Jet flavor identification for FCCee

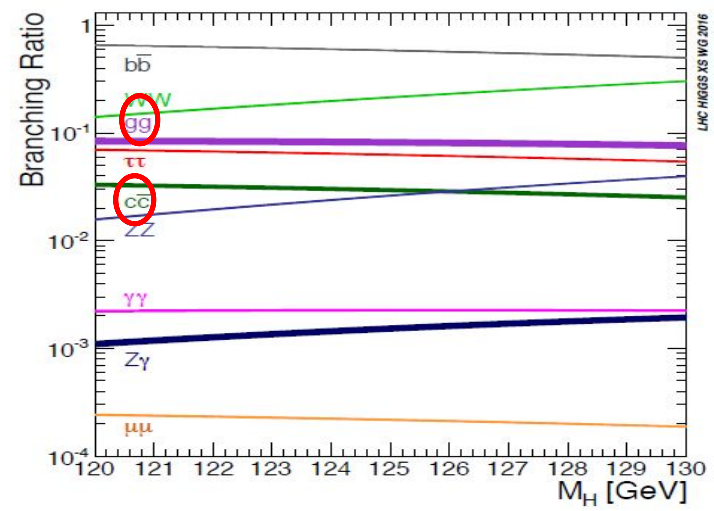
Michele Selvaggi
(CERN)

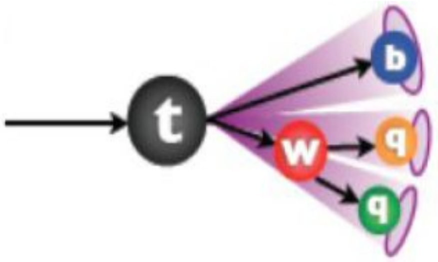
Andrea Del Vecchio (Rome), Dolores Garcia (CERN), Laurent Forthomme (CERN), Franco Bedeschi (INFN), Loukas Gouskos (Brown), Sara Auchmiller (CERN), Eli Marlin (BU)

[arXiv:2202.03285, EPJ C 82 646 (2022) link]

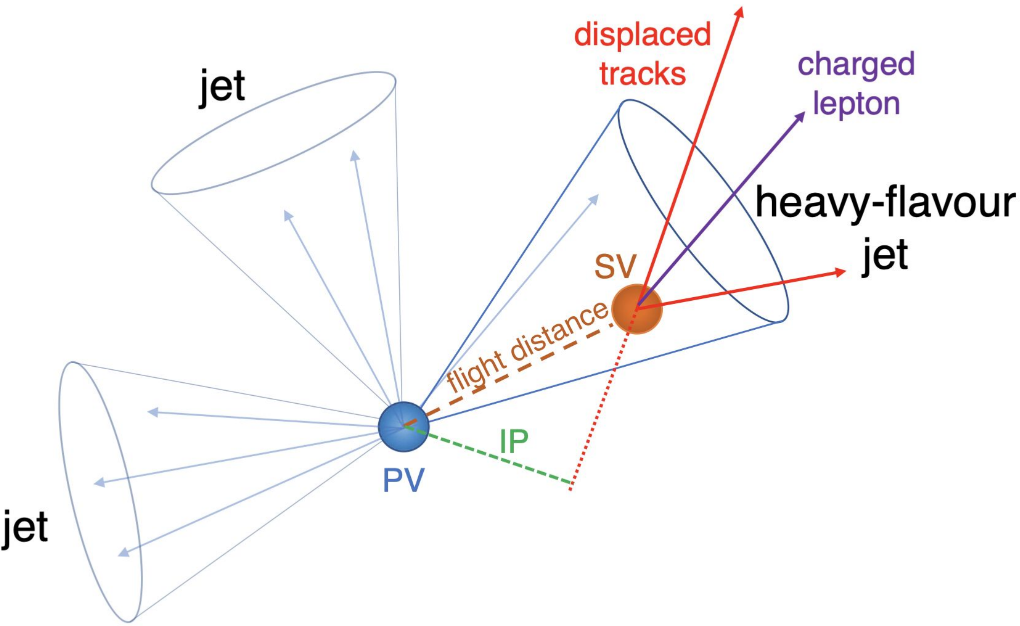
Physics motivation

- Flavour tagging essential for the e^+e^- program, e.g.:
  - Higgs Sector:
    - (HL-)LHC can access 3rd gen. couplings and a few of 2nd generation
    - Future e^+e^-: Measure Higgs particle properties and interactions in challenging decay modes
      - E.g. cc, 1st gen quarks/fermions, gg [?]
  - Top quark physics [if E_{CM} sufficient]
    - Precise determination of top properties [mass, width, Yukawa]
  - QCD Physics
    - strong coupling (a_S), event shapes ..
    - modelling of hadronization, MC tuning, …
  - ….





Basics of flavour tagging (b/c)



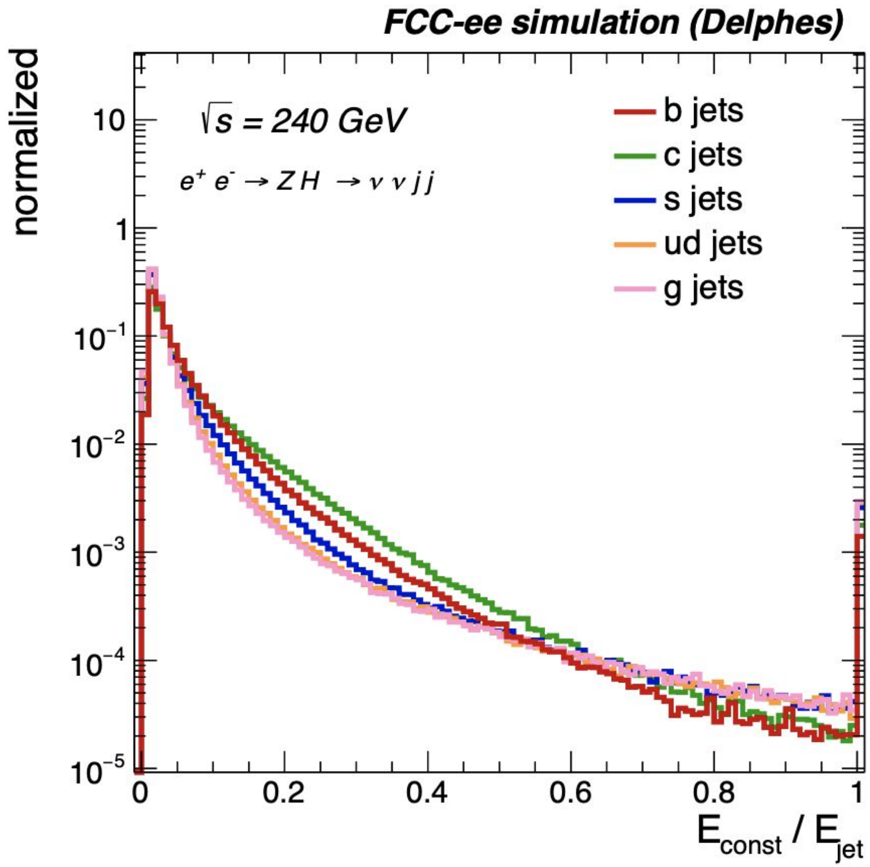
- Large lifetime
  - b (c) lifetime ~1 ps (~0.1ps)
  - b **(c)** decay length: ~500 μm
  - (~5) mm for ~50 GeV boost
- Displaced vertices/tracks
  - Large impact parameters
  - Tertiary vertices when B hadron decays to C hadron
- Large track multiplicity
  - ~5 (~2) charged tracks/decay
- Presence of non-isolated e/μ
  - ~20 (10)% in B (C) decays

<u>**Detector constraints:**</u>
Need power pixel/tracking detectors
- Good spatial resolution
- As little material as possible
- Precise track alignment

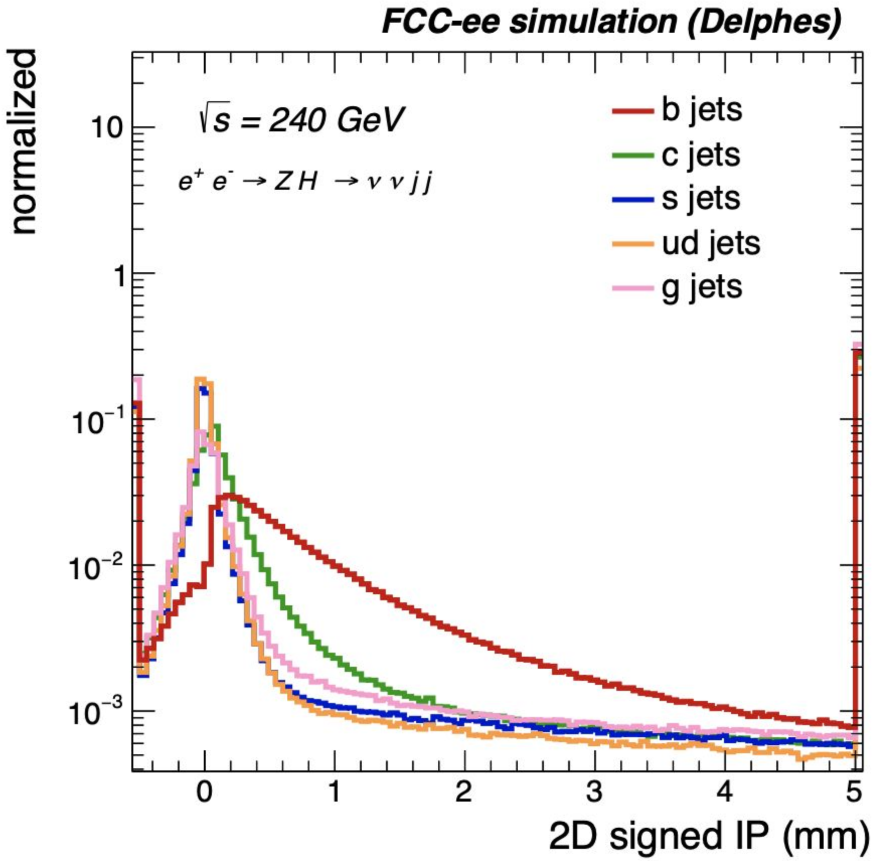
Input variables

- Comparison of input distributions for different jet flavors

Constituent relative energy



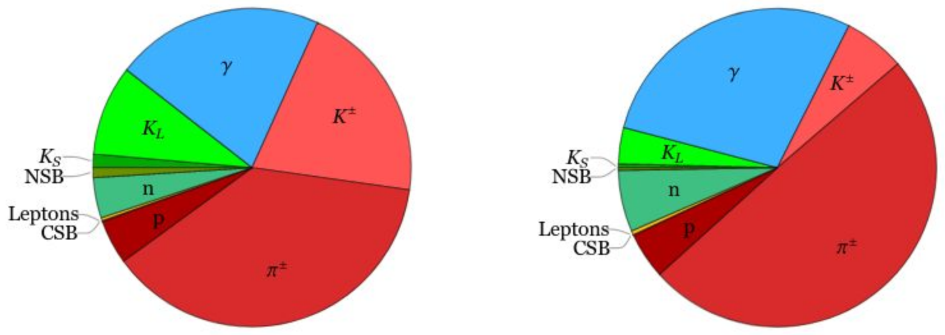
Impact parameter (d_0)



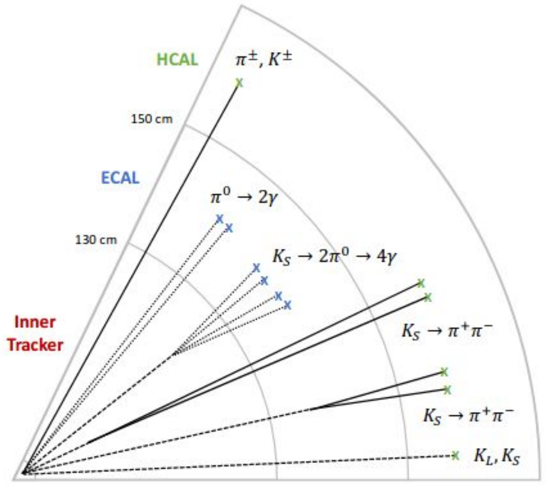
Basics of flavour tagging (strange)

[2003.09517]

Momentum weighted fraction:



Strange $p_T = 45\,\text{GeV}$ Down $p_T = 45\,\text{GeV}$



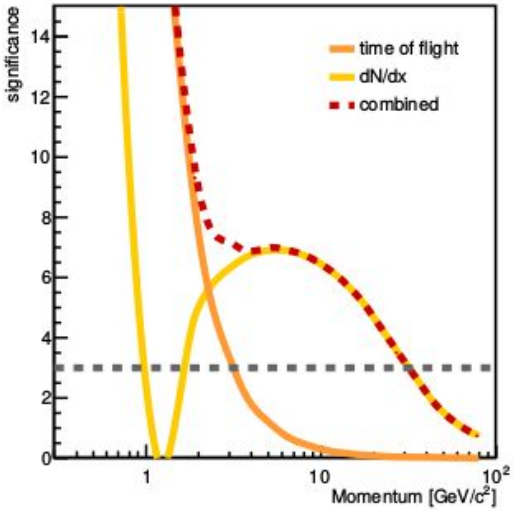
- Large Kaon content
  - Charged Kaon as track:
    - K/pi separation
      - TOF
      - dEdx/dNdx
  - Neutral Kaons:
    - $K_S \to \pi\pi$
      - Displaced 2 track vertex
      - 4 photons
    - K_L
      - TOF vs n ?

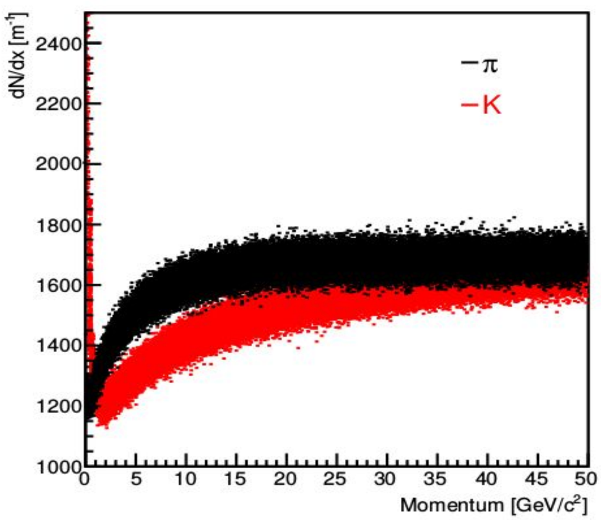
Detector constraints:

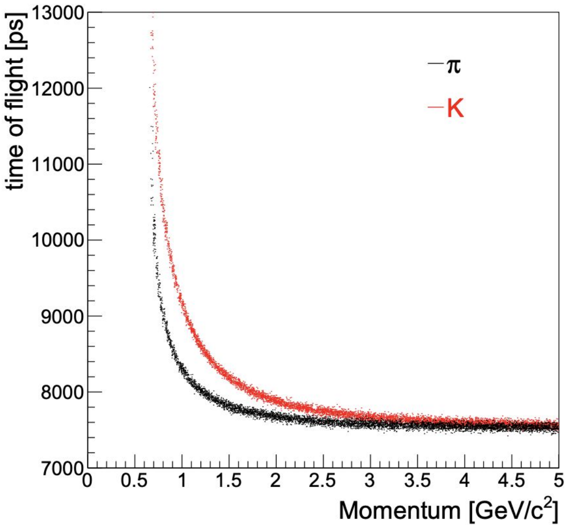
- timing detectors
- charged energy loss (gas/silicon)
- cherenkov detectors

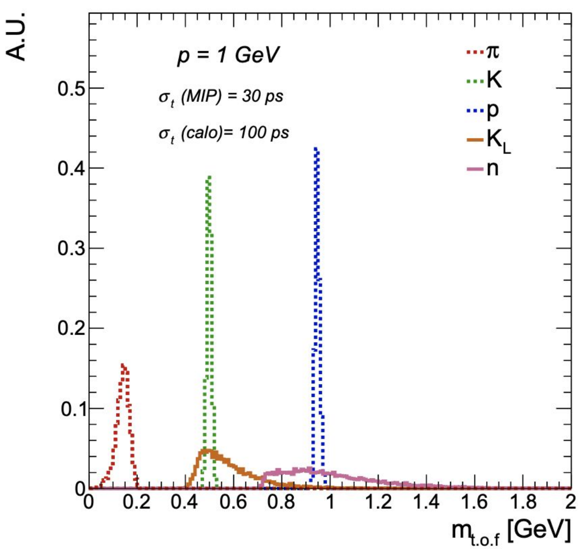
Particle ID: dN/dx and ToF

- Count number of **primary ionization** clusters along track path
- ToF results in good K/π separation at low-momenta
- Modules added in Delphes









Designing a Graph-based tagger

- **Jet representation:** critical for powerful jet tagging algorithms
  - **In theory:** A spray of particles produced by the hadronization of q and g
  - **Experimentally:** A cone of reconstructed particles in the detector

- Reminder: Current and future experiments have / will have a **PF-based** event reconstruction
  - **Output:** mutually exclusive list of particles
    - Rich set of info/particle
      - Energy/momentum, position
      - Displacement, particle type
      - timing
      - …

[O(50) properties/particle]
x [~50-100 particles/jet]
~O(1000) inputs/jet

- **Until recently:** Jet taggers based on human-inspired higher-level observables
  - Inputs to cut-based or simple ML-based algorithms
- Move to **particle-based jet tagging**: i.e. exploit directly the full list of jet constituents (ReconstructedParticles) and **new advances in ML**

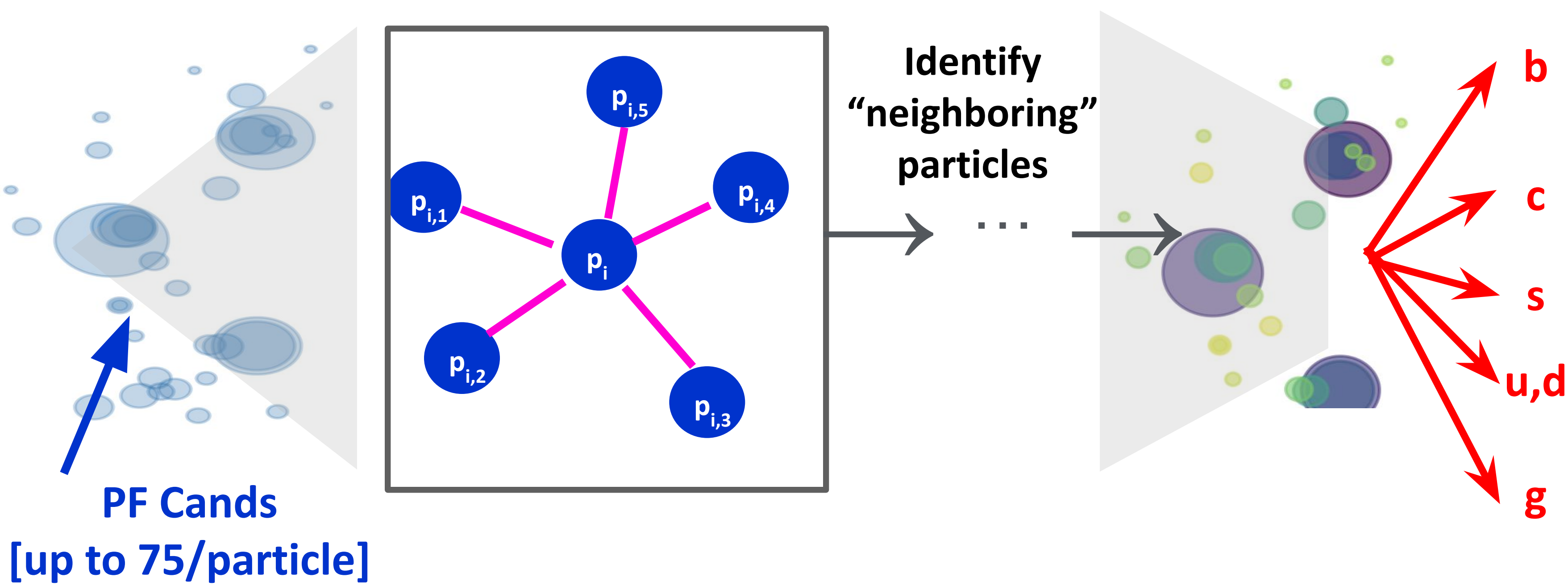
Full list of input variables

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

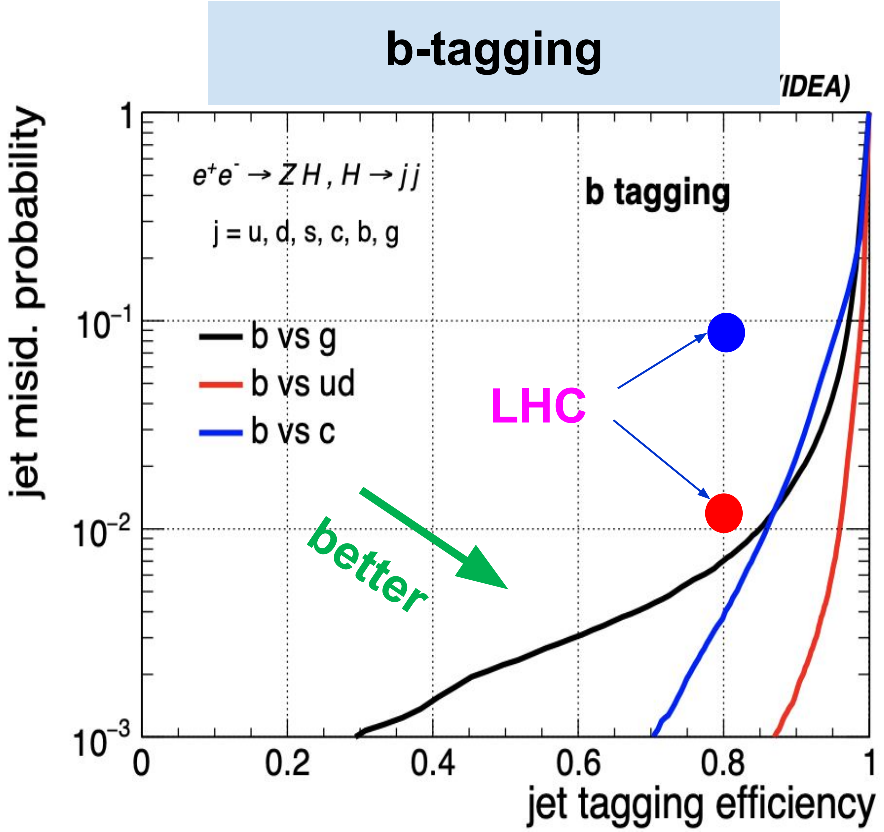
ParticleNet(-ee)

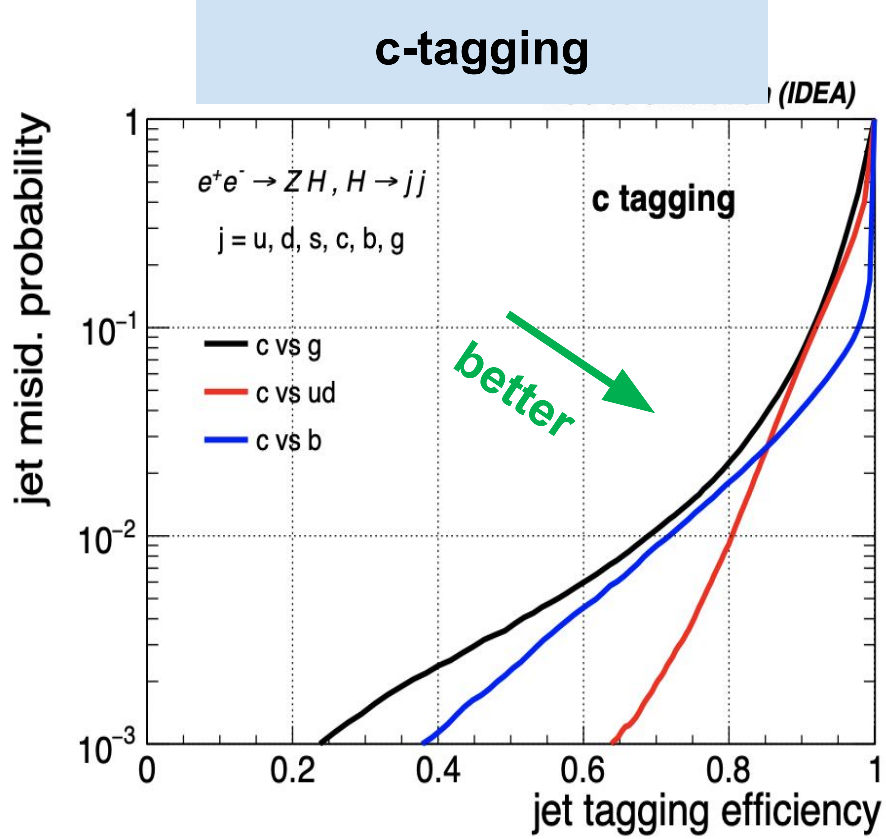
H. Qu and LG
PRD 101 056019 (2020)
F. Bedeschi, M. Selvaggi, LG
EPJ C 82 646 (2022)

- Jet representation: *"Point Cloud"*→ *"Particle Clouds"*
  - Treat the jet as an unordered set of particles
- Algorithm design: Graph Neural Networks
  - Particle cloud represented as a graph
    - Each particle: **node** of the graph; Connections between particles: the **edges**
- Follow a hierarchical learning approach
  - First learn local structures → then move to more global ones



ParticleNet@FCCee: b/c tagging



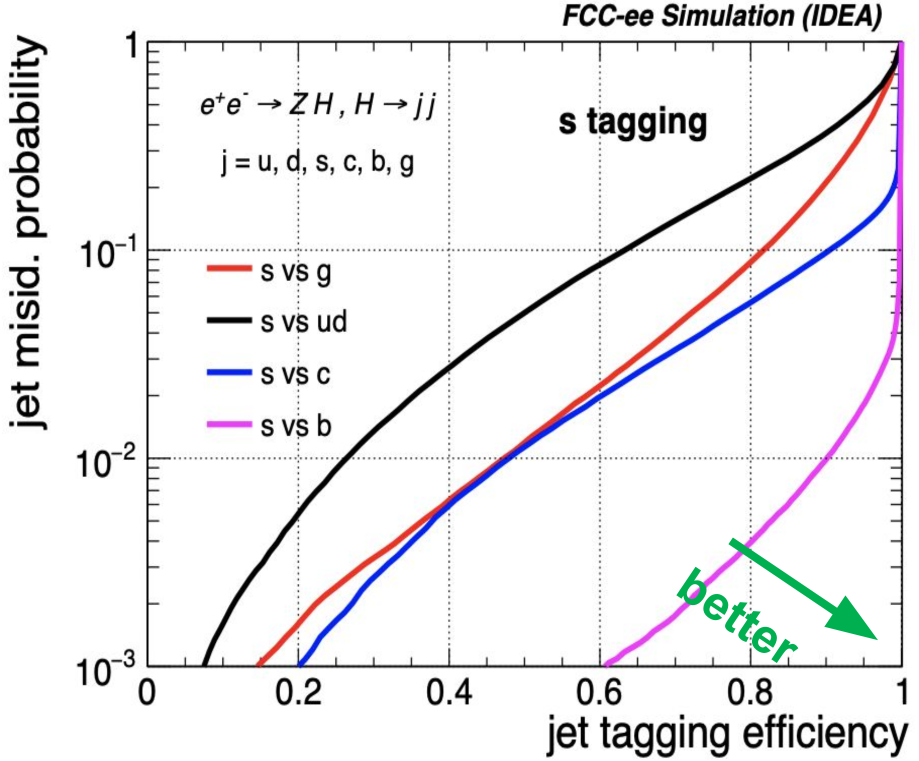


WP	Eff (b)	Mistag (g)	Mistag (ud)	Mistag (c)
Loose	90%	2%	0.1%	2%
Medium	80%	0.7%	<0.1%	0.3%

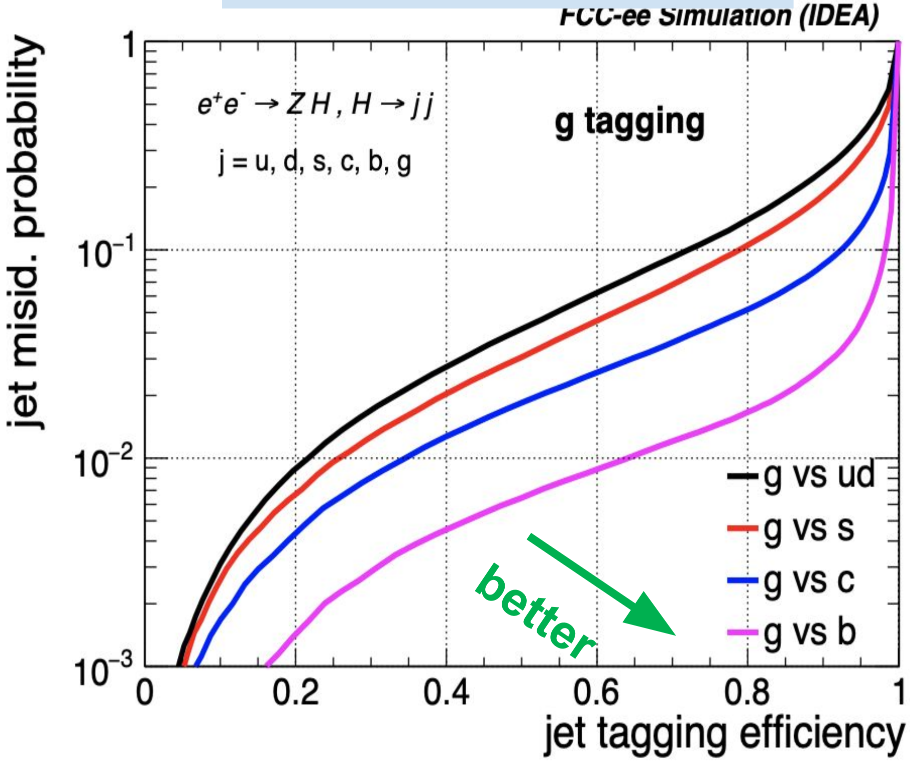
WP	Eff (c)	Mistag (g)	Mistag (ud)	Mistag (b)
Loose	90%	7%	7%	4%
Medium	80%	2%	0.8%	2%

ParticleNet@FCCee: s/g tagging

strange-tagging



gluon -tagging

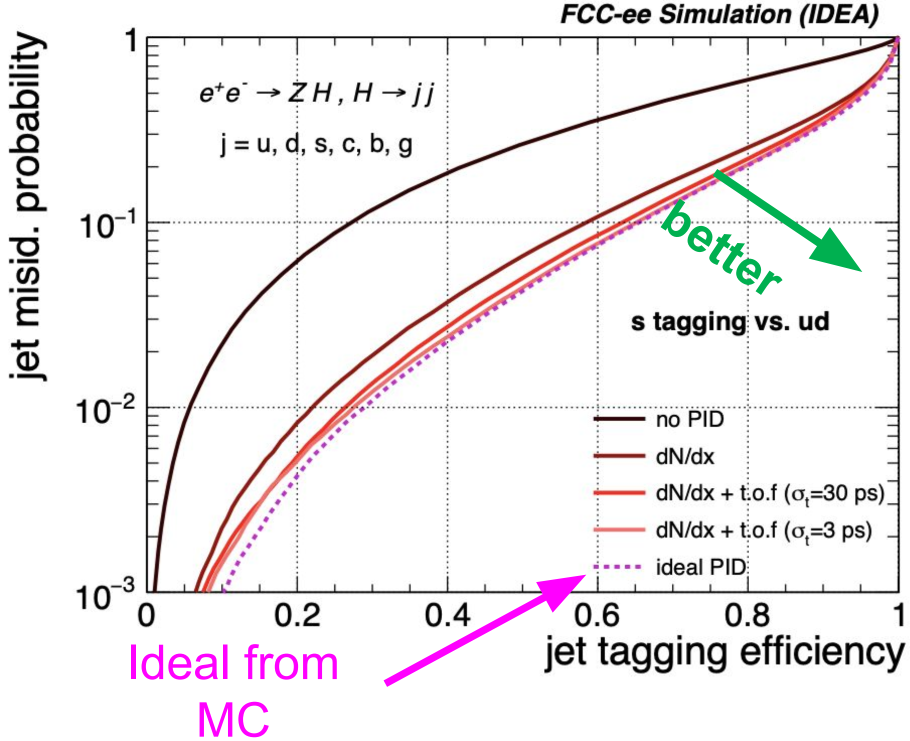


WP	Eff (s)	Mistag (g)	Mistag (ud)	Mistag (c)	Mistag (b)
Loose	90%	20%	40%	10%	1%
Medium	80%	9%	20%	6%	0.4%

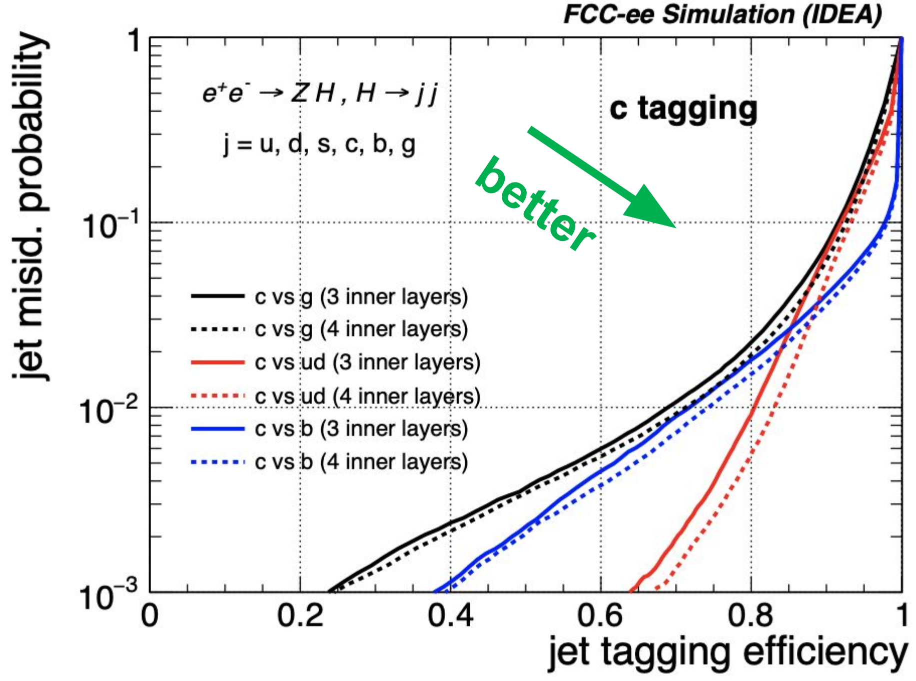
WP	Eff (g)	Mistag (ud)	Mistag (c)	Mistag (b)
Loose	90%	25%	7%	2.5%
Medium	80%	15%	5%	2%

Impact of detector configurations

Strange tagging [PID]



c-tagging [PIX layers]



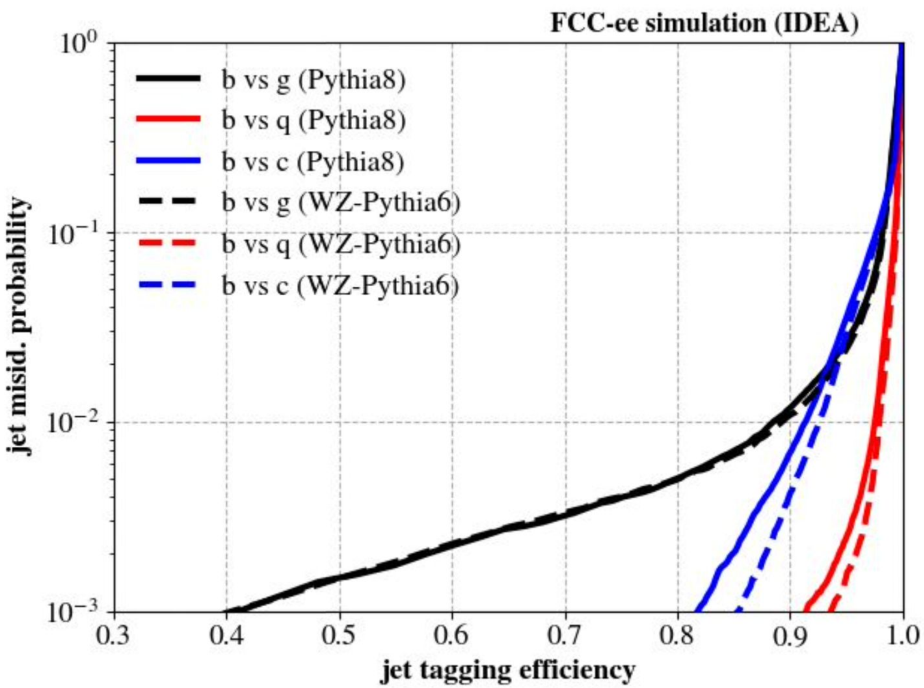
- dN/dx brings most of the gain additional gain w/ TOF (30ps)
  - TOF (3ps): marginal improvement
  - dN/dX + TOF(30ps) ~ perfect PID

- Additional pixel layer 1 cm from beam pipe vs 1.5 cm:
  - improved BKG rejection in c-tagging
  - marginal/no improvement in b-tagging

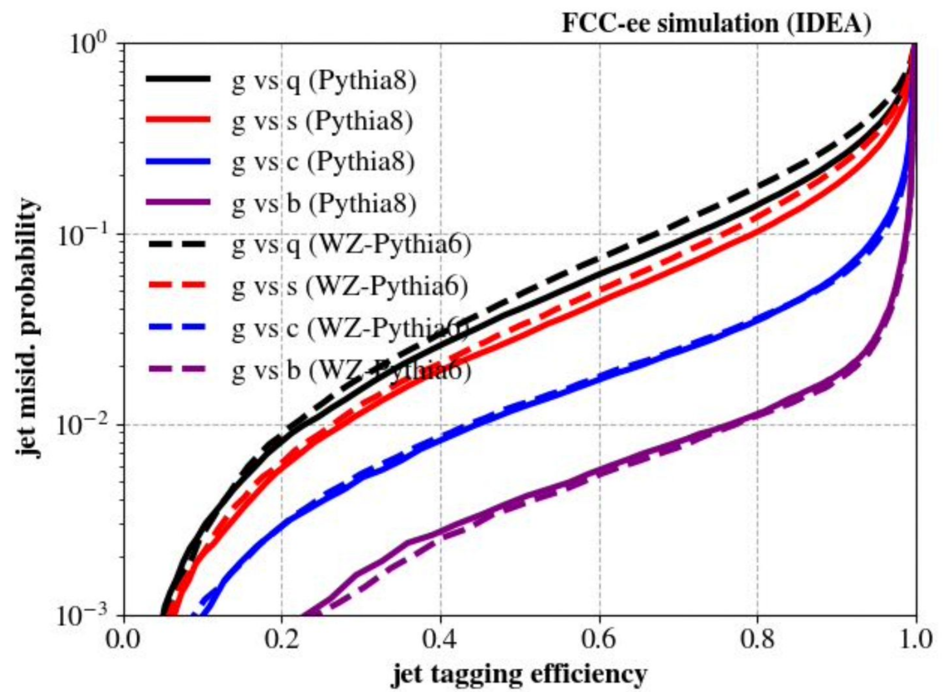
Robustness

- ParticleNet-ee trained using ***Pythia 8*** samples
  - tested on ***Pythia 8*** [solid lines]
  - tested on ***WZ-Pythia6*** [dashed lines]

b-tagging

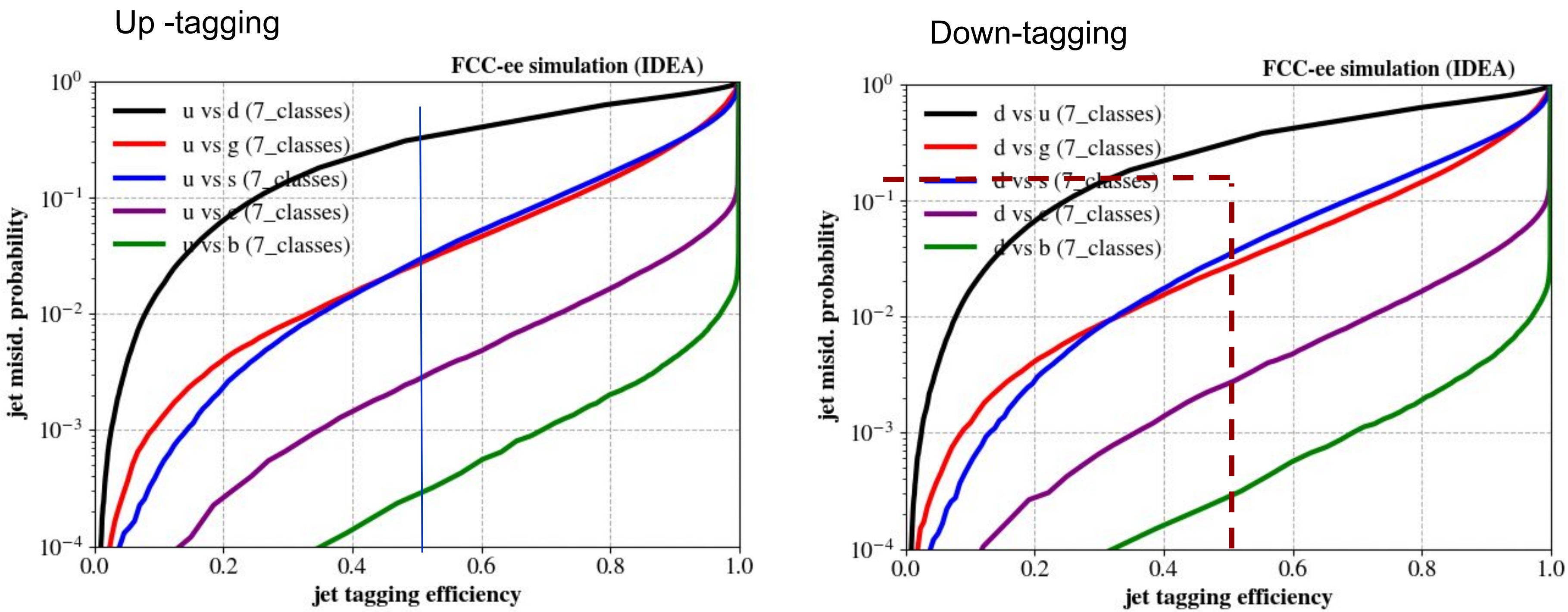


gluon -tagging



- Modest dependence on choice of generator
- More parton showers coming up (Herwig, Sherpa…)

Tagger update(up and down)



- Up vs Down discrimination seems possible thanks to jet charge
- 30% bkg eff at 50% signal (better than random coin toss)

Analysis strategy in a nutshell for H→bb/cc/ss/gg

- Signal: H ⟶ jj (j = b, c, s, g, τ)
- Background:
  - WW/ZZ/Z, qqH, HWW, HZZ

- Key ingredients:
  - Jets reconstruction
    - N = 2 Durham kt exclusive algorithm
  - ParticleNet jet tagger (4 categories: b, c, s, g)

- Analysis:
  1. Events pre-selection (lepton veto:orthogonalize with Z(ll)H analysis, cos theta),
  2. Categorization based on tagger scores
  3. Fit with floating 10% background normalisation uncertainty (to be constrained) and 4 signal strengths (Hbb, Hcc, Hss, Hbb)

Analysis strategy in a nutshell for H→bb/cc/ss/gg

Results using only **vvH** channel:

Hgg : 1.1%
Hss : 150 %
Hcc : 2.7%
Hbb : 0.5%

Event yields in the **S-like highest purity** category (with **122 < mvis < 128**) for **7.2 ab-1**

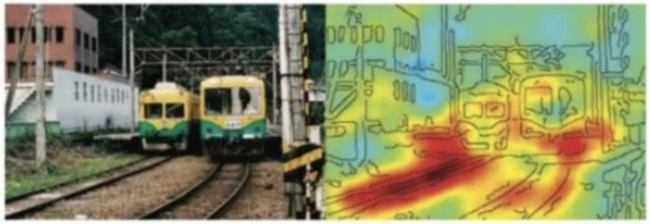
	Hss	Hgg	Hbb	Hcc	Htautau	HWW	HZZ	ZZ	WW	Zqq
N	10	10	0	0	0	8	10	300	150	80
S/B	1	1	0	0	0	1	1	1/30	1/15	1/10

Summary

- Powerful jet flavour identification important for the e^+e^- physics program

- Sophisticated jet tagging algorithms developed for e^+e^- experiments
  - Striking improvement in tagging performance compared to previous tools
    - allows us to explore more of the detector and event reconstruction potential
  - Integrated in FCCSW [data preparation, training, validation, inference, analysis] and used in FCCee physics analyses

- Still room for improvement / other ideas to try:
  - secondary tasks, secondary vertexing regression
  - new higher order graph architectures
  - improve explainability
  - resilience to modelling (more generators)
  - calibration (Z pole $\rightarrow$ ZH threshold extrapolation)



ParticleNet in FCCSW

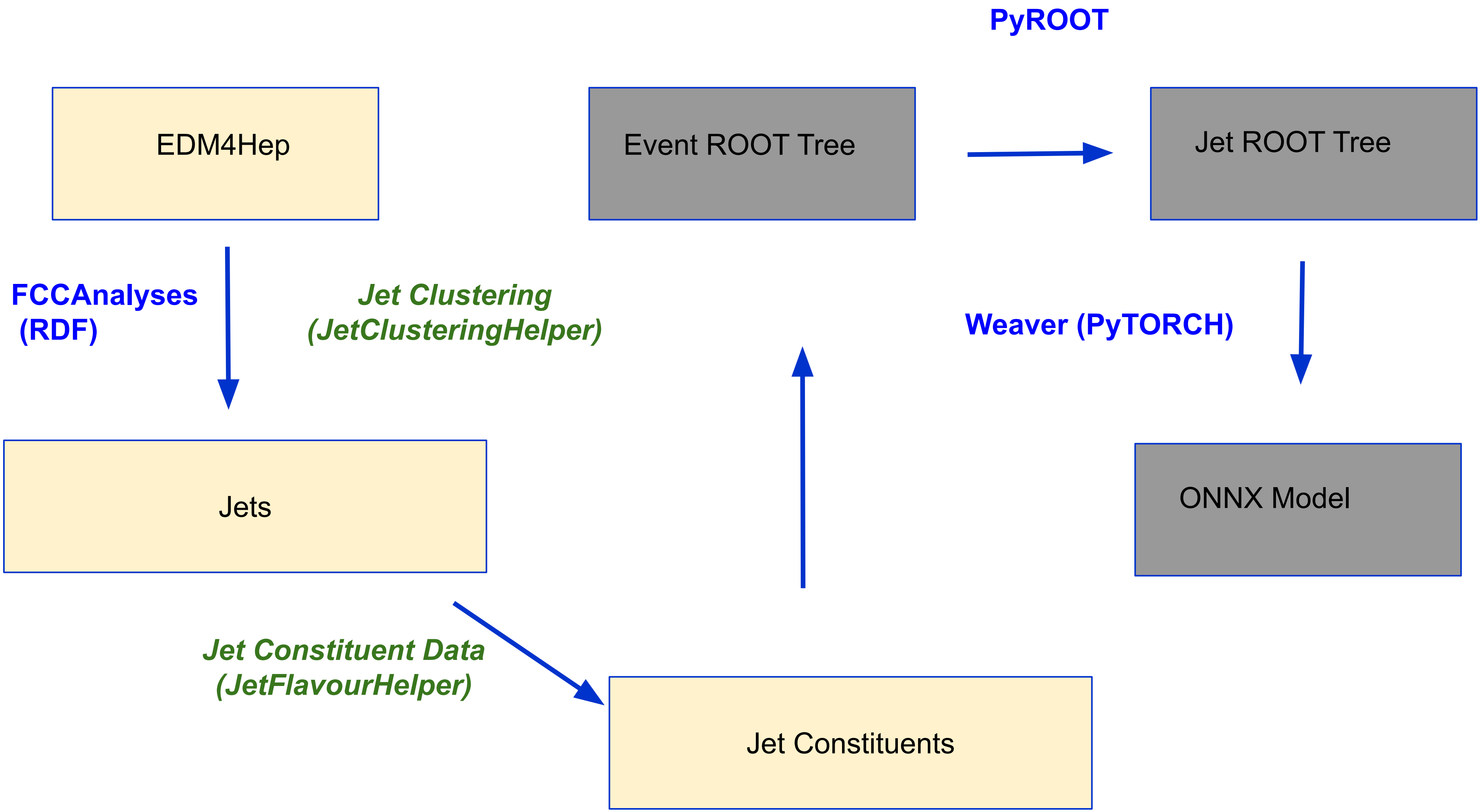
Sample Generation for training

- Generation of the samples in EDM4hep (whole event reconstructions, features for training not explicit)
- FCCAnalyses (wrapper RDataFrame)
  - Per-event $\longrightarrow$ per-jet structure
  - 2 stages. 1: read edm4hep and extract features. 2: produce n-tuples one per class.
  - final dataset: 5/7 classes and 10^6 events per class
  - trained on gpus (A100)

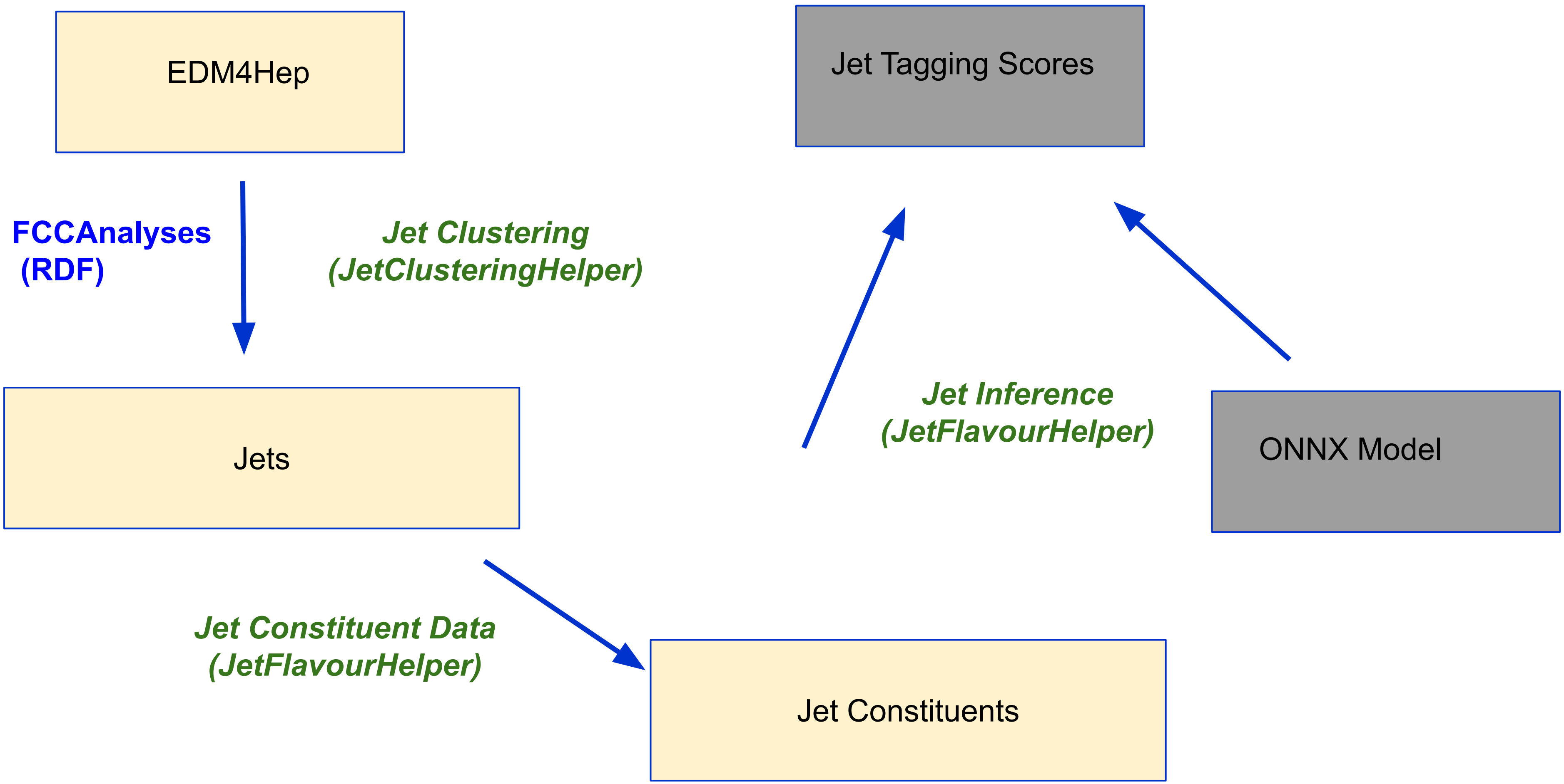
Inference

- Inference in FCCAnalyses:
  - load ONNX training files
  - Extract hard vertex and perform jet clustering
  - Extract jet constituents and compute observables
  - Evaluate NN $\longrightarrow$ output: **one probability per category**

Training the model



Inference


EDM4Hep
FCCAnalyses
(RDF)
Jet Clustering
(JetClusteringHelper)
Jets
Jet Constituent Data
(JetFlavourHelper)
Jet Constituents
Jet Tagging Scores
Jet Inference
(JetFlavourHelper)
ONNX Model

Inference with FCCAnalyses

```
## get local file, else download from url
weaver_preproc = get_file_path(url_preproc, local_preproc)
weaver_model = get_file_path(url_model, local_model)

from addons.ONNXRuntime.python.jetFlavourHelper import JetFlavourHelper
from addons.FastJet.python.jetClusteringHelper import ExclusiveJetClusteringHelper
```

Loading model parameters

```
## define jet clustering parameters
jetClusteringHelper = ExclusiveJetClusteringHelper(collections["PFParticles"], njets, tag)

## run jet clustering
df = jetClusteringHelper.define(df)

## define jet flavour tagging parameters
```

JetClusteringHelper

Jet clustering

```
jetFlavourHelper = JetFlavourHelper(
    collections,
    jetClusteringHelper.jets,
    jetClusteringHelper.constituents,
    tag,
)

## define observables for tagger
df = jetFlavourHelper.define(df)

## tagger inference
df = jetFlavourHelper.inference(weaver_preproc, weaver_model, df)

return df
```

JetFlavourHelper

Obtain input parameters

Run inference