likely far from optimal and could be extended to incorporate further parameters, including those related to jet-shape variables or the full covariance matrix. A primary focus would be to include more realistic PID assumptions based on a specific detector scenario. In Ref. [42], for instance, the mass calculated from the time-of-flight ($m_{t.o.f.}$) and the number of primary ionisation clusters along the track (dN/dx) are directly fed as inputs to the NN. On the other hand, it is also evident from the feature importance studies that there is some overlap in the current feature set, which could likely be reduced with marginal impact on the discriminative performance, thus lowering computational complexity if paired with a simplified architecture.

There is also significant room for hyperparameter tuning. The used batch size of 4000 is comparatively large, with typical values being less than 1024. The large batch size was chosen for training stability but has been shown to potentially lead to poorer generalisation. The chosen number of training jets of $O(10^6)$ can be considered a rough lower bound given the number of parameters in the network $\sim 10^6$. A natural next step would be to train the network on a much larger number of jets. Further improvements in the network architecture, including adjustments to layer parameters and network structure, are likely possible, though this was not explored in the context of these studies.

Subdividing jet flavours into categories with unique signatures, such as b-jets into those that decay hadronically and semi-leptonically, or $g \rightarrow b\bar{b}$ splittings that do not resemble the typical radiation pattern of a gluon jet, is likely to improve discrimination performance. Additional categories could likewise be included for anti-quarks, which would be helpful in discriminating dijet events where a quark-antiquark pair is expected, such as in $Z \rightarrow s\bar{s}$ decays. More generally, much could be gained from event-level tagging, particularly for s quark jets, where discrimination comes primarily from a high-momentum Kaon. Tagging an entire event could require not only a high-momentum Kaon in one jet but also an oppositely-charged high-momentum Kaon in the other, thus discriminating against Kaons produced during the dressing of a u or d quark, which will not have an oppositely-charged high-momentum Kaon in the other hemisphere.



(a)

(b)

(c)

(d)

Fig. 9 The reconstructed invariant mass of the dijet system before and after tagging both jets with `DeepJetTransformer`, corresponding to WP2 and WP3 in Table 8, for an assumed integrated luminosity of 125 ab^{-1}. Both jets are required to be tagged in each case. Shown are **a** the distribution without tagging applied, **b** after the rejection of b- and c-jets vs s-jets at 1% mistag rate, **c** the distribution after rejection of b- and c-jets at 1% and u- and d-jets vs s-jets at 10% mistag rate, **d** the distribution after rejection of b- and c-jets at 1% and u- and d-jets vs s-jets at 1% mistag rate

The updated design of the IDEA detector concept has the innermost layer of the vertex detector at 1.2 mm instead of 1.7 mm. It will improve the impact parameter resolution and, consequently, the displaced vertex resolutions, thus enhancing the performance of heavy flavour tagging. Further improvement is expected from an ultra-light ALICE ITS3-like vertex detector [102]. An updated version of CLD [70] is being developed with a dedicated compact-RICH PID detector, ARC, which is expected to aid in strange tagging.

A natural extension of isolating $Z \rightarrow s\bar{s}$ events would be to measure the branching fraction and coupling of the Z boson to the s quark and assess further flavour-dependent properties at the Z pole that are sensitive to extensions of the standard model. Extrapolating the excellent performance of `DeepJetTransformer` in discriminating strange jets and the continuing improvement of jet flavour taggers along with more sophisticated inputs, there is clear potential for the precise study of the light u and d quarks at the Z resonance at the FCC-ee.

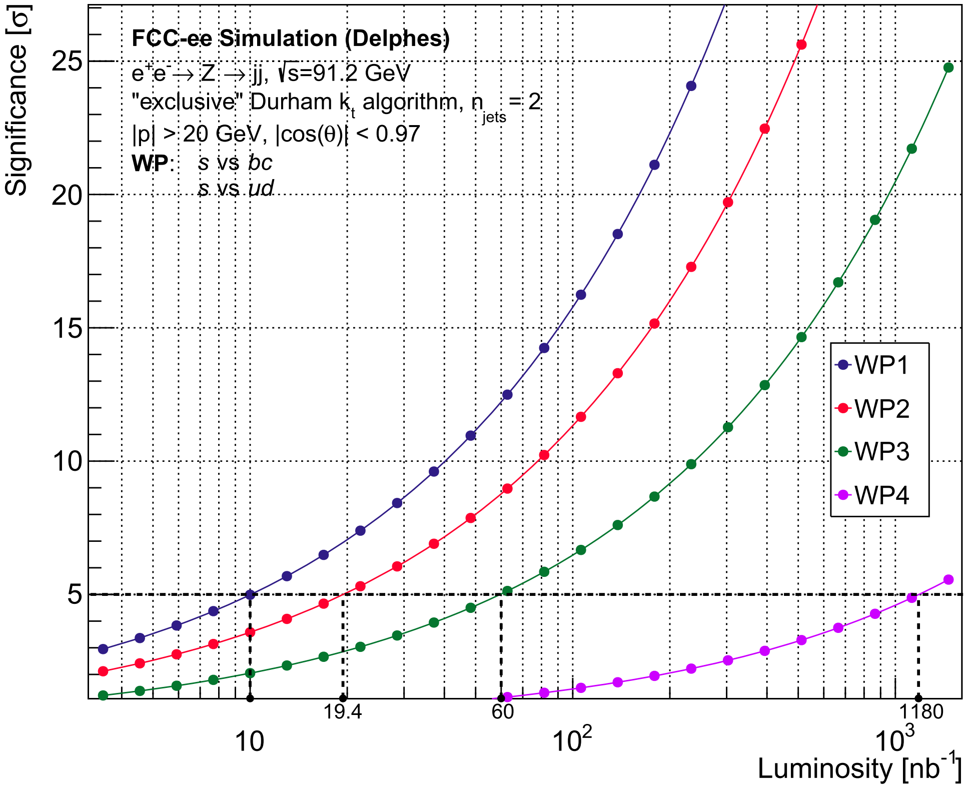


Fig. 10 Discovery significance vs luminosity for the four working points corresponding to Table 8. The points noted by the x axis intersecting with the dashed vertical lines are the luminosities required at the four respective working points to achieve the canonical discovery significance of 5σ

The similar performance in Higgsstrahlung events suggests the opportunity to measure the Yukawa coupling of the s quark, as attempted in Refs. [18,67], and the decent gluon discrimination, especially against heavy quarks, will make gluon final states accessible as well. The much larger Z boson cross-section will also provide opportunities for calibration and performance validation on data before the Higgs boson decay to s quarks is examined, which is likely to reduce experimental uncertainties.

9 Conclusion

Deep learning techniques have demonstrated excellent performance in analysing complex jet structures and extracting subtle flavour signatures in jet flavour identification. The short training time of `DeepJetTransformer` makes it uniquely suited for prospective studies of the developing detector concepts. It should be noted that even though this study focuses on FCC-ee and the IDEA detector, the conclusions are general, and `DeepJetTransformer` can also be utilised at other collider projects with appropriate adjustments, such as tuning to different detector geometries, jet clustering algorithms, or energy regimes.

These results show that modern jet flavour tagging techniques can isolate very pure samples of s quark decays originating from vector bosons. We hope that strange jet tagging will create opportunities for a new category of potential studies at future lepton colliders, including assessment of the feasibility of completely new or more precise measurements and enhancement of the sensitivity to new physics phenomena.

Acknowledgements We want to thank our CMS colleagues at the IIHE in Brussels, especially A.R. Sahasransu and Lode Vanhecke for their preparatory work, and Emil Bols for valuable discussions regarding `DeepJetTransformer`. We would also like to thank Kyle Cormier at the UZH for helpful discussions regarding the feature importance studies. We are grateful to Frank Gaede, Loukas Gouskos, and Michele Selvaggi for their feedback on the manuscript. This project is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 951754. Kunal Gautam and Eduardo Ploerer are supported by FWO (Belgium, Grant no. G0E2221N) and SNF (Switzerland, Grant no. 200021L). Freya Blekman acknowledges support from DESY (Hamburg, Germany), a member of the Helmholtz Association HGF, and support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2121 "Quantum Universe"—390833306. Armin Ilg is supported by SNF in Switzerland. Funding was supported by Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung (200021L), Fonds Wetenschappelijk Onderzoek (Grant no. G0E2221N).

Author contributions KG implemented and AI tested the vertex reconstruction algorithm. ADM designed and implemented the `DeepJetTransformer` architecture. EP adopted the model for the FCC-ee environment, performed input feature selection, and trained and evaluated the performance for several scenarios. KG assessed the classifier performance by isolating $Z \rightarrow s\bar{s}$ events from the exclusive $Z \rightarrow q\bar{q}$ decays. FB and AI supervised and reviewed the work throughout the study. FB, ADM, KG, AI, and EP contributed to the writing and editing of this paper.

Data Availability Statement Data will be made available on reasonable request. [Author's comment: The datasets generated during and/or analysed during the current study are not publicly available due to their large size and management by the FCC Collaboration, which requires appropriate permissions for access but are available from the corresponding author on reasonable request. Additionally, sufficient information has been provided to enable data reproduction.]

Code Availability Statement This manuscript has associated code/software in a data repository. [Author's comment: Github repository: https://github.com/Edler1/DeepJetFCC/tree/master/docs.]

Funded by SCOAP^3.

References

1. S.L. Glashow, The renormalizability of vector meson interactions. Nucl. Phys. **10**, 107 (1959). https://doi.org/10.1016/0029-5582(59)90196-8
2. A. Salam, J.C. Ward, Weak and electromagnetic interactions. Nuovo Cim. **11**, 568 (1959). https://doi.org/10.1007/BF02726525
3. S. Weinberg, A model of leptons. Phys. Rev. Lett. **19**, 1264–1266 (1967). https://doi.org/10.1103/physrevlett.19.1264
4. G. 't Hooft, M. Veltman, Regularization and renormalization of gauge fields. Nucl. Phys. B **44**, 189–213 (1972). https://doi.org/10.1016/0550-3213(72)90279-9
5. G. Aad, T. Abajyan, B. Abbott, J. Abdallah, S. Abdel Khalek, A. Abdelalim et al., Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. Phys. Lett. B **716**, 1 (2012). https://doi.org/10.1016/j.physletb.2012.08.020. arXiv:1207.7214
6. S. Chatrchyan, V. Khachatryan, A. Sirunyan, A. Tumasyan, W. Adam, E. Aguilo et al., Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. Phys. Lett. B **716**, 30 (2012). https://doi.org/10.1016/j.physletb.2012.08.021. arXiv:1207.7235
7. O.S. Brüning, P. Collier, P. Lebrun, S. Myers, R. Ostojic, J. Poole et al., LHC Design Report, CERN Yellow Reports: Monographs (CERN, Geneva, 2004). https://doi.org/10.5170/CERN-2004-003-V-1
8. M. Benedikt, A. Blondel, O. Brunner, M. Capeans Garrido, F. Cerutti, J. Gutleber et al., Fcc-ee: the lepton collider: future circular collider conceptual design report volume 2. Eur. Phys. J. Spec. Top (2018). https://doi.org/10.1140/epjst/e2019-900045-4
9. C. Adolphsen, M. Barone, B. Barish, K. Buesser, P. Burrows, J. Carwardine et al., The International Linear Collider Technical Design Report (CERN, Geneva, June 2013). arXiv:1306.6328
10. The CEPC Study Group, CEPC Conceptual Design Report: Volume 2—Physics & Detector. arXiv:1811.10545
11. L. Linssen, A. Miyamoto, M. Stanitzki, H. Weerts, eds., Physics and Detectors at CLIC: CLIC Conceptual Design Report. CERN, 2, 2012. https://doi.org/10.5170/CERN-2012-003. arXiv:1202.5940
12. The European Strategy Group, Deliberation document on the 2020 Update of the European Strategy for Particle Physics, (Geneva), 2020, CDS
13. J. de Blas, M. Cepeda, J. D'Hondt, R. Ellis, C. Grojean, B. Heinemann et al., Higgs Boson studies at future particle colliders. JHEP (2020). https://doi.org/10.1007/jhep01(2020)139. arXiv:1905.03764
14. D. d'Enterria, Higgs physics at the Future Circular Collider. PoS **ICHEP2016**, 434 (2017). https://doi.org/10.22323/1.282.0434. arXiv:1701.02663
15. P. Azzi, L. Gouskos, M. Selvaggi, F. Simon, Higgs and top physics reconstruction challenges and opportunities at FCC-ee. Eur. Phys. J. Plus (2021). https://doi.org/10.1140/epjp/s13360-021-02223-z. arXiv:2107.05003
16. F. An, Y. Bai, C. Chen, X. Chen, Z. Chen, J.G. da Costa et al., Precision Higgs physics at the CEPC. Chin. Phys. C **43**, 043002 (2019). https://doi.org/10.1088/1674-1137/43/4/043002. arXiv:1810.09037
17. D.M. Asner et al., ILC Higgs White Paper, in *Snowmass 2013: Snowmass on the Mississippi, 10* (2013). arXiv:1310.0763
18. A. Albert, M.J. Basso, S.K. Bright-Thonney, V.M.M. Cairo, C. Damerell, D. Egana-Ugrinovic et al., Strange quark as a probe for new physics in the Higgs sector. arXiv:2203.07535
19. H. Abramowicz, A. Abusleme, K. Afanaciev, N.A. Tehrani, C. Balázs, Y. Benhammou et al., Higgs physics at the CLIC electron-positron linear collider. Eur. Phys. J. C (2017). https://doi.org/10.1140/epjc/s10052-017-4968-5. arXiv:1608.07538
20. DELPHI Collaboration, b-tagging in DELPHI at LEP, Eur. Phys. J. C **32**, 185 (2004). https://doi.org/10.1140/epjc/s2003-01441-8. arXiv:hep-ex/0311003
21. J. Proriol, A. Falvard, P. Henrard, J. Jousset, B. Brandl, Tagging B quark events in ALEPH with neural networks: comparison of different methods. Int. J. Neural Syst. **3**, Supp. 267 (1991)
22. V. Abazov, B. Abbott, M. Abolins, B. Acharya, M. Adams, T. Adams et al., b-jet identification in the D0 experiment. Nucl. Instrum. Methods Phys. Res. Sect. A **620**, 490 (2010). https://doi.org/10.1016/j.nima.2010.03.118. arXiv:1002.4224
23. J. Freeman, T. Junk, M. Kirby, Y. Oksuzian, T. Phillips, F. Snider et al., Introduction to HOBIT, a b-jet identification tagger at the CDF experiment optimized for light Higgs boson searches. Nucl. Instrum. Methods Phys. Res., Sect. A **697**, 64 (2013). https://doi.org/10.1016/j.nima.2012.09.021. arXiv:1205.1812
24. ATLAS Collaboration, ATLAS flavour-tagging algorithms for the LHC Run 2 pp collision dataset. Eur. Phys. J. C **83**, 681 (2023). https://doi.org/10.1140/epjc/s10052-023-11699-1. arXiv:2211.16345
25. CMS Collaboration, Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV. JINST **13**, P05011 (2018). https://doi.org/10.1088/1748-0221/13/05/P05011. arXiv:1712.07158
26. H. Qu, L. Gouskos, ParticleNet: jet tagging via particle clouds. Phys. Rev. D **101**, 056019 (2020). https://doi.org/10.1103/PhysRevD.101.056019. arXiv:1902.08570
27. E. Bols, J. Kieseler, M. Verzetti, M. Stoye, A. Stakia, Jet flavour classification using DeepJet. J. Instrum. **15**, P12012 (2020). https://doi.org/10.1088/1748-0221/15/12/p12012. arXiv:2008.10519
28. ATLAS Collaboration, Performance of b-jet identification in the ATLAS Experiment. JINST **11**, P04008 (2016). https://doi.org/10.1088/1748-0221/11/04/P04008. arXiv:1512.01094
29. ATLAS Collaboration, ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} =$ 13 TeV. Eur. Phys. J. C **79**, 970 (2019). https://doi.org/10.1140/epjc/s10052-019-7450-8. arXiv:1907.05120

30. S. Mondal, L. Mastrolorenzo, Machine learning in high energy physics: a review of heavy-flavor jet tagging at the LHC. Eur. Phys. J. Spec. Top. (2024). https://doi.org/10.1140/epjs/s11734-024-01234-y. arXiv:2404.01071
31. H. Luo, M.-X. Luo, K. Wang, T. Xu, G. Zhu, Quark jet versus gluon jet: fully-connected neural networks with high-level features. Sci. China Phys. Mech. Astron. **62**, 991011 (2019). https://doi.org/10.1007/s11433-019-9390-8. arXiv:1712.03634
32. D. Guest, J. Collado, P. Baldi, S.-C. Hsu, G. Urban, D. Whiteson, Jet flavor classification in high-energy physics with deep neural networks. Phys. Rev. D **94**, 112002 (2016). https://doi.org/10.1103/PhysRevD.94.112002. arXiv:1607.08633
33. P.T. Komiske, E.M. Metodiev, M.D. Schwartz, Deep learning in color: towards automated quark/gluon jet discrimination. JHEP **01**, 110 (2017). https://doi.org/10.1007/JHEP01(2017)110. arXiv:1612.01551
34. ATLAS Collaboration, Quark versus gluon jet tagging using jet images with the ATLAS detector, CDS (2017)
35. V. Mikuni, F. Canelli, ABCNet: an attention-based method for particle tagging. Eur. Phys. J. Plus **135**, 463 (2020). https://doi.org/10.1140/epjp/s13360-020-00497-3. arXiv:2001.05311
36. ATLAS Collaboration, Graph neural network jet flavour tagging with the ATLAS Detector, (Geneva), CDS (2022)
37. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez et al., in *Attention Is All You Need, Advances in Neural Information Processing Systems 30 (NIPS 2017)* (2017). https://doi.org/10.48550/ARXIV.1706.03762. arXiv:1706.03762
38. A. Duperrin, Flavour tagging with graph neural networks with the ATLAS detector. arXiv:2306.04415
39. ATLAS Collaboration, Transformer neural networks for identifying boosted Higgs bosons decaying into $b\bar{b}$ and $c\bar{c}$ in ATLAS, (Geneva), 8, CDS (2023)
40. H. Qu, C. Li, S. Qian, Particle transformer for jet tagging. arXiv:2202.03772
41. F. Bedeschi, A detector concept proposal for a circular e+e- collider. PoS **ICHEP2020**, 819 (2021). https://doi.org/10.22323/1.390.0819
42. F. Bedeschi, L. Gouskos, M. Selvaggi, Jet flavour tagging for future colliders with fast simulation. Eur. Phys. J. C (2022). https://doi.org/10.1140/epjc/s10052-022-10609-1. arXiv:2202.03285
43. D. Parikh, B. Zhang, R. Kannan, V. Prasanna, C. Busart, Performance of graph neural networks for point cloud applications, in *2023 IEEE High Performance Extreme Computing Conference (HPEC)* (2023), pp. 1–7. arXiv:2304.07735
44. UA1 Collaboration, Experimental observation of lepton pairs of invariant mass around 95 GeV/c^2 at the CERN SPS Collider. Phys. Lett. B **126**, 398 (1983). https://doi.org/10.1016/0370-2693(83)90188-0
45. UA2 Collaboration, Evidence for $Z^0 \rightarrow e^+e^-$ at the CERN $\bar{p}p$ collider. Phys. Lett. B **129**, 130 (1983). https://doi.org/10.1016/0370-2693(83)90744-X
46. M. Woods, Review of weak mixing angle results at SLC and LEP, in *International Europhysics Conference on High-energy Physics (HEP 95), CDS* (1995), pp. 31–34, 10
47. M. Dam, Tests of the electroweak theory with the DELPHI detector at LEP, Ph.D. thesis, Oslo Univ.,. CDS (1995)
48. ALEPH Collaboration, Measurement of the Z resonance parameters at LEP. Eur. Phys. J. C **14**, 1 (2000). https://doi.org/10.1007/s100520000319
49. DELPHI Collaboration, Measurement of the mass and width of the Z^0 particle from multi - hadronic final states produced in e^+e^- annihilations. Phys. Lett. B **231**, 539 (1989). https://doi.org/10.1016/0370-2693(89)90706-5
50. OPAL Collaboration, Measurement of the Z^0 mass and width with the OPAL detector at LEP. Phys. Lett. B **231**, 530 (1989). https://doi.org/10.1016/0370-2693(89)90705-3
51. MarkII Collaboration, Initial Measurements of Z-boson resonance parameters in e^+e^- annihilation. Phys. Rev. Lett. **63**, 724 (1989). https://doi.org/10.1103/PhysRevLett.63.724
52. ALEPH, DELPHI, L3, OPAL, SLD Collaborations, Precision electroweak measurements on the Z resonance. Phys. Rep. **427**, 257 (2006). https://doi.org/10.1016/j.physrep.2005.12.006. arXiv:hep-ex/0509008
53. DELPHI Collaboration, Measurement of the strange quark forward-backward asymmetry around the Z^0 peak. Eur. Phys. J. C Part. Fields **14**, 613 (2000). https://doi.org/10.1007/s100520000378
54. Particle Data Group collaboration, Review of particle physics. PTEP **2022**, 083C01 (2022). https://doi.org/10.1093/ptep/ptac097
55. Y. Nakai, D. Shih, S. Thomas, Strange jet tagging. arXiv:2003.09517
56. J. Erdmann, A tagger for strange jets based on tracking information using long short-term memory. J. Instrum. **15**, P01021 (2020). https://doi.org/10.1088/1748-0221/15/01/P01021. arXiv:1907.07505
57. J. Erdmann, O. Nackenhorst, S.V. Zeißner, Maximum performance of strange-jet tagging at hadron colliders. JINST **16**, P08039 (2021). https://doi.org/10.1088/1748-0221/16/08/P08039. arXiv:2011.10736
58. SLD Collaboration, SLD Design Report, 5, (1984), iNSPIRE HEP
59. SLD Collaboration, First direct measurement of the parity-violating coupling of the Z^0 to the s quark. Phys. Rev. Lett. **85**, 5059 (2000). https://doi.org/10.1103/PhysRevLett.85.5059. arXiv:hep-ex/0006019
60. DELPHI Collaboration, Letter of intent: DELPHI detector (DEtector with Lepton Photon + Hadron Identification), (Geneva), CDS (1982)
61. OPAL Collaboration, Letter of intent: OPAL Detector, (Geneva), CDS (1982)
62. G.F. von Dardel, A.H. Walenta, E. Lorenz, P. Duinker, L3: Letter of intent; 1982 edn. (Geneva), CDS (1982)
63. ALEPH Collaboration, Letter of Intent: ALEPH detector—apparatus for LEP PHysics, (Geneva), CDS (1981)
64. C. Lippmann, Particle identification. Nucl. Instrum. Methods Phys. Res. Sect. A **666**, 148 (2012). https://doi.org/10.1016/j.nima.2011.03.009. arXiv:1101.3276
65. G. Cataldi, F. Grancagnolo, S. Spagnolo, Cluster counting in helium based gas mixtures. Nucl. Instrum. Methods A **386**, 458 (1997). https://doi.org/10.1016/S0168-9002(96)01164-3
66. W. Klempt, Review of particle identification by time-of-flight techniques. Nucl. Instrum. Methods A **433**, 542 (1999). https://doi.org/10.1016/S0168-9002(99)00323-X
67. J. Duarte-Campderros, G. Perez, M. Schlaffer, A. Soffer, Probing the Higgs-strange-quark coupling at e^+e^- colliders using light-jet flavor tagging. Phys. Rev. D **101**, 115005 (2020). https://doi.org/10.1103/physrevd.101.115005. arXiv:1811.09636
68. FCC Collaboration, FCC physics opportunities: Future Circular Collider conceptual design report volume 1. Eur. Phys. J. C **79**, 474 (2019). https://doi.org/10.1140/epjc/s10052-019-6904-3
69. B. Auchmann, W. Bartmann, M. Benedikt, J.-P. Burnet, P. Craievich, M. Giovannozzi et al., FCC midterm report, CDS (2024). https://doi.org/10.17181/ZH1GZ-52T41
70. N. Bacchetta, J.J. Blaising, E. Brondolin, M. Dam, D. Dannheim, K. Elsener et al., CLD—a detector concept for the FCC-ee. arXiv:1911.12230
71. B. Francois, Noble liquid calorimetry for a future FCC-ee experiment. Nucl. Instrum. Methods Phys. Res. Sect. A **1040**, 167035 (2022). https://doi.org/10.1016/j.nima.2022.167035
72. M. Selvaggi, F. Bedeschi, M. Ghilardi, FCC-ee IDEA detector Delphes card, GitHub

73. J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens et al., DELPHES 3: a modular framework for fast simulation of a generic collider experiment. JHEP (2014). https://doi.org/10.1007/jhep02(2014)057. arXiv:1307.6346
74. C. Bierlich, S. Chakraborty, N. Desai, L. Gellersen, I. Helenius, P. Ilten et al., A comprehensive guide to the physics and usage of PYTHIA 8.3. arXiv:2203.11601
75. M. Cacciari, G.P. Salam, G. Soyez, FastJet user manual. Eur. Phys. J. C (2012). https://doi.org/10.1140/epjc/s10052-012-1896-2. arXiv:1111.6097
76. S. Catani, Y.L. Dokshitzer, M. Olsson, G. Turnock, B.R. Webber, New clustering algorithm for multi-jet cross-sections in e+ e- annihilation. Phys. Lett. B **269**, 432 (1991). https://doi.org/10.1016/0370-2693(91)90196-W
77. M. Cacciari, G.P. Salam, G. Soyez, The anti-kt jet clustering algorithm. JHEP **2008**, 063–063 (2008). https://doi.org/10.1088/1126-6708/2008/04/063. arXiv:0802.1189
78. T. Suehara, T. Tanabe, LCFIPlus: a framework for jet analysis in linear collider studies. Nucl. Instrum. Methods Phys. Res. Sect. A **808**, 109 (2016). https://doi.org/10.1016/j.nima.2015.11.054. arXiv:1506.08371
79. K. Gautam, Flavour identification techniques, Master's thesis, University of Copenhagen, CDS (2020)
80. C. Helsens, E. Perez, M. Selvaggi, V. Volkl, L. Forthomme, J. Munch Torndal, HEP-FCC/FCCAnalyses: v0.9.0, Zenodo (2024). https://doi.org/10.5281/zenodo.10693709
81. F. Bedeschi, A vertex fitting package, CDS (2024). https://doi.org/10.17181/HVCPV-BK752. arXiv:2409.19326
82. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805
83. D. Buskulic, D. Casper, I. De Bonis, D. Decamp, P. Ghez, C. Goy et al., Performance of the ALEPH detector at LEP. Nucl. Instrum. Methods Phys. Res. Sect. A **360**, 481 (1995). https://doi.org/10.1016/0168-9002(95)00138-7
84. CMS Collaboration, Particle-flow event reconstruction in CMS and performance for jets, Taus, and MET, (Geneva), CERN, 4, CDS (2009)
85. A.J. Larkoski, J. Thaler, W.J. Waalewijn, Gaining (mutual) information about quark/gluon discrimination. JHEP **11**, 129 (2014). https://doi.org/10.1007/JHEP11(2014)129. arXiv:1408.3122
86. C. Caputo, G. Chiarello, A. Corvaglia, F. Cuna, B. D'Anzi, N. De Filippis et al., Particle identification with the cluster counting technique for the IDEA drift chamber. Nucl. Instrum. Methods Phys. Res. Sect. A **1048**, 167969 (2023). https://doi.org/10.1016/j.nima.2022.167969. arXiv:2211.04220
87. G. Wilkinson, Particle identification at FCC-ee. Eur. Phys. J. Plus (2021). https://doi.org/10.1140/epjp/s13360-021-01810-4
88. E. Nakano, Belle PID, Nucl. Instrum. Methods Phys. Res. Sect. A **494**, 402 (2002). https://doi.org/10.1016/S0168-9002(02)01510-3
89. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner et al., An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929
90. H. Xu, L. Xiang, H. Ye, D. Yao, P. Chu, B. Li, Permutation equivariance of transformers and its applications. arXiv:2304.07735
91. K. Fukushima, Visual feature extraction by a multilayered network of analog threshold elements. IEEE Trans. Syst. Sci. Cybern. **5**, 322 (1969). https://doi.org/10.1109/TSSC.1969.300225
92. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan et al., Pytorch: an imperative style, high-performance deep learning library. arXiv:1912.01703
93. M.R. Zhang, J. Lucas, G. Hinton, J. Ba, Lookahead optimizer: k steps forward, 1 step back. arXiv:1907.08610
94. L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao et al., On the variance of the adaptive learning rate and beyond. arXiv:1908.03265
95. K. Gautam, Jet-flavour tagging at FCC-ee, PoS **ICHEP2022**, 1147 (2022). https://doi.org/10.22323/1.414.1147. arXiv:2210.10322
96. R.D. Field, R.P. Feynman, A parametrization of the properties of quark jets. Nucl. Phys. B **136**, 1 (1978). https://doi.org/10.1016/0550-3213(78)90015-9
97. D. Krohn, M.D. Schwartz, T. Lin, W.J. Waalewijn, Jet charge at the LHC. Phys. Rev. Lett. **110**, 212001 (2013). https://doi.org/10.1103/PhysRevLett.110.212001. arXiv:1209.2421
98. L. Breiman, Random forests. Mach. Learn. **45**, 5 (2001). https://doi.org/10.1023/A:1010933404324
99. A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. **20**, 1 (2019). https://doi.org/10.48550/arXiv.1801.01489. arXiv:1801.01489
100. M. Cacciari, G.P. Salam, Pileup subtraction using jet areas. Phys. Lett. B **659**, 119 (2008). https://doi.org/10.1016/j.physletb.2007.09.077. arXiv:0707.1378
101. G. Cowan, E. Gross, Discovery significance with statistical uncertainty in the background estimate, ATLAS Statistics Forum (2008). https://www.pp.rhul.ac.uk/~cowan/stat/notes/SigCalcNote.pdf
102. L. Freitag, Benefits of minimizing the vertex detector material budget at the FCC-ee, Master's thesis, Zurich University, CDS (2023)