

One-to-one correspondence reconstruction at the electron-positron Higgs factory

Yuexin Wang^{1,2}, Hao Liang^{1,3,4}, Yongfeng Zhu⁵, Yuzhi Che^{1,3},
Xin Xia^{1,3}, Huilin Qu⁶, Chen Zhou⁵, Xuai Zhuang^{1,3},
Manqi Ruan^{1,3*}

^{1*}Institute of High Energy Physics, Chinese Academy of Sciences, 19B
Yuquan Road, Shijingshan District, Beijing, 100049, China.

²China Center of Advanced Science and Technology, Beijing, 100190,
China.

³University of Chinese Academy of Sciences, 19A Yuquan Road,
Shijingshan District, Beijing, 100049, China.

⁴Vanderbilt University Institute of Imaging Science, Vanderbilt
University Medical Center, Nashville, TN 37232, USA.

⁵State Key Laboratory of Nuclear Physics and Technology, School of
Physics, Peking University, Beijing, 100871, China.

⁶EP Department, CERN, CH-1211 Geneva 23, Switzerland.

*Corresponding author(s). E-mail(s): ruanmq@ihep.ac.cn;

Abstract

We propose one-to-one correspondence reconstruction for electron-positron Higgs factories. For each visible particle, one-to-one correspondence aims to associate relevant detector hits with only one reconstructed particle and accurately identify its species. To achieve this goal, we develop a novel detector concept featuring 5-dimensional calorimetry that provides spatial, energy, and time measurements for each hit, and a reconstruction framework that combines state-of-the-art particle flow and artificial intelligence algorithms. In the benchmark process of Higgs to di-jets, over 90% of visible energy can be successfully mapped into well-reconstructed particles that not only maintain a one-to-one correspondence relationship but also associate with the correct combination of cluster and track, improving the invariant mass resolution of hadronically decayed Higgs bosons by 25%. Performing simultaneous identification on these well-reconstructed particles, we observe efficiencies of 97% to nearly 100% for charged particles (e^\pm , μ^\pm , π^\pm , K^\pm , p/\bar{p}) and photons (γ), and 75% to 80% for neutral hadrons (K_L^0 , n , \bar{n}). For physics

measurements of Higgs to invisible and exotic decays, golden channels to probe new physics, one-to-one correspondence could enhance discovery power by 10% to up to a factor of two. This study demonstrates the necessity and feasibility of one-to-one correspondence reconstruction at electron-positron Higgs factories.

In collider experiments, high-energy particles collide and generate multiple final state particles, bringing information about underlying physics laws. These particles interact with the detector, excite detector sensors, and generate electronic signals that are recorded into data, typically in the form of detector hits. Correspondingly, data processing in collider experiments consists of two main steps: interpreting detector hits into physics objects, particularly final state particles, and performing physics measurements using these physics objects. The first step is called reconstruction, which can be regarded as the inverse process of detector hits generation.

The reconstruction establishes a mapping between visible particles and reconstructed ones, where the ultimate goal is a one-to-one (1-1) correspondence. The visible particles include both primary particles from the interaction point (IP) and secondary particles generated through interactions with detector materials upstream of the calorimeter, as far as they create detector hits. 1-1 correspondence reconstruction aims to not only successfully reconstruct each individual particle, but also correctly identify its species (a.k.a. the particle identification, PID). 1-1 correspondence is highly beneficial for collider experiments because it provides a universal foundation for reconstructing various physics objects and can significantly enhance the reconstruction performance for jets and missing energy/momentum.

Depending on the multiplicity of visible particles and the detector configuration, achieving 1-1 correspondence could be truly challenging. For proton-proton and heavy ion collisions at the Large Hadron Collider (LHC) [1] or future hadron colliders, each collision can generate $\mathcal{O}(10^2\text{--}10^4)$ visible particles, making 1-1 correspondence extremely difficult, especially in the detector forward region. In contrast, at electron-positron colliders with center-of-mass energies (\sqrt{s}) around 10 GeV or lower, such as the BelleII [2, 3] and BESIII [4, 5] experiments, the particle multiplicity is typically below 10, and their detectors usually emphasize on the PID performance, achieving 1-1 correspondence could be much easier. Our study focuses on the electron-positron Higgs factory [6–10], which is regarded as the highest-priority future collider as it offers excellent opportunity to discover new physics [11, 12], especially via precision measurements of the Higgs boson. It operates at center-of-mass energies ranging from 91.2 GeV to several TeV, typically producing $\mathcal{O}(100)$ visible particles distributed in narrow jets in hadronic events.

The concept of 1-1 correspondence evolves from the Particle Flow Algorithm (PFA) [13–18], with the key idea to trace each individual particle, classify and associate corresponding detector hits, and measure the particle 4-momentum using the most suitable sub-detector system. Originating from the ALEPH [16] experiment, the PFA has become the guiding principle for multiple conceptual detector designs at electron-positron Higgs factories [9, 10, 19] and detector upgrade projects at HL-LHC [20].

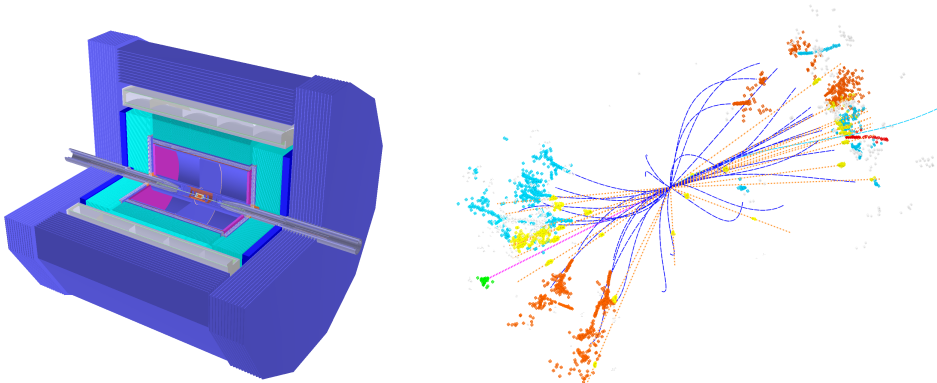


Fig. 1 AURORA detector geometry and an event display of $\nu\bar{\nu}H, H \rightarrow gg$ ($\sqrt{s} = 240$ GeV) event.

Benefiting from the rapid development of artificial intelligence, many machine learning algorithms have been implemented to enhance PFA performance [21–23]. The current PFA performance at electron-positron Higgs factories is primarily limited by confusion [14, 24], which includes the effects of fake particles, failures in track-cluster matching, particle loss due to shower overlap, etc. These confusion effects violate the 1-1 correspondence relationship. Therefore, the goal of 1-1 correspondence reconstruction is to approach confusion-free particle flow reconstruction and to identify the species of all visible particles.

The performance of PFA can be quantified using the Boson Mass Resolution (BMR) [9, 25], which represents the relative mass resolution of massive bosons (such as Higgs, Z , and W bosons) decaying into hadronic final states. At the electron-positron Higgs factory, a BMR smaller than 4% is required to separate the $q\bar{q}H$ signal from $q\bar{q}Z$ background using the recoil mass method [9, 26]. A smaller BMR is beneficial for all physics measurements with hadronic final states, especially for flavor physics measurements and new physics searches since they rely on accurate measurements of missing energy and momentum. Referring to the Circular Electron-Positron Collider (CEPC) Conceptual Design Report (CDR) [9], its baseline detector design achieves a BMR of 3.7% [25, 27], meeting the requirements for Higgs measurements. In terms of PID, the CEPC baseline detector demonstrates a typical efficiency of 99.5% and a misidentification rate around 1% for isolated lepton identification [28, 29]. It can also distinguish charged hadrons ($\pi^\pm, K^\pm, p/\bar{p}$) using techniques such as dE/dx and time-of-flight (TOF) measurements [30, 31].

To pursue 1-1 correspondence reconstruction, we develop the AURORA (Apparatus for RecOnstruction with Advanced algorithm) detector concept and the PROOF (Particle Reconstruction with One-to-One correspondence at Higgs Factory) reconstruction framework, based on the CEPC CDR baseline design [9]. AURORA features a high-granularity 5-dimensional (5D) calorimetry as its core, with a unified time resolution of 100 ps for each calorimeter hit. The left panel of Fig. 1 shows a cut-away view of AURORA, with more details provided in Methods. Combining the CEPC baseline PFA Arbor [14] and Particle Transformer (ParT) [32], PROOF performs a universal identification of all reconstructed particles, simultaneously distinguishing different

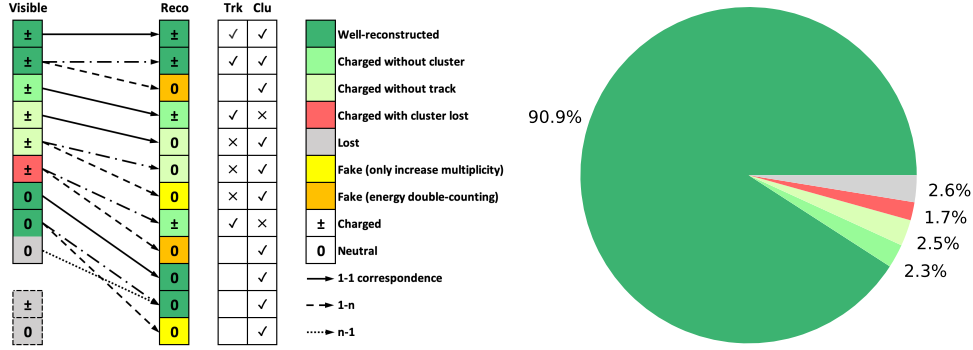


Fig. 2 Left: Schematic diagram of visible-reconstructed particle mapping. “Charged without cluster/track” means a charged particle does not create enough hits to form a cluster/track (mainly due to limited detector acceptance and efficiency), while “Charged with cluster lost” refers to cases where both a cluster and track are formed but do not match each other, resulting in fake particles. Right: Fraction of truth visible energy mapped into different categories.

confusion types and identifying particle species. ParT is a particle cloud- [33] and Transformer-based [34] machine learning model adapted for high-energy physics experiments, which has been applied in jet flavor tagging at the LHC [35–37] and jet origin identification (JOI) at the electron-positron Higgs factory [38]. Following the convention of CEPC performance studies [9, 27, 39], we simulate 1 million $\nu\bar{\nu}H, H \rightarrow gg$ ($\sqrt{s} = 240$ GeV) events with AURORA detector and reconstruct these events with Arbor. A reconstructed $\nu\bar{\nu}H, H \rightarrow gg$ event is displayed in the right panel of Fig. 1, and more details on simulation and reconstruction are described in Methods.

Using truth-level information in the Monte Carlo (MC) simulation, we establish a mapping between visible particles and reconstructed ones, as illustrated in the left panel of Fig. 2. The mapping consists of three main categories:

- 1-1 correspondence (solid arrows in Fig. 2): one visible particle is mapped to only one reconstructed particle. This category consists of three types, represented in shades of green in Fig. 2. One type is **well-reconstructed**, where a cluster is a must for neutral particles, and both track and cluster are required for charged ones. The other two types correspond to charged particles that generate only track or cluster, mainly caused by the limitations of detector acceptance (e.g. in the forward region) and efficiency (e.g. dead zone or energy/momentum measurement thresholds).
- Fake particles (in orange and yellow in Fig. 2): occur when one visible particle is mapped to multiple reconstructed particles. Due to detector inefficiency and particle interactions with upstream materials, the calorimeter hits generated by a single particle may be grouped into multiple clusters. For a charged particle, its track could be matched to only part of or even none of these clusters, while the remaining clusters are reconstructed as nearby extra neutral particles. These neutral particles that originate from charged particle shower fragments and un-associated clusters, represented in orange in Fig. 2, can cause double-counting in the total reconstructed energy. For simplicity, these neutral particles are called fake particles in this study.

- Lost particles (in gray in Fig. 2): visible particles with no corresponding reconstructed particles. Particles can be lost due to limited detector acceptance¹, selection criteria, and shower merging. Selection criteria are applied to improve the reconstruction performance, such as vetoing fake particles, but inevitably they could exclude genuine particles, usually low-energy ones. Shower merging (dotted arrow in Fig. 2) is another major cause of particle loss, particularly within high-energy jets where nearby or overlapping calorimeter showers can be merged into one cluster, especially at limited calorimeter granularity. For instance, the reconstruction of $\pi^0 \rightarrow \gamma\gamma$ can be particularly sensitive to shower merging.

According to this mapping with truth information, we label reconstructed particles into 15 types, including 10 types of visible particles (e^\pm , μ^\pm , π^\pm , K^\pm , p/\bar{p} , γ , K_L^0 , n , \bar{n} , and others including K_S^0 and Λ that not yet decay within the tracker volume) and 5 types of confusion (such as different kinds of fake particles, charged particle without track, etc). Each reconstructed particle is characterized by 55 observable variables as inputs to ParT. More details can be found in Methods.

The quantitative analysis [40] shows that fake particles are the leading contribution to the BMR. In our simulation, on average, a $\nu\bar{\nu}H, H \rightarrow gg$ event has 80 visible particles with a total energy of 135 GeV at the truth level, while the reconstructed fake particles lead to a double-counted energy of 6 GeV, severely degrading the BMR. PROOF can identify the fake particles with a typical efficiency of 77% and a purity of 97.5%, consequently suppressing the average double-counted energy from fake particles to 0.7 GeV (by nearly one order of magnitude) at the minor cost of mis-excluding visible particles with a total energy on the order of 0.5 GeV. Consequently, PROOF achieves a BMR of 2.75% (see the left panel of Fig. 3), where the effects of fake and mis-excluded particles get balanced. Compared to the CEPC CDR [41], the BMR is improved by 25%, where 10% is from the detector geometry optimization to AURORA [42], and 15% is from the fake particle suppression. The contributions from the three categories of visible-reconstructed particle mapping to the BMR are not proportional to their corresponding energy fractions. The contribution from the 1-1 correspondence category primarily stems from the detector resolution, which is orders of magnitude lower compared to the total energy of particles within this category. Meanwhile, the contributions from the fake and lost categories are comparable to their corresponding total energies. We found that the 1-1 correspondence category contributes to half of the BMR, while the fake and lost categories each contribute a quarter, combined quadratically.

Referring to the truth level energy of visible particles, we calculate the fraction of total visible energy corresponding to different categories of the mapping. Ignoring the remaining fake particles², we observe that over 90% of the total visible energy is

¹In this article, we define visible particles as those that create detector hits. On the other hand, the limited detector acceptance causes particle loss, i.e., AURORA has an overall geometry acceptance of $|\cos\theta| \approx 0.99$, where it cannot record particles traveling in the forward direction. Since particle loss due to detector acceptance certainly impacts the overall physics measurement, we list these particles in dashed boxes in Fig. 2 to complete the picture and also consider them when calculating visible energy fractions in the following text.

²The remaining fake particles could violate the 1-1 correspondence relationship, as one visible particle could be mapped to one leading reconstructed particle together with one or even several fake particles. However, as the energy of the fake particle is usually much lower than the leading one, the fake particle only has a tiny impact on the reconstruction and identification of the original visible particle. Meanwhile, the

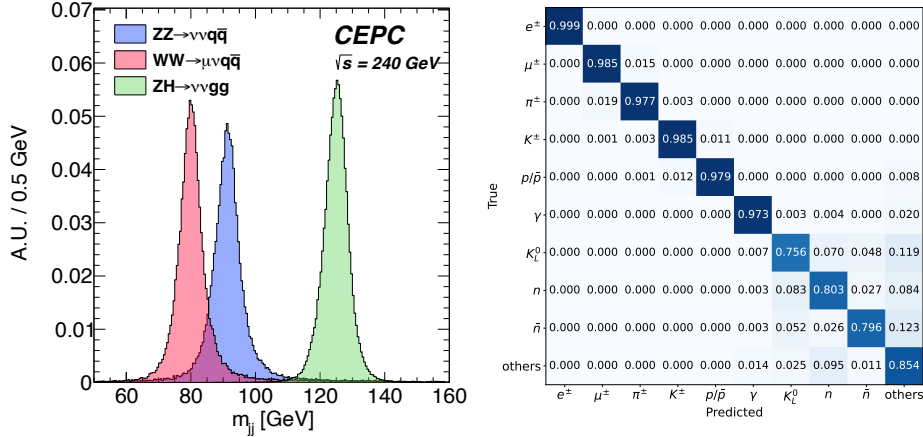


Fig. 3 Left: Invariant mass distributions of hadronically decayed Higgs, W , and Z bosons derived by the combination of AURORA and PROOF to approach 1-1 correspondence reconstruction. Right: Confusion matrix of **well-reconstructed** particles identification. Both are after removing identified fake particles.

mapped to the **well-reconstructed** particles, 4% to charged particles without clusters (with over half of this due to calorimeter acceptance/inefficiency), 2.5% to charged particles without tracks, and 2.6% of the energy is lost due to limited detector acceptance, shower merging, and mis-excluded genuine particles, as illustrated in the pie chart in Fig. 2. The detector acceptance dominates the confusion of charged particles without clusters or tracks. Summing the energy fraction of **well-reconstructed** particles and confusion type purely originating from detector acceptance, we consider that nearly 95% of the total visible energy maintains the 1-1 correspondence relationship.

Focusing on the well-reconstructed particles of 1-1 correspondence type, the PID performance of PROOF is shown in the confusion matrix in the right panel of Fig. 3. The identification efficiencies (diagonal matrix elements) range from 97% to nearly 100% for charged particles and photons, and from 75% to 80% for three types of neutral hadrons. Since the PID performance depends on particle kinematics, we extract the traces (i.e. sums of diagonal elements) of sub-matrices in different particle momentum and polar angle ranges, as shown in Fig. 4. For photons and charged particles, a high-trace plateau is observed within the momentum range of 0.8 to 20 GeV and $|\cos\theta| < 0.96$. Inefficiencies primarily arise due to limited cluster-level information in the low-momentum region and degraded PID performance in high-momentum and detector-forward regions. Similar behavior is observed for neutral hadrons, which shows that the trace is approaching 3 in a limited phase space, indicating that even neutral hadrons can be efficiently separated. These observations lead to the conclusion that 1-1 correspondence is feasible at future electron-positron Higgs factories.

It should be noted that many simplifications have been made to facilitate this analysis. To fully realize 1-1 correspondence reconstruction and maximize its impact

total energy of remaining fake particles only contributes to a sub-percentage level of total visible energy. Therefore, we ignore these remaining fake particles in our study.

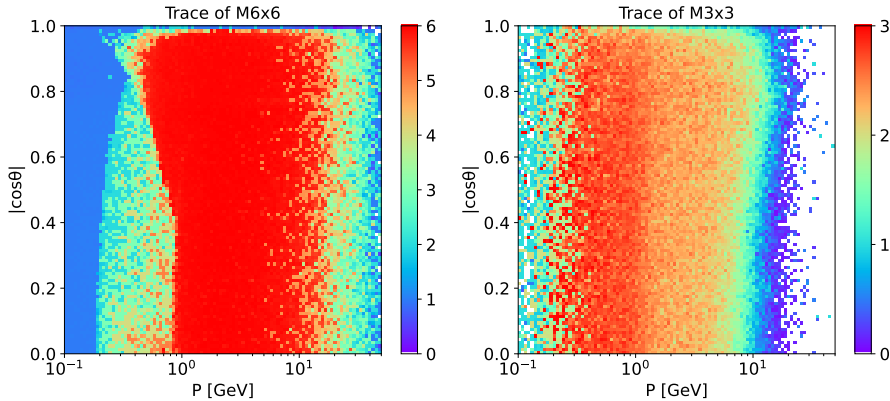


Fig. 4 Traces of 6×6 ($e^\pm, \mu^\pm, \pi^\pm, K^\pm, p/\bar{p}, \gamma$) (left) and 3×3 (K_L^0, n, \bar{n}) (right) sub-matrices as function of momentum and polar angle $|\cos\theta|$.

on physics exploration, dedicated studies are required in the future. Secondary particles can significantly affect reconstruction and physics measurements. For example, electron-positron pairs from photon conversions and muons from hadron decays could contribute to the background in measurements that rely on (semi-)leptonic decays. Secondary nucleons, generated in the interactions between primary particles and upstream materials, as well as backscattering, could impact the BMR. The reconstruction of the primary vertex is crucial for TOF measurements and affects PID performance. The impact of beam-induced background needs to be analyzed and presumably ameliorated. Event building at high event rates also needs to be developed. Detector optimization should be systematically performed, including the quantification of performance specification (particularly the TOF resolution of the calorimeter and dE/dx performance of the gaseous tracker), the survey of available and emerging technologies, and the geometry optimization. Modeling and verifying detailed detector responses are crucial for controlling systematic uncertainties, which is vital for precision measurements.

Prospects

Our study demonstrates the feasibility of 1-1 correspondence reconstruction at the electron-positron Higgs factory, it can significantly improve the BMR performance, and simultaneously identify 10 different kinds of visible particles. Meanwhile, the performance bottleneck of 1-1 correspondence in our current setup is identified to be the detector acceptance and neutral hadron identification, which could be ameliorated via dedicated algorithms and detector optimization. This section discusses the prospects for high-level reconstruction and physics measurements with an ideal 1-1 correspondence reconstruction.

Since 1-1 correspondence reconstruction provides the particle mass information via PID, the energy of visible particles can be determined by the TOF. 5D calorimetry

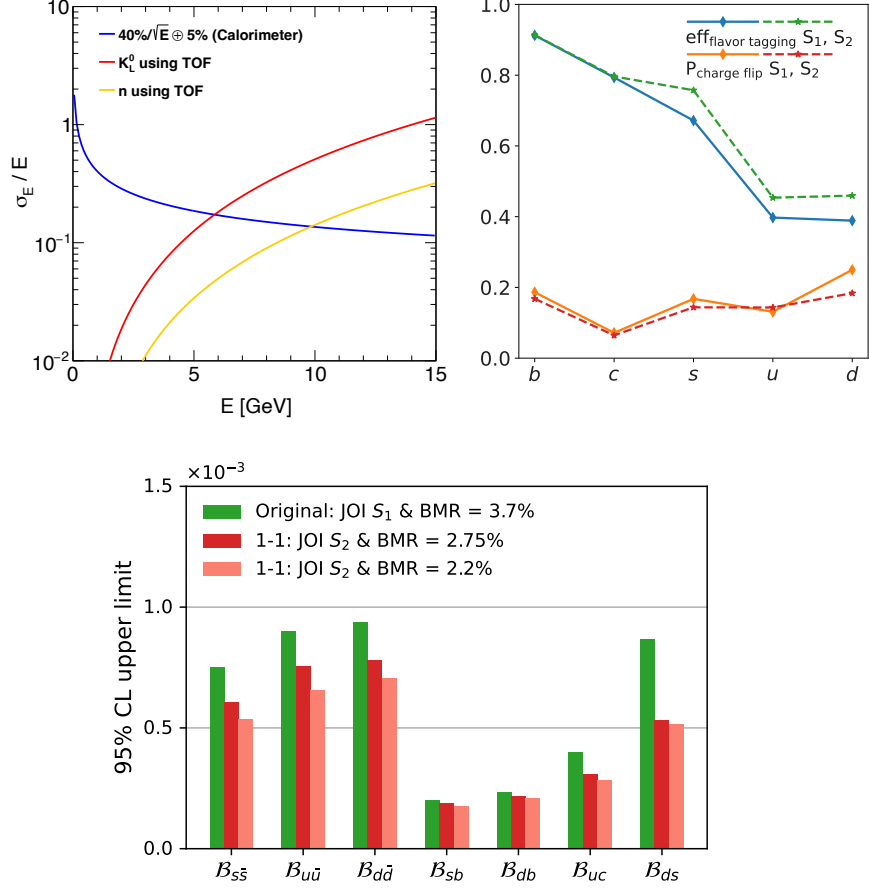


Fig. 5 Top left: Neutral hadron energy resolution using calorimeter with direct energy sum of each hit [42, 43] (blue line, derived by the full detector simulation) or TOF measurement with time resolution assumed to be 10 ps at particle level (red and yellow lines). Top right: Jet origin identification performance (characterized by flavor tagging efficiencies and charge flip rates) in two scenarios: Scenario 1 (S₁) assumes perfect identification of leptons and charged hadrons, while Scenario 2 (S₂) extends S₁ by further assuming perfect identification of K_L^0/K_S^0 , n/\bar{n} , and $\Lambda/\bar{\Lambda}$. Bottom: Anticipated upper limits on branching ratios of Higgs exotic and rare decays with different performances of JOI and BMR.

provides TOF measurements for both charged and neutral particles. For charged particles, their energies/momenta can be determined by three methods: shower energy via calorimeter, track momentum via tracker, and TOF via calorimeter, while the TOF method has a comparative advantage in the forward region. For neutral particles, their energies can be determined by both the shower energy and the TOF via calorimeter, with TOF showing a significant advantage for low-energy neutral hadrons, as demonstrated in the top left panel of Fig. 5. Therefore, by combining these measurements of particle energy, such as selecting the most suitable method according to the particle

species and their kinematics, we can significantly improve the energy/momentum resolution of particles, and consequently improve the BMR. Considering the improvement in particle energy measurement can also benefit pattern recognition and confusion control, we estimate that this approach could further enhance the BMR by 15–20%, meaning that a BMR of 2.2–2.4% could, in principle, be achieved.

1-1 correspondence reconstruction can also significantly enhance the JOI performance. JOI is the procedure to determine the type of quark/gluon (11 types are typically considered: $b, \bar{b}, c, \bar{c}, s, \bar{s}, u, \bar{u}, d, \bar{d}, g$) from which a jet originates. A recent study realizes JOI with the CEPC CDR baseline detector [38]. It can simultaneously identify b, c, s quarks with efficiencies of 70–90% and u, d quarks with efficiencies of 40%, meanwhile, the charge flip rates of quarks and anti-quarks are controlled to be 10–20%, with an ideal lepton and charged hadron identification. JOI significantly boosts the accuracy of flavor-sensitive physics measurements, such as $H \rightarrow s\bar{s}$ and $H \rightarrow sb$ decays [38], CKM matrix element $|V_{cb}|$ [44], CP -violating phase ϕ_s [45], and weak mixing angle [46] measurements. 1-1 correspondence provides the identification of K_L^0, n , and \bar{n} , while in principle the K_S^0 and Λ could also be identified [47], therefore, providing the PID information of all these neutral hadrons, we observe a significant improvement on JOI performance especially the identification performance of $u/d/s$ quarks, as shown in the top right panel of Fig. 5.

BMR and JOI are key performances of event reconstruction at electron-positron Higgs factories. Their improvements will benefit most physics measurements with hadronic final states. For example, the Higgs to invisible final states is a key portal to detect dark matter at Higgs factories. The original upper limit at 95% confidence level is quantified to be $\mathcal{O}(0.1\%)$ [48] with a BMR of 3.7%. Improving the BMR to 2.75%, or even 2.2%, will consequently enhance the upper limit by $\mathcal{O}(10\%)$ [49]. Higgs exotic and rare hadronic decays are also sensitive probes to new physics. Using JOI technology, these decay modes can be typically limited to 10^{-4} to 10^{-3} using $\nu\bar{\nu}H$ and $\ell\ell H$ processes [38]. With the improvement in JOI alone, the constraints could be enhanced by 10% to up to a factor of two. Considering also the improved BMR, the sensitivities are expected to be further enhanced by another $\mathcal{O}(10\%)$, see the bottom panel of Fig. 5. 1-1 correspondence also expands the methodology for detector monitoring: by reconstructing almost all visible particles, it provides an in-situ, holistic description of the detector response, which is critical for controlling systematic uncertainties in high-precision physics measurements. In short, 1-1 correspondence could significantly boost the discovery power of electron-positron Higgs factories.

Based on this study, we are confident that 1-1 correspondence reconstruction will be a new paradigm for event reconstruction at the high-energy frontier. For the electron-positron Higgs factory, 1-1 correspondence reconstruction is a goal that should be pursued, as it can significantly enhance the physics reach and discovery power. It is also a goal that could be achieved—presumably very soon—through integrating innovative detector technologies, advanced reconstruction algorithms, and artificial intelligence.

References

- [1] Evans, L., Bryant, P. (eds.): LHC Machine. JINST **3**, 08001 (2008) <https://doi.org/10.1088/1748-0221/3/08/08001>

[org/10.1088/1748-0221/3/08/S08001](https://doi.org/10.1088/1748-0221/3/08/S08001)

- [2] Altmannshofer, W., *et al.*: The Belle II Physics Book. PTEP **2019**(12), 123–01 (2019) <https://doi.org/10.1093/ptep/ptz106> arXiv:1808.10567 [hep-ex]. [Erratum: PTEP 2020, 029201 (2020)]
- [3] Abe, T., *et al.*: Belle II Technical Design Report (2010) arXiv:1011.0352 [physics.ins-det]
- [4] Yuan, C.-Z., Olsen, S.L.: The BESIII physics programme. Nature Rev. Phys. **1**(8), 480–494 (2019) <https://doi.org/10.1038/s42254-019-0082-y> arXiv:2001.01164 [hep-ex]
- [5] Ablikim, M., *et al.*: Design and Construction of the BESIII Detector. Nucl. Instrum. Meth. A **614**, 345–399 (2010) <https://doi.org/10.1016/j.nima.2009.12.050> arXiv:0911.4960 [physics.ins-det]
- [6] Behnke, T., Brau, J.E., Foster, B., Fuster, J., Harrison, M., Paterson, J.M., Peskin, M., Stanitzki, M., Walker, N., Yamamoto, H. (eds.): The International Linear Collider Technical Design Report - Volume 1: Executive Summary (2013) arXiv:1306.6327 [physics.acc-ph]
- [7] Linssen, L., Miyamoto, A., Stanitzki, M., Weerts, H. (eds.): Physics and Detectors at CLIC: CLIC Conceptual Design Report (2012) <https://doi.org/10.5170/CERN-2012-003> arXiv:1202.5940 [physics.ins-det]
- [8] Abdallah, W., *et al.*: CEPC Technical Design Report: Accelerator. Radiat. Detect. Technol. Methods **8**(1), 1–1105 (2024) <https://doi.org/10.1007/s41605-024-00463-y> arXiv:2312.14363 [physics.acc-ph]
- [9] Dong, M., *et al.*: CEPC Conceptual Design Report: Volume 2 - Physics & Detector (2018) arXiv:1811.10545 [hep-ex]
- [10] Abada, A., *et al.*: FCC-ee: The Lepton Collider: Future Circular Collider Conceptual Design Report Volume 2. Eur. Phys. J. ST **228**(2), 261–623 (2019) <https://doi.org/10.1140/epjst/e2019-900045-4>
- [11] Group, E.S.: Deliberation Document on the 2020 Update of the European Strategy for Particle Physics. <https://cds.cern.ch/record/2735179> (2020)
- [12] P5, U.: Report of the Particle Physics Project Prioritization Panel (P5). https://www.usparticlephysics.org/wp-content/uploads/2018/03/FINAL-P5_Report_053014.pdf (2014)
- [13] Marshall, J.S., Thomson, M.A.: Pandora Particle Flow Algorithm. In: International Conference on Calorimetry for the High Energy Frontier, pp. 305–315 (2013)

- [14] Ruan, M., Videau, H.: Arbor, a new approach of the Particle Flow Algorithm. In: International Conference on Calorimetry for the High Energy Frontier, pp. 316–324 (2013)
- [15] Sirunyan, A.M., *et al.*: Particle-flow reconstruction and global event description with the CMS detector. JINST **12**(10), 10003 (2017) <https://doi.org/10.1088/1748-0221/12/10/P10003> [arXiv:1706.04965](https://arxiv.org/abs/1706.04965) [physics.ins-det]
- [16] Buskulic, D., *et al.*: Performance of the ALEPH detector at LEP. Nucl. Instrum. Meth. A **360**, 481–506 (1995) [https://doi.org/10.1016/0168-9002\(95\)00138-7](https://doi.org/10.1016/0168-9002(95)00138-7)
- [17] Brient, J.C.: Improving the jet reconstruction with the particle flow method: An introduction. In: 11th International Conference on Calorimetry in High-Energy Physics (Calor 2004), pp. 445–451 (2004)
- [18] Videau, H., Brient, J.C.: Calorimetry optimised for jets. In: 10th International Conference on Calorimetry in High Energy Physics (CALOR 2002), pp. 747–760 (2002)
- [19] Abramowicz, H., *et al.*: The International Linear Collider Technical Design Report - Volume 4: Detectors (2013) [arXiv:1306.6329](https://arxiv.org/abs/1306.6329) [physics.ins-det]
- [20] Contardo, D., Klute, M., Mans, J., Silvestris, L., Butler, J.N. (eds.): Technical Proposal for the Phase-II Upgrade of the CMS Detector (2015) <https://doi.org/10.17181/CERN.VU8I.D59J>
- [21] Kakati, N., Dreyer, E., Ivina, A., Di Bello, F.A., Heinrich, L., Kado, M., Gross, E.: HGPflow: Extending Hypergraph Particle Flow to Collider Event Reconstruction (2024) [arXiv:2410.23236](https://arxiv.org/abs/2410.23236) [hep-ex]. <https://arxiv.org/abs/2410.23236>
- [22] Pata, J., Wulff, E., Mokhtar, F., Southwick, D., Zhang, M., Girone, M., Duarte, J.: Improved particle-flow event reconstruction with scalable neural networks for current and future particle detectors. Commun. Phys. **7**(1), 124 (2024) <https://doi.org/10.1038/s42005-024-01599-5> [arXiv:2309.06782](https://arxiv.org/abs/2309.06782) [physics.data-an]
- [23] Di Bello, F.A., *et al.*: Reconstructing particles in jets using set transformer and hypergraph prediction networks. Eur. Phys. J. C **83**(7), 596 (2023) <https://doi.org/10.1140/epjc/s10052-023-11677-7> [arXiv:2212.01328](https://arxiv.org/abs/2212.01328) [hep-ex]
- [24] Thomson, M.A.: Particle Flow Calorimetry and the PandoraPFA Algorithm. Nucl. Instrum. Meth. A **611**, 25–40 (2009) <https://doi.org/10.1016/j.nima.2009.09.009> [arXiv:0907.3577](https://arxiv.org/abs/0907.3577) [physics.ins-det]
- [25] Ruan, M., *et al.*: Reconstruction of physics objects at the Circular Electron Positron Collider with Arbor. Eur. Phys. J. C **78**(5), 426 (2018) <https://doi.org/10.1140/epjc/s10052-018-5876-z> [arXiv:1806.04879](https://arxiv.org/abs/1806.04879) [hep-ex]

- [26] Yu, D., Ruan, M., Boudry, V., Videau, H., Brient, J.-C., Wu, Z., Ouyang, Q., Xu, Y., Chen, X.: The measurement of the $H \rightarrow \tau\tau$ signal strength in the future e^+e^- Higgs factories. *Eur. Phys. J. C* **80**(1), 7 (2020) <https://doi.org/10.1140/epjc/s10052-019-7557-y>
- [27] Zhao, H., Zhu, Y.-F., Fu, C.-D., Yu, D., Ruan, M.-Q.: The Higgs Signatures at the CEPC CDR Baseline. *Chin. Phys. C* **43**(2), 023001 (2019) <https://doi.org/10.1088/1674-1137/43/2/023001> [arXiv:1806.04992](https://arxiv.org/abs/1806.04992) [hep-ex]
- [28] Yu, D., Ruan, M., Boudry, V., Videau, H.: Lepton identification at particle flow oriented detector for the future e^+e^- Higgs factories. *Eur. Phys. J. C* **77**(9), 591 (2017) <https://doi.org/10.1140/epjc/s10052-017-5146-5> [arXiv:1701.07542](https://arxiv.org/abs/1701.07542) [physics.ins-det]
- [29] Yu, D., Zheng, T., Ruan, M.: Lepton identification performance in Jets at a future electron positron Higgs Z factory (2021) <https://doi.org/10.1088/1748-0221/16/06/P06013> [arXiv:2105.01246](https://arxiv.org/abs/2105.01246) [hep-ex]
- [30] An, F., Prell, S., Chen, C., Cochran, J., Lou, X., Ruan, M.: Monte Carlo study of particle identification at the CEPC using TPC dE/dx information. *Eur. Phys. J. C* **78**(6), 464 (2018) <https://doi.org/10.1140/epjc/s10052-018-5803-3> [arXiv:1803.05134](https://arxiv.org/abs/1803.05134) [physics.ins-det]
- [31] Zhu, Y., Chen, S., Cui, H., Ruan, M.: Requirement analysis for dE/dx measurement and PID performance at the CEPC baseline detector. *Nucl. Instrum. Meth. A* **1047**, 167835 (2023) <https://doi.org/10.1016/j.nima.2022.167835> [arXiv:2209.14486](https://arxiv.org/abs/2209.14486) [physics.ins-det]
- [32] Qu, H., Li, C., Qian, S.: Particle Transformer for Jet Tagging (2022) [arXiv:2202.03772](https://arxiv.org/abs/2202.03772) [hep-ph]
- [33] Qu, H., Gouskos, L.: Jet Tagging via Particle Clouds. *Phys. Rev. D* **101**(5), 056019 (2020) <https://doi.org/10.1103/PhysRevD.101.056019> [arXiv:1902.08570](https://arxiv.org/abs/1902.08570) [hep-ph]
- [34] Vaswani, A.: Attention is all you need. *Advances in Neural Information Processing Systems* (2017)
- [35] A unified approach for jet tagging in Run 3 at $\sqrt{s}=13.6$ TeV in CMS (CMS-DP-2024-066) (2024)
- [36] Hayrapetyan, A., et al.: Measurement of inclusive and differential cross sections of single top quark production in association with a W boson in proton-proton collisions at $\sqrt{s} = 13.6$ TeV (2024) [arXiv:2409.06444](https://arxiv.org/abs/2409.06444) [hep-ex]
- [37] Brigljevic, V., *et al.*: HHH Whitepaper. In: HHH Workshop (2024)

- [38] Liang, H., Zhu, Y., Wang, Y., Che, Y., Zhou, C., Qu, H., Ruan, M.: Jet-
Origin Identification and Its Application at an Electron-Positron Higgs Factory.
Phys. Rev. Lett. **132**(22), 221802 (2024) [https://doi.org/10.1103/PhysRevLett.
132.221802](https://doi.org/10.1103/PhysRevLett.132.221802) arXiv:2310.03440 [hep-ex]
- [39] Hu, P., Wang, Y., Du, D., Hua, Z., Qian, S., Fu, C., Liu, Y., Ruan, M., Wang,
J., Wang, Y.: GSHCAL at future e+e- Higgs factories. Nucl. Instrum. Meth. A
1059, 168944 (2024) <https://doi.org/10.1016/j.nima.2023.168944>
- [40] Ruan, M.: Simulation and Performance of the CEPC: at CDR and recent update.
[https://indico.cern.ch/event/820586/contributions/3553194/attachments/
1909878/3155585/1_Simulation_and_Performance-Chicago.pdf](https://indico.cern.ch/event/820586/contributions/3553194/attachments/1909878/3155585/1_Simulation_and_Performance-Chicago.pdf) (2019)
- [41] Lai, P.-Z., Ruan, M., Kuo, C.-M.: Jet performance at the circular
electron-positron collider. JINST **16**(07), 07037 (2021) [https://doi.org/10.1088/
1748-0221/16/07/P07037](https://doi.org/10.1088/1748-0221/16/07/P07037) arXiv:2104.05029 [hep-ex]
- [42] Hu, P., Wang, Y., Du, D., Hua, Z., Qian, S., Fu, C., Liu, Y., Ruan, M., Wang,
J., Wang, Y.: GSHCAL at future e+e- Higgs factories. Nucl. Instrum. Meth. A
1059, 168944 (2024) <https://doi.org/10.1016/j.nima.2023.168944>
- [43] Li, H.: Progress of the CEPC GS-HCAL. [https://indico.ihep.ac.cn/event/
22941/contributions/171374/attachments/84638/108004/CLHCP2024_CEPC_
GSHCAL.pdf](https://indico.ihep.ac.cn/event/22941/contributions/171374/attachments/84638/108004/CLHCP2024_CEPC_GSHCAL.pdf) (2024)
- [44] Liang, H., Li, L., Zhu, Y., Shen, X., Ruan, M.: Measurement of CKM element
 $|V_{cb}|$ from W boson decays at the future Higgs factories (2024) arXiv:2406.01675
[hep-ph]
- [45] Li, X., Ruan, M., Zhao, M.: Prospect for measurement of the CP-violating
phase ϕ_s in the $B_s \rightarrow J/\psi\phi$ channel at a future Z factory. Eur. Phys. J. C **84**(8),
859 (2024) <https://doi.org/10.1140/epjc/s10052-024-13217-3> arXiv:2205.10565
[hep-ex]
- [46] Zhao, Z., Yang, S., Ruan, M., Liu, M., Han, L.: Measurement of the effective
weak mixing angle at the CEPC*. Chin. Phys. C **47**(12), 123002 (2023) <https://doi.org/10.1088/1674-1137/acf91f> arXiv:2204.09921 [hep-ex]
- [47] Zheng, T., Wang, J., Shen, Y., Cheung, Y.-K.E., Ruan, M.: Reconstructing K_S^0
and Λ in the CEPC baseline detector. Eur. Phys. J. Plus **135**(3), 274 (2020)
<https://doi.org/10.1140/epjp/s13360-020-00272-4>
- [48] Cheng, H., *et al.*: The Physics potential of the CEPC. Prepared for the US
Snowmass Community Planning Exercise (Snowmass 2021). In: Snowmass 2021
(2022)
- [49] Ruan, M.: Advanced reconstruction: PFA & Jet origin identification.

<https://indico.cern.ch/event/1335278/contributions/5733469/attachments/2785402/4856212/Advanced%20Reco%20-%20PFA%20-%20Jet%20origin%20-%20POST.pdf> (2024)

- [50] Liang, H., Zhu, Y., Lai, P.-Z., Ruan, M.: Optimization of tracker configuration for the CEPC. *JINST* **17**(11), 11016 (2022) <https://doi.org/10.1088/1748-0221/17/11/P11016> [arXiv:2209.00397](https://arxiv.org/abs/2209.00397) [physics.ins-det]
- [51] Kilian, W., Ohl, T., Reuter, J.: WHIZARD: Simulating Multi-Particle Processes at LHC and ILC. *Eur. Phys. J. C* **71**, 1742 (2011) <https://doi.org/10.1140/epjc/s10052-011-1742-y> [arXiv:0708.4233](https://arxiv.org/abs/0708.4233) [hep-ph]
- [52] Sjostrand, T., Mrenna, S., Skands, P.Z.: PYTHIA 6.4 Physics and Manual. *JHEP* **05**, 026 (2006) <https://doi.org/10.1088/1126-6708/2006/05/026> [arXiv:hep-ph/0603175](https://arxiv.org/abs/hep-ph/0603175)
- [53] Freitas, P.: Mokka, main guidelines and future. In: International Conference on Linear Colliders (LCWS 04), pp. 441–444 (2004). <https://core.ac.uk/download/pdf/46779218.pdf>
- [54] Fu, C.: Full Simulation Software at CEPC. CEPC Document Server (2017)
- [55] Ruan, M., Jeans, D., Boudry, V., Brient, J.-C., Videau, H.: Fractal Dimension of Particle Showers Measured in a Highly Granular Calorimeter. *Phys. Rev. Lett.* **112**(1), 012001 (2014) <https://doi.org/10.1103/PhysRevLett.112.012001> [arXiv:1312.7662](https://arxiv.org/abs/1312.7662) [physics.ins-det]

Methods

AURORA detector

AURORA, the conceptual detector used in this study, is a PFA-oriented detector design developed from the CEPC CDR baseline [9, 50]. It has an outer radius of 5.2 m, a length of 10.5 m, and a total weight of approximately 5000 tons, with over 65% of the weight contributed by the Yoke and nearly 25% by the HCAL. Fig. 6 displays two side views of AURORA, and specific geometry parameters are summarized in Table 1. In general, AURORA features a high-granularity calorimetry with both ECAL and HCAL, a high-precision tracker system with a low material budget, a high-precision vertex detector, and a large solenoid that encloses ECAL and HCAL. Compared to the CDR baseline, the major change is the HCAL, where we replace the Glass Resistive Plate Chamber (RPC) digital HCAL [9] with the Glass-Scintillator HCAL [39] and increase the HCAL thickness from 5 to 6 nuclear interaction lengths (λ) to reduce the longitudinal leakage. In addition, for both ECAL and HCAL, we assume each calorimeter cell can provide a time resolution of 100 ps, which is used for the PID and PFA confusion identification.

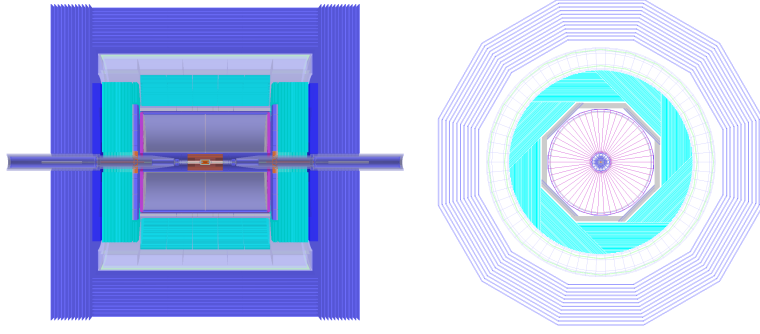


Fig. 6 The r - z (left) and r - ϕ (right) views of the AURORA detector geometry.

Table 1 AURORA detector geometry parameters.

Sub-detector	Thickness (mm)	Inner radius (mm)	Outer radius (mm)	Length (mm)	Volume (m ³)	Transverse cell size	#Layers	#Channels
Vertex	-	-	16–60	125–250	-	$25 \times 25 \mu\text{m}^2$	6	5.3×10^8
Si-strip Tracker	-	-	155 300 1810	736 1288 4600	-	$20 \mu\text{m} \times 2 \text{cm}$	3	3.0×10^7
TPC	-	300	1800	4700	47	$1 \times 6 \text{mm}^2$	220	2.9×10^6
ECAL	173	1845	2018	5250	15	$1 \times 1 \text{cm}^2$	30	2.5×10^7
HCAL	1145	2072	3250	7590	180	$2 \times 2 \text{cm}^2$	48	1.8×10^7
Solenoid	700	3275	3975	7750	120	-	-	-
Yoke	1200	4000	5200	10500	470	-	-	-

Monte Carlo simulation and reconstruction

A total of 1 million $\nu\bar{\nu}H, H \rightarrow gg$ events are generated and simulated using the Monte Carlo method. The simulation is conducted within the CEPC software framework [9], utilizing Whizard1.95 [51] and Pythia6.4 [52] for event generation, MokkaPlus [53, 54] for detector simulation, and Arbor [14, 25] for particle flow reconstruction. Using tracks and calorimeter hits as inputs, Arbor groups calorimeter hits into clusters based on the shower’s tree-like topology. With the high granularity of the AURORA calorimeter, Arbor can efficiently separate nearby showers.

Particle mapping

The visible-reconstructed particle mapping is constructed using the truth links recorded in the simulation sample. These truth links connect each tracker/calorimeter hit with the particles that excite the hits. The hits are also associated with reconstructed particles through the reconstruction procedure. The mapping is thus

established via hits. By analyzing the mapping, all reconstructed particles are labeled into 15 types:

- 10 types of 1-1 correspondence (in shades of green in Fig. 2): e^\pm , μ^\pm , π^\pm , K^\pm , p/\bar{p} , γ , K_L^0 , n , \bar{n} , and others such as K_S^0 and Λ .
- Charged particles with no track (in light green in Fig. 2): Since a track is an indispensable feature that provides more precise measurement and key information for identifying charged particles, the absence of a track will undoubtedly affect the identification of charged particles. We therefore classify this case as a separate category.
- Fake particles from charged shower fragments (in orange in Fig. 2).
- Fake particles from neutral shower fragments (in yellow in Fig. 2).
- Multi-track: more than one charged reconstructed particle (track) is mapped to one visible particle. Considering that the proportion of this type is less than 0.03%, we omit this case in the mapping schematic diagram for simplicity.
- In Arbor reconstruction, a small fraction ($< 0.1\%$) of reconstructed particles cannot be mapped to any visible particle. These particles are virtually reconstructed through Arbor’s supplementary “energy flow” strategy to balance energy during track and cluster matching. Since this is an algorithm-dependent case, it is also omitted in the mapping schematic diagram.

ParT

ParT [32] is used to classify reconstructed particles through an encoder-only Transformer architecture. This model processes both individual particle features and pairwise particle interaction features. The particle features are first embedded with a 3-layer multi-layer perceptron (MLP) and then passed through 10 layers of multi-head self-attention, where the pairwise interaction features introduce a pre-softmax bias to the attention mechanism. The final particle embeddings, obtained from the last attention layer, are further processed by three fully connected layers, each with 128 neurons and a dropout rate of 0.1. A final linear layer projects the embeddings into a space with dimensions corresponding to the number of classes, generating a set of scores that represent the classification probabilities for each particle.

Training setup

Each reconstructed particle is characterized by 55 observable variables at the particle, cluster, and track levels, as listed in Table 2. These variables serve as input features for ParT, which outputs the likelihoods for different types of reconstructed particles. The complete sample of 1 million reconstructed $\nu\bar{\nu}H, H \rightarrow gg$ events are split into three independent sets in a ratio of 6:2:2 for training, validation, and evaluation of the ParT model. A ParT model with approximately 2.2 million parameters is constructed and trained for 30 epochs.

Fake particle suppression

Fake particles are those extra reconstructed neutral particles arising from shower fragments of charged particles, denoted by f^\pm . Fig. 7 compares the output f^\pm -likelihoods

Table 2 Input variables of ParT.

Object level	Observable variables
Reconstructed particle	4-momentum (E, p_x, p_y, p_z) Direction (θ, ϕ) Number of tracks and clusters
Track	Number of hits Endpoint position 3-momentum ($ \vec{p} , p_x, p_y, p_z, p_T$) dE/dx (mean of 5–85% truncation and quartiles)
Cluster	Number of hits Energy Position of shower starting point Position of center of gravity Fractal dimension [55] Second moment (M_2) Distance between ECAL inner surface and shower starting point Distance between ECAL inner surface and center of gravity Distance between ECAL inner surface and the innermost hit Distance between ECAL inner surface and the outermost hit Maximum distance between cluster hits and the track helix (for charged particles) Maximum distance between cluster hits to the axis from the innermost hit to the center of gravity Average distance between cluster hits to the axis from the innermost hit to the center of gravity Hit time spectrum (the fastest time and quintiles)
Closest charged cluster	Minimum distance between cluster hits of each other Number of hits Energy Ratio of E_{cluster} to p_{track}

of different types of reconstructed neutral particles. Identified f^\pm is defined as those reconstructed neutral ones with f^\pm -likelihood larger than some threshold. By excluding the identified f^\pm , the optimal BMR of 2.75% is derived when f^\pm -likelihood > 0.85 .

Acknowledgements. We would like to express our sincere gratitude to Jianchun Wang for his continuous support throughout this work. In addition, we thank Hengyu

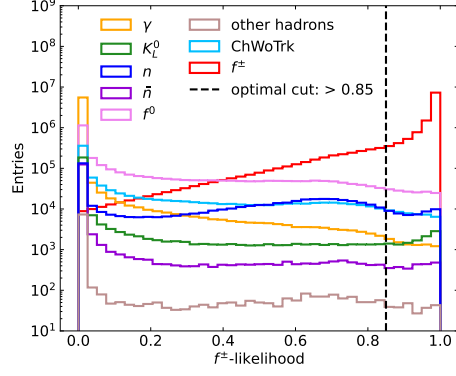


Fig. 7 Likelihoods of fake particles from charged shower fragments (f^\pm) of different types of reconstructed neutral particles. f^0 represents fake particles from neutral shower fragments. The dashed line corresponds to the optimal working point to minimize the BMR by excluding the identified f^\pm .

Wang for his study on HCAL resolution, which provided an important reference for our research. We are also grateful to Yuxuan Zhang for his review of the manuscript and valuable suggestions. This work is supported in part by the National Natural Science Foundation of China under grant No. 12042507, the National Key R&D Program of China under Contracts No. 2022YFE0116900, and the National Key Program for S&T Research and Development (2023YFA1606300). Additional support is provided by the Fundamental Research Funds for the Central Universities, Peking University.