Performance study of the full hadronic WW and ZZ events' separation at the CEPC

Yongfeng Zhu^{1,2}, Manqi Ruan^{1,2,a}

¹ Institute of High Energy Physics, Beijing, China

² University of Chinese Academy of Sciences, Beijing, China

Received: 21 December 2018 / Accepted: 24 February 2019 / Published online: 25 March 2019

© The Author(s) 2019

Abstract The full hadronic WW and ZZ events' separation is an important benchmark for the Circular Electron Positron Collider (CEPC) detector design and reconstruction algorithm development. This separation performance is determined by the intrinsic boson mass distributions, the detector performance, and the jet confusion. The latter refers to the uncertainties induced by the jet clustering and pairing algorithms. Using the CEPC baseline simulation, we demonstrate that the full hadronic WW and ZZ events can be efficiently separated. We develop an analytic method that quantifies the impact of each component and conclude that the jet confusion dominates the separation performance. The impacts of the Initial State Radiation (ISR) and the heavy flavor jets are also analyzed and confirmed to be critical for the separation performance.

1 Introduction

The CEPC is a proposed electron-positron collider with a total circumference of 100 km and two interaction points. It will be operated at center-of-mass energies from 91 GeV to 240 GeV and produces large samples of the W, Z, and Higgs bosons. Its nominal luminosity and massive boson yields are listed in Table 1 [1]. The CEPC can measure most of the Higgs boson properties with accuracies that exceed the ultimate precision of the HL-LHC by one order of magnitude and also boost current precision of the Electroweak (EW) measurements by one order of magnitude. The CEPC could also be upgraded to a proton-proton collider with a center-of-mass energy around 100 TeV.

At 240 GeV center-of-mass energy, the Higgs boson is mainly produced through the ZH process at the CEPC. The

leading di-boson Standard Model (SM) backgrounds for the CEPC Higgs measurements are the WW and ZZ processes, see Fig. 1 [1]. A successful separation between the Higgs signal and the di-boson backgrounds is essential for precise Higgs measurements. In addition, the separation of the WW and ZZ events is important for the QCD measurement, the Triplet Gauge Boson Coupling measurement, and the W boson mass measurement at continuum.

Half of these di-boson events decay into 4-jet final states. The separation between those 4-jet events is determined by the intrinsic boson mass distribution, the detector performance, and the jet confusion. The latter refers to the uncertainties induced by the jet clustering and pairing algorithms. Giving the relatively small mass difference between the W and Z boson, the separation between the WW and the ZZ events in the full hadronic final states is extremely demanding of the detector performance and the jet confusion control. Therefore it serves as a stringent benchmark for the detector design and reconstruction algorithm development. Using the CEPC baseline detector geometry and software, we investigate the separation performance of the full hadronic WW and ZZ events at the full simulation level. We confirm that these events can be clearly separated with the CEPC baseline detector. Through comparative analyses, we quantify the impacts of each component and conclude the jet confusion dominates the separation performance.

This paper is organized into five sections. Section 2 introduces the CEPC baseline detector geometry and software. The analysis method and the separation performance at various conditions are quantified and compared in Sect. 3. Using the Monte Carlo (MC) truth information, Sect. 4 further analyzes the jet confusion. The conclusion is summarized in Sect. 5.

^a e-mail: Manqi.ruan@ihep.ac.cn

Table 1 Running time, instantaneous and integrated luminosities at different values of the center-of-mass energy and anticipated corresponding boson yields at the CEPC. The Z boson yields of the Higgs factory and WW threshold scan operation are from the initial-state radiative

Operation mode	Z factory	WW threshold scan	Higgs factory
$\sqrt{s}$ (GeV)	91.2	158–172	240
Running time (years)	2	1	7
Instantaneous	17–32	10	3
Luminosity ($10^{34} \text{ cm}^{-2} \text{ s}^{-1}$)			
Integrated Luminosity (ab^{-1})	8–16	2.6	5.6
Higgs yield	–	–	10^6
W yield	–	10^7	10^8
Z yield	10^{11-12}	10^8	10^8

2 Detector geometry, software, sample and analysis method

The CEPC uses a Particle Flow oriented detector design as its baseline detector [1]. This detector reconstructs all the visible final state particles in the most-suited detector subsystems. For the CEPC physics measurements, this baseline detector could reconstruct all the core physics objects with high efficiency, high purity, and high precision [1, 2]. From inner to outer, the baseline detector is composed of a silicon pixel vertex detector, a silicon inner tracker, a Time Projection Chamber (TPC) surrounded by a silicon external tracker, a silicon-

tungsten sampling Electromagnetic Calorimeter (ECAL), a steel-Glass Resistive Plate Chambers (GRPC) sampling Hadronic Calorimeter (HCAL), a 3 Tesla superconducting solenoid, and a flux return yoke embedded with a muon detector. The structure of the CEPC detector is shown in Fig. 2.

In fact, the separation of vector boson scattering processes (with $\nu\nu WW$ and $\nu\nu ZZ$ final states) strongly motivates the Particle Flow oriented detector design [3, 4].

A baseline reconstruction software chain has been developed to evaluate the physics performance of the CEPC baseline detector, see Fig. 3. The data flow of CEPC baseline software starts from the event generators of Whizard [5, 6]

Fig. 1 The cross section for unpolarized e^+e^- collision, the right side shows the expected number of events at the nominal parameters of the CEPC Higgs runs at 240 GeV center-of-mass energy [1]

Fig. 2 The CEPC baseline detector. From inner to outer, the detector is composed of a silicon pixel vertex detector, a silicon inner tracker, a TPC, a silicon external tracker, an ECAL, an HCAL, a solenoid of 3 Tesla and a return yoke embedded with a muon detector. In the forward regions, five pairs of silicon tracking disks are installed to enlarge the tracking acceptance

Fig. 3 The information flow of the CEPC software chain

Fig. 4 The display of a reconstructed WW event. This event has 82 final state particles whose energy exceed 0.5 GeV, reconstructed by Arbor. The charged particles are represented by the curves (color represents particle charge) associated with calorimeter clusters. The photons are displayed as cyan straight lines associated with calorimeter clusters

and Pythia [7]. The detector geometry is implemented into the MokkaPlus [8], a GEANT4 [9] based full simulation module. The MokkaPlus calculates the energy deposition in the detector sensitive volumes and creates simulated hits. For each sub-detector, the digitization module converts the simulated hits into digitized hits by convoluting the corresponding sub-detector responses. The reconstruction modules include the tracking, the Particle Flow, and the high-level reconstruction algorithms. The digitized tracker hits are reconstructed into tracks via the tracking algorithms. The Particle Flow algorithm, Arbor [2], reads the reconstructed tracks and the calorimeter hits to build reconstructed particles. High-level reconstruction algorithms reconstruct com-

Table 2 The values of σ_B for different cases

σ_B/GeV	σ_W	σ_Z
GenJet	2.0	2.5
RecoJet	3.8	4.4

posite physics objects such as converted photons, τ s, jets, et al. and identify the flavor of the jets.

Using the CEPC baseline detector geometry and software chain, we simulated inclusive samples of 38k WW and 38k ZZ events. These samples include all the different quark flavors according to the SM decay branching ratios. To simplify the analysis, the interference between WW and ZZ is ignored. To analyze the impact of heavy flavors, we also produce light flavor samples for comparison. These light flavor samples are 30k WW $\rightarrow u\bar{d}u\bar{s}$ or $u\bar{s}u\bar{d}$ and 27k ZZ $\rightarrow u\bar{u}u\bar{u}$ events. Figure 4 displays a reconstructed $e^+e^- \rightarrow WW \rightarrow u\bar{u}s\bar{d}$ event using Druid [14]. All the samples are generated at the center-of-mass energy of 240 GeV.

Starting with the fully reconstructed WW/ZZ events, our analysis uses the jet clustering and pairing algorithm. The reconstructed particles are clustered into four RecoJets using the k_t algorithm for the e^+e^- collisions ($e^+e^-k_t$) with the FastJet package [10]. A minimal χ^2 method is used for the jet pairing. These four RecoJets are paired into two di-jet systems. Their masses are compared with the hypothesis of a WW or a ZZ event via the χ^2 defined as:

$$\chi^2 = \frac{(M_{12} - M_B)^2 + (M_{34} - M_B)^2}{\sigma_B^2}.$$

The quantity M_{12} and M_{34} refer to the masses of di-jet systems, and M_B is the reference mass of the Z or the W boson [11]. The σ_B is the convolution of the boson width and the detector resolution. According to [1], the detector resolution is set to be 4% of the boson mass. The values of the σ_B for different cases are listed in Table 2. Among all six possible combinations (corresponding to three different jet pairings and two values of M_B), the one with the minimal value of the χ^2 determines the event type and corresponding di-jet masses.

Using the same jet clustering and pairing algorithms and parameters for the RecoJets analysis, the visible particles at the MC truth level can be clustered into the GenJets and paired into di-jet systems. Since these GenJets are corresponding to the perfect detector, the separation performance using GenJets characterize the impacts of the intrinsic boson mass distribution and the jet confusion. In this paper, the analyses are performed using both the RecoJets and the GenJets.

Fig. 5 The RecoJet level distributions of $0.5 \times (M_{12} + M_{34})$ of the WW and ZZ events. The overlapping fraction is $57.8\% \pm 0.23\%$

3 Separation performance with overlapping fraction

Using the method introduced above, the masses of the di-jet systems (M_{12} and M_{34}) are calculated. Figure 5 shows the average reconstructed di-jet mass distributions of the inclusive WW and ZZ samples using the RecoJets, each normalized to unit area. Each distribution exhibits a clear peak at the anticipated boson mass and an artificial tail towards the other peak. These tails are induced by the jet pairing algorithm, the neutrinos generated in heavy flavor quark fragmentation, and the ISR photons. The peaks are clearly separated, however, the tails lead to significant confusion between the WW and ZZ events.

The confusion can be evaluated by the overlapping fraction between two distributions:

$$\text{Overlapping Fraction} = \sum_{\text{bins}} \min(a_i, b_i),$$

a_i and b_i are the bin contents of both distributions at the same bin. The overlapping fraction is approximately equal to the sum of two misidentification probabilities ($P_{WW \rightarrow ZZ} + P_{ZZ \rightarrow WW}$). An overlapping fraction of zero means no misidentification.

The overlapping fraction is sensitive to the jet clustering algorithm. In this paper, the jet clustering algorithm is selected via a parameter scan on the generalized k_t algorithm for the e^+e^- collision. This algorithm has two free parameters, the cone radius and the power index on the particle energy, denoted with R and P respectively. The scan shows that the minimal overlapping fraction on the inclusive WW and ZZ sample is achieved with $R = 2$ and $P = 1$, with which the generalized k_t algorithm converges to the $e^+e^-k_t$ algorithm. In addition, we also tried the Valencia

Fig. 6 The RecoJet level distribution of M_{12} versus M_{34}

Fig. 7 The GenJet level distributions of $0.5 \times (M_{12} + M_{34})$ of the WW and ZZ events. The overlapping fraction is $52.6\% \pm 0.25\%$

algorithm [12, 13], which gives similar performance compared to the $e^+e^-k_t$ algorithm.

Figure 5 has an overlapping fraction of $57.8\% \pm 0.23\%$. The correlation of M_{12} versus M_{34} using the RecoJet is shown in Fig. 6, the distributions of the WW and ZZ events are overlapped. Figure 6 has two separable peaks located on a large area of a flat plateau. The latter contributes significantly to the overlapping fraction.

The separation performance at the GenJet level is also analyzed. Figure 7 shows the distributions of the average di-jet mass which has an overlapping fraction of $52.6\% \pm 0.25\%$. Compared to the RecoJet distributions, Fig. 7 exhibits much narrower peaks but similar tails. That's to say, the peak width of the RecoJet distributions are mainly dominated by the

Fig. 8 The GenJet level distribution of M_{12} versus M_{34}

Fig. 9 The RecoJet level distribution of M_{12} versus M_{34} with the equal mass condition. The selection efficiency for WW/ZZ is 54%/44%

detector performance. The correlation between M_{12} versus M_{34} with the GenJets is shown in Fig. 8. Aside from two clearly separable peaks, Fig. 8 also has a plateau with similar contour and area compared to Fig. 6, the distribution at RecoJet level. Clearly, the common patterns of the GenJet and the RecoJet level distributions are induced by the intrinsic boson mass and the jet confusion.

The area of the plateau can be significantly reduced using the fact that WW and ZZ processes produce two equal mass bosons. We define an equal mass condition that requires the mass difference between the two di-jet systems to be smaller than 10 GeV ($|M_{12} - M_{34}| < 10 \text{ GeV}$). This condition vetos roughly half of the statistics. After applying this equal mass condition, the overlapping fractions are improved to $39.9\% \pm$

Fig. 10 The RecoJet level distributions of $0.5 \times (M_{12} + M_{34})$ with the equal mass condition, the overlapping fraction is $39.9\% \pm 0.40\%$

Fig. 11 The GenJet level distribution of M_{12} versus M_{34} with the equal mass condition. The selection efficiency for WW/ZZ is 59%/47%

0.40% and $27.1\% \pm 0.42\%$, corresponding to the RecoJet and the GenJet plots, see Figs. 9, 10, 11 and 12.

The overlapping fractions of the full hadronic WW and ZZ events can be compared with two reference values. The first one is the overlapping fraction of the semi-leptonic di-boson events, where the invariant mass of the hadronic decayed W and Z bosons can be reconstructed without any jet confusion. The second one is the overlapping fraction of the MC truth boson masses, which follow approximately the Breit-Wigner distributions. The first value provides a reference to the jet confusion evaluation, and the second one describes the impact of intrinsic boson mass distributions and is the lower limit of the overlapping fraction.

Fig. 12 The GenJet level distributions of $0.5 \times (M_{12} + M_{34})$ with the equal mass condition, the overlapping fraction is $27.1\% \pm 0.42\%$

Fig. 13 The distribution of the total invariant mass of the hadronic system of $ZZ \rightarrow \nu\nu q\bar{q}$, $ZH \rightarrow \nu\nu(Z)q\bar{q}(H)$, and $WW \rightarrow \mu\nu q\bar{q}$ [1]. The overlapping fraction of WW and ZZ is $47.32\% \pm 0.26\%$

The invariant hadronic mass distributions of semi-leptonic di-boson events ($ZZ \rightarrow \nu\nu q\bar{q}$, $ZH \rightarrow \nu\nu(Z)q\bar{q}(H)$, and $WW \rightarrow \mu\nu q\bar{q}$, inclusive sample) are shown in Fig. 13 [1]. It has clearly separated peaks at the anticipated masses. This semi-leptonic overlapping fraction is $47.3\% \pm 0.26\%$. It is significantly better than the inclusive full hadronic WW and ZZ events using the RecoJets ($57.8\% \pm 0.23\%$), but worse than the events with the equal mass constraint ($39.9\% \pm 0.40\%$).

The overlapping fractions of the MC truth boson masses of WW and ZZ events are extracted. For the full hadronic events, we calculate the average mass of two MC truth bosons and the overlapping fraction is $13.3\% \pm 0.34\%$. For the semi-

leptonic event, we extract the truth level value of the mass of the hadronic decay boson, and the overlapping fraction is 12.5%. In fact, those two values are close to the integration of two ideal Breit-Wigner distribution's overlapping fraction according to the W and the Z boson masses and widths (12%). For simplicity, the average value at full hadronic and semi-leptonic events (12.9%) is used in later discussion.

Energetic neutrinos can be generated via the semi-leptonic decays at the heavy-flavor jet fragmentation, leading to significant missing energy and momentum. At the full hadronic WW and ZZ samples, these energetic neutrinos can disturb the jet clustering and pairing performance and increase the jet confusion. Its impact is quantified using comparative analysis of the light jet sample. Compared to the inclusive sample, the overlapping fraction at light jet sample is reduced by 7.1% (from 39.9 to 32.8%) and 4.6% (from 57.8 to 53.2%), with and without the equal mass condition respectively.

At 240 GeV center-of-mass energy, a significant fraction of the WW and ZZ events have energetic ISR photons in their final states. These ISR photons, once incident into the ECAL ($|\cos(\theta)| < 0.995$ at the CEPC baseline), can be recorded as isolated energetic clusters. Those clusters may also increase of the jet confusion. We define an ISR veto condition that excludes events with ISR photons whose energy exceeds 0.1 GeV. Once applied on the light jet samples, the overlapping fraction can be further reduced by 3.4% (from 32.8 to 29.4%) and 3.6% (from 53.2 to 49.6%), with and without the equal mass condition respectively.

The same analysis is performed also with GenJets and the overlapping fraction is summarized in Table 3 and Fig. 14. Four lines, corresponding to the cases of the GenJet level or the RecoJet level, with or without the equal mass condition, are identified in Fig. 14. To be compared with two horizontal lines corresponding to the overlapping fraction of the truth level boson mass distribution (12.9%) and that of the semi-leptonic sample (47%). Several interesting conclusions can be drawn.

1. For the full reconstructed samples, the WW and ZZ events could be efficiently separated. The separation performance is slightly worse than the semi-leptonic events. However, the separation performance of the full hadronic events can exceed that of the semi-leptonic events, once the equal mass condition is applied.
2. It's actually the jet confusion that dominates the separation performance of the inclusive samples, as the GenJet level samples have already a significant overlapping fraction. The detector performance is significant on the boson peak width, but contributes only marginally to the overall separation performance. For the inclusive samples without the equal mass condition, the overlapping fraction only increases by 5% at the RecoJet level compared to

Table 3 The overlapping fractions with different conditions

	Light sample non energetic ISR	Light sample	Inclusive sample
RecoJet	$49.6\% \pm 0.30\%$	$53.2\% \pm 0.29\%$	$57.8\% \pm 0.23\%$
GenJet	$39.1\% \pm 0.33\%$	$48.9\% \pm 0.30\%$	$52.6\% \pm 0.25\%$
RecoJet with equal mass condition	$29.4\% \pm 0.71\%$	$32.8\% \pm 0.49\%$	$39.9\% \pm 0.40\%$
GenJet with equal mass condition	$16.0\% \pm 0.72\%$	$23.0\% \pm 0.51\%$	$27.1\% \pm 0.42\%$
Reference values			
Semi-leptonic, RecoJet		$47.3\% \pm 0.26\%$	
Intrinsic Boson Mass		$13.3\% \pm 0.34\%$

Fig. 14 The overlapping fractions for different cases. The X-axis indicates the different sample restrictive conditions: the light flavor samples with ISR veto condition, the light flavor samples, and the inclusive samples

that at the GenJet level. Meanwhile, their relative difference becomes more significant once the equal mass condition and other restrictive conditions are applied.

3. The equal mass condition can efficiently veto events contaminated by large jet confusion. After applying the equal mass condition, the overlapping fraction can be improved by roughly 20% for both the RecoJets and the Genjets; for the GenJets with the light jet samples and the ISR photons veto, the overlapping fraction is approaching to the physics lower limit of 12.9%. On the other hand, the equal mass condition has an efficiency of only 50%. The equal mass condition should be regarded as a tool to better understand the origin of the rather large overlapping ratios, while many methods, such as kinematic fits and Multiple Variable Analyses, could lead to better separation performance and higher efficiency.
4. The heavy flavor jets and the ISR photons contribute approximately a constant amount of overlapping fraction for all four different cases. In fact, the accumulated

impact of neutrinos and ISR photons are larger than that of the detector performance: for the light jet sample with the ISR veto, the RecoJet distribution overlapping fraction ($49.6\% \pm 0.30\%$) is smaller than that of the inclusive sample at the GenJet level ($52.6\% \pm 0.25\%$). Collectively, they contribute up to 10% of the overall overlapping fraction on the inclusive sample. Therefore, adequate jet flavor tagging and ISR photon finding algorithm can be applied to significantly improve the separation performance.

4 Quantification of the jet confusion

In this section, we analyze the correlation between the jet confusion and the overlapping fraction using the MC truth information. After the jet clustering and mapping, each di-boson event has two di-jet systems and two MC truth level bosons.

Fig. 15 The correlation of α_1 versus α_2 (unit in radians), the angular difference between reconstructed di-jet systems and the MC truth bosons of the inclusive WW samples

Fig. 16 The distribution of α ($\alpha = \alpha_1 \times \alpha_2$) of the WW sample using the RecoJets. There are four vertical lines to characterize α into five subsamples, each contains 20% of the statistics

The di-jet systems are then associated with the bosons, and the angle between the total momentum of the di-jet system and the MC truth level boson can be calculated. Among two different combinations, the one with the minimal value of the sum of the angles is selected.

These two angles ($\alpha_i = \text{angle}(\text{RecoJetPair}_i, \text{TruthBoson}_i)$, $i = 1, 2$) are used to characterize the jet confusion. Figure 15 shows the correlation of α_1 and α_2 in the inclusive WW sample. For α_1 and α_2 smaller than 0.1 radians, these two quantities are not correlated. The distribution actually reflects the jet angle resolution of the CEPC baseline detector. For α_1 and α_2 larger than 0.1 radians, a strong correlation is observed between these two quantities, corresponding to significant jet confusion.

We quantify the jet confusion using the product $\alpha = \alpha_1 \times \alpha_2$ as the order parameter, which increases with the jet confusion. Figure 16 shows the distribution of $\text{Log}_{10}(\alpha)$ at the RecoJet level, which exhibits a gaussian-like distribution up to $\text{Log}_{10}(\alpha) = -2$ and a flat plateau up to $\text{Log}_{10}(\alpha) = 0.4$. The plateau corresponds to the physics events with large jet confusion.

To quantify the impact of the jet clustering performance, the reconstructed WW sample is divided into five subsamples with equal statistics, see Fig. 16. A set of thresholds on α are extracted. The ZZ samples are divided also into five subsamples using the same thresholds, and the overlapping fractions of the same set of subsamples are calculated.

Figure 18 shows the average di-jet mass distributions of each set at the RecoJet and the GenJet level. Their overlapping fractions increase monotonically with the jet confusion, see Fig. 17. The relative difference between that of the

Fig. 17 The overlapping fractions of different sets sorted according to the jet confusion order parameter α . The red/blue lines is corresponding to the GenJet/RecoJet. The red/brown dashed horizontal line indicates the overlapping fraction of the semi-leptonic sample/intrinsic boson mass distributions, respectively

GenJets and the RecoJets, which reflects the detector performance, became less significant. In the first set - corresponding to 20% of the total statistics with the minimal jet confusion, the overlapping fraction of the GenJets is close to the lower limit, and that of the RecoJets is relatively 76% larger (14.1–24.8%). In the last set, for both GenJets and RecoJets, the distributions of the WW and ZZ events are similar. That's to say, the jet confusion eliminates almost completely the separation power for the last 20% of statistics with the worst jet confusion.

It's interesting that the jet confusion takes on a polarized pattern in this analysis. Sorting the inclusive samples with the jet confusion, the first 40% of the samples have only marginal jet confusion (as the overlapping fraction is close to the lower limit). However, the jet confusion soon grows to be the leading impact factor of WW/ZZ separation, and dominate the overlapping fraction for the last 40% of the samples. The critical point occurs at roughly half of the statistics. This S-curve in Fig. 17 may characterize profoundly the jet clustering and pairing performance, and can be used as a reference for corresponding performance evaluation and algorithm development (Fig. 18).

5 Conclusion

Using the CEPC baseline simulation tool, we analyze the full hadronic WW and ZZ events' separation at the CEPC Higgs runs. This separation performance is determined by

Fig. 18 The average di-jet mass distributions after dividing the inclusive sample into five subsamples. From left to right, the α is degrading. The distributions in the top row are using the RecoJets, the overlapping fraction is $24.8\% \pm 0.81\%$, $27.6\% \pm 0.77\%$, $39.1\% \pm 0.63\%$,

$74.1\% \pm 0.37\%$ and $91.1\% \pm 0.22\%$, respectively. The bottom distributions are corresponding to the GenJets, the overlapping fraction is $14.1\% \pm 0.89\%$, $15.0\% \pm 0.83\%$, $34.0\% \pm 0.65\%$, $74.4\% \pm 0.37\%$ and $91.9\% \pm 0.21\%$, respectively

the intrinsic boson mass distribution, the detector performance, and the jet confusion. We quantify the separation performance using the overlapping fraction and disentangle the impacts of different components through comparative analyses.

We confirm that the full hadronic WW and ZZ events can be clearly separated at the CEPC baseline detector and reconstruction software. Using the RecoJets, the overlapping fraction for the inclusive full hadronic WW and ZZ event samples at the CEPC is $57.8\% \pm 0.23\%$. An equal mass condition can reduce the overlapping fraction to $39.9\% \pm 0.40\%$. The overlapping fractions of the GenJet level distributions are $52.6\% \pm 0.25\%$ and $27.1\% \pm 0.42\%$, with and without the equal mass condition respectively. Though the separation performance with GenJets is significantly better than that with RecoJets, it's still much worse than the physics lower limit of 12.9%, the overlapping ratio of the MC truth boson mass distributions. Therefore, we conclude that the jet confusion plays a dominant role in the WW-ZZ separation with full hadronic final states, especially for the inclusive sample without equal mass condition.

The overlapping fraction for WW and ZZ events with semi-leptonic final state is estimated to be $47.3\% \pm 0.26\%$, which is between that of the inclusive full hadronic samples with and without the equal mass condition ($57.8\% \pm 0.23\%$ and $39.9\% \pm 0.40\%$). In other word, once the jet confusion is under control, the separation performance of the full hadronic events is better than that of semi-leptonic events, since the former can use mass information from both reconstructed bosons with independent detector response.

The neutrinos and ISR photons play an important role in the separation performance. Collectively, they contribute to

roughly 10% of the overall overlapping fraction. Therefore, the jet flavor tagging algorithm and the ISR photon identification algorithm are important for the full hadronic WW and ZZ event separation.

The jet confusion is further characterized by the reconstructed angle of bosons. The full hadronic WW and ZZ samples are divided into subsamples and sorted accordingly. For those subsamples, the jet confusion takes a polarized pattern. For the best 40% of the events, the difference between the reconstructed boson angle and the truth value is smaller than 0.1 radians, and the jet confusion is minimum. The overlapping fraction of the GenJet level distributions is close to the lower limit of 12.9%. The separation of those events are mainly dominated by the detector performance. For the last 40% of events, the jet confusion dominates the separation performance.

Control of the jet-confusion, or more generally, identification of the hadronic decayed color-singlets at multi-jet events, is essential for the physics reach of future Higgs factories. On top of the simple jet clustering and pairing algorithm used in this manuscripts, better color-singlet reconstruction performance is anticipated via the iterative jet clustering, the kinematic fits, the Multiple Variable Analyses, et al. The WW/ZZ separation analysis presented in this paper is an early step of these studies. It not only demonstrates the physics performance of the CEPC baseline but also provides the reference and a simple quantification method to evaluate different color-singlet reconstruction algorithms.

Acknowledgements We are in debt to Jianming Qian, Liantao Wang, Huaxing Zhu, and Haibo Li for their constructive suggestions. We thank Matthew Kurth for the careful reading and polish of this paper. We are grateful to Dan Yu, Xianghu Zhao, Hao Liang, and Yuxuan

Zhang for their supports and helps. We thank Gang Li and Chengdong Fu for producing the samples. This work was supported by National Key Program for S&T Research and Development (Grant No.: 2016YFA0400400), the National Natural Science Foundation of China (Grant No.: 11675202), the Hundred Talent Programs of Chinese Academy of Science (Grant No.: Y3515540U1).

Data Availability Statement This manuscript has associated data in a data repository. [Authors' comment: The data used in this article is generated with official CEPC simulation software, and currently stored at the CEPC data repository in IHEP.]

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. Funded by SCOAP³.

References

1. The CEPC Study Group, CEPC Conceptual Design Report, vol 2—physics and detector. [arXiv:1811.10545](https://arxiv.org/abs/1811.10545) [hep-ex]
2. M.Q. Ruan, Reconstruction of physics objects at the Circular Electron Positron Collider with Arbor. *Eur. Phys. J. C* **78**(5), 426 (2018)
3. T. Behnke et al., The International Linear Collider Technical Design Report, vol 4: detectors (2013). [arXiv:1306.6329](https://arxiv.org/abs/1306.6329)
4. The CLIC Collaboration, CLIC Conceptual Design Report (2012). CERN-2012-007
5. W. Kilian, T. Ohl, J. Reuter, WHIZARD: simulating multi-particle processes at LHC and ILC. *Eur. Phys. J. C* **71**, 1742 (2011)
6. M. Moretti, T. Ohl, J. Reuter, O'Mega: An optimizing matrix element generator, LC-TOOL-2001-040-rev. [arXiv:hep-ph/0102195](https://arxiv.org/abs/hep-ph/0102195)
7. The Pythia Group, An introduction to PYTHIA 8.2. *Comput. Phys. Commun.* **191**, 159–177 (2015). [arXiv:1410.3012](https://arxiv.org/abs/1410.3012) [hep-ph]
8. C.D. Fu, Full Simulation Software at CEPC. <http://cepcdoc.ihep.ac.cn/DocDB/0001/000167/001>. Accessed 23 Oct 2017
9. The GEANT4 Collaboration, S. Agostinelli et al., GEANT4: a simulation toolkit. *Nucl. Instrum. Methods A* **506**, 250–303 (2003)
10. M. Cacciari, G.P. Salam, G. Soyez, *Eur. Phys. J. C* **72**, 1896 (2012)
11. M. Tanabashi et al., Particle Data Group, *Phys. Rev. D* **98**, 030001 (2018)
12. M. Boronat et al., A robust jet reconstruction algorithm for high-energy lepton colliders. *Phys. Lett. B* **750**, 95–99 (2015)
13. M. Boronat et al., A new jet reconstruction algorithm for lepton colliders. [arXiv:1404.4294](https://arxiv.org/abs/1404.4294) [hep-ex] (2014)
14. M.Q. Ruan, Druid: event display for the linear collider. [arXiv:1303.3759](https://arxiv.org/abs/1303.3759) [physics.ins-det]