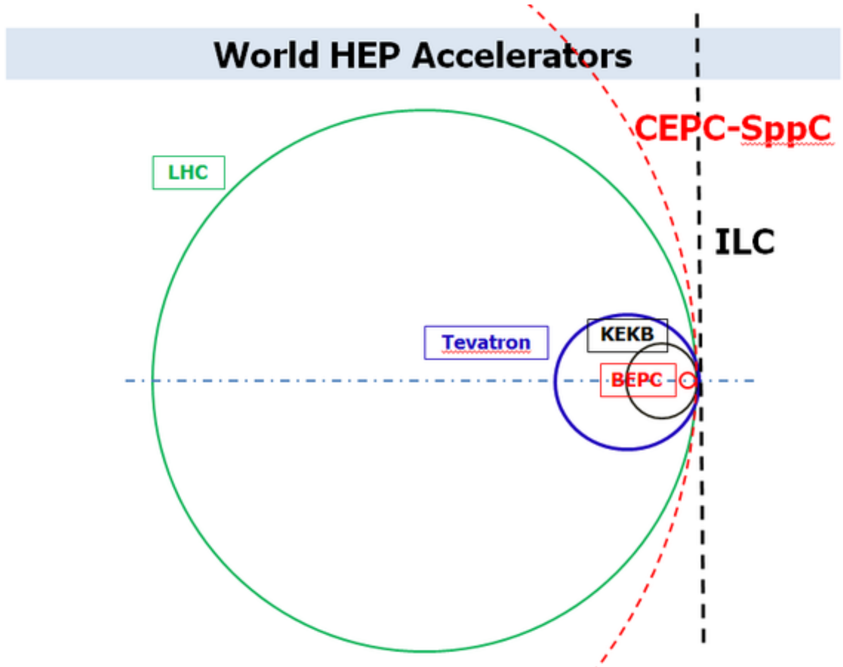
CEPC Workshop

Status and Outlook of H->bb/cc/gg Analysis

Yu Bai (from Southeast University, Nanjing)

On Behalf of CEPC Physics-Software Study Group

April 14, 2021



Why H->bb/cc/gg is important?

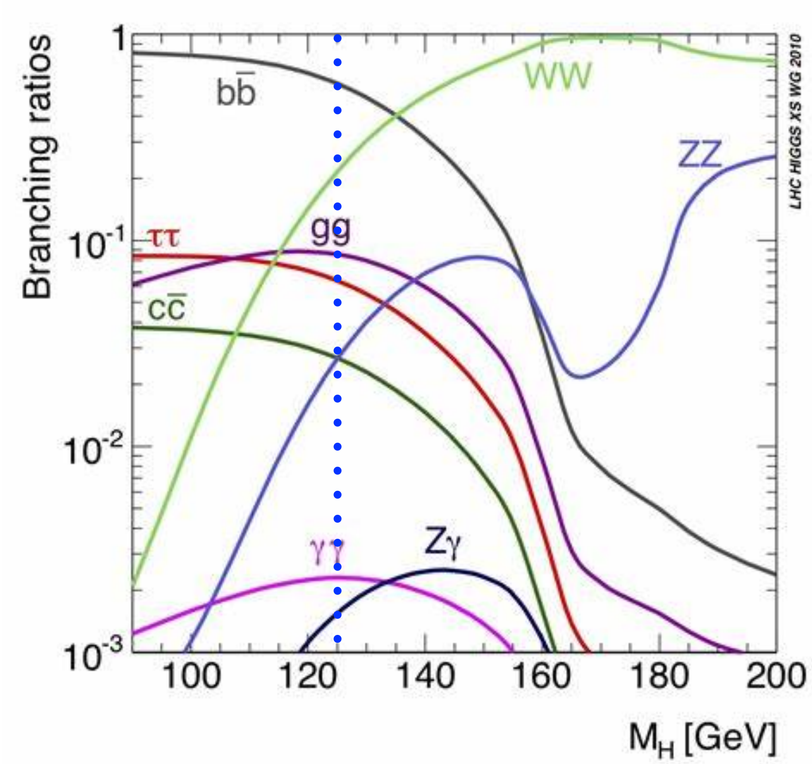
$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\bar{\Psi}\not{D}\psi + h.c. + \bar{\psi}_i y_{ij}\psi_j\phi + h.c. + |D_\mu\phi|^2 - V(\phi)$$

The only known sources of quarks' mass

A measurement on the couplings between Higgs and quarks will decode the origin of mass in great confidence

Current Status

H->bb/cc/gg is expected to be 57%/9%/3%



NOT bad with a 125 GeV Higgs mass

But that's not an easy task for our LHC colleges!

Direct H->bb measurement

- ***Observation of H->bb and VH production*** **by ATLAS Collaboration, Phys.Lett.B 786 (2018) 59**
- ***Observation of Higgs boson decay to bottom quarks*** **by CMS Collaboration, *Phys.Rev.Lett (2018) 121801***

The H->bb signal strength was measured with precision around 20%, consisted with SM prediction

Direct H->cc measurement:

- ***Search for the decay of the Higgs boson to Charm Quarks with the ATLAS Experiment Phys. Rev. Lett (2018), 211802***

A 95% CL upper limit set at about signal strength = 100

Gluon-gluon fusion Analysis

- **H->$\tau\tau$ in ATLAS**
- **H->WW in ATLAS**
- **H-> $\tau\tau$ in CMS**
- **H->$\gamma\gamma$ in CMS**

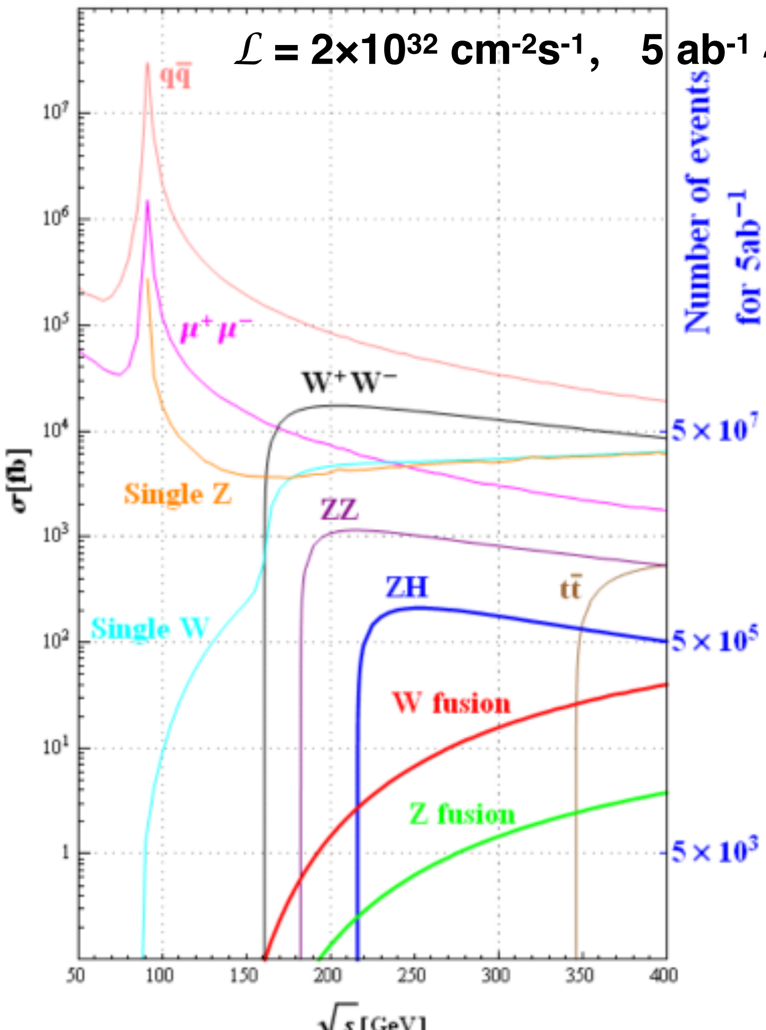
Uncertainty of O(10%)

Review of Previous Study

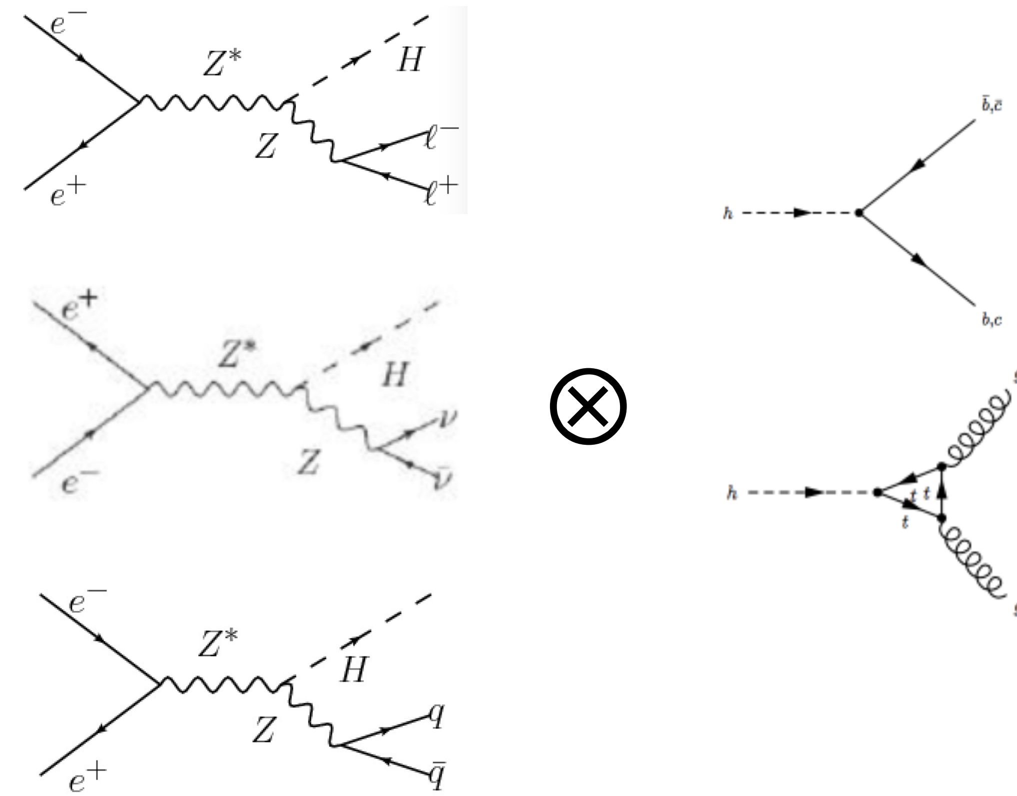
Study based on pre-CDR set up, using full simulated sample with sqrt(s)=250 GeV and old geometry (some extrapolate to sqrt(s) = 240 GeV)

H->bb/cc/gg at CEPC

$\mathcal{L} = 2\times10^{32}\ cm^{-2}s^{-1}$,   5 ab^{-1} ~10 years of running



~ 1 million higgs boson will be produced, 70% of them decay hadronically



Statistics normalized to 5/ab:

	eeH	$\mu\mu$H	$\tau\tau$H	qqH	$\nu\nu$H
bb	22k	20k	20k	410k	140k
cc	1.0k	1.0k	1.0k	20k	6.7k
gg	3.5k	3.2k	3.2k	65k	22k

We have explored:

- **96.7% of Higgs hadronic decay events**
- **2/3 of the over all Higgs events**

Analysis Strategy

H->bb/cc/gg Selection

Flavor components identification

Object and Event Selection

Chin.Phys.C, 43(4): 043002 (2019)

Flavor template fit

Particles are reconstruct with arborPFO algorithm.

Jets reconstructed by Durham algorithm

Chin.Phys.C, 44(1): 013001(2020)

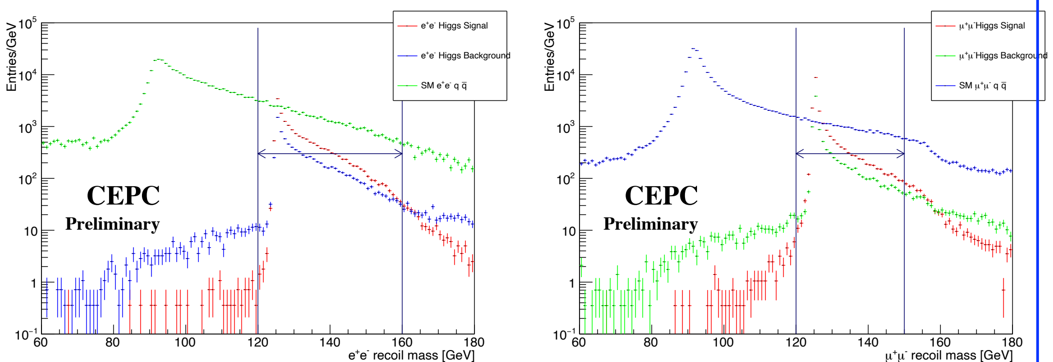
Flavor template × Recoil Mass fit

Leptons from W/Z need to pass isolation track selection

Dominant Backgrounds include the ZZ/WW events with same final states

Event Selection

llH



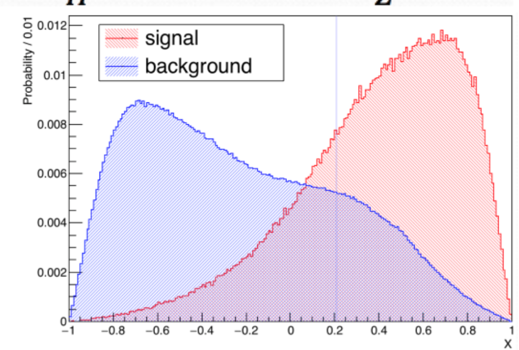
recoil mass is the crucial variable to distinguish signal and background

	signal	Signal eff	higgs bkg	non-higgs bkg
ee channel	9.15k	52.6%	1.10k	6.15k
$\mu\mu$ channel	12.8k	63.9%	1.48k	5.29k

qqH

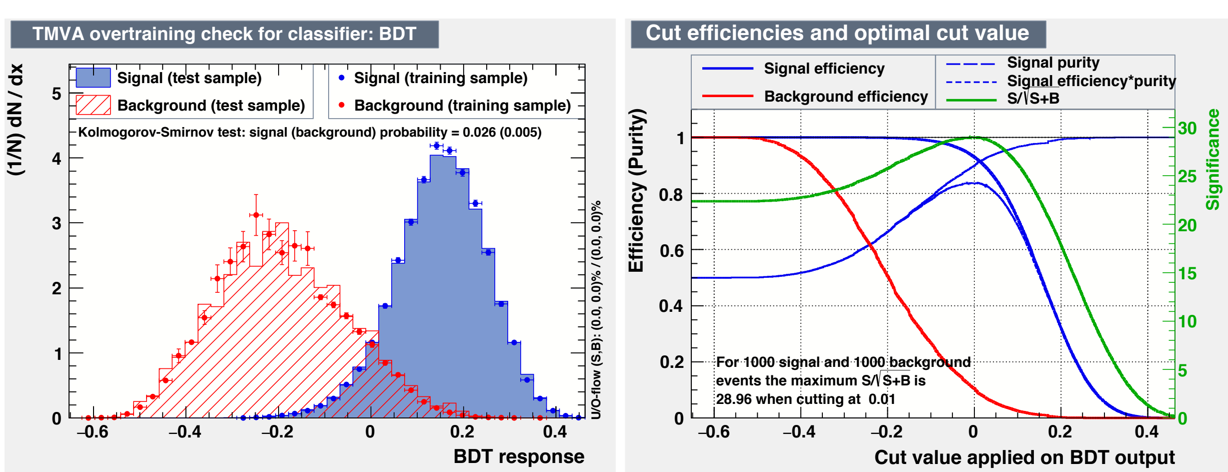
$$\chi^2 = \min_{i \neq j \neq a \neq b} \frac{(M_{ij} - M_H)^2}{\sigma_H^2} + \frac{(M_{ab} - M_Z)^2}{\sigma_Z^2}$$

$$X = \frac{\chi_W^2 - \chi_{ZH}^2}{\chi_W^2 + \chi_{ZH}^2}$$



Mass χ^2 with different hypothesis are combined to reject ZZ/WW events

$\nu\nu$H



BDT variable construct from nPFO, visible energy cos θ_{jj} and yin

qqH

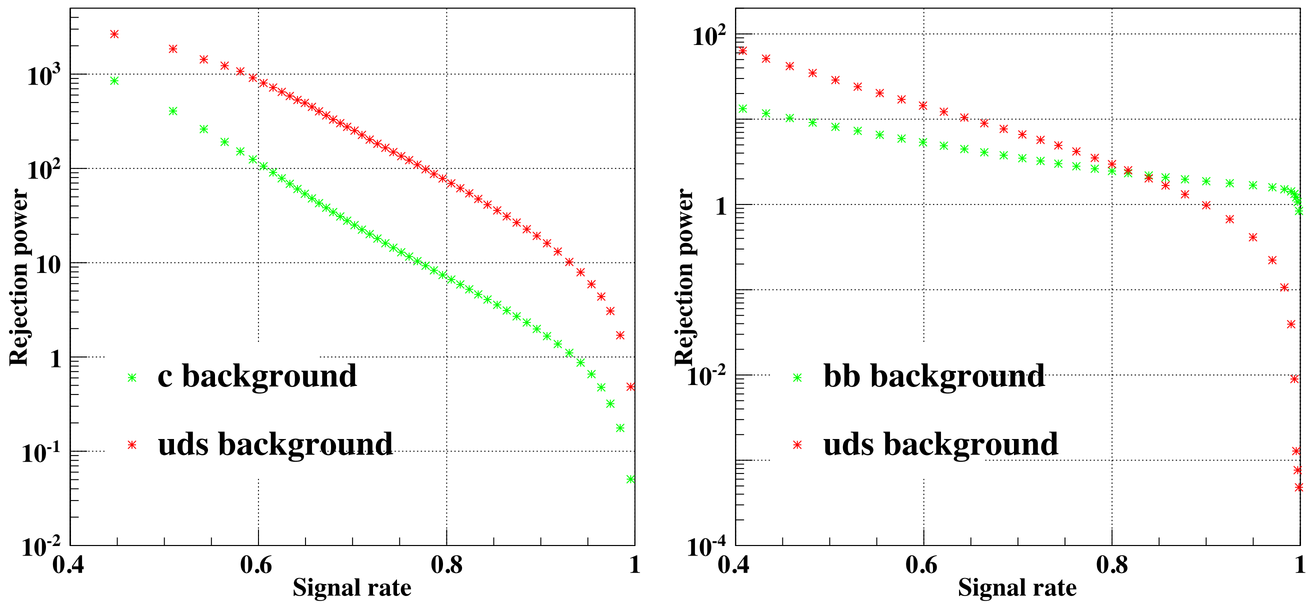
signal	Signal eff	higgs bkg	4f-hadronic	qq	4f-semilept
211.2k	42.8%	32.6k	1.08M	405.6k	0.58k

$\nu\nu$H

signal	Signal eff	higgs bkg	non-higgs bkg
85.8k	49.2%	1.96k	22.88k

Flavor Tagging and Flavor template

Flavor tagging performance

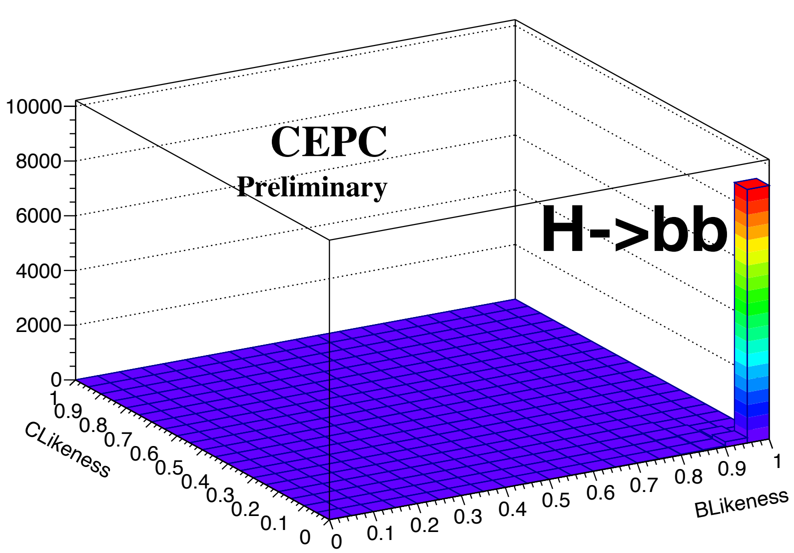


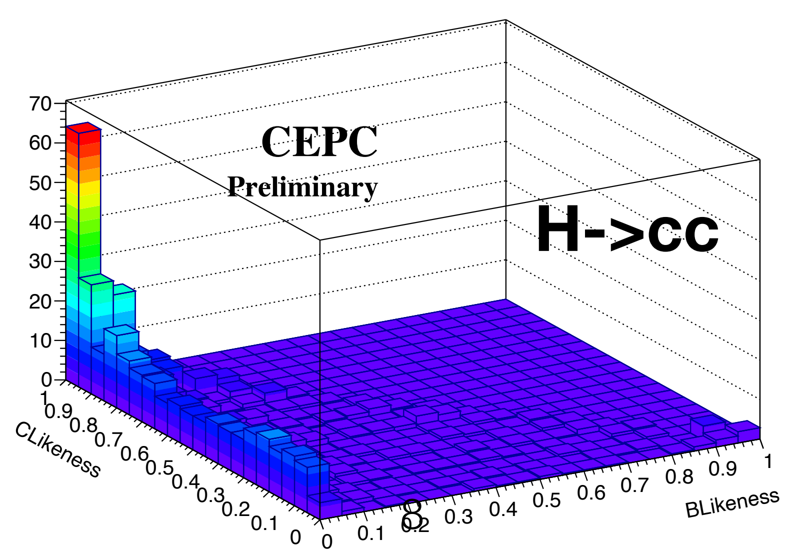
- 4 categories of events according to the soft-leptons and vertices multiplicity
- GBDT with tagging-sensitive variables method applied to each category

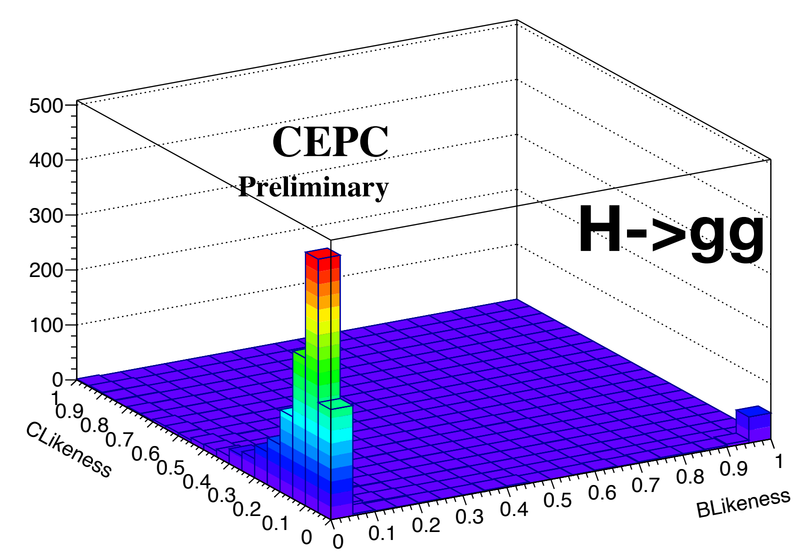
The performance of FT ensures the goal of precision

$$L_{qq} = \frac{qq\ pair}{qq\ pair\ +\ neither\ is\ q} = \frac{x_q^1\, x_q^2}{x_q^1 x_q^2 + (1 - x_q^1)(1 - x_q^1)}\ (qq = bb, cc)$$

Processes with different flavor components can be separated by template fit



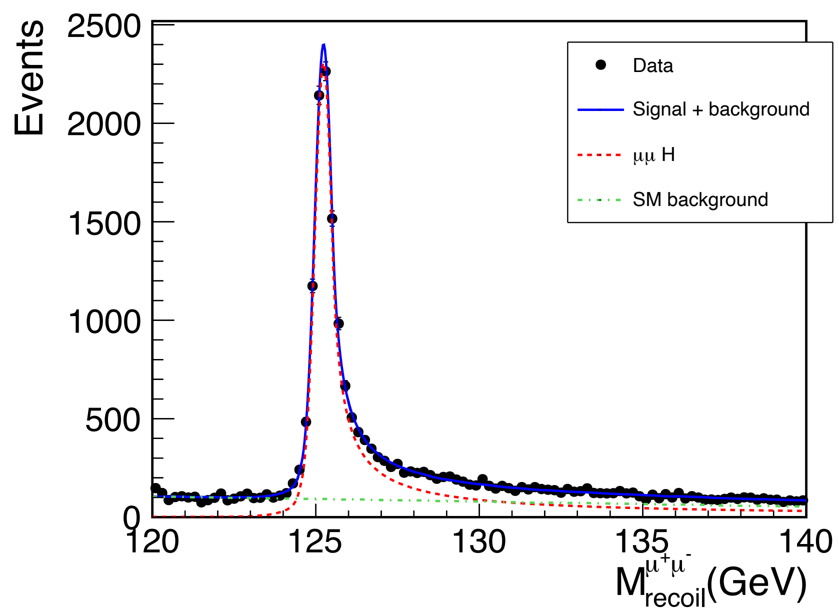


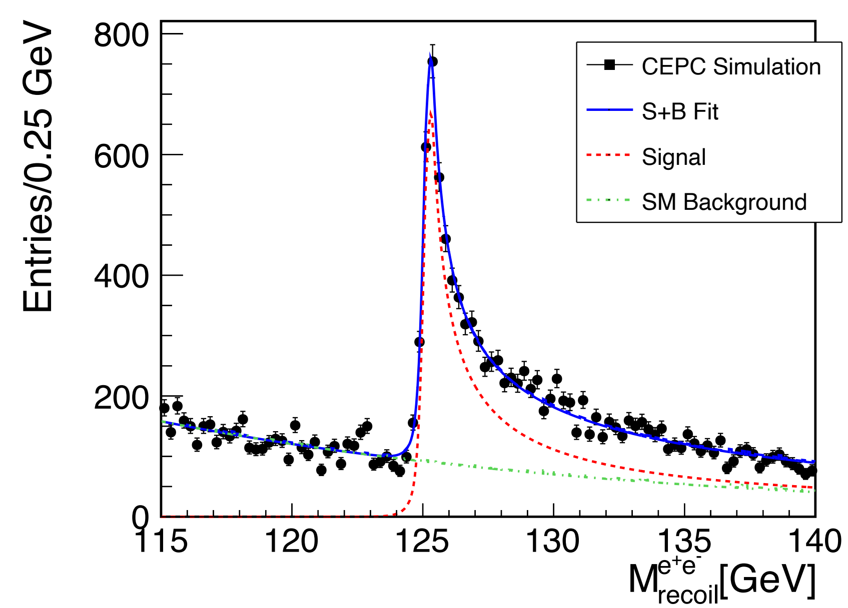


Template-Recoil Mass Combined Fit

Assuming lepton pair's recoil mass and jet flavor are independent in signal

$$\mathrm{PDF}^{3D}(X_B, X_C, M_{\mathrm{recoil}}) = \mathrm{PDF}^{flavor}(X_B, X_C) \times \mathrm{PDF}^{\mathrm{recoil\_mass}}(M_{\mathrm{recoil}})$$





- The shape parameter of recoil mass in signal and dominate background are float in the fit
- Reduce the dependency to the MC prediction
- Effects of systematic uncertainty also considered

Current Results

Combination of the 4 channels:

Statistic precision of σ(ZH)*Br(H->bb/cc/gg) is 0.3% 3.3% and 1.3%

Consistent with the goal expected in pre-CDR with full simulation samples

Decay mode	$\sigma(ZH) \times$ BR	BR
$H \to b\bar{b}$	0.28%	0.57%
$H \to c\bar{c}$	2.2%	2.3%
$H \to gg$	1.6%	1.7%

llH with 3D fit and systematic uncertainties considered:

Table 2. Uncertainties of $H \to b\bar{b}$, $H \to c\bar{c}$ and $H \to gg$

	$\mu^+\mu^-H$			e^+e^-H		
	$H \to b\bar{b}$	$H \to c\bar{c}$	$H \to gg$	$H \to b\bar{b}$	$H \to c\bar{c}$	$H \to gg$
Statistic Uncertainty	1.1%	10.5%	5.4%	1.6%	14.7%	10.5%
Fixed Background	-0.2% +0.1%	+4.1% -4.2%	7.6%	-0.2% +0.1%	+4.1% -4.2%	7.6%
Event Selection	+0.7% -0.2%	+0.4% -1.1%	+0.7% -1.7%	+0.7% -0.2%	+0.4% -1.1%	+0.7% -1.7%
Flavor Tagging	-0.4% +0.2%	+3.7% -5.0%	+0.2% -0.7%	-0.4% +0.2%	+3.7% -5.0%	+0.2% -0.7%
Non uniformity	$< 0.1\%$			$< 0.1\%$		
Combined Systematic Uncertainty	+0.7% -0.5%	+5.5% -6.6%	+7.6% -7.8%	+0.7% -0.5%	+5.5% -6.6%	+7.6% -7.8%

Analysis with more reliable approaches. Systematic uncertainties considered.

Study on New Data Sets

Outlook of New Data Sets

- Full simulation events with $\sqrt{s}$ = 240 GeV and new geometry
- Much larger statistics
- High performance PID based on LICH
- So far only llH channel are studied
    - Only consider Higgs and irreducible backgrounds

Soft lepton treatment

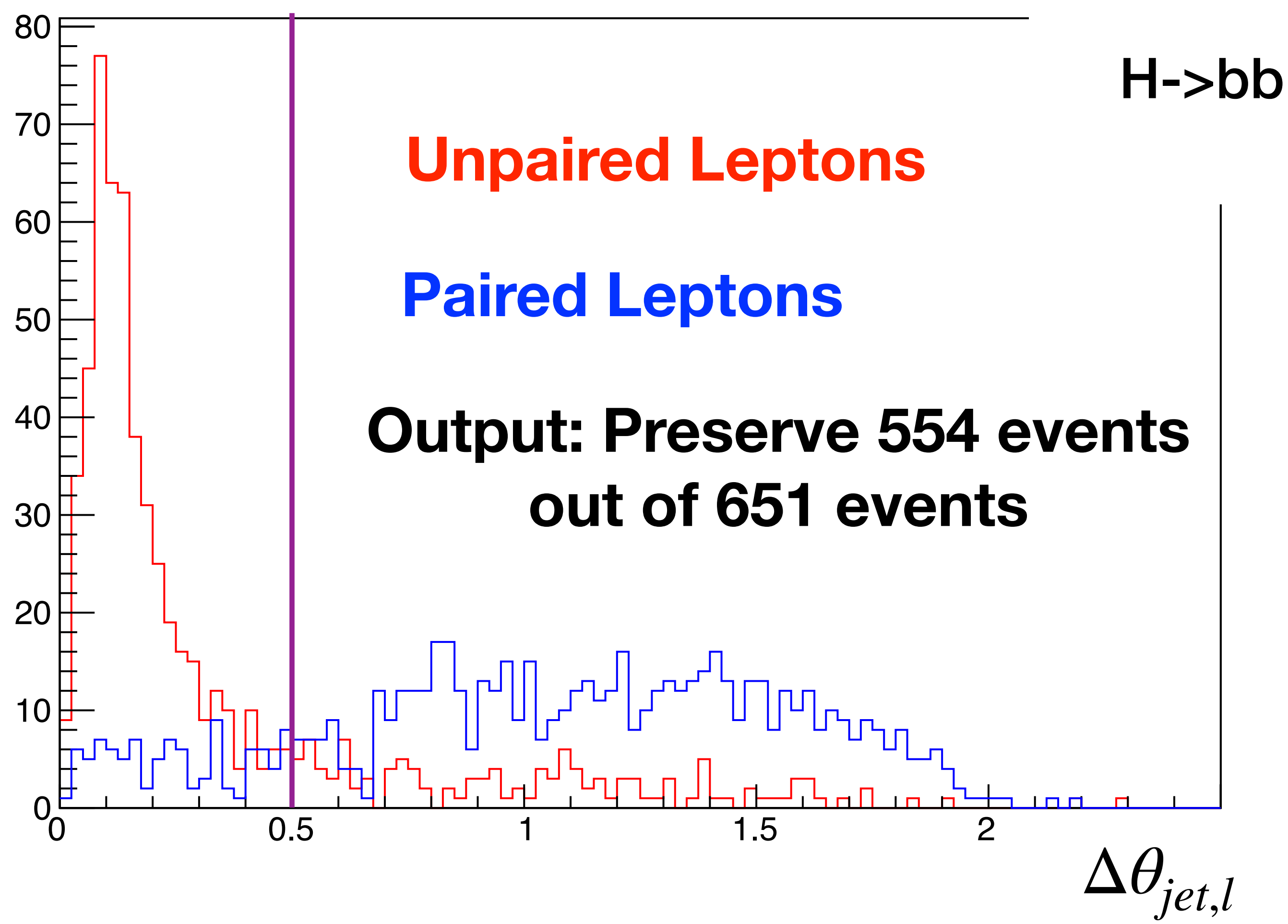
	H->bb		H->cc		H->gg	
	Old	New	Old	New	Old	New
Soft muon	0.40%	0.87%	0.96%	0.15%	0.03%	0.06%
Soft Electron	0.41%	1.99%	1.43%	2.22%	0.08%	2.19%

- **These leptons are here due to heavy flavor hadron decay or fake PID**
- **They background to the analysis**
- **Removal of these leptons can cause the drop in efficiency**

Lepton Pair Selection

- At least one pair of $\mu^+\mu^-$ or e^+e^-
- If there are additional muon/electron in $\mu^+\mu^-H$ or e^+e^-H channel:
  - Reject the pair which is out of recoil mass and Z mass window
  - The unpaired lepton should be close to a jet $\Delta\theta_{l,jet} < 0.5$
  - Jets absorb the closet unpaired leptons (in 4-momentum)

Lepton Pair Selection



Most of unpaired Leptons are from jets

Template Fit

$$
\begin{aligned}
&L(M^{l\bar{l}}_{\text{recoil}}, X_B, X_C; \vec{\theta}_s, \vec{\theta}_b, N^{\text{sig}}_{H\to b\bar{b}}, N^{\text{sig}}_{H\to c\bar{c}}, N^{\text{sig}}_{H\to gg}, N^{\text{bkg}}_{\text{irred\_}b\bar{b}}, N^{\text{bkg}}_{\text{irred\_}c\bar{c}}, N^{\text{bkg}}_{\text{irred\_}uds}, N^{\text{bkg}}_{l^+l^-H}, N_{\text{redu}}) \\
&= P_{\text{sig}}(M^{l\bar{l}}_{\text{recoil}}; \vec{\theta}_s)(N^{\text{sig}}_{H\to b\bar{b}} P^{H\to b\bar{b}}_{\text{flavor}}(X_B, X_C) + N^{\text{sig}}_{H\to c\bar{c}} P^{H\to c\bar{c}}_{\text{flavor}}(X_B, X_C) + N^{\text{sig}}_{H\to gg} P^{H\to gg}_{\text{flavor}}(X_B, X_C) + N^{\text{bkg}}_{l^+l^-H} P^{H\to \text{other}}_{\text{flavor}}(X_B, X_C)) \\
&+ P_{\text{irred}}(M^{ll}_{\text{recoil}}; \vec{\theta}_b)(N^{\text{bkg}}_{\text{irred\_}b\bar{b}} P^{\text{irred\_}bb}_{\text{flavor}}(X_B, X_C) + N^{\text{bkg}}_{\text{irred\_}c\bar{c}} P^{\text{irred\_}c\bar{c}}_{\text{flavor}}(X_B, X_C) + N^{\text{bkg}}_{\text{irred\_}uds} P^{\text{irred\_}uds}_{\text{flavor}}(X_B, X_C)) \\
&+ N_{\text{redu}} P_{\text{redu}}(M^{l\bar{l}}_{\text{recoil}}, X_B, X_C).
\end{aligned}
$$

Signal Recoil mass line shape

Signal Recoil Flavor Template

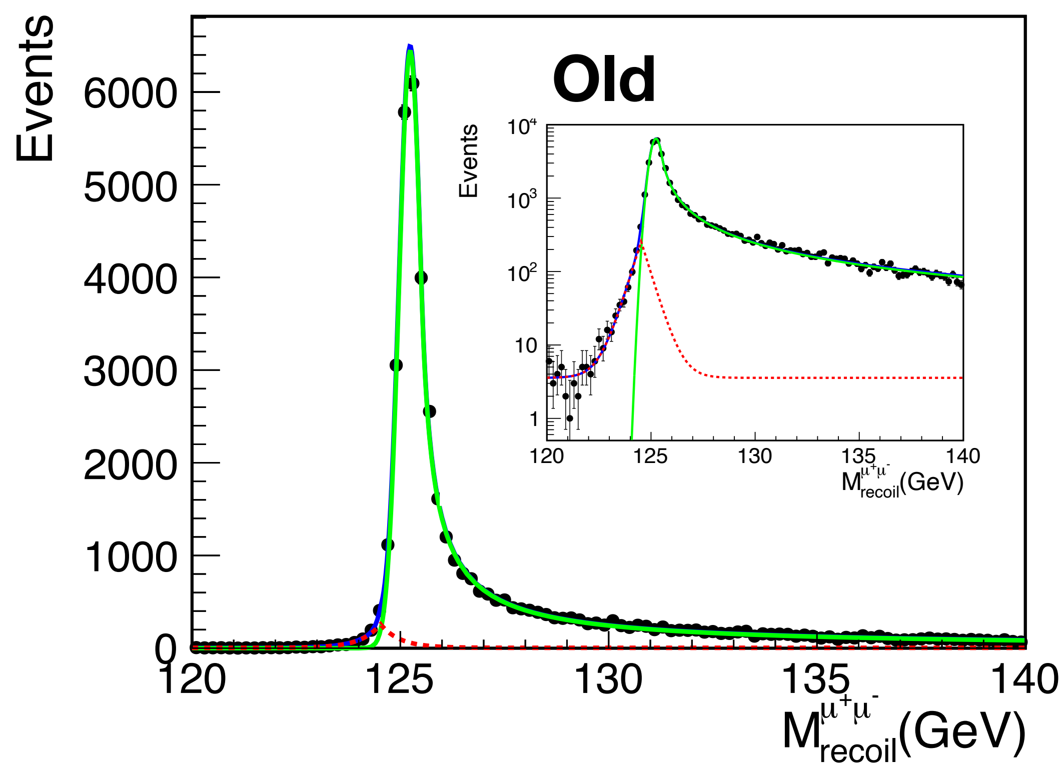
Background Recoil mass Lineshape

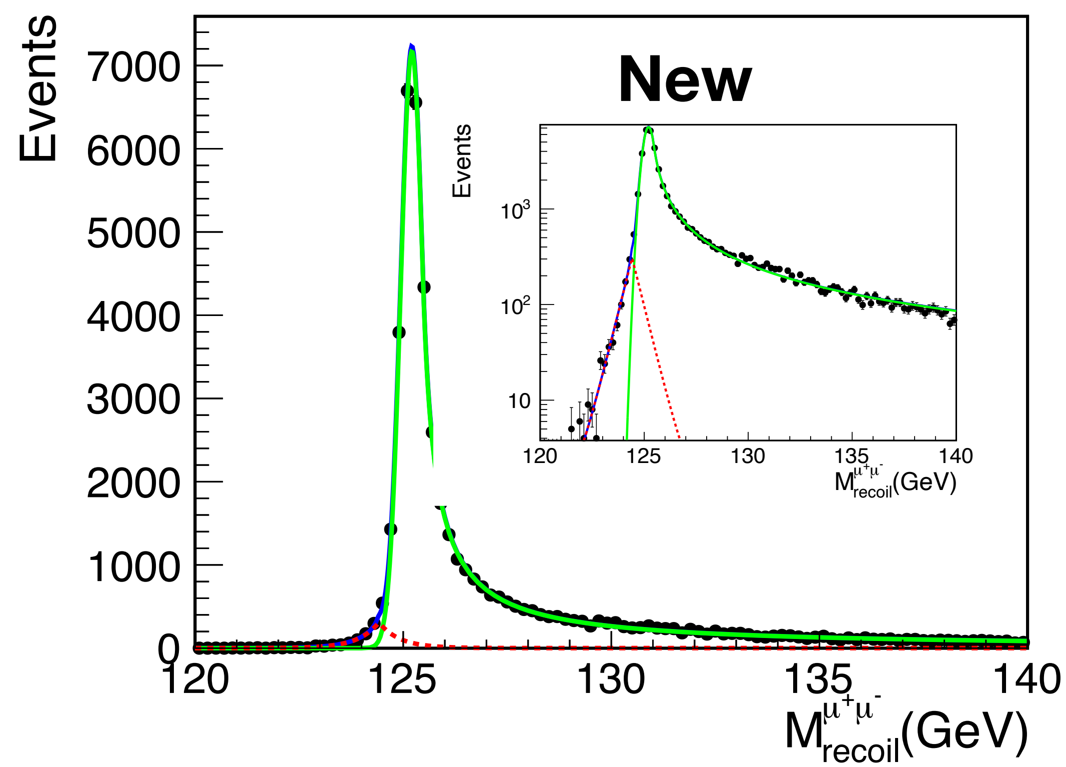
Background Flavor Template

- Signal Recoil line shape: Crystal ball function + exponential tail
- Background line shape: First order Chebychev polynomial (each flavor has different shape parameters)
- All the line shape parameters are float in the fit, as as event yields (Expect H->ww/ZZ)

Signal Line Shape of lepton recoil system mass: μμH

Signal Recoil Mass: Described by a Crystal ball function + double sided exponential Head



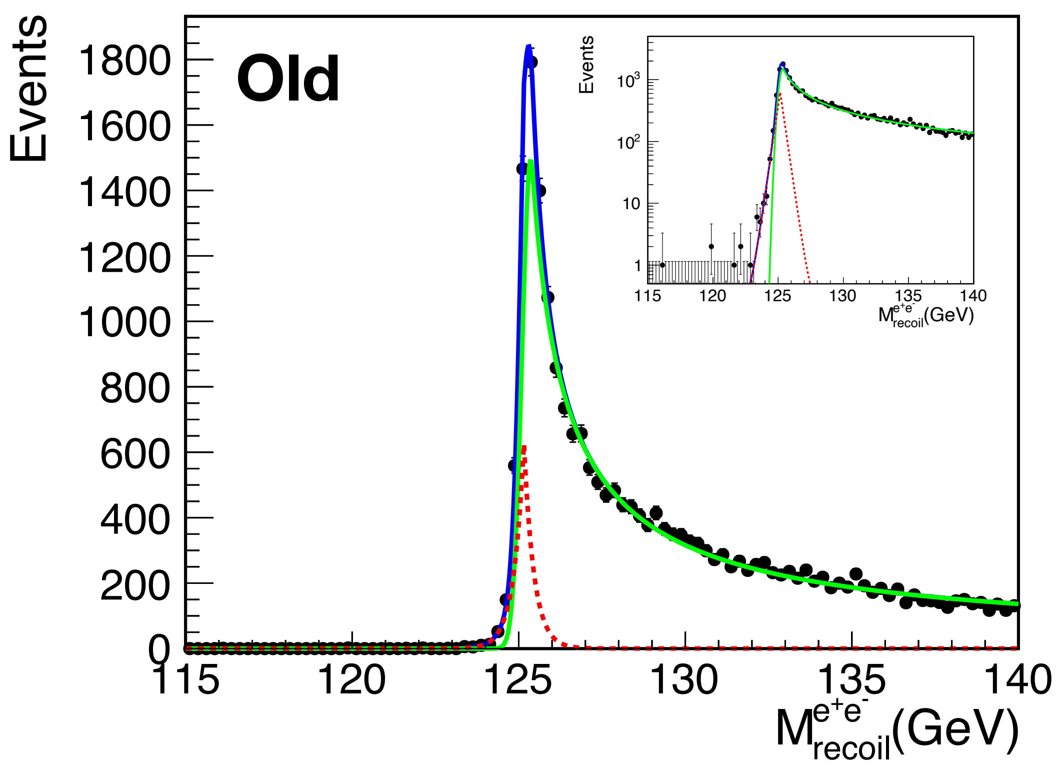


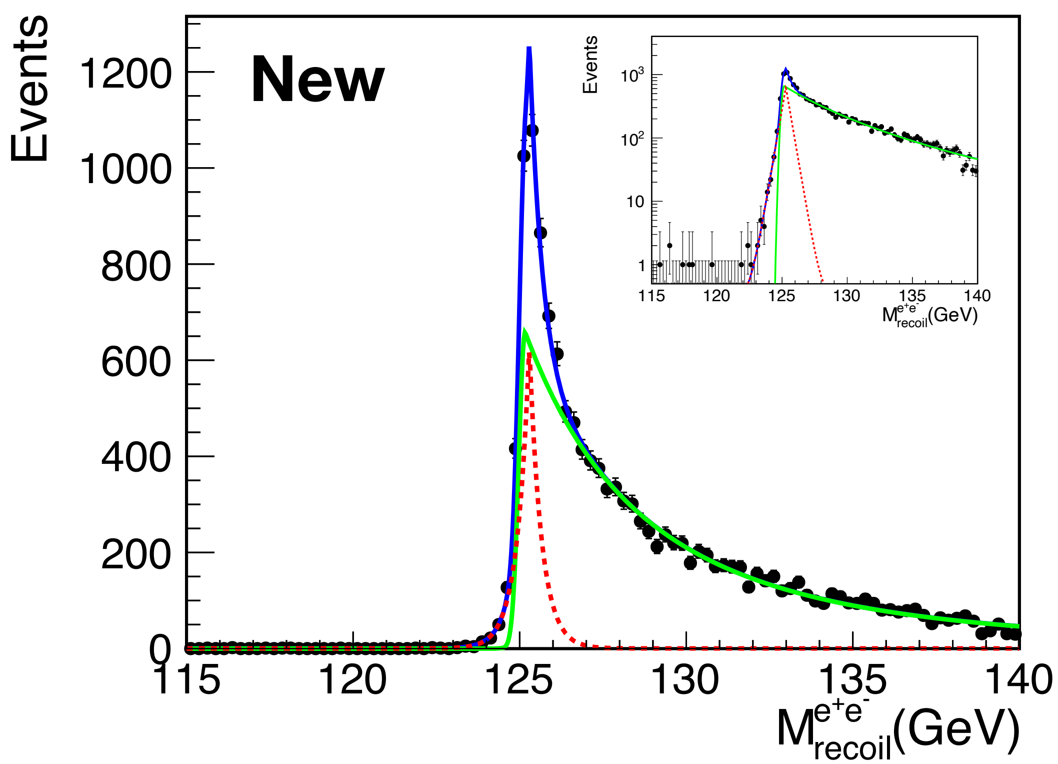
CB Parameters	old	new
a	-1.010 ± 0.016	-0.963 ± 0.014
n	0.939 ± 0.013	0.999 ± 0.012
sigma	0.269 ± 0.003	0.270 ± 0.003
mean	125.237 ± 0.003	125.210 ± 0.003

Head Parameters	old	new
peak	125.154 ± 0.05	124.42 ± 0.026
tau	0.465 ± 0.047	0.498 ± 0.024

Signal Line Shape of lepton recoil system mass: μμH

Signal Recoil Mass: Described by a Crystal ball function + double sided exponential Head

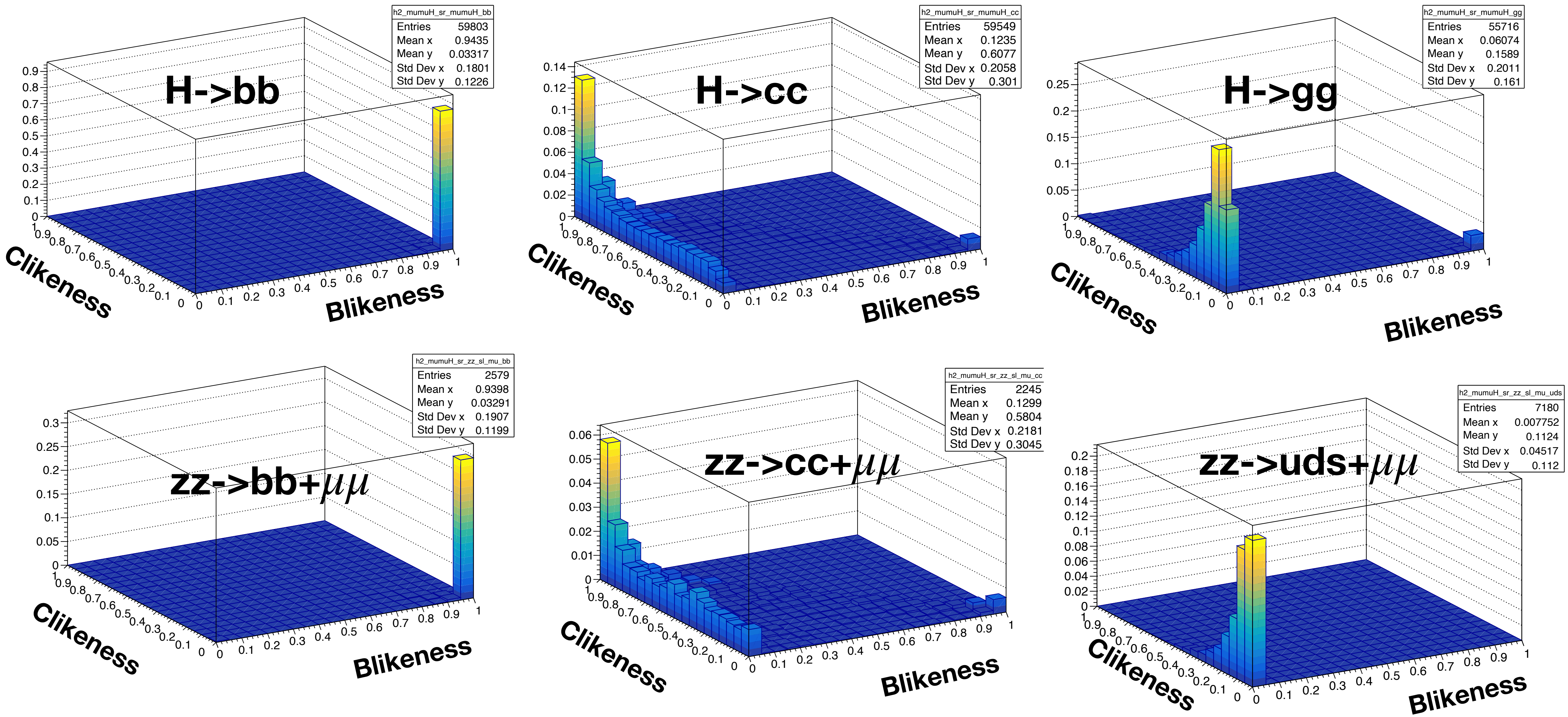




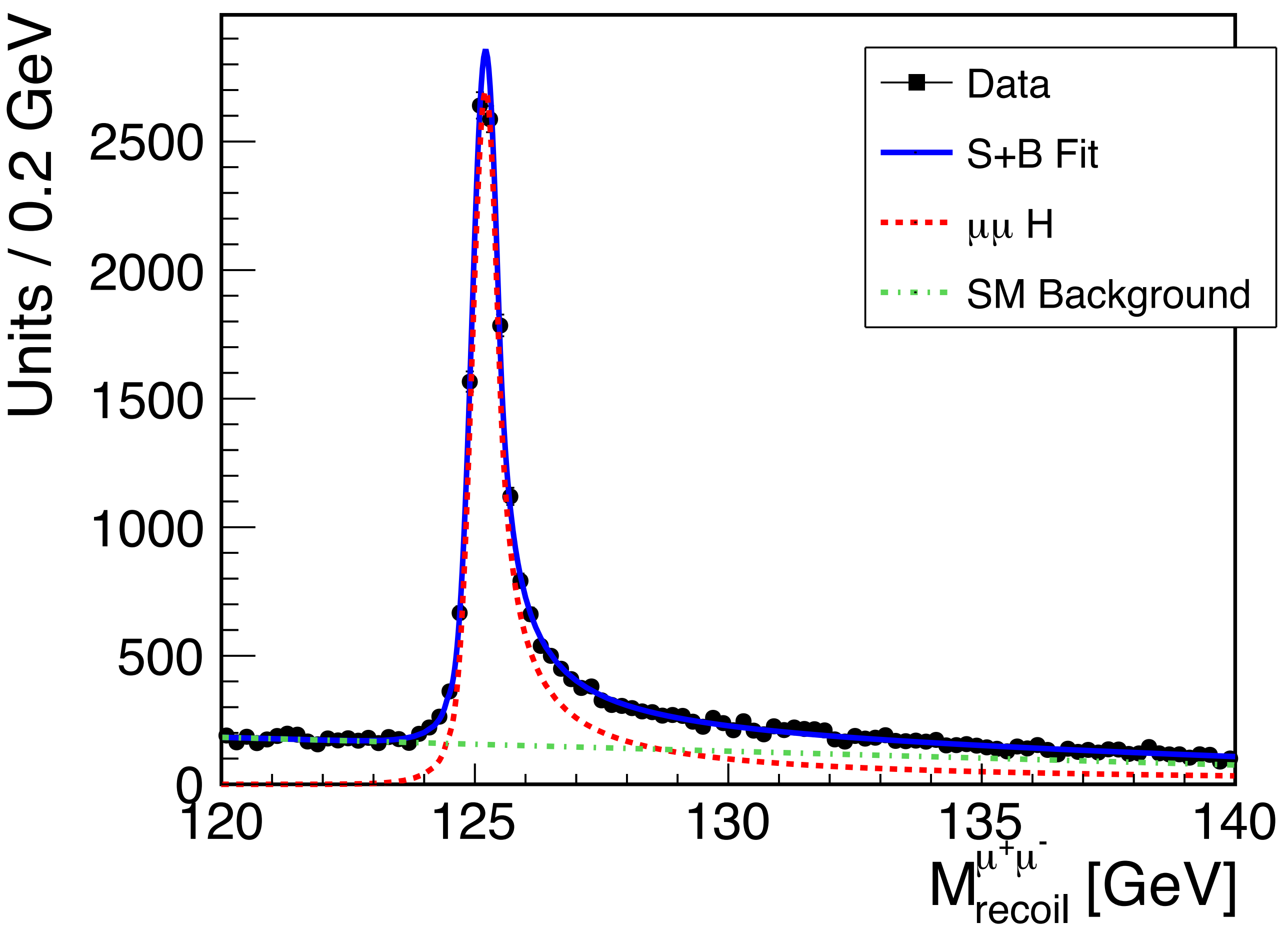
CB Parameters	old	new
a	-0.258±0.037	-0.053±0.008
n	0.800±0.032	3.03±0.46
sigma	0.260±0.013	0.186±0.020
mean	125.359±0.041	125.163±0.027

Head Parameters	old	new
peak	125.52±0.05	124.48±0.02
tau	0.465±0.047	0.434±0.022

Flavor Template



Results of Template Fit: μμH



Signal Variable	Fitted Value
$\mu_{\sigma(H\to bb)}$	0.989±0.012
$\mu_{\sigma(H\to cc)}$	0.987±0.11
$\mu_{\sigma(H\to gg)}$	1.04±0.061
a	-0.985±0.026
n	0.975±0.027
sigma	0.269±0.006
mean	125.214±0.005

old results:

	$\mu^+\mu^- H$		
	$H \to b\bar{b}$	$H \to c\bar{c}$	$H \to gg$
Statistic Uncertainty	1.1%	10.5%	5.4%

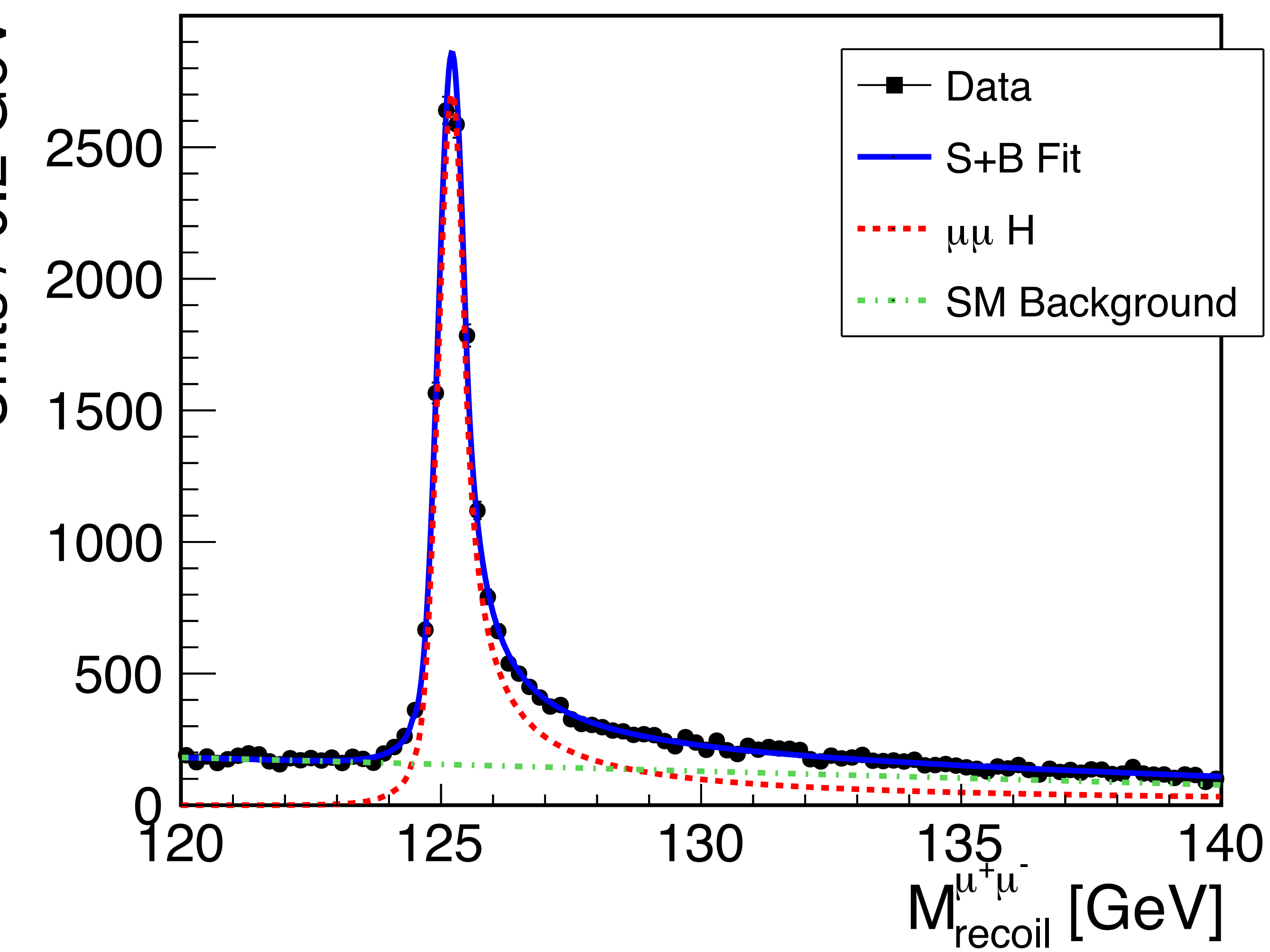
- **Good fit quality**
- **Slightly higher uncertainty than that in previous study**

Plan of Future

What can we learned from the previous analysis, and in which way we can move forward?

H->ww/zz and Hadronic Final states Separation

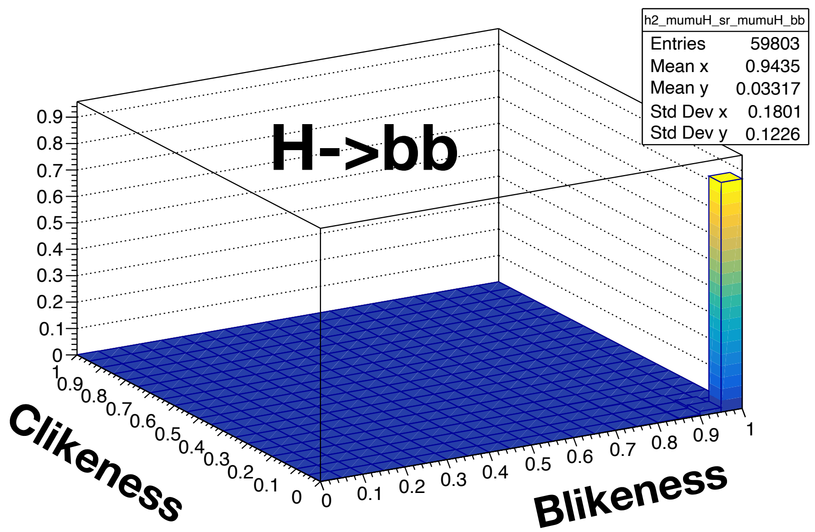
The event yields of H->ww/zz are fixed in the fit.
What happened if we make it float?

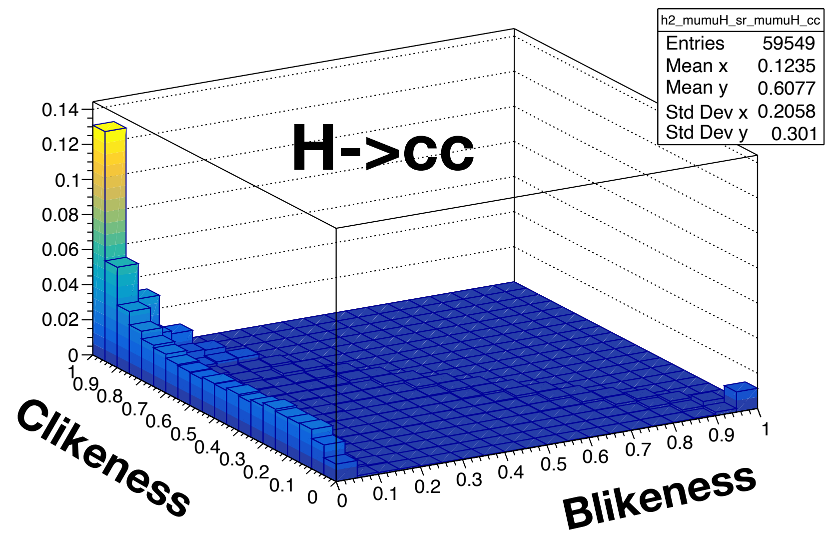


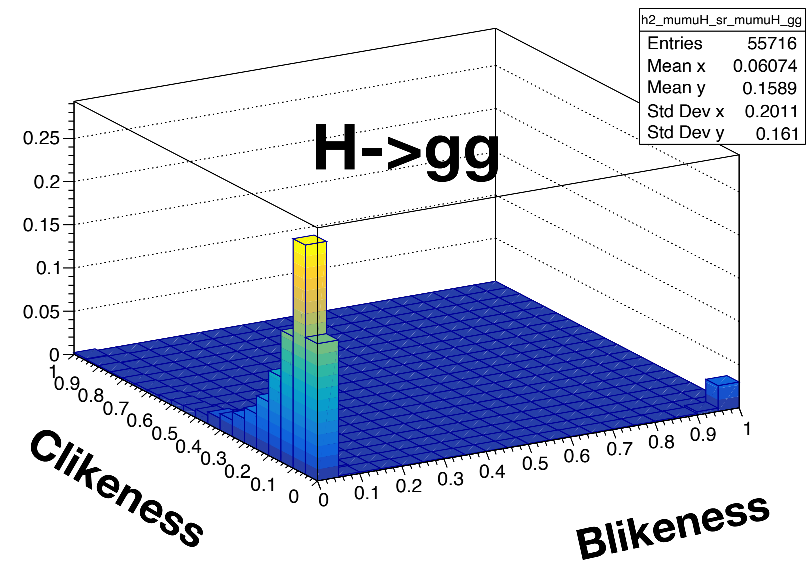
Signal Variable	Fitted Value
$\mu_{\sigma(H\to bb)}$	0.978±0.017
$\mu_{\sigma(H\to cc)}$	0.855±0.162
$\mu_{\sigma(H\to gg)}$	0.851±0.169
a	-0.970±0.032
n	0.990±0.034
sigma	0.268±0.006
mean	125.212±0.006

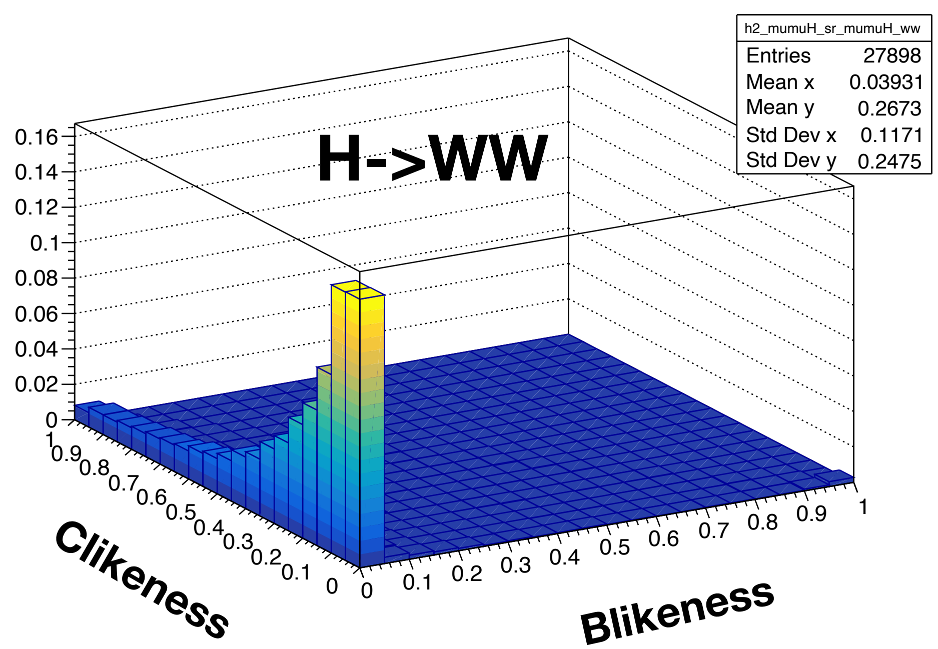
Significant deteriorate in H->cc/gg

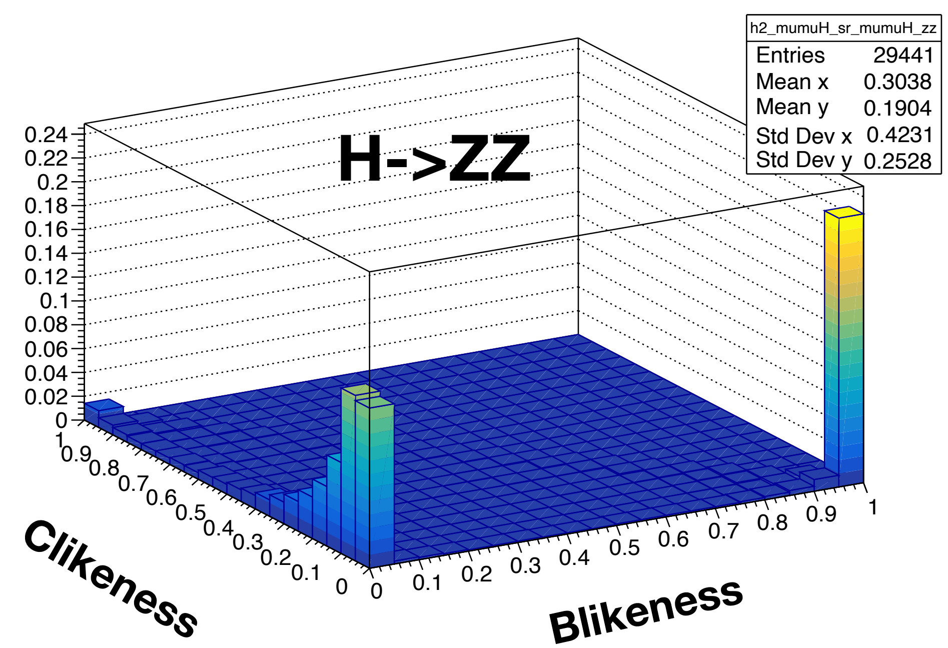
Flavor Template







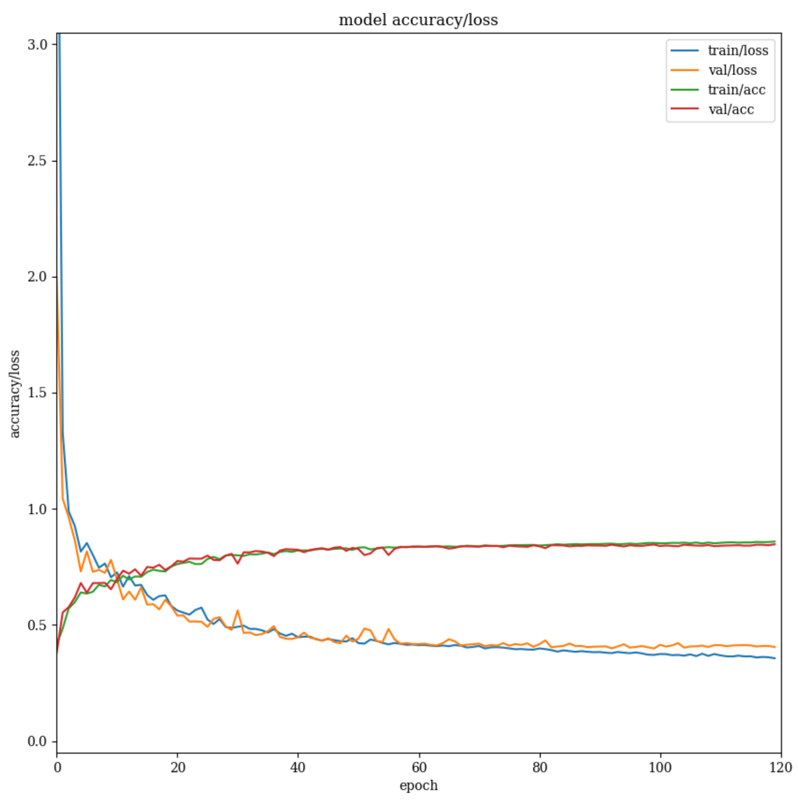


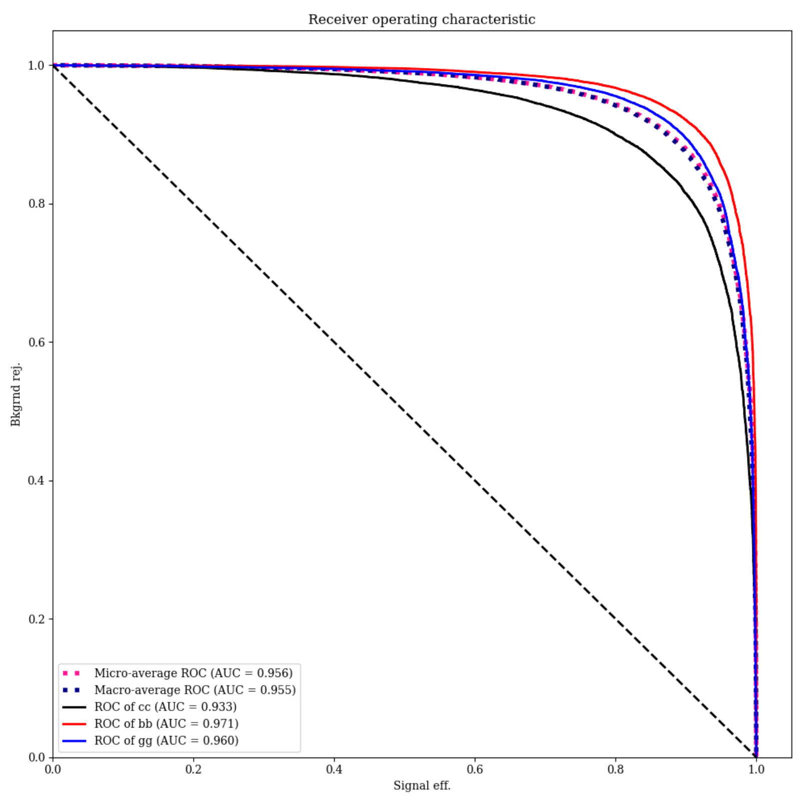


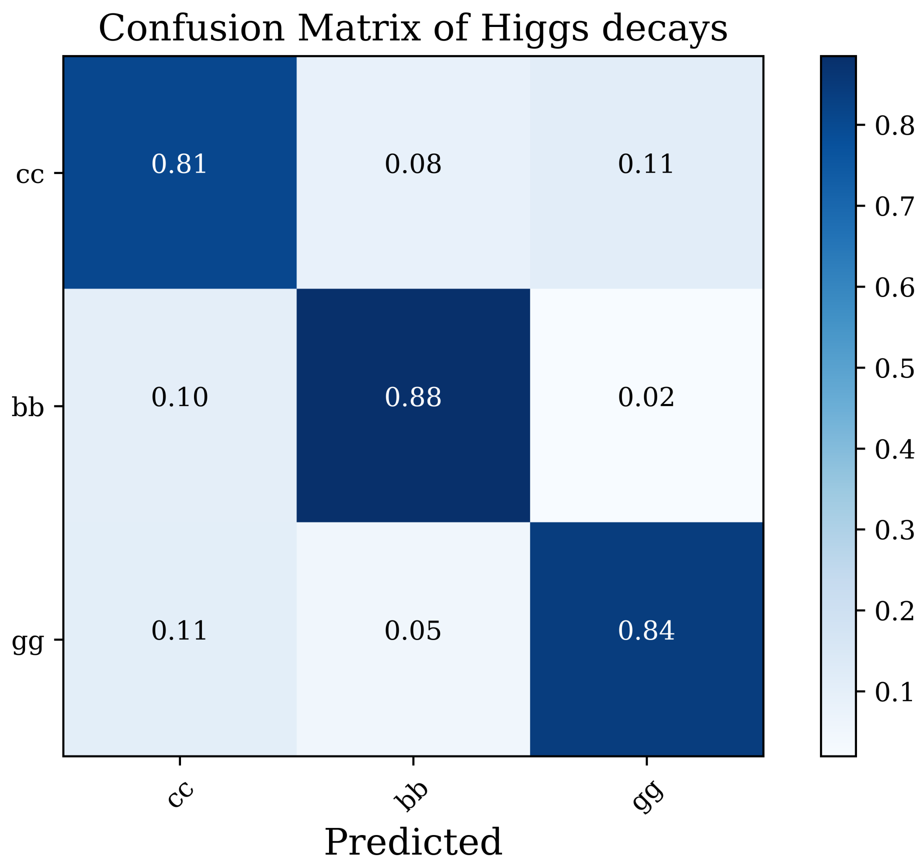
We cannot distinguish H->ww/zz just by flavor

Energy flow + Deep Set To Classify hadronic FS

3 Category: H->bb/cc/gg



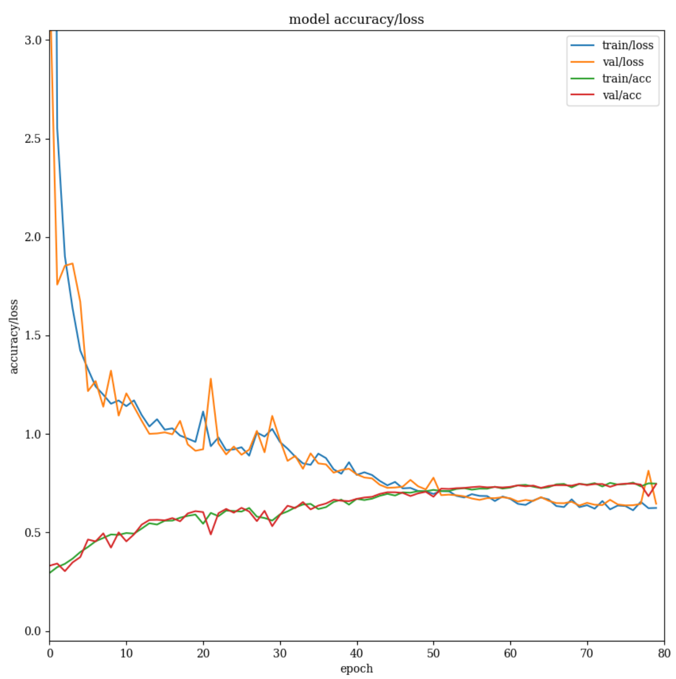


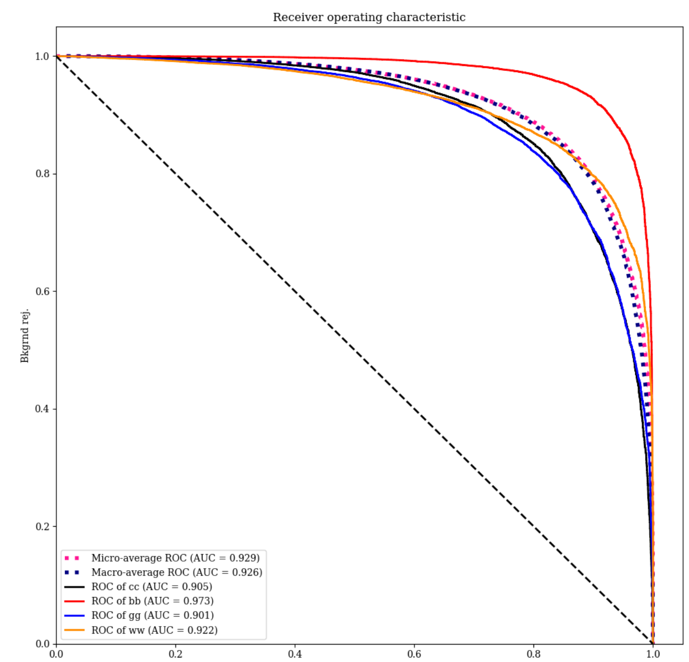


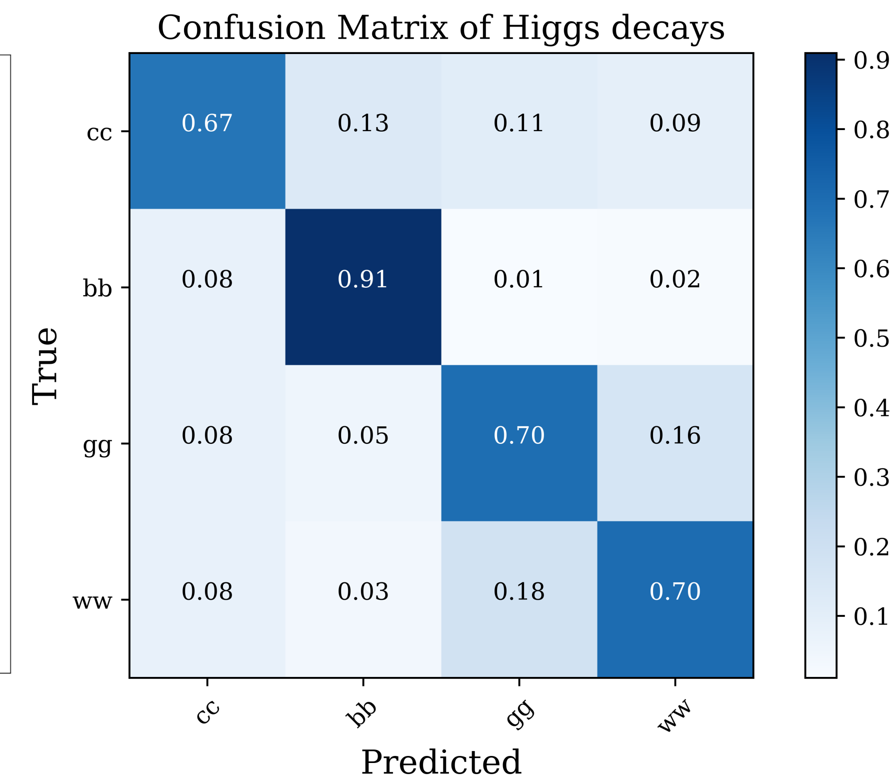
- Use information of 4-momentum, charge and impact parameter of particles in jets, full simulation
- Results seems promising

Energy flow + Deep Set To Classify hadronic FS

4 category : H->bb/cc/gg/ww(hadronic only)



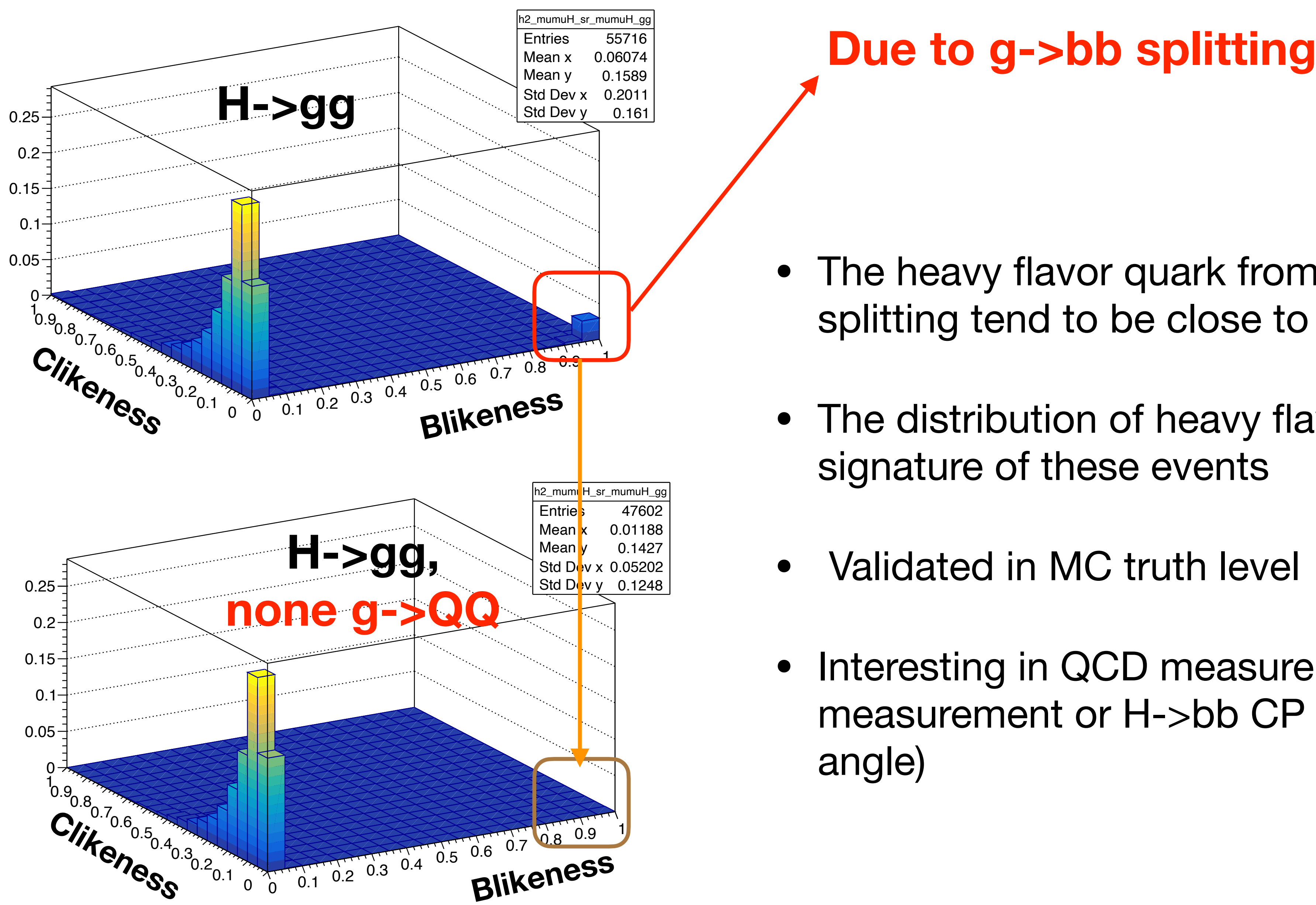




True \ Predicted	cc	bb	gg	ww
cc	0.67	0.13	0.11	0.09
bb	0.08	0.91	0.01	0.02
gg	0.08	0.05	0.70	0.16
ww	0.08	0.03	0.18	0.70

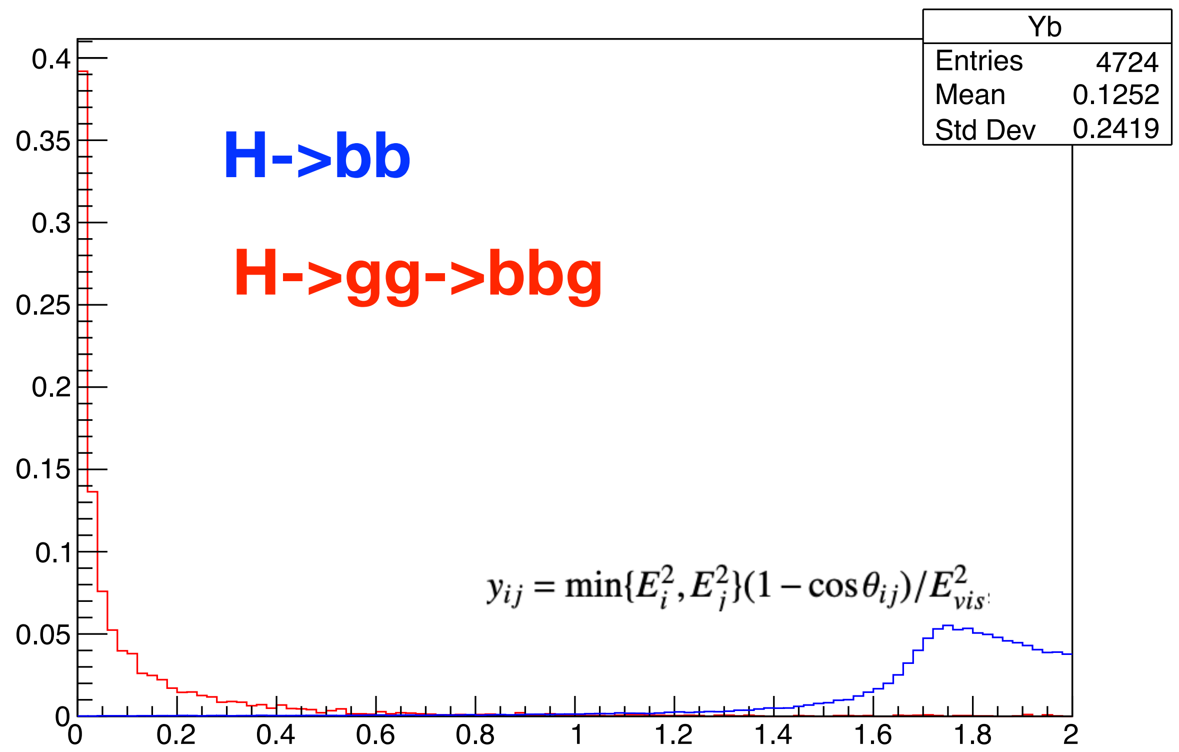
- Use information of 4-momentum, charge and impact parameter of particles in jets, full simulation
- Results seems promising, but LOTS of things to understand

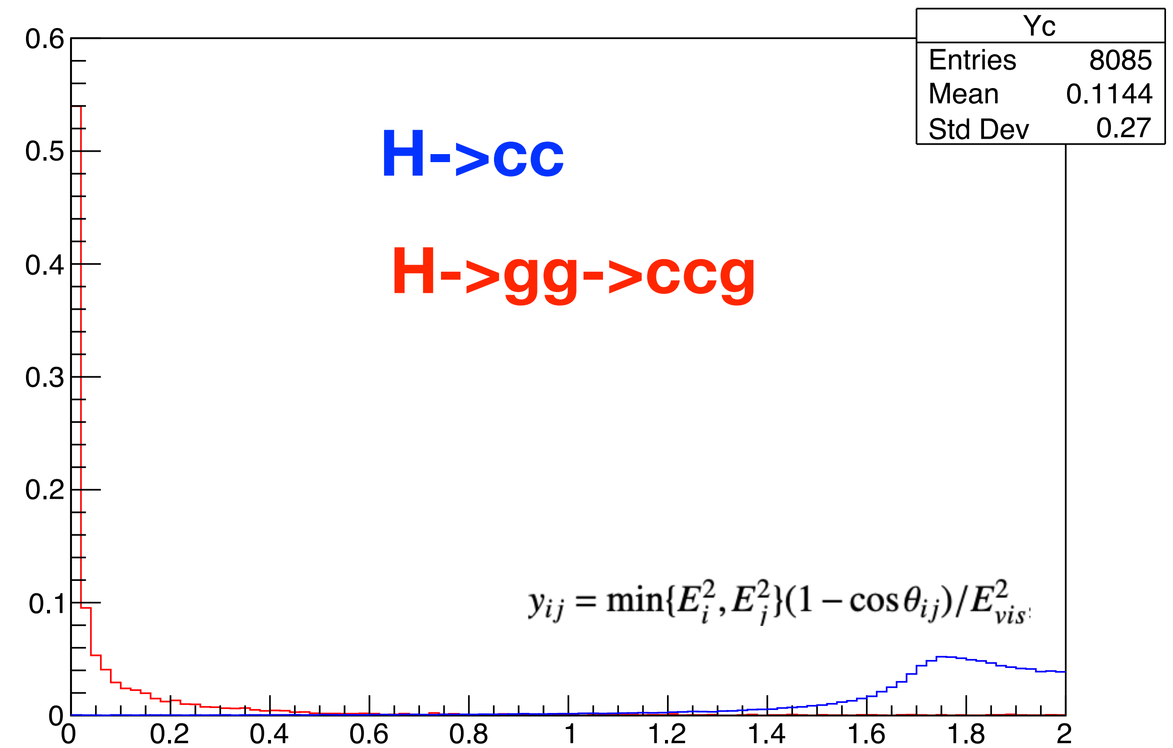
Jet Flavor Substructure



- The heavy flavor quark from gluon splitting tend to be close to each other
- The distribution of heavy flavor components in jets are signature of these events
- Validated in MC truth level
- Interesting in QCD measurement, precision Higgs measurement or H->bb CP phase (interference in small angle)

Heavy Flavor Components in Jet Substrue





- Clustering the have flavor components in truth level by ee-kt algorithm
- Significant difference between H->gg->QQ+g and H->QQ
- Working in reconstructed level now

Lepton Reconstruction with FSR

- Affect on lepton energy/momentum, lepton mass and lepton recoil system mass spectrum
- Affect on the lepton isolation
    - Lower lepton efficiency
- Affect on jet clustering
    - Photons clustered into jets
- Maybe Energy flow network is helpful ?

Summary

- Higgs -> bb/cc/gg has been explored with full simulation sample in most of the Higgs production mode in CDR stage
- Now the ZH-> ll+bb/cc/gg has with 240 GeV full simulation datasets verify the capability of Higgs precision measurement in CEPC
- New result has been improved in lepton pair selection, and modeling of lepton pair recoil system mass
- We are developing method to distinguish different hadronic Higgs decay, based on energy flow and deep set method
    - Hopefully these method can be applied to vvH and qqH bkg/sig speration
    - Hopefully these method can be used to reconstruct electrons by absorbing FS radiations
- We are developing method to not only tag heavy flavor jets, but heavy flavor sub structure in jets

Shall we left all the work to machine?

Thank You!

Backup

Full Result of mumuH Template fit

```
COVARIANCE MATRIX CALCULATED SUCCESSFULLY
FCN=-311062 FROM HESSE     STATUS=OK            229 CALLS        2078 TOTAL
                     EDM=0.031397    STRATEGY= 1      ERROR MATRIX ACCURATE
 EXT PARAMETER                                   INTERNAL      INTERNAL
 NO.   NAME      VALUE            ERROR          STEP SIZE       VALUE
  1  C            2.84275e-05   1.49011e-06   2.11406e-04   2.84275e-05
  2  a           -9.84678e-01   2.91026e-02   4.16962e-04  -9.86277e-02
  3  a1          -4.99944e-01   2.94977e-02   1.31093e-01  -1.58570e+00
  4  a2          -3.45382e-01   4.72285e-02   4.88108e-02  -7.62545e-01
  5  a3          -4.08923e-01   2.13520e-02   2.78585e-02  -9.57657e-01
  6  mean         1.25214e+02   5.38449e-03   1.64950e-03   1.26795e-01
  7  n            9.74954e-01   2.68661e-02   2.76803e-03   3.22181e-01
  8  nHbb         1.25347e+04   1.55693e+02   4.89837e-03   2.56230e-01
  9  nHcc         5.99540e+02   6.77704e+01   9.54471e-03  -1.07645e+00
 10  nHgg         1.77212e+03   1.04261e+02   1.41321e-02  -2.95700e-01
 11  nsig         9.70863e-01   4.50142e-03   1.40460e-02   7.87553e-01
 12  nzzsl_mu_bb    2.90810e+03   1.19358e+02   9.16978e-03  -1.70102e-01
 13  nzzsl_mu_cc    2.34943e+03   7.76258e+01   7.91869e-03  -3.35165e-01
 14  nzzsl_mu_uds     7.62886e+03   1.14594e+02   3.98275e-03  -2.39458e-01
 15  peak         1.24460e+02   3.91876e-02   1.86183e-02  -7.90802e-01
 16  sigma        2.69174e-01   5.81205e-03   4.12916e-03  -6.15130e-01
 17  tau          4.04828e-01   7.33186e-02   2.42746e-02  -8.42000e-01
                               ERR DEF= 0.5
EXTERNAL ERROR MATRIX.    NDIM=  25    NPAR= 17    ERR DEF=0.5
ELEMENTS ABOVE DIAGONAL ARE NOT PRINTED.
```