CPID: A Comprehensive Particle Identification Framework for Future $e^+e^-$ Colliders

Uli Einhaus
International Workshop on Future Linear Colliders
16.05.2023, SLAC
ulrich.einhaus@desy.de
The Landscape of Proposed Next-Gen Colliders / Future HTE Factories

- Many proposals under consideration – common tools desired, in particular software!
  → key4HEP / EDM4HEP
Increasing understanding that particle identification (PID), in particular charged hadron ID, is a very valuable observable at a Future Higgs Factories.

Recent studies focus on 90-250 GeV and precision flavour physics instead of direct (BSM) detection at TeV range → PID is more effective and more relevant there.

This work: new software framework for comprehensive PID.

https://arxiv.org/abs/1903.01629


https://arxiv.org/abs/2203.07535
Examples for PID Applications

- Z and W hadronic decay branching fractions via flavour tagging → make connection between quark flavour and jet composition
  [https://ediss.sub.uni-hamburg.de/handle/ediss/9634, https://ediss.sub.uni-hamburg.de/handle/ediss/9928]

- Forward-backward asymmetry in $e^+e^- \rightarrow q\bar{q}$ → study asymmetry in each flavour channel exclusively
  overview: [https://indico.desy.de/event/33640/contributions/127531]

- $H \rightarrow s\bar{s}$ with s-tagging → identify high-momentum kaons to tag s\bar{s} events
  [https://arxiv.org/abs/2203.07535]

- B and tau physics rely on excellent PID, mostly relevant at Z pole e.g. [https://indico.in2p3.fr/event/23012/contributions/89990]

- Kaon mass with TOF
  [https://pos.sissa.it/380/115/]

- Track refit with correct particle mass for better momentum and vertex
  [https://agenda.linearcollider.org/event/8498/]

- Overview: [https://indico.cern.ch/event/1256374/contributions/5338875/]

Uli Einhaus | LCWS | 16.05.2023 | Page 4
Possible PID Technology

- Gaseous trackers (Time Projection Chamber, Drift Chamber): specific energy loss $dE/dx$, via gas ionisation, up to 30 GeV

- Ring Imaging Cherenkov Detectors: Cherenkov angle, via imaging, 10 to 50 GeV

- Time of Propagation Counter: Cherenkov angle, via timing, up to 10 GeV

- Time of Flight: time, via Silicon timing, up to 5 GeV
Central Questions

- How do we combine the different technologies best?

- How can we create a general assessment of PID performance valid for all of them?

- Can we use machine learning to extract the best performance from the PID observables?

- Optimise detectors and compare them
  - At what timing resolution starts TOF to be relevant for flavour tagging?
  - How does my physics channel depend on the dE/dx (dN/dx) resolution?
  - What if we use a silicon tracker + RICH instead of a TPC in ILD?
New Framework: Comprehensive Particle Identification (CPID)

- Modularity as core philosophy:
  - observables algorithms
  - training methods (MVAs / NNs / etc.)

- Core code takes care of book keeping
  - simple, well defined data structures for storage (TTree) and interfaces (std::vector)

- For now, being implemented in LCIO / Marlin in iLCSoft
  - immediately usable in Key4HEP via ‘Marlin wrapper’
  - target: implement in EDM4HEP, make available to whole future colliders community

- In ILD: goal to replace current LikelihoodPIDProcessor
General Structure

for regular user
for module developer

PID observables
algorithms

steering file

processor parameters
input algorithms
algorithm parameters

input sample:
PFOs

Comprehensive
PIDProcessor

select events by
PFO properties

store observables
in TTrees

output:
PFO PID

MCParticles
Tracks, Clusters, etc.

TTrees

training &
evaluation
algorithm

weight file

PFOs

PFO collection

observables
values

for regular user
for module developer

performance
assessment
Example 1: $\pi/K$ Separation with Combined Observables

- $dE/dx + TOF$
- Single particles ‘calibration’ events, flat in $\log(p)$ and $\cos(\theta)$
- BDT with sig = K, bkg = $\pi$; train & infer per 12 mom bins and per used observable(s)
  $\rightarrow$ How do we calculate a separation power from a BDT score?
p-value Assessment

- Find cut with $\text{mis-ID} = 1 - \text{efficiency} = \text{p-value} \rightarrow \text{find Gaussian quantile}$
  $\rightarrow \text{compute } Z = 2 \times \text{quantile of standard Gauss}$

![Graph showing distributions of kaons and pions with BDT score vs. MVA_BDT_S](https://indico.psi.ch/event/7080/contributions/31950/attachments/22852/23789/pid_kgoetzen_separationpower.pdf)
p-value Assessment

- Find cut with $\text{mis-ID} = 1 - \text{efficiency} = p$-value $\rightarrow$ find Gaussian quantile
  $\rightarrow$ compute $Z = 2 \cdot \text{quantile of standard Gauss}$

$\text{p} = 6.2\%$

mis-ID = 1 - efficiency = 6.2%
p-value Assessment

- ‘Central tail split’ of BDT score is equivalent to crossing point of ROC curve with x=y line

![Graph showing BDT score distributions for kaons and pions](image)

![ROC curve showing backgr rejection vs signal eff](image)
Example 1: $\pi/K$ Separation with Combined Observables

- $dE/dx + TOF$
- Single particles ‘calibration’ events, flat in log($p$) and cos($\theta$)
- BDT with sig = K, bkg = $\pi$; train & infer per 12 mom bins and per used observable(s)

ILD Interim Design Report

CVID, BDT result

Example 1: $\pi/K$ Separation with Combined Observables

- $dE/dx$ very similar
- TOF levels out at low momenta due to misreconstructed events, which cause a constant finite background in the p-value assessment
  → still covers $dE/dx$ blind spot, $S > 4$ still good enough!
  → same performance, less ‘fancy’ way to show it, but more honest

analogue to ILD Interim Design Report

CPI, BDT result
Example 2: Multiclass Confusion Matrix

- \(\text{dE/dx + calorimeter cluster shapes}\)
- Single particle ‘calibration’ events, flat in \(\log(p)\) and \(\cos(\theta)\)
- \(e, \mu, \pi, K, p\); multiclass BDT; confusion matrix with \(\text{eff/pur}\) on diagonal
- CPID with simple BDT already competitive to LikelihoodPID
Example 2: Multiclass Confusion Matrix

- \( \frac{dE}{dx} + \) calorimeter cluster shapes
- Single particle ‘calibration’ events, flat in \( \log(p) \) and \( \cos(\theta) \)
- \( e, \mu, \pi, K, p; \) multiclass BDT; confusion matrix with \( \text{eff/pur} \) on diagonal
- Addition of TOF gives immediately better result – previously hard, easy in CPID

![Confusion Matrix Diagrams]

**LikelihoodPID**

- Reco PDG:
  - protons
  - kaons
  - pions
  - muons
  - electrons

- MC truth PDG:
  - electrons
  - muons
  - pions
  - kaons
  - protons
  - other

**CPID framework; + TOF**

- Reco PDG:
  - protons
  - kaons
  - pions
  - muons
  - electrons

- MC truth PDG:
  - electrons
  - muons
  - pions
  - kaons
  - protons
  - other
Summary

- New Comprehensive PID framework under development
- Aims to provide common platform for future e+e- Higgs factories
- Allow for
  - combining and comparing PID technologies
  - assessing on full detector level with robust performance quantities
  - easy-to-use retraining and flexible adaptation
- First performance indicators already comparable to state-of-the-art
- Application to ongoing ILD physics analyses under discussion
- Your feedback and input are welcome!
What is particle identification / PID

• Identification of the species of high energy particles
  → e.g. e, μ, γ, π, K, p, Λ, n, [whatever is detector-stable]

• Focus of dedicated PID systems is on charged hadron separation, specifically kaon ID

• Dedicated electron and muon ID relevant at lower momenta

• Also don’t forget V0s: K⁰_S, Λ
p-value assessment

Gaussian quantile is inverse of distribution function \( \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-z^2/2} \, dz \)

Due to our definition, sep. power \( Z = 2 \cdot \Phi^{-1}(1 - p) \)

\( p = 6.2\% \)

\( z = 1.54 \)

\[ \text{ROOT::Math::gaussian_cdf(z)} \]

\[ \text{ROOT::Math::gaussian_quantile_c(p, 1)} \]
Higgs to strange

- Study Higgs to strange coupling
- Cute-based analysis, final cut:
  developed strange tagger using $K^\pm$, $K^0_s$, $\Lambda^0$
  → allows to cut background by factor 3
- Results in upper limit on $\kappa_s < 6.7$

https://arxiv.org/abs/2203.07535
View of the steering file

- Steering file
  - input sample
  - observables algorithms
  - signal categories PDGs
  - evaluation algorithm
  - weight file
  - sample cuts etc.

```xml
<processor name="MyComprehensivePIDProcessor" type="ComprehensivePIDProcessor">
  <parameter name="PFOCollection" type="string" value="PandoraPFOs"/>
  <parameter name="inputAlgoSpecs" type="StringVec">
    dEdx
    TOF:TOF0
    TOF:TOF10
    TOF:TOF50
    dEdx_RCD:dEdx_RCD
  </parameter>
  <parameter name="dEdx.F" type="FloatVec" value="1 2 3"/>
  <parameter name="dEdx.S" type="StringVec" value="a b c"/>
  <parameter name="TOF0.S" type="StringVec" value="TOFEstimators0ps"/>
  <parameter name="TOF10.S" type="StringVec" value="TOFEstimators10ps"/>
  <parameter name="TOF50.S" type="StringVec" value="TOFEstimators50ps"/>
  <parameter name="dEdx_RCD.F" type="FloatVec">
    -1.28883368e-02 2.72959919e+01 1.10560871e+00 -1.74534200e+00 -9.84887586e-07
    6.49143971e-02 1.55775592e+03 9.31848047e+08 2.32201725e-01 2.50492066e-04
    6.54955215e-02 8.26239081e+04 1.92933904e+07 2.52743206e-01 2.46557525e-04
    7.52110150e-02 1.59711041e+04 1.79625604e+06 3.15315795e-01 2.30414997e-04
    7.92251260e-02 6.38129720e+04 3.82995071e+04 2.80793601e-01 7.14371743e-04
  </parameter>
</processor>
```