Jet-Origin Identification and Its Application at an Electron-Positron Higgs Factory

Hao Liang,[1,2,*] Yongfeng Zhu,[3,*] Yuexin Wang,[1,4] Yuzhi Che,[1,2] Manqi Ruan,[1,2,†] Chen Zhou,[3,‡] and Huilin Qu[5,§]

[1]Institute of High Energy Physics, Chinese Academy of Sciences, 19B Yuquan Road, Shijingshan District, Beijing 100049, China
[2]University of Chinese Academy of Sciences, 19A Yuquan Road, Shijingshan District, Beijing 100049, China
[3]State Key Laboratory of Nuclear Physics and Technology, School of Physics, Peking University, Beijing 100871, China
[4]China Center of Advanced Science and Technology, Beijing 100190, China
[5]CERN, EP Department, CH-1211 Geneva 23, Switzerland

(Received 16 October 2023; revised 26 April 2024; accepted 1 May 2024; published 31 May 2024)

To enhance the scientific discovery power of high-energy collider experiments, we propose and realize the concept of jet-origin identification that categorizes jets into five quark species (b, c, s, u, d), five antiquarks $(\bar{b}, \bar{c}, \bar{s}, \bar{u}, \bar{d})$, and the gluon. Using state-of-the-art algorithms and simulated $\nu\bar{\nu}H, H \to jj$ events at 240 GeV center-of-mass energy at the electron-positron Higgs factory, the jet-origin identification simultaneously reaches jet flavor tagging efficiencies ranging from 67% to 92% for bottom, charm, and strange quarks and jet charge flip rates of 7%–24% for all quark species. We apply the jet-origin identification to Higgs rare and exotic decay measurements at the nominal luminosity of the Circular Electron Positron Collider and conclude that the upper limits on the branching ratios of $H \to s\bar{s}, u\bar{u}, d\bar{d}$ and $H \to sb, db, uc, ds$ can be determined to 2×10^{-4} to 1×10^{-3} at 95% confidence level. The derived upper limit for $H \to s\bar{s}$ decay is approximately 3 times the prediction of the standard model.

DOI: 10.1103/PhysRevLett.132.221802

Introduction.—Quarks and gluons are standard model (SM) particles that carry color charges of the strong interaction. Because of the color confinement of quantum chromodynamics (QCD), colored particles cannot travel freely in spacetime and are confined to composite particles like hadrons. Once generated in high-energy collisions, quarks and gluons fragment into numerous particles that travel in directions approximately collinear to the initial colored particles. These collinear particles are called jets; see Fig. 1.

We define jet-origin identification as the procedure to determine from which colored particle a jet is generated and consider 11 different kinds: b, $\bar{b}$, c, $\bar{c}$, s, $\bar{s}$, u, $\bar{u}$, d, $\bar{d}$, and gluon. A successful jet-origin identification is critical for experimental particle physics at the energy frontier. At the Large Hadron Collider, successfully distinguishing quark jets from gluon ones could efficiently reduce the typically large background from QCD processes [2–8]. Jet flavor tagging is essential for the Higgs property measurements at the LHC [6,7,9,10]. The determination of jet charge [11,12] was essential for weak mixing angle measurements at both LEP and LHC [13], is critical for time-dependent CP measurements [14,15], and could have a significant impact on Higgs boson property measurements [16].

We realize the concept of jet-origin identification in physics events at an electron-positron Higgs factory using a GEANT4-based simulation [17] (referred to as full simulation for simplicity), since the electron-positron Higgs factory is identified as the highest-priority future collider project [18,19]. We develop the necessary software tools, Arbor [20,21] and ParticleNet [22], for the particle flow event reconstruction and the jet-origin identification. We

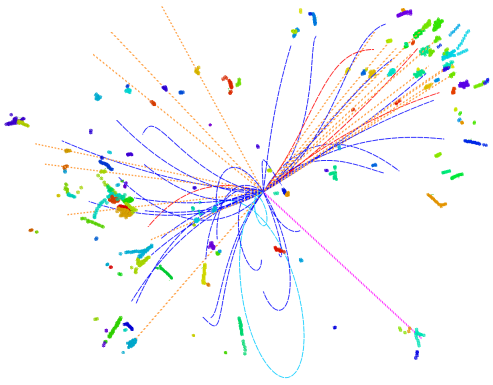

FIG. 1. Event display of an $e^+e^- \to \nu\bar{\nu}H \to \nu\bar{\nu}gg$ ($\sqrt{s} = 240$ GeV) event simulated and reconstructed with the CEPC baseline detector [1]. Different particles are depicted with colored curves and straight lines: red for $e^{\pm}$, cyan for $\mu^{\pm}$, blue for $\pi^{\pm}$, orange for photons, and magenta for neutral hadrons.

demonstrate the jet-origin identification performance using an 11-dimensional confusion matrix (referred to as M_{11} for simplicity), which exhibits the performance of jet flavor tagging and jet charge measurements. We apply the jet-origin identification to rare and exotic Higgs boson decay measurements under the Circular Electron Positron Collider (CEPC) nominal Higgs operation scenario. This scenario expects an integrated luminosity of 20 ab^{-1} at $\sqrt{s} = 240$ GeV and could accumulate 4×10^6 Higgs bosons [19,23]. We analyze the rare decays $H \to s\bar{s}$, $u\bar{u}$, and $d\bar{d}$ and the flavor-changing neutral current (FCNC) decays $H \to sb$, ds, db, and uc (here, sb denotes $s\bar{b}$ or $\bar{s}b$, and similarly for ds, db, and uc). We derive upper limits ranging from 10^{-3} to 10^{-4} for these seven processes. In the SM, the predicted branching ratio for the $H \to s\bar{s}$ process is 2.3×10^{-4} [24], and the derived upper limit corresponds to 3 times the SM prediction. The branching ratios for $H \to u\bar{u}$ and $d\bar{d}$ are expected to be smaller than 10^{-6} [24–27], while branching ratios for the above-mentioned FCNC processes are expected to be smaller than 10^{-7} from loop contributions [28].

Detector geometry and software tools.—We simulate $\nu\bar{\nu}H, H \to u\bar{u}$, $d\bar{d}$, $s\bar{s}$, $c\bar{c}$, $b\bar{b}$, and gg processes at 240 GeV center-of-mass energy with the CEPC baseline detector [1]. The CEPC baseline detector design is a particle-flow-oriented concept composed of a high-precision vertex system, a large-volume gaseous tracker, high granularity calorimetry, and a large-volume solenoid. We use PYTHIA6.4 [29] for the event generations and MOKKAPLUS [30,31] for the GEANT4-based detector simulation [17]. The simulated samples are processed with the Arbor particle flow algorithm that reconstructs all final-state particles and identifies their species. The reconstructed final-state particles in a physics event are clustered into two jets using the e^+e^--k_t algorithm [32,33]. For each jet, the kinematic and species information of all its final-state particles, including the track impact parameters associated with charged final-state particles, are input to a modified ParticleNet algorithm. The algorithm calculates the likelihoods corresponding to 11 different jet categories. For each process, one million physics events are simulated, where 600 000 events are used for training, 200 000 for validation, and 200 000 for testing. The model is trained for 30 epochs, and the epoch demonstrating the best accuracy on the validation sample is selected and applied to the testing sample to extract the numerical results.

Information on the species of the final-state particles is critical for jet-origin identification. We compare three scenarios to understand the impact of particle identification. The first scenario assumes perfect identification of charged leptons; i.e., $e^\pm$ and $\mu^\pm$ can be perfectly differentiated from each other and from charged hadrons. The second scenario further assumes perfect identification of the species of charged hadrons (proton, antiproton, $\pi^\pm$, and $K^\pm$). On top of the second scenario, the third one assumes perfect identification of K_S^0 and K_L^0. For simplicity, the assignment of particle identification is based on MC truth. On the other hand, full simulation performance studies show that the CEPC baseline detector could identify leptons with an efficiency of 99.5% with a hadron-to-lepton misidentification rate below 1% [34,35]. It could also distinguish different species of charged hadrons ($\pi^\pm$, $K^\pm$, proton, and antiproton) to better than 2σ [36–38] and reconstruct K_S^0 and Λ with a typical efficiency (purity) of 80% (90%) if they decay into charged particles [39]. Therefore, the second scenario is used as the default one, since it matches the CEPC baseline detector performance, while the third scenario is used for comparison, as the K_L^0 identification remains challenging.

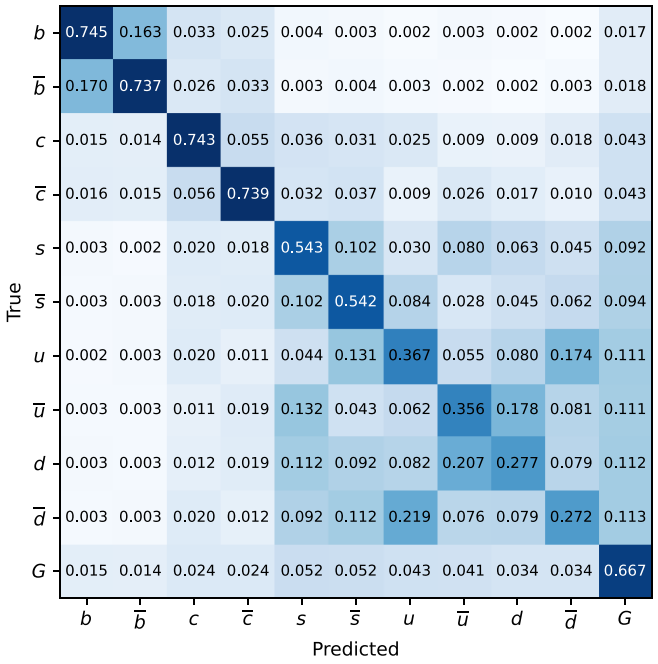

FIG. 2. The confusion matrix M_{11} with perfect identification of leptons and charged hadrons for $\nu\bar{\nu}H, H \to jj$ events at 240 GeV center-of-mass energy. The matrix is normalized to unity for each truth label (row).

Figure 2 shows the overall jet-origin identification performance with an 11-dimensional confusion matrix M_{11}, derived by classifying each jet into the category with the highest likelihood. In the quark sector, M_{11} is approximately symmetric and block diagonalized into 2×2 blocks, corresponding to each specific species of quark. Meanwhile, gluon jets can be identified with an efficiency of 67%.

The performance of the jet-origin identification can be studied in more detail via jet flavor tagging efficiencies and charge flip rates. For each jet, we compare the gluon likelihood and the five sums of quark and antiquark likelihoods of every kind. The jet flavor is then defined as the kind with the highest value. The jet charge is determined by comparing the likelihoods between the

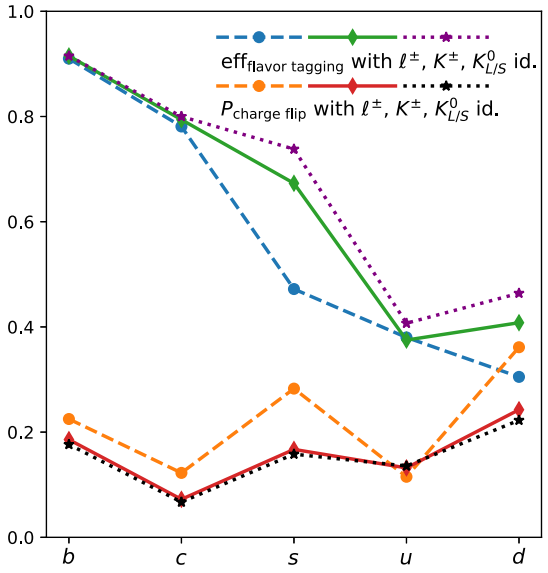

FIG. 3. Jet flavor tagging efficiencies and charge flip rates with perfect identification of leptons (the first scenario, denoted as $\ell^{\pm}$ in the legend) plus identification of charged hadrons (the second and default scenario, denoted as $K^{\pm}$) and neutral kaons (the third scenario, denoted as $K^0_{L/S}$).

quark and the antiquark. Figure 3 illustrates the derived jet flavor tagging efficiencies and charge flip rates, which slightly differ from M_{11} due to the different procedure described above.

Figure 3 additionally compares the performance under different particle identification scenarios. In the default scenario, represented by the solid lines, the $b/c/s$ jets could attain tagging efficiencies of 92%/79%/67% and charge flip rates of 19%/7%/17%, respectively. The identification of u and d jets is less accurate, amounting to tagging efficiencies of 37%–41% and jet charge flip rates of 13%–24%. Noticeably, the down-type jets have a significantly higher jet charge flip rate than the up-type jets, since the latter carries twice the absolute charge as the former. Of all types, the c jets have the lowest charge flip rate, as they are heavier and of the up type. Figure 3 also exhibits the impact of final-state particle identification on jet-origin identification. Compared to the scenario with only lepton identification, introducing charged hadron identification (the default scenario) enhances the s-tagging efficiency from 47% to 67%. Concurrently, it reduces the jet charge flip rates across all types except for u. Additionally, it significantly improves the d-tagging efficiency. The third scenario that includes neutral kaon information further enhances the s-tagging efficiency to 74%. However, the jet charge flip rates remain the same as in the second scenario, since K^0_S and K^0_L are superpositions of $|s\bar{d}\rangle$ and $|\bar{s}d\rangle$ states, meaning their identification has no impact on distinguishing quarks from antiquarks.

Benchmark physics analyses.—The precise measurement of Higgs boson properties is a central objective for particle physics. The anticipated precision of Higgs measurements at future Higgs factories has been extensively studied, showing that the major SM decay modes can be measured with a relative accuracy of 0.1%–1% at electron-positron Higgs factories [19,40–42], surpassing the expected precision at the High Luminosity-LHC (HL-LHC) by one order of magnitude [43]. Meanwhile, the rare and FCNC decays of the Higgs boson are of great interest to many new physics models [24,28,44–47].

We explore the anticipated upper limits of $H \to s\bar{s}$, $u\bar{u}$, $d\bar{d}$ and $H \to sb$, ds, db, uc at the CEPC, where Higgs bosons are mainly produced via the Higgsstrahlung (ZH) and vector boson fusion ($e^+e^- \to \nu_e\overline{\nu}_e H$, $e^+e^- \to e^+e^- H$) processes [48]. Our simulation analyses focus on the $\nu\bar{\nu}H$,

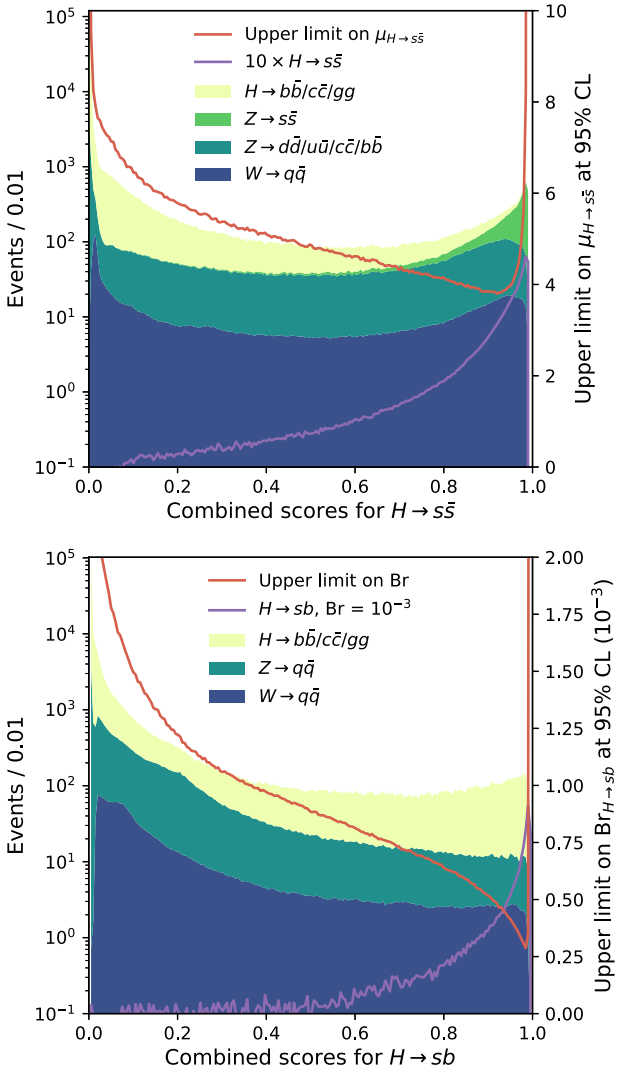

FIG. 4. The distributions of combined scores for signal and SM backgrounds, where the signals are (upper panel) $H \to s\bar{s}$ and (lower panel) $H \to sb$, respectively, in the $\nu\bar{\nu}H$ process, with CEPC nominal parameters.

TABLE I. Summary of background yields from $H \to b\bar{b}/c\bar{c}/gg$, Z, and W prior to the flavor-based event selection, along with the expected upper limits on Higgs decay branching ratios at 95% CL under the background-only hypothesis.

	Bkg (10^3)			Upper limits on Br (10^{-3})						
	H	Z	W	$s\bar{s}$	$u\bar{u}$	$d\bar{d}$	sb	db	uc	ds
$\nu\bar{\nu}H$	151	20	2.1	0.81	0.95	0.99	0.26	0.27	0.46	0.93
$\mu^+\mu^-H$	50	25	0	2.6	3.0	3.2	0.5	0.6	1.0	3.0
e^+e^-H	26	16	0	4.1	4.6	4.8	0.7	0.9	1.6	4.3
Comb.	⋯	⋯	⋯	0.75	0.91	0.95	0.22	0.23	0.39	0.86

$\mu^+\mu^-H$, and e^+e^-H channels, with expected event yields of 0.926×10^6, 0.135×10^6, and 0.141×10^6 under the CEPC nominal Higgs operation scenario, respectively.

We begin with the existing analyses of $\nu\bar{\nu}H, H \to b\bar{b}$, $c\bar{c}$, gg [49,50] at a center-of-mass energy of 240 GeV. These analyses consist of two stages: The first stage performs event selection to concentrate the Higgs to dijet signal in the entire SM data sample, and the second stage identifies different flavor combinations using the LCFIPlus [32] flavor tagging algorithm. For the Higgs rare and exotic decay analyses, we reoptimize the event selections in the first stage and replace the flavor tagging in the second stage with the jet-origin identification. After the event selections (described briefly in Appendix A), the leading SM backgrounds are mainly $\ell\bar{\nu}_\ell W$, $\nu\bar{\nu}Z$, and $\ell^+\ell^-Z$ events. Taking the $\nu\bar{\nu}H, H \to jj$ analyses as an example, the event selection in this stage has a final signal efficiency of 24% and reduces the backgrounds by 6 orders of magnitude, leading to a background yield of 23 000. A toy MC simulation is then applied to the remaining events to mimic the jet-origin identification, by sampling the 11 likelihoods of each jet according to its origin. A gradient boosting decision tree (GBDT) classifier [51] is trained to distinguish signal and background processes using the 22 likelihoods of the jet pair in a physics event.

For the $\nu\bar{\nu}H, H \to s\bar{s}$ analysis, the combined GBDT scores of the remaining events are illustrated in the upper panel in Fig. 4. Defining the signal strength as the ratio of the observed event yield to the SM prediction, the anticipated upper limit on the signal strength of $H \to s\bar{s}$ at 95% confidence level (CL) [52,53] as a function of cut value is shown in Fig. 4. With the optimal cut on the combined scores, there remain 37 events of $H \to s\bar{s}$ and 5100 background events, leading to an upper limit of 3.8 on the signal strength of $H \to s\bar{s}$ at 95% CL. A fit to the combined score distributions further improves the upper limit to 3.5. Combined with e^+e^-H and $\mu^+\mu^-H$ channels, an expected upper limit of 3.2 on the signal strength is achieved at 95% CL. It is worth noting that, in the analysis of $H \to s\bar{s}$, the branching ratios of all other Higgs decays are assumed to be at their SM predictions.

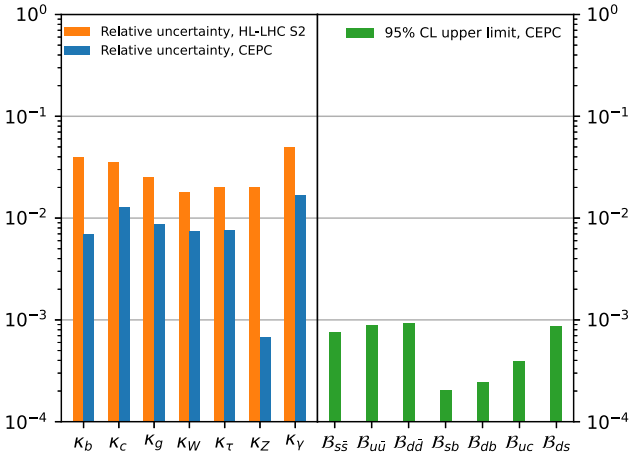

FIG. 5. Expected upper limits on the branching ratios of rare Higgs boson decays from this Letter (green) and the relative uncertainties of Higgs couplings anticipated at CEPC [19] (blue) and HL-LHC [43] (orange) under the kappa-0 fit scenario [54] and scenario S2 of systematics [55], as cited in Ref. [19]. The limit on $\mathcal{B}_{s\bar{s}}$ corresponds to an upper limit of 1.7 on the Higgs-strange coupling modifier κ_s (not shown).

We analyze $H \to u\bar{u}$ and $H \to d\bar{d}$ decay modes using the same method. By combining all three channels, the branching ratios of $H \to u\bar{u}$ and $d\bar{d}$ can be constrained to 0.091% and 0.095% at 95% CL, respectively. These results are less stringent than those for $H \to s\bar{s}$, since the identification of the u and d jets is much more challenging than s jets. We also analyze $H \to sb$, ds, db, and uc decay modes and obtain upper limits ranging from 0.02% to 0.1% for these decay modes. These results are summarized in Table I and Fig. 5.

Discussion and summary.—We propose the concept of jet-origin identification that distinguishes jets generated from 11 types of colored SM particles. State-of-the-art algorithms are developed to realize the concept of jet-origin identification at the future electron-positron Higgs factory, achieving jet flavor tagging efficiencies ranging from 67% to 92% for bottom, charm, and strange quarks and jet charge flip rates of 7%–24% for all species of quarks.

We analyze the impact of final-state particle identification on jet-origin identification and find that charged hadron identification is critical for both jet flavor tagging and charge measurement. The identification of neutral kaons further enhances jet flavor tagging performance but has no impact on jet charge measurement, as expected.

Utilizing jet-origin identification, we estimate the upper limits for seven rare and FCNC hadronic decay modes of the Higgs boson. We conclude that the branching ratios of these decay modes could be constrained to 0.02%–0.1% at 95% CL in the nominal CEPC Higgs operation scenario. For the $H \to s\bar{s}$ decay, the expected upper limit is approximately 3 times the SM prediction, which improves by more than a factor of 2 upon previous studies [24,45].

The improvement here is largely attributed to our state-of-the-art jet-origin identification algorithm, which is capable of exploiting the information of all particles in a jet, not just the kaon particles. The upper limits for $H \to u\bar{u}/d\bar{d}$ can be interpreted as constraints on the Higgs-quark couplings of < 101 and < 37 times the SM predictions, respectively (i.e., $\kappa_u < 101$ and $\kappa_d < 37$). This improves upon existing analyses by roughly one order of magnitude. Regarding the Higgs-boson FCNC decay, a previous study using DELPHES [56] fast simulation indicated that the branching fraction for $H \to sb$ ($H \to db$) could be constrained to 10^{-2} with an integrated luminosity of 30 ab^{-1} [57], while our results show an improvement of 2 (one) orders of magnitude. We also quantify the upper limits for $H \to uc$ and $H \to ds$ in Table I.

Many systematic and theoretical uncertainties are relevant to jet-origin identification, including detector performance, beam-induced backgrounds, the number of pileup events, jet kinematics, jet clustering algorithms, hadronization models, etc. Appendix B summarizes a series of relevant comparison studies. In short, we conclude that jet-origin identification performance is stable with respect to jet kinematics in the relevant energy range; see Fig. 6. We observe that the performances obtained from hadronic Z processes at 91.2 GeV and the $\nu\bar{\nu}H$ processes at 240 GeV are statistically consistent within the detector's fiducial region, as shown in Figs. 7 and 8. In other words, the jet-origin identification could be calibrated using the large number of events in the Z-pole sample to control the performance-relevant systematic uncertainties for the physics measurements including the Higgs property measurements. We observe comparable performance for different hadronization models with small but visible differences; see Fig. 9. These analyses lay the foundation for the application of jet-origin identification at the energy frontier, especially in physics measurements with relatively larger statistical uncertainties, while more dedicated studies are certainly needed.

The jet-origin identification algorithm reads critical information from all the reconstructed particles and provides much higher separation power between jets stemming from different species of colored SM particles. Consequently, this could significantly enhance the scientific discovery potential for physics measurements with multijet final states, such as those expected at future Higgs factories. Jet-origin identification appreciates a detector capable of efficiently distinguishing final-state particles and identifying their species information, as demonstrated in Fig. 3. Recent studies also suggest that a light, precise vertex detector located close to the interaction point is favorable for jet-origin identification [58,59]. Coevolving with state-of-the-art detector technology, reconstruction algorithms, and artificial intelligence, the jet-origin identification algorithm developed here indicates that colored SM particles could potentially be identified with comparable performance to leptons and photons.

We thank Christophe Grojean and Michele Selvaggi for the delightful discussions and Qiang Li, Gang Li, Congqiao Li, Yuxuan Zhang, and Sitian Qian for their support with the software tools. We are grateful to Xiaoyan Shen for her continuous support. This work is supported by the Innovative Scientific Program of the Institute of High Energy Physics, the National Natural Science Foundation of China under Grant No. 12342502, and the Fundamental Research Funds for the Central Universities, Peking University. We appreciate the Computing Center at the Institute of High Energy Physics for providing the computing resources.

Appendix A: Event selection of benchmark analyses.—This appendix describes the event selection for physics benchmark analyses presented in the Letter.

TABLE II. The event selection of $\nu\bar{\nu}H(H \to q\bar{q}/gg)$ when CEPC operates as a Higgs factory at the center-of-mass energy of 240 GeV and collects an integrated luminosity of 20 ab^{-1}. The $\gamma\gamma$ label is the abbreviation of $\gamma\gamma \to$ hadron process, and SW/SZ refers to single W and single Z processes. The units for mass, energy, and momentum are GeV/c^2, GeV, and GeV/c, respectively.

	$\nu\bar{\nu}Hq\bar{q}/gg$	$2f/\gamma\gamma$	SW/SZ	WW/ZZ	ZH
Total	6.4×10^5	4.6×10^9	1.1×10^8	2.8×10^8	3.4×10^6
$M_{\text{recoil}} \in (74, 131)$	5.6×10^5	2.8×10^8	1.4×10^7	2.4×10^7	2.7×10^5
$E_{\text{vis}} \in (109, 143)$	5.1×10^5	1.3×10^8	8.8×10^6	6.4×10^6	1.8×10^5
$E_{\text{leading lepton}} \in (0, 42)$	5.1×10^5	1.2×10^8	4.0×10^6	1.4×10^7	1.7×10^5
Multiplicity $\in (40, 130)$	5.1×10^5	1.0×10^8	2.7×10^6	1.3×10^7	1.5×10^5
$E_{\text{leading neutral}} \in (0, 41)$	5.0×10^5	9.2×10^7	2.5×10^6	1.2×10^7	1.5×10^5
$P_T \in (20, 60)$	4.3×10^5	8.9×10^5	1.4×10^6	6.9×10^6	1.3×10^5
$P_l \in (0, 50)$	4.2×10^5	1.9×10^5	6.4×10^5	3.0×10^6	1.2×10^5
$-\log_{10}(y_{23}) \in (3.375, +\infty)$	3.4×10^5	1.5×10^5	3.1×10^5	1.1×10^6	3.8×10^4
$M_{\text{invariant}} \in (110, 134)$	2.6×10^5	8.1×10^4	6.2×10^4	3.3×10^5	2.5×10^4
BDT $\in (0.1, +\infty)$	1.5×10^5	1.2×10^4	3.6×10^3	6.3×10^3	1.4×10^3

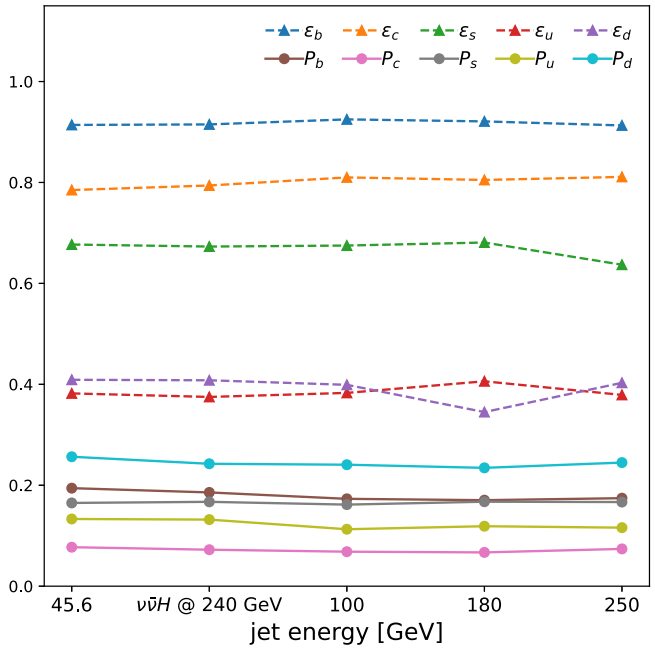

FIG. 6. The jet-origin identification performance: jet flavor tagging efficiencies (ε) and charge flip rates (P) for various jet energies. The error for each value is less than the per-thousand level.

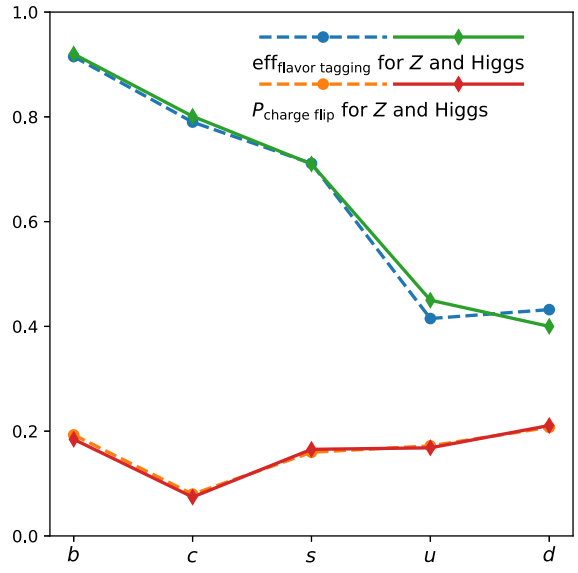

FIG. 7. The comparison of flavor tagging efficiencies and charge flip rates between the $Z \to q\bar{q}$ process (dashed lines) at 91.2 GeV center-of-mass energy and the $\nu\bar{\nu}H, H \to q\bar{q}$ process (solid lines) at 240 GeV. This result is obtained using a ten-category classification for quarks instead of an 11-category classification that includes a gluon category as presented in the main text.

We take as reference the existing full-simulation analysis of $\nu\bar{\nu}H, H \to b\bar{b}$, $c\bar{c}$, gg at the CEPC [49]. This reference simulation analysis considers a nominal luminosity of 5.6 ab^{-1}. It includes all major SM backgrounds, with a total of 4.6×10^7 physics events simulated and processed using the CEPC baseline software, and concludes that the signal strength of the $\nu\bar{\nu}H, H \to b\bar{b}$, $c\bar{c}$, gg processes can be measured with a relative precision of 0.49%, 5.8%, and 1.8%, respectively.

All benchmark analyses of $\nu\bar{\nu}H, H \to jj$ in this Letter use the same kinematic variables for the event selection as in the reference analysis. These kinematic variables include total recoil mass (M_{recoil}), total visible mass ($M_{\text{invariant}}$), total visible energy (E_{vis}), total transverse momentum (P_T), energies of the leading lepton candidate and leading neutral particle, and the Durham distance y_{23} [33] that describes the event topology. A loose cut is applied to the sample, with an efficiency of 40% on the $\nu\bar{\nu}H, H \to jj$ process and a reduction of the background to 495 000. A BDT cut that combines these kinematic and topological variables is applied, which further suppresses the SM background to 23 000 and has an efficiency of 24% on the $\nu\bar{\nu}H, H \to jj$ signal; see Table II.

The remaining events are then processed with toy MC to mimic the jet-origin identification and the GBDT classifier, leading to the distribution shown in Fig. 4.

Appendix B: Comparative analyses of jet-origin identification.—This appendix compares the performance of jet-origin identification for different samples. These

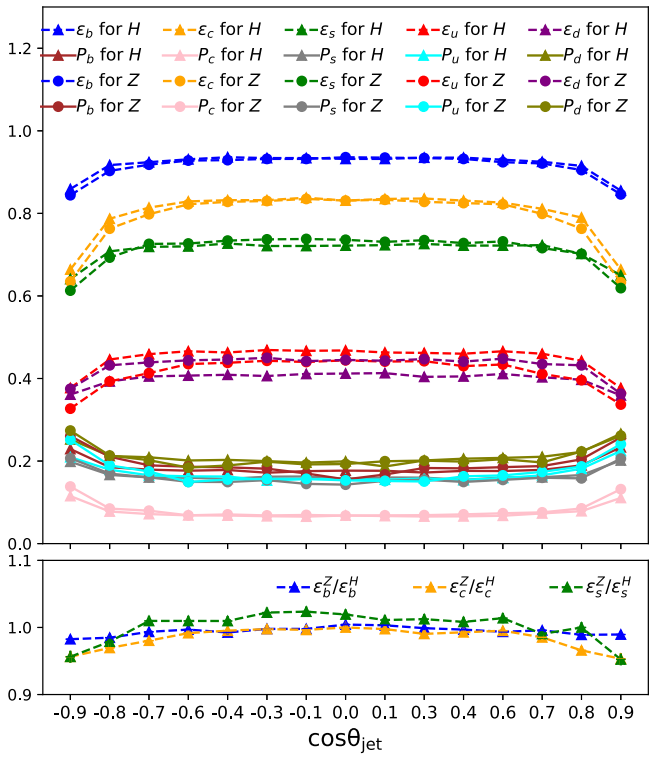

FIG. 8. The jet flavor tagging efficiencies (ε) and charge flip rates (P) at different jet polar angles, corresponding to both the $Z \to jj$ process at 91.2 GeV and the $\nu\bar{\nu}H, H \to jj$ process at 240 GeV. The lower panel displays the ratios of flavor tagging efficiencies for b, c, and s jets between these processes, showing the relative differences at the few-percent level, comparable to the statistical uncertainties.

samples are all full simulation samples using the CEPC baseline detector geometry and perfect lepton and charged hadron identification corresponding to the default scenario of particle identification.

Dependence on the jet energy and jet polar angle: We extract the jet flavor tagging efficiencies and charge flip rates for various jet energies and polar angles. On top of the $\nu\bar{\nu}H, H \to jj$ sample at 240 GeV center-of-mass energy, we simulate a Higgs boson at rest with changing mass, and the Higgs boson is forced to decay into a pair of jets. The Higgs boson mass is set to be 91.2, 200, 360, and 500 GeV, corresponding to jets with energies from 45.6 to 250 GeV. Figure 6 shows the performance at different jet energies, where the extracted jet tagging efficiencies and charge flip rates are rather stable. Figure 8 shows the performance versus the jet polar angle, which is flat in the barrel region of the detector ($|\cos\theta| < 0.8$) and exhibits slight degradation in the end cap region.

Comparison between different physics processes: We compare the jet-origin identification performance between the $Z \to q\bar{q}$ process at a center-of-mass energy of 91.2 GeV and the $\nu\bar{\nu}H, H \to q\bar{q}$ process at 240 GeV center-of-mass energy. We observe that the jet-origin identification performance agrees between these processes, especially in the fiducial barrel region of the detector for the flavor tagging performance of b, c, and s; see Figs. 7 and 8.

It should be noted that, since the Z boson does not decay into a pair of gluons, the gluon jet calibration is an open and interesting question, where dedicated QCD studies and usage of hadron collider data could be very helpful.

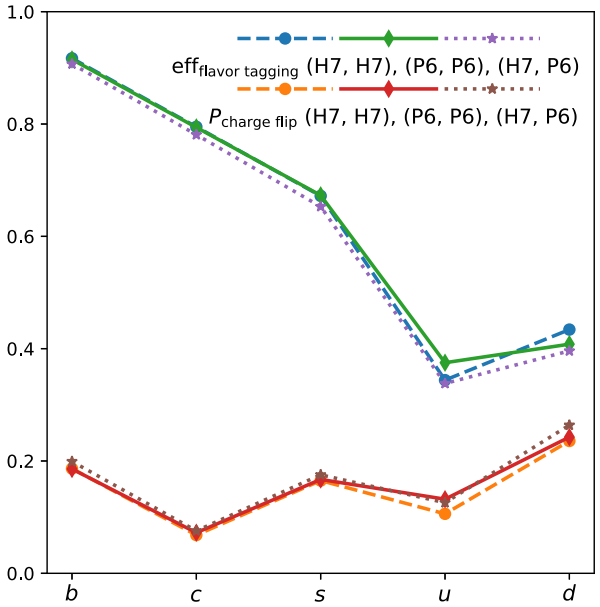

FIG. 9. The performance comparison of flavor tagging efficiencies and charge flip rates of the $\nu\bar{\nu}H, H \to jj$ process at 240 GeV center-of-mass energy using PYTHIA6.4 (P6) and HERWIG7.2.2 (H7). The legend brackets, i.e., (H7, P6), refer to the setup with training samples generated by Herwig and test samples generated by Pythia.

Comparison between different hadronization models: Jet-origin identification uses directly the information of reconstructed final-state particles, while the hadronization process is responsible for generating final-state particles from initial quarks or gluons. The dependence of jet-origin identification performance on the hadronization model is a natural concern.

We compare the jet-origin identification performance of samples derived from different hadronization models, namely, PYTHIA6.4 and HERWIG7.2.2 [60,61]. The predictions of the multiplicity of different final-state particles of these two hadronization models could be different by roughly 10% [62]. Figure 9 shows the performance with different training and test samples. To first order, the performance agrees between models, especially for b, c, and s jets. The performance exhibits small but visible differences for u and d jets.

These comparative analyses show that the jet-origin identification performance, especially for the heavy and strange quarks, is rather stable versus the jet kinematics (in the relevant energy range), different physics processes, and even different hadronization models. The observed stability is vital for applying jet-origin identification in real experiments. Meanwhile, it is a critical and challenging task to determine and validate the fragmentation behavior of colored particles at a future Higgs factory.

*These authors contributed equally to this work.
†ruanmq@ihep.ac.cn
‡czhouphy@pku.edu.cn
§huilin.qu@cern.ch

[1] CEPC Study Group, CEPC conceptual design report: Volume 2—physics & detector, arXiv:1811.10545.
[2] J. Gallicchio and M. D. Schwartz, Phys. Rev. Lett. **107**, 172001 (2011).
[3] CERN, CERN yellow reports: Monographs, Vol. 2 (2017): Handbook of LHC Higgs cross sections: 4. Deciphering the nature of the Higgs sector, 10.23731/CYRM-2017-002 (2017).
[4] R. Gauld, A. Huss, and G. Stagnitto, Phys. Rev. Lett. **130**, 161901 (2023).
[5] M. Aaboud *et al.* (ATLAS Collaboration), Phys. Lett. B **786**, 59 (2018).
[6] A. M. Sirunyan *et al.* (CMS Collaboration), Phys. Rev. Lett. **121**, 121801 (2018).
[7] A. Tumasyan *et al.* (CMS Collaboration), Phys. Rev. Lett. **131**, 061801 (2023).
[8] E. M. Metodiev and J. Thaler, Phys. Rev. Lett. **120**, 241602 (2018).
[9] G. Aad *et al.* (ATLAS Collaboration), Eur. Phys. J. C **81**, 537 (2021).
[10] G. Aad *et al.* (ATLAS Collaboration), Eur. Phys. J. C **82**, 717 (2022).
[11] Z.-B. Kang, A. J. Larkoski, and J. Yang, Phys. Rev. Lett. **130**, 151901 (2023).

[12] H. Cui, M. Zhao, Y. Wang, H. Liang, and M. Ruan, arXiv:2306.14089.
[13] A. M. Sirunyan *et al.* (CMS Collaboration), Eur. Phys. J. C **78**, 701 (2018).
[14] K. F. Chen *et al.* (Belle Collaboration), Phys. Rev. Lett. **98**, 031802 (2007).
[15] R. Aaij *et al.* (LHCb Collaboration), Eur. Phys. J. C **79**, 706 (2019); **80**, 601(E) (2020).
[16] H. T. Li, B. Yan, and C. P. Yuan, Phys. Rev. Lett. **131**, 041802 (2023).
[17] S. Agostinelli *et al.*, Nucl. Instrum. Methods Phys. Res., Sect. A **506**, 250 (2003).
[18] European Strategy Group, Deliberation document on the 2020 update of the European Strategy for Particle Physics, technical report, Geneva, 2020, 10.17181/ESU2020.
[19] H. Cheng *et al.*, arXiv:2205.08553.
[20] M. Ruan, arXiv:1403.4784.
[21] M. Ruan, H. Zhao, G. Li, C. Fu, Z. Wang, X. Lou, D. Yu, V. Boudry, H. Videau, V. Balagura, J.-C. Brient, P. Lai, C.-M. Kuo, B. Liu, F. An, C. Chen, S. Prell, B. Li, and I. Laketineh, Eur. Phys. J. C **78**, 426 (2018).
[22] H. Qu and L. Gouskos, Phys. Rev. D **101**, 056019 (2020).
[23] CEPC Accelerator Study Group, arXiv:2203.09451.
[24] J. Duarte-Campderros, G. Perez, M. Schlaffer, and A. Soffer, Phys. Rev. D **101**, 115005 (2020).
[25] R. L. Workman *et al.* (Particle Data Group), Prog. Theor. Exp. Phys. **2022**, 083C01 (2022).
[26] A. Denner, S. Heinemeyer, I. Puljak, D. Rebuzzi, and M. Spira, Eur. Phys. J. C **71**, 1753 (2011).
[27] F. Herren and M. Steinhauser, Comput. Phys. Commun. **224**, 333 (2018).
[28] J. F. Kamenik, A. Korajac, M. Szewc, M. Tammaro, and J. Zupan, Phys. Rev. D **109**, L011301 (2024).
[29] T. Sjöstrand, S. Mrenna, and P. Skands, J. High Energy Phys. 05 (2006) 026.
[30] P. Mora de Freitas, in *Proceedings of the International Conference on Linear Colliders (LCWS 04)*, https://core.ac.uk/download/pdf/46779218.pdf (2005), pp. 441–444.
[31] C. Fu (CEPC Software Group), CEPC document server, http://cepcdoc.ihep.ac.cn/DocDB/0001/000167/001/cepc-sim.pdf (2017).
[32] T. Suehara and T. Tanabe, Nucl. Instrum. Methods Phys. Res., Sect. A **808**, 109 (2016).
[33] S. Catani, Y. L. Dokshitzer, M. Olsson, G. Turnock, and B. R. Webber, Phys. Lett. B **269**, 432 (1991).
[34] D. Yu, M. Ruan, V. Boudry, H. Videau, J.-C. Brient, Z. Wu, Q. Ouyang, Y. Xu, and X. Chen, Eur. Phys. J. C **80**, 7 (2020).
[35] D. Yu, T. Zheng, and M. Ruan, J. Instrum. **16**, P06013 (2021).
[36] F. An, S. Prell, C. Chen, J. Cochran, X. Lou, and M. Ruan, Eur. Phys. J. C **78**, 464 (2018).
[37] Y. Zhu, S. Chen, H. Cui, and M. Ruan, Nucl. Instrum. Methods Phys. Res., Sect. A **1047**, 167835 (2023).
[38] Y. Che, V. Boudry, H. Videau, M. He, and M. Ruan, Eur. Phys. J. C **83**, 93 (2023); **83**, 470(E) (2023).
[39] T. Zheng, J. Wang, Y. Shen, Y.-K. E. Cheung, and M. Ruan, Eur. Phys. J. Plus **135**, 274 (2020).
[40] F. An *et al.*, Chin. Phys. C **43**, 043002 (2019).
[41] D. M. Asner *et al.*, arXiv:1310.0763.
[42] G. Bernardi *et al.*, arXiv:2203.06520.
[43] M. Cepeda *et al.*, CERN Yellow Rep. Monogr. **7**, 221 (2019).
[44] S. Bejar, F. Dilme, J. Guasch, and J. Sola, J. High Energy Phys. 08 (2004) 018.
[45] A. Albert *et al.*, arXiv:2203.07535.
[46] CMS Collaboration, CERN document server, https://cds.cern.ch/record/2853299 (2023).
[47] D. Barducci and A. J. Helmboldt, J. High Energy Phys. 12 (2017) 105.
[48] X. Mo, G. Li, M.-Q. Ruan, and X.-C. Lou, Chin. Phys. C **40**, 033001 (2016).
[49] Y. Zhu, H. Cui, and M. Ruan, J. High Energy Phys. 11 (2022) 100.
[50] Y. Bai, C. Chen, Y. Fang, G. Li, M. Ruan, J.-Y. Shi, B. Wang, P.-Y. Kong, B.-Y. Lan, and Z.-F. Liu, Chin. Phys. C **44**, 013001 (2020).
[51] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, in *Advances in Neural Information Processing Systems* (2017), pp. 3146–3154.
[52] A. L. Read, CERN Document Server, 10.5170/CERN-2000-005.81 (2000).
[53] A. L. Read, J. Phys. G **28**, 2693 (2002).
[54] J. de Blas *et al.*, J. High Energy Phys. 01 (2020) 139.
[55] J. De Blas, G. Durieux, C. Grojean, J. Gu, and A. Paul, J. High Energy Phys. 12 (2019) 117.
[56] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), J. High Energy Phys. 02 (2014) 057.
[57] M. Ilyushin, P. Mandrik, and S. Slabospitskii, Nucl. Phys. **B952**, 114921 (2020).
[58] Z. Wu, G. Li, D. Yu, C. Fu, Q. Ouyang, and M. Ruan, J. Instrum. **13**, T09002 (2018).
[59] Y. Zhu, H. Liang, Y. Wang, H. Qu, C. Zhou, and M. Ruan, Eur. Phys. J. C **84**, 152 (2024).
[60] M. Bahr *et al.*, Eur. Phys. J. C **58**, 639 (2008).
[61] J. Bellm *et al.*, Eur. Phys. J. C **76**, 196 (2016).
[62] M. Ruan, Advanced Reconstriction: Jet origin id, PFA, https://indico.cern.ch/event/1335278 (2024).