Detectors and Physics at a 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

An electron-positron linear collider is an option for future large particle accelerator projects. Such a collider would focus on precision tests of the higgs boson properties. This thesis describes several studies related to the optimisation of high granular calorimeters. Three main areas were covered.

The performance of photon reconstruction is improved. Photon reconstruction algorithms were developed within PandoraPFA, a world-leading pattern-recognition software for particle flow calorimetry. A sophisticated pattern recognition algorithm was implemented, which uses the topological properties of electromagnetic showers to identify photon candidates and separate them from nearby particles. It performs clustering of the energy deposits in the detector, followed by topological characterisation of the clusters, with the results being considered by a multivariate likelihood analysis. This algorithm leads to a significant improvement in the reconstruction of both single photons and multiple photons in high energy jets.

Reconstruction and classification of tau lepton decay modes were studied. Tau decay products, such as photons, were reconstructed as separate entities. Utilising high granular calorimeters, the resolution of energy and invariant mass of the tau decay products is improved. A hypothesis test was performed for expected decay final states. A multivariate analysis was trained to classify decay final states with a data-driven machine learning method. The performance of tau decay classification is used for the electromagnetic calorimeter optimisation at the ILC or CLIC.

Sensitivity of higgs couplings at the CLIC was studied, using simulated double Higgs boson production. Algorithms were developed to

identify isolated high energy leptons, and results were fed into a multivariate analysis. The study was done for two CLIC energy scenarios. This sensitivity study of triple and quartic Higgs self-couplings is a part of scientific cases for the CLIC. This work provides further motivation for high granular particle flow calorimetry for a future electron-positron linear collider.

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

There are many people that I would to thank for their help in my pursuit of a PhD degree. First of all, I would like express my most sincere gratitude to my parents, for their finical support and moral support. They have been supporting me for all this many years. Especially, when the PhD study became an intense and stressful exercise, they were able to put up with me and not abandon me. During a few months when I was really worried about not able to finish the PhD program and facing unemployment, they talked me through and gave me much consoling when I needed.

The next person I would like to thank is my supervisor, Mark Thomson. I was lucky to follow him to embark an incredible journey on an exciting project. I have received much useful guidance from him on numerous occasions. On one occasion, which influenced me greatly, was in the very early stage of my PhD study. I managed to make improvements to some algorithms. However, a study suggested that my improved algorithms were not as good as a rival algorithm by a certain metric. Feeling defeated and eager to prove myself, I wanted to repeat the studies just to prove that my algorithms are better. Mark suggested that it is more important to have a project to understand physics, rather than competing for the best performance defined by some arbitrary metrics, which taught me the importance of having the right priority in work, rather than engaging in meaningless competition, however tempting it may be.

I would also like to thank John Marshall for his constant support over the last four years. A large part of the improvement in coding skills is because of the help from John. There was a couple of months, where I had written my working algorithms in ugly codes, and had to rewrite my codes to meet PandoraPFA code standard. This refactorisation exercise indeed taught me a lot about the C++ coding concepts, as well as good coding habits. It was also him who introduced me to the wonderful world of git, which I hated in the beginning. Nevertheless, I was fortunate to have John as my second supervisor and coding mentor.

I was also extremely fortunate to have Steven Green as my colleague and my cherished friend. Other than the lovely, occasionally frustrating, four years that we spent in the same office, I was privileged to spend two years with Steve sampling the fine ale from local pubs on a regular basis. After the infamous "gin" incident, which was a great night, we continued to share our love of ale and pork scratchings in a much more civilised fashion. I was also honoured to be the usher on Steve's wedding. The wedding was great. And we should have more boardgame nights.

Before moving on to external collaborators, I would also like to thank Joris de Vries for providing entertainments in the office, for embarking on numerous pub trips together, and for suffering together in the "ceiling" incident. I would also like to thank Jack Anthony and Andy Smith for enduring me in the same office, and the rest of the Cambridge HEP group for their support.

I would like to thank Philipp Roloff for his teaching on various techniques in a physics analysis; Rosa Simoniello for collaborating on the double Higgs production analysis. The analysis would take much longer to finish without their help. I would also like to thank André Sailer and Marko Petric for their support with the CLIC grid computing system. At this time of this thesis is written, I should probably still be the top user on the grid system, in terms of the cpu time, much thanks to their help. I also have to thank André for introducing me to Café de l'aviation. It was the best steak that I had in Europe. My gratitude also goes to Lucie Linssen, who was very kind to fund several of my trips to CERN. It was an enjoyable experience to work in CERN and it would be impossible without Lucie's support. I would also like to thank the rest of CLICdp group in CERN for the friendly and the useful collaboration during my PhD.

My friends in Cambridge, whom I probably see on daily basis, deserve my a lot of my appreciation. It is them who made my PhD study in Cambridge lively and fun. I am again very luck not only to gain a PhD degree after another four years in Cambridge, but also to gain a group of good friends.

Apart from all the people that I have thanked above, there are a few extra people who proof-read my thesis: David Arvidsson, Sophie Morrison, and Laure-Anne Vincent. Thank you for the constructive suggestionss on my thesis.

Because of all the people that I have thanked, and those who I forget to thank, I was privileged to be able to spend four years to research on a topic that is truly interesting.

Contents

'A Higgs-Boson walks into a church,
the priest says
"We don't allow Higgs-Bosons in here.".
The Higgs-Boson says
"But without me, how can you have mass?"'

— Reddit

Chapter 1.

Introduction

'The journey of a thousand miles begins with a single step.'
— Lao Zi, 604 BC - 531 BC

Future electron-positron linear colliders are capable of making precise measurements of the Higgs sector, as well as the top quark sector [23, 24]. At a high centre-of-mass energy, the collider could search for new physics such as supersymmertry particles, and measure rare events, such as double Higgs production events. These measurements wold be difficult for the current proton-proton collider, limited by the underlying QCD interaction. Therefore, it is important to optimise the design of the future particle detector for the linear colliders to improve the event reconstruction and to perform physics simulation studies to demonstrate the superiority of the linear collider.

The thesis begins with overview of relevant theories on particle physics in chapter 2. Firstly a brief review of the current best particle theory, Standard Model of Particle Physics, is provided, including a short overview of the quantum electrodynamics, quantum chromodynamics, and the elctroweak interaction. The focus of the Standard Model discussion is on the Higgs mechanism and the Higgs boson in Standard Model. The discussion then moves on to theories beyond the Standard Model, with an example of a general parametrisation of the Higgs theory. The last part of the chapter dedicated to the discussion on identifying a Higgs boson from vector bosons using tau pair decay channel.

In chapter 3, the detector models used in the thesis are described in details. A general overview of two future electron-positron linear colliders, the International Linear Collider

(ILC) and the Compact Linear Collider (CLIC), is provided. After a short discussion on the physics program for these future colliders, a discussion of the impact of physics and other requirements on the detector design is presented. Afterwards, the International Large Detector, one detector option for the International Linear Collider, is discussed in details, followed by overviews on each sub-detector in the International Large Detector. The chapter finishes with a discussion on the modified International Large Detector detector concept for the Compact Linear Collider, where the modifications of the detector are highlighted.

In the next chapter, chapter 4, the software for event simulation and event reconstruction is discussed, followed by a discussion on the analysis software. Future linear colliders share common software framework. Hence, shared software for simulation and reconstruction is discussed first, with an emphasis on the PandoraPFA, a world-leading pattern-recognition software for particle flow calorimetry. Some CLIC specific issues are highlighted afterwards. Analysis software, including jet algorithms, is presented. Lastly, the multivariate analysis is discussed in details, where different fitting models, optimisation, and overfitting are discussed.

Chapter 5 describes several PandoraPFA algorithms regarding photon reconstruction. One algorithm performs the initial photon forming and photon ID test. Three algorithms are developed for the photon fragment removals. And one algorithm is developed to split the accidently merged photons. The core of identifying the photon is a two dimensional peaking finding algorithm. Having discussed the algorithms, performances of these algorithms are provided. Comparison with event reconstruction without photon reconstruction is also provided.

In chapter 6, a classification of the tau lepton decay modes is presented. The analysis contains the sample selection, pre-selection cuts, and the use of the multivariate classifier for the classification. The performance of the tau decay mode classification will be given, followed by an ECAL optimisation study using the tau decay mode classification. Lastly, the tau decay mode classification is further used in a proof-of-principle analysis to demonstrate the ability to use the tau pair polarisation correlation as a signature for Higgs boson.

In chapter 7, a full CLIC_ILD detector simulation study has been performed for the double Higgs production channel, $e^+e^- \to HH\nu_e\overline{\nu}_e$, via W^+W^- fusion. Event generation and simulation will be discussed first. An overview of the analysis, including lepton finding and jet reconstruction, is presented, followed by an optimised multivariate analysis

to distinguish signal from background processes. The optimised event selection is used to derive an estimate of the uncertainty on the cross section of double Higgs production at the CLIC. The event selection is further exploited to provide an estimate of the uncertainty on the measurements of trilinear Higgs self coupling and quartic coupling at the CLIC.

Chapter 2.

Theoretical overview

'I believe it is impossible to be sure of anything.'
— Han Fei Zi, 280 BC - 233 BC

This chapter provides a theoretical overview, which will 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, which supplies the theoretical background for the physics analysis in chapter 7. Lastly a theoretical discussion on the tau pair polarisation correlations as signature of Higgs boson is presented, which motivates the study in chapter 6.

2.1. Overview of the Standard Model

The Standard Model (SM) is a quantum field theory concerning three fundamental interactions of nature: the electromagnetic, the weak, and the strong interactions. The SM also describes the interactions between 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 triumphs of particle physics. The recent discovery of the Higgs boson in 2012 [1, 2] further verified the theory. This chapter summarises the SM based on the reviews of the SM presented in [3–6].

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 mediate

the fundamental forces between particles. For example, the photon is the force carrier of the electromagnetic force. W^+, W^-, and Z bosons are the force carriers of the weak force. And the gluon, g, is the force carrier of the strong force.

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 the same mass and spin, but opposite charge. Leptons and neutrinos have three generations. Each generation has the same interaction properties, but different masses. Although neutrinos could not be directly detected, measurements of the Z decay-width strongly suggested three generations of neutrinos [7]. Leptons and neutrinos experience weak forces as well as electromagnetic forces.

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 experience all three fundamental forces described by the SM.

The SM has enjoyed a great success with theoretical predictions being experimentally verified. Some highlights included the discovery of the top quark in 1995 [8], the tau neutrino in 2000 [9], and the Higgs boson in 2012 [1]. 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 exists yet. Another issue is that the SM does not allow neutrino masses and mixings. Because of these issues not fully explained by the SM, there are many theories beyond the Standard Model (BSM) trying to provide an explanation for these issues. One such example is the generalisation of the Higgs theory to allow non-SM coupling strengths. Many BSM theories also predict that the Higgs couples to tau leptons more strongly then other leptons. Therefore, tau pair polarisation correlations as a signature for Higgs boson is discussed.

The overview of the Standard Model starts with the quantum electrodynamics, and its generalisation to quantum chromodynamics. The unification of electromagnetism and the weak interaction, the electroweak gauge theory, will be discussed. Afterwards, the Higgs mechanism and Yukawa couplings will be introduced to explain masses of bosons and fermions 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 moves to an explanation of possible Higgs theories beyond the Standard Model, with their Lagrangian of the Higgs interaction and observables.

Lastly the discussion is provided on the tau pair polarisation correlation as a signature for Higgs boson. This signature could be used to identify Higgs boson if an excess of tau pair decay events observed among all lepton pair decay events.

2.2. Notations and conventions

Natural unit is used in this thesis: $\hbar = c = 1$. The metric is mostly-minus, $\eta^{\mu\nu} = \text{diag}(1, -1, -1, -1)$. The Dirac gamma matrices are represented with γ^μ, with $\mu \in \{0, 1, 2, 3\}$. $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$. $\overline{\psi} = \psi^\dagger\gamma^0$. Einstein summation convention is also used in this thesis.

This set of notations allow 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 with 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,

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

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. Here $D_\mu = \partial_\mu + ieA_\mu$ and $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. There are two free parameters in QED: m, the electron mass, and e, the electron charge. The mass term for the photon, $\nu^2 A_{nu}A^\nu$, is forbidden by gauge invariance.

QED has been verified experimentally. One of its greatest prediction is the spin magnetic dipole moment of the electron, g_s, which is predicted to be 2 from the Dirac

equations, plus small corrections to the value that comes from the electron's interaction with virtual photons, so called higher "loop" corrections in Feynman diagrams. The current best calculated QED prediction states $g = 2.001159652181643(764)$ [10], where the best experimental value and uncertainty is $g = 2.00115965218073(28)$ [11]. Precise agreement of the theoretical prediction and the experimental value, 1 part in 10^{12}, is a success of the QED.

2.4. Quantum chromodynamics

Like QED, quantum chromodynamics (QCD) is a quantum field theory. It explains strong interactions. There are eight gauge bosons, called gluons, coupling to nine fermions, called quarks. Unlike 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; and $G^a_{\mu\nu}$ 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 in the $G^a_{\mu\nu}$, comparing to the $F_{\mu\nu}$ in the QED, indicates the non-Abelian nature of QCD.

2.5. The electroweak interaction

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

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

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

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

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

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

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

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

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

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

For mass eigenstates, W^+ and W^- bosons only couple to left-handed fermions. Z boson and photon couple to both left-handed and right-handed fermions. Hence, by inspection, W^1 and W^2 are associated with W^+ and W^-. On the other hand, mass eigenstates for Z boson and photon, Z_μ and A_μ, are mixtures of W^3_μ and B_μ. Z_μ and A_μ are defined as:

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

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

where θ_W is the Weinberg mixing angle [12], which is determined experimentally.

So far the SU(2)⊗U(1) gauge theory explains the parity violating nature of the weak interaction. However, explicit mass terms are not allowed in the gauge symmetry.

2.6. Higgs Mechanism

The higgs mechanism via spontaneous symmetry breaking introduces mass terms for electroweak bosons. A complex scalar Higgs field, Φ_H, is added to the electroweak Lagrangian. Φ_H transforms as a doublet of SU(2) with hypercharge $Y = \frac{1}{2}$. The Higgs Lagrangian is therefore:

$$\mathcal{L}_{Higgs} = \left(D_\mu \Phi_{\mathrm{H}}\right)^\dagger \left(D^\mu \Phi_{\mathrm{H}}\right) - \mu^2 \Phi^\dagger_{\mathrm{H}} \Phi_{\mathrm{H}} - \lambda \left(\Phi^\dagger_{\mathrm{H}} \Phi_{\mathrm{H}}\right)^2, \tag{2.14}$$

with

$$D_\mu \Phi_{\mathrm{H}} = \left(\partial_\mu + ig\frac{\tau_i}{2} W^i_\mu + ig'\frac{1}{2} B_\mu\right) \Phi_{\mathrm{H}}. \tag{2.15}$$

For a negative μ^2, the Higgs field potential:

$$\mu^2 \Phi^\dagger_{\mathrm{H}} \Phi_{\mathrm{H}} + \lambda \left(\Phi^\dagger_{\mathrm{H}} \Phi_{\mathrm{H}}\right)^2, \tag{2.16}$$

is minimised with a Higgs vacuum expectation value:

$$\sqrt{\Phi_{\mathrm{H}} \Phi^\dagger_{\mathrm{H}}} = \frac{\nu}{\sqrt{2}} = \sqrt{\frac{\mu^2}{2\lambda}}. \tag{2.17}$$

Without the loss of generality, one can choose:

$$\left\langle \Phi_{\mathrm{H}} \right\rangle = \begin{pmatrix} 0 \\ \frac{\nu}{\sqrt{2}} \end{pmatrix}, \tag{2.18}$$

with a real ν. Therefore, after the spontaneous symmetry breaking, the $\mathcal{L}_{Higgs}$ provides the mass terms for W^+, W^-, Z and photon via terms in the Lagrangian:

$$\frac{(g\nu)^2}{4} W_\mu^+ W^{-\mu} + \frac{\left(g^2 + g^{'2}\right)\mu^2}{8} Z_\mu Z^\mu. \tag{2.19}$$

This provides an equal mass for W^+ and W^-, a mass term for Z, and no mass term for photon.

2.7. Yukawa couplings

The previous section 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 are given by:

$$\mathcal{L}_{Yukawa} = -\lambda^u \overline{q_L} \Phi_H^c u_R - -\lambda^d \overline{q_L} \Phi_H d_R - \lambda^e \overline{l_L} \Phi_H e_R + h.c., \tag{2.20}$$

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

$$m_u = \frac{\lambda^u \nu}{\sqrt{2}},\ m_d = \frac{\lambda^d \nu}{\sqrt{2}},\ m_e = \frac{\lambda^e \nu}{\sqrt{2}}. \tag{2.21}$$

2.8. Standard Model Higgs boson

So far, interactions between different fields in the Standard Model, as well as the boson and fermion masses obtaining mechanism, have been discussed. The only thing left for discussion is the Higgs boson, and interactions between the Higgs boson and 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 degrees of freedom are manifestly eaten. The Higgs field becomes:

$$H(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \nu + h(x) \end{pmatrix}, \tag{2.22}$$

where $h(x)$ is the real scalar field of the Higgs boson. It is not charged under electromagnetism as it is real. The Higgs boson interaction terms with other particles in the Lagrangian can be shown by replacing ν with $\nu + h(x)$ in previous Lagrangians. For example, for a fermions field, ψ_i, the Higgs boson interaction term is given by:

$$\mathcal{L} \supset -\frac{m_i}{\nu} h \overline{\psi}_i \psi_i, \tag{2.23}$$

where the m_i is the mass of the fermion i. The Higgs boson interaction terms with bosons can be shown from the equation 2.19 as:

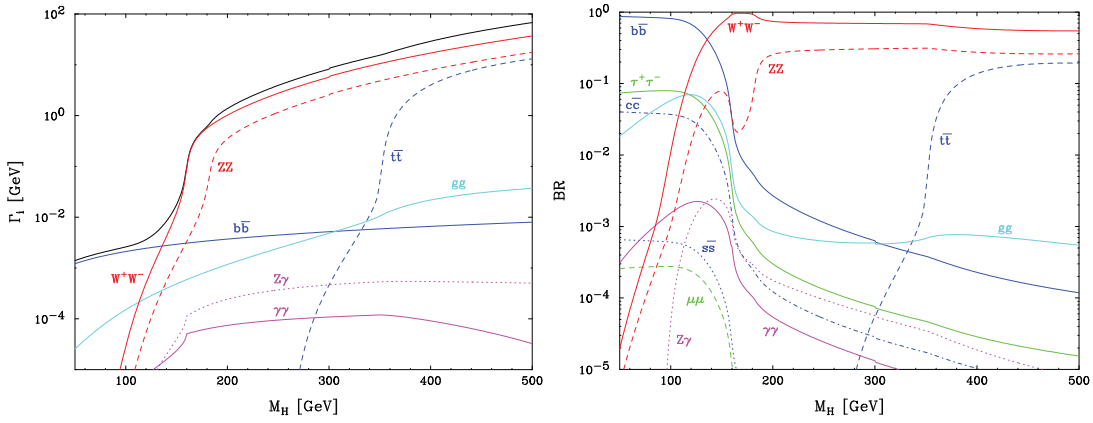
$$\mathcal{L} \supset m_W^2 \left(\frac{2h}{\nu} + \frac{h^2}{\nu^2} \right) W_\mu^+ W^{-\mu} + \frac{m_Z^2}{2} \left(\frac{2h}{\nu} + \frac{h^2}{\nu^2} \right) Z_\mu Z^\mu. \tag{2.24}$$

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

$$\mathcal{L} \supset \frac{\mu^2}{2} (\nu + h)^2 - \frac{\lambda}{4} (\nu + h)^4 \supset -\lambda \nu^2 h^2 - \lambda \nu h^3 - \frac{\lambda}{4} h^4 = -\frac{m_h^2}{2} h^2 - \frac{m_h^2}{2\nu} h^3 - \frac{m_h^2}{8\nu^2} h^4. \tag{2.25}$$

The Higgs boson mass, $m_{\mathrm{H}} = 2\lambda\nu^2$. The trilinear and quadlinear Higgs self interaction strengths are $-\frac{{m_{\mathrm{H}}}^2}{2\nu}$ and $\frac{{m_{\mathrm{H}}}^2}{8\nu^2}$, respectively. Once m_{H} is known, λ can be determined and the Higgs boson decay width and branching fraction can be approximately calculated. For example, figure 2.1a and figure 2.1b show the Higgs boson partial decay width and the branching ratios as a function of the Higgs boson mass.

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



(a) **(b)**

Figure 2.1.: Figure shows a) the Higgs boson partial decay widths, and b) selected Standard Model Higgs boson branching ratios. Both figures are shown a function of the Higgs boson mass, m_H. In figure a), the black curve shows the total decay width. Both figures are taken from [13].

2.9. Higgs beyond the Standard Model

Since the SM-like Higgs boson discovery in 2012, it has become important to understand the 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 the TeV scale. The SM Higgs boson 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 a low energy, but non SM Higgs behaviour at a high 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. 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 [14]. 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 [15]. In both scenarios, the interaction of Higgs becomes strong at a high energy. The coupling of the Higgs would deviate to those in the SM at a high energy.

An important physics channel for testing the Higgs theory is the double Higgs production via vector boson fusion at high energy [16–18]. For the composite Higgs scenario, the scattering amplitude increases with energy. For the dilaton scenario, no energy dependence 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 [17]. However, a multi-TeV linear electron-position collider, such as the Compact Linear Collider, would be able to precisely measure the cross section of the double Higgs production [19].

Following the assumption made in [17, 18], the self interaction of the light scalar Higgs, h, and its coupling to other SM bosons can be described by a Lagrangian using the notation in [18]. After the electroweak symmetry breaking, the bosonic part of the Lagrangian reads:

$$\mathcal{L} = \frac{1}{2}(\partial_\mu h) - V(h) + \left(m_W^2 W_\mu^+ W^{-\mu} + \frac{m_Z^2}{2} Z_\mu Z^\mu\right)\left[1 + 2a\frac{h}{\nu} + b\frac{h^2}{\nu^2} + \ldots\right], \tag{2.26}$$

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

$$V(h) = \frac{1}{2}m_h^2 h^2 + d_3\left(\frac{m_h^2}{2\nu}\right)h^3 + d_4\left(\frac{m_h^2}{8\nu^2}\right)h^4 + \ldots, \tag{2.27}$$

where a, b, d_3 and d_4 are 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 is the vector boson, $\mathrm{W}^\pm$ and Z. d_3 and d_4 are proportional to the trilinear and quadlinear h self coupling strength. Comparing with equation 2.24 and equation 2.25, the SM Higgs suggests $a = b = d_3 = d_4 = 1$, and all higher order terms vanish. Other BSM Higgs theory allow a, b, d_3, d_4 to take different values with different constraints. For example, the dilaton scenario imposes the relation, $a = b^2$.

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

$$A = a^2(A_{SM} + A_1\delta_b + A_2\delta_{d_3}), \tag{2.28}$$

where A_{SM} is the SM amplitude and:

$$\delta_b \equiv 1 - \frac{b}{a^2}, \tag{2.29}$$

$$\delta_{d_3} \equiv 1 - \frac{d_3}{a}. \tag{2.30}$$

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

$$\sigma = a^4\sigma_{SM}\Big(1 + A\delta_b + B\delta_{d_3} + C\delta_b\delta_{d_3} + D\delta_b^2 + E\delta_{d_3}^2\Big), \tag{2.31}$$

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 the threshold, allowing the extraction of δ_b, δ_{d_3} and hence the coupling strengths, g_{VVHH} and g_{HHH}. Suitable observables are variables that increase with the increasing of the centre-of-mass energies. Two examples of such variables are the invariant mass of the two Higgs system, m_{hh}, and the scalar sum of two Higgs transverse momenta, H_T. Figure 2.2 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 generator-level study performed in [18].

In the equation 2.31, a, which is proportional to g_{VVH}, only enters 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 ideally be a constant. For a multi-TeV electron-positron collider, the cross section for single Higgs production is far greater than that of the double Higgs production. Figure 2.3 shows the comparison of the cross section as a function of the centre-of-mass energy. Up to a centre-of-mass energy of $\sqrt{s} = 3\,\text{TeV}$, the cross sections of the signal Higgs are two orders of magnitude larger than the cross sections of the double higgs production. Therefore, g_{VVH}, and a, will be measured precisely using the single Higgs production before investigating the double Higgs production. Consequently, for the purpose of measuring g_{VVHH} and g_{HHH} via double Higgs production, a in the equation 2.31 can be treated as a constant.

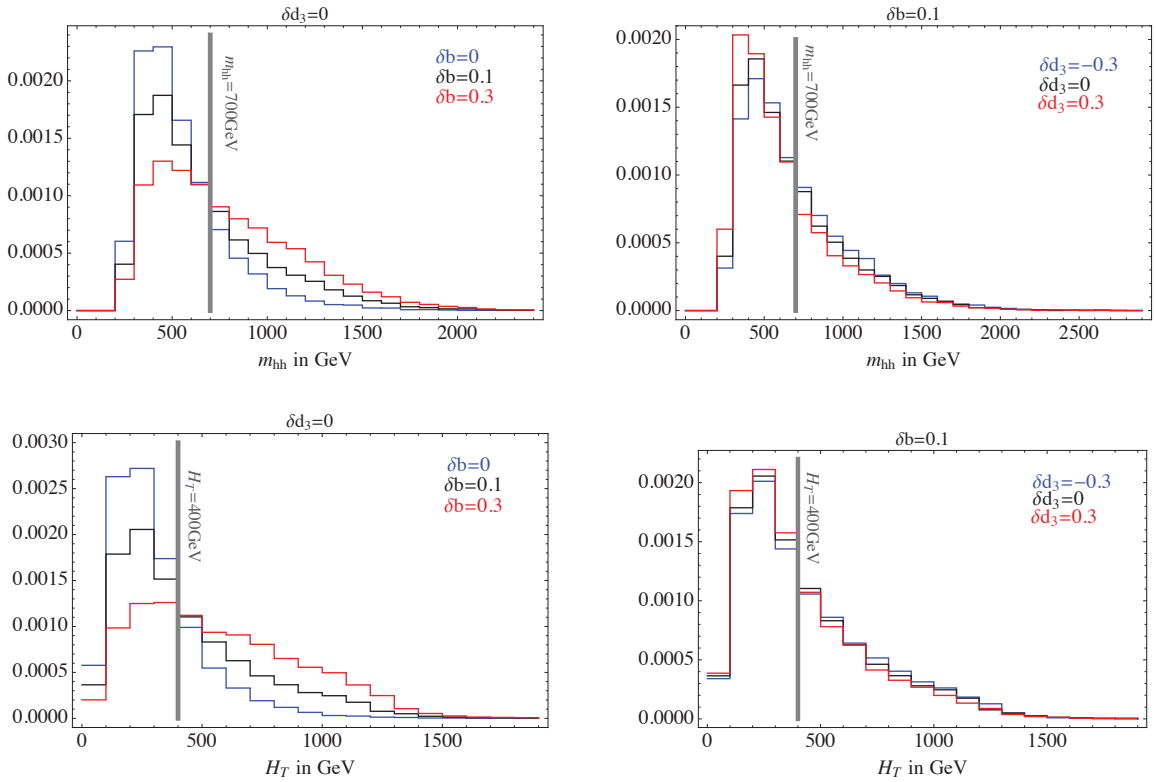


Figure 2.2.: Normalized differential cross sections $d\sigma/dm_{hh}$ and $d\sigma/dH_T$ for $e^+e^- \to \nu\bar{\nu}hh$ at the Compact Linear Collider, with $\sqrt{s} = 3\,\text{TeV}$ after the identification cuts, for several values of δ_b and δ_{d_3}. The plot is taken from [18].

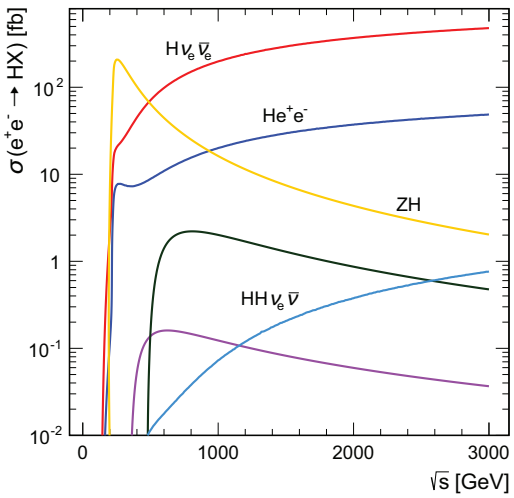


Figure 2.3.: Cross section as a function of centre-of-mass energy for the Higgs production processes at an electron-positron collider for a Higgs mass of 126GeV. The values shown correspond to unpolarised beams and do not include the effect of beamstrahlung. The plot is taken from [20].

2.10. Tau pair polarisation correlations as a 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. In these BSM theories, unlike vector bosons coupling to all flavours of leptons equally, the $H\tau^+\tau^-$ coupling would dominate the Higgs coupling to leptons. 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 [21] is followed and $\tau^- \to \pi^-\nu_\tau$ decay mode is used as the example. The boson to tau pair decay via $\tau^- \to \pi^-\nu_\tau$ can be represented as:

$$X \to \tau_\alpha^+\tau_\beta^- \to \pi^+\pi^- + \nu's, \tag{2.32}$$

where X is either H or Z; and 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:

$$\overline{z} = \frac{E_{\pi^+}}{E_{\tau^+}},\ z = \frac{E_{\pi^-}}{E_{\tau^-}}. \tag{2.33}$$

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

$$\frac{1}{\Gamma_\tau}\frac{d\Gamma}{dz} = Br_{\pi^-} f\left(\tau_\alpha^- \to \pi^-; z\right), \tag{2.34}$$

where Br_{π^-} is the branching fraction of $\tau^- \to \pi^-\nu_\tau$. The form f can be obtained from literature [22]:

$$f\left(\tau_\alpha^- \to \pi^-; z\right) = 1 + P_\alpha(2z-1), \tag{2.35}$$

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

$$\frac{d^2N(X \to \tau^+\tau^- \to \pi^+\pi^- + \nu's)}{dz\, d\bar{z}} = Br^2_{\pi^-} \sum_{\alpha,\,\beta} C^X_{\alpha\beta} f\Big(\tau^-_\alpha \to \pi^-; z\Big) f\Big(\tau^+_\beta \to \pi^+; \bar{z}\Big), \quad (2.36)$$

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

$$C^{\mathrm{H}}_{LL} = C^{\mathrm{H}}_{RR} = \frac{1}{2}, \quad (2.37)$$

and for $\mathrm{Z} \to \tau^+\tau^-$ the non-zero correlation coefficients are

$$C^{\mathrm{Z}}_{LR} = \frac{1}{2}\Big(1 - P_\tau\Big),\ C^{\mathrm{Z}}_{RL} = \frac{1}{2}\Big(1 + P_\tau\Big), \quad (2.38)$$

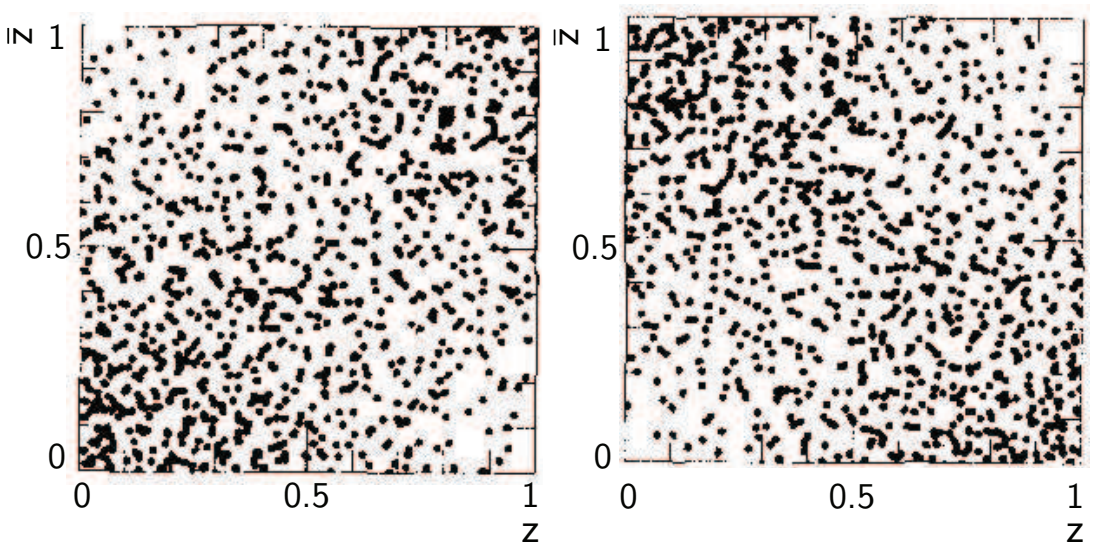
where tau polarisation, P_τ, arises from the non-parity-conserving tau decay in the SM. For Z:

$$P_\tau = \frac{-2va}{v^2 + a^2}, \quad (2.39)$$

$$v = -\frac{1}{2} + \sin^2\theta_{\mathrm{W}}, \quad (2.40)$$

$$a = -\frac{1}{2}, \quad (2.41)$$

where v and a are the vector and axial-vector $\mathrm{Z}\tau^+\tau^-$ couplings. Figure 2.4 shows two-dimensional distributions of tau pair polarisation correlation using both $\tau^- \to \pi^-\nu_\tau$ channel for $\mathrm{Z} \to \tau^+\tau^-$ and $\mathrm{H} \to \tau^+\tau^-$. The difference of the tau pair polarisation correlation between Z and H is clear. The distribution for $\mathrm{Z} \to \tau^+\tau^-$ has more entries along the diagonal, whilst the distribution for $\mathrm{H} \to \tau^+\tau^-$ has more entries along the anti-diagonal. Therefore, if an excess of $\tau^+\tau^-$ events over $\mathrm{e}^+\mathrm{e}^-$ or $\mu^+\mu^-$ events is observed, it can be easily identified where the excess is from H decay or Z decay, using the tau pair polarisation correlation.



(a) $Z \to \tau^+\tau^-$ **(b)** $H \to \tau^+\tau^-$

Figure 2.4.: Two-dimensional distribution of tau pair polarisation correlation using both $\tau^- \to \pi^-\nu_\tau$ channel for a) $Z \to \tau^+\tau^-$, and b) $H \to \tau^+\tau^-$. $\bar{z} = E_{\pi^+}/E_{\tau^+}$, $z = E_{\pi^-}/E_{\tau^-}$. Both figures are adapted from reference [22].

Chapter 3.

Detectors

'The great man is the one who does not lose his child's heart.'
— Mencius, 372 BC - 289 BC

Since the discovery of a particle consistent with the Standard Model Higgs boson at the LHC in 2012 [1, 2], the natural step for high energy physicists is to understand the Higgs. Yet limited by the underlying QCD interaction from proton-anti-proton collision, one has great difficulties in measuring the properties of the Higgs precisely. Next generation electron-positron linear colliders could hopefully make precise measurements in the Higgs sector, as well as the top quark sector [23, 24].

3.1. The ILC

Two leading candidates for next generation electron-positron linear colliders are the International Linear Collider (ILC) [23], and the Compact Linear Collider (CLIC) [24]. The ILC is a high-luminosity electron-positron linear collider with centre-of-mass energies from 200 GeV up to 1 TeV. The machine would be built at different stages. The first stage would have a centre-of-mass energy of 250/350 GeV. The second stage would have a centre-of-mass energy of 500 GeV with a possible upgrade to 1TeV. Thirty years of development leads to the technical design report in 2013 [25]. A layout of the collider complex is shown in figure 3.1. The proposal contains two detector concepts, the International Large Detector (ILD) and the Silicon Detector (SiD).

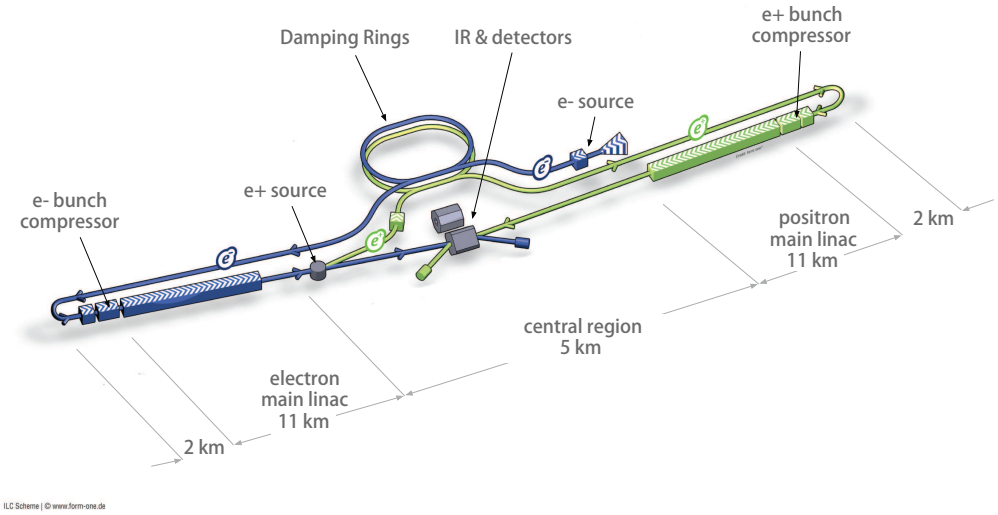


Figure 3.1.: A layout of the International Linear Collider complex, taken from [25].

3.2. The 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 to built in stages as well. The first stage, with a centre-of-mass energy of 380 GeV, is a compromise of precision measurement between top quark physics and Higgs physics. The final stage of a centre-of-mass energy of 3 TeV is motivated by the physics reach of detecting new physics and measuring rare decays of Higgs. The second stage has a centre-of-mass of 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 the CLIC detector concepts started with the ILC detector concepts. Two CLIC detector concepts, the CLIC_ILD and the CLIC_SiD, are developed based on the ILD and the SiD, respectively.

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, share 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, and search for new physics such as supersymmertry particles. The ILC can also operate at low

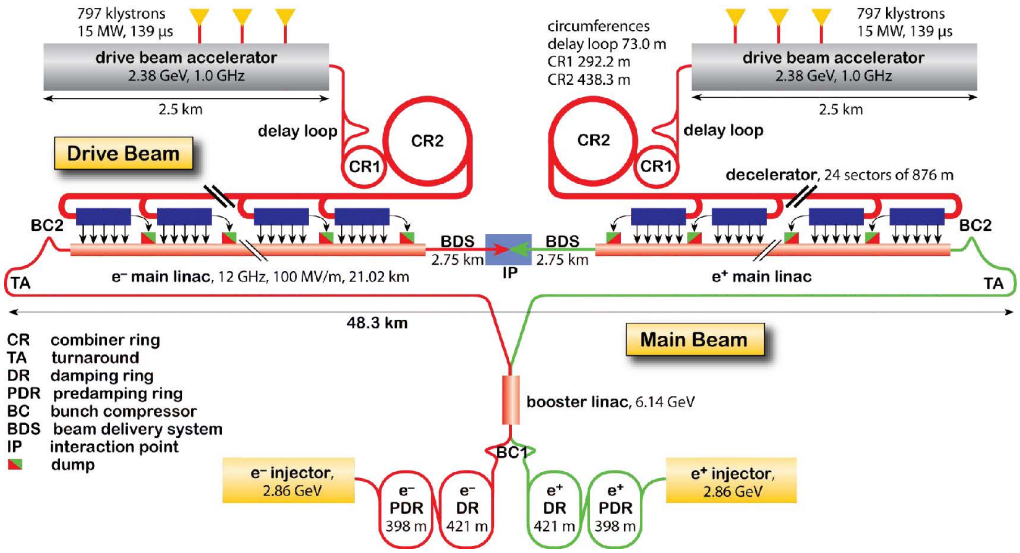


Figure 3.2.: A layout of the Compact Linear Collider at final stage of a centre-of-mass of energy of 3 TeV, taken from [26].

energy to be a Z and a H factory for ultra precise Z mass and H mass measurements. CLIC, on the other hand, has the advantage of a higher energy reach, which allows measurements of rare events, such as higgs trilinear self-couplings and quartic couplings.

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, using $W \to qq$ and $Z \to qq$ channels, by reconstructing the invariant mass via quark-jets [24, 25]. This translates to a requirement of 3.5-5% of the jet energy resolution. This level of precision is unlikely to be achieved with a traditional calorimetry design. A traditional energy flow approach 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 of the 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, the particle flow approach to calorimetry has demonstrated it ability to reach the goal [27, 28].

In a typical jet, using measurements of the particle composition from the LEP [29, 30], 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 combined. The jet energy resolution is thus limited by the energy resolution of the hadronic calorimeters, which typically is $\gtrsim 55\%/\sqrt{E(GeV)}$ [31].

The particle flow approach to calorimetry improves the jet energy resolution by fully reconstructing all visible particles in the detector. The jet energy is the sum of energy of individual particles, 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 approach, the hadronic calorimeter only measures about 10% of the energy, which would greatly improve 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 (hadron energy) is measured with $\sigma_E/E = 55\%/\sqrt{E(GeV)}$ [31], a jet energy of $\sigma_E/E = 19\%/\sqrt{E(GeV)}$ can be obtained. This satisfies the jet energy resolution requirement for separating W and Z via their hadronic decays. In reality, this level of performance is unattainable due to incorrect association of energy deposits to particles. At jet energies beyond tens of GeVs, the "confusion" rather than the intrinsic detector performance limits the particle flow performance.

The particle flow calorimetry requires to fully reconstruct particles and to associate calorimeter hits to tracks in tracking detectors. This is a demanding task 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. For the detector design, the detector needs to be highly granular for an excellent spatial resolution, which is needed to correctly associate calorimeter hits to the inner detector tracks. This imposes stringent requirements in the ECAL and the HCAL designs.

Figure 3.3 shows a typical topology of a 250 GeV jet, simulated with the CLIC_ILD detector concept. Particles consisting of 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, time projection chamber

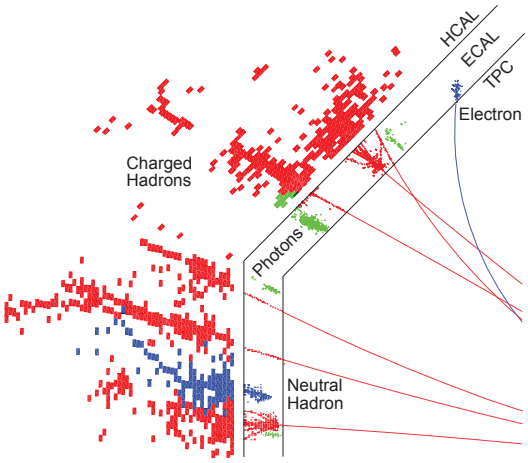


Figure 3.3.: A typical topology of a 250 GeV jet, simulated with thr CLIC_ILD detector concept, taken from [28].

(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. The inner tracking detector should be highly efficient and have 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 requirements for the detectors for the ILC and the CLIC are summarised in [24, 25]. Here important requirements are presented as motivations for detector designs.

The performance requirement of the vertex detector is determined by 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 in the form of:

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

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 $b \lesssim 15\mu m\, GeV$ [24, 31].

The requirement of tracking momentum resolution is driven by the Higgs boson mass resolution via the Higgsstrahlung process, $e^+e^- \to ZH$. The 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 $\sqrt{s} = 250\,\text{GeV}$, the momentum resolution needs to be $\sigma_{p_T}/{p_T}^2 \lesssim 5 \cdot 10^{-5} GeV^{-1}$ [31]. For the CLIC at high $\sqrt{s}$, the momentum resolution needs to be $\sigma_{p_T}/{p_T}^2 \lesssim 2 \cdot 10^{-5} GeV^{-1}$ [24].

The lepton identification should be over 95% for effective lepton tagging. The forward converge of the detector should be down to a very low angle, i.e. a few mrad, with respect to the beam axis. This is more critical for the CLIC as particles are boosted at a high centre-of-mass energy.

3.5. The International Large Detector

Two detector concepts have been designed for the ILC to deliver the physics program. The motivation for two detectors is to have multiple independent measurements within one collider for cross-checking, complementary measurements, and competition between collaborations. The two detectors are both designed to general purpose detectors. The Silicon Detector, SiD, is a compact detector with a large magnetic filed of 5 T. It uses silicon tracking modules. The second detector, 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

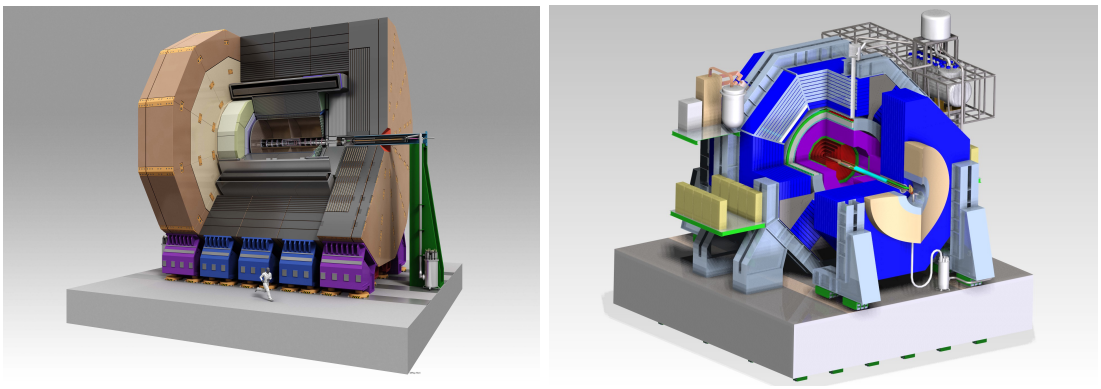
The International Large Detector, the ILD, is a detector concept at the ILC. The ILD detector concept has been optimised in the view of the particle flow techniques. Particle flow approach to event reconstruction has shown to deliver the best possible jet reconstruction with proof-of-principle implementation such as PandoraPFA (see chapter 4). Each individual particle is reconstructed with the particle flow algorithms. For charged particles, calorimeter hits are associated with the tracks. Therefore measurements of charged particles rely on an excellent tracking system resolution. Neutral particles reconstruction requires a high spatial resolution of the calorimeters.

The particle flow paradigm requires topological information of individual particle reconstructions. The sub-detector systems need to have the spatial resolution to separate charged particles from neutral particles. The result is a highly granular calorimeter

and a central tracking system with excellent momentum resolution. Longitudinal cross section of top quadrant of the ILD detector concept is shown in figure 3.11a. From the interaction point (IP) outwards, there is a tracking system comprising a large time projection chamber (TPC) augmented with silicon tungsten layers, highly granular electromagnetic calorimeters (ECAL) and hadronic calorimeters (HCAL), muon chambers, forward calorimeters (FCAL), magnetic coils and iron yokes.

The section below will describe the sub-systems of the ILD detector concept referred to as the ILD_o1_v05 option in MOKKA simulation in the ILD technical design report [25]. This detector concept has been optimised [32] and used in studies described in subsequent chapters.

The CLIC_ILD detector concept for the CLIC in the conceptual design report [24] is a modified version of the ILD, adapted to the CLIC colliding environment. As they share similarities, an overview of the ILD sub-detectors is provided, followed by a discussion on the difference between the ILD and the CLIC_ILD detector concepts. A comparison of the longitudinal cross section of the top quadrants of between ILD and CLIC_ILD can be seen in figure 3.11. A comparison of the key parameters between the ILD and the CLIC_ILD can be found in table 3.3.



(a) (b)

Figure 3.4.: Figure shows a) the International Large Detector, and b) the Silicon Detector for the International Linear Collider. Both figures are taken from [25].

3.6. Overview of ILD sub-detectors

The ILD detector concept is designed as a general purpose detector. As shown in figure 3.11a, closest to the interaction points are a precision vertex detector and a tracking

system. The tracking system consists of silicon tracking components and 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 also 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 vertices. As the TPC is the main tracking detector, the VTX mainly measures the impact parameter of tracks. Its structure is of three double layers with a barrel geometry. The double layers lower the material budget and improves the impact parameter measurements. The first double layer is of the half length of the other two, to avoid the high occupancy region of direct low momentum hits from the incoherent pair background. The baseline geometry of the vertex detector can be found in table 3.1.

	R	$\lvert z\rvert$	$\lvert\cos(\theta)\rvert$	σ	Readout time
Layer 1	16 mm	62.5 mm	0.97	2.8 μm	50 μs
Layer 2	18 mm	62.5 mm	0.96	6 μm	10 μs
Layer 3	37 mm	125 mm	0.96	4 μm	100 μs
Layer 4	39 mm	125 mm	0.95	4 μm	100 μs
Layer 5	58 mm	125 mm	0.91	4 μm	100 μs
Layer 6	60 mm	125 mm	0.90	4 μm	100 μs

Table 3.1.: Vertex detector parameters. The spatial resolution (σ) and readout times are for the CMOS option. The table is adapted from [31].

Figure 3.5 shows the impact parameter resolution as a function of the particle momentum for two different particle production angles. The desired impact parameter resolution (dashed line) is achievable.

3.6.2. Tracking Detectors

The hybrid tracking system consists of a large 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

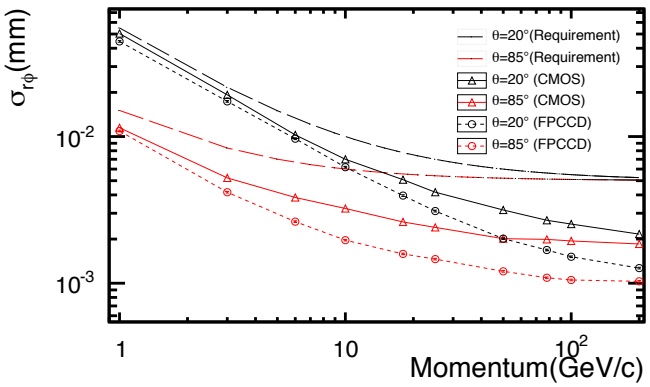


Figure 3.5.: Impact parameter resolution of the ILD vertex detector for two different particle production angles (20°and 85°), assuming the baseline point resolution given in table 3.1 for CMOS option (solid line), and the FPCCD option (dotted line). The curves with long dashes show the performance goal. The figure is taken from [31].

(ETD) in the forward region. The SIT, SET, and ETD are made up of 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 3.6. The main parameters of the silicon system and the TPC can be found in table 3.2.

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 performance and the best particle flow performance.

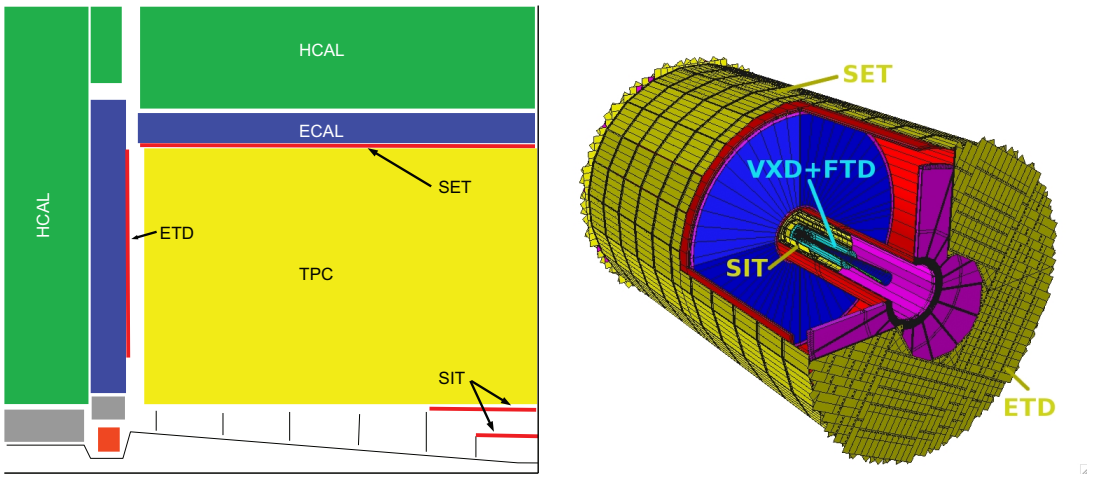
The barrel silicon trackers improve the overall momentum resolution. They provide additional high precision space points and additional redundancy between the TPC, the VTX, and the calorimeters. The ETD provide the low angle coverage, which is not covered by the TPC.

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 fine granular ECAL is located inside the HCAL. The ECAL measures photon energies and separates

	R	z	cos (θ)
SIT	153 mm	368 mm	0.910
SIT	300 mm	644 mm	0.902
SET	1811 mm	2350 mm	0.789
ETD	419-1822.7 mm	2420 mm	0.985-0.799
TPC	329-1808 mm	± 2350 mm	up to 0.98

Table 3.2.: Main parameters of the central silicon systems (SIT, SET, and ETD) and the TPC. The table is adapted from [31].



(a) **(b)**

Figure 3.6.: Plots for a) a top quadrant view of the ILD silicon envelope system, SIT, SET, ETD, and ETD, with TPC, ECAL, and HCAL, and b) a 3D detailed GEANT 4 simulation description of the silicon system as sketched in the quadrant view in a). Both plots are adapted from figures in [31].

photons from other particles. The ECAL also hosts the first part of the hadronic showers and greatly assists the separation of hadronic showers.

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 detailed 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 3.7a. 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. The choice of thin silicon layers offers a great spatial resolution at a cost of the energy resolution in favour of the particle flow. These silicon pads of 5.1 by 5.1 mm cover large areas, which are simple and reliable to operate.

The longitudinal segregation is a compromise between the cost and the performance. The total of 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 4.2 mm thick absorber plates. The test beam data with electrons shows the energy resolution of the ECAL concept to be $16.6/\sqrt{E(\ GeV)} \oplus 1.1\%$ [31], which is compatible with the values assumed for the full ILD detector simulation.

The optimisation of the ECAL design as a function of the number of longitudinal layers is performed, whilst keeping other geometry constant, using the jet energy resolution. The 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 $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$ 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 from semi-leptonic decay of heavy quarks. Using 90% of the entries is robust and focus on the Gaussian part of the distribution. The total jet energy is sampled at 91, 200, 360 and 500 GeV. Figure 3.7b shows the jet energy resolution for a single jet. For a 45 GeV jet, a degradation of 10% in the jet energy resolution is observed when the number of layers decreases from 30 to 20. The degradation in the jet energy resolution is

significant for number of layers fewer than 20, although the impact is smaller for high energy jets.

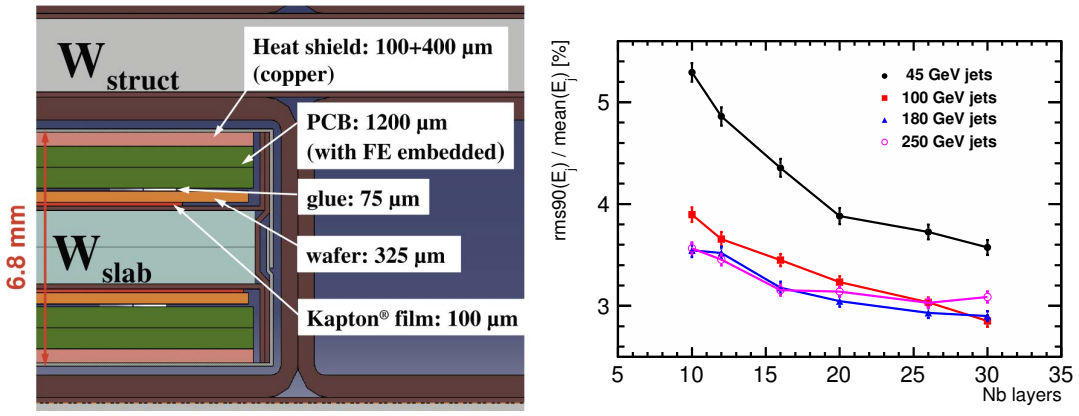


Figure 3.7.: Plot shows a) a cross section through the electromagnetic calorimeter layers, and b) jet energy resolution as a function of the total jet energy using $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$ sample at barrel region for optimisation of the ECAL design as a function of the number of longitudinal layers. Both plots are taken from [31].

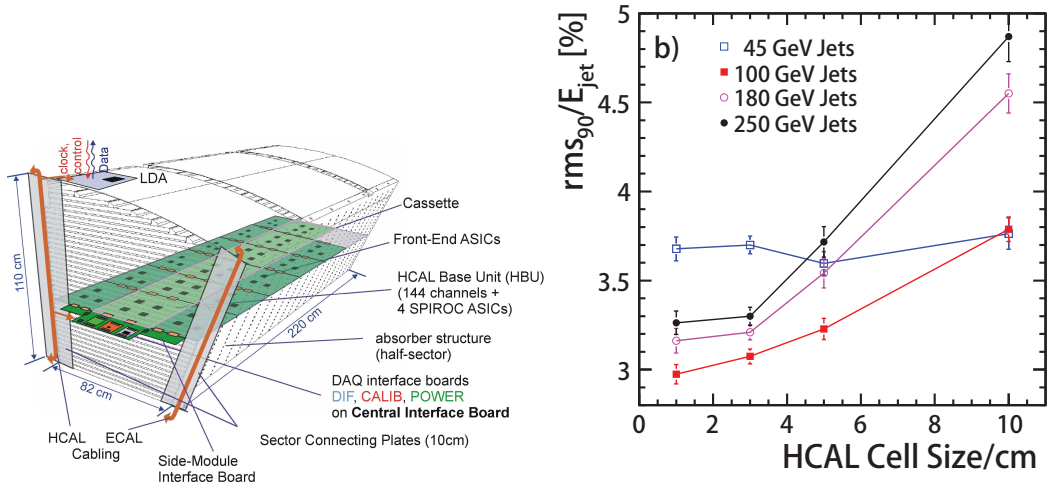
3.6.4. Hadronic Calorimeter

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

The principal 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 3 cm scintillator tiles, using an analogue read out system. The layout of a technological prototype, the "EUDET prototype" [33] is shown in figure 3.8a.

The longitudinal system including the ECAL provides about 6 radiation lengths, which is sufficient to contain the hadronic showers. The transverse cell sizes has been optimised for the best jet energy resolution. The jet energy resolution as a function of HCAL scintillator cell size for different jet energies is shown in figure 3.8b. There is no substantial gain in the jet energy resolution for cell sizes below 3 cm. However, the jet energy resolution degrades for cell sizes above 3 cm. Hence 3 cm cell size is chosen for the HCAL design.

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 3.8.: Figure a) shows the schematic view of a CALICE AHCAL technological prototype module. Figure b) shows the jet energy resolution as a function of the hadronic calorimeter scintillator cell sizes, with different energies. Both figures are taken from [25].

3.6.5. Solenoid, Yoke and Muon system

A large superconducting solenoid outside the calorimeters produces a nominal 3.5 T magnetic field. An iron yoke is instrumented with scintillator strips as active layers. The yoke returns the magnetic flux, and also acts as a muon detector and tail catcher calorimeter at the same time. The layout is shown in figure 3.9. The maximum magnetic field at 15 m radial distance from the detector is 50 Gauss to ensure safety [34]. A highly efficient muon detector is provided by the 3 by 3 cm scintillator strips. The first layer of the muon detector, also acting as a tail catcher calorimeter, catches the energy leakage from the HCAL and the ECAL. It has been shown that a 10% improvement of single particle energy resolution is possible with the tail catcher [35].

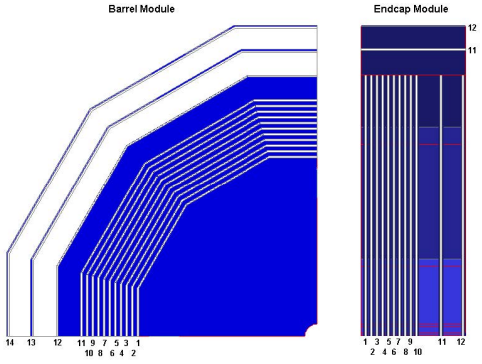


Figure 3.9.: Sensitive layers of the ILD muon system, taken from [25].

3.6.6. 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 a 500 GeV centre-of-mass energy. The beam calorimeter (BeamCAL), which is hit by many beamstrahlung pairs after each bunch corssing, extends the forward coverage. The BeamCAL also estimates a bunch-by-bunch luminosity. An additional hadron calorimeter at the forward region, LHCAL, extends the angular coverage of the HCAL to that of the LumiCAL. Electron tagging is possible with the very forward calorimeters [36], which aids event reconstruction at a high centre-of-mass energy.

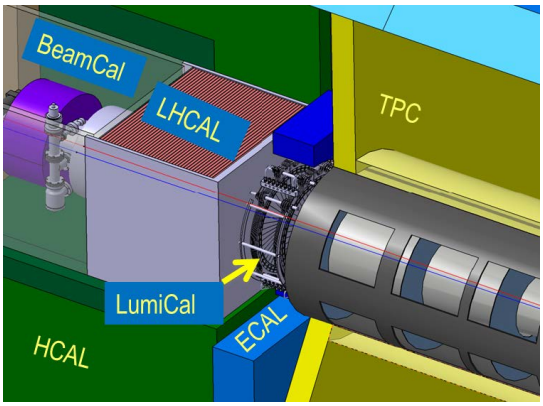


Figure 3.10.: The forward calorimeters of the ILD, taken from [25]. The LumiCAL, the BeamCAL, and the LHCAL are the luminosity calorimeter, the beam calorimeter, and the forward hadronic calorimeter, respectively.

3.7. The CLIC versus the ILC

The two main differences between the CLIC and the ILC are the high centre-of-mass energy and the high bunch charge density at the CLIC, which leads to significant beam related backgrounds. At the CLIC, within a bunch train, there is 0.5 ns between bunch crossings. There are two main sources of beam induced background at the CLIC colliding environment: 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.

3.7.1. The CLIC_ILD versus the ILD

There are two detector concepts studied in the CLIC conceptual design report [24], the CLIC_ILD and the CLIC_SiD. The CLIC_ILD detector concept is based on the ILD design. The CLIC_ILD and ILD share similarities due to similar physics motivations. Only the differences are highlighted here. A comparison of the CLIC_ILD and the ILD longitudinal cross sections can be seen in figure 3.11. A comparison of key parameters of the ILD and the CLIC_ILD detector concepts is shown in table 3.3.

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

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

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

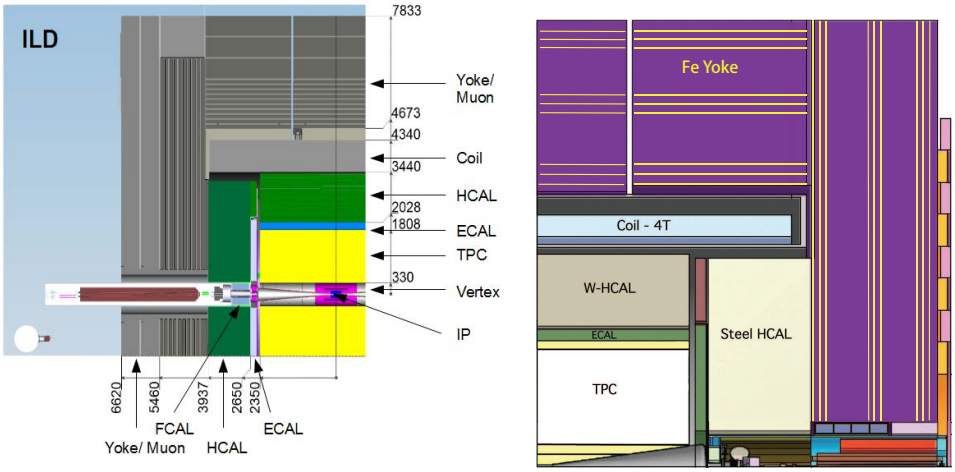
Concept	ILD	CLIC_ILD
Tracker	TPC/Silicon	TPC/Silicon
Solenoid Field (T)	3.5	4
Solenoid Field Bore (m)	3.3	3.4
Solenoid Length (m)	8.0	8.3
VTX Inner Radius (mm)	16	31
ECAL r_{min} (m)	1.8	1.8
ECAL Δr (mm)	172	172
HCAL Absorber B / E	Fe / Fe	Fe / W
HCAL Interaction Length	5.5	7.5
Overall Height (m)	14.0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

Table 3.3.: 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 [24].

For the HCAL at the CLIC_ILD, extra layers are added to contain the hadronic shower at a high centre-of-mass energy. The increased thickness is justified by the simulation studies [24], 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 is therefore increased to 230 cm.

The CLIC_ILD adopted a similar very forward calorimetry system as that of the ILD. The dimensions of the elements are changed due to a difference in the crossing angle (20 mrad for the CLIC and 14 mrad for the ILC). A comparison of the LumiCAL and the BeamCAL at the ILD and the CLIC_ILD is shown in table 3.4.



(a) **(b)**

Figure 3.11.: The longitudinal cross section of top quadrant of a) the ILD, taken from [25], and b) the CLIC_ILD, taken from and [24]. For both plots, from interaction point (IP) outwards, there is a tracking system comprising 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. The numbers are in units of mm.

		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 3.4.: Comparison of the LumiCAL and the BeamCAL at the ILD and the CLIC_ILD. The table is adapted from [24].

Chapter 4.

Simulation, reconstruction, and analysis software

'All the world's a stage, And all the men and women merely players; They have their exits and their entrances, And one man in his time plays many parts.'

— William Shakespeare, 1564 - 1616

In previous chapters, overviews of the particle physics theory and the detectors of the future linear collider experiments have been described. In this chapter, simulation, reconstruction and analysis software are discussed. Automated analysis is the only way to deal with the vast amount of data generated in high energy physics. Hence the software supporting the automated analysis are important. An analysis often consists of monte carlo event generation, event reconstruction, and using software to extract information of the event. Hence they are discussed together in this chapter.

Simulation and reconstruction software of the ILC and the CLIC shares a common software framework. Therefore, the shared simulation and reconstruction software is discussed first, and the CLIC specific issues are highlighted afterwards. The event reconstruction focuses on the PandoraPFA event reconstruction, which is the framework for the photon reconstruction algorithms in chapter 5. Lastly analysis software is presented. The multivariate analysis is discussed in lengthy details due to its complexity.

4.1. Monte Carlo event generation

Monte Carlo (MC) event generation is often the first step for the simulated study. Most events used in this thesis are generated with the WHIZARD software [37, 38]. Some simple events used in this thesis are generated by writing the event manually in the HEPEVT format [39]. The PYTHIA software [40] is used to describe parton showering, hadronisation and fragmentation. The parameters for the PYTHIA are tuned to OPAL data from the Large Electron-Positron Collider (LEP) [41]. The TAUOLA software [42] is used to describe the tau lepton decay with correct spin correlations of the decay products. The Initial State Radiation (ISR) effect is simulated in the WHIZARD, with the ISR photons being collinear with the beam direction. The Final State Radiation (FSR) is simulated in the PYTHIA.

4.2. Event Simulation

For all the simulated events used in this thesis, the simulation software used to simulate the interaction of particles through the detector material is the GEANT4 software [43]. The detector geometry description is provided by the MOKKA software [44]. The QGSP_BERT physics list is used to describe the hadronic shower decay in the detector.

4.3. Event Reconstruction

With simulated events (or real data in the future) as inputs, the next step is to reconstruct these events. The reconstruction software runs in the Marlin framework [45], as a part of the iLCSoft software package. The 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 [27, 28]. Details of the reconstruction can be found in [23, 24]. Here particle flow reconstruction via PandoraPFA will be discussed in details, as PandoraPFA provides the software framework for the photon reconstruction in chapter 5. The PandoraPFA event reconstruction is also used in the analyses in chapter 6 and chapter 7.

4.4. PandoraPFA event reconstruction

The tradition energy flow approach to calorimetry is unable to meet the mass and energy resolution requirements for future linear colliders. The particle flow approach to calorimetry with PandoraPFA has a proof-of-principle demonstration of its capability to reach the required jet energy resolution. The particle flow approach to calorimetry also puts 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 tracking detector, which has a much better resolution than the calorimeter. Small cell sizes of the calorimeters are required to identify calorimeter hits from different particles. The traditional sum of calorimeter cell energies is hence replaced by particle flow reconstruction algorithms - a complex pattern recognition problem.

Developed with the ILD detector concept, PandoraPFA has been adapted to the CLIC condition and shows its ability to deliver required energy resolutions [24]. There are over 60 electron-positron linear collider specific reconstruction algorithms. Each aims to address a particular topological issue in the reconstruction. In the recent development, the core base codes for basic object and memory managements are factorised in the Pandora C++ Software Development Kit [46].

In the subsequent sections, the main steps in the PandoraPFA reconstructions are summarised below. The details of the PandoraPFA event reconstruction can be found in [27, 28, 46]. The inputs of PandoraPFA are digitised calorimeter hits and reconstructed tracks, with some detector geometry information to aid the reconstruction. The output are reconstructed particles with four-momenta, also known as the Particle Flow Objects (PFOs).

4.4.1. Track selection

Tracks from the inner tracking detectors are important inputs of the PandoraPFA reconstruction. These tracks 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 passing 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 tracks of a “V0” shape. This is identified

by searching for a pair of tracks originated from a single point. Another topology is the "kinks" when a charged particle decays to a single charged particles with neutral particles. The last special topology are the "prongs" when a charged particle decays to multiple charged particles. This information about special topologies is stored and passed onto the subsequent reconstruction, along side with helical track fit (using last 50 reconstructed hits in the tracking detector) and the track projection to the front of the ECAL.

4.4.2. Calorimeter selection

The other important inputs of the PandoraPFA reconstruction are the calorimeter hits from calorimeters. The properties of a calorimeter hit include the 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 hits that pass the selection are used in later steps.

Extra information about calorimeter hits are calculated, stored and used in later steps. The extra information includes the geometry information of the hit and likelihood of the hit originated from a minimum ionising particle (MIP).

Isolated hits, often originating 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. However, these isolated hits participate in the reconstruction during the last step, particle flow object (PFO) creation, to contribute to the energy estimation.

4.4.3. Particle Identification

To improve the reconstruction of the charged particles, where calorimeter hits are associated to tracks, dedicated particle identification algorithms identify calorimeter hits associated with muons and photons. These calorimeter hits are removed from the subsequent reconstruction. As fewer hits are left to be reconstructed, the reconstruction of charged particles is improved and the pattern recognition problem is simplified.

Identified muons and photons do not participate in the later clustering and re-clustering stages, but re-enter the reconstruction at the fragment removal stage (see section 4.4.8). See chapter 5 for details on the photon reconstruction related algorithms..

4.4.4. Clustering

A cone based clustering algorithm is used to group calorimeter hits into clusters. The output clusters are further processed, merged, or split based on their topological properties. Since the cone clustering algorithm is widely used in many other reconstruction algorithms in the PandoraPFA, it is necessary to introduce the cone based clustering algorithm, before discussing the rest of the PandoraPFA reconstruction.

There are two main types of clustering algorithms: cone based algorithms and sequential combination algorithms (see section 4.6.2). The main type of clustering algorithms used in the PandoraPFA is the cone based algorithms. Illustrated in figure 4.1, cone based clustering algorithm identifies a seed first, shown as the yellow dot. The algorithm then forms a cone to include hits that are within a specified opening angle to the seed. Afterwards the cone with the associated hits form the cluster.

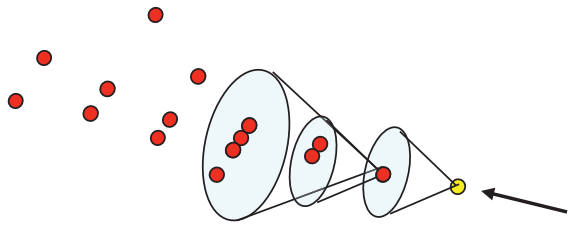


Figure 4.1.: Illustration of the cone based clustering algorithm, taken from [47]

The cone based clustering algorithm is preferred in PandoraPFA because the direction of the particle flow is largely unchanged from the originated particle, irrespective of the particle flow being an electromagnetic shower, QCD radiation, or hadronisation. Figure 4.2 shows the clustering algorithm used in the PandoraPFA. The seed for the cone clustering is typically the projection of a track to the front of the ECAL. A calorimeter hit can also be used as a seed. The initial cluster direction is taken as the direction of the seed. Afterwards, a cone with a specified opening angle and depth will be formed around the direction of the seed.

The building of the cone is iterated from the inner layer of the ECAL to the outer layer. At each layer, possible associations with calorimeters hits in previous layers and the same layer are made. If a calorimeter hit is not associated with the cone, the hit is used to seed a new cluster. The clustering algorithm produces basic working objects, clusters.

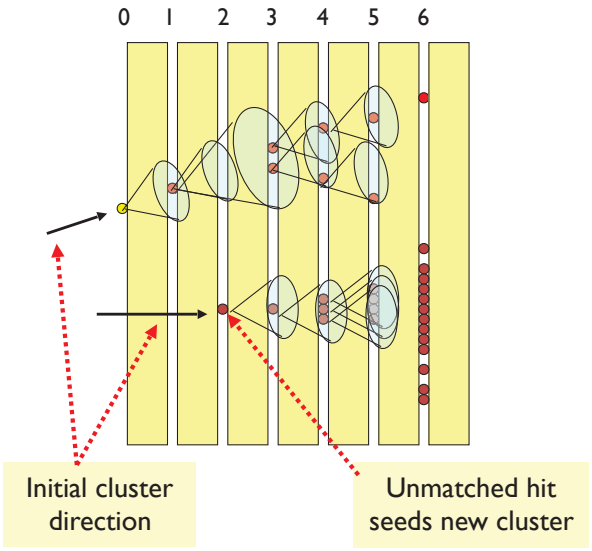


Figure 4.2.: Illustration of the clustering algorithm used in the PandoraPFA, taken from [47]

4.4.5. Topological cluster association

After the initial clustering, clusters are further refined using topological information of calorimeter hits in the calorimeters. This step is necessary because the initial clustering scheme tends to form small clusters. These small clusters are then merged based on clear topological signatures in this step. The merging signatures include combining track segments, connecting a track segment with gaps, connecting track segments to hadronic showers, and merging clusters when they are within close proximity. For example, association algorithms for looping track segments, back-scattered tracks from hadronic showers, and cone association are shown schematically in figure 4.3. In each case, the arrow indicates the tracks. The dark red dots represent the calorimeter hits in the associated cluster. The slightly fainter red dots represent the calorimeter hits in the neutral cluster. The black line represents the front of the ECAL.

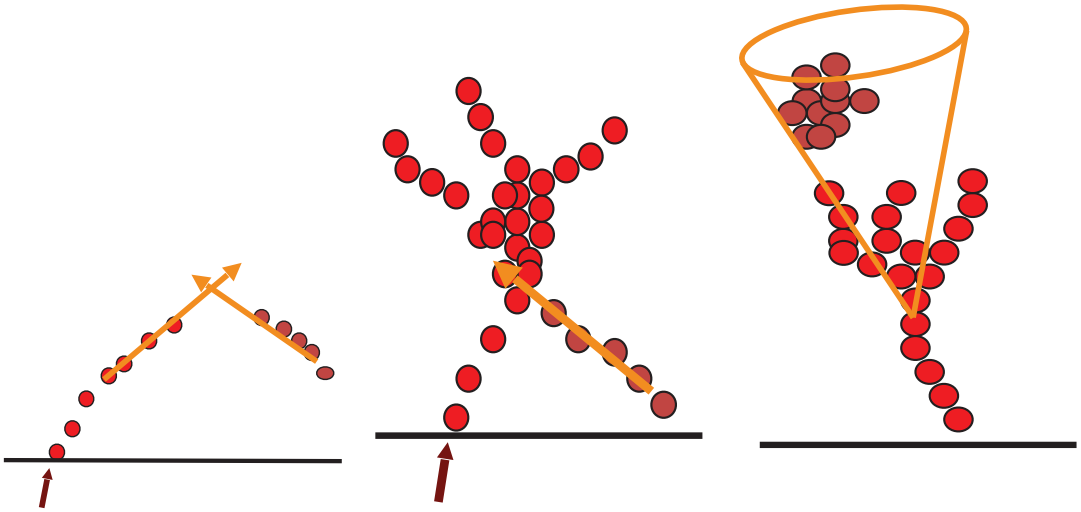


Figure 4.3.: Examples of topological association in the PandoraPFA. Rules for a) looping track segments, b) back-scattered tracks from hadronic showers, c) and cone association are shown. In all plots, the arrow indicates the tracks. The dark red dots represent the calorimeter hits in the associated cluster. The slightly fainter red dots represent the calorimeter hits in the neutral cluster. The black line represents the front of the ECAL. Figures are taken from [47].

4.4.6. Track-cluster association

Having refined the clusters in the calorimeter, the next step is to associate the clusters to the tracks obtained from the inner tracking detector. The associations are made according to the proximity of the first layer of the cluster and the track projection to the front of the ECAL. The consistency between the track direction and the initial cluster direction, as well as a match between the track momentum and the cluster energy, are required.

4.4.7. Re-clustering

The cluster association scheme described in the previous section works well for events with low-energy (less than 50 GeV) jets. For an environment with high-energy jets, electromagnetic and hadronic showers are boosted and are likely to overlap each other. Therefore, it is important to refine the track-cluster association based on the momentum and the energy information.

The re-clustering stage improves the compatibility of the cluster energy and the associated track momentum. It is performed on a statistical basis. If the cluster energy and the associated track momentum do not match, the cluster will be re-clustered either using the same clustering algorithm with different parameters, or different clustering algorithms. This re-clustering step creates many temporary clusters. Afterwards, out of many temporary clusters, the temporary cluster has the best track-momentum cluster-energy match is chosen, and the temporary cluster is associated with the track.

A schematic diagram of the re-clustering stage is shown in figure 4.4. In the figure, the black upright arrows indicate the tracks. The dark red dots represent the calorimeter hits in the associated cluster. The slightly fainter red dots represent the calorimeter hits in the neutral cluster. In this example, the initial 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 scheme tries different cone clustering algorithms by splitting the neutral cluster so that the topological association could make a correct association.

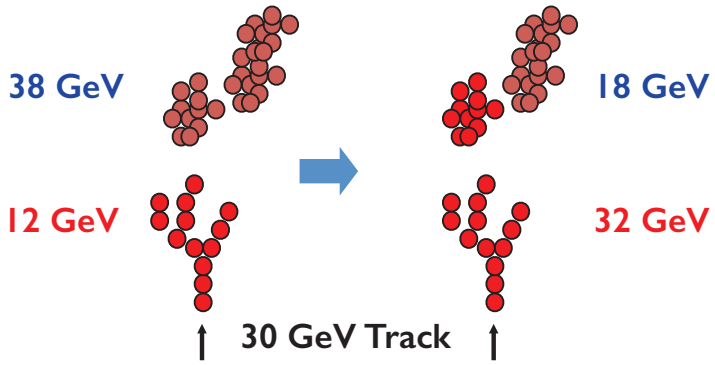


Figure 4.4.: Illustration of the re-clustering algorithm in PandoraPFA, taken from [47]. The arrow indicates the tracks. The dark red dots represent the calorimeter hits in the associated cluster. The slightly fainter red dots represent the calorimeter hits in the neutral cluster. The initial cluster 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 algorithms to split the neutral cluster so that the topological association could make a correct association.

4.4.8. Fragment removal

This stage of the PandoraPFA reconstruction will focus on merging low-energy clusters. These clusters are likely to be fragments of other particles. The merging criterion are

mostly based on the proximity of the fragment to the particle, and the energy comparison of the fragment to the particle. Algorithms dealing with photon fragment merging and photon splitting are described in details in chapter 5.

4.4.9. Particle Flow Object Creation

The last stage of the reconstruction is the creation of the output objects, Particle Flow Objects (PFOs). The PFOs contain clusters and associated tracks. Simple, but effective, particle identification for electrons and muons are applied.

The output objects, PFOs, contain information on positions, four-momenta and associated quantities. These PFOs are heavily used in physics analyses. The electron, muon and photon identifications associated with the PFOs are also used in physics analyses, for example, the analyses in chapter 6 and in chapter 7.

4.5. The CLIC specific simulation and reconstruction issue

There are a few simulation and reconstruction issues specific to the CLIC, which affect the analysis in chapter 7. One issue is that the luminosity spectrum for interactions with photon from Beamstrahlung is different to the luminosity spectrum of electron-positron interactions. A solution is presented in section 4.5.1 to correct for the differences. Another issue is that there is a large amount of beam induced background in the CLIC environment, which needs to be suppressed before physics analyses. The background suppression is described in section 4.5.2. Lastly simulated masses of particles are given in section 4.5.3.

4.5.1. Luminosity spectrum

The electron-photon interaction, where the photon is produced from initial state radiation via Beamstrahlung , has a different instantaneous luminosity than the electron-positron interaction. Hence, for the same time-frame, the total integrated luminosity of the electron-photon interaction is different to that of the electron-positron interaction. To correct for the difference in the luminosities, a simulated study [48] was performed with

the GUINEAPIG [49] and was simulated in the WHIZARD, to identify the ratio of the integrated luminosity of the electron-photon interaction to the electron-positron interaction. The results are summarised in table 4.1. For the physics analysis in chapter 7, event number for processes with initial-state photons from Beamstrahlung are corrected with the ratios 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
$L(e^\pm\gamma)$ / $L(e^+e^-)$	0.75	0.79
$L(\gamma e^\pm)$ / $L(e^+e^-)$	0.75	0.79
$L(\gamma\gamma)$ / $L(e^+e^-)$	0.64	0.69

Table 4.1.: Luminosity ratio for processes with initial-state photons from Beamstrahlung at the CLIC, at $\sqrt{s} = 1.4\,\text{TeV}$ and $3\,\text{TeV}$. The table summarises results from [48].

4.5.2. Beam induced backgrounds

The other issue considered when using the CLIC_ILD detector concept is the beam induced background. At a high centre-of-mass energy, the background becomes important for the event reconstruction. Therefore, the beam induced background are considered in the simulation.

There are different types of the beam induced background. The $\gamma\gamma \to$ hadrons is the dominant background in all calorimeters except inner part of the HCAL endcap. Another type of the background, the incoherent pairs, are ignored. Therefore $\gamma\gamma \to$ hadrons, intergraded over 60 bunch crossing, is overlayed onto the reconstruction.

The hadronisation of $\gamma\gamma \to$ hadrons background events are performed with the PYTHIA. The $\gamma\gamma \to$ hadrons background events are superimposed on the physics process simulations to save computational resources. The choice of 60 bunch crossings is an estimate of the amount of background in the experiment condition [50].

The beam induced background deposits significant amounts of energies in the detector. It needs to be suppressed for physics analyses. Two software have been developed to suppress these background: a track selector and a PFO selector [28].

The track selector aims to remove poor quality and fake tracks that are more likely from the beam induced background. It places a simple track-quality cut and a time-of-

arrival cut on tracks. If the arrival time of the track at the front of the ECAL, using the helical fit of the track, differs more than 50 ns from using a straight line fit, the track will be rejected.

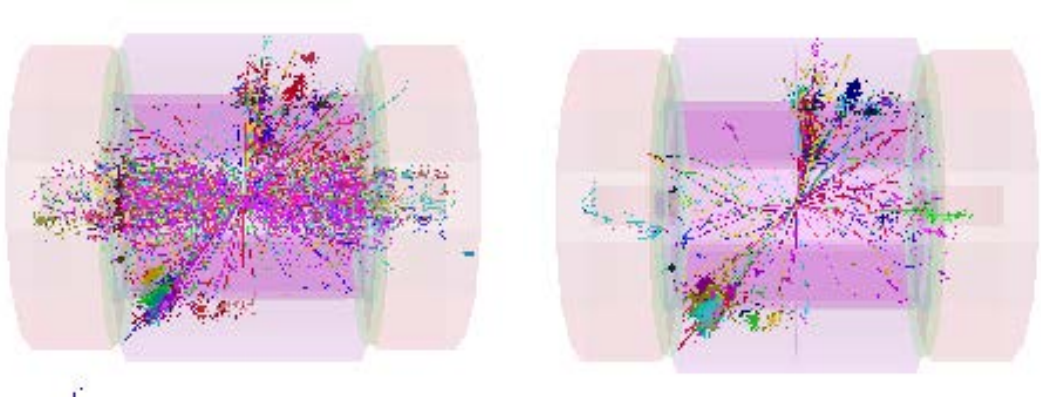
The PFO selector discards PFOs that are originated from the beam induced background from the event reconstruction, based on the transverse momentum (p_T) and time information of the PFOs. The PFOs from $\gamma\gamma \to$ hadrons often have low p_T and have a range of time-of-arrivals. In contrast, the PFOs from physics processes have a range of p_T, and the time-of-arrivals are close to the brunch crossing time. By utilising the high spatial resolution from the high granular calorimeter, individual PFOs can be tracked and reconstructed.

The PFO selector uses different p_T and time cuts for the central part of the detector and for the forward part of the detector. For the best performance, there are different cuts for different types of particles: photons, neutral PFOs, and charged PFOs. Three configurations of these cuts are developed: "loose", "normal", and "tight" selections. As the name suggested, "loose" selection corresponds to a looser cut of p_T and time-of-arrival, allows a larger value of p_T and time-of-arrival.

The optimal configuration depends on the centre-of-mass energy 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. Reconstructed particles in a simulated $e^+e^- \to HH \to t\bar{b}b\bar{t}$ event are integrated over a time window of 10 ns (100 ns in HCAL barrel) in the CLIC_ILD detector model, with 60 bunch crossings of $\gamma\gamma \to$ hadrons background overlaid in figure 4.5a. The effect of applying tight PFO section cuts is shown in figure 4.5b. The energy deposited in the detector by the background is reduced from 1.2 TeV to the level of 100 GeV.

4.5.3. CLIC simulated particle masses

Another information used in the analysis with the CLIC detector concept is the simulated mass and width of quarks and bosons. These values are listed in table 4.2 and are used for generating Monte Carlo samples.



(a) **(b)**

Figure 4.5.: Reconstructed particles in a simulated $e^+e^- \to HH \to t\bar{b}b\bar{t}$ are integrated over a time window of 10 ns (100 ns in HCAL barrel) event in the CLIC_ILD detector model, with 60 bunch crossings of $\gamma\gamma \to$ hadrons background overlaid in figure 4.5a. The effect of applying tight PFO section cuts is shown in figure 4.5b. The energy deposited in the detector by the background is reduced from 1.2 TeV to the level of 100 GeV. Figures are taken from [28].

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.: The masses and widths of quarks and bosons used for generating samples. The H mass is specified for individual samples. The table is taken from [24].

4.6. Analysis software

In the previous sections the automated reconstruction tools are described in details. This section is dedicated to the automated analysis software, which will be used in the analyses described in subsequent chapters.

4.6.1. Monte Carlo truth linker

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

4.6.2. Jet algorithms

For the linear collider, thanks to the high granular calorimeter and advanced PFA software, the starting point for analysis are individual Particle Flow Objects (PFOs), as well as individual tracks. Each of the PFOs encodes four-momentum and position information. At the same time, tracks would have momentum and position information. However, sometimes it is useful to group PFOs and tracks into jets, which are the results of hadronisation processes from high energy particles like quarks or gulons.

A jet is typically a visually obvious structure in an event display. The momentum and the direction of a jet tend to resemble the original particle. Despite the relative simplicity of identifying jets visually, it is a challenge for a pattern recognition program to identify jets effectively and efficiently. Early work on jet finding started in 1977 [51], where later development can be found in reviews [52–54]. This section is based on these reviews.

There are two large families of jet finding algorithm: cone based algorithms, and sequential combination algorithms. The cone based algorithms are briefly discussed in section 4.4.4 in the context of the PandoraPFA reconstruction. Here the focus is on the sequential combination algorithms.

Sequential combination algorithms typically calculate a pair-wise distance metric between a seed and a particle. The particle with the smallest metric is combined into the jet with the seed. The distance metric will be updated after a combination. This procedure is repeated until some stopping criterion are satisfied. The different jet algorithms typically differ in the definitions of distance metrics and stopping criterion.

The chosen jet algorithm implementation for this thesis is the FastJet C++ software package [55, 56]. The implementation in the Marlin software package is called the MarlinFastJet. The package provides a range of jet finding algorithms.The notations in the subsequent discussion follow the convention in [55].

4.6.3. Longitudinally-invariant k_t algorithm

Longitudinally-invariant k_t algorithm [57,58] is one of the common sequential combination algorithms used in the pp collider experiments. There are two variants of the algorithm: inclusive and exclusive. In the inclusive variant, the symmetrical pair-wise distance metric between particle i and j, d_{ij} or d_{ji}, and the beam distance, d_{iB}, are defined as

$$d_{ij} = d_{ji} = \min\left(p_{\mathrm{T}i}^2, p_{\mathrm{T}j}^2\right)\frac{\Delta R_{ij}^2}{R^2}, \tag{4.1}$$

$$d_{iB} = p_{\mathrm{T}i}^2, \tag{4.2}$$

where $p_{\mathrm{T}i}$ 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 $y_i = \frac{1}{2}\ln\frac{E_i + p_{z_i}}{E_i - p_{z_i}}$ 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. four-momentum of particle i is updated as the sum of the two particles. Otherwise if $d_{ij} \geqslant d_{iB}$, particle i is set to be a final jet. The above procedure is repeated until no particles are left.

The exclusive variant is similar to the inclusive variant. First difference is that when $d_{iB} < d_{ij}$, particle i forms part of the beam jet. The beam jet contains particles that are considered to be from the beam induced background. The beam jet is not used as a output. The second difference is that when both d_{ij} and d_{iB} are above some threshold, d_{cut}, the clustering will stop. In another word, the exclusive mode allows a specified number of jets to be found, where the d_{cut} is automatically determined. The inclusive mode, on the other hand, would find as many jets as the algorithm allows.

4.6.4. Durham algorithm

The Durham algorithm [59], also known as e^+e^- k_t algorithm, is commonly used for the e^+e^- collider experiments. It only has one distance metric:

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

where E_i is the energy of particle i; and θ_{ij} is the polar angle difference between particle i and j. The 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}.

Compared to the k_t algorithm, it uses energy instead of p_{T} in the distance metric, and it does not use the beam distance. This is because that for the e^+e^- collider at low centre-of-mass energies, the beam induced background is not significant.

4.6.5. 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 the Durham algorithm, are not suitable for the CLIC collision environment. Studies have shown that jet algorithms for the pp colliders give better performances for the CLIC [24, 60]. Therefore, longitudinal invariant k_t algorithm is often used in analyses with the CLIC environment.

4.7. Multivariate Analysis

Multivariate analysis (MVA) has become increasingly important in high energy physics. MVA is typically used in the physics analysis to classify signal events from background events. The MVA can be viewed as an advanced tool for regression or classification. Compared to the traditional cut-based method, modern machine learning techniques offer much improvement to the data analysis. The implementation of the machine learning techniques used in this thesis are provided by TMVA [61].

A typical machine learning MVA can be used for classification or regression. Classification classifies events into one of several classes. Regression gives an output in a

continuous range. The focus in this section is on the classification, as the MVA is often used to select one type events from another type.

A typical machine learning MVA classification involves two classes, also known as signal and background. A machine learning model 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. The response of the model is a classification with two-class outcome of signal or background.

This classification scheme can easily be extended to multiple classes, implemented in TMVA with the multiclass class. For example, The multiclass class is used in the tau decay mode classification in section 6.6 and in the flavour tagging classifier in section 7.5.

There should be three statistically independent samples for the MVA: one sample for the training; another sample for the validation, including optimisation and checking for overfitting; and the last sample for testing. However, due to technical reasons (TMVA only natively supports two samples), sometimes the same sample is used for the validation and the testing, which is an acceptable usage with samples of large data.

4.7.1. Optimisation and overfitting

One important concept with the MVA is the optimisation and the 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. Another way to describe the difference in a simple and a complex model is that a simple model typically has a low variance but a high bias, whilst a complex model would often have a low bias but a high variance.

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

Figure 4.6 shows an example of the model efficiency as function of the model complexity. One definition for overfitting is when the efficiency of the signal selection in the training

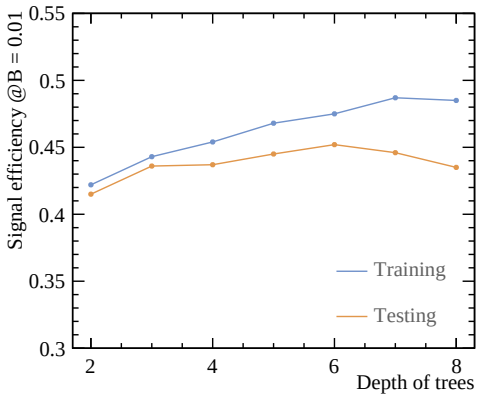


Figure 4.6.: Example of model efficiency as function of the model complexity. Here the model is a boosted decision tree. The model parameter reflecting the model complexity is the depth of tree. The y-axis is the signal efficiency when the background efficiency is 1%. From the tree depth of six onwards, overfitting occurs.

samples increases, but the efficiency in the testing sample decreases, with the increase of the model complexity. The example in figure 4.6 is chosen from the double Higgs analysis at $\sqrt{s} = 3\,\text{TeV}$, using the Boosted Decision Tree model. The efficiency of the signal selection is defined as the signal fraction when the background fraction is 1%, reported 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 a tree depth of two to six, the efficiency for both testing and training samples increases. From tree depth of six onwards, overfitting occurs. In this particular example, one should choose a tree depth fewer than seven to avoid overfitting.

4.7.2. Choice of models

The model can be as simple as a cut-based model, a likelihood estimator, or a linear regression model. The model can also be as complicated as a non-linear tree, a non-linear neutral network, or a support vector machine. Regardless of the model complexity, the choice of the most optimal classifier is often data driven to match the nature of the sample. For example, a non-linear model is the best to model a non-linear response. The comparison between different models without individual optimisation 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. A neutral network model could potentially be better than the boosted decision tree model, but it requires more

tuning, and it is less intuitive to interpret such a model. For these reasons, the boost decision tree model (BDT) is often the choice of machine learning model in high energy physics. And it is used in various physics analysis in this thesis. Before describing the BDT in detail, we will first visit some simpler models.

4.7.3. Rectangular Cut model

The rectangular cut method, probably the most intuitive model, optimise cuts to maximise some pre-defined metrics. 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 passing the rectangular cuts, respectively.

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

Because of its simplicity, the cut method is often performed manually, much more often at times pre-dating the spread of machine learning methods. It is still commonly used in the analyses in the pre-selection step before the MVA.

4.7.4. Projective Likelihood model

The projective likelihood model with probability density estimators (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.

The likelihood classifier calculates the probability density for each discriminative variable, for signal and background (hence PDE approach). The overall signal and background likelihood are defined as products of the individual probability density of each variable. 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, the likelihood model works better with decorrelated, gaussian like variables.

4.7.5. Decision Tree model

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

The decision tree is a binary tree, where each node, the splitting point, uses a single discriminative variable to decide whether an 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 after certain criterion are met. The stopping criterion 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 refers to the determination of the optimal cut at the node by minimising a metric. Assuming the probability of the cut producing the signal is p, three commonly used metrics for two-class classification are:

1. Misclassification error: $1 - \max(p, 1-p)$,
2. Gini index: $2p(1-p)$,
3. Cross-Entropy or deviance: $-p\log p - (1-p)\log(1-p)$.

The applying of a trained decision tree is performed by transversion the tree from the root node to the end node. 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 illustrates a simple example of a decision tree. The signal class is the PhD student and the background class is the undergraduate student. The depth of this imbalance binary tree is 2. A splitting node is represented by a diamond. The signal-like end node is represented by the red rectangle and the background-like end nodes are represented by blue rectangles. The tree is constructed with two possible cuts, "Party ends before 1am" and "Know where free pizza is". The attribute of samples is listed in table 4.3 and table 4.4. The Gini index metric to determine the optimal cut. If the first cut is "Party ends before 1am", the probability of the cut producing the signal, p, is $\frac{10}{13}$, as there are 10 PhD students and 3 undergraduate students who end part before 1 am. Gini index give $2p(1-p) \backsimeq 0.36$. If the first cut is "Know where free pizza is", $p = \frac{10}{15}$, as there are 10 PhD students and 5 undergraduate students who know where the free pizza

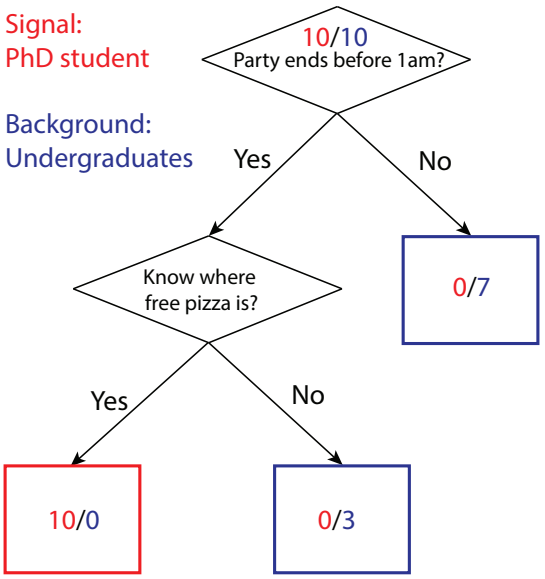


Figure 4.7.: Example of a decision tree. Numbers in each node represent number of PhD student (red) and number of undergraduate student (blue) after each cut. Diamond boxes represent splitting nodes. Rectangular boxes represent end nodes. Blue boxes are background-like end nodes. Red boxes are signal-like end-nodes.

PhD student	Party ends before 1 am	Party ends after 1 am
Know where free pizza is	10	0
Not know where free pizza is	0	0

Table 4.3.: The attribute of the PhD student class for the decision tree example shown in figure 4.7.

is located. Gini index is $2p(1-p) \backsimeq 0.44$. Therefore, by choosing the cut that minimise the Gini Index, 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. An example of applying the trained decision tree is provided. if there is s student who ends the party before 1 am and knows where a free pizza is located, then the student is classified as a PhD student.

4.7.6. To improve decision tree

The decision tree model has a low bias, but a high variance. This means it is very easy to construct a tree that fits the training data very well, but the tree would not be

Undergraduates	Party ends before 1 am	Party ends after 1 am
Know where free pizza is	0	5
Not know where free pizza is	3	2

Table 4.4.: The attribute of the undergraduates class for the decision tree example shown in figure 4.7.

optimal for the testing sample. To overcome the instability of the decision tree, many methods have been developed. Some of the most successful ones are boosting, bagging, and random forest.

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 weights and be more likely to be classified correctly. The boosting is done at every iteration, which can be a few hundred or a few thousand times. 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-sets of the training sample, and apply the model. In this case, every boosting iteration takes a bagged sample, rather than the whole sample.

Random Forest: when a tree is grown, a randomly selected sub-set of discriminative variables are used to grow the tree. This method is know to reduce the variance of the tree.

4.7.7. Boosted Decision Tree model

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 regarded as the best out-of-the-box machine learning method. There are two common boosting methods: adaptive boosting and gradient boosting. The adaptive boosting, first introduced in [62], is discussed in further details, as it is simpler to understand than the gradient boosting.

The basic idea of adaptive boosting is that the tree making procedure focuses on events which are difficult to classify correctly. By assigning a weight to each event, after

each tree growing iteration, the weights for misclassified events are gradually increased. Therefore misclassified events get more attention in the next iteration.

The adaptive boosting algorithm, adapted from [63], is outlined below:

- At the initialisation stage, event weight is initialised to $w = 1/N$ for every event, for N total events.
- Iterate 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 a testing event x, is a weighted vote from all M trees:

$$G(x) = \begin{cases} -1, & \text{if} \sum_{m=1}^{M} \alpha_m G_m(x) < 0, \\ 1, & \text{otherwise.} \end{cases} \tag{4.4}$$

The tree classifier output, G is denoted as -1 or 1. One can think of -1 as background and 1 as signal. There are N events and M iterations (trees). B represents if a event is misclassified. For the i^{th} event in the m^{th} tree, $B_{i,m} = 1$ if the event is misclassified and 0 if the event is correctly classified. $w_{i,m}$ represent the event weight for i^{th} event in m^{th} tree.

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 adaptive boosting is to dramatically improve the performance of a weak classifier. A weak classifier is a classifier which is gives a predictive performance sightly better than a 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 an event is classified as signal in the forest. The likelihood number is later used to select signal from background.

4.7.8. Optimisation of Boosted Decision Tree

Many parameters of the BDT can be optimised. The most important parameter is the depth of a tree, which determines how many end nodes the tree has, or the degrees of freedom of the tree. The related parameter is the number of trees. Experience shows that using many small trees yields the best result.

The number of trees 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 [63]. Therefore there is a debate on the metric to determine the optimal number of trees.

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 variants in TMVA implementation: adaptive boost and gradient boost.

The learning rate of the adaptive boost controls how fast the weight changes for events in each boosting iteration. Experience shows that a small learning rate ($\sim$0.1) with many trees works better than a large learning rate with fewer trees.

The shrinkage rate in the gradient boost is similar to the learning rate in the adaptive boost. The shrinkage rate controls how fast the weight changes for events in each boosting iteration. Again a small value ($\sim$0.1) is preferable.

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

The number of bins per variable is a necessary parameter to make tree growing efficient, because discretely binned variables are faster to compute than continuous variables. This parameter, however, does not impact the performance much. But because variables are binned, variables should be pre-processed before going into the model. For example, the variable should be limited to a sensible range to avoid the extreme values. The variable should also be transformed to obtain a more uniform distribution, if the original distribution is highly skewed.

For the end node, it can be determined as either signal-like or background-like, based on the majority of the training events 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 bagging fraction determines the fraction of randomly selected samples used in each boosting iteration. By choosing a small value, samples between each boosting iteration are less correlated. Hence the overall performance improves.

The DoPreSelection flag allows the classifier to throw away phase spaces where there are only background events.

4.7.9. Multiple classes

The above discussion assumes two classes - signal and background. The argument can be extended to multiple classes. There are two ways for the training multiple classes. “One versus one” scheme trains each class against each other class, and the overall likelihood is normalised. The second way to train these multiple classes is called “one versus all”, when each class is trained against all other classes.

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

TMVA multiclass implementation uses the “one versus all" scheme. For each class, the multiclass classifier will train the class as the signal against all other classes 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 an 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 classifier instead of a two-class classifier for multiple classes is that the correlation between different classes are accounted for. The classifier outputs are correctly adjusted for multiple classes. Hence one event can only be classified into one class. The issue with the multiclass is that powerful discriminative variables for each individual class need to enter the training stage simultaneously, resulting in a large number of variables in the multiclass classifier.

Chapter 5.

Photon Reconstruction in PandoraPFA

'When I walk along with two others, from at least one I will be able to learn.'

— Confucius, 551 BC - 479 BC

Many aspects of the photon reconstruction are important. The single photon energy resolution is necessary to reconstruct heavy particles using decay channels involving photons, such as tau lepton decay or π^0 decay. Another important aspect of the photon reconstruction is the photon separation resolution, which is the measure of minimum spatially closeness of two resolved photons. The photon separation resolution, together with the photon completeness and purity, is crucial for a photon counting experiment, where the number of the photon is used as a physics quantity. The most recent example of such a photon-counting example, benefited from this photon reconstruction, is the $H \to \gamma\gamma$ simulation study at $\sqrt{s} = 3\,\text{TeV}$ at the CLIC [64].

Having an efficient photon reconstruction in a dense jet environment also improves the overall event reconstruction. As the particle flow approach to the calorimetry aims to reconstruct each individual particle, by assigning correct calorimeter hits to photons, assigning the remaining hits to tracks for charged particles becomes an easier problem. Hence the correct track-cluster association can be achieved with fewer mistakes, and the jet energy resolution improves.

This chapter discusses a number of photon reconstruction algorithms within the PandoraPFA framework, followed by the performances of these algorithms. Part of this chapter has been published in a conference proceeding [65].

5.1. Overview of photon reconstruction in PandoraPFA

PandoraPFA is a multi-algorithm pattern recognition software package for the event reconstruction and an implementation of the particle flow approach to calorimetry. A detailed discussion of the PandoraPFA and the main steps of the PandoraPFA event reconstruction can be found in section 4.4. The multi-algorithm approach of the PandoraPFA allows each algorithm deals with a particular issue in the reconstruction. In the context of the photon reconstruction, there are five algorithms developed to tackle different issues in the photon reconstruction.

The most important photon algorithm is the Photon Reconstruction algorithm. It carefully reconstructs photons from calorimeter hits in the ECAL, including forming a photon candidate and applying a photon ID test, with special treatments for photons close to charged particles. This algorithm is implemented at the early stage of the reconstruction (see section 4.4.3), aiming to provide a clean environment for the subsequent charged-particle reconstruction.

Three algorithms are designed remove photon fragments at a later stage in the reconstruction. Two photon fragment removal algorithms merge fragments in the ECAL, and one algorithm merges fragments in the HCAL. These algorithms improve the single photon energy resolution.

The last photon algorithm is a photon splitting algorithm. The algorithm seeks to separate accidently merged photons, which helps photon separation resolution.

These algorithms together form the photon reconstruction in the PandoraPFA. This chapter will first introduce the photon-induced electromagnetic shower in a calorimeter, followed by the description of each algorithm. The performance of the photon reconstruction will be provided in the last part of the chapter.

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 properties of the EM shower is used as the basis of forming photon candidates, photon ID test, and photon separation.

The pair production and bremsstrahlung generate many low energy photons and electrons, producing a shower-like structure, hence the name EM shower. Two suitable length scales to describe the EM shower are the radiation length and the Molière radius. The 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 [66]. The Molière radius of a material is a suitable length describe the transverse shower profile.

Figure 5.1 shows the simulated longitudinal electromagnetic shower profiles as a function of radiation length for electrons and photons. The mean EM longitudinal shower profile can be described by the following function [67] :

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

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

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

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.

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 with a radius of one Molière radius. Transverse profile is often represented by a sum of two Gaussian function. The Gaussian nature of the transverse profile allows the separation of two EM showers using a two dimensional peak finding algorithm.

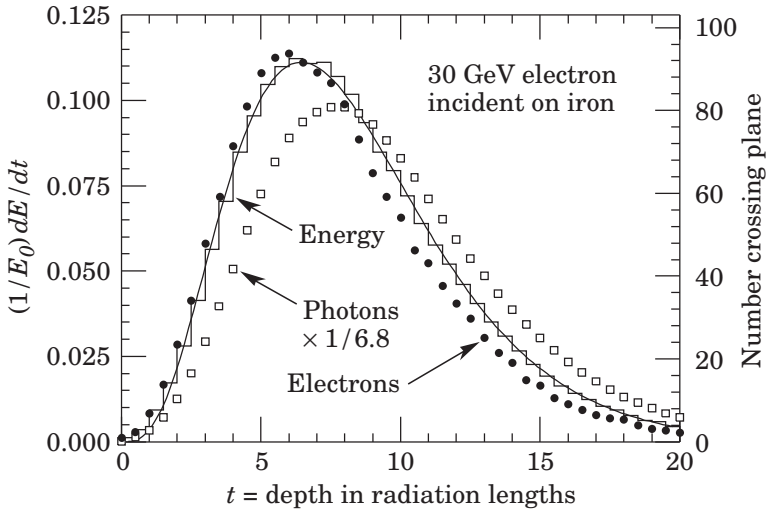


Figure 5.1.: An EGS4 simulation of a 30 GeV electron-induced electromagnetic shower in iron. The histogram shows fractional energy deposition as a function of radiation lengths, and the curve is a gamma-function fit to the distribution. Circles and squares are the number of electrons and photons respectively with total energy greater than 1.5 MeV crossing planes with scale on right. Plot is taken from [3].

5.3. Photon Reconstruction algorithm

The Photon Reconstruction algorithm is a photon reconstruction and identification algorithm at the early stage of the reconstruction. It correspondes to "Particle ID" stage in section 4.4.3 of the PandoraPFA reconstruction chain. Main steps of the Photon Reconstruction algorithm, shown in figure 5.2, are: coarsely forming photon clusters; finding photon candidates; photon ID test; and optional fragment removals. Finding photon candidates uses the transverse EM shower profile information, which requires a two dimensional peak finding algorithm, further explained 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 code base case class with another photon fragment removal algorithm. Hence two photon fragment removal algorithms are discussed together in section 5.6.

5.3.1. Form photon clusters

The inputs of the Photon Reconstruction algorithm are calorimeter hits in the ECAL, that have not been used in previous algorithms. For example, muon reconstruction

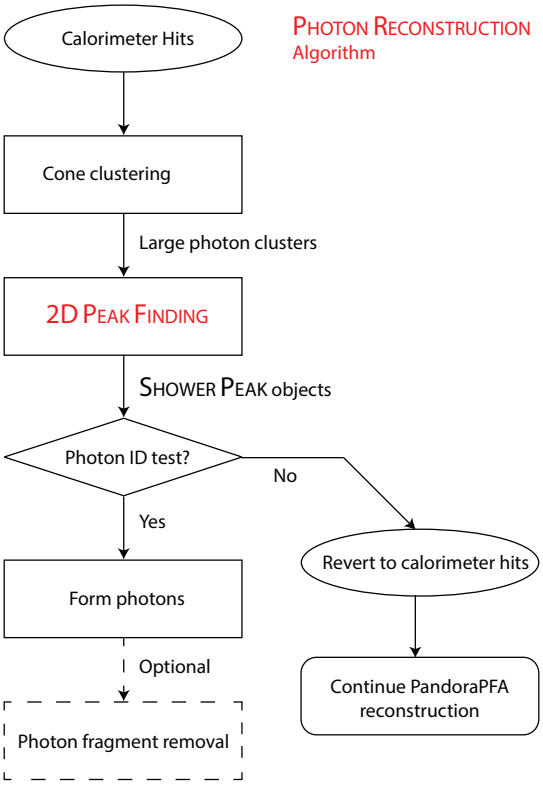


Figure 5.2.: A flow diagram showing main steps in the PHOTON RECONSTRUCTION algorithm.

algorithms will form muons and remove calorimeter hits associated with muons from the reconstruction. The muon associated calorimeter hits are not used in this step.

The PHOTON RECONSTRUCTION algorithm finds large potential photon clusters in the ECAL. The clusters are formed in a way such that a photon would not be split into two clusters, but one cluster may contain multiple photons. For simplicity, the algorithm opts to reuse the cone clustering algorithm (see section 4.4.4) 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.

5.3.2. Find photon candidates

The next stage is to refine large photon clusters into smaller photon candidates. The aim of this step is that each photon candidate should contain calorimeter hits from one photon only.

The method for refining clusters 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 EM showers, which is characterised by a narrow cone, widening while the shower develops. Along the direction of the photon, an EM shower can be modelled as a dense shower core with peripheral hits around the core. When the energy deposition is projected on to a plane, the EM shower core would appear as a mountain-like structure in the plane. One example a large photon cluster projected on to a two dimensional plane, where two EM showers are identified, is shown in Figure 5.3. Hence, by identifying a peak in the two dimensional plane, an EM shower core is identified.

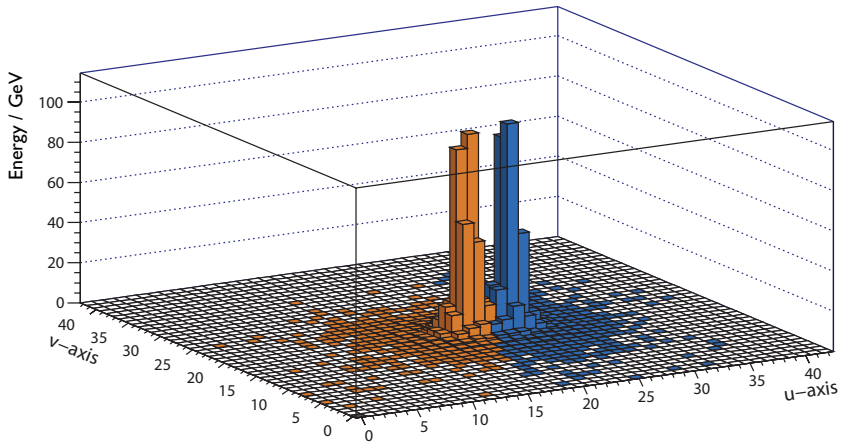


Figure 5.3.: Two 500 GeV photons (yellow and blue) belonged to a large photon cluster, just resolved in a transverse plane orthogonal to the direction of the flight by projecting their energy deposition in the ECAL. U and V are orthogonal axes in units of the ECAL cell sizes. Z axis is the sum of the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

By using the two-dimensional energy deposition projection, the three-dimensional cluster splitting problem is mapped to a two-dimensional peak finding problem with the projection translation. Therefore a high-performance two dimensional peak finding algorithm is the key to identify photon candidates within a photon cluster. Due the complexity of the projection and the peak finding procedure, a separate peak finding algorithm is developed and discussed in section 5.4. The output of the two dimensional peak finding is a collection of ShowerPeak objects. Each ShowerPeak object, which is used as the input for the next step, corresponds to a photon candidate and associated calorimeter hits.

5.3.3. Photon ID test

This step applies the photon ID test on the photon candidate, which is stored in the form of a SHOWER PEAK object. The photon ID test uses a multidimensional likelihood classifier, which needs to be trained before applying. A set of variables, which exploit features of electromagnetic showers, are used. The response from the classifier determines if a photon cluster is a photon. If it is a photon, the cluster would be tagged as a photon and the cluster does not participate in the event reconstruction until at the fragment removal stage after the charged particle reconstruction. If the cluster is not a photon, the cluster will be reverted to associated calorimeter hits. The hits will be passed on to the next stage of the reconstruction (see figure 5.2). The classifier is discussed in section 5.5, because the multidimensional likelihood classifier is complicated.

5.3.4. Photon Fragment removal

The photon fragment removal algorithm aims to merge small photon fragment to identified photons. The algorithm is optional as it is not run by the default setting. Since this step shares the same logic as another fragment removal algorithm, two algorithms are discussed togethers in a later 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. These photons do not participate in the event reconstruction until at the fragment removal stage (see section 4.4.8) after the charged particle reconstruction.

5.4. Two dimensional peak finding algorithm for photon candidate

As discussed in section 5.3.2, identifying photon candidates inside a cluster is translated to identifying peaks in a two-dimensional plane, with the two-dimensional peak finding algorithm (2D PEAK FINDING algorithm). An example of two photons resolved 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 calorimeter hits.

There are two variants of the 2D Peak Finding algorithm: the neutral cluster variant and the charged cluster variant. The base algorithm, the neutral cluster variant, treats all clusters as potential photon clusters. A flow chart of the main steps of the neutral cluster variant is shown in figure 5.4. 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. The neutral cluster variant is described first, followed by the modification for charged cluster variant.

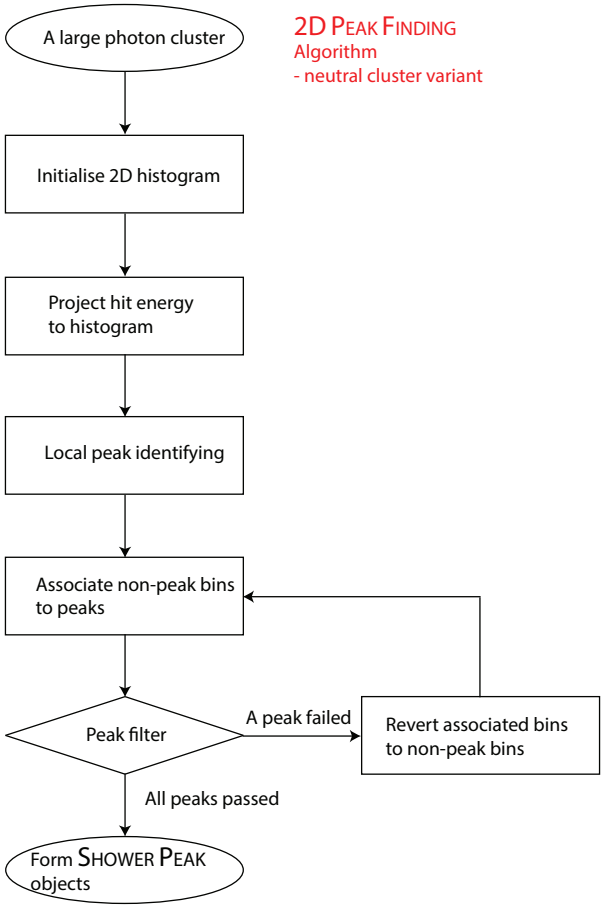


Figure 5.4.: Flow chart showing main steps in the neutral cluster variant of 2D Peak Finding algorithm.

5.4.1. Initialise the two-dimensional histogram

This step initialise a two-dimensional (2D) histogram to host the projection of the energy deposition of the photon cluster. For the best resolving power between photons, the projection direction is chosen to be the direction of the cluster. Two axes of the

two-dimensional histogram are chosen such that axes and the direction of the cluster form an orthogonal bases in the three dimensional space.

5.4.2. Project calorimeter hits to histogram

This step projects positions of the calorimeter hits, which are associated with the photon cluster, onto the 2D histogram initialised in the 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 along both axes corresponds to one ECAL square cell length. The relative distance between the calorimeter hit position and the cluster centroid position is converted into a distance vector. The distance vector is subsequently projected onto the histogram using the scalar product with the axes vectors. The distance vector, $\vec{s_i}$, of a hit i is obtained by:

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

where $\vec{a}$ is the three dimensional position of the hit i; $\langle\vec{a}\rangle$ is the centroid position of cluster a; and d_{cell} is the ECAL square cell length.

The projected position on the 2D histogram is binned at integer intervals. The bin height is the sum of the energies associated with the calorimeter hits that fall in that particular bin.

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 scanned linearly to identify all local peaks.

5.4.4. Associate non-peak bins to peaks

Having tagged all local peaks, this step associates non-peak bins to peaks based on the energy of the peak and the distance of the non-peak bin to the peak bin. The energy dependence is needed as the transverse EM shower width increases with the increase

of the energy. The distance dependence is needed because the EM showers have dense shower cores.

After all local peak bins are found, non-peak bins are associated to a peak bin. The peak bin is chosen by minimising the metric

$$\min_i \frac{d_i}{\sqrt{E_i}} \tag{5.4}$$

where d_i is the Euclidean distance between a non-peak bin and a peak bin i on the histogram, and E_i is the height of the peak bin i. The metric is iterated over all peak bins for each non-peak bin. Alternative metrics provided in the algorithm include d_i, $\frac{d_i}{E_i}$, and $\frac{d_i}{E_i^2}$. 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 the energy deposition. To select major peaks only, 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.

The Shower Peak object is created after peak filtering, which contains one peak bin and associated non-peak bins. The associated calorimeter hits with the bins are attached to the Shower Peak object as well. If multiple peaks are identified in a cluster, multiple Shower Peak objects are created as outputs.

This marks the end of the neutral clusters variant of the Photon Reconstruction algorithm, outlined in figure 5.4. The Shower Peak object is also referred to as the photon candidate.

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 EM showers, such as electromagnetic shower typically starting in first few layers of the ECAL with direction of the EM shower largely unchanged.

Figure 5.5 shows the main steps in the full 2D Peak Finding algorithm, including the treatment to clusters close to tracks. The "Close to track" step determines if a cluster is close to a track. If a cluster is less than 3 mm from the closest track projection, it is treated as a potential charged cluster.

The "Create and iterate sub-clusters" step performs the following. The ECAL is sliced longitudinally to help to identify photon candidates. For example, the default three slices will result in three ECAL fiducial spaces, which cover spaces from the front of the ECAL to a third, two thirds and the back of the ECAL, respectively. Three sub-clusters contained in each fiducial space are created. The neutral cluster variant of the 2D Peak Finding algorithm is then repeated for each sub-cluster. The Shower Peak objects created from each sub-cluster undergo the "Shower Peak filter" step. All peaks and associated Shower Peak objects from the first sub-cluster are preserved. Peaks and associated Shower Peak objects from subsequent sub-clusters are only preserved, if the peak bin position can be linked to a peak bin from the previous sub-cluster, whilst allowing a shift in position by no more than one neighbouring bin. Otherwise, the peak and the associated Shower Peak object are discarded in all sub-clusters. Furthermore, if a peak bin is within the eight neighbouring bins of the track projection, the peak is discarded in all sub-clusters. Only the peaks are preserved through the iteration of "Shower Peak filter" will form the final Shower Peak objects.

5.4.7. Inclusive mode

The two-dimensional histogram is iterated many times during the algorithm. The time complexity is $O(n^2)$ for a n bins by n bins histogram (Default $n = 41$). Therefore, for the purpose of speed, it is undesirable to have a large number of bins. Having a small

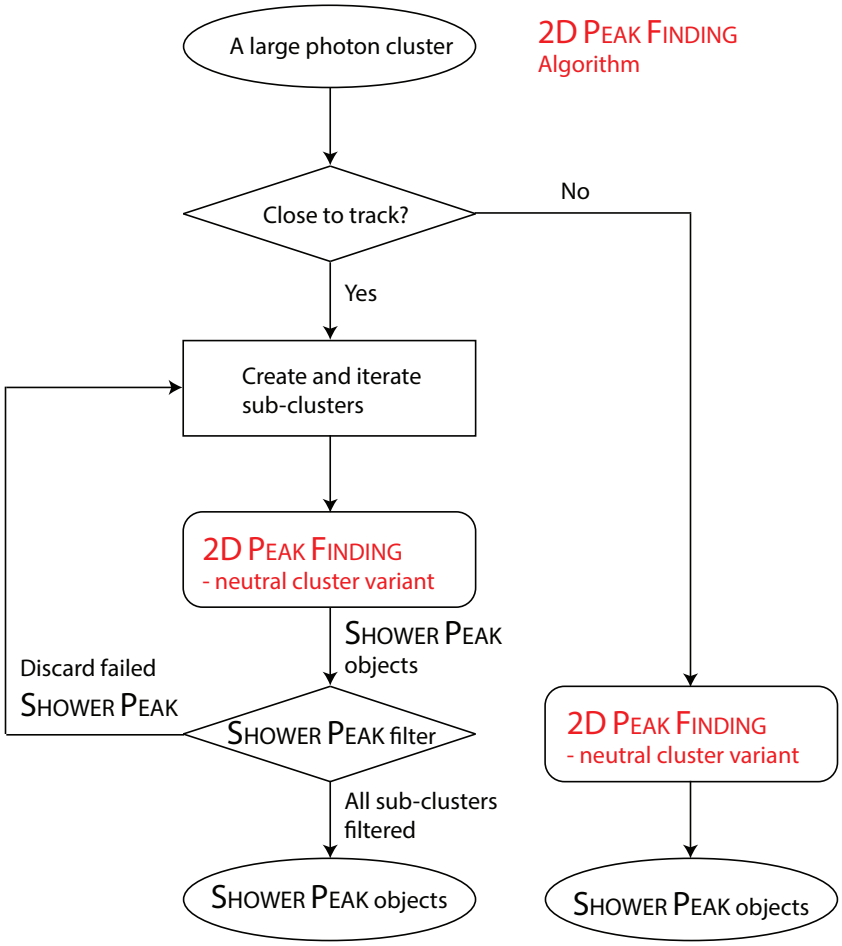


Figure 5.5.: Flow chart showing main steps in the 2D Peak Finding algorithm, including the charged cluster variant.

finite histogram speeds up the calculation. However, because of the finite size, only energy deposition projected onto 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 if the algorithm is only interested in finding the EM shower cores, for example, the Photon Reconstruction algorithm (section 5.3) and the photon fragment removal algorithms(section 5.6). However, for photon splitting (section 5.8), there should be no calorimeter hits loss from splitting a photon. Hence the inclusive mode of the 2D 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

In section 5.3.3, the photon ID test in the photon reconstruction algorithm was outlined. This section describes the multidimensional likelihood classifier used in the photon ID test in details, including variables used in the classier.

5.5.1. Variable used in the likelihood classifier

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.

Two variables exploit the longitudinal EM shower distribution: t_0 is the start layer from the longitudinal shower profile (see section 5.2), shown in figure 5.6a; and δl is fractional difference of the observed shower profile to the expected EM shower profile [27]:

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

δl is minimised as a function of the t_0. The δl distribution for photons and non-photons is shown in figure 5.6b. For a photon, t_0 and δl are expected to be small.

Three variables use the transverse shower information. $\langle w \rangle$ is the energy weighted root-mean-squared distance of all bins in a Shower Peak to its peak bin, shown in figure 5.6c. This is a measure of the transverse shower size. $\delta\langle w_{UV} \rangle$, is the smallest ratio of the two root-mean-squared distances of all bins in a Shower Peak to its peak bin

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 energy of the Shower Peak object to the cluster energy. This is a measure of the dominance of a photon in a cluster.

The last variable used in the classifier, d, is the distance between the photon candidate and the closest track projection in the front of the ECAL. The Shower Peak object is less likely to be a photon if it is close to a track. Its distribution for photons and non-photons is shown in figure 5.6d.

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.

5.5.2. Projective Likelihood classifier

Projective likelihood classifier with probability density estimators is used in PandoraPFA for the photon ID due to its simplicity and low requirement on computing resources. Details on projective likelihood classifier can be found in section 4.7.4.

The probability distribution of each variable for photons and non-photons are obtained in the training stage. 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 the variables distributions are stored separately for different photon energies. There are 8 energy bins by default by cutting the distribution of photon energies at 0.2, 0.5, 1, 1.5, 2.5, 5, 10, 20 GeV, which covers a good range of photon energies. The classifier training typically uses simulated 250 GeV jet events, with $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$. 250 GeV jets allow the training of photon candidates with energies greater than 20 GeV.

After training, for a given photon candidate with the energy in the energy bin α, the likelihood classifier output is given by

$$pid = \frac{N_p^\alpha \prod_i^6 P_{i,p}^\alpha}{N_p^\alpha \prod_i^6 P_{i,p}^\alpha + N_{np}^\alpha \prod_i^6 P_{i,np}^\alpha} \tag{5.6}$$

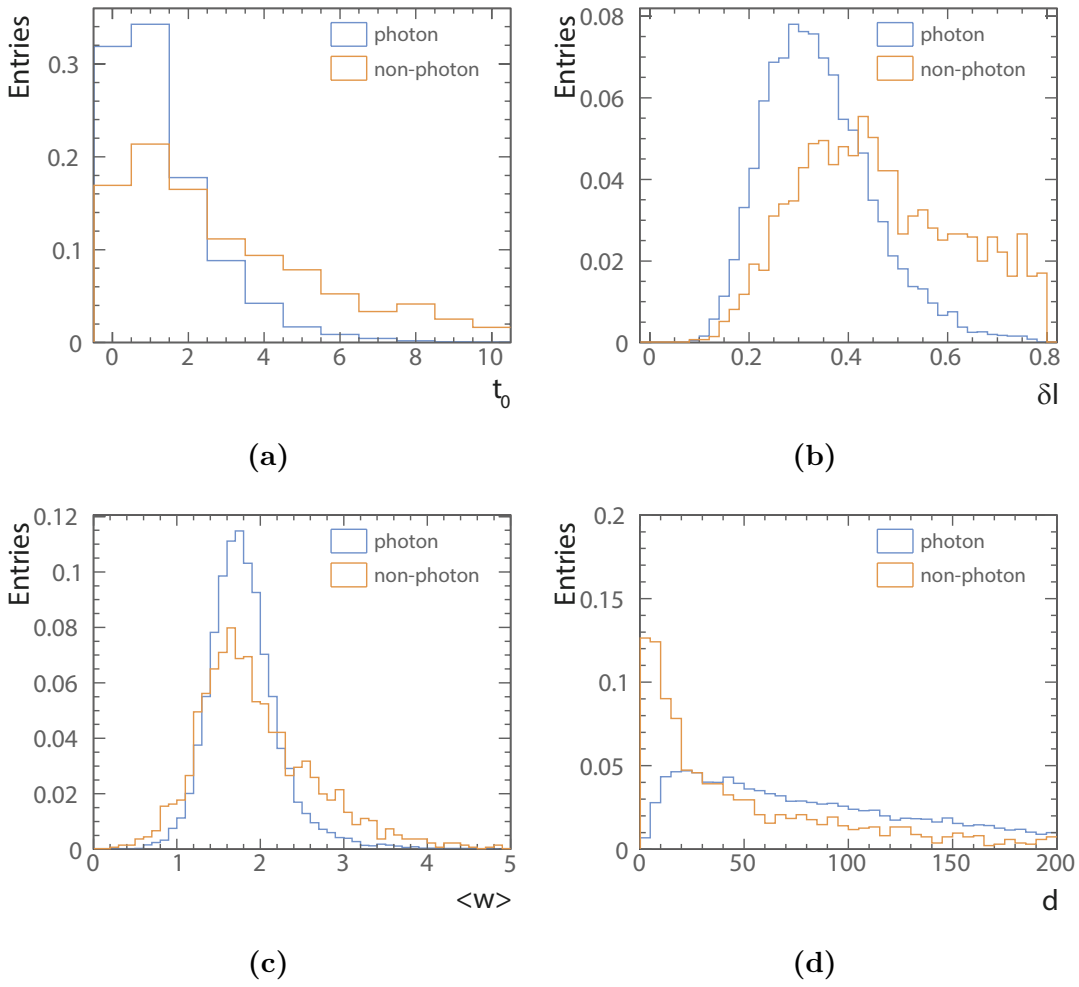


Figure 5.6.: Distributions for a) the start layer from the longitudinal shower profile (t_0), b) the fractional difference of the observed shower profile to the expected EM shower profile (δl), c) the energy weighted root-mean-squared distance of all bins in a SHOWER PEAK to its peak bin ($\langle w \rangle$), and d) the distance between the photon candidate and the closest track projection in the front of the ECAL (d) are shown. All plots are normalised, shown for photons and non-photons, where the particle ID is determined using the truth information. All plots are generated with simulated 250 GeV Z′ events, where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$.

where $P^{\alpha}_{i,p}$ and $P^{\alpha}_{i,np}$ are the probability of the i^{th} variable of the photon candidate fallen in the i^{th} variable probability distribution of the photon and non-photons in the energy bin α, respectively; N^{α}_{p} and N^{α}_{np} are the number of photons and non-photons in the energy bin α, respectively.

During classification, a photon candidate passes the photon ID test if

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

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

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), but the outer part of the shower is reconstructed as a separate particle, and wrongly identified as a photon or a neural hadron. Figure 5.7 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.

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

Figure 5.7.: An event display of a typical 10 GeV photon shown in a), reconstructed into a main photon shown in b), and a photon fragment shown in c).

Photon fragment removal algorithms can exist at difference places in the reconstruction: immediately after the PHOTON RECONSTRUCTION algorithm, or after the charged particle reconstruction. Since these algorithms share the same logics, they will be discussed together. The algorithm used after the charged particle reconstruction will be discussed here. The algorithm immediately after the PHOTON RECONSTRUCTION differs mostly in the default cut-off values in the 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 examined. The pair is merged when its properties pass a set of cuts, where cuts are 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 with the same photon pass the merging test, the pair with closest distance metric, d, will be merged.

Depending on whether the fragment is reconstructed as a photon or a neutral hadron, 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_c gives the distance between centroids of each PFO in the pair, which is a computationally 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 5.8):

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

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; and $E_{f,i}$ is the energy of the fragment in 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.

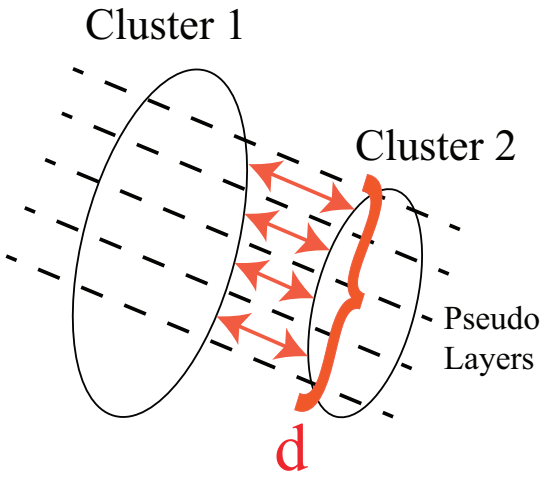


Figure 5.8.: Illustration of distance metric, d.

Other quantities used in the merging metric include E_m the main photon energy; E_f, the fragment energy; E_{p1} and E_{p2}, the energies of the two largest peaks and associated calorimeter hits respectively, found by the 2D Peak Finding algorithm (section 5.4), ordered by descending energy, using the pair as input; N_{calo}, the number of the ECAL hits in the fragment; and $|\cos(\theta_Z)|$, the absolute cosine of the polar angle of the main photon with respect to the beam direction.

One logic of merging is when the fragment has 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 energetic, demanding a small ratio of E_f to E_m.

The other logic of merging is when the pair looks like one photon in the two-dimensional energy deposition projection. The transverse shower comparison requires $\frac{E_{p1}}{E_m+E_f} > 0.9$, demanding most energy of the cluster contains in the first Shower Peak object.

Cuts for low energy fragments and high energy fragments are similar. Cuts for high energy fragments allow higher energy fragments to be merged.

Comparing photon-photon-fragment pair and photon-neutral-hadron-pair, the differences are due to the fact 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 pairs with $d < 80$ mm.

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	$N_{calo} < 40$, $d_c < 60$, $E_f < 0.6$, $\lvert\cos(\theta_Z)\rvert > 0.7$	-
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, respectively. E_m and E_f are the main photon energy and the fragment energy, respectively. E_{p1} and E_{p2} are the energies the two largest peaks, found by peak finding algorithm, ordered by descending energy, respectively. N_{calo} is the number of the calorimeter hits in the fragment. $\lvert\cos(\theta_Z)\rvert$ is the absolute cosine of the polar angle, where beam direction is the z-axis.

5.7. Photon fragment recovery algorithm in the HCAL

The previous section describes an effective algorithm to removal photon fragments that are peripheral to the main photon, the electromagnetic shower core. There is another

type of fragments originated from the leakage effect of the ECAL. 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 5.9. When a high energy EM shower is not fully contained in the ECAL, the 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. This photon fragment recovery algorithm is important when reconstructing high energy photons. For the ILD detector, this ECAL leakage effect appears when the photon energy is above 50 GeV.

Shown in figure 5.9, 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.

Figure 5.9.: 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).

This algorithm would collect photons in the ECAL and neutral hadrons in the HCAL as inputs. The algorithm then iterates 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. Photon-fragment pairs passing the cuts will be merged.

Fragments 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; and 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 root-mean-square 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 their shower widths should be similar. w_m^l and w_f^l are the root-mean-squared widths 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 cell size 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 respectively. 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 calorimeters.

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 event 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. This algorithm focuses on energetic photons with energy greater than 10 GeV.

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 root-mean-squared width of the last outer layer of the main photon and the first inner layer of the fragment. r_f is the root-mean-squared 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.

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.

Table 5.4 lists the cuts used in the algorithm. If an energetic photon is identified, and two energetic EM showers 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, which is defined as within 100 mm of the track projection on the front of the ECAL, extra care is taken by demanding a large value for second EM shower energy. E_{c1} and E_{c2}, the energy cut-off values, are 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.

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.

5.9. Characterise the performance

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.

Three different versions of the PandoraPFA are used to characterise the performance.

- No stand-alone photon reconstruction algorithms
- With a stand-alone photon reconstruction algorithm from PandoraPFA version 1.
- With full photon related algorithms described above, incorporated in PandoraPFA version 3.

First the performance with the full algorithms is compared with no photon algorithms, then compared with PandoraPFA version 1. Afterwards, the individual photon related algorithms are then characterised, followed by performance of the photon reconstruction in PandoraPFA version 3.

5.10. Compare with no photon reconstruction

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 photon events were generated with an uniform distribution in the solid angle of the photon. The two photon events were generated with an uniform distribution in the solid angle of the first photon, and an 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 in the tracking detector and the 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 5.10 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.

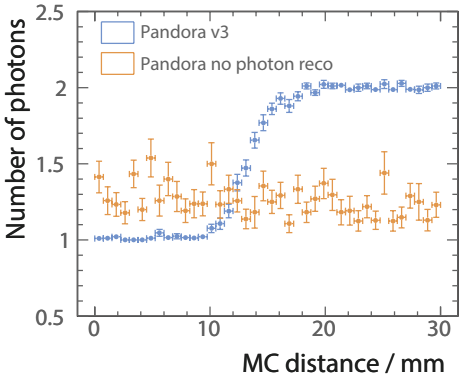


Figure 5.10.: Average number of photons using two photons of 500 and 50 GeV per event reconstructed without and with photon algorithms, as a function of the Monte Carlo distance separation between the photon pair.

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 $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$ 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 from semi-leptonic decay of heavy quarks. Using 90% of the entries is robust and focus on the Gaussian part of the distribution. The total jet energy is sampled at 91, 200, 360 and 500 GeV. Shown in figure 5.11, 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 mistakes. However, at low energy, the reconstruction is worse with photon algorithms, because photon algorithms are optimised for high energies.

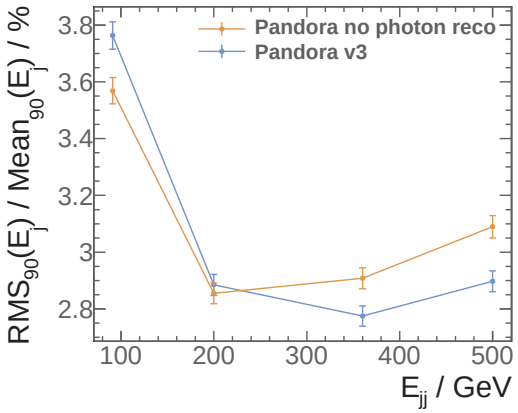


Figure 5.11.: Jet energy resolution as a function of the total jet energy using $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$ sample at barrel region. The top orange and bottom blue dots represent the reconstruction without and with photon related algorithms.

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	$\sqrt{s} = 91\,\text{GeV}$	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 total jet energy in the $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$ samples for reconstruction with and without photon algorithms.

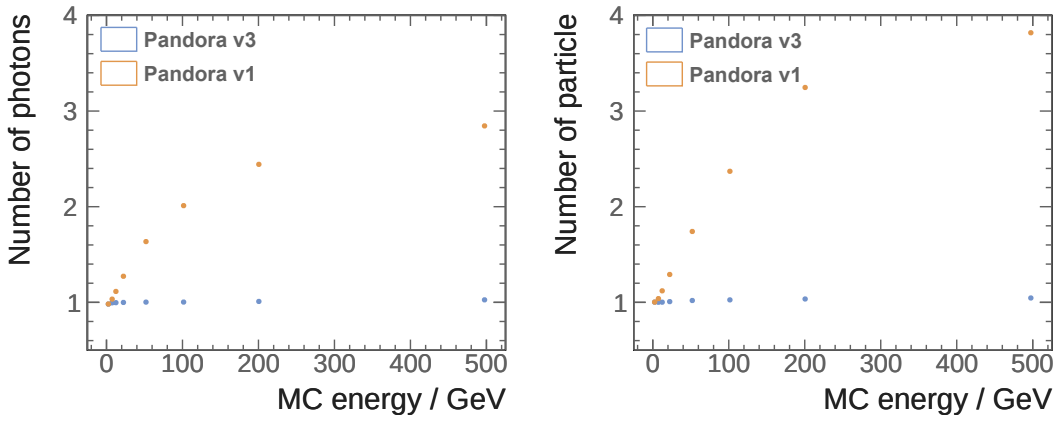
5.11. Compare with photon reconstruction in PandoraPFA version 1

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

Testing samples are generated in the same way as in the previous section. Figure 5.12a shows the reduction in fragments identified as photons, using a single photon per event sample. Indicating as the blue dots on the plots, 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 reduction in the number of reconstructed particles is shown in figure 5.12b, 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 from 2.4. For a 500 GeV photon, the number is reduced to 1.05 from 3.8.

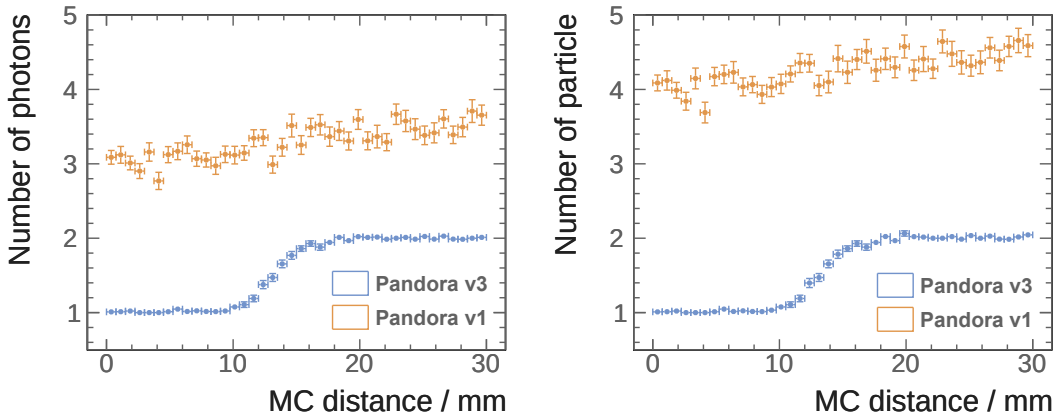
Figure 5.13 illustrates a reduction in the photon fragments and the neutral hadron fragments using two photons of 500 and 50 GeV per event sample. The figure is shown for the Monte Carlo distance separation of the photon pair from 0 to 30 mm, which corresponds to approximately 6 ECAL square cells of the default ILD detector model. 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 5.13a and figure 5.13b, the average numbers of



(a) (b)

Figure 5.12.: Average number of reconstructed a) photons, and b) reconstructed particles, as a function of their true energy using a single photon per event sample. For both figures, 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.

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.



(a) (b)

Figure 5.13.: Average number of reconstructed a) photons, and b) particles, as a function of the MC distance separation in the calorimeter, using two photons of 500 and 50 GeV per event sample. For both figures, the top orange and bottom blue dots present the reconstruction with PandoraPFA version 1 and version 3, respectively. The photon reconstruction is changed in PandoraPFA version 2.

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 5.14, 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.

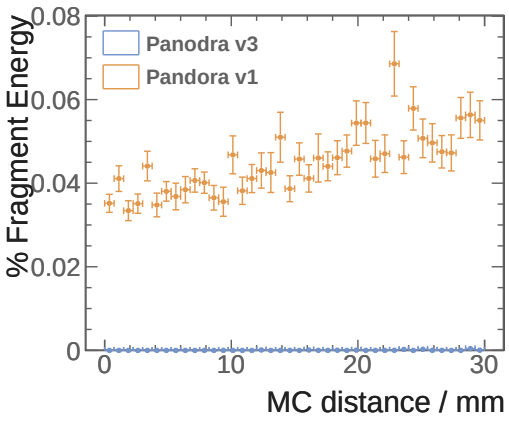


Figure 5.14.: Average fraction of fragments energies to the total energy, as a function of the Monte Carlo distance separation in the calorimeter, using two photons of 500 and 50 GeV per event sample. The top orange and bottom blue dots represent the reconstruction with PandoraPFA version 1 and version 3 respectively. The photon reconstruction is changed in PandoraPFA version 2.

The improvement in the completeness ofn 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 total jet energy is sampled at 91, 200, 360 and 500 GeV. Shown in figure 5.15, the jet energy resolutions are better at 360 and 500 GeV with improved photon reconstruction. This is due to more an aggressive photon reconstruction, 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.

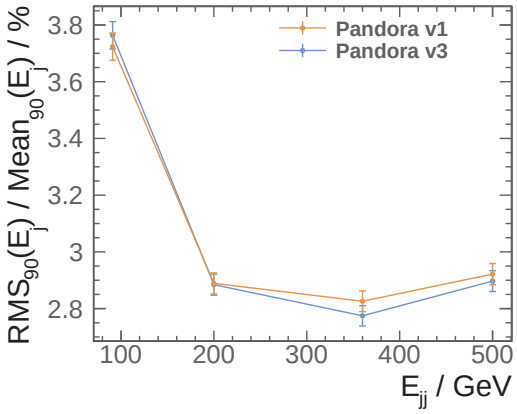


Figure 5.15.: Jet energy resolution as a function of the total jet energy using $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$ sample at barrel region. The top orange and bottom blue dots represent the reconstruction with PandoraPFA version 1 and version 3. The photon reconstruction is changed in PandoraPFA version 2.

5.12. Understand photon reconstruction improvement

To show the incremental improvement of individual algorithm, the average number of particle for a high energy photon pair, 500 - 500 GeV is shown in figure 5.16. Blue, orange, and green dots represent full photon reconstruction, only fragment removal algorithms in the ECAL, and fragment removal algorithms in the ECAL and the HCAL, respectively.

With fragment removal algorithm in the ECAL, the number of fragment is reduced significantly. With the additional 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.

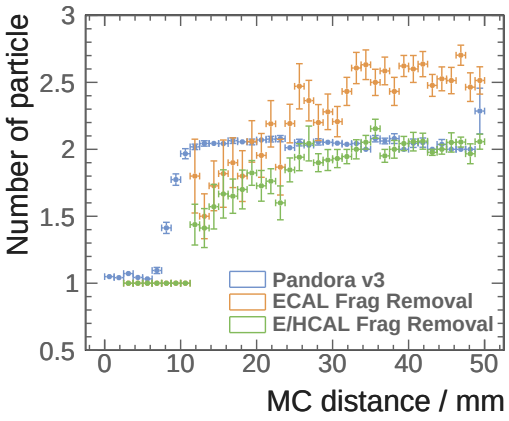


Figure 5.16.: Figure shows the average number of photons, as a function of the Monte Carlo distance separation between the photon pair in the calorimeter, using two photons of 500 and 500 GeV per event sample. The blue, orange, and green dots represent the reconstruction with PandoraPFA version 3, PandoraPFA 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.

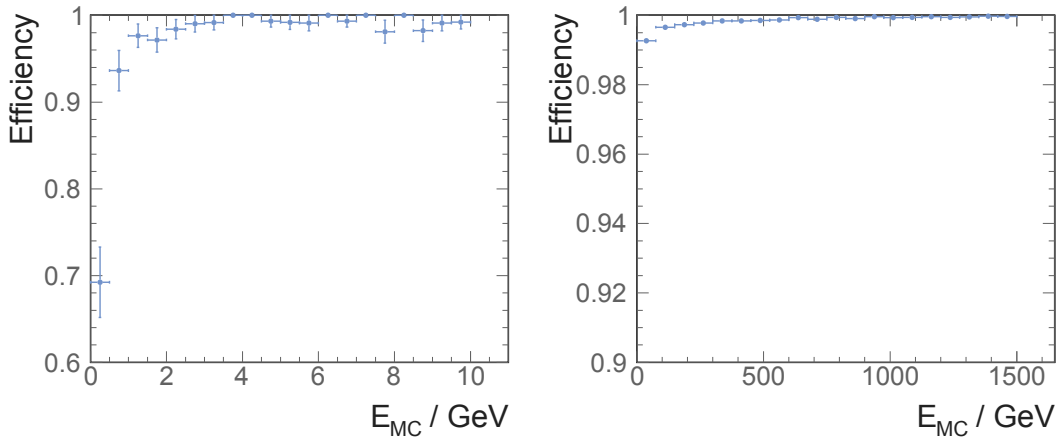
5.13. Current photon reconstruction performance

In section 5.11, 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 reconstruction and ID efficiency is demonstrated in figure 5.17. Using single photon samples, an event can have an efficiency of 1 or 0, depending on whether the photon reconstructed corresponds to the truth photon. 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 5.18 for 500 and 50 GeV photons pair sample, the average number of photons and particles beyond 20 mm apart is less than 2.05, 0.05 above the true value, 2.

The resolving power of a photon pair depends on energies of two photons. Figure 5.19 shows the average number of photon reconstructed for different photon pairs. When the energies of two photons are similar, the resolving power is greater. This is because that



(a) **(b)**

Figure 5.17.: Single photon reconstruction efficiency as a function of true energies. a) shows the low energy region and b) shows high energy region.

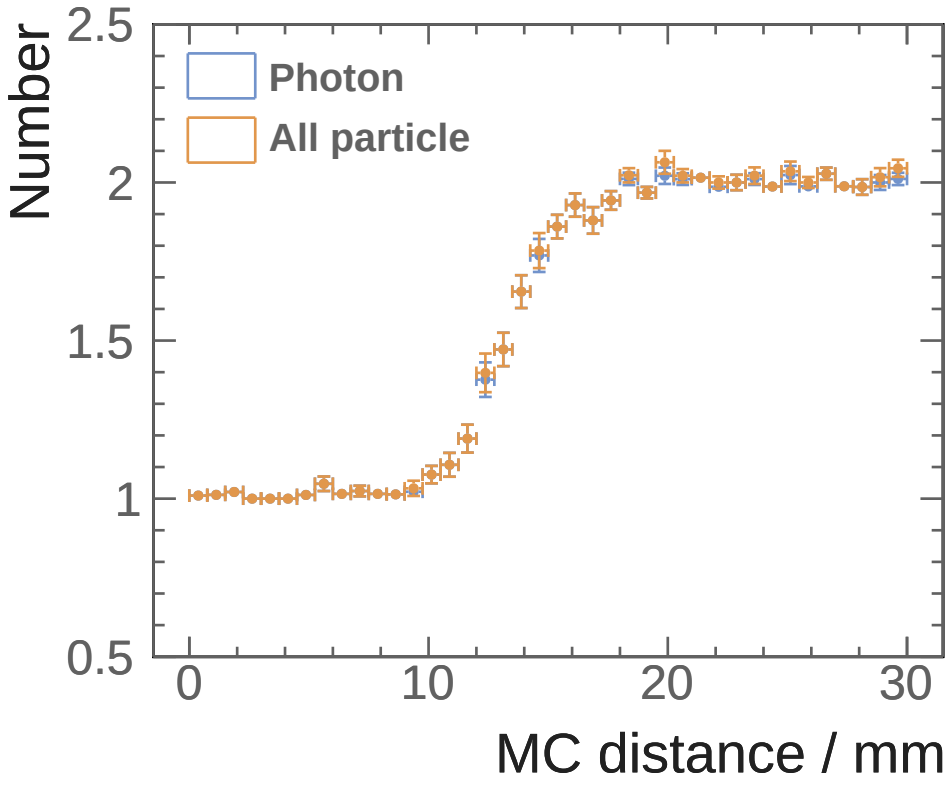


Figure 5.18.: Average numbers of photon (blue) and particle (orange), as a function of the Monte Carlo distance separation between the photon pair, using two photons of 500 and 50 GeV per event sample.

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 one ECAL cell length. Photon pairs with different energies, for example 500 - 50 GeV and 100 - 10 GeV pair, start to be resolved at 10 mm apart, which is about two ECAL cells length.

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 5.19. 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.

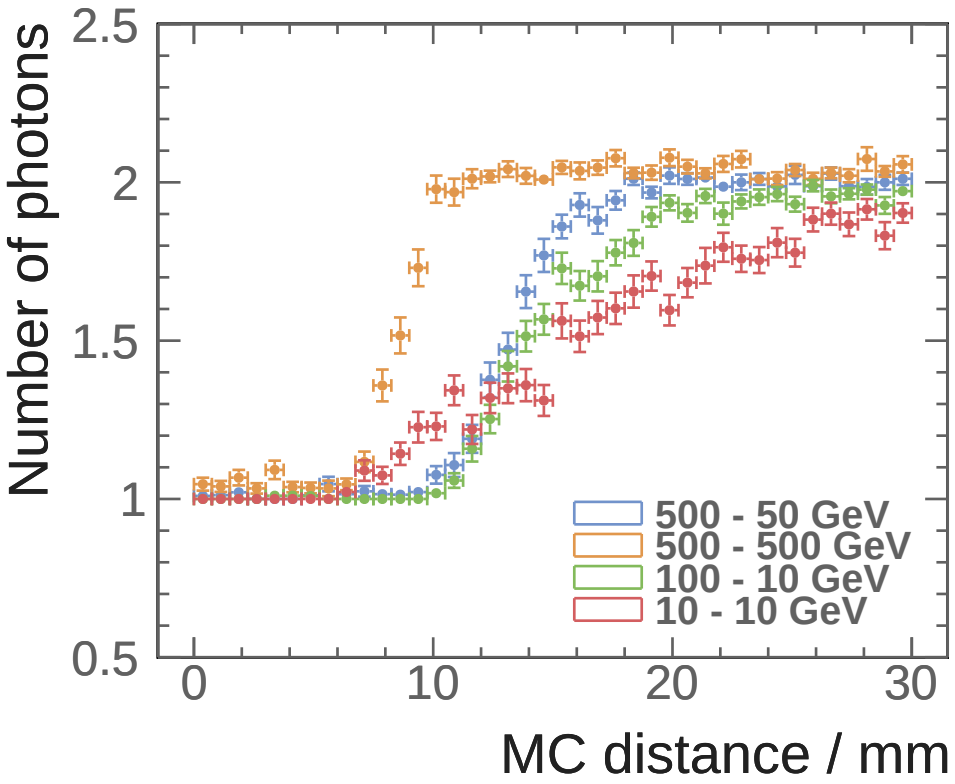


Figure 5.19.: Average numbers of photon for four different photon pairs: 500 - 50 (blue), 500 - 500 (orange), 100 - 10 (green), and 10 - 10 GeV (red), as a function of the Monte Carlo distance separation between the photon pair.

Chapter 6.

Tau Lepton Decay Modes Classification

'Give a man a fish and you feed him for a day; teach a man to fish and you feed him for a lifetime.'

—

The tau lepton has been studied extensively in the past at the Large Electron Positron Collider (LEP) [68]. The tau lepton spin state, which can be derived from kinematic properties of tau decay products, can be used to measure the CP (the product of charge conjugation and parity symmetries) of the Higgs, via $H \to \tau^+\tau^-$ channel [69]. The polarisation correlation of the tau pairs can be used to infer the spin of the parent boson, differentiating $H \to \tau^+\tau^-$ from $Z \to \tau^+\tau^-$.

The ability to identify tau decay mode can be also used as a benchmark for detector performance. Since tau lepton has a very short lifetime of 290 fs [70], only tau decay products can be detected with the tracking detectors and calorimeters. Therefore, the performances of the calorimetric and track systems determine the ability to reconstruct the tau lepton decay products and identify different tau decay modes.

The main challenge in the tau lepton decay modes classification is to reconstruct and separate spatially close photons. Many final states of the tau decay involves π^0, where $\pi^0 \to \gamma\gamma$. For some final states, the main difference in the topologies is the number of photons in the final state. At a high centre-of-mass energy, the decay product of the tau decay are often boosted. To reconstruct two photons from π^0 decay as separate

entities requires good pattern recognition algorithms for photons and a fine ECAL spatial resolution. Hence the photon reconstruction dedicated in chapter 5 is used in this study to identify photons.

This chapter is organised as follows. Firstly, the samples for the analysis will be presented, with the tau decay modes of interests identified. The pre-selection cuts and variables used in the MVA classification are discussed. The performance of the tau decay mode classification will be given, followed by the ECAL optimisation study using the tau decay mode classification. Lastly, the tau decay mode classification is further used in a proof-of-principle analysis to demonstrate the ability to identify H from Z using the tau pair decay channel.

6.1. Overview of the analysis

The analysis starts with defining the samples for study in section 6.2. Seven major tau lepton decay modes are chosen for the classification. The simulation and reconstruction of these tau lepton decays are described in section 6.3. Pre-selection of the events, discussed in section 6.4, are such that the reconstruction and detector effects, which do not vary with the ECAL cell sizes, do not affect this analysis. After defining discriminative variables used in the MVA classification in section 6.5, the classification is performed with a multivariate classifier. Since the decay products of a tau need to be classified into one of the seven decay modes, a multiclass classicisation, presented in section 6.6, is used to allow simultaneous classification between multiple decay modes. Afterwards, the performance of the classification is described in section 6.7.

The classification of the tau lepton decay modes are then repeated for different energies of tau lepton decay to access the impact of the energy on the classification. The classification is also used to study the impact of the ECAL cell sizes on the classification performance, discussed in section 6.8. Lastly, the classification is utilised to demonstrate the ability to separate H from Z using the tau pair decay channel in a proof-of-principle analysis, described in the section 6.9.

6.2. Samples for the analysis

The studied tau lepton decay channel is $e^+e^- \to \tau^+\tau^-$, with a centre-of-mass energy of 100 GeV. An event display of $e^+e^- \to \tau^+\tau^-$ interaction is shown in figure 6.1, simulated with the ILD detector model. The $e^+e^- \to \tau^+\tau^-$ channel contains two tau leptons travelling in opposite directions. Since the tau decay mode classification is applied on a per tau lepton basis, the decay products of two taus in one event are divided into two collections for separate classification. Each collection of particles corresponds to the decay products of one tau lepton.

The principle thrust axis vector is used to separation particles into two collections. Two collections are obtained based on the sign of the scalar product between the principle thrust axis vector and the momentum vector of a particle. The principle thrust axis vector, $\hat{t}$, is chosen by maximising the classical event shape thrust [71], T:

$$T = \max_{\hat{t}} \frac{\sum_i \left|\hat{t}\cdot\vec{p_i}\right|}{\sum_i |\vec{p_i}|} \tag{6.1}$$

where $\vec{p_i}$ is the momentum vector of the particle i; $\hat{t}$, is the unit principle thrust axis vector; and index i is summed over all particles in an event. The principle thrust axis vector is the axis that most particle aligned to.

6.2.1. Tau lepton decay modes

Tau lepton decays into a number of final states. To classify the predominant effect of the tau lepton decays, decay modes with branching ratio above 2% are studied. This results in seven tau lepton decay modes. Their branching ratios, along with decay modes and final states are shown in table 6.1. Thus the seven tau lepton decay modes, which cover 92.58 % of total branching ratio of tau decay, are studied in this analysis.

6.3. Simulation and reconstruction

Two million $e^+e^- \to \tau^+\tau^-$ events are simulated and reconstructed using the ILD detector model. As the study is aimed for optimisation study of the ECAL cell sizes, the beam

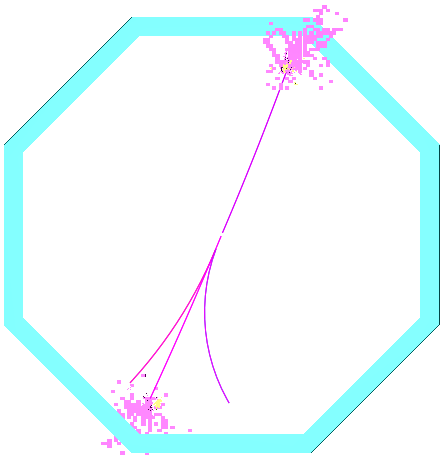


Figure 6.1.: An example event display of a simulated $e^+e^- \to \tau^+\tau^-$ event using the ILD detector model. The top half of the event is a tau lepton decaying into $\pi^-\pi^0$ final state and the bottom half of the event is a tau lepton decaying into $\pi^+\pi^-\pi^-\pi^0$ final state. The purple lines are the tracks left by $\pi^\pm$ in the tracking detectors. The purple clusters are the calorimeter hits if $\pi^\pm$ and the yellow clusters are the calorimeter hits of photon from $\pi^0 \to \gamma\gamma$. The blue region is the transverse cross section of the ECAL barrel part along the beam line direction.

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

Table 6.1.: Decay modes, detectable final state particles and branching ratios of the seven major τ^- decays, taken from [3]. τ^+ decays similarly to τ^-.

specific effects that are not affected by the varying ECAL cell sizes are not simulated. These effects include the initial state radiation (ISR) and the beam induced background.

The software used for simulation and reconstruction is described in chapter 4. Events are reconstructed with iLCSoft version v01-17-07 [72] and PandoraPFA version 3 [46], where the photon reconstruction is discussed in chapter 5.

6.4. Event pre-selection

Pre-selection cuts select events using the truth information. Since the analysis is aimed for the optimisation of the ECAL cell sizes, these pre-selection cuts are not affected by the changing of the ECAL cell sizes. These cuts allow the analysis to focus on the events with clear topologies. The pre-selection cuts are listed in table 6.2. The fraction of events passing each pre-selection cut for individual decay mode are listed in table 6.3.

One of the pre-selection cuts is to demand that the tau decay products do not have photons converted to electron pairs in the tracking detector, determined with the truth information. These discarded events would have fewer photons and more electrons than expected in the final states, which changes the topologies of the final states. Shown in table 6.3, only decay modes with photons in the final states are affected by this cut, as expected.

Another pre-selection cut requires the total energy of the non-neutrino tau decay products, $E_{vis,MC}$, to be greater than 5 GeV, based on the truth information. If most energy of a tau lepton is carried by neutrinos, non-neutrino decay products would have low energies and be difficult to be identified. Hence these events with low-energy non-neutrinos tau decay products are not used in the analysis. Decay modes with only one non-neutrino particle in the final states are mostly affected by this cut, indicated in table 6.3.

The last pre-selection cut is to discard events with tau decay products depositing energies in the gap region between barrel and the end cap part of the calorimeter. As the reconstruction does not attempt to recover reconstruction in the gap region, there is a significant drop in the particle reconstruction efficiency. The cut demands the absolute value of the polar angle of tau lepton, based on the truth information, $|\theta_{Z,MC}|$, is between 0.3 and 0.6 rad to be contained in the end cap region, or is between 0.8 and 1.57 to be

contained in the barrel region. All decay modes are affected almost equally by this cut, suggested by numbers in table 6.3.

Cuts	Values
Photon conversion in the tracking detector	No
Total energy of non-neutrino 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 modes classification.

Detectable final state	No photon conversion in the tracking detector	Total energy of non-neutrino decay products 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.: The fraction of events passing each pre-selection cut for individual decay mode. Cuts are presented in a "flow" fashion, where each cut contains all the cuts to its left.

6.5. Variables used in the MVA

Having pre-selected events, variables are carefully developed for the multivariate analysis (MVA). The full list of the variables are shown in table 6.4. The distribution of the four most power variables for selected tau decay modes are shown in figure 6.2.

6.5.1. PFOs number variables

The most crucial variables are the number of PFOs of different types of particles. There are five PFOs number variables used in MVA event selection: the number of charged particles (N_{χ^+}); the number of muons (N_μ); the number of electrons (N_e); the number of photons (N_γ); and the number of charged pions (N_{π^-}).

Figure 6.2a shows the distributions of the number of charged particles for selected tau decay modes. Whilst over 98% one-prong final states have one track reconstructed, around 95% three-prong final states have three tracks reconstructed. Figure 6.2b shows the distributions of the number of photons, which is powerful to distinguish final states with different numbers of π^0 s.

6.5.2. Invariant mass variables

Five invariant mass variables participate the MVA classification: the invariant mass of all non-neutrino decay products (m_{vis}); the invariant mass of all charged particles (m_{χ^+}); the invariant mass of all neutral particles (m_{χ^0}); the invariant mass of all photons (m_γ); and the invariant mass of all charged pions (m_{π^-}). Figure 6.2c shows distribution of the invariant mass of all non-neutrino decay products for selected tau decay modes. Resonance structures can be seen for ρ and a_1 decay modes in the figure.

6.5.3. Energy variables

Energy information helps to further separate different final states. Six energy variables are used in the MVA classification: the normalised total energy of all non-neutrino decay products ($\tilde{E}_{vis}$); the normalised total energy of the charged particles ($\tilde{E}_{\chi^+}$); the normalised total energy of the muons ($\tilde{E}_\mu$); the normalised total energy of the electrons ($\tilde{E}_\mathrm{e}$); the normalised total energy of the photons ($\tilde{E}_\gamma$); and the normalised total energy of the charged pions ($\tilde{E}_{\pi^-}$). All variables are normalised with respect to the energy of the associated tau lepton.

6.5.4. Calorimetric information variables

Two calorimetric information variable are used in the MVA classification: the fraction of the energy deposited in the ECAL over the energy deposited in the calorimeters for all charge particles ($\%E_{\chi^+}$) and the fraction of the energy deposited in the ECAL over the energy deposited in the calorimeters for all particles ($\%E$). These are two variables help to identify electrons and muons. An electron deposits most energy in the ECAL and a muon deposits 5% to 20% energy in the ECAL. The difference between two variables is that photons, which deposit most of their energy in the ECAL, do not participate in the calculation of $\%E_{\chi^+}$.

6.5.5. $\rho(\pi^-\pi^0)$ and $a_1(\pi^-\pi^0\pi^0)$ resonances reconstruction variables

$\rho(\pi^-\pi^0)$ and $a_1(\pi^-\pi^0\pi^0)$ decay modes are identified further using their invariant mass resonance structure. For example, $\rho(\pi^-\pi^0)$ decay mode contains a π^- and a π^0 decaying into two photons. By selecting π^- and photons consistent with ρ mass, $\rho(\pi^-\pi^0)$ decay mode could be identified. The $\rho(\pi^-\pi^0)$ decay mode hypothesis test performed by minimising a χ^2 function:

$$\chi^2 = \left(\frac{m_{tot} - m_{\rho}^{MC}}{\sigma_{\rho}^{MC}}\right)^2 + \left(\frac{m_{\gamma\gamma} - m_{\pi^0}^{MC}}{\sigma_{\pi^0}^{MC}}\right)^2, \tag{6.2}$$

where $m_{\gamma\gamma}$ is the invariant mass of two photons; m_{tot} is the invariant mass of the two photons and one π^-; m_{ρ}^{MC} and $m_{\pi^0}^{MC}$ are true masses of ρ and π^0, respectively, taken from [3]; and σ_{ρ}^{MC} and $m_{\pi^0}^{MC}$ are the half width of the invariant mass distribution of reconstructed ρ and π^0, respectively, obtained using the truth information. All combinations of photons and π^- are tested. Two variables obtained in this minimisation and used in the MVA classification are the invariant mass of the two photons in the fit, $m_{\pi^0}(\rho)$, and the invariant mass of the two photons and one π^-, m_{ρ}.

Similarly, $a_1(\pi^-\pi^0\pi^0)$ decay mode can be identified using an extended minimisation function:

$$\chi^2 = \left(\frac{m_{tot} - m_{a_1}^{MC}}{\sigma_{a_1}^{MC}}\right)^2 + \left(\frac{m_{\gamma 1\gamma 2} - m_{\pi^0}^{MC}}{\sigma_{\pi^0}^{MC}}\right)^2 + \left(\frac{m_{\gamma 3\gamma 4} - m_{\pi^0}^{MC}}{\sigma_{\pi^0}^{MC}}\right)^2, \tag{6.3}$$

where ρ has been replace by a_1 and other variables are defined in the same way as in the previous χ^2 function in equation 6.2. Two photon pairs and one π^- are needed for this minimisation. Both photon pairs are required to be consistent with the mass of π^0. Three variables obtained in this minimisation and used in the MVA classification are the invariant mass of the first two photons in the fit, $m_{\pi^0}(a_1)$; the invariant mass of the last two photons in the fit, $m^*_{\pi^0}(a_1)$; and the invariant mass of the four photons and one π^-, m_{a_1}. The first photon pair is defined to have a invariant mass closer to the invariant mass of the π^0 than the second photon pair. Figure 6.2d shows the distributions of the invariant mass of m_{a_1} under $a_1(\pi^-\pi^0\pi^0)$ hypothesis test for selected tau decay modes. Only the distribution for $a_1(\pi^-\pi^0\pi^0)$ decay mode has a resonance peak at a_1 mass.

The χ^2 functions for both hypothesis test are adapted for events where the event reconstruction fails to reconstruct enough photons. Relevant terms are dropped from the expression if there are fewer photons reconstructed than required in the χ^2 functions.

6.5.6. Separate e^- from π^-

The Particle ID obtained from PandoraPFA is used extensively to reconstruct variables, as PandoraPFA uses a wide range of information to determine electron the ID. However, extra variables are used in this analysis to help further identifying electrons, which could be 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 characterising the EM shower helps to identify e^-. Three variables are used in the MVA classification: the start layer of the longitudinal shower (t_0); the fractional difference between observed and expected longitudinal shower profile describe the longitudinal EM shower (δl); and $\langle w \rangle$, a measure of the EM shower transverse width. These variables are taken from the photon ID step in the photon reconstruction in PandoraPFA, described in section 5.5.

Another type of information to differentiate an EM shower from a early hadronic shower from a π^- is the calorimeter hit information. Two variables used in the MVA classification are: the average energy of a calorimeter hit ($\bar{E}_{hit}$) and the average fraction of possible minimum ionising calorimeter hit for all particles ($\%MIP$)

Last information used to separate e^- from π^- is the track-momentum-calorimeter-energy consistency check. The variable used in the MVA classification is the calorimeter energy divided by the track momentum for all particles ($\Delta E/P$).

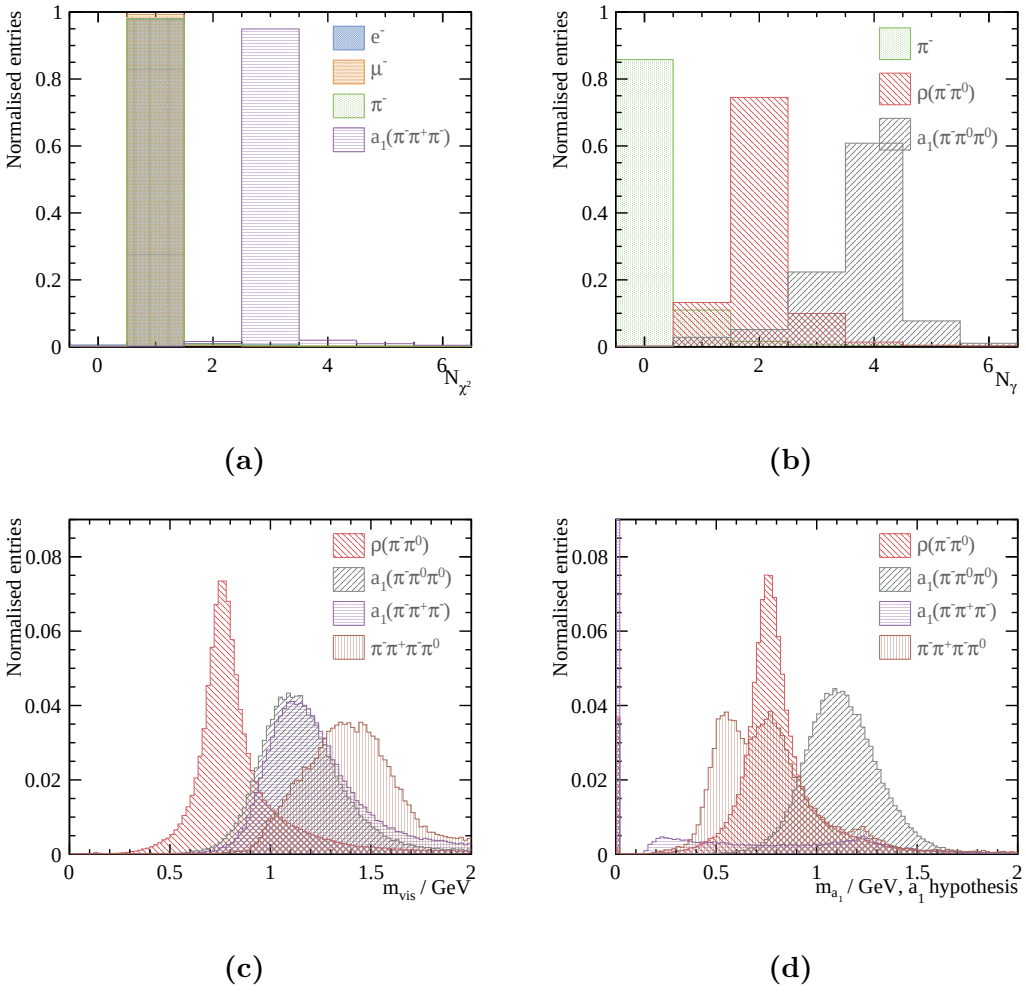


Figure 6.2.: Normalised distribution for a) the number of charged particle (N_{χ^+}); b) the number of photons (N_γ); c) the invariant mass of all non-neutrino decay products (m_{vis}); and d) the invariant mass of the a_1, reconstructed with $a_1(\pi^-\pi^0\pi^0)$ hypothesis. Decay modes in all plots are selected using the truth information.

Category	Variable
PFOs number	N_{χ^+}, N_{μ}, N_{e}, N_{γ}, N_{π^-}
Invariant mass	m_{vis}, m_{χ^+}, m_{χ^0}, m_{γ}, m_{π^-}
Energy	$\tilde{E}_{vis}$, $\tilde{E}_{\chi^+}$, $\tilde{E}_{\mu}$, $\tilde{E}_{\mathrm{e}}$, $\tilde{E}_{\gamma}$, $\tilde{E}_{\pi^-}$
Calorimetric info.	$\%E_{\chi^+}$, $\%E$
$\rho(\pi^-\pi^0)$ reconstruction	$m_{\pi^0}(\rho)$, m_{ρ}
$\mathrm{a}_1(\pi^-\pi^0\pi^0)$ reconstruction	$m_{\pi^0}(\mathrm{a}_1)$, $m^*_{\pi^0}(\mathrm{a}_1)$, 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 classification for the tau lepton decay mode classification.

6.6. Multivariate Analysis

For the multivariate analysis, the multiclass class of the TMVA package [73] was used to perform a multiclass classification, which classifies seven tau lepton decay final states simultaneously. The multiclass classifier is an extension of a standard two-class signal-background classifier. The discussion on multivariate analysis can be found in section 4.7. In particular the multiclass classifier is discussed in section 4.7.9.

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. An explanation of the variables can be found in section 4.7.8. Half of the randomly selected samples were used in the training process and the other half were used for testing.

6.7. Tau decay mode classification efficiency

The classification efficiencies for the seven tau decay modes are shown in table 6.6. Bold numbers show the correct classification probabilities.

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 4.7.8 for a detailed explanation of variables.

For the e^- decay mode, 99.8% correct classicisation efficiency is achieved. For μ^- decay mode, 99.5% correct classicisation efficiency is achieved, due to an effective track reconstruction and muon reconstruction algorithms.

For the π^- decay mode, 3.4% events were misclassified as $\rho(\pi^-\pi^0)$ decay events. If the reconstruction is unable to reconstruct the photon pair from π^0 decay in $\rho(\pi^-\pi^0)$ decay mode, the π^- and $\rho(\pi^-\pi^0)$ decay events would appear to similar and misclassification is caused. Only 0.9% of π^- decay events are misclassified as e^- decay, due to variables dedicated to separation between e^- and π^-.

For the $\rho(\pi^-\pi^0)$ decay mode, most misclassification comes from the confusion with $a_1(\pi^-\pi^0\pi^0)$ decay mode. If the reconstruction is unable to resolve the all photon pair from π^0 decay in $\rho(\pi^-\pi^0)$ and $a_1(\pi^-\pi^0\pi^0)$ decay mode, the two decay modes would have similar topologies.

For the $a_1(\pi^-\pi^0\pi^0)$ decay mode, the correct classification rate is the lowest among seven decay modes, as the $a_1(\pi^-\pi^0\pi^0)$ decay final state is the 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 photon 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, where the distribution overlaps with the distribution for $\rho(\pi^-\pi^0)$ decay mode. The $a_1(\pi^-\pi^0\pi^0)$ resonance reconstruction and the multiclass classifier reduce the confusion between two decay modes from 30% to 9.5%.

For the $a_1(\pi^+\pi^-\pi^-)$ decay mode, the biggest source of misclassification is with $\pi^+\pi^-\pi^-\pi^0$ decay mode. The biggest misclassification of $\pi^+\pi^-\pi^-\pi^0$ decay mode is with $a_1(\pi^+\pi^-\pi^-)$ decay mode.

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, using $e^+e^- \to \tau^+\tau^-$ channel at $\sqrt{s} = 100\,\text{GeV}$. Bold numbers show the correct classification probabilities. ν_τ are not shown in decay modes. - represents a number below 0.25%. Statistical uncertainties are less than 0.25%.

6.8. Electromagnetic calorimeter optimisation

In above sections, an analysis on tau decay mode classification is presented. Events used in the analysis were $e^+e^- \to \tau^+\tau^-$ events at $\sqrt{s} = 100\,\text{GeV}$ with the nominal ILD detector model. In this section, the analysis was repeated with varying ECAL square cell sizes at 3, 5, 7, 10, 15 and 20 mm, at four centre-of-mass energies of 100, 200, 500, 1000 GeV. Other ECAL dimensions are kept the same as the ILD nominal detector. The multivariate classifier was trained for each ECAL cell size and each centre-of-mass energy. Because the lepton reconstruction mostly relies on the tracking system, which was not varied in this study, only the hadronic tau decay modes were investigated. The correct classification efficiencies for tau hadronic decay final states as a function of the ECAL square cell sizes for different centre-of-mass energies are shown in figure 6.3.

As PandoraPFA is optimised for the nominal ILD detector, a re-optimisation is required when changing the ECAL square cell sizes. In particular, fragment removal algorithms have a large dependence on the ECAL cell sizes. For example, the PhotonFragmentRemoval algorithm which merges photon fragments uses a distance metric that depends on the ECAL cell sizes. Table 6.7 shows the optimised distance metrics as

a function of the ECAL square cell size. As cell sizes become larger, the distance cut for merging photons become larger, as expected.

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

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

As the centre-of-mass energy increases, tau decay products are often boosted. It is increasingly difficult to separate tau decay products. For example, the photon pair from π^0 decay becomes very challenging to separate at a high centre-of-mass energy. Therefore, the inability to separate photon pairs will degrade the classicisation performance.

An increase of the ECAL cell sizes has a similar effect of degrading the classicisation performance. The change in the ECAL cell size will change the transverse spatial resolution. Hence a large cell size will result in a low transverse spatial resolution, and leading to inability to separate photon pairs. Consequently, a worse classification performance is expected for a larger ECAL cell size.

Supported by figure 6.3, tau decay mode correct classification efficiencies generally decrease with an increase of centre-of-mass energies and an increase of ECAL cell sizes, as expected. This trend is observed for almost all tau decay modes.

For the $\rho(\pi^-\pi^0)$ decay mode,the efficiency for $\sqrt{s} = 500\,\text{GeV}$ increases as the cell sizes increases. This is because the multivariate classifier optimises for the overall classification efficiency, which balances the decrease of the efficiency of one decay mode by the increase of the efficiency of another decay mode. 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 an increasing ECAL cell size and increasing centre-of-mass energy 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 centre-of-mass energies.

For the $a_1(\pi^+\pi^-\pi^-)$ decay mode, the efficiencies are similar to that of the π^- decay mode. Both final states contain charged particles only. Therefore it is most sensitive to the tracking performance, which is not affected by varying 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 correct reconstruction efficiency of the tau leptonic decay 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 also used for the association to the tracks but it has a small impact on the lepton identification.

6.8.1. Tau hadronic decay correct classification efficiency

A single parameter for overall tau decay efficiency is constructed for two reasons. Firstly the multivariate classifier is trained to optimised to achieve the best the overall classification efficiency. Secondly it is easier to compare the impact of different detector models and different centre-of-mass energies with a single parameter.

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

$$\varepsilon_{had} = \frac{\sum_i^5 Br_i \varepsilon_i}{\sum_i^5 Br_i}\,, \tag{6.4}$$

where Br_i is the branching fraction of the hadronic decay mode i after the pre-selection cuts; ε_i is the correct reconstruction efficiency of the decay mode i; and index i is summed over five tau hadronic decay modes.

Figure 6.4 shows ε_{had} as a function of ECAL cell sizes with increasing centre-of-mass energies. The general trend for the ε_{had} is that ε_{had} decreases with the increase of centre-of-mass energies and the increase of ECAL cell sizes because it is increasingly difficult to reconstruct photons with boosted particles and lower ECAL transverse spatial resolutions.

At $\sqrt{s} = 100\,\text{GeV}$, the ε_{had} decreases from 94% at 3 mm cell size, to 91% at 20 mm cell size. The decrease is approximately linear to the increase in the cell size. The decrease in ε_{had} is greater at $\sqrt{s} = 200\,\text{GeV}$, where ε_{had} declined from 94% at 3 mm cell size, to

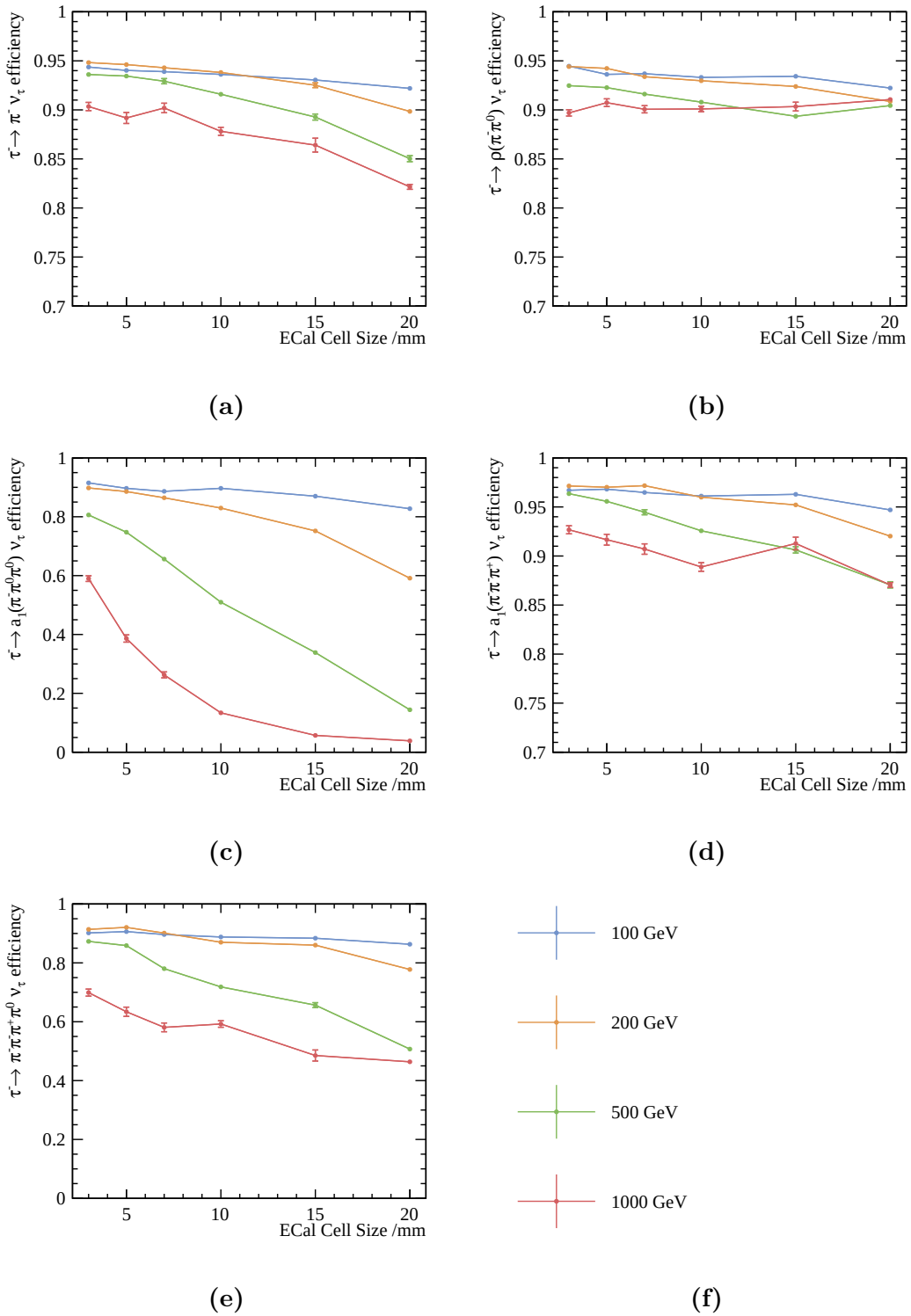


Figure 6.3.: The correct classification efficiencies for tau hadronic decay final states as a function of the ECAL square cell sizes for a) π^- decay mode, b) $\rho(\pi^-\pi^0)$ decay mode, c) $a_1(\pi^-\pi^0\pi^0)$ decay mode, d) $a_1(\pi^+\pi^-\pi^-)$ decay mode, and e) $\pi^+\pi^-\pi^-\pi^0$ decay mode. The legend is shown in f). All plots are produced using the ILD detector model with $\sqrt{s}$ = 100, 200, 500 and 1000 GeV.

86% for a ECAL cell size of 20 mm. Most significant decrease in the ε_{had} occurs at $\sqrt{s} =$ 500 GeV, where the ε_{had} decreases from 92% at 3 mm cell size, to 78% at 20 mm cell size. At $\sqrt{s} = 1000$ GeV, the ε_{had} drops from 85% at 3 mm cell size, to 75% at 20 mm cell size.

The increase in ECAL cell sizes has a larger impact in tau decay classification at high centre-of-mass energies. With decay products being spatially close at high centre-of-mass energies, it is more beneficial to have a smaller ECAL cell size to reconstruct individual particle.

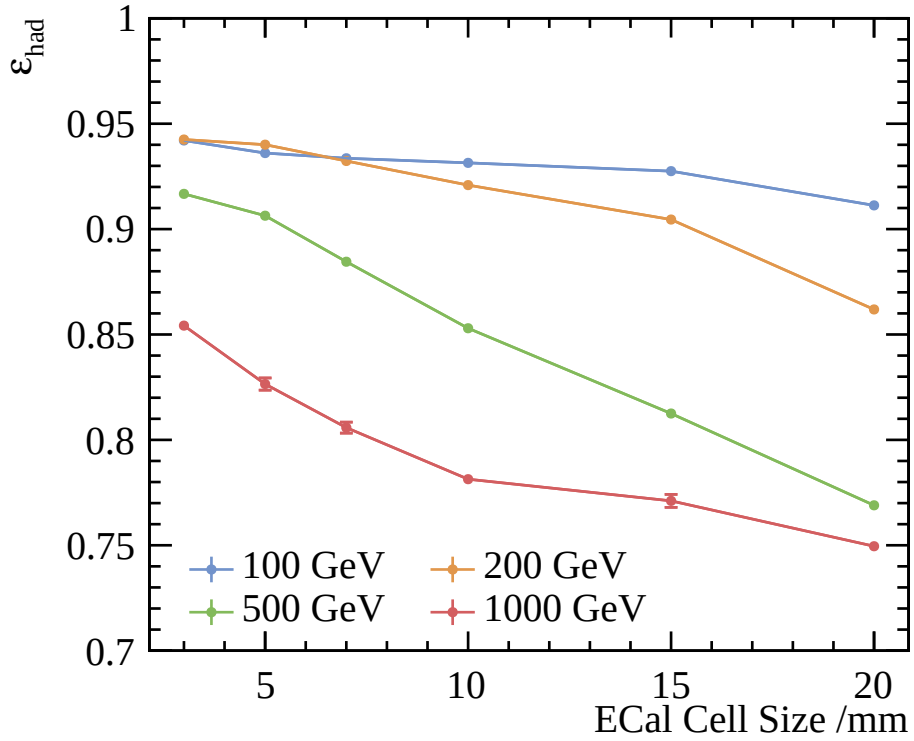


Figure 6.4.: The tau hadronic decay efficiency, ε_{had}, as a function of the ECAL cell sizes with different centre-of-mass energies using the ILD detector model. The blue, orange, green and red lines represent the efficiencies at $\sqrt{s} = 100$, 200, 500 and 1000 GeV respectively.

6.9. Tau pair polarisation correlations as a signature of Higgs boson

Many BSM theories predict the $H\tau^+\tau^-$ coupling would dominate the Higgs boson to leptons couplings. Therefore, if an experiment observes an excess of tau pair decay events, it could be an indication of the Higgs boson. Here, this section follows the theoretical discussion in section 2.10 to present a proof-of-principle analysis using tau pair polarisation correlation as a signature of Higgs boson.

H can be separated from Z using tau pair decay channel. Comparing $H \to \tau^+\tau^-$ and $Z \to \tau^+\tau^-$, the difference in the spin of the bosons reflects in the different polarisation correlation of the tau pair. By extracting the polarisation correlation of the tau pair, the parent boson can be identified.

The subsequent sections discuss the ability to reconstruct the polarisation correlation of the tau pair with $Z \to \tau^+\tau^-$ channel, where both $\tau^- \to \pi^-\nu_\tau$. The analysis starts with the event pre-selection, followed by identifying the tau decay products in the events. Afterwards, the tau decay mode classification is used to identify $\tau^- \to \pi^-\nu_\tau$ decays. Lastly the tau pair polarisation correlation is presented and compared to the distribution obtained with Monte Carlo simulation.

6.9.1. Event pre-selection

The channel to study 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.

The same seven tau decay modes in section 6.2 are studied. The $\tau^- \to \pi^-\nu_\tau$ decay mode is selected for the proof-of-principle analysis of H/ Z separation with tau pair decay channel.

The event pre-selection is similar to that in section 6.4. The cut on the total energy of non-neutrino decay products is not used, because a large fraction of $Z \to \tau^+\tau^-$ events, where $\tau^- \to \pi^-\nu_\tau$, has two low-energy charged pions. Therefore, the cut on the energy of non-neutrino decay products would lose many events.

6.9.2. Find tau decay products

The final state of the selected channel, $e^+e^- \to ZZ \to \tau^+\tau^- qq$, contains two tau leptons and two quark jets. Therefore, tau decay products can either be found by direct tau lepton searching, or by using jet algorithm to find tau decay products as jets. If a tau lepton decays into a few particles, then the direct tau searching would work better. If a tau lepton decays into many particles, finding tau decay products as a jet has a better performance, as jet clustering works better with more particles. Hence, two approaches are combined for the best tau pair identification.

Direct tau searching

Tau finder processer, IsolatedTauIdentifer, is a modified version of the one in section 7.3.2. The basic idea is to find tau decay products consistent with tau decay topologies, and requires the tau decay products to be isolated from the rest of the particles. Parameters chosen are set to find as many tau candidates as possible.

Table 6.8 lists all the cuts used in the IsolatedTauIdentifer. Particles with transverse momentum (p_T) less than 0.5 GeV 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 invariant mass of all particles inside the search cone less than 3 GeV. The maximum search cone opening angle (θ_S) is $\cos^{-1}(0.99)$. The isolation criteria states that the opening angle between the search cone and the 2^{nd} closest track ($\theta_{cone,2^{nd}X^+}$) is larger than 0.6 rad. If the criteria is satisfied, the search cone with the tau seed is identified as a tau lepton.

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

Table 6.8.: Optimised parameters of the modified IsolatedTauIdentifer.

Jet clustering

Tau hadronically decay products can be also identified as a small jet. The Durham algorithm, also known as the e^+e^- k_t algorithm [59], was used to form jets (see section 4.6.4). The jet algorithm runs in the exclusive mode to find four jets for $e^+e^- \to ZZ \to \tau^+\tau^- qq$ channel.

Select tau candidates

The best tau pair candidates are selected using kinematic constraints. In the $e^+e^- \to ZZ$, the energy of the Z is half of the centre-of-mass energy. The invariant mass of two quarks from Z should be close to Z mass. Therefore, the minimisation function utilising kinematic constraints is

$$\chi^2 = \frac{\left(m_{qq} - m_Z\right)^2}{\sigma^2_{m_{qq}}} + \frac{\left(E_{qq} - \frac{\sqrt{s}}{2}\right)^2}{\sigma^2_{E_{qq}}}, \tag{6.5}$$

where $\sqrt{s}$ is the centre-of-mass energy; m_Z is the mass of Z from reference [3]; $\sigma_{m_{qq}}$ and $\sigma_{E_{qq}}$ are the reconstructed mass resolution and energy resolution of the $Z \to qq$, respectively; m_{qq} and E_{qq} are defined differently for the direct tau searching and the jet clustering method; and the minimisation is iterated over all tau pairs from direct tau searching method or over all jets from jet clustering method. For the direct tau searching method, m_{qq} and E_{qq} are defined as the recoil momenta against two tau candidates, assuming collisions happening at $\sqrt{s}$. For the jet clustering method, m_{qq} and E_{qq} are defined as the mass and energy of two jets.

The χ^2 minimiser is repeated for the direct tau searching method and the jet clustering method. Each method selects a best tau pair candidate with the smallest χ^2. Hence two tau pair candidates with χ^2 are obtained. To find the best overall tau pair candidate, a set of conditions is used. If the best tau pair candidate from both methods satisfies the kinematic constraint:

$$\left|m_{qq} - m_Z\right| < \sigma_{m_{qq}} \ , \left|E_{qq} - \frac{\sqrt{s}}{2}\right| < \sigma_{E_{qq}}, \tag{6.6}$$

the tau pair candidate with smallest χ^2 is selected. Otherwise, if only one tau pair candidate satisfies the constraint in equation 6.6, 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 candidates from IsolatedTauIdentifer, then these two tau candidates are chosen. This is due to the fact that if one jet is close to the beam pipe, it is likely that some particles close to the jet are undetected, which leads to a failure in the kinematic constraint or the jet reconstruction. Lastly, if all conditions above are not satisfied, two smallest jets by the number of PFOs are chosen to be the best tau pair candidate.

6.9.3. Boost tau decay products to Z decay rest frame

The previous section describes the method to identify the tau pair decay products. To use the tau decay mode classifier, it is necessary to know the tau lepton energy. For the channel $Z \to \tau^+\tau^-$, the energy of the tau lepton can only be obtained in $Z \to qq$ decay rest frame, which is half of the Z energy in the $Z \to qq$ rest frame. Hence the tau decay products need to be boosted to the Z decay rest frame for the calculation of the variables used in the MVA classification. The boosting requires the four-momentum of the Z.

The four-momentum of the Z decaying to tau pair is calculated from the recoil momenta of non tau-decay-products:

$$p^{\mu}_{\tau\tau} = \begin{pmatrix} \sqrt{s} \\ \sqrt{s} \times \sin(\theta_{beam}) \\ 0 \\ 0 \end{pmatrix} - \sum_{i}^{non-\tau} p^{\mu}_{i}, \tag{6.7}$$

where θ_{beam} is the beam crossing angle; $\sqrt{s}$ is the centre-of-mass energy; p^{μ}_{i} is the four-momentum vector of the particle i; $p^{\mu}_{\tau\tau}$ is the four-momentum vector of the Z, where $Z \to \tau^+\tau^-$; and index i is summed over all non-tau-decay-product PFOs. Extra kinematic constraint fixes the energy of the $p^{\mu}_{\tau\tau}$ to be half of $\sqrt{s}$:

$$p^{\mu}_{\tau\tau,correct} \equiv p^{\mu}_{\tau\tau} \times \frac{\frac{1}{2}\sqrt{s}}{E_{\tau\tau}}, \tag{6.8}$$

where $E_{\tau\tau}$ is the energy of the vector $p^{\mu}_{\tau\tau}$ and other variables are defined in the same way as in the previous equation. $p^{\mu}_{\tau\tau,correct}$ is then treated as the four-momentum vector of Z, where $Z \to \tau^+\tau^-$. Tau decay products are boosted to the Z decay rest frame accordingly.

The calculation of the variables used in the MVA classifier are then performed in the $Z \to qq$ decay rest frame.

6.9.4. Variables used in the MVA

Variables used in the MVA classifier are a subset of the ones used in the previous analysis. Variables regarding EM shower profiles, calorimeter hit information and track information are not used (the last three rows in table 6.4) as the study focuses on the overall tau decay mode separation.

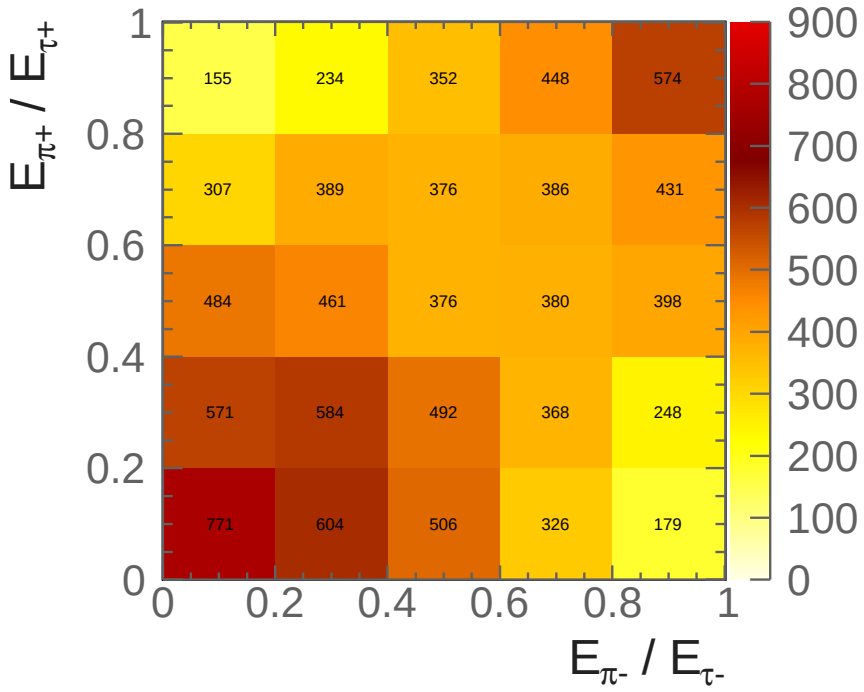
6.9.5. Multivariate analysis

Half of the randomly selected sample is used to train the multivariate classifier, which follows the procedure in section 6.6. The same classifier as in the previous analysis is used . In the classifier applying stage, $\tau^- \to \pi^- \nu_\tau$ decay mode is selected with an additional criteria that there is at least one $\pi^\pm$ among the tau decay products.

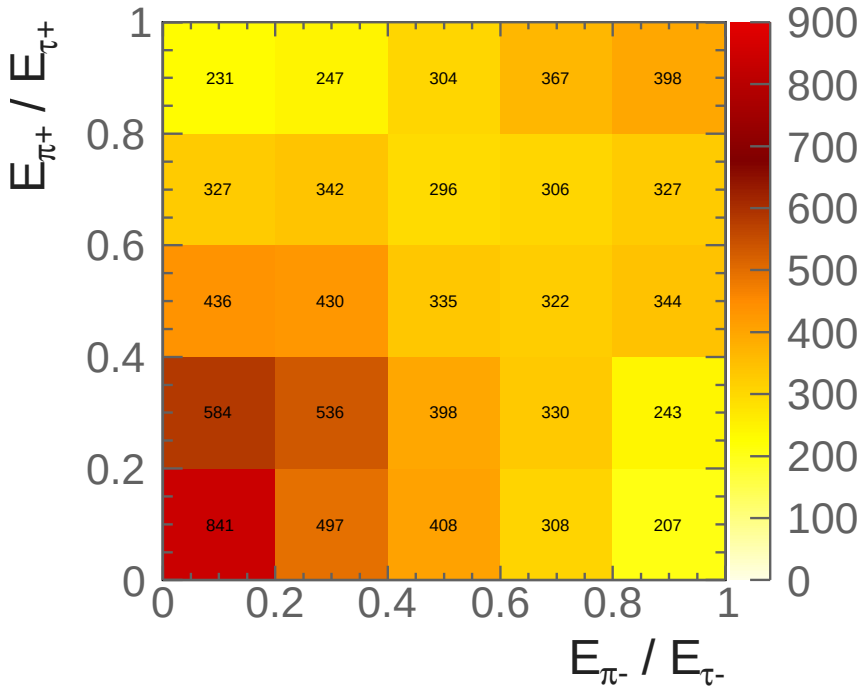
6.9.6. Result

Figure 6.5 shows the two-dimensional plot of tau pair polarisation correlations from Z decay, using $\tau^- \to \pi^- \nu_\tau$ decay mode, with $e^+e^- \to ZZ$ channel where one Z decays to a tau pair and the other Z decays hadronically. The energy fractions of the tau decay product to the tau lepton are the appropriate kinematic variables, motivated in the theoretical discussion in section 2.10. Figure 6.5a shows the distribution obtained with the Monte Carlo particles. Figure 6.5b shows the distribution using full detector simulation. Dark regions along the diagonal can be seen in both the distribution for the full detector simulation and the distribution for the Monte Carlo simulation. In the $Z \to \tau^+\tau^-$ decays, an energetic π^+ is likely to be associated with an energetic π^- and a low-energy π^+ is likely to be associated with a low-energy π^-. This trend is shown in both the distribution produced with the Monte Carlo particles and with the full detector simulation. Comparing the two figures inFigure 6.5, some events in the top right quadrant, resembling both $\pi^\pm$ being energetic, are not reconstructed correctly. This is due to the incorrect finding of the tau pair decay products (section 6.9.2).

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 ILD detector model. With a similar study of $H \to \tau^+\tau^-$, the tau polarisation correlations can be used to separate H from Z, and to identify Higgs boson in an experiment that observes the breaking of the lepton universality by favouring tau pair events.



(a) Monte Carlo particles



(b) Simulated and reconstructed particles

Figure 6.5.: Two-dimensional histograms of $E_{\pi^+}/E\tau^+$ as a function of $E_{\pi^-}/E\tau^-$ obtained with $Z \to \tau^+\tau^-$ channel , selecting $\tau^- \to \pi^-\nu_\tau$ decay mode for both taus, for a) Monte Carlo particles, and b) simulated and reconstructed particles.

Chapter 7.

Double Higgs Boson Production Analysis

'The supreme art of war is to subdue the enemy without fighting.'
— Sun Tzu, 544 BC - 496 BC

Since the Higgs was discovered at the LHC in 2012 [1, 2], for the particle physicists, it is crucial to understand the interaction between the Higgs and other particles. Another issue to investigate is whether the discovered Higgs is a Standard Model Higgs. A number of Higgs theories beyond the Standard Model may be tested via the double Higgs production in an electron-positron collider. Studying the double Higgs production allow the measurement of the Higgs trilinear self coupling, g_{HHH}, and the quartic coupling, g_{WWHH}. Monte Carlo studies have shown that the precision of the g_{HHH} reached by a multi-TeV linear collider, such as the Compact Linear Collider (CLIC), is superior to the LHC and the HL-LHC [18].

In e^+e^- collisions, there are two main constraints to study the double Higgs production, $e^+e^- \to HH\nu_e\overline{\nu}_e$. Firstly, the process has a small cross section. With 0.149 fb at $\sqrt{s} = 1.4\,\text{TeV}$ and 0.588 fb at $\sqrt{s} = 3\,\text{TeV}$, it is challenging to select signal events. The other challenge is that at high centre-of-mass energies, events are often boosted. Many final-state particles are in the forward region of the detector, where the reconstruction performance is inferior to the barrel region. Particles can escape detection, causing a degradation in the event reconstruction performance.

In this chapter, a full CLIC_ILD detector simulation study has been performed for the double Higgs production channel, $e^+e^- \to HH\nu_e\overline{\nu}_e$, via W^+W^- fusion. Event generation and simulation will be discussed first. An overview of the analysis, including lepton finding and jet reconstruction, is presented, followed by an optimised multivariate analysis to distinguish signal from background processes. The optimised event selection is used to derive an estimate of the uncertainty on g_{HHH} and g_{WWHH} measurements at the CLIC. Part of this analysis has been published in [20].

7.1. Analysis Straggly Overview

The study of the double Higgs production via W^+W^- fusion can probe the Higgs trilinear self coupling, g_{HHH} and quartic coupling, g_{WWHH}. Leading-order Feynman diagrams for double Higgs production via W^+W^- fusion are shown in figure 7.1. The diagram shown in figure 7.1a contains a triple Higgs vertex, which is sensitive to the Higgs trilinear self coupling g_{HHH}. The diagram in the figure 7.1b is sensitive to the quartic coupling g_{WWHH}. Figure 7.1c and figure 7.1d show the Feynman diagrams for irreducible background processes in the study of g_{HHH} and g_{WWHH}.

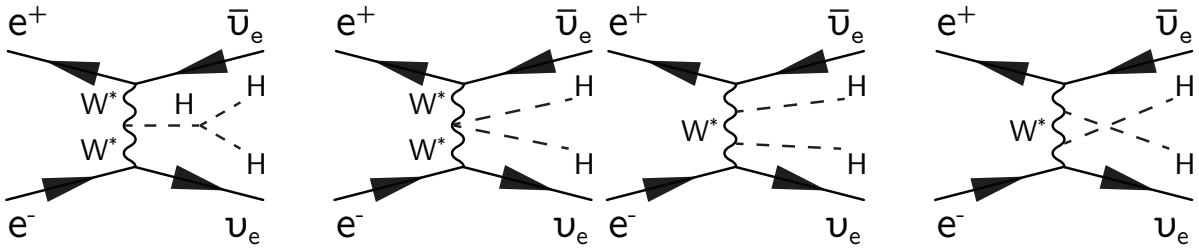


Figure 7.1.: The main Feynman diagrams for the leading-order $e^+e^- \to HH\nu_e\overline{\nu}_e$ processes at the CLIC.

Double Higgs production can be also produced via $e^+e^- \to ZHH$, where Z decays to $\nu\overline{\nu}$. This ZHH channel has been used for study at future e^+e^- colliders, for example, the ILC at $\sqrt{s} = 500\,\text{GeV}$ [25]. However, for the relevant CLIC energies of $\sqrt{s} = 1.4\,\text{TeV}$ and $3\,\text{TeV}$, its contribution to the $HH\nu\overline{\nu}$ final state is small compared to that of the W^+W^- fusion, and can be neglected.

From an experimental prospective, the two Higgs in the $e^+e^- \to HH\nu_e\overline{\nu}_e$ decay to a range of particles. Hence the double Higgs production has several distinct final-state

topologies. The sub-channel with the largest cross section, $HH \to b\bar{b}b\bar{b}$, is studied by collaborators in the CERN. In this chapter, the $HH \to b\bar{b}W^+W^-$ sub-channel is investigated. Firstly hadronic decay of the W^+W^- in the $HH \to b\bar{b}W^+W^-$ channel is studied, because the hadronic decay has the largest cross section in the $HH \to b\bar{b}W^+W^-$ channel . The hadronic decay sub-channel does not have neutrinos in the final state, which allows each W to be reconstructed. The semi-leptonic final state of the W^+W^- system in the $HH \to b\bar{b}W^+W^-$ is also studied. The presence of the neutrino in the final state makes it difficult to reconstruct the two Higgs bosons, as some momenta of one Higgs boson is carried by the neutrino.

The process, $HH\nu_e\bar{\nu}_e \to b\bar{b}W^+W^-\nu_e\bar{\nu}_e \to b\bar{b}qqqq\nu_e\bar{\nu}_e$, results in 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 events. The two b quarks in the final state can be identified statistically with b jet tagging.

A fast simulation study of the double Higgs production has already been performed using the CLIC_ILD detector model for $\sqrt{s} = 1.4\,\text{TeV}$ and $3\,\text{TeV}$ [24]. Firstly, suitable signal and background channels are identified. In order to select the signal events, events with identified isolated lepton are vetoed. Vertex information is used to identify b quark jets. The PFOs in an event are then clustered into jets depending on the number of quarks in the final states. Afterwards, the jets are used as inputs for preselection and multivariate analysis. The event reconstruction was performed using the Marlin framework and reconstruction package in iLCSoft v01-16. More details on the reconstruction software can be found in chapter 4.

7.2. Monte Carlo sample generation

For this simulation study, the first step is to generate Monte Carlo samples. A full list of generated samples with their cross sections can be found in table 7.1. The software for Monte Carlo sample generation used is described in section 4.1.

At high centre-of-mass energies, in addition to considering electron-electron interactions, electron-photon and photon-photon interactions are important as their interactions become significant. These photons are produced due to the high electric field generated by the colliding beams. Processes involving real photons from beamsstrahlung (BS)

and "quasi-real" photons are generated separately. For the "quasi-real" photon initiated processes, the Equivalent Photon Approximation (EPA) has been used [74].

Background processes with multiple quarks and missing momentum in the final states are challenging to reject, as the topologies are similar to that of the signal events. Two example background processes are $e^+e^- \to qqqq\nu\overline{\nu}$ and $e^\pm\gamma \to \nu qqqq$. For the same reason, single Higgs boson production, such as $e^+e^- \to qqH\nu\overline{\nu}$, has a similar final state to the signal events and it is also difficult to reject.

Some processes are not considered in this analysis because they either have very different event topologies to the signal, or they have very small cross sections. For example, $e^\pm\gamma \to qqH\ell$ is neglected as the cross section is very small, even at $\sqrt{s} = 3\,\text{TeV}$.

The background processes are generated according to the final states. The final states of background processes could also happen via Higgs production, which would result in double counting of Higgs events with signal sample. Therefore to separate Higgs production from other processes, all background processes are generated with a Higgs boson mass of 14 TeV to ensure a negligible Higgs contribution. Processes involving Higgs production are simulated with a Higgs boson mass of 126 GeV.

The cross section of the signal, $HH \to b\overline{b}W^+W^-$, is scaled according to values listed in [75], as the default Higgs branching ratios in the generator software are less accurate than the values listed in [75].

The simulation and reconstruction chain is described in chapter 4. For some background processes, events are generated requiring the invariant mass of the total momenta of all quarks above 50 GeV or 120 GeV, because the invariant masses of the visible momenta of signal events are mostly above 150 GeV.

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

7.3. Lepton identification

For the signal channel, $HH \to b\overline{b}W^+W^- \to b\overline{b}qqqq$, there is no primary lepton in the final state, whilst many background processes, such as $qqqq\ell\nu$, contain primary leptons in final states. Hence, efficiently rejecting events with primary leptons is an important

Channel	$\sigma(\sqrt{s} = 1.4\,TeV)$ / 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
$e^\pm\gamma(\mathrm{BS}) \to e^\pm qqqq$	2317
$e^\pm\gamma(\mathrm{EPA}) \to e^\pm qqqq$	574
$e^\pm\gamma(\mathrm{BS}) \to \nu qqqq$	159.1†
$e^\pm\gamma(\mathrm{EPA}) \to \nu qqqq$	34.7†
$e^\pm\gamma(\mathrm{BS}) \to qqH\nu$	31.5*
$e^\pm\gamma(\mathrm{EPA}) \to qqH\nu$	6.78*
$\gamma(\mathrm{BS})\gamma(\mathrm{BS}) \to qqqq$	21406.2*
$\gamma(\mathrm{BS})\gamma(\mathrm{EPA}) \to qqqq$	4018.7*
$\gamma(\mathrm{EPA})\gamma(\mathrm{BS}) \to qqqq$	4034.8*
$\gamma(\mathrm{EPA})\gamma(\mathrm{EPA}) \to qqqq$	753.0*

Table 7.1.: List of signal and background samples used in the double Higgs analysis with the corresponding cross sections at $\sqrt{s} = 1.4\,\mathrm{TeV}$. q can be u, d, s, b or t. Unless specified, q, ℓ and ν represent either particles or the 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 labeled with * and †, events are generated with the invariant mass of the total momenta of all quarks above 50 and 120 GeV, respectively.

step in the event selection. Primary leptons deposit energies in the tracking detector. The start of the energy deposition is typically very close to the interaction point. The primary leptons often have energies above 10 GeV and are isolated from other particles. Whilst electrons and muons are stable enough to deposit energies in the calorimeters, tau leptons are very short lived. With a typical decay lifetime of 290 fs [70], it decays before reaching the vertex detector. Therefore, only decay products of the tau leptons can be reconstructed.

7.3.1. Electron and muon identification

Two approaches to electron and muon identification were utilised, which are described below. The performance is summarised in table 7.6.

IsolatedLeptonFinder

The IsolatedLeptonFinder reconstruction package is used and optimised. This processor identifies high energy electrons and muons that are isolated from other particles. The algorithm parameters wer optimised using the signal channel and the $e^+e^- \to qqqq\ell\nu$ channel, which is a multi-quark final state containing a primary lepton.

Optimal values of the IsolatedLeptonFinder are listed in table 7.2. E is the energy of the lepton. E_{ECAL} is the energy of the lepton deposited in the ECAL in GeV. E_{cone} is the total energy within a cone of an opening angle of $\cos^{-1}(0.995)$ around the lepton, in GeV. d_0, z_0, and r_0 are the Euclidean distance in mm of the lepton track starting point to the interaction point in x-y plane, in z direction, and in x-y-z three dimensional space, respectively.

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

Table 7.2.: Optimised parameters of IsolatedLeptonFinder

IsolatedLeptonIdentifer

A complimentary electron finder, IsolatedLeptonIdentifer, is developed to further identify isolated electrons and muons. Compared to the IsolatedLeptonFinder, the main difference is that the IsolatedLeptonIdentifer utilises particle ID information provided by the PandoraPFA to identify leptons.

The processor uses two sets of cuts to identify isolated leptons. If a PFO passes either set of cuts, it will be identified by the processor. The first set of cuts uses the particle ID information from PandoraPFA, demanding a PandoraPFA electron or muon with high p_{T} and the associated track depositing energy close to the interaction point. The identified lepton should either have a very high transverse momentum or be isolated. The second set of cuts uses ECAL energy fraction, which for a PFO is the energy deposited in the ECAL divided by the total energy, to determine the lepton ID. The rest of the cuts are very similar to the first cuts. The second set cuts have stricter isolation criterion to reduce fake rate.

Table 7.3 lists the selection cuts for IsolatedLeptonIdentifer. Same variables used in the IsolatedLeptonFinder and the IsolatedLeptonIdentifer are defined in the same way. p_{T} is the transverse momentum in GeV. 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 in GeV.

7.3.2. Tau lepton identification

Tau lepton has a short lifetime and decays before reaching the vertex detector. It can only be identified through the reconstruction of its decay products. The leptonic decay of tau lepton can be identified using the isolated lepton finder processors described above. Therefore in this section, tau identification will focus on the hadronic decays.

The existing TauFinder [76] reconstruction package has been optimised. In addition, a package, IsolatedTauIdentifer, is developed to provide additional tau lepton identification.

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

Table 7.3.: Optimised parameters of IsolatedLeptonIdentifer. A PFO needs to pass either set of cuts to be identified as a lepton. Within a set of cuts, the PFO needs to satisfy either condition a) or b) to be identified as a lepton.

TauFinder

The TauFinder works by identifying tau lepton decay products, and requiring the decay products to be isolated from other PFOs. To find the decay products, the algorithm starts with the highest energy track selected as a seed for the cone clustering algorithm (see section 4.4.4). All tracks above 10 GeV are iterated as seeds. A cone with opening angle 0.03 rad with respect to the seed is formed. The PFOs within the cone are required to be consistent with the signature of a tau hadronic decay. The signature includes no more than 3 charged particles in the cone, invariant mass less than 2 GeV, and few than 10 PFOs in the cone. The cone is also required to be isolated from other particles. To reduce fake rate, low momentum (less than 1 GeV) and very forward particles ($|\cos(\theta_Z)| > 1.1$) do not participate in the tau finding, as they more likely come from $\gamma\gamma \to$ hadrons background.

The optimised parameters are listed in table 7.4. Variables are defined in the same way as in previous sections. θ_Z is the polar angle with respect to the beam axis. N_{X^+} and N_{tau} are the number of charged particles and the number of PFOs respectively in the tau cone. m_{tau} is the invariant mass of the sum of the PFOs in the tau candidate in

GeV. E_{cone} is the total energy of PFOs within a cone of an opening angle between 0.03 and 0.33 rad around tau seed in GeV.

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

Table 7.4.: Optimised parameters of TauFinder.

IsolatedTauIdentifer

The IsolatedTauIdentifer works in a similar way as the TauFinder. It identifies high momentum particles as tau seeds. Particles are iteratively added to a cone in the order of the ascending opening angle to the seed. The cone is called search cone, which contains tau decay products. After each particle addition, the temporary search cone is then considered as a temporary tau candidate and tested for isolation and consistency with a tau hadronic decay signature. The temporary tau candidate only needs to pass one of the isolation conditions to be identified as a tau candidate. There are multiple isolation conditions for tau 1-prong decay and 3-prong decay, reflecting different topologies of tau decay final states. The isolation criterion typically demand few particles around the search cone and the total p_{T} in the search cone to be greater than a threshold.

The iterative particle addition procedure stops when the cone opening angle is larger than a threshold. If multiple temporary tau candidates of the same tau 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 $\gamma\gamma \to$ hadrons background, particles with energies less than 1 GeV do not participate.

Table 7.4 lists the optimised parameters for IsolatedTauIdentifer. Variables are defined in the same way as those in previous sections: θ_S is the opening angle of the search cone in rad; *cone*1 and *cone*2 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 the Euclidean distance in mm of the tau seed associated track starting point to the interaction point in x-y-z three dimensional space.

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

Table 7.5.: Optimised parameters of IsolatedTauIdentifer

Compared to the TauFinder algorithm, 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.

7.3.3. Very forward electron identification

At the high centre-of-mass energy at the CLIC, particles are often boosted. It is important to identify leptons in the forward calorimeters to aid the signal selection because certain background channels, for example photon-electron interactions, can have energetic primary electrons in the forward calorimeters, the LumiCAL and/or the BeamCAL.

It is challenging to identify electrons in these forward calorimeters. Most particles in these forward calorimeters are from the beam induced background. Nevertheless, sufficiently high energy electrons can be efficiently identified in the BeamCAL and LumiCAL [36].

The reconstruction of electrons in the forward calorimeters uses a parametrisation approach. By parameterising the particle ID efficiency using MC particles, the equivalent performance to the full detector simulation approach can be achieved [77, 78] .

For the BeamCAL, an existing electron tagging algorithm was developed using $\sqrt{s} = 3\,\text{TeV}$ collision environment by comparing the simulated electron and background energy distributions [77]. An indicative performance plot of 500 GeV electron tagging efficiency as a function of polar angle is shown in figure 7.2a. An electron is tagged if the energy is significantly larger than the expected background energy distributions integrated over 40 bunch crossing.

This parametric tagging approach most likely underestimates efficiencies due to the coarse binning of energies. For example, a 650 GeV particle is treated in the same way as a 600 GeV particle, as the tagging efficiency for electrons with energy 500 to 1500 GeV are binned in histograms at an interval of 100 GeV. Also there is no tagging efficiency for electrons with energy below 500 GeV or above 1500 GeV.

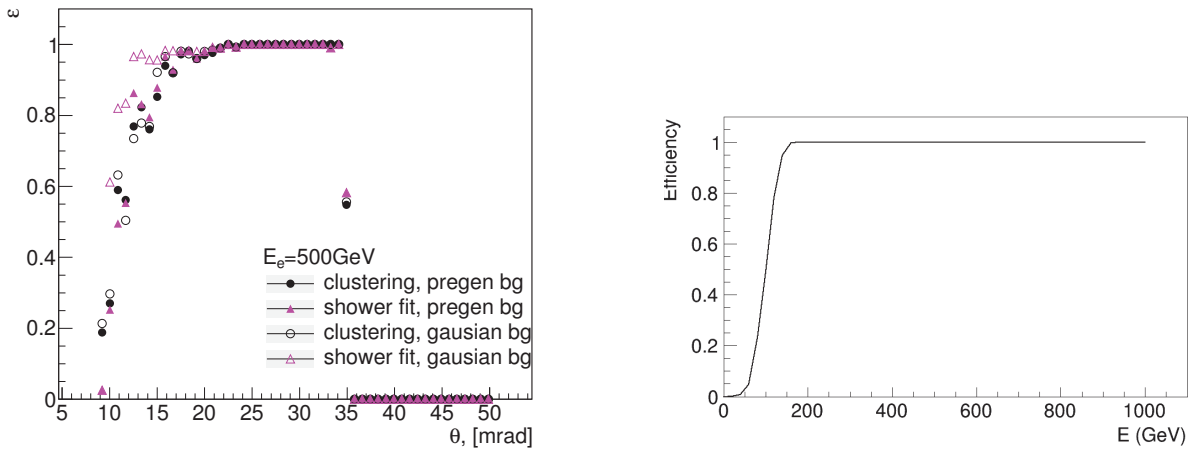
The input of the BeamCAL electron tagging algorithm is the four momenta of the MC electron. Since the algorithm assumes collisions at $\sqrt{s} = 3\,\text{TeV}$, for the $\sqrt{s} = 1.4\,\text{TeV}$ user case, the momenta of the MC electron is scaled down by a factor of $\frac{3}{1.4}$.

For the LumiCAL, 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 [78]. In this analysis, the LumiCAL electron tagging efficiency, ε, is parameterised as

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

where E is the energy of the electron 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 can not be differentiated as they have very similar EM shower profiles. Therefore, both photons and electrons are tagged by the algorithms above.



(a) **(b)**

Figure 7.2.: Figure a) shows 500 GeV electron tagging efficiency in the BeamCAL as a function of polar angle, with different methods to model backgrounds: pregenerated and Gaussian, and two methods to identify electrons: clustering algorithm and shower fitting algorithm, taken from [77]. Figure b) shows the electron tagging efficiency in the LumiCAL as a function of the electron energy, for polar angle $\theta = 50\,\text{mrad}$, taken from [78].

7.3.4. Lepton identification performance

The performances of the different lepton finding processors for signal events and the selected background processes are shown in table 7.6. Numbers in the table represent the fractions of events where no leptons are identified by the individual lepton finder. IsolatedLeptonIdentifer and IsolatedTauIdentifer reject more background events than the IsolatedLeptonFinder and TauFinder. By combining the processors, 86.6% of the signal events remain and 16.8% of the $e^+e^- \to qqqq\ell\nu$ events survive after rejecting events where leptons are identified.

The forward lepton finders are most effective at rejecting background events with primary leptons in the forward region. Table 7.6 shows the performance of the processors with the signal events and the $e^-\gamma(BS) \to e^-qqqq$ background events. 53.6% of the $e^-\gamma(BS) \to e^-qqqq$ background events survive after rejecting events with leptons identified by the forward lepton finder. The full list of fraction of events after rejecting events with leptons identified by any lepton finders for individual channel can be found in table 7.9.

The lepton finding processors are optimised with events at $\sqrt{s} = 1.4\,\text{TeV}$, and checked with events at $\sqrt{s} = 3\,\text{TeV}$. It was found that the same set of parameters for lepton identifiers works well under $\sqrt{s} = 1.4\,\text{TeV}$ and $3\,\text{TeV}$. The performances of the lepton

Efficiency (1.4 TeV)	Signal	$e^+e^- \to qqqq\ell\nu$	$e^-\gamma(\text{BS}) \to e^-qqqq$
IsolatedLeptonFinder	99.3%	50.3%	87.3%
IsolatedLeptonIdentifer	99.1%	39.9%	83.7%
TauFinder	97.5%	52.3%	90.4%
IsolatedTauIdentifer	89.7%	38.5%	78.5%
Forward Finder Processors	98.9%	95.1%	53.6%
Combined	86.6%	16.8%	30.8%

Table 7.6.: The performances of lepton finders on the signal events and selected background events at $\sqrt{s} = 1.4\,\text{TeV}$. Numbers represent the fractions of events where no leptons are identified by the individual lepton finder. γ (BS) represents a real photon from beamstrahlung (BS).

Efficiency (3 TeV)	Signal	$e^+e^- \to qqqq\ell\nu$	$e^-\gamma(\text{BS}) \to e^-qqqq$
IsolatedLeptonFinder	99.5%	66.8%	88.8%
IsolatedLeptonIdentifer	99.0%	52.5%	82.2%
TauFinder	97.7%	79.5%	76.7%
IsolatedTauIdentifer	86.3%	60.3%	92.6%
Forward Finder Processors	95.9%	80.7%	55.4%
Combined	81.0%	23.3%	33.4%

Table 7.7.: The performances of lepton finders on the signal events and selected background events at $\sqrt{s} = 3\,\text{TeV}$. Numbers represent the fractions of events where no leptons are identified by the individual lepton finder. γ (BS) represents a real photon from beamstrahlung (BS).

finders with the signal and selected background events at $\sqrt{s} = 3\,\text{TeV}$ are shown in table 7.7.

When comparing the lepton finding performances at $\sqrt{s} = 1.4\,\text{TeV}$ and $\sqrt{s} = 3\,\text{TeV}$, the performance for $\sqrt{s} = 1.4\,\text{TeV}$ is better. This is because at $\sqrt{s} = 3\,\text{TeV}$, particles are often boosted and spatial separation between particles is smaller. Therefore particles are less isolated from each other. The higher centre-of-mass energy at $\sqrt{s} = 3\,\text{TeV}$ also affects on the performance of the forward lepton finder. Whilst at $\sqrt{s} = 1.4\,\text{TeV}$, the forward finder only rejects 5% of the $e^+e^- \to qqqq\ell\nu$ background events and 1% of the signal events, at $\sqrt{s} = 3\,\text{TeV}$ it rejects 19% of events from the same background process and 4% of the signal events, as more leptons are boosted into the forward region.

7.4. Jet reconstruction

Another important aspect of this analysis is to pair jets to reconstruct the physical bosons. The signal channel, $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$, results in a six quark final state. In this section, the optimisation of the jet reconstruction is discussed.

7.4.1. Jet reconstruction optimisation

Jet reconstruction algorithms cluster particles into jets. For this analysis, longitudinal invariant, k_t, jet algorithm is chosen for the jet clustering, as discussed in section 4.6.2. The free parameter for k_t algorithm is the R parameter, which controls the radius of the jet. Another choice affecting the jet reconstruction is the choice of the PFO collection, which incorporates different level of timing and p_T cuts to reduce the beam induce background (see section 4.5.2).

The value of the R parameter and the PFO collection are chosen to optimise the invariant mass and mass resolution of H and W. To choose the optimal parameters, $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$ events are processed through k_t jet algorithm in the 6-jet exclusive mode. The six jets are paired using the MC truth information (see section 4.6.1) by examining the decay chain of MC particles. Four invariant mass distributions are obtained: two Higgs masses ($m_{H_{bb}}$ and $m_{H_{WW^*}}$) and two W masses (m_W and m_{W^*}). Here W^* indicates the off-mass-shell W boson.

Three invariant mass distributions are considered: $m_{H_{bb}}$, $m_{H_{WW^*}}$, and m_W. The optimal jet reconstruction should produce a sharp mass peak around the simulated particle masses (see section 4.5.3). For example, figure 7.3 shows the $m_{H_{bb}}$ invariant mass distribution for $R = 1.3$ using the loose PFO collection for samples at $\sqrt{s} = 3\,\text{TeV}$. An analytical functional form is fitted to quantitatively describe the shape. The fitting function is a Gaussian-like function. Free parameters are used in the fitting function to fit the tail distribution. The fitting function takes the form of

$$f(m) = Ae^{-\frac{(m-\mu)^2}{g}}, \tag{7.2}$$

$$g = \begin{cases} 2\sigma_L + \alpha_L(m-\mu), & \text{if } m < \mu, \\ 2\sigma_R + \alpha_R(m-\mu), & \text{if } m \geqslant \mu, \end{cases} \tag{7.3}$$

where: μ is the fitted mass peak position; σ_L and σ_R allow an asymmetrical width of the distribution; α_L and α_R fit the tail of the distribution; and A is the normalisation factor.

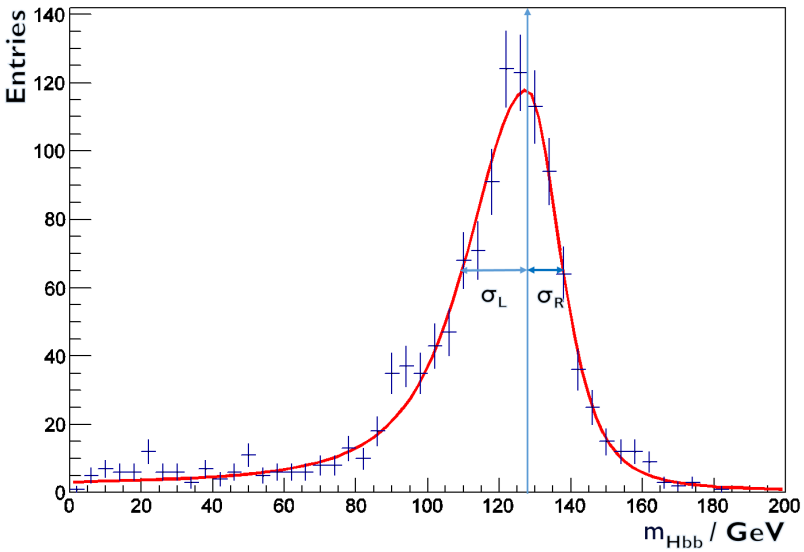


Figure 7.3.: A typical example of $m_{\mathrm{H_{bb}}}$ mass distribution for $R = 1.3$ using loose PFO collection for signal samples at $\sqrt{s} = 3\,\mathrm{TeV}$. Fitting function is superimposed in red. The arrow shows the fitted mean peak position.

The overall relative width is defined as $(\sigma_L + \sigma_R)/M$. A smaller width indicates a better mass resolution. The fitted $\mathrm{H_{bb}}$, $\mathrm{H_{WW^*}}$, and W masses are studied for R values between 0.5 and 1.3, at an interval of 0.1, and with three PFO collections: loose, normal, and tight.

Figure 7.4 shows the variation of the mass peak position and the relative width as a function of R and PFO collections, for $m_{\mathrm{H_{bb}}}$, $m_{\mathrm{H_{WW^*}}}$, and m_{W}. The mass peak position, μ, increases as R increases. This is because more particles are included in jets with increasing jet radius. For the relative width, the values for $\mathrm{H_{bb}}$ increase with increasing jet radius, but the values for $\mathrm{H_{WW^*}}$ decrease with increasing jet radius. This is due to a compensating effect. The invariant mass for $\mathrm{H_{WW^*}}$ is formed from four jets, which prefers a large jet radius. The invariant mass for $\mathrm{H_{bb}}$ is obtained from two jets, which favours a small jet radius.

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.

Based on the results summarised in figure 7.4, it was decided to use $R = 0.7$ with the selected PFO collection for this analysis. This choice gives good fitted mass peak positions for H_{bb}, H_{WW^*} and W. The extracted fitted parameters of optimal jet reconstructions are summarised in table 7.8.

Fitted jet parameters	$\sqrt{s} = 1.4\,\text{TeV}$	$\sqrt{s} = 3\,\text{TeV}$
$\mu_{H_{bb}}$	$122.3_{\pm 0.2}$	$119.1_{\pm 0.3}$
$\sigma_{L,H_{bb}}$	$15.2_{\pm 0.2}$	$15.0_{\pm 0.3}$
$\sigma_{R,H_{bb}}$	$7.55_{\pm 0.16}$	$8.4_{\pm 0.2}$
$\mu_{H_{WW^*}}$	$125.7_{\pm 0.2}$	$123.0_{\pm 0.3}$
$\sigma_{L,H_{WW^*}}$	$29.4_{\pm 0.3}$	$36.6_{\pm 0.6}$
$\sigma_{R,H_{WW^*}}$	$7.18_{\pm 0.17}$	$7.4_{\pm 0.2}$
μ_{W}	$80.5_{\pm 0.2}$	$78.1_{\pm 0.3}$
$\sigma_{L,W}$	$16.2_{\pm 0.3}$	$13.1_{\pm 0.4}$
$\sigma_{R,W}$	$9.03_{\pm 0.16}$	$9.5_{\pm 0.2}$

Table 7.8.: The fitted mass parameters for $\sqrt{s} = 1.4\,\text{TeV}$ analysis: $R = 0.7$ using the selected PFO collection, and for $\sqrt{s} = 3\,\text{TeV}$ analysis: $R = 0.7$ using the tight selected PFO collection.

For the jet reconstruction optimisation at $\sqrt{s} = 3\,\text{TeV}$, Figure 7.5 shows the variation of fitted mass peak positions and the relative mass resolutions for H_{bb}, H_{WW^*}, and W as function of R and PFO collections. The relative mass resolution of W worsen quickly with increasing R. The fitted mass peak positions also increases quicker with the increase of R, compared with the fitted positions at $\sqrt{s} = 1.4\,\text{TeV}$. This is due to at a higher centre-of-mass energy, more beam induced background particles are produced. The background particles, if included in the jets, will increase the invariant masses of the fitted physical bosons. Therefore, $R = 0.7$ with the tight selected PFO collection is chosen for the $\sqrt{s} = 3\,\text{TeV}$ analysis. With chosen parameters, the excellent mass resolutions compensate for the invariant masses slightly smaller than simulated values. The extracted fitted parameters of optimal jet reconstructions at $\sqrt{s} = 3\,\text{TeV}$ are summarised in table 7.8.

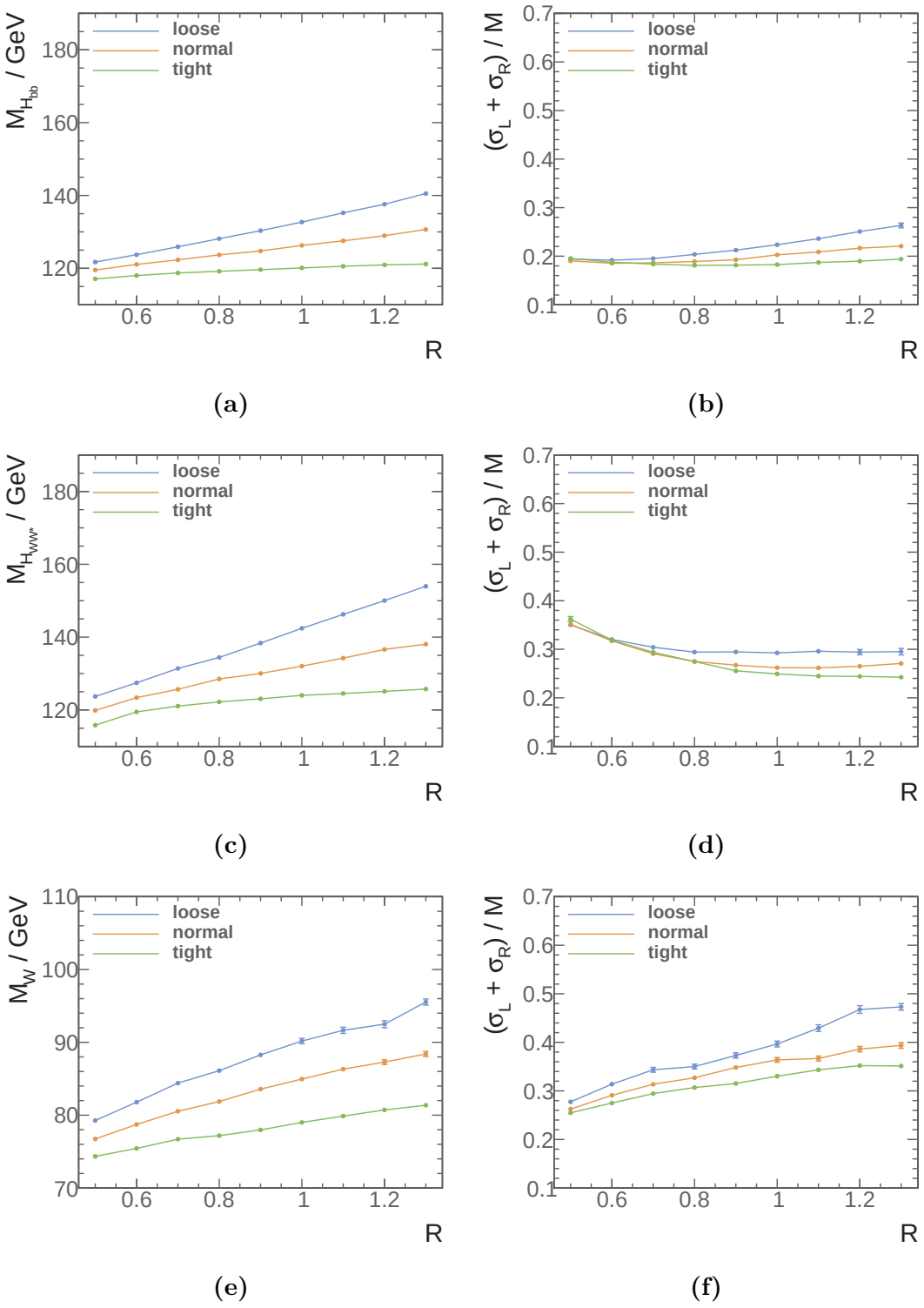


Figure 7.4.: Distributions of a) fitted mass peak positions of H_{bb}, b) relative mass peak width of H_{bb}, c) fitted mass peak positions of H_{WW^*}, and f) relative mass peak width of H_{WW^*}, e) fitted mass peak positions of W, b) relative mass peak width of W. All plots show the variation of the fitted masses and mass resolutions as a function of R for loose, normal and tight selected PFO collections at $\sqrt{s} = 1.4\,\text{TeV}$.

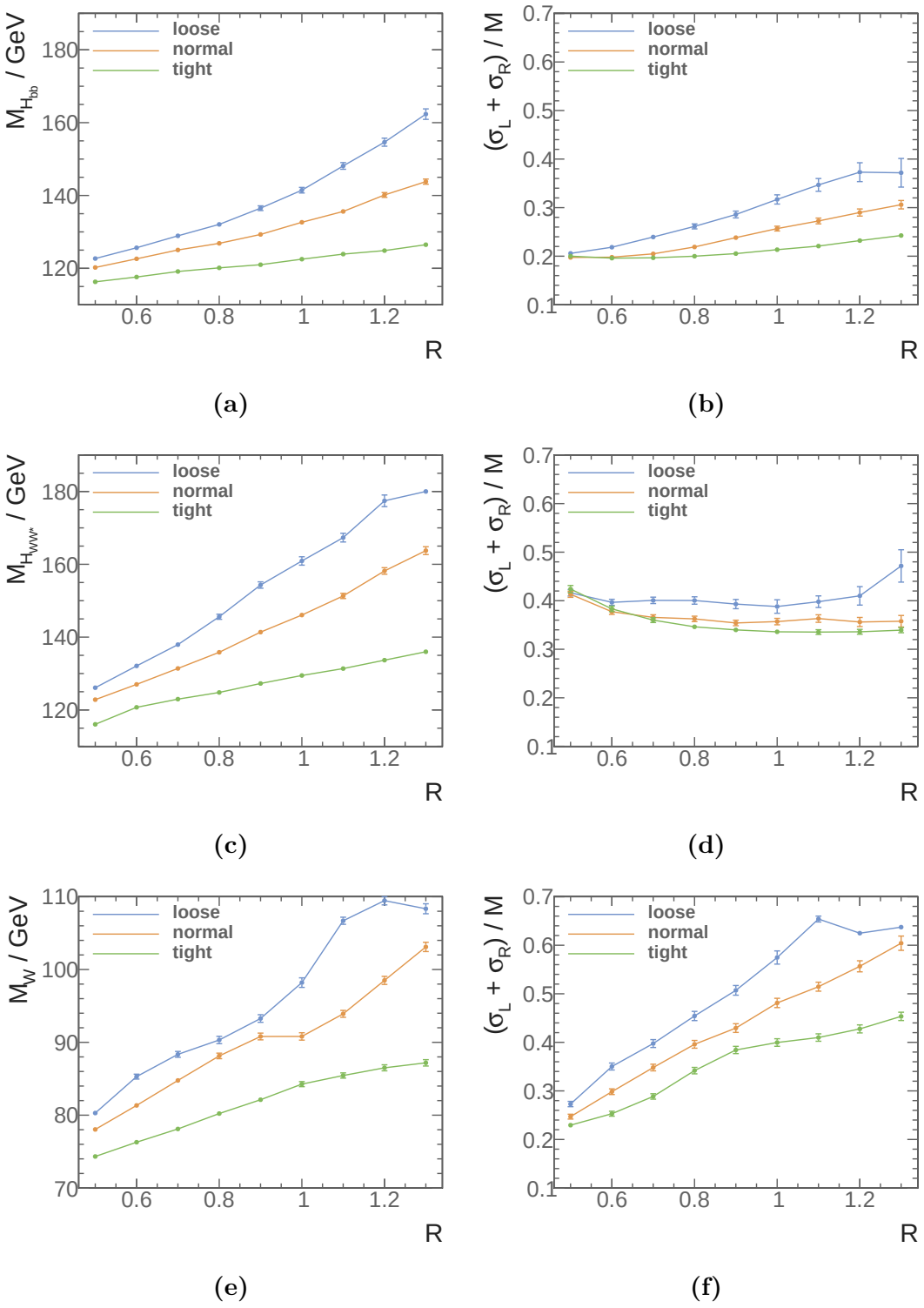


Figure 7.5.: Distributions of a) fitted mass peak positions of H_{bb}, b) relative mass peak widths of H_{bb}, c) fitted mass peak positions of H_{WW^*}, d) relative mass peak widths of H_{WW^*}, e) fitted mass peak positions of W, and f) relative mass peak widths of W. All plots show the variation of fitted masses and mass resolutions as a function of R for loose, normal and tight selected PFO collections at $\sqrt{s} = 3\,\text{TeV}$.

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, identifying jets originated from b quarks is an important part of the event selection. To establish the likelihood of a b jet (b-jet tag value), LCFIPlus [79] software package is used.

The flavour tagging processor, LCFIPlus [79] is based on the LCFIVertex package [80], which was used in the simulation studies for the ILC Letter of Intent [32, 81] and the CLIC Concept Design Report [24]. LCFIPlus is modular and can be used in any order. However here it will be described in the order that is used in this analysis.

The inputs are PFOs, after jet clustering using $R = 0.7$ and the selected PFO collection. The vertex finding algorithms in the LCFIPlus perform vertex fitting and identify primary and secondary vertices. There is a "V0" particle rejection step. A "V0" particle is a neutral particle that decays into pairs of charged particles. The topology of the "V0" particles can be similar to the decay of b or c quarks. Hence it is important to remove the "V0' particles to improve the b-quark and c-quark flavour tagging.

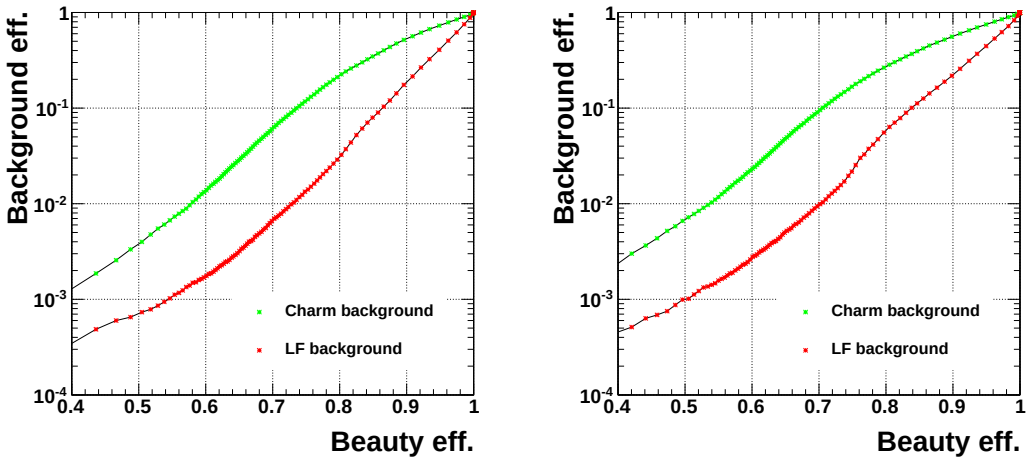
Once the primary and secondary vertices are found, PFOs are re-clustered into jets. This jet re-clustering scheme ensures that the secondary vertices and the muons which is identified from semi-leptonic decay of the quarks, fall into the same jet. Therefore, the topology of the jet is consistent with the hadronic decay of heavy quarks. The jet re-clustering scheme are modified Durham algorithms.

The next step is to refine vertices to improve the b jet sparation from the c jet. Additional information is applied to improve the vertex reconstruction. 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 the TMVA software package [61]. A number of flavour sensitive variables are calculated. The jets are is divided into four subsets: jets with zero, one, 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 subsets.

The events to train the multiclass classifier are $e^+e^- \to Z\bar{\nu}\nu$ event 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 selection efficiency of b jets and c jets with the training samples is shown in figure 7.6b. The normalised distribution of the highest b-jet tag value for the signal events is shown in figure 7.7.

For the $\sqrt{s} = 3\,\text{TeV}$, the flavour tagging processor is re-trained in the same way as in the $\sqrt{s} = 1.4\,\text{TeV}$ analysis. The training events are from the same channel but at $\sqrt{s} = 3\,\text{TeV}$. The performance of the flavour tagging with training samples is shown in figure 7.6a. Compared to the performance at $\sqrt{s} = 1.4\,\text{TeV}$, the performance is slightly worse, because at a high centre-of-mass energy, particles at more collimated and more difficult to separate. Therefore, the vertex identification and the flavour tagging performance are worse.



(a) $\sqrt{s} = 1.4\,\text{TeV}$ **(b)** $\sqrt{s} = 3\,\text{TeV}$

Figure 7.6.: Performance of b-jet tagging with training samples at a) $\sqrt{s} = 1.4\,\text{TeV}$, and b) $\sqrt{s} = 3\,\text{TeV}$.

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

Two $e^+e^- \to HH\nu_e\bar{\nu}_e$ final states with largest branching fractions are $HH \to b\bar{b}b\bar{b}$ (31.5%) and $HH \to b\bar{b}W^+W^-$ (25.9%). Two final states have different topologies and are subject of two analysis strategies. The $HH \to b\bar{b}W^+W^-$ final state is the subject of this thesis. The study of the $HH \to b\bar{b}b\bar{b}$ final state is the subject of an independent analysis. A set of cuts are designed to separate samples, for both signal and background events, into

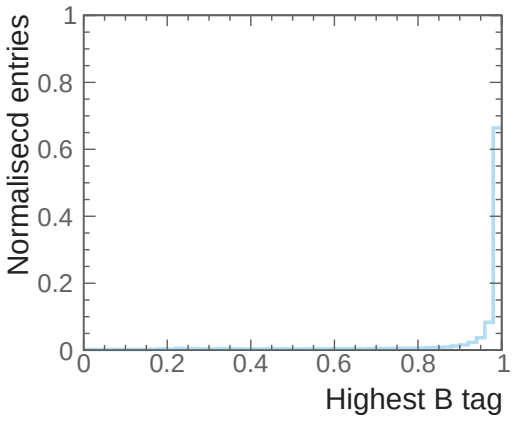


Figure 7.7.: The normalised distribution of the highest b-jet tag value for the signal events at $\sqrt{s} = 1.4\,\text{TeV}$

two mutually exclusive sets for two independent analyses. This ensures there are no correlations between two analyses.

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

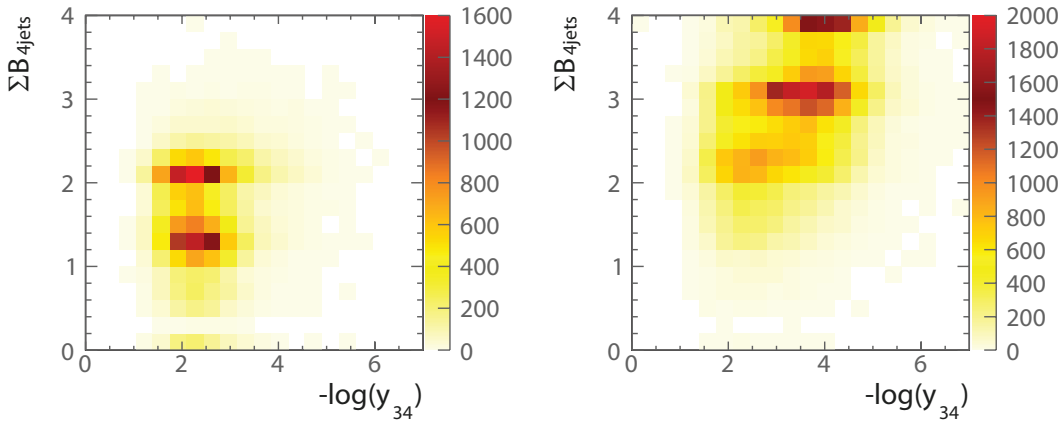
Figure 7.8 shows the sum of b-jet tag values, when the event is clustered into four jets, as a function of $-\log(y_{34})$ for the hadronic W^+W^- decay in $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$. y parameter is a measure of number of jets. As expected, two sub-channels can be clearly separated in this two dimensional phase space. A rectangular cut will separate the phase space into two spaces, donated as S and $\neg S$. The hadronic W^+W^- decay in $HH \to b\bar{b}W^+W^-$ events should be contained in S, and the $HH \to b\bar{b}b\bar{b}$ events should be contained in $\neg S$.

To maximise the separation of two sub-channels, a set of cuts are found by maximising the product of the fraction of the sub-channel events in each space:

$$\varepsilon = \frac{N_{HH \to b\bar{b}W^+W^-,\,\text{hadronic}} \in S}{N_{HH \to b\bar{b}W^+W^-,\,\text{hadronic}}} \times \frac{N_{HH \to b\bar{b}b\bar{b}} \in \neg S}{N_{HH \to b\bar{b}b\bar{b}}}, \tag{7.4}$$

where $N \in S$ indicates number of events in the phase space S.

Several variables are tested to maximise ε. The best cuts found defining S is $\Sigma B_{4jets} < 2.3$, $-\log(y_{34}) < 3.7$. With the cuts, 86% of the hadronic W^+W^- decay in $HH \to b\bar{b}W^+W^-$ events are in S and 78% of the $HH \to b\bar{b}b\bar{b}$ events are in $\neg S$. The



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

Figure 7.8.: Sum of b-jet tag values, when the event is clustered into four jets, as a function of $-\log(y_{34})$ at $\sqrt{s} = 1.4\,\mathrm{TeV}$, shown for a) hadronic $\mathrm{W^+W^-}$ decay of $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$, and b) $\mathrm{HH} \to \mathrm{b\bar{b}b\bar{b}}$ sub-channel.

full list of fraction of events after passing mutually exclusive cuts for individual background channel can be found in table 7.9.

7.6. Jet pairing

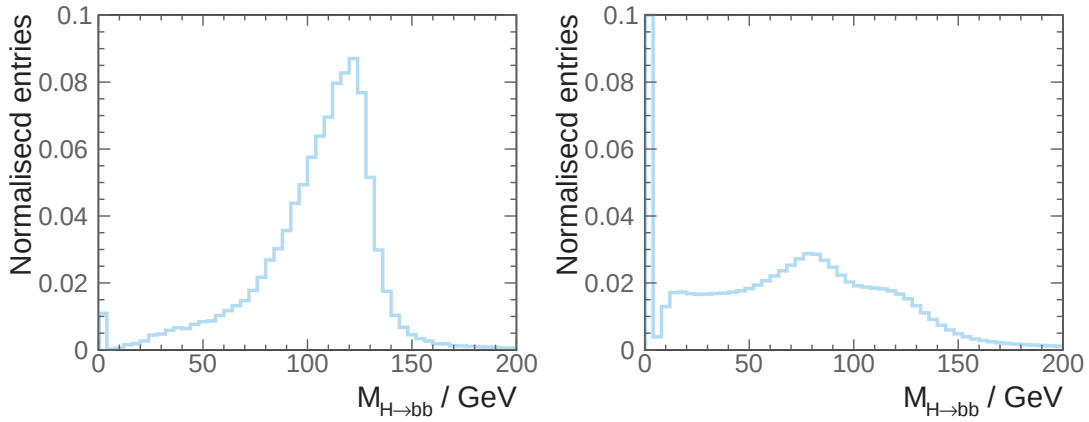
Having optimised the jet reconstruction, and obtained the six jets from the jet reclustering step in the LCFIPlus processor, the next step is to group jets according to signal event topology. Jets are paired up such that there are two jets for $\mathrm{H} \to \mathrm{b\bar{b}}$, two jets for hadronic decay of a W, two jets for hadronic decay of a W*, and the two Ws forming a H boson.

Six jets are associated to $\mathrm{H_{bb}}$, W and W*. There are 90 possible permutations. The best permutation is obtained by minimising a χ^2 based on expected masses and mass widthes:

$$\chi^2 = \left(\frac{m_{ij} - \mu_{\mathrm{H_{bb}}}}{\sigma'_{\mathrm{H_{bb}}}}\right)^2 + \left(\frac{m_{klmn} - \mu_{\mathrm{H_{WW^*}}}}{\sigma'_{\mathrm{H_{WW^*}}}}\right)^2 + \left(\frac{m_{kl} - \mu_{\mathrm{W}}}{\sigma'_{\mathrm{W}}}\right)^2, \tag{7.5}$$

$$\sigma'_{\mathrm{H_{bb}}} = \begin{cases} \sigma_{L,\mathrm{H_{bb}}}, & \text{if } m_{ij} < \mu_{\mathrm{H_{bb}}}, \text{ etc.} \\ \sigma_{R,\mathrm{H_{bb}}}, & \text{otherwise,} \end{cases} \tag{7.6}$$

where the indices i to l indicate the one of the six jets; 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 the definition of $\sigma'_{\mathrm{H_{bb}}}$, $\sigma'_{\mathrm{H_{WW^*}}}$, and σ'_{W}. The jet pairing is only considered when at least one of the jets associated to the $\mathrm{H_{bb}}$ decay with a b-jet tag $>$ 0.2. Of these combinations of jets, the jet pairing giving smallest χ^2 is selected. The normalised distribution of $m_{\mathrm{H_{bb}}}$ after jet pairing, for the signal channel, $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$ and the sum of all background channels can be seen in figure 7.9. For the signal channel, the distribution peaks around the expected mass of $m_{\mathrm{H_{bb}}}$. It can also be seen from the figures that around 1% of the signal events have no solutions for the jet pairing, as no jet has a b-jet tag $>$ 0.2. The full list of fraction of events after jet pairing selection can be found in table 7.9.



(a) Signal **(b)** Background

Figure 7.9.: The normalised distribution of $m_{\mathrm{H_{bb}}}$ after jet pairing, for a) signal channel, hadronic $\mathrm{W^+W^-}$ decay of $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$, b) sum of all background channels. All plots are shown for $\sqrt{s} = 1.4\,\mathrm{TeV}$.

Channel / Efficiency $\sqrt{s}$ = 1.4 TeV	Expected number of events	Lepton evto	Mutually exclusive	Jet Pairing
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$				
$b\overline{b}W^+W^-\nu_e\overline{\nu}_e$, hadronic	27.9	89.7%	79.1%	78.3%
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$				
$b\overline{b}b\overline{b}\nu_e\overline{\nu}_e$	67.6	90.8%	18.0%	18.0%
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$ other	128.0	40.8%	35.8%	31.2%
$e^+e^- \to q_\ell q_\ell H\nu\overline{\nu}$	1290	72.8%	69.7%	57.7%
$e^+e^- \to c\overline{c}H\nu\overline{\nu}$	540	74.7%	59.8%	52.7%
$e^+e^- \to b\overline{b}H\nu\overline{\nu}$	465	74.3%	32.2%	31.8%
$e^+e^- \to qqqq$	1867650	79.9%	64.0%	38.6%
$e^+e^- \to qqqq\ell\ell$	93150	8.9%	8.2%	4.7%
$e^+e^- \to qqqq\ell\nu$	165600	16.5%	14.6%	13.3%
$e^+e^- \to qqqq\nu\overline{\nu}$	34800	87.6%	82.0%	46.8%
$e^+e^- \to qq$	6014250	81.0%	57.8%	39.0%
$e^+e^- \to qq\ell\nu$	6464550	22.5%	17.0%	10.5%
$e^+e^- \to qq\ell\ell$	4088700	19.4%	18.6%	12.4%
$e^+e^- \to qq\nu\nu$	1181550	91.8%	74.0%	47.3%
$e^\pm\gamma(\mathrm{BS}) \to e^\pm qqqq$	2606625	34.2%	33.5%	22.9%
$e^\pm\gamma(\mathrm{EPA}) \to e^\pm qqqq$	861000.0	16.4%	15.8%	10.7%
$e^\pm\gamma(\mathrm{BS}) \to \nu qqqq$	178987.5	85.6%	81.3%	54.4%
$e^\pm\gamma(\mathrm{EPA}) \to \nu qqqq$	52050	44.5%	42.0%	27.4%
$e^\pm\gamma(\mathrm{BS}) \to qqH\nu$	35437.5	70.7%	65.0%	55.4%
$e^\pm\gamma(\mathrm{EPA}) \to qqH\nu$	10170	37.0%	33.8%	28.8%
$\gamma(\mathrm{BS})\gamma(\mathrm{BS}) \to qqqq$	2054951.5	85.6%	81.3%	54.0%
$\gamma(\mathrm{BS})\gamma(\mathrm{EPA}) \to qqqq$	4521037.5	49.6%	48.5%	32.9%
$\gamma(\mathrm{EPA})\gamma(\mathrm{BS}) \to qqqq$	4539150	49.6%	48.5%	32.9%
$\gamma(\mathrm{EPA})\gamma(\mathrm{EPA}) \to qqqq$	1129500	31.0%	30.1%	20.5%

Table 7.9.: Number of events and fraction of events passing lepton veto, the mutually exclusive cuts, and the jet pairing for the signal and background events at $\sqrt{s} = 1.4\,\mathrm{TeV}$, assuming an integrated luminosity of $1500\,fb^{-1}$. The selection efficiencies are presented in a "flow" fashion. Every selection cut contains all the cuts to the left of it. q can be u, d, s, b or t. Unless specified, q, ℓ and ν represent either particles or the 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.

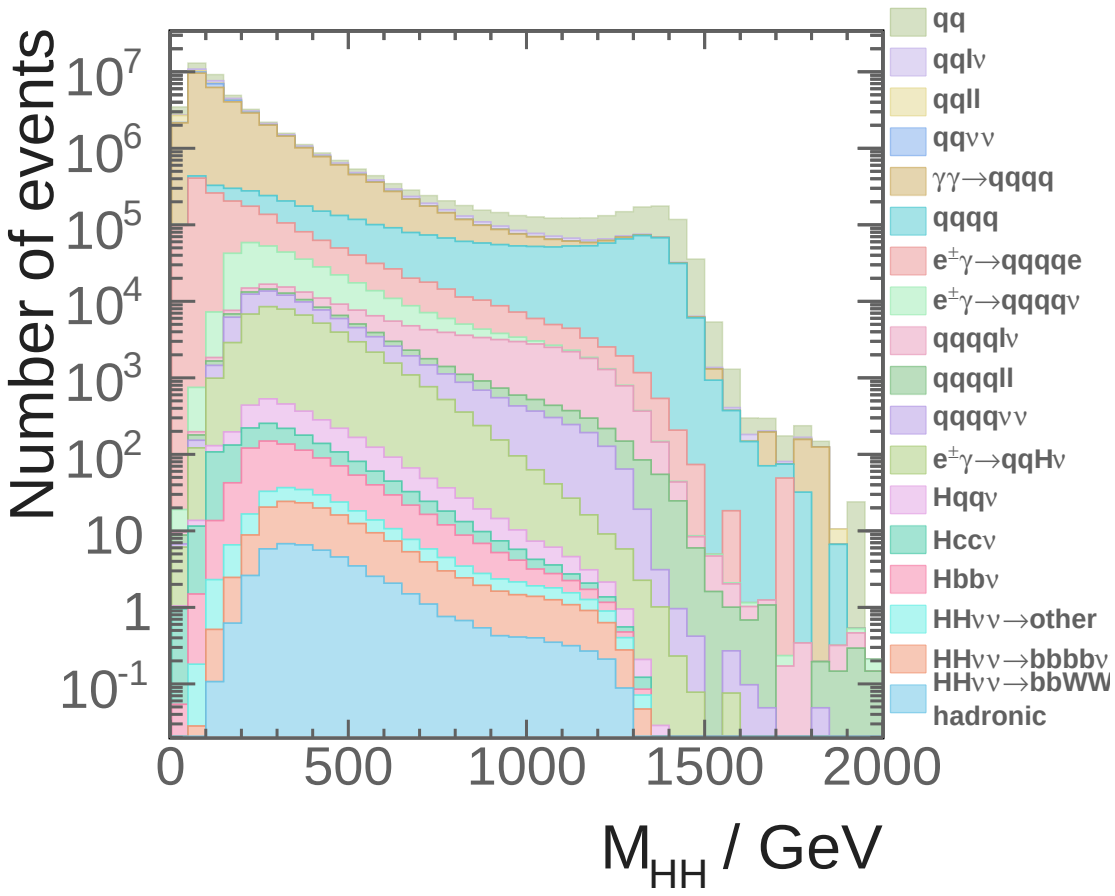


Figure 7.10.: Distributions of the invariant mass of the two Higgs system for $\sqrt{s} = 1.4\,\text{TeV}$, assuming an intergraded luminosity of 1500 fb^{-1}.

7.7. Pre-selection

After the association of jets to candidate bosons are made under hypothesis that an event is signal, kinematic and topological variables can be calculated. A set of pre-selection cuts are placed to discard the phase space dominated by background events. Cuts on p_{T}, b-jet tag, and invariant mass of the double Higgs system are used.

Since both Higgs bosons are on mass shell, the invariant mass of the double Higgs system is large. Consequently, a cut on $m_{\text{HH}} > 150\,\text{GeV}$, as shown in Figure 7.10, removes a small amount of signal events, but discard lots of background events, especially $\gamma\gamma \to qqqq$ events.

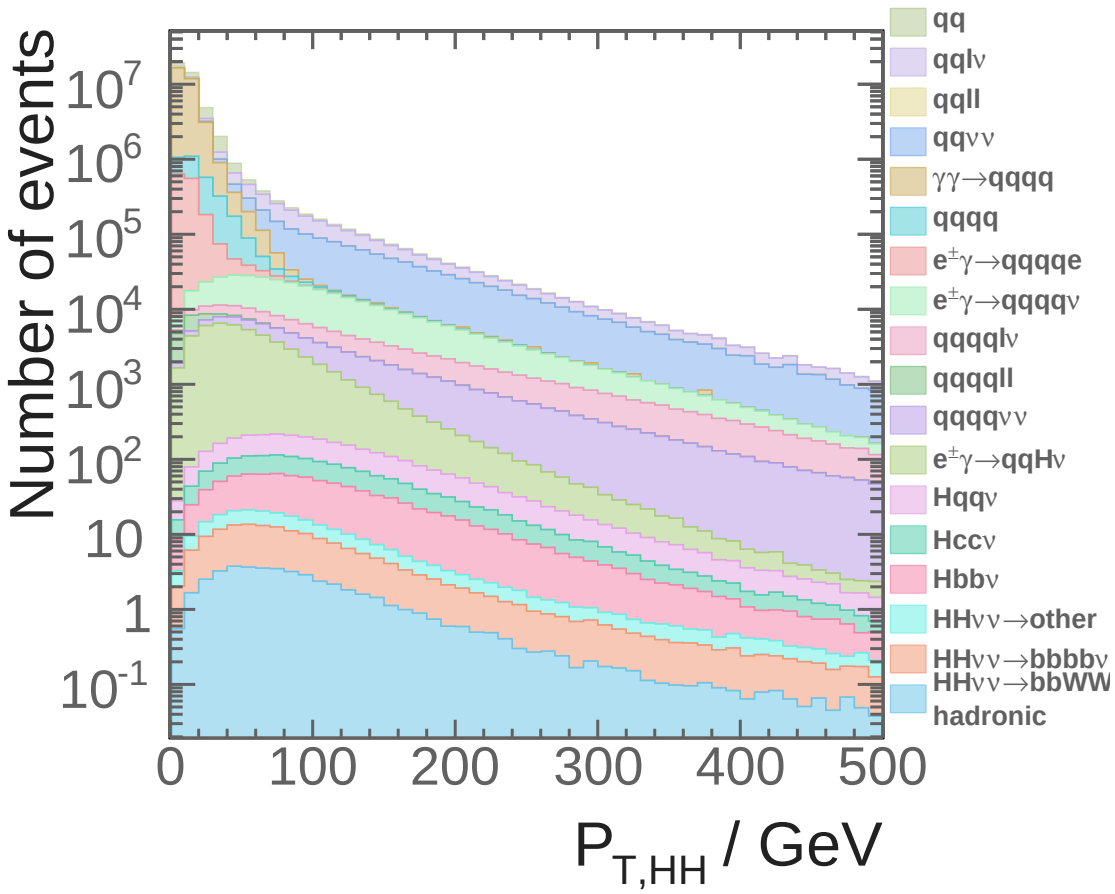


Figure 7.11.: Distributions of the second highest b-jet tag value for $\sqrt{s} = 1.4\,\text{TeV}$, assuming an intergraded luminosity of 1500 fb^{-1}.

Many background events do not have b-quark jets in the final state. Therefore, by requiring the second highest b-jet tag value greater than 0.2, as shown in Figure 7.11, background events with no b-jets in final states are removed.

The signal final states have neutrinos and hence missing momentum in the events. Therefore, the transverse momentum of the two Higgs system is non zero. A cut of $p_{\text{T}} > 30\,\text{GeV}$, as shown in figure 7.12, is extremely effective against background channels with no neutrinos in the final state.

The full list of fraction of events after each pre-selection cut can be found in table 7.10.

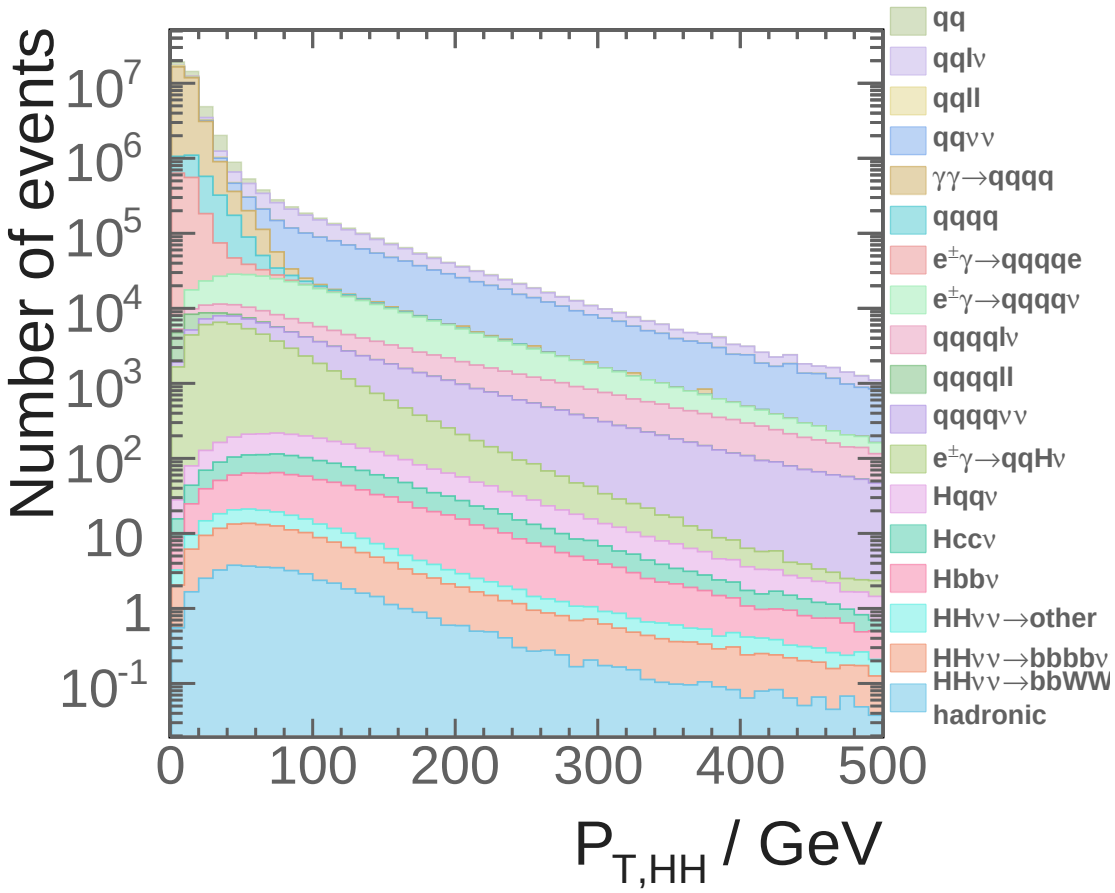


Figure 7.12.: Distributions of the transverse momentum of the two Higgs system for $\sqrt{s} = 1.4\,\text{TeV}$, assuming an intergraded luminosity of 1500 fb^{-1}.

Channel	m_{HH}>150GeV	B_2>0.2	p_{T} >30GeV
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$			
$b\overline{b}W^+W^-\nu_e\overline{\nu}_e$, hadronic	78.1%	66.3%	59.7%
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$			
$b\overline{b}b\overline{b}\nu_e\overline{\nu}_e$	17.8%	17.4%	15.4%
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$ other	30.5%	23.0%	20.5%
$e^+e^- \to q_l q_l H\nu\overline{\nu}$	56.8%	42.3%	39.5%
$e^+e^- \to c\overline{c}H\nu\overline{\nu}$	44.8%	34.1%	31.7%
$e^+e^- \to b\overline{b}H\nu\overline{\nu}$	30.7%	27.0%	25.2%
$e^+e^- \to qqqq$	36.1%	13.2%	3.4%
$e^+e^- \to qqqq\ell\ell$	4.7%	1.5%	0.3%
$e^+e^- \to qqqq\ell\nu$	13.2%.	10.7%	9.8%
$e^+e^- \to qqqq\nu\overline{\nu}$	46.1%	17.7%	16.6%
$e^+e^- \to qq$	8.1%	3.7%	0.8%
$e^+e^- \to qq\ell\nu$	3.1%	1.2%	0.9%
$e^+e^- \to qq\ell\ell$	0.7%	0.4%	0.1%
$e^+e^- \to qq\nu\nu$	9%	4.3%	4.0%
$e^\pm\gamma(\mathrm{BS}) \to e^\pm qqqq$	10.1%	4.1%	0.4%
$e^\pm\gamma(\mathrm{EPA}) \to e^\pm qqqq$	5.1%	2.0%	0.3%
$e^\pm\gamma(\mathrm{BS}) \to \nu qqqq$	53.0%	28.0%	25.1%
$e^-\gamma(\mathrm{EPA}) \to \nu qqqq$	26.7%	13.8%	12.5%
$e^\pm\gamma(\mathrm{BS}) \to qqH\nu$	54.3%	40.3%	30.6%
$e^-\gamma(\mathrm{EPA}) \to qqH\nu$	28.2%	20.9%	16.1%
$\gamma(\mathrm{BS})\gamma(\mathrm{BS}) \to qqqq$	23.1%	9.2%	0.3%
$\gamma(\mathrm{BS})\gamma(\mathrm{EPA}) \to qqqq$	13.6%	5.4%	0.4%
$\gamma(\mathrm{EPA})\gamma(\mathrm{BS}) \to qqqq$	13.6%	5.4%	0.3%
$\gamma(\mathrm{EPA})\gamma(\mathrm{EPA}) \to qqqq$	8.6%	3.5%	0.3%

Table 7.10.: List of signal and background events after each pre-selection cut at $\sqrt{s} = 1.4\,\mathrm{TeV}$. The selection efficiencies are presented in a "flow" fashion. Every selection cut contains all the cuts to the left of it and all cuts in table 7.9. q can be u, d, s, b or t. Unless specified, q, ℓ and ν represent either particles or the 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.

7.8. MVA variables

Having extracted information about leptons, b-jets, and jet pairing, a number of variables are used to differentiate the signal and the background events. These variables are the basis of the subsequent MVA event selection. The variables used are listed in table 7.11. The distributions of the four most power discriminators are show in figure 7.13.

7.8.1. Invariant mass variables

Four invariant masses are used in the MVA event selection: the invariant mass of H_{bb} ($m_{H_{bb}}$), the invariant mass of H_{WW^*} ($m_{H_{WW^*}}$), the invariant mass of W (m_W), and the invariant mass of the double Higgs system (m_{HH}).

After the jet pairing under the hypothesis of the signal events, the distributions of the invariant mass of the physical bosons of the signal events have peaks around the expect masses, where the distributions of the background events do not have such resonance structure. Shown in the figure 7.13a, the distributions of the invariant mass of the H_{bb} is different to the distributions of the background events. Similarly, the distributions of the invariant mass of the H_{WW^*}, shown in figure 7.13b have a different peak position to the distributions of the background events. The invariant mass of the double Higgs system in the signal events is large due to the presence of two on-mass-shell Higgs bosons, which is also different to the distribution of the background events

7.8.2. Energy and momentum variables

Six energy and momentum variables participate in the MVA event selection: the energy of the off-mass-shell W (E_{W^*}), the energy of the missing momenta (E_{mis}), the transverse momentum of H_{bb} ($p_{TH_{bb}}$), the transverse momentum of H_{WW^*} ($p_{TH_{WW^*}}$), the transverse momentum of W (p_{TW}), and the transverse momentum of the double Higgs system (p_{THH}).

For the off-mass-shell W, the energy is used instead of the invariant mass, as invariant mass distribution of W^* does not have a resonance structure. The energy of the missing momenta is powerful against background events with no neutrinos in the final states. The missing momenta is calculated by assuming the collision at $\sqrt{s}$ and a beam crossing angle of 20 mrad. Other momentum variables correspond to the same physical bosons or

the double Higgs system used in the invariant mass variables, for the same reason that the distributions of these momentum variables are different for the signal events and the background events.

7.8.3. Lab-frame angle variables

Four lab-frame angle variables are in the MVA event selection: the pseudorapidity of the missing momenta (η_{mis}), the acollinearity of the two jets associated with H_{bb} ($A_{H_{bb}}$), the acollinearity of the two jets associated with H_{WW^*} ($A_{H_{WW^*}}$), and the acollinearity of the two Higgs bosons(A_{HH}).

The pseudorapidity of the missing momenta is used, instead of the polar angle, because the forward polar angles are transformed to a larger range in the pseudorapidity. The pseudorapidity of the missing momenta is defined as

$$\eta_{mis} \equiv -\ln\left[\tan\left(\frac{\theta_{mis}}{2}\right)\right], \tag{7.7}$$

where θ_{mis} is the polar angle of the missing momenta measured in a spherical polar coordinate system.

Acollinearity is a measures the angle between the two momenta. The definition for the acollinearity for momenta i and momenta j is

$$A_{ij} = \pi - \cos^{-1}\left(\hat{\mathbf{p}}_i \cdot \hat{\mathbf{p}}_j\right), \tag{7.8}$$

where $\hat{\mathbf{p}}_i$ is the unit momentum three-vector of momenta i. The distribution of the $A_{H_{bb}}$, shown in figure 7.13c, peaks at the value of 0 or π for many background events, which are not the same as the signal events. For the same reason, the distributions of $A_{H_{WW^*}}$ and A_{HH} are different for the signal and the background events.

7.8.4. Boosted-frame angle variables

The MVA event selection also uses five boosted-frame angle variables: the angle between two jets associated with H_{bb} in the H_{bb} decay rest frame ($\cos\left(\theta^*_{H_{bb}}\right)$), the angle between two W associated with H_{WW^*} in the H_{WW^*} decay rest frame ($\cos\left(\theta^*_{H_{WW^*}}\right)$), the angle between two jets associated with W in the W decay rest frame ($\cos\left(\theta^*_{W}\right)$), the angle

between two jets associated with W* in the W* decay rest frame ($\cos\left(\theta^*_{\mathrm{W}^*}\right)$), and the angle between two Higgs bosons in two Higgs bosons decay rest frame ($\cos\left(\theta^*_{\mathrm{HH}}\right)$).

These variables are some of the most powerful variables. For example, $\cos\left(\theta^*_{\mathrm{H_{bb}}}\right)$ for the signal events has a unform distribution, shown in figure 7.13d, as it is equally likely for two quarks to decay in any open angle in the $\mathrm{H_{bb}}$ decay rest frame. For the background events, by pairing jets under the signal event hypothesis, $\cos\left(\theta^*_{\mathrm{H_{bb}}}\right)$ does not have a flat distribution. The $\cos\left(\theta^*_{\mathrm{H_{bb}}}\right)$ distribution for the background events peaks at 1.

7.8.5. Event shape variables

Five event shapes variables are used in the MVA event selection: the absolute value of the sphericity ($|\mathbf{S}|$), the negative logarithm of y_{23} ($-\ln(y_{23})$), the negative logarithm of y_{34} ($-\ln(y_{34})$), the negative logarithm of y_{45} ($-\ln(y_{45})$), the negative logarithm of y_{56} ($-\ln(y_{56})$).

The sphericity, **S**, is a measurement of the spherically symmetry of the event, which will be different for the signal and background events. The sphericity is derived from the sphericity tensor [82]. The sphericity tensor is defined as

$$\boldsymbol{S}^{\alpha\beta} = \frac{\sum_i p_i^\alpha p_i^\beta}{\sum_i |\vec{p_i}|^2}, \tag{7.9}$$

where $\vec{p_i}$ is the momentum vector of the particle i; index i is summed over all particles in the event; and α and β refer to the x, y, z coordinate axis. Eigenvalues of $\boldsymbol{S}$ tensor, denoted with λ_1, λ_2, λ_3, can be found via diagonalisation of the matrix $\boldsymbol{S}$. 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,

$$\mathbf{S} = \frac{3}{2}(\lambda_1+\lambda_2). \tag{7.10}$$

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

The y parameter is a measure of the number of jets in an event. It describes the transition of the 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. Numerically the y parameter is often much smaller than 1. A typically way to convert the small number to a machine acceptable range is to take the negative logarithm of the number. See section 4.6.2 for discussion on jet algorithms and d_{cut}.

7.8.6. b and c tag variables

Six b-jet and c-jet tag variables are used in the MVA event selection: the highest b-jet tag value of the two jets associated with H_{bb} ($B_{1,H_{bb}}$), the lowest b-jet tag value of the two jets associated with H_{bb} ($B_{2,H_{bb}}$), the highest b-jet tag value of the two jets associated with W ($B_{1,W}$), the highest b-jet tag value of the two jets associated with W* (B_{1,W^*}), the highest c-jet tag value of the two jets associated with H_{bb} ($C_{1,H_{bb}}$), and the highest c-jet tag value of the two jets associated with W ($C_{1,W}$).

As mentioned in the flavour tagging section, these b-jet and c-jet tag variables are useful to separate the signal events from the background events which do not have b-quark jets in the final states.

7.8.7. PFOs number variables

The last four variables used in the MVA event selection are the PFOs number variables: the number of PFOs associated with H_{bb} (H_{bb}), the number of PFOs associated with H_{WW^*} (H_{WW^*}), the number of PFOs associated with W (W), the number of PFOs associated with W* (W*). These variables are effective to differentiate the signal events from the background events with fewer than six 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.

7.8.8. Cuts to aid the MVA

A set of cuts reduce the range of invariant masses variables in order to increase the effectiveness of the MVA event selection. Occasionally, extreme values of the invariant masses variables skew the distributions. Therefore by limiting the range of the variables, the MVA classifier could focus on the phase spaces with high event densities. The cuts

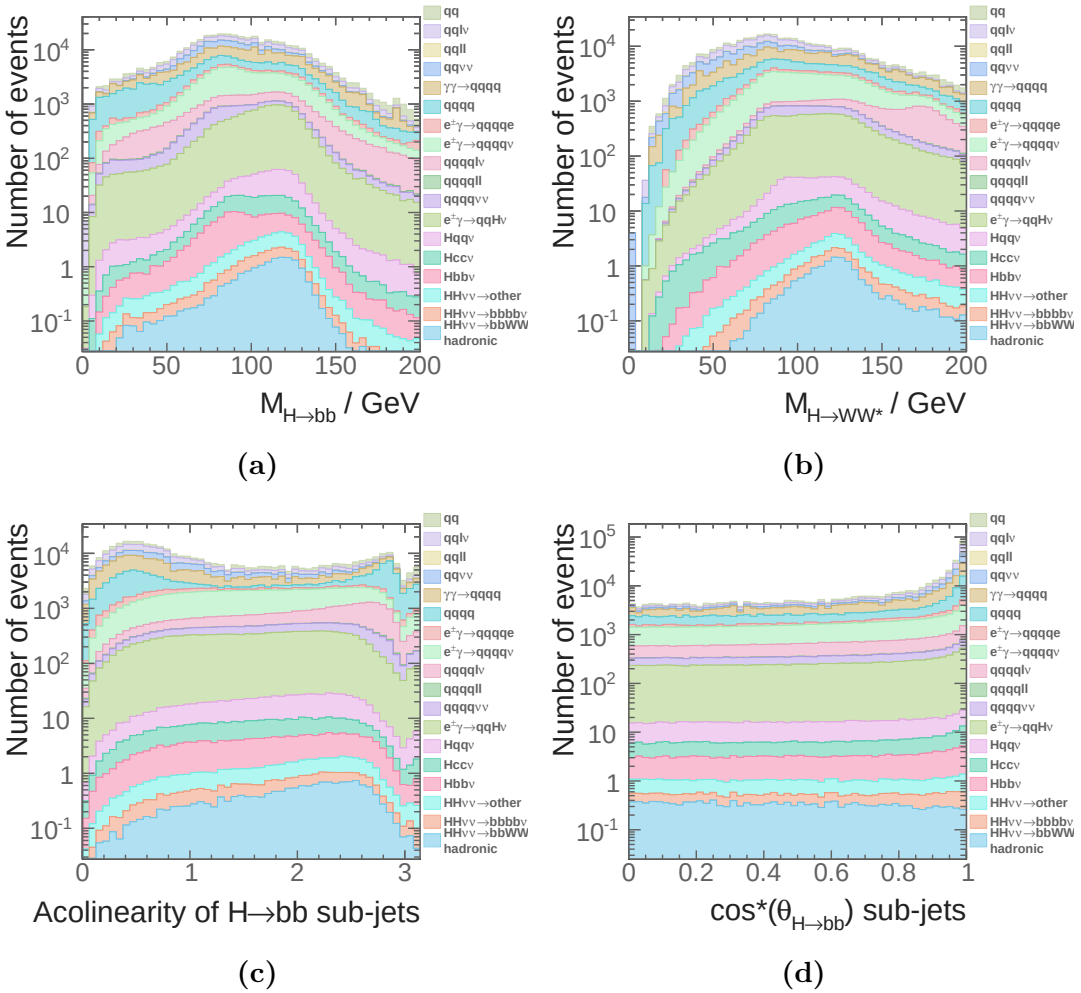


Figure 7.13.: Distributions of the four variables with highest discriminating power: a) the invariant mass of H_{bb}, b) the invariant mas of H_{WW^*}, c) the acolinearity of the two jets associated with H_{bb}, and d) the opening angles of the two jets associated with H_{bb} in the decay rest frame of the H_{bb}. All plots assumes an intergraded luminosity of 1500 fb^{-1} at $\sqrt{s} = 1.4$ TeV after all pre-selection cuts applied before the MVA.

Category	Variable
Invariant mass	$m_{\mathrm{H_{bb}}}$, $m_{\mathrm{H_{WW^*}}}$, m_{W}, m_{HH}
Energy and momentum	$E_{\mathrm{W^*}}$, E_{mis}, $p_{\mathrm{T\,H_{bb}}}$, $p_{\mathrm{T\,H_{WW^*}}}$, $p_{\mathrm{T\,W}}$, $p_{\mathrm{T\,HH}}$
Lab-frame angles	η_{mis}, $A_{\mathrm{H_{bb}}}$, A_{W}, A_{HH}
Boosted-frame angles	$\cos\left(\theta^*_{\mathrm{H_{bb}}}\right)$, $\cos\left(\theta^*_{\mathrm{H_{WW^*}}}\right)$, $\cos\left(\theta^*_{\mathrm{W}}\right)$, $\cos\left(\theta^*_{\mathrm{W^*}}\right)$, $\cos\left(\theta^*_{\mathrm{HH}}\right)$
Event shape	$\lvert\mathbf{S}\rvert$, $-\ln(y_{23})$, $-\ln(y_{34})$, $-\ln(y_{45})$, $-\ln(y_{56})$
b and c tag	$B_{1,\mathrm{H_{bb}}}$, $B_{2,\mathrm{H_{bb}}}$, $B_{1,\mathrm{W}}$, $B_{1,\mathrm{W^*}}$, $C_{1,\mathrm{H_{bb}}}$, $C_{1,\mathrm{W}}$
PFOs number	$N_{\mathrm{H_{bb}}}$, $N_{\mathrm{H_{WW^*}}}$, N_{W}, $N_{\mathrm{W^*}}$

Table 7.11.: Variables used in the MVA event selection for $\sqrt{s} = 1.4\,\mathrm{TeV}$

require the invariant mass of the $\mathrm{H_{bb}} < 500\,\mathrm{GeV}$, the invariant mass of the $\mathrm{H_{WW^*}} < 800\,\mathrm{GeV}$, the invariant mass of the $\mathrm{W} < 200\,\mathrm{GeV}$, and the invariant mass of the double Higgs system $< 1400\,\mathrm{GeV}$.

7.9. Multivariate analysis

After gathering information and applying pre-selection cuts, signal events are selected using the multivariate analysis (MVA) with Boosted Decision Tree classifier (BDT), as implemented in the TMVA [61]. The parameters for boosted decision tree were optimised and checked for overtraining, following the strategy outlined in section 4.7. Half of the events were used for training, and the other half used for testing and classifier optimisation. The optimised parameters are listed in table 7.12.

After dividing all events into a training set and a testing set, in the training stage of the MVA classifier, the training signal events are the hadronic $\mathrm{W^+W^-}$ decay of the $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$ events in the training set. The training background events are all events without double higgs production in the training set. However, for the extraction of the $\mathrm{g_{HHH}}$ and $\mathrm{g_{WWHH}}$, all events with double higgs production are sensitive to the couplings Therefore, at the applying stage of the MVA classifier, all events in the testing set are used.

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 7.12.: Optimised parameters for the boosted decision tree classifier used in the MVA event selection. See section 4.7.8 for detailed explanations of variables.

7.10. Signal selection results

Number of events passed the MVA event selection at $\sqrt{s} = 1.4\,\text{TeV}$, assuming a luminosity of 1500 fb^{-1} are listed in table 7.13 for individual channels. A few background channels have non-zero events after the MVA event selection. $e^+e^- \to q\bar{q}H\nu\overline{\nu}$ events are difficult to discard because its topology, one Higgs plus neutrino, is very similar to the signal event topology. Similarly, $e^+e^- \to qqqq\ell\nu$ events can be confused with the signal events when the lepton is undetected in the forward region, or the energy of the lepton is too low to be tagged. $e^+e^- \to qqqq\nu\overline{\nu}$ events can also have a similar topology to the signal events. Other background channels that are not discarded after the MVA are the electron-photon and photon interactions with the same final states as the channels above.

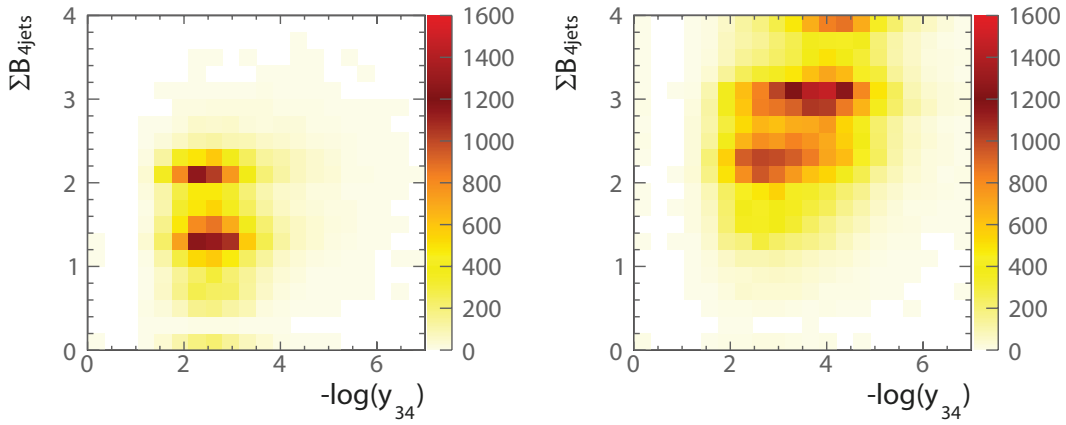
Before interpreting the result for analysis at $\sqrt{s} = 1.4\,\text{TeV}$, the analyses at $\sqrt{s} = 3\,\text{TeV}$ and the semi-leptonic channel of $e^+e^- \to HH\nu_e\overline{\nu}_e \to b\bar{b}W^+W^-\nu_e\overline{\nu}_e$ are presented.

7.11. $\sqrt{s} = 3\,\text{TeV}$ analysis

The the hadronic W^+W^- decay of the $e^+e^- \to HH\nu_e\overline{\nu}_e \to b\bar{b}W^+W^-\nu\overline{\nu}$ at $\sqrt{s} = 3\,\text{TeV}$ analysis follows the same strategy as the analysis at $\sqrt{s} = 1.4\,\text{TeV}$. Lepton finding, jet

$\sqrt{s} = 1.4\,\text{TeV}$	N	ε_{presel}	ε_{MVA}	N_{MVA}
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$				
$b\overline{b}W^+W^-\nu_e\overline{\nu}_e$, hadronic	27.9	59.8%	8.2%	1.29
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$				
$b\overline{b}b\overline{b}\nu_e\overline{\nu}_e$	67.6	15.4%	0.5%	0.05
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$ other	128.0	20.4%	1.7%	0.45
$e^+e^- \to q_l q_l H\nu\overline{\nu}$	1290	39.5%	0.05%	0.29
$e^+e^- \to c\overline{c}H\nu\overline{\nu}$	540	31.6%	0.1%	0.16
$e^+e^- \to b\overline{b}H\nu\overline{\nu}$	465	24.7%	0.3%	0.37
$e^+e^- \to qqqq$	1867650	3.3%	-	-
$e^+e^- \to qqqq\ell\ell$	93150	0.3%	-	-
$e^+e^- \to qqqq\ell\nu$	165600	9.8%	0.01%	2.06
$e^+e^- \to qqqq\nu\overline{\nu}$	34800	16.5%	0.002%	0.10
$e^+e^- \to qq$	6014250	0.8%	-	-
$e^+e^- \to qq\ell\nu$	6464550	0.9%	-	-
$e^+e^- \to qq\ell\ell$	4088700	0.08%	-	-
$e^+e^- \to qq\nu\nu$	1181550	4.0%	-	-
$e^\pm\gamma(BS) \to e^\pm qqqq$	2606625	0.3%	-	-
$e^\pm\gamma(EPA) \to e^\pm qqqq$	861000	0.3%	-	-
$e^\pm\gamma(BS) \to \nu qqqq$	178987.5	25.7%	0.005%	2.05
$e^\pm\gamma(EPA) \to \nu qqqq$	52050	12.5%	0.004%	0.27
$e^\pm\gamma(BS) \to qqH\nu$	35437.5	30.7%	0.02%	2.16
$e^\pm\gamma(EPA) \to qqH\nu$	10170.0	16.1%	0.06%	0.95
$\gamma(BS)\gamma(BS) \to qqqq$	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 7.13.: List of signal and background events with selection efficiency and number of events at $\sqrt{s} = 1.4\,\text{TeV}$, assuming a luminosity of $1500 fb^{-1}$. The number of events (N), the selection efficiencies of pre-selection cuts (ε_{presel}), the selection efficiencies of MVA after pre-selection cuts (ε_{MVA}), and the number of events after MVA (N_{MVA}) are shown. - represents a number less than 0.01. q can be u, d, s, b or t. Unless specified, q, ℓ and ν represent either particles or the 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.



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

Figure 7.14.: Sum of b-jet tag, when the event is clustered into four jets, as a function of y_{34} at $\sqrt{s} = 3\,\mathrm{TeV}$, shown for a) hadronic $\mathrm{W^+W^-}$ decay of $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$, and b) $\mathrm{HH} \to \mathrm{b\bar{b}b\bar{b}}$ sub-channels.

pairing and flavouring tagging have been discussed in previous sections. The differences, which have not been mentioned, will be highlighted in this section.

Cross sections of used samples are listed in table 7.14. The mutually exclusive cuts to separate events into two independent sets are almost identical to the cuts used in the $\sqrt{s} = 1.4\,\mathrm{TeV}$ analysis. Figure 7.14 shows the sum of b-jet tag values, when the event is clustered into four jets, as a function of $-\log(y_{34})$ for the hadronic $\mathrm{W^+W^-}$ decay in $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$ and $\mathrm{HH} \to \mathrm{b\bar{b}b\bar{b}}$ sub-channels. The optimised cuts are $\Sigma B_{4jets} < 2.3$, $-\log(y_{34}) < 3.6$. The selection efficiencies of evens after lepton veto, the mutually exclusive cuts and the jet pairing for individual channel are shown in table A.1.

The pre-selection cuts at $\sqrt{s} = 3\,\mathrm{TeV}$ use the same cut on m_{HH}. The cut on b-jet tag is different because the performance of flavour tagging is worse at $\sqrt{s} = 3\,\mathrm{TeV}$ in comparison to the performance at $\sqrt{s} = 1.4\,\mathrm{TeV}$. Figure 7.16 shows the distribution of the highest b-jet tag value, where the cut above 0.7 helps to reduce background events with no b-jet in final states. Figure 7.15 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. The fraction of events passing each pre-section cut for individual channel are listed in table A.2.

The cuts to aid the MVA at $\sqrt{s} = 3\,\mathrm{TeV}$ are largely the same as the ones at $\sqrt{s} = 1.4\,\mathrm{TeV}$, apart from the difference on the cut of the invariant mass of HH due to a higher

Channel	$\sigma(\sqrt{s} = 3\,TeV)$ / 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
$e^+e^- \to q_l q_l H\nu\overline{\nu}$	3.06
$e^+e^- \to c\overline{c}H\nu\overline{\nu}$	1.15
$e^+e^- \to b\overline{b}H\nu\overline{\nu}$	1.78
$e^+e^- \to qqqq$	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
$e^\pm\gamma$(BS) $\to e^\pm qqqq$	2536.3*
$e^\pm\gamma$(EPA) $\to e^\pm qqqq$	575.7*
$e^\pm\gamma$(BS) $\to \nu qqqq$	524.8*
$e^\pm\gamma$(EPA) $\to \nu qqqq$	108.4*
$e^\pm\gamma$(BS) $\to qqH\nu$	117.1*
$e^\pm\gamma$(EPA) $\to qqH\nu$	22.4*
γ(BS)γ(BS) $\to qqqq$	13050.3*
γ(BS)γ(EPA) $\to qqqq$	2420.6*
γ(EPA)γ(BS) $\to qqqq$	2423.1*
γ(EPA)γ(EPA) $\to qqqq$	402.7*

Table 7.14.: List of signal and background samples used in the double Higgs analysis with the corresponding cross sections at $\sqrt{s} = 3\,\text{TeV}$. q can be u, d, s, b or t. Unless specified, q, ℓ and ν represent either particles or the 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 labelled with *, events are generated with the invariant mass of the total momenta of all quarks above 50 GeV.

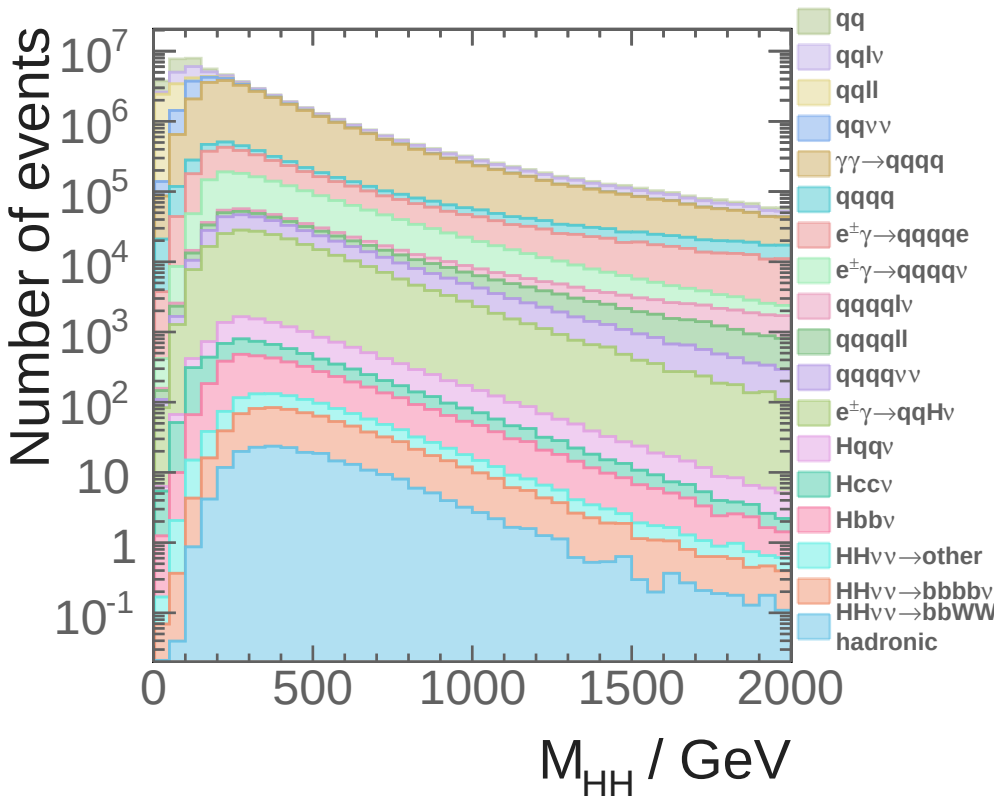


Figure 7.15.: Distributions of the invariant mass of the two Higgs system for $\sqrt{s} = 3\,\text{TeV}$, assuming an intergraded luminosity of 2000 fb^{-1}.

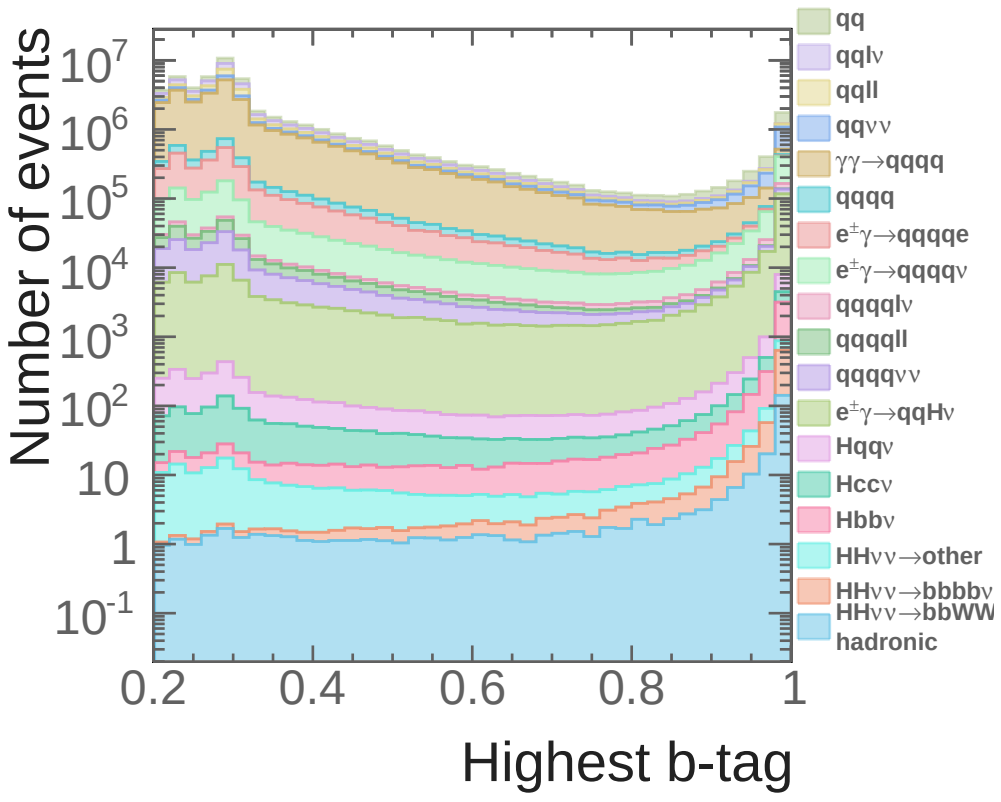


Figure 7.16.: Distributions of the highest b-jet tag value for $\sqrt{s} = 3\,\text{TeV}$, assuming an intergraded luminosity of 2000 fb^{-1}.

$\sqrt{s}$. The cuts are the invariant mass of the $H_{bb} < 500\,\text{GeV}$, the invariant mass of the $H_{WW^*} < 800\,\text{GeV}$, the invariant mass of the $W < 200\,\text{GeV}$, and the invariant mass of the double Higgs system $< 3000\,\text{GeV}$.

The same set of variables are used in the MVA as in the analysis at $\sqrt{s} = 1.4\,\text{TeV}$. The optimised parameters for the Boosted Decision Tree classifier are the same. The efficiencies of the MVA event selections and the number of events after the MVA event selection are listed in table 7.15. Background channels that are dominant after the MVA event selection are almost identical to those at $\sqrt{s} = 1.4\,\text{TeV}$. Hence see section 7.10 for discussion.

7.12. Semi-leptonic decay at $\sqrt{s} = 3\,\text{TeV}$ analysis

The final analysis is the semi-leptonic W^+W^- decay of $e^+e^- \to HH\nu_e\overline{\nu}_e \to b\bar{b}W^+W^-\nu\overline{\nu}$ at $\sqrt{s} = 3\,\text{TeV}$. The semi-leptonic decay analysis at $\sqrt{s} = 1.4\,\text{TeV}$ was also performed. However there are not enough signal events to have a meaningful discussion for the analysis at $\sqrt{s} = 1.4\,\text{TeV}$. Hence, only the semi-leptonic decay analysis at $\sqrt{s} = 3\,\text{TeV}$ is presented.

The strategy of the semi-leptonic decay 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 four quarks instead of six. H_{bb} and W can not be reconstructed due to the leptonic decay of one of the W. Hence, the signal events are selected when there is one identified lepton using the same lepton finding processors. The jet reconstruction parameters are the same as hadronic decay analysis at the $\sqrt{s} = 3\,\text{TeV}$. There are no mutually exclusive cuts since there is no semi-leptonic analysis in the $HH \to b\bar{b}b\bar{b}$ analysis.

The pre-selection cuts are similar to the cuts in the hadronic analysis. The invariant mass of the double Higgs system is required to be above 150 GeV. The highest b-jet tag value is higher than 0.2. The transverse momentum of the double Higgs system is higher than 30 GeV.

Variables used in the MVA classifier, listed in table 7.16, belong to a reduced set of the variables used in the hadronic decay analysis, as H_{bb} and W can not be reconstructed in the semi-hadronic decay analysis. For the same reason, the cuts to aid the MVA are

$\sqrt{s} = 3\,\text{TeV}$	N	ε_{presel}	ε_{MVA}	N_{MVA}
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$				
$b\overline{b}W^+W^-\nu_e\overline{\nu}_e$, hadronic	146.0	61.7%	11.6%	9.89
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$				
$b\overline{b}b\overline{b}\nu_e\overline{\nu}_e$	355.0	18.8%	1.5%	1.05
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$ other	675.0	20.0%	3.6%	4.51
$e^+e^- \to q_lq_lH\nu\overline{\nu}$	6120	36.0%	0.4%	9.42
$e^+e^- \to c\overline{c}H\nu\overline{\nu}$	2300	26.3%	0.5%	3.13
$e^+e^- \to b\overline{b}H\nu\overline{\nu}$	3560	25.8%	1.2%	6.82
$e^+e^- \to qqqq$	1093000	1.4%	0.01%	1.43
$e^+e^- \to qqqq\ell\ell$	338600	0.6%	-	-
$e^+e^- \to qqqq\ell\nu$	213200	7.3%	0.05%	8.35
$e^+e^- \to qqqq\nu\overline{\nu}$	143000	9.0%	0.05%	6.35
$e^+e^- \to qq$	5897800	1.4%	-	-
$e^+e^- \to qq\ell\nu$	11121800	0.1%	-	-
$e^+e^- \to qq\ell\ell$	6639200	0.4%	-	-
$e^+e^- \to qq\nu\nu$	2635000	3.1%	-	-
$e^\pm\gamma(BS) \to e^\pm qqqq$	4007354	0.7%	-	-
$e^\pm\gamma(EPA) \to e^\pm qqqq$	1151200	0.4%	-	-
$e^\pm\gamma(BS) \to \nu qqqq$	829184	16.4%	0.04%	61.0
$e^\pm\gamma(EPA) \to \nu qqqq$	216800	7.6%	0.04%	6.0
$e^\pm\gamma(BS) \to qqH\nu$	185018	30.2%	0.2%	121.7
$e^\pm\gamma(EPA) \to qqH\nu$	46800.0	15.3%	0.2%	18.1
$\gamma(BS)\gamma(BS) \to qqqq$	18009414	1.6%	-	-
$\gamma(BS)\gamma(EPA) \to qqqq$	3824548	1.0%	-	-
$\gamma(EPA)\gamma(BS) \to qqqq$	3828498	1.0%	-	-
$\gamma(EPA)\gamma(EPA) \to qqqq$	805400	0.6%	-	-

Table 7.15.: List of signal and background events with selection efficiency and number of events at $\sqrt{s} = 3\,\text{TeV}$, assuming a luminosity of $2000 fb^{-1}$. The number of events (N), the selection efficiencies of pre-selection cuts (ε_{presel}), the selection efficiencies of MVA after pre-selection cuts (ε_{MVA}), and the number of events after MVA (N_{MVA}) are shown. - represents a number less than 0.01. q can be u, d, s, b or t. Unless specified, q, ℓ and ν represent either particles or the 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.

Category	Variable
Invariant mass	$m_{\mathrm{H_{bb}}}$, m_{W}, m_{HH}
Energy and momentum	E_{mis}, $p_{\mathrm{T\,H_{bb}}}$, $p_{\mathrm{T\,W}}$, $p_{\mathrm{T\,HH}}$
Lab-frame angles	θ_{mis}, $A_{\mathrm{H_{bb}}}$, A_{W}, A_{HH}
Boosted-frame frames	$\cos\left(\theta^*_{\mathrm{H_{bb}}}\right)$, $\cos\left(\theta^*_{\mathrm{HH}}\right)$
Event shape	$\lvert\mathbf{S}\rvert$, $-\ln(y_{23})$, $-\ln(y_{34})$, $-\ln(y_{45})$, $-\ln(y_{56})$
b and c tag	$B_{1,\mathrm{H_{bb}}}$, $B_{2,\mathrm{H_{bb}}}$, $B_{1,\mathrm{W}}$, $C_{1,\mathrm{H_{bb}}}$, $C_{1,\mathrm{W}}$
PFOs number	$N_{\mathrm{H_{bb}}}$, N_{W}

Table 7.16.: Variables used in the MVA event selection for the semi-leptonic $\mathrm{W^+W^-}$ decay of $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$ analysis at $\sqrt{s} = 3\,\mathrm{TeV}$.

reduced to the invariant mass of $\mathrm{H_{bb}} < 500\,\mathrm{GeV}$ and the invariant mass of the double Higgs system $< 3000\,\mathrm{GeV}$.

Figure 7.17 lists the selection efficiency and number of events after the MVA event selection for individual channel at $\sqrt{s} = 3\,\mathrm{TeV}$, assuming a luminosity of 2000 fb^{-1}. Almost all background channels are non-zero after the MVA event selection. Nevertheless, dominant background channels are almost identical to the hadronic decay analysis at $\sqrt{s} = 3\,\mathrm{TeV}$. Hence discussion of the MVA event selection is provided in section 7.10.

7.13. Result interpretation

The results of analyses at the $\sqrt{s} = 1.4\,\mathrm{TeV}$ and $\sqrt{s} = 3\,\mathrm{TeV}$ are summarised in table 7.18. For the hadronic $\mathrm{W^+W^-}$ decay of $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$ analyses at $\sqrt{s} = 1.4\,\mathrm{TeV}$ and $\sqrt{s} = 3\,\mathrm{TeV}$, numbers of the signal events passing the MVA event selection are 1.79 and 15.45, respectively; the numbers of background events passing the MVA event selection are 8.41 and 242.28, respectively. For the semi-leptonic $\mathrm{W^+W^-}$ decay of $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$ analyses $\sqrt{s} = 3\,\mathrm{TeV}$, the number of the signal events passing the MVA event selection is 31.24, whilst the numbers of background events passing the MVA event selection is 3612.39.

$\sqrt{s} = 3\,\text{TeV}$	N	ε_{presel}	ε_{MVA}	N_{MVA}
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$				
$b\overline{b}W^+W^-\nu_e\overline{\nu}_e$, semi-leptonic	96.8	44.6%	21.9%	13.11
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$				
$b\overline{b}b\overline{b}\nu_e\overline{\nu}_e$	355.0	13.3%	10.9%	5.38
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$ other	724.2	13.1%	13.6%	12.75
$e^+e^- \to q_l q_l H\nu\overline{\nu}$	6120	7.4%	13.7%	62.63
$e^+e^- \to c\overline{c}H\nu\overline{\nu}$	2300	6.3%	12.1%	17.10
$e^+e^- \to b\overline{b}H\nu\overline{\nu}$	3560	15.9%	5.1%	18.03
$e^+e^- \to qqqq$	1093000	0.6%	0.2%	15.04
$e^+e^- \to qqqq\ell\ell$	338600	1.0%	0.06%	1.85
$e^+e^- \to qqqq\ell\nu$	213200	27.6%	0.5%	270.33
$e^+e^- \to qqqq\nu\overline{\nu}$	143000	1.9%	1.6%	43.78
$e^+e^- \to qq$	5897800	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	0.6%	0.2%	84.14
$e^+e^- \to qq\nu\nu$	2635000	0.4%	0.9%	92.55
$e^\pm\gamma(\text{BS}) \to e^\pm qqqq$	4007354	1.2%	-	-
$e^\pm\gamma(\text{EPA}) \to e^\pm qqqq$	1151200	1.1%	-	-
$e^\pm\gamma(\text{BS}) \to \nu qqqq$	829184	3.6%	1.5%	452.45
$e^\pm\gamma(\text{EPA}) \to \nu qqqq$	216800	11.0%	0.9%	200.65
$e^\pm\gamma(\text{BS}) \to qqH\nu$	185018	7.9%	10.4%	1521.93
$e^-\gamma(\text{EPA}) \to qqH\nu$	46800	22.8%	7.1%	750.85
$\gamma(\text{BS})\gamma(\text{BS}) \to qqqq$	18009414	0.4%	-	-
$\gamma(\text{BS})\gamma(\text{EPA}) \to qqqq$	3824548	1.0%	-	-
$\gamma(\text{EPA})\gamma(\text{BS}) \to qqqq$	3828498	1.0%	0.08%	28.85
$\gamma(\text{EPA})\gamma(\text{EPA}) \to qqqq$	805400	1.1%	-	-

semi-leptonic W^+W^- decay of $HH \to b\overline{b}W^+W^-$

Table 7.17.: List of signal and background events with selection efficiency and number of events at $\sqrt{s} = 3\,\text{TeV}$ for semi-leptonic W^+W^- decay of $HH \to b\overline{b}W^+W^-$ analysis , assuming a luminosity of $2000 fb^{-1}$. The number of events (N), the selection efficiencies of pre-selection cuts (ε_{presel}), the selection efficiencies of MVA after pre-selection cuts (ε_{MVA}), and the number of events after MVA (N_{MVA}) are shown. - represents a number less than 0.01. q can be u, d, s, b or t. Unless specified, q, ℓ and ν represent either particles or the 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.

The expected uncertainty on the measurement of 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:

$$\frac{\Delta\left[\sigma\left(\text{HH}\nu_\text{e}\overline{\nu}_\text{e}\right)\right]}{\sigma\left(\text{HH}\nu_\text{e}\overline{\nu}_\text{e}\right)} = \begin{cases} 179\%, & \text{at } \sqrt{s} = 1.4\,\text{TeV}, \\ 92\%, & \text{at } \sqrt{s} = 3\,\text{TeV}, \end{cases} \tag{7.11}$$

where N_S is the number of $\text{e}^+\text{e}^- \to \text{HH}\nu_\text{e}\overline{\nu}_\text{e}$ events passing the MVA event selection and N_B is the number of background events passing the MVA event selection. The result at $\sqrt{s} = 3\,\text{TeV}$ combines both the hadronic and semi-leptonic decay sub-channels.

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

Table 7.18.: Number of signal and background events, and significance after MVA for all $\text{HH} \to \text{b}\overline{\text{b}}\text{W}^+\text{W}^-$ analyses.

As previously stated, the double Higgs production cross section is sensitive to the Higgs trilinear self coupling g_HHH. The relative uncertainty on the coupling can be related to the uncertainty on the coupling via:

$$\frac{\Delta\text{g}_\text{HHH}}{\text{g}_\text{HHH}} \approx \kappa \cdot \frac{\Delta\left[\sigma\left(\text{HH}\nu_\text{e}\overline{\nu}_\text{e}\right)\right]}{\sigma\left(\text{HH}\nu_\text{e}\overline{\nu}_\text{e}\right)}, \tag{7.12}$$

κ can be extracted by varying the g_HHH and parameterising the cross section. Figure 7.17 shows the cross sections as a function of the coupling at generator level for $\sqrt{s} = 1.4\,\text{TeV}$ and $\sqrt{s} = 3\,\text{TeV}$ [20]. The negative gradient indicates that the Feynman diagram that is sensitive to the g_HHH experiences destructive interferences with other SM Feynman diagrams. At the SM g_HHH value, the κ is 1.22 and 1.47 at $\sqrt{s} = 1.4\,\text{TeV}$ and $3\,\text{TeV}$ respectively.

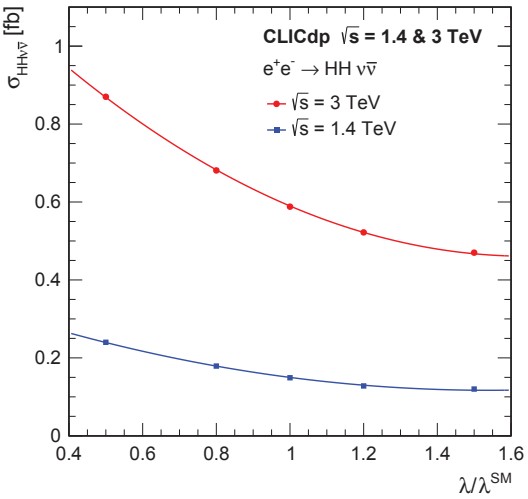


Figure 7.17.: Cross section for the $e^+e^- \to HH\nu_e\overline{\nu}_e$ process as a function of the ratio λ/λ_{SM} at $\sqrt{s} = 1.4\,\text{TeV}$ and 3 TeV, taken from [20]. Here λ is the Higgs trilinear self coupling, g_{HHH}.

The uncertainty on measurement of the Higgs trilinear self coupling, g_{HHH}, from $e^+e^- \to HH\nu_e\overline{\nu}_e \to b\bar{b}W^+W^-\nu_e\overline{\nu}_e$ analysis is obtained via equation 7.12:

$$\frac{\Delta g_{HHH}}{g_{HHH}} = \begin{cases} 218\%, & \text{at } \sqrt{s} = 1.4\,\text{TeV}, \\ 135\%, & \text{at } \sqrt{s} = 3\,\text{TeV}. \end{cases} \tag{7.13}$$

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 $P(e^-) = 80\%$, the uncertainty of g_{HHH} via equation 7.12 becomes:

$$\frac{\Delta g_{HHH}}{g_{HHH}} = \begin{cases} 163\%, & \text{at } \sqrt{s} = 1.4\,\text{TeV}, \\ 97\%, & \text{at } \sqrt{s} = 3\,\text{TeV}. \end{cases} \tag{7.14}$$

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:

$$\frac{\Delta\left[\sigma\left(HH\nu_e\bar{\nu}_e\right)\right]}{\sigma\left(HH\nu_e\bar{\nu}_e\right)} = \begin{cases} 44\%, & \text{at } \sqrt{s} = 1.4\,\text{TeV}, \\ 20\%, & \text{at } \sqrt{s} = 3\,\text{TeV}, \end{cases} \tag{7.15}$$

This translates to the uncertainty on the Higgs trilinear self coupling g_{HHH}, via equation 7.12, without electron polarisation:

$$\frac{\Delta g_{HHH}}{g_{HHH}} = \begin{cases} 54\%, & \text{at } \sqrt{s} = 1.4\,\text{TeV}, \\ 29\%, & \text{at } \sqrt{s} = 3\,\text{TeV}. \end{cases} \tag{7.16}$$

7.15. Simultaneous couplings extraction

As stated in the beginning of the chapter the study of the double Higgs production via W^+W^- fusion can probe the Higgs trilinear self coupling, g_{HHH} and quartic coupling, g_{WWHH}. Therefore, a simultaneous extraction on the coupling uncertainty can be performed by extending the method in the previous sections. Once a relationship between g_{HHH}, g_{WWHH} and difference in kinematic variable distributions is established, a contour of the uncertainty of the measurements of g_{HHH} and g_{WWHH} in 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 a fitting. The integrated luminosity is assumed to be 3000 fb^{-1} to reflect the updated CLIC running scenario.

The normalised cross section of the $e^+e^- \to HH\nu_e\bar{\nu}_e$ as a function of g_{HHH} and g_{WWHH} is shown in figure 7.18. The SM cross section is normalised to 1. Around the SM coupling value, the cross section increases with the decrease of g_{HHH} and with the increase of g_{WWHH}. Hence the cross sections along the anti-diagonal are nearly constant, 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 scalar sum of the two Higgs transverse momentum, H_T.

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. Figure 7.19 shows the signal significance of the double Higgs events with hadronic $\mathrm{W^+W^-}$ decay of $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$ sub-channel as a function of $\mathrm{g_{HHH}}$ and $\mathrm{g_{WWHH}}$.

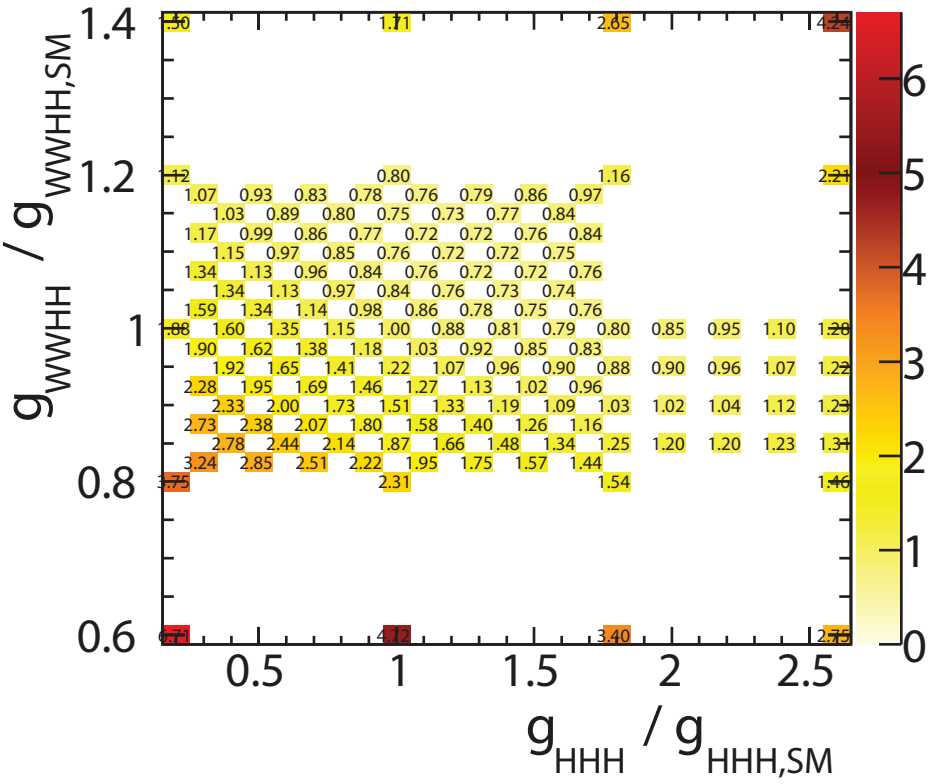


Figure 7.18.: Normalised cross section for the $\mathrm{e^+e^- \to HH\nu_e\bar{\nu}_e}$ process as a function of the $\mathrm{g_{HHH}/g_{HHH}}_{SM}$ and $\mathrm{g_{WWHH}/g_{WWHH}}_{SM}$ at $\sqrt{s} = 3\,\mathrm{TeV}$.

The selected events are divided into 8 kinematic bins. Two bins in H_T are obtained by dividing the H_T distribution at 200 GeV. Four bins in m_{HH} are obtained by dividing the m_{HH} distribution at 400, 560, and 720 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

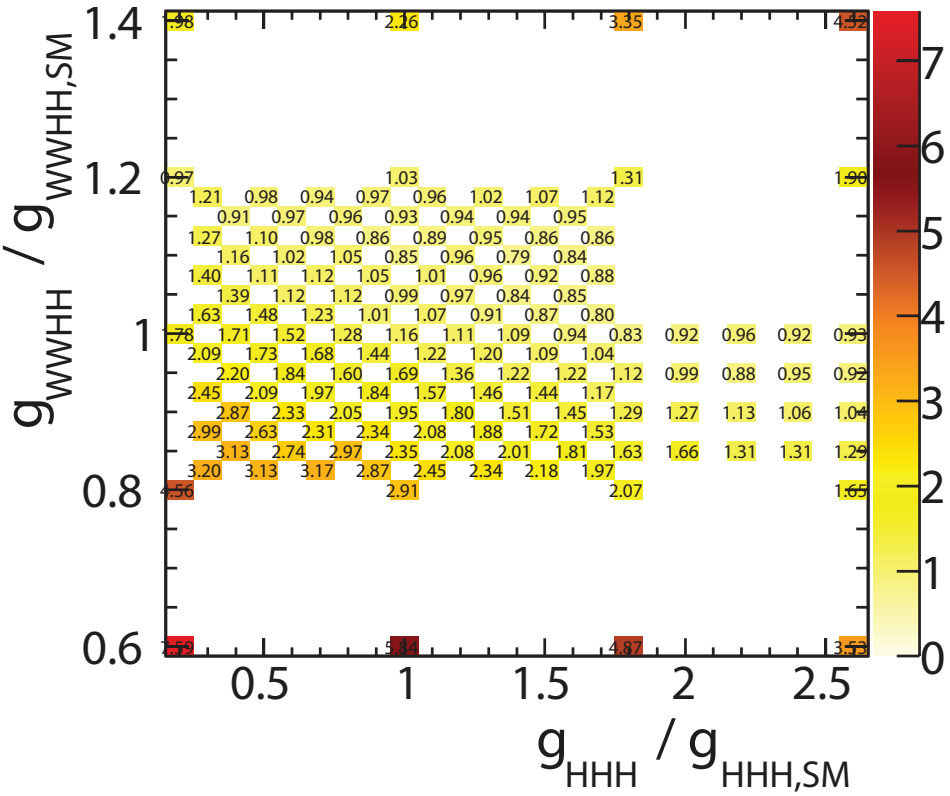


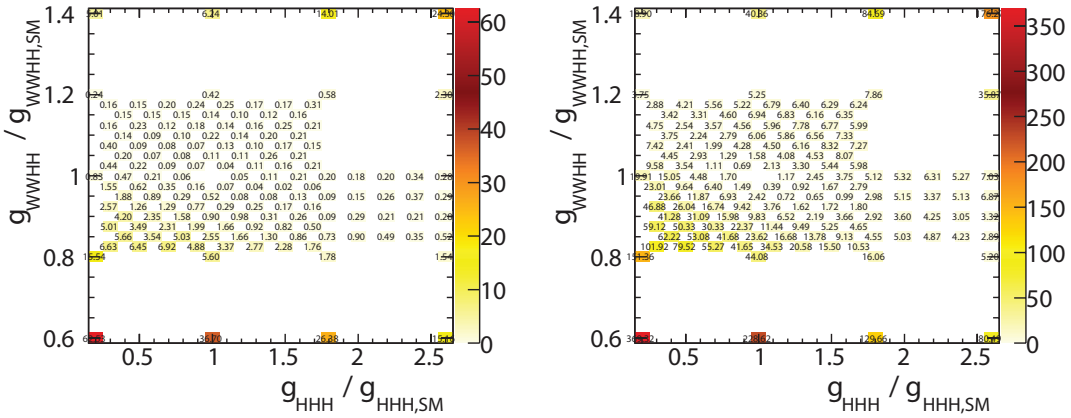
Figure 7.19.: The significance for the $e^+e^- \to HH\nu_e\overline{\nu}_e$ process as a function of the g_{HHH}/g_{HHHSM} and g_{WWHH}/g_{WWHHSM} at $\sqrt{s} = 3\,\text{TeV}$, using sub-channel hadronic W^+W^- decay of $HH \to b\bar{b}W^+W^-$, assuming an integrated luminosity of $3000 fb^{-1}$.

coupling sample, defined as:

$$\chi^2 = \sum_{i}^{bins} \frac{(N_i - N_{i,observed})^2}{N_i}, \tag{7.17}$$

where N_i is the number of event expected in a kinematic bin i for a non-SM coupling sample; and $N_{i,observed}$ is the number of event observed in a kinematic bin i. Here the observed set is the SM coupling sample. The expression is summed over all kinematic bins. By construction, the SM coupling point has a χ^2 of 0. Figure 7.20 shows the χ^2 as a function of g_{HHH} and g_{WWHH} for two sub-channels; hadronic W^+W^- decay of $HH \to b\bar{b}W^+W^-$ 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 as the $HH \to b\bar{b}b\bar{b}$ is more sensitive to the couplings.

Two sub-channels, hadronic W^+W^- decay of $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$, are combined to increase the statistical precision on the coupling measurements. To avoid statistical fluctuations in the sample, a toy MC experiment is performed. The SM



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

Figure 7.20.: 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 a) hadronic W^+W^- decay of $HH \to b\bar{b}W^+W^-$, b) and $HH \to b\bar{b}b\bar{b}$ sub-channel, assuming an integrated luminosity of 3000 fb^{-1}.

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 7.21 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 7.22 shows the contour. The counter can be sliced one dimensionally to extract the uncertainty of the measurements of one coupling for a given value of the other coupling. For example:

$$\frac{\Delta g_{WWHH}}{g_{WWHH}} \simeq 4.9\% \text{ for } g_{HHH} = g_{HHH,SM} \tag{7.18}$$

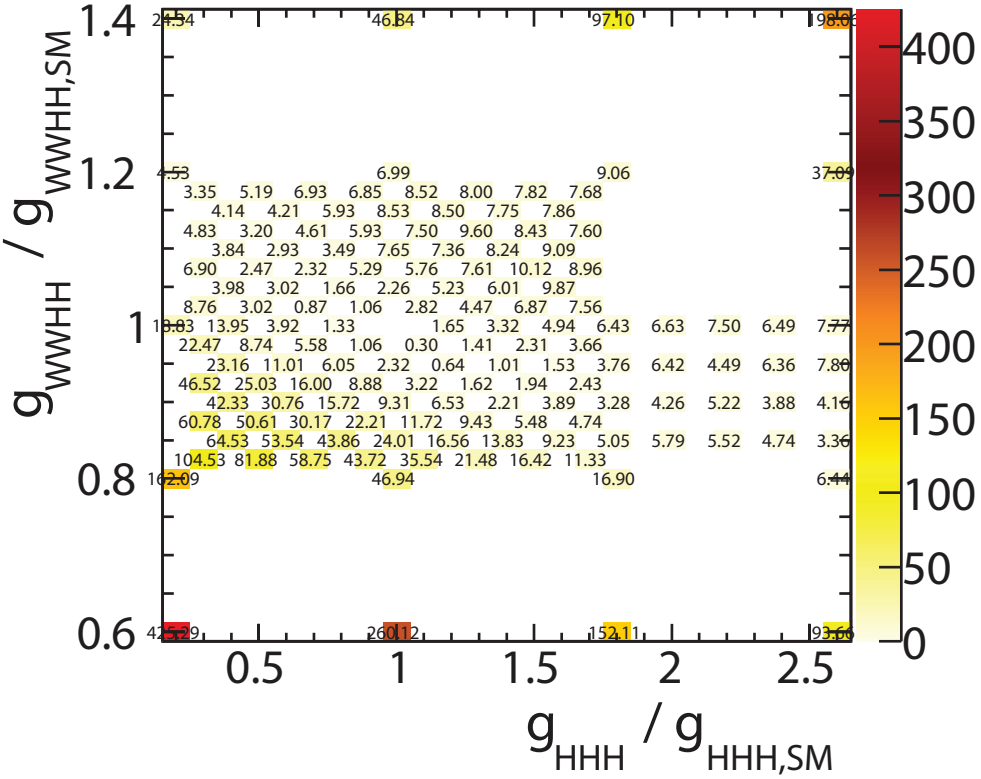


Figure 7.21.: Normalised χ^2, after averaging the toy MC experiments, as a function of g_{HHH}/g_{HHHSM} and g_{WWHH}/g_{WWHHSM}, combining hadronic decay $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$ sub-channels, assuming an integrated luminosity of $3000 fb^{-1}$.

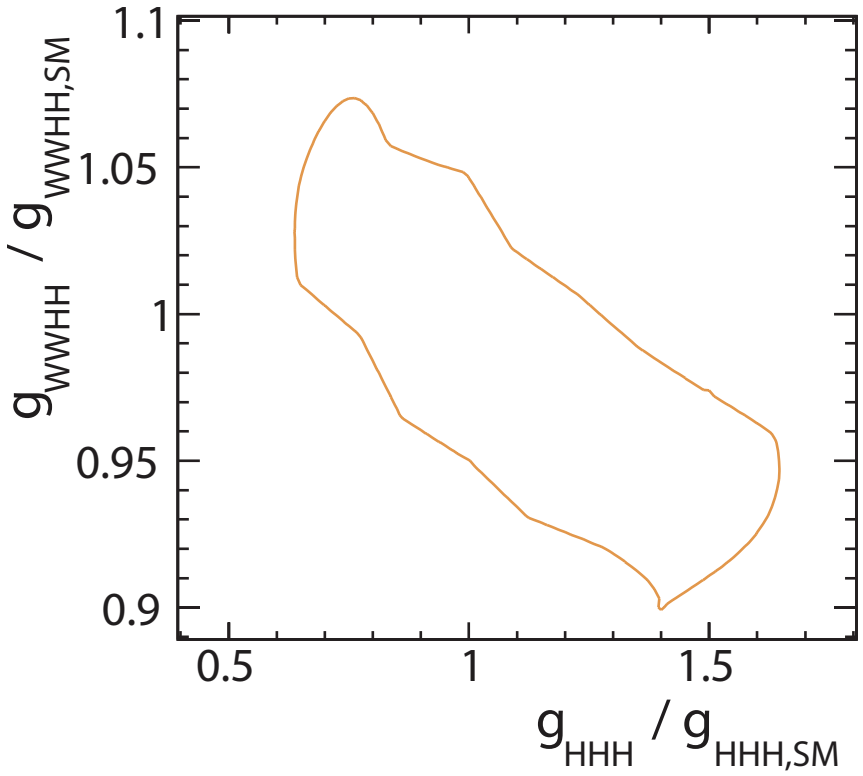


Figure 7.22.: Contour plot of 68% confidence ($\chi^2 = 2.3$) , after averaging the toy MC experiments, as a function of $g_{HHH}/g_{HHH_{SM}}$ and $g_{WWHH}/g_{WWHH_{SM}}$, combining hadronic W^+W^- decay of $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$ sub-channels, assuming an integrated luminosity of $3000 fb^{-1}$.

$$\frac{\Delta g_{HHH}}{g_{HHH}} \simeq 29\% \text{ for } g_{WWHH} = g_{WWHH,SM} \tag{7.19}$$

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 [17].

Chapter 8.

Summary

'If you know the enemy and know yourself, you need not fear the result of a hundred battles.'

— Sun Tzu, 544 BC - 496 BC

This chapter summarises key results presented in analyses in previous chapters. In chapter 5, a set of photon reconstruction algorithms developed in PandoraPFA are presented. The photon fragments produced during the event reconstruction have been greatly reduced. The photon separation power and the jet energy resolution have improved, as a result of a better photon reconstruction.

For the single photon reconstruction, the efficiency is above 98% for photons with energies above 2GeV, and above 99.5% for photons with energies above 100 GeV. To quantise the photon fragment reduction, for a 500 - 50 GeV photons pair sample, the average number of photons and particles beyond 20 mm apart are both less than 2.05, where the true value is 2. For the photon separation power, 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. For photon pairs with different energies, for example 500 - 50 GeV pair and 100 - 10 GeV pair, start to be resolved at 10 mm apart, which is about 2 ECAL cells. At 20 mm apart, two photons in 500 - 500 GeV pair are fully resolved, where approximately 60% of two photons in 10 - 10 GeV pair are resolved.

In chapter 6, a high classification rate of the tau lepton seven major decay modes is achieved. The classification is applied to different electromagnetic calorimeter cell sizes with different centre-of-mass energies. The tau hadronic decay correct classification

efficiency, ε_{had}, is used as the performance metric. At $\sqrt{s} = 100\,\text{GeV}$, the ε_{had} decreases from 94% at 3 mm cell size, to 91% at 20 mm cell size. Most significant decrease in the ε_{had} occurs at $\sqrt{s} = 500\,\text{GeV}$, where the ε_{had} decreases from 92% at 3 mm cell size, to 78% at 20 mm cell size. The increase in ECAL cell sizes has a larger impact in tau decay classification at high centre-of-mass energies. With decay products spatially close at high centre-of-mass energies, it is more beneficial to have a smaller ECAL cell size to reconstruct individual particle.

With the developed tau decay mode classification, a proof-of-principle analysis shows the tau pair polarisation correlations with $\text{Z} \to \tau^+\tau^-$ decay where $\tau^- \to \pi^-\nu_\tau$ can be observed with ILD detector model. A good match of the tau pair polarisation correlations between the reconstruction and the Monte Carlo simulation is also achieved. With a similar study of $\text{H} \to \tau^+\tau^-$, the tau polarisation correlations can be used to as a signature to identify Higgs boson from Z boson.

In chapter 7, the analyses of the $\text{e}^+\text{e}^- \to \text{HH}\nu_\text{e}\overline{\nu}_\text{e} \to \text{b}\overline{\text{b}}\text{W}^+\text{W}^-\nu_\text{e}\overline{\nu}_\text{e}$ channel for the Compact Linear Collider at $\sqrt{s} = 1.4\,\text{TeV}$ and $\sqrt{s} = 3\,\text{TeV}$ are performed. The significance of the signal events are 0.56 and 1.09, assuming an integrated luminosity of $1500\,fb^{-1}$ and $2000\,fb^{-1}$, for $\sqrt{s} = 1.4\,\text{TeV}$ and $\sqrt{s} = 3\,\text{TeV}$ respectively. The uncertainty on measurement of the Higgs trilinear self coupling, g_HHH, from $\text{e}^+\text{e}^- \to \text{HH}\nu_\text{e}\overline{\nu}_\text{e} \to \text{b}\overline{\text{b}}\text{W}^+\text{W}^-\nu_\text{e}\overline{\nu}_\text{e}$ analysis is obtained:

$$\frac{\Delta \text{g}_\text{HHH}}{\text{g}_\text{HHH}} = \begin{cases} 218\%, & \text{at } \sqrt{s} = 1.4\,\text{TeV}, \\ 135\%, & \text{at } \sqrt{s} = 3\,\text{TeV}. \end{cases} \tag{8.1}$$

When analysis at both $\text{HH} \to \text{b}\overline{\text{b}}\text{W}^+\text{W}^-$ and $\text{HH} \to \text{b}\overline{\text{b}}\text{b}\overline{\text{b}}$ sub-channels are combined at $\sqrt{s} = 3\,\text{TeV}$ to improve the measurement, the simultaneous extraction of the uncertainty on the measurement of the g_HHH and g_WWHH yeilds:

$$\frac{\Delta \text{g}_\text{WWHH}}{\text{g}_\text{WWHH}} \simeq 4.9\% \text{ for } \text{g}_\text{HHH} = \text{g}_{\text{HHH},SM} \tag{8.2}$$

$$\frac{\Delta \text{g}_\text{HHH}}{\text{g}_\text{HHH}} \simeq 29\% \text{ for } \text{g}_\text{WWHH} = \text{g}_{\text{WWHH},SM} \tag{8.3}$$

Appendix A.

Double Higgs Boson Production Analysis

'I was an adventurer like you, then I took an arrow in the knee.'
— The town guard, Skyrim, 2011

Here are extra tables and plots for the chapter 7.

A.1. Hadronic decay at $\sqrt{s} = 3\,\text{TeV}$ analysis

$\sqrt{s} = 3\,\text{TeV}$	Expected number of events	Lepton veto	Mutually exclusive	Jet pairing
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$				
$b\overline{b}W^+W^-\nu_e\overline{\nu}_e$, hadronic	146.0	80.9%	72.8%	72.1%
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$				
$b\overline{b}b\overline{b}\nu_e\overline{\nu}_e$	355.0	83.5%	20.5%	20.5%
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$ other	675.0	40.1%	34.3%	20.5%
$e^+e^- \to q_\ell q_\ell H\nu\overline{\nu}$	6120	67.7%	61.9%	61.9%
$e^+e^- \to c\overline{c}H\nu\overline{\nu}$	2300	69.1%	53.0%	48.8%
$e^+e^- \to b\overline{b}H\nu\overline{\nu}$	3560	70.1%	30.9%	30.6%
$e^+e^- \to qqqq$	1093000	62.4%	44.9%	34.9%
$e^+e^- \to qqqq\ell\ell$	338600	21.4%	19.6%	13.3%
$e^+e^- \to qqqq\ell\nu$	213200	23.3%	19.5%	16.3%
$e^+e^- \to qqqq\nu\overline{\nu}$	143000	80.7%	71.4%	50.7%
$e^+e^- \to qq$	5897800	72.9%	63.9%	55.4%
$e^+e^- \to qq\ell\nu$	11121800	34.0%	24.7%	20.5%
$e^+e^- \to qq\ell\ell$	6639200	43.1%	41.7%	37.0%
$e^+e^- \to qq\nu\nu$	2635000	84.6%	63.8%	53.2%
$e^\pm\gamma(BS) \to e^\pm qqqq$	4007354	31.0%	28.2%	21.1%
$e^\pm\gamma(EPA) \to e^\pm qqqq$	1151200	15.9%	14.5%	10.9%
$e^\pm\gamma(BS) \to \nu qqqq$	829184	78.3%	68.8%	53.3%
$e^\pm\gamma(EPA) \to \nu qqqq$	216800	39.6%	35.0%	26.9%
$e^\pm\gamma(BS) \to qqH\nu$	185018.0	64.0%	55.4%	49.8%
$e^\pm\gamma(EPA) \to qqH\nu$	46800	32.9%	28.8%	25.9%
$\gamma(BS)\gamma(BS) \to qqqq$	18009414	71.6%	65.5%	49.4%
$\gamma(BS)\gamma(EPA) \to qqqq$	3824548	44.3%	40.6%	30.6%
$\gamma(EPA)\gamma(BS) \to qqqq$	3828498	44.3%	40.7%	30.7%
$\gamma(EPA)\gamma(EPA) \to qqqq$	805400	29.0%	26.7%	20.1%

Table A.1.: Number of events and fraction of events passing lepton veto, the mutually exclusive cuts, and the jet pairing for the signal and background events at $\sqrt{s} = 3\,\text{TeV}$, assuming an integrated luminosity of 2000 fb^{-1}. The selection efficiencies are presented in a "flow" fashion. Every selection cut contains all the cuts to the left of it. q can be u, d, s, b or t. Unless specified, q, ℓ and ν represent either particles or the 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.

Channel	m_{HH}>150GeV	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	71.7%	61.8%
$e^+e^- \to HH\nu_e\bar{\nu}_e \to$		
$b\bar{b}b\bar{b}\nu_e\bar{\nu}_e$	20.2%	18.8%
$e^+e^- \to HH\nu_e\bar{\nu}_e \to$ other	30.2%	20.0%
$e^+e^- \to q_l q_l H\nu\bar{\nu}$	53.1%	36.0%
$e^+e^- \to c\bar{c}H\nu\bar{\nu}$	43.8%	26.3%
$e^+e^- \to b\bar{b}H\nu\bar{\nu}$	29.6%	25.9%
$e^+e^- \to qqqq$	26.5%	1.7%
$e^+e^- \to qqqq\ell\ell$	12.8%	0.7%
$e^+e^- \to qqqq\ell\nu$	16.0%	7.9%
$e^+e^- \to qqqq\nu\bar{\nu}$	49.7%	9.0%
$e^+e^- \to qq$	8.3%	1.4%
$e^+e^- \to qq\ell\nu$	6.0%	0.1%
$e^+e^- \to qq\ell\ell$	1.9%	0.4%
$e^+e^- \to qq\nu\nu$	16.6%	3.1%
$e^\pm\gamma(\mathrm{BS}) \to e^\pm qqqq$	19.4%	0.7%
$e^\pm\gamma(\mathrm{EPA}) \to e^\pm qqqq$	9.9%	0.4%
$e^\pm\gamma(\mathrm{BS}) \to \nu qqqq$	51.3%	16.4%
$e^\pm\gamma(\mathrm{EPA}) \to \nu qqqq$	26.0%	7.7%
$e^\pm\gamma(\mathrm{BS}) \to qqH\nu$	47.9%	30.3%
$e^-\gamma(\mathrm{EPA}) \to qqH\nu$	25.0%	15.8%
$\gamma(\mathrm{BS})\gamma(\mathrm{BS}) \to qqqq$	44.5%	1.7%
$\gamma(\mathrm{BS})\gamma(\mathrm{EPA}) \to qqqq$	27.4%	1.0%
$\gamma(\mathrm{EPA})\gamma(\mathrm{BS}) \to qqqq$	27.5%	1.0%
$\gamma(\mathrm{EPA})\gamma(\mathrm{EPA}) \to qqqq$	18.0%	0.7%

Table A.2.: List of signal and background events after each pre-selection cut at $\sqrt{s} = 3\,\mathrm{TeV}$. The selection efficiencies are presented in a "flow" fashion. Every selection cut contains all the cuts to the left of it and all cuts in table A.1. q can be u, d, s, b or t. Unless specified, q, ℓ and ν represent either particles or the 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.

Colophon

This thesis was made in LaTeX 2_ε using the "hepthesis" class [83].

Bibliography

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

[2] CMS Collaboration, S. Chatrchyan *et al.*, Phys.Lett. **B716**, 30 (2012), 1207.7235.

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

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

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

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

[7] 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.

[8] D0, S. Abachi *et al.*, Phys. Rev. Lett. **74**, 2632 (1995), hep-ex/9503003.

[9] DONUT, K. Kodama *et al.*, Phys. Lett. **B504**, 218 (2001), hep-ex/0012035.

[10] T. Aoyama, M. Hayakawa, T. Kinoshita, and M. Nio, Phys. Rev. **D91**, 033006 (2015), 1412.8284.

[11] D. Hanneke, S. F. Hoogerheide, and G. Gabrielse, Phys. Rev. **A83**, 052122 (2011), 1009.4831.

[12] S. Weinberg, Phys. Rev. Lett. **19**, 1264 (1967).

[13] 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.

[14] D. B. Kaplan and H. Georgi, Phys. Lett. **B136**, 183 (1984).

[15] W. D. Goldberger, B. Grinstein, and W. Skiba, Phys. Rev. Lett. **100**, 111802 (2008), 0708.1463.

[16] G. F. Giudice, C. Grojean, A. Pomarol, and R. Rattazzi, JHEP **06**, 045 (2007), hep-ph/0703164.

[17] R. Contino, C. Grojean, M. Moretti, F. Piccinini, and R. Rattazzi, JHEP **05**, 089 (2010), 1002.1011.

[18] R. Contino, C. Grojean, D. Pappadopulo, R. Rattazzi, and A. Thamm, JHEP **02**, 006 (2014), 1309.7038.

[19] V. Barger, T. Han, P. Langacker, B. McElrath, and P. Zerwas, Phys. Rev. **D67**, 115001 (2003), hep-ph/0301097.

[20] H. Abramowicz *et al.*, (2016), 1608.07538.

[21] B. K. Bullock, K. Hagiwara, and A. D. Martin, Phys. Lett. **B273**, 501 (1991).

[22] Y.-S. Tsai, Phys. Rev. **D4**, 2821 (1971), [Erratum: Phys. Rev.D13,771(1976)].

[23] J. Brau *et al.*, (2007).

[24] L. Linssen, A. Miyamoto, M. Stanitzki, and H. Weerts, (2012), 1202.5940.

[25] H. Baer *et al.*, (2013), 1306.6352.

[26] M. Aicheler *et al.*, (2012).

[27] M. Thomson, Nucl.Instrum.Meth. **A611**, 25 (2009), 0907.3577.

[28] J. S. Marshall, A. MÍźnnich, and M. A. Thomson, Nucl. Instrum. Meth. **A700**, 153 (2013), 1209.4039.

[29] I. G. Knowles and G. D. Lafferty, J. Phys. **G23**, 731 (1997), hep-ph/9705217.

[30] M. Green, *Electron-Positron Physics at the Z*Studies in high energy physics, cosmology, and gravitation (Taylor & Francis, 1998).

[31] H. Abramowicz *et al.*, (2013), 1306.6329.

[32] Linear Collider ILD Concept Group -, T. Abe *et al.*, (2010), 1006.3396.

[33] CALICE, C. Adloff, (2011), 1105.0511.

[34] B. Parker *et al.*, (2009).

[35] CALICE, JINST **7**, P04015 (2012), 1201.1653.

[36] 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).

[37] W. Kilian, T. Ohl, and J. Reuter, European Physical Journal C **71** (2011).

[38] M. Moretti, T. Ohl, and J. Reuter, p. 1981 (2001), hep-ph/0102195.

[39] G. Altarelli, R. Kleiss, and C. Verzegnassi, editors, *Z PHYSICS AT LEP-1. PROCEEDINGS, WORKSHOP, GENEVA, SWITZERLAND, SEPTEMBER 4-5, 1989. VOL. 3: EVENT GENERATORS AND SOFTWARE*, 1989.

[40] T. Sjostrand, (1995), hep-ph/9508391.

[41] OPAL, G. Alexander *et al.*, Z. Phys. **C69**, 543 (1996).

[42] S. Jadach, Z. Was, R. Decker, and J. H. Kuhn, Comput. Phys. Commun. **76**, 361 (1993).

[43] GEANT4, S. Agostinelli *et al.*, Nucl.Instrum.Meth. **A506**, 250 (2003).

[44] P. Mora de Freitas and H. Videau, p. 623 (2002).

[45] F. Gaede, Nucl. Instrum. Meth. **A559**, 177 (2006).

[46] J. S. Marshall and M. A. Thomson, Eur. Phys. J. **C75**, 439 (2015), 1506.05348.

[47] J. S. Marshall, Presentation on pandorapfa with lc reconstruction, https://github.com/PandoraPFA/Documentation/blob/master/Pandora_LC_Reconstruction.pdf, 2017.

[48] A. Sailer, Luminosities for ee, eg, and gg interactions, https://indico.cern.ch/event/233706/contributions/499053/attachments/390186/542711/130514_LuminosityNormalisation.pdf, 2013.

[49] D. Schulte, (1999).

[50] T. Barklow, D. Dannheim, M. O. Sahin, and D. Schulte, (2012).

[51] G. F. Sterman and S. Weinberg, Phys. Rev. Lett. **39**, 1436 (1977).

[52] S. Moretti, L. Lonnblad, and T. Sjostrand, JHEP **08**, 001 (1998), hep-ph/9804296.

[53] G. P. Salam, Eur. Phys. J. **C67**, 637 (2010), 0906.1833.

[54] A. Ali and G. Kramer, Eur. Phys. J. **H36**, 245 (2011), 1012.2288.

[55] M. Cacciari, G. P. Salam, and G. Soyez, Eur. Phys. J. **C72**, 1896 (2012), 1111.6097.

[56] M. Cacciari and G. P. Salam, Phys. Lett. **B641**, 57 (2006), hep-ph/0512210.

[57] S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber, Nucl. Phys. **B406**, 187 (1993).

[58] S. D. Ellis and D. E. Soper, Phys. Rev. **D48**, 3160 (1993), hep-ph/9305266.

[59] S. Catani, Y. L. Dokshitzer, M. Olsson, G. Turnock, and B. R. Webber, Phys. Lett. **B269**, 432 (1991).

[60] M. Battaglia and F. P., CERN Report No. LCD-Note-2010-006, 2010 (unpublished).

[61] A. Hocker *et al.*, PoS **ACAT**, 040 (2007), physics/0703039.

[62] Y. Freund and R. E. Schapire, Journal of Computer and System Sciences **55**, 119 (1997).

[63] 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).

[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] 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.

[66] E. Segrè, *Nuclei and particles: an introduction to nuclear and subnuclear physics* (W. A. Benjamin, 1977).

[67] E. Longo and I. Sestili, Nucl. Instrum. Meth. **128**, 283 (1975), [Erratum: Nucl. Instrum. Meth.135,587(1976)].

[68] ALEPH collaboration, S. Schael *et al.*, Phys. Rept. **421**, 191 (2005).

[69] S. Berge, W. Bernreuther, and S. Kirchner, Phys. Rev. **D92**, 096012 (2015).

[70] DELPHI collaboration, P. Abreu *et al.*, Phys. Lett. **B267**, 422 (1991).

[71] E. Farhi, Phys. Rev. Lett. **39**, 1587 (1977).

[72] F. Gaede and J. Engels, EUDET Report (2007).

[73] TMVA Core Developer Team, J. Therhaag, AIP Conf.Proc. **1504**, 1013 (2009).

[74] D. Lyth, Journal de Physique Colloques **35**, C2 (1974).

[75] S. Dittmaier *et al.*, (2012), 1201.3084.

[76] A. MÍźnnich, CERN Report No. LCD-Note-2010-009, 2010 (unpublished).

[77] CLICdp, A. Sailer and A. Sapronov, (2017), 1702.06945.

[78] 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.

[79] T. Suehara and T. Tanabe, Nucl. Instrum. Meth. **A808**, 109 (2016), 1506.08371.

[80] LCFI, D. Bailey *et al.*, Nucl. Instrum. Meth. **A610**, 573 (2009), 0908.3019.

[81] H. Aihara *et al.*, (2009), 0911.0006.

[82] G. Hanson *et al.*, Phys. Rev. Lett. **35**, 1609 (1975).

[83] A. Buckley, The hepthesis LaTeX class.

List of Figures

List of Tables