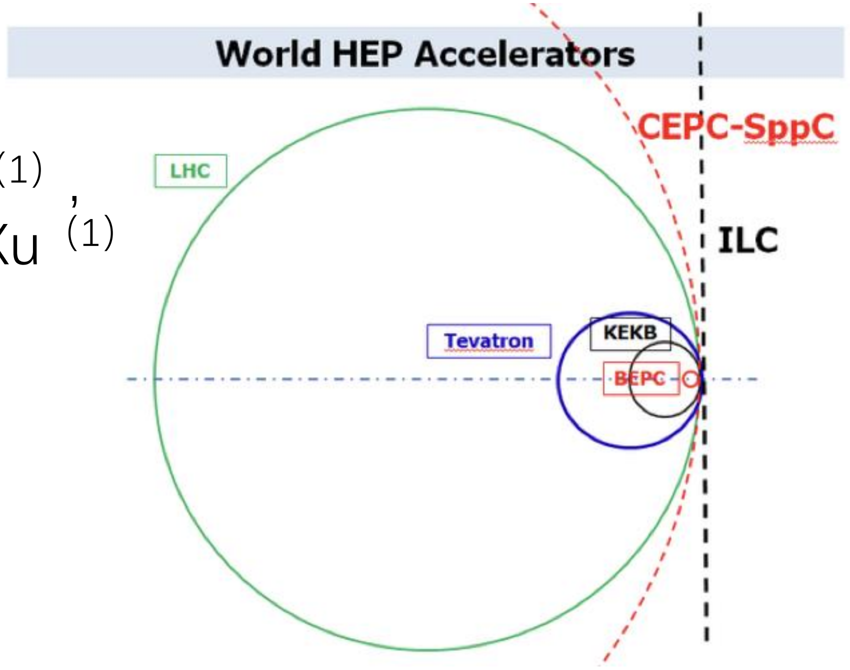
H→bb/cc/gg measurement in CEPC with modified PFN

Yu Bai [1] , Gang Li [2] , JianPeng Deng [1] , JianYu Huang [1] , WeiHan Tan [1] , HeYu Meng [1] , Ke Wang [1] , NengXuan Xu [1]

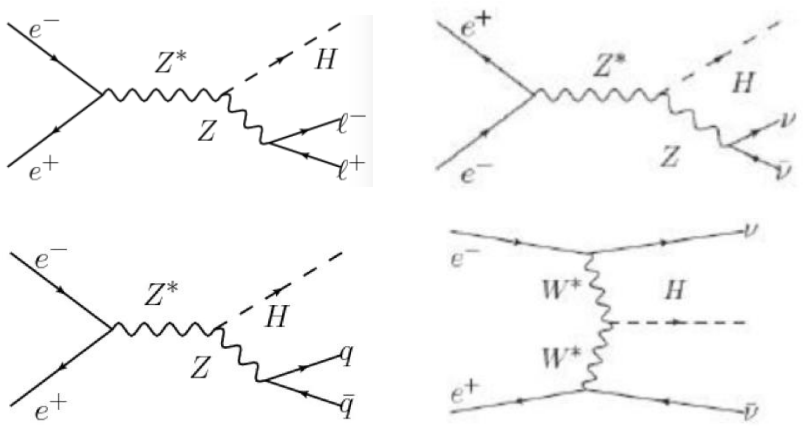
July 4, 2023

(1) Southeast University
(2) Institute of high energy physics

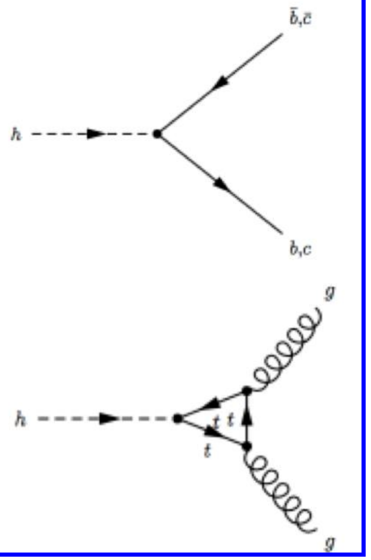


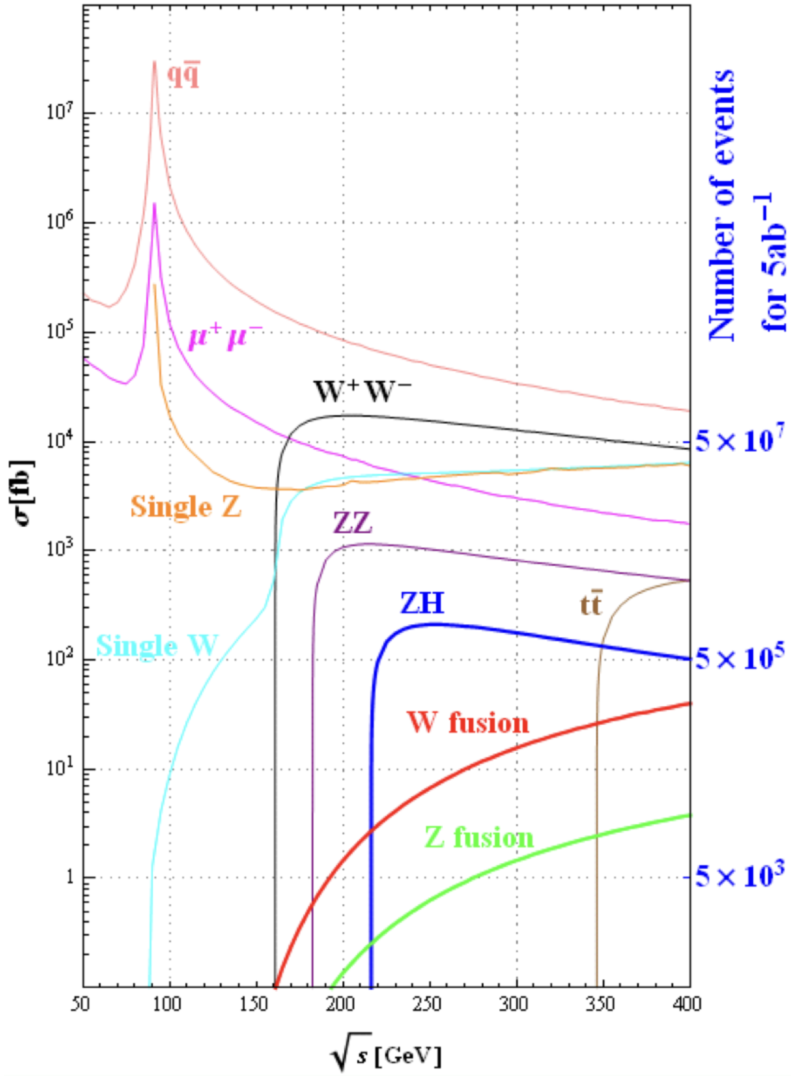
H->bb/cc/gg in CEPC

Higgs Production in CEPC



H->bb/cc/gg





- Understanding Yukawa coupling between Higgs boson and quarks
- CEPC is an ideal place to precisely measure Higgs decay
- An important benchmark measurement of CEPC

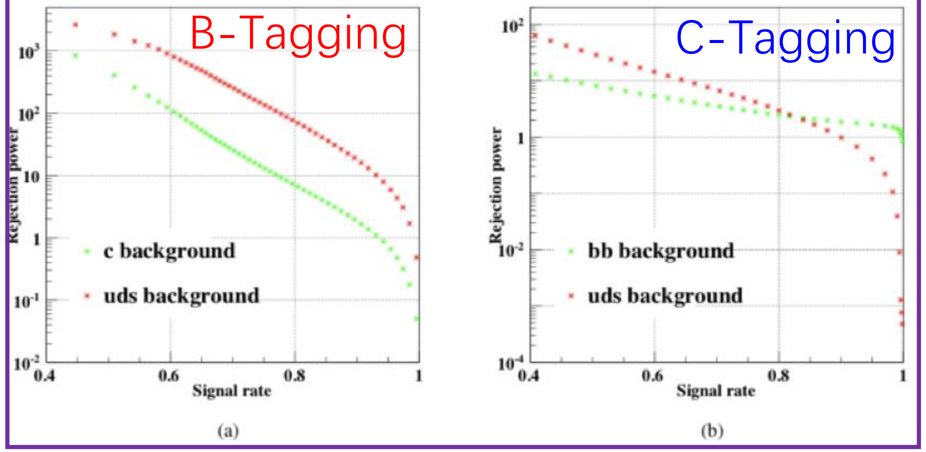
Previous measurement

[Chinese Physic C, 043002, 2019]
[Chinese Physic C, 013001, 2020]

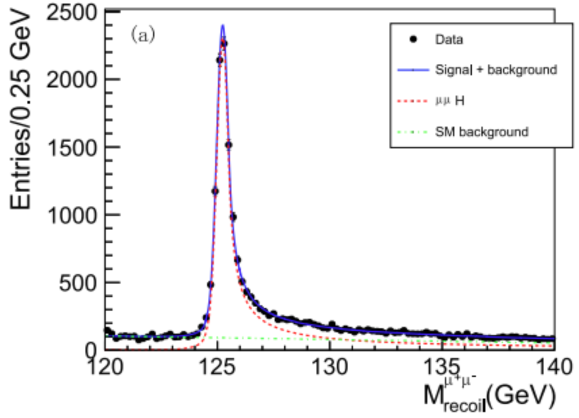
$$y_{ij} = \min\{E_i^2, E_j^2\}(1-\cos\theta_{ij})/E_{vis}^2$$

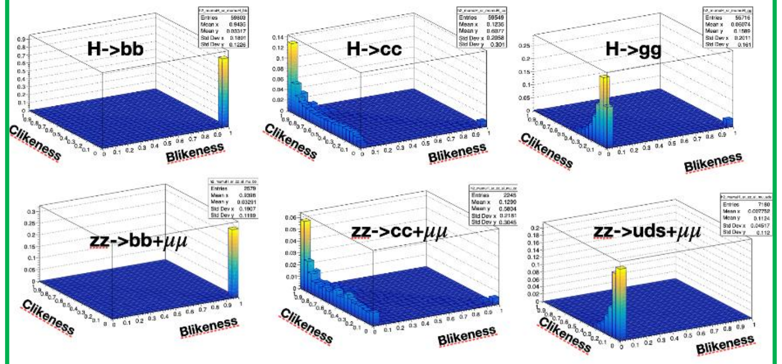
Merge Jets according to 'distance'

GBDT Flavor Tagging



3D-Fitting





- ee-kt jet algorithm
- BDT Flavor tag
- Mrecoil × Flavor Tagging template fit

Key performance to be improved

- Jet algorithm
  - Separation between with different parton multiplicity
    - H->bb/cc/gg and H->ww/zz->4q/2q+2l
- Flavor tagging
  - B/L separation is good
  - Improve c-tag performance(especially contaminations from B)
  - Gluon/quark jet separation

Particle flow network (PFN)

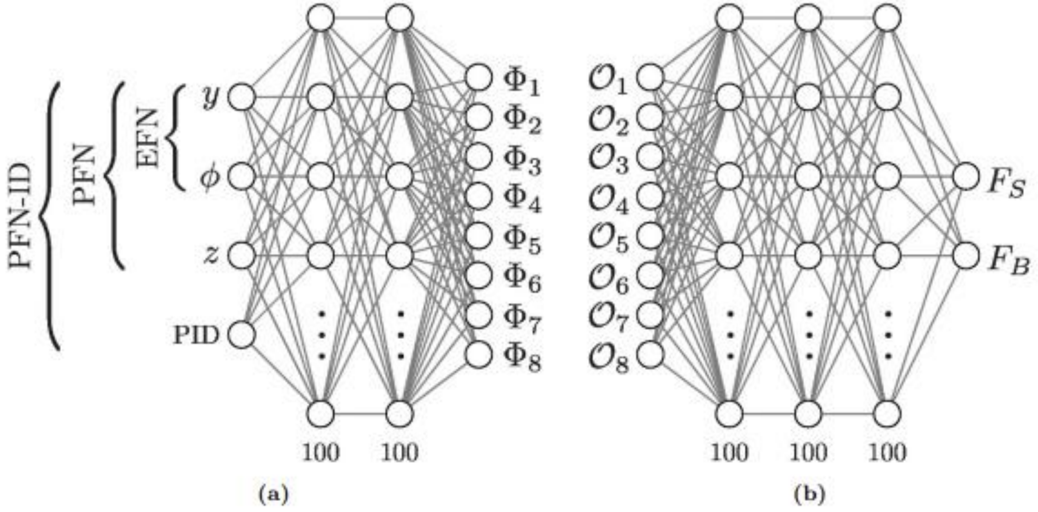


Figure 4. The particular dense networks used here to parametrize (a) the per-particle mapping Φ and (b) the function F, shown for the case of a latent space of dimension $\ell = 8$. For the EFN, the latent observable is $\mathcal{O}_a = \sum_i z_i \Phi_a(y_i, \phi_i)$. For the PFN family, the latent observable is $\mathcal{O}_a = \sum_i \Phi_a(y_i, \phi_i, z_i, \mathrm{PID}_i)$, with different levels of particle-ID (PID) information. The output of F is a softmaxed signal (S) versus background (B) discriminant.

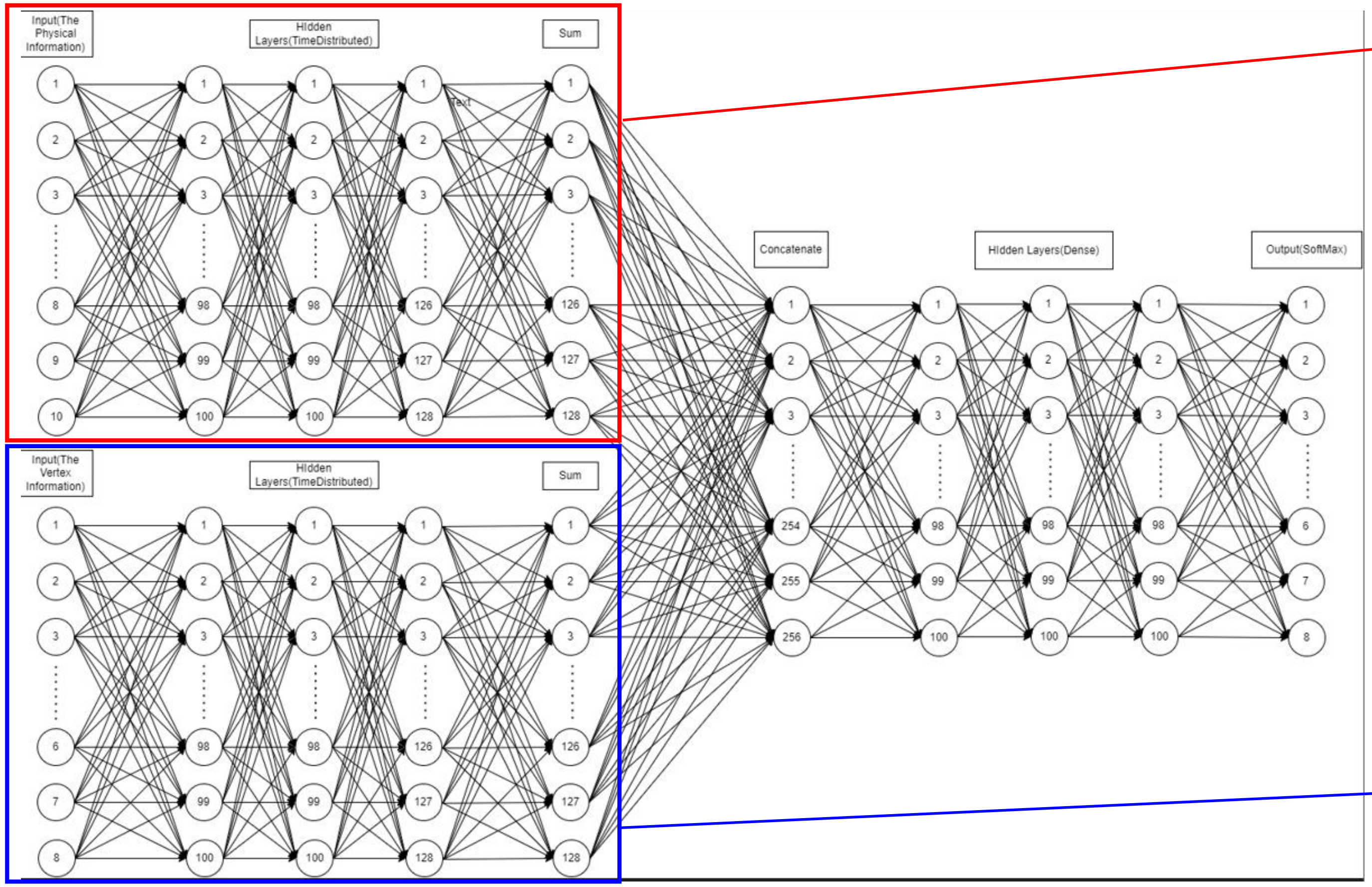
JHEP01, 2019, 121

z : Particle momentum
Variables added as Φ input:

- PID (recoded)
- D0, Z0
- D0, Z0 significance
- Track likelihood (from primary vertices)
- Energy
- Charge

- Φ-Layer: DNN layer for each particle
- O-Layer: A summation over particles (dot production between z and Φ)

Modification of PFN



z : Particle momentum
Variables added as Φ input:

- PID (recoded)
- D0, Z0
- D0, Z0 significance
- Track likelihood (from primary vertices)
- Energy
- Charge

Secondary Vertices Information:

- Vertices position
- Vertices momentum

- Adding Φ-Layers for vertices to improve tagging performance

Datasets and Pre-selection

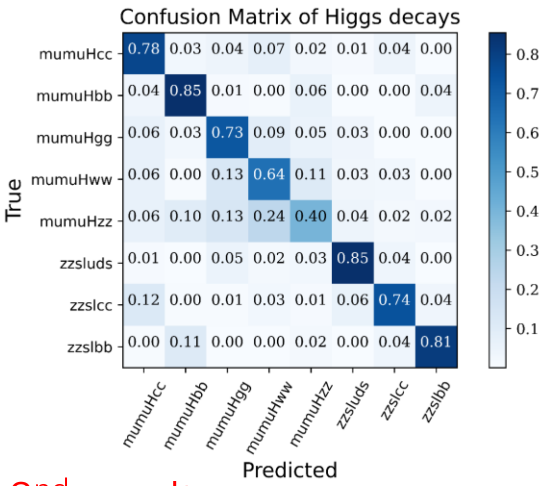
- Event selection are the same as in previous work
- High statistic samples are necessary for model training
  - 400k-700k events for μμH, H->bb/cc/gg/ww/zz each
  - High statistics ZZ-> μμ +qq background, generated with event filter

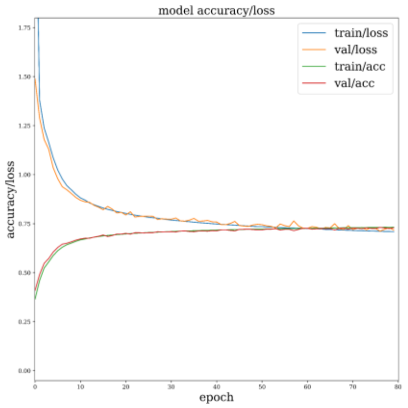
sample	ZZ->μμbb	ZZ->μμcc	ZZ->μμss	ZZ->μμuu/dd
Cross section(fb)	45.45	43.73	45.48	89.20
Events in 5 ab^{-1}	227.3k	218.7k	227.4k	446.0k
Generated events	100M	100M	100M	100M
Filtered Events	4.807M	4.492M	4.812M	4.649M
Simulated and reconstructed	4.684M	4.378M	4.687M	454.2M
Passing cut	1.778M	1.633M	1.712M	1.644M
Selection Eff	1.825%	1.676%	1.758%	1.630%

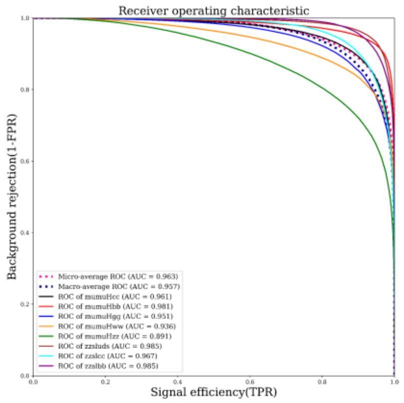
Filter Eff 98% reached

Training Results

1st results

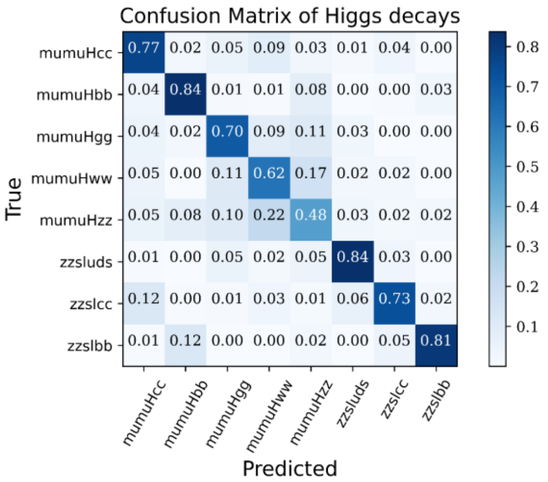


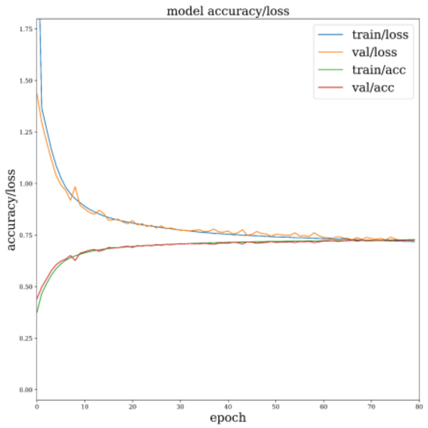


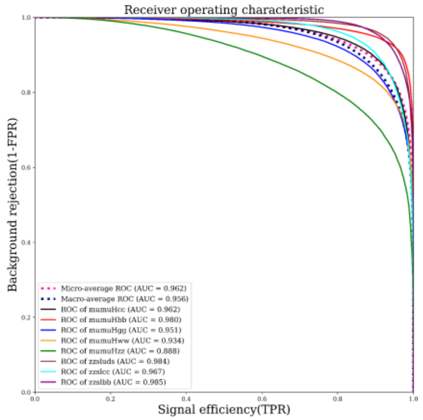


- 400k events per-sample (3.2M all)
- Training:Validation:Test = 8:1:1

2nd results







80 epochs,
72%-73% overall accuracy
no over training is found

Results of H->bb/cc/gg with Modified PFN

$$\overset{\text{Observed}}{\begin{bmatrix} N^o_{s1} \\ N^o_{s2} \\ \cdots \\ N^o_{b1} \\ N^o_{b2} \\ \cdots \end{bmatrix}} = \Big(\underbrace{M^{cnf}}_{\text{Confusion Matrix}} \overbrace{M^s}^{\text{Selection Eff}}\Big) \times \overset{\text{Process Event Yields}}{\begin{bmatrix} N^p_{s1} \\ N^p_{s2} \\ \cdots \\ N^p_{b1} \\ N^p_{b2} \\ \cdots \end{bmatrix}}$$

Denoting $M = M^{cnf} M^s$

$$\begin{aligned} N^p &= M^{-1} N^o \\ \underbrace{\Sigma(N^p)}_{\text{Uncertainty of } N^p} &= M^{-1} \Sigma(N^o) (M^T)^{-1} \end{aligned}$$

process	μμH->μμbb	μμH->μμcc	μμH->μμgg	μμH->μμww	μμH->μμzz
1st results	1.0%	9.1%	3.5%	3.1%	29.%
2nd results	1.0%	8.9%	3.5%	3.2%	31.5%

Previous results

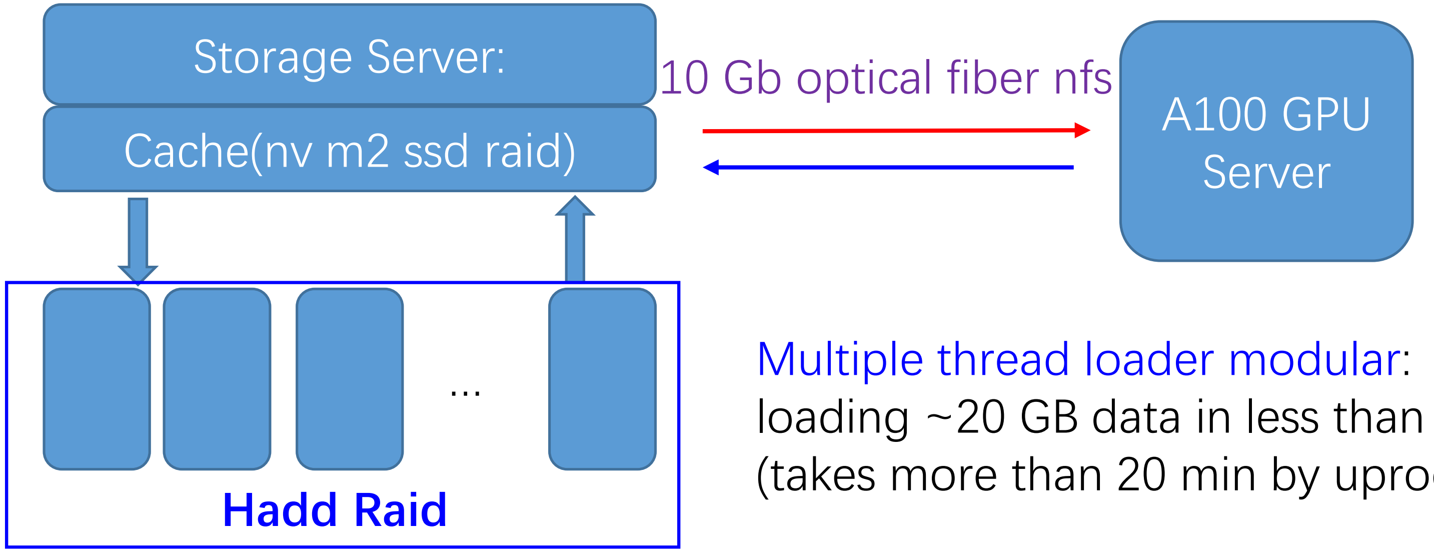
Higgs boson production	$\mu^+\mu^- H$		
Higgs boson decay	$H \to b\bar{b}$	$H \to c\bar{c}$	$H \to gg$
statistic uncertainty	1.1%	10.5%	5.4%
fixed background	−0.2%	+4.1%	7.6%
	+0.1%	−4.2%	
event selection	+0.7%	+0.4%	+0.7%
	−0.2%	−1.1%	−1.7%
flavor tagging	−0.4%	+3.7%	+0.2%
	+0.2%	−5.0%	−0.7%
combined systematic uncertainty	+0.7%	+5.5%	+7.6%
	−0.5%	−6.6%	−7.8%

- Improvement performance in H->bb/cc/gg, especially for H->gg and H->cc

Some technical notes

工欲善其事，必先利其器
–孔子
To Do something well, sharpen your tools first
-Confucius

- Model training requires computing power
- Solution: Storage server + fiber connection + GPU server + LoadingModular



~30 min training 3.2M events in 80 epoches

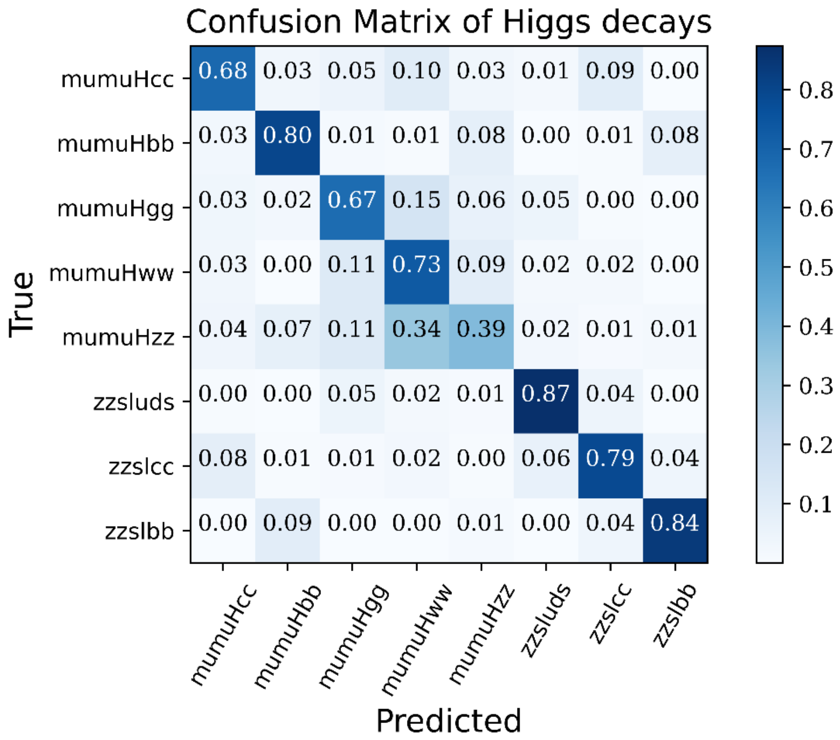
Multiple thread loader modular:
loading ~20 GB data in less than 20 seconds
(takes more than 20 min by uproot4)

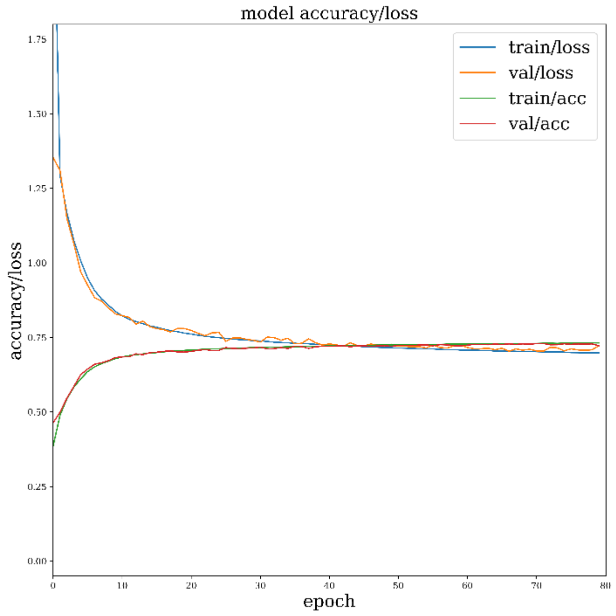
Summary and Prospect

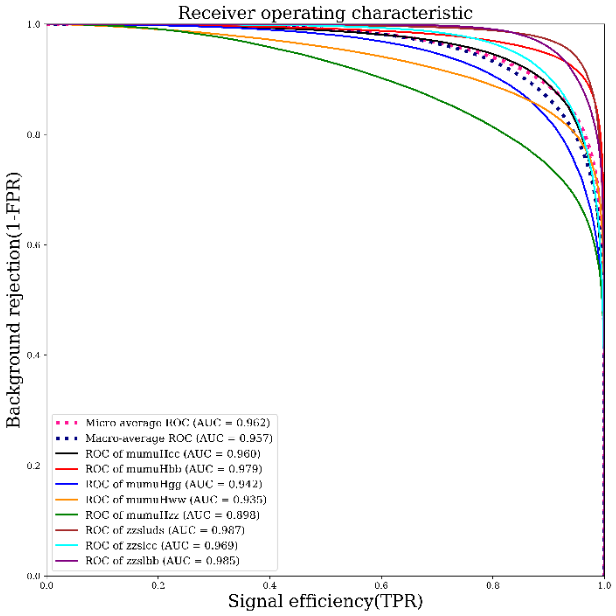
- PFN is demonstrated to be a powerful tool to identify multiple jets FS
  - Kinematics information: particle momentum/angular distribution, multiplicity etc.
  - Flavor information: D0/Z0 and their significance
- The PFN Φ-Layer can integrate other type of object to improve performance
  - E.g. vertices information for jet flavor identification
- Huge size of training sample are necessary
  - Event filter is a solution to the problem of huge computing resources required
- Fast loading tool is very helpful to cooperate with hardware
- Architecture improvement:
  - A vector of weight implemented in Φ-Layer summation? (1D weight->mult-Dim W)

Backups

Training results with full hadronic Hww/zz







Comparison between different configuration

	$\sigma^{c\bar{c}}_{\mu^+\mu^- H}$	$\sigma^{bb}_{\mu^+\mu^- H}$	$\sigma^{gg}_{\mu^+\mu^- H}$
Cut-based + Template fit	10.5%	1.1%	5.4%
PFN	9.8%	1.0%	4.1%
PFN-VTX	9.1%	0.9%	3.6%

Table 2. The statistical uncertainty of the signal cross section

Vertex information improve
H->gg/cc precision significantly

Variables	Accuracy	AUC	$\sigma^{c\bar{c}}_{\mu^+\mu^- H}$	$\sigma^{bb}_{\mu^+\mu^- H}$	$\sigma^{gg}_{\mu^+\mu^- H}$
Mom + THETA + PHI	0.45	0.87	36.66	1.99	9.32
M + PDGID + Energy + IsPhoton + Charge	0.53	0.90	24.76	1.47	6.56
$M + F$ + D0 + Z0 + Prob	0.74	0.96	9.52	0.95	3.76
$M + F + \theta$ + Vertex	0.75	0.96	9.11	0.94	3.61

Table 3. The performance with changed variables

D0/Z0 is essential to Flavor tagging