Transformer-based Jet Flavor Tagging in Full Simulation for CLD at FCC-ee

FUTURE CIRCULAR COLLIDER

3rd ECFA workshop on e^+e^- Higgs, Top & ElectroWeak Factories

9. October 2024

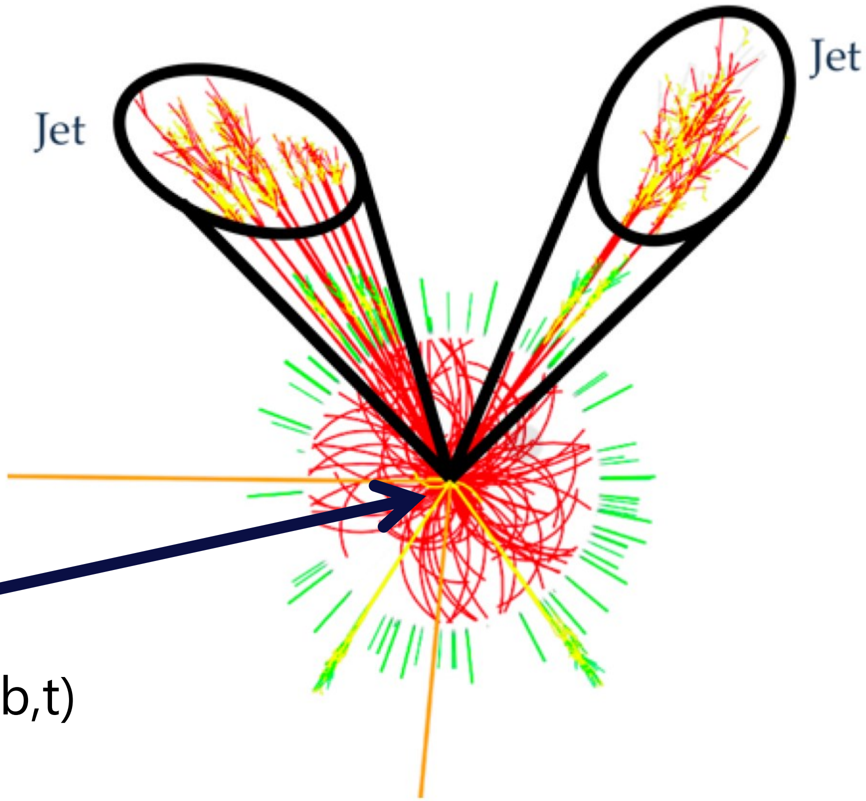
Sara Aumiller, Dolores Garcia, Michele Selvaggi

Jet-Flavor Tagging explained

FUTURE CIRCULAR COLLIDER

Jet

= narrow cone of multiple particles produced by hadronization



Jet origin:

- quarks (u,d,s,c,b,t)
- gluons (g)
- leptons (τ)

Jet-flavor tagging

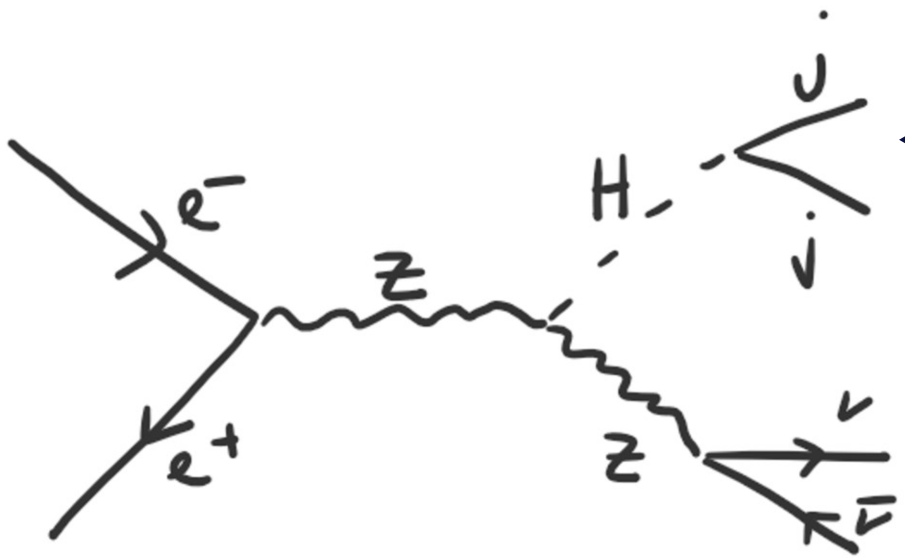
=

Which particle caused the jet?

Why jet-flavor tagging?

Future Colliders = Higgs factories for precious measurements

→ Study Higgs decay modes



Event signature: N jets (here: 2)

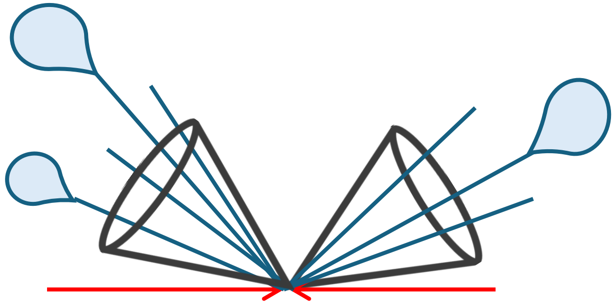
or $l\bar{l}$, jj

Jet-flavor tagging set-up

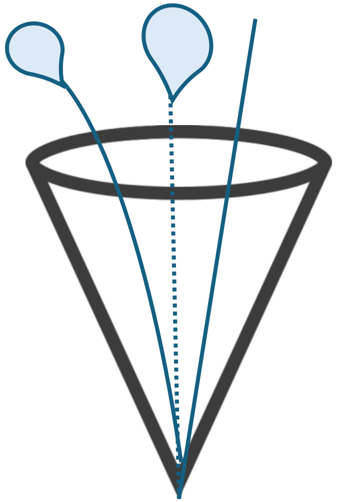
reconstruction → Jet description → Neural network → Jet flavor → Higgs physics

reconstruction

Find two jets per event



Jet description

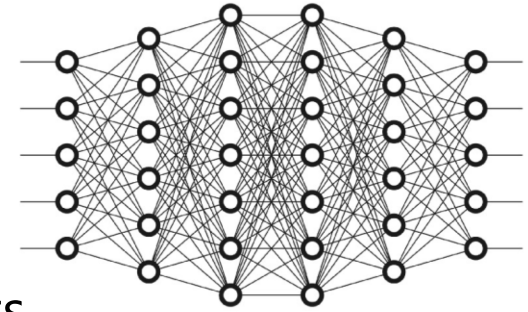


constituents
Kinematics
Track displacements
PID
...

Neural network

Particle Transformer

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



Jet flavor

7 flavors (u,d,s,c,b, g, τ)

Higgs physics

Branching ratios

Couplings to fermions

...

Jet description in fast & full simulation at CLD

Fast simulation*: provides time & computational efficient early-stage feasibility studies

Full simulation: more realistic description of detector concept and reconstruction algorithms

Jet-description observables**:

Kinematics (3)	Identification (7)	Track displacements (23)
log E_{rel}, θ_{rel}, ϕ_{rel}	reco PID, charge, PID flags	d_0, z_0, covariance matrix c_{ij}, SIP in 2D, 3D (& significance), Jet-track distance d_{3D} (& sig.)

*IDEA fast sim with silicon tracker, here referred to as "CLD fast sim"

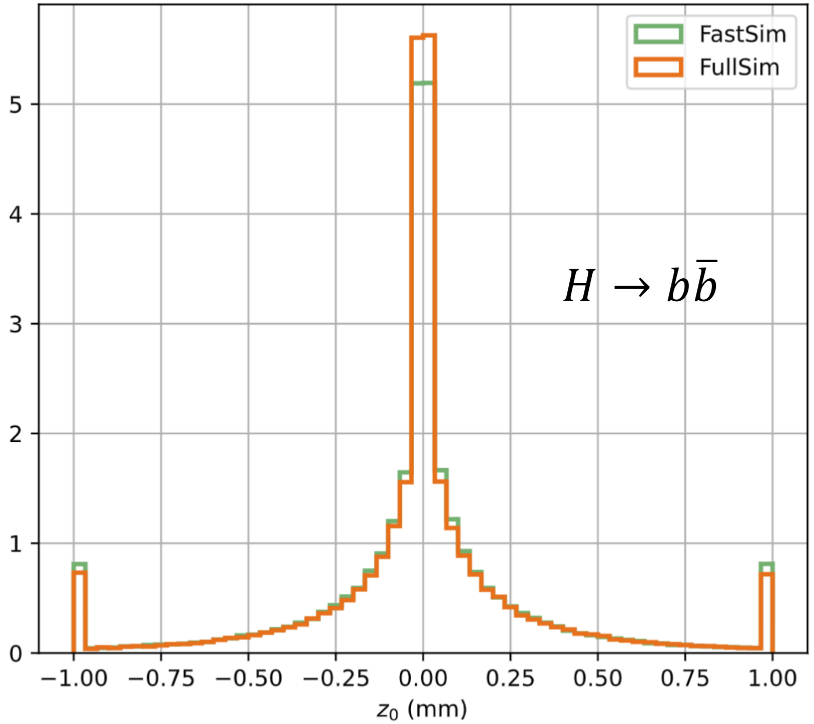
**see IDEA fast sim tagging

Jet description in fast & full simulation at CLD

Fast simulation*: provides time & computational efficient early-stage feasibility studies

Full simulation: more realistic description of detector concept and reconstruction algorithms

→ Comparison of jet description shows mostly good agreement.



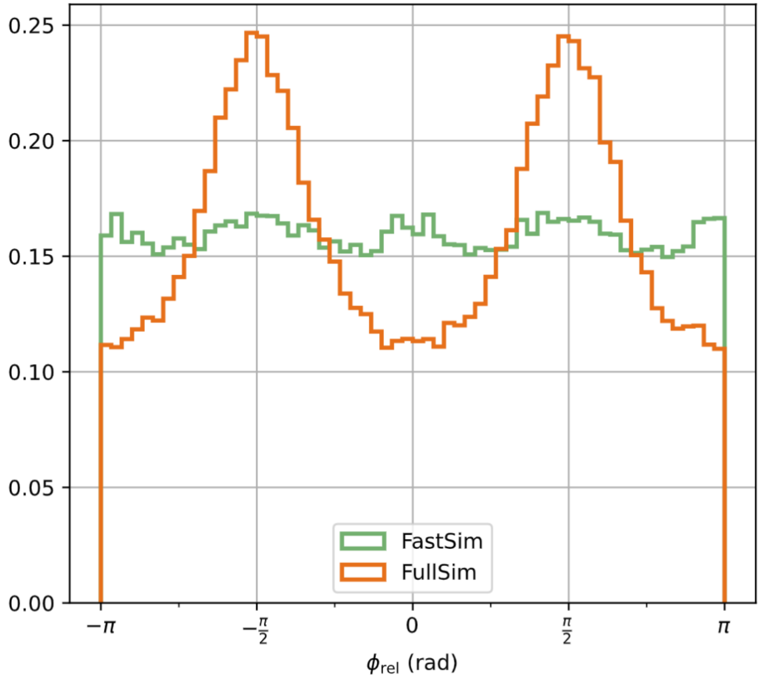
Two major differences when moving to full simulation:

1. Fake neutrals
2. Lost tracks

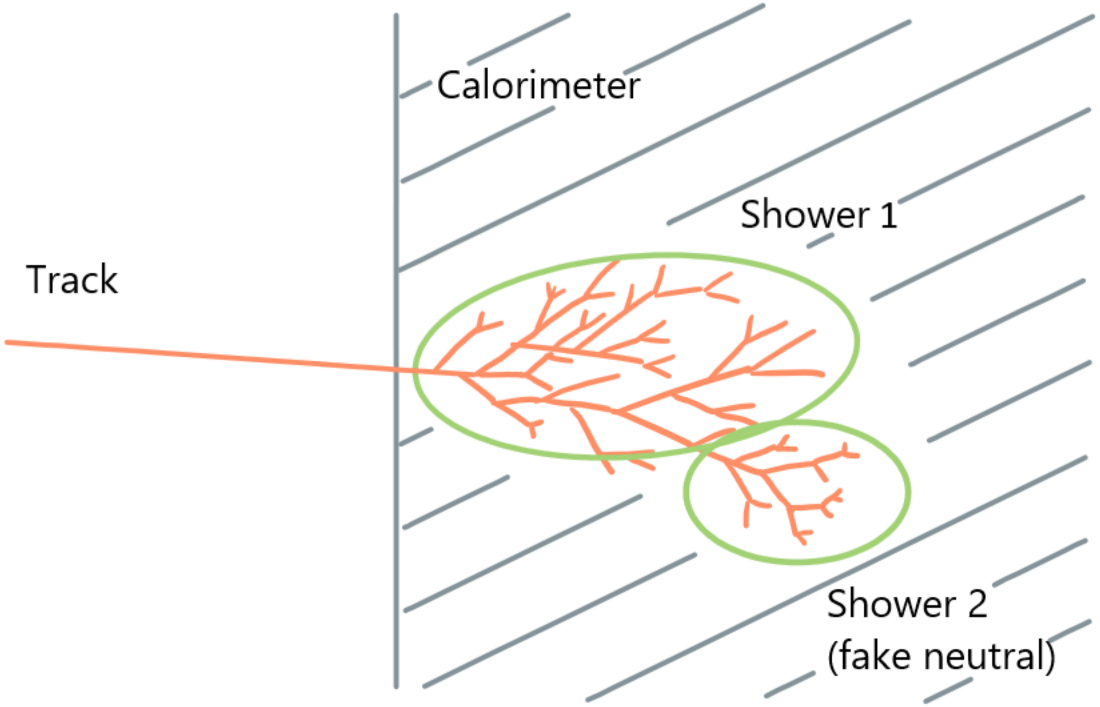
*IDEA fast sim with silicon tracker, here referred to as "CLD fast sim"

(1) Fake neutrals in full sim

- More neutral hadrons in full than in fast simulation
- Relative angle ϕ of neutral jet constituents shows discrepancy
- Peaks at $\pm\frac{\pi}{2}$ indicate high-energy charged particles (at MC level) which are wrongly reconstructed



leading neutral hadronic jet constituents

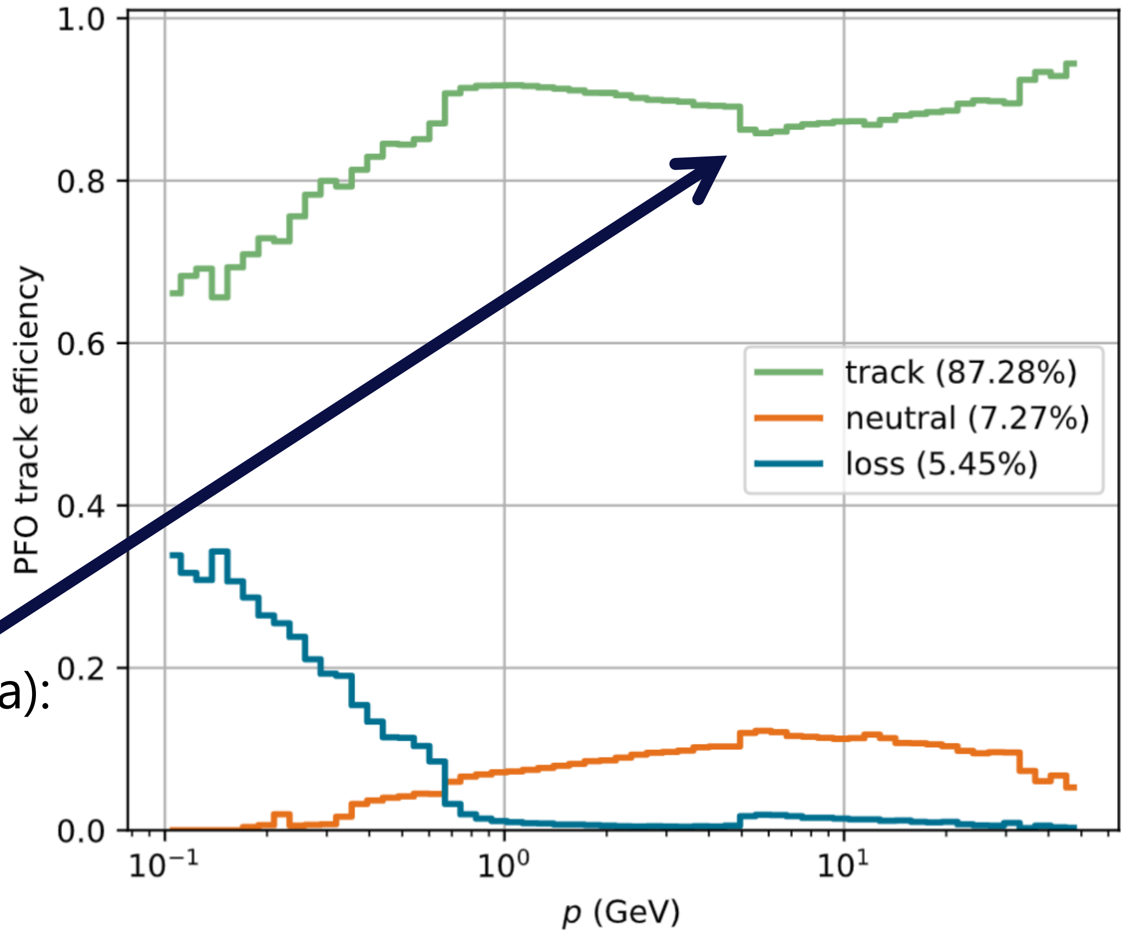


(2) Lost tracks in full simulation

Some charged particles are wrongly reconstructed as neutrals in full sim.

→ problematic as tracks are crucial for jet flavor tagging

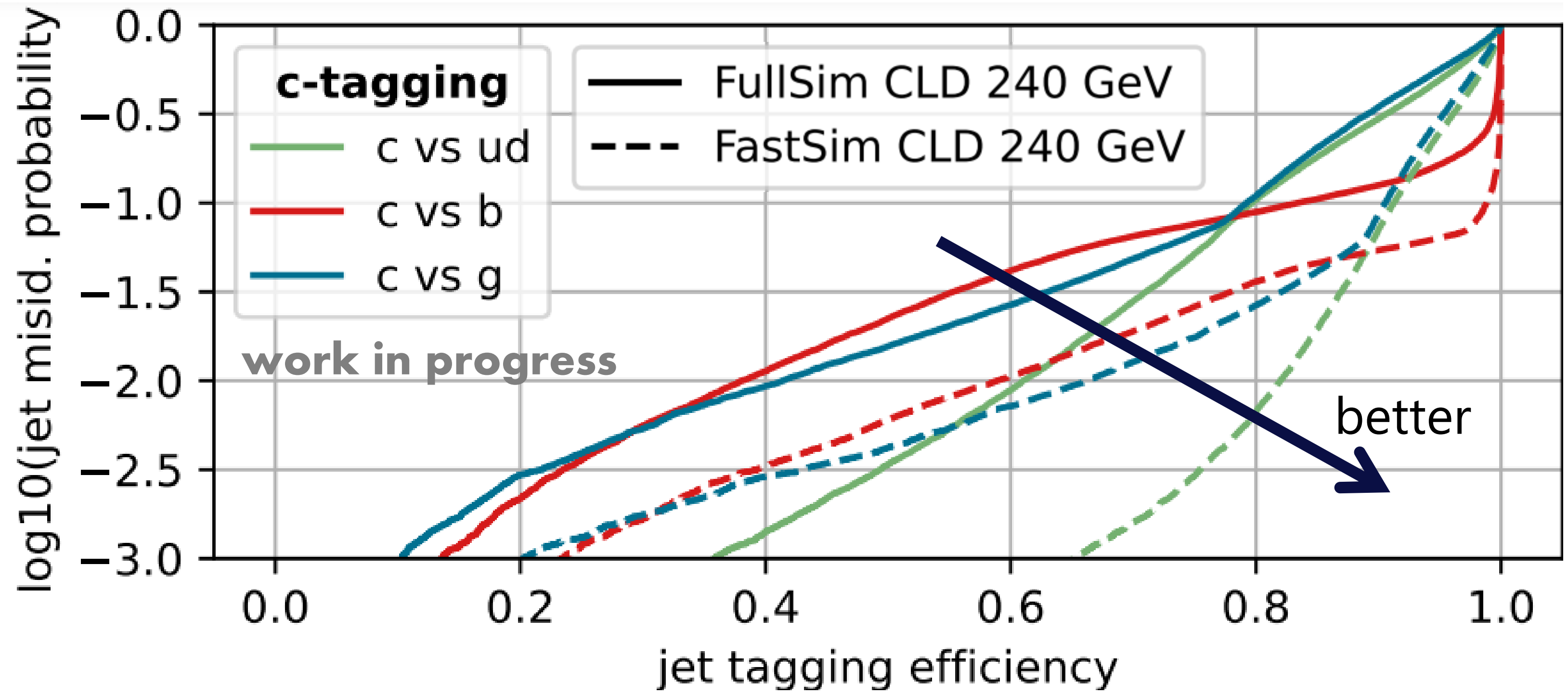
MC charged hadrons ($H \to b\bar{b}$)



Reconstruction constraint (from pandora): above 5 GeV charged particles must have cluster associated

→ reconstruction could be improved

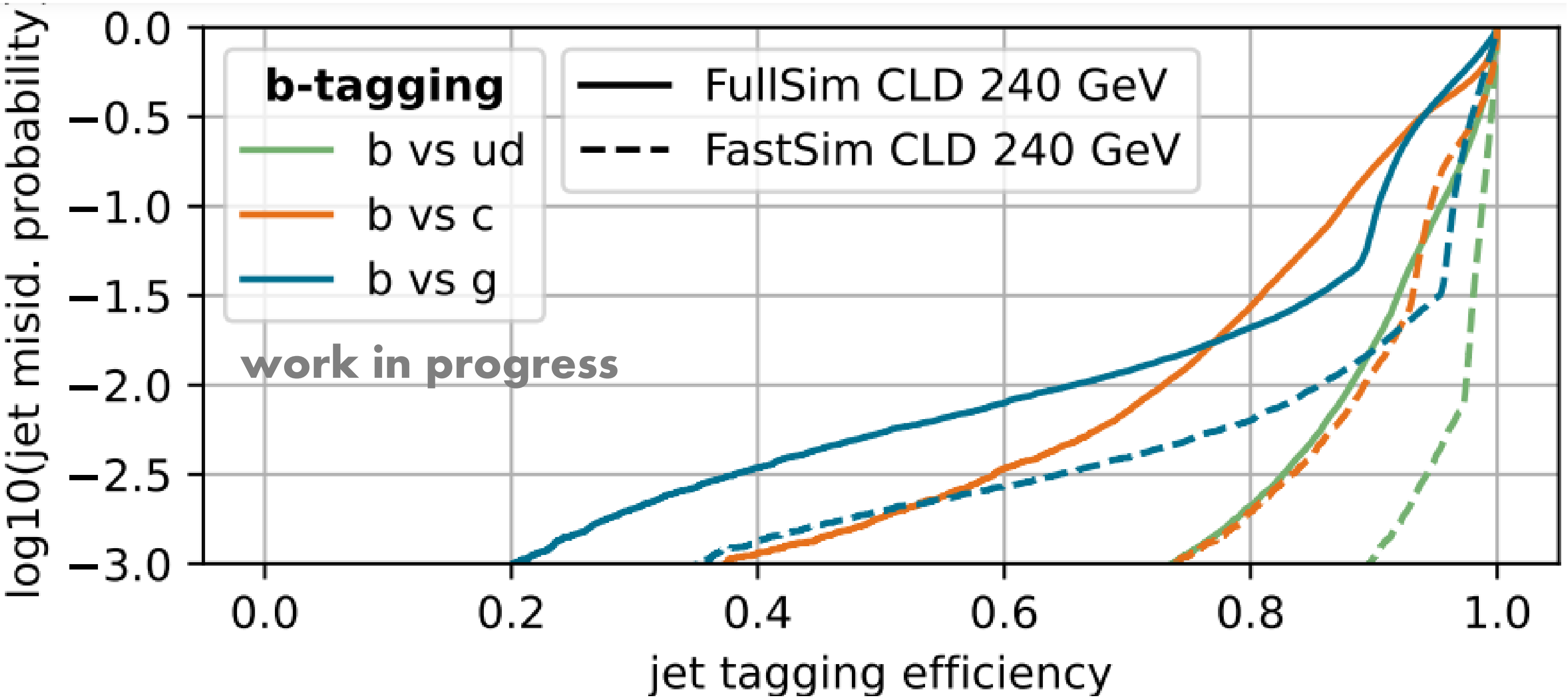
Full vs. fast simulation



Loss in performance when moving from fast to full simulation

e.g. misidentification probability of 10^{-2} for c vs. ud: 82% (fast sim) / 61% (full sim)

Full vs. fast simulation



Loss in performance when moving from fast to full simulation

e.g. misidentification probability of 10^{-2} for b vs. ud: 97% (fast sim) / 88% (full sim)

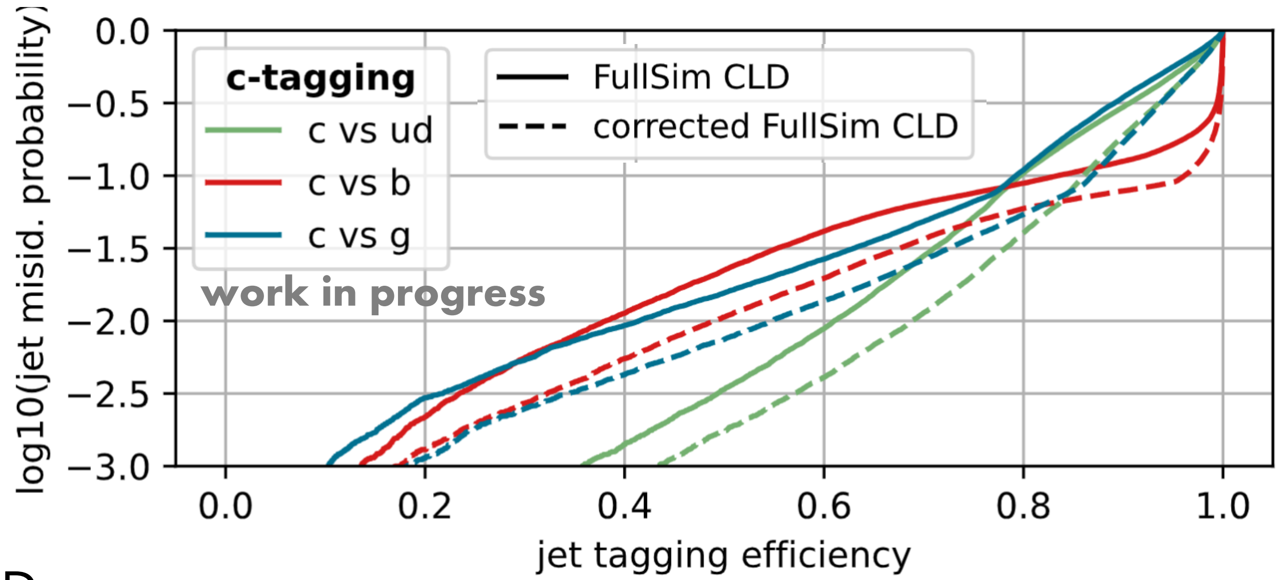
Improving full sim tagging?

- Improve input data to neural network
- Use all tracks available!
- Ignore fake neutrals

Idea:

Instead of PFOs (particle flow objects) use

- Tracks for charged particles
- PFOs for neutral particles but check MC PID to avoid double counting



Large improvement:
e.g. misidentification probability of 10^{-2} for c vs. ud:
Improves from 61% to 69% (fast sim: 82%)

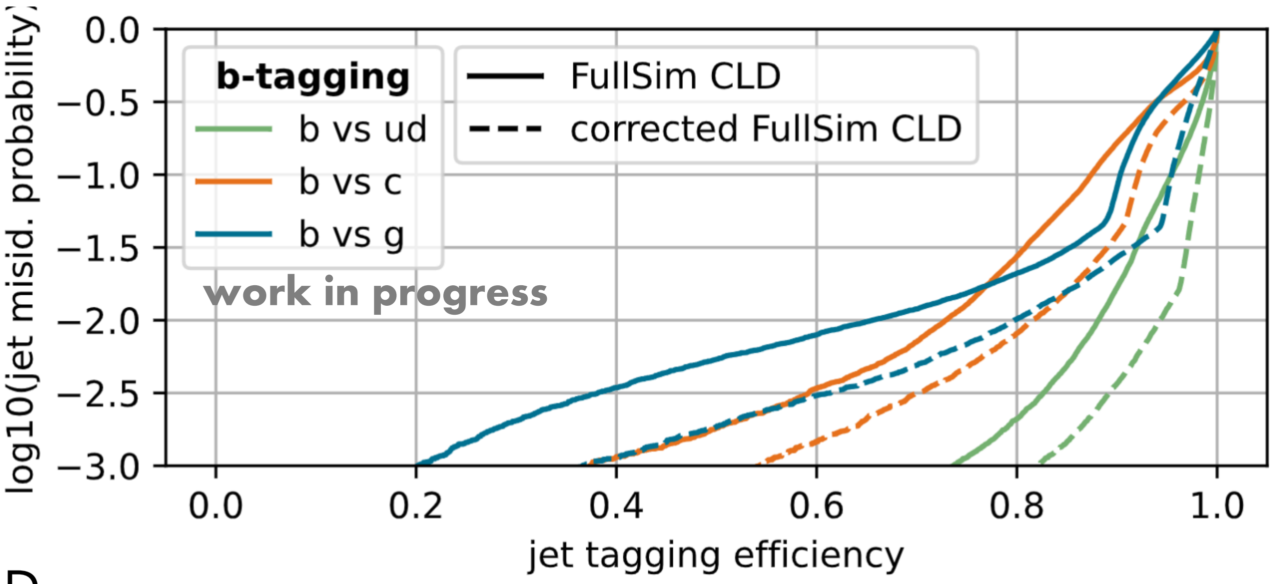
Improving full sim tagging?

- Improve input data to neural network
- Use all tracks available!
- Ignore fake neutrals

Idea:

Instead of PFOs (particle flow objects) use

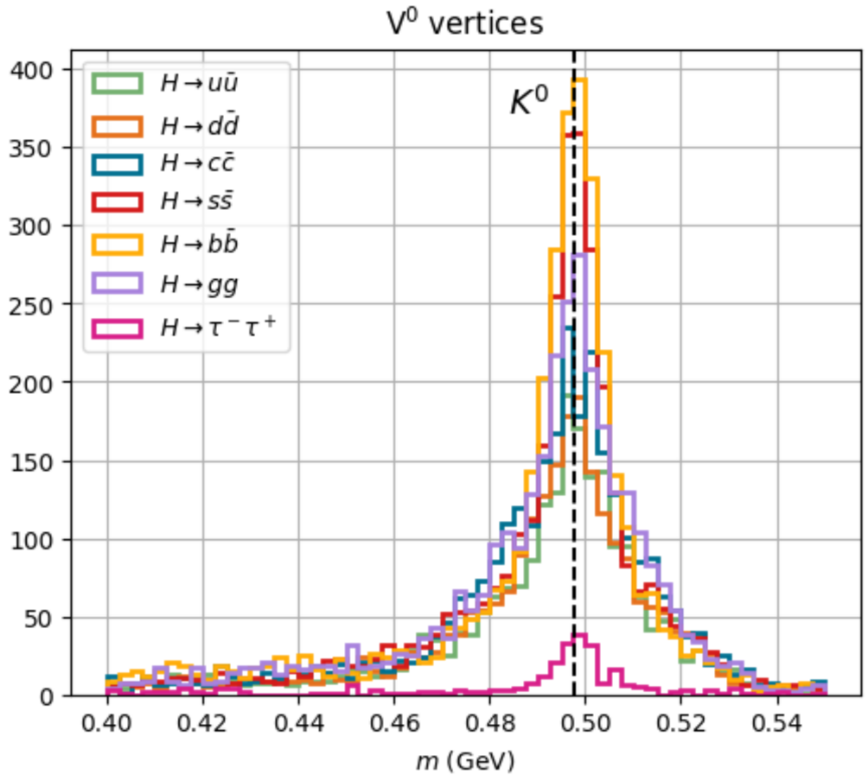
- Tracks for charged particles
- PFOs for neutral particles but check MC PID to avoid double counting

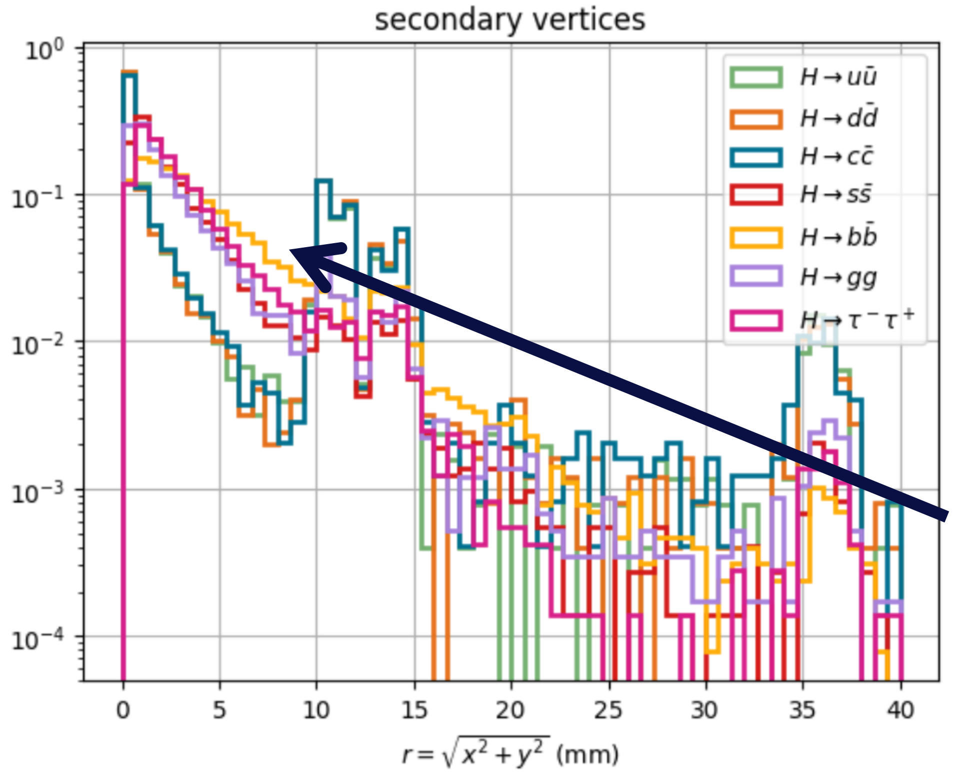


Large improvement:
e.g. misidentification probability of 10^{-2} for b vs. ud:
Improves from 88% to 94% (fast sim: 97%)

Adding vertex information

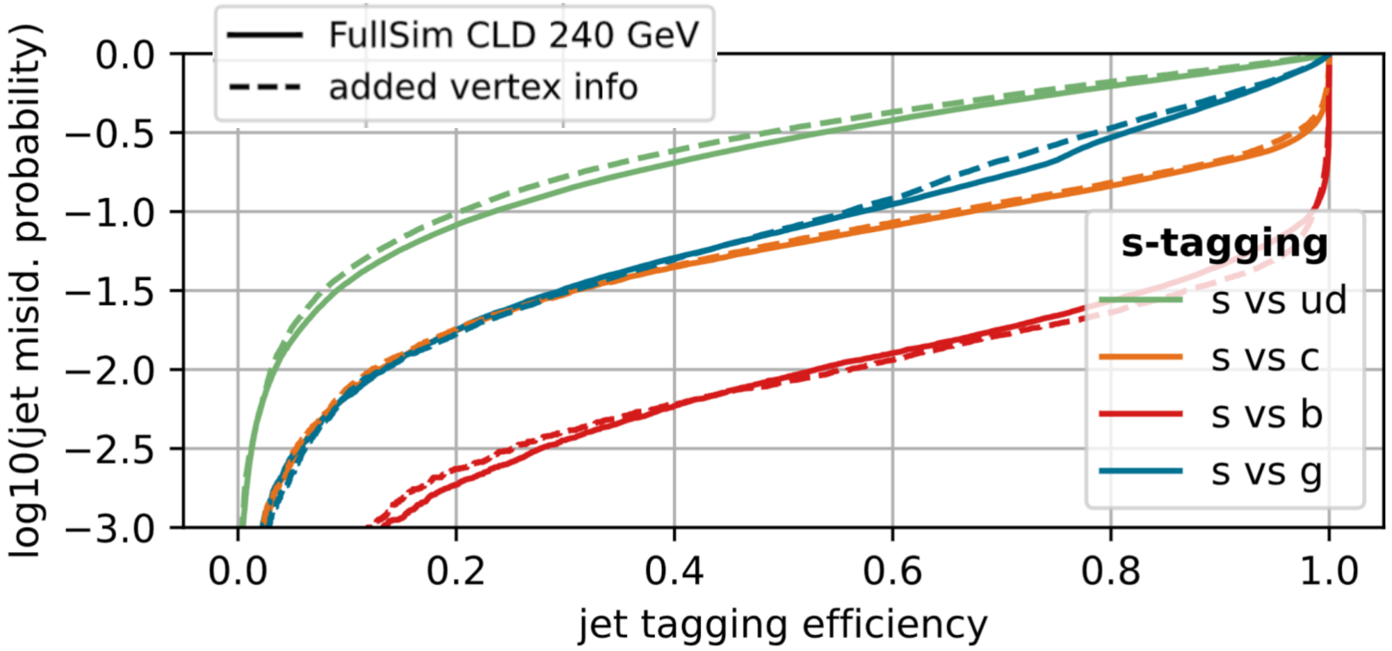
- Use V^0s and secondary vertices
- Add vertex position and invariant mass as input to neural network





Larger vertex displacements for b vs light quarks

Adding vertex information



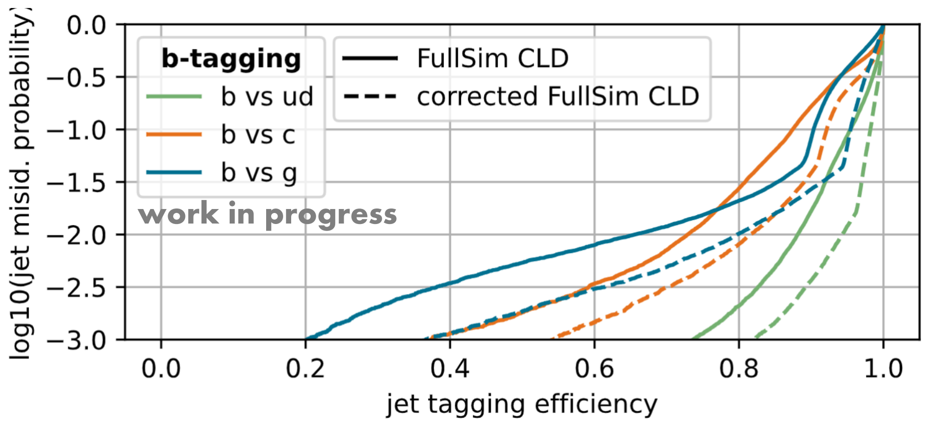
→ performance does not improve

→ network learns information by itself e.g.

- vertex positions through track displacements
- invariant mass at vertices through track kinematics

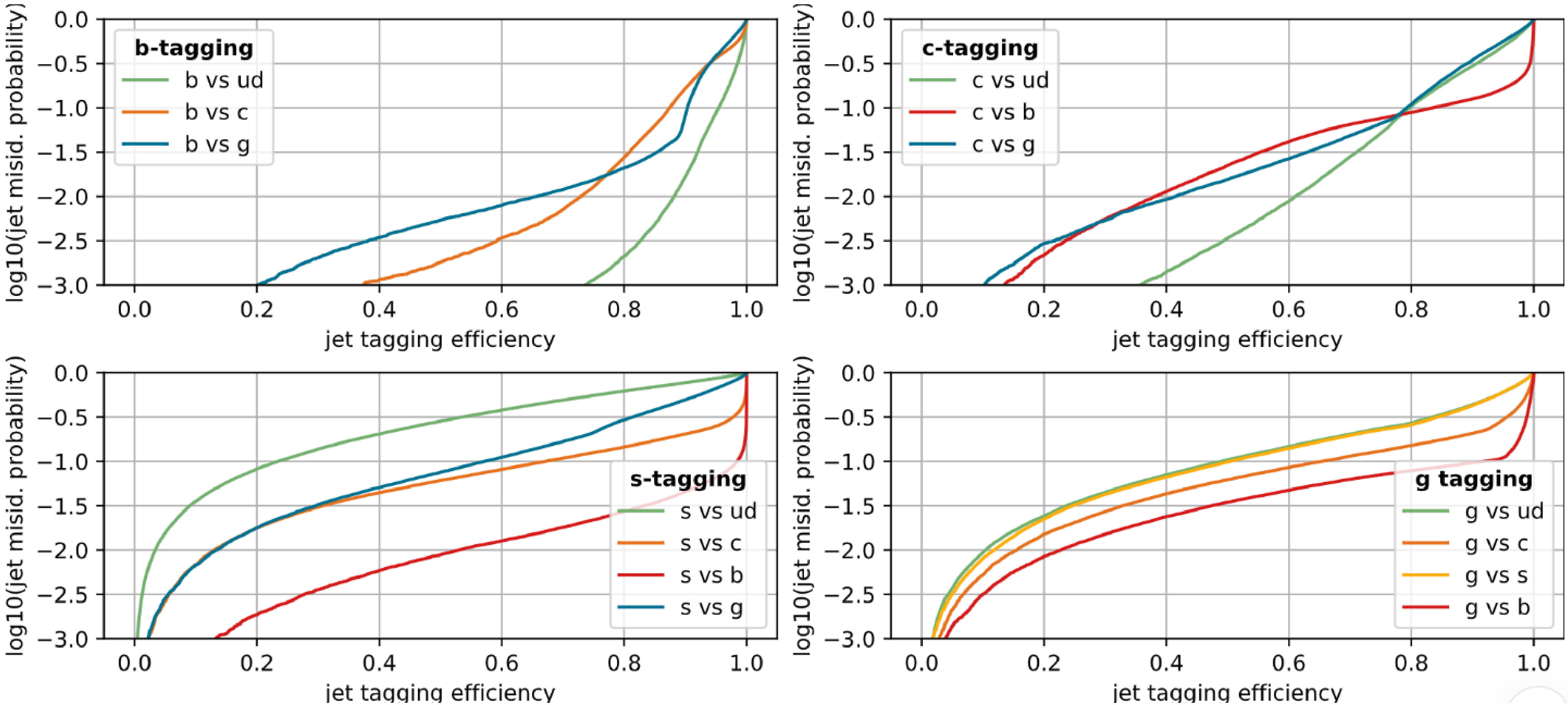
Summary

- Full vs. fast sim jet description in good agreement but
  - Fake neutrals in full sim
  - Lost tracks in full sim
- First implementation of jet flavor tagging in full sim at CLD
- Possible improvement of full sim tagging through better reconstruction
- Adding vertex information does not improve tagging performance



Backup

Full sim jet flavor tagging



Input parameters to network

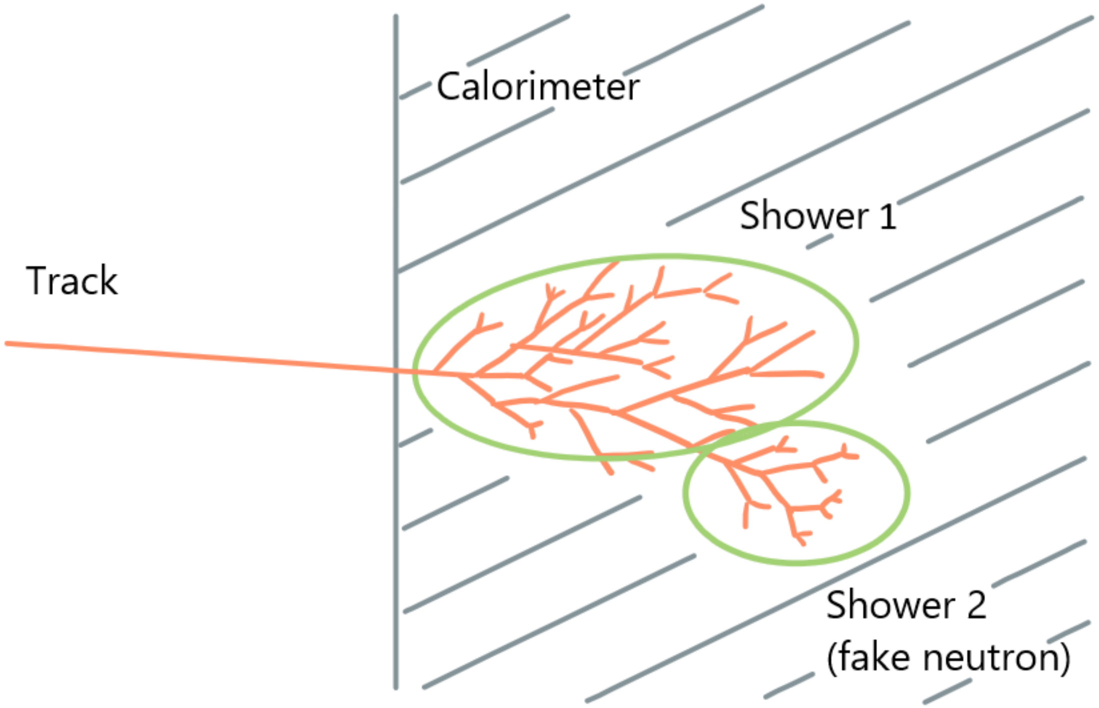
FUTURE CIRCULAR COLLIDER

Table 1. Set of input variables

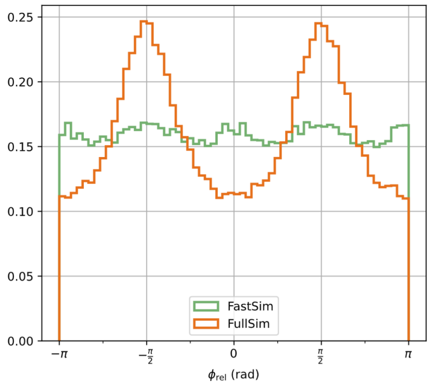
Variable	Description
	Kinematics
$E_{\mathrm{const}}/E_{\mathrm{jet}}$	energy of the jet constituent divided by the jet energy
θ_{rel}	polar angle of the constituent with respect to the jet momentum
ϕ_{rel}	azimuthal angle of the constituent with respect to the jet momentum
	Displacement
d_{xy}	transverse impact parameter of the track
d_z	longitudinal impact parameter of the track
$\mathrm{SIP}_{\mathrm{2D}}$	signed 2D impact parameter of the track
$\mathrm{SIP}_{\mathrm{2D}}/\sigma_{\mathrm{2D}}$	signed 2D impact parameter significance of the track
$\mathrm{SIP}_{\mathrm{3D}}$	signed 3D impact parameter of the track
$\mathrm{SIP}_{\mathrm{3D}}/\sigma_{\mathrm{3D}}$	signed 3D impact parameter significance of the track
d_{3D}	jet track distance at their point of closest approach
$d_{\mathrm{3D}}/\sigma_{d_{\mathrm{3D}}}$	jet track distance significance at their point of closest approach
C_{ij}	covariance matrix of the track parameters
	Identification
q	electric charge of the particle
~~$m_{\mathrm{t.o.f.}}$~~	~~mass calculated from time-of-flight~~
~~dN/dx~~	~~number of primary ionisation clusters along track~~
`isMuon`	if the particle is identified as a muon
`isElectron`	if the particle is identified as an electron
`isPhoton`	if the particle is identified as a photon
`isChargedHadron`	if the particle is identified as a charged hadron
`isNeutralHadron`	if the particle is identified as a neutral hadron

from IDEA fast sim tagging

From ϕ_{rel} to fake neutrons

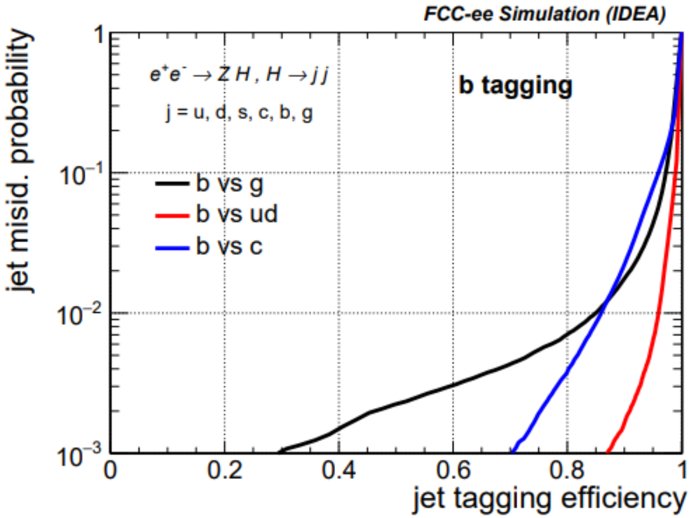


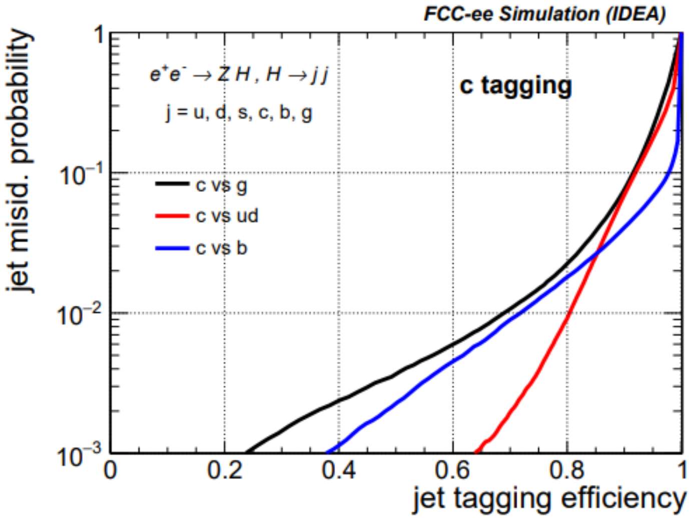
- If constituents and jet have similar ϕ, θ then $\phi_{rel} \rightarrow \pm\frac{\pi}{2}$
- High energetic charged particle dominate jet kinematics
- Fake neutron similar angles as charged particle, so also similar angles to jet → peaks in distribution



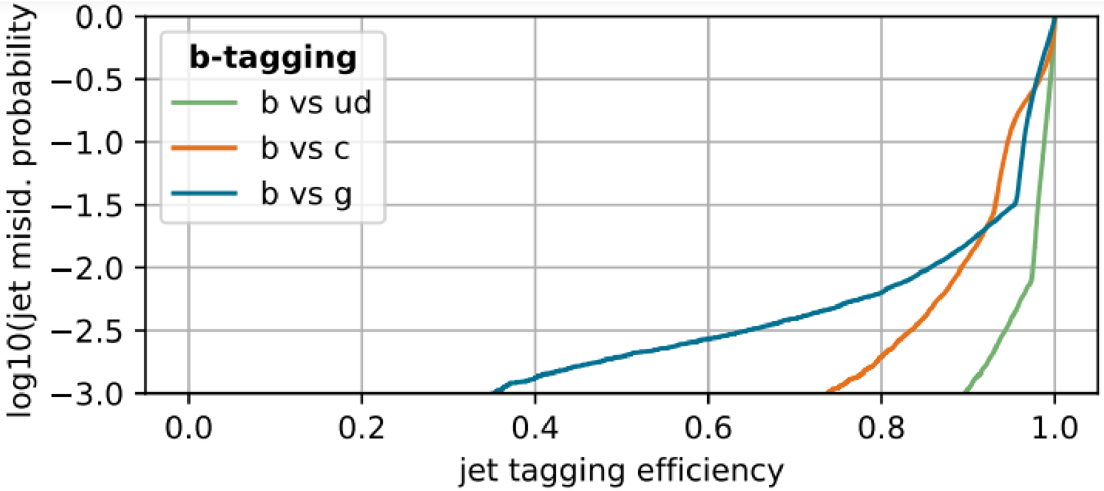
Fast sim CLD* vs. IDEA

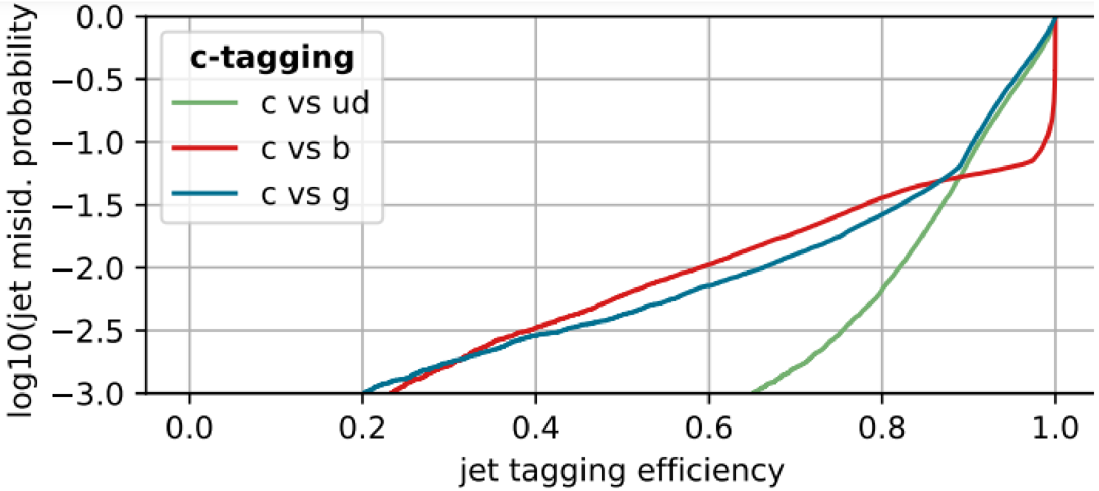
FUTURE CIRCULAR COLLIDER





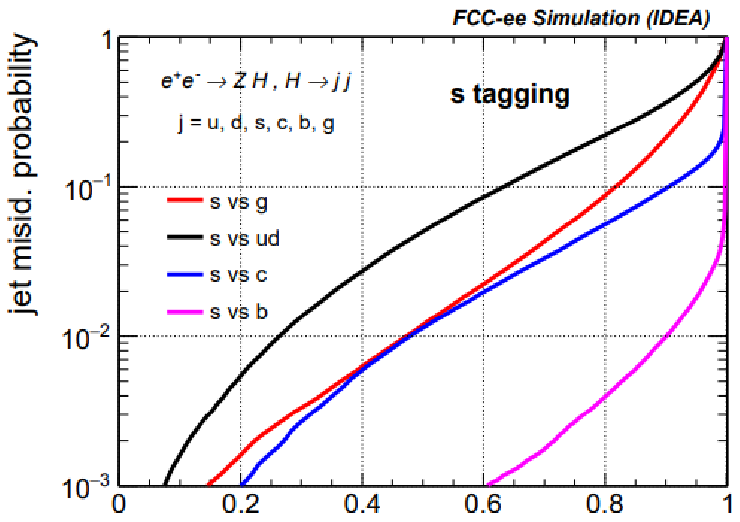
*"CLD fast sim" is IDEA fast sim with silicon tracker

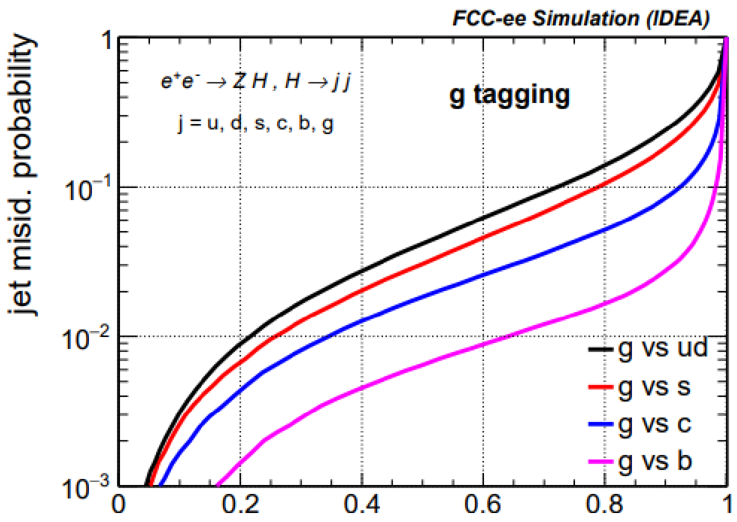




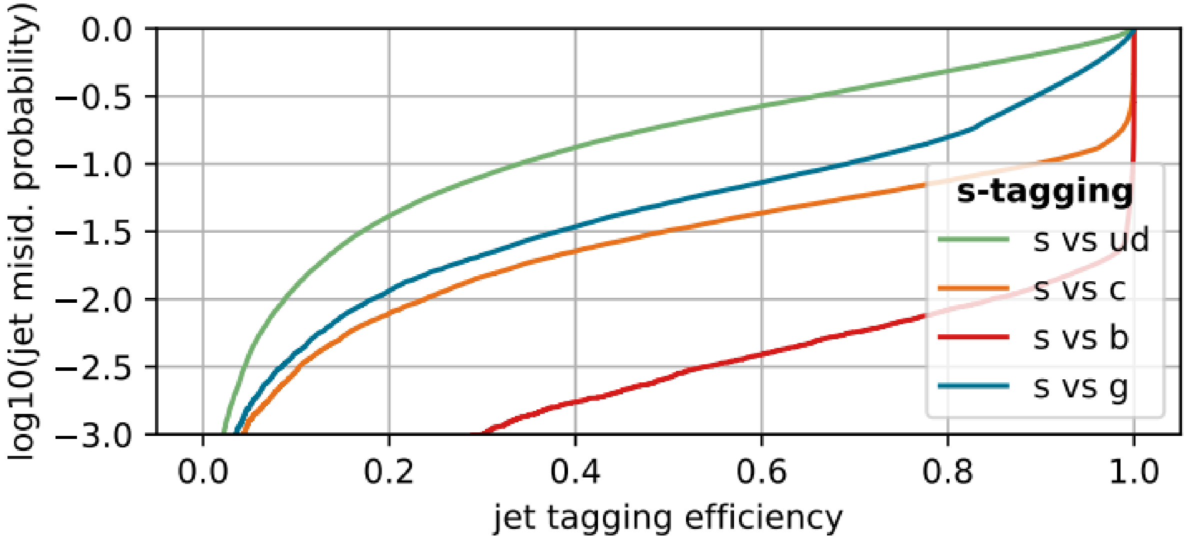
Fast sim CLD* vs. IDEA

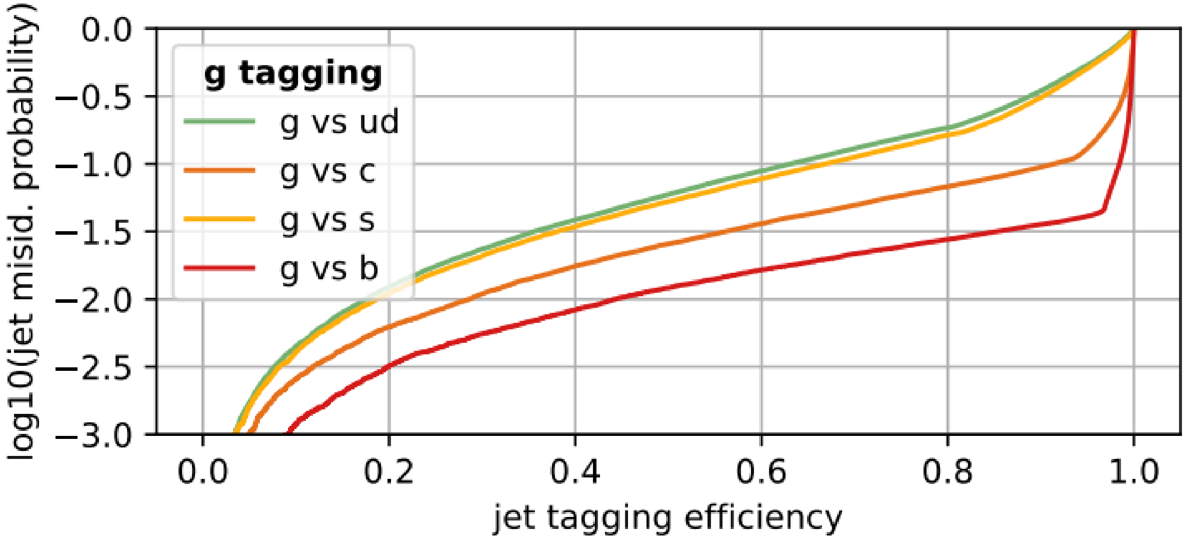
FUTURE CIRCULAR COLLIDER





*"CLD fast sim" is IDEA fast sim with silicon tracker





Multiplicities