Flavor-tagging using BDT

2025 0317

Content

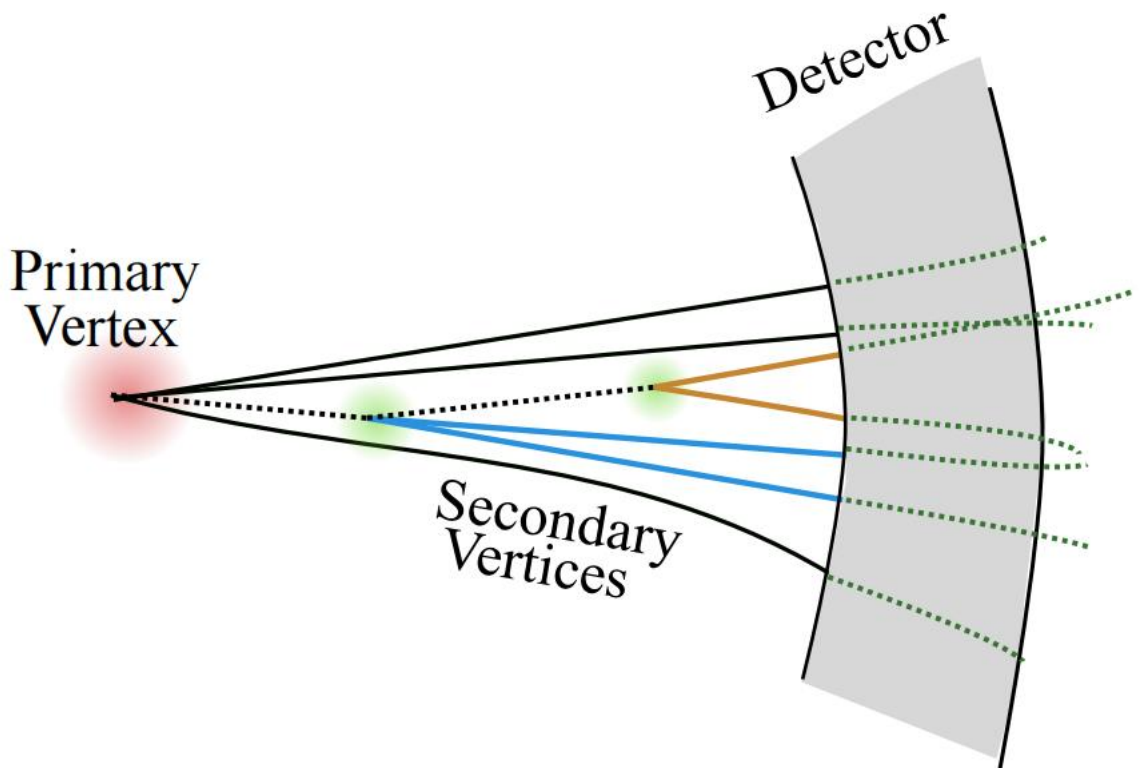
- **Motivation**
- **Features**
- **BDT results**
- **Truth match**
- **Lepton ID**

from yifan

from changhua

Motivation

Higgs		
Observable	HL-LHC projections	CEPC precision
M_H	20 MeV	3 MeV
Γ_H	20%	1.7%
$\sigma(ZH)$	4.2%	0.26%
$B(H \to bb)$	4.4%	0.14%
$B(H \to cc)$	-	2.0%
$B(H \to gg)$	-	0.81%
$B(H \to WW^*)$	2.8%	0.53%
$B(H \to ZZ^*)$	2.9%	4.2%
$B(H \to \tau^+\tau^-)$	2.9%	0.42%
$B(H \to \gamma\gamma)$	2.6%	3.0%
$B(H \to \mu^+\mu^-)$	8.2%	6.4%
$B(H \to Z\gamma)$	20%	8.5%
$B\text{upper}(H \to \text{inv.})$	2.5%	0.07%

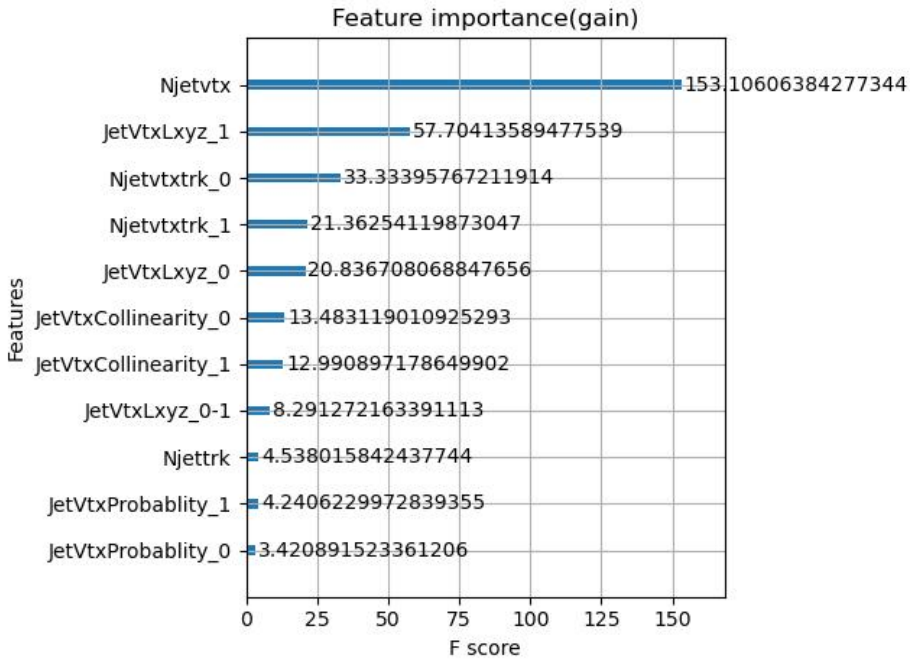


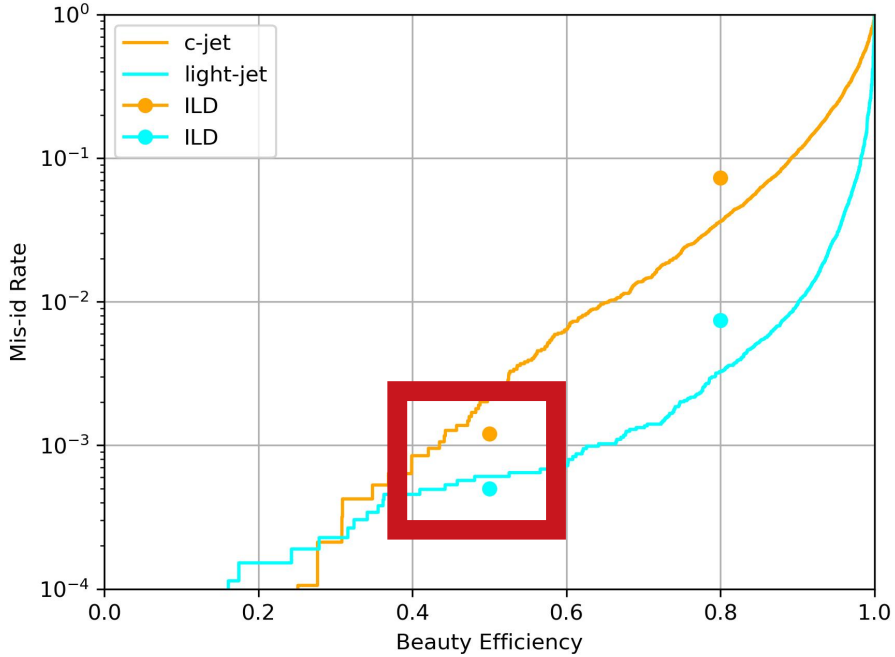
>90% of ZH events contains jets

Jet Flavor Tagging is critical for revealing fundamental physical processes

Motivation

from previous job





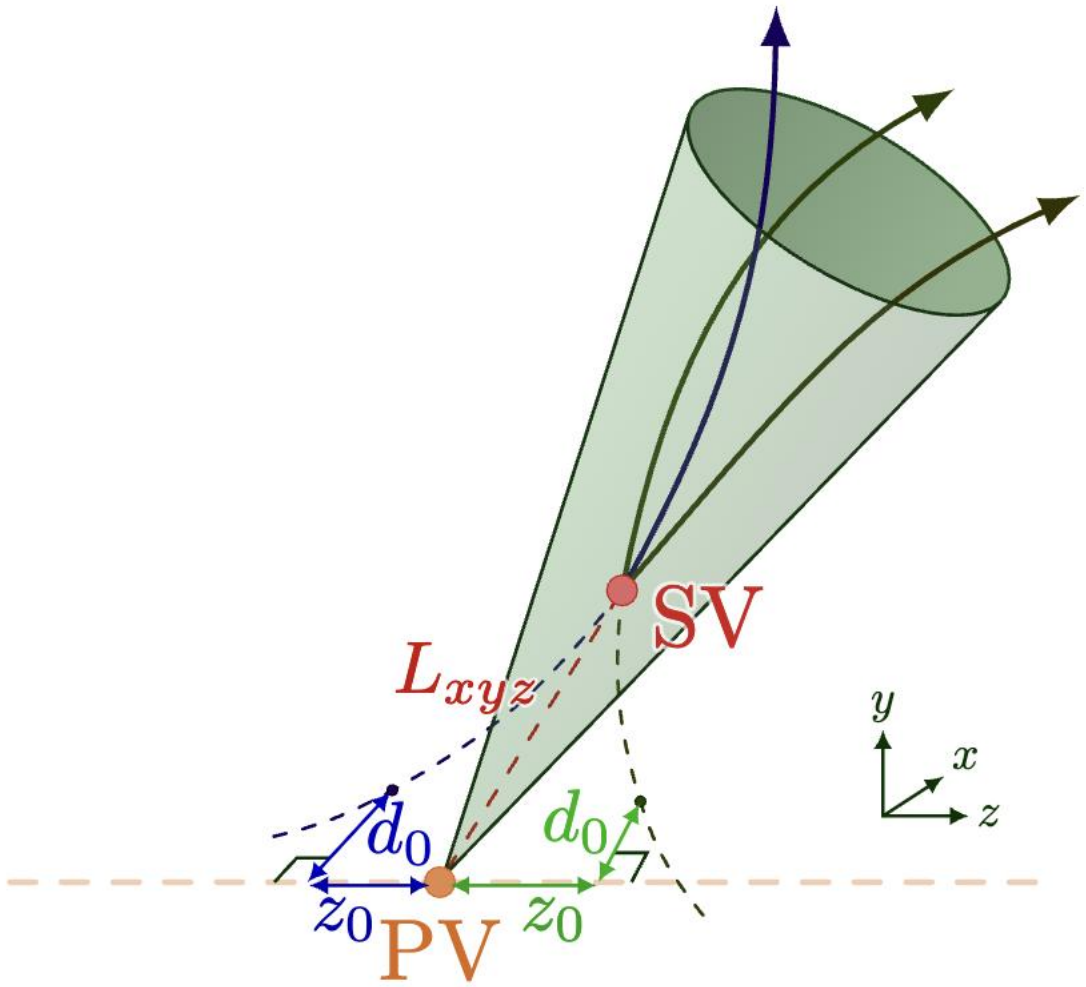
the performance not as good as those in ILD

Need to add more variables refer to ILD

ILD paper

https://arxiv.org/abs/1506.08371

Features



track level features
- Number of tracks in a jet
- Number of tracks associated to a vertex
- track parameters(d0/z0) ...

vertex level features
- Number of vertices in a jet
- Flight of distance of a vertex ...

number of features:
- 33 used from ILD
- add 7 from ours

*Only the first two vertices information used when using vector-liked features

Feature categorize

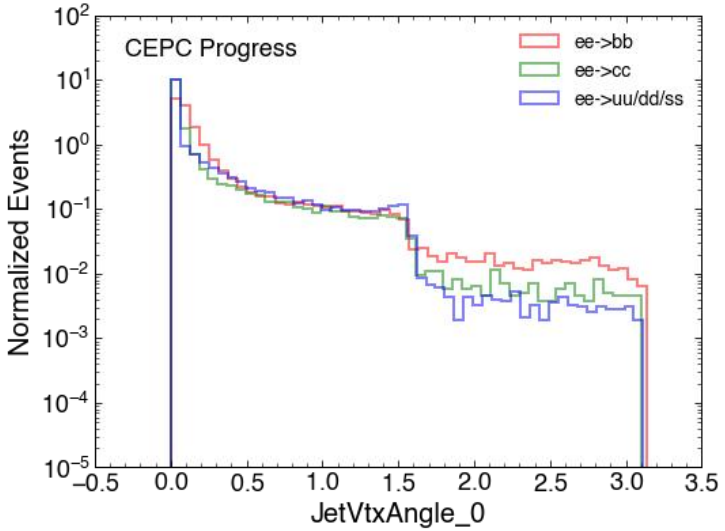
ild_features: choose the features same as those in ILD paper

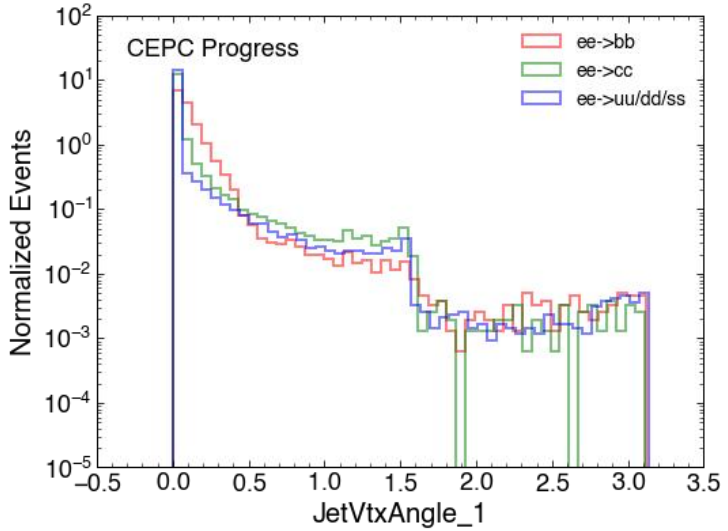
features_ild = ["SingleVtxProb", "JetVtxLxyz_0", "JetVtxLxyz_1",
"vtxlen12","JetVtxAngle_0", "JetVtxAngle_1", "vtxsig_0", "vtxsig_1","vtxsig12", "vtxmult",
"Njetvtxtrk_0", "Njetvtxtrk_1", "JetVtxMomenta_0", "JetVtxMomenta_1" ,"JetVtxMass_0",
"JetVtxMass_1", "trk1d0sig", "trk2d0sig", "trk1z0sig", "trk2z0sig", "trk1pt", "trk2pt",
"nelectron", "nmuon", "JetTotalVtxMass","allTrackD0bprob", "allTrackD0cprob",
"allTrackD0qprob", "allTrackZ0bprob", "allTrackZ0cprob",
"allTrackZ0qprob","allTrackMass", "JetVtxProbablity_0" ,"JetVtxProbablity_1"]

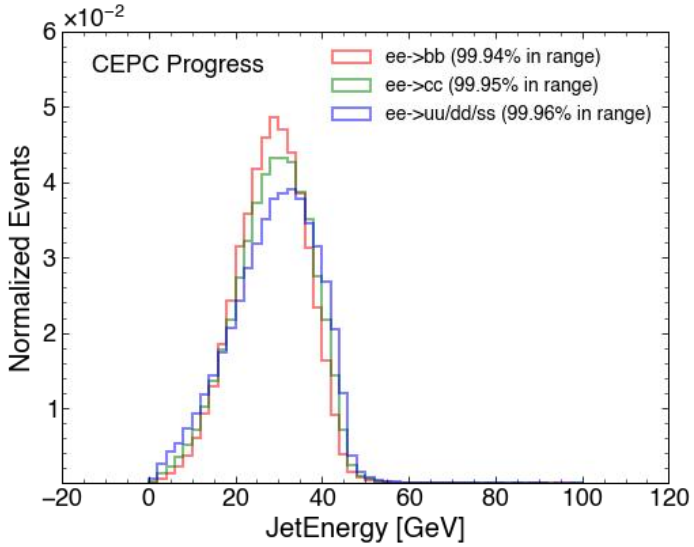
all_features: choose both of ILD_features and features added

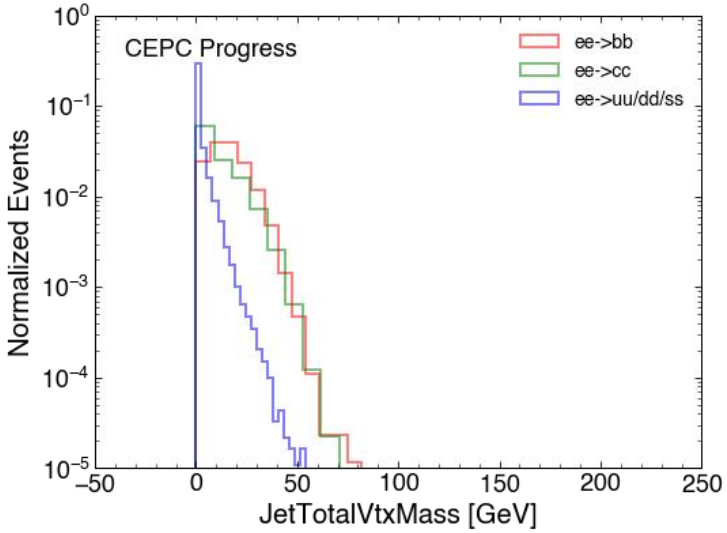
features_ild +
Njettrk, Njetvtx, Njetvtxfrac, JetVtxCollinearity, JetVtxChi2, JetVtxEnergy, JetVtxKin

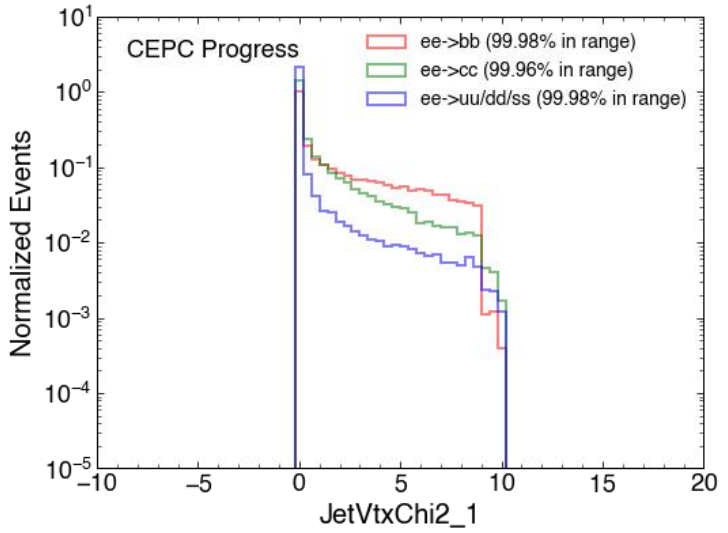
Jet Tagging Vars Distribution

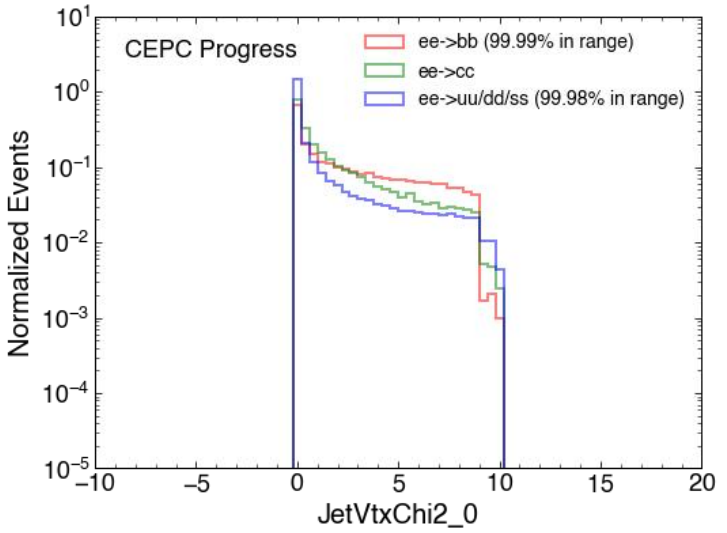






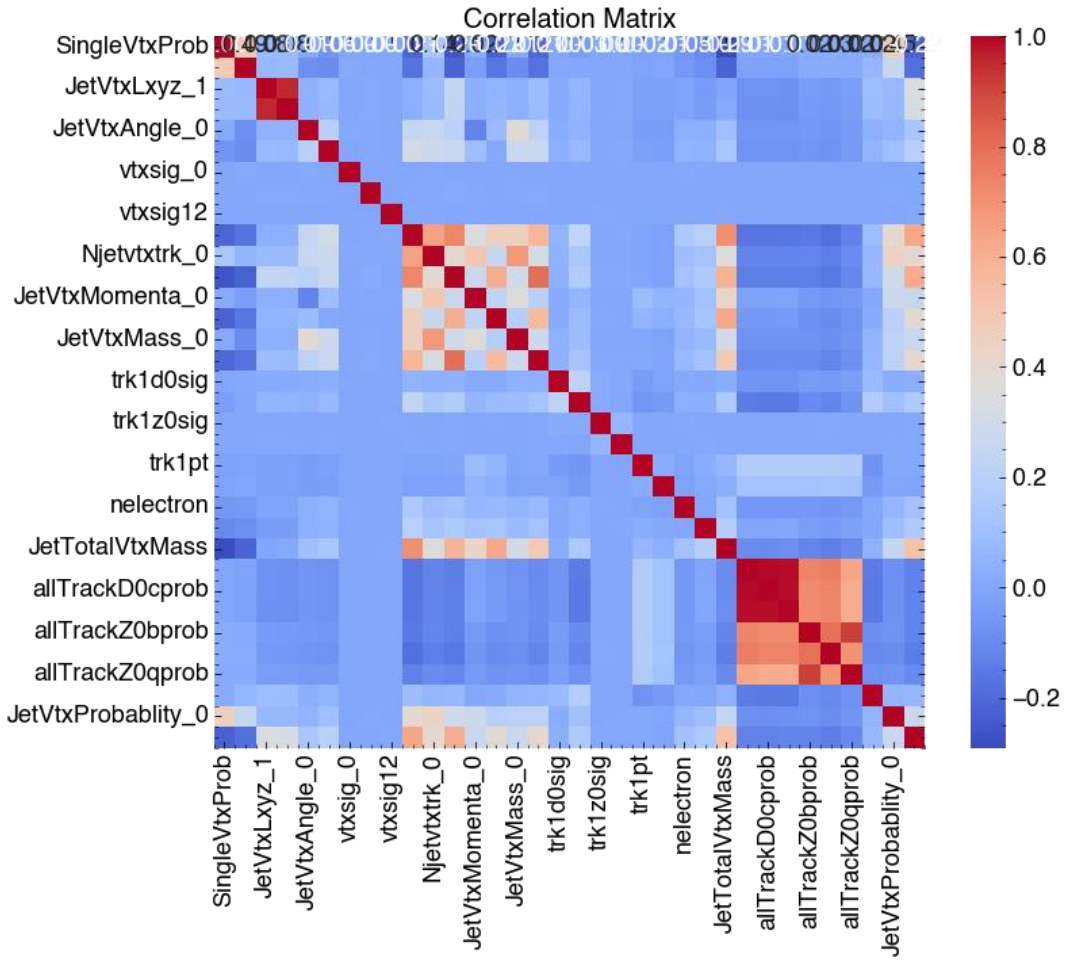


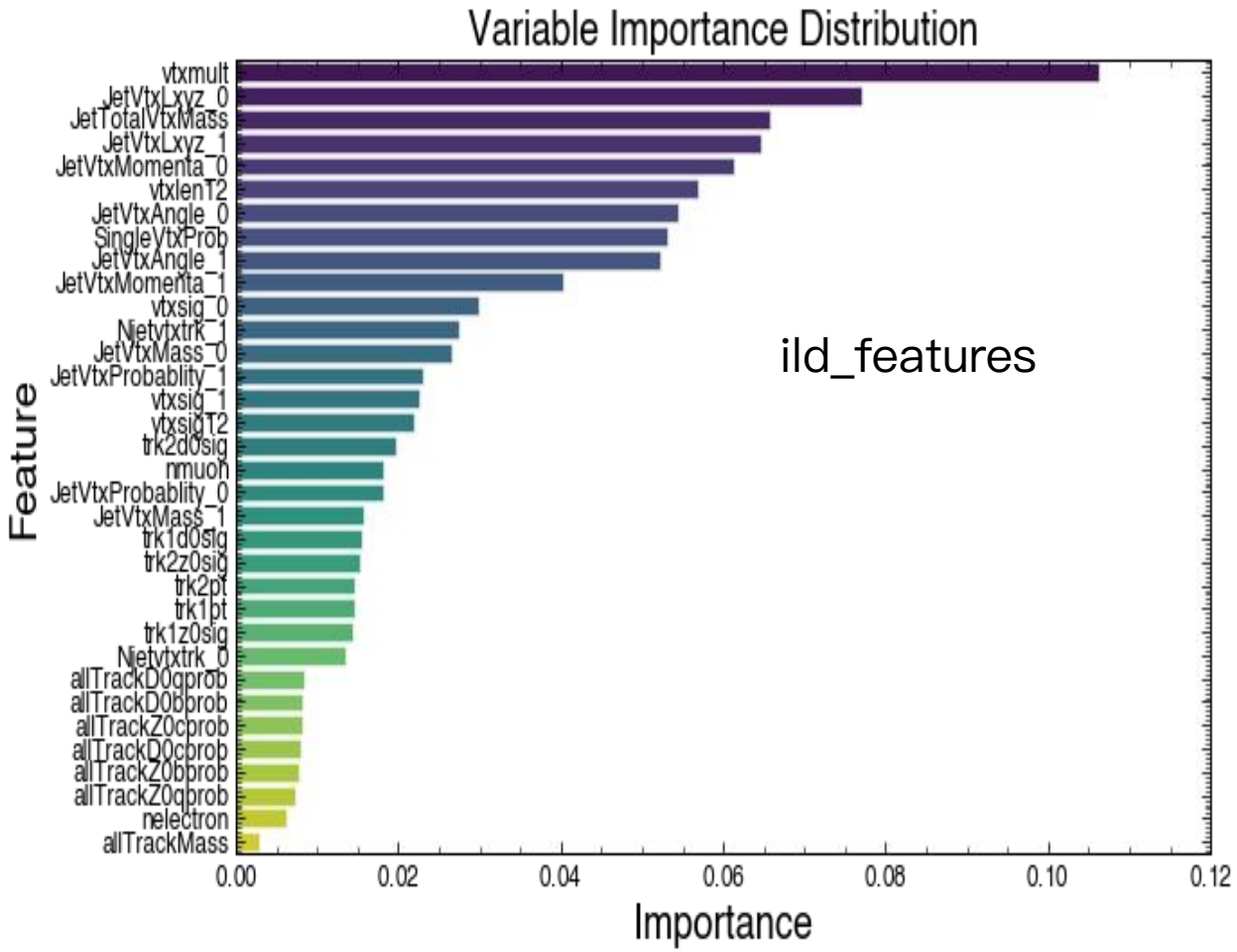




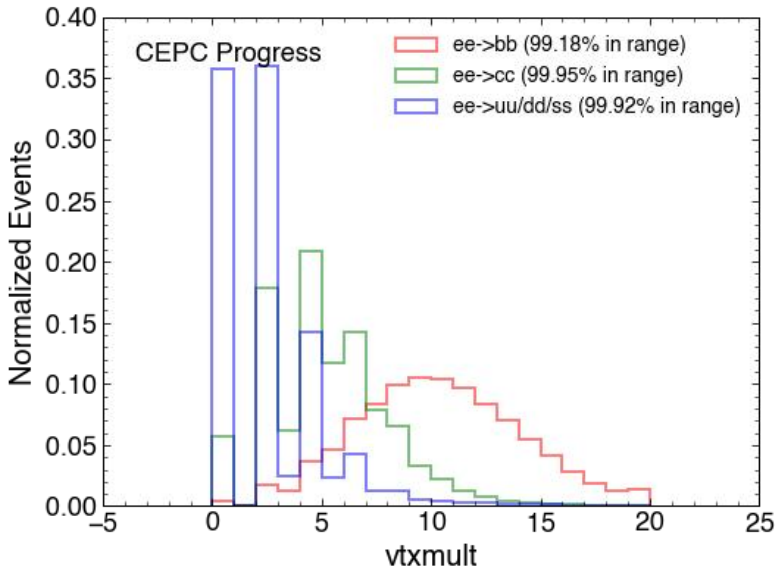
The distribution and definition of all variables can be found in the backup

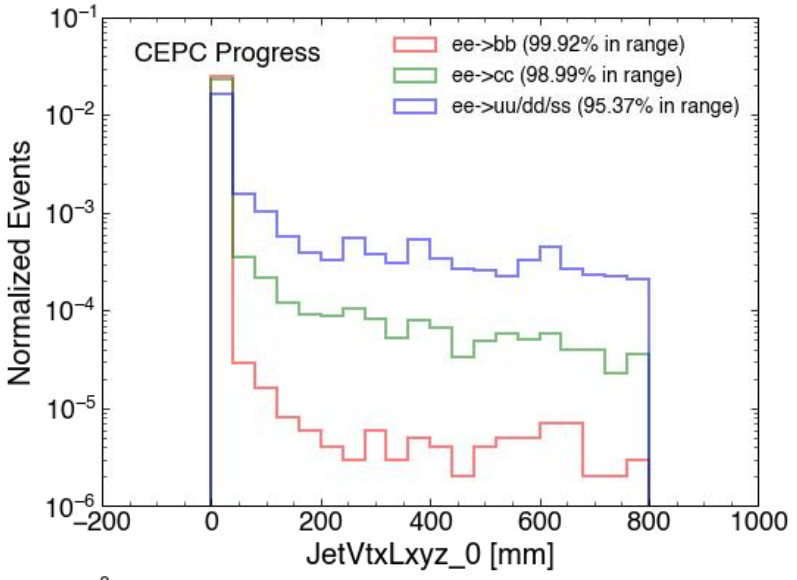
Vars importance and correlation

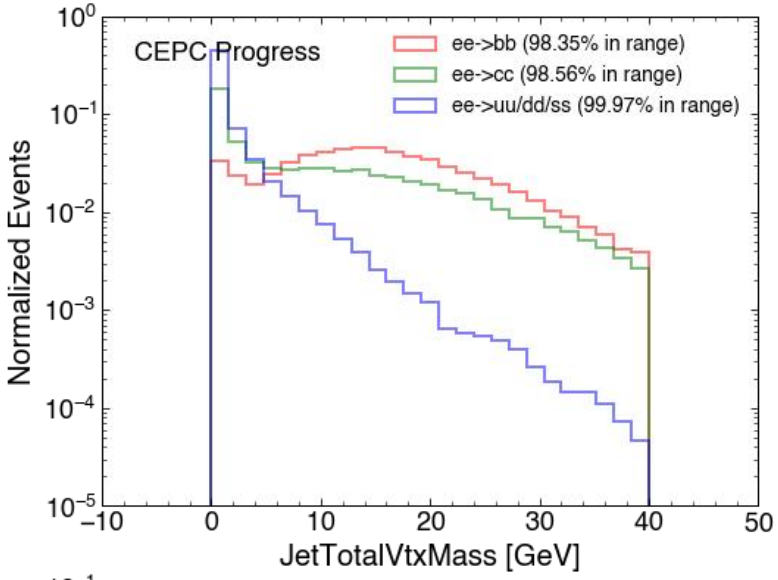


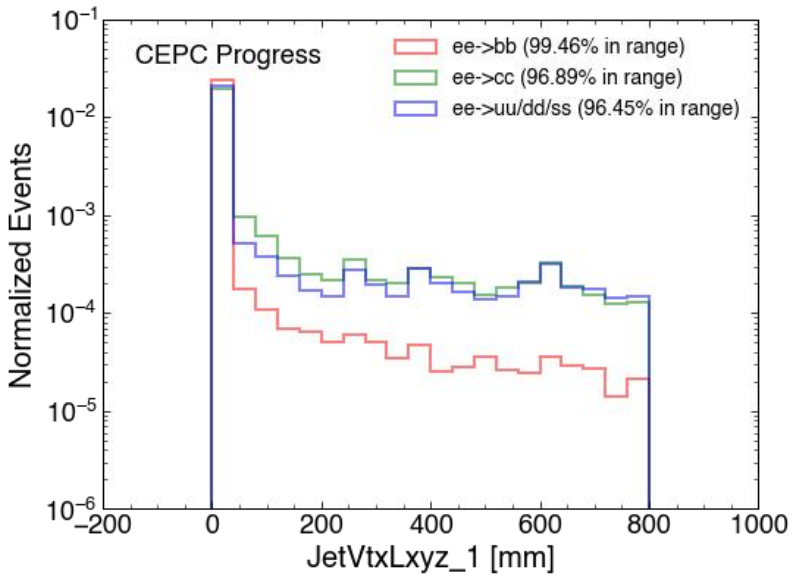


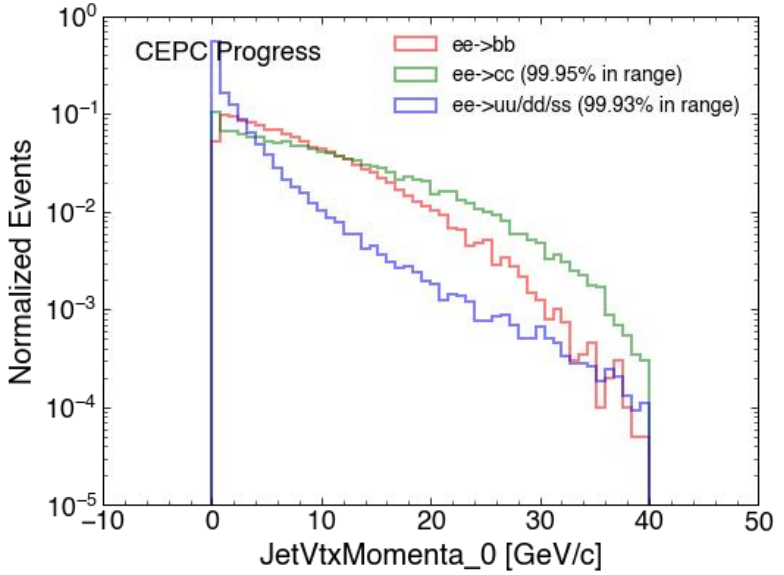
Discriminating vars

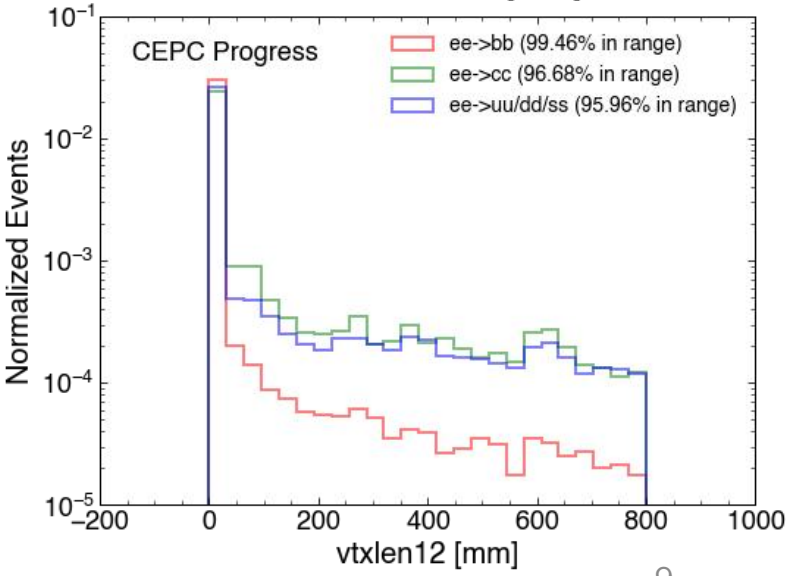








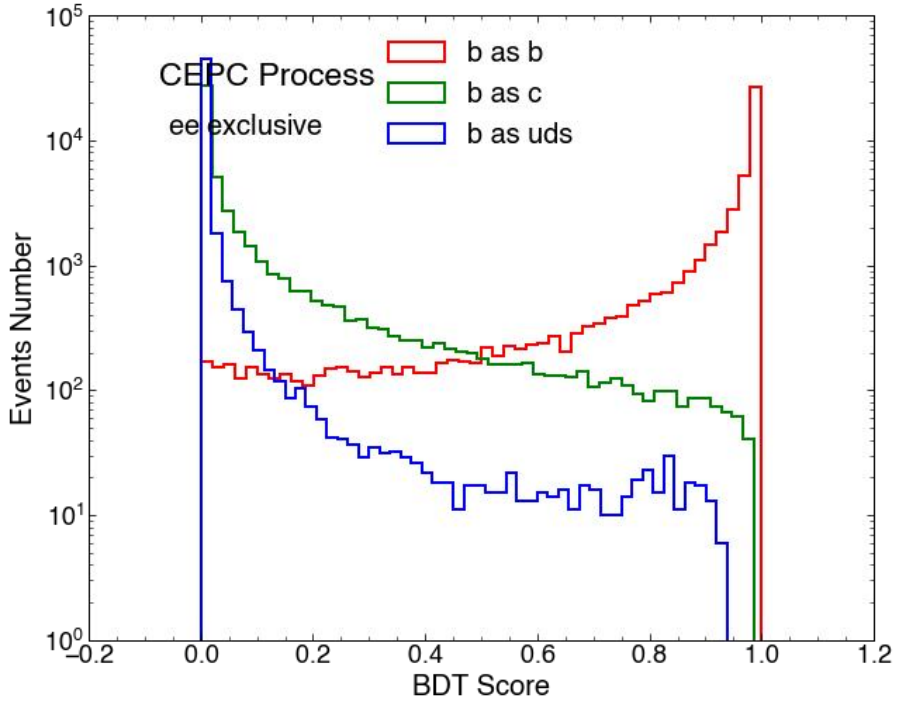




BDT results

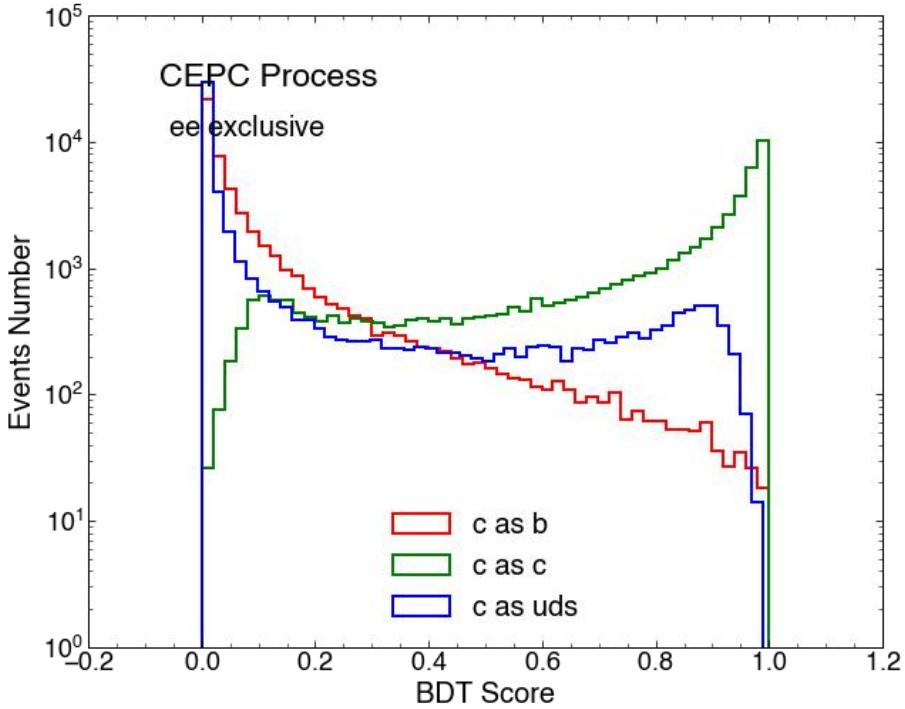
ild_features

ee->bb



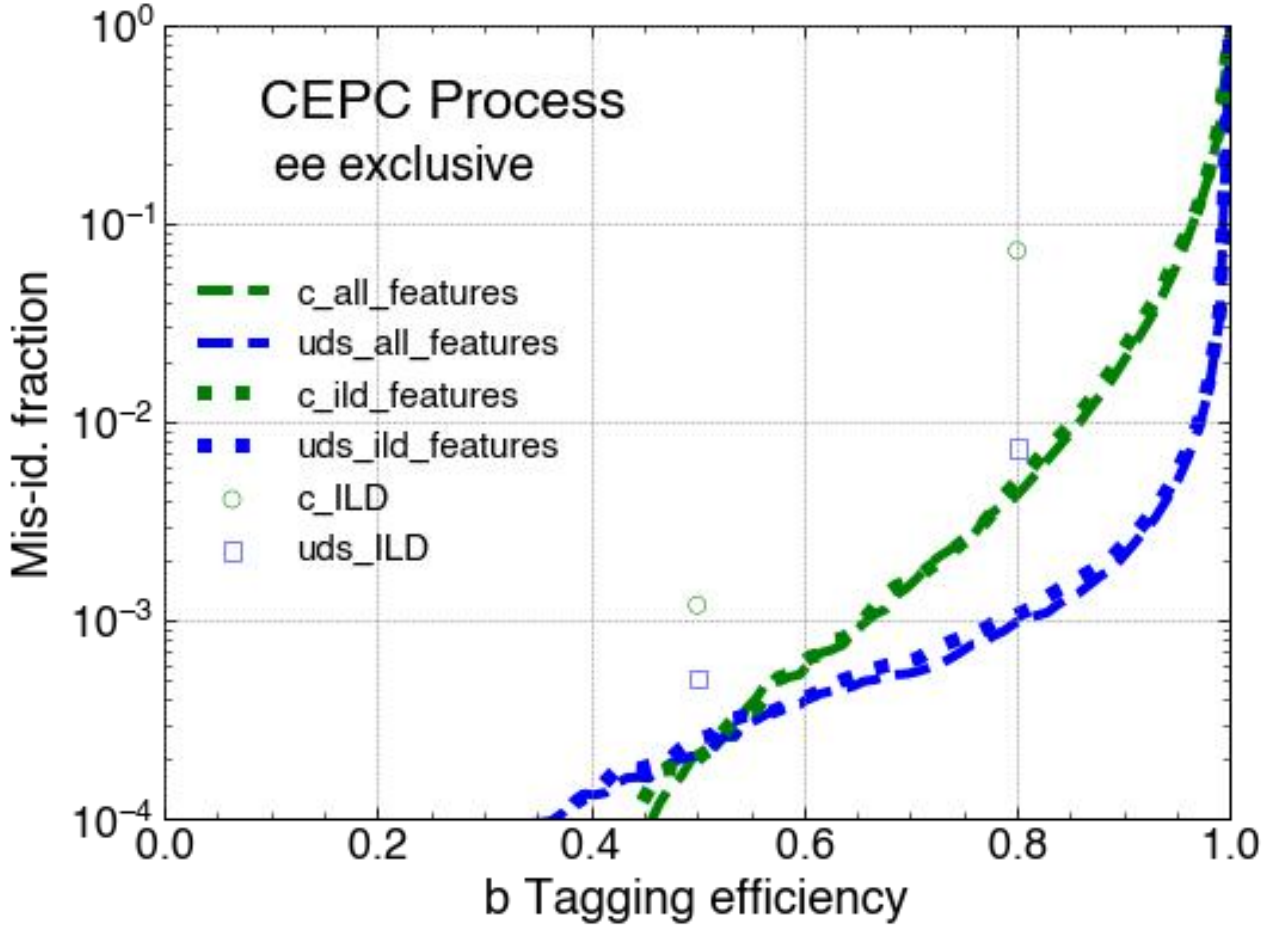
Score	b_eff	c_rej_eff	uds_rej_eff
0.30	0.95	0.89	0.98
0.60	0.90	0.96	0.99
0.90	0.75	0.99	0.99

ee->cc



Score	c_eff	b_rej_eff	uds_rej_eff
0.30	0.88	0.91	0.81
0.60	0.75	0.97	0.88
0.90	0.5	0.99	0.98

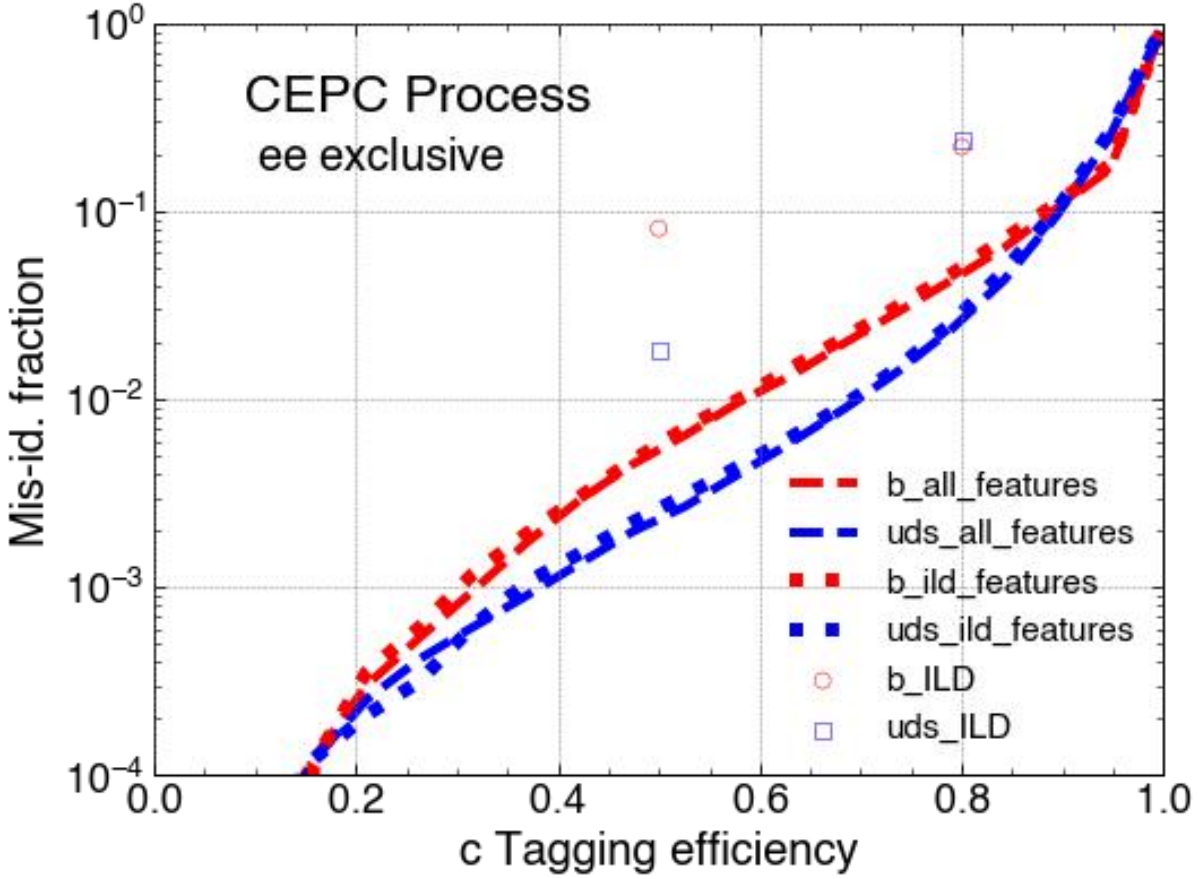
BDT results



the c–jet rejection improves by a factor of **17 (5)** at a b–tagging efficiency of 0.8 (0.5).

the uds–jet rejection improves by a factor of **7 (2)** at a b–tagging efficiency of 0.8 (0.5).

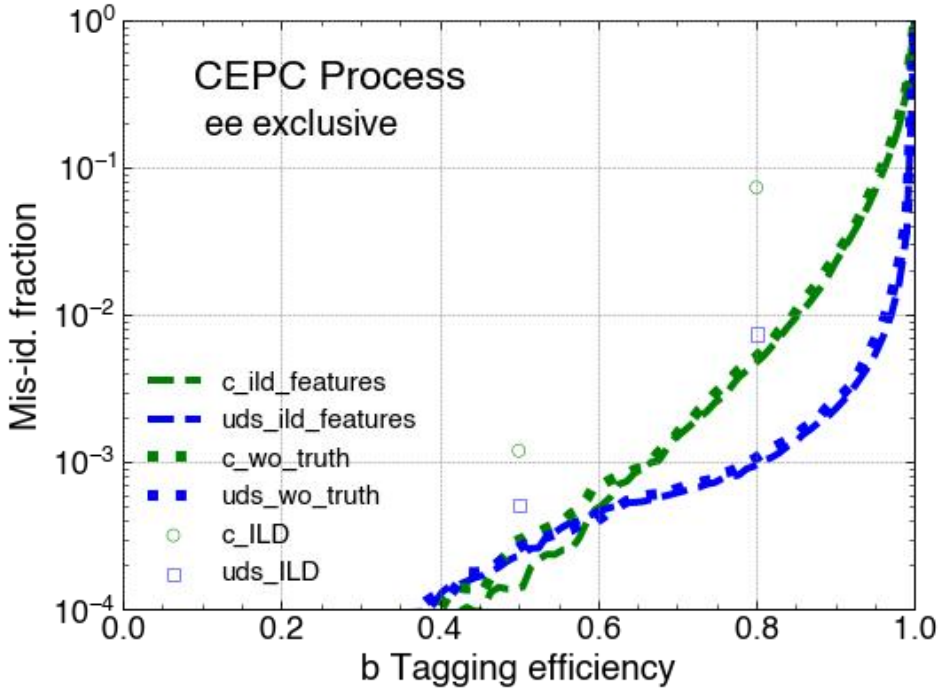
BDT results

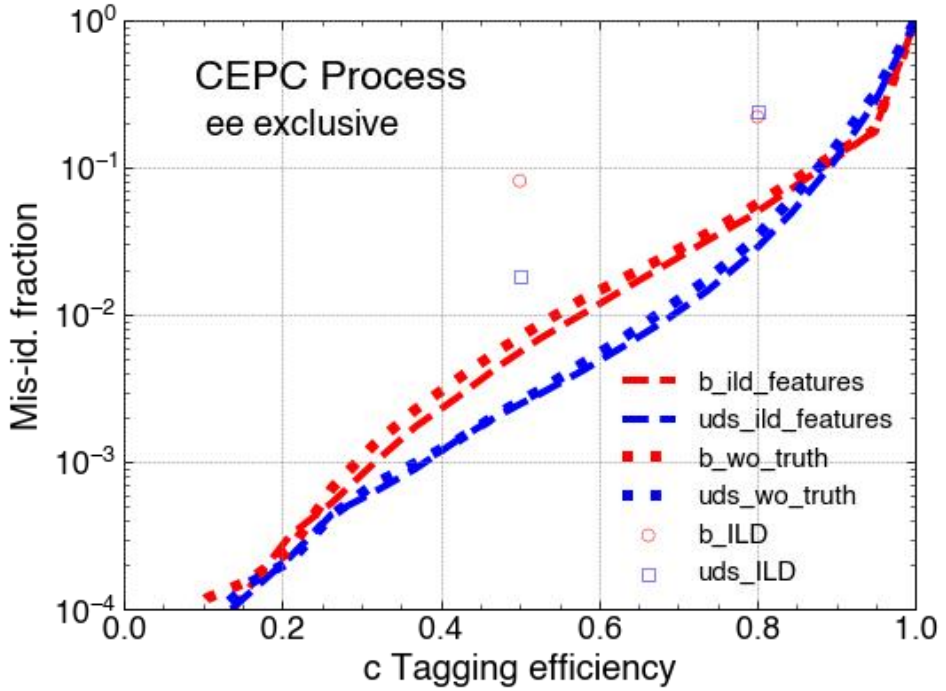


the b–jet rejection improves by a factor of **4 (15)** at a b–tagging efficiency of 0.8 (0.5).

the uds–jet rejection improves by a factor of **8 (7)** at a b–tagging efficiency of 0.8 (0.5).

Remove the truth lepton PID

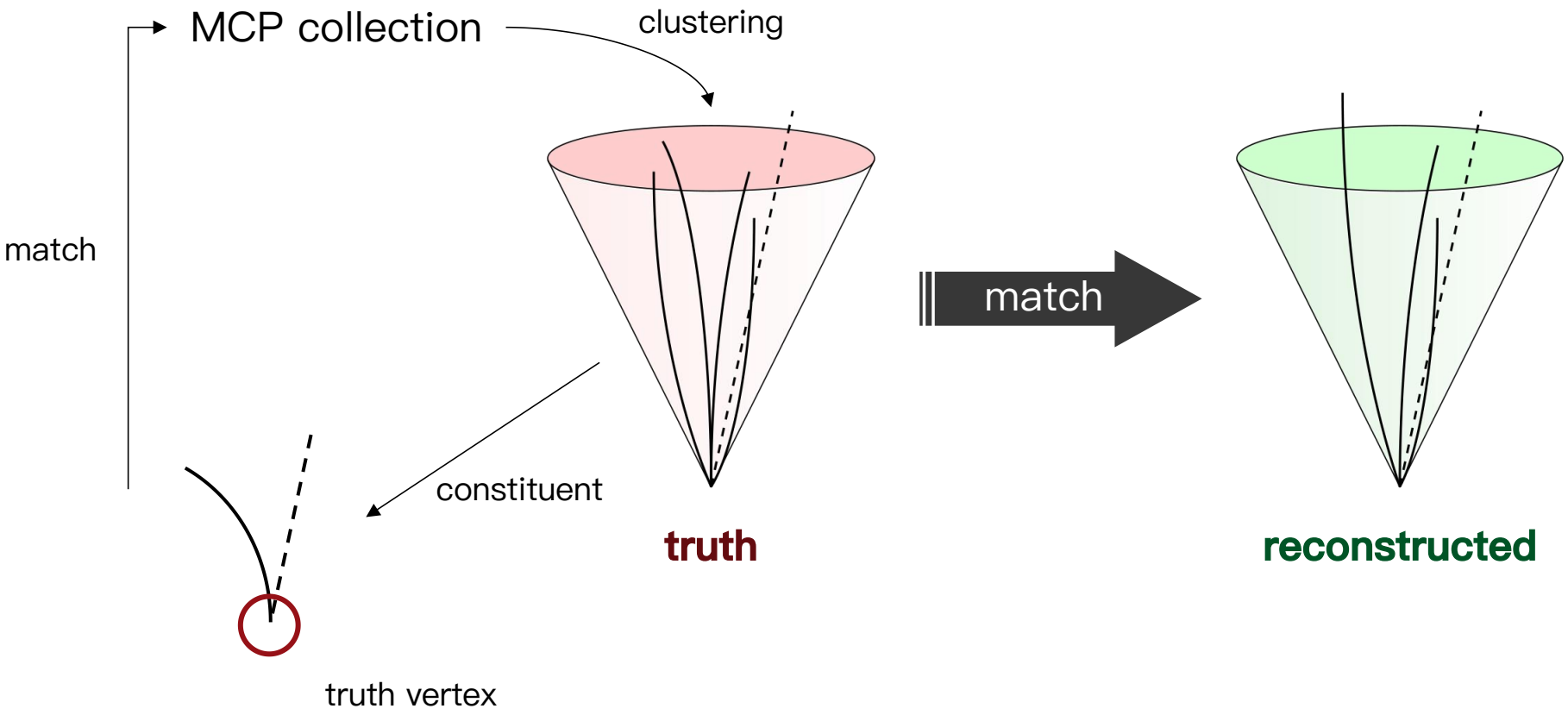




Truth matching idea

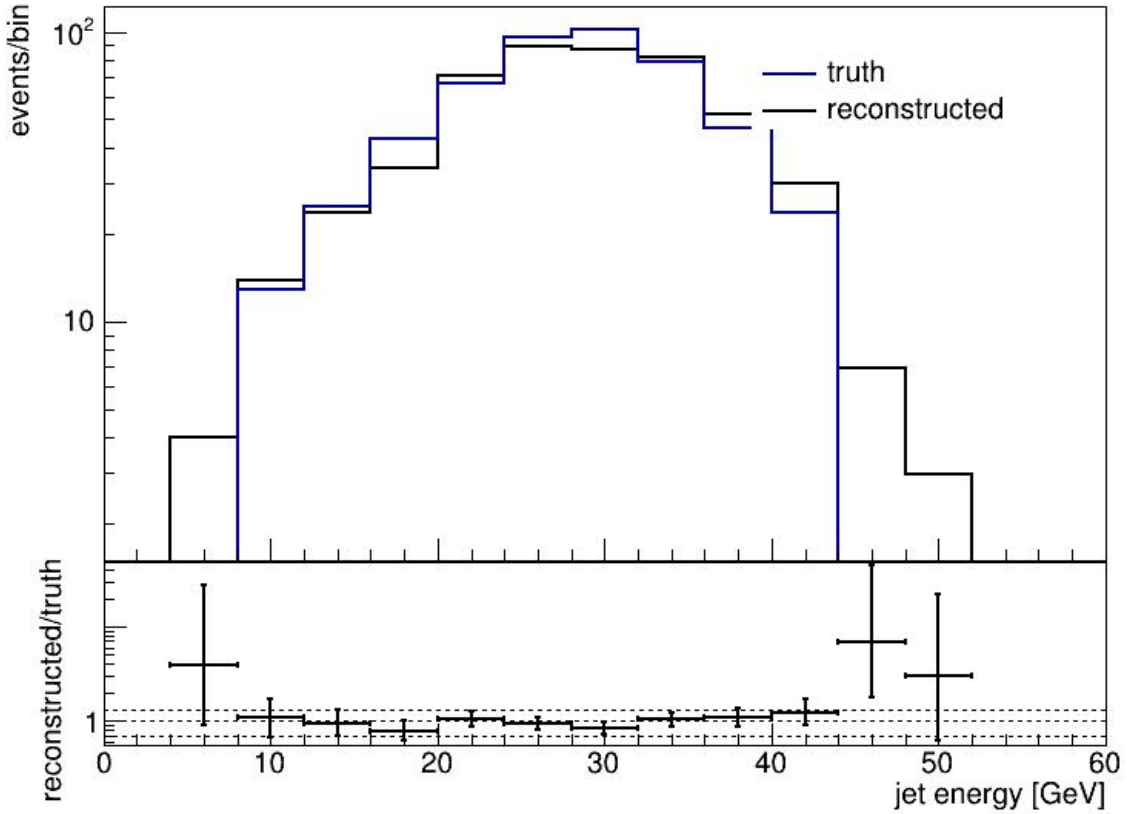
- Reconstructed jet: clustered tracks
- Truth jet: clustered MCP (charged & stable)
- Match truth jet to reconstructed jet
- Truth vertex: vertex of MCP in matched truth jet

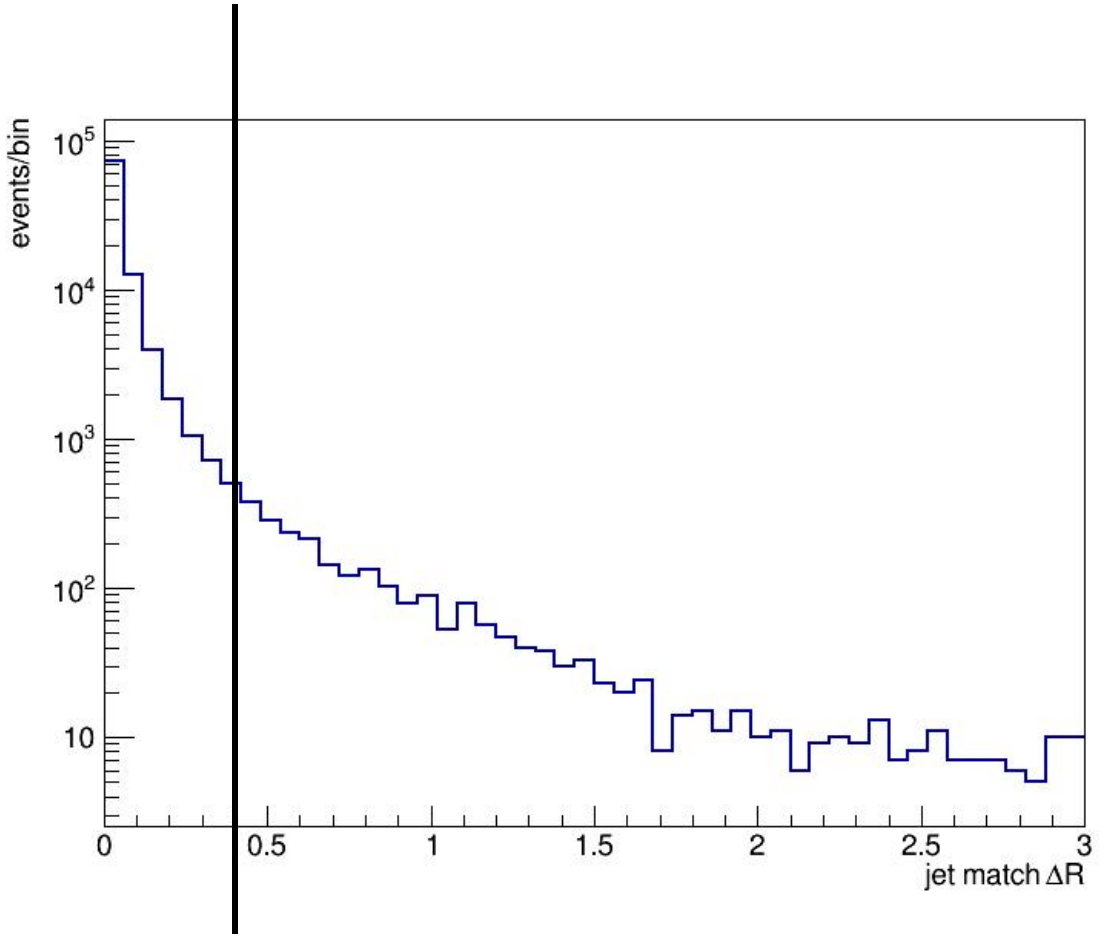
from yifan



Jet matching performance

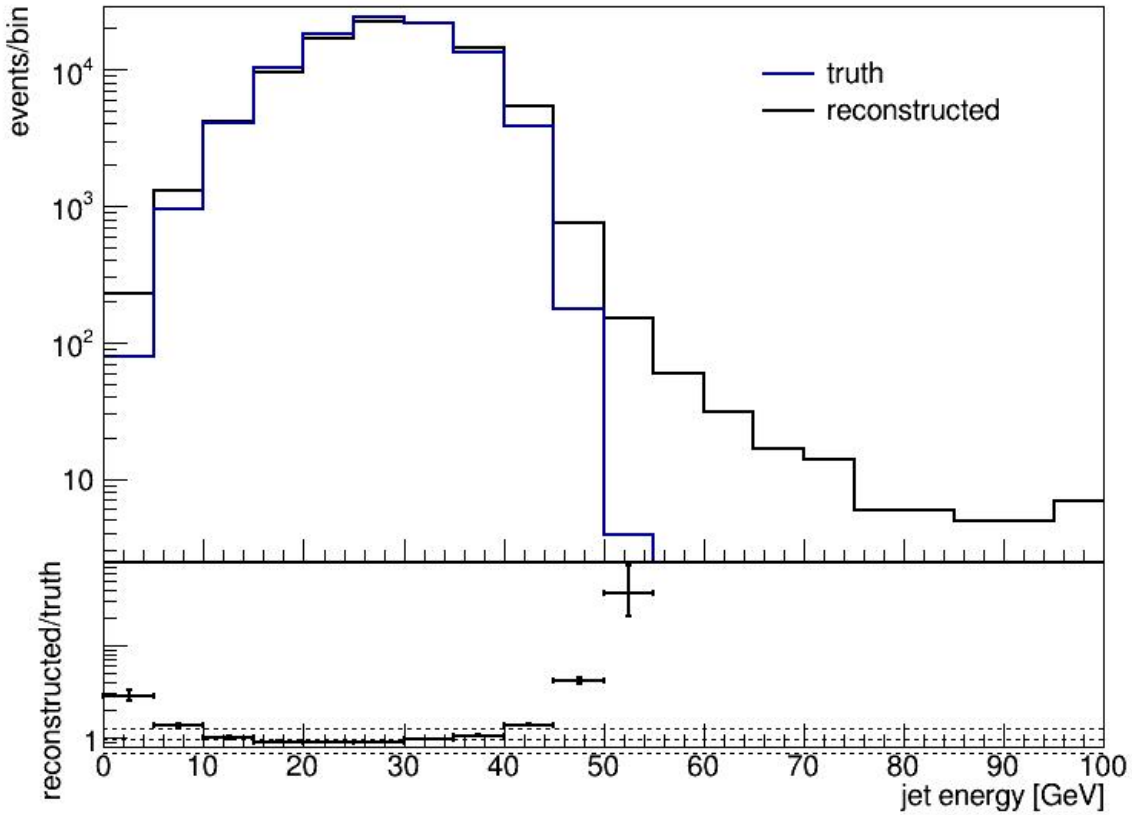
- ee->bb @ 240GeV
- Match jet by ΔR
- ΔR and energy match well

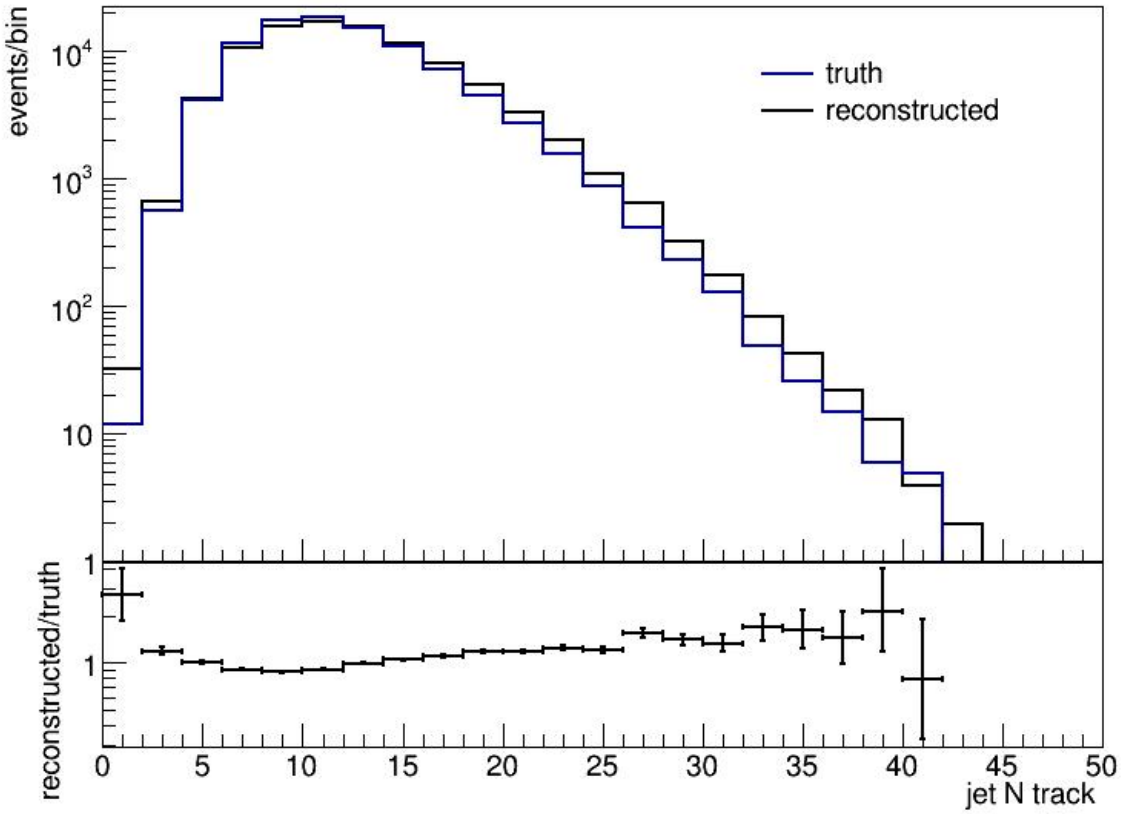




Jet vars

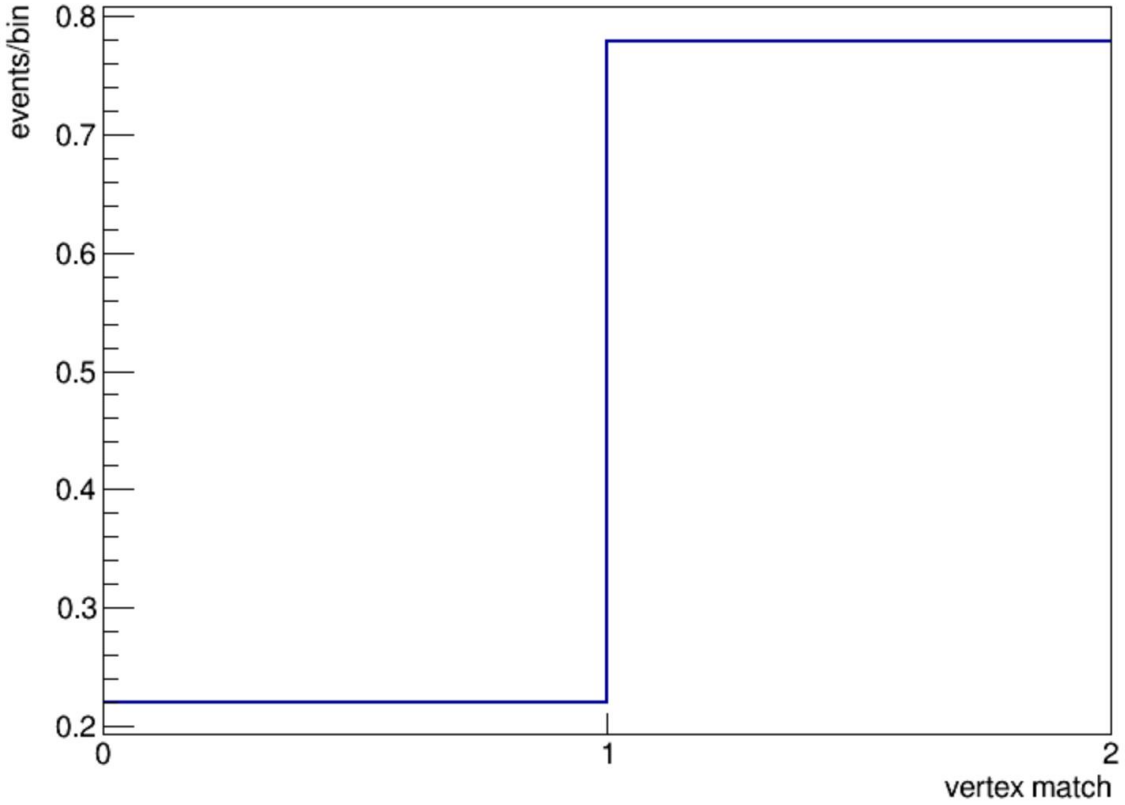
- ΔR < 0.4
- N tracks in jet matches well



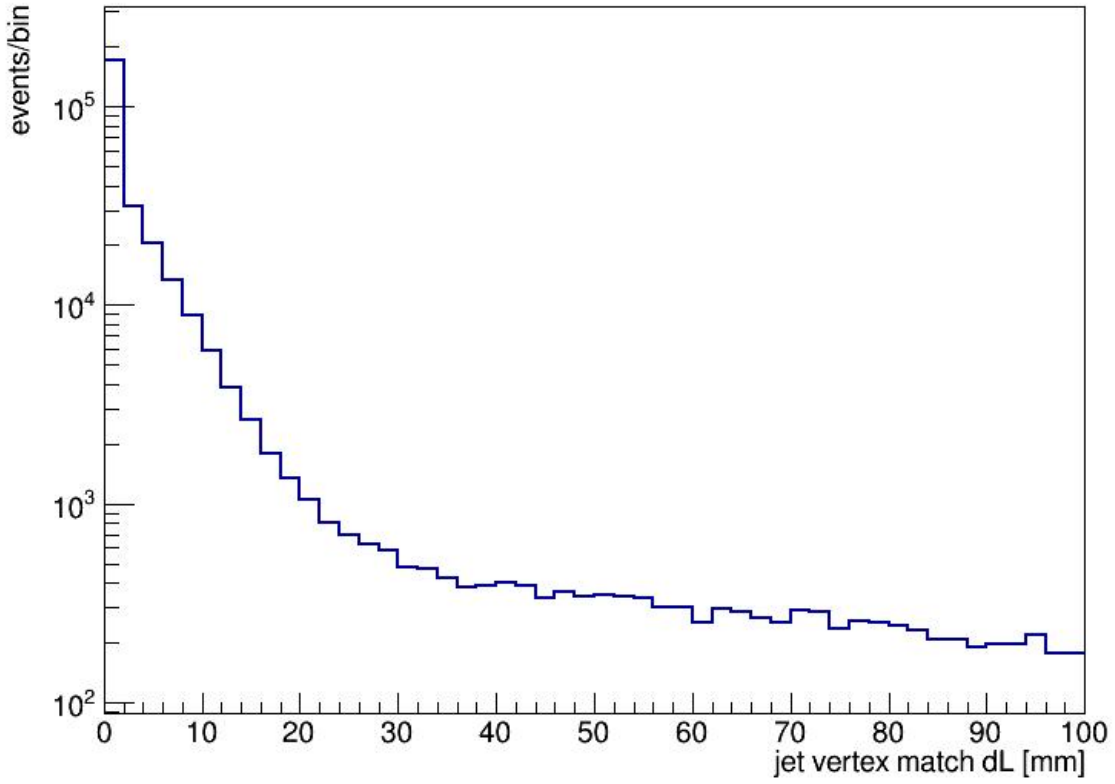


Vertex matching performance

- Criteria: Match vertex by distance (jet $\Delta R < 0.4$)
- Resolution: ~cm



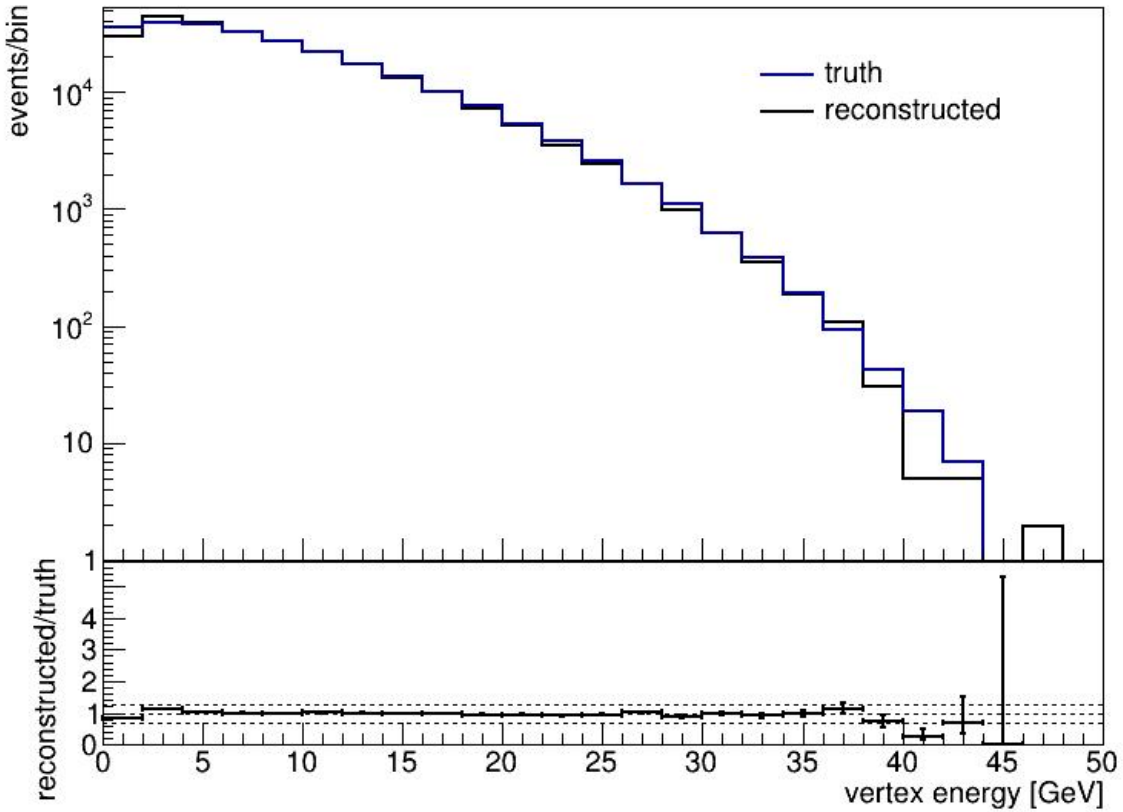
unmatched

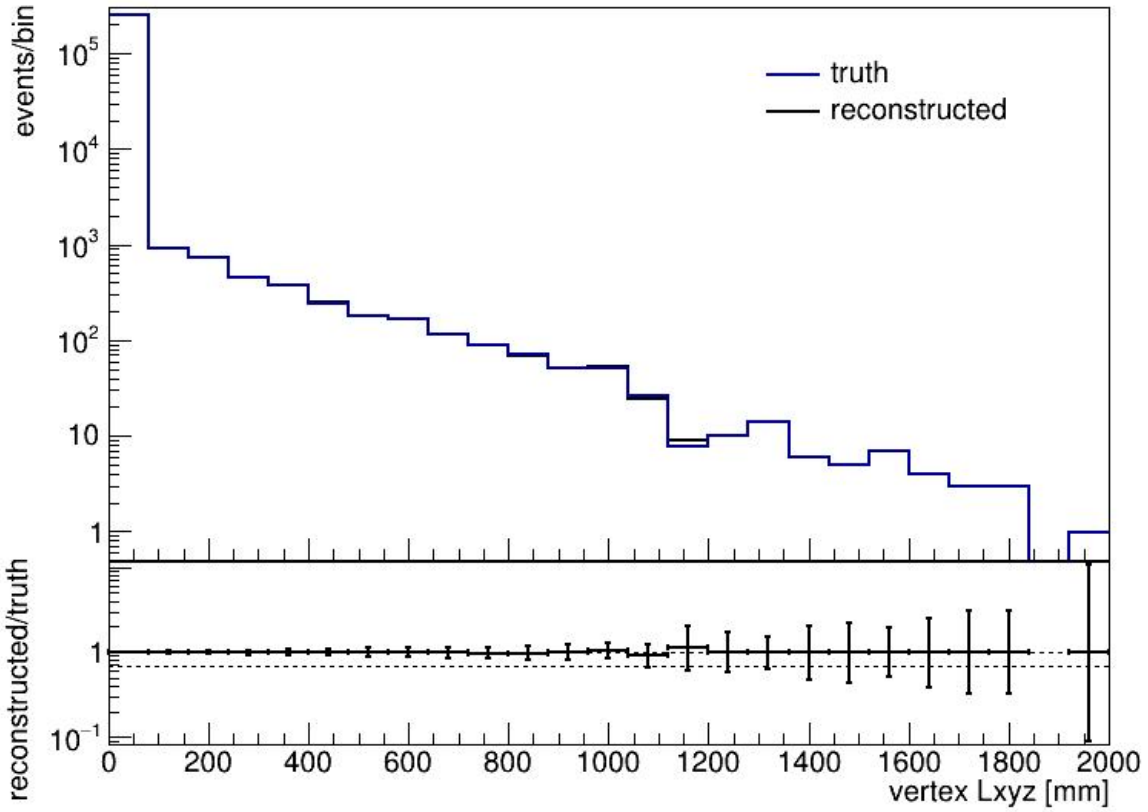


matched

Vertex vars

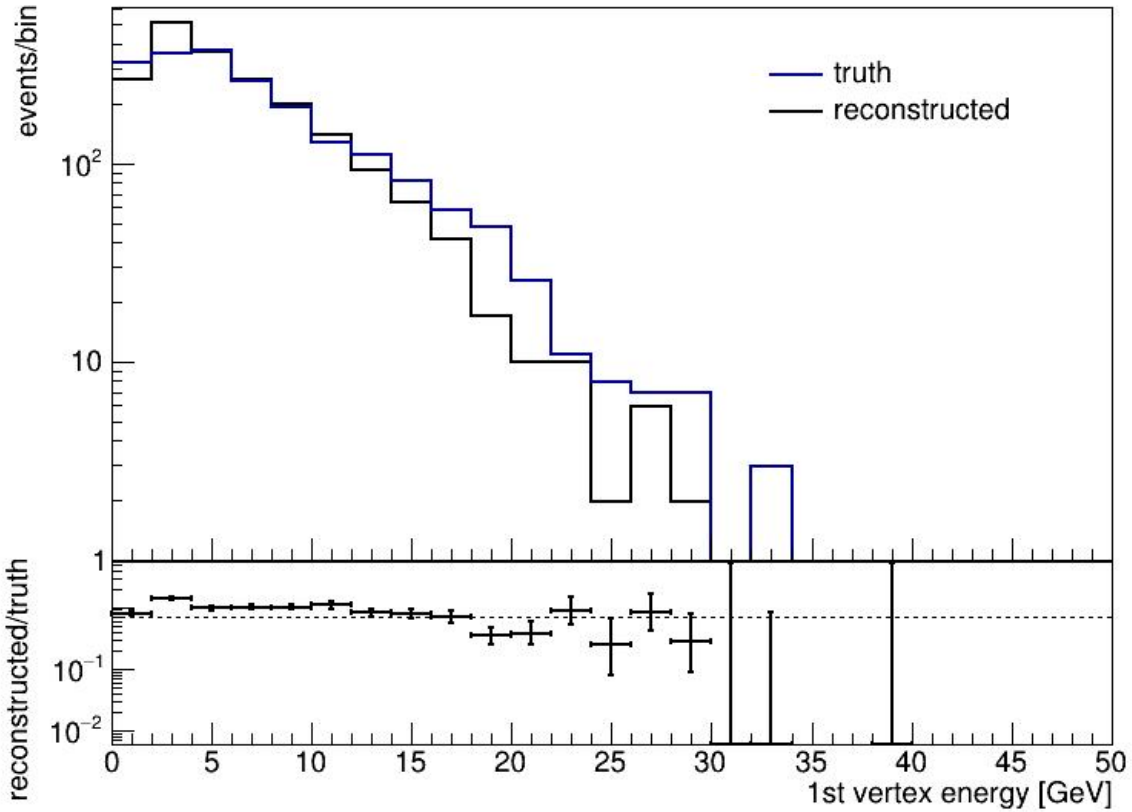
- dL(reco, truth) < 20mm
- Good agreement

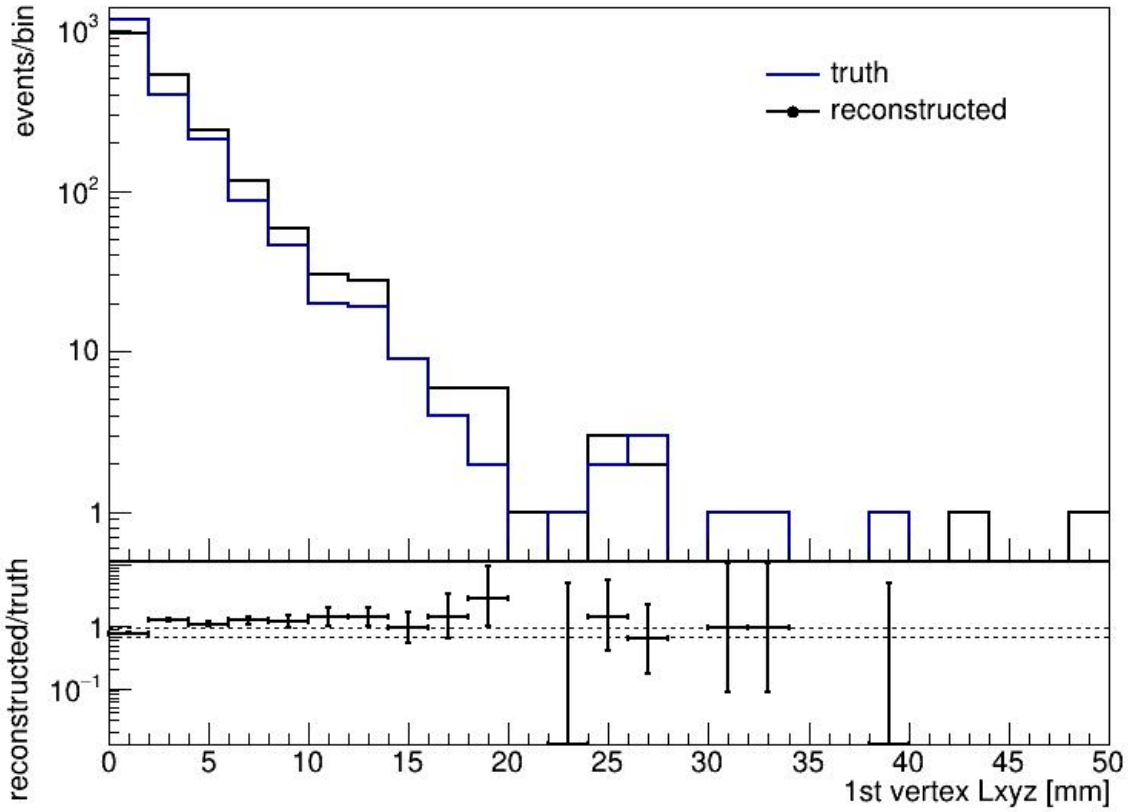




First secondary vertex vars

- dL(reco, truth) < 20mm
- Secondary vertex closest to IP/primary vertex
- Good agreement

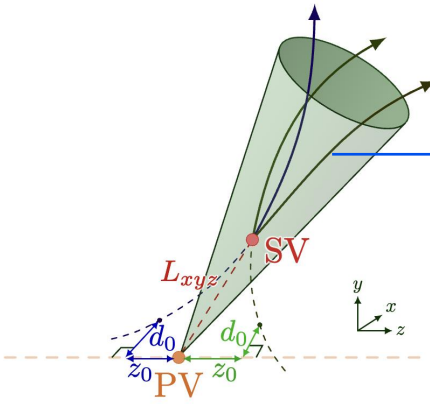




Lepton ID

from changhua

- Basic Idea



track

chi2_dndx
chi2_tof
Eecal
Ehcal
Rhad
p
theta

chi2

output

“e/mu library”

chi2_dnd’
chi2_tof’
Eecal’
Ehcal’
Rhad’

the smaller the chi2, the higher the similarity between the reco particle and the records in the library

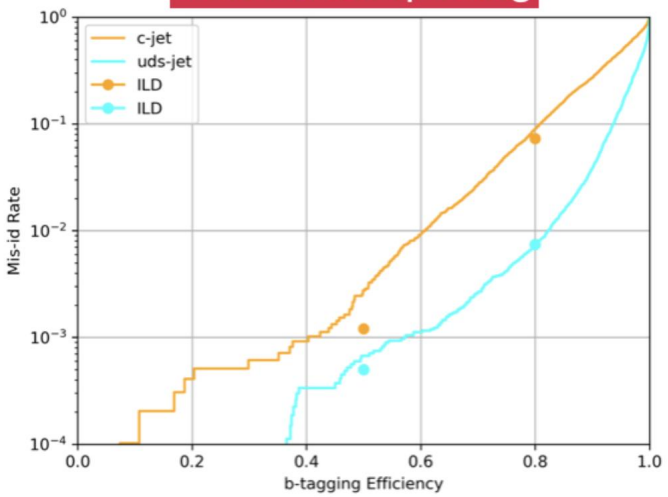
edm4hep::RecDqdxCollection, RecTofCollection, ReconstructedParticleCollection

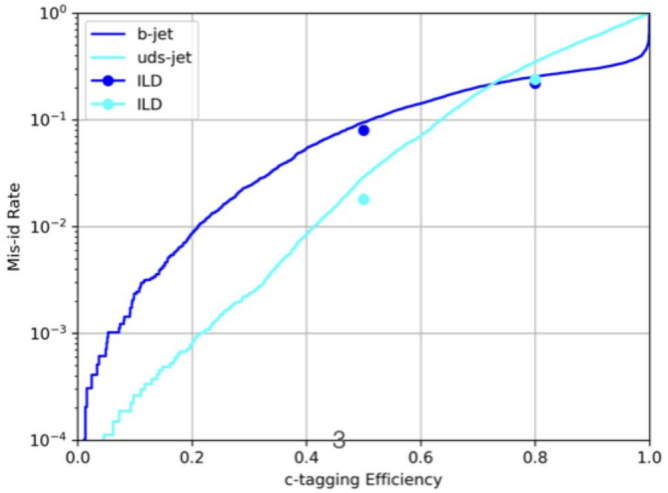
still in process

Njetvtxfrac

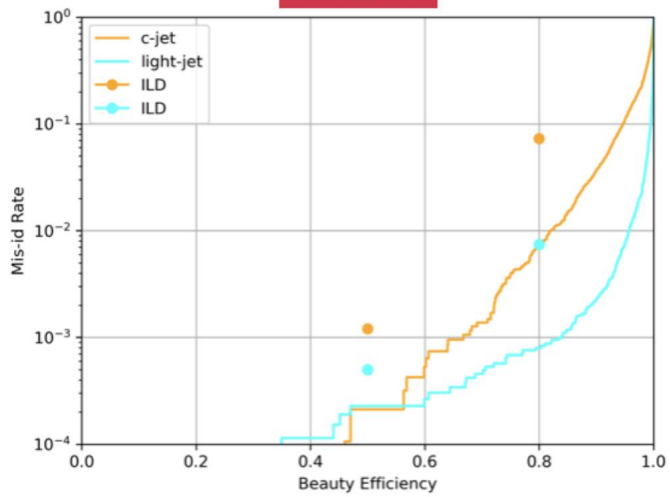
- Jet clustering performed after vertex reconstruction
  - Some tracks connected to the same vertex are assigned to different jets
  - Without vertex splitting, all vertices and corresponding splitting fraction as BDT inputs
  - With vertex splitting, after jet-clustering, split those vertices accordingly, splitting fraction = 1 always

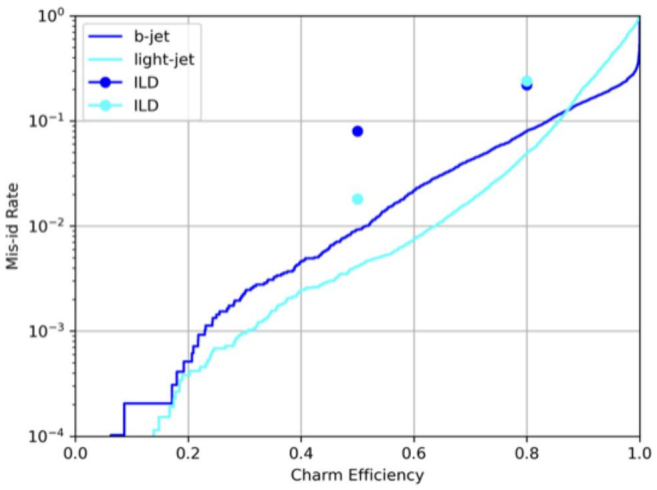
with vertex splitting





without

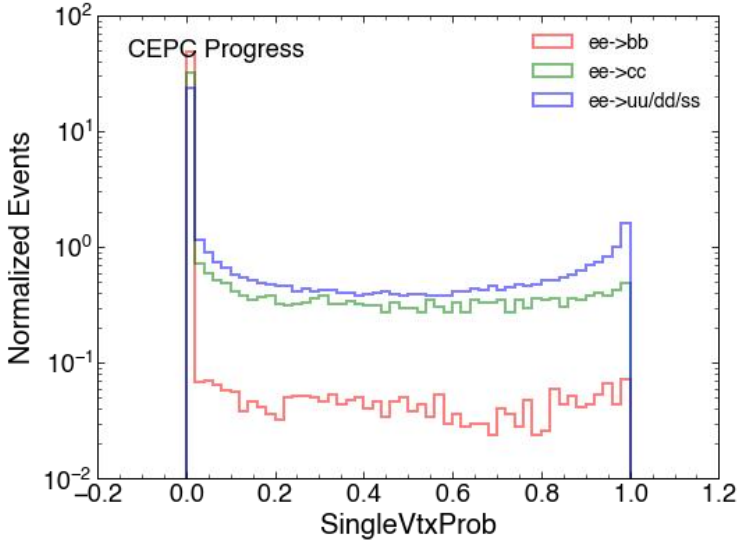


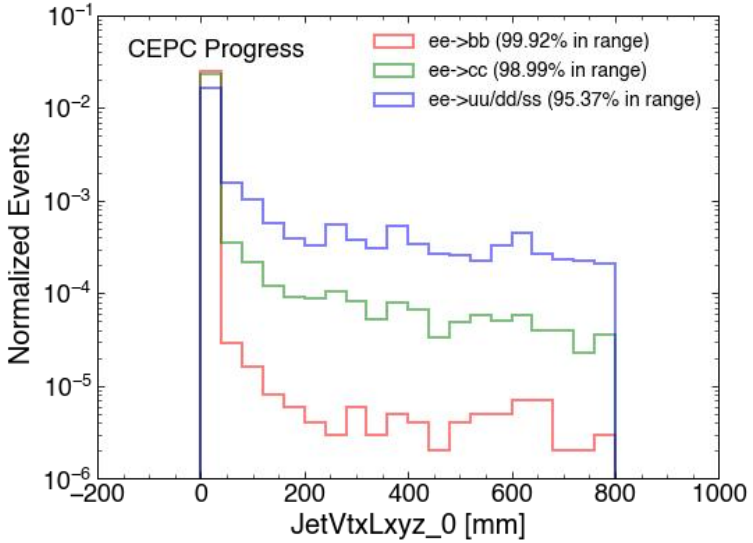


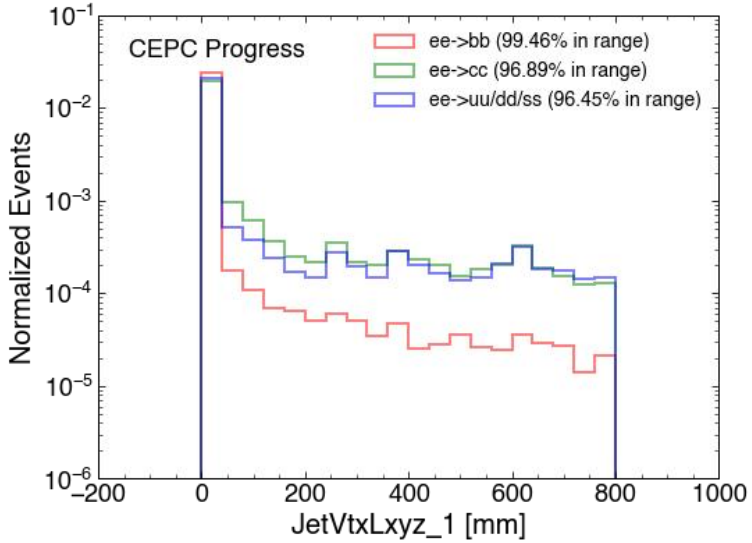
Thank You !

Backup

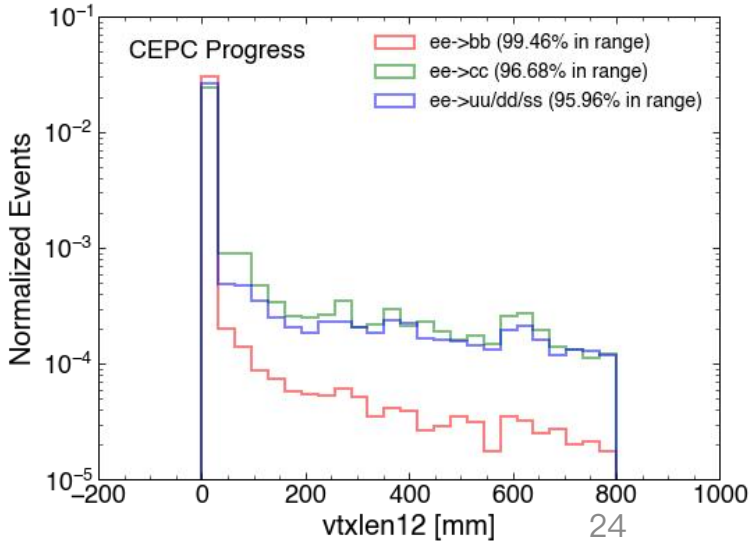
Features in ILD



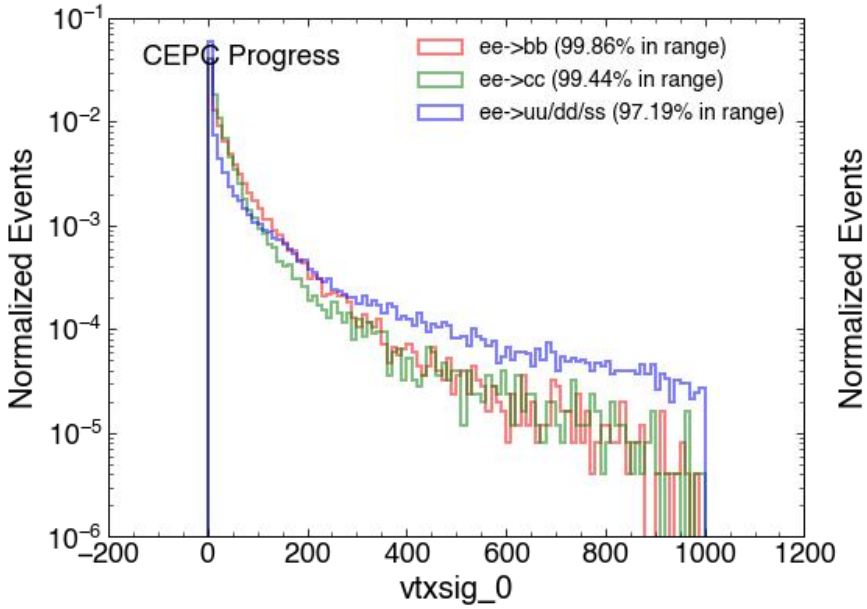


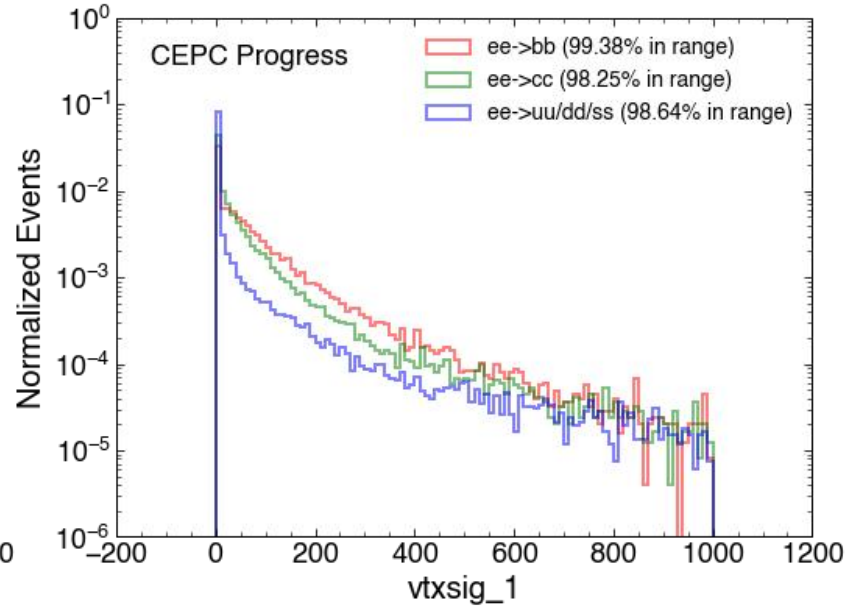


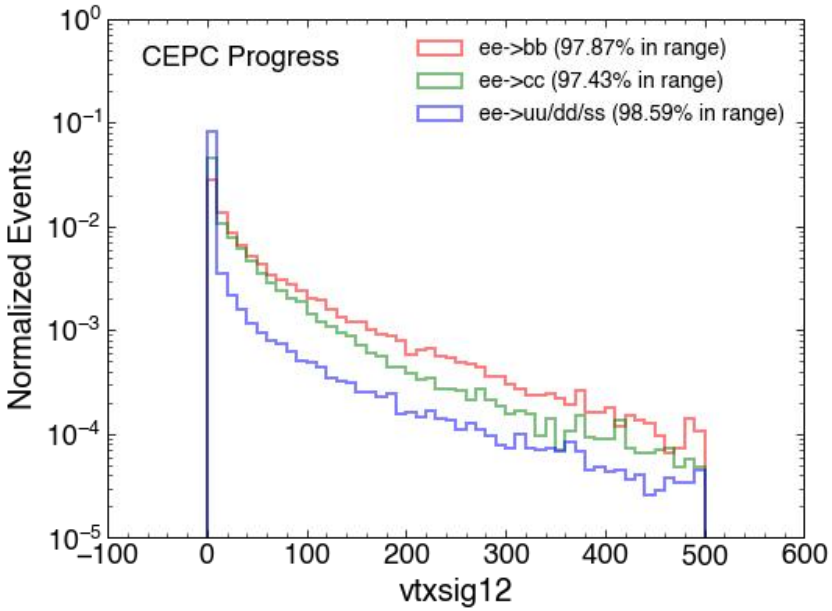
1vtxprob	vertex probability with all tracks associated in vertices combined
vtxlen1	decay length of the first vertex in the jet (zero if no vertex is found)
vtxlen2	decay length of the second vertex in the jet (zero if number of vertex is less than two)
vtxlen12	distance between the first and second vertex (zero if number of vertex is less than two)



Features in ILD

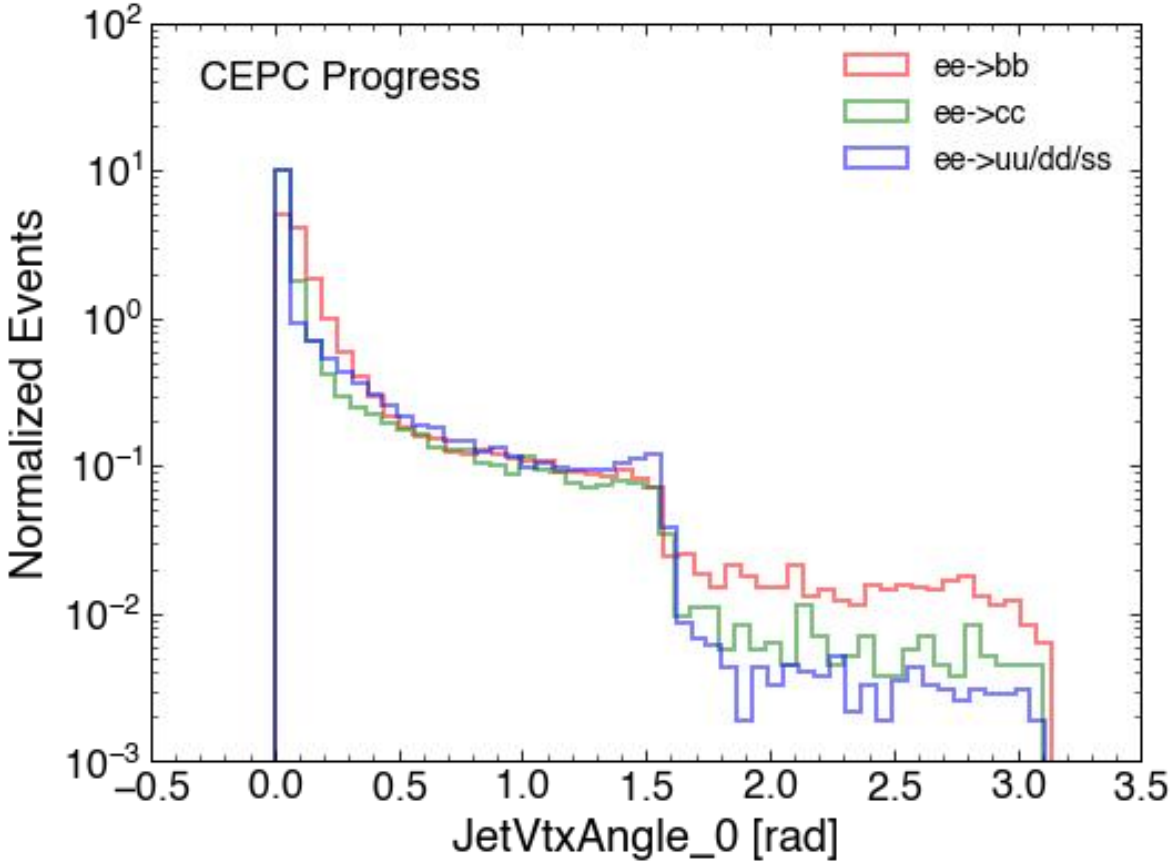


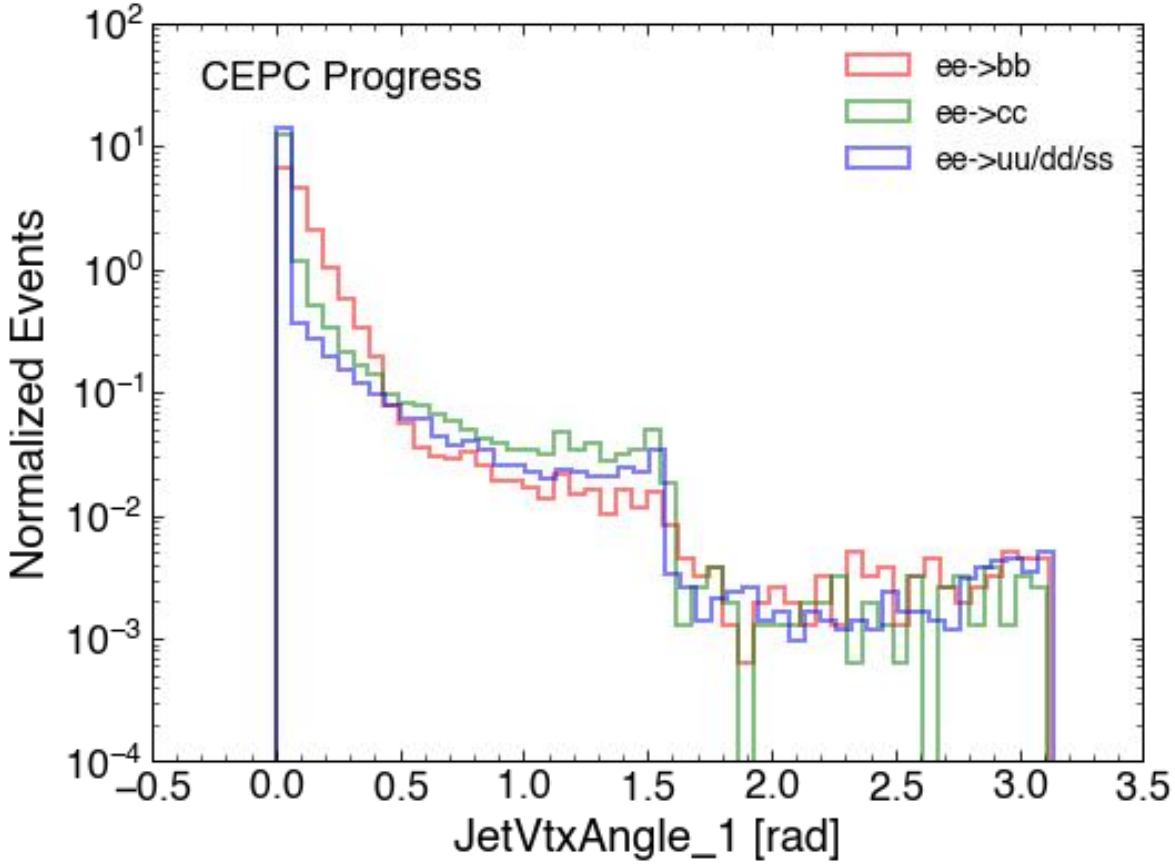




vtxsig1	decay length significance of the first vertex in the jet (zero if no vertex is found)
vtxsig2	decay length significance of the second vertex in the jet (zero if number of vertex is less than two)
vtxsig12	vtxlen12 divided by its error as computed from the sum of the covariance matrix of the first and second vertices, projected along the line connecting the two vertices

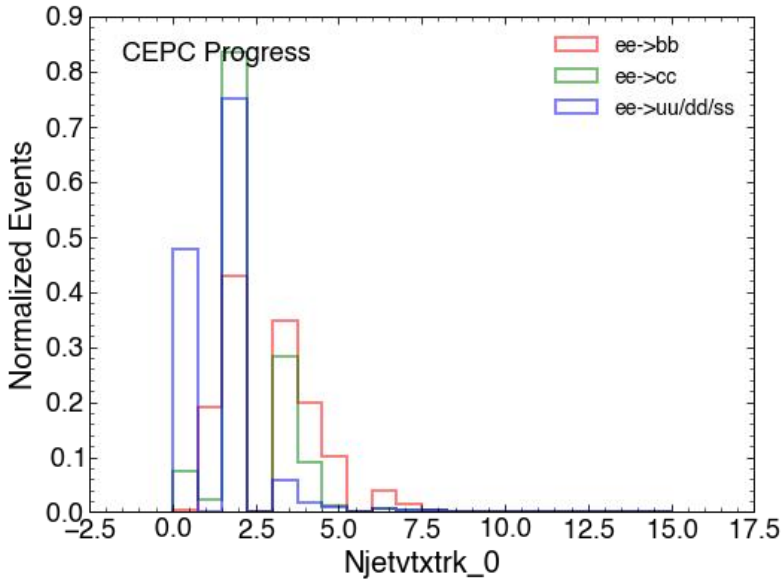
Features in ILD

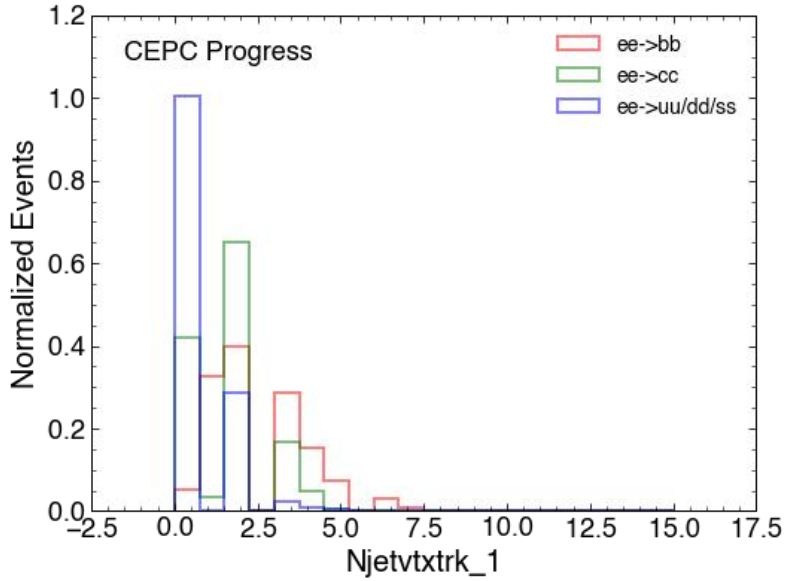


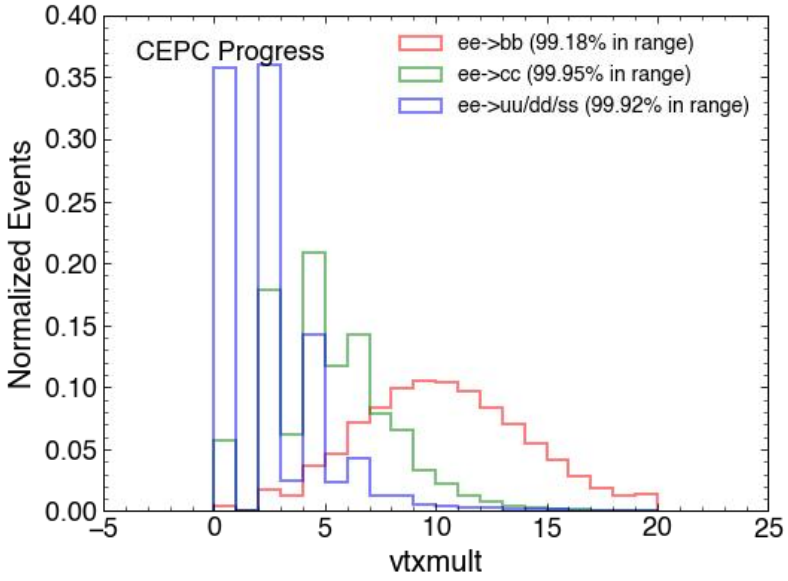


vtxdirang1	the angle between the momentum (computed as a vector sum of track momenta) and the displacement of the first vertex
vtxdirang2	the angle between the momentum (computed as a vector sum of track momenta) and the displacement of the second vertex

Features in ILD

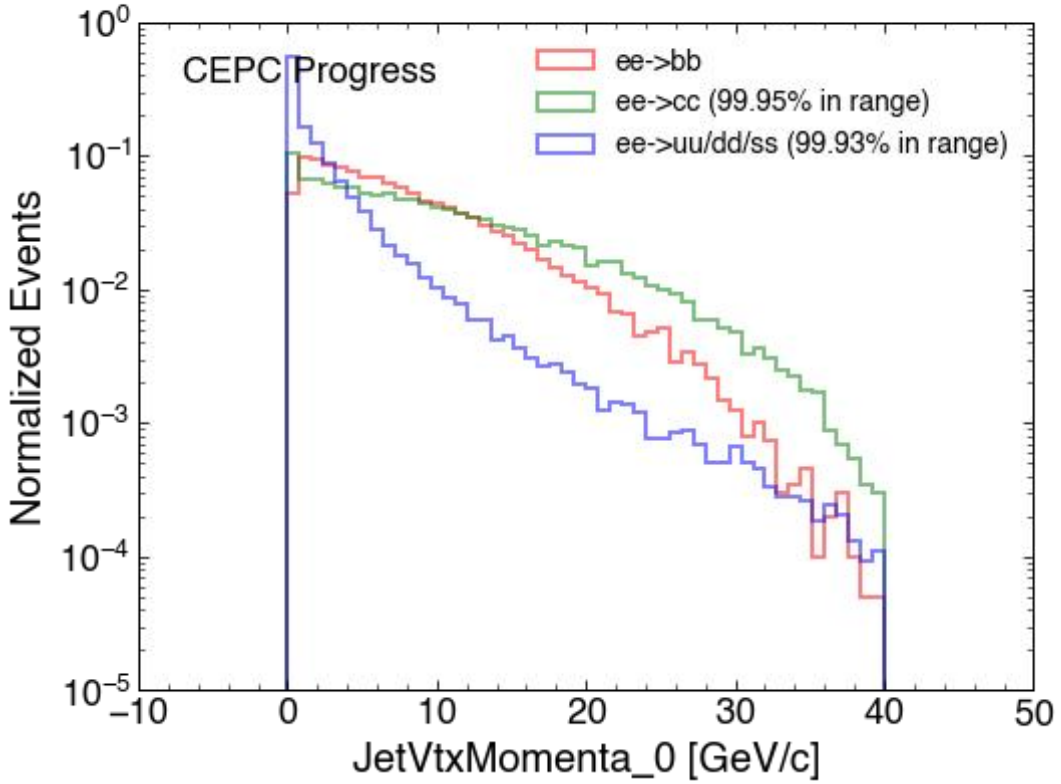


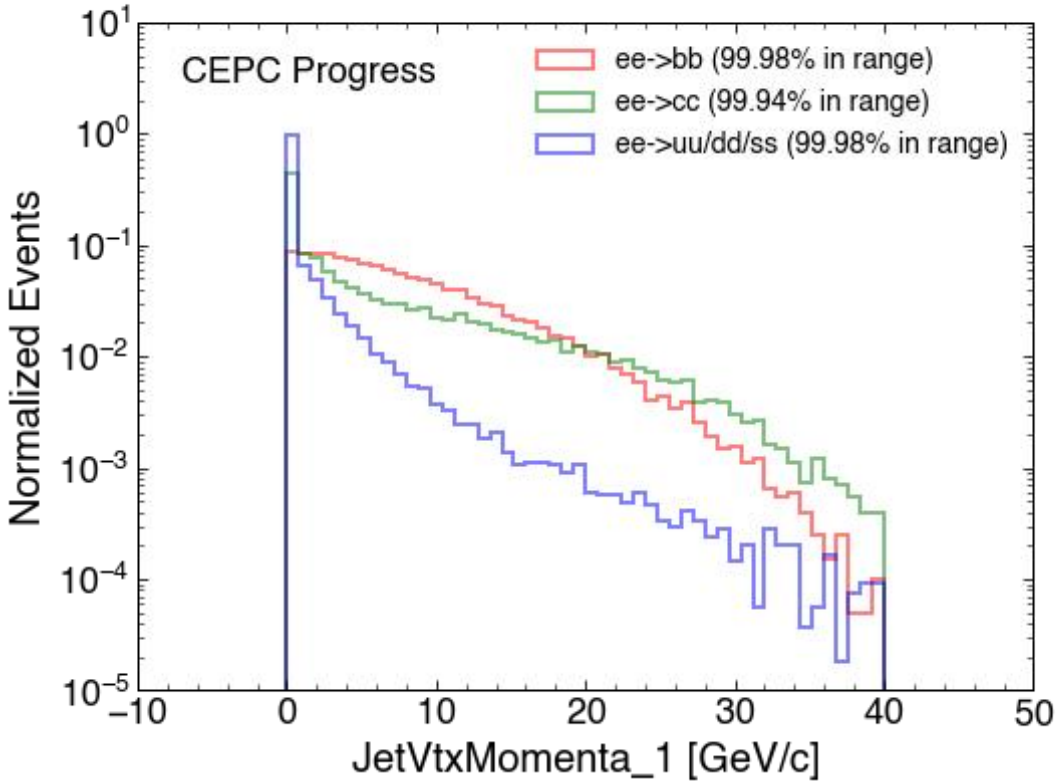




vtxmult1	number of tracks included in the first vertex (zero if no vertex is found)
vtxmult2	number of tracks included in the second vertex (zero if number of vertex is less than two)
vtxmult	number of tracks which are used to form secondary vertices (summed for all vertices)

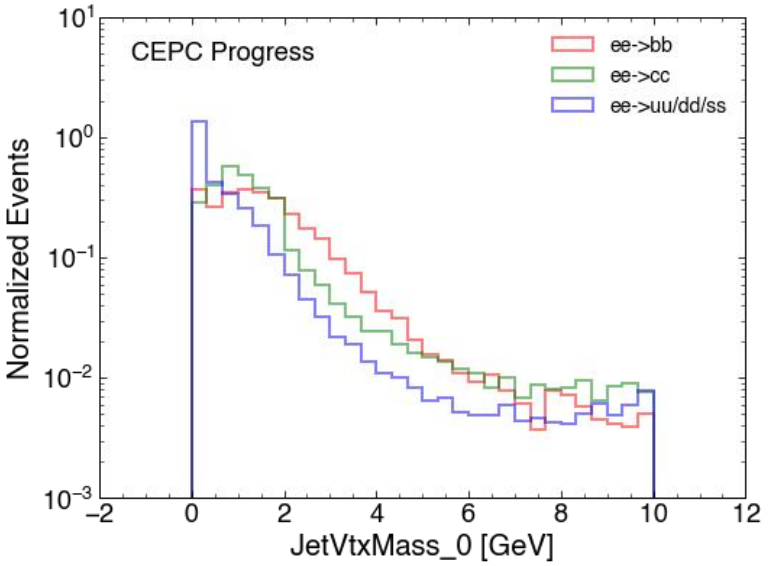
Features in ILD

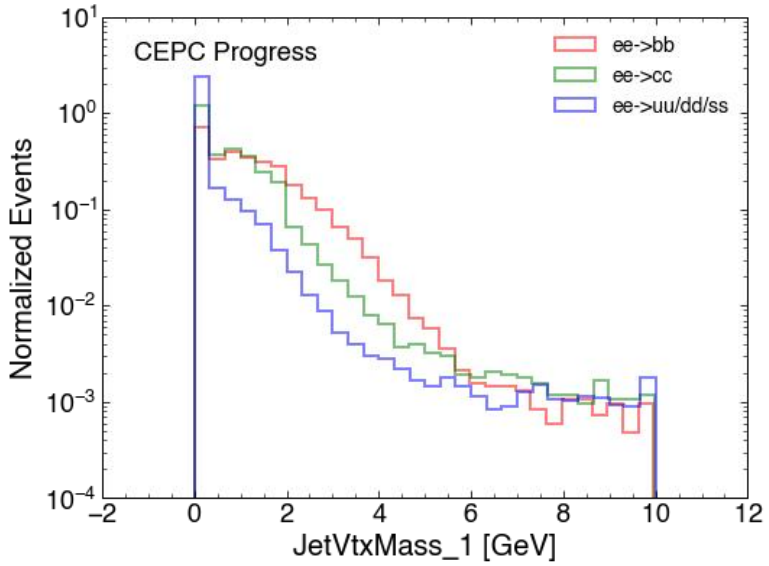


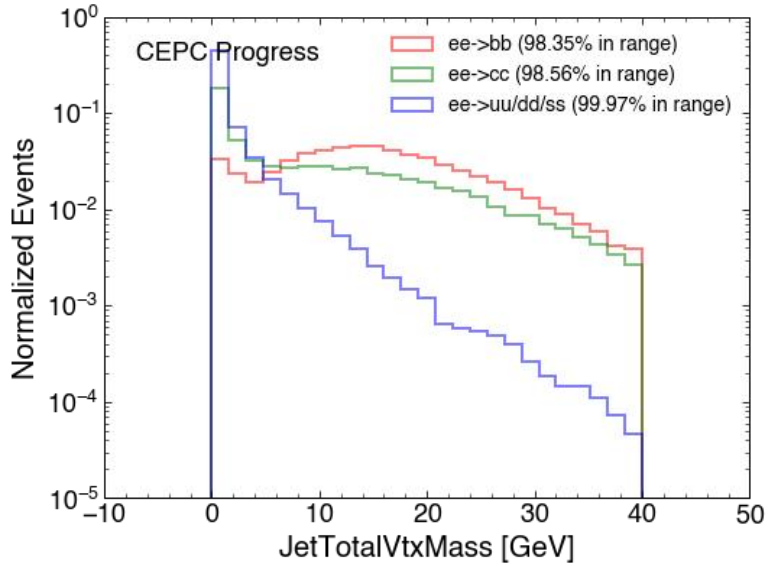


vtxmom1	magnitude of the vector sum of the momenta of all tracks combined into the first vertex
vtxmom2	magnitude of the vector sum of the momenta of all tracks combined into the second vertex

Features in ILD

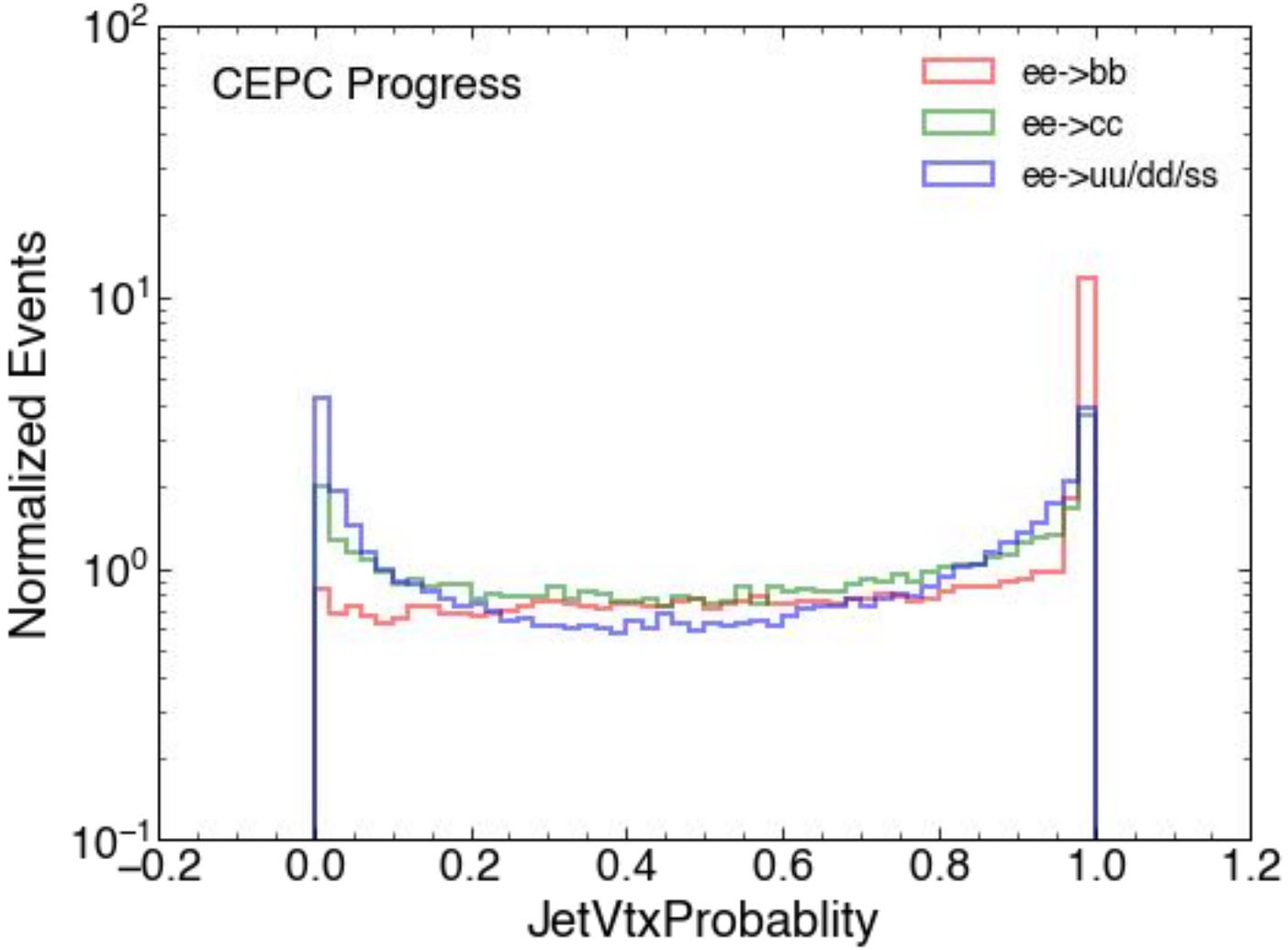






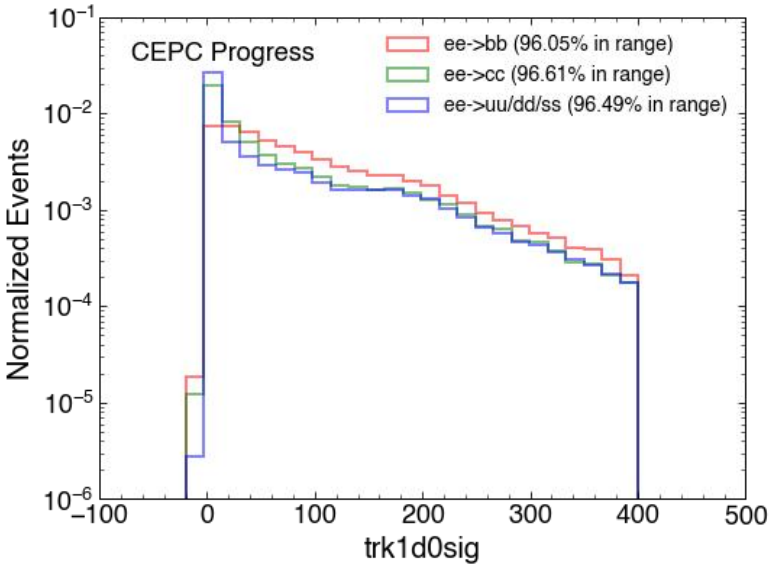
vtxmass1	mass of the first vertex computed from the sum of track four-momenta
vtxmass2	mass of the second vertex computed from the sum of track four-momenta
vtxmass	vertex mass as computed from the sum of four momenta of all tracks forming secondary vertices

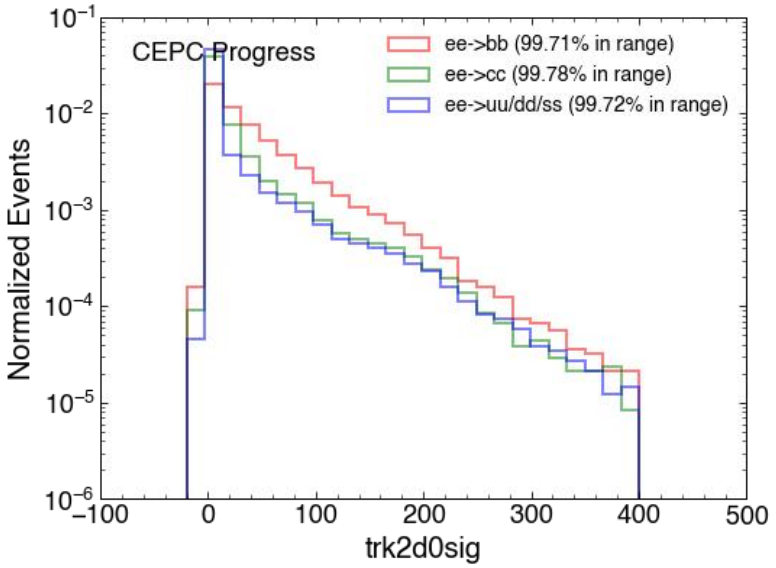
Features in ILD

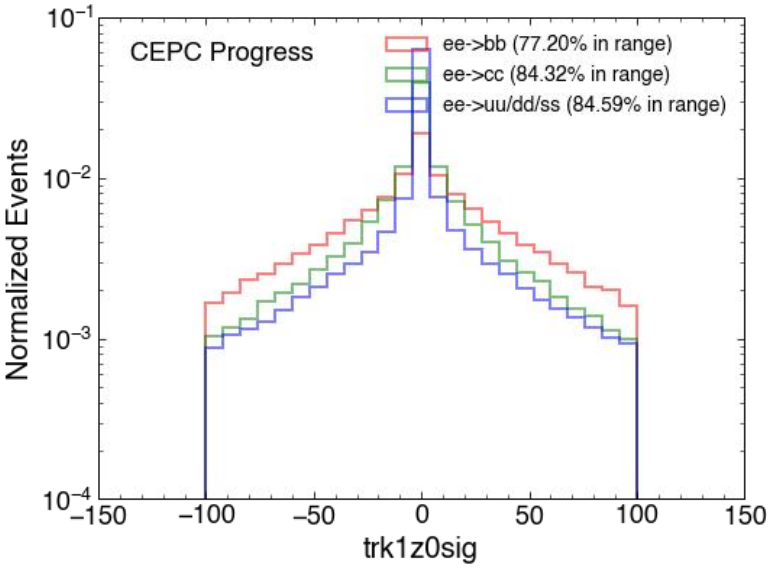


vtxprob	vertex probability; for multiple vertices, the probability P is computed as 1-P = (1-P1)(1-P2)...(1-PN)

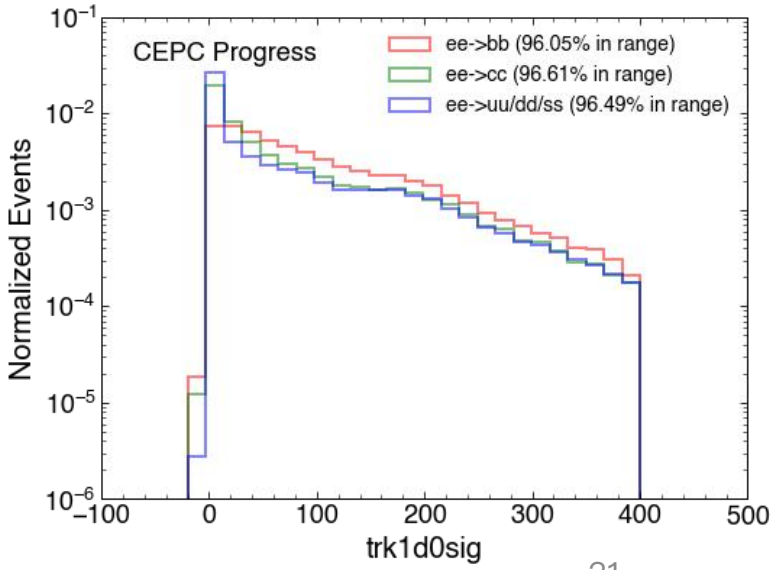
Features in ILD



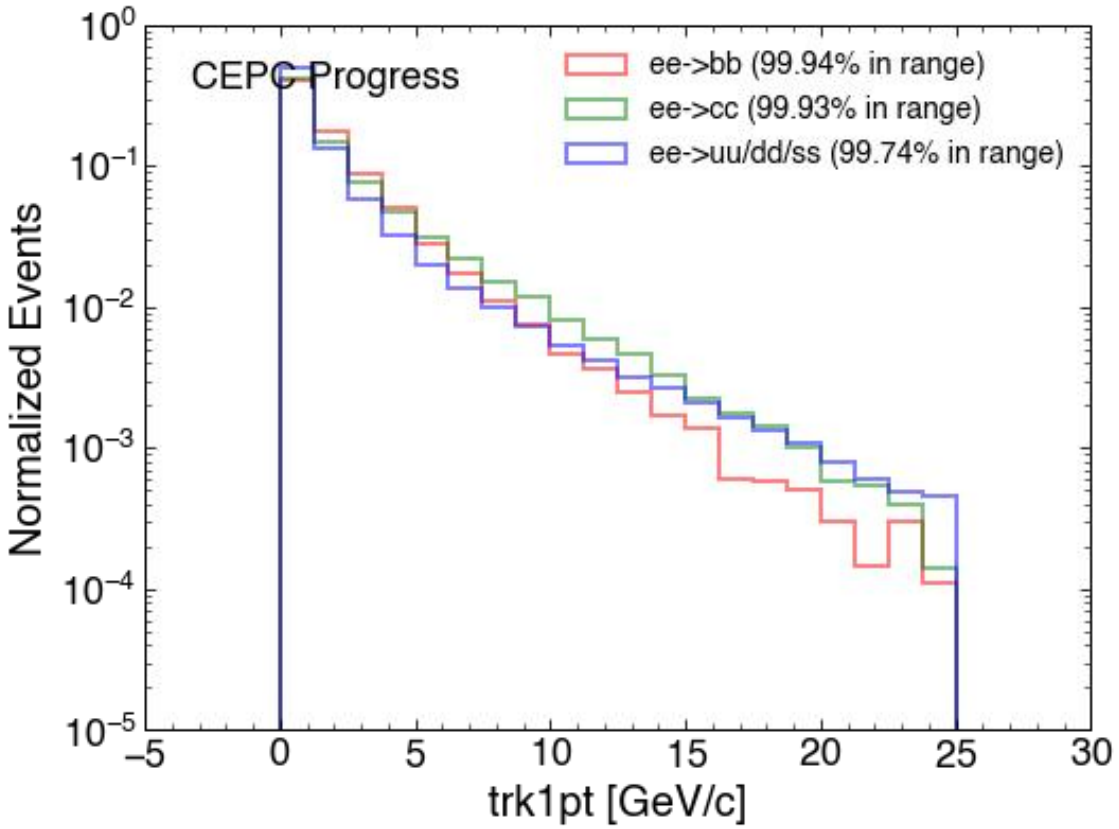


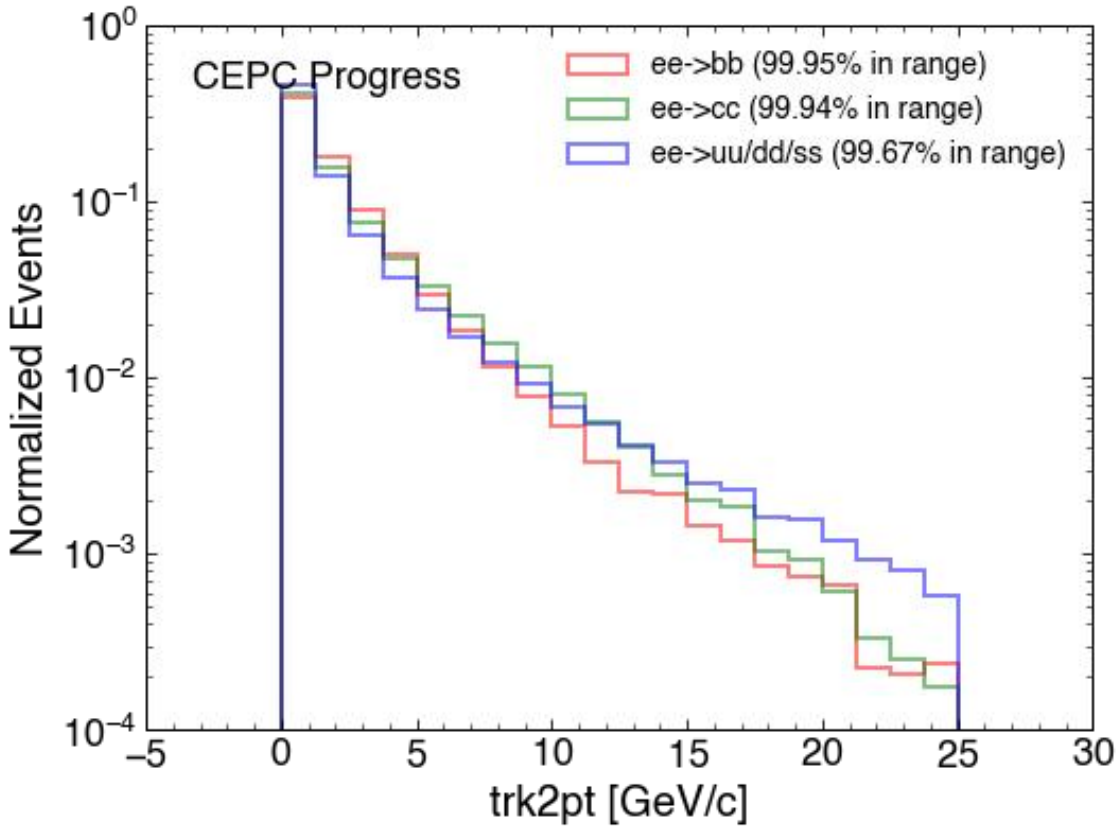


trk1d0sig	d0 significance of track with highest d0 significance
trk2d0sig	d0 significance of track with second highest d0 significance
trk1z0sig	z0 significance of track with highest d0 significance
trk2z0sig	z0 significance of track with second highest d0 significance



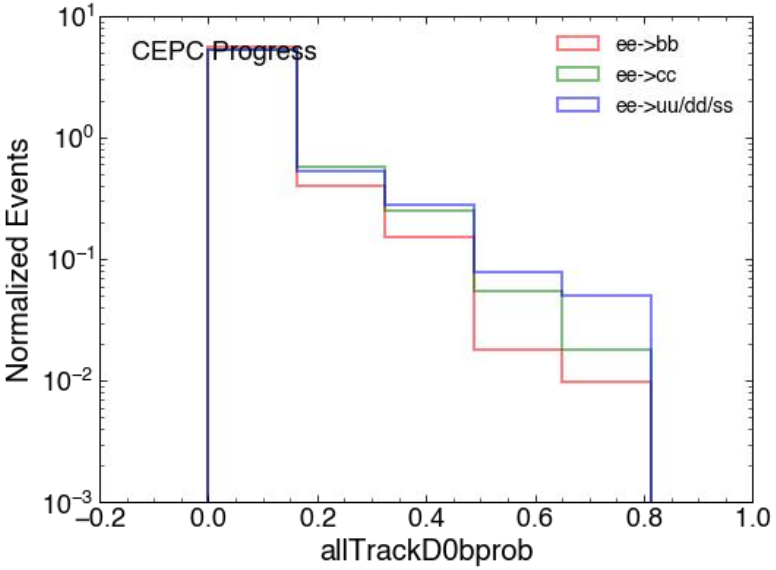
Features in ILD

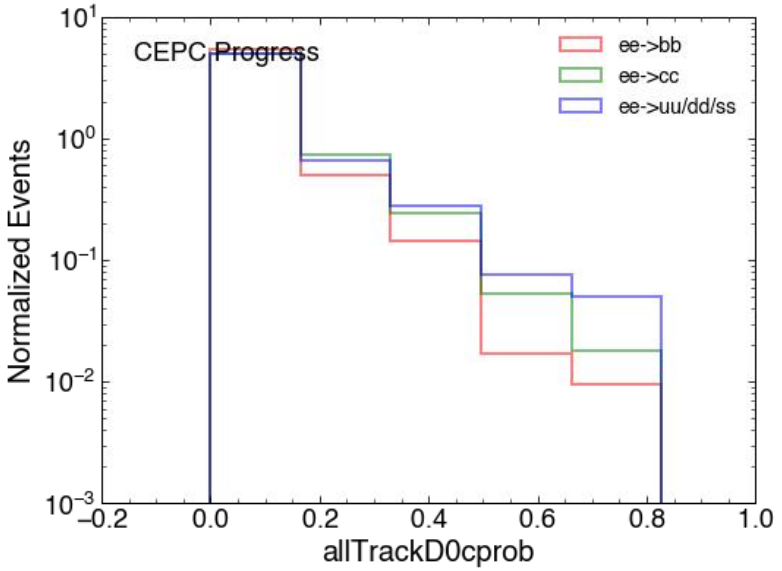


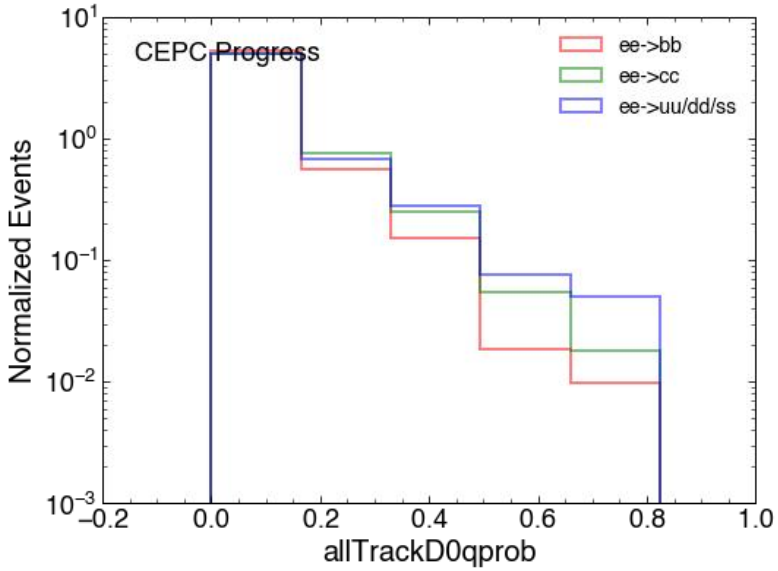


trk1pt	transverse momentum of track with highest d0 significance
trk2pt	transverse momentum of track with second highest d0 significance

Features in ILD

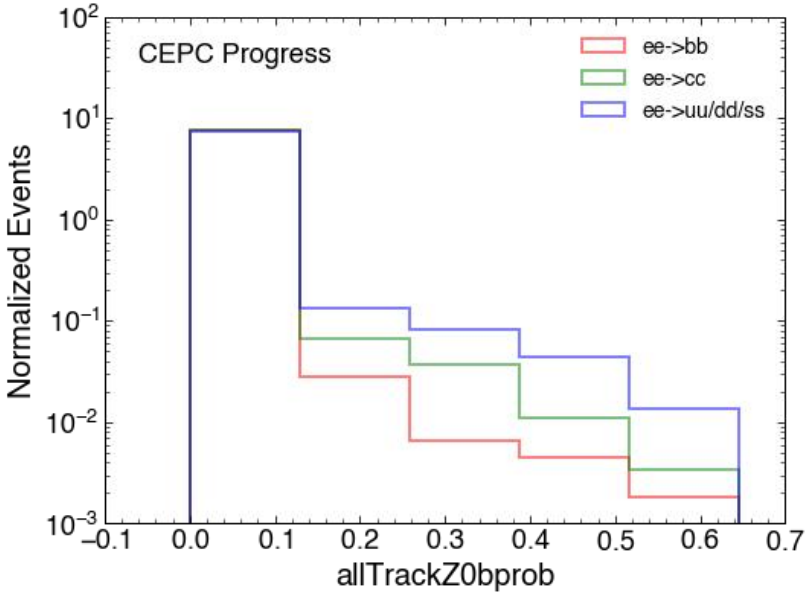


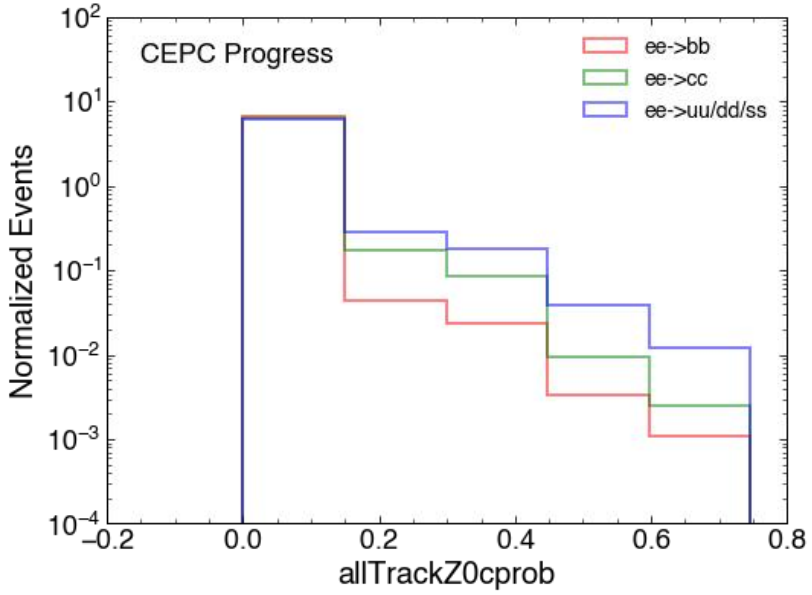


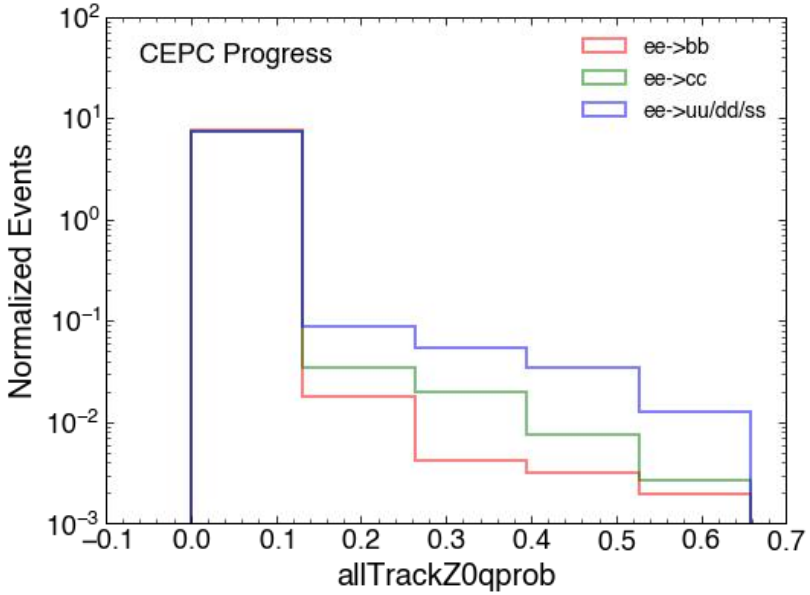


d0bprob	product of b-quark probabilities of d0 values for all tracks, using b/c/q d0 distributions
d0cprob	product of c-quark probabilities of d0 values for all tracks, using b/c/q d0 distributions
d0qprob	product of q-quark probabilities of d0 values for all tracks, using b/c/q d0 distributions

Features in ILD

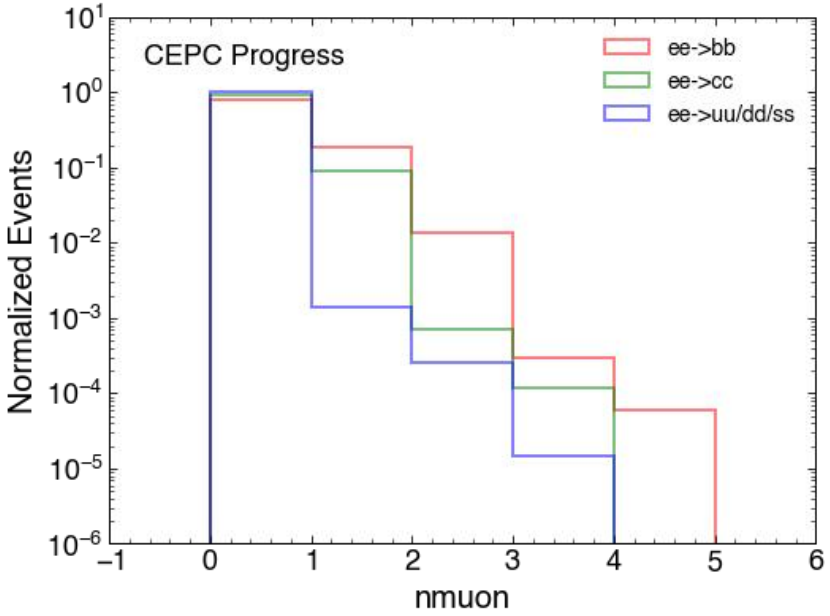


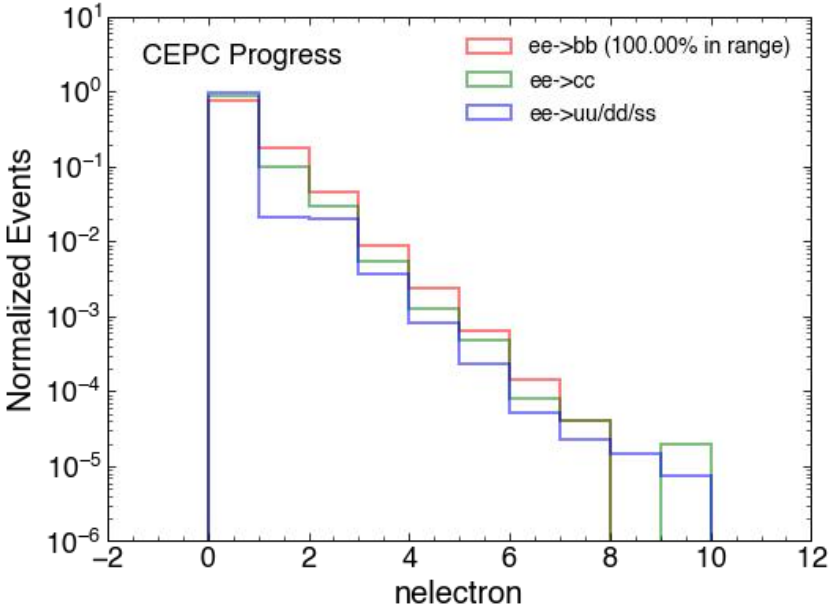


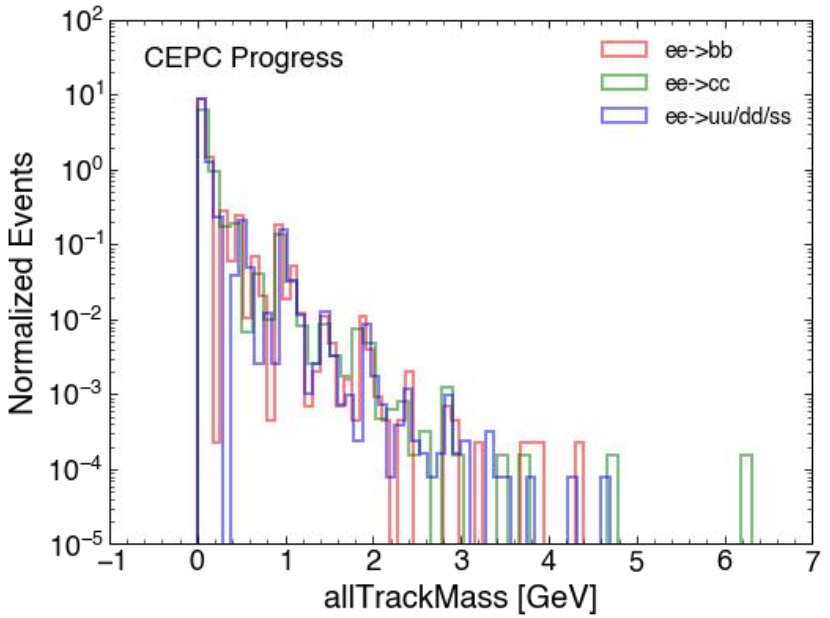


z0bprob	product of b-quark probabilities of z0 values for all tracks, using b/c/q z0 distributions
z0cprob	product of c-quark probabilities of z0 values for all tracks, using b/c/q z0 distributions
z0qprob	product of q-quark probabilities of z0 values for all tracks, using b/c/q z0 distributions

Features in ILD

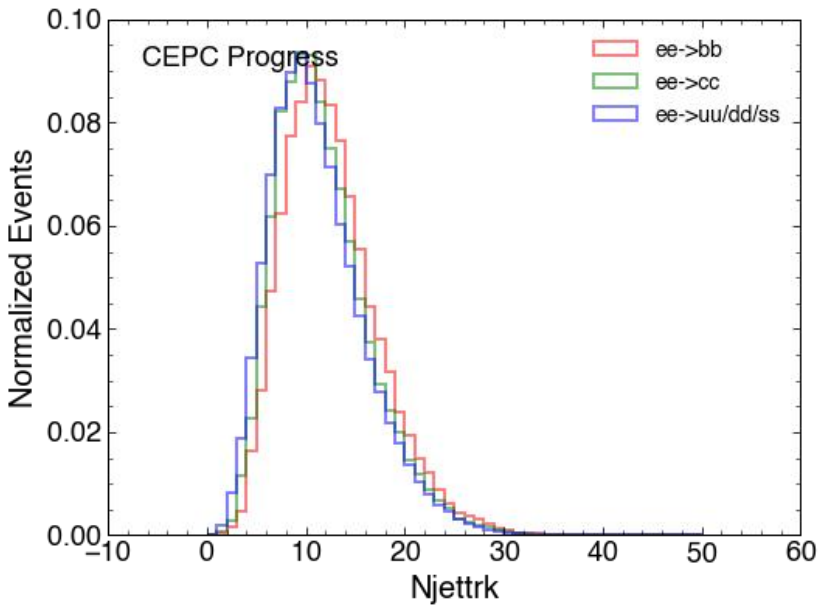


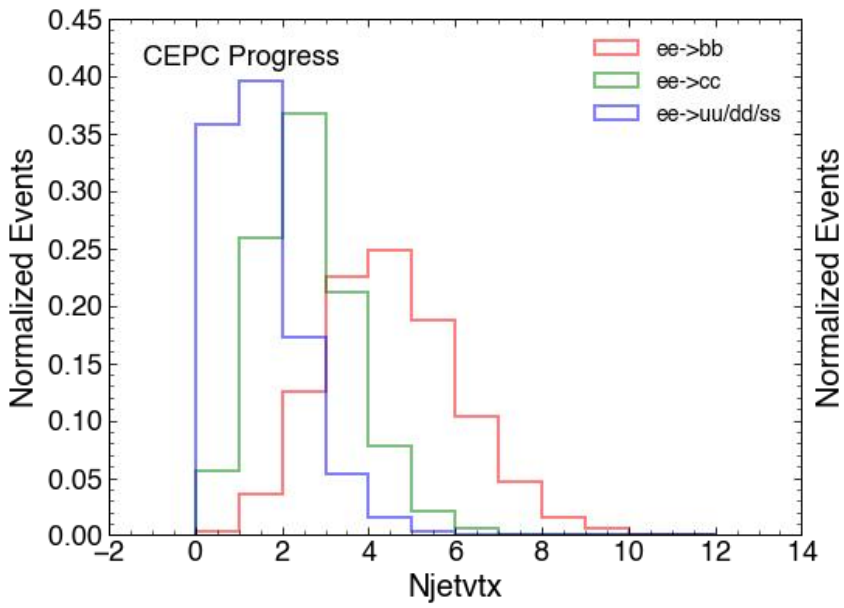


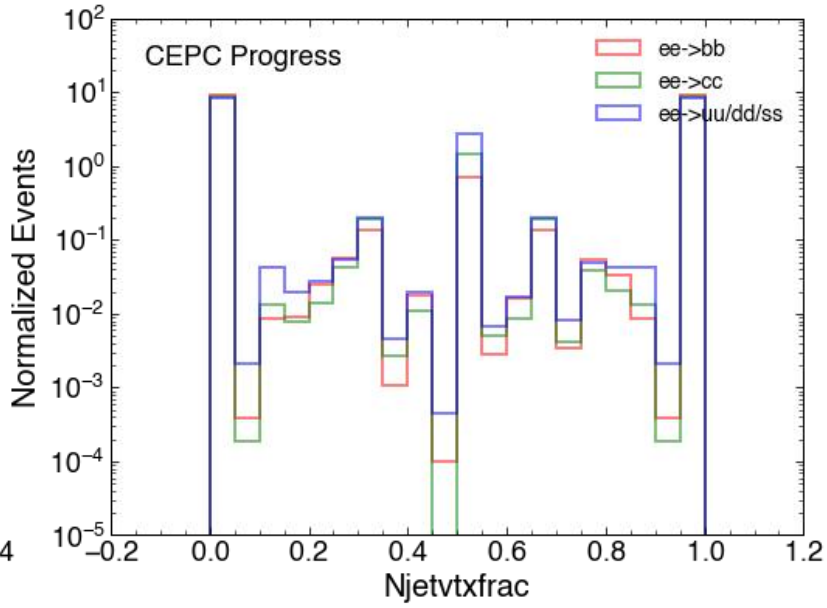


nmuon	number of identified muons
nelectron	number of identified electrons
trkmass	mass of all tracks exceeding 5 sigma significance in d0/z0 values

Features added

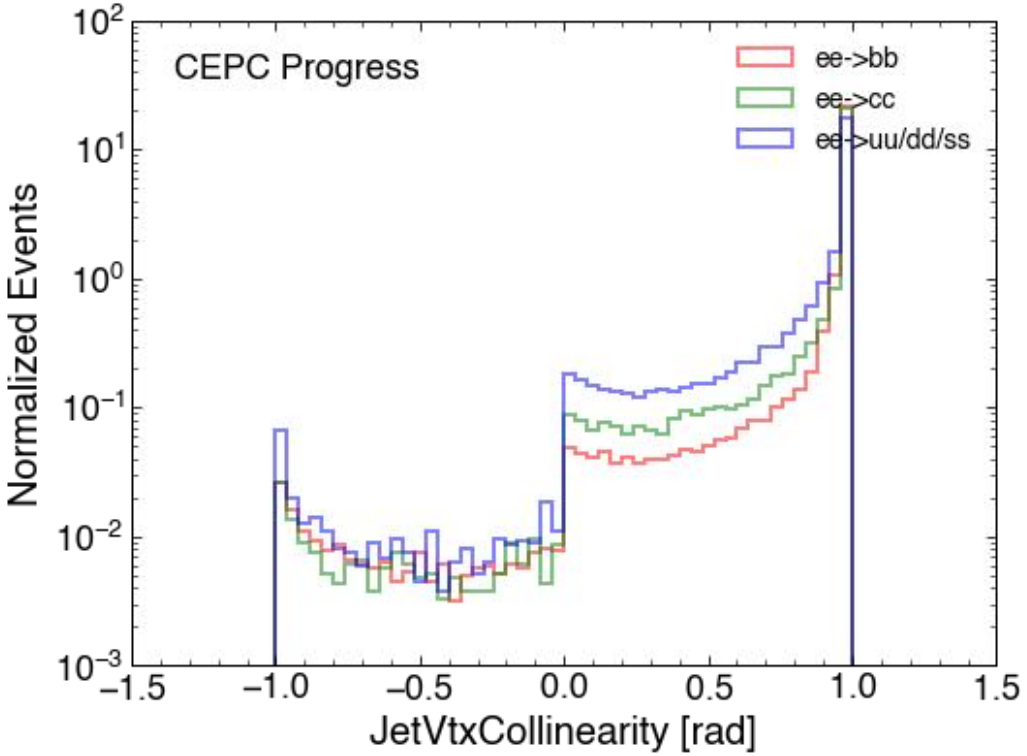


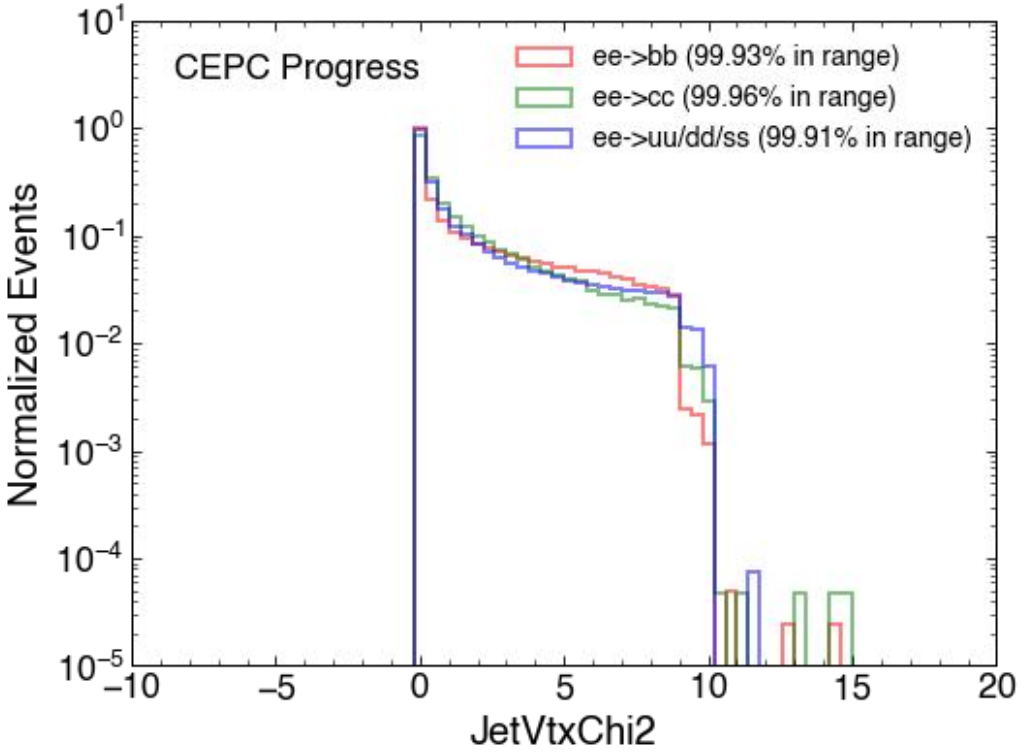




Njettrk: the number of tracks in the jet
Njetvtx: the number of secondary vertexes in the jet
Njetvtxfrac: the fraction of a given secondary vertex's tracks that fall within that jet's region.

Features added

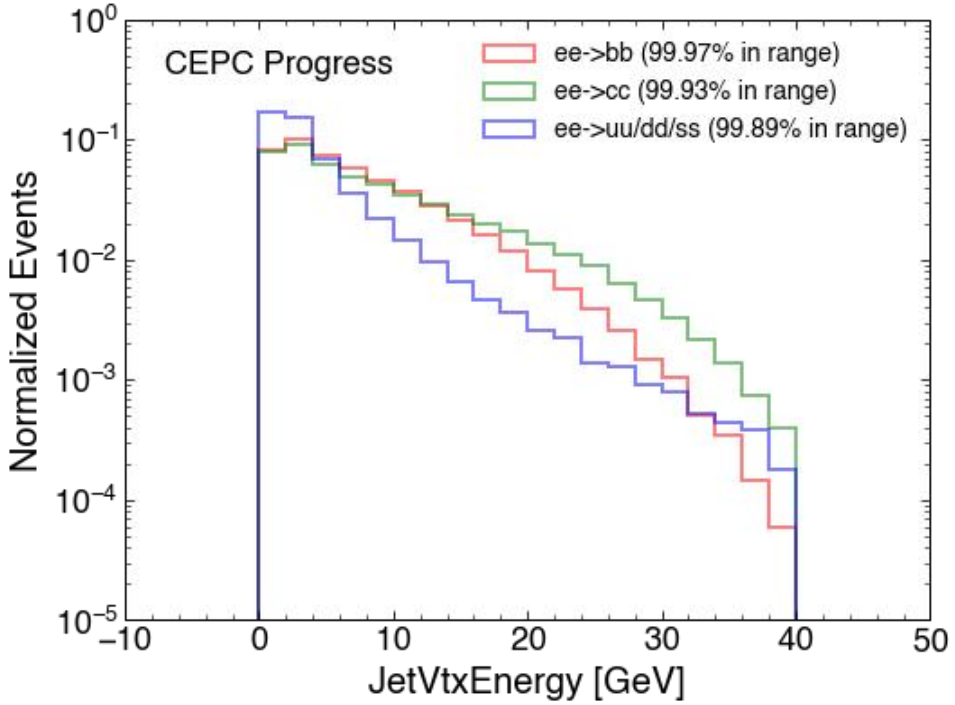


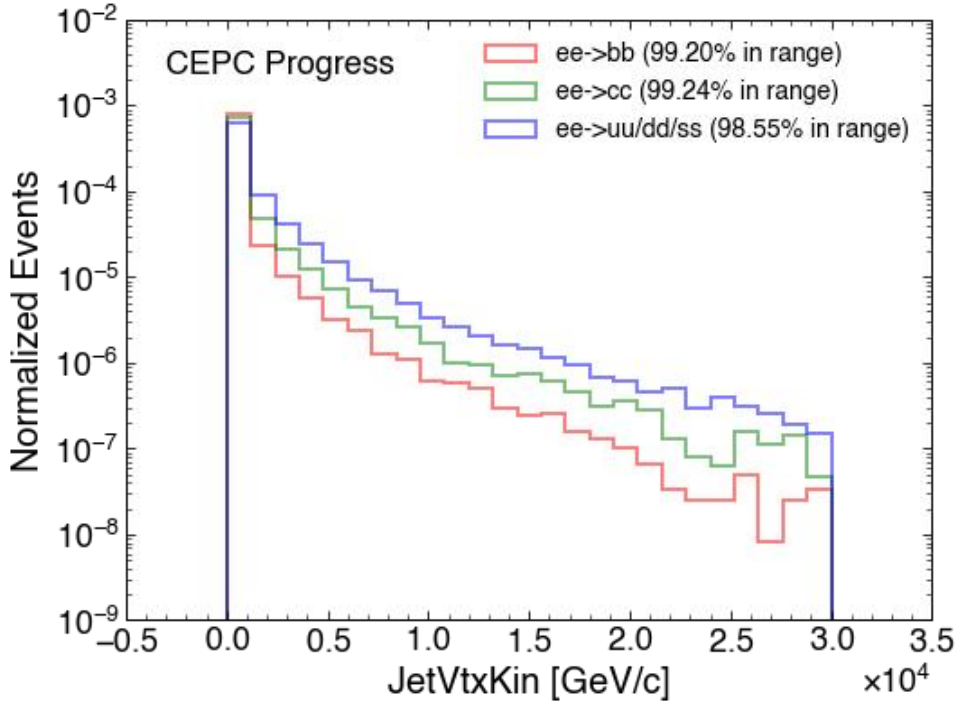


JetVtxCollinearity: Collinearity between the secondary vertex position and the sum of momenta of associated tracks

JetVtxChi2: the χ2 of the secondary vertices

Features added





JetVtxEnergy: the sum of the tracks' energy from the secondary vertex
JetVtxKin: the dot product of the tracks' momentum with the secondary vertex displacement vector.

Content