Jet Flavour Tagging at FCC-ee with a Transformer-based Neural Network: DeepJetTransformer

Freya Blekman (DESY+UHH), Florencia Canelli (UZH), Alexandre De Moor (VUB), Kunal Gautam (VUB+UZH), Armin Ilg (UZH), Anna Macchiolo (UZH), Eduardo Ploerer (VUB+UZH)

Flavour Tagging

Identification of hadronic final states is an essential to collider experiments

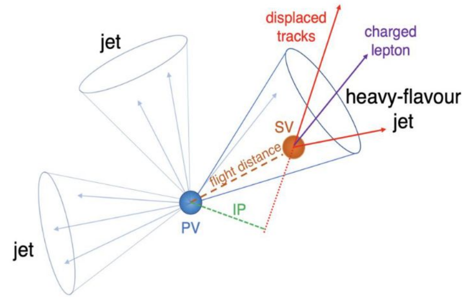
Future lepton collider such as FCC-ee offer much cleaner environment than hadronic collisions (Initial state kinematics known, no PDFs, no QCD ISR, …)

Distinguishing features:

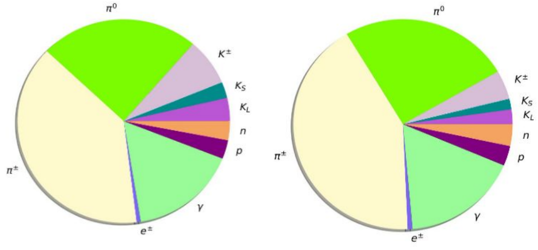
$$C_A/C_F = \frac{9}{4}$$

- Differing colour factors for q vs g
- Displaced SVs for b/c's
- Kaon excess for s
- Jet charge for up/down

$$Q_\kappa = \frac{1}{p_{T,jet}^{\kappa}} \Sigma_j q_j (p_T^j)^\kappa$$



ML has established history for jet-tagging



Strange jets Down jets

Experimental Environment

Spring2021 samples corresponding to

Pythia for event generation

Delphes used for reconstruction assuming IDEA detector concept

Jet clustering performed with exclusive e+e- kT algorithm

Physics process

Z->qqbar

Z(->vv)H(->qqbar)

Experimental Environment

Spring2021 samples corresponding to

Pythia for event generation

Delphes used for reconstruction assuming IDEA detector concept

Jet clustering performed with exclusive e+e- kT algorithm

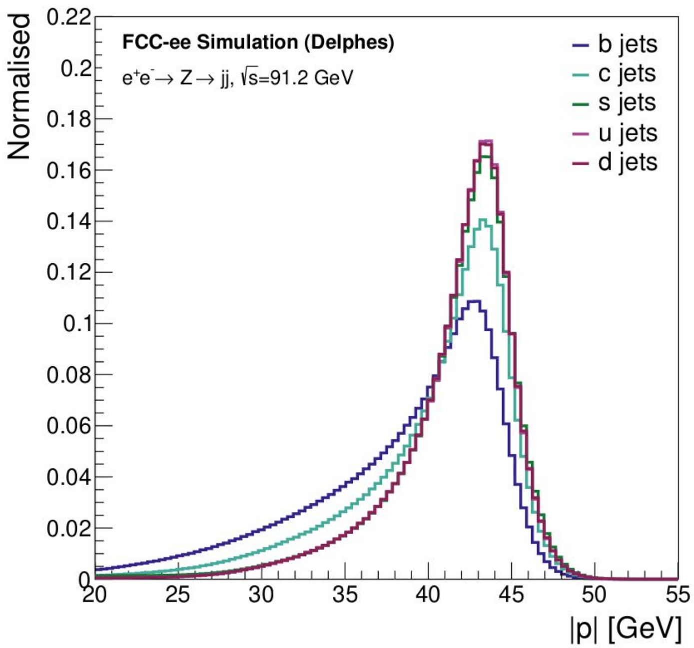
Physics process

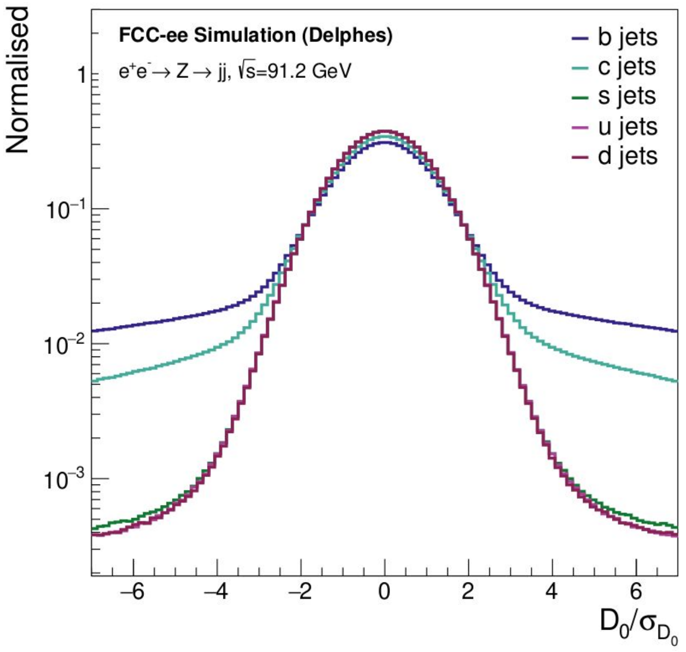
Z->qqbar

Z(->vv)H(->qqbar)

Jet flavour defined via flavour of quarks from decaying Z boson

Experimental Handles

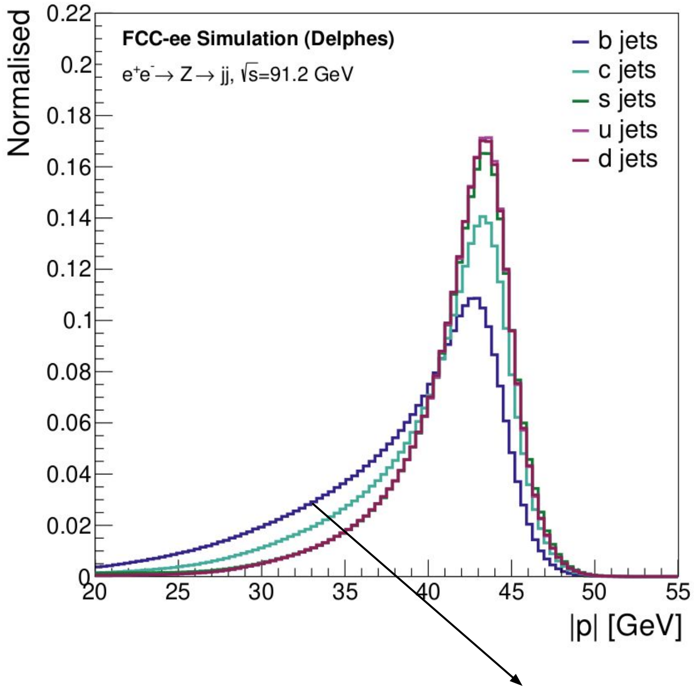


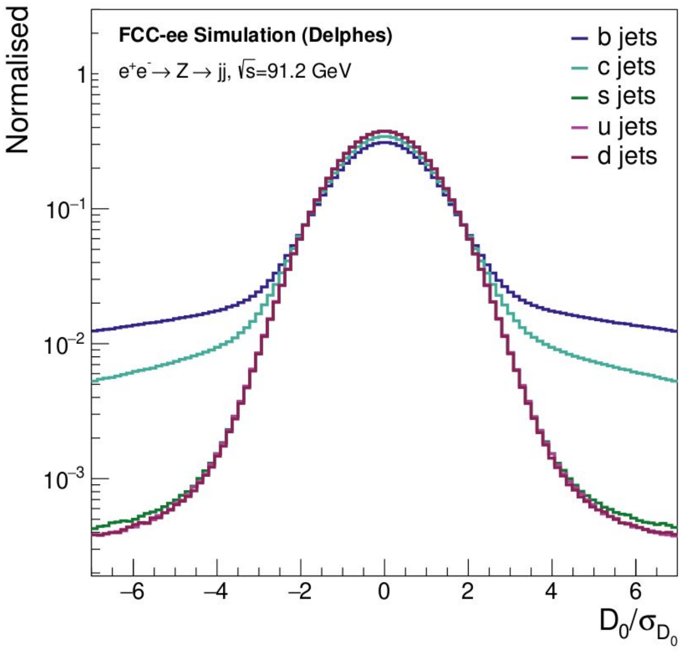


Low-level information already shows distinctions amongst different jets flavours

Can be optimally exploited by ML algorithm

Experimental Handles

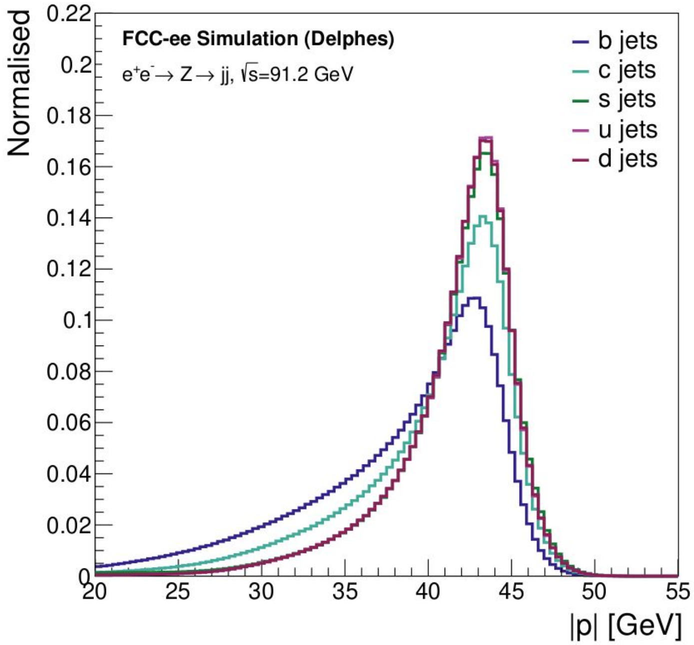


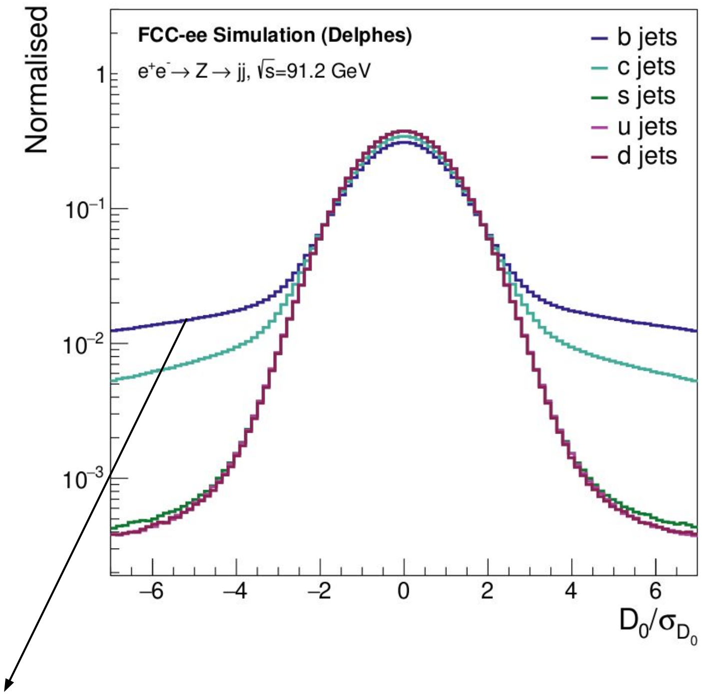


b-jets have much more pronounced tail due to longer decay chain

More momentum can be lost through neutrinos than in light jets

Experimental Handles

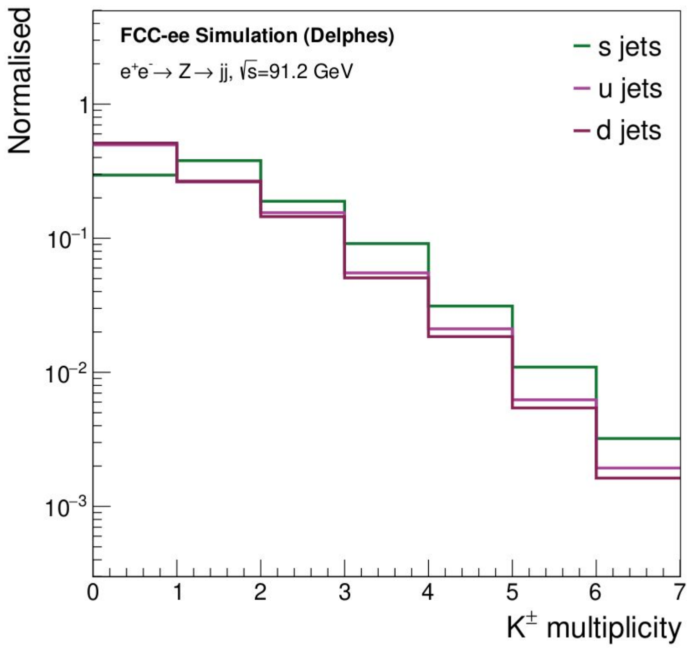


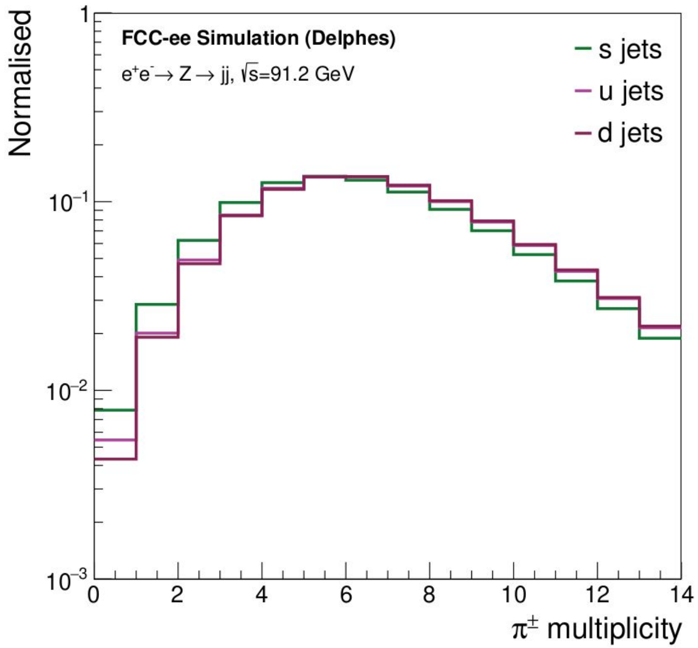


Similarly, decaying B hadrons have longer lifetime than D or light hadrons

Displaced vertices show up in larger transverse impact parameter

Multiplicities

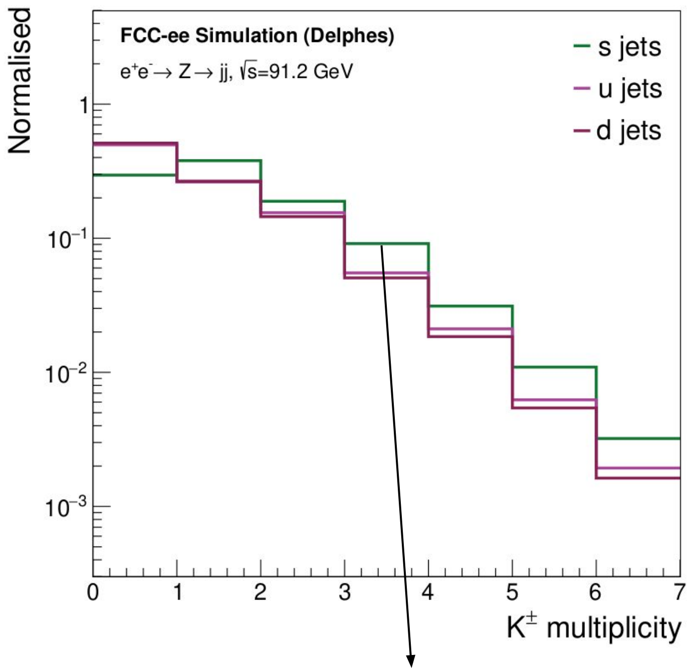


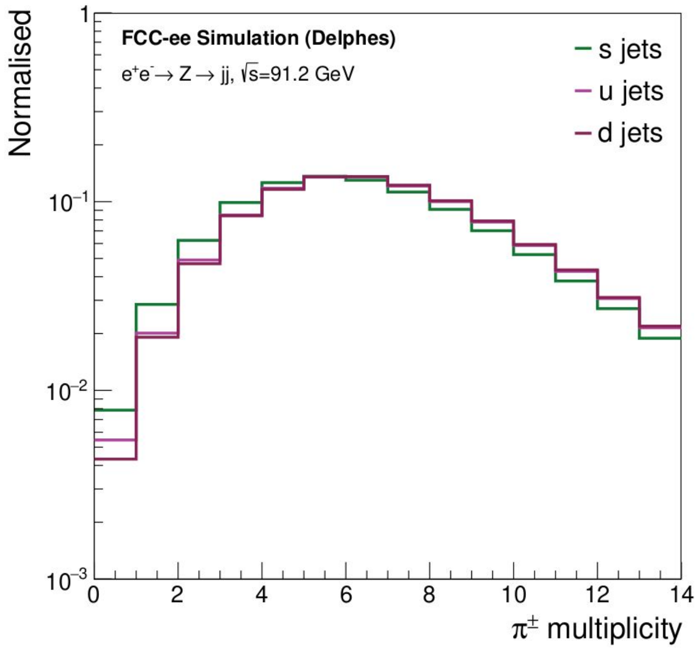


Conservation of strangeness during hardonization of jets shows up as higher Kaon multiplicity for strange jets

Conversely, a lower pion multiplicity

Multiplicities

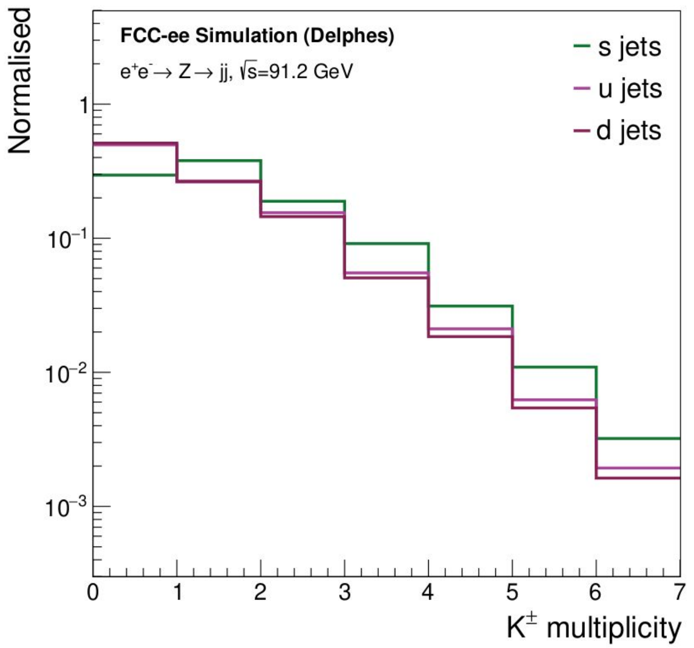


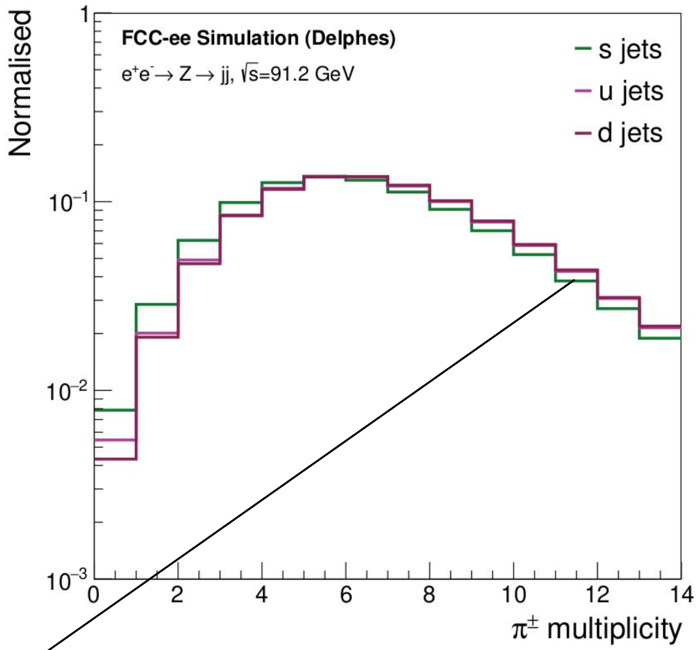


Conservation of strangeness during hardonization of jets shows up as higher Kaon multiplicity for strange jets

Conversely, a lower pion multiplicity

Multiplicities

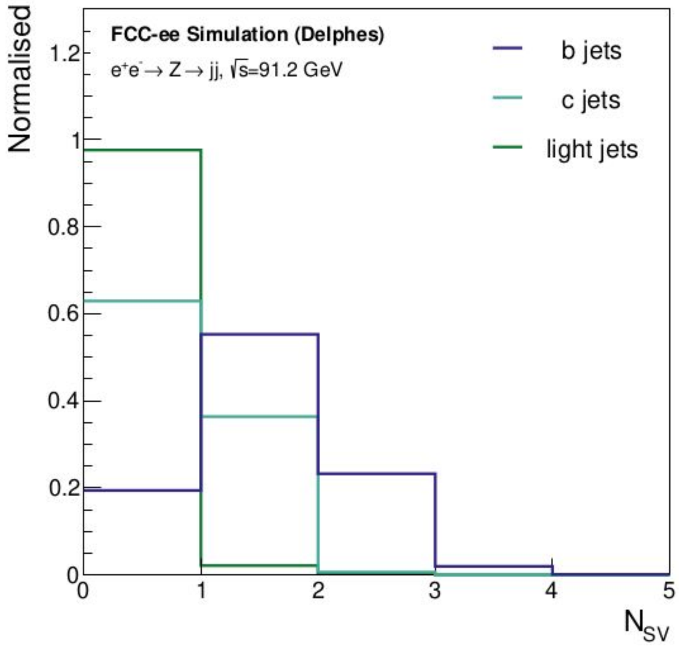


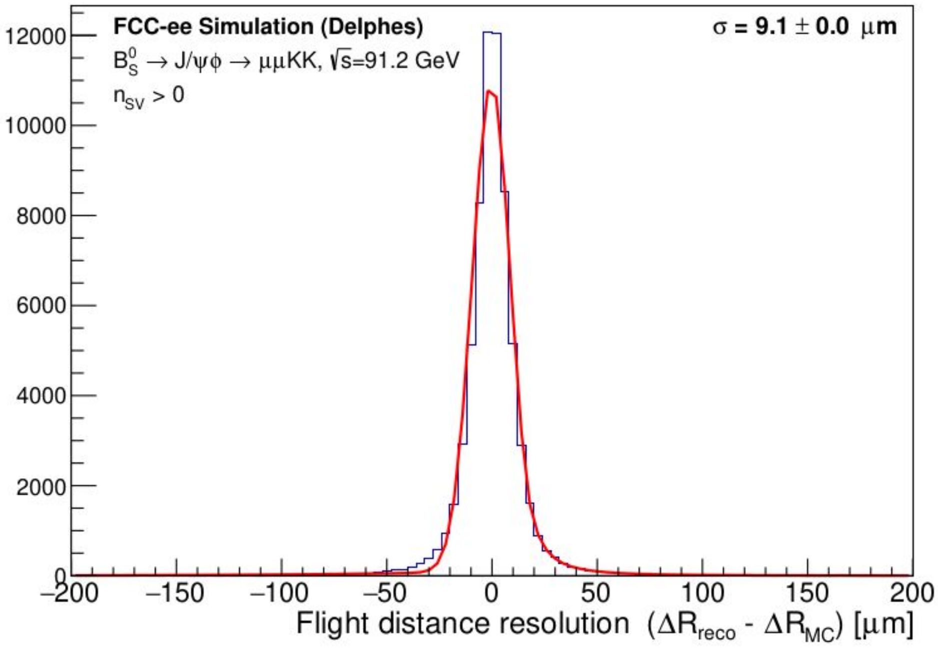


Conservation of strangeness during hardonization of jets shows up as higher Kaon multiplicity for strange jets

Conversely, a lower pion multiplicity

Vertexing

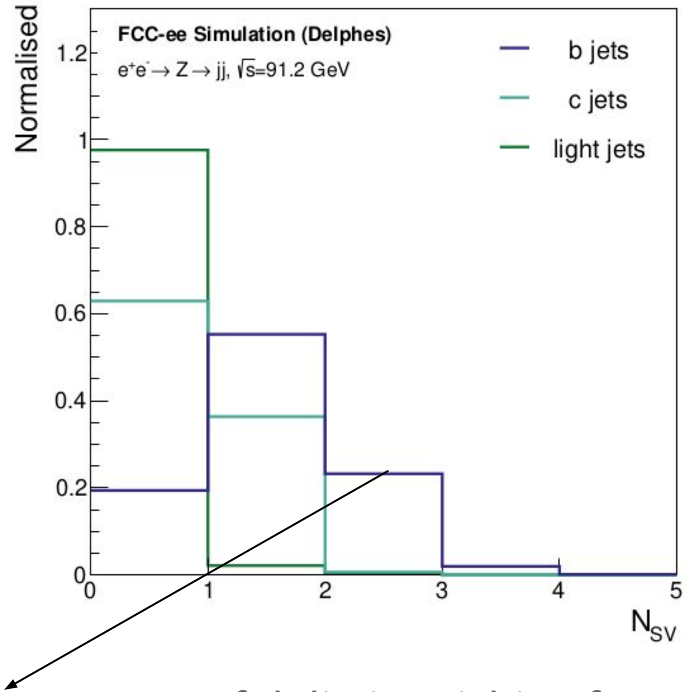


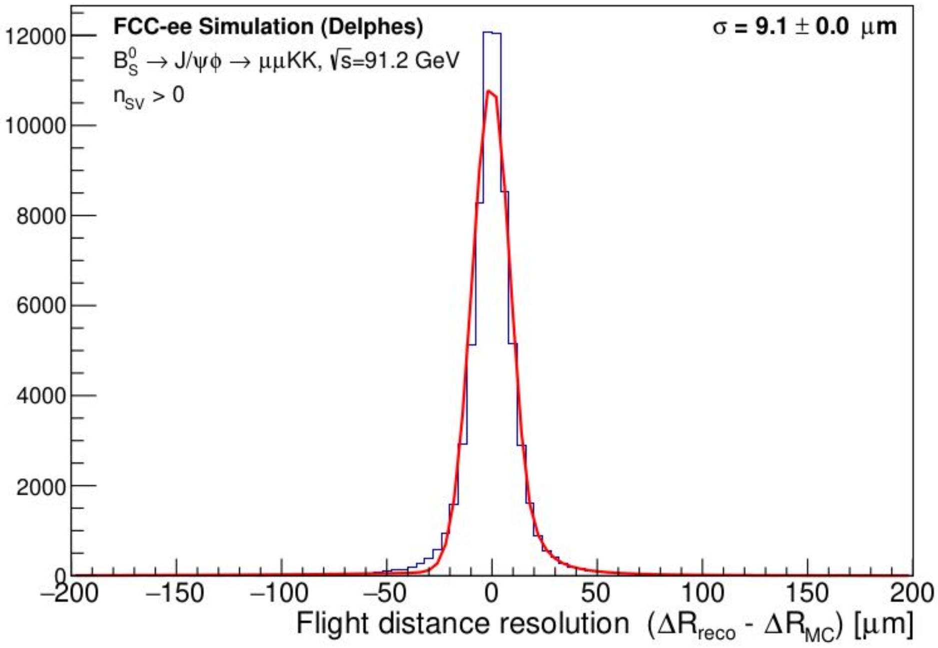


Implemented vertexing algorithm in FCCAnalyses to extract distinguishing features more explicitly

Details can be found here

Vertexing

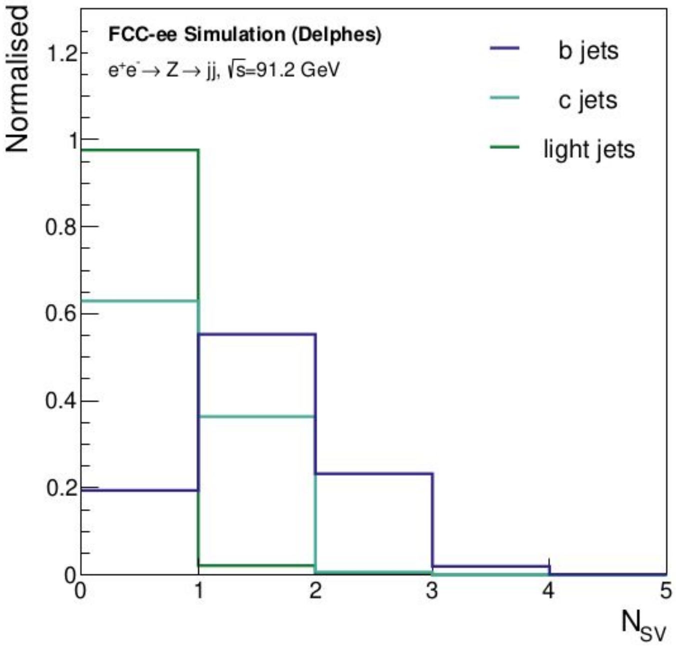


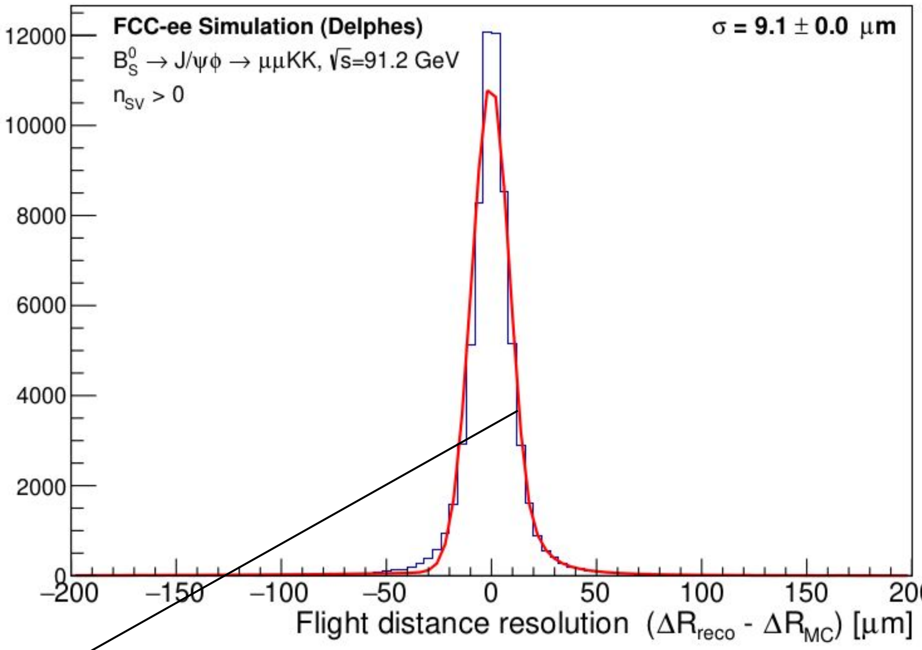


Nsv as a powerful distinguishing feature

Can achieve a resolution of 9 microns in B^0s decays using this reconstruction

Vertexing

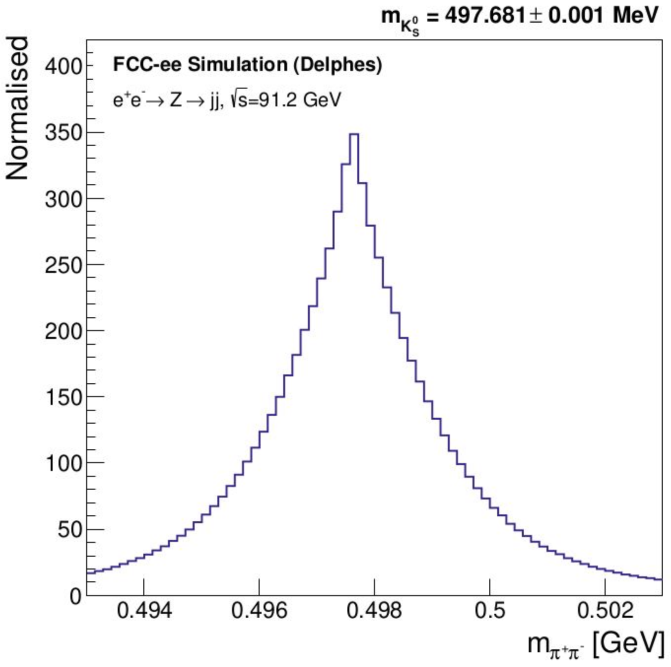


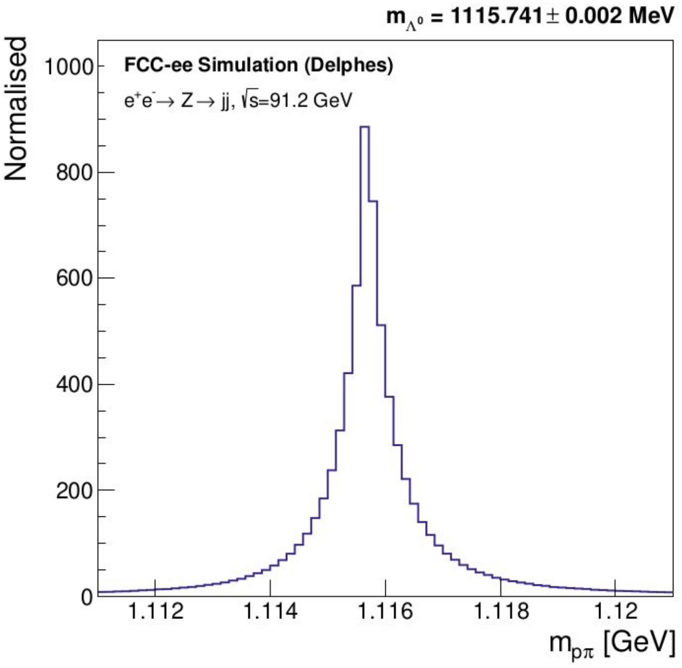


Nsv as a powerful distinguishing feature

Can achieve a resolution of 9 microns in B^0s decays using this reconstruction

V⁰ Reconstruction





Added track PID criterion by considering invariant mass of track pair using different mass hypotheses

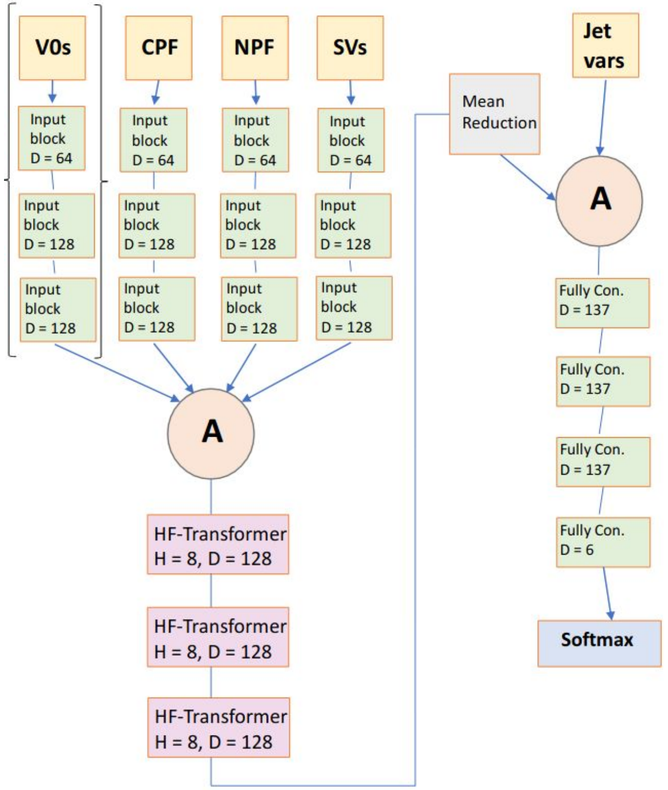
Reconstruct particles carrying strangeness

DeepJetTransformer

DeepJetTransformer is a transformer-based architecture achieving state-of-the-art performance, but using an encoder-decoder architecture

Self-attention allows dynamic assignment of weights to individual elements within the jet capturing intricate dependencies across the entirety of the jet structure

More lightweight/still performant (~1M trainable weights, only 65k per encoder layer)



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

Training

Trained network with 10^6 Z ->qqbar jets (80%/20% train/validation), evenly split into b, c, s, u, d

Implemented in Pytorch (v1.10.1)

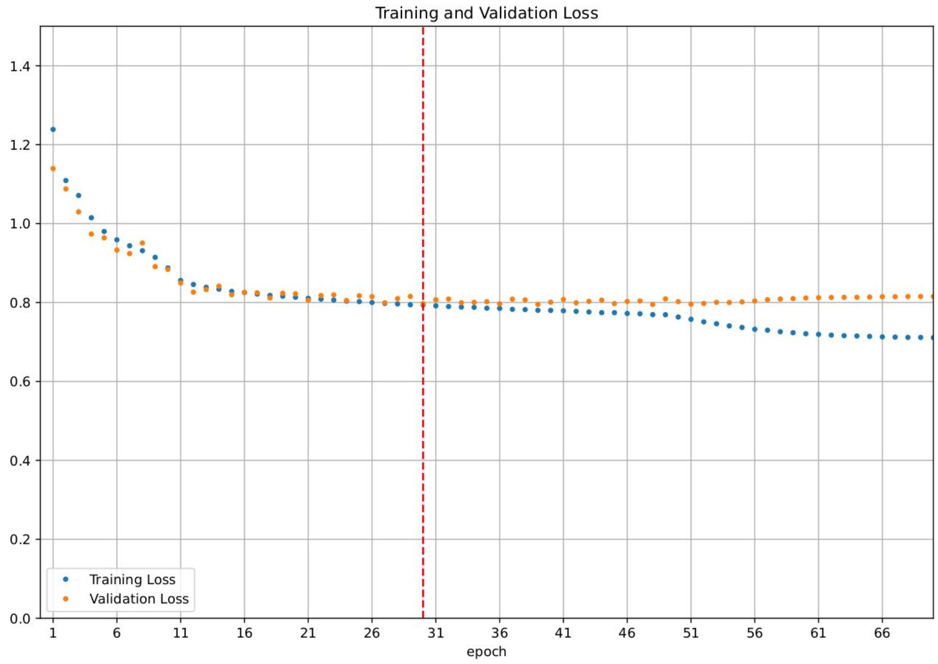
70 epochs w/ batch size of 4000 trained in

~2 hours

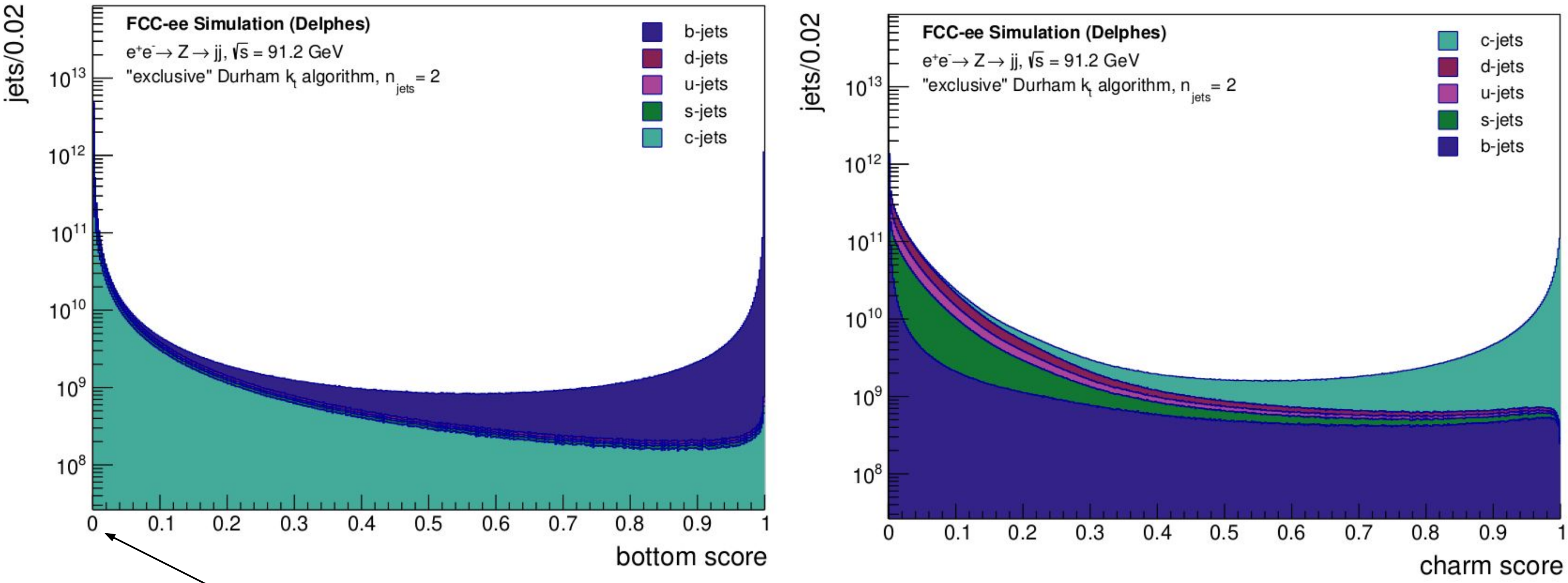
-> No obvious overfitting/overtraining

Categorical cross entropy as loss function

$$L(\mathbf{y}, \mathbf{p}) = -\Sigma_i^C y_i log(p_i)$$



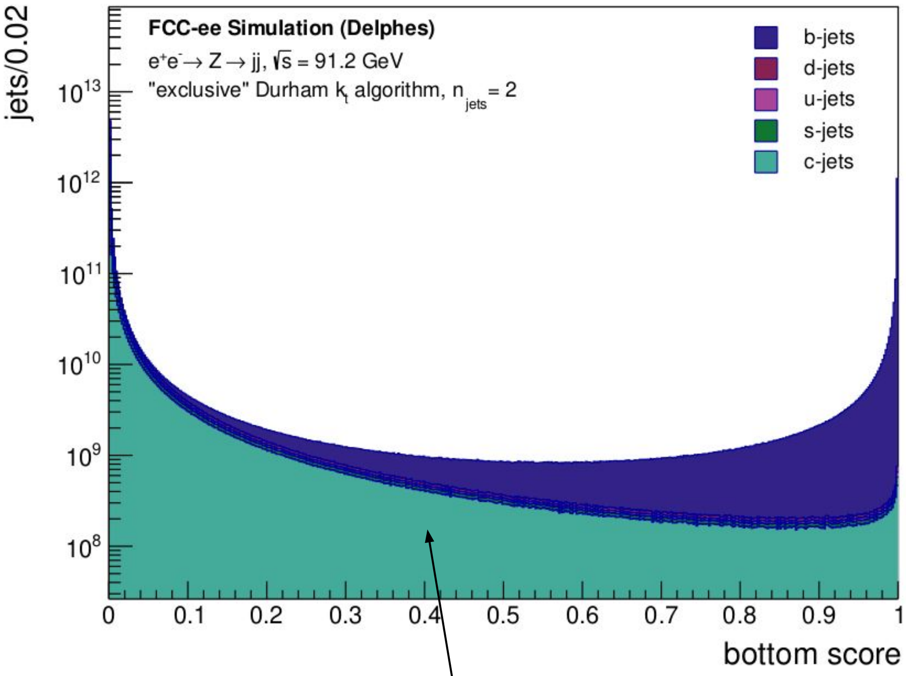
Classifier Distributions: bottom and charm

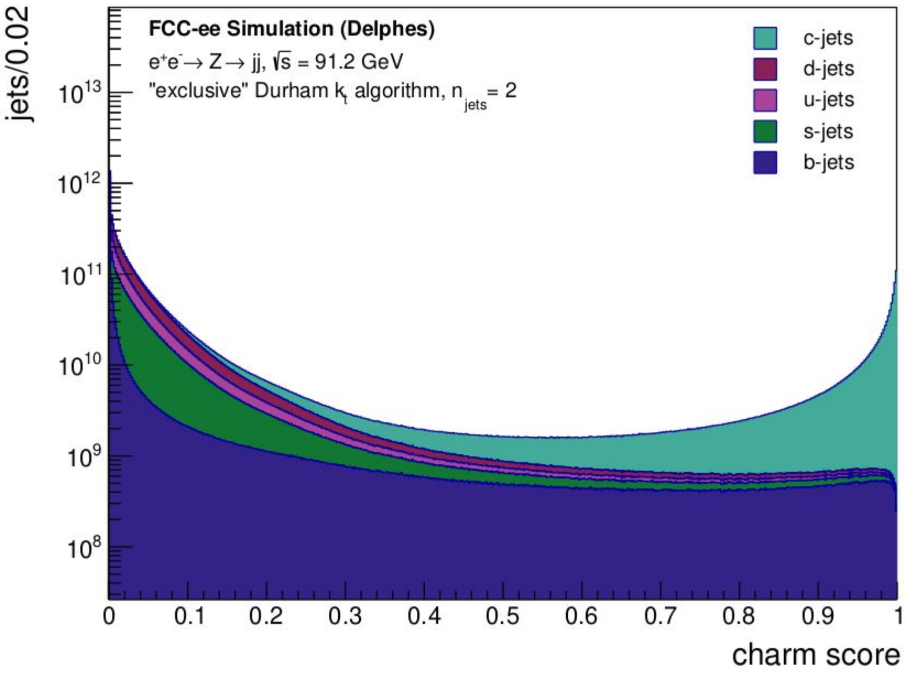


Vast majority of light jets at 0 bin

Charm jets are only significant background to b jets

Classifier Distributions: bottom and charm

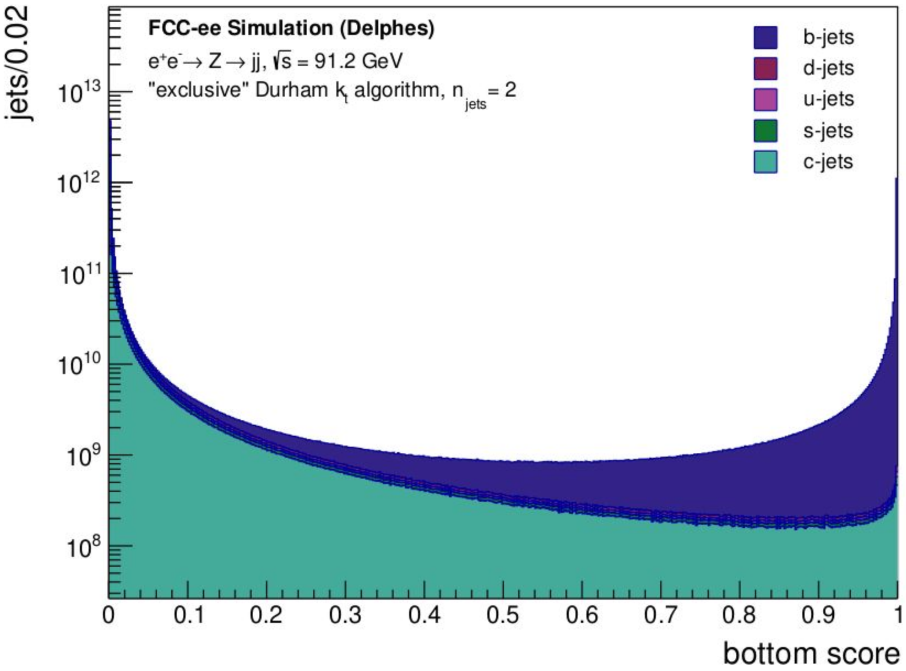


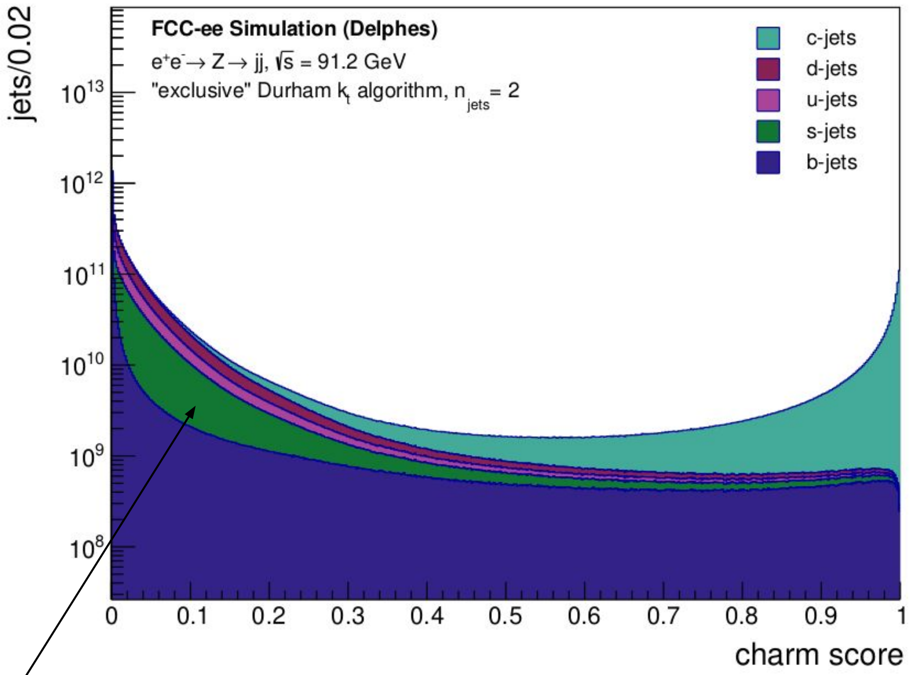


Vast majority of light jets at 0 bin

Charm jets are only significant background to b jets

Classifier Distributions: bottom and charm





Charm score distribution not nearly as pure

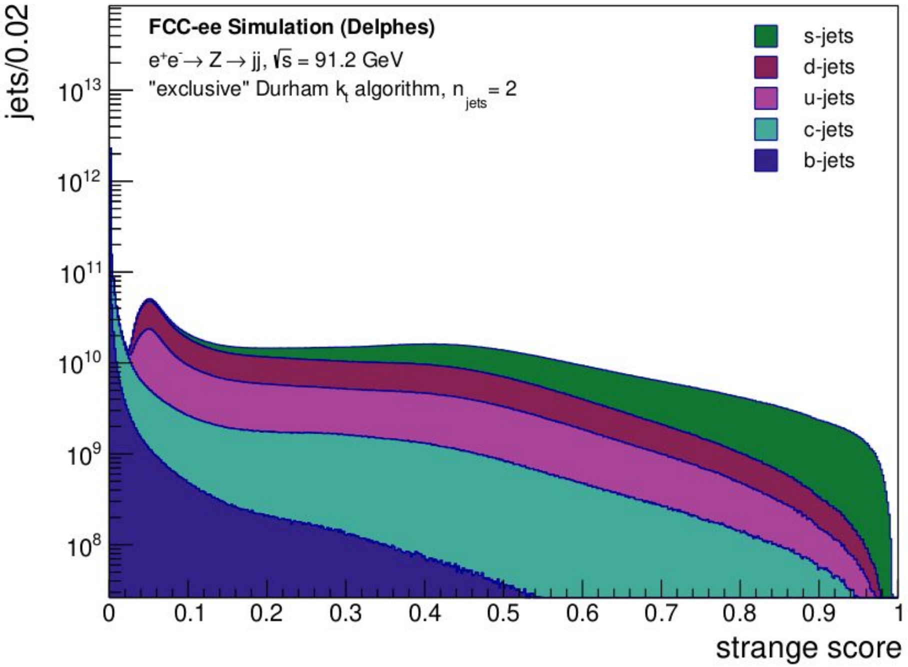
s-jets form significant background at low classifier scores, before only b-jets remain

Classifier Distributions: strange

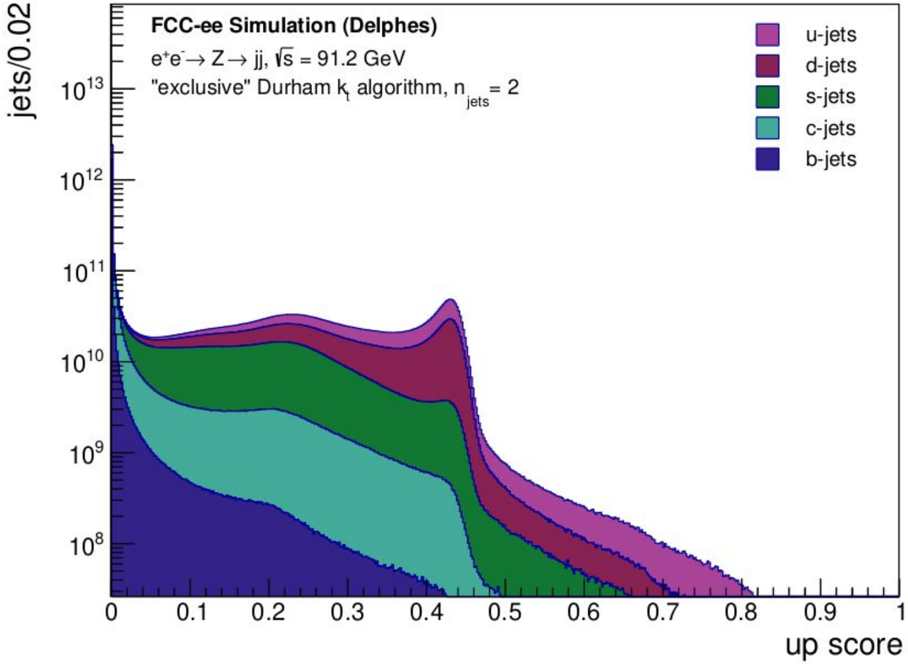
Strange quark discrimination much more non-trivial

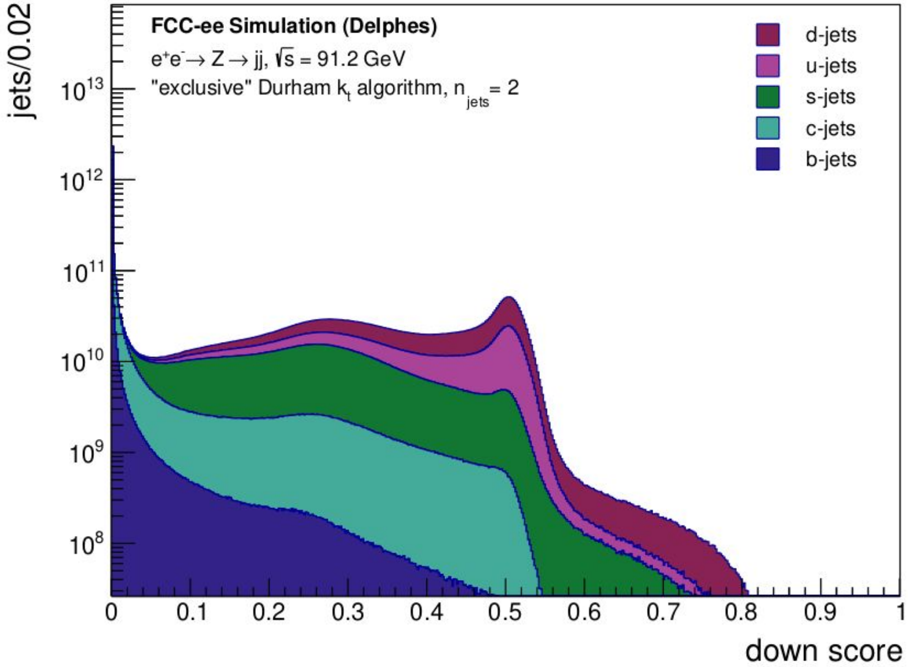
At high purity only u and d remain as backgrounds

Classifier distribution does not peak as distinctly as for heavy flavours, suggesting less confidence in discriminating power



Classifier Distributions: up and down



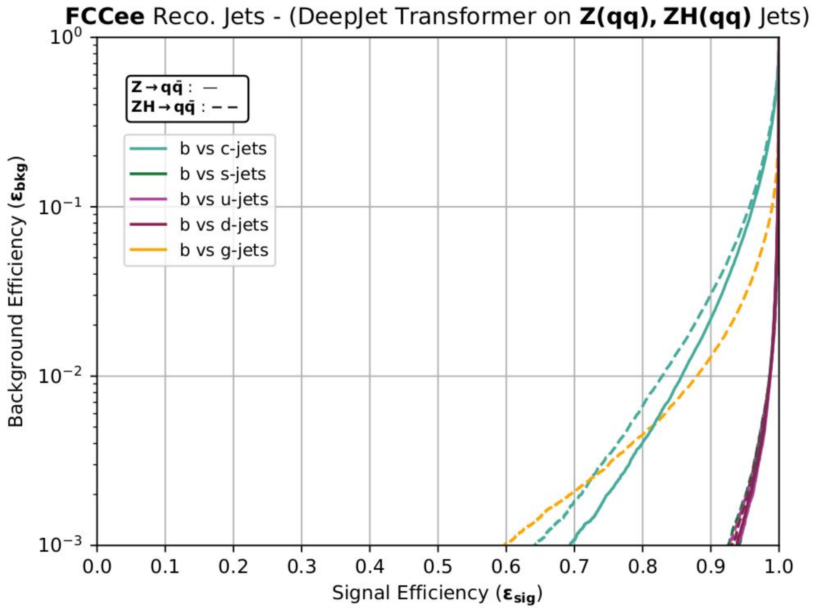


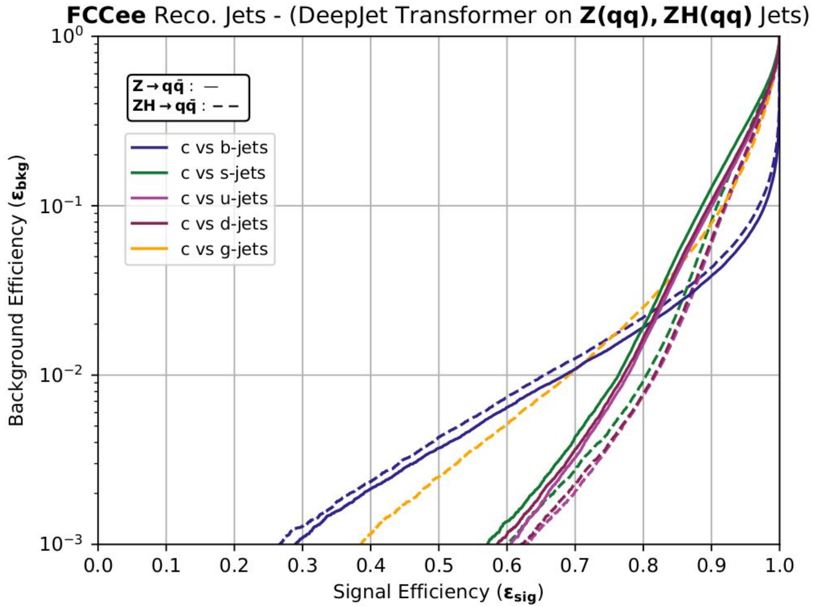
For up and down quarks discrimination is considerably worse

Peak at ~0.5 likely due to softmaxed output of classifier score being split between up and down

Classifier Performance: b and c

$$S_{ij} = \frac{S_i}{S_i + S_j}$$



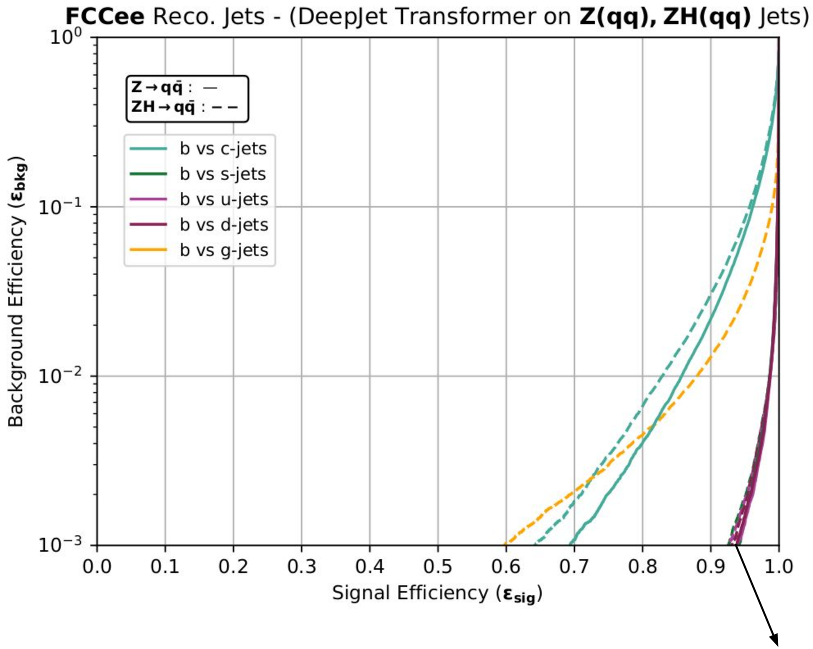


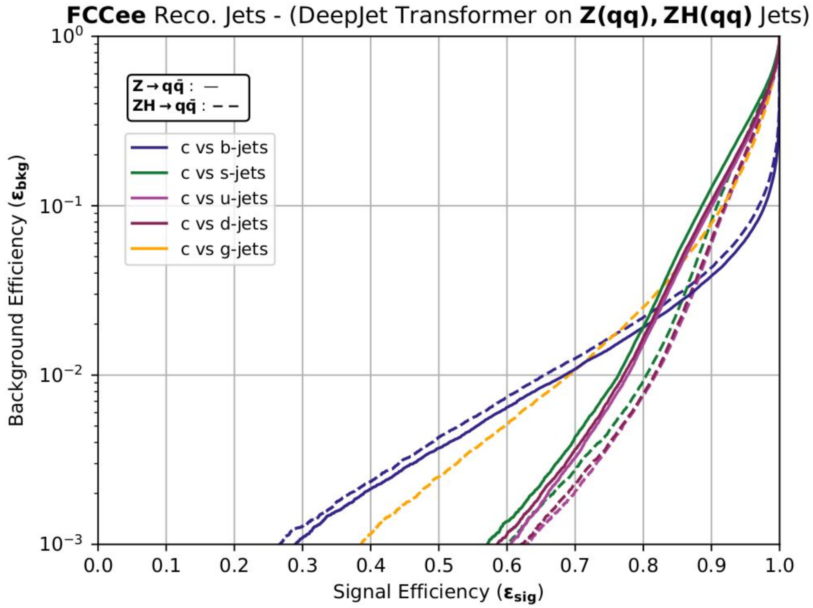
Excellent discrimination of b jets wrt light jets w/ 90%+ at bkg eff 0.1%

c jets as largest background together with gluon jets

Classifier Performance: b and c

$$S_{ij} = \frac{S_i}{S_i + S_j}$$



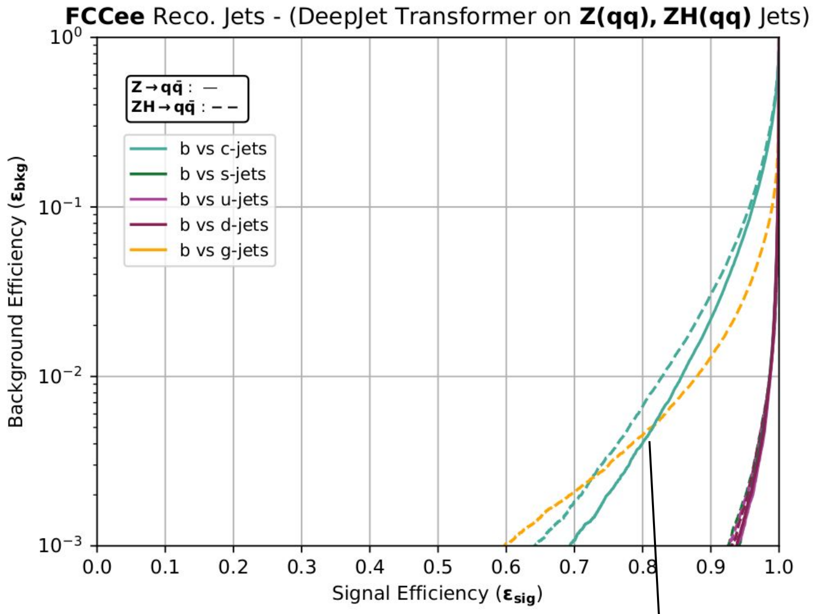


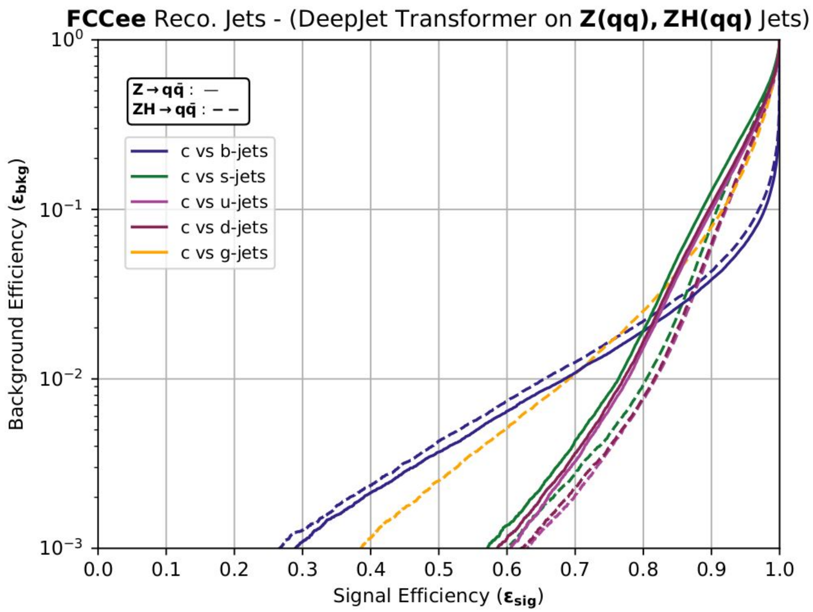
Excellent discrimination of b jets wrt light jets w/ 90%+ at bkg eff 0.1%

c jets as largest background together with gluon jets

Classifier Performance: b and c

$$S_{ij} = \frac{S_i}{S_i + S_j}$$



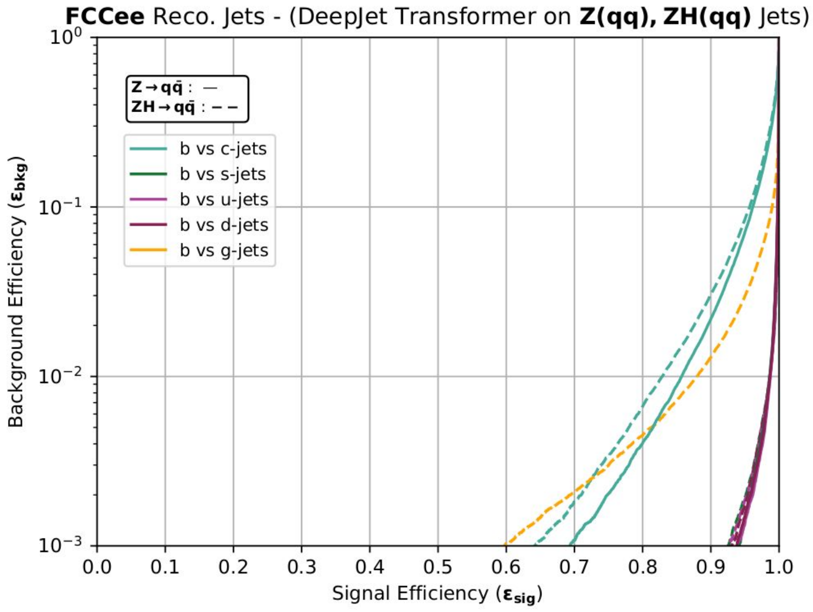


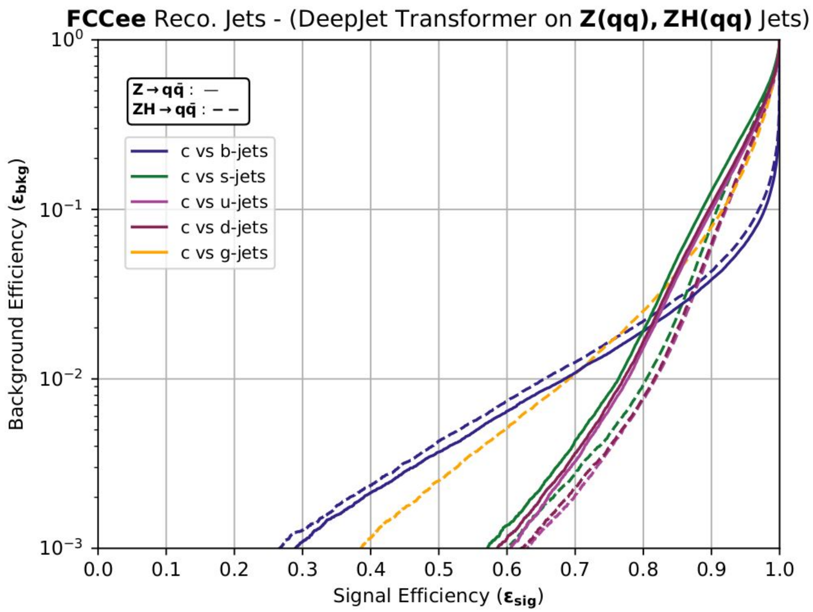
Excellent discrimination of b jets wrt light jets w/ 90%+ at bkg eff 0.1%

c jets as largest background together with gluon jets

Classifier Performance: b and c

$$S_{ij} = \frac{S_i}{S_i + S_j}$$





c jets likewise show strong performance, with b jets and gluons acting as background

Classifier Performance: s

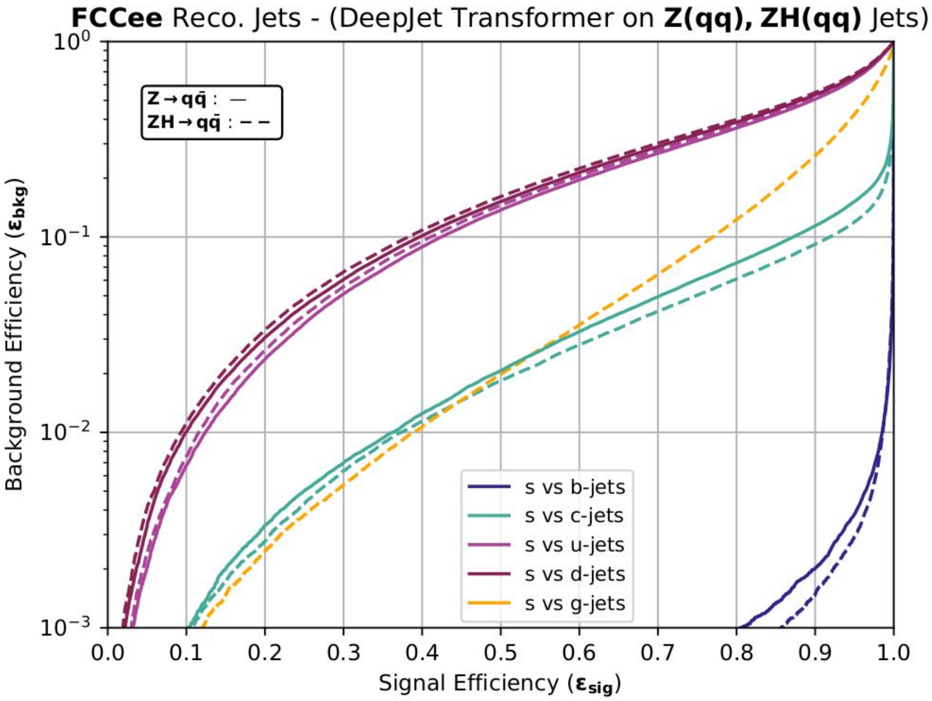
$$S_{ij} = \frac{S_i}{S_i + S_j}$$

For s-tagging, up and down jets present by far most challenging background

- PID is central to this type of discrimination

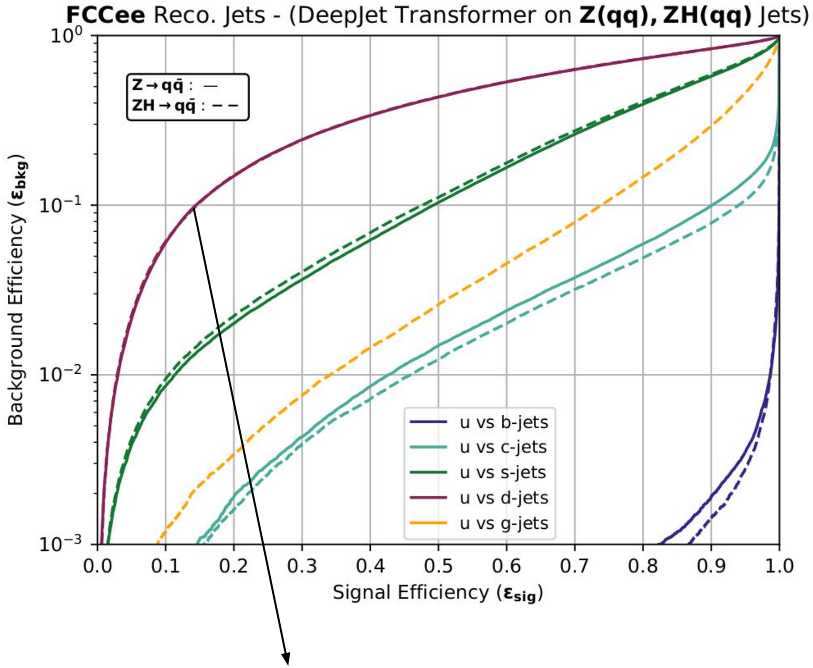
Charm and gluon jets present second most challenging, likely due to

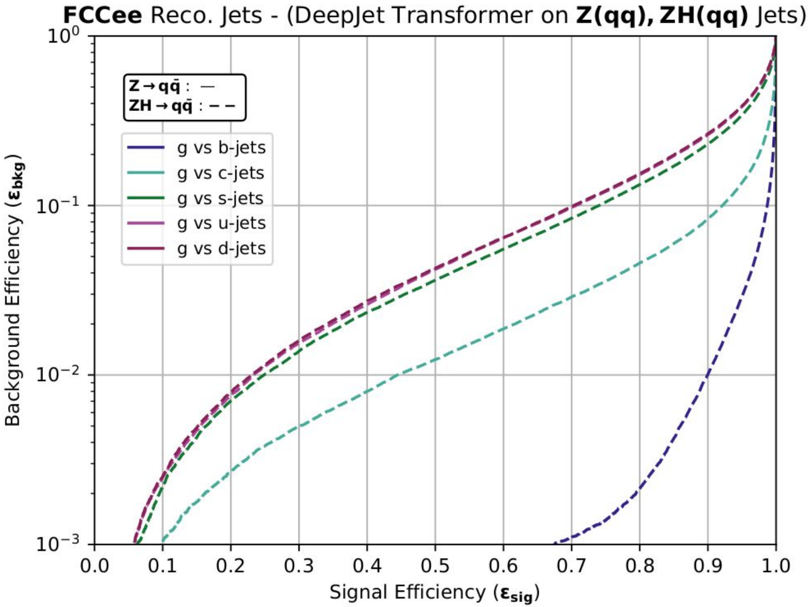
- Charm hadron decay to strange hadron
- g->ss



Classifier Performance: u and gluons

$$S_{ij} = \frac{S_i}{S_i + S_j}$$

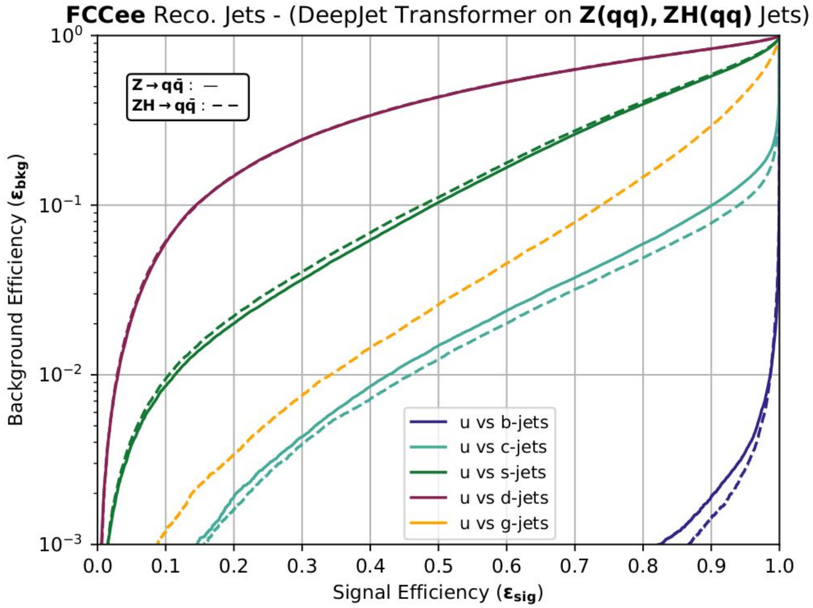


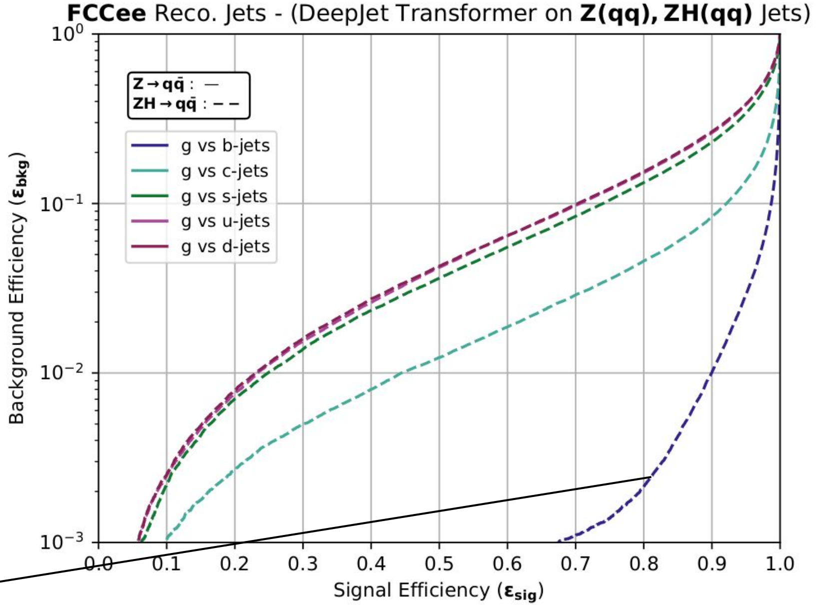


up jet vs down jet discrimination not much better than random classifier with sig eff ~15% and bkg eff 10%

Classifier Performance: u and gluons

$$S_{ij} = \frac{S_i}{S_i + S_j}$$



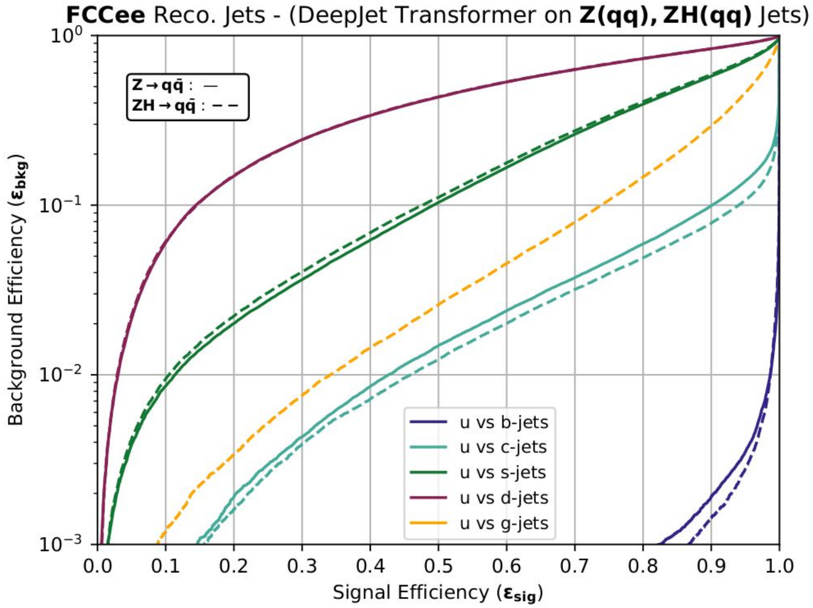


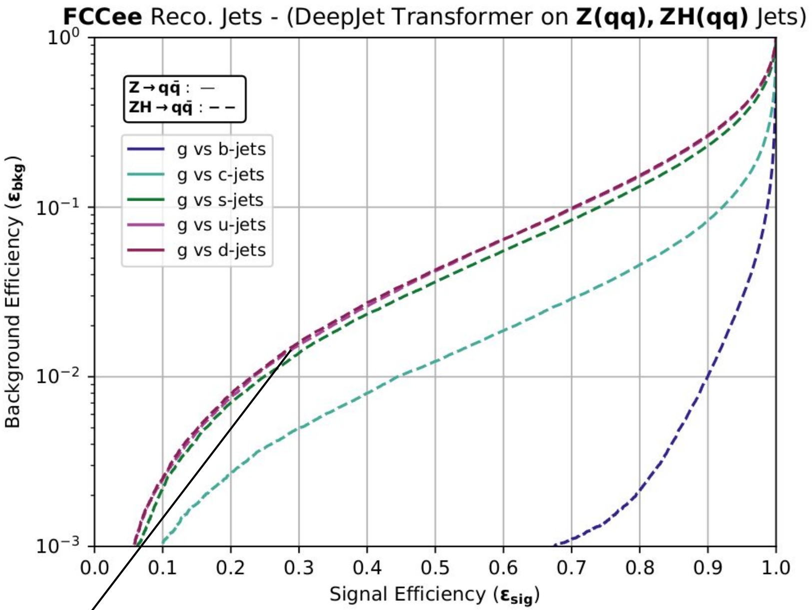
Best gluon discrimination is against b quarks

uds challenging due to similar jet composition

Classifier Performance: u and gluons

$$S_{ij} = \frac{S_i}{S_i + S_j}$$

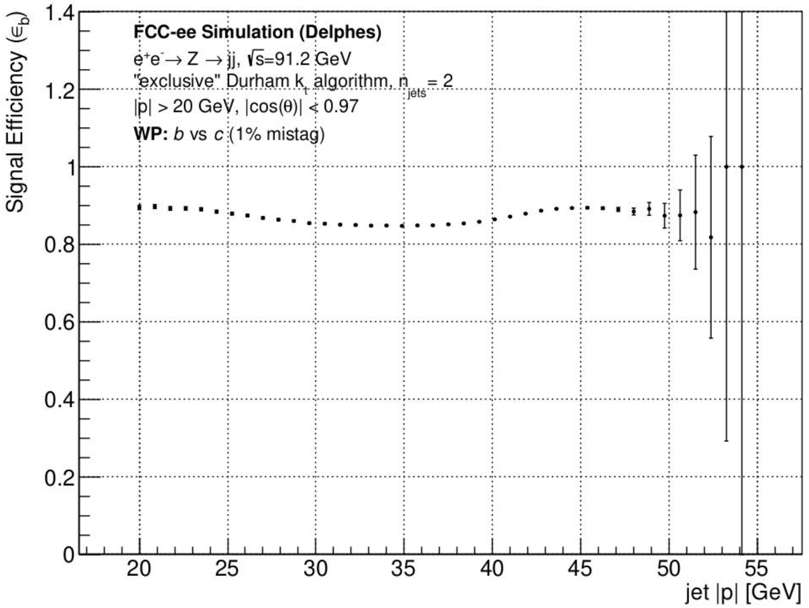


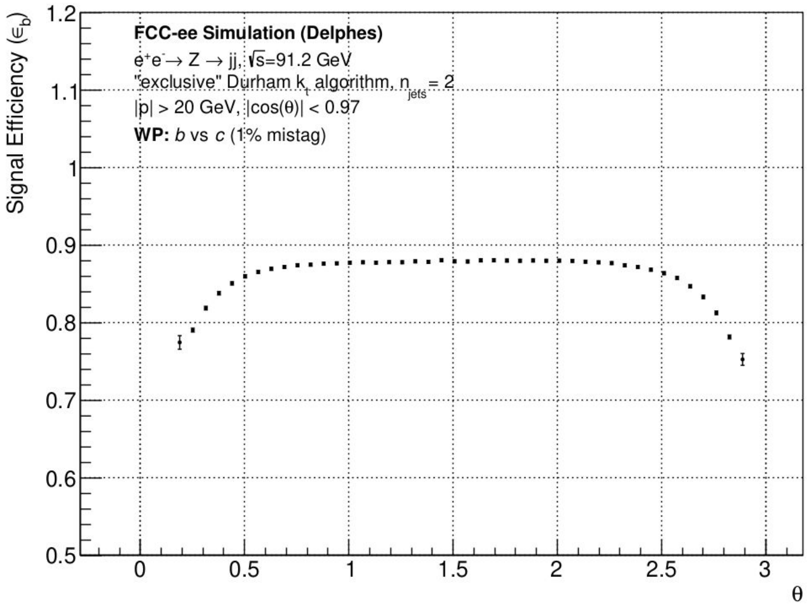


Best gluon discrimination is against b quarks

uds challenging due to similar jet composition

Tagging Efficiencies of b vs c

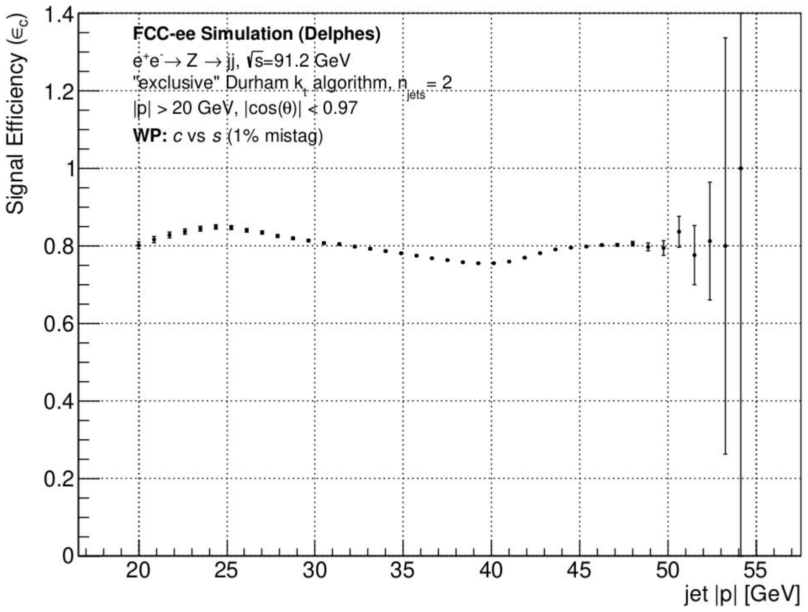


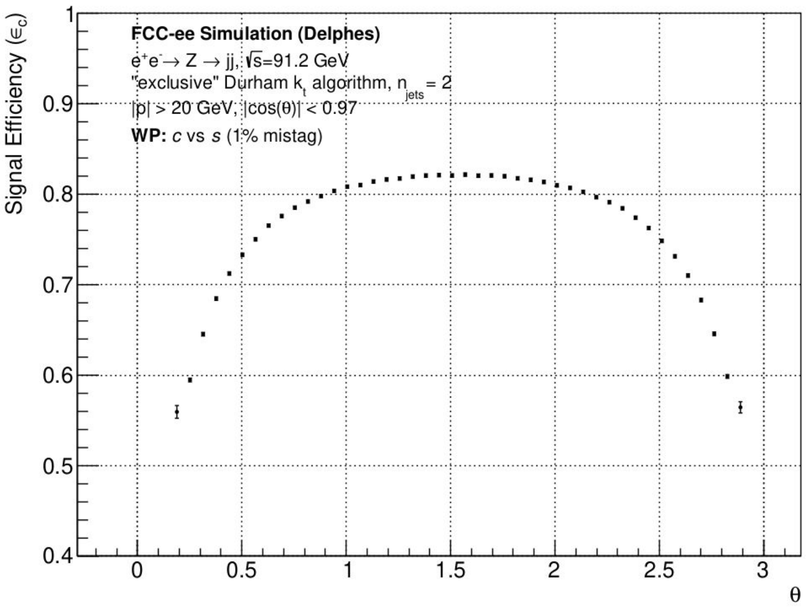


Efficiency mostly uniform across jet |p|

Theta shows drop off at extremes, due to jet constituents being lost to fiducial cuts

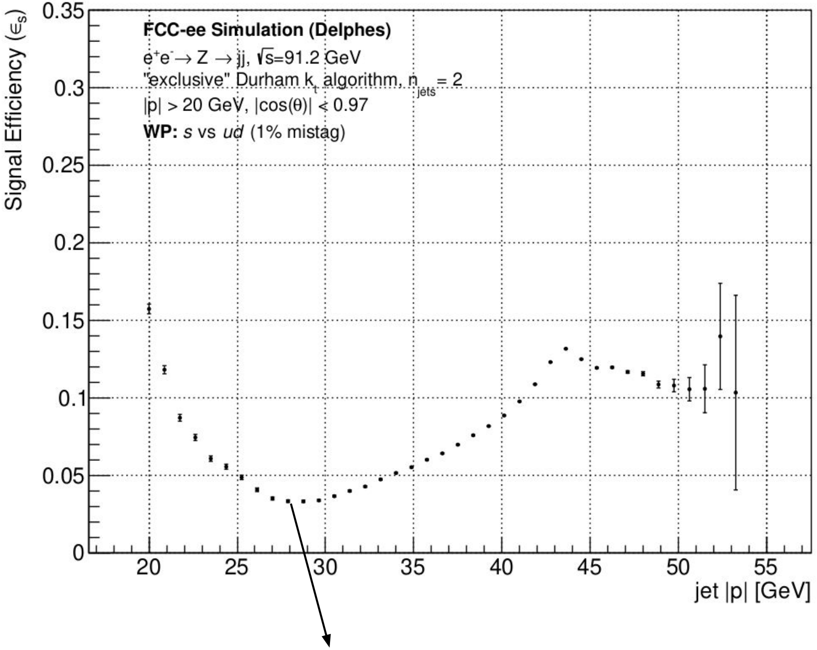
Tagging Efficiencies of c vs s

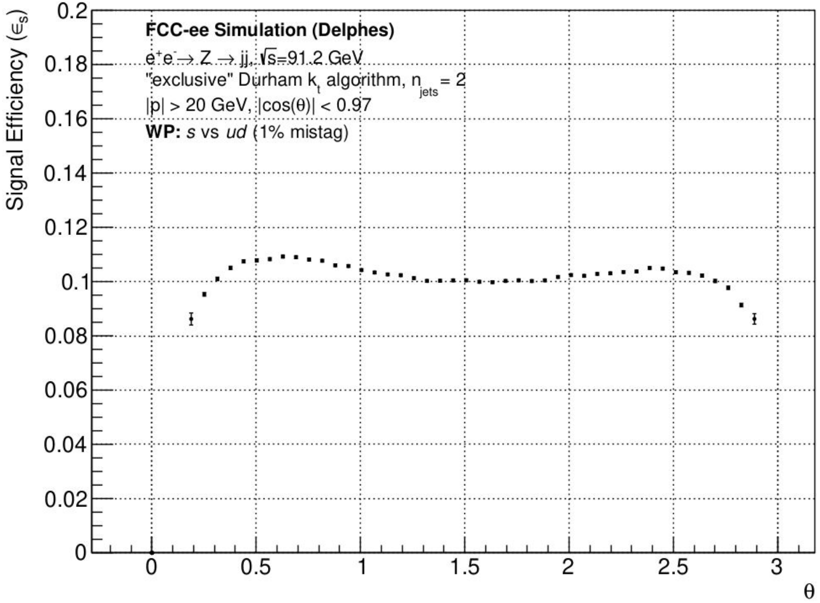




Virtually same trends as for b vs c discrimination, with uniform efficiencies

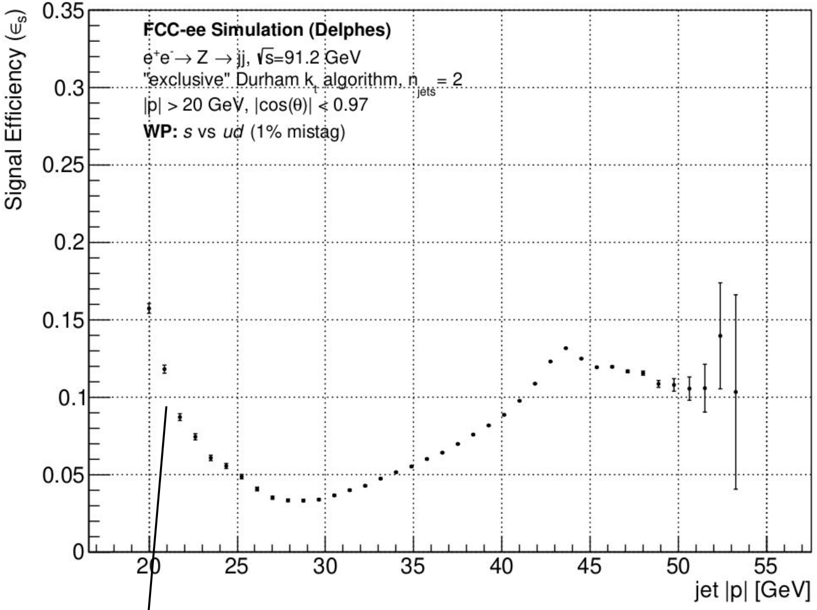
Tagging Efficiencies of s vs ud

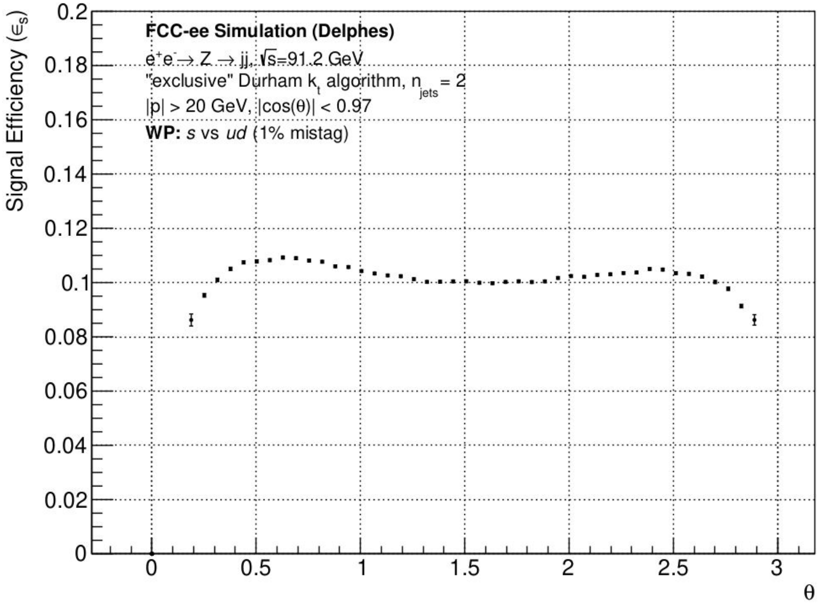




Low momentum strange jets have lower K^{+-} multiplicities, leading to reduced tagging efficiency

Tagging Efficiencies of s vs ud



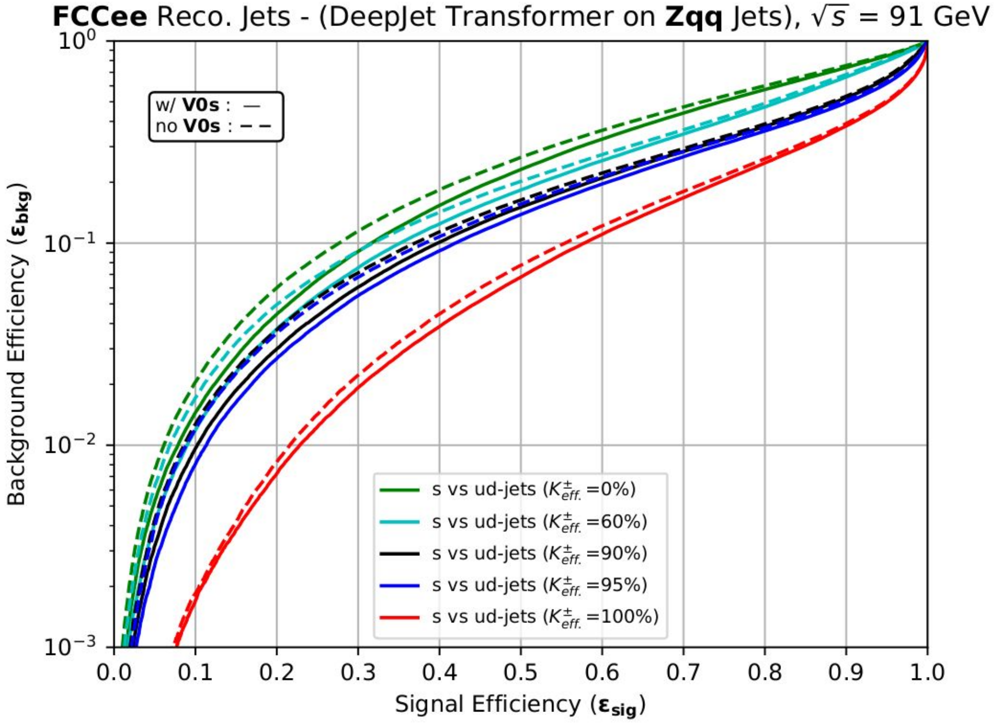


Very low momentum strange jets have low particle multiplicities overall, where a single reconstructed V^0 becomes a distinguishing feature

Dependence of s-tagging on Kaon ID

s-tagging performance w/ ud-jets as background is extremely sensitive to K^{+-} ID

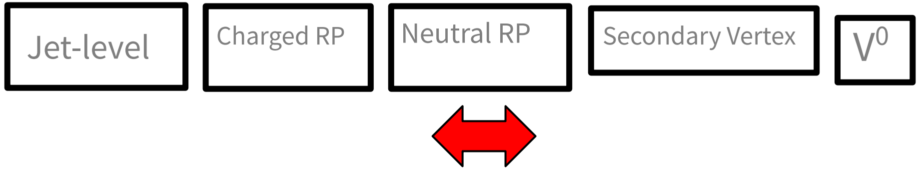
Further gains possible through inclusion of V^0 variables



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

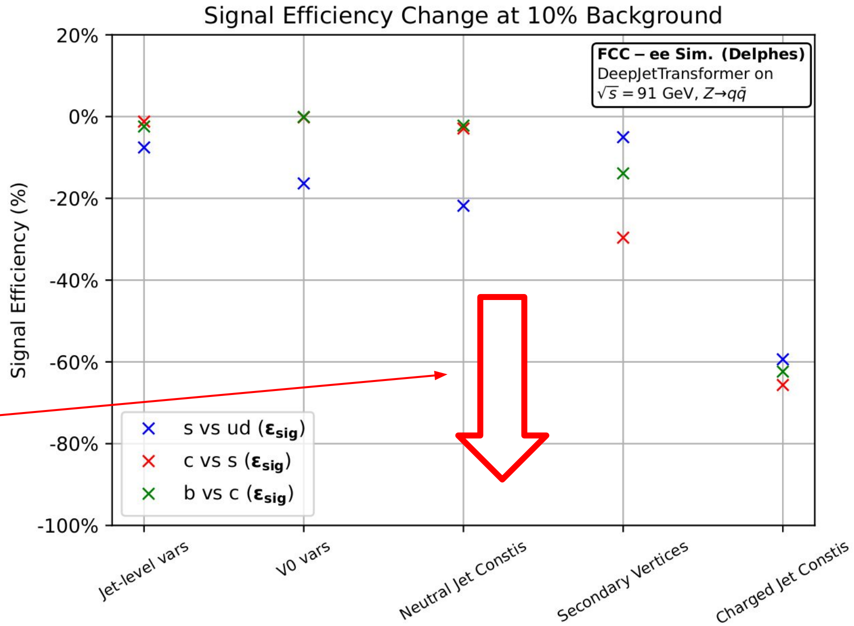
Importance of Variable Classes

Shuffle entire group of variables (e.g. Neutral RP variables) amongst different jets to estimate importance



Consider % change in signal efficiency at fixed background efficiency of 10% for s vs ud, c vs s, b vs c:

Lower = more impactful, bounded below by 100% (50% for AUC), which
is worse than a random classifier

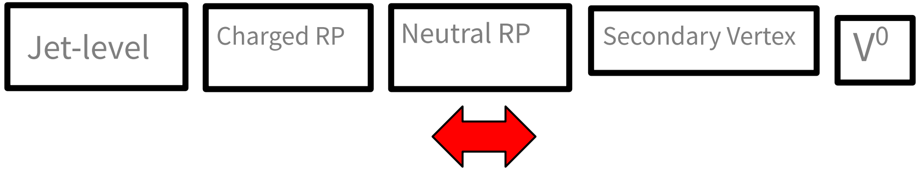


Importance of Variable Classes

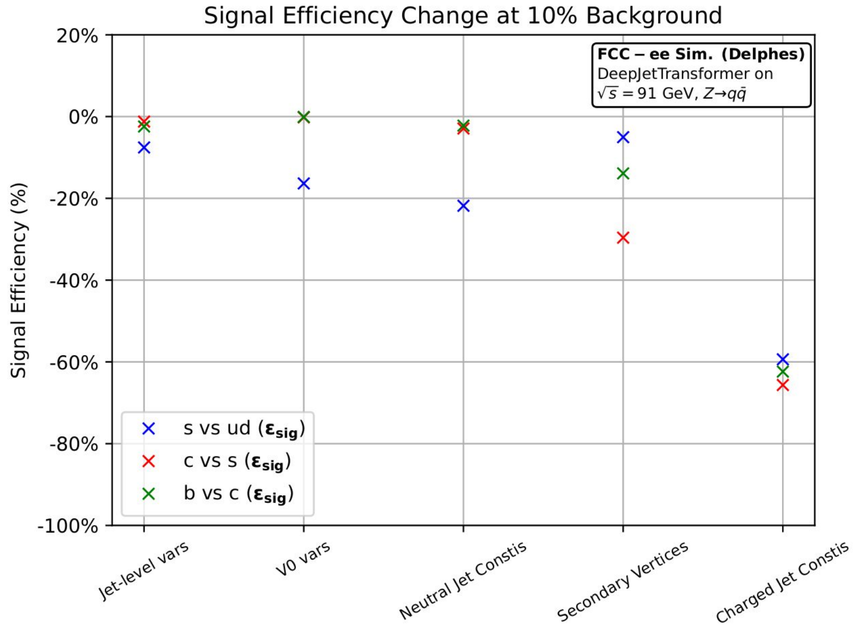
Shuffle entire group of variables (e.g. Neutral RP variables) amongst different jets to estimate importance

Consider % change in signal efficiency at fixed background efficiency of 10% for s vs ud, c vs s, b vs c:

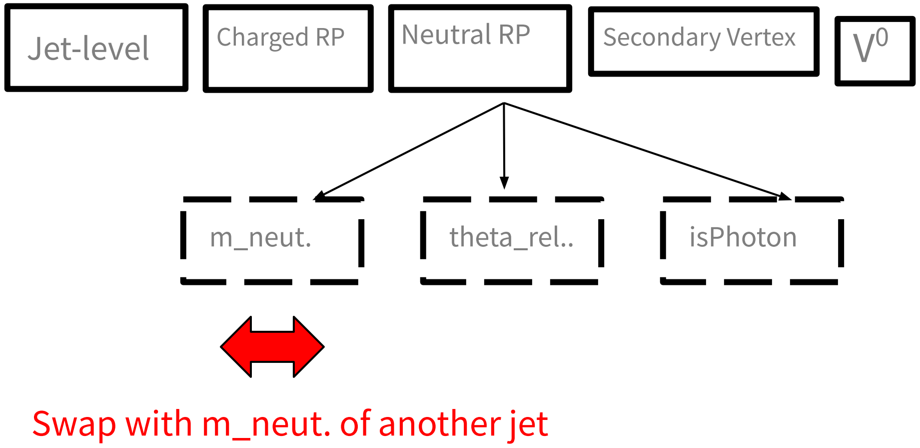
- Charged jet constituents most impactful for all three flavour combinations
- s vs ud seems to benefit from the other three types of variables (jet-level, V0, neutral jet constituents), while heavy flavour tagging does not



Swap with those of another jet



Importance of Individual Variables



There are roughly 60 sub-variables belonging to the 5 variable types (sv, v0, …)

Plotted the 10 most impactful ones:

- Kinematic variables of charged particle constituents are generally impactful
- Track variables are likewise impactful
- PID variables matter massively for s vs ud

Importance of Individual Variables

Jet-level | Charged RP | Neutral RP | Secondary Vertex | V^0

Kinematic $(|p|,\ E,\ p/p_{jet.},\ \theta,\ \Delta\theta,\ \dots)$

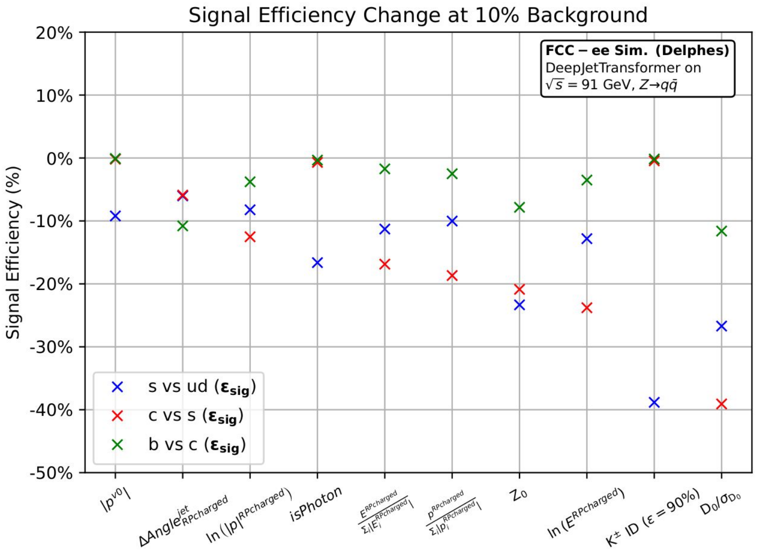
PID $(isPhoton,\ K^{\pm}\mathrm{ID},\ \dots)$

Track $(D_0,\ Z_0,\ \dots)$

There are roughly 60 sub-variables belonging to the 5 variable types (sv, v0, …) →

Plotted the 10 most impactful ones:

- Kinematic variables of charged particle constituents are generally impactful
- Track variables are likewise impactful
- PID variables matter massively for s vs ud



The Z boson at the FCC-ee

Z bosons decay relatively uniformly to 5 quark flavours, providing ideal case study for strange tagging

Performed by SLD to measure A_s

First Direct Measurement of the Parity-Violating Coupling of the Z^0 to the s Quark

Koya Abe *et al.* (The SLD Collaboration)
Phys. Rev. Lett. **85**, 5059 – Published 11 December 2000

Performed also by DELPHI

Measurement of the strange quark forward-backward asymmetry around the Z^0 peak

Experimental physics | Published: June 2000

Volume 14, pages 613–631, (2000) Cite this article

With 6×10^{12} visible decays during its 4 year Z pole run, the FCC-ee is uniquely suited

Event Selection

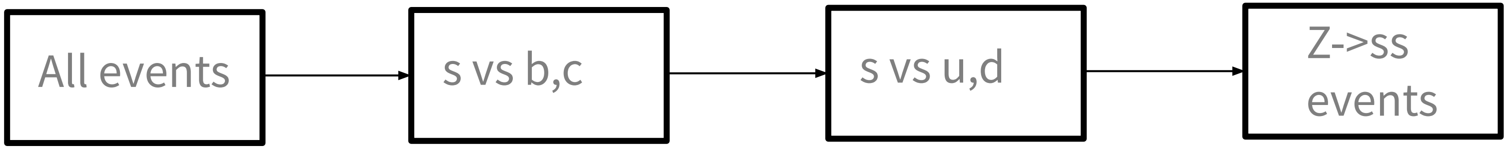
Exclusive clustering of Z->qqbar events into 2 jets using e+e- kT algorithm

Impose |p|>20 GeV & cos(theta)<0.972

Define classifier thresholds at 4 Working Points wrt **per-jet** background efficiency of two sequential cuts

- s vs bc
- s vs ud

Both jets in event required to pass cuts on s-jets



Performance for all Working Points

Lumi = 125 ab^{-1}

		Mistag Rate [%]	Efficiency [%]	N_{sig}	N_{bkg}
WP1	s vs bc	10.01	98.93 ± 0.03	7.35×10^{11}	1.35×10^{12}
	s vs ud	10.03	40.03 ± 0.04	1.45×10^{11}	3.25×10^{10}
WP2	s vs bc	1.02	54.18 ± 0.04	2.38×10^{11}	2.06×10^{11}
	s vs ud	10.03	39.28 ± 0.06	5.10×10^{10}	5.57×10^{9}
WP3	s vs bc	1.02	54.18 ± 0.04	2.38×10^{11}	2.06×10^{11}
	s vs ud	1.0	10.05 ± 0.11	1.12×10^{10}	4.77×10^{9}
WP4	s vs bc	0.11	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}	3.39×10^{6}

per-jet (Mistag Rate, Efficiency) — per-event (N_{sig}, N_{bkg})

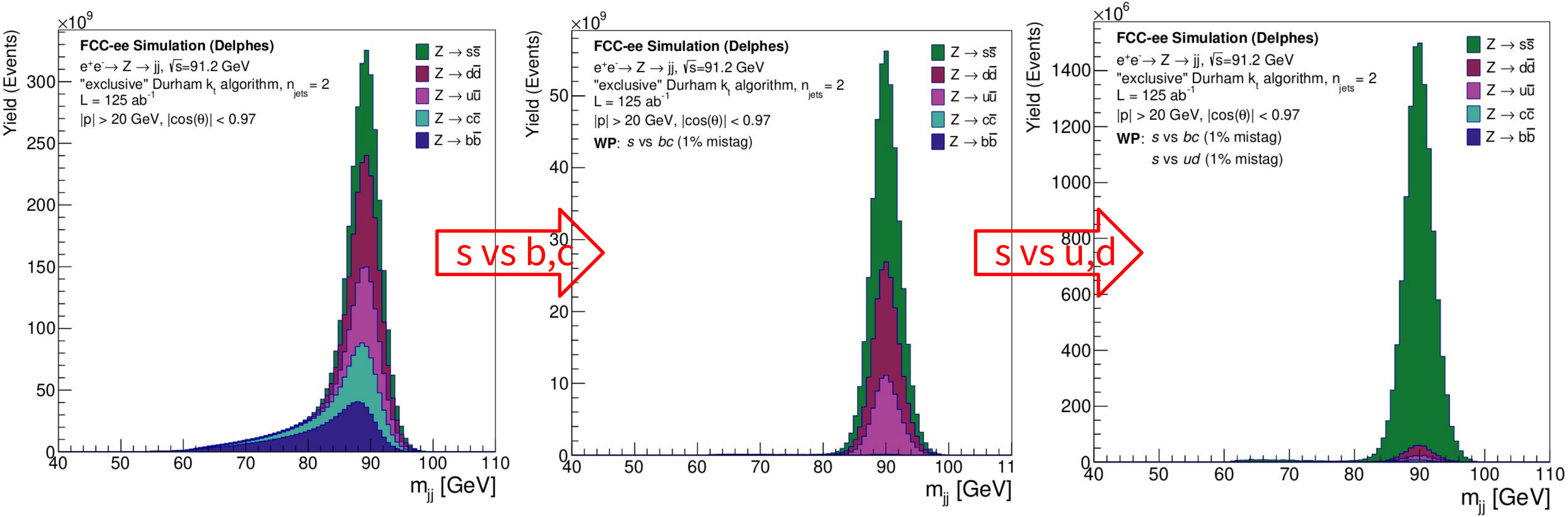
Performance for all Working Points

Lumi = 125 ab^{-1}

		Mistag Rate [%]	Efficiency [%]	N_{sig}	N_{bkg}
WP1	s vs bc	10.01	98.93 ± 0.03	7.35×10^{11}	1.35×10^{12}
	s vs ud	10.03	40.03 ± 0.04	1.45×10^{11}	3.25×10^{10}
WP2	s vs bc	1.02	54.18 ± 0.04	2.38×10^{11}	2.06×10^{11}
	s vs ud	10.03	39.28 ± 0.06	5.10×10^{10}	5.57×10^{9}
WP3	s vs bc	1.02	54.18 ± 0.04	2.38×10^{11}	2.06×10^{11}
	s vs ud	1.0	10.05 ± 0.11	1.12×10^{10}	4.77×10^{9}
WP4	s vs bc	0.11	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}	3.39×10^{6}

per-jet per-event

Performance at WP3



Obtain very pure resonance of s-jets at 1% working point

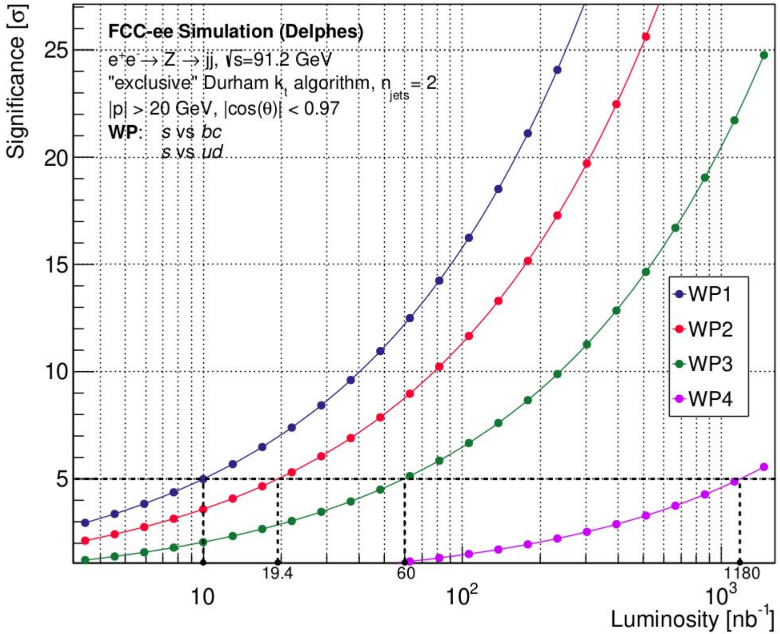
Results for other working points in backup

Significance

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

For these studies we neglect backgrounds!

For WP3, 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



Outlook

A wish-list (beyond the scope of our paper):

Improvements in current feature set

- Could be extended to include jet-shape variables and full covariance matrix
- Include more realistic PID assumptions like ParticleNetIDEA (mass from time-of-flight, dN/dx)
- Reduce degeneracy/overlap in current input feature set

Outlook II

Physically-motivated sub-division of flavours

- Hadronic vs semi-leptonic b-jets
- g->bb splittings
- Quarks vs Anti-quarks
- (Event-level tagging)

Updated detector concepts

- IDEA w/ innermost layer of vertex moving from 1.7mm to 1.3mm
- CLD w/ dedicated RICH PID detector

Conclusions

Flavour tagging essential for future colliders

DeepJetTransformer as lightweight + performant alternative to competing architectures

Not unique to FCC-ee, other collider projects with appropriate adjustments

Excellent discrimination of

- b, c vs s, u, d
- s vs ud feasible but very dependent on K+-/pi+- separation and V^0 reconstruction

Showed that Z->ssbar can be efficiently isolated from other hadronic decays of Z boson

Plan to submit paper in arxiv in time scale of few weeks

Draft already available in CDS (internal): https://new-cds.cern.ch/records/x5sc0-01010

- Many thanks to Loukas and Michele for agreeing to have a look!

Backup

Sample Preparation and Training

https://github.com/Edler1/DeepJetFCC/tree/master/docs