Detector optimisation for future linear collider

Boruo Xu
of King's College

A dissertation submitted to the University of Cambridge
for the degree of Doctor of Philosophy

Abstract

This is my abstract. To be or not to be.

Declaration

This dissertation is the result of my own work, except where explicit reference is made to the work of others, and has not been submitted for another qualification to this or any other university. This dissertation does not exceed the word limit for the respective Degree Committee.

Boruo Xu

Acknowledgements

Of the many people who deserve thanks, some are particularly prominent, such as my supervisor...

Preface

This will be my preface. Where is Wolly?

Contents

"Two bags of pork scratchings are worth a bag of gold."

— Joris the Dutch

Chapter 1

Let's make introduction great again

"The journey of a thousand miles begins with a single step."
— Lao Zi

âĂŢ Introduction

1.1 Future Linear Colliders

Basic intro. LHC.

Next challenge

Future Options. FCC vs LC

LC options

1.2 Motivation

Photon - passage through matter. Photon electromagnetic shower

Since Higgs discovery in the LHC in 2012, Higgs

Ha there is a higgs.

We found higgs. Higgs is cool. It explains mass.

Why double higgs. Double higgs coulpling is unique to linear collider. It can revel much about the BSM models.

Generator level study has performed. ILC has done this this and that. gHHH in CLIC before

Here we do things differently. First subchannels, then extract both couplings simultaneously.

Chapter 2

Theoretical overview

"ILC will be built next year"
— Mysterious person

This chapter provides a theoretical overview which would be used in the subsequent chapters. A short review of the Standard Model of Particle Physics, the current best particle theory, is provided, with an emphasis on the Higgs mechanism and the Higgs boson. A general parametrisation of the Higgs theory, beyond the Standard Model, is discussed, and supplies the theoretical background for the physics analysis in the chapter 7.

2.1 Overview of the Standard Model

The Standard Model (SM) is a quantum field theory concerning three fundamental interactions of nature: the electromagnetic, weak and strong integrations. The SM also describes the interactions between the sub-atomic particles. The deployment and the experimental verification of the SM throughout the second half of the 20th century is one of the greatest triumph of the particle physics. The most recent discovery of the Higgs boson in 2012 [1] further verified the theory. This chapter summarise the SM based on the summaries of the SM [2–5].

The fundamental particles in the SM consist of three categories: force exchange bosons, leptons and neutrinos, and quarks. In the SM, the force exchange bosons carries the fundamental forces between particles. For example, photon is the force carrier of the

electromagnetic force. W^+, W^-, and Z are the force carriers of the weak force. Gluon, g, is the force carrier of the strong force. These bosons will be discussed further in section 2.5. There is also the Higgs boson from the spontaneous symmetry breaking of the Higgs filed, which is discussed in section 2.8

Another category of fundamental particles contains leptons and neutrinos. These particles are fermions. For each fermion in the SM, there is an anti-fermion with same mass and spin, but opposite charge. Leptons and neutrinos have three generations. Each generation has same interaction properties, but different masses. Although neutrinos could not be directly detected, measurements of the Z decay width strong suggested the three generations of neutrinos [6]. Leptons and neutrinos experience weak forces as well as electromagnetic forces, which will be further discussed in section 2.5.

The last category of fundamental particles are quarks, which are also fermions and have three generations. Each generation has a positively charged up type quark and a negatively charged down type quark. Quarks experiences all three fundamental forces described by the SM.

The SM has enjoyed great success with theoretical predictions being experimentally varied. Some highlights included the discovery of the top quark in 1995 cite, the tau netrino in 2000 cite and the Higgs bosons in 2012 citeAad:2012tfa. However, there are observations which are not explained by the SM. One issue is that the SM does not incorporate the gravitational force. There have been attempts to modify the SM but no conclusive theory yet. Another issue is that the SM does not allow neutrino masses and mixings. There have been many theories beyond the Standard Model (BSM). One such example is the generalisation of the Higgs theory to allow non SM coupling strengths. This will be discussed in section 2.9.

Notations and conventions will be introduced. The overview of the Standard Model starts with the quantum electrodynamics, and its generalisation to quantum chromodynamics. The unification of electromagnetism and weak interaction, electroweak gauge theory will be discussed. Afterwards, Higgs mechanism and Yukawa couplings will be introduced to explain bosons and fermions masses whilst preserving the Lagrangian symmetry. This will be followed by a detailed discussion on the Standard Model Higgs boson, its mass and interactions with other particles. The chapter finishes with explanation for possible Higgs theories beyond the Standard Model, the Lagrangian of the Higgs interaction, and observables, which forms the theoretical background for the analysis on the double Higgs production in the chapter 7.

2.2 Notations and conventions

The natural unit is used in this document, $\hbar = c = 1$. The metric is mostly-minus, $\eta^{\mu\nu} = \mathrm{diag}(1, -1, -1, -1)$. The Dirac gamma matrices are represented with γ^μ, with μ goes from 0 to 3. $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$. $\overline{\psi} = \psi^\dagger\gamma^0$. Einstein summation convention is used as well this document.

This set of notations allows a contracted pair to be a Lorentz invariant. For a Weyl spinor, ψ_α, the mass term in the lagrangian is of the form $\psi^\alpha\psi_\alpha$, which is the Majorana mass term. The contracted pair between two different Weyl spinors would form a Dirac mass term.

2.3 Quantum electrodynamics

The natural starting point to introduce the SM is the quantum electrodynamics (QED). The QED is a quantum field theory explaining electromagnetic interactions. The theory involves a spin-half Dirac (electron) field ψ and a vector (photon) field A_μ. When the local (gauge) symmetry is imposed, which is equivalent to the Lagrangian invariance under transformations,

$$\psi \rightarrow e^{ie\phi(x)}, \; A_\mu \rightarrow A_\mu - \partial^\mu\phi(x), \tag{2.1}$$

the Lagrangian is fixed to be

$$\mathcal{L}_{QED} = \overline{\psi}\left(i\gamma^\mu D_\mu - m\right)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}, \tag{2.2}$$

if up to cubic terms are allowed in the fields. There are two free parameters in the QED, m the electron mass, and e the electron charge. The mass term for the photon, $v^2 A_{nu}A^\nu$, is forbidden by gauge invariance.

The QED has been verified experimentally. One of the greatest prediction is the spin magnetic dipole moment of the electron, defined as $\vec{\mu} = g_s\frac{Qq}{2m}\vec{s}$. The g_s is predicted to be 2 by the Dirac equations. The small corrections to the value comes from the electron's interaction with virtual photons, so called higher "loop" corrections in Feynman diagrams. The precise agreement of the theoretical prediction and the experimental value is a success of the QED.

2.4 Quantum chromodynamics

Like the QED, the quantum chromodynamics (QCD) a quantum field theory explaining strong interactions. There are eight gauge bosons, gluons, coupling to nine fermions, quarks. Unlike the QED, the theory is invariant under local non-Abelian SU(3) transformations. Gulons can interact with other gluons, and carry colour charge (red, green, and blue). Nine quarks transform as colour triplets. The QCD Lagrangian is

$$\mathcal{L}_{QCD} = \sum_{f\in u,d,s,c,b,t} \overline{\psi}\left(i\gamma^{\mu}\partial_{\mu} - g_s\gamma^{\mu}G^a_{\mu}\frac{\lambda^a}{2} - m_f\right)\psi - \frac{1}{4}G^a_{\mu\nu}G^{a\mu\nu}, \tag{2.3}$$

where g_s is the strong coupling constant. a is the colour charge. λ is the Gell-Mann matrices. $G^a_{\mu\nu}$ is the gluon field strength, given by

$$G^a_{\mu\nu} = \partial_{\mu}\gamma^a_{\nu} - \partial_{\nu}\gamma^a_{\mu} - g_s f_{abc}G^b_{\mu}G^c_{\nu}. \tag{2.4}$$

The last extra term comparing to QED indicates the non-Abelian nature of the QCD.

2.5 The electroweak interaction

The electroweak interaction can be thought of a extension to the QED to incorporate the weak force, and to explain different coupling strength to left-handed and right-handed fermions. However, the Lagrangian does not allow the massive electroweak force exchange bosons and fermions, which are explained by the Higgs mechanism and the Yukawa interactions.

There are four vector boson fields in the theory, 3 W and 1 B field. The Lagrangian can be divided into two parts: the bosonic self interaction and the couplings to the fermions.

$$\mathcal{L}_{Electroweak} = \mathcal{L}_{Boson} + \mathcal{L}_{Fermion} \tag{2.5}$$

The bosonic self interaction Lagrangian, $\mathcal{L}_{Boson}$, is given by

$$\mathcal{L}_{Boson} = -\frac{1}{4}W^i_{\mu\nu}W^{i\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}, \tag{2.6}$$

where

$$W^i_{\mu\nu} = \partial_\nu W^i_\mu - \partial_\mu W^i_\nu - g\varepsilon^{ijk} W^j_\mu W^k_\nu \tag{2.7}$$

$$B_{\mu\nu} = \partial_\nu B_\mu - \partial_\mu B_\nu \tag{2.8}$$

B field is invariant under U(1). W field is invariant under non-Abelian SU(2) transformations. g is the coupling strength of the W field. The indices, i, j, and k indicates 3 W fields, going from 1 to 3.

The fermionic part of the Lagrangian, $\mathcal{L}_{\text{Fermion}}$, has different components for the left-handed and right-handed fermions, given by

$$\mathcal{L}_{\text{Fermion}} = \sum_{\psi \in \text{fermions}} \overline{\psi}_L \gamma^\mu D^L_\mu \psi_L + \overline{\psi}_R \gamma^\mu D^R_\mu \psi_R \tag{2.9}$$

D^L_μ and D^R_μ are defined as

$$D^L_\mu = \partial_\mu + ig\frac{\tau_i}{2} W^i_\mu + ig' Y_\psi B_\mu \tag{2.10}$$

$$D^R_\mu = \partial_\mu + ig' Y_\psi B_\mu \tag{2.11}$$

This Lagrangian allows W and B field to couple with left-handed fermions, but only B field couples to right-handed fermions. The τ_i matrices are the generators of the SU(2). Pauli spin matrices are one of the representations. Y_ψ is the hypercharge associating with the fermion field ψ. g' is the B field strength.

Physical bosons W^+, W^- only couples to left-handed fermions. Z and γ couples to both left-handed and right-handed fermions. Hence W^1 and W^2 are associated to W^+ and W^-. Mass eigenstates for Z and γ, Z_μ and A_μ are mixture of W^3_μ and B_μ.

$$Z_\mu = \cos(\theta_W) W^3_\mu - \sin(\theta_W) B_\mu \tag{2.12}$$

$$A_\mu = \sin(\theta_W) W^3_\mu + \cos(\theta_W) B_\mu \tag{2.13}$$

θ_W is the Weinberg mixing angle [7], which is determined experimentally. So far the SU(2)⊗U(1) gauge theory explains the parity violating nature of the weak interaction.

The explicit fermion mass are not allowed in the gauge symmetry. The higgs mechanism via spontaneous symmertry breaking would introduce mass terms for fermions.

2.6 Higgs Mechanism

A complex scalar Higgs field, Φ_H, is added to the electroweak Lagrangian. Φ_H transforms as a doublet of SU(2) with hypercharge $Y = \frac{1}{2}$

$$\mathcal{L}_{\text{Higgs}} = (D_\mu \Phi_H)^\dagger (D^\mu \Phi_H) - \mu^2 \Phi_H^\dagger \Phi_H - \lambda \left(\Phi_H^\dagger \Phi_H\right)^2 \tag{2.14}$$

with

$$D_\mu \Phi_H = \left(\partial_\mu + ig\frac{\tau_i}{2} W_\mu^i + ig'\frac{1}{2} B_\mu\right) \Phi_H \tag{2.15}$$

For negative μ^2, the Higgs filed potential

$$\mu^2 \Phi_H^\dagger \Phi_H + \lambda \left(\Phi_H^\dagger \Phi_H\right)^2 \tag{2.16}$$

is minimised with a Higgs vacuum potential $\frac{v}{\sqrt{2}} = \sqrt{\frac{v^2}{2\lambda}}$. After the symmetry breaking, the $\mathcal{L}_{\text{Higgs}}$ provides the mass terms for W^+, W^-, Z and γ via terms in the Lagrangian:

$$\frac{(gv)^2}{4} W_\mu^+ W^{-\mu} + \frac{\left(g^2 + g'^2\right)\mu^2}{8} Z_\mu Z^\mu \tag{2.17}$$

This provides equal mass for W^+ and W^- with massless photon.

2.7 Yukawa couplings

Section 2.6 explains the Higgs mechanism for gauge bosons gaining masses. The fermions gain masses in a similar fashion. Consider a Higgs field transforming as a doublet of SU(2) with hypercharge $Y = \frac{1}{2}$, the Yukawa couplings is given

$$\mathcal{L}_{\text{Yukawa}} = -\lambda^u \overline{q_L} \Phi_H^c u_R - -\lambda^d \overline{q_L} \Phi_H d_R - \lambda^e \overline{l_L} \Phi_H e_R + \text{h.c.} \tag{2.18}$$

$\Phi_H^c \equiv i\sigma^2 H^*$ is an SU(2) double field with hypercharge $Y = -\frac{1}{2}$ and σ is the Pauli spin matrix. u, d, and e are fields for up-type quark, down-type quark and leptons. The Lagrangian is summed over all possible quarks and leptons. The interactions terms in the $\mathcal{L}_{Yukawa}$ become mass terms when the Higgs vacuum expectation value is substituted. The fermion masses are given by

$$m_u = \frac{\lambda^u v}{\sqrt{2}}, \; m_d = \frac{\lambda^d v}{\sqrt{2}}, \; m_e = \frac{\lambda^e v}{\sqrt{2}} \tag{2.19}$$

2.8 Standard Model Higgs boson

So far, interactions between different fields in the Standard Model, as well as the mass obtaining mechanism have been discussed. Only left for discussion is the Higgs bosons, and its interactions with other fields.

For the Higgs doublet complex field in the SM, there are four real scalar degrees of freedom. By choosing the unitary gauge, three degree of freedoms are manifestly eaten. The Higgs field becomes

$$H(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix} \tag{2.20}$$

$h(x)$ is real scalar field for the Higgs boson. It is not charged under electromagnetism as it is real. The Higgs boson interaction terms in the Lagrangian with other particles can be shown by replacing v with $v + h(x)$ in previous expressions. For fermions field, ψ_i, the Higgs boson interaction term is given by

$$\mathcal{L} \supset -\frac{m_i}{v} h \bar{\psi}_i \psi_i \tag{2.21}$$

From the equation 2.17, the Higgs boson interaction terms for bosons can be shown as

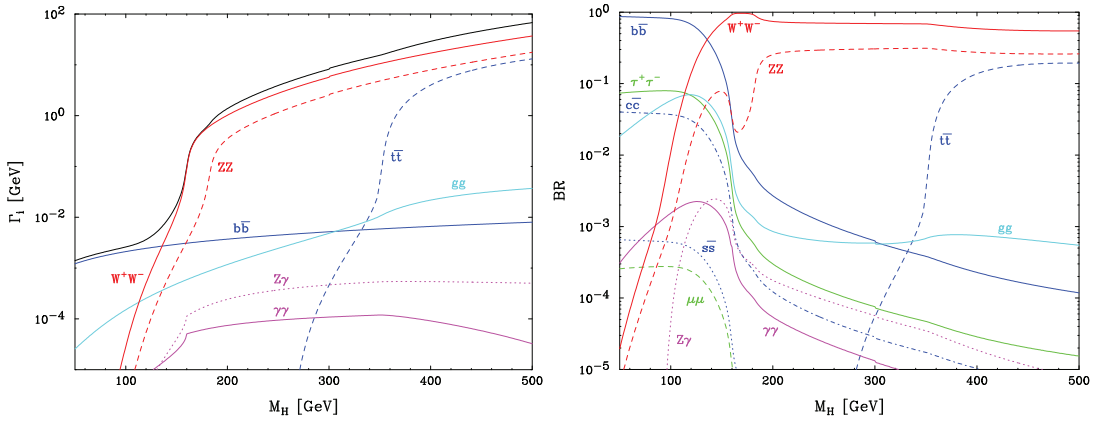
$$\mathcal{L} \supset m_W^2 \left(\frac{2h}{v} + \frac{h^2}{v^2} \right) W_\mu^+ W^{-\mu} + \frac{m_Z^2}{2} \left(\frac{2h}{v} + \frac{h^2}{v^2} \right) Z_\mu Z^\mu \tag{2.22}$$

The Higgs boson self interactions are obtained from the Higgs field potential

$$\mathcal{L} \supset \frac{\mu^2}{2}(v+h)^2 - \frac{\lambda}{4}(v+h)^4 \supset -\lambda v^2 h^2 - \lambda v h^3 - \frac{\lambda}{4}h^4 = -\frac{m_h^2}{2}h^2 - \frac{m_h^2}{2v}h^3 - \frac{m_h^2}{8v^2}h^4. \tag{2.23}$$

The Higgs boson mass, m_h is $2\lambda v^2$. The triple and quartic self interaction strengths are $-\frac{m_h^2}{2v}$ and $\frac{m_h^2}{8v^2}$. Once the m_h is determined, λ can be worked out. The Higgs boson decay width and branching fraction can be roughly worked out. For example, figure 2.1a and figure 2.1b show partial decay width and the branching ratios as a function of m_h.

The Higgs boson decaying to a pair of heavier particles, such as W^+W^- or Z Z is forbidden kinematically. However, figure 2.1 shows that the Higgs decaying to W^+W^- dominates before the mass threshold $m_H = 2m_W \sim 160$ GeV. This is because one of the $W^\pm$ gauge bosons is virtual and not on the mass shell, which is allowed by the quantum field theory. The virtual gauge boson subsequently decays to real on-shell particles.



(a) **(b)**

Figure 2.1: figure 2.1a shows Standard Model Higgs boson partial widths as a function of its mass, M_H. The total width is the black curve figure 2.1b shows selected Standard Model Higgs boson branching ratios as a function of its mass, M_H. Both plots are taken from [8]

2.9 Higgs beyond the Standard Model

Since the SM like Higgs boson discovery in 2012, it becomes important to understand of role of the Higgs boson in the electroweak spontaneous symmetry breaking. In the absence of the Higgs boson, the coupling strength of the longitudinally polarised vector

bosons grows with energy and becomes strong at TeV scale. The SM Higgs bosons moderates the interacting strength, allowing the extraction of the weak coupling at short distances. In this scenario, the SM Higgs couplings are constrained and predicted in one parameter only, the Higgs mass. But other alternative scenarios could allow the behaviour of the SM Higgs at low energy. One such example, motivated by the hierarchy problem and the electroweak data, is that the light and narrow Higgs-scalar is a composite bound state of some strongly interacting sector at the TeV scale. The couplings of the Higgs to fermions and bosons would be different to those in the SM. If the composite Higgs is the pseudo Nambu-Goldstone boson from a spontaneous global symmetry breaking, the Higgs can be naturally light [9]. Another scenario is that a composite dilaton, the pseudo Nambu-Goldstone boson arose from a spontaneous scale invariance breaking, partially behaves like a light Higgs [10]. In both scenarios, the interaction of Higgs becomes strong at high energy. The coupling of the Higgs would deviate to those in the SM.

An important physics channel for testing the Higgs theory is the double Higgs production via vector boson fusion at high energy [11–13]. For the composite Higgs scenario, the scattering amplitude increases with the energy. For the dilaton scenario, no energy defence on the scattering amplitude is expected. It is difficult for the Large Hardon Collider to measure the cross section due to the large SM background rate [12]. However, a multi-TeV linear electron position collider, such as Compact Linear Collider, would be able to precisely measure the cross section [14].

Following the assumption made in the [12, 13], the self interaction of the light scalar Higgs, h, and its coupling to other SM bosons can be described by the following Lagrangian. The notation in the [13] is followed. After the electroweak symmetry breaking, the bosonic part of the Lagrangian reads:

$$\mathcal{L} = \frac{1}{2}(\partial_\mu h) - V(h) + \left(m_W^2 W_\mu^+ W^{-\mu} + \frac{m_Z^2}{2} Z_\mu Z^\mu\right)\left[1 + 2a\frac{h}{v} + b\frac{h^2}{v^2} + \ldots\right], \tag{2.24}$$

where $V(h)$ is the h field potential,

$$V(h) = \frac{1}{2}m_h^2 h^2 + d_3\left(\frac{m_h^2}{2v}\right)h^3 + d_4\left(\frac{m_h^2}{8v^2}\right)h^4 + \ldots. \tag{2.25}$$

a, b, d_3 and d_4 are arbitrary dimensionless parameters. Higher order terms in h are omitted. a and b are proportional to the coupling strength of the VVh and $VVhh$ vertices, where $V = W^\pm, Z$. d_3 and d_4 are proportional to the triple and quartic h self coupling strength. Comparing with equation 2.22 and equation 2.23, the SM Higgs

suggests $a = b = d_3 = d_4 = 1$ and all higher order terms vanish. The dilaton scenario imposes the relation, $a = b^2$.

The scattering amplitude for $V_L V_L \to hh$ can be written as

$$A = a^2(A_{SM} + A_1\delta_b + A_2\delta_{d_3}), \tag{2.26}$$

where A_{SM} is the SM amplitude and

$$\delta_b \equiv 1 - \frac{b}{a^2}, \tag{2.27}$$

$$\delta_{d_3} \equiv 1 - \frac{d_3}{a}. \tag{2.28}$$

A_1 grows like energy squared at large center-of-mass energy, $E \gg m_V$. A_{SM} and A_2 has no energy depended. Therefore, δ_b controls the magnitude of the scattering amplitude increasing as a function of energy. δ_{d_3}, however, determines the magnitude at threshold. In an electron-positron collider, this scattering process and be studied via $e^+e^- \to \nu\bar{\nu}hh$ channel. The cross section of the channel can be written as

$$\sigma = a^4\sigma_{SM}\left(1 + A\delta_b + B\delta_{d_3} + C\delta_b\delta_{d_3} + D\delta_b^2 + E\delta_{d_3}^2\right), \tag{2.29}$$

where σ_{SM} is the cross section predicted by the SM. With suitable kinematic cuts, high-energy behaviour can be disentailed from the physics at threshold, allowing the extraction of δ_b, δ_{d_3} and hence the coupling strength g_{VVHH} and g_{HHH}. Suitable observables are variables that increases with increasing centre-of-mass energy. Two examples of such variables are the invariant mass of the two Higgses system, m_{hh}, and the sum of their transverse momenta, H_T. figure **??** shows that the m_{hh} and H_T distributions are sensitive to the values of δ_b and δ_{d_3}. The figure shows the result of a general level study performed in [13].

In the equation 2.29, a, which is proportional to g_{VVH}, only entres as an overall factor. However, a also appears in the definition of δ_b and δ_{d_3}. To extract δ_b and δ_{d_3}, a should be a constant ideally. For a multi-TeV electron-positron collider, the cross section for single Higgs production is far greater than that of the double Higgs. Figure 2.3 shows the comparison of the cross section as a function of the centre-of-mass energy. Therefore, g_{VVH} and a will be measured precisely before investigating the double Higgs production.

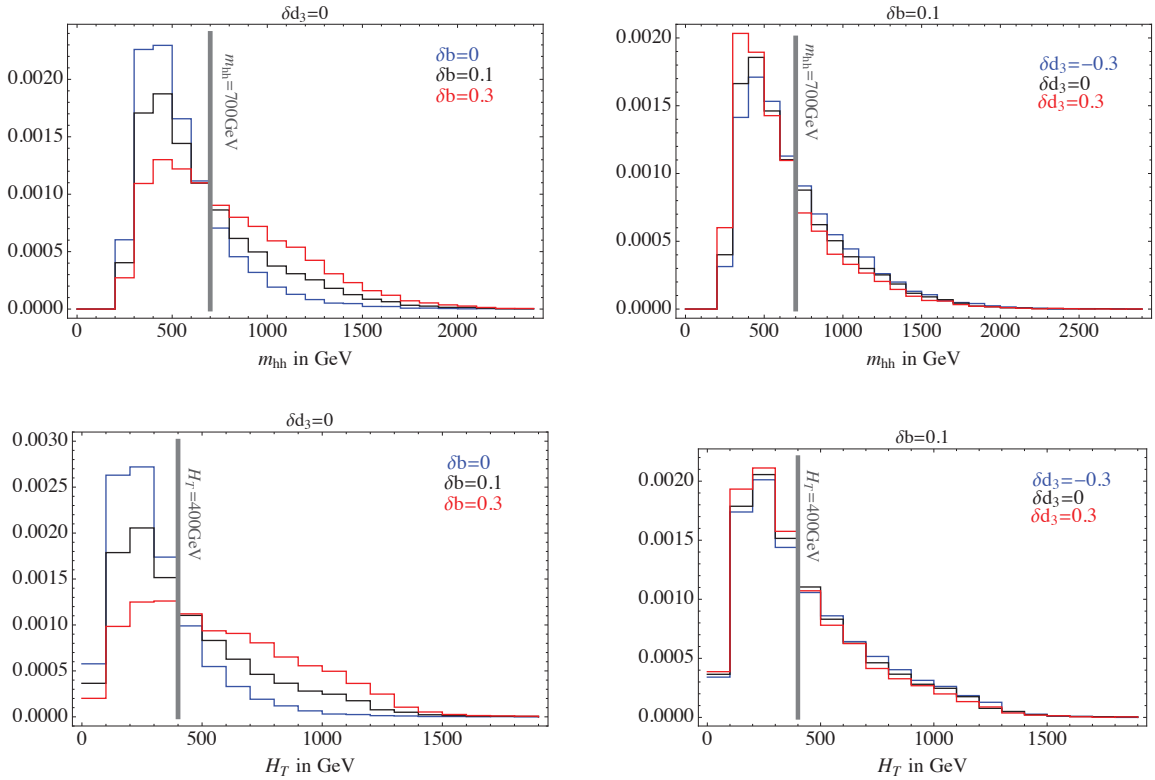


Figure 2.2: Normalized differential cross sections $d\sigma/dm_{hh}$ and $d\sigma/dH_T$ for $e^+e^- \to \nu\bar{\nu}hh$ at the Compact Linear Collider, with $\sqrt{s} = 3\,\text{TeV}$ after the identification cuts, for several values of δ_b and δ_{d_3}. Plot is taken from [13].

For the purpose of measuring g_{VVHH} and g_{HHH} via double Higgs production, a in the equation 2.29 can be treated as a constant.

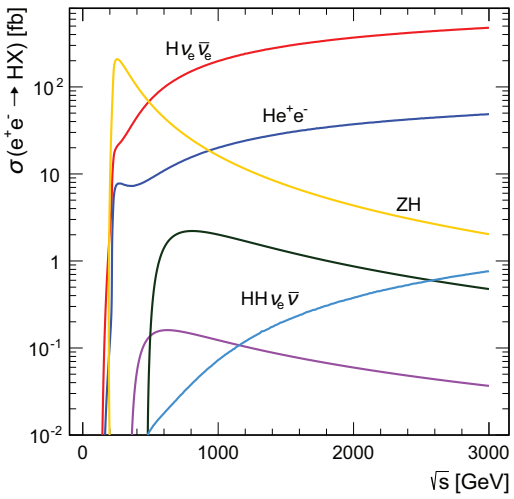


Figure 2.3: Cross section as a function of centre-of-mass energy for the Higgs production processes at an electron-positron collider for a Higgs mass of 126GeV. The values shown correspond to unpolarised beams and do not include the effect of beamstrahlung. Plot is taken from [15].

2.10 Tau pair polarisation correlations as signature of Higgs boson

For many theories beyond the Standard Model, a common feature is that the coupling of the Higgs particle to leptons increases with the increase of the lepton mass. Unlike Z and γ vector bosons which couples to leptons equally, the $H\tau^+\tau^-$ coupling would dominate. Therefore, if an experiment observes the breaking of the lepton universality by favouring $\tau^+\tau^-$ events, it could indicates the existence of a scalar Higgs. When such a breaking is observed, a helicity correlation test can be used to show that the $\tau^+\tau^-$ pair is from a scalar boson or a vector boson. In particular, the polarisation correlations of tau leptons are different for $H \to \tau^+\tau^-$ and $Z \to \tau^+\tau^-$, as scalar Higgs decays to $\tau^+_L\tau^-_L$ or $\tau^+_R\tau^-_R$ and Z decays to $\tau^+_L\tau^-_R$ or $\tau^+_R\tau^-_L$, where L, R denotes the tau lepton helicities.

Tau pair polarisation correlations can be studied using various tay decay modes. Here reference [72] is followed and $\tau^- \to \pi^-\nu_\tau$ decay mode is used as the example. The boson

decay can be represented as:

$$X \to \tau_\alpha^+ \tau_\beta^- \to \pi^+\pi^- + \nu's, \tag{2.30}$$

where X is either H or Z. α, β are the helicities, L or R. In the collinear limit where $m_\tau^2/m_X^2 \ll 1$, the appropriate kinematic variables are the energy fractions:

$$\bar{z} = \frac{E_{\pi^+}}{E_{\tau^+}},\ z = \frac{E_{\pi^-}}{E_{\tau^-}}. \tag{2.31}$$

For a single tau decay, the collinear distribution can be written as:

$$\frac{1}{\Gamma_{Ptau}}\frac{d\Gamma}{dz} = Br_{\pi^-} f(\tau_\alpha^- \to \pi^-; z), \tag{2.32}$$

where Br_{π^-} is the branching fraction of $\tau^- \to \pi^-\nu_\tau$. The form f can be obtained from literature [16]:

$$f(\tau_\alpha^- \to \pi^-; z) = 1 + P_\alpha(2z - 1), \tag{2.33}$$

where $P_L = -1$ and $P_R = +1$. For the tau pair decay, the collinear distribution is of form:

$$\frac{d^2N(X \to \tau^+\tau^- \to \pi^+\pi^- + \nu's)}{dz\, d\bar{z}} = Br_{\pi^-}^2 \sum_{\alpha,\beta} C_{\alpha\beta}^X f(\tau_\alpha^- \to \pi^-; z) f\left(\tau_\beta^+ \to \pi^+; \bar{z}\right), \tag{2.34}$$

where the only non-zero correlation coefficients $C_{\alpha\beta}$ for the party-conserving $H \to \tau^+\tau^-$ are:

$$C_{LL}^H = C_{RR}^H = \frac{1}{2}, \tag{2.35}$$

and for $Z \to \tau^+\tau^-$ the coefficients are

$$C_{LR}^Z = \frac{1}{2}\left(1 - P_\tau\right),\ C_{RL}^Z = \frac{1}{2}\left(1 + P_\tau\right), \tag{2.36}$$

where tau polarisation, P_τ, is due to the non-parity-conserving tau decay in the SM. For Z:

$$P_\tau = \frac{-2va}{v^2 + a^2}, \tag{2.37}$$

$$v = -\frac{1}{2} + \sin^2\theta_W, \tag{2.38}$$

$$a = -\frac{1}{2}, \tag{2.39}$$

where v and a are the vector and axial-vector $Z\tau^+\tau^-$ couplings. Figure 2.4 shows the two-dimensional distribution of $Z \to \tau^+\tau^-$ and $H \to \tau^+\tau^-$ as a function of the energy fractions z and $\bar{z}$. The difference between Z and H decaying to a tau pair is clear. Therefore, if an excess of $\tau^+\tau^-$ events over e^+e^- or $\mu^+\mu^-$ events is observed, it can be easily identified from the tau pair polarisation correlation, where the excess is from H decay or Z decay.

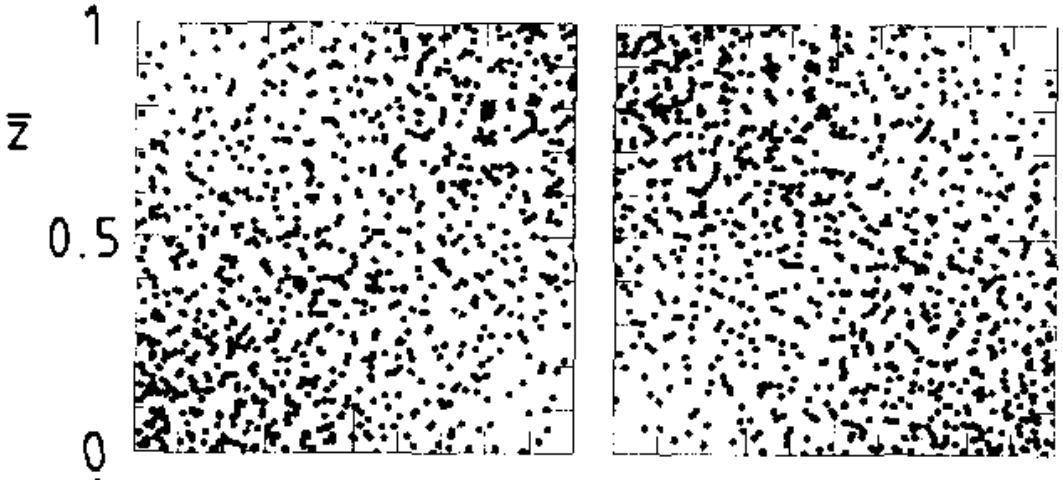


Figure 2.4: Two-dimensional distribution of $Z \to \tau^+\tau^-$ on the left, and $H \to \tau^+\tau^-$ on the right as a function of the energy fractions $\bar{z} = E_{\pi^+}/E_{\tau^+}$, $z = E_{\pi^-}/E_{\tau^-}$. The figure is adapted from reference [16].

Chapter 3

Detector and Physics at Future Linear Colliders

"ILC will be built next year"
— Mysterious person

Since the discovery of a particle consistent with being the SM Higgs boson in LHC at 2012 [1, 17], our understanding of Standard Model has improved greatly. Yet limited by the underlying QCD interaction from proton-anti-proton collision, one has great difficulty to measure the properties of the Higgs precisely. Next generation electron-positron linear collider could hopefully make precision measurements of the Higgs sector and the Top quark sector [18, 19].

3.1 ILC

Two leading candidates for next generation electron-positron linear colliders are the International Linear Collider (ILC) [18], and the Compact Linear Collider (CLIC) [19]. The ILC is a high-luminosity electron-positron linear collider with centre-of-mass energy from 200 GeV up to 1 TeV. The machine would be build at different stages. The first stage would have a centre-of-mass energy of 250/350 GeV. The second stage would be 500 GeV with a possible upgrade to 1TeV. Thirty years of development leads to the technical design report in 2013 [20]. A layout of the collider complex is shown in figure 3.1.

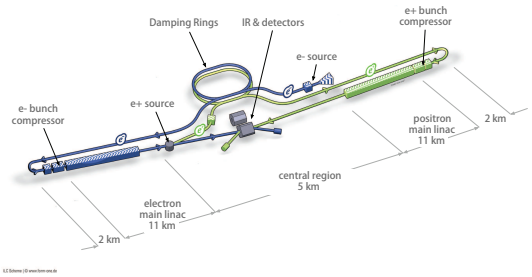


Figure 3.1: A layout of the International Linear Collider complex, taken from [20].

3.2 CLIC

The other potential next generation electron-positron linear collider, the Compact Linear Collider (CLIC), has a higher reach of the centre-of-mass energy up to 3 TeV. The CLIC is designed as a staged machine. The first stage, with centre-of-mass energy 380 GeV, is a compromise of precision measurement between both top quark and Higgs physics. The final stage 3 TeV is motivated by the physics reach of detecting new physics, and measurement rare decays of Higgs. The second stage is around 1.4 TeV, which bridges between the first stage and the final stage. A layout of the CLIC complex is shown in figure 3.2. Due to the similarities of the two linear collider programs, the development with CLIC detector concepts start with ILC detector concepts. CLIC_ILD and CLIC_SiD are developed based on ILD and SiD.

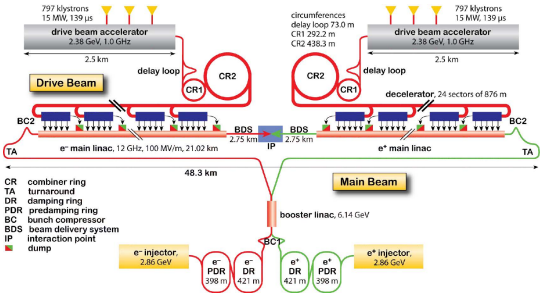


Figure 3.2: A layout of the Compact Linear Collider at 3 TeV, taken from [21].

3.3 Physics at future linear colliders

The physics program for the CLIC and the ILC, which is a driving force for the detector design, have some common goals. ILC has a reach of centre-of-mass from 200 GeV to 1 TeV, whilst CLIC can reach from 350 GeV to 3 TeV. Both machines are capable of

precision higgs coupling measurements, top mass and coupling measurements, search for new physics such as supersymmertry particles. The ILC can also operate at low energy to be a Z and a H factory for ultra precision Z mass and H mass measurement. CLIC, however, has the advantage of higher energy reach, which allows measurements of rare events, such as higgs triple self-couplings and quartic couplings. The discovery potential for the CLIC is also greater.

3.4 Impact of physics requirements on the detector design

3.4.1 Jet energy resolution requirements on the detector design

The physics goal of jet energy resolution at the ILC and the CLIC is to separate W and Z hadronic decays via reconstruction their di-jet masses [19, 20]. This translate to a requirement of 3.5-5% of the energy resolution. This level of precision is unlikely to be achieved with a traditional calorimetry design. A traditional energy flow to calorimetry measures jet energies as a sum of the energy in the calorimeters. The jet energy resolution is parameterised by

$$\frac{\sigma_E}{E} = \frac{\alpha}{\sqrt{E(GeV)}} \oplus \beta \tag{3.1}$$

The stochastic term a is typically greater than 60% and b is order of a few percents. For the jet energy resolution of 3.5%, a should be less than 30% which is unlikely to be achieved by a traditional calorimeter. On the contrary, particle flow approach has demonstrated it ability to reach the goal [22, 23].

In a typical jet, using measurements on the particle composition from the LEP [24, 25], about 62% of the jet energy is from charged particles, 27% from photons, 10% from long-lived neutral hadrons, and 1.5% from neutrinos. In a traditional approach to calorimetry, about 72% jet energy is measured in the electromagnetic (ECAL) and the hadronic (HCAL) calorimeters combines. The jet energy resolution is thus limited by the energy resolution of the hadronic calorimeters, which typically is $\gtrsim 55\%/\sqrt{E(GeV)}$.

The particle flow approach to calorimetry improves the jet energy resolution by fully reconstructing the four momenta of all visible particles in the detector. The jet energy is the sum of the individual particles' energies, where the energy of the charge particles are measured in the tracking detectors, and the energy of neutral particles are measured in calorimeters. In this manner, the hadronic calorimeter only measures about 10% of the energy, which would greatly improves the overall energy measurement. Assuming 30% of the jet energy, photon energy, is measured with $\sigma_E/E = 15\%/\sqrt{E(GeV)}$, and 10% of the jet energy , which are hadrons, measured with $\sigma_E/E = 55\%/\sqrt{E(GeV)}$, a jet energy of $\sigma_E/E = 19\%/\sqrt{E(GeV)}$ can be obtained. This satisfies the jet energy resolution for separating W and Z hadronic decays. In reality, this level of performance is unattainable due to incorrect association of energy deposits to particles. At jet energy beyond tens of GeVs, the "confusion" rather than the intrinsic detector performance limits the particle flow performance. This has stringent requirements in the ECAL and the HCAL design.

The particle flow calorimetry requires to fully reconstruct particles and associate calorimeter hits to tracks. This is demanding for the software design and the detector design. The software details of the PandoraPFA, which is a successful particle flow implementation, are described in section 4.4. The detector needs to be highly granular for the excellent spatial resolution to be able to correctly associate calorimeter hits to the inner detector tracks. This forms the main motivation in the calorimeter design.

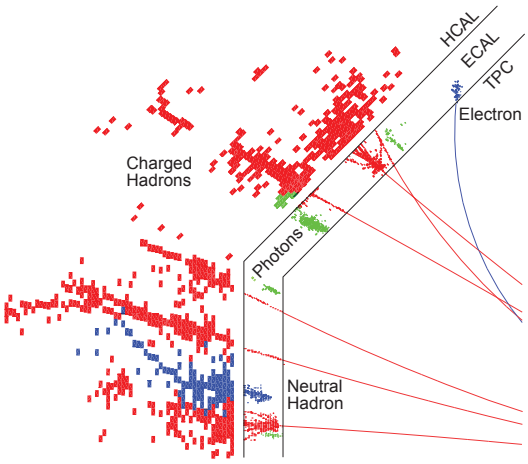


Figure 3.3: A typical topology of a 250 GeV jet, simulated with CLIC_ILD detector concept, taken from [23].

Figure 3.3 is a typical topology of a 250 GeV jet, simulated with CLIC_ILD detector concept. The particles with the calorimeter hits and tracks are labelled with different colours. Clusters of calorimeter hits in the highly granular ECAL and HCAL are

associated with tracks from the inner tracking detector, TPC. Photons are identified using the characteristics longitudinal and transverse electromagnetic shower profiles. Hadronic showers are separated from electromagnetic showers due to the small transverse spread of the electromagnetic shower. Therefore, the inner tracking detector should be highly efficient and has very little material. For the calorimeter, the ECAL and HCAL, both should be highly granular. The material of the calorimeter should be dense and has a large ratio of interaction length to radiation length.

3.4.2 Other requirements on the detector design

Other physics requirement for the detectors for the ILC and the CLIC are summarised from [19, 20].

The requirement of tracking momentum resolution is driven by the Higgs boson mass resolution via Higgsstrahlung process, $e^-e^+ \to ZH$. Higgs mass can be reconstructed precisely as the recoil mass against the Z momenta, which is obtained via $Z \to \mu^+\mu^-$. For the ILC operating at $\sqrt{s} = 250\,\mathrm{GeV}$, the momentum resolution needs to be $\sigma_{p_T}/{p_T}^2 \lesssim 5 \cdot 10^{-5}\mathrm{GeV}^{-1}$. For the CLIC at high $\sqrt{s}$, the momentum resolution needs to be $\sigma_{p_T}/{p_T}^2 \lesssim 2 \cdot 10^{-5}\mathrm{GeV}^{-1}$.

The performance requirement of the vertex detector is determined by efficient b-quark and c-quark tagging. The ability to identify secondary vertices and tracks, which are not originated from the interaction point, is the prerequisite for the flavour tagging. The impact parameter resolution can be written as

$$\sigma_{d_0}^2 = a^2 + \frac{b^2}{p^2 \sin^2(\theta)} \tag{3.2}$$

where a is related to the point resolution and b is related to multiple scattering. The requirements for both the ILC and the CLIC detectors are $a \lesssim 5\mu m$ and $b \lesssim 15\mu m\mathrm{GeV}$

The lepton identification are should be over 95% for effective lepton tagging. The forward converge of the detector should be down to a very low angle. This is more critical for the CLIC as particles are boosted at high $\sqrt{s}$.

3.5 International Large Detector

Two detectors concepts have been designed for the ILC to deliver the physics program. Precision tests for Standard Model requires an excellent jet energy resolution and di-jet mass reconstruction. Particle Flow Algorithms based event reconstruction meets the requirement and motivates the detector designs. For the best performance of the PFA, high granular calorimeter systems and highly efficient tracking systems are designed. The requirement to separate W and Z bosons in di-jet final states requires a jet energy resolution below 3.5%. The momentum resolution of 5×10^{-5} GeV is motivated by the Higgs boson recoil reconstruction in the Higgs-strahlung.

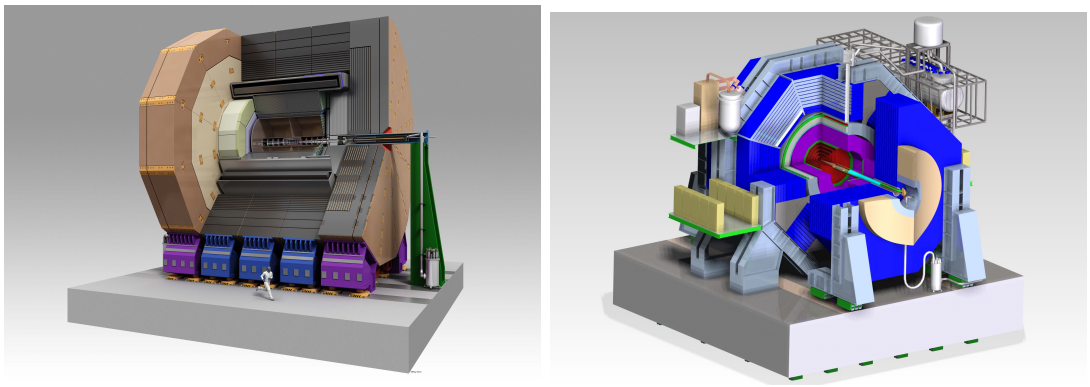
Motivation for two detectors is to have multiple independent measurements within one collider for cross-checking, complementary measurements and competition between collaborations. Two detectors are both general purpose detectors. Silicon Detector, SiD, is a compact detector with a large magnetic filed of 5 T. It uses silicon tracking modules. The International Large Detector, ILD, is a larger detector with a time projection chamber as the main tracking unit. Both detectors have high granular calorimeters optimised for the particle flow. A view of both detector concepts can be seen in figure 3.4

The International Large Detector, ILD, is a detector concept at the International Linear Collider, ILC. The ILD detector concept has been optimised in the view of the particle flow techniques. Particle flow approach to event reconstruction has shown to deliver the best possible jet reconstruction with proof-of-principle implementation such as PandoraPFAchapter 4. Each individual particles are reconstructed with the particle flow approach. For charged particles, calorimeter hits are associated with the tracks. The measurement of charged particle relies on the excellent tracking system resolution. Neutral particle reconstruction require fine spatial resolution of the calorimeters. These form the requirements for the detector designs and optimisations.

The particle flow paradigm requires topological information for individual particle reconstruction. The sub-detector systems need to have the spatial resolution to separate charged particles from neutral particles. The result is a highly granular calorimeters with a central tracking system with excellent momentum resolution. Longitudinal cross section of top quadrant of the ILD detector concept, taken from [18], is shown in figure 3.10a From interaction point (IP) outwards, there is a tracking system compromising a large time projection chamber (TPC) augmented with silicon tungsten layer, highly granular electromagnetic calorimeters (ECAL) and hadronic calorimeters (HCAL), muon

chambers, forward calorimeters (FCAL), magnetic coils and iron yokes. Numbers are in units of mm.

This section will describe the sub-systems of the ILD detector concept in the ILD technical design report [], the ILD_o1_v05 option in MOKKA simulation. This detector concept has been used in studies described in subsequent chapters. The ILD detector concept has been optimised and documented in previous documents, such as the letter of intent []. The CLIC_ILD detector concept for the CLIC in the conceptual design report [19] is a modified version of the ILD, adapted to the CLIC colliding environment. The differences between ILD and CLIC_ILD can be seen in figure 3.10 and table 3.1, and are addressed in the discussion below.



(a) **(b)**

Figure 3.4: Figure 3.4a and figure 3.4b show the view of the International Large Detector and the Silicon Detector for the International Linear Collider, taken from [20].

3.6 Overview of ILD sub-detectors

The ILD detector concept is designed as a general purpose detector. Closest to the interaction points are the precision a vertex detector and a tracking system. The tracking system consists of silicon tracking with a time projection chamber. Surrounding the tracking system is a high granular calorimeter system. The outer solenoid provides a magnetic field of 3.5 T. The most outer iron return yoke acts as a muon calorimeter.

3.6.1 Vertex Detector

The pixel-vertex detector(VTX) needs to be close to the interaction point to reconstruct secondary vertex. As the TPC is the main tracking detector, the VTX mainly measures the impact parameter of tracks. The structure is three double layers with a barrel geometry. Double layer lowers the material budget and improves the impact parameter measurements. The first double layer is half length of the other two to avoid the high occupancy region of direct low omentulum hits from the incoherent pair background.

3.6.2 Tracking Detectors

The hybrid tracking system is consists of a large volume time projection chamber (TPC), a Silicon Inner Tracker (SIT), a Silicon External Tracker (SET) in the barrel region, a end cap tracking component (ETD) behind the endplate of the TPC, and a silicon forward tracker (ETD) in the forward region. The SIT, SET, and ETD are made up two single-sided strip layers tilted by a small angle. The ETD is a system of two silicon-pixel disks and five silicon-strip disks. The silicon envelope tracking system and the TPC are shown in figure 3.5.

The main part of the tracking system, the TPC, can measure a large number of three dimensional spatial points. Continuous tracking allows precise reconstruction of non-pointing tracks. The TPC is optimised for point resolution and minimum material, as required for the best calorimeter and particle flow performance.

The barrel silicon trackers improve the overall momentum resolution. They provide additional high precision space points and additional redundancy between the TPC, the VTX, and the calorimeters. The ETD provide the low angle coverage which is not covered by the TPC.

3.6.3 Electromagnetic Calorimeter

The Silicon-Tungsten sampling electromagnetic calorimeters in the ILD consist of a nearly cylindrical barrel and two end cap systems, optimised for particle flow. The ECAL measures photon energies and separates photons from other particles. The fine granular ECAL also sits inside the HCAL, which hosts the first part of the hadronic showers and greatly helps to separate hadronic showers.

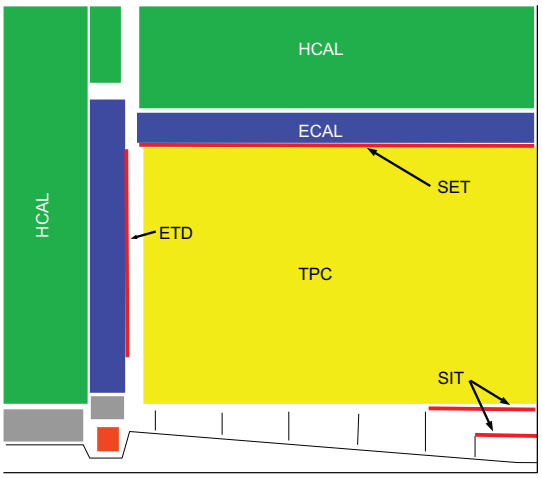


Figure 3.5: A top quadrant view of the ILD silicon envelope system, SIT, SET, ETD, and ETD as included in MOKKA full simulation, adapted from the figure in [20].

The particle flow paradigm has a large impact on the ECAL design with many requirements. In addition for the ECAL to measure and separate photons, it also needs to reconstruct detail shower profiles to separate electromagnetic showers from hadronic showers, as approximately 50% of hadronic showers starts in the ECAL. These requirements can be fulfilled with an excellent three dimensional granular ECAL.

From test beam data and simulation studies, a sampling calorimeter with longitudinal and transverse segmentation below one Molière radius and below one radiation length at the front the calorimeter is needed. The most compact design is realised with Tungsten as absorber material and silicon pad diodes as active material. A cross section of the ECAL is shown in figure 3.6. Tungsten is a dense material with a large ratio of interaction length to radiation length. This helps to separate electromagnetic showers from hadronic showers by making electromagnetic showers transversely narrow. Silicon pad size of 5.1 by 5.1 mm cover large areas. They are simple and reliable to operate. The choice of thin silicon layers offers a great spatial resolution at a cost of the energy resolution in favour of the particle flow.

The longitudinal segregation is a compromise between the cost and the performance. The total 30 layers, which is about 20 cm, provides about 24 radiation lengths. The first 20 layers use 2.1 mm thick absorber plates, which is twice finer sampling than the last 10 layers with 4.2mmthick absorber plates. The test beam data with electron shows the energy resolution of the ECAL concept to be $16.6/\sqrt{E(\ GeV)} \oplus 1.1\%$, which is compatible with the values assumed for the full ILD detector simulation.

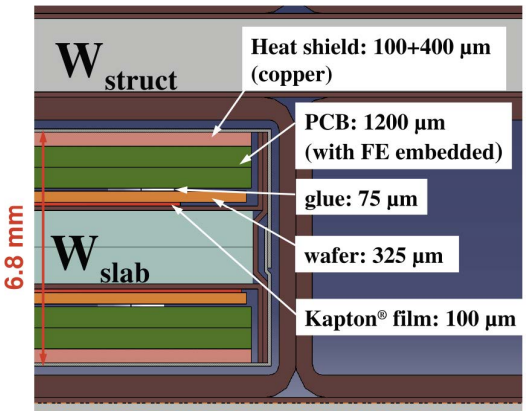


Figure 3.6: A cross section through electromagnetic calorimeter layers, taken from [20].

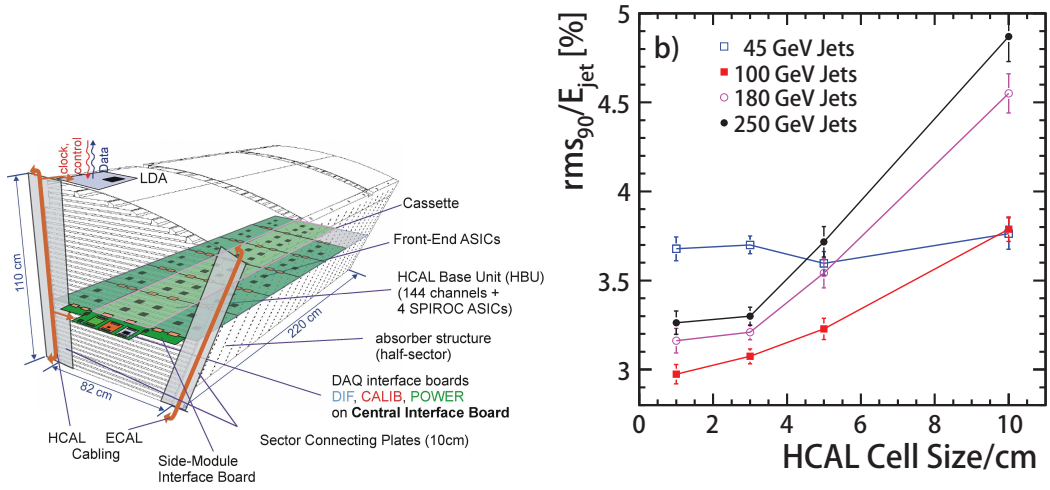
3.6.4 Hadronic Calorimeter

The requirements of sampling hadronic calorimeter is, again, driven by the need of the particle flow. The need of three dimensional granularity in transverse and logistically direction is satisfied by a sampling calorimeter.

The principle role of the HCAL is to separate neutral hadron showers from other particles, and to measure neutral hadron energies. The neutral hadron contribution of the jet energy is around 10% on average. A moderate fine granular HCAL is a good balance between cost and performance. The chosen layout is 48 longitudinal layers with 3 by 3cmscintillator tiles, using an analogue read out system. The layout of a technological prototype, the "EUDET prototype" is shown in figure 3.7a [26].

The longitudinal system provide about 6 radiation lengths including the ECAL, which is sufficient to contain the hadronic showers. The transverse cell sizes has been optimised for the best jet energy resolution. It is found that no substantial gain below 3 cm and performance degradation above 3 cm. Hence 3 cm cell size is chosen for the HCAL. The jet energy resolution as a function of HCAL scintillator cell size with different jet energies is shown in figure 3.7b.

For the absorber material, stainless steel is chosen for mechanical and calorimetric reasons. Steel allows a self-supporting structure without auxiliary supports. Also steel has a moderate ratio of interaction length to radiation length.



(a) (b)

Figure 3.7: Figure 3.7a shows the schematic view of a CALICE AHCAL technological prototype module. Figure 3.7b shows the jet energy resolution as a function of the hadronic calorimeter scintillator cell sizes, with different energies. Both figures are taken from [20].

3.6.5 Solenoid

A large superconducting solenoid outside the calorimeters produces a nominal 3.5 T magnetic field.

3.6.6 Yoke and Muon system

An iron yoke instrumented with scintillator strips active layers returns the magnetic flux, and acts as a muon detector and tail catcher calorimeter at the same time. The layout is shown in figure 3.8. The agreed maximum magnetic field at 15 m radial distance from the detector is 50 Gauss to ensure safety [27]. A highly efficient muon detector is provided by the 3 by 3 cm scintillator strips. As a tail catcher calorimetry, the first layer of the muon detector, catches the energy leakage from the HCAL and the ECAL. It has been shown a 10% improvement of single particle energy resolution is possible with the tail catcher [28].

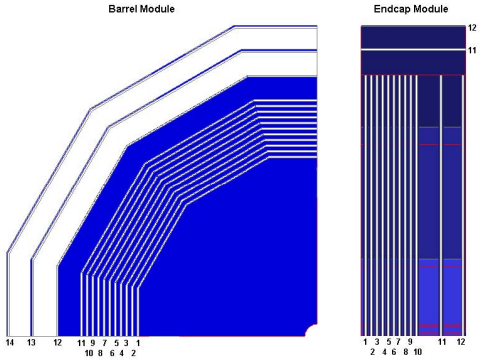


Figure 3.8: Sensitive Layers of the ILD muon system, taken from [20].

3.6.7 Very Forward Calorimeters

The forward region detectors provide luminosity measurements and forward coverage of calorimeters. A system of precision and radiation resistent calorimeters are required. The luminosity calorimeter counts Bhabha scattering to measure the luminosity to precision of 10^{-3} at 500 GeV centre-of-mass energy. The beam calorimeter (BeamCAL) extend the forward coverage, which are hit by many beamstrahlung pairs after each bunch corssing. BeamCAL estimates a bunch-by-bunch luminosity. An additional hadron calorimeter, LHCAL, at the forward region extends the angular coverage of the HCAL to that of the LumiCAL. Electron tagging is possible with the very forward calorimeters [29], which aids event reconstruction at high centre-of-mass energy.

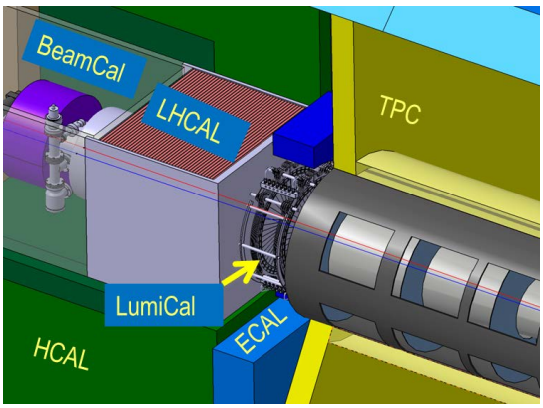


Figure 3.9: The forward calorimeters of the ILD, taken from [20]. LumiCAL, BeamCAL, and LHCAL are the luminosity calorimeter, beam calorimeter, and forward hadronic calorimeter.

3.7 CLIC versus ILC

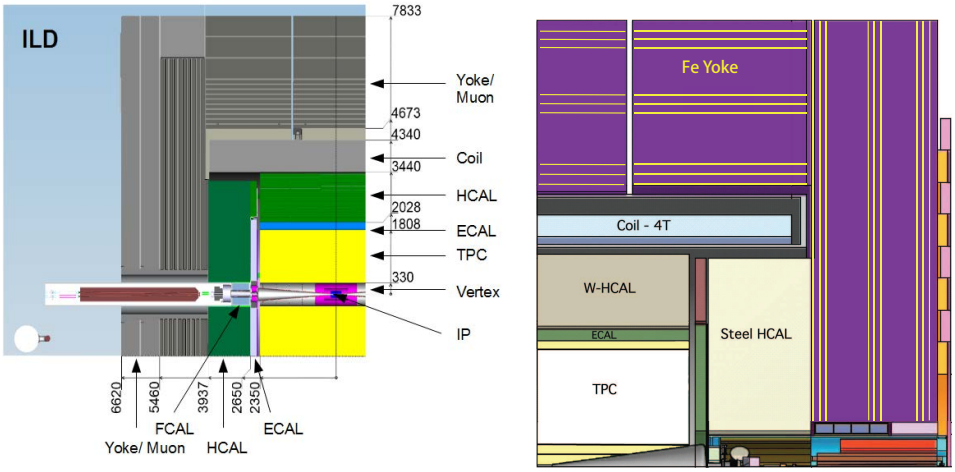
The two main differences between CLIC and ILC are the high centre-of-mass energy and the high bunch charge density leading to significant beam related backgrounds. Within a bunch train, there is 0.5 ns between bunch crossings at CLIC. There are two main sources of beam induced background: incoherent electron pairs from photon (real or virtual) interactions with individual particles of the other beam, and interactions of two photons from the colliding beams. These differences leads to a modification in the detector design and the reconstruction software for the CLIC. A comparison of CLIC_ILD and ILD longitudinal cross sections can be seen in figure 3.10. A comparison of key parameters of the ILD and CLIC_ILD detector concepts is shown in table 3.1.

Concept	ILD	CLIC_ILD
Tracker	TPC/Silicon	TPC/Silicon
Solenoid Field (T)	3.5	4
Solenoid Field Bore (m)	3.3	3.4
Solenoid Length (m)	8.0	8.3
VTX Inner Radius (mm)	16	31
ECAL r_{min} (m)	1.8	1.8
ECAL Δr (mm)	172	172
HCAL Absorber B / E	Fe	Fe / W
HCAL Interaction Length	5.5	7.5
Overall Height (m)	14.0	14.0
Overall Length (m)	13.2	12.8

Table 3.1: A comparison of key parameters of the ILD and CLIC_ILD detector concepts. ECAL r_{min} is the smallest distance from the calorimeter to the main detector axis. HCAL Absorber B / E indicates the absorber material for the barrel (B) and the endcap (E). The table is adapted from [19].

3.7.1 CLIC_ILD versus ILD

There are two detector concepts studied in the CLIC conceptual design report [19], CLIC_ILD and CLIC_SiD. The CLIC_ILD detector concept is based on the ILD design. They share similarities due to similar physics motivations. Only differences are highlighted here.



(a) **(b)**

Figure 3.10: Figure 3.10a and Figure 3.10a shows longitudinal cross section of top quadrant of the ILD and the CLIC_ILD detector concepts, taken from [20] and [19] respectively. From interaction point (IP) outwards, there is a tracking system compromising a large time projection chamber (TPC) augmented with silicon tungsten layer, highly granular electromagnetic calorimeters (ECAL) and hadronic calorimeters (HCAL), muon chambers, forward calorimeters (FCAL), magnetic coils and iron yokes. Numbers are in units of mm.

For the vertex detector, the first layer is moved outwards by 15 mm due to a larger high occupancy region with higher centre-of-mass energy. The detector is also required to provided time stamping at nanoseconds level, which would be a different detector.

For the tracking detector, the same silicon-TPC hybrid structure is used. The outer silicon tracking system is more important at the CLIC to achieve a high momentum resolution at high centre-of-mass energy, as it is challenging using a TPC to sperate two tracks in high energy jets and to identify events in the collection of 312 bunch crossings in 156 ns. The solid angle coverage of the tracking detector is $12^\circ \lesssim \theta \lesssim 168^\circ$

The same ECAL from the ILD is assumed, as the requirements of a CLIC detector are satisfied. The increased centre-of-mass energy results in extra energy leakage. But only a small fraction of particles are affected and the leakage is controlled by the HCAL.

For the HCAL, extra layers are added to contain the hadronic shower at high energy. The increased thickeners is justified by the simulation studies, where the jet energy resolution degrades quickly for a thinner HCAL. To sustain the same inner bore radius, a more dense material, Tungsten, is chosen as the absorber material in the HCAL barrel.

The magnetic filed is increased to 4 T for a better performance at a high centre-of-mass energy. Due to the different magnetic field strength, the iron yoke thickness increase to 230 cm.

The CLIC_ILD adopted a similar very forward calorimetry system as that of the ILD. Dimension of the elements are changed due to a difference in the crossing angle (20 mrad for CLIC and 14 mrad for ILC). A comparison of LumiCAL and BeamCAL at ILD and CLIC_ILD is shown in table 3.2.

Modifications to the design due to CLIC 3 TeV centre-mass-of energy can be found in [19].

		ILD	CLIC_ILD
LumiCAL	geometrical acceptance (mrad)	31 - 77	38 - 110
	fiducial acceptance (mrad)	41 - 67	44 - 80
	z (start) (mm)	2450	2654
	number of layers (W + Si)	30	40
BeamCAL	geometrical acceptance (mrad)	5 - 40	10 - 40
	z (start) (mm)	3600	3281
	number of layers (W + sensor)	30	40
	graphite layer thickness (mm)	100	100

Table 3.2: Comparison of LumiCAL and BeamCAL at ILD and CLIC_ILD. The table is adapted from [19].

Chapter 4

Simulation and Reconstruction

"How to open a pandora box?"
— A wise Chinese

Automated analysis is the only way to deal with the vast amount of data generated in the high energy physics. An analysis involves an event reconstruction and using software to extract information of the event. Reconstruction software and analyses software have the same purpose to extract information. Hence they are discussed together in this chapter. Since the work presented in this document is on future colliders, simulation and monte carlo method is used throughout the document and presented in this chapter as well.

In previous chapters, overviews of the theory and the future linear collider experiments have been described. In this chapter, the simulation and event reconstruction chain are discussed, followed by discussion on common analyses software. Simulation and reconstruction of events for the future Linear Colliders, ILC and CLIC, share common software framework. Therefore shared reconstruction is discussed first, and the CLIC specific issues are highlighted afterwards. The event reconstruction with emphasis on the PandoraPFA event reconstruction, which is the framework for the photon reconstruction algorithms in chapter 4. Lastly multivariate analysis is presented in a separate section because of its complexity. An overview of the multivariate with focus on boosted decision tree is provided.

4.1 Monte Carlo event generation

Monte Carlo (MC) event generation is the first step for the simulated study. Most events, electron-positron interaction, are generated with WHIZARD software, [30, 31], with no polarisation of the electron and positrons. Some simple events are generated with HEPEVT. PYTHIA [32] is used to describe parton showering, hadronisation and fragmentation. The parameters for PYTHIA are tuned to OPAL data from the LEP [33]. TAUOLA [34] debrides the tau lepton decay with correct spin correlations of the day products. The Initial State Radiation (ISR) is simulated in WHIZARD with the ISR photons being collinear with the beam direction. The Final State Radiation (FSR) is simulated with default parameters in PYTHIA.

4.2 Event Simulation

The event simulation software is GEANT4 [35], and the detector geometry description is provided by MOKKA [36]. QGSP_BERT physics list is used to describe the hadronic showers decay in the detector.

4.3 Event Reconstruction

Reconstruction software runs in Marlin framework [37], as a part of the iLCSoft. Event reconstruction contains following steps: digitisation of simulated calorimeter hits, reconstruction of tracks in the tracking system using pattern recognition algorithms, and particle flow objects (PFOs) reconstruction with PandoraPFA [22, 23]. Details of the reconstruction can be found in [18, 19]. Particle flow reconstruction via PandoraPFA will be discussed in details, which provides the software framework for the photon reconstructions in PandoraPFA in chapter 5.

4.4 PandoraPFA reconstruction

Tradition calorimetric approach is unable to meet the mass and energy resolution requirements for future linear colliders. The particle flow approach with PandoraPFA has a proof-of-principle demonstration of its capability to reach required resolution. The

particle flow approach also put stringent requirements on the detector design, which is described in section 3.4.1. By associating calorimeter hits to the tracks, around 60% of the jet energy from charged particles is measured by the tracker, which has a much better resolution than the calorimeter. Small cell sizes of the calorimeters are required to identify hits from different particles. The traditional sum of calorimeter cell energies is replaced by particle flow reconstruction algorithms, a complex pattern recognition problem. The PandoraPFA algorithm has been developed and used in the ILC and CLIC simulation studies.

Developed with the ILD detector concept, PandoraPFA has been adapted to the CLIC condition and shows its ability to deliver required energy resolutions [19]. Recent the code base of the PandoraPFA has been restructured. The core base codes for basic object and memory managements are factorised in the Pandora C++ Software Development Kit [38]. There are over 60 linear collider specific reconstruction algorithms, each aims to address a particular topological in the reconstruction.

In the subsequent paragraphs, the main steps in the PandoraPFA reconstructions are described. The details of the reconstruction can be found in [22, 23, 38].

Inputs of PandoraPFA are digitised calorimeter hits and reconstructed tracks. The output are reconstructed particles with four-momenta, Particle Flow Objects (PFOs).

4.4.1 Track selection

Tracks from the tracking system are selected based on their topological properties, how likely they are from physical processes, and whether they are consistent with the tracker resolution. Only tracks passed the selection are used for the subsequent reconstruction.

Special topologies of tracks are identified, such as when a neutral particle decays or converts into a pair of charged tracks, leaving a “V0” shape tracks. This is identified by searching for a pair of tracks originated from a single point. Another topology is “kinks” , when a charged particle decays to a single charged particles with neutral particles. The last special topology is “prongs”, when a charged particles decays to multiple charged particles. This information are stored and passed on to the subsequent reconstruction, along side with helical track fit (using last 50 reconstructed tracker hits) and the track projection to the front of the ECAL.

4.4.2 Calorimeter selection

The information of a calorimeter hit is its position, its layer in the calorimeter and its energy response from the calorimeter digitiser.

Calorimeter hits are selected based on a series of criterion. The selected hits need to have energies above certain thresholds, measured in minimum ionising particle (MIP) equivalent, or measured in directly converted energy. Similar to tracks, only calorimeter passed the selection are used in later steps.

Geometry information and likelihood of the hit originated from a minimum ionising particle (MIP) are calculated.

Isolated hits, often originated from low energy neutrons in a hadronic shower, are difficult to associate to the correct hadronic shower. They are identified and not used during the clustering stage. They participate the reconstruction during the last step, particle flow object (PFO) creation.

4.4.3 Cone Clusters Algorithm

Before discussing the rest of the PandoraPFA reconstruction, it is necessary to introduce the cone based clustering algorithm, which is widely used in the calorimeter in PandoraPFA. The clustering algorithm produces basic working objects, Clusters.

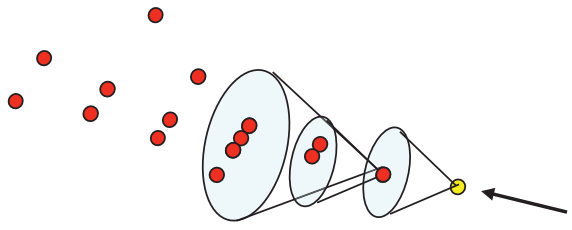


Figure 4.1: Illustration of the cone based clustering, taken from [39]

There are two main types of clustering algorithms: cone based and sequential combination (see section 4.6.2). The main clustering scheme PandoraPFA is cone based clustering to group calorimeter hits. Illustrated in figure 4.1, cone clustering has a specified opening angle of the seed hit. Because the direction of particle flows is largely unchanged from the originated particle, whether it is a electromagnetic shower, QCD radiation or hadronisation, these cone clusters have similar direction and energy to the

originated particle. Therefore it is applicable to use cone based clustering algorithms for building clusters.

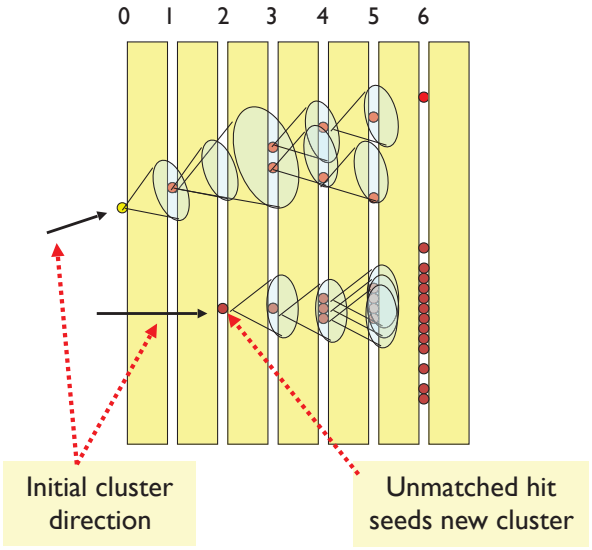


Figure 4.2: Illustration of the clustering algorithm in PandoraPFA, taken from [39]

The seed for the cone clustering is typically the projection of a energetic track to the front of the ECAL. A high energy calorimeter hit can also be used as a seed. A cone with a specified opening angle and depth will be formed around the seed. The four-momentum of calorimeter hits sum to the cone's four-momentum. The cone is built up from the inner layer of the ECAL to the outer layer. At each layer, possible associations with hits in previous layers and same layer are checked. If a hit is not associated, it is used to seed a new cluster. This process is illustrated in figure 4.2.

4.4.4 Particle Identification

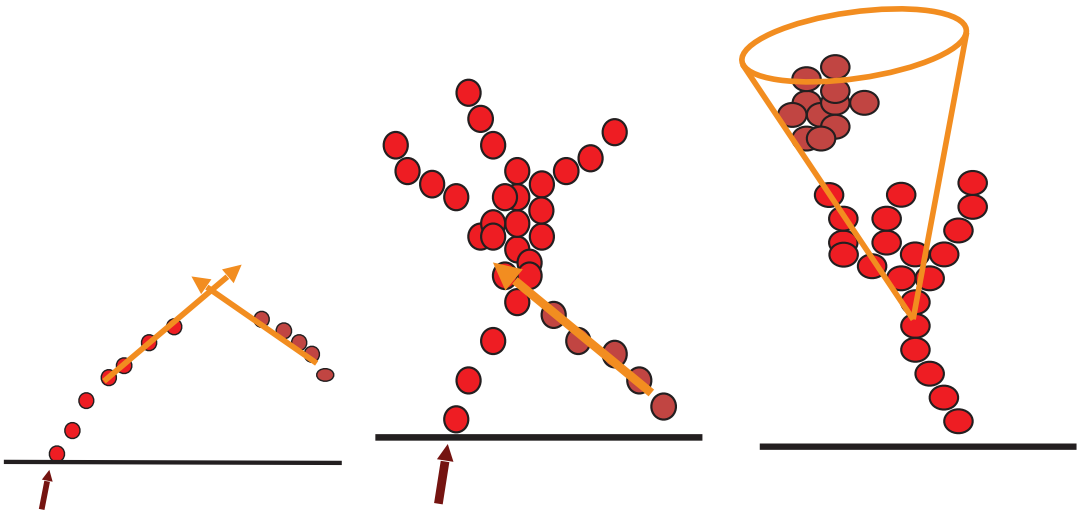
Dedicated particle identification algorithms aim to identify muons and photons before associating calorimeter hits to tracks. The details of the photon reconstruction algorithms and photon related algorithms are described in chapter 4. By removing the hits from muons and photons, the reconstruction of charged particles is improved as the pattern recognition problem is reduced with fewer hits. Identified muons and photons do not participate in the clustering and re-clustering stages, but re-entre the construction at the fragment removal stage (see section 4.4.9).

4.4.5 Clustering

The cone clustering algorithm described in section 4.4.3 is used to group calorimeter hits from innermost to outmost pseduo-layer. The output Clusters are further processed, merged or split based on their topological properties.

4.4.6 Topological cluster association

Initial clustering scheme is aggressive at splitting clusters. Small clusters are merged based on clear topological signatures. These merging signatures include combining track segments, connecting track segments with gaps, connecting track segment to a hadronic shower, and merging clusters when they are within close proximity. Some association algorithms are shown schematically in figure 4.3.



(a) (b) (c)

Figure 4.3: Examples of topological association in PandoraPFA. Figure 4.3a, figure 4.3b, and figure 4.3c show rules for looping tack segments, back-scattered tracks from hadronic showers, and cone association. In each case, the arrow indicates the tracks. The dark red dots represent the calorimeter hits in the associated cluster. The pink dots represent the calorimeter hits in the neutral cluster. The black line represents the front of the ECAL. Figures are taken from [39].

4.4.7 Track-cluster association

Clusters are associated to tracks, according to the proximity of the first layer of the cluster and the track projection to the front of the ECAL. The track and initial cluster directions are required to consistent, as well as a match between track momentum and cluster energy.

4.4.8 Re-clustering

The cluster association scheme work well for low energy (less than 50 GeV) jet. For a high energy jet, particles and the subsequent hadronic showers are more boosted and more likely to overlap each other. Therefore, it is important to re-cluster based on the compatibility of the cluster energy and the associated track momentum. A cluster may be split into two or more clusters. Two clusters maybe be re-clustered based on the track-cluster association. The split up clusters would attempt to be associated using topological association algorithms. The re-clustering scheme is applied iteratively to find a more correct clustering of calorimeter hits. A schematic diagram is shown in figure 4.4.

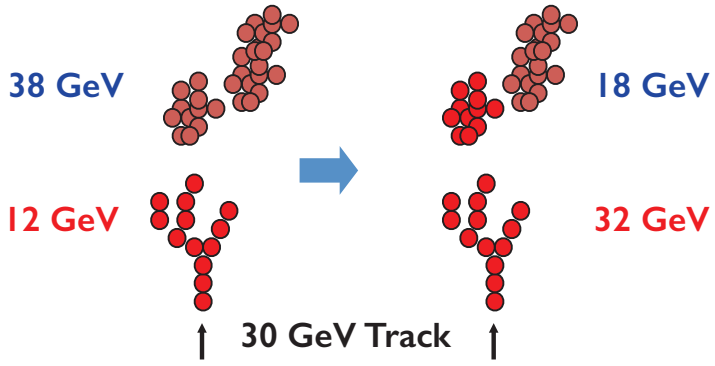


Figure 4.4: Illustration of the re-clustering algorithm in PandoraPFA, taken from [39]. The arrow indicates the tracks. The dark red dots represent the calorimeter hits in the associated cluster. The pink dots represent the calorimeter hits in the neutral cluster. The cluster energy is less than the associated track momentum. The topological association algorithms did not add the natural cluster, as it would have formed a cluster with too much energy. The re-clustering algorithm tries different cone clustering to split the neutral cluster so that the topological association could make correct association.

4.4.9 Fragment removal

The late stage of the reconstruction will focus on merging low energy clusters, especially non-photon neutral clusters. These neutral clusters are likely to be fragments of charged clusters, instead of being a physical particle. The merging criterion are mostly based on the proximity and the energy comparison. There are algorithms dealing with photon fragment merging and photon clusters splitting, which are described in details in chapter 5.

4.4.10 Particle Flow Object Creation

Particle Flow Objects (PFOs) are created at the last step. Tracks are associated to the clusters based on the proximity. Simple but effective particle identification for electrons, muons are applied. Photon identifications have been applied at various stages of the reconstruction.

PFOs are the output of the PandoraPFA reconstruction, providing information on positions, four-momenta and other associated information. These PFOs are used heavily in physics analyses. The electron, muon and photon identification are also used in physics analyses in chapter 6 and in chapter 7.

4.5 CLIC specific simulation and reconstruction

There are a few simulation and reconstruction specific to CLIC, that are used in chapter 7. Reconstruction does not include calorimeters hits in the forward calorimeters, due to computational reasons. Instead, a fast simulation using MC particles is used and details are laid out in section 7.3.3.

The luminosity spectrum for interactions with photon from Beamstrahlung is different to electron-positron interaction. Therefore it is discussed in section 4.5.1. There is a large amount of beam induced background in CLIC, which needs to be suppressed before physics analyses. The background suppression is described in section 4.5.2. Simulated masses of particles are given in table 4.2, which will be used in section 7.4.1.

4.5.1 Luminosity spectrum

The electron-photon interaction where the photon is via Beamstrahlung of the initial state radiation has a different instantaneous luminosity than the electron-positron interaction. During the same time-frame the total integrated luminosity are different. A simulated study has performed [40] with GUINEAPIG [41] and simulated in WHIZARD. The results are summarised in table 4.1. For physics analysis in chapter 7, event number for processes with initial-state photons from Beamstrahlung should be corrected by factors in table 4.1.

Luminosity ratio	$\sqrt{s} = 1.4\,\text{TeV}$	$\sqrt{s} = 3\,\text{TeV}$
$L(e^+e^-)\ /\ L(e^+e^-)$	1	1
$L()/(\text{L}) \to (\ e^+e^-)$	0.75	0.79
$L(\gamma e^\mp)\ /\ L(e^+e^-)$	0.75	0.79
$L()(/)\text{L}(() \to e^+e^-)$	0.64	0.69

Table 4.1: Luminosity ratio for processes with initial-state photons from Beamstrahlung for CLIC at $\sqrt{s} = 1.4\,\text{TeV}$ and $3\,\text{TeV}$. The table summarises results in [40].

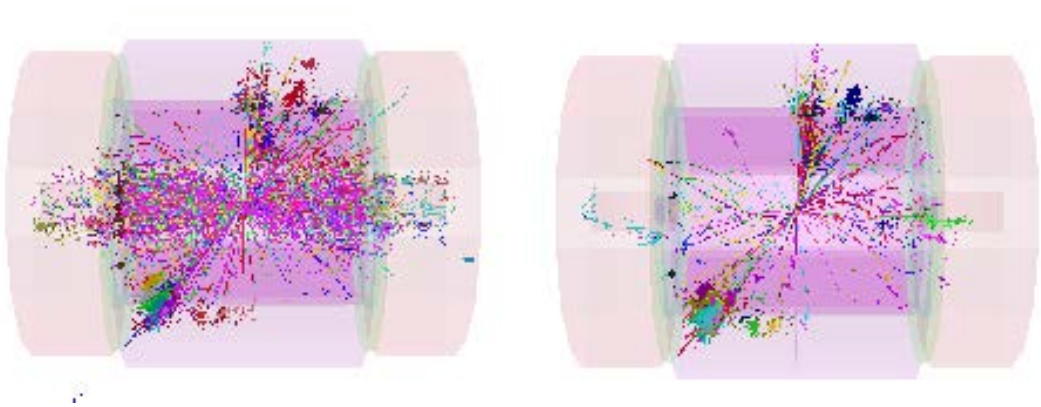
4.5.2 Beam induced backgrounds

The beam induced background are considered in the simulation. $\gamma\gamma \to \texttt{hadrons}$ intergraded over 60 bunch crossing has been overlayed onto the reconstruction. The incoherent pairs are ignored as the $\gamma\gamma \to \texttt{hadrons}$ is the dominant background in all calorimeters except inner part of the HCAL endcap. These $\gamma\gamma \to \texttt{hadrons}$ background events are hadronised with PYTHIA, and superimposed on the physics process simulations to save computational resources. The choice of 60 bunch crossing is a conservative estimate of the amount of the background [42, 42]. These background deposit significant amount of energies in the detector and need to be suppressed for physics analysis in chapter 7.

Two Marlin process has been developed to suppress these background, a track selector and a PFO selector [23].

The track selector aims to remove poor quality and fake tracks. It places simple quality cut and a simple time of arrival cut. If the arrival time of the track at the front of the ECAL, using the helical fit, differs more than 50 ns from using a straight line fit, the track will be rejected.

The PFO selector utilise the high spatial resolution from the high granular calorimeter. PFOs from $\gamma\gamma \to \texttt{hadrons}$ often have low p_T and have a range of time. PFOs from physics processes have a range of p_T, and have time close to the brunch crossing time. These two distinctive features allow $\gamma\gamma \to \texttt{hadrons}$ background to be separated. The optimal suppression uses different p_T and time cuts for the central part of the detector, and for the forward part of the detector, and uses different cuts for photons, neutral PFOs and charged PFOs. Three configurations of these cuts are developed, namely "loose", "normal", and "tight" selections. As the name suggested, "loose" selection corresponds to a looser cut of p_T and time. The optimal configuration depends on the $\sqrt{s}$ of the collision, and the physics process to study. Figure 4.5 shows the effect of the suppression of the background with the tight PFO selection.



(a) **(b)**

Figure 4.5: Reconstructed particles for a time window of 10 ns (100 ns in HCAL barrel) in a simulated $e^+e^- \to HH \to t\bar{b}b\bar{t}$ event in the CLIC_ILD dectro, with 60 bunch crossings of $\gamma\gamma \to \texttt{hadrons}$ background overlaid in figure 4.5a. The effect of applying tight PFO section cuts is shown in figure 4.5b. Figures are taken from [23].

4.5.3 CLIC simulated particle masses

Mass and width of quarks and bosons used for generating Standard Model samples given in table 4.2. This information will be used in section 7.4.1.

Particle	Mass (GeV/c^2)	Width (GeV/c^2)
u, d, s quarks	0	0
c quark	0.54	0
b quark	2.9	0
t quark	174	1.37
W	80.45	2.071
Z	91.188	2.478

Table 4.2: Masses of quarks and bosons used for generating Standard Model samples. H mass is specific for individual sample. Table is taken from [19].

4.6 Reconstruction Processors

In the last chapter we described the automated reconstruction tools in details. This chapter is dedicated to the common automated analysis software, which will be used in the analysis described in subsequent chapters.

4.6.1 MC truth linker

It is extremely useful to be able to associate reconstructed objects to the MC particles to develop algorithms or to optimise event selection. The MC truth linker processor provides the link between a MC particle and a reconstructed calorimeter hit. From the link, the main MC particle contributed to a reconstructed PFO or a group of PFOs (jet) can be determined.

4.6.2 Jet algorithm

For the linear collider, thanks to the high granular calorimeter, the starting point for analysis are individual Particle Flow Objects, as well as individual tracks. Each of the PFOs encodes four-momentum and position information. For tracks, they would have momentum and position information. However, sometimes it is interesting to group PFOs and tracks into jets, which is the result of hadronisation process from high energy particles like quarks or gulons.

A jet is typically a visually obvious structure in a event display. The momentum and the direction of a jet tend to resemble the originated particle. Despite the relative easiness

of identifying jets visually, it presents a challenge for a pattern recognition program to identify jets effectively and efficiently.

Early work on jet finding started in 1977 [43], where later development can be found in reviews [44–46].

There are two large families of jet finding algorithm, cone based algorithms, and sequential combination algorithms. Cone based algorithm is briefly discussed in section **??** in the context of the PandoraPFA reconstruction.

Sequential combination algorithms typically calculate a pair-wise distance metric. Pairs with the smallest metric will be combined. The metric will be calculated and updated after a combination. This procedure will be repeated until some stopping criterion are satisfied. The different jet algorithms typically differ in the distance metric and stopping criterion.

The chosen jet algorithm implementation is FastJet C++ software package [47, 48], providing a wide range of jet finding algorithms. The implementation in Marlin software package is called MarlinFastJet. The symbols in the subsequent discussion follow the convention in [47].

k_t algorithm Longitudinally-invariant k_t algorithm [49, 50] is one of the common sequential combination algorithms for $\overline{p}p$ collider experiment. In the inclusive variant, the symmetrical pair-wise distance metric between particle i and j, and the beam distance, are defined as

$$d_{ij} = d_{ji} = \min\left(p_{Ti}^2, p_{Tj}^2\right)\frac{\Delta R_{ij}^2}{R^2}, \tag{4.1}$$

$$d_{iB} = p_{Ti}^2, \tag{4.2}$$

where p_{Ti} is the transverse momentum of particle i with respect to the beam (z) direction, and ΔR_{ij}^2 is the measurement of angular separation of particle i and j, defined as $\Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$, where $y_i = \frac{1}{2}\ln\frac{E_i + p_{zi}}{E_i - p_{zi}}$ and ϕ_i are particle i's rapidity and azimuthal angle. R is a free parameter controlling the jet radius.

If $d_{ij} < d_{iB}$, particle i and j are merged, with the four-momentum of particle i updated as the sum of the two particles. Otherwise, particle i is set to be a final jet, and deleted from the particle list. The above procedure is repeated until no particle left.

The exclusive variant is similar. First difference is that when $d_{iB} < d_{ij}$, the particle i is discarded and part of the beam jet. The second difference is that when both d_{ij} and d_{iB} are above some threshold, d_{cut}, the clustering will stop. In practise, exclusive mode allows a specified number of jets to be found, which will automatically choose the d_{cut}. The inclusive mode would find as many jets as the algorithm allows. The exclusive k_t algorithm is used in section 7.4.1.

Durham algorithm Durham algorithm [51], also known as e^+e^- k_t algorithm, is commonly used e^+e^- collider experiment. It has a single distance metric:

$$d_{ij} = 2\min\left(E_i^2, E_j^2\right)(1 - \cos(\theta_{ij})), \tag{4.3}$$

where E_i is the energy of particle i. θ_{ij} is the polar angle difference between particle i and j. Durham algorithm can only be run at exclusive mode, which means that the clustering will stop when d_{ij} is above some threshold, d_{cut}.

Comparing to k_t algorithm, it uses energy instead of p_T in the distance metric, and it did not have a beam jet. This is because that for the e^+e^- collider in the past, the beam induced background was not severe and collisions energy is known, $\sqrt{s}$.

Jet algorithm for the CLIC Although CLIC is a e^+e^- collider, the significant beam-induced background adds a large amount of energy. Therefore, traditional e^+e^- jet algorithms, like Durham algorithm, is not suitable for the CLIC environment. Studies has shown that jet algorithms for $\bar{p}p$ colliders have better performance for CLIC [19, 52].

A more recent attempt at marrying merits from both Durham and k_t algorithms has resulted in Valencia jet algorithm [53]. It had shown promising improvement comparing to k_t algorithm, which is used in the parallel $HH \to b\bar{b}b\bar{b}$ sub-channel analysis in chapter 7 by collaborators.

y parameter y parameter is calculated for each specific jet algorithm. It is a measure of the number of jets in an event. y parameter describes the transition of exclusive jet algorithm going from N clustered jets to $N+1$ clustered jets. For example, y_{23} would be the d_{cut} value for a exclusive jet algorithm, above which the jet algorithm returns 2 jets, below which the jet algorithm returns 3 jets (see section **??** for jet algorithm).

Numerically y parameter is often much smaller than one. A typically way to convert the small number to a human acceptable range is to take the minus logarithm of the number.

4.6.3 LCFIPlus

Another useful analysis technique is to identify jets from b and c quarks. These jets have signatory topologies. A combination of vertex finding and multivariate analysis is used to identify b and c jets.

The flavour tagging processor, LCFIPlus [54] is based on the LCFIVertex package [55], which was used in the simulation studies for ILC Letter of Intent [56, 57] and CLIC Concept Design Report [19]. Current software is built in mind of a future e^+e^- collider. Although the software is modular and can be used in any order, here it will be described in the order used in a physics analysis.

The input are PFOs. The vertex finding algorithms perform vertex fitting and identify primary and secondary vertex. There is a "V0" particle rejection step, which is neutral particles decaying or converting into a pair of charged tracks. The topology is similar to the decay of b or c hadrons. Hence it is important to remove the V0 particles to improve the heavy quark flavour tagging (see section 4.4.1 for a similar V0 rejection).

Once the primary and secondary vertices are found, PFOs are clustered in to jets. This jet clustering scheme ensures that the secondary vertices and the muons identified from semi-leptonic decay fall in the same jet. Therefore, it is consistent with the hadronic decay. Jet algorithms used are Durham and Durham modified algorithms(see section 4.6.2).

The next step is to refine vertices finding to improve the b jet identification from c jet. Since the existence of two close by vertices is strongly correlated to a b jet, the vertices refining step will reconstruct as many secondary vertices correctly as possible.

The last step is to gather the information about vertices and jets, and deploy a multivariate analysis. The multivariate classier used, Boosted Decision Tree, is implemented in TMVA software package [58], which is discussed later in section **??**. A series of flavour sensitive variables are calculated, and the classification is divided into four subset: jet with zero, one, or two properly reconstructed vertices, or a single-track pseudo-vertex. For each subset, a jet can either be classified to a b jet, a c jet, or a light flavour quark jet (u, d or s). The multiclass classifier's response is normalised across different subset,

and they will be referred in the subsequent physics analysis as the tag value. See section **??** for a discussion on multiclass classifier.

The flavour tagging is performed after the initial jet reconstruction, and all the PFOs in the reconstructed jets are the input to the LCFIPlus flavour tagging processor. Therefore, the classifier in the LCFIPlus processor is trained for a specific PFO collection and a specific jet reconstruction algorithm. The output of the processor for a jet is three values, corresponding to the likelihood of the jet being a b jet, a c jet, or a light flavour quark jet.

4.6.4 Event shape variables

Event shape variables are some useful global variables to describe the shape of the event, for example whether it is back-to-back, or homogenous in the solid angle.

The classical event shape thrust [59], is defined as

$$T = \max_{\hat{t}} \frac{\sum_i \left|\hat{t}\cdot\vec{p_i}\right|}{\sum_i |\vec{p_i}|} \tag{4.4}$$

where $\vec{p_i}$ is the momentum vector of the particle i. Summation is over all particles in the event. Thrust axis, $\hat{t}$, is a unit vector. (Principle) Thrust value, T, is 1 for a perfect pencillike back-to-back two-jet event, and 0.5 for a perfect spherical event. The thrust value is useful in picking out back-to-back two-jet event. Thrust axis is useful to separate each jet in a back-to-back two-jet event.

A related variable , sphericity is derived from the sphericity tensor [60]. The sphericity tensor is defined as

$$S^{\alpha\beta} = \frac{\sum_i p_i^\alpha p_i^\beta}{\sum_i |\vec{p_i}|^2}, \tag{4.5}$$

where $\vec{p_i}$ is the momentum vector of the particle i. Summation is over all particles in the event. α and β refer to the x, y, z coordinate axis. Eigenvalues of tensor $\mathbf{S}$ can be found, or in this case diagonalisation of the matrix $\mathbf{S}$, denoted with λ_1, λ_2, λ_3. The normalisation condition requires $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3$ and $\lambda_1+\lambda_2+\lambda_3=1$. Sphericity, S, is defined

in terms of λ,

$$\mathbf{S} = \frac{3}{2}(\lambda_1+\lambda_2). \tag{4.6}$$

S, is 0 for a perfect pencil-like back-to-back two-jet event, and 1 for a perfect spherically symmetric event.

Aplanarity is another useful event shape variable that distinguishes spherical symmetrical events from planar and linear events. The definition is

$$S = \frac{3}{2}(\lambda_1), \tag{4.7}$$

where λ_1 is the largest eigenvalue in the diagonalised sphericity tensor, $\mathbf{S}^{\alpha\beta}$.

4.7 Multivariate Analysis

Multivariate analysis (MVA) has become increasingly common in high energy physics. MVA is typically used as the last step of the physics analysis to select signal from background. It can be viewed as an advanced tool for regression or classification. Comparing to the traditional cut based method, modern machine learning technique offers much improvement in data analysis. Software package for MVA used throughout this document is TMVA [58].

A typical machine learning MVA classification involves two classes, also known as signal and background. A machine learning model, also known as a classifier in TMVA, needs to be trained with training data. The model requires a set of discriminative variables, which separate the signal from background. The trained model will be applied onto the testing data for signal extraction. Response of the model could be a classification of signal or background, or could a response in a continuous spectrum, where the user decides the value to separate signal from background.

Strictly, there should be three statistically independent samples for the MVA. One sample is for the training. Another sample for the validation, including optimisation and checking for overfitting. The last sample is for testing. However, due to technical reason (TMVA only natively supports two samples), sometimes the same sample is used for the validation and the testing, which is acceptable with large statistics.

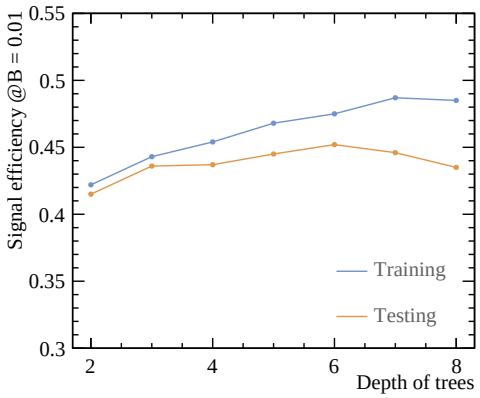


Figure 4.6: Example of model efficiency as a complexity of model parameter. Here the model is boosted decision tree. The model parameter is depth of tree. From tree depth 6 onwards, overfitting occurs.

This classification scheme can be easily extended to multiple classes, implemented in TMVA with multiclass class. The multiclass class is used in the tau decay mode classification in section **??** and in the flavour tagging classifier in section 4.6.3.

4.7.1 Optimisation and overfitting

The optimisation of the model refers to selecting the optimal free parameters of the model. One could build a complex model which fits the training samples very well, but it would not be optimal for another testing sample. A simple model is less prone to statistical fluctuation of samples, however, it might be too simple to achieve the optimal modeling. The former case is known as overfitting, or overtraining. The latter case is called underfitting, or undertraining.

The compromise is clear. The optimal model is the one between overfitting and underfitting. In practice, this involves building the model with increasing complexity, and finding the point where overfitting occurs.

Figure 4.6 shows a typical overfitting plot. Overfitting is defined when the efficiency of signal selection in the training samples increases, but the efficiency in the testing sample decreases. The example in figure 4.6 is chosen from double Higgs analysis at $\sqrt{s}$ = 3 TeV, using Boosted Decision Tree model. The efficiency of signal selection is defined as the signal fraction when background fraction is 1%, report by the TMVA training process. In figure 4.6 , the depth of the tree, or the number of layers in the tree, reflects

the complexity of the model. From tree depth 2 to 5, the efficiency for both testing and training samples increases. From tree depth 6 onwards, the overfitting occurs. In this particular example, one should choose a tree depth fewer than 7 to avoid overfitting.

Figure 4.6 can be repeated with different split of training and testing samples to avoid the statistical fluctuation in samples. This allows a better estimation of where overfitting occurs. However this method is not used as the TMVA does not support such a method.

4.7.2 Choice of models

The model, also known as the classifier in TMVA, can be as simple as cut based, likelihood or linear regression. It can also be as complicated as non linear tree, non linear neutral network or support vector machine. Regardless of model complexity, the choice of most optimal classifier is often data driven. Also, given the free parameters in each model, the comparison between different models without individual tuning is not rigourous. Nevertheless, as researchers in the machine learning suggested, the boosted decision tree is probably the best out-of-the-box machine learning method. Neutral network could potentially be better than the boosted decision, but it requires more tuning, and it is less intuitive to interpret the model. For these reasons, boost decision tree (BDT) is often the choice of machine learning model in the high energy physics. And it is used in various physics analysis in this document.

Before describing BDT in detail, we will first visit some simple models.

Rectangular Cut

Probably the most intuitive model, the rectangular cut method optimise cuts to maximise some specific metric. The metric could be the signal efficiency for a particular background efficiency. Alternatively, the metric can be the significance, $\frac{S}{\sqrt{S+B}}$, where S and B are signal and background numbers, respectively.

Discriminative variables gives better separation power when they are gaussian-like and statistically independent. Therefore it is common to decoorelate the variables and gaussian transform them before using the rectangular cut MVA.

Because its simplicity, the cut method is often performed manually, much more often in the time pre-date the wide spread of machine learning methods. It is still commonly

used for the pre-selection step before the MVA (see section 7.7), and other simple cases. Unless specified, the optimal cuts proposed in this document for various physics analysis are found using the rectangular cut method manually.

Projective Likelihood

Projective likelihood model (PDE) is used in PandoraPFA for the photon ID due to its simplicity and low requirement on computing resources. The PandoraPFA implementation is discussed in section 5.5.1

PDE implemented in the TMVA calculates the probability density for each discriminative variable, for signal and background. The overall signal and background likelihood are defined as products of the individual probability density. The likelihood ratio, R, is then defined as the signal likelihood over signal plus background likelihood. TMVA implementation also fits an underlying function to the probability density.

Similarly to the rectangular cut method, PDE works better with decorrelated, gaussian like variables.

Decision tree

Before discussing Boost decision tree (BDT), it is necessary to introduce Decision tree. Decision tree is a non linear tree based model. Its rather complex nature requires a careful explanation of many concepts.

Decision tree is a binary tree, where each node, the splitting point, uses a single discriminative variable to decide whether a event is signal-like ("goes down by a layer to the left"), or background-like ("goes down by a layer to the right"). At each node, samples are divided into signal-like and background-like sub-samples. The tree growing starts at the root node, and stops at certain criterion, which could be the minimum number of events in a node, the number of layers of the tree, or a minimum/maximum signal purity.

The training of the decision tree is to determine the optimal cut at the node by minimising the metric. The probability of the cut producing the signal is p. Three commonly used metrics for two-class classification are

1. Misclassification error: $1 - \max(p, 1-p)$,

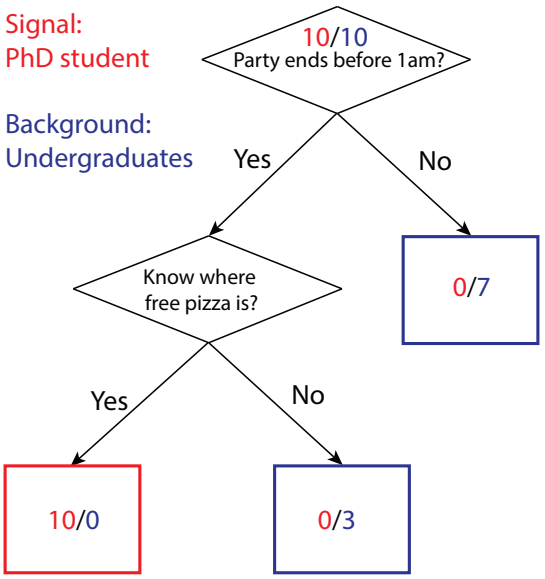


Figure 4.7: Example of a decision tree. Numbers in each node represent number of PhD student (red) and number of undergraduate student (blue) after each cut.

2. Gini index: $2p(1-p)$,

3. Cross-Entropy or deviance: $-p\log p-(1-p)\log(1-p)$.

The using of a trained decision tree is to transverse along the tree. The event is classified as signal or background depending on whether it falls in the signal-like or background-like end node.

Figure 4.7 illustrate a simple example of a decision tree. The signal is the PhD student and the background is the undergraduate student. The depth of this imbalance binary tree is 2. A node is represented by a diamond. The signal-like node is the red rectangle and the background-like nodes are blue rectangles. The tree is constructed with two possible cut, "Party ends before 1am" and "Know where free pizza is". The attribute of samples is listed in table 4.3. To demonstrate the choice of the first layer cut, the Gini index metric is used. If the first cut is "Party ends before 1am", the probability of the cut producing the signal, p, is $\frac{10}{13}$. Gini index is $2p(1-p) \backsimeq 0.36$. If the first cut is "Know where free pizza is", $p = \frac{10}{15}$. Gini index is $2p(1-p) \backsimeq 0.44$. Therefore, the first cut is "Party ends before 1am".

The simple tree in figure 4.7 is grown fully as each end node contains signal or background only. To use the trained decision tree, if there is student who ends party before 1am and knows where free pizza is, then the student is classified as a PhD student.

Number	Party ends before 1am	Know where free pizza is
PhD student	10	10
Undergrad student	3	5

Table 4.3: The attribute of samples for the decision tree example.

Improve decision tree Decision tree has a low bias, but high variance. This means it is very easy to construct a tree, which is also the best tree, that fits the training data very well, but the tree would not be optimal for the testing sample. To overcome the instability of the decision tree, many methods have been developed. The most successful one is boosting and bagging.

Boosting: it is a technique where the misclassified events receives a higher weight than the correctly classified events. Therefore, when the training is iterated, the misclassified events would receive higher and higher wights and more likely to classify correctly. The boosting is done at every iteration, which can be few hundred or few thousand time. This will create a "forest" of many trees. The final output could be a majority vote, by transversing the event to the end node for each tree in the forest.

Bagging: also known as boot-strap, it is a method that select a simple random sub-set of the training sample, and apply the model. In this case, every boosting iteration takes a bagged sample, rather than the whole sample.

Boosted decision tree

Boosted decision tree (BDT) contains a forest of decision trees , where each tree is iterated many times using a technique called boosting. By overcoming the instability of a single decision tree, BDT is often regards as the best out-of-the-box machine learning method. There are two common boosting methods: adaptive boosting and gradient boosting. First introduced in [61], the adaptive boosting is discussed in further details, as it is simpler to understand than gradient boosting.

The basic idea of the adaptive boosting is such that the tree making procedure focuses on events which are difficult to classify correctly. By assigning a weight to each event, after each tree making iteration, the weights for misclassified events are gradually increased. Therefore misclassified events gets more attention.

A simple example is provided. Assuming tree classifier output is -1 or 1. One can think of -1 is background and 1 is signal. Suppose there are N events and M iterations (trees). For i^{th} event in m^{th} tree, $B_{i,m} = 1$ if the event is misclassified, 0 otherwise.

The adaptive boosting algorithm, adapted from [62], is outlined below.

- For N events, event weight is $w = 1/N$ for every event.
- Iterate for M times. M is the total number of trees. For iteration m:
  - Create a m^{th} tree with weighted samples.
  - Update m^{th} tree error function, $err_m = \frac{\sum_{i=1}^{N} w_{i,m-1} B_{i,m}}{\sum_{i=1}^{N} w_{i,m-1}}$.
  - Update m^{th} tree weight, $\alpha_m = \log\left(\frac{1-err_m}{err_m}\right)$
  - Update i^{th} event weight, $w_{i,m} = w_{i,m-1} e^{\alpha_m B_{i,m}}$.
- The output $G(x)$ for testing event x is a weighted vote from all M trees:

$$G(x) = \begin{cases} -1, & \text{if } \sum_{m=1}^{M} \alpha_m G_m(x) < 0, \\ 1, & \text{otherwise.} \end{cases} \tag{4.8}$$

In each iteration, if the i^{th} event is misclassified, the weight increases by a factor of $(1 - err_m)/(err_m)$. Otherwise, the event weight does not change.

The power of the ablative boosting to dramatically improve the performance of a weak classifier. A weak classifier is a classifier is sightly better than random guessing. A small decision tree would be a weak classifier. By sequentially applying many weak classifier with weighted samples, the final "forest" is very robust with very good performance.

TMVA implementation of the BDT for the output is using a likelihood estimator, depending on how often a event is classified as signal in the forest. The likelihood number is later used to select signal from background.

Optimisation of Boosted Decision Tree Many parameters of the BDT can be tuned. The tuned parameters are described below.

The most important parameter is the depth of a tree, which determines how many end nodes a tree has, or the degrees of freedom of a tree. The related parameter is the

number of trees. Experience shows that using many small trees yields the best result. The performance as a function of the depth is shown in figure 4.6.

The number of tree is another important parameter. Intuitively large number of trees leads to overfitting. However, it has been shown that a large number does not lead to overfitting, using the definition above. Therefore there is a debate on the metric to determine the optimal number of tree.

The minimum number of events in a node, which is a stopping criteria for tree growing, affects the size of the tree. But it is less influential than the depth of the tree.

The boosting has two variant in TMVA implementation, adaptive boost and gradient boost.

The learning rate of the adaptive boost, which controls how fast the weight changes for events in each boosting iteration. Experience shows small learning rate with many trees work better than large learning rate with few trees.

The shrinkage rate of the gradient boost is similar to the learning rate of the adaptive. It controls how fast the weight changes for events in each boosting iteration. Again a small value is preferable.

The usual choice of the metric for the optimal cuts is either Gini index or cross-entropy. Typically Gini index metric is chosen. It makes little difference to performances, comparing to the cross-entropy metric.

Number of bins per variables for the cut is necessary to make tree growing efficient. Discrete binned variables are faster to computer than continuous variables. The parameter does not impact the performance much. However, variables should be pre-processed before going into the model. For example, the variable should be limited to a sensible range to avoid the extremes. The variable should also be transformed to obtain a more uniform distribution, if the original distribution is highly skewed.

For the end node, it is determined as either signal-like or background-like, based on the majority for the training event in the end node. Numerically, it corresponds to 1/0. However, the end node could also use signal purity as the output, resulting in a continues spectrum of [0,1]. The adaptive boosting algorithm is modified for the output value continues spectrum.

Bagging fraction determines the fraction of randomly selected samples used in each boosting iteration. By choosing a smaller value, samples between each boosting iteration are less correlated. Hence the overall performance improves.

DoPreSelection flag allows the classifier to throw away phases spaces where there is only background events.

4.7.3 Multiple classes

The above discussion is done assuming two classes - signal and background. The argument can be easily extended to multiple classes. There are two ways for the training. “One v.s. one” is each class is trained against each other class. And the overall likelihood is normalised. The second way to train is called “one v.s. all”, which is when each class is trained against all other classes.

Using a three-class example, A, B and C, “one v.s. one" scheme trains A against B, B against C, and C against A. Then the likelihood is normalised. “One v.s. all" would train A against B plus C, B against A plus C, and C against A plus B.

TMVA multiclass implementation uses “one v.s. all" scheme. For each class, the multiclass classifier will train the class as the signal against all other final states as the background. This process is repeated for each class. The classifier output for a single event is a normalised response using all trained classifier, where the sum is one. The response of each class in a event can be treated as the likelihood. In the classicisation stage, The event is classified into a particular class if that class has the highest classifier output response.

The advantage of using the multiclass is that the correlation between different classes are accounted for and the classifier output are correctly adjusted for multiple class. Hence one event can only be classified into one final state. The issue with the multiclass is that discriminative variables all classes need enter the training stage, resulting in a large number of variables.

TMVA multiclass implementation uses "one v.s. all" scheme. Multiclass is used in flavour tagging of jets, section 4.6.3, and in the tau lepton final state separation study, section **??**.

Chapter 5

Photon Reconstruction in PandoraPFA

"When it is obvious that the goals cannot be reached, don't adjust the goals, adjust the action steps."

— Confucius

Photon reconstruction is an important part of PandoraPFA reconstruction. A good photon reconstruction should provide a good single photon completeness and purity, as well as a good photon separation resolution. For many physics processes, heavy particles decaying into photons, such as τ lepton and π^0. Photon reconstruction is crucial for reconstructing these heavy particles.

The photon reconstruction presented in this chapter has improved the photon reconstruction completeness by reducing the fragments. The photon separation resolution has also been improved. This work has been published in a conference proceeding [63]. The improved photon reconstruction has benefited many physics analyses involving photons. The most recent example is the $H \rightarrow \gamma\gamma$ analysis at $\sqrt{s} = 3\,\text{TeV}$ at CLIC [64].

5.1 Overview of photon reconstruction in PandoraPFA

PandoraPFA provides a framework for particle reconstruction [38], as described in section 4.4. In the linear collider user case, it has a vast library of algorithms developed over years by many people. Each algorithm addresses one topological issue in the particle reconstruction. The essential part of the PandoraPFA is track-cluster association to find the best track-cluster pair, and re-clustering to find the best cluster consistent with the track. Other algorithms that identifies trackless clusters, such as muon clusters or photon clusters, would provide a cleaner environment for the track-cluster association, hence improving the jet energy resolution.

Photon identification in PandoraPFA has two main mechanisms. The basic mechanism performs photon identification at the last step of the reconstruction (see section 4.4.10). The second more sophisticated photon identification is performed at an early stage of the reconstruction (see section 4.4.4). The second algorithm identifies photon electromagnetic shower cores carefully in a dense jet environment. By removing photons from the environment, fewer calorimeters hits are left for charged particle reconstruction. Hence the overall reconstruction improves.

The Photon Reconstruction algorithm in PandoraPFA version 1 improves jet energy resolution by correctly identifying photon electromagnetic shower cores and leaving a cleaner environment for the track-cluster association. However, peripheral calorimeter hits to the shower cores may be reconstructed as separate particles (fragments). This lowers the reconstructed photon completeness and makes the number of reconstructed photons a less useful physical quantity. Also, the algorithm in PandoraPFA version 1 leaves rooms for improvement of photon separation resolution.

This chapter presents a solution to photon reconstruction issues. The newly introduced algorithms reduces photon fragments and improves the photon separation resolution.

Firstly an overview of electromagnetic showers is presented. The Photon Reconstruction algorithm will be described next, followed by fragment removal algorithms and photon splitting algorithms. This chapter will end with a discussion on performances of these photon related algorithms, including comparisons with the previous photon algorithm.

5.2 Electromagnetic shower

An electromagnetic (EM) shower refers to the pair production and bremsstrahlung when a high energy photon or electron passing though a thick absorber. The interaction generates many low energy photons and electrons, producing a shower-like structure. Two suitable length scales to describe the EM shower are radiation length and Molière radius. A radiation length of a material is used to describe the longitudinal shower profile, defined as the mean distance travelled where an electron loses its energy by a factor of $1/e$ via bremsstrahlung. It is also defined as the mean free path for pair production by a high energy photon [65]. A Molière radius is used to describe the transverse shower profile.

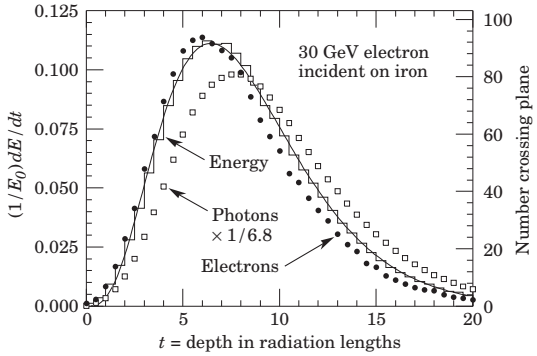


Figure 5.1: An EGS4 simulation of a 30 GeV electron-induced electromagnetic shower in iron. The histogram shows fractional energy deposition as a function of radiation lengths, and the curve is a gamma-function fit to the distribution. Circles and squares are the number of electrons and photons respectively with total energy greater than 1.5 MeV crossing planes with scale on right. Plot is taken from [2].

Figure 5.1 shows simulated longitudinal electromagnetic shower profile as a function of radiation length for electrons and photons. The mean EM longitudinal shower profile can be described by the following function [66] :

$$\frac{dE}{dt} = E_0 b \frac{b t^{a-1} e^{-bt}}{\Gamma(a)}, \tag{5.1}$$

where t is the number of radiation lengths. a and b are free parameters. E_0 is the shower energy. b varies slightly with material but it is sufficient to use $b = 0.5$. a can be calculated and is:

$$a = 1.25 + 0.5 \ln\left(\frac{E_0}{E_c}\right), \tag{5.2}$$

where E_c is the critical energy. This parametrisation should only be used to describe an average behaviour of the EM shower, as the fluctuation is important. For the photon identification, a comparison with the parametrisation is used.

The transverse shower profile can be described as a narrow cone widening as the shower develops. 90% of the energy is contained in a fiducial cylinder of 1 Molière radius. Transverse profile is often represented by a sum of two Gaussian function. This allows the two dimensional peak finding to separate EM showers using transverse shower profiles.

5.3 Photon Reconstruction algorithm

The Photon Reconstruction algorithm refers to the more sophisticated photon identification at the early stage of the reconstruction. It corresponding to "Particle ID" stage in section 4.4.4 in the PandoraPFA reconstruction chain. Main steps of Photon Reconstruction are: coarsely forming photon clusters, finding photon candidates, photon ID test, and optional fragment removals. The step to find photon candidates uses a two dimensional peak finding algorithm, which requires further explanation in section 5.4. The photon ID test involves a multi dimensional likelihood classifier, which is described in section 5.5. The optional fragment removal algorithm shares a common base case with another photon fragment removal algorithm. Hence they are discussed together in section 5.6 . A flow diagram of the Photon Reconstruction algorithm is shown in figure 5.2.

5.3.1 Form photon clusters

The input of the Photon Reconstruction algorithm is a collection of calorimeter hits. Muon clusters have been removed prior to this step. This step finds large potential photon clusters in the ECAL, which may contain multiple photons. For simplicity, the algorithm opts to reuse the cone clustering algorithm (see section 4.4.3) provided inside PandoraPFA to find large clusters. Since the target for reconstruction is neutral electromagnetic shower in the ECAL, the cone cluster algorithm is set to use high energy calorimeter hits as initial seeds. The parameters for the cone clustering are generous, forming large clusters for further process.

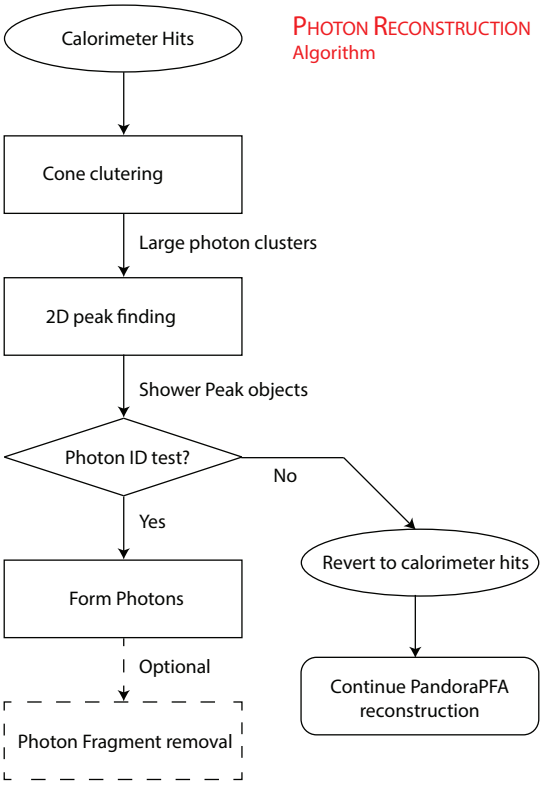


Figure 5.2: A flow diagram of the PHOTON RECONSTRUCTION algorithm.

5.3.2 Find photon candidates

The next stage is to refine the large photon clusters into smaller photon candidates. Each photon candidate should contain calorimeter hits from one particle only. The aim for this step is to split a three dimensional cluster into several smaller clusters. The three dimensional splitting problem is hard. Therefore, a translation is needed to map the three dimensional problem to a more manageable two dimensional problem. The translation relies on the transverse distribution of characteristic electromagnetic (EM) showers. When an energetic photon or electron hits the absorber layers of the ECAL, it initiates an electromagnetic shower, where electron pair production and bremsstrahlung produce low energy photons and electrons. The transverse distribution is characterised by a narrow cone, widening while the shower develops. Therefore, along the direction of the photon shower, an EM shower appears as a dense shower core with peripheral hits. If the energy deposition is projected on to a plane, the EM shower core would appear as a peak. By identifying a peak in the two dimensional plane, an EM shower core is identified. Hence with the projection translation, three dimensional cluster splitting problem is mapped to a two dimensional peak finding problem. Figure 5.3 shows an

example of a large photon cluster projected on to a two dimensional plane, where two EM showers are identified.

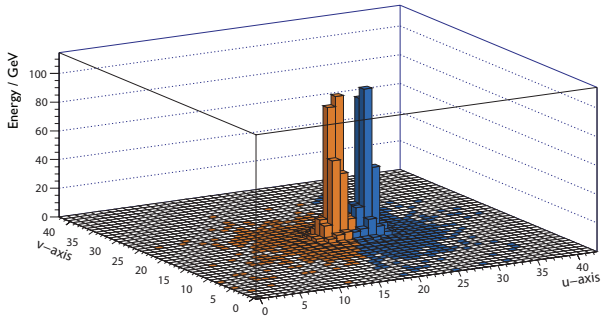


Figure 5.3: Two 500 GeV photons (yellow and blue), just resolved in a transverse plane orthogonal to the direction of the flight by projecting their energy deposition in the electromagnetic calorimeter. U and V are orthogonal axes. Z axis is the sum of the calorimeter hit energy in GeV.

By using the two-dimensional energy deposition projection, separating photons translates to separating peaks in the projection. Therefore a high-performance two dimensional peak finding algorithm is the key to identify multiple photons. Due its complexity, the peak finding algorithm is discussed separately in section 5.4. The output of the two dimensional peak finding is a collection of Shower Peak objects. Each Shower Peak object corresponds to a photon candidate and contains associated calorimeter hits.

5.3.3 Photon ID test

Shower Peak objects are tested for photon tagging. A set of discriminating variables that exploit features of electromagnetic showers are calculated. A multidimensional likelihood classifier is implemented, which needs to be trained before applying. The response from the classifier determines if a Shower Peak object is a photon. If it is a photon, it would be tagged and separated from the event. If it is not a photon, the calorimeter hits of the Shower Peak object will be passed on to the next stage of the reconstruction (see figure 5.2). Because the classifier is complicated, it is discussed in a separate section 5.5.

5.3.4 Photon Fragment removal

The optional photon fragment removal algorithm aims to merge small photon fragment to identified photons. Since this step shares the same logic as the fragment removal algorithm in section 5.6, this step is described in details in section 5.6.

This step marks the end of the photon reconstruction algorithm. The output are a collection of reconstructed photons, separated from non-photon calorimeter hits. Figure 5.2 summarises key steps in the PHOTON RECONSTRUCTION algorithm.

5.4 Two dimensional peak finding algorithm for photon candidate

As discussed in section 5.3.2, separating photon candidates from a cluster is translated to identifying peaks in a two dimensional histogram. An example of two photons in a two dimensional plane is shown in the figure 5.3. The 2D PEAK FINDING algorithm aims to correctly identify peak positions in a two dimensional histogram and associate other bins. A flow chart for the algorithm is shown in figure 5.4.

The base algorithm treats all clusters as potential photon clusters, hence the neutral cluster variant. Since charged hadrons would deposit tracks in the tracking system, extra care is taken when a cluster is close to the projection of the track in the front of the ECAL (see section 5.4.6). The neutral cluster variant is described first, followed by the modification for charged cluster variant.

5.4.1 Initialise 2D histogram

A two dimensional histogram is initialised. For the best resolving power between photons, the projection direction is chosen to be cluster's direction. Therefore, two axes for the 2D histogram are chosen such that axes and the direction form a orthogonal bases in three dimensional space. The axes are labelled as U and V axis in figure 5.3.

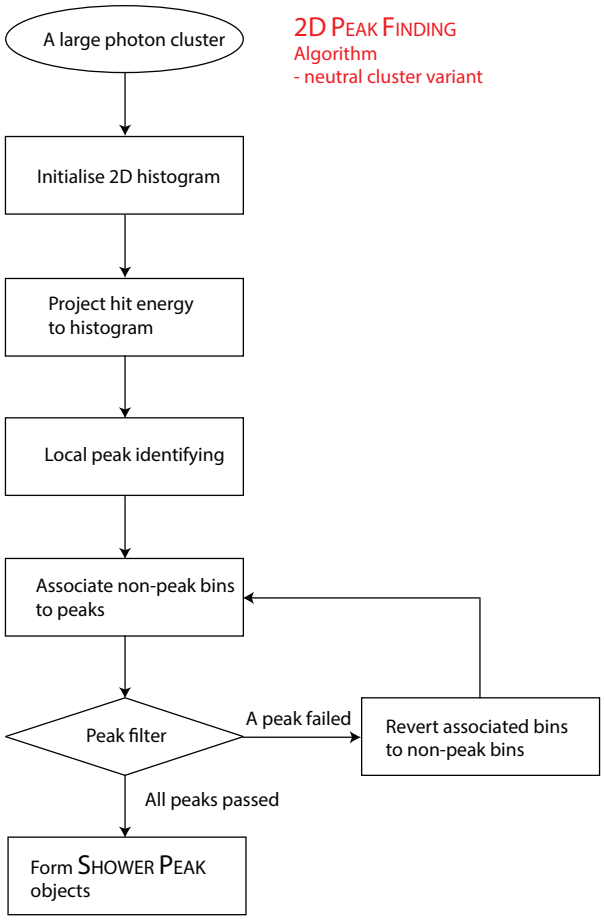


Figure 5.4: Flow chart for 2D PEAK FINDING algorithm neutral cluster variant.

5.4.2 Project hit energy to histogram

This step projects calorimeter hits positions onto the 2D histogram defined in previous step. For a finite sized 2D histogram, the projection is chosen such that the cluster centroid position is at the centre of the histogram. The bin size corresponds to one ECAL square cell length. The relative distance between the calorimeter hit position is converted into a two dimensional coordinates, and subsequently projected onto the histogram. The coordinate of a hit is obtained by:

$$\vec{s_i} = \frac{\vec{a_i} - \langle\vec{a}\rangle}{d_{cell}}, \tag{5.3}$$

where $\vec{a}$ is the position of the hit i. $\langle\vec{a}\rangle$ is the centroid position of cluster a. d_{cell} is the ECAL square cell length. $\vec{s_i}$ is the three dimensional coordinate, subsequently projected to U and V axes via the scalar product.

The projected position on the 2D histogram is binned at integer intervals. The bin height is the sum of the calorimeter hits energy in that particular bin. The issue with a finite histogram size is discussed in section 5.4.7.

5.4.3 Local peak identifying

This step identifies all local peaks in the 2D histogram. For example, in figure 5.3, there are clearly two peaks, both colour coded. A local peak is defined as a bin where its height is above all eight neighbouring bins. The 2D histogram is linearly scanned.

5.4.4 Associate non-peak bins to peaks

Having tagged all local peaks, this step associates non-peak bins to peaks. A high energy EM shower has a wider shower width. Therefore bins should be associated based on its distance to the peak and the energy of the peak. This determines the choice of the metric for associating bins. After all local peak bins are found, non-peak bins are associated to a peak bin, by choosing the peak bin that minimise the metric

$$\frac{d}{\sqrt{E_{peak}}} \tag{5.4}$$

where d is the Euclidean distance between a non-peak bin and a peak bin on the histogram, and E_{peak} is the height of the peak bin. Alternative metrics provided in the algorithm include d, $\frac{d}{E_{peak}}$, and $\frac{d}{E^2_{peak}}$. The default metric is chosen due to a good balance between distance and energy of the peak.

5.4.5 Peak filtering

The performance of the two dimensional peaking finding algorithm is improved by clever programming and physics arguments. For a given two dimensional histogram, such as the one in figure 5.3, major peaks most likely correspond to physical photons, while the minor peaks more likely come from fluctuations in energy deposition. To select major peaks, every time after all non-peak bins are associated with peak bins, minor peaks with fewer than three bins associated (including the peak bin) are discarded. These discarded bins are re-associated with other peak bins. This iterative process stops when all peak bins have at least three bins associated.

The peak filtering step also allows bins with height below a critical value to not participate in the peak finding. The default value is set such that only empty bins are not used.

This marks the end of the Photon Reconstruction algorithm with neutral clusters, outlined in figure 5.4.

5.4.6 Candidate close to track projection

If a cluster or a photon candidate is close to the projection of the track in the front of the ECAL, it is more likely that the cluster or the candidate is a charged hadron. Misidentifying a charged hadron as a photon leads to significant degradation in reconstruction performance. However, if a photon next to a charged hadron is carefully reconstructed, the overall reconstruction is improved. Hence this step aims to carefully identifies photon candidates next to charged hadrons, by using track information and features of electromagnetic showers, such as electromagnetic shower typically starting in first few layers with direction largely unchanged.

Figure 5.5 shows the flow chart for the full 2D Peak Finding algorithm, including the treatment to clusters close to tracks. If a cluster is less than 3 mm from the closest track projection, it is treated as a potentially charged cluster. The ECAL is sliced

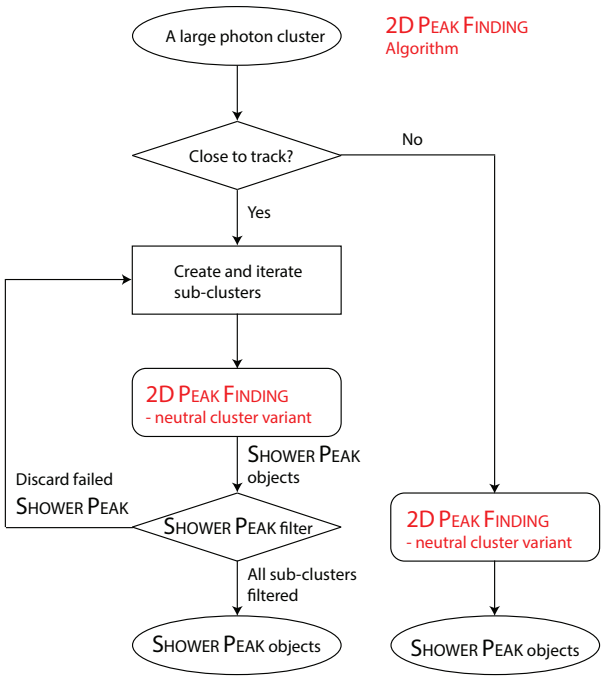


Figure 5.5: Flow chart for 2D Peak Finding algorithm.

longitudinally to help identify photon candidates. For example, the default three slices will result in three ECAL fiducial spaces, which cover from the front of the ECAL to a third, two thirds and the back of the ECAL, respectively. The peaking finding algorithm is repeated for sliced clusters contained in each fiducial space. A peak and corresponding Shower Peak object are only preserved as a photon candidate if the peak exists for evert sliced cluter, and if its position is shifted by no more than one neighbouring bin between sliced clusters. Furthermore, if a peak bin is within the eight neighbouring bins of the track projection, the peak and its associated bins are flagged as charged particles (non-photon).

5.4.7 Inclusive mode

The two dimensional histogram is iterated during the algorithm. The time complexity is $O(n^2)$ for a $n \times n$ histogram (Default $n = 41$). Therefore, for the purpose of speed, it is undesirable to have a very large histogram. Having a small finite histogram speeds up the calculation. However, because of the finite size, only energy deposition projected on the histogram would be considered for peak finding. Calorimeter hits outside the histogram would be lost when Shower Peak objects are constructed. This behaviour is suitable for Photon Reconstruction algorithm (section 5.3) and for photon fragment

removal (section 5.6). However, for photon splitting (section 5.8), there should be no calorimeter hits loss from splitting a photon. Hence inclusive mode of the peak finding algorithm is developed, and it allows energy deposition projected outside the histogram to be associated with identified peaks.

5.5 Likelihood classifier for photon ID

Section 5.3.3 outlines the photon ID test in the photon reconstruction algorithm. This section describes the multidimensional likelihood classifier in details, including discriminating variables. For each photon candidate, a set of variables are calculated and used to as inputs to the classifier.

5.5.1 Overview of Projective Likelihood

Projective likelihood model (PDE) is used in PandoraPFA for the photon ID due to its simplicity and low requirement on computing resources.

PDE implemented calculates the probability density for each discriminative variable, for signal and background. The overall signal and background likelihood are defined as products of the individual probability density. The likelihood ratio, R, is then defined as the signal likelihood over signal plus background likelihood.

To use the likelihood ratio, one way is to fit an underlying function to the probability density, which is implemented the TMVA software package. The other way is to use binned likelihood ratio as the output, due to its simplicity. This is implemented in the PandoraPFA. Similarly to classifiers like the rectangular cut method, PDE works better with decorrelated, gaussian like variables. The PandoraPFA implementation does not decorrelate nor transform the variables, to keep implementation fast.

5.5.2 Projective Likelihood in PandoraPFA

Kinematic variables exploit the differences between a characteristic electromagnetic shower and a hadronic shower, and the fact that a photon is more likely to be isolated from other showers and charged tracks. A full list of variables can be found in table 5.1.

Categories	Variables
Longitudinal shower profile	δl,t_0
Transverse shower profile	$\langle w \rangle$,$\delta\langle w_{UV}\rangle$, $\delta E_{cluster}$
Distance to track	d

Table 5.1: List of variables for the likelihood based photon ID test.

Two variables exploit the longitudinal EM shower distribution. t_0 is the start layer from the longitudinal shower profile (see section 5.2). δl is fractional difference from the expected shower profile [22]:

$$\delta l = \frac{1}{E_0} \sum_i \left| \Delta E^i_{obs} - \Delta E^i_{EM} \right|. \tag{5.5}$$

δl is minimised as a function of the t_0. For a proper photon, t_0 and δl are expected to be close to 0.

Three variables use the transverse shower information. $\langle w \rangle$ is the energy weighted r.m.s. distance of a bin to its peak. This is a measure of the transverse shower size. $\delta\langle w_{UV}\rangle$ is the smallest ratio of the two r.m.s. distances in each U, V axis direction, a measure of the circularity of the transverse shower. Last variable, $\delta E_{cluster}$, is the ratio between the photon candidate energy to the big cluster energy. This is a measure of the dominance of a photon in a large cluster.

Last variable for the classifier, d, is the distance between the photon candidate and the closest track projection. It is less likely to be a photon if it is close to a track.

The distributions of these variables are normalised to probability distribution, stored in binned histograms. The classifier is improved by realising the variable distributions varies with photon energies. Thus these distributions are divided by bins of photon candidate energy. The default energy bins edges are 0.2, 0.5, 1, 1.5, 2.5, 5, 10, 20 GeV, which covers a good range of photon energies. Candidate with energy below 0.2 GeV would not be examined in this step, as it is very unlikely to be a photon. The classifier training typically uses simulated 250 GeV jet events. High energy jets allow the training for high energy photon candidates.

After training, for a given photon candidate, the likelihood classifier output is given by

$$\mathrm{pid} = \frac{N \prod_i P_i}{N \prod_i P_i + N' \prod_i P'_i} \tag{5.6}$$

where P_i and P'_i are the probability of i^{th} variable of photon and non-photon candidates. N and N' are the number photon and non-photon candidates, found in the training step.

During classification, a candidate passes the photon ID test if

$$\begin{cases} \mathrm{pid} > 0.6, & \text{if } 0.2 < E < 0.5\,\mathrm{GeV} \\ \mathrm{pid} > 0.4, & \text{if } E \geqslant 0.5\,\mathrm{GeV} \end{cases} \tag{5.7}$$

where E is the candidate energy. Two values of the pid cuts reflect the confidence of the photon ID test with different candidate energy. The test is more cautious with low energy candidate.

5.6 Photon fragment removal algorithm in the ECAL

During the reconstruction, it is possible that a core of the photon electromagnetic shower is identified as a photon (the main photon). The outer part of the shower is reconstructed as a separate particle, and wrongly identified as a photon or a neural hadron. Figure 5.6 shows a typical creation of such a photon fragment. A fragment does not have the electromagnetic shower structure, and typically it has much lower energy than a proper photon. If a photon-fragment pair is merged, the pair should be consistent with an one-particle profile. These characteristics are used to merge fragments to main photons.

Photon fragment removal algorithms can exist in multiple step in the reconstruction: at the end of the Photon Reconstruction algorithm (see Section 5.3.4), or at the end of the PandoraPFA reconstruction. Since these algorithms share the same base class, they will be discussed together. The latter algorithm will be discussed here. The former algorithm differs mostly in the default cut-off values for merging metrics.

Spatially close photon and a potential fragment form a pair of particles (photon-fragment pair). Kinematic and topological properties of a photon-fragment pair are

(a) **(b)** **(c)**

Figure 5.6: An event display of a typical 10 GeV photon (figure 5.6a), reconstructed into a main photon (figure 5.6b) and a photon fragment (figure 5.6c).

examined. The pair is merged when the properties pass a set of cuts, developed by comparing true photon-fragment pairs and non photon-fragment pair. This merging test is iterated over all possible photon-fragment pairs. If multiple photon-fragment pairs pass the merging test, the pair with closest distance metric, d, will be merged.

The photon-fragment pairs is classified into photon-photon-fragment pairs and photon-neutral-hadron-fragment pairs, because they have different kinematic and topological distributions. The pairs are further classified into low energy and high energy pairs, depending on whether the fragment energy (E_p) is above 1 GeV. The cuts for merging pairs are listed in table 5.2.

Table 5.2 lists cuts for merging photon-photon-fragment pairs and photon-neutral-hadron-fragment pairs for both low energy and high energy fragments. d, d_c and d_h are mean energy weighted intra-layer distance within the pair, distance between two centroids, and minimum distance between calorimeter hits of each PFO in the pair, respectively. Three distance measurements have subtle difference. d_c gives the distance between centroids of each PFO in the pair, which is a quick but crude measurement. d_h is the minimum distance between calorimeter hits of each PFO in the pair. For a true photon-fragment, d_h should be close to zero as the pair should be spatially close. d is the mean energy weighted intra-layer distance between each PFO in the pair (see figure 5.7):

$$d = \frac{\sum_i^{layers} d_{l,i}\ E_{f,i}}{\sum_i^{layers} E_{f,i}} \tag{5.8}$$

where i indicates i^{th} pseudo-layer of the ECAL. $d_{l,i}$ is the minimum distance between calorimeter hits of the pair in the i^{th} pseudo-layer. $E_{f,i}$ is the energy of the fragment in

Low E_f	Photon-photon	Photon-neutral-hadron
transverse shower comparison	$d < 30$, $\frac{E_{p1}}{E_m+E_f} > 0.9$, $\frac{E_{p2}}{E_f} < 0.5$, $E_{p1} > E_m$	-
close proximity	-	$d < 20$, $d_c < 40$
low energy fragment	$d < 20$, $E_p < 0.4$	-
small fragment 1	$d < 30$, $N_{calo} < 40$, $d_c < 50$	$d < 50$, $N_{calo} < 10$, $d_h < 50$
small fragment 2	$d < 50$, $N_{calo} < 20$	-
small fragment forward region	$N_{calo} < 40$, $d_c < 60$, $E_f < 0.6$, $\|\cos(\theta_Z)\| > 0.7$	-
relative low energy fragment	$d < 40$, $d_h < 20$, $\frac{E_f}{E_m} < 0.01$	$d < 40$, $d_h < 15$, $\frac{E_f}{E_m} < 0.01$
High E_f	**Photon-photon**	**Photon-neutral-hadron**
transverse shower comparison	$\frac{E_{p1}}{E_m+E_f} > 0.9$, $E_{p2} = 0$ or ($\frac{E_{p2}}{E_f} < 0.5$, $E_{p1} > E_m$)	$\frac{E_{p1}}{E_m+E_f} > 0.9$, $E_{p2} = 0$ or ($\frac{E_{p2}}{E_f} < 0.5$, $E_{p1} > E_m$)
relative low energy fragment 1	$d < 40$, $d_h < 20$, $\frac{E_f}{E_m} < 0.02$	$d < 40$, $d_h < 20$, $\frac{E_f}{E_m} < 0.02$
relative low energy fragment 2	-	$d < 40$, $d_h < 20$, $\frac{E_f}{E_m} < 0.1$, $E_f > 10$
relative low energy fragment 3	-	$d < 20$, $d_h < 20$, $\frac{E_f}{E_m} < 0.2$, $E_f > 10$

Table 5.2: The cuts for merging photon-photon-fragment pairs and photon-neutral-hadron-fragment pairs for both low energy and high energy fragments. d, d_c and d_h are the mean energy weighted intra-layer distance of the pair, the distance between centroids, the minimum distance between calorimeter hits of the pair. E_m and E_f are the main photon energy and the fragment energy. E_{p1} and E_{p2} are the two largest peaks, found by peak finding algorithm, ordered by descending energy. N_{calo} is the number of the calorimeter hits in the fragment. $|\cos(\theta_Z)|$ is the absolute cosine of the polar angle, where beam direction is the z-axis.

the the i^{th} pseudo-layer. d is a better measurement of the closeness of the pair. All three distance metrics should be small to merge a photon-fragment pair.

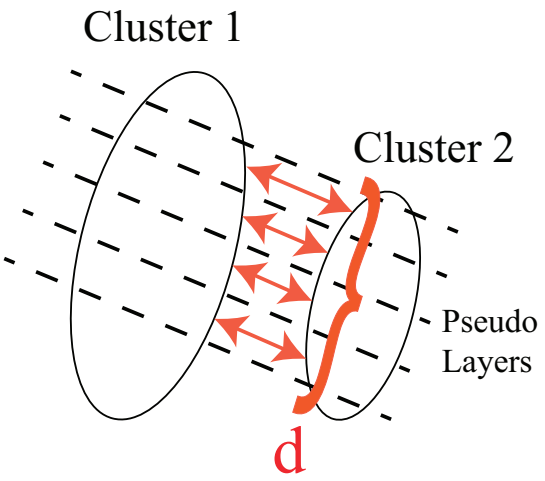


Figure 5.7: Illustration of distance metric, d.

E_m and E_f are the main photon energy and the fragment energy. E_{p1} and E_{p2} are the two largest peaks and associated calorimeter hits, found by the 2D PEAK FINDING algorithm (section 5.4), ordered by descending energy, using the pair as input. N_{calo} is the number of the ECAL hits in the fragment. $|\cos(\theta_Z)|$ is the absolute cosine of the polar angle of the main photon, where beam direction is the z-axis.

One logic for merging is when the fragment is small with low energy and is close to the main photon. Hence E_f and N_{calo} are required to be small. Alternatively the fragment should be relatively low energy, demanding a small ratio of E_f to E_m.

The other logic is when the pair looks like one photon in two-dimensional energy deposition projection (see section 5.3.2 and figure 5.3). The transverse shower comparison requires $\frac{E_{p1}}{E_m+E_f} > 0.9$, where most energy contains in the first SHOWER PEAK object.

Comparing low E_f and high E_f cut, the cuts are similar. High E_f cuts are more relaxed on the energy comparison for small fragment test.

Comparing photon-photon-fragment pair and photon-neutral-hadron-pair, cuts for photon-neutral-hadron-pair are more conservative for low E_f, but more relaxed for high E_f. This reflects that the neutral hadron fragments originated from charged particles are more likely to be low energy, whilst high energy neutral fragments are more likely to be photon fragments.

Since all possible photon-fragment pairs are compared, this is a costly cooperation with $O(n^2)$ time complexity for n particles. The speed is improved by considering only the pairs with $d < 80$ mm.

5.7 Photon fragment recovery algorithm in the HCAL

Section 5.6 descried an effective algorithm to removal photon fragments that are peripheral to the main photon, the electromagnetic shower core. An example of such fragment is shown in figure 5.6. There is another type of fragments originated from the leakage effect of the ECAL. When a high energy EM shower is not fully contained in the ECAL, shower deposits energy in the HCAL, which often forms a neutral hadron in the HCAL. Previous algorithms only consider calorimeter hits in the ECAL. Therefore no attempts have been made to recover photon fragments in the HCAL. This section presents an algorithm to merge photon fragments in the HCAL

An example of a 500 GeV photon reconstructed into a main photon in the ECAL (yellow) and a neutral hadron fragment in the HCAL (blue) is shown in figure 5.8. For the ILD detector, this ECAL leakage effect appears when the photon energy is above 50 GeV.

Figure 5.8: An event display of a typical 500 GeV photon, reconstructed into a main photon in the ECAL (yellow) and a neutral hadron fragment in the HCAL (blue).

Shown in figure 5.8, photon fragments in the HCAL is spatially close to the main photon. A fitted cone from the main photon, if extended to the HCAL, covers most of

the fragment. These features allow a set of cuts developed to merge fragments in the HCAL, listed in table 5.3

This algorithm would collect photons in the ECAL and neutral hadrons in the HCAL as inputs. The algorithm would iterate over all pairs of reconstructed photons and neutral hadrons. For each pair, a set of variables are calculated and compared to a set of cuts (table 5.3). Photon-fragment pairs passing the cuts will be merged.

High energy fragment recovery	Cuts
distance comparison	$d_c^l \leqslant 173$ mm, $d_{fit}^l \leqslant 100$ mm, $d_{fit} \leqslant 100$ mm
shower width comparison	$0.3 \leqslant \frac{w_f^l}{w_m^l} \leqslant 5$
projection comparison	$r_f \leqslant 45$ mm
energy comparison	$\frac{E_f}{E_m} \leqslant 0.1$
cone comparison	$\%N \geqslant 0.5$

Table 5.3: The cuts for merging high energy photon fragment in the HCAL to the main photon in the ECAL. d_c^l is the distance between centroids of the last outer layer of the main photon and the first inner layer of the fragment. d_{fit}^l is the distance between fitted directions using the last outer layer of the main photon and the first inner layer of the fragment. d_{fit} is the distance between fitted directions using the main photon and the fragment. w_m^l and w_f^l are the r.m.s. width of the last outer layer of the main photon and the first inner layer of the fragment. r_f is the r.m.s. mean energy weighted distance of a calorimeter hit in the fragment to the direction of the main photon. E_m and E_f are the main photon energy and the fragment energy. $\%N_{calo,cone}$ is the fraction of the calorimeter hits in the fragment in the extended fitted cone of the main photon.

Fragment in the HCAL should be spatially close to the main photon, measured by three metrics. d_c^l is the distance between centroids of the last outer layer of the main photon and the first inner layer of the fragment. d_{fit}^l is the distance between fitted directions using the last outer layer of the main photon and the first inner layer of the fragment. d_{fit} is the distance between fitted directions using the main photon and the fragment. These three distances should be small for merging.

The direction of the fragment should be similar to the direction of the main photon. r_f, the r.m.s. mean energy weighted distance of a calorimeter hit in the fragment to the direction of the main photon, has to be small for merging.

Another feature of the fragment and the main photon is that the shower width should be similar. w_m^l and w_f^l are the r.m.s. width of the last outer layer of the main photon and the first inner layer of the fragment. The ratio $\frac{w_f^l}{w_m^l}$ needs to be in the range of 0.3 to 5. The generous upper bound is because the HCAL cell size is much larger than that of the ECAL.

When a fitted cone from the main photon is extended to the HCAL, the cone should contain a significant amount of the fragment. %N, the fraction of the calorimeter hits in the fragment in the extended fitted cone of the main photon, has to be no less than 0.5 for the merging.

The last criteria is the fragment should has low energy relative to the main photon. E_m and E_f are the main photon energy and the fragment energy. The ratio, $\frac{E_f}{E_m}$, has to be less than 0.1 for the merging.

If multiple photon-fragment pairs pass the cuts with the same fragment, the pair with highest %N will be merged.

This HCAL fragment removal algorithm occurs after the first pass of topological association in the reconstruction, which connects tracks to clusters in the ECAL and the HCAL.

5.8 Photon splitting algorithm

Algorithms described above deal with forming photons from calorimeter hits in the ECAL, merging photon fragments in the ECAL and the HCAL. Another aspect in photon reconstruction is splitting accidentally merged photons. During the particle reconstruction, it is possible that photons are accidentally merged if they are spatially close. Hence another algorithm at the end of the particle reconstruction addresses this issue and tries to split merged photons.

Merged photons are typically energetic. A merged photon should be consistent with topologies of a spatially closed photon pair. Extra care should be taken if the photon is close to a charged tracks. Many PandoraPFA algorithms deal with track clusters association and there is a greater confidence in clusters associated with tracks. These features form logics behind the algorithm.

Photon splitting	Cuts
Cuts	$E > E_{c1}$, $E_{p2} > E_{c2}$, $N_p < 5$
E_{c1} and E_{c2} values	
0 charged PFO nearby	$E_{c1} = 10$, $E_{c2} = 1$
1 charged PFO nearby	$E_{c1} = 10$, $E_{c2} = 5$
> 1 charged PFO nearby	$E_{c1} = 20$, $E_{c2} = 10$

Table 5.4: The cuts for splitting photons, and the values for energy cut-off points. E is the photon energy. E_{p2} is energy if the second largest peak from the two dimensional peak finding. N_p is the number of peaks identified by the peak finding. E_{c1} and E_{c2} are the energy cut-off values, determined by the number of nearby charged PFOs.

The table 5.4 shows the cuts for the algorithm. E is the photon energy. E_{p2} is energy of the second largest peak from the 2D Peak Finding algorithm. If an energetic photon (cut on E) is identified, and two energetic EM showers (cut on E_{p2}) can be found at the same time, the photon should be split according to the 2D Peak Finding results.

When the candidate is close to a charged track, extra care is taken by demanding a large value for second EM shower energy. E_{c1} and E_{c2} are the energy cut-off values, determined by the number of nearby charged track.

The restraint on N_p, the number of peaks identified by the peak finding, is needed because one reconstructed photon is unlikely to be merged from more than four photons.

5.9 Compare with no photon reconstruction

Motivations and implementations of four different photon related algorithms have been described above. The main photon reconstruction algorithm in section 5.3 improves the photon completeness and the photon pair resolution, due to the improved two dimensional peak finding algorithm in section 5.4. The fragment removal algorithms in section 5.6 and section 5.7 further reduce the photon fragments in the ECAL and the HCAL. The photon splitting algorithm in section 5.8 exploits the peak finding algorithms to separate photons using transverse shower information, which improves the photon separation resolution.

Photon reconstruction improves single photon resolutions. It also improves jet energy resolution at high $\sqrt{s}$ because of the high photon reconstruction completeness,. This section compares the performance with and without photon related algorithms using photon pair and jet samples. The nominal ILD detector model is used. The single and two photon events were generated with a uniform distribution in the solid angle for a range of the opening angles between the pair. Events are selected such that there is no early photon conversion and the Monte Carlo photon deposits energies in the calorimeter. The events are further restricted to photon decaying in barrel and end cap region only, to minimise the detector effect.

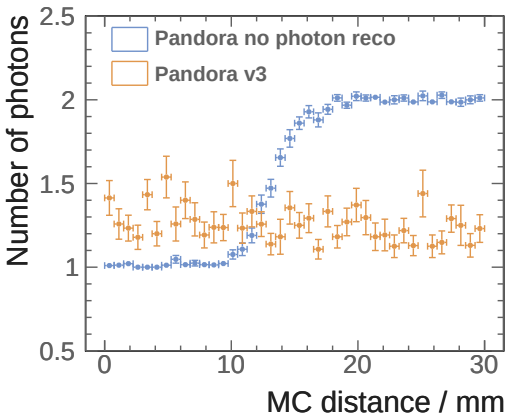


Figure 5.9: Average number of photons using two photons of 500 and 50 GeV per event sample reconstructed without and with photon algorithms.

Figure 5.9 shows the photon reconstruction for two spatially close photons reconstructed without and with photon algorithms.. Without the photon related algorithms, fragments are produced. The number of photons between 0 and 10 mm separation is around 1.5. The true photon number for that region should be 1, as it is extremely challenging to separate photons less than two cells apart. Without the photon related algorithms, photon separation resolution is much worse, as the number of photon appears to be flat between 0 to 30 mm separation. With the photon related algorithms, two photons start to be separated at 10 mm and fully separated at 20 mm separation.

The improvement in completeness and resolution in photon reconstruction, leads to a considerable improvement in the jet energy resolution at high energy. Jet energy resolution is defined as the root mean squared divided by the mean for the smallest width of distribution that contains 90% of entries, using $Z' \to u/d/s$ sample at barrel region. The angular cut is to avoid the barrel/endcap overlap region. The light quark decay of the Z is used as PandoraPFA does not attempt to recover missing momentum

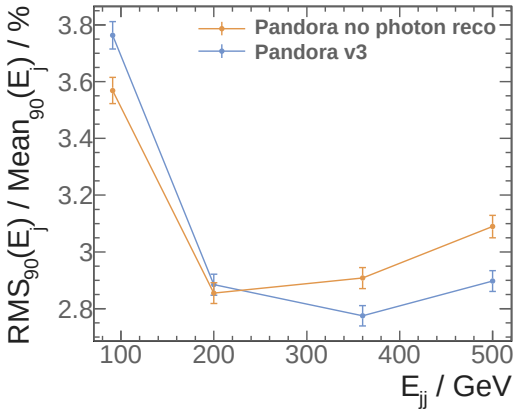


Figure 5.10: Jet energy resolution as a function of the di-jet energy using $Z' \to u/d/s$ sample at barrel region. The top orange and bottom blue dots are reconstructed without and with photon related algorithms.

from semi-leptonic decay of heavy quarks. The di-jet energy is sampled at 91, 200, 360 and 500 GeV. Shown in figure 5.10, the jet energy resolutions are much better at 360 and 500 GeV with improved photon reconstruction. By identifying photons before reconstructing charged particles in a dense jet environment, the reconstruction task is easier and less likely to make mistake. However, at low energy, the reconstruction is worse with photon algorithms, because photon algorithms are optimised for high energies.

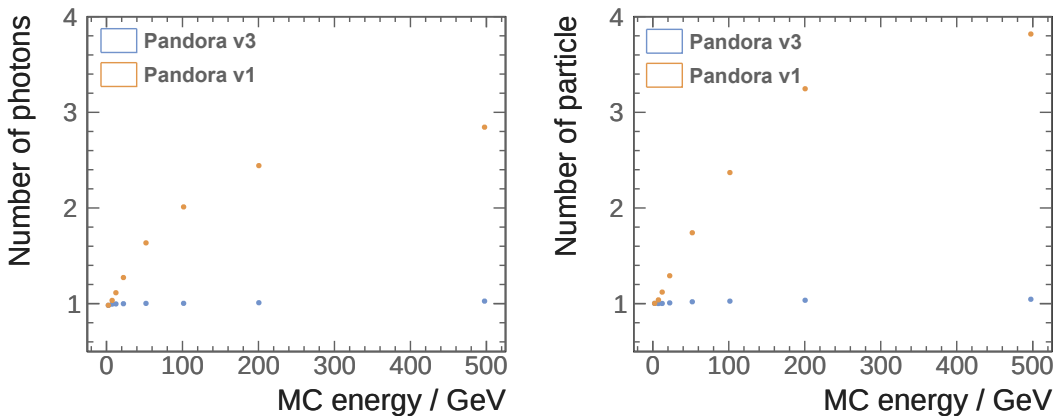
To access the impact of photon algorithms on jet energy resolution, perfect photon reconstruction is used to compare the performances, which identifies calorimeter hits from truth information. Same jet samples are used. Photon confusion, which is defined as the quadrature difference between a normal reconstruction and a perfect photon confusion, is listed in table 5.5. Photon confusion term, except for $\sqrt{s} = 91$ GeV, has been reduced to 0.9% with the photon algorithms.

Photon confusion	$\sqrt{s} = 91$ GeV	200 GeV	360 GeV	500 GeV
PandoraPFA without photon algorithms / %	0.7	0.9	1.3	1.4
PandoraPFA with photon algorithms / %	1.4	0.9	0.9	0.9

Table 5.5: Photon confusion as a function of energy for reconstruction with and without photon algorithms.

5.10 Compare with photon reconstruction in PandoraPFA version 1

This section reviews the performance improvement with the introduced algorithms, using single photon, photon pair and jet samples. There is a photon reconstruction algorithm in PandoraPFA version 1. Since the changes to the photon reconstruction is made in PandoraPFA version 2, version 3 contains all the improvement of the photon reconstruction. This section concentrates on the improvement from version 1 to version 3.

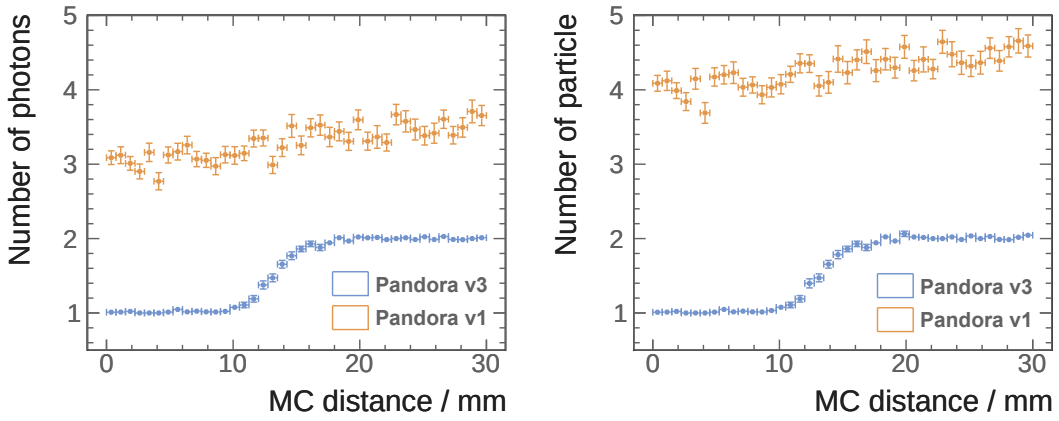


(a) (b)

Figure 5.11: Figure 5.11a and figure 5.11b shows the average number of reconstructed photons and reconstructed particles, as a function of their true energy using a single photon per event sample. The top orange and bottom blue dots are reconstructed with PandoraPFA version 1 and version 3. The photon reconstruction is changed in PandoraPFA version 2.

Figure 5.11a shows the reduction in fragments identified as photons, using a single photon per event sample. For the blue dots, average number of photon stays below 1.05 even at 500 GeV (true value 1). For a 100 GeV photon, the average number of photons is reduced to 1 from 2. For a 500 GeV photon, the number is reduced to 1.05 from 2.8.

A similar trend is shown in figure 5.11b, where the extra fragments identified as neutral hadrons have also taken into account. For a 100 GeV photon, the average number of particles is reduced to 1 2.4. For a 500 GeV photon, the number is reduced to 1.05 from 3.8.



(a) (b)

Figure 5.12: Figure 5.12a and figure 5.12b shows the average number of reconstructed photons and reconstructed particles, as a function of the MC distance separation in the calorimeter, using two photons of 500 and 50 GeV per event sample. The top orange and bottom blue dots are reconstructed with PandoraPFA version 1 and version 3. The photon reconstruction is changed in PandoraPFA version 2.

Figure 5.12 illustrates a reduction in the photon fragments and the neutral hadron fragments using two photons of 500 and 50 GeV per event sample. The figure shows the MC distance separation from 0 to 30 mm, which corresponds to approximately 6 ECAL square cells. This is a difficult test for fragment removal as high energy photons are more likely to create fragments and the imbalance in the two photon energies makes it more difficult to separate correctly. In both figure 5.12a and figure 5.12b, the average numbers of photon and particle are below 2.05 at 30 mm apart, which is significantly better than the reconstruction in PandoraPFA version 1. Two photons start to be resolved at 10 mm apart, and fully resolved at 20 mm apart. The resolution is better than reconstruction in PandoraPFA version 1, which is difficult to extract due to excess fragments.

Another metric to reflect the improvement in photon reconstruction is the fraction of the fragment energy to the total energy as function of the distance separation. Shown in figure 5.13, using two photons of 500 and 50 GeV per event sample, a reduction in fragment energy can be seen clearly. With improved reconstruction, the average fragment energy fraction is below 0.1% up to 30 mm apart, whilst around 5% energy would be reconstructed in fragments with PandoraPFA version 1.

The improvement in completeness and resolution in photon reconstruction, as shown in single photon and double photon reconstruction, leads to a small improvement in the jet energy resolution at high energy. The di-jet energy is sampled at 91, 200, 360

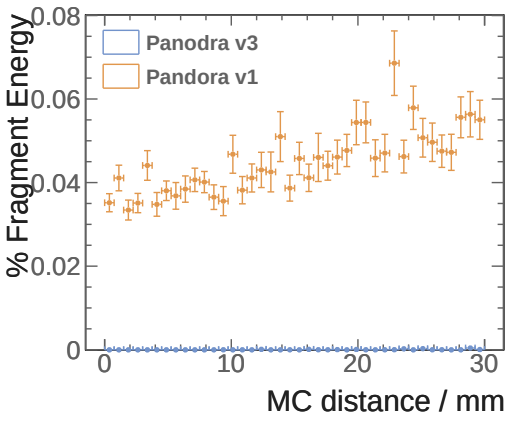


Figure 5.13: Average fraction of fragments energies to the total energy, as a function of the MC distance separation in the calorimeter, using two photons of 500 and 50 GeV per event sample. The top orange and bottom blue dots are reconstructed with PandoraPFA version 1 and version 3 respectively. The photon reconstruction is changed in PandoraPFA version 2.

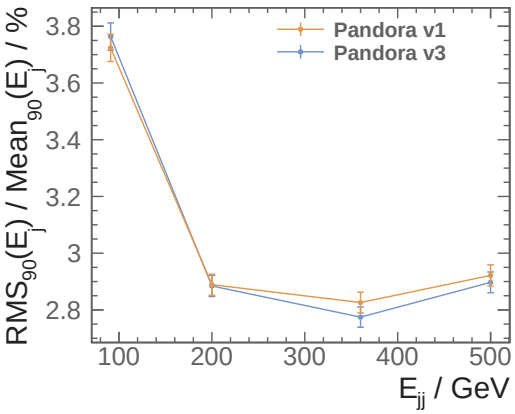


Figure 5.14: Jet energy resolution as a function of the di-jet energy using $Z' \rightarrow u/d/s$ sample at barrel region. The top orange and bottom blue dots are reconstructed with PandoraPFA version 1 and version 3. The photon reconstruction is changed in PandoraPFA version 2.

and 500 GeV. Shown in figure 5.14, the jet energy resolutions are better at 360 and 500 GeV with improved photon reconstruction. This is due to more aggressive photon reconstruction especially nearby tracks, which is more useful at a high-energy dense jet environment.

The improvement of the photon is also demonstrated in chapter 6, where tau lepton decay modes are classified. Excellent photon reconstruction leads to a high classification rate.

5.11 Understand photon reconstruction improvement

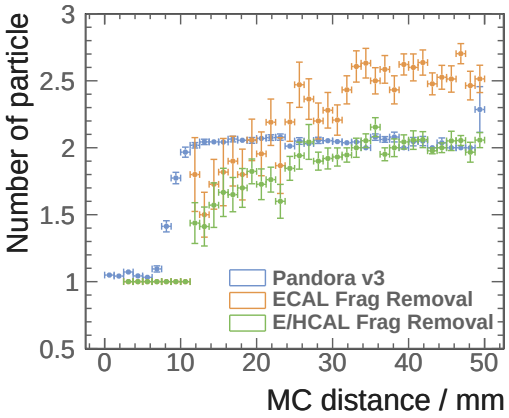


Figure 5.15: Figure shows the average number of photons, as a function of the MC distance separation in the calorimeter, using two photons of 500 and 500 GeV per event sample. The blue, orange, and green dots are reconstructed with PandoraPFA version 3, PandoraPFA version 1 with fragment removal in the ECAL, and PandoraPFA version 1 with fragment removal in the ECAL and the HCAL. The photon reconstruction is changed in PandoraPFA version 2.

As stated before, photon reconstruction algorithm in section 5.3 and photon splitting algorithm in section 5.8 improves the photon completeness and the photon pair resolution. The fragment removal algorithm in section 5.6 removes fragments in the ECAL. High energy fragment removal algorithm in section 5.7 removes fragments in the HCAL. To show the incremental improvement, the average number of particle for a high energy photon pair, 500 - 500 GeV is shown in figure 5.15.

With fragment removal algorithm in the ECAL, the number of fragment is reduced significantly (comparing with figure 5.12b). With the energy fragment removal in the HCAL, the number of fragments are reduced further. At 40 mm apart, with both fragment removal algorithms (green dots), there is less than 0.05 fragment per photon pair.

The introduction of the revised photon reconstruction and photon splitting improves the photon separation resolution. Photons pair starts to be resolved at 5 mm apart for 500 - 500 GeV pair and fully resolved at 15 mm apart. With previous photon reconstruction in PandoraPFA version 1, the same photon pair starts to be resolved at 10 mm apart and fully resolved at around 40 mm apart.

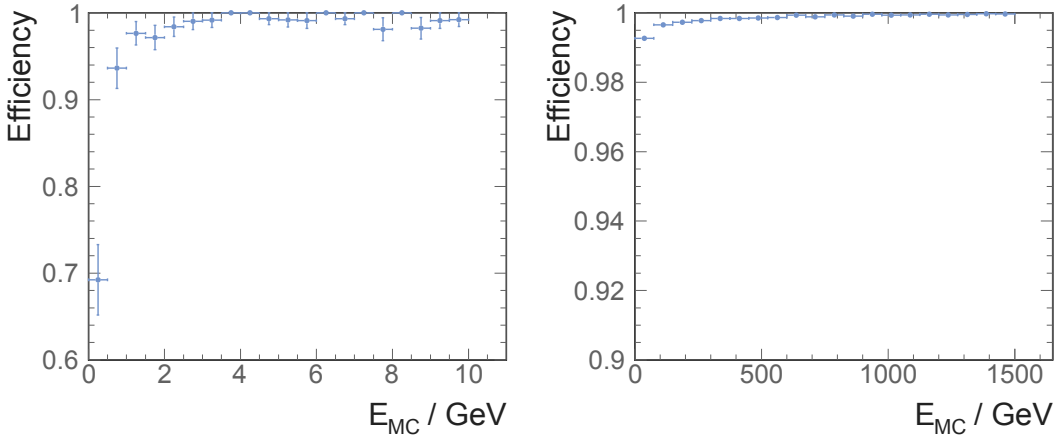
5.12 Current photon reconstruction performance

In section 5.10, the improved performance of the photon reconstruction is demonstrated with different metrics, using single photon, double photons and jet samples. In this section, the performance of the photon reconstruction as a function of photon energies will be described.

Single particle efficiency is demonstrated in figure 5.16. Using single photon samples, an event can have an efficiency of 1, if the photon reconstructed corresponds to the truth photon, or an efficiency of 0. The low efficiency in the first bin, from 0 to 0.25 GeV is because photon reconstruction does not attempt to reconstruct photons below 0.2 GeV. The single photon reconstruction efficiency is above 98% for photons above 2GeVand above 99.5% for photons above 100 GeV.

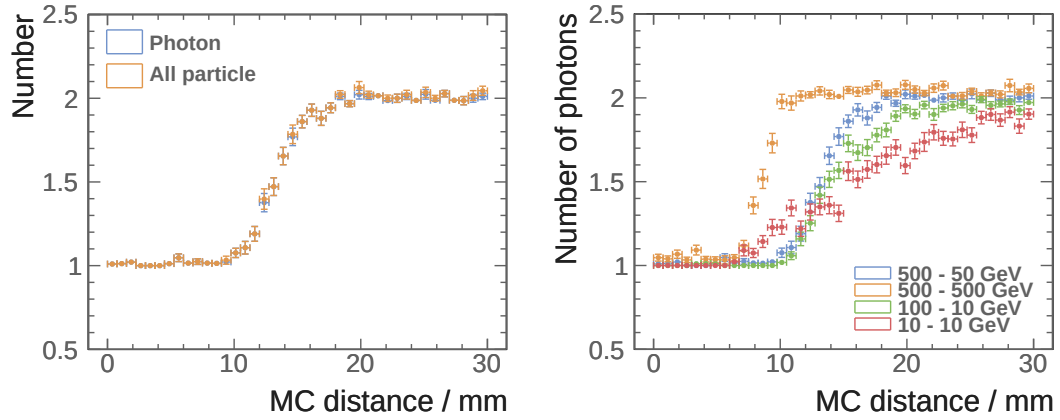
For simple samples such as two photons per event, there are very few fragments. Shown in figure 5.17a for 500 and 50 GeV photons pair sample, the average number of photons and particles beyond 20 mm apart is less than 0.05 above the true value, 2.

The resolving power of a photon pair depends on energies of two photons. Figure 5.17b is an example of average number of photon reconstructed for differen photon pairs. When the energies of two photons are similar, the resolving power is greater. This is because that the two photon showers have similar sizes, and the 2D Peak Finding algorithm can exploit the symmetry in the size of the EM showers. For example, 500 - 500 GeV photon pair and 10 - 10 GeV photon pair start to be resolved at 6 mm apart, which is about 1 ECAL cell. The asymmetrical photon pair, 500 - 50 GeV and 100 - 10 GeV pair, starts to be resolved at 10 mm apart, which is about 2 ECAL cells.



(a) **(b)**

Figure 5.16: Single photon reconstruction efficiency as a function of energy. figure 5.16a shows the low energy region and figure 5.16b shows high energy region.



(a) **(b)**

Figure 5.17: Figure 5.17a shows the average numbers of photon and particle using two photons of 500 and 50 GeV per event sample. Figure 5.17b shows the average numbers of photon for four different photon pairs: 500 - 50, 500 - 500, 100 - 10 and 10 - 10 GeV.

For an energetic photon, it is more difficult to remove fragments, but it is easier to identify the photon. The electromagnetic shower core is more dominant than the peripherals. Therefore separating two energetic photons is easier than separating two low energy photons. This can be seen in figure 5.17b. At 20 mm apart, two photons in 500 - 500 GeV pair are fully resolved, where approximately only 60% of two photons in 10 - 10 GeV pair are resolved.

Chapter 6

Tau Lepton Final State Classification

"MVA: Turn numbers into gold."
— TMVA

The tau lepton has been studied extensively in the past at the Large Electron Positron Collider (LEP) [67]. The tau lepton spin state, which can be derived from kinematic properties of its decay products, can be used to measure the CP (the product of charge conjugation and parity symmetries) of the Higgs, via $H \rightarrow \tau^+\tau^-$ channel [68]. The polarisation correlation of the tau pairs can be used to infer the spin of the parent boson, departing $H \rightarrow \tau^+\tau^-$ from $Z \rightarrow \tau^+\tau^-$.

The ability to identify tau decay mode is also a benchmark for detector performance. Since tau lepton has a very short life time [69], only its decay products can be detected via the tracking detectors and calorimeters. Therefore the performance of the calorimetric and track systems determines the ability to separate different tau decay modes.

The main difficulty in the classification is to classify 1-prong final states. These final states involves π^0, where two photons from π^0 decay could be poorly reconstructed. At high energy, π^0 is boosted making the reconstruction more challenging. The ability to reconstruct the two photons as separate entities requires good pattern recognition algorithm for photons and ECAL spatial resolution. Hence the improved photon reconstruction in chapter 5 is used in this study. The impact of the ECAL transverse spatial resolution on the tau decay classification is demonstrated as well.

This chapter is organised in the following sections. Firstly, the analysis chain to identify tau decay modes will be described, followed by the ECAL optimisation study using the decay mode classification as a benchmark. Lastly, the classification is further used in a proof-of-principle analysis to demonstrate the ability to identify H from Z using tau pair decay channel.

6.1 Overview of the analysis

The study of the tau final state classification in the context for the ECAL optimisation allows the analysis to discard reconstruction and detector issue that do not vary with the ECAL design. For example, the early photon conversion that happens in the tracking detector would complicate the event topology. But since it is affected by the tracker design, it can be ignored in this analysis. Section 6.3 and section 6.4 discuss the pre-selection cuts to choose the signal samples and events for this analysis.

The classification is performed with a multivariate classifier. Discriminating variables are calculated before feeding into the classifier. A multiclass classicisation is used to allow simultaneous classification between multiple final states, described in section **??**.

This classification is repeated for different energies of tau lepton decay to access the impact of energy on classification. The impact of the ECAL design is studied afterwards, where the ECAL square cell size is varied. An overall tau hadronic decay classification efficiency is constructed to allow direct and easy comparison between different detector design and different energies of the tau decay.

A proof-of-principle analysis to demonstrate the ability to identify H from Z using tau pair decay channel is presented in the section **??**. The difference in the spin of the boson reflects in the different spin correlation of the tau pair. By extracting the spin correlation, parent bosons can be separated.

The follow sections on the analysis use the a 50 GeV tau lepton decaying sample, reconstructed with nominal the ILD detector model.

6.2 Decay modes

Tau lepton decays into a numerous number of final states. To study the predominant effect of the tau lepton decays, decay modes with branching ratio above 1% are studied. This results in seven tau lepton decay modes. Their branching ratios, along with decay mode and final states are shown in table 6.1, which in total covers 92.58 % tau decay. The most difficult final states to separate are $\pi^-\ \pi^0$ and $\pi^-\ \pi^0\ \pi^0$, where photons from boosted π^0 are very challenging to reconstruct correctly.

Decay modes	Detectable final states	Branching ratio
$e^-\ \overline{\nu}_e\ \nu_\tau$	e^-	17.83 ± 0.04
$\mu^-\ \overline{\nu}_\mu\ \nu_\tau$	μ^-	17.41 ± 0.04
$\pi^-\ \nu_\tau$	π^-	10.83 ± 0.06
$\rho\ \nu_\tau$	$\pi^-\ \pi^0$	25.52 ± 0.09
$a_1 \nu_\tau$	$\pi^-\ \pi^0\ \pi^0$	9.30 ± 0.11
$a_1 \nu_\tau$	$\pi^+\ \pi^-\ \pi^-$	8.99 ± 0.06
$\pi^+\ \pi^-\ \pi^-\ \pi^0\ \nu_\tau$	$\pi^+\ \pi^-\ \pi^-\ \pi^0$	2.70 ± 0.08

Table 6.1: Decay modes, detectable final state particles and branching ratios of the seven major τ^- decays, taken from [2]. τ^+ decays similarly to τ^-.

6.3 Simulation and reconstruction

$e^+e^- \rightarrow \tau^+\tau^-$ channel is chosen to study the tau lepton decay mode classification, due to its the simplest channel for tau decay. The choice of the detector model for simulation is ILD. The initial state radiation (ISR) and the beam induced background were not simulated, as they are not significant. These beam specific effects does not vary with the ECAL design. Hence they are not simulated for the detector optimisation study. Another reason for not simulating ISR is that the energy of the tau lepton would not be a constant with the ISR contribution.

Two million events were reconstructed with iLCSoft version v01-17-07 [70] and PandoraPFA version 3 [38], where the photon reconstruction is described in chapter 5.

6.4 Event pre-selection

Some events are very difficult or almost impossible to reconstruct correctly. This can happen if particles are in the forward beam pipe direction, where the forward calorimeters are not simulated due to computational limitations, or if particles carry too little energies to be more significant than noises. Another type of reconstruction failure is when a photon converts into an electron pair in the tracking systems. Instead of reconstructing as a photon, an electron pair is reconstructed, which alters the event topology and decrease the classification efficiencies. One more issue is that for the ILD detector model concept, the gap region between barrel part of the calorimeter and the end cap part causes a significant drop in the particle reconstruction efficiency. The PandoraPFA particle reconstruction does not attempt to recover reconstruction in the gap region. These reconstruction and detector issues do not vary with change of the ECAL design in the later optimisation study. Therefore, these un-reconstructible events are discarded at MC level from the analysis. The selection cuts are listed in table 6.2 and the selection efficiency of each cut is presented in table 6.3.

Cuts	Values
No photon early conversion	-
Visible energy of decay products	$E_{vis,MC} > 5\,GeV$
Polar angle acceptance	$0.6 > \lvert\theta_{Z,MC}\rvert > 0.3$ or $1.57 > \lvert\theta_{Z,MC}\rvert > 0.8$

Table 6.2: Pre-selection cuts for tau lepton decay final state classification.

The reason for no photon early conversion in the tracker has been stated above. The polar angle acceptance requires the MC tau lepton travelling to barrel or endcap only. The visible energy cut requires there is enough energy deposited in the calorimeter to have a proper reconstruction. The visible energy of the MC tau lepton decay product is defined as the energy of the MC tau minus the energy of the MC tau neutrino. These cuts use MC truth information. Discarded events do not participate in the subsequent reinstruction and analysis.

The no photon early conversion cuts only effective against final states with π^0, as π^0 decays to a photon pair. For final states with one π^0, about 77% events survived. For final states with two π^0, about 61% events survived, which is roughly 0.77^2. For the visible angle acceptance, final states with only one particle are affected the most, whilst

Final state	No photon conversion	Visible energy acceptance	Polar angle acceptance
e^-	100.0	84.7	66.2
μ^-	100.0	85.2	66.7
π^-	100.0	88.3	60.9
$\pi^- \pi^0$	77.1	76.9	61.9
$\pi^- \pi^0 \pi^0$	61.3	61.2	50.5
$\pi^+ \pi^- \pi^-$	100.0	100.0	78.0
$\pi^+ \pi^- \pi^- \pi^0$	77.0	77.0	61.8

Table 6.3: MC level pre-selection cut efficiencies for tau lepton decay final state classification in percentages. Cuts are presented in a "flow" fashion, where each column contains all the cuts in columns to its left.

final states with more than one particles typically have more visible energies. The polar angle acceptance efficiencies depend on the final states, as light final states are boosted and more likely in the forward region.

6.5 Select single tau decay

The $e^+e^- \to \tau^+\tau^-$ channel contains two tau leptons travelling in opposite directions. Since the tau decay final state separation is applicable to single tau, PFOs in one event are divided into two collections, each corresponding to one tau. Separation of PFOs uses the principle thrust axis vector, explained in the section 4.6.4, which is the axis that most PFOs aligned to. Two collections are obtained based on the sign of the scalar product between the thrust axis and the PFO momentum vector. An example of the simulated $e^+e^- \to \tau^+\tau^-$ event is shown in figure 6.1.

6.6 Discriminative variables

Having pre-selected events, a set of discriminating variables are carefully developed for the multivariate analysis (MVA). The full list of variables are shown in table 6.4.

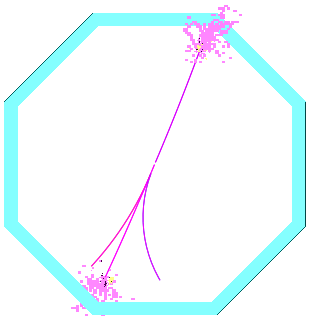


Figure 6.1: An example event display of simulated $e^+e^- \to \tau^+\tau^-$ event. The top tau decay is $\pi^-\ \pi^0$ final state and the bottom is $\pi^+\ \pi^-\ \pi^-\ \pi^0$ final state. Purple clusters are $\pi^\pm$ and yellow clusters are photons. Blue region is the transverse cross section of the ECAL barrel, looking along the beam line direction.

Category	Variable
Number of PFOs	N_{χ^+}, N_μ, N_e, N_γ, N_{π^-}
Invariant mass	m_{vis}, m_{χ^+}, m_{χ^0}, m_γ, m_{π^-}
Calorimetric info.	$\%E_{\chi^+}$, $\%E$
Energy	$\tilde{E}_{vis}$, $\tilde{E}_{\chi^+}$, $\tilde{E}_\mu$, $\tilde{E}_e$, $\tilde{E}_\gamma$, $\tilde{E}_{\pi^-}$
$\rho(\pi^-\pi^0)$ reconstruction	$m_{\pi^0}(\rho)$, m_ρ
$a_1(\pi^-\pi^0\pi^0)$ reconstruction	$m_{\pi^0}(a_1)$, $m^*_{\pi^0}(a_1)$, m_{a_1}
EM shower profile	δl, t_0, $\langle w \rangle$
Calorimeter hit info.	$\bar{E}_{hit}$, $\%MIP$
Track info.	$\Delta E/P$

Table 6.4: Variables used in the MVA

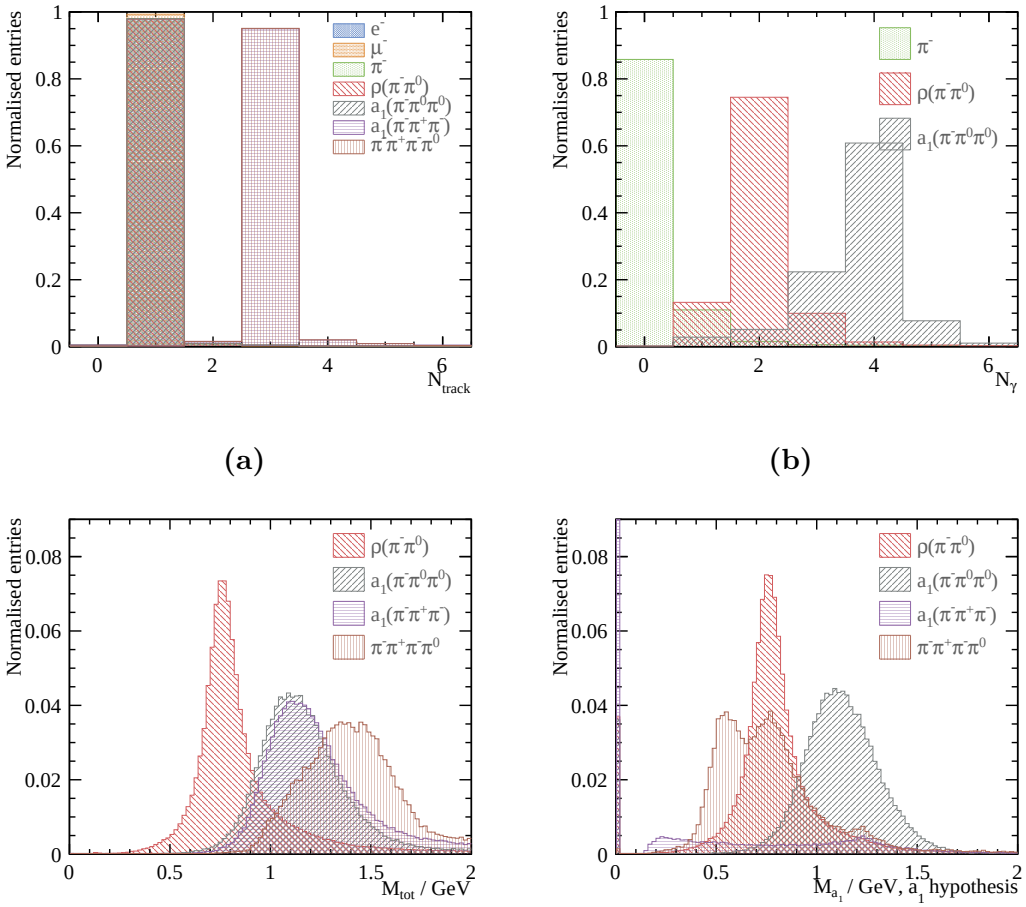


Figure 6.2: Normalised distribution for selected variables showing for seven final states, e^- , μ^- , π^- , $\rho(\pi^-\pi^0)$, $a_1(\pi^-\pi^0\pi^0)$, $a_1(\pi^+\pi^-\pi^-)$ and π^+ π^- π^- π^0 , separated using truth information. Figure 6.2a, figure 6.2b, figure 6.2c, and figure 6.2d show distributions for the number of tracks, the number of photons, the invariant mass of visible PFOs, and the invariant mass of a_1 for $a_1(\pi^-\pi^0\pi^0)$ hypothesis test, respectively. The area for each final state is normalised to 1.

The most crucial variables are the number of PFOs of different particles. Figure 6.2a shows the distribution of N_{χ^+}, the number of charged PFOs for different tau final states. Whilst over 98% 1-prong final states have one track reconstructed, around 95% 3-prong final states have three tracks reconstructed. This is an excellent variable to separate 1-prong and 3-prong final states. An orthogonal measurement is the number of reconstructed photons, N_γ, shown in figure 6.2b. The overlap between π^- and $\rho(\pi^-\pi^0)$ final states is about 15% and the confusion between $\rho(\pi^-\pi^0)$ and $a_1(\pi^-\pi^0\pi^0)$ is around 15%. N_γ can also separate two 3-prong final states. N_μ, N_e, N_{π^-} are useful to identify two leptonic final states, and further separate 3-prong final states from 1-prong final states.

Invariant masses of different particles are good at characterising different final states. Figure 6.2c shows the invariant mass of the system. Clear reasonable peaks can be seen for ρ and a_1. The mass peak of $\pi^+\,\pi^-\,\pi^-$ are much higher. m_{χ^+} and m_{χ^0} are invariant masses of charged and neutral particles respectively. They separate final states with neutral particles from those without neutral particles. Similarly, m_γ and m_{π^-} identify final states with photons and with π^- respectively.

Calorimetric information is used to identify electron and muon. An electron deposits most energy in the ECAL and a muon deposits 5 to 20% energy in the ECAL. Two variables, $\%E_{\chi^+}$ and $\%E$, are the fraction of energy deposited in the ECAL over total energy in the calorimeter, of charged particles and all PFOs respectively. Energy fraction for the charged particles does not include contribution from photons, which also deposit most energy in the ECAL, like electrons.

Energy information further separate different final states. Variables are normalised to the expect tau energy. For example, $\tilde{E}_\gamma$, normalised energy of photons are different for final states with and without photons.

Extra variables are used to differentiate specific final states.

6.6.1 $\rho(\pi^-\pi^0)$ and $\rho(\pi^-\pi^0)$ resonances reconstruction

$\rho(\pi^-\pi^0)$ and $a_1(\pi^-\pi^0\pi^0)$ are identified further using their resonance structure. For example, $\rho(\pi^-\pi^0)$ final state contains a π^- and a π^0 decaying to two photons. By selecting π^- and photons consistent with ρ decay pattern, $\rho(\pi^-\pi^0)$ final state could be

separated. The particle selection is via minimising a χ^2 function, defined as:

$$\chi^2 = \left(\frac{m_{tot} - m_{\rho}^{MC}}{\sigma_{\rho}^{MC}}\right)^2 + \left(\frac{m_{\gamma\gamma} - m_{\pi^0}^{MC}}{\sigma_{\pi^0}^{MC}}\right)^2, \tag{6.1}$$

where $m_{\gamma\gamma}$ is the invariant mass of two photons and m_{tot} is the invariant mass of the photon pair and one π^-. All combinations of photons and π^- are tested. m_{ρ}^{MC} and $m_{\pi^0}^{MC}$ are expected masses of ρ and π^0, taken from [2]. σ_{ρ}^{MC} and $m_{\pi^0}^{MC}$ are the half width of the invariant mass distribution of reconstructed ρ and π^0 using the truth information. This minimisation allows $\rho(\pi^-\pi^0)$ final state to be separated from $a_1(\pi^-\pi^0\pi^0)$, where failure to reconstruct photons causes $a_1(\pi^-\pi^0\pi^0)$ events to be simualr to that of $\rho(\pi^-\pi^0)$.

Similarly, $a_1(\pi^-\pi^0\pi^0)$ can be separate using a extended minimisation with two extra photons:

$$\chi^2 = \left(\frac{m_{tot} - m_{a_1}^{MC}}{\sigma_{a_1}^{MC}}\right)^2 + \left(\frac{m_{\gamma1\gamma2} - m_{\pi^0}^{MC}}{\sigma_{\pi^0}^{MC}}\right)^2 + \left(\frac{m_{\gamma3\gamma4} - m_{\pi^0}^{MC}}{\sigma_{\pi^0}^{MC}}\right)^2, \tag{6.2}$$

where ρ has been replace by a_1. Selected particles for the test are two photon pairs and one π^-. Both photon pairs are required to be consistent with π^0 decaying to two photons. Figure 6.2d shows the distribution of m_{tot} from $a_1(\pi^-\pi^0\pi^0)$ consistency test for selected final states. The main feature is that only $a_1(\pi^-\pi^0\pi^0)$ final state has a mass peak at a_1 resonance position. Comparing to simple invariant mass distribution in figure 6.2c, $a_1(\pi^-\pi^0\pi^0)$ mass peak is enhanced.

The χ^2 functions for both consistency test are adapted for cases where reconstruction fails to resolve photon pairs. Relevant terms are dropped from the expression if there are fewer photons than required.

6.6.2 Separate e^- from π^-

Particle ID reported from PandoraPFA is used extensively to reconstruct discriminating variables. PandoraPFA uses a wide range of information to determine electron ID. However, extra information is used in this analysis to help identifying electrons, which are mistaken as π^- by PandoraPFA reconstruction.

An electron leaves a characteristic electromagnetic (EM) shower in the ECAL (see section 5.2), whilst π^- doesn't. Variables defining the EM shower helps to identify e^-,

taken from photon reconstruction in section 5.5. t_0, the start layer of the longitudinal shower and δl, the fractional difference between observed and expected longitudinal shower profile describe the longitudinal EM shower. $\langle w \rangle$ is a measure of the EM shower transverse width.

Another type of information to differentiate an EM shower from a e^- and a early hadronic shower from a π^- is the calorimeter hit information. $\bar{E}_{hit}$, the average energy of a hit, and %MIP, fraction of possible minimum ionising particles, are different for EM and hadronic showers. Last information used is the track-calorimeter consistency check. $\Delta E/P$ is the calorimeter energy divided by the track momentum. This variable is found to help to differentiate e^- from π^- final state.

6.7 Multivariate Analysis

For the multivariate analysis, the multiclass class of the TMVA package [71] was used to perform a multiclass classification, which trains the seven final states simultaneously. The multiclass class is an extension of the standard two-class signal-background classifier. The discussion on multivariate analysis can be found in section 4.7. The multiclass classifier is discussed in section 4.7.3.

The multiclass classifier used is Boosted Decision Tree with Gradient boost (BDTG). The optimisation of the BDTG classifier follows the strategy in section 4.7.1. The optimised parameters are listed in table 6.5. Explanation of variables can be found in section 4.7.2. Half of the randomly selected samples were used in the training process and the other half were used for testing.

6.8 Classification Efficiency

The reconstruction efficiencies for the seven tau decay final states are shown in table 6.6. Bold numbers show the correct classification probability. For example, 99.8% of true e^- are reconstructed correctly.

For the e^- decay mode, a high correct classicisation is achieve, due to using particle ID from PandoraPFA and calorimetric information. Similarly for μ^- decay mode, 99.5%

Parameter	Value
Depth of tree	5
Number of trees	3000
Boosting	gradient boost
learning rate of the gradient boost	0.1
metric for the optimal cuts	Gini Index
bagging fraction	0.5
Number of bins per variables	100
End node output	yes/no

Table 6.5: Optimised parameters for the Boosted Decision Tree with Gradient boost multiclass classifier. See section 4.7.2 for detailed explanation of variables.

correct rate is achieved, due to efficient tracking detector and muon reconstruction algorithm.

For the π^- decay mode, 3.4% confusion with $\rho(\pi^-\pi^0)$ decay mode is to due to differentiate two event topologies. Around 15% of π^- events have at least one photon reconstructed, mostly due to the FSR. If the reconstruction is unable to identify the photon pair from π^0 decay in $\rho(\pi^-\pi^0)$ decay mode, two decay modes would appear to similar and confusion is caused. The confusion with e^- is at percent level, which is low due to the usage of EM shower variables. The percent level confusion with $a_1(\pi^+\pi^-\pi^-)$ is because the tracking efficiency is at 98%, where 2% π^- events have more than one track reconstructed.

For the $\rho(\pi^-\pi^0)$ decay mode, biggest confusion is with $a_1(\pi^-\pi^0\pi^0)$ because of the similar reason of unable to reconstruct all photons form π^0 decaying.

For the $a_1(\pi^-\pi^0\pi^0)$ decay mode, the correct classicisation rate is the lowest as it is most challenging to reconstruct correctly: two photon pairs and one $\pi^\pm$. The 9.5% confusion with $\rho(\pi^-\pi^0)$ is due to the same reconstruction failure issue. It should be noted that figure 6.2b suggests that 30% of $a_1(\pi^-\pi^0\pi^0)$ events have fewer than four photons reconstructed, which overlap with $\rho(\pi^-\pi^0)$ distribution. The $a_1(\pi^-\pi^0\pi^0)$ resonance reconstruction (section 6.6.1) and the multiclass classifier reduce the confusion to 9.5% from 30%.

For the $a_1(\pi^+\pi^-\pi^-)$ decay mode, most confusion is with π^+ π^- π^- π^0 , due to inability of separating photons. And the reason is the same for the confusion of π^+ π^- π^- π^0

decay mode classified as $a_1(\pi^+\pi^-\pi^-)$ decay mode. The unprecedented high classification rate has been achieved. The improvement of photon reconstruction described in section **??** improved the ability to separate 1-prong final state. Most notably, figure **??** shows number of photons have a high correct reconstruction efficiency.

Reco ↓ Truth →	e^-	μ^-	π^-	$\rho(\pi^-\pi^0)$	$a_1(\pi^-\pi^0\pi^0)$	$a_1(\pi^+\pi^-\pi^-)$	$\pi^+\ \pi^-\ \pi^-\ \pi^0$
e^-	**99.7**	-	0.9	0.6	0.4	-	-
μ^-	-	**99.5**	0.6	-	-	-	-
π^-	-	0.3	**94.0**	0.8	-	0.4	-
$\rho(\pi^-\pi^0)$	-	-	3.4	**93.6**	9.5	0.6	2.3
$a_1(\pi^-\pi^0\pi^0)$	-	-	-	4.5	**89.7**	-	0.6
$a_1(\pi^+\pi^-\pi^-)$	-	-	0.9	-	-	**96.8**	6.4
$\pi^+\ \pi^-\ \pi^-\ \pi^0$	-	-	-	0.3	-	2.0	**90.6**

Table 6.6: Classification efficiency in percentage for tau decay modes using the nominal ILD detector model , for $\sqrt{s} = 100\,\text{GeV}$. Bold numbers show the correct classification probability. Numbers below 0.25%, and ν_τ are not shown in decay modes, for display purposes. Statistical uncertainties are less than 0.25%.

6.9 Electromagnetic calorimeter optimisation

In the above section, an analysis on tau decay mode classification is presented. Events are $e^+e^- \to \tau^+\tau^-$ at $\sqrt{s} = 100\,\text{GeV}$ with nominal ILD detector model. This classification can be applied for different $\sqrt{s}$ and with different detector models. The main reconstruction issue is to separate boosted photon pair from π^0 decay. High $\sqrt{s}$ would photon pair more collimated, and therefore more challenging to separate. The ECAL cell size is crucial in separating the EM showers from photons. As discussed in previous chapter, separating photons requires a high transverse spatial resolution. Therefore, the ECAL square cell size is expected to affect the classification efficiency.

The main difficulty in the classification is to classify 1-prong final states. These final states involves π^0, where two photons from π^0 decay could be poorly reconstructed. At high energy, π^0 is boosted making the reconstruction more challenging. The ability to reconstruct the two photons as separate entities requires good pattern recognition algorithm for photons and ECAL spatial resolution. Hence the improved photon recon-

struction in chapter 5 is used in this study. The impact of the ECAL transverse spatial resolution on the tau decay classification is demonstrated as well.

The analysis is repeated with varying ECAL square cell sizes at 3, 5, 7, 10, 15 and 20 mm, at four $\sqrt{s}$ = 100, 200, 500, 1000 GeV. The multivariate classifier is trained for each cell size and each $\sqrt{s}$. Other ECAL dimensions are kept the same as the ILD nominal detector. Because the lepton reconstruction mostly relies on the tracking system, which is not varied in this study, only the hadronic tau decays were investigated and compared between different ECAL square cell sizes.

As PandoraPFA is optimised for the nominal ILD detector, fragment removal algorithms have a large dependence on the ECAL cell sizes. One algorithm is re-optimised for the varying ECAL cell sizes. The PhotonFragmentRemoval algorithm which merges photon fragments uses a distance metric and it is optimised with values listed in the table 6.7. As cell sizes become larger, the distance cut for merging photons are larger.

ECAL square cell size (mm)	3	5	7	10	15	20
ClosestHitDistance	5	10	10	10	20	20

Table 6.7: Optimised parameters of PhotonFragmentRemoval algorithm as a function of ECAL square cell size.

Figure 6.3 shows the correct classification efficiencies for tau hadronic decay final states as a function of the ECAL square cell sizes. For example, plotted values for 5 mm cell size at $\sqrt{s}$ = 100 GeV are the same as the bold numbers in table 6.6.

In general, the correct classification efficiencies generally decreases with increase of $\sqrt{s}$ and increase of ECAL cell sizes. As the $\sqrt{s}$ increases, tau decay products are boosted. It is increasingly difficult to separate identical decay products. For example, the photon pair from π^0 decay become very challenging to separate at high $\sqrt{s}$. Increase of the ECAL cell sizes has a similar effect. With a bigger cell size, photons may deposit energy in a same cell, making separating these photons almost impossible.

For the π^- decay mode, the general trend is followed. The efficiency for $\sqrt{s}$ = 200 GeV is slightly better than that of the $\sqrt{s}$ = 100 GeV for small cell sizes.

For the $\rho(\pi^-\pi^0)$ decay mode,the efficiency for $\sqrt{s}$ = 500 GeV increases as the cell sizes increases. This is because the multivariate classifier optimises for the overall classification

efficiency, which may compensate one decay mode for another. In this case, the small increase in efficiency for $\rho(\pi^-\pi^0)$ at $\sqrt{s} = 500\,\text{GeV}$ is compensated by the drastic decrease in efficiency for $a_1(\pi^-\pi^0\pi^0)$ at $\sqrt{s} = 500\,\text{GeV}$.

For the $a_1(\pi^-\pi^0\pi^0)$ decay mode, the loss of efficiency with increase of the cell size and increase in $\sqrt{s}$ is most significant comparing to other decay modes. With most number of particles in the final state, it is the most challenging decay channel to reconstruct and thus most sensitive to the change in cell sizes and $\sqrt{s}$.

For the $a_1(\pi^+\pi^-\pi^-)$ decay mode, the efficiencies are similar to that of the π^- . Both final states contain charged particles only. Therefore it is most sensitive to the tracking performance, which is not affected by the ECAL cell sizes.

For the $\pi^+\ \pi^-\ \pi^-\ \pi^0$ decay mode, the decrease in efficiencies are more significant for $\sqrt{s} = 500\,\text{TeV}$ and $1000\,\text{GeV}$.

The leptonic decay correct reconstruction efficiency is not used as a metric as they are similar across different ECAL cell sizes. This is because the $e^\pm$ and $\mu^\pm$ identifications mostly rely on the tracking system, which was not varied in this study. The energy deposited in the calorimeter are used for the association to the tracks but it has a small impact on the lepton identification.

6.9.1 Tau hadronic decay correct classification efficiency

There are two reasons for constructing a single parameter for the overall Tau decay efficiency. First is that multivariate classifier is trained to optimised for the overall classification efficiency. Second is that it is easier to compare the impact of different detector models and different $\sqrt{s}$. The choice of the hadronic decay is because reconstructing hadronic decays is sensitive to the ECAL cell sizes, which is relevant for this study.

The constructed Tau hadronic decay correct classification efficiency, ε_{had}, is a weighted classification efficiency for all hadronic decay modes:

$$\varepsilon_{\text{had}} = \frac{\sum_i^5 \text{Br}_i \varepsilon_i}{\sum_i^5 \text{Br}_i}, \tag{6.3}$$

where Br_i is the branching fraction of the hadronic decay mode i after the generator level cut (section 6.4). ε_i is the correct reconstruction efficiency of the decay mode i, and i is summing over five tau hadronic decay modes.

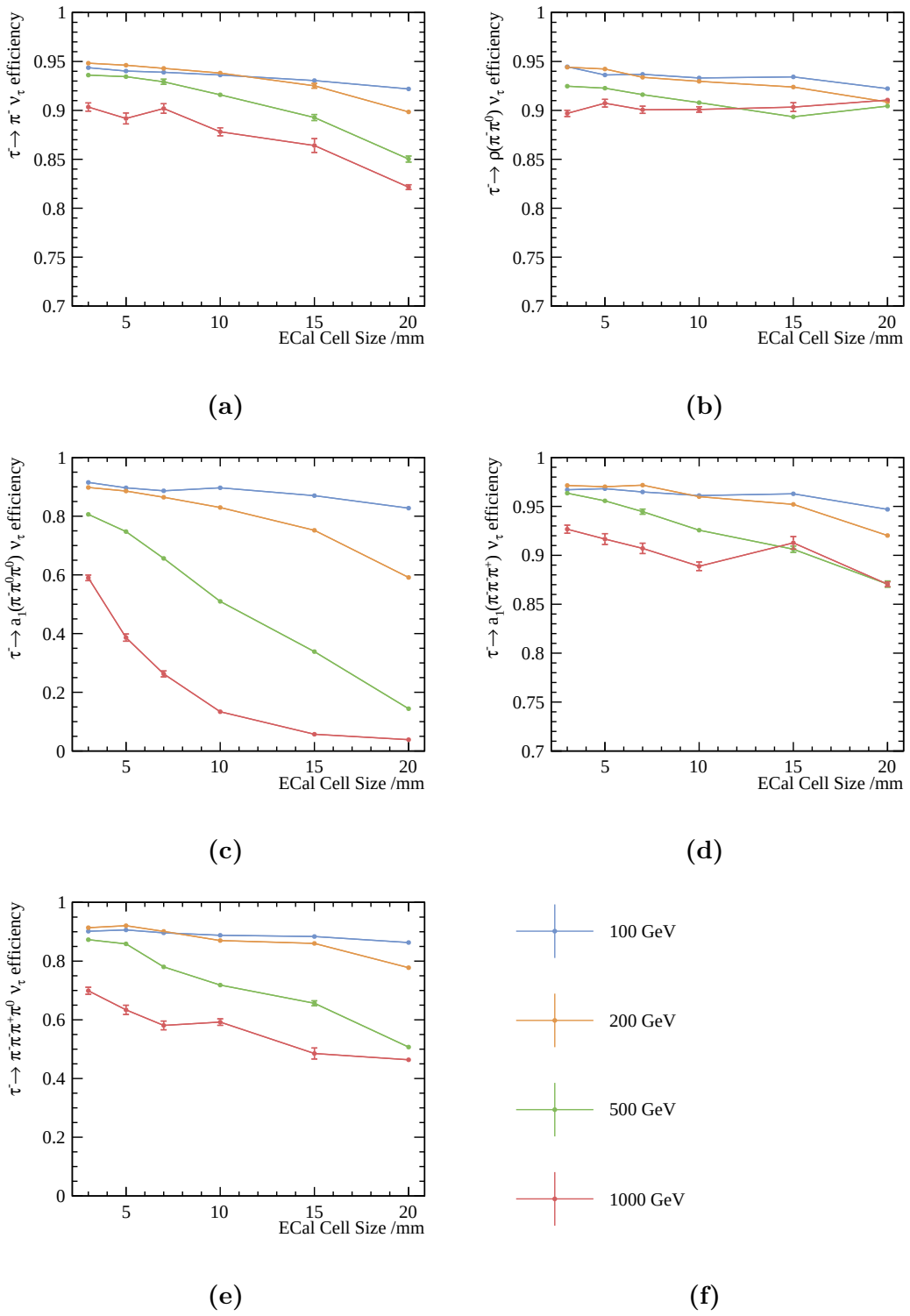


Figure 6.3: The correct classification efficiencies for tau hadronic decay final states as a function of the ECAL square cell sizes, using the ILD detector model with $\sqrt{s}$ = 100, 200, 500 and 1000 GeV. Figure 6.3a, 6.3b, 6.3b, 6.3c, and 6.3d show the π^- , $\rho(\pi^-\pi^0)$, $a_1(\pi^-\pi^0\pi^0)$, $a_1(\pi^+\pi^-\pi^-)$, π^+ π^- π^- π^0 decay modes, respectively.

Figure 6.4 shows ε_{had} as a function of ECAL cell sizes with different $\sqrt{s}$. The general trend for the ε_{had} is that ε_{had} decreases with increase of $\sqrt{s}$ and increase of ECAL cell sizes for the same reasons stated in the previous section. As the $\sqrt{s}$ increases, tau decay products are boosted and it is challenging to separate identical decay products. Similarly, increasing ECAL cell sizes makes particle separation more difficult.

For ε_{had} at $\sqrt{s} = 100\,\text{GeV}$, the efficiency decreases from 94% at 3 mm cell size, to 91% at 20 mm cell size. The decrease is approximately proportional to the increase in the cell size.

For ε_{had} at $\sqrt{s} = 200\,\text{GeV}$, the efficiency decreases from 94% at 3 mm cell size, to 86% at 20 mm cell size.

For ε_{had} at $\sqrt{s} = 500\,\text{GeV}$, the efficiency decreases from 92% at 3 mm cell size, to 78% at 20 mm cell size. Most significant decrease occurs at this $\sqrt{s}$.

For ε_{had} at $\sqrt{s} = 1000\,\text{GeV}$, the efficiency decreases from 85% at 3 mm cell size, to 75% at 20 mm cell size.

The increase in ECAL cell sizes has a larger impact on high $\sqrt{s}$. With decay products spatially close at high $\sqrt{s}$, it is more beneficial to have a smaller ECAL cell size to reconstruct individual particle.

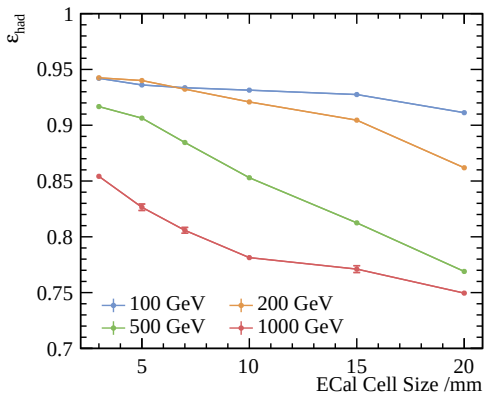


Figure 6.4: The tau hadronic decay efficiency, ε_{had}, as a function of the ECAL cell sizes at different $\sqrt{s}$ with the nominal ILD detector model. The blue, orange, green and red lines are representing the $\sqrt{s} = 100$, 200, 500 and 1000 GeV respectively.

6.10 Separate H from Z with tau pair decay

H can be separated from Z using tau pair decay channel. The difference in the spin of the boson reflects in the different polarisation correlation of the tau pair. By extracting the polarisation correlation, parent bosons can be separated. A proof-of-principle analysis is presented to demonstrate the physics usage of the tau decay mode classification, motivated by theoretical studies such as in [72]. The theoretical motivation is described in section 2.10. The subsequent sections discuss the ability to reconstruct the polarisation correlation of the Z using tau pair decay, and to match with the truth information.

The analysis largely follows the same procedure for the tau decay mode classification. Differences are highlighted in sections below.

The channel is $e^+e^- \to ZZ$, where one Z decays hadronically and the other Z decays to a tau lepton pair. The samples were generated at $\sqrt{s} = 350\,\text{GeV}$ without ISR contribution for this proof-of-principle study.

6.10.1 Event pre-selection

Same seven tau decay modes in section 6.2 are studied. The $\tau^- \to \pi^-\nu_\tau$ decay mode is used to demonstrate tau pair polarisation correlation.

The event pre-selection is similar to that in section 6.4. The cut on the visible energy of decay products is not used because a large fraction of $Z \to \tau^+\tau^-$ events where $\tau^- \to \pi^-\nu_\tau$ have two low-energy $\pi^\pm$. Therefore, the cut on the visible energy of decay products would bias the distribution.

6.10.2 Identify tau pairs

The final state of $e^+e^- \to ZZ$ channel contains two tau leptons and two quark jets. Therefore, tau decay products can either found by direct tau searching, or by finding two jets and working out the recoil. If a tau lepton decays to a few particles, then the direct tau searching would work. If a tau lepton decays to many particles, finding tau decay products as a jet has a better performance. Here two approaches are combined for the best tau pair identification.

Direct tau search

Tau finder processer, IsolatedTauIdentifer, is developed by the author, which is a modified version from the one in section 7.3.2. The basic idea is to find tau decay products consistent with decay topologies, and require the decay products to be isolated from rest of particles in an event. Parameters chosen are loose to find as many as possible tau candidates. The filtering of the candidates is via kinematic constraints.

The modified IsolatedTauIdentifer works as following. Low p_T are not considered. A seed particle is chosen and a search cone is formed around the seed, which requires one or three tracks with the search cone invariant mass less than 3 GeV. The isolation criteria states that the opening angle between the search cone and the 2^{nd} closest track is larger than 0.6 rad. If all criterion are satisfied, the search cone with the tau seed is identified as a tau lepton.

Modified IsolatedTauIdentifer	Selection
Veto low p_T (GeV)	$p_T < 0.5$
Seed particle (GeV)	$p_T > 1$
Maximum search cone opening angle (rad)	$\theta_S \leqslant \cos^{-1}(0.99)$
Tau candidate rejection	$N_{X^+} \neq 1, 3,\ m_{PFO} > 3$
Isolation (rad)	$\theta_{cone,2^{nd}X^+} > 0.6$

Table 6.8: Optimised parameters of modified IsolatedTauIdentifer.

Jet clustering

Tau hadronically decay can be also identified as a small jet. Durham algorithm [51] was used to jet clustering, also known as e^+e^- k_t algorithm. (see section 4.6.2). The jet algorithm runs in the exclusive mode to find four jets.

Select tau candidates

The best tau pair candidates are selected using kinematic constraints. Half of $\sqrt{s}$ is in the Z. The invariant mass of two quarks from Z should be close to Z mass. Therefore,

the minimisation function is

$$\chi^2 = \frac{\left(m_{qq} - m_Z\right)^2}{\sigma^2_{m_{qq}}} + \frac{\left(E_{qq} - \frac{\sqrt{s}}{2}\right)^2}{\sigma^2_{E_{qq}}}, \tag{6.4}$$

where m_Z is the mass of Z from reference [2]. $\sigma_{m_{qq}}$ and $\sigma_{E_{qq}}$ are the reconstructed resolution of mass and energy of the Z, respectively. m_{qq} and E_{qq} are defined differently for the direct tau searching and the jet clustering method. For the direct tau search method, m_{qq} and E_{qq} are defined as the recoil against two tau candidates, assuming collision at $\sqrt{s}$, iterating over all tau candidates. For the jet clustering method, m_{qq} and E_{qq} are defined as the mass and energy of two jets, iterating over all possible jets.

The χ^2 minimiser is repeated for the direct tau search and the jet clustering method. Each method produces a best tau pair candidate. Hence two tau pair candidates are obtained. To find the over best tau candidate, a set of conditions is used. If best tau pair candidates for both methods satisfy the kinematic constraint:

$$\left|m_{qq} - m_Z\right| < \sigma_{m_{qq}} \ , \left|E_{qq} - \frac{\sqrt{s}}{2}\right| < \sigma_{E_{qq}}, \tag{6.5}$$

the pair with smallest χ^2 is chosen.

If only one candidate satisfies the constraint in equation 6.5, that candidate is chosen. If none of the candidates satisfies the constraint, and if one jet from the jet clustering is close to the beam pipe and there are exactly two tau leptons from IsolatedTauIdentifer, then these two tau leptons are chosen. This is because if one jet is close to the beam pipe, it is likely that some particles are undetected, which leads to a failure in the kinematic constraint. Lastly, if all conditions above are not satisfied, two smallest jets by number of PFOs are chosen to be tau leptons decay products.

Boost to Z decay rest frame

To use the tau decay mode classifier, it is necessary to know the tau lepton energy before decaying. For the channel Z decaying to a tau pair, the tau energy can be calculated in Z decay rest frame, which is half of the Z energy. To calculate the energies of tau decay products in the rest frame, the decay products need to be boosted to the rest frame, which requires the four-momentum of the Z, the tau pair system.

The previous section describes the method to identify the tau pair decay products. The four-momentum of the Z decaying to tau pair is calculated from the recoil of non tau decay products:

$$p^{\mu}_{\tau\tau} = \begin{pmatrix} \sqrt{s} \\ \sqrt{s} \times \sin(\theta_{beam}) \\ 0 \\ 0 \end{pmatrix} - \sum_{i}^{non-\tau} p^{\mu}_{i}, \tag{6.6}$$

where θ_{beam} is the beam crossing angle. Index i is summing over all non tau decay PFOs. Extra kinematic constraint fixes the energy of the $p^{\mu}_{\tau\tau}$ to be half of $\sqrt{s}$:

$$p^{\mu}_{\tau\tau,correct} \equiv p^{\mu}_{\tau\tau} \times \frac{\frac{1}{2}\sqrt{s}}{E_{\tau\tau}}. \tag{6.7}$$

$p^{\mu}_{\tau\tau,correct}$ is used as the boost vector to boost tau decay products in the rest frame. The calculation of the variables for the MVA classifier are preformed in the rest frame.

6.10.3 Variables

Variables for the MVA classifier are largely the same as the ones in table 6.4. Variables regarding EM shower profiles, calorimeter hit information and track information are not used (last three rows in table 6.4) as the study focus on the overall tau decay mode separation. Also for the computational reason, it was not feasible to use these variables for the MVA. e^+ and π^+ separation could be improved if these extra variables are included.

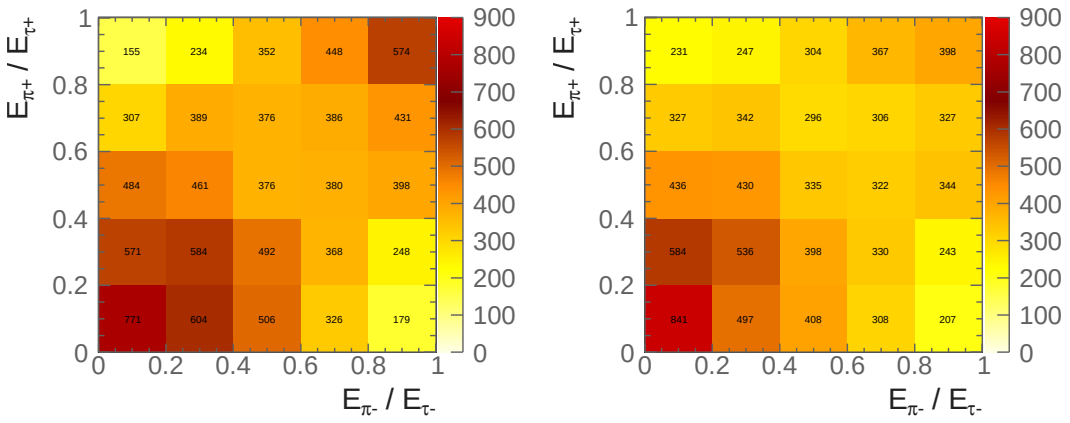
6.10.4 Multivariate analysis

Half of the sample is used to train the multivariate classifier, which follows the procedure in section 6.7. In the applying stage, $\tau^- \to \pi^- \nu_\tau$ decay mode is selected with additional criteria that at least one $\pi^{\pm}$ is identified in the tau decay products.

6.10.5 Result

Figure 6.5 shows the two-dimensional plot of tau pair polarisation correlations from Z decay, using $\tau^- \to \pi^- \nu_\tau$ decay mode. The energy fractions are the appreciate kinematic variables, motivated in section 2.10. Figure 6.5a shows the distribution using the truth information. Figure 6.5b shows the distribution using full detector simulation. In the $Z \to \tau^+\tau^-$ decays, an energetic $\pi^\pm$ is likely to be associated with an energetic $\pi^\pm$, shown in both figure 6.5a and figure 6.5b. Comparing two figures, some events in the top right quadrant, resembling both $\pi^\pm$ being energetic, are not reconstructed correctly. This is due to the incorrect identification of the tau pair decay products.

This proof-of-principle analysis shows the tau polarisation correlations with $Z \to \tau^+\tau^-$ decay where $\tau^- \to \pi^- \nu_\tau$ can be observed. With a similar study of $H \to \tau^+\tau^-$, such the tau polarisation correlations can be used to sperate H from Z.



(a) Truth info. **(b)** Simulated

Figure 6.5: Two-dimensional plot of tau pair polarisation correlations from Z decay, using $\tau^- \to \pi^- \nu_\tau$ decay mode.

Chapter 7

Double Higgs Bosons Production Analysis

"Two is better than one"
— Sir Steve Orange, 1785–1854

Since the discovery of Higgs boson at the LHC in 2012 [1,17], it is crucial to understand the properties of the Higgs boson and test if it is a Standard Model Higgs. The Higgs mechanism and the Higgs boson in the Standard Model have been explained in chapter 2. Beyond the Standard Model, a number of theories may be tested via the double Higgs production in an electro-positron collider (see section 2.9). Generator level studies have shown that the precision reached by a multi-TeV linear collider, such as the Compact Linear Collider (CLIC), is superior to that of the Large Hadron Collider (LHC) - even with $3000\mathrm{fb}^{-1}$ of data [13].

The first challenge for the double Higgs bosons production analysis is that events are rare. The cross section is very small compared to other background physics processes, making it difficult to select signal events. The second challenge is that at high centre-of-mass, events are boosted and many particles are in the forward region of the detector, where the reconstruction performance is inferior to the barrel region and particles can escape detection.

In this chapter, a full CLIC_ILD detector simulation study has been performed for the double Higgs production, $e^-e^+ \to HH\nu_e\overline{\nu}_e$, via WW fusion. Event generation and simulation will be discussed first. An overview of the analysis, including lepton finding

and jet reconstruction, is presented. This is followed by a multivariate analysis with results on the selection efficiency. The results of the signal selection are interpreted in the context of the Higgs self coupling. The results of this analysis have been publish in reference [15].

7.1 Analysis Straggly Overview

The leading order Feynman diagrams of double Higgs production via WW fusion are shown in figure 7.1. Figure 7.1a is sensitive to Higgs triple self coupling g_{HHH}. Figure 7.1b is sensitive to quartic coupling g_{WWHH}. Figure 7.1c and figure 7.1d are the irreducible background processes for the study of g_{HHH} and g_{WWHH}.

Another channel of the double Higgs production is the $e^-e^+ \to ZHH$, where Z decays to $\nu\,\overline{\nu}$. This ZHH channel can also be used to study the Higgs triple self couplings, and has been studied at the ILC for $\sqrt{s} = 500\,\text{GeV}$ [20]. However, its contribution to the $HH\nu\overline{\nu}$ final state is small compared to the WW fusion in figure 7.1, for the relevant CLIC energies $\sqrt{s} = 1.4\,\text{TeV}$ and $3\,\text{TeV}$. The cross section of $e^-e^+ \to ZHH$ is one order of magnitude smaller than $e^-e^+ \to HH\nu_e\overline{\nu}_e$ via the WW fusion, shown in figure 2.3,. Therefore, the effect of $e^-e^+ \to ZHH$ present in $e^-e^+ \to HH\nu_e\overline{\nu}_e$ channel at $\sqrt{s} = 1.4\,\text{TeV}$ and $3\,\text{TeV}$ is negligible.

The double Higgs production $e^-e^+ \to HH\nu_e\overline{\nu}_e$ is divided into sub-channel $HH \to b\overline{b}W^+W^-$ and $HH \to b\overline{b}b\overline{b}$ to target the specific kinematic properties if each final state, which provides cross-validation between two sub-channels and an improvement in signal selection when combined. In this chapter, sub-channel $HH \to b\overline{b}W^+W^-$ is investigated. Firstly the hadronic decay mode of the $HH \to b\overline{b}W^+W^-$ channel is studied. It is chosen because the hadronic decay has the largest cross section and does not produce neutrinos, which allows each W to be reconstructed using di-jet. The semi-leptonic final state is also considered. However, extra neutrinos in the final states present greater difficulty to reconstruct the two Higgs bosons, as some momenta of one Higgs boson is missing. This channel will be discussed briefly and is adapted from the hadronic decay analysis.

The $HH \to b\overline{b}b\overline{b}$ sub-channel is studied independently by collaborators. However, there is collaboration between the two studies. The two analyses are combined on the final couplings extractions.

The signal channel final state, $HH\nu_e\bar{\nu}_e \to b\bar{b}W^+W^-\nu_e\bar{\nu}_e \to b\bar{b}qqqq\nu_e\bar{\nu}_e$, is a six quark final state with missing momentum. The high number of quarks requires an efficient jet reconstruction and a jet pairing algorithm to select the signal. The two b quarks in the final states allow the analysis to use b jet tagging. Since the final state does not contain leptons, event-level lepton finding - typically for energetic isolated leptons - would improve the signal selection efficiency.

A proof-of-principle, generator-level study was performed at CLIC using CLIC_ILD detector model at $\sqrt{s} = 1.4\,\text{TeV}$ and $3\,\text{TeV}$ [19]. In this chapter a full CLIC_ILD detector simulation study is presented. Firstly, suitable signal and background channels are identified. In order to select the signal, events with light lepton and tau lepton are vetoed. B-jet tagging information is used to identify b quark jets. PFOs in an event are clustered into jets depending on the final state followed by pre-selection cuts and multivariate analysis. This analysis optimised independently according to the two $\sqrt{s} = 1.4\,\text{TeV}$ and $\sqrt{s} = 3\,\text{TeV}$. The $HH \to b\bar{b}W^+W^-$ hadronic decay will be presented first, followed by the semi-leptonic sub-channel analysis.

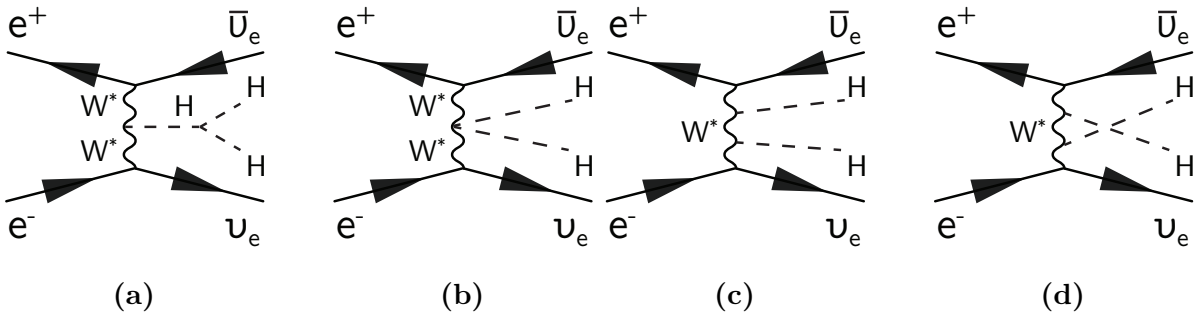


Figure 7.1: Figures show Feynman diagrams of leading-order $e^-e^+ \to HH\nu_e\bar{\nu}_e$ processes at CLIC, without considering $e^-e^+ \to ZHH$. Figure 7.1a is sensitive to the Higgs triple self coupling g_{HHH}. Figure 7.1b is sensitive to quartic coupling g_{WWHH}.

7.2 Monte Carlo Sample Generation

Background samples considered in this analysis are listed in Table 7.1. The signal channel is $e^-e^+ \to HH\nu_e\bar{\nu}_e \to b\bar{b}W^+W^-$ where both W's decay hadronically.

Background processes with many quarks with missing energies would be challenging to reject. Two examples are $e^-e^+ \to qqqq\nu\bar{\nu}$ and $e^\pm\gamma \to \nu qqqq$. Single Higgs boson production, such as $e^-e^+ \to qqH\nu\bar{\nu}$, would also be difficult to remove.

Some background channels are not considered because they either have different event topologies, or they have very small cross sections. For example, $e^{\pm}\gamma() \to qqH\ell$ is ignored as the cross section is very small - even at $\sqrt{s} = (\,\text{TeV}\ 3)$ (0.07 fb with photon from EPA, 0.6 fb with photon from BS). Some background channels are not simulated due to computational limitations.

Electron-photon and photon-photon interactions are considered in this analysis. They are important as their interactions become significant at high $\sqrt{s}$. These photons are produced due to the high electric field generated by the colliding beams. Processes involving real photons from bremsstrahlung (BS) and "quasi-real" photons are generated separately. For the "quasi-real" photon initiated processes, the Equivalent Photon Approximation (EPA) has been used.

For processes involving Higgs production explicitly, simulated Higgs mass is 126 GeV. For other processes, they are generated assuming a Higgs mass of 14 TeV. This is to produce negligible double Higgs production cross sections to decouple the Higgs production. The cross section of the signal, such as $HH \to b\bar{b}W^+W^-$, is corrected manually according to [73], as the internal value used in the event generation software (PYTHIA) is inaccurate.

The simulation and reconstruction chain are described in chapter 4. For most background processes, events are simulated requiring invariant mass of quarks are above 50 GeV. For electron-photon interaction with $qqqq\nu$ final state at $\sqrt{s} = (\,\text{TeV}\ 1.4)$, events are simulated requiring invariant mass of quarks above 120 GeV.

Finally, the main beam induced background $\gamma\gamma \to hadrons$ is simulated and overlayed to all samples. Details can be found in section 4.5.2.

7.3 Lepton identification

The reconstruction is done via Marlin in iLCSoft v01-16. The latest functioning flavour tagging processor exist in iLCSoft v01-16. Thus newer versions of iLCSoft can not be used in this analysis. Separate software packages (processors) exist for lepton identification and for jet reconstruction. New processors have been developed and existing processors have been optimised for signal selection and background rejection.

Channel	$\sigma(\sqrt{s} = 1.4\,\text{TeV})$ / fb
$e^-e^+ \to HH\nu_e\overline{\nu}_e$	0.149
$e^-e^+ \to HH\nu_e\overline{\nu}_e \to b\overline{b}W^+W^-$,hadronic	0.018
$e^-e^+ \to HH\nu_e\overline{\nu}_e \to b\overline{b}b\overline{b}\nu_e\overline{\nu}_e$	0.047
$e^-e^+ \to HH \to$ others	0.085
$e^-e^+ \to q_l q_l H\nu\overline{\nu}$	0.86
$e^-e^+ \to c\overline{c}H\nu\overline{\nu}$	0.36
$e^-e^+ \to b\overline{b}H\nu\overline{\nu}$	0.31
$e^-e^+ \to qqqq$	1245.1
$e^-e^+ \to qqqq\ell\ell$	62.1*
$e^-e^+ \to qqqq\ell\nu$	110.4*
$e^-e^+ \to qqqq\nu\overline{\nu}$	23.2*
$e^-e^+ \to qq$	4009.5
$e^-e^+ \to qq\ell\nu$	4309.7
$e^-e^+ \to qq\ell\ell$	2725.8
$e^-e^+ \to qq\nu\nu$	787.7
$e^-\gamma(BS) \to e^-qqqq$	1160.7
$e^+\gamma(BS) \to e^+qqqq$	1156.3
$e^-\gamma(EPA) \to e^-qqqq$	287.1
$e^+\gamma(EPA) \to e^+qqqq$	286.9
$e^-\gamma(BS) \to \nu qqqq$	79.8†
$e^+\gamma(BS) \to \overline{\nu}qqqq$	79.3†
$e^-\gamma(EPA) \to \nu qqqq$	17.4†
$e^+\gamma(EPA) \to \overline{\nu}qqqq$	17.3†
$e^-\gamma(BS) \to qqH\nu$	15.8*
$e^+\gamma(BS) \to qqH\nu$	15.7*
$e^-\gamma(EPA) \to qqH\nu$	3.39*
$e^+\gamma(EPA) \to qqH\nu$	3.39*
$\gamma(BS)\gamma(BS) \to qqqq$	21406.2*
$\gamma(BS)\gamma(EPA) \to qqqq$	4018.7*
$\gamma(EPA)\gamma(BS) \to qqqq$	4034.8*
$\gamma(EPA)\gamma(EPA) \to qqqq$	753.0*

Table 7.1: List of signal and background samples with the corresponding cross sections at $\sqrt{s} = 1.4\,\text{TeV}$. q can br u, d, s, b or t. Unless specified, q, ℓ and ν represent particles and its corresponding anti-particles. γ (BS) represents a real photon from beamstrahlung (BS). γ (EPA) represents a "quasi-real" photon, simulated with the Equivalent Photon Approximation. For processes involving Higgs production explicitly, simulated Higgs mass is 126 GeV. Otherwise, Higgs mass is set to 14 TeV. For processes labeled with * and †, the generator-level cut requires invariant mass of quarks greater than 50 and 120 GeV, respectively.

For the signal channel, $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$, there is no lepton in the final state, whilst many background final states contain primary leptons, such as $qqqq\ell\nu$. Hence effectively rejecting events with leptons would improve the signal selection efficiency. Primary leptons leave tracks which originate very close to the interaction point. They are typically energetic and often isolated from other particles. Whilst electrons and muons are stable enough to deposit energies in calorimeters, tau leptons are very short lived and decay in the tracker. Therefore, only decay products of the tau leptons can be reconstructed. These characteristics provide the basis for the isolated lepton finding.

7.3.1 Electron and muon identification

In this section, two processors are described followed by their performances. As the signal is very rare comparing to the background, it is necessary to develop high performance isolated lepton finder to veto events with leptons and improve the signal selection efficiency. An existing lepton finder is optimised in section 7.3.1 and a separate lepton finder is developed by the author in section 7.3.1.

IsolatedLeptonFinderProcessor

IsolatedLeptonFinderProcessor in Marlin package is used, modified, and optimised. This processor identifies high energy electrons and muons that are isolated from other particles'. The optimal parameters below are chosen in collaboration and tested using the signal channel and the $e^-e^+ \to qqqq\ell\nu$ channel.

Electron induced electromagnetic showers are mostly contained in the ECAL, while muons would deposit some energies in the ECAL. The inner detector tracks of electrons and muons from the electron-positron interaction are primary tracks, which are very close to the interaction point. The isolation criteria requires the lepton to be spatially separated from other high energy particles. This forms the logic behind the IsolatedLeptonFinderProcessor. Optimal values of the processor are listed in table 7.2. E_{ECAL} is the energy deposited in the ECAL. E_{cone} is the total energy of PFOs within a cone of an opening angle of $\cos^{-1}(0.995)$ around the lepton. d_0, z_0, and r_0 are the Euclidean distance of the track starting point to the interaction point in x-y plane, in z direction, and in x-y-z three dimensional space. Performance of the processor is shown in table 7.6.

IsolatedLeptonFinderProcessor	Selection
High Energy (GeV)	$E > 15$
$e^{\pm}$ ID	$\frac{E_{ECAL}}{E} > 0.9$
$\mu^{\pm}$ ID	$0.25 > \frac{E_{ECAL}}{E} > 0.05$
Primary Track (mm)	$d_0 < 0.02, z_0 < 0.03, r_0 < 0.04$
Isolation (GeV)	$E^2_{cone} \leqslant 5.7 \times E - 50$

Table 7.2: Optimised parameters of IsolatedLeptonFinderProcessor

IsolatedLeptonIdentifer

IsolatedLeptonFinderProcessor has strict criterion to find high-energy isolated leptons to avoid mistakes. However, since the signal cross section is low in this analysis, it would be beneficial to reject more events with leptons identified to improve the signal to background ratio. Hence another isolated lepton finder is developed. The main feature of the IsolatedLeptonIdentifer is that it utilises calorimetric information provided by PandoraPFA.

The processor uses two sets of cuts. Its logic is similar to that of the IsolatedLeptonFinderProcessor. The first set of cuts uses the particle ID information from PandoraPFA. The cuts demand a PandoraPFA electron or muon with high p_T and is from a primary track. The lepton should either have a very high transverse momentum or be isolated. The second set of cuts uses ECAL energy fraction to determine light lepton ID. The rest of the cuts are very similar to the first cuts. The isolation criterion is stricter than it in the first set to reduce fake rate.

Table 7.3 lists values for the IsolatedLeptonIdentifer lepton selection cut. Variables are defined similarly as those in section 7.3.1. p_T is the transverse momentum. E_{cone1} and E_{cone2} are the total energy of PFOs within a cone of an opening angle of $\cos^{-1}(0.995)$ and $\cos^{-1}(0.99)$ respectively around the lepton. The performance of the processor is shown in table 7.6.

Comparison: IsolatedLeptonFinderProcessor versus IsolatedLeptonIdentifer

The two processors share similar criterion for light lepton identification. The main difference is that the IsolatedLeptonIdentifer uses particle identification from

IsolatedLeptonIdentifer	Selection
High Energy (GeV)	$E > 10$
$e^{\pm}$ ID	PandoraPFA reconstructed & $\frac{E_{ECAL}}{E} > 0.95$
$\mu^{\pm}$ ID	PandoraPFA reconstructed
Primary Track (mm)	$r_0 < 0.015$
a) High Transverse Momentum (GeV)	$p_T > 40$
b) Isolation (GeV)	$E \geqslant 23 \times \sqrt{E_{cone1}} + 5$
High Energy (GeV)	$E > 10$
$e^{\pm}$ ID	$\frac{E_{ECAL}}{E} > 0.95$
$\mu^{\pm}$ ID	$0.2 > \frac{E_{ECAL}}{E} > 0.05$
Primary Track (mm)	$r_0 < 0.5$
a) High Transverse Momentum (GeV)	$p_T > 40$
b) Isolation (GeV)	$E \geqslant 28 \times \sqrt{E_{cone2}} + 30$

Table 7.3: Optimised parameters of IsolatedLeptonIdentifer

PandoraPFA, which takes into account extra calorimetric information to determine the particle ID than simple ECAL energy fraction. IsolatedLeptonIdentifer also allows high p_T light leptons to be identified in a non-isolated environment, which leads to the more aggressive nature of the IsolatedLeptonIdentifer.

7.3.2 Tau identification

Tau leptons have a short life-time and decay before reaching the detector and can only be identified through the reconstruction of their decay products. The leptonic decay of tau can be identified using the two isolated lepton finder processors in section 7.3.1 and section 7.3.1. Therefore tau identification will focus on the hadronic decay. To improve signal selection, a tau lepton identifier is developed by the author in section 7.3.2.

The basic logic of a tau finder is similar to a cone clustering algorithm (see section 4.4.3). A high energy track is selected as a seed and a small cone is formed around it. The PFOs inside the cone are required to be consistent with a tau hadronic decay: no more than 3 charged particles, invariant mass close to tau mass and few PFOs in the cone. The cone then forms a tau candidate. Like the isolated lepton in the isolated light lepton finder, the tau candidate is required to be isolated from other particles. To reduce

fake rate, low momentum and very forward particles do not participate in the tau finding, as they more likely come from $\gamma\gamma \to \mathtt{hadrons}$ background.

TauFinderProcessor

TauFinderProcessor [74], a processor in Marlin package, has been tuned in collaboration and tested. The tuned parameters are listed in table 7.4. Variables are defined similarly as in previous sections. θ_Z is the polar angle w.r.t. the beam axis. $N_{\chi^+} > 3$ and N_{tau} are the number of charged particles and the number of PFOs respectively in the tau candidate. m_{tau} is the invariant mass of the sum of the PFOs in the tau candidate. E_{cone} is the total energy of PFOs within a cone of an opening angle between 0.03 and 0.33 rad around the tau seed. The performance of the processor is shown in table 7.6.

TauFinderProcessor	Selection
Veto $\gamma\gamma \to \mathtt{hadrons}$ (GeV)	$p_T < 1$, $\lvert\cos(\theta_Z)\rvert > 1.1$
Seed particle (GeV)	$p_T > 10$
Tau candidate cone opening angle (rad)	0.03
Tau candidate rejection	$N_{\chi^+} > 3$, $N_{tau} > 10$, $m_{tau} > 2$
Isolation (GeV)	$E_{cone} < 3$

Table 7.4: Optimised parameters of TauFinderProcessor

IsolatedTauIdentifer

For the similar reason of developing a more aggressive light lepton finder, a new more aggressive tau lepton selection processor is developed by the author. IsolatedTauIdentifer identifies a high momentum particle as a tau seed. Particles are iteratively added to the search cone according to the opening angle to the seed. After each particle addition, the temporary search cone is then considered as a temporary tau candidate and required to be isolated and consistent with tau hadronic decay signature. The temporary tau candidate only needs to pass one of the isolation conditions. The iterative particle addition would stop when the cone opening angle is bigger than a threshold. If multiple temporary tau candidates of the same seed pass the selection, the one with smallest opening angle is chosen to form the final tau candidate. To reduce fake tau

decay products from $\gamma\gamma \to \mathtt{hadrons}$ background, low energy particles do not participate in the tau finding.

Table 7.4 lists the optimised parameters for IsolatedTauIdentifer. Variables are defined similarly as those in previous sections. θ_S is the opening angle of the search cone. $\mathtt{cone1}$ and $\mathtt{cone2}$ are defined as a cone of an opening angle of $\cos^{-1}(0.95)$, and $\cos^{-1}(0.99)$ respectively around the tau seed. r_0 is referring to tau seed particle. The performance of the processor is shown in table 7.6.

IsolatedTauIdentifer	Selection
Veto $\gamma\gamma \to \mathtt{hadrons}$ (GeV)	$E < 1$
Seed particle (GeV)	$p_T > 5$
Maximum search cone opening angle (rad)	$\theta_S \leqslant \cos^{-1}(0.999)$
Tau candidate rejection	$N_{X^+} \neq 1, 3$, $m_{PFO} > 3$
Isolation (GeV) 1	$N_{cone1} = 0$, $p_{T\,cone} \geqslant 10$
Isolation (GeV) 2	$N_{X^+} = 1$, $N_{cone1} = 1$, $r_0 > 0.01$
Isolation (GeV) 3	$N_{X^+} = 3$, $N_{cone1} = 1$, $p_{T\,cone} \geqslant 10$, $\theta_S < \cos^{-1}(0.9995)$
Isolation (GeV) 4	$N_{X^+} = 1$, $N_{cone2} = 0$, $r_0 > 0.01$, $p_{T\,cone} \geqslant 10$
Isolation (GeV) 5	$N_{X^+} = 3$, $N_{cone2} = 0$, $p_{T\,cone} \geqslant 10$, $\theta_S < \cos^{-1}(0.9995)$

Table 7.5: Optimised parameters of IsolatedTauIdentifer

Comparison: TauFinderProcessor v.s. IsolatedTauIdentifer

The two processors share similar logic: searches cone and isolation cone. The main difference is that the IsolatedTauIdentifer has an iterative approach to build up a tau candidate, which allows a dynamic tau search cone size. The IsolatedTauIdentifer also has smaller cut values on minimum p_T and invariant mass, but stricter isolation criterions. The performances of processors are shown in table 7.6.

7.3.3 Very forward electron identification

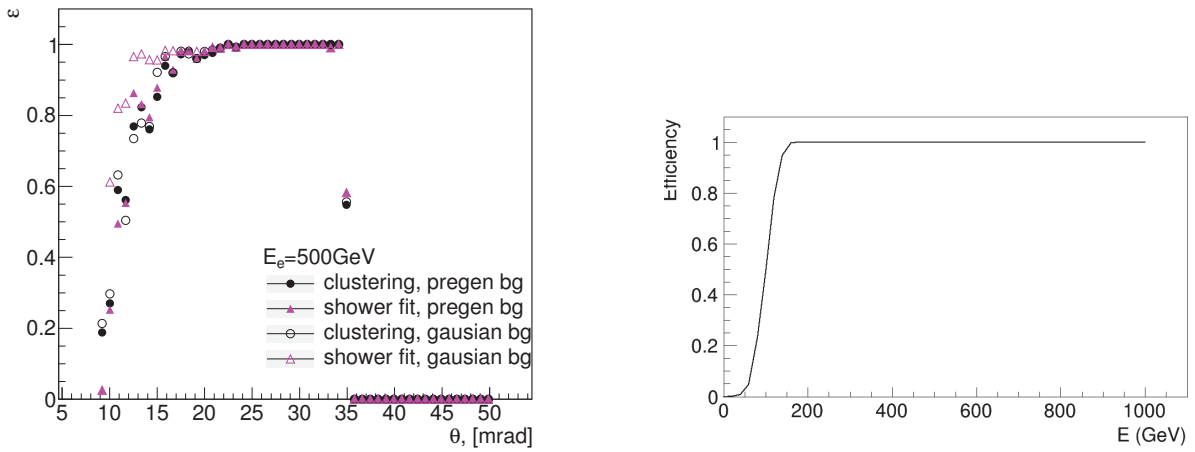
At high $\sqrt{s}$, particles are boosted and it is important to extract information in the forward calorimeters to aid signal selection. Certain background channels, for example photon-electron interactions, can have energetic electrons in the forward calorimeters, the LumiCAL and the BeamCAL. As forward calorimeters have very low angular acceptances, most particles in these forward detector would be very forward particles from beam induced background. However, [29] shows that sufficiently high energy electrons can be efficiently identified in the BeamCAL and LumiCAL.

For the CLIC_ILD detector concept, energies deposited in the LumiCAL and the BeamCAL are not reconstructed in simulation. This is because the thousands of beam induced background particles per bunch crossing requires expensive computational resources. However, previous studies [75, 76] using parameterise the particle Id efficiency give equivalent performance to the full detector simulation approach.

For the BeamCAL, [75] describes an electron tagging algorithm developed using $\sqrt{s} = 3\,\text{TeV}$ collision environment by comparing the simulated electron and background energy distributions. An electron is tagged if the energy is significantly larger than the expected background energy distributions. Events are overlaid with background energy deposition integrated over 40 bunch crossing. The tagging efficiency for electrons with energy 500 to 1500 GeV are binned in histograms at an interval of 100 GeV. There is no tagging for electrons with energy below 500 GeV or about 1500 GeV. In addition, this efficacy tagging approach most likely underestimates efficiencies due to the coarse binning of energies. I.e. a 650 GeV particle is treated the same as a 600 GeV particle. An indicative performance plot of 500 GeV electron tagging efficiency as a function of polar angle is shown in figure 7.2a.

The input of the BeamCAL electron tagging algorithm is the four momenta of the MC electron. Since the algorithm assumes collision at $\sqrt{s} = 3\,\text{TeV}$, for the $\sqrt{s} = 1.4\,\text{TeV}$ user case, the momenta of the MC electron is scaled down by a factor of $\frac{3}{1.4}$.

For the LumiCAL, the $H \to \mu\mu$ analysis in [77, 78] has developed an algorithm for electron tagging in the LumiCAL with similar logic to the algorithm for the BeamCAL. Figure 7.2 shows the LumiCAL electron tagging efficiency as a function of the electron energy for polar angle $\theta = 50\,\text{mrad}$ where events are overlaid with background energy deposition integrated over 100 bunch crossings. The LumiCAL electron tagging in this analysis is based on the performance plot in figure 7.2.



(a) **(b)**

Figure 7.2: Figure 7.2a shows BeamCAL 500 GeV electron tagging efficiency as a function of polar angle with different methods to model backgrounds and fittings, taken from [75]. Figure 7.2b shows the LumiCAL electron tagging efficiency as a function of the electron energy, for polar angle $\theta = 50\,\mathrm{mrad}$, taken from [76].

Assuming that the LumiCAL electron tagging efficiency is the same as in figure 7.2 for all polar angles and for $\sqrt{s} = 1.4\,\mathrm{TeV}$ and 3 TeV, LumiCAL electron tagging efficiency, ε is parameterised as

$$\varepsilon = \begin{cases} 0, & \text{if } E < 50\,\mathrm{GeV}, \\ 0.99 \times \frac{(\mathrm{erf}(E-100)+1)}{2}, & \text{otherwise}, \end{cases} \tag{7.1}$$

where E is the energy of the electron or the photon and erf is the error function. For each MC electron in the LumiCAL a random number between 0 and 1 is generated. If the random number is less than ε the MC electron is tagged.

Due to lack of tracking ability in the forward region, electrons and photons would have the same electromagnetic shower profile (see section 5.2) for the ECAL spatial resolution. Therefore, photons and electrons appear indistinguishable to the BeamCAL and LumiCAL and both photons and electrons are tagged by the above algorithms.

7.3.4 Lepton identification performance

The performance of all lepton finding processors on the signal and selected background samples is shown in table 7.6. The percentages represent the events remaining. IsolatedLeptonIdentifer and IsolatedTauIdentifer are more aggressive at re-

Efficiency (1.4 TeV)	Signal	$e^+e^- \to qqqq\ell\nu$
IsolatedLeptonFinderProcessor	99.3%	50.3%
IsolatedLeptonIdentifer	99.1%	39.9%
TauFinderProcessor	97.5%	52.3%
IsolatedTauIdentifer	89.7%	38.5%
Forward Finder Processors	98.9%	95.1%
Combined	86.6%	16.8%

Table 7.6: Isolated lepton finder processors performance on the signal and selected background samples at $\sqrt{s} = 1.4\,\text{TeV}$.

Selection / Efficiency (1.4 TeV)	Signal	$e^-\gamma(BS) \to e^-qqqq$
Combined light lepton finder	87.6%	67.5%
Forward Finder Processors	98.9%	53.6%
Combined	86.6%	30.8%

Table 7.7: Very forward electron and photon finder performance on the signal and selected background samples at $\sqrt{s} = 1.4\,\text{TeV}$.

jecting background than the IsolatedLeptonFinderProcessor and TauFinderProcessor. By combining the processors, 86.6% signal events remain and 16.8% of $e^+e^- \to qqqq\ell\nu$ events survive after rejecting events where leptons are identified.

The ForwardFinderProcessor is most effective at rejecting backgrounds with leptons in the forward region. Table 7.7 shows the performance of the signal and the $e^-\gamma(BS) \to e^-qqqq$ background. 53.6% of $e^-\gamma(BS) \to e^-qqqq$ background shrived after the lepton veto with ForwardFinderProcessor.

7.3.5 Other lepton identification processors

Other isolated lepton identification processors have been tested, including IsolatedLeptonTagging and TauJetClustering. The results were unsatisfactory after parameter optimisation. They either performed poorly comparing to the processors above, or became redundant after using the processors above. Therefore, these other lepton selection processors are not used in this analysis.

7.4 Jet reconstruction

For the signal channel, $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$, one Higgs boson decays to two b quarks, resulting in two jets from hadronisation. Similarly the other Higgs boson decays to two W bosons, where each W boson decays into two quarks. Therefore, the expected number of jets is six. Physical bosons, W and H, can be reconstructed from suitable jets thus allowing information to be extracted about the signal channel. Therefore, it is important to have efficient jet reconstruction to have maximum information extraction. In this section, the optimisation of the jet reconstruction is discussed.

7.4.1 Jet reconstruction optimisation

An overview of the jet algorithms can be found in section 4.6.2. Longitudinal invariant, k_t, jet algorithm is chosen for the jet clustering. The free parameters for k_t algorithm is the R parameter, which controls the radius of the jet. The other parameter to be optimised is the choice of the PFO collection, which incorporates different level of timing and p_T cuts to reduce beam induce background (see section 4.5.2).

The metric for optimising the R parameter, and the PFO collection, is the invariant mass and mass resolution of H and W. With the signal events, jets will be paired to give physical bosons using cheated MC truth information (see section 4.6.1).

The sample for the optimisation is $e^-e^+ \to HH\nu_e\bar{\nu}_e$. The signal channel, hadronic decay of $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$, is chosen using the MC truth information by examining the decay chain of MC particles. The signal event is then processed through k_t jet algorithm in 6-jet exclusive mode. The six jets are paired up using MC truth information to the corresponding Higgs and W boson. Four invariant mass distributions are obtained: two Higgs masses, $m_{H_{bb}}$, $m_{H_{WW^*}}$, and two W masses m_W, m_{W^*}. W^* indicates the off-mass-shell W boson since when a Higgs decays into two W bosons one W is off the mass shell, as the Higgs mass is less than the sum two W masses (see section 2.8).

Mass resolution fit Three invariant mass resolutions are worth comparing for optimising jet reconstruction; $m_{H_{bb}}$, $m_{H_{WW^*}}$, and m_W. The optimal jet reconstruction should produce a sharp mass peak around the particle's simulated mass (see section 4.5.3). An example of $m_{H_{bb}}$ invariant mass distribution is shown in figure 7.3. A function is fitted to quantitatively access the mass distribution. The basic fitting function is

a Gaussian function. Although the underlying mass distribution of particles like m_W is a Breit-Wigner distribution, the overall mass distribution is Gaussian like. This is because the resolution of the detector itself is worse than the W width and the overall mass distribution is a convolution of a narrow W Breit-Wigner distribution and a wide Gaussian distribution for the detector resolution. The $m_{H_{bb}}$ mass distribution is gaussian like but with asymmetrical width. This is due to b quarks decaying to neutrinos leading to a loss of detectable particles and loss of momentum. Therefore, there are more events with lower invariant mass and thus asymmetrical width is obtained. As only the peak region of the mass distribution is Gaussian like, tail parameters are added to the fitting function in order to fit the whole range of the mass distribution. The fitting function takes the form of

$$f(m) = Ae^{-\frac{(m-\mu)^2}{g}} \begin{cases} g = 2\sigma_L + \alpha_L(m-\mu), & \text{if } m < \mu, \\ g = 2\sigma_R + \alpha_R(m-\mu), & \text{if } m \geqslant \mu, \end{cases} \tag{7.2}$$

where m is binned mass distribution with 50 bins in range $[0, 200]$ GeV. μ is the fitted mass peak. σ_L and σ_R allow asymmetrical width of the distribution. α_L and α_R control the fit of tails. A is the normalisation factor.

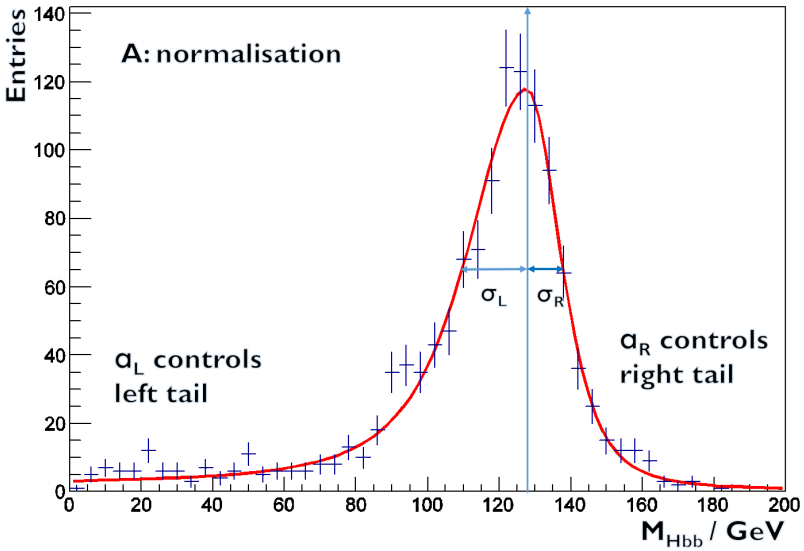


Figure 7.3: A typical example of $m_{H_{bb}}$ mass distribution with the superimposed fitting function in red.

Optimal R and PFO collection The mass fit is performed for $m_{H_{bb}}$, $m_{H_{WW^*}}$, and m_W distributions. The optimal jet reconstruction should have the mass peak close to the particle's simulated mass and a narrow peak width. Due to the asymmetrical fit, the overall relative width is defined as $(\sigma_L + \sigma_R)/M$. Smaller width indicates better mass resolution. This mass fit is repeated for reconstruction with R values of 0.5 to 1.3, at interval of 0.1, and with three PFO collections: loose, normal, and tight (see section 4.5.2).

Figure 7.4 shows the mass peak and the relative width as a function of R and PFO collections, for $m_{H_{bb}}$, $m_{H_{WW^*}}$, and m_W. The mass peak value increases as R increases. This is because more particles are included in jets with increasing R. Hence a larger invariant mass is obtained. For the relative width, values for H_{bb} increase with increasing R, but values for H_{WW^*} decrease. This is due to a compensating effect. H_{WW^*} decays to 4 jets, which prefers a large jet radius, whilst H_{bb} decays to 2 jets, which prefers a small jet radius. Similarly, W prefers a small jet radius and the relative width increases with increasing R.

The choice of PFO collection impacts number of PFOs in the event. The loose PFO selection has the most PFOs in the event and, therefore, the largest invariant mass and worst mass resolution. This trend is consistent when comparing loose to normal to tight PFO collections.

The optimal choice, normal selected PFO collection with $R = 0.7$ gives a good fitted mass for H_{WW^*} and W. The mass is slightly too low for H_{bb}. The small R is good for $m_{H_{bb}}$ and m_W resolution. $m_{H_{WW^*}}$ resolution is relatively flat when $R > 0.7$. Hence normal selected PFO collection with $R = 0.7$ is the optimal choice. The extracted, fitted parameters of optimal jet reconstructions are summarised in table 7.8.

7.5 Jet flavour tagging

As the signal channel, $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$ has two b quarks in the final state, information can be extracted to determine the likelihood of a jet originated from a b quark. To establish the likelihood of a b jet, also known as b tag value, LCFIPlus [54] software package is used. An overview of the flavour tagging processor is in section 4.6.3.

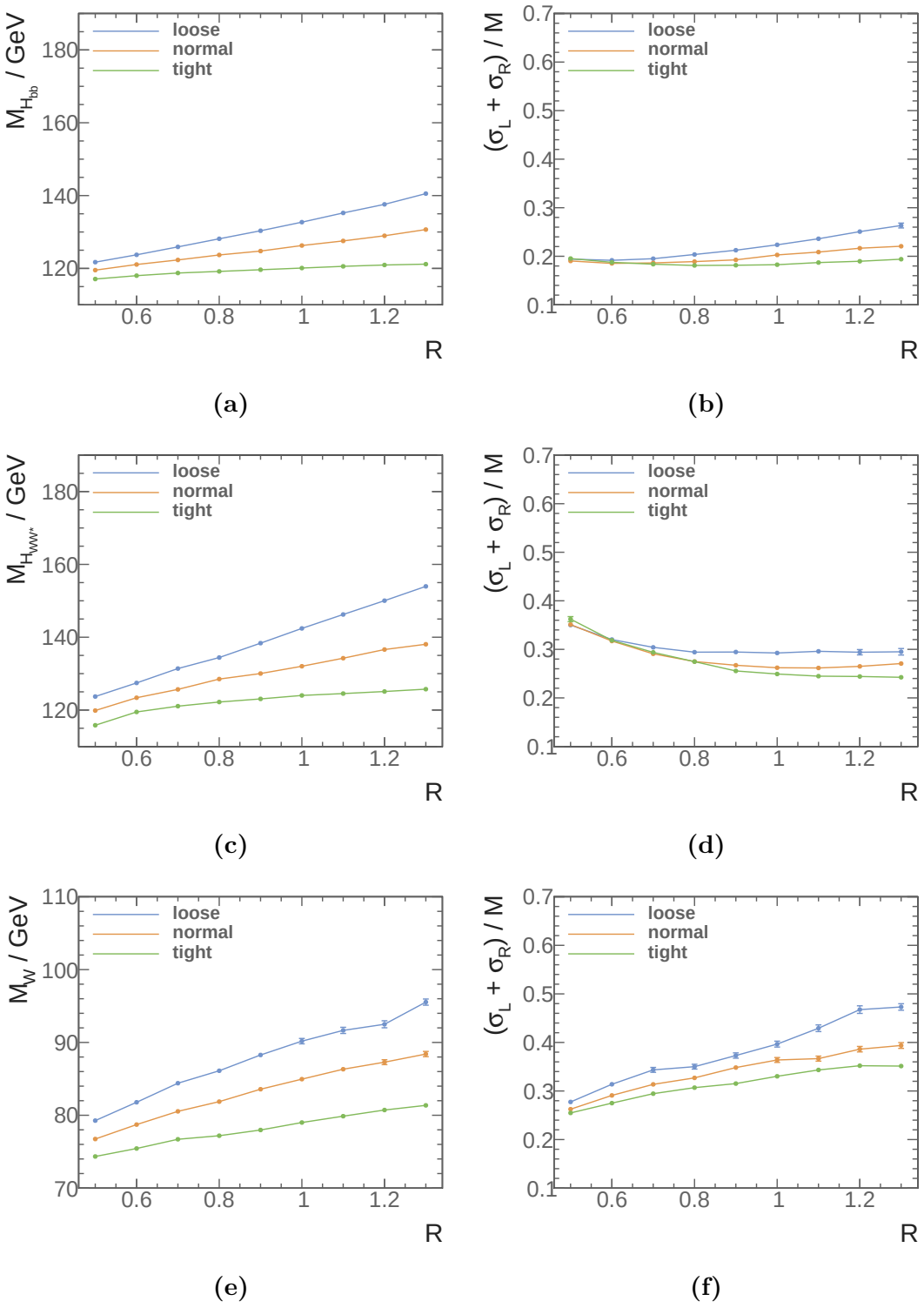


Figure 7.4: Figure 7.4a, 7.4c, and 7.4e show fitted mass of H_{bb}, H_{WW^*}, and W as a function of R, for loose, normal and tight selected PFO collection at $\sqrt{s} = 1.4\,\text{TeV}$. Figure 7.4b, 7.4d, and 7.4f show relative width of H_{bb}, H_{WW^*}, and W as a function of R for loose, normal and tight selected PFO collection at $\sqrt{s} = 1.4\,\text{TeV}$.

Fitted jet parameters	$\sqrt{s} = 1.4\,\text{TeV}$
$\mu_{H_{bb}}$	$122.3_{\pm 0.2}$
$\sigma_{L,H_{bb}}$	$15.2_{\pm 0.2}$
$\sigma_{R,H_{bb}}$	$7.55_{\pm 0.16}$
$\mu_{H_{WW^*}}$	$125.7_{\pm 0.2}$
$\sigma_{L,H_{WW^*}}$	$29.4_{\pm 0.3}$
$\sigma_{R,H_{WW^*}}$	$7.18_{\pm 0.17}$
μ_W	$80.5_{\pm 0.2}$
$\sigma_{L,W}$	$16.2_{\pm 0.3}$
$\sigma_{R,W}$	$9.03_{\pm 0.16}$

Table 7.8: The fitted parameters of optimal jet reconstruction, normal selected PFO collection with $R = 0.7$, at $\sqrt{s} = 1.4\,\text{TeV}$.

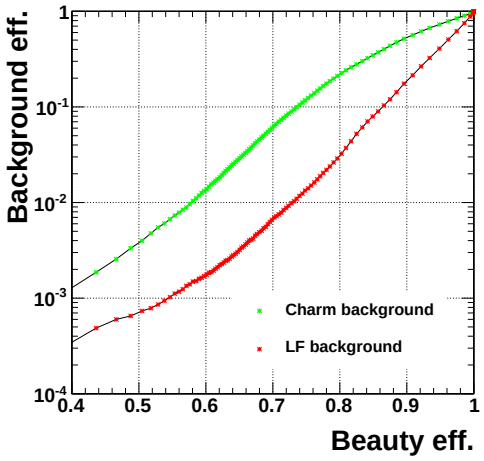


Figure 7.5: Performance of b-jet tagging with training samples at $\sqrt{s} = 1.4\,\text{TeV}$.

The flavour tagging is performed after the initial jet reconstruction, and PFOs in the reconstructed jets are the input to the flavour tagging processor. LCFIPlus includes a multiclass classifier which needs to be trained. The samples for training the multiclass classifier are $e^+e^- \to Z\bar{\nu}\nu$ at $\sqrt{s} = 1.4\,\text{TeV}$, where Z decays to $b\bar{b}$, $c\bar{c}$, or $u\bar{u}/d\bar{d}/s\bar{s}$. The classifier in the LCFIPlus processor is trained with the optimal jet parameters. The jet clustering step is set to find two jets. The output of the processor for a jet is three values, corresponding to the likelihood of the jet being a b jet, a c jet, or a light flavour quark jet. The selection efficiency of b jets and c jets with training samples is shown in figure 7.5.

To use the LCFIPlus, all the PFOs in the initial reconstructed jet are fed into the processor. The jet clustering step in the LCFIPlus is set to find six jets. For each jet, values for the likelihood of a b jet and a c jet are obtained.

7.6 Jet pairing

Having optimised the jet reconstruction, and obtained the six jets from the jet clustering step in the LCFIPlus processor, the next step is to group jets according to event topology. The hadroinc decay of the $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$ has six quarks in the final state, which results in six jets in the event. Similar to the jet pairing scheme in the jet reconstruction optimisation section 7.4.1, jets are paired up such that there are two jets for $H \to b\bar{b}$, two jets for hadroinc decay of a W^*, two jets for of a $W*$, and the two W forming a H. W^* indicates the off-mass-shell W boson, because when a Higgs decays into two W bosons, one W is off the mass shell, as the Higgs mass is less than the sum of two W masses (see section 2.8).

The jet pairing should reconstruct the invariant mass H_{bb}, H_{WW^*}, and W, with the expected mass resolution. The best possible reconstructed mass peak and the mass resolution are obtained with the MC truth information, listed in the table 7.8. These numbers are used as in jet pairing metric:

$$\chi^2 = \left(\frac{m_{ij} - \mu_{H_{bb}}}{\sigma'_{H_{bb}}}\right)^2 + \left(\frac{m_{klmn} - \mu_{H_{WW^*}}}{\sigma'_{H_{WW^*}}}\right)^2 + \left(\frac{m_{kl} - \mu_W}{\sigma'_W}\right)^2, \tag{7.3}$$

$$\sigma'_{H_{bb}} = \begin{cases} \sigma_{L,H_{bb}}, & \text{if } m_{ij} < \mu_{H_{bb}} \\ \sigma_{R,H_{bb}}, & \text{otherwise} \end{cases} \tag{7.4}$$

$$\sigma'_{H_{WW^*}} = \begin{cases} \sigma_{L,H_{WW^*}}, & \text{if } m_{klmn} < \mu_{H_{WW^*}} \\ \sigma_{R,H_{WW^*}}, & \text{otherwise} \end{cases} \tag{7.5}$$

$$\sigma'_W = \begin{cases} \sigma_{L,W}, & \text{if } m_{kl} < \mu_W \\ \sigma_{R,W}, & \text{otherwise} \end{cases} \tag{7.6}$$

where i to l indicate the one of the six jets with all possible combinations. μ and σ are the fitted invariant mass and the fitted width from table 7.8. The asymmetrical structure of the fitting function is reflected in the jet pairing metric. The jet pairing with minimal χ^2 is chosen with an additional requirement of at least one of two jets forming H_{bb} having a b-jet tag greater than 0.2.

7.7 Pre-selection

With reconstructed jets paired to the physical bosons, kinematic and topological variables can be calculated for the signal selection. A set of pre-selection cuts are placed to aid the multivariate analysis. The pre-selection cuts falls into three categories: discriminative pre-selection cuts, loose cuts for the MVA, and the mutually exclusive cuts with the $HH \to b\bar{b}qqqq$ sub-channel.

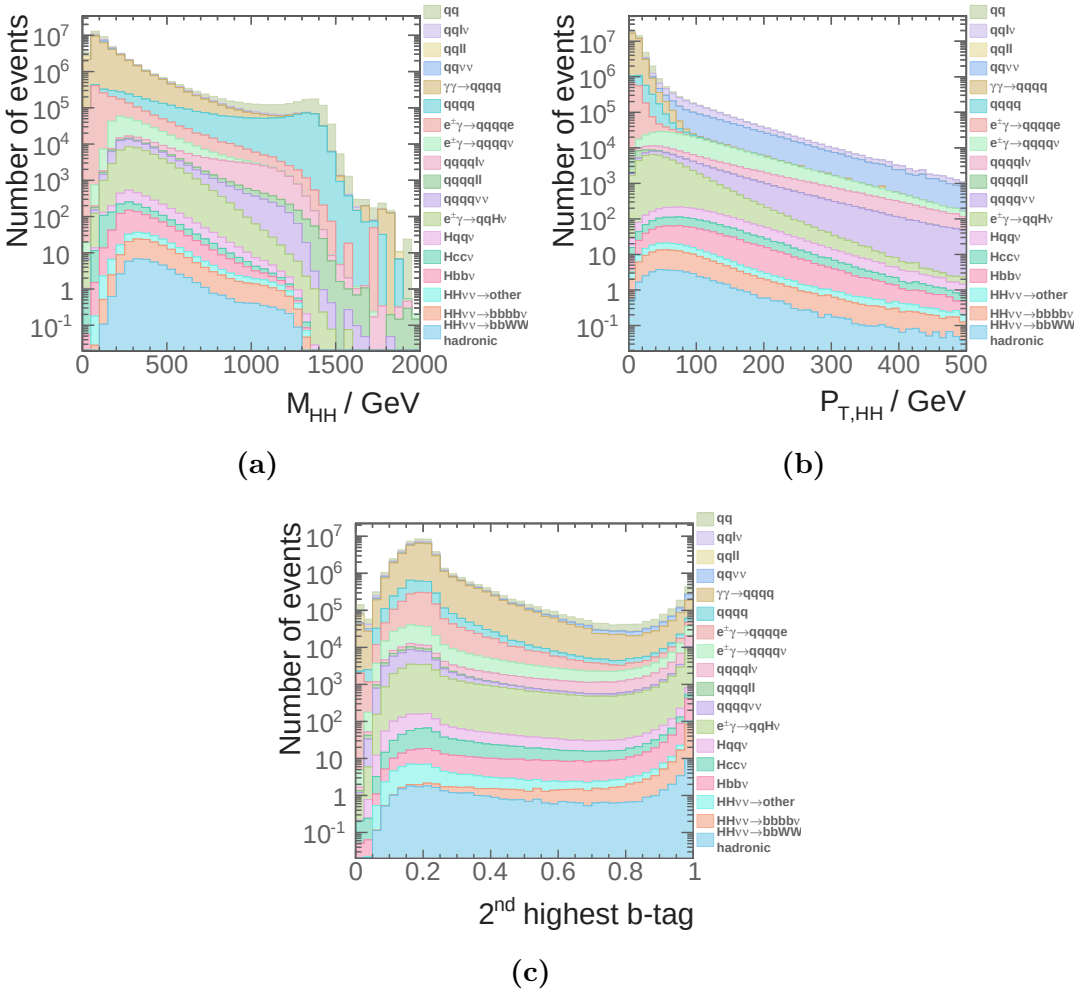
7.7.1 Discriminative pre-selection cuts

This set of pre-selection cuts are designed to discard the phase space which is dominated by background events. The double Higgs system should have substantial invariant mass as both H are real. The two b jets in the signal final state is a clear signature. The final state contains neutrinos. Therefore, there is missing momentum in the signal event. These form the logics behind the pre-selection cuts. The cuts are listed in table 7.9 and shown in figure 7.6.

Figure 7.6a shows the distribution of the invariant mass of the two Higgs system, where the cut above 150 GeV is effective against samples with two quark final states. Figure 7.6c shows the distribution of the second highest b-jet tag, where the cut above 0.2 helps to reduce background events with no b-jets in final states. Figure 7.6b shows the distribution of the p_T of the two Higgs system, where the cut above 30 GeV is extremely effective against background channels with no neutrinos in the final state.

Pre-selection	$\sqrt{s} = 1.4\,\text{TeV}$
Discriminative pre-selection	$m_{HH} > 150\,\text{GeV}$, $B_2 > 0.2$, $p_{T\,HH} > 30\,\text{GeV}$
Loose cuts for MVA	$m_{H_{bb}} < 500\,\text{GeV}$, $m_{H_{WW^*}} < 800\,\text{GeV}$, $m_W < 200\,\text{GeV}$, $m_{HH} < 1400\,\text{GeV}$
Mutually exclusive	$\Sigma B_{4\text{jets}} < 2.3$, $y_{34} < 3.7$

Table 7.9: Pre-selection cuts at $\sqrt{s} = 1.4\,\text{TeV}$.



(a) **(b)**

(c)

Figure 7.6: Discriminative pre-selection variables for $\sqrt{s} = 1.4\,\text{TeV}$, after rejecting events with leptons, and jet pairing.

The selection efficiency of the lepton veto and the pre-selection is shown in table 7.10. These pre-selections are very aggressive. The reason being that the cross sections of the signal channel is extremely small compared to the background. Hence only the signal events with very clear characteristic topologies would be able to pass the final MVA selection, which is optimised for the signal significance. Therefore, aggressive pre-selection cuts would not be detrimental to the final signal selection. On the contrary, this would improve final signal selection efficiency as the MVA can focus on the difficult background events where their topologies are similar to the signal events.

7.7.2 Mutually exclusive cuts for $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$

This set of cuts is designed to divide samples, both signal and background, into two mutually exclusive sets for the parallel analyses of two sub-channels; $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$ and $HH \to b\bar{b}b\bar{b}$. This eases the difficulty of combining sub-channels as correlations between sub-channels do not need to be considered where samples are mutually exclusive.

The most distinctive difference between the two sub-channels is the different number of jets and the different number of b-jets in the final state. Variables relating to the number of b-jets and number of overall jets are suitable for separating two sub-channels.

As demonstrated in figure 7.7, two sub-channels can be clearly separated in the two dimensional parameter space. The optimal rectangular cuts were selected by scanning the two parameters and maximising a variant of Gini Index (see section 4.7.2):

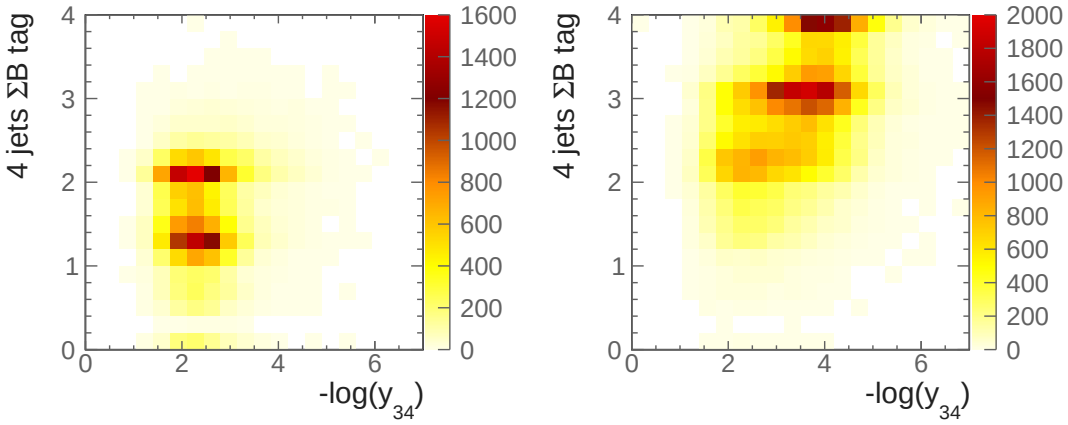
$$\varepsilon = P(\mathtt{subchannel}_1|\mathtt{selection}) \times P(\mathtt{subchannel}_2|\neg\mathtt{selection}), \tag{7.7}$$

where $\mathtt{selection}$ represents the mutually exclusive cuts, $\neg\mathtt{selection}$ indicates the phase space not covered by the $\mathtt{selection}$.

Variables tested include ΣB_{4jets}, $\sum_1^3 B_{4jets}$, y_{34}, y_{45}, y_{56}, y_{67}. The best separation is summarised in table 7.11. The ΣB_{4jets} is the sum of the b tag values when clustering an event to four jets. y parameters measures the number of jets in an event (see section 4.6.2).

Channel / Efficiency $\sqrt{s} = 1.4\,\text{TeV}$	Expected number of events	Lepton ID and jet pairing	$m_{HH} > 150\text{GeV}$	$B_2 > 0.2$	$p_T > 30\text{GeV}$
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$ $b\bar{b}W^+W^-\nu_e\bar{\nu}_e$, hadronic	27.9	85.8%	85.6%	73.7%	66.4%
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$ $b\bar{b}b\bar{b}\nu_e\bar{\nu}_e$	67.6	90.8%	90.5%	90.1%	80.6%
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$ other	128.0	36.2%	35.3%	27.7%	24.7%
$e^-e^+ \to q_l q_l H\nu\bar{\nu}$	1304.0	60.7%	59.8%	44.9%	42.0%
$e^-e^+ \to c\bar{c}H\nu\bar{\nu}$	546.1	67.4%	57.7%	46.5%	43.4%
$e^-e^+ \to b\bar{b}H\nu\bar{\nu}$	463.0	73.9%	72.6%	68.7%	64.2%
$e^-e^+ \to qqqq$	1867650.0	48.8%	46.1%	17.3%	4.7%
$e^-e^+ \to qqqq\ell\ell$	93150.0	5.0%	4.9%	1.5%	0.3%
$e^-e^+ \to qqqq\ell\nu$	165600.0	15.1%	15.1%	12.4%	11.4%
$e^-e^+ \to qqqq\nu\bar{\nu}$	34800.0	50.7%	50.0%	20.1%	18.8%
$e^-e^+ \to qq$	6014250.0	54.5%	17.5%	8.4%	2.2%
$e^-e^+ \to qq\ell\nu$	6464550.0	14.1%	5.3%	2.0%	1.6%
$e^-e^+ \to qq\ell\ell$	4088700.0	13.0%	1.1%	0.6%	0.1%
$e^-e^+ \to qq\nu\nu$	1181550.0	60.1%	12.3%	6.2%	5.8%
$e^-\gamma(BS) \to e^-qqqq$	1305787.5	23.3%	10.6%	4.4%	0.4%
$e^+\gamma(BS) \to e^+qqqq$	1300837.5	23.4%	10.5%	4.3%	0.4%
$e^-\gamma(EPA) \to e^-qqqq$	430650.0	11.1%	5.4%	2.2%	0.3%
$e^+\gamma(EPA) \to e^+qqqq$	430350.0	11.1%	5.3%	2.1%	0.3%
$e^-\gamma(BS) \to \nu qqqq$	89775.0	58.3%	56.8%	31.0%	27.7%
$e^+\gamma(BS) \to \bar{\nu}qqqq$	89212.5	57.6%	56.1%	30.3%	27.3%
$e^-\gamma(EPA) \to \nu qqqq$	26100.0	29.6%	28.9%	15.4%	13.9%
$e^+\gamma(EPA) \to \bar{\nu}qqqq$	25950.0	29.2%	28.5%	15.0%	13.7%
$e^-\gamma(BS) \to qqH\nu$	17775	61.0%	59.8%	45.5%	34.6%
$e^+\gamma(BS) \to qqH\nu$	17662.5	61.1%	60.0%	45.6%	34.6%
$e^-\gamma(EPA) \to qqH\nu$	5085	31.8%	31.2%	23.7%	18.2%
$e^+\gamma(EPA) \to qqH\nu$	5085	31.9%	31.3%	23.8%	18.4%
$\gamma(BS)\gamma(BS) \to qqqq$	2054951.5	56.3%	23.9%	9.6%	0.3%
$\gamma(BS)\gamma(EPA) \to qqqq$	4521037.5	33.6%	14.2%	5.7%	0.4%
$\gamma(EPA)\gamma(BS) \to qqqq$	4539150.0	33.7%	14.2%	5.7%	0.4%
$\gamma(EPA)\gamma(EPA) \to qqqq$	1129500.0	21.1%	9.1%	3.7%	0.4%

Table 7.10: Pre-selection cut efficiency for signal and background samples at $\sqrt{s} = 1.4\,\text{TeV}$, assuming a luminosity of $1500\,\text{fb}^{-1}$. The selection efficiencies are presented in a "flow" fashion, as the every selection cut contains all the cuts to the left of it.



(a) $HH \to b\bar{b}W^+W^-$, hadronic **(b)** $HH \to b\bar{b}b\bar{b}$

Figure 7.7: Sum of b tag against y_{34} at $\sqrt{s} = 1.4\,\text{TeV}$, shown for hadronic decay of $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}W^+W^-$ sub-channel.

Selection efficiency	$HH \to b\bar{b}W^+W^-$, hadronic	$HH \to b\bar{b}b\bar{b}$
$\Sigma B_{4jets} < 2.3$ and $y_{34} < 3.7$	86%	78%

Table 7.11: Mutually exclusive cuts at $\sqrt{s} = 1.4\,\text{TeV}$ for hadronic decay of $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$ sub-channels.

The selection efficiencies, after mutually exclusive cuts are made, are listed in table 7.12. As desired, the mutually exclusive cuts reject most $HH \to b\bar{b}b\bar{b}$ events.

7.7.3 Loose cuts for the MVA

A set of physics-motivated loose cuts aims to reduce the range of invariant masses to increase the effectiveness of MVA. (See section **??**). The invariant masses of physical bosons are required to be within a certain range to avoid the effect of extreme values on the MVA. The cuts are listed in table 7.9 and the performance is shown in figure 7.12.

7.8 Discriminative variables for MVA

A series of discriminative variables are calculated to differentiate the signal and background events. These variables are fed into MVA for signal selection. A full list of variables can be found in table 7.13. These are grouped into several categories.

Invariant mass variables are very effective at selecting signal events as no background events have double Higgs bosons in final states. Figure 7.8a and figure 7.8b show the distributions of $m_{H_{bb}}$ and $m_{H_{WW^*}}$ after all pre-selection cuts.

For the off-shell W*, energy is used as its mass distribution does not have a resonance. For the recoil momenta, which is calculated by assuming the collision at $\sqrt{s}$ and a beam crossing angle 20 mrad, the pseudorapidity is used to focus on the forward region. The pseudorapidity, η, is defined as:

$$\eta \equiv -\ln\left[\tan\left(\frac{\theta}{2}\right)\right], \tag{7.8}$$

where θ is the polar angle. A_{12} measures the angle between the two constituent jets, defined as:

$$A_{12} = \pi - \cos^{-1}\left(\hat{\mathbf{p}}_1 \cdot \hat{\mathbf{p}}_2\right), \tag{7.9}$$

where $\hat{\mathbf{p}}_1$ is the unit momentum vector of jet 1. $\cos(\theta^*_{12})$ is the cosine of the angle between the two constituent jets in their decay rest frame. Figure 7.8c and figure 7.8d compares the $A_{H_{bb}}$ and $\cos\left(\theta^*_{H_{bb}}\right)$. Both show clear differences for the signal and background channels.

Channel	Previous cuts and loose cuts	Mutually exclusive
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$		
$b\bar{b}W^+W^-\nu_e\bar{\nu}_e$, hadronic	66.4%	59.7%
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$		
$b\bar{b}b\bar{b}\nu_e\bar{\nu}_e$	80.6%	15.4%
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$ other	24.7%	20.5%
$e^-e^+ \to q_l q_l H\nu\bar{\nu}$	42.0%	39.5%
$e^-e^+ \to c\bar{c}H\nu\bar{\nu}$	43.4%	31.7%
$e^-e^+ \to b\bar{b}H\nu\bar{\nu}$	64.2%	25.2%
$e^-e^+ \to qqqq$	4.6%	3.4%
$e^-e^+ \to qqqq\ell\ell$	3.3%	3.1%
$e^-e^+ \to qqqq\ell\nu$	11.4%.	9.8%
$e^-e^+ \to qqqq\nu\bar{\nu}$	18.8%	16.6%
$e^-e^+ \to qq$	2.0%	0.8%
$e^-e^+ \to qq\ell\nu$	1.6%	0.9%
$e^-e^+ \to qq\ell\ell$	0.1%	0.1%
$e^-e^+ \to qq\nu\nu$	5.8%	4.0%
$e^-\gamma(BS) \to e^-qqqq$	0.4%	0.3%
$e^+\gamma(BS) \to e^+qqqq$	0.4%	0.4%
$e^-\gamma(EPA) \to e^-qqqq$	0.3%	0.2%
$e^+\gamma(EPA) \to e^+qqqq$	0.3%	0.3%
$e^-\gamma(BS) \to \nu qqqq$	27.7%	25.3%
$e^+\gamma(BS) \to \bar{\nu}qqqq$	27.3%	24.9%
$e^-\gamma(EPA) \to \nu qqqq$	13.9%	12.6%
$e^+\gamma(EPA) \to \bar{\nu}qqqq$	13.7%	12.3%
$e^-\gamma(BS) \to qqH\nu$	34.6%	30.6%
$e^+\gamma(BS) \to qqH\nu$	34.6%	30.6%
$e^-\gamma(EPA) \to qqH\nu$	18.2%	16.0%
$e^+\gamma(EPA) \to qqH\nu$	18.4%	16.1%
$\gamma(BS)\gamma(BS) \to qqqq$	0.3%	0.3%
$\gamma(BS)\gamma(EPA) \to qqqq$	0.4%	0.3%
$\gamma(EPA)\gamma(BS) \to qqqq$	0.4%	0.3%
$\gamma(EPA)\gamma(EPA) \to qqqq$	0.4%	0.3%

Table 7.12: List of signal and background samples after loose cuts and mutually exclusive cuts at $\sqrt{s} = 1.4\,\text{TeV}$. The selection efficiencies are presented in a "flow" fashion, as the every selection cut contains all the cuts to the left of it.

Category	Variable
Invariant mass	$m_{H_{bb}}$, $m_{H_{WW^*}}$, m_W, m_{HH}
Energy and momentum	E_{W^*}, E_{mis}, $p_{TH_{bb}}$, $p_{TH_{WW^*}}$, p_{TW}, p_{THH}
Angles in lab frame	η_{mis}, $A_{H_{bb}}$, A_W, A_{HH}
Angles in boosted frames	$\cos\left(\theta^*_{H_{bb}}\right)$, $\cos\left(\theta^*_{H_{WW^*}}\right)$, $\cos\left(\theta^*_W\right)$, $\cos\left(\theta^*_{W^*}\right)$, $\cos\left(\theta^*_{HH}\right)$
Event shape	$\lvert\mathbf{S}\rvert$, $-\ln(y_{23})$, $-\ln(y_{34})$, $-\ln(y_{45})$, $-\ln(y_{56})$
b and c tag	$B_{1,H_{bb}}$, $B_{2,H_{bb}}$, $B_{1,W}$, B_{1,W^*}, $C_{1,H_{bb}}$, $C_{1,W}$
Number of PFOs	$N_{H_{bb}}$, $N_{H_{WW^*}}$, N_W, N_{W^*}

Table 7.13: Variables used in MVA at $\sqrt{s} = 1.4\,\text{TeV}$

For the signal, $\cos\left(\theta^*_{H_{bb}}\right)$ has a flat distribution, as expected from a back-to-back decay of $H \to b\bar{b}$. For the background, $\cos\left(\theta^*_{H_{bb}}\right)$ peaks at 1.

The global event shape variables includes y variables (see section **??**), and the sphericity, **S**. **S** is a measurement of the spherically symmetry of the event (see section **??**).

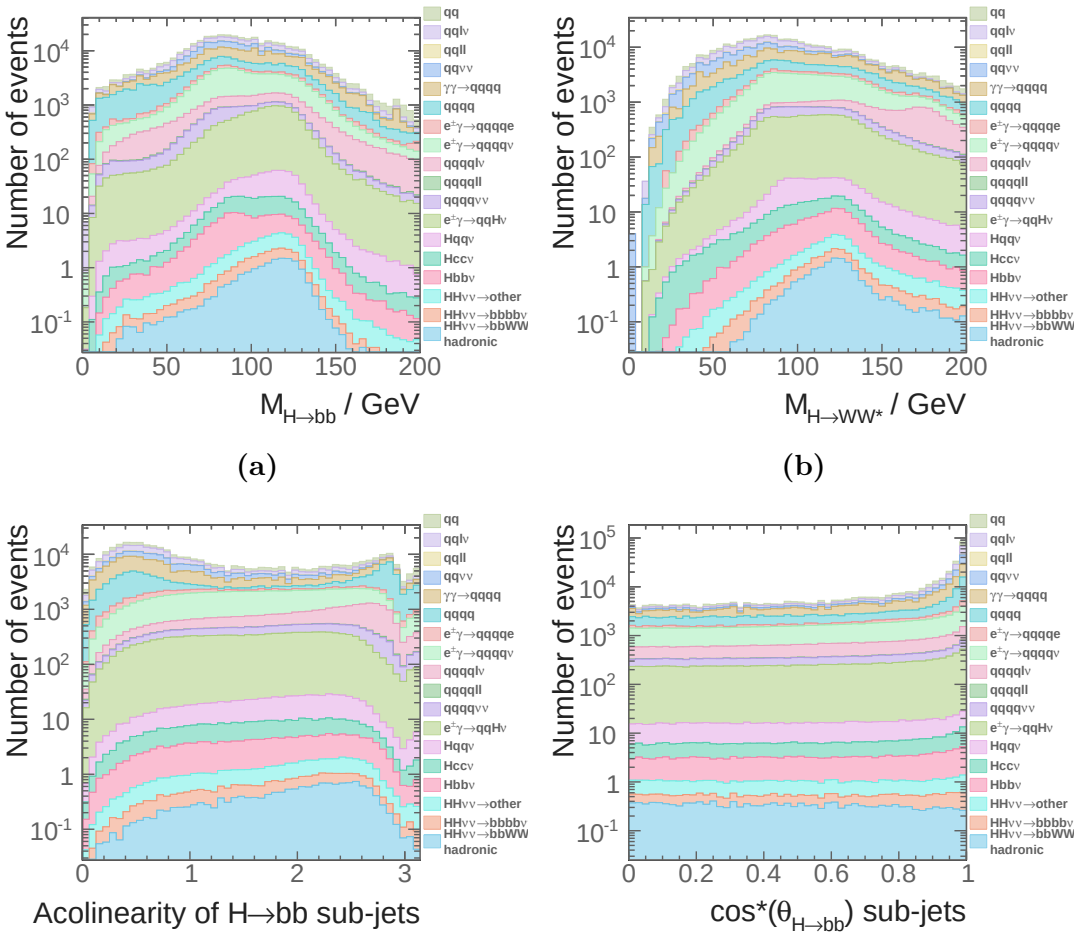
The flavour tagging variables are discussed in section 7.5. For example, $B_{1,H_{bb}}$ denotes the highest b jet tag value of two jets forming H_{bb}. The number of PFOs variables are effective against background events with fewer quarks in final states.

An optimal set of 32 variables are chosen for the best MVA performance, whilst no strong ($> 80\%$) pair-wise correlation exists between any two variables.

7.9 Multivariate analysis

After gathering information and applying pre-selection cuts, signal selection is performed using multivariate analysis (MVA) with Boosted Decision Tree classifier (BDT). The parameters for boosted decision tree are optimised and checked for overtraining. A brief discussion on the MVA, classifier, and overtraining can be found in section 4.7.

The optimisation of the BDT follows the strategy in section 4.7.2. The optimised parameters are listed in table 7.14. The optimal values are obtained by choosing the



(a) (b) (c) (d)

Figure 7.8: Stacked plots for discriminative variables for the MVA at $\sqrt{s} = 1.4\,\text{TeV}$ after all pre-selection cuts. Figure 7.8a and figure 7.8b show the mass distributions of H_{bb} and H_{WW^*}. Figure 7.8c and figure 7.8d show the acollinearity and the opening angle in the decay rest frame of the two jets forming H_{bb}.

Parameter	Value
Depth of tree	4
Number of trees	4000
The minimum number of events in a node	0.25% of the total events
Boosting	adaptive boost
learning rate of the adaptive boost	0.5
metric for the optimal cuts	Gini Index
bagging fraction	0.5
Number of bins per variables	40
End node output	$x \in [0, 1]$
Do-PreSelection	yes

Table 7.14: Optimised parameters for the boosted decision tree classifier. See section 4.7.2 for detailed explanation of variables.

best performance without overfitting at $\sqrt{s} = 3\,\text{TeV}$. The same optimised parameters are used for $\sqrt{s} = 1.4\,\text{TeV}$ analysis.

Half of the samples were used for training, and the other half used for testing and classifier optimisation. The signal for the MVA is the hadronic decay of $e^-e^+ \to HH\nu_e\bar{\nu}_e \to b\bar{b}W^+W^-\nu_e\bar{\nu}_e$. $e^-e^+ \to HH\nu_e\bar{\nu}_e$ decaying to other final states does not participate in the MVA training step. Although they are from the Feynman diagrams as the signal (see section 7.1), event topologies are different. They participate in the MVA applying stage.

7.10 Signal selection results

Efficiencies of the pre-selection cut and the MVA are listed in table 7.15, alongside with number of events after the MVA selection. A few background channels survive after the MVA selection. $e^-e^+ \to q\bar{q}H\nu\bar{\nu}$ survives as its topology, one single Higgs plus neutrino, is very similar topology to the signal. Similarly, $e^-e^+ \to qqqq\ell\nu$ can be confused as the signal when the lepton is undetected in the forward region, or the lepton's energy is too low to be tagged. $e^-e^+ \to qqqq\nu\bar{\nu}$ can also have a similar topology to the signal. Other background channels that survive the MVA are the electron-photon and photon interactions with the same final states as the above channels.

Before interpreting the result, the analyses at $\sqrt{s} = 3\,\text{TeV}$, and with the semi-leptonic channel of $e^-e^+ \to HH\nu_e\bar{\nu}_e \to b\bar{b}W^+W^-\nu_e\bar{\nu}_e$ are presented.

7.11 $e^-e^+ \to HH\nu_e\bar{\nu}_e \to b\bar{b}W^+W^-\nu\bar{\nu}$ hadronic decay at $\sqrt{s} = 3\,\text{TeV}$ analysis

The $e^-e^+ \to HH\nu_e\bar{\nu}_e \to b\bar{b}W^+W^-\nu\bar{\nu}$ hadronic decay at $\sqrt{s} = 3\,\text{TeV}$ analysis follows the same strategy as the analysis at $\sqrt{s} = 1.4\,\text{TeV}$. A brief discussion of each step and the results are provided and differences highlighted. Cross sections of used samples are listed in table 7.16.

The lepton finding processors are either developed or optimised with samples at $\sqrt{s}$ = 1.4 TeV, and checked against samples at $\sqrt{s} = 3\,\text{TeV}$ (see section 7.3). It was found that the same set of parameters for lepton identifiers works well under $\sqrt{s} = 1.4\,\text{TeV}$ and 3 TeV. The performance of the lepton processors is shown in table 7.17.

When comparing $\sqrt{s} = 1.4\,\text{TeV}$ and $\sqrt{s} = 3\,\text{TeV}$, the lepton finding performance is better for $\sqrt{s} = 1.4\,\text{TeV}$. This is because at $\sqrt{s} = 3\,\text{TeV}$, particles are boosted and separation between particles is smaller. The effect of high $\sqrt{s}$ also reflects on the performance of the ForwardFinderProcessor. Whilst at $\sqrt{s} = 1.4\,\text{TeV}$, the processor only rejects 5% $e^+e^- \to qqqq\ell\nu$ background and 1% signal, at $\sqrt{s} = 3\,\text{TeV}$ it rejects 19% background and 4% signal, suggesting many leptons are in the forward region. These processors are complimentary and a good rejection rate is achieved with combined processors.

For the jet reconstruction optimisation, the strategy outlined in section 7.4.1 is used. Figure 7.9 shows fitted mass for H_{bb}, H_{WW^*}, and W, along with the relative mass resolution. The relative resolution of W rises sharply with increasing R, hence favouring a small R and tight selected PFO collection. The optimal jet reconstruction parameters is tight selected PFO collection with $R = 0.7$. The invariant mass is smaller than simulated value to compensate for the better mass resolution. The fitted values of the chosen jet reconstruction are listed in table 7.18

Channel / Efficiency $\sqrt{s} = 1.4\,\text{TeV}$	N	$\varepsilon_{\text{presel}}$	ε_{MVA}	N_{MVA}
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$				
$b\bar{b}W^+W^-\nu_e\bar{\nu}_e$, hadronic	27.9	59.8%	8.2%	1.29
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$				
$b\bar{b}b\bar{b}\nu_e\bar{\nu}_e$	67.6	15.4%	0.5%	0.05
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$ other	128.0	20.4%	1.7%	0.45
$e^-e^+ \to q_l q_l H\nu\bar{\nu}$	1304.0	39.5%	0.05%	0.29
$e^-e^+ \to c\bar{c}H\nu\bar{\nu}$	546.1	31.6%	0.1%	0.16
$e^-e^+ \to b\bar{b}H\nu\bar{\nu}$	463.0	24.7%	0.3%	0.37
$e^-e^+ \to qqqq$	1867650.0	3.3%	-	-
$e^-e^+ \to qqqq\ell\ell$	93150.0	0.3%	-	-
$e^-e^+ \to qqqq\ell\nu$	165600.0	9.8%	0.01%	2.06
$e^-e^+ \to qqqq\nu\bar{\nu}$	34800.0	16.5%	0.002%	0.10
$e^-e^+ \to qq$	6014250.0	0.8%	-	-
$e^-e^+ \to qq\ell\nu$	6464550.0	0.9%	-	-
$e^-e^+ \to qq\ell\ell$	4088700.0	0.08%	-	-
$e^-e^+ \to qq\nu\nu$	1181550.0	4.0%	-	-
$e^-\gamma(BS) \to e^-qqqq$	1305787.5	0.3%	-	-
$e^+\gamma(BS) \to e^+qqqq$	1300837.5	0.4%	-	-
$e^-\gamma(EPA) \to e^-qqqq$	430650.0	0.3%	-	-
$e^+\gamma(EPA) \to e^+qqqq$	430350.0	0.3%	-	-
$e^-\gamma(BS) \to \nu qqqq$	89775.0	25.4%	0.005%	1.09
$e^+\gamma(BS) \to \bar{\nu}qqqq$	89212.5	24.9%	0.004%	0.96
$e^-\gamma(EPA) \to \nu qqqq$	26100.0	12.6%	-	-
$e^+\gamma(EPA) \to \bar{\nu}qqqq$	25950.0	12.4%	0.008%	0.27
$e^-\gamma(BS) \to qqH\nu$	17775	30.8%	0.02%	1.00
$e^+\gamma(BS) \to qqH\nu$	17662.5	30.6%	0.02%	1.16
$e^-\gamma(EPA) \to qqH\nu$	5085	16.0%	0.04%	0.33
$e^+\gamma(EPA) \to qqH\nu$	5085	16.2%	0.08%	0.62
$\gamma(BS)\gamma(BS) \to qqqq$	2054951.5	0.2%	-	-
$\gamma(BS)\gamma(EPA) \to qqqq$	4521037.5	0.4%	-	-
$\gamma(EPA)\gamma(BS) \to qqqq$	4539150.0	0.3%	-	-
$\gamma(EPA)\gamma(EPA) \to qqqq$	1129500.0	0.3%	-	-

Table 7.15: List of signal and background samples with selection efficiency and number of events at $\sqrt{s} = 1.4\,\text{TeV}$, assuming a luminosity of $1500\,\text{fb}^{-1}$. The number of events, selection efficiency of pre-selection, selection efficiency of MVA after pre-selection, number of events after MVA are shown. - represents a number less than 0.01.

Channel	$\sigma(\sqrt{s} = 3\,\mathrm{TeV})$ / fb
$e^-e^+ \to HH\nu_e\overline{\nu}_e$	0.588
$e^-e^+ \to HH\nu_e\overline{\nu}_e \to b\overline{b}W^+W^-$,hadronic	0.07
$e^-e^+ \to HH\nu_e\overline{\nu}_e \to b\overline{b}b\overline{b}\nu_e\overline{\nu}_e$	0.19
$e^-e^+ \to HH \to$ others	0.34
$e^-e^+ \to q_l q_l H\nu\overline{\nu}$	1.78
$e^-e^+ \to c\overline{c}H\nu\overline{\nu}$	1.12
$e^-e^+ \to b\overline{b}H\nu\overline{\nu}$	1.91
$e^-e^+ \to qqqq$	546.5*
$e^-e^+ \to qqqq\ell\ell$	169.3*
$e^-e^+ \to qqqq\ell\nu$	106.6*
$e^-e^+ \to qqqq\nu\overline{\nu}$	71.5*
$e^-e^+ \to qq$	2948.9
$e^-e^+ \to qq\ell\nu$	5561.1
$e^-e^+ \to qq\ell\ell$	3319.6
$e^-e^+ \to qq\nu\nu$	1317.5
$e^-\gamma(BS) \to e^-qqqq$	1268.7*
$e^+\gamma(BS) \to e^+qqqq$	1267.6*
$e^-\gamma(EPA) \to e^-qqqq$	287.9*
$e^+\gamma(EPA) \to e^+qqqq$	287.8*
$e^-\gamma(BS) \to \nu qqqq$	262.5*
$e^+\gamma(BS) \to \overline{\nu}qqqq$	262.3*
$e^-\gamma(EPA) \to \nu qqqq$	54.2*
$e^+\gamma(EPA) \to \overline{\nu}qqqq$	54.2*
$e^-\gamma(BS) \to qqH\nu$	58.6*
$e^+\gamma(BS) \to qqH\nu$	58.5*
$e^-\gamma(EPA) \to qqH\nu$	11.7*
$e^+\gamma(EPA) \to qqH\nu$	11.7*
$\gamma(BS)\gamma(BS) \to qqqq$	13050.3*
$\gamma(BS)\gamma(EPA) \to qqqq$	2420.6*
$\gamma(EPA)\gamma(BS) \to qqqq$	2423.1*
$\gamma(EPA)\gamma(EPA) \to qqqq$	402.7*

Table 7.16: List of signal and background samples with the corresponding cross sections at $\sqrt{s}$ = 3 TeV. q can be u, d, s, b or t. Unless specified, q, ℓ and ν represent particles and their corresponding anti-particles. γ (BS) represents a real photon from bremsstrahlung (BS). γ (EPA) represents a "quasi-real" photon simulated with the Equivalent Photon Approximation. For processes involving Higgs production explicitly, simulated Higgs mass is 126 GeV. Otherwise, Higgs mass is set to 14 TeV. For processes labelled with *, the generator level cut requires invariant mass of quarks greater than 50.

Efficiency (3 TeV)	Signal	$e^+e^- \to qqqq\ell\nu$
IsolatedLeptonFinderProcessor	99.5%	66.8%
IsolatedLeptonIdentifer	99.0%	52.5%
TauFinderProcessor	97.7%	79.5%
IsolatedTauIdentifer	86.3%	60.3%
Forward Finder Processors	95.9%	80.7%
Combined	81.0%	23.3%

Table 7.17: Isolated lepton finder processors performance with the signal and selected background samples at $\sqrt{s} = 3\,\text{TeV}$.

Jet Parameters	$\sqrt{s} = 3\,\text{TeV}$
$\mu_{H_{bb}}$	$119.1_{\pm 0.3}$
$\sigma_{L,H_{bb}}$	$15.0_{\pm 0.3}$
$\sigma_{R,H_{bb}}$	$8.4_{\pm 0.2}$
$\mu_{H_{WW^*}}$	$123.0_{\pm 0.3}$
$\sigma_{L,H_{WW^*}}$	$36.6_{\pm 0.6}$
$\sigma_{R,H_{WW^*}}$	$7.4_{\pm 0.2}$
μ_W	$78.1_{\pm 0.3}$
$\sigma_{L,W}$	$13.1_{\pm 0.4}$
$\sigma_{R,W}$	$9.5_{\pm 0.2}$

Table 7.18: The extracted fitted parameters of optimal jet reconstructions for tight selected PFO collection with $R = 0.7$ at $\sqrt{s} = 3\,\text{TeV}$.

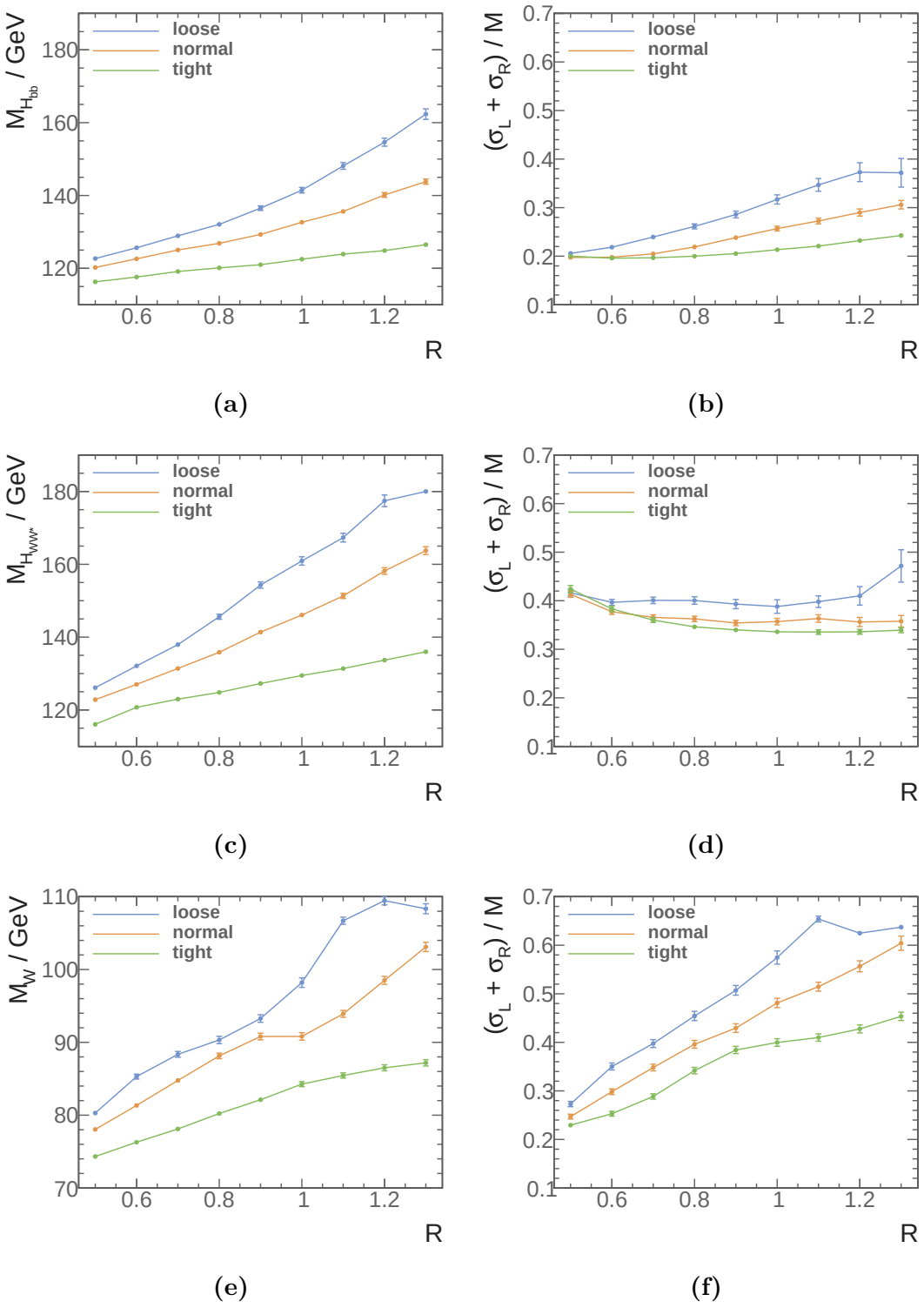


Figure 7.9: Figure 7.9a, 7.9c, and 7.9e show fitted mass of H_{bb}, H_{WW^*}, and W, respectively, for loose, normal and tight selected PFO collection as a function of R, at $\sqrt{s} = 3\,\text{TeV}$. Figure 7.9b, 7.9d, and 7.9f show relative mass resolutions of H_{bb}, H_{WW^*}, and W, respectively, for loose, normal and tight selected PFO collection as a function of R, at $\sqrt{s} = 3\,\text{TeV}$.

The flavour tagging processor is trained with the optimal jet parameters at $\sqrt{s} = 3\,\text{TeV}$. The performance of the flavour tagging with training samples is shown in figure 7.10. Comparing this to the performance at $\sqrt{s} = 1.4\,\text{TeV}$ shows that the flavour tagging does not perform as favourably. At high $\sqrt{s}$, particles at more collimated and more difficult to separate hence the performance degrades.

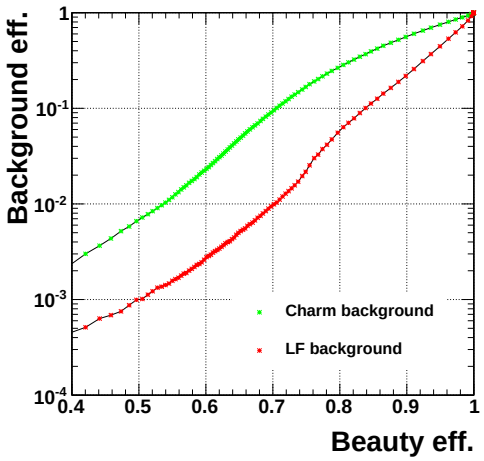


Figure 7.10: Performance of b-jet tagging with training samples at $\sqrt{s} = 3\,\text{TeV}$.

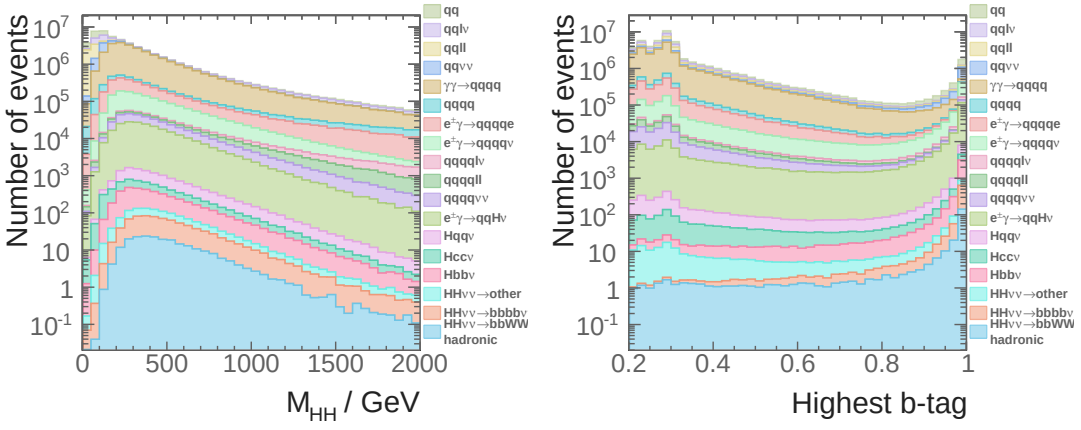
The pre-selection cuts at $\sqrt{s} = 3\,\text{TeV}$ are largely the same as ones for $\sqrt{s} = 1.4\,\text{TeV}$ analysis, listed in table 7.19. The reason for using a differnt b tag cut is because the performance of flavour tagging is poorer at high $\sqrt{s}$. A different cut is need. Figure 7.11b shows the distribution of the highest b-jet tag, where the cut above 0.7 helps to reduce background events with no b-jets in final states. The cut is aggressive to compensate for the worse performance of the flavour tagging at high $\sqrt{s}$. Figure 7.11a shows the distribution of the invariant mass of the two Higgs system, where the cut above 150 GeV is effective against samples with two quark final states.

Pre-selection	$\sqrt{s} = 3\,\text{TeV}$
Discriminative pre-selection	$m_{HH} > 150\,\text{GeV}$, $B_1 > 0.7$, $p_{T\,HH} > 30\,\text{GeV}$
Loose cuts for MVA	$m_{H_{bb}} < 500\,\text{GeV}$, $m_{H_{WW^*}} < 800\,\text{GeV}$, $m_W < 200\,\text{GeV}$, $m_{HH} < 3000\,\text{GeV}$
Mutually exclusive	$\Sigma B_{4\text{jets}} < 2.3$, $y_{34} < 3.6$

Table 7.19: Pre-selection cuts at $\sqrt{s} = 3\,\text{TeV}$.

Channel / Efficiency $\sqrt{s} = 3\,\text{TeV}$	Expected number of events	Lepton ID and jet pairing	$m_{HH} > 150\text{GeV}$	$B_1 > 0.7$
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$				
$b\bar{b}W^+W^-\nu_e\bar{\nu}_e$, hadronic	146.0	80.2%	79.9%	69.7%
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$				
$b\bar{b}b\bar{b}\nu_e\bar{\nu}_e$	355.0	83.4%	82.9%	81.2%
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$ other	675.0	36.7%	35.8%	25.2%
$e^-e^+ \to q_l q_l H\nu\bar{\nu}$	6115.4	59.5%	58.5%	40.4%
$e^-e^+ \to c\bar{c}H\nu\bar{\nu}$	2249.9	64.8%	58.4%	39.3%
$e^-e^+ \to b\bar{b}H\nu\bar{\nu}$	2197.7	69.7%	68.4%	64.2%
$e^-e^+ \to qqqq$	1093000.0	48.5%	39.7%	3.0%
$e^-e^+ \to qqqq\ell\ell$	338600.0	14.7%	14.2%	0.7%
$e^-e^+ \to qqqq\ell\nu$	213200.0	19.7%	19.4%	10.0%
$e^-e^+ \to qqqq\nu\bar{\nu}$	143000.0	58.4%	57.3%	11.9%
$e^-e^+ \to qq$	5897800.0	62.8%	13.2%	2.7%
$e^-e^+ \to qq\ell\nu$	11121800	28.3%	11.9%	0.3%
$e^-e^+ \to qq\ell\ell$	6639200.0	38.3%	2.9%	0.7%
$e^-e^+ \to qq\nu\nu$	2635000.0	71.4%	24.1%	5.3%
$e^-\gamma(BS) \to e^-qqqq$	2004388.1	23.3%	21.5%	0.8%
$e^+\gamma(BS) \to e^+qqqq$	2002334.1	23.4%	21.6%	0.8%
$e^-\gamma(EPA) \to e^-qqqq$	575600.0	12.0%	11.0%	0.5%
$e^+\gamma(EPA) \to e^+qqqq$	575600.0	12.0%	10.9%	0.4%
$e^-\gamma(BS) \to \nu qqqq$	414750.0	61.7%	59.5%	20.4%
$e^+\gamma(BS) \to \bar{\nu}qqqq$	414434.0	61.2%	59.1%	19.4%
$e^-\gamma(EPA) \to \nu qqqq$	108400.0	30.9%	29.9%	9.6%
$e^+\gamma(EPA) \to \bar{\nu}qqqq$	108400.0	30.7%	29.7%	9.1%
$e^-\gamma(BS) \to qqH\nu$	92588.0	58.3%	56.2%	37.3%
$e^+\gamma(BS) \to qqH\nu$	92430.0	58.1%	56.0%	37.1%
$e^-\gamma(EPA) \to qqH\nu$	23400.0	30.1%	29.2%	19.4%
$e^+\gamma(EPA) \to qqH\nu$	23400.0	29.7%	28.6%	18.8%
$\gamma(BS)\gamma(BS) \to qqqq$	18009413.9	54.2%	49.2%	1.9%
$\gamma(BS)\gamma(EPA) \to qqqq$	3824548.1	33.5%	30.2%	1.2%
$\gamma(EPA)\gamma(BS) \to qqqq$	3828498.1	33.7%	30.3%	1.2%
$\gamma(EPA)\gamma(EPA) \to qqqq$	805400.0	22.0%	19.8%	0.8%

Table 7.20: List of signal and background samples with selection efficiency and event numbers after the pre-selection cuts at $\sqrt{s} = 3\,\text{TeV}$, assuming a luminosity of 2000fb^{-1}. The selection efficiencies are presented in a "flow" fashion, as the every selection cut contains all the cuts to the left of it.



(a) (b)

Figure 7.11: Variable distributions in discriminative pre-selection cuts at $\sqrt{s} = 3\,\text{TeV}$, after rejecting events with identified leptons and jet pairing.

The mutually exclusive cuts divide samples - both signal and background - into two mutually exclusive sets for the parallel analyses of two subchannels; $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$ and $HH \to b\bar{b}b\bar{b}$. The cuts are obtained using the same strategy in section 7.7.2. The values are listed in table 7.19. The two dimensional spaces for two sub-channels are shown in figure 7.12. The selection efficiencies after the loose cuts and the mutually exclusive cuts are shown in table 7.21.

The loose cuts for MVA at $\sqrt{s} = 3\,\text{TeV}$ are largely the same as the ones at $\sqrt{s} = 3\,\text{TeV}$, apart from the difference on the cut of the invariant mass of HH due to higher $\sqrt{s}$. The selection efficiency of the lepton veto and the pre-selection is shown in table 7.20. These pre-selection cuts are still very aggressive.

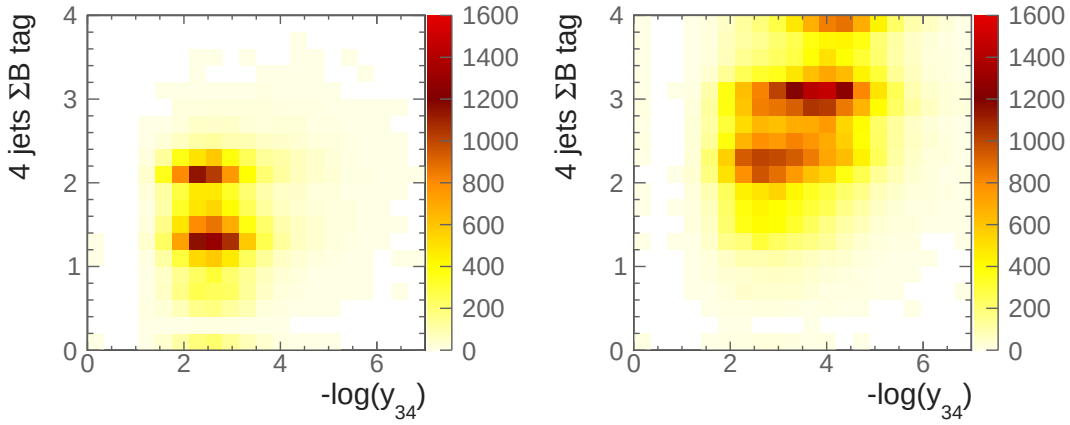
The same set of variables for the MVA are used at $\sqrt{s} = 3\,\text{TeV}$. The Boosted Decision Tree classifier is optimised at $\sqrt{s} = 3\,\text{TeV}$. The efficiencies of the pre-selection cuts and the efficiencies of the MVA selections are listed in table 7.22, alongside the number of events after the MVA selection. Background channels that are dominant after the MVA selection are almost identical to those at $\sqrt{s} = 1.4\,\text{TeV}$. Hence see section 7.10 for discussion.

Channel	Previous cuts and loose cuts	Mutually exclusive
$e^-e^+ \to HH\nu_e\overline{\nu}_e \to$		
$b\bar{b}W^+W^-\nu_e\overline{\nu}_e$, hadronic	69.5%	61.7%
$e^-e^+ \to HH\nu_e\overline{\nu}_e \to$		
$b\bar{b}b\bar{b}\nu_e\overline{\nu}_e$	81.1%	18.8%
$e^-e^+ \to HH\nu_e\overline{\nu}_e \to$ other	25.1%	20.0%
$e^-e^+ \to q_l q_l H\nu\overline{\nu}$	40.3%	35.9%
$e^-e^+ \to c\bar{c}H\nu\overline{\nu}$	39.2%	26.2%
$e^-e^+ \to b\bar{b}H\nu\overline{\nu}$	64.2%	25.9%
$e^-e^+ \to qqqq$	2.5%	1.4%
$e^-e^+ \to qqqq\ell\ell$	0.7%	0.6%
$e^-e^+ \to qqqq\ell\nu$	9.2%	7.2%
$e^-e^+ \to qqqq\nu\overline{\nu}$	11.8%	9.0%
$e^-e^+ \to qq$	2.5%	1.4%
$e^-e^+ \to qq\ell\nu$	0.3%	0.1%
$e^-e^+ \to qq\ell\ell$	0.7%	0.4%
$e^-e^+ \to qq\nu\nu$	5.3%	3.1%
$e^-\gamma(BS) \to e^-qqqq$	0.8%	0.7%
$e^+\gamma(BS) \to e^+qqqq$	0.8%	0.7%
$e^-\gamma(EPA) \to e^-qqqq$	0.4%	0.4%
$e^+\gamma(EPA) \to e^+qqqq$	0.4%	0.3%
$e^-\gamma(BS) \to \nu qqqq$	20.3%	16.8%
$e^+\gamma(BS) \to \overline{\nu}qqqq$	19.3%	15.9%
$e^-\gamma(EPA) \to \nu qqqq$	9.4%	7.8%
$e^+\gamma(EPA) \to \overline{\nu}qqqq$	8.9%	7.3%
$e^-\gamma(BS) \to qqH\nu$	37.2%	30.2%
$e^+\gamma(BS) \to qqH\nu$	37.1%	30.2%
$e^-\gamma(EPA) \to qqH\nu$	19.0%	15.7%
$e^+\gamma(EPA) \to qqH\nu$	18.4%	15.2%
$\gamma(BS)\gamma(BS) \to qqqq$	1.9%	1.7%
$\gamma(BS)\gamma(EPA) \to qqqq$	1.1%	1.0%
$\gamma(EPA)\gamma(BS) \to qqqq$	1.1%	1.0%
$\gamma(EPA)\gamma(EPA) \to qqqq$	0.7%	0.6%

Table 7.21: List of signal and background samples with the selection efficiencies after loose cuts and mutually exclusive cuts at $\sqrt{s} = 3\,\text{TeV}$. The selection efficiencies are presented in a "flow" fashion, as the every selection cut contains all the cuts to the left of it.

Channel / Efficiency $\sqrt{s} = 3\,\text{TeV}$	N	$\varepsilon_{\text{presel}}$	ε_{MVA}	N_{MVA}
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$				
$b\bar{b}W^+W^-\nu_e\bar{\nu}_e$, hadronic	146.0	61.7%	11.6%	9.89
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$				
$b\bar{b}b\bar{b}\nu_e\bar{\nu}_e$	355.0	18.8%	1.5%	1.05
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$ other	675.0	20.0%	3.6%	4.51
$e^-e^+ \to q_l q_l H\nu\bar{\nu}$	6115.4	36.0%	0.4%	9.42
$e^-e^+ \to c\bar{c}H\nu\bar{\nu}$	2249.9	26.3%	0.5%	3.13
$e^-e^+ \to b\bar{b}H\nu\bar{\nu}$	2197.7	25.8%	1.2%	6.82
$e^-e^+ \to qqqq$	1093000.0	1.4%	0.01%	1.43
$e^-e^+ \to qqqq\ell\ell$	338600.0	0.6%	-	-
$e^-e^+ \to qqqq\ell\nu$	213200.0	7.3%	0.05%	8.35
$e^-e^+ \to qqqq\nu\bar{\nu}$	143000.0	9.0%	0.05%	6.35
$e^-e^+ \to qq$	5897800.0	1.4%	-	-
$e^-e^+ \to qq\ell\nu$	11121800	0.1%	-	-
$e^-e^+ \to qq\ell\ell$	6639200.0	0.4%	-	-
$e^-e^+ \to qq\nu\nu$	2635000.0	3.1%	-	-
$e^-\gamma(BS) \to e^-qqqq$	2004388.1	0.7%	-	-
$e^+\gamma(BS) \to e^+qqqq$	2002334.1	0.7%	-	-
$e^-\gamma(EPA) \to e^-qqqq$	575600.0	0.4%	-	-
$e^+\gamma(EPA) \to e^+qqqq$	575600.0	0.3%	-	-
$e^-\gamma(BS) \to \nu qqqq$	414750.0	16.8%	0.04%	30.7
$e^+\gamma(BS) \to \bar{\nu}qqqq$	414434.0	15.9%	0.05%	30.3
$e^-\gamma(EPA) \to \nu qqqq$	108400.0	7.8%	0.04%	3.37
$e^+\gamma(EPA) \to \bar{\nu}qqqq$	108400.0	7.3%	0.03%	2.63
$e^-\gamma(BS) \to qqH\nu$	92588.0	30.2%	0.2%	67.5
$e^+\gamma(BS) \to qqH\nu$	92430.0	30.3%	0.2%	54.2
$e^-\gamma(EPA) \to qqH\nu$	23400.0	15.4%	0.2%	7.88
$e^+\gamma(EPA) \to qqH\nu$	23400.0	15.2%	0.3%	10.2
$\gamma(BS)\gamma(BS) \to qqqq$	18009413.9	1.6%	-	-
$\gamma(BS)\gamma(EPA) \to qqqq$	3824548.1	1.0%	-	-
$\gamma(EPA)\gamma(BS) \to qqqq$	3828498.1	1.0%	-	-
$\gamma(EPA)\gamma(EPA) \to qqqq$	805400.0	0.6%	-	-

Table 7.22: List of signal and background samples selection efficiencies and event numbers after MVA at $\sqrt{s} = 3\,\text{TeV}$, for a luminosity of $2000\,\text{fb}^{-1}$. The number of events, selection efficiency of pre-selection, selection efficiency of MVA after pre-selection, number of events after MVA are shown. - represents a number less than 0.01.



(a) $HH \to b\bar{b}W^+W^-$, hadronic **(b)** $HH \to b\bar{b}b\bar{b}$

Figure 7.12: Sum of b tag against y_{34} at $\sqrt{s} = 3\,\text{TeV}$, shown for hadronic decay of $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$ and $HH \to b\bar{b}W^+W^-$ sub-channels.

7.12 $e^-e^+ \to HH\nu_e\bar{\nu}_e \to b\bar{b}W^+W^-\nu\bar{\nu}$ semi-leptonic decay at $\sqrt{s} = 3\,\text{TeV}$ analysis

Before interpreting the results, the final analysis is on the $e^-e^+ \to HH\nu_e\bar{\nu}_e \to b\bar{b}W^+W^-\nu\bar{\nu}$ semi-leptonic decay at $\sqrt{s} = 3\,\text{TeV}$. The corresponding semi-leptonic analysis at $\sqrt{s} = 1.4\,\text{TeV}$ was also performed. Since the selected event number is too low, there are not enough signal events to have a meaningful discussion. Hence, only the $\sqrt{s} = 3\,\text{TeV}$ analysis is presented.

The strategy of the analysis is very similar to the hadronic decay analysis. The main difference are that there is one lepton in the final state and the final state has 4 quarks instead of 6. H_{bb} and W can not be reconstructed due to the leptonic decay of one of the W. Hence, events are selected when there is one identified lepton using the same lepton identifiers. The jet reconstruction parameters are the same as the $\sqrt{s} = 3\,\text{TeV}$ hadronic decay analysis. The pre-selection cuts are the same, listed in table 7.23. There are no mutually exclusive cuts since there is no semi-leptonic analysis in the parallel analysis. The jet pairing still tries to reconstruct all the physical bosons. Parameters for the MVA are listed in table 7.24. The efficiencies of the pre-selection cuts, the efficiencies of the MVA selections are listed in table 7.25, alongside the number of events after the MVA selection. Since there are three neutrinos in the final state, reconstructing the correct event topology is more difficult. The MVA performance is worse and almost all

background channels survive the MVA. Nevertheless, dominant background channels are almost identical to those at $\sqrt{s} = 1.4\,\text{TeV}$. Refer to section 7.10 for discussion.

Pre-selection	$\sqrt{s} = 3\,\text{TeV}$
Discriminative pre-selection	$m_{HH} > 150\,\text{GeV}$, $B_1 > 0.2$, $p_{T\,HH} > 30\,\text{GeV}$
Loose cuts for MVA	$m_{H_{bb}} < 500\,\text{GeV}$, $m_{HH} < 3000\,\text{GeV}$

Table 7.23: Pre-selection cuts at $\sqrt{s} = 3\,\text{TeV}$ for semi-leptonic analysis.

7.13 Result interpretation

The results of the $\sqrt{s} = 1.4\,\text{TeV}$ and $\sqrt{s} = 3\,\text{TeV}$ analyses are summarised in table 7.26. N_S is all $e^-e^+ \to HH\nu_e\bar{\nu}_e$ events passed the MVA. The expected precisions on the cross sections, which is roughly $\sqrt{N_S + N_B}/N_S$ at $\sqrt{s} = 1.4\,\text{TeV}$ and $3\,\text{TeV}$, are:

$$\frac{\Delta\left[\sigma\left(HH\nu_e\bar{\nu}_e\right)\right]}{\sigma\left(HH\nu_e\bar{\nu}_e\right)} = \begin{cases} 179\%, & \text{at } \sqrt{s} = 1.4\,\text{TeV}, \\ 92\%, & \text{at } \sqrt{s} = 3\,\text{TeV}, \end{cases} \tag{7.10}$$

where $\sqrt{s} = 3\,\text{TeV}$ result combines hadronic and semi-leptonic decay sub-channels.

As previously stated, the double Higgs production cross section is sensitive to the Higgs triple self coupling λ. The relative uncertainty on the coupling can be related to

Category	Variable
Invariant mass	$m_{H_{bb}}$, m_W, m_{HH}
Energy and momentum	E_{mis}, $p_{T\,H_{bb}}$, $p_{T\,W}$, $p_{T\,HH}$
Angles in lab frame	θ_{mis}, $A_{H_{bb}}$, A_W, A_{HH}
Angles in boosted frames	$\cos\left(\theta^*_{H_{bb}}\right)$, $\cos\left(\theta^*_{HH}\right)$
Event shape	$\lvert S\rvert$, $-\ln(y_{23})$, $-\ln(y_{34})$, $-\ln(y_{45})$, $-\ln(y_{56})$
b and c tag	$B_{1,H_{bb}}$, $B_{2,H_{bb}}$, $B_{1,W}$, $C_{1,H_{bb}}$, $C_{1,W}$
Number of PFOs	$N_{H_{bb}}$, N_W

Table 7.24: Variables used in MVA for semi-leptonic analysis at $\sqrt{s} = 3\,\text{TeV}$.

Channel / Efficiency $\sqrt{s} = 1.4\,\text{TeV}$	N	$\varepsilon_{\text{presel}}$	ε_{MVA}	N_{MVA}
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$				
$b\bar{b}W^+W^-\nu_e\bar{\nu}_e$, semi-leptonic	96.8	44.6%	21.9%	13.11
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$				
$b\bar{b}b\bar{b}\nu_e\bar{\nu}_e$	355.0	13.3%	10.9%	5.38
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to$ other	724.2	13.1%	13.6%	12.75
$e^-e^+ \to q_l q_l H\nu\bar{\nu}$	6115.4	7.4%	13.7%	62.63
$e^-e^+ \to c\bar{c}H\nu\bar{\nu}$	2249.9	6.3%	12.1%	17.10
$e^-e^+ \to b\bar{b}H\nu\bar{\nu}$	2197.7	15.9%	5.1%	18.03
$e^-e^+ \to qqqq$	1093000.0	0.6%	0.2%	15.04
$e^-e^+ \to qqqq\ell\ell$	338600.0	1.0%	0.06%	1.85
$e^-e^+ \to qqqq\ell\nu$	213200.0	27.6%	0.5%	270.33
$e^-e^+ \to qqqq\nu\bar{\nu}$	143000.0	1.9%	1.6%	43.78
$e^-e^+ \to qq$	5897800.0	0.4%	0.3%	60.82
$e^-e^+ \to qq\ell\nu$	11121800	0.3%	0.08%	21.24
$e^-e^+ \to qq\ell\ell$	6639200.0	0.6%	0.2%	84.14
$e^-e^+ \to qq\nu\nu$	2635000.0	0.4%	0.9%	92.55
$e^-\gamma(BS) \to e^-qqqq$	2004388.1	1.2%	-	-
$e^+\gamma(BS) \to e^+qqqq$	2002334.1	1.2%	-	-
$e^-\gamma(EPA) \to e^-qqqq$	575600.0	1.1%	-	-
$e^+\gamma(EPA) \to e^+qqqq$	575600.0	1.1%	-	-
$e^-\gamma(BS) \to \nu qqqq$	414750.0	3.7%	1.5%	226.77
$e^+\gamma(BS) \to \bar{\nu}qqqq$	414434.0	3.5%	1.6%	225.68
$e^-\gamma(EPA) \to \nu qqqq$	108400.0	11.2%	0.9%	107.90
$e^+\gamma(EPA) \to \bar{\nu}qqqq$	108400.0	10.7%	0.8%	92.75
$e^-\gamma(BS) \to qqH\nu$	92588.0	7.9%	10.7%	779.36
$e^+\gamma(BS) \to qqH\nu$	92430.0	7.9%	10.1%	741.57
$e^-\gamma(EPA) \to qqH\nu$	23400.0	22.9%	6.9%	369.52
$e^+\gamma(EPA) \to qqH\nu$	23400.0	22.7%	7.2%	381.33
$\gamma(BS)\gamma(BS) \to qqqq$	18009413.9	0.4%	-	-
$\gamma(BS)\gamma(EPA) \to qqqq$	3824548.1	1.0%	-	-
$\gamma(EPA)\gamma(BS) \to qqqq$	3828498.1	1.0%	0.08%	28.85
$\gamma(EPA)\gamma(EPA) \to qqqq$	805400.0	1.1%	-	-

Table 7.25: List of signal and background samples selection efficiencies and event numbers after MVA for semi-leptonic analysis at $\sqrt{s} = 3\,\text{TeV}$, assuming a luminosity of 2000fb^{-1}. The number of events, selection efficiency of pre-selection, selection efficiency of MVA after pre-selection, number of events after MVA are shown. - represents a number less than 0.01.

Channel	N_S	N_B	$N_S/\sqrt{N_S+N_B}$
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to b\bar{b}W^+W^-\nu_e\bar{\nu}_e$, hadronic, $\sqrt{s} = 1.4\,\text{TeV}$	1.79	8.41	0.56
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to b\bar{b}W^+W^-\nu_e\bar{\nu}_e$, hadronic, $\sqrt{s} = 3\,\text{TeV}$	15.45	242.28	0.96
$e^-e^+ \to HH\nu_e\bar{\nu}_e \to b\bar{b}W^+W^-\nu_e\bar{\nu}_e$, semi-leptonic, $\sqrt{s} = 3\,\text{TeV}$	31.24	3612.39	0.52

Table 7.26: Number of signal and background events, and significance after MVA for all $HH \to b\bar{b}W^+W^-$ analyses.

the uncertainty on the coupling via:

$$\frac{\Delta\lambda}{\lambda} \approx \kappa \cdot \frac{\Delta\left[\sigma\left(HH\nu_e\bar{\nu}_e\right)\right]}{\sigma\left(HH\nu_e\bar{\nu}_e\right)}. \tag{7.11}$$

κ can be extracted by varying the λ and parameterising the cross section change at a general level. Figure 7.13 shows the cross section as a function of the coupling at generator level for $\sqrt{s} = 1.4\,\text{TeV}$ and $\sqrt{s} = 3\,\text{TeV}$. The negative gradient indicates that the dependence on λ experiences the destructive interference with other SM Feynman diagrams. At the SM λ value, the κ is 1.22 and 1.47 at $\sqrt{s} = 1.4\,\text{TeV}$ and $3\,\text{TeV}$ respectively. Since κ is extracted from the relation at generator level, the fully simulated reconstruction selection may favour certain Feynman diagrams, therefore affecting the sensitivity to λ .

Without electron polarisation, the uncertainty on the Higgs triple self coupling λ, from $e^-e^+ \to HH\nu_e\bar{\nu}_e \to b\bar{b}W^+W^-\nu_e\bar{\nu}_e$ analysis is:

$$\frac{\Delta\lambda}{\lambda} = \begin{cases} 218\%, & \text{at } \sqrt{s} = 1.4\,\text{TeV}, \\ 135\%, & \text{at } \sqrt{s} = 3\,\text{TeV}. \end{cases} \tag{7.12}$$

Since the Feynman diagrams for the double Higgs boson productions include t-channel WW-fusion, the cross section can be enhanced by using polarised electron beam. For

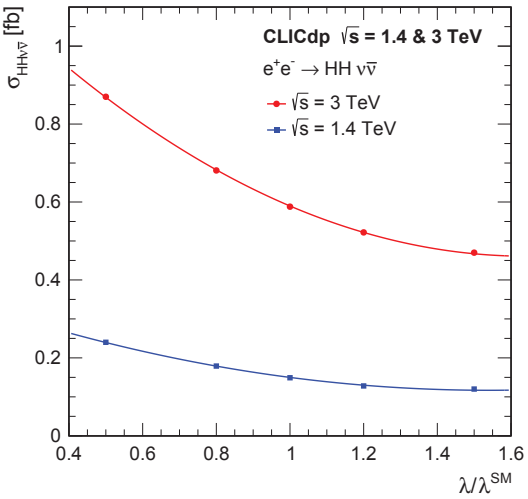


Figure 7.13: Cross section for the $e^-e^+ \to HH\nu_e\bar{\nu}_e$ process as a function of the ratio λ/λ_{SM} at $\sqrt{s} = 1.4\,\text{TeV}$ and 3 TeV, taken from [15].

$P(e^-) = 80\%$, the uncertainty of λ becomes:

$$\frac{\Delta\lambda}{\lambda} = \begin{cases} 163\%, & \text{at } \sqrt{s} = 1.4\,\text{TeV}, \\ 97\%, & \text{at } \sqrt{s} = 3\,\text{TeV}. \end{cases} \tag{7.13}$$

When both $\sqrt{s}$ are combined, the statistical precision on λ increases to 99% for the unpolarised beam, and 87% for the polarised beam with $P(e^-) = 80\%$.

7.14 Combined results

When $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$ sub-channels are combined, the expected precisions on the cross sections are:

$$\frac{\Delta\left[\sigma\left(HH\nu_e\bar{\nu}_e\right)\right]}{\sigma\left(HH\nu_e\bar{\nu}_e\right)} = \begin{cases} 44\%, & \text{at } \sqrt{s} = 1.4\,\text{TeV}, \\ 20\%, & \text{at } \sqrt{s} = 3\,\text{TeV}, \end{cases} \tag{7.14}$$

This translates to the uncertainty on the Higgs triple self coupling λ, without electron polarisation:

$$\frac{\Delta\lambda}{\lambda} = \begin{cases} 54\%, & \text{at } \sqrt{s} = 1.4\,\text{TeV}, \\ 29\%, & \text{at } \sqrt{s} = 3\,\text{TeV}. \end{cases} \tag{7.15}$$

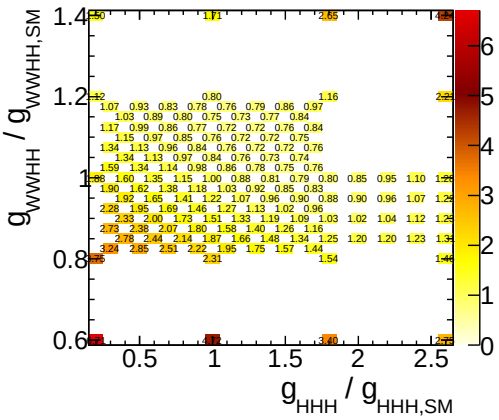


Figure 7.14: Normalised cross section for the $e^-e^+ \to HH\nu_e\overline{\nu}_e$ process as a function of the g_{HHH}/g_{HHHSM} and g_{WWHH}/g_{WWHHSM} at $\sqrt{s} = 3\,\text{TeV}$.

7.15 Simultaneous couplings extraction

The double Higgs production, $e^-e^+ \to HH\nu_e\overline{\nu}_e$, can occur via processes in figure 7.1. As previously stated, Figure 7.1a is sensitive to Higgs triple self coupling g_{HHH} while Figure 7.1b is sensitive to quartic coupling g_{WWHH}. Therefore, a simultaneous extraction on the coupling uncertainty can be performed by extending the method in the section 7.13. Once a relationship between g_{HHH}, g_{WWHH} and difference in kinematic variable distributions is established, a contour of the uncertainty in g_{HHH} and g_{WWHH} two dimensional phase space can be obtained.

This two dimensional template fitting is performed at $\sqrt{s} = 3\,\text{TeV}$, as the precision at $\sqrt{s} = 1.4\,\text{TeV}$ is too low to support such fitting. The luminosity is assumed to be 3000fb^{-1} to reflect the updated CLIC running scenario.

The normalised cross section of the $e^-e^+ \to HH\nu_e\overline{\nu}_e$ as a function of g_{HHH} and g_{WWHH} is shown in figure 7.14. The SM cross section is normalised to 1. Around the SM coupling value, the cross section increases with the decrease of g_{HHH} and the increase of g_{WWHH}. The cross sections along the anti-diagonal do not vary much, which would be difficult to precisely determine the statistical uncertainty on the coupling measurements.

To determine the uncertainty on the coupling measurements, the variables proposed in the generator-level study in section 2.9 are used: the invariant mass of the two Higgs system, m_{HH}, and the sum of their transverse momenta, H_T. By choosing kinematic bins, high-energy behaviour can be disentailed from the physics at threshold, allowing the extraction of the coupling strength g_{WWHH} and g_{HHH}.

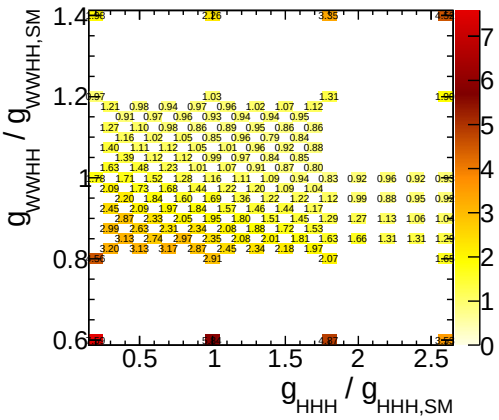


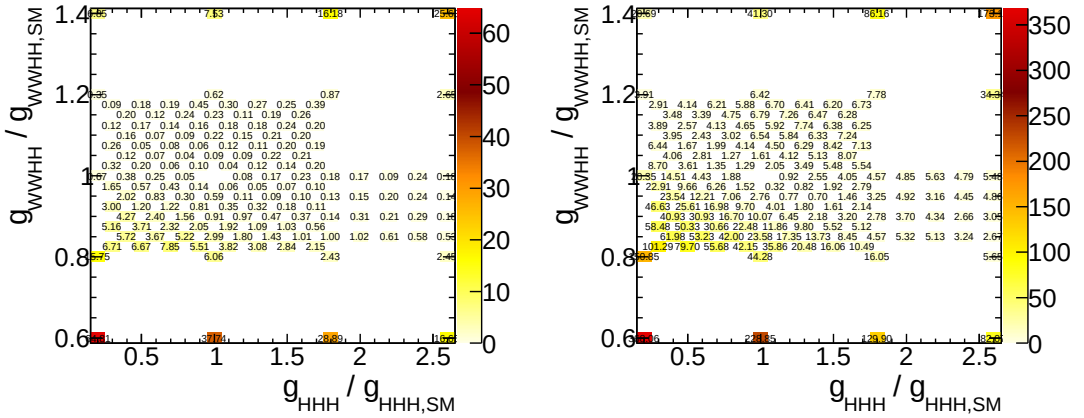
Figure 7.15: TODO TO UPDATE The significance for the $e^-e^+ \to HH\nu_e\bar{\nu}_e$ process as a function of the g_{HHH}/g_{HHHSM} and g_{WWHH}/g_{WWHHSM} at $\sqrt{s} = 3\,\text{TeV}$, using subchannel hadronic decay $HH \to b\bar{b}W^+W^-$, assuming a luminosity of 3000fb^{-1}.

The strategy for coupling extraction is described below. Simulated events with non-SM couplings are generated and reconstructed. These events went through the analysis chain discussed in this chapter with the same cuts and the same MVA classifier trained with the SM coupling sample. The signal significance of the double Higgs events with sub-channel hadronic decay $HH \to b\bar{b}W^+W^-$ as a function of g_{HHH} and g_{WWHH} is shown in figure 7.15.

The selected events are classified into 8 kinematic bins. 2 bins in H_T are cut at 200 GeV. 4 bins in m_{HH} are cut at 500, 700, 1000 GeV. A χ^2 function is constructed to access the difference of the m_{HH} and H_T distributions for non-SM coupling comparing to SM coupling sample. χ^2 is:

$$\chi^2 = \sum_i^{\text{bins}} \frac{(N_i - N_{i,\text{observed}})^2}{N_i}, \tag{7.16}$$

where N_i is the number of event expected in a kinematic bin i for a non-SM coupling sample. $N_{i,\text{observed}}$ is the number of event observed in a kinematic bin i. Here the observed set can be the SM coupling sample. The expression is summing over all kinematic bins. By construction, the SM coupling point has a χ^2 of 0. Figure 7.16 shows the χ^2 as a function of g_{HHH} and g_{WWHH} for two sub-channels; hadronic decay $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$. The χ^2 values for the $HH \to b\bar{b}b\bar{b}$ sub-channel are larger due to larger signal significance obtained with this sub-channel.



(a) $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$, hadronic **(b)** $HH \to b\bar{b}b\bar{b}$

Figure 7.16: TODO TO UPDATE The χ^2 for the $e^-e^+ \to HH\nu_e\bar{\nu}_e$ process as a function of the g_{HHH}/g_{HHHSM} and g_{WWHH}/g_{WWHHSM} at $\sqrt{s} = 3\,\text{TeV}$, using sub-channel hadronic decay $HH \to b\bar{b}W^+W^-$ and sub-channel, assuming a luminosity of 3000fb^{-1}.

Two sub-channels, hadronic decay $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$, are combined to increase the statistical precision on the coupling measurements. Two χ^2 surfaces are summed. To avoid statistical fluctuations in the sample, a toy MC experiment is performed. The SM coupling samples are treated as a data template set. 100000 data sets are generated by fluctuating the event number in each kinematic bin in the data template set according to Poisson distribution. The χ^2 is performed and summed using these generated data sets as the observed data. The summed χ is then averaged over the number of data sets (100000) and normalised such that the χ^2 at the SM coupling is 0. Since only the difference between the non-SM and SM χ^2 is used for the coupling measurements, the normalisation does not affect the measurements and helps to ease the visualisation. Figure 7.17 shows the normalised χ^2 after averaging over the toy MC experiments as a function of g_{HHH}/g_{HHHSM} and g_{WWHH}/g_{WWHHSM}. The χ^2 changes slowly along the anti-diagonal which is similar to the cross section plot.

Since there are two couplings in this χ^2 surface, the degree of freedom for this fit is 2. A contour of 68% confidence ($\chi^2 = 2.3$) can be drawn by interpolating between points on the surface. Figure 7.18 shows the contour. The counter can be sliced one dimensionally to extract the uncertainty of one coupling for a given value of the other coupling. For example:

$$\frac{\Delta g_{WWHH}}{g_{WWHH}} \simeq 4.9\% \text{ for } g_{HHH} = g_{HHHSM} \tag{7.17}$$

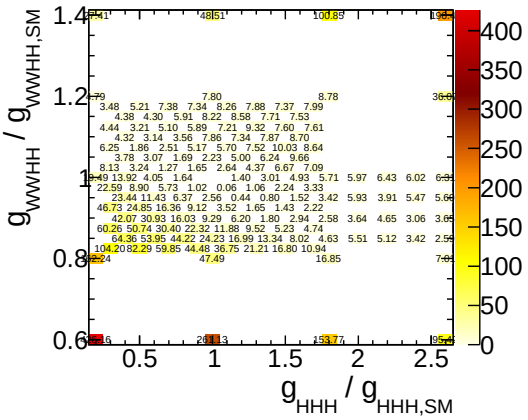


Figure 7.17: TODO TO UPDATE Normalised χ^2, after averaging the toy MC experiments, as a function of g_{HHH}/g_{HHHSM} and g_{WWHH}/g_{WWHHSM}, combining hadronic decay $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$ sub-channels, assuming a luminosity of $3000 fb^{-1}$.

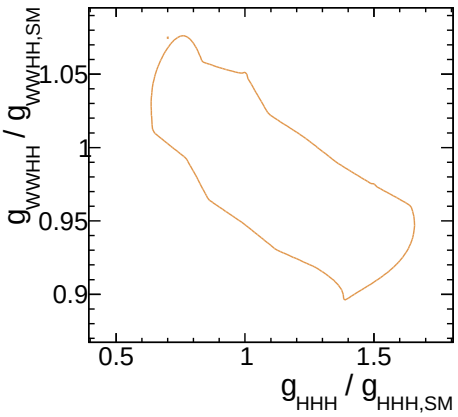


Figure 7.18: TODO Contour plot of χ^2, after averaging the toy MC experiments, as a function of g_{HHH}/g_{HHHSM} and g_{WWHH}/g_{WWHHSM}, combining hadronic decay $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$ sub-channels, assuming a luminosity of $3000 fb^{-1}$.

$$\frac{\Delta g_{HHH}}{g_{HHH}} \simeq 29\% \text{ for } g_{WWHH} = g_{WWHH_{SM}} \tag{7.18}$$

The statistical precisions on g_{WWHH} and g_{HHH} are much better at the CLIC than at the current LHC, or at the high luminosity upgraded LHC [12].

Colophon

This thesis was made in LaTeX 2_ε using the "hepthesis" class [79].

Bibliography

[1] ATLAS Collaboration, G. Aad *et al.*, Phys.Lett. **B716**, 1 (2012), 1207.7214.

[2] Particle Data Group, K. A. Olive *et al.*, Chin. Phys. **C38**, 090001 (2014).

[3] M. Thomson, *Modern particle physics* (Cambridge University Press, New York, 2013).

[4] D. Tong, Lectures on quantum field theory, 2006.

[5] B. Gripaios, Lectures on gauge field theory, 2017.

[6] SLD Electroweak Group, DELPHI, ALEPH, SLD, SLD Heavy Flavour Group, OPAL, LEP Electroweak Working Group, L3, S. Schael *et al.*, Phys. Rept. **427**, 257 (2006), hep-ex/0509008.

[7] S. Weinberg, Phys. Rev. Lett. **19**, 1264 (1967).

[8] D. Rainwater, Searching for the Higgs boson, in *Proceedings of Theoretical Advanced Study Institute in Elementary Particle Physics : Exploring New Frontiers Using Colliders and Neutrinos (TASI 2006): Boulder, Colorado, June 4-30, 2006*, pp. 435–536, 2007, hep-ph/0702124.

[9] D. B. Kaplan and H. Georgi, Phys. Lett. **B136**, 183 (1984).

[10] W. D. Goldberger, B. Grinstein, and W. Skiba, Phys. Rev. Lett. **100**, 111802 (2008), 0708.1463.

[11] G. F. Giudice, C. Grojean, A. Pomarol, and R. Rattazzi, JHEP **06**, 045 (2007), hep-ph/0703164.

[12] R. Contino, C. Grojean, M. Moretti, F. Piccinini, and R. Rattazzi, JHEP **05**, 089 (2010), 1002.1011.

[13] R. Contino, C. Grojean, D. Pappadopulo, R. Rattazzi, and A. Thamm, JHEP **02**,

006 (2014), 1309.7038.

[14] V. Barger, T. Han, P. Langacker, B. McElrath, and P. Zerwas, Phys. Rev. **D67**, 115001 (2003), hep-ph/0301097.

[15] H. Abramowicz *et al.*, (2016), 1608.07538.

[16] Y.-S. Tsai, Phys. Rev. **D4**, 2821 (1971), [Erratum: Phys. Rev.D13,771(1976)].

[17] CMS Collaboration, S. Chatrchyan *et al.*, Phys.Lett. **B716**, 30 (2012), 1207.7235.

[18] J. Brau *et al.*, (2007).

[19] L. Linssen, A. Miyamoto, M. Stanitzki, and H. Weerts, (2012), 1202.5940.

[20] H. Baer *et al.*, (2013), 1306.6352.

[21] M. Aicheler *et al.*, (2012).

[22] M. Thomson, Nucl.Instrum.Meth. **A611**, 25 (2009), 0907.3577.

[23] J. S. Marshall, A. MÍźnnich, and M. A. Thomson, Nucl. Instrum. Meth. **A700**, 153 (2013), 1209.4039.

[24] I. G. Knowles and G. D. Lafferty, J. Phys. **G23**, 731 (1997), hep-ph/9705217.

[25] M. Green, *Electron-Positron Physics at the Z*Studies in high energy physics, cosmology, and gravitation (Taylor & Francis, 1998).

[26] CALICE, C. Adloff, (2011), 1105.0511.

[27] B. Parker *et al.*, (2009).

[28] CALICE, JINST **7**, P04015 (2012), 1201.1653.

[29] A. Sailer, *Radiation and Background Levels in a CLIC Detector Due to Beam-beam Effects: Optimisation of Detector Geometries and Technologies* (Humboldt Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät I, 2012).

[30] W. Kilian, T. Ohl, and J. Reuter, European Physical Journal C **71** (2011).

[31] M. Moretti, T. Ohl, and J. Reuter, p. 1981 (2001), hep-ph/0102195.

[32] T. Sjostrand, (1995), hep-ph/9508391.

[33] OPAL, G. Alexander *et al.*, Z. Phys. **C69**, 543 (1996).

[34] S. Jadach, Z. Was, R. Decker, and J. H. Kuhn, Comput. Phys. Commun. **76**, 361 (1993).

[35] GEANT4, S. Agostinelli *et al.*, Nucl.Instrum.Meth. **A506**, 250 (2003).

[36] P. Mora de Freitas and H. Videau, p. 623 (2002).

[37] F. Gaede, Nucl. Instrum. Meth. **A559**, 177 (2006).

[38] J. S. Marshall and M. A. Thomson, Eur. Phys. J. **C75**, 439 (2015), 1506.05348.

[39] J. S. Marshall, Presentation on pandorapfa with lc reconstruction, https://github.com/PandoraPFA/Documentation/blob/master/Pandora_LC_Reconstruction.pdf, 2017.

[40] A. Sailer, Luminosities for ee, eg, and gg interactions, https://indico.cern.ch/event/233706/contributions/499053/attachments/390186/542711/130514_LuminosityNormalisation.pdf, 2013.

[41] D. Schulte, (1999).

[42] T. Barklow, D. Dannheim, M. O. Sahin, and D. Schulte, (2012).

[43] G. F. Sterman and S. Weinberg, Phys. Rev. Lett. **39**, 1436 (1977).

[44] S. Moretti, L. Lonnblad, and T. Sjostrand, JHEP **08**, 001 (1998), hep-ph/9804296.

[45] G. P. Salam, Eur. Phys. J. **C67**, 637 (2010), 0906.1833.

[46] A. Ali and G. Kramer, Eur. Phys. J. **H36**, 245 (2011), 1012.2288.

[47] M. Cacciari, G. P. Salam, and G. Soyez, Eur. Phys. J. **C72**, 1896 (2012), 1111.6097.

[48] M. Cacciari and G. P. Salam, Phys. Lett. **B641**, 57 (2006), hep-ph/0512210.

[49] S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber, Nucl. Phys. **B406**, 187 (1993).

[50] S. D. Ellis and D. E. Soper, Phys. Rev. **D48**, 3160 (1993), hep-ph/9305266.

[51] S. Catani, Y. L. Dokshitzer, M. Olsson, G. Turnock, and B. R. Webber, Phys. Lett. **B269**, 432 (1991).

[52] M. Battaglia and F. P., CERN Report No. LCD-Note-2010-006, 2010 (unpublished).

[53] M. Boronat, J. Fuster, I. Garcia, E. Ros, and M. Vos, Phys. Lett. **B750**, 95 (2015),

1404.4294.

[54] T. Suehara and T. Tanabe, Nucl. Instrum. Meth. **A808**, 109 (2016), 1506.08371.

[55] LCFI, D. Bailey *et al.*, Nucl. Instrum. Meth. **A610**, 573 (2009), 0908.3019.

[56] Linear Collider ILD Concept Group -, T. Abe *et al.*, (2010), 1006.3396.

[57] H. Aihara *et al.*, (2009), 0911.0006.

[58] A. Hocker *et al.*, PoS **ACAT**, 040 (2007), physics/0703039.

[59] E. Farhi, Phys. Rev. Lett. **39**, 1587 (1977).

[60] G. Hanson *et al.*, Phys. Rev. Lett. **35**, 1609 (1975).

[61] Y. Freund and R. E. Schapire, Journal of Computer and System Sciences **55**, 119 (1997).

[62] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*Springer Series in Statistics (Springer New York, 2009).

[63] B. Xu, Improvement of photon reconstruction in PandoraPFA, in *Proceedings, International Workshop on Future Linear Colliders (LCWS15): Whistler, B.C., Canada, November 02-06, 2015*, 2016, 1603.00013.

[64] G. Kačarević, Prelection for h $\rightarrow$ γ γ at 3 tev, https://indico.cern.ch/event/577810/contributions/2485070/attachments/1424897/2185427/GoranKacarevic.pdf, 2017.

[65] E. Segrè, *Nuclei and particles: an introduction to nuclear and subnuclear physics* (W. A. Benjamin, 1977).

[66] E. Longo and I. Sestili, Nucl. Instrum. Meth. **128**, 283 (1975), [Erratum: Nucl. Instrum. Meth.135,587(1976)].

[67] ALEPH collaboration, S. Schael *et al.*, Phys. Rept. **421**, 191 (2005).

[68] S. Berge, W. Bernreuther, and S. Kirchner, Phys. Rev. **D92**, 096012 (2015).

[69] DELPHI collaboration, P. Abreu *et al.*, Phys. Lett. **B267**, 422 (1991).

[70] F. Gaede and J. Engels, EUDET Report (2007).

[71] TMVA Core Developer Team, J. Therhaag, AIP Conf.Proc. **1504**, 1013 (2009).

[72] B. K. Bullock, K. Hagiwara, and A. D. Martin, Phys. Lett. **B273**, 501 (1991).

[73] S. Dittmaier *et al.*, (2012), 1201.3084.

[74] A. MÍźnnich, CERN Report No. LCD-Note-2010-009, 2010 (unpublished).

[75] CLICdp, A. Sailer and A. Sapronov, (2017), 1702.06945.

[76] S. Lukić, Forward electron tagging in the h $\rightarrow$ μ μ analysis at 1.4 tev, http://indico.cern.ch/event/262809/contributions/1595499/attachments/464689/643931/electronTagging.pdf, 2013.

[77] G. Milutinović-Dumbelović *et al.*, (2014), 1412.5791.

[78] C. Grefe, T. Lastovicka, and J. Strube, Light Higgs Studies for the CLIC CDR, in *Helmholtz Alliance Linear Collider Forum: Proceedings of the Workshops Hamburg, Munich, Hamburg 2010-2012, Germany*, pp. 258–264, Hamburg, 2013, DESY, DESY, 1205.3908.

[79] A. Buckley, The hepthesis LaTeX class.

List of figures

List of tables