Lepton identification at particle flow oriented detector for the future e+e- Higgs factories (Postprint)

Authors: Dan Yu, Manqi Ruana, Vincent Boudry, Henri Videau

Date: 2017-11-10T00:00:00+00:00

Abstract

The lepton identification is essential for the physics programs at high-energy frontier, especially for the precise measurement of the Higgs boson. For this purpose, a Toolkit for Multivariate Data Analysis (TMVA) based lepton identification (LICH1) has been developed for detectors using high granularity calorimeters. Using the conceptual detector geometry for the Circular Electron-Positron Collider (CEPC) and single charged particle samples with energy larger than 2 GeV, LICH identifies electrons/muons with efficiencies higher than 99.5% and controls the mis-identification rate of hadron to muons/electrons to better than 1%/0.5%. Reducing the calorimeter granularity by 1-2 orders of magnitude, the lepton identification performance is stable for particles with E > 2 GeV. Applied to fully simulated eeH/mmH events, the lepton identification performance is consistent with the single particle case: the efficiency of identifying all the high energy leptons in an event, is 95.5-98.5%.

Full Text

Abstract

Lepton identification is essential for physics programs at the high-energy frontier, particularly for precise measurements of the Higgs boson. To this end, a Toolkit for Multivariate Data Analysis (TMVA) based lepton identification algorithm called LICH (Lepton Identification for Calorimeter with High granularity) has been developed for detectors employing high-granularity calorimeters. Using the conceptual detector geometry for the Circular Electron-Positron Collider (CEPC) and single charged particle samples with energies above 2 GeV, LICH identifies electrons and muons with efficiencies exceeding 99.5% while controlling the hadron-to-muon/electron mis-identification rates to better than 1%/0.5%. Even when the calorimeter granularity is reduced by 1-2 orders of

magnitude, the lepton identification performance remains stable for particles with E > 2 GeV. Applied to fully simulated eeH/μμH events, the lepton identification performance is consistent with the single-particle case: the efficiency for identifying all high-energy leptons in an event ranges from 95.5% to 98.5%.

1 Introduction

Following the discovery of the Higgs boson, the precise determination of its properties has become the central focus of particle physics experiments. Phenomenological studies demonstrate that physics at the TeV scale could be revealed if Higgs couplings can be measured with percent-level accuracy [1][2]. While the LHC is a powerful Higgs factory, the precision of Higgs measurements there is limited by enormous QCD backgrounds and large theoretical and systematic uncertainties. Furthermore, Higgs signals at the LHC are typically tagged by their decay products, making these measurements inherently model-dependent. Consequently, the precision of Higgs couplings at the HL-LHC is generally limited to the 5–10% level, depending on theoretical assumptions [3][4].

Electron-positron colliders play a complementary role to hadron colliders for Higgs measurements, offering distinct advantages. Numerous electron-positron Higgs factories have been proposed, including the International Linear Collider (ILC), the Compact Linear Collider (CLIC), the Future e+e- Circular Collider (FCC-ee), and the CEPC [1][5][6]. These proposed facilities can select and reconstruct Higgs events with nearly 100% efficiency while determining the absolute values of Higgs couplings. Compared to the LHC, they achieve far superior accuracy in measuring the Higgs total width and searching for exotic Higgs decays, with statistical errors dominating the measurement uncertainties. For instance, the Circular Electron-Positron Collider (CEPC) is expected to produce one million Higgs bosons during its Higgs operation, enabling Higgs coupling measurements with percent-level or even per-mille-level precision [6].

Lepton identification is crucial for precise Higgs measurements. The Standard Model Higgs boson has approximately a 10% probability to decay into final states containing leptons, such as H → WW* →   / qq, H → ZZ* →   qq, H → , H → μμ, etc. The Higgs also has a branching ratio Br(H → bb) = 58%, and lepton identification provides essential input for jet flavor tagging and jet charge measurement. Moreover, the Higgs boson has a significant probability of being produced in association with leptons. For example, in ZH events—the dominant Higgs production process at 240–250 GeV electron-positron collisions—approximately 7% of Higgs bosons are produced with a pair of leptons (Br(Z → ee) and Br(Z → μμ) = 3.36%). At electron-positron colliders, ZH events with Z decaying into a lepton pair are considered the golden channel for HZZ coupling and Higgs mass measurements [7]. Additionally, leptons serve as intensive trigger signals at proton colliders to select physics events from overwhelming QCD backgrounds.

The Particle Flow Algorithm (PFA) has become the paradigm for detector design at the high-energy frontier [8,9,6,12]. Its key principle is to reconstruct every final-state particle in the most suitable sub-detectors and build all physics objects from these particles. PFA-oriented detectors achieve high efficiency in reconstructing physics objects such as leptons, jets, and missing energy. The PFA also significantly improves jet energy resolution since charged particles, which contribute the majority of jet energy, are measured with much better precision in trackers than in calorimeters [14,9,10,11,13].

To reconstruct every final-state particle, the PFA requires excellent separation capability, achieved through highly granular calorimeters. In detector designs such as the International Large Detector (ILD) or the Silicon Detector (SiD) [1,15], the total number of readout channels in calorimeters reaches 10^8. In addition to cluster separation, these detectors provide detailed spatial, energy, and even temporal information on shower development. Accurate interpretation of this recorded information enhances the physics performance of the full detector [16].

Using information from high-granularity calorimeters and dE/dx information from the tracker, LICH—a dedicated lepton identification algorithm for Higgs factories—has been developed. Using the CEPC conceptual detector geometry [6] (based on ILD) and the Arbor reconstruction package [14], its performance has been tested on both single particles and physics events. For single particles with energy above 2 GeV, LICH achieves better than 99.5% efficiency in identifying muons and electrons, and 98% for pions. Its performance on physics events (eeH/μμH) yields final efficiencies consistent with the single-particle level.

This paper is organized as follows. Section 2 presents the detector geometry and samples. Section 3 summarizes the discriminant variables measured from reconstructed charged particles and describes the algorithm architecture. Section 4 presents LICH performance on single-particle events. Section 5 explores correlations between LICH performance and calorimeter geometry. Section 6 studies LICH performance on ZH events where Z decays into ee or μμ pairs, comparing results with single-particle events. Section 7 summarizes the results and discusses the impact of calorimeter granularity.

2 Detector Geometry and Sample

The reference geometry in this paper is the CEPC conceptual detector [6], developed from the ILD geometry [1]. ILD is a PFA-oriented detector designed for center-of-mass energies up to 1 TeV, equipped with a low-material tracking system and extremely high-granularity calorimeter systems.

In this CEPC conceptual detector design, the forward region and yoke thickness have been adjusted for the CEPC collision environment relative to the ILD detector. The core component is a large 3.5 Tesla solenoid with an inner radius of 3.4 meters and length of 8.05 meters, housing both the tracker and calorimeter systems. The tracking system comprises a TPC as the main tracker, a vertex system, and silicon tracking devices. The material in front of the calorimeter is limited to approximately 5% radiation length. Both the electromagnetic calorimeter (ECAL) and hadronic calorimeter (HCAL) employ sampling structures with extremely high granularity. The ECAL uses tungsten as absorber and silicon as sensor, divided into 30 layers in depth, with each layer segmented into 5 × 5 mm² cells in the transverse direction. The HCAL uses stainless steel absorber and GRPC (Glass Resistive Plate Chamber) sensor layers, with 10 × 10 mm² cells and 48 total layers.

As a Higgs factory, the CEPC will operate at 240–250 GeV center-of-mass energy. To study lepton identification performance, we simulated single-particle samples (, µ , and e) over an energy range of 1-120 GeV (1, 2, 3, 5, 7, 10, 20, 30, 40, 50, 70, 120 GeV). At each energy point, 100k events were simulated for each particle type, following a flat distribution in and over the 4 solid angle.

These samples were reconstructed using Arbor (version 3.3). To isolate lepton identification performance from PFA reconstruction effects and geometry defects, we selected events where only one charged particle was reconstructed. The total number of such events is recorded as N Particle, and the number identified with correct particle types as N Particle,T. The lepton identification performance is expressed as a migration matrix in Table 2 , where diagonal elements represent identification efficiencies (defined as N Particle,T/N Particle) and off-diagonal elements P represent the probability of a type i particle being mis-identified as type j.

3 Discriminant Variables and Output Likelihoods

LICH takes individual reconstructed charged particles as input, extracts 24 discriminant variables for lepton identification, and calculates the corresponding likelihoods for being an electron or muon. These variables fall into five categories:

dE/dx: For a track in the TPC, the energy loss per unit distance follows a Landau distribution. The dE/dx estimator used here is the average value after truncating the tails at both edges of the Landau distribution (first 7% and last 30%). dE/dx provides strong discriminating power to distinguish electron tracks from others at low energies (below 10 GeV) (Figure 1 [Figure 1: see original paper]).

Fractal Dimension: The fractal dimension (FD) describes the self-similar behavior of shower spatial configurations. Following the original definition in [16], fractal dimension is directly linked to shower compactness. At fixed energy, electromagnetic showers are much more compact than muon or hadron showers, resulting in larger FD values. Muon showers typically exhibit a one-dimensional MIP (Minimum Ionizing Particle) track configuration, yielding FD values close to zero. Hadronic shower FD values lie between EM and MIP tracks since they contain both components. A typical FD distribution for 40 GeV showers is shown in Figure 2 [Figure 2: see original paper]. For any calorimeter cluster, LICH calculates five different FD values: from ECAL hits, HCAL hits, hits in the first 10 or 20 layers of ECAL, and all calorimeter hits.

Energy Distribution: LICH constructs variables from shower energy information, including the fraction of energy deposited in the first 10 ECAL layers relative to the entire ECAL, and energy deposited in cylinders around the incident direction with radii of 1 and 1.5 Molière radii.

Hit Information: This includes the number of hits in ECAL and HCAL, number of ECAL (HCAL) layers hit by the shower, and number of hits in the first 10 ECAL layers.

Shower Shape and Spatial Information: Spatial variables include the maximum distance between a hit and the extrapolated track, maximum and average distances between shower hits and the shower axis (defined by the innermost point and center of gravity), depth (perpendicular to detector layers) of the center of gravity, and shower depth defined as the distance between innermost and outermost hits.

The correlation of these variables at 40 GeV is summarized in Figure 3 [Figure 3: see original paper]; all variable definitions are listed in Appendix A. Notably, dE/dx measured from tracks does not correlate with any calorimeter-based variables. Some variables are highly correlated, such as FD_ECAL (FD calculated from ECAL hits) and EcalNHit (number of ECAL hits), but all are retained because their correlations vary with energy and polar angle.

LICH employs TMVA [17] methods to combine these input variables into two likelihoods corresponding to electrons and muons. Multiple TMVA methods were tested, and Boosted Decision Trees with Gradient boosting (BDTG) was selected for superior performance. The e-likeness (L) and μ-likeness (Lμ) for different particles in a 40 GeV sample are shown in Figure 4 [Figure 4: see original paper].

4 Performance on Single-Particle Events

The phase space spanned by lepton likelihoods (L and Lμ) can be divided into domains corresponding to different particle categories, adjustable according to physics requirements. This paper demonstrates lepton identification performance on single-particle samples using the following categories: - **Muon:** Lμ > 0.5 - **Electron:** L > 0.5 - **Pion:** 1 - (Lμ + L) > 0.5 - **Undefined:** Lμ < 0.5 & L < 0.5 & 1 - (Lμ + L) < 0.5

The probability of undefined particles is very low (<10 ^{3}) in single-particle samples with these categories.

Since the distribution of these variables depends on the polar angle of the initial particle ( ), the TMVA was trained independently on four subsets: - **Barrel 1:** Middle of barrel (|cos | < 0.3) - **Barrel 2:** Edge of barrel (0.3 < |cos | < 0.7) -

Overlap: Overlap region of barrel and endcap (0.7 < |cos | < 0.8) - **Endcap:** (0.8 < |cos | < 0.98)

Taking the 40 GeV charged particle sample as an example, the migration matrix is shown in Table 2 . Compared to ALEPH results for energetic taus [18], efficiencies are improved and hadron-to-lepton mis-identification rates are significantly reduced.

The lepton identification efficiencies (diagonal elements of the migration matrix) across different energies are presented in Figure 5 [Figure 5: see original paper] for the four regions. Identification efficiencies saturate at 99.9% for particles with energy above 2 GeV. For energies below 2 GeV, performance drops significantly, particularly in barrel 2 and overlap regions. In the overlap region, complex geometry limits performance; in barrel 2, charged particles with p < 0.97 GeV cannot reach the barrel and instead hit the endcaps at large incident angles, making their signals more difficult to categorize.

Regarding off-diagonal migration matrix elements, the probabilities of electrons being mis-identified as muons or pions are negligible (P μ, P < 10^{3}), with crosstalk observed at even lower levels. However, pion-to-lepton mis-identification probabilities (P μ) are of order 1% and energy-dependent. These mis-identifications primarily arise from irreducible physics effects: pion decay and production via -nucleon collisions. Meanwhile, muons have a small probability of being mis-identified as pions at energies below 2 GeV. Figure 6 [Figure 6: see original paper] shows the significant crosstalk terms (P) as a function of particle energy in the endcap region. The green shaded band indicates the probability of pion decay before reaching the calorimeter, which is roughly comparable to P μ and Pμ .

5 Lepton Identification Performance for Different Geometries

The power consumption and electronic cost of the calorimeter system scale with the number of readout channels, making it important to evaluate physics performance across different calorimeter granularities. LICH performance was analyzed over the following parameter ranges: - Number of ECAL layers: 20, 26, 30 - Number of HCAL layers: 20, 30, 40, 48 - ECAL cell size: 5×5 mm², 10×10 mm², 20×20 mm², 40×40 mm² - HCAL cell size: 10×10 mm², 20×20 mm², 40×40 mm², 60×60 mm², 80×80 mm²

In general, lepton identification performance is extremely stable across the scanned parameter space. Only for HCAL cell sizes larger than 60×60 mm² or HCAL layer numbers fewer than 20 is marginal performance degradation observed: muon identification efficiency degrades by 1-2% for low-energy particles (E 2 GeV), and pion identification efficiency degrades slightly across the full energy range (Figure 7 [Figure 7: see original paper]).

6 Performance on Physics Events

The Higgs boson is primarily produced through the Higgsstrahlung process (ZH) and, more marginally, through vector boson fusion at electron-positron Higgs factories. A significant fraction of Higgs bosons are produced with a pair of leptons (electrons and muons) from Z boson decay in the ZH process. For electrons, additional production occurs in Z boson fusion events (Figure 8 [Figure 8: see original paper]). At the CEPC, 3.6 × 10 µµH events and 3.9 × 10 eeH events are expected for an integrated luminosity of 5 ab ^{1}. In these events, the particles are rather isolated.

The eeH and µµH events provide excellent access to model-independent Higgs boson measurements using the recoil mass method [7]. The recoil mass spectrum for eeH and µµH events is shown in Figure 9 [Figure 9: see original paper], exhibiting a high-energy tail from radiation effects (ISR, FSR, bremsstrahlung, beamstrahlung, etc.), though beamstrahlung is negligible at CEPC. Bremsstrahlung effects for muons are significantly smaller than for electrons, resulting in a higher maximum and smaller tail.

Figure 10 [Figure 10: see original paper] shows the energy spectrum for all reconstructed charged particles in 10k eeH/µµH events. Leptons can be classified into two categories: initial leptons (produced with the Higgs boson) and those from Higgs decay cascades. For eeH events, the initial electron energy spectrum shows a small low-energy peak corresponding to Z fusion events. Precise identification of these initial leptons is the key physics objective for detector lepton identification performance.

Since lepton identification performance depends on particle energy and most initial leptons have energies above 20 GeV, we focused on studying high-energy particle performance for two different calorimeter cell size configurations. The µ-likeness and e-likeness distributions for electrons, muons, and pions in eeH and µµH events are shown in Figures 11 [Figure 11: see original paper] and 12 [Figure 12: see original paper]. Table 3 summarizes lepton definitions and corresponding performance under different conditions.

Identification efficiencies for initial leptons degrade by 1-2% relative to the single-particle case, primarily due to shower overlap—more significant for electrons since electromagnetic showers are much wider than muon showers, increasing overlap probability. Electrons in µµH events (and vice versa) originate from Higgs decay; their identification efficiency and purity remain reasonable. For charged leptons with energy below 20 GeV, performance degrades by about 10% due to high background statistics and cluster overlap.

The event identification efficiency, defined as the probability of successfully identifying both initial leptons, is presented in the last row of Table 3 . Event identification efficiency is roughly the square of the initial lepton identification efficiency. Comparing both geometries, reducing the number of readout channels by a factor of four degrades event reconstruction efficiency by 1.3% and

1.7% for μμH and eeH events, respectively.

7 Conclusion

High-granularity calorimetry is a promising technology for detectors at High Energy Frontier collider facilities. It provides excellent separation between different final-state particles essential for PFA reconstruction and records shower spatial development and energy profiles with unprecedented detail for energy measurement and particle identification.

To exploit the lepton identification capability of high-granularity calorimeters and provide a viable toolkit for future Higgs factories, LICH—a TMVA-based lepton identification package dedicated to high-granularity calorimeters—has been developed. Using primarily shower description variables extracted from the high-granularity calorimeter plus dE/dx information from the tracker, LICH calculates e-likeness and μ-likeness for each reconstructed charged particle, enabling lepton identification according to different physics requirements.

Applied to single-particle samples simulated with the CEPC_v1 detector geometry, typical identification efficiencies for electrons and muons exceed 99.5% for energies above 2 GeV, while pion efficiency reaches 98%. These efficiencies are comparable to ALEPH performance, but mis-identification rates are significantly improved. Ultimately, performance is limited by irreducible confusions: muon-to-electron and electron-to-muon mis-identification probabilities are negligible, while pion-to-muon mis-identification is dominated by pion decay.

The tested geometry employs ultra-high granularity calorimeters with 1×1 cm^2 cell sizes and 30/48 ECAL/HCAL layers. To reduce the total channel count, LICH was applied to more modest granularities, revealing that lepton identification performance degrades only for particle energies below 2 GeV when HCAL cell size exceeds 60×60 mm^2 or HCAL layer number falls below 20.

LICH performance was also tested on the most important physics events at CEPC. In these events, multiple final-state particles are produced in single collisions, potentially degrading particle identification performance through nearby particle overlap. Lepton identification in eeH/μμH events at 250 GeV collision energy shows single-lepton identification efficiency consistent with single-particle results. The efficiency for finding both leptons decreases by 1–2% when cell size doubles, implying the detector would need 2–4% more statistics. In eeH events, performance degradation occurs because the clustering algorithm requires further optimization.

The ultra-high granularity calorimeter designed for ILC provides excellent lepton identification capability for operation near the ZH threshold. It may be slightly over-engineered for CEPC, and slightly reduced granularity could offer a better compromise. LICH, the dedicated lepton identification algorithm for future e+e- Higgs factories, is prepared for this challenge.

References

1. T. Behnke, J.E. Brau, P.N. Burrows, et al, The International Linear Collider Technical Design Report-Volume 4: Detectors[J]. arXiv preprint arXiv:1306.6329, 2013.
2. M.E. Peskin, Physics goals of the linear collider[J]. arXiv preprint hep-ph/9910521, 1999.
3. ATLAS collaboration, Physics at a High-Luminosity LHC with ATLAS[J]. arXiv preprint arXiv:1307.7292, 2013.
4. CMS collaboration, Projected Performance of an Upgraded CMS Detector at the LHC and HL-LHC: Contribution to the Snowmass Process[J]. arXiv preprint arXiv:1307.7135, 2013.
5. CLIC CDR, A multi-TeV linear collider based on CLIC technology: CLIC Conceptual Design Report[J]. edited by M. Aicheler, P. Burrows, M. Draper, T. Garvey, P. Lebrun, K. Peach, N. Phinney, H. Schmickler, D. Schulte and N. Toge, CERN-2012-007, 2012.
6. M. Ahmad et al (The CEPC-SPPC Study Group), CEPC-SppC Preliminary Conceptual Design Report: Physics and Detector, http://cepc.ihep.ac.cn/preCDR/main_preCDR.pdf, retrieved 4th May 2015
7. Z. Chen, Y. Yang, M. Ruan, et al, Study of Higgsstrahlung Cross Section and Higgs Mass Measurement Precisions with ZH ($Z \rightarrow \mu \mu$) events at CEPC[J]. arXiv preprint arXiv:1601.05352, 2016.
8. CMS collaboration, Technical proposal for the phase-II upgrade of the CMS detector[J]. CERN, CERN-LHCC-2015-010. LHCC-P-008, 2015.
9. M.A. Thomson, Particle flow calorimetry and the PandoraPFA algorithm[J]. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 2009, 611(1): 25-40.
10. F. Beaudette, The CMS Particle Flow Algorithm[J]. arXiv preprint arXiv:1401.8155, 2014.
11. J.C. Brient, Improving the Jet Reconstruction with the Particle Flow Method; an Introduction[J]. arXiv preprint physics/0412149,
12. J.C. Brient, H. Videau, The calorimetry at the future e+ e- linear collider[J]. arXiv preprint hep-ex/0202004, 2002.
13. H. Videau, Energy flow or Particle flow-The technique of energy flow for pedestrians[C]//International Conference on Linear Colliders-LCWS04. Ecole Polytechnique Palaiseau, 2004: 105-

14. M. Ruan, Arbor, a new approach of the Particle Flow Algorithm. arXiv:1403.4784 (2014).
15. T. Abe, ILD Concept Group-Linear Collider Collaboration. The International Large Detector: Letter of Intent, 2010[J]. arXiv preprint arXiv:1006.3396, 4(10).
16. M. Ruan, D. Jeans, V. Boudry, J.C. Brient, & H. Videau, (2014), Fractal Dimension of Particle Showers Measured in a Highly Granular Calorimeter, Physical review letters, 112(1), 012001.
17. A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, H. Voss, ···& D. Dannheim (2007), TMVA-Toolkit for multivariate data analysis. arXiv preprint physics: 0703039.
18. Aleph Collaboration, Measurement of the Tau Polarisation at LEP[J]. arXiv preprint hep-ex/0104038, 2001.

Appendix A: Variable Definitions

List and meaning of variables used in TMVA not mentioned in the text: - **NH_ECALF10:** Number of hits in the first 10 layers of ECAL - **FD_ECALL20:** FD calculated using hits in the last 20 layers of ECAL - **FD_ECALF10:** FD calculated using hits in the first 10 layers of ECAL - **AL_ECAL:** Number of ECAL layer groups (each five layers forms a group) with hits - **av_NHH:** Average number of hits in each HCAL layer group (each five layers forms a group) - **rms_Hcal:** RMS of hits in each HCAL layer group (each five layers forms a group) - **EEClu_r:** Energy deposited in a cylinder around the incident direction with radius of 1 Molière radius - **EEClu_R:** Energy deposited in a cylinder around the incident direction with radius of 1.5 Molière radii - **EEClu_L10:** Energy deposited in the first 10 layers of ECAL - **MaxDisHel:** Maximum distance between a hit and the helix - **minDepth:** Depth of the innermost hit - **cluDepth:** Depth of the cluster position - **graDepth:** Depth of the cluster gravity center - **EcalEn:** Energy deposited in ECAL - **avDisHtoL:** Average distance between a hit and the axis from the innermost hit to the gravity center - **maxDisHtoL:** Maximum distance between a hit and the axis from the innermost hit to the gravity center - **NLHcal:** Number of HCAL layers with hits - **NLEcal:** Number of ECAL layers with hits - **HcalNHit:** Number of HCAL hits - **EcalNHit:** Number of ECAL hits

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv —Machine translation. Verify with original.