

ML-based Jet-Flavor Tagging at FCC-ee



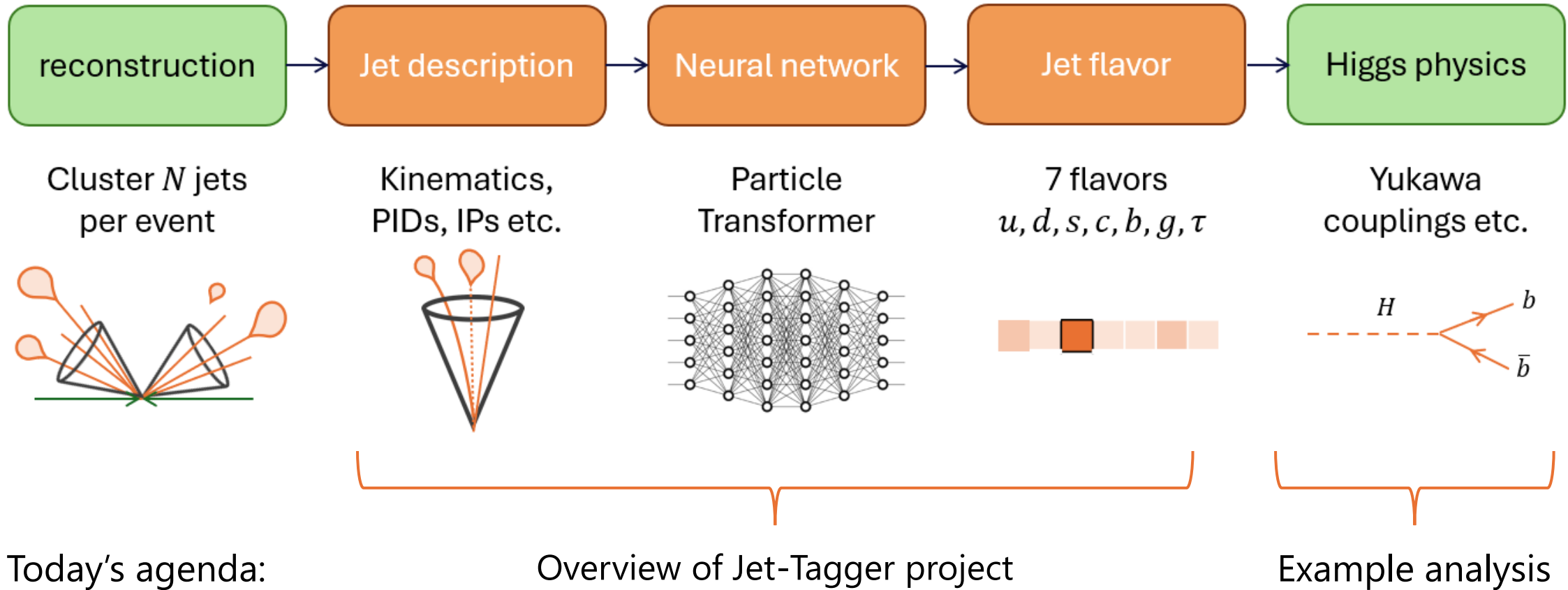
FUTURE
CIRCULAR
COLLIDER

FCC week

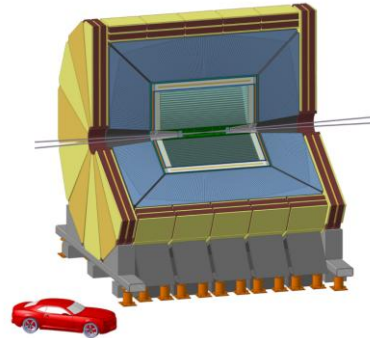
21. May 2025

Sara Aumiller, Michele Selvaggi, Dolores Garcia

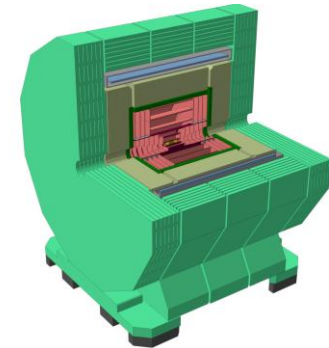
Jet-Flavor Tagging Set-Up



Jet-Flavor Tagging Status



IDEA



CLD

Fast simulation

Ready to use on-the-fly in FCCAnalyses since 2022

Performance tested in 2024

Full simulation

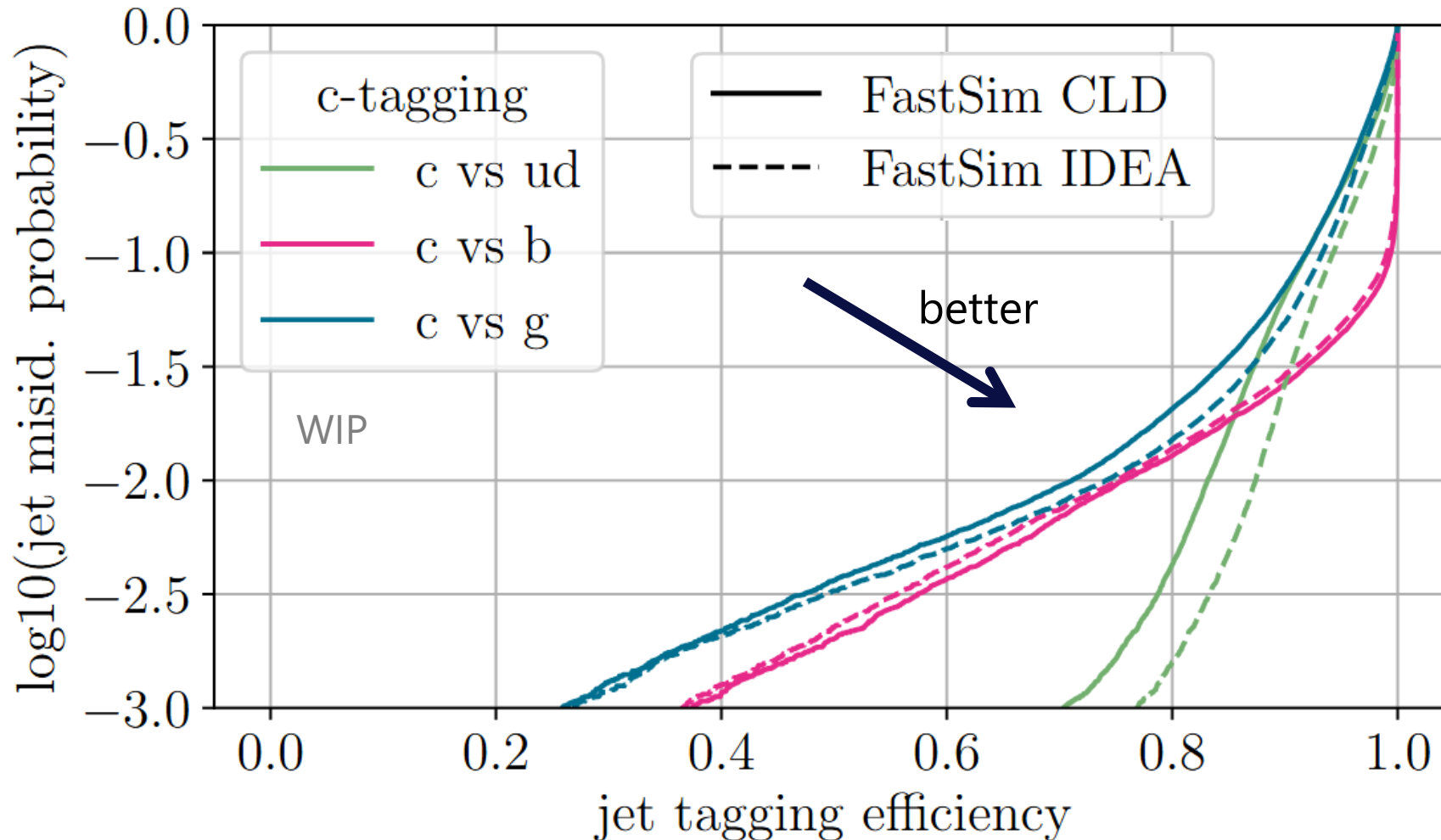
x

Available in key4hep (2025) and central production (WIP)



Further links: Tagging performance [note](#), FCCAnalyses fast sim [example](#), [GitHub](#) full sim implementation

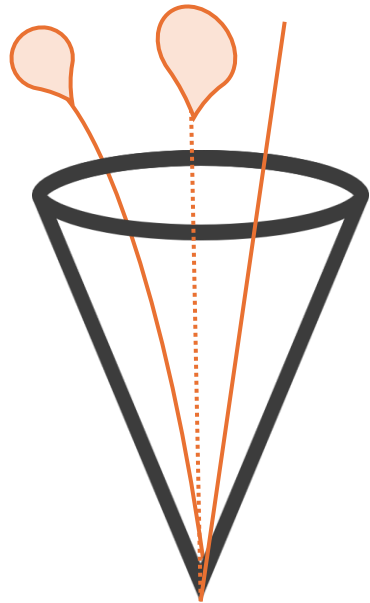
Jet Tagging in Fast Simulation



Good agreement in performance between CLD and IDEA

 *c vs. ud*: IDEA has PID!

Jet Description in Full Simulation



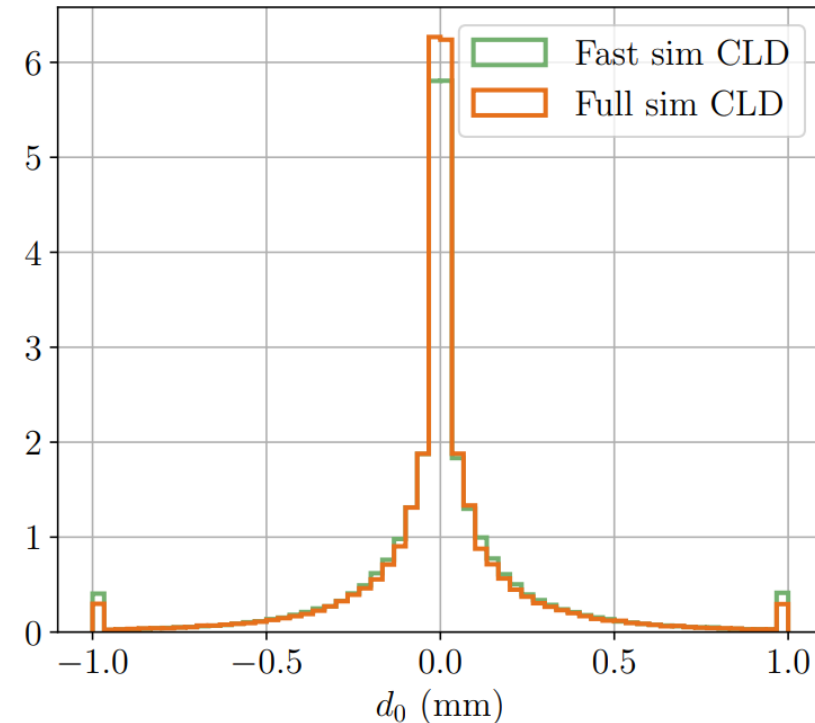
We need to **check the jet description** in **full simulation**.

Are there differences to fast simulation?

→ Comparison shows mostly good agreement:

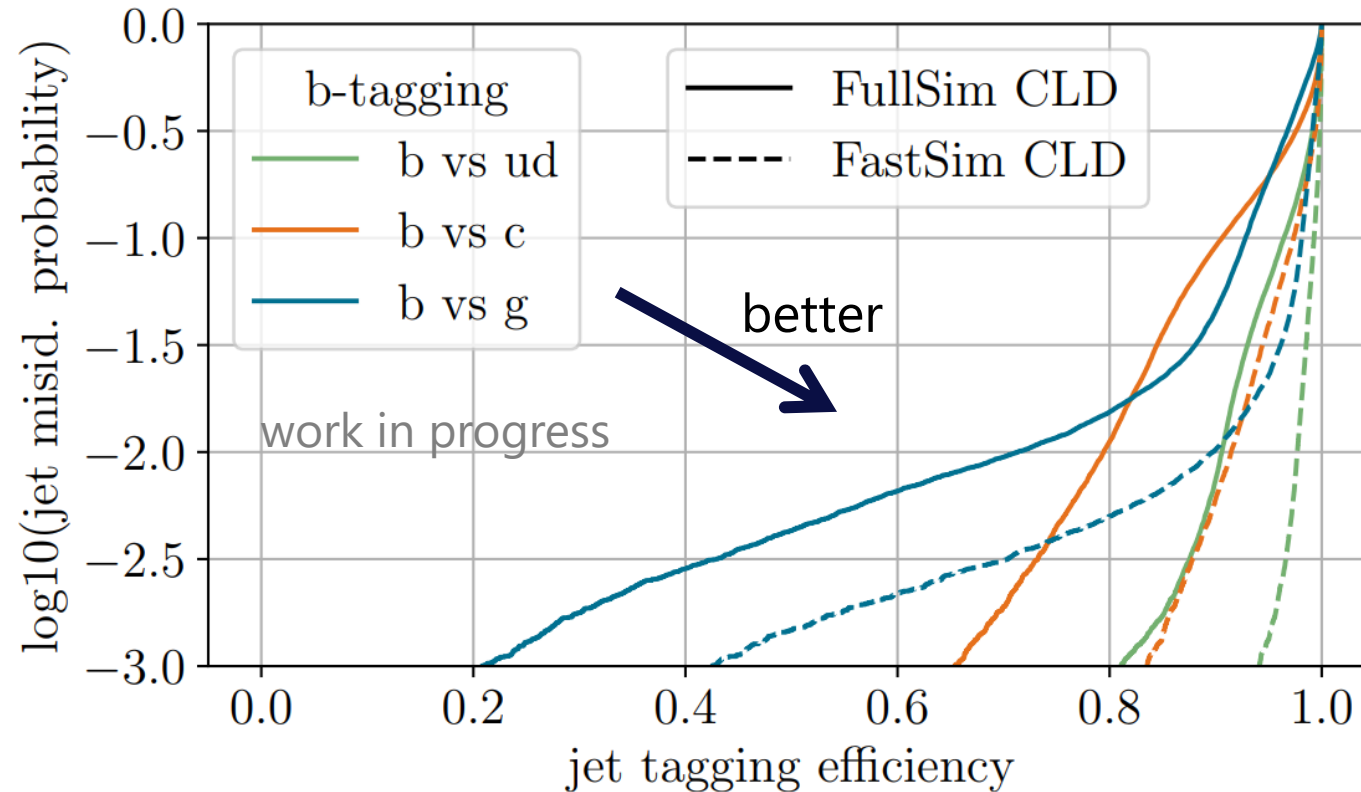
Key-problems identified in full simulation:

1. Unassociated tracks to PFOs
2. Fake neutrals
3. Suboptimal vertex reconstruction



Transverse impact parameter of leading tracks for $H \rightarrow b\bar{b}$

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)

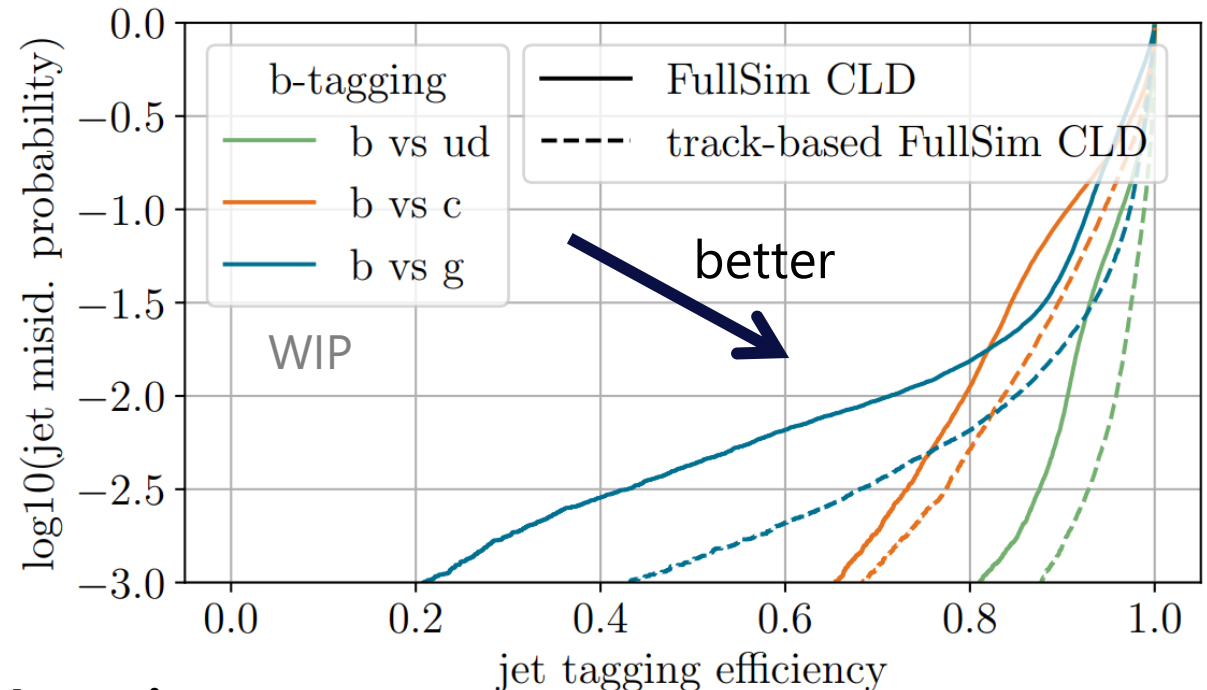
Improving Full Sim Tagging

- Improve **input data** to neural network → fix mentioned problems in full simulation!
- Sanity check: Does the performance of the tagger improve with manual fixes?

Manual fixes:

Instead of PFOs (particle flow objects) use


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



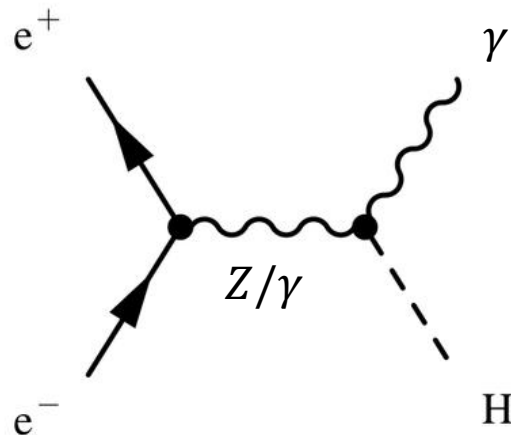
Large improvement:

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

Using Jet-Tagging in Full Sim

- **Key4hep implementation** of ML-based tagging: Gaudi functional [k4MLJetTagger](#)
 - Verified performance
 - Available in the software stack ("nightlies") since April 2025 
- WIP: jet-tags in central production
- WIP: example analysis with full simulation jet-tagger in FCCAnalyses

Effective $HZ\gamma$ coupling via $H\gamma$ production (see Lena Herrmann's [note](#))



Signatures:

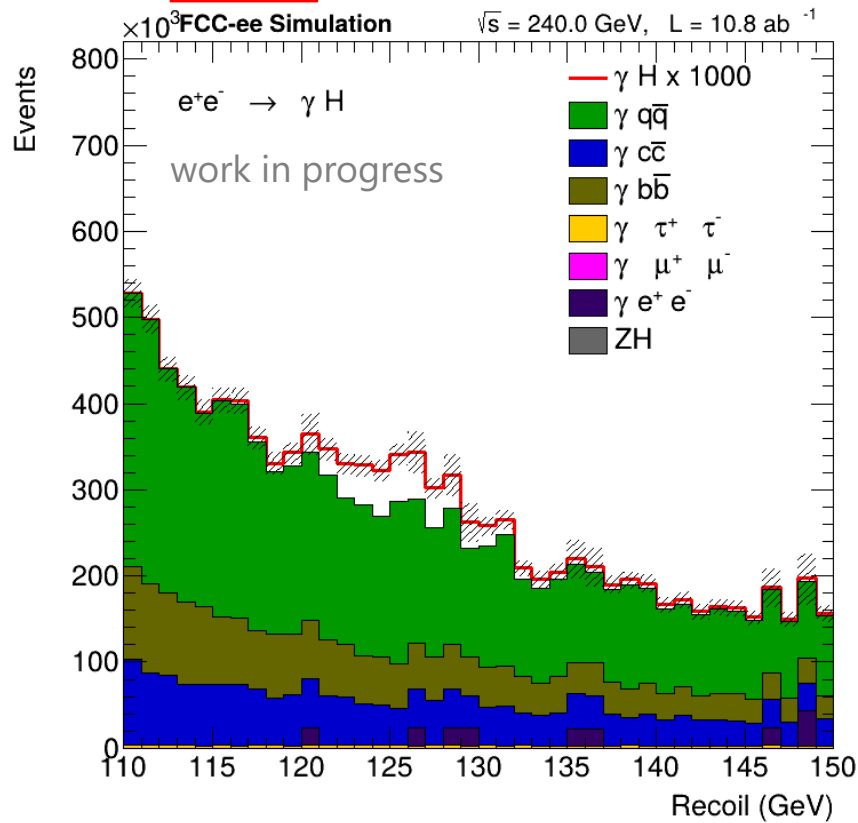
- Isolated photon
- Fixed photon momentum
- Recoil mass of the Higgs via the photon



Possible analysis handle:
tagging!

Effective $HZ\gamma$ coupling analysis

CLD (full simulation)

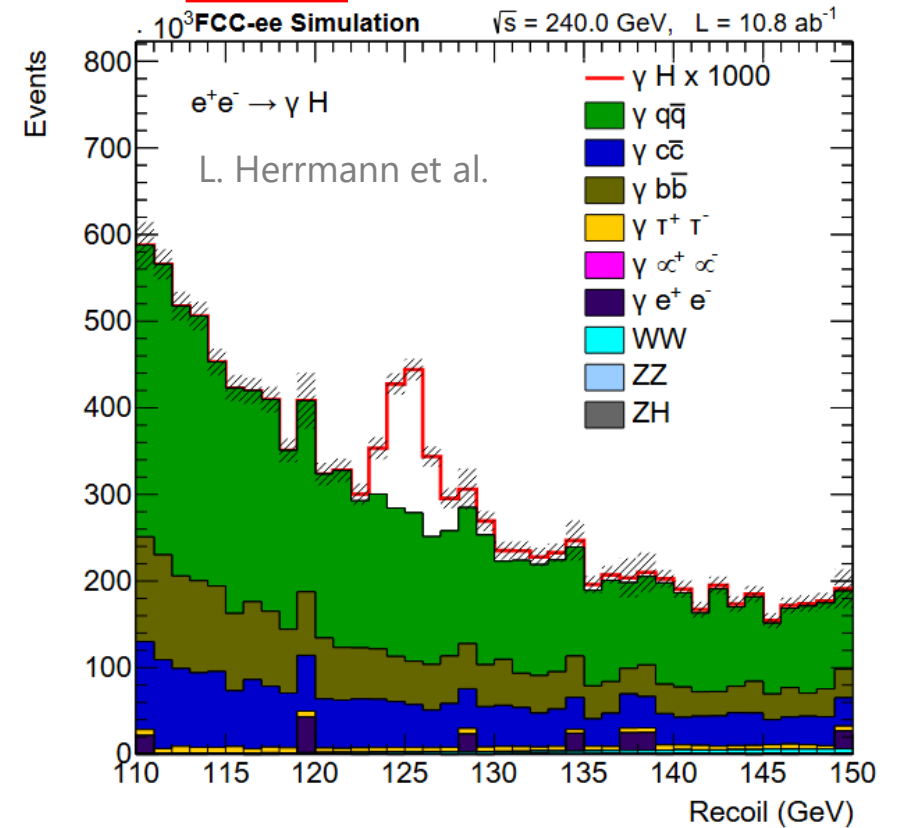


Precision (after recoil cut): 401%



CLD has ≈ 4 times worse ECAL resolution
 \rightarrow factor 2 worse precision!

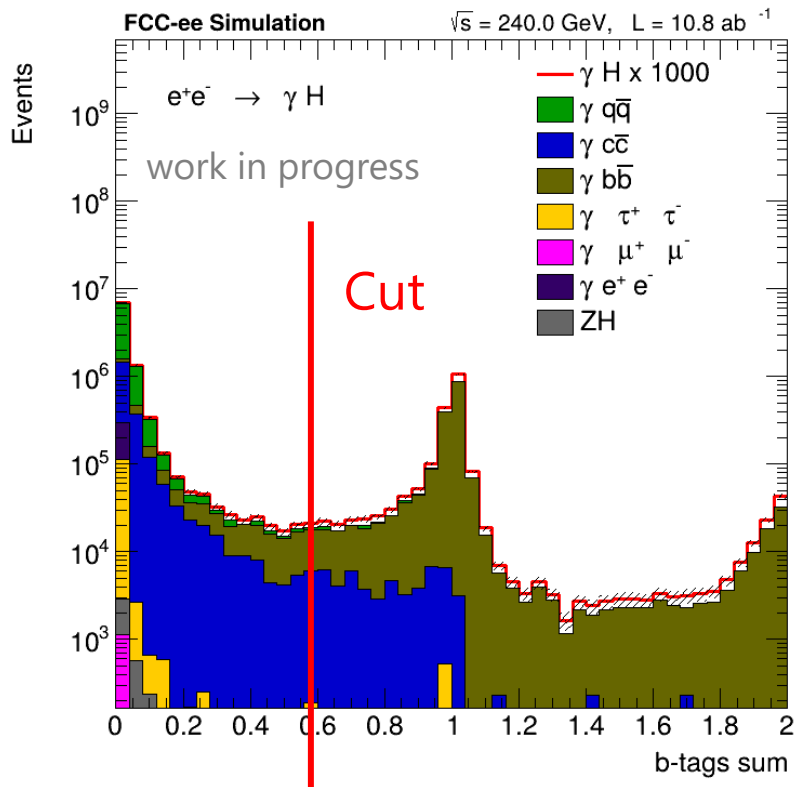
IDEA (fast simulation)



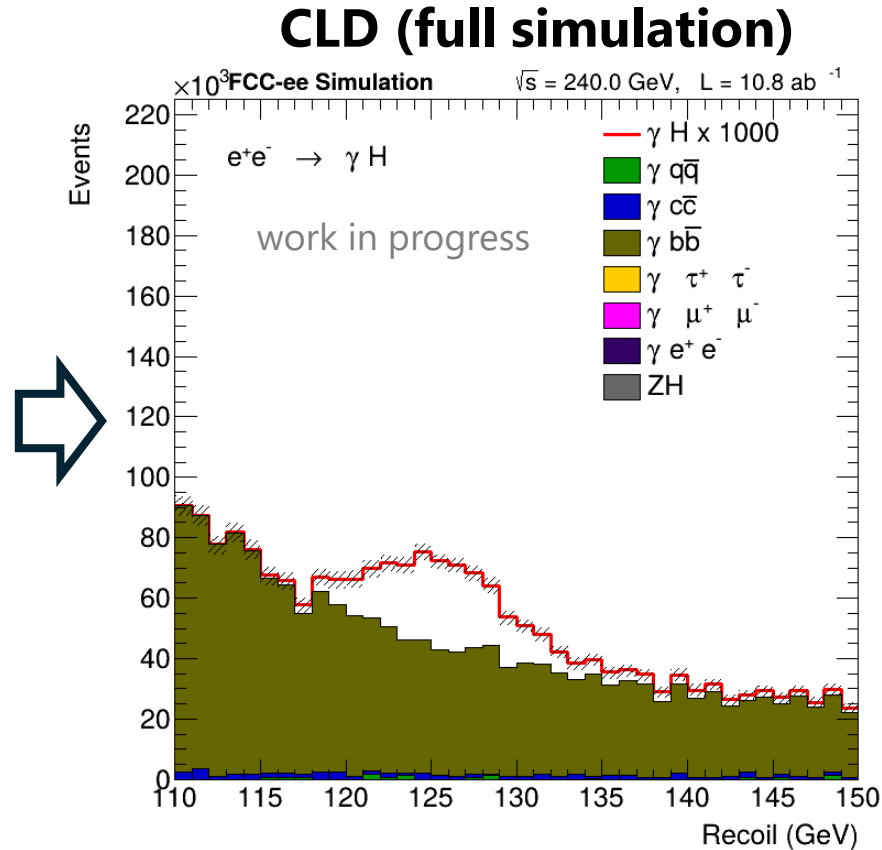
Precision (after recoil cut): 228%

Effective $HZ\gamma$ coupling analysis

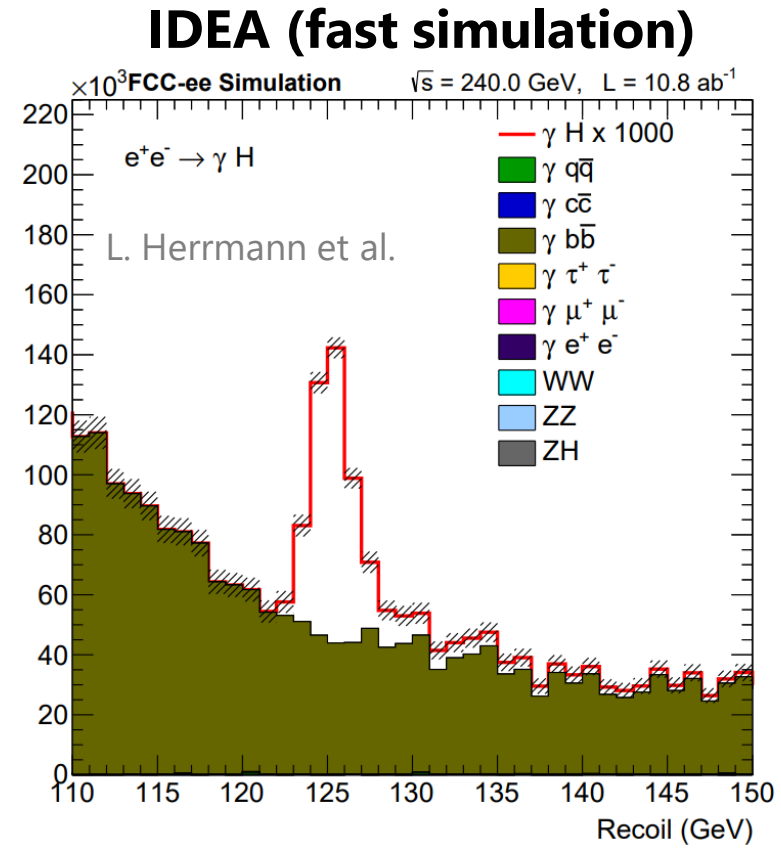
Apply jet-tagging as a handle for analysis: Select events with summed b-scores > 0.6



DISCLAIMER: This full sim analysis still contains (known) bugs! WIP!



Precision (after recoil cut): 299%

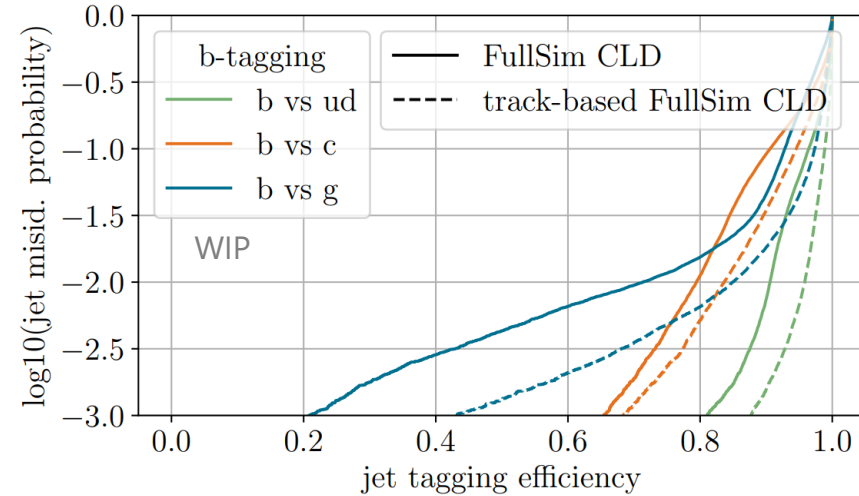


Precision (after recoil cut): 156%


DISCLAIMER: analysis cuts differ between the two plots!

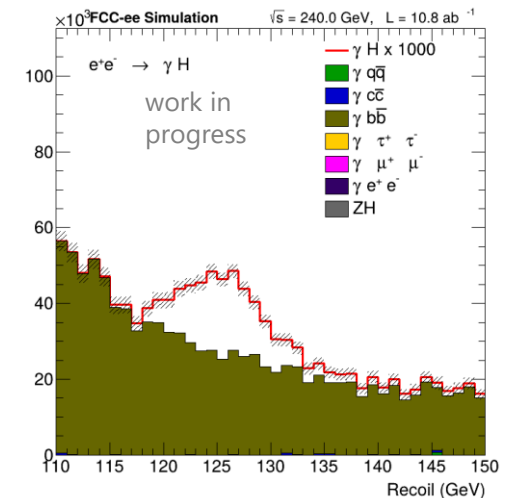
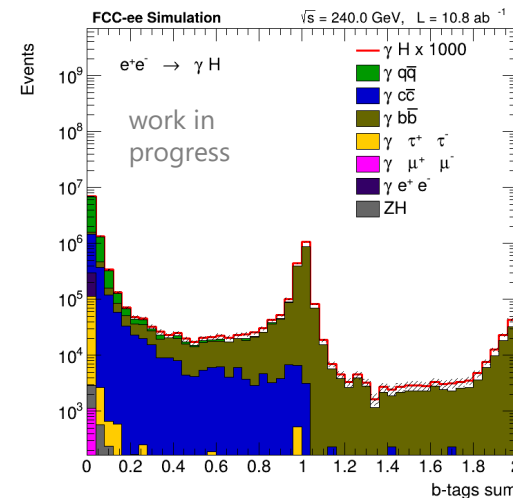
Summary – Jet Tagging at FCC-ee

- Jet-tagging available for
 - IDEA: fast simulation implementation via FCCAnalyses
 - CLD: **full simulation implementation via key4hep**
- Better performance if ParticleFlow, tracking and reconstruction improve
- Example **full simulation analysis with tagging**: effective $HZ\gamma$ coupling



For more information see [note on CDS!](#)

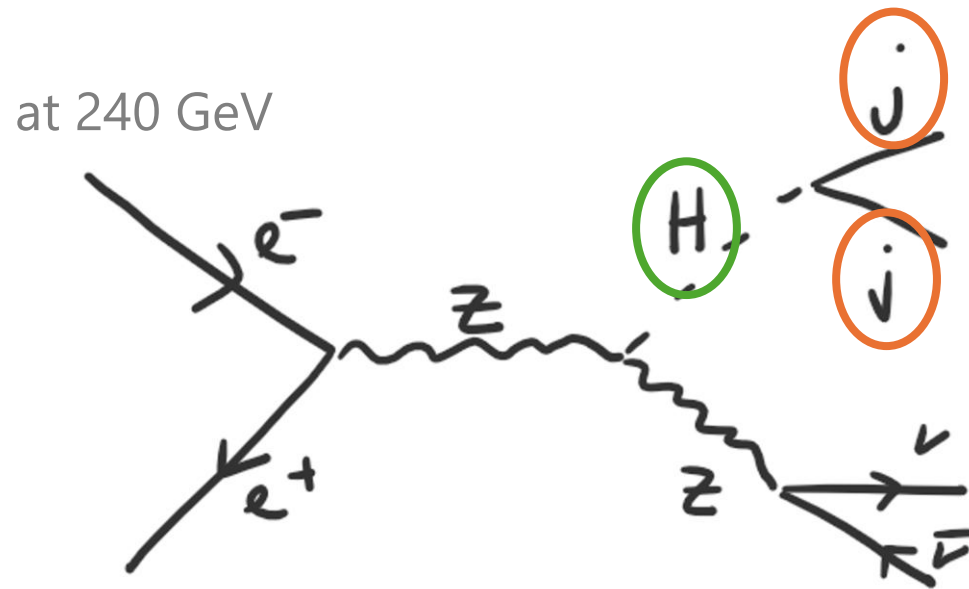
Or ask me questions over coffee ;) 



Backup



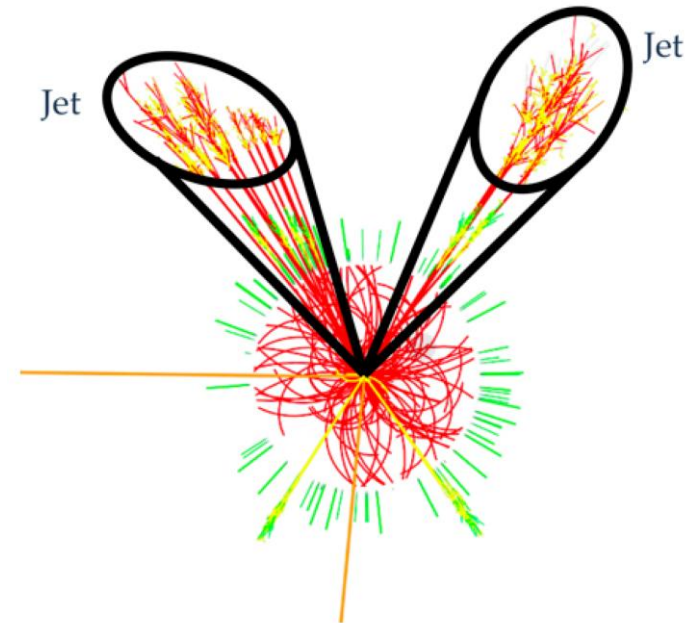
Why Jet-Flavor Tagging?



Future Colliders

=

Higgs factories for precious measurements



Particles causing jets:

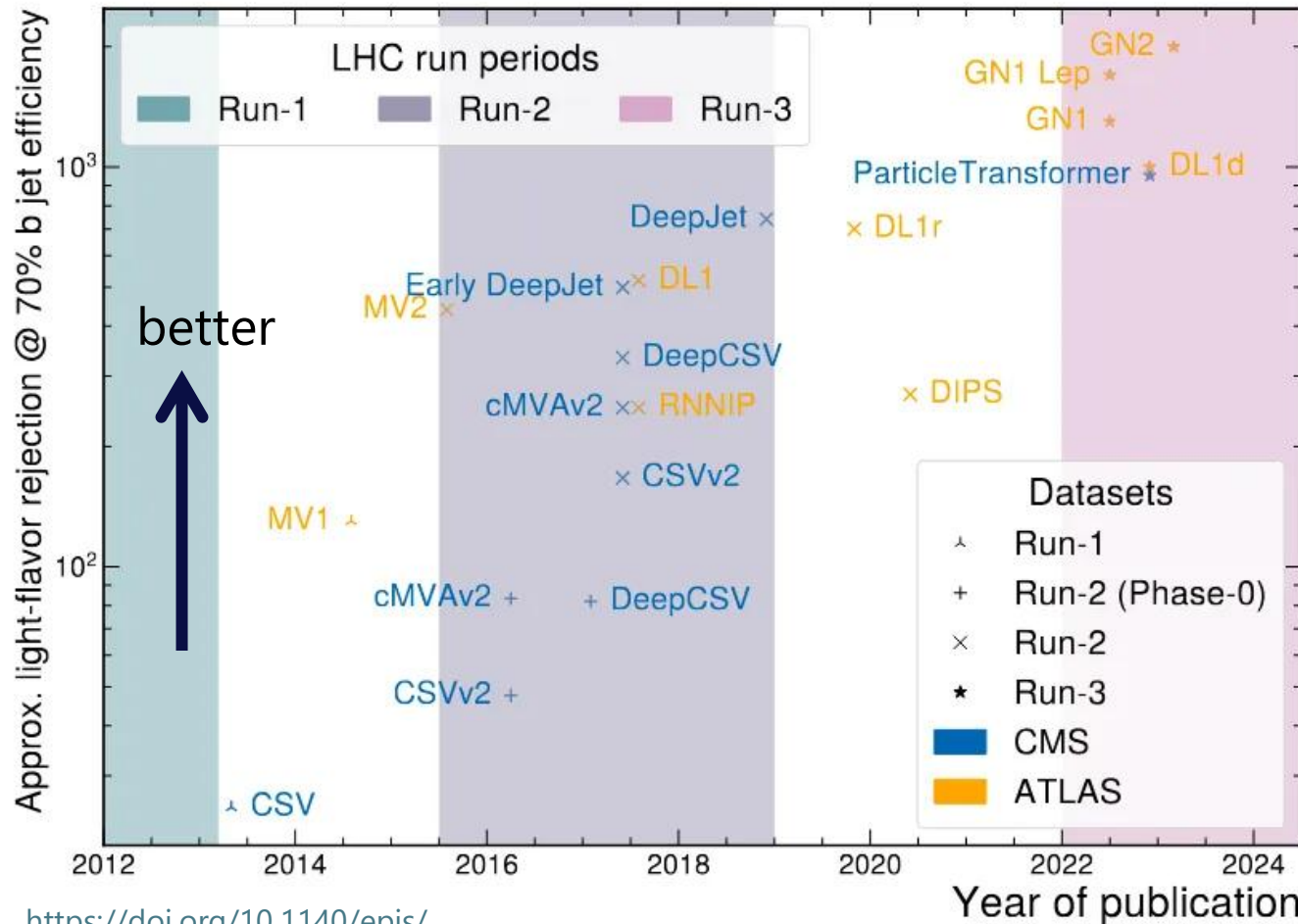
- quarks (u,d,s,c,b)
- gluons (g)
- leptons (τ)

Why use Machine Learning (ML)?



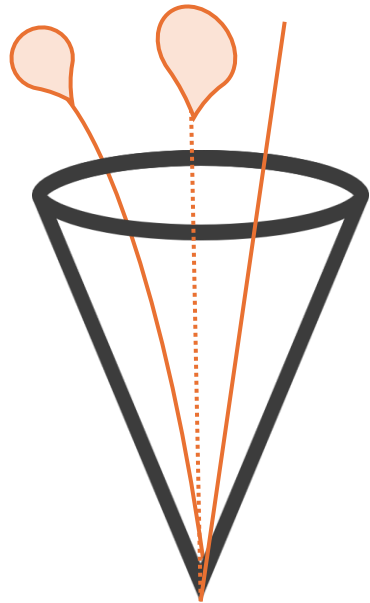
2045?

Performance of the tagger



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

Jet Description



We characterize the **jet constituents**:

Kinematics (3)	Identification (7)	Track displacements (23)
$\log E_{rel}, \theta_{rel}, \phi_{rel}$	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.)

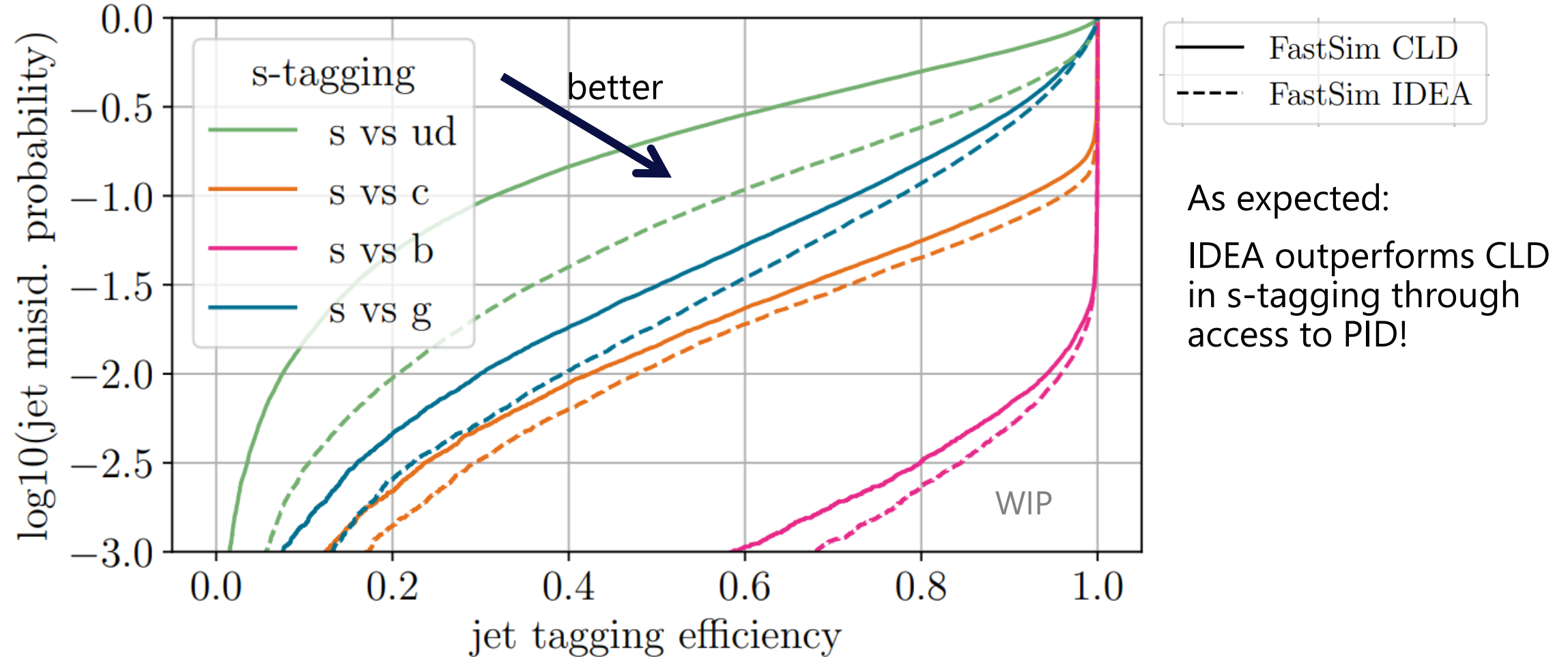
Input parameters to the network

Table 1. Set of input variables

Variable	Description
Kinematics	
$E_{\text{const}}/E_{\text{jet}}$	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
SIP _{2D}	signed 2D impact parameter of the track
SIP _{2D} / σ_{2D}	signed 2D impact parameter significance of the track
SIP _{3D}	signed 3D impact parameter of the track
SIP _{3D} / σ_{3D}	signed 3D impact parameter significance of the track
d_{3D}	jet track distance at their point of closest approach
$d_{3D}/\sigma_{d_{3D}}$	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
$m_{\text{t.o.f.}}$	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](#)

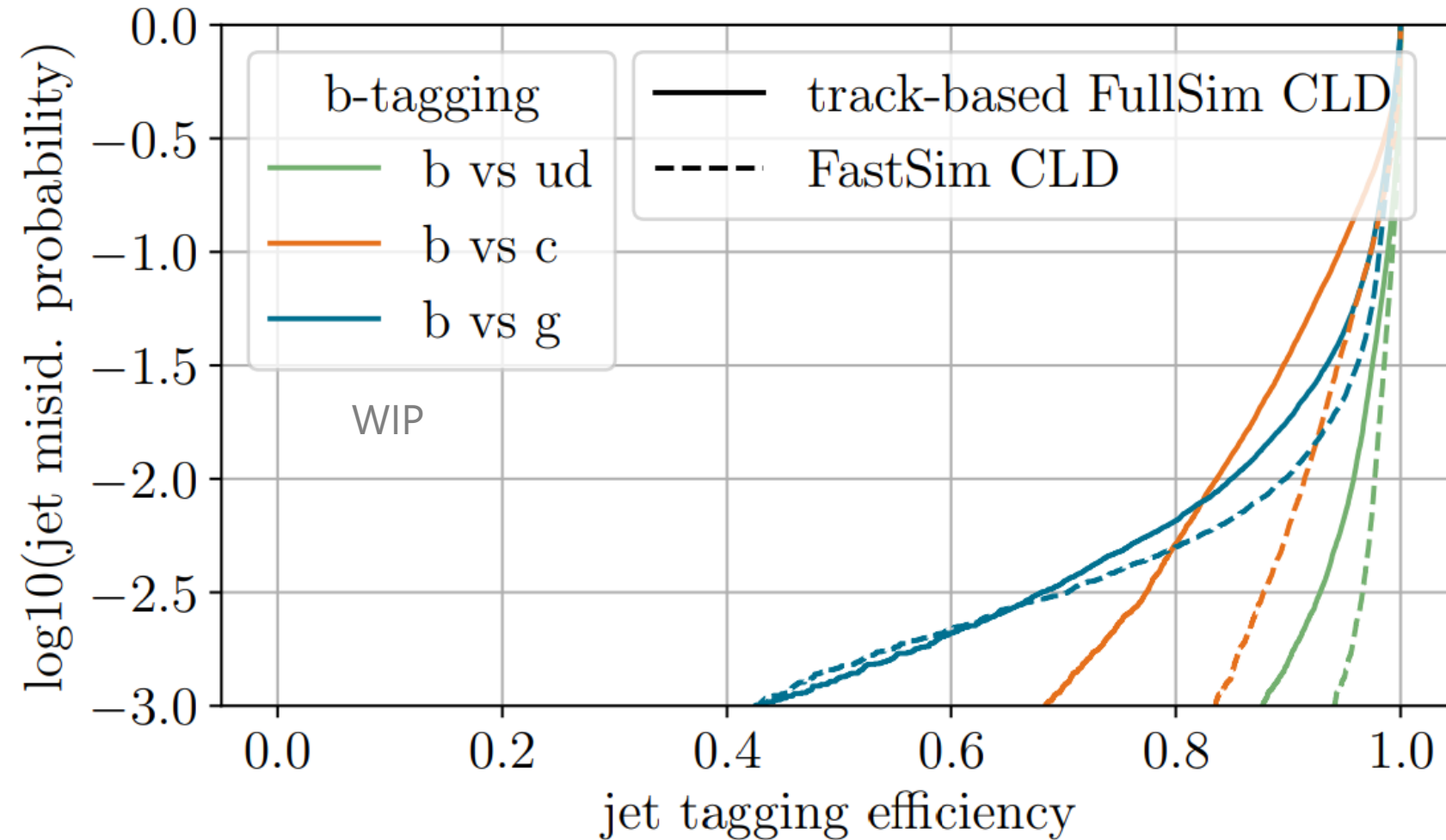
Jet Tagging in Fast Simulation



As expected:

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

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

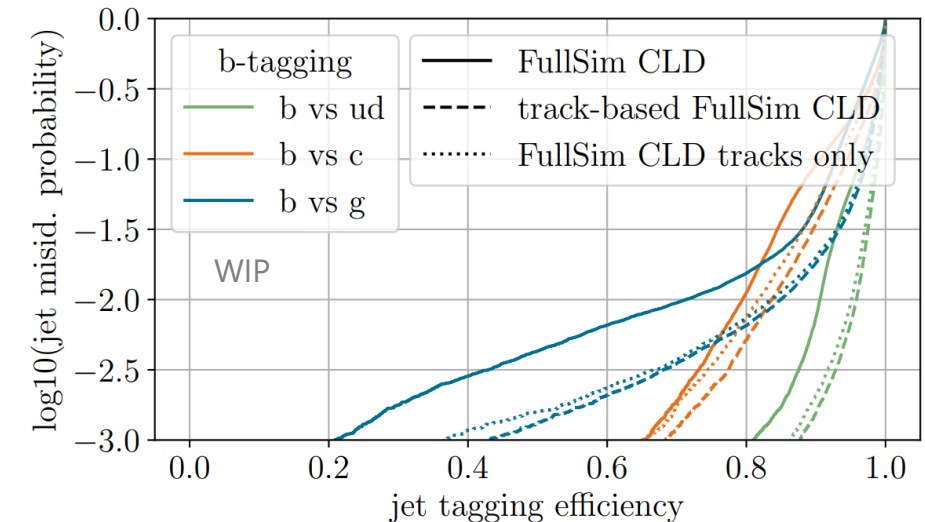
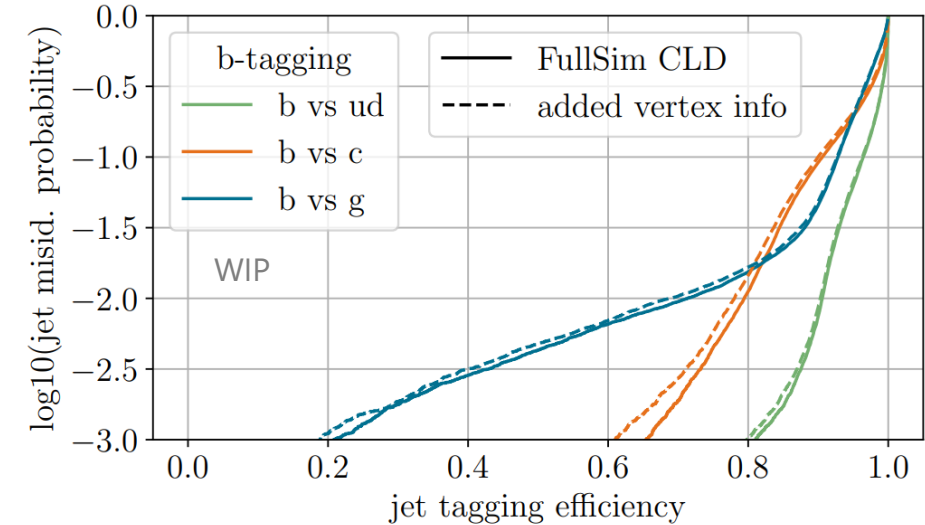
1. Adding Vertex Information

- 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](#)

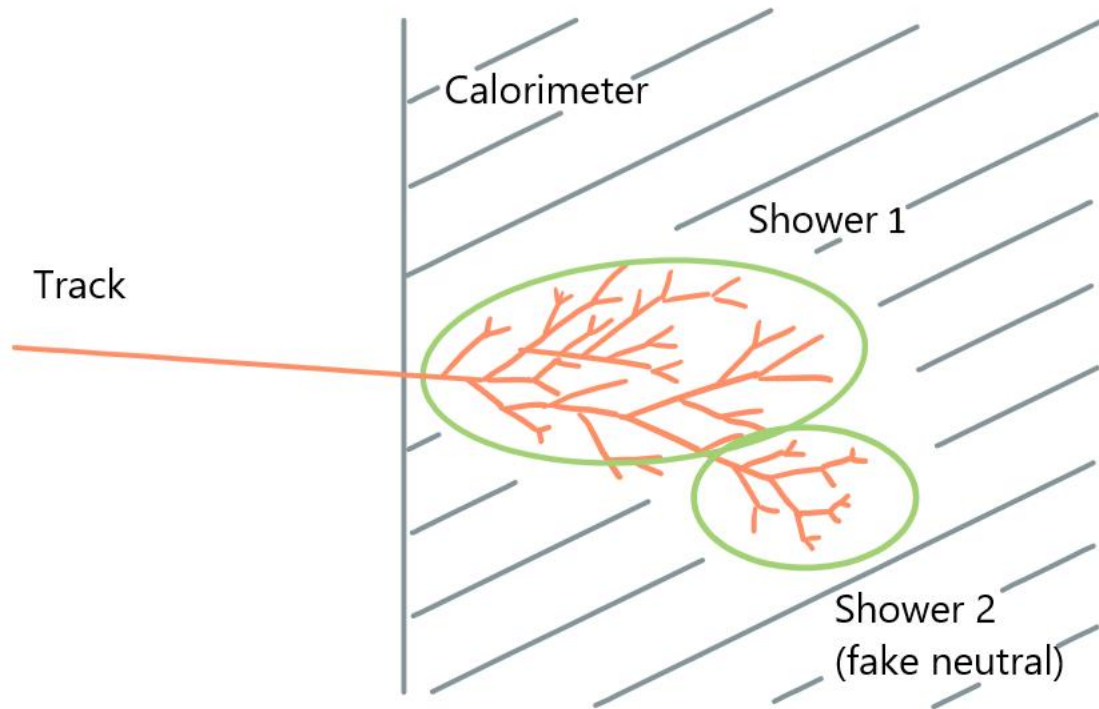


Backup

Full simulation issues

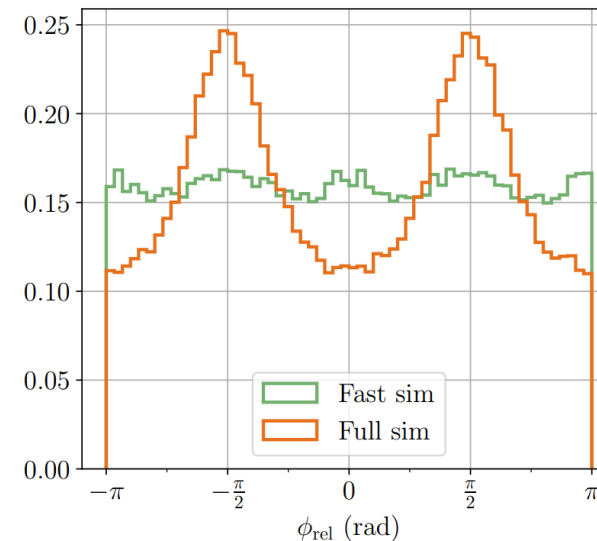


(1) Fake neutrals in full sim



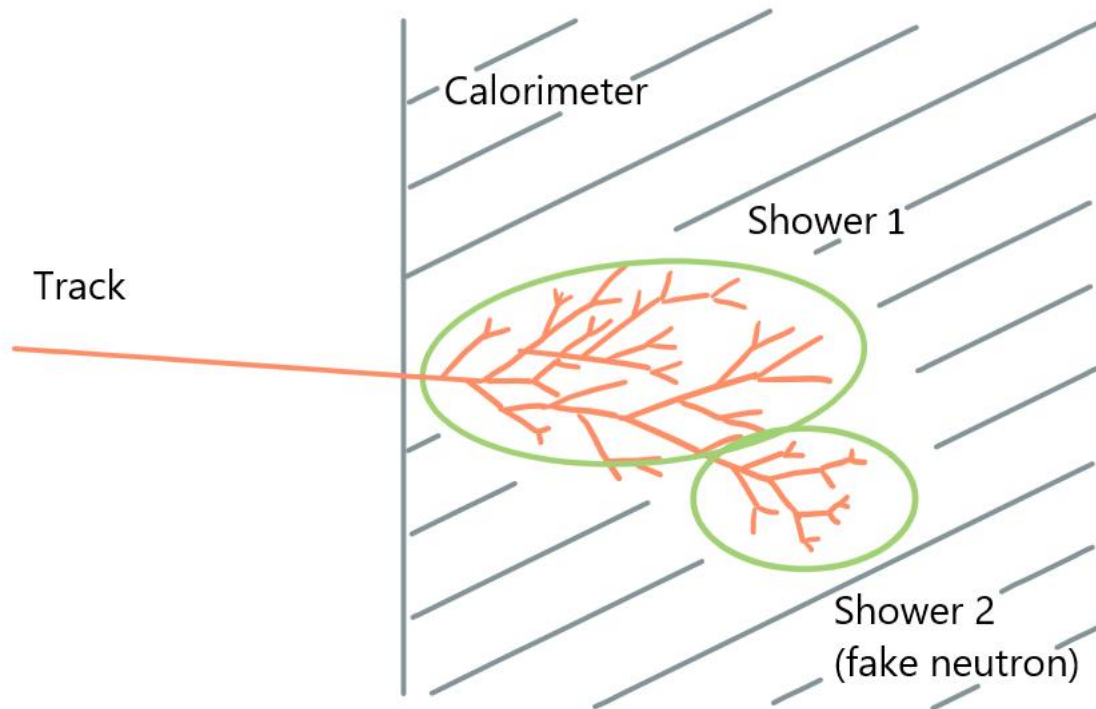
Artificially split cluster of high-energy charged particles (at MC level) creates **fake neutral**.

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

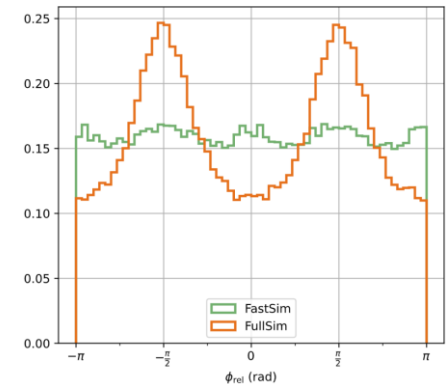


leading neutral hadronic jet constituents

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 \rightarrow peaks in distribution



(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.

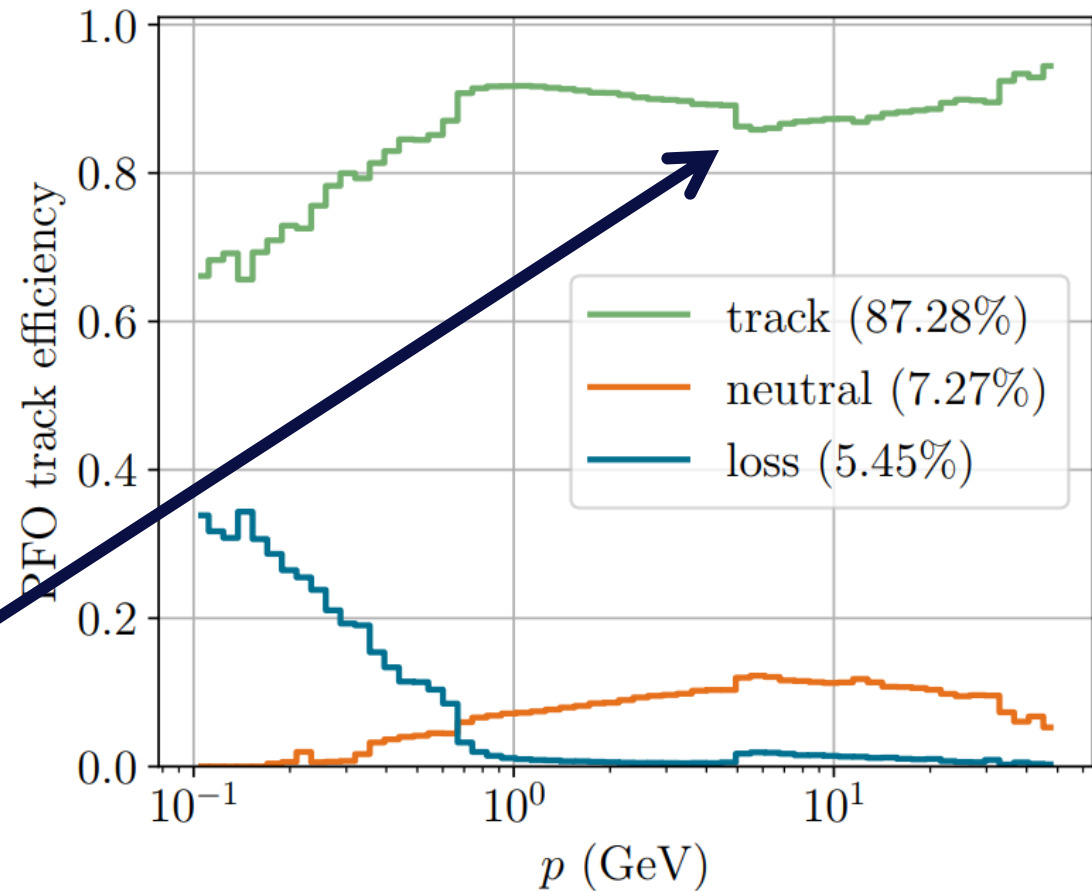
→ **track-cluster association fails**

→ problematic as tracks are crucial for jet flavor tagging

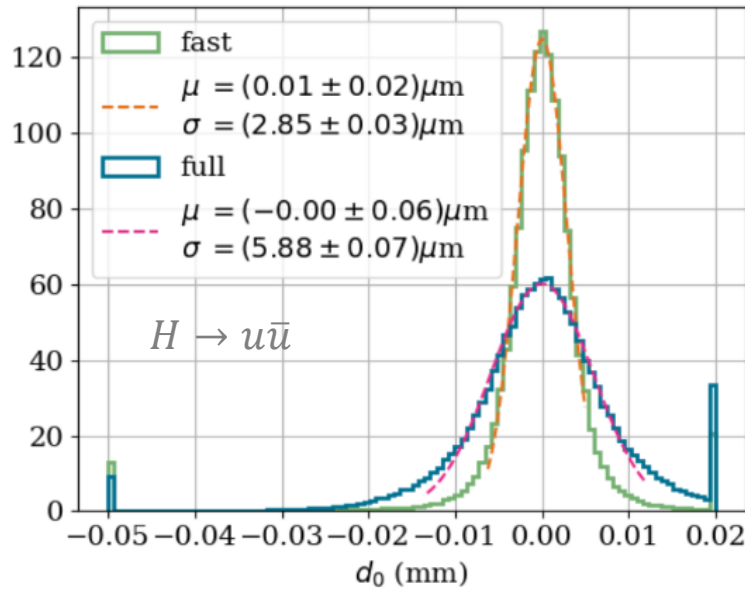
Reconstruction constraint (from pandora): above 5 GeV charged particles must have cluster associated

→ reconstruction could be improved

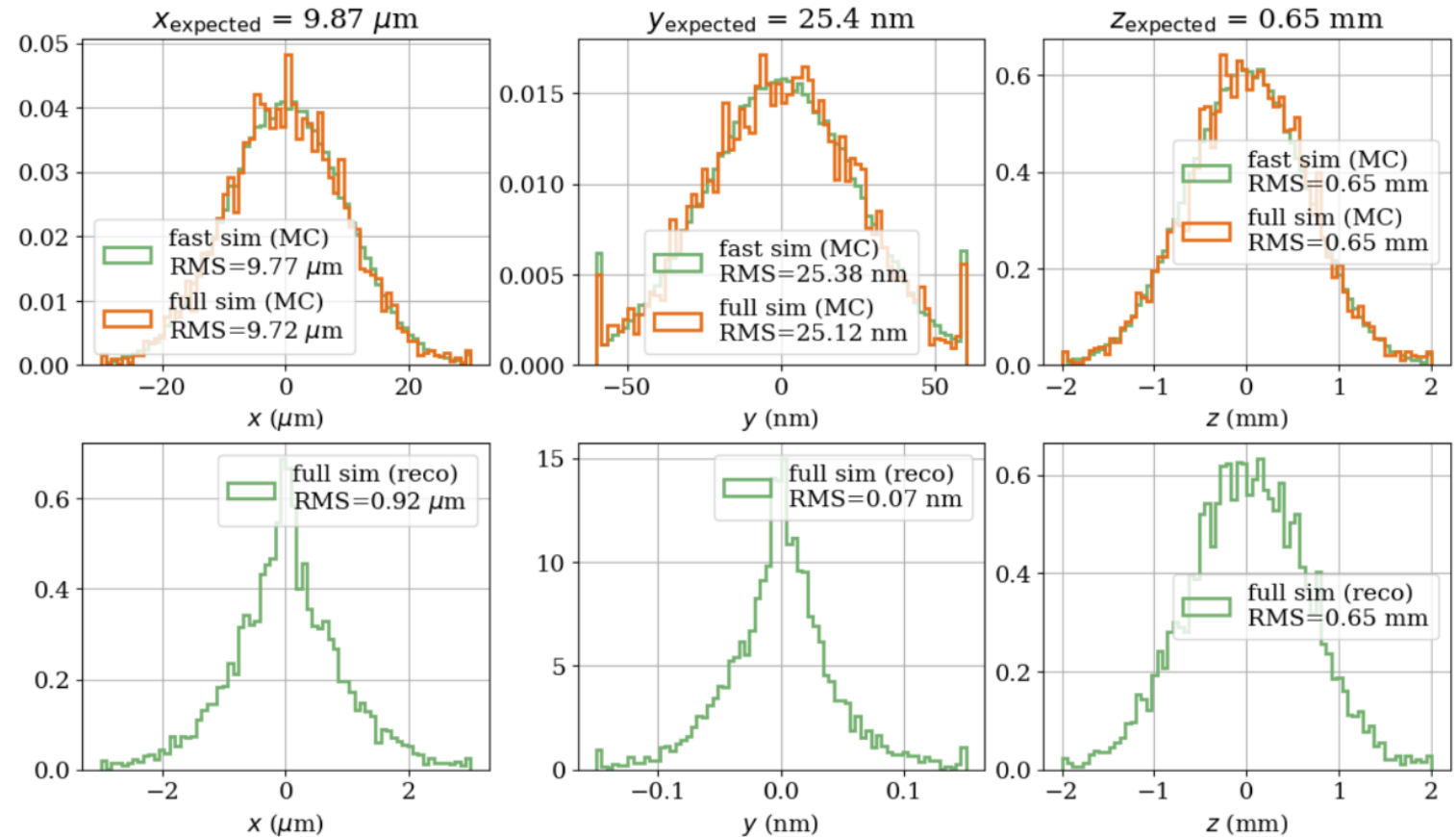
MC charged hadrons ($H \rightarrow b\bar{b}$)



(3) Vertex reconstruction



Discrepancy in d_0 led to the primary vertex. The distributions in reco full sim are too narrow. Most likely source: beam spot constraint in PV fit wrongly implemented.



For further information, see [github issue](#).

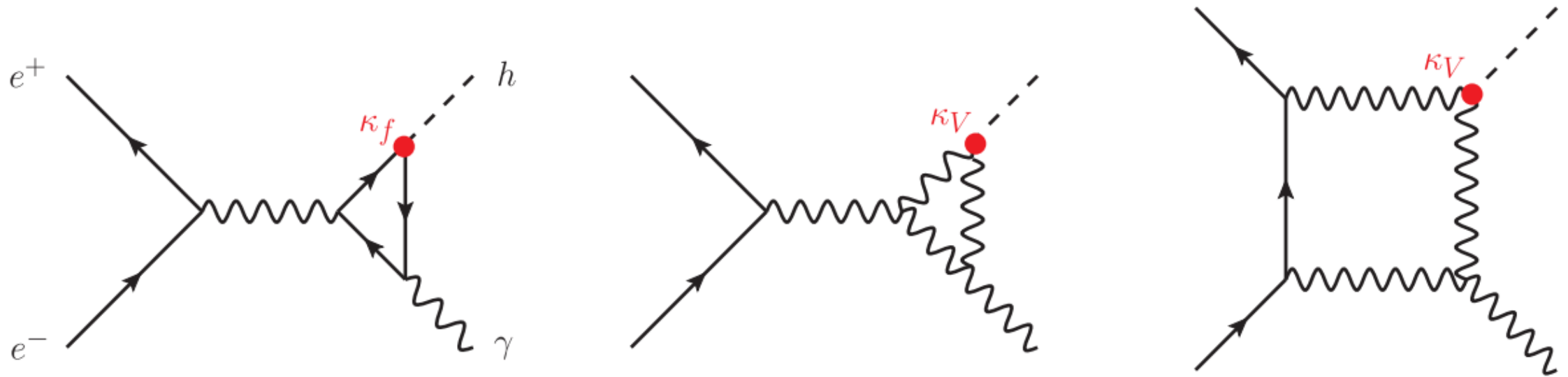
Backup

HγZ effective coupling full simulation analysis



FUTURE
CIRCULAR
COLLIDER

Feynman diagrams



<https://journals.aps.org/prd/abstract/10.1103/PhysRevD.99.035023>

WIP: Effective $HZ\gamma$ coupling analysis

Selection Cut	γH	γqq	γcc	γbb	$\gamma \tau\tau$	$\gamma \mu\mu$	γee	ZH prec (%)
All events	886	74520000	23.2M	25.4M	8.3M	8.6M	2052M	2.2M 5286.3%
iso(γ) < 0.2	883 (99.7)%	74.5M (99.9)%	23.2M (99.9)%	25.4M (100.0)%	7.4M (88.8)%	6.6M (76.4)%	918.1M (44.7)%	2.1M (98.5)% 3680.5%
60 < p_γ < 100 GeV	780 (88.1)%	26.4M (35.4)%	8.0M (34.3)%	9.1M (35.7)%	1.9M (22.4)%	1.8M (21.1)%	115.7M (5.6)%	8095 (0.4)% 1635.5%
$ \cos(\theta)_\gamma < 0.9$	678 (76.5)%	13.5M (18.1)%	4.1M (17.8)%	4.6M (18.1)%	1.0M (11.7)%	1.0M (11.0)%	16.7M (0.8)%	7222 (0.3)% 943.0%
# particles > 9	627 (70.8)%	13.5M (18.1)%	4.1M (17.8)%	4.6M (18.1)%	0.2M (2.7)%	2074 (0.0)%	0.5M (0.0)%	5302 (0.2)% 763.7%
120 < m_{recoil} < 132 GeV	454 (51.3)%	2.0M (2.7)%	0.6M (2.5)%	0.6M (2.6)%	37.3k (0.4)%	345 (0.0)%	82.0k (0.0)%	720 (0.0)% 401.4%
b score sum > 1	175 (19.8)%	0 (0.0)%	928 (0.0)%	322.5k (1.3)%	0 (0.0)%	0 (0.0)%	0 (0.0)%	21 (0.0)% 324.1%

- Obvious cuts: isolated photon with certain momentum
- Cuts against background: angle of the photon & number of particles
- Cut on recoil mass: 401% precision on the measurement (cross-check: with IDEA fast sim 228%. Explanation: CLD has ≈ 4 times worse photon resolution \rightarrow factor 2 in precision)
- Cut on b-tags: improvement of the analysis to 324%