ML-based Jet Flavor Tagging in Fast and Full Simulation

FUTURE CIRCULAR COLLIDER

8th FCC Physics Workshop

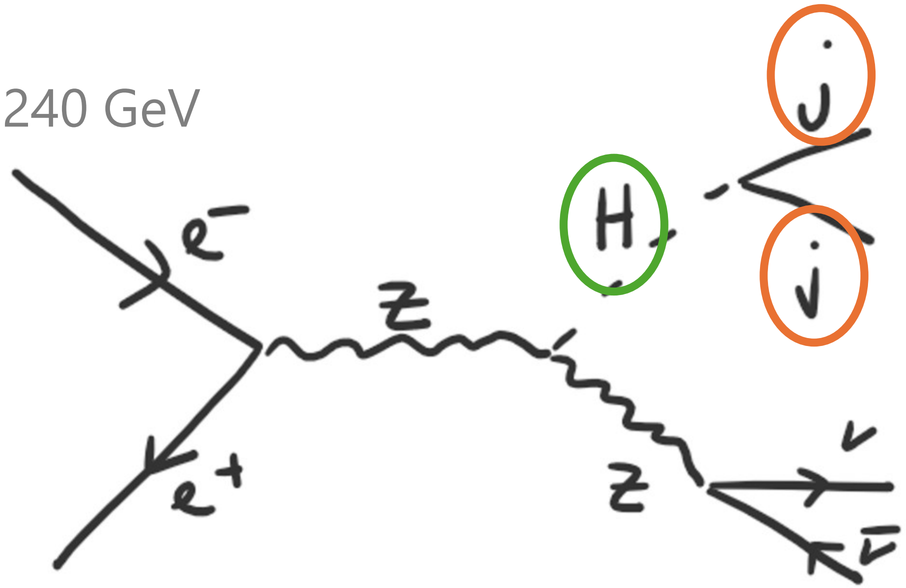
16. January 2025

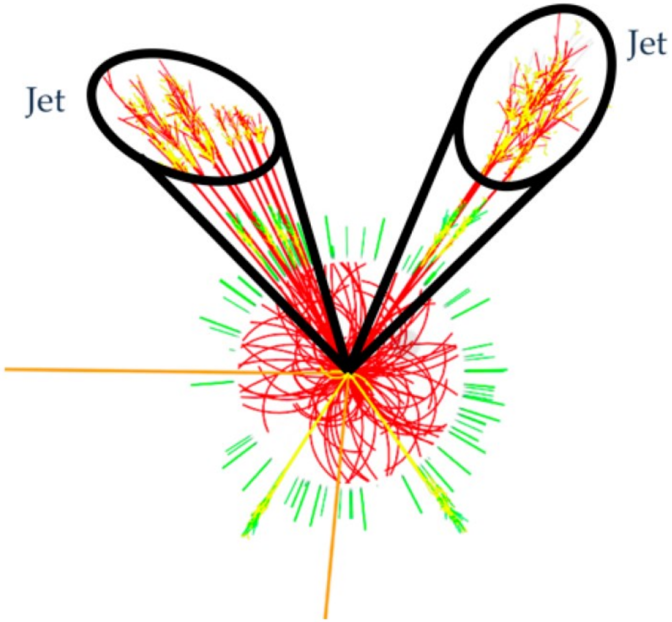
Sara Aumiller, Dolores Garcia, Michele Selvaggi

FUTURE CIRCULAR COLLIDER

Why Jet-Flavor Tagging?

at 240 GeV





Future Colliders

=

Higgs factories for precious measurements

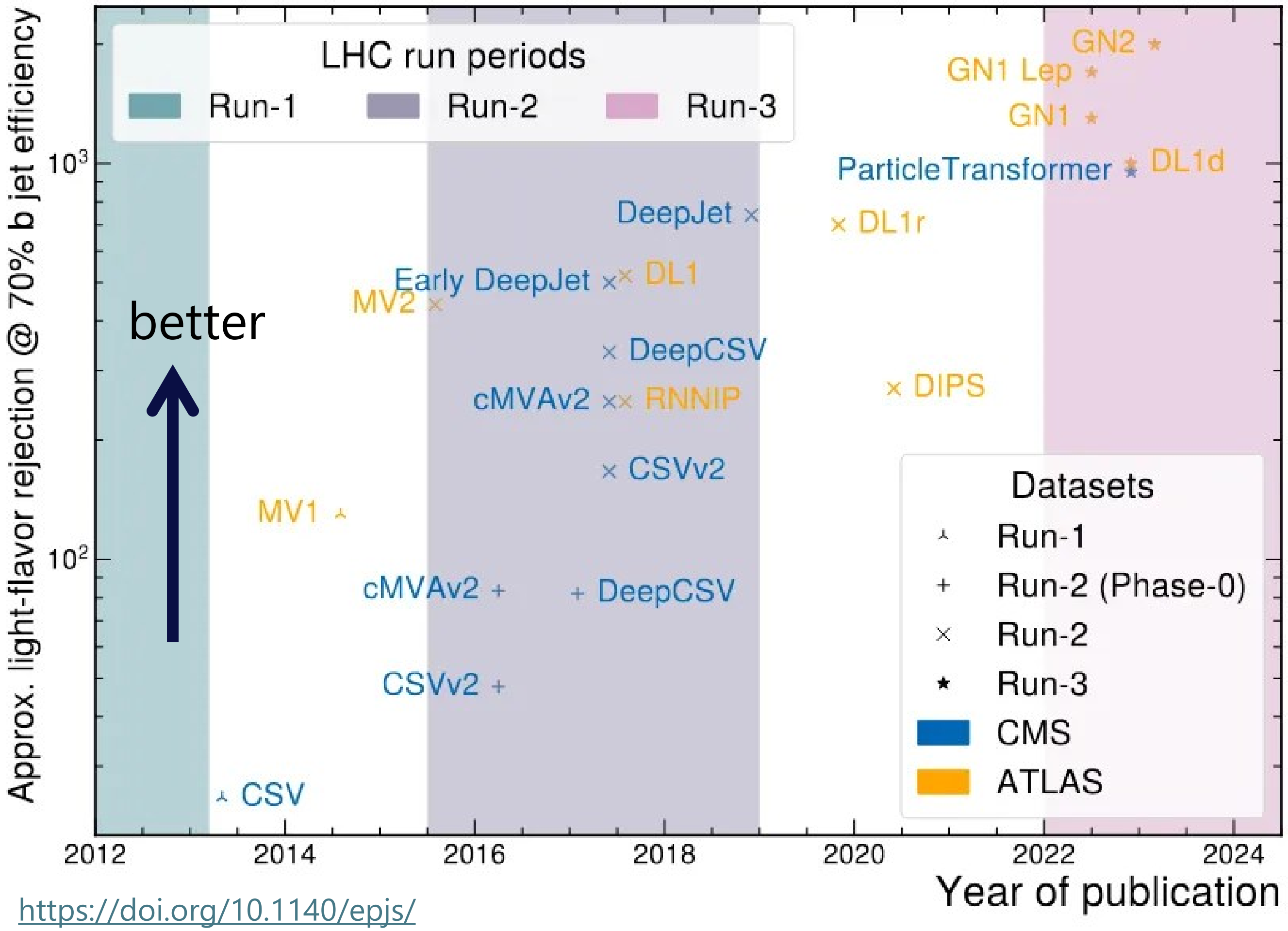
Particles causing jets:

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

Why use Machine Learning (ML)?

2045?

Performance of the tagger

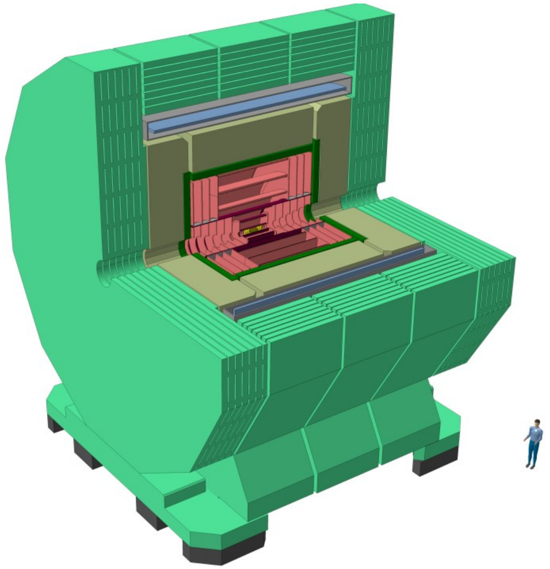


Stunning improvement of jet-flavor tagging through ML over the last decade

https://doi.org/10.1140/epjs/s11734-024-01234-y

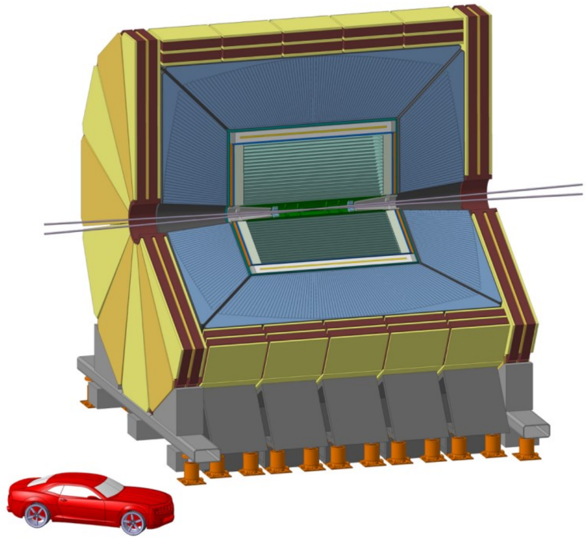
Fast & Full Simulation at FCC-ee

https://arxiv.org/pdf/1911.12230



CLD

https://doi.org/10.22323/1.414.0337



IDEA

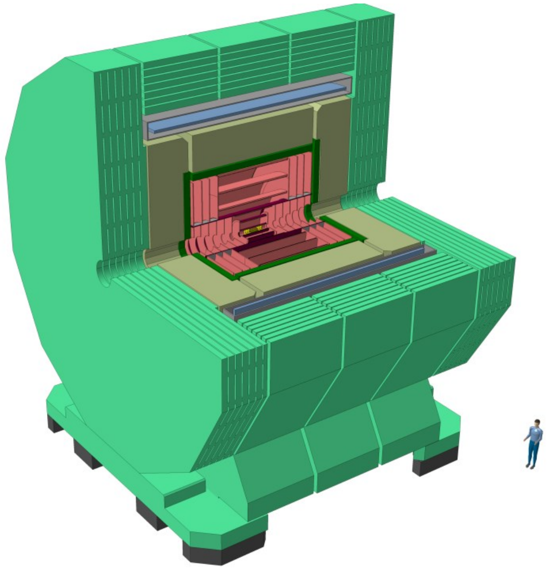
Fast simulation

time & computational efficient early-stage feasibility studies

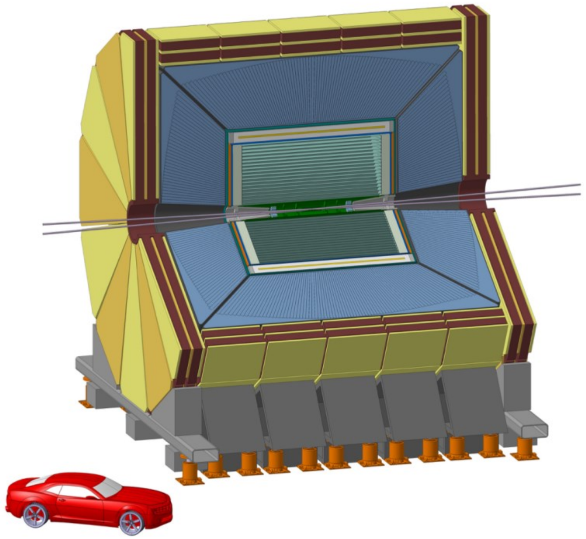
Full simulation

more realistic description of detector concept and reconstruction algorithms

Fast & Full Simulation at FCC-ee



https://arxiv.org/pdf/1911.12230



https://doi.org/10.22323/1.414.0337

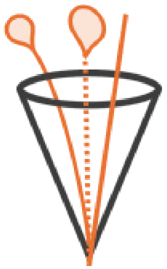
	Fast simulation	**Full simulation**
	time & computational efficient early-stage feasibility studies	more realistic description of detector concept and reconstruction algorithms
CLD	IDEA fast simulation with silicon tracker ✓	Using GEANT4 and Marlin ✓
IDEA	Using DELPHES ✓	-

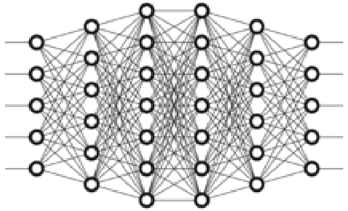
Major difference between CLD & IDEA:
IDEA can retreive **PID** via dNdx & ToF

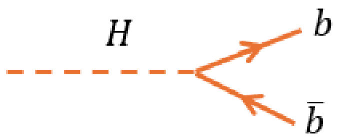
Jet-Flavor Tagging Set-Up

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

reconstruction	Jet description	Neural network	Jet flavor	Higgs physics
Cluster N jets per event	# constituents, kinematics etc.	Particle Transformer	7 flavors u, d, s, c, b, g, τ	Yukawa couplings etc.

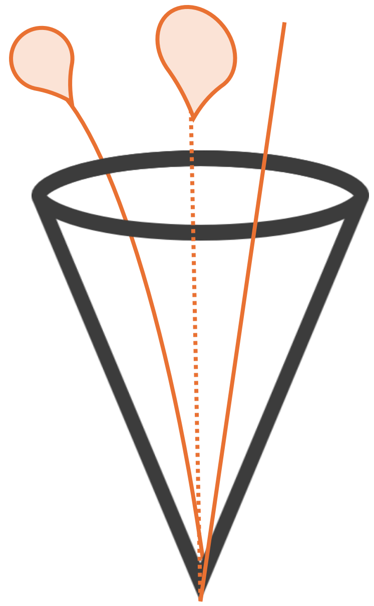






Focus of this work

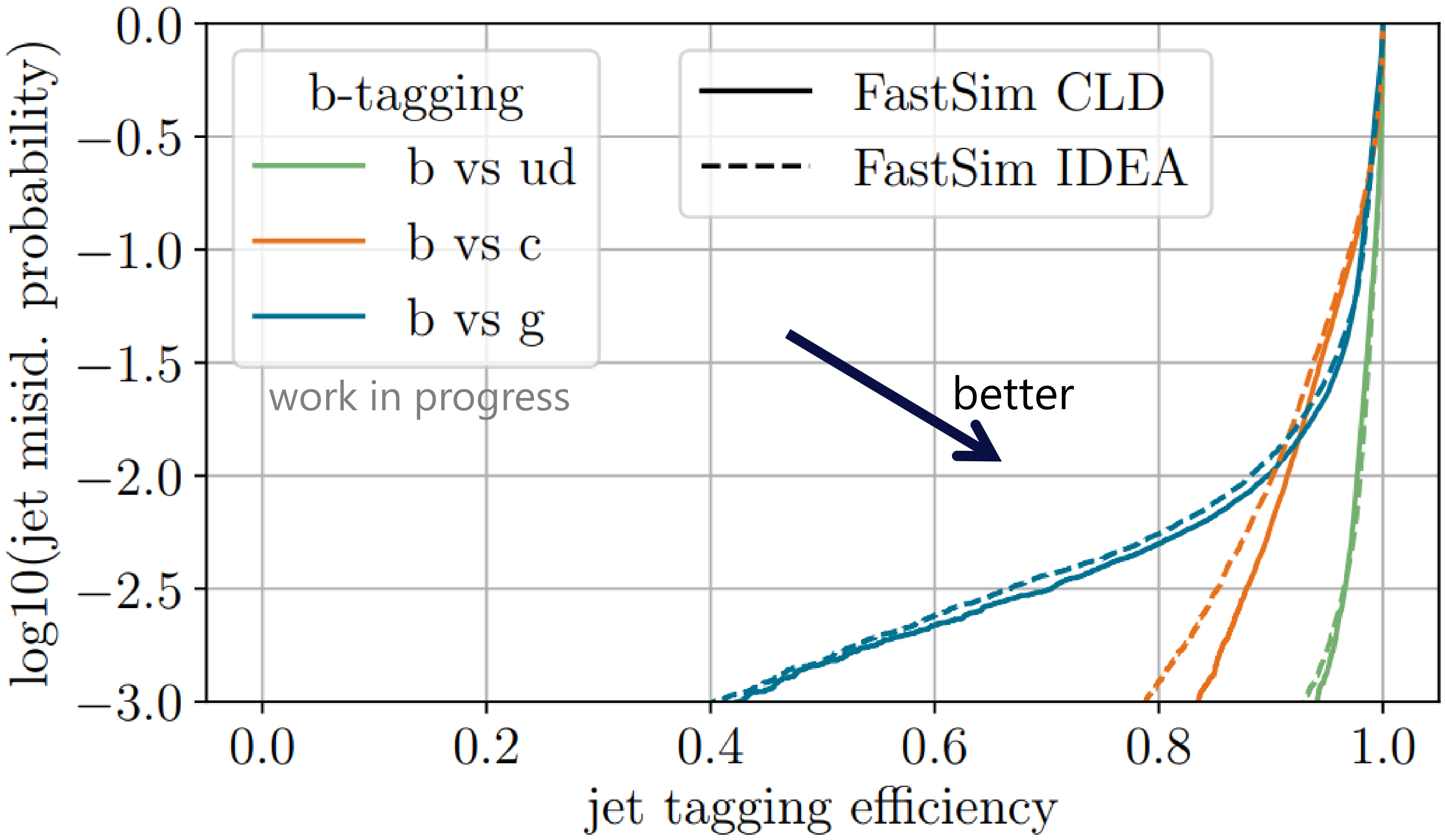
Jet Description



We characterize the **jet constituents**:

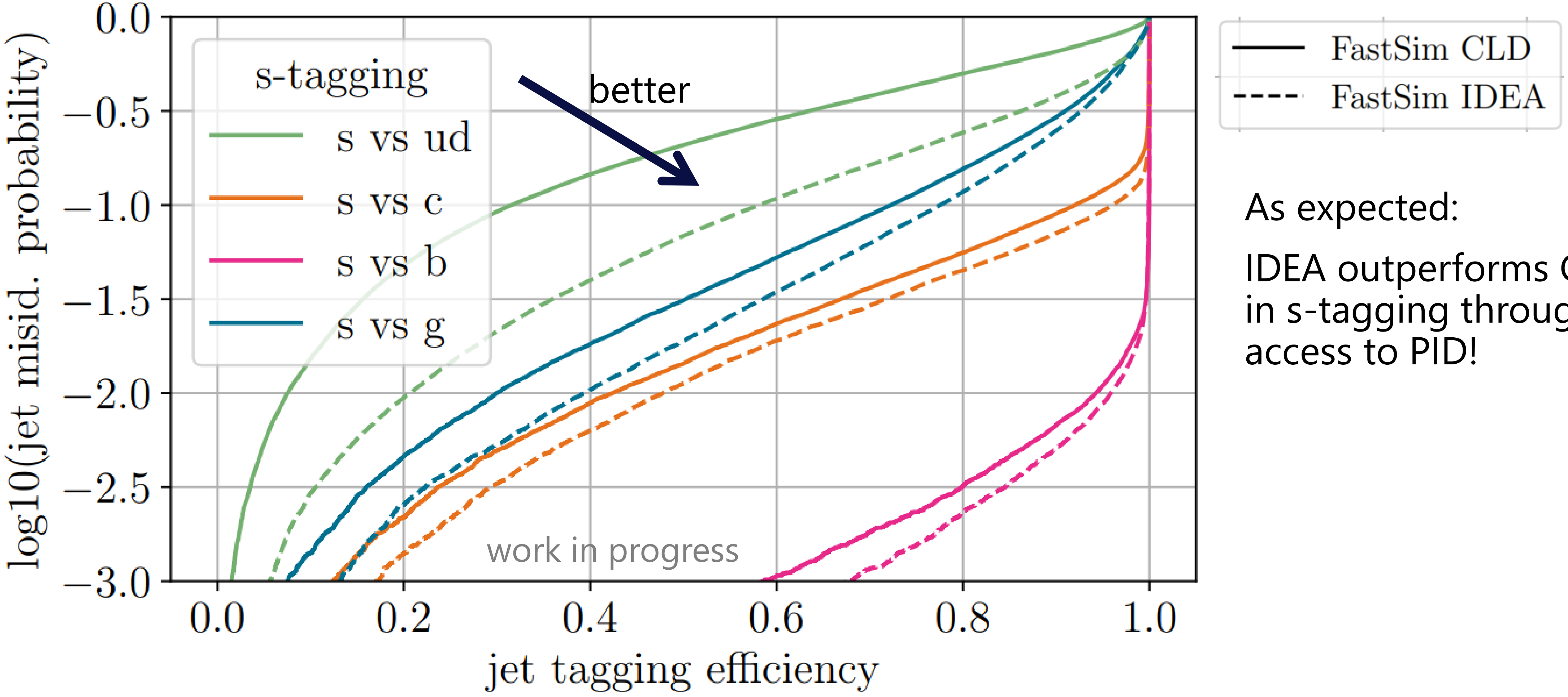
Kinematics (3)	Identification (7)	Track displacements (23)
$\log E_{rel}$, θ_{rel}, ϕ_{rel}	reco PID, charge, PID flags, (dNdx, ToF for IDEA)	d_0, z_0, covariance matrix c_{ij}, SIP in 2D, 3D (& significance), Jet-track distance d_{3D} (& sig.)

Jet Tagging in Fast Simulation



Good agreement in performance between CLD and IDEA

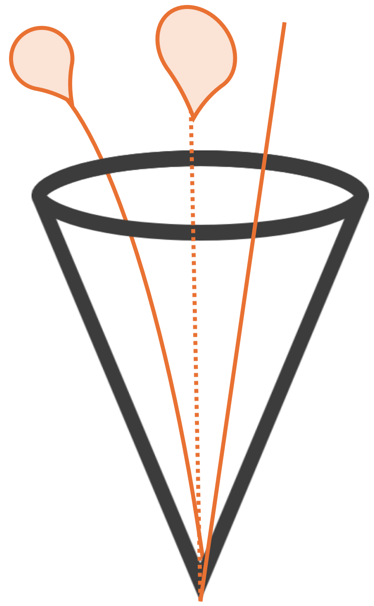
Jet Tagging in Fast Simulation



As expected:

IDEA outperforms CLD in s-tagging through access to PID!

Jet Description in Full Simulation

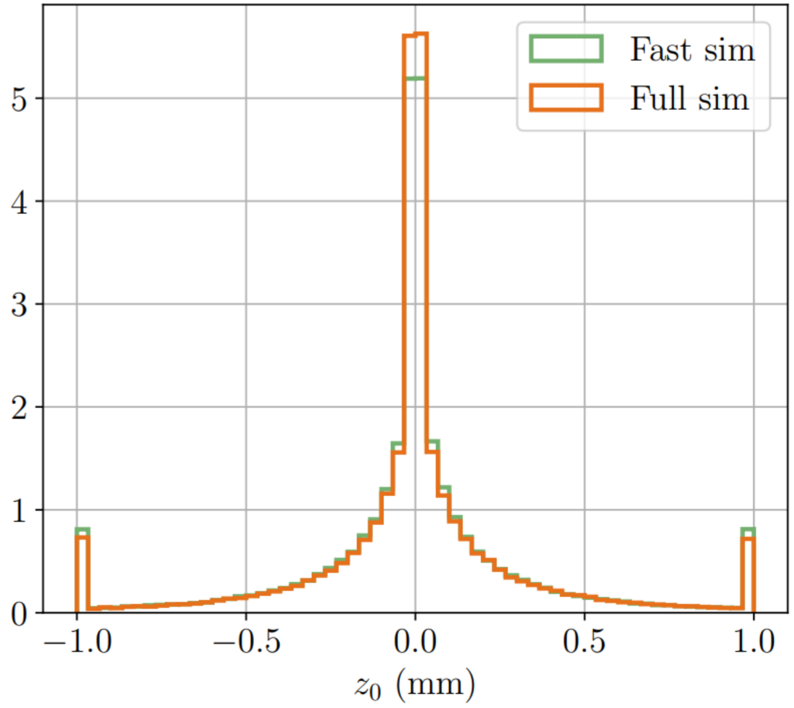


We need to **validate the jet description** in **full simulation** comparing it to fast simulation!

→ Comparison shows mostly good agreement:

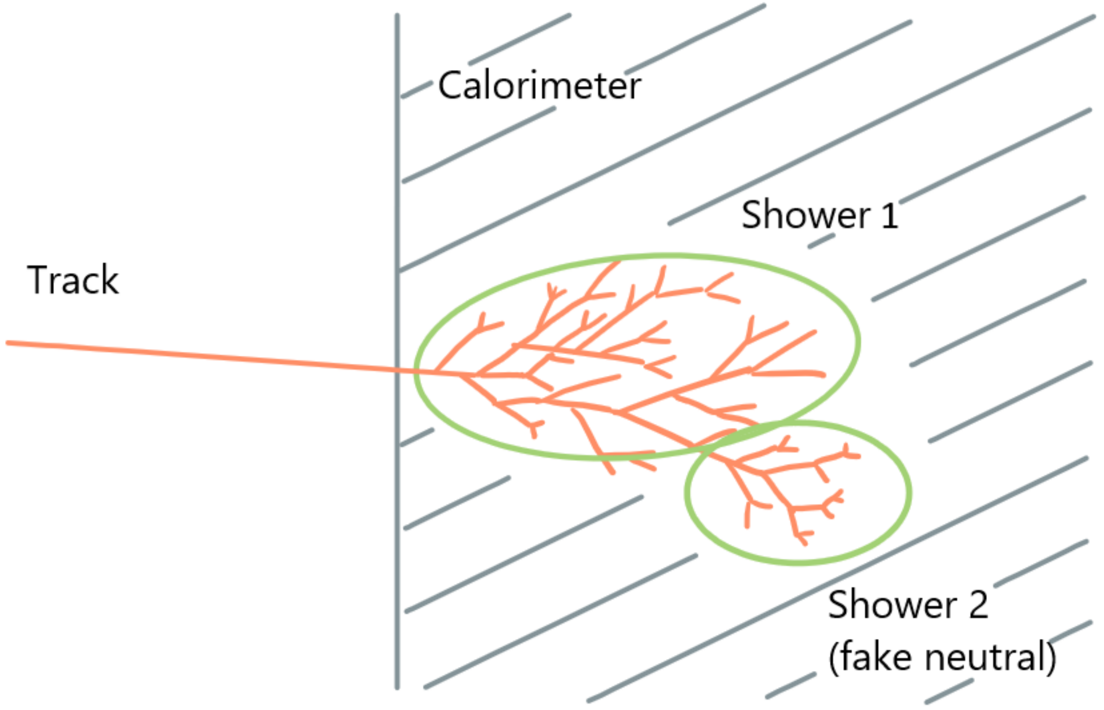
Two major differences in full simulation:

1. Fake neutrals
2. Unassociated tracks to PFOs



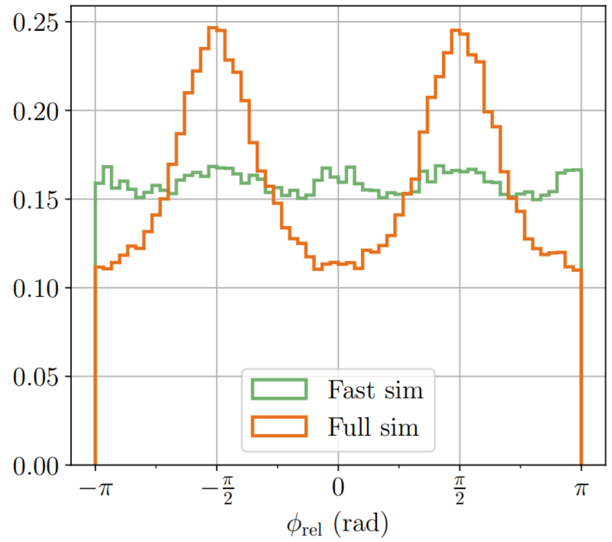
Longitudinal impact parameter of leading tracks for $H \to b\bar{b}$

(1) Fake neutrals in full sim



Artificially split cluster of high-energy charged particles (at MC level) creates **fake neutral.**

- More neutral hadrons in full than in fast simulation
- Relative angle ϕ of neutral jet constituents shows discrepancy



leading neutral hadronic jet constituents

(2) Unassociated tracks to PFOs in full simulation

Some **charged particles** are wrongly reconstructed as **neutral PFOs** in full sim although the track efficiency is high.

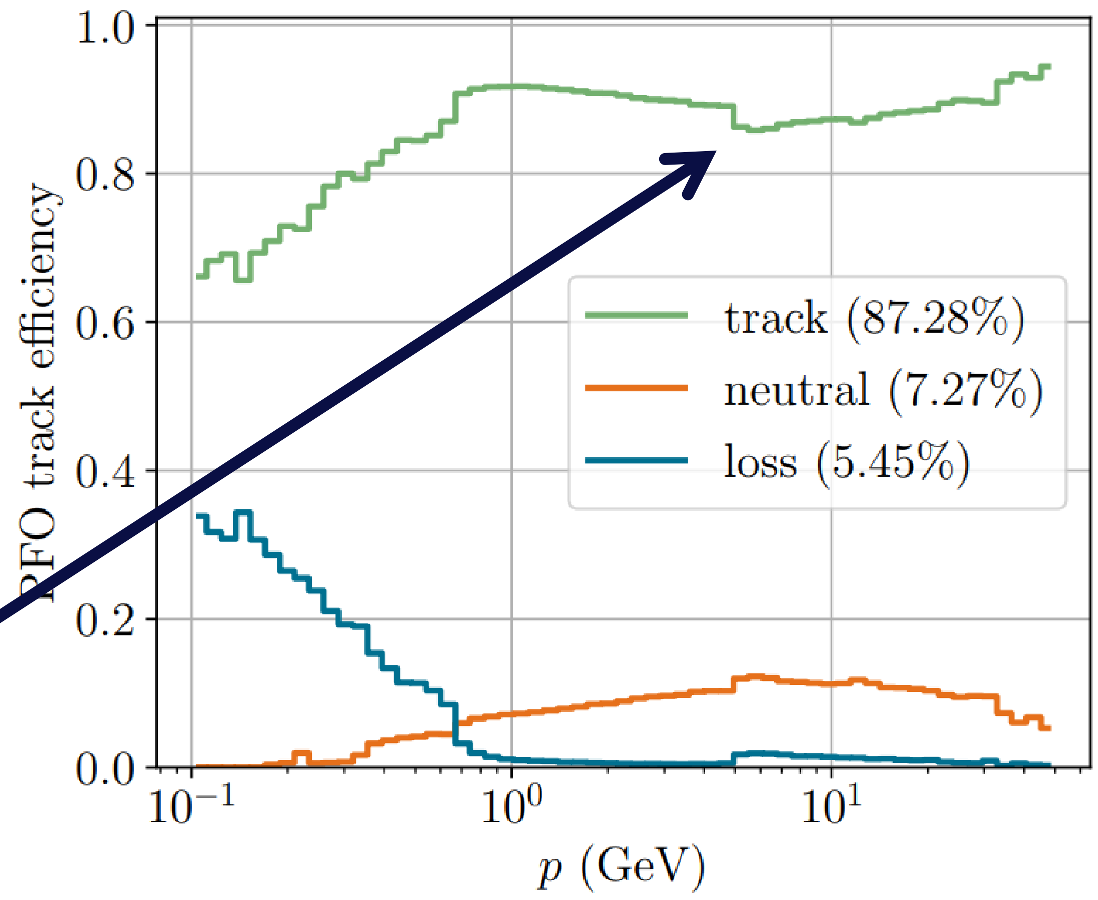
→ **track-cluster association fails**

→ problematic as tracks are crucial for jet flavor tagging

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

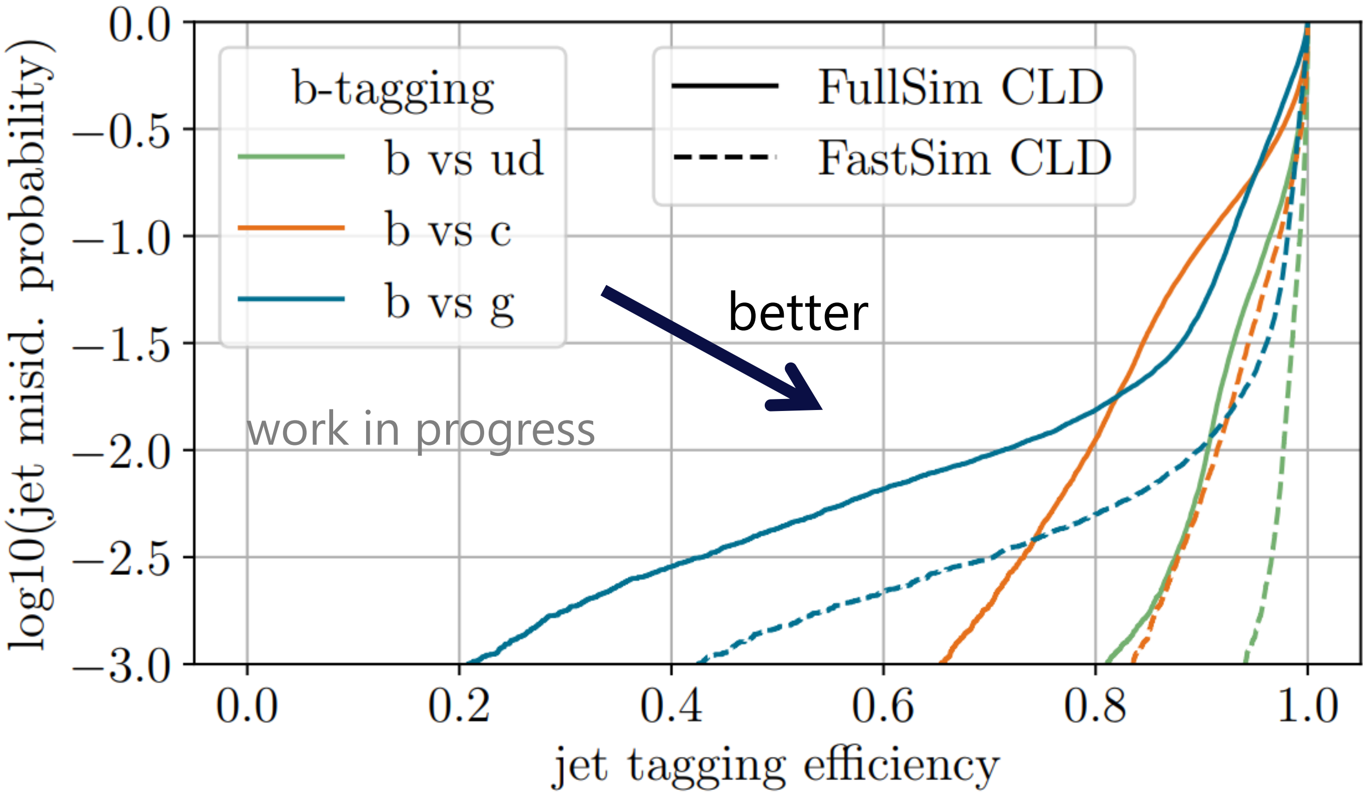
→ reconstruction could be improved

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



See Anna Zaborowska talk!

Full vs. Fast Simulation CLD



Loss in performance in full simulation

e.g. at a misidentification probability of 10^{-2} for b vs. ud:

Efficiency drops from 97% (fast sim) / 90% (full sim)

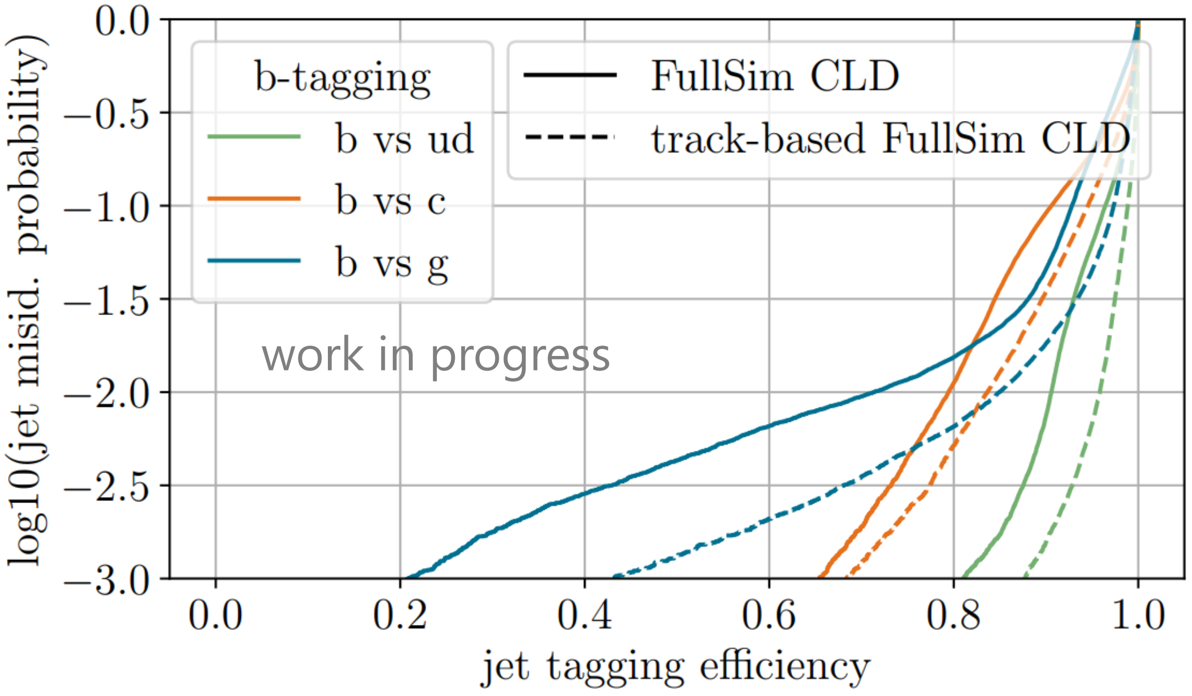
Improving Full Sim Tagging

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

Idea:

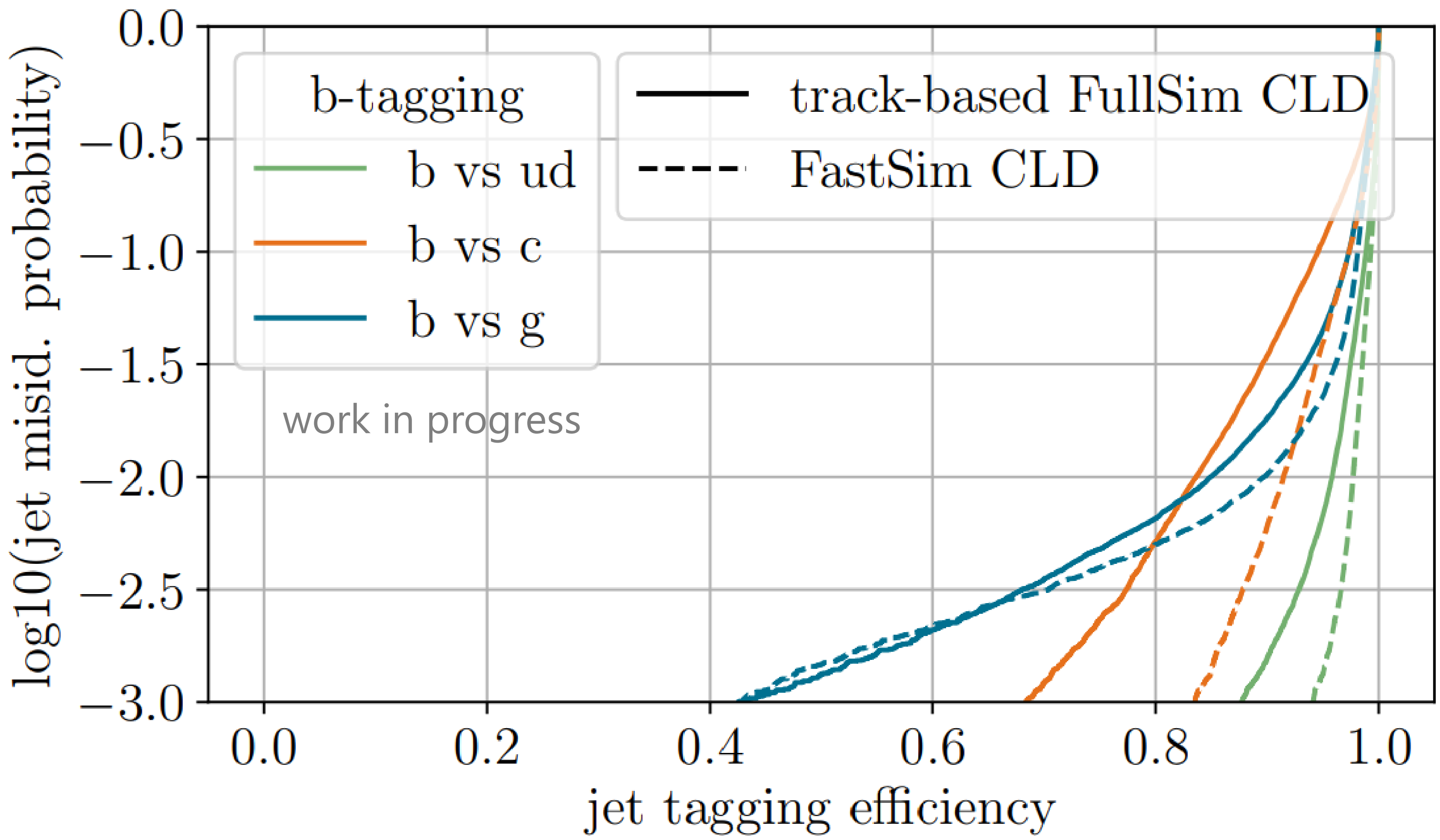
Instead of PFOs (particle flow objects) use

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



Large improvement:
e.g. at a misidentification probability of 10^{-2} for b vs. ud:
Efficiency improves from 90% to 95% (fast sim: 97%)

Fast vs. track-based Full Sim



Using

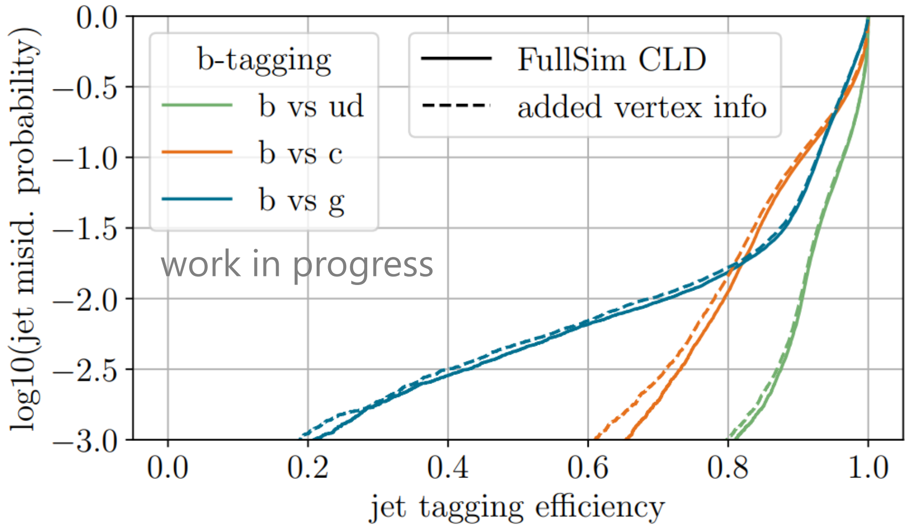
- Tracks for charged particles
- PFOs for neutral particles
- Plus some MC checking

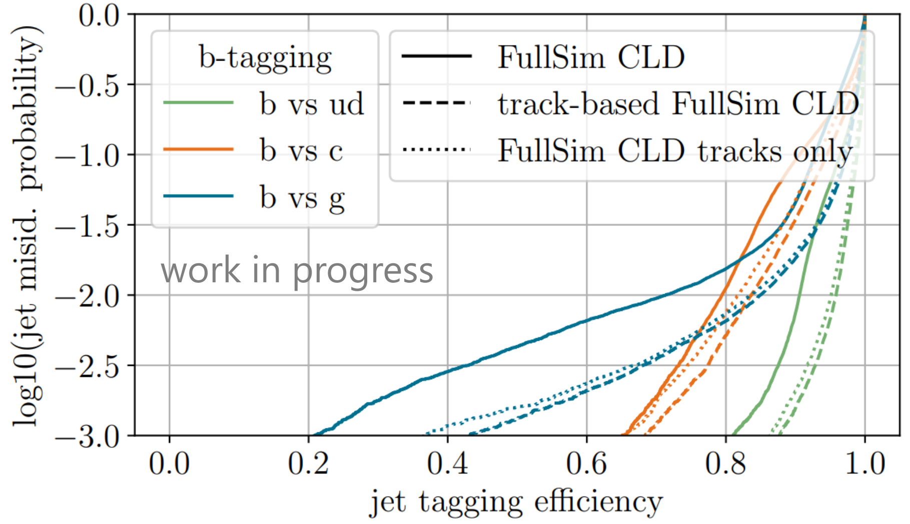
Before:

- Purely PFO based

Other Studies

1. Adding Vertex Information
   - Important for b- and c-tagging
   - Added location and mass of secondary vertices and V0
   - **Performance**: Does not improve
   - **Conclusion:** Network learns information on its own
2. Using tracks only
   - Resolves the problem of lost tracks in PFO creation
   - Solves the problem of fake neutrals
   - **Performance**: Good for b-tagging but in other cases not as good as corrected PFO input
   - **Conclusion**: Work on PF algorithm encouraged





Further details in FCC note

Full Sim Key4Hep Implementation

We want to make full simulation tagging at CLD **available to everyone** by implementing it to key4hep.

https://github.com/saracreates/JetTagging

Include the JetTagger in the steering file:

Adds 7 PID collections "RefindJetTag_X" with flavors X

```
flavor_collection_names = ["RefinedJetTag_G", "RefinedJetTag_U", "RefinedJetTag_S", "RefinedJetTag_C", "RefinedJetTag_B",
"RefinedJetTag_D", "RefinedJetTag_TAU"]
transformer = JetTagger("JetTagger",
                        model_path="/afs/cern.ch/work/s/saaumill/public/onnx_export/fullsimCLD240_2mio.onnx",
                        json_path="/afs/cern.ch/work/s/saaumill/public/onnx_export/preprocess_fullsimCLD240_2mio.json",
                        flavor_collection_names = flavor_collection_names, # to make sure the order and nameing is correct
                        InputJets=["RefinedVertexJets"],
                        InputPrimaryVertices=["PrimaryVertices"],
                        OutputIDCollections=flavor_collection_names)
```

Uses a k4FWCore Transformer

FUTURE CIRCULAR COLLIDER

Full Sim Key4Hep Implementation

We want to make full simulation tagging at CLD **available to everyone** by implementing it to key4hep.

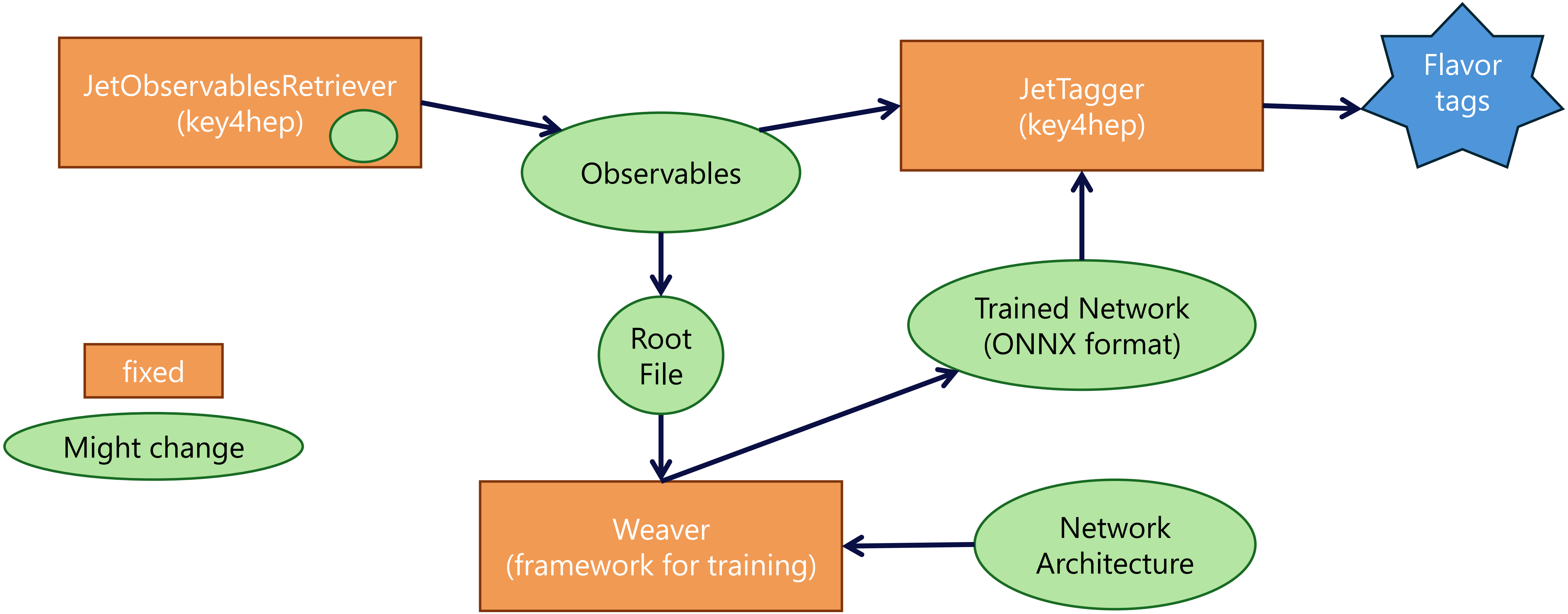
https://github.com/saracreates/JetTagging

Special **thanks** to
Brieuc Francois,
Thomas Madlener
and especially
Leonhard Reichenbach
for their support!

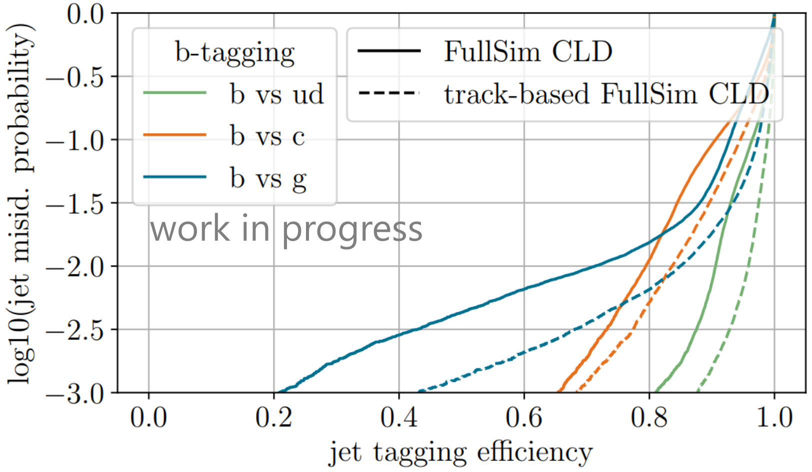
Status:

- Implementation is done! 🎉
- **Currently**: **Validating** the performance
    - run inference on the key4hep pipeline
    - Recreate ROC curves and compare performance
- **Outlook:** Implementing **full life cycle of tagging** for quick adjustments in the future
    - Retrieve jet constituent variables / network input conveniently from key4hep for easy retraining of a neural network
    - Validate whole life cycle
    - Add thorough documentation

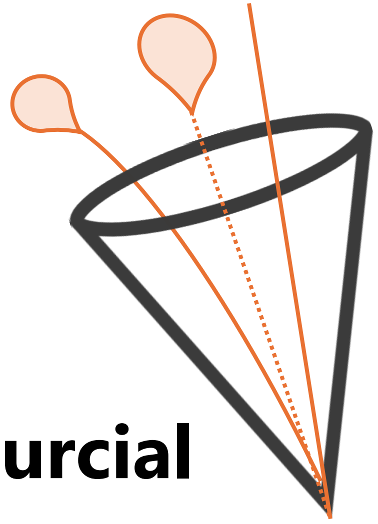
Full cycle for the future



Summary



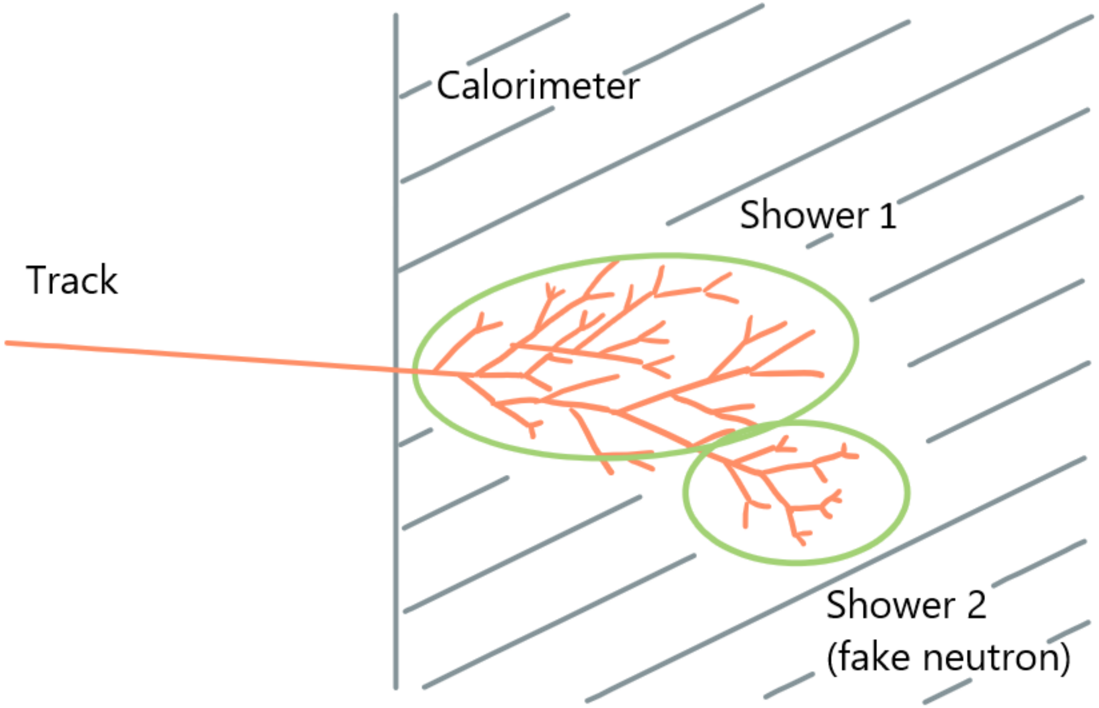
- Studied **jet tagging** in full (CLD) and fast (CLD, IDEA) simulation
- Crystalized two main challenges in full simulation:
  - Fake neutrals
  - Unassociated tracks to PFOs
- Studied options to improve tagging performance in full simulation: **improvement of Pandora Particle Flow is curcial**
- Work-in-progress **key4hep implementation** of CLD full simulation jet tagging for 7 flavors



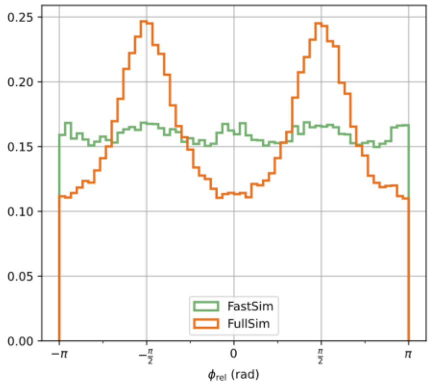
Backup

FUTURE CIRCULAR COLLIDER

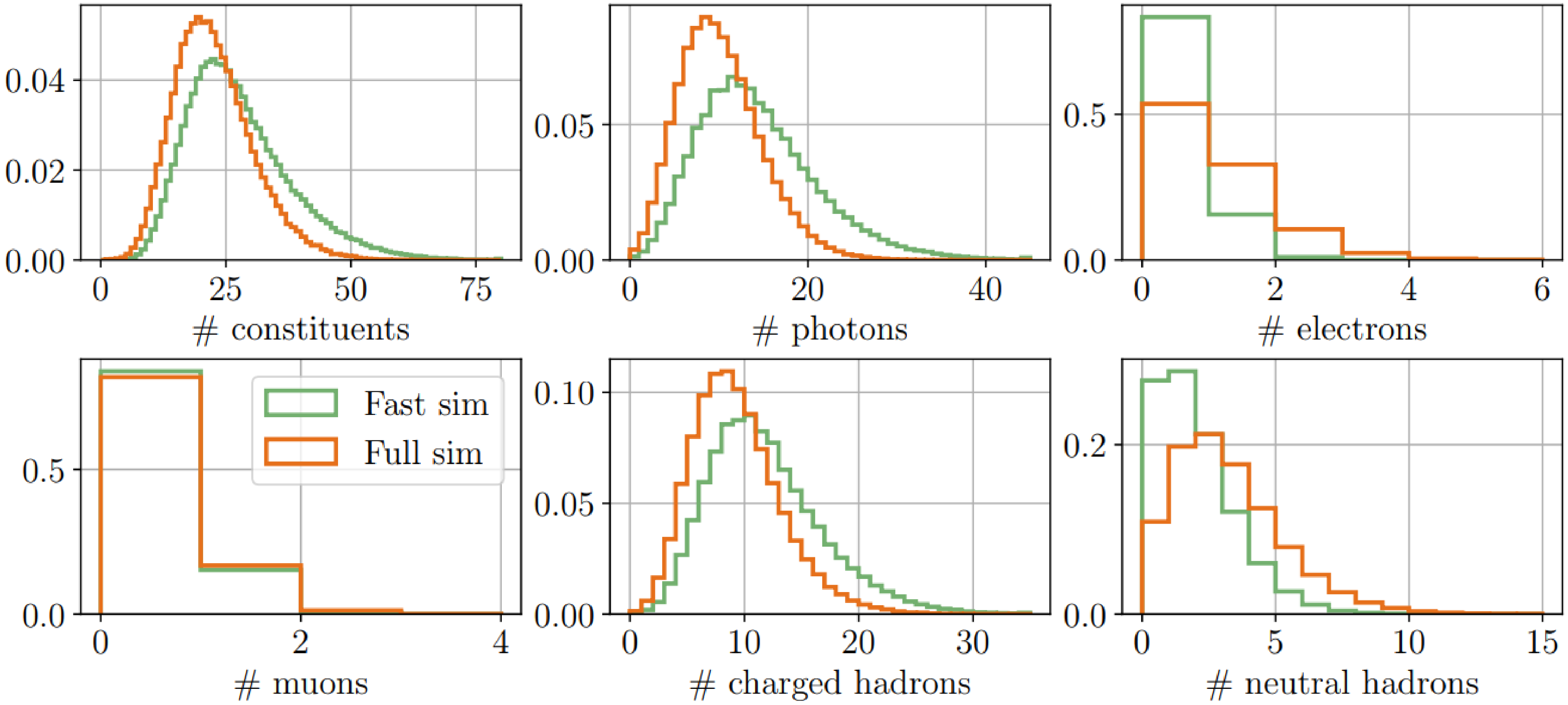
From ϕ_{rel} to fake neutrons



- 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



Multiplicities



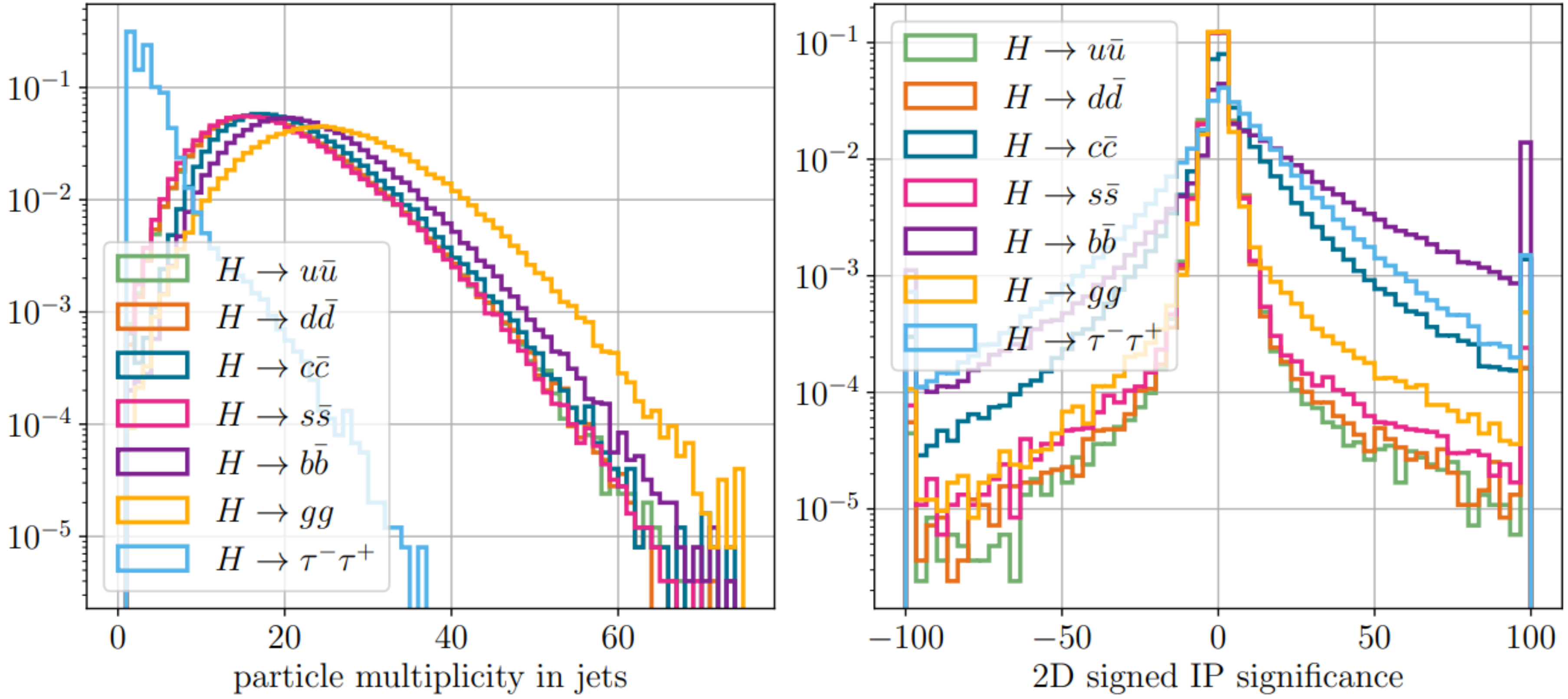
Input parameters to the network

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
$\mathrm{SIP}_{2\mathrm{D}}$	signed 2D impact parameter of the track
$\mathrm{SIP}_{2\mathrm{D}}/\sigma_{2\mathrm{D}}$	signed 2D impact parameter significance of the track
$\mathrm{SIP}_{3\mathrm{D}}$	signed 3D impact parameter of the track
$\mathrm{SIP}_{3\mathrm{D}}/\sigma_{3\mathrm{D}}$	signed 3D impact parameter significance of the track
$d_{3\mathrm{D}}$	jet track distance at their point of closest approach
$d_{3\mathrm{D}}/\sigma_{d_{3\mathrm{D}}}$	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]

Comparison of Higgs channels



(a) Particle multiplicities

(b) Distribution of track displacement parameter for leading charged particles