The measurement of Rb/Rc with ParticleFlow Network method at the CEPC

Libo Liao, Gang Li, Weiming Song, Shudong Wang, Zhaoling Zhang

December 1, 2021

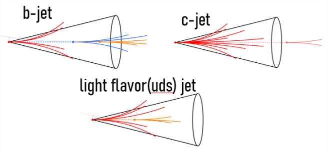
Outline

Motivation

Jets

Jets: Initiated by their ancestral particles via parton shower within a cone.

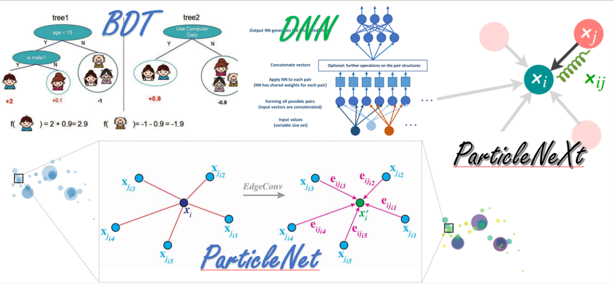
- ◇ Domainant decays of $H/Z/W$;
- ◇ Key physics objects in $H/Z/W$ studies;
- ◇ To understand the QCD process;
- ◇ Important to new physics.



Jet clustering, jet tagging, jet charge, ...

Several methods

Cut-based → TMVA(BDT, XGBoost, etc.) → Machine Learning[1] (DNN, CNN, Graph, Sets, ...)



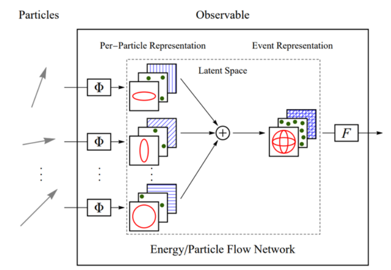
[1] https://github.com/iml-wg/HEPML-LivingReview

ParticleFlow Network[2]

Particle Flow Networks, which allow for general energy dependence and the inclusion of additional particle-level information such as charge and flavor.

$$\hat{O}(p_1, \ldots, p_M) = F\left(\sum_{i=1}^{M} \Phi(p_i)\right)$$

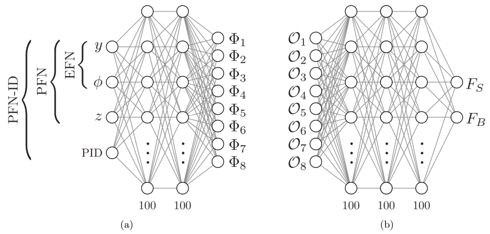
Each particle is mapped by a function Φ to latent space, then mapped by a continuous function F, softmaxed function, to the value of observable.



A visualization of the decomposition of an observable

[2]arXiv:1810.05165

ParticleFlow Network



This is a particular dense neural network to parameterize the functions F and Φ.

The latent observable is $\mathcal{O}_a = \Sigma_i \Phi_a(y_i, \phi_i, z_i, PID_i)$.
The output of F is a softmaxed signal (S) versus background (B) discriminant.

Both applied in single jet-tagging and event-tagging.

Application of jet tagging

Relative decay width(R_b, R_c, R_{uds}):

$$R_{q(q=b,c,uds)} = \frac{\Gamma_{Z\to qq}}{\Gamma_{Z\to had.}}$$

Status of R_b[Bin Yan’ s report [3]]:
Uncertainties of measurement and prediction are both 10^{-4} level.

	measured value	SM prediction
R_b	0.21629 ± 0.00066	0.21578 ± 0.00011

[3] https://ihepco.yonsei.ac.kr/event/130/contributions/531/attachments/394/616/Zbb_BY.pdf

Relative decay width

Different measurement methods:

- ◇ LEP: Double-tagging method
- ◇ Template method: See Bo Li' s report[4];
- ◇ New method: Global analysis[5].

[4] https://indico.ihep.ac.cn/event/14938/session/12/contribution/152/material/slides/0.pdf

[5] https://indico.ihep.ac.cn/event/14938/session/4/contribution/179/material/slides/0.pdf

Analysis methods

Global analysis

A general event selection:

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

Estimating N_i: inverse the equation:

$$\begin{pmatrix} N_1 \\ N_2 \\ N_3 \end{pmatrix} = \begin{pmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{pmatrix}^{-1} \begin{pmatrix} n_1 \\ n_2 \\ n_3 \end{pmatrix}$$

Efficiency matrix is the key.

It is confusion matrix in ML equivariantly.

Global analysis

Compare to double-tagging method and template method, global analysis has several advantages:

- ◇ No necessary to consider the correlations between jet pair, such as $b\bar{b}$ and $c\bar{c}$;
- ◇ Smaller stat. uncertainty with multinomial distribution.

$$B_i = \frac{N_i}{N_1 + N_2 + N_3 + \cdots}, \sigma_N = \sqrt{N \times p \times (1-p)}$$

- ◇ "On-shop" measurement: all R_is measured simultaneously is more efficient.

Preliminary results

Data sets

Full simulation with CEPC_V4 at Z-pole.
85% for training & validation, 15% for test.
For jet tagging

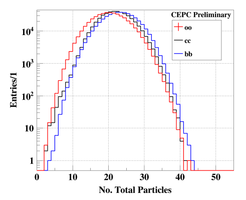
- 900k jets for each flavor(b, c, uds);
- Jet clustering by $ee-kt$ algorithm in the LCFIPlus framework.

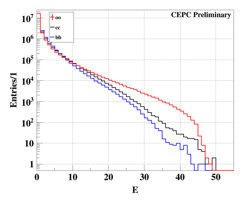
For R_b, R_c, R_{uds} measurement with event-tagging:

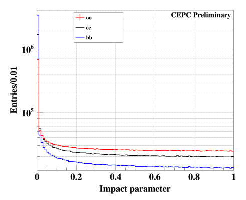
- 450k events for each decay: $Z \to b\bar{b}, Z \to c\bar{c}, Z \to (uds)$;
- No jet clustering.

Distributions of data sets

Variables list: Impact parameter, Energy, Momentum, Charge, Angle information, PID, 12 features for each (charged) particle in total.

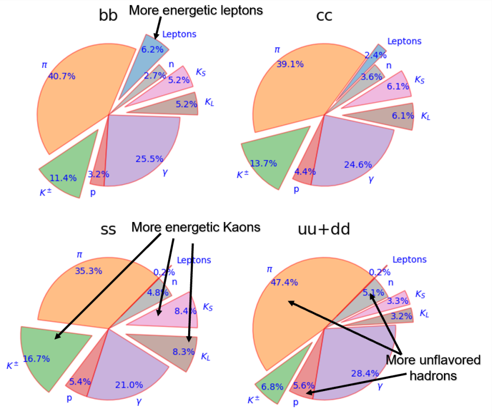




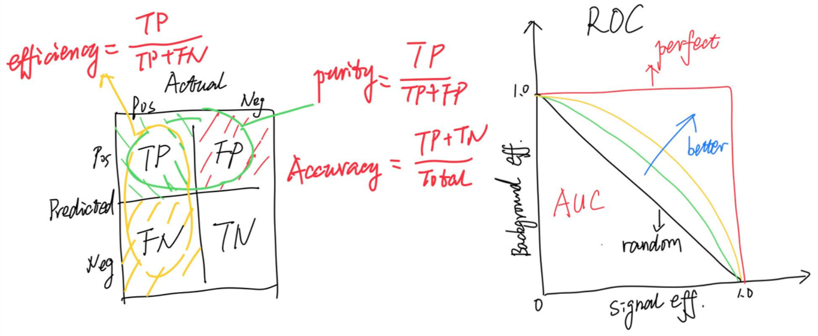


Ingredients of each event

Ingredients of each event weighted by momentum.



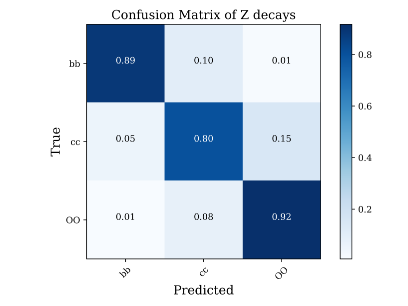
Performence metrics

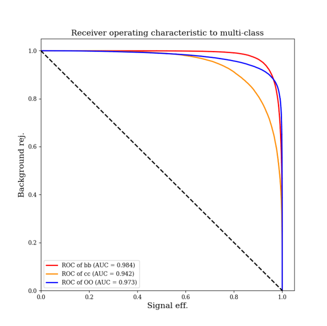


Confusion matrix and ROC curve

The production of efficiency and purity($\epsilon \times \rho$) is popular in HEP.

The preliminary results of jet tagging





The confusion matrix and ROC plot of jet tagging with PFN.

The performence of jet tagging improved significantly, especially for b and uds jets.

Compared with previous study

Algorithm	DNN	BDT	GBDT	gcforest	XGBoost	PFN
Accuracy	0.788	0.776	0.794	0.785	0.801	0.867

Fan Yang's report[6] apart from PFN in November 7, 2017

The best performence is XGBoost, others are close to 0.8. And the accuracy of PFN is 0.867, improved more than 8%.

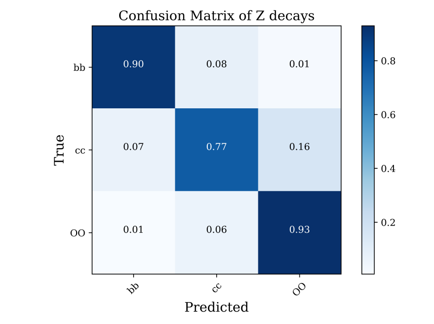
[6] https://indico.ihep.ac.cn/event/6618/session/19/contribution/136/material/slides/0.pdf

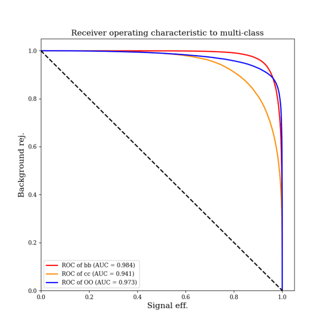
Compared with previous study

tag	ϵ_S	$\epsilon \times \rho$ XGBoost	$\epsilon \times \rho$ PFN	Improvement
b	50%	-	49.9%	-
	90%	79.7%	86.6%	8.7%
	95%	72.8%	87.0%	19.5%
c	50%	-	49.0%	-
	90%	58.5%	74.4%	27.2%
	95%	50.3%	72.4%	43.9%
uds	50%	-	48.7%	-
	90%	-	83.3%	-
	95%	-	85.6%	-

XGBoost versus PFN.

The preliminary results of event-tagging





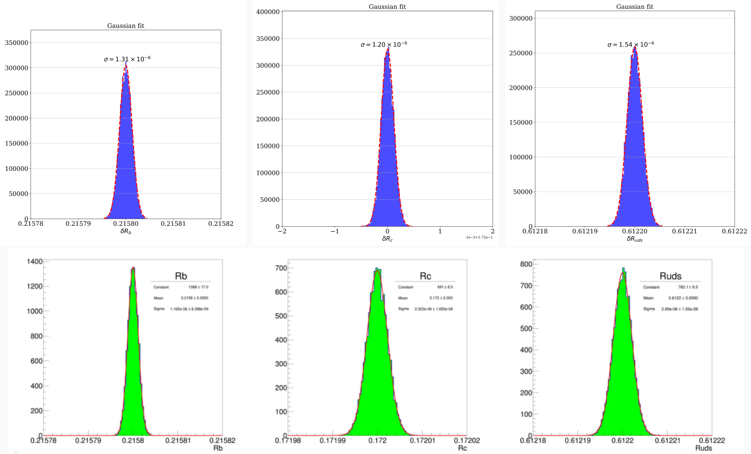
The confusion matrix and ROCs of event-tagging with PFN.

The performence of event-tagging is comparable with jet-tagging.

The preliminary results of R_b & R_c & R_{uds}

Assuming 10^{11} statistics with
$R_b = 0.2158, R_c = 0.1720, R_{uds} = 0.6122 \rightarrow$ the Ni sampled $\rightarrow n_i \rightarrow$ estimate N_i.
Repeat the procedure 10,000 times.

The prelimimary results of R_b & R_c & R_{uds}



The stat. uncertainties of R_b & R_c & R_{uds}

	σ_{R_b}/R_b	σ_{R_c}/R_c	$\sigma_{R_{uds}}/R_{uds}$
LEP	3100	-	-
Bo Li’ s	5.40	13.38	3.41
This study	6.07	6.97	2.51
Diff.	$\downarrow$	$\uparrow$	$\uparrow$

Relative uncertainty(10^{-6}).

CEPC can improve the Rb presicion by a factor of 500.
Difference between this two results should be studied further.

Summary & Outlook

Summary & Outlook

Summary:

- The performence of jet tagging with PFN improved 8%.
- The statistic uncertainty of R_b improved a factor of 500 and helpful to clarify the Rb problem.
- Need more validation for our study.

Outlook:

- Optimize the ML model & feature selection;
- Try more ML architectures, such as ParticleNet;
- Try strange tagging;
- Optimize the 4th concept detector according to jet tagging.

Thank you!