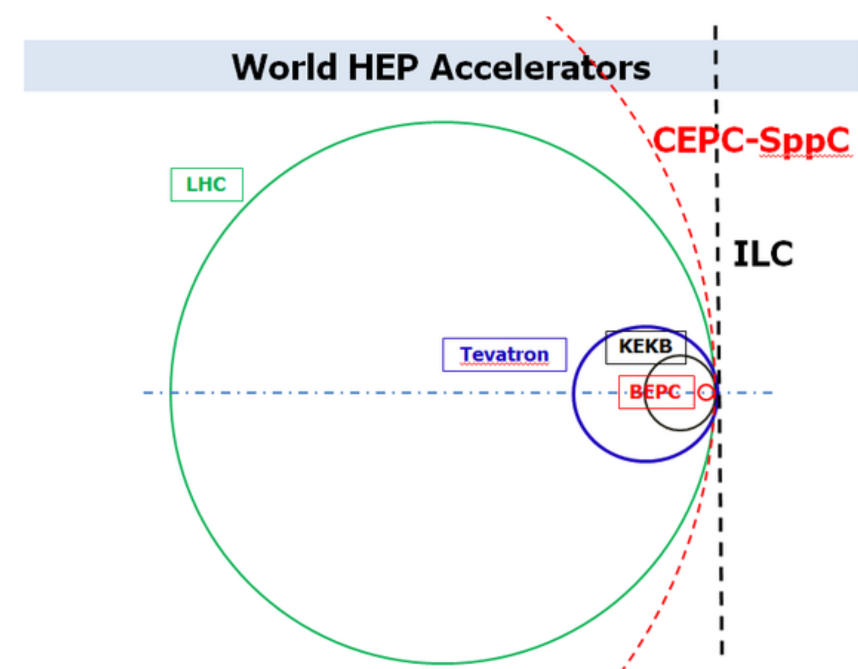
CEPC Workshop

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

Yu Bai (from Southeast University, Nanjing)

On Behalf of CEPC Physics-Software Study Group

November 19, 2019



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

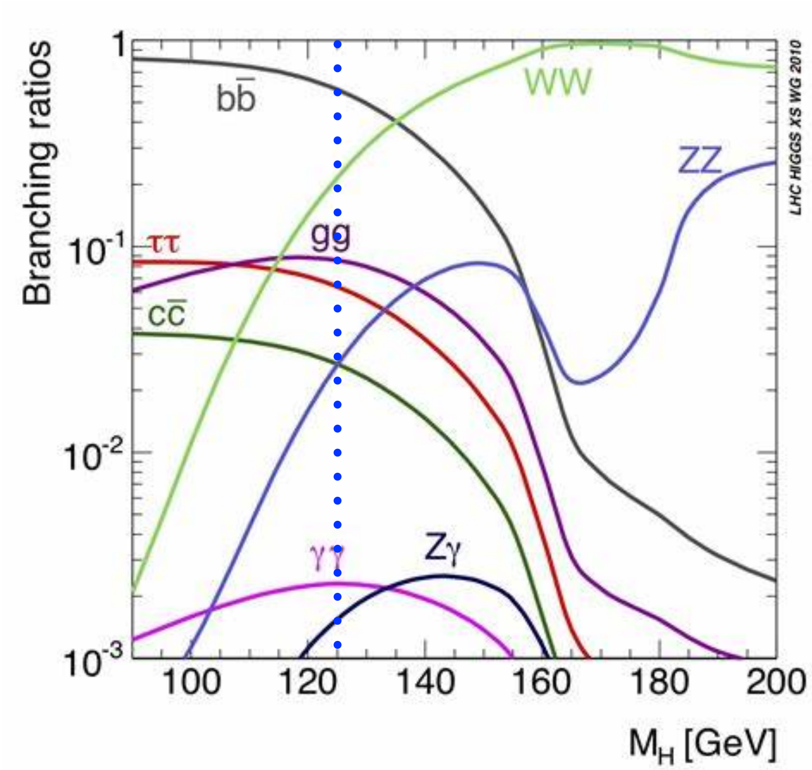
$$
\begin{aligned}
\mathcal{L} = & -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} \\
& + i \bar{\psi} \not{D} \psi + h.c. \\
& + \bar{\psi}_i y_{ij} \psi_j \phi + h.c. \\
& + |D_\mu \phi|^2 - V(\phi)
\end{aligned}
$$

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

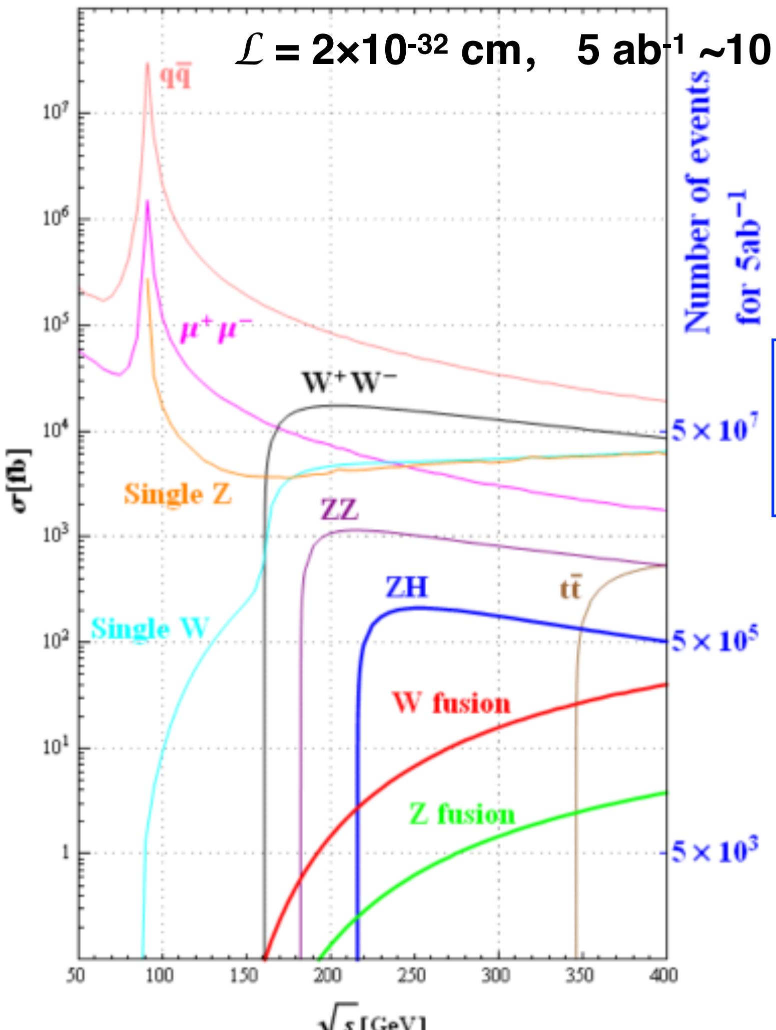
- **H->$\tau\tau$ in ATLAS**
- **H->WW in ATLAS**
- **H-> $\tau\tau$ in CMS**
- **H->$\gamma\gamma$ in CMS**

Uncertainty of O(10%)

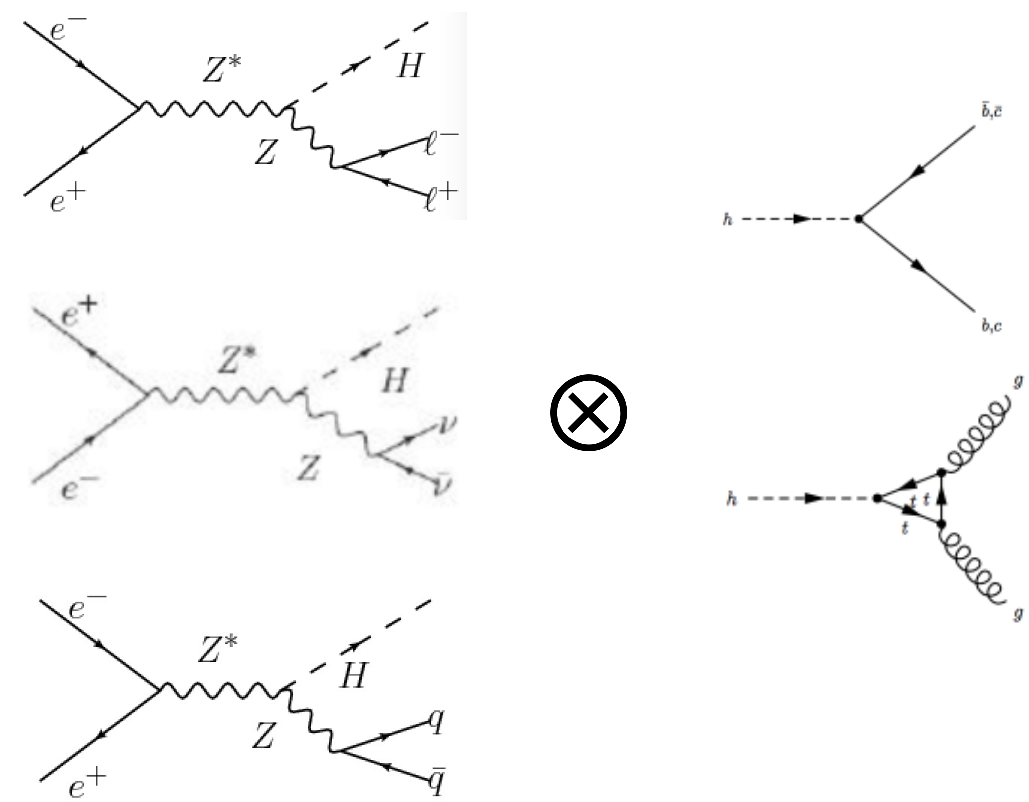
Review of Previous Study

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



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



Statistics normalized to 5/ab:

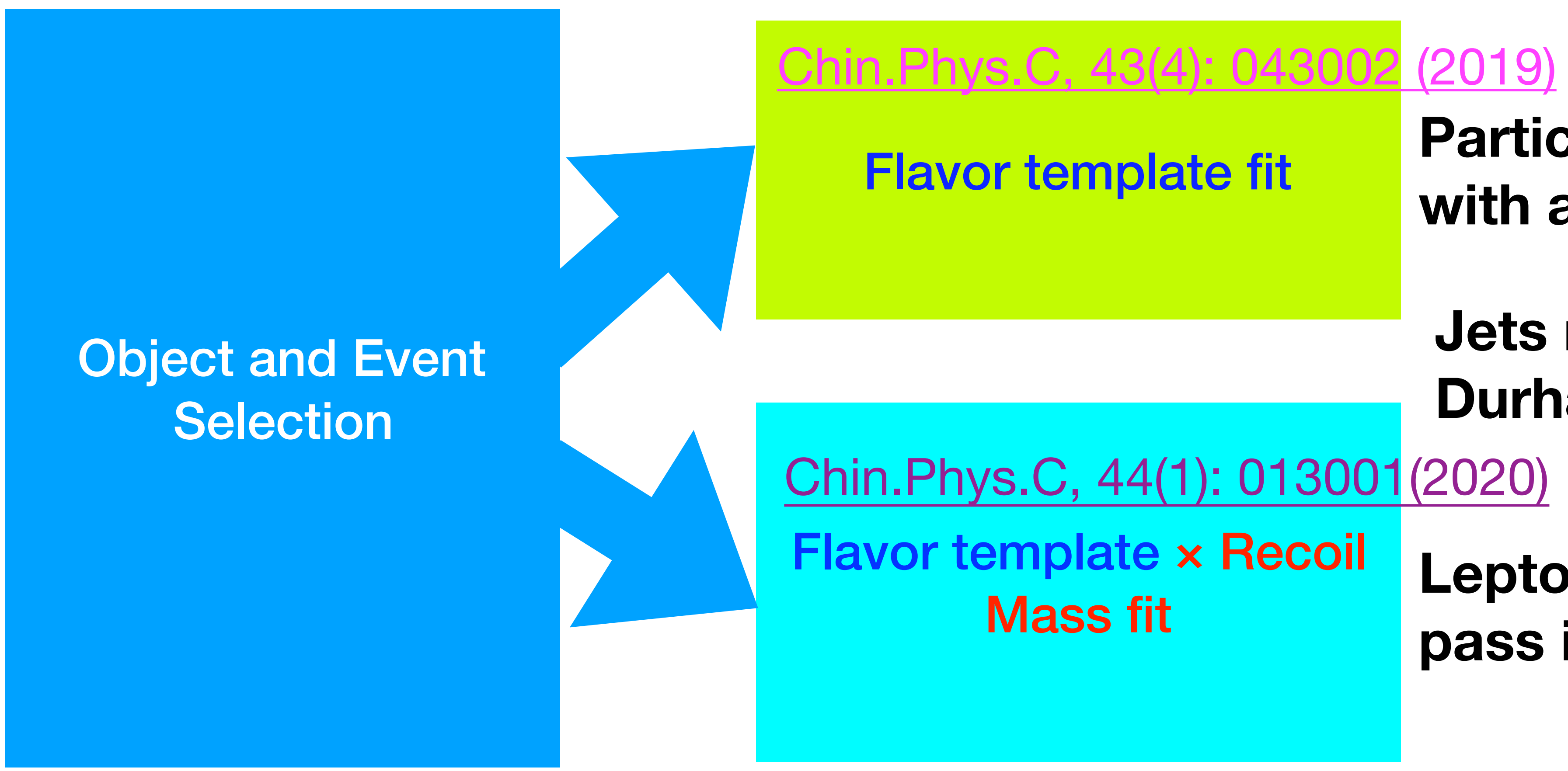
	eeH	$\mu\mu$H	$\tau\tau$H	qqH	$\nu\nu$H
bb	22k	20k	20k	410k	140k
cc	1.0k	1.0k	1.0k	20k	6.7k
gg	3.5k	3.2k	3.2k	65k	22k

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**



Particles are reconstruct with arborPFO algorithm.

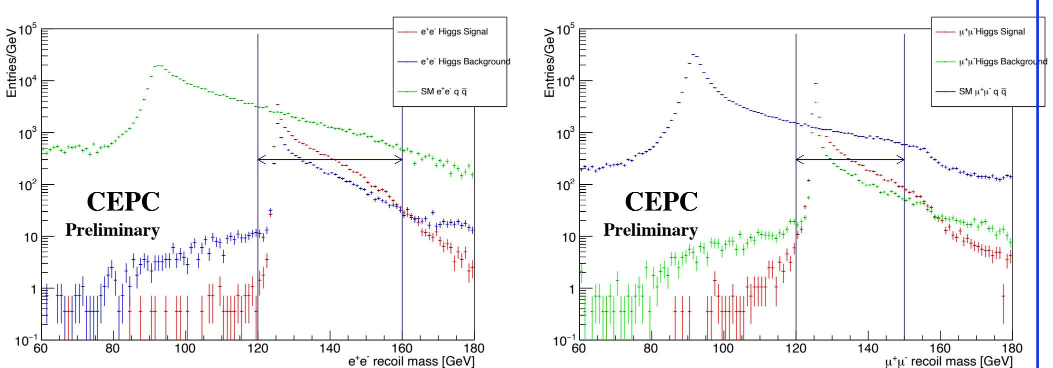
Jets reconstructed by Durham algorithm

Leptons from W/Z need to pass isolation track selection

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

Event Selection

llH



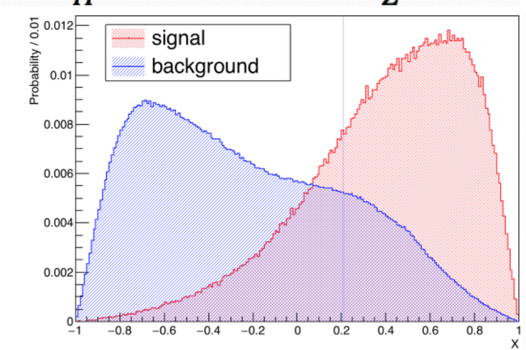
recoil mass is the crucial variable to distinguish signal and background

	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

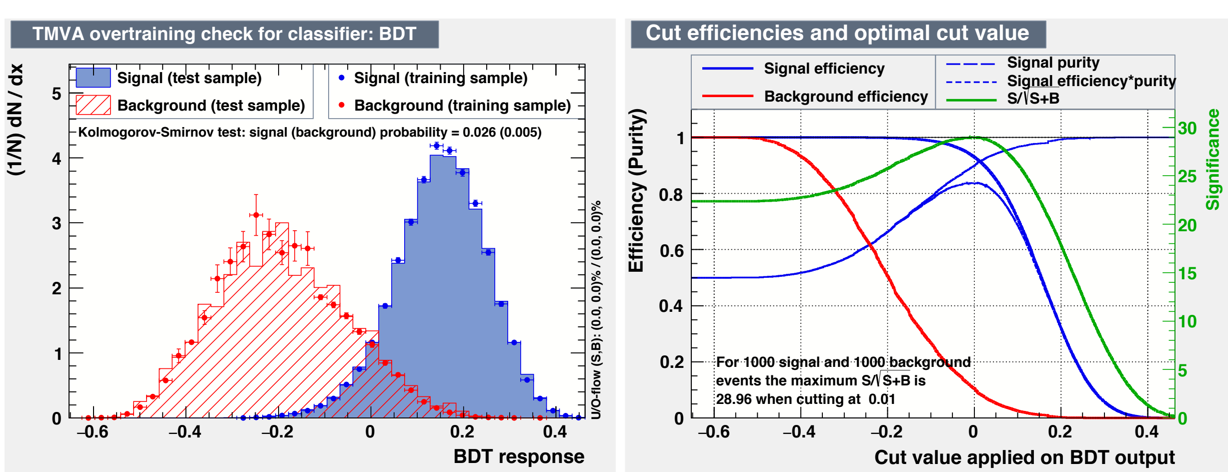
$$\chi^2 = \min_{i \neq j \neq a \neq b} \frac{\left(M_{ij} - M_H\right)^2}{\sigma_H^2} + \frac{\left(M_{ab} - M_Z\right)^2}{\sigma_Z^2}$$

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



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

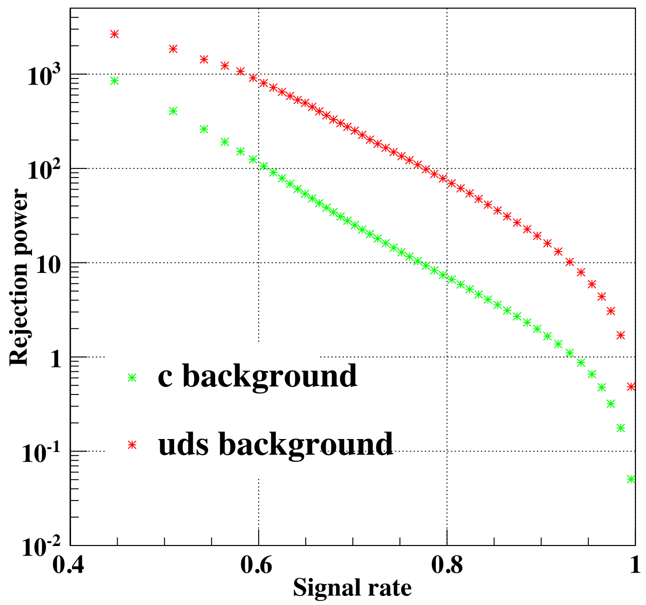
signal	Signal eff	higgs bkg	4f-hadronic	qq	4f-semilept
211.2k	42.8%	32.6k	1.08M	405.6k	0.58k

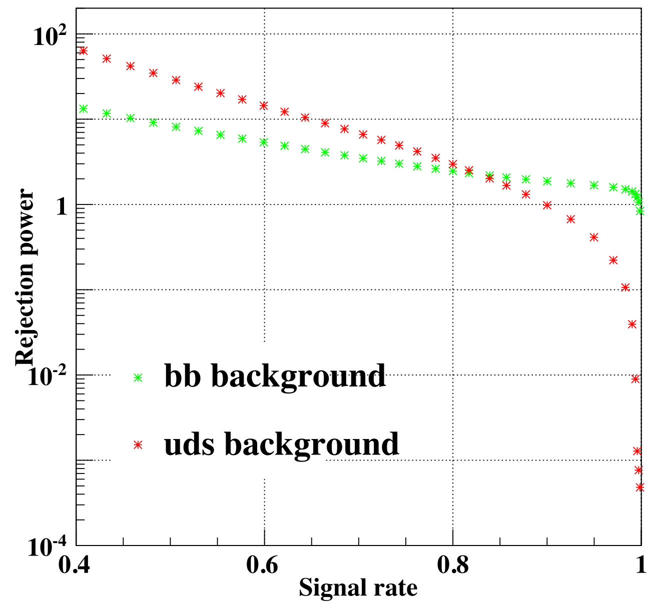
$\nu\nu$H

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

Flavor Tagging and Flavor template

Flavor tagging performance



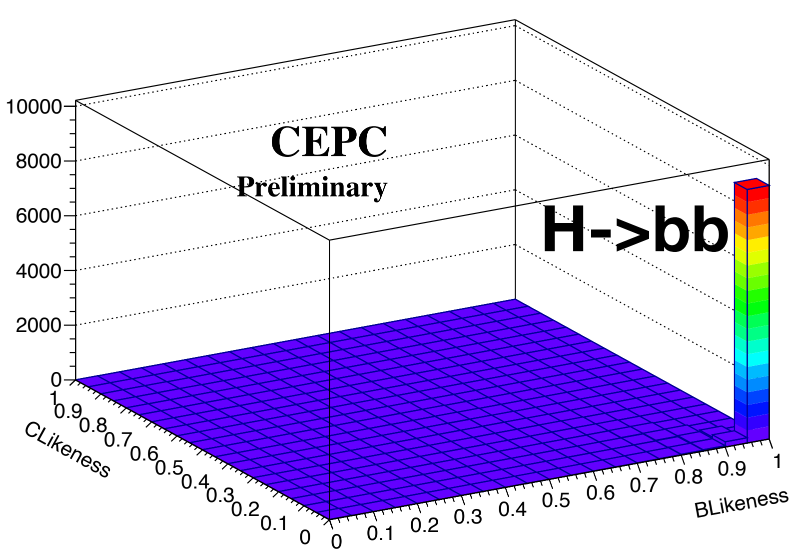


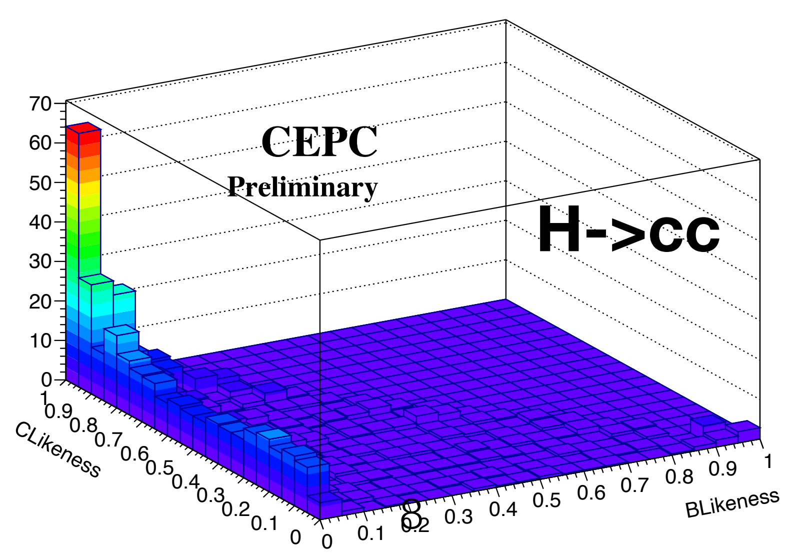
- 4 categories of events according to the soft-leptons and vertices multiplicity
- GBDT with tagging-sensitive variables method applied to each category

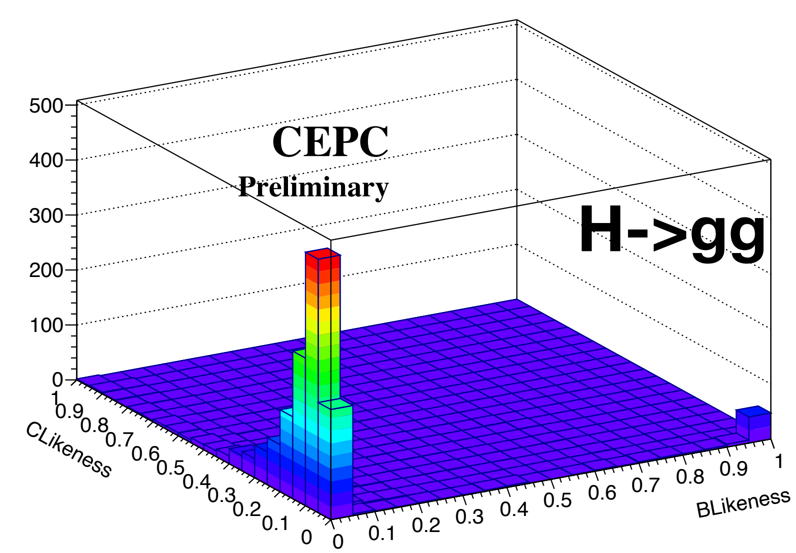
The performance of FT ensures the goal of precision

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

Processes with different flavor components can be separated by template fit



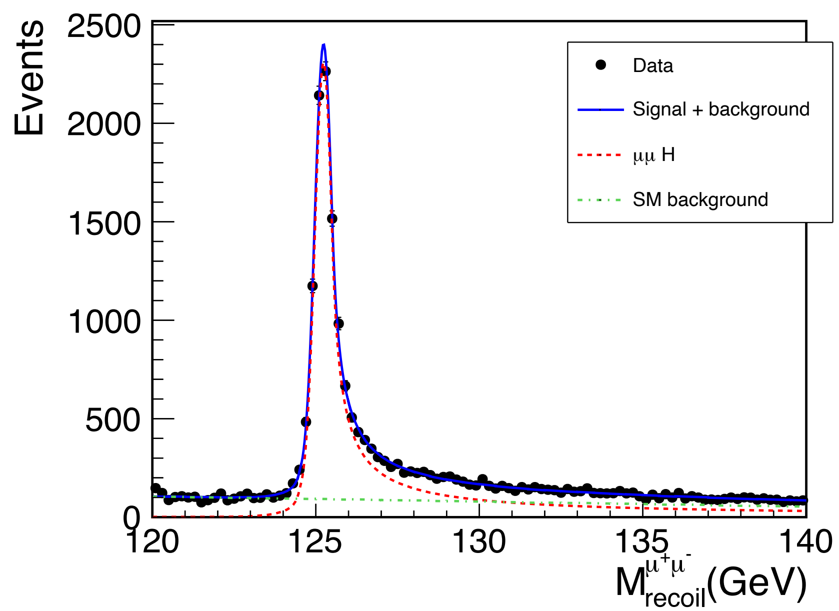


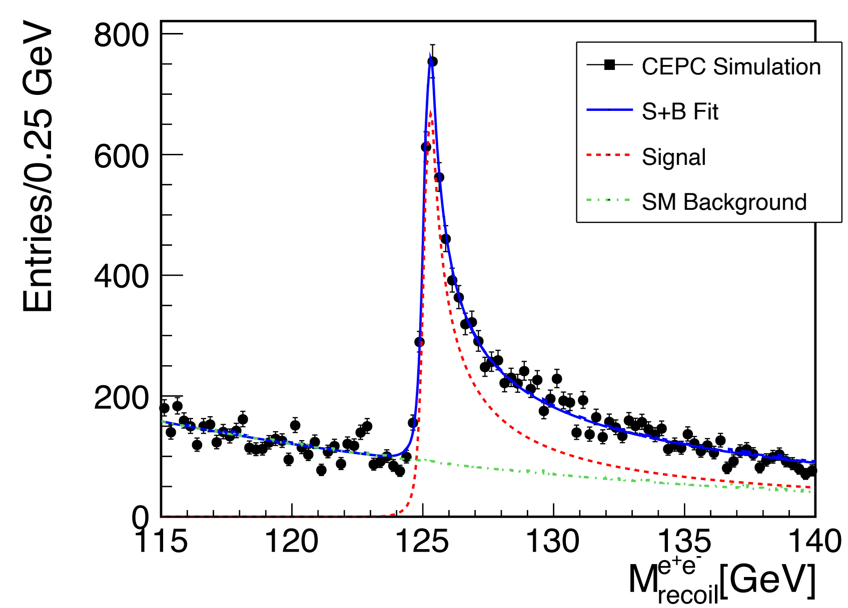


Template-Recoil Mass Combined Fit

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

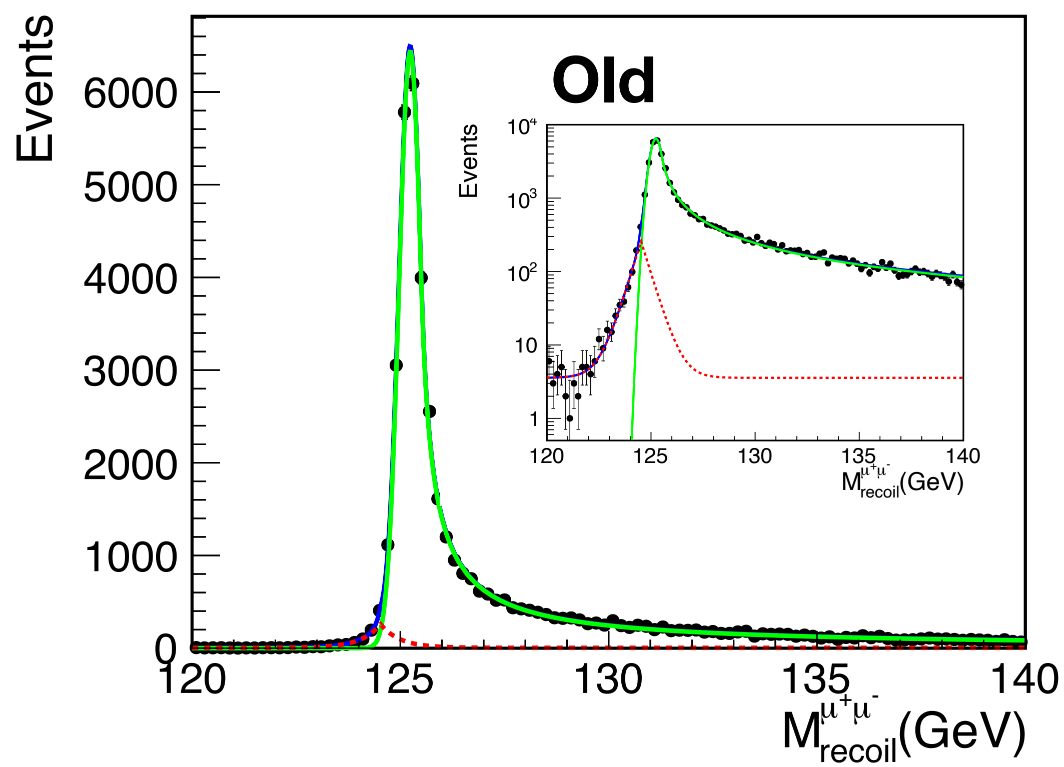
- Full simulation events with $\sqrt{s}$ = 240 GeV and new geometry
- Much larger statistics than previous samples:(in 2 fermions and single W semi-lepton sample)
- High performance PID based on LICH
- So far only llH channel are studied
    - Only consider Higgs and semi-leptonic backgrounds

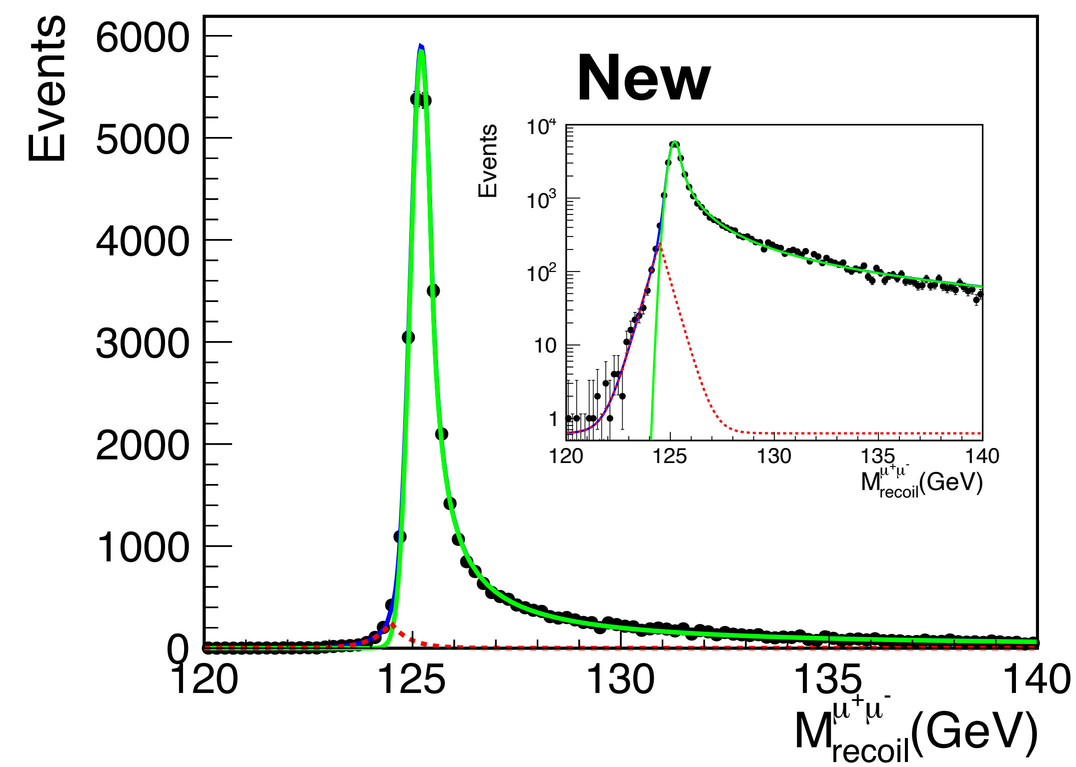
Efficiencies

- **Efficiency of selecting 2 muons dropped from 87% to 75% (we expect efficiency around 95%)**
- **Efficiency of selecting 1 electron and positron dropped from65% to 55% (we expect efficiency around 90%)**
- **Fake leptons in single W semi-lepton sample rising by 50 times(0.1%-5%)**
- **Wrong PID are used…☹️ May be solved soon**

Signal Line Shape of lepton recoil system mass: μμH

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



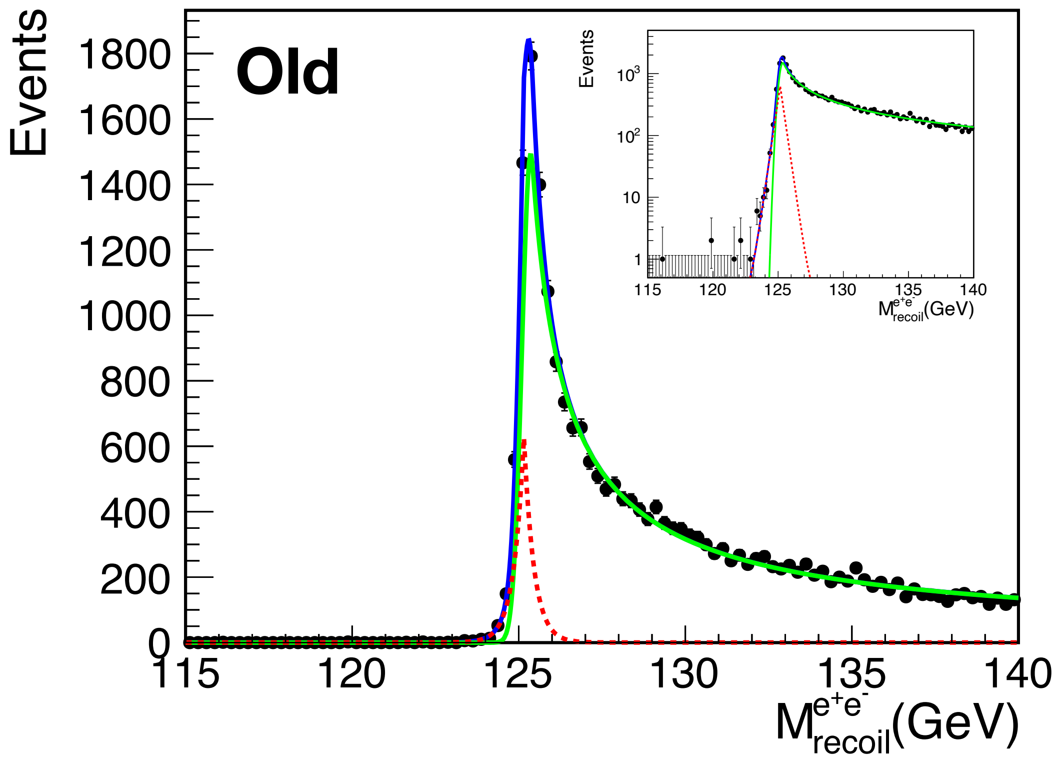


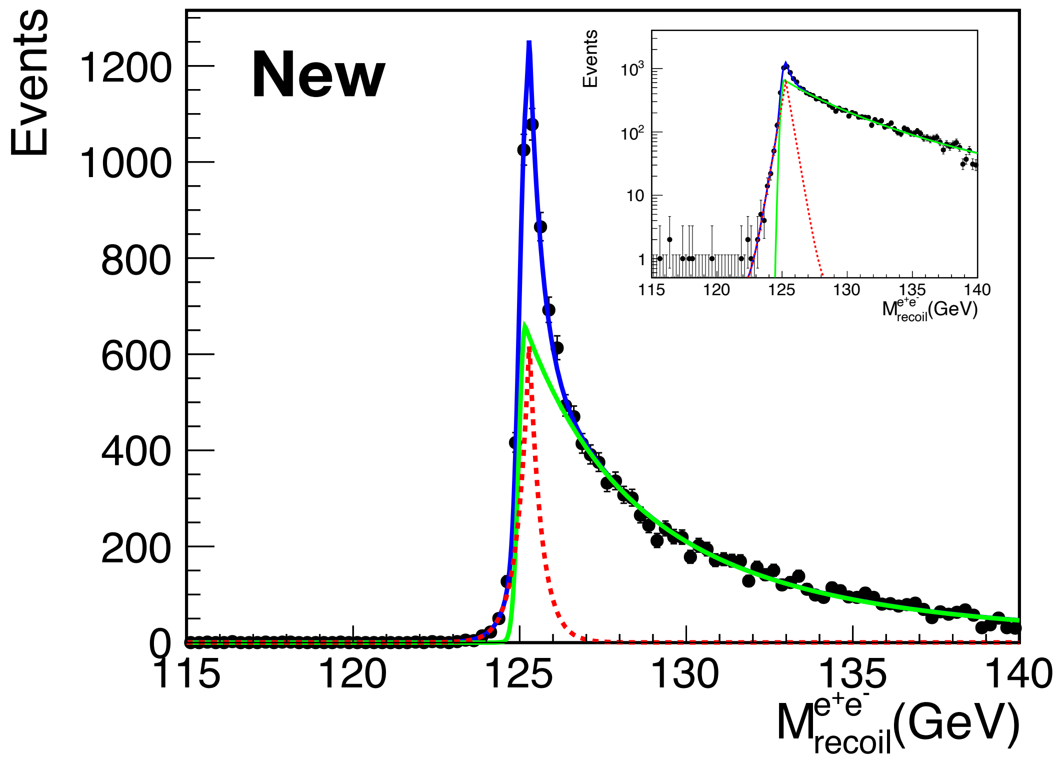
CB Parameters	old	new
a	-1.010 ± 0.016	-0.947 ± 0.016
n	0.939 ± 0.013	1.044 ± 0.014
sigma	0.269 ± 0.003	0.265 ± 0.003
mean	125.237 ± 0.003	125.211 ± 0.003

Head Parameters	old	new
peak	125.154 ± 0.05	124.48 ± 0.02
tau	0.465 ± 0.047	0.434 ± 0.022

Signal Line Shape of lepton recoil system mass: μμH

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

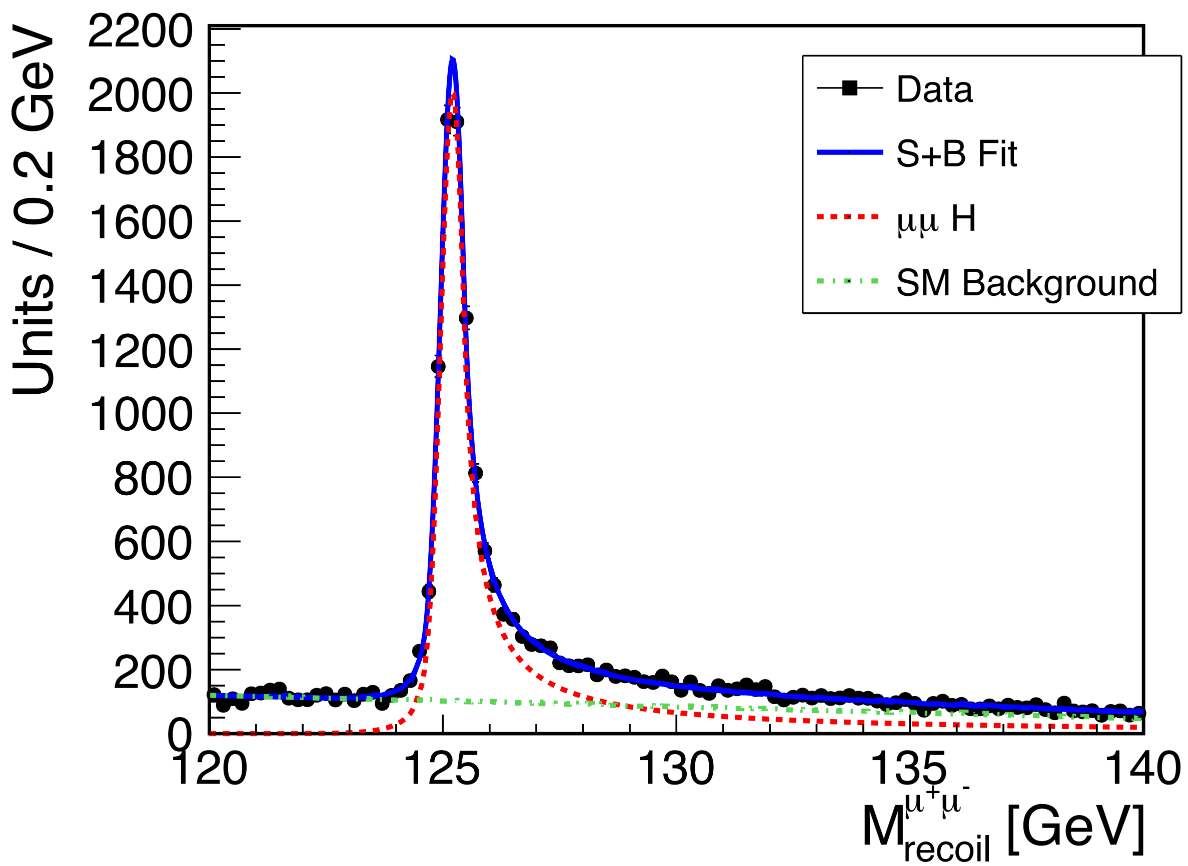




CB Parameters	old	new
a	-0.258±0.037	-0.053±0.008
n	0.800±0.032	3.03±0.46
sigma	0.260±0.013	0.186±0.020
mean	125.359±0.041	125.163±0.027

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

Results of Template Fit: μμH



Signal Variable	Fitted Value
$\mu_{\sigma(H\to bb)}$	0.983±0.014
$\mu_{\sigma(H\to cc)}$	1.02±0.13
$\mu_{\sigma(H\to gg)}$	0.993±0.063
a	-0.932±0.032
n	1.073±0.035
sigma	0.262±0.005
mean	125.211±0.006

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**

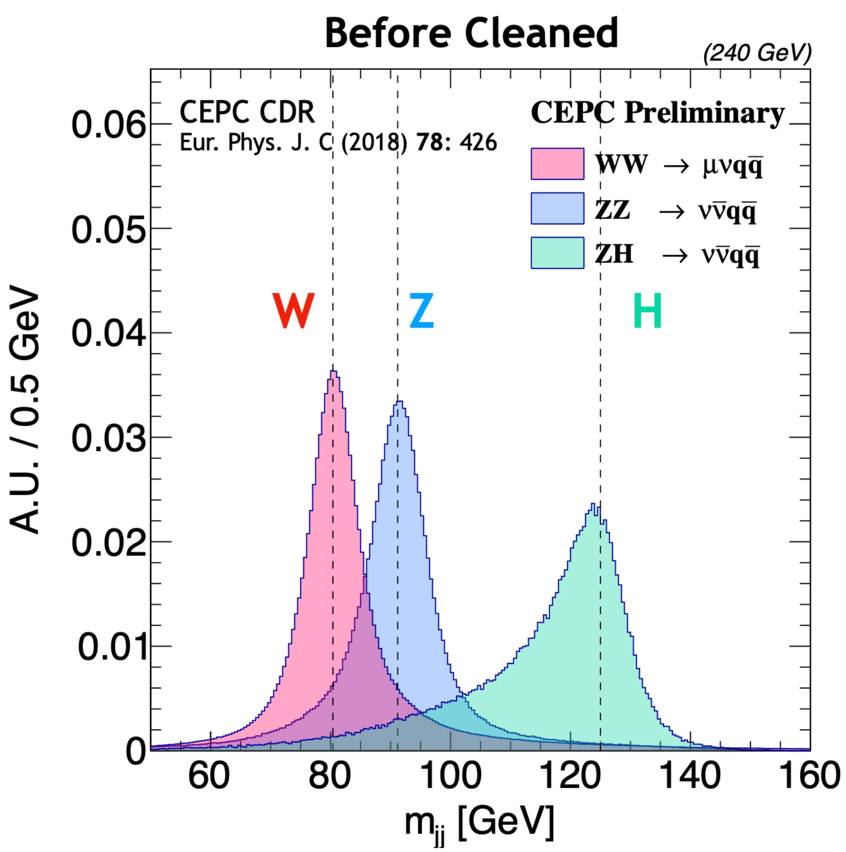
Plan of Future

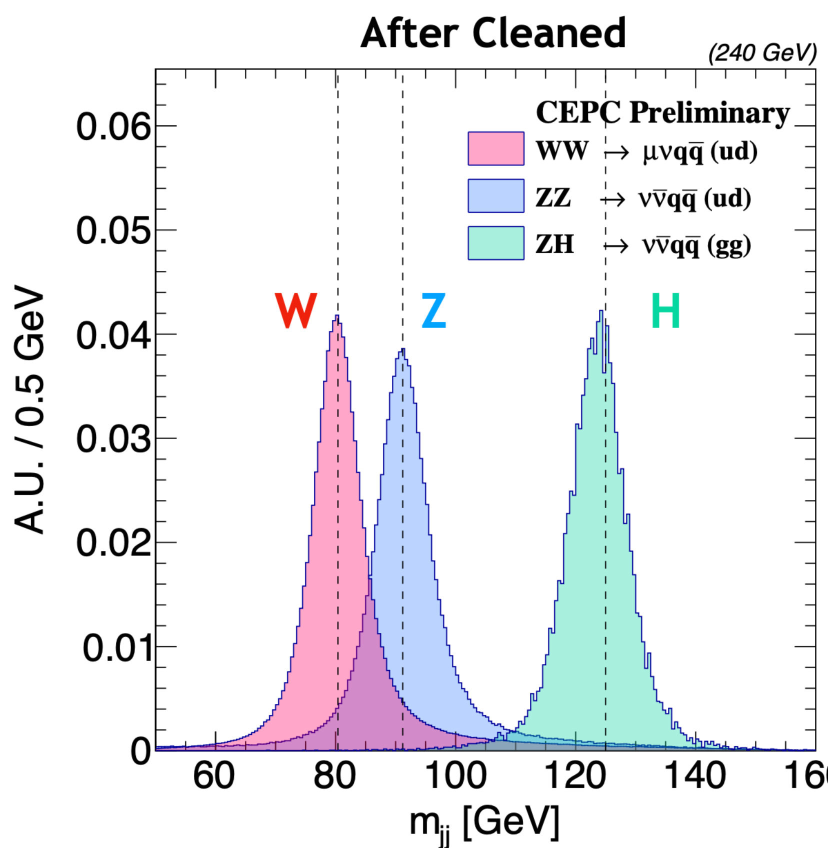
What can we learned from the previous analysis, and in which way we can move forward?

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
- We need a robust algorithm to remove radiated photons and recover their energy/momentum into the leptons

New Strategy on qqH analysis





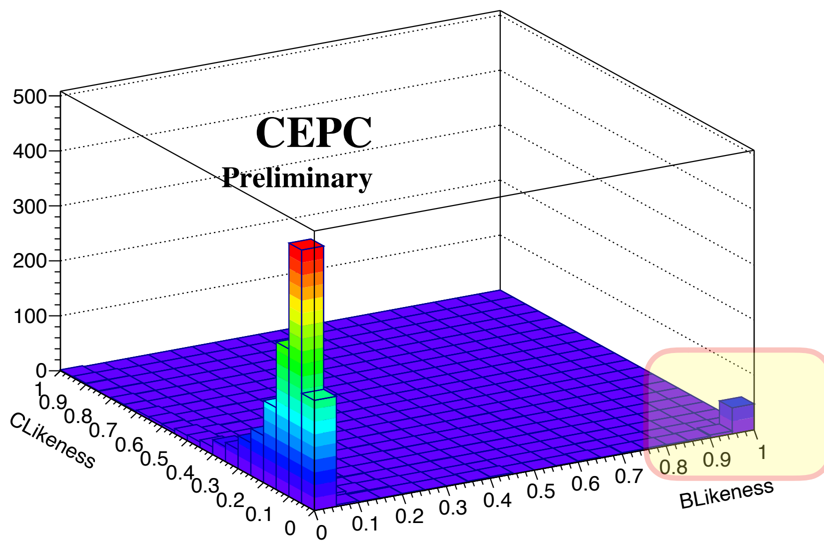
See Lai's talk yesterday

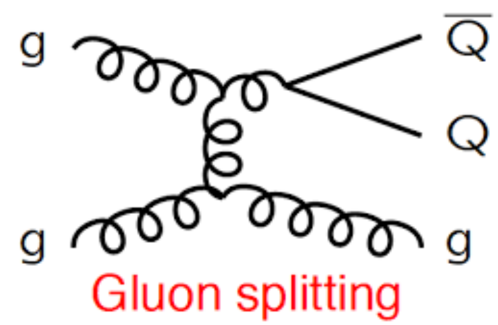
Semi-leptonic decay of heavy flavor hadrons affect the jet pair mass resolution significantly

How might we take this into account?

- **Study the qqH->qq+jj by Z and H flavor categories: 9 categories, explore each of them with different mass pair resolution**
- **Explicitly tag jets with soft leptons(heavy flavor with semi-leptonic decay), try to recovery the missing energy/momentum**

Gluon Jets Issues: Contamination in Heavy Flavor





About 4 % of H->gg in bb flavor region

H->gg template not orthogonal to H->bb/cc

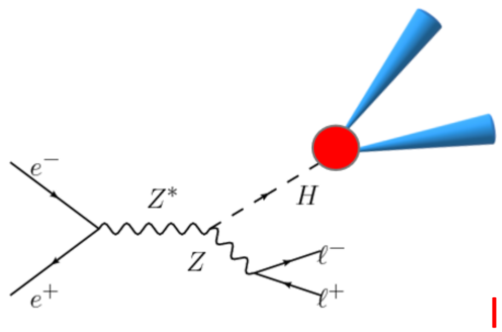
- We need to not only count the flavors, but also consider their distributions in sub-jet level

Gluon Issues: H->gg Analysis

Now H->bb/cc/gg are combined as a ‘signal’ unless the template fit was applied.

- **H->bb has the largest fraction, the optimization of event selection might not so good to the other two channels**
- **Is it possible to apply analysis optimized only for H->gg**

An explicit H->gg analysis



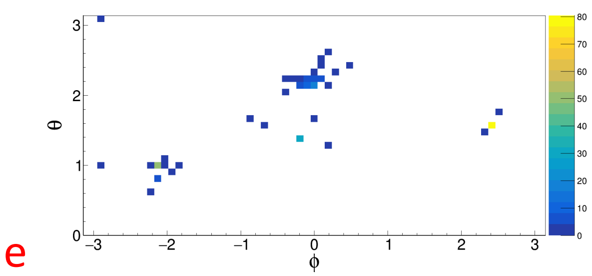
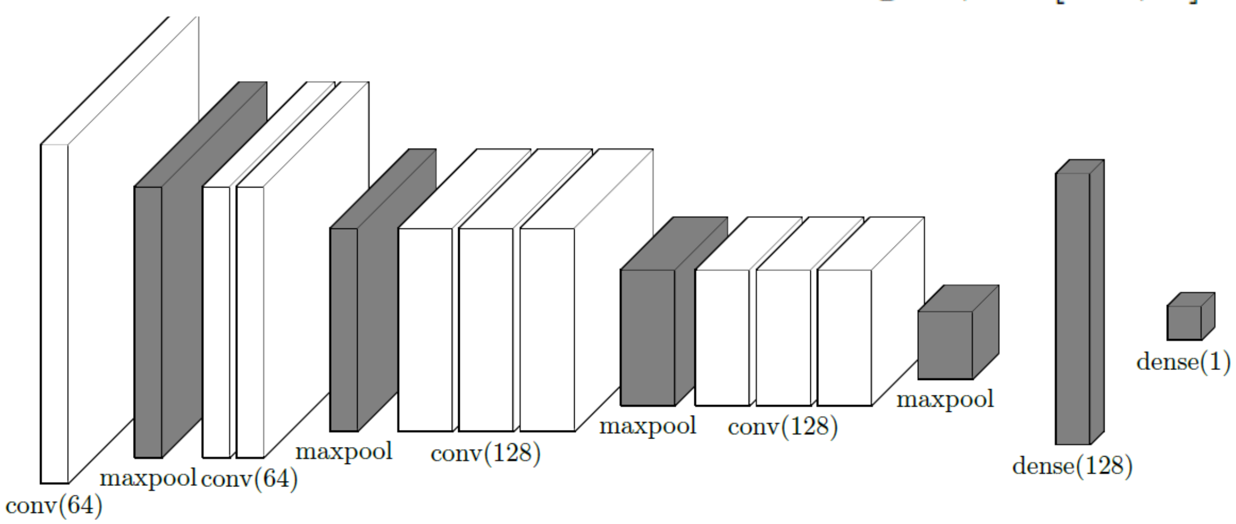


Image is constructed with the energy of all the final state stable particles in an event.

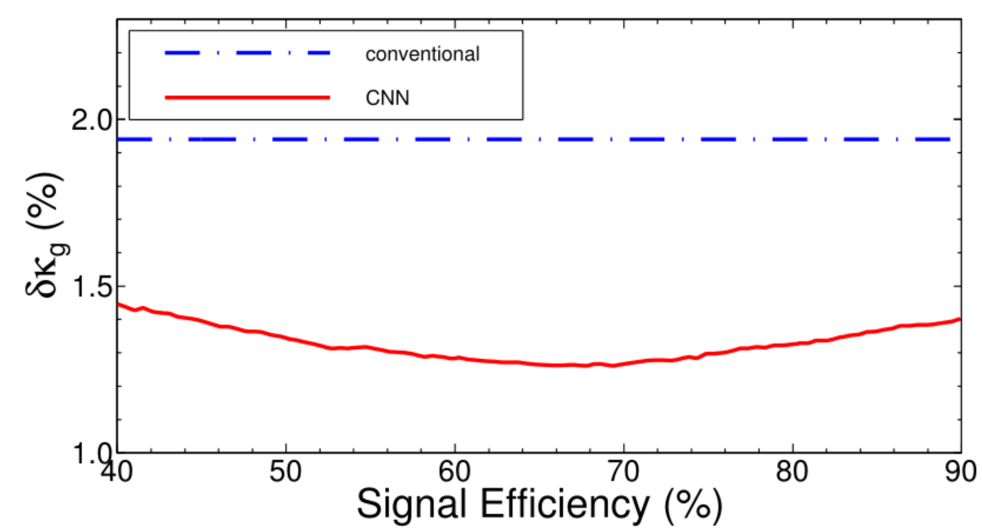
by Wang Yan, Li Zhao and Li Gexing

azimuthal angle $\phi \in [-\pi, \pi]$ and polar angle $\theta \in [0, \pi]$



hgg

not hgg



Each convolutional layer is consisted of 64 or 128 filters with filter size 3*3 and a ReLU activation.

Each maxpooling layer performed a 2*2 down-sampling with a stride length of 2.

Gluon Issues: gluon control sample

- **Like other final states, gluon jets control sample to commissioning the distributions of gluon**
  - **Particle number, flavor etc., need high purity and high efficiency**
- **Unlike other final states, this is not easy to find**
  - **Three jets events?**

Systematic Uncertainty of Flavor Tagging

A lot of things need to be measured in data, and compare to MC prediction: Track multiplicity in jets, tracks' impact parameters, secondary vertex variables, B/C-likeness, X_B-X_C distributions(correlation of tagging variables)

We don't have data yet.

We need to demonstrate how well we could measure those variables in data.

A simple example

To Estimate the systematic of b-tagging in $\mu\mu H$ channel

- **The uncertainty are directly from the difference in 'data' and MC-prediction of the templates**
- **Assuming we use ZZ->**$\mu\mu$+bb **to calibrate the bb-templates, and MC prediction bias for some systematic uncertainty reason(H->bb should also have such bias).**
- **Select a ZZ->$\mu\mu$+qq sample. The purity of sample is 99.6%, with more than 20k ZZ->**$\mu\mu$+bb events.
- Estimate how big a difference between MC and data would be detected, with a limit in the data statistic uncertainty and the precision of R_b prediction
- Estimate the impact to the branch fraction result with such difference

c-jets control sample

- b-c jets separation is relatively worse than c-light jet separation
- A control sample with low b-jets contamination is required
- Semi-leptonic WW sample might be a choice
- Selection on WW-> μ ν+qq events gives high statistic and high purity control sample

	μ+qq	τ+qq	zz->$\mu\mu$+qq	Other
Event Yields	2.2M	38k	10k	8.3k
Fraction	97.5%	1.7%	0.43%	0.36%

- **tau+qq with tau -> mu decay fakes to WW-> μ ν+qq, but they should have similar W hadronic decay components**
- **WW semi-leptonic events has a purity over 99%, and**

About half of them contains a c-jets

Summary

- H->bb/cc/gg measurements are very important in understanding quark's mass origin, and are important BM analysis in CEPC
- Previous work with old geometry verified the capability of high precision measurements of H->bb/cc/gg in CEPC
- PID need to be corrected. But the fit seems promising.
- A lot of things to try and improve: leptons reconstruction, jets paring, gluon jets study, flavor tagging calibration…

Let's keep up the good work!

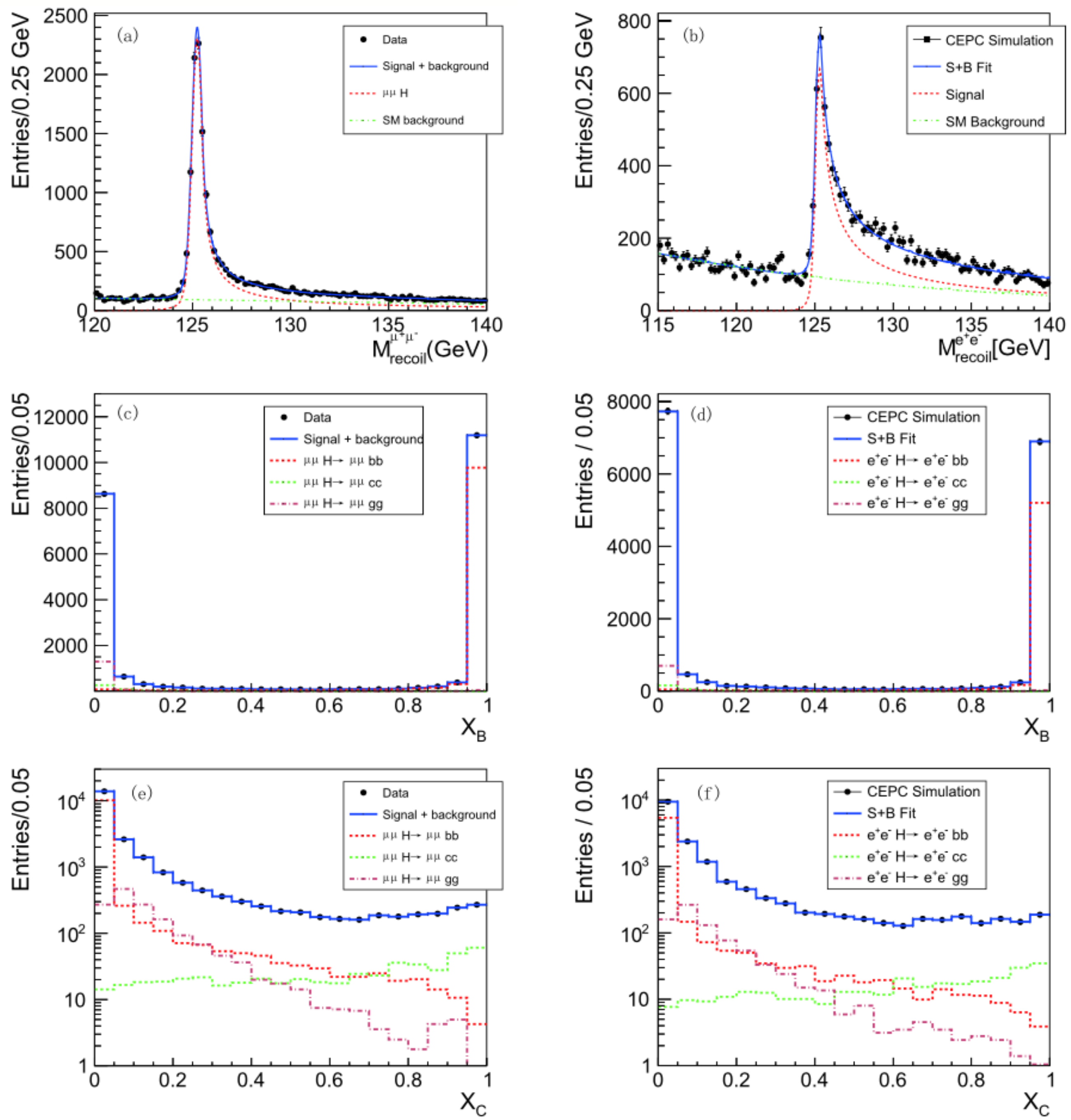
Thank You!

Backup

Likelihood Function of 3D Fit

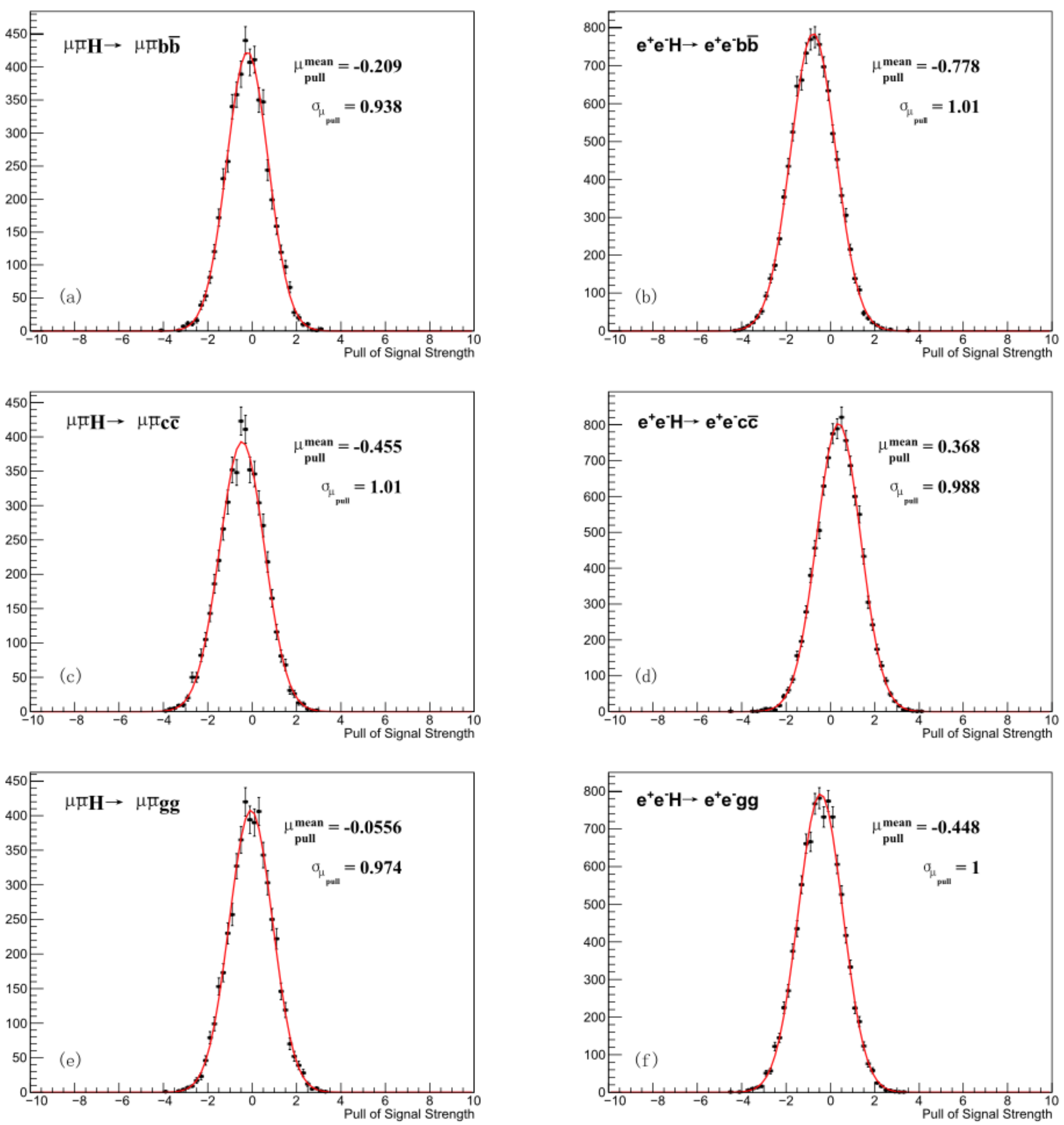
$$
\begin{aligned}
&L(M_{\text{recoil}}^{l\bar{l}}, X_B, X_C; \vec{\theta_s}, \vec{\theta_b}, N_{H\to b\bar{b}}^{\text{sig}}, N_{H\to c\bar{c}}^{\text{sig}}, N_{H\to gg}^{\text{sig}}, N_{\text{irred\_}b\bar{b}}^{\text{bkg}}, N_{\text{irred\_}c\bar{c}}^{\text{bkg}}, N_{\text{irred\_}uds}^{\text{bkg}}, N_{l^+l^-H}^{\text{bkg}}, N_{\text{redu}}) \\
=&P_{\text{sig}}(M_{\text{recoil}}^{l\bar{l}}; \vec{\theta_s})(N_{H\to b\bar{b}}^{\text{sig}} P_{\text{flavor}}^{H\to b\bar{b}}(X_B, X_C) + N_{H\to c\bar{c}}^{\text{sig}} P_{\text{flavor}}^{H\to c\bar{c}}(X_B, X_C) + N_{H\to gg}^{\text{sig}} P_{\text{flavor}}^{H\to gg}(X_B, X_C) + N_{l^+l^-H}^{\text{bkg}} P_{\text{flavor}}^{H\to \text{other}}(X_B, X_C)) \\
&+ P_{\text{irred}}(M_{\text{recoil}}^{l\bar{l}}; \vec{\theta_b})(N_{\text{irred\_}b\bar{b}}^{\text{bkg}} P_{\text{flavor}}^{\text{irred\_}b\bar{b}}(X_B, X_C) + N_{\text{irred\_}c\bar{c}}^{\text{bkg}} P_{\text{flavor}}^{\text{irred\_}c\bar{c}}(X_B, X_C) + N_{\text{irred\_}uds}^{\text{bkg}} P_{\text{flavor}}^{\text{irred\_}uds}(X_B, X_C)) \\
&+ N_{\text{redu}} P_{\text{redu}}(M_{\text{recoil}}^{l\bar{l}}, X_B, X_C).
\end{aligned}
$$

3D fit Plots



(a)
Data
Signal + background
μμ H
SM background
Entries/0.25 GeV
$M_{recoil}^{\mu^+\mu^-}$(GeV)
(b)
CEPC Simulation
S+B Fit
Signal
SM Background
Entries/0.25 GeV
$M_{recoil}^{e^+e^-}$[GeV]
(c)
Data
Signal + background
μμ H→ μμ bb
μμ H→ μμ cc
μμ H→ μμ gg
Entries/0.05
X_B
(d)
CEPC Simulation
S+B Fit
e^+e^- H→ e^+e^- bb
e^+e^- H→ e^+e^- cc
e^+e^- H→ e^+e^- gg
Entries / 0.05
X_B
(e)
Data
Signal + background
μμ H→ μμ bb
μμ H→ μμ cc
μμ H→ μμ gg
Entries/0.05
X_C
(f)
CEPC Simulation
S+B Fit
e^+e^- H→ e^+e^- bb
e^+e^- H→ e^+e^- cc
e^+e^- H→ e^+e^- gg
Entries / 0.05
X_C

ToyMC Test of 3D Fit



H->gg conventional results

Simulation

Whizard → $\sqrt{s} \sim 250GeV, \int Ldt \sim 5ab^{-1}$

⇓

PYTHIA/HERWIG → Shower and Hadronization

⇓

FASTJET → Anti-kt, cone size R = 1.5

⇓

kinematic cuts and b tagging

⇓

uncertainties of Kg in different channels (conventional results)

	$Z \to e^+e^-$	$Z \to \mu^+\mu^-$	$Z \to \ell^+\ell^-$
PYTHIA	2.93%	2.53%	1.94%
HERWIG	2.67%	2.47%	1.82%

Correlation of FT on two jets

- **We need to tag two jets in a events: the tagging/mis-tagging rate will be correlated**
- **Correlation is small (up to a few percent), but need to be considered to achieve high precision**

The correlation of FT are considered in R_b measurement in ALEPH:

- **Phys.Lett.B 401(1997) 163**
- **Phys.Lett.B 401(1997) 150**

Think about di-jet sample:

- $f_s = R_b\varepsilon_b + R_c\varepsilon_c + (1-R_b-R_c)\varepsilon_{uds}$
- $f_d = R_b\varepsilon_b^2 (1+\varrho_b)+ R_c\varepsilon_c^2 + (1-R_b-R_c)\varepsilon_{uds}^2$
- f_s, f_d measured from data
- R_c from other measurement
- ε_c, ε_{uds} and ϱ_b taken from MC
- Work out R_b and ε_b

- $\mathbf{f_s}$: Single Hemisphere Tagged
- $\mathbf{f_d}$: Both Hemisphere tagged
- $\boldsymbol{\varepsilon_x}$: Efficiency for flavor x
- ϱ_b : Hemispheres Correlation

$\varrho_b\langle\varepsilon_b\rangle^2 = \langle\varepsilon_b^2\rangle-\langle\varepsilon_b\rangle^2$

In less symmetric case:

$\boldsymbol{\varrho_b\langle\varepsilon_1\rangle\langle\varepsilon_2\rangle = \langle\varepsilon_1\varepsilon_2\rangle-\langle\varepsilon_1\rangle\langle\varepsilon_2\rangle}$