LCFIVertex+ : flavor tagging for linear colliders

Tomohiko Tanabe, Taikan Suehara, Satoru Yamashita

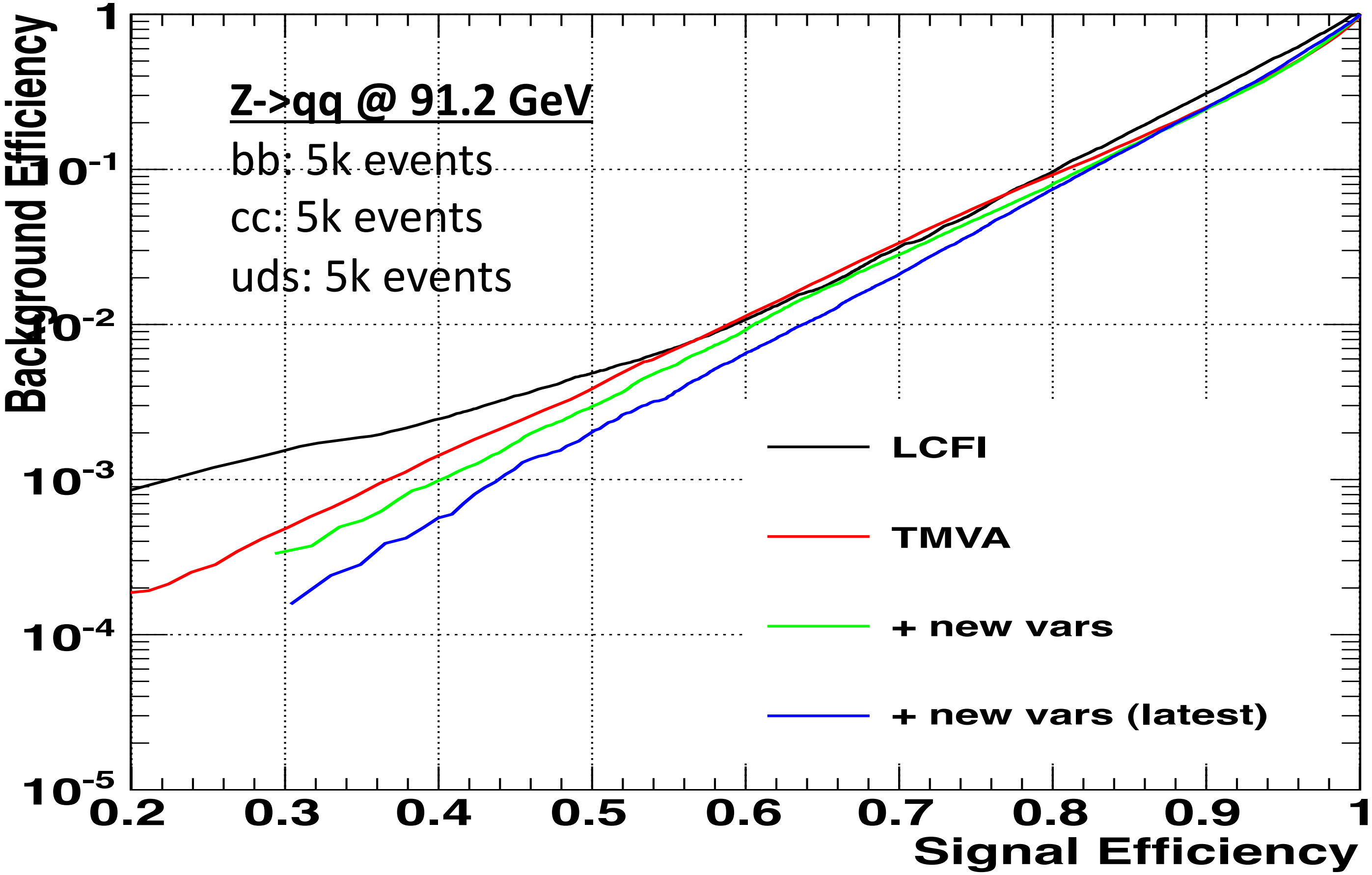
ICEPP, The Univ. of Tokyo

September 28, 2011

LCWS 11, Granada, Spain

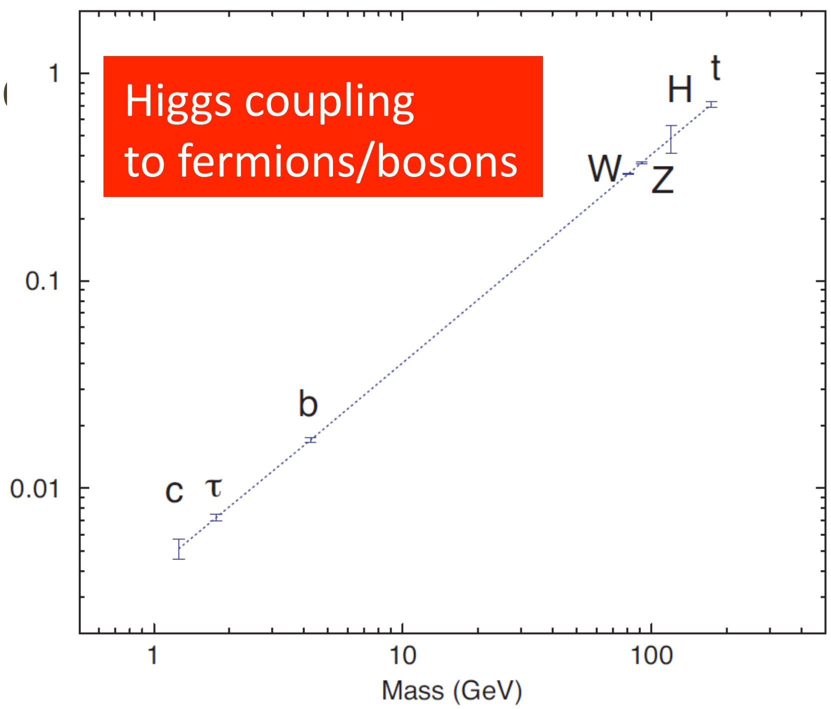
topics

- introduction of software framework
- improvements in vertex finding, jet clustering, flavor tagging

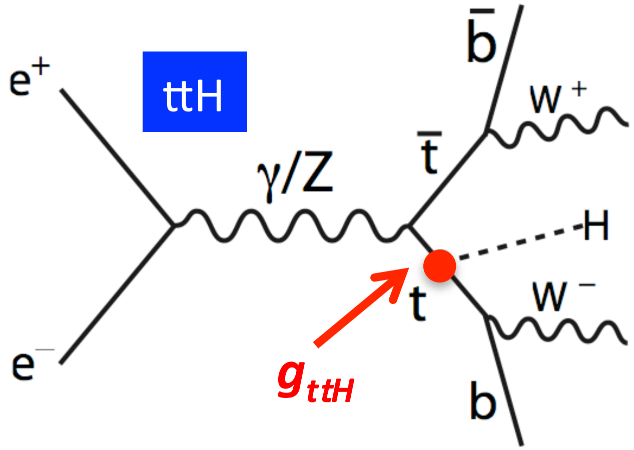


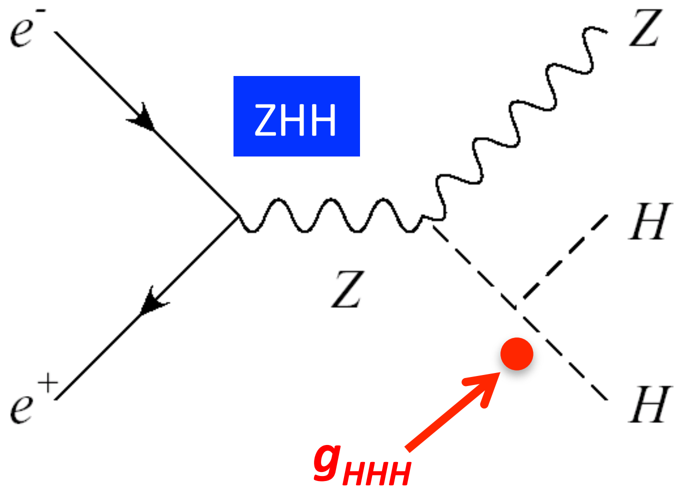
motivation

- Many important physics process
have multiple heavy flavor jets
  - Higgs BF : H -> bb, H -> cc
  - Higgs self-coupling : ZHH -> qqbbb
  - top-Yukawa : ttH -> bWbWbb
  - top physics : tt -> bWbW



Higgs coupling to fermions/bosons

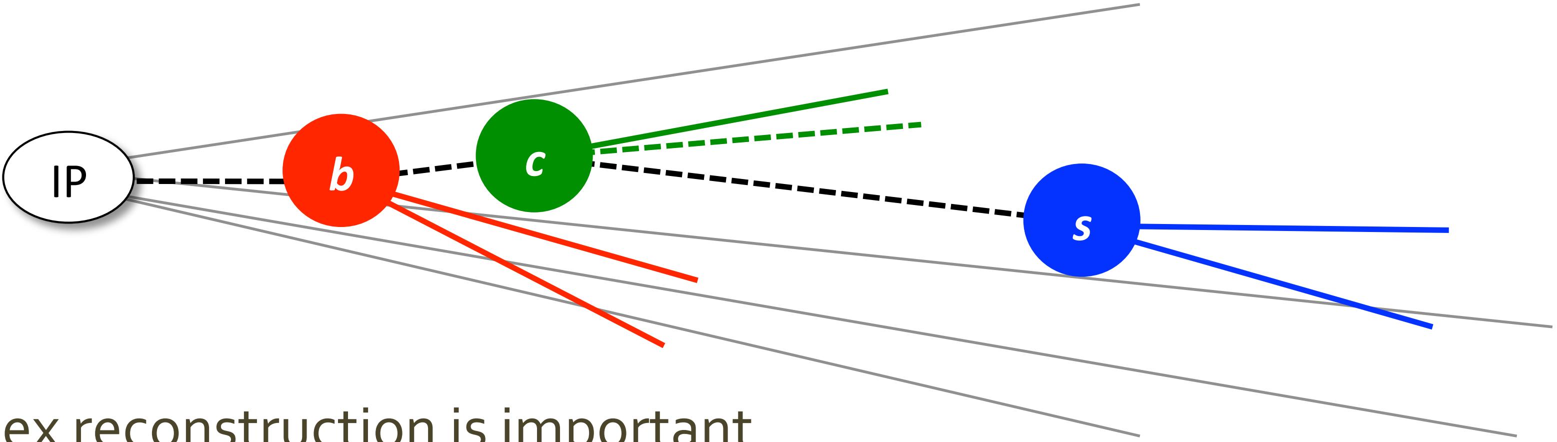




Z→qq (70%)
ll (30%)
W→qq (65%)
lν (35%)
H→bb (65%)
(m_H=120GeV)

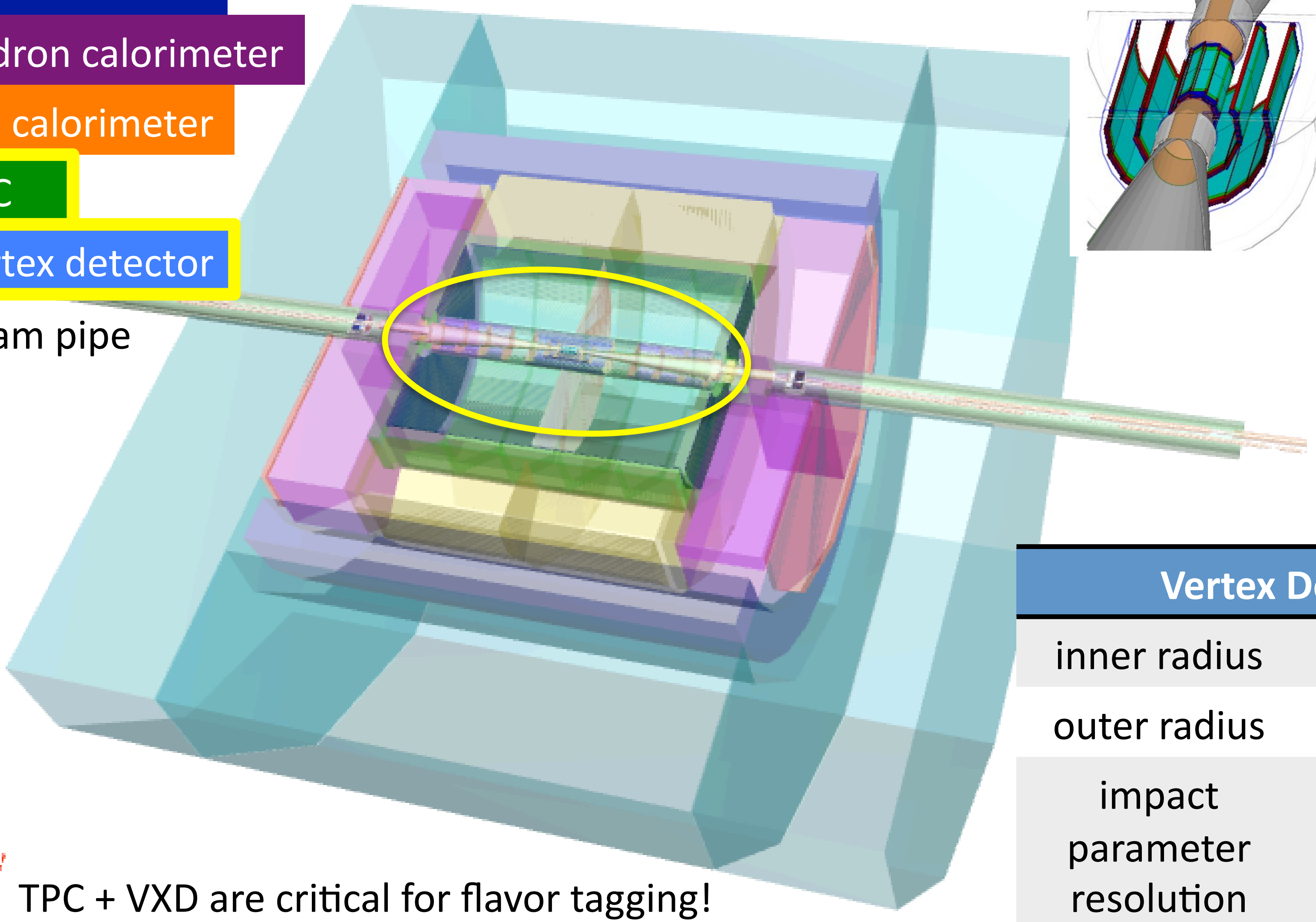
ideal flavor tagging

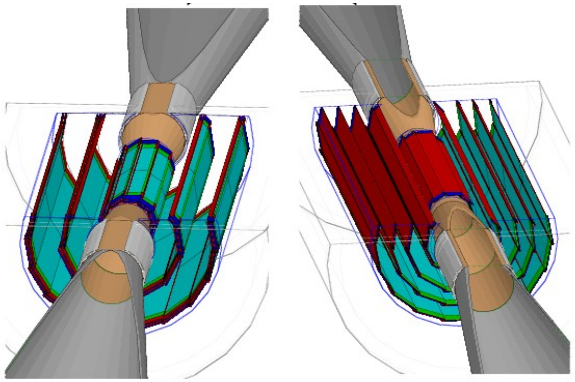
- reconstruct the entire decay chain (b -> c -> s) in a jet



  - vertex reconstruction is important
    - but vertices cannot be made from a single track
      - use track measurements (impact parameter)
  - presence of neutral particles
    - missing correction by using p_T
  - lepton ID: energetic/isolated leptons is a sign of heavy quark decays
- key is variable combination
  - likelihood, multivariate analysis (e.g. neural net, BDTs)
  - event categorization (discrete variables)

ILD Detector

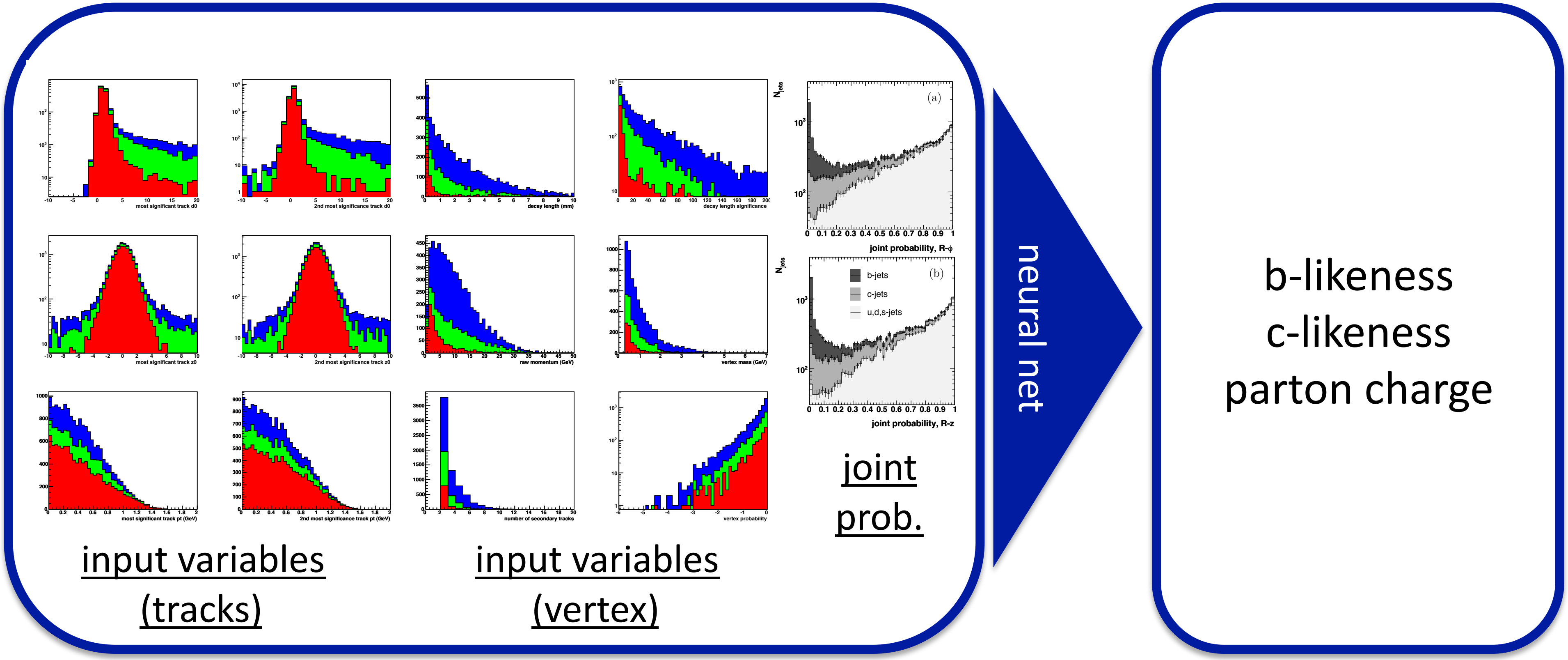




TPC + VXD are critical for flavor tagging!

Vertex Detector	
inner radius	15 mm
outer radius	60 mm
impact parameter resolution	< 5 mm (high momentum)

LCFIVertex



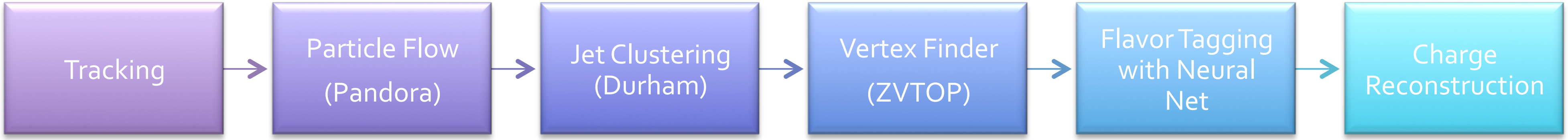
input variables (tracks)

input variables (vertex)

joint prob.

neural net

b-likeness
c-likeness
parton charge



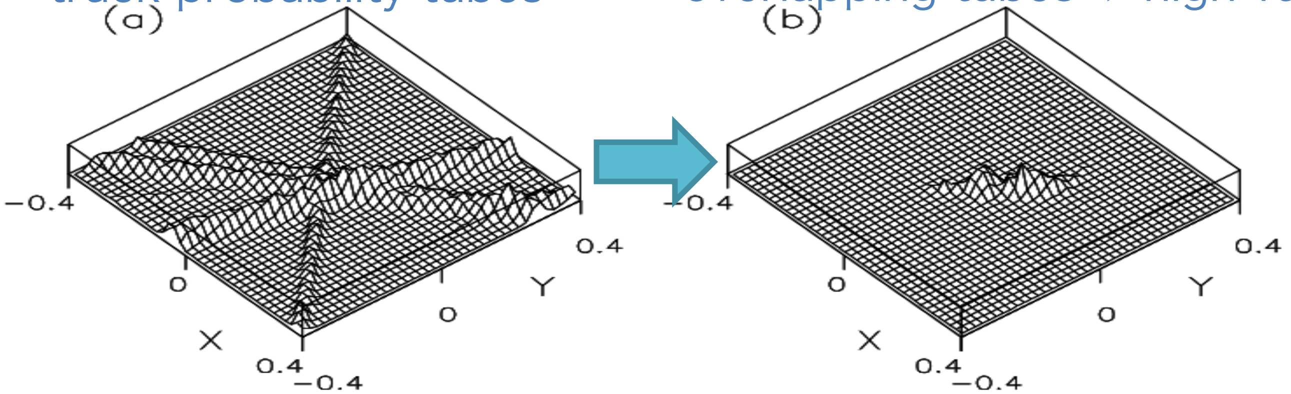
Tracking → Particle Flow (Pandora) → Jet Clustering (Durham) → Vertex Finder (ZVTOP) → Flavor Tagging with Neural Net → Charge Reconstruction

Vertex Finder

- topological vertex finder (ZVTOP)

track probability tubes

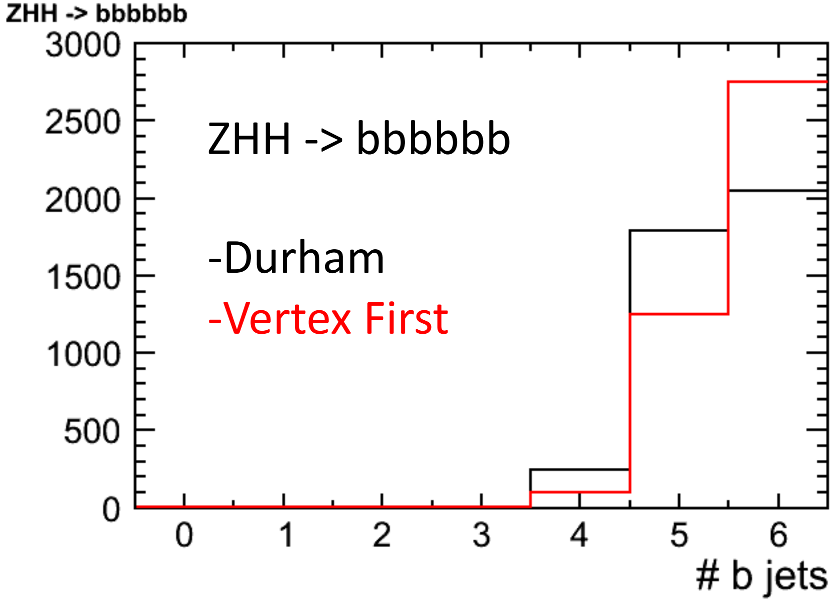
vertex function:
overlapping tubes -> high values



  - can find vertices for arbitrary topology with any number of tracks
  - takes jet direction as input to bias the secondary vertex search away from the primary vertex
- tear-down algorithm
  - start from a set of tracks, remove tracks which are inconsistent (large chi-squared contribution)
  - if the primary tracks are properly removed, vertices can be found with high efficiency
  - used for primary vertex finding
- build-up algorithm
  - using track pairs as seed, attach other tracks
  - good seeds lead to good vertices
  - used for secondary vertex finding

vertex splitting

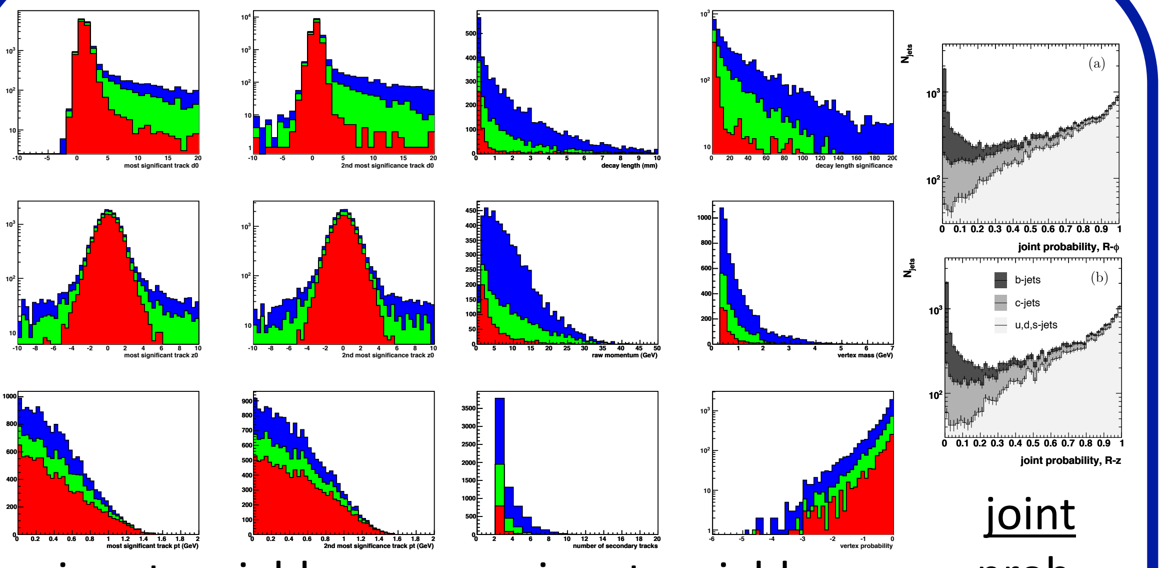
- tracks from secondary decays can get combined into wrong jets
  - reduces the b-tagging probability of real b-jets
  - increases the mis-tagging probability for fake b-jets
- this effect becomes significant in many jet events!
- solution: perform vertex finding <u>before</u> jet clustering
  - challenge: secondary vertex search using all tracks in the event
- build-up type vertex finder optimized for high purity shown to improve performance



count of jets containing >50% of secondary tracks originating from a b-hadron (MC information)

for details: see talk by T. Suehara

LCFIVertex+



input variables (tracks)

+many more

input variables (vertex)

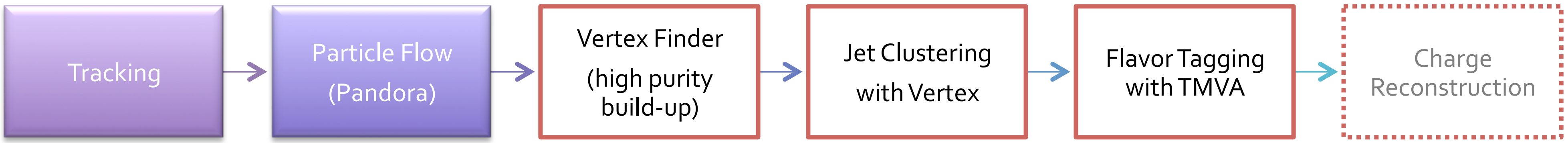
+many more

joint prob.

neural net

b-likeness

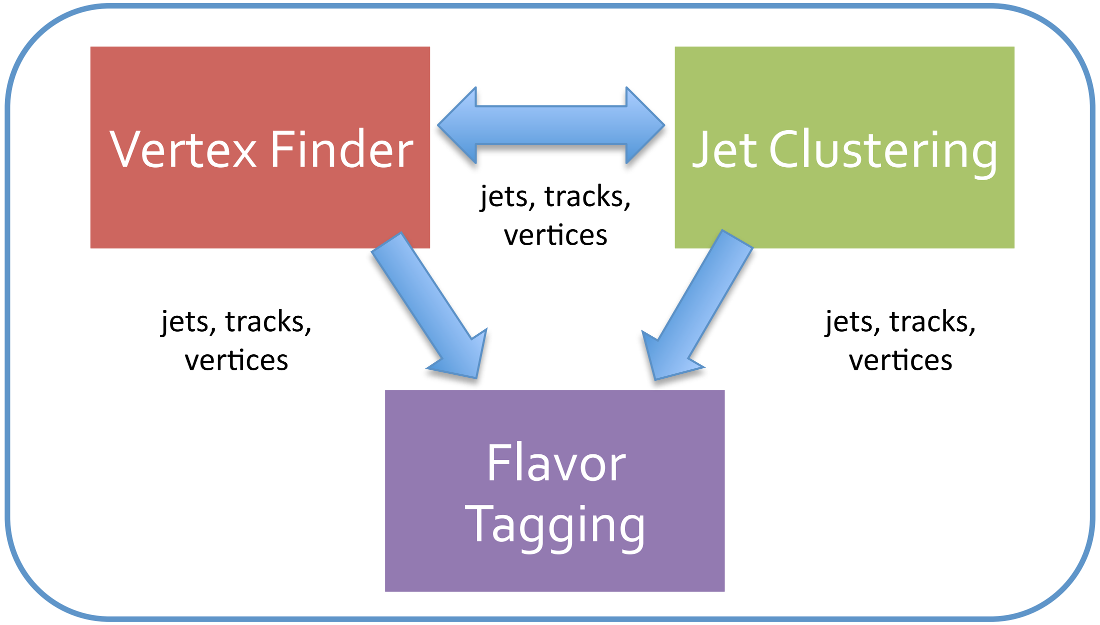
(c-likeness and vertex charge not yet implemented in the extended framework)



- VF <-> JC order switched, critical for many jet events!
- new multivariate analysis framework with TMVA

LCFIVertex framework

- improvements in **vertex finding, jet clustering, flavor tagging** in a unified way
  - creation of a new framework suited to this task
    - data types: event, track, neutral, mcparticle, jet, vertex
    - algorithms: vertex finding, jet clustering, flavor tagging



software framework

- Marlin processor
  - takes as input PandoraPFOs, Vertex (optional), ...
  - outputs jets, vertices (primary/secondary), flavor tags (PIDHandler)
- implementation:
  - dedicated data types for jet clustering & flavor tagging
  - modular design of algorithms (e.g. vertex finder, jet clustering) which can be switched on/off & combined in any order
  - control via xml steering file

Task	Status
Marlin interface	done
Primary vertex finder	done
Secondary vertex finder	done
Flavor tagging variables	done
TMVA interface	in progress!
Documentation	not started

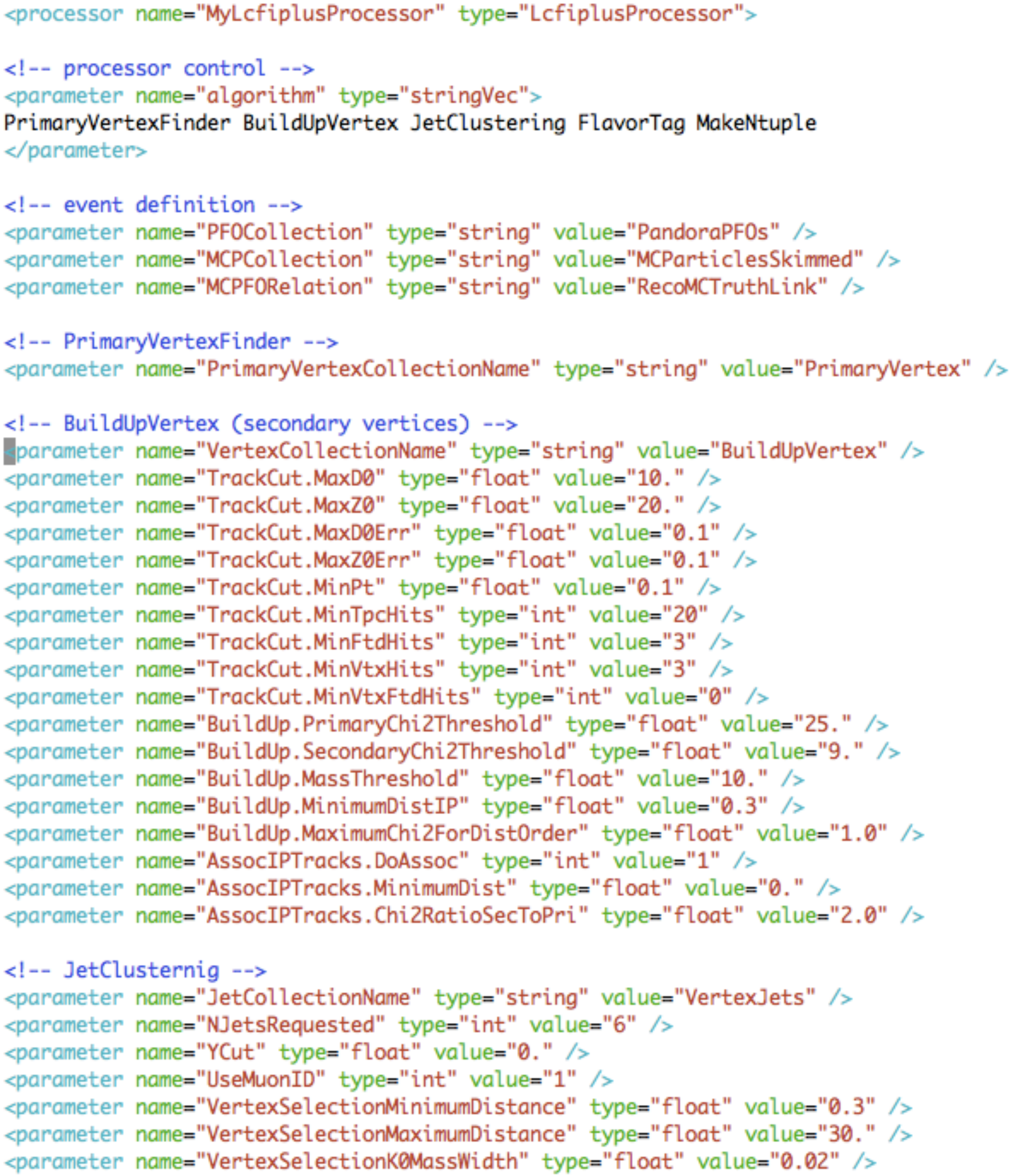
computing intensive: should be done with mass production!

Name of package: LCFIPlus
Latest code is in DESY SVN
First release expected in 1-2 weeks

choose which algorithm to run

parameters for vertex reconstruction

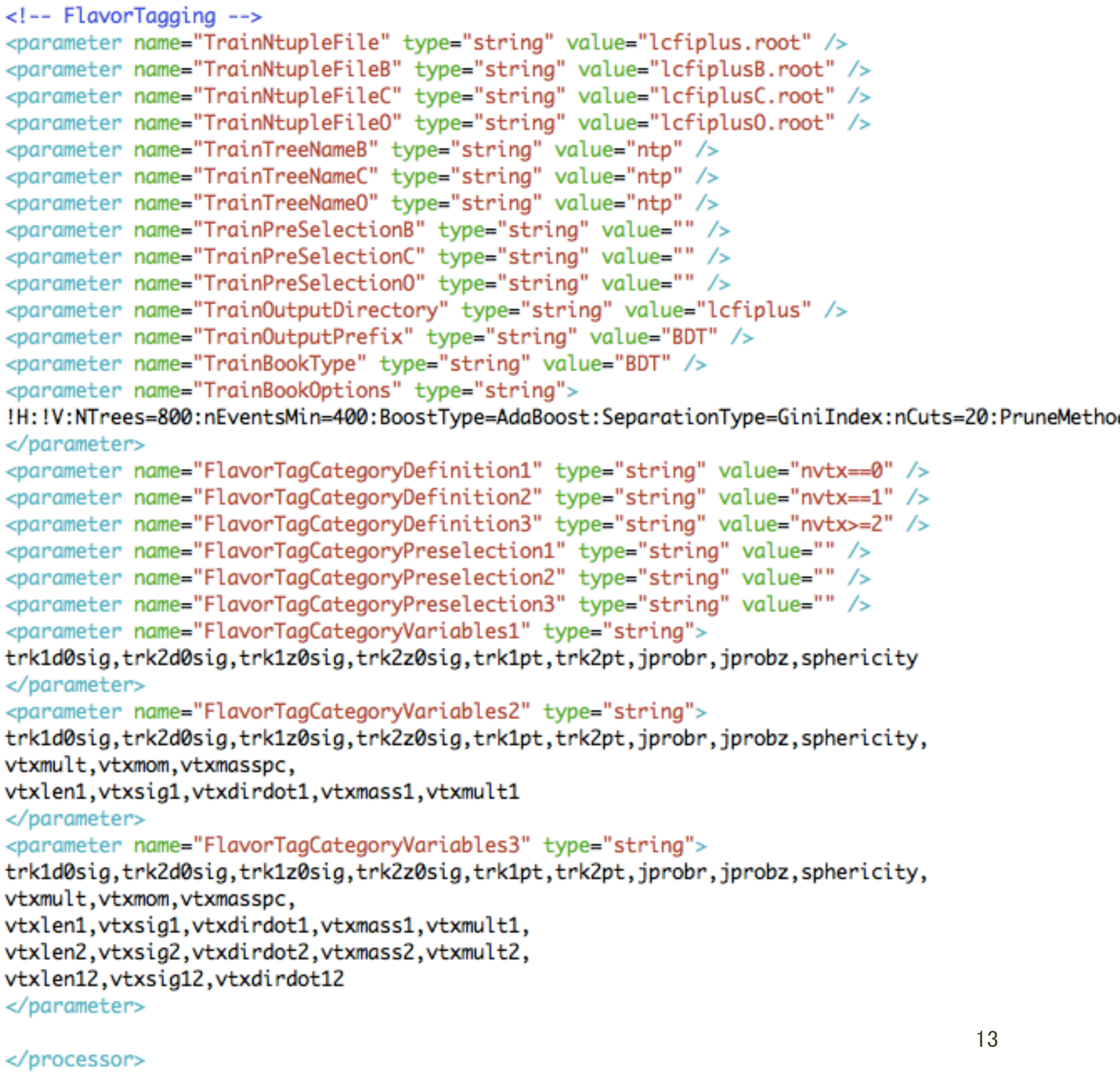
parameters for jet clustering



```
<processor name="MyLcfiplusProcessor" type="LcfiplusProcessor">

<!-- processor control -->
<parameter name="algorithm" type="stringVec">
PrimaryVertexFinder BuildUpVertex JetClustering FlavorTag MakeNtuple
</parameter>

<!-- event definition -->
<parameter name="PFOCollection" type="string" value="PandoraPFOs" />
<parameter name="MCPCollection" type="string" value="MCParticlesSkimmed" />
<parameter name="MCPFORelation" type="string" value="RecoMCTruthLink" />

<!-- PrimaryVertexFinder -->
<parameter name="PrimaryVertexCollectionName" type="string" value="PrimaryVertex" />

<!-- BuildUpVertex (secondary vertices) -->
<parameter name="VertexCollectionName" type="string" value="BuildUpVertex" />
<parameter name="TrackCut.MaxD0" type="float" value="10." />
<parameter name="TrackCut.MaxZ0" type="float" value="20." />
<parameter name="TrackCut.MaxD0Err" type="float" value="0.1" />
<parameter name="TrackCut.MaxZ0Err" type="float" value="0.1" />
<parameter name="TrackCut.MinPt" type="float" value="0.1" />
<parameter name="TrackCut.MinTpcHits" type="int" value="20" />
<parameter name="TrackCut.MinFtdHits" type="int" value="3" />
<parameter name="TrackCut.MinVtxHits" type="int" value="3" />
<parameter name="TrackCut.MinVtxFtdHits" type="int" value="0" />
<parameter name="BuildUp.PrimaryChi2Threshold" type="float" value="25." />
<parameter name="BuildUp.SecondaryChi2Threshold" type="float" value="9." />
<parameter name="BuildUp.MassThreshold" type="float" value="10." />
<parameter name="BuildUp.MinimumDistIP" type="float" value="0.3" />
<parameter name="BuildUp.MaximumChi2ForDistOrder" type="float" value="1.0" />
<parameter name="AssocIPTracks.DoAssoc" type="int" value="1" />
<parameter name="AssocIPTracks.MinimumDist" type="float" value="0." />
<parameter name="AssocIPTracks.Chi2RatioSecToPri" type="float" value="2.0" />

<!-- JetClusternig -->
<parameter name="JetCollectionName" type="string" value="VertexJets" />
<parameter name="NJetsRequested" type="int" value="6" />
<parameter name="YCut" type="float" value="0." />
<parameter name="UseMuonID" type="int" value="1" />
<parameter name="VertexSelectionMinimumDistance" type="float" value="0.3" />
<parameter name="VertexSelectionMaximumDistance" type="float" value="30." />
<parameter name="VertexSelectionK0MassWidth" type="float" value="0.02" />
```

configuration
for training
(PRELIMINARY)

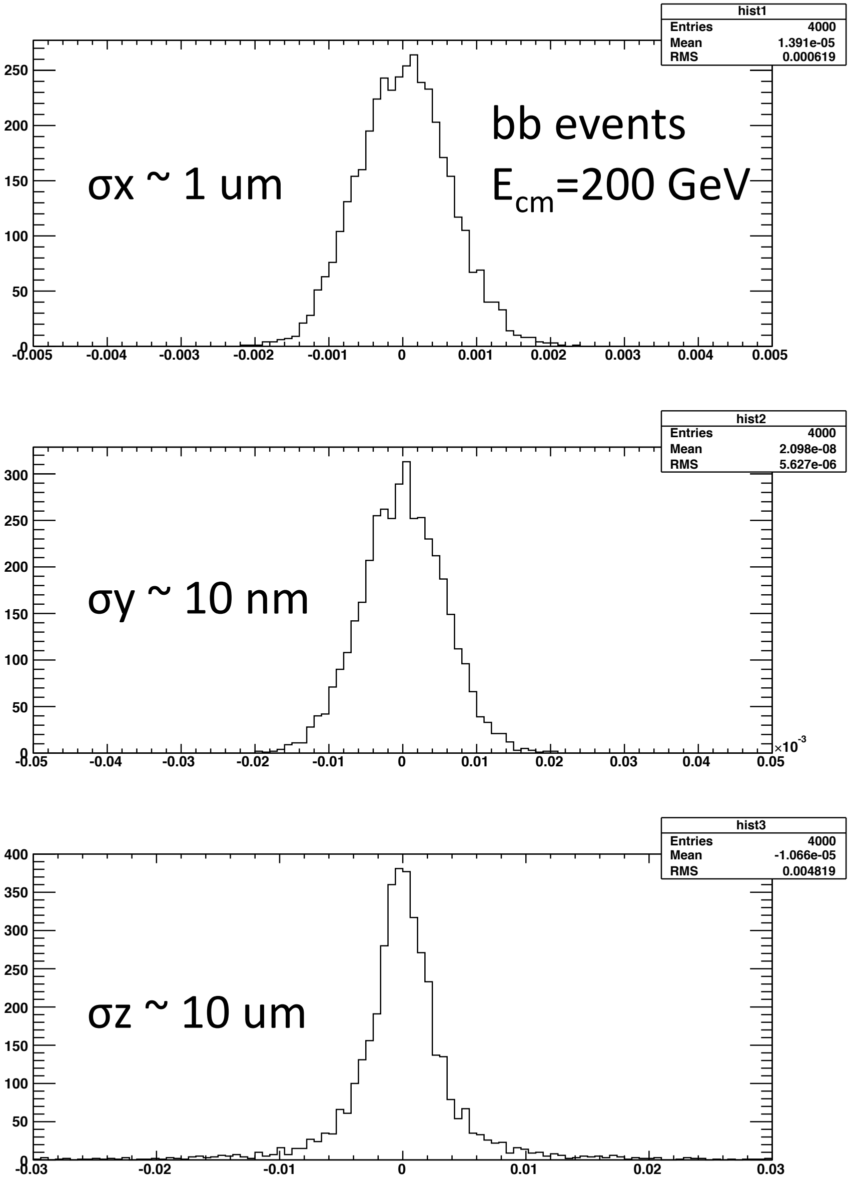
definition of
flavor tagging
categories &
variables



```
<!-- FlavorTagging -->
<parameter name="TrainNtupleFile" type="string" value="lcfiplus.root" />
<parameter name="TrainNtupleFileB" type="string" value="lcfiplusB.root" />
<parameter name="TrainNtupleFileC" type="string" value="lcfiplusC.root" />
<parameter name="TrainNtupleFileO" type="string" value="lcfiplusO.root" />
<parameter name="TrainTreeNameB" type="string" value="ntp" />
<parameter name="TrainTreeNameC" type="string" value="ntp" />
<parameter name="TrainTreeNameO" type="string" value="ntp" />
<parameter name="TrainPreSelectionB" type="string" value="" />
<parameter name="TrainPreSelectionC" type="string" value="" />
<parameter name="TrainPreSelectionO" type="string" value="" />
<parameter name="TrainOutputDirectory" type="string" value="lcfiplus" />
<parameter name="TrainOutputPrefix" type="string" value="BDT" />
<parameter name="TrainBookType" type="string" value="BDT" />
<parameter name="TrainBookOptions" type="string">
!H:!V:NTrees=800:nEventsMin=400:BoostType=AdaBoost:SeparationType=GiniIndex:nCuts=20:PruneMetho
</parameter>
<parameter name="FlavorTagCategoryDefinition1" type="string" value="nvtx==0" />
<parameter name="FlavorTagCategoryDefinition2" type="string" value="nvtx==1" />
<parameter name="FlavorTagCategoryDefinition3" type="string" value="nvtx>=2" />
<parameter name="FlavorTagCategoryPreselection1" type="string" value="" />
<parameter name="FlavorTagCategoryPreselection2" type="string" value="" />
<parameter name="FlavorTagCategoryPreselection3" type="string" value="" />
<parameter name="FlavorTagCategoryVariables1" type="string">
trk1d0sig,trk2d0sig,trk1z0sig,trk2z0sig,trk1pt,trk2pt,jprobr,jprobz,sphericity
</parameter>
<parameter name="FlavorTagCategoryVariables2" type="string">
trk1d0sig,trk2d0sig,trk1z0sig,trk2z0sig,trk1pt,trk2pt,jprobr,jprobz,sphericity,
vtxmult,vtxmom,vtxmasspc,
vtxlen1,vtxsig1,vtxdirdot1,vtxmass1,vtxmult1
</parameter>
<parameter name="FlavorTagCategoryVariables3" type="string">
trk1d0sig,trk2d0sig,trk1z0sig,trk2z0sig,trk1pt,trk2pt,jprobr,jprobz,sphericity,
vtxmult,vtxmom,vtxmasspc,
vtxlen1,vtxsig1,vtxdirdot1,vtxmass1,vtxmult1,
vtxlen2,vtxsig2,vtxdirdot2,vtxmass2,vtxmult2,
vtxlen12,vtxsig12,vtxdirdot12
</parameter>

</processor>
```

primary vertex finder

- teardown type algorithm
  - preselected tracks: d_0<20 mm, z_0<20 mm, require hit in first VXD, 5 hits in VXD + FTD
  - remove tracks with the worst contribution to chi2 until threshold value is reached
  - beam constraint implemented via "straight" tracks with errors defined according to the beam spot size
  - Gaussian smearing of initial position
    - since all events are generated at (0,0,0) at the moment



secondary vertex finding

- build-up type vertex
  - start with track pair, associate additional tracks which are compatible
  - jet direction is not used (essential!)
  - vertex quality selection based on vertex mass
- evaluation of secondary vertex performance is complicated because of vertex splitting & merging
- we use track-based evaluation
  - take the tracks used to form a vertex and look at their origin (primary, b, c, other)
  - good vertex algorithm: fewer primary, more b/c

secondary vertex finding

qqhh, 500 GeV

Trks.	# tracks	ZVTOP (Durham 6-jets)			Build-up vertex finder		
		All	Good	Pure	All	Good	Pure
Primary	10231	160			54		
b	2037	1399	1344	857	1309	1303	919
c	2433	1653	1618	1181	1571	1562	1197
Others	587	159	45	34	46	18	13
All	15288	3371	3007	2072	2980	2883	2129

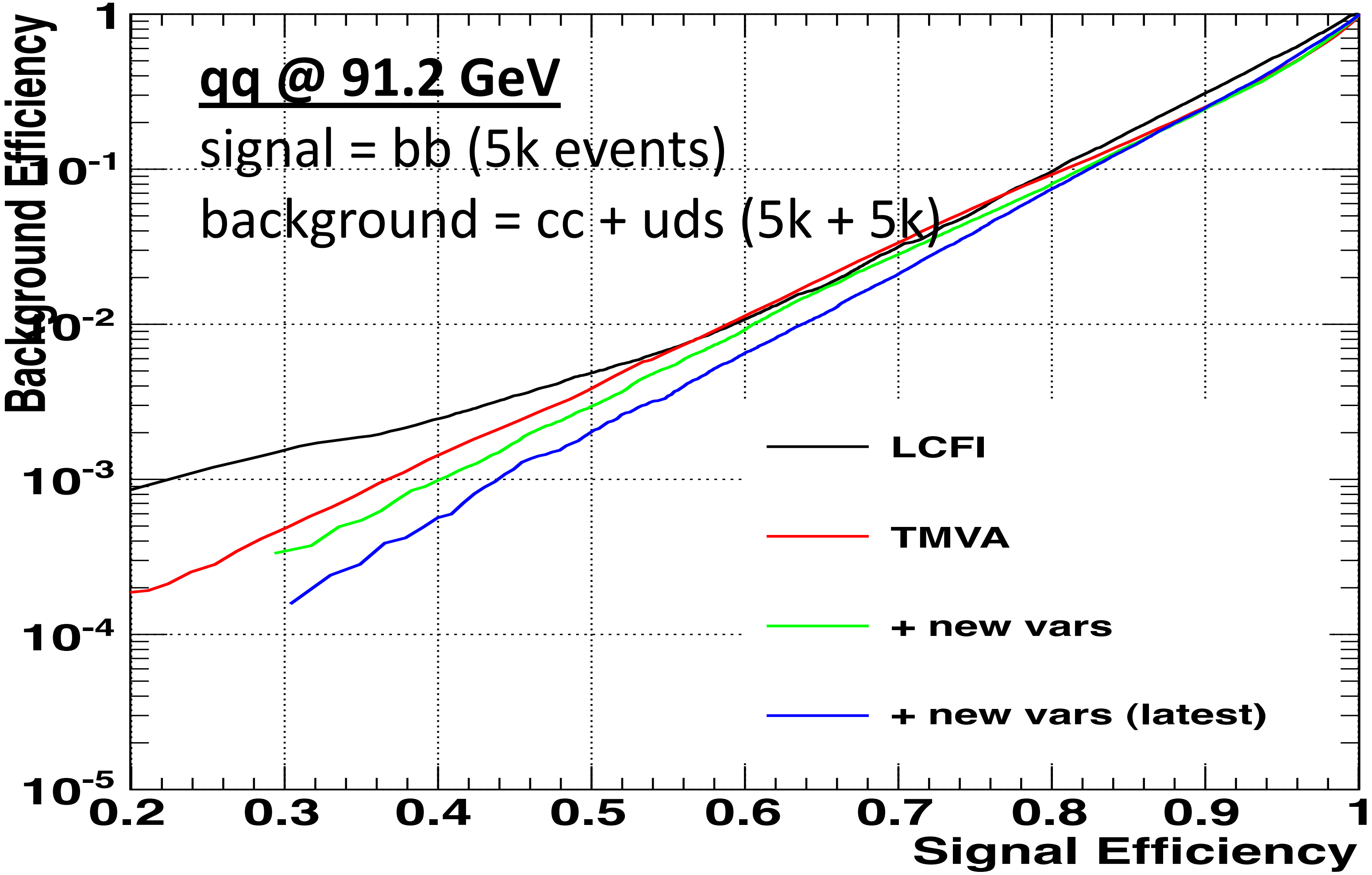
bbcssc, 500 GeV

Trks.	# tracks	ZVTOP (Durham 6-jets)			Build-up vertex finder		
		All	Good	Pure	All	Good	Pure
Primary	6980	76			14		
b	893	612	593	405	579	573	413
c	1627	1086	1052	878	1045	1035	874
Others	430	119	28	24	53	19	15
All	9930	1893	1673	1307	1691	1627	1302

GOOD vertex: require all tracks originate from the same b/c tree

PURE: b/c must be separated

example performance of flavor tagging



Improvement over existing algorithm is seen in all regions of signal efficiency.

summary

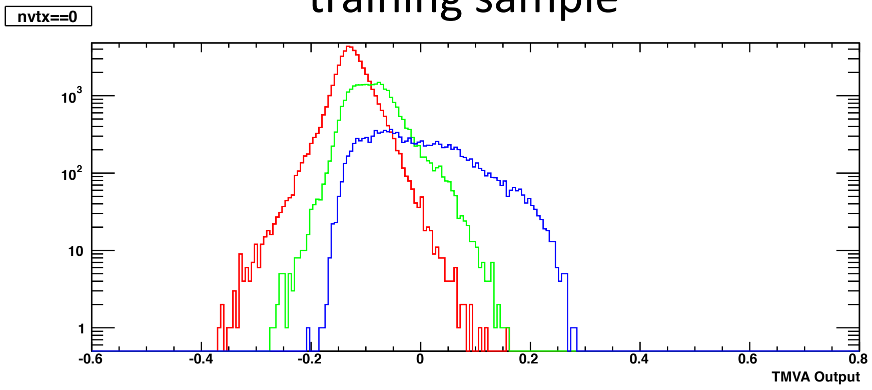
- we have extended the flavor tagging framework for linear colliders
  - improvements in vertex finding, jet clustering, and flavor tagging
- future work will focus on:
  - first release of software (a.s.a.p.)
  - optimization at higher energies & different detector configurations
  - inclusion of backgrounds
  - lepton ID within jets (with new Pandora)
  - application to physics analysis (ZHH, ttH, …)

backup

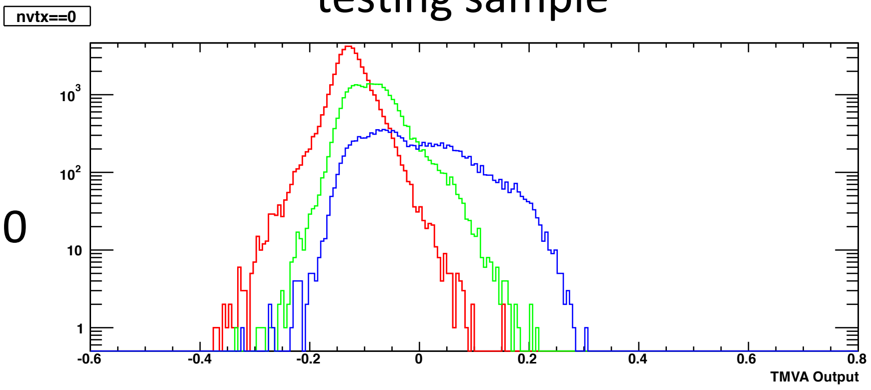
BDT response

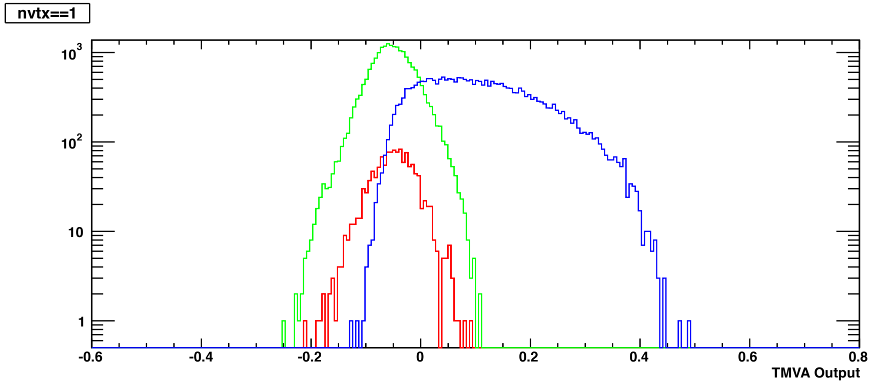
training sample

testing sample

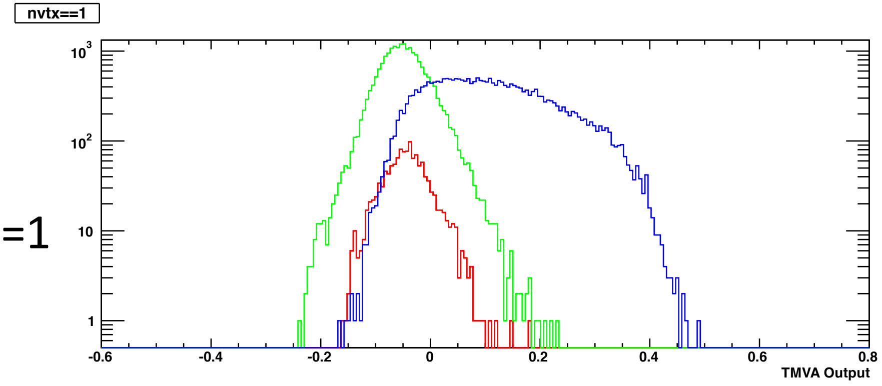


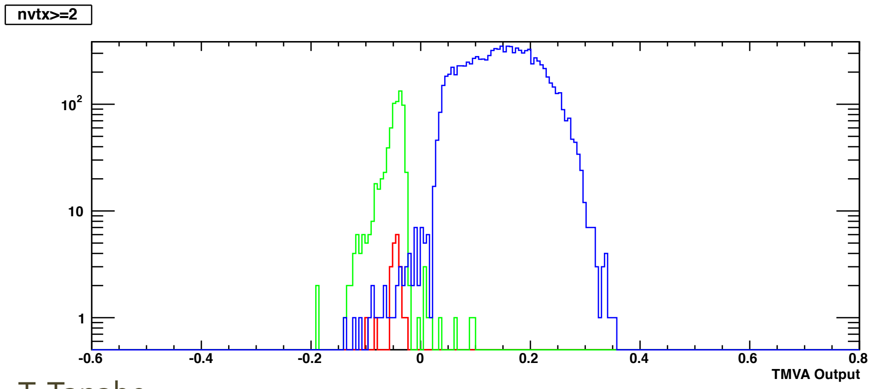
#vtx=0



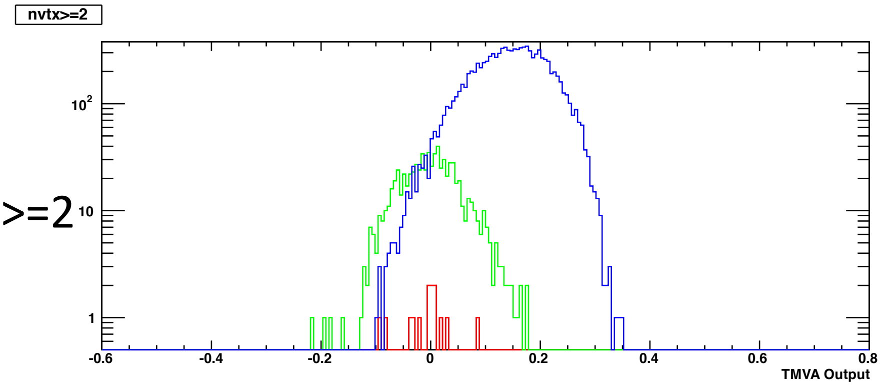


#vtx=1





#vtx>=2

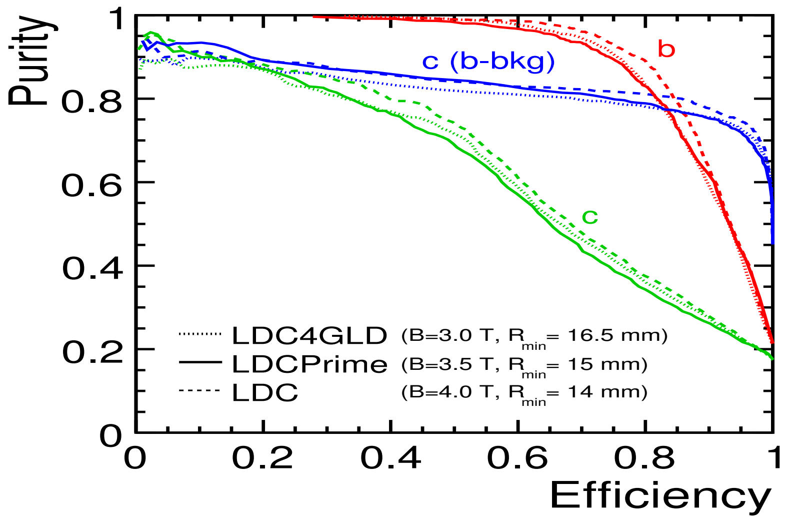


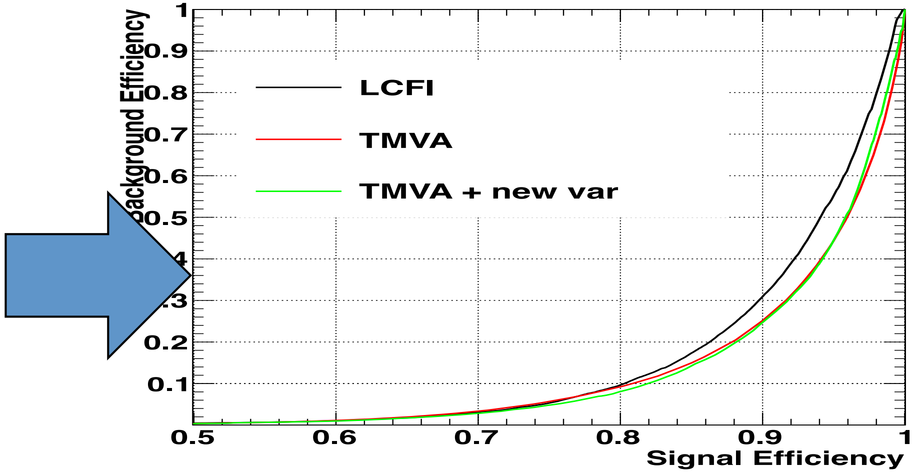
new input variables

- the new variables are trained by BDT
  - for # vertex = 0 (9 variables):
    - d_0 impact parameter (1)
    - d_0 impact parameter (2)
    - z_0 impact parameter (1)
    - z_0 impact parameter (2)
    - track momentum (1)
    - track momentum (2)
    - d_0 joint probability
    - z_0 joint probability
    - boosted sphericity
  - for # vertex = 1 (17 variables):
    - d_0 impact parameter (1)
    - d_0 impact parameter (2)
    - z_0 impact parameter (1)
    - z_0 impact parameter (2)
    - track momentum (1)
    - track momentum (2)
    - d_0 joint probability
    - z_0 joint probability
    - boosted sphericity
    - vertex decay length
    - vertex decay length significance
    - vertex momentum
    - vertex mass (pt-corrected)
    - vertex mass (not pt-corrected)
    - vertex multiplicity
    - vertex probability from the fitter
    - vertex disp/momentum angle
  - for # vertex >= 2 (29 variables):
    - d_0 impact parameter (1)
    - d_0 impact parameter (2)
    - z_0 impact parameter (1)
    - z_0 impact parameter (2)
    - track momentum (1)
    - track momentum (2)
    - d_0 joint probability
    - z_0 joint probability
    - boosted sphericity
    - vertex #1 decay length
    - vertex #2 decay length
    - distance between vertex #1 & #2
    - vertex #1 decay length significance
    - vertex #2 decay length significance
    - separation significance between vertex #1 & #2
    - vertex #1 momentum
    - vertex #2 momentum
    - vertex momentum (combined)
    - vertex #1 mass (not pt-corrected)
    - vertex #2 mass (not pt-corrected)
    - vertex mass (combined, pt-corrected)
    - vertex #1 multiplicity
    - vertex #2 multiplicity
    - vertex multiplicity (combined)
    - vertex probability from the fitter
    - vertex #1 disp/momentum angle
    - vertex #2 disp/momentum angle
    - vertex #1/#2 disp/momentum angle
    - vertex #1/#2 angle

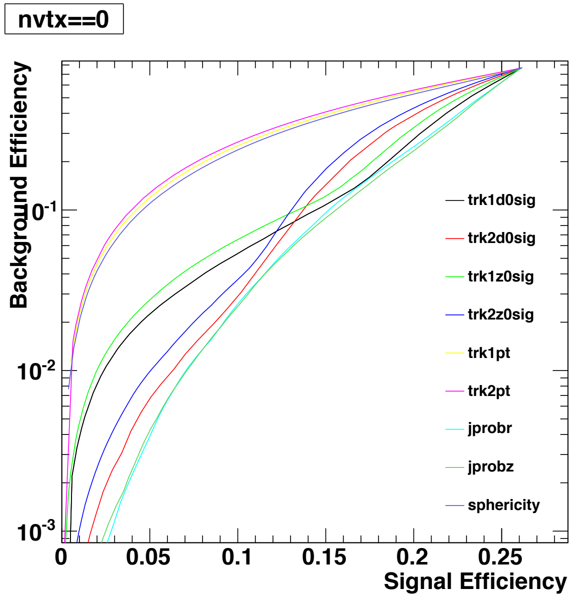
evaluating classifier response

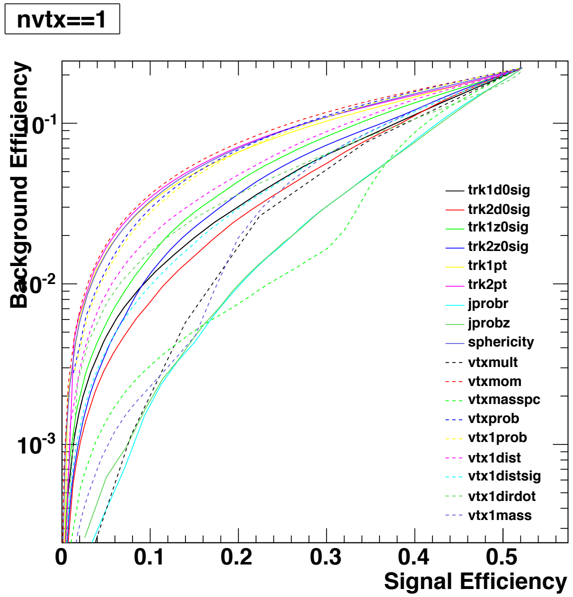
- LoI flavor-tagging evaluation produced purity-efficiency plots
  - but this depends on the fraction of heavy jets, which changes from sample to sample
    - BF(Z->bb)=15%, BF($H_{120\text{ GeV}}$->bb)=68%
- better to use a fraction-independent measure: evaluate using background efficiency versus signal efficiency instead

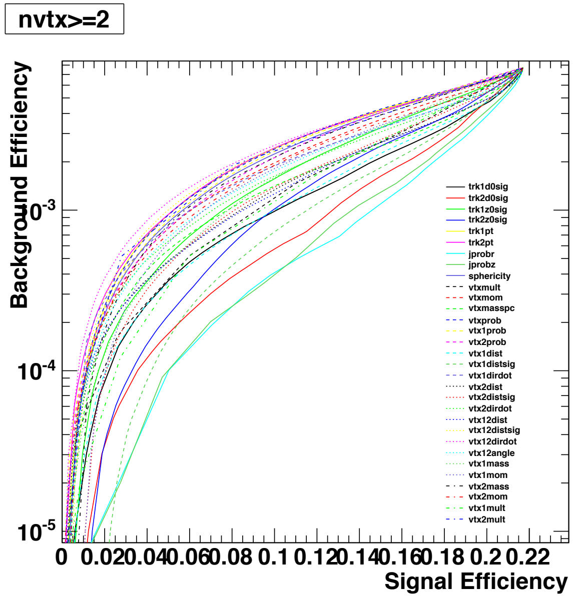




variable ranking



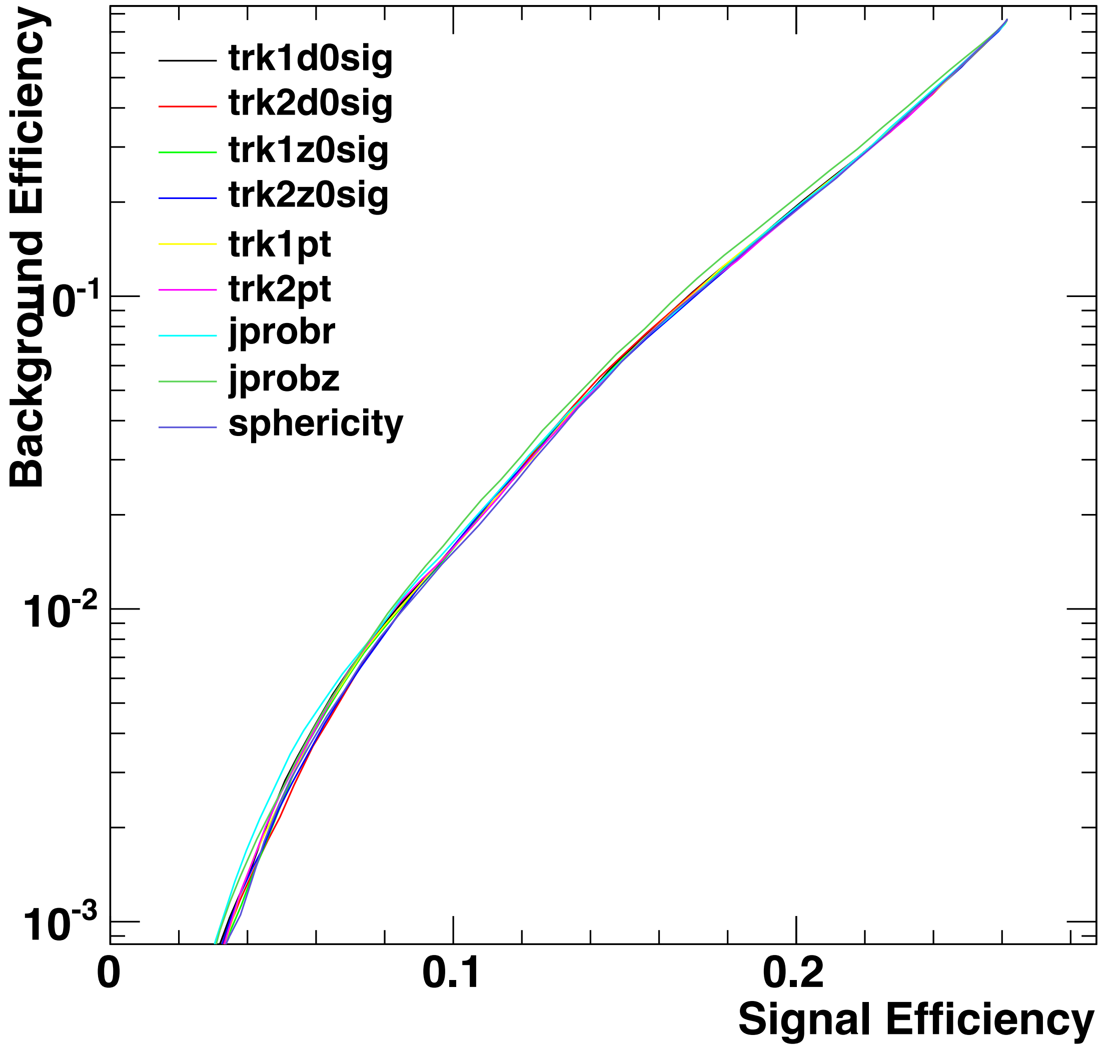




- single variable ranking shows the most useful variables (on their own)
  - joint probabilities, vertex mass
- however, any other uncorrelated variable can help; these plots do not show this effect

variable ranking (correlation)

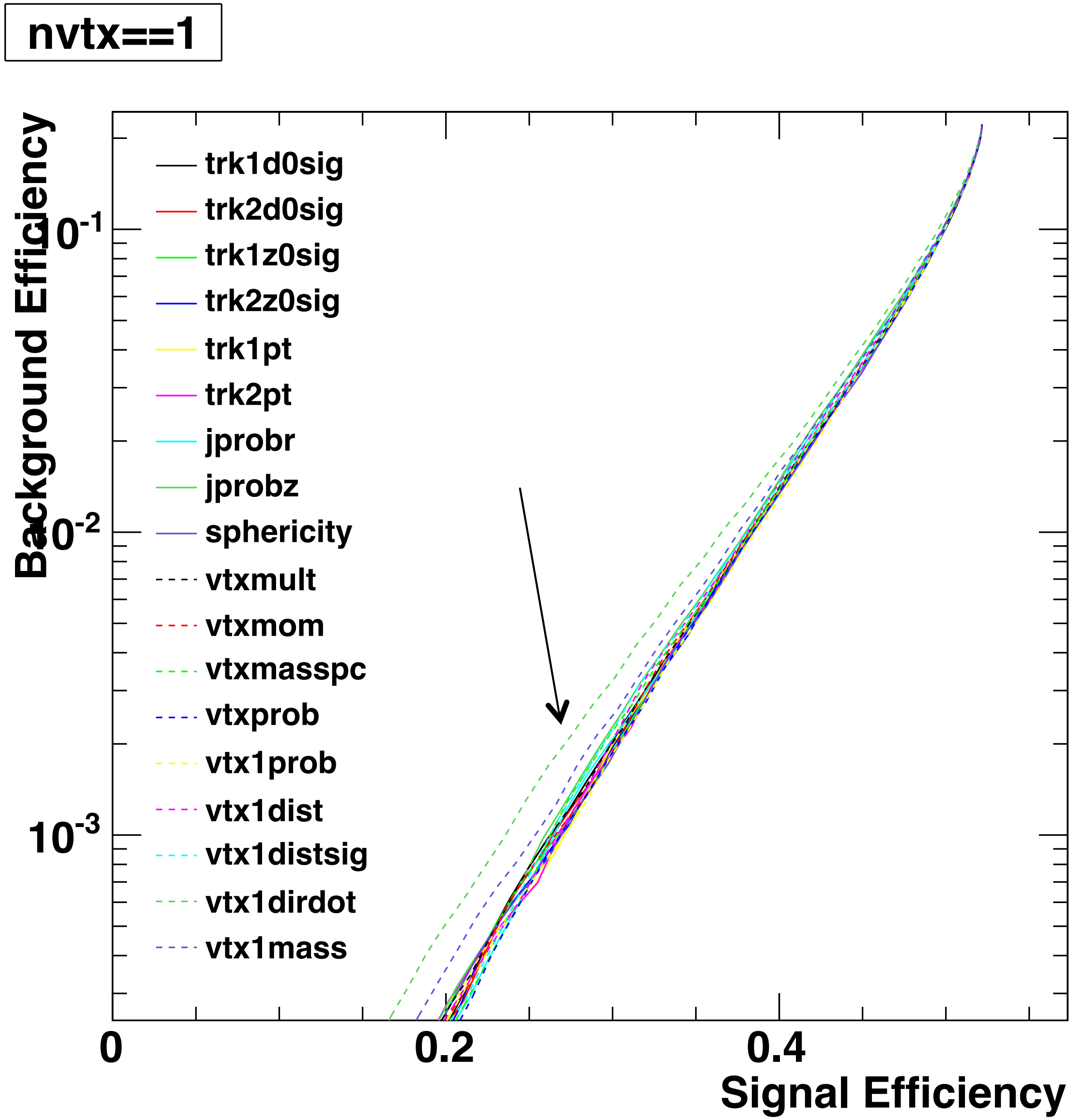
nvtx==0



- result of training after removing a variable
- this shows how “unique” this variable is in terms of uncorrelated classifying power
- significantly worse performance after removing the variable shows that it’s effective
- for nvtx=0, joint probabilities (both d_0 & z_0) are the most powerful as expected

T. Tanabe

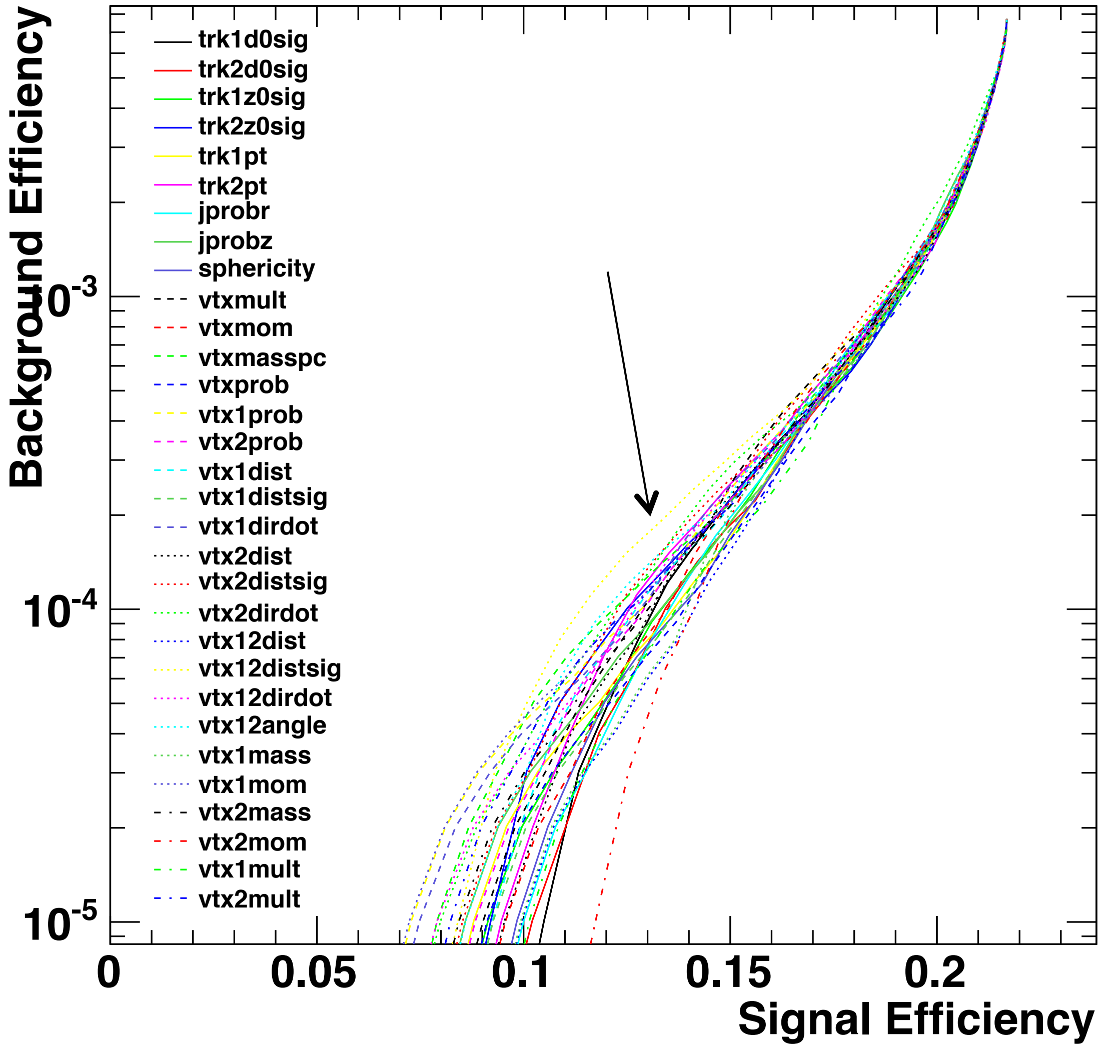
variable ranking (correlation)



- for nvtx=1, the most effective variables are:
  - displacement/ momentum angle of the vertex
  - uncorrected mass of the vertex
- newly added variables are shown to be effective!!!

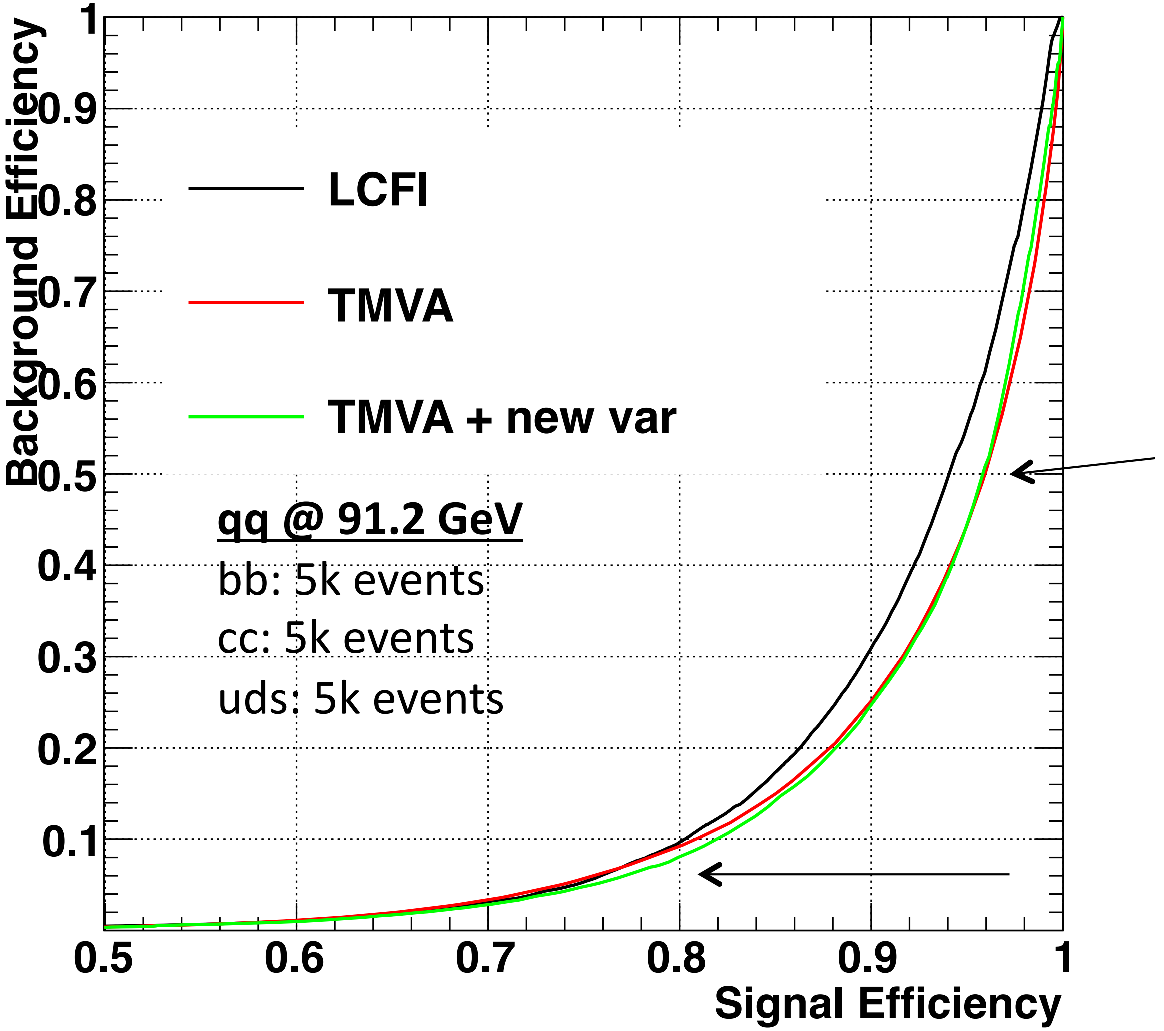
variable ranking (correlation)

nvtx>=2



- for nvtx>=2, the most effective variables are:
  - separation significance between the 1st and 2nd vertices
- AGAIN: newly added variables are shown to be effective!!!

results



background = cc & uds mixed equally

there is already improvement merely by switching to TMVA (signal eff > 0.8)

more improvement by adding new variables (signal eff > 0.75)

rejection of V° particles

- despite having V° taggers in the Marlin reconstruction chain, our vertex finders still find V°'s (K_S, Lambda, conversions) for two-track vertices
- we apply the following cuts to reject V°'s (reduce uds contamination):
  - cut on the angle θ between the vertex displacement from IP and the V° direction
  - mass requirements
  - K_S: $\cos\theta > 0.999$ & mass 15 MeV within PDG value
  - Lambda: $\cos\theta > 0.99995$ & mass 20 MeV within PDG value
  - conversions: $\cos\theta > 0.99995$ & less than 10 MeV for conversion mass, where the mass is geometrically corrected so that it is calculated using the track dip angles

$$m^2_{\mathrm{conv}} = 2|\vec{p}_1||\vec{p}_2|(1 - \cos\Delta\lambda_{12})$$

	before cut	after cut
K_S	3205	623
Lambda	1482	371
conversions	2544	278
other two-track reco vertices	30747	30333