Enhancing flavor tagging and R_b measurement at the CEPC

Libo Liao,
Gang LI, Weimin Song, Shudong Wang, and Zhaoling Zhang

2022.5

Outline

- Introduction
- Deep learning architectures & Datasets
- Evaluation metrics
- Tagging
  - Jet tagging
  - Event tagging
- R_b & R_c measurements
- Conclusion

Introduction

- Relative decay width plays a key role in SM testing and experiment of precision measurement of Z boson

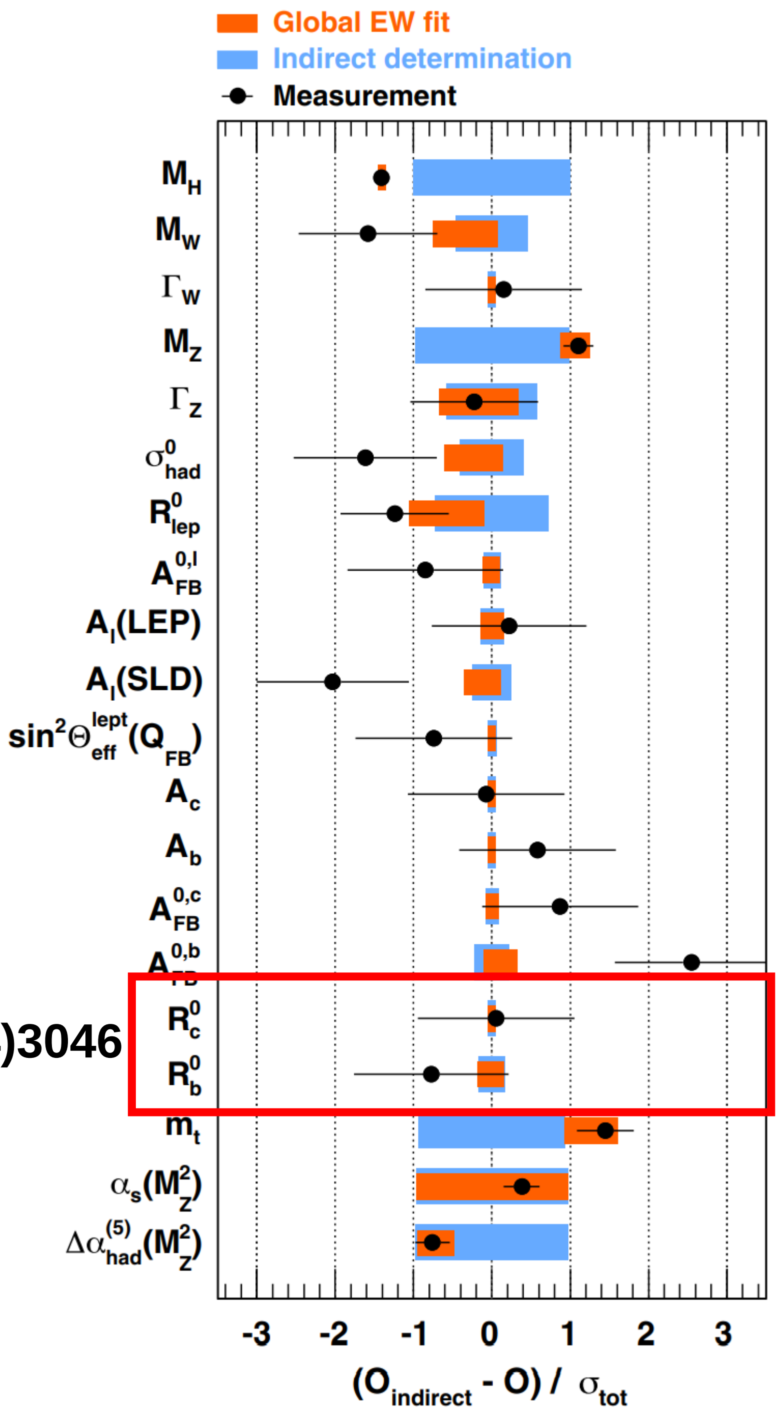
$$R_q = \Gamma_q/\Gamma_h$$

$$\Gamma(Z \to q\bar{q}) = 12\boxed{\Gamma_0}(\boxed{g_{Aq}^2}\boxed{R_{Aq}} + \boxed{g_{Vq}^2}\boxed{R_{Vq}})$$

- Status of R_b and R_c measurements in experiment and theory
  - 1.3σ deviations in R_b

	Experiment	Theory
R_b	0.21629 ± 0.00066	0.21578 ± 0.00011
R_c	0.1721 ± 0.0030	$0.17226^{+0.00009}_{-0.00008}$

Gfitter Group: EPJC74 (2014)3046



Introduction

- Key object: Jet
  - Vertex finding
  - Clustering
  - Tagging
- Methods for tagging
  - Technique: Cut-based->TMVA->Deep Learning
- CEPC baseline detector

  CDR: 80% effi. & 90% purity in b-tagging

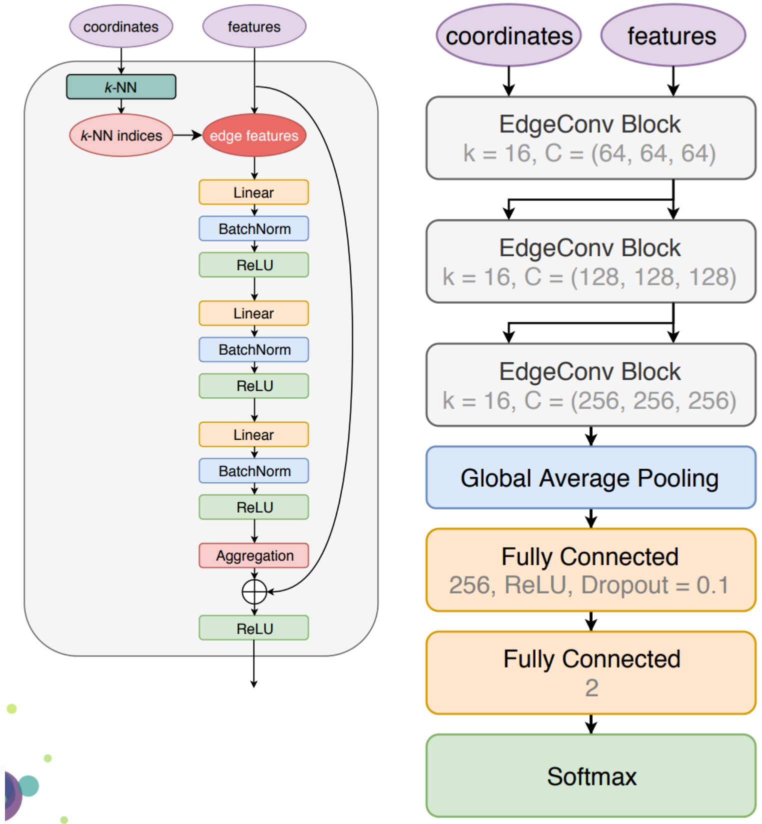
  60% effi. & 60% purity in c-tagging
- A new attempt: Event tagging

Deep learning architectures

[Jet tagging via particle clouds]

➢ ParticleNet

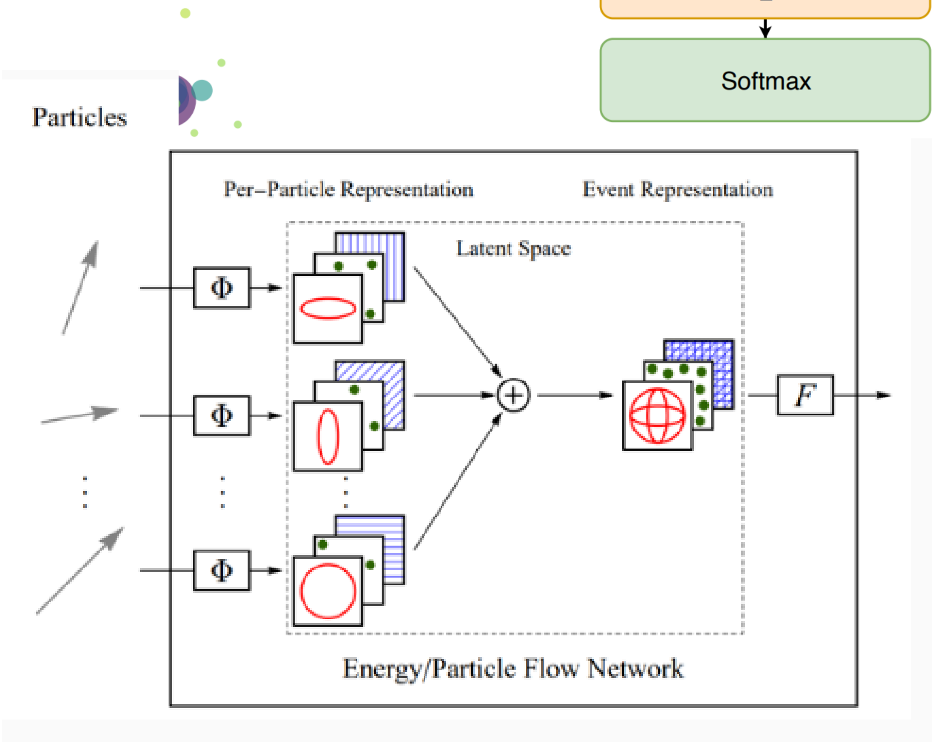
- Treating a jet as an unordered set of particles in space
- Using permutation-invariant graph neural networks



[Energy flow networks: deep sets for particle jets]

➢ Particle Flow Network (PFN)

- Based on "point clouds"
- As a cross check



Datasets

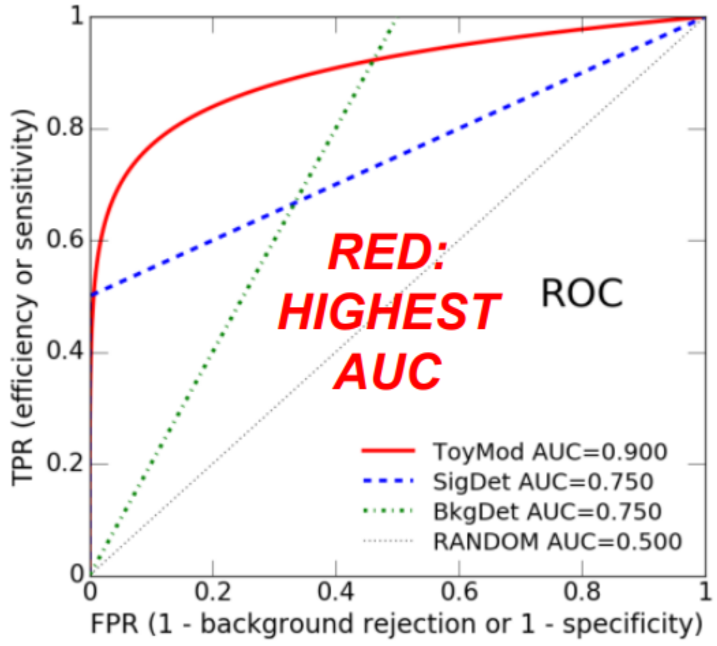
- Full simulation with CEPC baseline detector at Z-pole
- PID used as a feature by matching reconstruction and MC truth
- In flavor tagging
  - 900k jets for each flavor($b, c, o = uds$);
  - Clustered by $ee - kt$ into 2 jets
- In event tagging(same samples are used)
  - $Z \to b\bar{b}, c\bar{c}, o\bar{o}$, 450k events for each channel
  - The main background $Z \to \tau\bar{\tau}$ considered
  - No jet clustering (directly classify events into different category)
- Train:validation:test = 7:1.5:1.5

Evaluation metrics

- Efficiency ϵ_s = TP/(TP+FN)
- Purity ρ_s = TP/(TP+FP)
- Accuracy = (TP+TN)/ALL
- ROC/AUC
- $\epsilon_s \times \rho_s$: between 0 and 1
  - The higher, the better
  - Proportional to 1/error2

$$\boxed{(\Delta R_i)^2 = \frac{R_i}{\mathcal{L}\sigma_h \epsilon_i \rho_i} \propto \frac{1}{\epsilon_i \rho_i},}$$

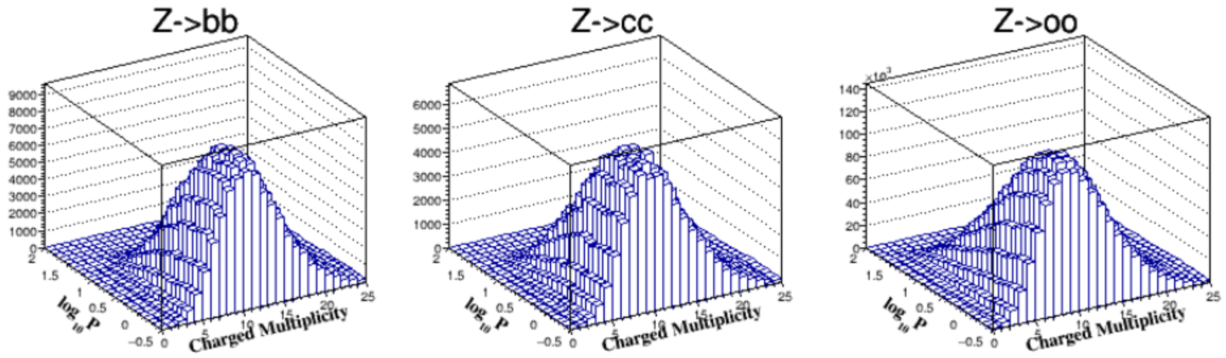
TP (S_{sel}), FP (B_{sel}), FN (S_{rej}), TN (B_{rej})	TP (S_{sel}), FP (B_{sel}), FN (S_{rej}), TN (B_{rej})	TP (S_{sel}), FP (B_{sel}), FN (S_{rej}), TN (B_{rej})
$\mathbf{TPR} = \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FN}}$	$\mathbf{PPV} = \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FP}}$	$\mathbf{TNR} = \frac{\mathbf{TN}}{\mathbf{TN}+\mathbf{FP}} = \mathbf{1}-\mathbf{FPR}$
HEP: "efficiency" $\epsilon_s = \frac{S_{\text{sel}}}{S_{\text{tot}}}$	HEP: "purity" $\rho = \frac{S_{\text{sel}}}{S_{\text{sel}}+B_{\text{sel}}}$	HEP: "background rejection" $1-\epsilon_b = 1-\frac{B_{\text{sel}}}{B_{\text{tot}}}$



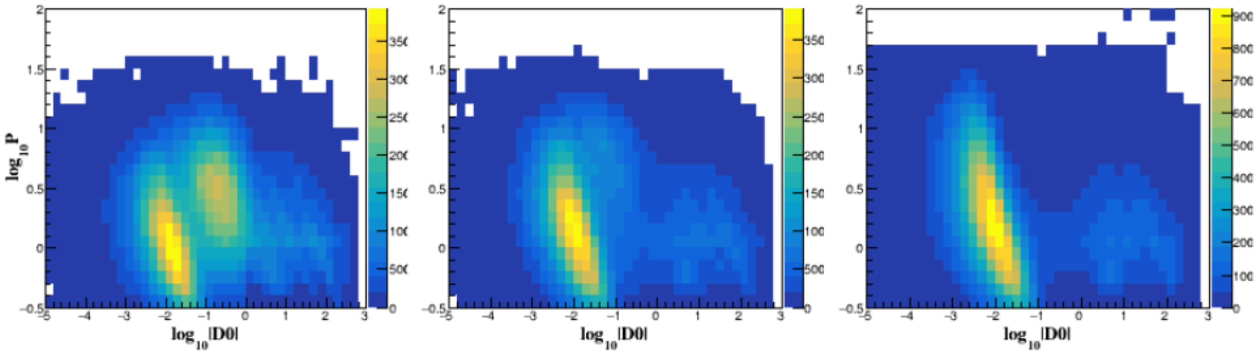
Jet tagging

Jet features

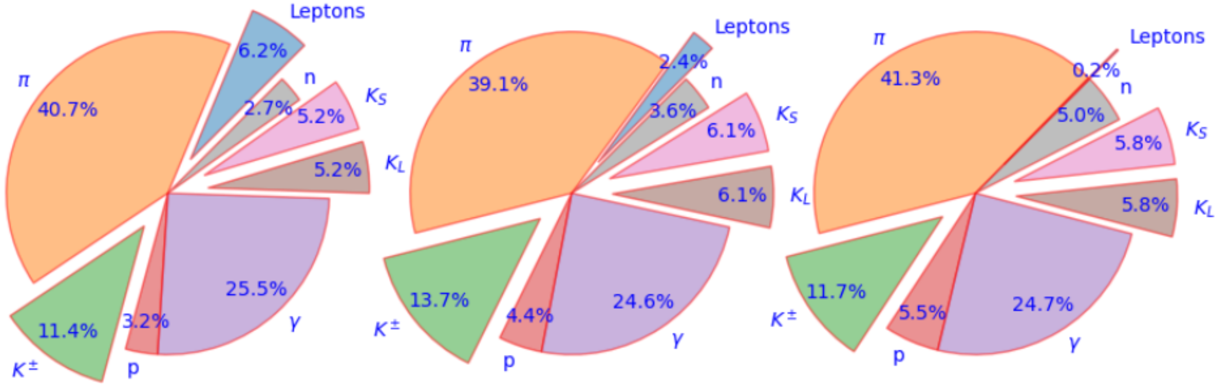
- (a) The multiplicity versus momenta of tracks
  - The number of tracks: $b > c > o$
- (b) The distribution of impact parameters versus momenta
  - Larger impact parameters and energetic tracks in b
- (c) The weighted fractions of different particle type
  - Far more energetic leptons in b
  - Slightly more energetic K in c



(a) Log(P) vs. Ntrks



(b) Log(P) vs. log(D0)



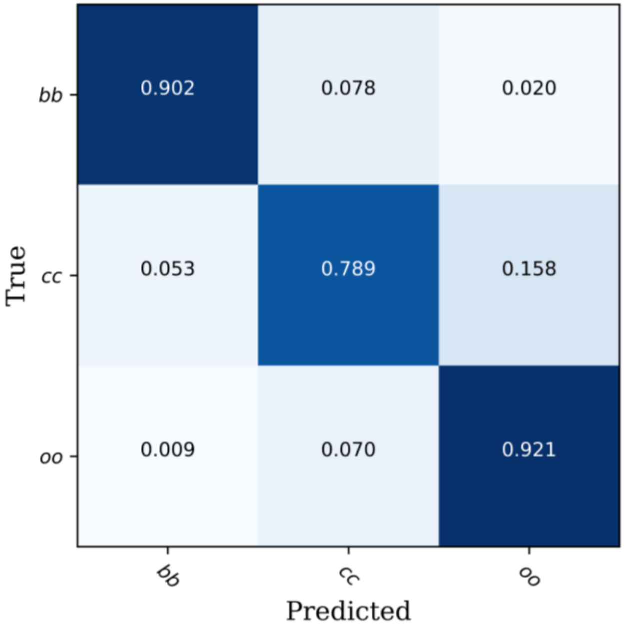
(c)

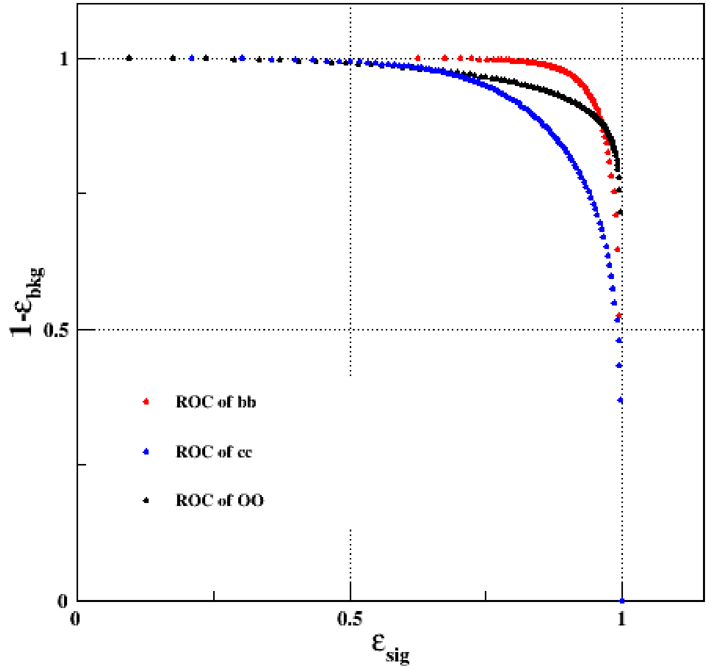
Jet tagging

Fan Yang: Flavor Tagging Using Machine Learning Algorithms

Algorithm	ParticleNet	PFN	DNN	BDT	GBDT	gcforest	XGBoost
Accuracy	0.872	0.850	0.788	0.776	0.794	0.785	0.801

- At least 9% improvement in ParticleNet at global accuracy
- The performance of b-tagging and o-tagging are much better than c-tagging
- ParticleNet is better than the PFN
  - Consistent with the study by Qu, et al





tag	ParticleNet		PFN	
	Efficiency	AUC	Efficiency	AUC
b	0.902	0.984	0.870	0.979
c	0.789	0.948	0.765	0.930
o	0.921	0.972	0.911	0.966

Physics impacts of jet tagging

- Several working points, LCFIPlus & XGBoost as reference
  - ParticleNet & PFN are better than the other two, especially in c
- Improved statistical uncertainty
  - roughly 30% in counting c jets(compare to CDR)

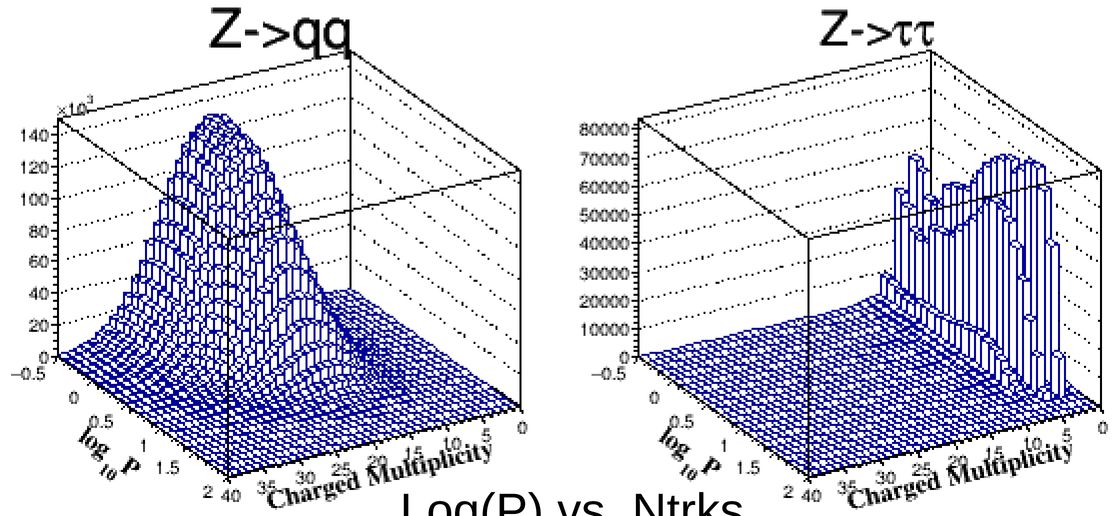
tag	$\epsilon_S(\%)$	$\epsilon \times \rho$			
		LCFIPlus	XGBoost	ParticleNet	PFN
b	80	-	0.747	0.780	0.763
	90	0.72	0.713	0.810	0.752
c	60	0.36	-	0.548	0.485
	70	-	-	0.589	0.497
	80	-	0.345	0.584	0.467
	90	-	0.292	0.516	0.402

Applied in R_q measurement

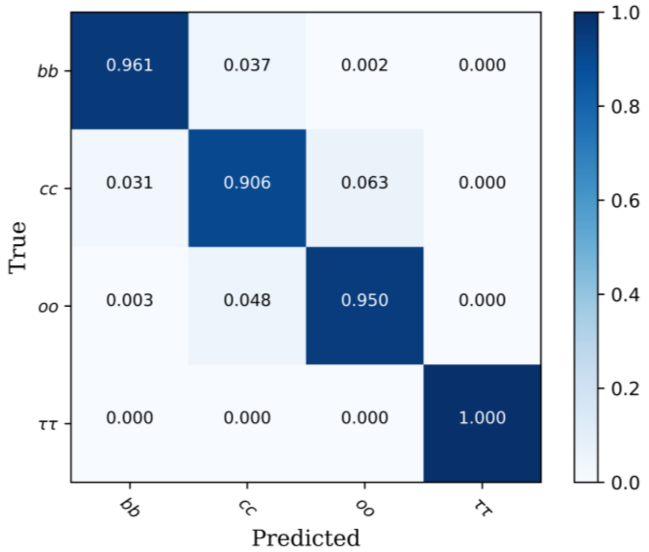
Event tagging

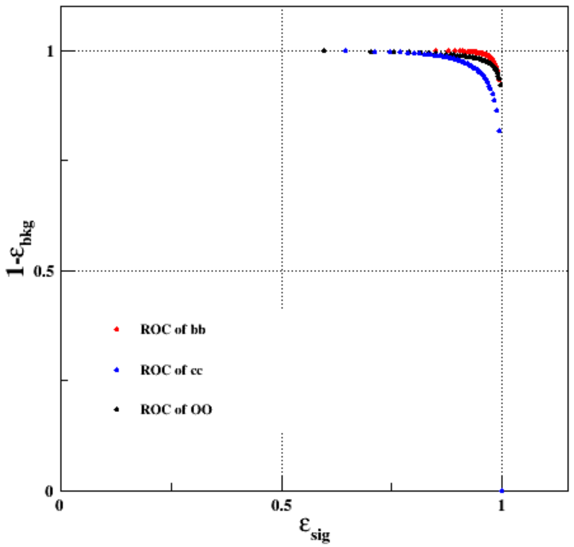
- ➢ The multiplicity versus momenta of tracks
  - Few tracks and more energetic in $\tau\bar{\tau}$
  - Easy to discriminate $q\bar{q}$ & $\tau\bar{\tau}$
- Approximate to diagonal matrix
- Good performance

tag	ParticleNet		PFN	
	Efficiency	AUC	Efficiency	AUC
b	0.961	0.997	0.930	0.993
c	0.905	0.989	0.832	0.976
o	0.950	0.995	0.945	0.992



Log(P) vs. Ntrks





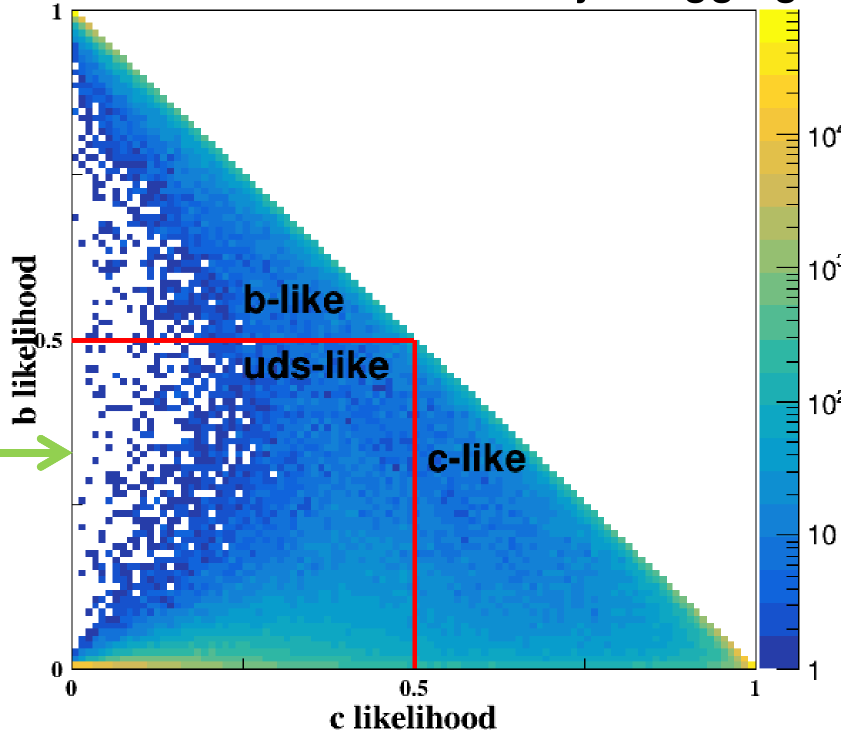
Next: R_q measurement with two methods

R_b & R_c measurement

$$f_s = \varepsilon_b R_b + \varepsilon_c R_c + \varepsilon_{uds} R_{uds}$$

- Double tagging: $f_d = C_b R_b \varepsilon_b^2 + C_c R_c \varepsilon_c^2 + C_{uds} R_{uds} \varepsilon_{uds}^2$
  - Neglect the correlation of jets
  - Choose the working point
  - Solved 6 equations by the least square method

b likelihood vs. *c* likelihood in jet tagging



- Confusion matrix:
  - To solve the right equations
  - Calculate $R_q = \frac{N_q}{\Sigma_i N_i}$

$$\begin{pmatrix} n_1 \\ n_2 \\ n_3 \end{pmatrix} = \begin{pmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{pmatrix} \begin{pmatrix} N_1 \\ N_2 \\ N_3 \end{pmatrix}$$

R_b & R_c measurement

	σ_{R_b}/R_b	σ_{R_c}/R_c	σ_{R_o}/R_o
LEP/SLD	3051.46	17431.73	-
Template fit	5.38	13.37	3.41
Double tag	6.06	6.88	-
Confusion matrix	6.35	8.32	-

- References
  - LEP/SLD:
    a) Limited by statistics & flavor tagging
  - Template fit:
    a) Much larger statistics & flavor tagging in CEPC CDR

- results
  - Double tag & Confusion matrix:
    a) Statistic of 10^{11} Z bosons, same as template fit
    b) Comparable in R_b
    c) Improved more than 60% in R_c measurement

R_b & R_c measurement

- Systematic uncertainty
  - Dominant in future colliders
  - Efficiency
    a) Arise by MC models
    b) Reduced by orders of magnitudes, since much improved knowledge on the production and decay of B & D mesons
  - Correlation between jets
    a) Reduced by improved tagging efficiency
    b) Cancel in confusion matrix since there is no jet clustering

Conclusion

- Two novel deep learning methods are used to enhance the performance of jet flavor tagging
  - Significant improvement in jet tagging, especially for c tagging
  - Maximize the usage of information in a jet/event

- R_q measurement is taken to demonstrate the physics impacts
  - Improved more than 60% in R_c measurement
  - Cancel an important systematic uncertainty by confusion matrix method

Thank you!

Backup

$$N_{bo} = L\sigma_h R_b \epsilon\rho$$

$$R_b = \frac{N_{bo}}{L\sigma_h \epsilon\rho}$$

$$\Delta R_b = \frac{\sqrt{N_{bo}}}{L\sigma_h \epsilon\rho}$$

$$\Delta R_b = \frac{\sqrt{L\sigma_h R_b \epsilon\rho}}{L\sigma_h \epsilon\rho} = \sqrt{R_b/(L\sigma_h \epsilon\rho)}$$

$$(\Delta R_b)^2 = \frac{R_b}{L\sigma_h \epsilon\rho} \propto \frac{1}{\epsilon\rho}$$